Oigonucleotide (Affyx) Array Basics

Joseph Nevins
Holly Dressman
Mike West

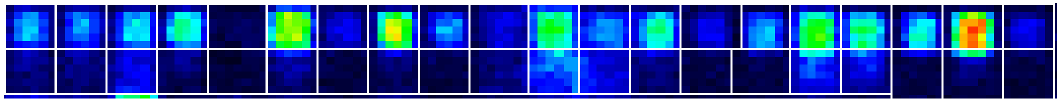
Duke University

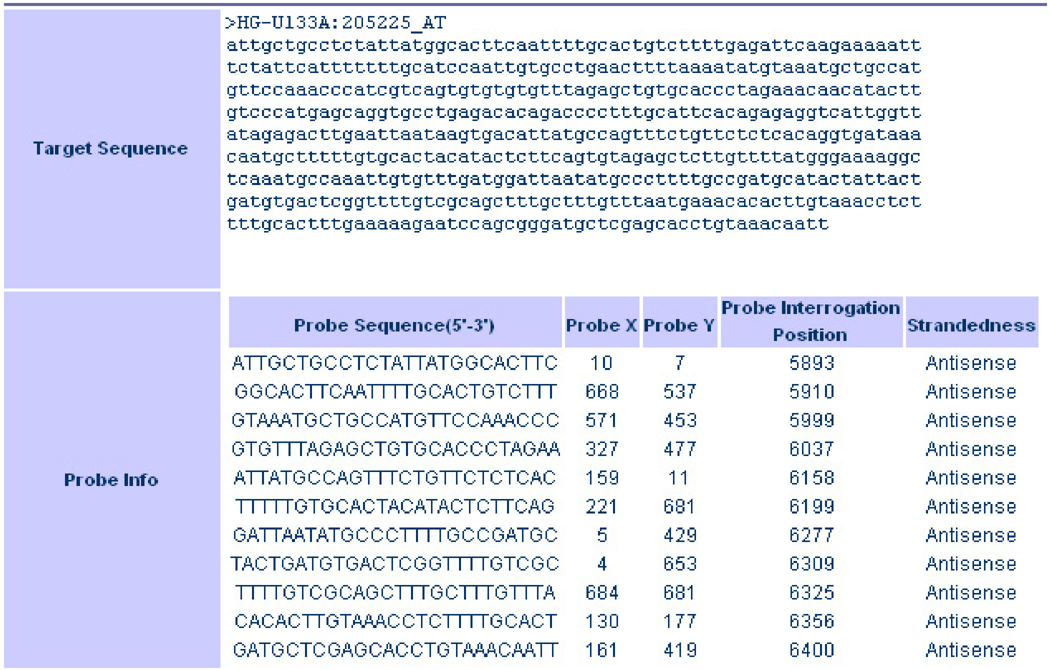
Affymetrix GeneChip© Arrays

Human U95a 12k genes
U133a 22k genes
U133+ 50k genes(?)

