Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

This document describes the sequential steps that were applied to select the list of primary actives that we suggest for subsequent analysis. Each section contains comments from the person that analyzed the data in bold/violet color font.

Delivered data package structure	
subfolder	content
analysis\<assay_name>	PrimaryScreeningReport.pdf (this file), PlateReports.pdf (graphical representation of the plate data), sampledata.txt (all numeric statistically analyzed data suitable for spreadsheets), primary_actives.sdf (structure data file containing the suggested compounds for cherry picking and further validation).
converter	contains .png files from quality checks during assay data collection
documents	contains the manually written protocol and project application forms
raw_data	contains .txt or .asc files containing the raw data as they come from the instruments

Step1: value type selection and removal of non-numeric results
Following value types where selected for analysis:

measurement_type	annotation
-{0s_1_A3P;0s_1_A2P}	fluorescence increase after P1 addition
statistic	slope NADH fluorescence increase

Of 31960 samples, 19 were found to contain non-numeric values and were discarded. 31941 samples have passed

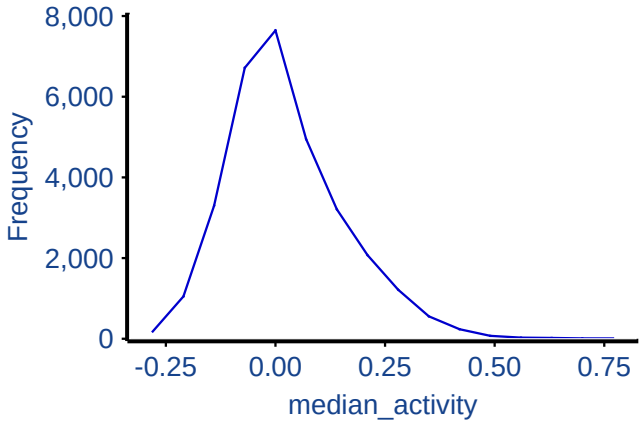
Reasons for non-numeric values:
-> saturation of PMT due to autofluorescent compounds

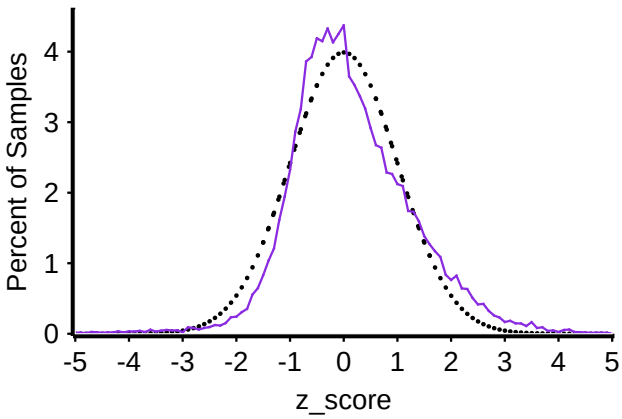
Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

Step2: selection of valid measurements

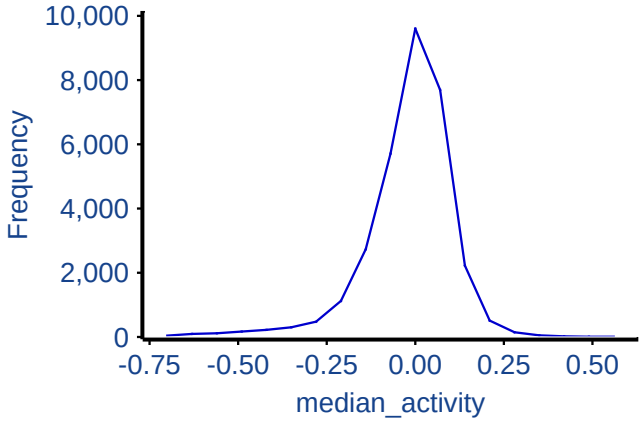
For all samples on a plate and for all selected measurement types, the normalization properties "z_score" and "median_activity" are calculated (see appendix). To be able to define limits for identifying valid measurements, the median_activity and z_score distribution is plotted for each measurement type (the dotted line in the z-score distribution is the normal distribution reference), and the z-score correlation between the different measurement types are plotted.

