Copy-number estimation using Robust Multichip Analysis

-

Supplementary materials for the aroma.affymetrix lab session

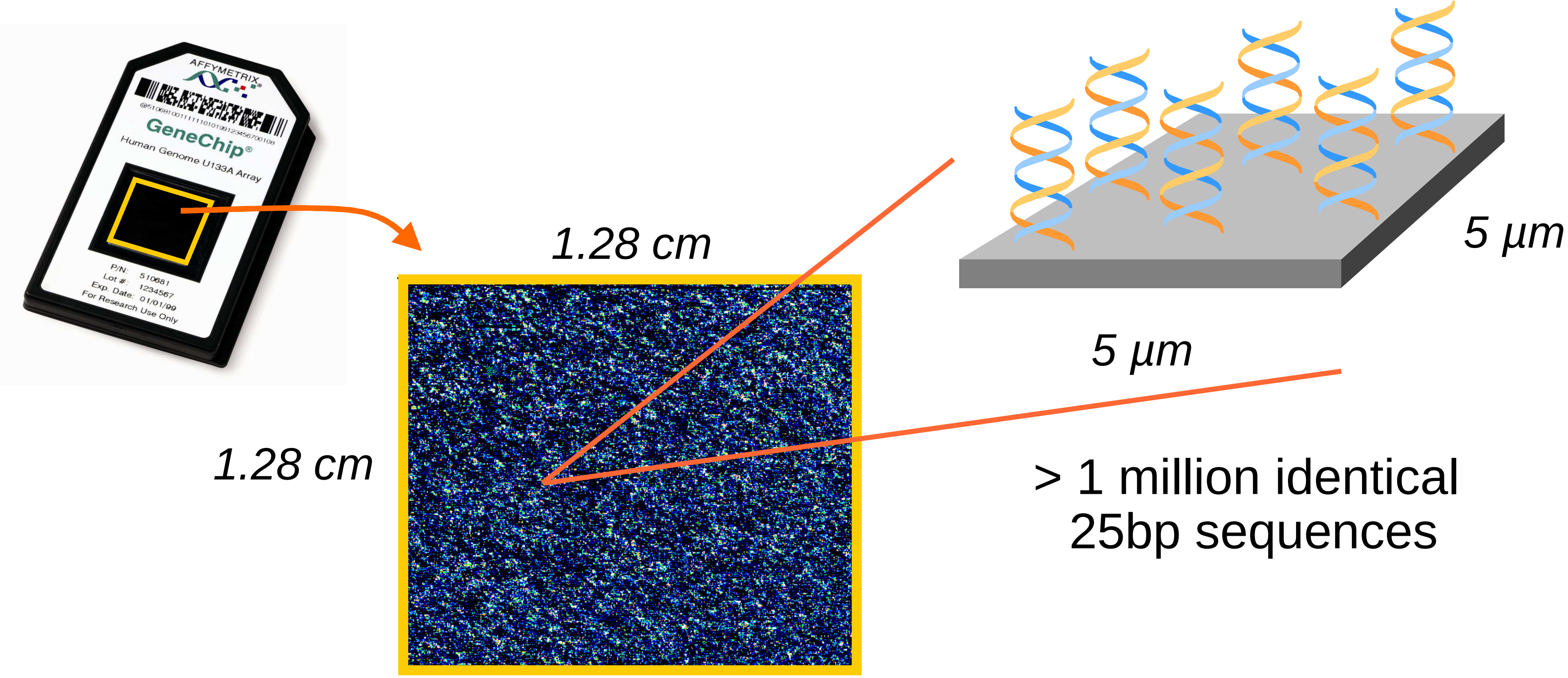
Henrik Bengtsson & Terry Speed
Dept of Statistics, UC Berkeley

August 7, 2007

BioC 2007

Affymetrix chips

Generic Affymetrix chip



1.28 cm

1.28 cm

5 µm

5 µm

> 1 million identical 25bp sequences

6.5 million probes/chip

Feature size: *100µm* to *18µm* to *11µm* and now *5µm.*
Soon: 1µm, 0.8*µm*, with a huge increase in number of probes.

