XiaLab Analytics

Empowering researchers through trainings, tools and AI

https://www.xialab.ca • contact@xialab.ca

Omics Data Science Training Course

Winter 2024

Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:40	NGS omics
10:45 – 10:15	MS omics
10:20 – 10:55	Enrichment analysis
11:00 – 11:15	Pathway & network analysis
Discussion & Next Lecture	

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
Poland
Belarus
Germany
Ukraine
Romania
Italy
Spain
Portugal
Greece
Türkiye
Kazakhstan
Mongolia
Kyrgyzstan
Turkmenistan
Syria
Iraq
Iran
Afghanistan
Pakistan
China
South Korea
Japan
Sea of Japan
Sea of Okhotsk
East China Sea
Tunisia
Morocco
Algeria
Libya
Egypt
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Saudi Arabia
Oman
Yemen
Gulf of Aden
Arabian Sea
Nepal
India
Myanmar (Burma)
Thailand
Vietnam
Bay of Bengal
Laccadive Sea
Malaysia
Indonesia
Philippines
Philippine Sea
Burkina Faso
Guinea
Ghana
Nigeria
Gulf of Guinea
Ethiopia
Somalia
Kenya
Gabon
DRC
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
South Africa
Indian Ocean
Australia
Papua New Guinea
Banda Sea
Arafura Sea
Great Australian Bight

Our Resources

Recordings & slides

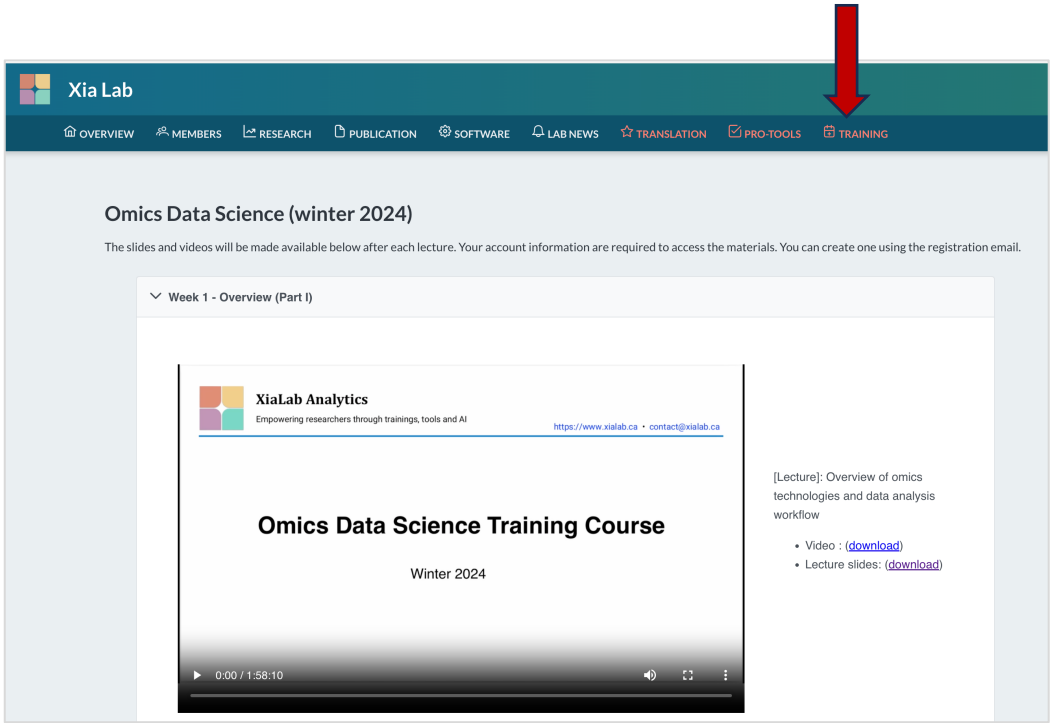
- https://www.xialab.ca under the "**Training**" tab within 3 days after lecture
- Faster with Firefox

Community tool:

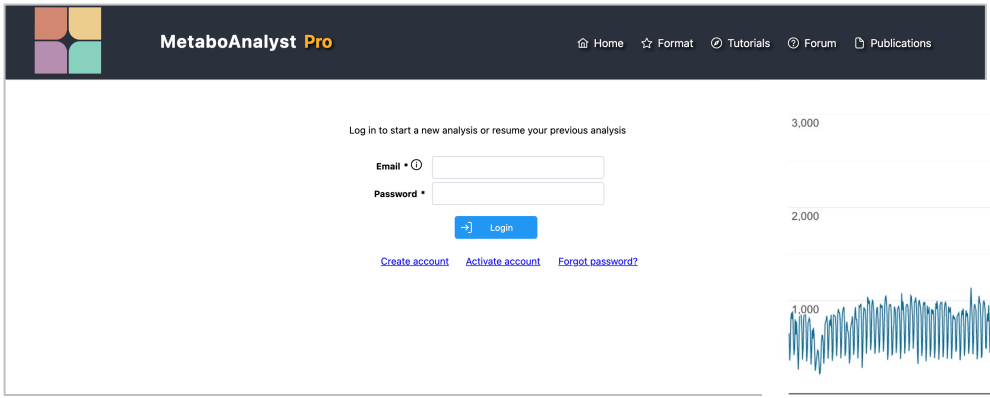
- https://www.####.ca

Pro Tools:

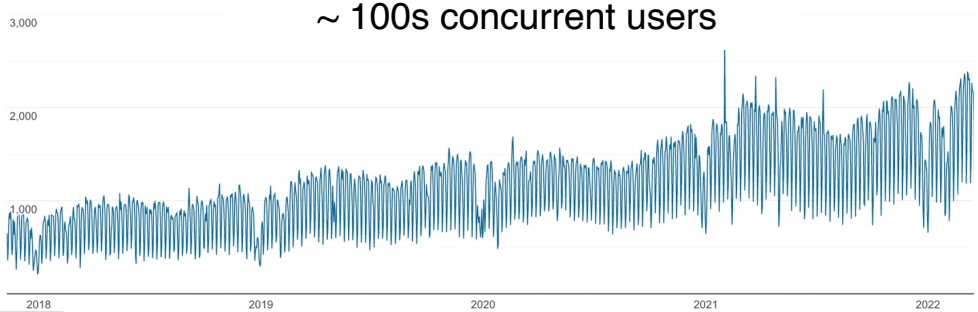
- https://pro.####.ca
  - You will be assigned to one of the three nodes



More on the “pro” tools



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



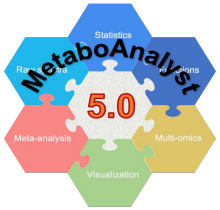
A total of six tools have their “pro” versions

- 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. More features
3. More stable
4. Better support
5. Towards AI (coming …)

Raw data ➔ statistics ➔ functions

LC-MS spectra

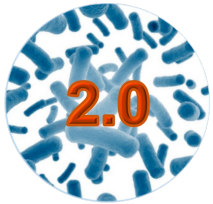


MetaboAnalyst

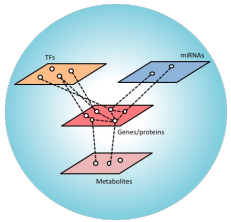
Bulk RNA-seq

ExpressAnalyst

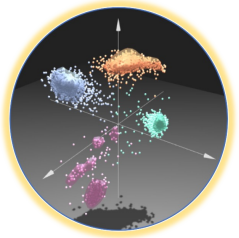
Microbiome



MicrobiomeAnalyst



OmicsNet



OmicsAnalyst

Towards conversational analytics

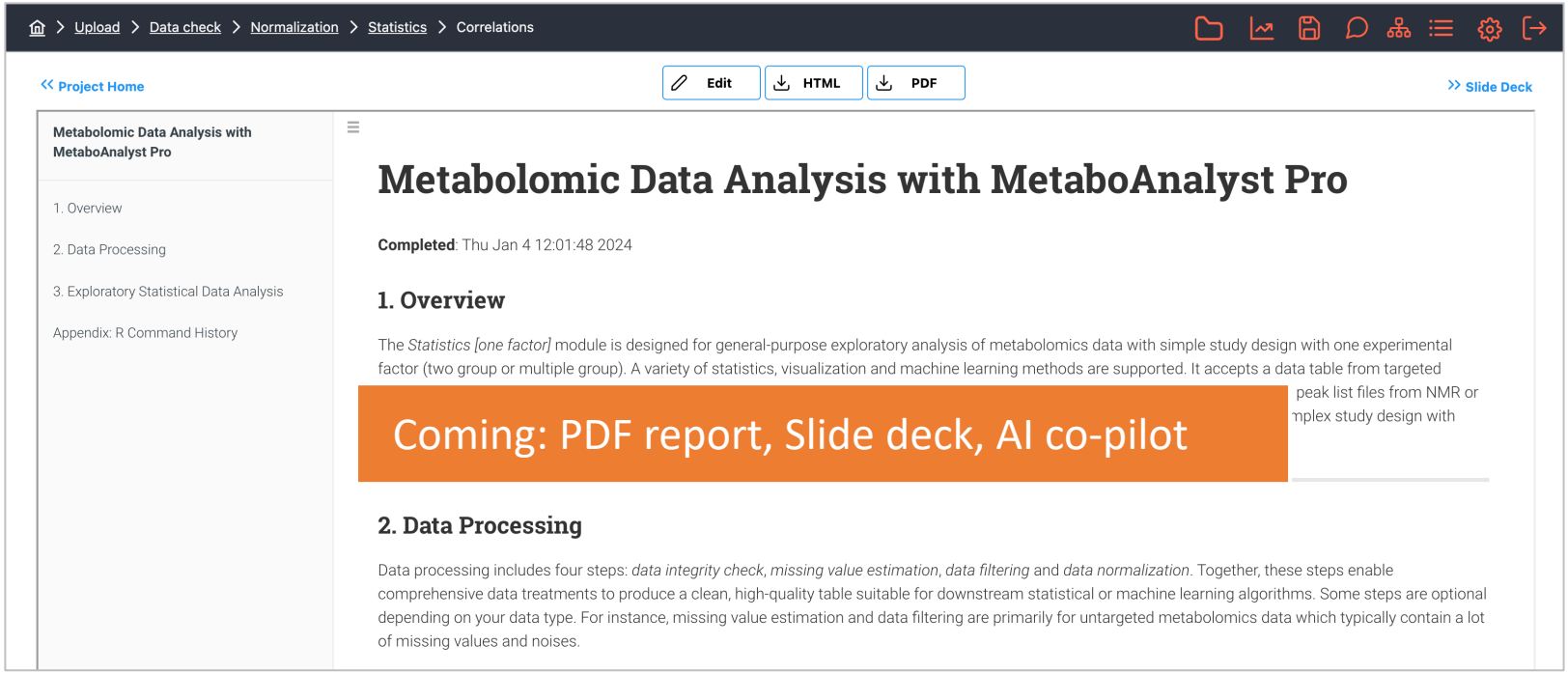
You

1. Drop your data
2. Provide data information
3. Describe analysis goal
7. Approve plan
10. Review report

OmicsBot

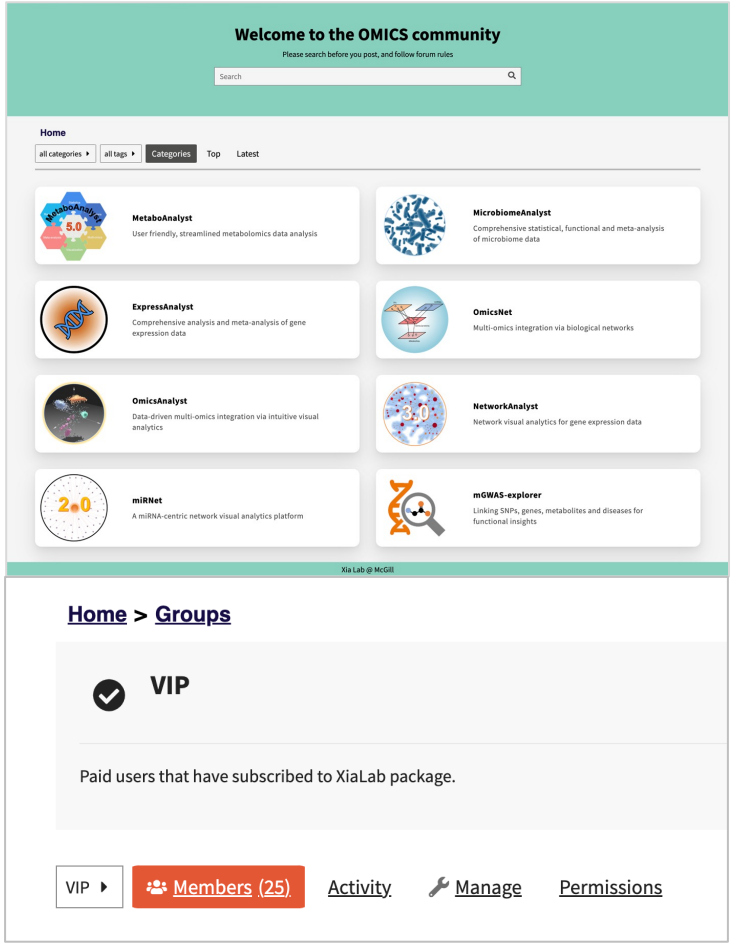
4. Confirm data format
5. Confirm analysis goal
6. Propose analysis plan
8. Execute plan
9. Generate report

Project Saving & Live Report



Our Support

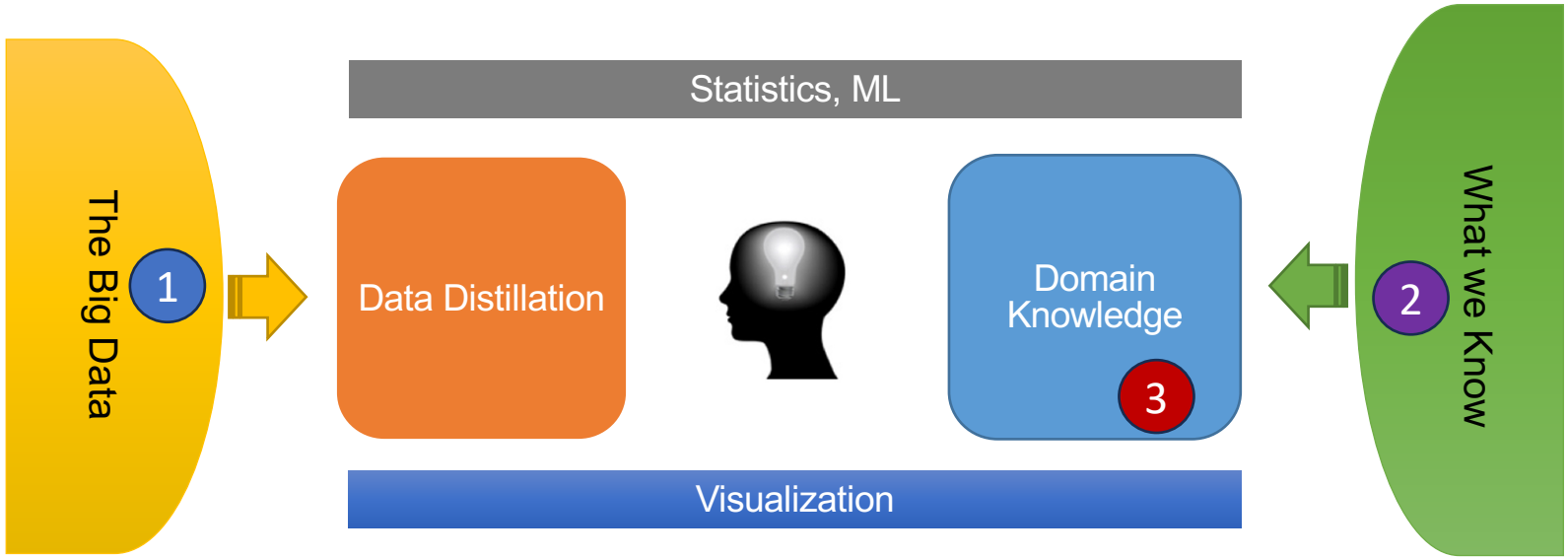
- User forum
  - https://omicsforum.ca
  - Please register with subscription email! We will add a **VIP** label associated with your account
- Through email
  - support@xialab.ca
- We need sufficient information to reproduce the issue.
  1. Analysis goal
  2. Tool + input data
  3. Steps + parameters



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

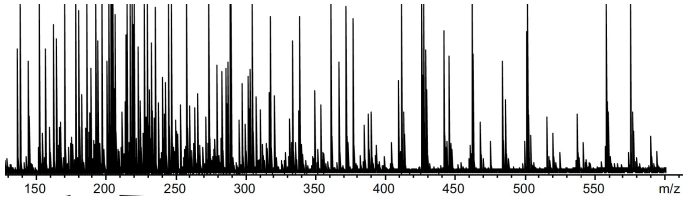
The “big picture” of omics data analysis



Two main types of raw omics data

1. NGS-omics: next-generation sequencing (NGS) data
   - Genomics
   - Epigenomics
   - Transcriptomics
2. MS-omics: mass spectrometry (MS) data
   - Proteomics
   - Metabolomics
   - Exposomics

Reads

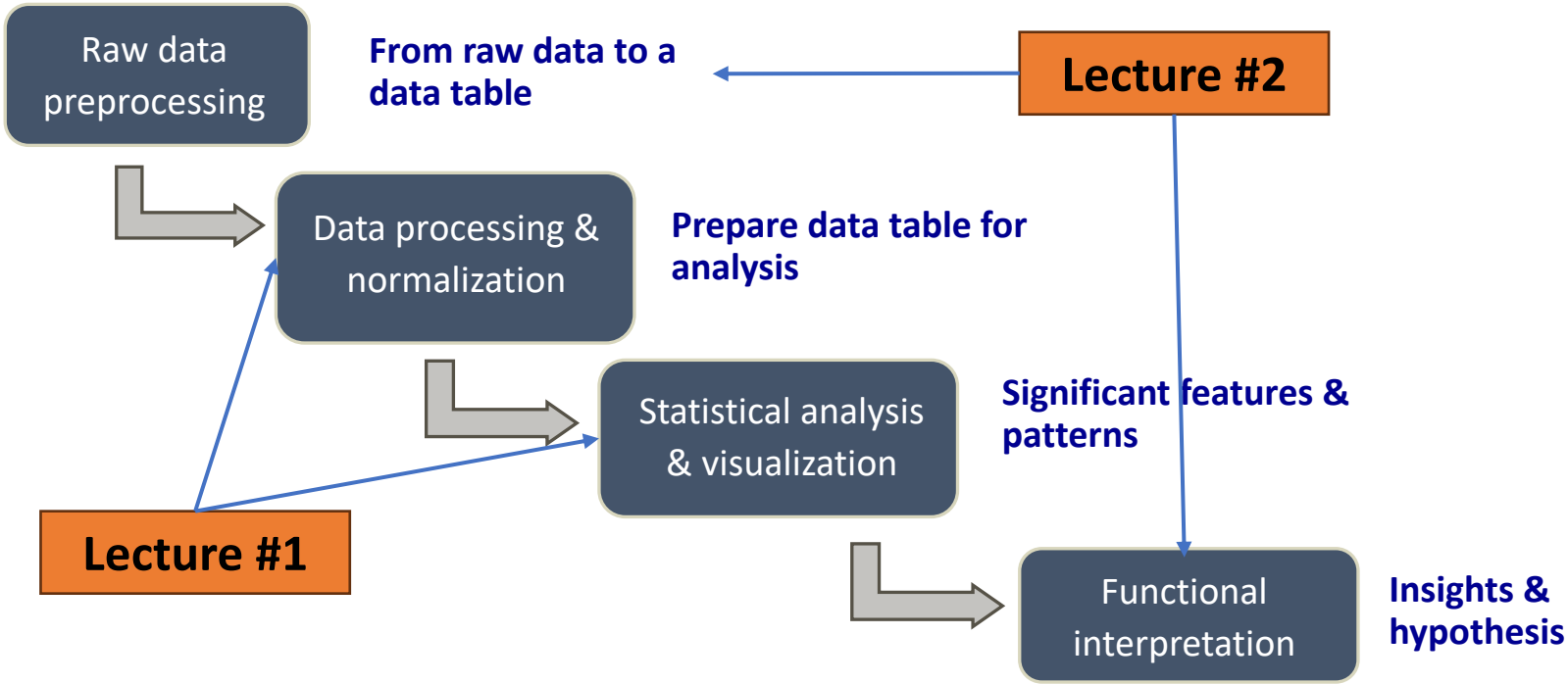


Spectra

I have spent > $10,000 on sequencing (or metabolomics), and we have received > 100s GB data

Now What?

