Microarray data analysis

Mihaela Martis

NBIS & Faculty of Medicine and Health Sciences
Division Cell Biology, IKE

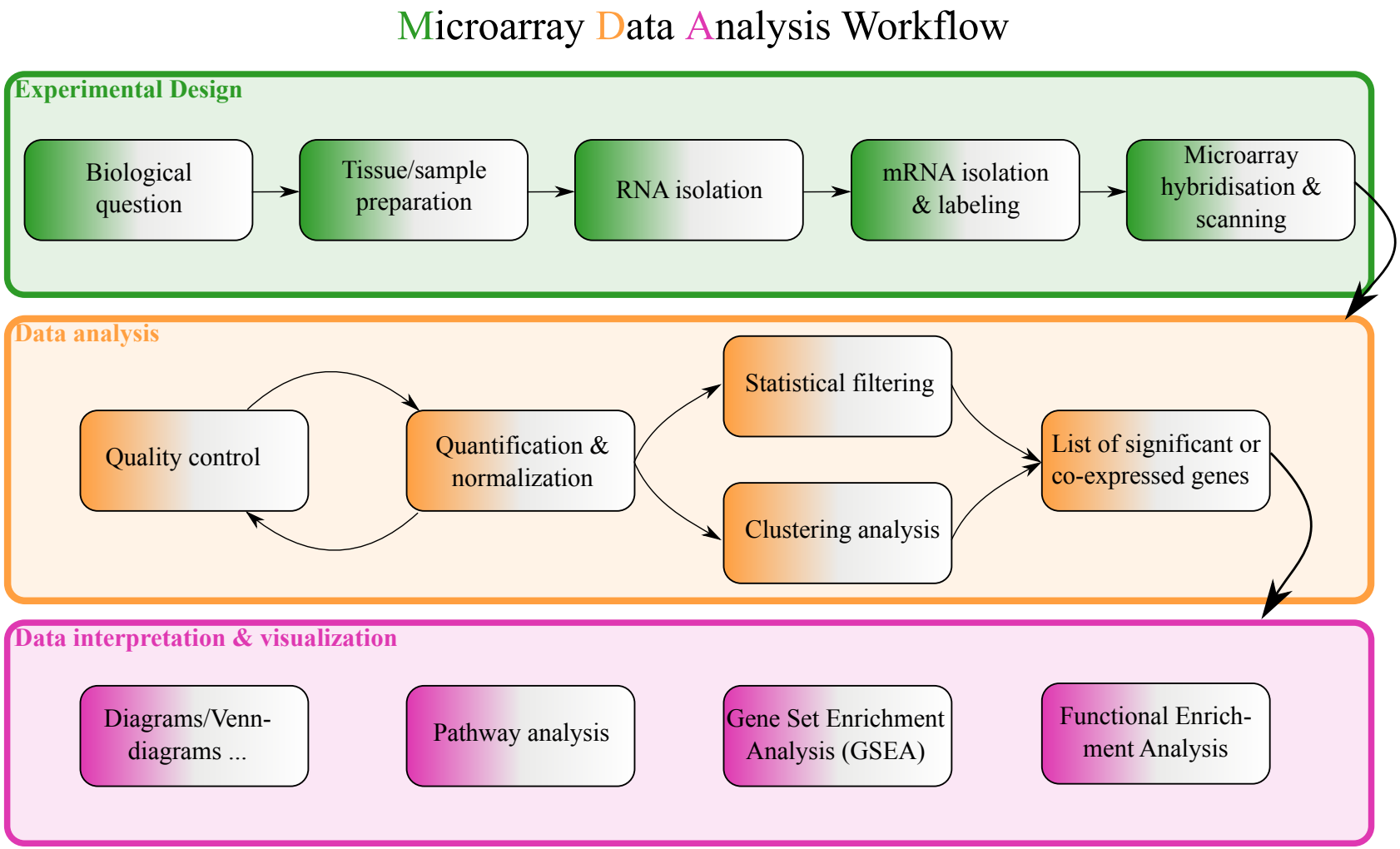
NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

elixir SWEDEN elixir SLOVENIA

SciLifeLab

LINKÖPING UNIVERSITY

Workflow

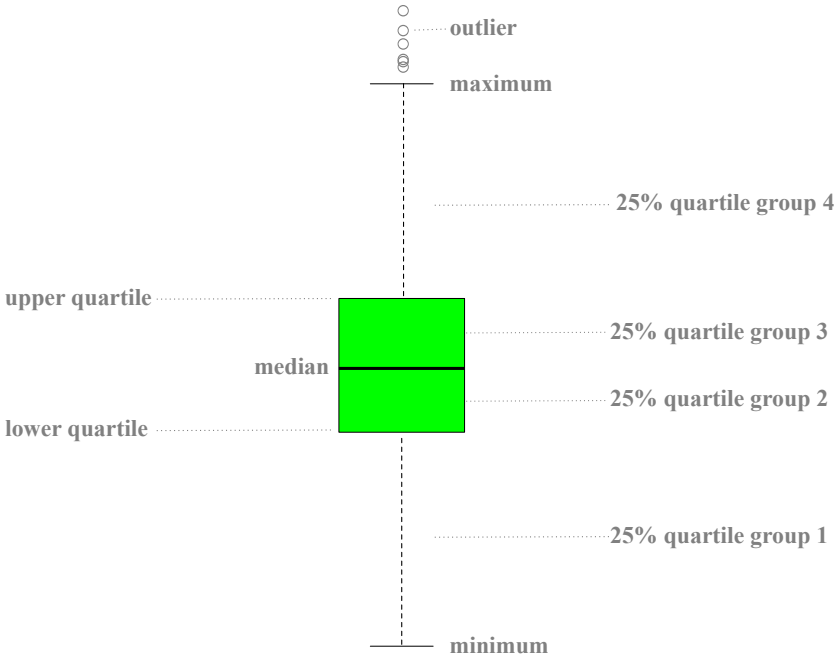

Microarray Data Analysis Workflow
Experimental Design
Biological question
Tissue/sample preparation
RNA isolation
mRNA isolation & labeling
Microarray hybridisation & scanning
Data analysis
Quality control
Quantification & normalization
Statistical filtering
Clustering analysis
List of significant or co-expressed genes
Data interpretation & visualization
Diagrams/Venn-diagrams ...
Pathway analysis
Gene Set Enrichment Analysis (GSEA)
Functional Enrich-ment Analysis

