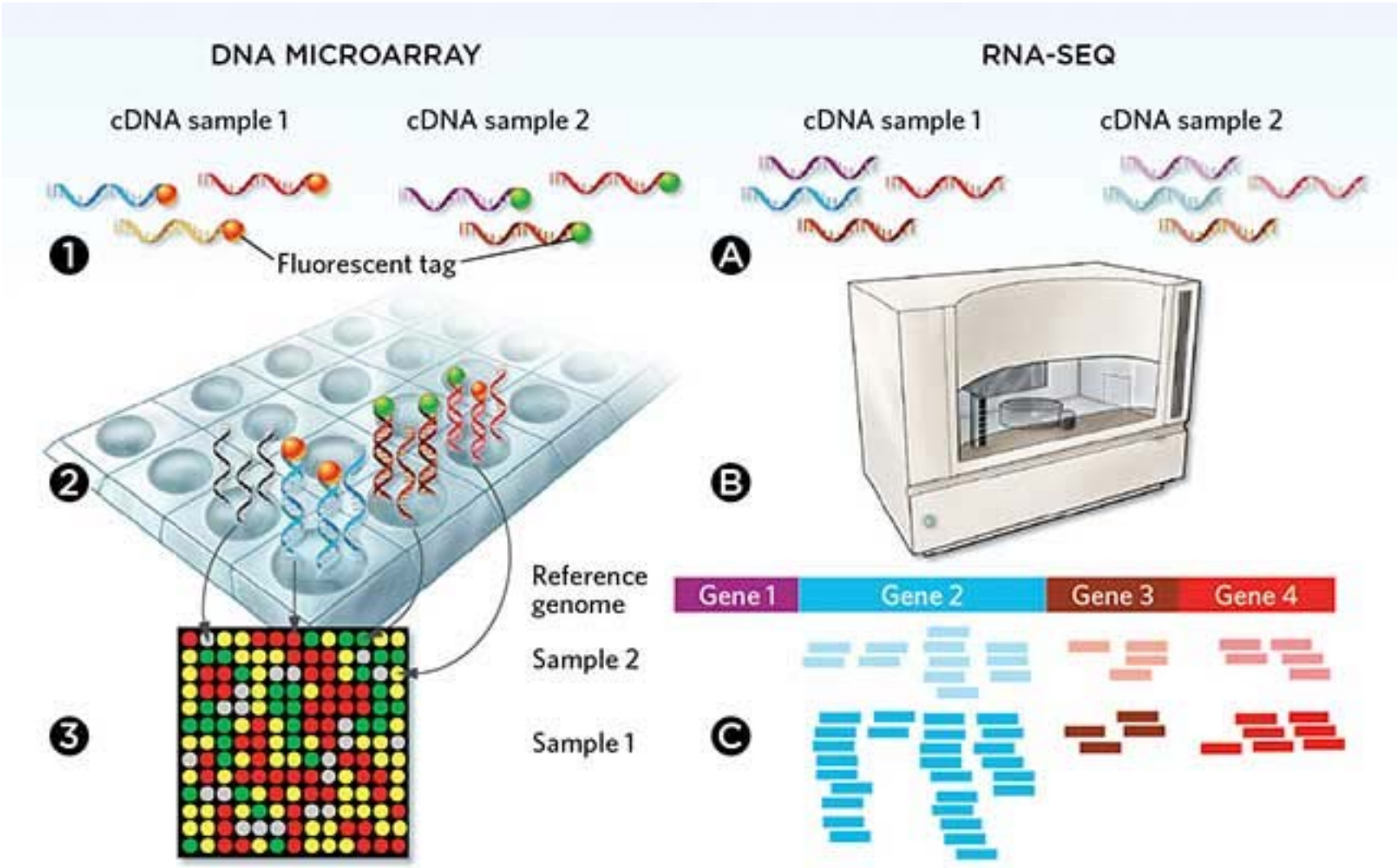
PIBI:

Differential Gene Expression Analysis

Microarrays vs. RNA-seq



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Our data set: microarray

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Generated on Affymetrix Mouse Gene 1.0 ST platform												
2								Fold	Fold	Fold			
3	Original_O	MATCH_ORDER	Gene	Probeset	Gene	Protein		Range CD	Range HFD	Range_All	C57BL/6J_Liver_CD	DBA/2J_Liver_CD	BXD43_Liver_CD
4	12643		Igh	10403036	Igh	0		1,50	1,72	1,72	5,138	5,296	5,061
5	12646		Igh	10403043	Igh	0		2,94	1,99	3,60	7,983	7,529	7,645
6	12653		Igh	10403063	Igh	0		3,72	1,86	4,26	7,169	7,023	7,083
7	8759	1458	Hcfc2	10365242	Hcfc2	A0AUN4		1,48	1,47	1,54	9,241	8,99	8,967
8	18740	8191	Ccdc112	10458794	Ccdc112	A0AUP1		1,83	1,94	2,00	6,761	6,656	6,729
9	16962		EG635895	10441899	EG635895	A0AUV4		1,94	1,70	2,17	6,689	6,456	6,528
10	25569	12857	Srp72	10522676	Srp72	A0JLN1		1,58	1,42	1,58	11,442	11,313	11,359
11	29717	15643	Ceacam15	10560294	Ceacam15	A0JLX4		1,77	1,79	1,84	6,464	6,546	6,57
12	21302	10042	4930578N16	10483536	4930578N16	A0JLY1		1,53	1,43	1,73	6,763	6,444	6,701
13	26618	13525	Ccdc64	10533007	Ccdc64	A0JNT9		1,66	1,81	1,81	7,525	7,747	7,819
14	12174	3901	Aspg	10398795	Aspg	A0JNU3		2,83	2,00	2,86	11,036	10,897	10,361
15	11143	3199	Efcab5	10388625	Efcab5	A0JP43		1,57	1,48	1,57	5,682	6,019	5,869
16	23261		Bclp2	10501007	Bclp2	A0JP54		1,61	1,44	1,64	5,902	6,109	6,33
17	22458	10842	Sprr2e	10493867	Sprr2e	A0PJN0		2,30	2,23	2,36	6,463	6,463	6,027
18	13599		3110006E14	10410460	3110006E14	A0PJN4		2,34	2,31	2,41	7,143	7,618	7,502
19	11368	3370	Krtap2-4	10390911	Krtap2-4	A0PK51		1,98	1,84	2,00	8,862	9,531	9,336
20	34031	18366	Olfr239	10598105	Olfr239	A0PK55		2,42	1,71	2,42	8,869	9,153	9,144
21	19543	8787	Olfr1436	10466298	Olfr1436	A0PK57		2,31	1,97	2,40	5,503	5,333	5,246
22	12493	4115	Zdhhc22	10401702	Zdhhc22	A0PK84		2,02	1,92	2,02	6,167	6,536	6,098
23	12494		Zdhhc22	10401705	Zdhhc22	A0PK84		1,94	1,72	2,08	7,594	7,667	7,792
24	24060	11857	Oscp1	10508099	Oscp1	A0ZV96		1,79	1,43	2,02	6,969	7,283	7,358
25	22977	11186	Veph1	10498547	Veph1	A1A535		1,53	1,42	1,53	6,895	7,088	6,995
26	31360	16620	Isx	10572865	Isx	A1A546		1,78	1,76	1,94	6,937	7,337	7,264
27	22452	10839	Pglyrp3	10493842	Pglyrp3	A1A547		1,46	1,55	1,55	6,551	6,669	6,62
28	27988	14437	Tcf3	10545458	Tcf3	A1A549		1,59	1,41	1,64	8,475	8,601	8,491
29	34724	18750	Tbc1d25	10603478	Tbc1d25	A1A5B6		1,84	1,62	1,90	8,255	8,109	8,212
30	25590	12871	Odam	10522875	Odam	A1E960		1,33	1,31	1,58	5,295	5,329	5,511
31	12381	4033	Fscb	10400564	Fscb	A1EGX6		1,58	1,82	1,82	6,63	6,526	6,52
32	18783		4933429F08	10459210	4933429F08	A1IGU4		2,62	2,68	3,11	8,131	7,427	7,606
33	7103	336	Faim3	10349593	Faim3	A1KXC4		1,70	1,94	1,94	7,183	7,518	7,541
34	26741	13622	Srcrb4d	10534537	Srcrb4d	A1L0T3		1,43	1,51	1,55	8,652	8,993	8,869
35	15335	5864	Olfr282	10426592	Olfr282	A1L1B4		2,20	1,67	2,22	8,018	8,162	8,189

Data normalization: Why?

Data normalization: Why?

an earthquake

Data normalization: Why?

High-level answer:

Things which do not have anything to do with the condition under study
are different from sample to sample (experiment to experiment)

Classical application of Murphy's law ...

Data normalization: Why?

High-level answer:

Things which do not have anything to do with the condition under study
are different from sample to sample (experiment to experiment)

Low-level examples:

- different amounts of total RNA used for different arrays
- different dye behaviour (for whatever reason)
- if time passed between measurements:
  - room temperature, air humudity, lab staff, reagent batches
- an earthquake

Data normalization: Why?

High-level answer:

Things which do not have anything to do with the condition under study
are different from sample to sample (experiment to experiment)

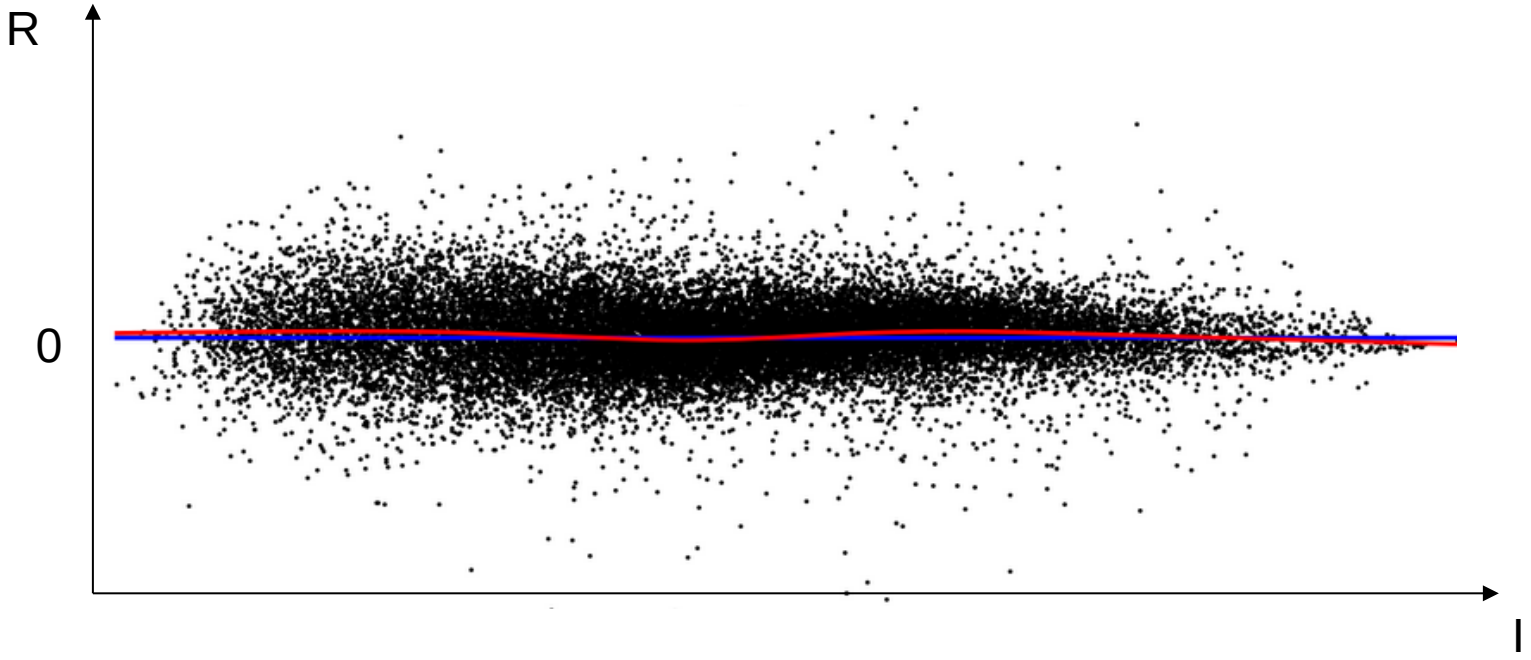
Low-level examples:

- different amounts of total RNA used for different arrays
- different dye behaviour (for whatever reason)
- if time passed between measurements:
  - room temperature, air humudity, lab staff, reagent batches
- an earthquake

http://www.people.vcu.edu/~mreimers/OGMDA/normalize.expression.html

Data normalization: What happens?

