Microarray data analysis: SNP arrays applied to calculate Copy Number alterations in cancer samples

EuGESMA

Bioinformatics Training School

23-24.March.2010

Celia FONTANILLO and Javier DE LAS RIVAS

Grupo de Investigación Bioinformática y Genómica Funcional

Centro de Investigación del Cáncer

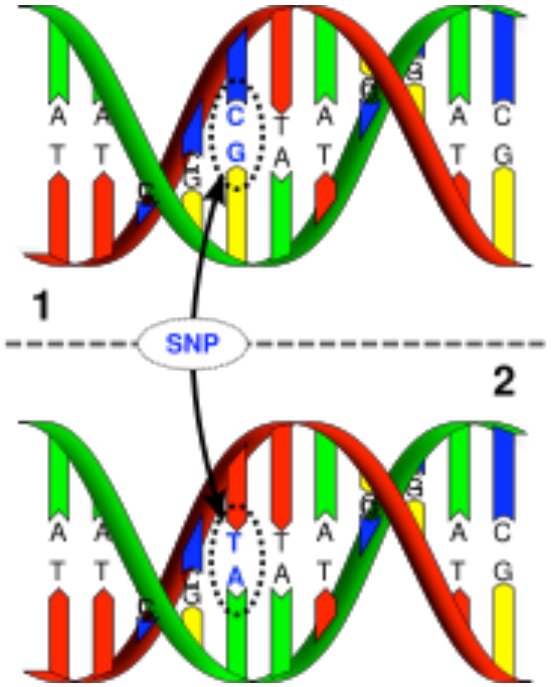
(CIC, CSIC-USAL, Salamanca, España)

SNP Microarrays

- What are SNPs?
- SNPs Databases
- SNPs detection Platforms
- SNPs microarrays: Affymetrix
- SNPs microarray Data Analysis
- aroma.affymetrix
  - Genotyping
  - Copy Number analysis

What are SNPs?

SNP (pronounced “snip”) :
Single Nucleotide Polimorfism



- Definition: Variations in **single base pairs** that are randomly dispersed throughout the genome (every 100 to 300 bases along the 3-billion-base human genome)
- Point mutations stablished in at least **>1% of a given population**
- Act as measures of **genetic diversity** within the specie (i.e. 90% of human genetic variation)
- SNPs can occur in both **coding** (genes) and **non-coding regions** of the genome
- **Many** SNPs have **no effect** on cell function, but **others** could predispose people to disease or influence their response to a drug or other factor

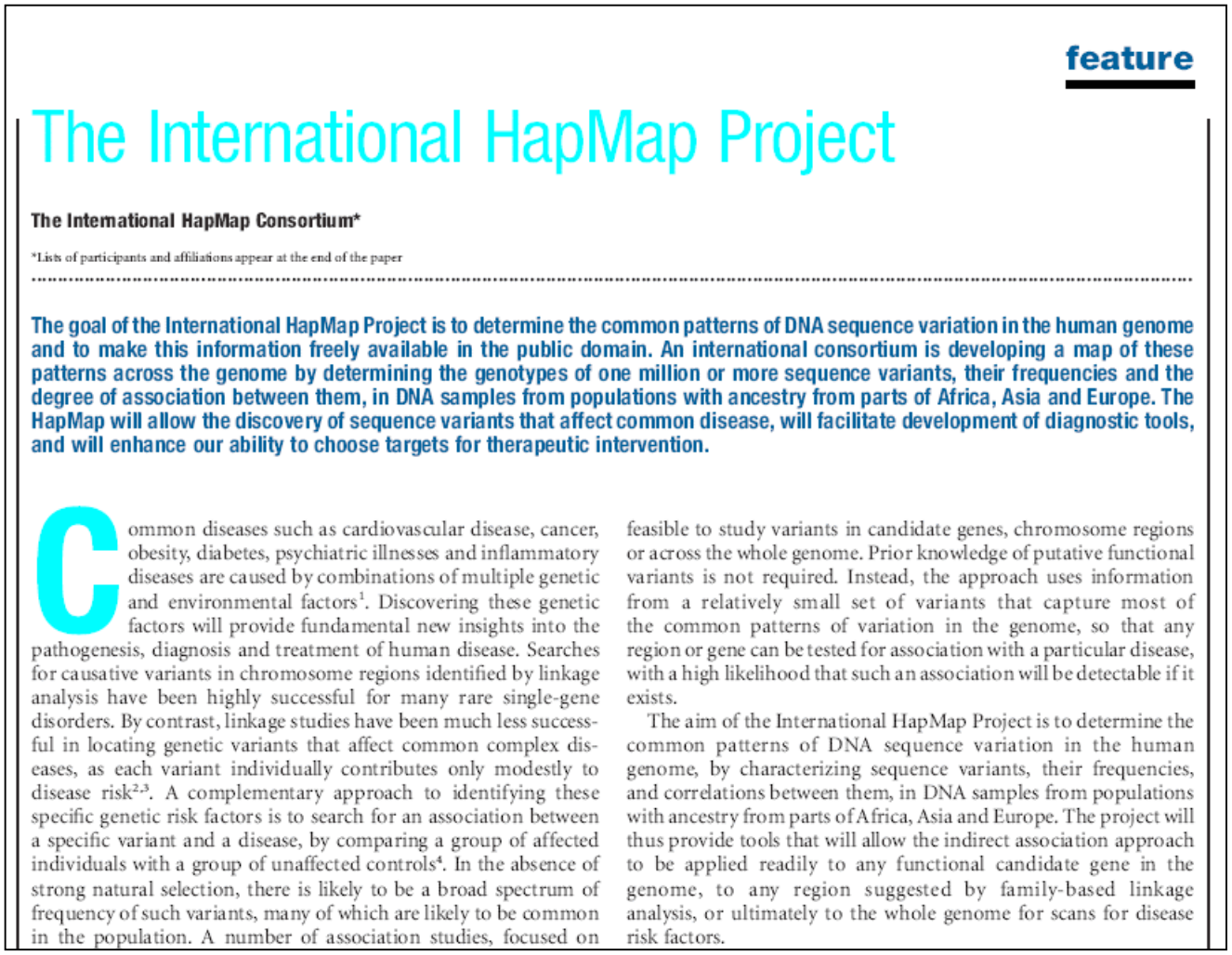
Some SNPs Databases

- **HapMap** (http://hapmap.org/) :
  - International project involving scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria,and the United States. It is the reference database in many SNP studies

- **dbSNP** (http://www.ncbi.nlm.nih.gov/projects/SNP/) :
  - The largest SNP database and is hosted at the National Center for Biotechnology Information

- **SNPSeek** (http://snp.wustl.edu/cgi-bin/SNPseek/index.cgi) :
  - Has more than 90000 coding genes in exons of known genes
- **SNP500Cancer** (http://snp500cancer.nci.nih.gov/home_1.cfm) :
  - Dedicated to the identification, validation and caracterization of polimorphism in cancer related genes
- **Ensembl** (http://www.ensembl.org/) :
  - SNPView
- **SNPedia** (http://www.snpedia.com/index.php?title=SNPedia)

SNPs: HapMap project (2003...)



feature

The International HapMap Project

The International HapMap Consortium*

*Lists of participants and affiliations appear at the end of the paper

The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.

Common diseases such as cardiovascular disease, cancer, obesity, diabetes, psychiatric illnesses and inflammatory diseases are caused by combinations of multiple genetic and environmental factors[1]. Discovering these genetic factors will provide fundamental new insights into the pathogenesis, diagnosis and treatment of human disease. Searches for causative variants in chromosome regions identified by linkage analysis have been highly successful for many rare single-gene disorders. By contrast, linkage studies have been much less successful in locating genetic variants that affect common complex diseases, as each variant individually contributes only modestly to disease risk[2,3]. A complementary approach to identifying these specific genetic risk factors is to search for an association between a specific variant and a disease, by comparing a group of affected individuals with a group of unaffected controls[4]. In the absence of strong natural selection, there is likely to be a broad spectrum of frequency of such variants, many of which are likely to be common in the population. A number of association studies, focused on feasible to study variants in candidate genes, chromosome regions or across the whole genome. Prior knowledge of putative functional variants is not required. Instead, the approach uses information from a relatively small set of variants that capture most of the common patterns of variation in the genome, so that any region or gene can be tested for association with a particular disease, with a high likelihood that such an association will be detectable if it exists.

The aim of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome, by characterizing sequence variants, their frequencies, and correlations between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The project will thus provide tools that will allow the indirect association approach to be applied readily to any functional candidate gene in the genome, to any region suggested by family-based linkage analysis, or ultimately to the whole genome for scans for disease risk factors.

HapMap: Large project to identify SNPs in humans. Started 2003. Now is a catalog of **common genetic variants** that occur in human beings.

It describes **what** these variants are, **where** they occur in our DNA, and **how often** they are distributed among people within populations and among populations in different parts of the world.

URL:
http://www.hapmap.org/

SNPs: HapMap project (2003...)

Normal individuals genotyped (**270**)
(diff. labs & various technologies):

- **90** CEU individuals (Utah/Europe, 30 trio families)
- **90** YRI individuals (Nigeria; 30 trio families)
- **45** CHB (China; unrelated)
- **45** JPT (Japan; unrelated)

Publicly available:

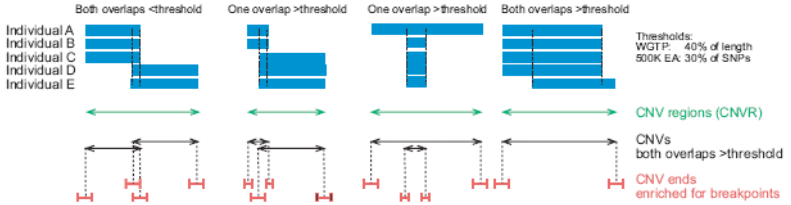
- High quality data.
- Raw data, e.g. Affymetrix CEL files.
- Genotypes.
- Studied by many groups.

SNPs: mapping different bio-types

Not all the **polymorphism** have the same *biological strength*.
Different types of SNPs with different **biological meaning**:

1. SNPs that map just on **genome** non-transcribed/non-gene sites.
2. SNPs that map on **gene regulatory** related **sites**
   (e.g. promoters, etc).
3. SNPs that map on **transcribed locus** corresponding to **ncRNAs**
   (e.g. miRNAs, snRNAs, etc).
4. SNPs that map on **gene locus** (**mRNA**) on non-coding **exons**
   (e.g. UTRs).
5. SNPs that map on **gene locus** (**mRNA**) on **protein-coding** exons
   (nucleotides with synomin effect = do not affect protein aa seq.).
6. SNPs that map on **gene locus** (**mRNA**) on **protein-coding** exons
   (nucleotides with non-synomin effect = affect the **protein aa seq.**).

CNVs: common Copy Number polymorphisms and variants present in normal people (2005...)



Large-Scale Copy Number Polymorphism in the Human Genome

Jonathan Sebat,[1] B. Lakshmi,[1] Jennifer Troge,[1] Joan Al
Janet Young,[2] Pär Lundin,[3] Susanne Månér,[3] Hillary
Megan Walker,[2] Maoyen Chi,[1] Nicholas Navin,[1] Rober
John Healy,[1] James Hicks,[1] Kenny Ye,[4] Andrew Re
T. Conrad Gilliam,[5] Barbara Trask,[2] Nick Patters
Anders Zetterberg,[3] Michael Wigler[1]*

The extent to which large duplications and deletions contribute to hu
variation and diversity is unknown. Here, we show that large-scale
polymorphisms (CNPs) (about 100 kilobases and greater) contribute
to genomic variation between normal humans. Representational oli
microarray analysis of 20 individuals revealed a total of 221 copy nu
ences representing 76 unique CNPs. On average, individuals differed
and the average length of a CNP interval was 465 kilobases. We ob
number variation of 70 different genes within CNP intervals, incl
involved in neurological function, regulation of cell growth, regulatio
olism, and several genes known to be associated with disease.

Many of the genetic differences between humans and other primates are a result of large duplications and deletions (*1–3*). From these observations, it is reasonable to expect that differences in gene copy number could be a significant source of genetic variation between humans. A few examples of large duplication polymorphisms have

In our previous studie
with the use of representa
tide microarray analysis
detected many genomic
deletions in tumor genome
comparison to an unrelat
(*5*), but some of these

nature Vol 444|23 November 2006|doi:10.1038/nature05329

ARTICLES

Global variation in copy number in the human genome

Richard Redon[1], Shumpei Ishikawa[2,3], Karen R. Fitch[4], Lars Feuk[5,6], George H. Perry[7], T. Daniel Andrews[1], Heike Fiegler[1], Michael H. Shapero[4], Andrew R. Carson[5,6], Wenwei Chen[4], Eun Kyung Cho[7], Stephanie Dallaire[7], Jennifer L. Freeman[7], Juan R. González[8], Mònica Gratacòs[8], Jing Huang[4], Dimitrios Kalaitzopoulos[1], Daisuke Komura[3], Jeffrey R. MacDonald[5], Christian R. Marshall[5,6], Rui Mei[4], Lyndal Montgomery[1], Kunihiro Nishimura[2], Kohji Okamura[5,6], Fan Shen[4], Martin J. Somerville[9], Joelle Tchinda[7], Armand Valsesia[1], Cara Woodwark[1], Fengtang Yang[1], Junjun Zhang[5], Tatiana Zerjal[1], Jane Zhang[4], Lluis Armengol[8], Donald F. Conrad[10], Xavier Estivill[8,11], Chris Tyler-Smith[1], Nigel P. Carter[1], Hiroyuki Aburatani[2,12], Charles Lee[7,13], Keith W. Jones[4], Stephen W. Scherer[5,6] & Matthew E. Hurles[1]

Copy number variation (CNV) of DNA sequences is functionally significant but has yet to be fully ascertained. We have constructed a first-generation CNV map of the human genome through the study of 270 individuals from four populations with ancestry in Europe, Africa or Asia (the HapMap collection). DNA from these individuals was screened for CNV using two complementary technologies: single-nucleotide polymorphism (SNP) genotyping arrays, and clone-based comparative genomic hybridization. A total of 1,447 copy number variable regions (CNVRs), which can encompass overlapping or adjacent gains or losses, covering 360 megabases (12% of the genome) were identified in these populations. These CNVRs contained hundreds of genes, disease loci, functional elements and segmental duplications. Notably, the CNVRs encompassed more nucleotide content per genome than SNPs, underscoring the importance of CNV in genetic diversity and evolution. The data obtained delineate linkage disequilibrium patterns for many CNVs, and reveal marked variation in copy number among populations. We also demonstrate the utility of this resource for genetic disease studies.

Genetic variation in the human genome takes many forms, ranging from large, microscopically visible chromosome anomalies to single-nucleotide changes. Recently, multiple studies have discovered an abundance of submicroscopic copy number variation of DNA segments ranging from kilobases (kb) to megabases (Mb) in size[1–9]. Deletions, insertions, duplications and complex multi-site variants[9],

at genes at which other types of mutation are strongly associated with specific diseases: CHARGE syndrome[21] and Parkinson's and Alzheimer's disease[22,23]. Furthermore, CNVs can influence gene expression indirectly through position effects, predispose to deleterious genetic changes, or provide substrates for chromosomal change in evolution[1,9,11,17,24].

SNPs detection Platforms

- Taqman assay
  - Applied Biosystems

    (http://www3.appliedbiosystems.com/AB_Home/index.htm)
- SNPStream assay
  - Orchid Cellmark/Beckman Coulter

    (http://www.beckmancoulter.com/products/instrument/geneticanalysis/ceq/genomelab_snpstream_dcr.asp)
- iPLEX assay
  - Sequenom

    (www.sequenom.com/iPLEX/index2.php)
- GoldenGate genotyping microarray
  - Illumina

    (www.illumina.com/pages.ilmn?ID=11)
- Infinium genotyping microarray
  - Illumina

    (www.illumina.com/pages.ilmn?ID=12)
- GeneChip Human Mapping microarray
  - Affymetrix

    (http://www.affymetrix.com/products/arrays/index.affx)

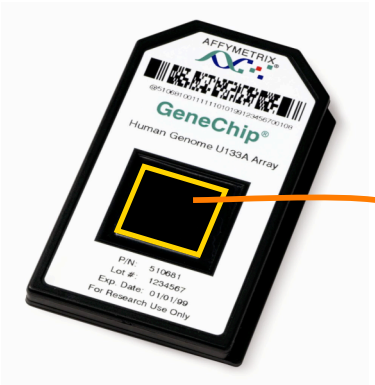
Affymetrix SNP microarrays
(Mapping 10K, 100K, 500K)
(Genome-Wide Human 5.0 and 6.0)

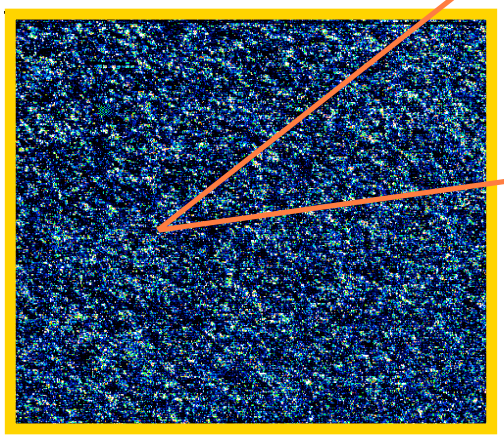
GeneChip® Mapping Assay Kit

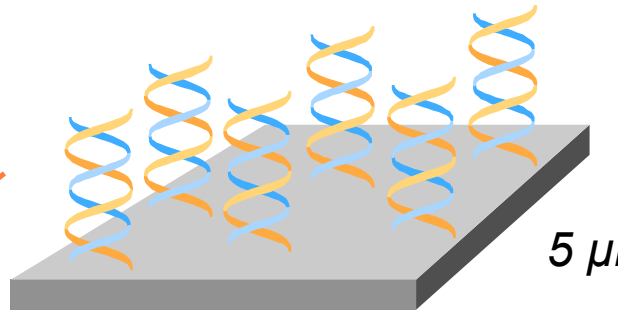
GeneChip® Mapping 100K/500K Set

GCS 3000 7G

Affymetrix microarrays
high-density oligonucleotide chips







1.28 cm

1.28 cm

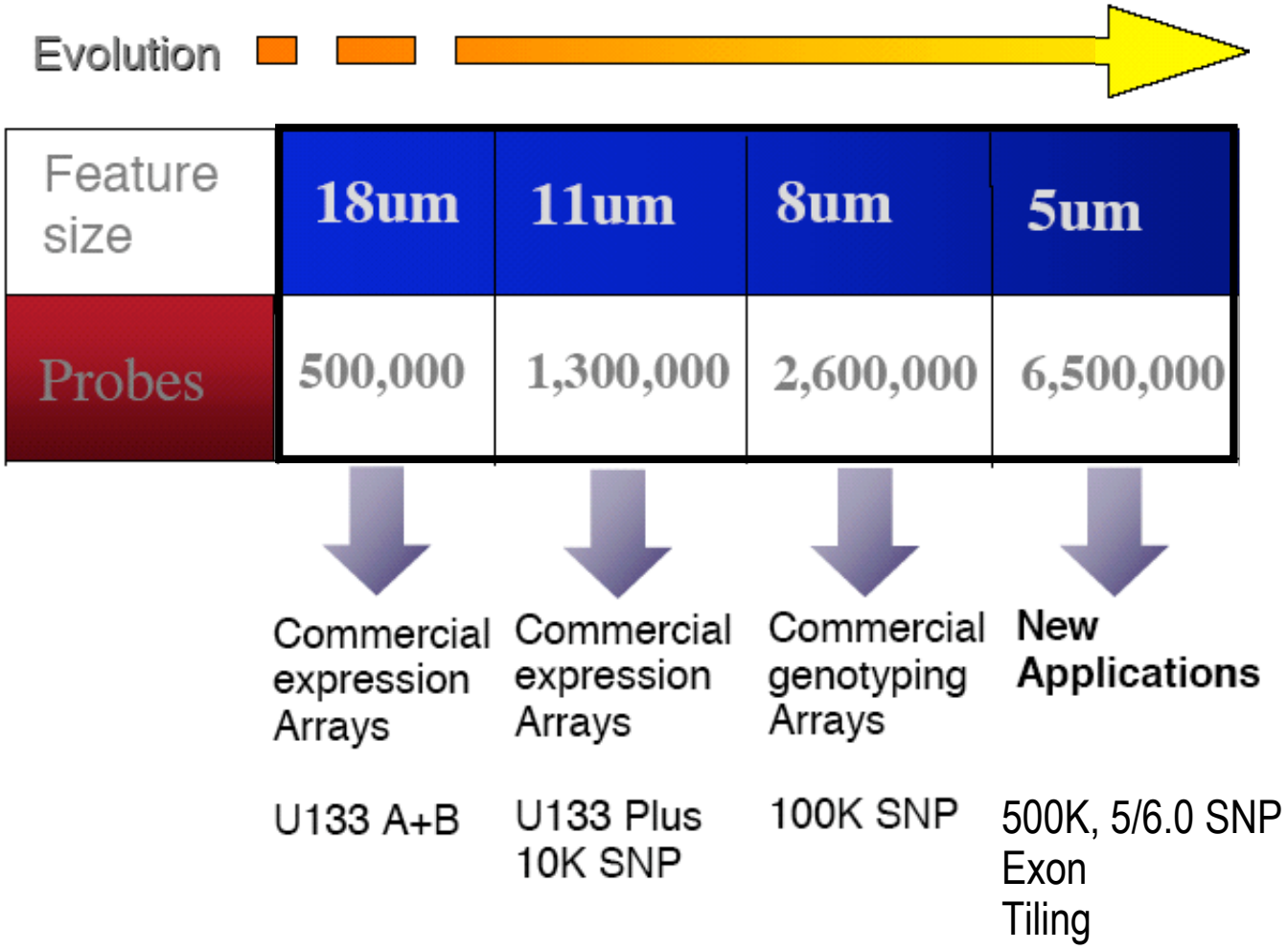
5 µm

5 µm

1 million identical
25-mer sequences

6.5 million probes/ chip

Affymetrix microarrays
great progress on chips capacity



Affymetrix SNP microarrays

- Genome-Wide Human SNP Array 6.0
- Genome-Wide Human SNP Array 5.0
- Mapping 500K Array Set (2 x 250K, Nsp & Sty)
- Mapping 100K Array Set (2 x 50K, Nsp & Sty)
- Mapping 10K 2.0 Array

	10K	100K	500K	5.0	6.0
Released	July 2003	April 2004	Sept 2005	Feb 2007	May 2007
# SNPs	10,204	116,204	500,568	500,568	934,946
# CNPs	-	-	-	340,742	946,371
# loci	10,204	116,204	500,568	841,310	1,878,317
Distance	294kb	25.8kb	6.0kb	3.6kb	1.6kb
Price / chip set	**65 USD**	**400 USD**	**300 USD**	**175 USD**	**300 USD**
# loci / cup of coffee ($1.35)	**116 loci**	**215 loci**	**1236 loci**	**3561 loci**	**4638 loci**

Price source: Affymetrix Pricing Information [http://store.affymetrix.com/] and Berkeley Coffee Shops, Dec 2008.