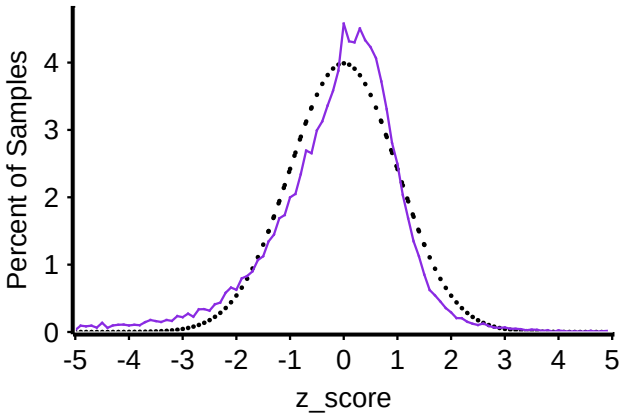
Measurement type: -{0s_1_A3P;0s_1_A2P},



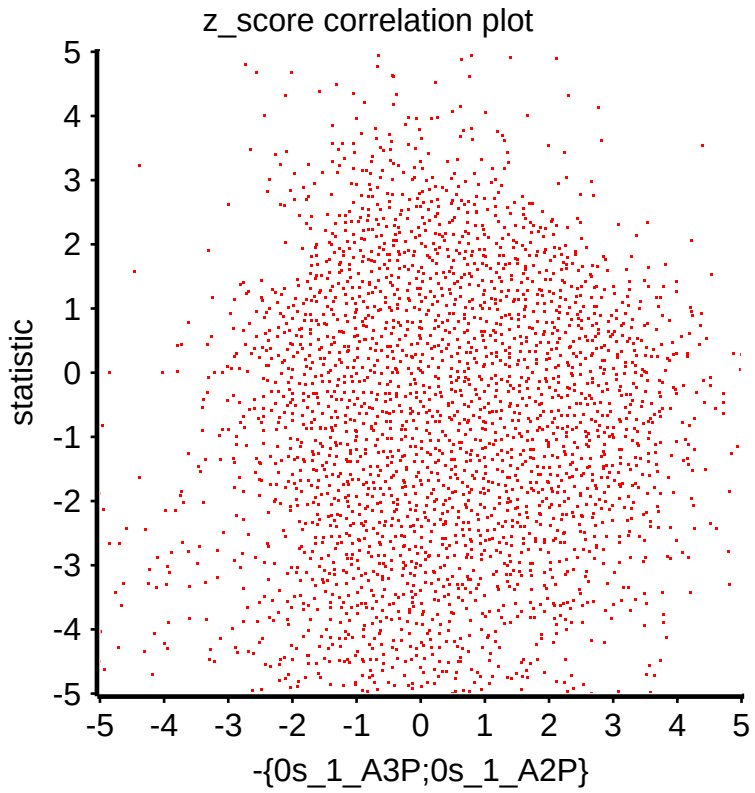


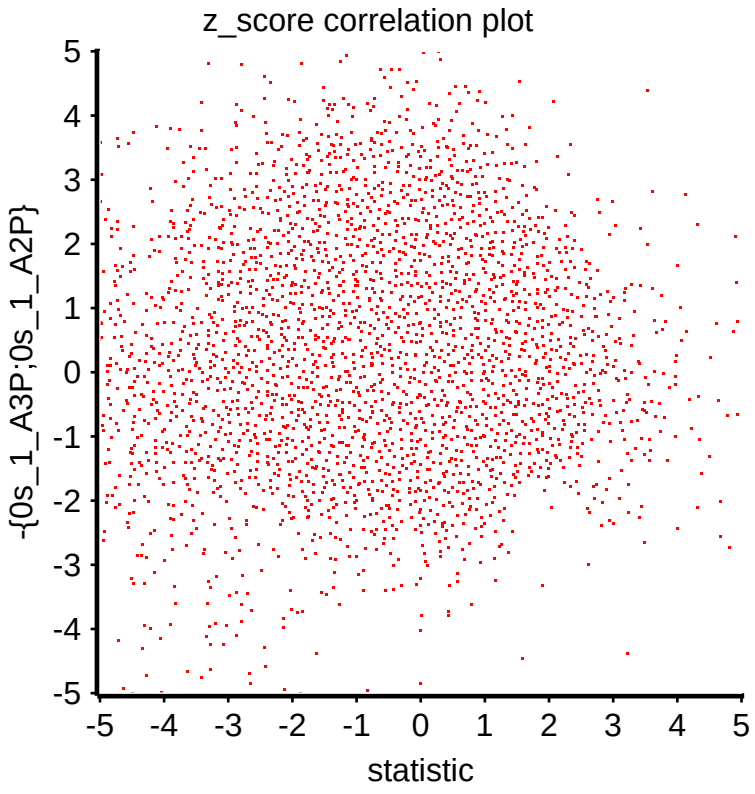
Measurement type: statistic,





Assay Name: T_111031_UOOMuster_JohnDoolittle_P53





Correlation Matrix		
Property	-{0s_1_A3P;0s_1_A2P}	statistic
-{0s_1_A3P;0s_1_A2P}	1	0.040880
statistic	0.040880	1

Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

Reason for choosing specific limits for the selection of valid samples:

-> no more than 25% deviation from median value allowed

Applied Filter Rules	
applied_filter	number_of_samples_failing
-{0s_1_A3P;0s_1_A2P}: z_score	0
-{0s_1_A3P;0s_1_A2P}: median_activity>-0.25 AND<0.25	2864
statistic: z_score	0
statistic: median_activity	0

Of 31941 samples, 29077 pass all of the filters defined above.

The z-scores, median-activity and control-based relative activities are re-calculated for the valid measurements.
Following measurement type was selected for statistic analysis: statistic

Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

Step3: Summary of single plates analytics

Graphical representation of each plate data set is provided in the "PlateReports.pdf" file delivered together with this report. For the classification criteria of the z´-factors, see the literature references in the appendix.

Plate_ID	z_factor	z_factor_classification	accept	plate_comment
111031A1P1001	0.84	excellent	yes	
