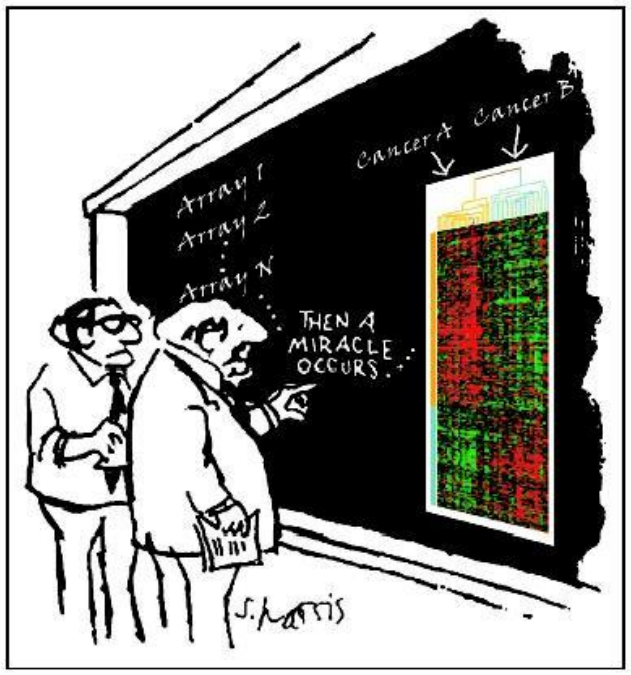
Maastricht University Leading in Learning!

Introduction practical 3

using public omics data and online databases

MBS 1002



"I think you should be more explicit here in step two."

Part 1: Omics data repositories and data processing

Omics data

- Nowadays, we can perform many high-throughput measurements in molecular biology, called 'omics'

- This can be:
  - Genetic variations: genomics
  - Gene expression: transcriptomics
  - Protein abundance: proteomics
  - Metabolite abundance: metabolomics
  - Epigenetic modifications: epigenomics
  - ...

Reuse of omics data

- Omics data are often hypothesis generating
- They contain more information than has been used for the original research or paper
- They may be reused to answer other research questions or to be explored in a different way
- They may be integrated with newly generated data or compared to that

Repositories of publicly available omics data

→ Sharing data is important (and even often obligatory)
- To be able to validate original results and conclusions
- But also to reuse data for other studies

- To support easy of use, repositories commonly make use of standardised data formats

- Also proper annotation (metadata) is required, otherwise the data cannot be understood or reused
  - Metadata should be detailed enough
  - Provide information on the samples, also on possible covariates
  - Using standardised formats as much as possible
  - Using standardised terminology as much as possible to make it easier to find and compare studies

Raw and processed data

- Omics data has to be processed after generation, in order to be used for statistical and biological evaluation and interpretation (as will be discussed shortly)

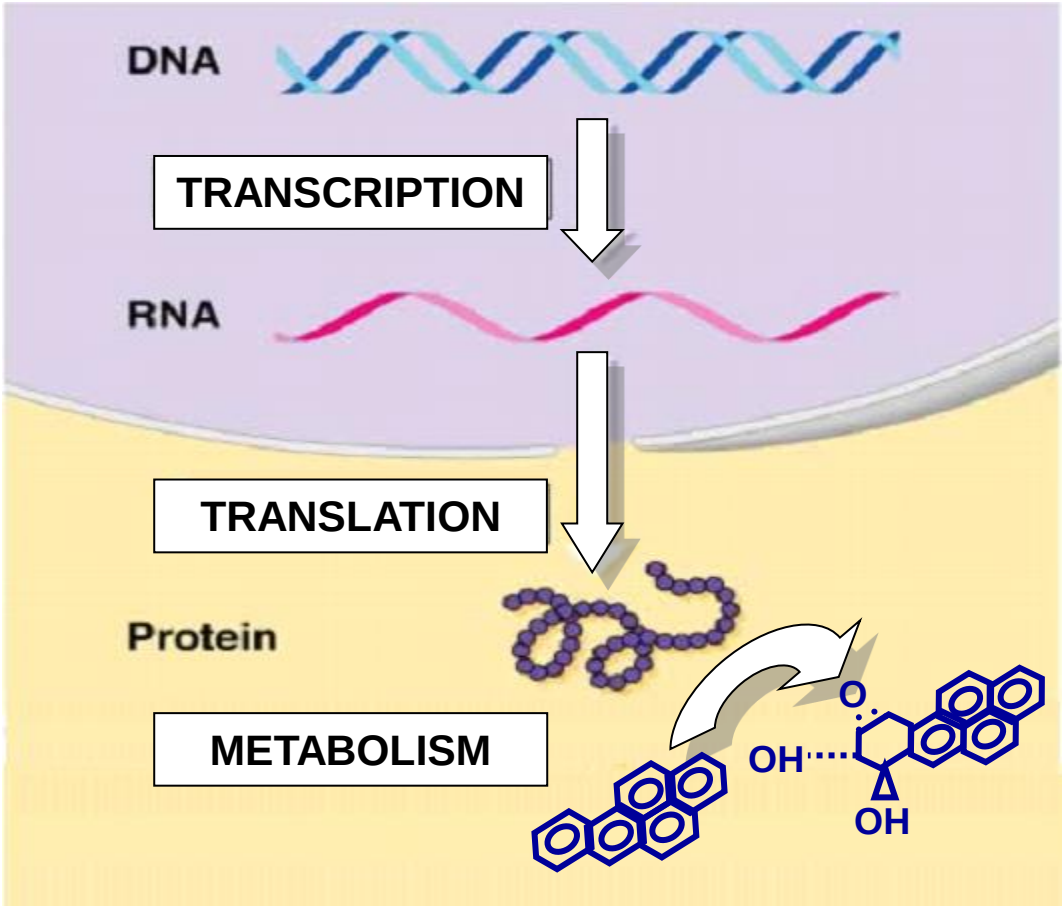
- Data may be made available at different levels:
  - Raw or unprocessed data per sample
  - Processed or 'normalised' data per sample
  - Statistically analysed data, *e.g.* Comparisons between experimental groups

- Repositories generally contain raw and/or processed data
  - The statistically analysed data is often provided with the paper (as this depends on the research questions asked)

General concepts of processing of omics data: transcriptomics data

- One of the most commonly used methods is transcriptomics
- This related to the fact that it is easier to measure mRNA expression than abundance of proteins

The central dogma and -omics technologies



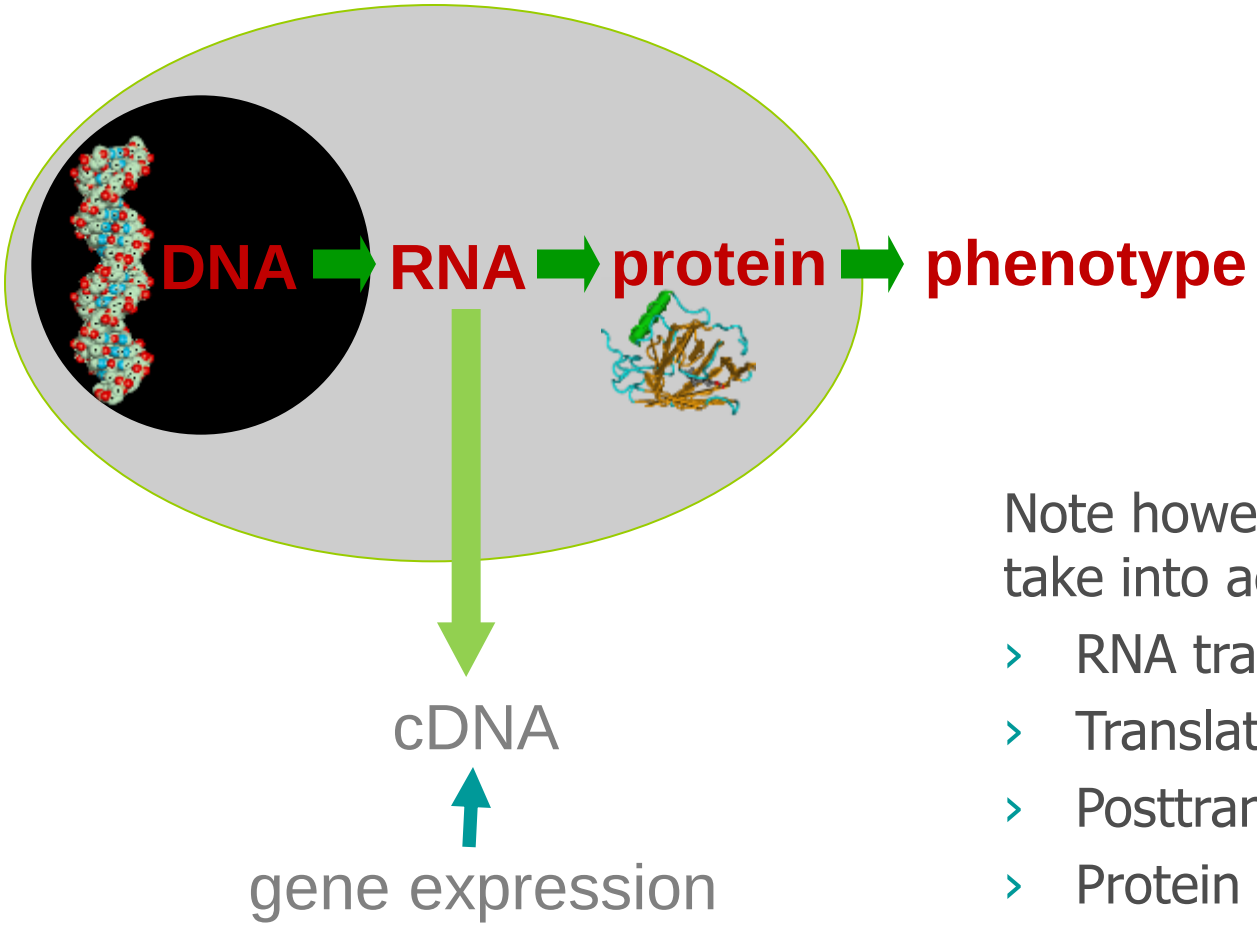
Genome
Genomics

Transcriptome
Transcriptomics

Proteome
Proteomics

Metabolome
Metabolomics

Gene expression is an indirect measure of effect



Note however, that this does not take into account:

- RNA transport
- Translation to protein
- Posttranslational modification
- Protein localisation
- Protein degradation

Technologies for transcriptomics

- The two most commonly used technologies for high-throughput gene expression measurement are microarrays and RNA-sequencing

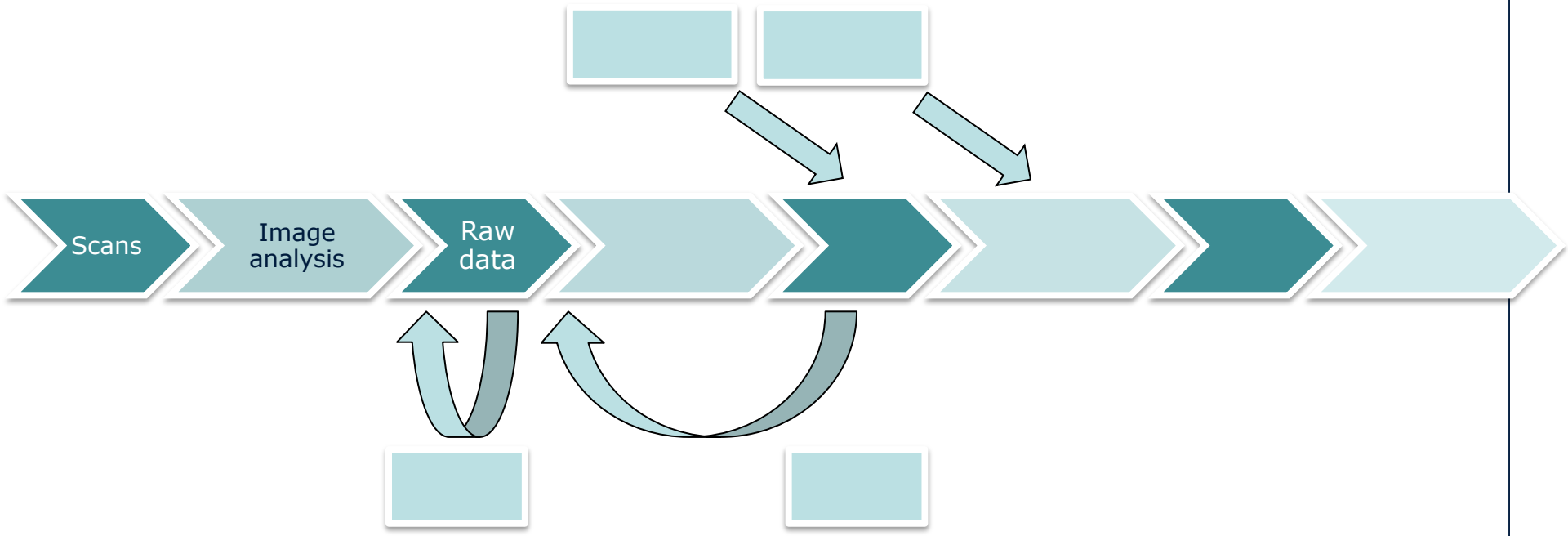
- Microarrays contain predesigned probes to detect a large number of gene transcripts (mostly tens of thousands)
  - Can only measure genes for which the probes (short complementary sequences) have been designed

- RNA-sequencing sequences all mRNAs present and quantifies the number of molecules detected (read counts)
  - Can measure everything present

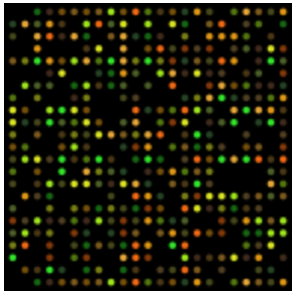
Repositories for transcriptomics data

- The most known general repositories for transcriptomics data are ArrayExpress (Europe) and Gene Expression Omnibus (GEO, VS)

Data processing workflow for microarrays



scan



scanner software

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ORF	t0 green	t0 green bk	t0 red	t0 red bkg	t0.5 green	t0.5 green	t0.5 red	t0.5 red bk	t2 green	t2 green bk	t2 red	t2 red bk
2	YHR007C	3570	1132	3643	692	3858	1213	5102	1052	2477	1351	3850	795
3	YOL109W	7534	1159	12218	622	7016	1386	5418	576	6119	1470	8272	872
4	YAL056W	1441	996	1043	569	2873	1062	2465	384	1984	1361	1537	858
5	YAL058W	2145	1168	1740	631	2623	1291	1768	670	2122	1535	1486	926
6	YAL059W	1894	1109	1578	575	2145	1052	801	442	1784	1385	1069	789
7	YAL060W	7927	1143	8770	694	9361	1484	5820	772	6740	1586	4029	978
8	YAL061W	5208	1171	5664	756	5914	1108	6008	494	3492	1376	3517	759
9	YAL062W	8258	1224	9527	664	5637	1836	22504	2094	4015	1474	21303	873
10	YAR002W	2374	1308	1838	752	3632	1156	2451	511	2675	1168	1881	643
11	YAR003W	2131	1230	1397	636	2668	1368	2265	580	1848	1184	1652	632
12	YAR007C	2183	1373	1553	794	3170	1179	6450	508	2191	1209	5920	650
13	YAR008W	1702	1214	964	603	2106	1397	1160	590	1635	1250	1743	662
14	YAR009C	4848	1356	4079	748	6508	1277	5457	493	4770	1191	3480	619
15	YAR010C	10550	1361	9306	748	11736	1503	10471	687	9254	1363	7756	742
16	YAL001C	1530	1118	1018	607	2221	1151	1233	421	1818	1407	1171	798
17	YAL002W	2302	1104	1881	614	2705	1493	2307	746	2102	1460	1603	892
18	YAL003W	6897	1160	7621	705	12021	1244	3263	479	6281	1450	2750	762
19	YAL004W	10306	1187	13176	718	12818	1568	8520	804	13036	1506	7086	811
20	YAL005C	9570	1305	13796	857	11039	1308	8848	594	9246	1470	4087	855
21	YAL007C	3041	1142	2768	665	4013	1530	2306	800	2629	1404	2471	834
22	YAL008W	3649	1274	3850	706	5321	1200	3721	557	5284	1675	5655	899
23	YAL009W	2067	1179	1572	634	4709	1406	3768	718	2600	1445	2019	826
24	YAL010C	2596	1144	2396	724	2807	1229	2026	756	2203	1498	1226	808
25	YAL011W	3971	1166	3777	668	5128	1360	3203	670	3017	1373	2448	778
26	YAL012W	3394	1239	2964	712	2653	1108	4221	611	3068	1430	1695	773
27	YAL013W	2812	1032	2763	568	2766	1320	2216	644	2085	1370	1347	808
28	YAL014C	2500	1324	1954	728	3683	1314	3212	536	2610	1121	1941	578
29	YAL015C	3010	1374	2236	753	3838	1120	2546	409	2646	1238	1570	644
30	YAL016W	4777	1260	4243	667	6863	1147	5379	449	5054	1183	2807	580
31	YAL017W	2534	1362	1828	735	3102	1214	1933	460	2659	1318	1758	706

