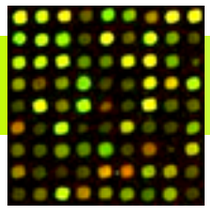
Quality Assessment of the Affymetrix U133A&B probe sets by target sequence mapping and expression data analysis

Yuri L. Orlov et al., August 2007

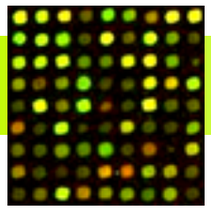
Special Topics on Bioinformatics
SS 2010

J. Palme



Topic

- carefull checking of probe sequences of microarray
- verification of probe sequences against reference data in DBs => identification of unreliable probes
- remove data of problematic probes before analysis
- reevaluation of three corrected cancer data sets
- improvements in selection of differentially expressed genes, clustering and construction of co-regulatory expression networks expected



U133A & B

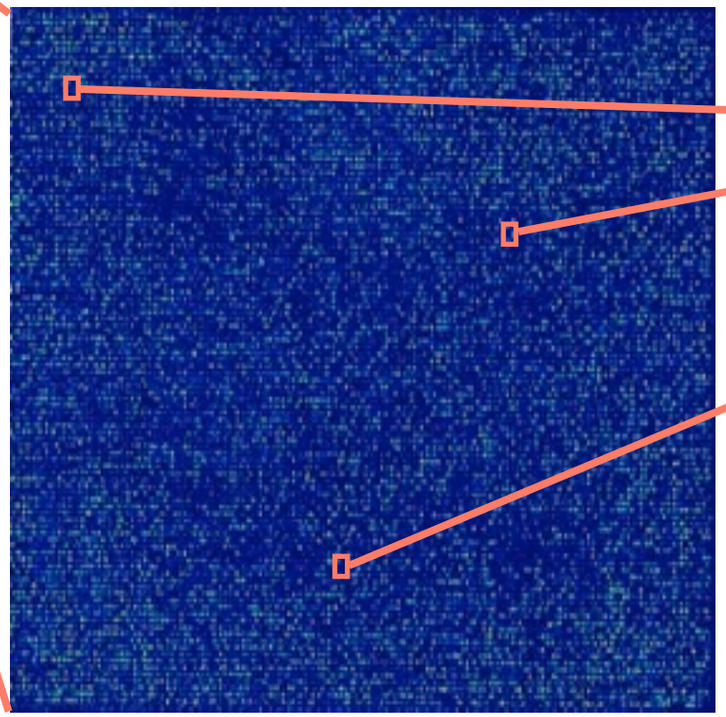
This research work is based on 2 specific microarrays

- Affymetrix Human Genome U133A Array
- Affymetrix Human Genome U133B Array
- Official annotation files for both arrays from Affymetrix, currently release 30 from 15/11/09
  http://www.affymetrix.com/support/technical/annotationfilesmain.affx?hightlight=true&rootCategoryId=
- idea for this work seems to come from the Micro Array Quality Control project (MAQC) initiated by the FDA

U133A

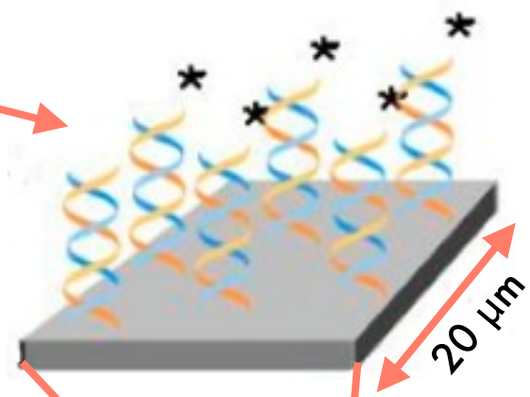
Human Genome U133A GeneChip® Array

(1) Probe Array



(4) Spot

Each spot contains ~4x10^8 copies of a specific probe which is complementary to the genetic material of interest. One probe is a single stranded, flourescently labelled oligonucleotide (25 mers)



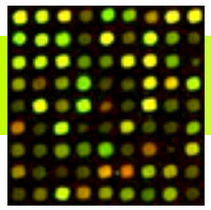
(3) Probe Pair

Each Perfect Match (PM) and MisMatch (MM) spots are associated by pairs.