Affymetrix SNP microarrays
2 generations

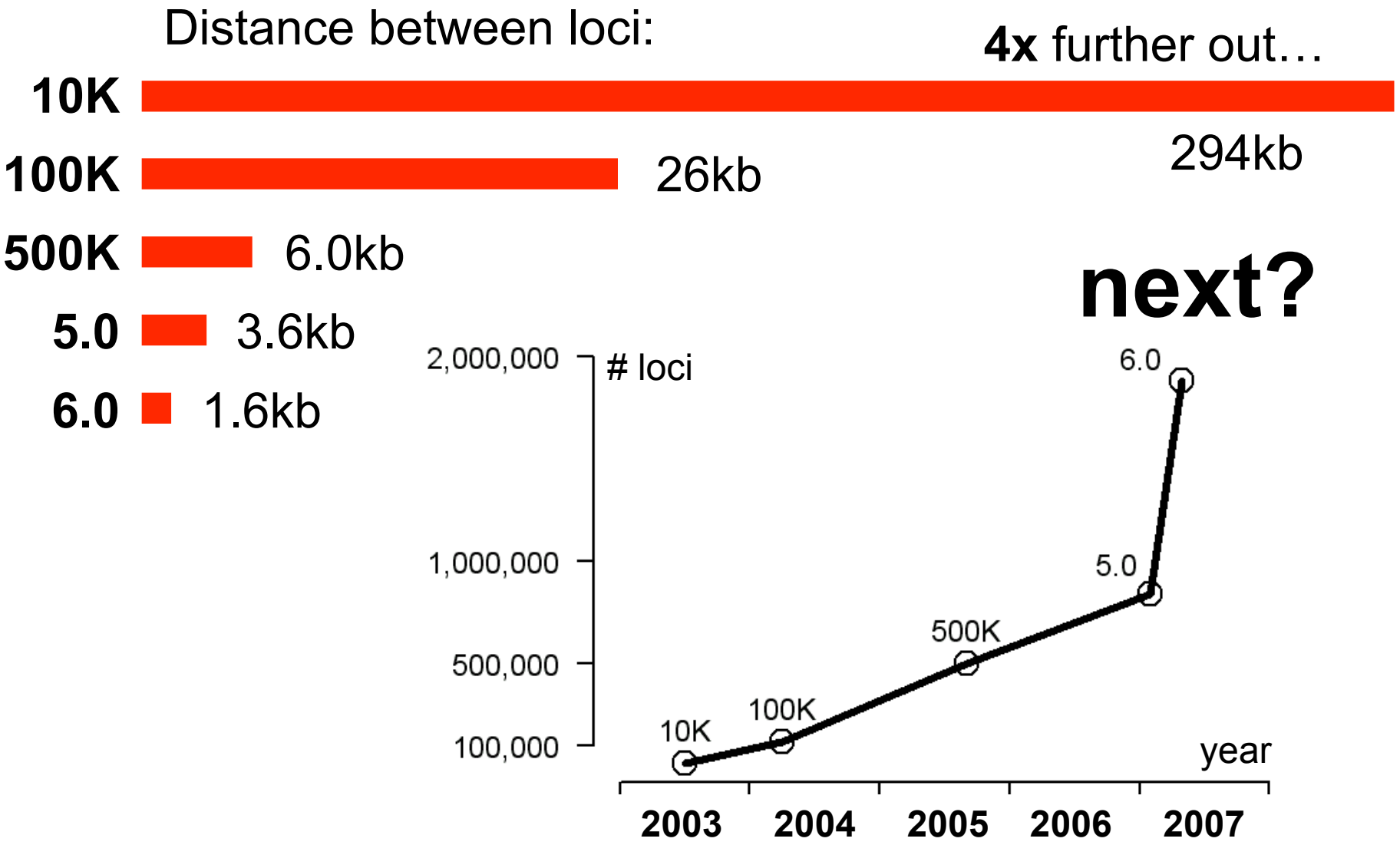
1st generation:
chips interrogating **SNPs**

- Mapping 10K 2.0 Array
- Mapping 100K Array Set (2 x 50K, Nsp & Sty)
- Mapping 500K Array Set (2 x 250K, Nsp & Sty)

2nd generation:
chips interrogating **SNPs** and **CNVs** (Copy Number Variants)

- Genome-Wide Human SNP Array 5.0
- Genome-Wide Human SNP Array 6.0

Affymetrix SNP microarrays
2 generations



Affymetrix SNP microarrays
2 generations

1st generation:
chips interrogating **SNPs**

- Mapping 10K 2.0 Array
- Mapping 100K Array Set (2 x 50K, Nsp & Sty)
- Mapping 500K Array Set (2 x 250K, Nsp & Sty)

2nd generation:
chips interrogating **SNPs** and **CNVs** (Copy Number Variants)

- Genome-Wide Human SNP Array 5.0
- Genome-Wide Human SNP Array 6.0

The *Affymetrix* Genome-Wide Human SNP Array **6.0** features **1.8 million genetic markers**, including **906,600** single nucleotide polymorphisms (**SNPs**) and more than **946,000** probes for the detection of copy number variation (**CNVs**).

Affymetrix SNP microarrays
last generation (6.0)

- > **906,600 SNPs**:
  - Unbiased selection of **482,000 SNPs**: historical SNPs from the SNP Array 5.0 (== 500K)
  - Selection of additional **424,000 SNPs**:
    - Tag SNPs
    - SNPs from chromosomes X and Y
    - Mitochondrial SNPs
    - Recent SNPs added to the dbSNP database
    - SNPs in recombination hotspots

- > **946,000 CNVs** (copy-number probes):
  - **202,000 probes** targeting **5,677 CNV regions** from the *Toronto Database of Genomic Variants*. Regions resolve into **3,182 distinct, non-overlapping segments**; on average 61 probe sets per region.
  - **744,000 probes**, evenly spaced along the genome.

Affymetrix SNP microarrays
2 generations

1st generation:
chips interrogating **SNPs**

- Mapping 10K, 100K, 500K Arrays

1.- set of two arrays: for two enzymes (**Nsp** & **Sty**)
2.- perferct match & mismatch probes: **PM** & **MM**
3.- probe pairs slightly **shifted** relative to each other
4.- probes mapping on both strands of the DNA

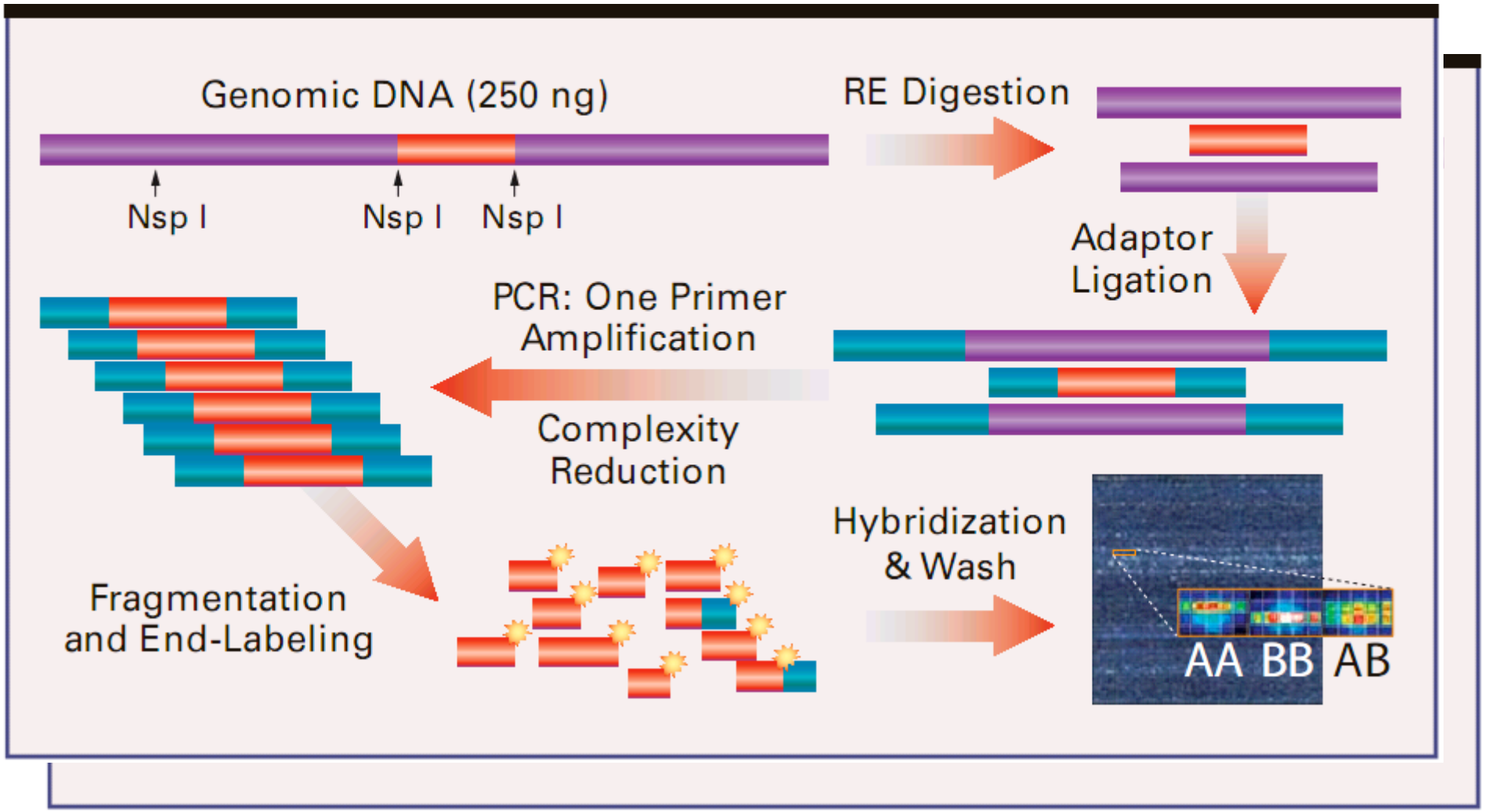
2nd generation:
chips interrogating **SNPs** and **CNVs** (Copy Number Variants)

- Genome-Wide Human SNP Arrays 5.0 / 6.0

1.- only one array: for enzymes **Nsp** & **Sty**
2.- only perfect match probes: **PM**
3.- probe pairs are **no** longer **shifted** (tech. replicates)
4.- probes mapping on only one strand of the DNA

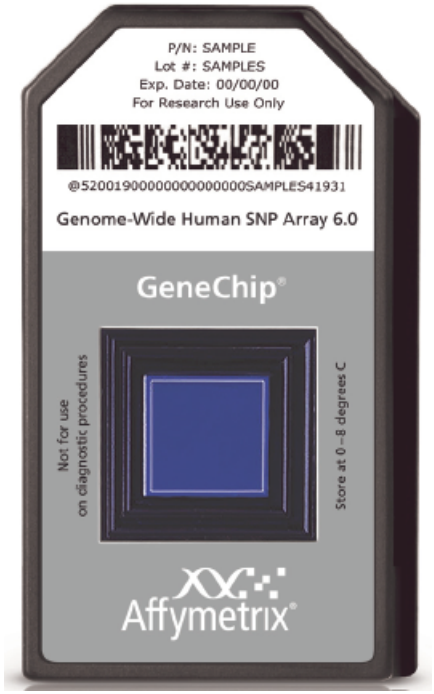
Affymetrix SNP microarrays
100K, 500K Arrays (sets of two: Nsp / Sty)

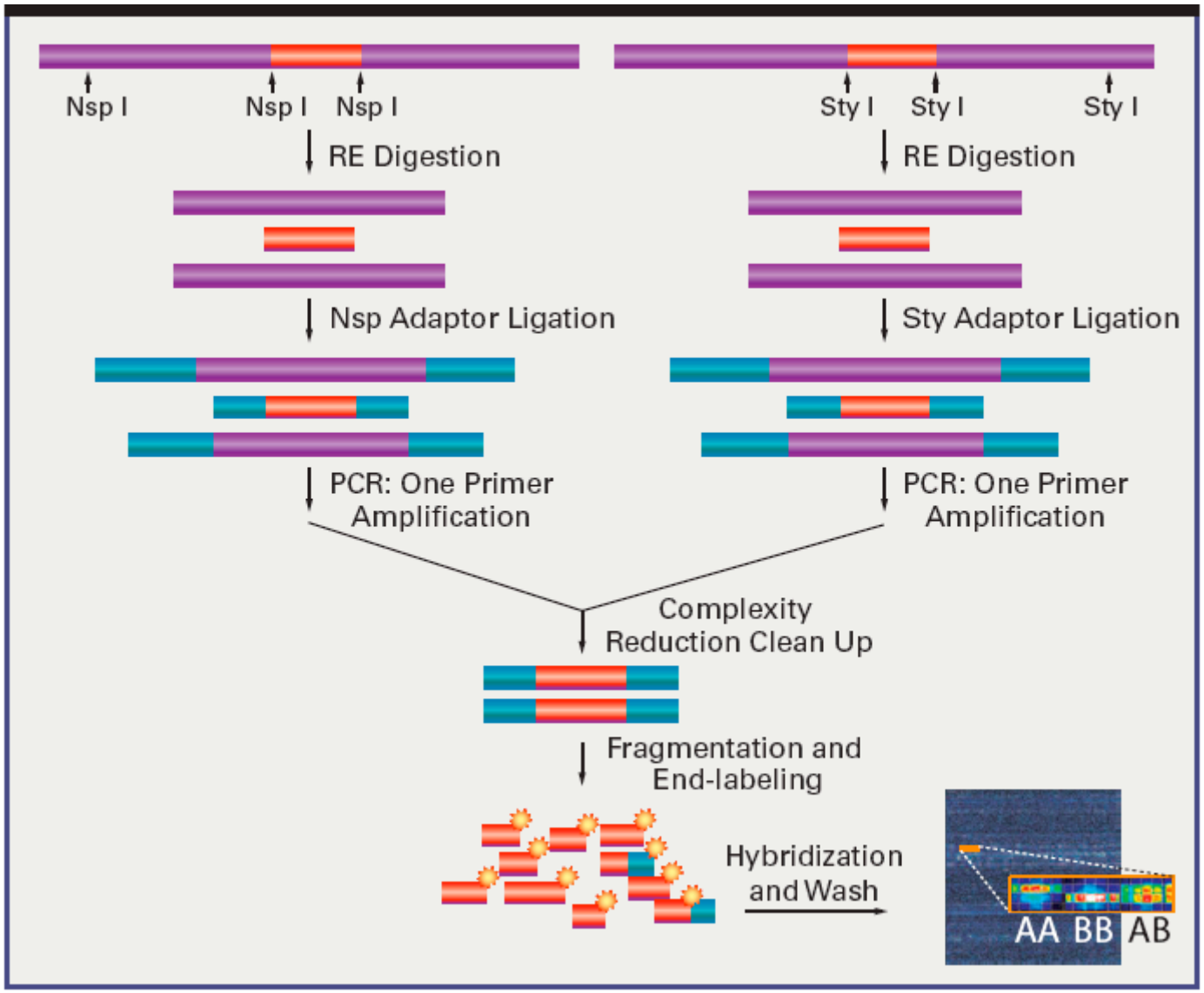
Genome-wide SNP Assay overview



Affymetrix SNP microarrays
5.0 / 6.0 Arrays (one single: Nsp+Sty)

Genome-wide SNP Assay overview





Affymetrix SNP microarrays
2 generations

1st generation:
chips interrogating **SNPs**

- Mapping 10K, 100K, 500K Arrays

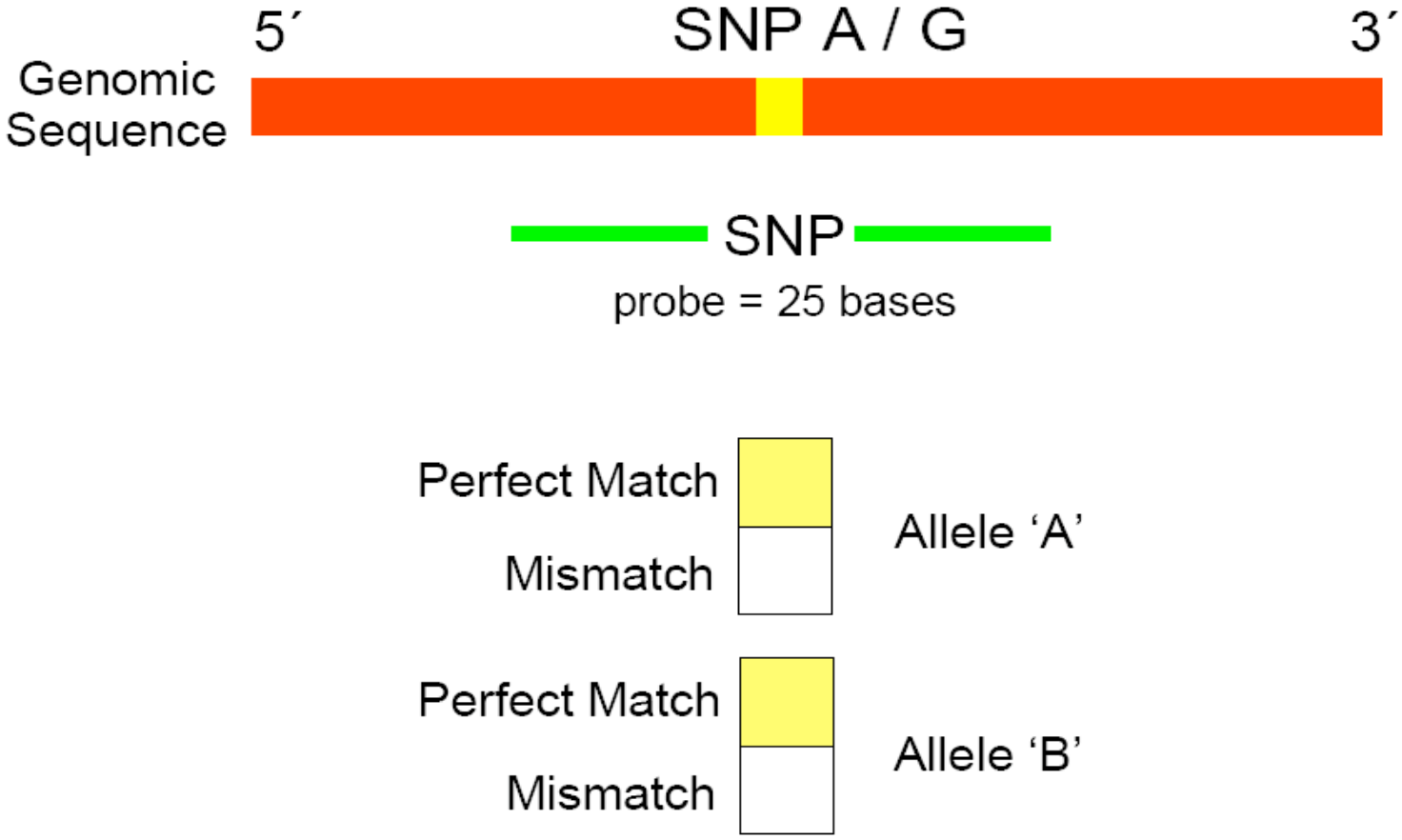
1.- set of two arrays: for two enzymes (**Nsp** & **Sty**)
2.- perferct match & mismatch probes: **PM** & **MM**
3.- probe pairs slightly **shifted** relative to each other
4.- probes mapping on both strands of the DNA

2nd generation:
chips interrogating **SNPs** and **CNVs** (Copy Number Variants)

