

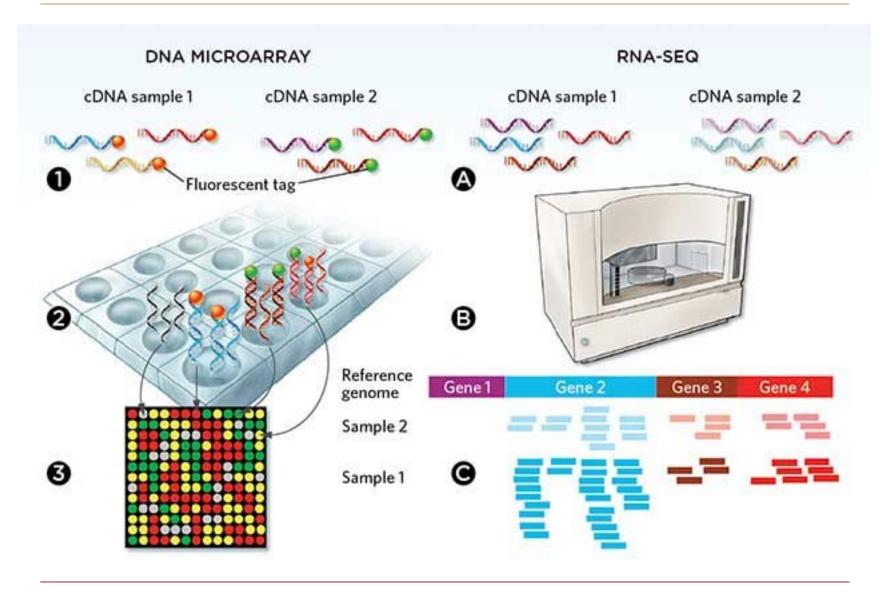


# PIBI:

**Differential Gene Expression Analysis** 



# Microarrays vs. RNA-seq





# Our data set: microarray

	Α	В	С	D	E	F		G	Н		I	J	K	L	M
1	Generated or	n Affymetrix Mouse Ge	ene 1.0 ST pla	tform											
2		,						F	old	F		Fold			
3	Original O	MATCH_ORDER -	Gene 🔻	Probeset 🔻	Gene ▼	Protein	~	▼ P	lange CD					DBA/2J_Liver_CD 🔻	
4	12643		lgh	10403036	lgh		0		1,	,50	1,72				
5	12646		lgh	10403043	lgh		0			,94	1,99	3,60			
6	12653		lgh	10403063	lgh		0		3,	,72	1,86				
7	8759	1458	Hcfc2	10365242	Hcfc2	A0AUN4				,48	1,47	1,54			
8	18740	8191	l Ccdc112	10458794	Ccdc112	A0AUP1			1,	,83	1,94				
9	16962		EG635895	10441899	EG635895	A0AUV4			1,	,94	1,70				
10	25569	12857	7 Srp72	10522676	Srp72	A0JLN1			1,	,58	1,42				
11	29717	15643	3 Ceacam15	10560294	Ceacam15	A0JLX4			1,	,77	1,79				
12	21302	10042	2 4930578N16	10483536	4930578N16	A0JLY1			1,	,53	1,43				
13	26618	13525	5 Ccdc64	10533007	Ccdc64	A0JNT9			1,	,66	1,81	1,81	7,525		
14	12174	3901	l Aspg	10398795	Aspg	A0JNU3			2,	,83	2,00				
15	11143	3199	9 Efcab5	10388625	Efcab5	A0JP43			1,	,57	1,48	1,57			
16	23261		Bclp2	10501007	Bclp2	A0JP54			1,	,61	1,44	1,64	5,902		
17	22458	10842	2 Sprr2e	10493867	Sprr2e	A0PJN0			2,	,30	2,23	2,36			
18	13599		3110006E14	10410460	3110006E14	A0PJN4			2,	,34	2,31	2,41			
19	11368		0 Krtap2-4	10390911	Krtap2-4	A0PK51			1,	,98	1,84	2,00			
20	34031	18366	6 Olfr239	10598105	Olfr239	A0PK55			2,	,42	1,71	2,42	8,869		
21	19543	8787	7 Olfr1436	10466298	8 Olfr1436	A0PK57			2,	,31	1,97	2,40	5,503		
22	12493	4115	5 Zdhhc22	10401702	Zdhhc22	A0PK84			2,	,02	1,92	2,02	6,167		
23	12494		Zdhhc22	10401705	Zdhhc22	A0PK84			1,	,94	1,72				
24	24060		7 Oscp1	10508099	Oscp1	A0ZV96			1,	,79	1,43				
25	22977		6 Veph1	10498547	Veph1	A1A535			1,	,53	1,42				
26	31360	16620	0 lsx	10572865	Isx	A1A546			1,	,78	1,76	1,94			
27	22452	10839	9 Pglyrp3	10493842	Pglyrp3	A1A547			1,	,46	1,55				
28	27988			10545458	3 Tcf3	A1A549			1	,59	1,41	1,64			
29	34724		0 Tbc1d25	10603478	Tbc1d25	A1A5B6			1,	,84	1,62				
30	25590		1 Odam	10522875	Odam	A1E960			1	,33	1,31	1,58			
31	12381		3 Fscb	10400564	1 Fscb	A1EGX6			1	,58	1,82	1,82	6,63		
32	18783		4933429F08	10459210	4933429F08	A1IGU4			2	,62	2,68	3,1	8,131		
33	7103		6 Faim3	10349593	Faim3	A1KXC4			1	,70	1,94	1,94	7,183		
34	26741		2 Srcrb4d	10534537	7 Srcrb4d	A1L0T3			1	,43	1,51	1,5			
35	15335		4 Olfr282	10426592	Olfr282	A1L1B4			2	,20	1,67	2,2	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 ...



