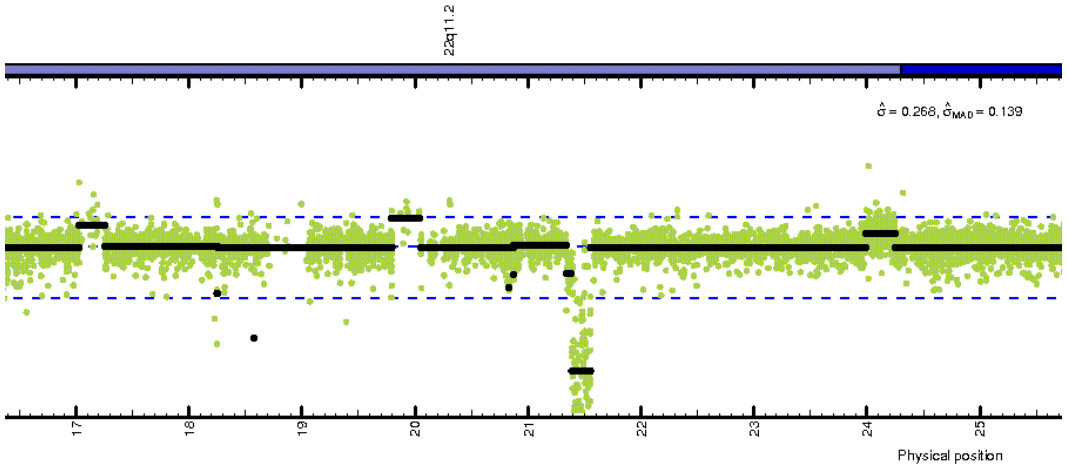
Copy-number estimation on the latest generation of high-density oligonucleotide microarrays



Henrik Bengtsson
(work with Terry Speed)
Dept of Statistics, UC Berkeley

January 24, 2008

Postdoctoral Seminars, Mathematical Biosciences Institute, The Ohio State University

Acknowledgments

UC Berkeley:
James Bullard
Kasper Hansen
Elizabeth Purdom
Terry Speed

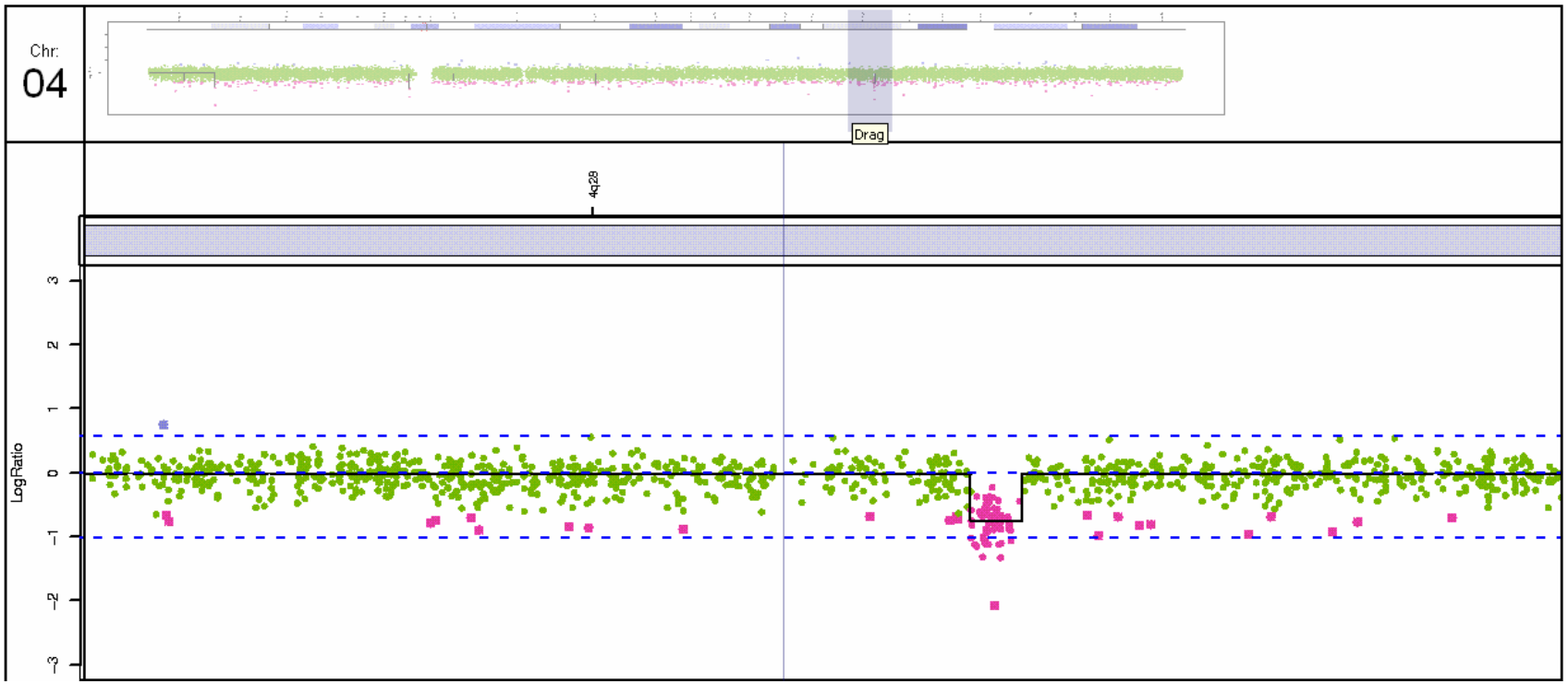
WEHI, Melbourne:
Mark Robinson
Ken Simpson

ISREC, Lausanne:
“Asa” Wirapati

John Hopkins:
Benilton Carvalho
Rafael Irizarry

Affymetrix, California:
Ben Bolstad
Simon Cawley
Luis Jevons
Chuck Sugnet
Jim Veitch

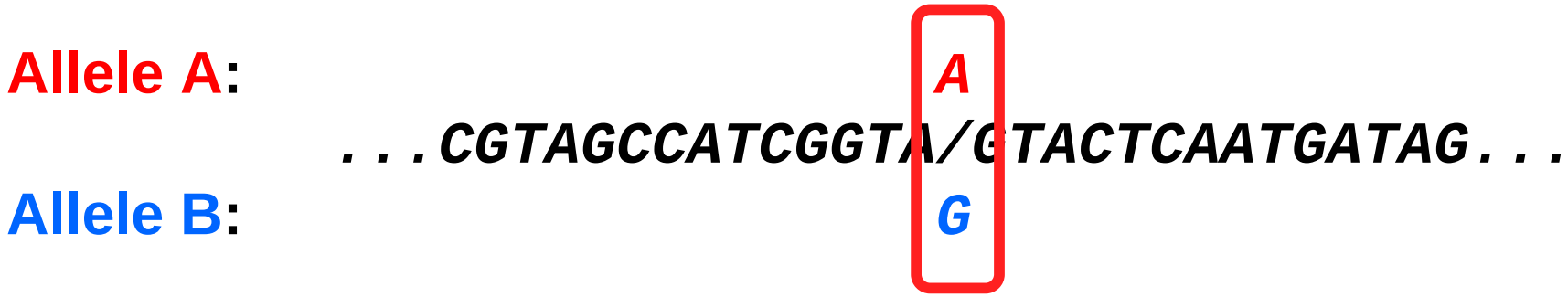
Copy number analysis is about finding "aberrations" in a person's genome.



Size = 264 kb, Number of loci = 72

Single Nucleotide Polymorphisms (SNPs) make us unique

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



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

Human Genetic Variation: Breakthrough of the Year 2007 (Science)

- 3 billion DNA bases.
- First sequenced 2001.

- HapMap: 270 individuals genotyped.
  3 million known SNPs (places where one base differ from one person to another).
  Estimate: 15 million SNPs.
- Genomewide association studies take over (over linkage analysis).

- Copy Number Polymorphism:
  - 1,000s to millions of bases lost or added.
  - Estimate: 20% of differences in gene activity are due to copy-number variants; SNPs (genotypes) account for the rest.

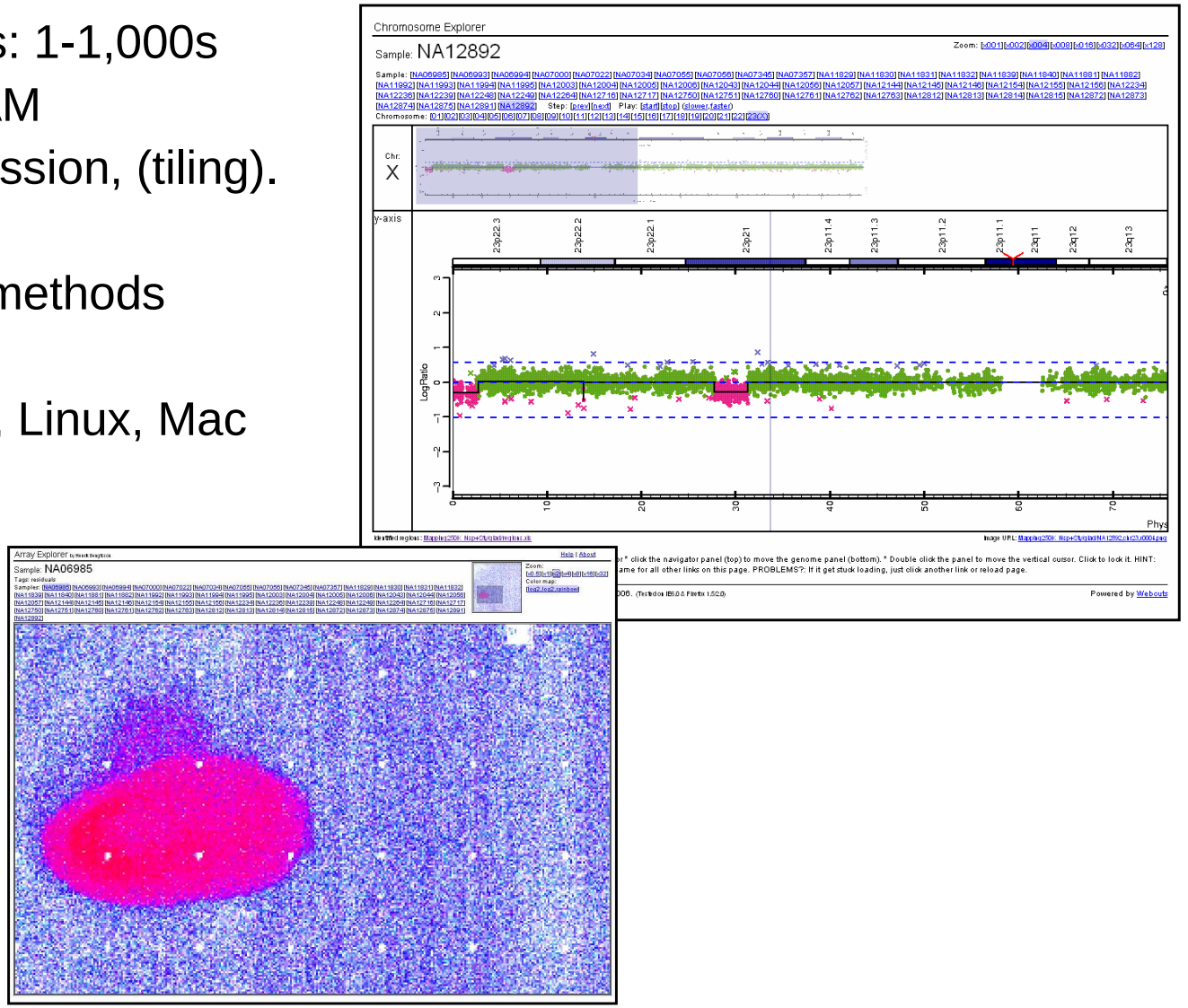
- January 22, 2008: The 3-year "1,000 Genomes Project" will sequence 1,000 individuals. This follows the HapMap Project (SNPs).

Objectives of this presentation

- Total copy number estimation/segmentation
- Estimate single-locus CNs well
  (segmentation methods take it from there)
- All generations of Affymetrix SNP arrays:
  - SNP chips: 10K, 100K, 500K
  - SNP & CN chips: 5.0, 6.0
- Small and very large data sets

Available in aroma.affymetrix

“Infinite” number of arrays: 1-1,000s
Requirements: 1-2GB RAM
Arrays: SNP, exon, expression, (tiling).
Dynamic HTML reports
Import/export to existing methods
Open source: R
Cross platform: Windows, Linux, Mac