Example 1: ratio-intensity plots



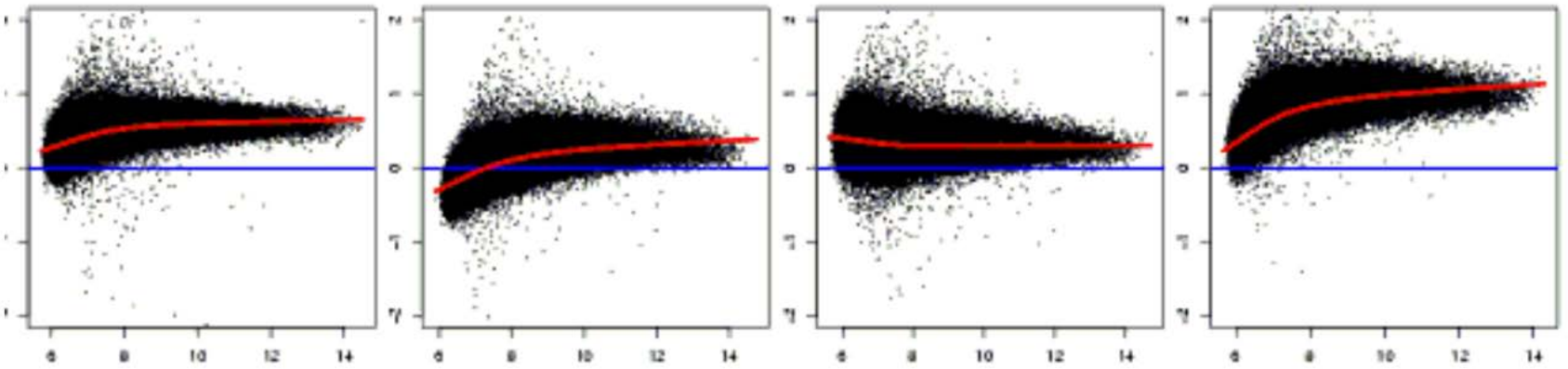
R = log(expression in sample 1 / expression in sample 2)

I = (log(expression in sample 1)
+ log(expression in sample 2)) / 2

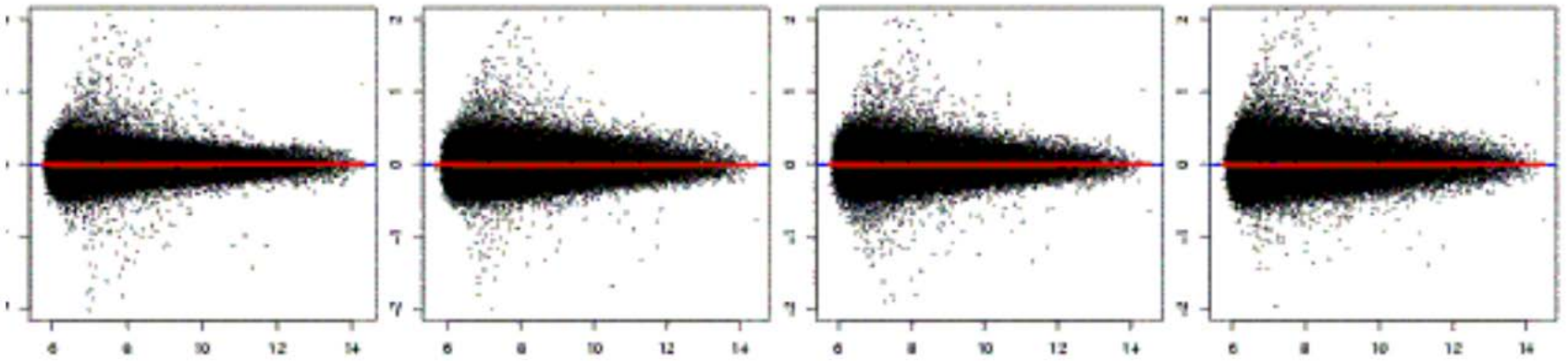
Data normalization: What happens?

Example 1: ratio-intensity plots

Before normalization:

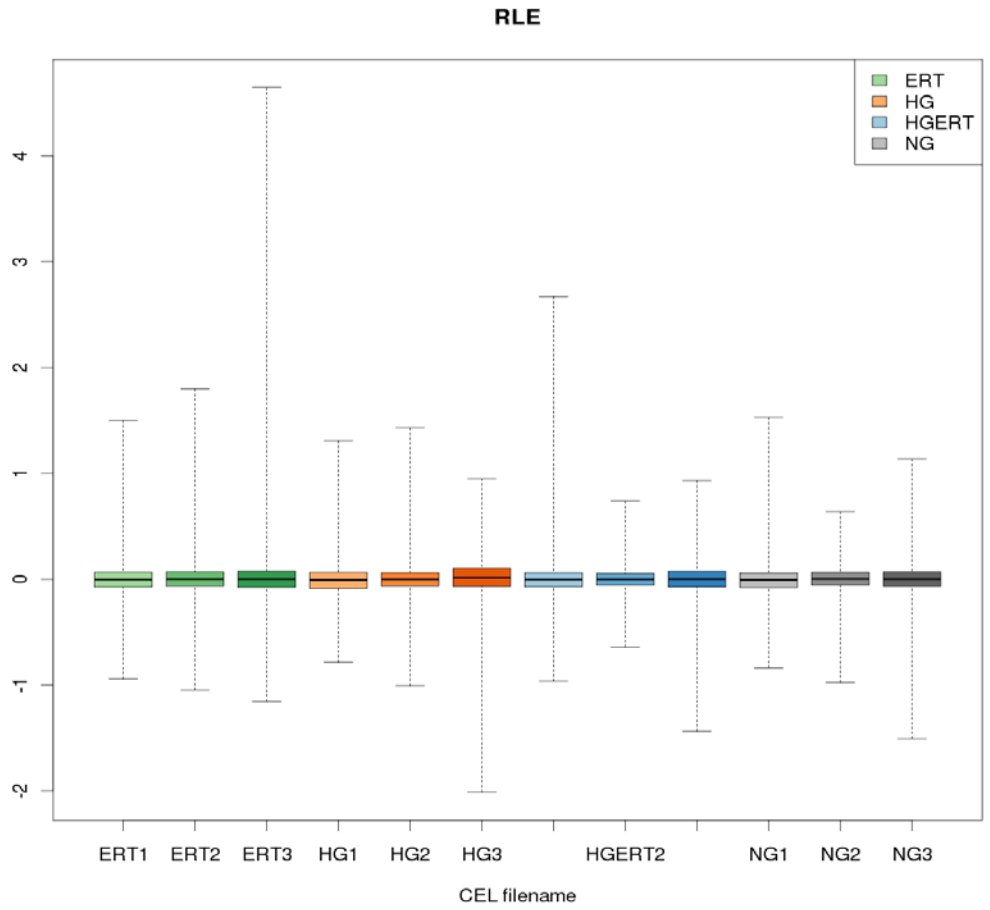


After normalization:



Data normalization: What happens?

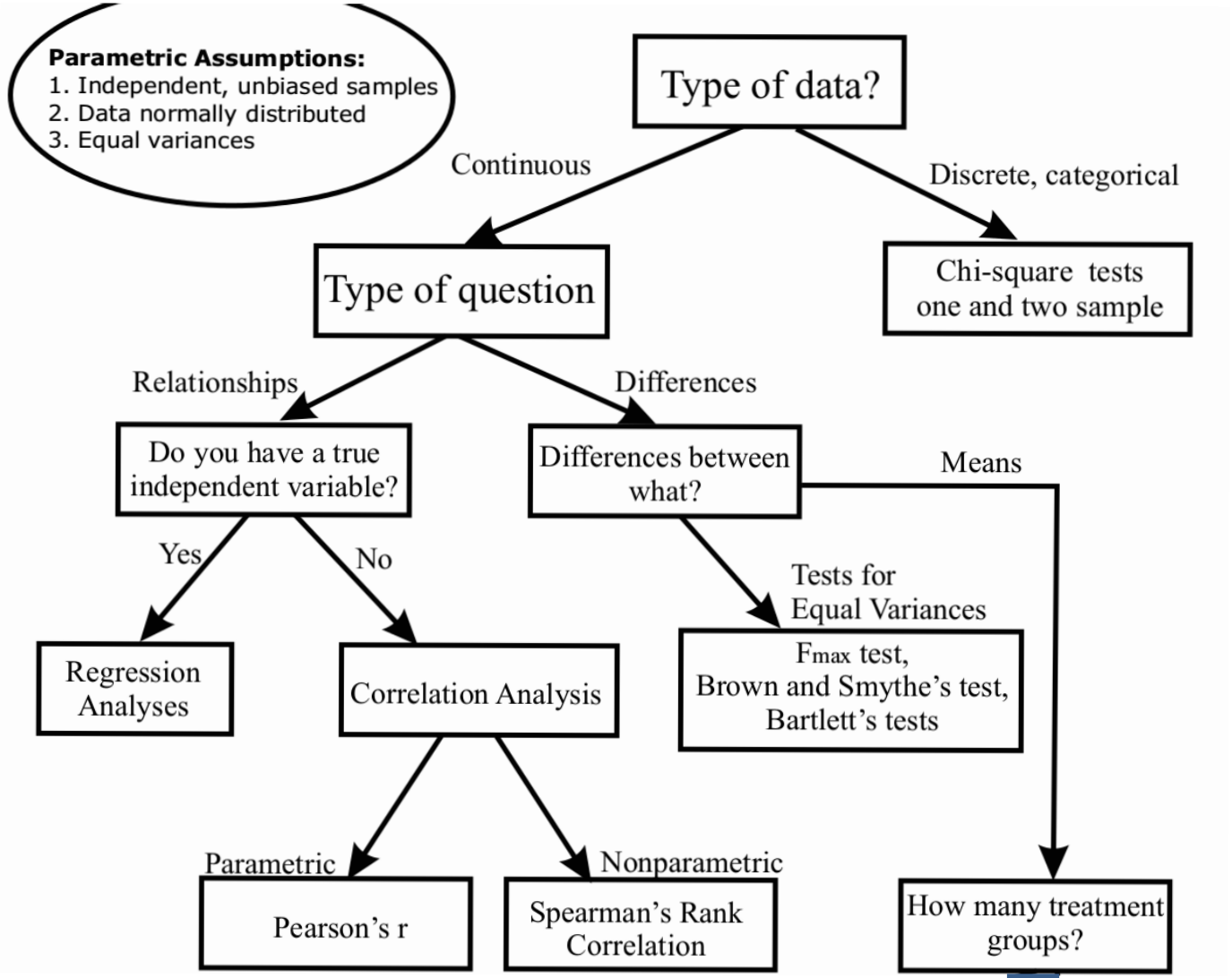
Example 2: relative log expression plot (RLE plot)



- For any gene, calculate the Median across all samples
- For a given sample divide the expression values by the corresponding median
- For every sample plot a box plot

