

Automated analysis of high-throughput siRNA screening data

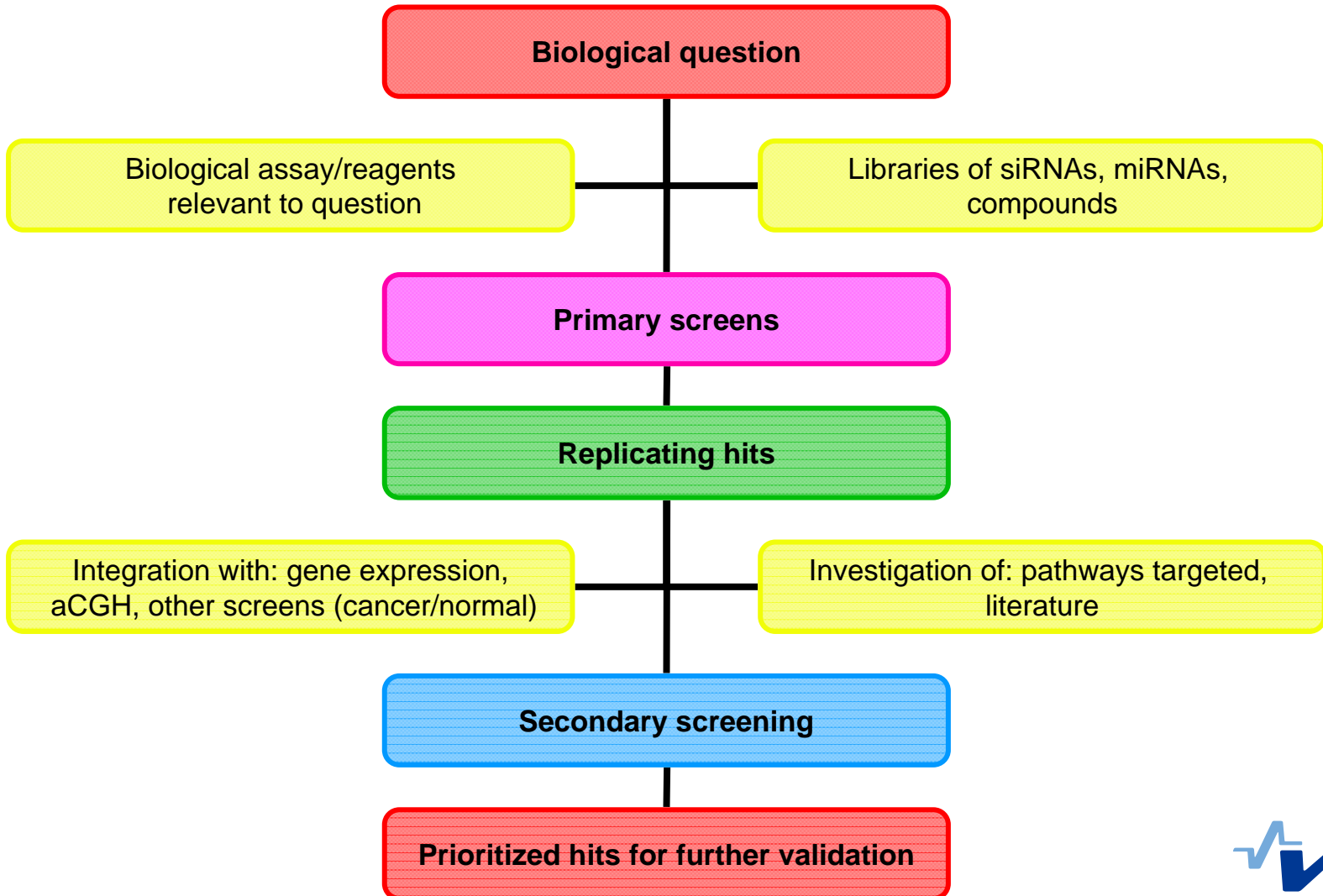
APO-SYS WP7 Workshop “Biostatistics” ,
January 19, 2009

Pekka Kohonen, Vidal Fey
VTT Medical Biotechnology

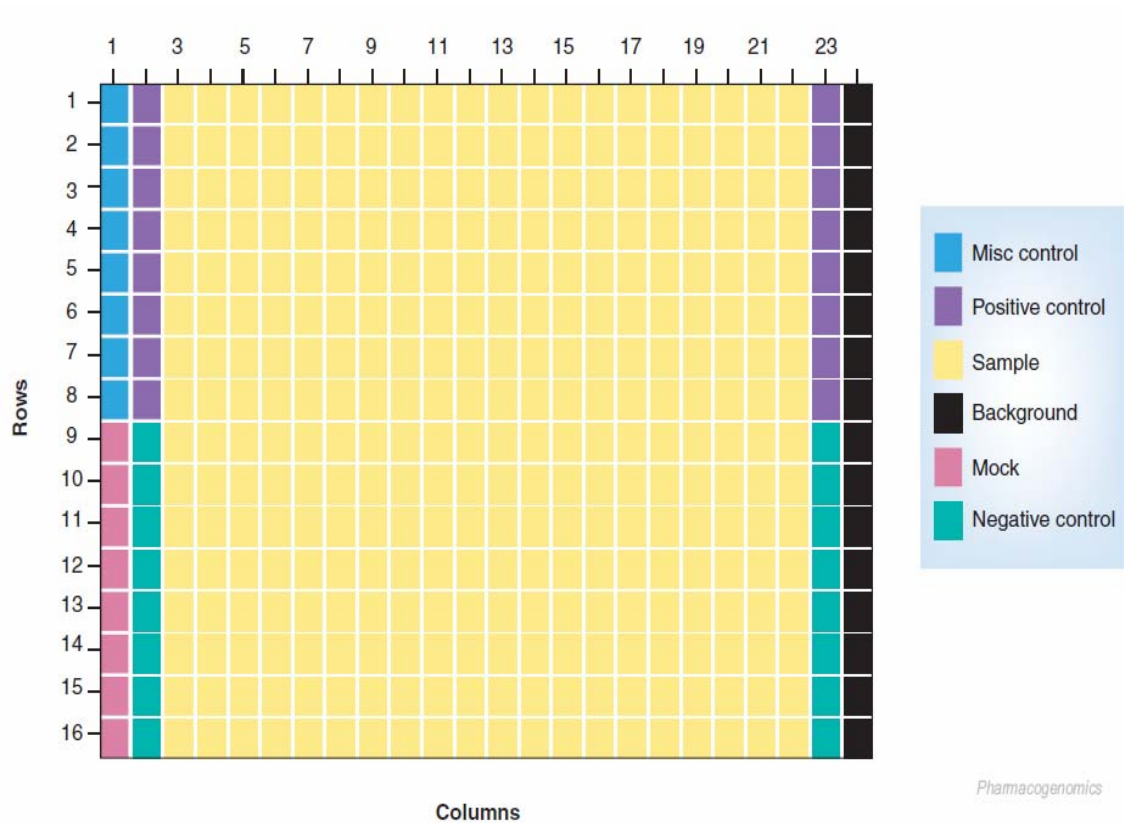
Presentation overview

1. Screening workflow
2. Design considerations:
 - Which genes to assay: biological question
 - Sources of error in the screens:
 - Experimental (Optimization of the biological materials)
 - Off-target effects and how to avoid them: specific/non-specific
3. Statistical concepts important for RNAi screens
4. RNAi screening data analysis **software** (siRna, cellHTS2)
5. Sensitization analysis (synthetic lethal screens)
6. siRNA screenign is not the end but the beginning.