Affymetrix chips

Running the assay
take 4-5 working days

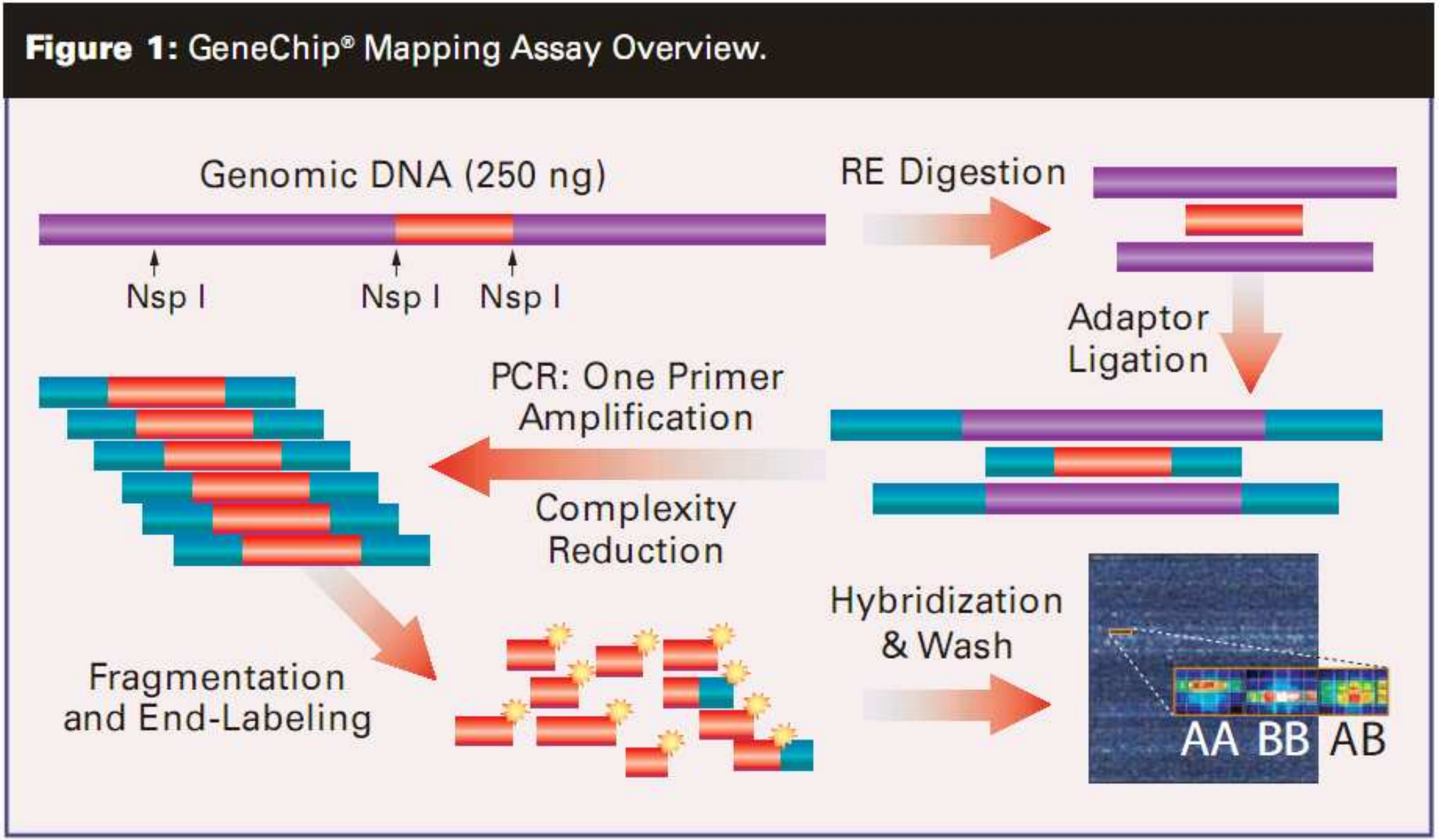
1. Start with target ***gDNA*** (genomic DNA) or ***mRNA*****.**

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

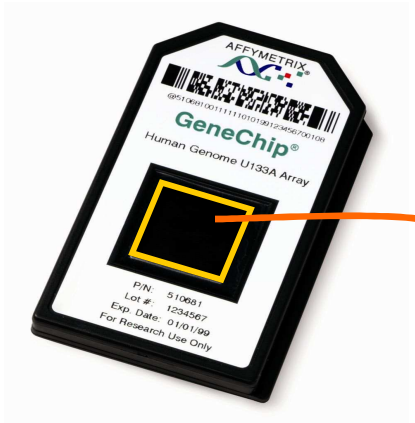
3. After hybridization, washing, and scanning we get a **digital image**.

4. Image summarized across pixels to ***probe-level intensities*** before we begin. Thisis our "**raw data**".

Restriction enzymes digest the DNA, which is then amplified and hybridized



The Affymetrix GeneChip is a synthesized high-density 25-mer microarray



1.28 cm

1.28 cm

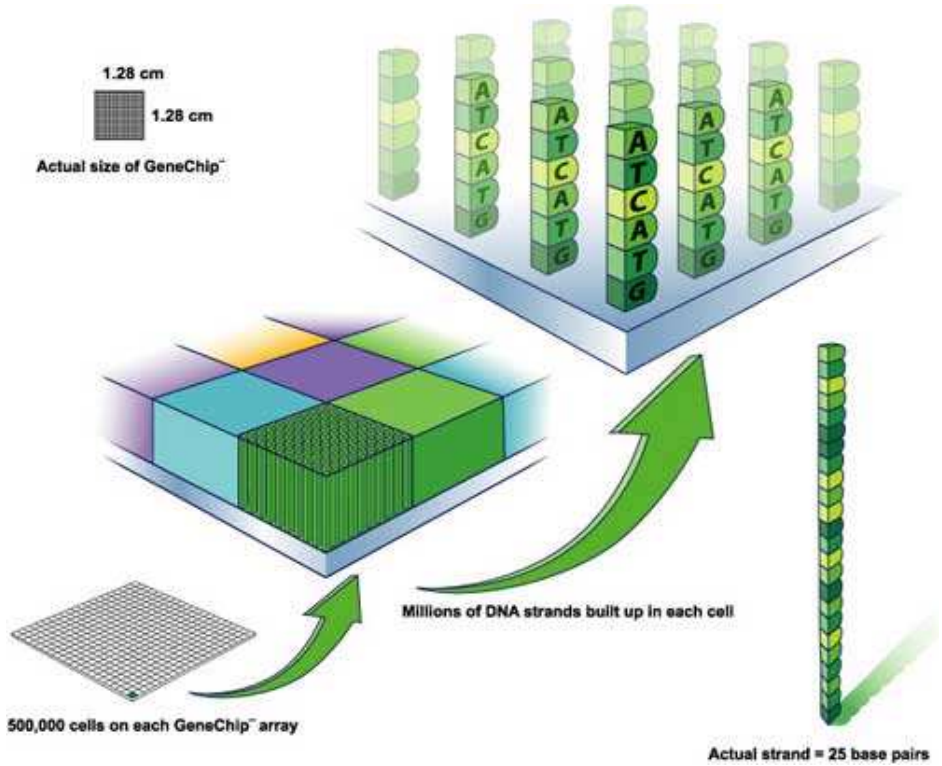
5 µm

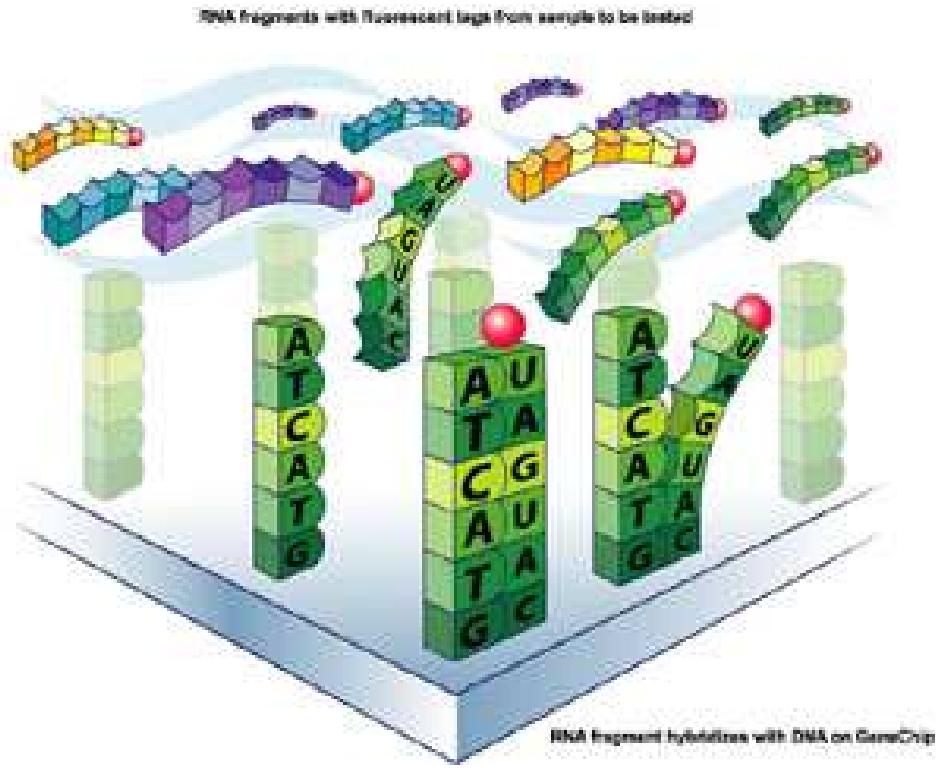
5 µm

> 1 million identical 25bp sequences

6.5 million probes/chip

Target DNA find their way to complementary probes by massive parallel hybridization





Hybridization + Scanning

DAT File(s)
[Image, pixel intensities]

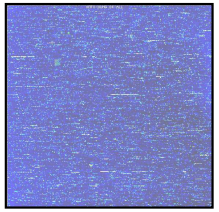


Image analysis

workable raw data

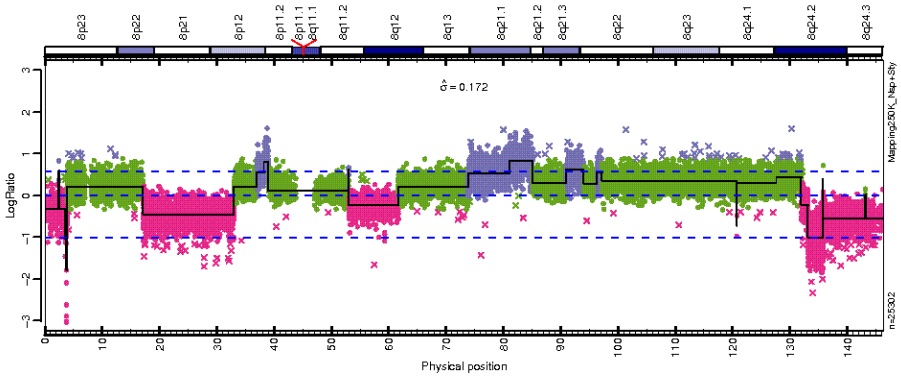
CEL File(s)
[Probe Cell Intensity]

+

CDF
[Chip Description File]

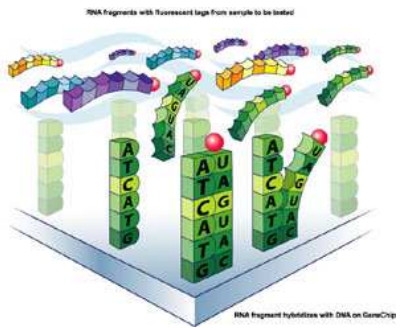
Pre-processing

Segmentation



Affymetrix copy-number & genotyping arrays

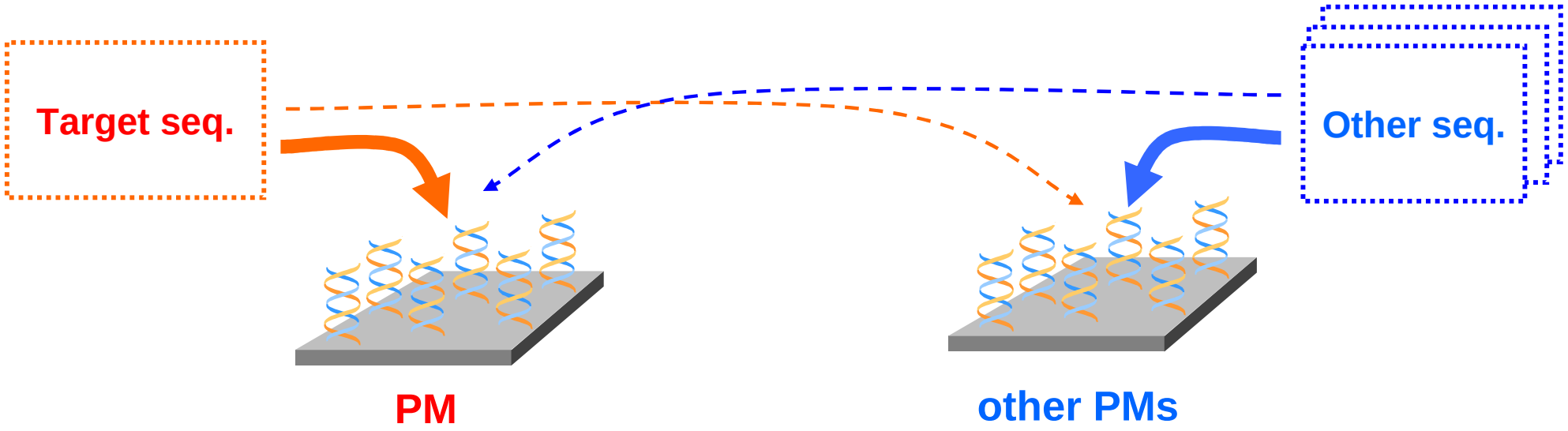
Terminology



Target sequence: ***...CGTAGCCATCGGTAAGTACTCAATGATAG...***

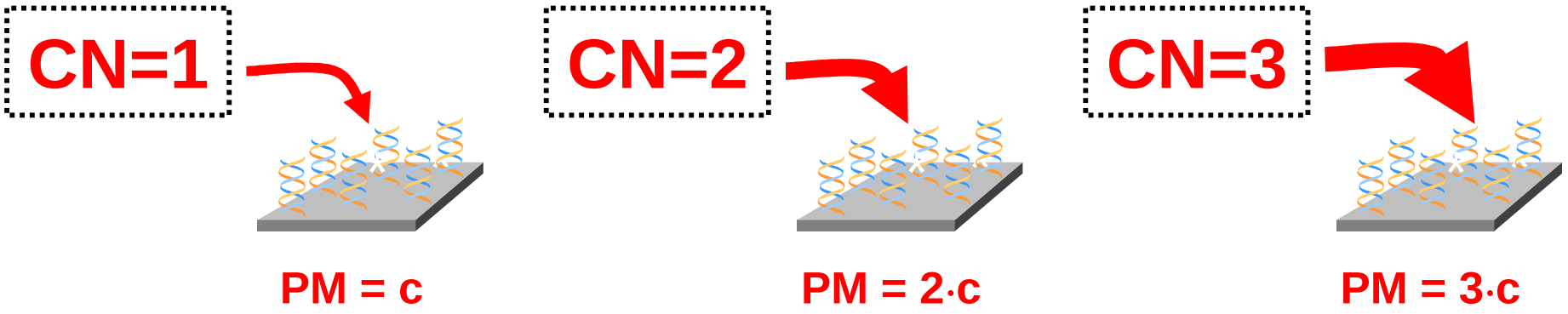
Perfect match (PM): **ATCGGTAGCCATTCATGAGTTACTA**