Abbreviated generic assay description

1. Start with target ***gDNA*** (genomic DNA) or ***mRNA*.**

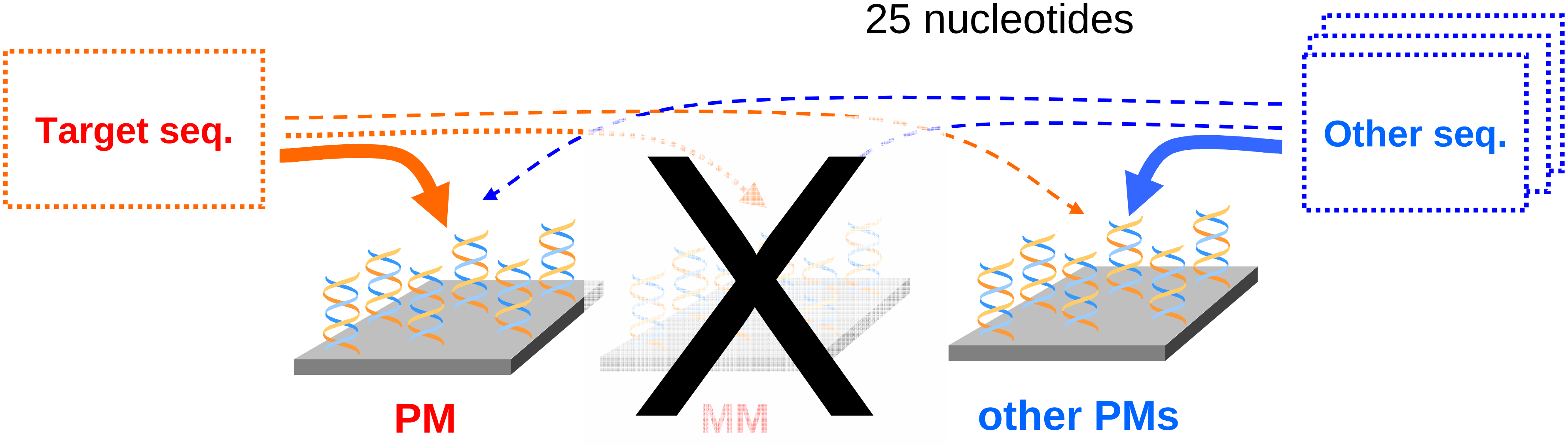
2. Obtain ***labeled single-stranded*** target DNA fragments for hybridization to the probes on the chip.

3. After hybridization, washing, staining and scanning we get a **digital image**. This is summarized across pixels to ***probe-level intensities*** before we begin. They are our **raw data**.

Affymetrix probe terminology

```
Target DNA:          ...CGTAGCCATCGGTAAGTACTCAATGATAG...
                          |||||||||||||||||||||||||
Perfect match (PM):       ATCGGTAGCCATTCATGAGTTACTA
Mis-match (MM):           ATCGGTAGCCATACATGAGTTACTA
```

25 nucleotides



Affymetrix SNP chips

(Mapping 10K, 100K, 500K)

Single Nucleotide Polymorphism (SNP)

Definition:
A sequence variation such that two chromosomes may differ by a single nucleotide (A, T, C, or G).

Allele A: ***A***

...CGTAGCCATCGGTA/GTACTCAATGATAG...

Allele B: ***G***

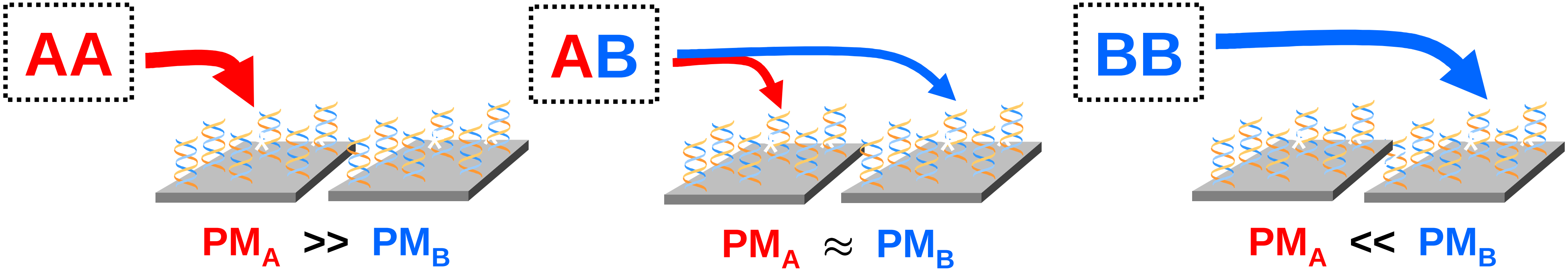
A person is either AA, AB, or BB at this SNP.

Probes for SNPs

PM_A: **ATCGGTAGCCATTCATGAGTTACTA**
Allele A: ***...CGTAGCCATCGGTAAGTACTCAATGATAG...***

Allele B: ***...CGTAGCCATCGGTAGGTACTCAATGATAG...***
PM_B: **ATCGGTAGCCATCCATGAGTTACTA**

(Also MMs, but not in the newer chips, so we will not use these!)



AA: PM_A >> PM_B

AB: $PM_A \approx PM_B$

BB: PM_A << PM_B

Copy-number analysis with SNP arrays

Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (or quantile)
Total CN	$PM = PM_A + PM_B$
Summarization *(SNP signals θ)*	log-additive PM only
Post-processing	fragment-length (GC-content)
Raw total CNs *R = Reference*	$M_{ij} = \log_2(\theta_{ij} / \theta_{Rj})$ chip *i*, probe *j*

