Statistical issues in reproducible research

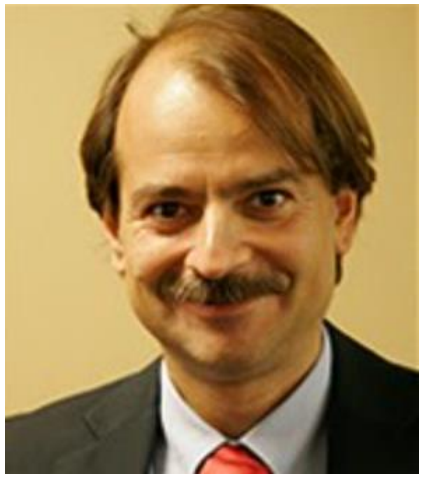
Qunhua Li

Dept of Statistics

Penn State University

June 6, 2016

Why Most Published Research Findings Are False



- Reliance on p-values
- Lack of replication
- Large number of relationships tested without preliminary findings
- Publication bias: Only “interesting” events are published.
- Detection bias: Selective or distorted reporting, conflicts of interest, deliberate manipulation
- Selection of most significant events instead of proper meta-analysis when there is replication

Ioannidis (2005) PLoS Medicine
slide adapted from Naomi Altman

What can we do?

- Improve study power:
  - do sample size calculation, better experimental design
- Improve pre-study odds:
  - preliminary studies, literature search
- Reduce reporting bias:
  - Correct research practice
- Appropriate use of statistical methods

Statistical issues

- Data preparation
- Inappropriate use of statistical methods (random sampling, model assumption, etc.)
- False discovery, multiple testing, p-hacking, overuse and misuse of p-values
- Model robustness to parameter changes and data perturbations (replicates)

Data preparation is crucial

Small changes in data filtering decisions and preparation steps can dramatically affect the data analytical outcome

- How were outliers identified and managed?
- How were missing values handled?
- How were duplicate entries handled?
- What are the input parameters?

Remedy

- Document preprocessing steps
- Use software to keep analysis history (e.g. Galaxy, R sweave – later in bootcamp)
- Describe them in the supplementary materials at the level of details that people can reproduce
- Upload preprocessed data
- Upload analysis scripts with data

Journal Requirements

In January 2014 Science enacted new policies. Check for:

1. a “data-handling plan” i.e. how outliers will be dealt with,
2. sample size estimation for effect size,
3. whether samples are treated randomly,
4. whether experimenter blind to the conduct of the experiment.

Statisticians added to the Board of Reviewing Editors.

More journals are implementing similar requirements.

Proper use of statistical methods

I will focus on

- p-value
- correlation

Caution about P-values

What is a p-value?

- What does p-value mean?

- If P value = 0.01, does it say that there was just a 1% chance of the result being false?

