Data analysis approaches in high throughput screening

Bei Li

Technical Journal Club, 8th April 2014

Overview

❖ **Part-1: introduction of the main aspects of HTS data analysis**

❖ **Part-2: data analysis approaches in RNAi HTS**

❖ **Part-3: case studies with two papers**

Role of statistics in HTS design

Normalization

QC methods

"Hit" selection methods

A typical HTS process

Informatics

- Performance validation •
- Signal-to-noise, signal-to-background, •
  signal window, assay variability ratio,
  Z'-factor, SSMD

- Systematic error detection •
- Data normalization •
- Experiment-, plate-, well-wise QC •
- "Hit" selection •
- Annotation •

- Structure-activity-relations, scaffold clustering •
  (small molecule)
- Pathway analysis, clusterograms, dendrograms •
  (RNAi)

Target Identification

Assay Development

Primary screen

Confirmatory Screen

Hit Follow-up

Application

- Assay read-out: fluorescence, luminescence or absorbance
- Positive, negative and reference controls
- Cell density, media conditions (cell-based)
- Protein concentration, buffer conditions (biochemical)
- Incubation times, signal stability, assay tolerance to solvent

- Test sample: pooled or individual siRNA or compound
- Screen at single copy

- Deconvolution: 1 siRNA/well
  (when pooled siRNAs are used in primary screen)
- Dose-response (small molecule)
- Screen in replicates

- Small scale screening
- On-bench follow-up experiments

Robust statistics: data distribution

- One of the main assumptions when analyzing HTS data is that the data is **normally distributed**. Or it complies with **the central limit theorem**: under certain (fairly common) conditions, the sum of many random variables will have an approximately normal distribution.

- **Log transformations** are often applied to the data in the pre-processing stage to achieve more symmetrically distributed data around the mean as in a normal distribution.
  - Especially for cell growth assays: to represent the relationship between variables in a more linear way; to make an efficient use of the assay quality assessment parameters.

In HTS practices, the presence of outliers is generally experienced.

- **Outliers**: data points that do not fall within the range of the rest of the data.
- Outliers in control wells can be easily identified, but in the test sample may be misinterpreted as real "hits" instead of random errors.

Two approaches for statistical analysis of data sets to remove outliers:

1) **Classical statistics**: use the truncated (or trimmed) mean, a statistical measure of central tendency.

2) **Robust statistics**: use the median and median absolute deviation (MAD).

$$\text{Median} \pm k\ \text{MAD}, \text{ where } \text{MAD} = \text{median}_i (\ |X_i\text{-median}_j(X_j)|\ )$$

- K is a constant scale factor depending on the distribution. For normally distributed data, K is taken to be $1/(\Phi^{-1}(3/4)) \approx 1.4826$
- more popular in HTS data analysis for its insensitivity to outliers.
- robust to outliers in RNAi data.

Part-1: Introduction of the main aspects of HTS data analysis

Role of statistics in HTS design

Normalization

QC methods

"Hit" selection methods

Normalization for assay variability

Assay variability: data variability between plates due to random fluctuations in assay performance, which is separated from the systematic errors that can be linked to a known reason.

Two main approaches for normalization of HTS data:

1) **Controls-based**: positive and negative controls are used as the upper (100%) and lower (0%) bounds of the assay activity. The activities of the test samples are calculated with respect to these values.

2) **Non-controls-based**: most of the samples are assumed to be inactive in order to serve as their own "negative controls".

   ? However, may be misleading when the majority of the wells in a plate consist of true "hits" (e.g., screening a library of bioactive molecules) or siRNAs (e.g., focused library).

Normalization for assay variability

Controls-based

Percent of control
$$PC = \frac{S_i}{mean(C)} \times 100$$
Normalized percent inhibition
$$NPI = \frac{mean(C_{high}) - S_i}{mean(C_{high}) - mean(C_{low})} \times 100$$

The normalization method that fits best with one's experimental results should be carefully chosen to perform efficient "hit" selection with low FDR.

Non-controls-based

Percent of samples	Z-score
$$PS = \frac{S_i}{mean(S_{all})} \times 100$$	$$Z\text{-score} = \frac{S_i - mean(S_{all})}{std(S_{all})}$$
Robust percent of samples	Robust Z-score
$$RPS = \frac{S_i}{median(S_{all})} \times 100$$	$$\text{Robust Z-score} = \frac{S_i - median(S_{all})}{MAD(S_{all})}$$ $$MAD(S_{all}) = 1.4826 \times median(|S_i - median(S_{all})|)$$

Role of statistics in HTS design

Normalization

QC methods

"Hit" selection methods

Quality control methods

Quality control (QC): to determine if the data collected from each plate meet the minimum quality requirements.

assay development

screen

Signal-to-background (S/B)	$\dfrac{mean(C_{pos})}{mean(C_{neg})}$
Signal-to-noise (S/N)	$\dfrac{mean(C_{pos}) - mean(C_{neg})}{std(C_{neg})}$ or $\dfrac{mean(C_{pos}) - mean(C_{neg})}{\sqrt{std(C_{pos})^2 + std(C_{neg})^2}}$
Signal window (SW)	$\dfrac{\|mean(C_{pos}) - mean(C_{neg})\| - 3 \times (std(C_{pos}) + std(C_{neg}))}{std(C_{pos})}$ or $\dfrac{\|mean(C_{pos}) - mean(C_{neg})\| - 3 \times (std(C_{pos}) + std(C_{neg}))}{std(C_{neg})}$
Assay variability ratio (AVR)	$\dfrac{3 \times std(C_{pos}) + 3 \times std(C_{neg})}{\|mean(C_{pos}) - mean(C_{neg})\|} = 1 - Z'\text{-factor}$
Z'-factor	$1 - \dfrac{3 \times std(C_{pos}) + 3 \times std(C_{neg})}{\|mean(C_{pos}) - mean(C_{neg})\|}$
Z-factor	$1 - \dfrac{3 \times std(C_{pos}) + 3 \times std(S_{all})}{\|mean(C_{pos}) - mean(S_{all})\|}$
SSMD **(Strictly standardized mean difference)**	$\dfrac{mean(C_{pos}) - mean(C_{neg})}{\sqrt{std(C_{pos})^2 + std(C_{neg})^2}}$

Quality control methods

- **Signal-to-background (S/B):**

$$S/B= \frac{\text{mean}(C_{pos})}{\text{mean}(C_{neg})}$$

- **Signal-to-noise (S/N):**

$$S/N= \frac{\text{mean}(C_{pos})-\text{mean}(C_{neg})}{\text{std}(C_{neg})}$$ or $$S/N= \frac{\text{mean}(C_{pos})-\text{mean}(C_{neg})}{\sqrt{\text{std}(C_{pos})^2+\text{std}(C_{neg})^2}}$$

   – Applied for the establishment of an assay.

   – Both S/B and S/N are considered weak parameters to represent dynamic signal range for an HTS and are rarely used.

The most widely used QC criterion in HTS:

- **Z'-factor**: The separation between positive (C_{pos}) and negative (C_{neg}) controls is calculated as a measure of the signal range of a particular assay in a single plate.

$$Z' = 1 - \frac{3 \text{ x std}(C_{pos}) + 3 \text{ x std}(C_{neg})}{|\text{mean}(C_{pos}) - \text{mean}(C_{neg})|}$$

- Z'-factor has its basis on normality assumption, and the use of 3x std of the mean of the group comes from the 99.73% confidence limit.

- The magnitude of the Z'-factor does not necessarily correlate with the hit confirmation rates.

- Require very strong controls to get "good" Z'-factor, thus more applicable for small-molecule screens.

- **Z-factor**: The modified version of the Z'-factor, where the mean and std of the negative control are substituted with the ones for the test samples (S_{all}).

$$Z = 1 - \frac{3 \text{ x std}(C_{pos}) + 3 \text{ x std}(S_{all})}{|\text{mean}(C_{pos}) - \text{mean}(S_{all})|}$$

  - In a focused library, in which many possible "hits" are clustered in certain plates, Z-factor would not be an appropriate QC parameter.

The Z' (or Z) parameter based QC criteria:

$0.5 \leq Z' < 1$, a very good assay.