25 nucleotides



Copy-number probes are used to quantify the amount of DNA at known loci

CN locus: ***. . .CGTAGCCATCGGTAAGTACTCAATGATAG. . .***
PM: **ATCGGTAGCCATTCATGAGTTACTA**



Raw copy numbers
- log-ratios relative to a reference

From the preprocessing, we obtain for sample i=1,2,...,I, CN locus j=1,2,...,J:

$$\text{Observed signals: } (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iJ})$$

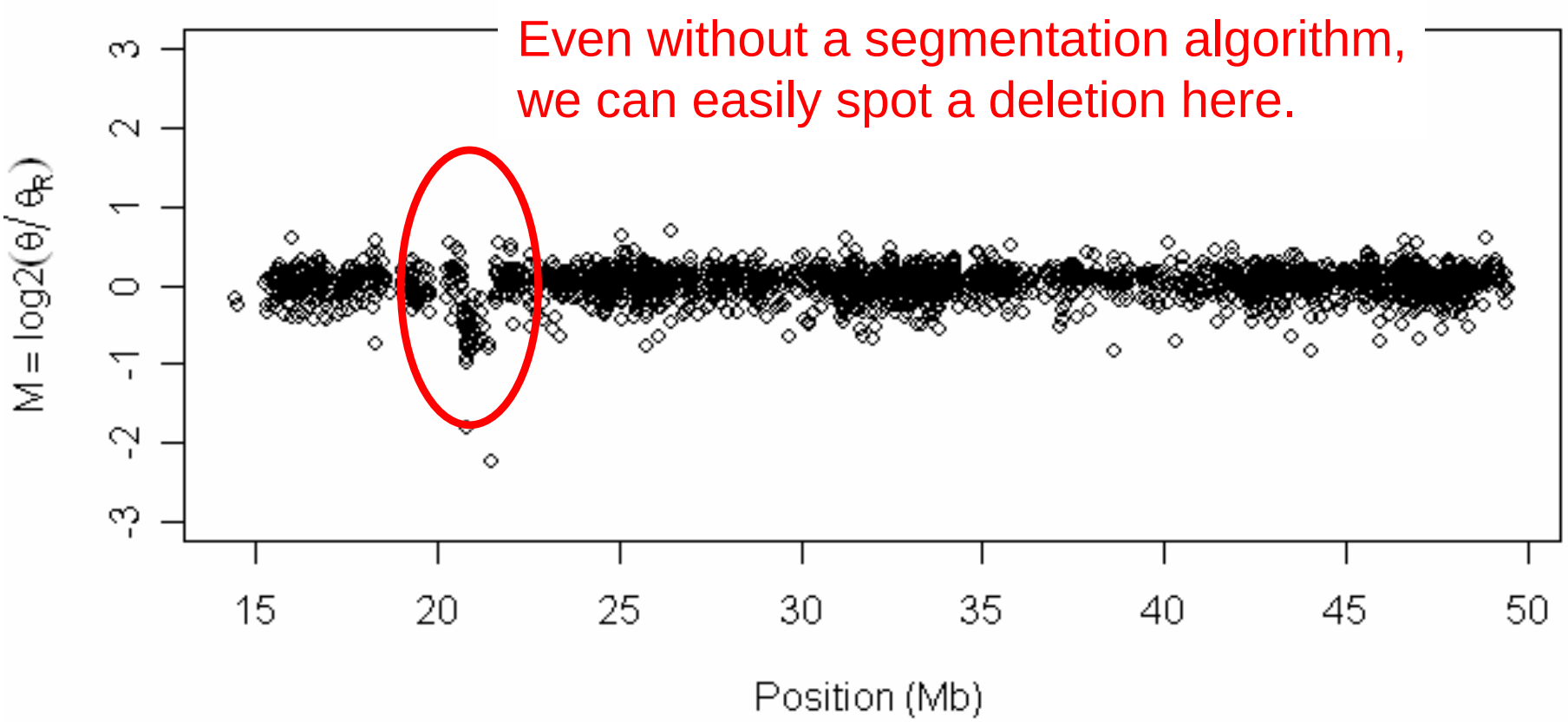
These are not absolute copy-number levels. In order to interpret these, we compare each of them to a reference "**R**", i.e. $\theta_{ij} / \theta_{Rj}$, but even better **"raw copy numbers"**:

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

The reference can be from normal tissue, or from a pool of normal samples.

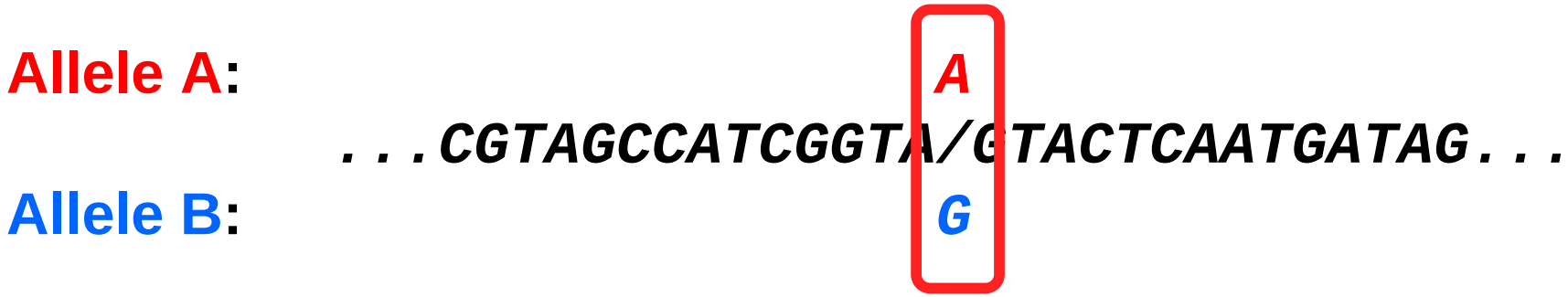
Copy number regions are found by lining up estimates along the chromosome

Example: Log-ratios for one sample on Chromosome 22.



Single Nucleotide Polymorphisms (SNPs) make us unique

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



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

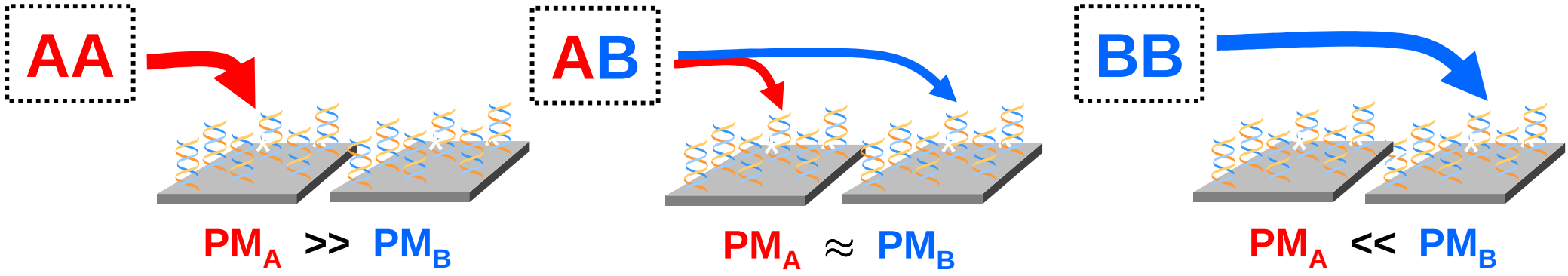
Affymetrix probes for a SNP
- can be used for genotyping

PM_A: ATCGGTAGCCATTCATGAGTTACTA

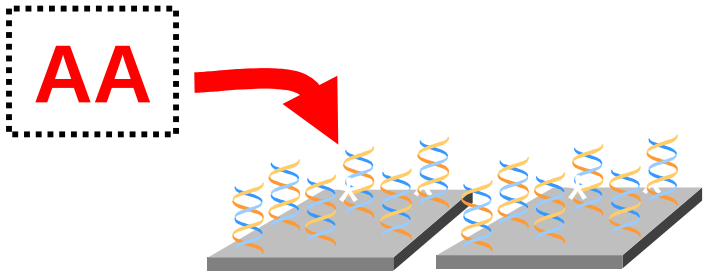
Allele A: ...CGTAGCCATCGGTAAGTACTCAATGATAG...

Allele B: ...CGTAGCCATCGGTAGGTACTCAATGATAG...

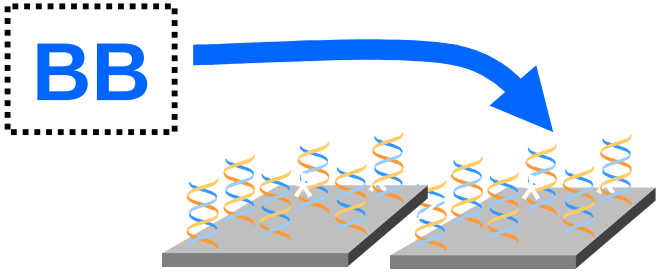
PM_B: ATCGGTAGCCATCCATGAGTTACTA



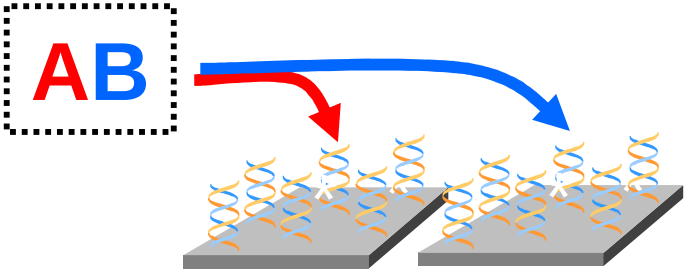
SNPs can also be used for estimating copy numbers