Screening work-flow

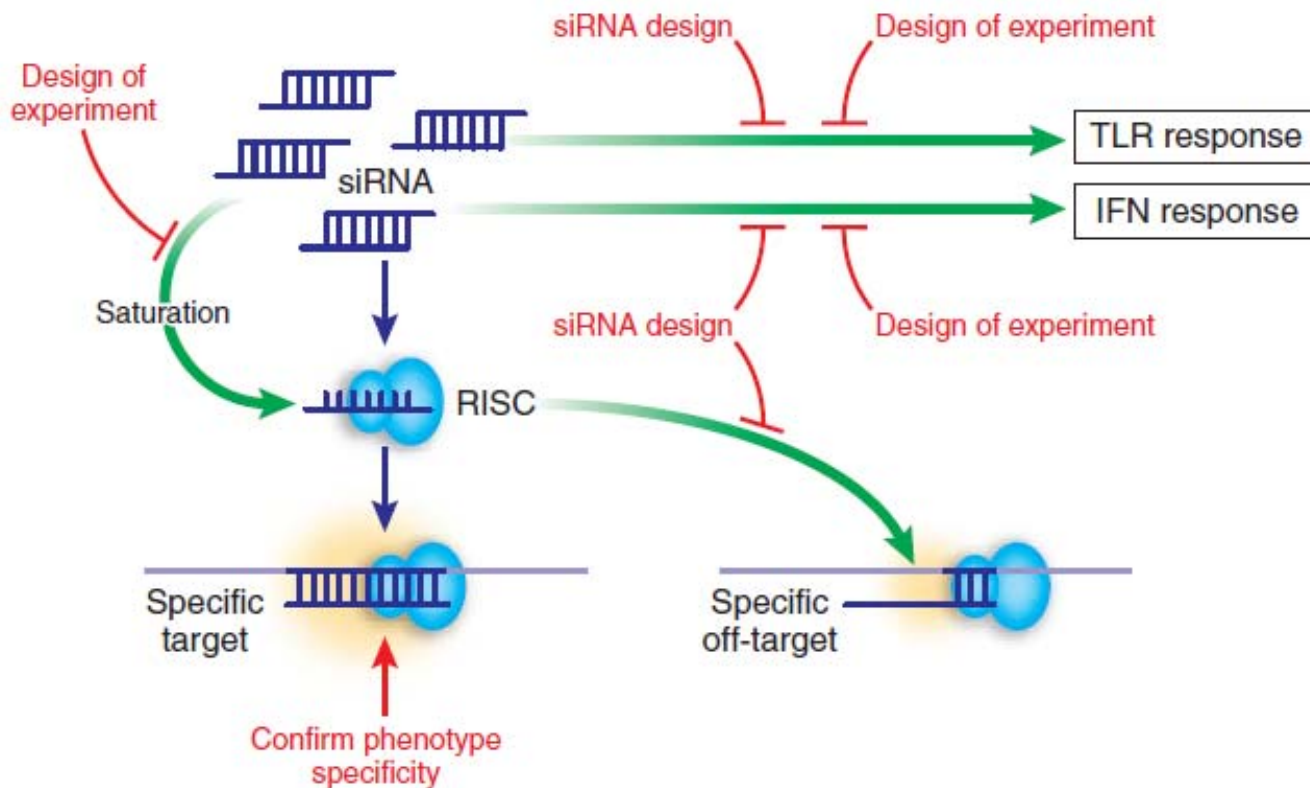


Design considerations: Identity and placement of controls



- Positive control: PLK1, KIFF11(EG5) or **specific for the assay**
- Negative control: scramble siRNA, non-hitting genes?
- Mock transfection/lipid only
- Background/ cells-only
- Enough controls that you can calculate standard deviation from them: 8-16
- Placement of controls on the array should be random, but in practice at least some alternation of rows/columns is good

Design considerations: Off-target effects



All off-target effects are cell line and reagent specific

- Non-sequence specific off-target effects:
 - Interferon response
 - siRNA causing miRNA machinery saturation
 - **Lipid toxicity**
- Specific:
 - Effects on related mRNAs
 - **miRNA mechanism based off-target effects**

Design considerations: how to minimize off-target effects (OTE)

1. **Optimize** screens (not covered on this course).
 - Use **fresh cells** and minimal amounts of siRNA.
2. Use **replicate screens and redundant** siRNAs (2-8).
 - Different siRNAs are unlikely to have the same off-targets
 - Pools can also reduce OTE but redundant siRNAs = better
3. Assay as **many parameters** as possible.
4. **Normalize** the screen properly.
5. **Ratios** of measurements can help:
 - Ratios control often for cell amount etc...(self-normalizing)
6. Knockdown of the mRNA/protein needs to be **confirmed**.
7. Always treat screening data with **caution**.

Robust statistics reduces the effect of outliers on the results

- Robustness: method is relatively insensitive to outliers, anomalous values, in the series
- Mean: The average of the values in the series
- Median: the center of the values in the series.
- Standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

- Median absolute deviation:

$$\text{MAD} = \text{median}_i (|X_i - \text{median}_j(X_j)|)$$

$$\text{median}(|2-3.25|, \dots, |8-3.25|)$$

- $\text{sd} = 1.4826 * \text{MAD}$

- $s1 = (2, 2.5, 3, 3.5, 3.5, 4)$

- $s2 = (2, 2.5, 3, 3.5, 3.5, 8)$

- Mean:

- $\text{mean}(s1) = 3.08$

- $\text{Mean}(s2) = 3.75$

- Median

- $\text{median}(s1) = 3.25$

- $\text{median}(s2) = 3.25$

- Standard deviation:

- $\text{sd}(s1) = 0.74$

- $\text{sd}(s2) = 2.16$

- Median absolute deviation:

- $\text{mad}(s1) = 0.74$

- $\text{mad}(s2) = 0.74$

Z-scores and mad-scores

**Z-score is the distance of the value from mean
in standard deviation units.**

- Z-score: $x - \text{mean}(\text{sample})/\text{sd}(\text{sample})$
- MAD-score: $x - \text{median}(\text{sample})/\text{MAD}(\text{sample})$
- Robust Z-score: $x - \text{median}(\text{sample})/\text{MAD}(\text{sample}) * 1.4826$

**If robust metrics are not used it is possible to miss
significant changes in activity.**

> $(s1 - \text{mean}(s1))/\text{sd}(s1)$

[1] -1.4719601 -0.7925939 -0.1132277 0.5661385 0.5661385 1.2455047

> $(s1 - \text{median}(s1))/\text{mad}(s1)$

[1] -1.6862269 -1.0117361 -0.3372454 0.3372454 0.3372454 1.0117361

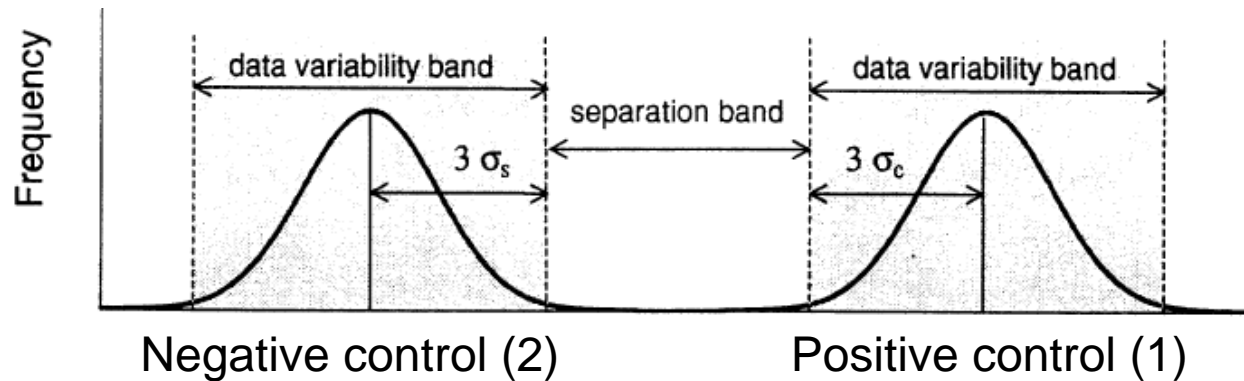
> $(s2 - \text{mean}(s2))/\text{sd}(s2)$

[1] -0.8093703 -0.5781216 -0.3468730 -0.1156243 -0.1156243 **1.9656135**

> $(s2 - \text{median}(s2))/(\text{mad}(s2) * 1.4826)$

[1] -1.1373445 -0.6824067 -0.2274689 0.2274689 0.2274689 **4.3219090**

Quality control parameters for screens



Signal to noise ratio:

$$S/N = \frac{\mu_1 - \mu_2}{\sigma_2} \quad \text{and} \quad S/B = \frac{\mu_1}{\mu_2}$$

mean/median of *positive controls*
 (1) - mean/median of *negative controls* (2) divided by the standard deviation of the negative controls

Z' - factor:

$$Z_f = 1 - \frac{3(\sigma_1 + \sigma_2)}{|\mu_1 - \mu_2|}$$

Z' = 1 (ideal screen)

1 > Z' > 0.5 (excellent screen)

0.5 > Z' > 0 (doable screen)

Z < 0 (screen has failed)

Software for siRNA data analysis

- R/Bioconductor:
 - **CellHTS2**
 - Has been around for a longer time
 - Html reports of the results
 - **RNAither**
 - More developed between screen normalizations
- **siRna** (developed at VTT)
 - A workflow oriented approach
 - Minimal command typing and hands on time
 - one command carries out entire analysis automatically
 - Minimal knowledge of R needed
 - A great deal of graphical output is generated
 - Creates Excel readable output files

