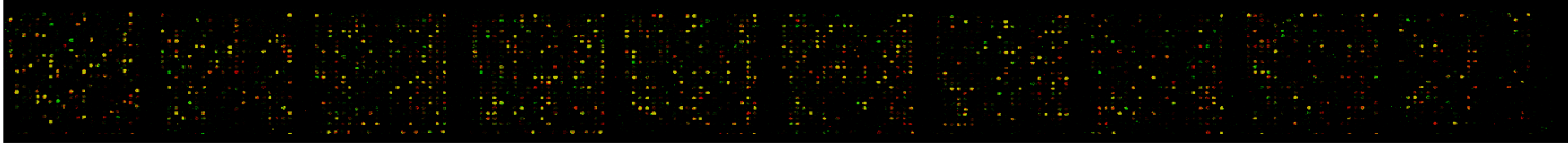
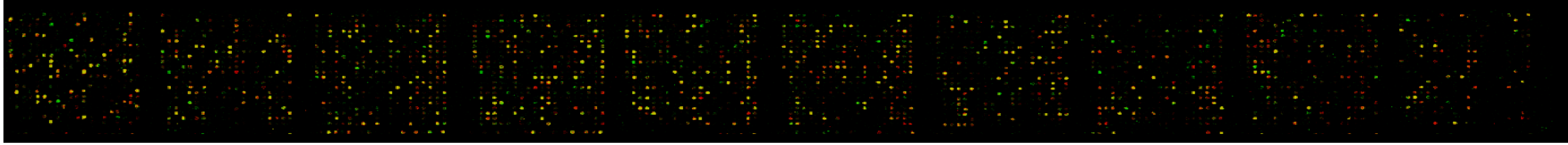


Transcriptomics –

Global transcript analysis

Dr Peter Kille
Cardiff School of Biosciences



-ome refers to totality of something (https://en.wikipedia.org/wiki/List_of_omics_topics_in_biology)

‘Omics

Hypothesis independent investigation

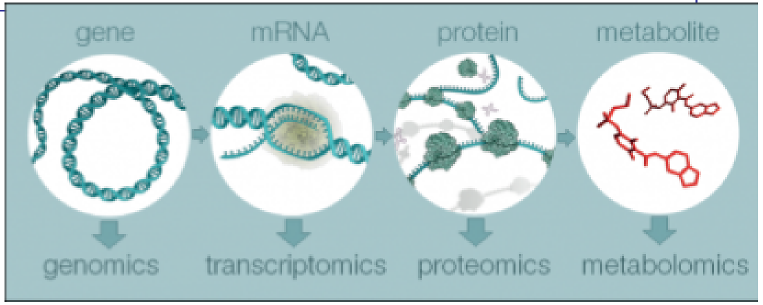
- ‘Treatment’ Vs Control
- Comparative not quantitative (qPCR)

Omics

Omics = study of *[fill in the word]* content within cell/tissues/ organism

> characterization and quantification

> structure, function, evolution



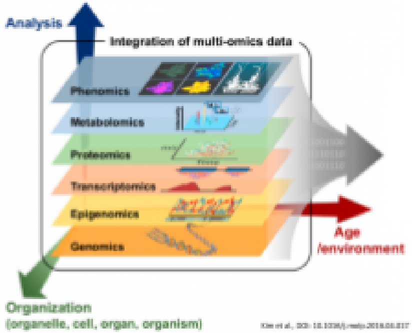
Genomics (*DNA and genetic information*)

Transcriptomics (*RNA*)

Proteomics (*Protein*)

Metabolomics (*small molecules* = *metabolites*)

Epigenomics *(reversible genome modifications ~ "chemical tags")*



Kim et al., DOI: 10.1016/j.molp.2016.03.017

Comparative genomics (~ evolution)

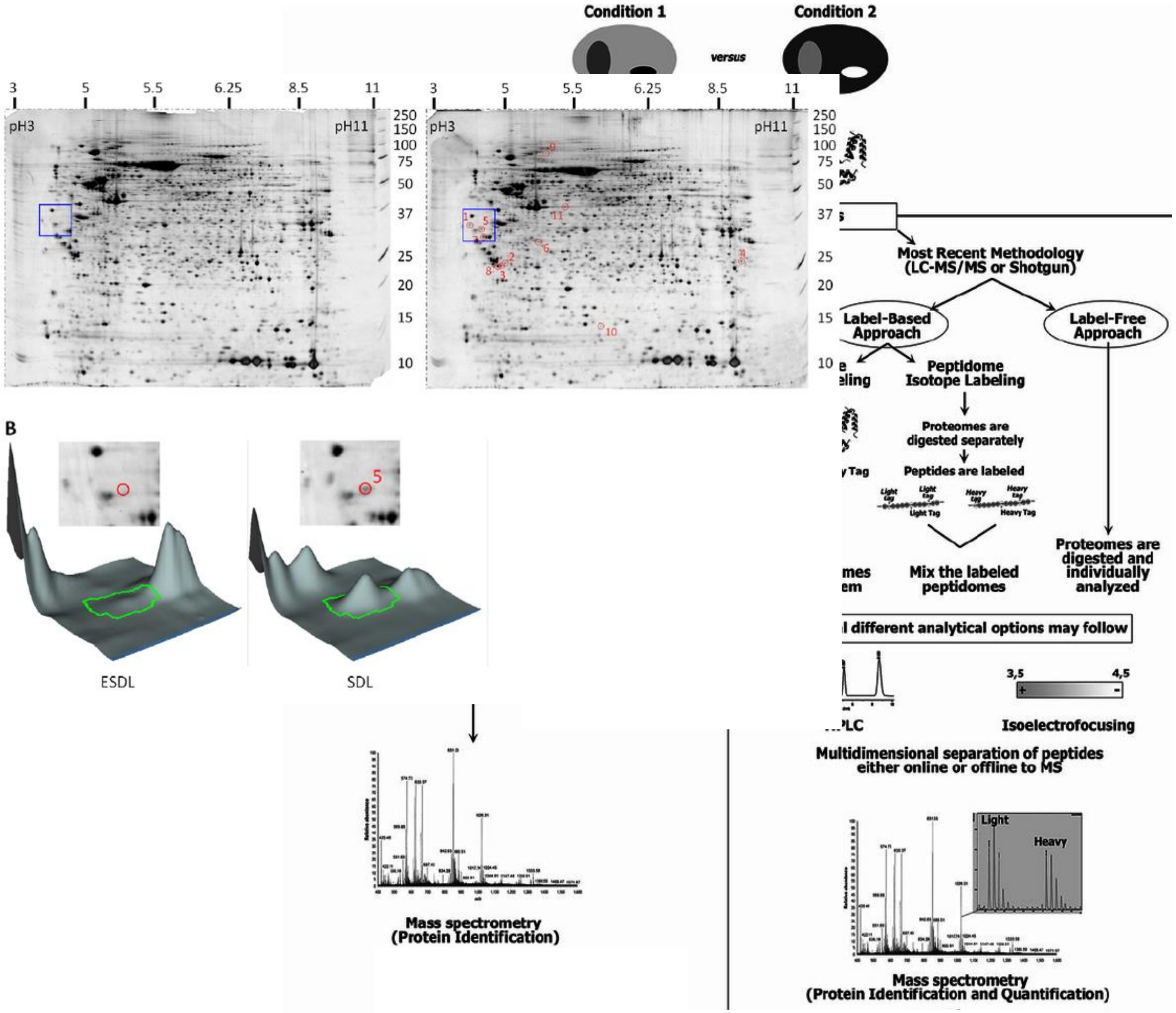
Metagenomics (~ environmental genomics)

New terms are popping up all the times: e.g. nutrigenomics (relation between diet and genes)

Transcriptomics = global analysis of gene expression (RNA)

- Qualitatively identify genes are expressed/not
- Quantitatively measure the varying levels of expression

Proteomics: 2D gels Vs Liquid Chromatography
PMF Vs MS-MS



Most Recent Methodology (LC-MS/MS or Shotgun)

Label-Based Approach

Label-Free Approach

e
eling

Peptidome Isotope Labeling

Proteomes are digested separately

y Tag

Peptides are labeled

mes
em

Mix the labeled peptidomes

Proteomes are digested and individually analyzed

l different analytical options may follow

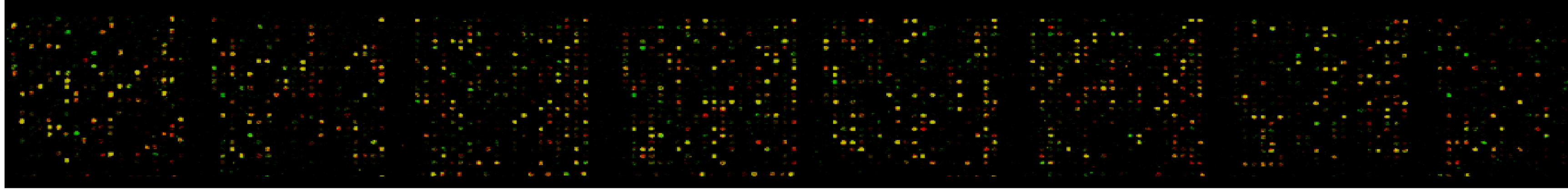
...PLC

Isoelectrofocusing

Multidimensional separation of peptides either online or offline to MS

Mass spectrometry (Protein Identification)

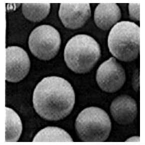
Mass spectrometry (Protein Identification and Quantification)

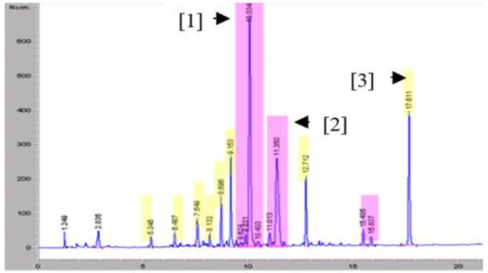


Metabolomics

HPLC

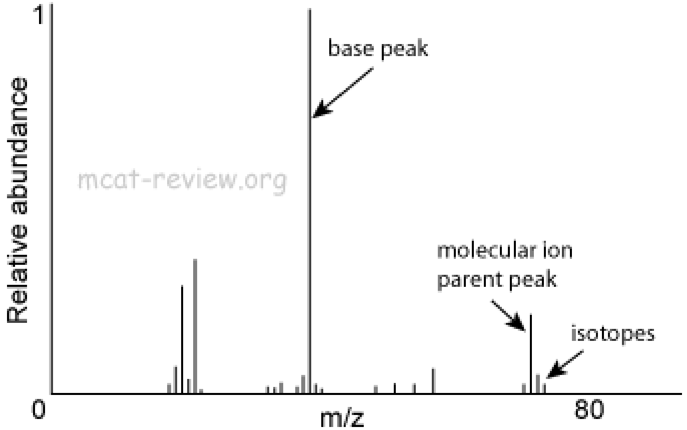
high pressure (performance) liquid chromatography





MS

Mass spectroscopy

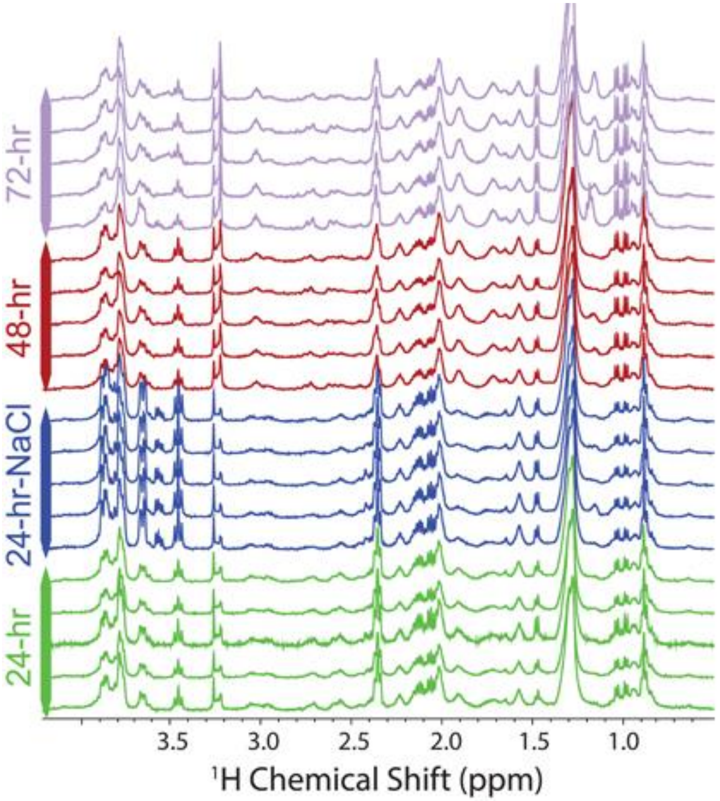


GC

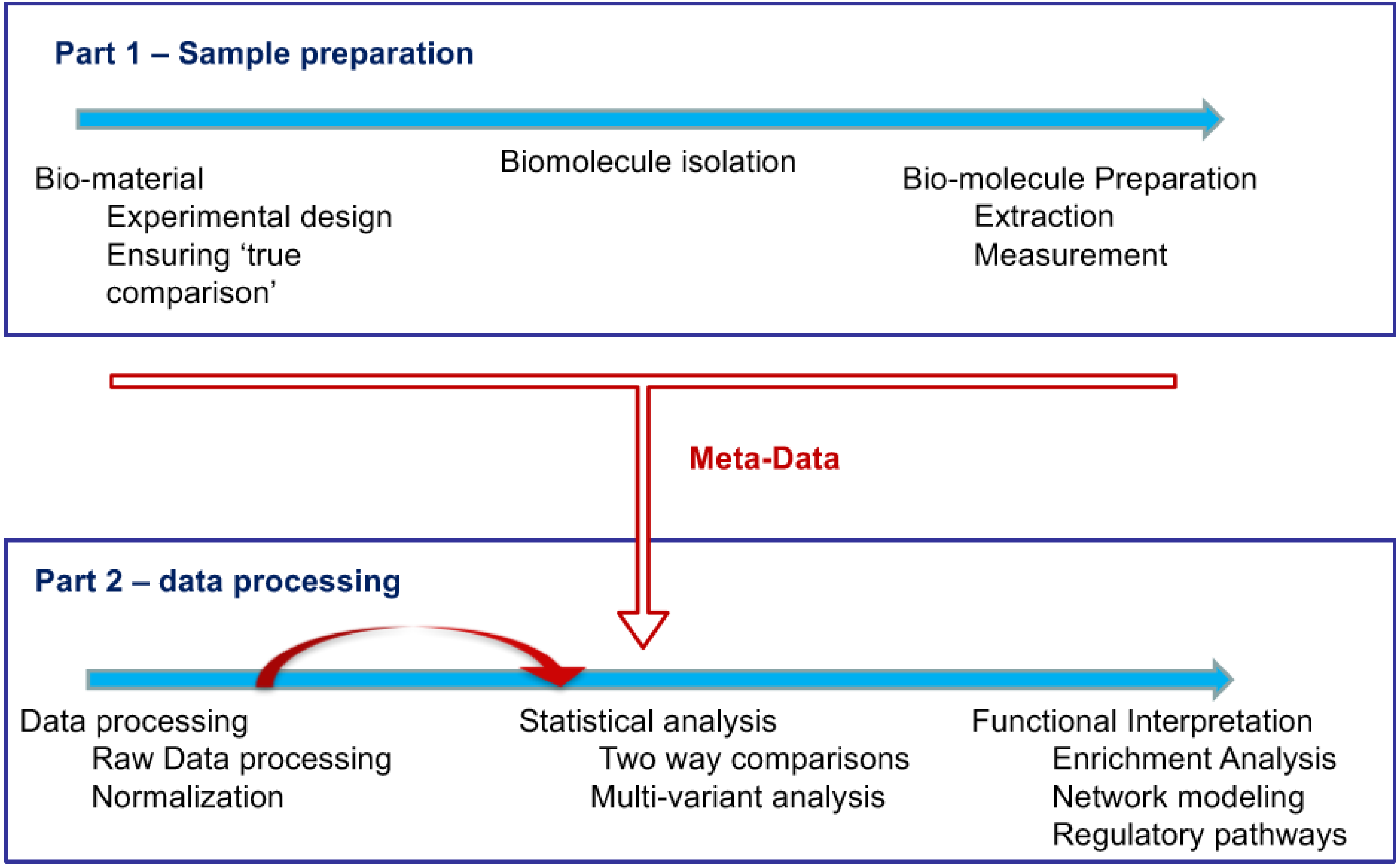
Gas chromatography

NMR

Nuclear Magnetic Resonance



The ‘Omics Process



Part 1 – Sample preparation

Bio-material
- Experimental design
- Ensuring ‘true comparison’

Biomolecule isolation

Bio-molecule Preparation
- Extraction
- Measurement

Meta-Data

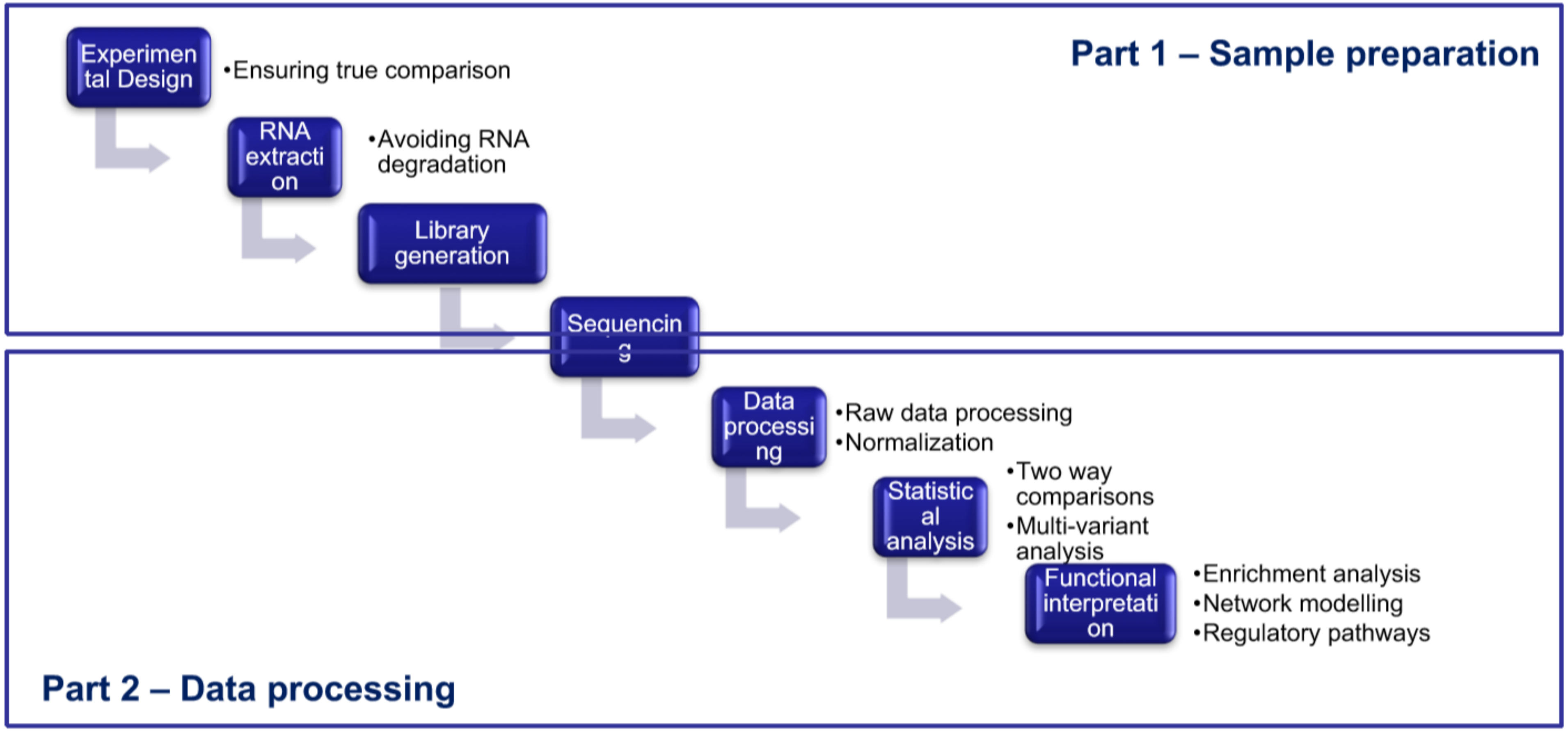
Part 2 – data processing

Data processing
- Raw Data processing
- Normalization

Statistical analysis
- Two way comparisons
- Multi-variant analysis

Functional Interpretation
- Enrichment Analysis
- Network modeling
- Regulatory pathways

Transcriptomics process



Part 1 – Sample preparation

Experimental Design
- Ensuring true comparison

RNA extraction
- Avoiding RNA degradation

Library generation

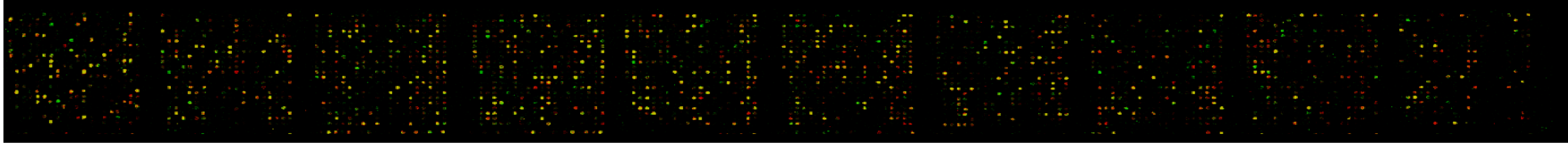
Sequencing

Data processing
- Raw data processing
- Normalization

Statistical analysis
- Two way comparisons
- Multi-variant analysis

Functional interpretation
- Enrichment analysis
- Network modelling
- Regulatory pathways

Part 2 – Data processing



Experimental design is everything

Biosample

Biomaterial preparation: Ruling out confounding factors

Consider:

Sampling time

- Circadian influence
- Dark/Light

Operator

- Animal handle

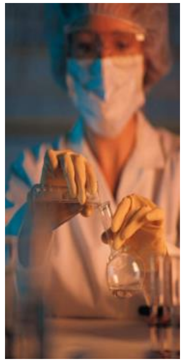
Underlying Physiology

- Sex
- Genetic similarity
- General stressors

Experimental setup

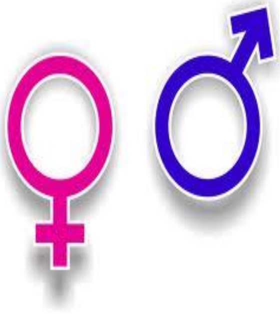
- Position in incubator
- Time to sample tissue

Sampling Time:
Circadian Influence



Operator:
Animal Handler

Underlying Physiology:
Sex



Experimental Infrastructure:
Position in incubator

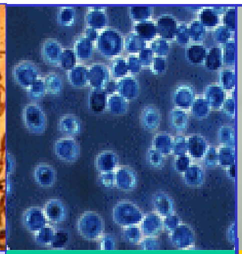
Biomaterial preparation: Sample heterogeneity

Whole organisms

Tissues

Cell populations

Bacteria


Yeast

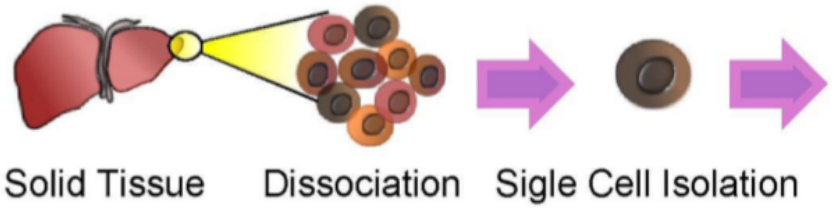
Nematode

Liver

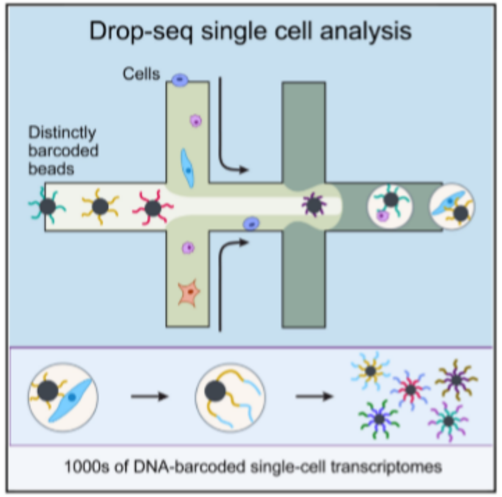
Kidney

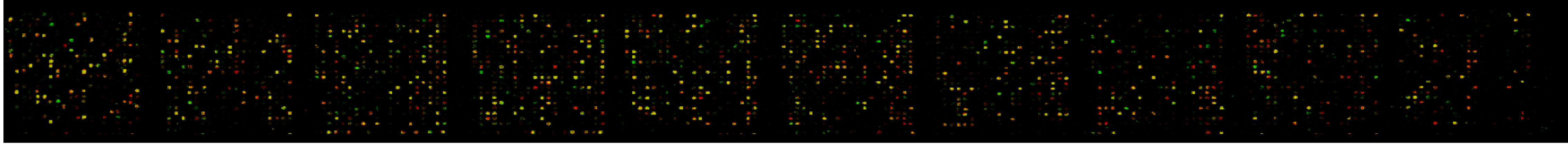
Brain

Single cell transcriptomics



Many techniques (e.g. Drop-seq)


Drop-seq single cell analysis
Cells
Distinctly barcoded beads
1000s of DNA-barcoded single-cell transcriptomes

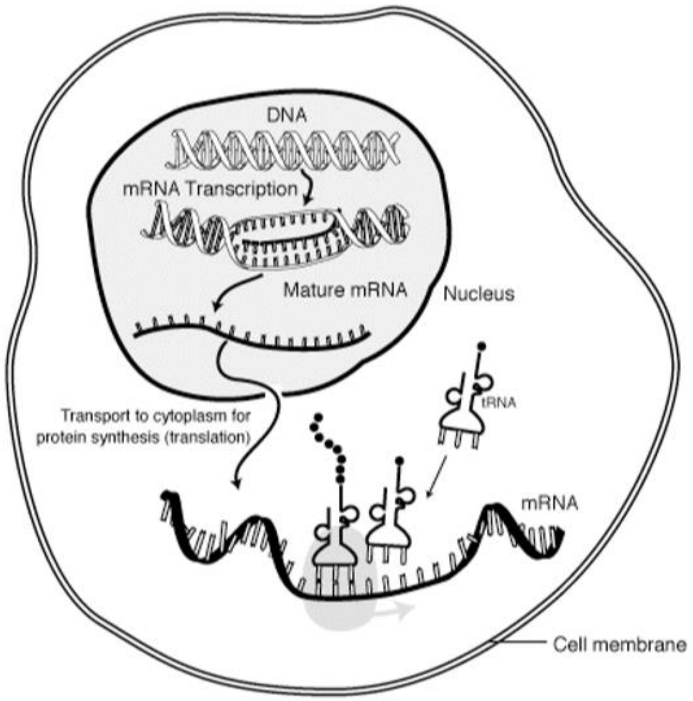


Quality and quantity

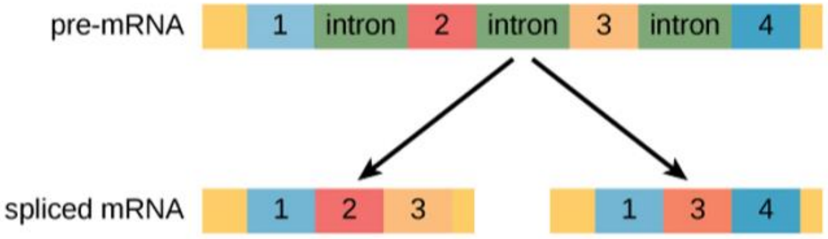
Biomolecule Isolation, Quantification & Identification

RNA

Central Dogma
DNA -> RNA -> Protein

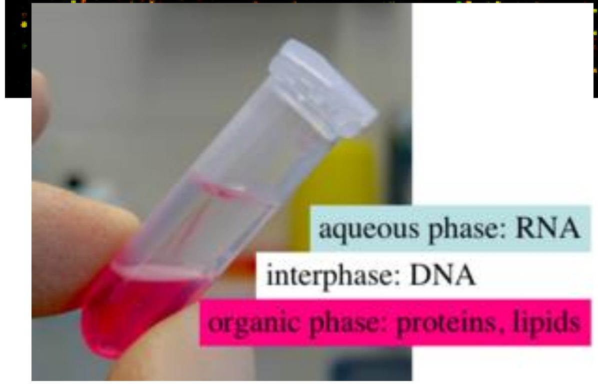


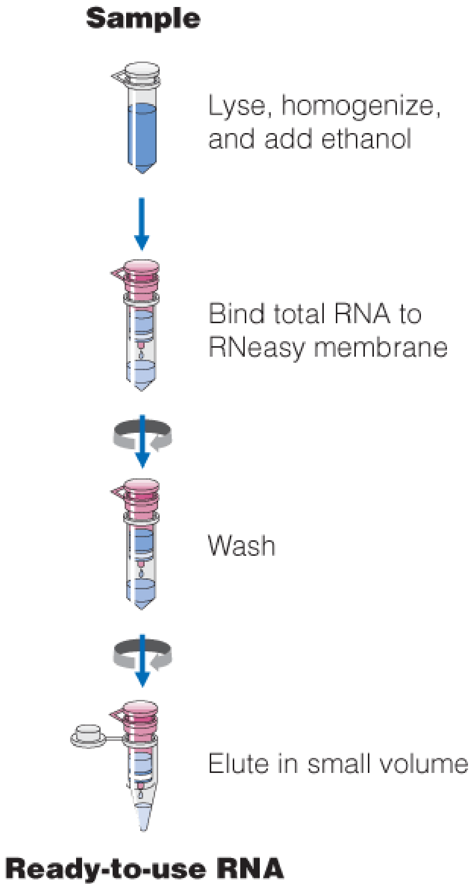
Splicing

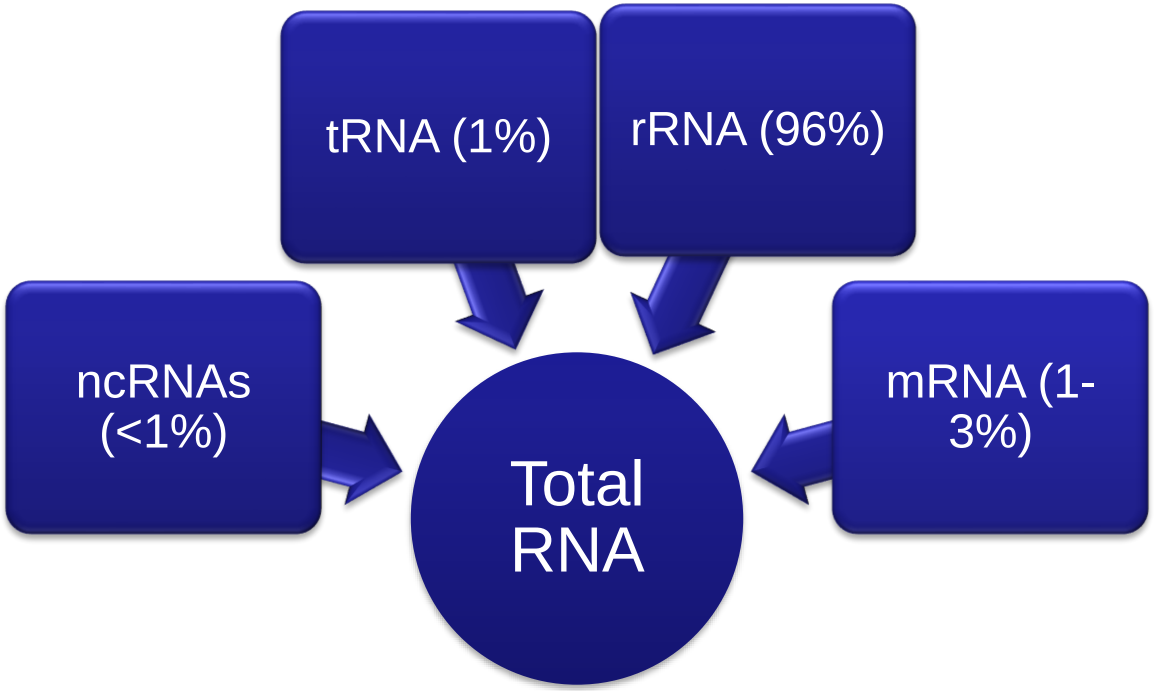


Transcript Isolation: Total RNA

Tri reagent

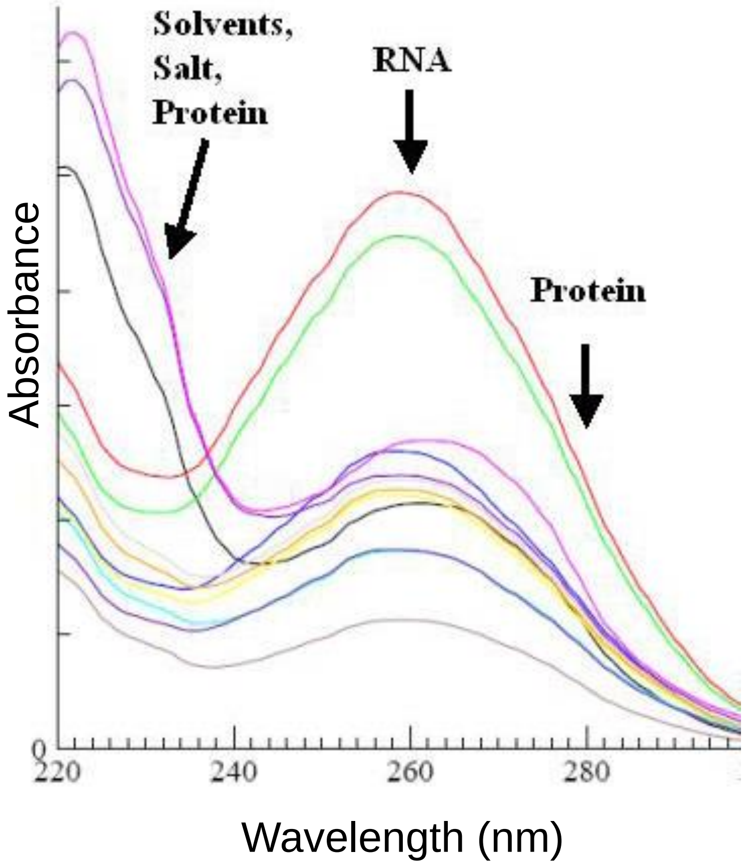


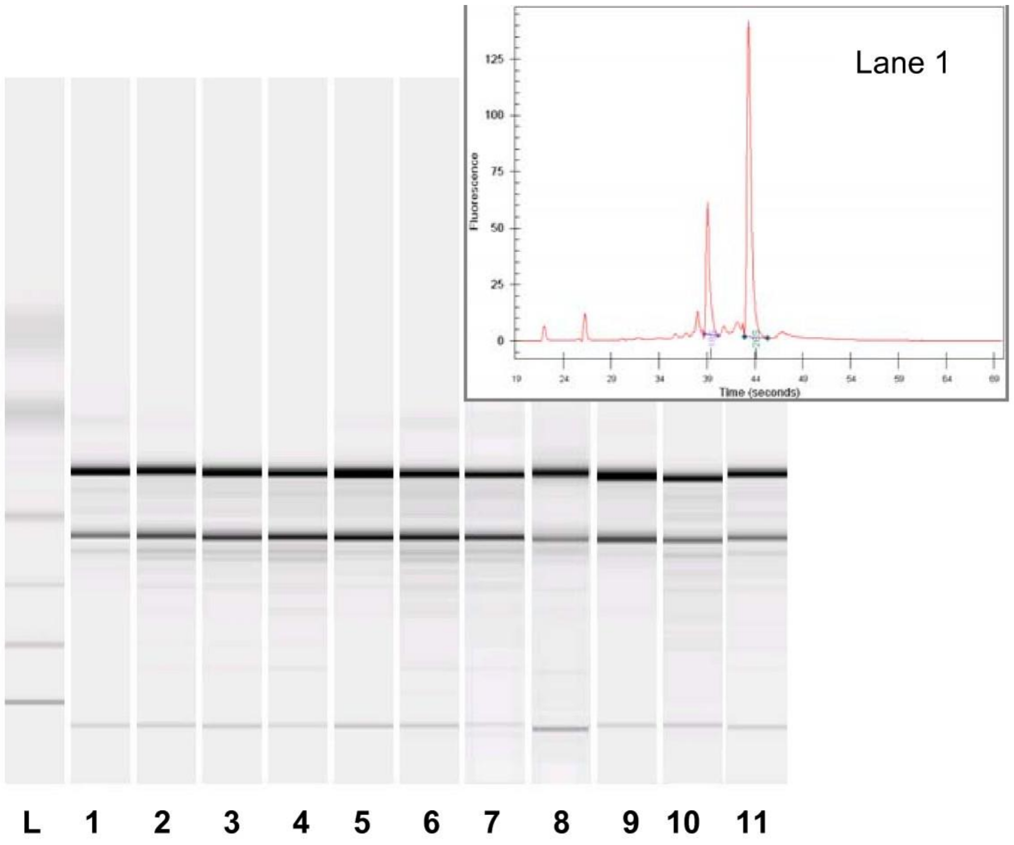

Sample
Lyse, homogenize, and add ethanol
Bind total RNA to RNeasy membrane
Wash
Elute in small volume
Ready-to-use RNA



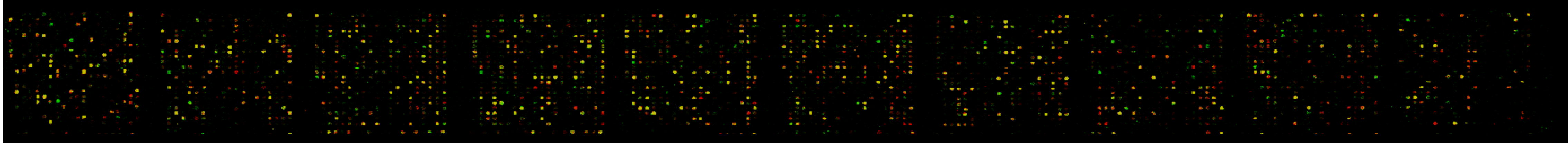
Transcript Isolation: Purity and Integrity

Purity: Spectral Analysis





Integrity: Molecular Weight Profiling



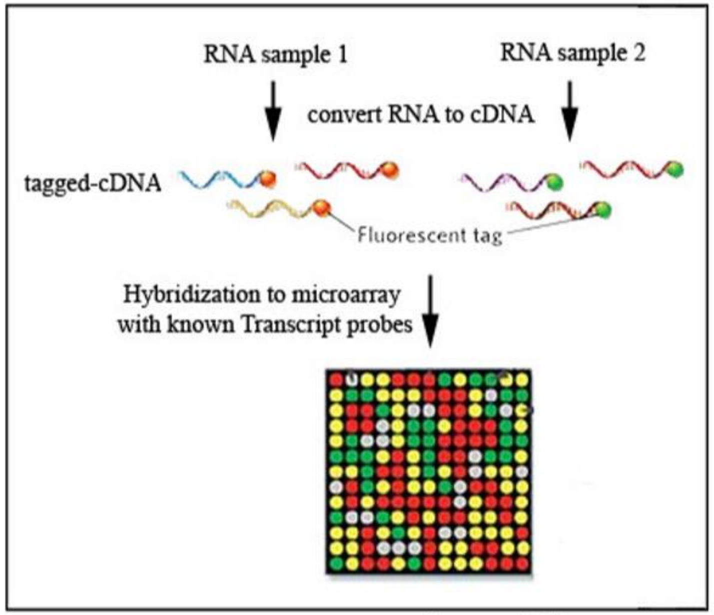
From counting decay to counting reads

Transcript quantification

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

Transcriptomics: Arrays Vs RNAseq

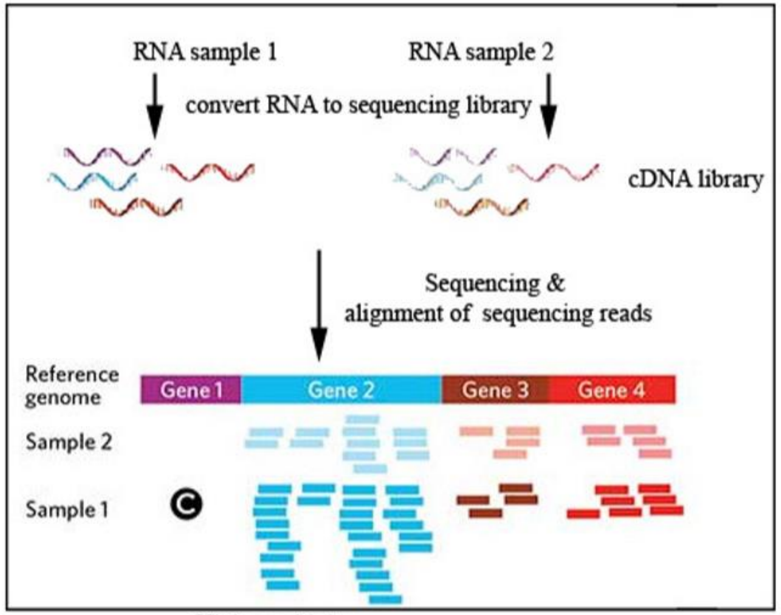
Arrays



relative intensity
=
expression levels

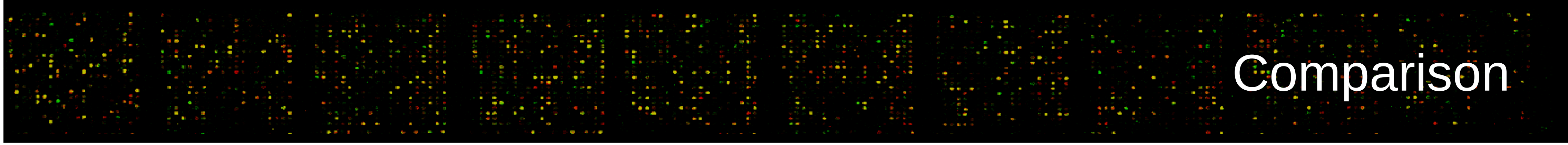
Low sensitivity
Low dynamic range
known transcript only
No alternative splicing information
lower cost

RNAseq



High sensitivity
High dynamic range
Novel transcripts sequences identified
structural variation & alternative splicing revealed
unlimited sample comparisons

Sequencing Reads
=
expression levels



Comparison

	Arrays	RNAseq
👍	Gold standards Low cost Low cost analysis “Easier”	Any organism Novel transcript identification Spliced variants Reproducible Low input ok (down to1ng)
👎	Limited to probes Input ~ 300ng-3ug	Data analysis • More complicated • Need computing power & storage capacity More expensive

Comparison

Arrays

👍

- ~~Gold standards~~
- Low cost
- Low cost analysis
- “Easier”

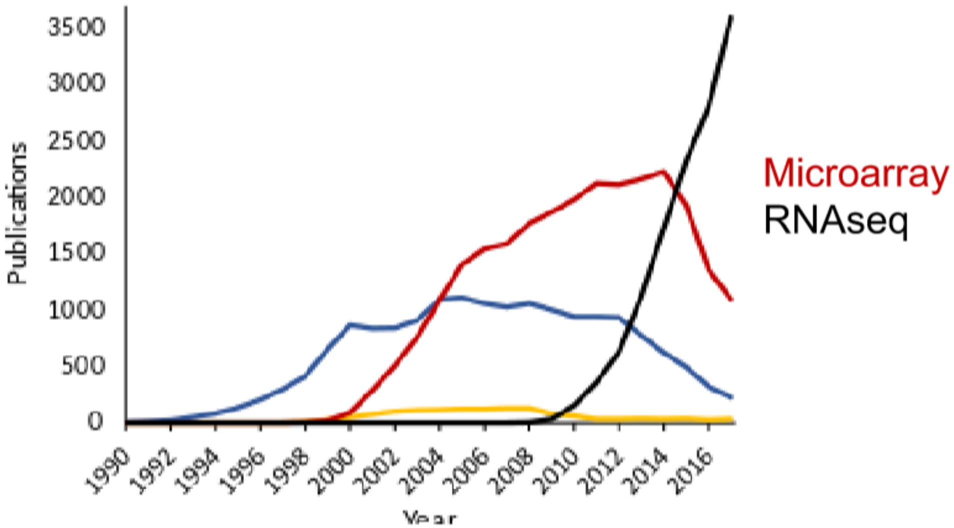
RNAseq

- ☀ Any organism
- ☀ Novel transcript i
- Spliced variants
- Reproducible
- ☀ Low input ok (down to1ng)

Exploratory work

👎

With time..



