Quantitative gene transcript expression profiling

Selection of current challenges & opportunities

David P. Kreil

Chair of Bioinformatics

Boku University Vienna

Workshop MPI Plön – 10th May, 2011

Peter Sykacek

Smriti Shridhar

Alexandra Posekany

Brian Godsey

Paweł Łabaj

Germán Leparc

Ulrike Mückstein

Nancy Stralis-Pavese

WWTF

www.bioinf.boku.ac.at

Baxter

AIT

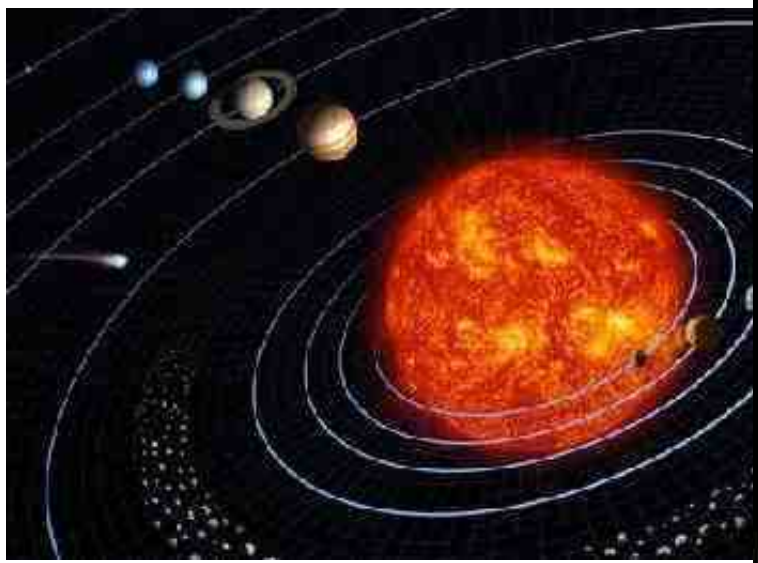
BOKU

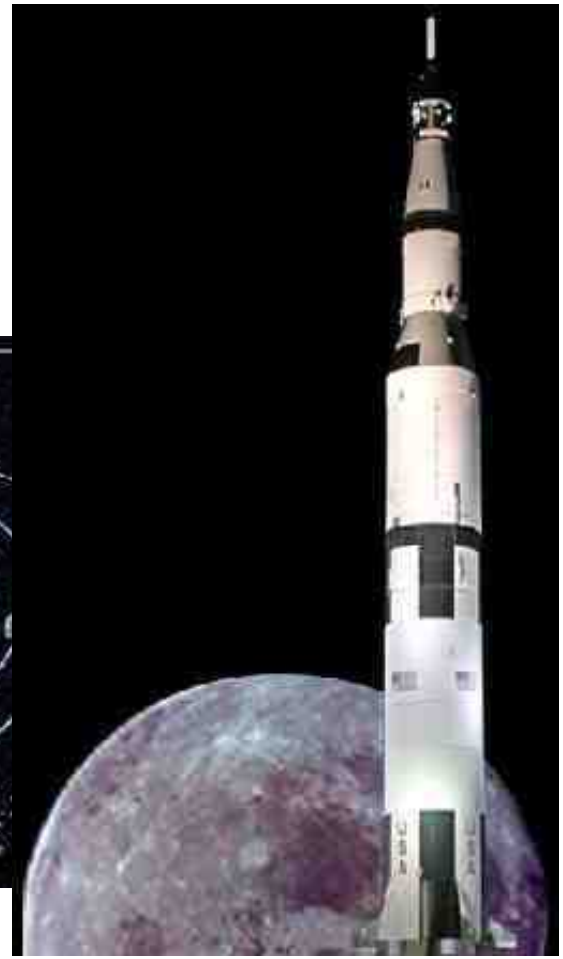
Reductionism

Understanding complex systems

Reduction to the ***essential***

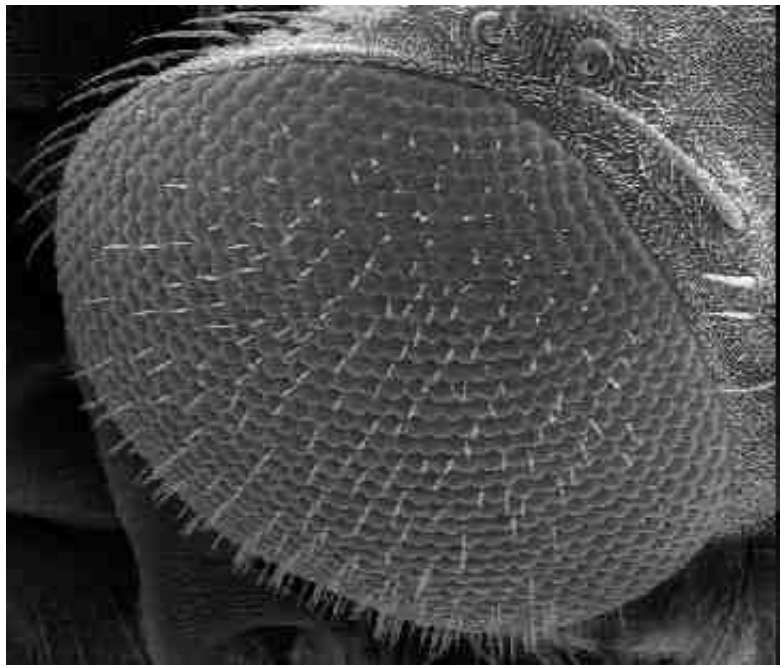
Success in physics:



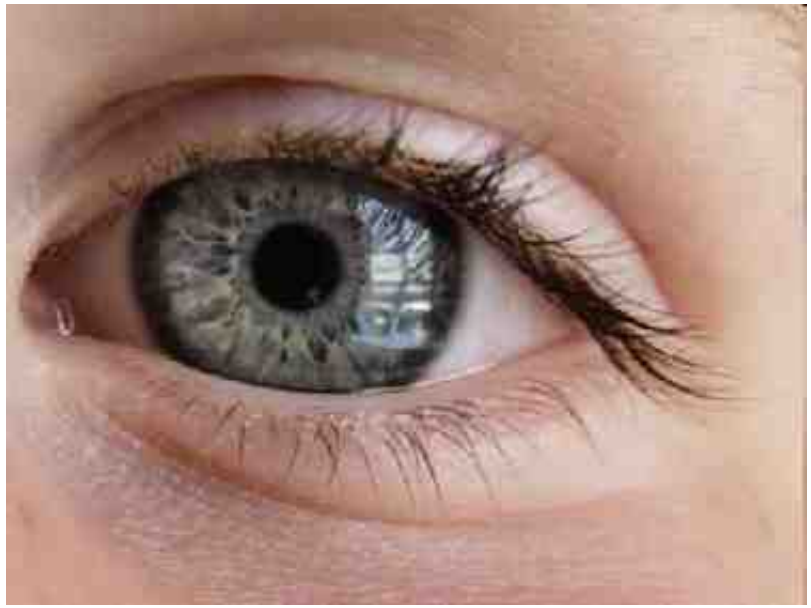


Astrology *vs* Newton's mechanics

Success in biology

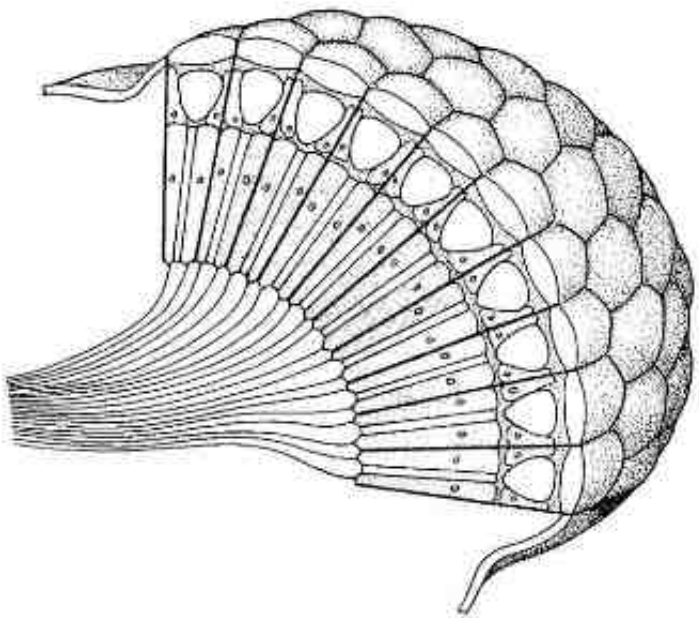


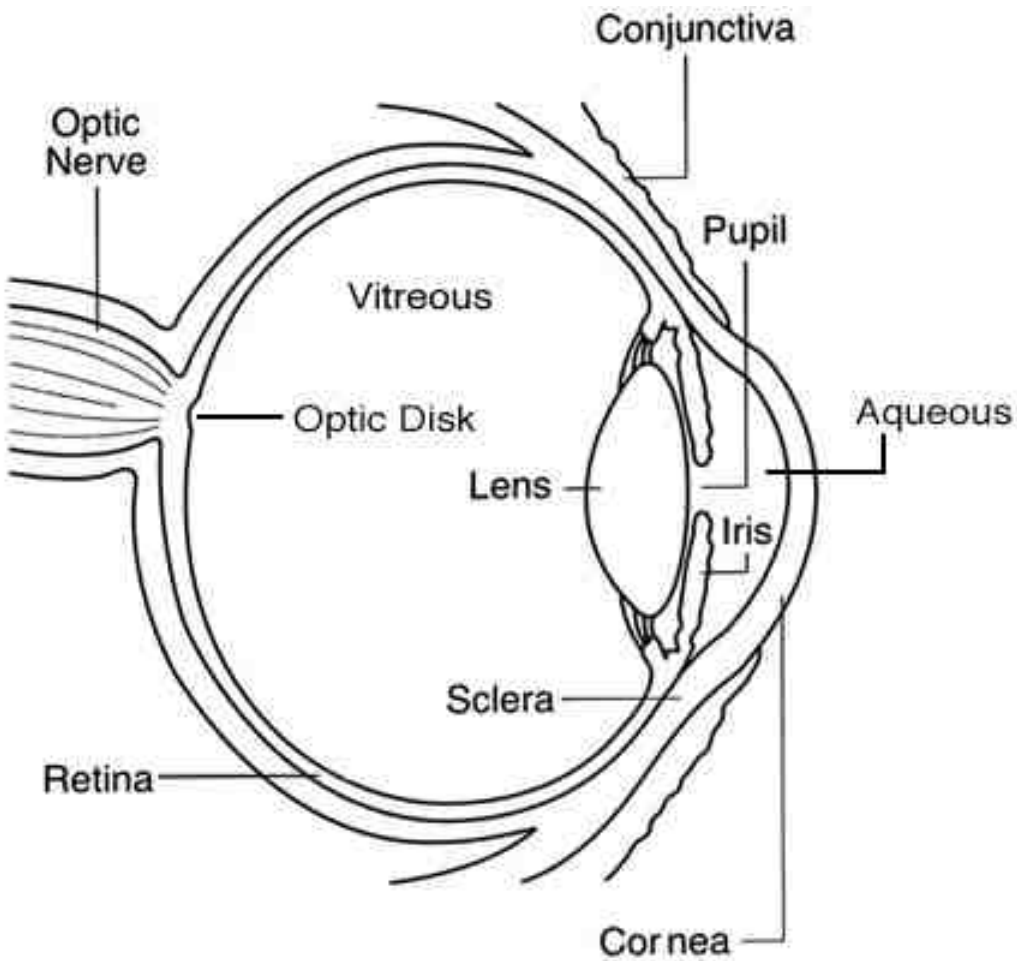
Compound eye



Camera eye

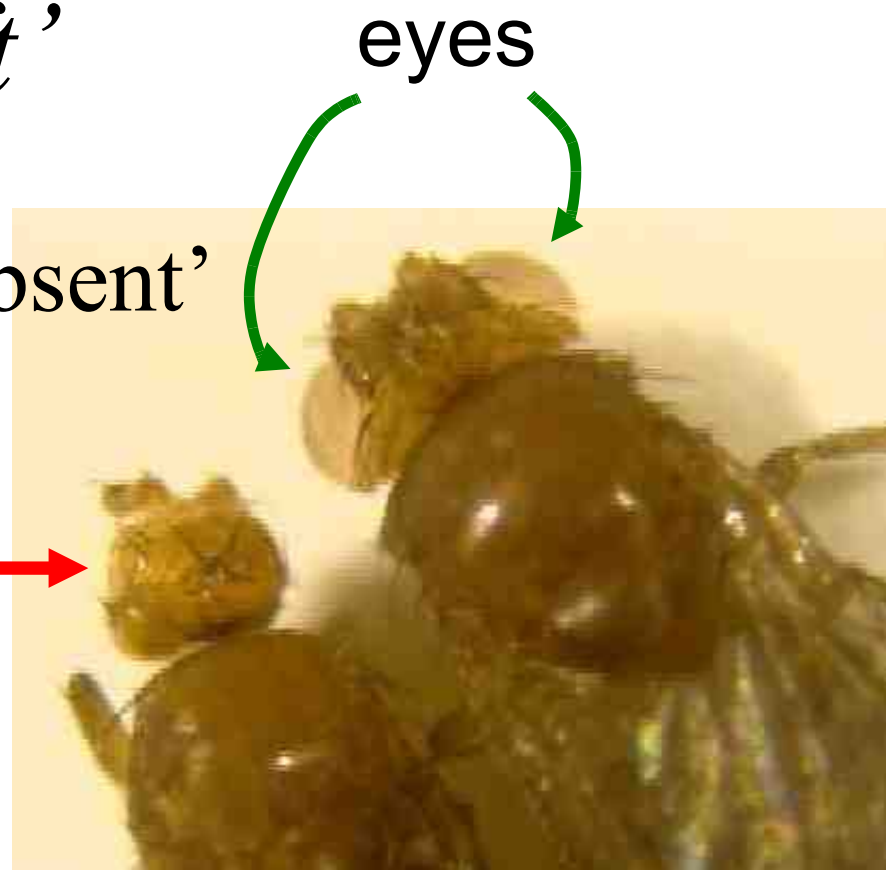
Completely different structures!





Drosophila 'eya/clift'

- Mutations in the 'Eyes Absent' gene
  → Flies **without eyes**.
  (Bonini *et al*., 1993)



- Mutations in the human homologue '*eya1*'
  → Eye defects in humans.
  (Azuma *et al*., 2000)

Marker identification

- Comparative profiling *Screens*

  → Highly discriminating small sets of genes by Feature selection

  Example: Lung cancer classification:

  – 98−100% accuracy with 1 gene for most types

  – Adenocarcinoma:

81%	1
89%	2
94%	3
97% accuracy for	4 features

- Markers / insight?