Quality control (QC)

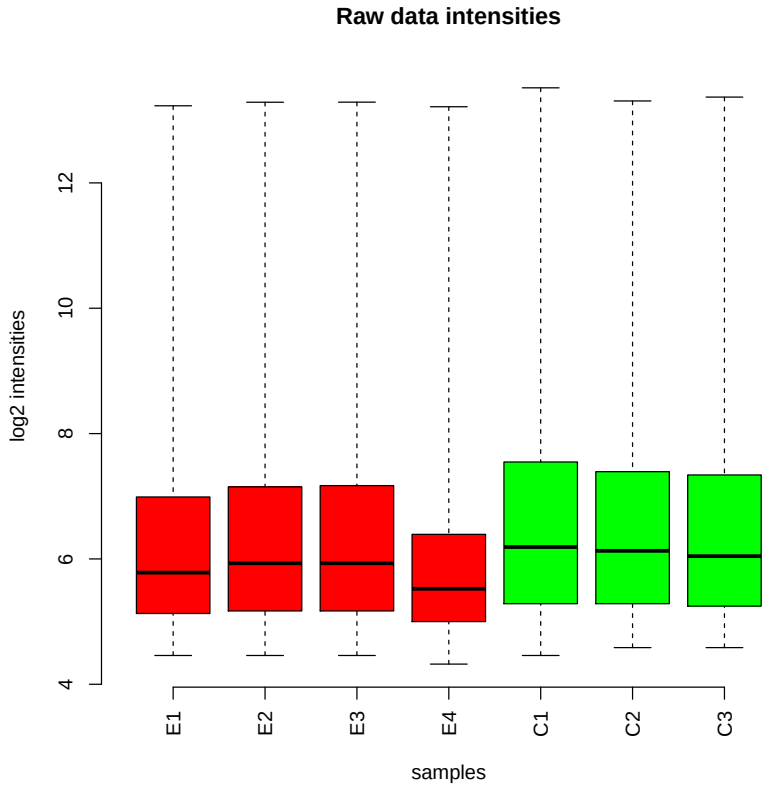
- is crucial to the reliability of the findings concluded from the analysis
- exclude samples of low quality
- several QC methods focussing on different aspects of the data
  - signal distribution $\rightarrow$ histogram, boxplots
  - array comparisons $\rightarrow$ PCA plots, array-array intensity correlation, hierarchical clustering

Intensity distribution: boxplots

- is a graphical representation of statistical measures

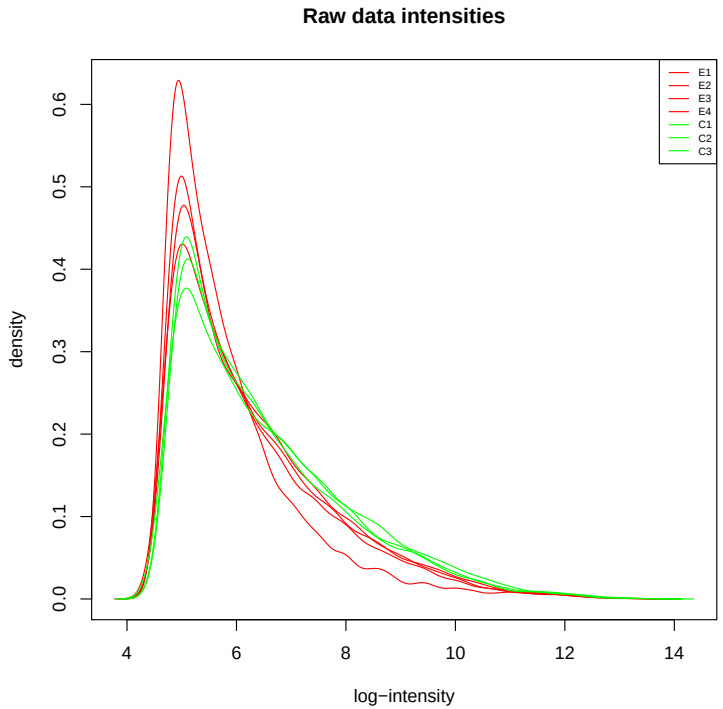


- used to detect major variations between the arrays



Density plot

- a smoothed histogram $\rightarrow$ shows intensity distribution of each array
- identifies arrays that need to be carefully examined before using in further analysis



