Improved precision and accuracy for microarrays using updated probe set definitions

Special Topics in Bioinformatics, SS 2010

Andrea Salfinger, 0455957

Report about ...

BMC Bioinformatics

BioMed Central

Correspondence

Open Access

Improved precision and accuracy for microarrays using updated probe set definitions

Rickard Sandberg[1] and Ola Larsson*[2]

Address: [1]Massachusetts Institute of Technology, Department of biology, 68-211, Cambridge, MA 02139, USA and [2]University of Minnesota, Department of Medicine, MMC 276, Minneapolis, MN 55455, USA

Email: Rickard Sandberg - sandberg@mit.edu; Ola Larsson* - larss004@tc.umn.edu

* Corresponding author

Published: 8 February 2007

BMC Bioinformatics 2007, 8:48 doi:10.1186/1471-2105-8-48

Received: 26 September 2006

Accepted: 8 February 2007

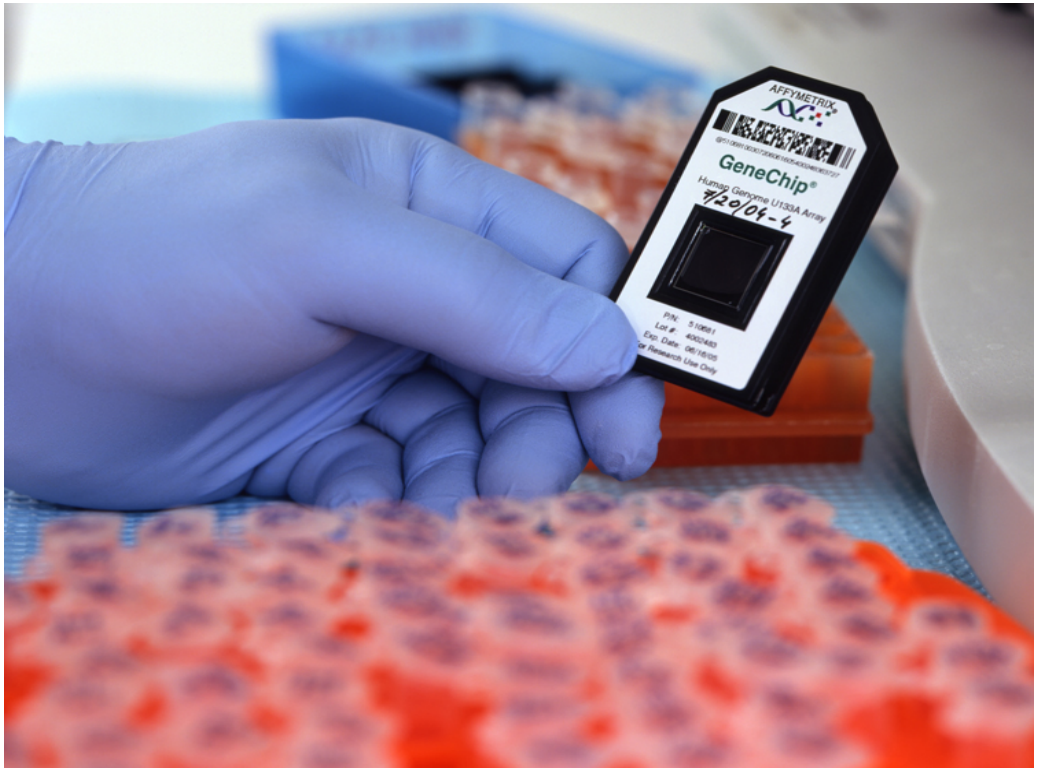
This article is available from: http://www.biomedcentral.com/1471-2105/8/48

© 2007 Sandberg and Larsson; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Outline

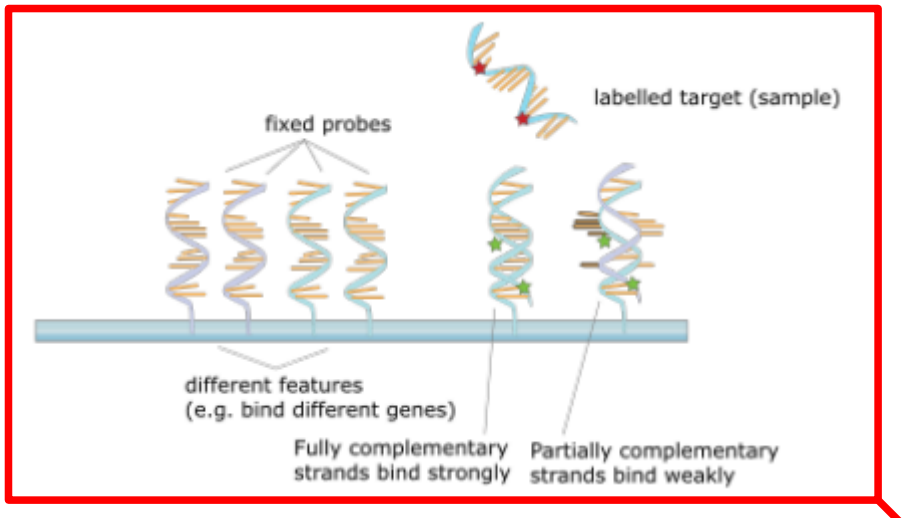
- Recap: Microarrays
- State of the art: Problems
- Updated Probe Set Definitions
- Motivation
- Purpose of this study
- Study Design
- Methods
- Experimental Data
- Precision
- Accuracy
- Results
- Summary