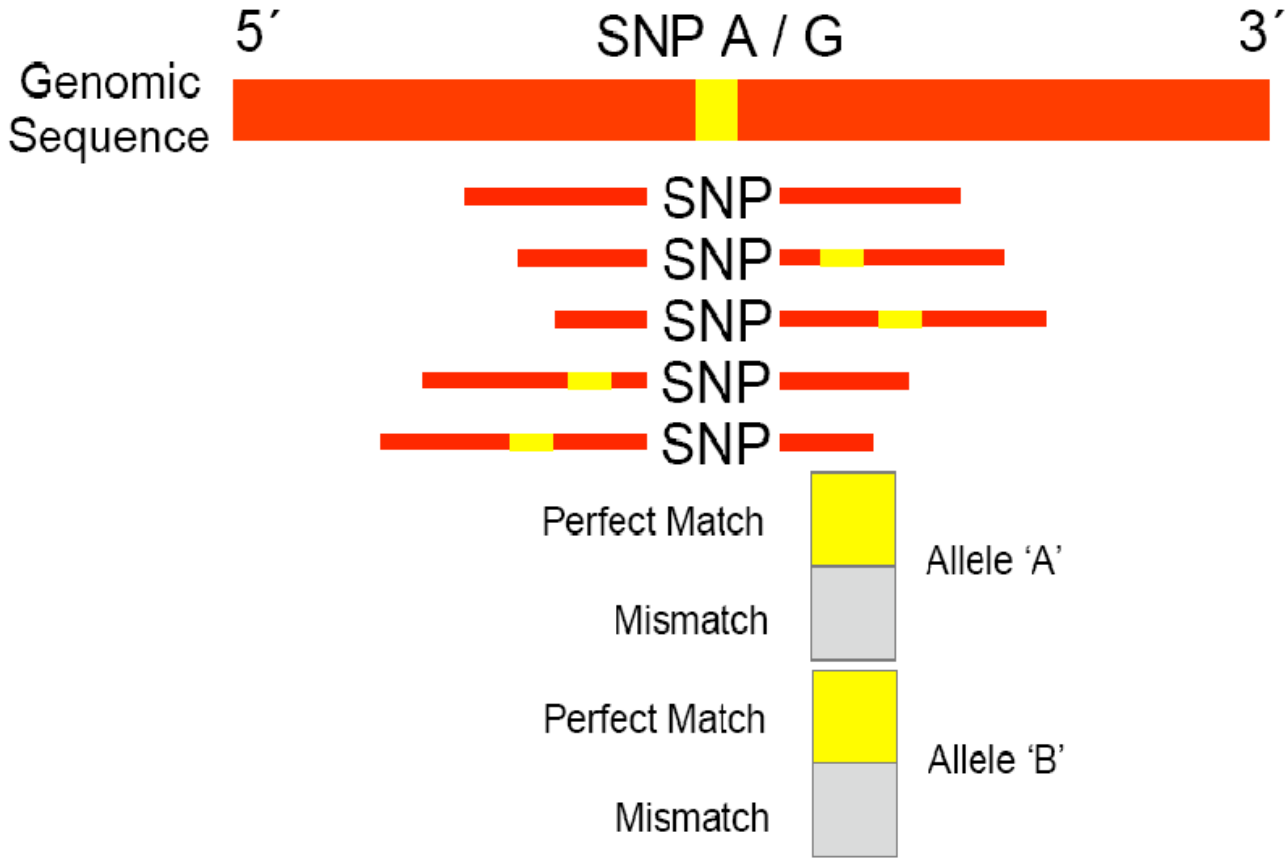
- Genome-Wide Human SNP Arrays 5.0 / 6.0

1.- only one array: for enzymes **Nsp** & **Sty**
2.- only perfect match probes: **PM**
3.- probe pairs are **no** longer **shifted** (tech. replicates)
4.- probes mapping on only one strand of the DNA

Affymetrix SNP arrays :: probes design for chips 10 K, 100K and 500K

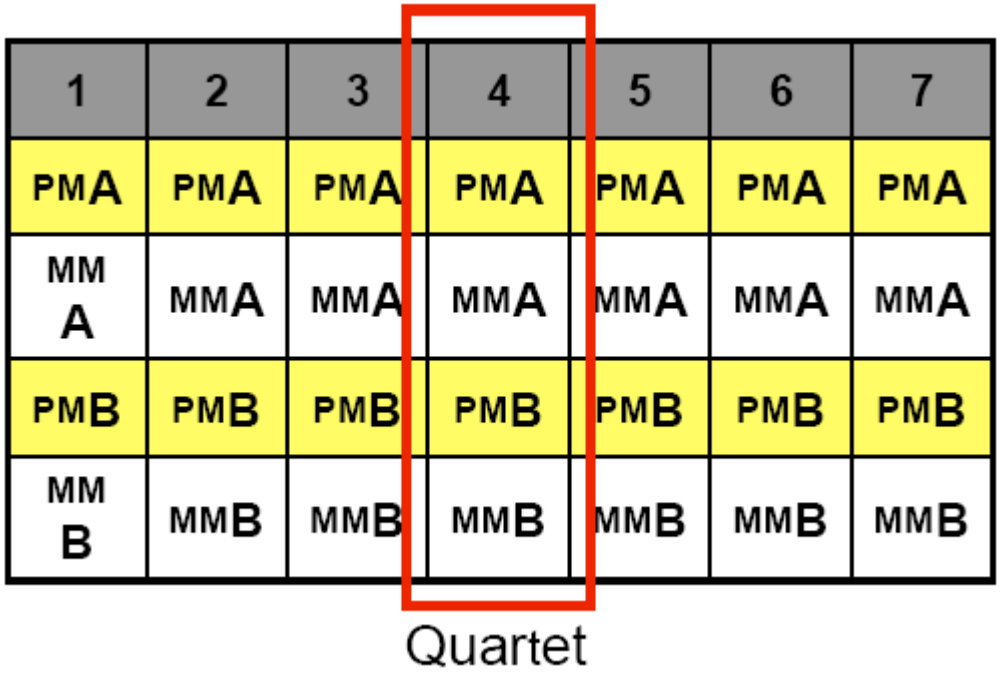


Affymetrix SNP arrays :: probes design for chips 10 K, 100K and 500K



- Each SNP is interrogated on the **Forward** and on the **Reverse strand** by 5 **quarters**
- Therefore a total of 2 (F&R) x 5 (positions) x 4 (quarters, A&B PM&MM) makes a set of **40 probes** to **interrogate a particular SNP**.

Affymetrix SNP arrays :: probes design for chips 10 K, 100K and 500K



- 14 **quartets** are evaluated and the 6 best performing were chosen to represent each SNP. The PM/MM correspond to oligo cells in the microarray.
- The 6 **quartets** interrogate each SNP either the Forward and/or the Reverse strand. Some probesets have all probes on one strand.
- In general, **24 probes** total are used to interrogate a single SNP.
- Some SNPs use **40 probes** (i.e the maximum) usually the ones of high genetic importance.

Affymetrix SNP microarrays
2 generations

1st generation:
chips interrogating **SNPs**

- Mapping 10K, 100K, 500K Arrays

1.- set of two arrays: for two enzymes (**Nsp** & **Sty**)
2.- perferct match & mismatch probes: **PM** & **MM**
3.- probe pairs slightly **shifted** relative to each other
4.- probes mapping on both strands of the DNA

2nd generation:
chips interrogating **SNPs** and **CNVs** (Copy Number Variants)

- Genome-Wide Human SNP Arrays 5.0 / 6.0

1.- only one array: for enzymes **Nsp** & **Sty**
2.- only perfect match probes: **PM**
3.- probe pairs are **no** longer **shifted** (tech. replicates)
4.- probes mapping on only one strand of the DNA

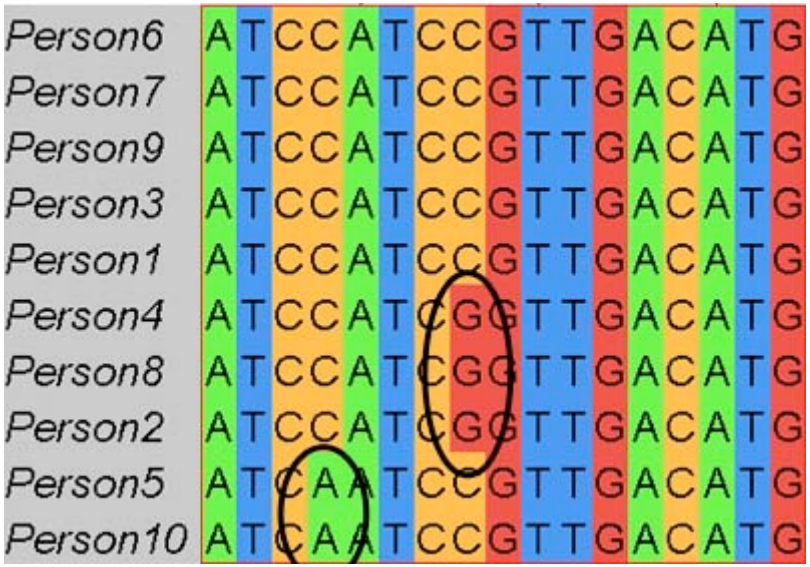
Affymetrix SNP microarrays
detection of SNPs and CpNb

Single **N**ucleotide **P**olymorphism (**SNP**)

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

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

Affymetrix SNP microarrays
detection of SNPs and CpNb



Definition:
Natural variation between people (**A**, **T**, **C**, or **G**).

Allele A: ...CGTAGCCATCGGTA**A**/GTACTCAATGATAG...

Allele B: ...CGTAGCCATCGGTA**G**/GTACTCAATGATAG...

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

Affymetrix SNP microarrays
detection of SNPs and CpNb

Probes for SNPs

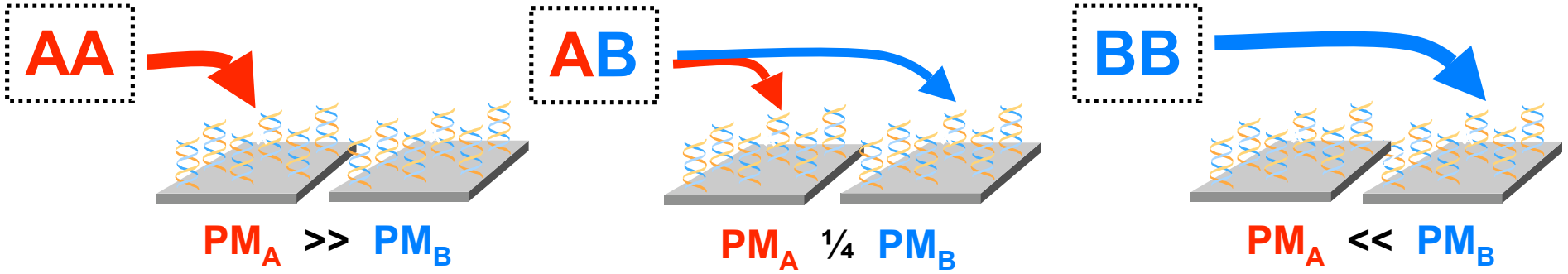
PM_A: ATCGGTAGCCATTCATGAGTTACTA

Allele A: ...CGTAGCCATCGGTAAGTACTCAATGATAG...

Allele B: ...CGTAGCCATCGGTAGGTACTCAATGATAG...

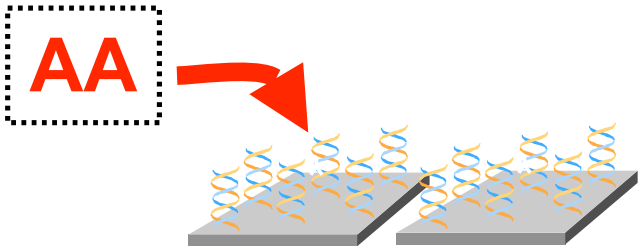
PM_B: ATCGGTAGCCATCCATGAGTTACTA

(The old chips include also MMs)

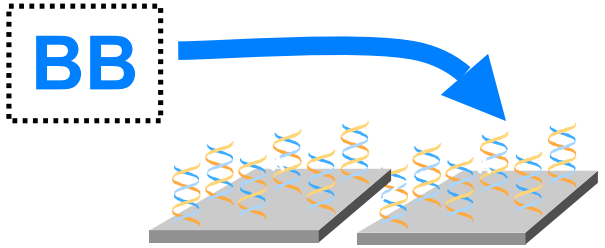


Affymetrix SNP microarrays
detection of SNPs and CpNb

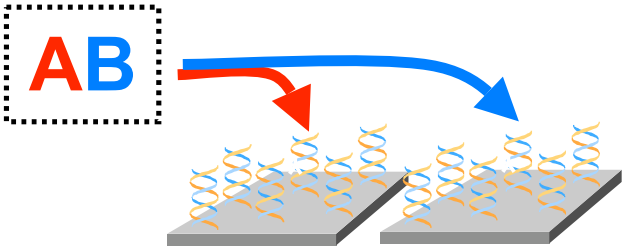
SNP probes can also be used to
estimate total copy numbers



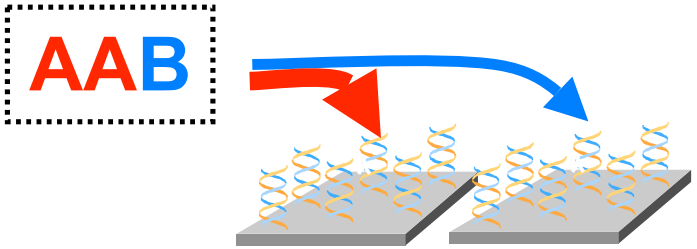
PM = PM_A + PM_B = 2c



PM = PM_A + PM_B = 2c



PM = PM_A + PM_B = 2c



PM = PM_A + PM_B = 3c

SNPs Microarray Data Analysis

Three types of possible analysis:

- **Genotyping** (SNPs and other P) analysis
- **Specific SNPs** sets analysis (≈ genotyping)
- **Copy number** and **LOH** analysis

SNPs Microarray Data Analysis

Three types of possible software:

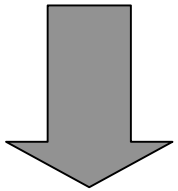
- ***Affymetrix*** software

- ***Affymetrix*** compatible comercial software

- ***Affymetrix*** compatible non-comercial software

SNPs Microarray Data Analysis
Affymetrix software tools

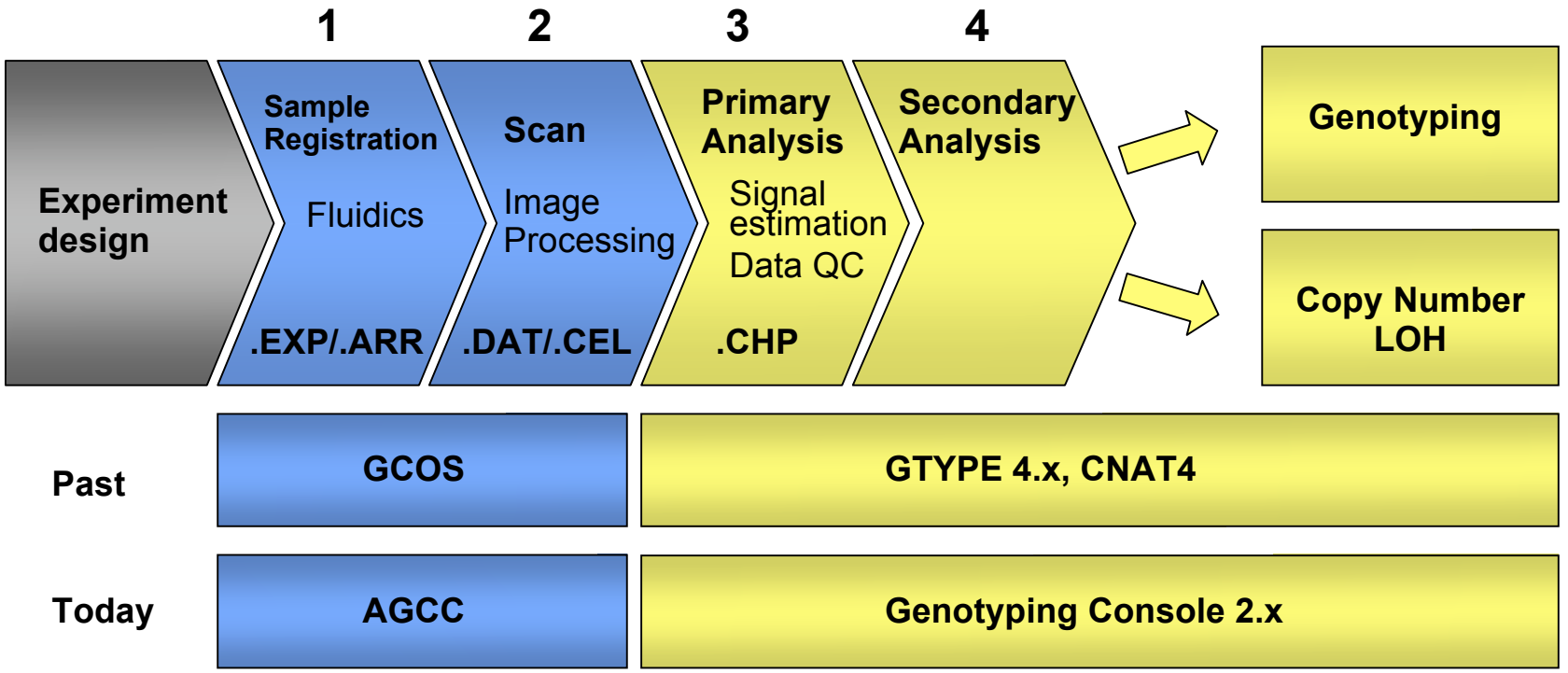
- **GCOS** (**G**ene**C**hip **O**perating **S**oftware)
  - **DTT** (**D**ata **T**ransfer **T**ool)
- **GTYPE 4.x** (includes BRLMM Analysis Tool)
- **CNAT 4** (**C**opy **N**umber **A**nalysis **T**ool)



- **AGCC** (**A**ffymetrix **G**eneChip **C**ommand **C**onsole)
- **Genotyping Console 2.x**
  - **GTC Browser** (Visualization Tool)

At present the current version supported by Affymetrix
(http://www.affymetrix.com/products/software/specific/genotyping_console_software.affx)

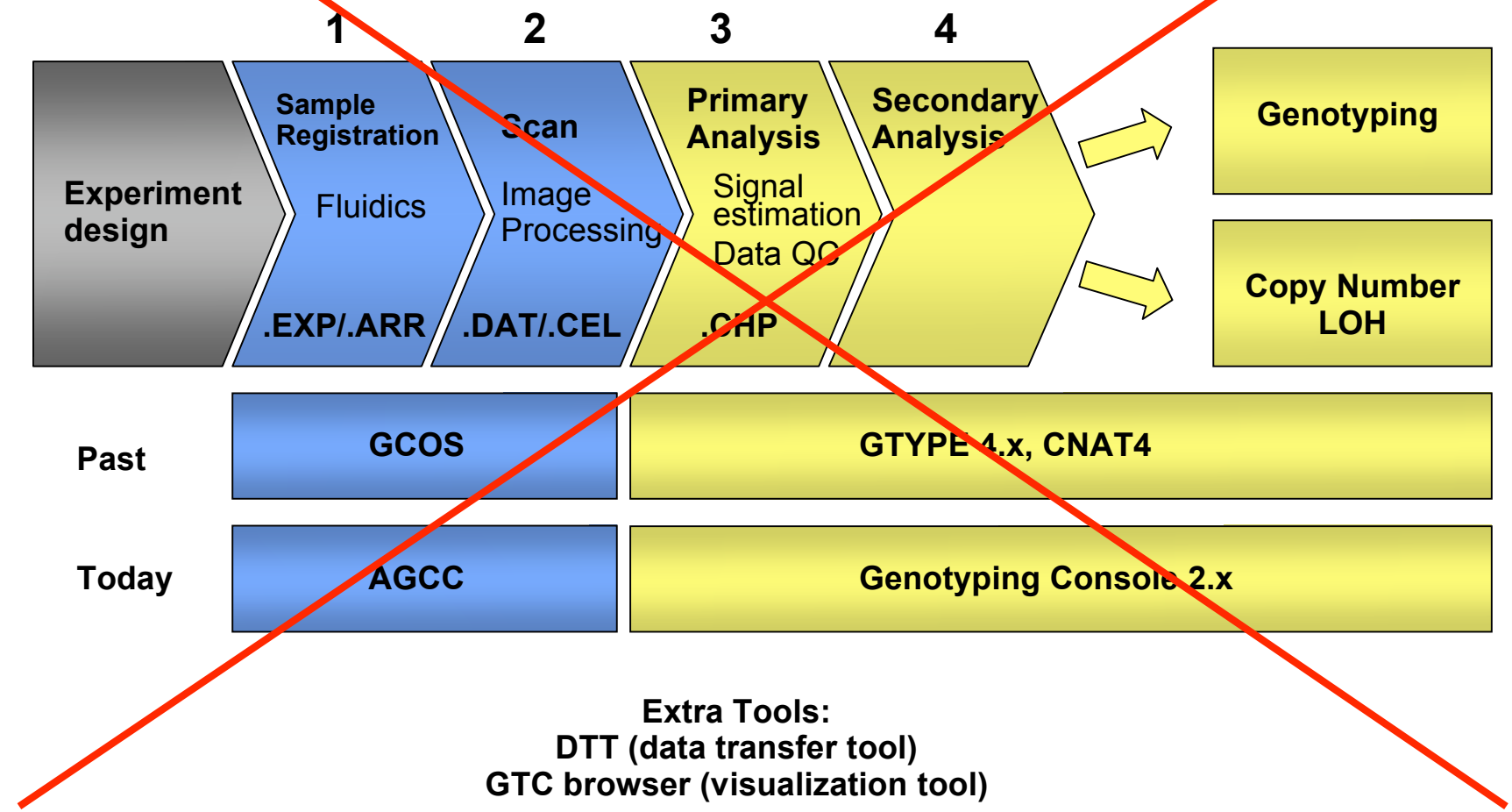
SNPs Microarray Data Analysis
Affymetrix software workflow



Extra Tools:
DTT (data transfer tool)
GTC browser (visualization tool)

SNPs Microarray Data Analysis

Affymetrix software workflow



Extra Tools:
DTT (data transfer tool)
GTC browser (visualization tool)

SNPs Microarray Data Analysis

Three types of possible software:

- ***Affymetrix*** software
- ***Affymetrix*** compatible comercial software
- ***Affymetrix*** compatible non-comercial software

SNPs Microarray Data Analysis
Affymetrix compatible software

Comercial Software

- **Partek® Genomics Solution™** is the first commercial software application available for copy number analysis in support of Affymetrix SNP arrays
- **Nexus-CGH** from Biodiscovery tha supports many platforms including Affymetrix SNP data
- **HelixTree® Genetics Analysis Software** from Golden Helix
- **GeneSense** from InforSense
- **Exemplar Genotyping Analysis Suite** from Sapio Sciences
- **JMP® Genetics** from SAS

Non-comercial software

- **CNAG** developed at the University of Tokyo
  Ref: Nannya Y. *et al*. Cancer Res. (2005) 65(14):6071-9
- **dChip SNP** developed at Dana Farber Cancer Institute (Harvard University)
  Ref: Lin M. *et al*. Bioinformatics (2004) 20(8):1233-40
- **R and BioC packages** (SNPchip, oligo, aroma.affymetrix, VanillaICE, …)

SNPs Microarray Data Analysis
Affymetrix compatible software

Comercial Software

- **Partek® Genomics Solution™** is the first commercial software application available for copy number analysis in support of Affymetrix SNP arrays
- **Nexus-CGH** from Biodiscovery tha supports many platforms including Affymetrix SNP data
- **HelixTree® Genetics Analysis Software** from Golden Helix
- **GeneSense** from InforSense
- **Exemplar Genotyping Analysis Suite** from Sapio Sciences
- **JMP® Genetics** from SAS

Non-comercial software

- **CNAG** developed at the University of Tokyo
  Ref: Nannya Y. *et al*. Cancer Res. (2005) 65(14):6071-9
- **dChip SNP** developed at Dana Farber Cancer Institute (Harvard University)
  Ref: Lin M. *et al*. Bioinformatics (2004) 20(8):1233-40
- **R and BioC packages** (SNPchip, oligo, aroma.affymetrix, VanillaICE, …)

Microarray Data Analysis workflow

Preparation
+ Hybridization
+ Scanning

DAT File(s)
[Image, pixel intensities]

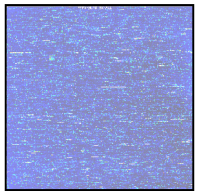


Image analysis

workable raw data

CEL File(s)
[Probe Cell Intensity]

+

CDF
[Chip Description File]

Analysis step 1
Analysis step 2
Analysis step 3
...