Hierarchical clustering

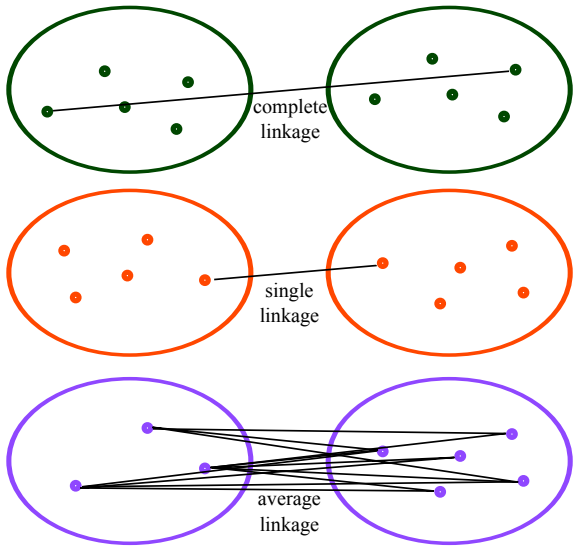
- widely used data analysis tool to combine or identify objects that are close or similar to each other into clusters
- cluster $\rightarrow$ a collection of data objects similar to one another within the same cluster & disimmilar to the objects in the other clusters
- idea $\rightarrow$ build a binary tree (dendrogram) of the data that successively merges similar groups of points
- requires measures of similarity/distances between individual points and groups of data points

Distance metrics

- distances are a numerical description of how far apart objects are
- choosing the right distance measure is a critical step in clustering
- examples:
  - Euclidean distance
  - Manhattan distance
  - $1 -$ correlation

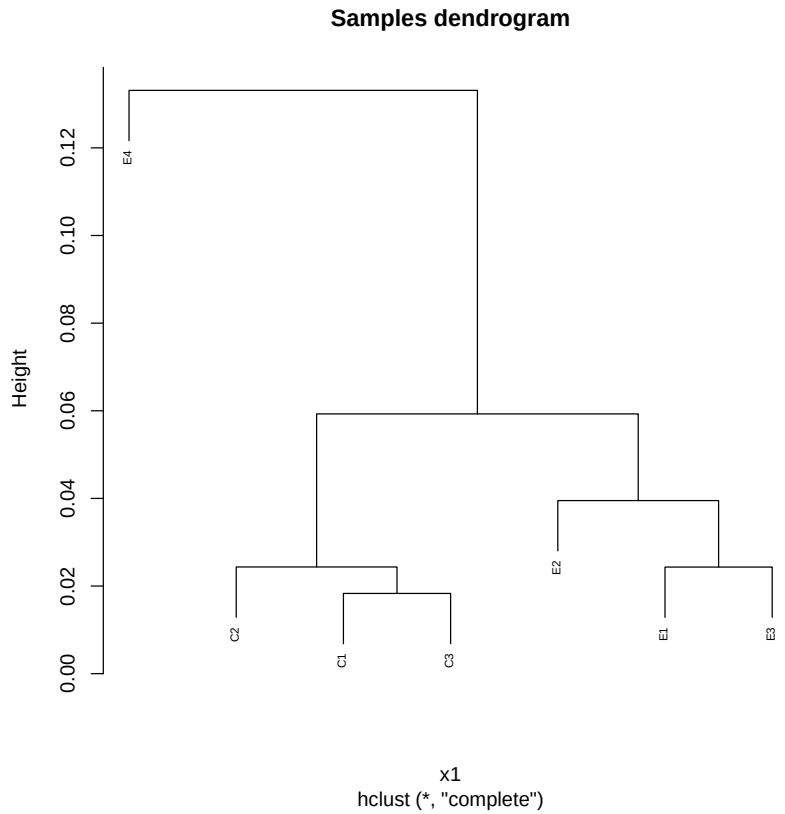
Group similarity

- is a measure of strength of relationship between two objects
- several methods to define intergroup similarity:
  1. single-linkage → is the minimum distance between any 2 objects, one from each cluster
  2. complete-linkage → is the maximum distance between 2 objects, one from each cluster
  3. average-linkage → is the average of all pairwise distances between the members of both clusters



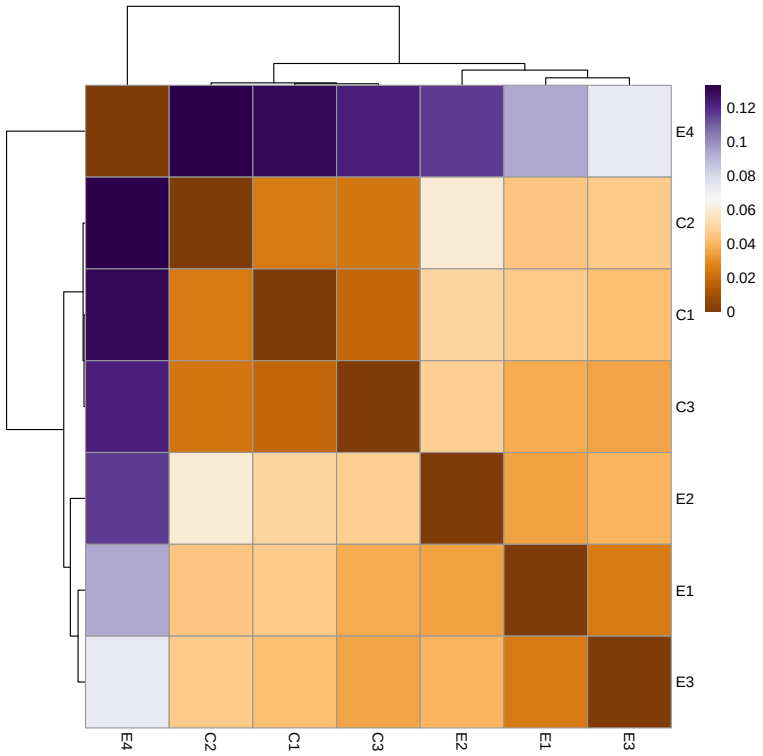
Dendrogram

- a tree that defines the relationships between objects and the distance between clusters