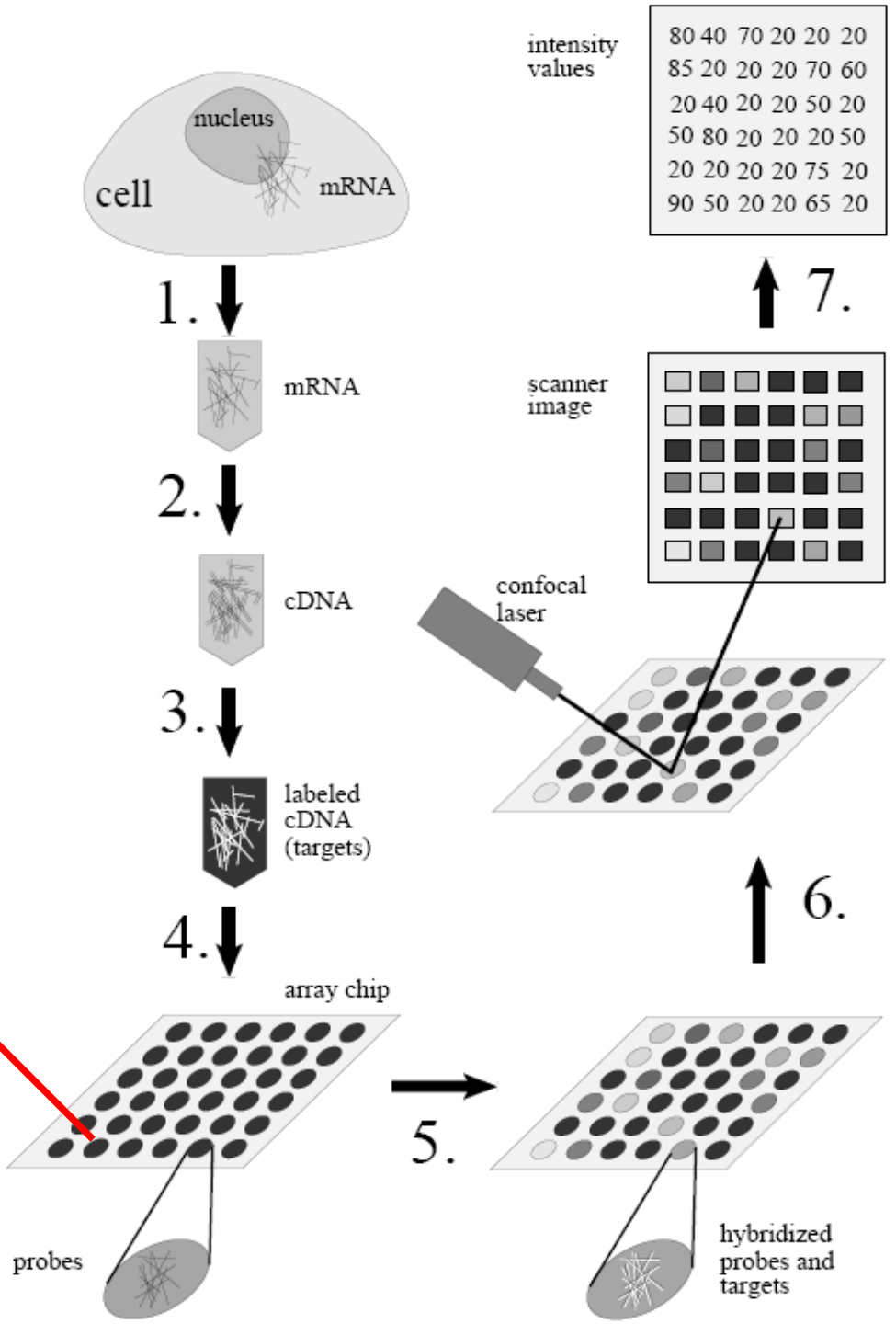


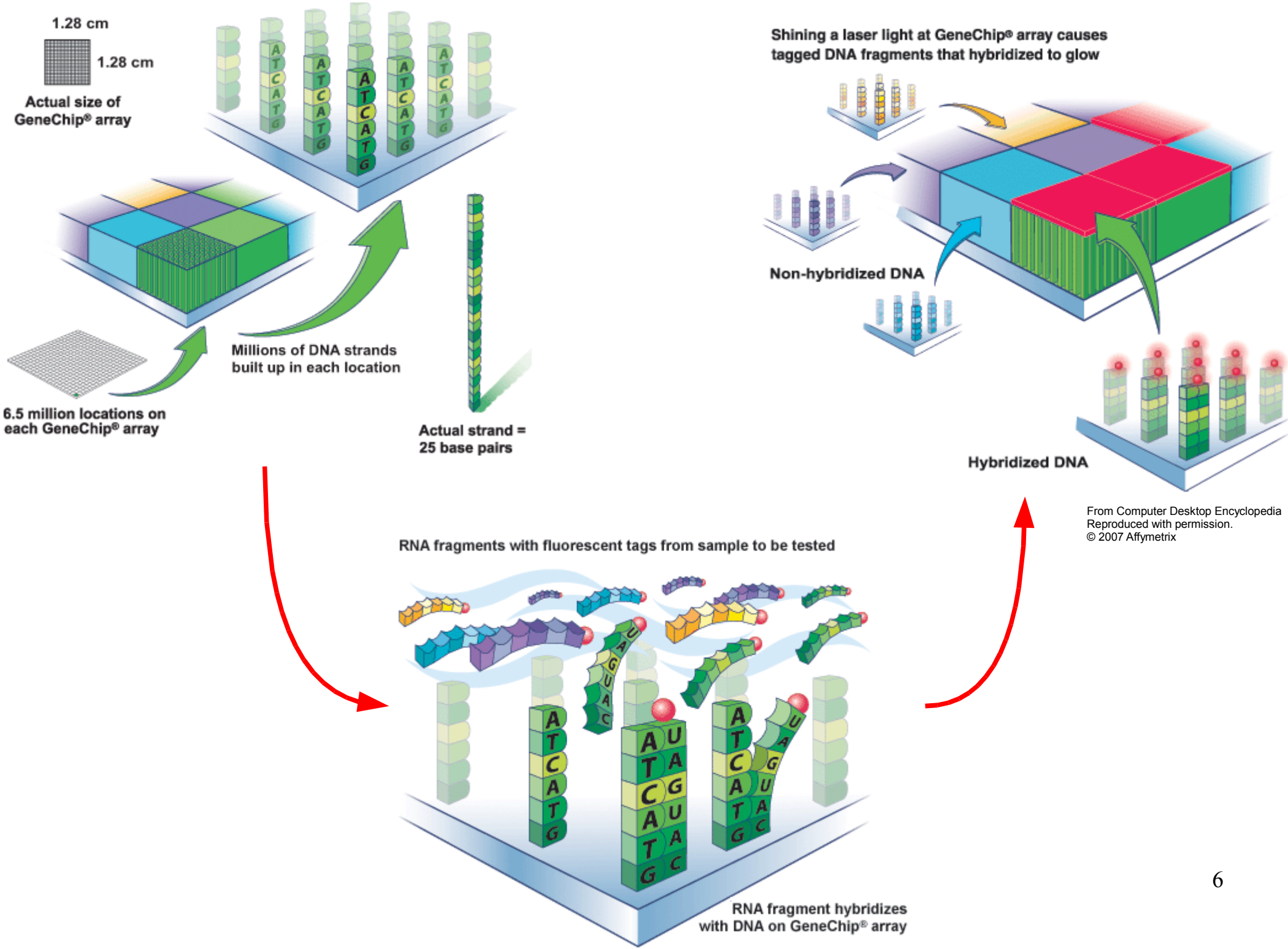
Recap: Microarrays (1)

- When?
  - introduced in 1995
- What?
  - measure extracted mRNA transcripts of a cell
- How?
  - hybridize the amplified and labeled transcripts to template strands on a solid surface (microarray), measure which and how many transcripts have bound
- Why?
  - to measure gene expression levels of a cell

- How it works:

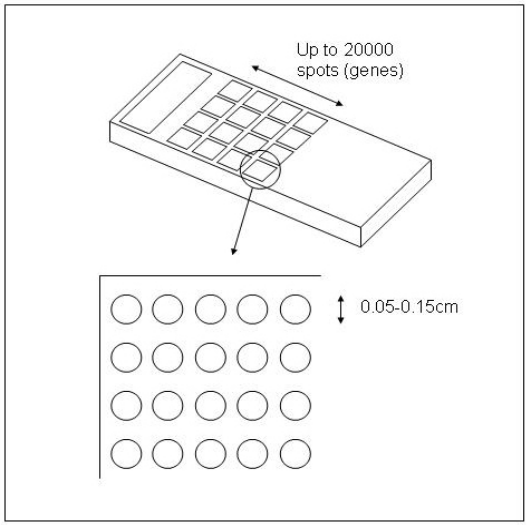

labelled target (sample)
fixed probes
different features
(e.g. bind different genes)
Fully complementary
strands bind strongly
Partially complementary
strands bind weakly

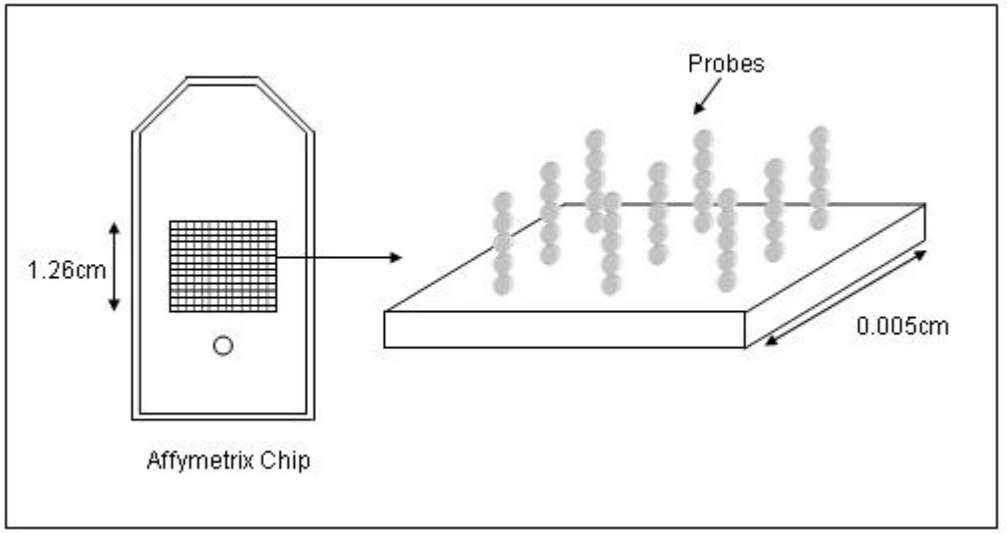

intensity
values
80 40 70 20 20 20
85 20 20 20 70 60
20 40 20 20 50 20
50 80 20 20 20 50
20 20 20 20 75 20
90 50 20 20 65 20
7.
scanner
image
confocal
laser
nucleus
cell
mRNA
1.
mRNA
2.
cDNA
3.
labeled
cDNA
(targets)
4.
array chip
6.
5.
probes
hybridized
probes and
targets