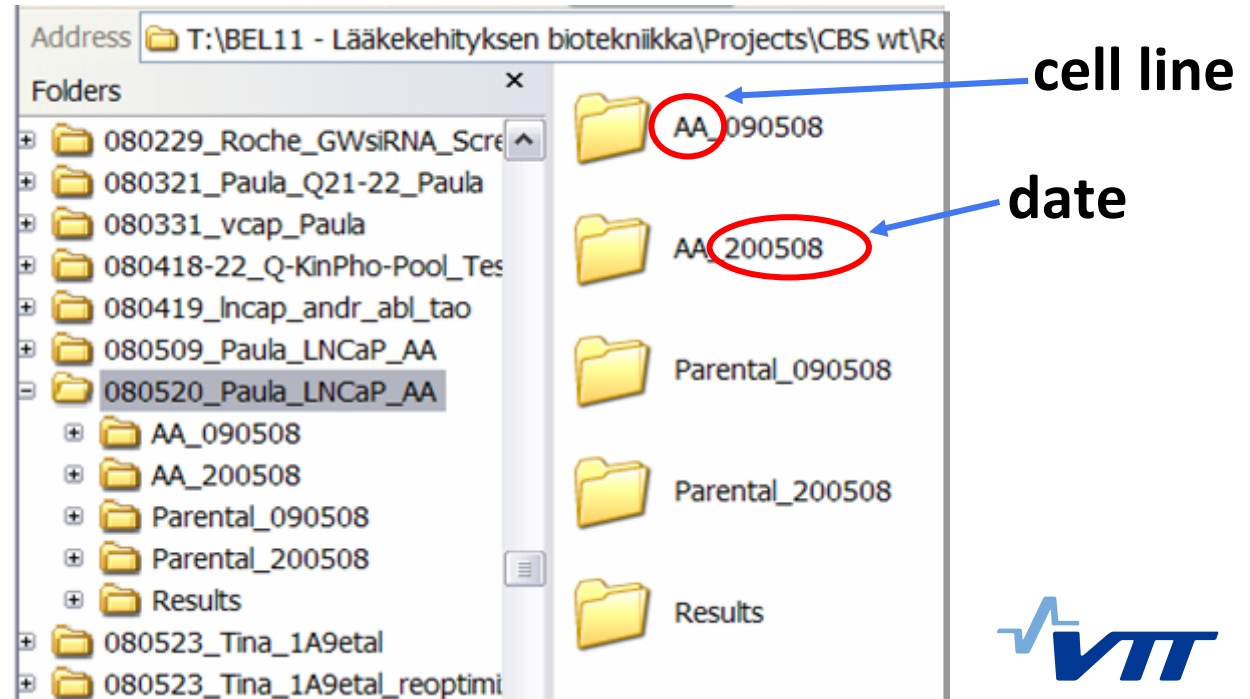
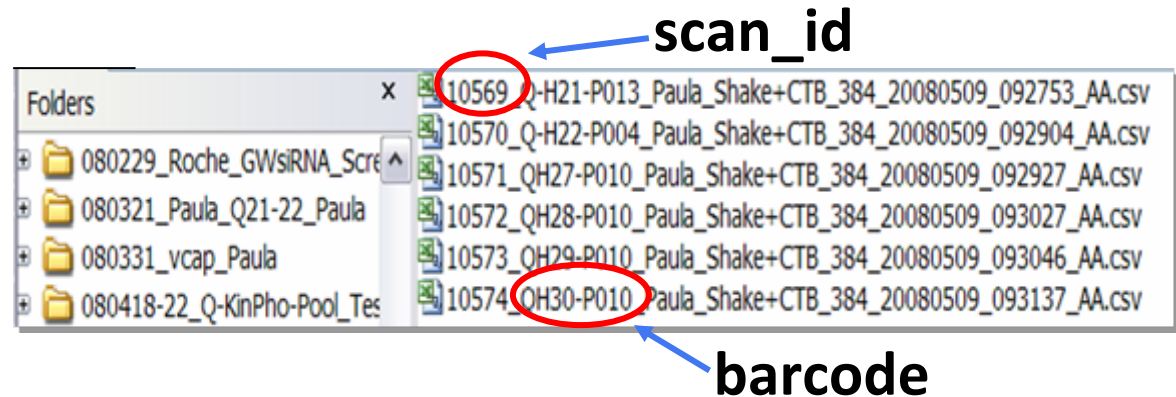
Outline of an analysis run using *siRna*

- Reading in the data
 - Correct format is important
 - Defining the siRNAs (Control/Sample/**use of barcodes**)
 - Exceptional annotations handling
- Normalisation of the data
- Plate series plot: for each screen/normalisation
- Plate image plot: for each plate/normalisation
- Data distribution plots:
 - tests for normal distribution and artefacts
 - q-norm plot, histogram, boxplot
- A combitable for each normalisation (Excel compatible)
- Log file containing analysis verbose output is created

Reading in the data

Raw data file:

	A	B	C
1			PltRpt=1 Gr
2	Barcode	Well	Signal
3	Q-H21-P013	A01	3687162
4	Q-H21-P013	A02	4165002
5	Q-H21-P013	A03	3922653
6	Q-H21-P013	A04	3830668
7	Q-H21-P013	A05	4210147
8	Q-H21-P013	A06	4230210
9	Q-H21-P013	A07	4467150
10	Q-H21-P013	A08	4420726
11	Q-H21-P013	A09	4201349
12	Q-H21-P013	A10	4621194



Results can be displayed in a plate series plot

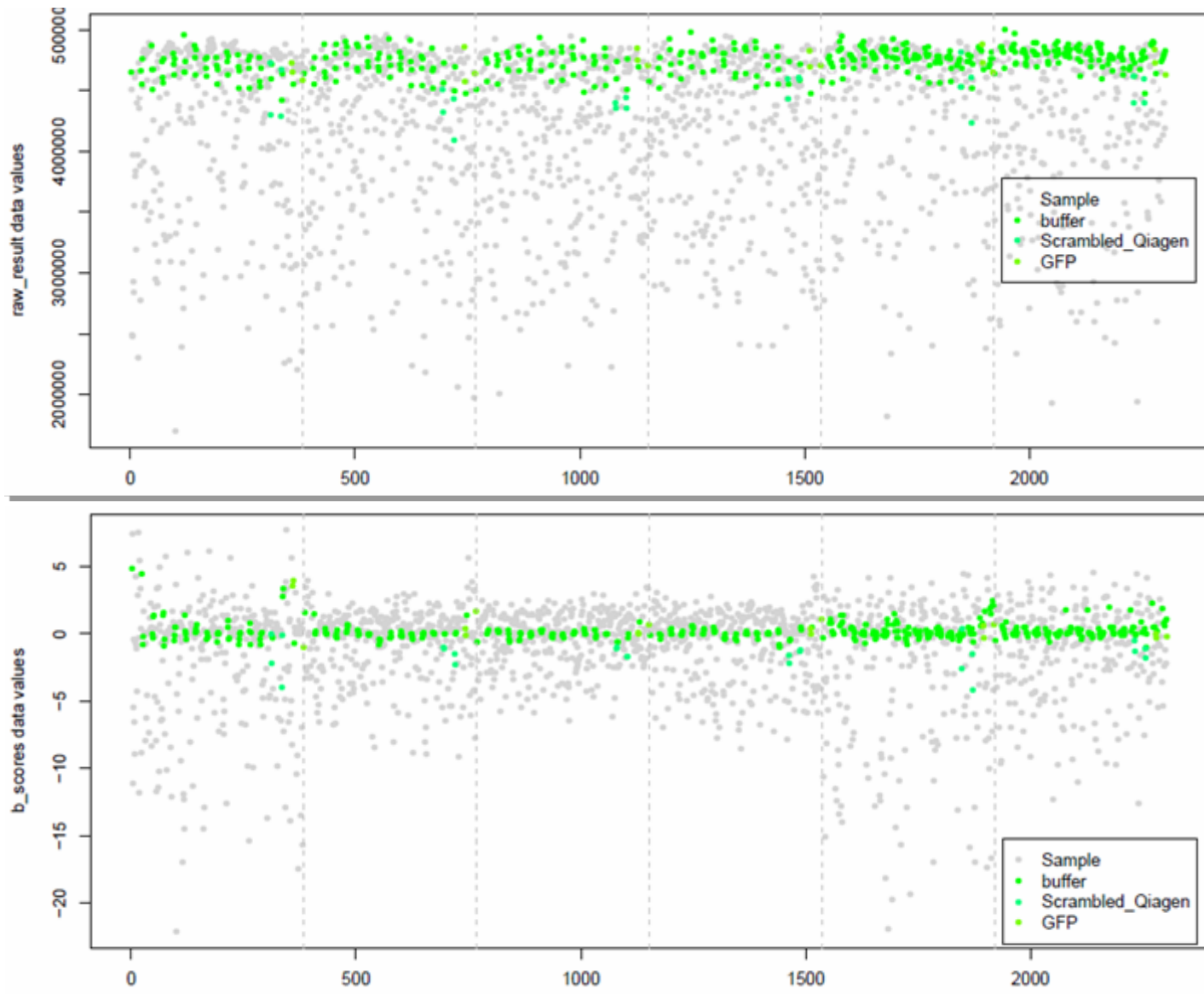
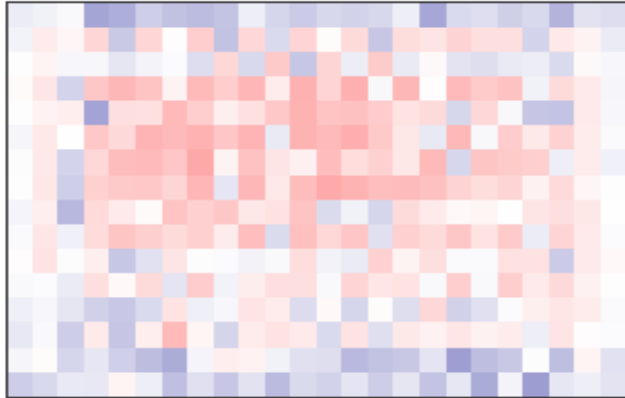
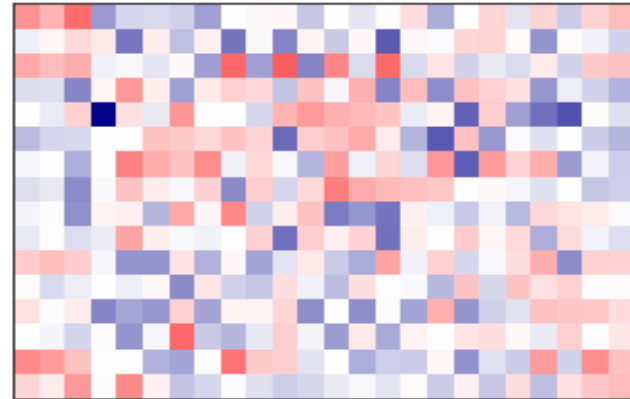