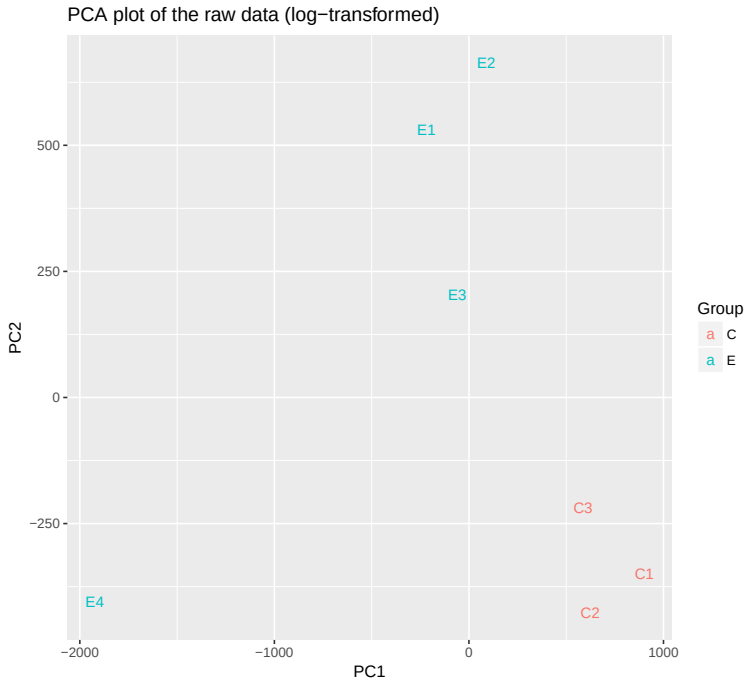
Array-array intensity correlation

- displays the correlation of all pairwise samples using heatmaps



Principle component analysis

- transforms the data from a high-dimensional space into a 2 or 3-dimensional one without losing much of the variation in the original values
- dimenisonality reduction allows visual inspection of the data
- idea $\rightarrow$ samples with similar intensities should cluster together



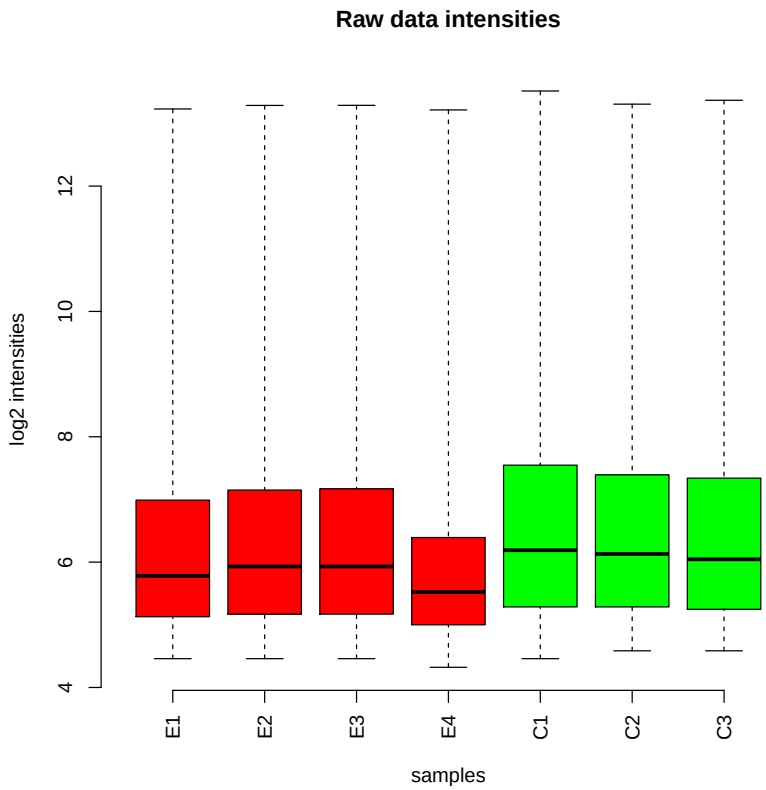
Normalisation

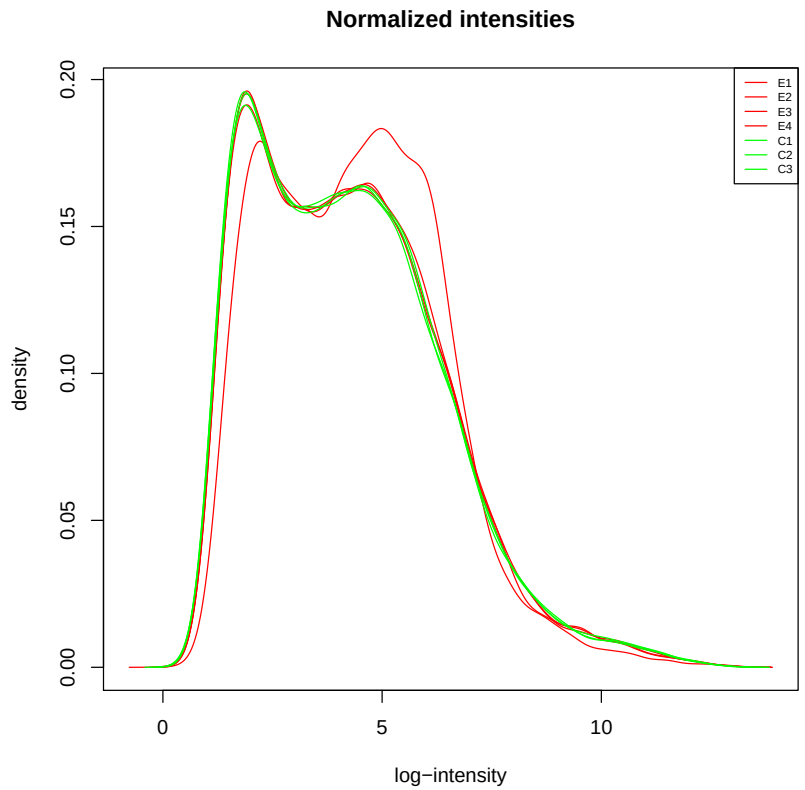
- compensates for systematic technical differences between chips
  - unequal quantities of starting RNA
  - different amounts of labelling
  - varied hybrdisation conditions across the physical extent of one array
  - different scanner settings
- normalisation techniques:
  - scale normalisation
  - lowess normalisation
  - MAS 5.0
  - RMA (quantile normalisation)

