



Yale University



CANCER RESEARCH UK



igmm

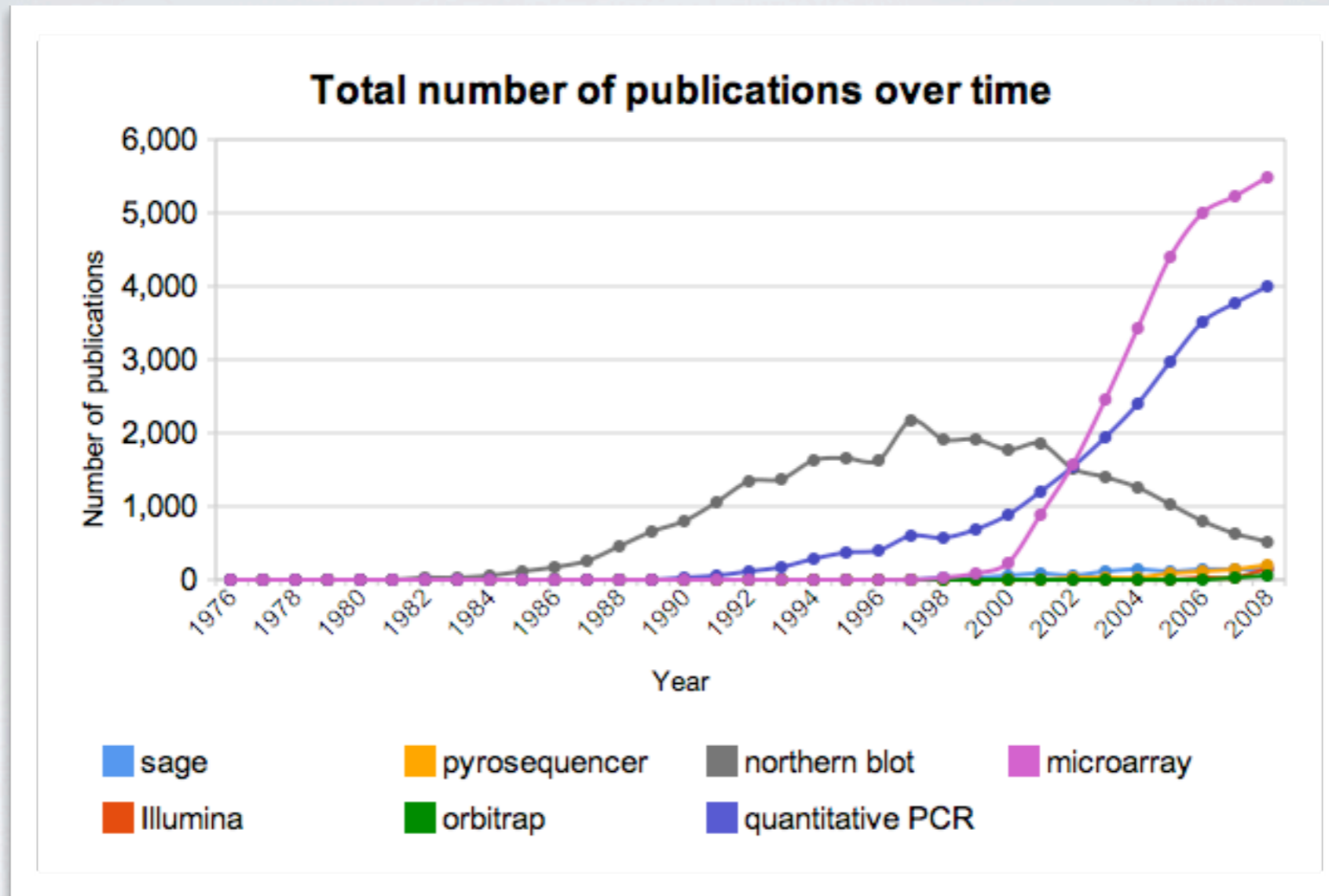
INSTITUTE OF GENETICS  
& MOLECULAR MEDICINE

# Past, present, and future of high-throughput genomics

Rob Kitchen

11 / 02 / 2011

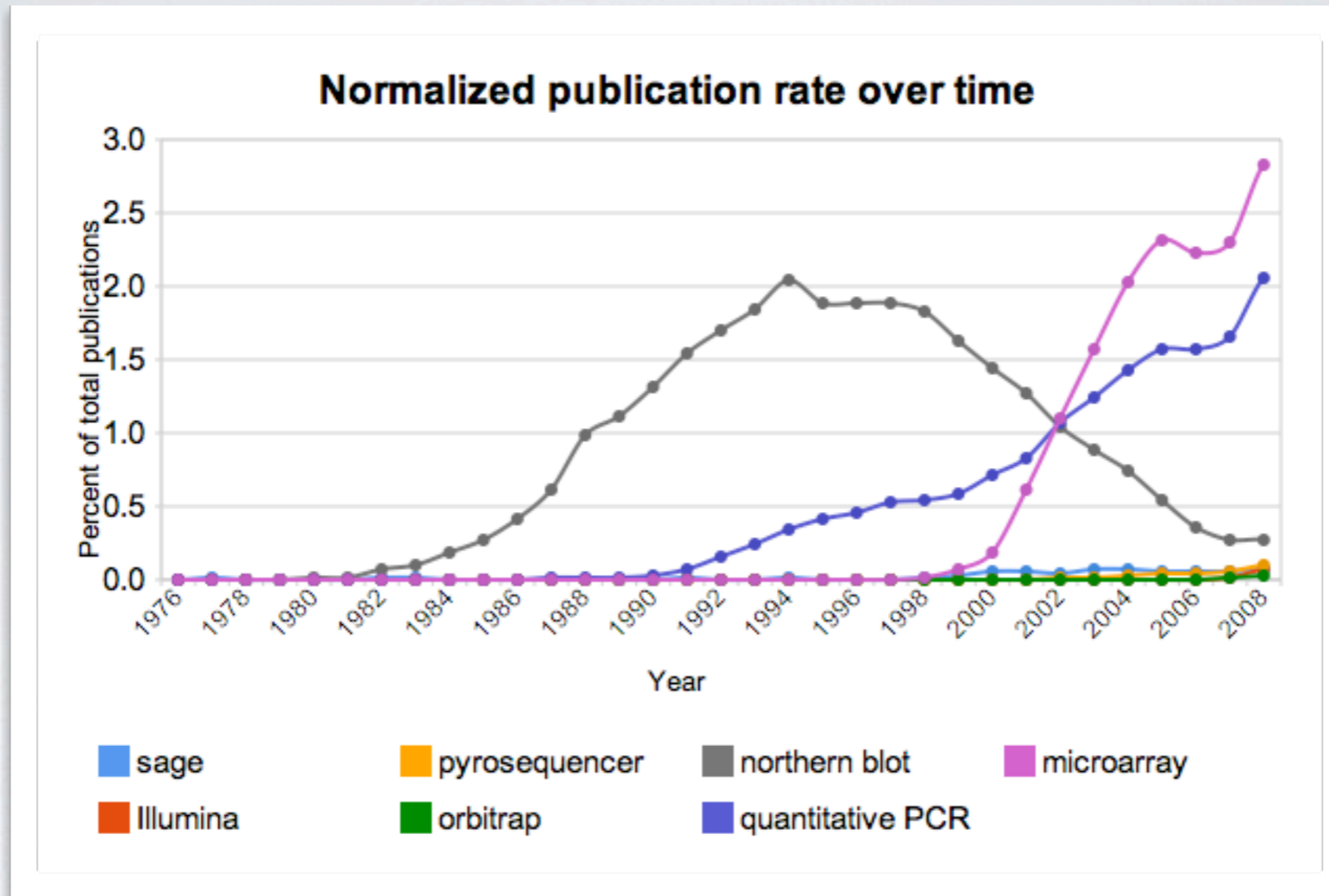
# outline



source: scitrends.com

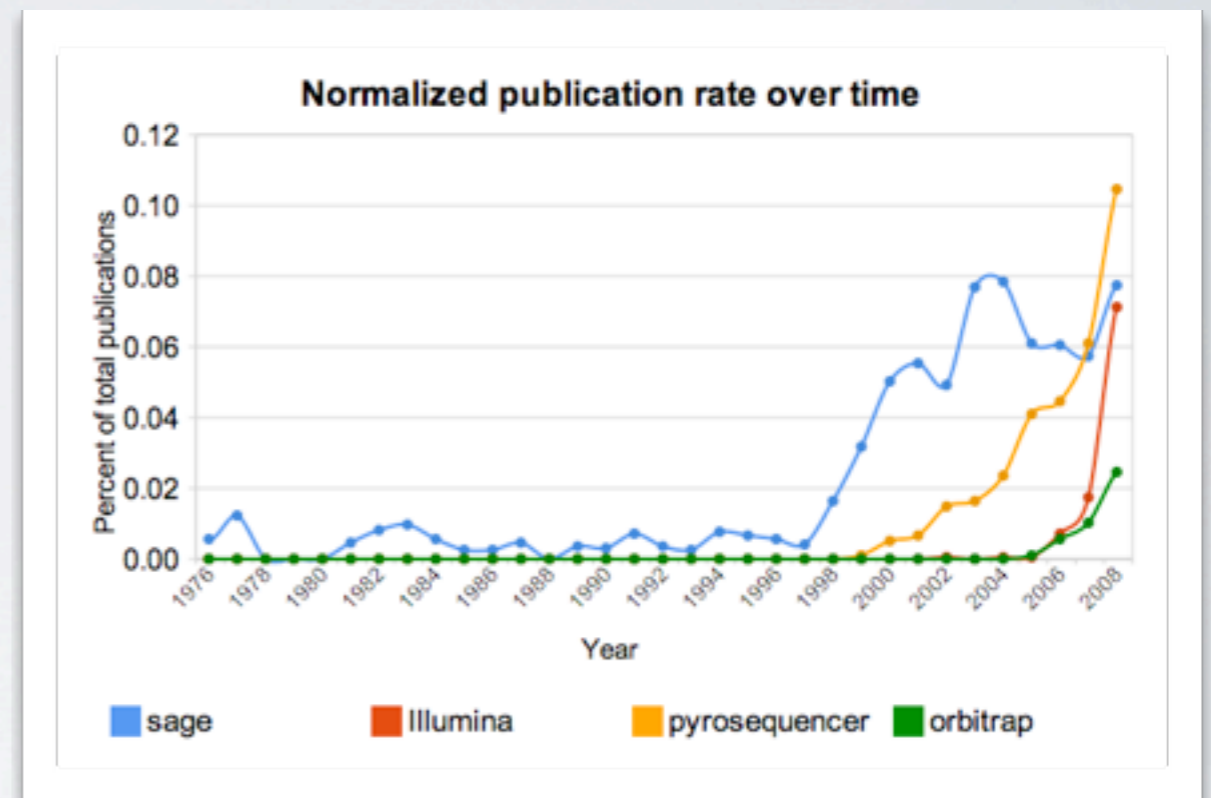
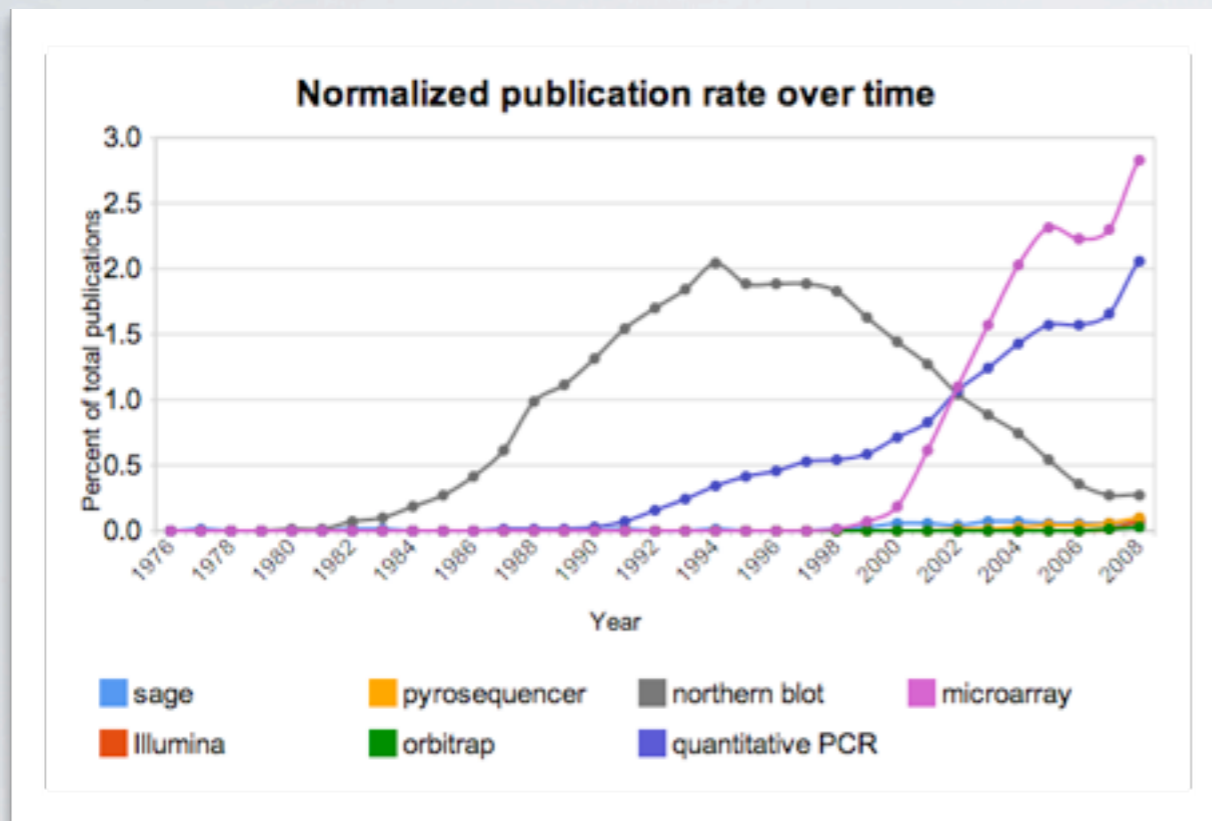


# outline



source: scitrends.com

# outline



past / present

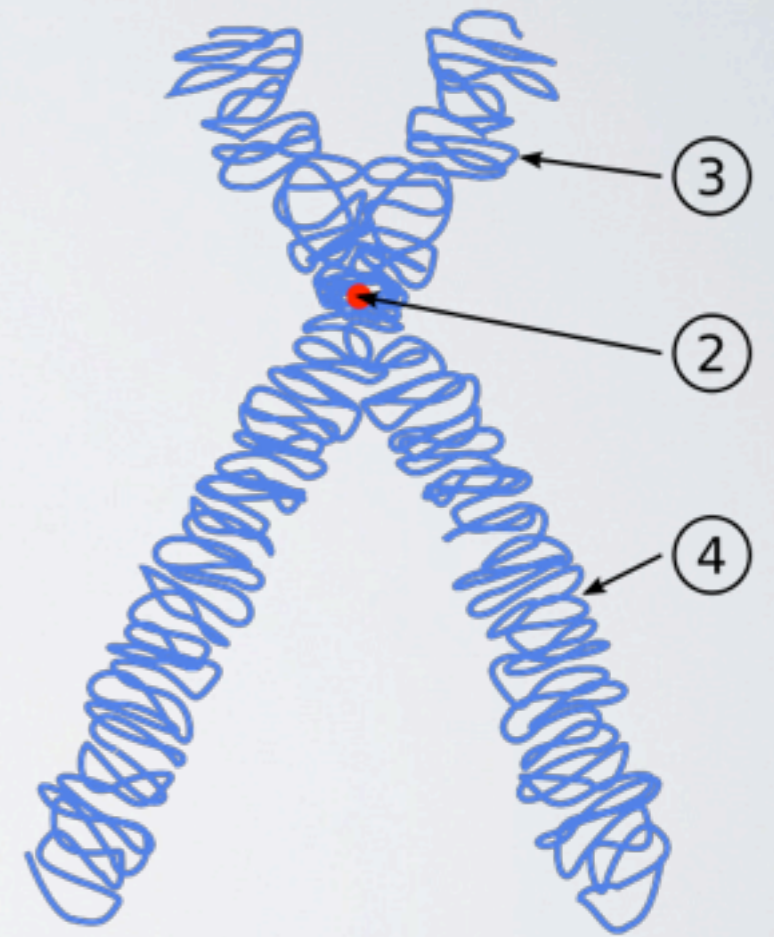
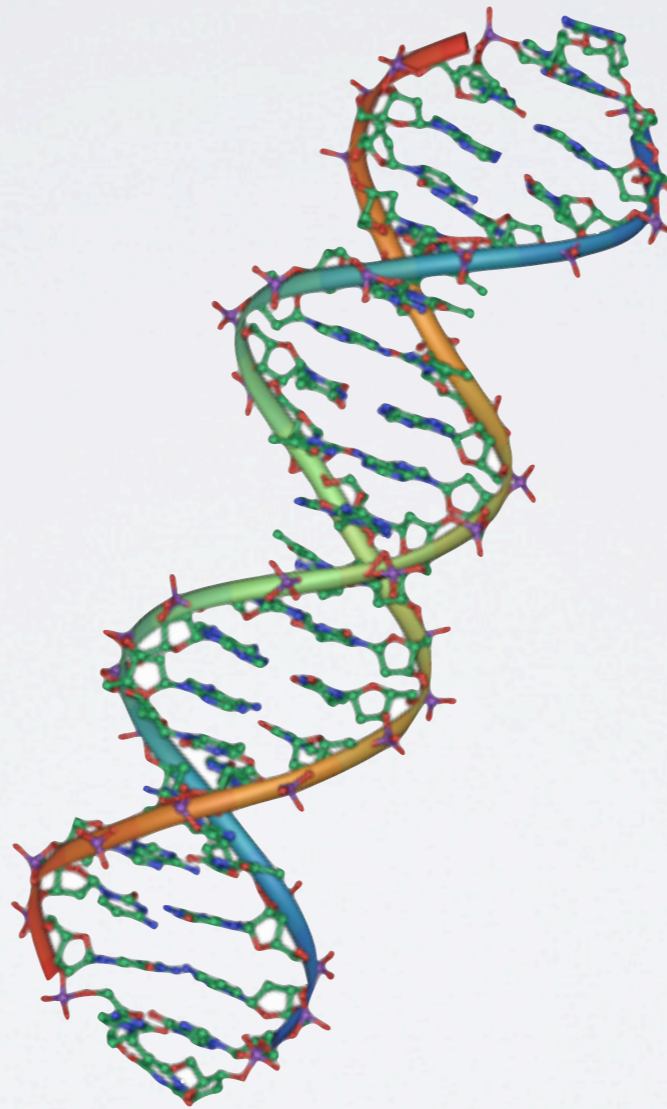
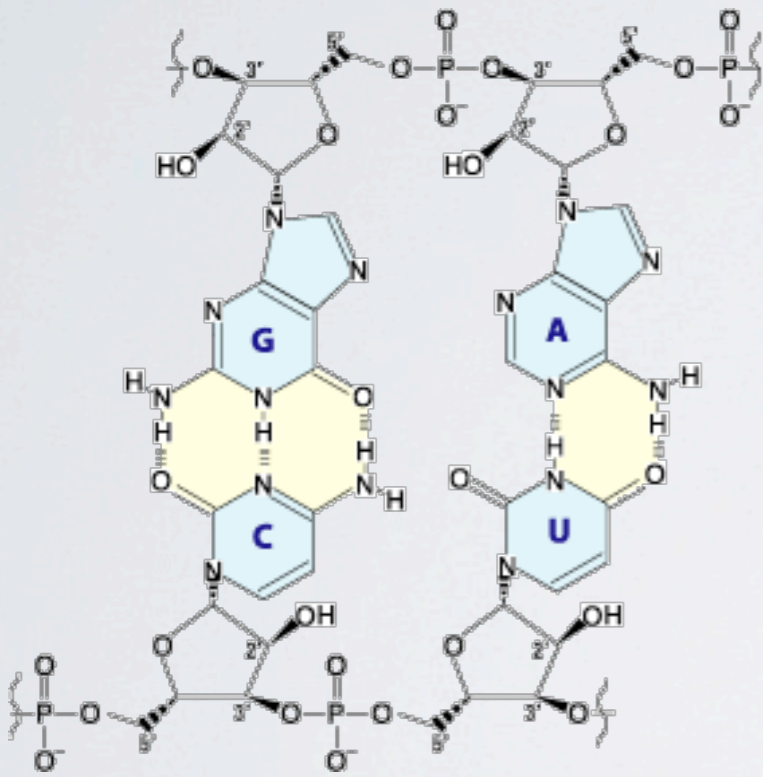
present / future

source: scitrends.com



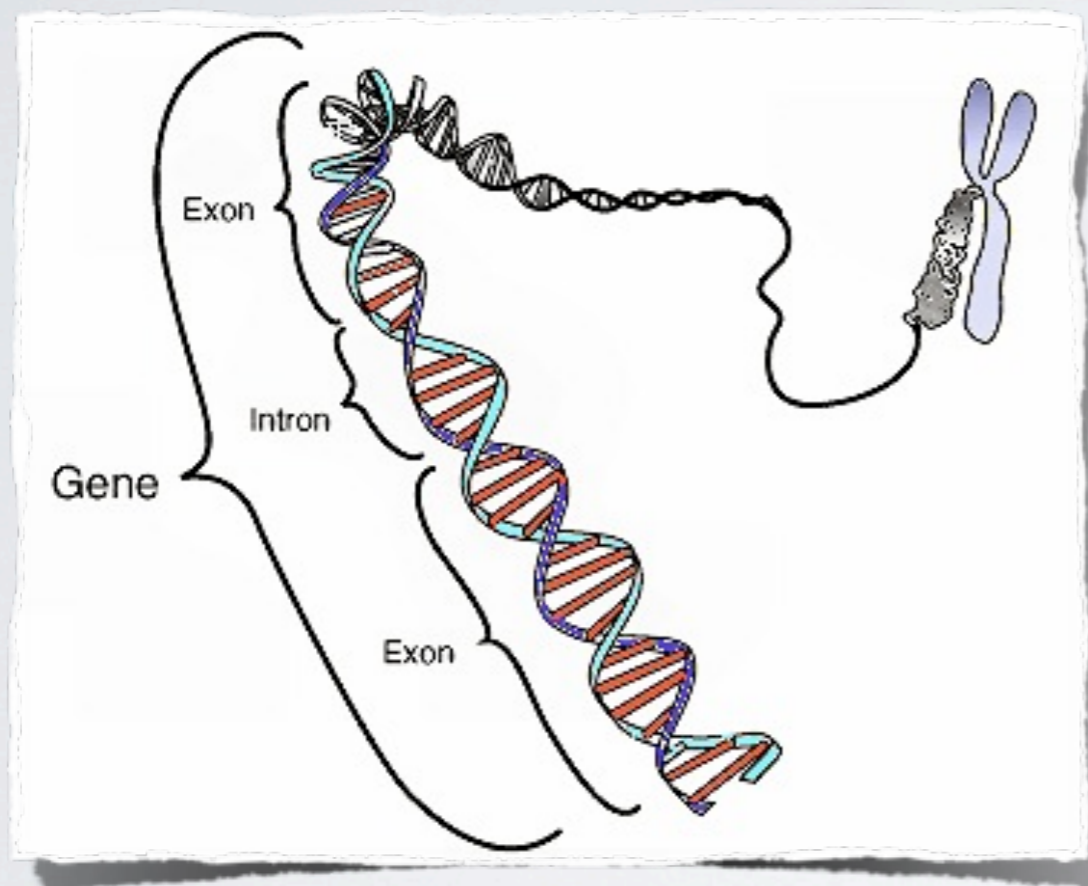
# brief background

## DNA



# brief background

## genes



DNA subdivided into  
**chromosomes**

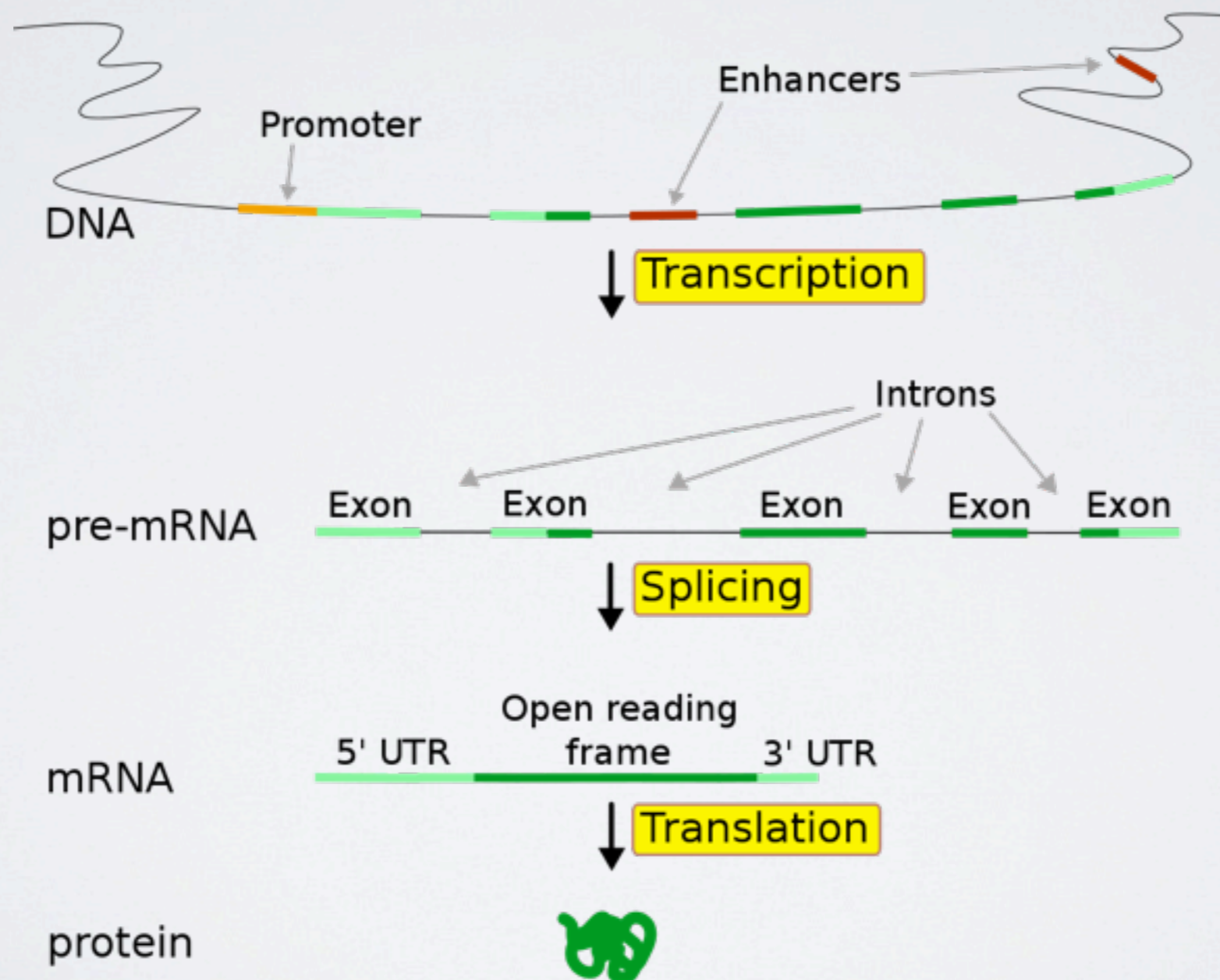
chromosomes further divided  
into regions called **genes**

~**20-25,000** genes in humans  
(exact number still unknown!)