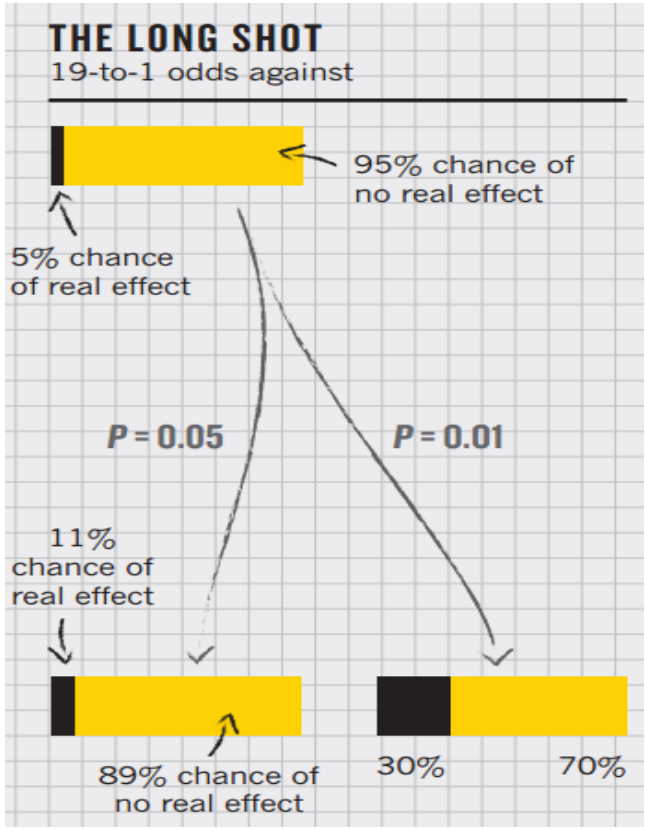
- Definition of p-value

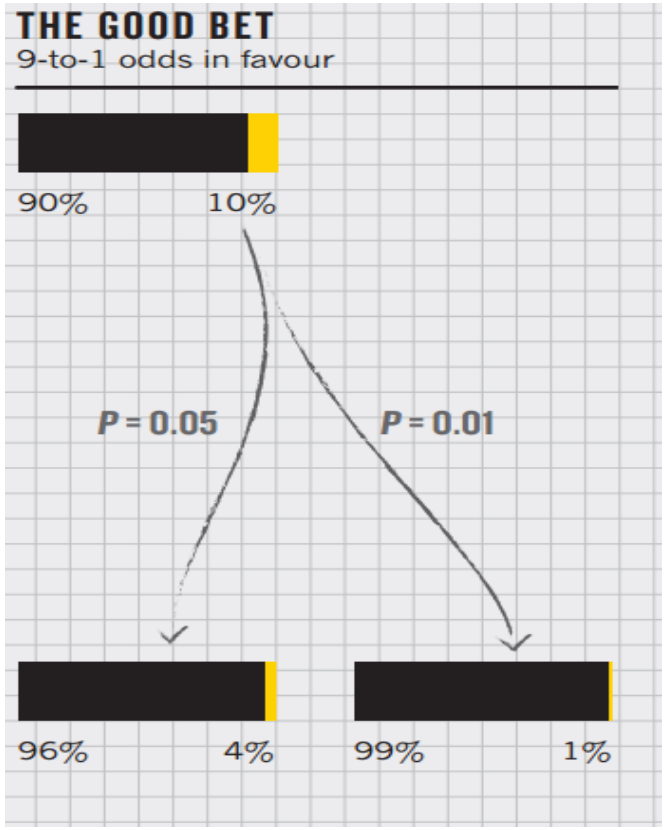
- First have a null hypothesis
- If the null hypothesis is true, p-value is the chance of getting results at least as extreme as what was actually observed
- Small p-value indicates stronger evidence against null hypothesis

P-value depends on the odds that a real effect was there at the first place

- The more implausible the hypothesis, the greater the chance that an exciting finding is a false alarm, no matter what the P value is.

A small P-value can make a hypothesis more plausible, but the difference may not be dramatic.





Nuzzo (2014) Nature

P-value depends on the underlying model

- It is likely to get different p-values using different data analytic software on the same data

Previous reports have relied heavily on the statistical significance (*P* value) rather than on the actual measured quantity of differential expression (fold change or ratio) in identifying differentially expressed genes. This strict reliance on *P* values alone has resulted in the apparent lack of agreement between sites and microarray platforms[20,26]. Our results from analyzing the MAQC human data sets

nature biotechnology

Article

Nature Biotechnology **24**, 1151 - 1161 (2006)
Published online: 8 September 2006 | doi:10.1038/nbt1239

The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements

MAQC Consortium

P-value depends on sample size

Significance is no indicator of practical relevance

- When sample size is small, real effects may not be statistically significant; when sample size is big, trivial effects can have significant p-value

Marital satisfaction and break-ups differ across on-line and off-line meeting venues

John T. Cacioppo[a,1], Stephanie Cacioppo[a], Gian C. Gonzaga[b], Elizabeth L. Ogburn[c], and Tyler J. VanderWeele[c]

Conclusion: the findings for marital break-up and marital satisfaction remained significant. These data suggest that the Internet may be altering the dynamics and outcomes of marriage itself.