Omics data analysis in a nutshell



Task #1: from raw data to data table

FASTQ

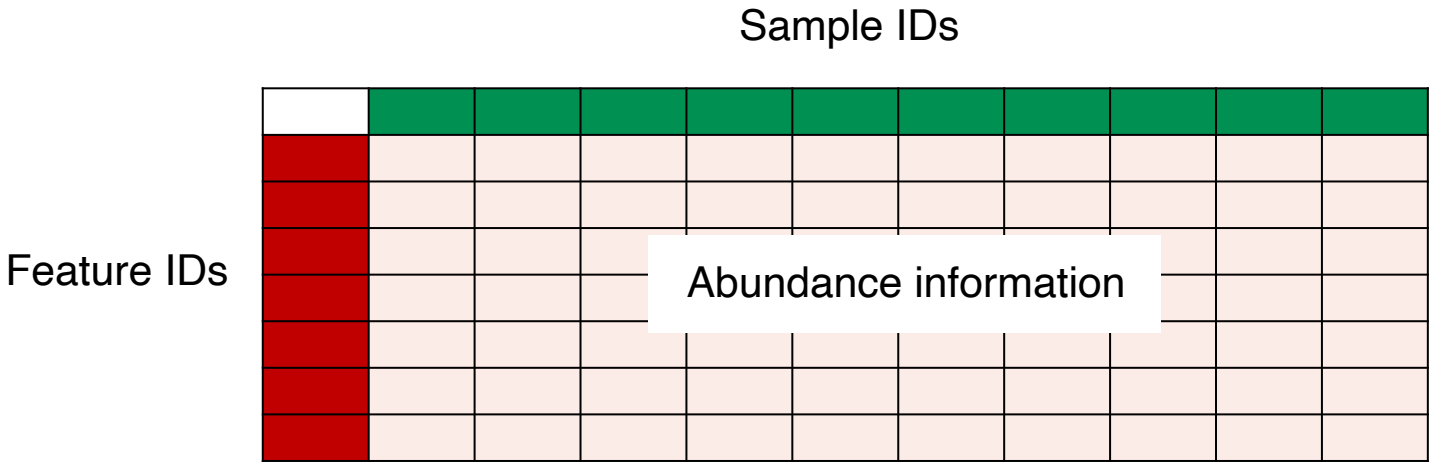
Plekhg1	485	462	467	562	347	296	458	487
Plekhg6	126	172	206	198	137	164	110	152
4930592I03R	7	5	9	7	2	7	0	7
Plekhg4	0	0	0	0	0	0	0	0
Plekhg5	117	150	138	95	83	109	105	91
Nsa2	1540	1472	1577	1792	1419	1580	1288	1670
Nampt	2882	3103	3487	3090	2400	3148	2343	2569
Vmo1	43	68	51	62	24	30	22	26
Man2a1	11979	13268	15988	14592	7416	9708	7285	10071

mzML

Data Table

200.1/2926	147887.53	451600.71	65290.38	56540.93	175177.08	82619.48	51951.61
205/2791	1778569	1567038	1482796	1039130	1950287	1466781	1572679
206/2791	237993.6	269714	201393.4	150107.3	276541.8	222366.2	211717.7
207.1/2719	380873	460629.7	351750.1	219288	417169.6	324892.5	277990.7
219.1/2524	235544.92	173623.38	82364.59	79480.4	244584.47	161184.05	72029.38
231/2516	117649.77	48960.63	222609.07	286232.15	465898.01	61234.44	96841.46
233/3023	399145.3	356951.3	410550.7	198416.5	397107.8	271252.1	334459.9
234/3024	76880.87	99526.27	97493.76	53461.71	65215.64	55952.44	73781.01
235.1/2695	171995.22	128945.16	155442.48	115286.25	199981.49	30028.6	156968.3
236.1/2524	252282.04	206031.93	71763.79	73602.47	253791.07	187225.65	79389.63

What’s inside a table ……



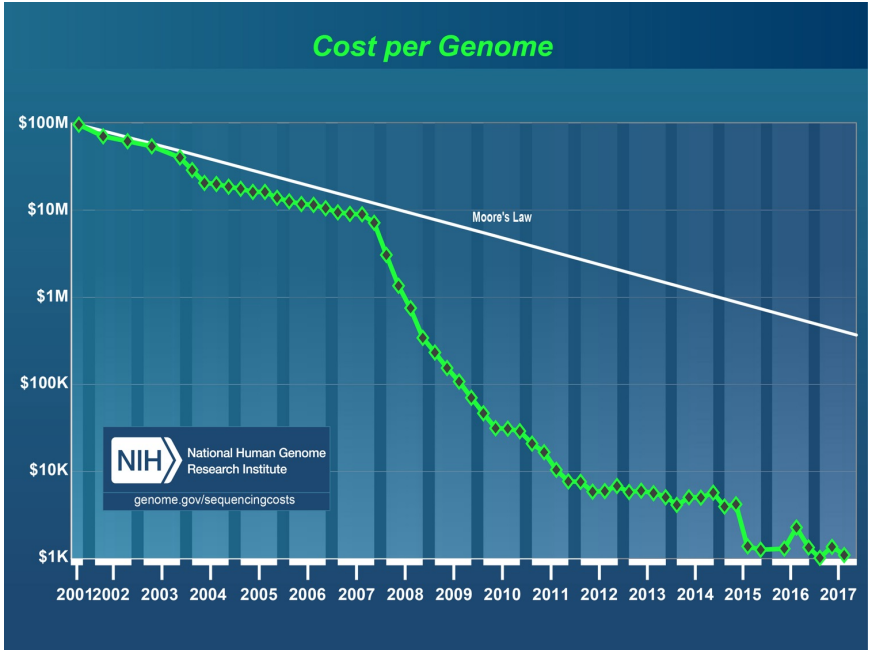
Raw Data Processing =
Feature **Identification & Quantification** across samples

Schedule for today

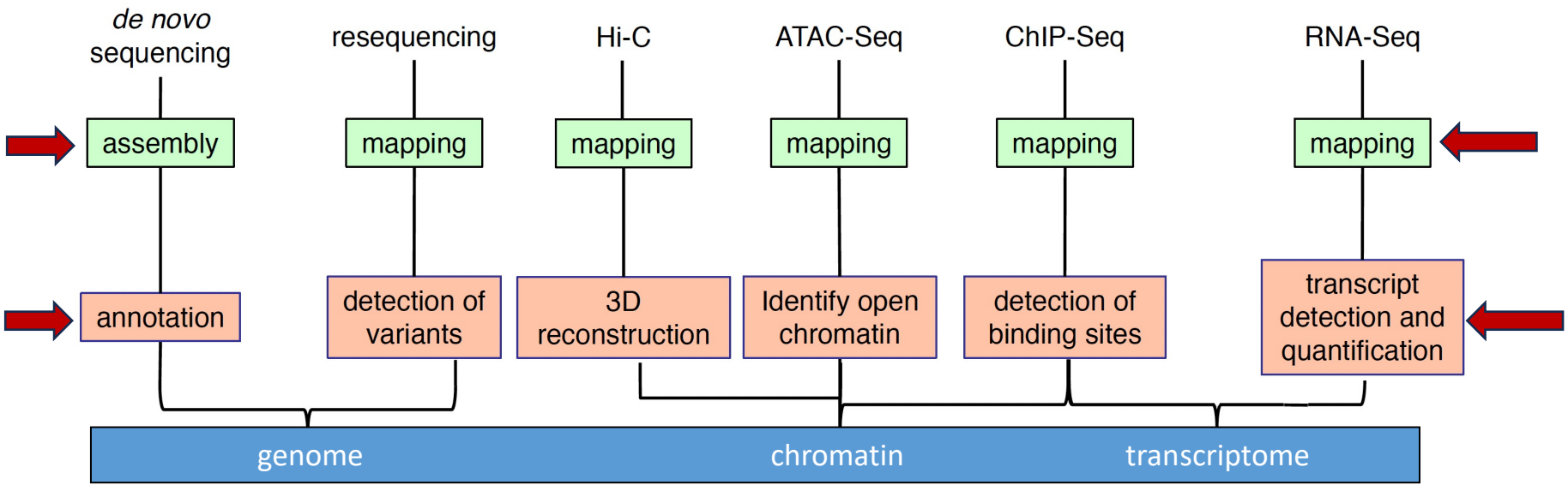
Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:40	NGS omics
10:45 – 10:15	MS omics
10:20 – 10:55	Enrichment analysis
11:00 – 11:15	Pathway & network analysis
Discussion & Next Lecture	

Next Generation Sequencing (NGS)

- **Deprecated**
  - 454
  - SOLiD
- **Supported, not used much in genome assembly**
  - Ion Torrent (Ion PGM)
  - Ion Proton
- **Current workhorses**
  - Illumina (short-reads)
  - Pacific Biosciences (long reads)
- **Third gen**
  - Oxford Nanopore
  - 10x Genomics - GemCode



Applications of NGS



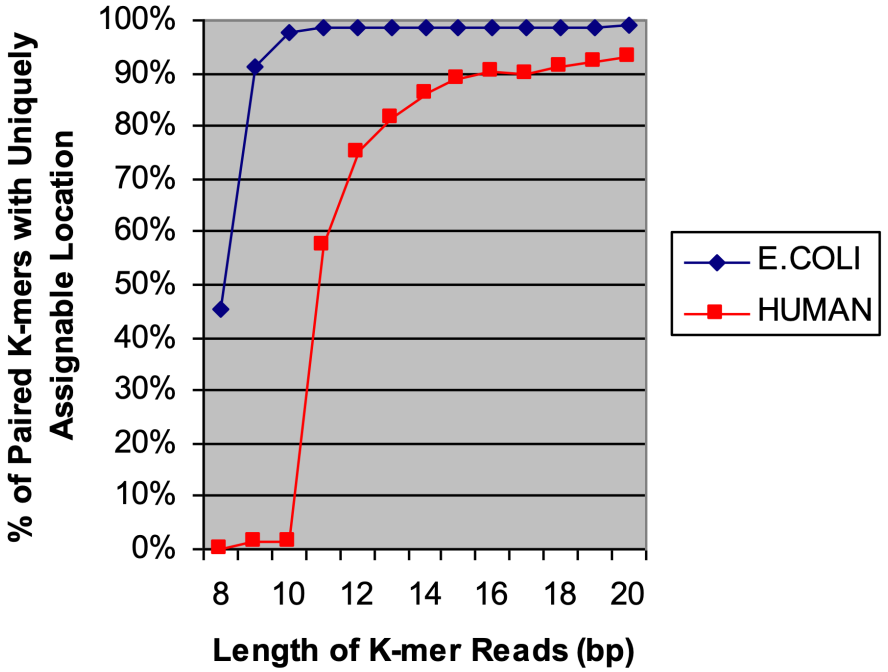
Short Reads vs Long Reads

Long reads is for genome de novo assembly (no reference genome) or detecting structural variants

- Novel species

Short reads is for genome **mapping**

- Already has a reference genome
- Detection & quantification
  - ✓ RNA-seq
  - ✓ Metagenomics



We will discuss **k-mer** later

Illumina Systems

Pros:

- Huge yield, cheap, reliable
- Read length “long enough” (100-300 bp)
- Industry standard=huge amount of available software

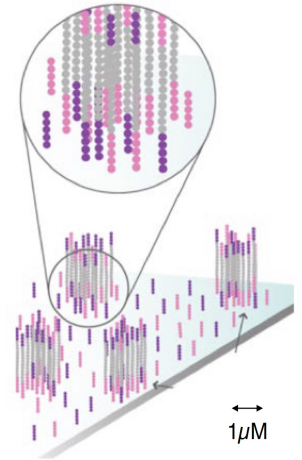
Cons:

- GC-problems, quality-dip at end of reads,
- Long running time for Hi-Seq, short insert-sizes

Key Innovation

Massively parallel detection on immobilized “molecular colonies”

1. Using PCR to grow clusters of a DNA template (no cloning with bacteria or yeast!);
2. Imaging while sequencing by synthesis (SBS) in each cycle



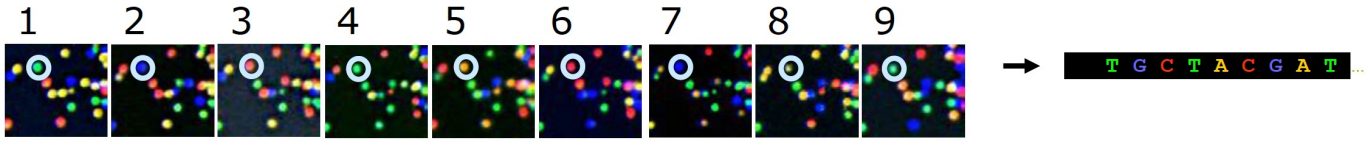
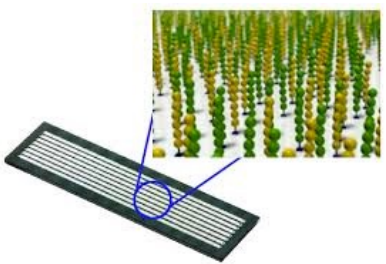
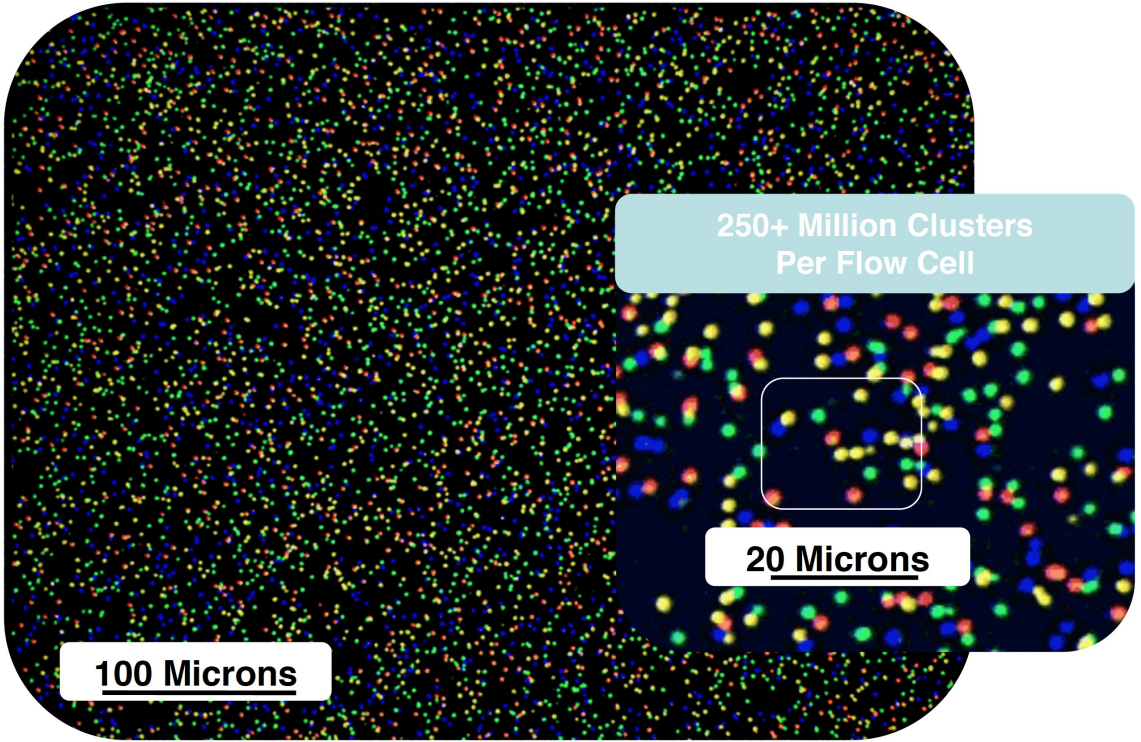


Image acquisition **Base calling**

Massively Parallel



Illumina Flow Cells



What the output look like (fastq)?

Developed at the Wellcome Trust Sanger Institute to bundle a FASTA formatted sequence and its quality scores

Show the top 50 lines in a fastq file
> head –n 50 ####.fastq

```
jeffxia@Jeffs-MacBook-Pro-2 testdata % head -n 50 D2.CE2-H2-LT_R1.fastq
@A00266:275:HLFTWDSXX:2:2640:17562:28651 1:N:0:CCAGTATC+AGCGAGAT
AGGGAGGATCATGAACTCCTGCATAGCCAGCTTATTGCCAGCATGGGAGCCACCATTGATCACATTGAAAGCAGGAACTGGAAGAATGACTTCTGCATTCC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:1528:28257:14137 1:N:0:CCAGTATC+AGCGAGAT
CTTACCTCTCTCTCTCTGCCAGCCCCACACACGCTGTAAAACTGACTCCCTTTAGCTCGGAAAACTTTTACAGATACTCAGCGTGAAACTAGGGCCGGCGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:2602:14443:5368 1:N:0:CCAGTATC+AGCGAGAT
TTCCTTCTCTCCCAGGAGTACCTGGTTGTCCTGGGAAACCAGGATCTCCCTTTTCACCTTTCAGGATAATATCATAAGAAAAATCTCCTGGCTCCCCTTTT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:1101:4580:1000 1:N:0:CCAGTATC+AGCGAGAT
TGCTGTTCTACTTGATGATGGAAAAGAGAAAATAGCATAACCAAAAAATCCCAGTGGACGGTGGAAGCCGCCTTCCACGTGGTACCAGGTCACTTTTACGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:1335:6867:32863 1:N:0:CCAGTATC+AGCGAGAT
GGAGCGCTGGGCTGACAGCAGCCGGAGGGGACGTGCCCCTGAACGACGAGGCTGAACCGACAGCTACTGGTGGAGATGAACAGTAAGGACAAAACTAAACG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FF
@A00266:275:HLFTWDSXX:2:2338:3513:27524 1:N:0:CCAGTATC+AGCGAGAT
GGGCCATCAACTGAAATCTTCCAAAACACACGGCGATGAATTGTCCAGGAGATGATAGTATTTCTCAGGAAATGCTTCATTGAAGGTAGAGCTTGGAGAAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFF
@A00266:275:HLFTWDSXX:2:1678:19162:1611 1:N:0:CCAGTATC+AGCGAGAT
CAGCAGGAACAGTGCGGAGAAGAGTAGTGACAGTTGCTACAGCTACATCCTGAGGGGTGTACTTCATGGGGCAGGAATGCCCAGGTGTCACCATGTTGGGT
+
FF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFF
@A00266:275:HLFTWDSXX:2:2223:27814:19570 1:N:0:CCAGTATC+AGCGAGAT
GGGGGTTGGCATGGAAGCTCCATCATCAGATGTCATTCCAAACAGGTTAAACAAATGTCTTCTGTCAATAGCTGATCCTGCCTTAGGGTACTAAGTACAGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFF:FFFFFF:FFFFFFFF:FF
@A00266:275:HLFTWDSXX:2:1660:15881:8609 1:N:0:CCAGTATC+AGCGAGAT
TGATAGGTCAAGGATTTGTCACTTGGGTCGTCTCCAAAATCAAAGCCATCAAAGGCATCCCCAGCTTCACCACCAAGAAAGTAGGCATAAGTGAGTCGATA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:2624:4906:4820 1:N:0:CCAGTATC+AGCGAGAT
GGCCGCCACCGCCACCGCCCCAGGGTCAGGCTGCAGCAACTGGGATGAGCTGATGTAGCTGGCTCTGCCCGCCCCGGCTTCCATATGCCTGGTGGCCTCCG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

How to understand fastq file?

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATTTGTTGGGGGAGACATTTTTGTGATTGCCTTGAT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BBBBBBBBBBBBBBBBBBBBBB
```

'@' followed by a sequencer Identifier

The **sequence**

'**+**', optionally followed by a sequence Identifier

The **quality scores**

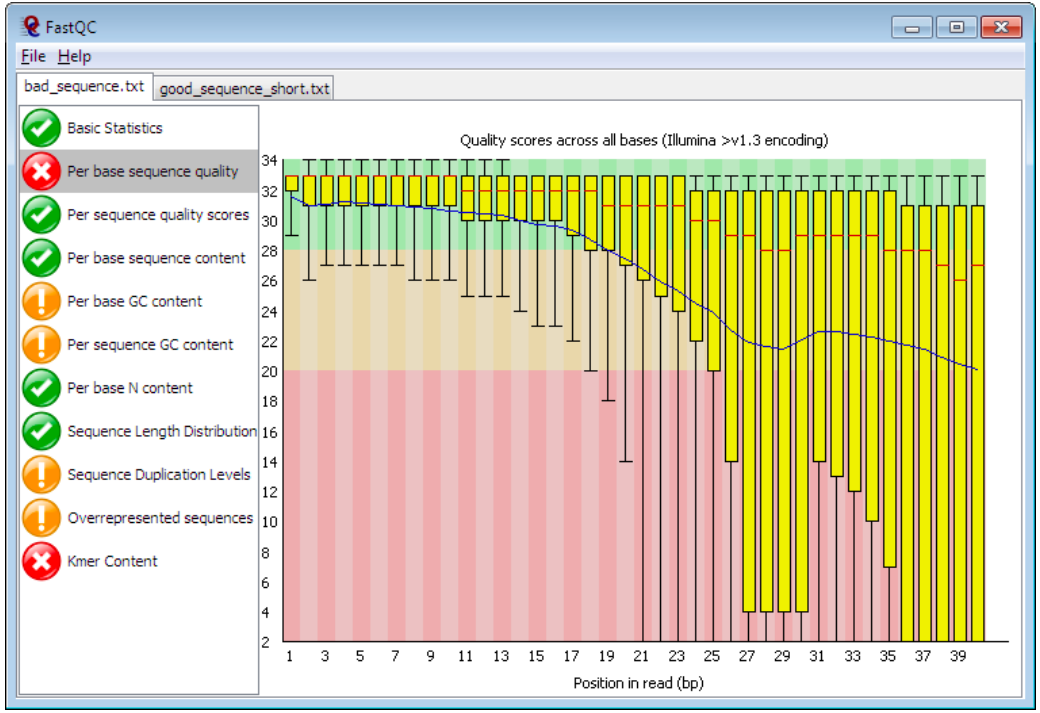
What is a base quality score?

