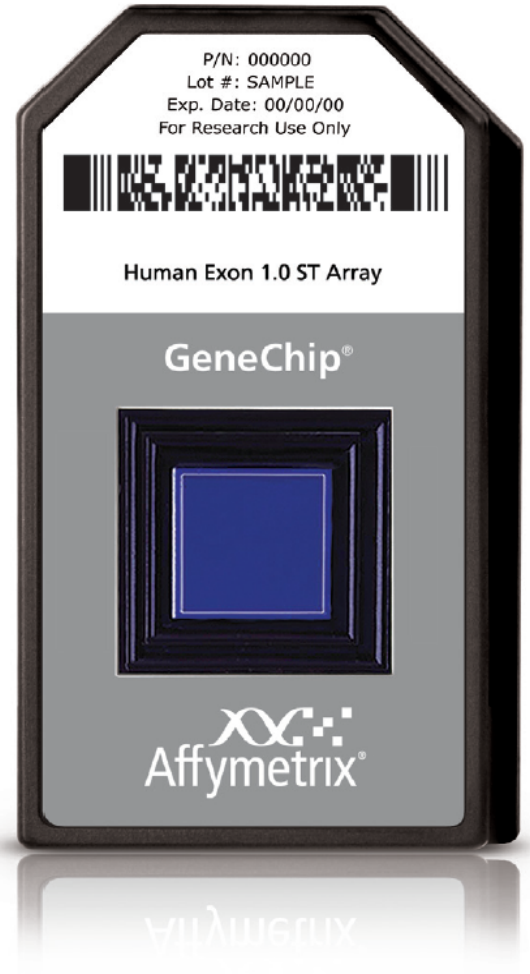


Cross-hybridization modeling on Affymetrix exon arrays

Special Topics in Bioinformatics, SS 2010

Andrea Salfinger, 0455957

Report about ...

BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 24 2008, pages 2887–2893
doi:10.1093/bioinformatics/btn571

Gene expression

Cross-hybridization modeling on Affymetrix exon arrays

Karen Kapur[1], Hui Jiang[2], Yi Xing[3] and Wing Hung Wong[1,4,*]

[1]Department of Statistics, [2]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, [3]Department of Internal Medicine and Department of Biomedical Engineering, University of Iowa, Iowa City, IA and [4]Department of Health Research and Policy, Stanford University, Stanford, CA, USA

Received on July 15, 2008; revised on October 3, 2008; accepted on October 30, 2008

Advance Access publication November 4, 2008

Associate Editor: Trey Ideker

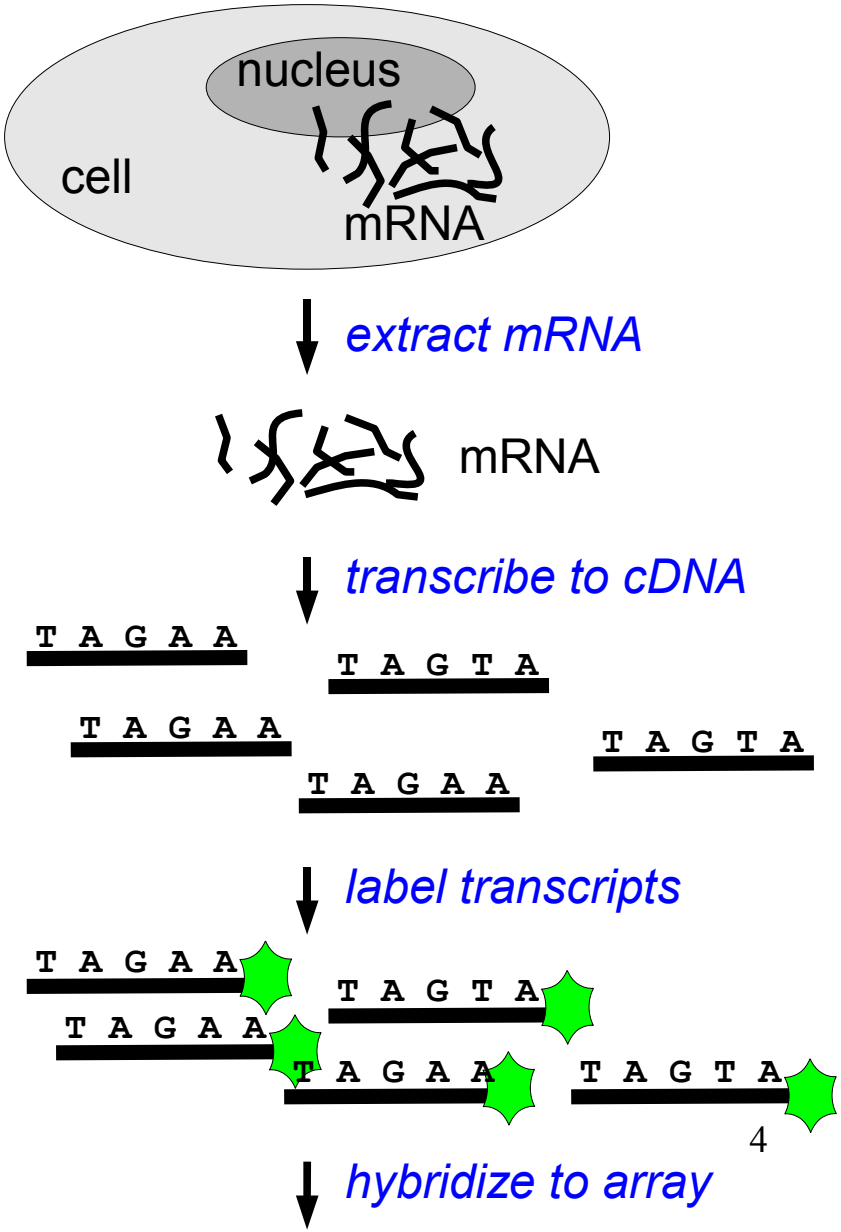
Outline

- Introduction
  - Recap: Microarrays → State of the Art → Motivation → Idea
- Analysis, Methods & Algorithms
  - Algorithm
  - Methods
  - The Cross-Hybridization Model
- Results
  - **GeneBASE-xhyb**
  - Evaluation of GeneBASE-xhyb
  - Comparison to Other Methods
  - Summary