### **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



### **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)

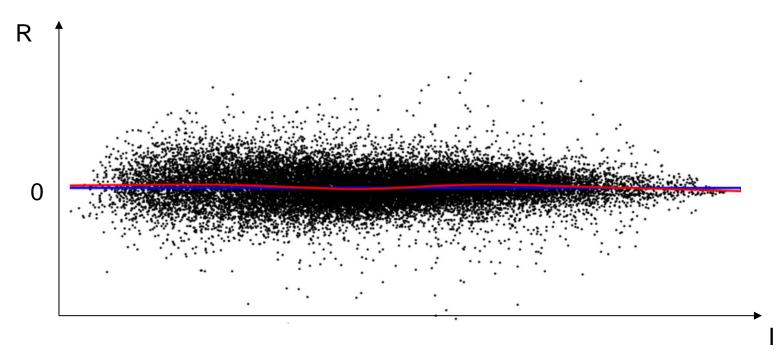
#### Low-level examples:

- 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**



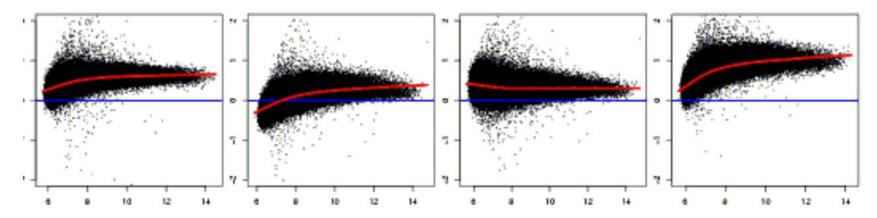
R = log( expression in sample 1 / expression in sample 2 )

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

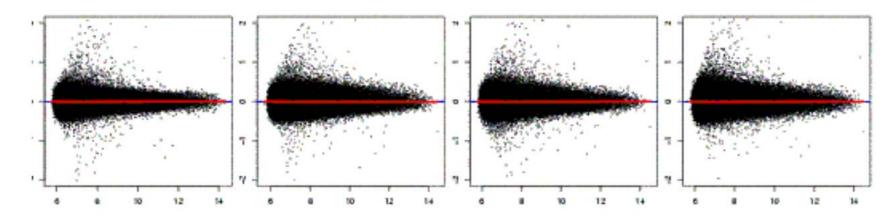
# **Data normalization: What happens?**

### **Example 1: ratio-intensity plots**

#### Before normalization:



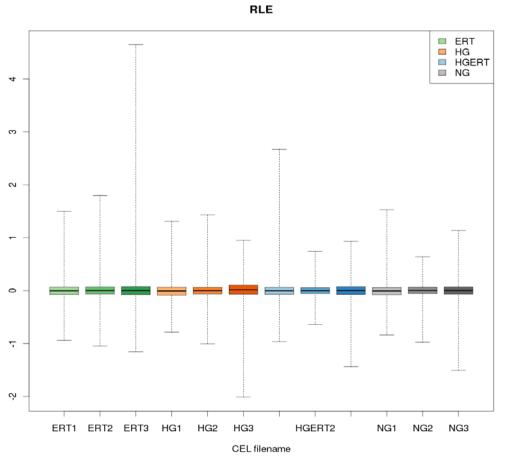
#### After normalization:





# Data normalization: What happens?

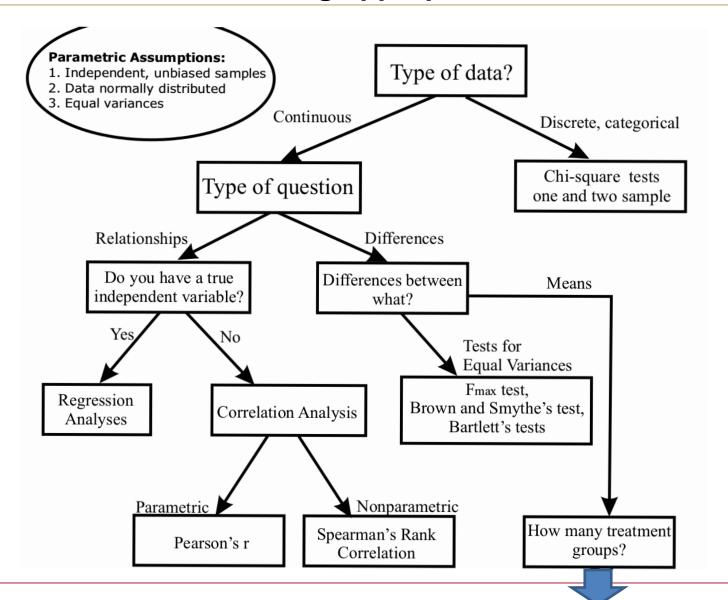
### **Example 2: relative log expression plot (RLE plot)**