(2) Probe Set

Each probe set contains 11 probe pairs (PM:MM) of different probes which intend to target the same transcript

The Human Genome U133A GeneChip® array represents more than 22.000 full length genes and EST clusters

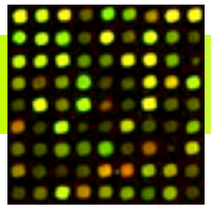


U133A & B

Critical Specifications for GeneChip® Human Genome Products

	Cartridge Format		Plate Format	
	Human Genome U133 Plus 2.0 Array	**Human Genome U133A 2.0 Array**	**Human Genome U133 A Array Plate**	**Human Genome U133 B Array Plate**
Number of transcripts	~47,400	~18,400	~18,400	~20,600
Number of genes	>38,500	>14,500	>14,500	>18,500
Number of probe sets	>54,000	>22,000	>22,000	>22,000
Feature size	11 µm	11 µm	8 µm	8 µm
Oligonucleotide probe length	25-mer	25-mer	25-mer	25-mer
Probe pairs/sequence	11	11	11	11
Control sequences included:				
Hybridization controls	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*	*bioB, bioC, bioD, cre*
Poly-A controls	*dap, lys, phe, thr*	*dap, lys, phe, thr*	*dap, lys, phe, thr*	*dap, lys, phe, thr*
Normalization control set	100 probe sets	100 probe sets	100 probe sets	100 probe sets
Housekeeping/Control genes	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)	GAPDH, beta-Actin, ISGF-3 (STAT1)
Detection sensitivity	1:100,000*	1:100,000*	1:100,000*	1:100,000*

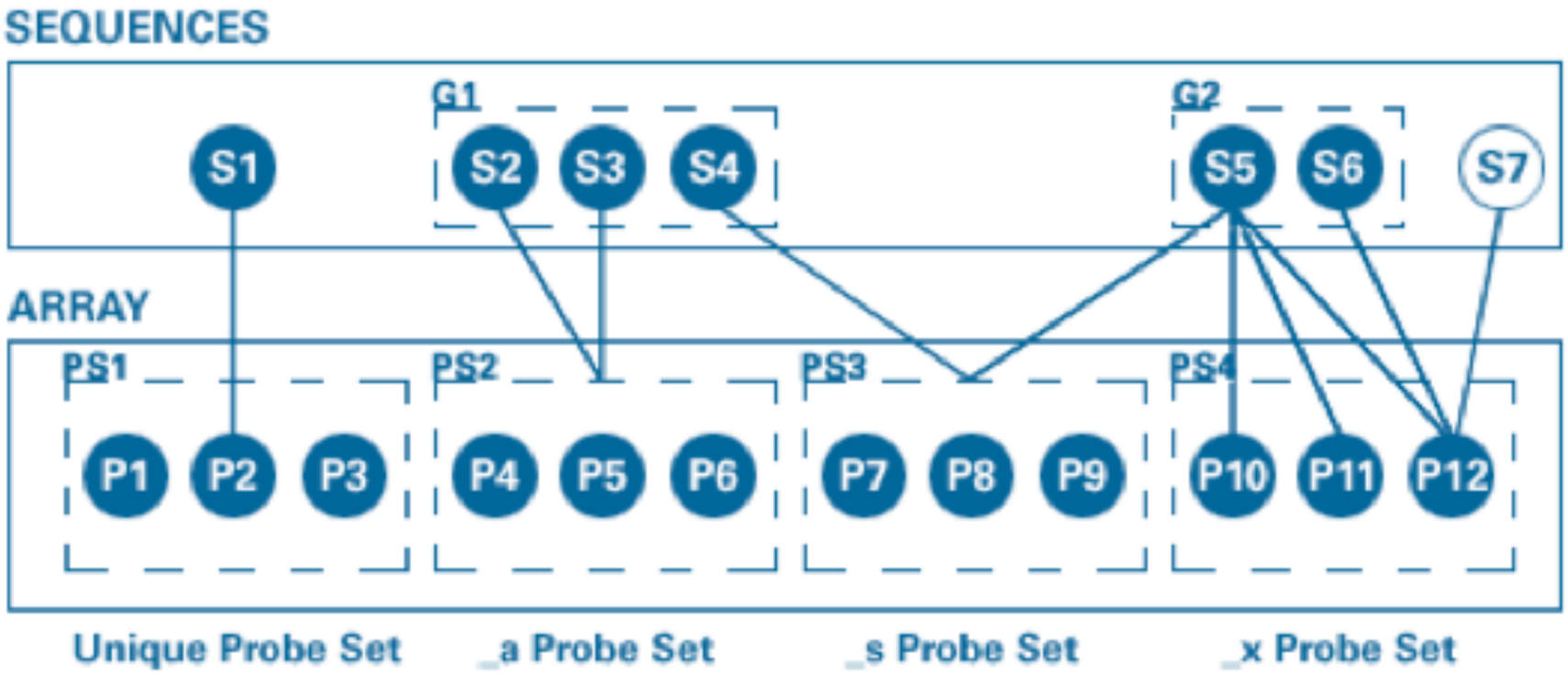
*As measured by detection of pre-labeled transcripts derived from human cDNA clones in a complex human background.