Robust multiarray average (RMA)

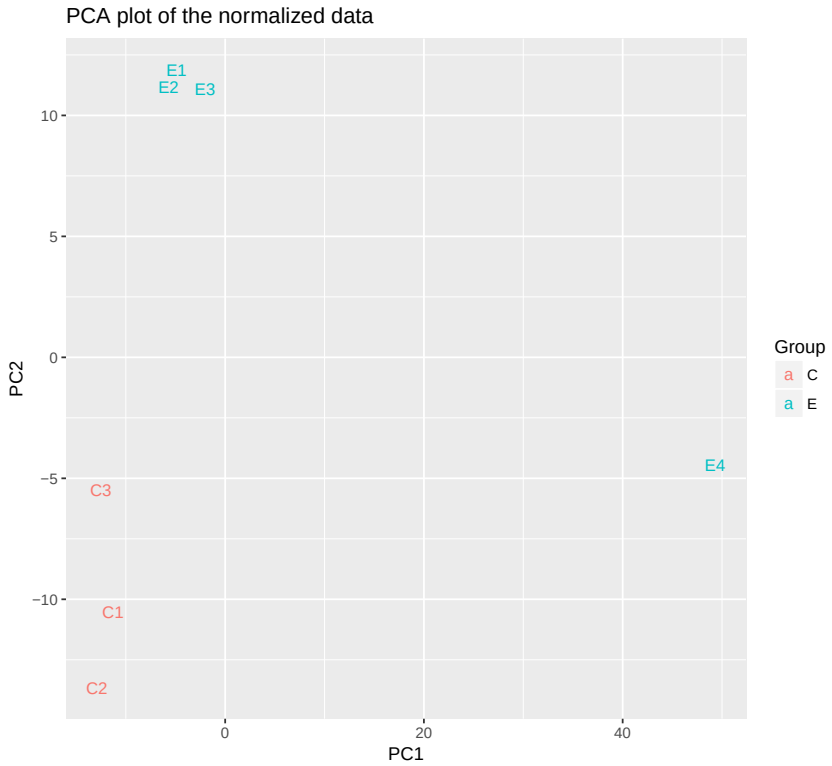
- one of the widest used normalisation methods for Affymetrix arrays
- uses only the PM probes on the chip
- normalizing at probe level avoids the loss of information
- 4 steps:
  - background correction
  - quantile normalisation
  - probe level intensity calculation
  - probe set summarization ('median polishing')

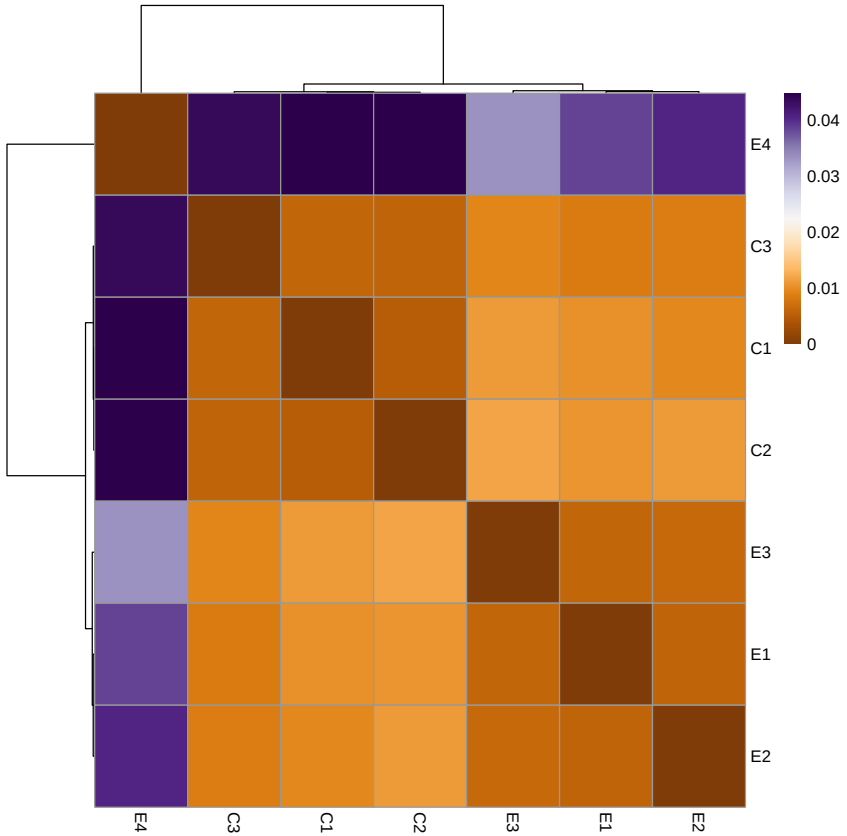
QC after normalisation (I)





QC after normalisation (II)