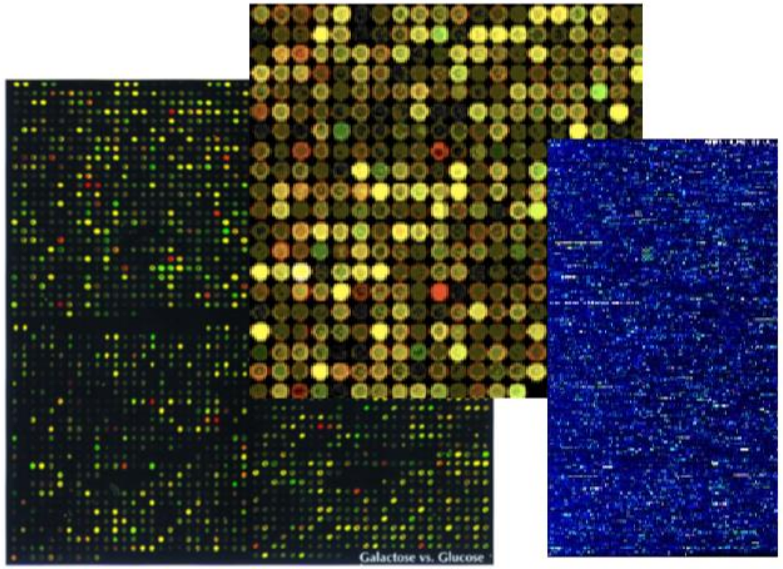
Arrays have there place

BUT

Arrays are:

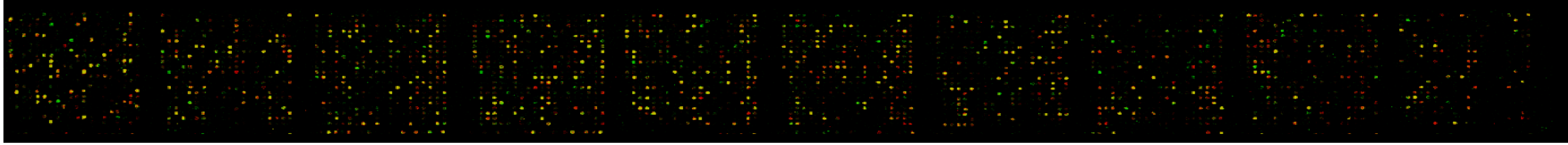
- Everywhere in the literature
- Responsible for some very pretty pictures you'll never see with RNAseq…

Arrays



VS

RNAseq



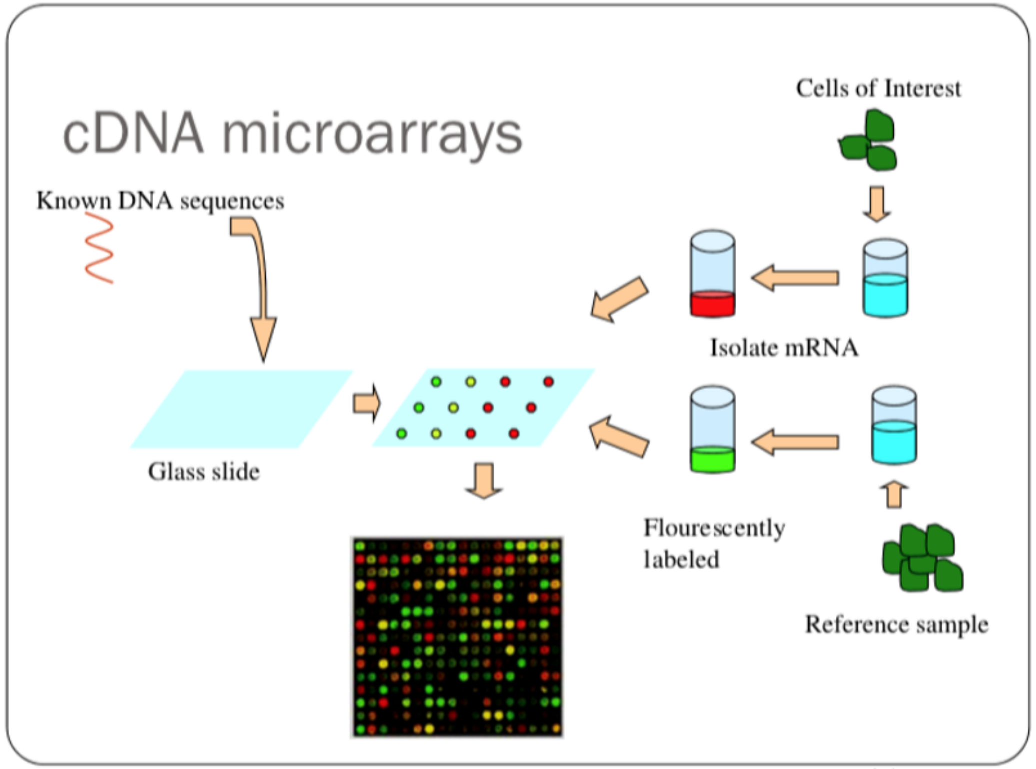
Seeing spots before your eyes

Microarrays

Arrays (Gene arrays)

- Solid supports (“DNA chip”) upon which a collection of gene-specific nucleic acids have been placed at defined locations, either by spotting or direct synthesis.
- Sample of interest is hybridized with the gene-specific targets on the array
- One-color vs two-color arrays

e.g. : two-color array

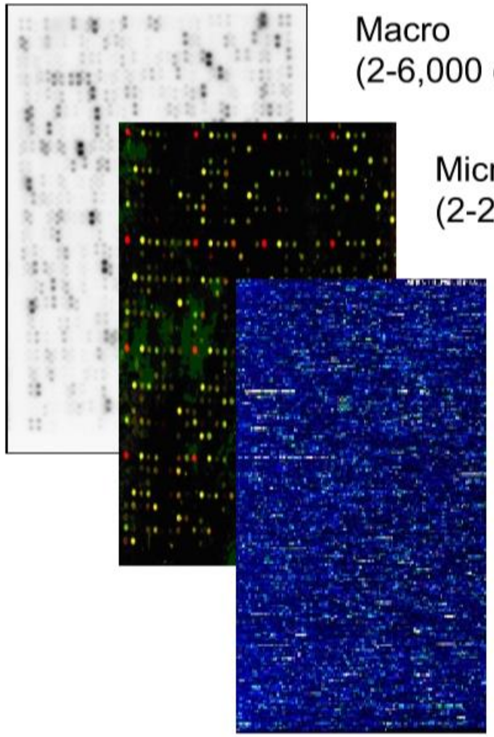


@Guy Zinman

From macro arrays to microarrays

Arrays (Gene arrays)

- Solid supports (“DNA chip”) upon which a collection of gene-specific nucleic acids have been placed at defined locations, either by spotting or direct synthesis.
- Sample of interest is hybridized with the gene-specific targets on the array



Macro
(2-6,000 elements)

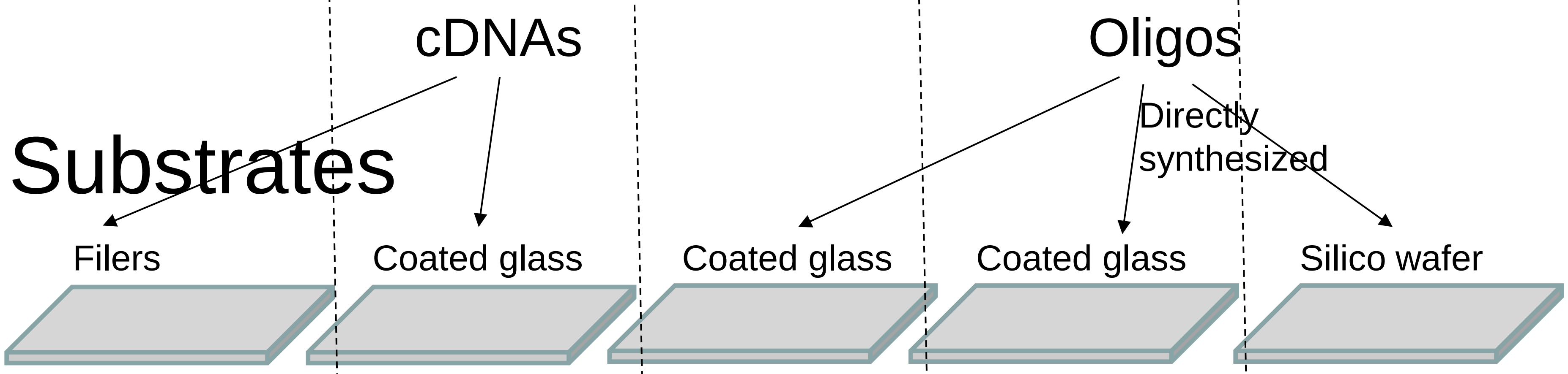
Micro (spotted)
(2-20,000 elements)

Micro (Affymetrix)
(1M-100,000 elements)

Probes, Substrates, and Target Labelling

Probes

Substrates



Target (cDNA) labelling

Radioactive
(P^{32})

Single

Fluorescence
(Cy3/Cy5)

Duel

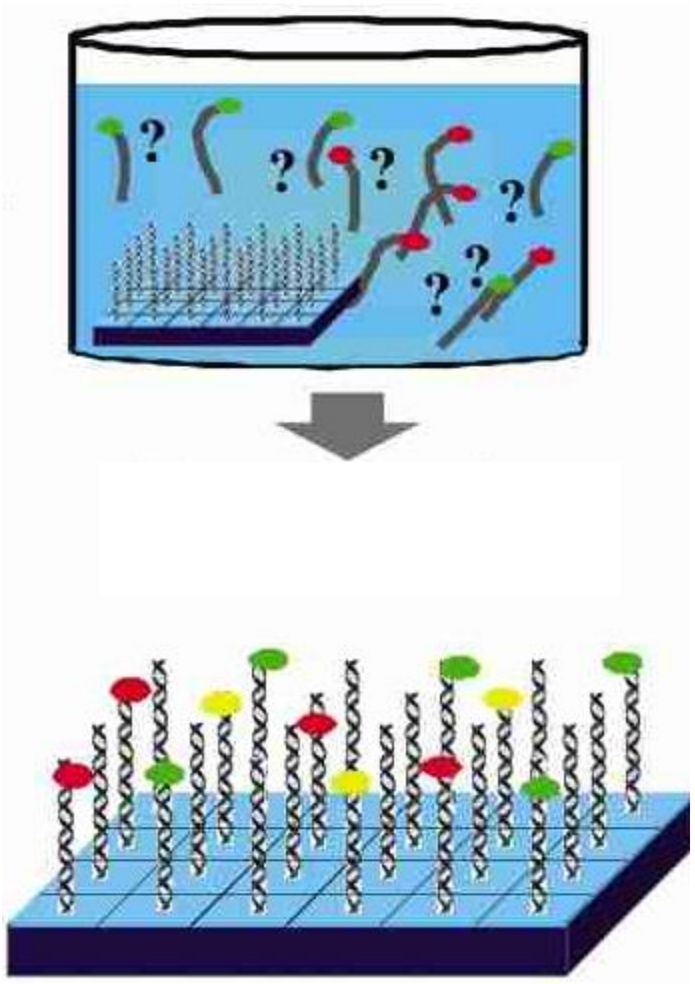
Fluorescence
(Cy3)

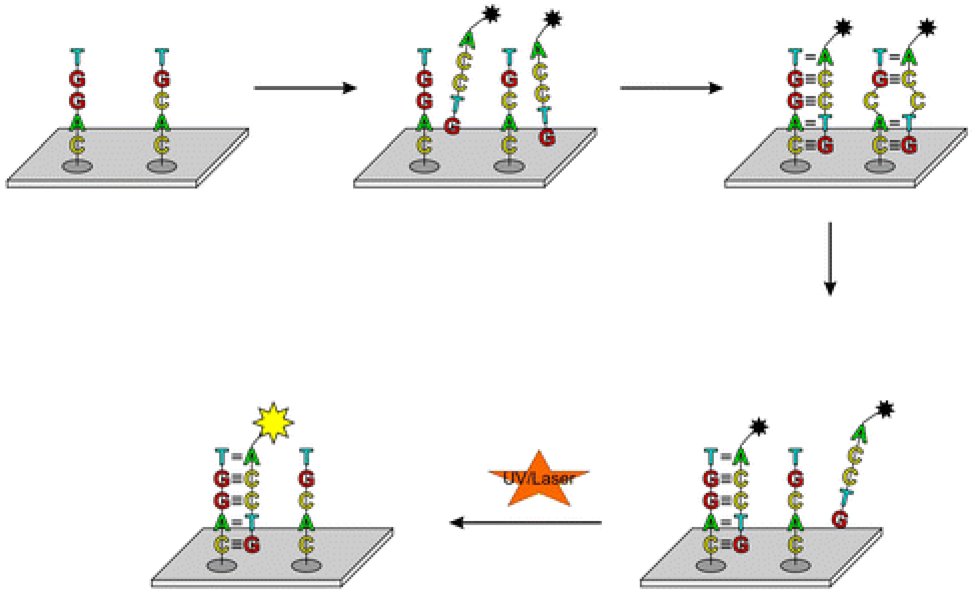
Single

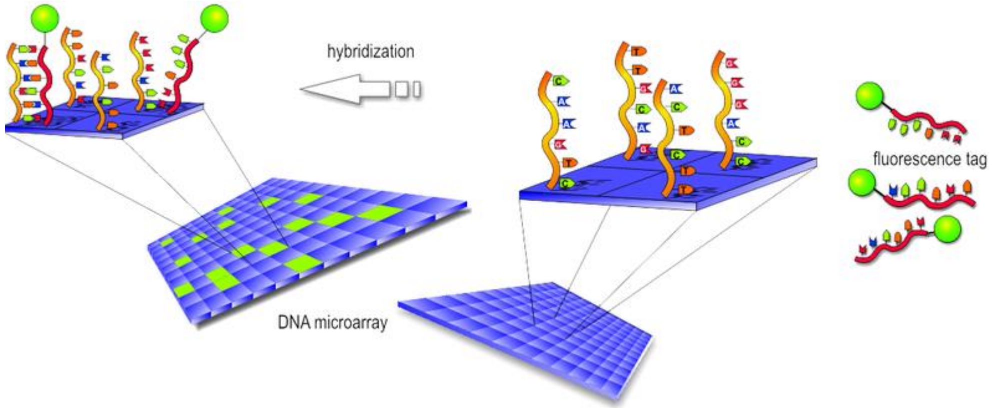
Biotin

Single

Hybridisation

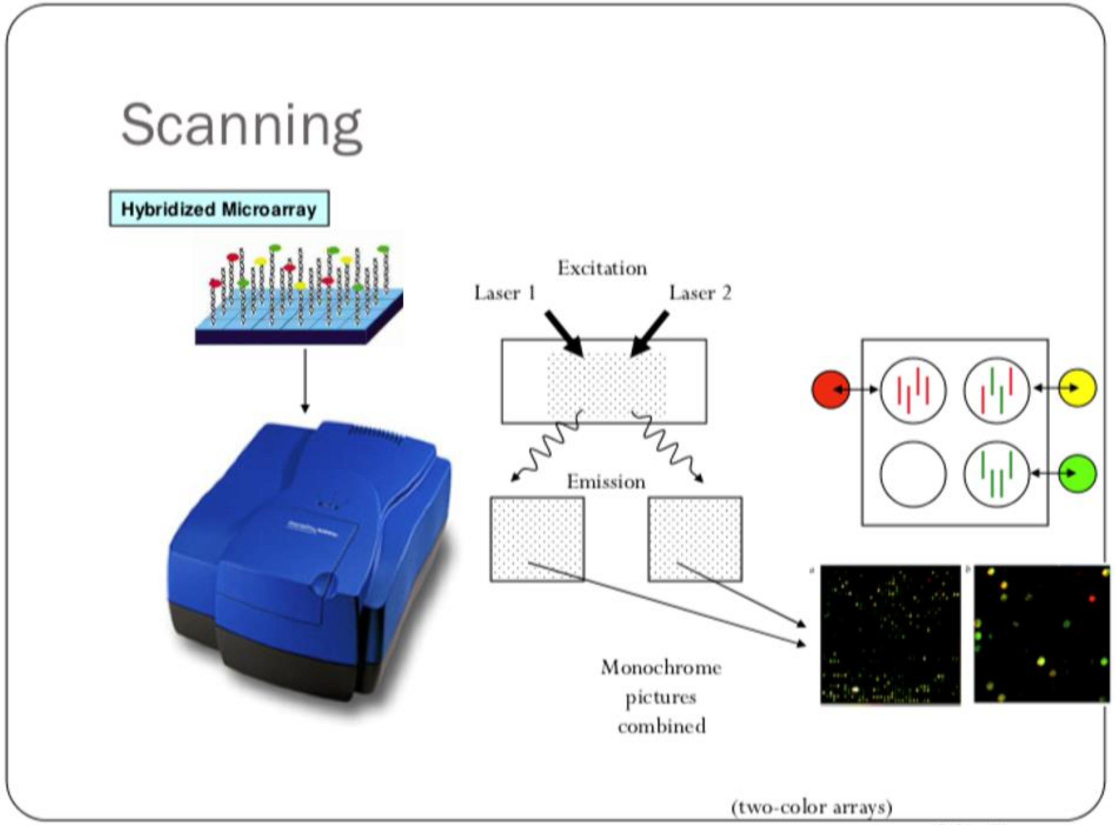






Scanning

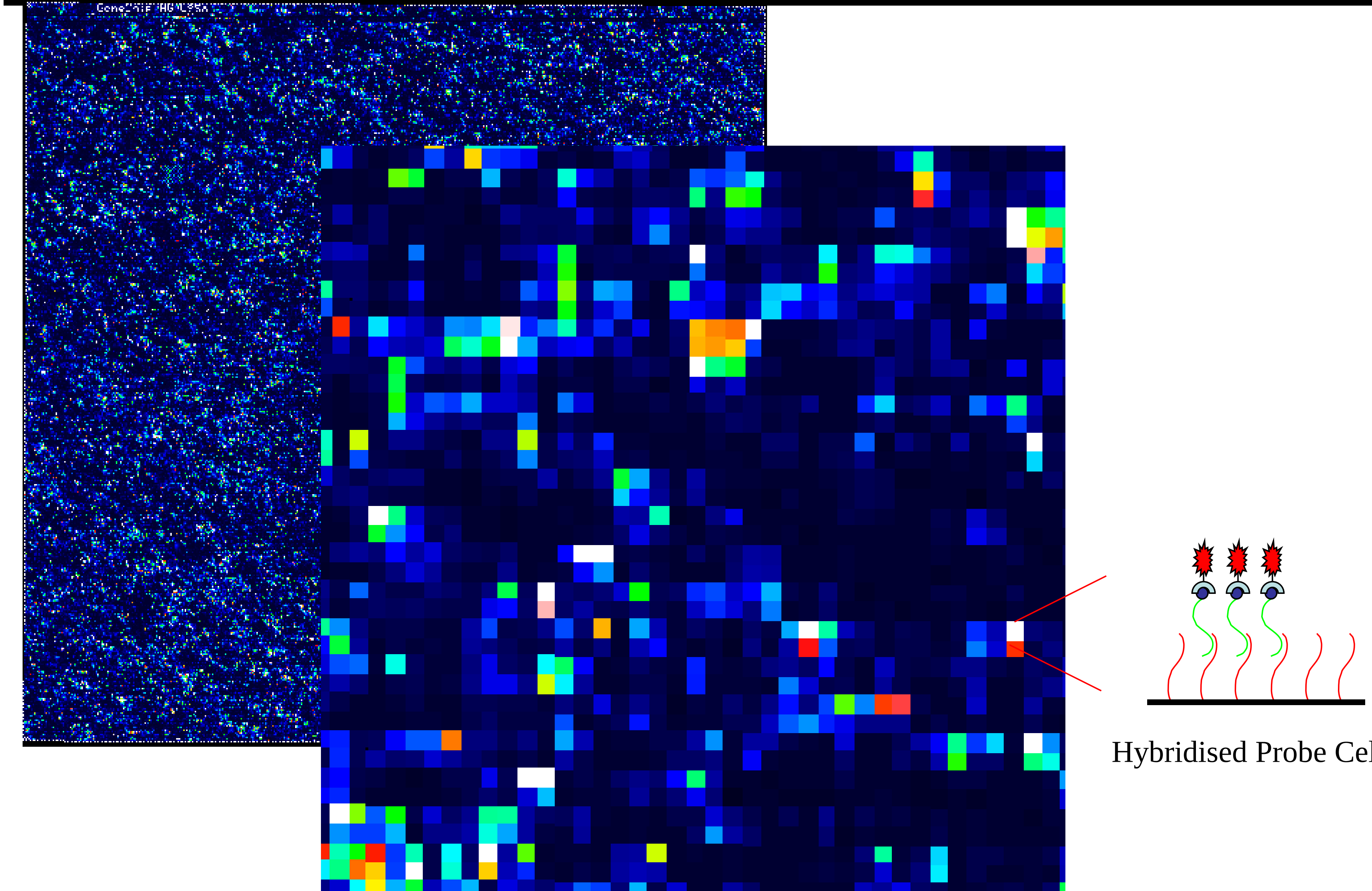
- Array is scanned to measure fluorescent label
- Ratio of red vs green ~ relative abundance of the two samples (two color array)