Probe Set Naming I

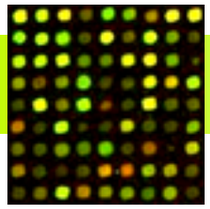
- ..._at: (anti-sense target) detects antisense strand of given gene, these are unique probes
- _a_at: (gene family probeset)
  recognize multiple transcripts of same gene
  _s_at: (identical probeset)
  recognize multiple transcripts from different genes
  _x_at: (mixed probe set)
  cross-hyb. with other sequences used for design
- a, s, x derived from gene cluster + gene family info

Probe Set Naming II



Legend

G: a set of sequences belonging to the same gene family
S: a sequence
PS: a probe set
P: a probe in a probe set

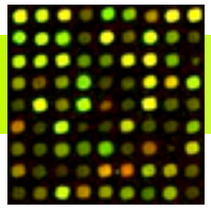


Verifications

- Verification of probe quality - as annotated! (assuming identity of probe and annotation)
- Get official sequence information for the probes from Affymetrix annotation files
- Verify unique sequence to gene mapping
- Verify that sequence maps into human genome
- Verify correct strand orientation
- Consideration of repeated exon elements

BLAT Analysis

- BLAT sequence search at 90% simularity level
- check of overlap with exonic regions in RefSeq and mRNA / spliced EST variants in hg17 and hg18
- mapping of probe sets to gene sequence blocks based on initial target sequences
- results of analysis (chromosom coordinates, orientation, overlapping details with exons / repeats...) stored in local DB against probe set ID



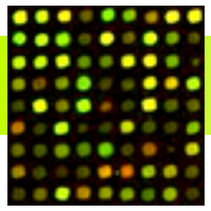
Used Reference Data

- NCBI Human Genome (hg17 and hg18)
- RefSeq
- NCBI GEO GSE4922 breast cancer data sets
- NCBI GEO GDS1962 brain cancer data sets
- NCBI GEO GGSE586 lung cancer data sets

Used Tooling

- For sequence verification:
  - BLAT (Difference to BLAST)
  - UCSC Genome Browser
  - Software developed at GIS (Genom Institute of Singapore) and BII (Bioinformatics Inst. Singapore)

- For statistical evaluations:
  - SAM 3.1 (Statistical Analysis of Microarrays)
  - Statistica 6
  - StatXact 6



Probe Quality Criteria

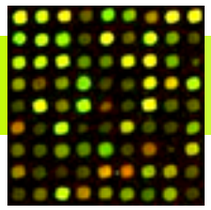
1. sequence with with unique locus in human genome
2. perfect match of transcript
3. correspond to sequence of transcribed strand at this locus including correct strand orientation
4. no overlapping with any other non-gene sequence
5. correspond to mature RNA (only exons included)

Problematic Sequences

# locations(Hg18)	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6+	Tag0	Total
#Affymetrix IDs	42708	450	129	67	42	84	1212	44692
%	95.56	1.0	0.28	0.14	0.09	0.18	2.71	100

Sequences in U133A and U133B

- Tag0 - not found in human genome
- Tag1 - found exactly once => correct sequence
- Tag2+ - found at multiple loci