Identification of differentially expressed genes

- identify DE genes and apply statistical tests to assess the significance of the observed associations
- statistical tests:
  - comparison of two conditions: *Student's t-test* or *Wilcoxon rank sum test*
  - multiple/nested conditions: *ANOVA*
  - Linear models for microarray data (*LIMMA*)
- multiple testing $\rightarrow$ p-value adjustments (*FDR, FWER*)

Identification of differentially expressed genes

- identify DE genes and apply statistical tests to assess the significance of the observed associations
- statistical tests:
  - comparison of two conditions: *Student's t-test* or *Wilcoxon rank sum test*
  - multiple/nested conditions: *ANOVA*
  - Linear models for microarray data (*LIMMA*)
- multiple testing $\rightarrow$ p-value adjustments (*FDR, FWER*)
- statistical significance is not necessarily the same as biological significance

Linear models for microarray data (LIMMA)

- is an R package designed to analyse complex experiments involving comparisons between many samples simultaneously
- operates on a matrix of expression values
- it allows different levels of variability between genes and between samples

Linear models for microarray data (LIMMA)

- is an R package designed to analyse complex experiments involving comparisons between many samples simultaneously
- operates on a matrix of expression values
- it allows different levels of variability between genes and between samples
- simplified approach:
  - construct a linear model to describe the relation between observed values and experimental conditions
  - fit the linear model to each row of data to estimate the fold changes
  - apply empirical Bayes to calculate moderated t-statistics
  - output: moderated t-statistics

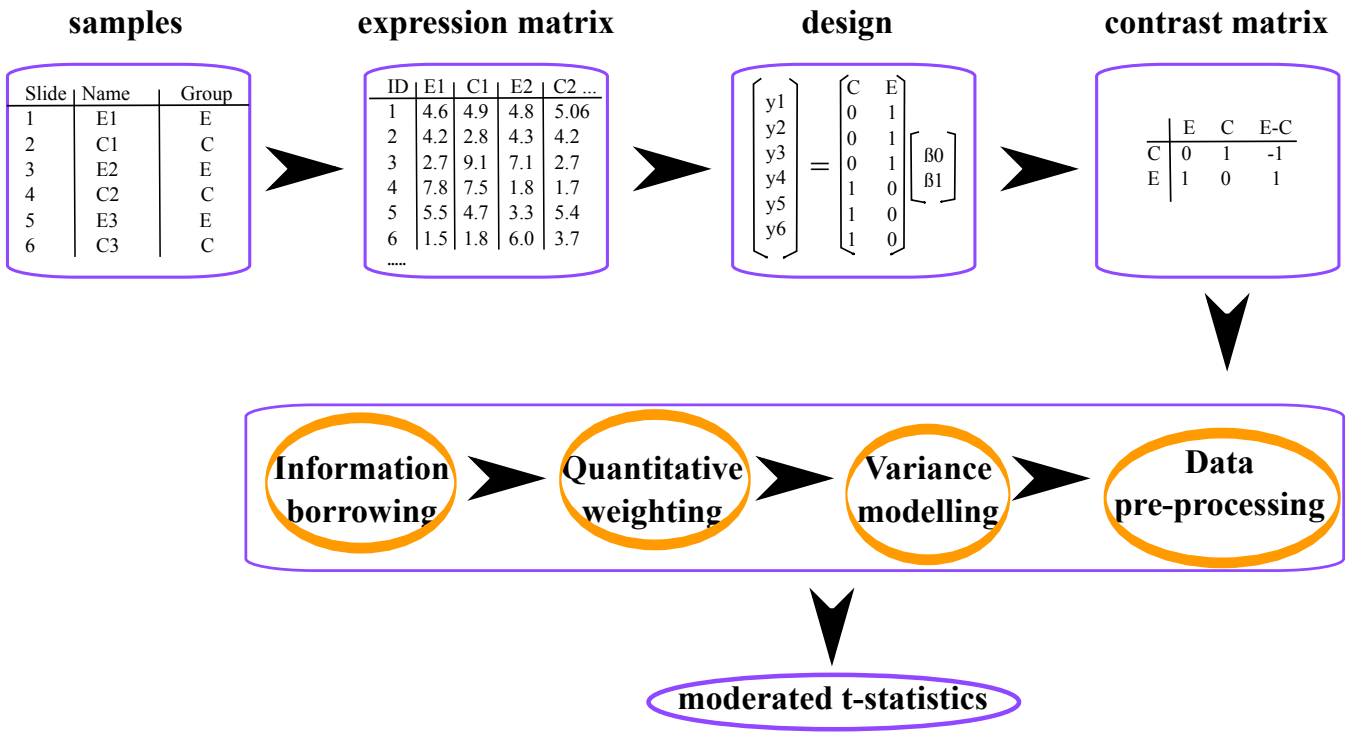
Linear models

- most used statistical methods
- can be used to compare 2 or more groups and for multifactorial designs
- requires a design matrix and a contrast matrix
  - design matrix $\rightarrow$ states which samples are allocated to which conditions
  - contrast matrix $\rightarrow$ describes which comparisons are of interest
- $y = X\beta + \epsilon$

y: vector of observed data, X: design matrix, β: vector of parameters to estimate

Example

- 2x3 factorial design: 2 conditions, each replicated three times
- goal: find differentially expressed genes between 2 conditions