- 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

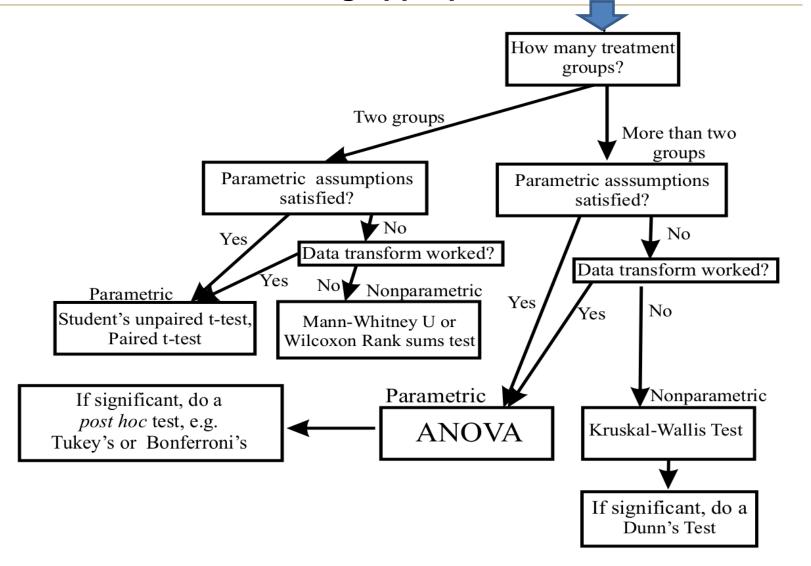


## **Choosing appropriate statistical methods**



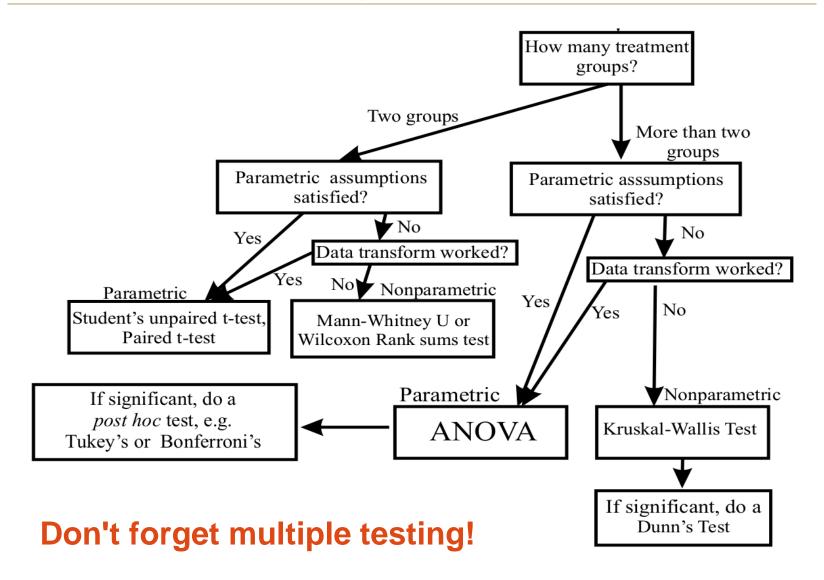


## **Choosing appropriate statistical methods**





# Choosing appropriate statistical methods





## **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.

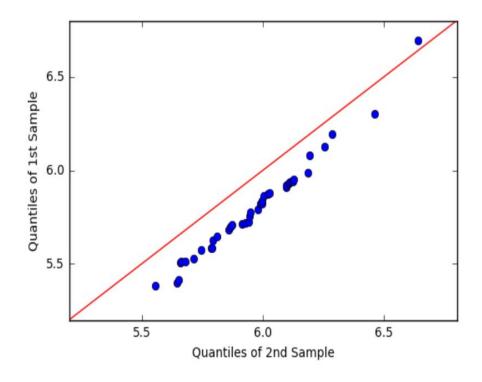


## **Comparing distributions: Q-Q plots**

#### 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.





# Our data set

	Α	В	С	D	E	F		G	Н	I	J		K	L	M
1	Generated or														
2		•							Fold	Fold	Fold				
3	Original O	MATCH_ORDER	→ Gene -	Probeset -	Gene •	Protein	~	~	Range CD	Range HFD	▼ Range_Al			DBA/2J_Liver_CD 🔻	
4	12643		lgh	10403036	lgh		0		1,50			1,72	5,138		
5	12646		lgh	10403043	lgh		0		2,9	1,	99	3,60	7,983	7,529	
6	12653		lgh	10403063	lgh		0		3,7	2 1,	36	4,26	7,169	7,023	
7	8759	145	8 Hcfc2	10365242	Hcfc2	A0AUN4			1,48	3 1,	47	1,54	9,241	8,99	
8	18740	819	1 Ccdc112	10458794	Ccdc112	A0AUP1			1,83	3 1,	94	2,00	6,761	6,656	
9	16962		EG635895	10441899	EG635895	A0AUV4			1,9	1 1,	70	2,17	6,689		
10	25569	1285	7 Srp72	10522676	Srp72	A0JLN1			1,58	1,	42	1,58	11,442		
11	29717	1564	3 Ceacam15	10560294	Ceacam15	A0JLX4			1,7	7 1,	79	1,84	6,464		
12	21302		2 4930578N1	0 10483536	4930578N1	A0JLY1			1,5	3 1,	43	1,73	6,763		
13	26618	1352	25 Ccdc64	10533007	Ccdc64	A0JNT9			1,6	5 1,	81	1,81	7,525		
14	12174		1 Aspg	10398795	Aspg	A0JNU3			2,83	3 2,	00	2,86	11,036		
15	11143		99 Efcab5	10388625	Efcab5	A0JP43			1,5	7 1,	48	1,57	5,682		
16	23261		Bclp2	10501007	7 Bclp2	A0JP54			1,6	1,	44	1,64	5,902		
17	22458	1084	12 Sprr2e	10493867	7 Sprr2e	A0PJN0			2,3	2,	23	2,36	6,463		
18	13599		3110006E14	10410460	3110006E14	1-A0PJN4			2,3	1 2,	31	2,41	7,143		
19	11368	337	70 Krtap2-4	10390911	LKrtap2-4	A0PK51			1,9	3 1,	84	2,00	8,862		
20	34031		6 Olfr239	10598105	Olfr239	A0PK55			2,4	2 1,	71	2,42	8,869	9,153	
21	19543	878	37 Olfr1436	10466298	3 Olfr1436	A0PK57			2,3	1 1,	97	2,40	5,503		
22	12493		L5 Zdhhc22	10401702	Zdhhc22	A0PK84			2,0	2 1,	92	2,02	6,167	6,536	
23	12494		Zdhhc22	10401705	Zdhhc22	A0PK84			1,9	4 1,	72	2,08	7,594		
24	24060		7 Oscp1	10508099	Oscp1	A0ZV96			1,7	1,	43	2,02	6,969		
25	22977		36 Veph1	10498547		A1A535			1,5	3 1,	42	1,53	6,895		
26	31360		20 lsx	10572865	Isx	A1A546			1,7	3 1,	76	1,94			
27	22452		39 Pglyrp3	10493842		A1A547			1,4	5 1,	55	1,55	6,551		
28	27988		37 Tcf3	10545458		A1A549			1,5	9 1,	41	1,64			
29	34724		50 Tbc1d25	10603478	BTbc1d25	A1A5B6			1,8	4 1,	62	1,90	8,255		
30	25590		71 Odam	10522875	Odam Odam	A1E960			1,3	3 1,	31	1,58			
31	12381		33 Fscb	10400564		A1EGX6			1,5	В 1,	82	1,82	6,63		
32	18783		4933429F08		4933429F08	3 A1IGU4			2,6	2 2,	68	3,11	8,131	7,427	
33	7103		36 Faim3	10349593		A1KXC4			1,7	0 1,	94	1,94	7,183		
34	26741		22 Srcrb4d	1053453		A1L0T3			1,4	3 1,	51	1,55	8,652	8,993	
35	15335		54 Olfr282	10426592	Olfr282	A1L1B4			2,2	0 1,	67	2,22	8,018	8,162	8,189