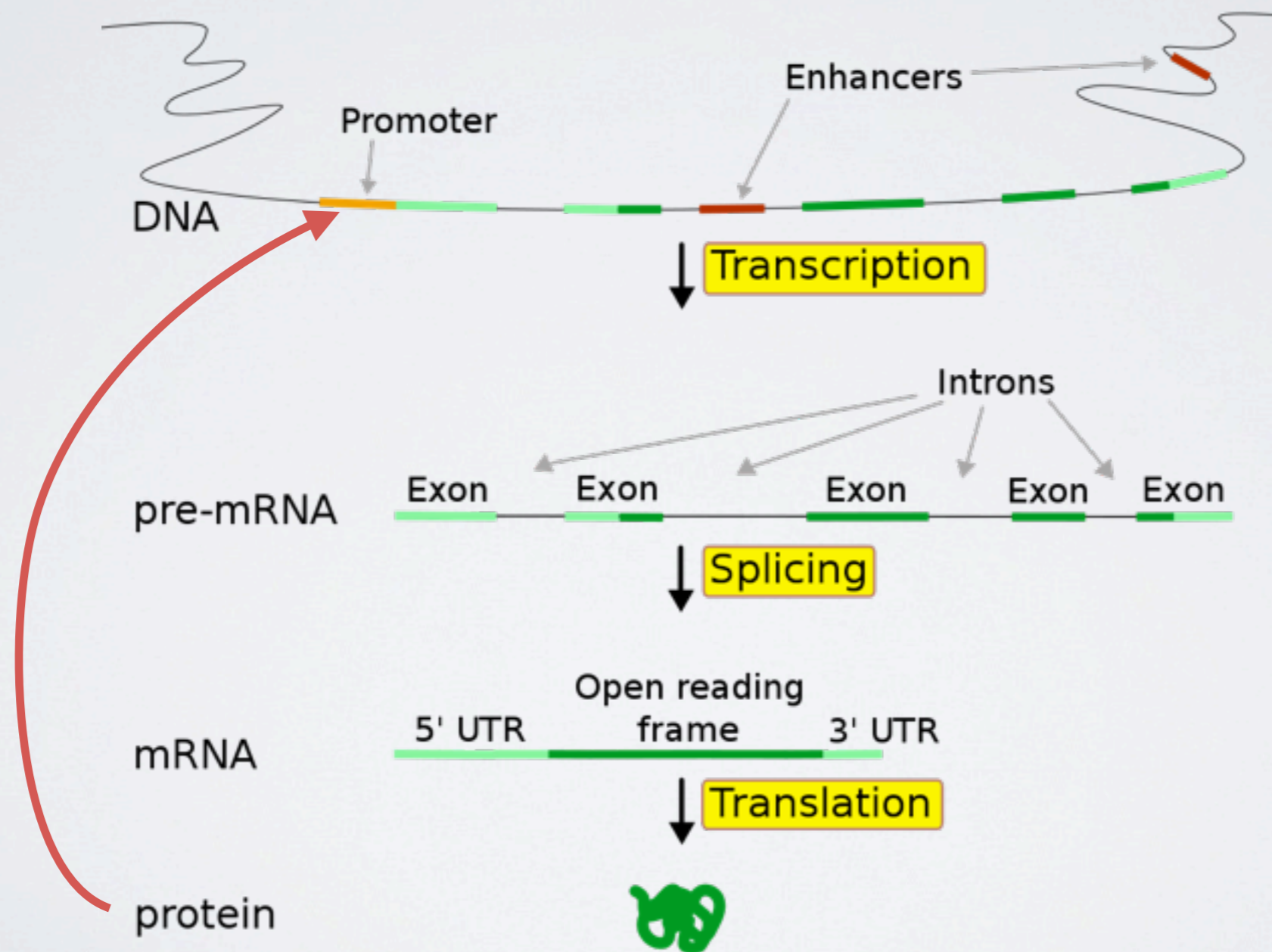
# brief background

## gene-expression (vastly simplified)



# brief background

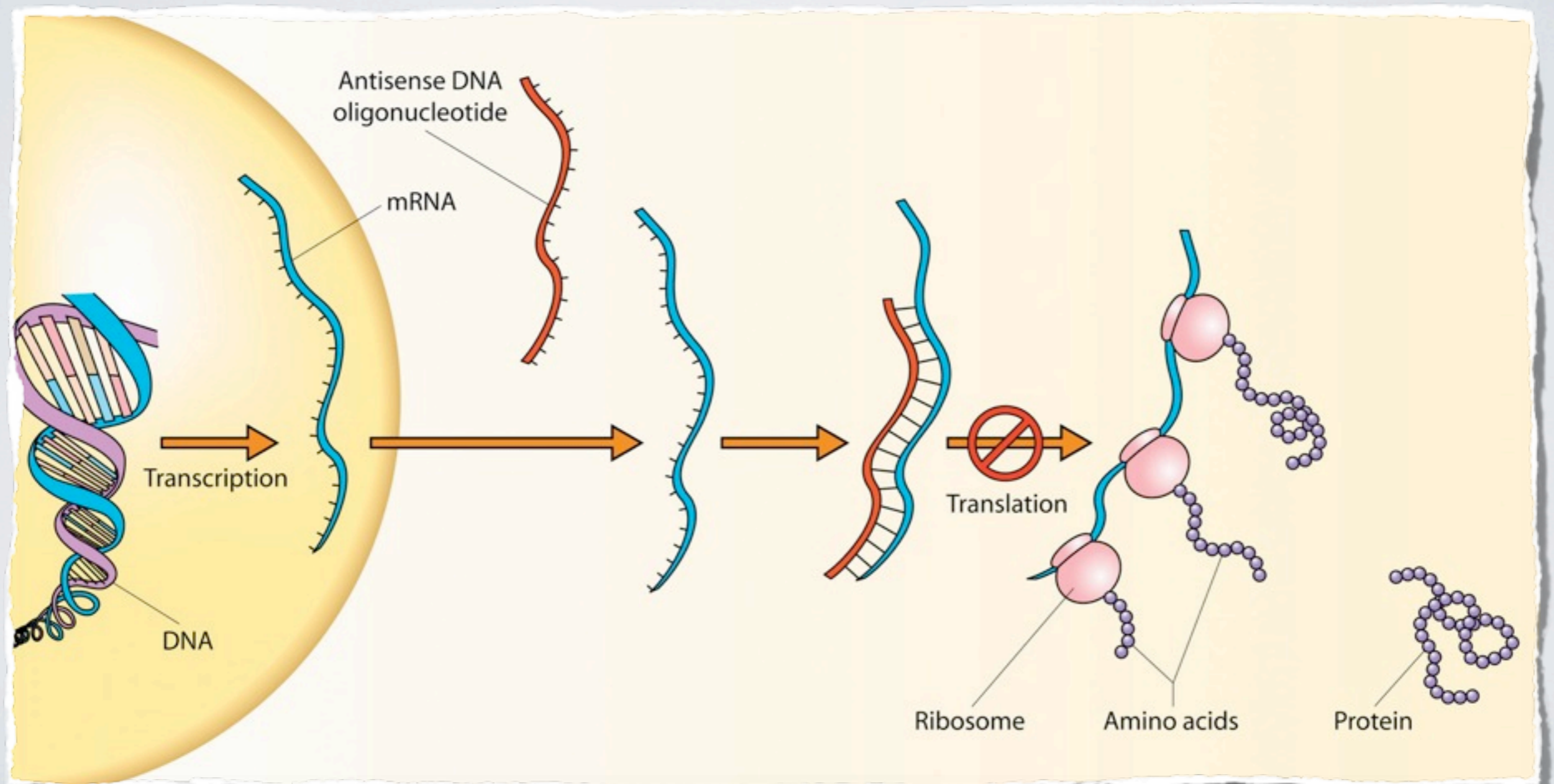
## gene-expression (vastly simplified)



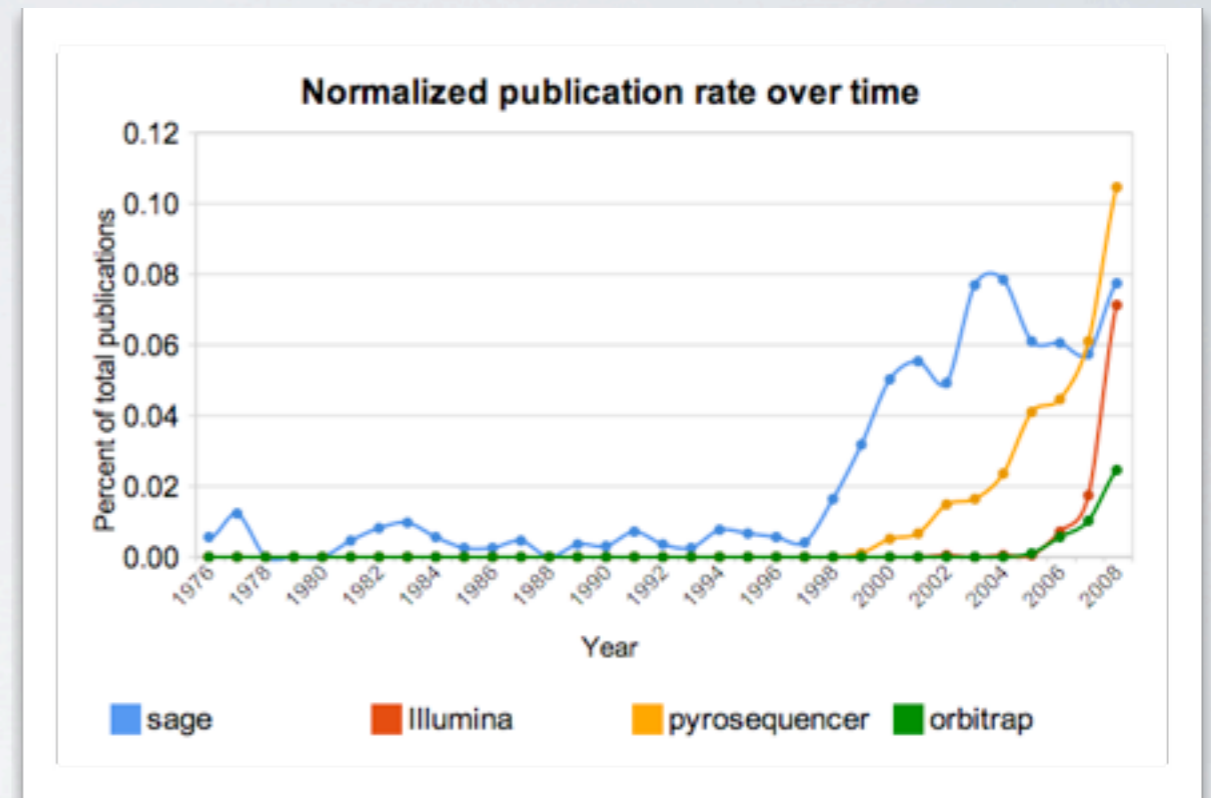
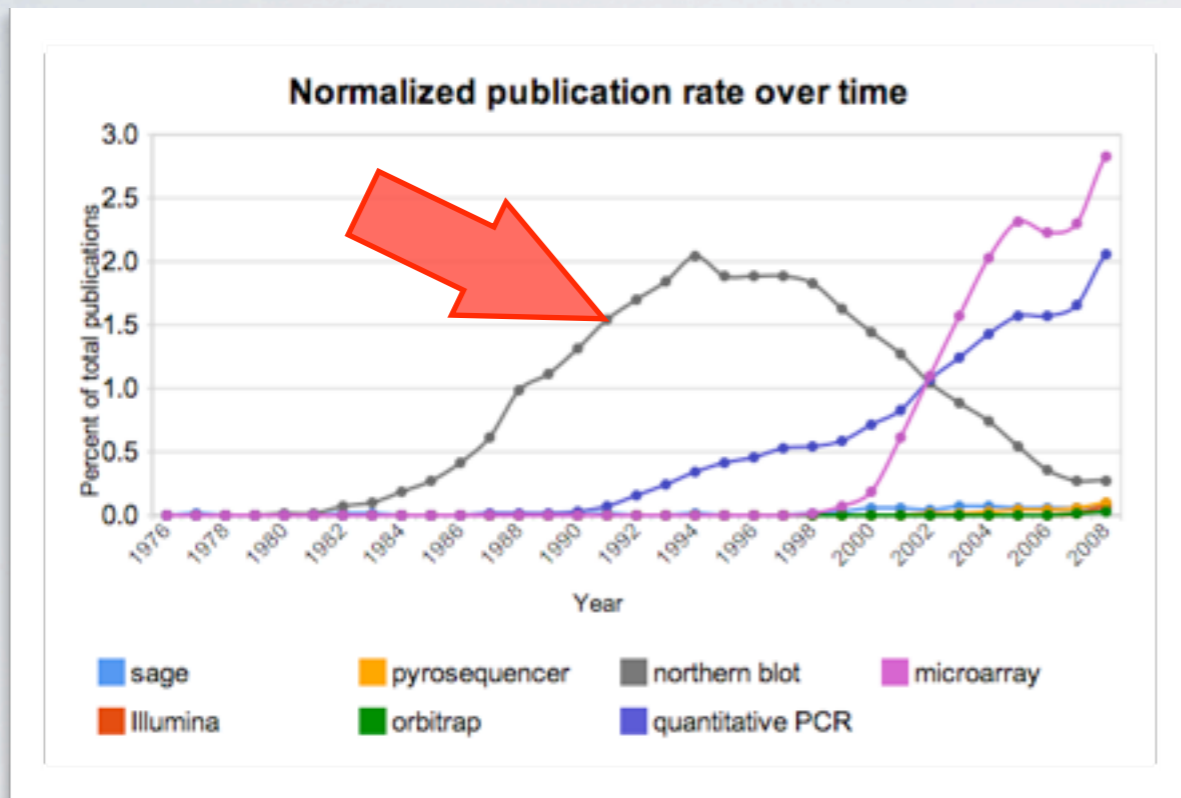


# brief background

## gene-expression (vastly simplified)



past

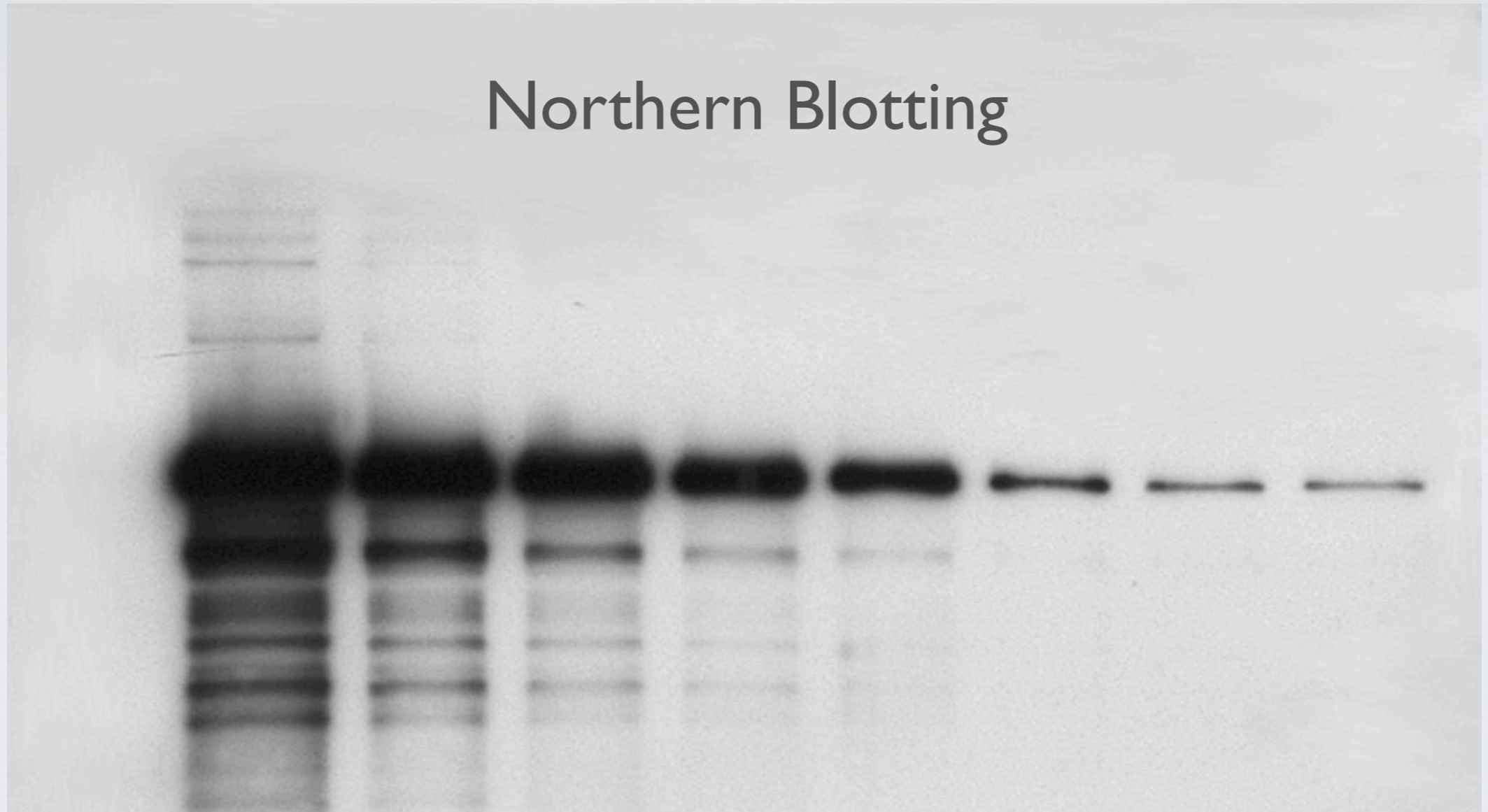





past

# Northern Blotting

E



dilutions

1

0.9

0.8

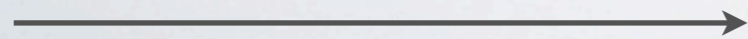
0.7

0.6

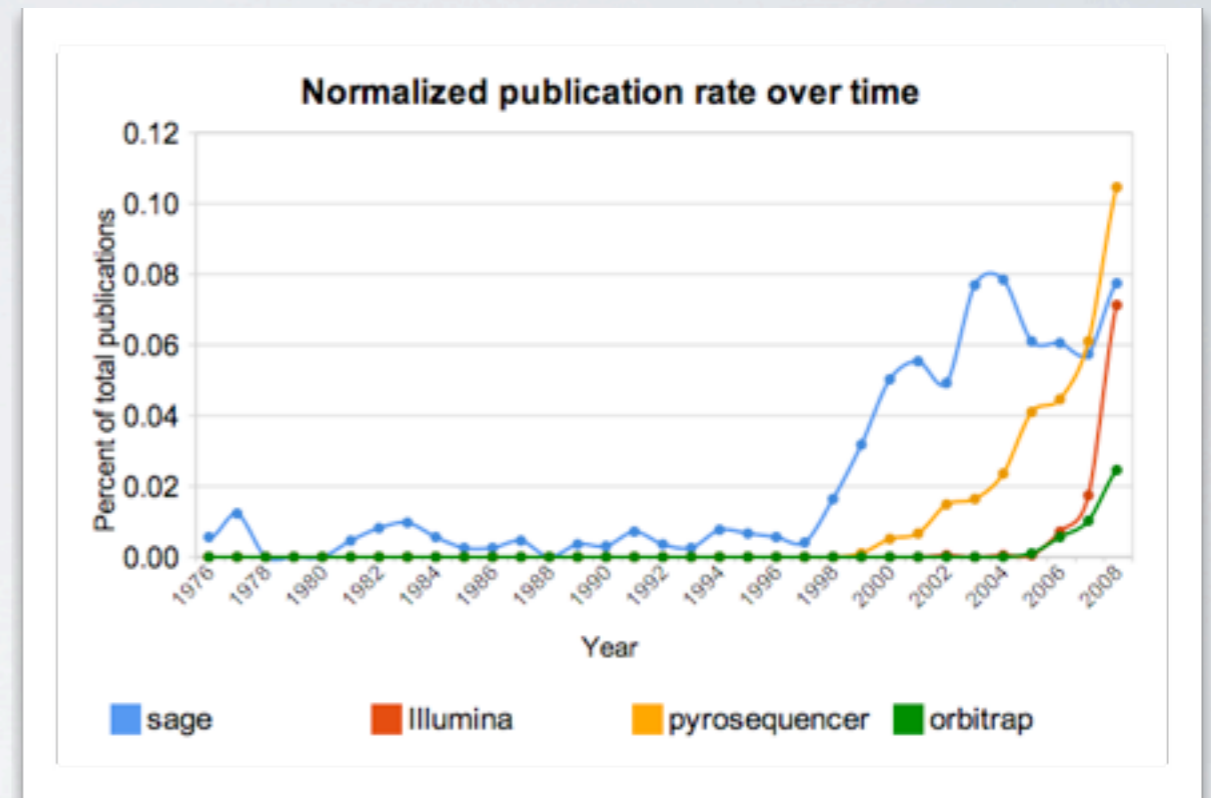
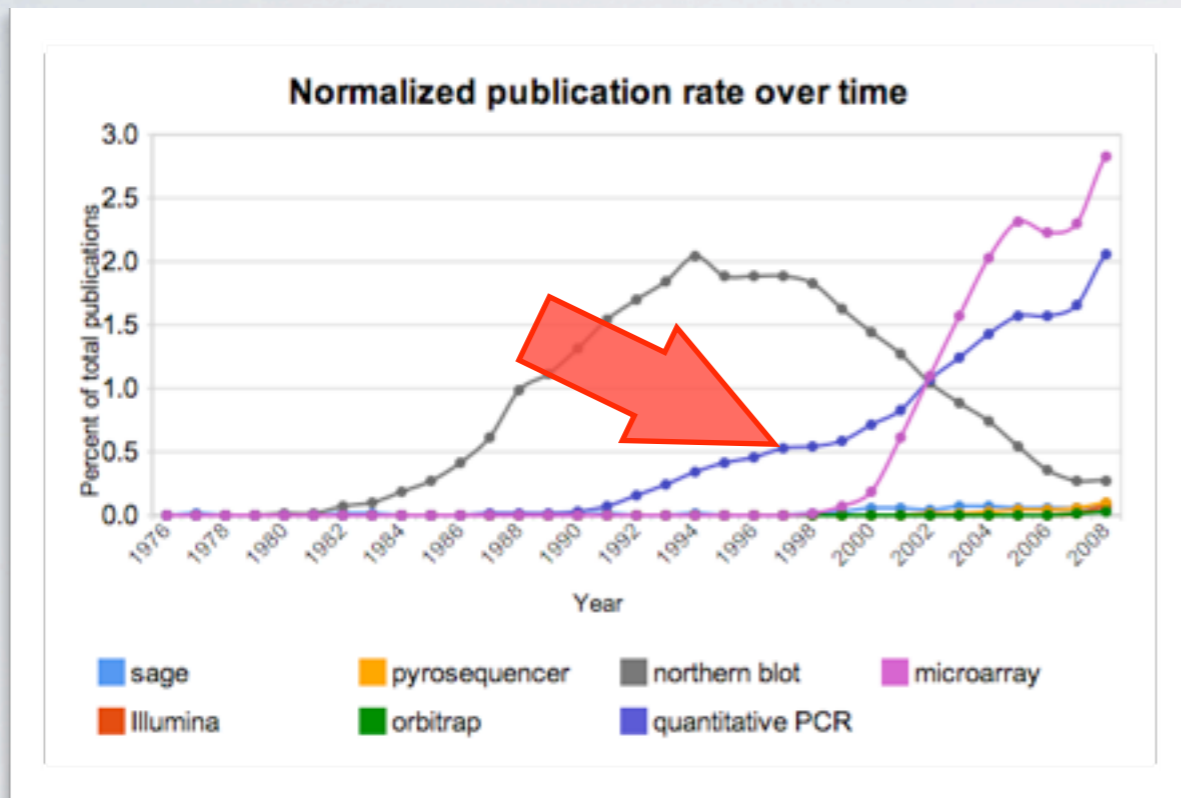
0.5