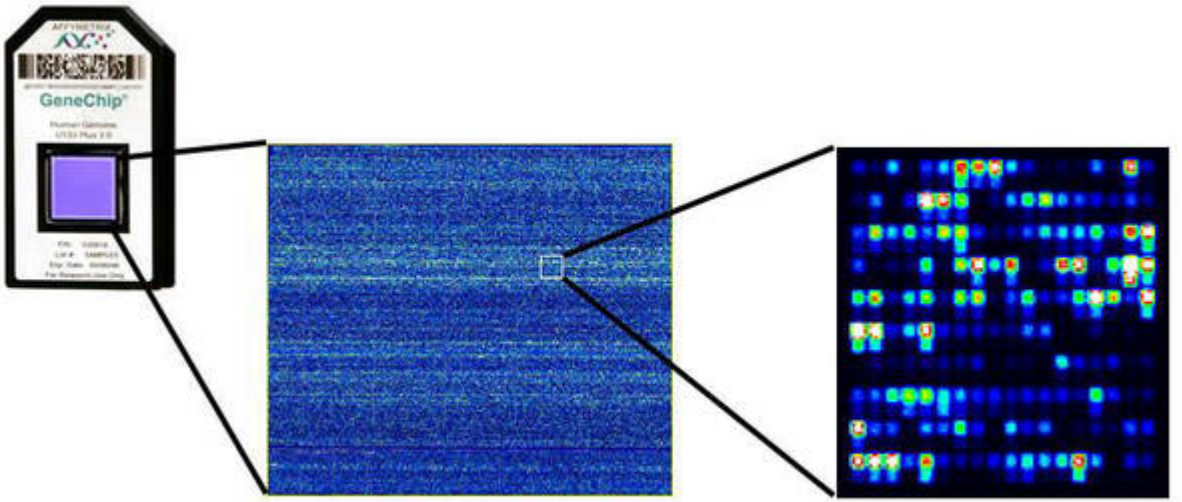


From Computer Desktop Encyclopedia
Reproduced with permission.
© 2007 Affymetrix

Recap: Microarrays (4)

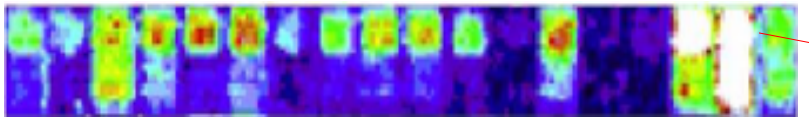






Recap: Microarrays (5)

- What is measured?

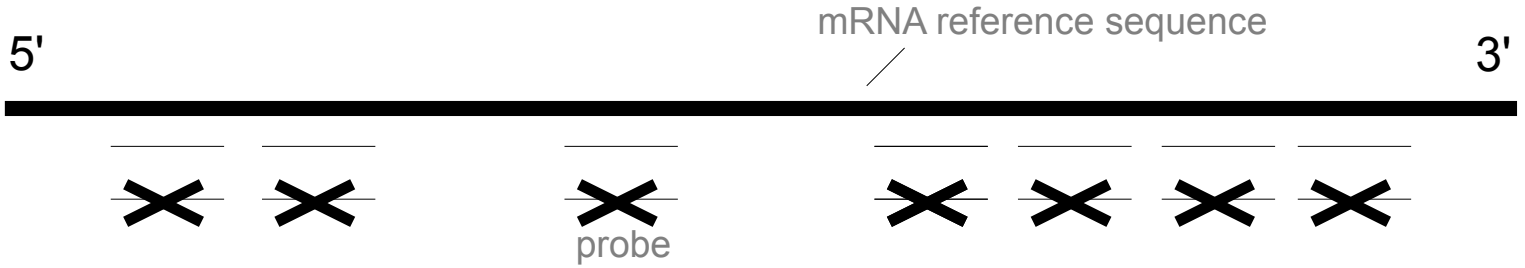


DNA fragments of
25 bp length

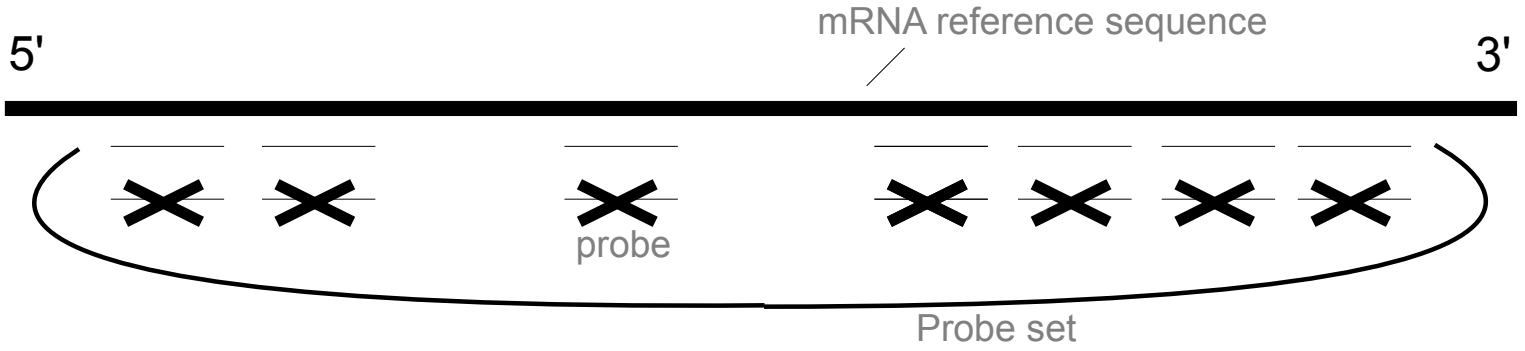
- How to get gene expression level from this, when only DNA fragments of 25 bp length are measured?

Recap: Microarrays (6)