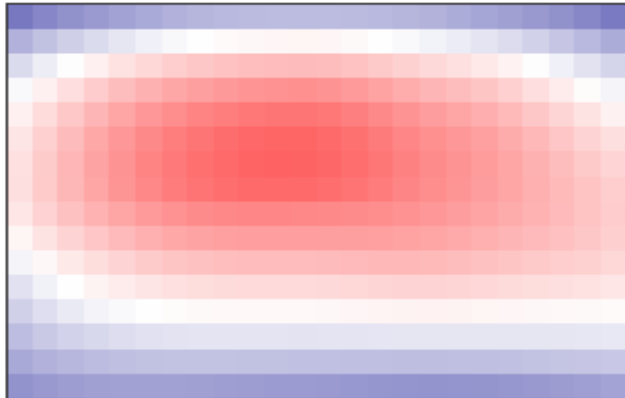
Plate picture plots make it easy to compare normalizations



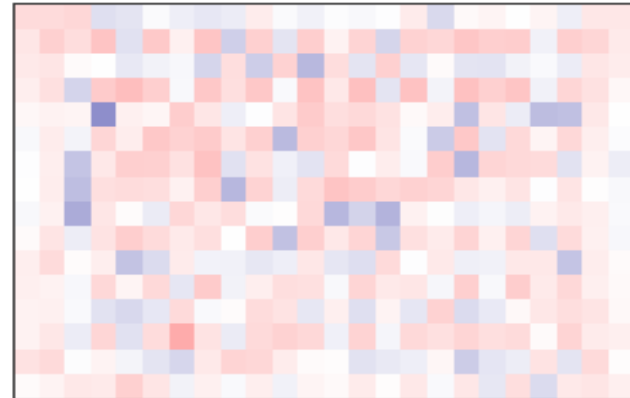
Q-F01A-P059

raw data

Q-F01A-P059

B-score

Q-F01A-P059

loess fit

Q-F01A-P059

loess-log

Lower signal values in **blue**, higher in **red** and neutral in **white**

Our loess method estimates hits before smoothing

- Statistical outliers are defined using the IQR (inter-quartile range) and given a low weight

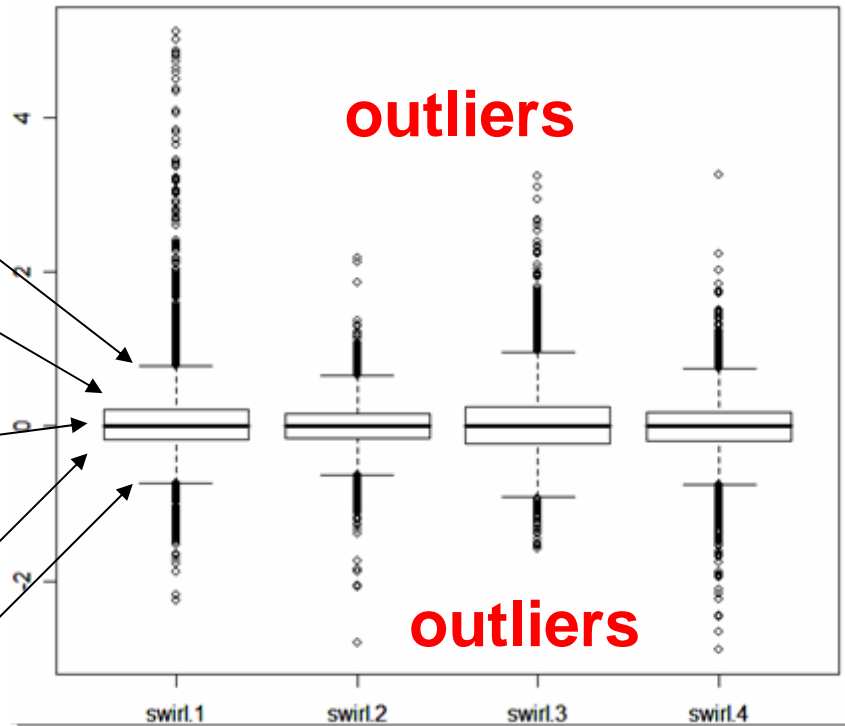
○ maximum

○ upper quartile

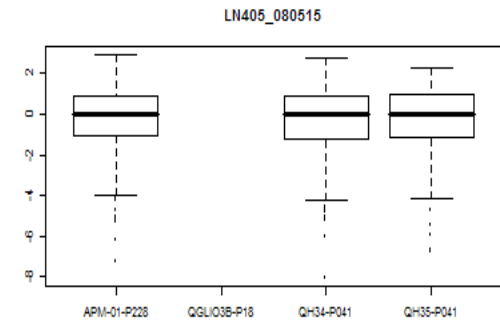
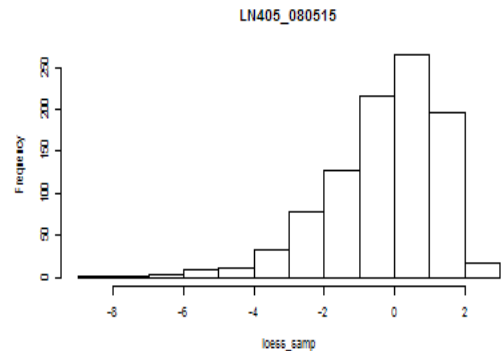
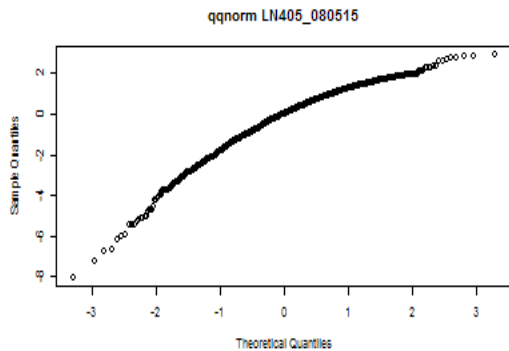
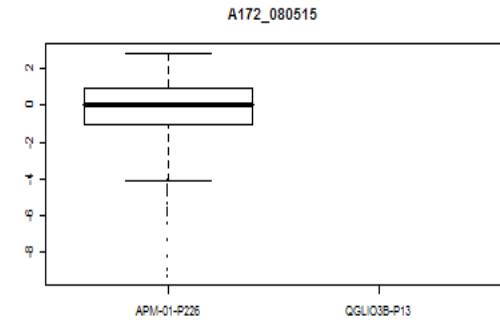
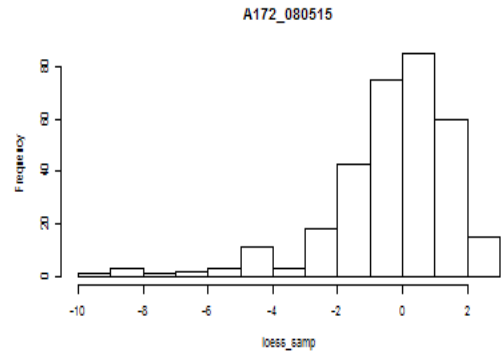
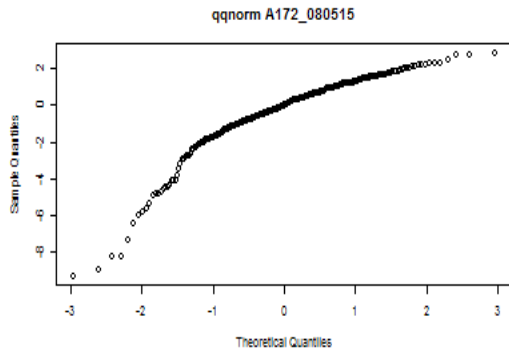
○ median