Results
with statistical and biological significance

Methods implemented in aroma.affymetrix

Method	*References*	*Multiarray model?*	*Bounded in memory*
Pre-summarization calibration & normalization			
Optical background correction	Wu *et al.* (2004)	single	yes
GCRMA background correction	Wu *et al.* (2004)	single	yes
RMA background correction	Irizarry *et al.* (2003)	single	yes
Rank-based quantile normalization*	Bolstad *et al.* (2003)	multi	yes
Function-based quantile normalization	Bengtsson *et al.* (2008)	multi	yes
Allelic crosstalk calibration	Bengtsson *et al.* (2008)	single	yes
Probe-level summarization			
Averaging model	(to be published)	single	yes
Log-additive model*	Irizarry *et al.* (2003)	multi	yes
Multiplicative model*	Li and Wong (2001)	multi	yes
Affine (multiplicative) model	Bengtsson *et al.* (2004)	multi	yes
Post-summarization calibration & normalization			
PCR fragment-length normalization	Bengtsson *et al.* (2008)	multi	yes
GC-content normalization	Bengtsson *et al.* (2008)	multi	yes
Equivalent fragment class normalization	(to be published)	multi	yes
Copy-number segmentation			
(Fast) CBS segmentation	Venkatraman and Olshen (2007)	single	yes
GLAD segmentation	Hupé *et al.* (2004)	single	yes
Alternative splicing			
FIRMA model	Purdom *et al.* (2008)	multi	yes

SNPs Microarray Data Analysis workflow

Preparation
+ Hybridization
+ Scanning

DAT File(s)
[Image, pixel intensities]

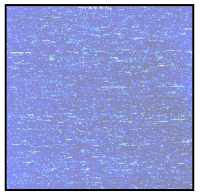


Image analysis

workable raw data

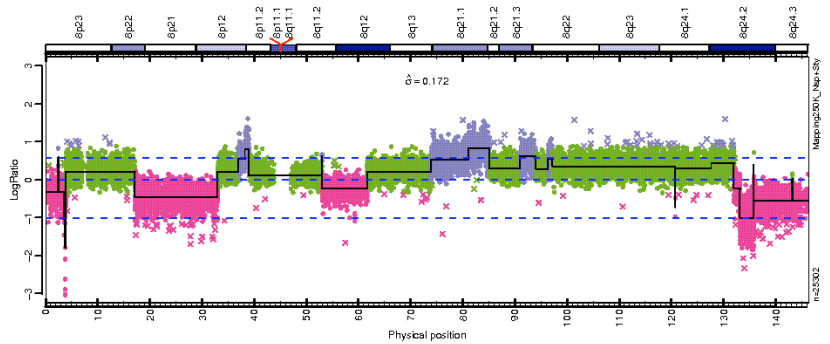
CEL File(s)
[Probe Cell Intensity]

+

CDF
[Chip Description File]

Low-level analysis = CRMAv2

Segmentation = CBS



Estimate copy numbers from SNP arrays
aroma.affymetrix

BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 6 2008, pages 759–767
doi:10.1093/bioinformatics/btn016

Genome analysis

Estimation and assessment of raw copy numbers at the single locus level

H. Bengtsson[1,*], R. Irizarry[2], B. Carvalho[2] and T. P. Speed[1,3]

[1]Department of Statistics, University of California, Berkeley, USA, [2]Department of Biostatistics, Johns Hopkins University, Baltimore, USA and [3]Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Australia

Received on September 18, 2007; revised on December 25, 2007; accepted on January 9, 2008

Advance Access publication January 19, 2008

Associate Editor: Alfonso Valencia

CRMA (2008)

BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 17 2009, pages 2149–2156
doi:10.1093/bioinformatics/btp371

Genome analysis

A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6

Henrik Bengtsson[1,*], Pratyaksha Wirapati[2] and Terence P. Speed[3]

[1]Department of Statistics, University of California, Berkeley, USA, [2]Bioinformatics Core Facility, Swiss Institute of Bioinformatics, Lausanne, Switzerland and [3] Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Parkville, Australia

Received on November 4, 2008; revised on June 1, 2009; accepted on June 11, 2009

Advance Access publication June 17, 2009

Associate Editor: John Quackenbush

CRMAv2 (2009)

Estimate copy numbers from SNP arrays
aroma.affymetrix (notation)

Indices:

Arrays/samples: i = 1, 2, …, I

Loci/**SNP**s/CN units: j = 1, 2, …, J

Replicated probes for SNP: k = 1, 2, …, K

Probe signals:

CN locus: $y_{ij} = PM_{ij}$ (single-probe units)

SNP allele pair k: $(y_{ijkA}, y_{ijkB}) = (PM_{ijkA}, PM_{ijkB})$

Summarized signals ("chip effects"):

CN locus: θ_{ij}

SNP: $(\theta_{ijA}, \theta_{ijB})$

Estimate copy numbers from SNP arrays
aroma.affymetrix (obtain CN estimates)

- Calculate non-polymorphic SNP summaries:
  - For each **array** i=1,…,I and **SNP** j=1,…,J:
    - Probe **allele** pairs: (PM_{ijkA}, PM_{ijkB}); k=1,…,K
    - For both alleles, average across probes:
      $\theta_{ijA} = \text{median}_k \{PM_{ijkA}\}$, $\theta_{ijB} = \text{median}_k \{PM_{ijkB}\}$
    - Sum both alleles: $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$
- Calculate reference θ_{Rj} across all arrays:
  - For each **SNP** j=1,…,J:
    - $\theta_{Rj} = \text{median}_i \{\theta_{ij}\}$
- Calculate CN log-ratios:
  - For each **array** i=1,…,I and **SNP** j=1,…,J:
    - $M_{ij} = \log_2 (\theta_{ij} / \theta_{Rj})$

Estimate copy numbers from SNP arrays

	CRMA v2
Preprocessing *(probe signals)*	1. Allelic crosstalk calibration 2. Probe-sequence normalization
Summarization	Robust averaging: CN probes: $\theta_{ij} = PM_{ij}$ SNPs: $\theta_{ijA} = \text{median}_k(PM_{ijkA})$ $\theta_{ijB} = \text{median}_k(PM_{ijkB})$ array *i*, loci *j*, probe *k.*
Post-processing	PCR fragment-length normalization
Transform	$(\theta_{ijA},\theta_{ijB}) \Rightarrow (\theta_{ij}, \beta_{ij})$ $\theta_{ij} = \theta_{ijA}+\theta_{ijB}$, $\beta_{ij} = \theta_{ijB} / \theta_{ij}$
Allele-specific & Total Copy Nbs	$C_{ijA} = 2*(\theta_{ijA} / \theta_{Rj})$ and $C_{ijB} = 2*(\theta_{ijA} / \theta_{Rj})$ $C_{ij} = 2*(\theta_{ij} / \theta_{Rj})$ reference *R*

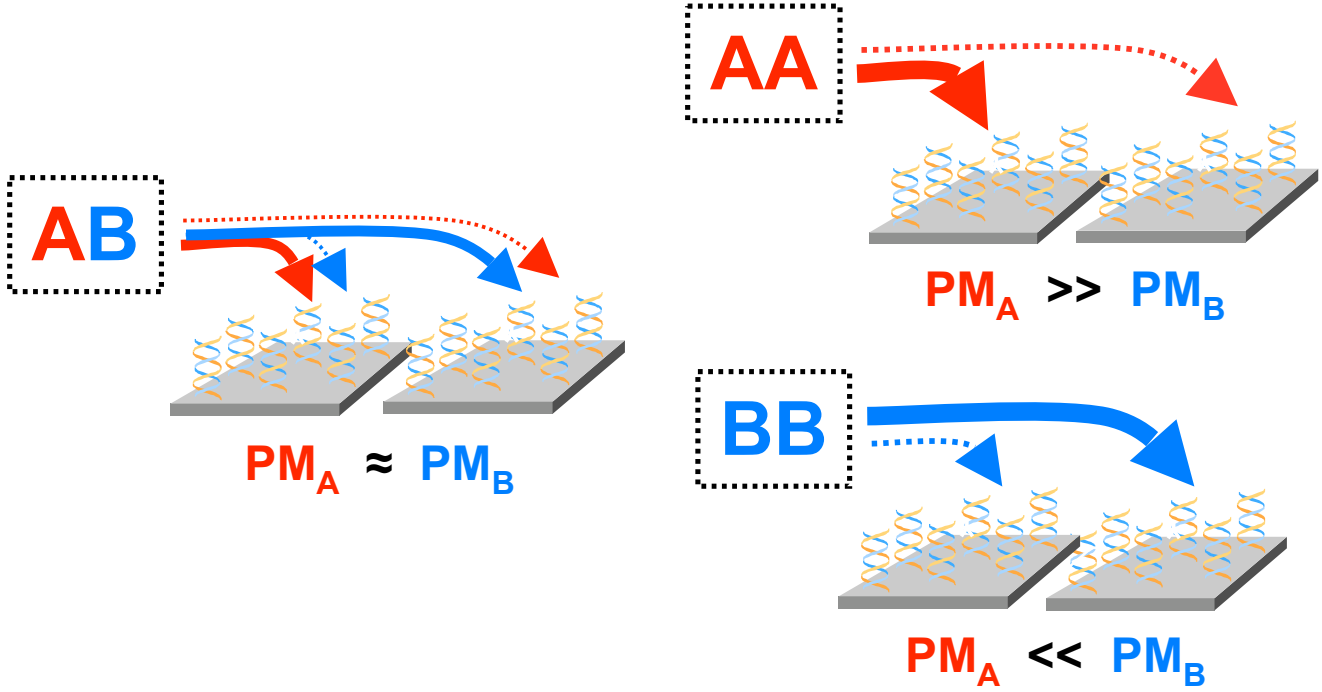
Allelic crosstalk calibration

Allelic crosstalk calibration (CRMAv2)

Crosstalk between alleles
(adds significant artifacts to signals)

Cross-hybridization:

Allele A: TCGGTAAGTACTC
Allele B: TCGGTATGTACTC



Allelic crosstalk calibration (CRMAv2)

There are six possible allele pairs

- Nucleotides: {**A, C, G, T**}

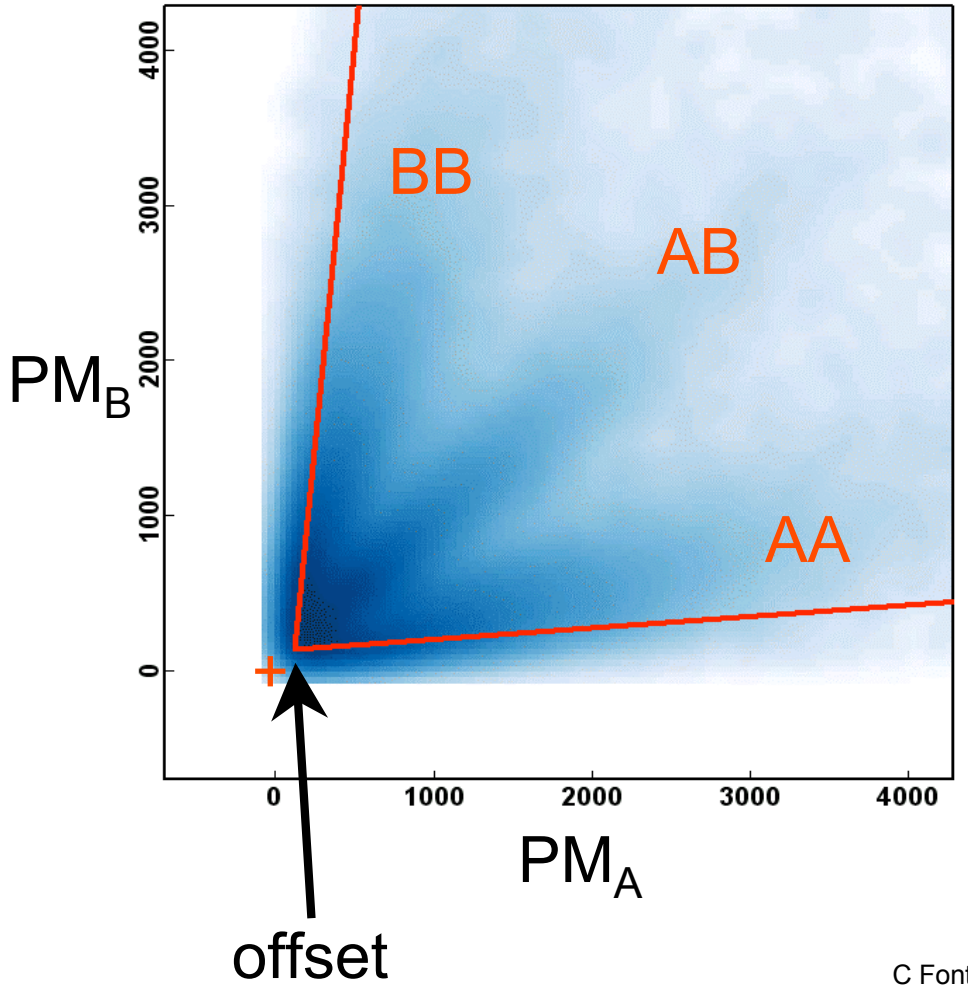
- Ordered pairs (6 possible):

  (**A,C**), (**A,G**), (**A,T**), (**C,G**), (**C,T**), (**G,C**)

- Because of different nucleotides bind differently, the crosstalk from A to C might be very different from A to T.

Allelic crosstalk calibration (CRMAv2)

Crosstalk between alleles is easy to spot

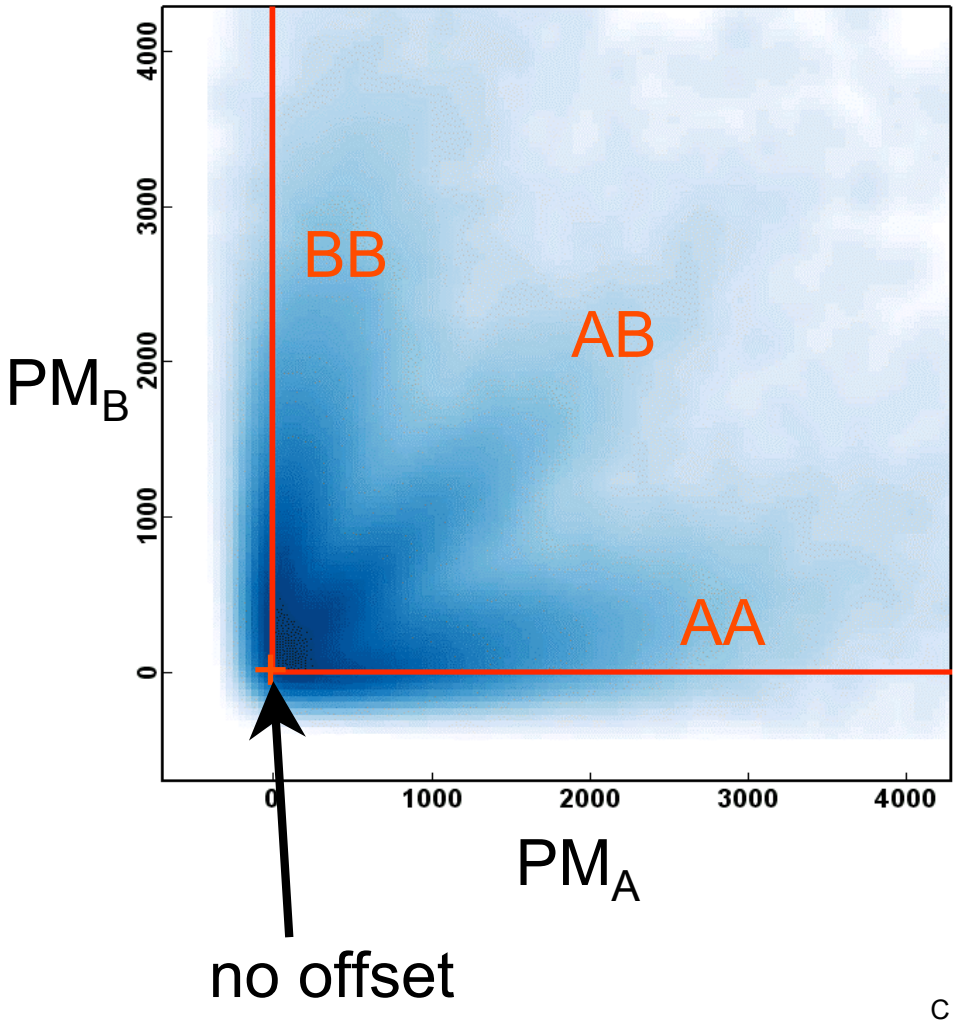


Example:

Data from one array.
Probe pairs (PM_A, PM_B) for nucleotide pair (A,T).

Allelic crosstalk calibration (CRMAv2)

Crosstalk between alleles can be estimated and corrected for



What is done:

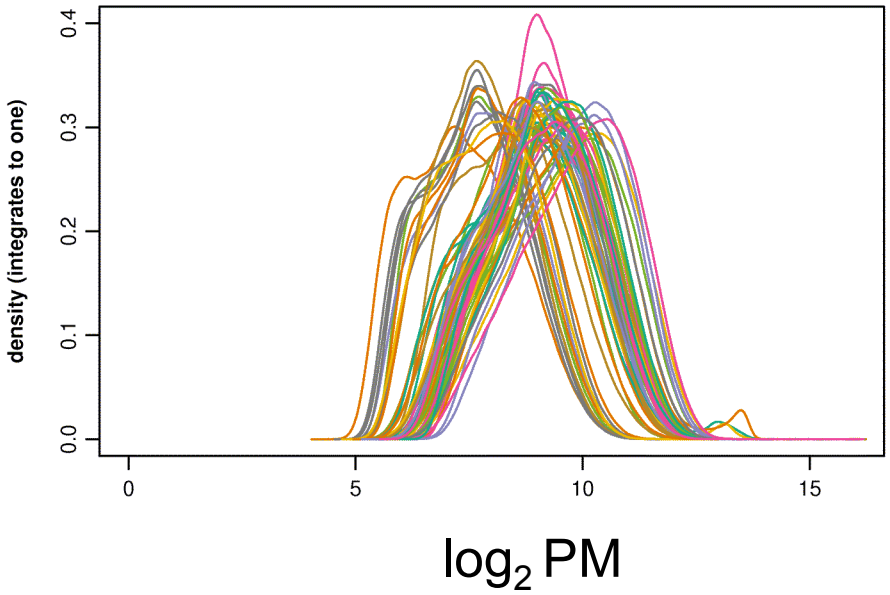
1. **Offset is removed** from SNPs and CN units.

2. **Crosstalk is removed** from SNPs.