~ 24 slides in total

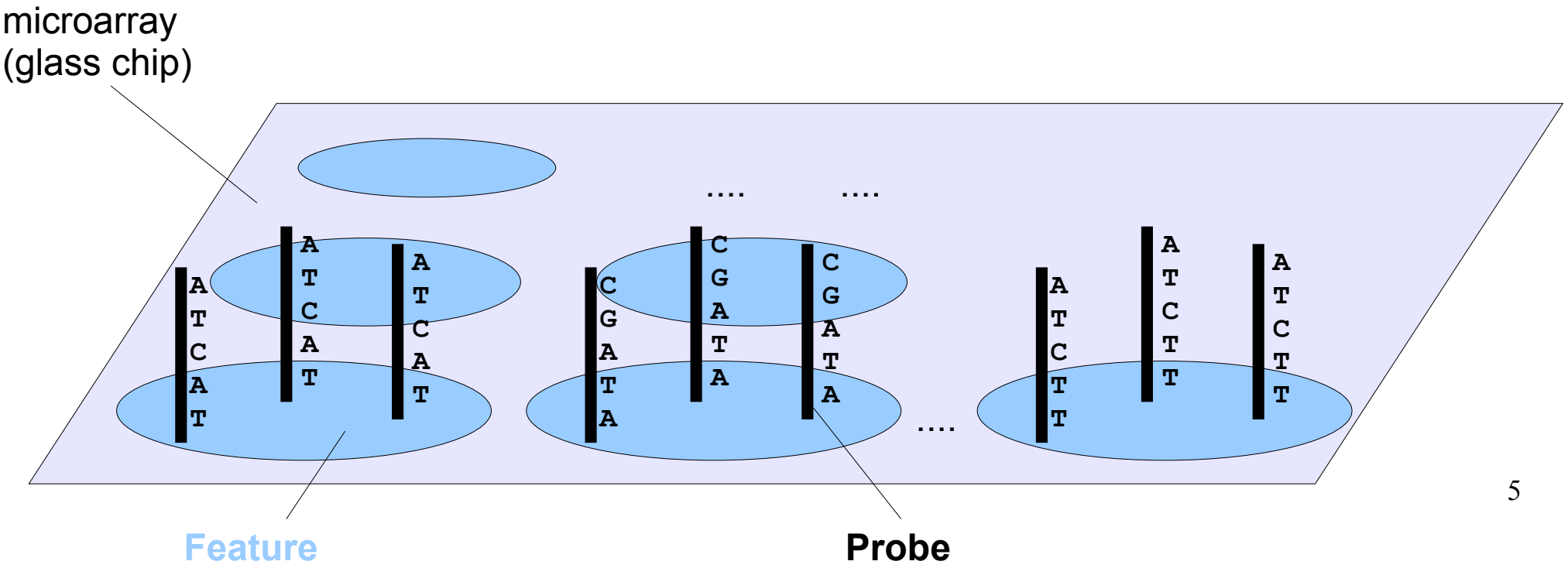
Recap: Microarrays (1)

- 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



Recap: Microarrays (2)

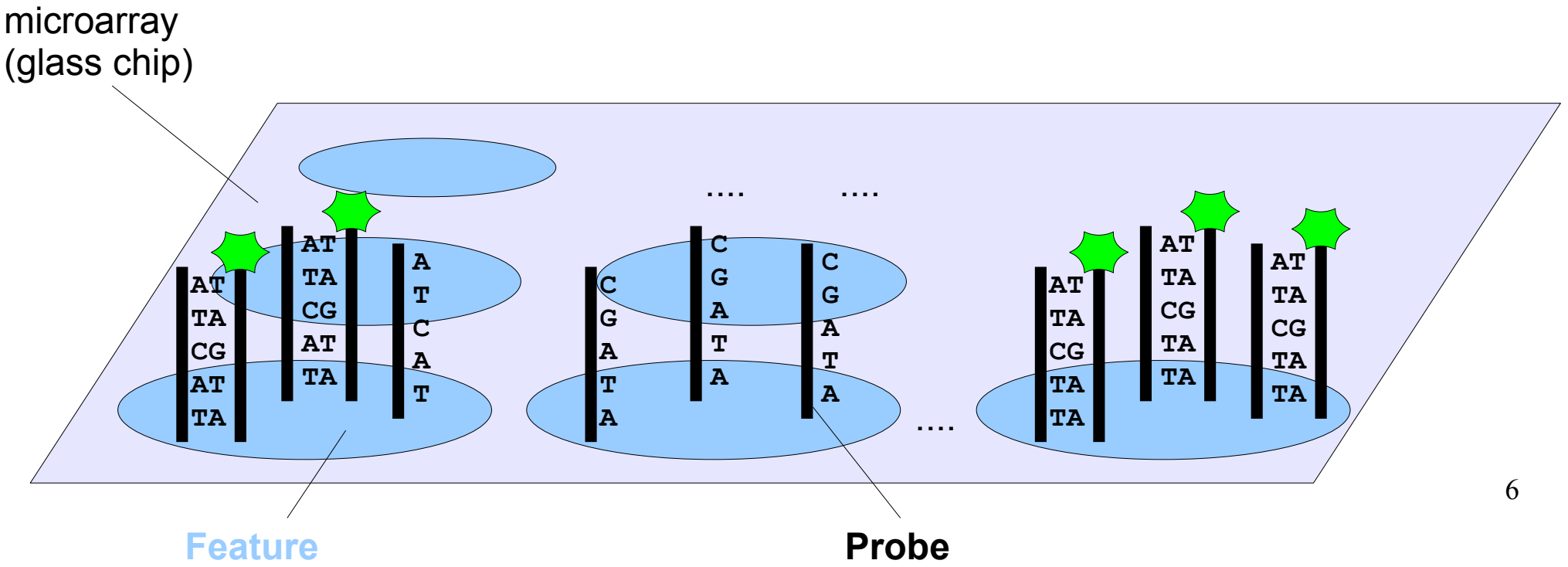
- DNA fragments (= „probes“) are attached to a microarray slides
- each probe has exact nucleotide sequence complementarity to a specific transcript