0.4

0.3



# present

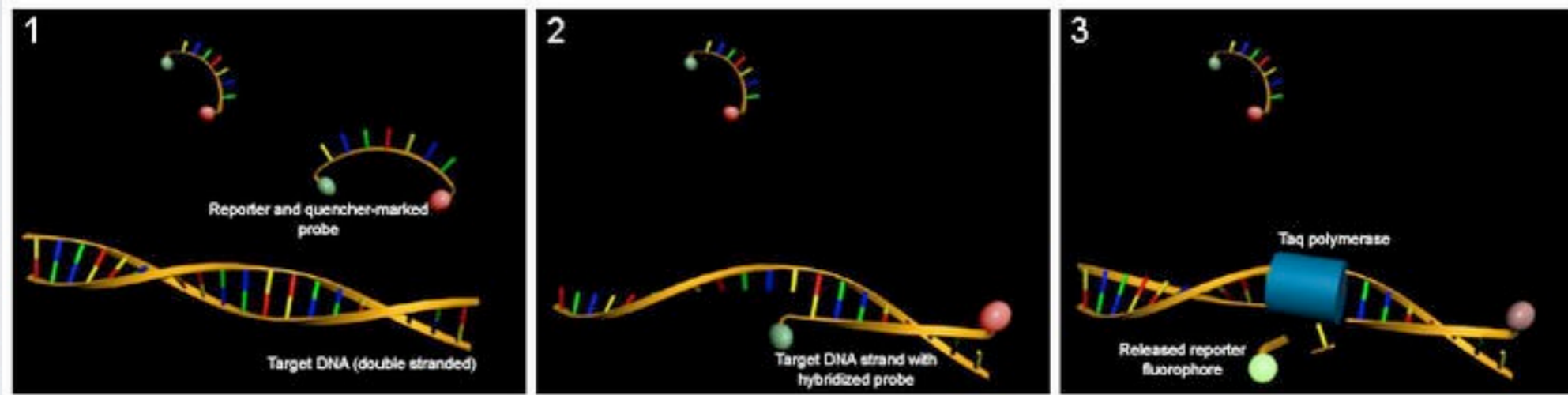




# quantitative PCR

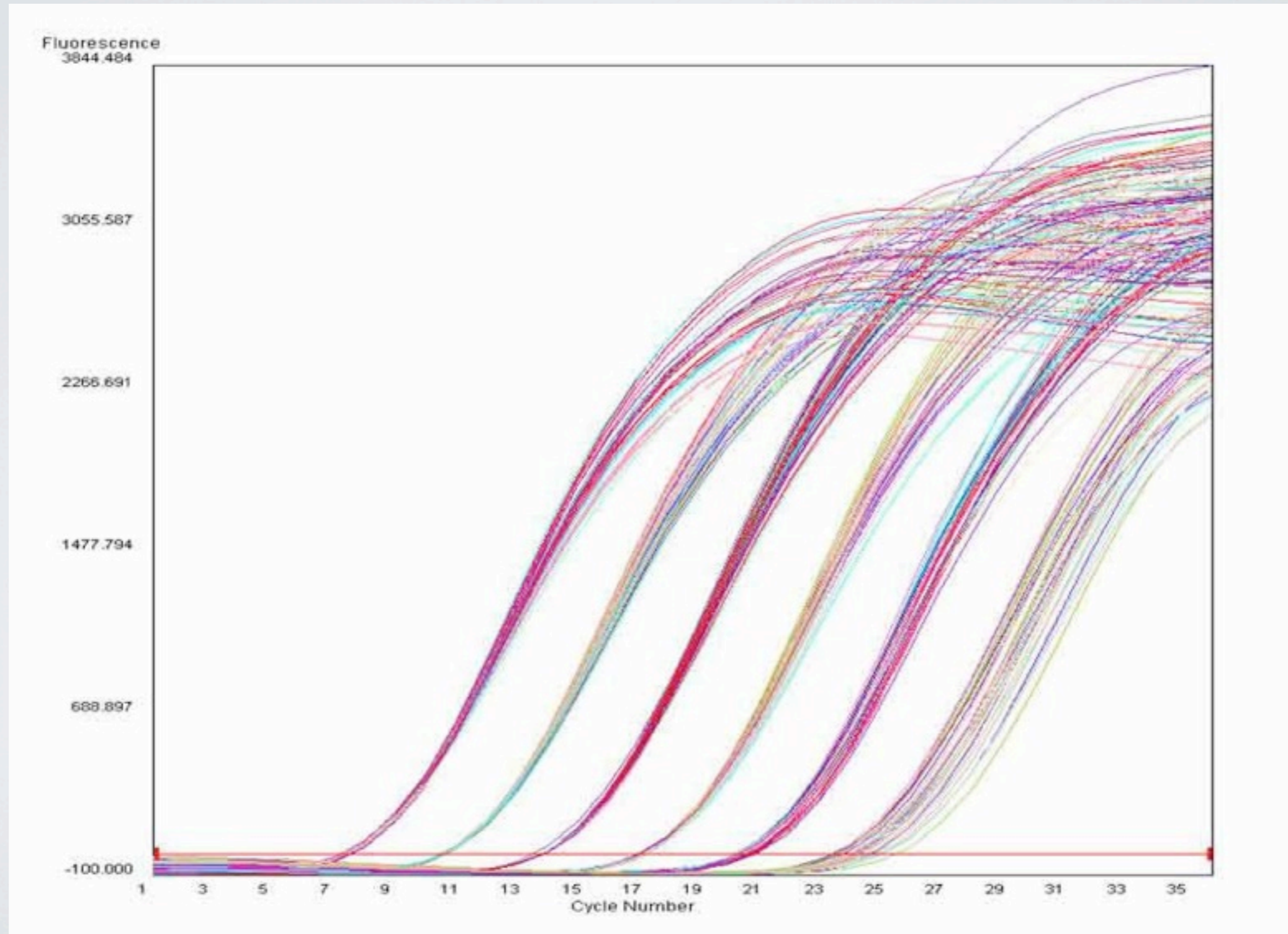


# quantitative PCR

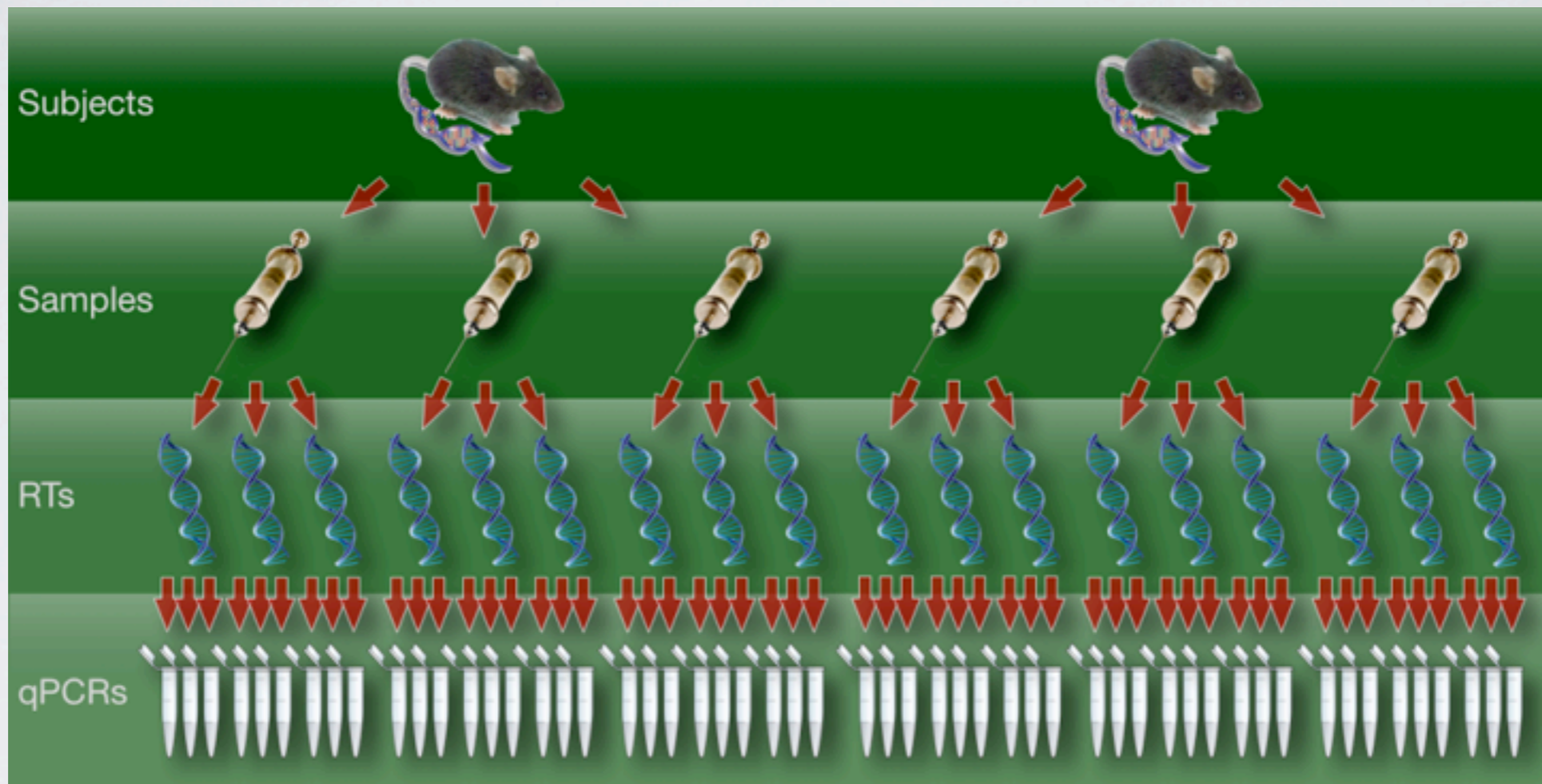




# quantitative PCR



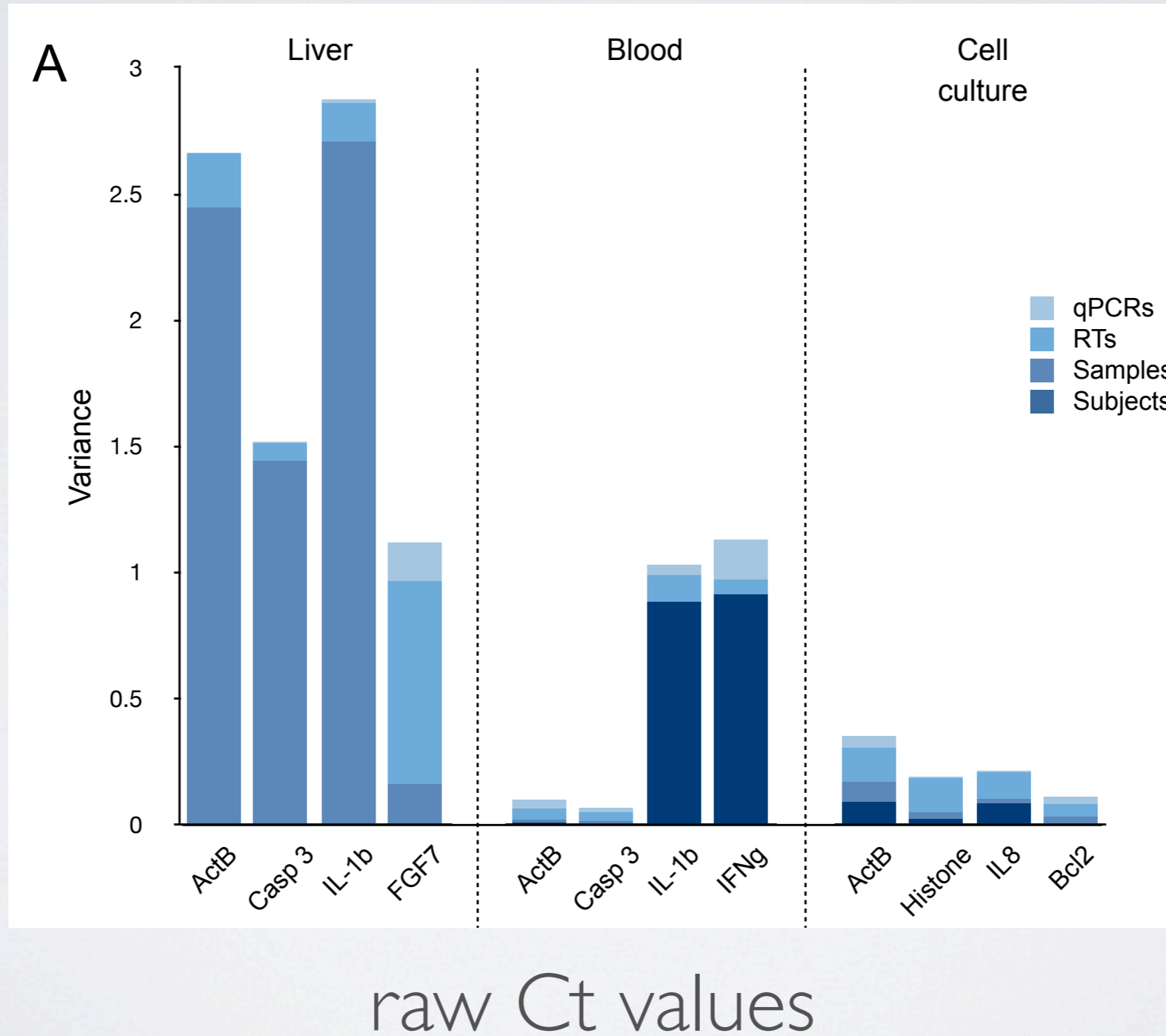
# quantitative PCR



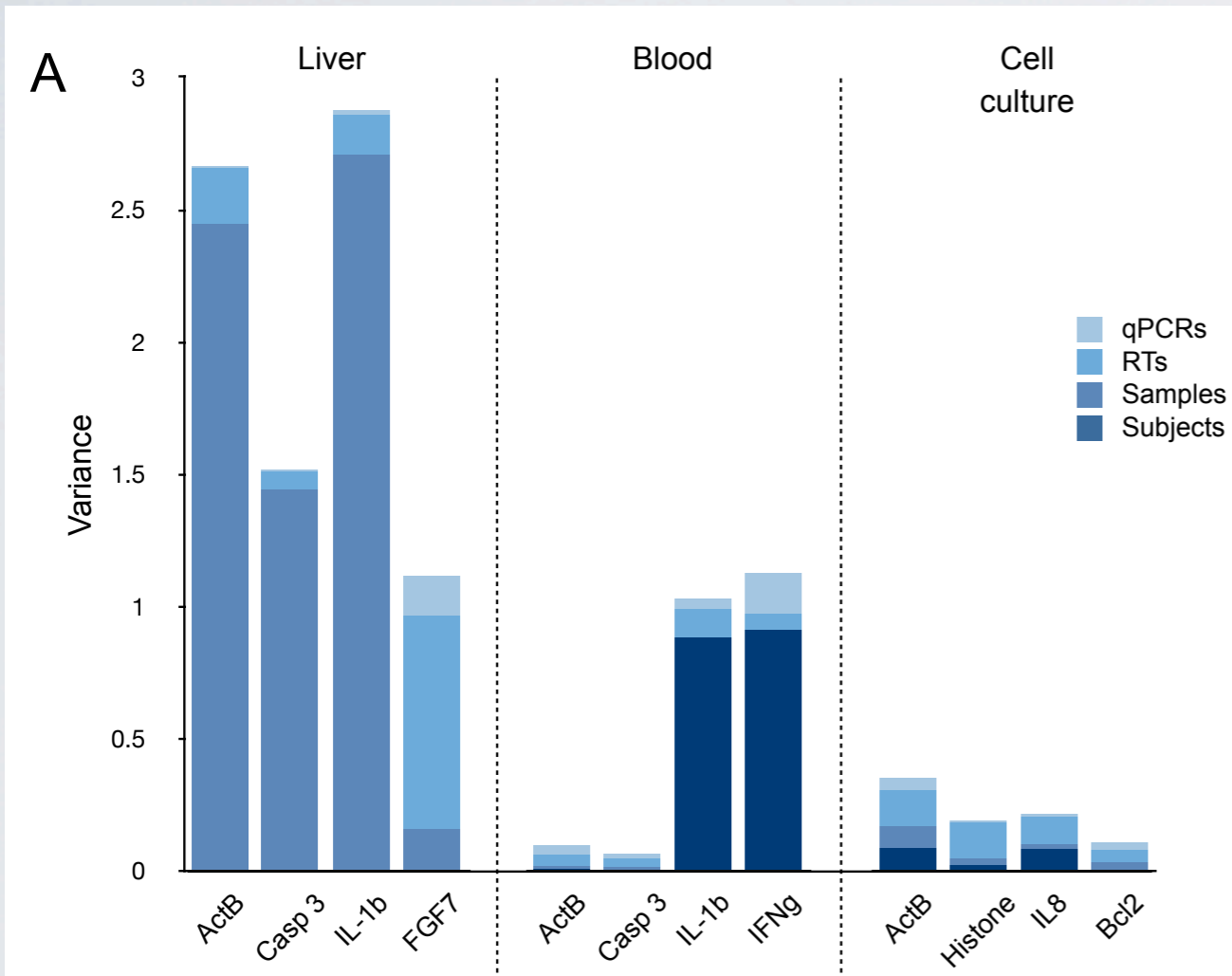
interested in exploring the **noise introduced** at each stage of sample-prep on reported results