○ lower quartile

○ minimum



Data distribution plots



qqnorm: values vs.
the normal
distribution

histogram

boxplot: each plate
separately

Hit determination: threshold method

Screen-wise threshold

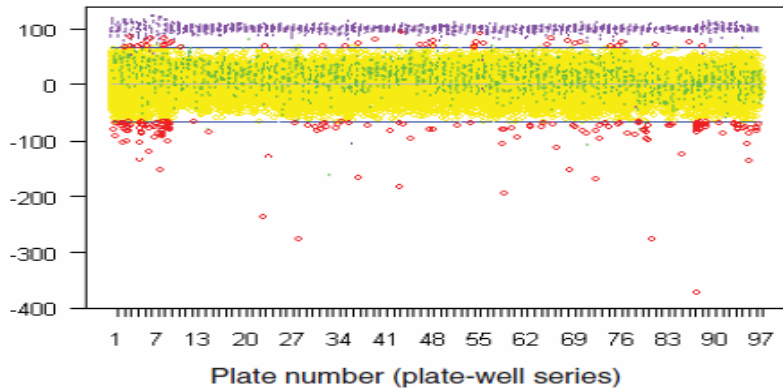
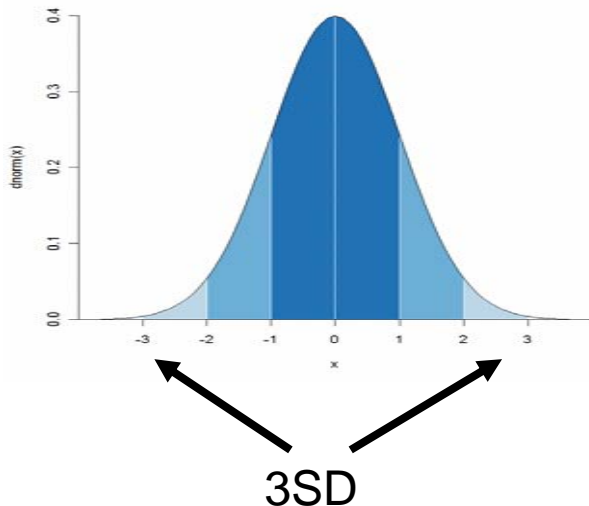
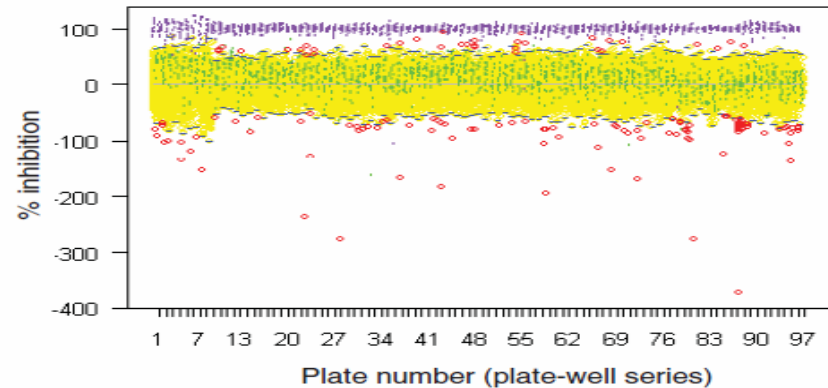
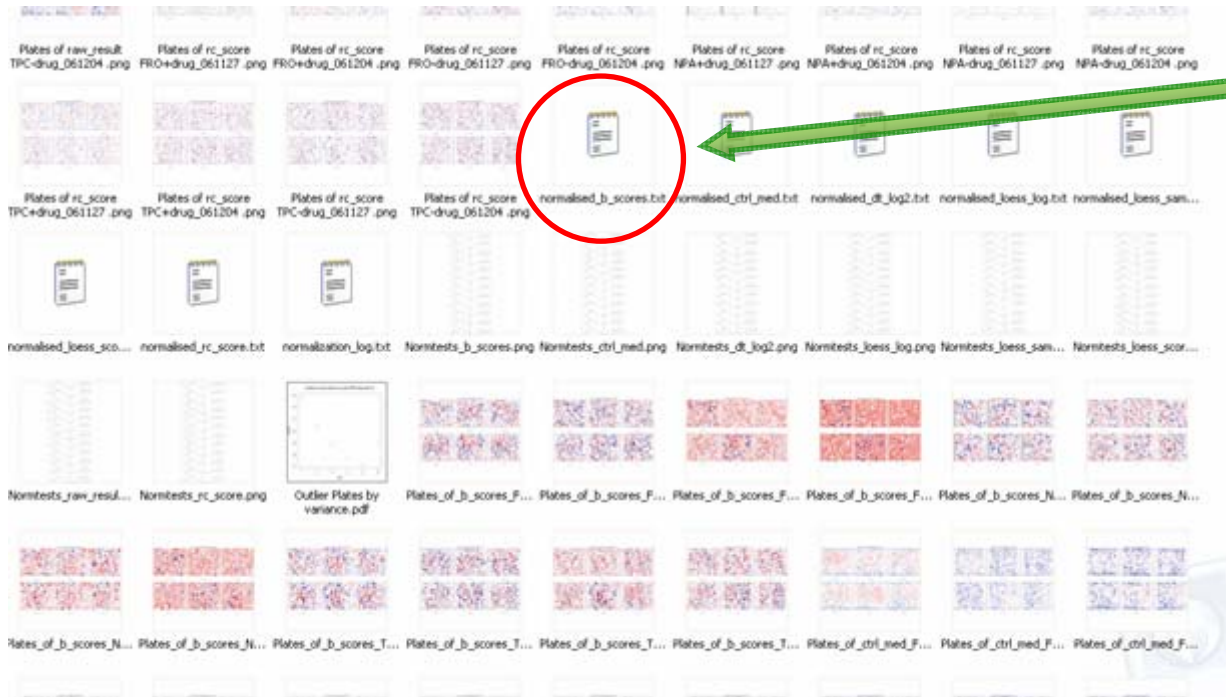


Plate-wise threshold



- A threshold of mean/median \pm 2-3SD is often used to define hits
- If the data is normally distributed then 95% of the data is within 2SD and 99.7% within 3SD (thresholds are only indicative).
- Thresholds are usually chosen so that there is a degree of repeatability in the hits.
- We use screen-wise thresholds but plate-wise can be useful sometimes.

Comprehensive sorted output



One file for each normalisation

Commercial use (Test)

Bedrijfs Formatering - Formateren - Als Tabel - Zelfformateren - Invoegen - Lijsten - Formaat - Sorteren - Zoeken en - en Filteren - Approeven - Formaatopties - Zellen - Bewerken

J	K	L	M	N	O	P	Q
Sense_Seq	Concentratic	drugposi	tion	raw_result	F raw_result	F raw_result	F raw_result
NA	NA	Q-F01A-A01	3559668	4315832	3090684	3526173	3064853
NA	NA	Q-F01A-A02	4333893	4645808	3163903	3797549	3271376
NA	NA	Q-F01A-A03	3675011	4510450	3302924	3596631	3222831
NA	NA	Q-F01A-A04	2029084	2486850	2173496	2666945	2843720
NA	NA	Q-F01A-A05	1840935	2472571	2283971	2927630	3147921
NA	NA	Q-F01A-A06	3203133	3974622	2674907	3148308	3550637
NA	NA	Q-F01A-A07	2067064	2926059	2510961	2598074	3488108
NA	NA	Q-F01A-A08	2559767	2840655	2439316	2649712	2982668
NA	NA	Q-F01A-A09	2677406	3533838	2528064	2557421	3169951
NA	NA	Q-F01A-A10	3772861	4606770	3132475	3885272	3290213
NA	NA	Q-F01A-A11	2674727	3799568	2789043	3296600	2890474
NA	NA	Q-F01A-A12	3154029	3836596	2630271	3003948	3384542
NA	NA	Q-F01A-A13	2378365	3194378	2788184	3007399	3153442
NA	NA	Q-F01A-A14	3179617	4401237	2728186	3228526	3436339
NA	NA	Q-F01A-A15	3484107	4645354	2801174	3574444	3442968
NA	NA	Q-F01A-A16	3460072	3973291	3127566	3437075	3683796
NA	NA	Q-F01A-A17	1968346	2300819	2208886	2511608	2718307
NA	NA	Q-F01A-A18	4093461	4723403	2836209	3581329	3925309
NA	NA	Q-F01A-A19	3399363	4355329	2892097	3339731	3660657
NA	NA	Q-F01A-A20	3830726	4442844	2677046	3420570	3772809
NA	NA	Q-F01A-A21	3184678	3886036	2813497	3296633	3622462
NA	NA	Q-F01A-A22	2459788	2774145	2338583	2609873	3648431
NA	NA	Q-F01A-A23	4282221	4667030	2945844	3949421	3672942
NA	NA	Q-F01A-A24	3805735	4472206	2879114	3701845	3423448
NA	NA	Q-F01A-B01	4363115	4547557	3186734	3813265	3582383
NA	NA	Q-F01A-B02	4889654	4757700	3707347	4280839	4124448
NA	NA	Q-F01A-B03	4023056	4531411	3521830	4096615	4009281
NA	NA	Q-F01A-B04	5084303	4910520	4117945	4633683	4346633