Scanning

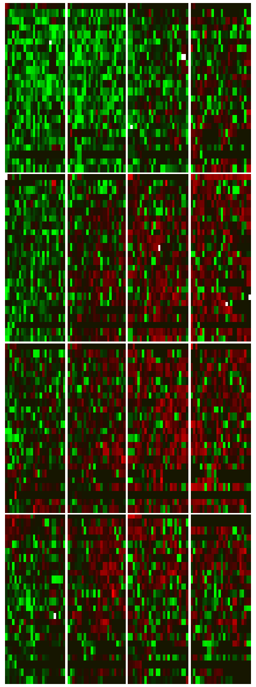
@Guy Zinman

Human Genome U95A Genechip Array

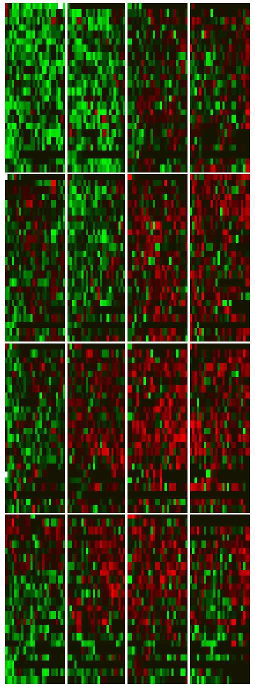


Hybridised Probe Cell

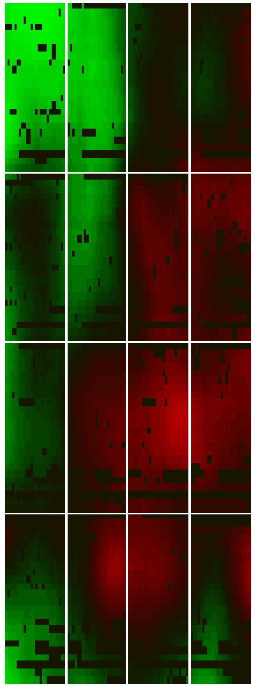
Background Correction (Normlaisation)



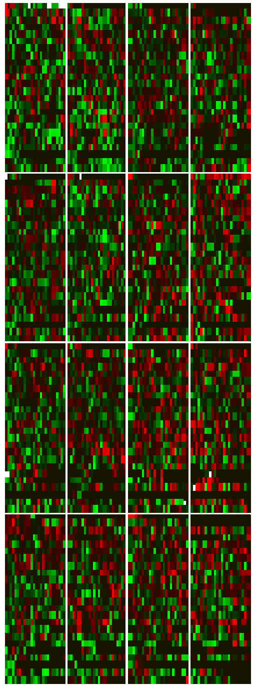
Raw data



After intensity normalization

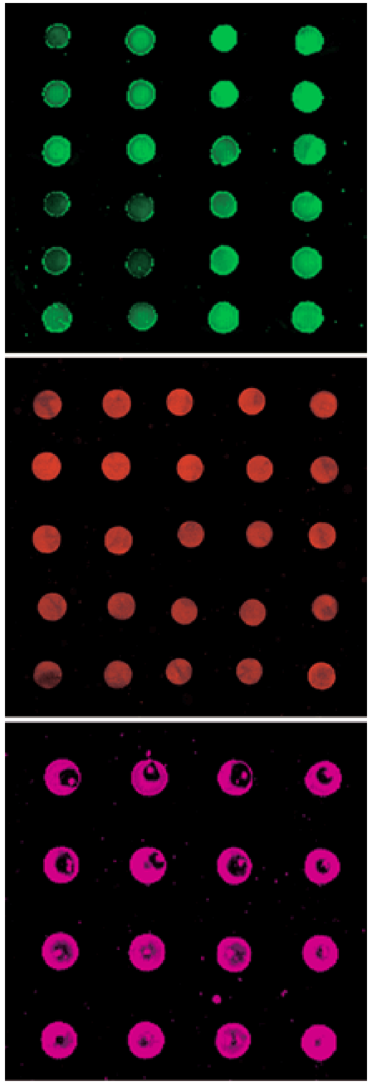


Spatial bias estimate



After spatial normalization

Spot Morphology:
Signal Vs background / Spot filtering



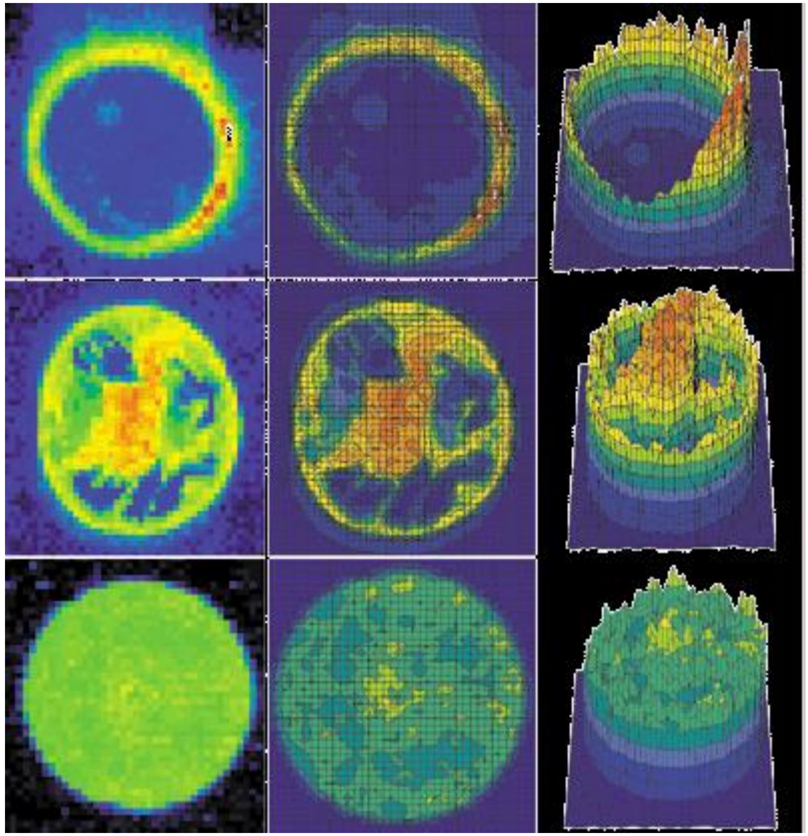
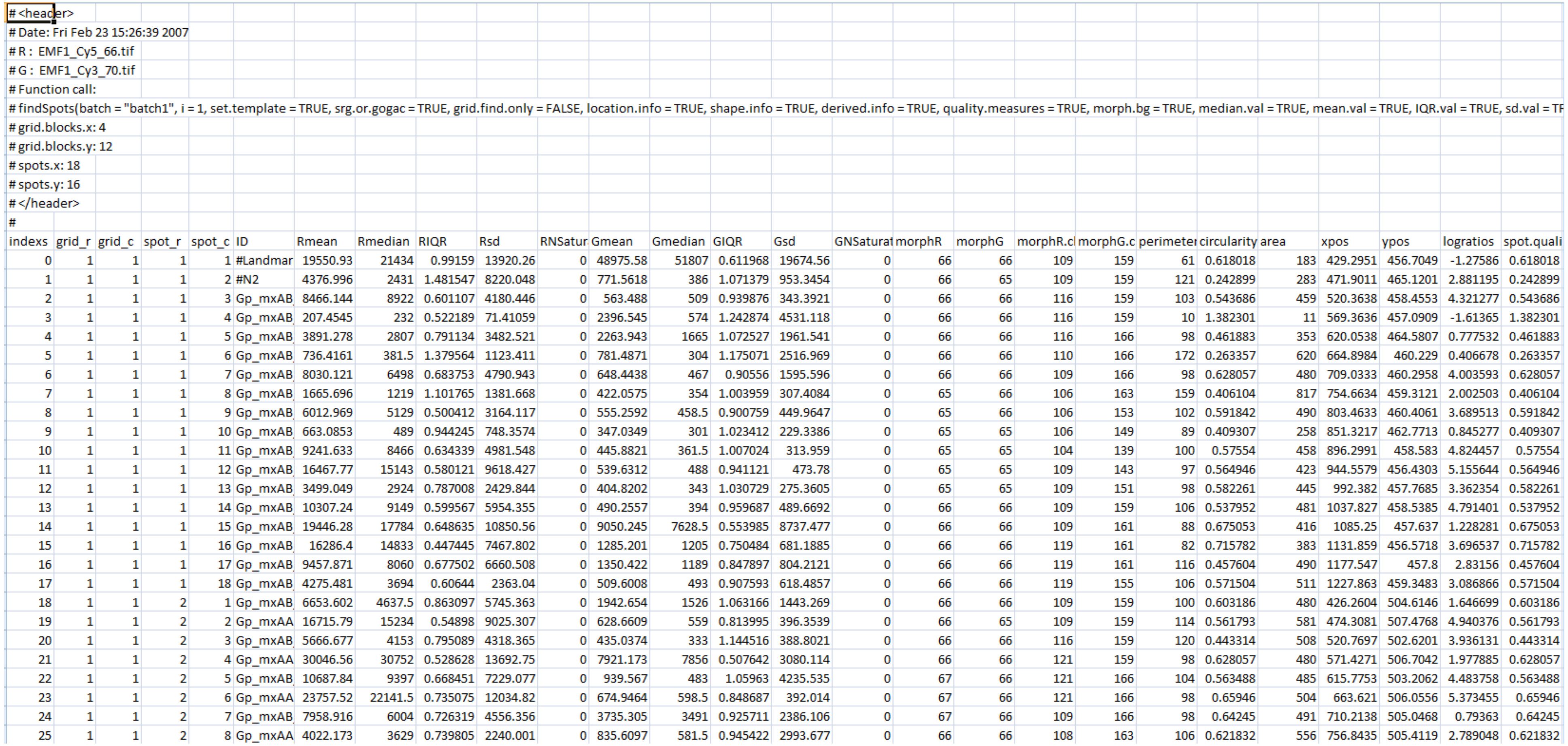
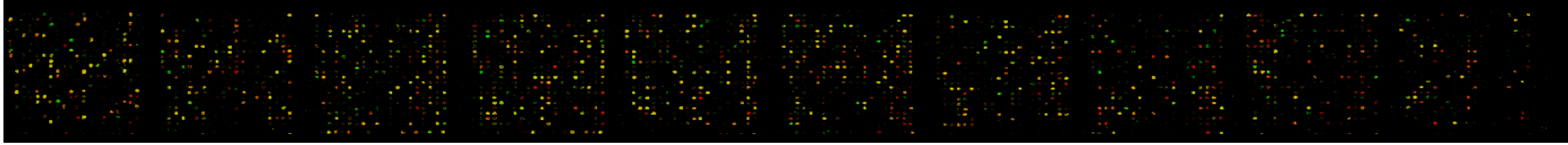


Image to numbers: Mean, median and backgrounds



```
# <header>
# Date: Fri Feb 23 15:26:39 2007
# R : EMF1_Cy5_66.tif
# G : EMF1_Cy3_70.tif
# Function call:
# findSpots(batch = "batch1", i = 1, set.template = TRUE, srg.or.gogac = TRUE, grid.find.only = FALSE, location.info = TRUE, shape.info = TRUE, derived.info = TRUE, quality.measures = TRUE, morph.bg = TRUE, median.val = TRUE, mean.val = TRUE, IQR.val = TRUE, sd.val = TF
# grid.blocks.x: 4
# grid.blocks.y: 12
# spots.x: 18
# spots.y: 16
# </header>
#
```

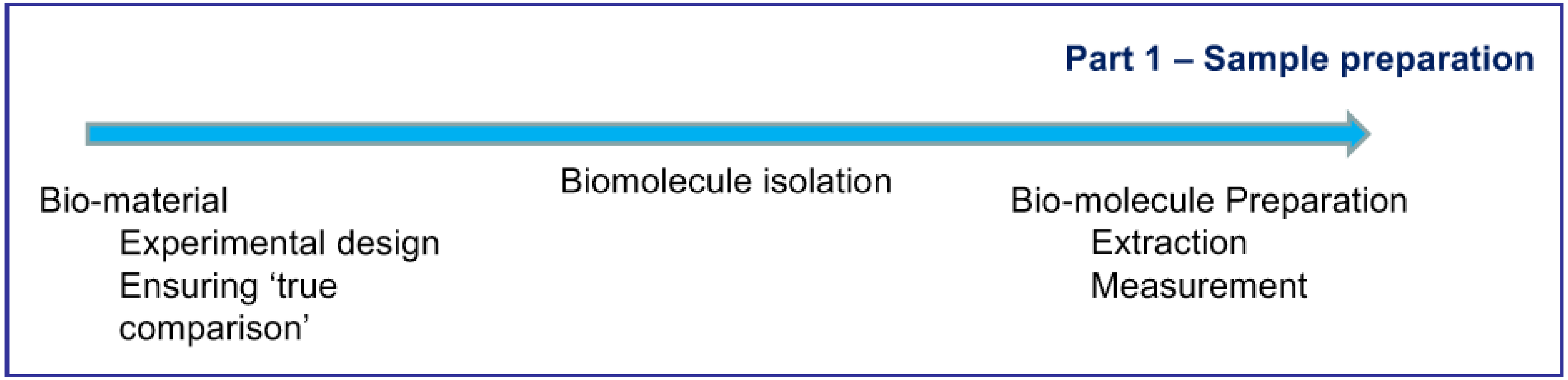
indexs	grid_r	grid_c	spot_r	spot_c	ID	Rmean	Rmedian	RIQR	Rsd	RNSatur	Gmean	Gmedian	GIQR	Gsd	GNSaturat	morphR	morphG	morphR.cl	morphG.c	perimeter	circularity	area	xpos	ypos	logratios	spot.quali
0	1	1	1	1	#Landmar	19550.93	21434	0.99159	13920.26	0	48975.58	51807	0.611968	19674.56	0	66	66	109	159	61	0.618018	183	429.2951	456.7049	-1.27586	0.618018
1	1	1	1	2	#N2	4376.996	2431	1.481547	8220.048	0	771.5618	386	1.071379	953.3454	0	66	65	109	159	121	0.242899	283	471.9011	465.1201	2.881195	0.242899
2	1	1	1	3	Gp_mxAB	8466.144	8922	0.601107	4180.446	0	563.488	509	0.939876	343.3921	0	66	66	116	159	103	0.543686	459	520.3638	458.4553	4.321277	0.543686
3	1	1	1	4	Gp_mxAB	207.4545	232	0.522189	71.41059	0	2396.545	574	1.242874	4531.118	0	66	66	116	159	10	1.382301	11	569.3636	457.0909	-1.61365	1.382301
4	1	1	1	5	Gp_mxAB	3891.278	2807	0.791134	3482.521	0	2263.943	1665	1.072527	1961.541	0	66	66	116	166	98	0.461883	353	620.0538	464.5807	0.777532	0.461883
5	1	1	1	6	Gp_mxAB	736.4161	381.5	1.379564	1123.411	0	781.4871	304	1.175071	2516.969	0	66	66	110	166	172	0.263357	620	664.8984	460.229	0.406678	0.263357
6	1	1	1	7	Gp_mxAB	8030.121	6498	0.683753	4790.943	0	648.4438	467	0.90556	1595.596	0	66	66	109	166	98	0.628057	480	709.0333	460.2958	4.003593	0.628057
7	1	1	1	8	Gp_mxAB	1665.696	1219	1.101765	1381.668	0	422.0575	354	1.003959	307.4084	0	65	66	106	163	159	0.406104	817	754.6634	459.3121	2.002503	0.406104
8	1	1	1	9	Gp_mxAB	6012.969	5129	0.500412	3164.117	0	555.2592	458.5	0.900759	449.9647	0	65	66	106	153	102	0.591842	490	803.4633	460.4061	3.689513	0.591842
9	1	1	1	10	Gp_mxAB	663.0853	489	0.944245	748.3574	0	347.0349	301	1.023412	229.3386	0	65	65	106	149	89	0.409307	258	851.3217	462.7713	0.845277	0.409307
10	1	1	1	11	Gp_mxAB	9241.633	8466	0.634339	4981.548	0	445.8821	361.5	1.007024	313.959	0	65	65	104	139	100	0.57554	458	896.2991	458.583	4.824457	0.57554
11	1	1	1	12	Gp_mxAB	16467.77	15143	0.580121	9618.427	0	539.6312	488	0.941121	473.78	0	65	65	109	143	97	0.564946	423	944.5579	456.4303	5.155644	0.564946
12	1	1	1	13	Gp_mxAB	3499.049	2924	0.787008	2429.844	0	404.8202	343	1.030729	275.3605	0	65	65	109	151	98	0.582261	445	992.382	457.7685	3.362354	0.582261
13	1	1	1	14	Gp_mxAB	10307.24	9149	0.599567	5954.355	0	490.2557	394	0.959687	489.6692	0	66	66	109	159	106	0.537952	481	1037.827	458.5385	4.791401	0.537952
14	1	1	1	15	Gp_mxAB	19446.28	17784	0.648635	10850.56	0	9050.245	7628.5	0.553985	8737.477	0	66	66	109	161	88	0.675053	416	1085.25	457.637	1.228281	0.675053
15	1	1	1	16	Gp_mxAB	16286.4	14833	0.447445	7467.802	0	1285.201	1205	0.750484	681.1885	0	66	66	119	161	82	0.715782	383	1131.859	456.5718	3.696537	0.715782
16	1	1	1	17	Gp_mxAB	9457.871	8060	0.677502	6660.508	0	1350.422	1189	0.847897	804.2121	0	66	66	119	161	116	0.457604	490	1177.547	457.8	2.83156	0.457604
17	1	1	1	18	Gp_mxAB	4275.481	3694	0.60644	2363.04	0	509.6008	493	0.907593	618.4857	0	66	66	119	155	106	0.571504	511	1227.863	459.3483	3.086866	0.571504
18	1	1	2	1	Gp_mxAB	6653.602	4637.5	0.863097	5745.363	0	1942.654	1526	1.063166	1443.269	0	66	66	109	159	100	0.603186	480	426.2604	504.6146	1.646699	0.603186
19	1	1	2	2	Gp_mxAA	16715.79	15234	0.54898	9025.307	0	628.6609	559	0.813995	396.3539	0	66	65	109	159	114	0.561793	581	474.3081	507.4768	4.940376	0.561793
20	1	1	2	3	Gp_mxAB	5666.677	4153	0.795089	4318.365	0	435.0374	333	1.144516	388.8021	0	66	66	116	159	120	0.443314	508	520.7697	502.6201	3.936131	0.443314
21	1	1	2	4	Gp_mxAA	30046.56	30752	0.528628	13692.75	0	7921.173	7856	0.507642	3080.114	0	66	66	121	159	98	0.628057	480	571.4271	506.7042	1.977885	0.628057
22	1	1	2	5	Gp_mxAB	10687.84	9397	0.668451	7229.077	0	939.567	483	1.05963	4235.535	0	67	66	121	166	104	0.563488	485	615.7753	503.2062	4.483758	0.563488
23	1	1	2	6	Gp_mxAA	23757.52	22141.5	0.735075	12034.82	0	674.9464	598.5	0.848687	392.014	0	67	66	121	166	98	0.65946	504	663.621	506.0556	5.373455	0.65946
24	1	1	2	7	Gp_mxAB	7958.916	6004	0.726319	4556.356	0	3735.305	3491	0.925711	2386.106	0	67	66	109	166	98	0.64245	491	710.2138	505.0468	0.79363	0.64245
25	1	1	2	8	Gp_mxAA	4022.173	3629	0.739805	2240.001	0	835.6097	581.5	0.945422	2993.677	0	66	66	108	163	106	0.621832	556	756.8435	505.4119	2.789048	0.621832



Elegance of counting

RNAseq

RNAseq workflow

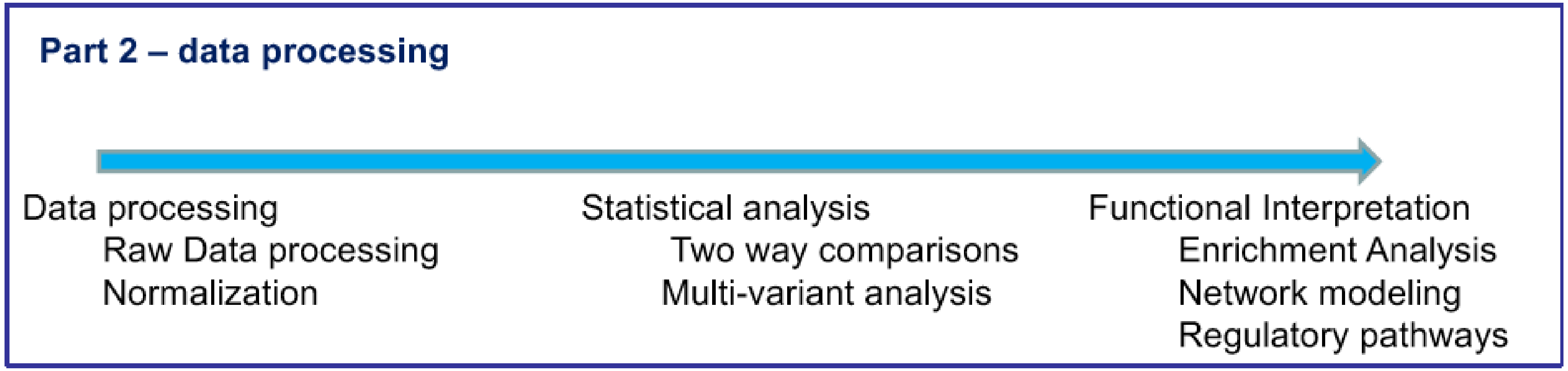


Part 1 – Sample preparation

Bio-material
- Experimental design
- Ensuring ‘true comparison’