Tag0 and Tag2+

- Tag0:
  - 45% "xenosequence/non-human"
    (cow, mouse, pathogens, rat ...)
  - 17% classified as low-accuracy
- Tag2+:
  - 81737_**at** found at 22 different locations
  - 213089_**at** found at more than 11 locations
- Probedesign based on Genbank sequences without verification of mapping to human genome

Repeats in Tag I

Set of genome repeats	Repeat class	# in target sequences
Simple repeats	Simple repeat, Low complexity	3233
Short transposons (<300 bp)	DNA, SINE/Alu, SINE/MIR	4347
Long transposes (>300 bp)	LINE/CR1, LINE/L1, LTR/ERV1/ERVK/ERVL/MaLR	5420
Non-transposons and satellites	Other, RNA, rRNA, Satellite, scRNA, snRNA, srpRNA	80

- lead to cross hybridization and wrong detection of expressed genes

Inverse Sequences

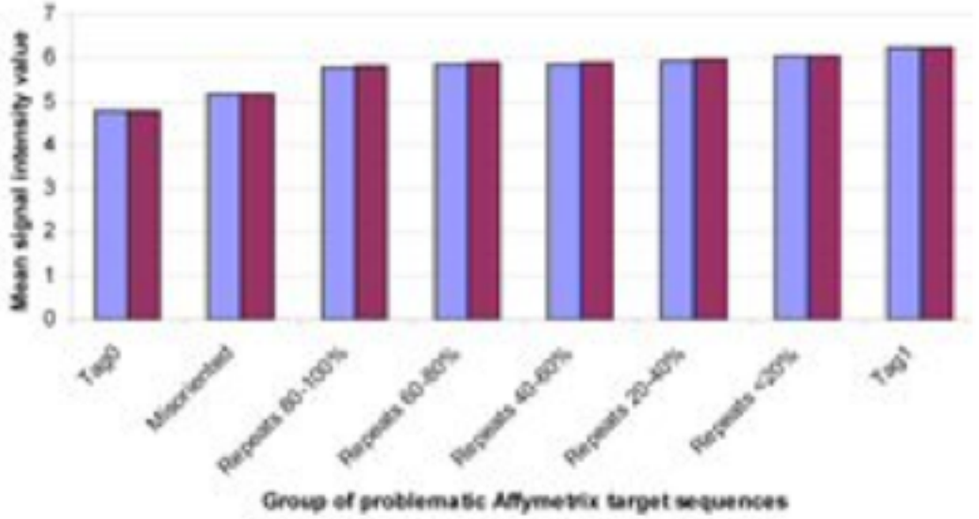
Sets	Correct orientation		Misoriented to intended transcript in as verified by				Total in Tag1
	Total	Match to NAST	Manual curation	RefSeq	mRNA	EST	
# Target seq	41898	13260	370	138	302	487	42708
%	93.74	29.66	0.82	0.3	0.67	1.08	95.65

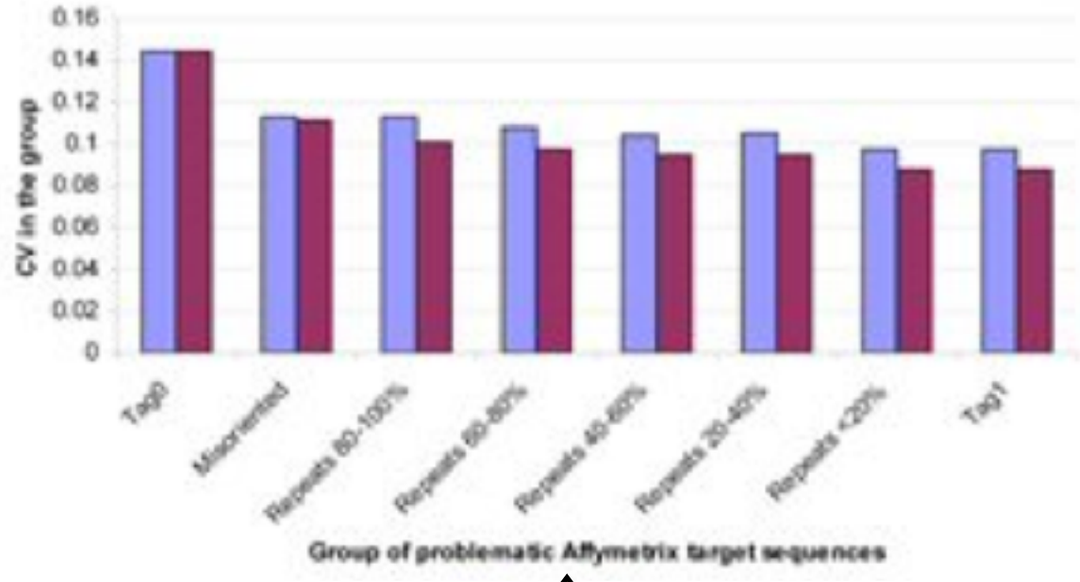
- inversely oriented if it matches the opposite strand of RefSeq, mRNA or EST related gene
- match to NAST (natural antisense transcript) 30%