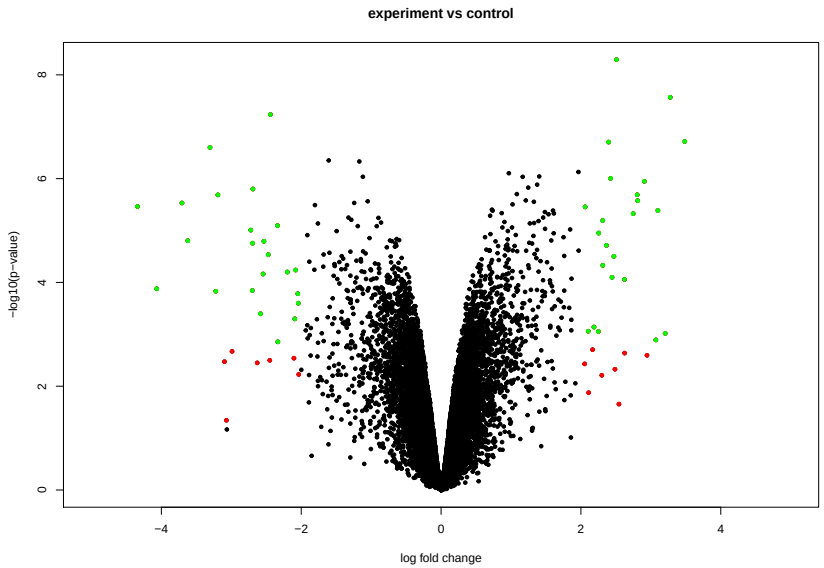


List of significantly differentially expressed genes

PROBEID	ENTREZID	SYMBOL	GENENAME	logFC	AveExpr	t	P.Value	adj.P.Val	B	E1	E2	E3	C1	C2	C3
TC01000017.hg.1	400728	FAM87B	family with sequence similarity 87 member B	-0.5108676	5.282430	-5.830297	7.381490e-04	0.0334267866	-0.057279721	5.082846	5.092984	4.905159	5.502273	5.673162	5.438157
TC01000113.hg.1	9563	H6PD	hexose-6-phosphate dehydrogenase (glucose 1-deh...	-0.4735258	7.655717	-5.053993	1.648450e-03	0.0467099019	-0.899082448	7.502889	7.309566	7.444406	7.936681	8.012458	7.728300
TC01000151.hg.1	5351	PLOD1	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 1	-0.5105864	6.892207	-5.476685	1.054191e-03	0.0386410376	-0.429817745	6.790599	6.673173	6.446968	7.162915	7.158692	7.120893
TC01000184.hg.1	23254	KAZN	kazrin, periplakin interacting protein	0.3506924	6.551819	5.353727	1.197689e-03	0.0409859801	-0.563584091	6.746468	6.774337	6.660690	6.430708	6.293886	6.404824
TC01000193.hg.1	6248	RSC1A1	regulatory solute carrier protein, family 1, member 1	-0.2682336	6.626843	-5.023315	1.704401e-03	0.0477554758	-0.934185613	6.500803	6.488452	6.488923	6.745079	6.733196	6.804605
TC01000371.hg.1	677774	SCARNA1	small Cajal body-specific RNA 1	-1.2962790	3.983974	-10.113995	2.556184e-05	0.0078580955	3.314046365	3.534069	3.035438	3.437998	4.641666	4.637176	4.617499
TC01000408.hg.1	347735	SERINC2	serine incorporator 2	-0.3554747	6.278629	-5.709592	8.321886e-04	0.0353095024	-0.182447479	6.115187	6.130025	6.057464	6.513044	6.481094	6.374961
TC01000424.hg.1	8668	EIF3I	eukaryotic translation initiation factor 3 subunit I	-0.3568492	8.093274	-5.177178	1.443535e-03	0.0445867614	-0.759565140	7.930683	7.993479	7.820386	8.322752	8.207641	8.284703
TC01000497.hg.1	64064	OXCT2	3-oxoacid CoA-transferase 2	-0.3896637	4.672660	-5.610016	9.199577e-04	0.0366517120	-0.287254955	4.544900	4.477226	4.411360	4.849566	4.794054	4.958857
TC01000536.hg.1	339539	LOC339539	uncharacterized LOC339539	0.6883077	5.218392	10.184796	2.445011e-05	0.0077400436	3.355994713	5.618360	5.462320	5.606956	4.822785	4.874720	4.925208
TC01000550.hg.1	6487	ST3GAL3	ST3 beta-galactoside alpha-2,3-sialyltransferase 3	0.3884467	5.964524	5.974156	6.413111e-04	0.0316322927	0.089253120	6.145311	6.078243	6.252688	5.746718	5.764427	5.799757
TC01000564.hg.1	6202	RPS8	ribosomal protein S8	-0.3657236	6.695883	-6.843642	2.879934e-04	0.0212561857	0.916921727	6.520853	6.478339	6.539872	6.870361	6.860791	6.905082
TC01000565.hg.1	94161	SNORD46	small nucleolar RNA, C/D box 46	-0.4804855	11.325241	-7.370833	1.840225e-04	0.0177412913	1.373998443	11.117220	11.003706	11.134068	11.518699	11.545139	11.632612
TC01000668.hg.1	26027	ACOT11	acyl-CoA thioesterase 11	0.4607009	5.485165	8.153243	9.907859e-05	0.0139638874	1.996975995	5.691272	5.727595	5.727680	5.266671	5.300785	5.196988
TC01000730.hg.1	3953	LEPR	leptin receptor	-0.2800242	5.489537	-5.210109	1.393693e-03	0.0439250668	-0.722656530	5.326151	5.383075	5.339350	5.631077	5.657286	5.600285
TC01000745.hg.1	1647	GADD45A	growth arrest and DNA damage inducible alpha	0.3319442	6.718726	5.824295	7.425322e-04	0.0334267866	-0.063455075	6.903872	6.926172	6.824051	6.572825	6.524426	6.561011
TC01000781.hg.1	256435	ST6GALNAC3	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-...	-0.7128701	3.674617	-6.640652	3.446840e-04	0.0231384560	0.732236807	3.521368	3.242929	3.190250	4.139107	3.906683	4.047368
TC01000874.hg.1	729970	LOC729970	hCG2028352-like	-0.5080706	5.747263	-5.653954	8.800092e-04	0.0358800205	-0.240834250	5.326493	5.645737	5.507452	5.951125	6.046520	6.006250
TC01000883.hg.1	58155	PTBP2	polypyrimidine tract binding protein 2	0.7157137	5.906043	9.355395	4.191649e-05	0.0101177096	2.841605787	6.163183	6.383313	6.245203	5.617368	5.540461	5.486728