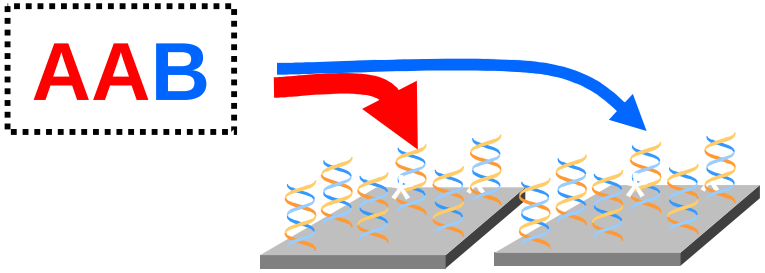
$PM = PM_A + PM_B = 2c$



$PM = PM_A + PM_B = 2c$

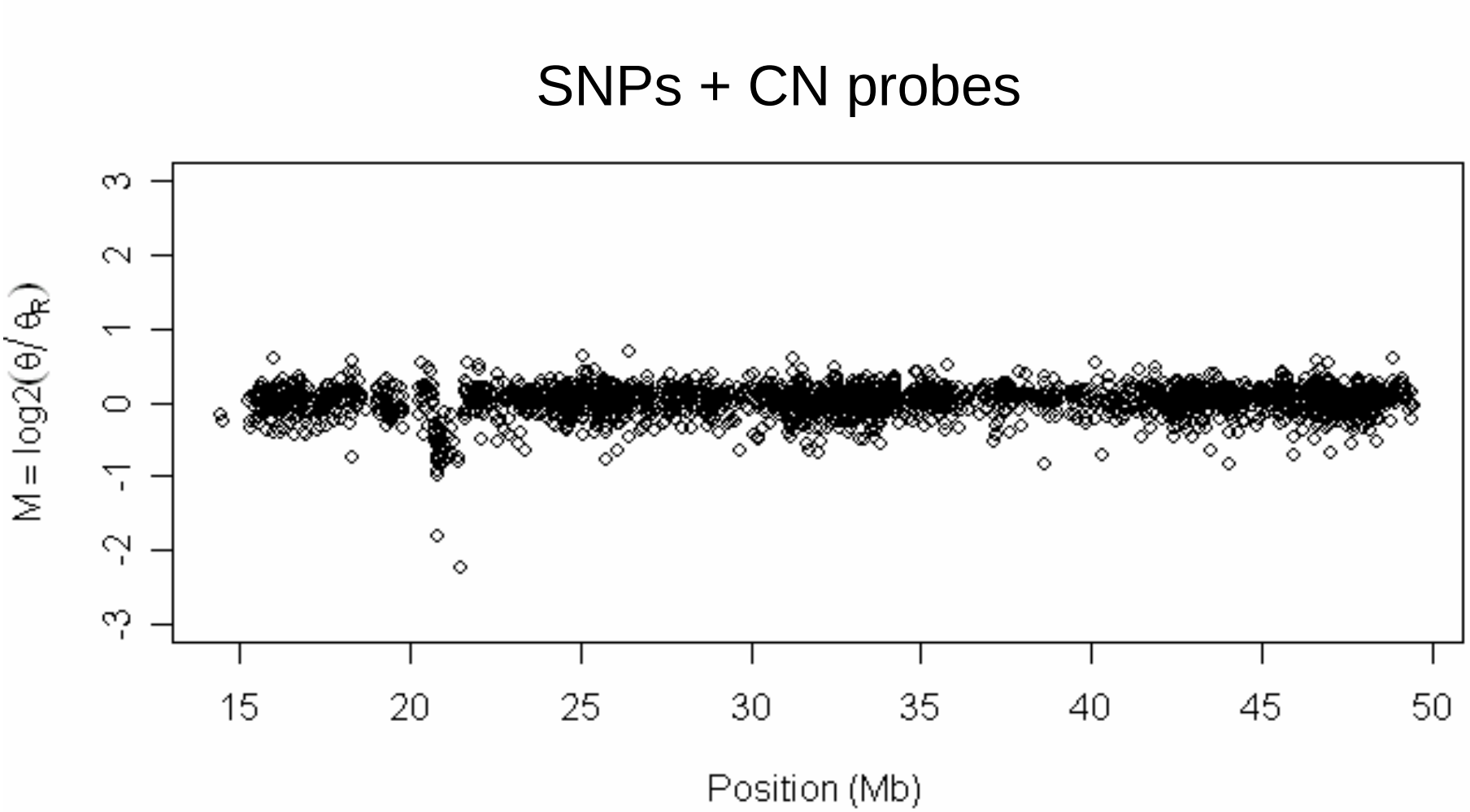


$PM = PM_A + PM_B = 2c$



$PM = PM_A + PM_B = 3c$

Combing CN estimates from SNPs and CN probes means higher resolution



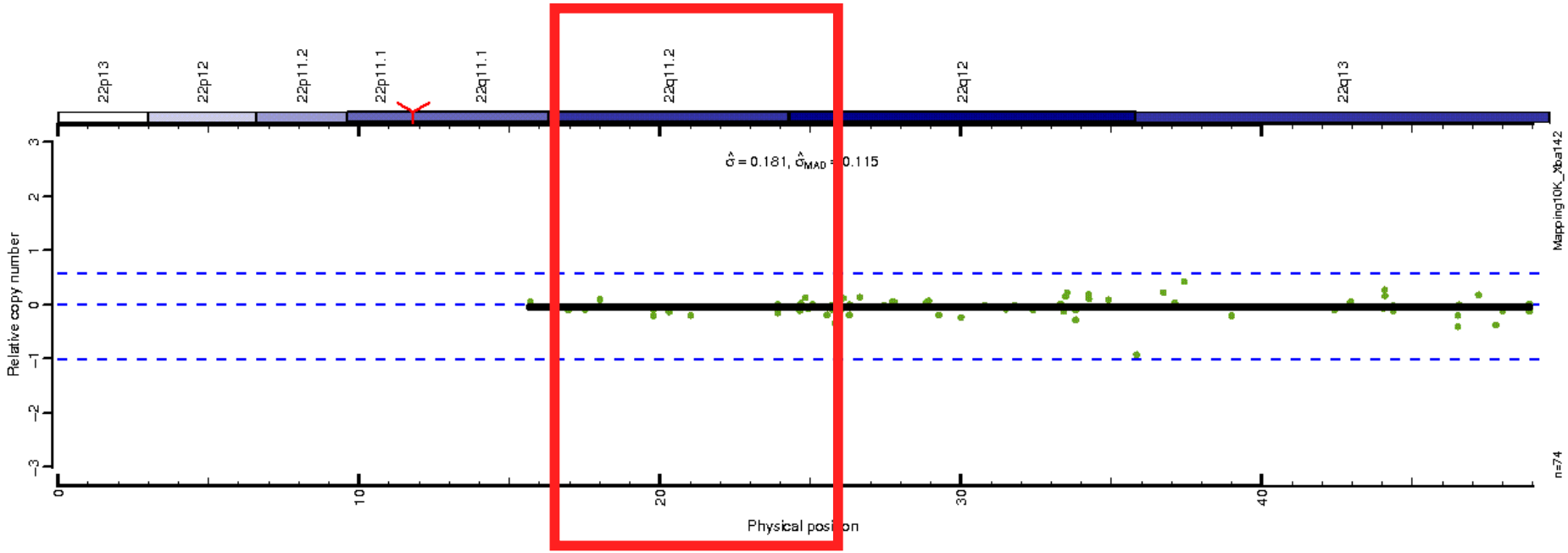
A brief history...

Genome-Wide Human SNP Array 6.0 *is the state-of-the-art array*

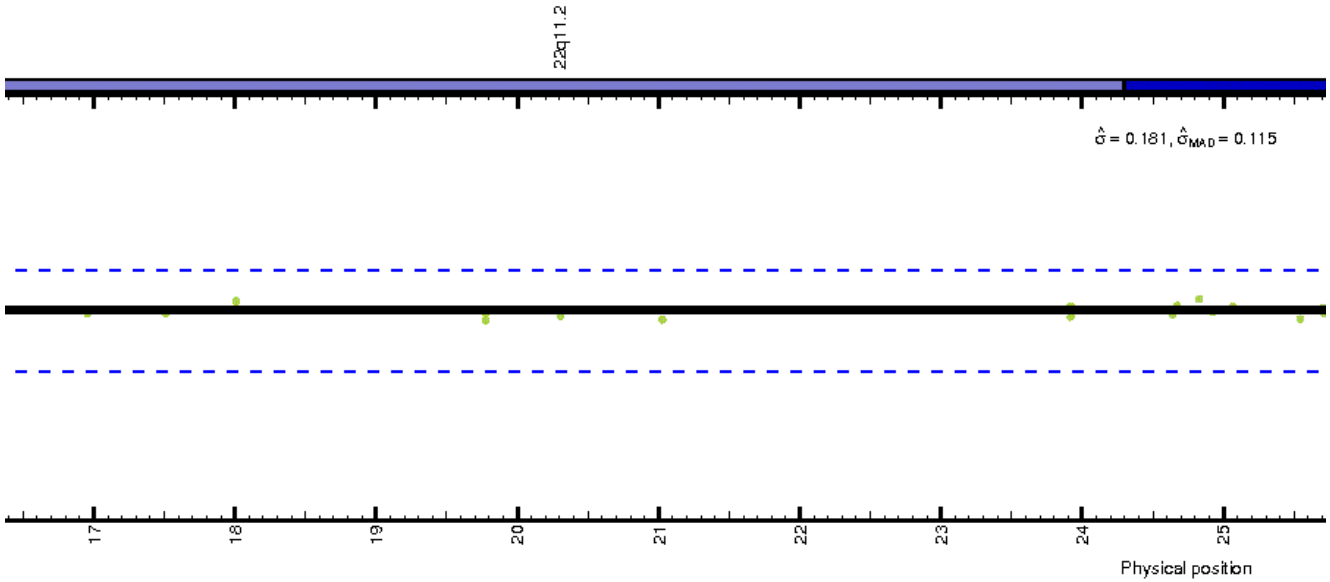
- **> 906,600 SNPs:**
  - Unbiased selection of 482,000 SNPs: historical SNPs from the SNP Array 5.0 (== 500K)
  - Selection of additional 424,000 SNPs:
    - Tag SNPs
    - SNPs from chromosomes X and Y
    - Mitochondrial SNPs
    - Recent SNPs added to the dbSNP database
    - SNPs in recombination hotspots
- **> 946,000 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

How did we get here?

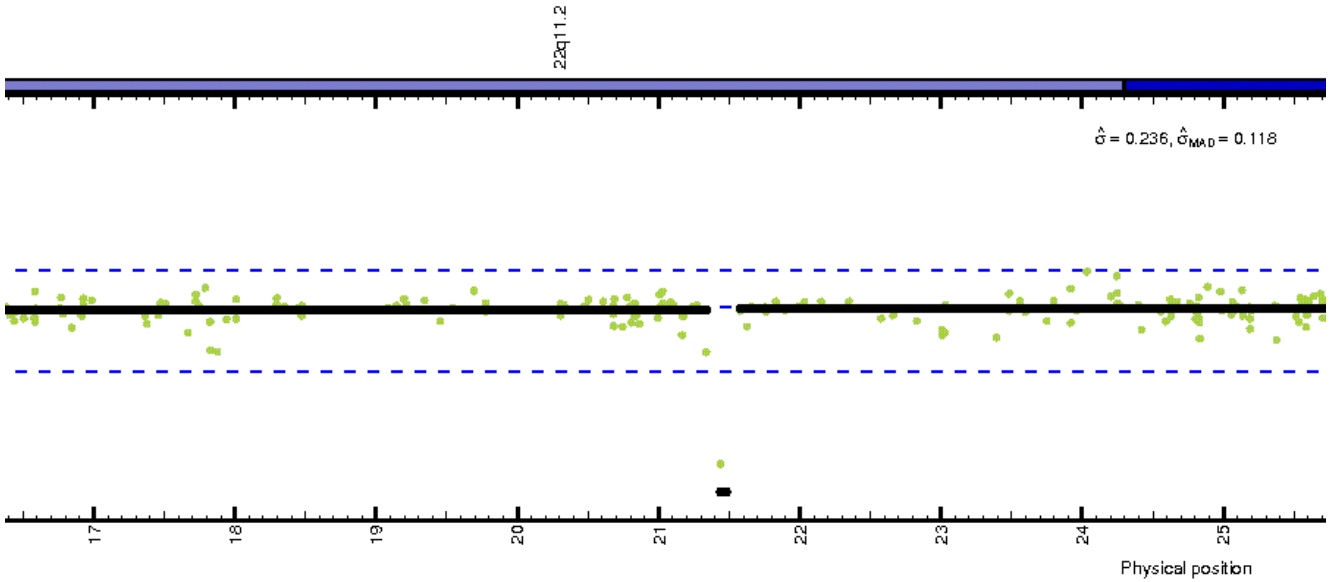
Data from 2003 on Chr22 (on of the smaller chromosomes)