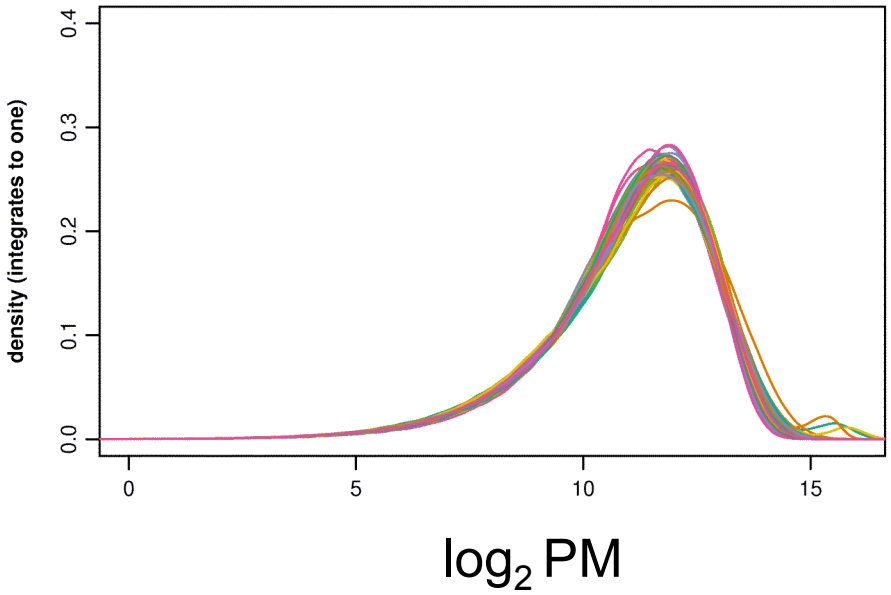
Allelic crosstalk calibration (CRMAv2)

Crosstalk calibration corrects for differences in distributions too

Before removing crosstalk
the arrays differ significantly



After removing **offset** & **crosstalk**
differences goes away.



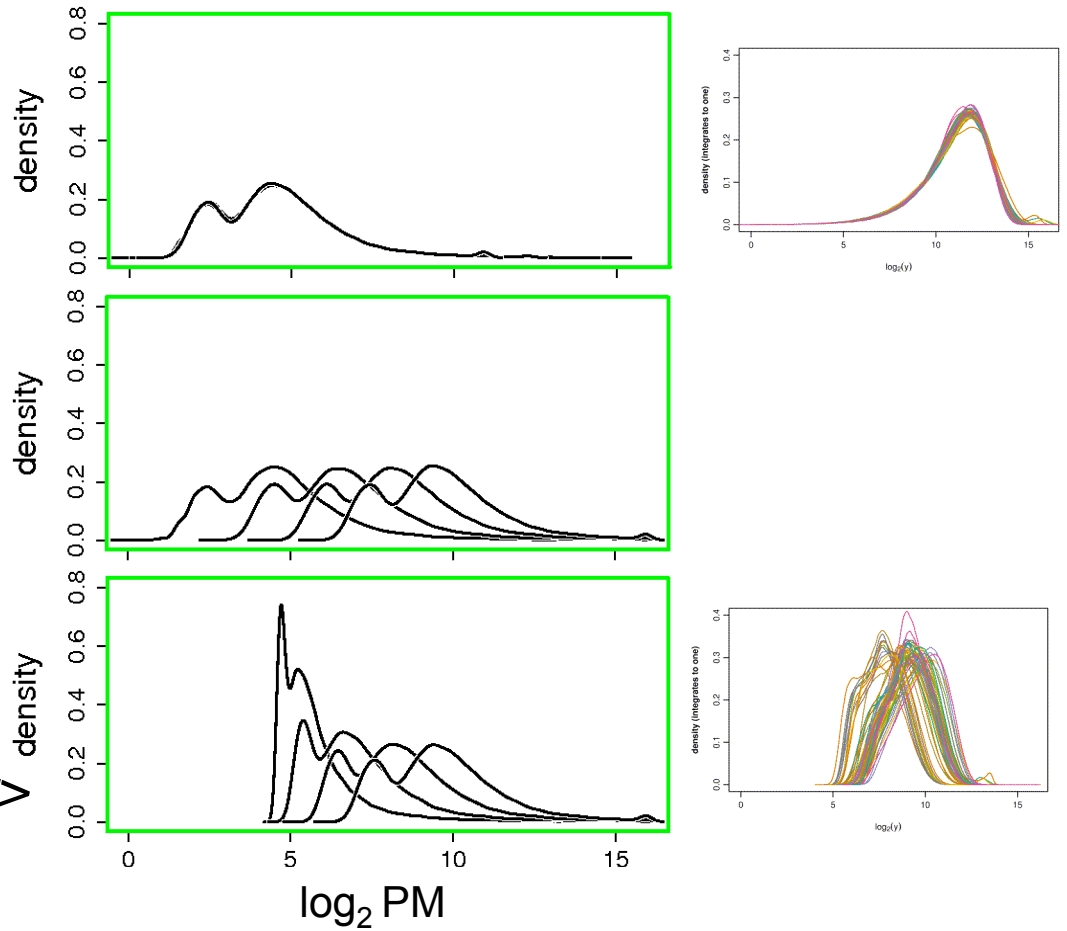
Allelic crosstalk calibration (CRMAv2)

How can a translation and a rescaling make such a big difference?

4 measurements
of the **same thing**:

With **different scales**:
*log(b*PM) = log(b)+log(PM)*

With **different scales**
and **some offset**:
*log(a+b*PM) = <non-linear>*



Allelic crosstalk calibration (CRMAv2)

Allelic crosstalk calibration

... controls for:

- Offset in signals

- Crosstalk between allele A and allele B

aroma.affymetrix

You will need:

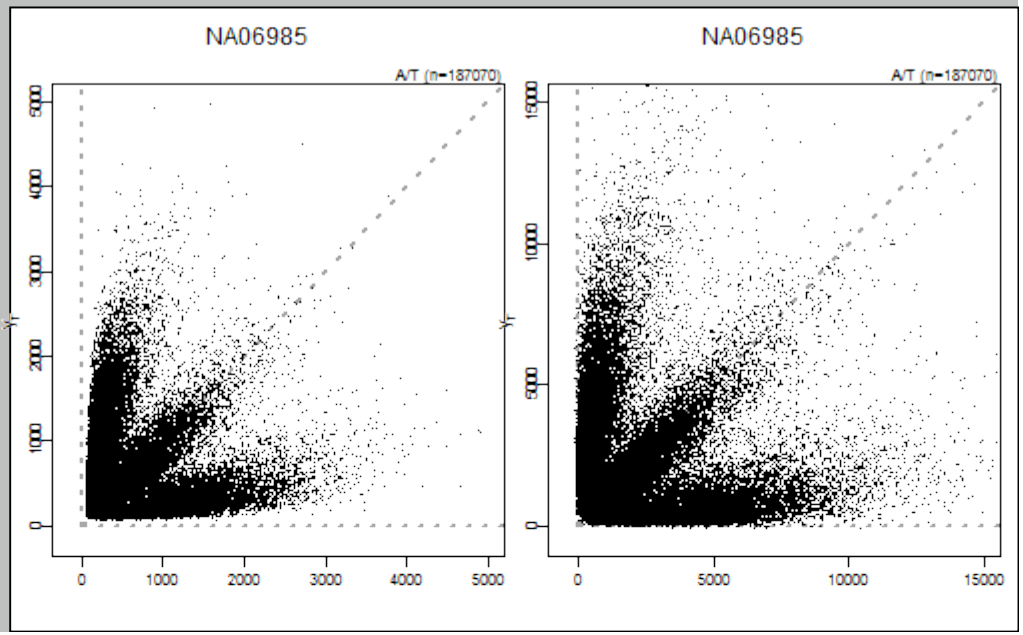
- Affymetrix CDF, e.g. GenomeWideSNP_6.cdf
- Probe sequences*, e.g. GenomeWideSNP_6.acs

Calibrate CEL files:

```
cdf <- AffymetrixCdfSet$byChipType("GenomeWideSNP_6")
csR <- AffymetrixCelSet$byName("HapMap", cdf=cdf)
acc <- AllelicCrosstalkCalibration(csR, model="CRMAv2")
csC <- process(acc)
```

To plot:

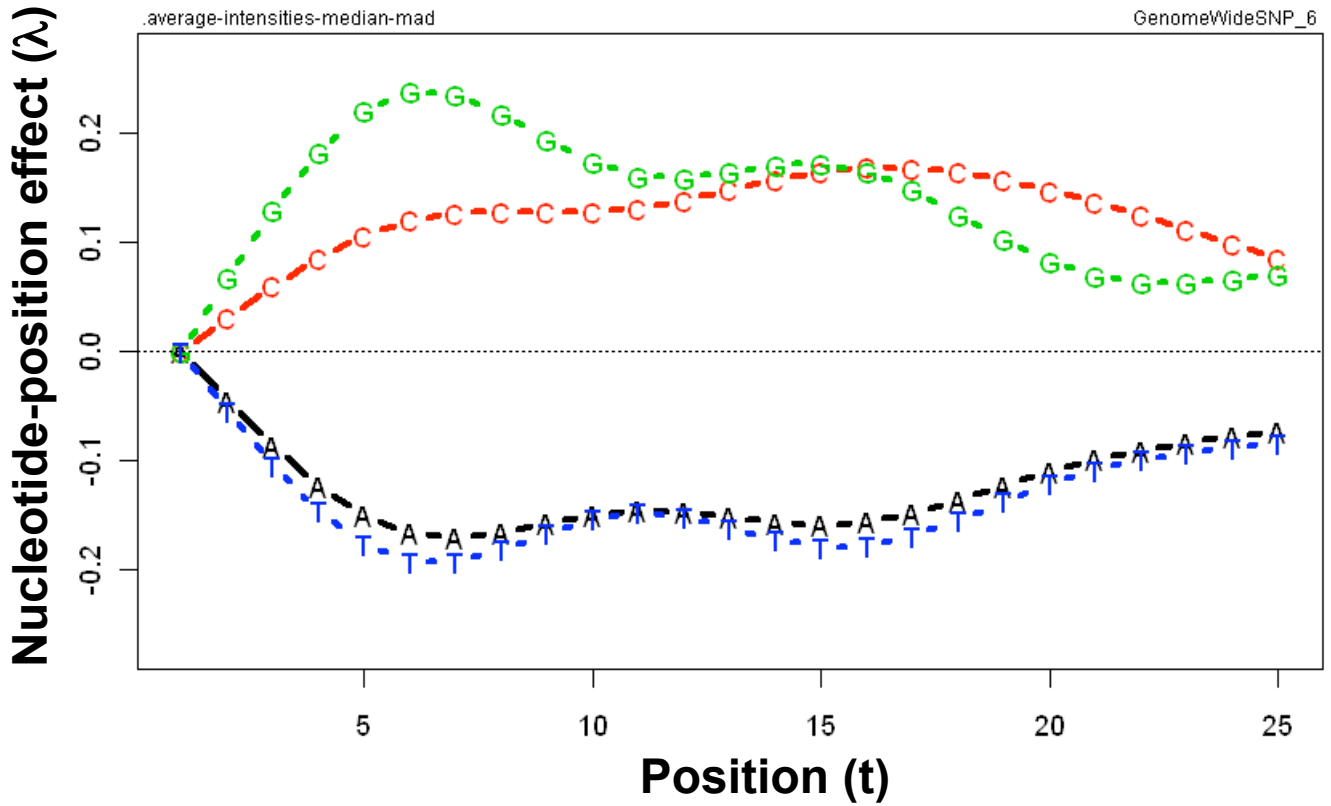
```
plotAllelePairs(acc, array=1)
plotAllelePairs(acc, array=1,
                what="output")
```



Probe sequence normalization

Probe sequence normalization (CRMAv2)

Nucleotide-Position Model



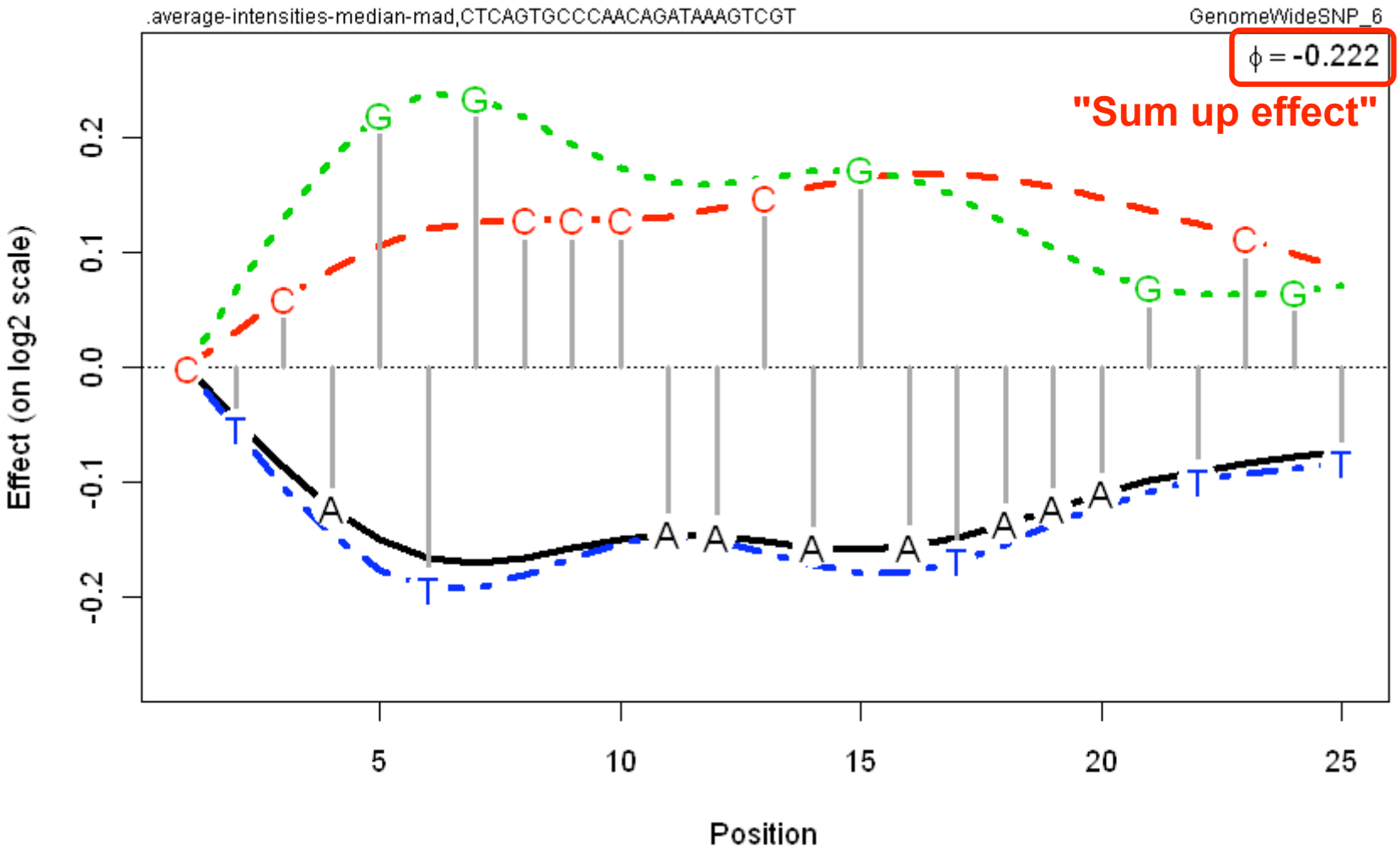
Probe-position ($\log_2$) affinity for probe k:

$$\phi_k = \phi((b_{k,1},b_{k,2},\ldots,b_{k,25})) = \sum_{t=1..25} \sum_{b=\{ACGT\}} I(b_{k,t}=b)\lambda_{b,t}$$

Probe sequence normalization (CRMAv2)

Example: Probe-position affinity for

CTCAGTGCCCAACAGATAAAGTCGT



Probe sequence normalization (CRMAv2)

The normalization helps to correct effects

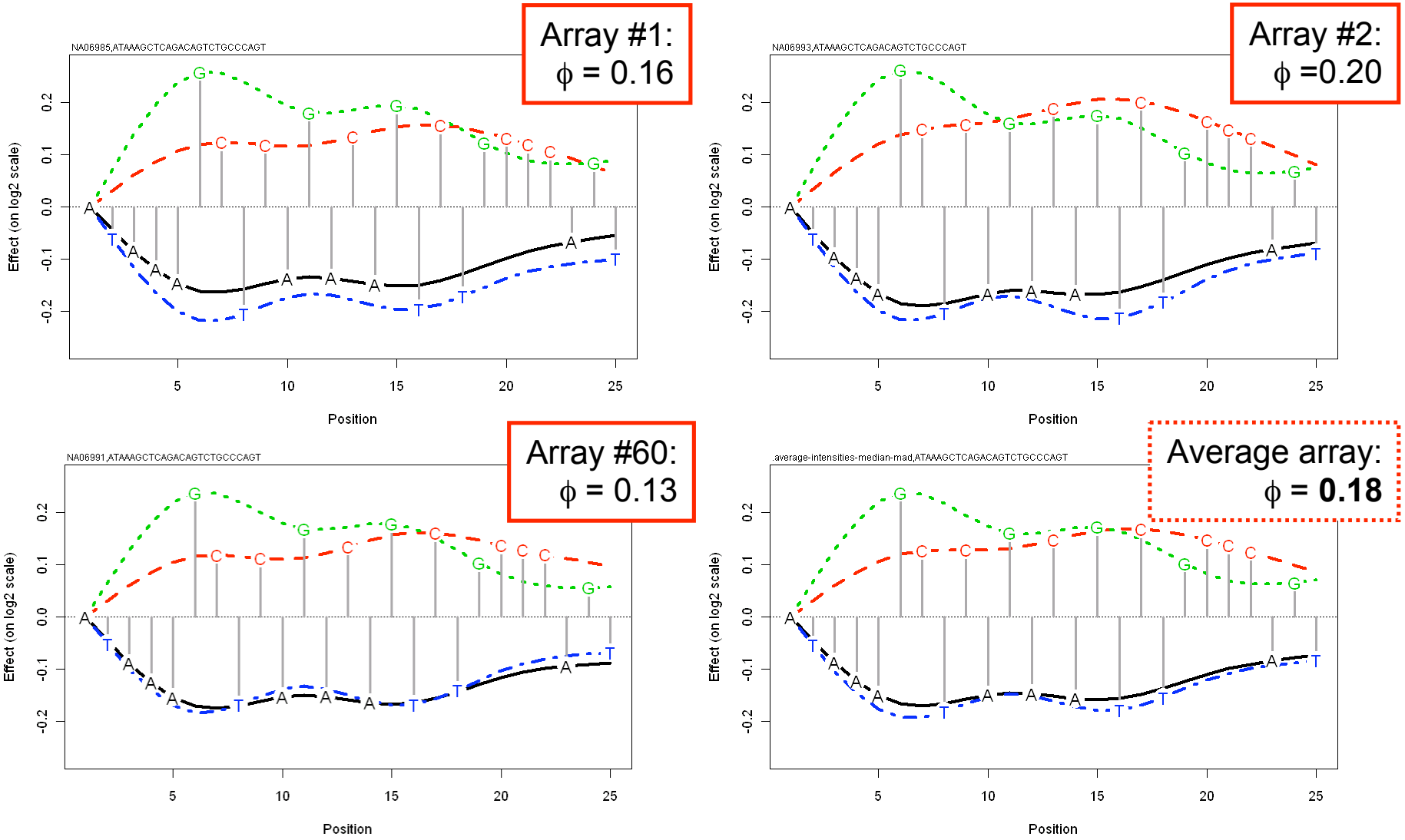
1. The effects differ slightly across arrays:
   - adds extra across-array variances
   - *will be removed*

2. The effects differ between PM_A and PM_B:
   - introduces genotypic imbalances such that PM_A+PM_B will differ for AA, AB & BB.
   - *will be removed*

1. BPN controls for across array variability

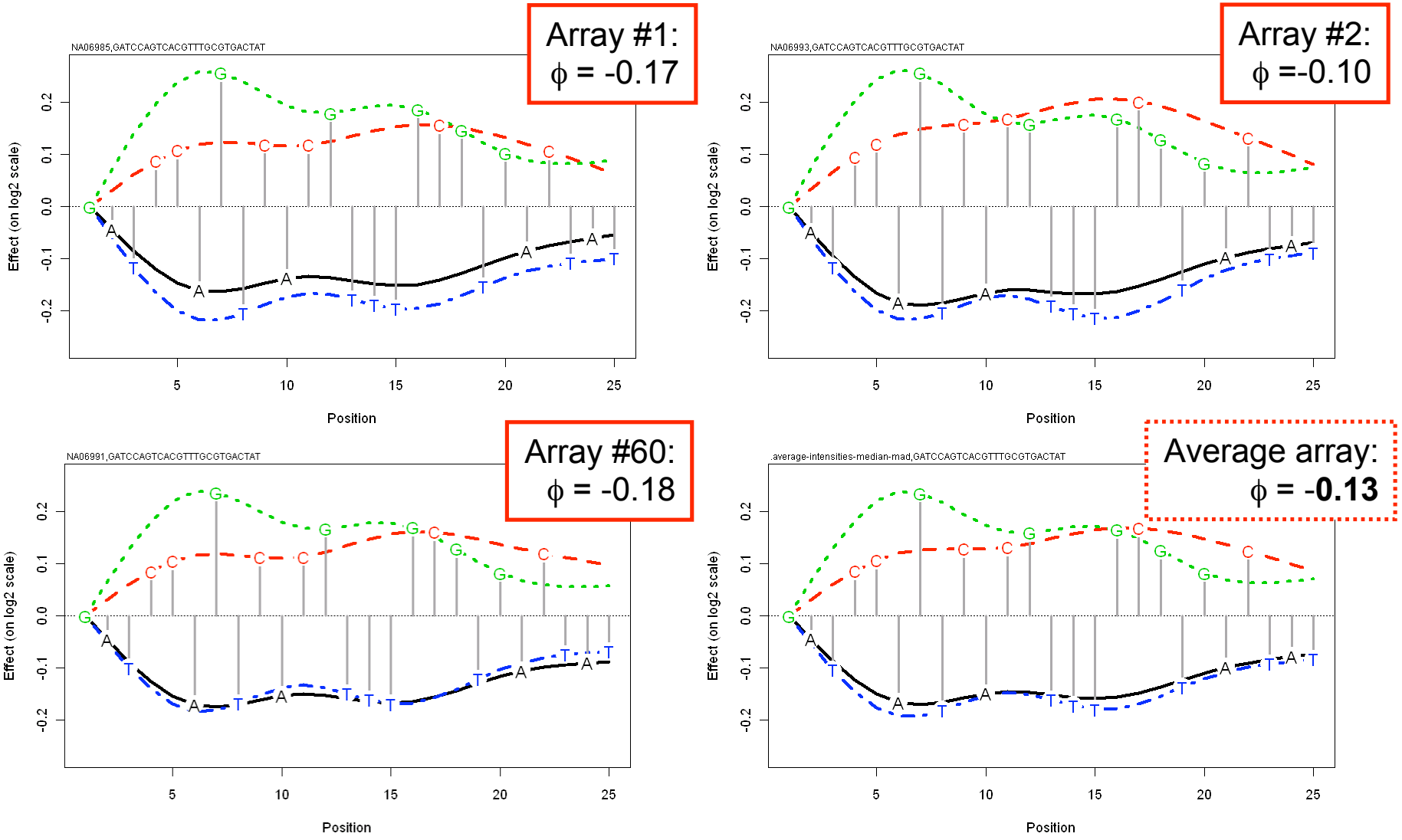
Probe sequence normalization (CRMAv2)