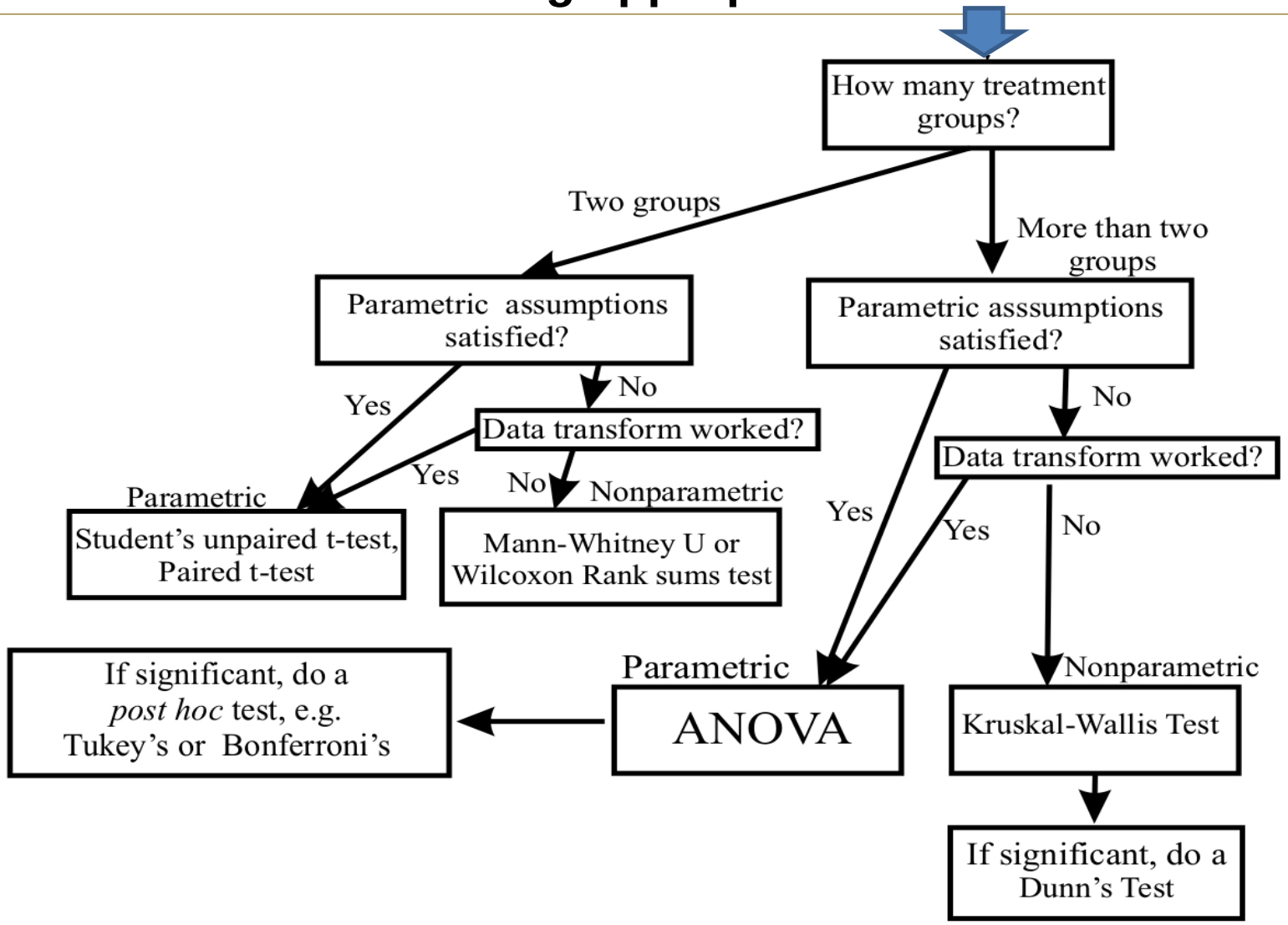
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Choosing appropriate statistical methods



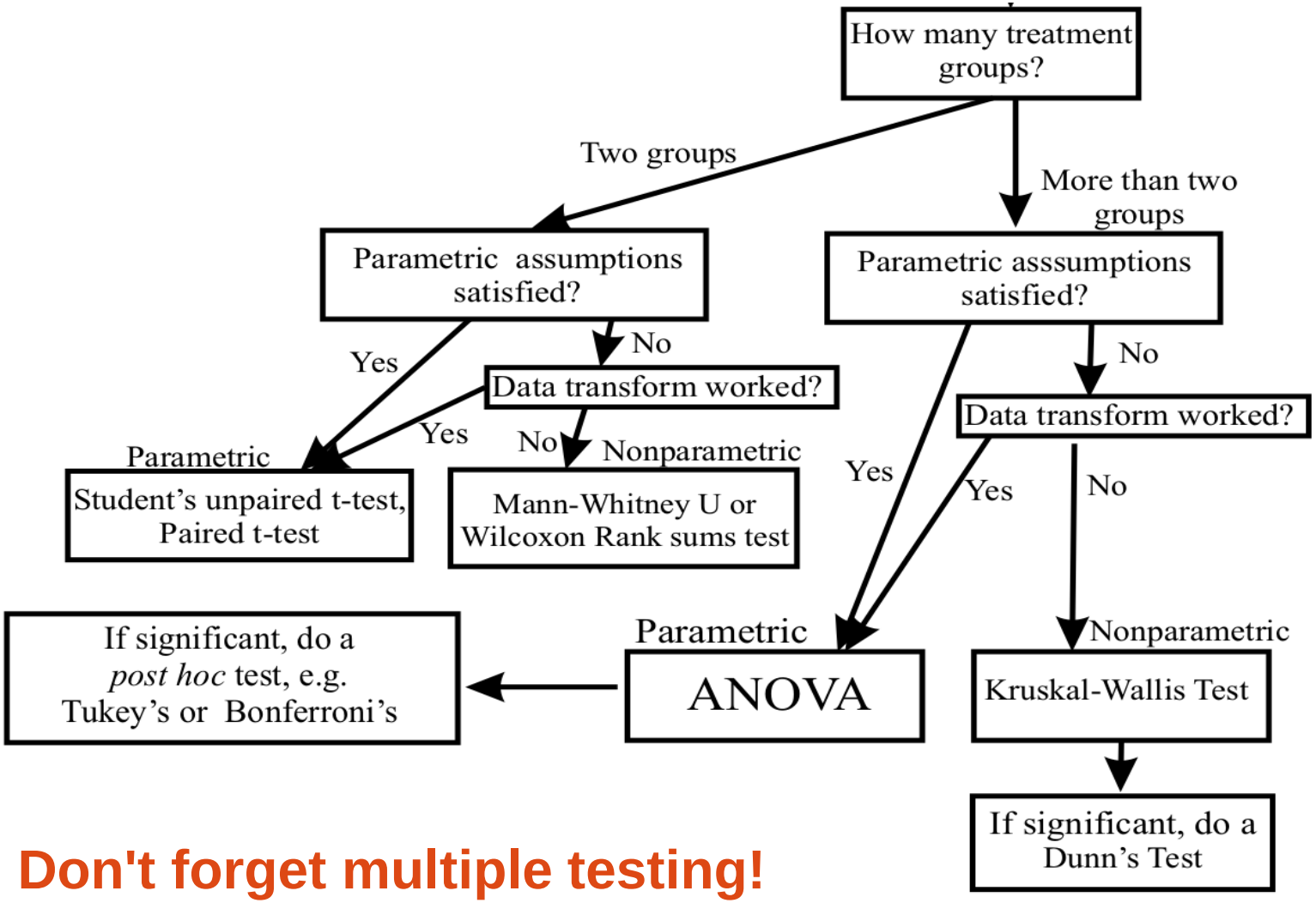
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Choosing appropriate statistical methods