Splice variants, alternative oligos

Gene: 25mer oligo Probe set



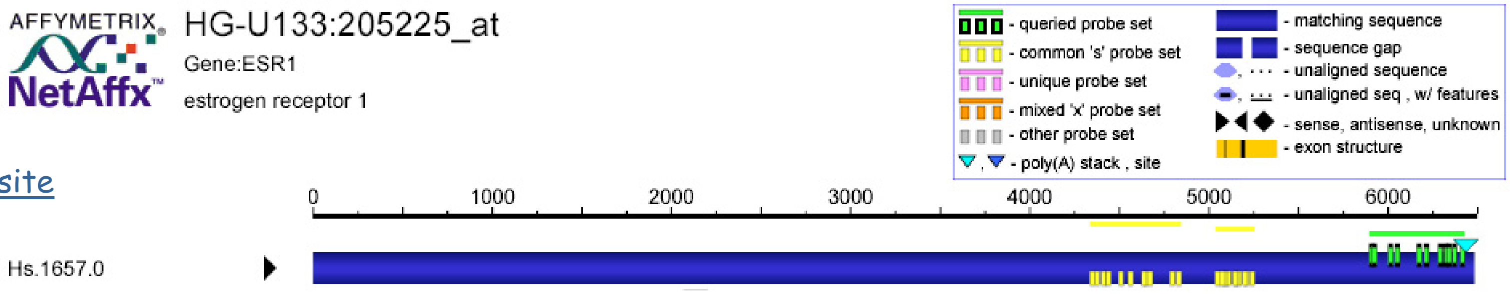


Target Sequence

```
>HG-U133A:205225_AT
attgctgcctctattatggcacttcaattttgcactgtcttttgagattcaagaaaatt
tctattcatttttttgcatccaattgtgcctgaacttttaaaatatgtaaatgctgccat
gttccaaacccatcgtcagtgtgtgtgtttagagctgtgcaccctagaaacaacatactt
gtcccatgagcaggtgcctgagacacagacccctttgcattcacagagaggtcattggtt
atagagacttgaattaataagtgacattatgccagtttctgttctctcacaggtgataaa
caatgctttttgtgcactacatactcttcagtgtagagctcttgttttatgggaaaaggc
tcaaatgccaaattgtgtttgatggattaatatgcccttttgccgatgcatactattact
gatgtgactcggttttgtcgcagctttgctttgtttaatgaaacacacttgtaaacctct
tttgcactttgaaaaagaatccagcgggatgctcgagcacctgtaaacaatt
```

Probe Info

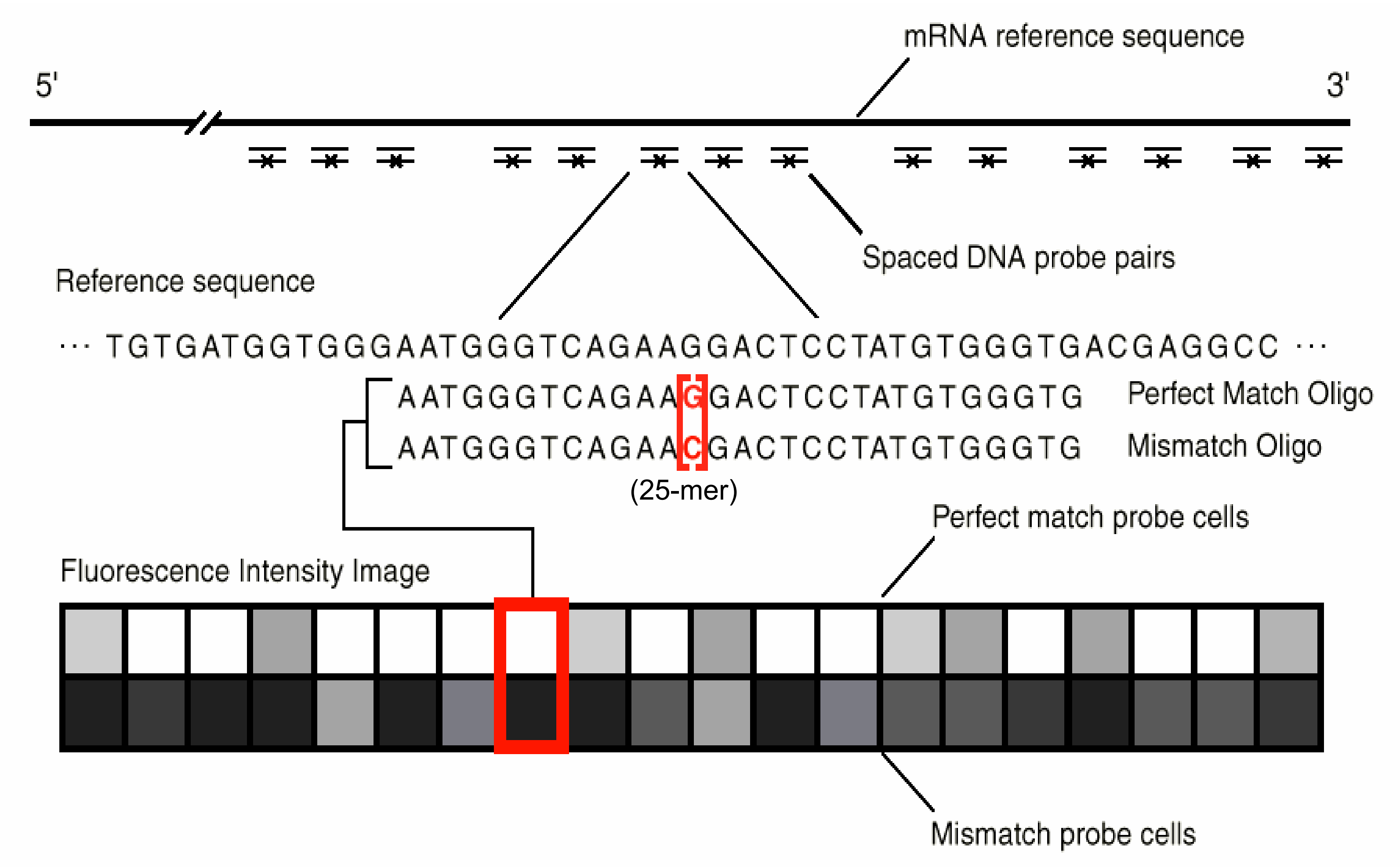
Probe Sequence(5'-3')	Probe X	Probe Y	Probe Interrogation Position	Strandedness
ATTGCTGCCTCTATTATGGCACTTC	10	7	5893	Antisense
GGCACTTCAATTTTGCACTGTCTTT	668	537	5910	Antisense
GTAAATGCTGCCATGTTCCAAACCC	571	453	5999	Antisense
GTGTTTAGAGCTGTGCACCCTAGAA	327	477	6037	Antisense
ATTATGCCAGTTTCTGTTCTCTCAC	159	11	6158	Antisense
TTTTTGTGCACTACATACTCTTCAG	221	681	6199	Antisense
GATTAATATGCCCTTTTGCCGATGC	5	429	6277	Antisense
TACTGATGTGACTCGGTTTTGTCGC	4	653	6309	Antisense
TTTTGTCGCAGCTTTGCTTTGTTTA	684	681	6325	Antisense
CACACTTGTAAACCTCTTTTGCACT	130	177	6356	Antisense
GATGCTCGAGCACCTGTAAACAATT	161	419	6400	Antisense