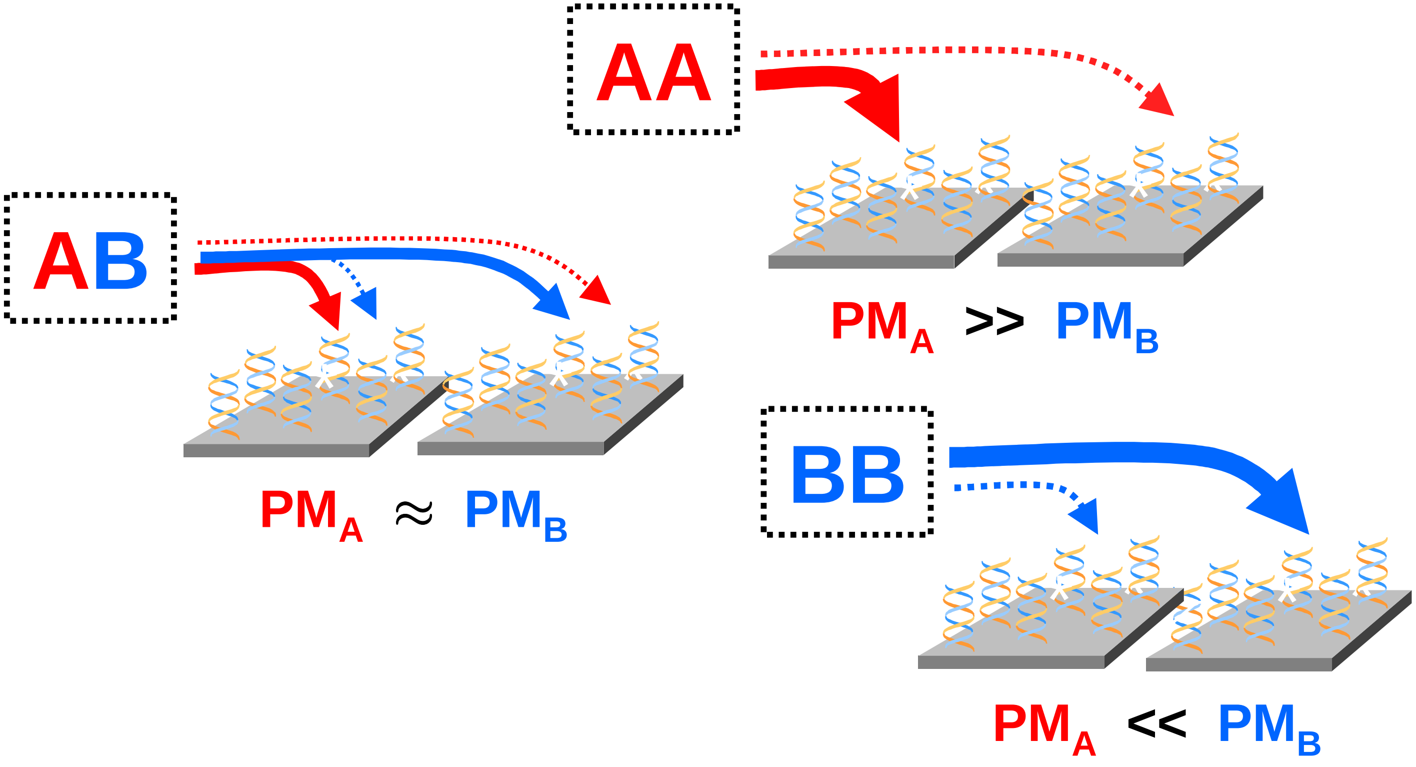
Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log (PM
Post-processing	frag (GC
Raw total CNs	M_{ij}

Cross-hybridization:

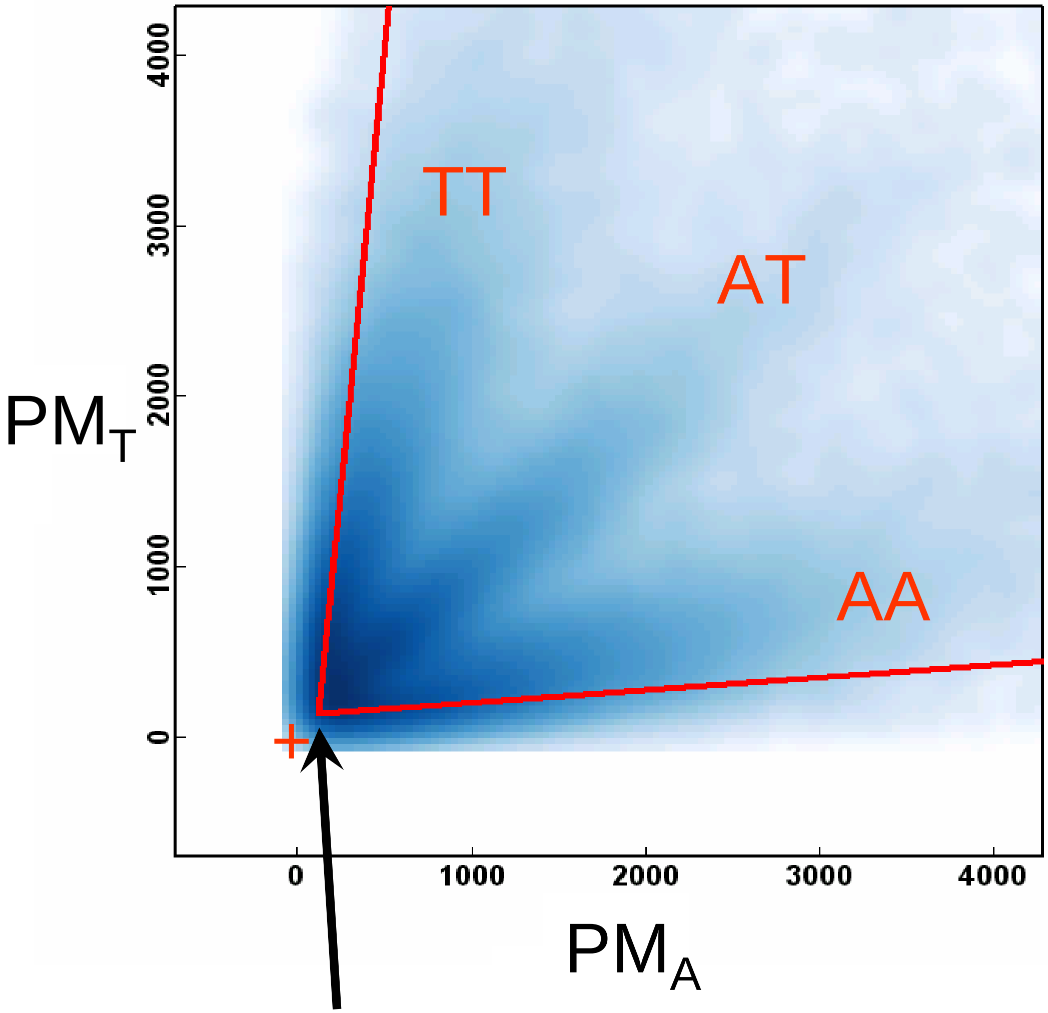
Allele A: ***TCGGTAAGTACTC***

Allele B: ***TCGGTATGTACTC***



Copy-number estimation using Robust Multichip Analysis (CRMA)

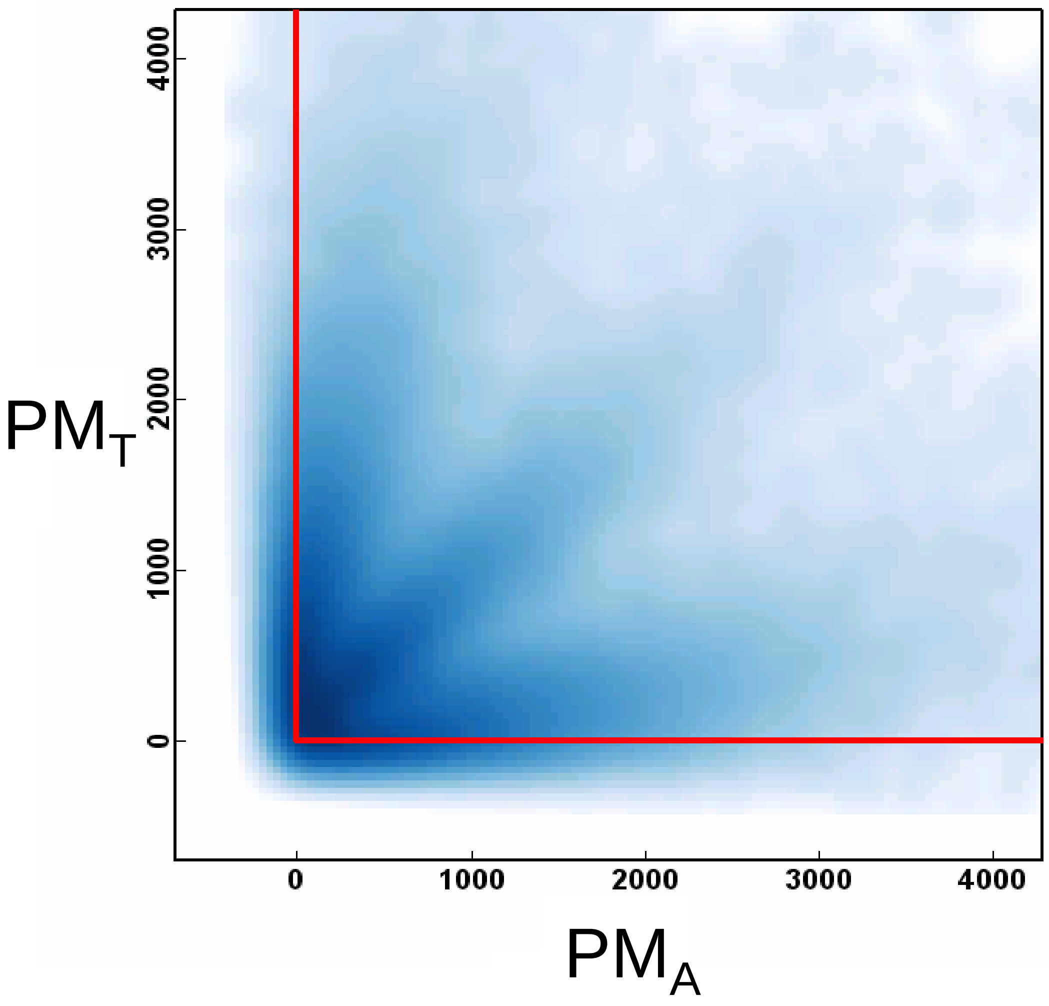
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$



offset

Copy-number estimation using Robust Multichip Analysis (CRMA)