Recap: Microarrays (2)

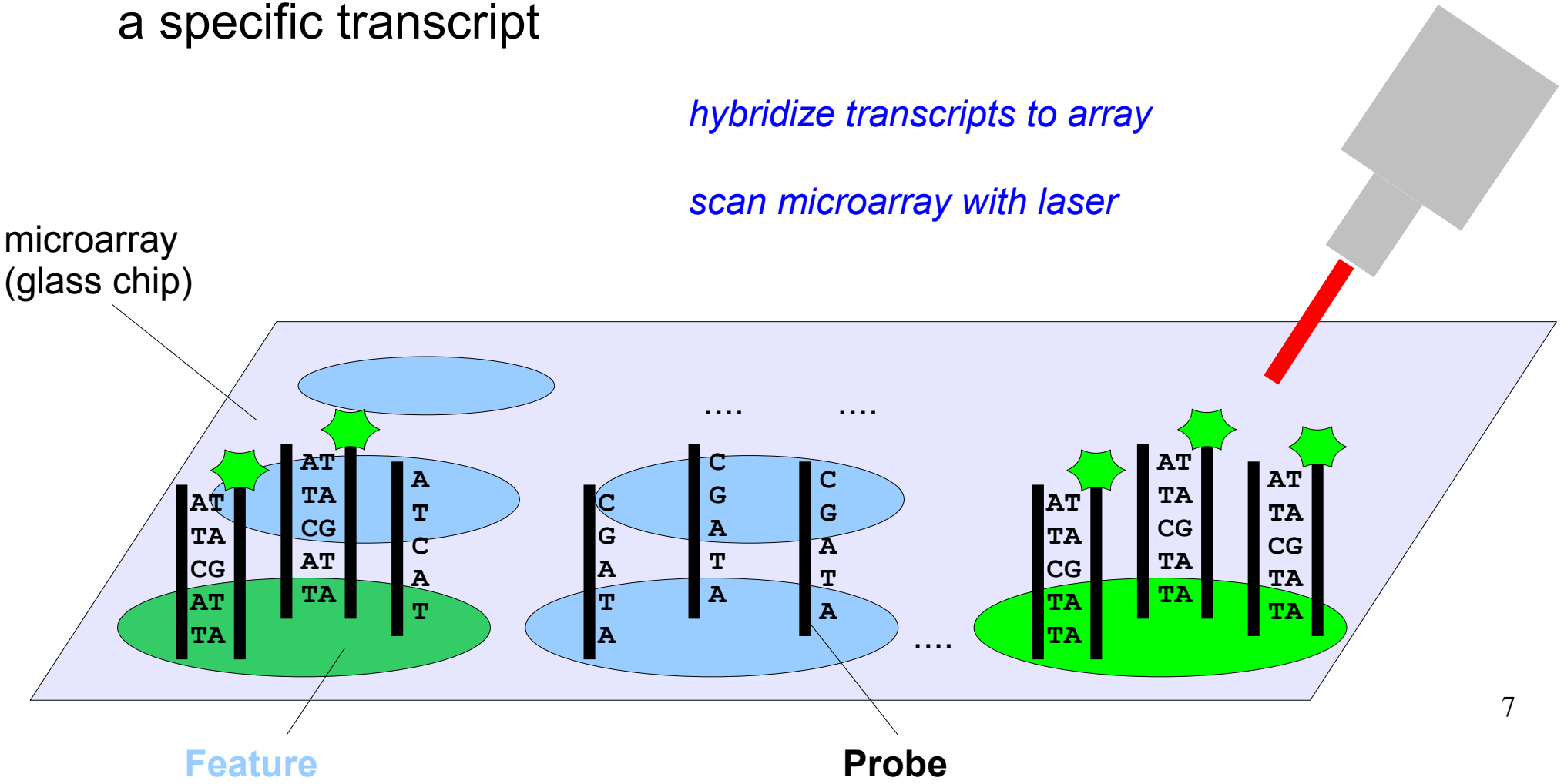
- DNA fragments (= „probes“) are attached to a microarray slides
- each probe has exact nucleotide sequence complementarity to a specific transcript

hybridize transcripts to array



Recap: Microarrays (2)

- DNA fragments (= „probes“) are attached to a microarray slides
- each probe has exact nucleotide sequence complementarity to a specific transcript