Lots more!

foreground intensity

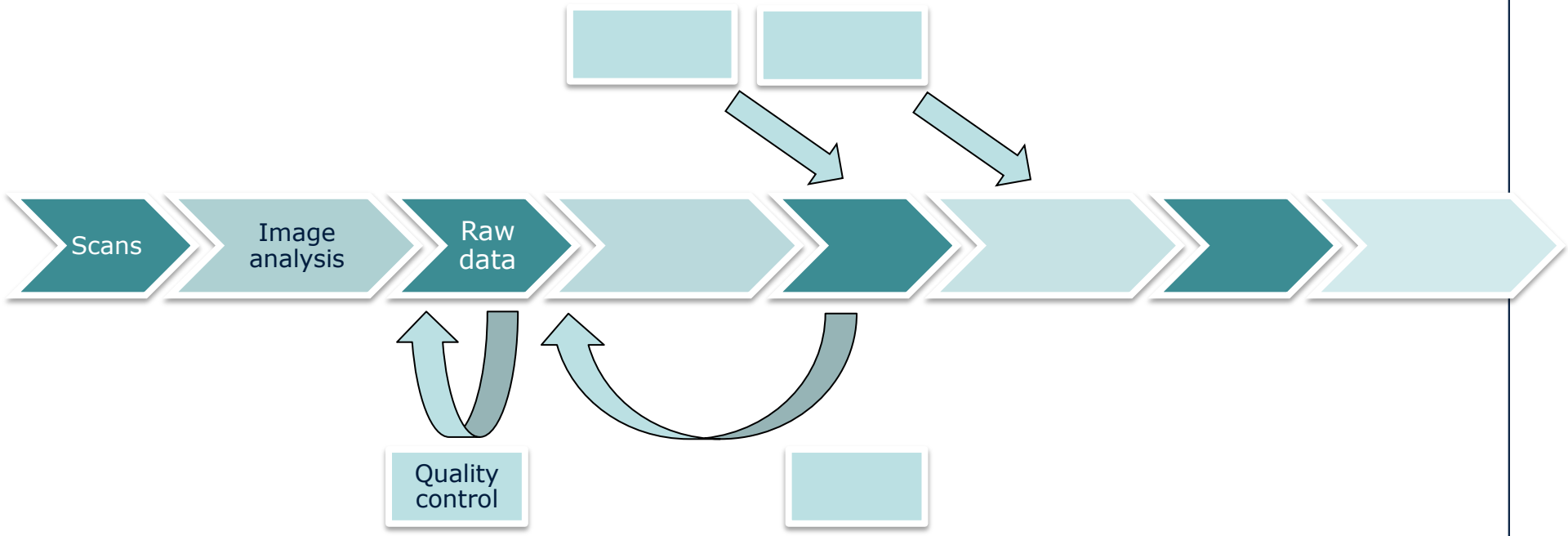
background intensity

?

computer-aided quality control (QC)

Images: arrayit.com; agilent.com; wikimedia.org

Generic data processing workflow for microarrays

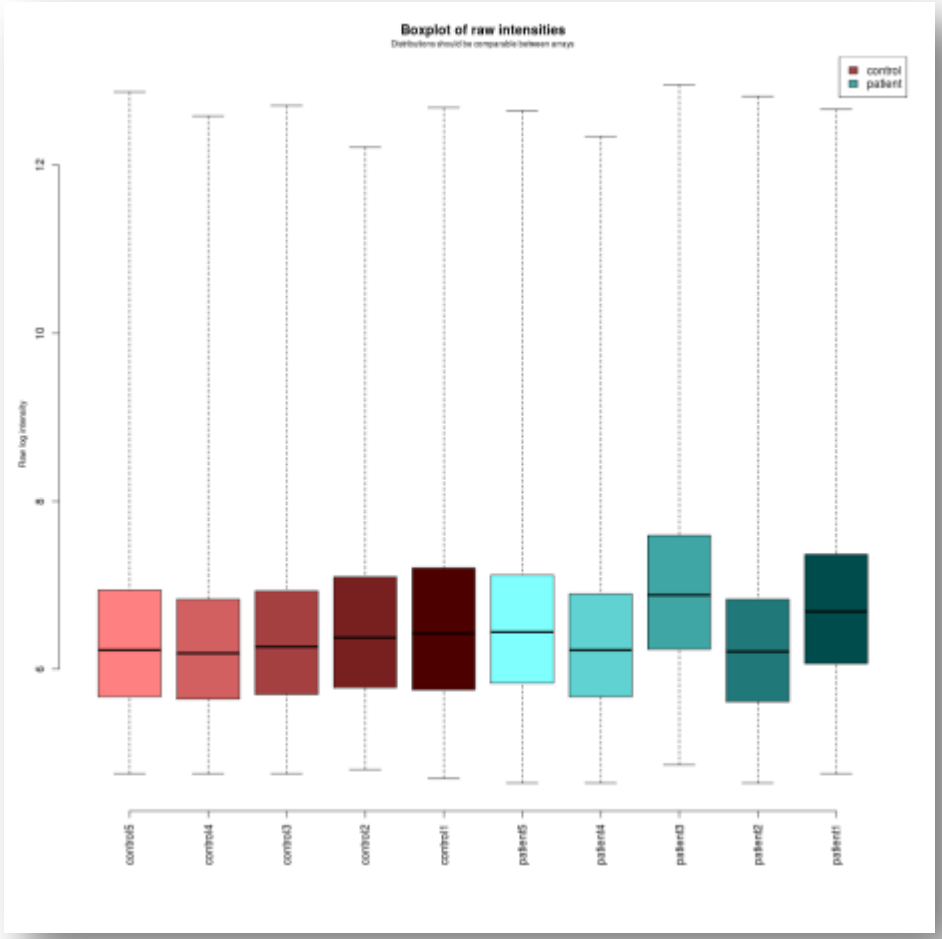


Quality control (QC)

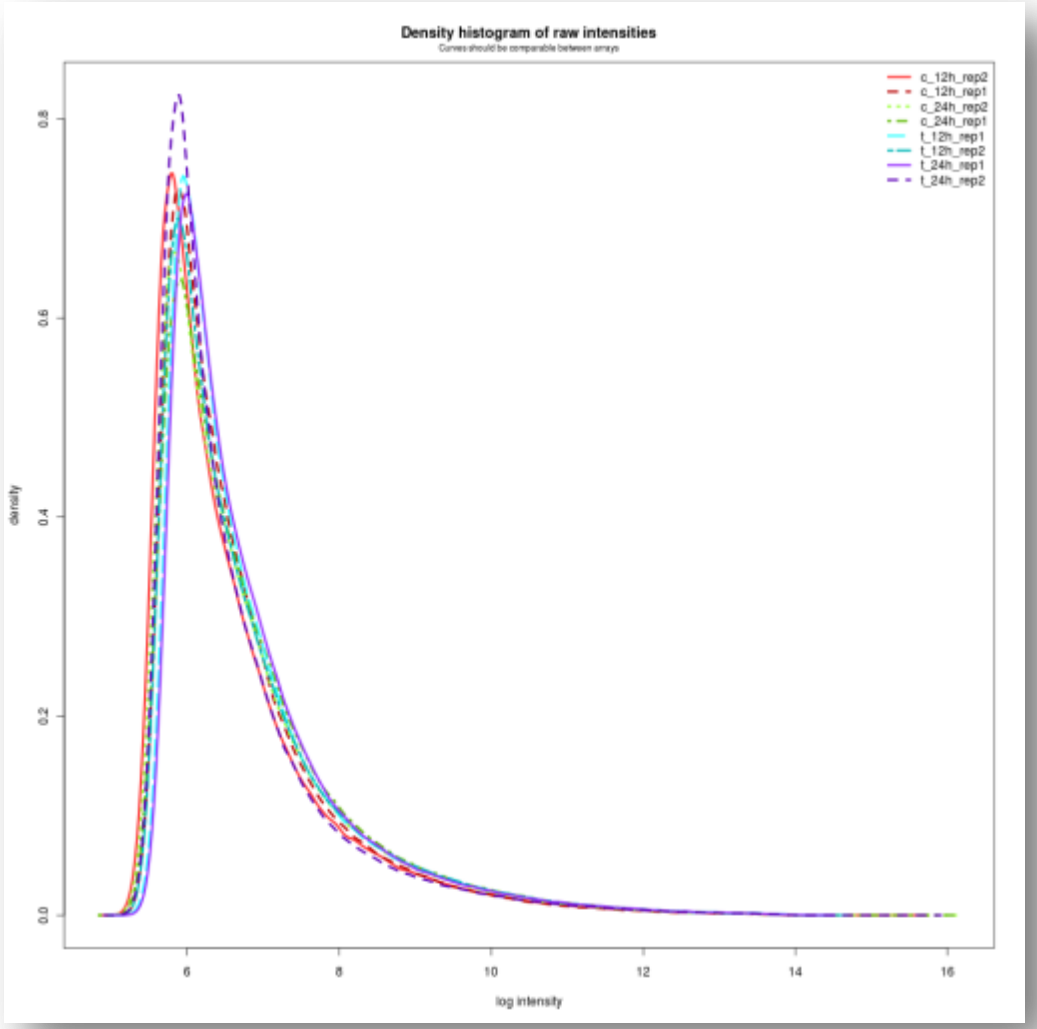
- Ensure comparable signals for all samples:
  - Degraded / low quality sample
  - Failed hybridisation
  - Too low or high overall intensity

- Some differences can be corrected for, others require removal of data from the set

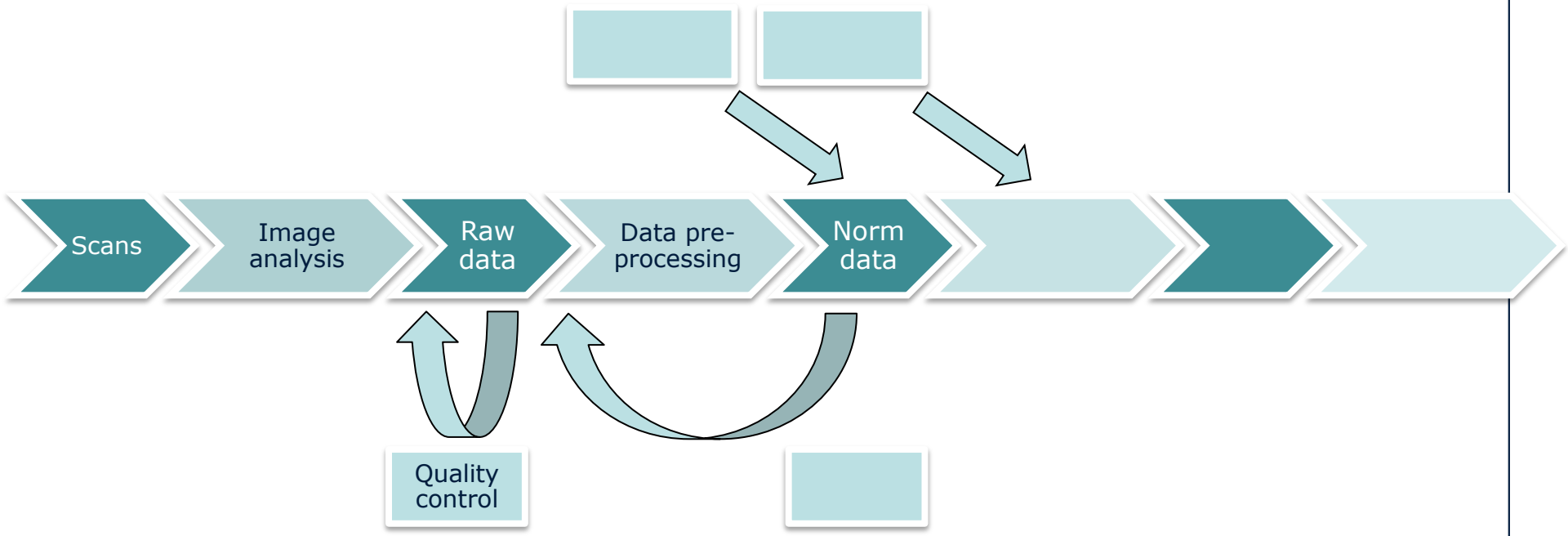
Boxplot


Boxplot of raw intensities
control
patient
Raw log intensity
6
8
10
12
control5
control4
control3
control2
control1
patient5
patient4
patient3
patient2
patient1

Density plot



Data processing workflow



Pre-processing: normalisation

- After discarding bad samples, remaining differences not related to the biology, need to be corrected for
  - Differences in signal strength between samples (*e.g.* because of different amount of starting material)
  - Experimental artifacts
  - Batch effects

- This is mostly done based on the assumption that the overall distribution of the signals of all measured genes should not change between samples

- This is quite robust!
  - Not always true: in such cases one can use other methods (for example using added artificial controls)

Log transformation

- Generally, the measurements are first ^{2}log-transformed
  - The distribution of the logged intensities is more 'normal' than on the original scale
  - ^{2}log is common in biology for reasons of interpretation
- Check whether processed data is given on a log scale or not
- After logging and normalisation one can compute the difference in means ('logFC') between several experimental groups
  - The difference is easier to handle statistically (additive model)

The log Fold Change

- We are normally interested in the ratio of a gene's expression between experimental groups, called the fold change → a / b

- This transforms to a difference on the log scale, the log fold change → **logFC** = $^{2}\log(a/b)$ = **$^{2}\log(a) - {}^{2}\log(b)$**