Marital satisfaction

- Effect size: 5.48 (offline) to 5.64 (online) on a 7-point scale
- P-value <0.0001
- Why? n > 19,000 people

Multiple testing and P-hacking

- The more test you run, the more likely you get a significant p-value by chance. Need adjust for multiple testing.
- P-hacking is trying multiple things until you get the desired result
- P-hacking is especially likely in the studies that chase small effects hidden in noisy data

- Simonsohn (2013) found evidence that many published psychology papers report P values that cluster suspiciously around 0.05, just as would be expected if researchers fished for significant P values until they found one.

Remedy

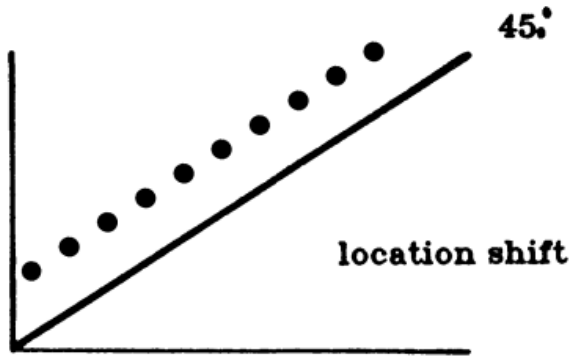
- Report effect sizes and confidence intervals. These convey what a P value does not: the magnitude and relative importance of an effect.
- Do literature search and preliminary study to improve pre-study odds
- Bring in scientific judgement about the plausibility of a hypothesis and study limitations

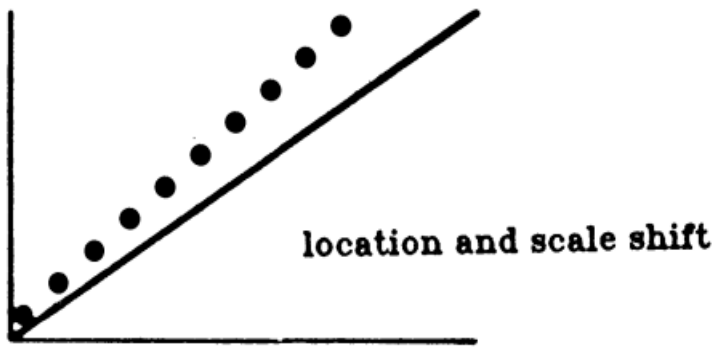
Caution about correlation

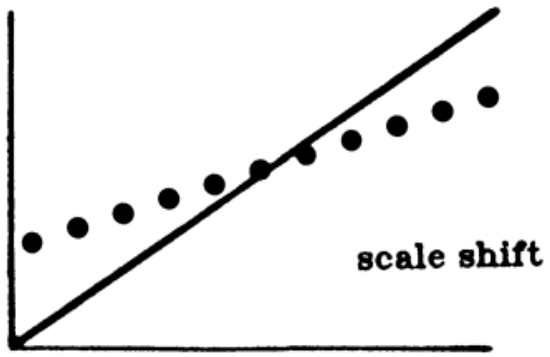
Correlation ≠ agreement

ENCODE RNA-seq data standard document (2011):
A typical R^2 (Pearson) correlation of gene expression (RPKM) between two biological replicates, for RNAs that are detected in both samples using RPKM or read counts, should be between 0.92 to 0.98. Experiments with biological correlations that fall below 0.9 should be either be repeated or explained.