HG-U133:205225_at
Gene:ESR1
estrogen receptor 1

Affymetrix web site

Affymetrix Probes



mRNA reference sequence

5'

3'

Spaced DNA probe pairs

Reference sequence

··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC ···

AATGGGTCAGAAGGACTCCTATGTGGGTG Perfect Match Oligo

AATGGGTCAGAACGACTCCTATGTGGGTG Mismatch Oligo

(25-mer)

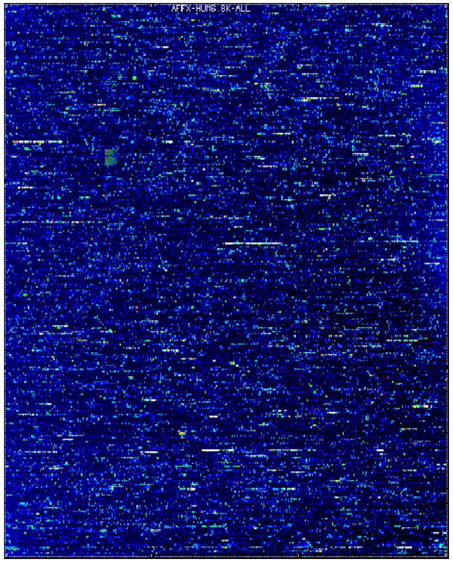
Perfect match probe cells

Fluorescence Intensity Image

Mismatch probe cells

Affymetrix.com

"Low Level" Data Processing



Scanning

Background noise - corrections

Gross errors, defects

Gene probe sets:

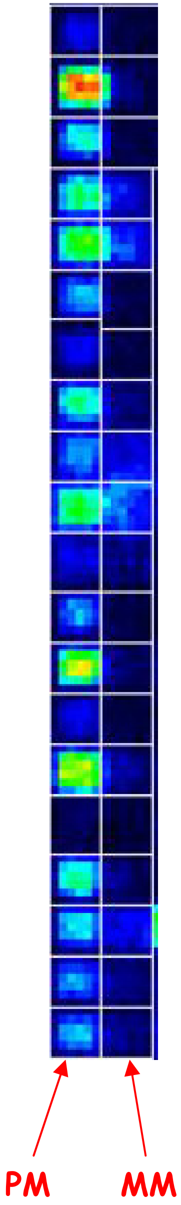
Perfect Match & MisMatch

Probe intensities: 75%tile of central pixels

Probe effects evident

Cross-hybridization: MM signals

Overall mRNA expression estimate?