$0 < Z' < 0.5$, an acceptable assay.

$Z' < 0$, an unacceptable assay.

- **Strictly standardized mean difference (SSMD)**: the ratio between the difference of the means and the standard deviation of the difference between controls.

$$\text{SSMD} = \frac{\text{mean}(C_{pos}) - \text{mean}(C_{neg})}{\sqrt{\text{std}(C_{pos})^2 + \text{std}(C_{neg})^2}}$$

- SSMD was developed primarily for RNAi screens with a solid probabilistic and statistical basis.

- SSMD has both clear original and probability meanings for evaluating the differentiation between positive and negative controls.

- the SSMD-based QC criteria obtain consistent QC results for multiple positive controls with different effect sizes.

Quality control methods

Table 1. SSMD-Based QC Criteria in RNAi HTS Assays Taking into Account Effect Size of a Positive Control in an Assay under Two Situations

$C_{pos} > C_{neg}$ *Situation I   The intensity of a positive control is theoretically greater than that of a negative reference.*

Quality Types	Criterion Ia: For a Moderate Control	Criterion Ib: For a Fairly Strong Control	Criterion Ic: For a Strong Control	Criterion Id: For a Very Strong Control
Excellent	$\hat{\beta} \geq 2$	$\hat{\beta} \geq 3$	$\hat{\beta} \geq 4.7$	$\hat{\beta} \geq 6.67$
Good	$2 > \hat{\beta} \geq 1$	$3 > \hat{\beta} \geq 2$	$4.7 > \hat{\beta} \geq 3$	$6.67 > \hat{\beta} \geq 4.7$
Inferior	$1 > \hat{\beta} \geq 0.5$	$2 > \hat{\beta} \geq 1$	$3 > \hat{\beta} \geq 2$	$4.7 > \hat{\beta} \geq 3$
Poor	$\hat{\beta} < 0.5$	$\hat{\beta} < 1$	$\hat{\beta} < 2$	$\hat{\beta} < 3$

$C_{pos} < C_{neg}$ *Situation II   The intensity of a positive control is theoretically less than that of a negative reference.*

Quality Types	Criterion IIa: For a Moderate Control	Criterion IIb: For a Fairly Strong Control	Criterion IIc: For a Strong Control	Criterion IId: For a Very Strong Control
Excellent	$\hat{\beta} \leq -2$	$\hat{\beta} \leq -3$	$\hat{\beta} \leq -4.7$	$\hat{\beta} \leq -6.67$
Good	$-2 < \hat{\beta} \leq -1$	$-3 < \hat{\beta} \leq -2$	$-4.7 < \hat{\beta} \leq -3$	$-6.67 < \hat{\beta} \leq -4.7$
Inferior	$-1 < \hat{\beta} \leq -0.5$	$-2 < \hat{\beta} \leq -1$	$-3 < \hat{\beta} \leq -2$	$-4.7 < \hat{\beta} \leq -3$
Poor	$\hat{\beta} > -0.5$	$\hat{\beta} > -1$	$\hat{\beta} > -2$	$\hat{\beta} > -3$

$\hat{\beta}$ denotes estimated strictly standardized mean difference (SSMD) value.

Quality control methods

Table 2. Strategies for Quality Controls in an HTS Experiment with One or Two Positive Controls

Strategy 1: Using 1 positive control,
    a plate passes QC if it has good or excellent quality;
    a plate fails QC if it has poor or inferior quality.

Strategy 2: Using 2 positive controls,
    a plate passes QC if it has good or excellent quality in both positive
        controls;
    a plate fails QC if it has inferior or poor quality in both positive controls;
    depending on experimental need and cost, a plate may pass or fail QC if it
        has good or excellent quality in 1 positive control and inferior or poor
        quality in the other positive control.

Part-1: Introduction of the main aspects of HTS data analysis

Role of statistics in HTS design

Normalization

QC methods

"Hit" selection methods

"Hit" selection in primary screen

"Hit": test sample with value above or below the cutoff.

Erroneous assessment of "hits" and "non-hits" is likely, highly depending on

- the specificity and sensitivity of an HT assay.
- the analysis methods used for "hit" selection.

False discovery rate (FDR): the ratio of false discoveries to the total number of discoveries.

- The FDR can be inherently reduced by selection of the appropriate controls for screen quality assessment as well as for "hit" selection.

"Hit" selection in primary screen

The "hit" selection methods for primary HTS:

- **Percent inhibition cut-off**
  - preferred for small molecule screens with strong controls.

- **SSMD and robust SSMD**
  - Mean and std in the standard SSMD formula is substituted with median and MAD in the robust version.

- **Mean ± k std**

- **Median ± k MAD**
  - robust to outliers and has been shown to identify weak hits in RNAi data more effectively than mean ± k std while still capturing the strong hits and controlling false positives.

……

The "hit" selection methods for confirmatory screen:

- **Dose-response analysis**

  - Defined by top asymptote (maximal response), bottom asymptote (baseline response), slope (Hill slope), and the EC_{50} value.

  - Common used in small molecule screens.

  $$signal = B + \frac{T-B}{1 + \left(\frac{EC_{50}}{x}\right)^h}$$

- **SSMD**

- **Multiple t-tests**

- **Redundant siRNA activity (RSA)**

- **Rank product**

......

p-value:
in replicates to statistically calculate the significance of the sample activity in relation to a negative reference group.

Part-2: Data analysis approaches in RNAi HTS

- ➢ **Comparison of RNAi and small-molecule screens**

- ➢ **RNAi screening data analysis**

Features of RNAi and small-molecule HTS data

Comparison of existing datasets from more than 19 optimized siRNA screens and more than 13 mammalian cell-based small-molecule screens:

Assay Parameter	Small-Molecule Assays	siRNA Assays
Median Signal to Background (S/B; Note 1)	5.6 (n=14)	2.9 (n=18)
Median Coefficient of Variation (CV; Note 2)	13.4% (n=21)	26.5% (n=25)
Normality of Data (Note 3)	5% Normally Distributed (n=21)	26% Normally Distributed (n=25)
Z'-factor	Generally > 0.5	Generally < 0.5

- Screens were carried out at the ICCB-Longwood Screening Facility at Harvard Medical School.
- 384-well format. siRNA screen with 6 positive and 4 negative controls per plate. Small-molecule screen with 6 positive and 16 negative controls per plate.

In general, the siRNA screens were less robust by these standard measures than small-molecule screens.

Differences between RNAi and small molecule HTS

❖ **Library design:**

- In theory, most siRNA reagents in a genome-scale library should have an expressed cellular target and knockdown of each target may have physiological effects.

- RNAi reagents have a higher intrinsic probability for impacting the overall network biology.

- In contrast, most compounds in small-molecule screening libraries have no effect on cell pathways because they lack cellular targets and are de facto negative controls.

Differences between RNAi and small molecule HTS

❖ **Transfection:**

- Even under well-controlled conditions, the transfection process, especially transfection efficiency, is a major source of variability for siRNA screens.

- Transfection also causes cell stress and can affect cell viability, which may have variable and indirect phenotypic effects in cellular assays.

Differences between RNAi and small molecule HTS

❖ **Kinetics and mechanism of RNAi action:**

- RNAi reagents generally require 48–72 hours for maximal effect, whereas small molecules can directly affect their protein targets within hours.
    - the increased time between cell plating and assay endpoints leads to greater impact of cell culture and environmental variation on phenotypes and more assay variability.

- The "window" of maximal RNAi effect likely varies for each gene, but typically a single endpoint is chosen for a screening assay.
    - Assaying too early may produce a false negative.
    - Assaying too late may lead to false positives because of downstream effects.

Differences between RNAi and small molecule HTS

❖ **Number and quality of controls:**

- RNAi positive controls generally have weaker effects (variations in transfection efficiencies) than small-molecule positive controls.

- Small-molecule screens use vehicle-only wells as negative controls, no such universal negative control exists for siRNA screens. Even 'nontargeting' siRNA controls may exhibit off-target effects in some cell lines.