The **nucleotide-position** effect differ between arrays



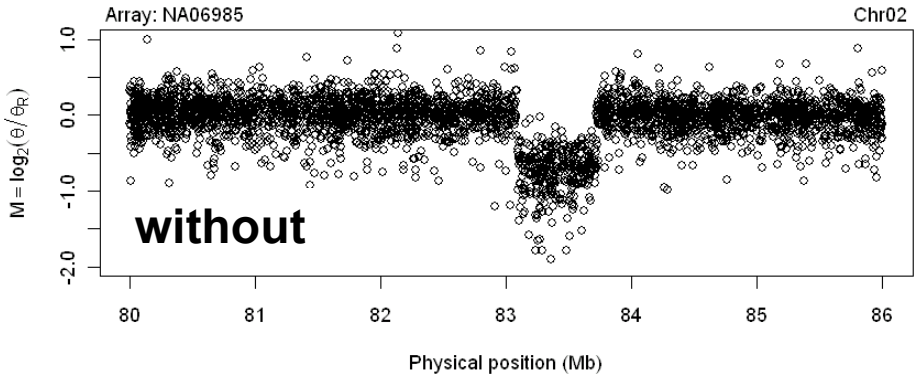
Probe sequence normalization (CRMAv2)

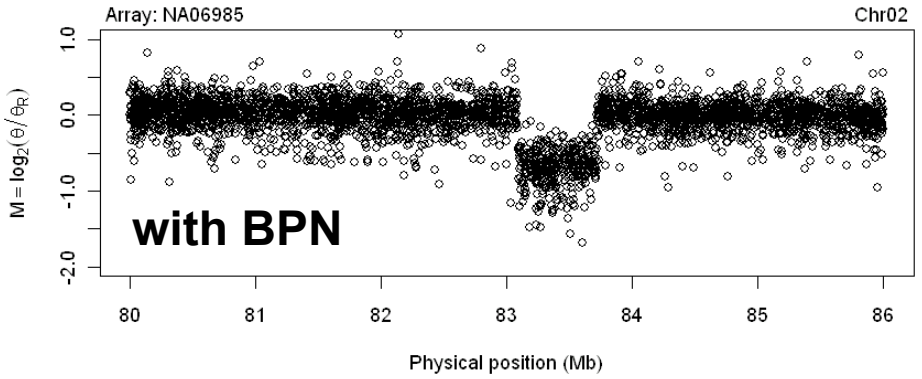
The **impact** of these effects varies with probe sequence

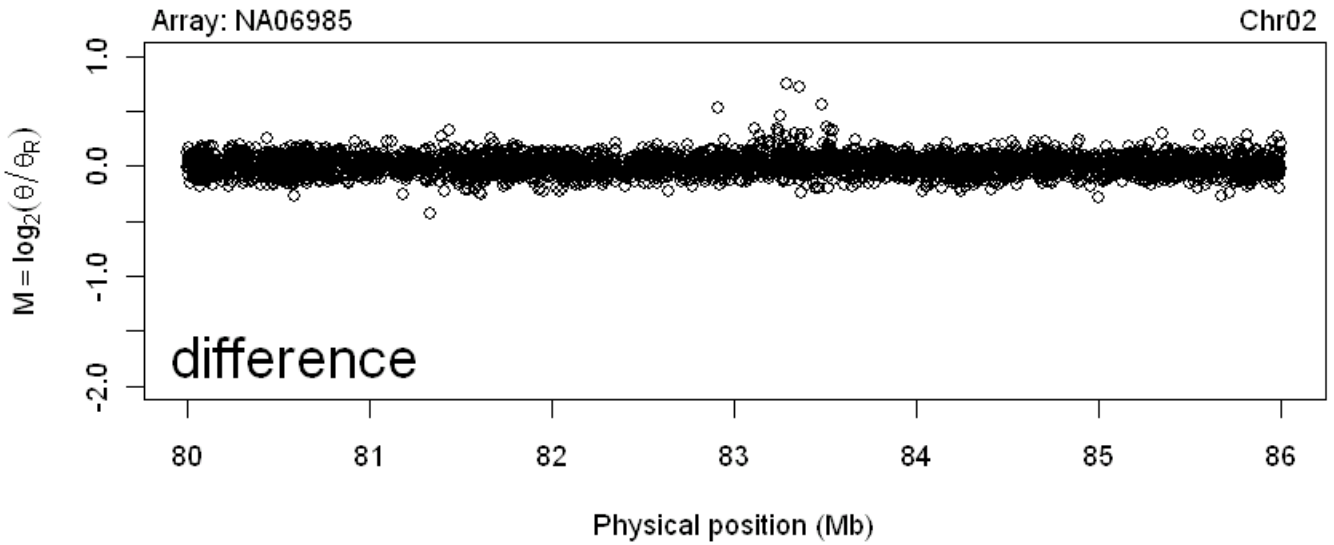


Probe sequence normalization (CRMAv2)

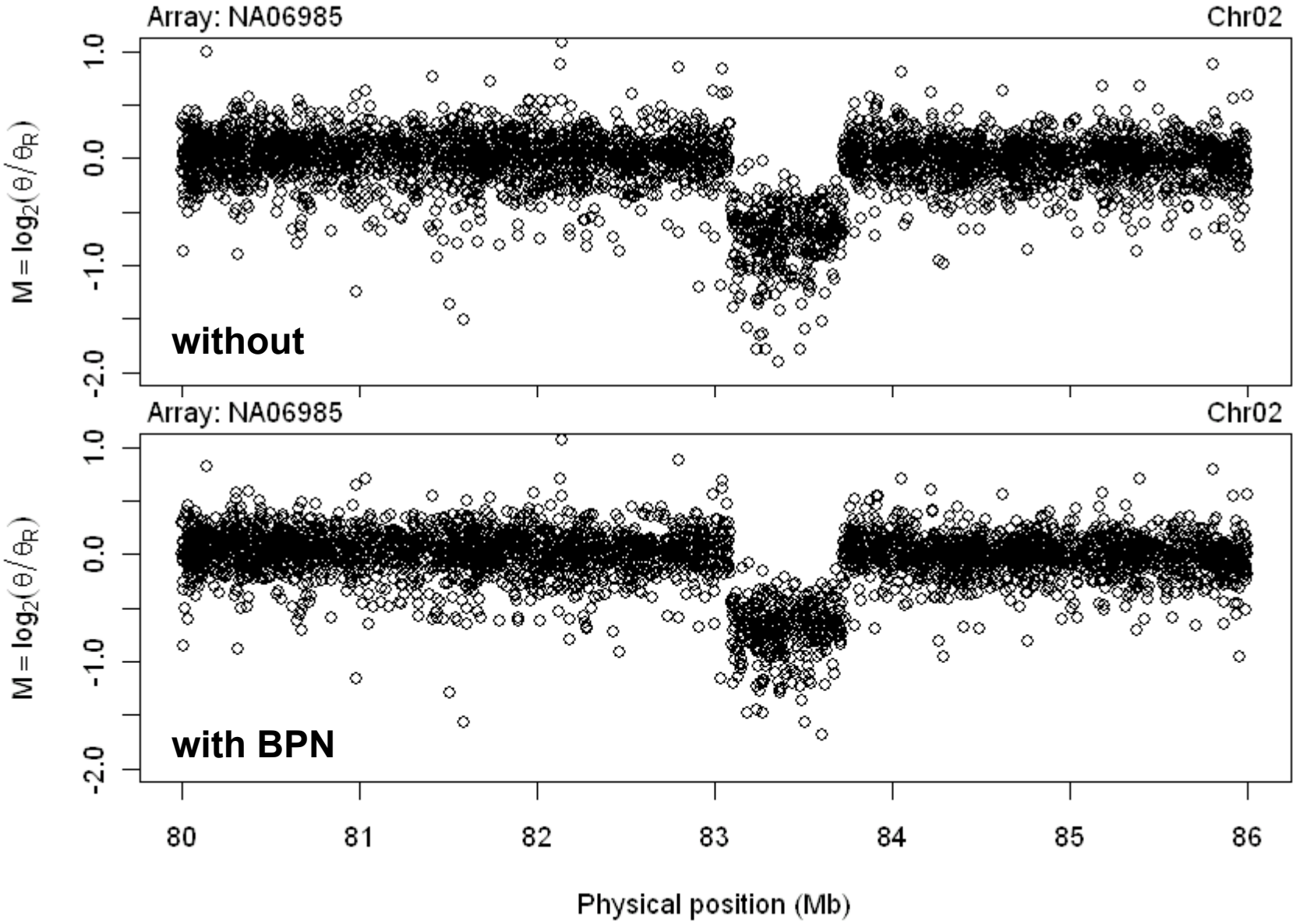
There is a noticeable difference in raw CNs before and after normalization





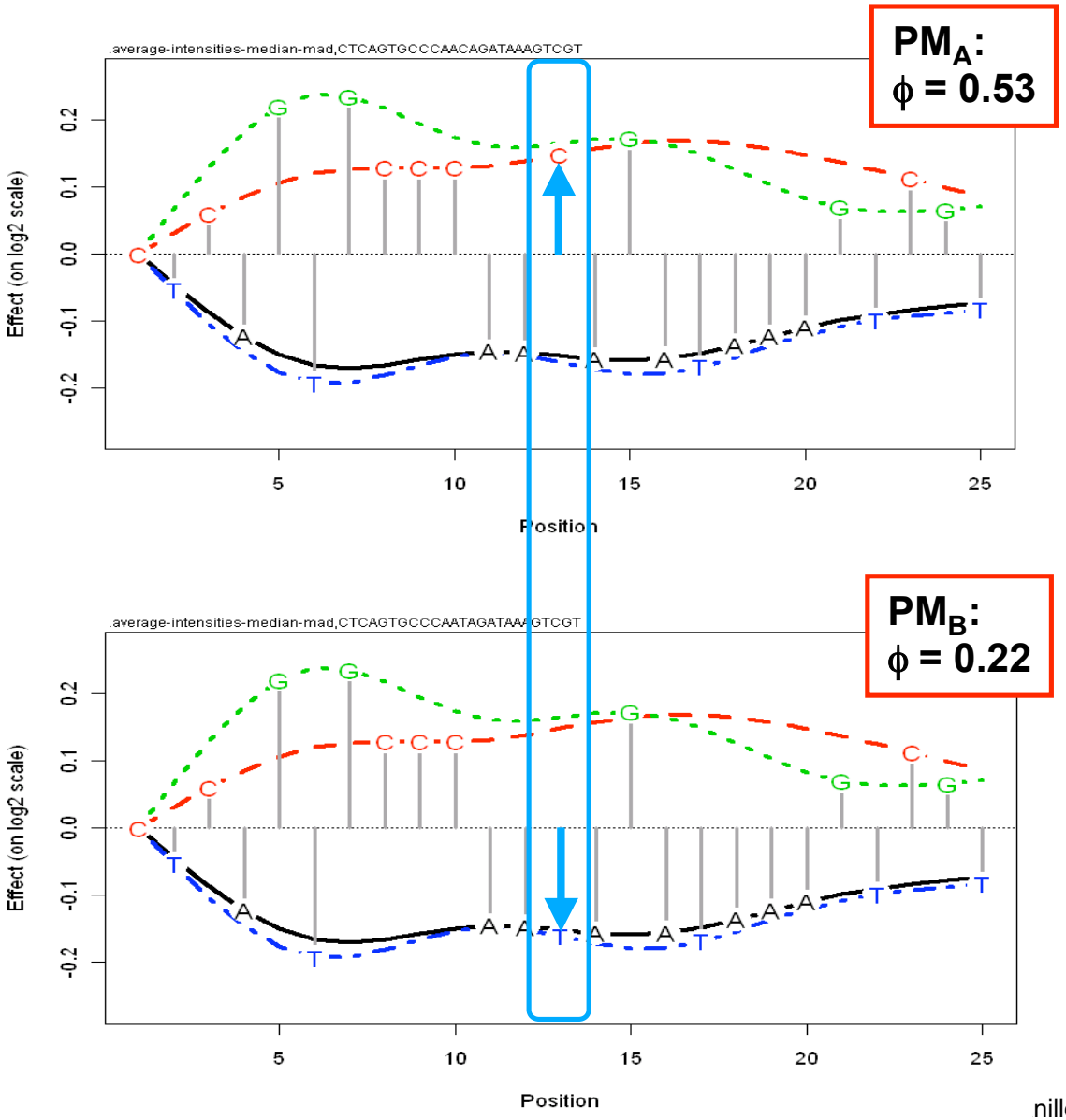


There is a noticeable difference in raw CNs before and after normalization



2. BPN controls for allele A and allele B imbalances

Nucleotide-position normalization controls for imbalances between allele A & allele B

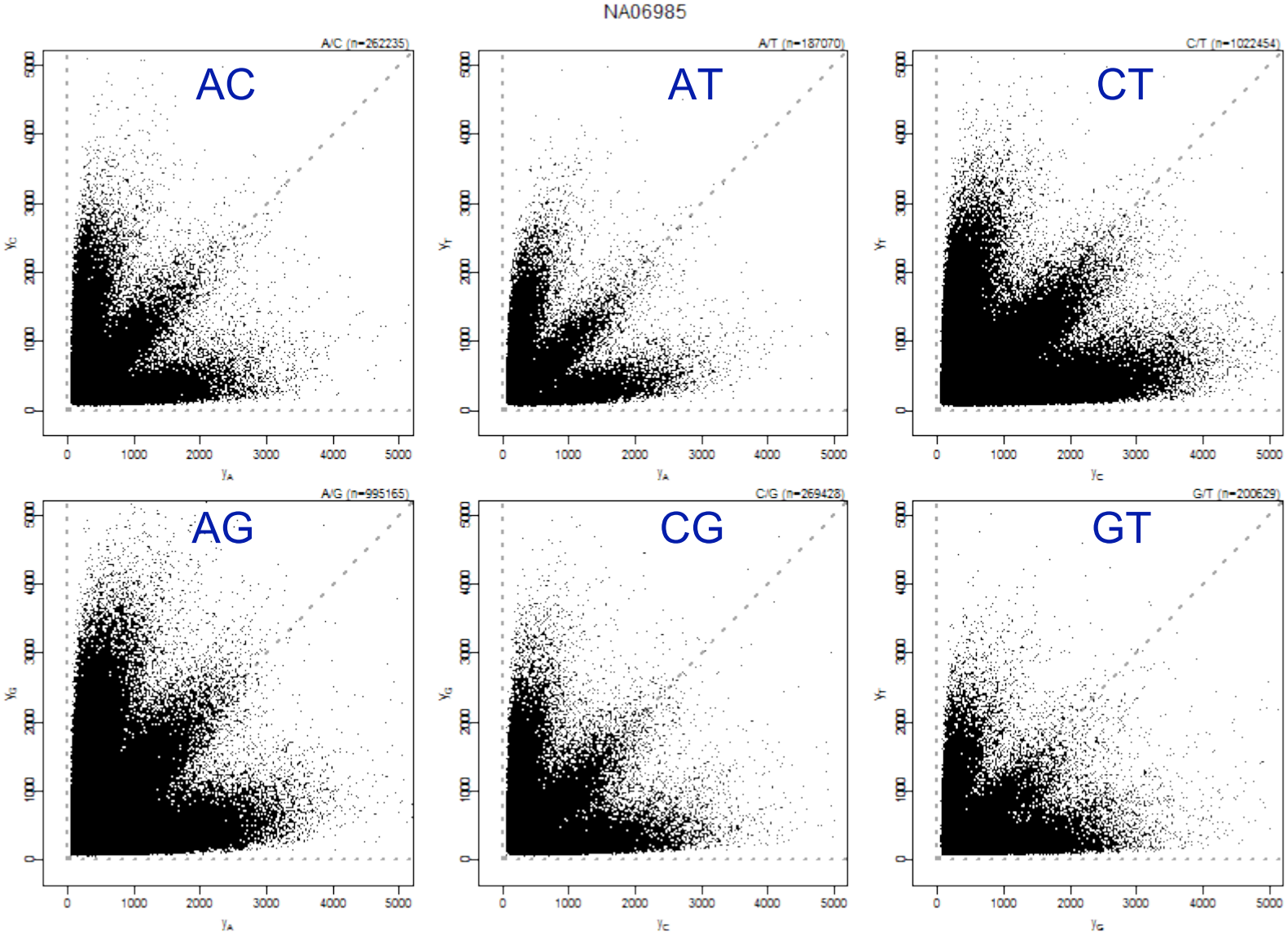


Genotypic imbalances:

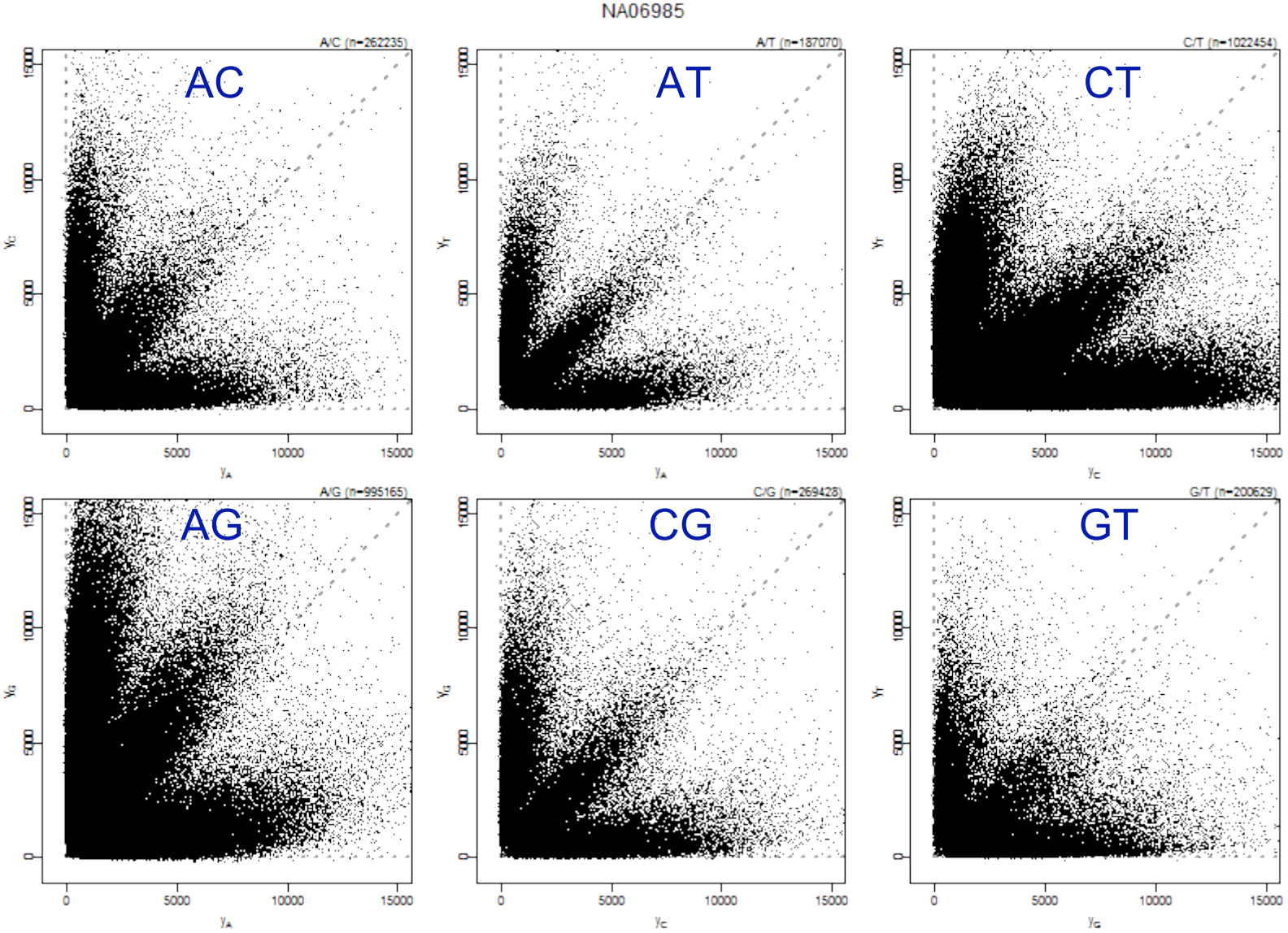
$PM=PM_A+PM_B$:
AA: 0.53+0.53 = 1.06
AB: 0.53+0.22 = 0.75
BB: 0.22+0.22 = 0.44

Thus, AA signals are 2^(1.06-0.44) = 2^0.62 = 1.54 times stronger than BB signals.