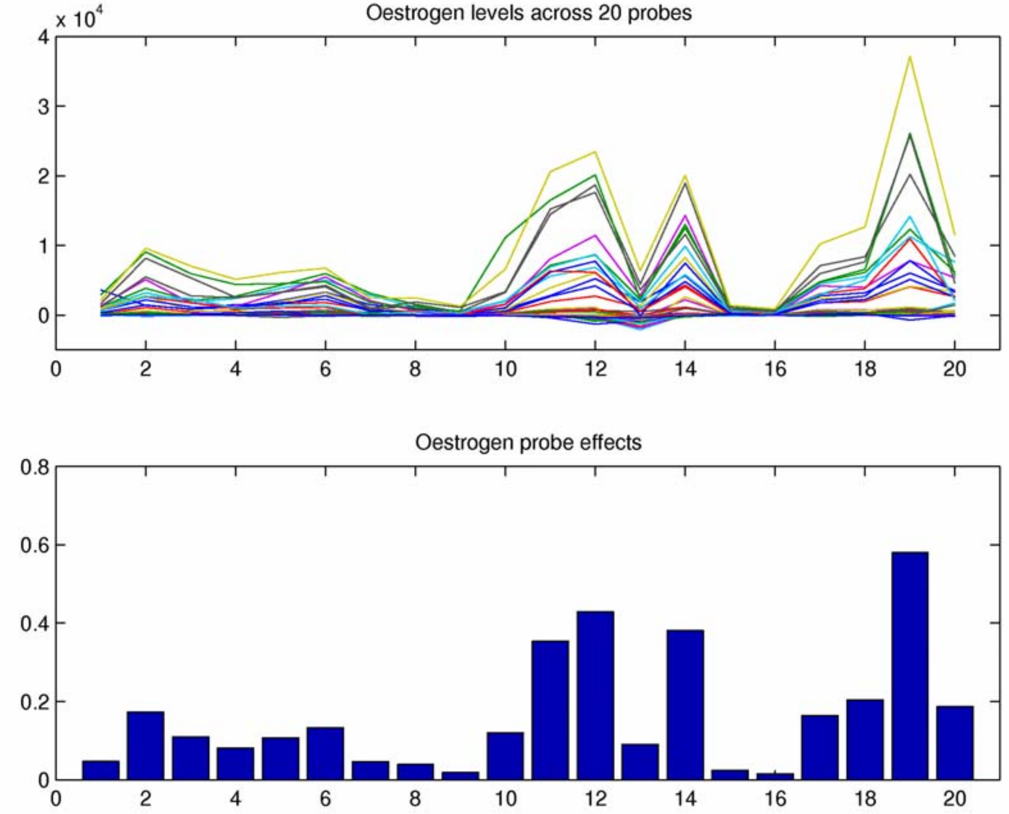


Affymetrix web site

Rafael Irizarry web site: Affy package for R

Bioconductor R software

Oligo Array Data: Probe Effects



ER gene

20 probes
Early HU6800 array

50 breast tumours

Concordant patterns

Empirical estimates
of probe effects

Affymetrix Array Data


Duke0145
2003-11-21
Affy Image
Data

=*.DAT file, image file of chip in Affymetrix software format 43,754 KB


Duke0115
2003-07-18
Affy
Standard
Output

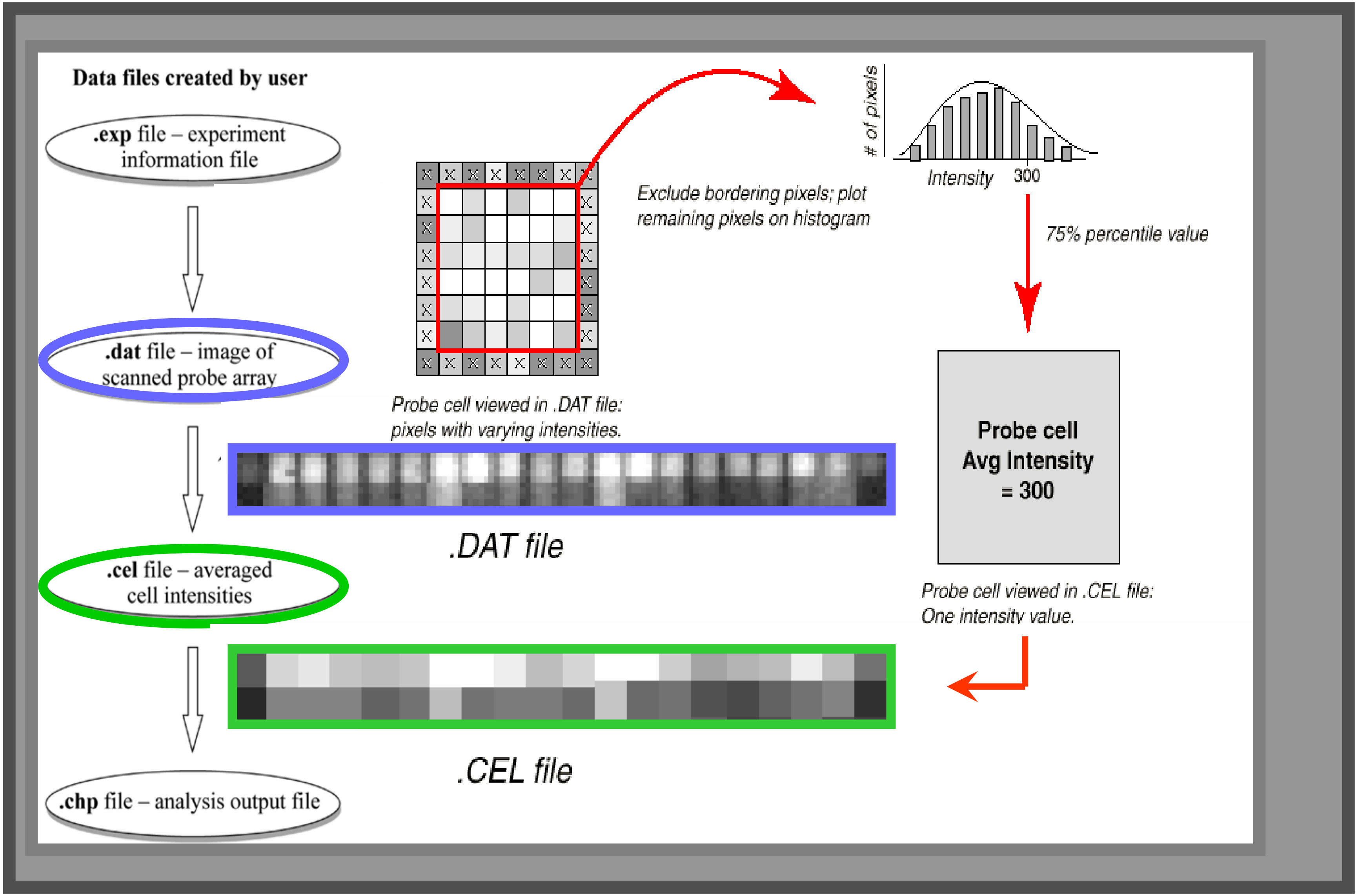
= *.CEL file, contains individual probe cell measurements 32,148 KB
*.RPT file, contains control information 3 KB
*.CHP file, contains results in Affymetrix software format 11,953 KB
*.EXP file, contains info about wash/stain/scan 2 KB


Duke0115
2003-11-19
EA
Generated
Output

= *_v5a.txt, contains control info with scaling factor for chip 1 KB
*_v5p.txt, contains raw data for chip 14,005 KB

Average total size of data for one array (one sample) = 101,866 KB

.DAT file and .CEL file



Data files created by user

.exp file – experiment information file

.dat file – image of scanned probe array

.cel file – averaged cell intensities

.chp file – analysis output file

Exclude bordering pixels; plot remaining pixels on histogram

75% percentile value

Probe cell viewed in .DAT file: pixels with varying intensities.