- Sequence-specific off-target effects of RNAi resulting from partial sequence complementarity between siRNA and mRNA complicate interpretation of RNAi experiments. (mechanisms of off-target effects are not yet completely understood)

Part-2: Data analysis approaches in RNAi HTS

- ➢ **Comparison of RNAi and small-molecule screens**

- ➢ **RNAi screening data analysis**

Successful data analysis of RNAi screening depends critically on:

- The careful experimental design and assay development before the primary screen.

- Plate layout and placement of control reagents.

- Whole genomes can be carried out in <30,000 wells and thus can be performed routinely in duplicate or higher replicate numbers to decrease both false positive and false negative rates.

Data triage

- **Plate visualization**: an effective technique for uncovering undesirable patterns that might indicate technical problems.

  – Data should be examined while the screen is in progress to allow the identification of potential problems as they occur.
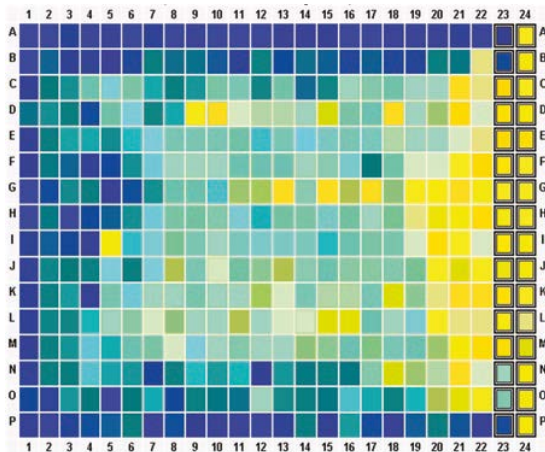
"Heat map" of raw values from a 384-well plate with strong edge effects.
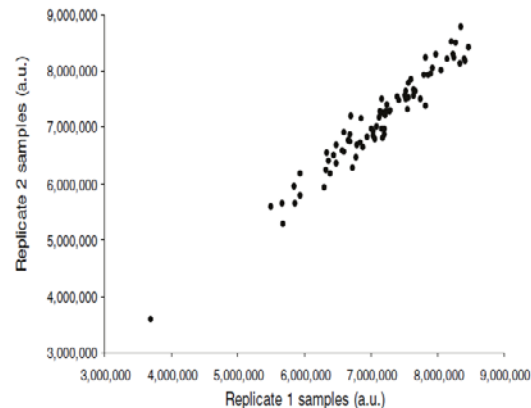
"Replicate correlation plots" of raw values from 2 replicates of the same master plate showing good agreement and suggesting overall good reproducibility.
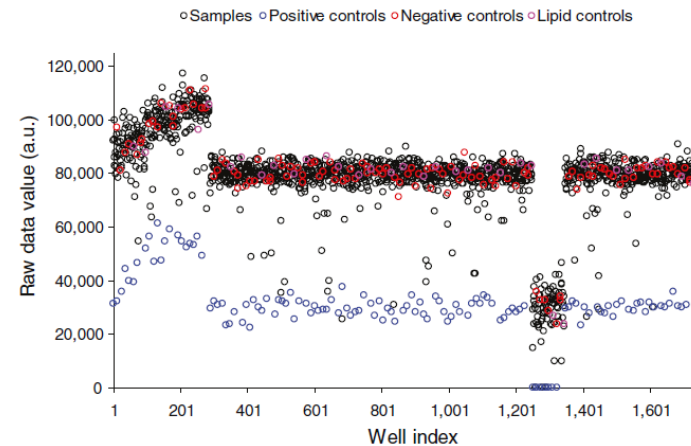
"Plate-well scatter plot" of raw values from a screen of 18x 96-well plates including evidence of drift in the first three plates and a plate with unusually low values near the end of the screen

Normalization

Normalization Methods	Formula	Advantages	Disadvantages
% of control mean	**% = sample signal *100** / **mean of control**	**• easy to calculate** **• easy to interpret** **• biologists are accustomed to this type of normalization**	**• very sensitive to outliers** **• need many replicates of controls** **• does not incorporate information on control variation** **• does not adjust for positional effects within plates**
% of sample median	**% = sample signal * 100** / **median of samples**	**• easy to calculate** **• easy to interpret** **• not very sensitive to outliers** **• biologists are accustomed to this type of normalization** **• does not need many replicates of controls, because use samples as '*de facto*' negative controls**	**• does not incorporate information on sample variation** **• does not adjust for positional effects within plates**

Normalization

Normalization Methods	Formula	Advantages	Disadvantages
z-score	z = <u>sample value - sample mean</u> sample standard deviation	• easy to calculate • incorporates information on sample variation • easy to use results in hit identification : typically use threshold of z >=2 or 3 • does not need many replicates of controls because **use samples as '*de facto*' negative controls**	• **can be sensitive to outliers** • plate-based z scores can be skewed if hits are unevenly distributed on plates (common) • does not adjust for positional effects within plates
robust z-score (MAD z-score)	z = <u>sample value – sample median</u> sample median absolute deviation	• easy to calculate • incorporates information on sample variation • **not very sensitive to outliers** • does not need many replicates of controls because **use samples as '*de facto*' negative controls**	• does not adjust for positional effects within plates • **some biologists may not be accustomed this type of normalization**

Calculation of quality metrics

Quality Metrics	Formula	Advantages	Disadvantages
Z′-factor	$Z' = 1 - \dfrac{(3*SD\ of\ h.v.\ control + 3*SD\ of\ l.v.\ control)}{\|h.v.\ control\ mean - l.v.\ control\ mean\|}$ **(SD, Standard Deviation h.v., high-value; l.v., low-value.)**	• easy to calculate • takes dynamic range and data variation into account.	• can get "good" Z′-factor using very strong controls, but this may not be representative of positives from screen • used extensively for small molecule assays; may need to adjust thresholds for siRNA screens
Z-factor	$Z = 1 - \dfrac{(3*SD\ of\ sample + 3*SD\ of\ control)}{\|sample\ mean - control\ mean\|}$	• easy to calculate • takes signal dynamic range and data variation into account	• used extensively for small molecule assays; may need to adjust thresholds for siRNA screens

Calculation of quality metrics

Quality Metrics	Formula	Advantages	Disadvantages
SSMD (Strictly Standardized Mean Difference)	SSMD = $\dfrac{\text{h.v. control mean} - \text{l.v. control mean}}{\sqrt{\text{SD}^2 \text{ of h.v. control} + \text{SD}^2 \text{ of l.v. control}}}$ (SD, Standard Deviation h.v., high-value; l.v., low-value.)	• easy to calculate • takes dynamic range and data variation into account • less conservative estimator than Z/Z′-factor • has rigorous statistical estimator for non-normal data • different thresholds available for different control strengths	• not yet as widely recognized as Z/Z′-factor

Hit identification

Strategies	Formula	Advantages	Disadvantages
Mean ± k std	**Hit with increased activity = any sample whose value is >= sample mean + k std;** **Hit with decreased activity = any sample whose value is <= sample mean – k std**	• **easy to calculate** • **easily linked to hit p-values**	• **sensitive to outliers** • **can miss weak positives** • **requires multiple comparison corrections if using p-values**
Median ± k MAD **(Median Absolute Deviation)**	**Hit with increased activity = any sample whose value is >= sample median + k MADs;** **Hit with decreased activity = any sample whose value is <= sample median – k MADs**	• **easy to calculate** • **can identify weaker hits** • **not very sensitive to outliers**	• **not easily linked to hit p-values**

Hit identification

Strategies	Formula	Advantages	Disadvantages
Multiple t-tests	**Hit = any reagent for which t-test result between samples at two conditions < threshold (usually p = 0.05 or p = 0.01)**	**• easy to calculate** **• provides hit p-values**	**• requires triplicates, at minimum** **• sensitive to outliers** **• inappropriate if data is not normally distributed** **• requires multiple comparison corrections of p-values**
SSMD **(Strictly standardized mean difference)**	**Appropriate equations depend on whether goal is control of rates of false negatives, false positives or both**	**• allows control of both false positive and false negative rate** **• linked to rigorous probability interpretation**	**• not available in most analysis software** **• not intuitive for many biologists**
RSA **(Redundant siRNA activity)**	**Iterative ranking algorithm that cannot be reduced to a single equation**	**• can identify weaker hits** **• not sensitive to outliers** **• provides hit p-values** **• may help reduce false-positives due to off-target effects of single reagents**	**• difficult to calculate** **• may have limited utility for pool-based screens**