10	Q-F01A-A09	Q006486	BMP2	bone morph	659	NA	AAGCACCGAA9	NA	NA	NA	Q-F01A-A09
11	Q-F01A-A10	Q006568	EPHB3	EPH receptor	2049	ENSG000001	CCGCACTG.A10	NA	NA	NA	Q-F01A-A10
12	Q-F01A-A11	Q006494	CALM2	calmodulin 2	805	ENSG000001	ACAAAAGGAA11	NA	NA	NA	Q-F01A-A11
13	Q-F01A-A12	Q006576	FER	fer (fps/Yes r	2241	ENSG000001	CAGAAACAAC12	NA	NA	NA	Q-F01A-A12
14	Q-F01A-A13	Q006502	CDC2L1	cell division	984	ENSG000001	CAAGATCTAA13	NA	NA	NA	Q-F01A-A13
15	Q-F01A-A14	Q006584	FLT3	fms-related	2322	ENSG000001	CCGGCTTGAA14	NA	NA	NA	Q-F01A-A14
16	Q-F01A-A15	Q006510	CDK9	cyclin-deper	1025	ENSG000001	AACCGCTGC.A15	NA	NA	NA	Q-F01A-A15
17	Q-F01A-A16	Q006592	GCK	glucokinase	2645	ENSG000001	CAGGACTTTA16	NA	NA	NA	Q-F01A-A16
18	Q-F01A-A17	Q006518	CHEK1	CHK1 checkp	1111	ENSG000001	AAGAAAGAA17	NA	NA	NA	Q-F01A-A17
19	Q-F01A-A18	Q006600	GSK3B	glycogen tyr	2932	ENSG000001	AACACTGGTA18	NA	NA	NA	Q-F01A-A18
20	Q-F01A-A19	Q006526	CKS1B	CDC28 prose	1163	ENSG000001	AACATCTTCA19	NA	NA	NA	Q-F01A-A19
21	Q-F01A-A20	Q006608	IRAK1	interleukon-	3654	ENSG000001	CCGGGCAAT.A20	NA	NA	NA	Q-F01A-A20
22	Q-F01A-A21	Q006534	CSK	c-src tyrosin	1445	ENSG000001	CCGGTACAG.A21	NA	NA	NA	Q-F01A-A21
23	Q-F01A-A22	Q006616	JAK3	janus kinase	3718	ENSG000001	AGGGCTTAA.A22	NA	NA	NA	Q-F01A-A22
24	Q-F01A-A23	C	C	C	NA	NA	A23	NA	NA	NA	Q-F01A-A23
25	Q-F01A-A24	C	C	C	NA	NA	A24	NA	NA	NA	Q-F01A-A24
26	Q-F01A-B01	C	C	C	NA	NA	B1	NA	NA	NA	Q-F01A-B01
27	Q-F01A-B02	C	C	C	NA	NA	B2	NA	NA	NA	Q-F01A-B02
28	Q-F01A-B03	Q006626	LYN	v-yes-1 Yamu	4067	ENSG000001	CCGGGACGA.B3	NA	NA	NA	Q-F01A-B03
29	Q-F01A-B04	Q006798	DRK4L1	ncrtein kin	3567	ENSG000001	CCACGATA.B4	NA	NA	NA	Q-F01A-B04

"Customer" is provided with an interactive Excel file

normalised_b_sc-loess_log_Mariolina_081112.xls [Kompatibilitätsmodus] - Microsoft Excel nichtkommerzielle Verwendung (Test)

	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
1								Hit determination														
2																						
3								threshold		2												
4	-0,0987	0,024499	-0,12362	-0,02532	-0,21585	0,10931	-0,08034															
5	0,16283	0,102322	0,176154	0,255625	0,336017	0,205366	0,274998															
6	0,226959	0,229143	0,228688	0,485932	0,456186	0,520041	0,469658	0	0	22	2	5	0	31	6	1	5	42	10	0	2	
7	-0,42436	-0,18014	-0,47593	-0,53657	-0,88788	-0,30142	-0,63033	178	352	92	138	170	296	109	147	178	117	107	111	163	65	
8																						
9	NPA-drug	NPA-drug	NPA-drug	TPC-drug	TPC-drug	TPC-drug	TPC-drug	FRO-drug	FRO-drug	FRO-drug	FRO-drug	NPA-drug	NPA-drug	NPA-drug	NPA-drug	TPC-drug	TPC-drug	TPC-drug	TPC-drug	FRO-drug	FRO-drug	NPA-drug
10	-0,09071	NA	-0,12857	-0,36465	-0,35375	0,091648	0,011059	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
11	-0,01362	NA	-0,00037	0,219017	-0,03628	0,171273	0,132433	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
12	-0,1485	0,01919	-0,25296	0,101073	-0,34796	0,160446	-0,03271	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
13	-0,44311	-0,04835	-0,37193	-0,2403	-0,44811	0,015021	-0,19546	DOWN	DOWN	--	--	DOWN	DOWN	--	--	--	--	--	--	DOWN-HIT	--	DOWN-H
14	-0,09534	-0,06063	0,014616	-0,6265	-0,66339	-0,44307	-0,28726	DOWN	DOWN	--	--	--	--	DOWN	--	DOWN	--	DOWN	--	DOWN-HIT	--	--
15	-0,4	0,032556	-0,25895	0,035948	-0,39746	-0,01396	-0,22583	--	--	--	--	--	--	--	--	--	--	--	--	--	DOWN-HIT	--
16	-0,10826	-0,02191	-0,01989	-0,12738	-0,23267	-0,24031	-0,2411	DOWN	DOWN	--	--	--	--	--	--	--	--	--	--	--	DOWN-HIT	--
17	-0,28993	-0,14374	-0,39947	-0,17613	-0,48286	-0,12851	-0,26116	--	DOWN	--	--	DOWN	--	--	--	--	--	--	--	--	--	--
18	-0,37736	-0,14953	-0,32317	-0,2223	-0,40611	0,147191	-0,05633	--	--	DOWN	--	--	--	--	--	--	--	--	--	--	--	--
19	-0,02255	-0,04856	-0,07635	0,140991	0,018134	0,197243	0,220214	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
20	-0,26648	-0,10606	-0,19592	-0,33953	-0,26887	-0,27206	-0,30608	--	--	DOWN	--	--	--	--	--	--	--	--	--	--	--	--
21	-0,17008	-0,10863	-0,07541	0,03048	-0,73271	-0,0152	-0,32656	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
22	-0,3355	-0,16138	-0,3762	-0,46071	-0,49473	-0,30057	-0,27584	--	DOWN	--	DOWN	--	--	--	--	--	--	--	--	--	--	--
23	-0,08208	0,001969	-0,01669	0,099678	-0,42711	0,14456	-0,00031	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
24	-0,06341	-0,06313	-0,08508	-0,14102	-0,17974	-0,06351	-0,03956	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
25	-0,09775	0,026883	-0,1346	-0,21133	-0,21814	-0,03478	-0,19543	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
26	-0,84592	-0,03242	-0,539	-0,49459	-0,38845	-0,25242	-0,24756	DOWN	DOWN	--	--	DOWN	DOWN	--	DOWN	--	--	--	--	DOWN-HIT	--	DOWN-H
27	0,057469	0,036777	-0,03315	-0,01886	-0,03731	-0,0685	0,001807	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
28	-0,22962	0,035306	-0,1846	-0,07076	-0,47476	-0,06777	-0,44001	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
29	-0,11479	-0,02401	-0,28245	0,255064	0,1344	0,017521	0,101295	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
30	-0,12642	0,079621	0,0627	-0,14678	-0,47913	0,08075	-0,1364	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
31	-0,42313	-0,09766	-0,31933	-0,08438	-0,67137	-0,06329	-0,27281	DOWN	DOWN	--	--	--	--	--	--	--	--	--	--	DOWN-HIT	--	--
32	-0,04422	0,064042	-0,01261	0,009063	-0,00587	0,147126	0,00625	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