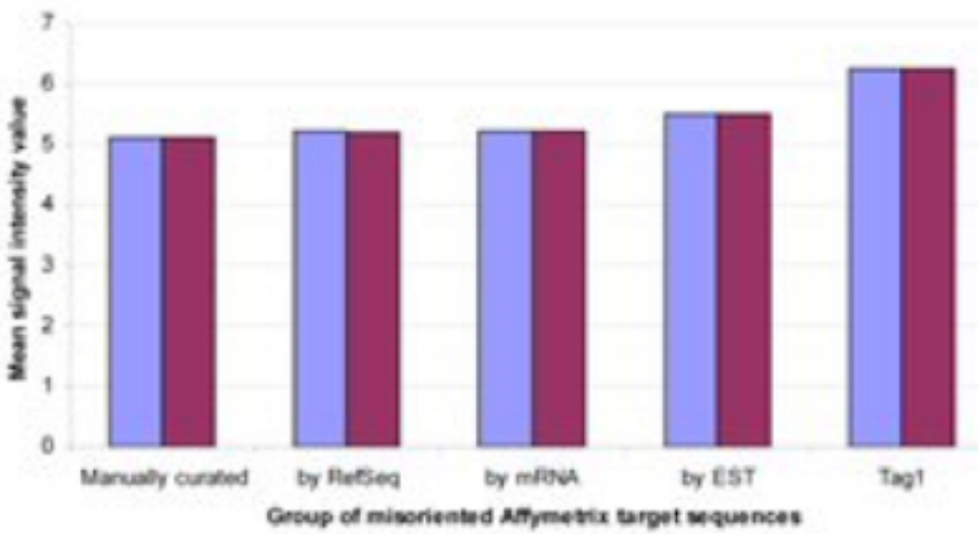
Total Picture

Target sequences groups	Non-redundant # of probesets	%
Total # of non-Tag1 sequences, including:	**1984**	**4.43**
Tag0	1212	2.71
Tag2+	772	1.72
Total # misoriented target sequences, including:	**810**	**1.81**
RefSeq IDs (by blocks)	138	0.3
mRNA GenBank (by blocks)	302	0.67
Manual curation protocol	370	0.82
Total # of target sequences overlapped with repeats, including:	**3387**	**7.57**
Overlap 80-100% of target sequence length	761	1.7
Overlap 60-80%	936	2.09
Overlap 40-60%	1690	3.78
Total # of useful Tag1 target sequences, including:	**38511**	**86.16**
Overlaps with observed transcripts in opposite strand	13260	29.66
Misoriented to ESTs in Tag1	487	1.08
Target sequences with 20-40% of repeats	2409	5.39
Target sequences with <20% of repeats	1210	2.7
TOTAL # of Affymetrix target sequences	**44692**	**100**

Expression Levels



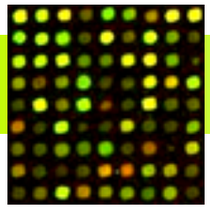




CV per sequence category

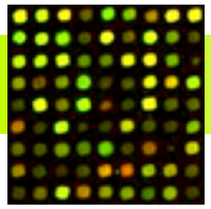
Mean per sequence category

Mean for inverted sequences



Repeats in Sequences

- Data for G1 and G3 grade breast cancer which show expression for at least 4000 probe sets
- Through overlaps with repeats - less specific binding, this means: portion of these probe sets in statistically significant expressed genes decreases => decreased recognition of cancer signals
- Only relevant for longer repeat overlaps
- No impact for simple repeats and low complexity sequences