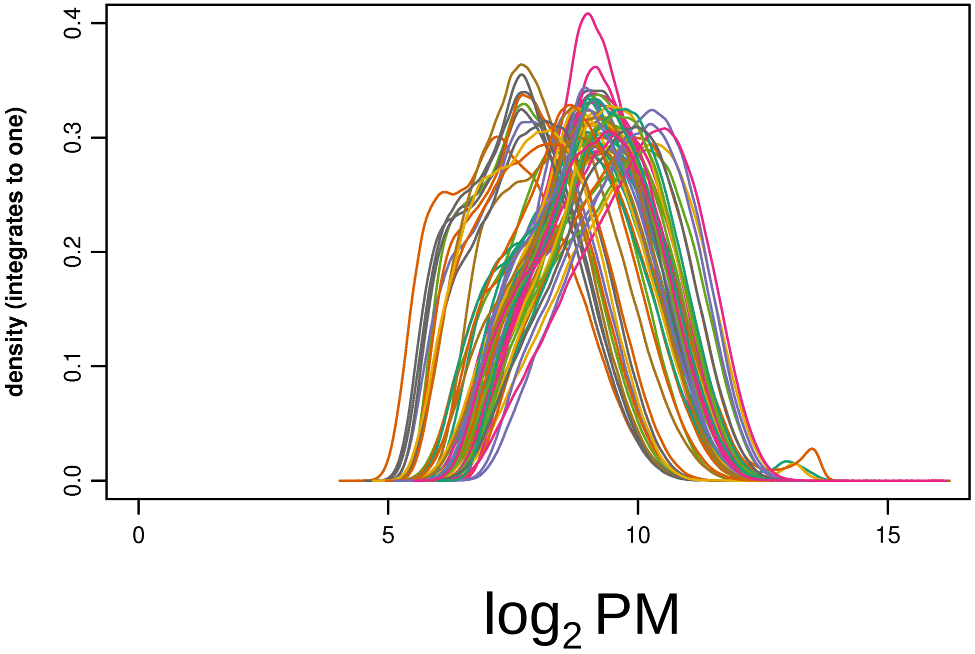
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM = PM_A + PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$



Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

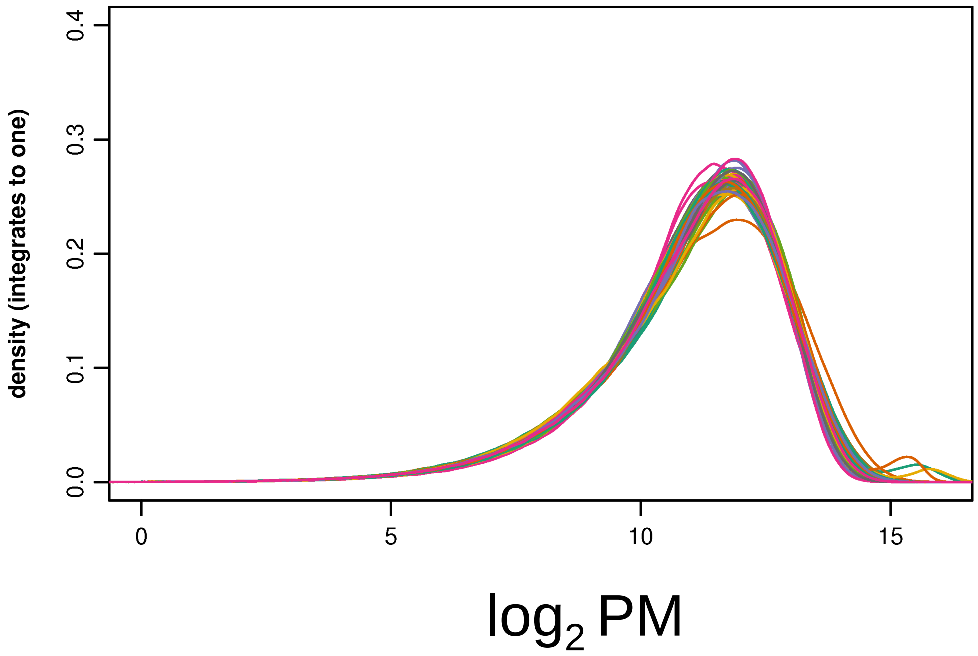
Crosstalk calibration corrects for differences in distributions too



Copy-number estimation using Robust Multichip Analysis (CRMA)