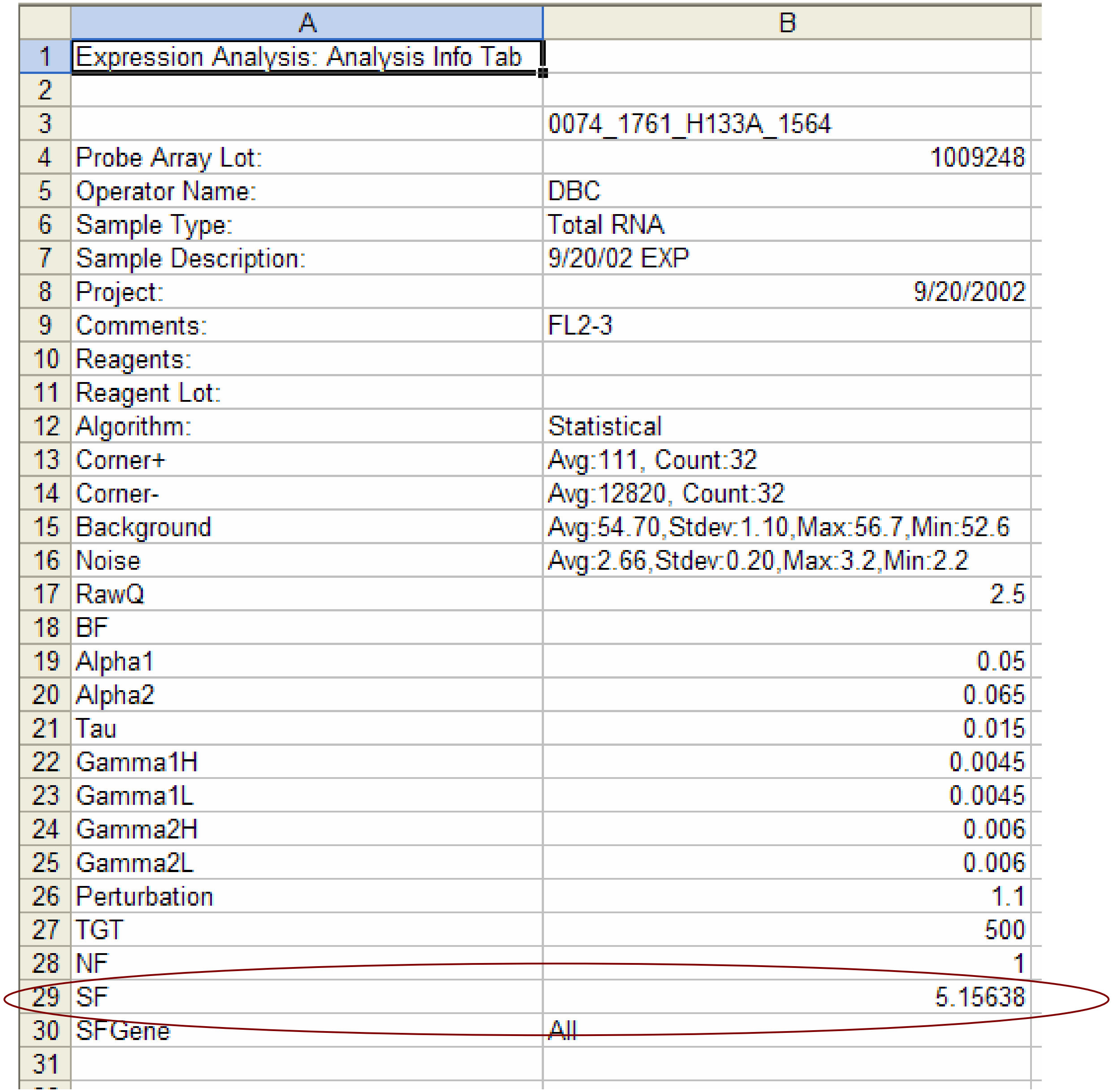
.DAT file

Probe cell Avg Intensity = 300

Probe cell viewed in .CEL file: One intensity value.