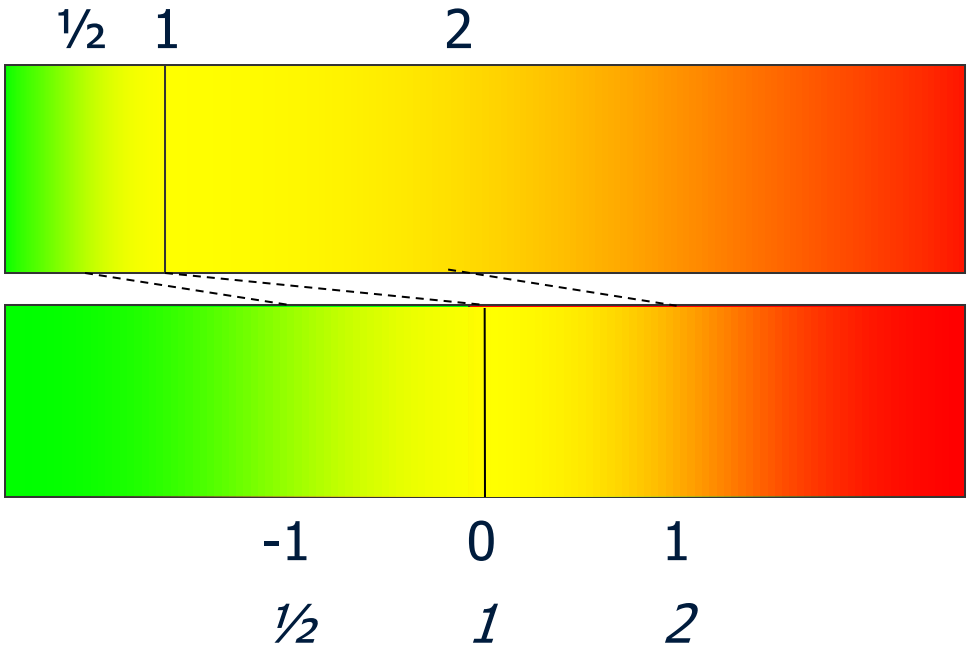
- 2^{logFC} computes the ratio on original scale →

  $2^{logFC} = 2^{(2\log(a) - 2\log(b))} = 2^{(2\log(a/b))} = a / b$

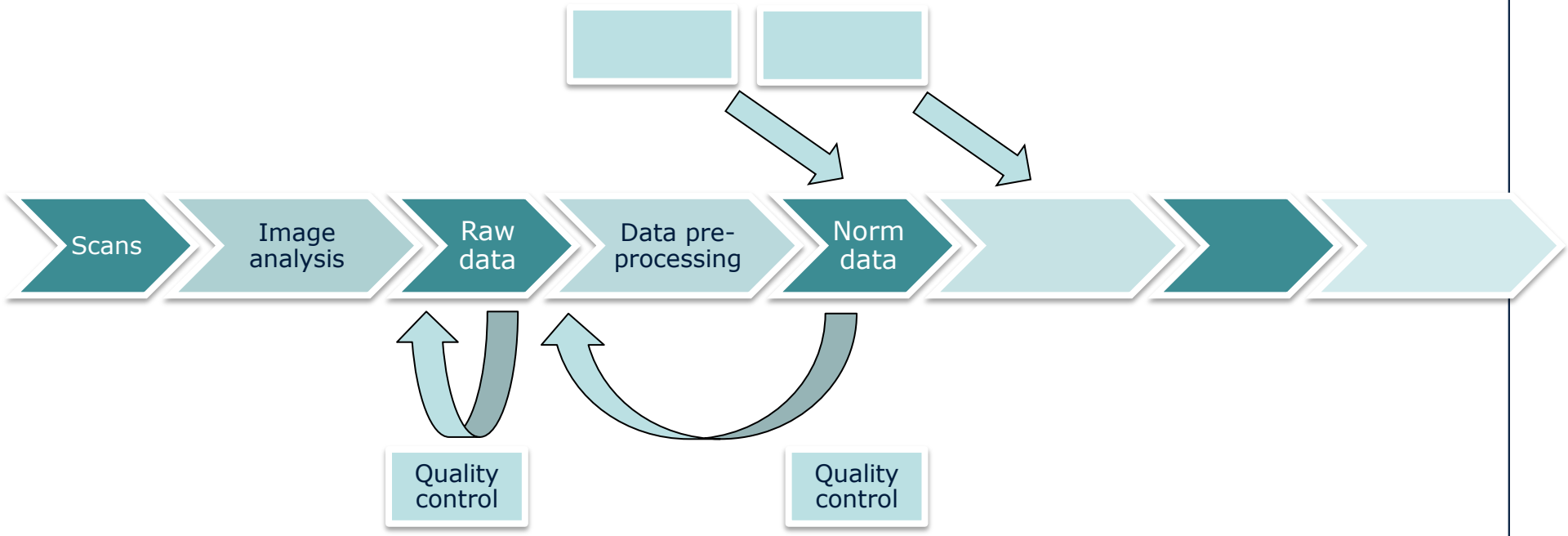
The log Fold Change

- The logFC 'spreads out' the data and offers symmetry

- 'raw' ratio (FC)
- log ratio (logFC)

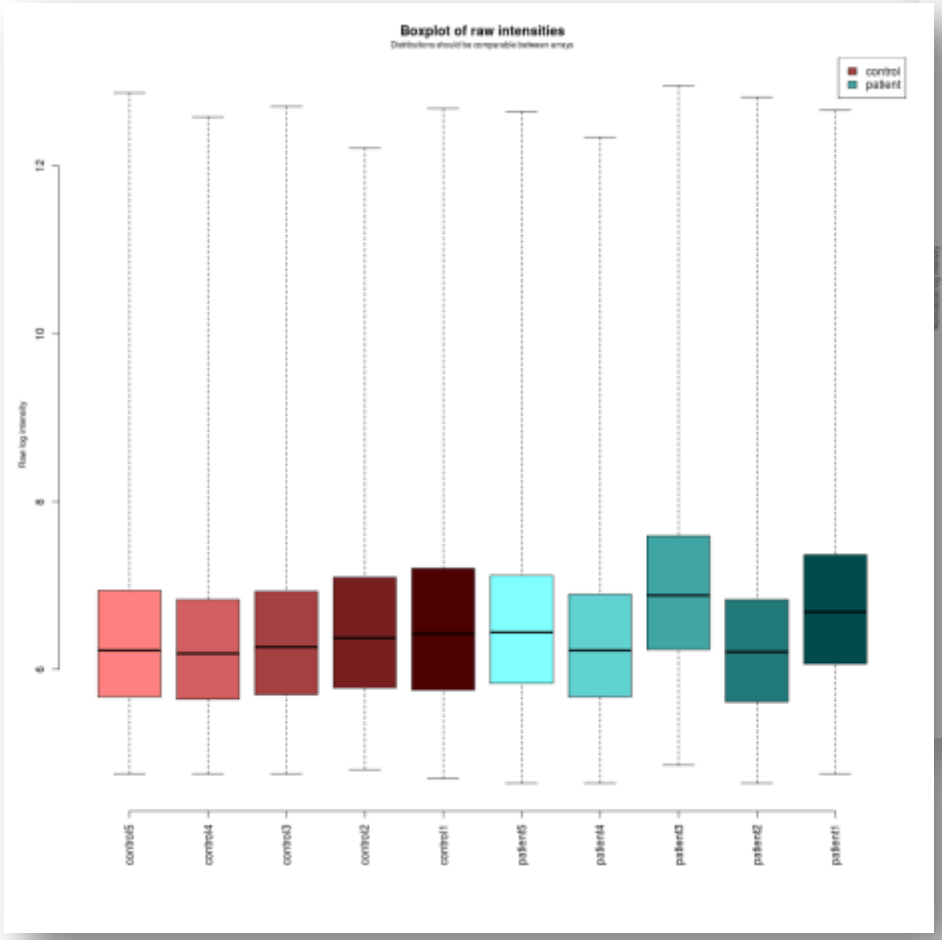


Data processing workflow

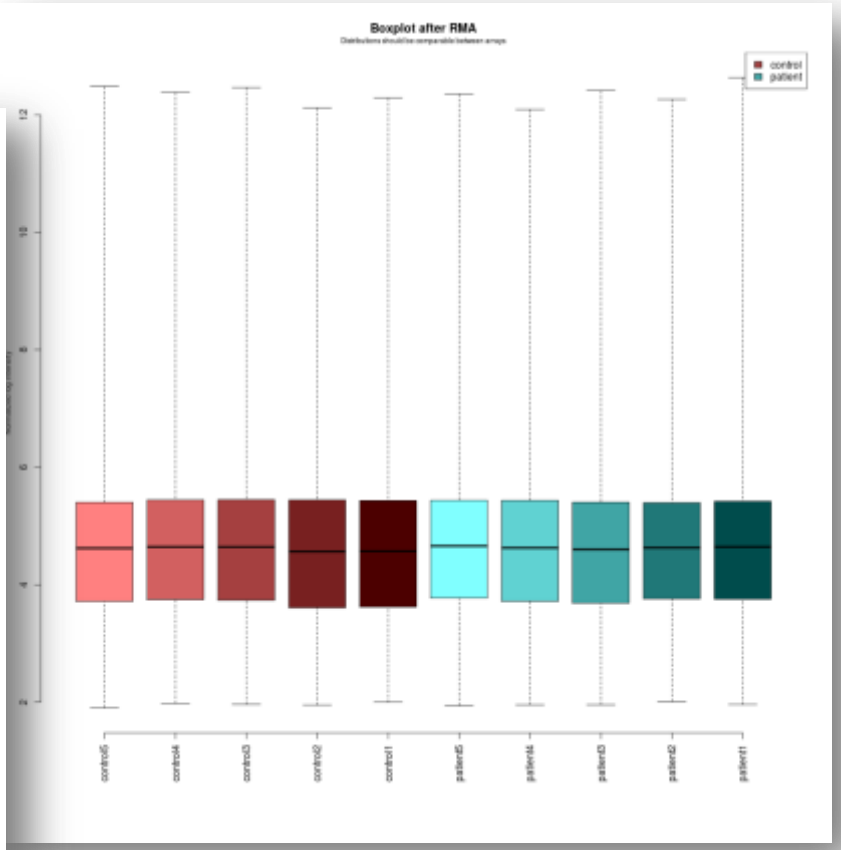


Boxplots

before

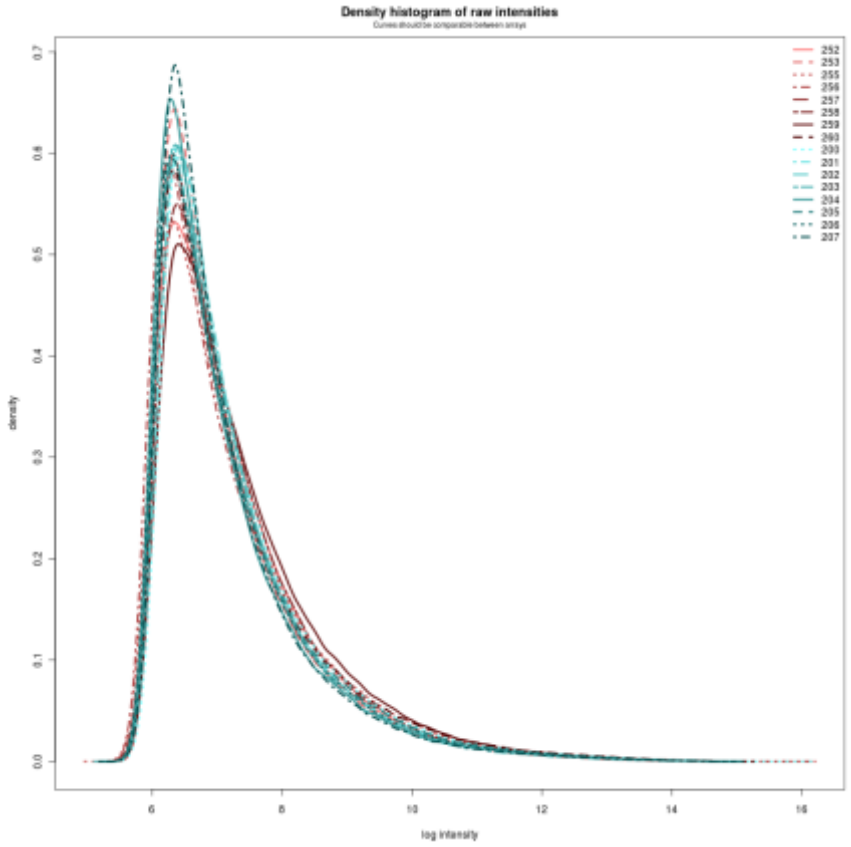


after

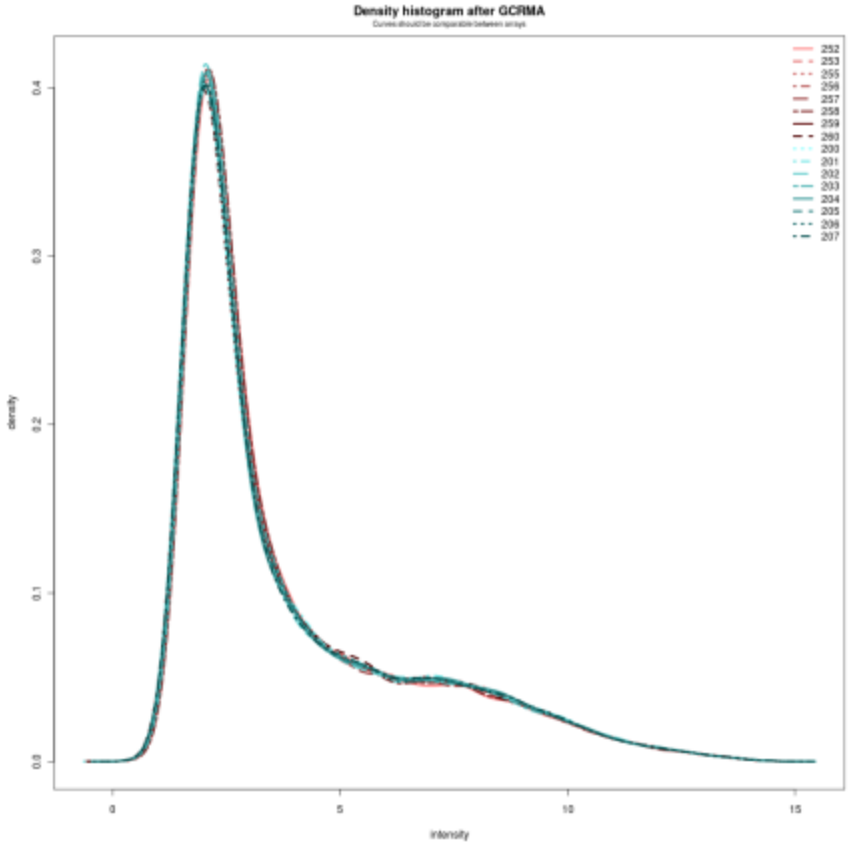


Density plots

before

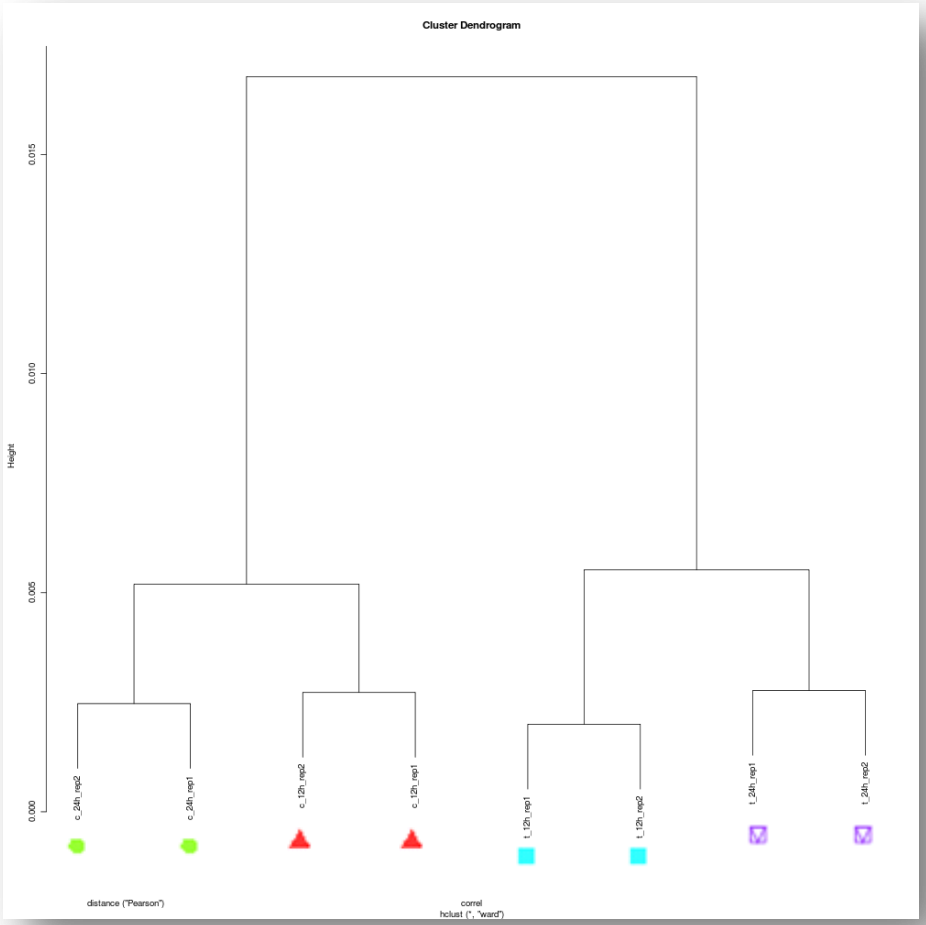


after

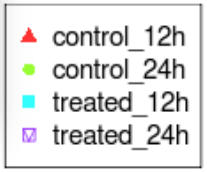


Recall: The assumption is that most genes do not change!