Same correlation for all three examples



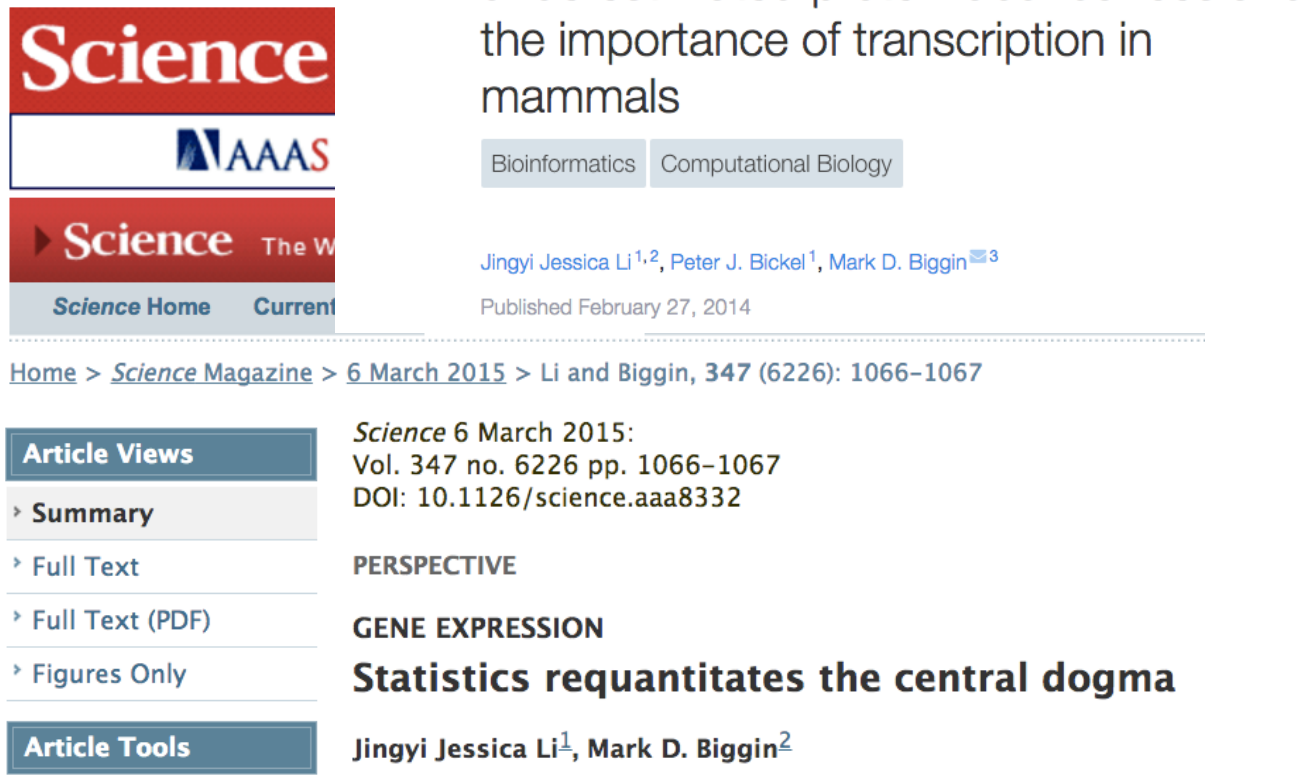




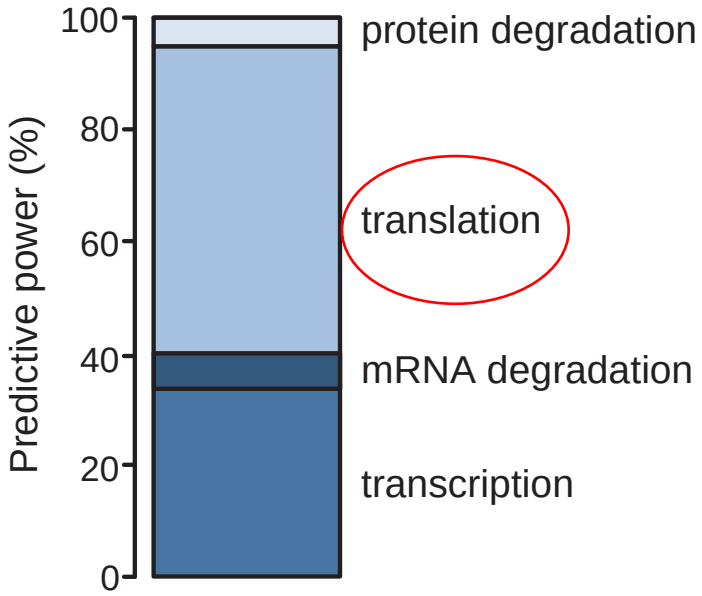
Lin (1989)