Unique drug screening approach for prion diseases identifies tacrolimus and astemizole as antiprion agents

Yervand Eduard Karapetyan[a,1,2], Gian Franco Sferrazza[a,1,3], Minghai Zhou[a], Gregory Ottenberg[a,3], Timothy Spicer[b], Peter Chase[b], Mohammad Fallahi[c], Peter Hodder[b], Charles Weissmann[a,4], and Corinne Ida Lasmézas[a,4]

[a]Department of Infectious Diseases, [b]Lead Identification, and [c]Information Technology and Informatics, The Scripps Research Institute, Scripps Florida, Jupiter, FL 33458
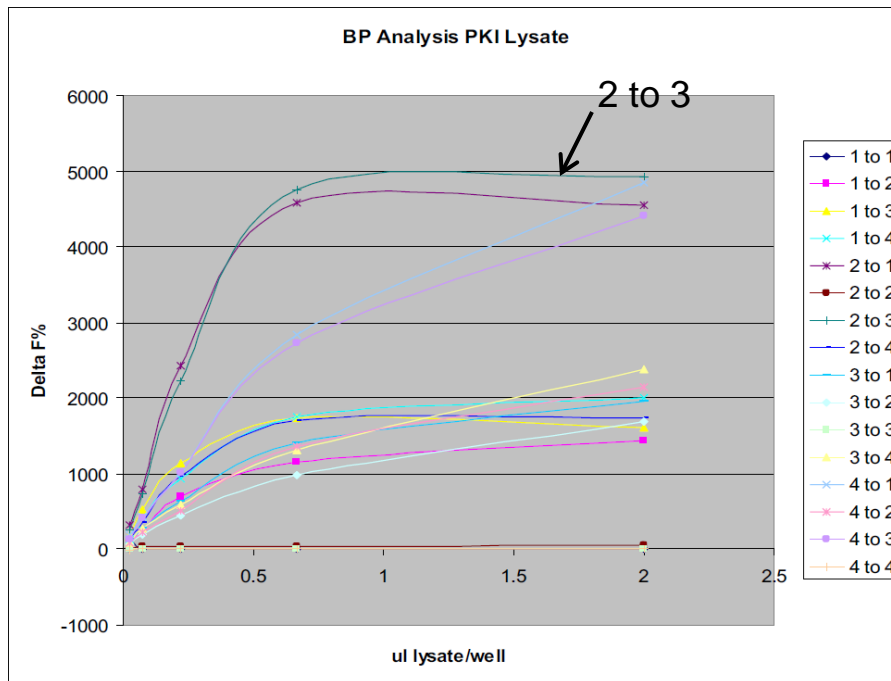
- Prion diseases are characterized by the accumulation, in brain and lymphoid tissue of PrP^{Sc}, a misfolded, aggregated form of the host prion protein (PrP^{C}).

- Because PrP^{C} is necessary for prion replication, this paper applied HTS for compounds that decrease PrP^{C} expression on living cell surface.

Features of this screen

- **Primary assay:** PrP-FEHTA (FRET-based high throughput assay)

- **Cell based**: LD9 cells (a subclone of L929 fibroblasts)

- **Componds**: US Drug Collection that comprises 1,280 drugs approved for use in humans.

- **Normalization:** % of control mean
  - normalized to compound solvent control.

- **QC**: Z′-factor
  - Positive control: maximum signal, LD9 cells or LD9 cells + compound solvent DMSO
  - Negative control: minimum signal, Hpl PrP$^{0/0}$ cells or LD9 cells + pharmacological control brefeldine A (BFA, blocks progression of proteins from the ER to the Golgi apparatus)

- **Hit selection**: threshold set at ≤50% PrP^C expression and ≤10% toxicity.

Assay development

- **PrP-FEHTA**: FRET-enabled high throughput assay, which uses two antibodies recognizing distinct domains of PrP^C for quantification of a cell surface protein.

BP Analysis PKI Lysate

2 to 3

Legend:
- 1 to 1
- 1 to 2
- 1 to 3
- 1 to 4
- 2 to 1
- 2 to 2
- 2 to 3
- 2 to 4
- 3 to 1
- 3 to 2
- 3 to 3
- 3 to 4
- 4 to 1
- 4 to 2
- 4 to 3
- 4 to 4

Mouse neuroblastoma cells (PK1) lysates contained 18.6 µg/mL protein.
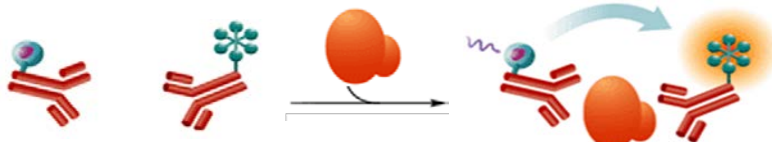
Antibody key:
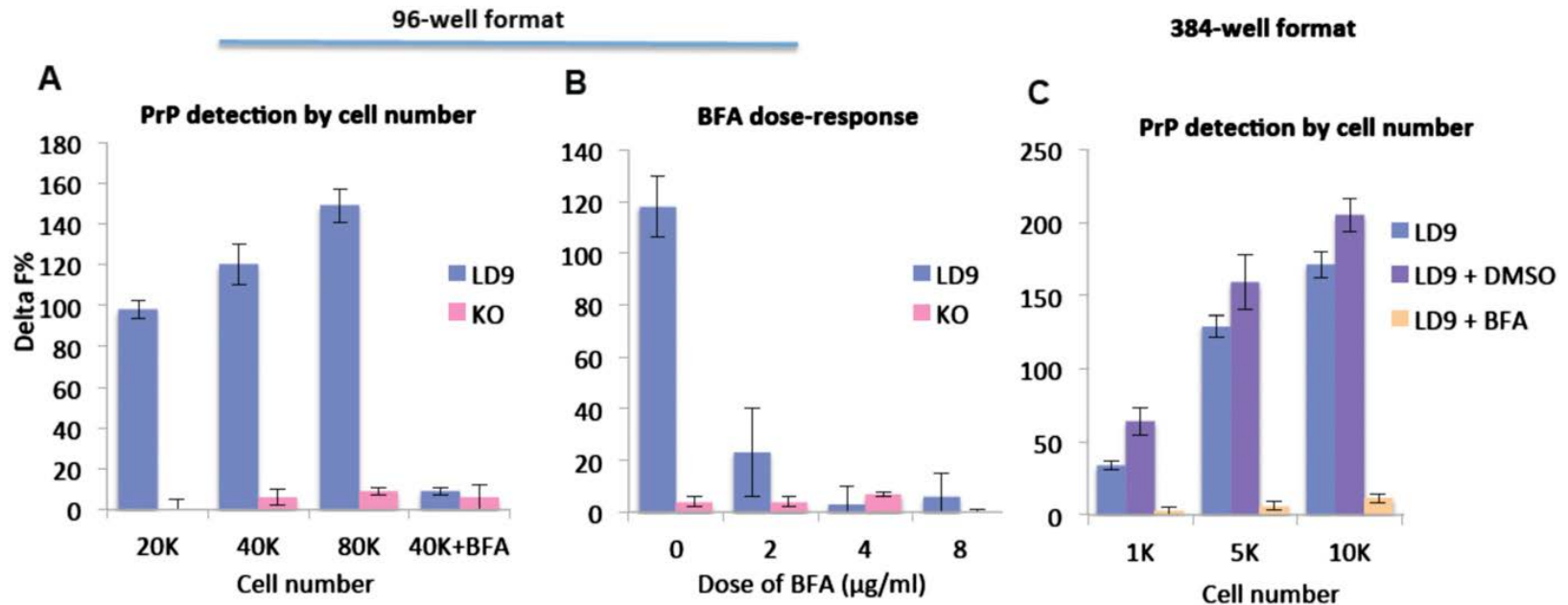1, POM19;
2, SAF32;
3, D18;
4, D13.

Donor   Akzeptor

The best signal was obtained with anti-PrP antibodies SAF32 (amino acids 53–93) and D18 (amino acids 133–157)

$$\Delta F\% = [(R_{C+} - R_{C-})/R_{C-}] \times 100$$

where R_{C+} and R_{C-} are the FRET (665/620) ratios of the positive and negative controls (Hpl PrP$^{0/0}$).