Biomolecule isolation

Bio-molecule Preparation
- Extraction
- Measurement



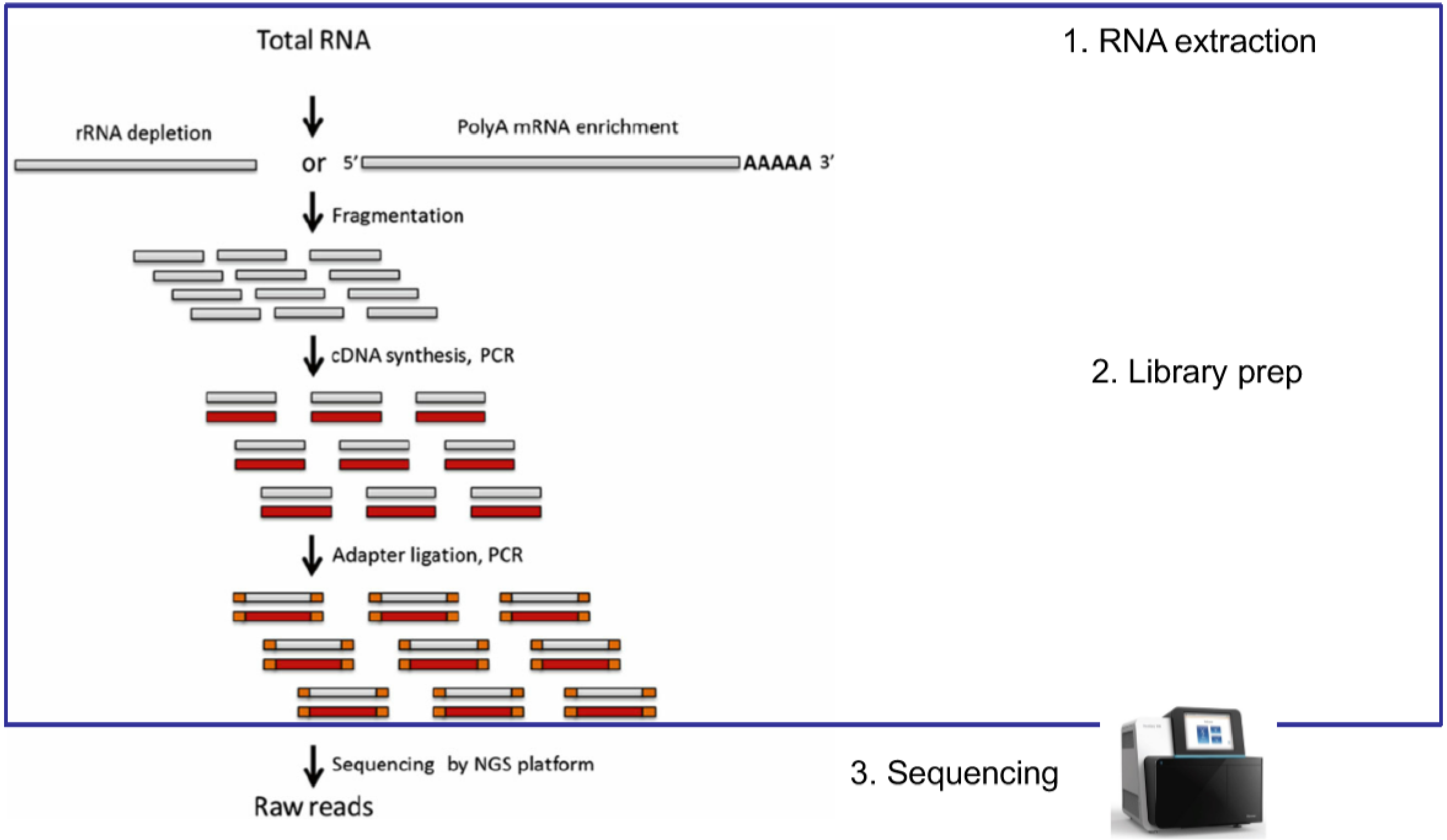
Part 2 – data processing

Data processing
- Raw Data processing
- Normalization

Statistical analysis
- Two way comparisons
- Multi-variant analysis

Functional Interpretation
- Enrichment Analysis
- Network modeling
- Regulatory pathways

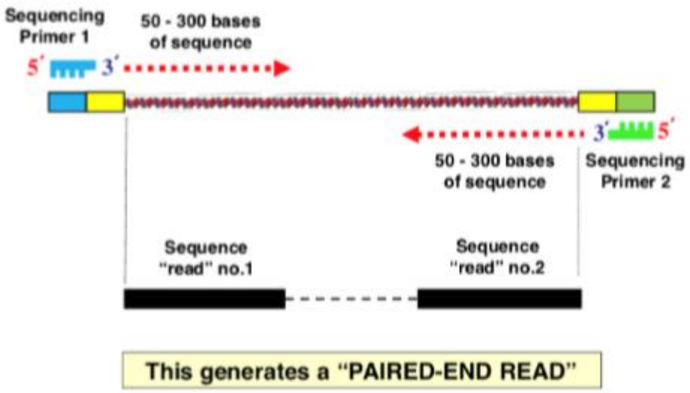
Part 1 – Sample preparation



Sequencing Technology

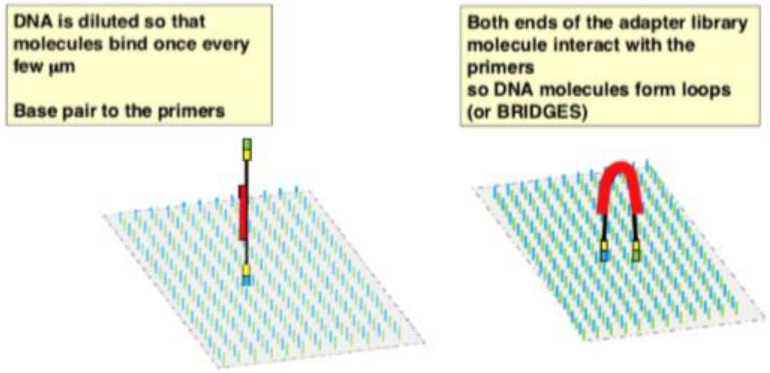
See **Prof. Hilary Rogers** lecture : Sanger & **NGS sequencing**

Illumina NGS: paired end sequencing



This generates a "PAIRED-END READ"

Illumina NGS: adaptor library binds to the primers

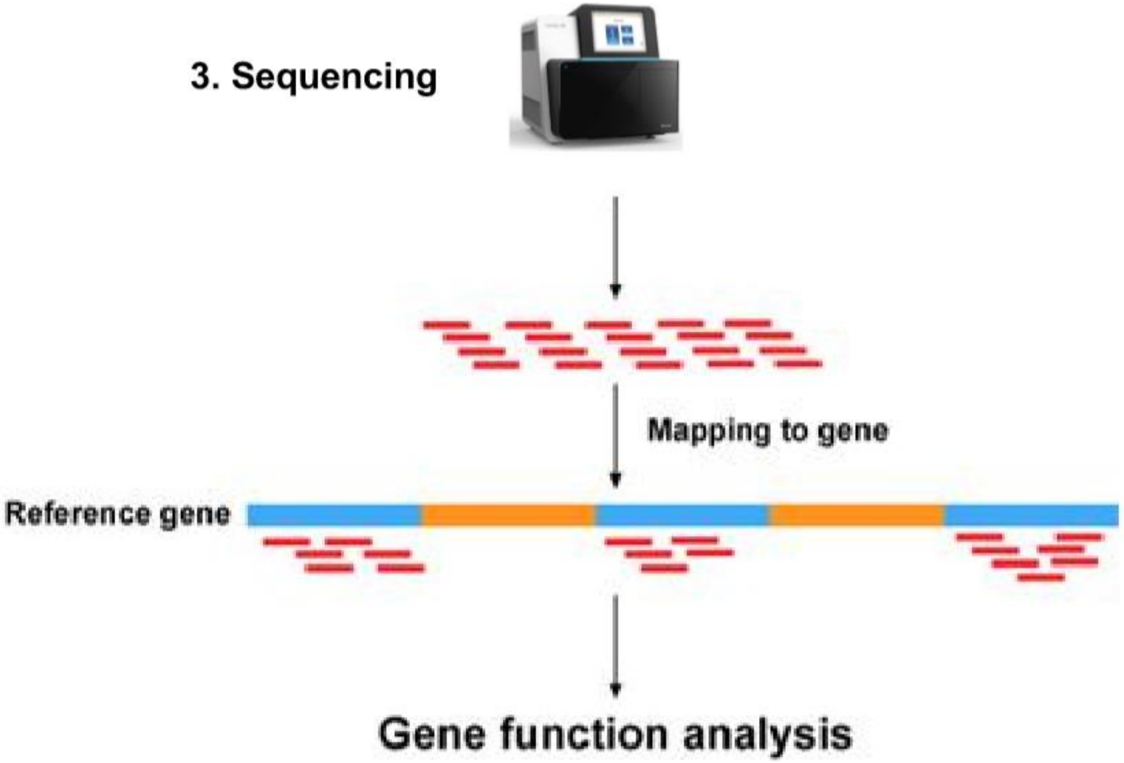


DNA is diluted so that molecules bind once every few µm

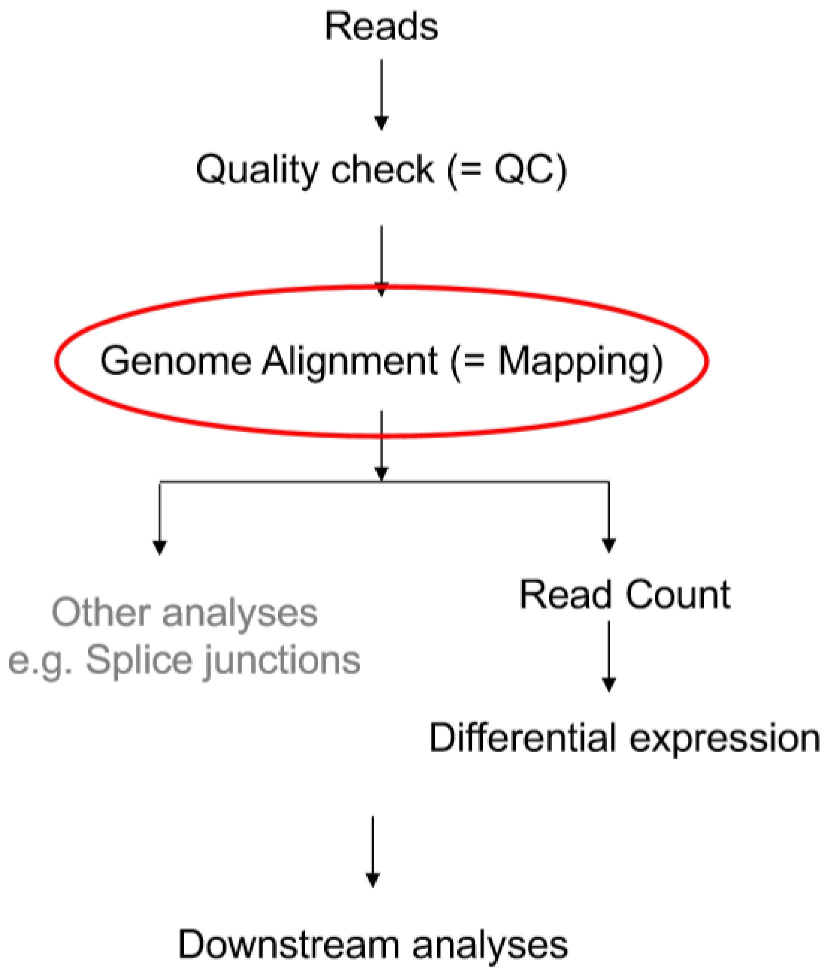
Base pair to the primers

Both ends of the adapter library molecule interact with the primers
so DNA molecules form loops (or BRIDGES)

RNAseq – data processing the concept



RNAseq – data processing the concept



Reads

Quality check (= QC)

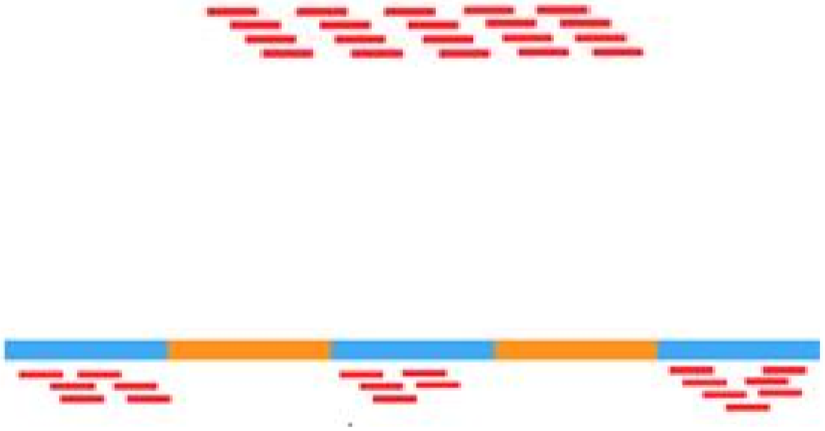
Genome Alignment (= Mapping)

Other analyses
e.g. Splice junctions

Read Count

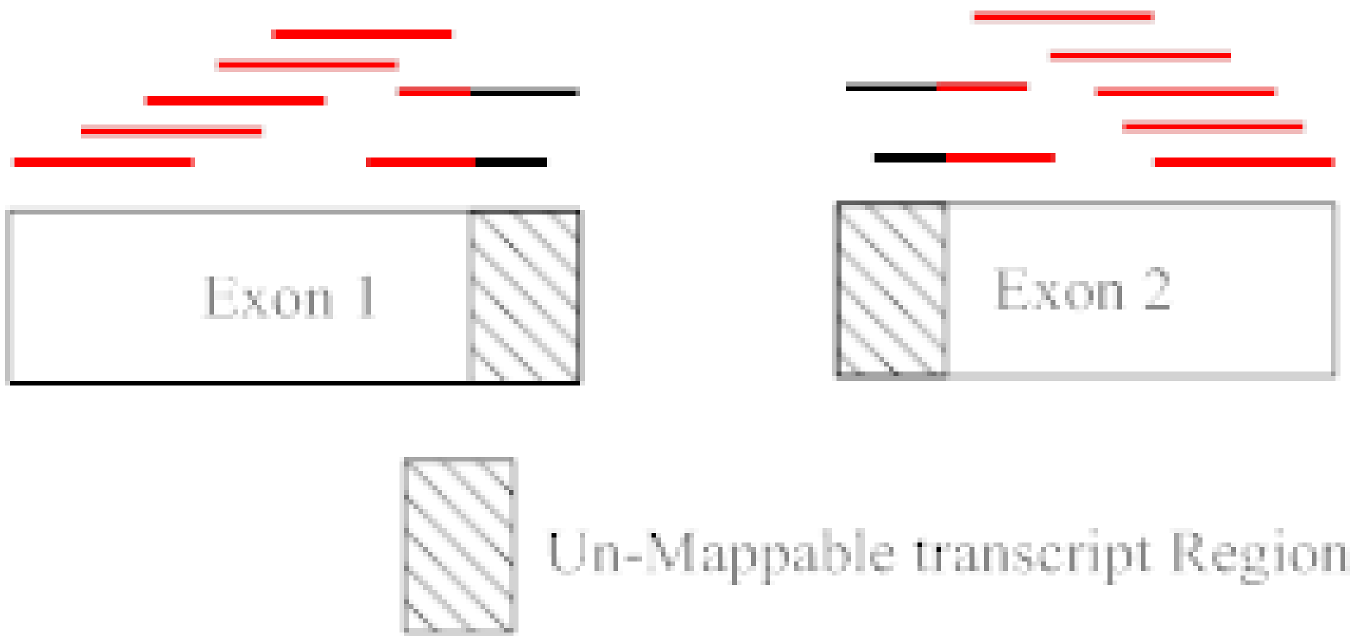
Differential expression

Downstream analyses

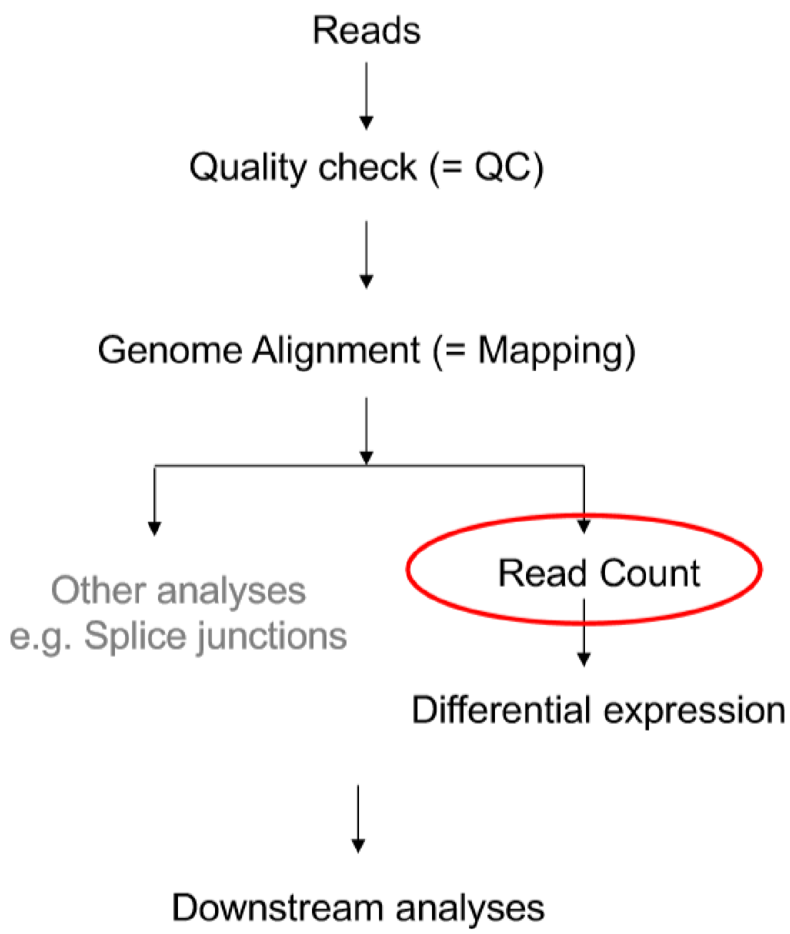


Mapping

- Map to Genome (= Gene) or Transcriptome (= Varient)



Read Count



Reads

Quality check (= QC)

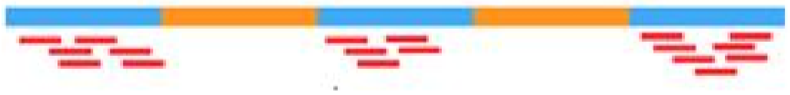
Genome Alignment (= Mapping)

Other analyses
e.g. Splice junctions

Read Count

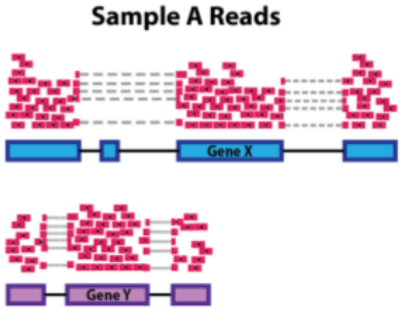
Differential expression

Downstream analyses



Calibration = Normalisation = Scaling

Aim: Counting reads ▬ = 1 read


Sample A Reads

1. **Number of mapped reads (“Sequencing depth”)**

Sample A has double number of reads than sample B

2. **Gene length**

Gene X longer than Gene Y

3. **RNA composition**

Gene DE takes up most of the reads in sample A but not in sample B

Normalization of Counting

Several common normalization methods exist to account for these differences:

FPKM, Deseq2, EdgeR

FPKM

- Length of Gene
- Number of mapped reads
  - Unique/multiple hit
  - Genome Vs cDNA
  - Paired mapping