$Q = -10 \log_{10} p$

- p is the probability that the base call is wrong

Base Quality (Q)	P_{error}(obs. base)
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %

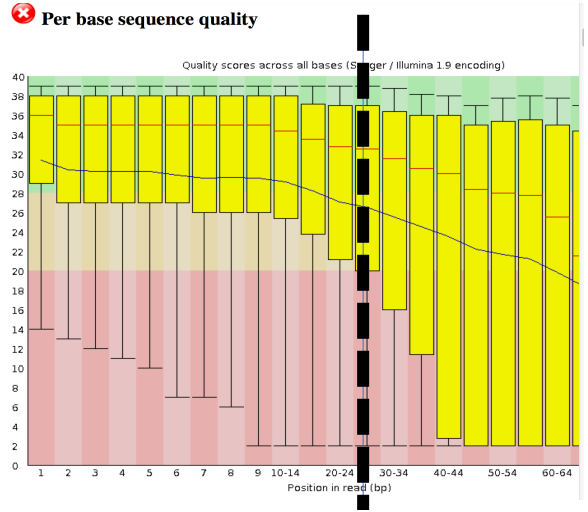
Quality checking (FastQC)



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Trimming based on quality scores

- **fastp**
  - ✓https://github.com/OpenGene/fastp
- **Trim Galore**
  - ✓https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
- **FASTX**
  - ✓http://hannonlab.cshl.edu/fastx_toolkit



fastp: an ultra-fast all-in-one FASTQ preprocessor

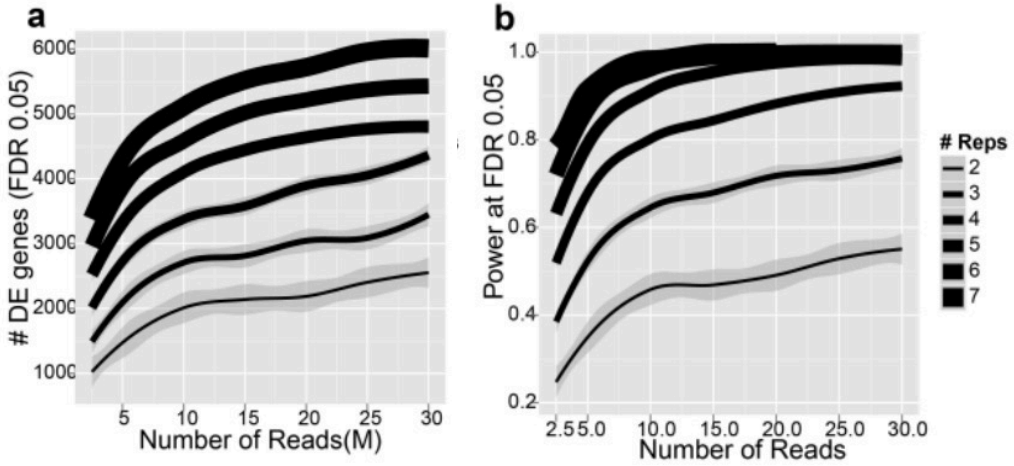
by S Chen · 2018 · Cited by 10270 — **fastp** is a versatile tool that can perform quality profiling, read filtering, read pruning, adapter trimming, polyG/polyX tail trimming, UMI ...

Abstract · Introduction · Materials and methods · Results

Sequencing types & depth

Single end (SE) read

Paired-end (PE) read
Forward & Reverse sequencing



Bioinformatics (2013) doi: 10.1093/bioinformatics/btt688

For human studies (most cost effective):
5 replicates, 10M reads, single-end

Multiplexing to reduce cost

Due to massive parallel sequencing, the reads length in one illumina run can cover a bacterial genome >1000x times!

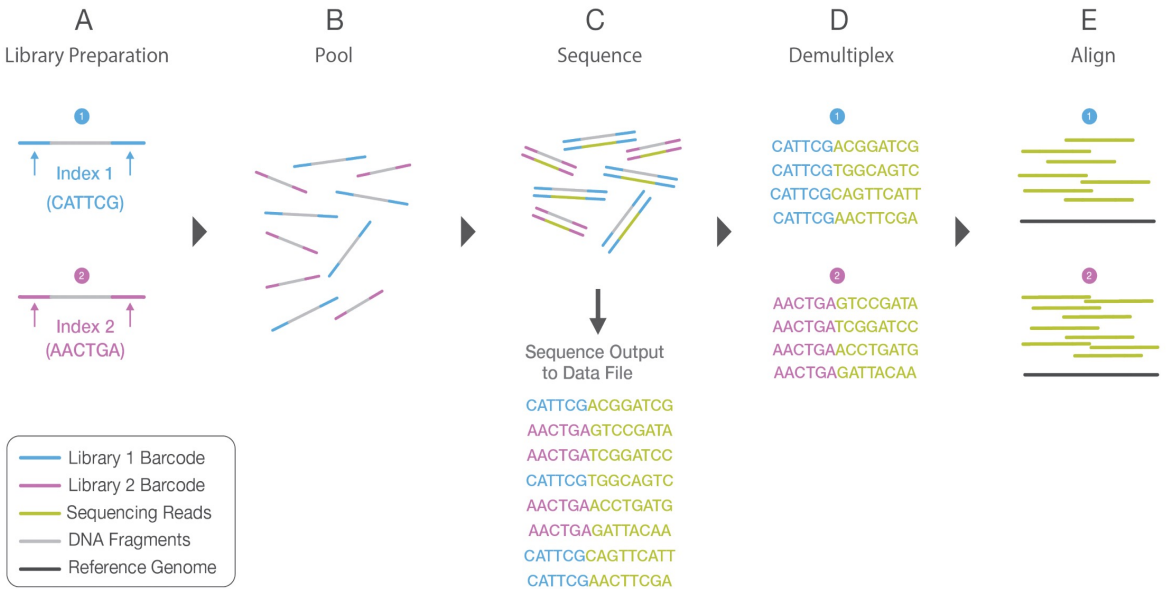
- 30~40X coverage for genome assembly

Multiplexing

- Simultaneously measures multiple samples/genomes in a single run
- Primer (adaptor key in the end) + tag + template specific primer + sequence
  - Adaptor contains a biotin tag and is needed during sequencing of the template
  - Tag allows discrimination between samples (barcoding)
  - Primer is needed for amplification of the sequence