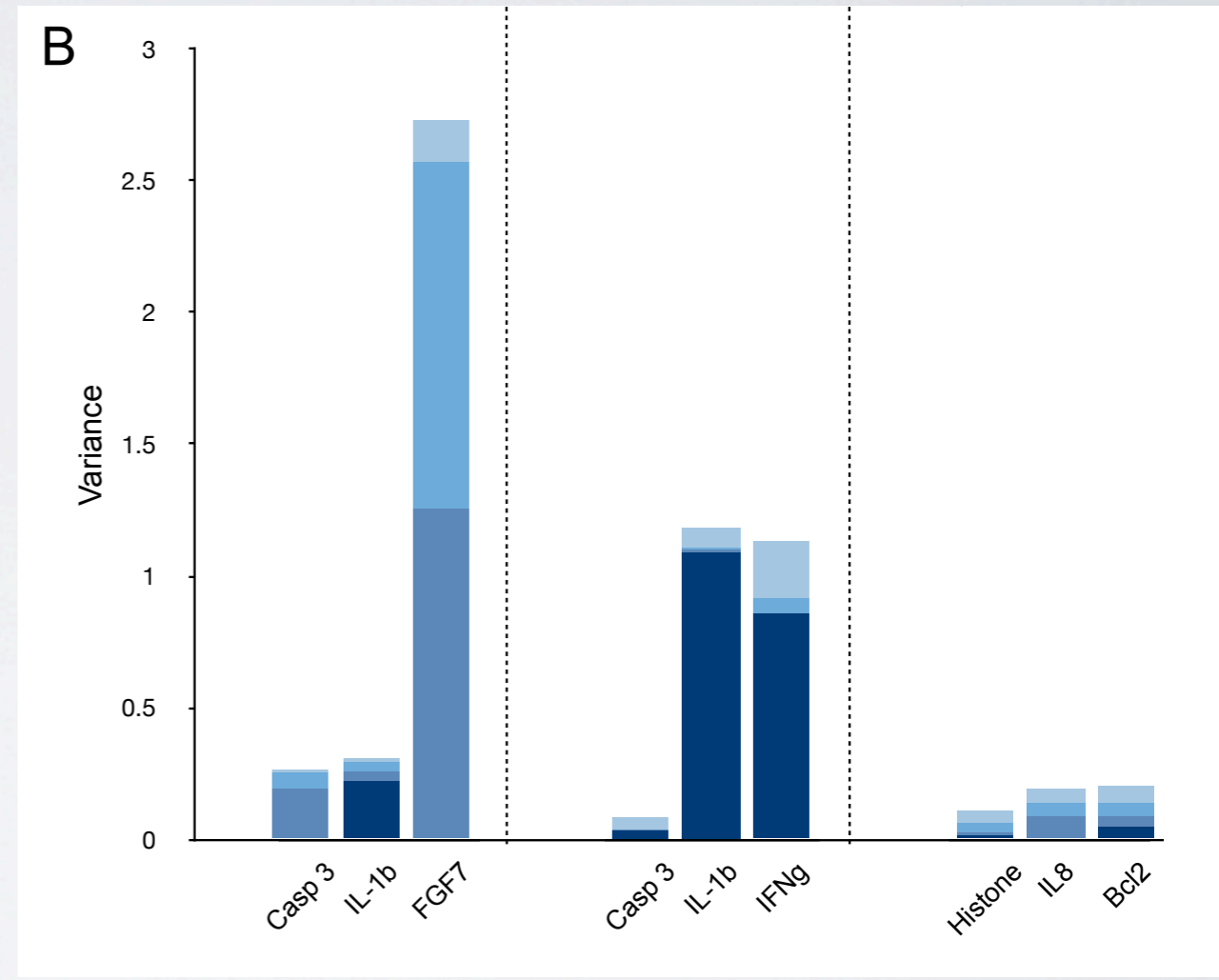
# quantitative PCR



# quantitative PCR



raw Ct values

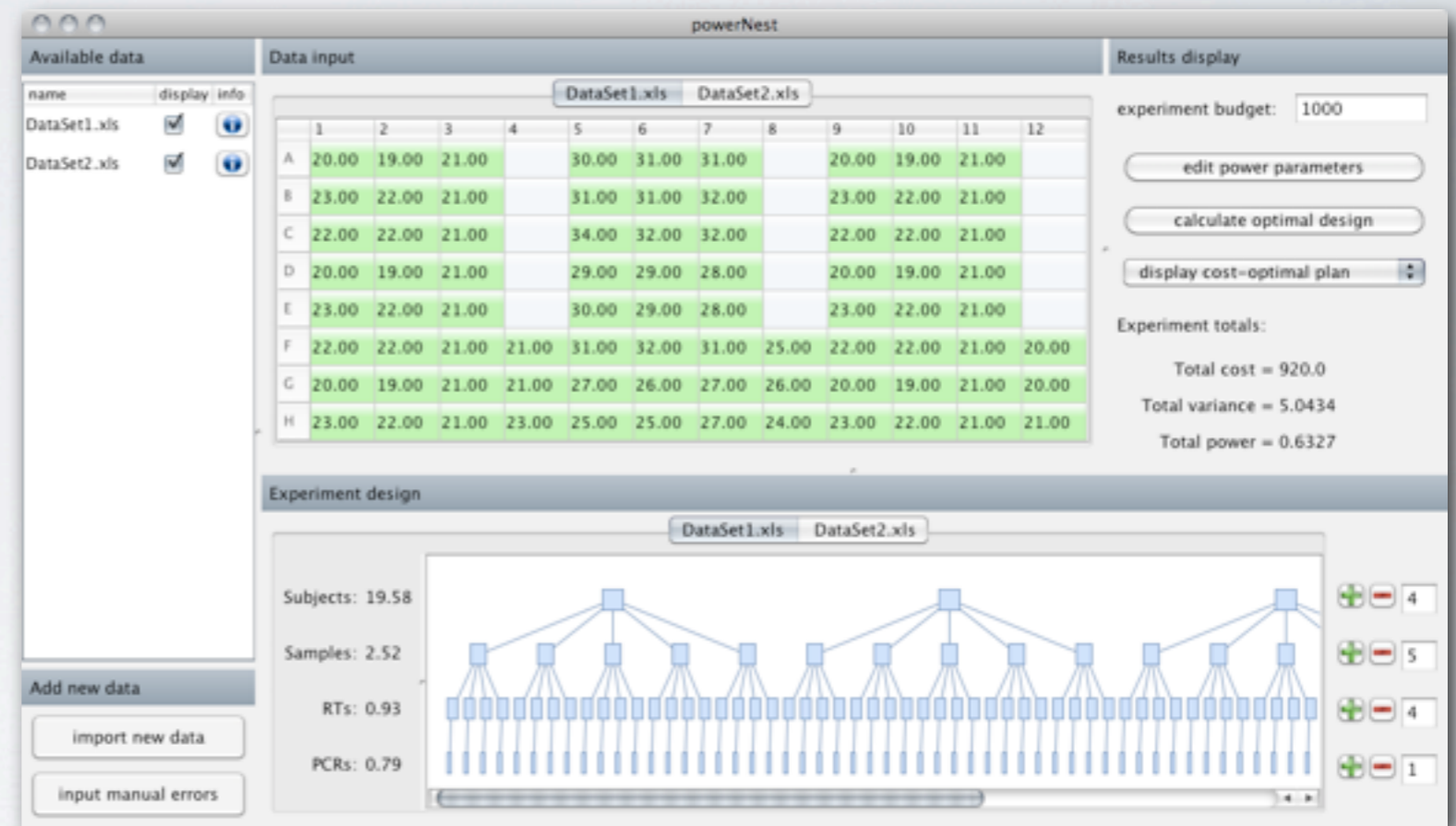


$\Delta$ Ct values



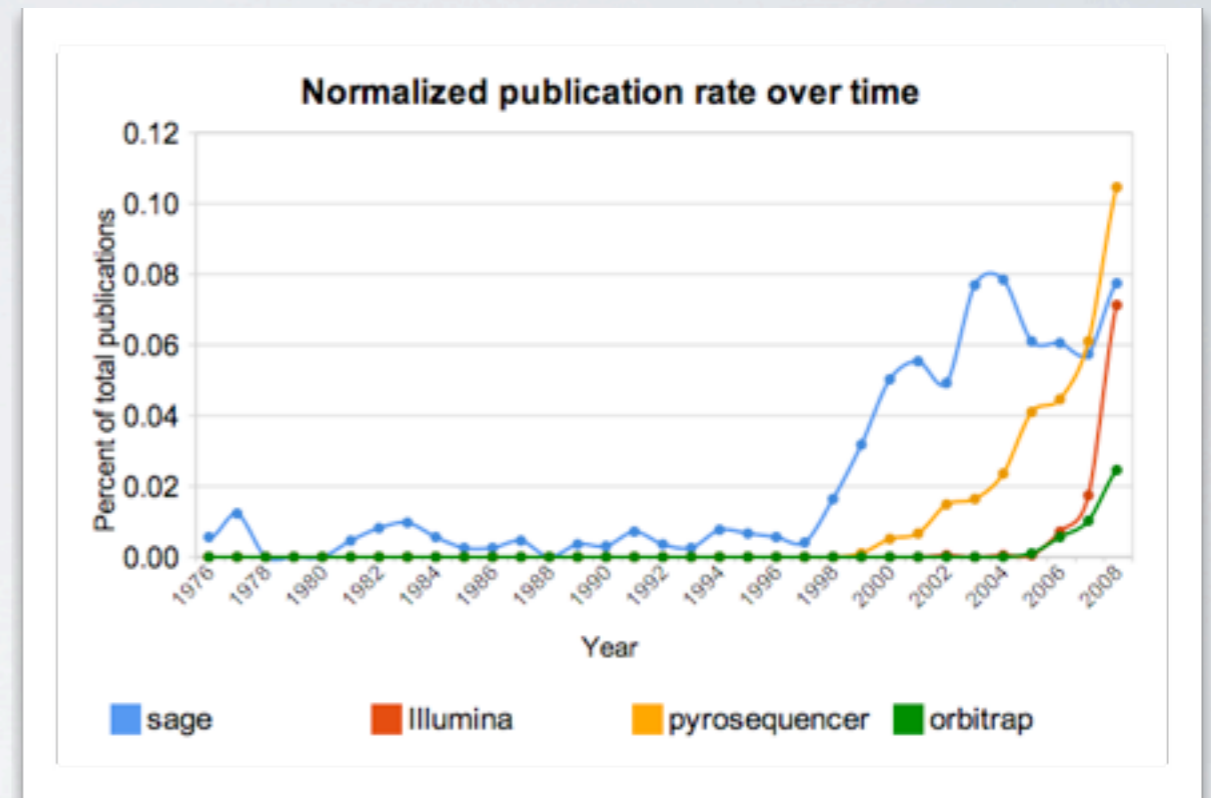
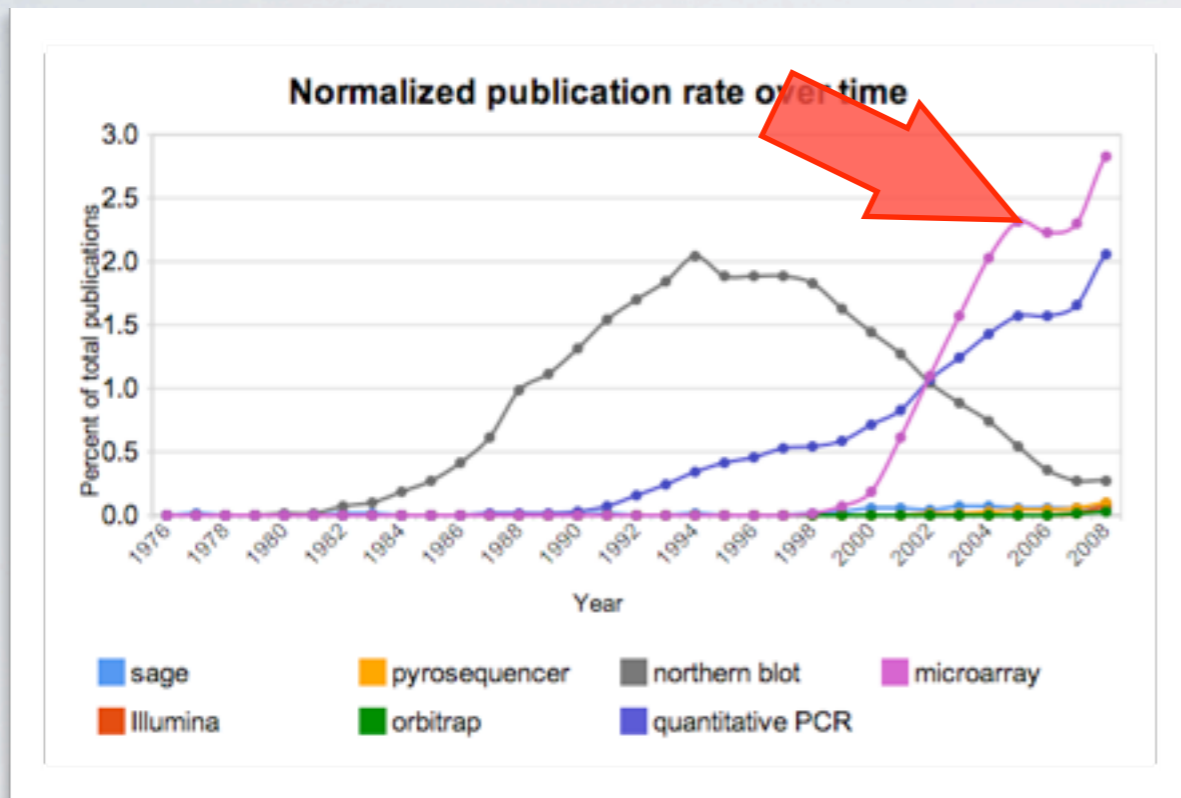
# quantitative PCR

shameless plug #1:  
(for further reading)



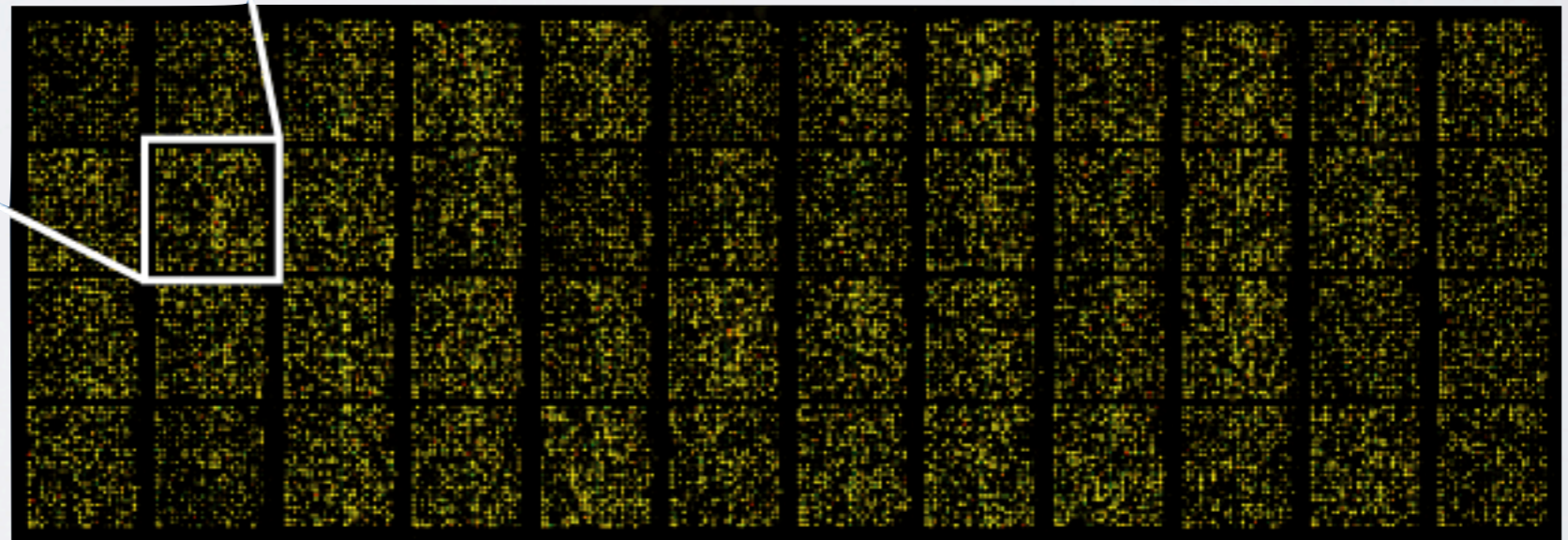
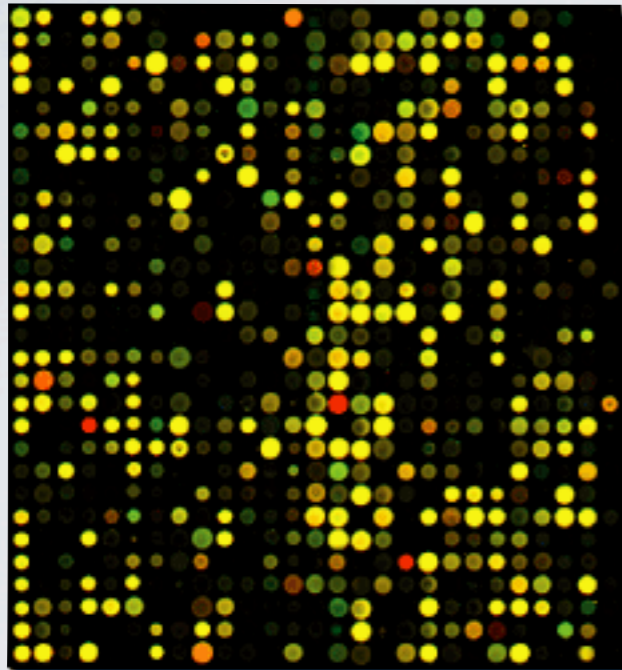
Kitchen et al. Methods (2010) vol. 50 (4)

present



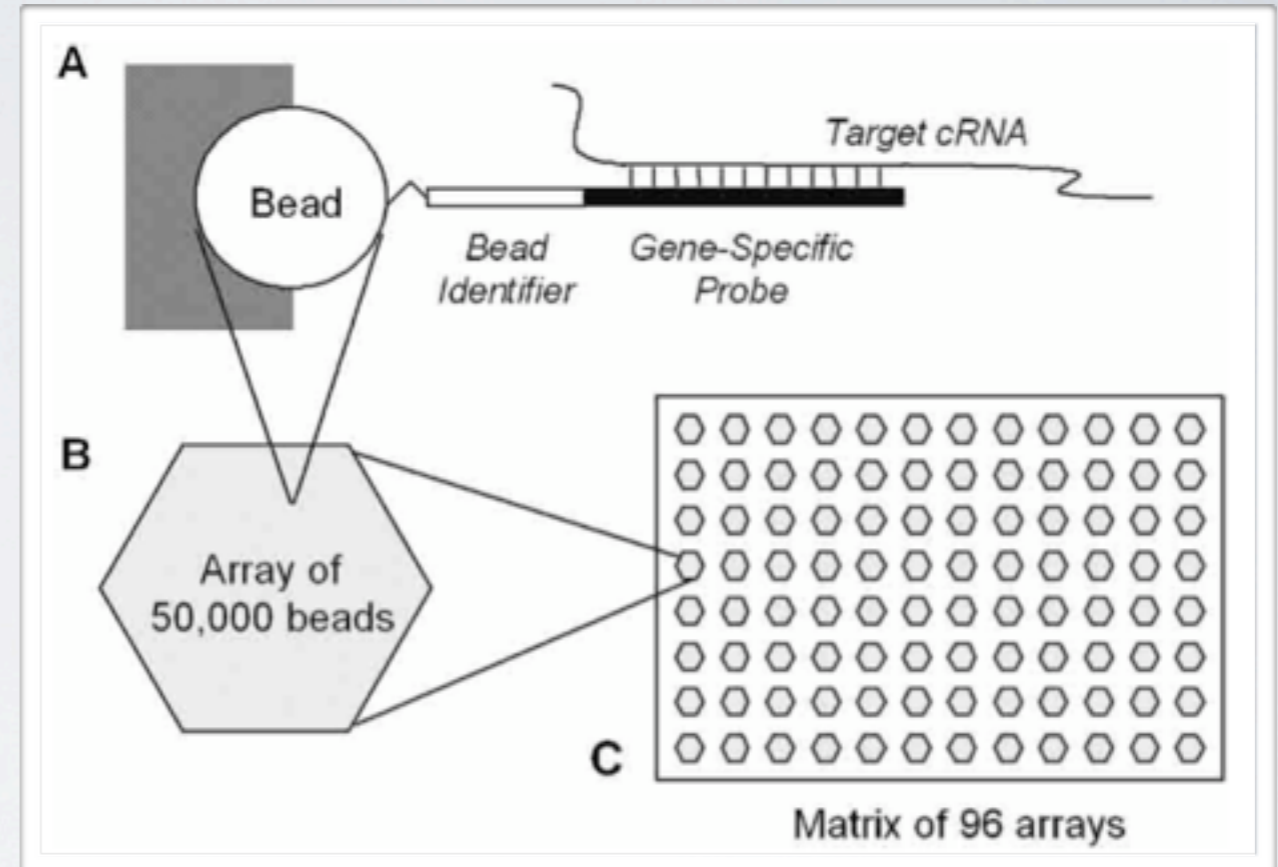
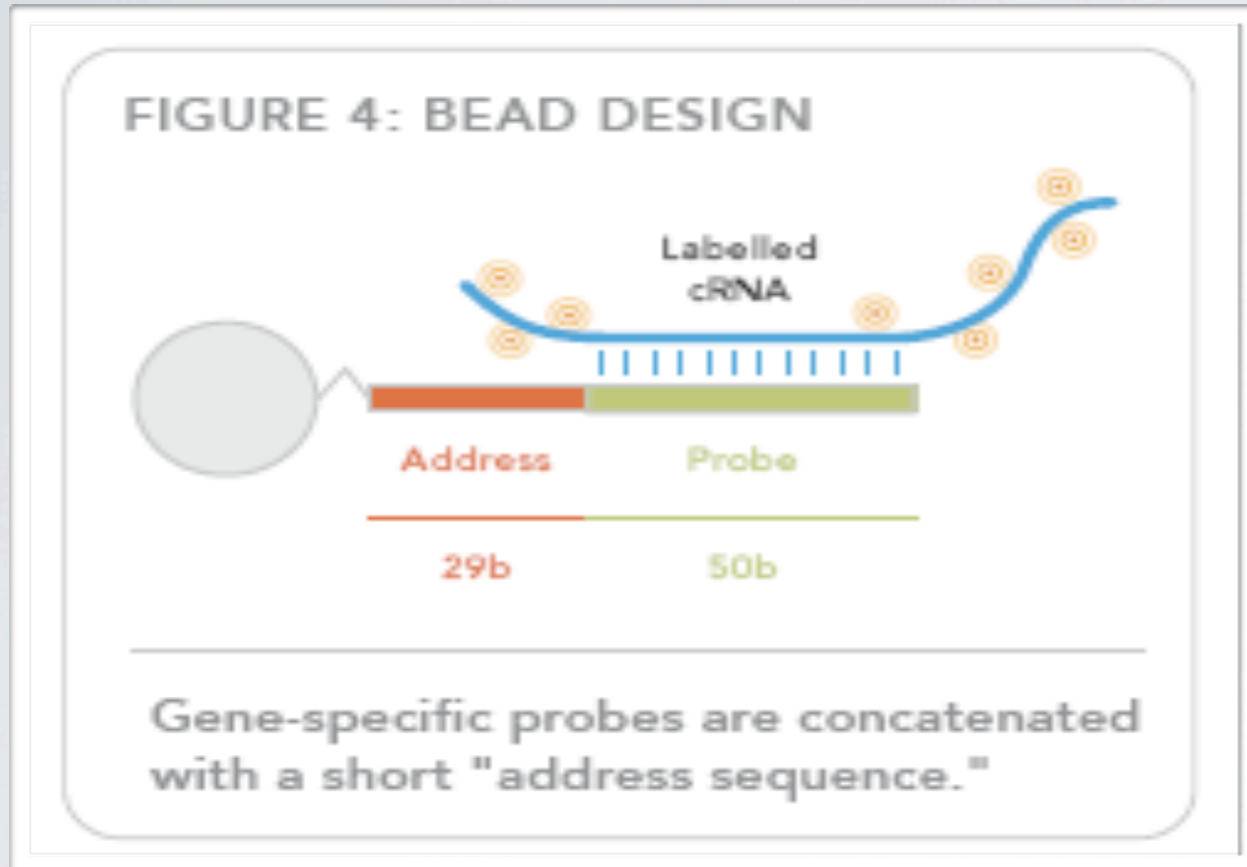


# microarray





# microarray

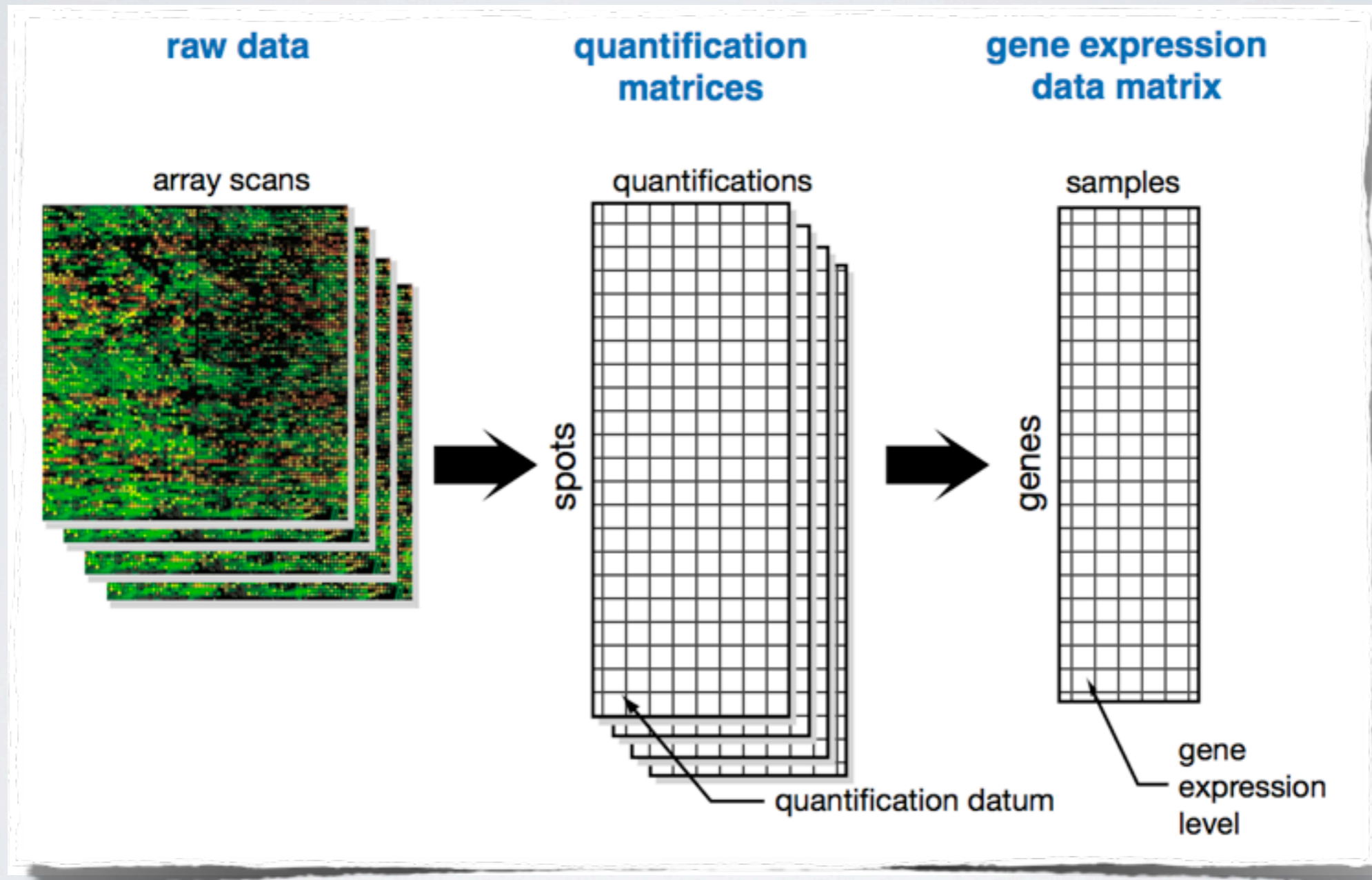


gene activity proportional to number of target molecules captured by bead





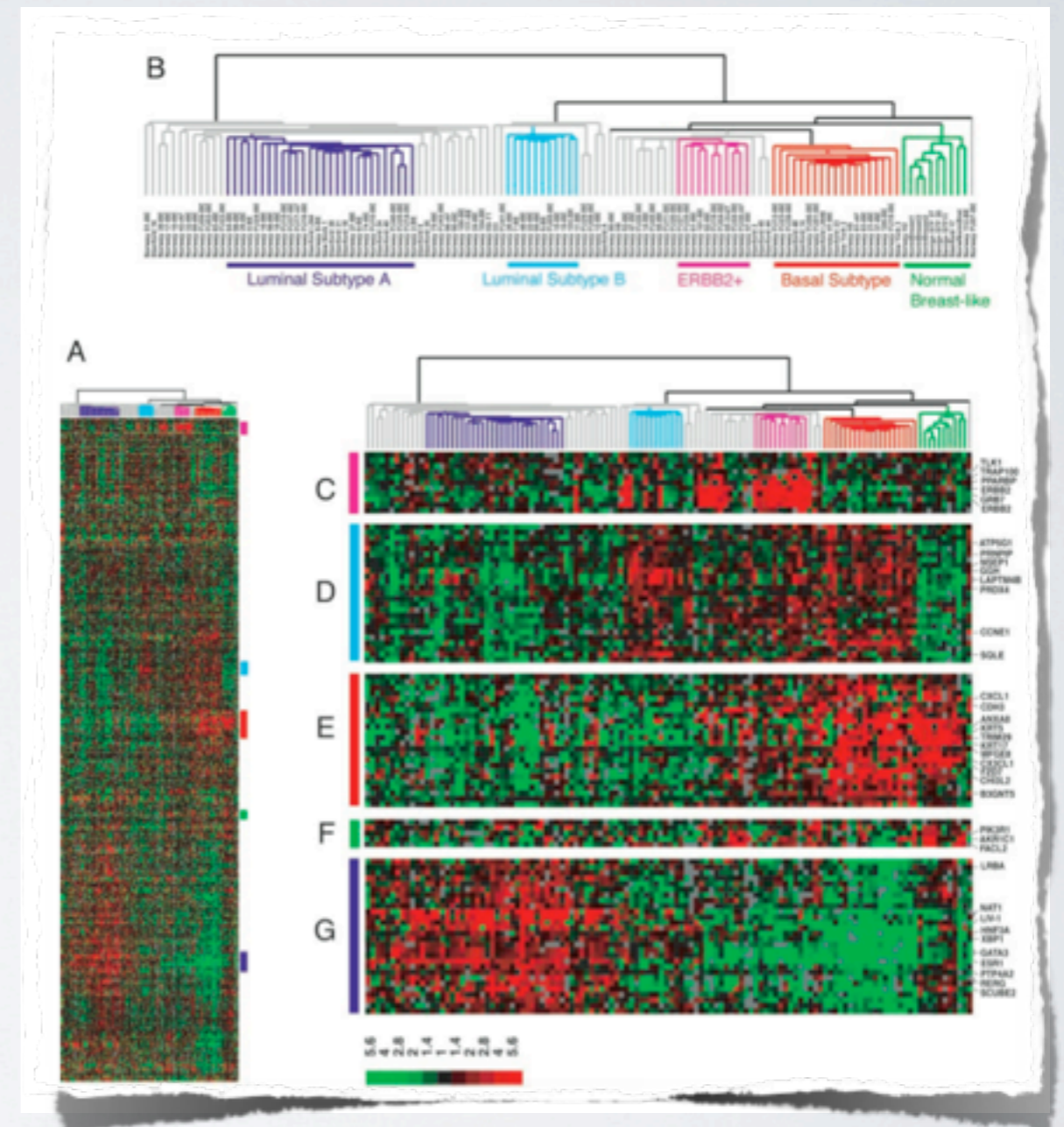
# microarray



# microarray

aim to generate profiles of genes that can **differentiate** between disease states