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Choosing appropriate statistical methods



Don't forget multiple testing!

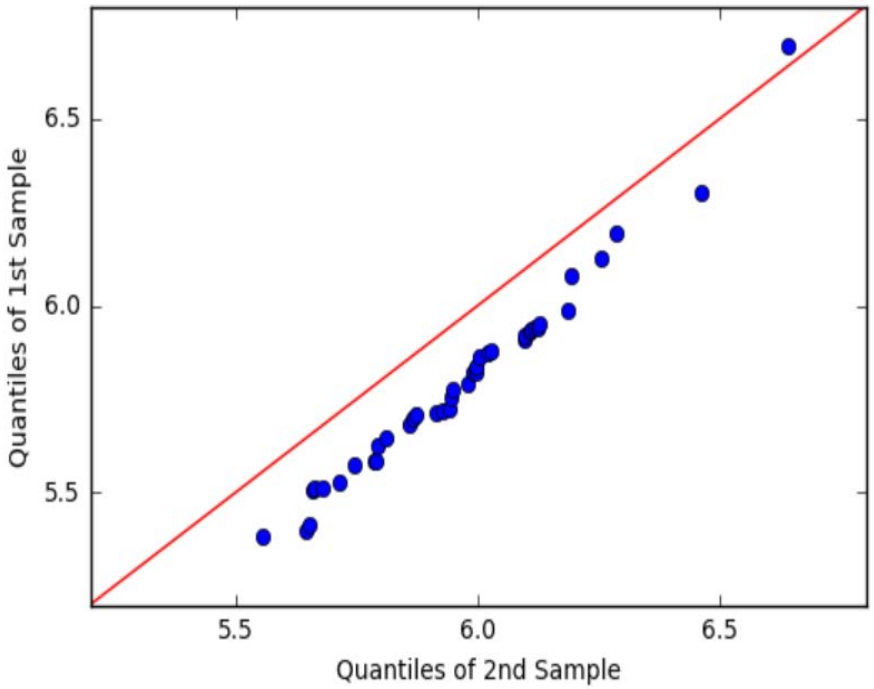
Comparing distributions: Q-Q plots

p-quantile:
value q at which (either theoretically or empirically)
a value drawn from the distribution will be less than q
with probability p

Q-Q plot:
given a point x from the first distribution, calculate the probability p that some other point is less than x.
Then plot x against y, where y is the p-quantile of the second distribution.

Comparing distributions: Q-Q plots

Q-Q plot:
given a point x from the first distribution, calculate the probability p that some other point is less than x.
Then plot x against y, where y is the p-quantile of the second distribution.



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Our data set

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Generated on Affymetrix Mouse Gene 1.0 ST platform												
2								Fold	Fold	Fold			
3	Original_O	MATCH_ORDER	Gene	Probeset	Gene	Protein		Range CD	Range HFD	Range_All	C57BL/6J_Liver_CD	DBA/2J_Liver_CD	BXD43_Liver_CD
4	12643		Igh	10403036	Igh	0		1,50	1,72	1,72	5,138	5,296	5,061
5	12646		Igh	10403043	Igh	0		2,94	1,99	3,60	7,983	7,529	7,645
6	12653		Igh	10403063	Igh	0		3,72	1,86	4,26	7,169	7,023	7,083
7	8759	1458	Hcfc2	10365242	Hcfc2	A0AUN4		1,48	1,47	1,54	9,241	8,99	8,967
8	18740	8191	Ccdc112	10458794	Ccdc112	A0AUP1		1,83	1,94	2,00	6,761	6,656	6,729
9	16962		EG635895	10441899	EG635895	A0AUV4		1,94	1,70	2,17	6,689	6,456	6,528
10	25569	12857	Srp72	10522676	Srp72	A0JLN1		1,58	1,42	1,58	11,442	11,313	11,359
11	29717	15643	Ceacam15	10560294	Ceacam15	A0JLX4		1,77	1,79	1,84	6,464	6,546	6,57
12	21302	10042	4930578N16	10483536	4930578N16	A0JLY1		1,53	1,43	1,73	6,763	6,444	6,701
13	26618	13525	Ccdc64	10533007	Ccdc64	A0JNT9		1,66	1,81	1,81	7,525	7,747	7,819
14	12174	3901	Aspg	10398795	Aspg	A0JNU3		2,83	2,00	2,86	11,036	10,897	10,361
15	11143	3199	Efcab5	10388625	Efcab5	A0JP43		1,57	1,48	1,57	5,682	6,019	5,869
16	23261		Bclp2	10501007	Bclp2	A0JP54		1,61	1,44	1,64	5,902	6,109	6,33
17	22458	10842	Sprr2e	10493867	Sprr2e	A0PJN0		2,30	2,23	2,36	6,463	6,463	6,027
18	13599		3110006E14	10410460	3110006E14	A0PJN4		2,34	2,31	2,41	7,143	7,618	7,502
19	11368	3370	Krtap2-4	10390911	Krtap2-4	A0PK51		1,98	1,84	2,00	8,862	9,531	9,336
20	34031	18366	Olfr239	10598105	Olfr239	A0PK55		2,42	1,71	2,42	8,869	9,153	9,144
21	19543	8787	Olfr1436	10466298	Olfr1436	A0PK57		2,31	1,97	2,40	5,503	5,333	5,246
22	12493	4115	Zdhhc22	10401702	Zdhhc22	A0PK84		2,02	1,92	2,02	6,167	6,536	6,098
23	12494		Zdhhc22	10401705	Zdhhc22	A0PK84		1,94	1,72	2,08	7,594	7,667	7,792
24	24060	11857	Oscp1	10508099	Oscp1	A0ZV96		1,79	1,43	2,02	6,969	7,283	7,358
25	22977	11186	Veph1	10498547	Veph1	A1A535		1,53	1,42	1,53	6,895	7,088	6,995
26	31360	16620	Isx	10572865	Isx	A1A546		1,78	1,76	1,94	6,937	7,337	7,264
27	22452	10839	Pglyrp3	10493842	Pglyrp3	A1A547		1,46	1,55	1,55	6,551	6,669	6,62
28	27988	14437	Tcf3	10545458	Tcf3	A1A549		1,59	1,41	1,64	8,475	8,601	8,491
29	34724	18750	Tbc1d25	10603478	Tbc1d25	A1A5B6		1,84	1,62	1,90	8,255	8,109	8,212
30	25590	12871	Odam	10522875	Odam	A1E960		1,33	1,31	1,58	5,295	5,329	5,511
31	12381	4033	Fscb	10400564	Fscb	A1EGX6		1,58	1,82	1,82	6,63	6,526	6,52
32	18783		4933429F08	10459210	4933429F08	A1IGU4		2,62	2,68	3,11	8,131	7,427	7,606
33	7103	336	Faim3	10349593	Faim3	A1KXC4		1,70	1,94	1,94	7,183	7,518	7,541
34	26741	13622	Srcrb4d	10534537	Srcrb4d	A1L0T3		1,43	1,51	1,55	8,652	8,993	8,869
35	15335	5864	Olfr282	10426592	Olfr282	A1L1B4		2,20	1,67	2,22	8,018	8,162	8,189