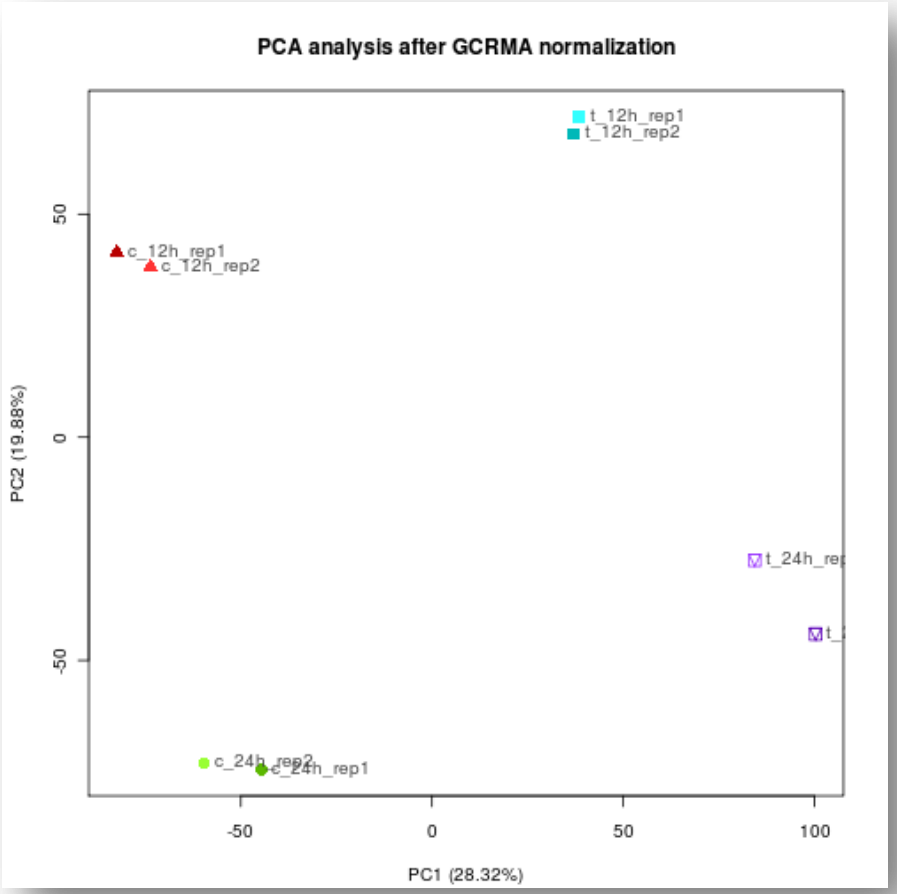
Clustering plot



- Outliers
- Grouping as expected?
- Wrongly grouped samples
- What determines grouping
  - For example: maybe not treatment but sex



PCA plot shows high dimensional data in 2 or 3 dimensions

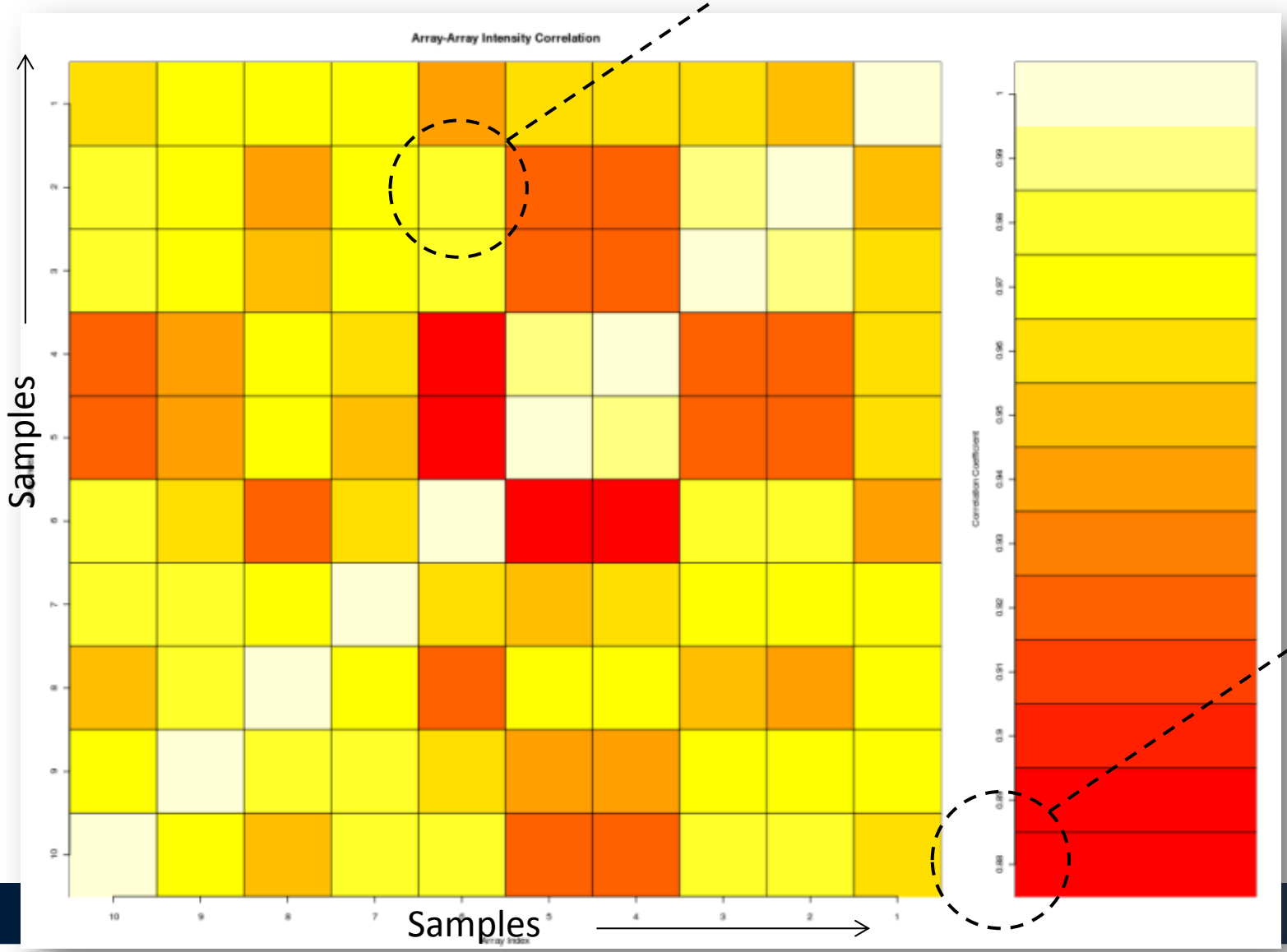


- Outliers
- Grouping as expected?
- Wrongly grouped samples
- What determines grouping
  - For example: maybe not treatment but sex

control_12h
control_24h
treated_12h
treated_24h

Sample correlation plot

Correlation coefficient between two arrays (always between -1 and 1)

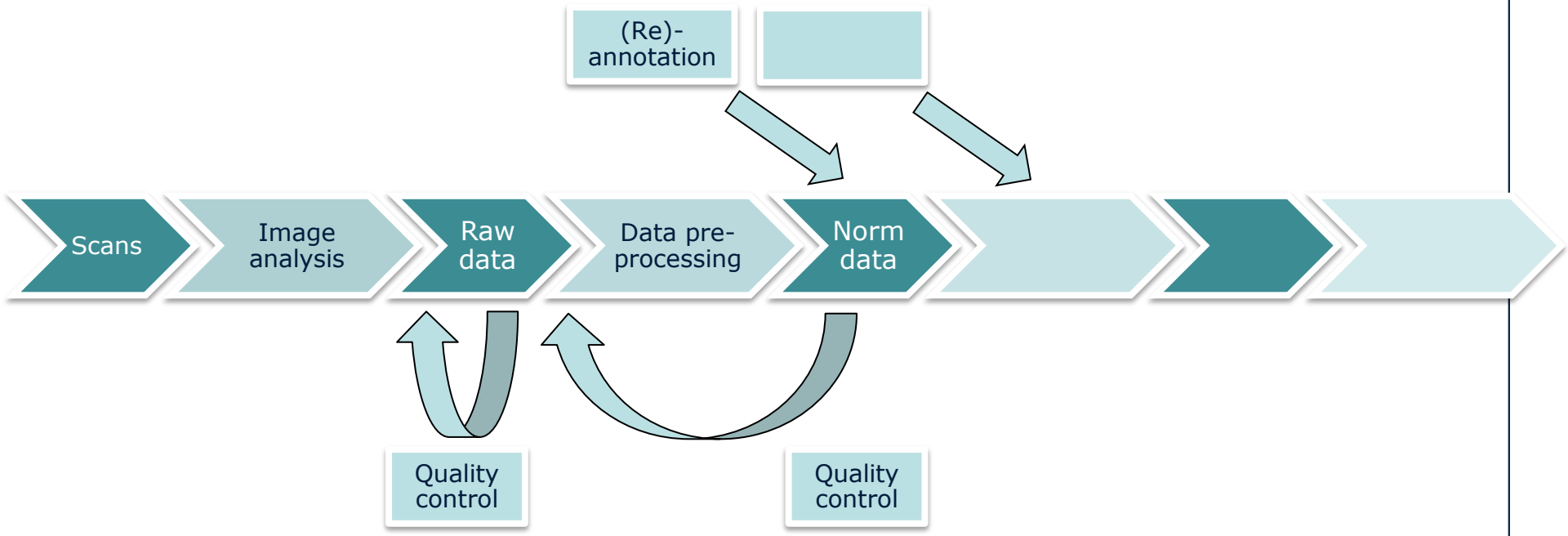


Note what the scale is

Pre-processing: normalisation

- The procedure is cyclic
  - Several QC plots are made before and after normalisation
  - Whether normalisation can correct an artifact may influence decision to discard or not
  - After data selection, the QC and normalisation should be run again
    - Some abberations may have been masked by larger ones
    - Normalised signals should not depend on low quality arrays

Data processing workflow



Normalised data: which genes did we measure?