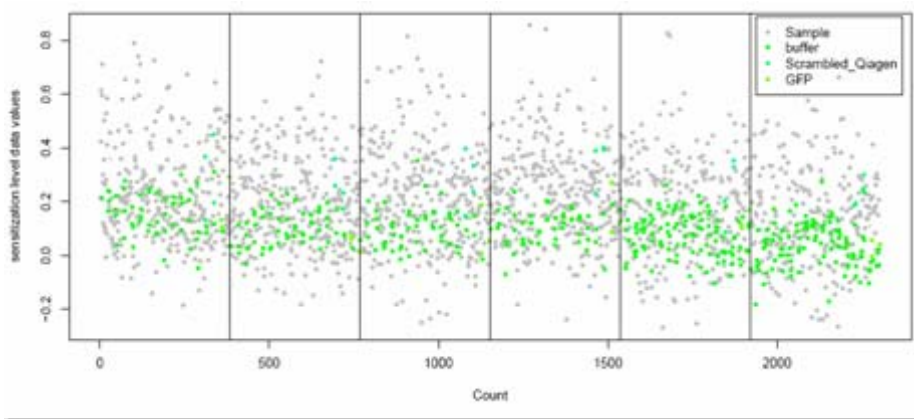
Can be generated automatically with R.

Criteria: Sensitization hits

- Two screens, each with and without drug.
- Difference of '+drug' and '-drug' of each screen:
 $\text{diff1} = 'S1_+\text{drug}' - 'S1_-\text{drug}'$ (resp. for diff2)
- **Plate-based** normalisation of the sensitisation score by dividing the median score of the plate
- **Screen-based** (across-plates) normalisation using Quantile normalisation (limma)
- **Sensitisation score** = $\text{mean}(\text{diff1}, \text{diff2})$
- P-values for the probability that the sensitisation score lies within the 'normal-like' set of the score distribution (outliers)
- Q-values using FDR (false discovery rate) method for multiple testing correction (qvalue)

Plate series plots to illustrate steps...

raw data 1



raw data 2

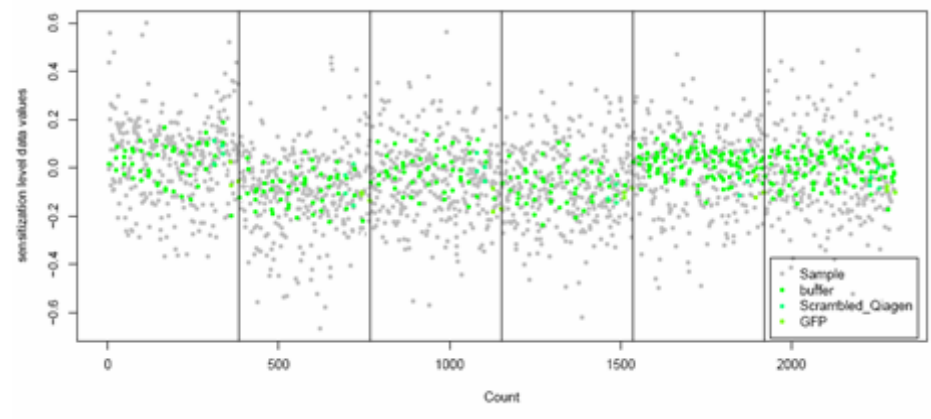


plate-normalised 1

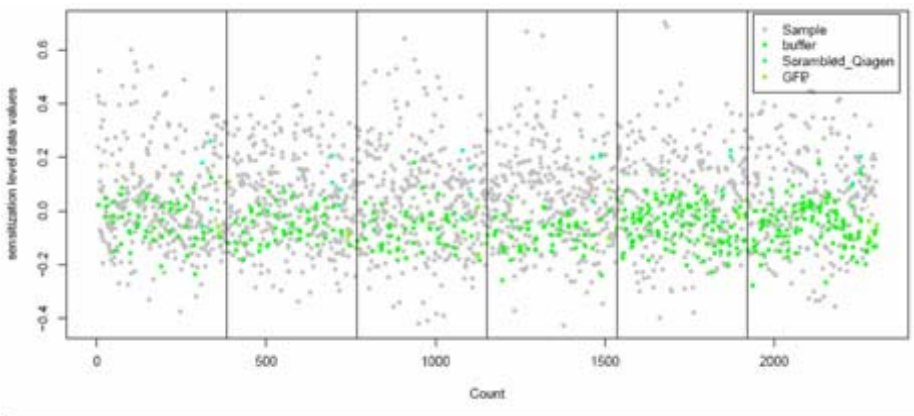
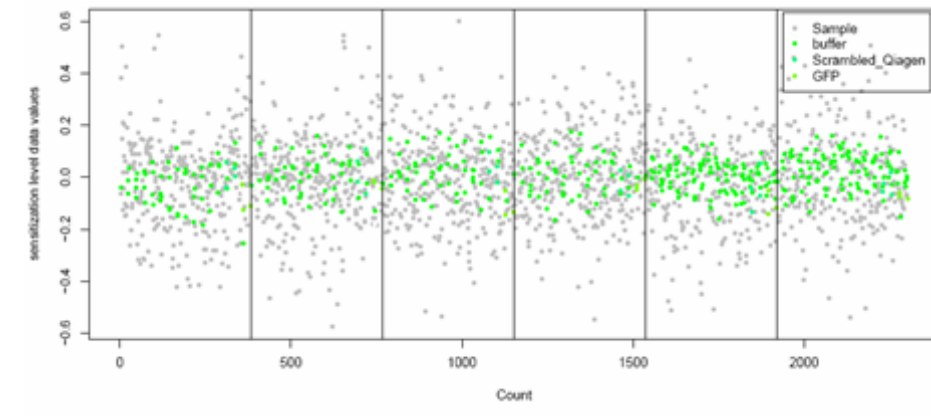


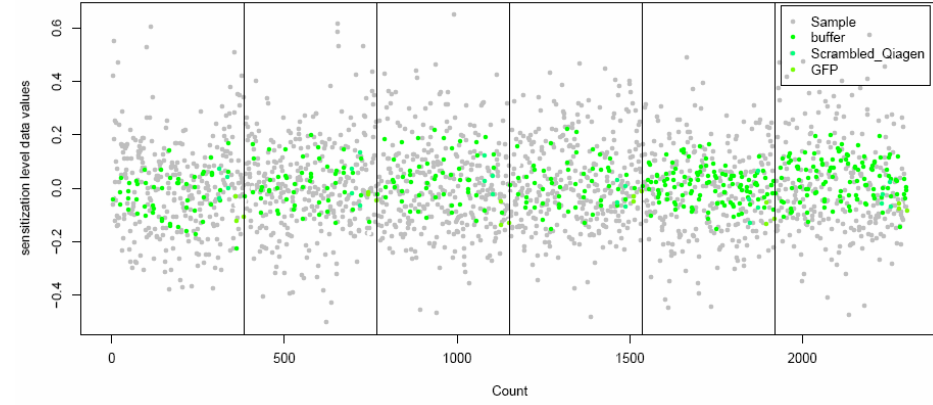
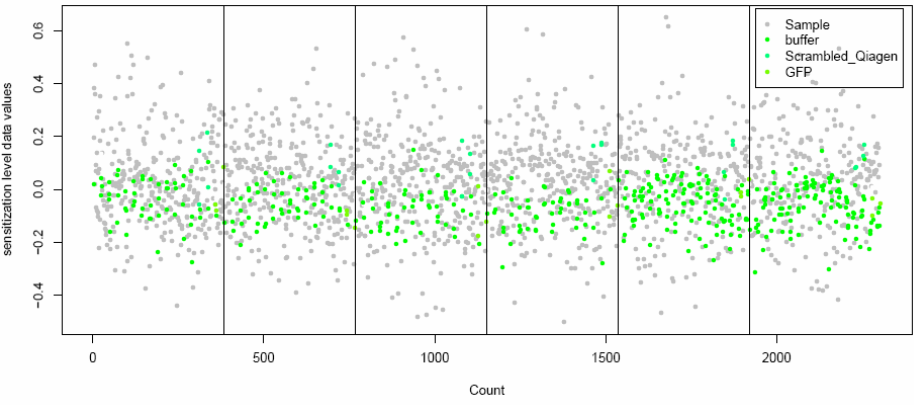
Plate-normalised 2