Example: TCAGTACTCGGCCTACGGGAGGCAGCAG

Multiplexing & de-multiplexing



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Task overview

NGS sequence file

```
jeffxia@Jeffs-MacBook-Pro-2 testdata % head -n 50 D2.CE2-H2-LT_R1.fastq
@A00266:275:HLFTWDSXX:2:2640:17562:28651 1:N:0:CCAGTATC+AGCGAGAT
AGGGAGGATCATGAACTCCTGCATAGCCAGCTTATTGCCAGCATGGGAGCCACCATTGATCACATTGAAAGCAGGAACTGGAAGAATGACTTCTGCATTCC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:1528:28257:14137 1:N:0:CCAGTATC+AGCGAGAT
CTTACCTCTCTCTCTCTGCCAGCCCCACACACGCTGTAAAACTGACTCCCTTTAGCTCGGAAAACTTTTACAGATACTCAGCGTGAAACTAGGGCCGGCGT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:2602:14443:5368 1:N:0:CCAGTATC+AGCGAGAT
TTCCTTCTCTCCCAGGAGTACCTGGTTGTCCTGGGAAACCAGGATCTCCCTTTTCACCTTTCAGGATAATATCATAAGAAAAATCTCCTGGCTCCCCTTTT
+
FFFFFFFFFFFFFFFF                                                                                FFFFFF:FFFFFFFFFFFF
@A00266:275:HLFTW                                                                                GTACCAGGTCACTTTTACGT
+
FFFFFFFFFFFFFFFF                                                                                FFFFFFFFFFFFFFFFFFFF
@A00266:275:HLFTW                                                                                AGTAAGGACAAAACTAAACG
+
FFFFFFFFFFFFFFFF                                                                                FFFFFFFFFFFFFFFF:FF
@A00266:275:HLFTW                                                                                GAAGGTAGAGCTTGGAGAAT
+
FFFFFFFFFFFFFFFF                                                                                FFFFFFFFFF:FFFFFFFF
@A00266:275:HLFTW                                                                                CAGGTGTCACCATGTTGGGT
+
FF:FFFFFFFFFFFFF                                                                                FFFFFFFFFFFF:FFFFFF
@A00266:275:HLFTW
GGGGGTTGGCATGGAAGCTCCATCATCAGATGTCATTCCAAACAGGTTAAACAAATGTCTTCTGTCAATAGCTGATCCTGCCTTAGGGTACTAAGTACAGG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFF:FFFFFF:FFFFFFF:FF
@A00266:275:HLFTWDSXX:2:1660:15881:8609 1:N:0:CCAGTATC+AGCGAGAT
TGATAGGTCAAGGATTTGTCACTTGGGTCGTCTCCAAAATCAAAGCCATCAAAGGCATCCCCAGCTTCACCACCAAGAAAGTAGGCATAAGTGAGTCGATA
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00266:275:HLFTWDSXX:2:2624:4906:4820 1:N:0:CCAGTATC+AGCGAGAT
GGCCGCCACCGCCACCGCCCCAGGGTCAGGCTGCAGCAACTGGGATGAGCTGATGTAGCTGGCTCTGCCCGCCCCGGCTTCCATATGCCTGGTGGCCTCCG
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

- Demultiplexed
- Removed adaptor, primers tags
- Trimmed/filtered low-quality base pairs

?

Gene count table

NAME	X454039	X454051	X454057	X454066	X454081	X454090	X454102	X454108
ENSMUSG00000023868	219	49	42	50	6	17	22	21
ENSMUSG00000029538	424	376	191	436	529	1172	914	1014
ENSMUSG00000064141	32	4	4	22	76	165	163	170
ENSMUSG00000020986	198	167	100	203	158	475	406	369
ENSMUSG00000024811	475	255	194	442	384	1365	1148	1219
ENSMUSG00000038781	2	2	2	2	2	5	8	3
ENSMUSG00000022159	96	31	29	59	102	222	154	171
ENSMUSG00000002103	888	505	298	638	541	1570	1351	1517
ENSMUSG00000024608	1504	868	613	1521	1083	3077	2111	4917
ENSMUSG00000022037	57	32	4	12	0	0	1	0
ENSMUSG00000027108	410	323	184	392	396	1145	752	965
ENSMUSG00000033540	359	138	90	279	483	1382	1462	1524
ENSMUSG00000021910	4407	2380	1497	3471	7348	19703	15845	18788
ENSMUSG00000039156	2518	1181	706	1821	383	948	821	966
ENSMUSG00000021959	345	228	134	277	700	1720	1274	1616
ENSMUSG00000020364	0	0	4	0	16	36	16	8
ENSMUSG00000020415	586	958	439	915	212	464	516	707
ENSMUSG00000015340	4749	2689	1880	4618	1582	5098	4652	4618
ENSMUSG00000025875	574	156	101	231	514	1389	1259	1438
ENSMUSG00000017734	122	46	27	64	231	450	441	509
ENSMUSG00000022787	110	66	32	66	178	336	284	248
ENSMUSG00000031834	25	17	15	29	259	700	564	649
ENSMUSG00000002428	85	61	44	109	246	763	642	705
ENSMUSG00000026064	123	79	30	93	31	111	76	96
ENSMUSG00000041420	44	13	5	3	116	214	254	161
ENSMUSG00000027297	0	6	0	5	0	6	0	0
ENSMUSG00000014771	97	76	43	130	102	247	240	291
ENSMUSG00000045980	146	106	48	118	546	1334	1078	1174
ENSMUSG00000025591	547	396	187	506	15	36	41	32

Understanding NGS data

With reference genome/transcriptome (i.e. model organisms)

- Gene mapping (exact)

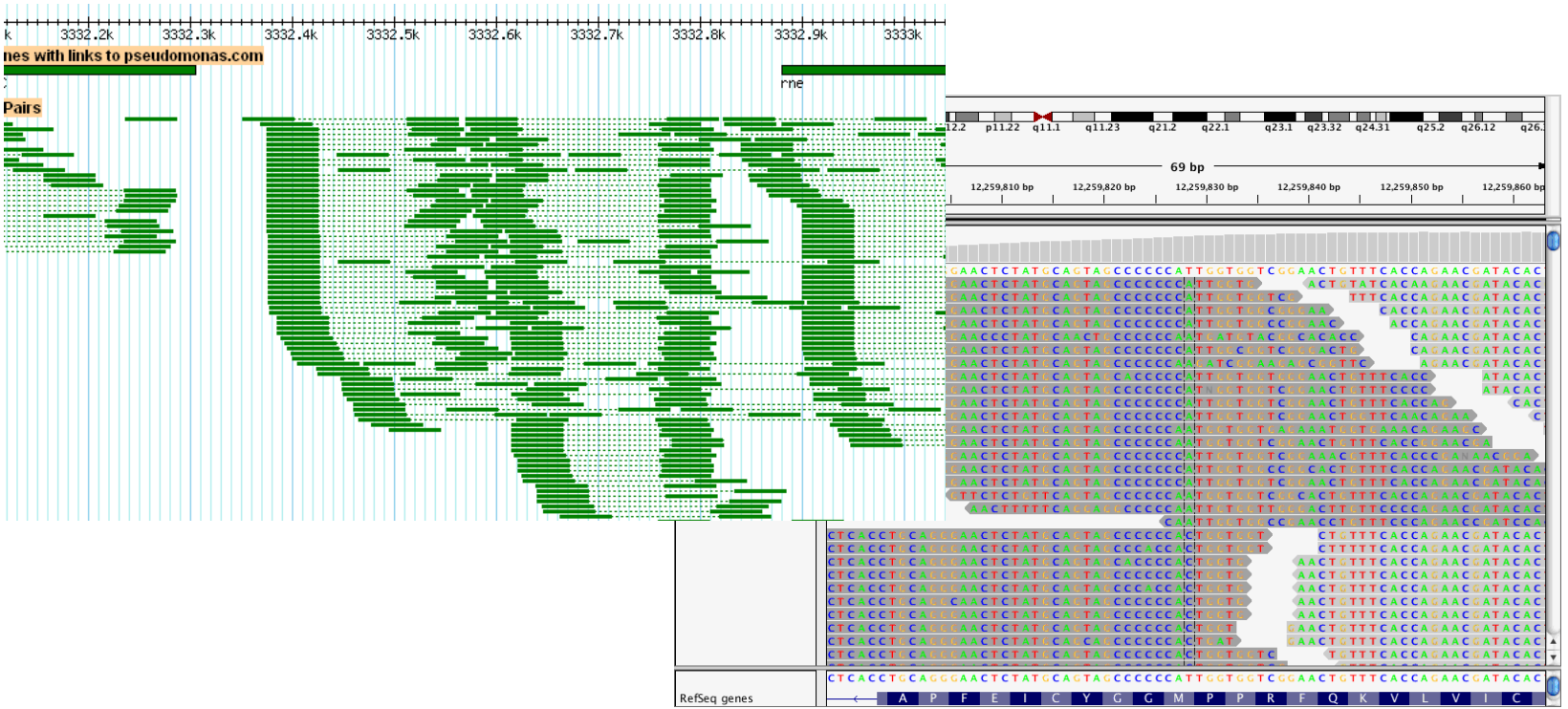
Lecture #3

Without reference genomes (non-model organisms)

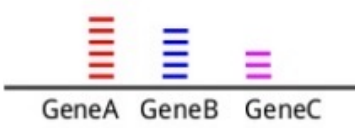
- Ortholog mapping (allow gaps / mismatches)
- It is possible to do genome assembly followed by genome annotation (will touch this later)

Lecture #4

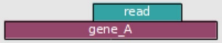
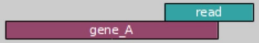
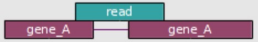
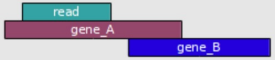
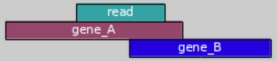
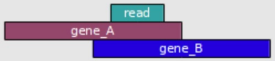
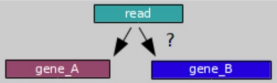
Mapping to genome (a.k.a alignment)



From “mapped reads” to gene counts



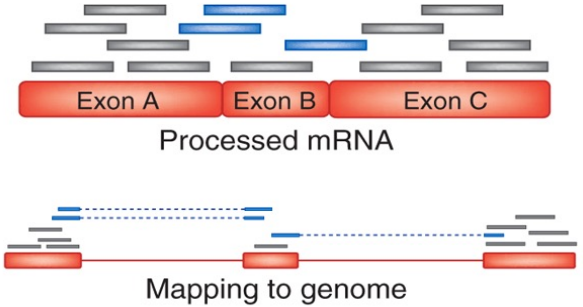
Things can get “complicated”

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

https://htseq.readthedocs.io/en/latest/htseqcount.html

NGS table generation via “reads mapping”

1. Identification
   - Coordinates (locations) define the ID (gene/exon/intron)
2. Quantification
   - Count how many times the reads assigned to the same ID



FASTQ

Plekhg1	485	462	467	562	347	296	458	487
Plekhg6	126	172	206	198	137	164	110	152
4930592I03R	7	5	9	7	2	7	0	7
Plekhg4	0	0	0	0	0	0	0	0
Plekhg5	117	150	138	95	83	109	105	91
Nsa2	1540	1472	1577	1792	1419	1580	1288	1670
Nampt	2882	3103	3487	3090	2400	3148	2343	2569
Vmo1	43	68	51	62	24	30	22	26
Man2a1	11979	13268	15988	14592	7416	9708	7285	10071

K-mer & Genome Assembly

Unique challenges with short-read NGS:

Pros:

Cost effective, high-quality and high-throughput , ideal for detection and quantification

Cons:

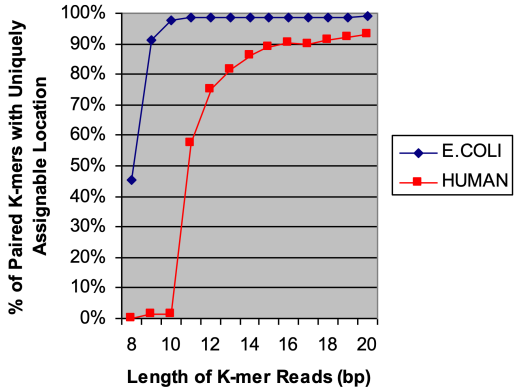
- Short & large amount
- **Computational cost** is very high for both mapping (and genome assembly)

➔ **we want to run data analysis on our own laptop!**

K-mer based approach

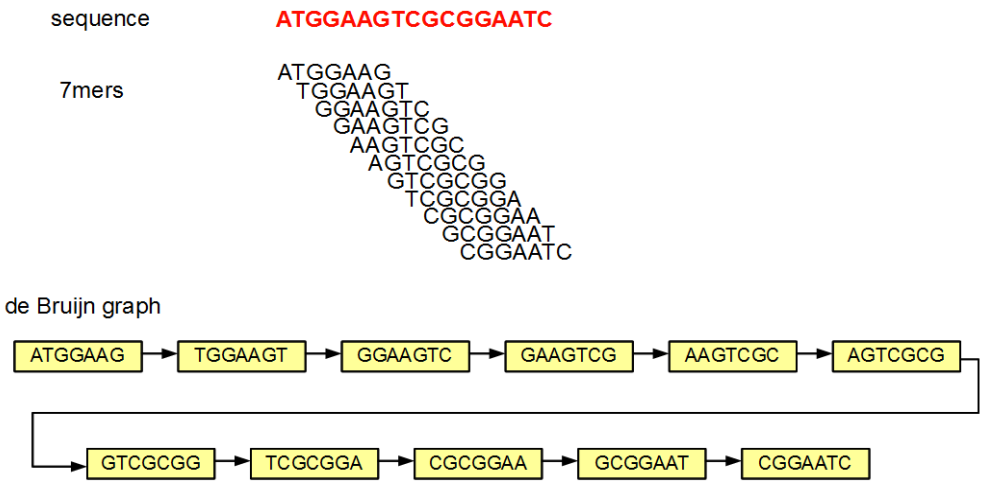
- A k-mer is a substring of length k
- A string of length L has (L-k+1) k-mers
  - A moving window
  - Example read L=8 and k= 4

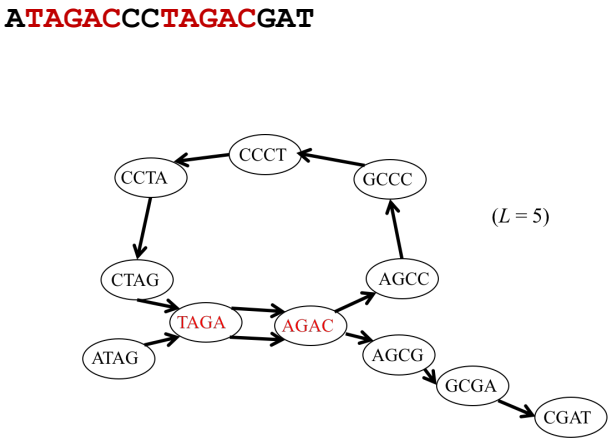
```
AGATCCGT
AGAT
  GATC
    ATCC
      TCCG
        CCGT
```



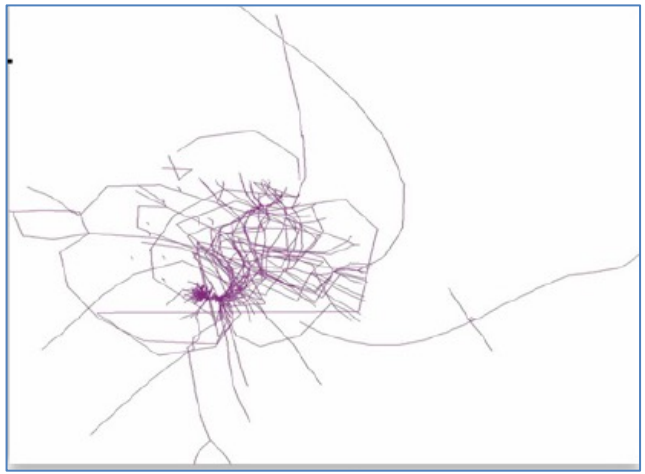
- The **frequency** of a set of k-mers can be used as a "signature" of the underlying sequence.
- Comparing these frequencies is computationally easier.
- Widely used in both genome assembly and mapping (“pseudo-”alignment)

K-mer – “natural” for genome assembly





Visualize the assembly



Need to explore a range of k values

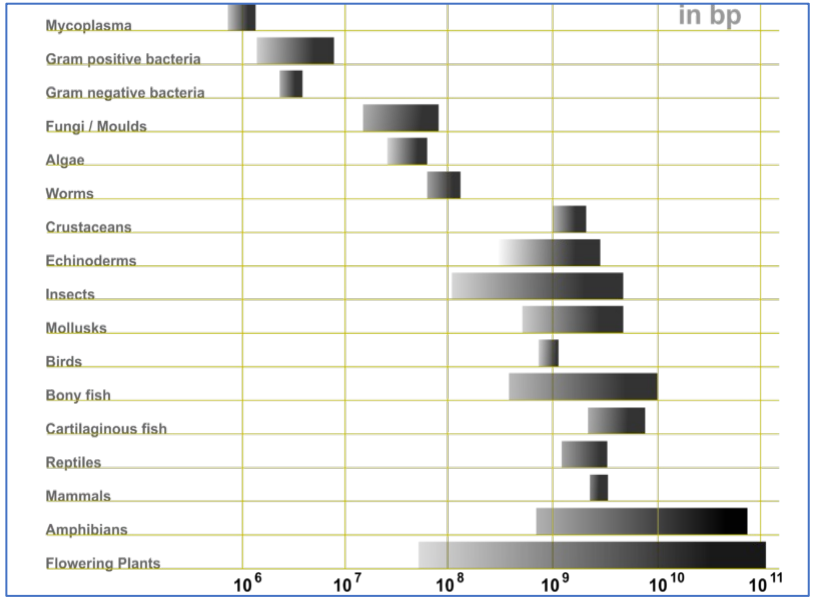
- For very different reads, use shorter size
- For more similar reads, user longer size
- Starting from published parameters

Genome assembly is getting easier

Do it yourself

Give it a try

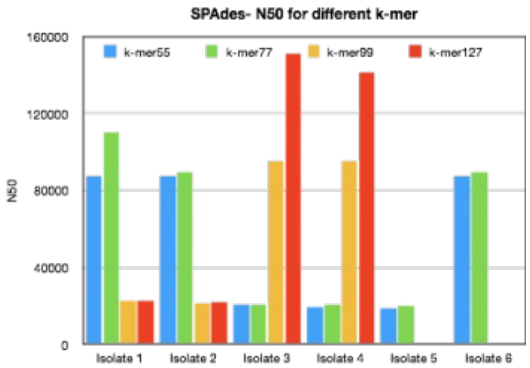
Find a collaborator

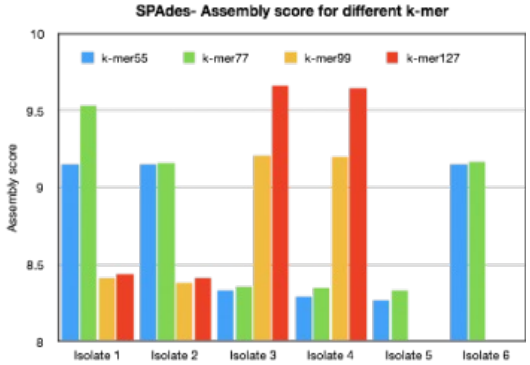


SPAdes – short read assembly for small genomes

Open source de novo assembler suitable for single-cell, isolate, RNA, plasmid genomes and metagenome

- https://cab.spbu.ru/software/spades/



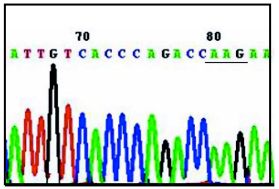


Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:40	NGS omics
10:45 – 10:15	MS omics
10:20 – 10:55	Enrichment analysis
11:00 – 11:15	Pathway & network analysis
Discussion & Next Lecture	

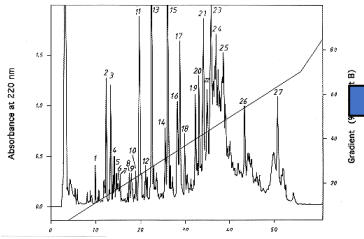
5 minutes break

Our goals



Mapping to reference genomes

Gene/protein IDs + Transcript Abundance



?

Compound IDs + Concentrations

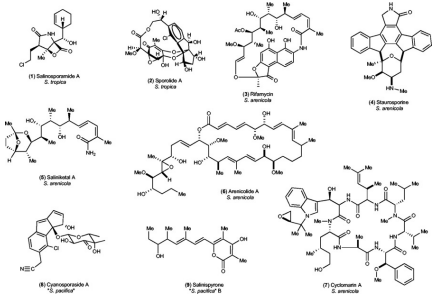
MS-omics: interfacing with the environment

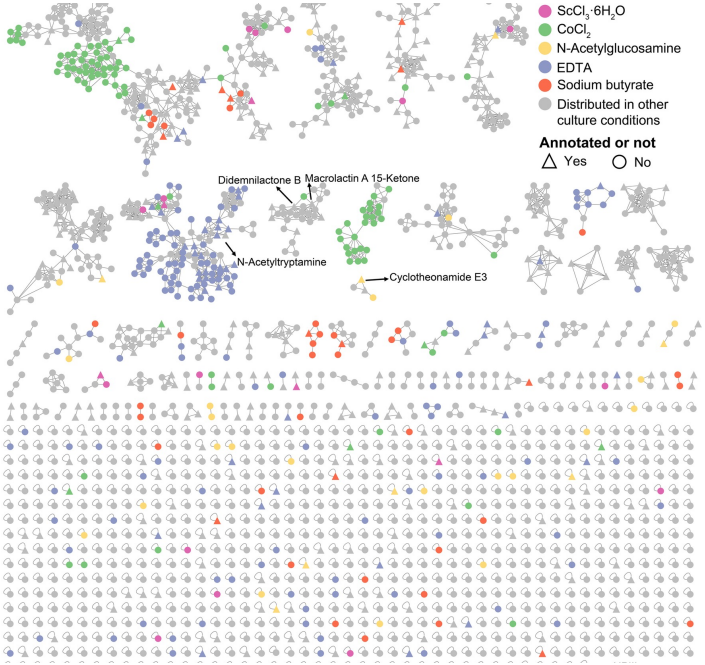
Endogenous and from environment

- food & drugs
- Microbiome
- pollutants

(Bio)chemical space is vast, highly dynamic

- Known knowns
- Known unknowns
- Unknown unknowns

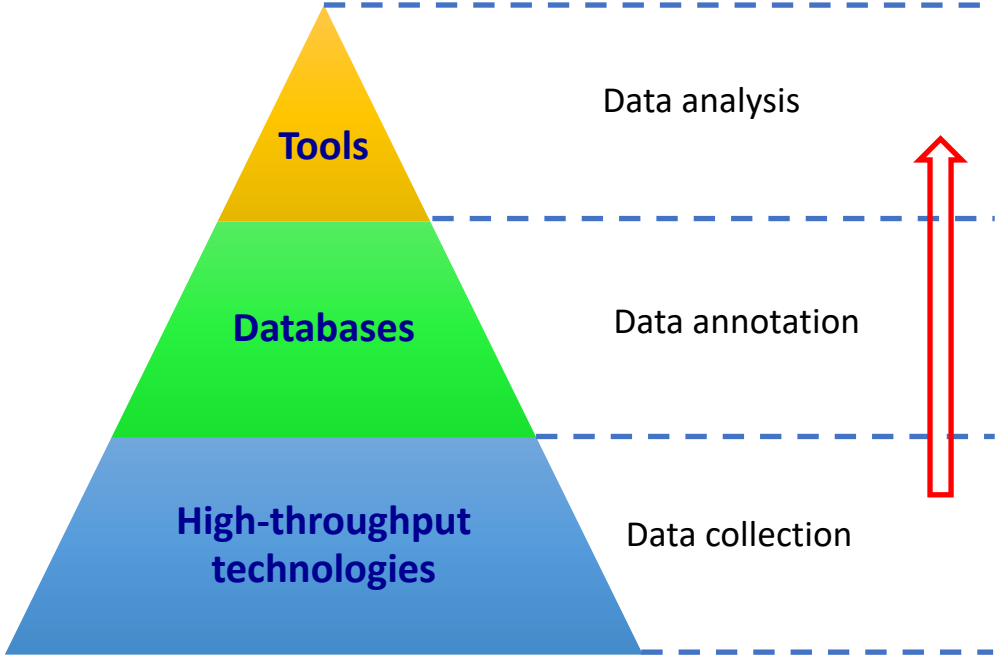




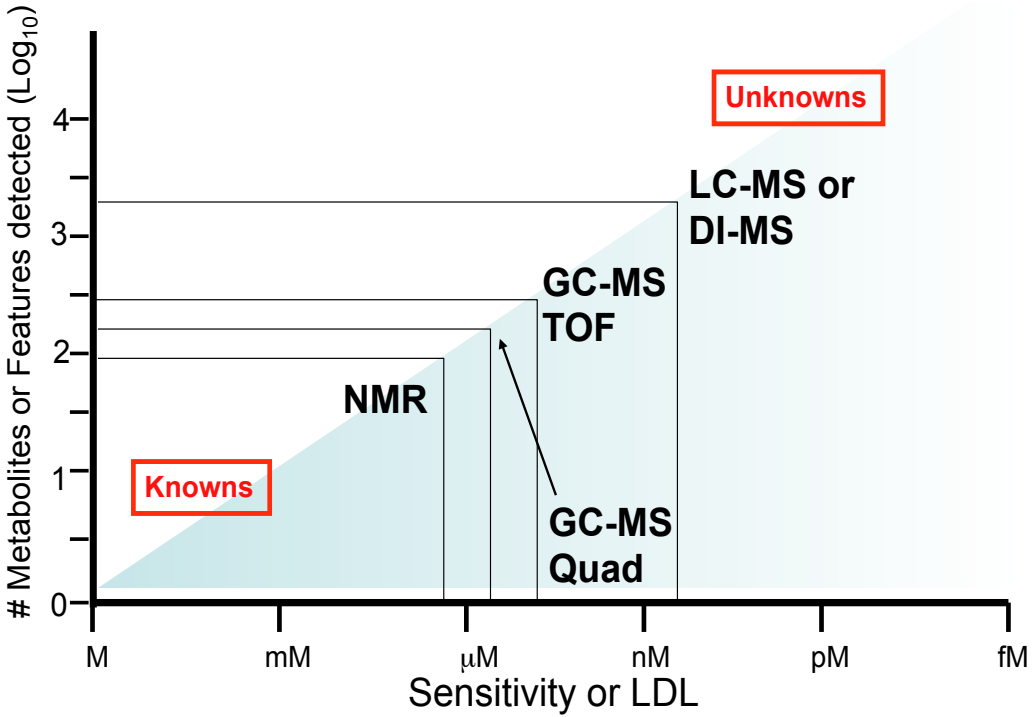
A map generated from GNPS

A relatively new omics field

- High-throughput
  - Measure a lot samples
- Cover a broad range
  - Measure a lot of features
- High-resolution
  - Accurate identification
- Cost effective
  - Instrument
  - Automation
- Bioinformatics infrastructure
  - Broadly accessible
  - Standardized?

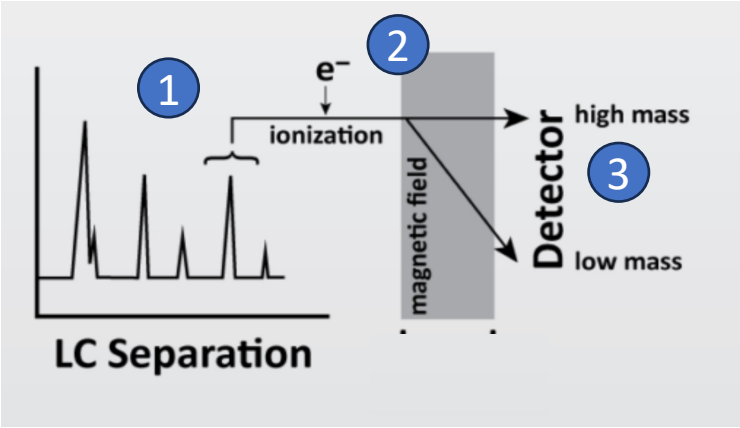


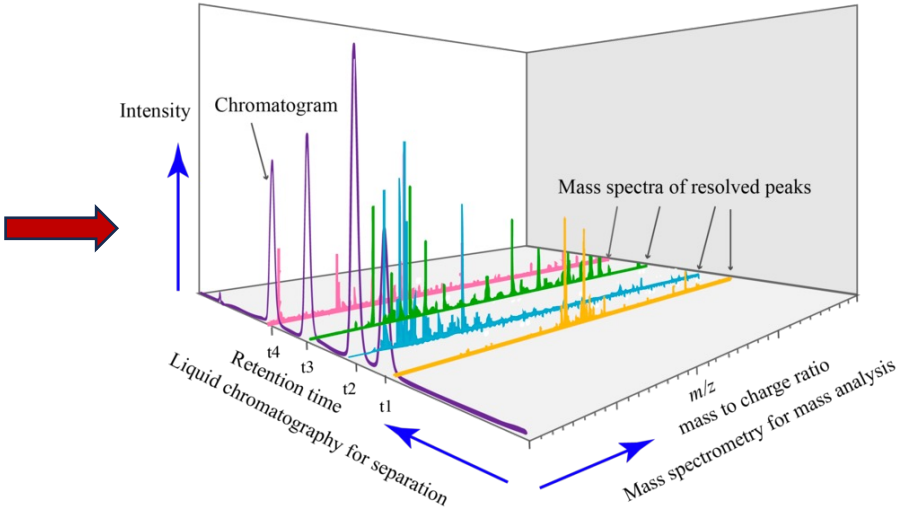
Analytical technologies



Wishart DS et al. (2006)

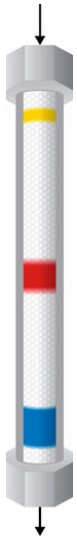
Overview of LC-MS

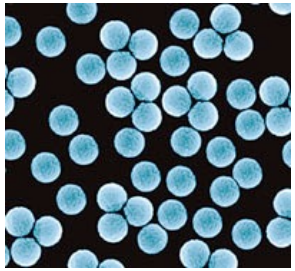


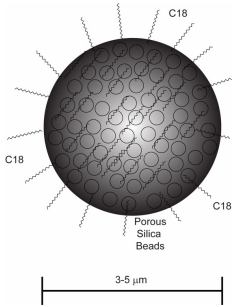


LC modalities

- **Reversed phase –** for separation of non-polar molecules (non-polar stationary phase, polar mobile phase)
- **Normal phase –** for separation of non-polar molecules (polar stationary phase, non-polar/organic mobile phase)
- **HILIC** – hydrophilic interaction liquid chromatography for separation of polar molecules (polar stationary phase, mixed polar/nonpolar mobile phase)







What we get from LC: Retention Time/Index (RT/RI)

Retention time (RT) is the time taken by an analyte to pass through a column

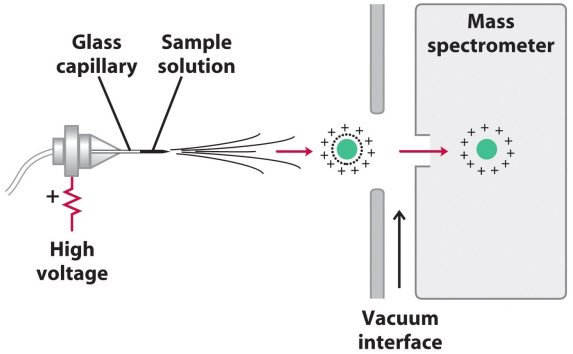
- RT is affected by compound, column (dimensions and stationary phase), flow rate, pressure, carrier, temp.
- Comparing RT from a standard sample to an unknown allows compound ID

Retention index (RI) is the retention time normalized to the retention times of adjacently eluting n-alkanes

- “standardized” - comparable cross different columns
- Mainly used in GC

Electrospray Ionization (ESI)

- Used in LC-MS
- Very sensitive, requires less than a pico-mole of material
- Strongly affected by salts & detergents
- Two ionization modes:
  - ✓ Positive ion mode
    - Add formic acid to solvent
  - ✓ Negative ion mode
    - Add ammonia to solvent



What we get from ESI? Adduct Ions

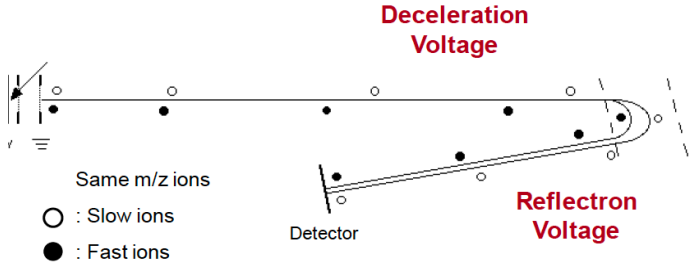
- Adducts are formed with ions derived from inorganic salts or residual water, inherently present in all biological samples.
- The most common adduct ions in LC-MS are
  - Protonated ion [M+H]+ (positive mode)
  - Deprotonated ion [M-H]- (negative mode)

Ionization	Formation	Ion Mass
Positive	$[M+H]^+$	m+1.0073
	$[M+2H]^{2+}$	m/2+1.0073
	$[M+Na]^+$	m+22.9892
	$[M+K]^+$	m+38.9632
	$[M+NH4]^+$	m+18.03382
Negative	$[M-H]^-$	m-1.0073
	$[M-2H]^{2-}$	m/2-1.0073
	$[M-2H+Na]^-$	m+20.9747
	$[M-2H+K]^-$	m+36.9486

M is the molecule with molecular weight m

MS Detectors

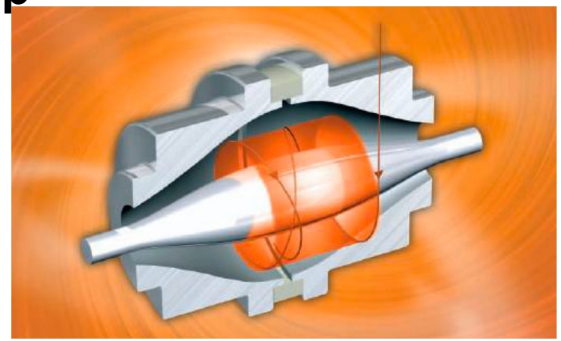
Time of Flight (TOF)



Time of Flight

- Ions fly along a long tube
- Diagram shows a reflector
- Smaller ions get to the detector 1st
- Analogous to a gel

Orbitrap



Orbitrap (FT)

- Ions are trapped between a central & an outer electrode
- Ions oscillate to produce a current
- Current is transformed into a m/z
- High sensitivity, good range
- Very High Resolution & Accuracy

Mass Accuracy

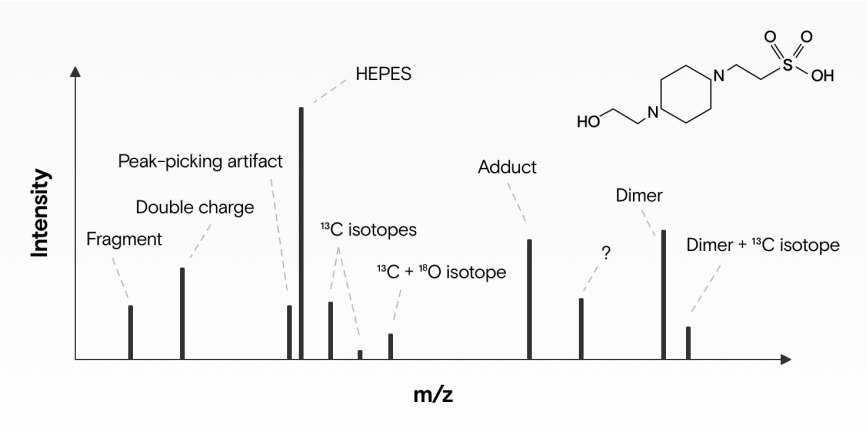
A metric describing the difference between the measured m/z (m_{exp}) of an ion and the real, exact m/z (m_{calc}) of that ion

$$ppm = \left(\frac{m_{exp} - m_{calc}}{m_{exp}}\right) * 1E+6$$

Type	Mass Accuracy
FT-ICR-MS	0.1 - 1 ppm
Orbitrap	0.5 - 1 ppm
Magnetic Sector	1 - 2 ppm
TOF-MS	3 - 5 ppm
Q-TOF	3 - 5 ppm
Triple Quad	3 - 5 ppm

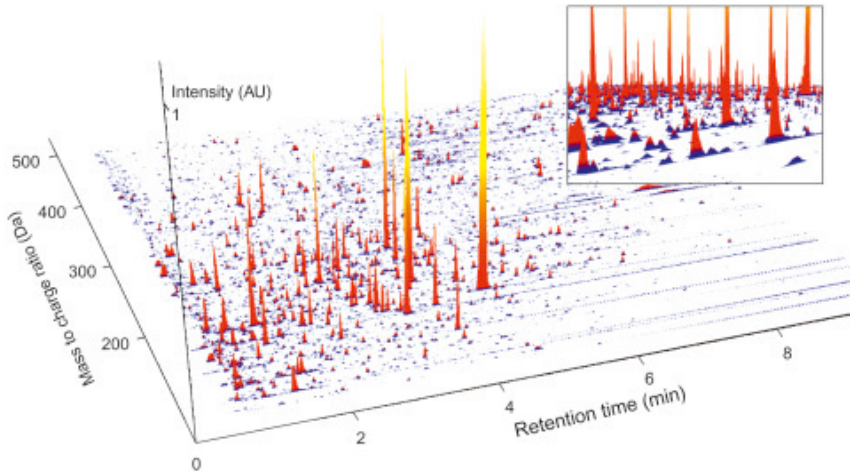
What we get from MS? mass spectrum

- Characterized by sharp, narrow peaks
- X-axis position indicates the m/z ratio of a given ion (for singly charged ions this corresponds to the mass of the ion)
- Height of peak indicates the relative abundance of a given ion
  - Not reliable for absolute quantitation
  - Peak intensity indicates the ion's ability to desorb or "fly" (some fly better than others)



LC-MS feature annotation with binner

Mass chromatograms of biological mixtures

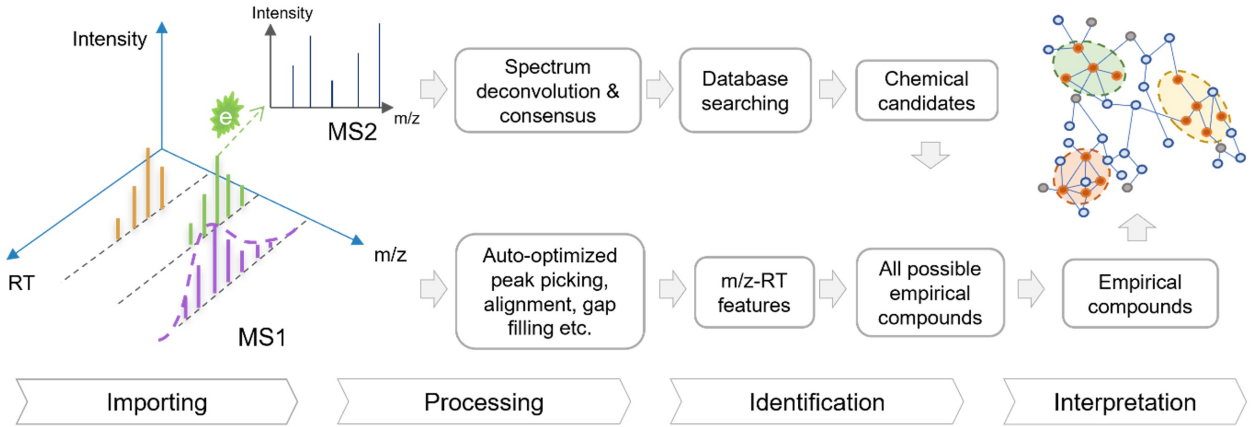


- Containing >10,000 peaks (for complex samples such as blood)
- Each peaks are characterized by RT and m/z

Processing MS Spectra

Lecture #8
Global Metabolomics

1. Raw data is large and noisy
2. Identify, quantify, and align all possible features (peaks) across samples
3. Output: A table of features (RT, m/z) with their quantitative information for subsequent statistical analysis



What is the table look like?

Feature ID:

m/z__retention time

They are usually not unique for each compounds

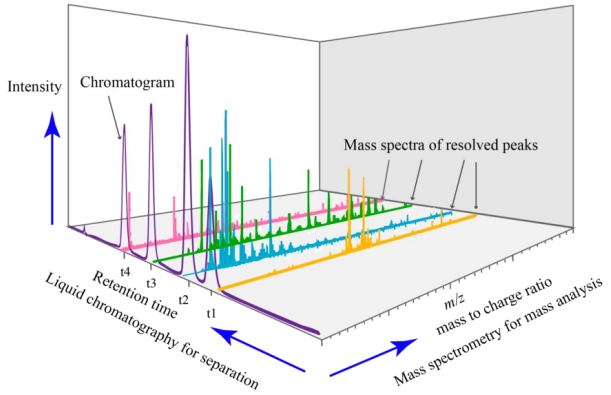
- One compounds can give multiple peaks (adducts, fragments, isotopes)
- One peak can be contributed from different compounds

Sample	Naive_007	Naive_027	Naive_071	Naive_109	Naive_127	Naive_139	Semi_025	Semi_045
85.065__24.64	166020.82174	120086.08327	19509.14147	118112.9565	87039.61996	104898.32903	152651.82014	65984.36172
85.0843__167.65	1645008.20686	1190622.95729	638456.21528	0	1629532.41894	425340.44229	2314313.37182	2527459.04192
86.0602__37.16	18142.47554	16616.03235	10886.52278	0	0	40055.58588	1212986.02473	661407.86422
86.0602__129.2	757420.68035	1401437.3743	986969.31026	2559178.25865	1720322.70316	927659.48141	3910272.07896	3044111.48558
86.0603__31.13	40168.41032	28760.74122	40007.49015	2115.15429	30082.36671	83055.68507	1880331.86961	925490.77154
86.0966__58.82	95507679.62103	88613771.10909	71913129.46904	0	2871361.94795	120978190.73858	124289527.10497	193556499.1973
86.0966__93.94	24447611.88604	33839015.07872	18306222.10061	0	0	38881672.64366	264723805.50055	119992641.57424
86.0966__82.57	16330118.15869	29280720.01315	30329806.59365	0	2432433.31135	22107141.53254	118751749.01004	93636719.91825
86.0966__68.05	175048972.67449	211835241.96262	195847604.48591	324856733.47678	241282511.16983	288748953.47116	122900362.28296	209581146.42487
86.0966__112.47	232174121.40282	296434730.20178	249900170.5577	318182655.37461	320382689.28819	285853906.32259	62344728.00885	87990374.29879
86.3224__113.03	7869.68786	86403.08999	140194.5321	122428.33792	119821.68133	16382.27583	0	0
86.4142__99.45	0					0	194280.97559	104339.076
86.9088__78.17	216418.99578	366778.9				587	2565704.48103	1946411.1865
86.9089__119.63	358477.04559	471082				562	437767.57505	338888.61801
87.0555__125.67	195542.43909	438179.				259	1707623.47516	1785615.80662
87.0807__24.28	10726572.6923	15861006.58178	4908177.35316	15718191.42655	8369954.2801	21065169.96481	19605613.83141	17342193.77814
87.1__94.12	1782287.21956	2340270.52594	2199636.80059	0	1838237.43785	2658906.7147	17021522.88232	7194992.10607
87.1__82.33	606445.59884	779267.39812	1070621.04313	0	945785.90837	1069344.12945	7694163.1766	5724586.55159
87.1001__59.14	6353575.18596	5317603.85364	4658363.32391	6300115.1542	5076827.74772	7669117.8019	8172228.67049	12122819.67458
87.1001__112.34	15818007.26817	17441204.29839	17078378.05002	19122679.91077	19371530.70183	17250463.5533	4720043.01702	158567.7644
87.1001__68.33	10536674.7611	13724928.85807	12572554.71905	13167562.0864	12927906.28045	17704361.0455	8080735.09674	12087543.16651
88.0396__118.11	582908.82363	473551.43372	526186.84049	0	993740.18125	695491.17967	4108599.87774	3961933.86831
88.0634__161.45	5990652.35725	5565827.30796	23231827.96063	3879300.38088	6591815.81478	2658625.76775	7665461.22479	11917821.56697
88.084__24.45	600429.24374	804082.55302	272354.79181	2549387.97523	148499.20848	2126344.35834	1035719.95574	1000048.4845
88.5093__90.56	10665122.95898	10036657.78865	17542485.32793	0	7128972.6611	10329434.35436	9708306.02631	10942692.94755
88.5094__24.81	671314.2294	845725.38143	426586.15385	882046.35578	627668.46376	964648.55011	738398.32934	438146.60468

Peak intensity values

Summary of MS table generation

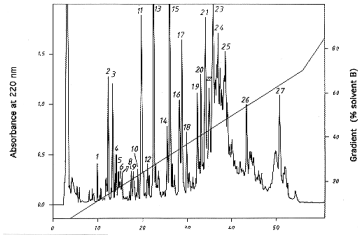
1. Identification
   - Coordinates (m/z and RT) define the feature ID
2. Quantification
   - Relative quantification (peak height or peak area)