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Our data set

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Generated on Affymetrix Mouse Gene 1.0 ST platform												
2													
3	Original_O	MATCH_ORDER	Gene	Probeset	Gene	Protein		Fold Range CD	Fold Range HFD	Fold Range_All	C57BL/6J_Liver_CD	DBA/2J_Liver_CD	BXD43_Liver_CD
4	12643		Igh	10403036	Igh	0		1,50	1,72	1,72	5,138	5,296	5,061
5	12646		Igh	10403043	Igh	0		2,94	1,99	3,60	7,983	7,529	7,645
6	12653		Igh	10403063	Igh	0		3,72	1,86	4,26	7,169	7,023	7,083
7	8759	1458	Hcfc2	10365242	Hcfc2	A0AUN4		1,48	1,47	1,54	9,241	8,99	8,967
8	18740	8191	Ccdc112	10458794	Ccdc112	A0AUP1		1,83	1,94	2,00	6,761	6,656	6,729
9	16962		EG635895	10441899	EG635895	A0AUV4		1,94	1,70	2,17	6,689	6,456	6,528
10	25569	12857	Srp72	10522676	Srp72	A0JLN1		1,58	1,42	1,58	11,442	11,313	11,359
11	29717	15643	Ceacam15	10560294	Ceacam15	A0JLX4		1,77	1,79	1,84	6,464	6,546	6,57
12	21302	10042	4930578N16	10483536	4930578N16	A0JLY1		1,53	1,43	1,73	6,763	6,444	6,701
13	26618	13525	Ccdc64	10533007	Ccdc64	A0JNT9		1,66	1,81	1,81	7,525	7,747	7,819
14	12174	3901	Aspg	10398795	Aspg	A0JNU3		2,83	2,00	2,86	11,036	10,897	10,361
15	11143	3199	Efcab5	10388625	Efcab5	A0JP43		1,57	1,48	1,57	5,682	6,019	5,869
16	23261		Bclp2	10501007	Bclp2	A0JP54		1,61	1,44	1,64	5,902	6,109	6,33
17	22458	10842	Sprr2e	10493867	Sprr2e	A0PJN0		2,30	2,23	2,36	6,463	6,463	6,027
18	13599		3110006E14	10410460	3110006E14	A0PJN4		2,34	2,31	2,41	7,143	7,618	7,502
19	11368	3370	Krtap2-4	10390911	Krtap2-4	A0PK51		1,98	1,84	2,00	8,862	9,531	9,336
20	34031	18366	Olfr239	10598105	Olfr239	A0PK55		2,42	1,71	2,42	8,869	9,153	9,144
21	19543	8787	Olfr1436	10466298	Olfr1436	A0PK57		2,31	1,97	2,40	5,503	5,333	5,246
22	12493	4115	Zdhhc22	10401702	Zdhhc22	A0PK84		2,02	1,92	2,02	6,167	6,536	6,098
23	12494		Zdhhc22	10401705	Zdhhc22	A0PK84		1,94	1,72	2,08	7,594	7,667	7,792
24	24060	11857	Oscp1	10508099	Oscp1	A0ZV96		1,79	1,43	2,02	6,969	7,283	7,358
25	22977	11186	Veph1	10498547	Veph1	A1A535		1,53	1,42	1,53	6,895	7,088	6,995
26	31360	16620	Isx	10572865	Isx	A1A546		1,78	1,76	1,94	6,937	7,337	7,264
27	22452	10839	Pglyrp3	10493842	Pglyrp3	A1A547		1,46	1,55	1,55	6,551	6,669	6,62
28	27988	14437	Tcf3	10545458	Tcf3	A1A549		1,59	1,41	1,64	8,475	8,601	8,491
29	34724	18750	Tbc1d25	10603478	Tbc1d25	A1A5B6		1,84	1,62	1,90	8,255	8,109	8,212
30	25590	12871	Odam	10522875	Odam	A1E960		1,33	1,31	1,58	5,295	5,329	5,511
31	12381	4033	Fscb	10400564	Fscb	A1EGX6		1,58	1,82	1,82	6,63	6,526	6,52
32	18783		4933429F08	10459210	4933429F08	A1IGU4		2,62	2,68	3,11	8,131	7,427	7,606
33	7103	336	Faim3	10349593	Faim3	A1KXC4		1,70	1,94	1,94	7,183	7,518	7,541
34	26741	13622	Srcrb4d	10534537	Srcrb4d	A1L0T3		1,43	1,51	1,55	8,652	8,993	8,869
35	15335	5864	Olfr282	10426592	Olfr282	A1L1B4		2,20	1,67	2,22	8,018	8,162	8,189