Correlation measures linear trend

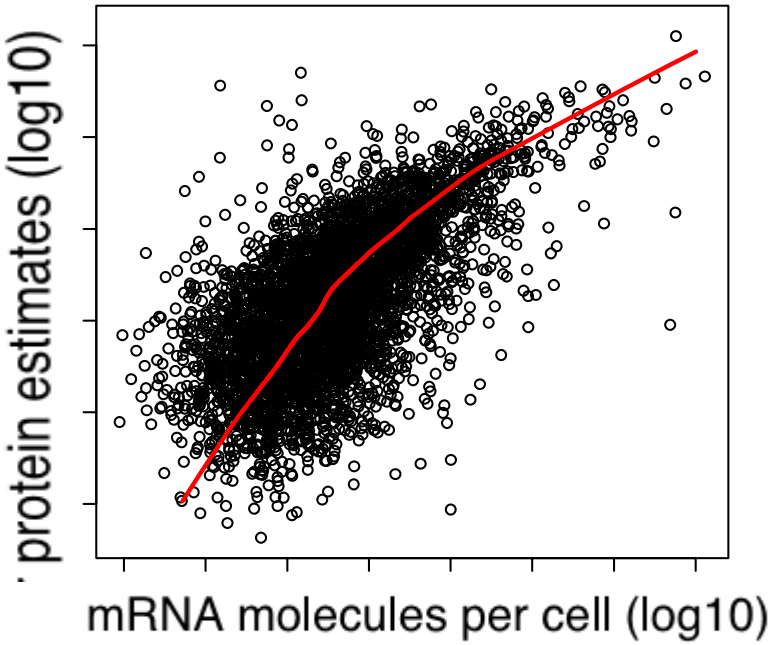
- Which steps in central dogma determine protein levels in animals?



Schwanhausser et al. suggest that translation rates are the most important, because protein and mRNA have poor correlation



measured protein vs measured mRNA

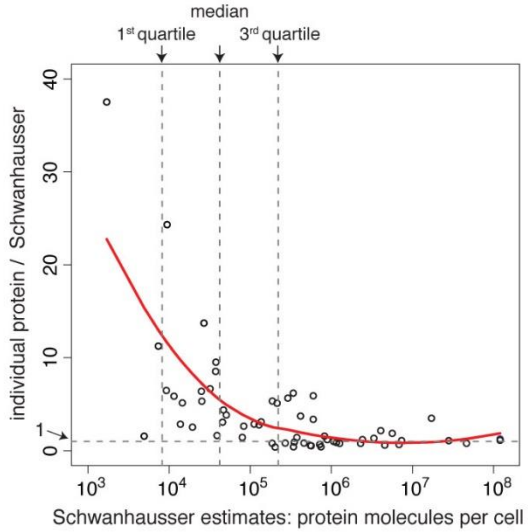


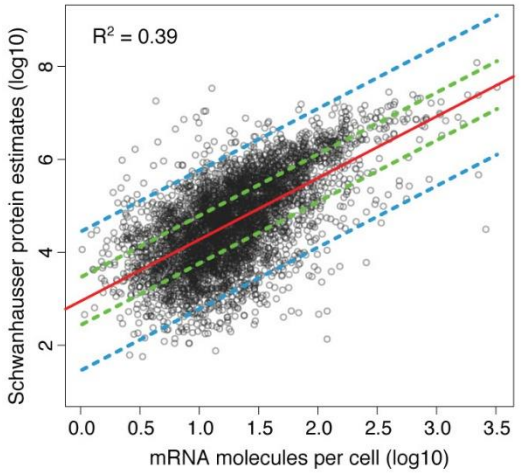
Schwanhausser et al. (2011) Nature 473, p 337