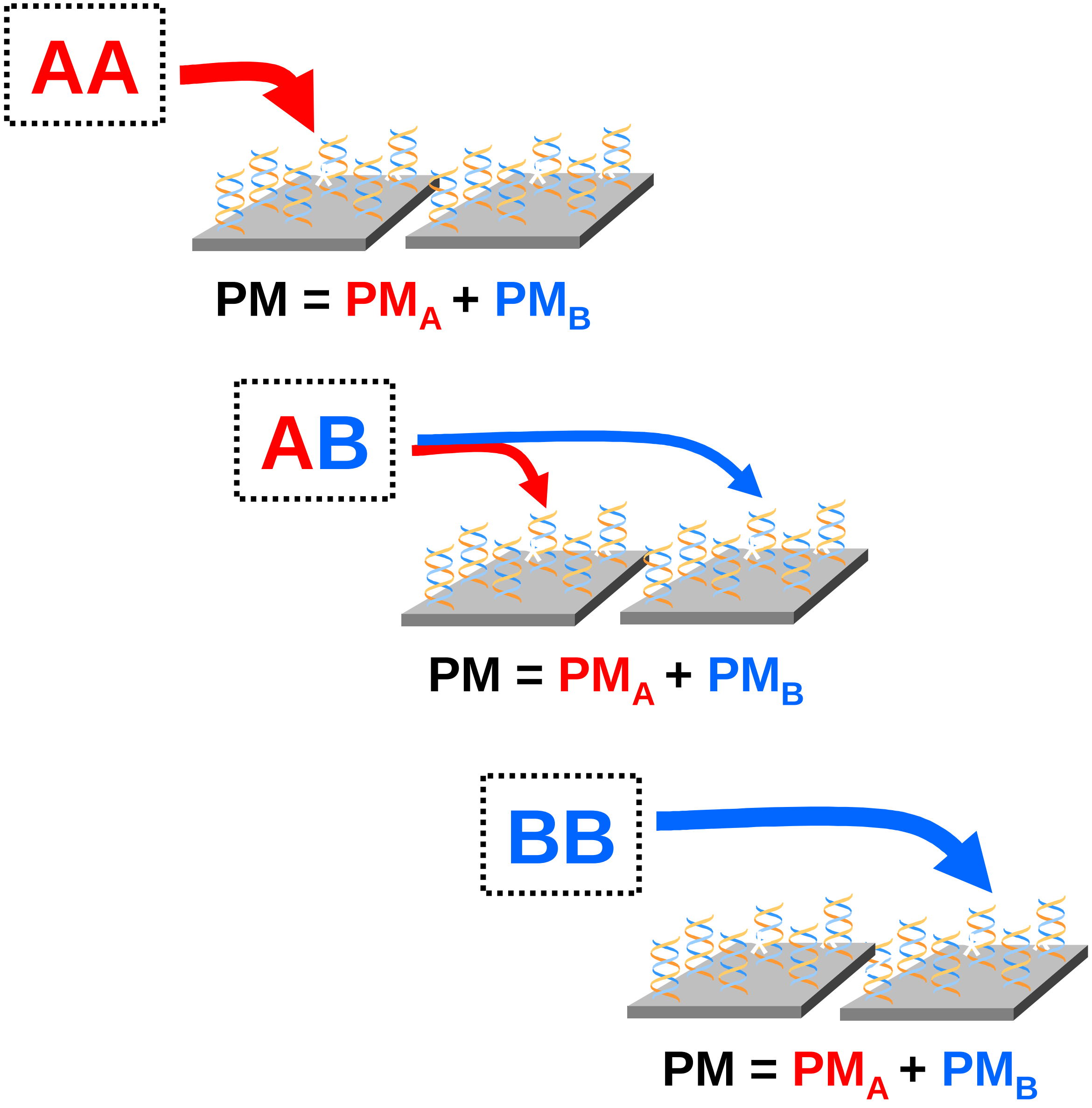
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

Crosstalk calibration corrects for differences in distributions too



Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$



PM = PM$_A$ + PM$_B$

PM = PM$_A$ + PM$_B$

PM = PM$_A$ + PM$_B$

Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

The log-additive model:

$$\log_2(PM_{ijk}) = \log_2 \theta_{ij} + \log_2 \phi_{jk} + \varepsilon_{ijk}$$

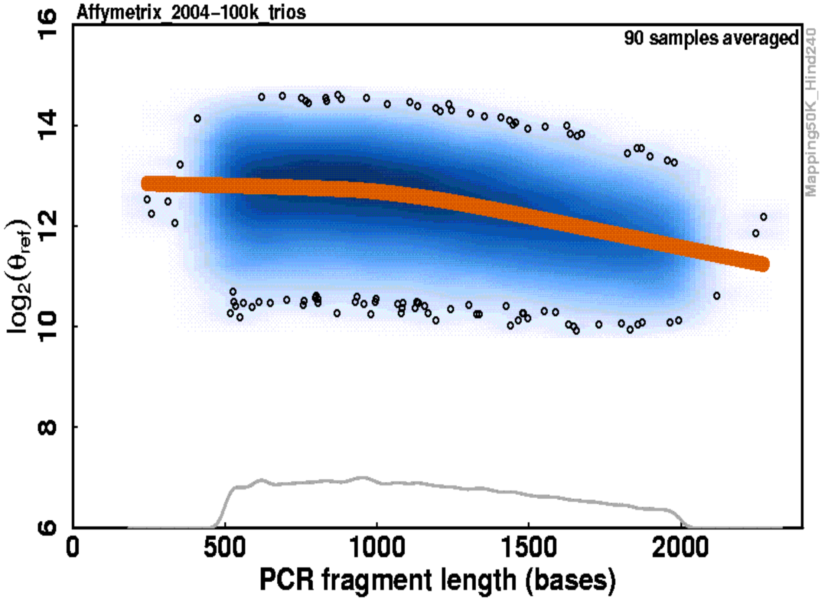
sample *i*, SNP *j*, probe *k*.