- Sometimes multiple “probesets“ for a single gene …

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Our data set

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Generated on Affymetrix Mouse Gene 1.0 ST platform												
2								Fold	Fold	Fold			
3	Original_O	MATCH_ORDER	Gene	Probeset	Gene	Protein		Range CD	Range HFD	Range_All	C57BL/6J_Liver_CD	DBA/2J_Liver_CD	BXD43_Liver_CD
4	12643		Igh	10403036	Igh	0		1,50	1,72	1,72	5,138	5,296	5,061
5	12646		Igh	10403043	Igh	0		2,94	1,99	3,60	7,983	7,529	7,645
6	12653		Igh	10403063	Igh	0		3,72	1,86	4,26	7,169	7,023	7,083
7	8759	1458	Hcfc2	10365242	Hcfc2	A0AUN4		1,48	1,47	1,54	9,241	8,99	8,967
8	18740	8191	Ccdc112	10458794	Ccdc112	A0AUP1		1,83	1,94	2,00	6,761	6,656	6,729
9	16962		EG635895	10441899	EG635895	A0AUV4		1,94	1,70	2,17	6,689	6,456	6,528
10	25569	12857	Srp72	10522676	Srp72	A0JLN1		1,58	1,42	1,58	11,442	11,313	11,359
11	29717	15643	Ceacam15	10560294	Ceacam15	A0JLX4		1,77	1,79	1,84	6,464	6,546	6,57
12	21302	10042	4930578N16	10483536	4930578N16	A0JLY1		1,53	1,43	1,73	6,763	6,444	6,701
13	26618	13525	Ccdc64	10533007	Ccdc64	A0JNT9		1,66	1,81	1,81	7,525	7,747	7,819
14	12174	3901	Aspg	10398795	Aspg	A0JNU3		2,83	2,00	2,86	11,036	10,897	10,361
15	11143	3199	Efcab5	10388625	Efcab5	A0JP43		1,57	1,48	1,57	5,682	6,019	5,869
16	23261		Bclp2	10501007	Bclp2	A0JP54		1,61	1,44	1,64	5,902	6,109	6,33
17	22458	10842	Sprr2e	10493867	Sprr2e	A0PJN0		2,30	2,23	2,36	6,463	6,463	6,027
18	13599		3110006E14	10410460	3110006E14	A0PJN4		2,34	2,31	2,41	7,143	7,618	7,502
19	11368	3370	Krtap2-4	10390911	Krtap2-4	A0PK51		1,98	1,84	2,00	8,862	9,531	9,336
20	34031	18366	Olfr239	10598105	Olfr239	A0PK55		2,42	1,71	2,42	8,869	9,153	9,144
21	19543	8787	Olfr1436	10466298	Olfr1436	A0PK57		2,31	1,97	2,40	5,503	5,333	5,246
22	12493	4115	Zdhhc22	10401702	Zdhhc22	A0PK84		2,02	1,92	2,02	6,167	6,536	6,098
23	12494		Zdhhc22	10401705	Zdhhc22	A0PK84		1,94	1,72	2,08	7,594	7,667	7,792
24	24060	11857	Oscp1	10508099	Oscp1	A0ZV96		1,79	1,43	2,02	6,969	7,283	7,358
25	22977	11186	Veph1	10498547	Veph1	A1A535		1,53	1,42	1,53	6,895	7,088	6,995
26	31360	16620	Isx	10572865	Isx	A1A546		1,78	1,76	1,94	6,937	7,337	7,264
27	22452	10839	Pglyrp3	10493842	Pglyrp3	A1A547		1,46	1,55	1,55	6,551	6,669	6,62
28	27988	14437	Tcf3	10545458	Tcf3	A1A549		1,59	1,41	1,64	8,475	8,601	8,491
29	34724	18750	Tbc1d25	10603478	Tbc1d25	A1A5B6		1,84	1,62	1,90	8,255	8,109	8,212
30	25590	12871	Odam	10522875	Odam	A1E960		1,33	1,31	1,58	5,295	5,329	5,511
31	12381	4033	Fscb	10400564	Fscb	A1EGX6		1,58	1,82	1,82	6,63	6,526	6,52
32	18783		4933429F08	10459210	4933429F08	A1IGU4		2,62	2,68	3,11	8,131	7,427	7,606
33	7103	336	Faim3	10349593	Faim3	A1KXC4		1,70	1,94	1,94	7,183	7,518	7,541
34	26741	13622	Srcrb4d	10534537	Srcrb4d	A1L0T3		1,43	1,51	1,55	8,652	8,993	8,869
35	15335	5864	Olfr282	10426592	Olfr282	A1L1B4		2,20	1,67	2,22	8,018	8,162	8,189

- Sometimes multiple “probesets“ for a single gene …
- Otherwise expression intensity values for probesets and strains + condition (Strain_Liver_Condition columns)

Questions for today

- **Is the data normalized? If not you should normalize it …:**

1) Visual exploration whether data is already normalized

2) If not, which normalization method to choose?

Questions for today

- **Is the data normalized? If not you should normalize it …:**

  1) Visual exploration whether data is already normalized

  2) If not, which normalization method to choose?

- **Which statistical procedures do we choose to test for differential Expression:**

  1) Visual exploration of data properties important for choosing a suitable test procedure

  2) Multiple testing correction

Questions for today

- **Is the data normalized? If not you should normalize it …:**

  1) Visual exploration whether data is already normalized

  2) If not, which normalization method to choose?

- **Which statistical procedures do we choose to test for differential Expression:**

  1) Visual exploration of data properties important for choosing a suitable test procedure

  2) Multiple testing correction

Day56_DE_Analysis.ipynb