Assay development

96-well format

A

PrP detection by cell number

(Delta F% vs Cell number: 20K, 40K, 80K, 40K+BFA; legend: LD9, KO)

B

BFA dose-response

(Dose of BFA (µg/ml): 0, 2, 4, 8; legend: LD9, KO)

384-well format

C

PrP detection by cell number

(Cell number: 1K, 5K, 10K; legend: LD9, LD9 + DMSO, LD9 + BFA)

PrP-FEHTA optimization:

– (A and B) Optimization of LD9 cell density and pharmacological control **BFA.**

– (A) **Z′ values** were **0.7** for 20,000, **0.6** for 40,000, and **0.8** for 80,000 LD9 cells (**LD9 cells providing the maximal**, **PrP KO cells the minimal** signal intensity). Z′ was **0.7** when 40,000 **LD9 cells provided the maximal signal** and 40,000 **BFA-treated LD9 cells the minimal signal** (4 µg/mL BFA).

– (B) BFA (4 and 8 µg/mL) reduced the PrP signal to background.

– (C) Optimization of cell density and DMSO tolerance in the 384-well format. **Z′ values** were calculated for **LD9+DMSO** versus **LD9+BFA** and were equal to **0.4** (1K), **0.6** (5K), and **0.8** (10K).

Primary screen
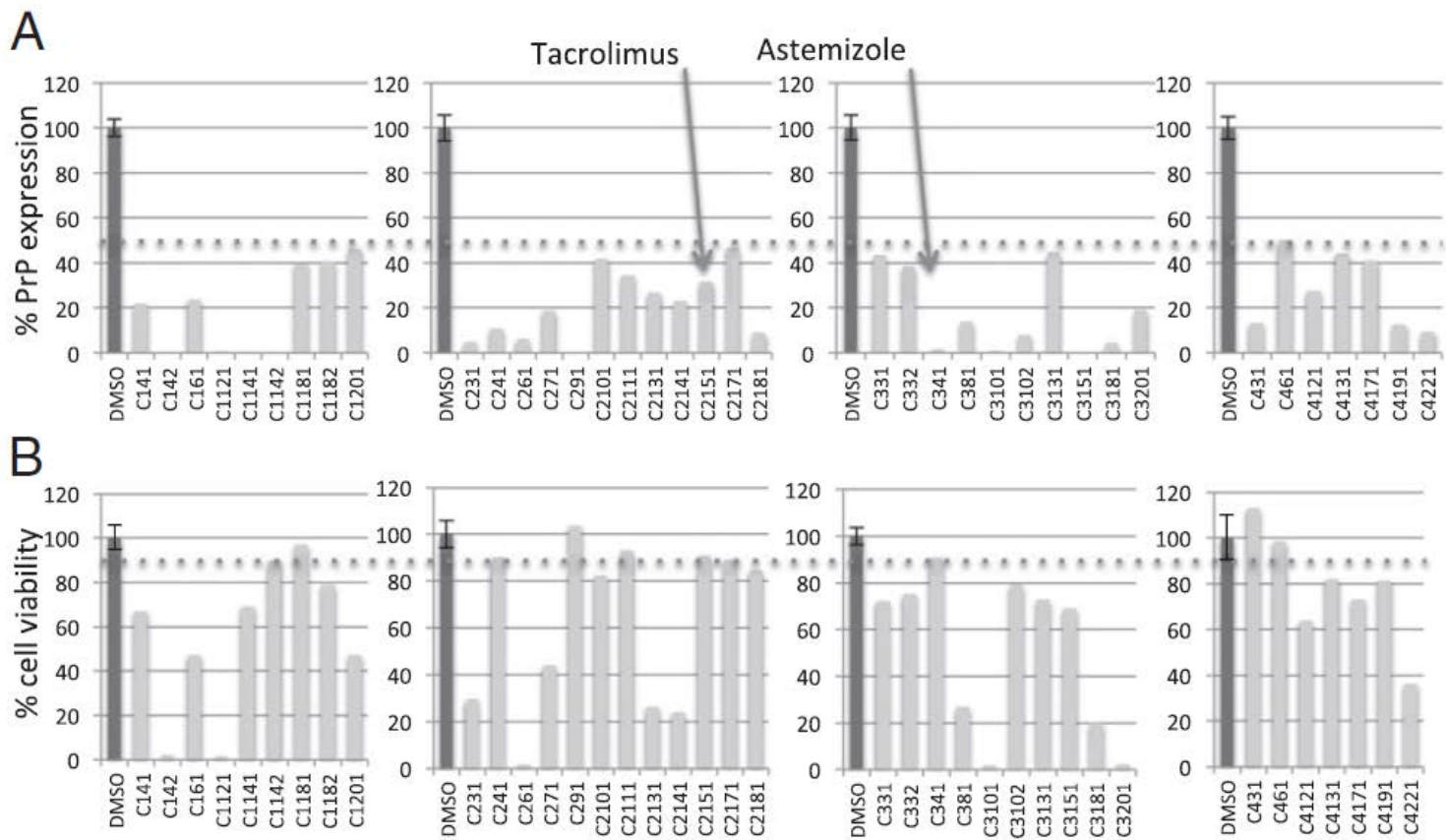
- Cutoffs are indicated by dotted lines: ≤ 50% PrP^C expression, ≤10% toxicity.

- The screen yielded **9** hits (0.7%) from 1,280 drug compounds.

Hit validation
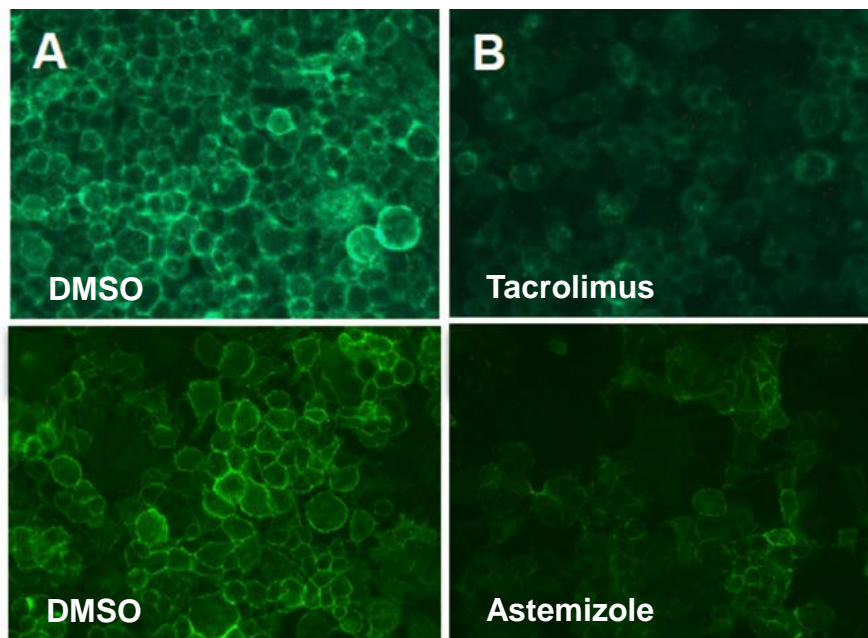
They selected two hits, **tacrolimus** and **astemizole**, based on their activity in orthogonal assays.
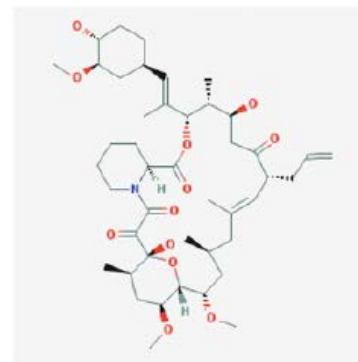
Tacrolimus

Astemizole

Hit validation by immunofluorescence on neuroblastoma cells and high-content analysis:
N2a cells were treated with the control DMSO (0.5%) or with the drug (20 µM) for 24h. PrP staining with D18 mAb on living cells, postfixation, and epifluorescence analysis.