.CEL file

*_v5a.txt



	A	B
1	Expression Analysis: Analysis Info Tab	
2		
3		0074_1761_H133A_1564
4	Probe Array Lot:	1009248
5	Operator Name:	DBC
6	Sample Type:	Total RNA
7	Sample Description:	9/20/02 EXP
8	Project:	9/20/2002
9	Comments:	FL2-3
10	Reagents:	
11	Reagent Lot:	
12	Algorithm:	Statistical
13	Corner+	Avg:111, Count:32
14	Corner-	Avg:12820, Count:32
15	Background	Avg:54.70,Stdev:1.10,Max:56.7,Min:52.6
16	Noise	Avg:2.66,Stdev:0.20,Max:3.2,Min:2.2
17	RawQ	2.5
18	BF	
19	Alpha1	0.05
20	Alpha2	0.065
21	Tau	0.015
22	Gamma1H	0.0045
23	Gamma1L	0.0045
24	Gamma2H	0.006
25	Gamma2L	0.006
26	Perturbation	1.1
27	TGT	500
28	NF	1
29	SF	5.15638
30	SFGene	All
31		

Methods & Models: Basics

background correct PM, MM

ave(PM-MM)

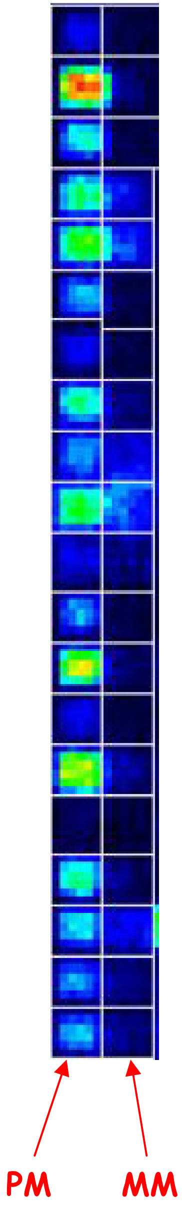
robust ave(PM-MM)

log scale (base 2 – fold change)

Affyx MAS 5.0 Signal:

robust ave (log (PM-mm))

mm=f(MM,background)



Affymetrix web site

Rafael Irizarry web site: Affy package for R

Bioconductor R software

Methods & Models: Normalization

Comparability across samples

Variability in process: Normalize

Affyx scaling: Scale Factor

Average level of set of genes same across n samples

SF as a crude quality metric

Linear scaling

NCI Genomics & Bioinformatics web site

Affymetrix web site

Rafael Irizarry web site: Affy package for R

Bioconductor R software

Nonlinear Normalization:

Match some or all quantiles across samples

Match selected quantiles to median across samples

Match empirical CDF of each chip to median chip

- West et al PNAS 2001
- Wong & Li (dChip) 2001
- Speed, Bolstad, Irizarry (2001-now)
- Bioconductor site

Methods & Models: Normalization

Comparability across samples

Variability in process: Normalize

Affyx scaling: Scale Factor

Average level of set of genes same across n samples

SF as a crude quality metric

Linear scaling

NCI Genomics & Bioinformatics web site

Affymetrix web site

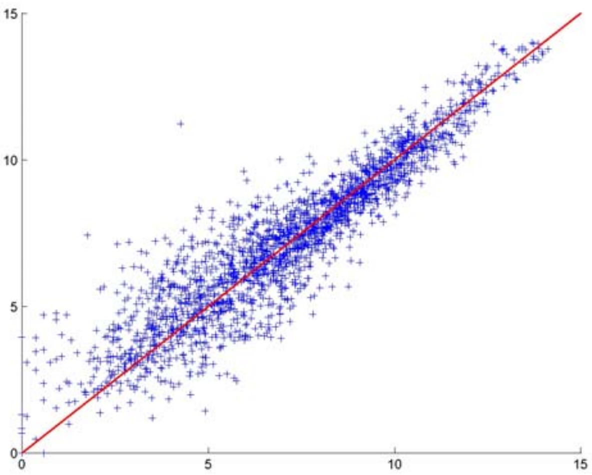
Rafael Irizarry web site: Affy package for R