Crimins & *al.*, *CAMDA*, 2003

Qualitative expression profiling

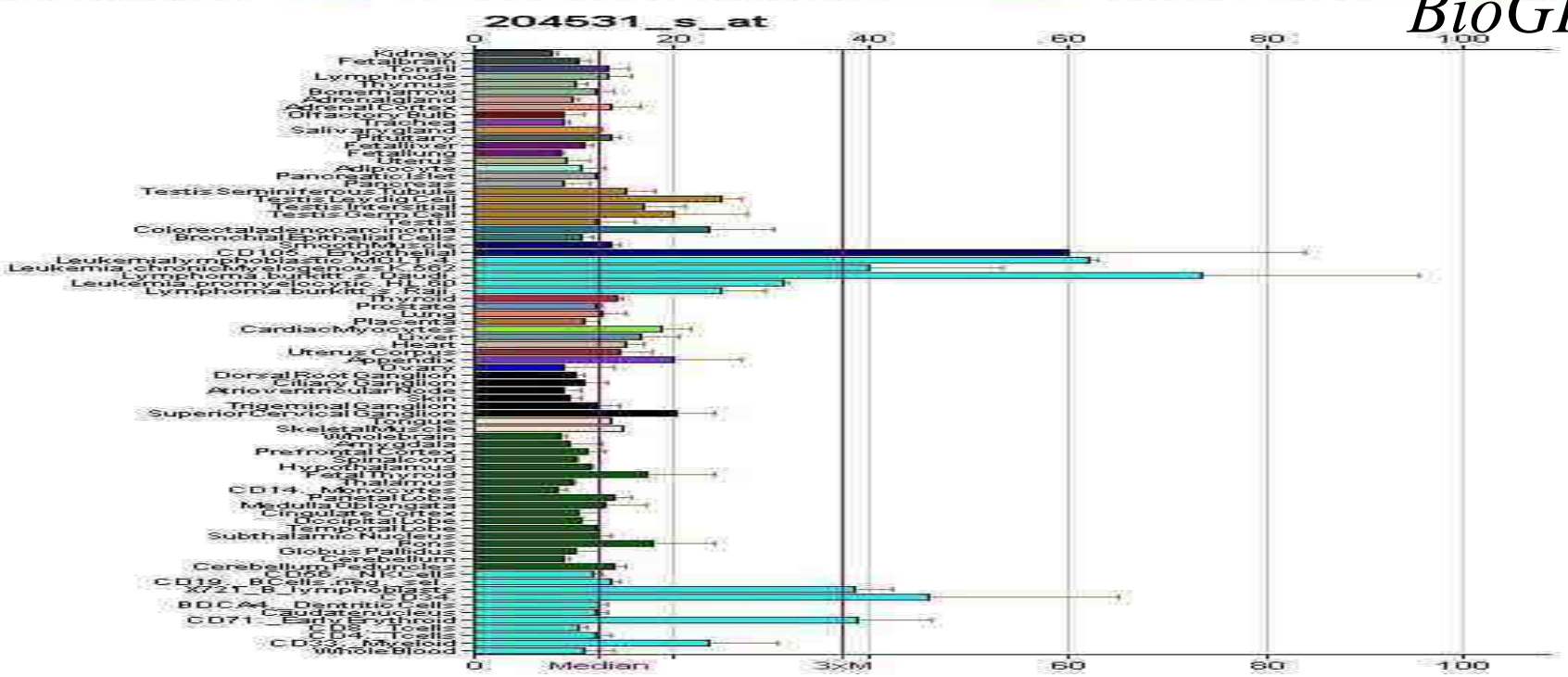
- Application: tissue-specific expression atlas
  - Colour coded or bar-plots, presence maps
  - *Example:* Novartis GPS,
    `https://biogps.gnf.org/`

BRCA1 (breast cancer 1, early onset)

Gene expression/activity chart

Human (672) → GeneAtlas U133A, gcrma → 204531_s_at | Scale image to

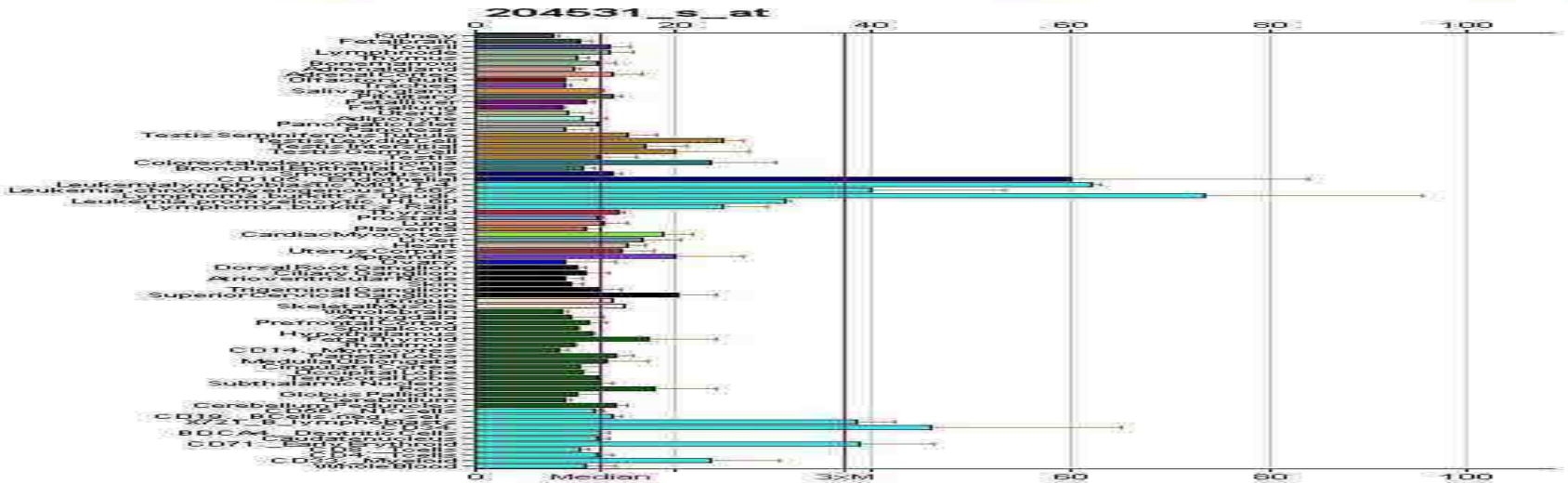
BioGPS, Sample



BRCA1 (breast cancer 1, early onset)

Gene expression/activity chart

Human (672) → GeneAtlas U133A, gcrma → 204531_s_at | Scale image to



Limits of Gene-by-gene approaches

- Traditional gene-by-gene approaches often

No phenotype  /  Lethal

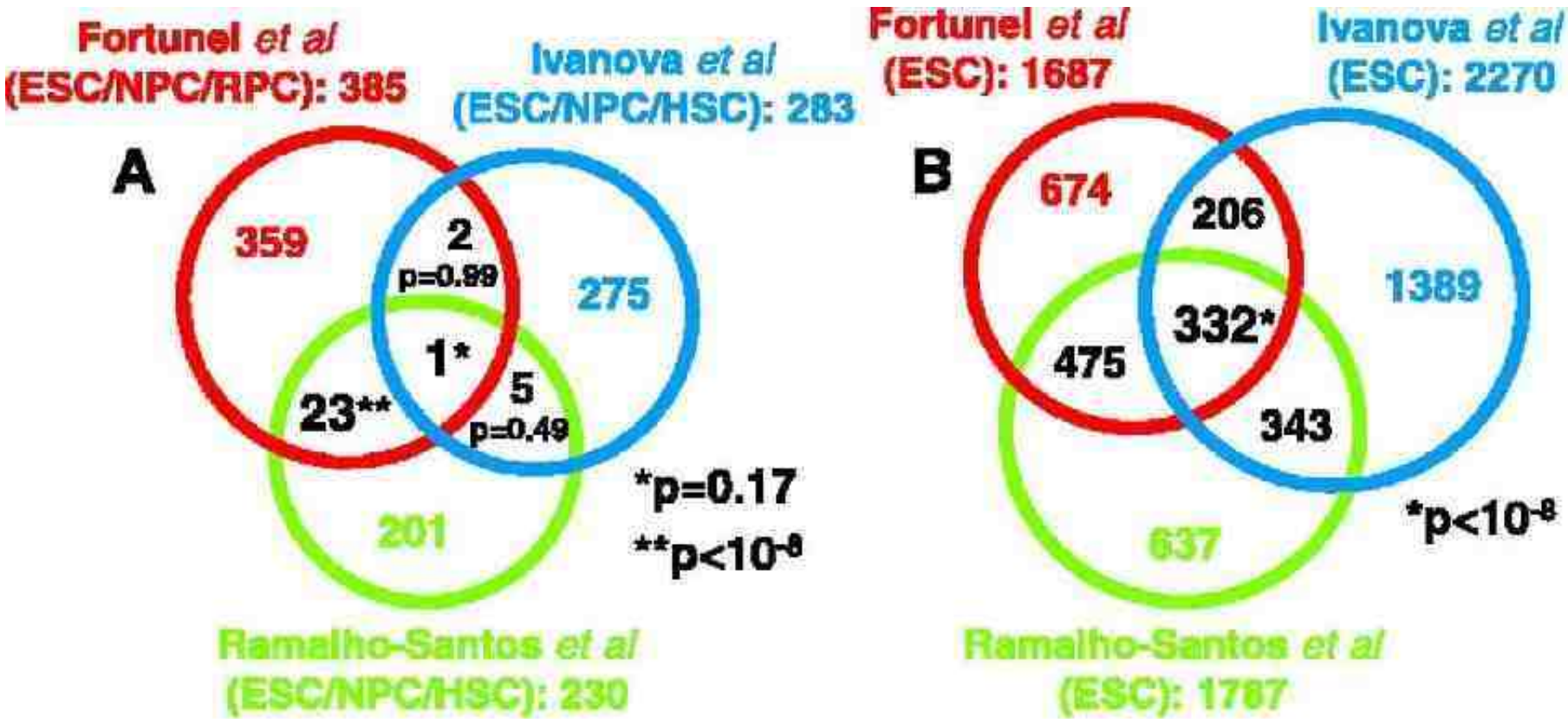
Gene involved in process of interest?

Gene irrelevant or important? (Redundancies!)

- *Similar:* '-omics' screens with '1 gene'-mindset

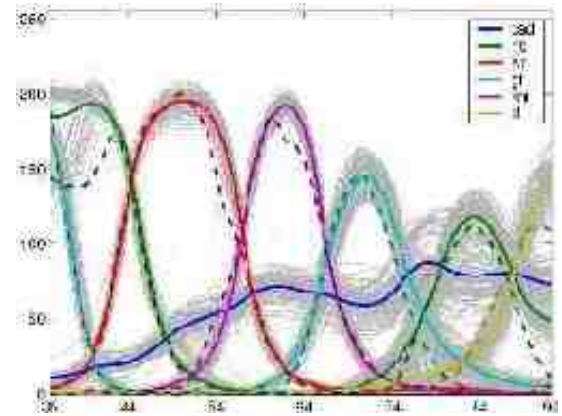
Limitations of qualitative screens

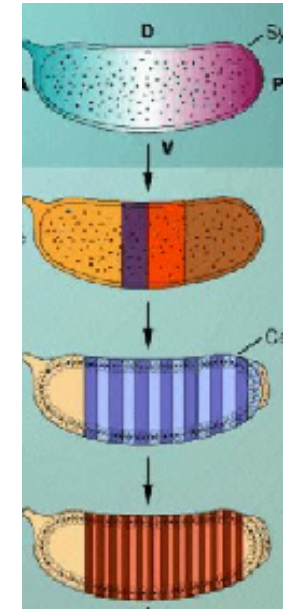
Example: `Stemness genes´: (Fortunel *& al.,* 2003)



Understanding systems…

interactions

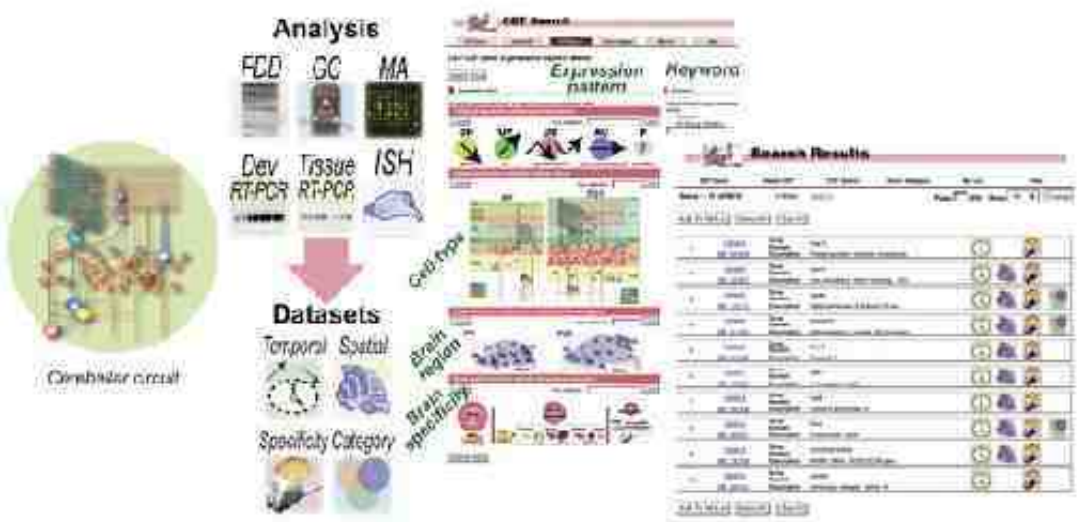


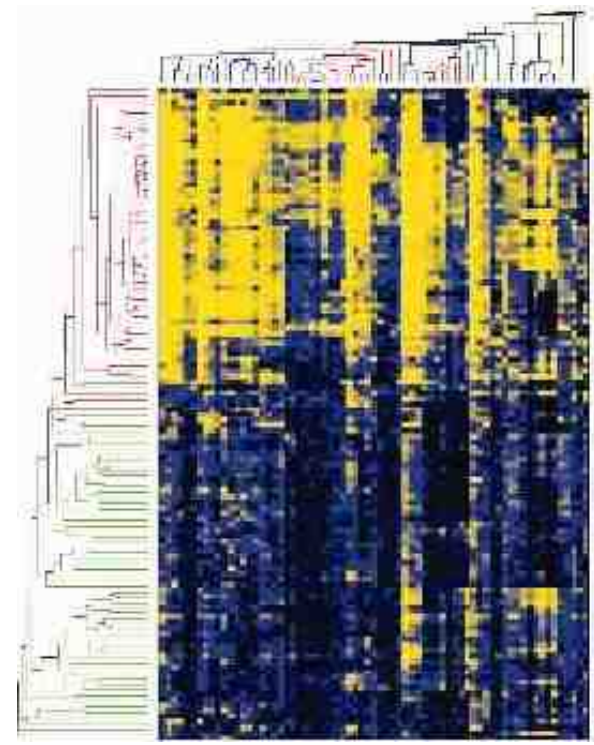


Complementary approaches:

model driven: $\frac{d[x]_{i,j}}{dt} = synthesis - decay \pm transformations \pm transport$

data driven:





Analyses of Quantitative expression profiles

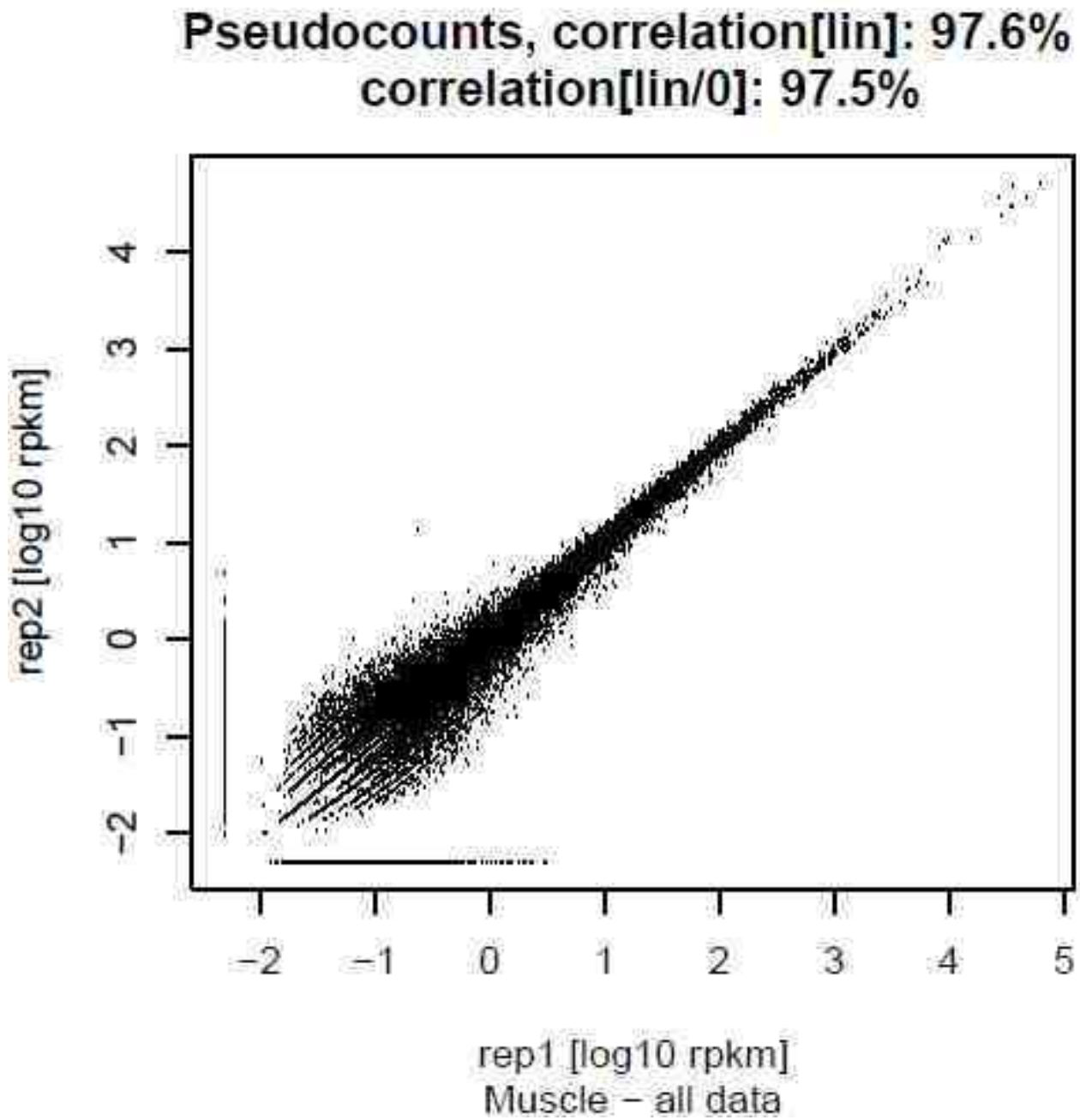
- Permit the detection of subtle multi-dimensional patterns
  → Groups of co-acting genes
- Can identify subtle conditional dependencies on regulators or events
  → Pinpoint opportunities for upstream intervention
- Technical challenge:
  – Can be sensitive to distortion and noise

Understanding gene function

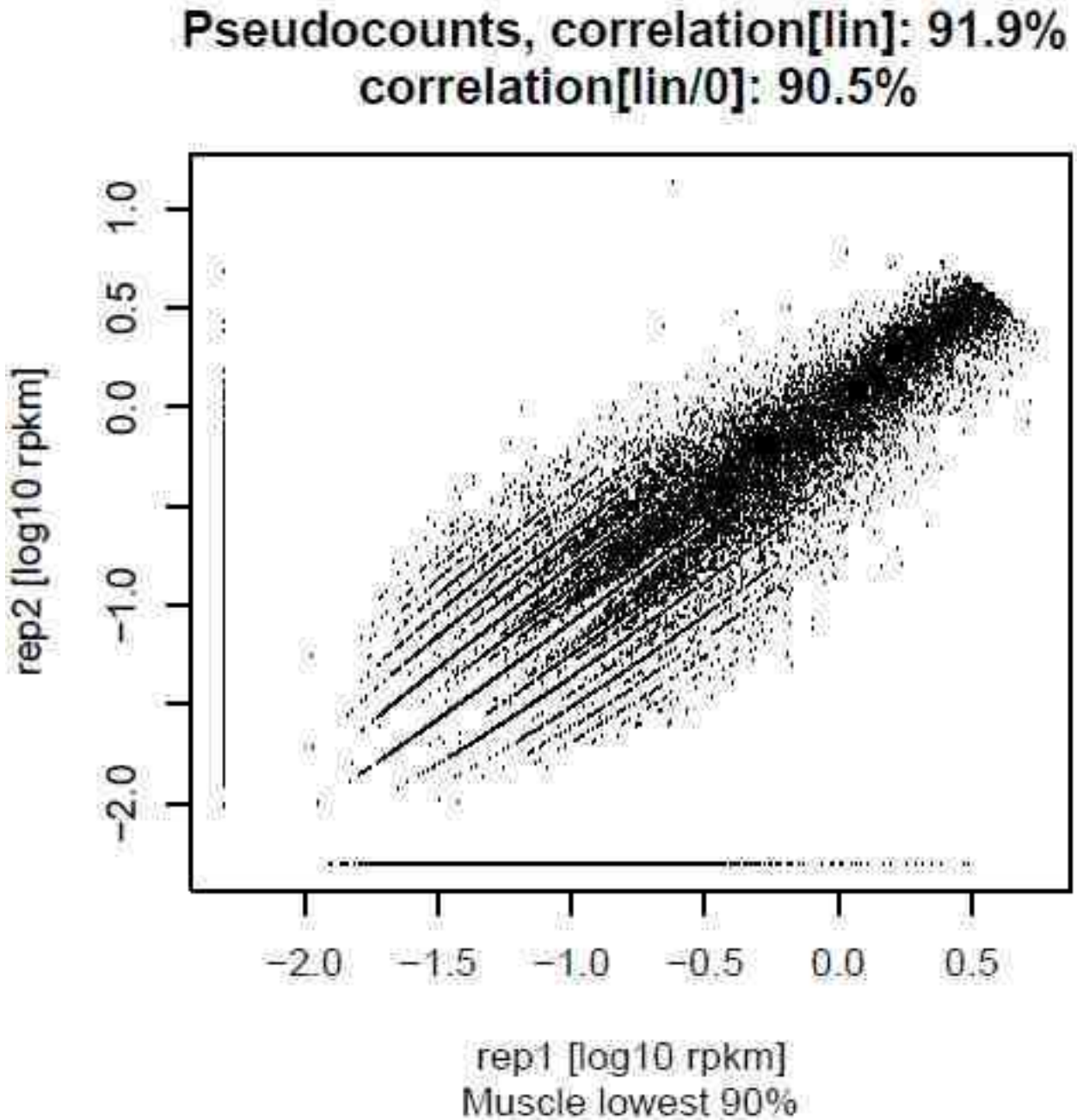
- Sequence → Structure → Function?

- 99% sequence identity in the largest parts of the human and chimpanzee genomes
- More differences in
  - Alternative splicing
  - Regulatory elements, affecting ***gene activity*** (expression)