```
<?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="PT0.095565" 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>
```

200.1/2926	147887.53	451600.71	65290.38	56540.93	175177.08	82619.48	51951.61
205/2791	1778569	1567038	1482796	1039130	1950287	1466781	1572679
206/2791	237993.6	269714	201393.4	150107.3	276541.8	222366.2	211717.7
207.1/2719	380873	460629.7	351750.1	219288	417169.6	324892.5	277990.7
219.1/2524	235544.92	173623.38	82364.59	79480.4	244584.47	161184.05	72029.38
231/2516	117649.77	48960.63	222609.07	286232.15	465898.01	61234.44	96841.46
233/3023	399145.3	356951.3	410550.7	198416.5	397107.8	271252.1	334459.9
234/3024	76880.87	99526.27	97493.76	53461.71	65215.64	55952.44	73781.01
235.1/2695	171995.22	128945.16	155442.48	115286.25	199981.49	30028.6	156968.3
236.1/2524	252282.04	206031.93	71763.79	73602.47	253791.07	187225.65	79389.63

Have we reached our goal?



?

Feature IDs +
Relative abundance

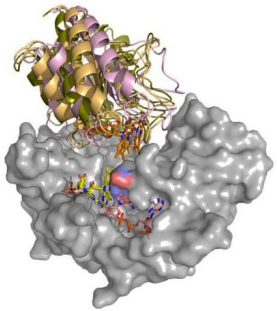
Compound IDs +
Concentrations

Accurate compound ID is unattainable with current LC-MS technologies that allow cost effective, high-throughput analysis

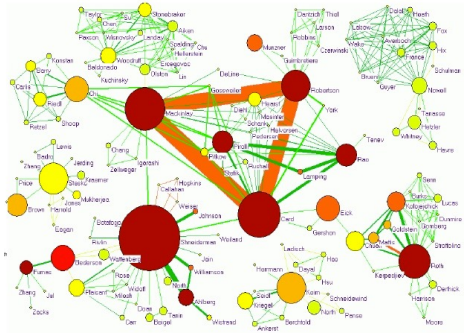
In Omics Data Analysis, accuracy at individual feature level is usually not the goal. Instead, we should focus on robust patterns, trends and functions. All we need is "**Approximately Correct**"

Goals of omics data analytics

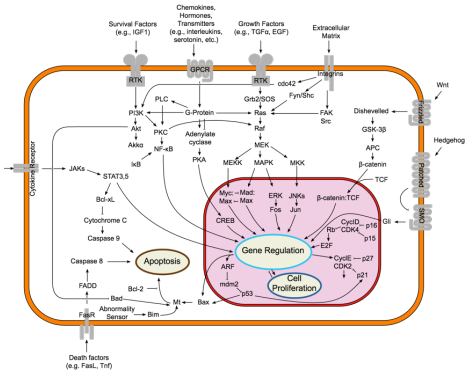
Individual Molecules



Omics & Multi-omics



Systems Biology



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

Levels of Metabolite Identification

1. Positively identified compounds
   - ➢ Confirmed by match to known standard

Accurate

2. Putatively identified compounds
   - ➢ Match to MS + RT or MS/MS + RT

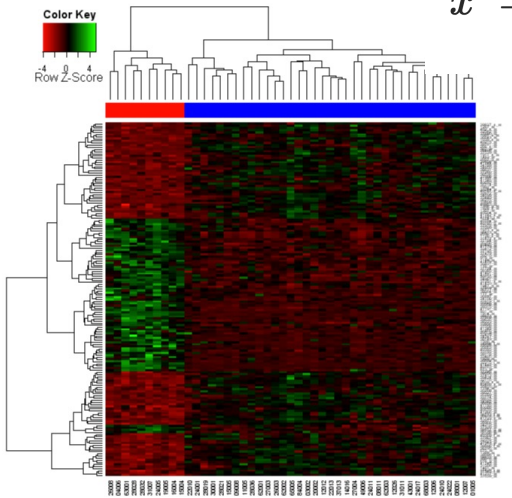
Approximately Correct

3. Compounds putatively identified in a compound class
4. Unknown compounds

Relative abundance are all we need (most time)

- Absolute concentrations are useful as reference values across different platforms
  - ✓ Regulatory requirements
- Relative quantification are sufficient:
  - ✓ Compare same features under different conditions
- Not a concern in NGS-omics – all relative qualifications
  - ✓ Gene counts are related to sequence depth

$$x' = \frac{x - x_{mean}}{\sigma}$$



Even we start with absolute concentration, normalization will convert into relative concentrations before statistical analysis.

If you look for absolute ID & quantification

- Most MS-based metabolomics studies are not absolutely quantitative
- Absolute quantitation requires spiked addition of ^{2}H, ^{13}C or ^{15}N isotopic standards
- Also requires that the labeled compound is of the same chemical type or very nearly the same chemical type as target compound
- Single Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM) is used to ensure correct compound ID

Schedule for today

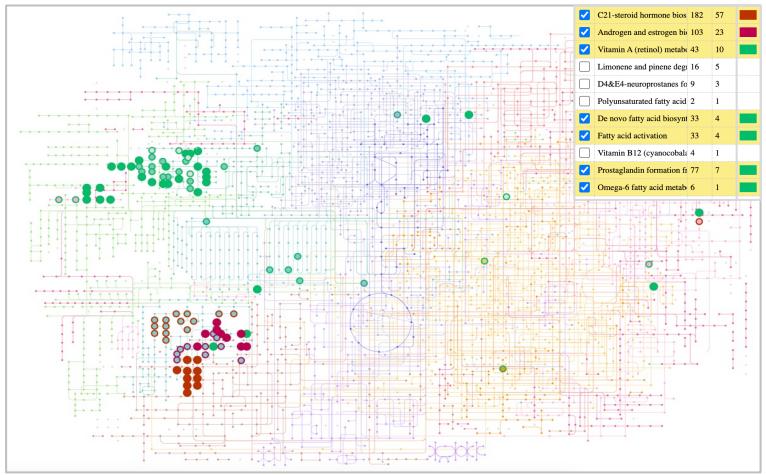
Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:40	NGS omics
10:45 – 10:15	MS omics
10:20 – 10:55	Enrichment analysis
11:00 – 11:15	Pathway & network analysis
Discussion & Next Lecture	

5 minutes break

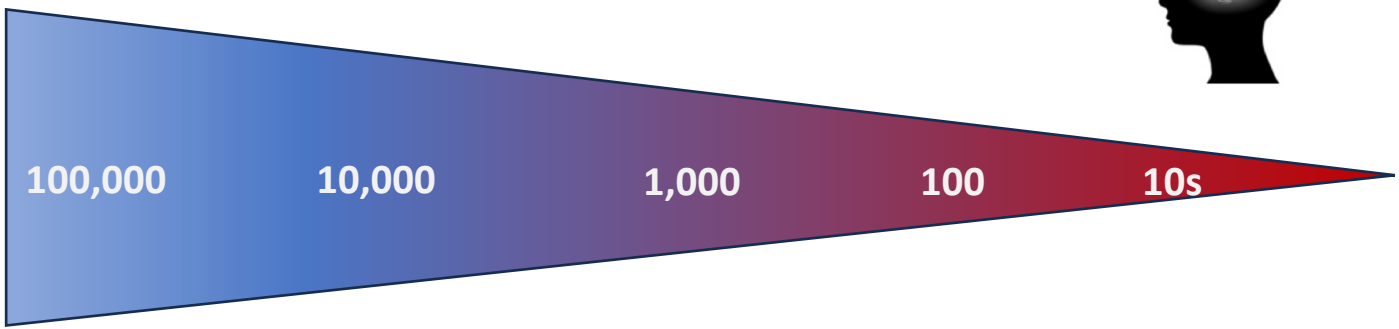
Task #2: from data table to functions

Plekhg1	485	462	467	562	347	296	458	487
Plekhg6	126	172	206	198	137	164	110	152
4930592I03R	7	5	9	7	2	7	0	7
Plekhg4	0	0	0	0	0	0	0	0
Plekhg5	117	150	138	95	83	109	105	91
Nsa2	1540	1472	1577	1792	1419	1580	1288	1670
Nampt	2882	3103	3487	3090	2400	3148	2343	2569
Vmo1	43	68	51	62	24	30	22	26
Man2a1	11979	13268	15988	14592	7416	9708	7285	10071

200.1/2926	147887.53	451600.71	65290.38	56540.93	175177.08	82619.48	51951.61
205/2791	1778569	1567038	1482796	1039130	1950287	1466781	1572679
206/2791	237993.6	269714	201393.4	150107.3	276541.8	222366.2	211717.7
207.1/2719	380873	460629.7	351750.1	219288	417169.6	324892.5	277990.7
219.1/2524	235544.92	173623.38	82364.59	79480.4	244584.47	161184.05	72029.38
231/2516	117649.77	48960.63	222609.07	286232.15	465898.01	61234.44	96841.46
233/3023	399145.3	356951.3	410550.7	198416.5	397107.8	271252.1	334459.9
234/3024	76880.87	99526.27	97493.76	53461.71	65215.64	55952.44	73781.01
235.1/2695	171995.22	128945.16	155442.48	115286.25	199981.49	30028.6	156968.3
236.1/2524	252282.04	206031.93	71763.79	73602.47	253791.07	187225.65	79389.63



How do we (human) understand data?



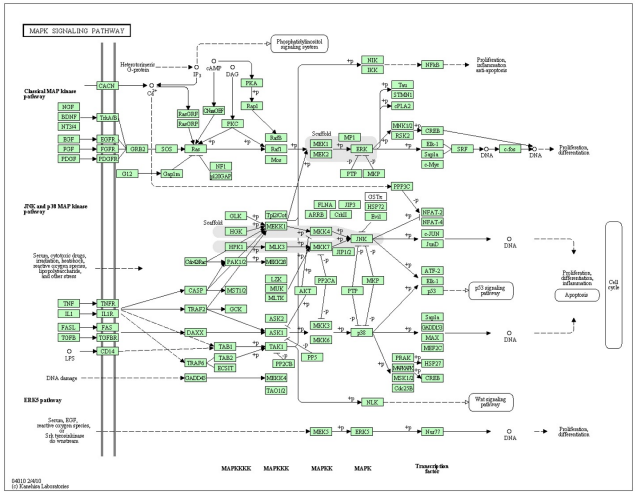
- NGS omics: Reads => genes => functions => enrichment tests
- MS omics: Peaks => compounds => functions => enrichment tests

"Biological" dimensionality reduction!

What are functions?