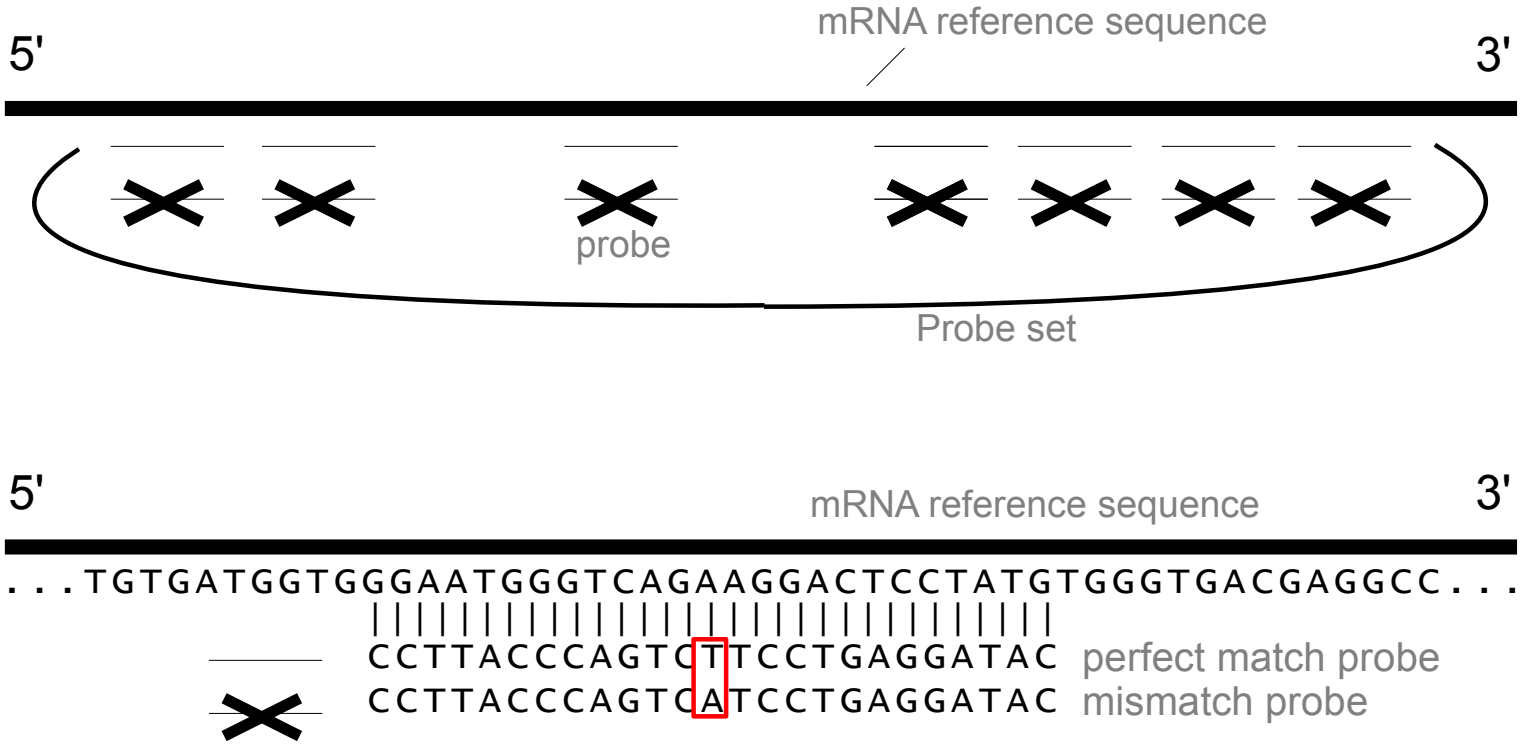
Recap: Microarrays (6)



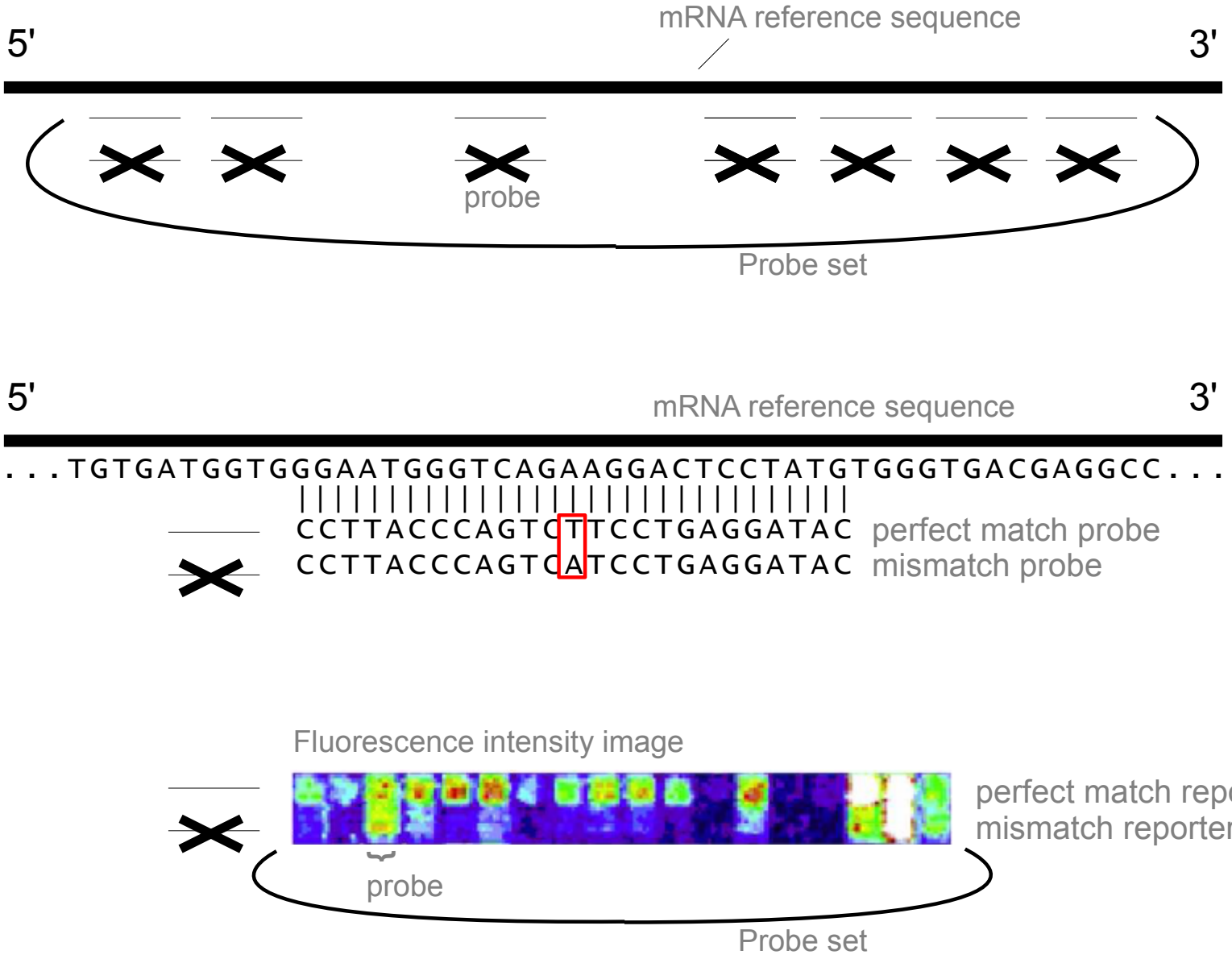
Recap: Microarrays (6)



Recap: Microarrays (6)



Recap: Microarrays (6)



State of the art: Problems (1)

- Affymetrix platforms currently in use were designed before genomes were fully sequenced
  - many probes were designed after consensus sequence of clusters of Expressed Sequence Tags (ESTs)
- Problem: design of probes based on (meanwhile) outdated genomic knowledge!

  (we know it better now)

State of the art: Problems (2)

- Problem:
  - accurate probe set definitions essential for integrating probe signals into expression levels!
  - many old Affymetrix chips still in use in research community
  - it takes a lot of time to develop new chips
  - What to do with the studies that have been carried out with older chip-generations?
- Solution:
  - analyze the data using updated probe set definitions