Gene 1

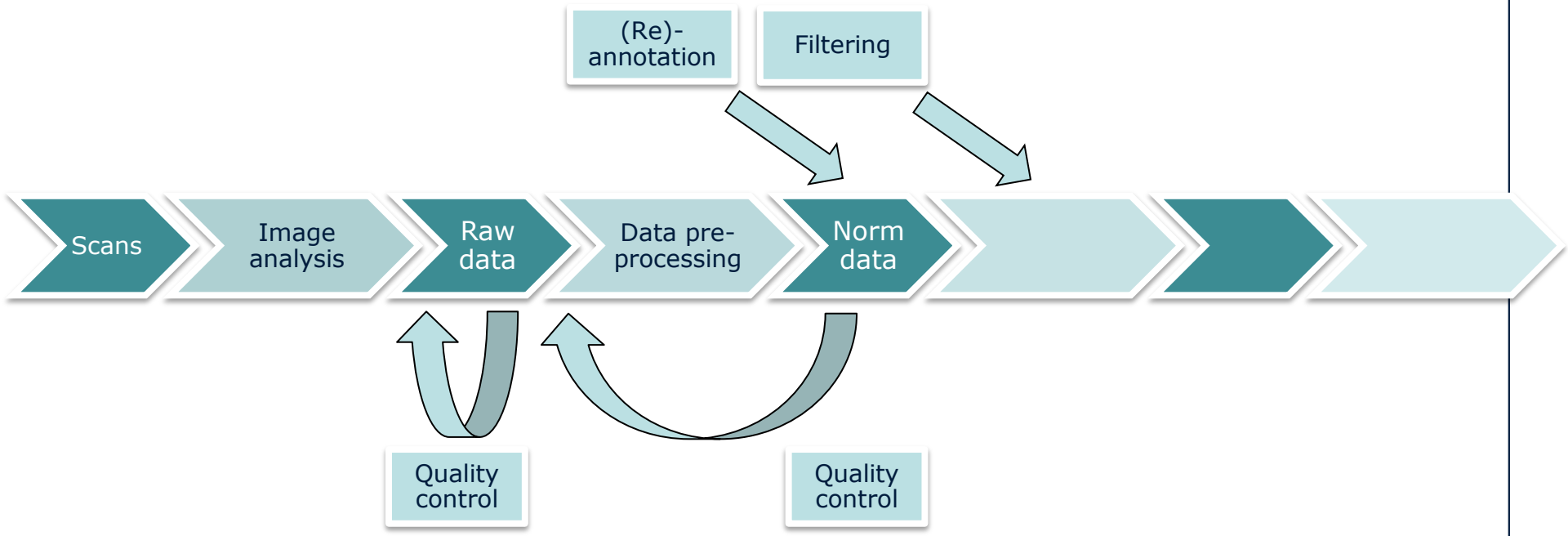
Gene 2

-
-
-

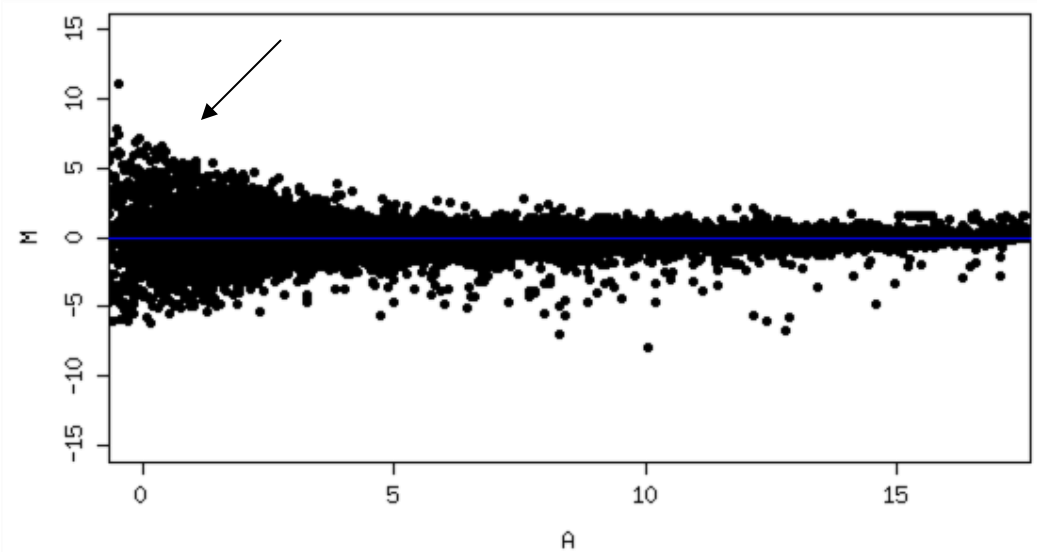
Gene n

	E	F	G	H	I	J	K	L	M	N	O	P
34	0.723067	0.696334	5.306525	2.708209	0.596605	0.595187	0.029721	3.903883	1.415842	3.44363	2.809893	5.065
15	2.335591	0.894245	1.973899	5.080506	1.511485	3.822928	0.446955	4.7461	5.168643	1.145423	4.570865	3.657
07	0.318848	4.792404	2.1378	2.73099	2.057794	0.558325	0.856889	4.393741	1.377838	5.484752	2.485976	2.546
03	3.57681	4.502568	4.119507	4.273115	0.623597	3.64981	4.969225	4.009376	3.362778	4.513491	5.394772	2.548
57	0.366612	2.963851	1.878136	4.950512	5.309345	4.746938	1.777224	3.53458	2.626429	1.692264	4.053316	0.221
42	3.382857	4.816893	4.134014	5.357933	3.338145	3.987972	5.112476	3.495564	1.906187	4.460554	1.444951	3.288
06	0.06643	4.197093	1.023985	5.309899	5.068731	5.247064	4.665507	2.278859	2.859749	1.065216	1.670334	2.254
02	2.169436	0.041848	1.594635	3.90809	3.372297	4.342395	5.489928	3.977514	2.826189	0.683588	2.293742	3.896
54	2.590829	2.793649	3.278129	2.974495	3.964388	3.251174	0.034284	0.325612	1.536994	5.342694	4.968363	3.528
05	3.974013	4.216572	0.465578	3.869911	1.670959	2.752999	0.086357	4.92117	2.85334	0.666545	0.133212	4.813
17	3.727204	0.515586	1.102518	3.35618	4.926224	4.112016	4.657633	0.001114	1.144036	3.622775	0.335591	1.737
57	1.254372	3.988419	3.362662	1.037414	4.636872	0.331022	5.39625	0.012493	3.284902	0.18064	4.35422	2.136
93	3.615515	2.580219	4.667467	2.419086	4.938206	0.499771	3.61686	3.222779	3.887891	4.040124	5.261997	1.476
52	0.695838	0.746584	4.406426	1.030825	0.772952	3.928176	2.162931	1.466699	4.197605	4.417046	2.43303	1.504
29	1.594255	1.170663	2.334343	0.366403	5.155987	4.15595	3.888023	2.284582	1.99963	2.432864	1.249872	1.537
41	2.79293	3.379797	3.154978	1.439589	2.11463	2.893611	1.065311	0.063606	1.681535	3.574217	4.791971	1.497
93	2.631804	0.467399	5.001239	0.949755	3.027005	1.476476	4.408012	0.416456	1.040352	0.619943	0.217544	5.414
56	2.102892	4.242676	2.628069	4.882908	0.380213	1.895572	3.48001	0.957321	0.776458	3.722438	0.491269	1.49
97	0.333237	1.878758	5.445539	0.014734	3.962481	3.945479	5.259968	0.586827	4.845621	1.059331	3.350349	5.472
18	3.974133	4.200288	0.123608	5.447872	0.035139	5.024607	2.764382	2.070159	2.988902	1.071201	3.220618	0.059
39	1.332244	3.27435	2.687692	1.754777	5.065625	2.707253	0.844946	3.880804	1.884298	5.016752	5.110541	0.891
34	0.239654	2.564695	4.927973	0.110017	3.225023	5.238135	2.220898	5.217151	1.165678	3.611153	2.33169	1.993
91	4.074218	1.548933	4.472477	4.482922	3.678754	5.280298	0.735437	0.592868	1.068242	1.221479	0.774905	0.188
07	0.182579	3.470729	3.605185	2.418245	4.535783	3.826456	3.977994	5.492359	4.352	4.630467	0.836987	3.358

Data processing workflow



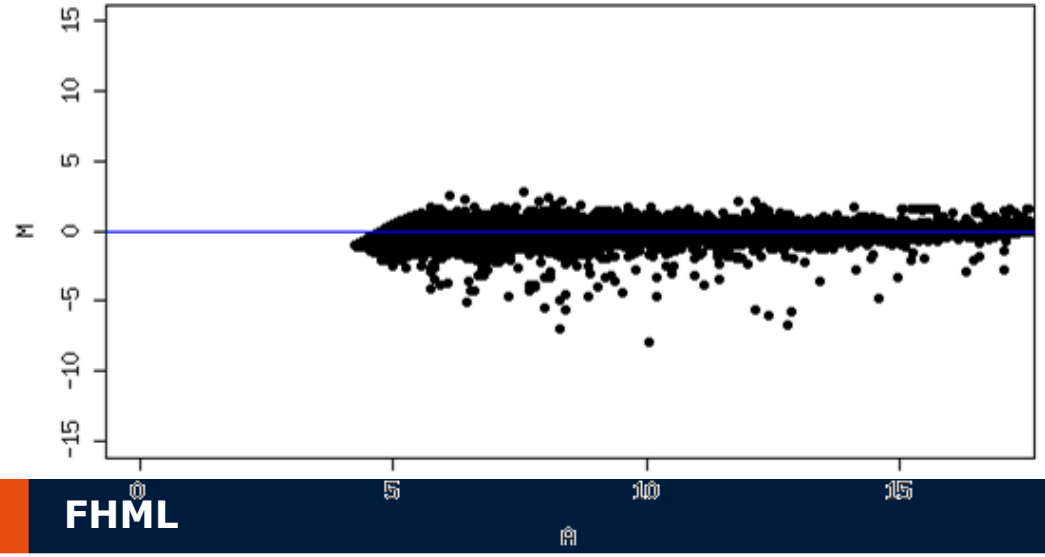
Low intensity filtering



- Before filtering

difference between groups

average intensity



- After filtering
- Low intensity spots are more affected by noise signal

Low intensity (unexpressed) genes may be removed

Gene 1

Gene 2

-

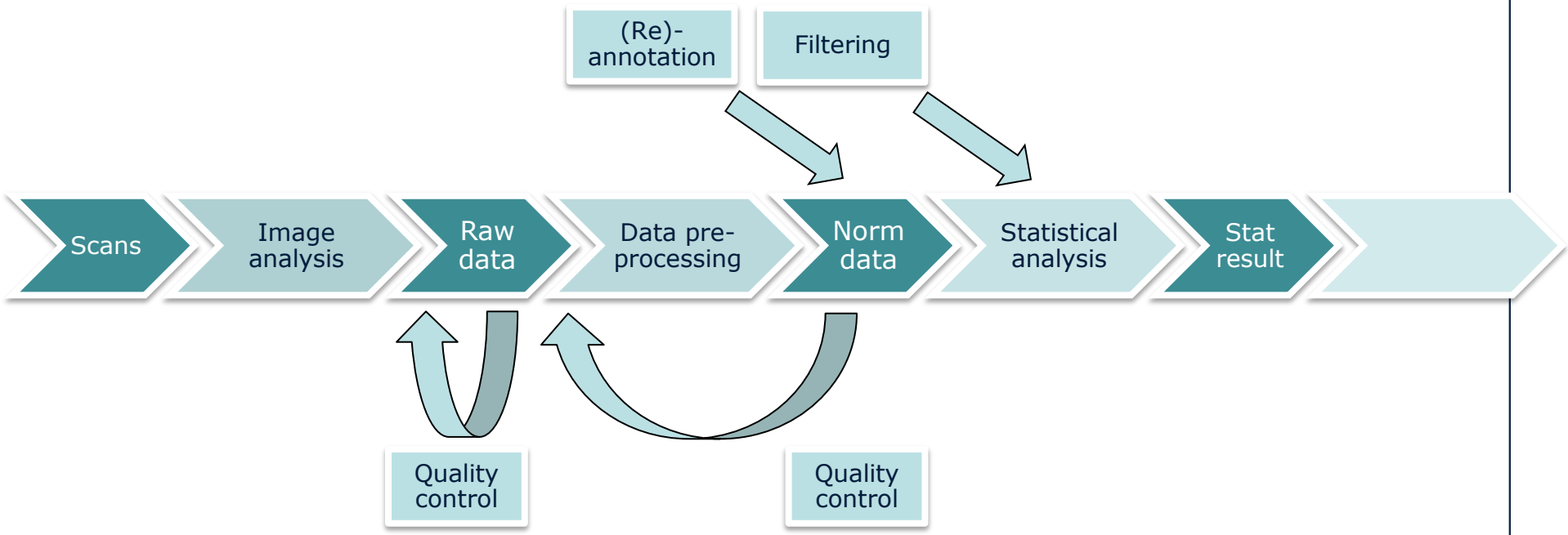
-

-

Gene n

	E	F	G	H	I	J	K	L	M	N	O	P
34	0.723067	0.696334	5.306525	2.708209	0.596605	0.595187	0.029721	3.903883	1.415842	3.44363	2.809893	5.065
15	2.335591	0.894245	1.973899	5.080506	1.511485	3.822928	0.446955	4.7461	5.168643	1.145423	4.570865	3.657
07	0.318848	4.792404	2.1378	2.73099	2.057794	0.558325	0.856889	4.393741	1.377838	5.484752	2.485976	2.546
03	3.57681	4.502568	4.119507	4.273115	0.623597	3.64981	4.969225	4.009376	3.362778	4.513491	5.394772	2.548
57	0.366612	2.963851	1.878136	4.950512	5.309345	4.746938	1.777224	3.53458	2.626429	1.692264	4.053316	0.221
42	3.382857	4.816893	4.134014	5.357933	3.338145	3.987972	5.112476	3.495564	1.906187	4.460554	1.444951	3.288
06	0.06643	4.197093	1.023985	5.309899	5.068731	5.247064	4.665507	2.278859	2.859749	1.065216	1.670334	2.254
02	2.169436	0.041848	1.594635	3.90809	3.372297	4.342395	5.489928	3.977514	2.826189	0.683588	2.293742	3.896
54	2.590829	2.793649	3.278129	2.974495	3.964388	3.251174	0.034284	0.325612	1.536994	5.342694	4.968363	3.528
05	3.974013	4.216572	0.465578	3.869911	1.670959	2.752999	0.086357	4.92117	2.85334	0.666545	0.133212	4.813
17	3.727204	0.515586	1.102518	3.35618	4.926224	4.112016	4.657633	0.001114	1.144036	3.622775	0.335591	1.737
57	1.254372	3.988419	3.362662	1.037414	4.636872	0.331022	5.39625	0.012493	3.284902	0.18064	4.35422	2.136
93	3.615515	2.580219	4.667467	2.419086	4.938206	0.499771	3.61686	3.222779	3.887891	4.040124	5.261997	1.476
52	0.695838	0.746584	4.406426	1.030825	0.772952	3.928176	2.162931	1.466699	4.197605	4.417046	2.43303	1.504
29	1.594255	1.170663	2.334343	0.366403	5.155987	4.15595	3.888023	2.284582	1.99963	2.432864	1.249872	1.537
41	2.79293	3.379797	3.154978	1.439589	2.11463	2.893611	1.065311	0.063606	1.681535	3.574217	4.791971	1.497
93	2.631804	0.467399	5.001239	0.949755	3.027005	1.476476	4.408012	0.416456	1.040352	0.619943	0.217544	5.414
56	2.102892	4.242676	2.628069	4.882908	0.380213	1.895572	3.48001	0.957321	0.776458	3.722438	0.491269	1.49
97	0.333237	1.878758	5.445539	0.014734	3.962481	3.945479	5.259968	0.586827	4.845621	1.059331	3.350349	5.472
18	3.974133	4.200288	0.123608	5.447872	0.035139	5.024607	2.764382	2.070159	2.988902	1.071201	3.220618	0.059
39	1.332244	3.27435	2.687692	1.754777	5.065625	2.707253	0.844946	3.880804	1.884298	5.016752	5.110541	0.891
34	0.239654	2.564695	4.927973	0.110017	3.225023	5.238135	2.220898	5.217151	1.165678	3.611153	2.33169	1.993
91	4.074218	1.548933	4.472477	4.482922	3.678754	5.280298	0.735437	0.592868	1.068242	1.221479	0.774905	0.188
07	0.182579	3.470729	3.605185	2.418245	4.535783	3.826456	3.977994	5.492359	4.352	4.630467	0.836987	3.358

Data processing workflow



Finding interesting genes

- Once we have numbers for the measurements of the genes, we need a way to find the genes which are interesting to us

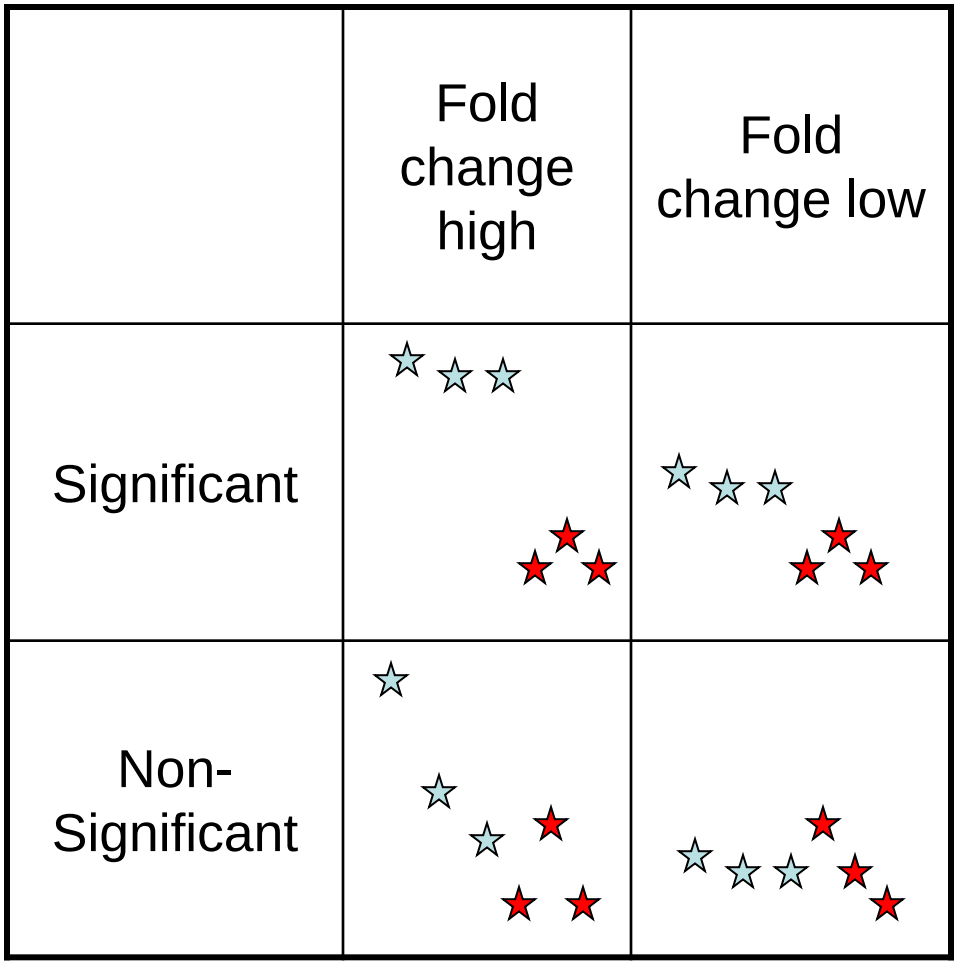
- We need to compare groups!
  - Significance in a statistical test looking for differences between two or more groups
  - Fold change between two conditions
  - Correlation to another feature of interest

- But we also need to take the *multiple testing problem* in to account

Dr. Rachel Cavill, Department of Knowledge Engineering, FHS, kindly provided some of these slides

Using fold change and statistical significance

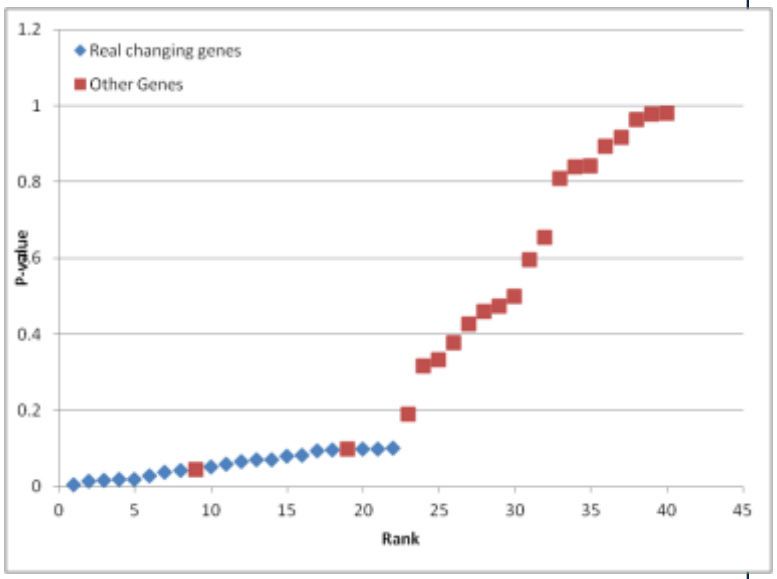
Often people use both fold change and statistical significance between two groups to determine the list of significant genes



Why do we need to worry about multiple testing?