Tacrolimus and astemizole inhibit prion replication in cell culture

A 9 days post-RML infection | 18 days post-infection (6 days without treatment)
Tac 20 Tac 6.7 Ast 2 PPS CTRL | Tac 20 Tac 6.7 Ast 2 PPS CTRL

B 9 days post-22L infection | 18 days post-infection (6 days without treatment)
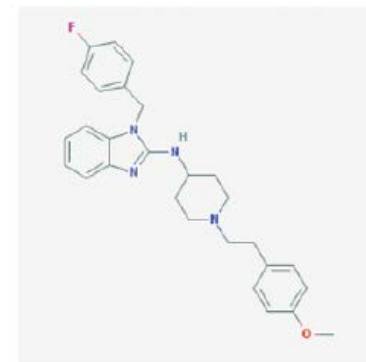Tac 20 Tac 6.7 Ast 2 PPS CTRL | Tac 20 Tac 6.7 Ast 2 PPS CTRL

Cells were analyzed for pK-resistant PrP^{Sc} by Western blot
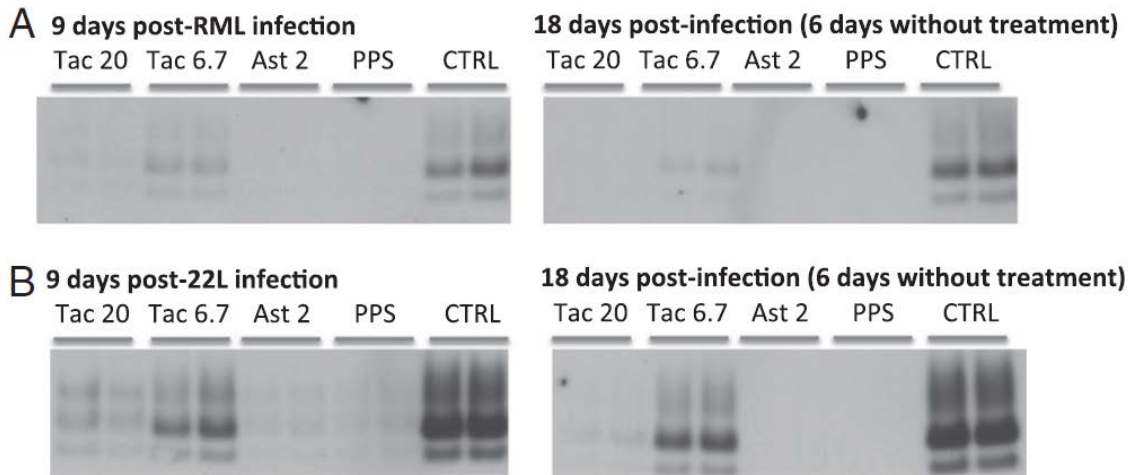
- Tacrolimus at 20µM, but not 6.7µM, strongly inhibited the replication of both RML and 22L prions.

- Astemizole blocked replication of both strains at 2µM.

Prion infection:
RML(A) or 22L(B)

Western blot   Western blot

-3 day   0 day   9 day   12 day   18 day

Drug treatment

Cells were split every 3 days
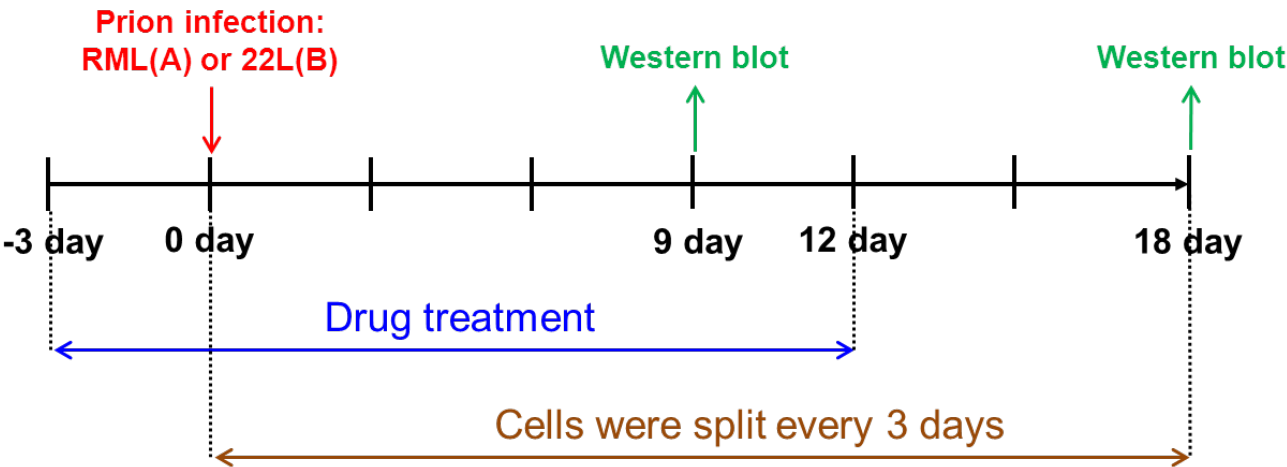
Proposed mechanism of action of tacrolimus and astemizole

Both cell surface PrP^C (biotinylated fraction) and intracellular PrP^C (unbiotinylated fraction) were analyzed by Western blot.

- Tacrolimus treatment (20µM) reduced both membrane and intracellular PrP^C by >70%, but mRNA levels were unaffected. → tacrolimus inhibits PrP^C translation.

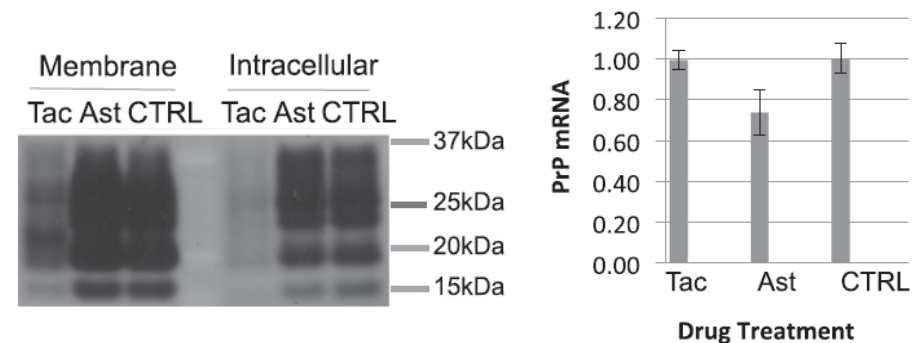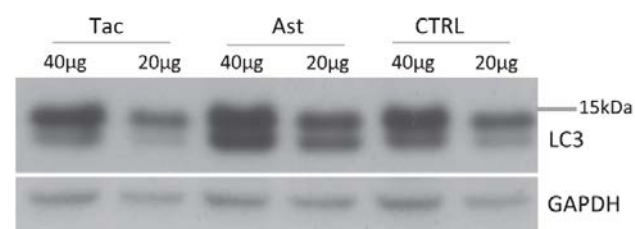- Astemizole treatment (2µM) didn't significantly affect the surface and intracellular levels of PrP^C as well as mRNA. → astemizole, at the prion inhibitory dose, inhibited prion replication by a mechanism independent of PrP expression.

Treatment	Protein	Band density	Ratio LC3-II/I
Tac	LC3-I	58.9	0.41
	LC3-II	24.2	
Ast	LC3-I	87.6	**0.93**
	LC3-II	81.4	
CTRL	LC3-I	67.9	0.56
	LC3-II	37.7	

An indicator of autophagy induction: the conversion of the cytosolic protein LC3-I to the conjugated, autophagosome-bound form LC3-II

- Tacrolimus did not significantly modify the LC3-II/I ratio. →The antiprion effect of tacrolimus is linked to a nontranscriptional regulation of PrP steady-state levels.

- The ratio doubled after astemizole treatment, indicating autophagy induction. → The antiprion effect of astemizole acts by enhancing autophagic function and thereby, prion clearance.