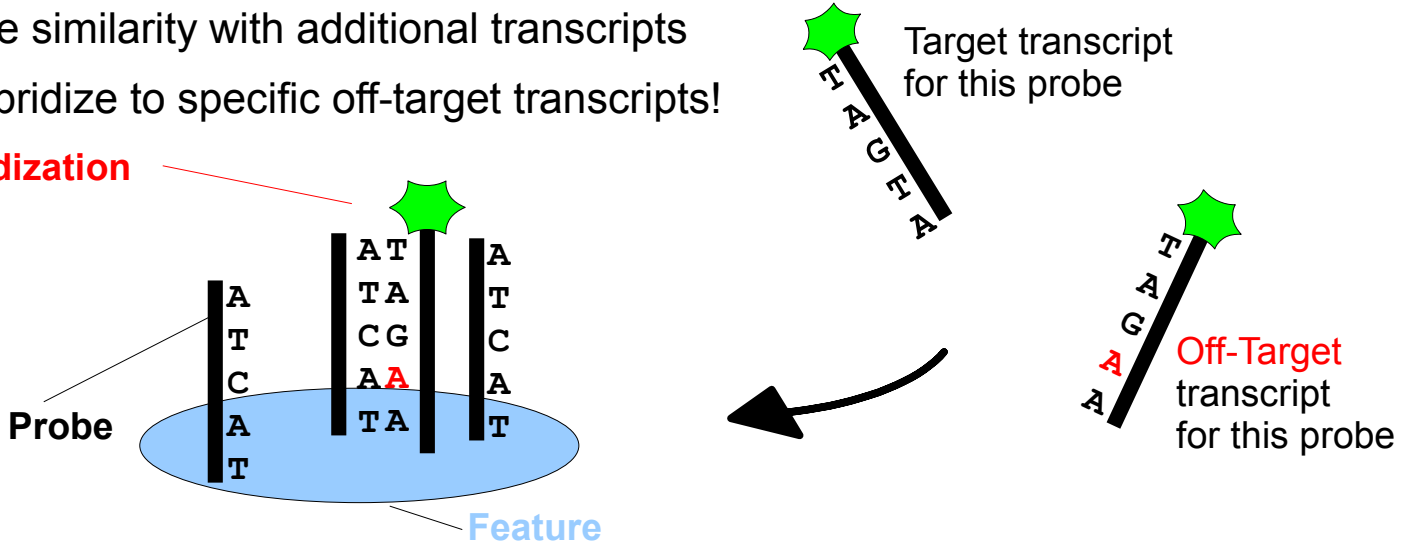
State of the Art: Problems (1)

- Gene expression microarray probes are designed with perfect complementarity to target mRNA transcripts

  BUT:

  - Share sequence similarity with additional transcripts
  - Probes may hybridize to specific off-target transcripts!

    = **cross-hybridization**



- Therefore:
  - artifacts of cross-hybridization = <u>important source of noise</u>
  - overall gene expression estimate suffers from biases of probe cross-hybridization

State of the Art: Problems (2)

- estimation of gene-level expression with strategies for avoiding a cross-hybridization bias:
  - dChip (Li and Wong, 2001)
    - Outlier removal procedure removes probes from a probe set with intensities which differ substantially from the other probes of this probe set
  - Robust estimation procedure (e.g. Robust Multichip Average)
    - Mitigate the effects of a small number of cross-hybridizing probes
  - Probe selection strategies
    - Select subset of probes which show highly correlated intensities across multiple samples

- Problem:
  - These methods work well when only a minority of probes are affected by cross-hybridization!

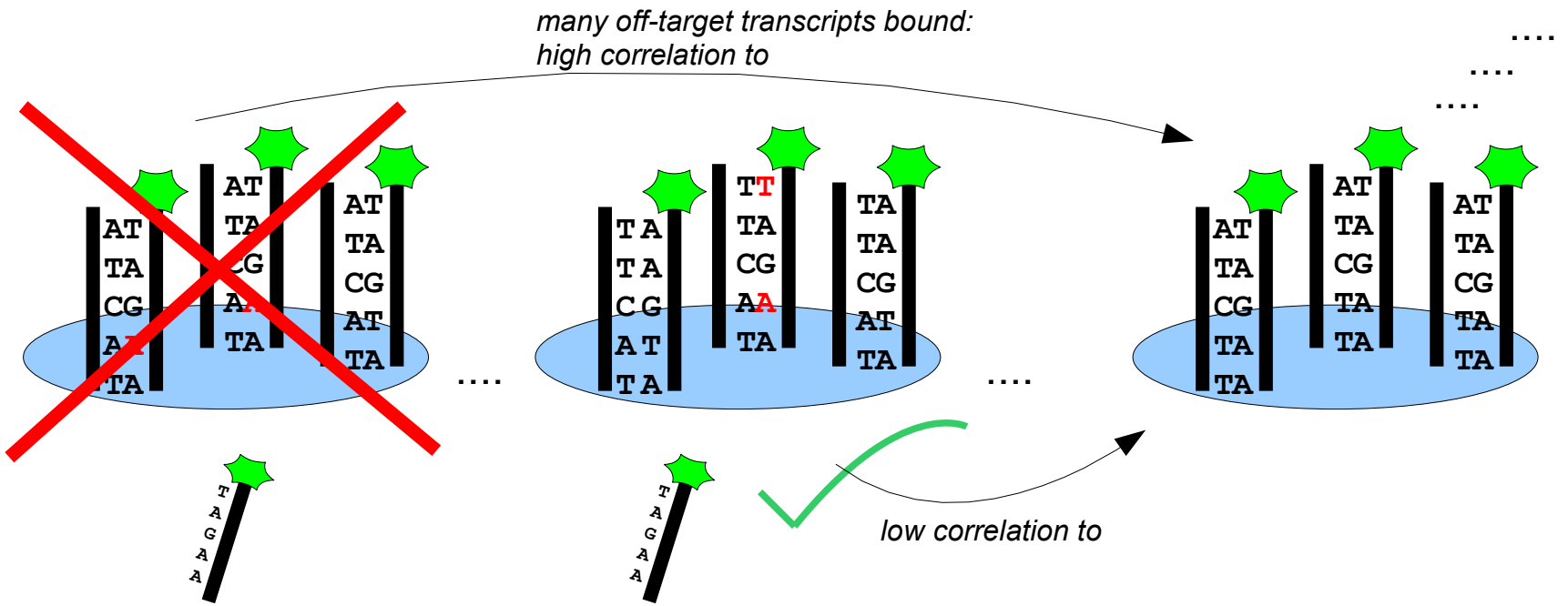
Motivation

- What to do when a large % of probes (of a probe set) bind to off-target transcripts?
  - no reliable estimate of overall gene expression possible

- Especially criticial for
  - Closely related genes
    - Possibly substantial number of potentially cross-hybridizing probes
  - Cross-hybridization is a major cause of false predictions of differential alternative splicing
  - Increasing oligonucleotide density on microarray chips
    - Specific features targeted by a small number of probes (e.g. exon arrays)

Idea

- systematically investigate cross-hybridization
  - assess how many probes actually are affected by potential cross-hybridization
- develop a model that corrects for cross-hybridization:
  - correlation-based filtering method to detect and remove probes showing sequence-specific cross-hybridization to off-target transcripts