Updated Probe Set Definitions

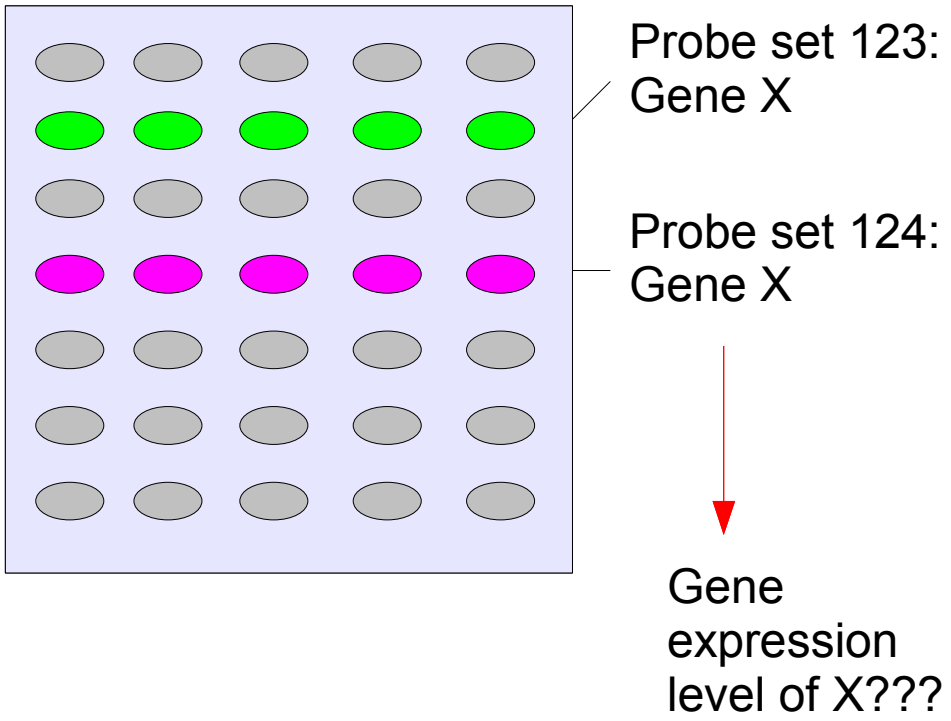
= reannotation of the the existing probes on Affymetrix platforms

- reannotation better reflects transcript information and gene annotations available today

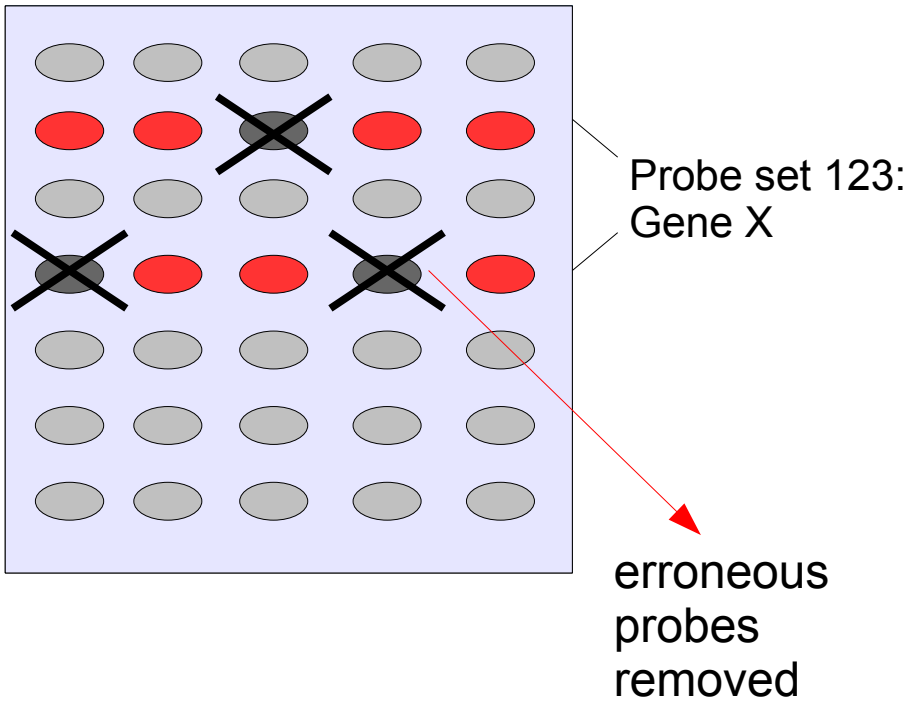
- Updated probe set definitions map probes to transcript annotations, e.g.
  - EnsEMBL transcripts
  - Refseq
  - Entrez GeneID

Updated Probe Set Definitions

original probe set definitions



updated probe set definitions



Motivation (1)

- original Affymetrix probe set definitions:
  - many probe sets map to the same gene
    - integrative microarray studies use ad-hoc heuristics to integrate these values into a single expression estimate
- updated probe set definitions
  - erroneous or non-specific probes are removed
  - probe sets targeting the same gene/transcript are pooled
    - fewer probe sets compared with the original
    - number of probe pairs per probe set is no longer identical
      - better/worse ?!

Motivation (2)

- Studies showed:
  - Updated probe set definitions affect approximately 20-30% of all probe sets → affect a large portion of gene estimates!
  - Genes identified as differentially expressed using original and updated probe set definitions only show 50% overlap!
  - Improvement of cross-platform reproducibility of microarray experiments when using updated probe set defintions

Purpose of this study

- Such updated probe set definitions are now available
  - can easily be integrated into bioconductor packages (*affy*, *gcrma*)
  - map the platform probe signals to genes, transcripts and even exon expression levels
- Aim of this study
  - evaluate the impact of updated probe set definitons on precision and accuracy in estimated expression levels

Study Design

- What?
  - Re-analysis of raw data using updated probe set definitions
  - Comparison of results with original Affymetrix probe set definitions
- Why?
  - Does usage of updated probe set definitions improve precision and accuracy of results?
    - Hypothesis: variable number of probe pairs integrated into each probe set → might have negative impact on precision? (fewer probes in some probe sets)

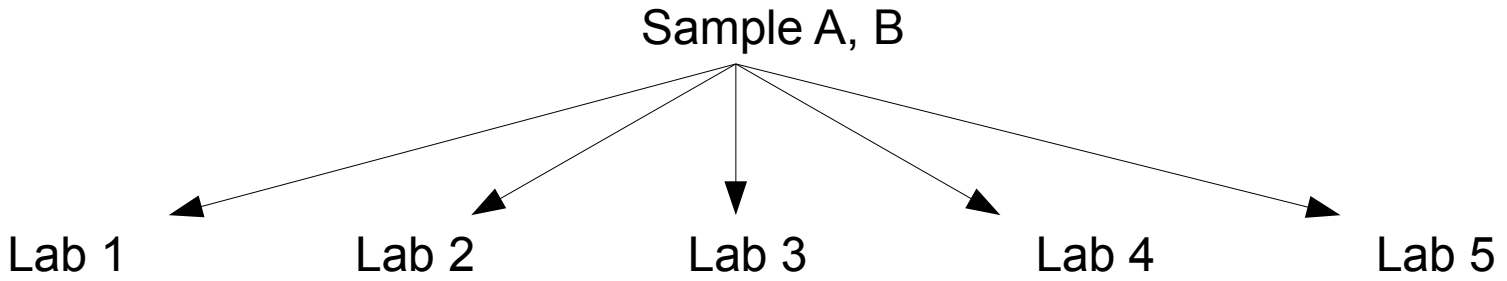
Methods (1)

- How?
  - Re-analysis of a gene expression data set using
    - the original NetAffx probe set definitions
    - 6 updated probe set definitions (custom CDFs)
  - Dataset:
    - 2 RNA samples which differ in expression of only a few genes
    - both samples were hybridized to 2 HG-U133A Affymetrix arrays each by 5 different labs

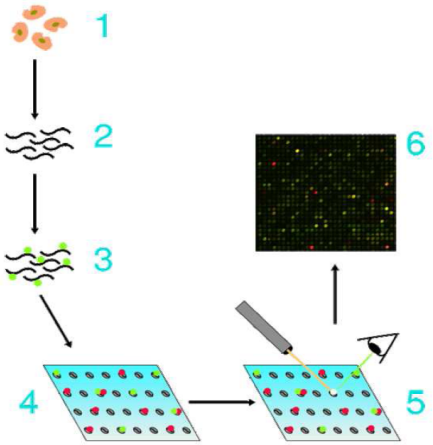
Methods (2)

Sample A, B

Methods (2)



Methods (2)