- Group behavior, collective changes

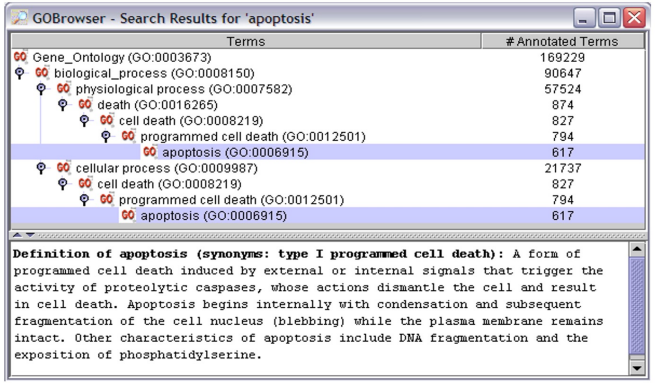
KEGG is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemicals.



Gene Ontology

GO has three independent partitions, which are not interconnected:

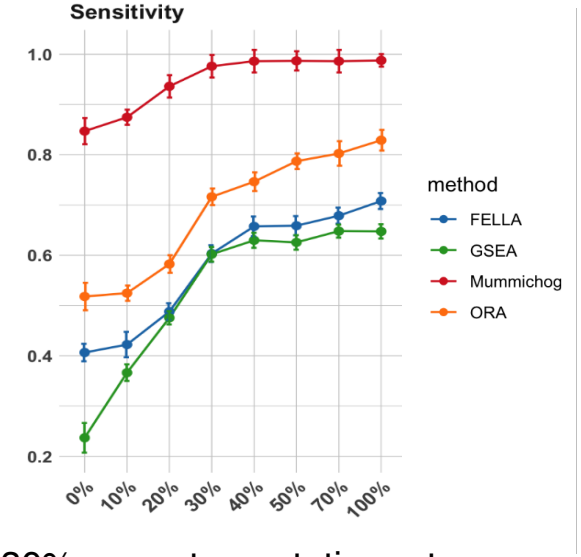
- **Molecular Function**
  - Describes biochemical activities, in-vitro binding specificities, etc…
  - Example: *Ligase Activity*, *Kinase Activity*, *DNA Binding*
- **Cellular Component**
  - Describes parts of the cell
  - Example: *Mitochondrion*, *Spindle Microtubule*
- **Biological Process**
  - Describes processes at the intra-cellular and organism level
  - Example: *DNA Replication*, *Apoptosis*, *Development*



Key observation

- Biological systems showing coordinated changes or group behaviors
- Leveraging this “collective power” inherent in biological systems can better tolerate the random errors/inaccuracies based on individual features

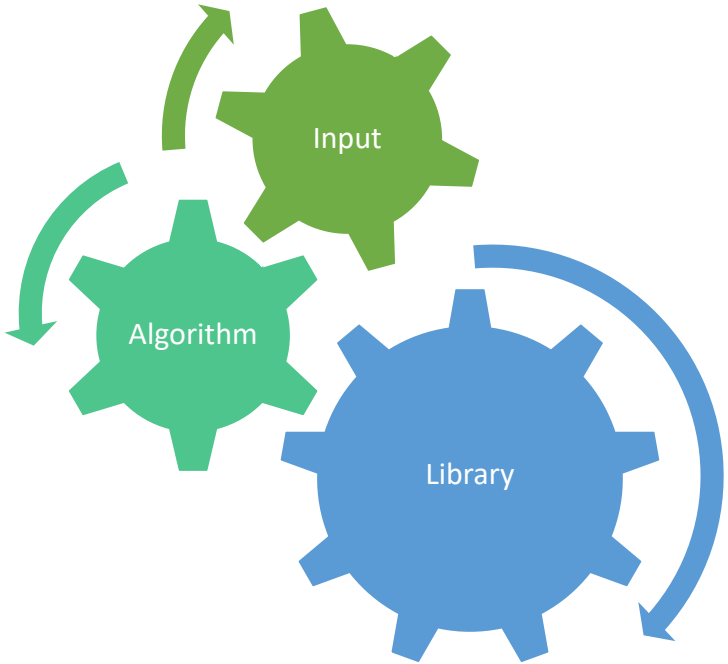
Approximately correct at feature levels allows detection of **accurate functional level** changes



30% correct annotation rate (remaining random assignment) enable ~100% pathway detection

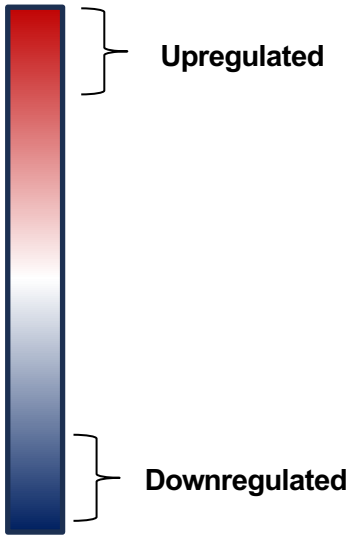
https://doi.org/10.1093/bib/bbac553

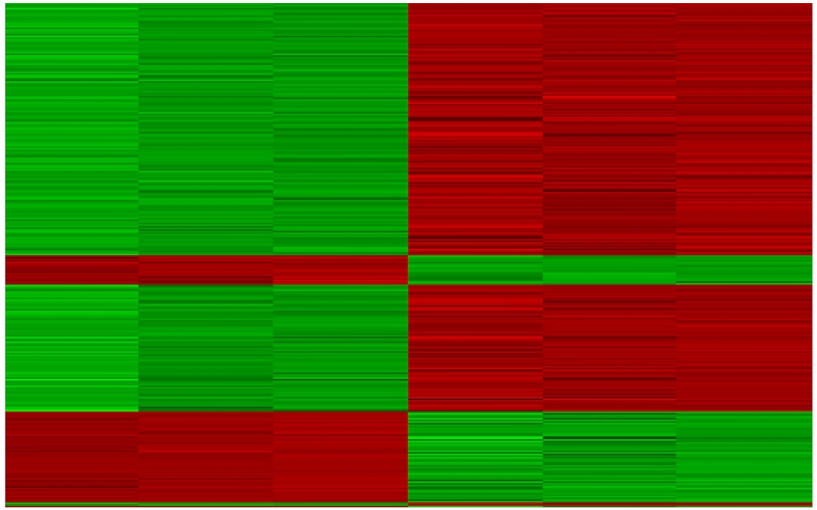
Three components for functional analysis



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

Input for enrichment analysis



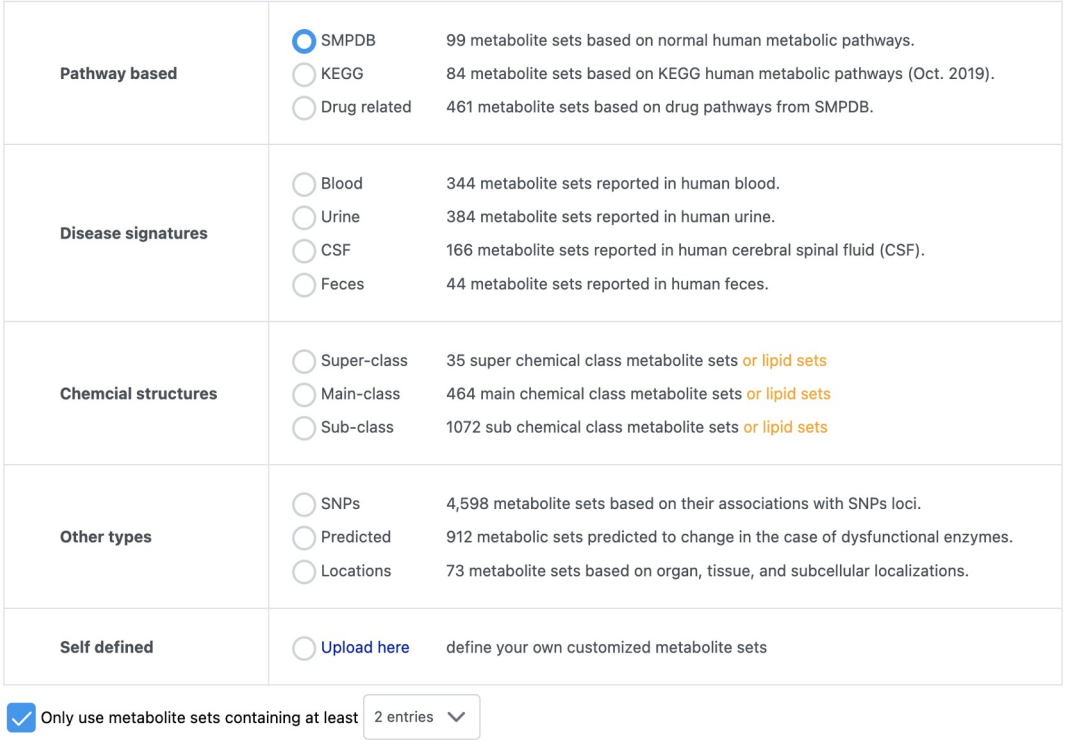


Clustered (co-regulated) compounds

Functional libraries

Biologically meaningful groups such as:

- Pathways
- Disease associated metabolite signatures
- Localization
- SNP associated metabolites
- …..



Enrichment Tests

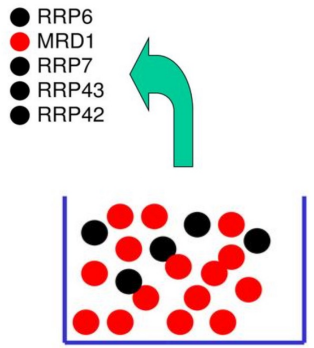
We need to calculate the null distribution - how often we can see this by random chance? If the chance is low, then we think it is enriched

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

Hypergeometric distribution

A discrete probability distribution that describe the probability of success in draws, without replacement

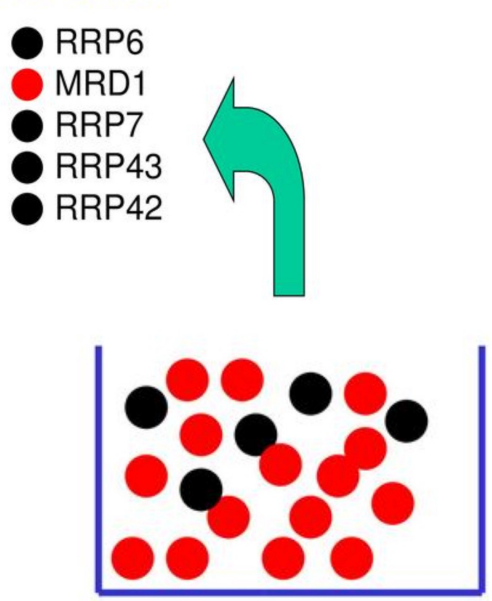
$$f_X(k|N,K,n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$



	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)

1. An input gene list is created using a certain threshold or criteria
    - FDR < 0.05 and/or
    - FC > 2
2. For each pathway, input genes that are part of the pathway are counted
3. Test for probability using hypergeometrical tests
4. Repeat for every pathway
5. Multiple testing adjustment



Issues with ORA approach

1. The input list is subject to some arbitrary cutoff
   - Different cutoff values can lead to different results
2. The “background universe” (all measurable features by the platform) may not be well defined
   - The transcriptome / proteome / metabolome?
   - All those defined in the functional library?
3. Selection bias: all balls in the urn should have equal chance of being selected) issue
   - Every gene has equal chance being sequenced by NGS?
   - Every compound has equal chance being measured by MS?

The issues are more outstanding with MS-omics

Enrichment tests in global metabolomics

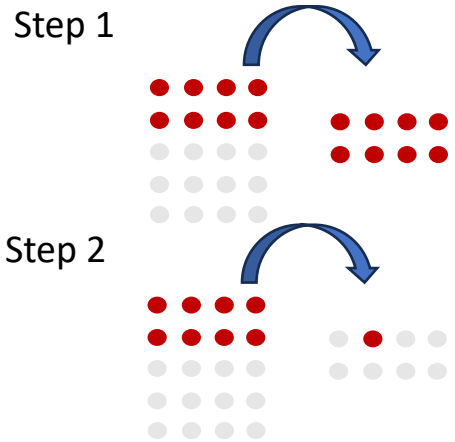
Input requirement:

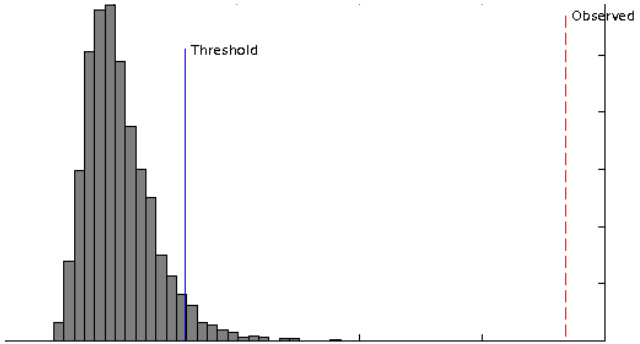
- Significant list: selected by t-test or fold change
- Reference list: all features detected

Step 1: Match the peaks to putative metabolites. Looked up all the significant metabolites in each pathway and calculate the p-value using hypergeometric tests

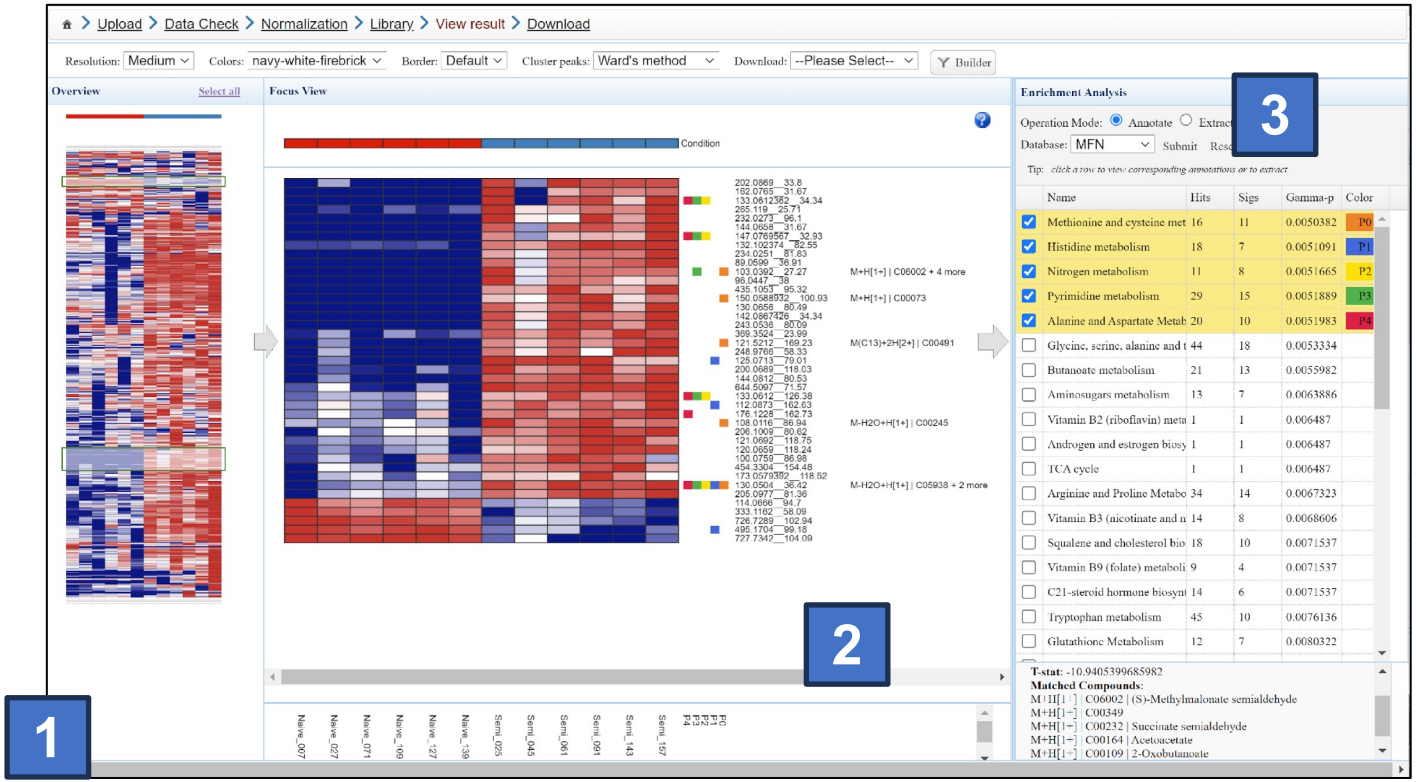
Step 2: Randomly draw features with the length same as the significant ones from the reference list, and repeat Step 1 for thousand times

Step3: Test if certain pathways are enriched in the significant peaks as compared to null models





From patterns to functions to matching details

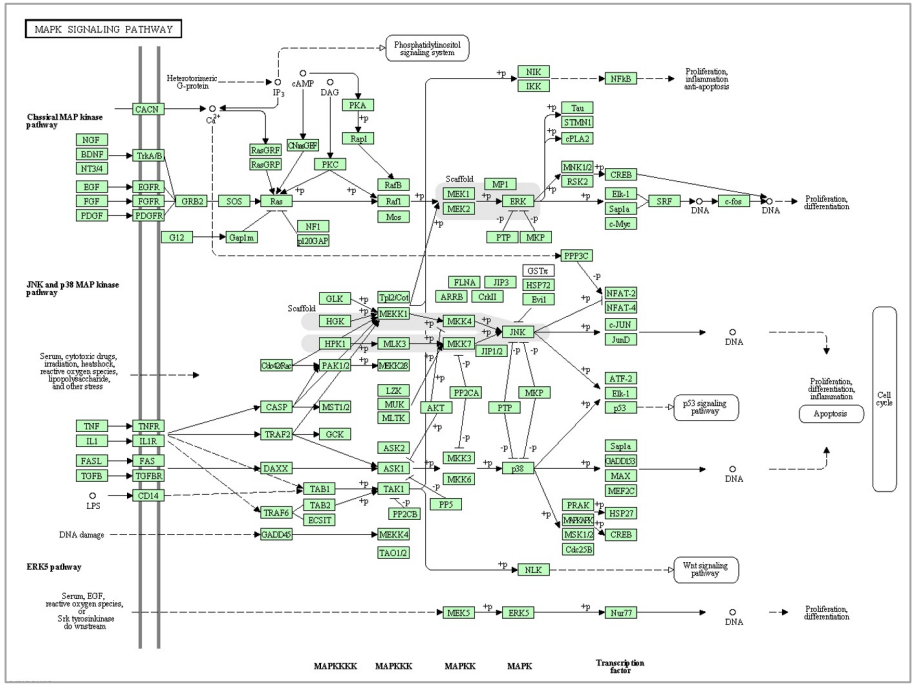


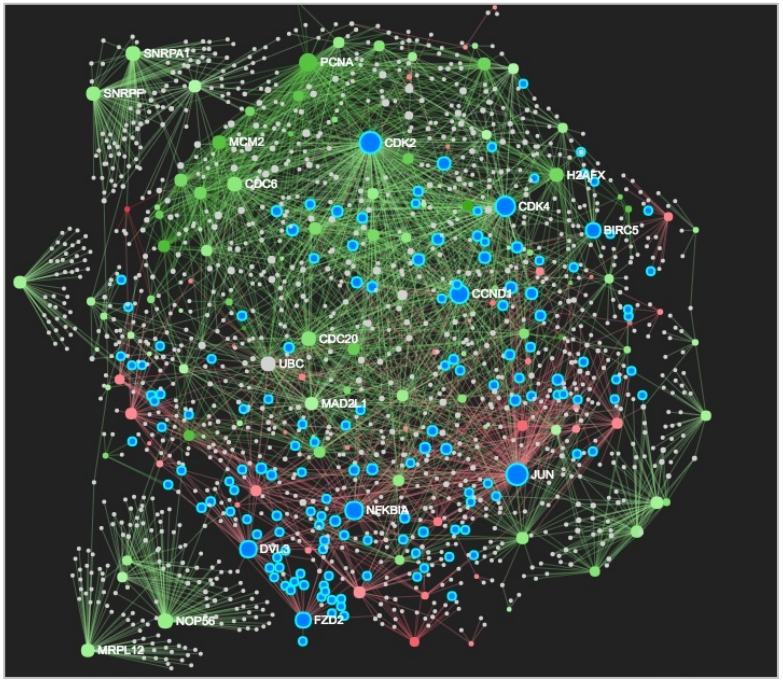
Schedule for today

Time	Topics
9:00 – 9:10	Introduction
9:10 – 10:40	NGS omics
10:45 – 10:15	MS omics
10:20 – 10:55	Enrichment analysis
11:00 – 11:15	Pathway & network analysis
Discussion & Next Lecture	

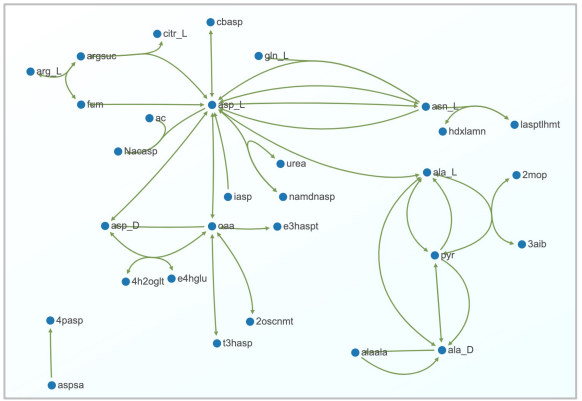
5 minutes break

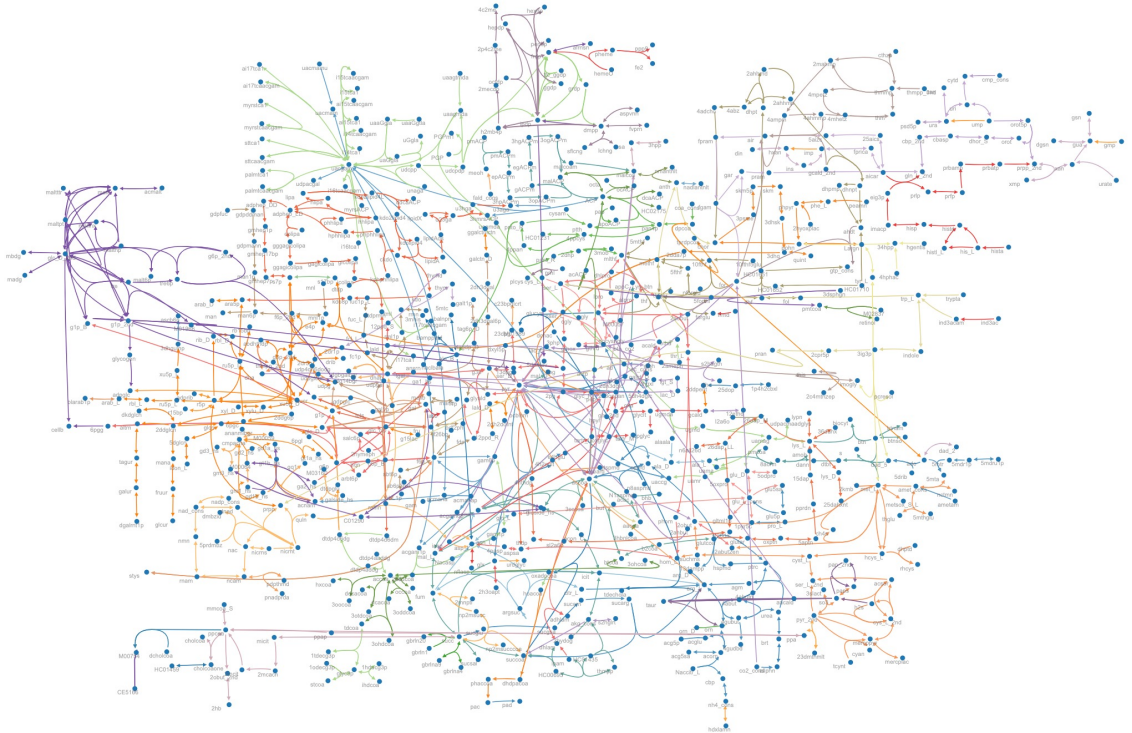
Pathways & Networks (current)





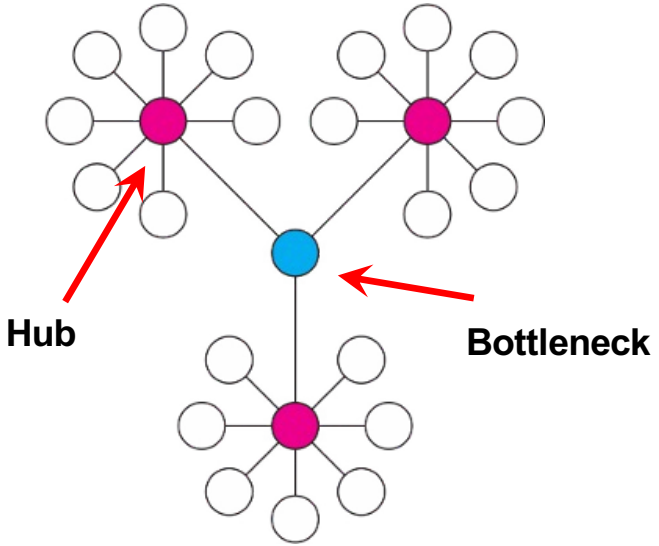
Based on genome scale metabolic models (GEMs)





Pathway analysis – what's new?

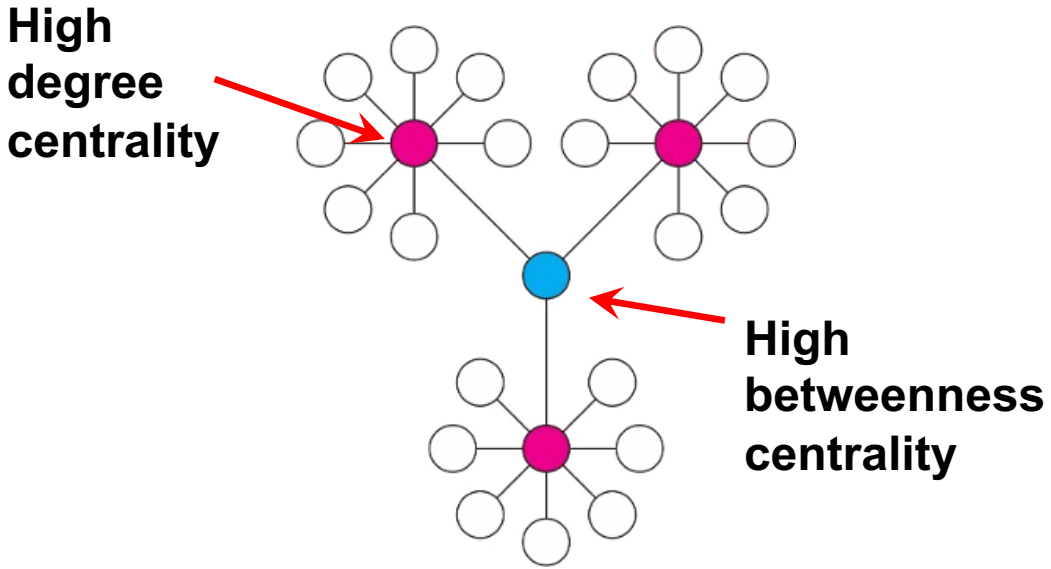
- Which positions are important?
  - Hubs
    - Nodes that are highly connected (red ones)
  - Bottlenecks
    - Nodes on many shortest paths between other nodes (blue ones)