Algorithm

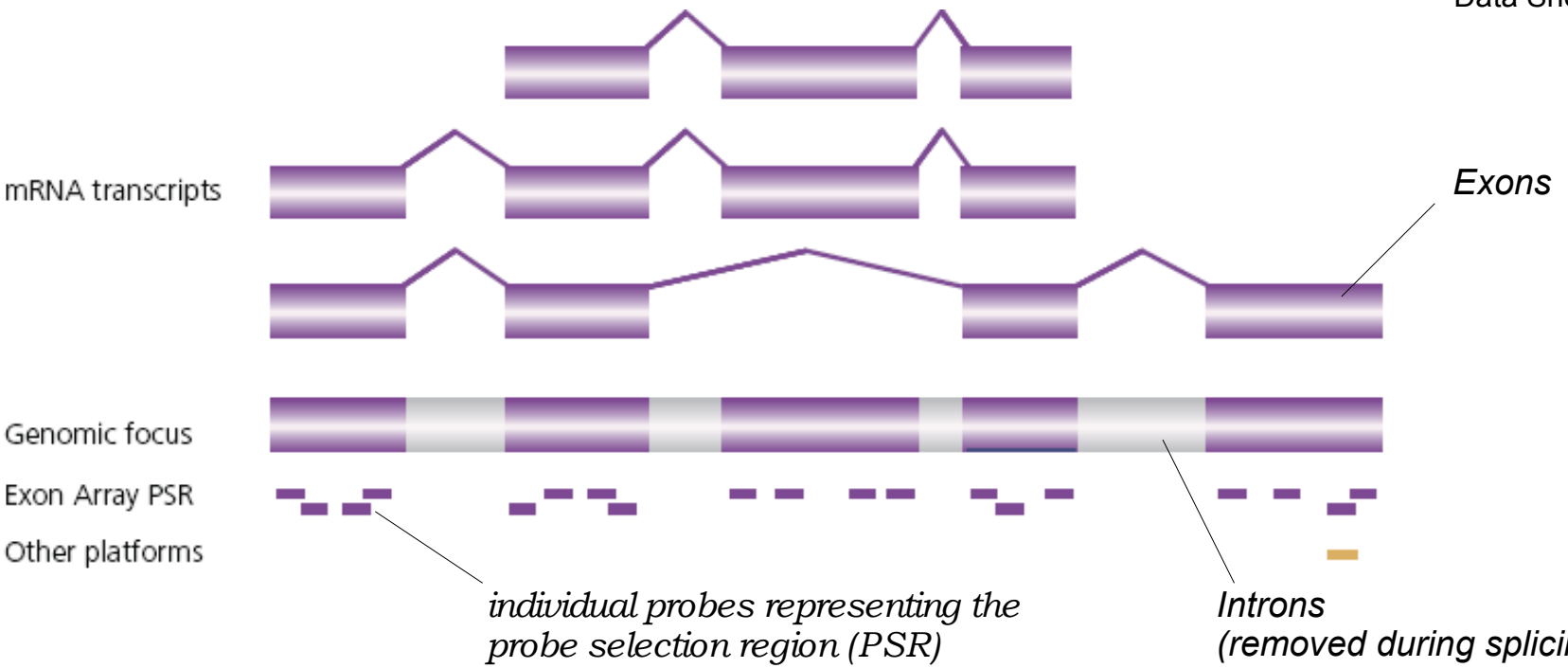
- Step (1)
  - identify potentially cross-hybrizing transcripts
    - perform sequence matching between probes and transcripts
- Step (2)
  - correct cross-hybridization bias:
    - when match between a probe to an off-target transcript is found
      - compute correlation between observed probe intensity with expression pattern of putative cross-hybridizing transcript
      - remove probes which follow the off-target expression pattern, keep remaining probes
- Step (3)
  - compute gene-level expression from the retained, informative probes

Methods (1)

- Affymetrix GeneChip Exon 1.0 ST Array used
  - contains ~ 6.5 million probes
    - to target all annotated and predicted exons in the genome

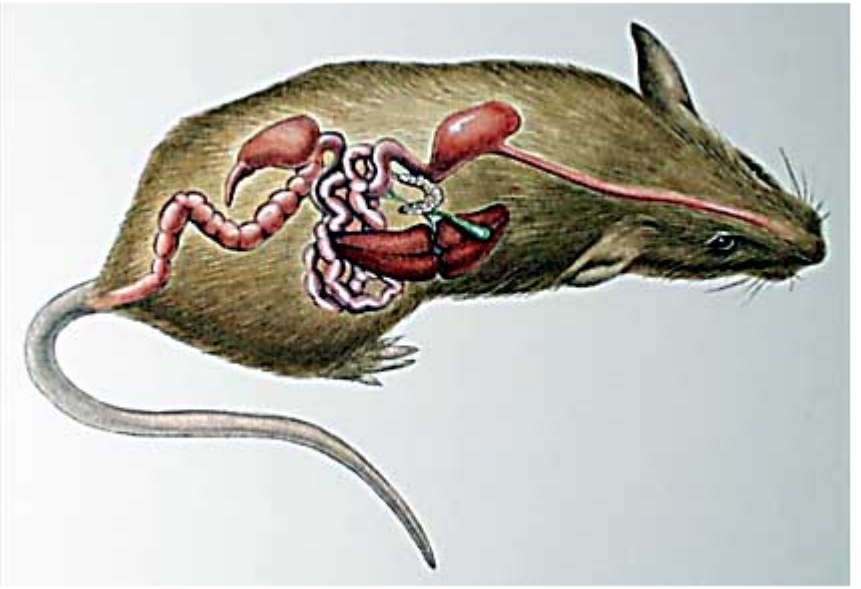
Figure 1: Coverage of probes across the entire length of the transcript.*

*Taken from the Affymetrix Exon ArrayDesign Data Sheet



Methods (2)