111031A1P1002	0.84	excellent	yes	
111031A1P1003	0.83	excellent	yes	
111031A1P1004	0.85	excellent	yes	
111031A1P1005	0.81	excellent	yes	
111031A1P1006	0.82	excellent	yes	
111031A1P1007	0.76	excellent	yes	
111031A1P1008	0.82	excellent	yes	
111031A1P1009	0.78	excellent	yes	
111031A1P1010	0.69	excellent	yes	
111031A1P2001	0.83	excellent	yes	
111031A1P2002	0.75	excellent	yes	samples with negative slopes
111031A1P2003	0.79	excellent	yes	
111031A1P2004	0.83	excellent	yes	
111031A1P2005	0.84	excellent	yes	
111031A1P2006	0.88	excellent	yes	
111031A1P2007	0.84	excellent	yes	
111031A1P2008	0.86	excellent	yes	
111031A1P2009	0.82	excellent	yes	
111031A1P2010	0.87	excellent	yes	
111031A1P2011	0.81	excellent	yes	
111031A1P2012	0.74	excellent	yes	
111031A1P2013	0.83	excellent	yes	
111031A1P2014	0.79	excellent	yes	
111031A1P2015	0.65	excellent	yes	
111031A1P2016	0.79	excellent	yes	
111031A1P2017	0.77	excellent	yes	
111031A1P2018	0.78	excellent	yes	
111031A1P2019	0.75	excellent	yes	
111031A1P2020	0.64	excellent	yes	
111031A1P2021	0.80	excellent	yes	
111031A1P2022	0.80	excellent	yes	
111031A1P2023	0.82	excellent	yes	
111031A1P2024	0.74	excellent	yes	
111031A1P2025	0.80	excellent	yes	
111031A1P2026	0.81	excellent	yes	
111031A1P2027	0.75	excellent	yes	
111031A1P2028	0.75	excellent	yes	
111031A1P2029	0.80	excellent	yes	
111031A1P2030	0.85	excellent	yes	
111031A1P2031	0.71	excellent	yes	
111031A1P2032	0.76	excellent	yes	

Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

111031A1P2033	0.79	excellent	yes	
111031A1P2034	0.76	excellent	yes	
111031A1P2035	0.66	excellent	yes	
111031A1P2036	0.82	excellent	yes	
111031A1P2037	0.73	excellent	yes	
111031A1P2038	0.67	excellent	yes	
111031A1P2039	0.63	excellent	yes	
111031A1P2040	0.66	excellent	yes	
111031A1P2041	0.82	excellent	yes	
111031A1P2042	0.80	excellent	yes	
111031A1P2043	0.85	excellent	yes	
111031A1P2044	0.88	excellent	yes	
111031A1P2045	0.84	excellent	yes	
111031A1P2046	0.83	excellent	yes	
111031A1P2047	0.73	excellent	yes	
111031A1P3001	0.83	excellent	yes	
111031A1P3002	0.76	excellent	yes	
111031A1P3003	0.83	excellent	yes	
111031A1P3004	0.76	excellent	yes	
111031A1P3005	0.76	excellent	yes	
111031A1P3006	0.76	excellent	yes	
111031A1P3007	0.69	excellent	yes	
111031A1P3008	0.69	excellent	yes	
111031A1P3009	0.72	excellent	yes	
111031A1P3010	0.79	excellent	yes	
111031A1P3011	0.78	excellent	yes	
111031A1P3012	0.76	excellent	yes	
111031A1P3013	0.62	excellent	yes	
111031A1P4001	0.78	excellent	yes	
111031A1P4002	0.84	excellent	yes	
111031A1P4003	0.87	excellent	yes	
111031A1P4004	0.83	excellent	yes	
111031A1P4005	0.86	excellent	yes	
111031A1P4006	0.80	excellent	yes	
111031A1P4007	0.80	excellent	yes	
111031A1P4008	0.80	excellent	yes	
111031A1P4009	0.79	excellent	yes	
111031A1P4010	0.68	excellent	yes	
111031A1P4011	0.70	excellent	yes	
111031A1P4012	0.42	poor	yes	
111031A1P5001	0.75	excellent	yes	
111031A1P5002	0.78	excellent	yes	
111031A1P5003	0.76	excellent	yes	
111031A1P5004	0.83	excellent	yes	
111031A1P5005	0.72	excellent	yes	
111031A1P6001	0.77	excellent	yes	

Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

111031A1P6002	0.82	excellent	yes	
111031A1P6003	0.86	excellent	yes	
111031A1P6004	0.78	excellent	yes	

Average Z´Factor: 0.78 ± 0.071

General comments about the plate data:

-> some samples visible with negative slopes

Of 91 plates, 91 were accepted.

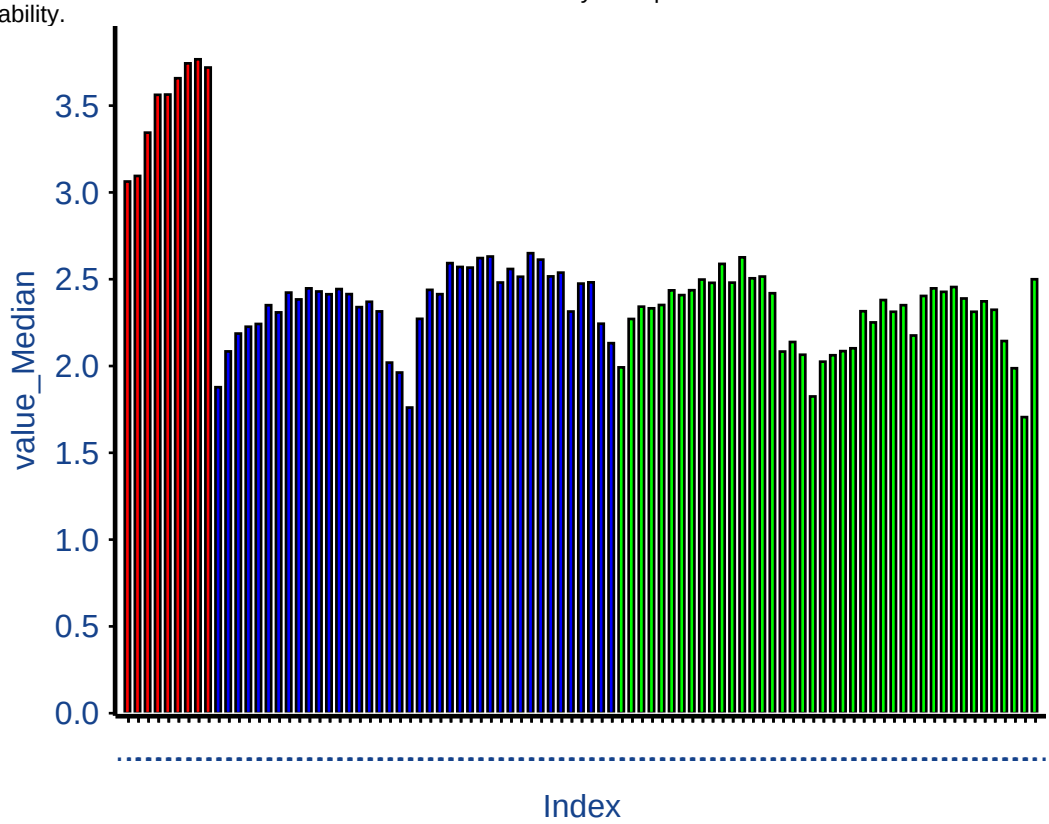
Of 29077 samples, 29077 samples passed visual inspection of the plate data.