Slide from Jessica Li

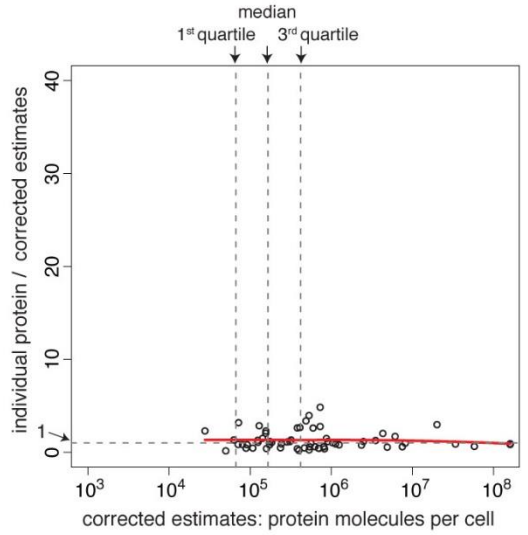
Corrected protein estimates increases the correlation of protein and mRNA levels

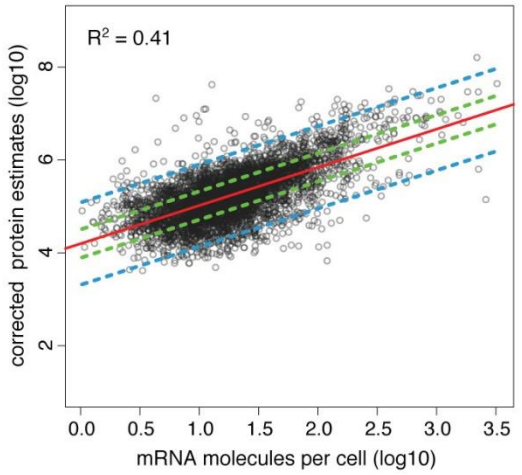
Schwanhausser's protein estimates





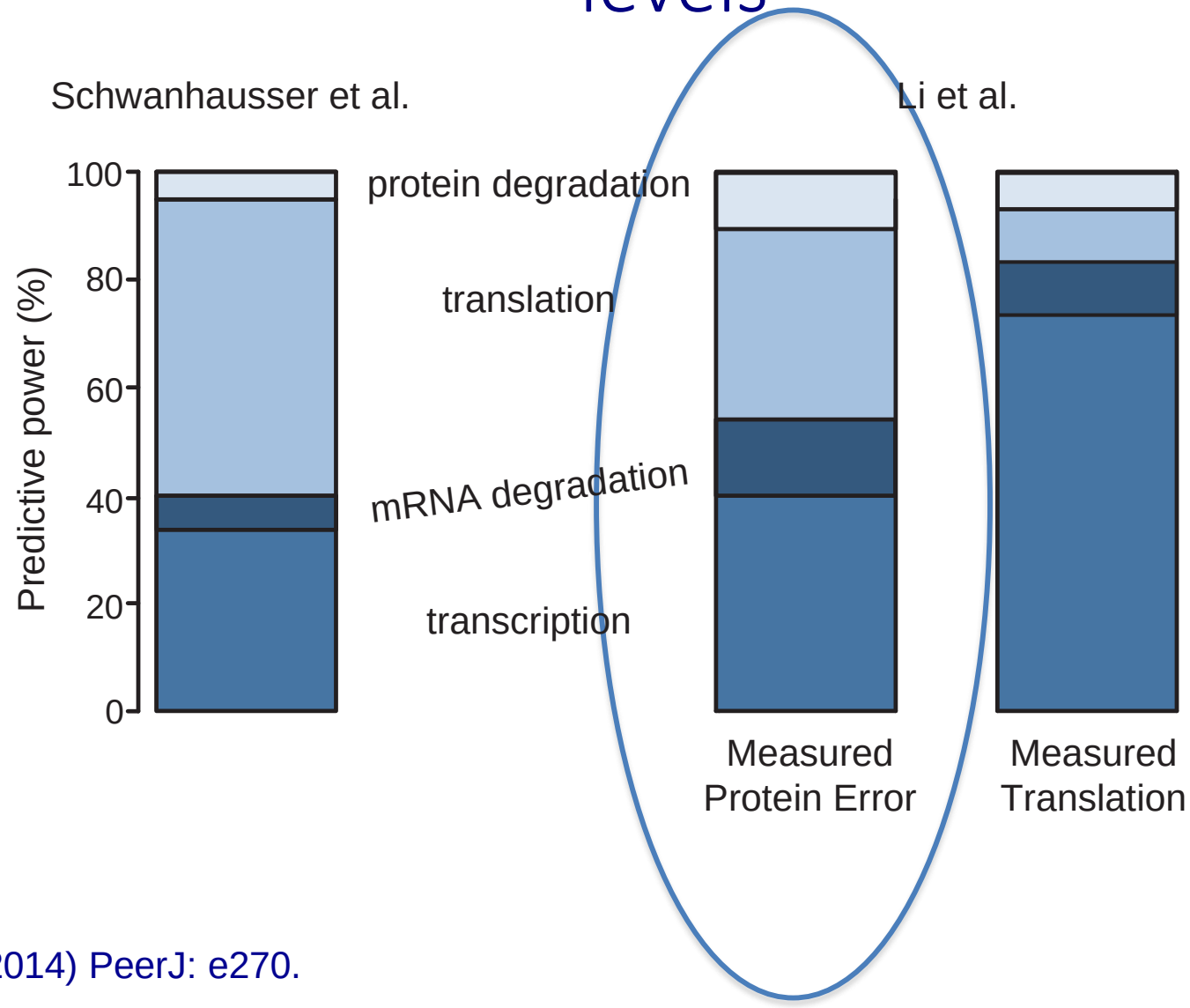
Li's corrected protein estimates





Slide from Jessica Li