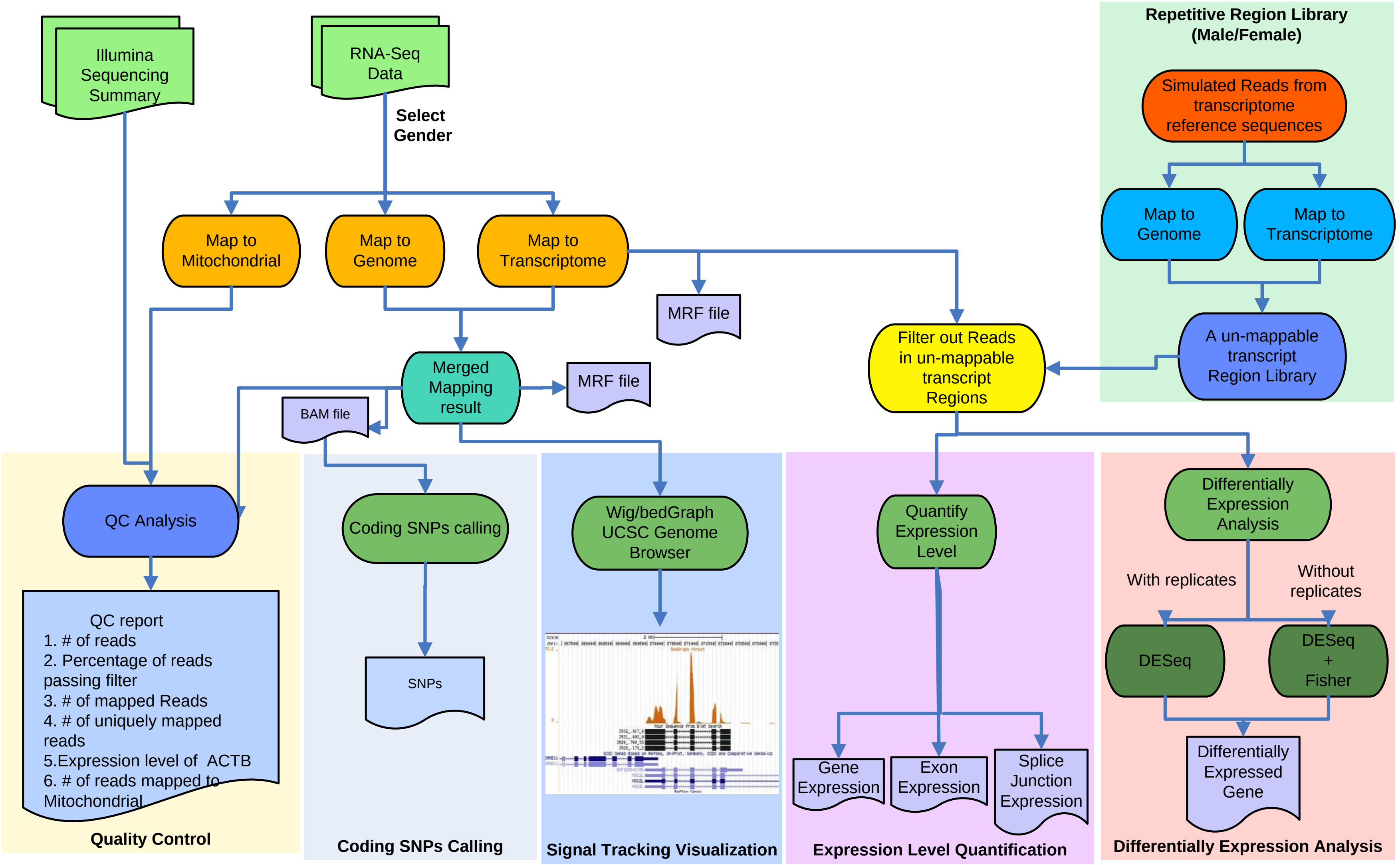
Frequency Per Kilobase Million

FPKM

What information can you acquire apart from transcript frequencies

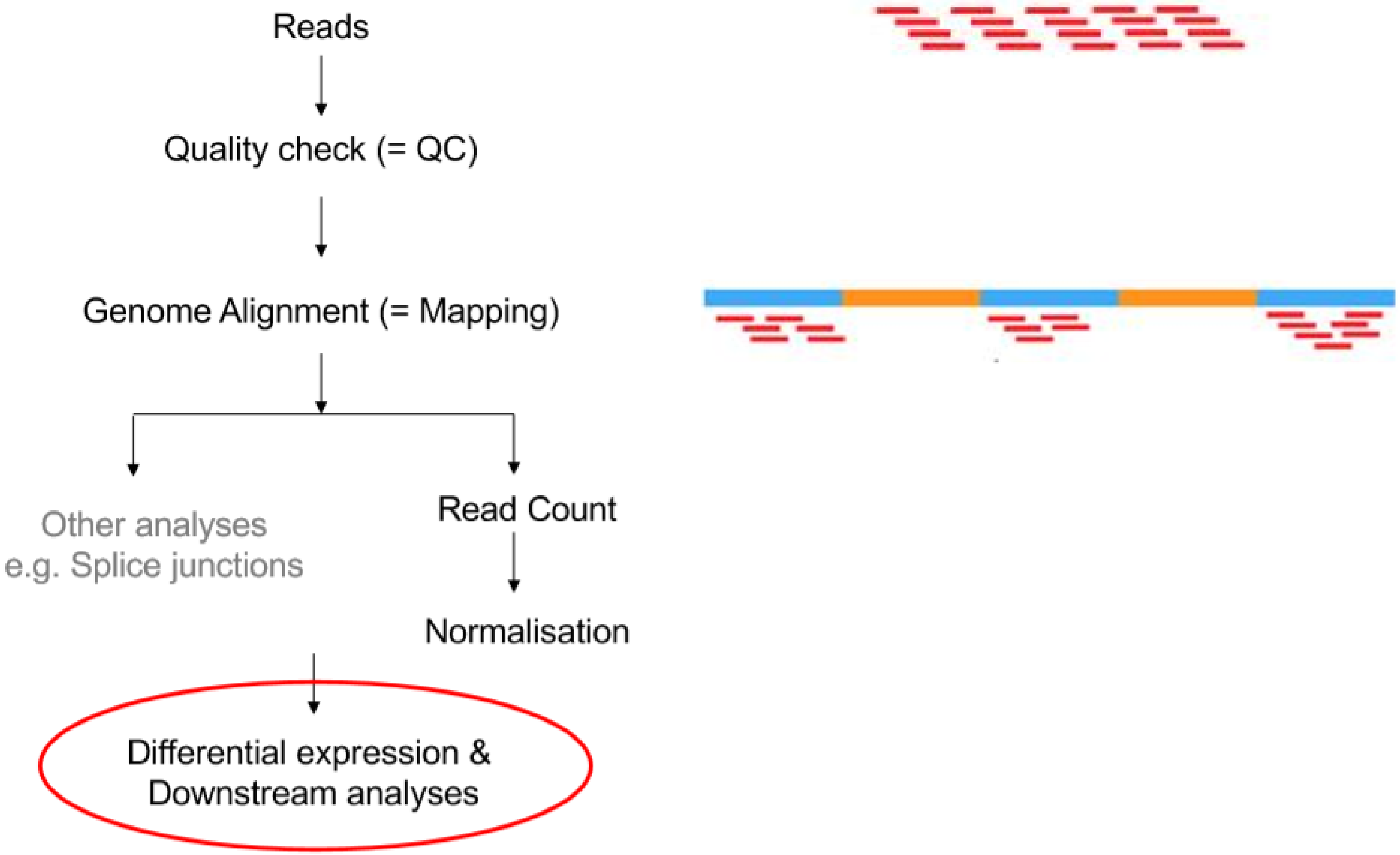
- *De novo* transcript assembly
- SNPs
- SSRs
- Novel splice varients
- miRNAs (depends on library production)

RNAseq analysis: Simple, not so much


Illumina Sequencing Summary
RNA-Seq Data
Select Gender
Repetitive Region Library (Male/Female)
Simulated Reads from transcriptome reference sequences
Map to Genome
Map to Transcriptome
A un-mappable transcript Region Library
Map to Mitochondrial
Map to Genome
Map to Transcriptome
MRF file
Filter out Reads in un-mappable transcript Regions
Merged Mapping result
MRF file
BAM file
QC Analysis
QC report
1. # of reads
2. Percentage of reads passing filter
3. # of mapped Reads
4. # of uniquely mapped reads
5.Expression level of ACTB
6. # of reads mapped to Mitochondrial
Quality Control
Coding SNPs calling
SNPs
Coding SNPs Calling
Wig/bedGraph UCSC Genome Browser
Signal Tracking Visualization
Quantify Expression Level
Gene Expression
Exon Expression
Splice Junction Expression
Expression Level Quantification
Differentially Expression Analysis
With replicates
Without replicates
DESeq
DESeq + Fisher
Differentially Expressed Gene
Differentially Expression Analysis

Wang et. al. 2011 Bioinformatics 27:2598-2600

Differentials Analysis



Reads

Quality check (= QC)

Genome Alignment (= Mapping)

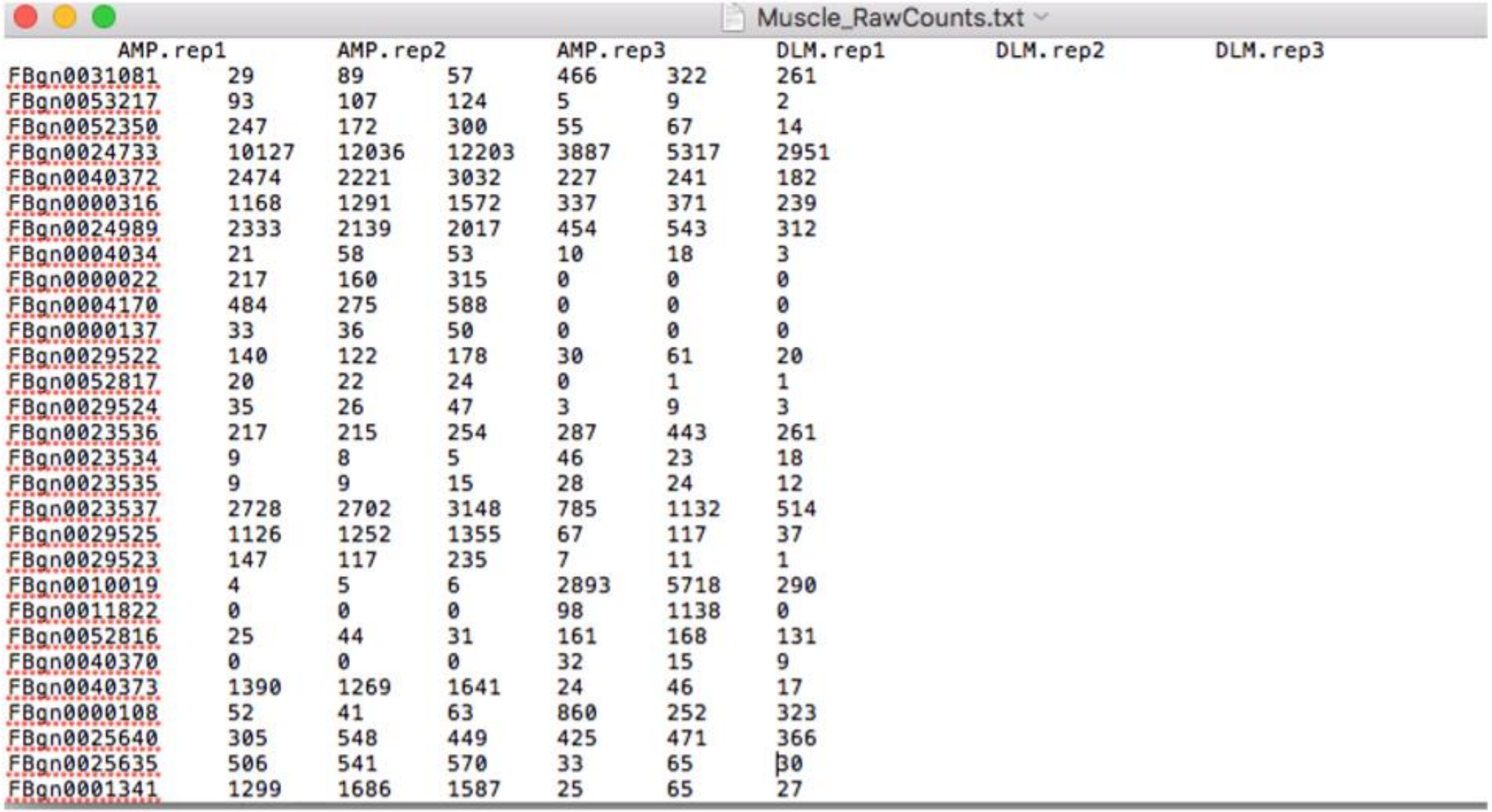
Other analyses
e.g. Splice junctions

Read Count

Normalisation

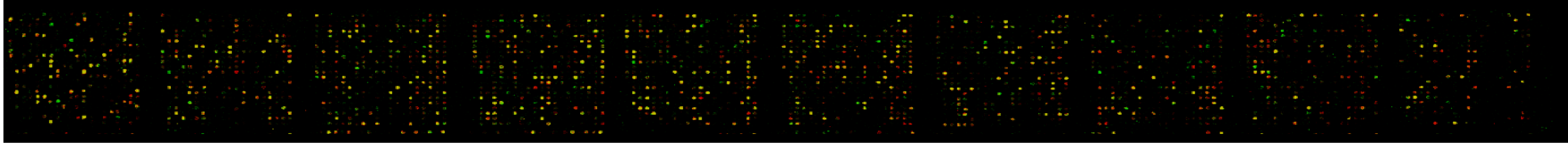
Differential expression &
Downstream analyses

Start with count file



Muscle_RawCounts.txt

```
                AMP.rep1        AMP.rep2        AMP.rep3        DLM.rep1        DLM.rep2        DLM.rep3
FBgn0031081     29      89      57      466     322     261
FBgn0053217     93      107     124     5       9       2
FBgn0052350     247     172     300     55      67      14
FBgn0024733     10127   12036   12203   3887    5317    2951
FBgn0040372     2474    2221    3032    227     241     182
FBgn0000316     1168    1291    1572    337     371     239
FBgn0024989     2333    2139    2017    454     543     312
FBgn0004034     21      58      53      10      18      3
FBgn0000022     217     160     315     0       0       0
FBgn0004170     484     275     588     0       0       0
FBgn0000137     33      36      50      0       0       0
FBgn0029522     140     122     178     30      61      20
FBgn0052817     20      22      24      0       1       1
FBgn0029524     35      26      47      3       9       3
FBgn0023536     217     215     254     287     443     261
FBgn0023534     9       8       5       46      23      18
FBgn0023535     9       9       15      28      24      12
FBgn0023537     2728    2702    3148    785     1132    514
FBgn0029525     1126    1252    1355    67      117     37
FBgn0029523     147     117     235     7       11      1
FBgn0010019     4       5       6       2893    5718    290
FBgn0011822     0       0       0       98      1138    0
FBgn0052816     25      44      31      161     168     131
FBgn0040370     0       0       0       32      15      9
FBgn0040373     1390    1269    1641    24      46      17
FBgn0000108     52      41      63      860     252     323
FBgn0025640     305     548     449     425     471     366
FBgn0025635     506     541     570     33      65      30
FBgn0001341     1299    1686    1587    25      65      27
```

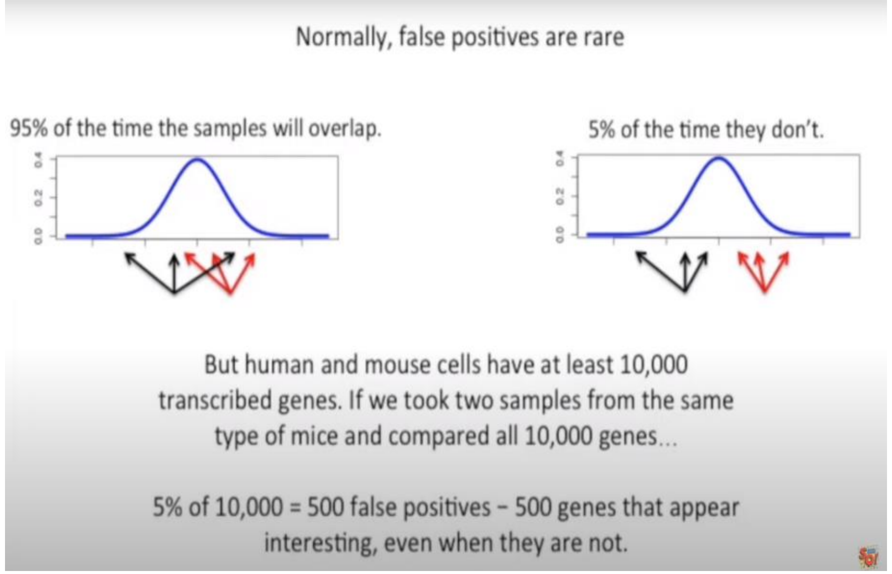


T-test to multi-variant – False discover

Statistical analysis

Statistical concepts

False Discovery Rate



Normally, false positives are rare

95% of the time the samples will overlap.

5% of the time they don't.

But human and mouse cells have at least 10,000 transcribed genes. If we took two samples from the same type of mice and compared all 10,000 genes…

5% of 10,000 = 500 false positives – 500 genes that appear interesting, even when they are not.

[StatQuest - FDR](https://www.youtube.com/watch?v=K8LQSvtjcEo)

https://www.youtube.com/watch?v=K8LQSvtjcEo

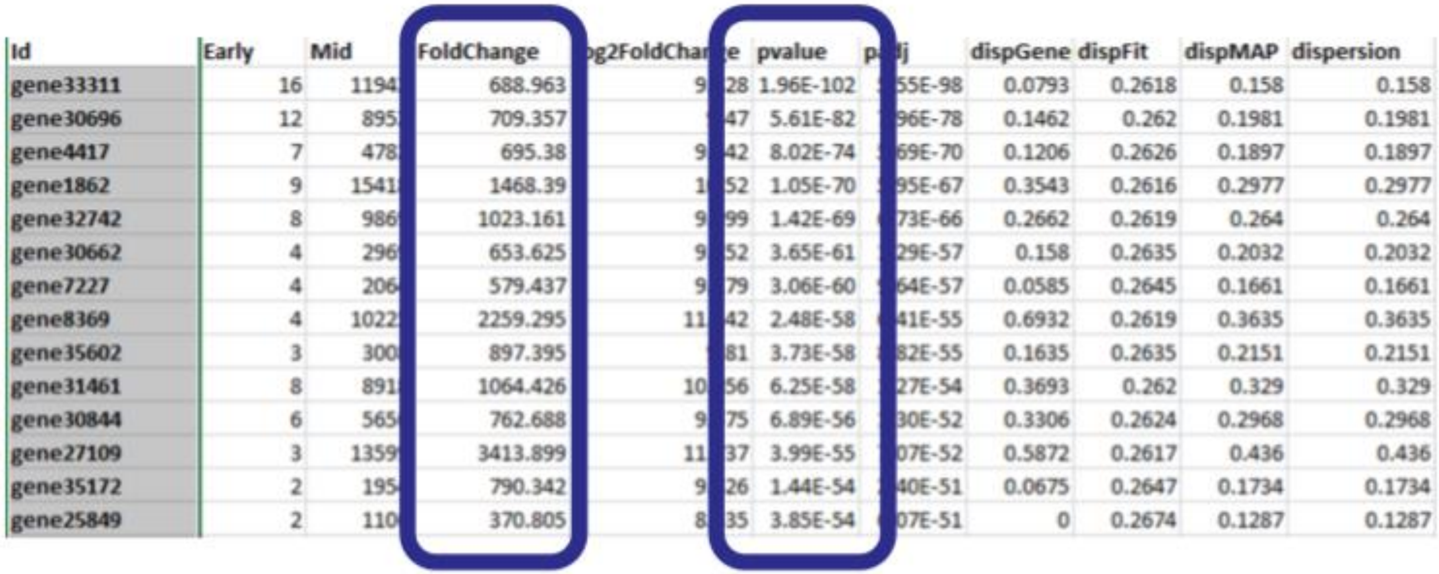
Enrichment analysis (GSEA)

[StatQuest - **Fisher's Exact Test**](https://statquest.org/statquickie-fishers-exact-test-and-enrichment-analysis/)

https://statquest.org/statquickie-fishers-exact-test-and-enrichment-analysis/

P-Values and FDR (Adjusted P-value)

Differentially expressed genes: THE table



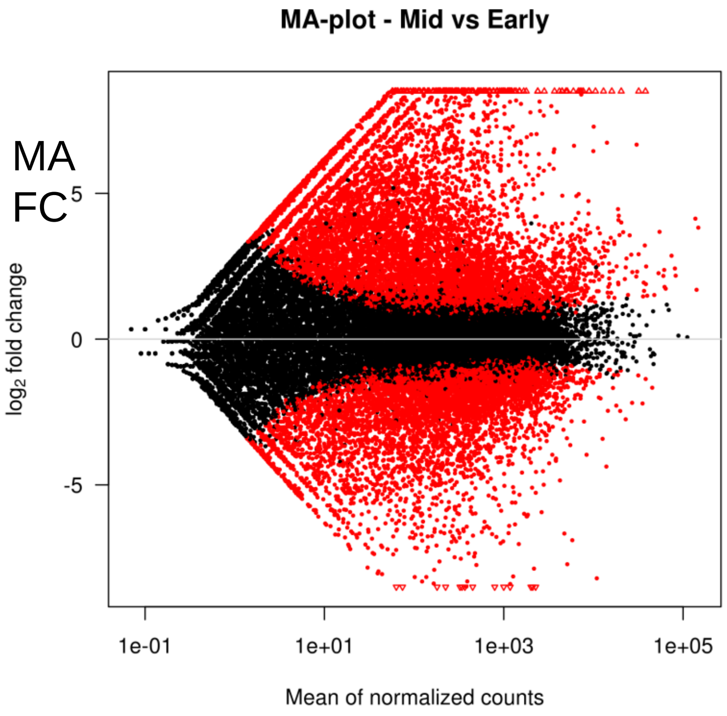
Id	Early	Mid	FoldChange	og2FoldChange	pvalue	p dj	dispGene	dispFit	dispMAP	dispersion
gene33311	16	1194	688.963	9 28	1.96E-102	55E-98	0.0793	0.2618	0.158	0.158
gene30696	12	895	709.357	47	5.61E-82	96E-78	0.1462	0.262	0.1981	0.1981
gene4417	7	478	695.38	9 42	8.02E-74	69E-70	0.1206	0.2626	0.1897	0.1897
gene1862	9	1541	1468.39	1 52	1.05E-70	95E-67	0.3543	0.2616	0.2977	0.2977
gene32742	8	986	1023.161	9 99	1.42E-69	73E-66	0.2662	0.2619	0.264	0.264
gene30662	4	296	653.625	9 52	3.65E-61	29E-57	0.158	0.2635	0.2032	0.2032
gene7227	4	206	579.437	9 79	3.06E-60	64E-57	0.0585	0.2645	0.1661	0.1661
gene8369	4	1022	2259.295	11 42	2.48E-58	41E-55	0.6932	0.2619	0.3635	0.3635
gene35602	3	300	897.395	81	3.73E-58	82E-55	0.1635	0.2635	0.2151	0.2151
gene31461	8	891	1064.426	10 56	6.25E-58	27E-54	0.3693	0.262	0.329	0.329
gene30844	6	565	762.688	9 75	6.89E-56	30E-52	0.3306	0.2624	0.2968	0.2968
gene27109	3	1359	3413.899	11 37	3.99E-55	07E-52	0.5872	0.2617	0.436	0.436
gene35172	2	195	790.342	9 26	1.44E-54	40E-51	0.0675	0.2647	0.1734	0.1734
gene25849	2	110	370.805	8 35	3.85E-54	07E-51	0	0.2674	0.1287	0.1287

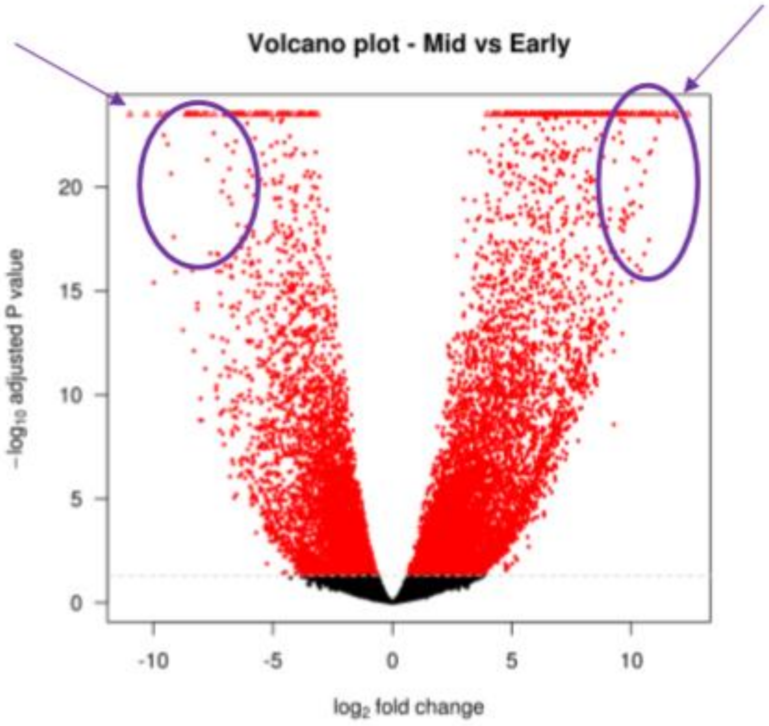
- P-value <0.05 commonly used but…
- **False Discovery Rate (FDR)** vital