- analyzed exon array data for a panel of mouse tissues
  - brain, embryo, heart, kidney, liver, lung, muscle, ovary, spleen, testes, thymus
    - each with 3 replicates
- probe sequence-specific background correction
- normalization



Identify Cross-hybridization Candidates (1)

- map each 25-bp probe of Affymetrix exon array to transcripts
  - using the SeqMap algorithm
  - try to find off-target transcripts
    - which differ by any combination of mismatches or insertions/deletions (indels)
    - match edit distance = ∑ of mismatches and indels between the 2 sequences = Levenshtein distance
    - match edit distance between the 2 sequences is ≤ 3 bp

example of a match between a probe and transcript

Identify Cross-hybridization Candidates (2)

- Result of sequence matching:
  - most core probes uniquely match their target transcripts when allowing a matching distance of up to 3 bp

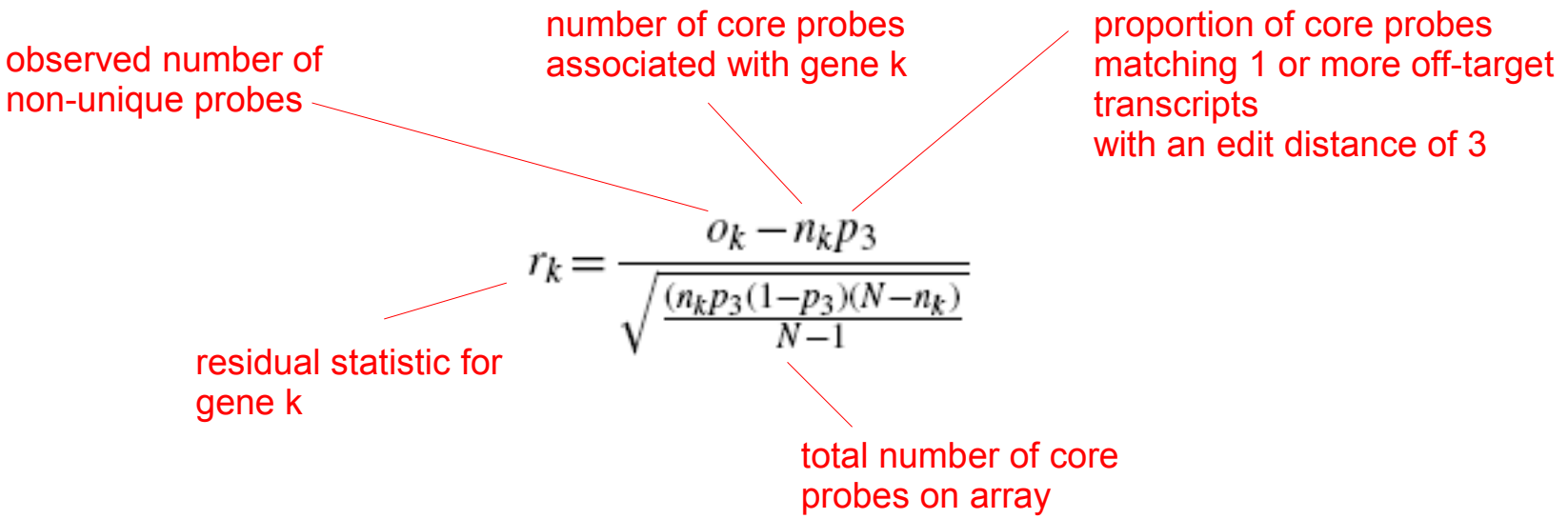
Table 1. The number of matches between core probes and off-target transcripts, allowing variable matching edit distances

Distance	Number matching transcripts				
	0	1	2	3	4+
0	839580 (98.03%)	11312 (1.32%)	3937 (0.46%)	573 (0.07%)	1069 (0.12%)
1	834693 (97.46%)	13174 (1.54%)	5501 (0.64%)	1042 (0.12%)	2061 (0.24%)
2	831059 (97.03%)	14534 (1.70%)	6395 (0.75%)	1438 (0.17%)	3045 (0.36%)
3	774502 (90.43%)	43623 (5.09%)	25673 (3.00%)	5083 (0.59%)	7590 (0.89%)

no off-target transcripts found

Identify Cross-hybridization Candidates (3)

- only a small number of probes matches to off-target transcripts
- Are we done yet? No!
  - Individual genes may have a large number of such probes!
  - Standardized residual statistic reveals genes with large proportions of non-unique probes:



- This statistic gives the observed minus expected number of non-unique probes, divided by the SD of a hypergeometric distribution.

Identify Cross-hybridization Candidates (4)

- detecting genes enriched for non-unique probes:
  - majority of transcript clusters tend to have residuals near 0
  - large number of transcript clusters that are enriched for potentially cross-hybridizing probes (residuals > 0)
    - 1136 transcript clusters having residuals > 7.0
    - Reason for this?
      - investigate whether these genes belong to paralog families
        - using the Ensembl compara homology database of paralog predictions
        - 65.29 % of genes were classified as belonging to a paralog familiy