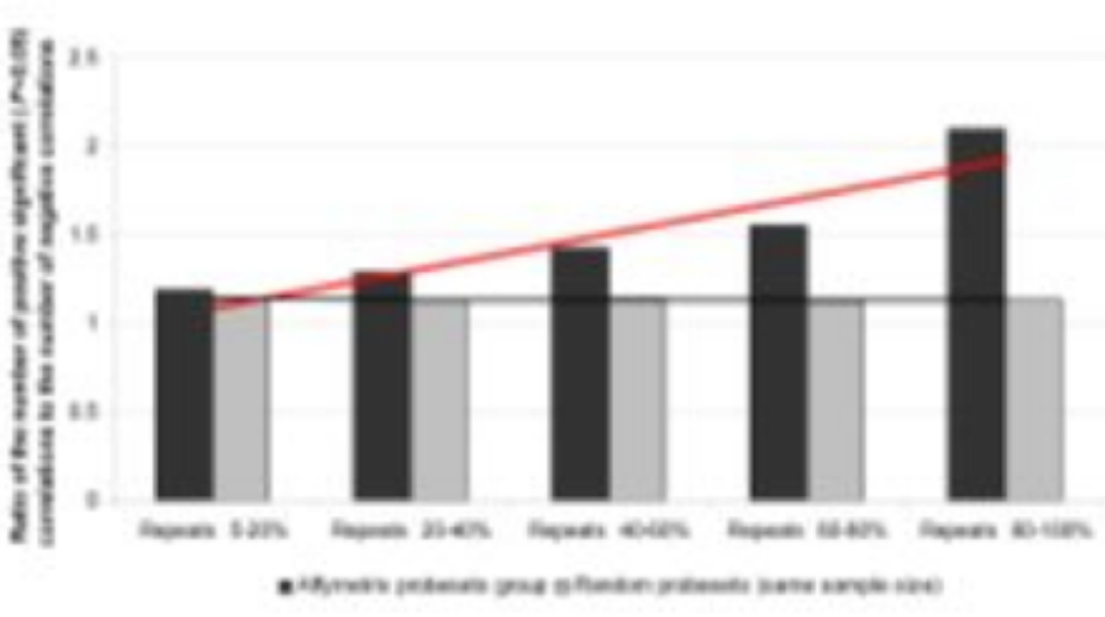


Cross-Hybridization I

- Kendall τ rank correlation between probe set values
- Comparison of problematic groups with randomly selected groups
- Same correlation behavior for whole array and randomly selected groups
- Higher positive correlation for problematic target sequence groups compared to control groups
  => can lead to spurious correlation

Cross-Hybridization II

- Ratio of positive to negative correlation values was about 1 for random groups
- For probe sets with repeat coverage the ratio increases with length of overlap (trend not visible for same size subgroup of normal probe-sets (Tag 1)