  (because: many variables (genes) but small sample set (few replicates))

MA and Volvcano plots: binary analysis

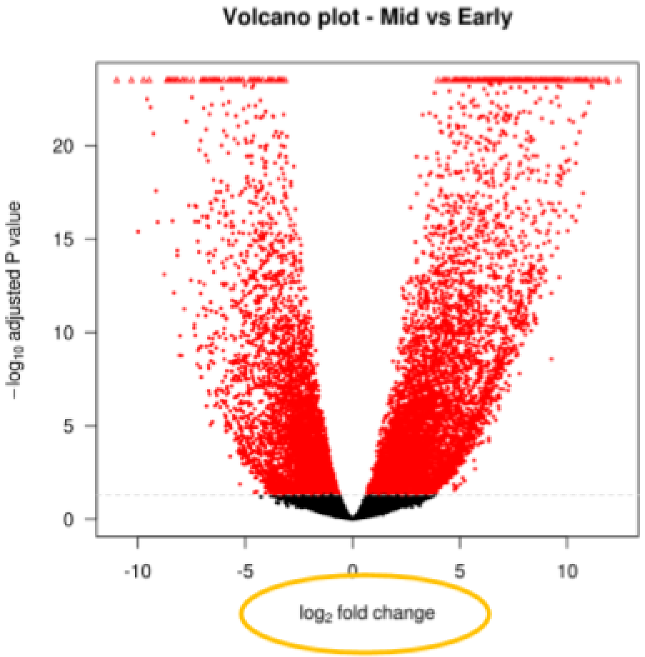
Log MA
Log FC

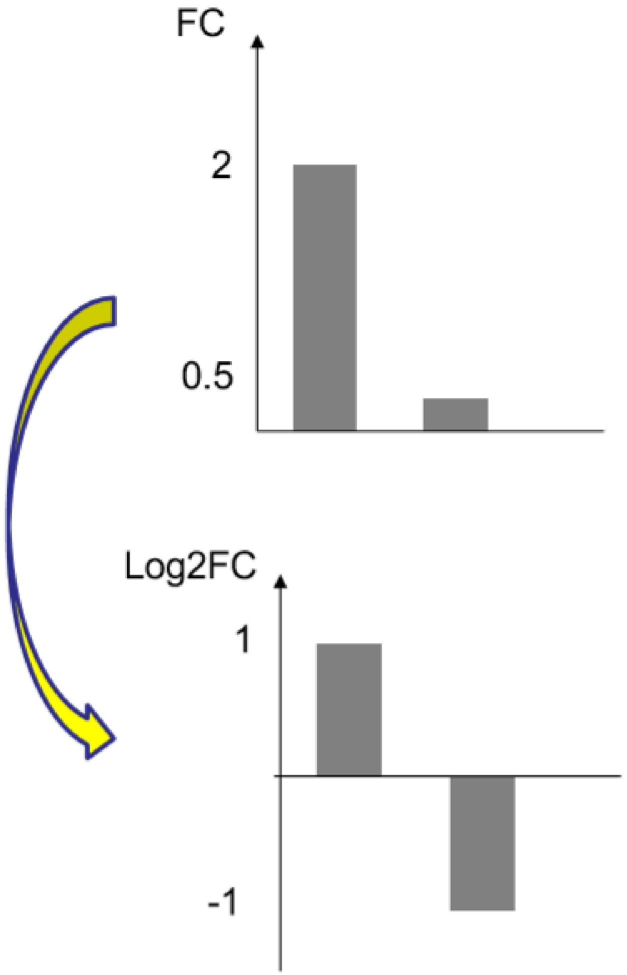




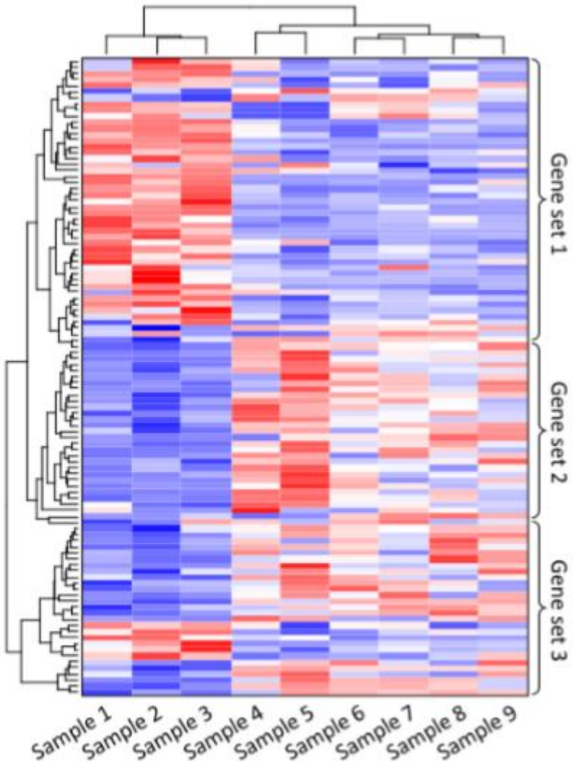
Why use log?

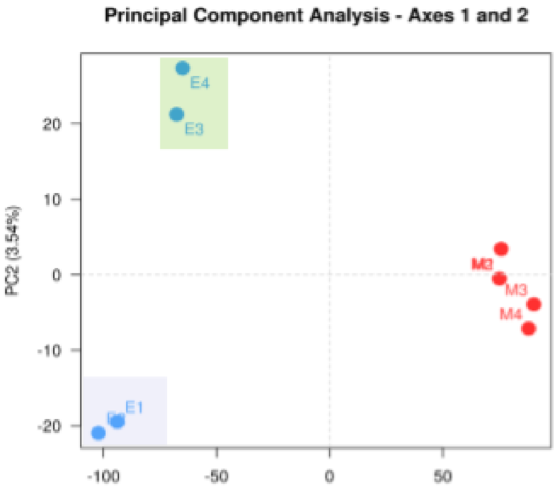
Why using **log**FC?

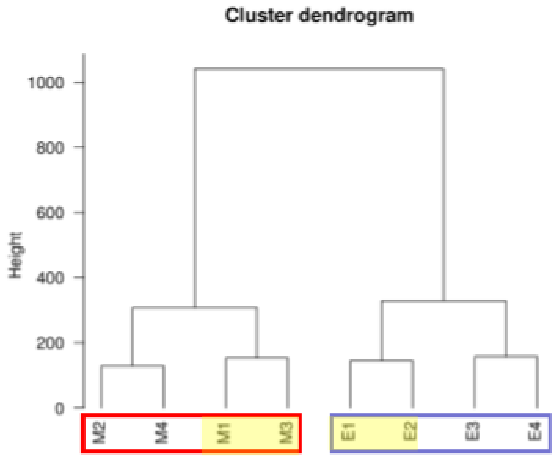




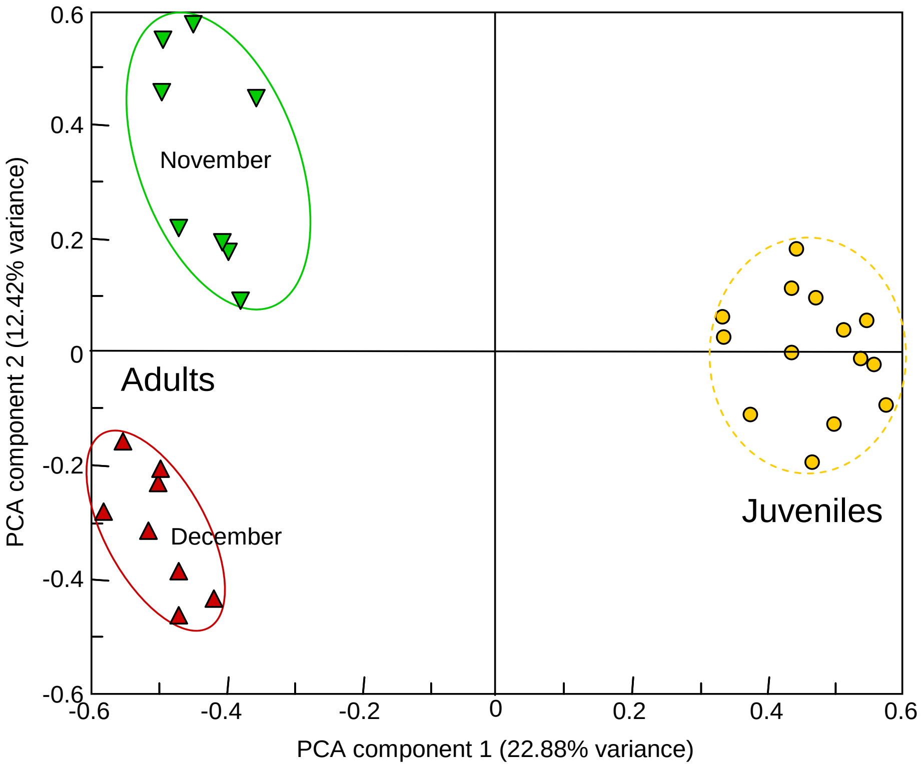
Hierarchical clustering, PCA and Dendrograms

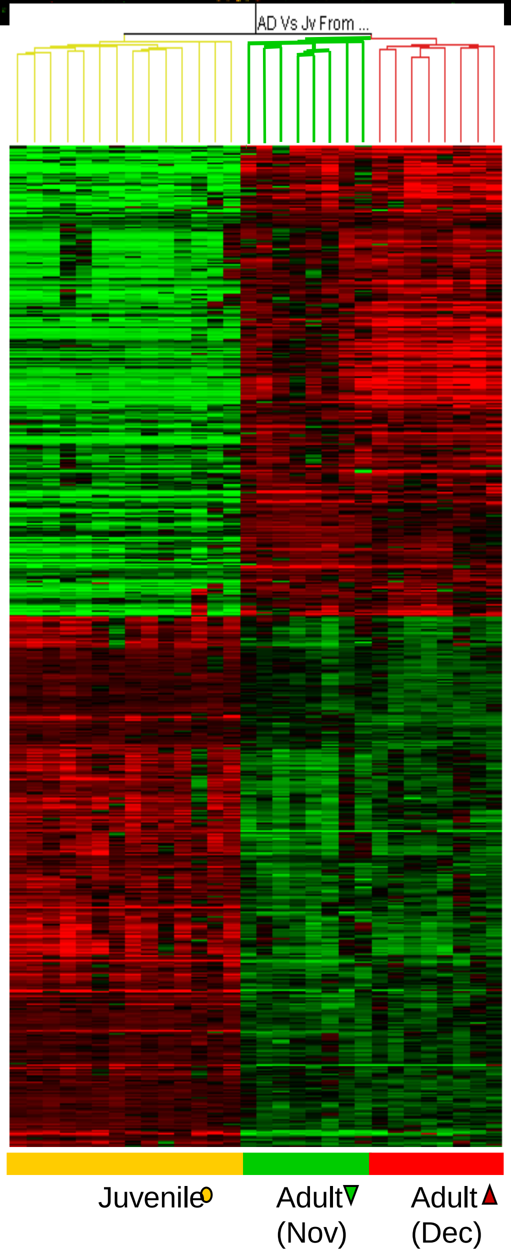




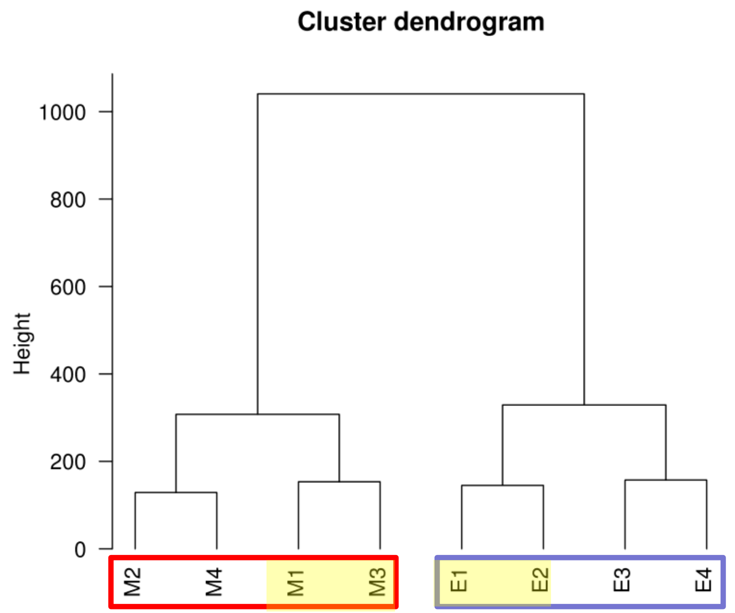


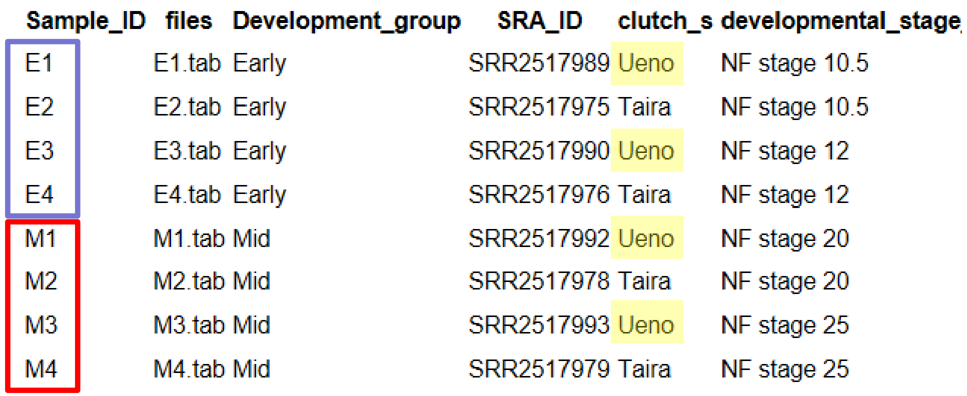
Example of PCA and Hierarchical clustering





Dendograms

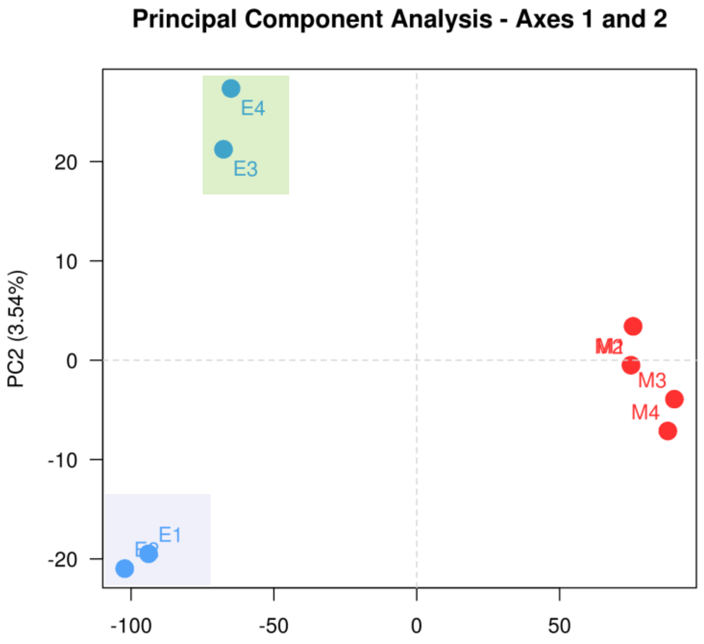


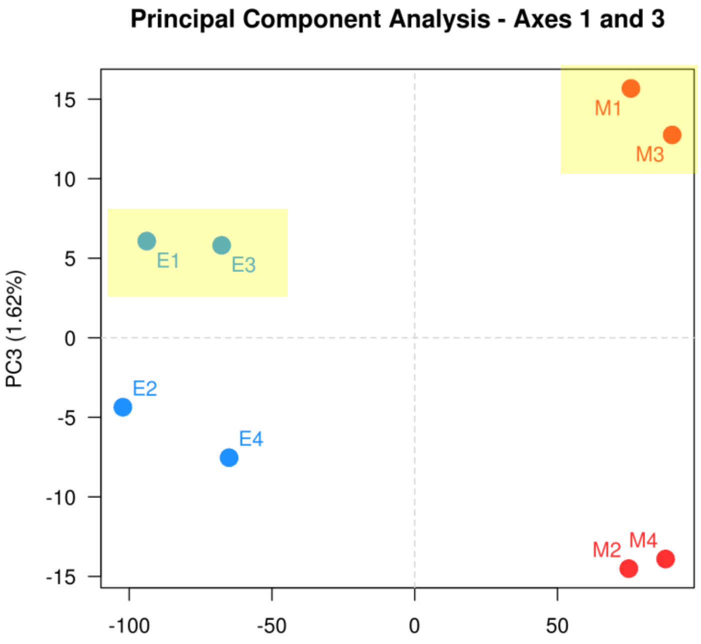


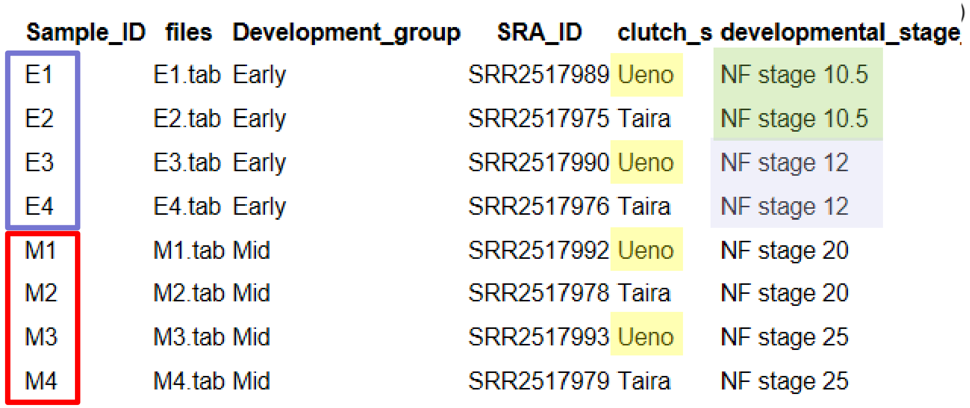
Sample_ID	files	Development_group	SRA_ID	clutch_s	developmental_stage
E1	E1.tab	Early	SRR2517989	Ueno	NF stage 10.5
E2	E2.tab	Early	SRR2517975	Taira	NF stage 10.5
E3	E3.tab	Early	SRR2517990	Ueno	NF stage 12
E4	E4.tab	Early	SRR2517976	Taira	NF stage 12
M1	M1.tab	Mid	SRR2517992	Ueno	NF stage 20
M2	M2.tab	Mid	SRR2517978	Taira	NF stage 20
M3	M3.tab	Mid	SRR2517993	Ueno	NF stage 25
M4	M4.tab	Mid	SRR2517979	Taira	NF stage 25

Table 1: Data files and associated biological conditions.

PCA

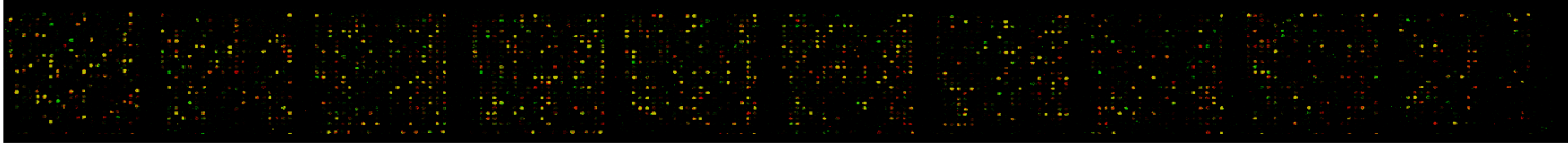






Sample_ID	files	Development_group	SRA_ID	clutch_s	developmental_stage
E1	E1.tab	Early	SRR2517989	Ueno	NF stage 10.5
E2	E2.tab	Early	SRR2517975	Taira	NF stage 10.5
E3	E3.tab	Early	SRR2517990	Ueno	NF stage 12
E4	E4.tab	Early	SRR2517976	Taira	NF stage 12
M1	M1.tab	Mid	SRR2517992	Ueno	NF stage 20
M2	M2.tab	Mid	SRR2517978	Taira	NF stage 20
M3	M3.tab	Mid	SRR2517993	Ueno	NF stage 25
M4	M4.tab	Mid	SRR2517979	Taira	NF stage 25

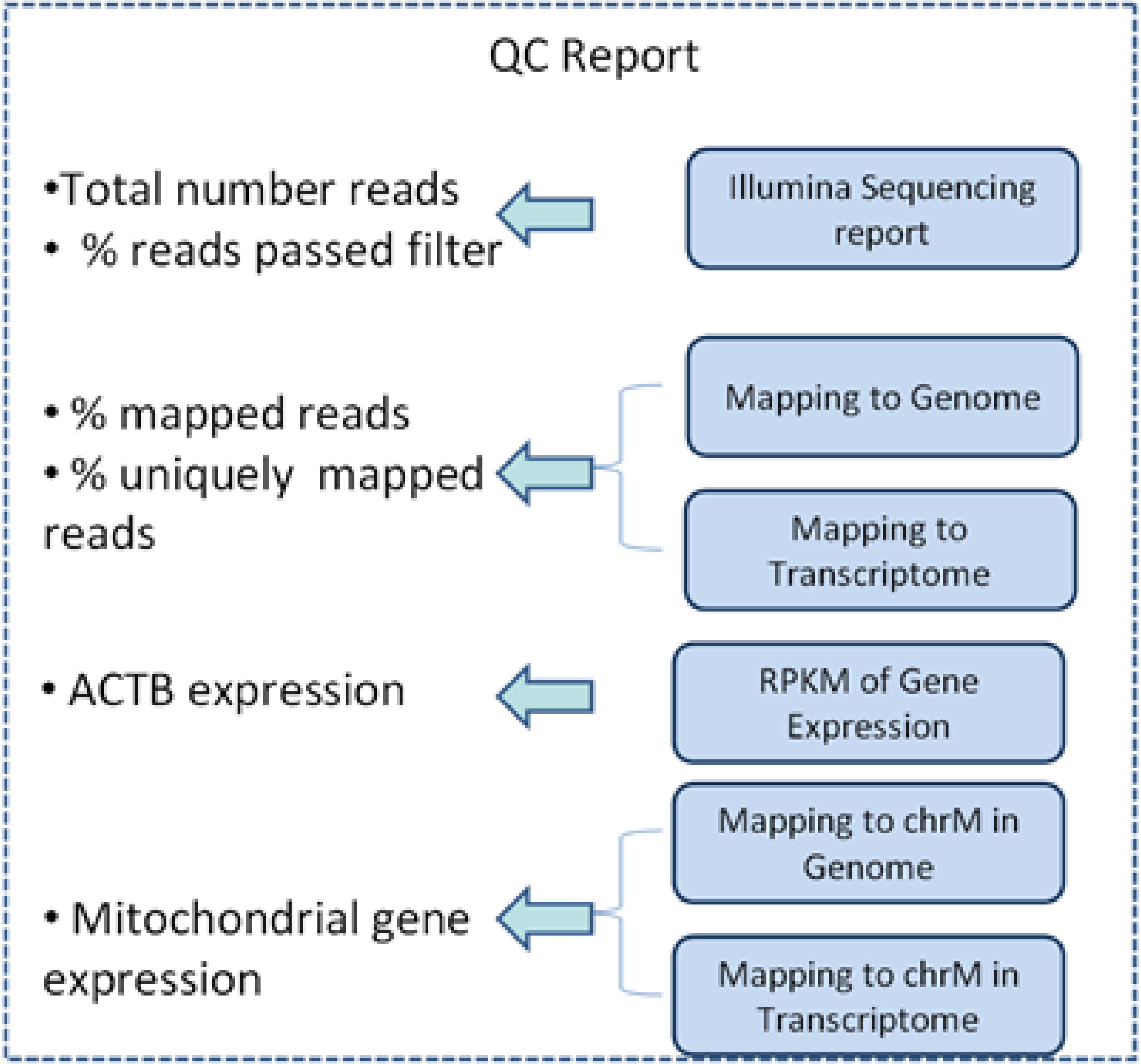
Table 1: Data files and associated biological conditions.