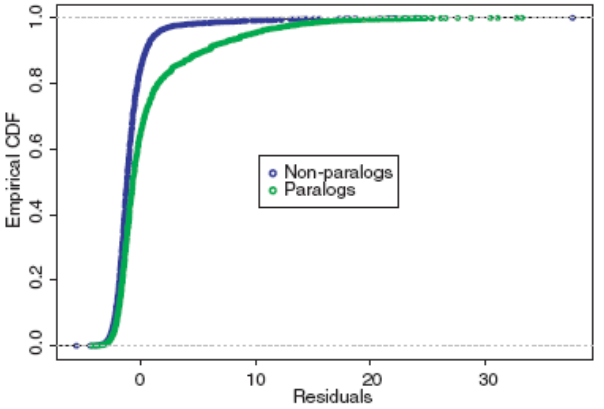
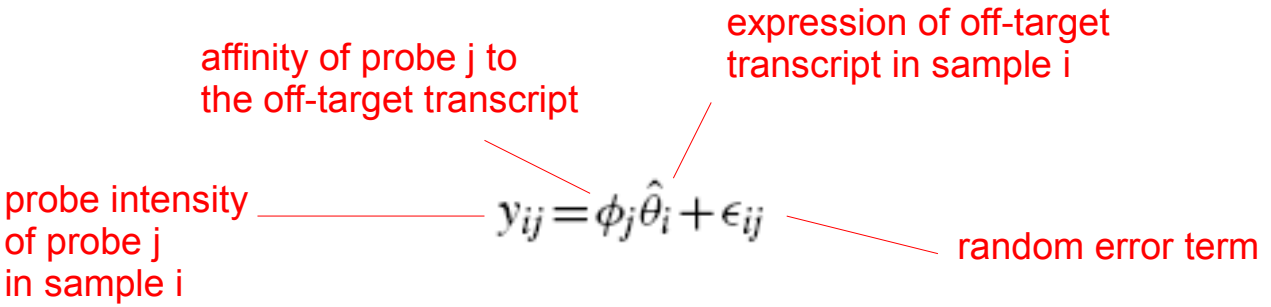


Fig. 2. The empirical cdf of gene standardized residuals, separated by paralog classification.

  - Result of this analysis:
    - Genes belonging to paralog gene families are enriched for probes with sequence similarity to off-target transcripts!

The Cross-Hybridization Model

- Can the expression of an off-target transcript explain the observed probe intensity?
  - probe intensity y_{ij} of probe j in sample i is modeled as:



- R^2 statistic of this model represents the % of probe intensity variance which can be explained by cross-hybridization to the off-target transcript

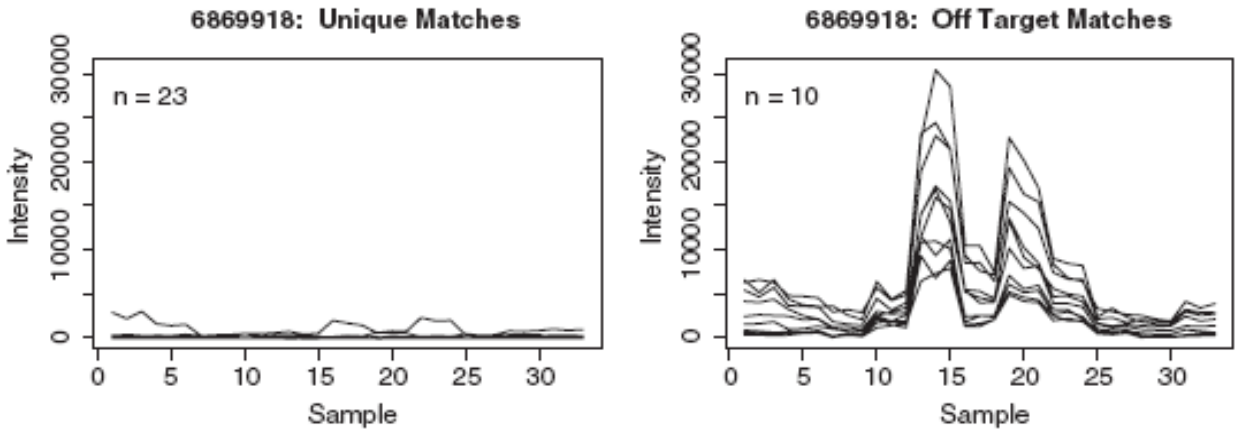
Correlation-Based Filtering (1)

- Example:

probes uniquely matching target transcript

probes matching off-target transcript

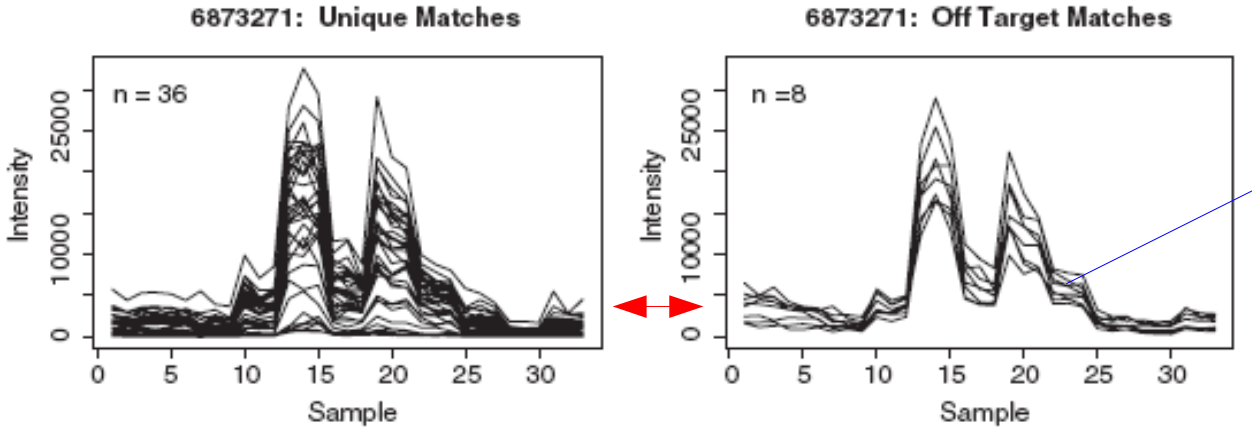
Gene Scd3



expression pattern of Scd1!

→ some of the Scd3 probes are cross-hybridizing to Scd1 transcripts!