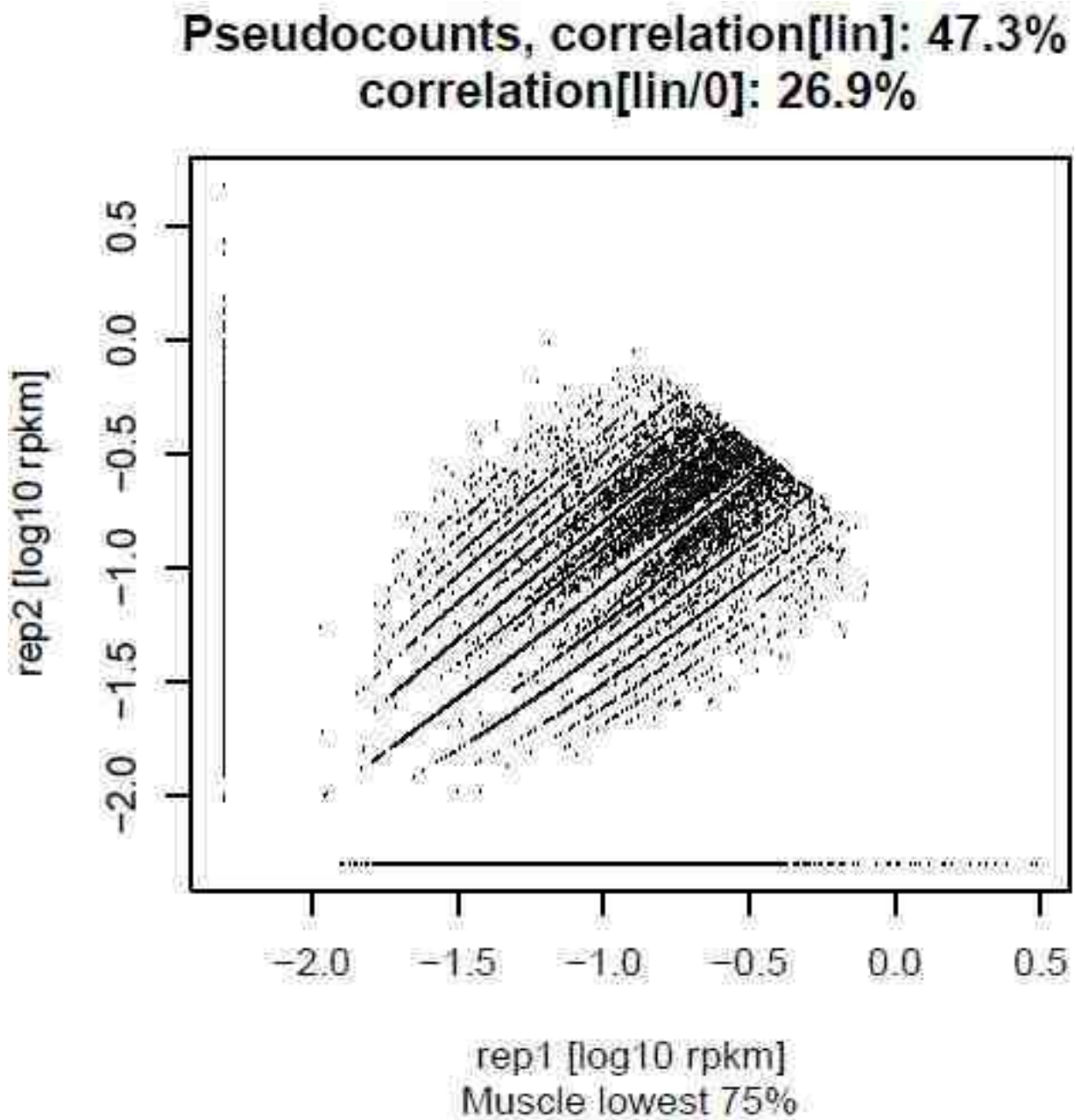
Notes on noise & correlation



Notes on noise & correlation

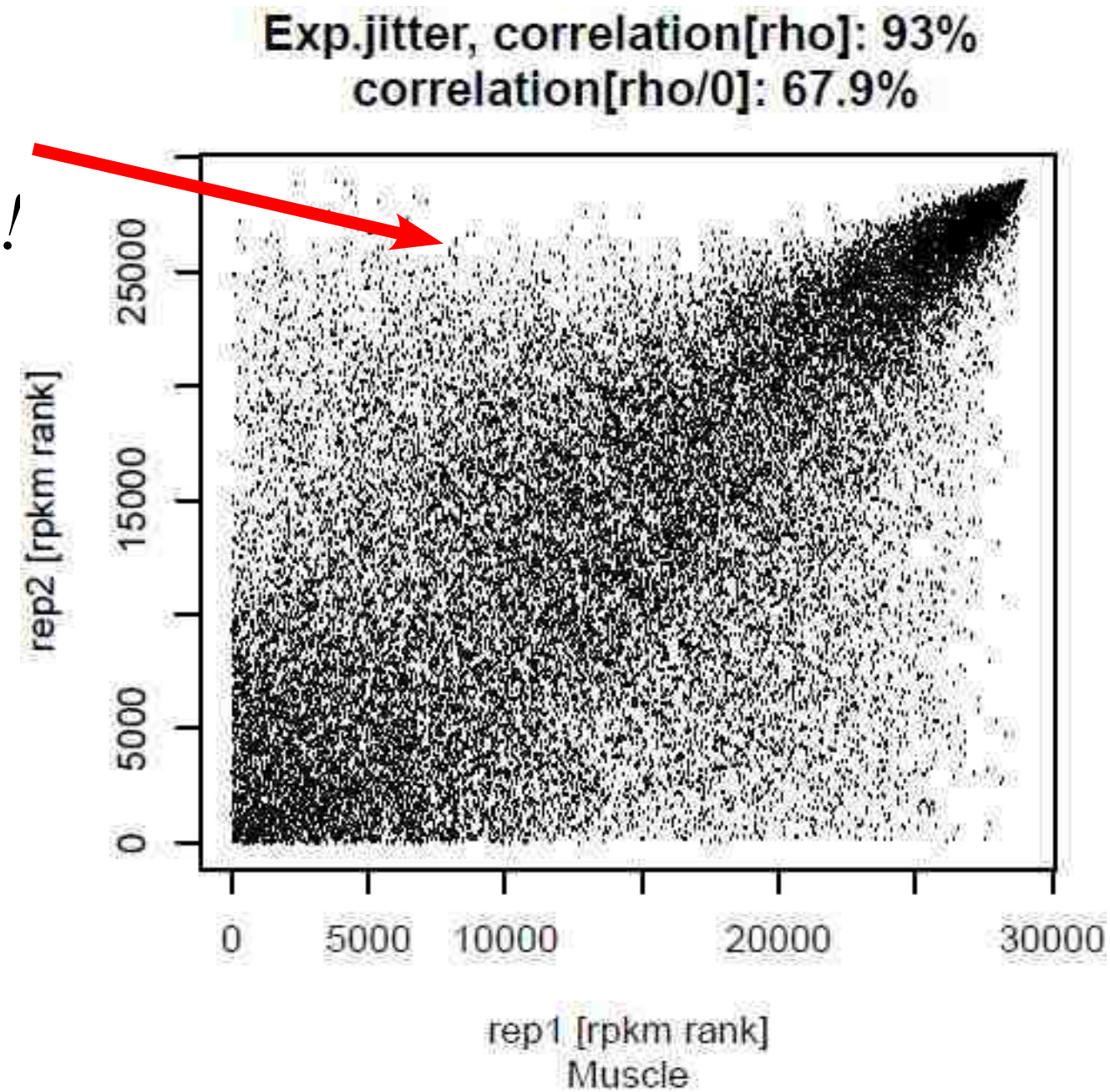


Notes on noise & correlation

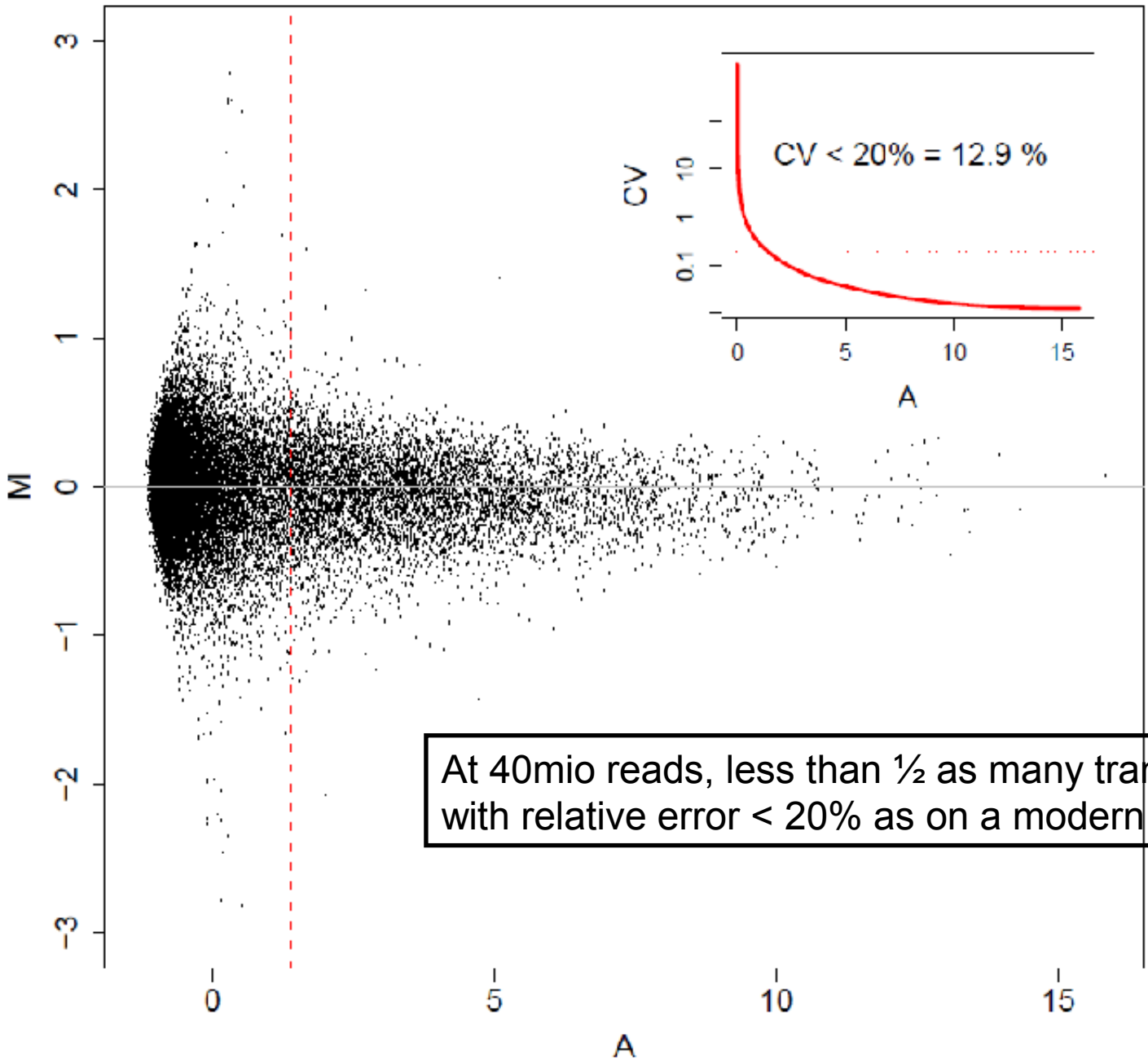


Notes on noise & correlation

- *16 x noise!*

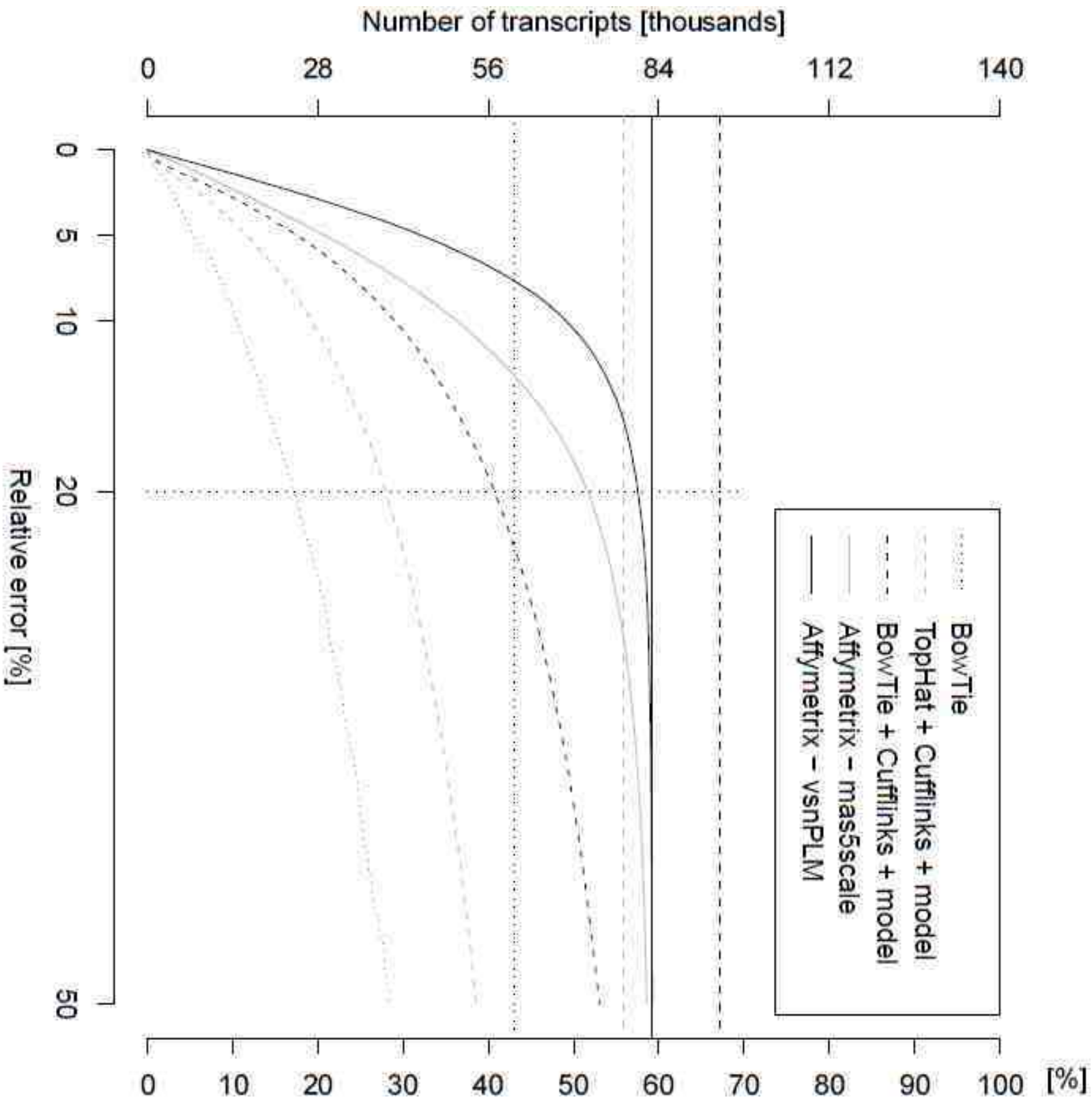


RNA-seq Precision



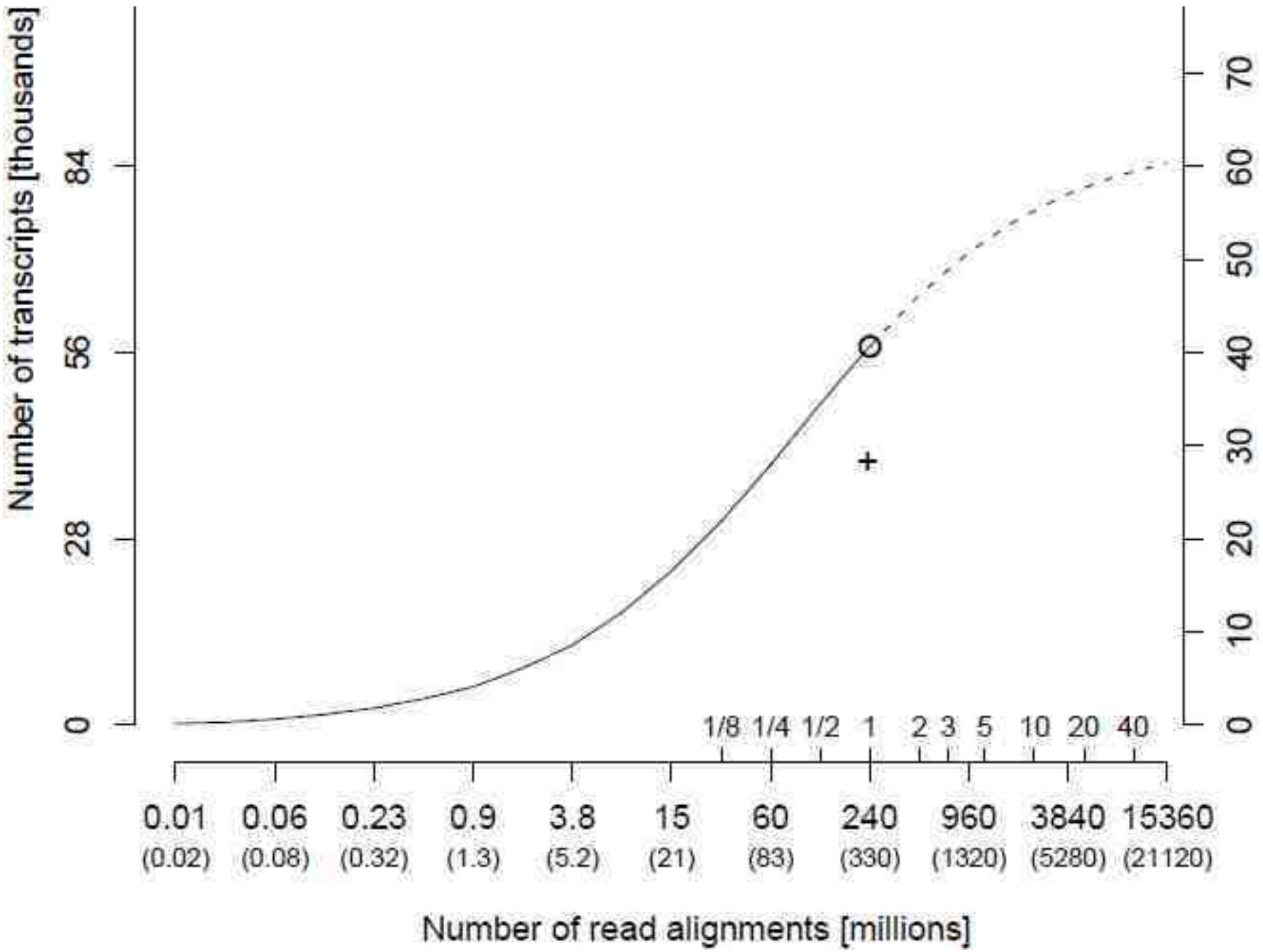
At 40mio reads, less than ½ as many transcripts assessed with relative error < 20% as on a modern microarray!

Labaj *et al.*, ISMB 2011



Labaj *et al.*, ISMB 2011

RNA-seq Precision – Fast Forward



Labaj *et al.*, unpublished

Multi-level source of bias

(M. Sammeth, CNAG Barcelona)

Sources of bias:

platform specific

protocol chemistry 'version' specific

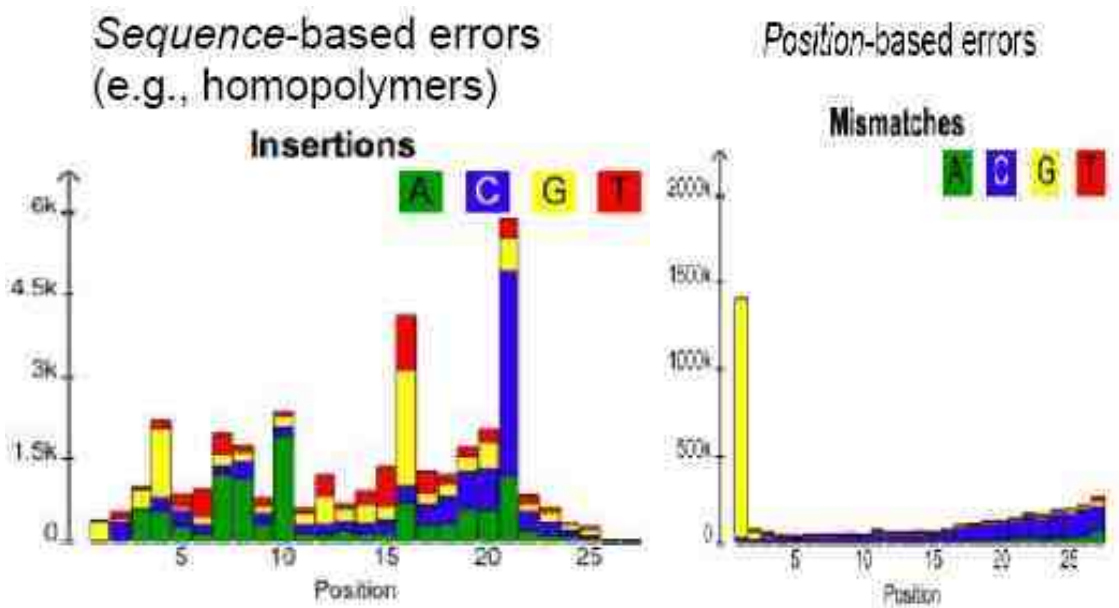
fragment size dependent

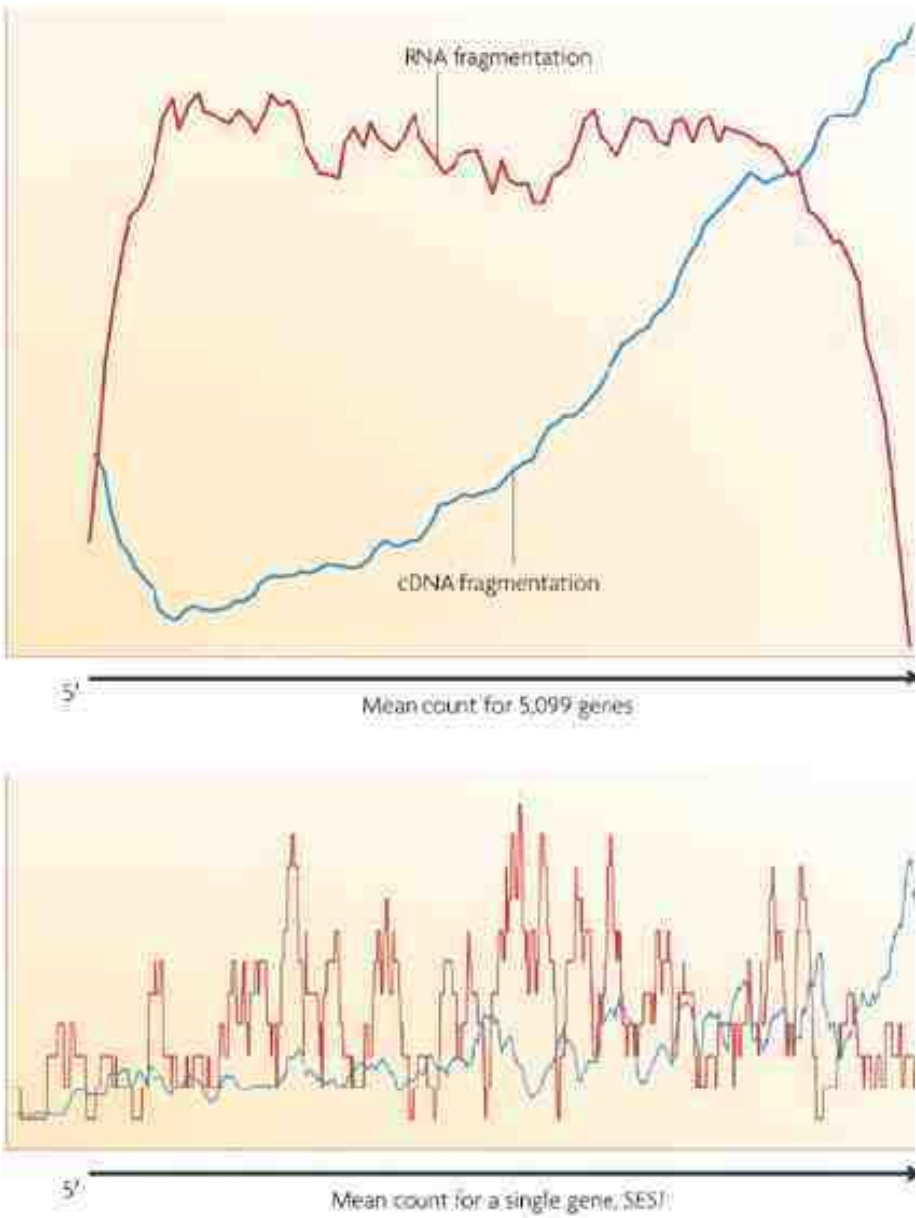
position specific call / insert errors

http://bioinf.boku.ac.at/StatSeq

Note:

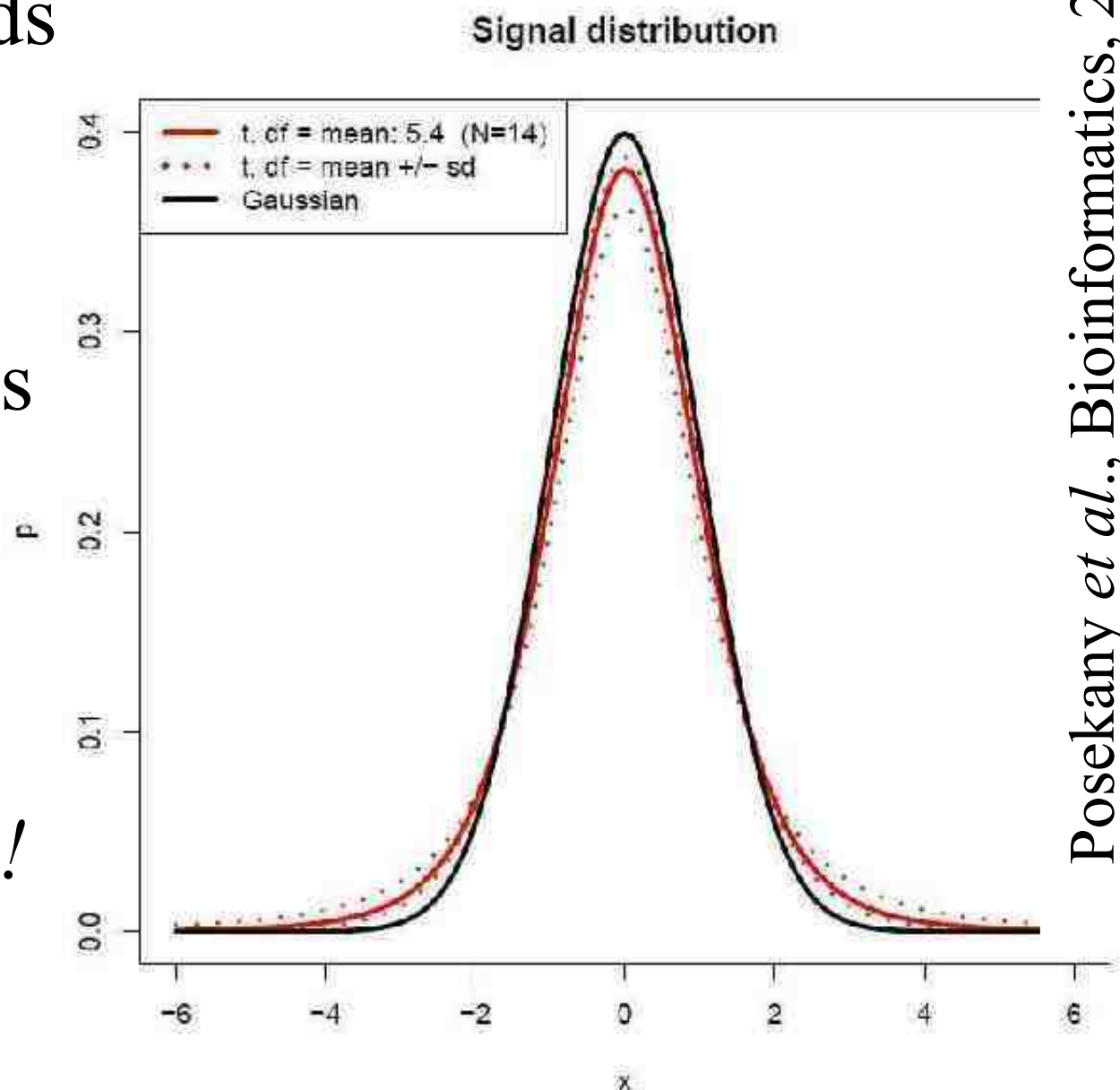
trends *vs* individual coverage patterns





The right noise model?

- Valid inference needs an appropriate noise model.
- Microarray noise has *heavy tails!*
- ...even affects outcome strongly at the ***pathway*** *level!*



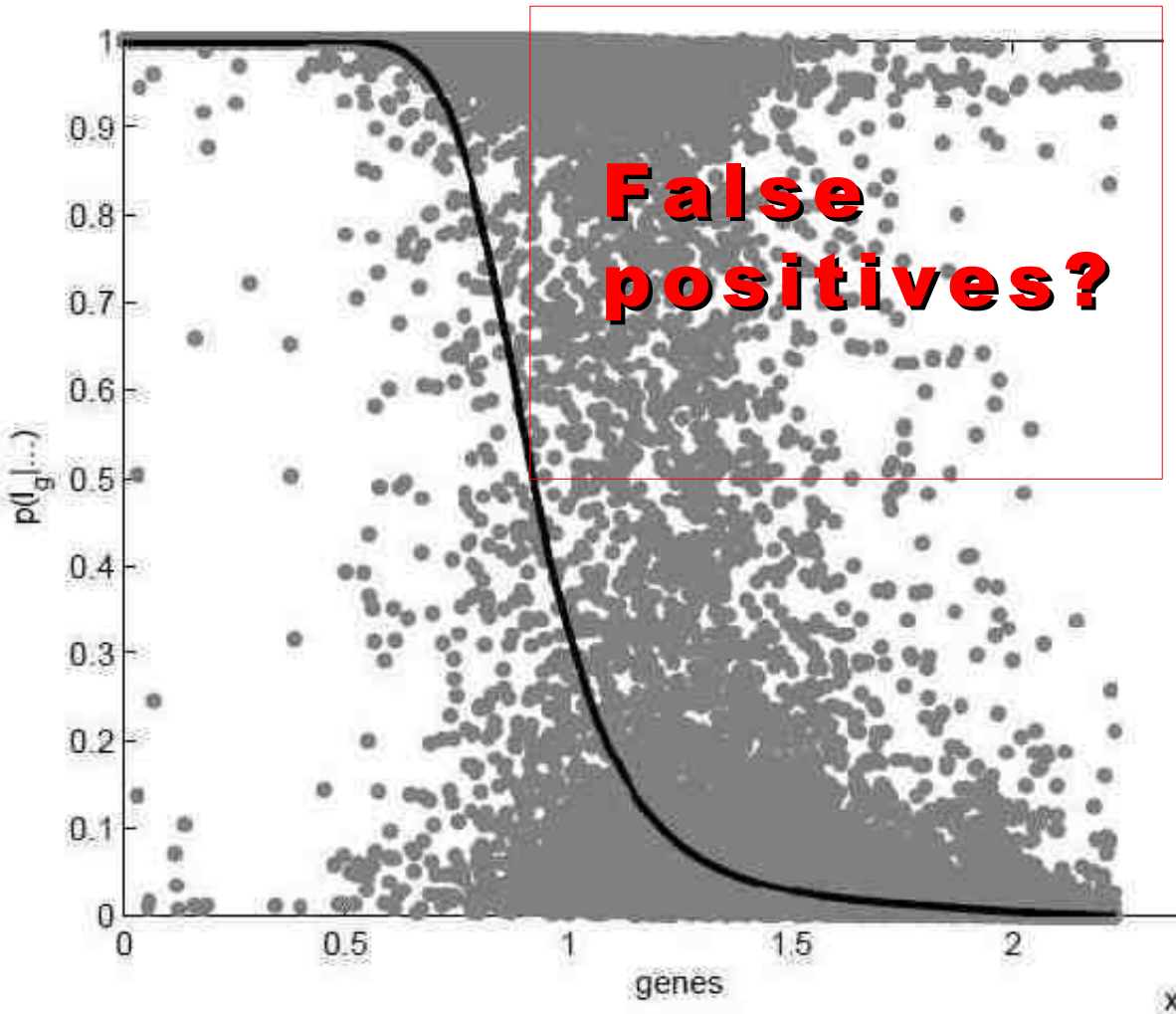
Posekany *et al.*, Bioinformatics, 2010

Model effects on outcome

- Comparison of *p*-values across models:
  - grey: *t*-distribution
  - black: Gaussian

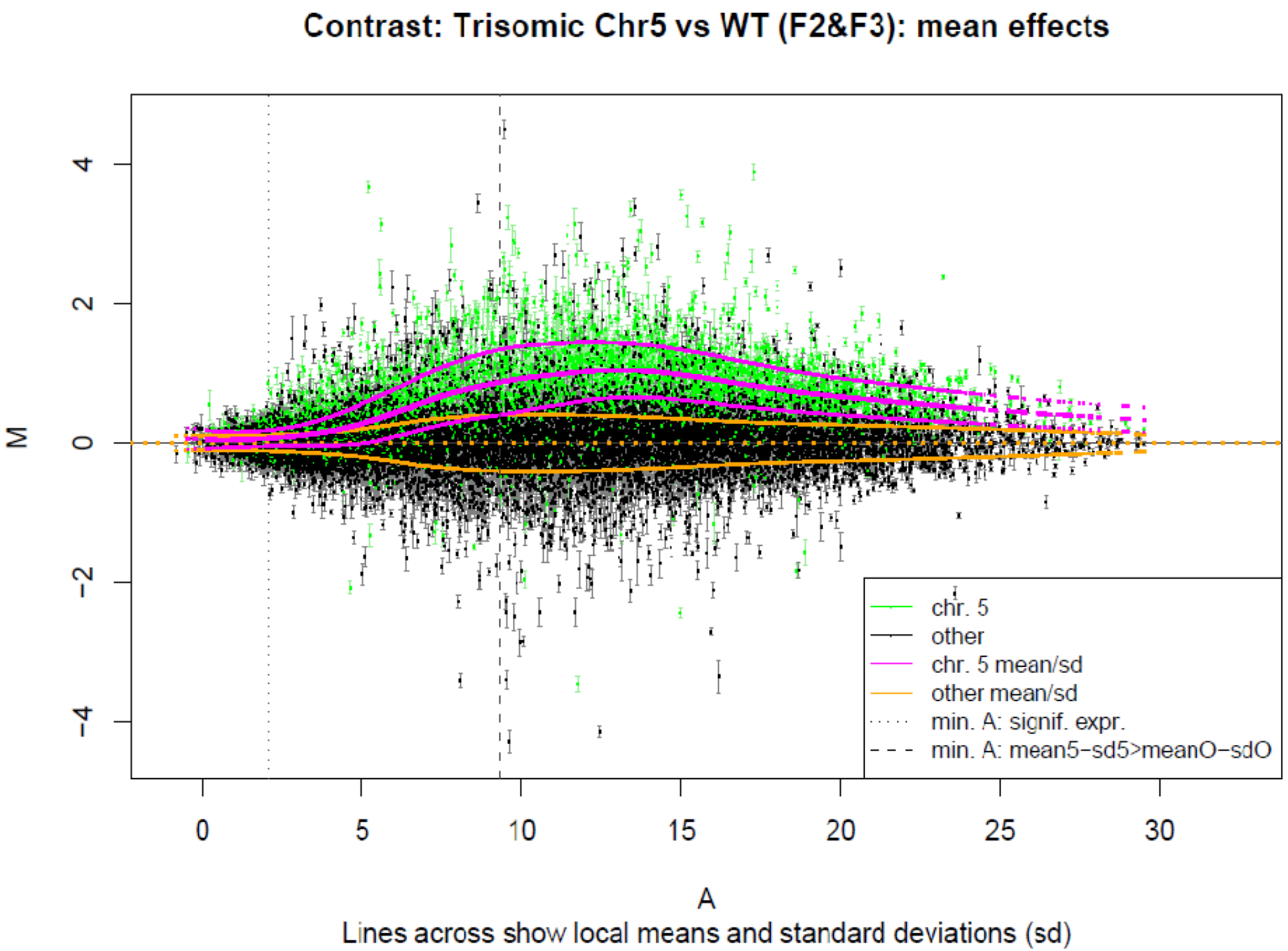
(both sorted by Gaussian distribution)

→ *affects most genes!*



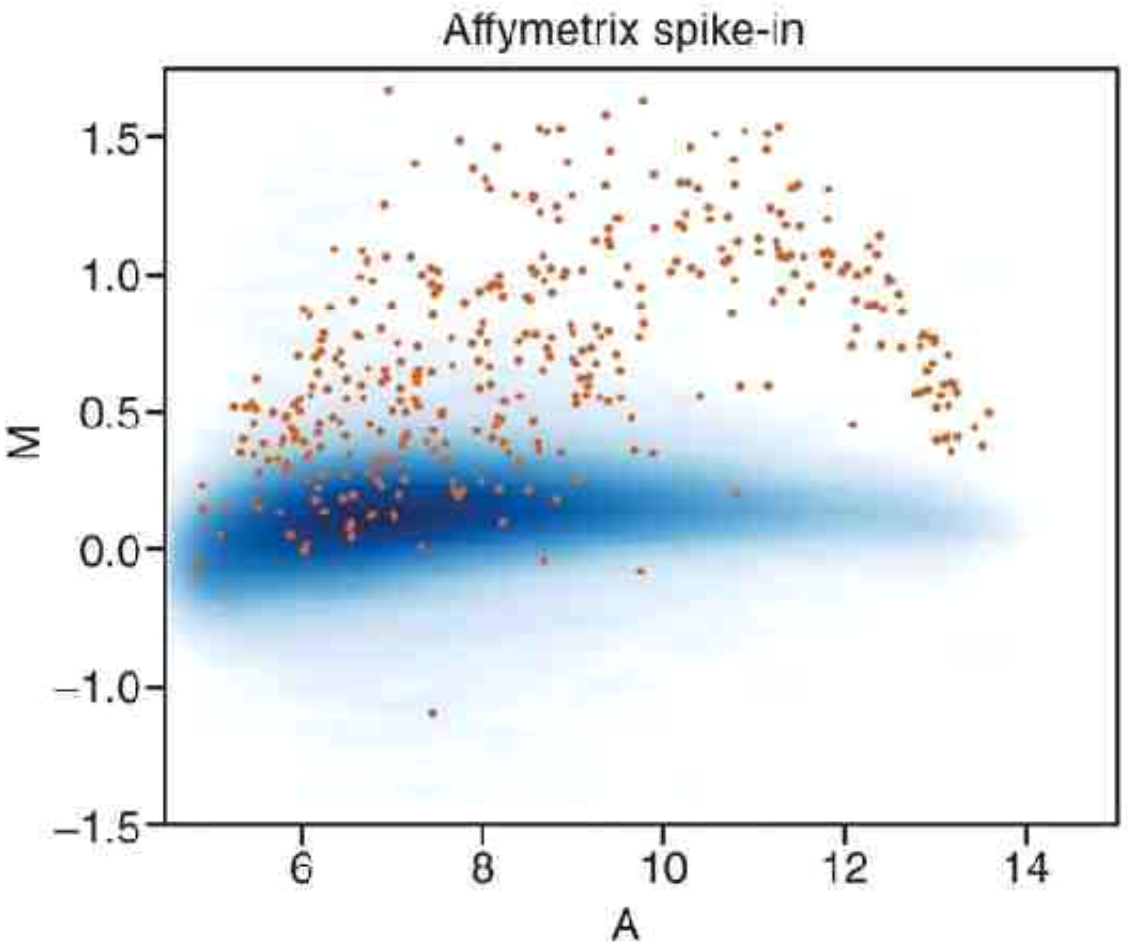
Posekany *et al.*, Bioinformatics, 2010

Nature's very own Calibration experiment:



Hüttel,* Kreil,* Matzke & Matzke, *PLoS Genetics*, 2008

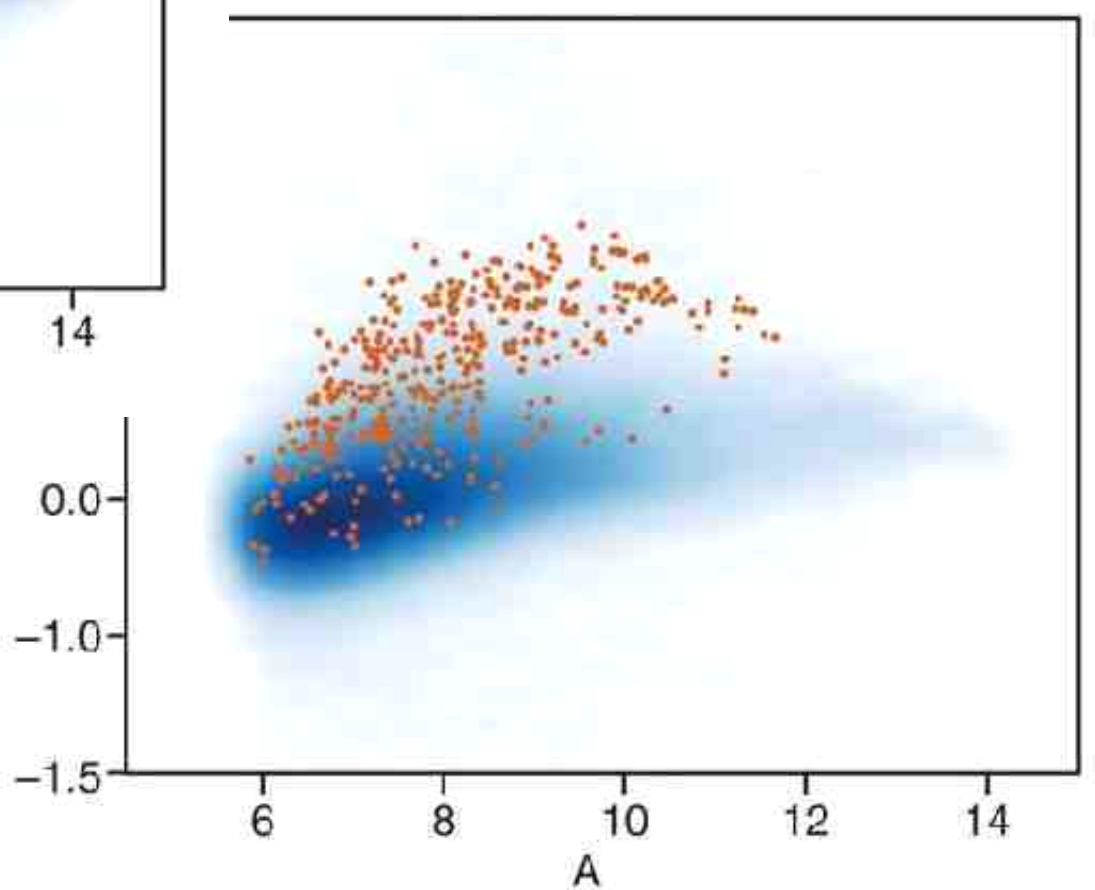
Nature's very own Calibration experiment *vs* Spike-in Series:



Human Trisomy-21
(Down Syndrome)

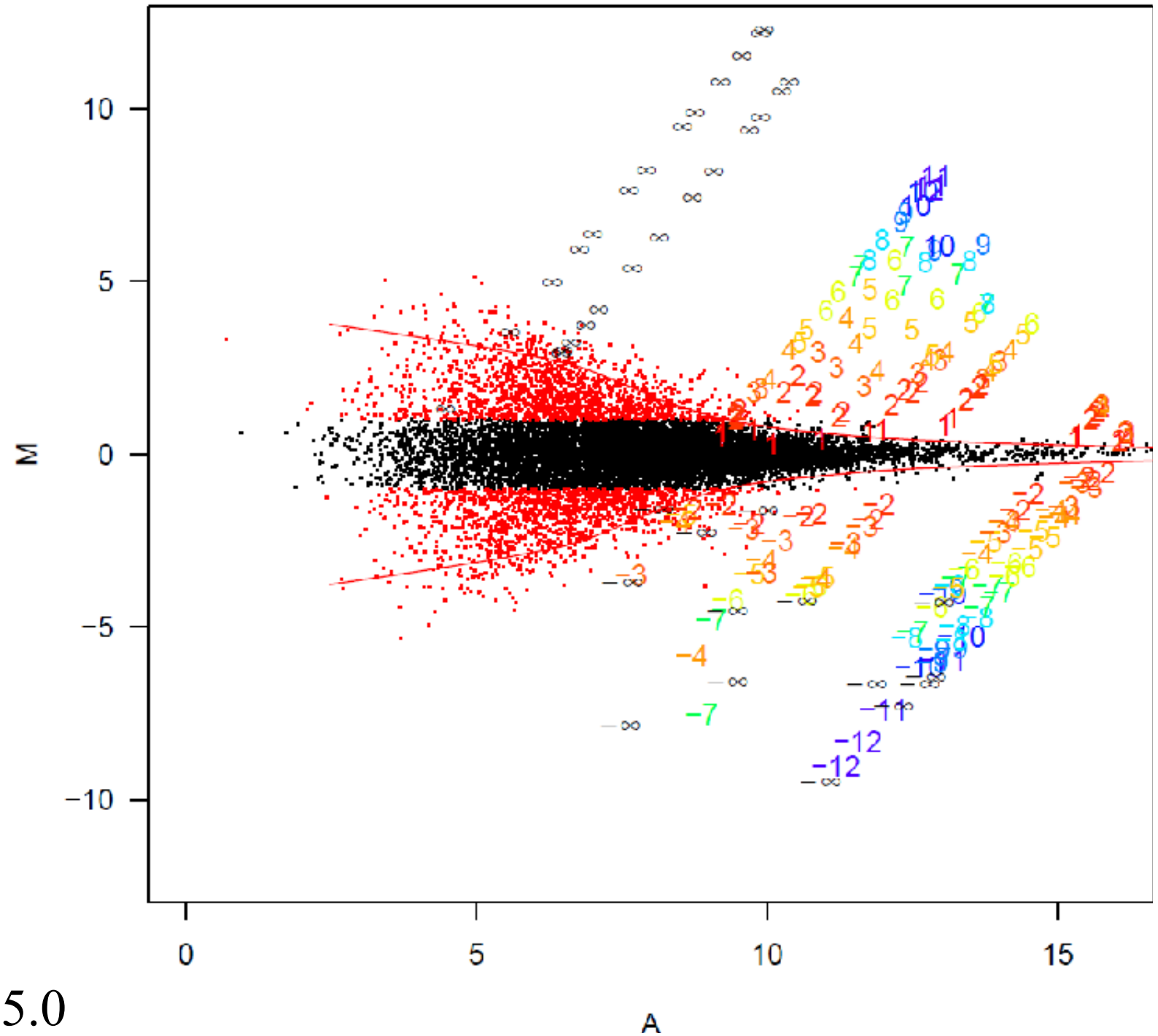
Nominal 1.5-fold change.

Nominal 2.0-fold change spike-in data set.



Irizarry, Cope, & Wu, *Genome Biology*, 2006

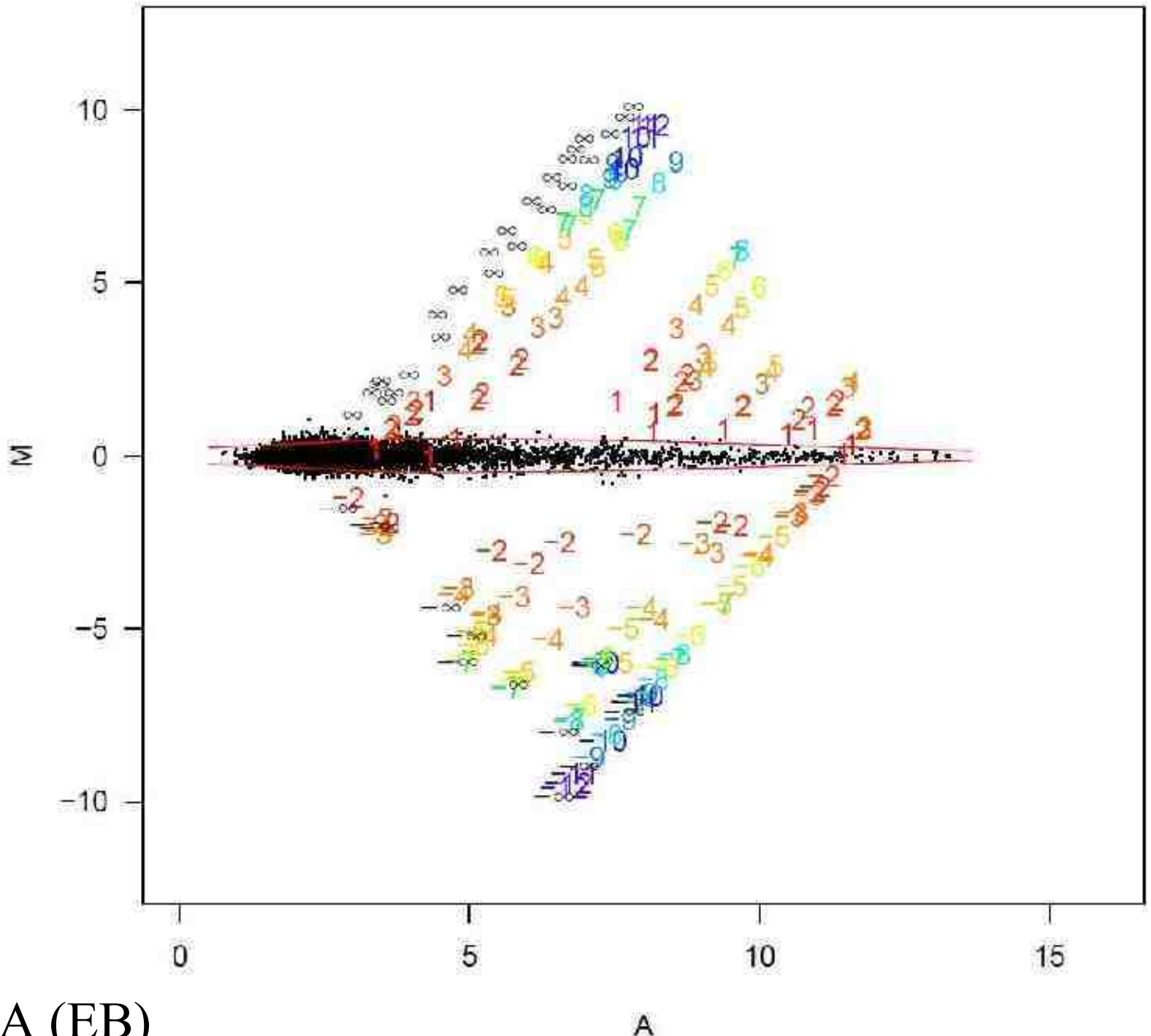
Effect of processing model – Spike-in Series



MAS 5.0

Wu *et al.*, Working Paper 1, 2006

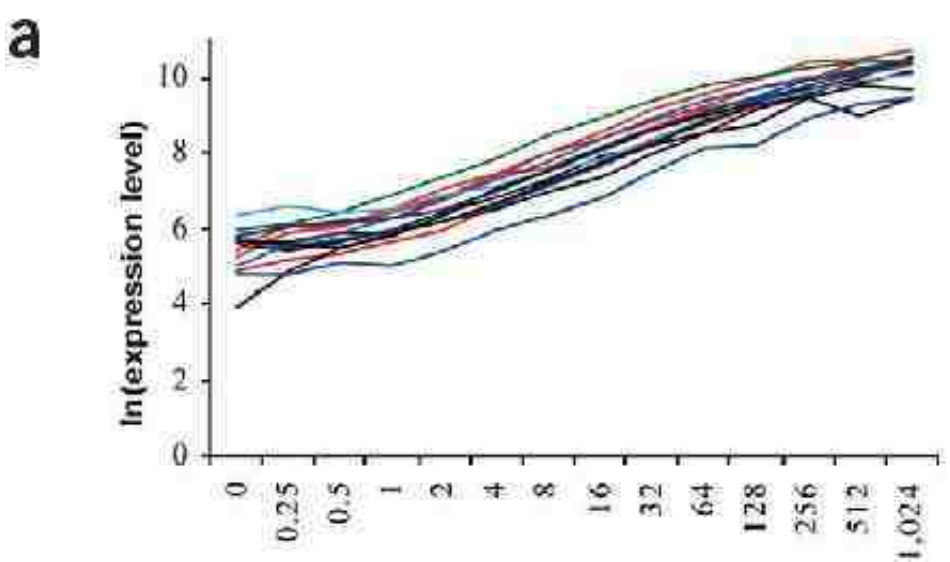
Effect of processing model – Spike-in Series

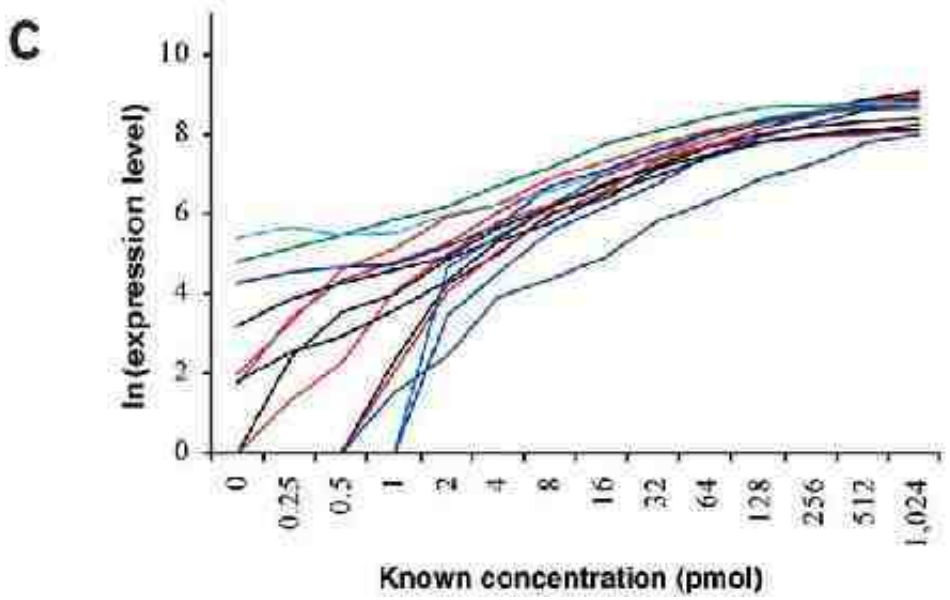


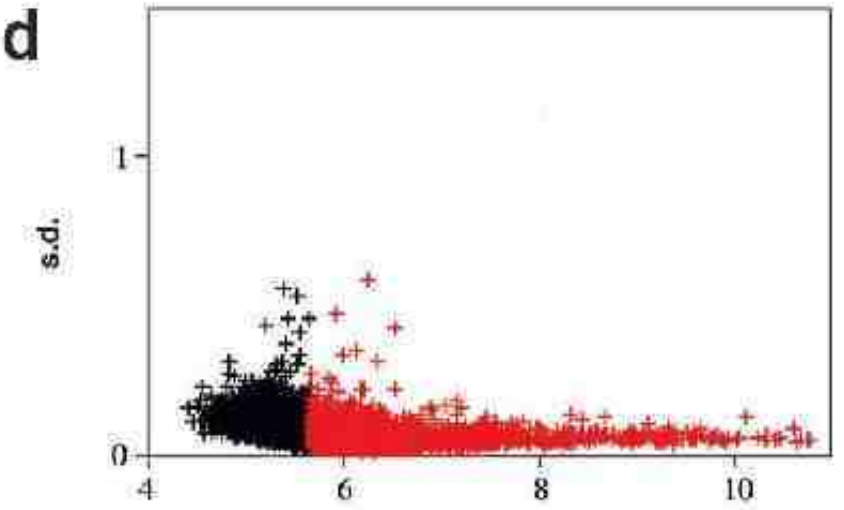
gcRMA (EB)