### Our data set

	Α	ВС	D E	F	G	Н	ı	J	K	L	M
1	Generated or	Affymetrix Mouse Gene 1.0 ST plan	tform								
2	Generated 5	7.11.7.11.00.11.11.11.11.11.11.11.11.11.11.11.				Fold	Fold	Fold			
3	Original O	MATCH ORDER Gene	Probeset Gene	Protein -	· ·	Range CD 🔻	Range HFD 🖃	Range_All 🔻	C57BL/6J_Liver_CD		BXD43_Liver_CD
4	12643	lgh	10403036 lgh	(	0	1,50	1,72	1,72	5,138	5,296	5,061
5	12646	lgh	10403043 lgh	(	0	2,94	1,99	3,60	7,983	7,529	
6	12653	lgh	10403063 lgh	(	0	3,72	1,86	4,26	7,169		
7	8759	1458 Hcfc2	10365242 Hcfc2	A0AUN4		1,48	1,47	1,54		8,99	
8	18740	8191 Ccdc112	10458794 Ccdc112	A0AUP1		1,83	1,94	2,00		6,656	
9	16962	EG635895	10441899 EG635895	A0AUV4		1,94	1,70	2,17	6,689		
10	25569	12857 Srp72	10522676 Srp72	A0JLN1		1,58	1,42	1,58			
11	29717	15643 Ceacam15	10560294 Ceacam15	A0JLX4		1,77	1,79	1,84			
12	21302	10042 4930578N16	10483536 4930578N16	A0JLY1		1,53	1,43	1,73			
13	26618	13525 Ccdc64	10533007 Ccdc64	A0JNT9		1,66	1,81	1,81			
14	12174	3901 Aspg	10398795 Aspg	A0JNU3		2,83	2,00	2,86			
15	11143	3199 Efcab5	10388625 Efcab5	A0JP43		1,57	1,48				
16	23261	Bclp2	10501007 Bclp2	A0JP54		1,61	1,44	1,64			
17	22458	10842 Sprr2e	10493867 Sprr2e	A0PJN0		2,30	2,23	2,36			
18	13599	3110006E14	10410460 3110006E14	A0PJN4		2,34	2,31	2,41			
19	11368	3370 Krtap2-4	10390911 Krtap2-4	A0PK51		1,98	1,84	2,00			
20	34031	18366 Olfr239	10598105 Olfr239	A0PK55		2,42	1,71	2,42			
21	19543	8787 Olfr1436	10466298 Olfr1436	A0PK57		2,31	1,97	2,40			
22	12493	4115 Zdhhc22	10401702 Zdhhc22	A0PK84		2,02	1,92	2,02			
23	12494	Zdhhc22	10401705 Zdhhc22	A0PK84		1,94	1,72	2,08			
24	24060	11857 Oscp1	10508099 Oscp1	A0ZV96		1,79					
25	22977	11186 Veph1	10498547 Veph1	A1A535		1,53	1,42	1,53			
26	31360	16620 lsx	10572865 Isx	A1A546		1,78					
27	22452	10839 Pglyrp3	10493842 Pglyrp3	A1A547		1,46					
28	27988	14437 Tcf3	10545458 Tcf3	A1A549		1,59		1,64			
29	34724	18750 Tbc1d25	10603478 Tbc1d25	A1A5B6		1,84	1,62				
30	25590	12871 Odam	10522875 Odam	A1E960		1,33					
31	12381	4033 Fscb	10400564 Fscb	A1EGX6		1,58					
32	18783	4933429F08	10459210 4933429F08	A1IGU4		2,62					
33	7103	336 Faim3	10349593 Faim3	A1KXC4		1,70					
34	26741	13622 Srcrb4d	10534537 Srcrb4d	A1L0T3		1,43	1,51				
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 ...