TC01000892.hg.1	54873	PALMD	palmdelphin	-1.5961253	7.425419	-10.037111	2.683487e-05	0.0081723454	3.268093656	6.877048	6.704516	6.300505	8.343710	8.292006	8.034728
TC01000909.hg.1	1901	S1PR1	sphingosine-1-phosphate receptor 1	-0.7190823	6.329539	-5.878584	7.039221e-04	0.0326016120	-0.007776798	5.983407	5.941898	5.984688	6.452707	6.664954	6.949578
TC01001004.hg.1	56944	OLFML3	olfactomedin like 3	-0.3748706	6.466112	-5.435179	1.100358e-03	0.0393837151	-0.474724388	6.322097	6.200433	6.313499	6.570661	6.732215	6.657765
TC01001026.hg.1	8458	TTF2	transcription termination factor, RNA polymerase II	0.3761420	5.344338	5.353874	1.197504e-03	0.0409859801	-0.563422243	5.473276	5.511736	5.612214	5.220922	5.062530	5.185349
TC01001030.hg.1	54855	FAM46C	family with sequence similarity 46 member C	0.8025345	5.933439	7.559013	1.578103e-04	0.0167272092	1.529641277	6.425485	6.324733	6.253900	5.664675	5.312671	5.619168
TC01001053.hg.1	647121	EMBP1	embigin pseudogene 1	-0.9099513	4.783384	-5.778300	7.771126e-04	0.0343129955	-0.110949190	4.425107	4.577769	3.982349	5.118014	5.375836	5.221229
TC01001201.hg.1	10962	MLLT11	myeloid/lymphoid or mixed-lineage leukemia; transl...	-0.9376877	4.805903	-5.914541	6.795930e-04	0.0324783243	0.028875887	4.440566	4.448112	4.122501	4.956562	5.372954	5.494726
TC01001262.hg.1	4881	NPR1	natriuretic peptide receptor 1	0.4207164	6.176434	5.328641	1.229578e-03	0.0414553795	-0.591148043	6.452776	6.303448	6.404152	6.060845	5.996651	5.840730
TC01001274.hg.1	57198	ATP8B2	ATPase phospholipid transporting 8B2	0.4062779	6.365881	5.245222	1.342654e-03	0.0433614839	-0.683480409	6.688421	6.477253	6.541387	6.219719	6.191483	6.077025
TC01001348.hg.1	3428	IFI16	interferon, gamma-inducible protein 16	-0.7519609	6.766634	-7.434972	1.745707e-04	0.0173431723	1.427477250	6.290559	6.514843	6.366559	6.986561	7.133814	7.307469
TC01001377.hg.1	4817	NIT1	nitrilase 1	0.5312612	6.647760	6.077412	5.806069e-04	0.0302714482	0.192685120	6.760562	6.950099	7.029510	6.387677	6.298021	6.460690
TC01001422.hg.1	6675	UAP1	UDP-N-acetylglucosamine pyrophosphorylase 1	-0.4832461	6.574137	-5.148068	1.489258e-03	0.0451432083	-0.792327520	6.440064	6.268701	6.288776	6.908747	6.641255	6.897278
TC01001423.hg.1	4921	DDR2	discoidin domain receptor tyrosine kinase 2	-0.8498982	8.758723	-7.295497	1.958727e-04	0.0181157270	1.310608299	8.573590	8.318348	8.109385	9.167569	9.153386	9.230063
TC01001425.hg.1	4921	DDR2	discoidin domain receptor tyrosine kinase 2	-0.8240146	8.148635	-6.678313	3.332765e-04	0.0227676052	0.766880315	7.761718	7.793431	7.654735	8.530624	8.334582	8.816721
TC01001464.hg.1	100313835	MIR1255B2	microRNA 1255b-2	0.7645595	3.698043	7.833916	1.267770e-04	0.0155307630	1.750284059	3.980791	4.022401	4.237778	3.443932	3.189162	3.314197
TC01001484.hg.1	5396	PRRX1	paired related homeobox 1	0.9293720	5.456657	10.533458	1.971960e-05	0.0072212701	3.557543997	5.864014	5.815694	6.084322	4.915966	4.989748	5.070200

$\rightarrow$ easy to sort and filter out significantly differentially expressed genes: adj.P.Val < 0.05 and logFC $> \pm 2$

Volcano plot



p-value < 0.05 & logFC $> \pm 2$
adj.P.Val < 0.05 & logFC $> \pm 2$

- widely used visualisation technique to inspect the result of the statistical analysis
- large difference in expression $\rightarrow$ the more extreme the points will lie on the x-axis
- significant difference $\rightarrow$ the smaller the p-value & the higher the -$\log_{10}$(p-value)

Software - data analysis

- R & Bioconductor
  - *agilp* (Agilent Expression Arrays)
  - *oligo, exonmap* (Gene/Exon ST Arrays)
  - *affy* (3'biased Arrays)
  - *lumi, beadarray* (Illumina Expression Arrays)
  - *limma*
- Affymetrix Transcriptome Analysis Console (TAC)
- GeneSpring GX (Agilent, Affymetrix, Illumina arrays)

Thank you for your attention!