After correction of nonlinear trend: Translation is less important in determining protein expression levels

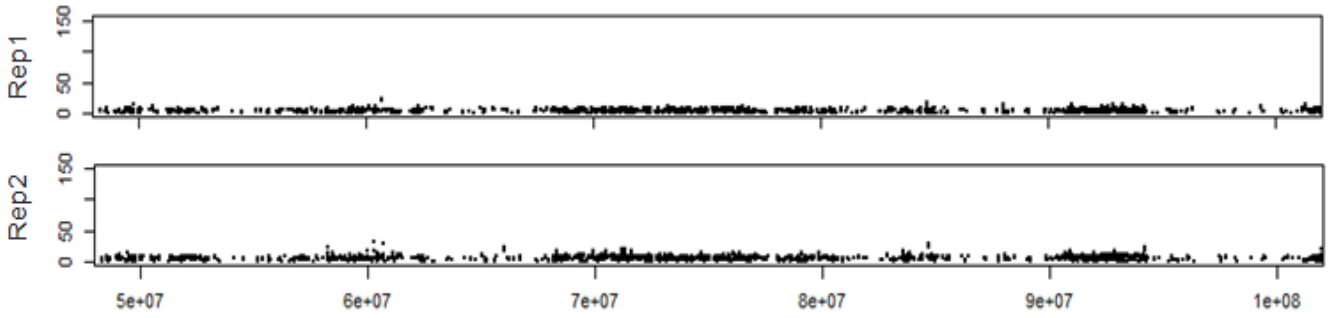


After more correction of measurement errors: Transcription is the dominant step determining protein levels

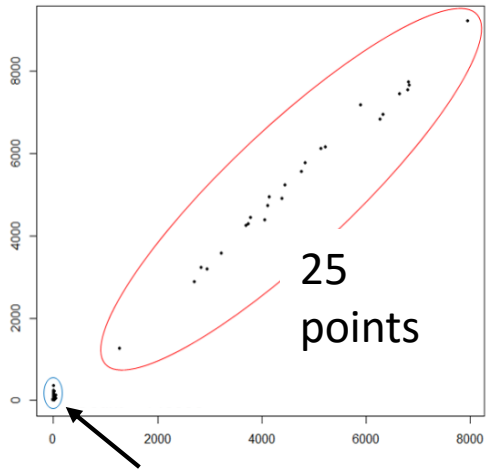
Li et al. (2014) PeerJ: e270.