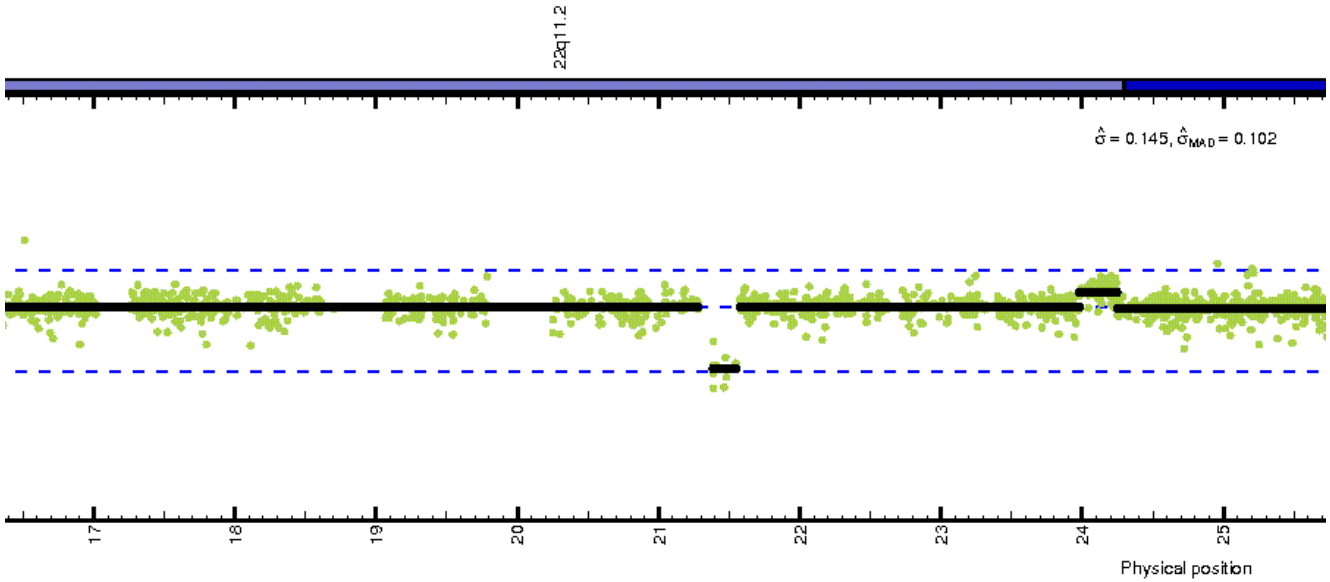
2003: 10,000 loci x1



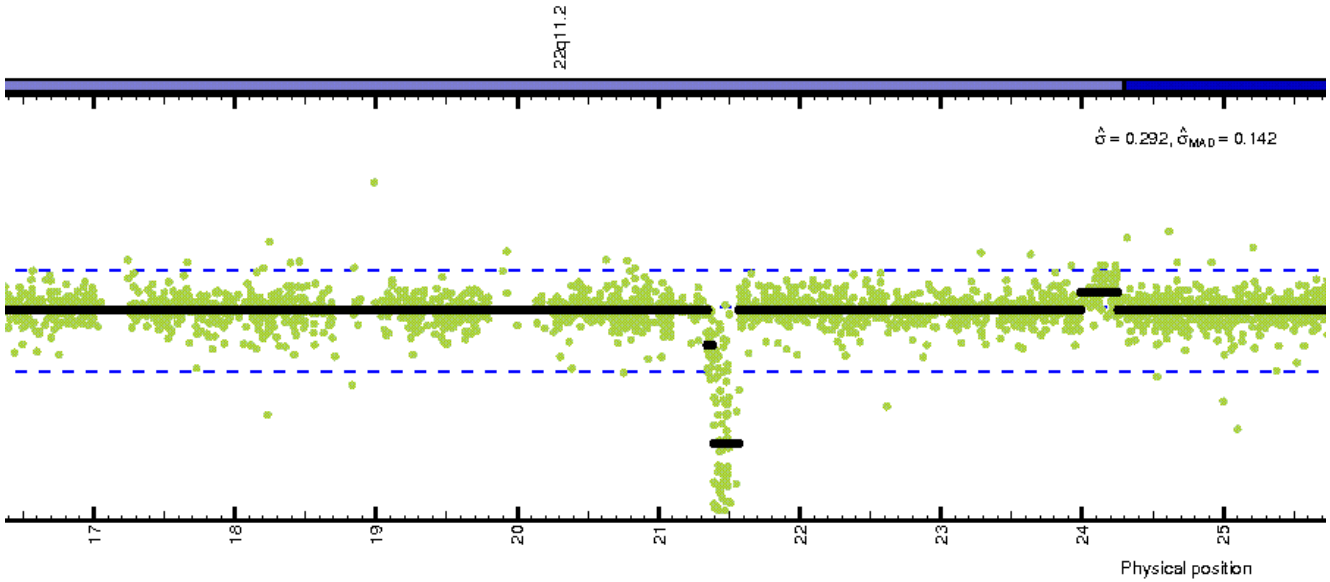
2004: 100,000 loci x10



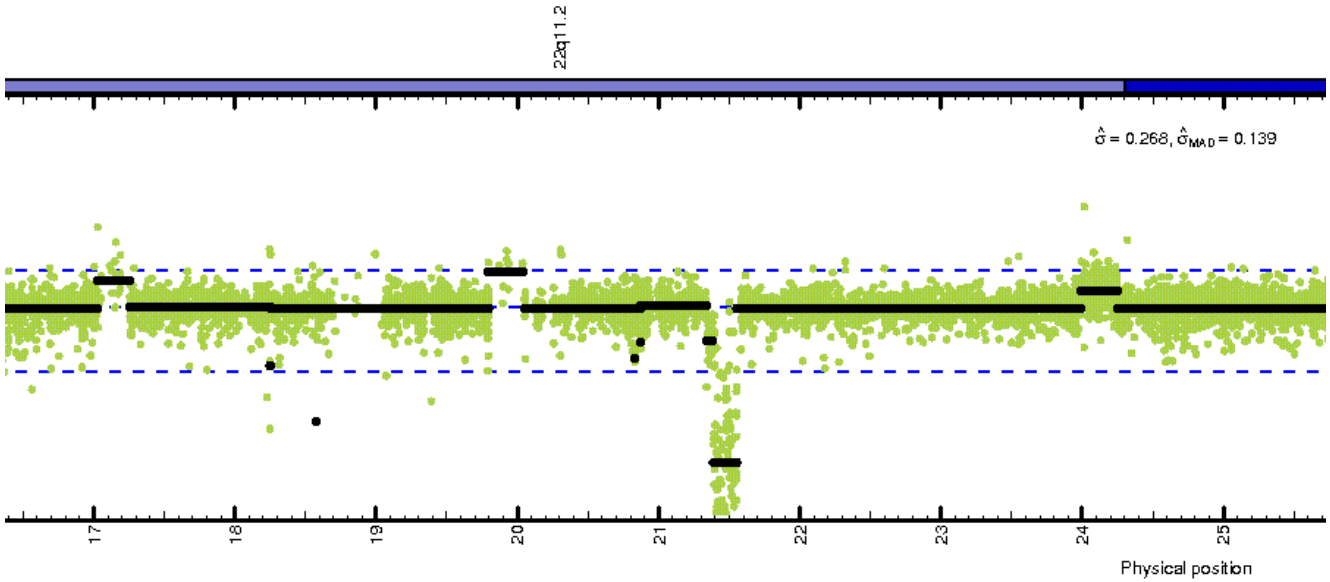
2005: 500,000 loci x50



2006: 900,000 loci x90



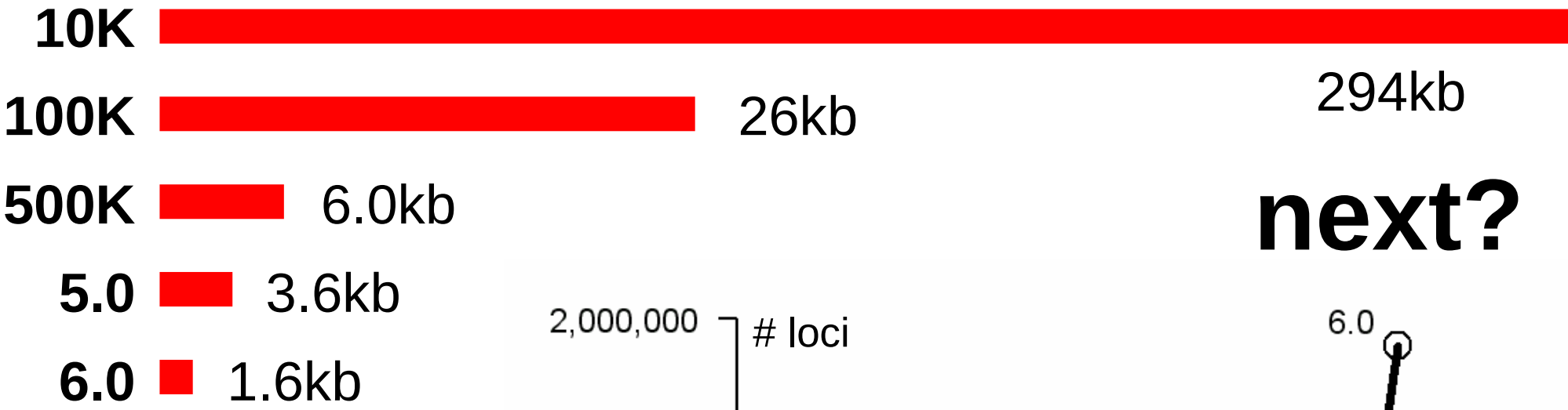
2007: 1,800,000 loci x180



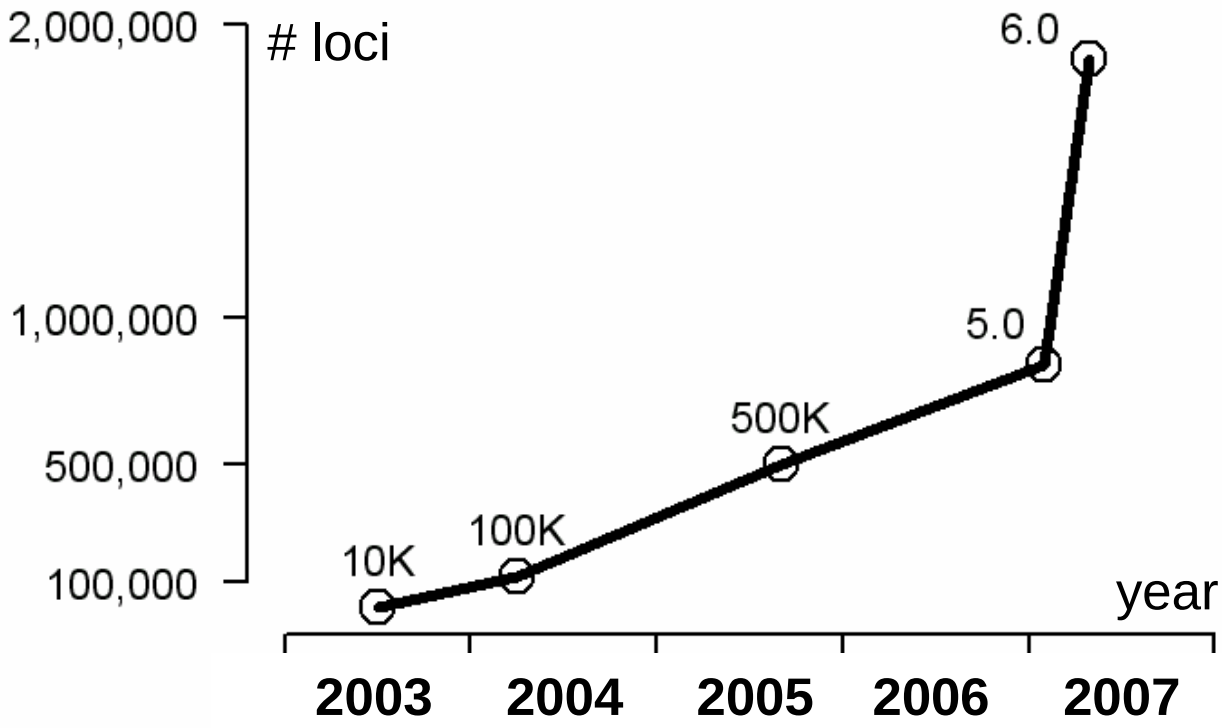
Rapid increase in density

Distance between loci:

4× further out…



next?



Affymetrix & Illumina are competing
- we get more bang for the buck (cup)

	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**	**260 USD**	**175 USD**	**300 USD**
# loci / cup of espresso ($1.35)	**116 loci**	**216 loci**	**1426 loci**	**3561 loci**	**4638 loci**

Price source: Affymetrix Pricing Information, http://www.affymetrix.com/, January 2008.

Preprocessing for copy-number analysis

Copy-number estimation using
Robust **M**ultichip **A**nalysis (CRMA)

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*

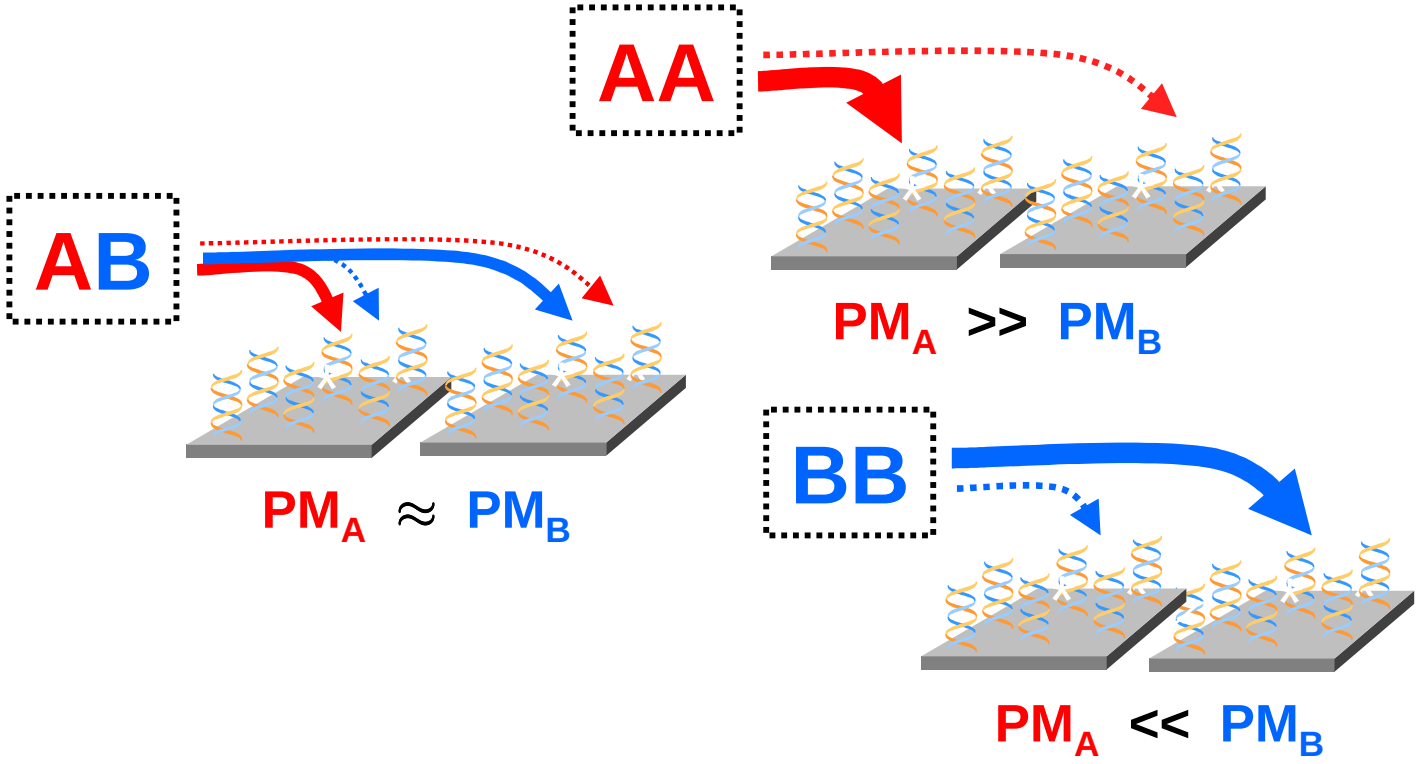
Crosstalk between alleles adds significant artifacts to signals

	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:

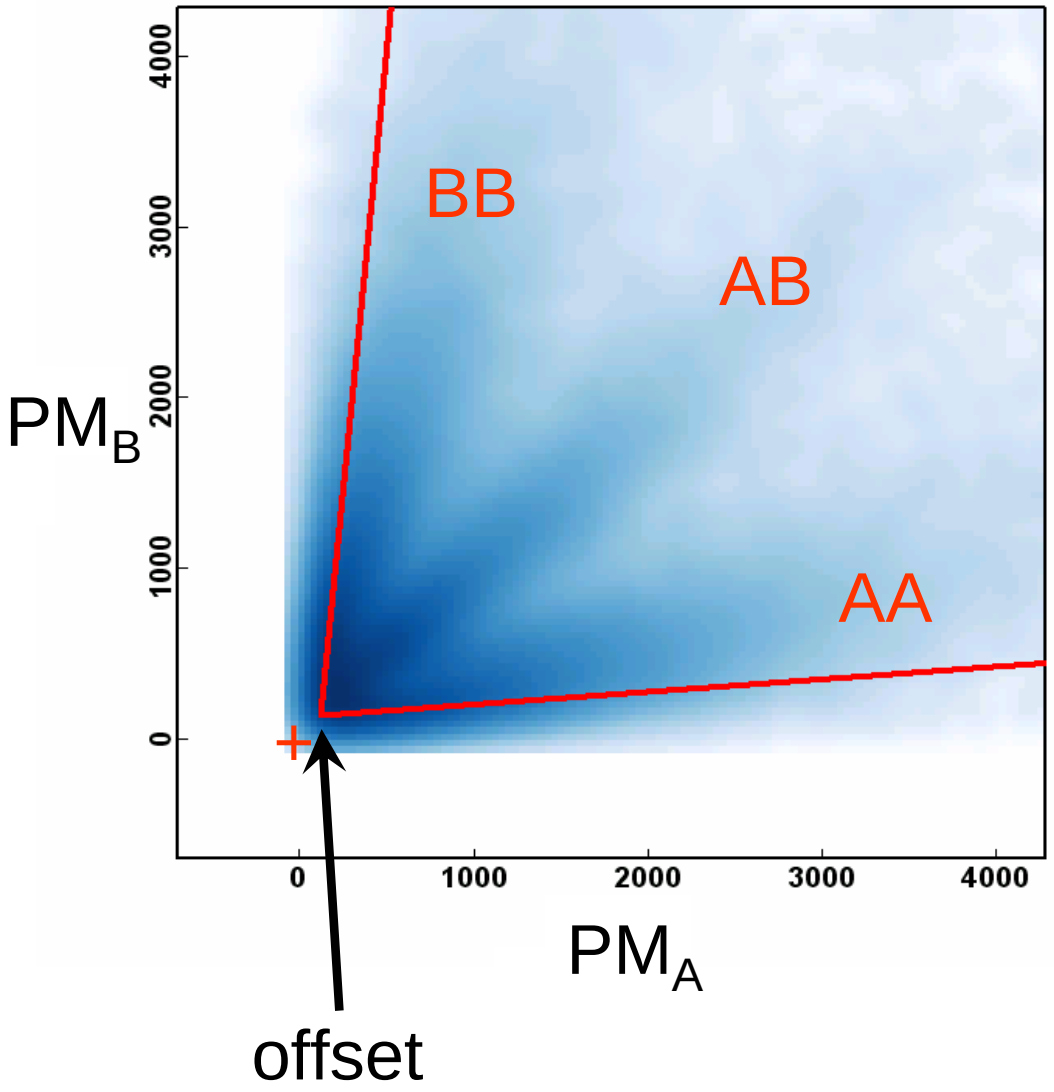
Allele A: ***TCGGTAAGTACTC***

Allele B: ***TCGGTATGTACTC***



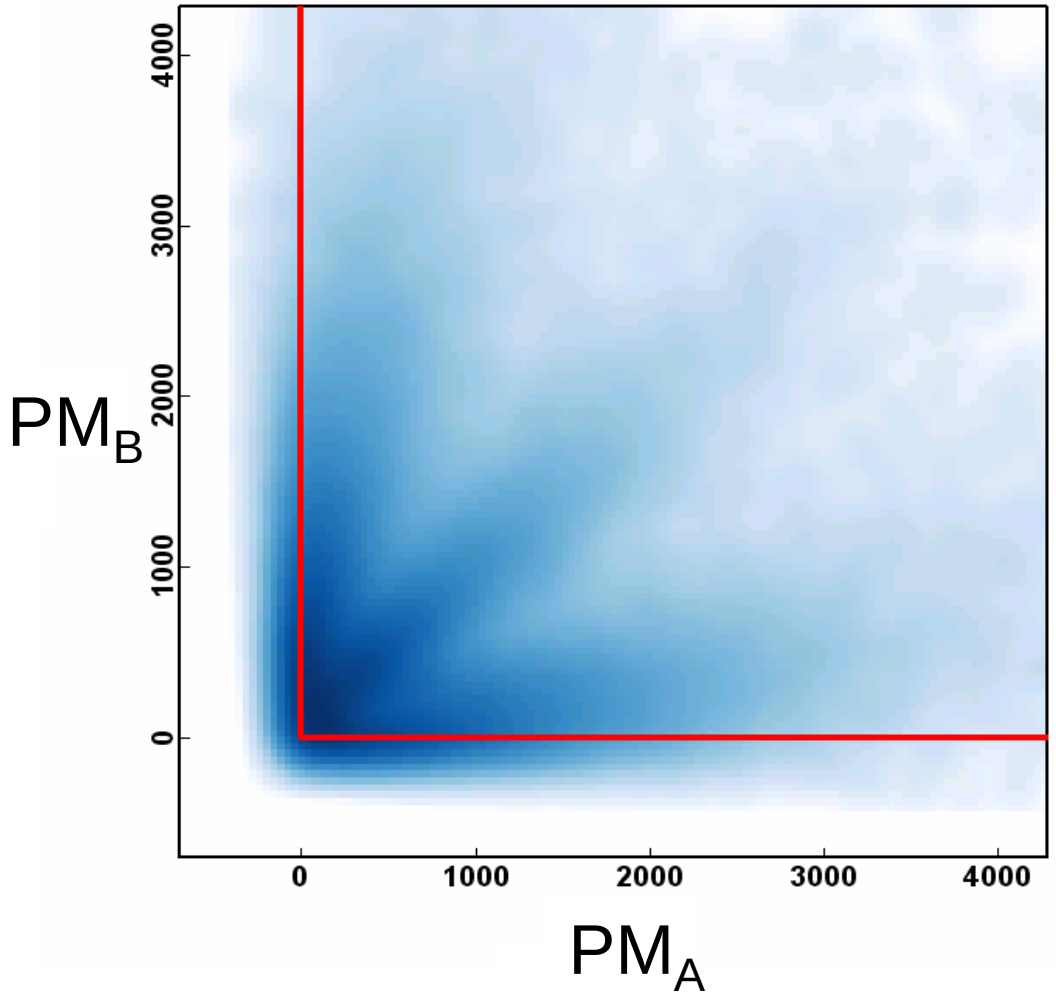
Crosstalk between alleles is easy to spot

	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 between alleles can be estimated and corrected for

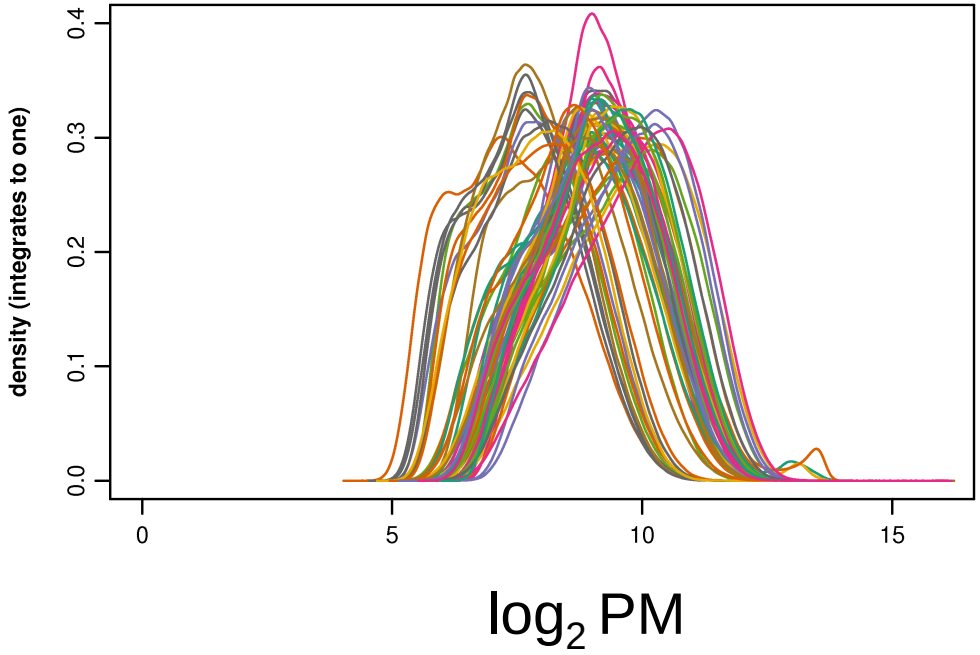
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$



Before removing crosstalk the arrays differ significantly...

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

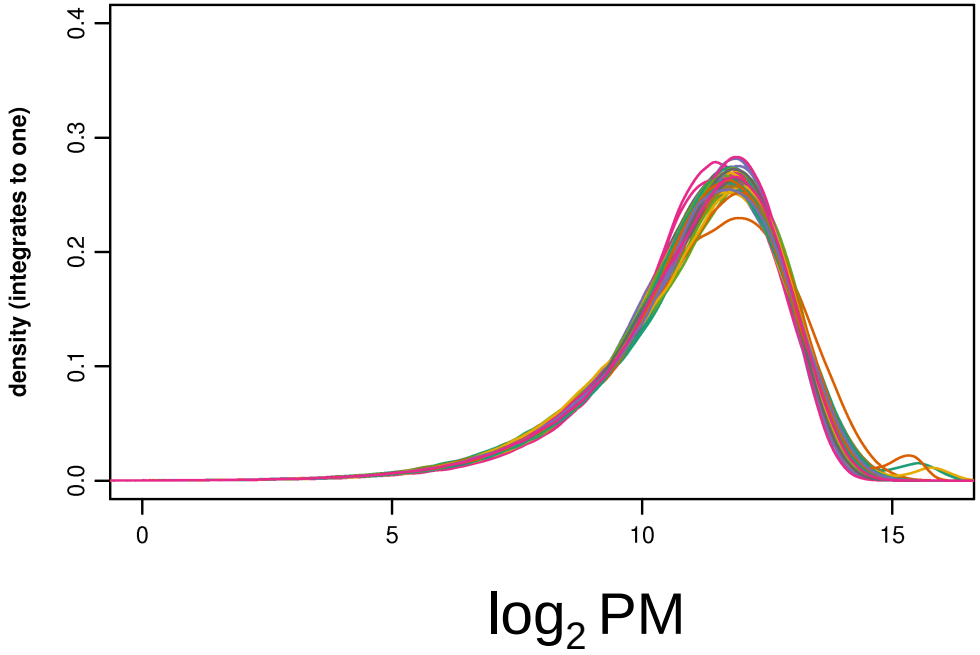
Crosstalk calibration corrects for differences in distributions too



When removing crosstalk system differences between arrays goes away

	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

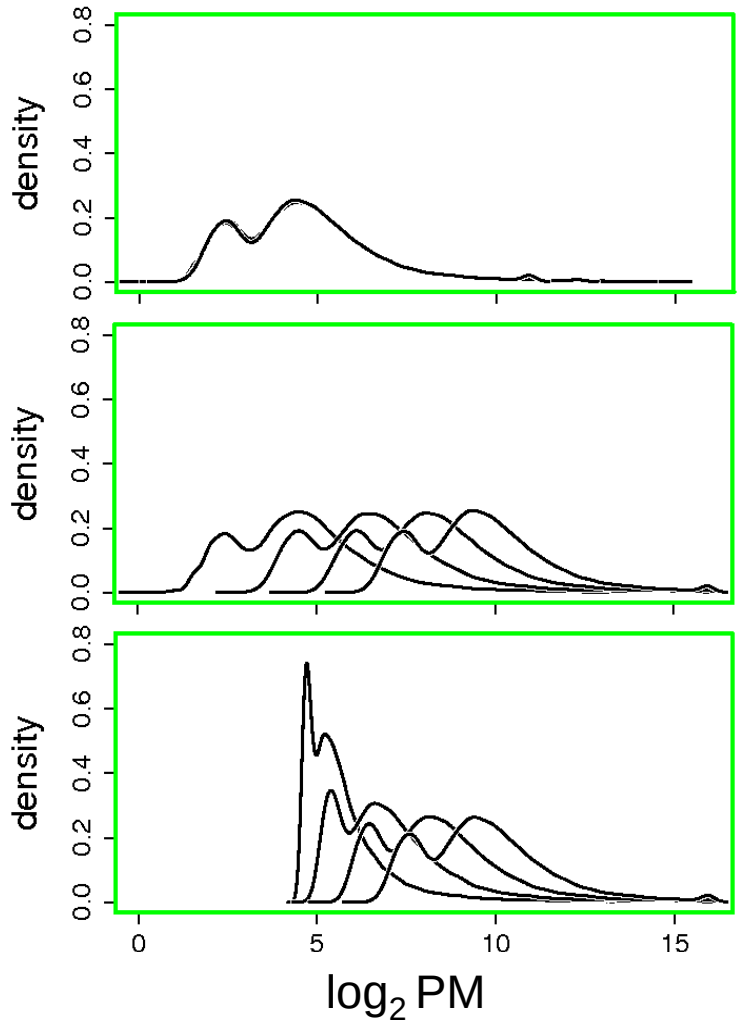


How can a translation and a rescaling make such a big difference?

Four measurements of the **same thing**:

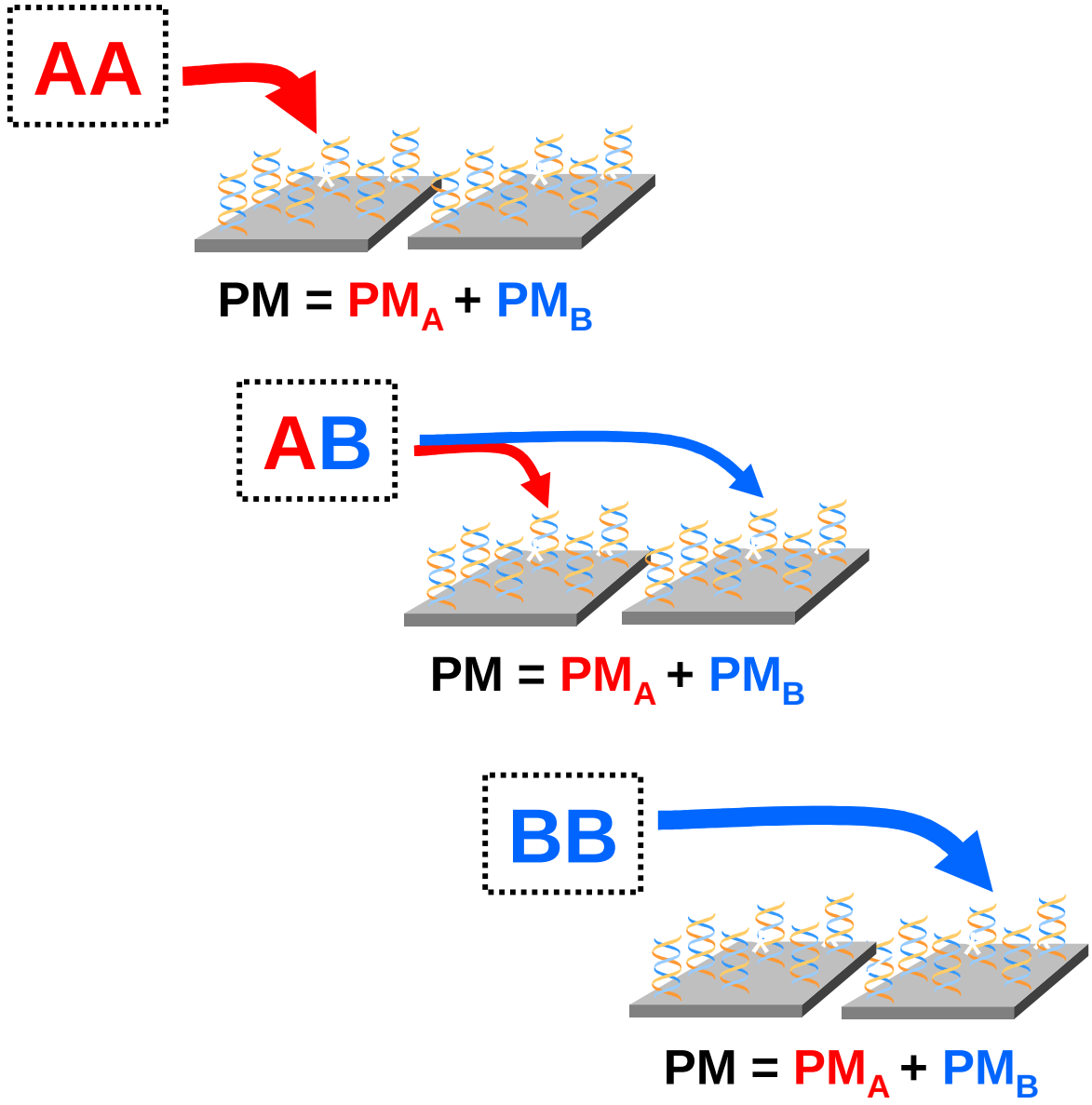
With **different scales**:
*log(b*PM) = log(b)+log(PM)*