**single samples** used often due to low availability of primary material and/or limited budget



Sørli et al. 2003 PNAS



# microarray

we're interested in **reproducibility** of microarray experiments

**microarray quality control consortium** (MAQC) engaged in 3-stage project to assess arrays and RNA-seq

stage I [Nat Biotechnol (2006) vol. 24 (9)] looked at **cross-platform** and **inter-laboratory** consistency using dilution series of two reference RNA samples

# microarray

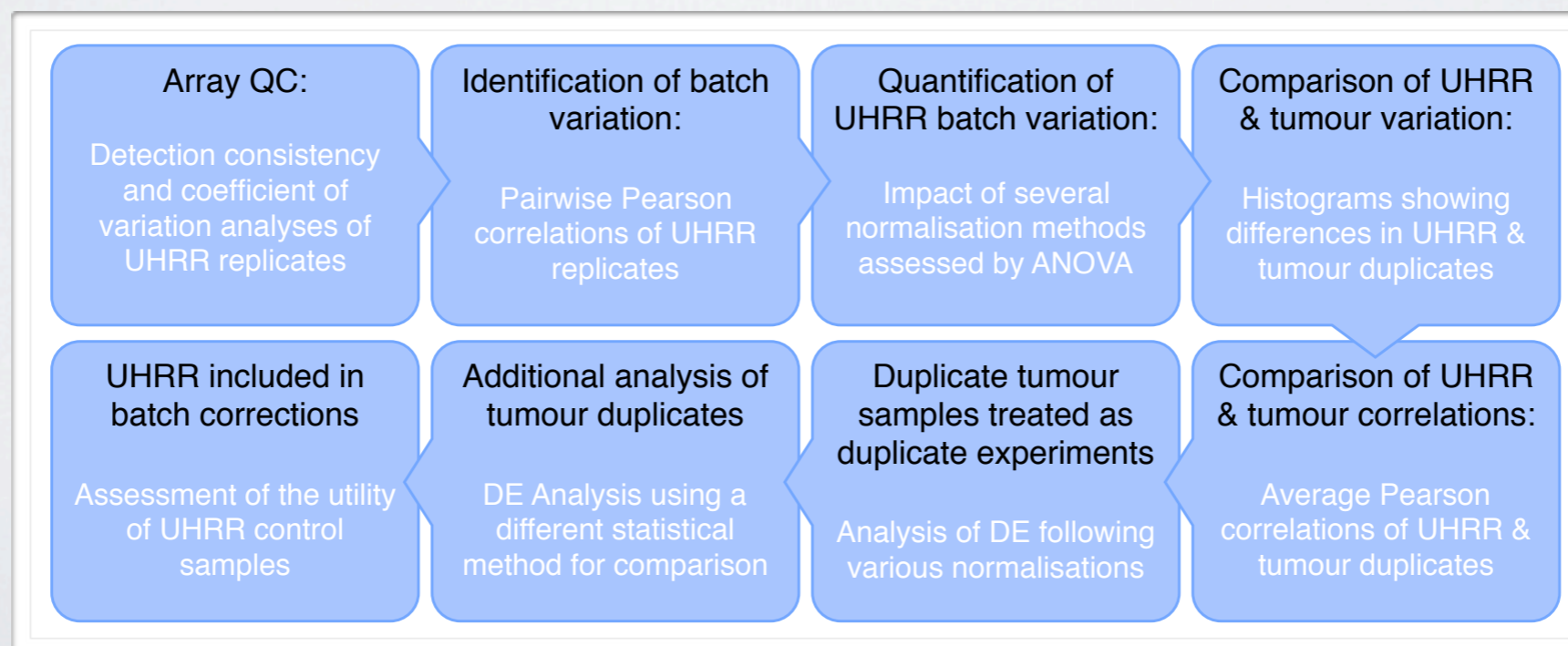
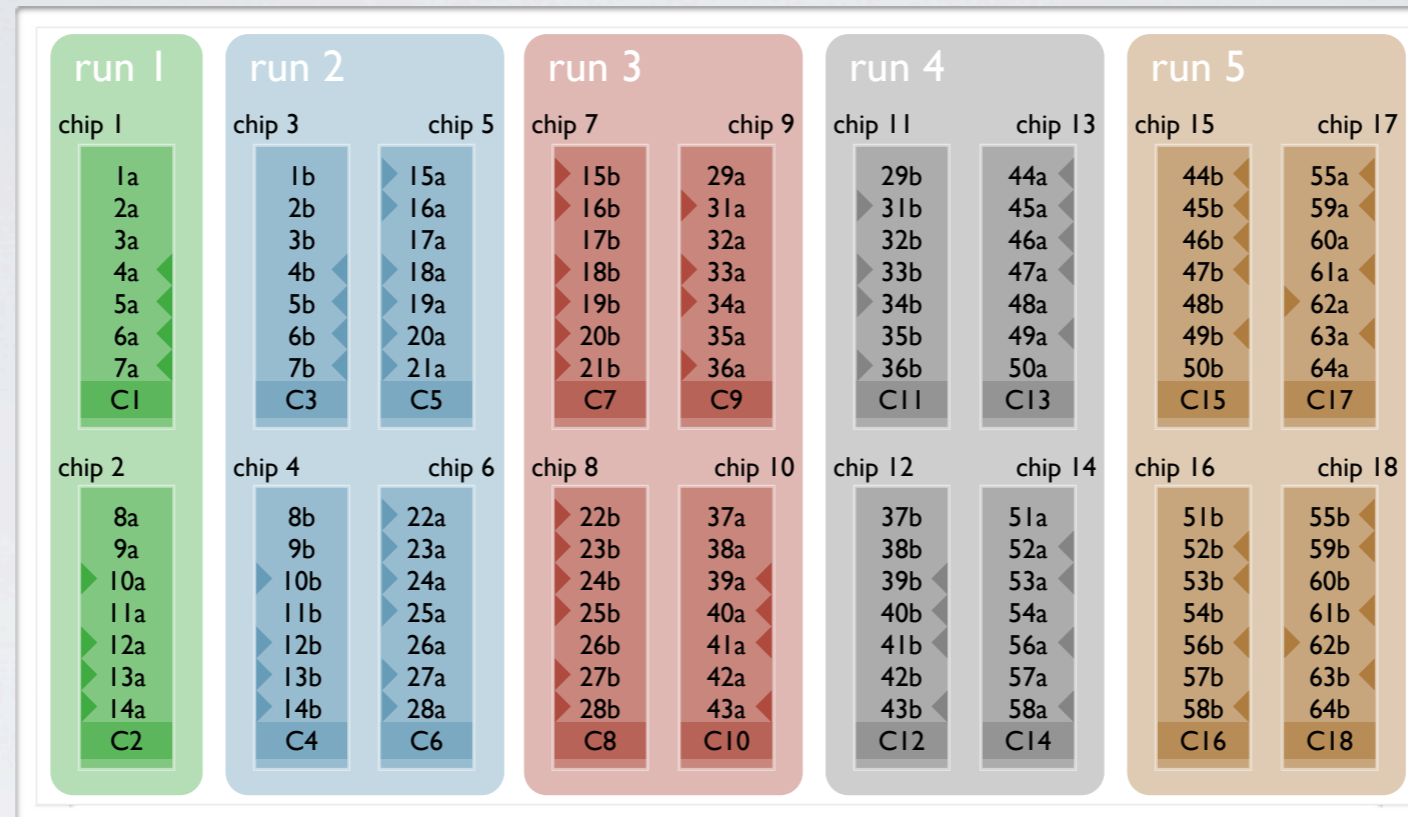
we took a slightly different approach with a specific goal:

*profile intra-experiment technical variation*

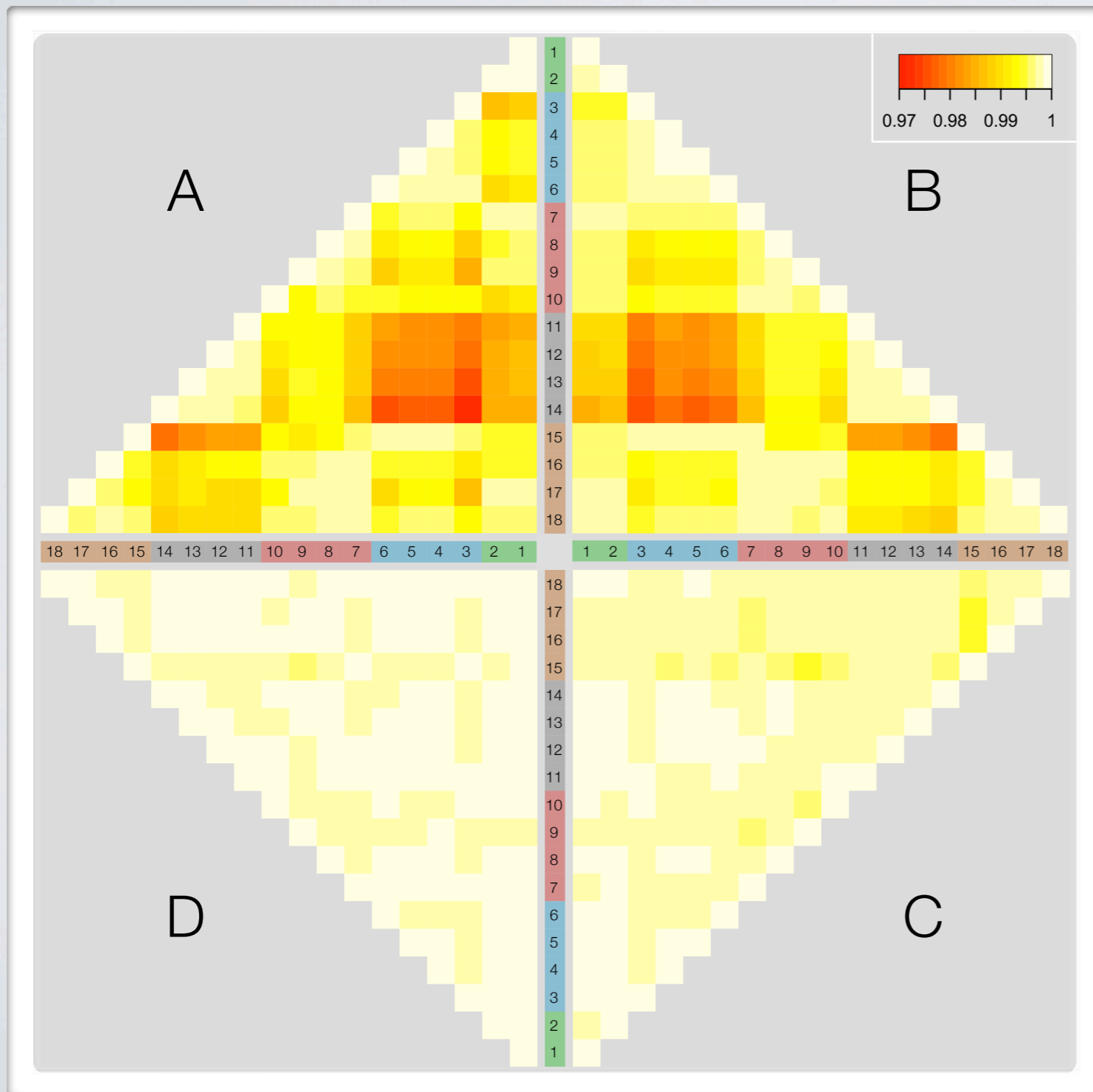
- ✓ same laboratory
- ✓ same technology
- ✓ same type of RNA



# microarray



# microarray



**A** raw

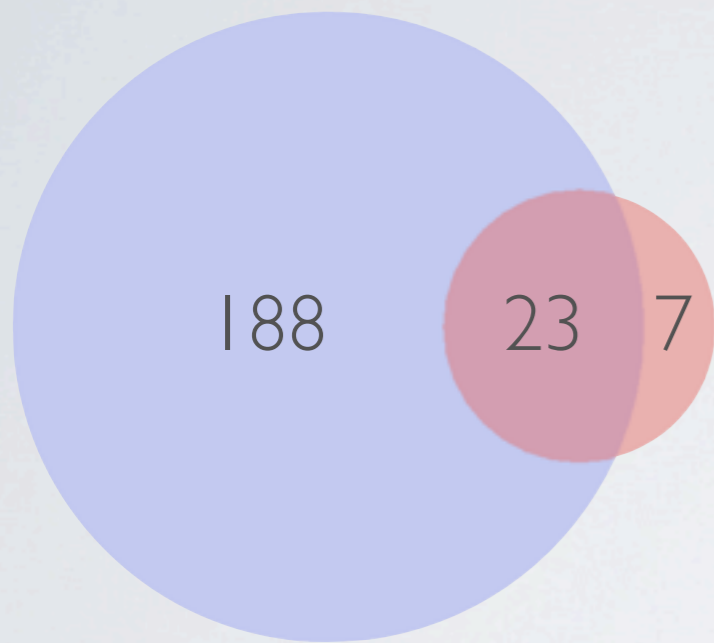
**B** standard  
normalisation

**C** mean-centred

**D** mean-centred &  
variance-adjusted

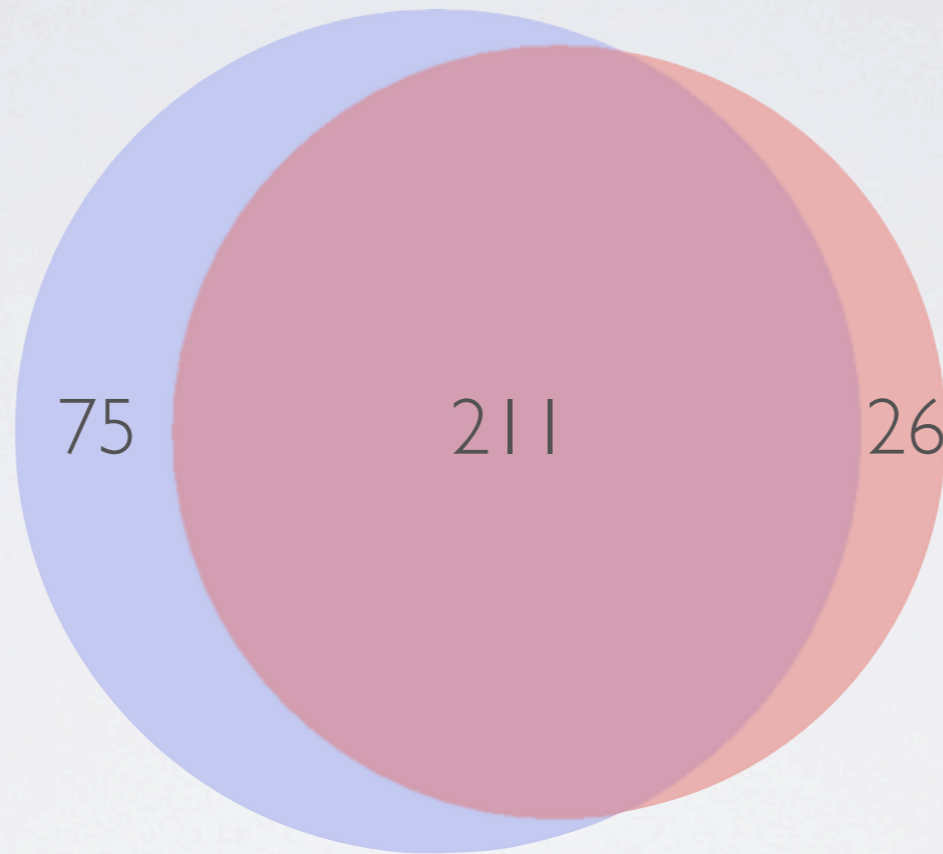


# microarray



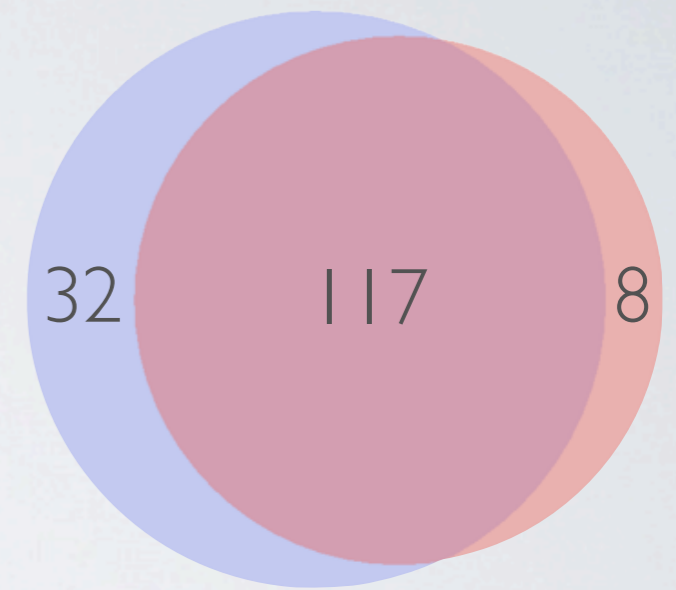
standard  
analysis:

10.6%



batch-corrected  
analysis:

67.6%



batch-corrected  
analysis + ref:

74.5%

# microarray










The image shows a screenshot of a research article page from BMC Genomics. The page features the journal's logo, an Impact Factor of 3.93, and navigation links. The article title is "Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles". The authors listed are Robert R Kitchen, Vicky S Sabine, Andrew H Sims, E JANE Macaskill, Lorna Renshaw, Jeremy S Thomas, Jano I van Hemert, J MICHAEL Dixon, and John M S Bartlett. The article is published in BMC Genomics 2010, volume 11, issue 134, on February 24, 2010. The abstract is provisional. The background section discusses the challenges of microarray technology and the use of ComBat for batch correction. The results section shows that ComBat significantly improved the consistency of gene lists. The conclusion states that single samples can generate reliable data after batch correction.

**BMC Genomics** **IMPACT FACTOR 3.93** [Welcome University of Edinburgh \(Subscriptions\) \(Log on / register\)](#) [Feedback](#) | [Support](#) | [My details](#)

[Home](#) | [Journals A-Z](#) | [subject areas](#) | [advanced search](#) | [authors](#) | [reviewers](#) | [libraries](#) | [about](#) | [my BioMed Central](#)

Research article Open Access

## Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles

Robert R Kitchen , Vicky S Sabine , Andrew H Sims , E JANE Macaskill , Lorna Renshaw , Jeremy S Thomas , Jano I van Hemert , J MICHAEL Dixon  and John M S Bartlett 

BMC Genomics 2010, **11**:134 doi:10.1186/1471-2164-11-134  
Published: 24 February 2010

**Abstract (provisional)**

### Background

Microarray technology is a popular means of producing whole genome transcriptional profiles, however high cost and scarcity of mRNA has led many studies to be conducted based on the analysis of single samples. We exploit the design of the Illumina platform, specifically multiple arrays on each chip, to evaluate intra-experiment technical variation using repeated hybridisations of universal human reference RNA (UHRR) and duplicate hybridisations of primary breast tumour samples from a clinical study.

### Results

A clear batch-specific bias was detected in the measured expressions of both the UHRR and clinical samples. This bias was found to persist following standard microarray normalisation techniques. However, when mean-centering or empirical Bayes batch-correction methods (ComBat) were applied to the data, inter-batch variation in the UHRR and clinical samples were greatly reduced. Correlation between replicate UHRR samples improved by two orders of magnitude following batch-correction using ComBat (ranging from 0.9833-0.9991 to 0.9997-0.9999) and increased the consistency of the gene-lists from the duplicate clinical samples, from 11.6% in quantile normalised data to 66.4% in batch-corrected data. The use of UHRR as an inter-batch calibrator provided a small additional benefit when used in conjunction with ComBat, further increasing the agreement between the two gene-lists, up to 74.1%.

### Conclusion


In the interests of practicalities and cost, these results suggest that single samples can generate reliable data, but only after careful correction for technical bias in...

**BMC Genomics**  
Volume 11

**Viewing options:**

- Abstract
- PDF (6.1MB)

**Associated material:**

- Readers' comments 

**Related literature:**

- Other articles by authors
  - on Google Scholar
  - on PubMed
- Related articles/pages on Google Scholar

**Tools:**

- Download citation(s)
- Email to a friend
- Order reprints
- Post a comment

**Post to:**

- Citeulike
- Connotea
- Del.icio.us
- Facebook
- Mendeley
- Twitter

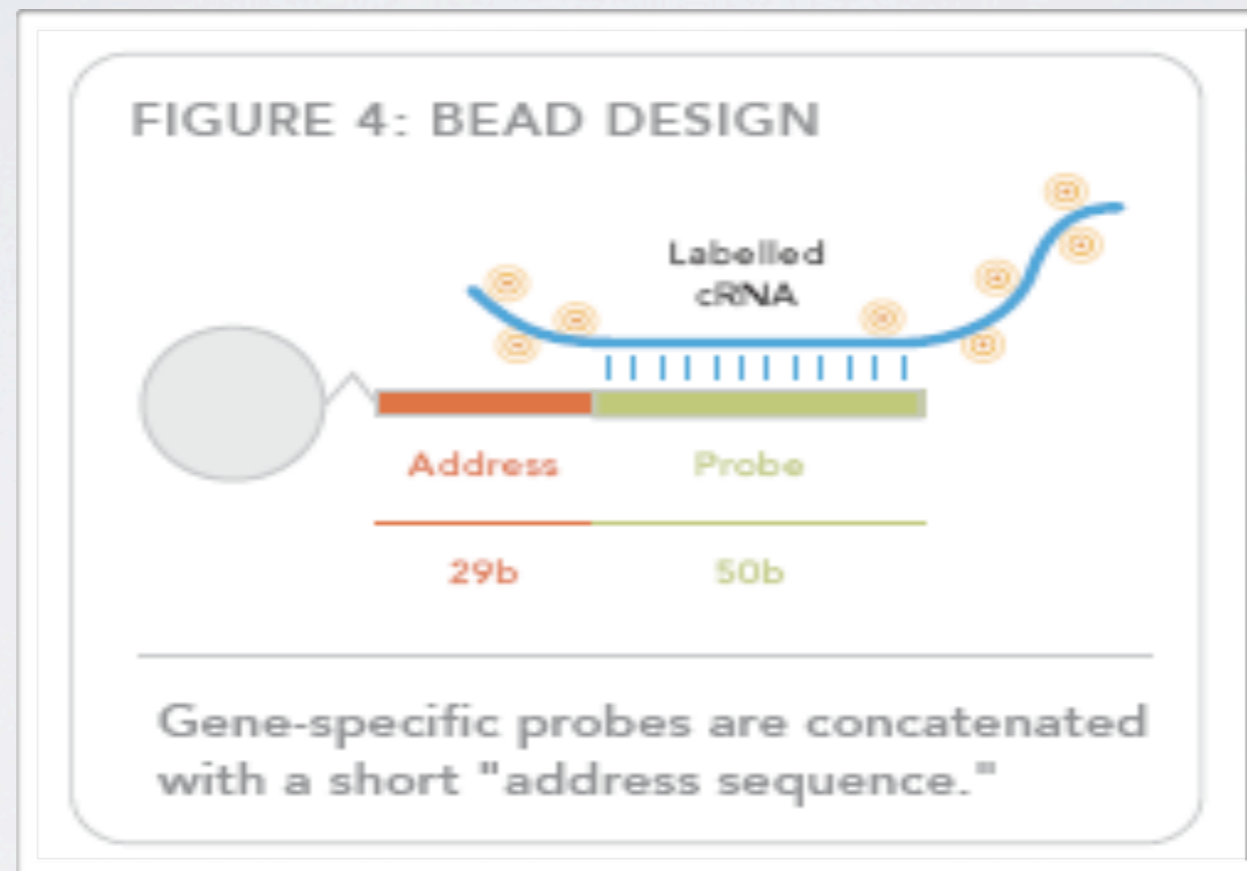
shameless plug #2:  
(for further reading)

Kitchen et al. BMC Genomics (2010) vol. 11 (1)