Correlation needs normal assumption

Input ChIP-seq data

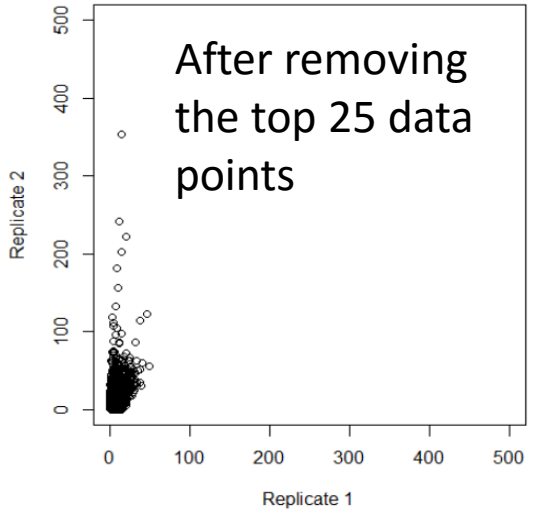


Pearson=0.997



146993 points

Pearson=0.086

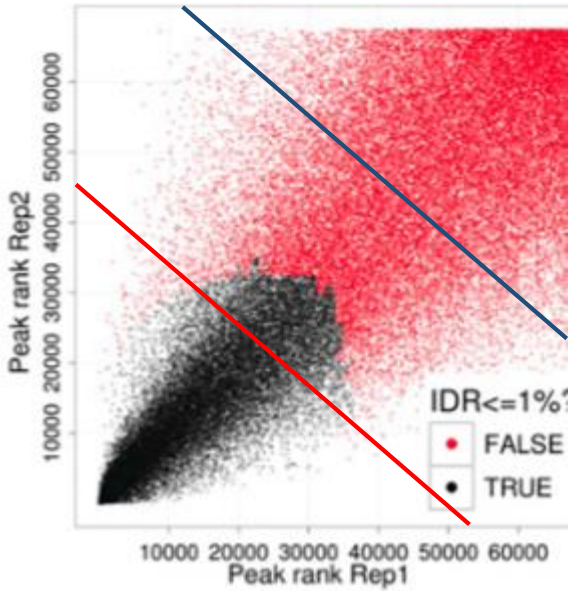


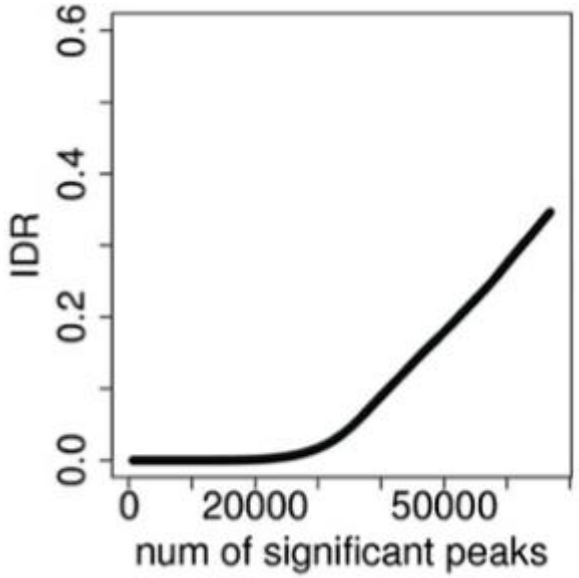
A high correlation can be caused by a fake linearity due to heterogeneity

This is common in high-throughput data, as it is a mixture of signal and noise

Correlation between findings depends on the threshold of significance

	Rank(X)	X	Y	Rank(Y)	
	1	10	31	1	
	2	9	30	2	good
Signal	3	8	27.9	4	agreement
	4	7.5	28.1	3	
		...	...		
		...	...		
	98	0.6	10.7	90	
Noise	99	0.5	10.8	85	bad
	100	0.4	11.1	75	agreement





An assessment based on correlation is confounded by threshold choices
IDR (irreproducible discovery rate) is more suitable

Li et al (2011)

Remedy

- Visualization is important -- scatterplots
- Always do some exploratory data analysis
- Understand your definition of reproducibility
- When unsure, ask a statistician

Reproducible ≠ Correct

Errors are often reproducible if they comes from the same source

- Systematic biases are always reproducible
- If we always accept our own beliefs

Correct is more important!