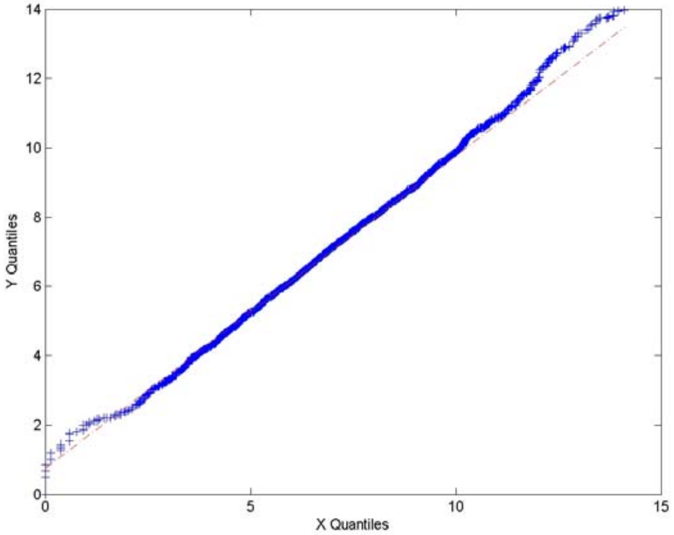
Bioconductor R software

Nonlinear Normalization:

WHY?

- West et al PNAS 2001
- Wong & Li (dChip) 2001
- Speed, Bolstad, Irizarry (2001-now)
- Bioconductor site



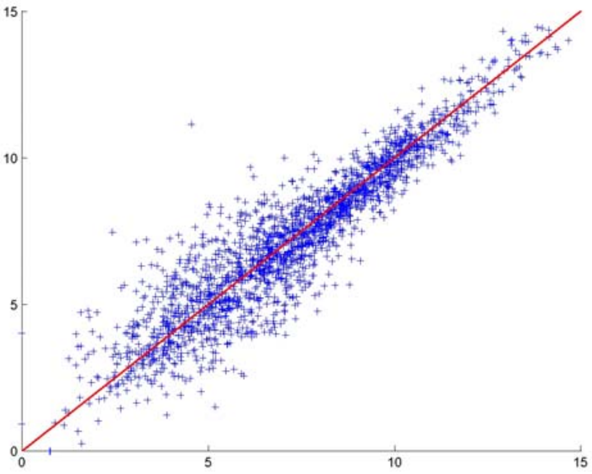


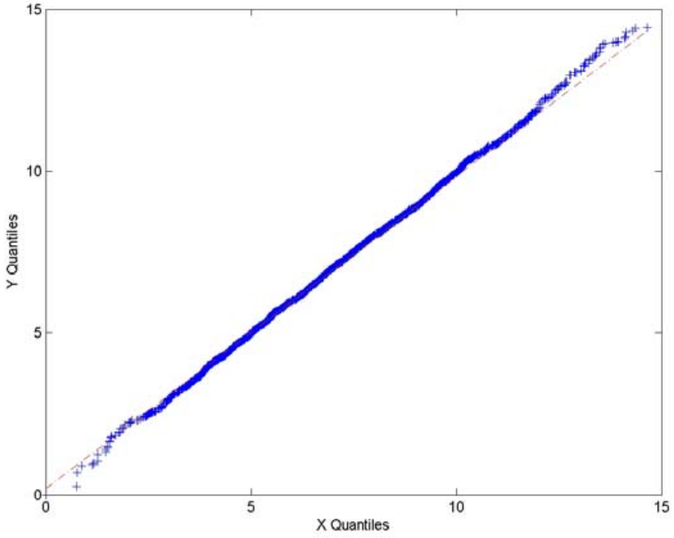
2000 genes
Replicate chips

MAS5.0 Signal
Scaled with SF
(0.52, 0.56)

Mouse tumour data: Huang et al 2004 Nature Genetics

Effects of Quantile Normalization





same genes, chips

quantile normalized

'Low Level' Statistical Models

dChip (Li & Wong 2001+)

$$PM_{ij} - MM_{ij} = E_i\phi_j + \varepsilon_{ij}$$

One gene

Sample(chip) *i*, probe *j*

Expression E on sample i

Expression levels and probe effects

Similar to MAS 5.0

Cross-Hybridization issues

Signal in MM

Under development

Bioconductor

'Low Level' Statistical Models

RMA (Speed, Irizzary, Bolstad 2001+)

$$\log_2(pm_{ij}) = A_i + B_j + \varepsilon_{ij}$$

Background corrected, probe level q-normalized

Robust fitting: outlier probes (10-15%)

(chip) *i*, probe *j*

Expression index *A*

Similar to MAS 5.0

Improved resolution at low levels of expression

Bias at low-moderate levels

Under development

Bioconductor

ABSS04