With **different scales** and **some offset**:
*log(a+b*PM) = ...*

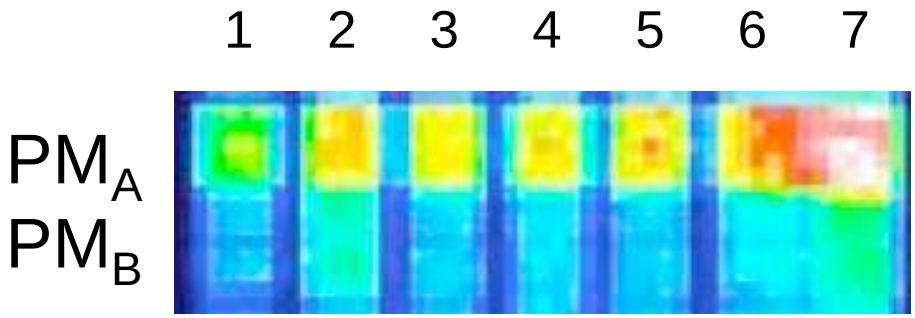


Copy-number estimation using Robust Multichip Analysis (CRMA)

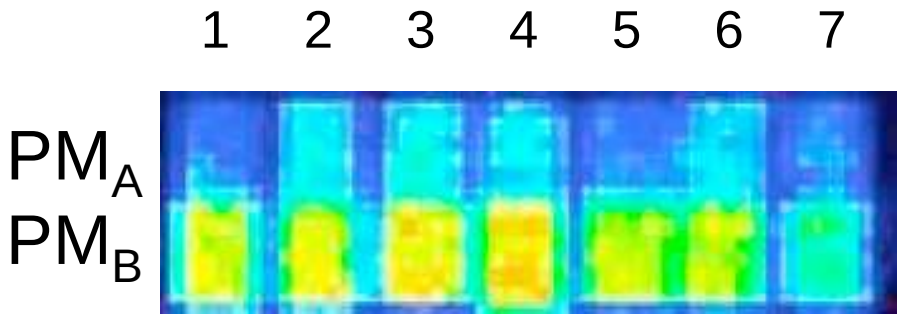
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$



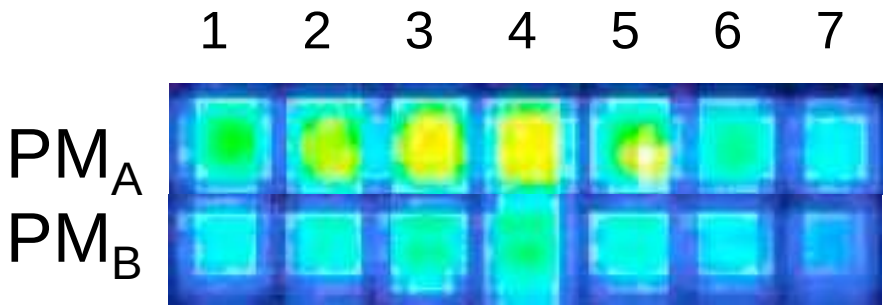
For robustness (against outliers), there are multiple probes per SNP



Genotype AA



Genotype BB



Genotype AB

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)

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

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

where ξ_{ik} is noise.

Probe-level summarization
- the log-additive model

For one SNP, the model is:

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

Take the logarithm on both sides:

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

Sample $i=1,2,\ldots,I$, and probe $k=1,2,\ldots,K$.

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:

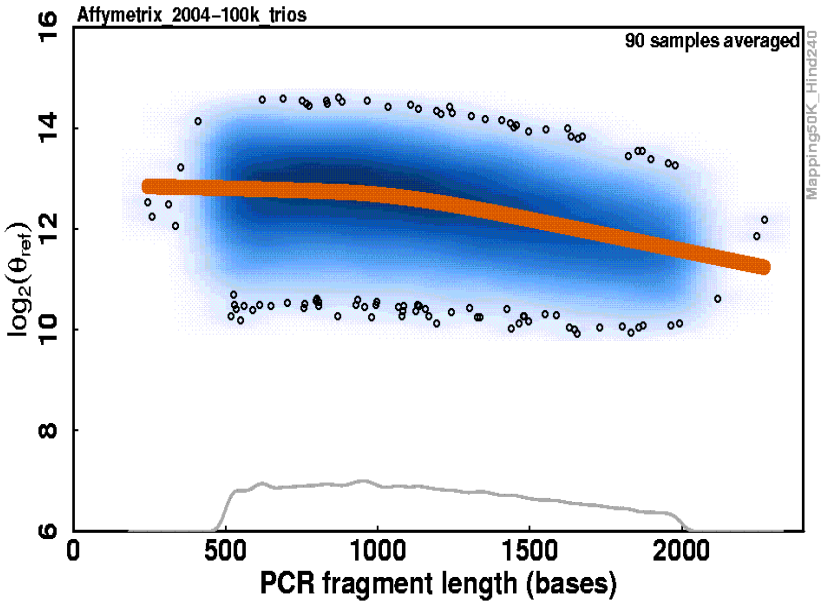
$$\log_2(PM_{ik}) = \log_2\phi_k + \log_2\theta_i + \varepsilon_{ik}$$

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

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

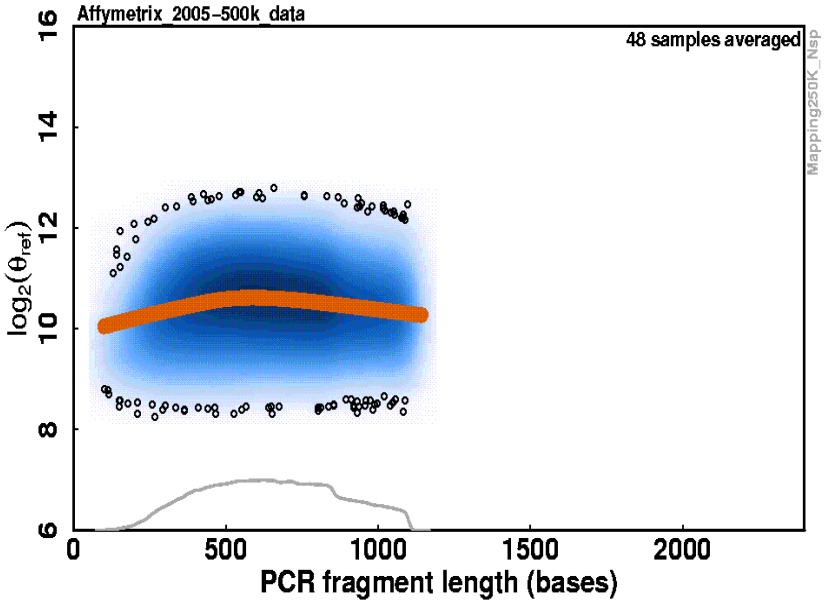


100K

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

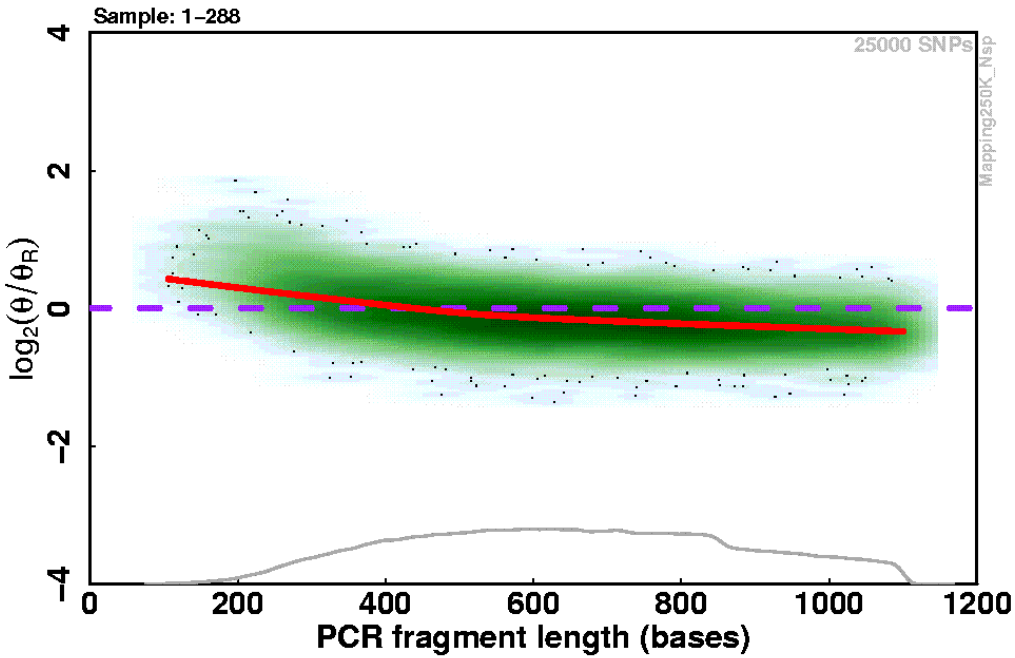


500K

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

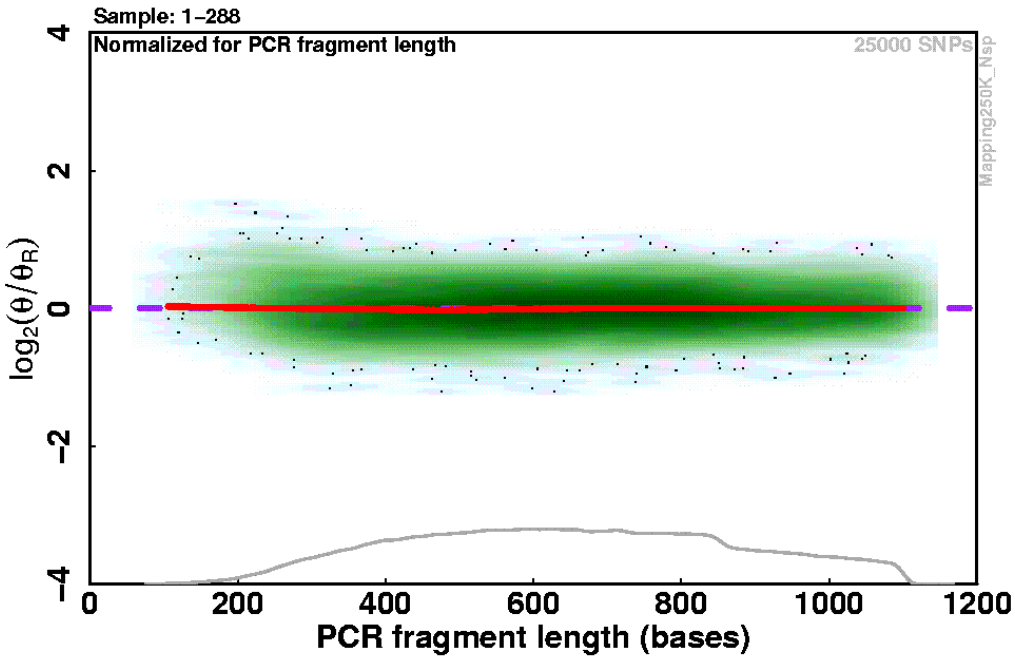
Normalize to get same fragment-length effect for all hybridizations



Copy-number estimation using Robust Multichip Analysis (CRMA)

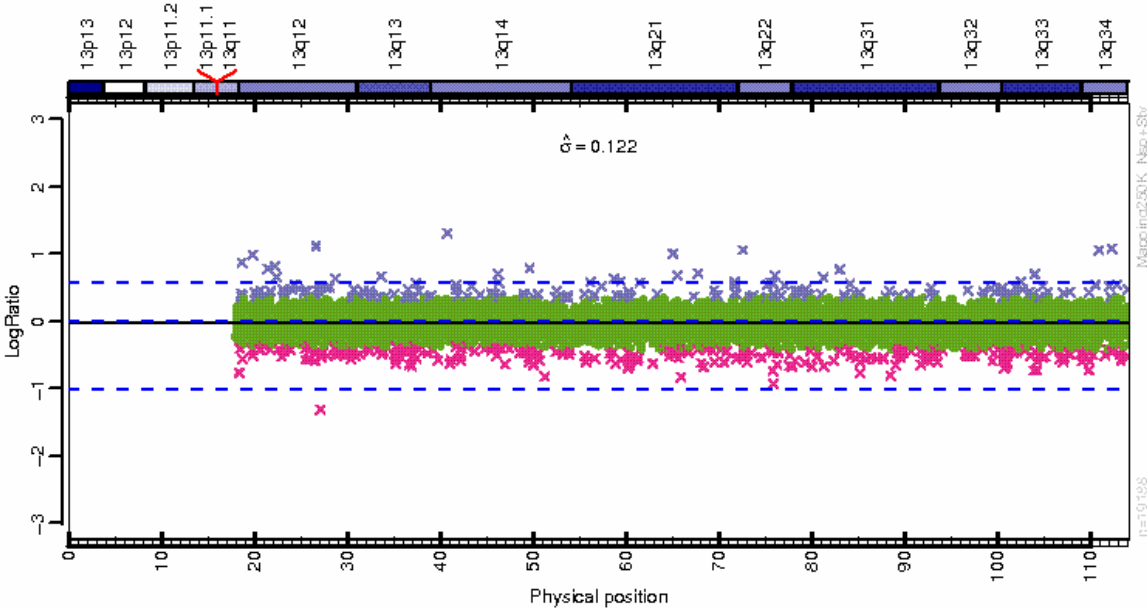
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

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



Results

(comparing with other methods)

Other methods

	CRMA	***dChip*** **(Li & Wong 2001)**	***CNAG*** **(Nannya et al 2005)**	***CNAT v4*** **(Affymetrix 2006)**
Preprocessing *(probe signals)*	allelic crosstalk (quantile)	invariant-set	scale	quantile
Total CNs	$PM=PM_A+PM_B$	$PM=PM_A+PM_B$ $MM=MM_A+MM_B$	$PM=PM_A+PM_B$	$\theta=\theta_A+\theta_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)	multiplicative (PM-MM)	sum (PM-only)	log-additive (PM-only)
Post-processing	fragment-length (GC-content)	-	fragment-length GC-content	fragment-length GC-content
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

How well can be differentiate between one and two copies?