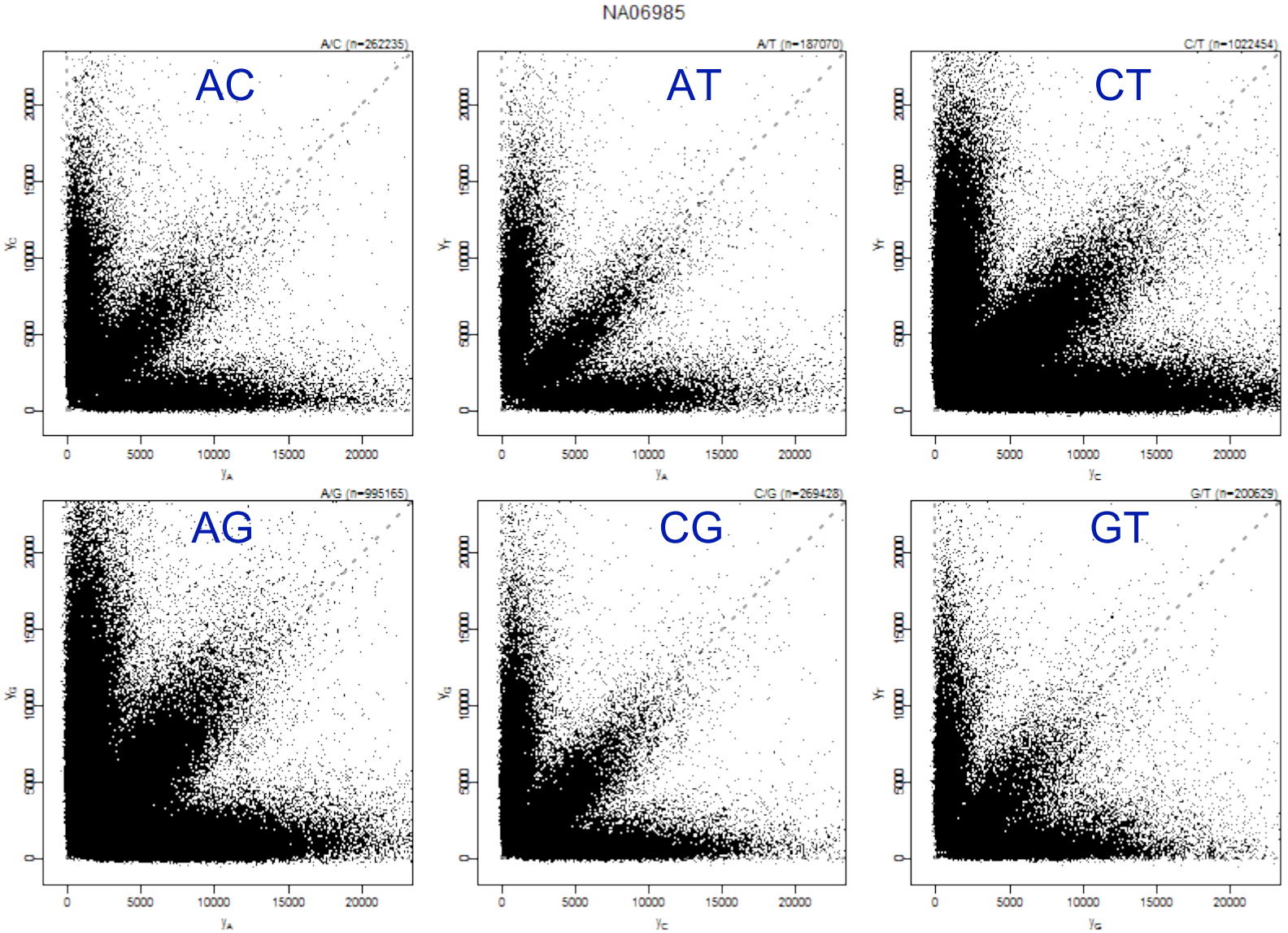
(i) Before calibration there is **crosstalk**
6 *pairs AC, AG, AT, CG, CT & GT*



(ii) After calibration the homozygote arms are more **orthogonal** (note heterozygote arm!)



(iii) After sequence normalization the heterozygote arms are more **balanced**



aroma.affymetrix

You will need:

- Affymetrix CDF, e.g. GenomeWideSNP_6.cdf
- Probe sequences*, e.g. GenomeWideSNP_6.acs

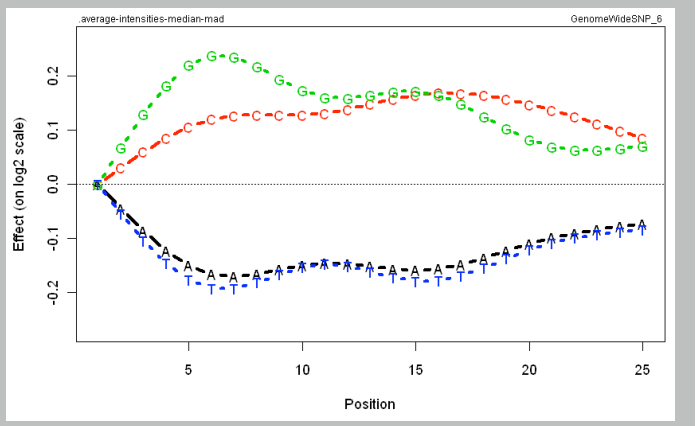
Normalize CEL files:

```
bpn <- BasePositionNormalization(csC, target="zero")
csN <- process(bpn)
```

Works with any chip type, e.g. resequencing,
exon, expression, SNP.

To plot:

```
fit <- getFit(bpn, array=1)
plot(fit)
```



Probe summarization

Probe summarization (CRMAv2)

Summarization on the 2nd generation arrays

- CN units: All single-probe units:
  - Chip-effect estimate: $\theta_{ij} = PM_{ij}$
- SNPs: Identically replicated probe pairs:
  - Probe pairs: (PM_{ijkA}, PM_{ijkB}); k=1,2,3
  - Allele-specific estimates:
    - $\theta_{ijA} = \text{median}_k\{PM_{ijkA}\}$
    - $\theta_{ijB} = \text{median}_k\{PM_{ijkB}\}$

Probe summarization (CRMAv2)

Probe-level summarization (10K-500K)
(if) replicated probes respond differently

For a particular SNP we now have K added signals:

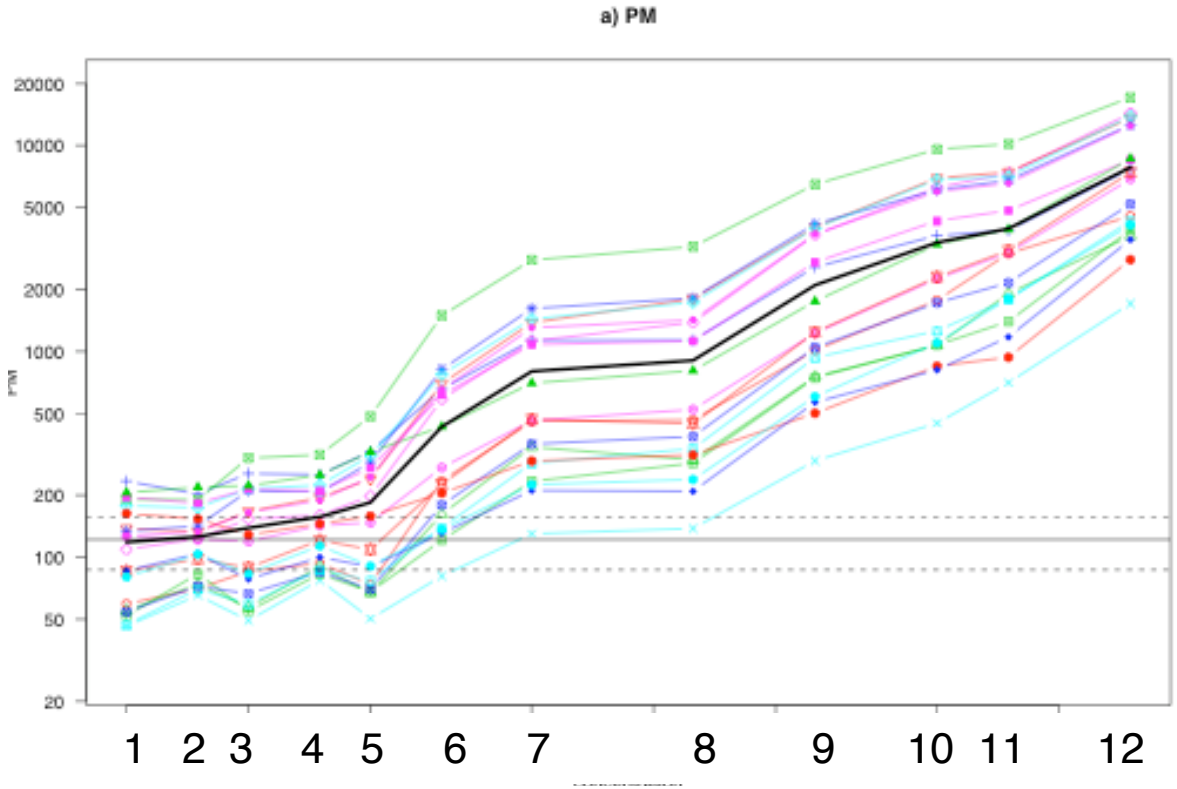
$$(PM_1, PM_2, ..., PM_K)$$

... which are measures of the same thing - the CN. However, they have slightly different sequences, so their hybridization efficiency might differ.

Probe summarization (CRMAv2)

Probe-level summarization
different probes respond differently

log(PM)



12 arrays with different expression levels

18 probes for the same probe set

Example:
$\log_2(PM_1) = \log_2(PM_2) + a_1$
=>
$PM_1 = \phi_1 * PM_2$
$(\phi_1 = 2^{a_1})$

Probe summarization (CRMAv2)

Probe-level summarization
probe affinity model

For a particular SNP, the total CN signal for sample *i=1,2,...,I* is: θ_i

Which we observe via K probe signals: $(PM_{i1}, PM_{i2}, ..., PM_{iK})$

rescaled by probe affinities: $(\phi_1, \phi_2, ..., \phi_K)$

A **multiplicative model** for the observed PM signals is then:

$$PM_{ik} = \phi_k * \theta_i + \xi_{ik}$$

where ξ_{ik} is noise.

Probe summarization (CRMAv2)

Probe-level summarization
the log-affinity model

For one SNP, the model is:

$$PM_{ik} = \phi_k * \theta_i + \xi_{ik}$$

Take the logarithm on both sides:

$$\log_2(PM_{ik}) = \log_2(\phi_k * \theta_i + \xi_{ik})$$
$$\approx \log_2(\phi_k * \theta_i) + \varepsilon_{ik}$$
$$= \log_2\phi_k + \log_2\theta_i + \varepsilon_{ik}$$

Sample *i=1,2,...,I*, and probe *k=1,2,...,K*.

Probe summarization (CRMAv2)

Probe-level summarization
the log-additive model

With multiple arrays i=1,2,...,I, we can estimate the probe-affinity parameters $\{\phi_k\}$ and therefore also the "chip effects" $\{\theta_i\}$ in the model:

$$\log_2(PM_{ik}) = \log_2\phi_k + \log_2\theta_i + \varepsilon_{ik}$$

Conclusion: We have summarized signals (PM_{Ak}, PM_{Bk}) for probes k=1,2,...,K into **one signal θ_i per sample**.

aroma.affymetrix

You will need:

- Affymetrix CDF, e.g. GenomeWideSNP_6.cdf

Summarizing probe signals:

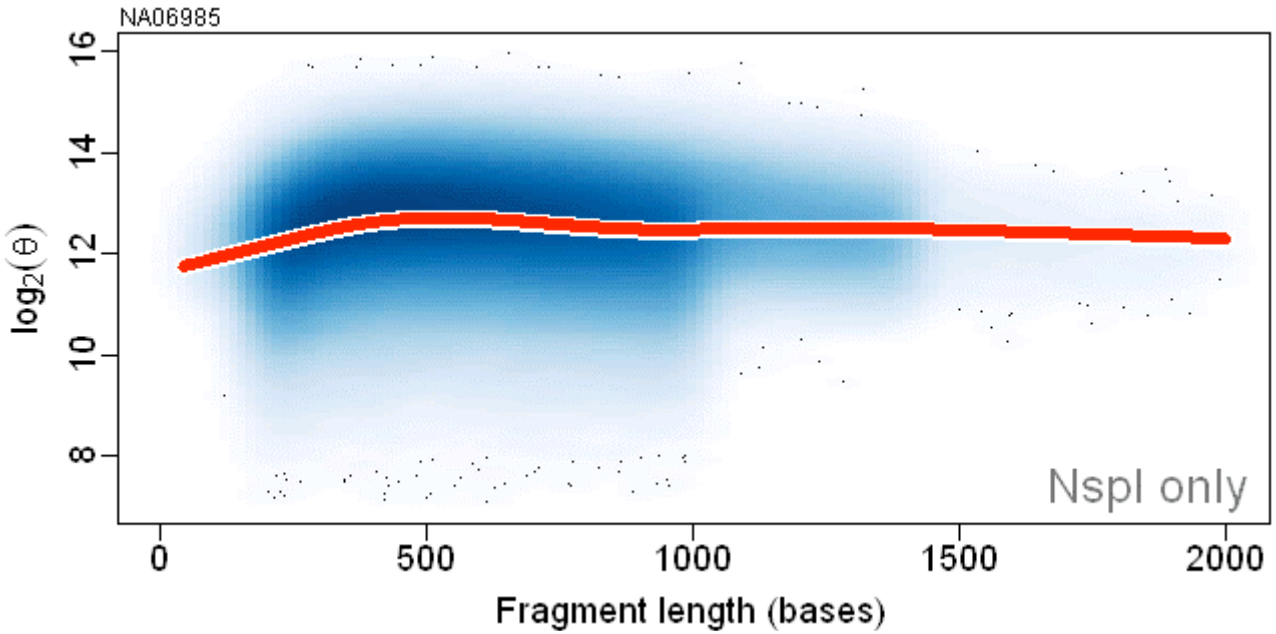
```
plm <- AvgCnPlm(csN, combineAlleles=FALSE)
fit(plm)

ces <- getChipEffectSet(plm)
theta <- extractTheta(ces)
```

Fragment length normalization

Fragment length normalization (CRMAv2)

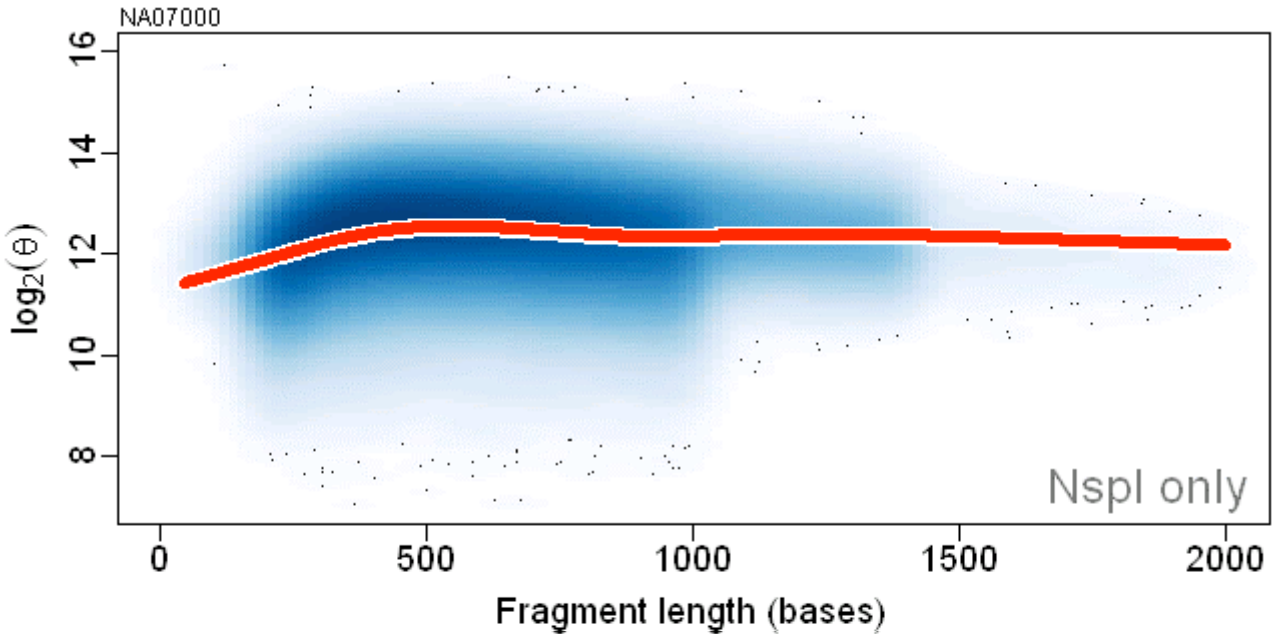
Longer fragments are amplified less by PCR
Observed as weaker θ signals



Note, here we study the effect on non-polymorphic signals, that is, for SNPs we first do $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$.

Fragment length normalization (CRMAv2)

Slightly different effects between arrays adds extra variation



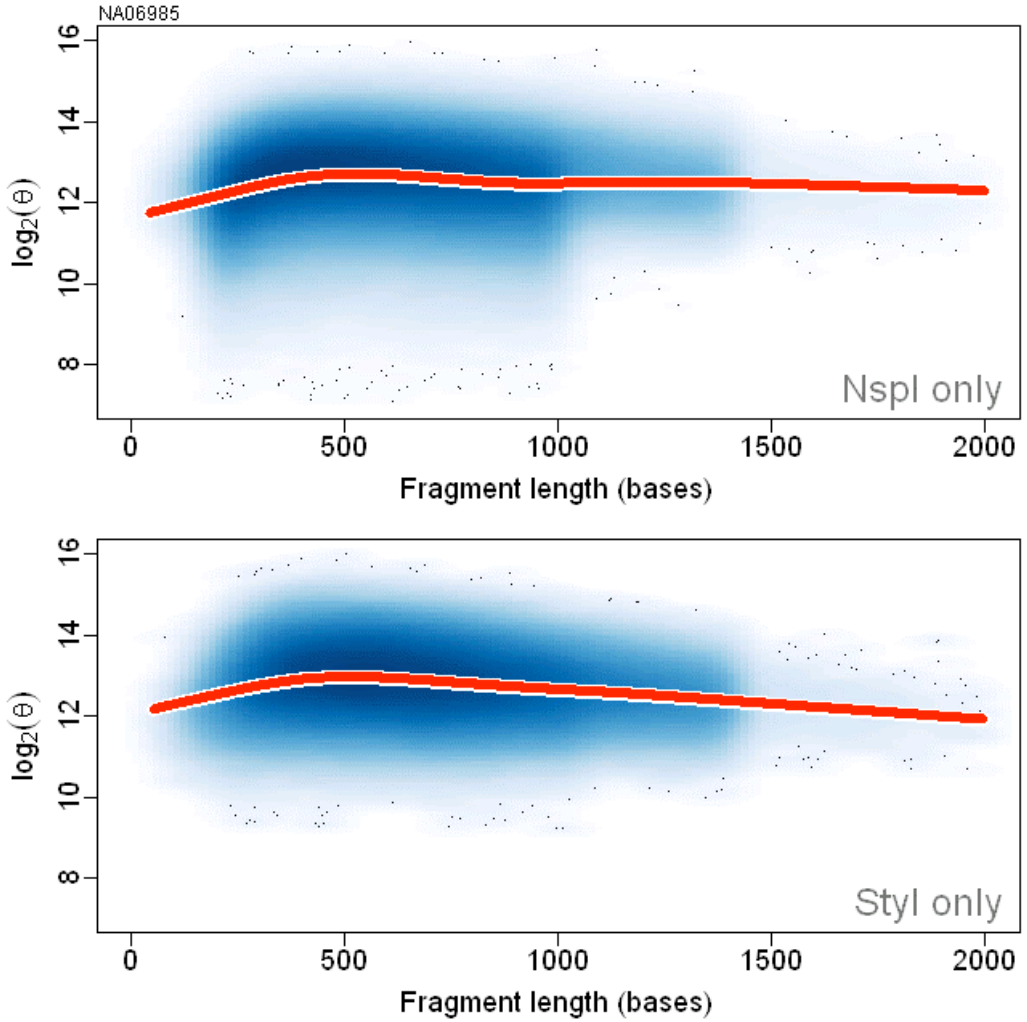
Fragment length normalization (CRMAv2)

Fragment-length normalization for multi-enzyme hybridizations

- For **GWS5.0 and 6.0**, the DNA is fragmented using **two enzymes.**
- For all **CN probes**, all targets originate from ***Nsp*I** digestion.
- For **SNP probes**, some targets originate exclusively from *Nsp*I, exclusively from *Sty*I, or from **both *Nsp*I** and ***Sty*I**.