At this point, after visually checking the plausibility of the plate data, the "sampledata.txt" file containing the sample and control values and its normalized values is written. This file can be used for your own data mining purposes.

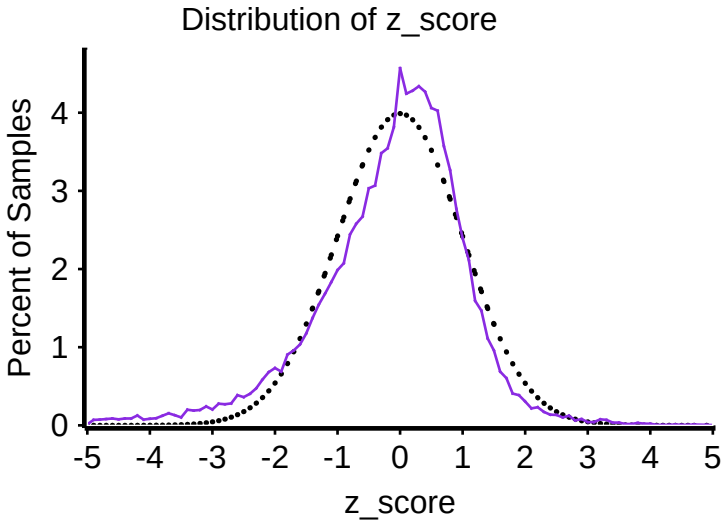
Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

Step4: Statistic analysis of valid numeric results

The median of the sample signals on each plate are plotted against the sequence of plates (plates sorted by measurement time). A different color is used for each different measurement day. The plot shows trends for batch to batch variations and component stability.



The z-score distribution of the valid samples shows whether the systematic variations between plates could be removed by normalization and gives hints on where to set the filter limits for selection of actives based on z-score. The dotted line represents an ideal normal distribution for comparison



Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

Step5: Final selection of primary actives

-> tailing towerds lower z-scores visible, z-score limit increased from 3 to 5. To exclude samples with negative slopes, rel_activ must be larger than -0.1

Applied Filter Rules	
applied_filter	number_of_samples_failing
statistic: z_score<-5	28527
statistic: median_activity	0
statistic: rel_activ>-0.1	125

Of 29077 samples, 28652 samples fail in at least one of the filters defined above, leaving 425 samples.

Then, the samples were sorted ascending by rel_activ and the top 352 samples were kept, giving the final list of 352 selected actives.

The chemical structures, normalized data and links to PubChem can be found in the file "primary_actives.sdf"

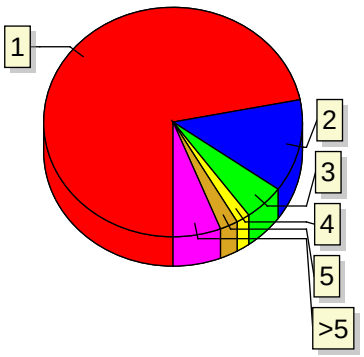
Assay Name: T_111031_UOOMuster_JohnDoolittle_P53

Step6: Characterisation of the primary actives list

The basic statistics table analyzes the normalised properties z_score, median_activities and control based relative activities. Use this table to check whether the final selection criteria where reasonable (especially sorting order)

basic statistics of the primary actives list	
Property	value
z_score_Mean	-7.62
z_score_Min	-16.00
z_score_Max	-5.00
median_activity_Mean	-0.69
median_activity_Min	-1.08
median_activity_Max	-0.45
rel_activ_Mean	0.29
rel_activ_Min	-0.08
rel_activ_Max	0.52

Using Tanimoto similarity search (Tanimoto cut-off at 0.6 using the ECFP_4 Fingerprint, see references table at the end of this document), the number of related structures within the primary actives list is determined for each compound sample, and a summary saying how many compounds are in the list with how many related structures is displayed in the pie chart and its accompanying table. Since similar structure should lead to similar activities, appearance of structures with related structures is a sign for success of the screening process