Therapeutic effect of astemizole in prion-infected mice
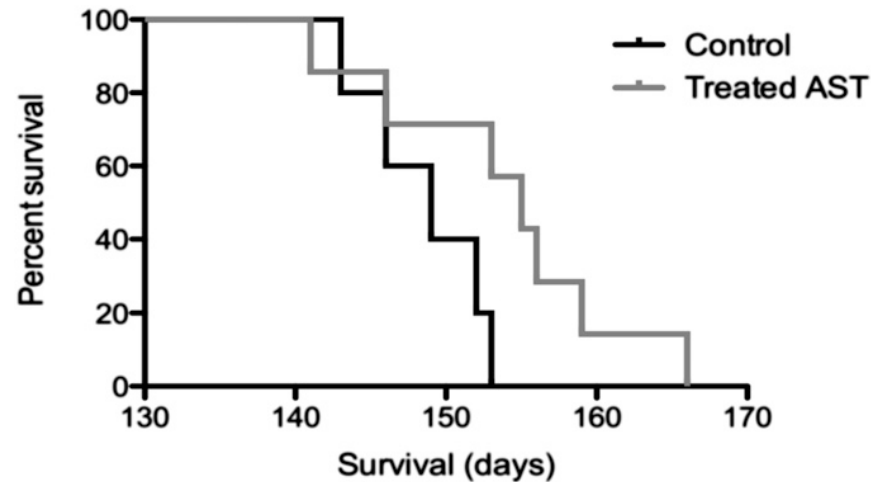
Mice were intracerebrally infected with RML prions and treated from 20 to 50d after infection by i.p. injection of tacrolimus at 1.5 mg/kg or astemizole at 3 mg/kg.

- The astemizole-treated mice showed a clear increase in survival time.

- No increase in survival times was observed in the tacrolimus-treated group.

LETTER

nature

doi:10.1038/nature12748

High–content genome–wide RNAi screens identify regulators of parkin upstream of mitophagy

Samuel A. Hasson[1,2]*, Lesley A. Kane[1]*, Koji Yamano[1], Chiu–Hui Huang[1], Danielle A. Sliter[1], Eugen Buehler[2], Chunxin Wang[1], Sabrina M. Heman-Ackah[3], Tara Hessa[1], Rajarshi Guha[2], Scott E. Martin[2] & Richard J. Youle[1]

[1]Surgical Neurology Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, Maryland 20892, USA.

- Recent studies of the Parkinson's disease associated genes PINK1 and PARK2 indicate that they may act in a quality control pathway preventing the accumulation of dysfunctional mitochondria.

- This paper elucidated regulators that have an impact on parkin translocation to damaged mitochondria with genome-wide siRNA screens coupled to high-content microscopy.

Strategy
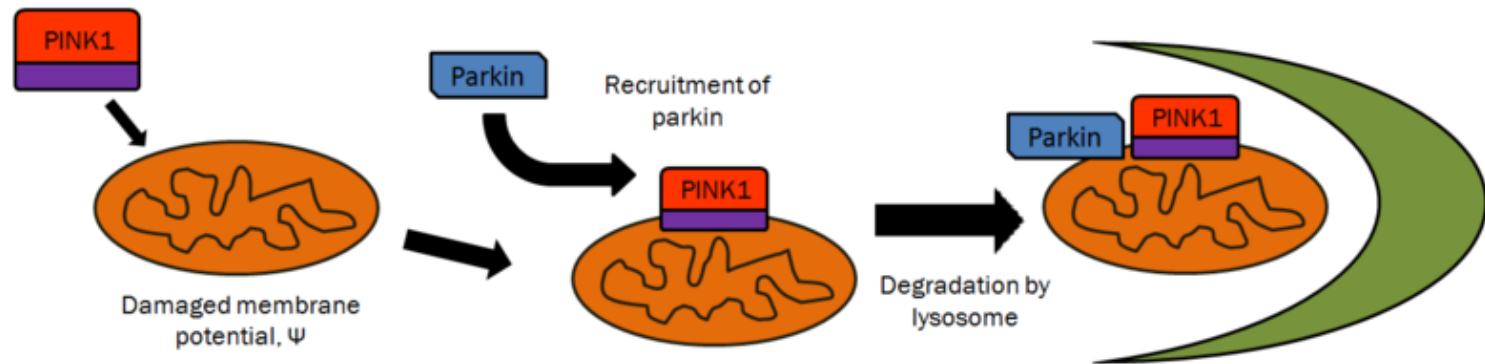
- PINK1 (positive siRNA control) <u>protects</u> cells from stress-induced mitochondrial dysfunction, by recruiting Parkin protein to bind to depolarized mitochondria to induce autophagy (mitophagy) of those mitochondria.

1) Hela cells stably expressing
   - GFP-Parkin (AcGFP1-Parkin fusion protein)
   - mito-dsRed (a mitochondrial-targeted red fluorescent protein)

2) siRNA treatment on cells for 48h, in 384-well format.

3) CCCP (carbonyl cyanidem-chlorophenyl hydrazine) treatment on cells for 2.5h, for chemically depolarizing of mitochondria.

→ Introduce Parkin translocation to mitochondria.

Features of this screen

- **Primary assay:** automated high-content microscopy assay

- **Cell based**: HeLa cells

- **Genome wide human siRNA libraries:**
    1) Ambion non-pooled library: 3 unique siRNA per gene
    2) Dharmacon pooled library: pool of 4 siRNA per gene
    In total, there were 83,000 siRNA targeting 21,993 genes

- **Normalization:**
    1) Normalization of raw data to negative control
    2) MAD Z-score of normalized data
    3) Log transformation of MAD Z-score to remove bias

- **QC**: Z'-factor
    – Positive control: Pink1 siRNA
    – Negative control: non-target siRNA
    – Plate with a Z'<0.5 was repeated.

- **Hit selection:**
    1) Toxicity filter: cell count MAD Z-score
    2) Mitochondrial filter: mitochondrial density MAD Z-score
    3) Candidate selection: parkin translocation MAD Z-score

Overview of data analysis workflow

- Raw translocation measurements
- Raw cell counts
- Raw mitochondrial intensity measurements

- Plate control <u>normalized</u> translocation measurements
- Plate control <u>normalized</u> cell counts
- Plate control <u>normalized</u> mitochondrial intensity measurements

- Conversion of normalized translocation, cell count, and mitochondrial intensity measurements from screen-level datasets <u>to median absolute deviation (MAD) *Z*-scores</u>

- <u>Log transformation translocation</u> MAD *Z*-scores to remove bias towards inhibitors in screen-level dataset

- <u>Remove all siRNAs with a cell/nuclei count MAD *Z*-score of < -2</u> to filter out toxic phenotypes and out-of-focus segmentation failures.

Overview of data analysis workflow

- Remove all siRNAs with a mitochondrial intensity MAD Z-score of <-2 to filter out mitochondrial depletion and mitophagy inducing siRNA phenotypes. The mitochondrial filter also eliminated less catastrophic focusing issues that may have interfered with translocation scoring.

- Candidate selections used the screen-specific cutoffs detailed in the supplementary methods

- Qualitative and quantitative tools used to select a small subset of interesting candidates for follow-up analysis. Since we could not pursue all candidates, tools such as common seed analysis, gene pathway enrichment with GeneGO or database searches of GO terms, and functional annotation. STRING database analysis was also applied to 'mitochondria' and 'ubiquitin' gene enrichments.

- A subset of the candidates identified from both screens were selected for follow-up verification with 4 additional siRNA reagents and secondary screening.