- Graph theory
  - Degree centrality
  - Betweenness centrality



Junker *et al. BMC Bioinformatics* 2006

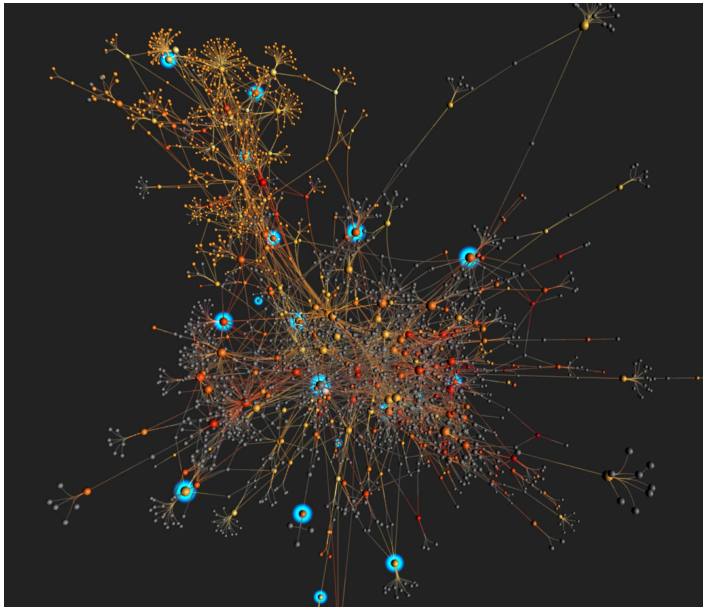
Which node is more important?

- Hubs
- Bottlenecks
- Modules
- Shortest paths
- .....

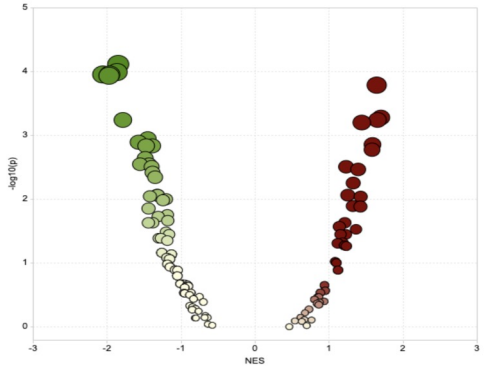


Leveraging networks to advance omics research

- Knowledge representation
  - Gene regulatory networks
  - Protein-protein interaction networks
  - Metabolic networks
- Case studies
  - Accelerate discovery to illustrate unknown
    - Global metabolomics
  - Contextualized analysis
    - Microbiome
  - Multi-omics integration



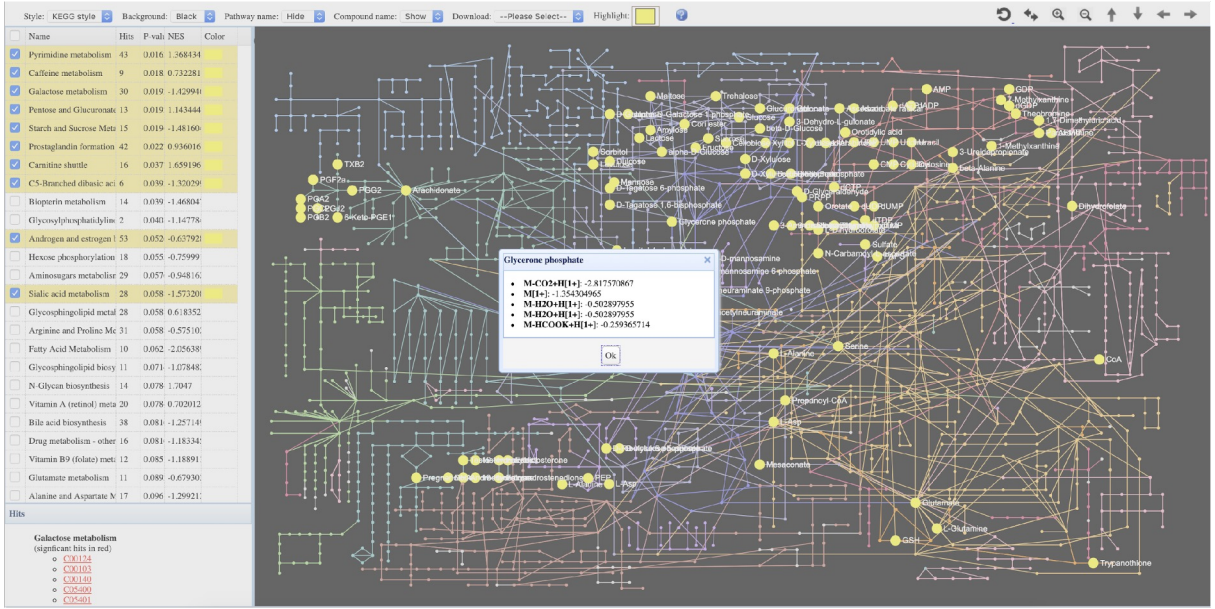
Functional insights & matching details

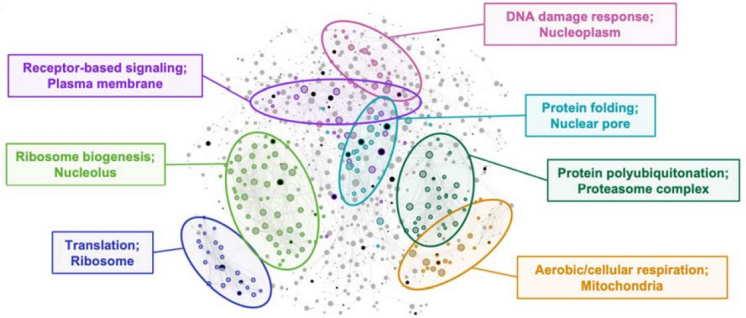


The red compounds indicate all potential matches from the user's input to the selected pathway.

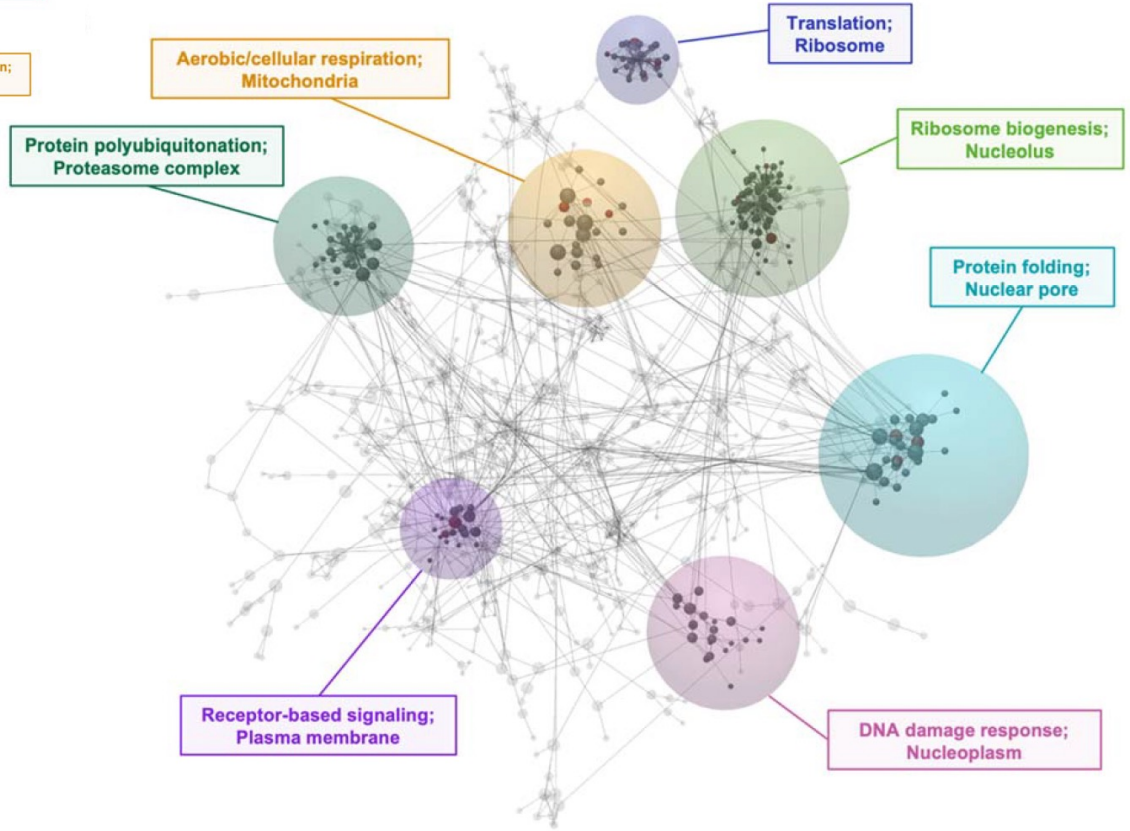
Pathway	Metabolites
Prostaglandin formation from arachidonate	**CE5703**; CE5700; CE6238; **C00425**; CE5707; **CE5704**; **CE5705**; CE5724; CE5708; CE1243; CE6236; C00584; **CE5535**; **C00959**; **CE5537**; **CE5925**; **CE5533**; C13856; **CE0955**; **CE4980**; C01312; CE5304; **C04686**; CE4938; **C05963**; **C04741**; **CE5931**; **CE0737**; **C01041**; CE4878; **CE5926**; **C00051**; **CE5924**; **C00219**; C00704; **CE5928**; **CE5929**; **C00020**; CE4876; **C00010**; **C05954**; C00024; **C05962**; C00030; **C00072**; CE7107; CE7105; C00028; C00696; **CE1447**; **C04685**; **C11695**; CE6244; CE6245; CE6242; CE6243; **CE6240**; CE6241; **C05953**; C00027; **C05956**; **C05957**; C00427; **C05955**; CE4877; CE6235; **C05959**; **C00639**; CE5730; **CE5534**; **CE5930**; C00282; **CE3481**; C00189; CE7054; C11304; C02198; **CE5828**

OK



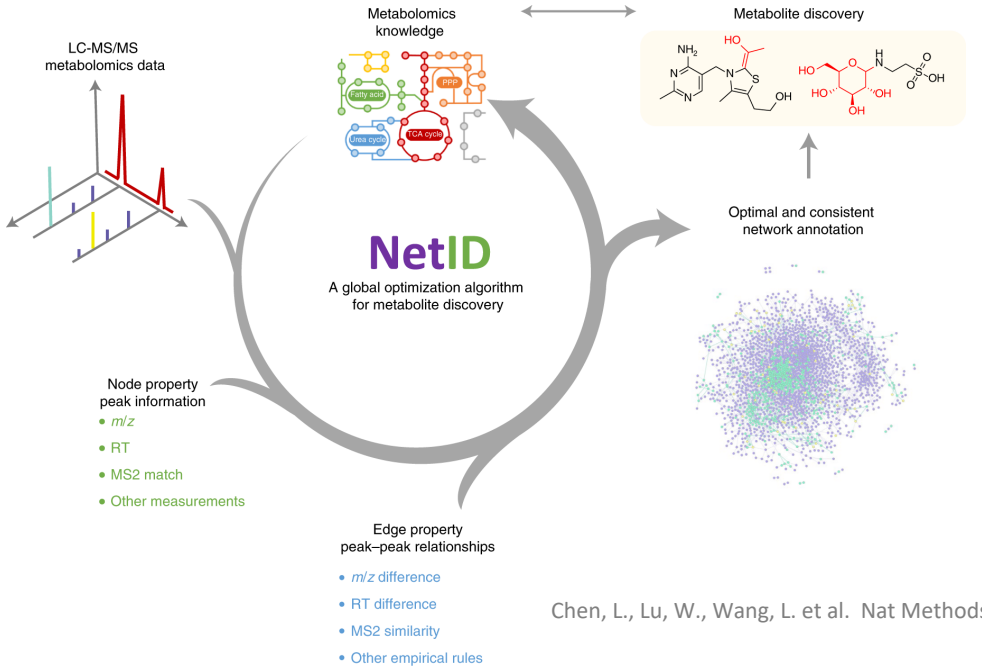


Module detection



Global optimization on networks for annotation

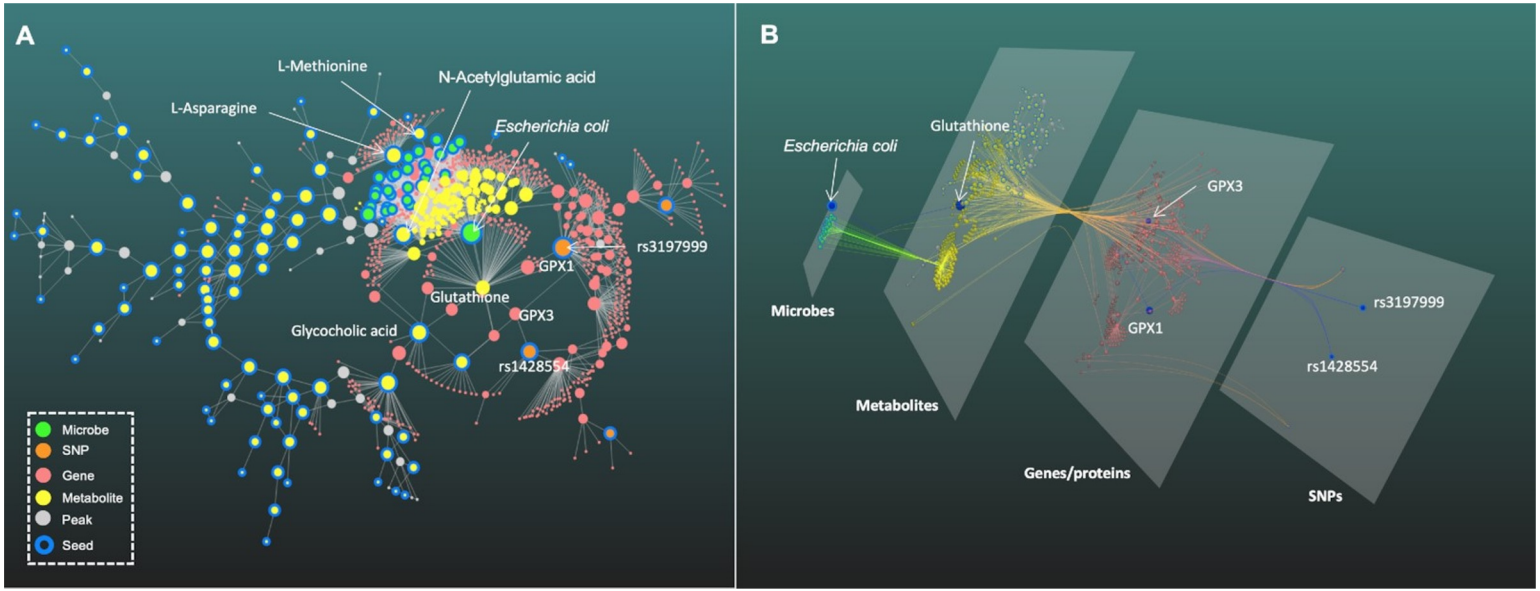
Integrating contextual knowledge (biochemical rules) – multiple signals converging to the real assignment



Chen, L., Lu, W., Wang, L. et al. Nat Methods (2021)

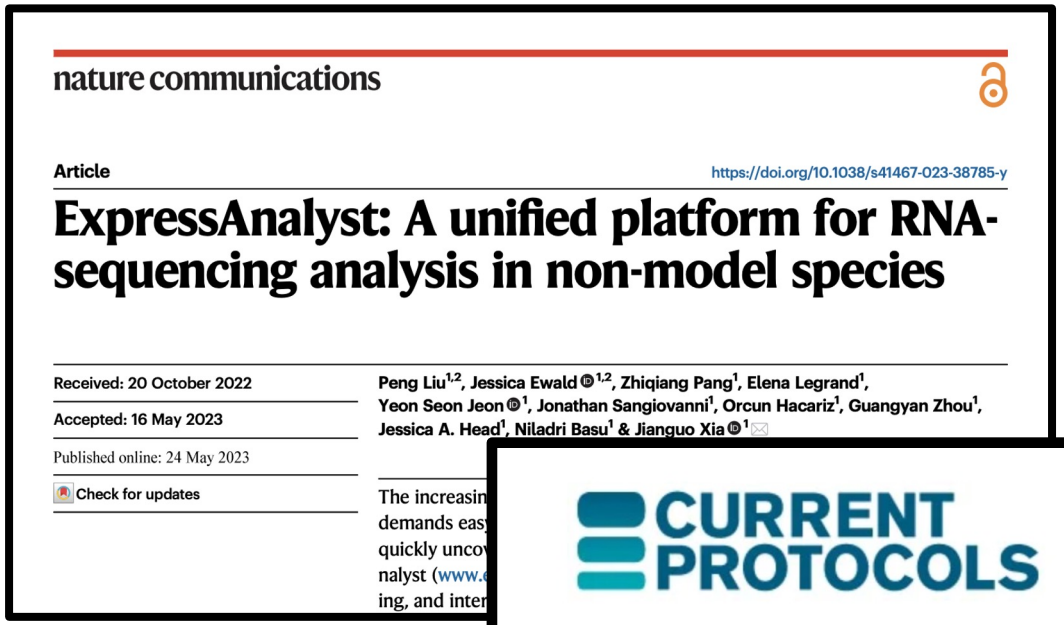
Multi-omics based on networks

❖ Integrating SNPs, Taxa, LC-MS peaks to our knowledge framework



Next Lecture

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



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

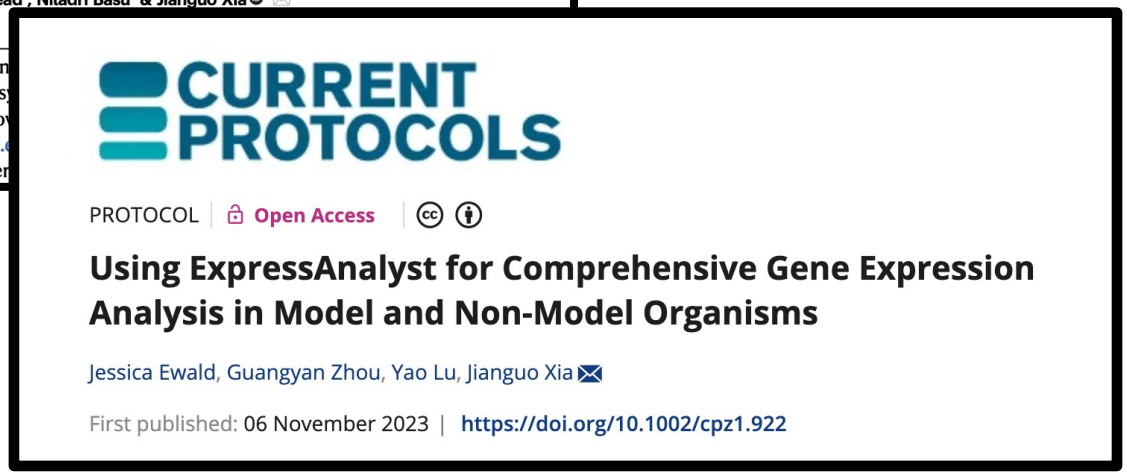
Accepted: 16 May 2023

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 increasin
demands eas
quickly uncov
nalyst (www.
ing, and inter



CURRENT PROTOCOLS

PROTOCOL | Open Access

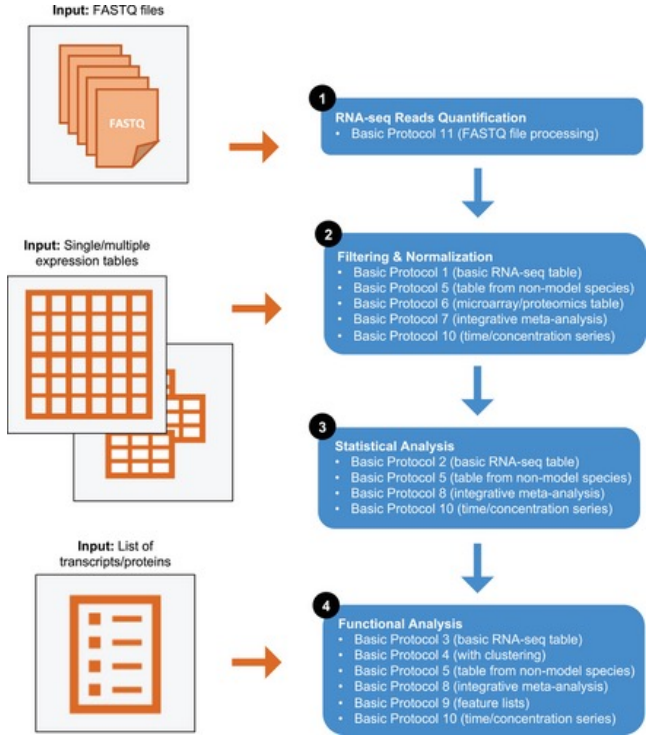
Using ExpressAnalyst for Comprehensive Gene Expression Analysis in Model and Non-Model Organisms

Jessica Ewald, Guangyan Zhou, Yao Lu, Jianguo Xia

First published: 06 November 2023 | https://doi.org/10.1002/cpz1.922

Comprehensive protocols (76 pages)

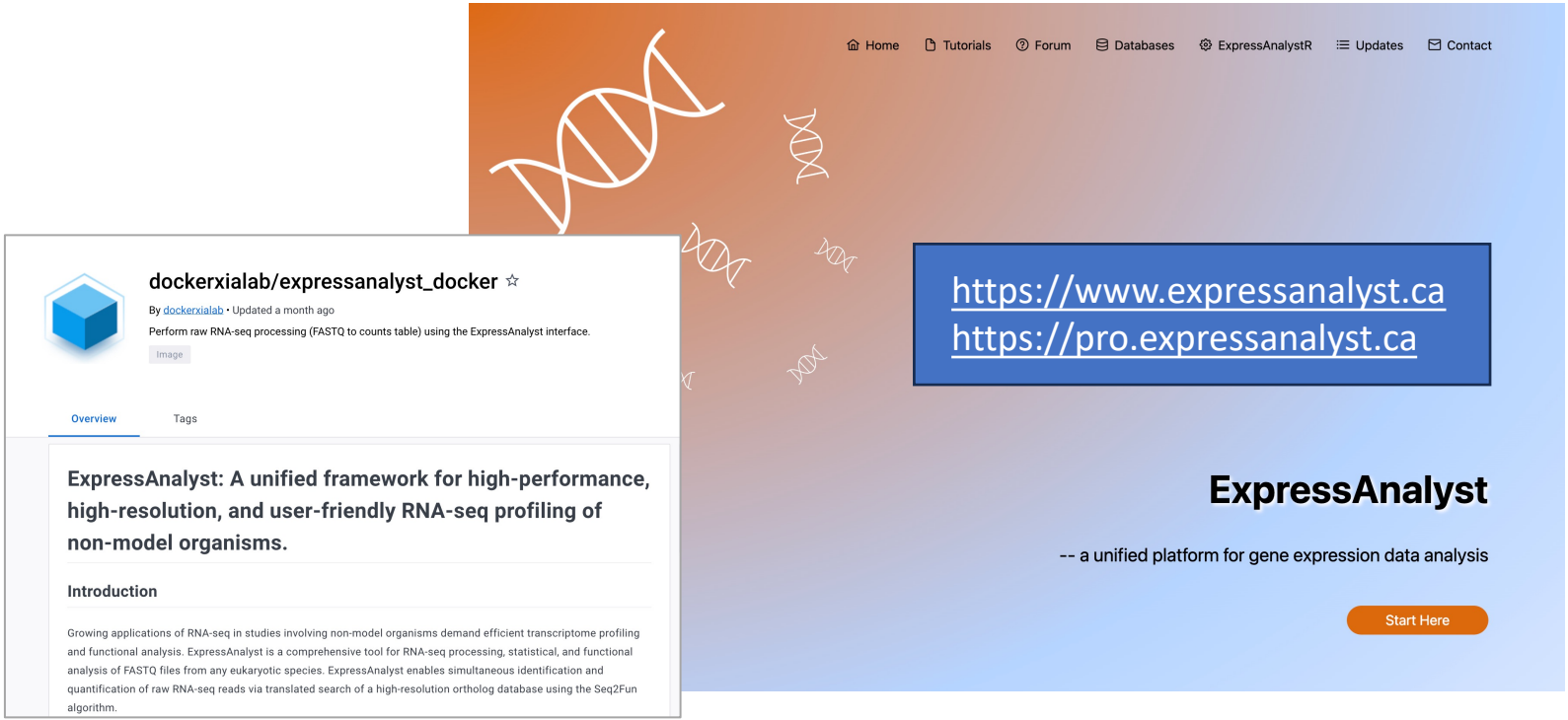
Basic Protocol 1: RNA-seq count table uploading, processing, and normalization
Basic Protocol 2: Differential expression analysis with linear models
Basic Protocol 3: Functional analysis with volcano plot, enrichment network, and ridgeline visualization
Basic Protocol 4: Hierarchical clustering analysis of transcriptomics data using interactive heatmaps
Basic Protocol 5: Cross-species gene expression analysis based on ortholog mapping results
Basic Protocol 6: Proteomics and microarray data processing and normalization
Basic Protocol 7: Preparing multiple gene expression tables for meta-analysis
Basic Protocol 8: Statistical and functional meta-analysis of gene expression data
Basic Protocol 9: Functional analysis of transcriptomics signatures
Basic Protocol 10: Dose-response and time-series data analysis
Basic Protocol 11: RNA-seq reads processing and quantification with and without reference transcriptomes



Assignment: Basic Protocol #11

➔ RNA-seq reads processing and quantification with reference transcriptomes (the “kallisto” pipeline)

ExpressAnalyst



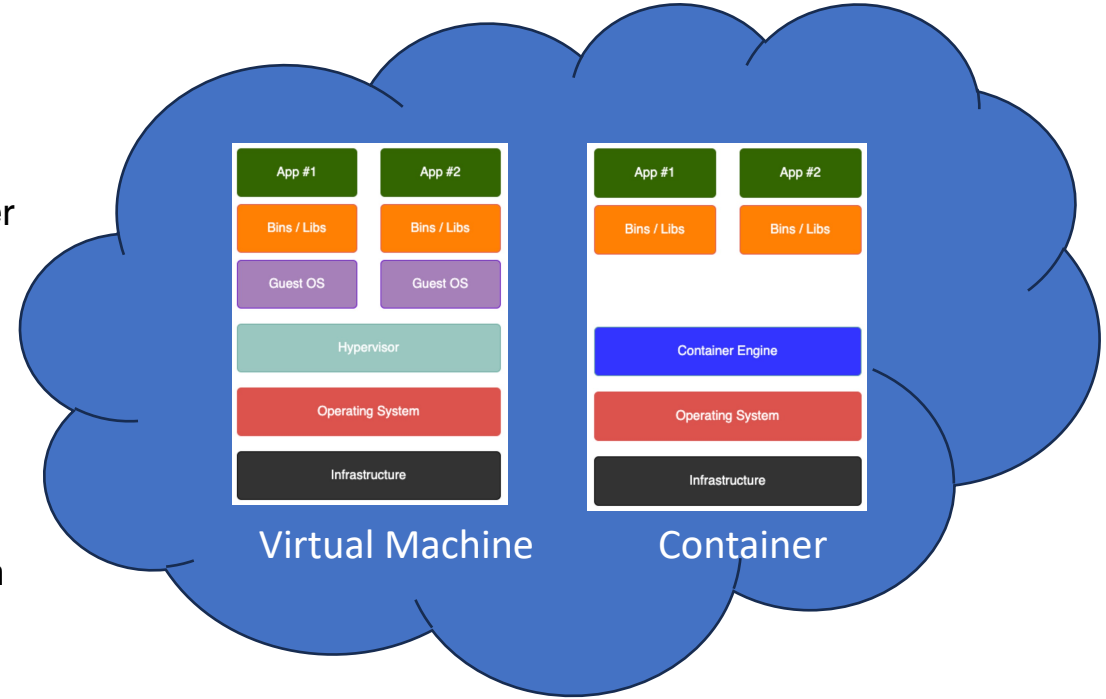
Cloud, Virtual Machine & Container (Docker)

Virtual Machine (VM)

- Hardware is virtualised.
- Each VM has its own operating system.
- Multiple VMs on a large server
- Slower start-up time

Container

- Operating system is virtualised
- It is run from an image containing all dependencies
- Multiple workloads can be run with one OS.
- Faster start-up time



ExpressAnalyst Docker for RNAseq processing

1. Download & install Docker Desktop
   - https://www.docker.com/products/docker-desktop/
2. Install ExpressAnalyst Docker
   > docker pull dockerxialab/expressanalyst_docker:latest