If we have 10,000 measurements for each item in 2 groups, with a t-test we find measurements different between the two groups…

- We will expect 500 of the measurements to be significantly different in the t-test ($p<0.05$)



With 10,000+ genes measured by each microarray, we can get many **false positive** results.

How do we deal with multiple testing?

- Examples of multiple testing correction methods:

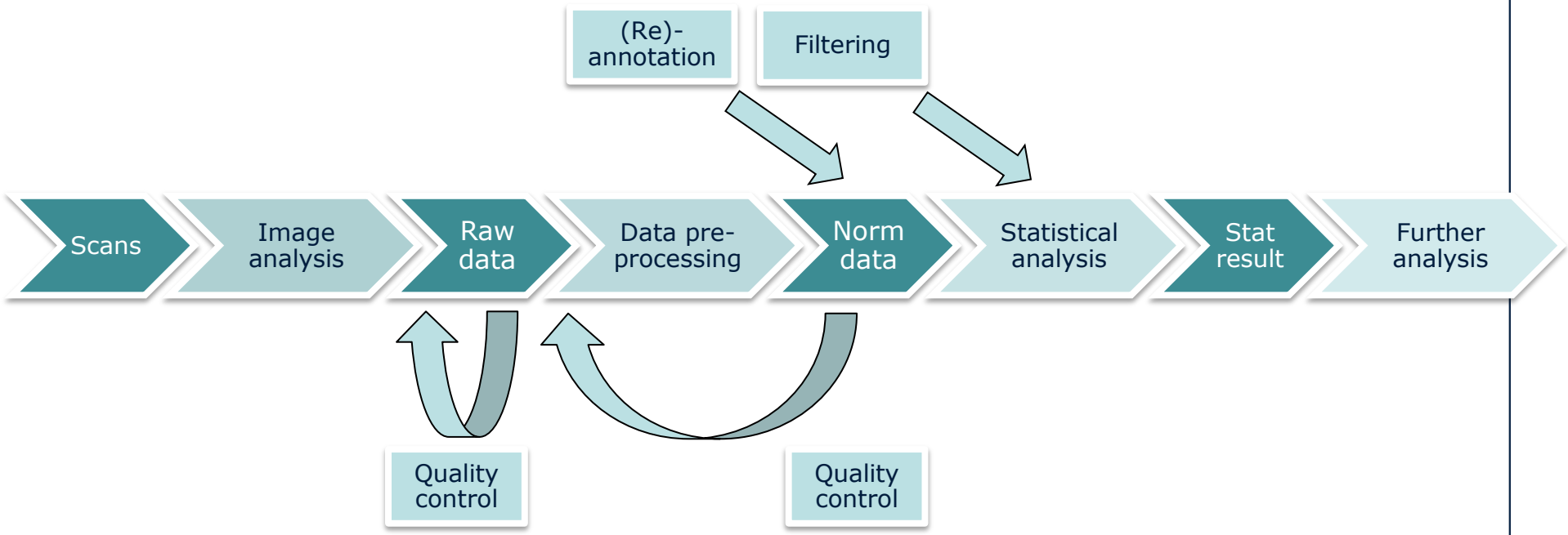
  • **Bonferroni** – a very strict correction, very few false positives remain, but we will discount many true positives too.

  Adjusted p-value = calculated p-value * number of tests done

  E.g. when we test 100 genes to see if they are different between the two groups. A certain gene gives a p-value of 0.002, the adjusted p-value is;
  0.002 * 100 = 0.20 – not significant.

  • **Benjamini-Hochberg** – we set the % of results which we can tolerate as false positives (False Discovery Rate or FDR control)

Data processing workflow



What next?

- Once we have found a list of genes which are correlated or significantly changed between two groups, we often still have 1,000's of genes to consider.
- We may use only the most changed genes only to further study the differences between the groups
- Or search the literature
- Or better…apply:
  - Clustering methods
  - Correlation methods
  - Classification methods
  - Pathway analysis
  - Gene Set Enrichment Analysis
  - Gene Ontology analysis
  - Network analysis
  - …

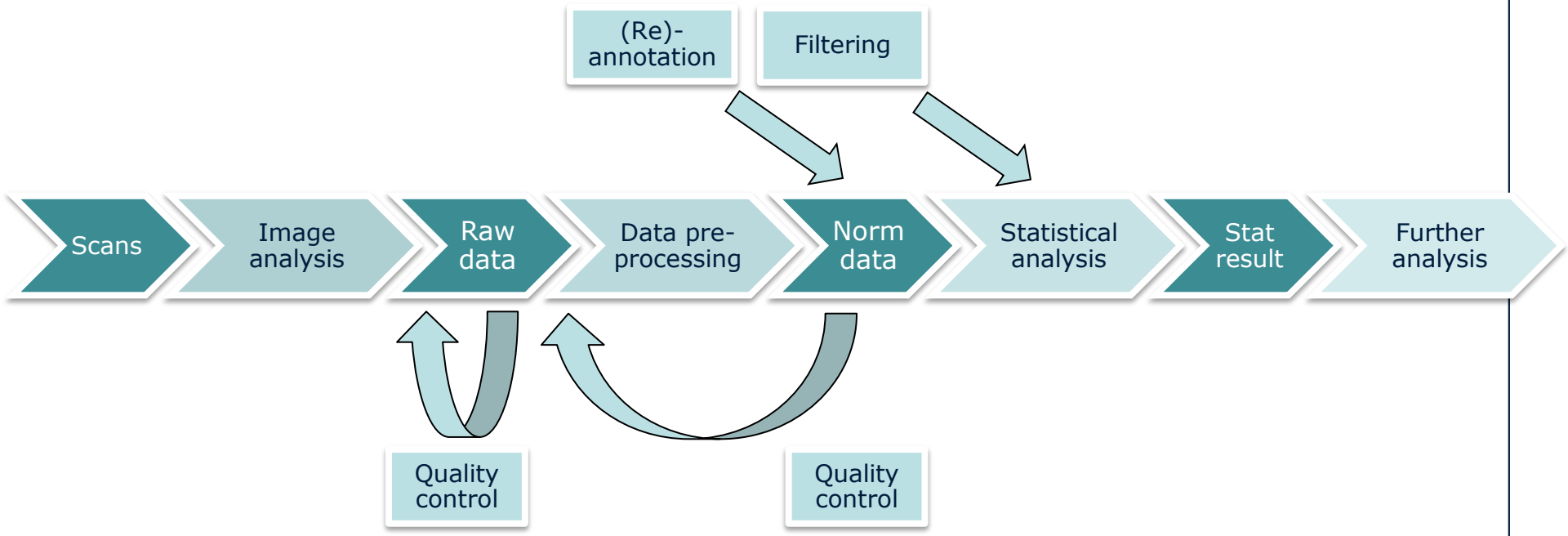
→ some of those you will see in the next practical

Recall: hypothesis generating

→ eventually, get back to the lab or study subjects to biologically verify findings

Data processing workflow

The basic principles are generic:



The details are different:

Dependency on technology

Dependency on biological question

Next Generation Sequencing (NGS)

- A more recent development is the use of Next-Generation Sequencing (NGS) or High Throughput Sequencing (HTS) to measure mRNA

→ RNA-Seq

→ It "reads" all the mRNA fragments provided, giving us counts of the frequency of each mRNA across the whole genome

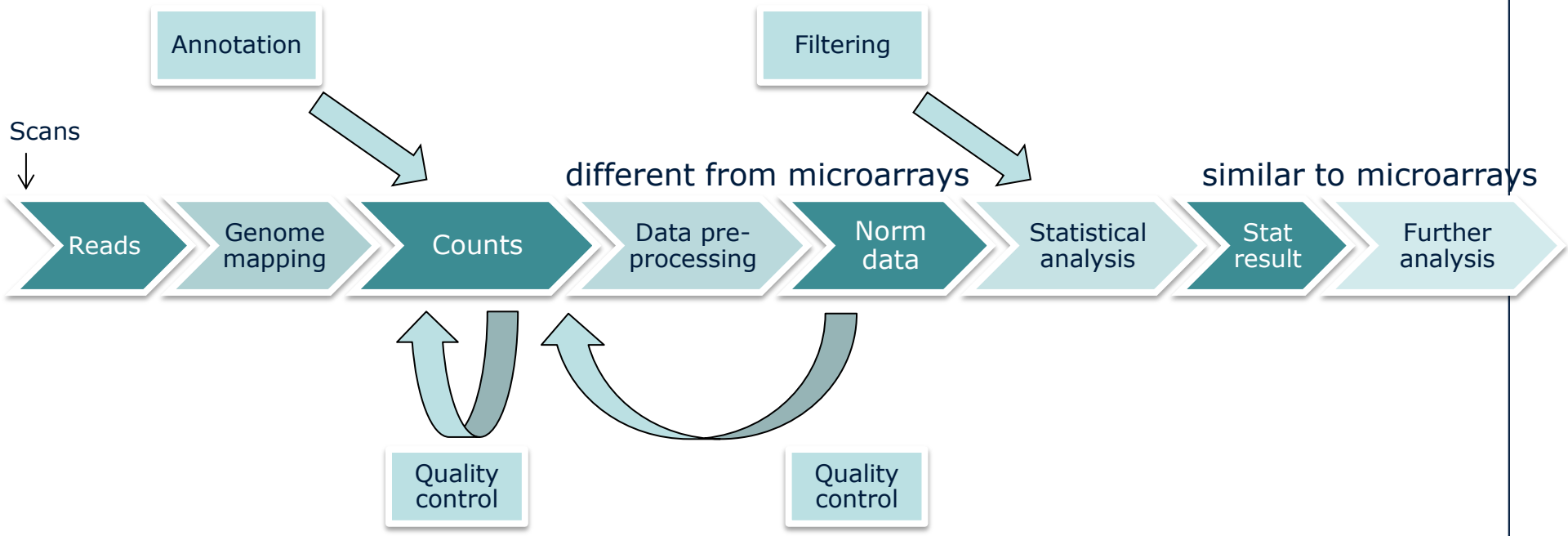
- And to identify genetic variants

→ DNA-Seq

→ Also possible for coding regions based on RNA-Seq

- And many other applications (not discussed now)

Data processing workflow for RNA-seq



http://www.arrayanalysis.org

ARRAY ANALYSIS .ORG

- Friendly solutions for standardised high throughput data analysis -

Get started | Download sources | QC Modules description | Documentation | Bug tracker

QUICK LINKS

[Affymetrix QC & pre-processing]

[Illumina QC & pre-processing] NEW!

[Statistical analysis]

[Pathway analysis] NEW!

BiGCaT bioinformatics

Member of: DTL

In collaboration with: EdgeLeap, TNO

Welcome to ArrayAnalysis.org !

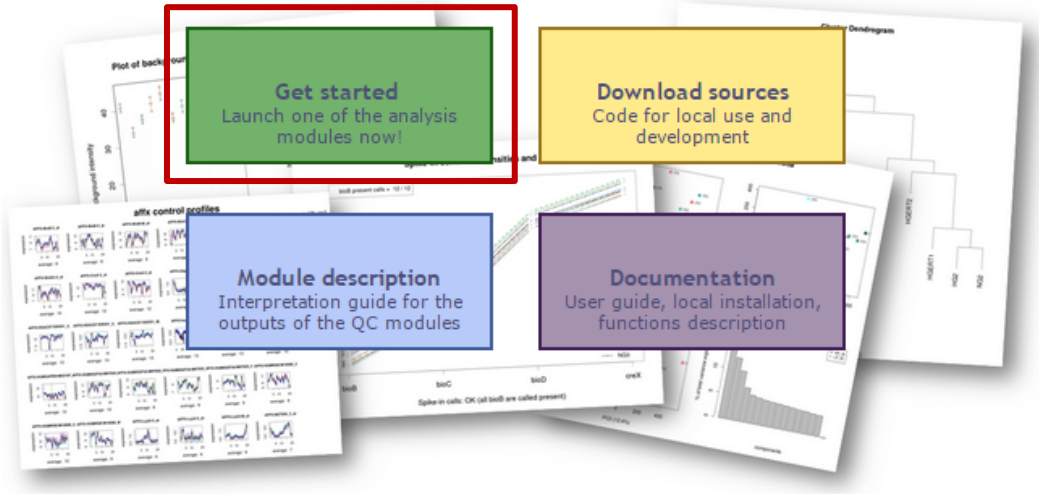
Cite ArrayAnalysis

ArrayAnalysis offers user-friendly solutions for gene expression data analysis, from raw data to biological pathways. It contains modules of three types that can be lauched individually or successively as an integrated workflow.

[QC & pre-processing] module gathers a complete panel of QC plots and indicators: a variety output plots or tables help you determine sample quality, hybridisation and overall signal quality, signal comparability and bias diagnostic and array correlation. Pre-processing methods combine probe set re-annotation, background correction and normalisation. Currently, modules are available for Affymetrix and Illumina arrays.