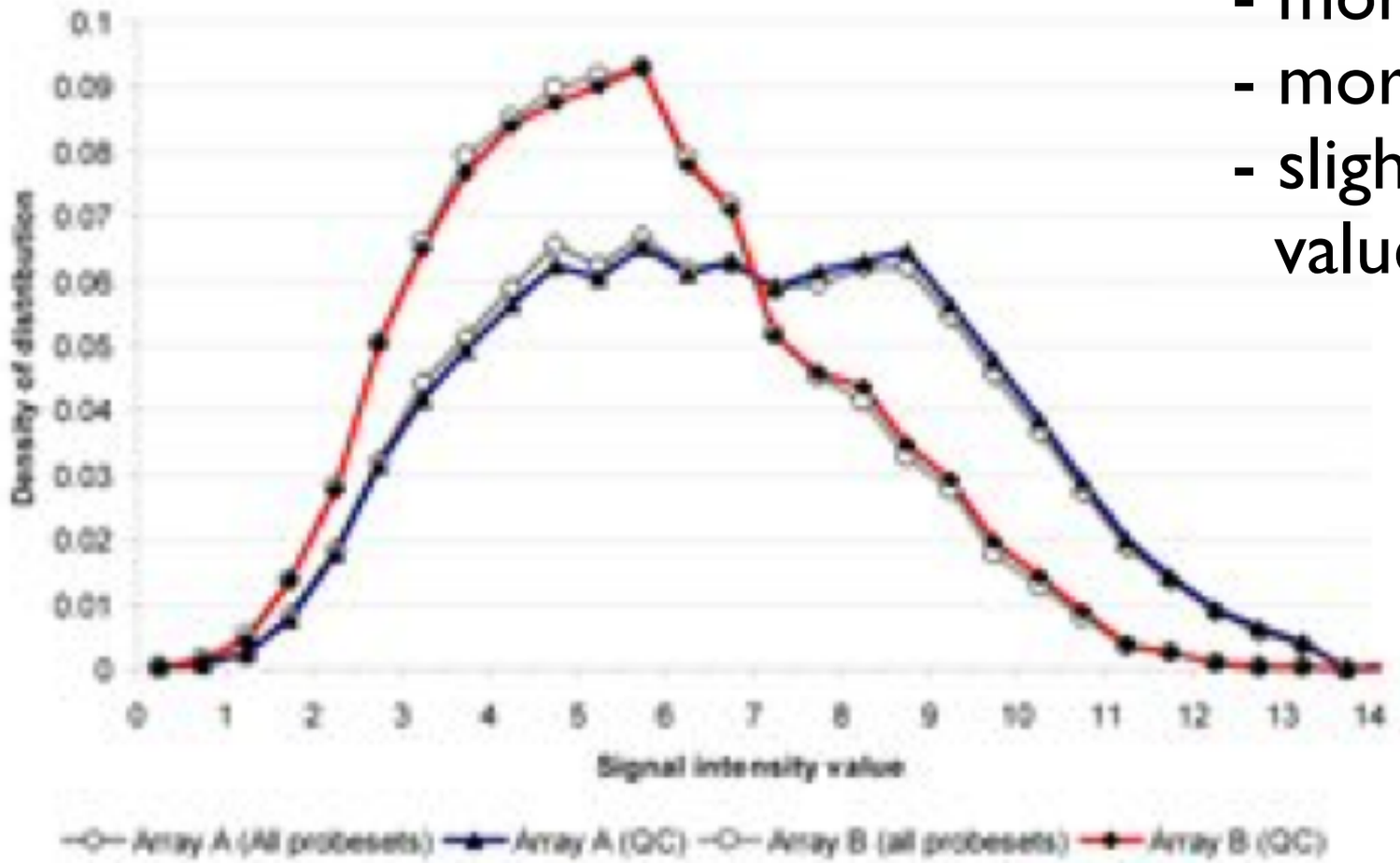


Comparison U113A&B

	# Probesets	# Correct probesets (passed QC)	% of correct probesets (passed QC)
A and B	100	98	98.0
Service probesets	68	N.A.	N.A.
Array U133A	22115	19753	89.3
Array U133B	22477	18660	83.0