#### Our data set

		D		D	_	E		G	Н					K	L	M
	A Concepted on	n Affymetrix Mouse Ge	no 1 0 CT pla		E	Г		u	п			,		- R	-	
1	Generated on	i Affymetrix Mouse Ge	ne 1.0 51 pia	uom					Fold	F	old	Fold	_			
2	Original OF	MATCH_ORDER 🔽	Gene ▼	Probeset 🔻	Gene	Protein	*		Range CD				ŢC'	57BL/6I Liver CD	DBA/2] Liver CD 🔽	BXD43 Liver CD
3	12643			10403036		rioteiii	0			50	1,72	1,7		5,138		
4	12646		lgh lgh	10403030			0			94	1,99	3.6		7,983	7,529	7,645
5	12653		lgh	10403043			0			72	1.86	4,2		7,169		7,083
7	8759		Hcfc2	10365242		A0AUN4	-			48	1,47	1,5		9.241	8,99	
8	18740		Ccdc112	10458794		A0AUP1				83	1,94	2,0		6,761	6,656	
9	16962		EG635895		EG635895	A0AUV4				94	1,70	2,1		6,689		
10	25569	12857		10522676		A0JLN1				58	1,42	1.5		11,442		11,359
11	29717		Ceacam15		Ceacam15	A0JLX4				77	1,79	1,8		6,464	6,546	
12	21302		4930578N16		4930578N16					53	1,43	1,7		6,763	6,444	6,701
13	26618		Ccdc64	10533007		AOINT9				,66	1,81	1,8		7,525	7,747	7,819
14	12174	3901		10398795		A0JNU3				,83	2,00	2,8		11,036	10,897	
15	11143		Efcab5	10388625		A0JP43				.57	1,48	1,5	57	5,682	6,019	5,869
16	23261		Bclp2	10501007		A0JP54			1,	,61	1,44	1,6	64	5,902	6,109	6,33
17	22458		Sprr2e	10493867		A0PJN0			2,	,30	2,23	2,3	36	6,463		
18	13599		3110006E14		3110006E14				2,	,34	2,31	2,4	41	7,143		
19	11368	3370	Krtap2-4	10390911	Krtap2-4	A0PK51			1,	,98	1,84	2,0	00	8,862		
20	34031		Olfr239	10598105	Olfr239	A0PK55			2,	,42	1,71	2,4	42	8,869		
21	19543	8787	Olfr1436	10466298	Olfr1436	A0PK57			2,	,31	1,97	2,4	40	5,503		
22	12493	4115	Zdhhc22	10401702	Zdhhc22	A0PK84			2,	,02	1,92		02	6,167	6,536	
23	12494		Zdhhc22	10401705	Zdhhc22	A0PK84			1,	,94	1,72		80	7,594		
24	24060	11857	Oscp1	10508099	Oscp1	A0ZV96				,79	1,43		02	6,969		
25	22977	11186	Veph1	10498547	Veph1	A1A535				,53	1,42		53	6,895		
26	31360	16620	Isx	10572865	Isx	A1A546				,78	1,76		94	6,937		
27	22452	10839	Pglyrp3	10493842	Pglyrp3	A1A547				,46	1,55		55	6,551		
28	27988	14437	Tcf3	10545458	Tcf3	A1A549				,59	1,41	1,0		8,475		
29	34724	18750	Tbc1d25	10603478		A1A5B6				,84	1,62	1,9		8,255		
30	25590	12871	Odam	10522875	Odam	A1E960				,33	1,31		58	5,295		
31	12381	4033	Fscb	10400564		A1EGX6				,58	1,82		82	6,63		
32	18783		4933429F08	10459210	4933429F08	A1IGU4				,62	2,68		11	8,131		
33	7103	336	Faim3	10349593	Faim3	A1KXC4				,70	1,94		94	7,183		
34	26741	13622	Srcrb4d	10534537	Srcrb4d	A1L0T3				,43	1,51		55	8,652		
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)





- 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?



- 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



- 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