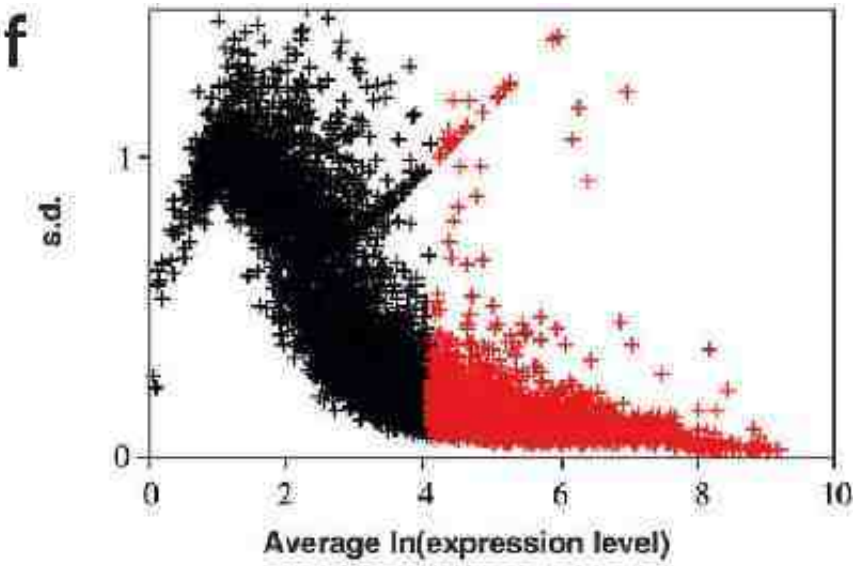
Wu *et al.*, Working Paper 1, 2006

Example of progress achievable by returning to 'low-level' analysis:









Zhang &*al*, 2003

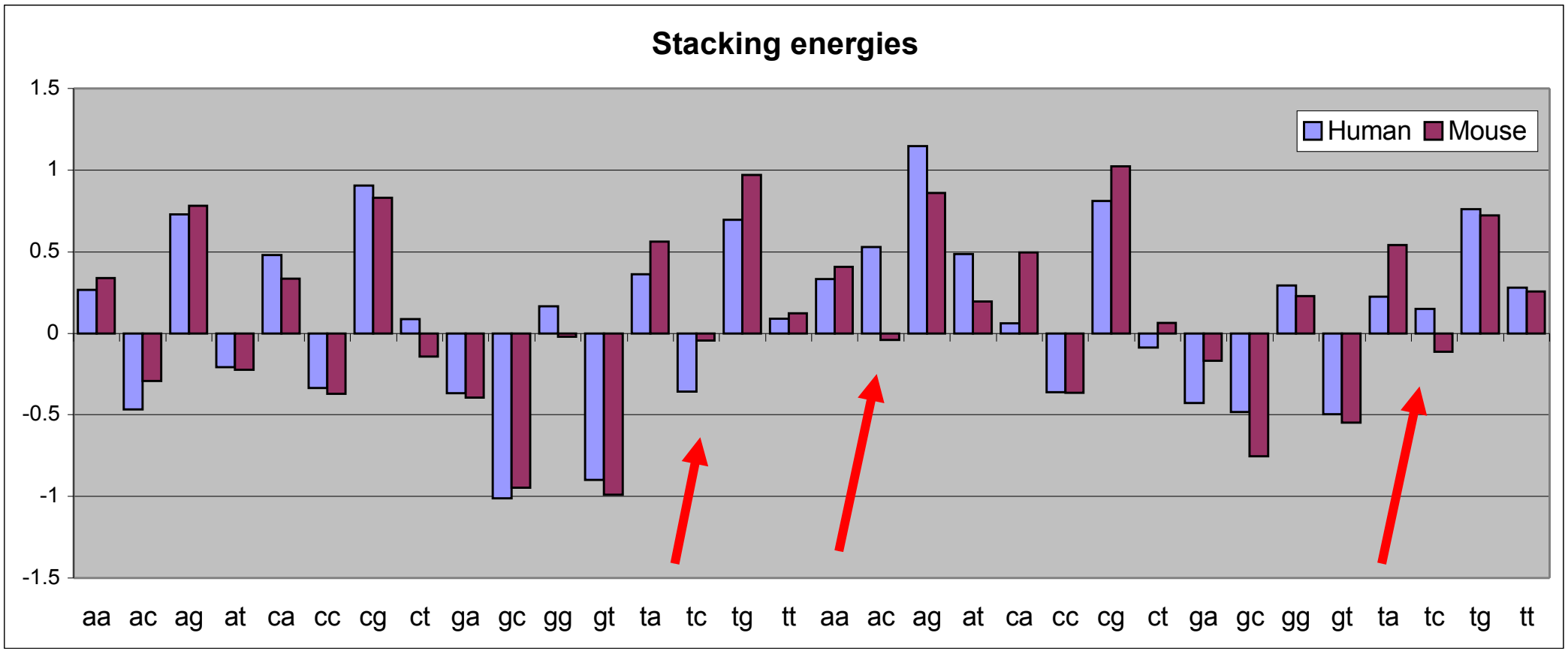
MAS 5.0

→ Dilution series benchmark data show a substantial improvement!

* Similar models should already be used in the design of oligo probes, can be an iterative cycle of better probes & better signal interpretation! (Work in progress at our lab. *Cf.* our paper Leparc *et al.*, NAR, 2009)

* Limitations / challenges: The model fits well, it does not explain well:

Base stacking energies differ between chips, *e.g.*:



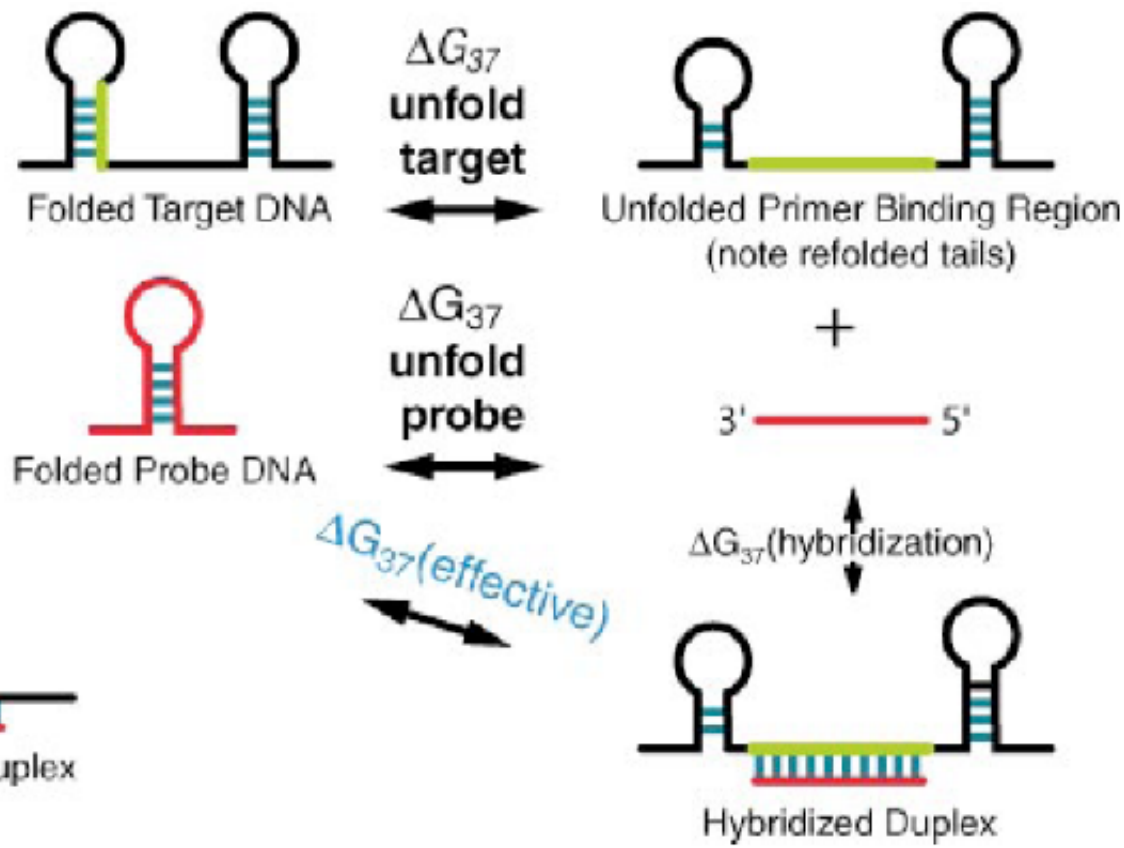
(unpublished)

Some current challenges for probe-level modelling

* Heterogeneous probes (due to *in-situ* synthesis)
* Surface-specific effects
* Complexity of multi-state models

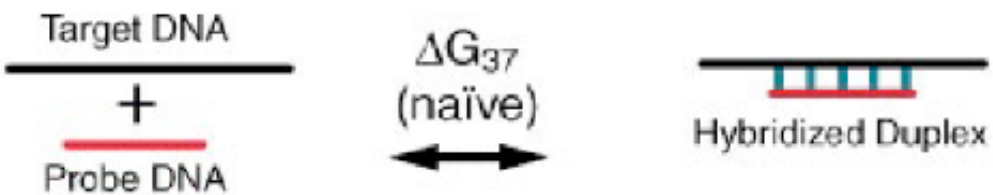
(cf. Mückstein / Kreil, BMC Bioinf., 2010)

Multi-State Model



(Diagrams, SantaLucia & Hicks, 2004)

Two-State Model



Impact of Target structure

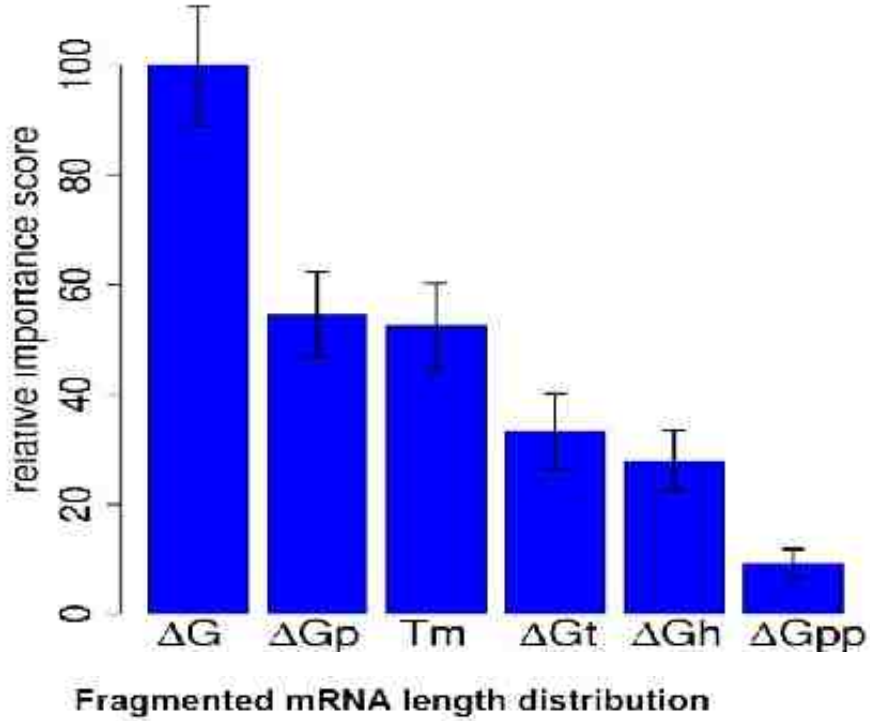
(Mückstein / Kreil, BMC Bioinf, 2010)

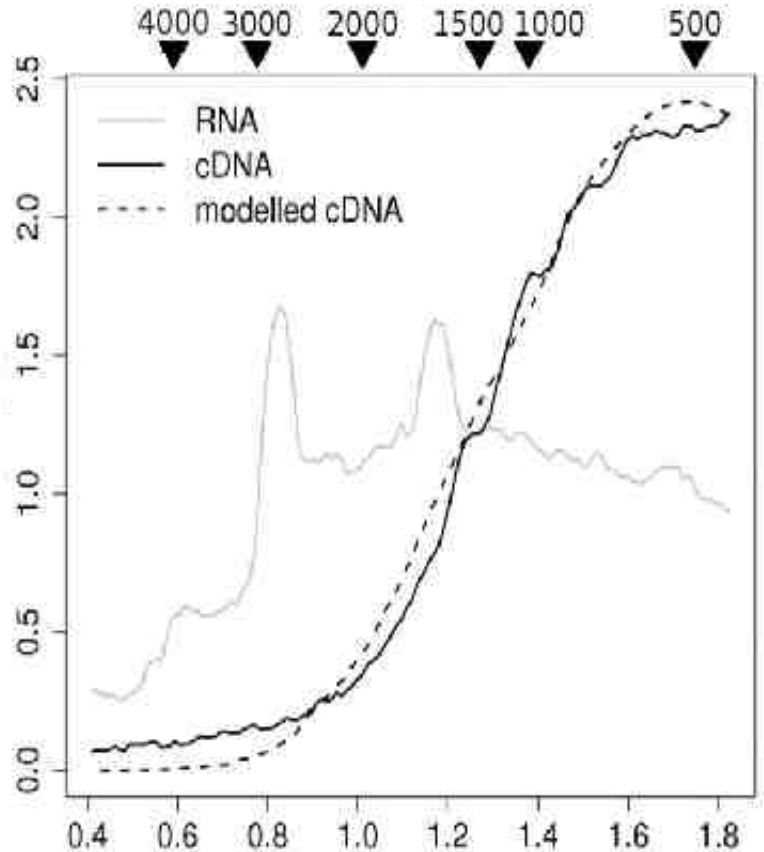
Labelling RT

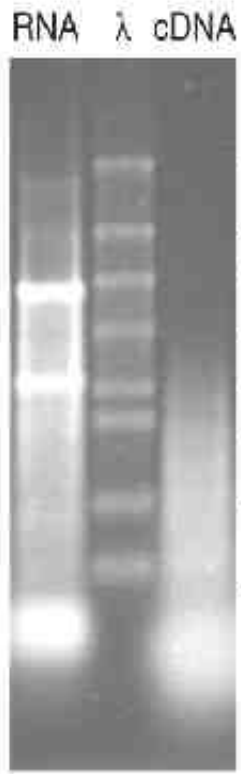
(Leparc / Kreil, NAR, 2009)