Fit using robust linear models (rlm)

Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	PM=PM$_A$+PM$_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

Longer fragments ⇒
less amplified by PCR ⇒
weaker SNP signals θ

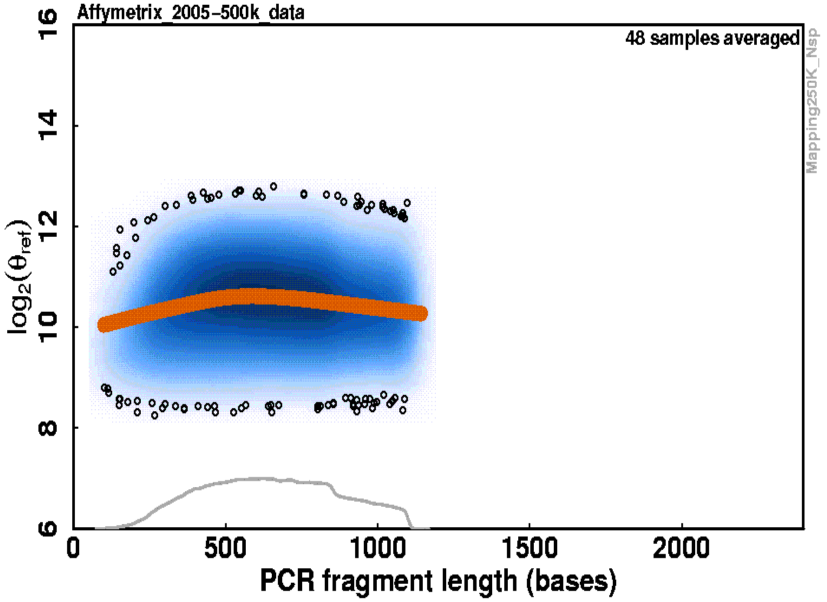


100K

Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

Longer fragments ⇒
less amplified by PCR ⇒
weaker SNP signals θ

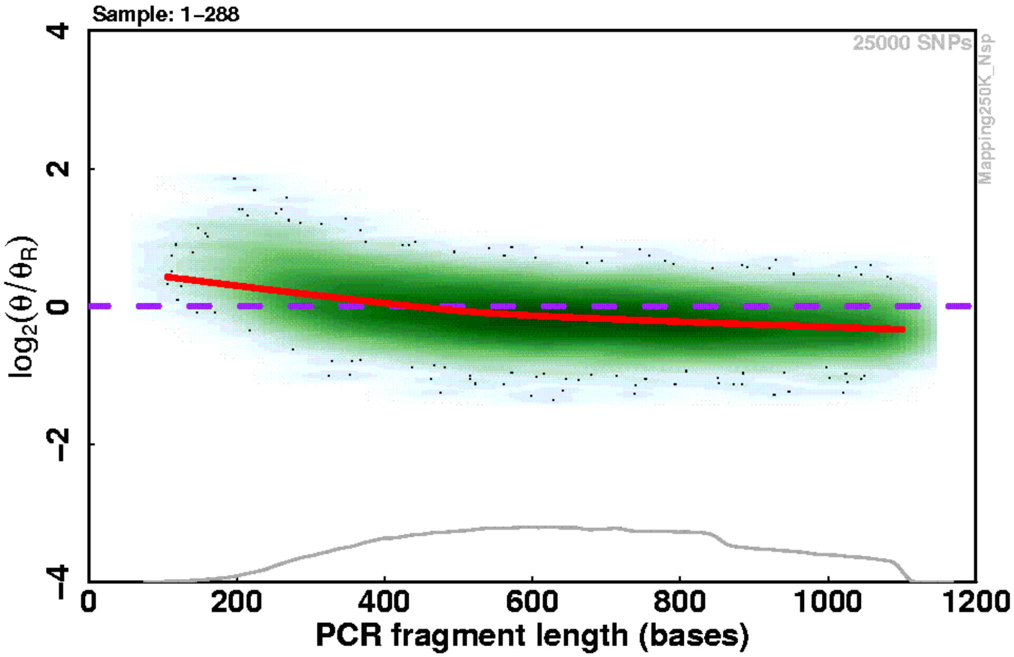


500K

Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	PM=PM$_A$+PM$_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