# microarray

current interest is in correlating observed experiment noise with specific **probe properties**:



eg: TGGGAAAGAACACAGAGGAATCCAGCCATTTCCACAGCGTCCAGCTCTGC

# microarray

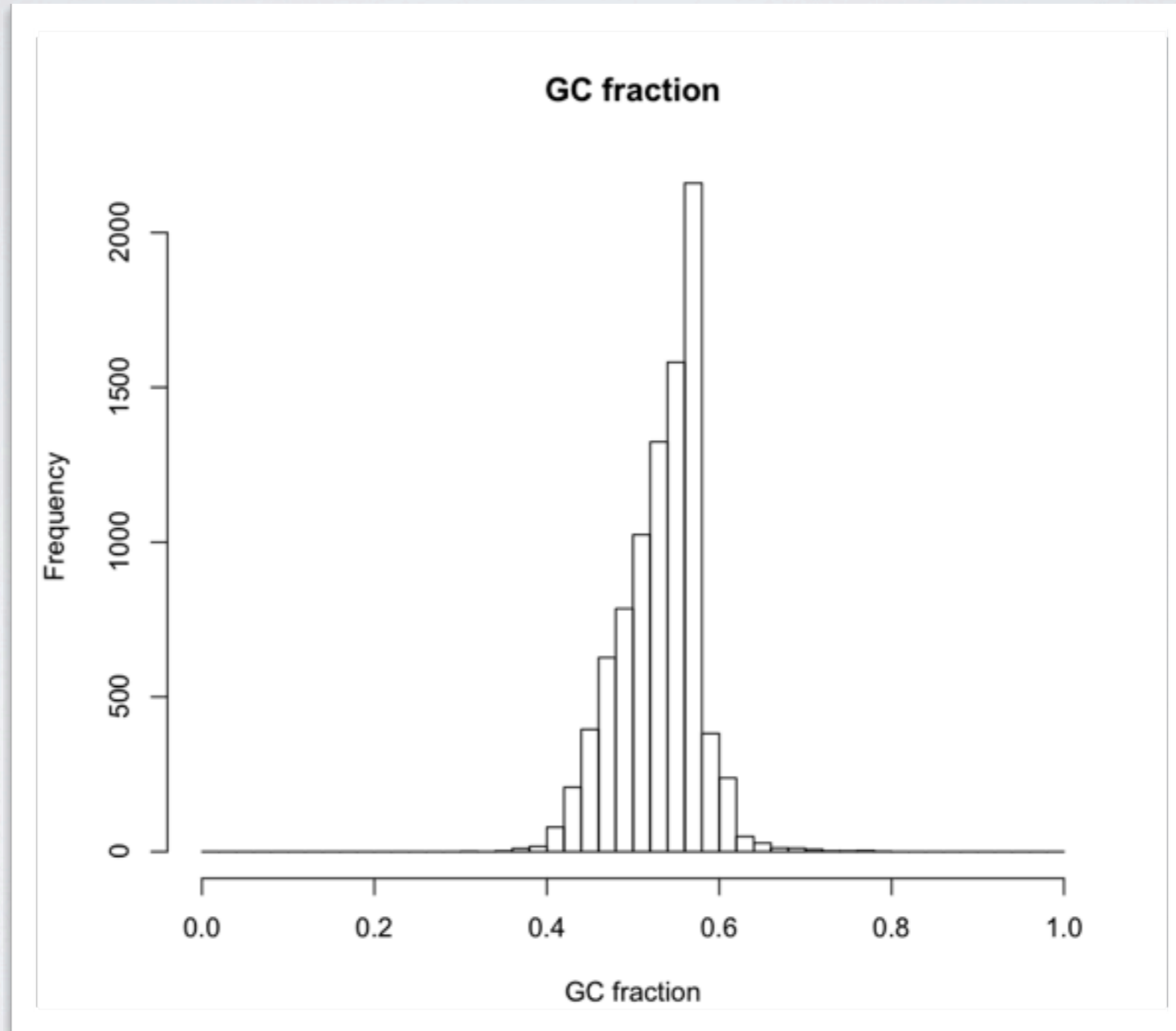
current interest is in correlating observed experiment noise with specific **probe properties**:

for example:

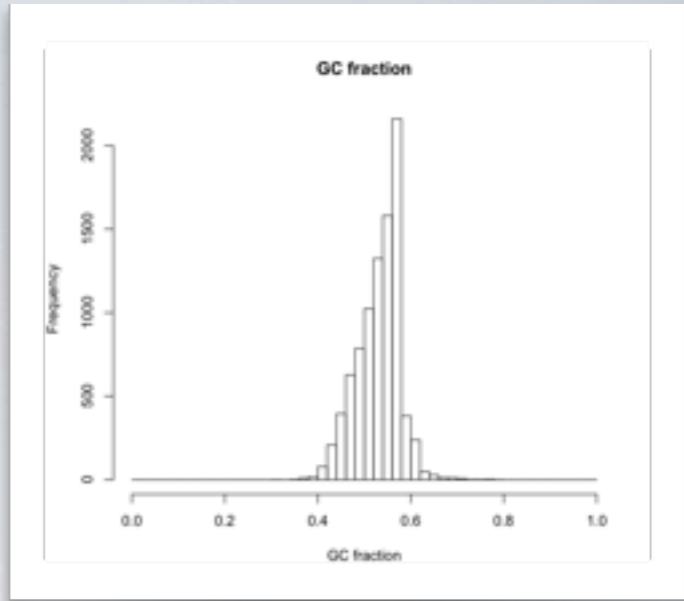
- probe position within target gene
- number of transcripts consecutively hit by the probe
- GC content
- CpG count
- rough total length of target gene



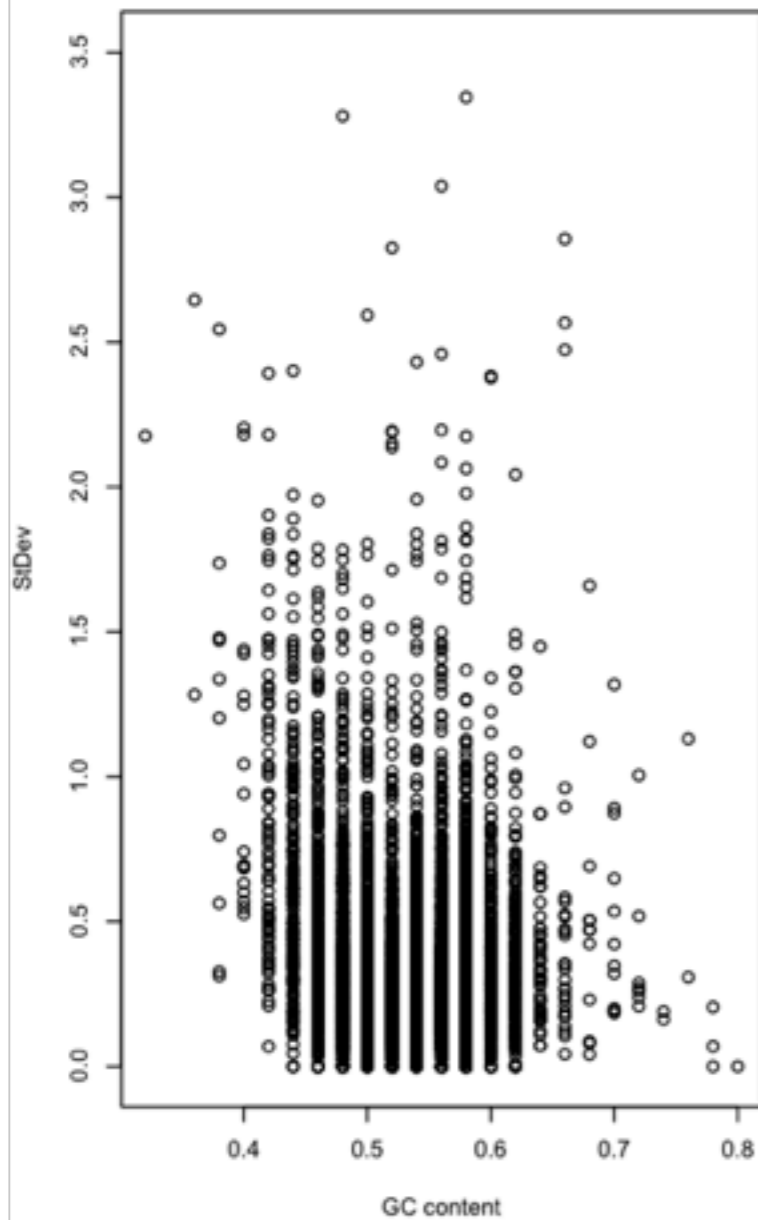
# microarray



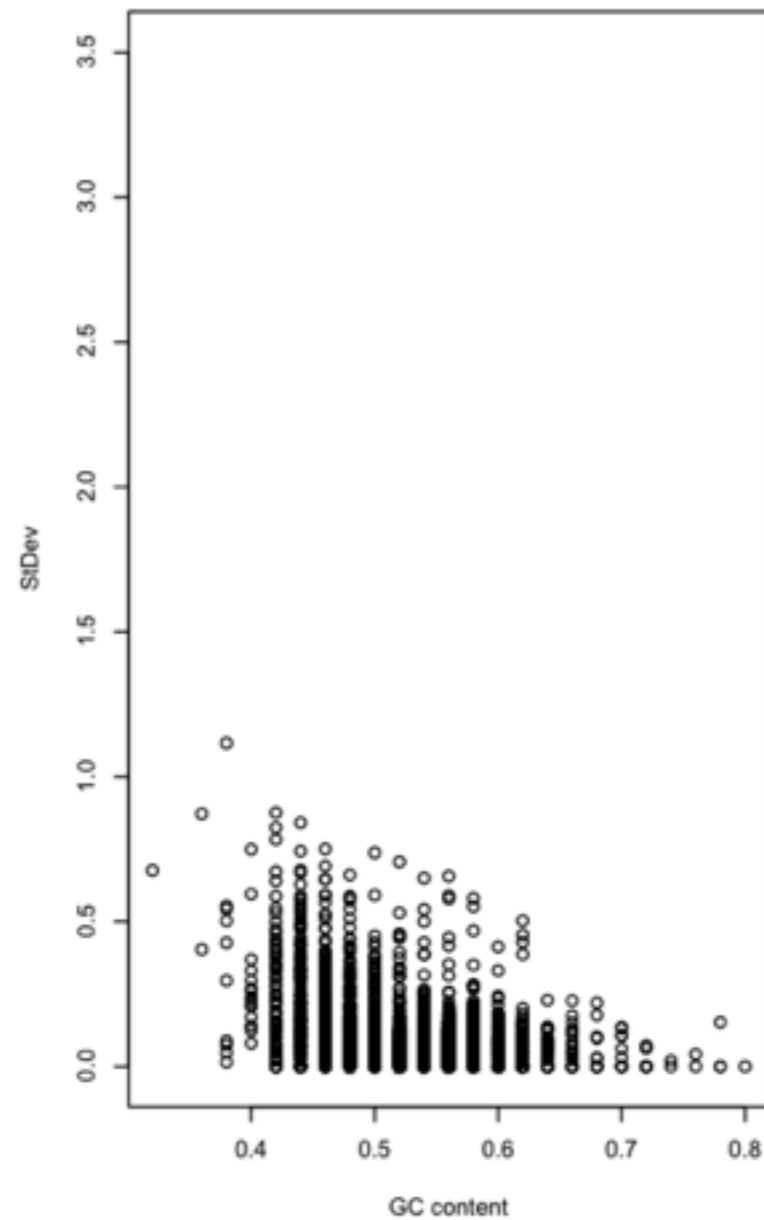
# microarray



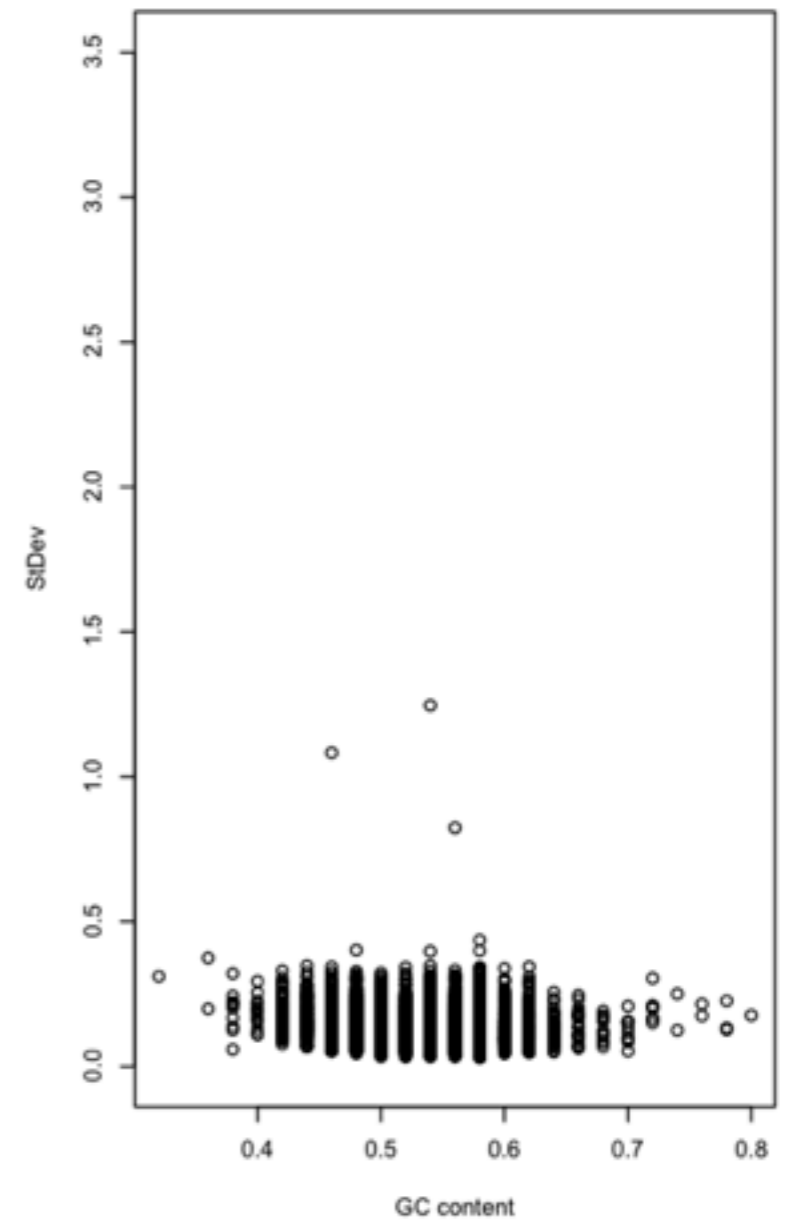
(DF QN) GC content vs. inter-expt SD



(DF QN) GC content vs. inter-run SD

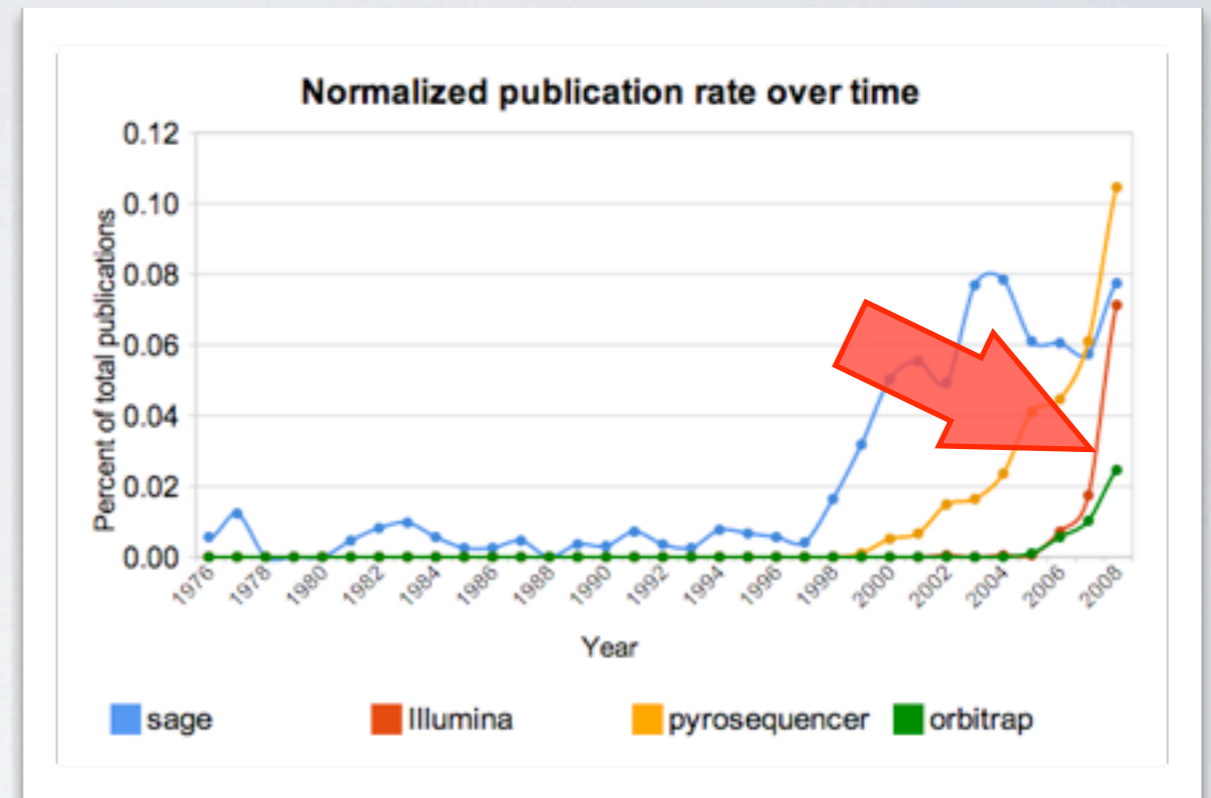
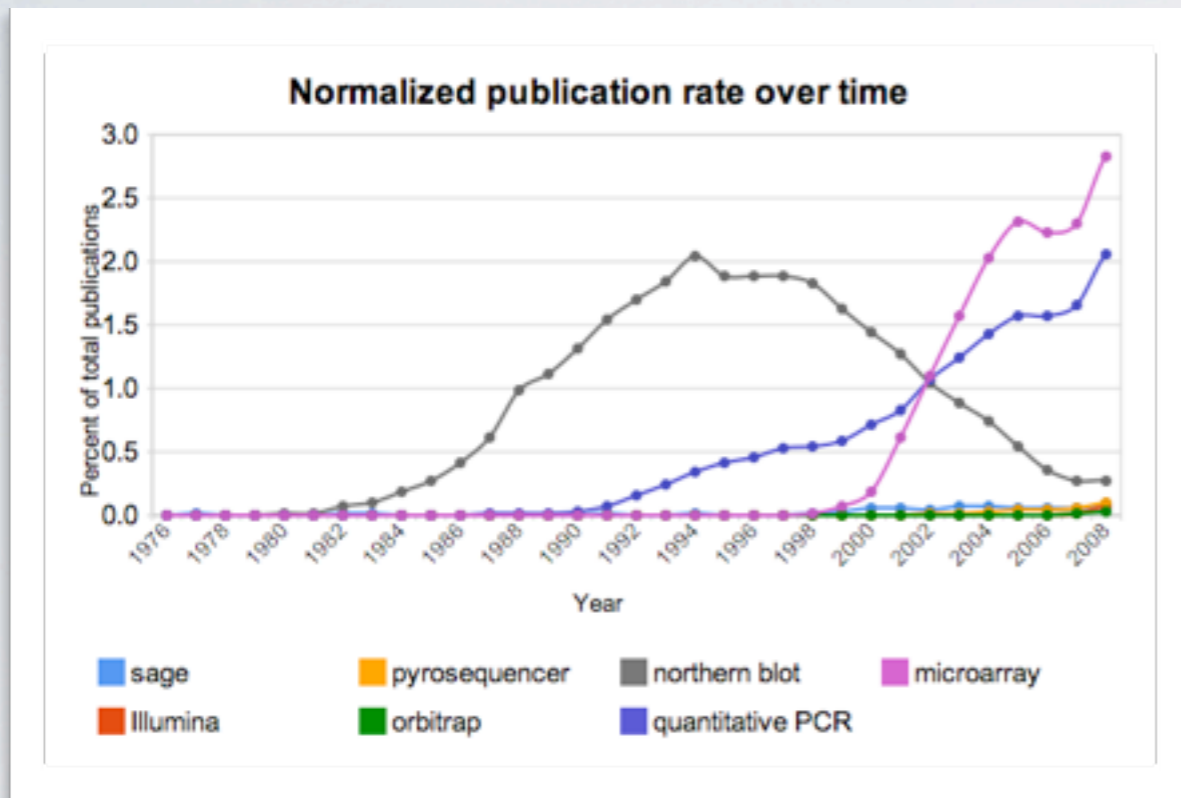


(DF QN) GC content vs. inter-chip SD

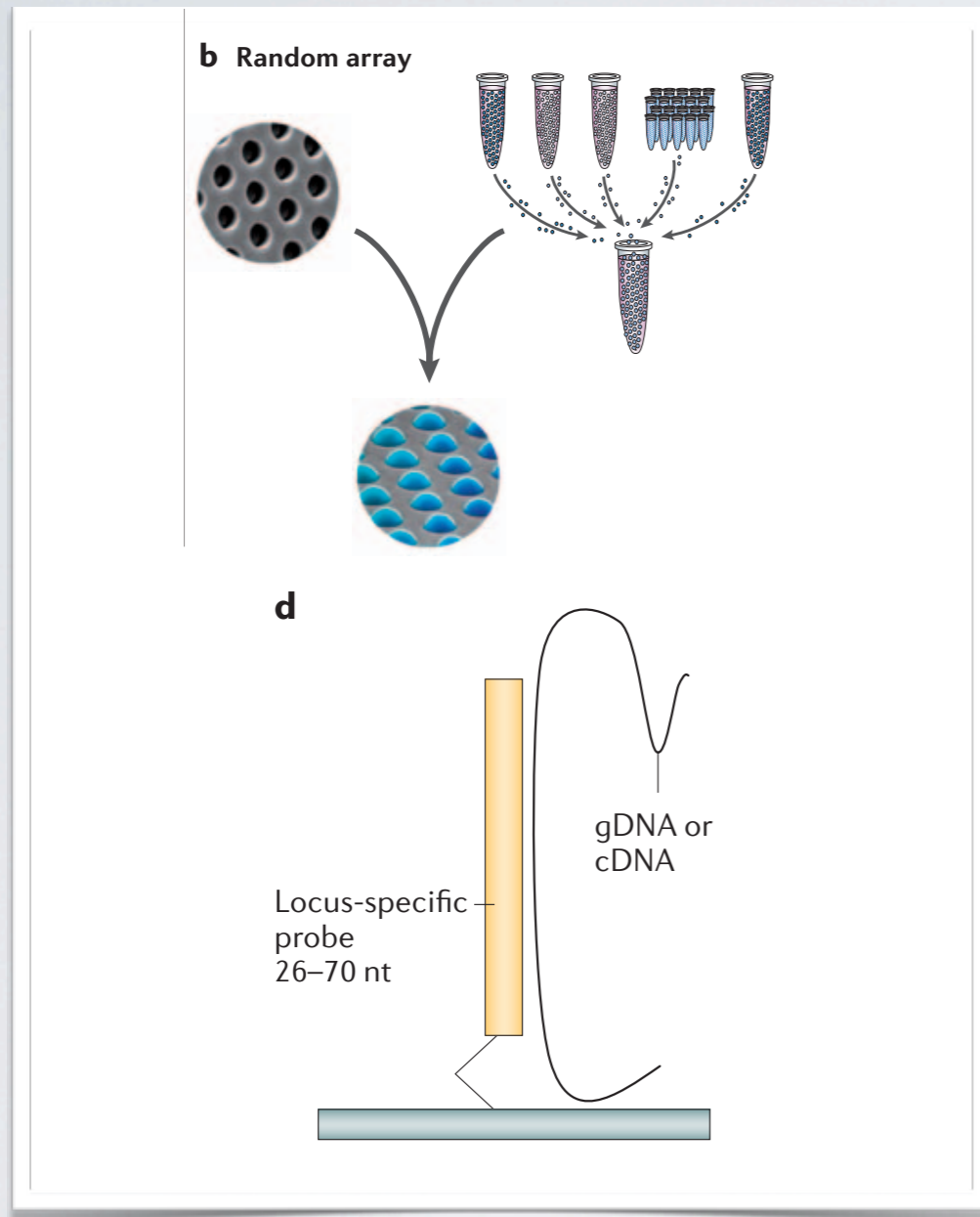




# present

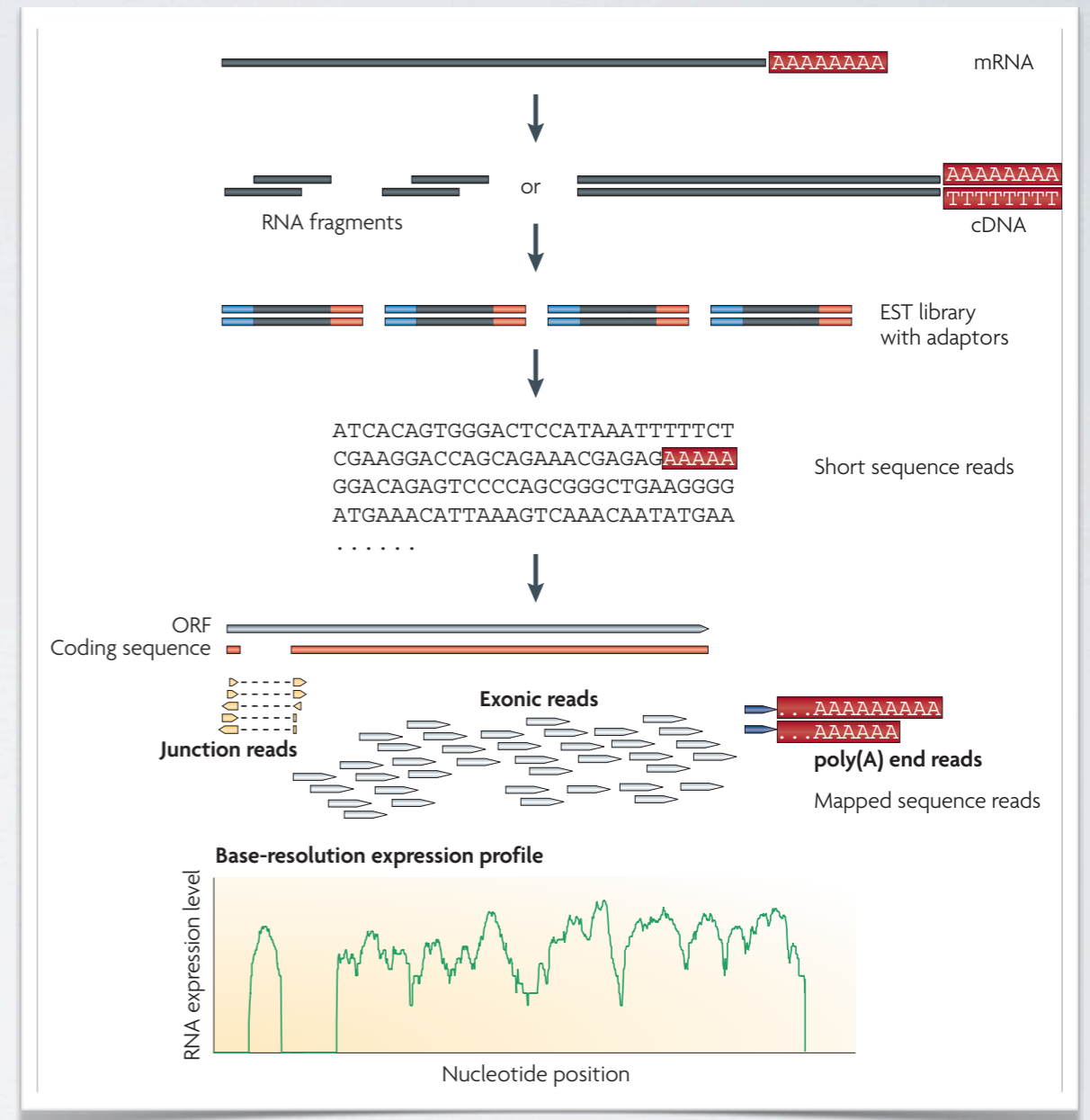


# 2<sup>nd</sup> generation sequencing



microarray

Fan, Chee, Gunderson (2006)