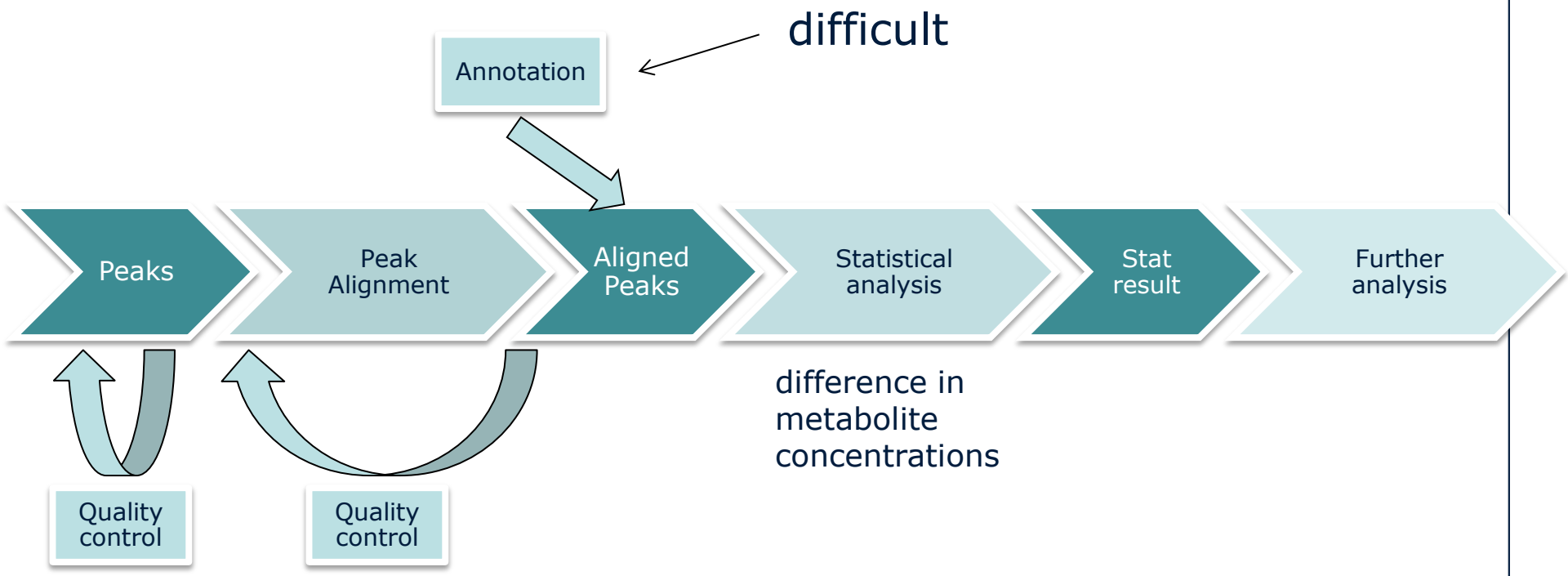
[Statistical analysis] module models your gene expression data using a linear model applied at the probe set level. You are given the possibility to custom your analysis and computing several models on a run. For a quick interpretation of the output result, P-Value and Fold change histograms can be computed as well as custom summary tables.

[Pathway analysis] module allows to quickly and easily visualise your statistics results on a biological pathway basis and identify significantly changed processes using PathVisio technology. This module will be activated soon, for now a mock-up module is in place that shows the possibilities using an example data sets.



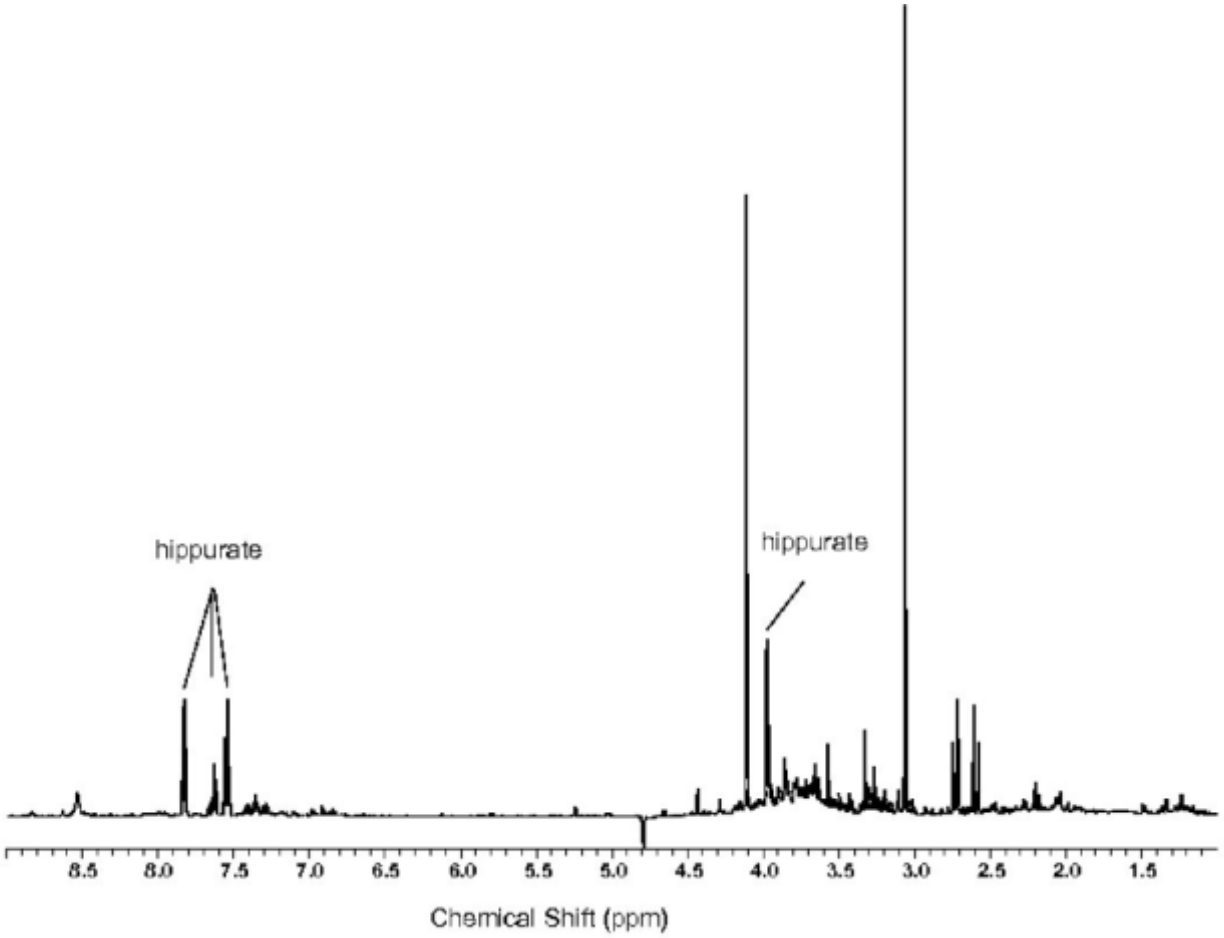
We gratefully acknowledge all authors of R/BioConductor packages used by ArrayAnalysis.org.

Data processing workflow for (untargeted*) metabolomics



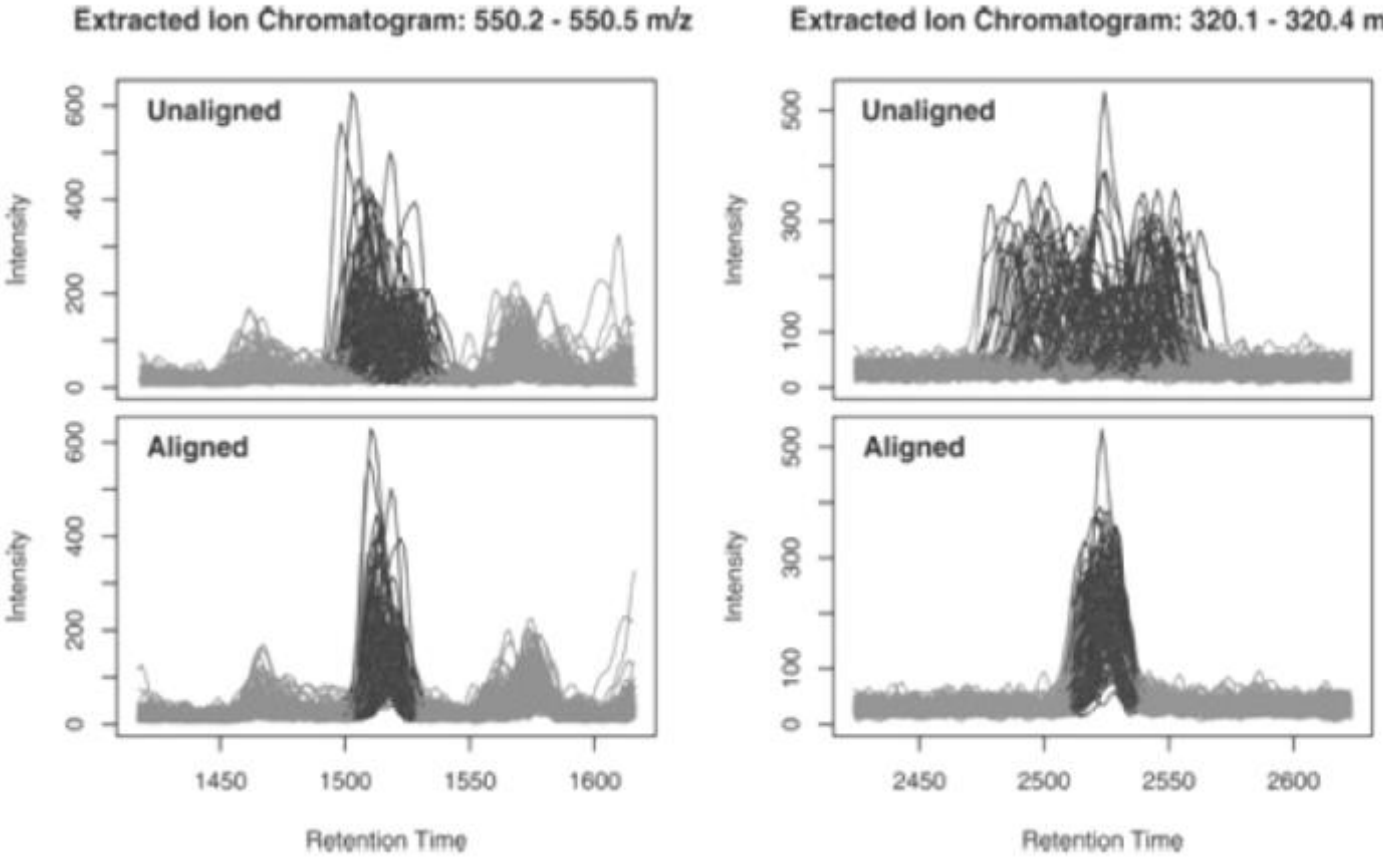
(* targeted metabolomics is easier: you know what you are measuring, and get (raw) values directly)

Visible Metabolome: NMR of urine



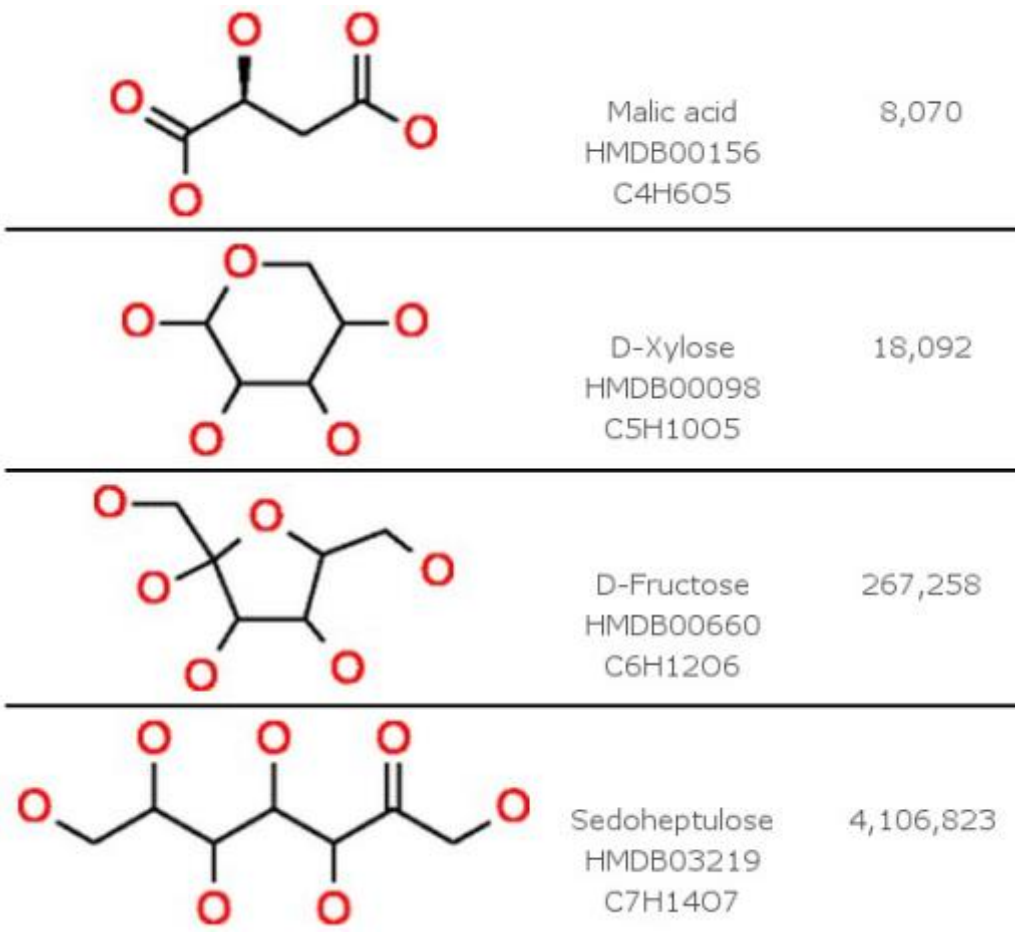
Bryan et al. BMC Bioinformatics 2008 9:470 doi:10.1186/1471-2105-9-470

Metabolomics LC/MS & GC/MS: Peak Alignment



Smith et al., XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification, 2006, Anal. Chem.

Identification: M/Z to structure



Note: some analysis is already possible without identification (clustering, PCA, classification)

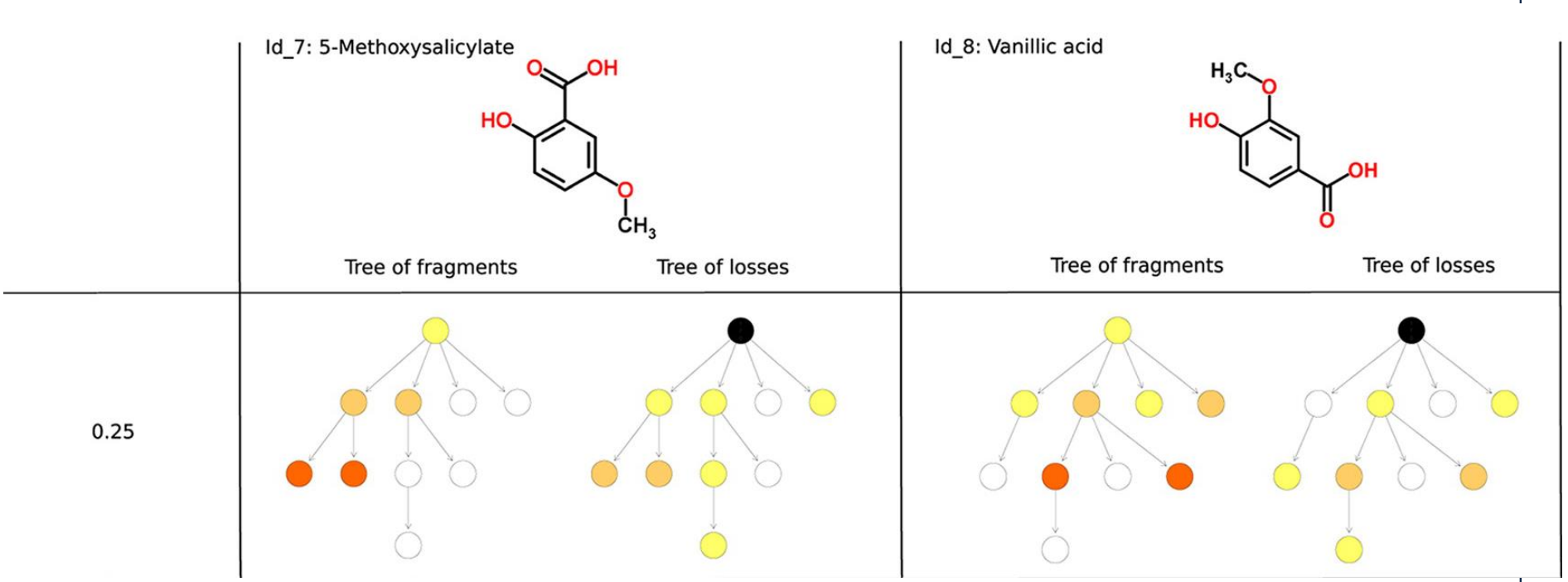
But to biologically interpret we need identified metabolites

Note 2: adducts may change the M/Z value

Peironcely et al. JChemInf 2012 4:21 doi:10.1186/1758-2946-4-21

Which identiy is the correct one?