Gene Scd1



probe intensities across the 33 tissue panel experiments

Correlation-Based Filtering (2)

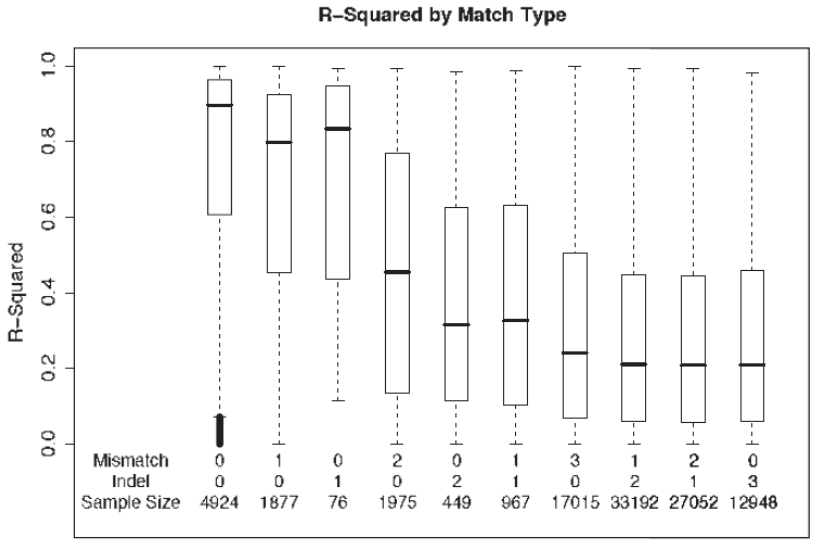
- Some of the Scd3 probes are cross-hybridizing to the Scd1 transcript
- Resulting gene-level expression pattern changes depends on set of probes used for summarization
  - only a relatively small number of cross-hybridizing probes can result in large biases of gene-level expression

Therefore:

- Identiy potentially cross-hybridizing probes
  - remove cross-hybridization bias → use uniquely matching probes to estimate expression level of Scd3

Match Type Effects

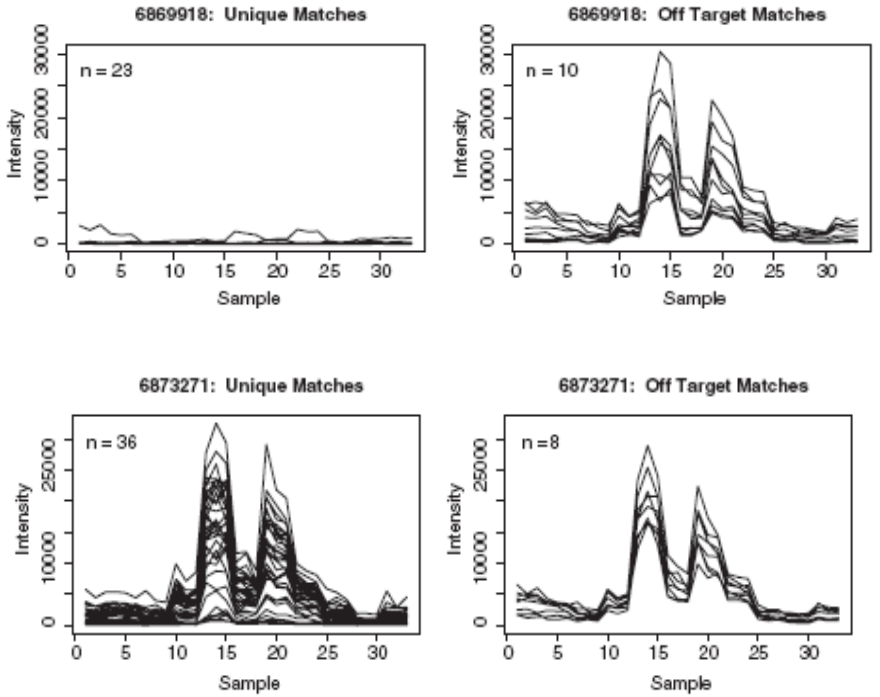
- Assess effects of match type (number of mismatches/indels) on cross-hybridization behaviour:



  - decreasing correlation between probes and expression patterns of matching off-target transcripts as match edit distance between probe and transcript is increased
  - allowing an edit distance of 3 bp between probes and off-target transcripts → probes may show strong signals of cross-hybridization, compared with signals expected by chance

GeneBASE-xhyb

- GeneBASE = program for generating gene-level expression estimates
- GeneBASE-xhyb extends GeneBASE with cross-hybridization correction:



- remove probes showing strong evidence of cross-hybridization
  - Probes with up to 2-bp mismatches/ indels and correlation > 0.7 with the off-target transcript expression level are excluded from gene-level summarization.
- retain probes which have matches to off-target transcripts but show little evidence of cross-hybridization

Evaluation of GeneBASE-xhyb (1)

- compare GeneBASE and GeneBASE-xhyb
  - to estimates obtained by Solexa sequencing of RNA sequences
    - for mouse liver, skeletal muscle and brain
    - Independent samples pooled from adult mouse tissues
  - generated estimates of gene-level expression from sequencing reads
    - by counting the reads per kilobase gene exon per million mapped reads
  - for each RefSeq transcript:
    - generated expression estimate by counting the number of normalized reads which fall in exonic regions

Evaluation of GeneBASE-xhyb (2)

- Is GeneBASE-xhyb more concordant with Solexa expression estimates than GeneBASE?
  - Result: yes
  - Implication:
    - Correction for cross-hybridization leads to a significant improvement of gene expression estimates!
- Comparison with other methods:

Table 2. Spearman correlation between exon array estimates of gene expression and ultra high-throughput sequencing estimates for different summarization methods

	GeneBASE-xhyb	GeneBASE	RMA	Plier	IterPlier
Liver (N = 12 339)	0.8539	0.8521	0.8064	0.8198	0.8125
Muscle (N = 13 136)	0.8500	0.8481	0.8072	0.8109	0.8080
Brain (N = 13 783)	0.7542	0.7535	0.7443	0.7275	0.7132

Summary

- Proposed a correlation-based filtering method:
  - to detect and remove probes showing sequence-specific cross-hybridization to off-target transcripts.
- Takes advantage ot the tiling of probes of all transcribed regions
  - to compare the observed probe intensity with the expression pattern of the putative cross-hybridizing transcript
- Include as many informative probes as possible for the summarization of gene-level expression
- Predictions of gene-level expression were validated using Solexa sequencing data
  - cross-hybridization modeling improves estimates of gene-level expression