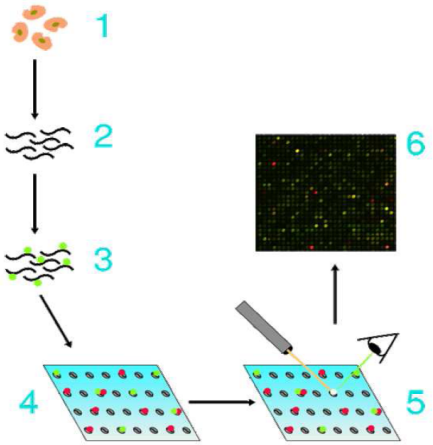


Sample A, B

Hybridized to
2 arrays
(HG-U133A)

Lab 1 Lab 2 Lab 3 Lab 4 Lab 5

Methods (2)

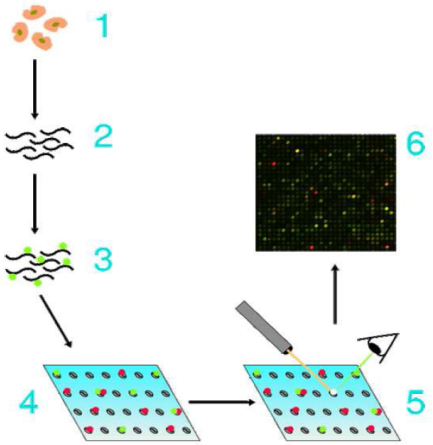


Sample A, B

Hybridized to
2 arrays
(HG-U133A)

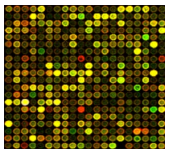
Lab 1	Lab 2	Lab 3	Lab 4	Lab 5
A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2

Methods (2)

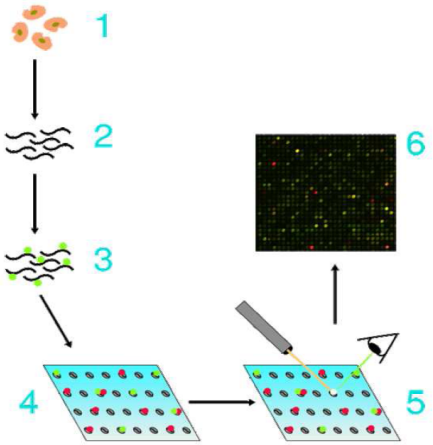


Sample A, B

Hybridized to
2 arrays
(HG-U133A)

	Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	
	A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	Raw data files (CEL)

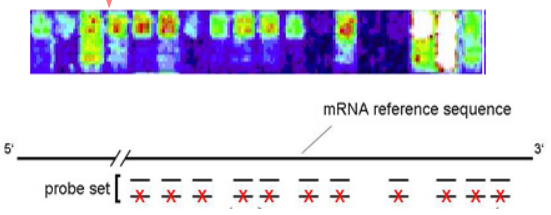
Methods (2)



Sample A, B

Hybridized to
2 arrays
(HG-U133A)

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	
A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	A1, A2 B1, B2	Raw data files (CEL)



Apply probe set definition

ensEMBL exons
NetAffx

ensEMBL transcripts

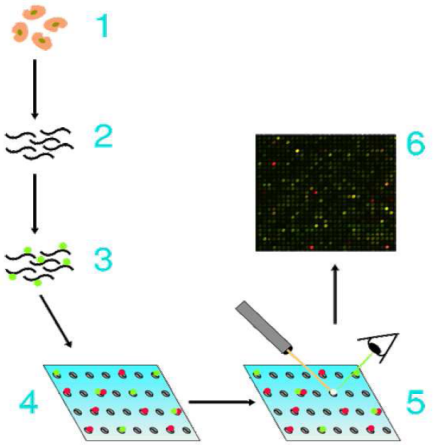
ensEMBL genes

Entrez Gene IDs

UniGene

RefSeq

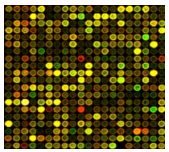
Methods (2)



Sample A, B

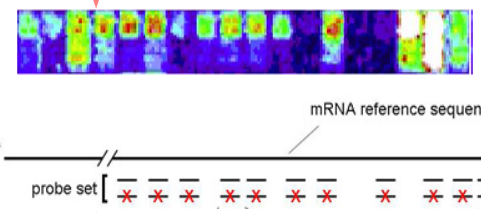
Hybridized to 2 arrays (HG-U133A)

Lab 1 | Lab 2 | Lab 3 | Lab 4 | Lab 5



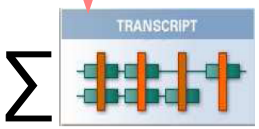
A1, A2 B1, B2 (Lab 1)
A1, A2 B1, B2 (Lab 2)
A1, A2 B1, B2 (Lab 3)
A1, A2 B1, B2 (Lab 4)
A1, A2 B1, B2 (Lab 5)

Raw data files (CEL)



Apply probe set definition

ensEMBL exons
NetAffx
ensEMBL transcripts
ensEMBL genes
Entrez Gene IDs
RefSeq
UniGene

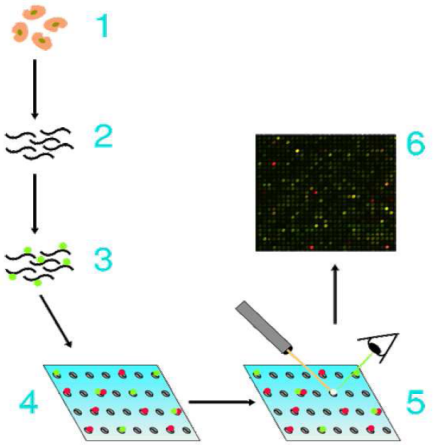


Transcript X = 80

Calculate probe set summaries

MAS5
RMA
GCRMA

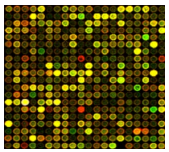
Methods (2)



Sample A, B

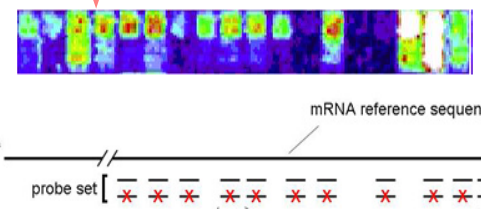
Hybridized to 2 arrays (HG-U133A)

Lab 1 | Lab 2 | Lab 3 | Lab 4 | Lab 5



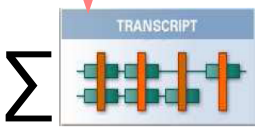
A1, A2 B1, B2 (Lab 1)
A1, A2 B1, B2 (Lab 2)
A1, A2 B1, B2 (Lab 3)
A1, A2 B1, B2 (Lab 4)
A1, A2 B1, B2 (Lab 5)

Raw data files (CEL)



Apply probe set definition

ensEMBL exons
NetAffx
ensEMBL transcripts
ensEMBL genes
Entrez Gene IDs
RefSeq
UniGene

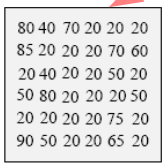


Transcript X = 80

MAS5
RMA
GCRMA

Calculate probe set summaries

21 probe set estimates per array!!!



A1/B1 21x A2/B2 21x
A1/B1 21x A2/B2 21x
A1/B1 21x A2/B2 21x
A1/B1 21x A2/B2 21x
A1/B1 21x A2/B2 21x

Experimental Data

- 4 arrays per lab
  - 2x sample A (A1, A2)
  - 2x sample B (B1, B2)
  - computed expression values using 7 different probe set definitions (default vs. updated) and 3 different probe set summarization algorithms
- Within each lab:
  - pairs of replicates used to estimate precision and accuracy
- 5 labs performed identical experiment:
  - estimated precision and accuracy obtained in each lab can be summarized to provide a robust assessment of the effects

Precision (1)

- Measures data reproducibility and variability
- Precision = correlation between the relative log2 expression ratios of the 2 RNA samples using the 2 pairs of replicate pairs:
  - A1/B1 vs. A2/B2
- Clear indication of experiment performance
  - Correlation of 1 → perfect precision
  - Correlation of 0 → no precision

Precision (2)

- Comparison of precision between original probe set definitions and updated probe set definitions:

Table 1: Improved precision using update probe set definitions

	ensEMBL exon	ensEMBL gene	ensEMBL transcript_	Entrez	RefSeq	UniGene
MAS5	-0.014 p = 0.094	0.013 p = 0.057	0.013 p = 0.0092	0.014 p = 0.056	0.019 p = 0.018	0.023 p = 0.052
RMA	-0.035 p = 0.0041	0.053 p = 0.00025	0.030 p = 0.0071	0.059 p = 0.00011	0.045 p = 0.00045	0.047 p = 2.9E-06
GCRMA	-0.11 p = 0.000051	0.023 p = 0.045	-0.0063 p = 0.28	0.040 p = 0.019	0.031 p = 0.062	0.028 p = 0.0071

Precision better for all updated probe set definitions but exons → why?