Fragment length normalization (CRMAv2)

Fragment-length effects for co-hybridized enzymes are assumed to be additive



Fragment length normalization (CRMAv2)

Fragment-length normalization for co-hybridized enzymes

Multi-enzyme normalization model:

$$\log_2\theta_j^* \leftarrow \log_2\theta_j - \delta^*$$

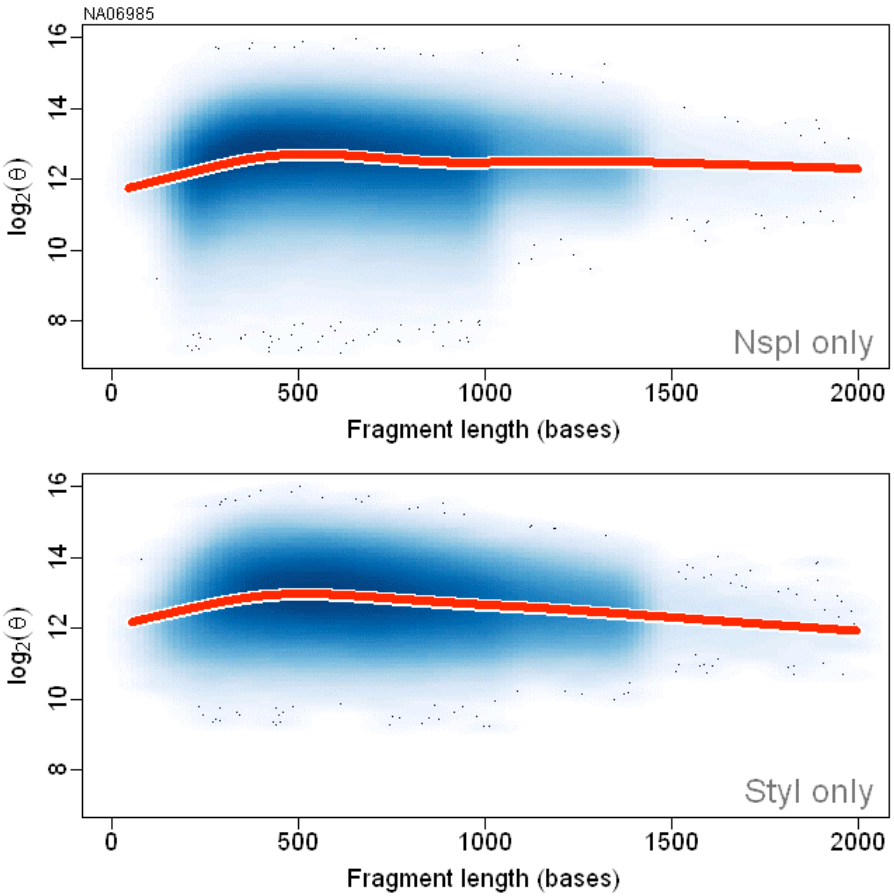
$$\delta^* = \delta(\lambda_{Nsp,j}, \lambda_{Sty,j}) = \text{correction}$$

λ_{Nsp},λ_{Sty} = fragment lengths in *Nsp*I and *Sty*I.

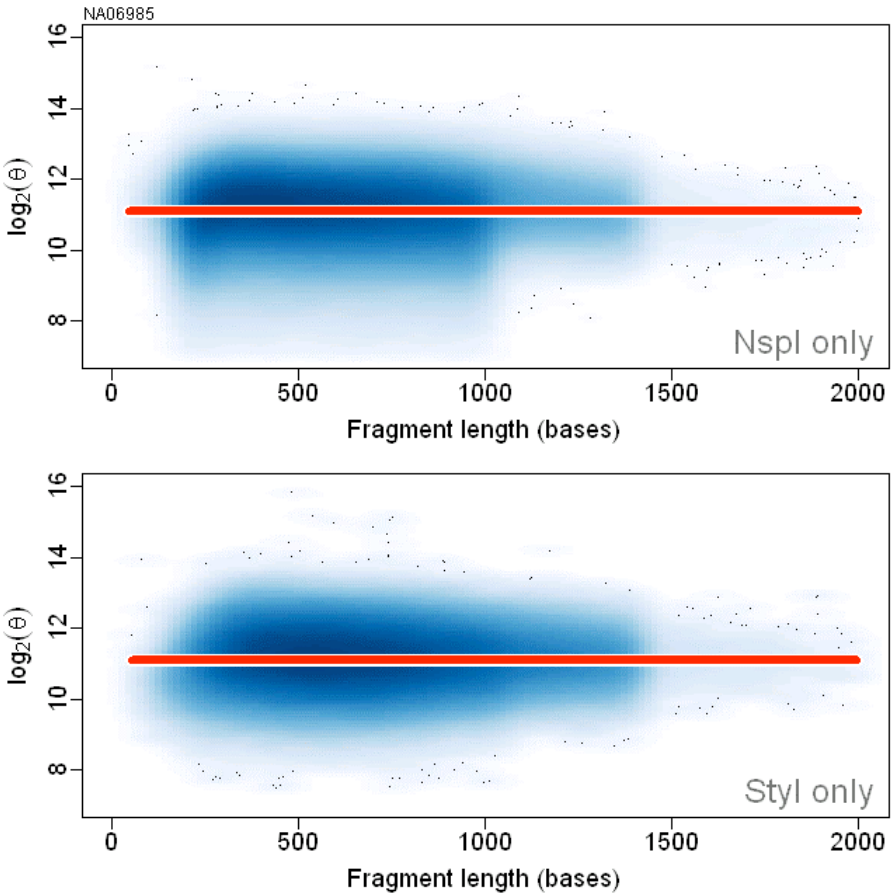
Fragment length normalization (CRMAv2)

Multi-enzyme fragment-length normalization removes the effects

Array #1 before



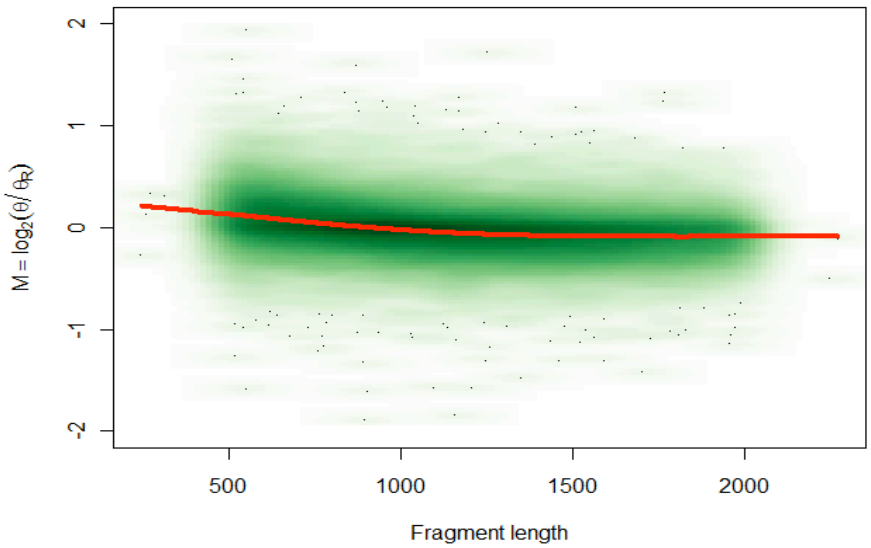
Array #1 after



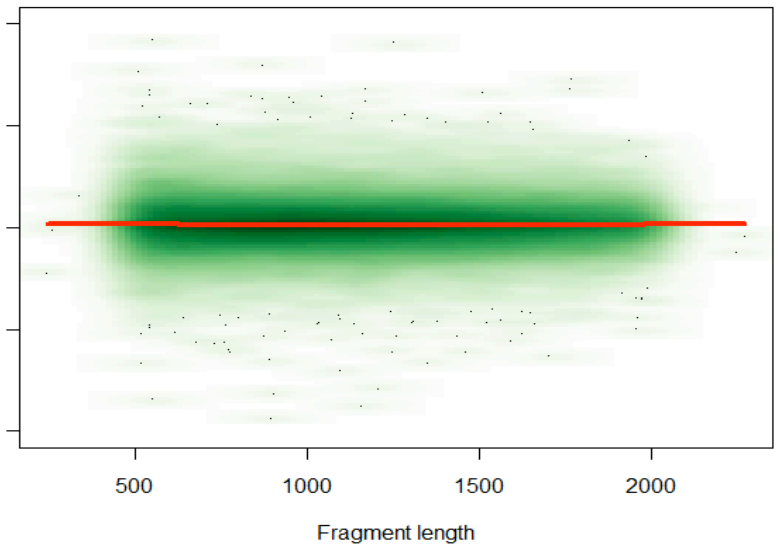
Fragment length normalization (CRMAv2)

Removing the effect on the chip effects, will also remove the effect on CN log ratios

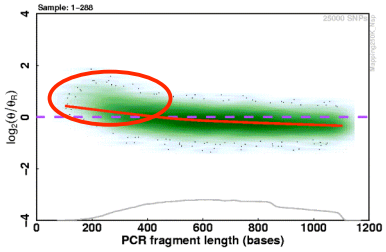
Before



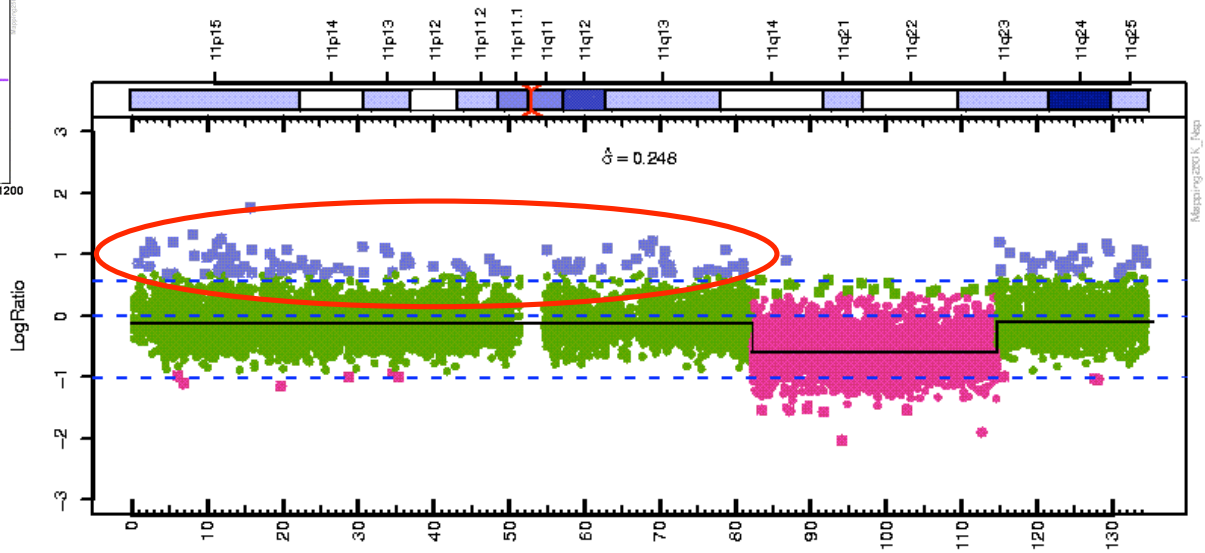
After



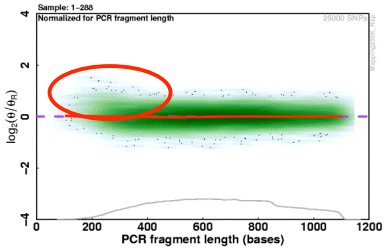
Fragment length normalization (CRMAv2)



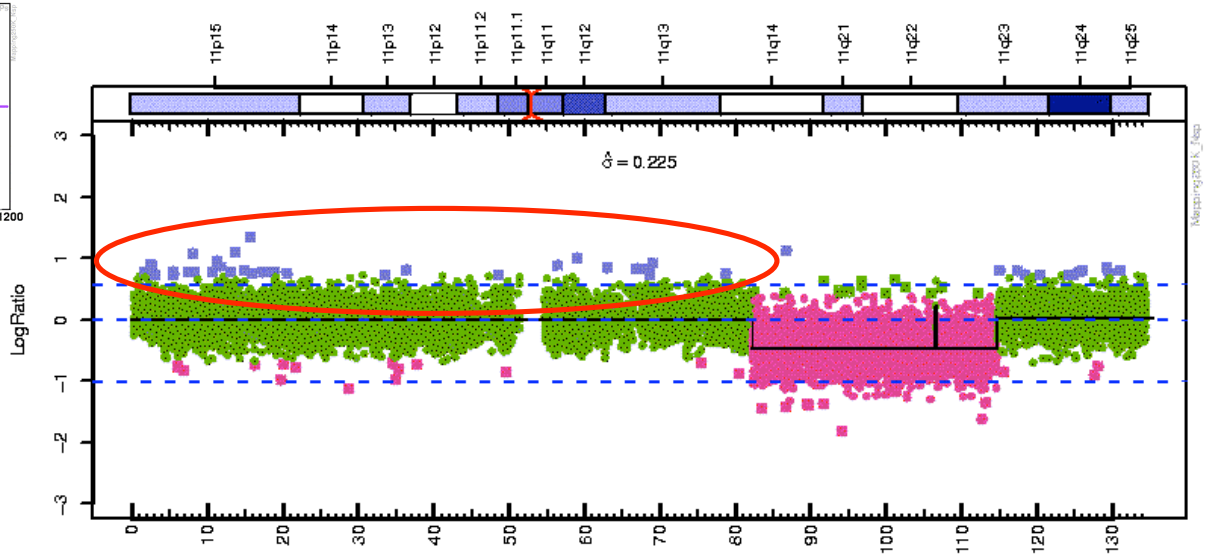
Before



σ = 0.246



After



σ = 0.225

aroma.affymetrix

You will need:

- Affymetrix CDF, e.g. GenomeWideSNP_6.cdf
- A Unit Fragment Length file, e.g. GenomeWideSNP_6.ufl

```
fln <- FragmentLengthNormalization(ces, target="zero")
cesN <- process(fln)
```

Finally, a convenient transform

Convenient transform (CRMAv2)

Bijective transform of (θ_{ijA},θ_{ijB}) in to (θ_{ij}, β_{ij}).

Transform (θ_{ijA},θ_{ijB}) to (θ_{ij}, β_{ij}) by:

Non-polymorphic SNP signal: $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$

Allele B frequency signal: $\beta_{ij} = \theta_{ijB} / \theta_{ij}$

A CN probe does not have a β_{ij}. However, both CN probes and SNPs have a non-polymorphic signal θ_{ij}.

We expect the following:

Genotype BB: $\theta_{ijB} >> \theta_{ijA}$ => $\beta_{ij} \approx 1$

Genotype AA: $\theta_{ijB} << \theta_{ijA}$ => $\beta_{ij} \approx 0$

Genotype AB: $\theta_{ijB} \approx \theta_{ijA}$ => $\beta_{ij} \approx ½$

Thus, θ_{ij} carry information on CN and β_{ij} on genotype.

Convenient transform (CRMAv2)

Copy numbers are estimated relative to a reference

Relative copy numbers:

$$C_{ij} = 2*(\theta_{ij} / \theta_{Rj})$$

Alternatively, log-ratios:

$$M_{ij} = \log_2(\theta_{ij} / \theta_{Rj})$$

Note: C_{ij} is defined also when $\theta <= 0$, but M_{ij} is not.

Array i=1,2,...,I. Locus j=1,2,...,J.

Convenient transform (CRMAv2)

Allele-specific copy numbers

Allele-specific copy numbers (C_{ijA},C_{ijB}):

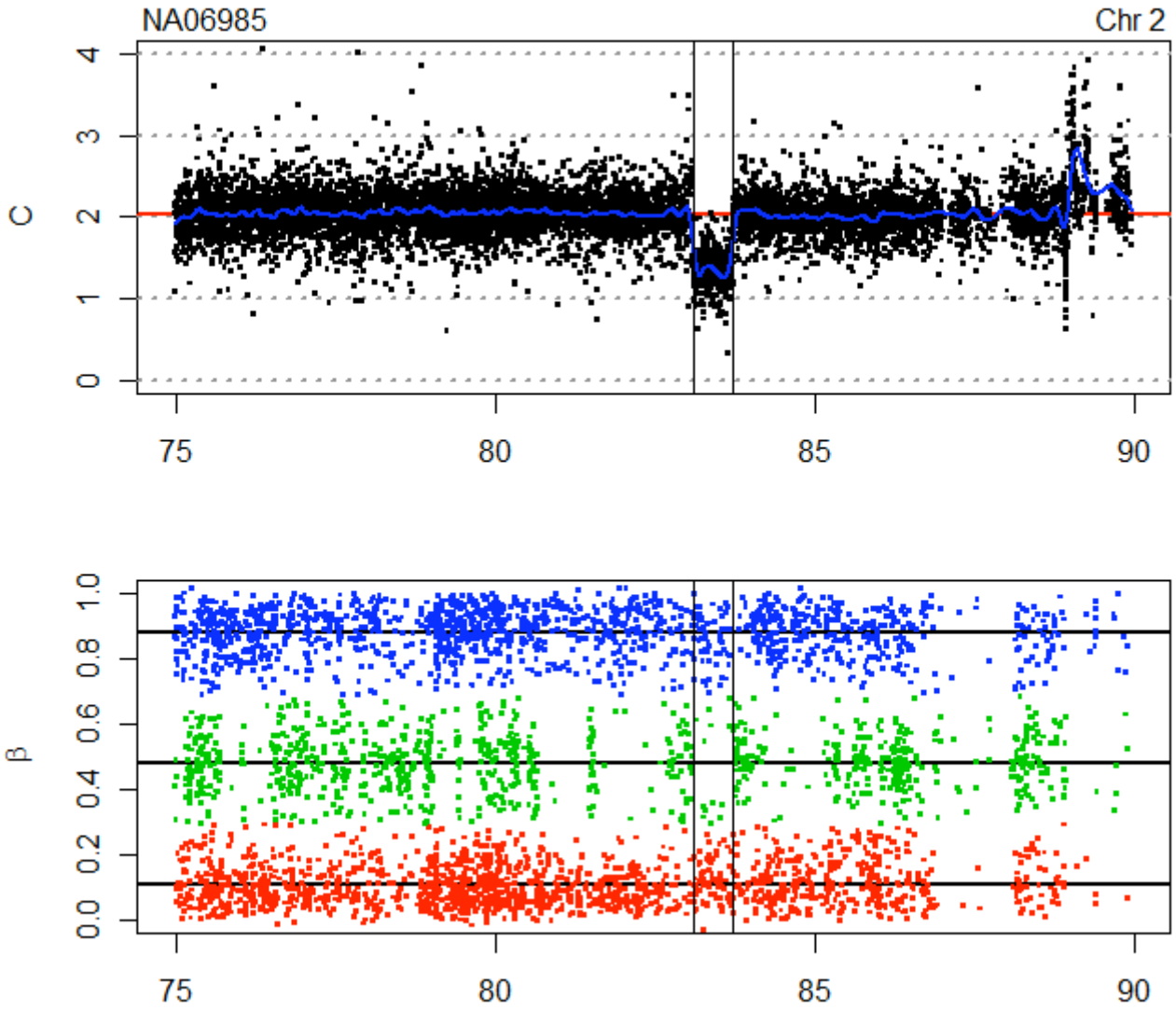
$$C_{ijA} = 2*(\theta_{ijA} / \theta_{Rj})$$
$$C_{ijB} = 2*(\theta_{ijB} / \theta_{Rj})$$

Note that,

1. $C_{ij} = C_{ijA}+C_{ijB} = 2*(\theta_{ijA}+ \theta_{Rj}) / \theta_{Rj} = 2*(\theta_{ij} / \theta_{Rj})$
2. $C_{ijB}/C_{ij} = [2*(\theta_{ijB} / \theta_{Rj})] / [2*(\theta_{ij} / \theta_{Rj})] = \theta_{ijB} / \theta_{ij} = \beta_{ij}$
3. $C_{ijB} = 2*(\theta_{ijB} / \theta_{ij}) * (\theta_{ij}/ \theta_{Rj}) = \beta_{ij} * C_{ij}$

Convenient transform (CRMAv2)

CN and freqB - (C,β) - along genome



aroma.affymetrix

You will need:

- Affymetrix CDF, e.g. GenomeWideSNP_6.cdf
- A Unit Genome Position file, e.g. GenomeWideSNP_6.ugp

```
data <- extractTotalAndFreqB(cesN)
theta <- data[,"total",]
freqB <- data[,“freqB",]

Plot Array 3 along chromosome 2
gi <- getGenomeInformation(cdf)
units <- getUnitsOnChromosome(gi, 2)
pos <- getPositions(gi, units)
plot(pos, theta[units,3])
plot(pos, freqB[units,3])
```

Estimate copy numbers from SNP arrays

	CRMA v2
Preprocessing *(probe signals)*	1. Allelic crosstalk calibration 2. Probe-sequence normalization
Summarization	Robust averaging: CN probes: $\theta_{ij} = PM_{ij}$ SNPs: $\theta_{ijA} = \text{median}_k(PM_{ijkA})$ $\theta_{ijB} = \text{median}_k(PM_{ijkB})$ array *i*, loci *j*, probe *k.*
Post-processing	PCR fragment-length normalization
Transform	$(\theta_{ijA},\theta_{ijB}) \Rightarrow (\theta_{ij}, \beta_{ij})$ $\theta_{ij} = \theta_{ijA}+\theta_{ijB}$, $\beta_{ij} = \theta_{ijB} / \theta_{ij}$
Allele-specific & Total Copy Nbs	$C_{ijA} = 2*(\theta_{ijA} / \theta_{Rj})$ and $C_{ijB} = 2*(\theta_{ijA} / \theta_{Rj})$ $C_{ij} = 2*(\theta_{ij} / \theta_{Rj})$ reference *R*