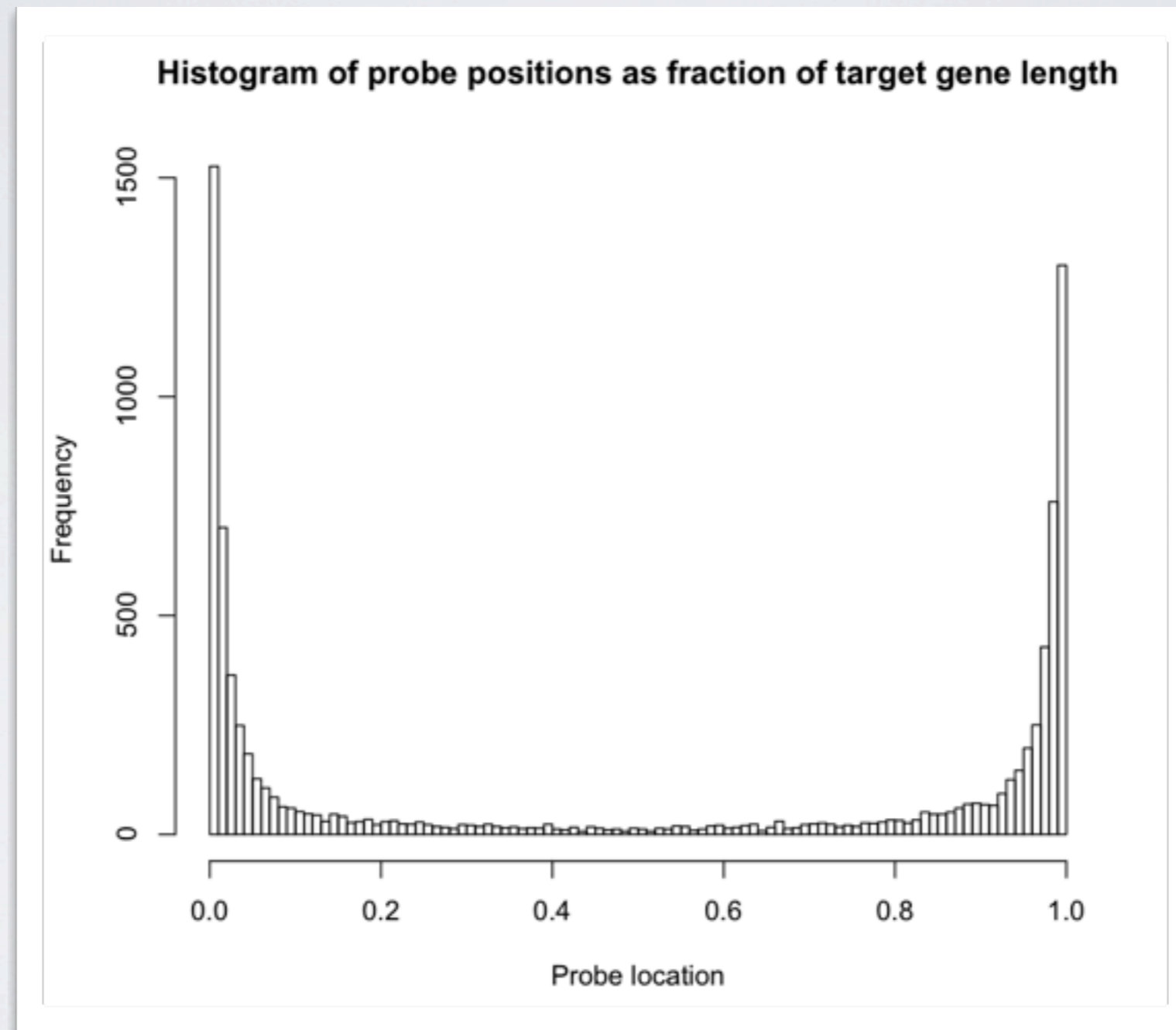


RNA-seq

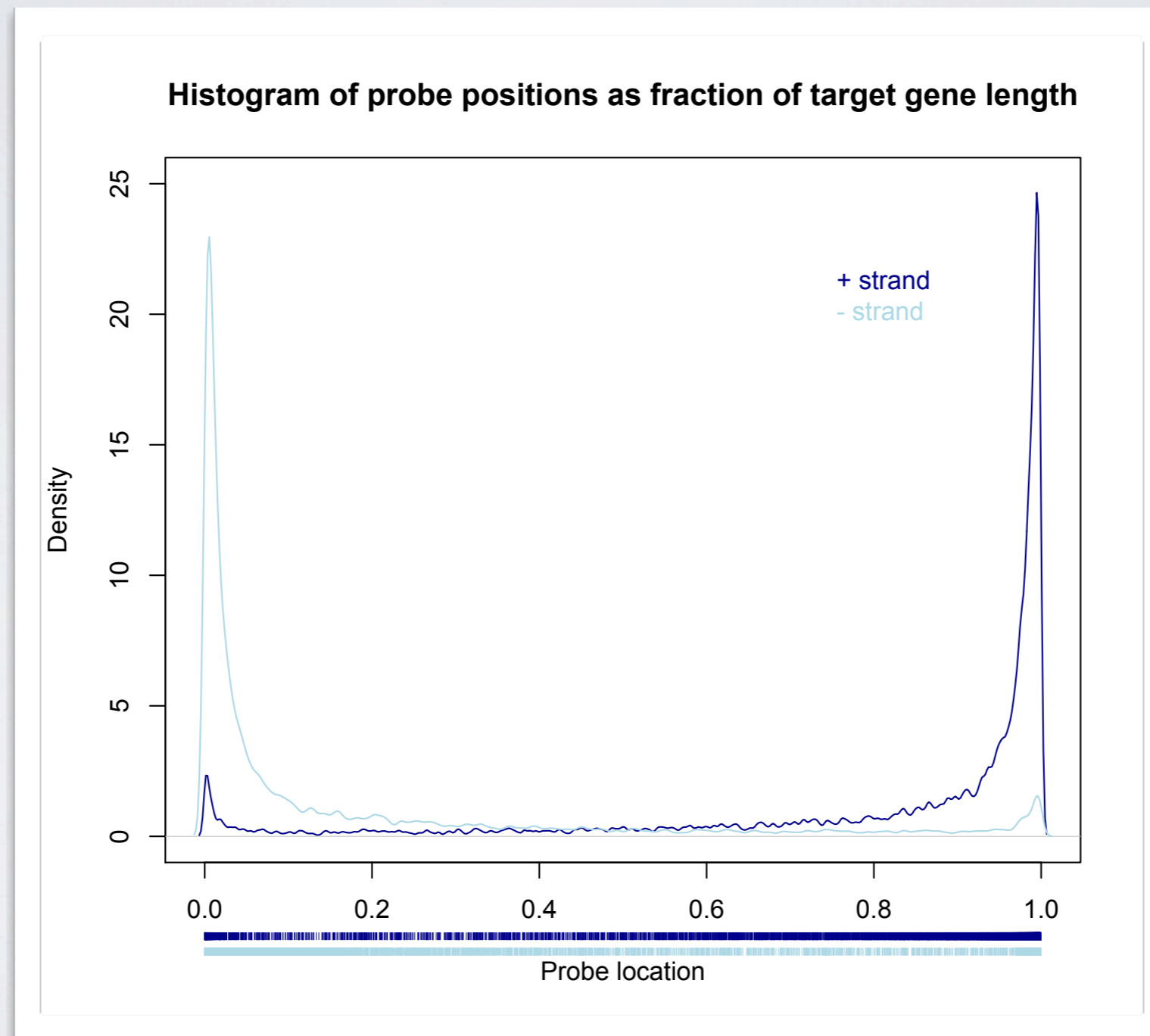
Wang, Gerstein, Snyder (2009)



# 2<sup>nd</sup> generation sequencing



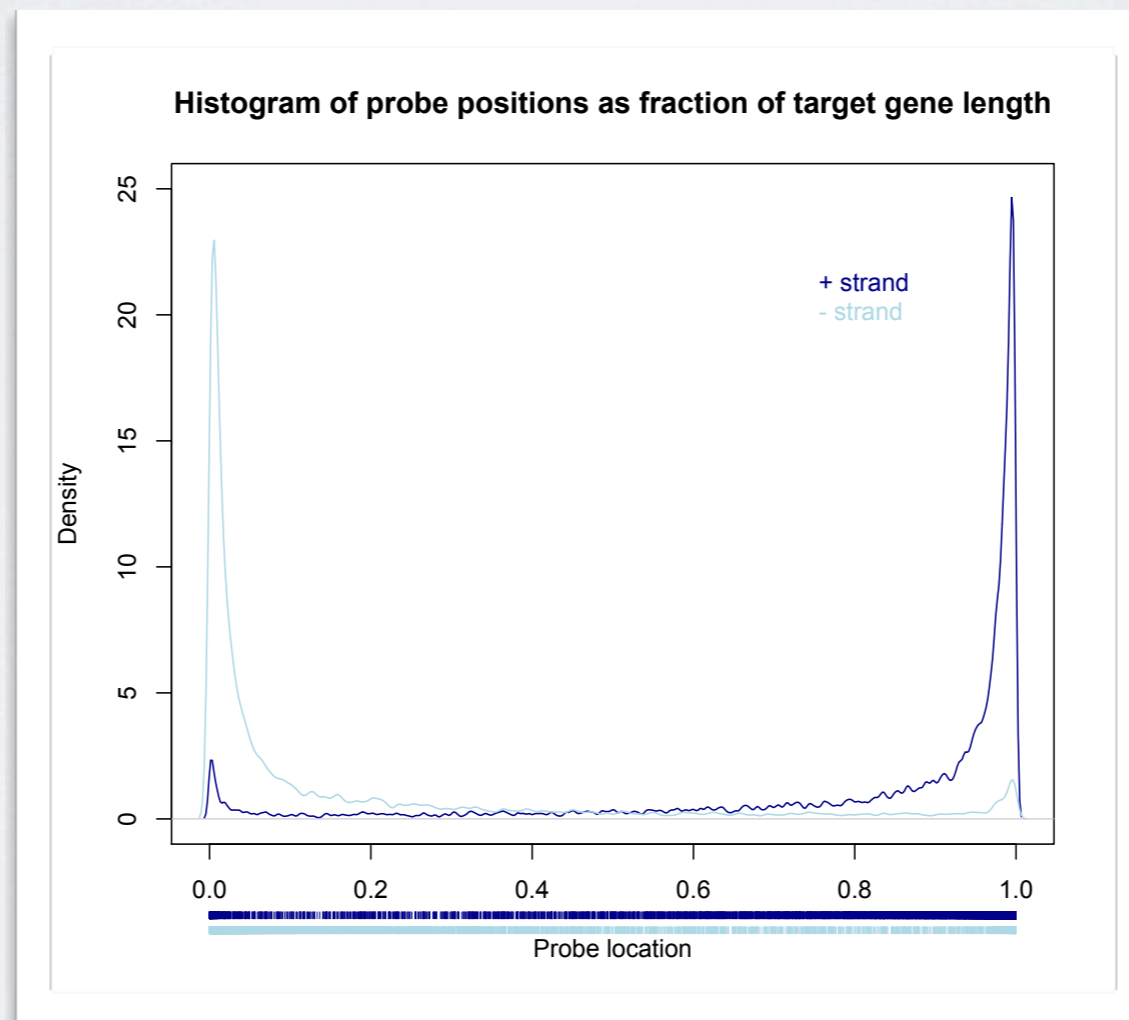
# 2<sup>nd</sup> generation sequencing





# 2<sup>nd</sup> generation sequencing

gene:



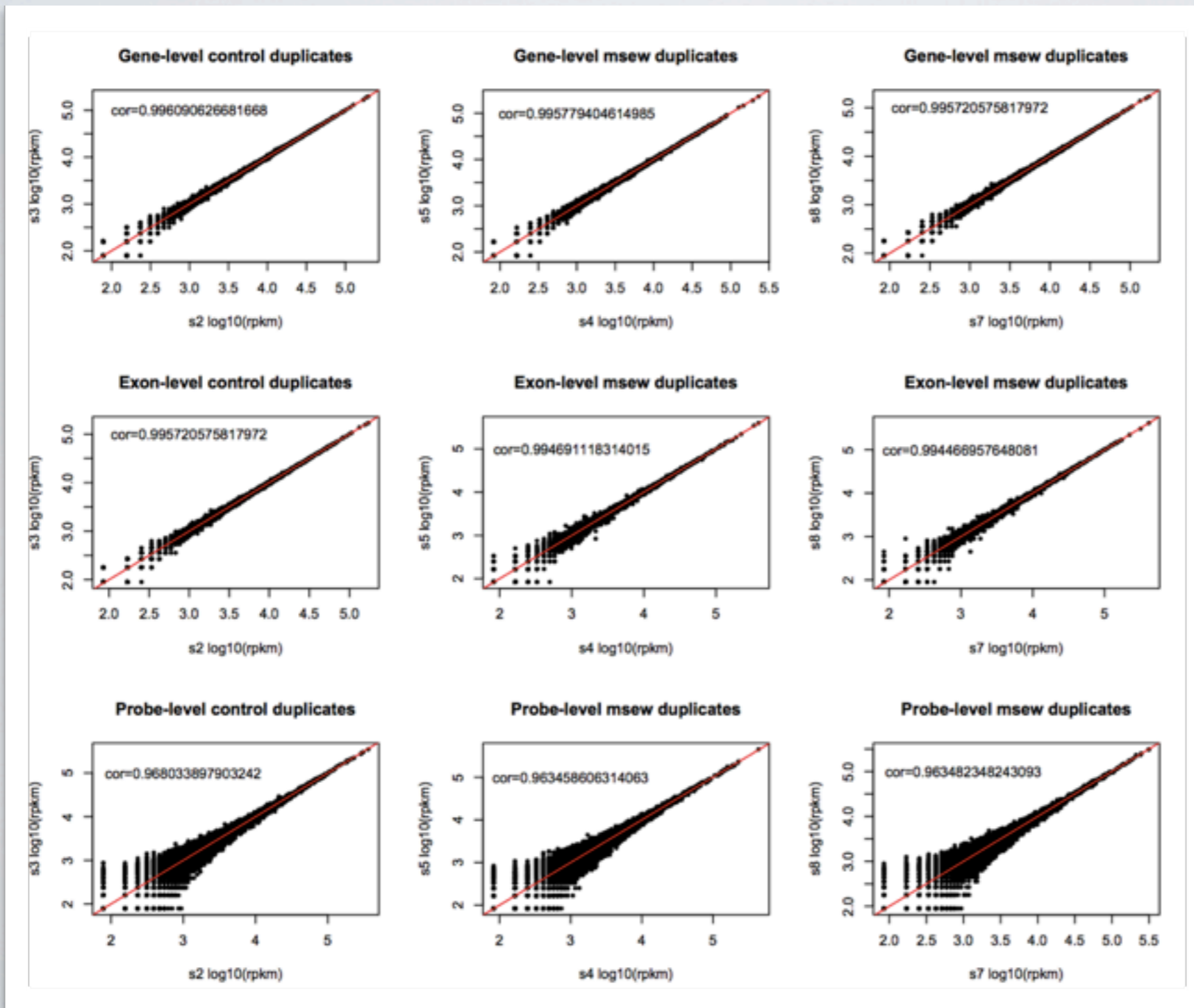
# 2<sup>nd</sup> generation sequencing

correlation

99.6%

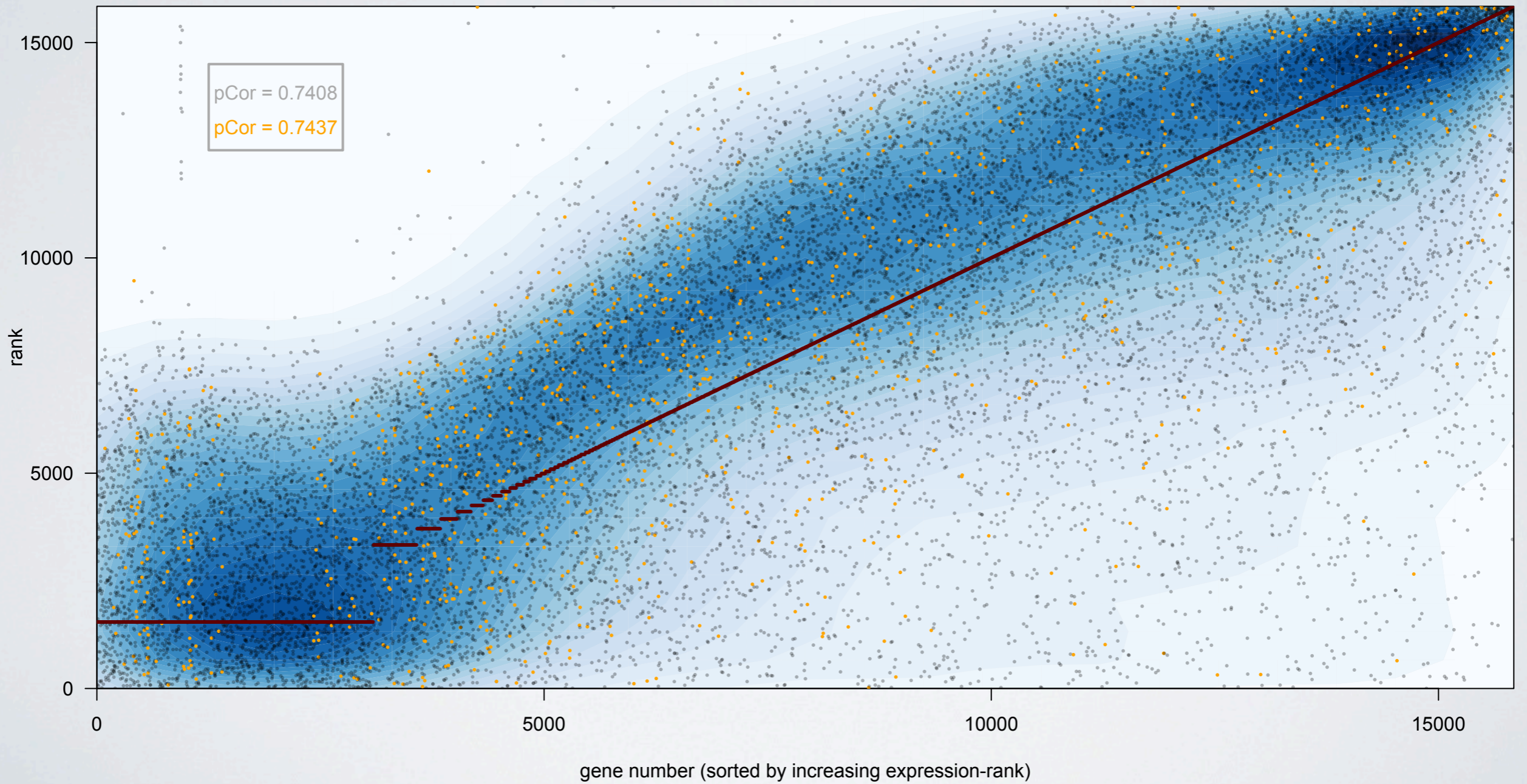
99.5%

96.4%

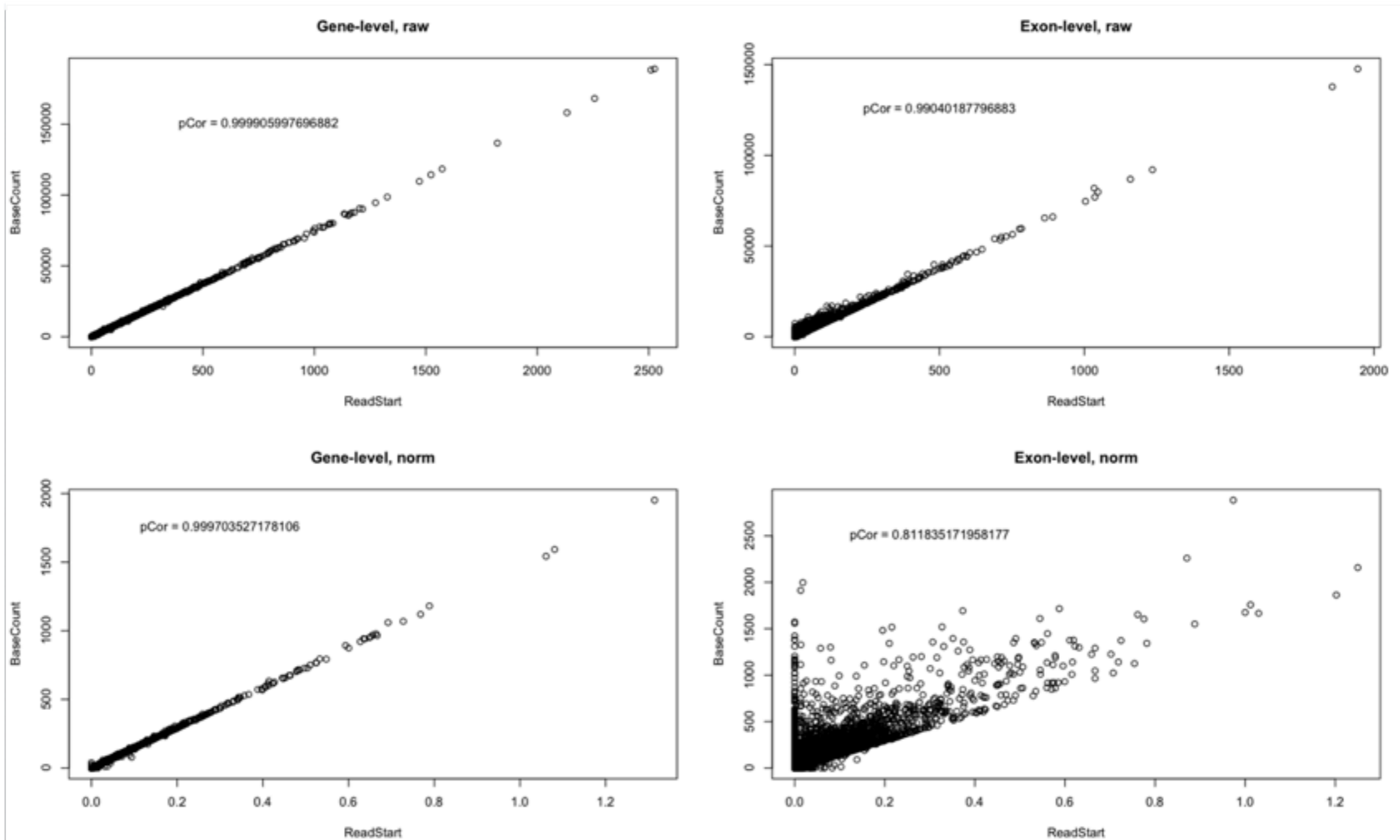




# 2<sup>nd</sup> generation sequencing

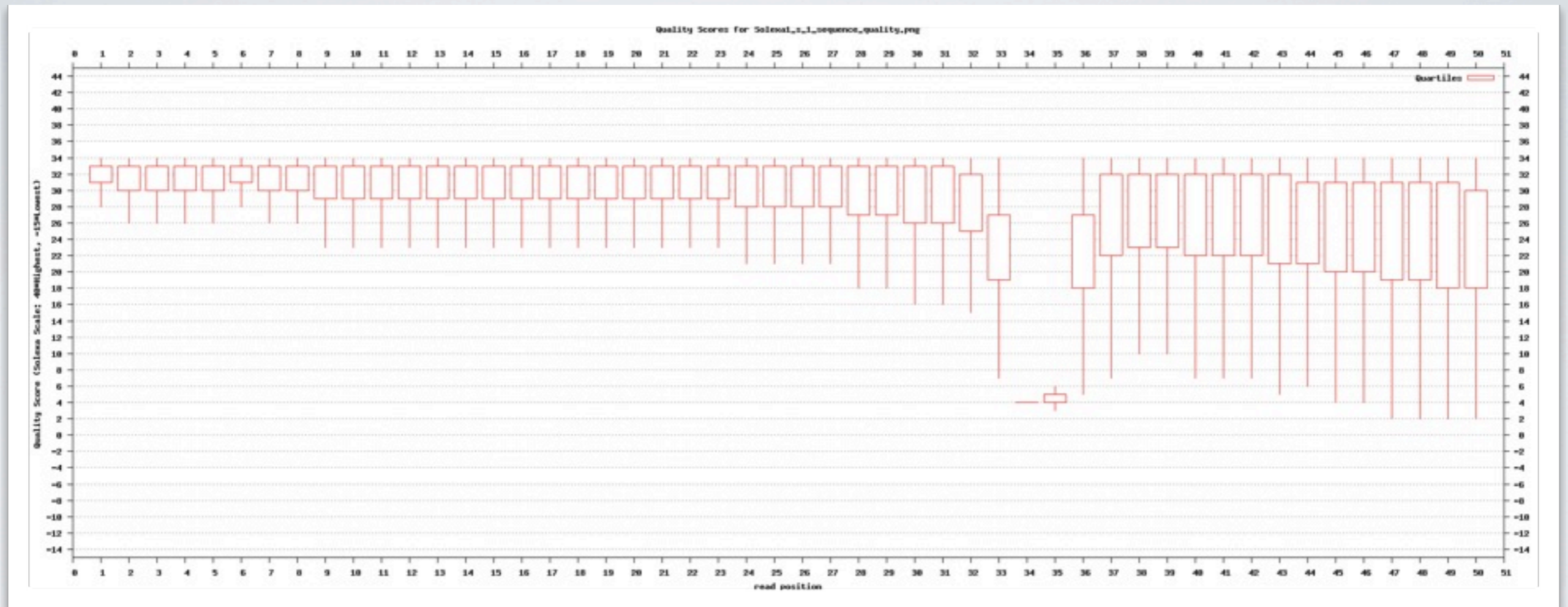


# 2<sup>nd</sup> generation sequencing



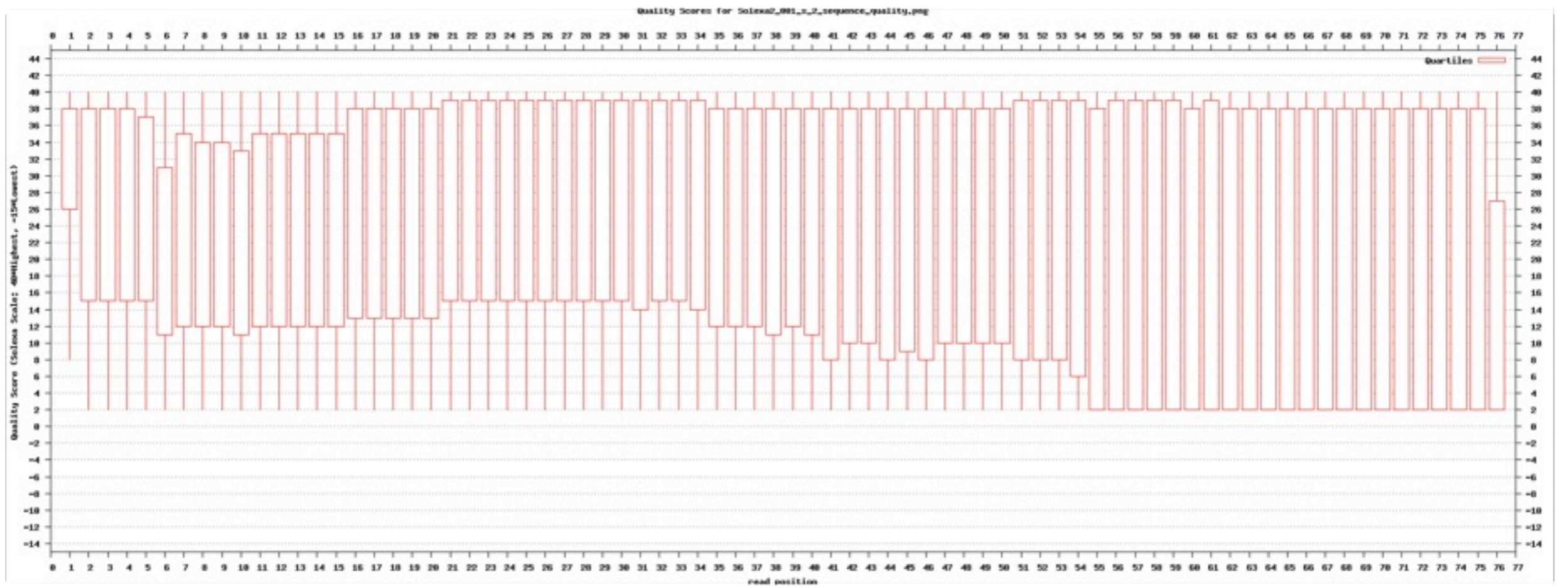


# 2<sup>nd</sup> generation sequencing



first run

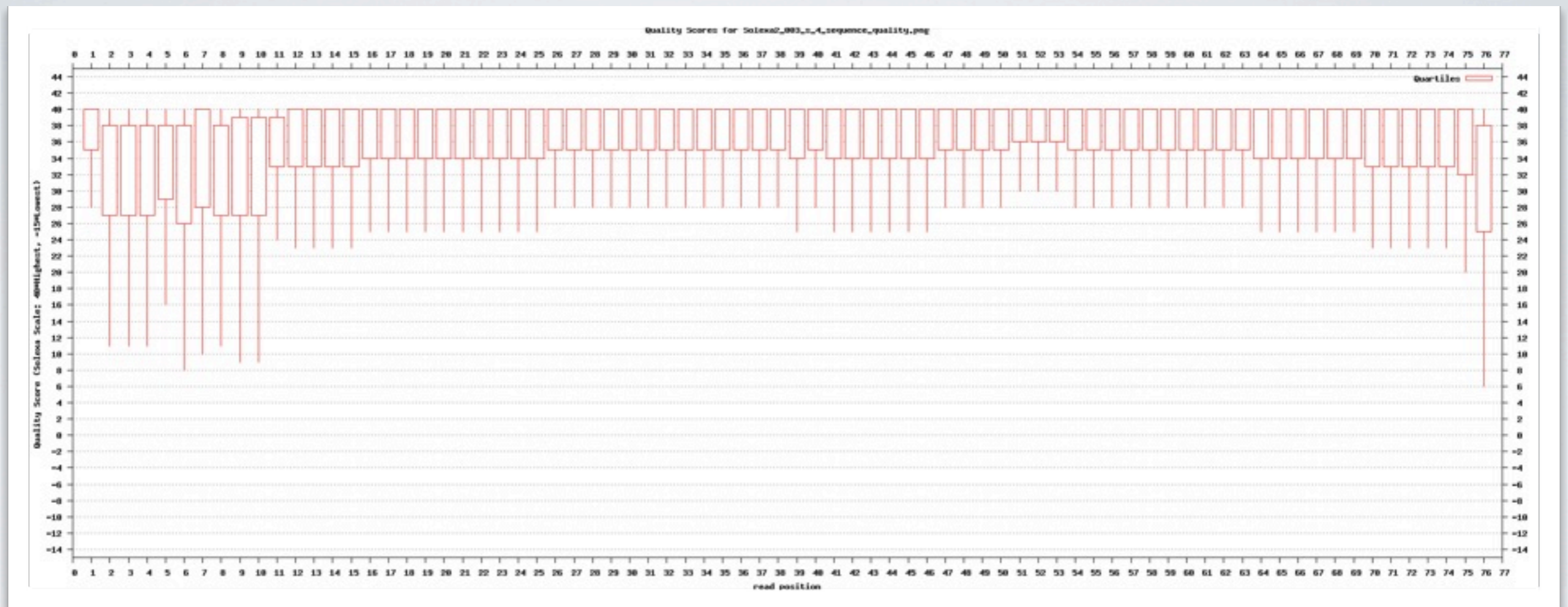
# 2<sup>nd</sup> generation sequencing



second run



# 2<sup>nd</sup> generation sequencing



second run - fresh samples

# label-free proteomics

similar, in principle, to RNA-seq:

## RNAseq

- ✓ many short fragments
- ✓ assign fragments to parent RNA transcript
- ✓ incomplete reference

## LF-MS/MS proteomics

- ✓ many short fragments
- ✓ assign fragments to parent protein
- ✓ very incomplete reference



# label-free proteomics

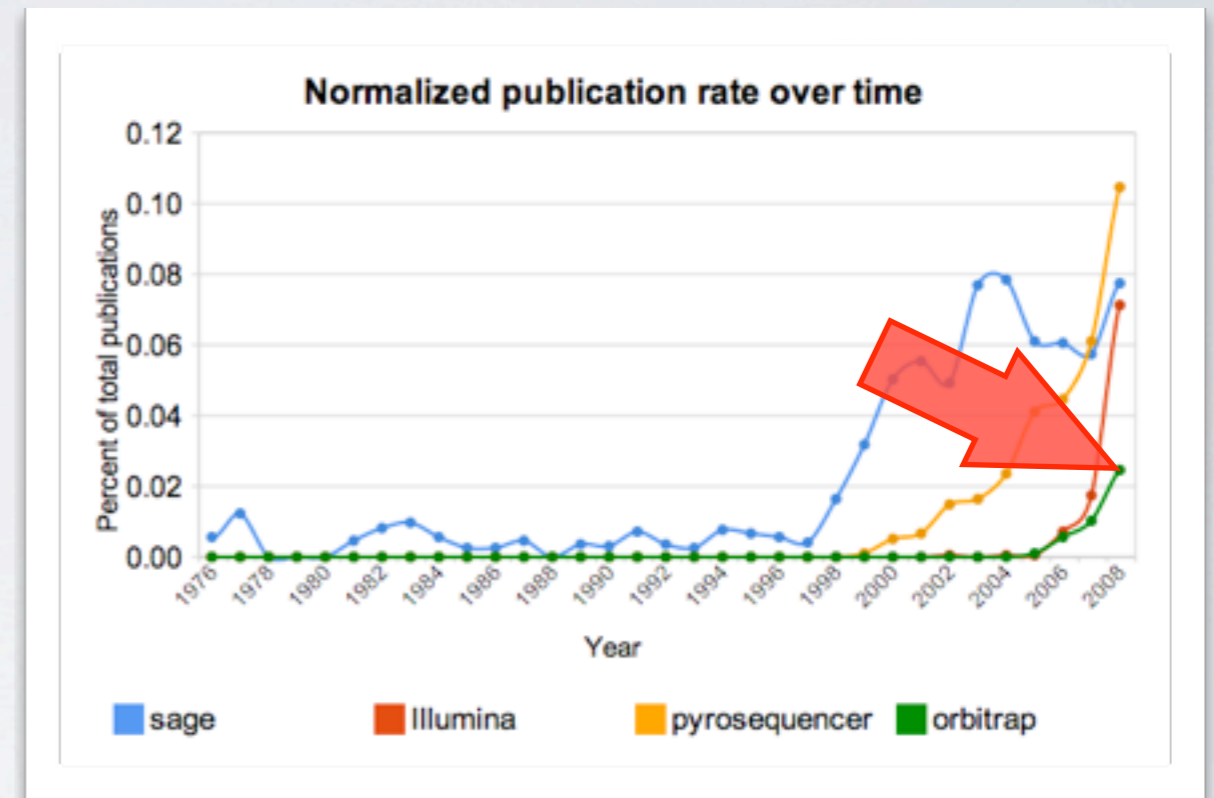
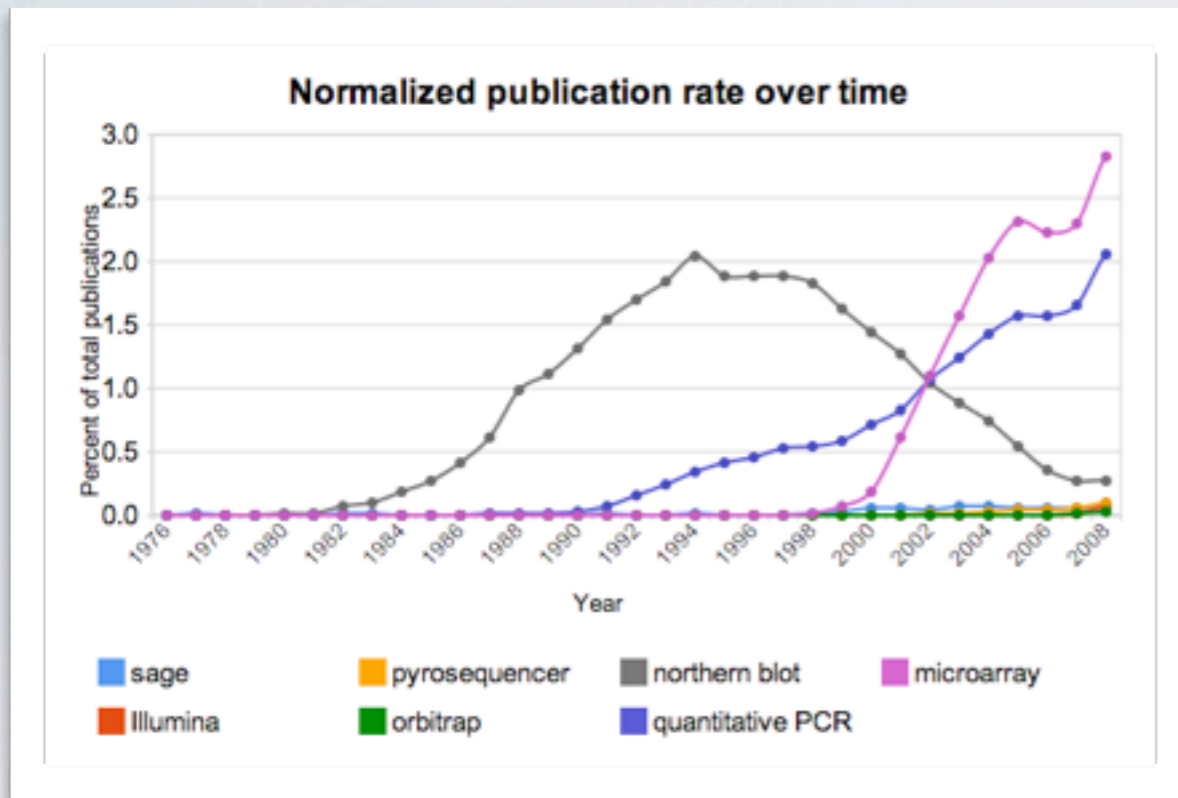
LF-MS/MS proteomics experiments currently rely on **curated protein reference databases** for peptide assignment

such reliance limits its effectiveness as a discovery tool

from our data, a typical rate of peptide **assignment ~30%**

Bitton et al. (2010) used an *in-silico* **translation of the entire human genome** & identified 346 'putative' novel peptides

# future





# integrated RNAseq / proteomics

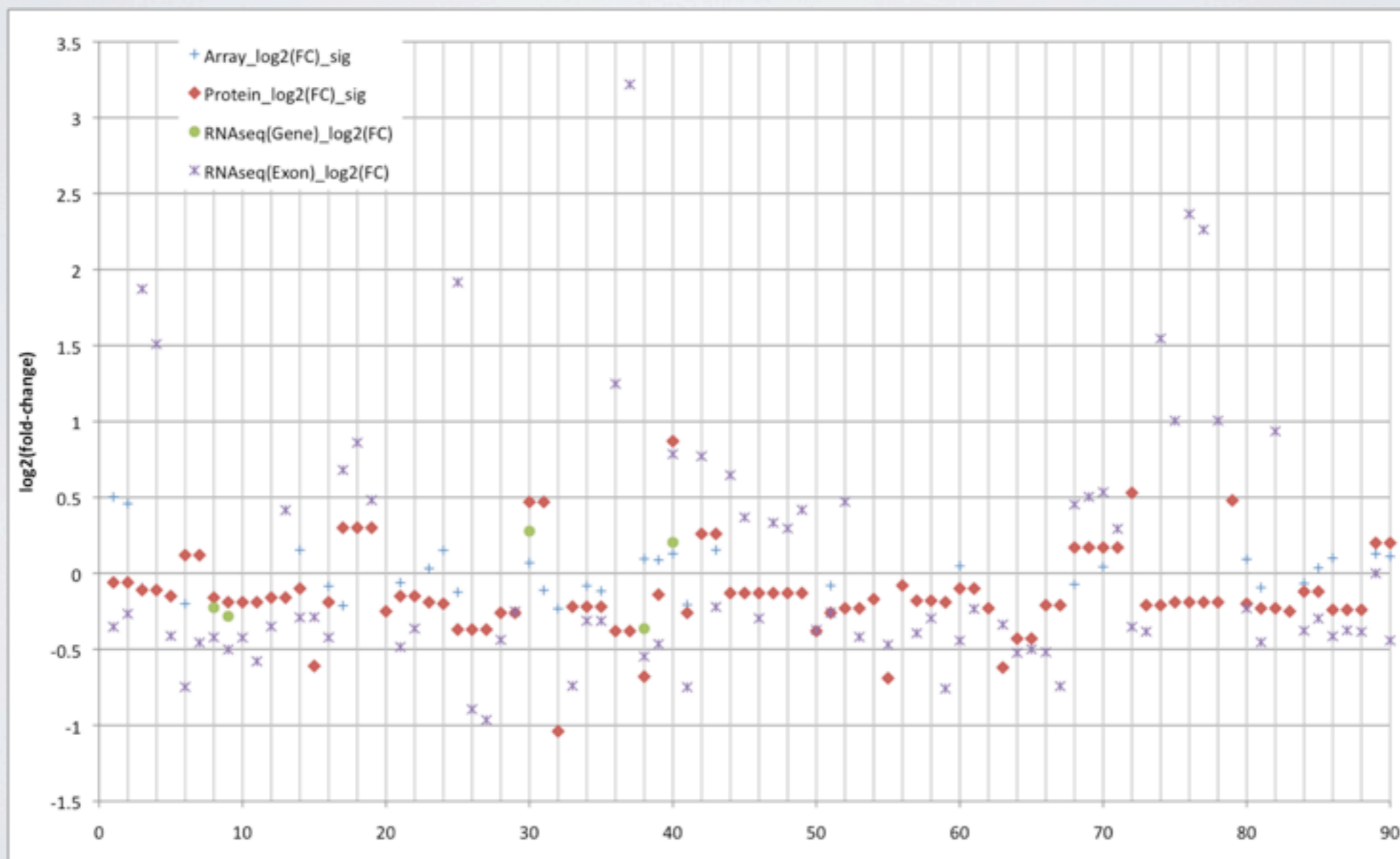
RNAseq data lends itself perfectly as a source of protein reference

perform *in-silico* translation of RNA known to be present in a given set of samples -- map peptides to this reference

RNA/protein **expression analysed simultaneously**, reduce false-positives etc.

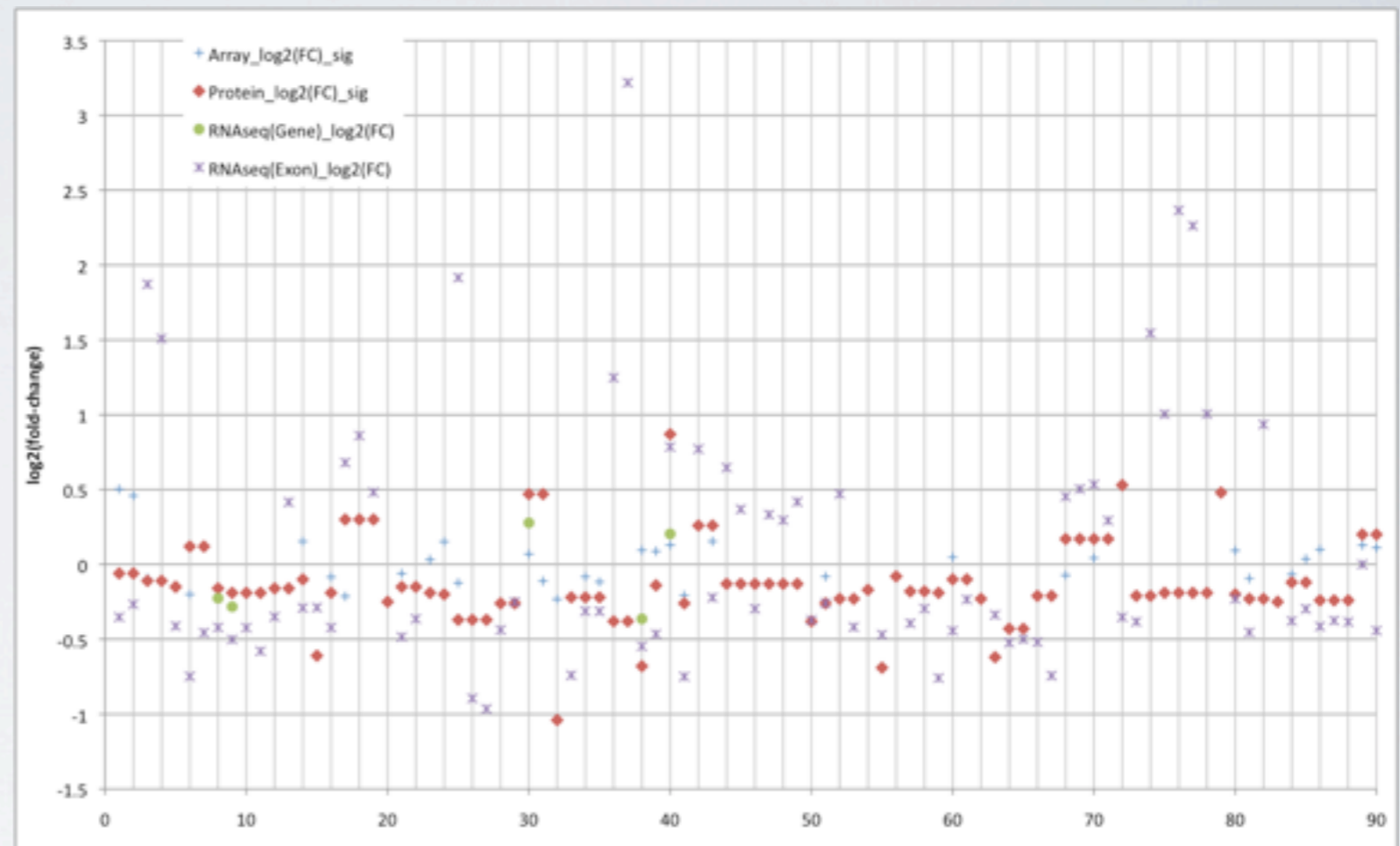
# integrated RNAseq / proteomics

our own, very preliminary, very naïve analysis shows promise:





# integrated RNAseq / proteomics



compared to protein  
direction of change:

53% of array probes agree ( $\chi^2$  pVal=0.886)

71% of exon-level RNA-seq agree ( $\chi^2$  pVal=0.012)

100% of gene-level RNA-seq agree (only 5 of them...)

# acknowledgements

## Edinburgh

Jano van Hemert  
Andrew Sims  
Varrie Ogilvie  
Peter Clarke

Malcolm Atkinson  
John Bartlett  
Vicky Sabine  
Jeremy Thomas  
David Porteous

## Other

Michael Pfaffl  
Ales Tichopad  
Jo Vandesompele  
Mikael Kubista  
Anders Ståhlberg

## Yale

Arthur Simen  
Becky Carlyle  
Kelly Bordner  
Libby George  
everyone else in the Simen lab

Angus Nairn  
TuKiet Lam  
Christopher Colangelo  
Nicholas Carriero  
Robert Bjornson

[rob.kitchen@ed.ac.uk](mailto:rob.kitchen@ed.ac.uk)  
[rob.kitchen@yale.edu](mailto:rob.kitchen@yale.edu)  
linkedin : rkitchen