Mean precision difference for updated probe set definition compared with the original probe set

Significance of difference (2-tailed paired t-test)

Precision (3)

- Possible reason for decrease in precision for probe set definition to ensEMBLE exons?
  - mean number of probes per probe set is lower for ensEMBL exons
  - using fewer probes when estimating an expression level likely increases variance and lowers precision

Precision (4)

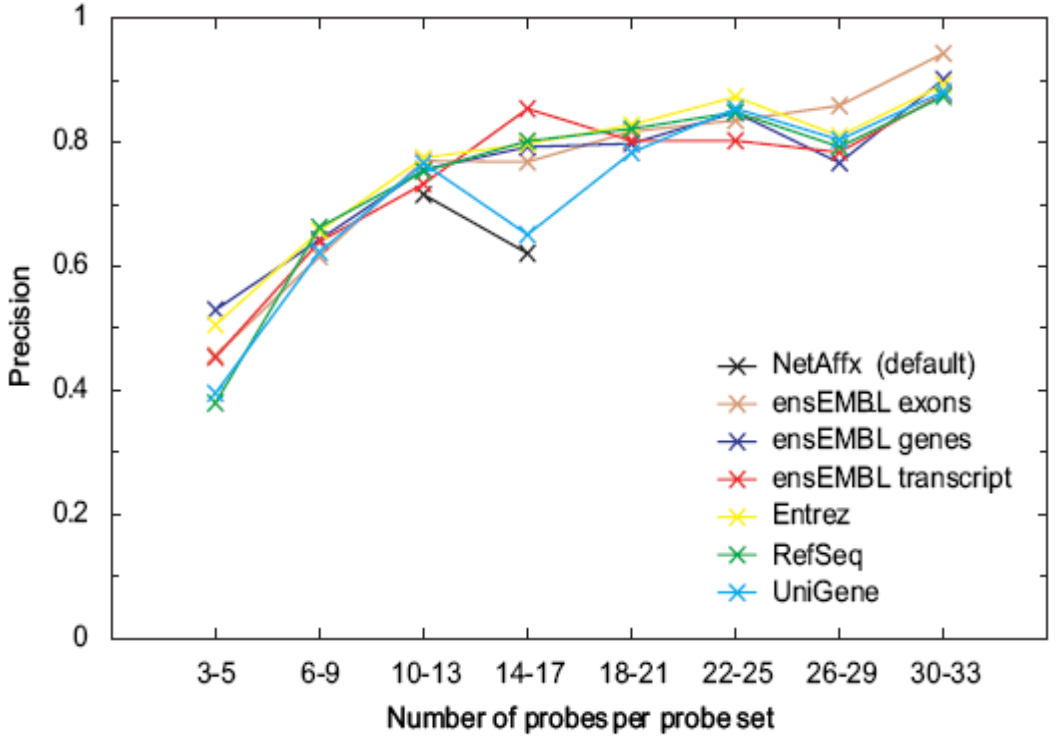
- Improved precision for other updated probe set definitions could be due to larger number of probes mapping to each probe set
  - mean number of probes higher than for original probe set definition

Table 2: Characteristics of probe set definitions

Probe set definition	Number of probe sets	Mean number of probe pairs per probe set
NetAffx (original)	22283	11.1
ensEMBL exon	35191	9.3
ensEMBL gene	18671	14.0
ensEMBL transcript	36174	13.9
Entrez	12132	14.1
RefSeq	17880	14.9
UniGene	11694	15.0

Precision (5)

- Precision was analyzed as a function of the number of probes used to estimate each probe set
  - Positive correlation between the number of probes per probe set and precision found



Precision (6)

- But:
  - Updated probe set definitions appear to achieve better precision, even when similar number of probes were integrated into signal estimates!

→ Updated probe set definitions have significant improvements in precision!

  - removal of erroneous or non-specific probes that otherwise add noise

Accuracy (1)

- Accuracy estimates how close microarry estimates are to the „real expression“ changes
  - most often „real“ expression is measured using RT-PCR (real time PCR)
- To assess accuracy the updated probe set definitions achieved:
  - Differential expression detected with microarrays was compared to those measured with RT-PCR for 16 genes (for the different probe set definitions)

Accuracy (2)

- Accuracy was defined as the slope after a linear regression between RT-PCR and microarry data
  - An accuracy of 1.0 is optimal
- Difference in accuracy for each lab between the updated probe set definitions and the standard probe set definition was calculated
  - Test: is mean accuracy difference (averaged across the 5 labs) significant?
    - Using paired t-test (2-tailed distribution)

Accuracy (3)

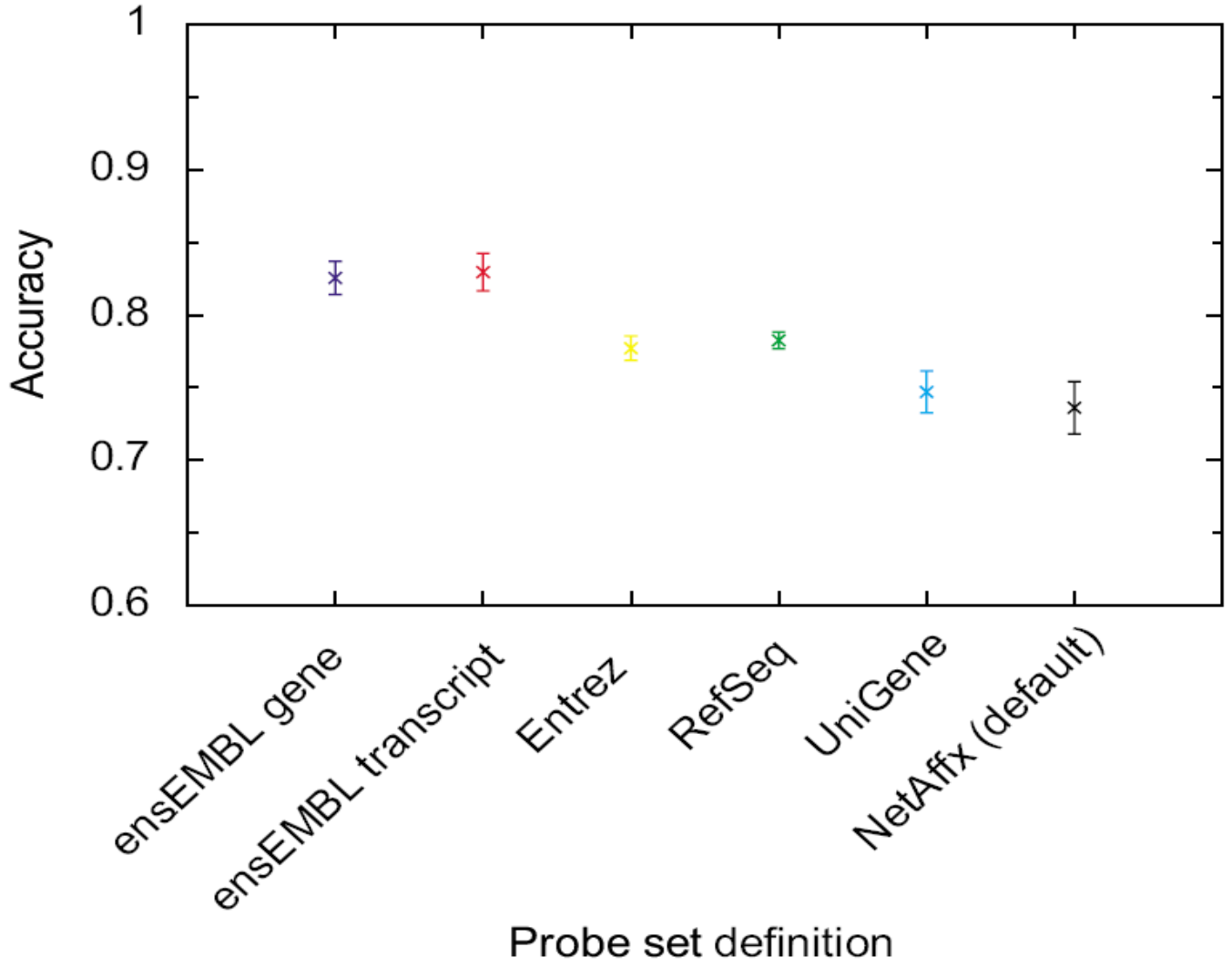
- Significant improvements in accuracy were observed (when data was normalized using RMA) for all but UniGene definition