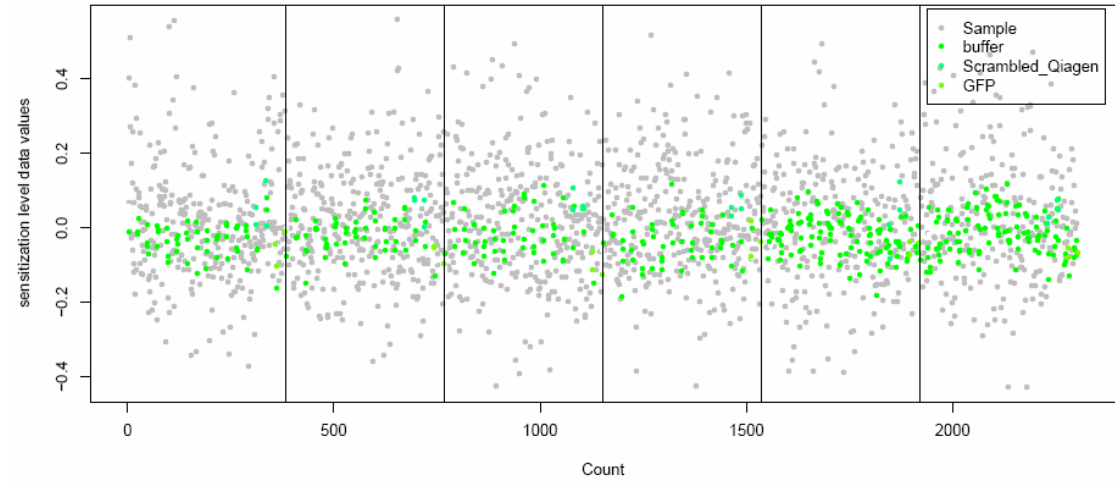
To the sensitization score

quantile-normalised 1

quantile-normalised 2

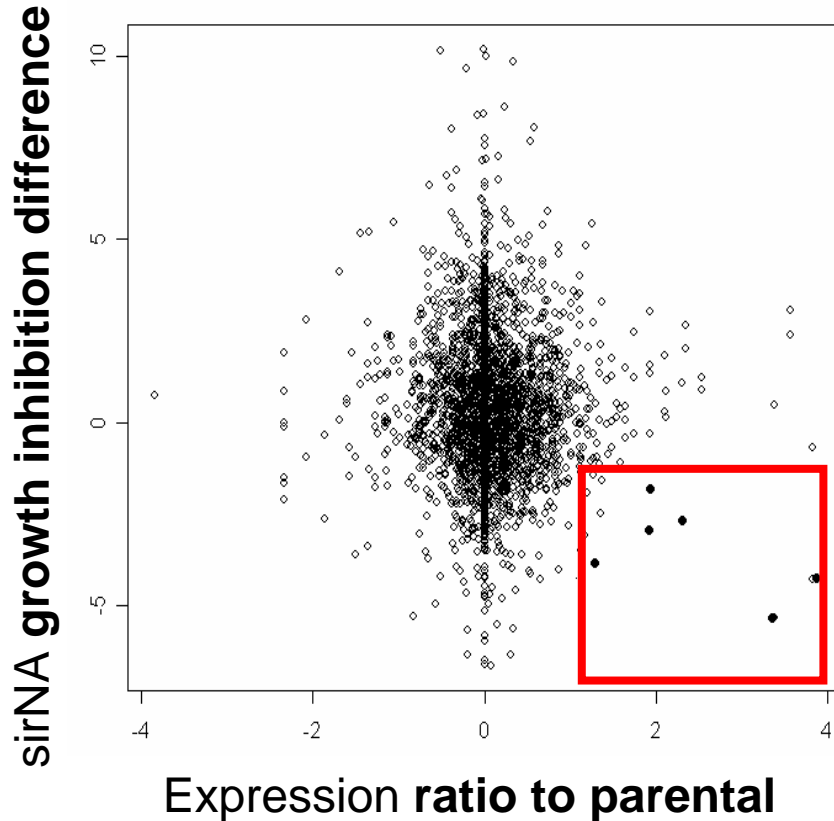


Sensitisation score



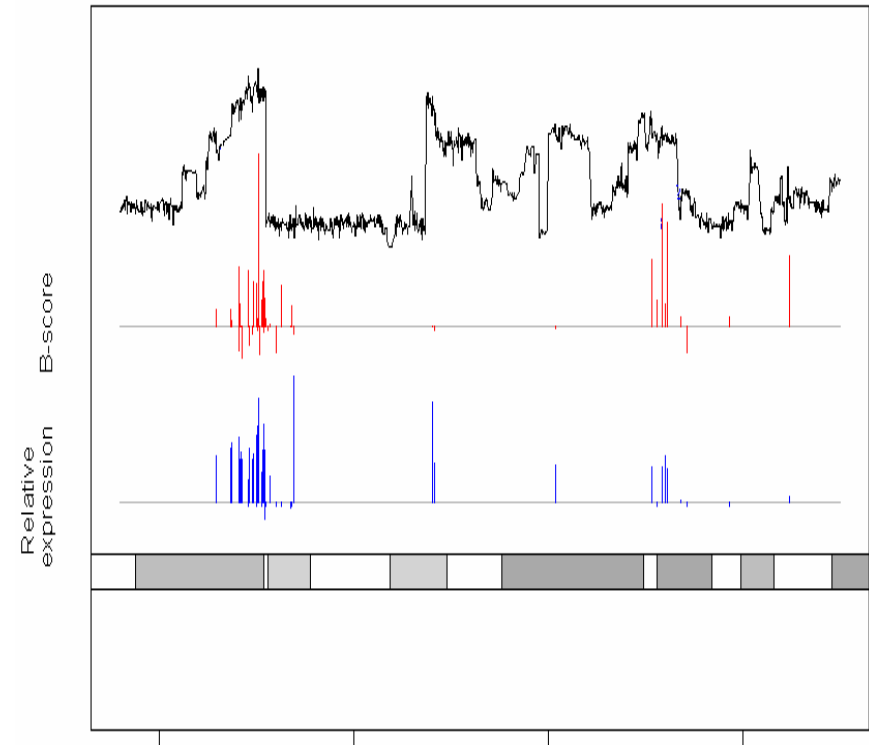
Integration of data from other sources

Two cell lines: GE+siRNA



Increased gene expression
and greater siRNA growth
inhibition

One cell line: GE+siRNA+aCGH



Gene amplification, siRNA
growth inhibition and gene
expression increase

by Henrik Edgren

High Throughput Screening Database: Multiple Assays of the same Model System

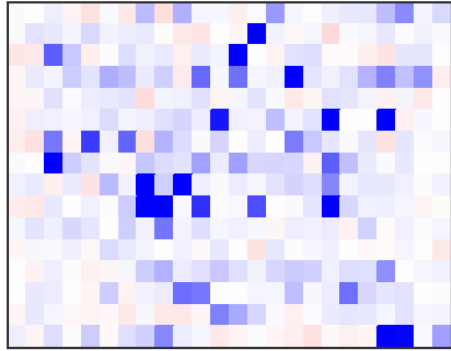
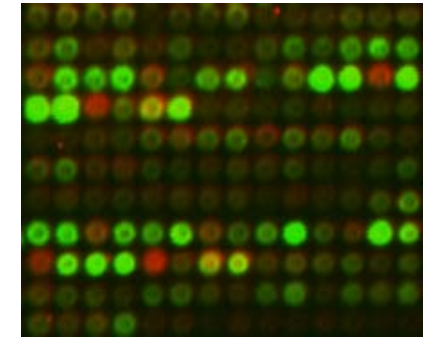


Plate based:

- CTB
- CellTiter-Glo™
- ApoOne™
- luciferase assays



Lysate arrays:

- up to 3 channels
- multiple endpoints
- use of ratios

Cell Arrays:

- up to 5 channels
- uHTS (10000's)
- improved repeatability
- use ratios for normalization

Supporting Data:

- gene expression
- aCGH
- miRNA expression

References:

- siRNA statistics:
 - Malo N, Hanley et al. Statistical practice in high-throughput screening data analysis. *Nat Biotechnol.* (24)2:16775, 2006.
 - Zhang et al. Robust statistical methods for hit selection in RNAi high-throughput screening experiments. *Pharmacogenomics*, 7(3): 299, 2006.
 - Zhang et al. A simple statistical parameter for use in evaluation and Validation of Hight-throughput screening assays. *J Biomol Screening*, 4(2):67, 1999.
- Avoiding off-targets
 - Echeverri CJ, Perrimon N. High-throughput RNAi screening in cultured cells: a user's guide. *Nat. Rev. Genet*, 7(5):37384, 2006.
 - Cullen BR, Enhancing and confirming specificity of RNAi experiments. *Nature Methods*, 3(9): 677, 2006.
 - Echeverri CJ, *Nature Methods*, 3(10): 777, 2006.