Fragment peaks may help



Rojas-Cherto et al. Anal. Chem., 2012, 84 (13), p 5524–5534, DOI: 10.1021/ac2034216

Part 2: Finding information about genes and metabolites

Genome databases

- Ensembl (Europe)

- NCBI (US)

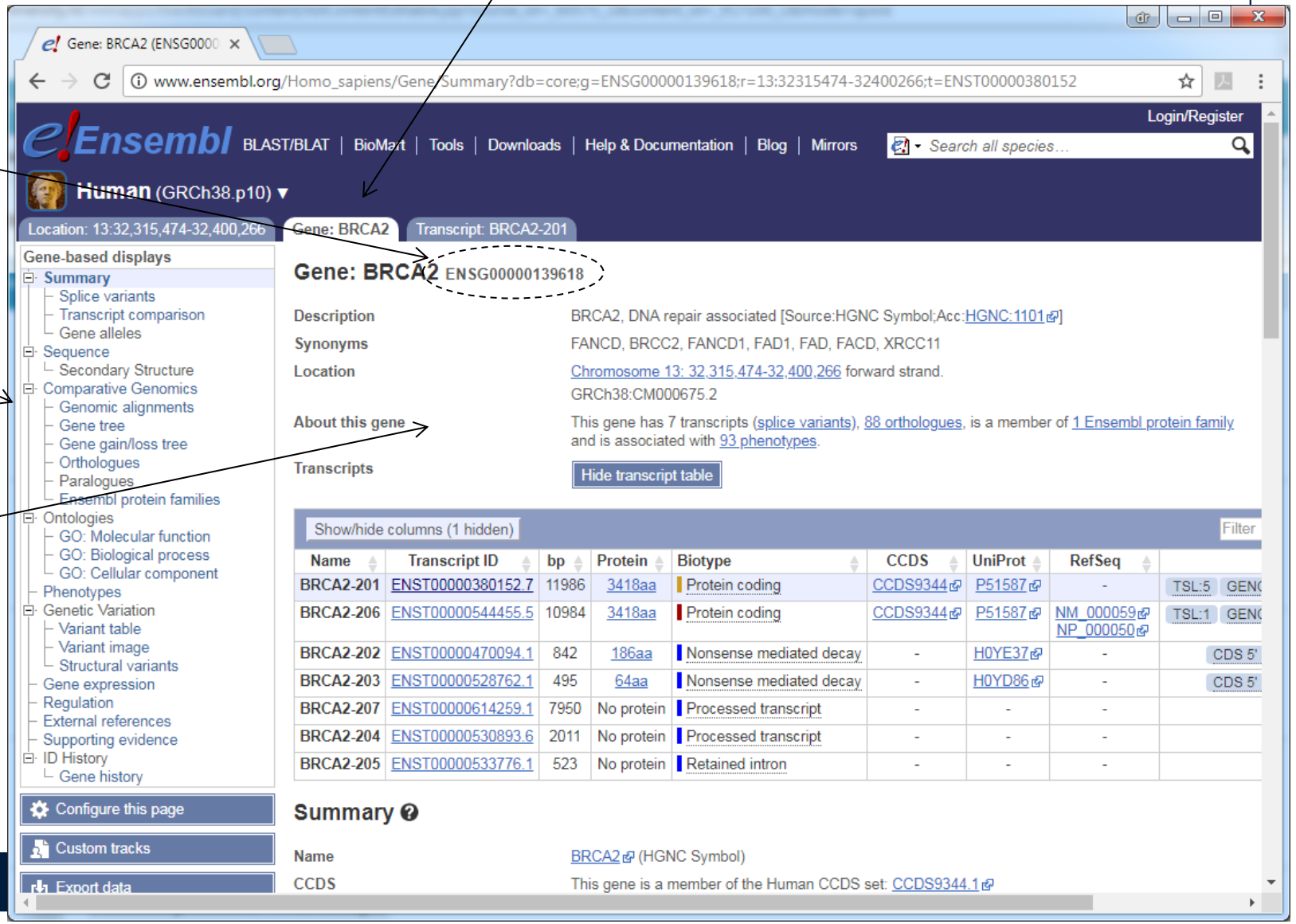
Ensembl interface

Tabs (location, gene, transcript, ...

Ensembl gene identifier

Menu

Information pane



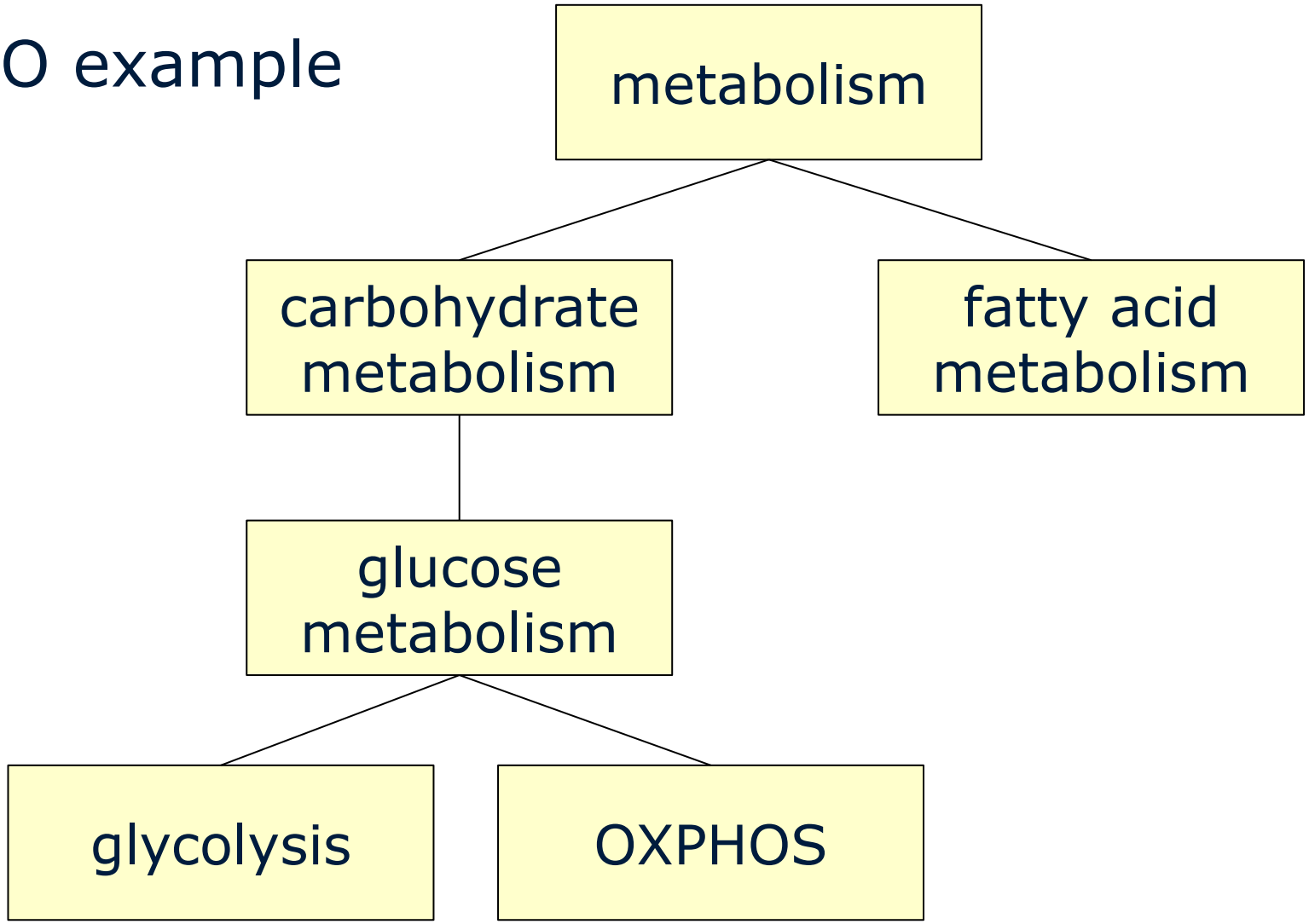
Gene (protein) function: Gene Ontology (GO) terms

- These are systematic terms that indicate:
  - biological processes
  - molecular functions
  - cellular localisations

  which proteins are involved in

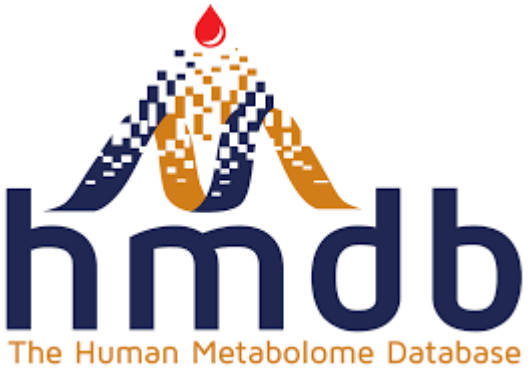
- These terms are in a tree-structure
  - nested, dependent!

GO example

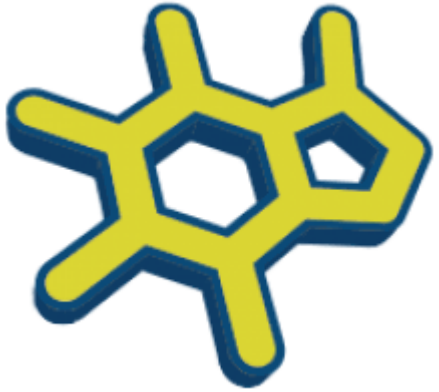

metabolism
carbohydrate metabolism
fatty acid metabolism
glucose metabolism
glycolysis
OXPHOS

Metabolite databases

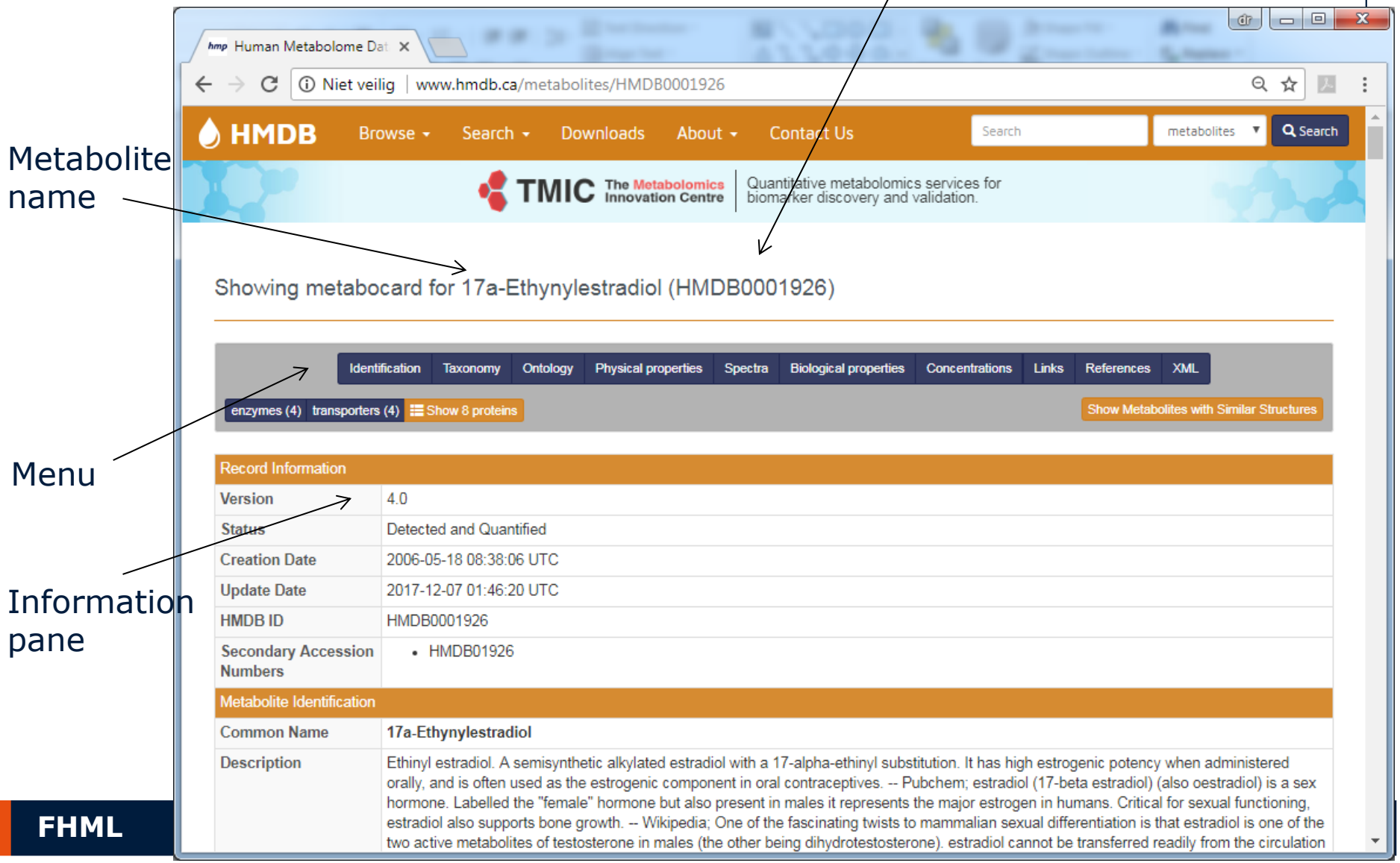
- HMDB contains human metabolites (Human Metabolome Database)



- ChEBI contains all metabolites



HMDB interface



Today's practical

- Look up a dataset related to thyroid neoplasia in ArrayExpress and check which information is provided
- Look at some QC images for this dataset*
- Evaluate statistical results from this dataset*
- Find information about a strongly changed gene in Ensembl
- Look at metabolomics results from another study on thyroid neoplasia
- Look up some information about T3 and T4 in HMDB

* These we have generated for you using ArrayAnalysis

Organisational aspects

- At the end of the practical, make sure you get **signed off**
- This will be done digitally by any of the supervisors

- If you finish early and want te leave, you have to show your answers to the supervisor before being signed off

- The practical takes 4 hours
  - 9.00 am – 1.00 pm or 1.30 pm – 5.30 pm
- After 2 hours there is a 15 minute break (return on time!)
  - At 11.00 am or 3.30 pm

Practical coordinator

Lars Eijssen
l.eijssen@maastrichtuniversity.nl

GOOD LUCK!