NumberOfCompoundsThatHave	NumberOfRelatedStructures
253	1
46	2
18	3
6	4
8	5
21	>5

Summary and Outlook:

with the residual samples, proceed to IC50 determination and include the counterscreen for coupling enzyme inactivation

Appendix. Property legends

Properties of the sampledata.txt file	
Property Name	Legend
Comp_ID	The internal compound identifier used by the screening unit
Plate_ID	The plate barcode
concentration_uM	The concentration of the compound samples
Replica_ID	The number of the replica
Lib_ID	The 4-digit library ID identifies a specific set of 352 compounds
SampleType	"sample" for compound samples, "pos_Co" for high-value controls, "neg_Co" for low-value controls, and "custom_CoX" for other control samples
Well_ID	The alphanumeric well location code
value	The measured raw data value
value_description	The measurement type description
measurement_date	Date and time when the plate was measured
Instrument_ID	The serial number of the instrument that was used to record the data
value_Median	The median of the measured raw data values, determined for each SampleType separately on each plate
value_MAD	The median absolute deviation, determined for each SampleType separately on each plate
z_score	The standard score (z-score), formula: (value - value_Median)/(1.48258*MAD). The distance to the median in units of standard deviation. -4 for example means that the signal is 4 standard deviations lower than the median of the other signals on the plate.
median_activity	The median activity, formula: (value - value_Median)/value_Median. The distance to the median in units of signal median. A value of -0.5 for example means: 50% lower than the median of the other signals on the plate.
rel_activ	The relative activity compared to the control samples. 1: like positive control 0: like negative control
rel_activ_type	Depending on the availability of controls, the calculation of the relative activity is adapted. If both positive and negative controls are present, the value_Median of the positive controls equals rel_activ = 1, the value_Median of the negative controls equals rel_activ = 0 ("relative to positive and negative controls"). If the negative controls are missing, the value_Median of the missing negative controls is assumed to be 0 ("relative to positive control"). If the positive controls are missing, the value is divided by the value_Median of the negative controls ("relative to negative control").
z_factor	The z´factor calculated using positive and negative control samples for this plate. Formula: z_factor = 1 - 3*1.48258*(pos_Co_MAD + neg_Co_MAD)/Abs(pos_Co_Median - neg_Co_Median).
accept	The judgement based upon the graphical platemap whether the plate data can be accepted
plate_comment	Comments about the plate data quality entered by the user

The sampledata.txt file can be analyzed using common spreadsheet software. Make sure the decimal delimiter is set to point, not to comma.

Additional properties of the PrimaryActives.sdf file	
Property	Legend
Origin	The Supplier/Donator of the compound
ID_Number	The Supplier ID/Donator Labjournal page number
NumberOfClosest	The total number of compounds with similar structure within the primary actives list
ClosestNames	The Comp_ID identifiers of the compounds with similar structure
PubChem_CID	The PubChem compound identifier
PubChem_Link	An http link to PubChem for this compound - copy and paste to your browser

The PrimaryActives.sdf file can be analyzed for example using the freeware "MarvinView"

Literature references	
Reference	Topic
A simple statistical parameter ffor use in evaluation and validation of high throughput screening assays. Zhang JH, Chung TD, Oldenburg KR. J Biomol Screen. 1999;4(2):67-73.	Describes the theory of the z´factor assay evaluation
Improved statistical methods for hit selection in high-throughput screening.Brideau C, Gunter B, Pikounis B, Liaw A.J Biomol Screen. 2003 Dec;8(6):634-47.	Describes the advantages of using robust statistics for HTS analysis
Design of chemical libraries with potentially bioactive molecules applying a maximum common substructure concept.Lisurek M, Rupp B, Wichard J, Neuenschwander M, von Kries JP, Frank R, Rademann J, Kühne R.Mol Divers. 2010 May;14(2):401-8. Epub 2009 Aug 15.	Describes the design principle of the CBB1 part of the FMP compound library.
High-throughput screening follow-up using extended-connectivity fingerprints with laplacian-modified bayesian analysis in high-throughput screening follow-Up, David Rogers, Robert D. Brown and Mathew Hahn, J Biomol Screen 2005; 10; 682	Describes the ECFP and FCFP fingerprints for structure similarity determination