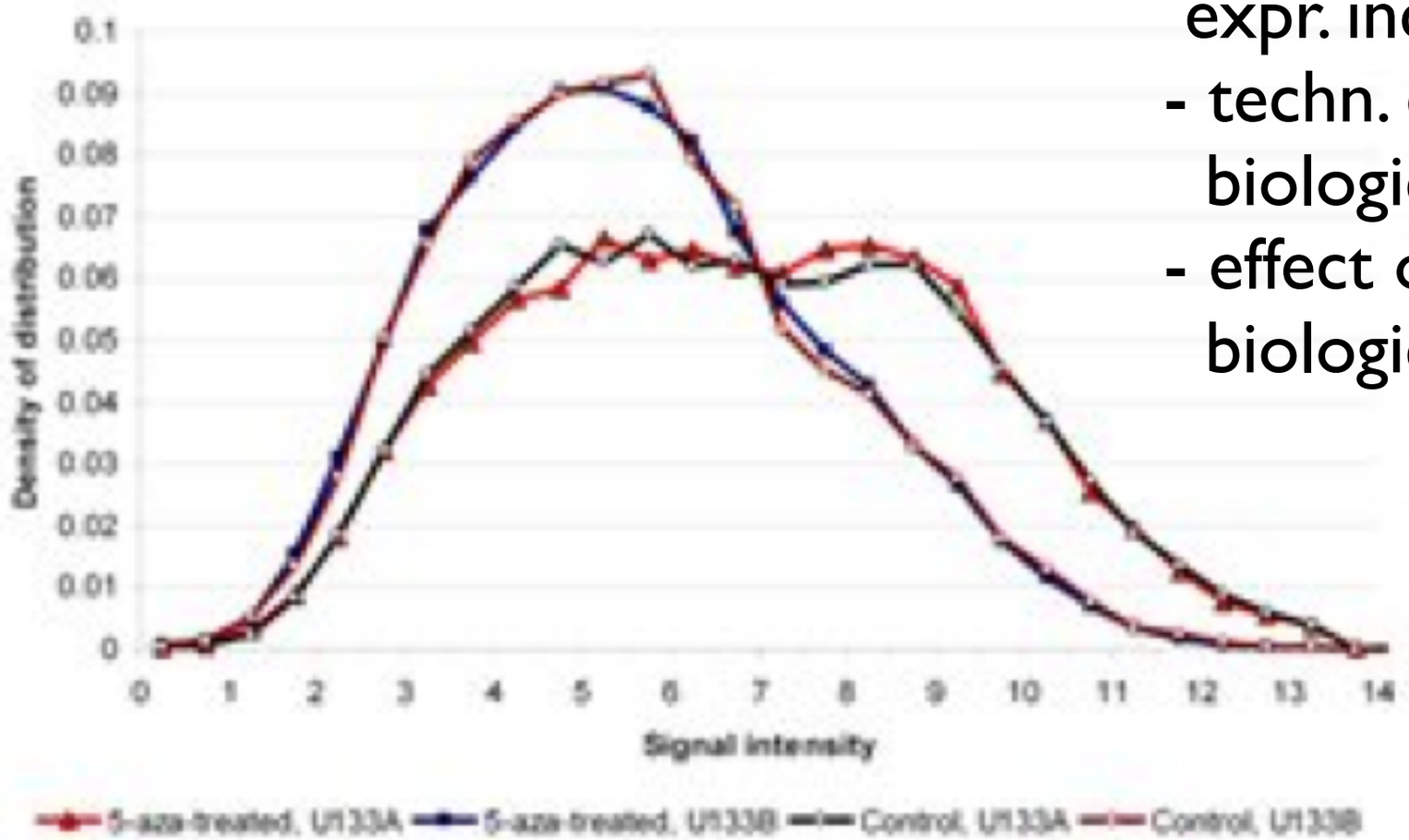
- Passed QC means: tag1, correct orientation on chromosome and repeat coverage is less than 40%
- U113A:
  - more probes with annotation quality
  - higher signal intensity and lower noise

Signal Intensity Distrib.

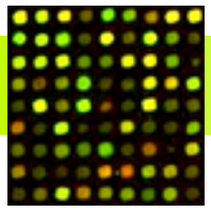


- lung cancer cell lines:
  - higher expr. values on A
  - more specific expr. on A
  - more noise on B
  - slightly higher expr. values for QC probesets

Signal Intensity Distrib.

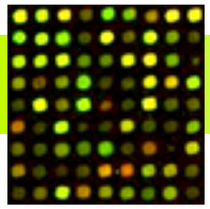


- 5-aza treated samples (11a.) vs. untreated control samples (10a.): (5-aza: higher methylation and expr. increase for many genes)
  - techn. differences larger than biological differences
  - effect of QC filtering similar to biological variation



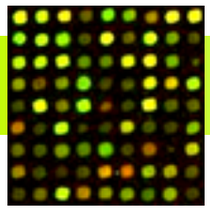
SUMMARY I

- Carefull analysis reveals probe sequence problems: non-human, multi-locus, misoriented, non-specific sequences
- Only 86% well designed sequences should be used for data analysis
- U113A has higher number of correct probesets and performs better - related to expression level and noise



SUMMARY II

- Unreliable sequences lower average expression level and add noise
- False correlation for probesets with higher repeat coverage
- Prerequisite for microarray data analysis: Check sequences with most recent annotation files against current references



SUMMARY III

- Excellent research work presented in paper
- No of unreliable probe sets can and will increase with new releases of annotation files and progress in gene definition
- 4 G probes: with 4 or more guanine bases
- chimeric RNA - transcribed from 2 genes ...

  => stay alert when using microarrays