Table 3: Improved accuracy using updated probe set definitions

RMA	Mean Slope	p-value	Std
NetAffx (original)	0.74		0.02
ensEMBL gene	0.83	0.00040	0.01
ensEMBL transcript	0.83	0.00053	0.01
Entrez	0.78	0.00430	0.01
RefSeq	0.78	0.00381	0.01
UniGene	0.75	0.07085	0.01

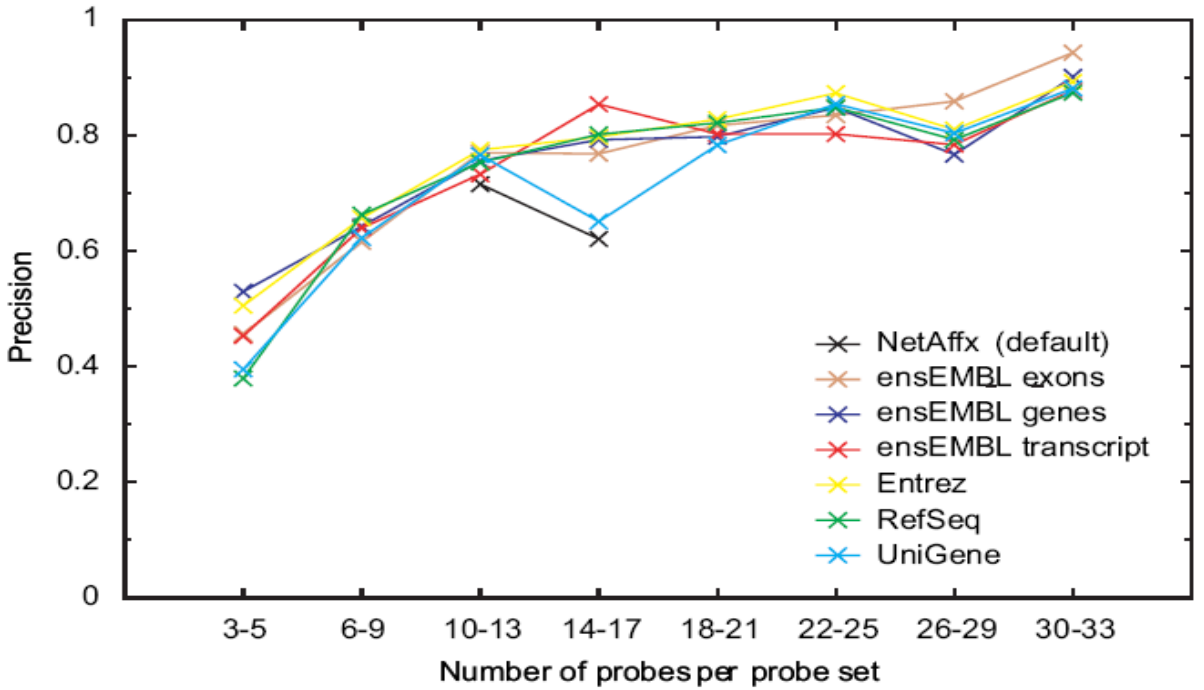
The slopes estimated from the 5 different labs are in good agreement.

Accuracy (4)

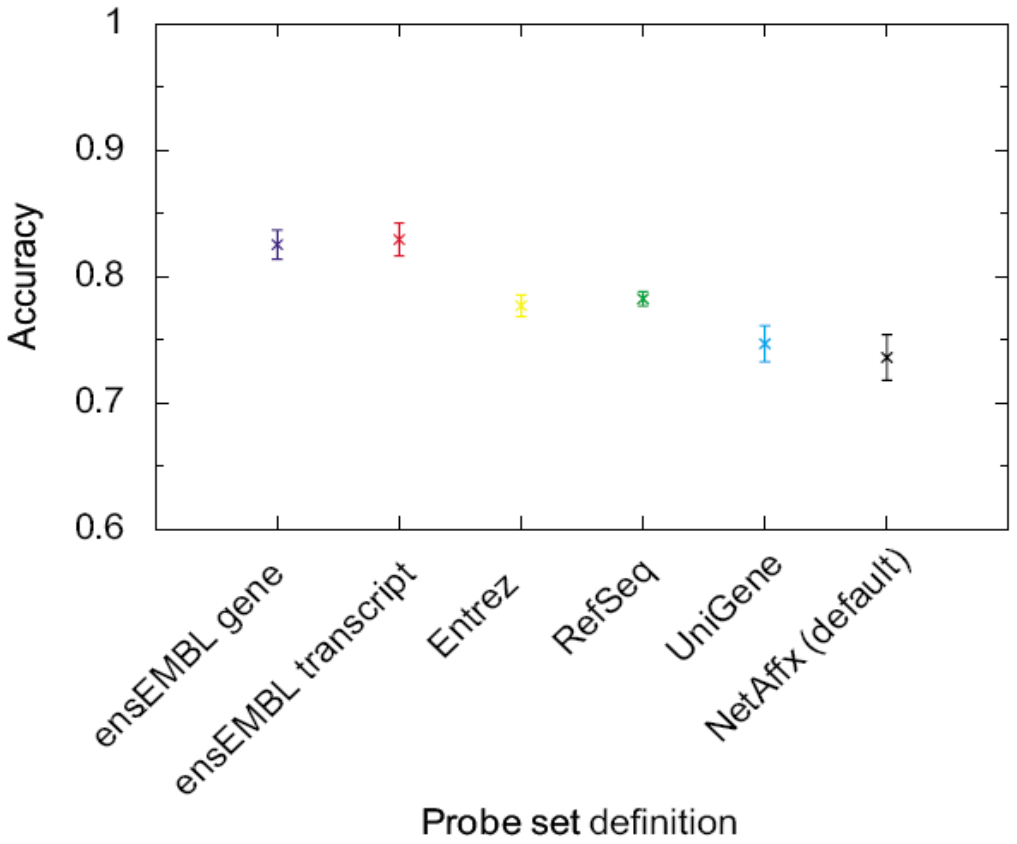

1
0.9
0.8
0.7
0.6
Accuracy
ensEMBL gene
ensEMBL transcript
Entrez
RefSeq
UniGene
NetAffx (default)
Probe set definition

Results (1)

- Significant improvement in <u>precision</u> using updated probe set definitions!



Results (2)



- Significant improvement in accuracy using updated probe set definitions!

Summary

- Updated probe set definitions offer expression levels that are more accurately associated to genes and transcripts.
  - It could be shown in this study that they also improve the estimated transcript levels:
    - using updated probe set definitions improves both the precision and accuracy of the relative expression level estimates
- The results of this study encourage a wide spread use of updated probe set definitions.