```
latest: Pulling from dockerxialab/expressanalyst_docker
63b65145d645: Pull complete
e1f77cd6449b: Pull complete
186f0d3f72b8: Pull complete
4301afcc17dd: Pull complete
9d2c30f7c5e0: Pull complete
1d991a6ec73f: Pull complete
db6f7732d7bd: Pull complete
a42b58f51227: Pull complete
bed8083a3a9a: Pull complete
be4ec316ba0b: Pull complete
aec28cf050c0: Pull complete
4f4fb700ef54: Pull complete
Digest: sha256:066ca3a648d4c532189d3f498d5a500892b70e127fdbc2e4f273cae51f348a26
Status: Downloaded newer image for dockerxialab/expressanalyst_docker:latest
docker.io/dockerxialab/expressanalyst_docker:latest
```

3. Find your home location

   > pwd

   - /Users/jeffxia

4. Start the Docker

   > docker run -ti --rm -p 8080:8080 -v /Users/jeffxia:/data dockerxialab/expressanalyst_docker:latest

   - The command tells Docker to attach the container to your home directory, and to create a new, temporary directory called /data to use while it is running.
   - The web is accessible: http://localhost:8080/ExpressAnalystSA

5. Create a folder “Process_Kallisto” in your Desktop
6. Download the example FASTAQ files
   https://www.xialab.ca/api/download/expressanalyst/protocol_2023_fastq.zip
   - This dataset contains 18 sub-sampled FASTQ files from double-crested cormorant.
   - Save it under the folder “FASTQ” inside “Process_Kallisto”
7. Download the reference transcriptome
   https://www.expressanalyst.ca/ExpressAnalyst/docs/Databases.xhtml
   - On the “With a Reference Transcriptome” tab, find the Gallus gallus (chicken) reference transcriptome
   - Save it under the folder “DATABASE” inside “Process_Kallisto”
8. Upload the FASTQ files and map reads to the reference transcriptome

We would like to hear your comment & feedback

contact@xialab.ca

See (most of) you next week!