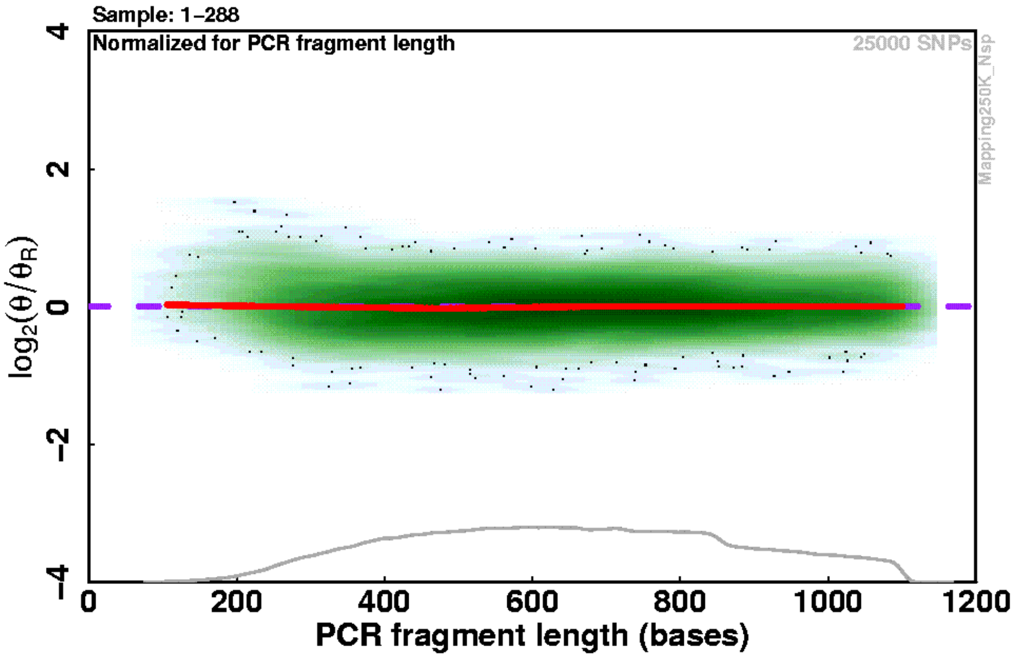
Normalize to get same fragment-length effect for all hybridizations



Copy-number estimation using Robust Multichip Analysis (CRMA)

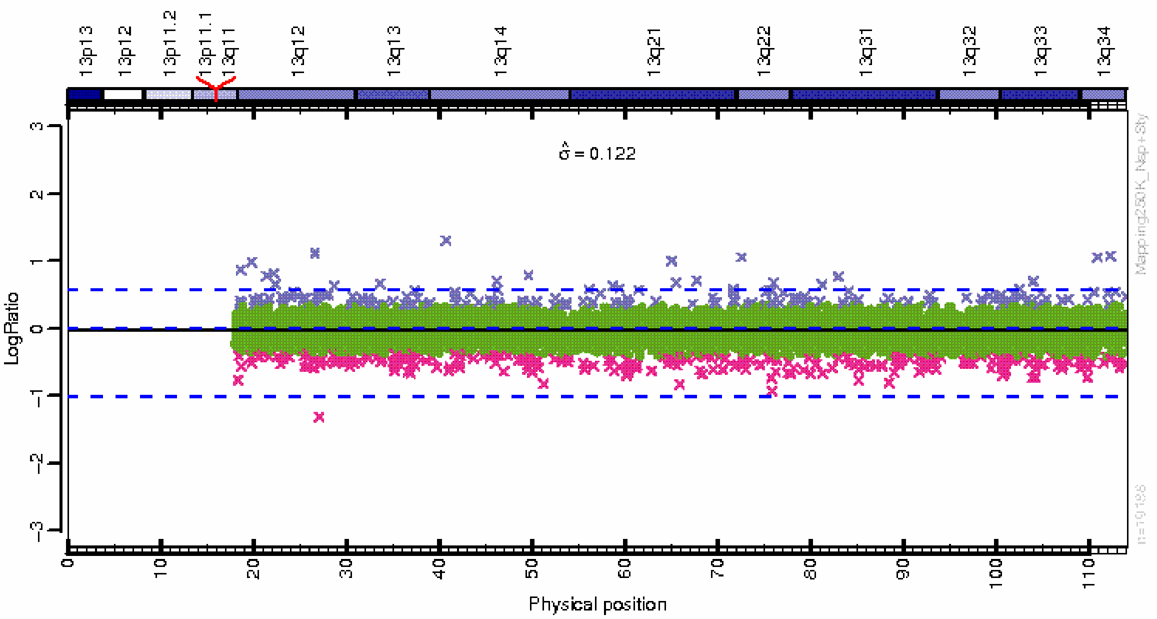
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

Normalize to get same fragment-length effect for all hybridizations



Copy-number estimation using Robust Multichip Analysis (CRMA)

	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$


13p13
13p12
13p11.2
13p11.1
13q11
13q12
13q13
13q14
13q21
13q22
13q31
13q32
13q33
13q34
$\hat{\sigma}$ = 0.122
Mapping250K_Nsp+Sty
n=10196
LogRatio
Physical position