Enrichment to functional networks

Functional Interpretation

Quality Control



QC Report

- Total number reads
- % reads passed filter

Illumina Sequencing report

- % mapped reads
- % uniquely mapped reads

Mapping to Genome

Mapping to Transcriptome

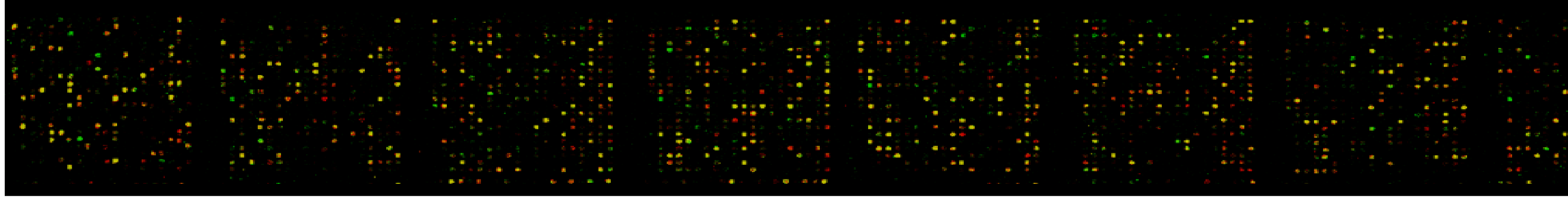
- ACTB expression

RPKM of Gene Expression

- Mitochondrial gene expression

Mapping to chrM in Genome

Mapping to chrM in Transcriptome



Pattern analysis

Analysis of ‘groups’ of genes identifying collective/coordinated changes associated with underlying biological process.

- Gene Set Enrichment Analysis
- Gene Ontology analysis

Uses Prior Knowledge (databases, pathways, ect)

- “Guilt by association” => if unknown gene ***i*** is similar in expression to known gene ***j,*** maybe they are involved in the same/related pathway
- Dimensionality reduction: datasets are too big to be able to get information out without reorganizing the data

Functional data analysis: Gene annotation enrichment analysis

- Enrichment
- The occurrence of a specific gene annotation within a subset of genes occurs at a statistically higher frequency than would be expected if the parental population was sampled by random.

Think about selected red and blue balls randomly taken from a bag.

The proportion of red and blue balls in any sub-set should represented the ratio of the original mix if sampling is **random** and all balls are **identical**.

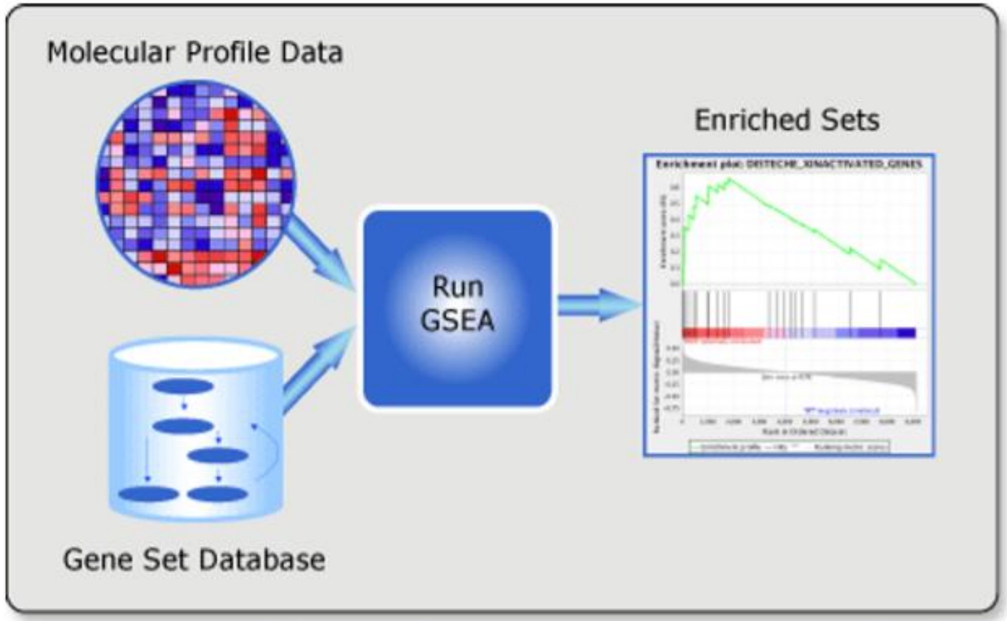
The statistical **probability** of any deviation from the original ratio can be mathematically calculated.

The greater the **deviation from the expected ratio** the more probable that the sampling is not random i.e. The sampling technique is selecting or enriching.

Gene Set Enrichment Analysis

Idea: Overlap gene sets

Rank genes by a criteria (e.g. FC or expression level)
Walk down your list and compute overlap with known gene sets (database)



Gene Ontology

Idea: **Gene ontology** ~ definition of gene sets by:

- Biological process (BP)

e.g. DNA replication

- Molecular Function (MF)

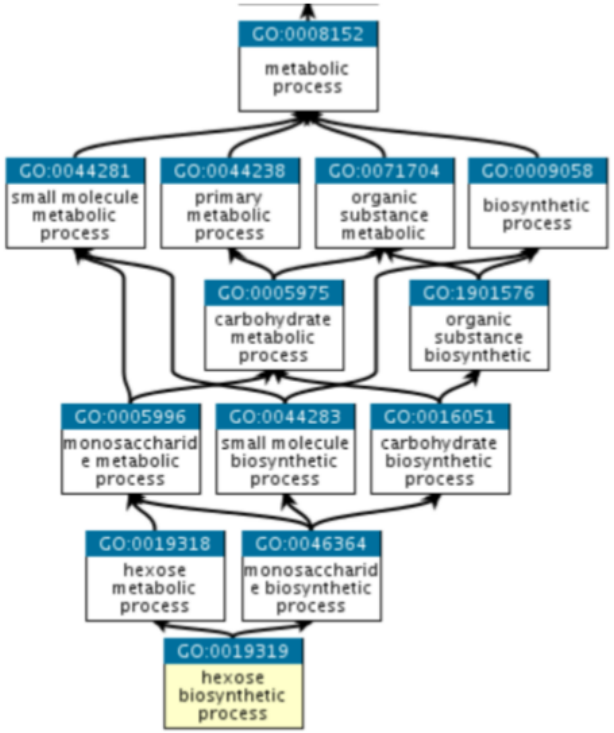
e.g. enzyme, DNA-binding

- Cellular Component (CC)

e.g. cell membrane

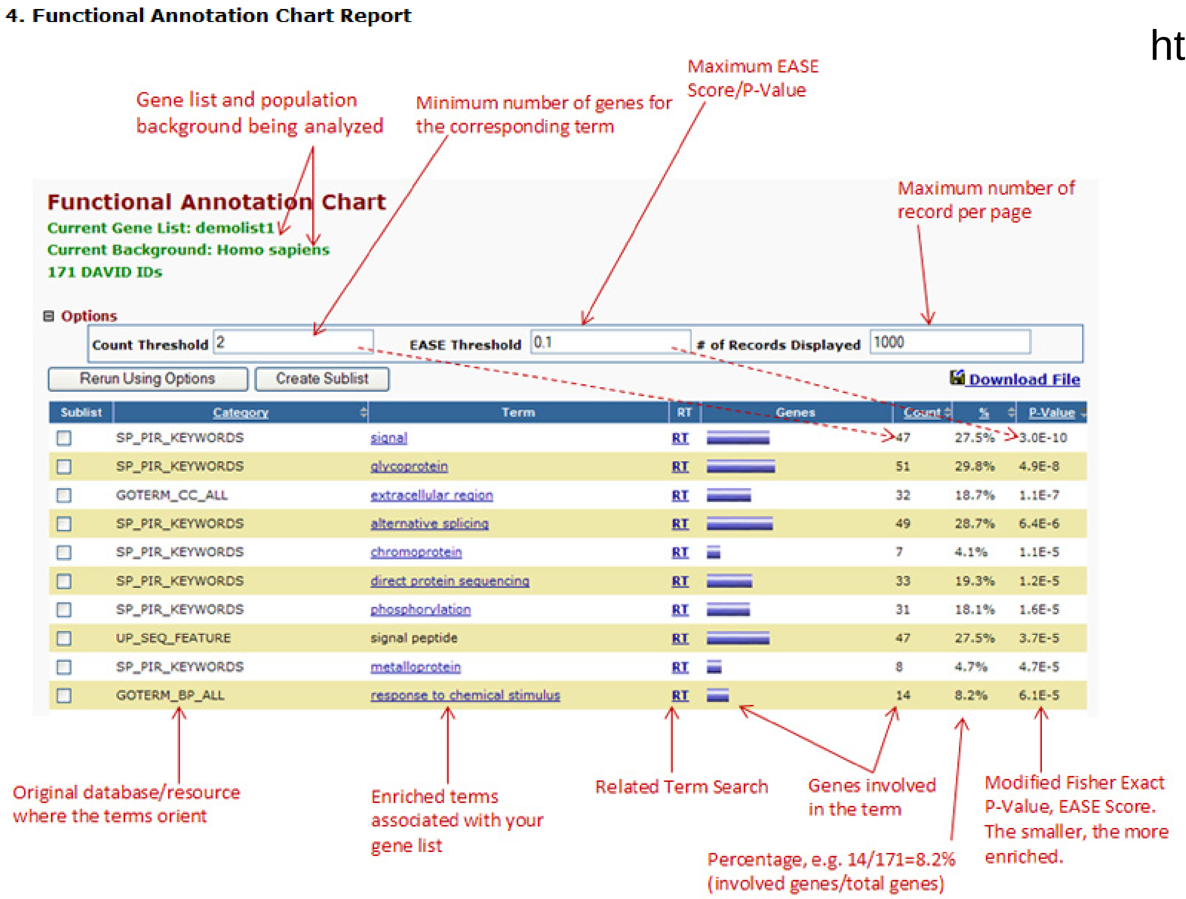
Different level

Work across organisms > powerful


GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

Gene ontology and representational analysis

http://david.abcc.ncifcrf.gov/



4. Functional Annotation Chart Report

Functional Annotation Chart
Current Gene List: demolist1
Current Background: Homo sapiens
171 DAVID IDs

⊟ **Options**

Count Threshold 2 EASE Threshold 0.1 # of Records Displayed 1000

Sublist	Category	Term	RT	Genes	Count	%	P-Value
□	SP_PIR_KEYWORDS	signal	RT		47	27.5%	3.0E-10
□	SP_PIR_KEYWORDS	glycoprotein	RT		51	29.8%	4.9E-8
□	GOTERM_CC_ALL	extracellular region	RT		32	18.7%	1.1E-7
□	SP_PIR_KEYWORDS	alternative splicing	RT		49	28.7%	6.4E-6
□	SP_PIR_KEYWORDS	chromoprotein	RT		7	4.1%	1.1E-5
□	SP_PIR_KEYWORDS	direct protein sequencing	RT		33	19.3%	1.2E-5
□	SP_PIR_KEYWORDS	phosphorylation	RT		31	18.1%	1.6E-5
□	UP_SEQ_FEATURE	signal peptide	RT		47	27.5%	3.7E-5
□	SP_PIR_KEYWORDS	metalloprotein	RT		8	4.7%	4.7E-5
□	GOTERM_BP_ALL	response to chemical stimulus	RT		14	8.2%	6.1E-5

Rerun Using Options
Create Sublist
Download File
Gene list and population background being analyzed
Minimum number of genes for the corresponding term
Maximum EASE Score/P-Value
Maximum number of record per page
Original database/resource where the terms orient
Enriched terms associated with your gene list
Related Term Search
Genes involved in the term
Percentage, e.g. 14/171=8.2% (involved genes/total genes)
Modified Fisher Exact P-Value, EASE Score. The smaller, the more enriched.

Functional Annotation Chart:
Chart Report is an annotation-term-focused view which lists annotation terms and their associated genes under study. To avoid over counting duplicated genes, the Fisher Exact statistics is calculated based on corresponding DAVID gene IDs by which all redundence in original IDs are removed. All result of Chart Report has to pass the thresholds (by default, Max.Prob.<=0.1 and Min.Count>=2) in Chart Option section to ensure only statistically significant ones displayed.

EASE Score Threshold (Maximum Probability):
The threshold of EASE Score, a modified Fisher Exact P-Value, for gene-enrichment analysis. It ranges from 0 to 1. Fisher Exact P-Value = 0 represnts perfect enrichment. Usually P-Value is equal or smaller than 0.05 to be considered strongly enriched in the annotation categories. Default is 0.1. More details.

Count Threshold (Minimum Count):
The threshold of minimum gene counts belonging to an annotation term. It has to be equal or greater than 0. Default is 2. In short, you do not trust the term only having one gene involved.

RT (Related Term Search):
Related Term Search can identify other similar terms. More details.

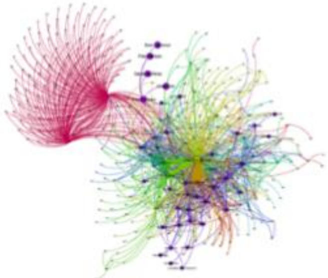
Live demonstration – for live session

Meta Analysis of published data - Geo

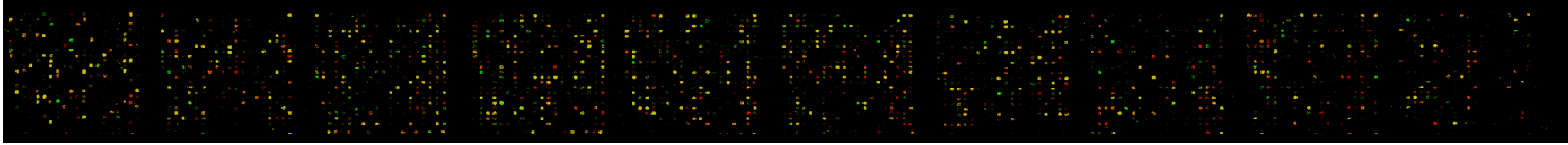
Enrichment analysis – gprofiler

Semantic similarity - Revigo

Network analysis- String



- Pdf instructions accompanying presentation



Transcripts vs transcription complex (promoters/enhances)

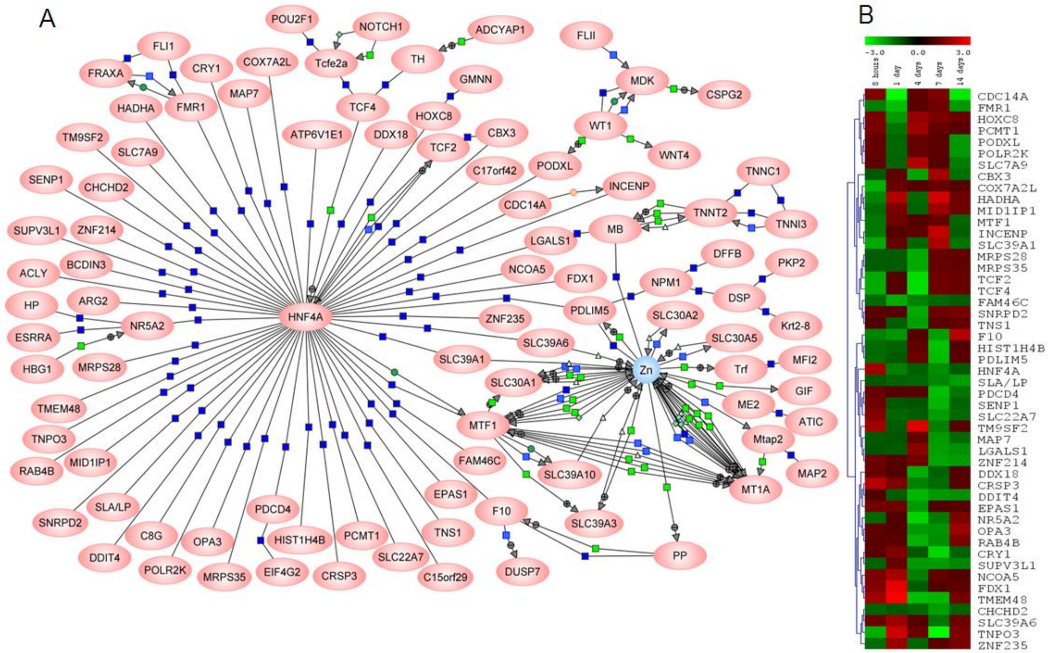
Sequence Based Analysis

Sequence based data analysis

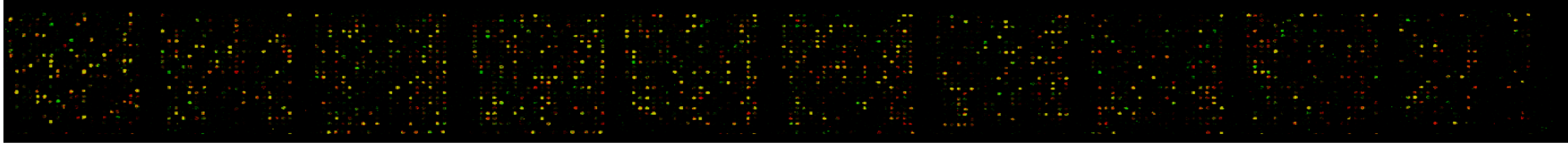
- Special arrangement bias
  - Genes with common regulation occur in specific areas of the genome i.e. In “operons” or areas under same “locus control”.
- Promoter Analysis
  - Gene with common regulation share a common Transcription Factor Binding Site(s) (TFBS) within their promoters.

Network Analysis

- Determine evidence based relationship between members of a specific transcriptome…..proteomic and metabolomic data can also be integrated.



http://www.genmapp.org/ http://www.cytoscape.org/



Only as good as you experimental descriptors

Meta data and data repositories

Data management and MIAME

1. Transcriptomic Meta Data

2. MIAME (Data Standards,

(www.mged.org/Workgroups/MIAME/miame.html)

3. MGED Ontology (Vocabulary,

www.mged.org/)

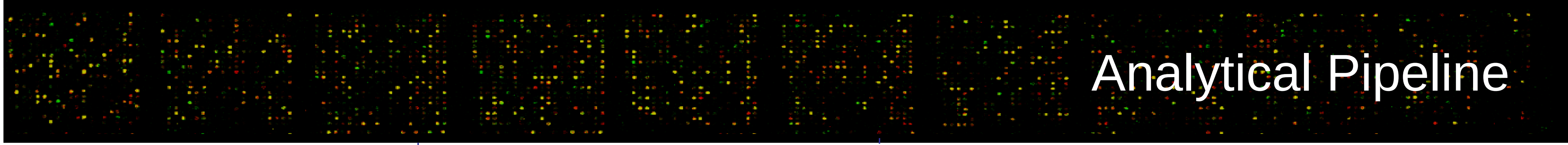
4. MAGE (standard representation,

www.mged.org/Workgroups/MAGE/mage.html)

Data repositories

- Non-specific data warehouses
  - GEO (http://www.ncbi.nlm.nih.gov/gds/)
  - Array express (http://www.ebi.ac.uk/microarray-as/ae/)

...and many other specific sites



Analytical Pipeline

Process	Approach	Software
Differential Data Analysis	T-test Anova Time course analysis	R / GEO
Functional Interpretation	Enrichment Analysis (EASE)	David gprofiler
Systems Level thinking	Semantics Or Network Analysis	Revigo Cytoscape STRING

Take Home

RNAseq & microarray:

- RNA content = transcript analysis
- Microarray = a subset of targets
- RNAseq = ALL => careful about experiment design
- RNA > counts > normalised counts > stats > analyses

Read a paper– what do you want to know

Critical thinking: does it make sense to you?

○ Experience design : replicates? Controls? Confounding factors?

○ Normalization – all samples together, what methods, how

○ Statistics? Significant – how?

Analyses:

What is your question?

○ How to subsample

○ What to test (and how to test it)