Workflow diagram elements:
- Mitochondrial Filter (using nuclei count MAD Z-scores)
- Ambion Dataset (final)
- Dharmacon Dataset (final)
- Preliminary Selections
- Preliminary Selections
- Removal of candidate genes not present in current genome annotation
- Candidate List
- Candidate List
- Common seed analysis
- Annotation lookup
- Seed Adjusted List
- Pathway Enrichment & GeneGO
- STRING
- Category 1
- Category 2
- Category 4
- Category 3
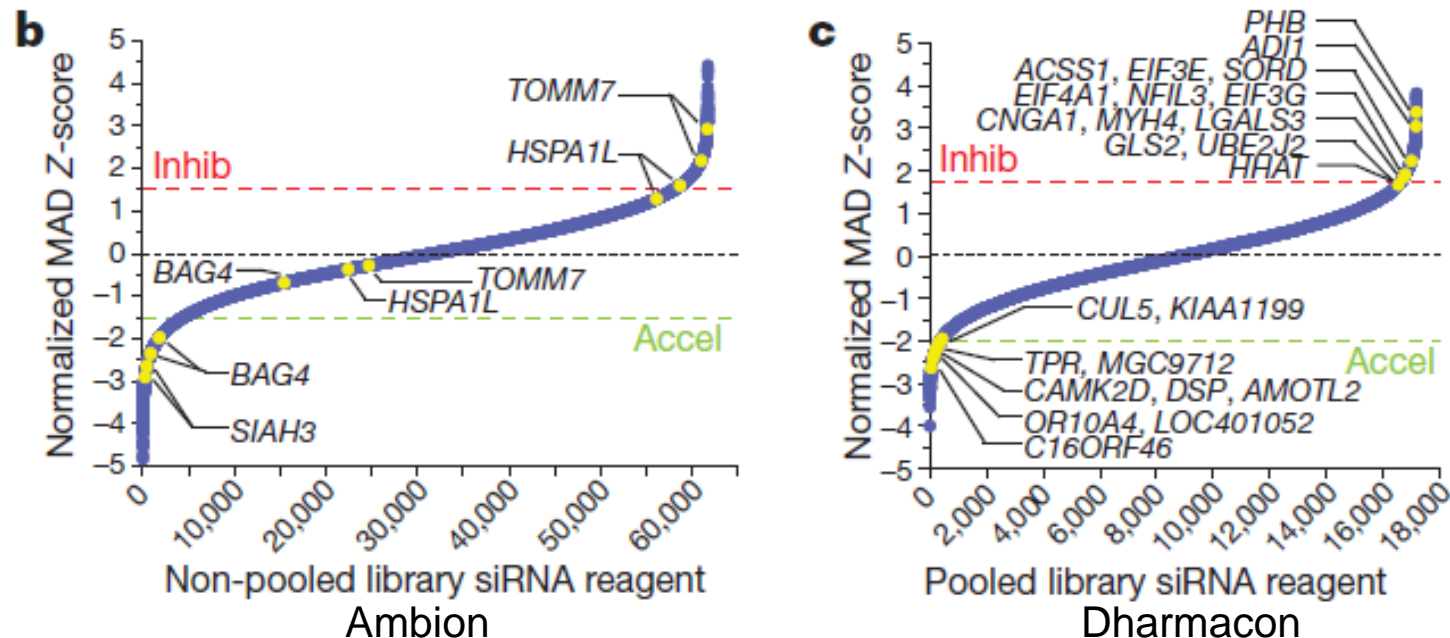- Genes selected for follow-up analysis

Candidate gene selection

Representative images of parkin translocation assay.
NTC (non-targeting control), PINK1 (inhibited translocation) and LMAN1
(accelerated translocation) siRNA transfected cells from genome-wide screens.
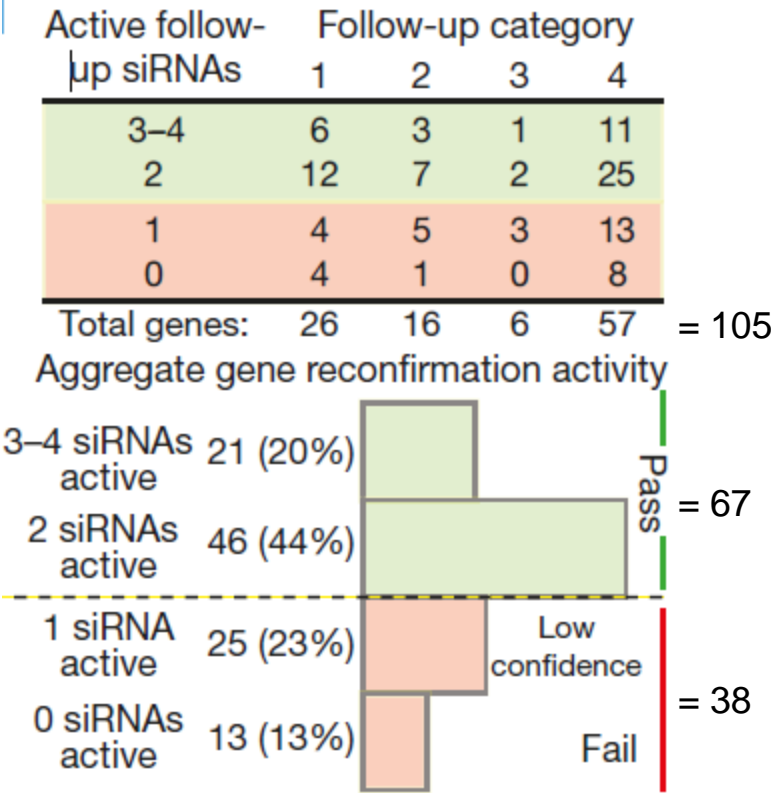
Candidate gene selection

- Filters: active siRNAs were required to maintain a cell count of MAD>-2 to remove toxic phenotypes and a "mitochondrial stain intensity" MAD>-2 to remove mitochondria depleting phenotypes.

- Candidate selection criteria:
  - non-pooled: at least two siRNAs from the data set were active based on phenotypic evaluation of parkin translocation (±1.5 MAD).
  - pooled: thresholds were set to yield similar numbers of genes as the non-pooled screen (~ ±2 MAD).

Candidate gene selection

Primary screen	Ambion siRNA library	Dharmacon siRNA library	Overlap
Positive regulator (inhibitory siRNA phenotypes)	290 genes (1.3%)	340 genes (1.5%)	14 genes
Negative regulator (accelerating siRNA phenotypes)	418 genes (2.3%)	367 genes (2.2%)	10 genes

Validation of candidate gene with follow-up secondary screen

Active follow-up siRNAs	Follow-up category			
	1	2	3	4
3–4	6	3	1	11
2	12	7	2	25
1	4	5	3	13
0	4	1	0	8
Total genes:	26	16	6	57 = 105

Aggregate gene reconfirmation activity

3–4 siRNAs active   21 (20%)
2 siRNAs active   46 (44%)   Pass = 67
1 siRNA active   25 (23%)   Low confidence
0 siRNAs active   13 (13%)   Fail = 38

- Validation screen used 4 additional siRNAs from a different vendor.

- Confirmation of knockdown by qPCR.

- Knockout (TALENs technology) to confirm the effects:
  - TOMM7: essential for stabilizing PINK1 on the outer mitochondrial membrane following mitochondrial damage.
  - HSPA1L: role in the regulation of parkin translocation.
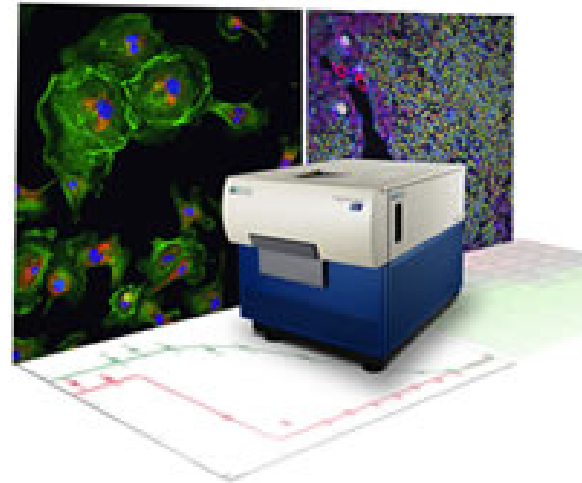
Agilent BioCel 1800 (siRNA preparation and cell culture)

Thermo Scientific WellMate dispenser (cell seeding and splitting)

BioTek EL406 dispenser (cell fixing and washing for imaging)

Molecular Devices ImageXpress Micro (image analysis)

- HTS is a comprehensive process to discover new drug targets using siRNA and drug candidates from small molecule libraries.

- High-throughput RNAi offers a wealth of exciting new data but also provides new challenges in the statistical interpretation.

- Statistical techniques and software for these HTS assays are still in their infancy; no doubt this area will see considerable development effort in the coming years.

- It should be emphasized that rigorous secondary screening is necessary to validate primary screen findings. Even the best statistics cannot establish the biological importance of genes identified as primary hits.

Math is just a horrible thing, But....

Thank you for listening!