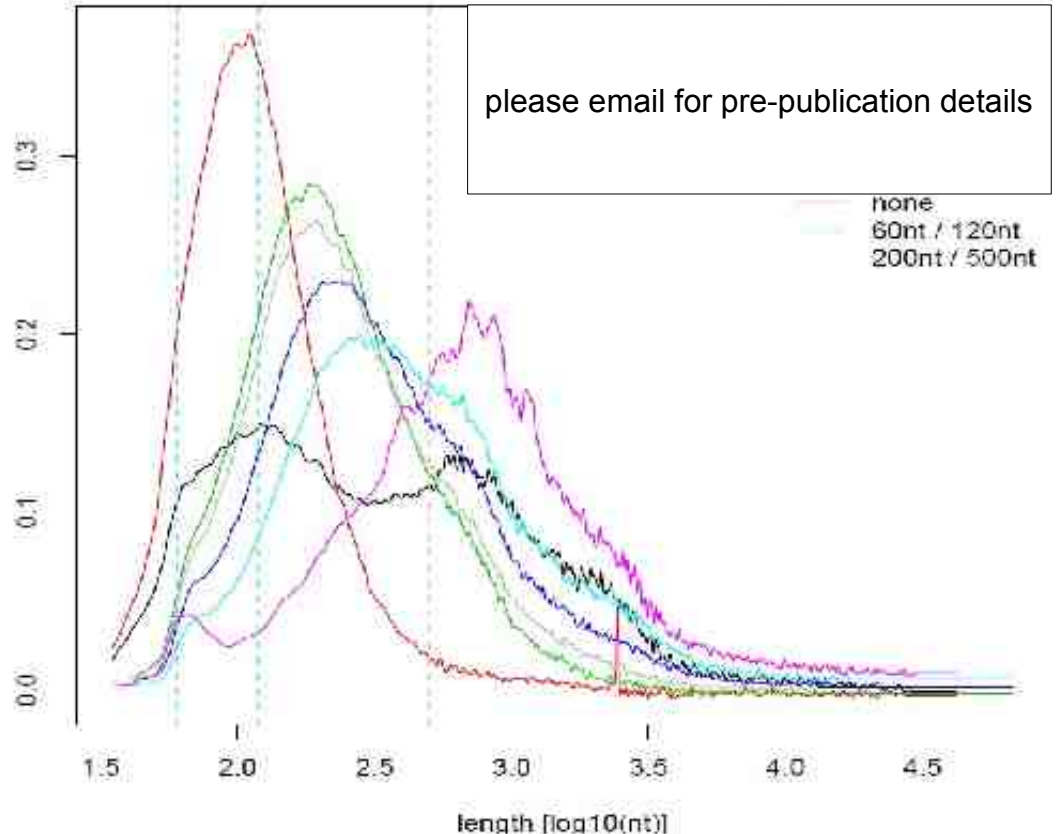
Fragmentation → x-target hybs?

(unpublished)



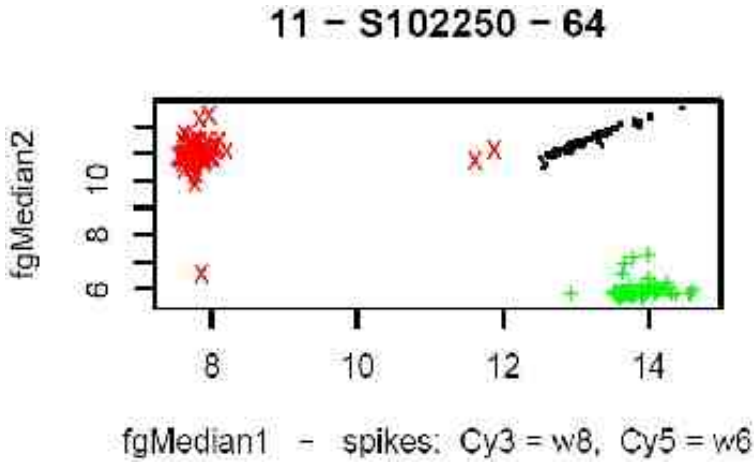






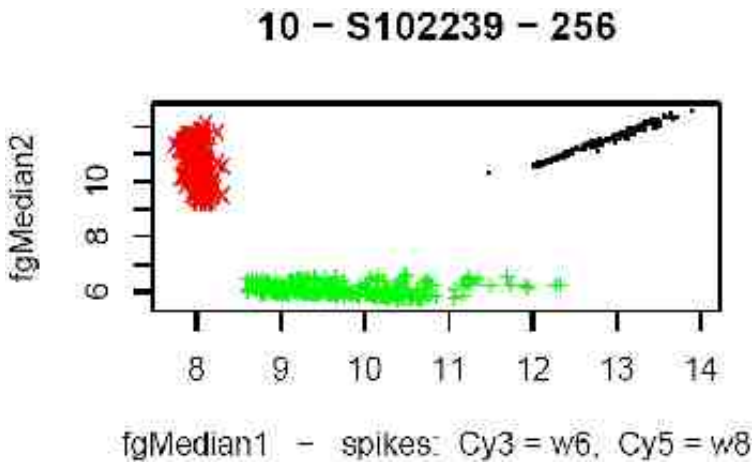
Non-specific binding:

(Kreil et al., unpublished)

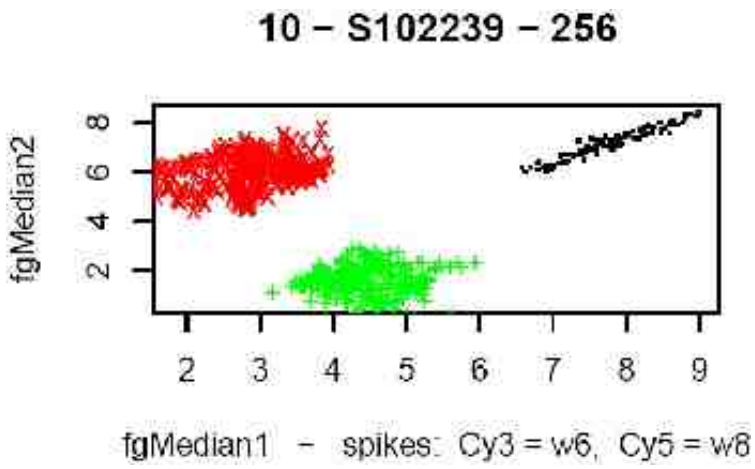


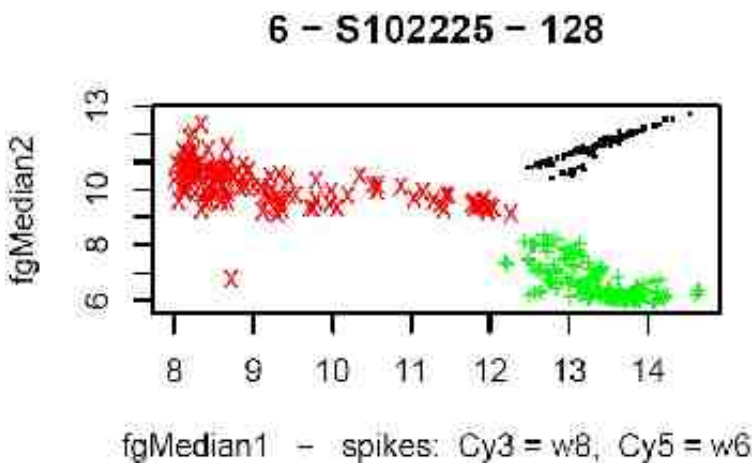
... the good ...

... with Lowess:

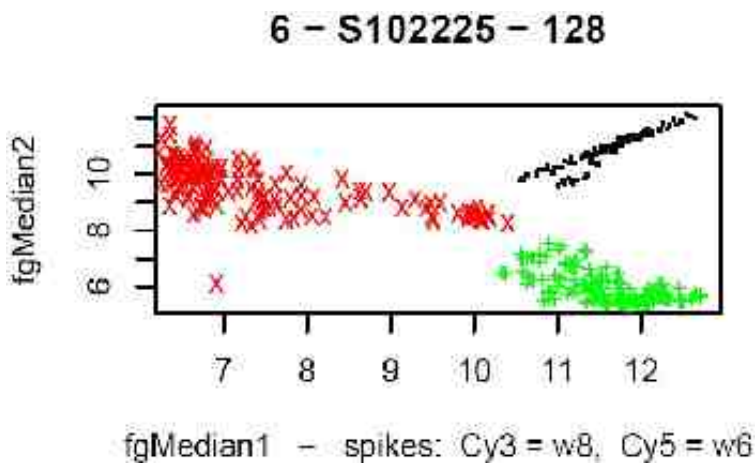


... the bad ...



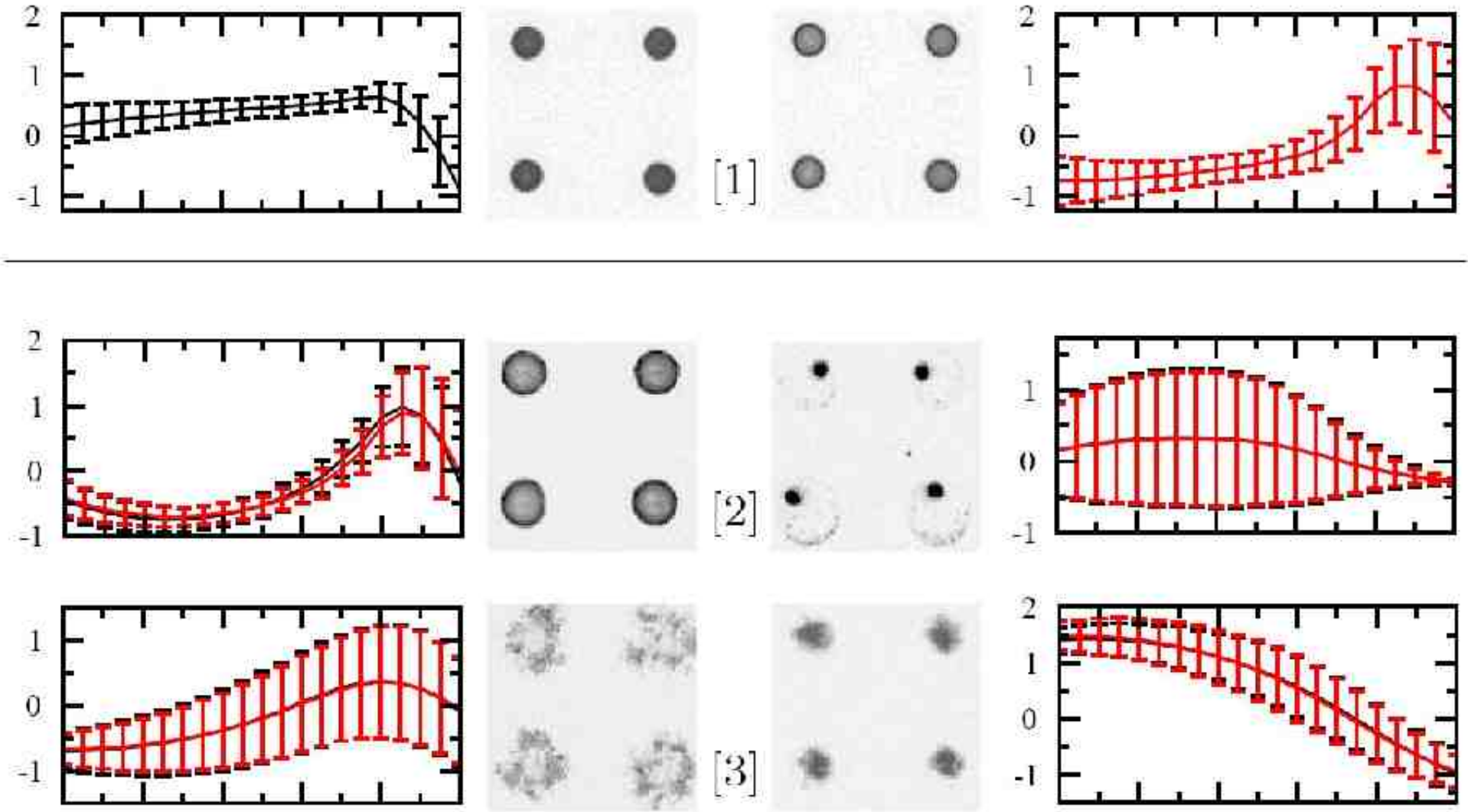


... and the ugly!



Impact of buffer / slide chemistries

Dye Separation [1] and other nasties...



(Kreil et al., unpublished)

15–16 July, www.camda.info, ISMB 17 July...

CAMDA

Massive Critical Assessment of ~~Microarray~~ Data Analysis

CAMDA 2011

MAIN MENU

- CAMDA 2011
- What is CAMDA
- Important Dates
- Registration
- Call for Papers
- Call for Posters
- Contest Datasets
- Agenda
- Scientific committee
- Organizers
- Travel and accommodation
- See and do
- Contact us
- Sponsors
- Previous CAMDA Editions
- CAMDA publications

Agenda

Table of Contents

The scientific program includes keynotes by leading researchers in the field and selected presentations of contest dataset analyses.

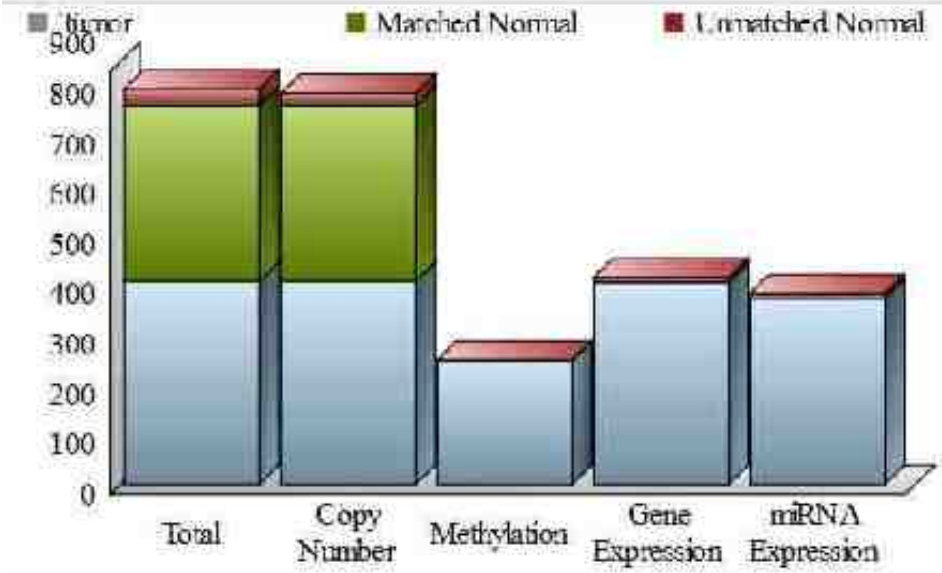
Current plans are to start the meeting Friday afternoon, 15 July, as well as encourage informal discussions during a reception / dinner.

The program will continue all day, Saturday 16 July, with the conference closing early evening.

Sign up to our low-volume CAMDA 2011 announcements mailing list!

Keynotes

We are delighted to welcome you to the CAMDA keynote talks by established experts in the field.



Terry Speed

Professor Terry Speed heads the Bioinformatics division at the Walter and Eliza Hall Institute of Medical Research (WEHI), in Melbourne, Australia.

Terry has made key contributions to microarray analysis, and has early identified the need for thorough low level analysis of the data. His research interests include a large variety of applications, such as his recent contributions to the study of confounding factors in genome wide DNA methylation measurements or his research work on base calling for resequencing chips.

John Storey

Professor John D. Storey heads the Genomics research group at the Lewis-Sigler Institute for Integrative Genomics of Princeton University.

John's group is interested in the analysis of high-dimensional data sets, such as large scale genotyping or gene expression profiles. He has developed efficient approaches to the multiple-testing problem central to the field. His recent research addresses challenges of integrating multiple genome wide data sources and the identification of confounding structures.