HapMap (CEU):

Mapping250K Nsp data (one half of the "500K")

30 males and 29 females (no children; one excl. female)

Chromosome X is known:

Males (CN=1) & females (CN=2)

5,608 SNPs

Classification rule:

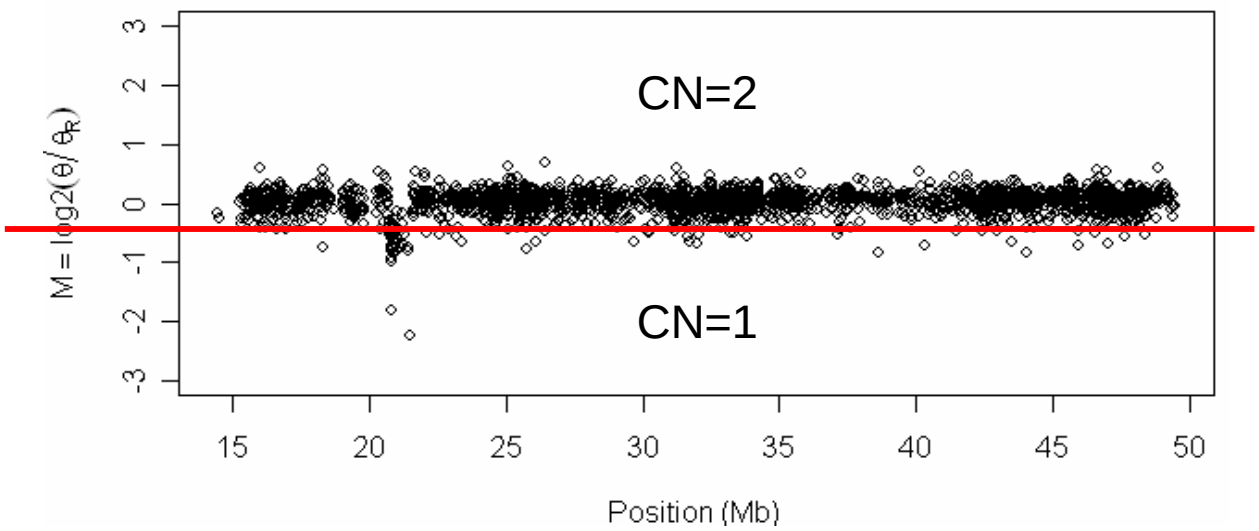
M_{ij} < threshold $\Rightarrow$ CN_{ij} =1, otherwise CN_{ij} =2.

Number of calls: 59$\times$5,608 = 330,872

Classification rule for loci on X
- use raw CNs to call CN=1 or CN=2

Classification rule:

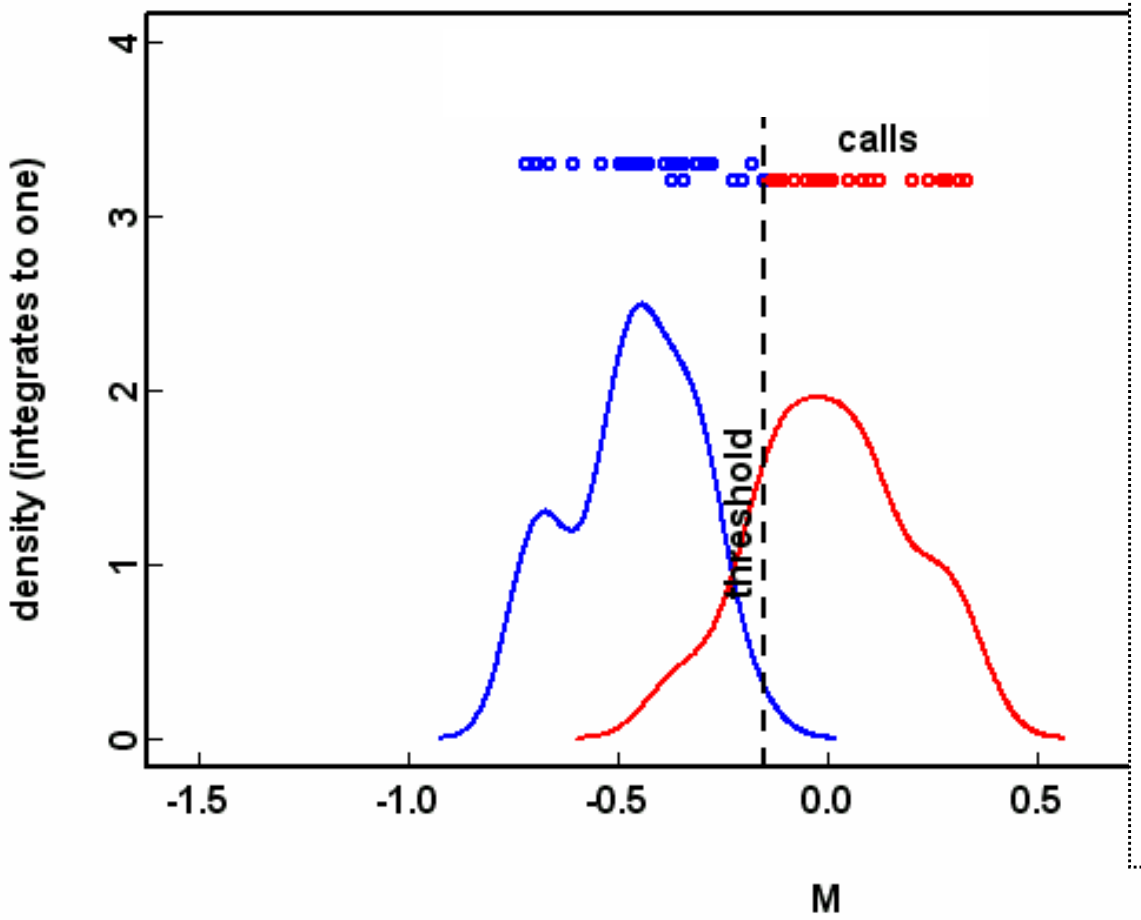
M_{ij} < threshold $\Rightarrow$
CN_{ij}=1, else CN_{ij}=2.



Number of calls per locus (SNP): 59 (one per samples)

Across Chromosome X: $59 \times 5{,}608$ loci = 330,872

Calling samples for SNP_A-1920774



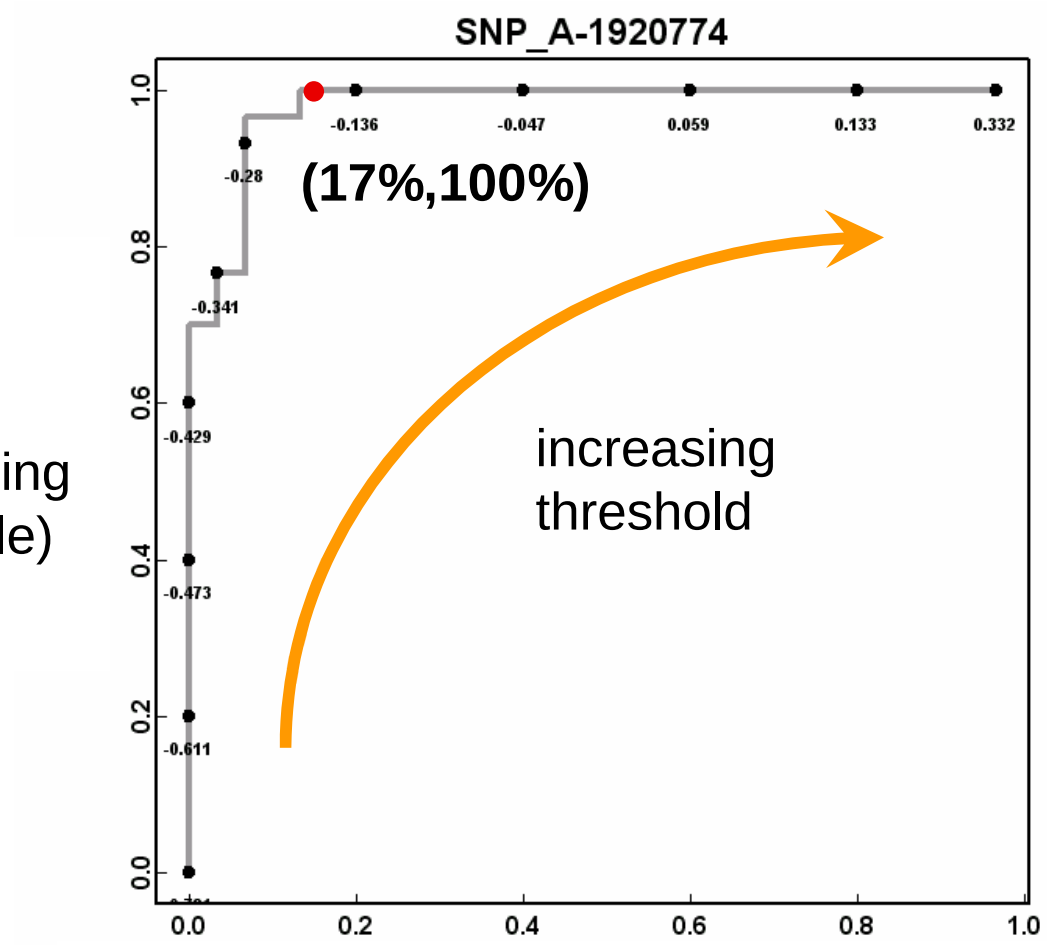
males: 30
females: 29

Call rule:
If M_i < threshold, a **male**

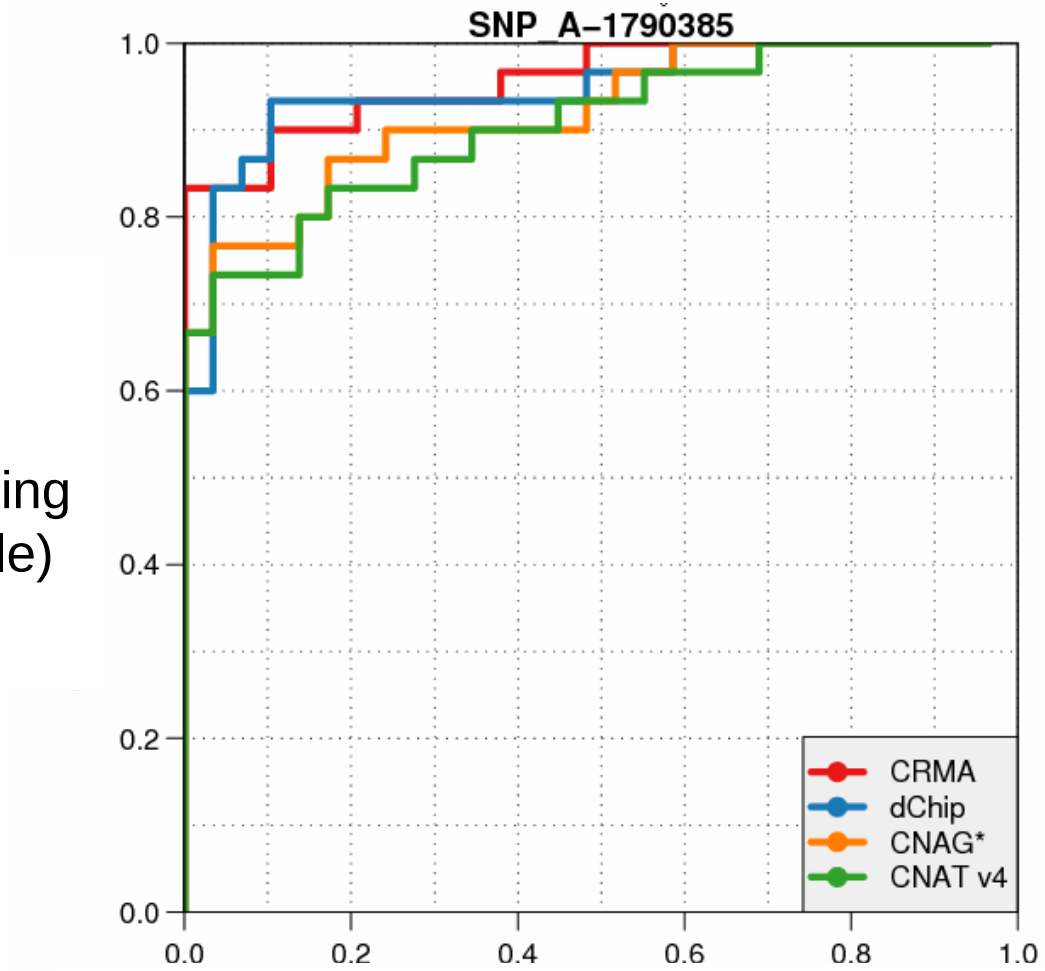
Calling a male male:
#True-positives: 30
TP rate: 30/30 = 100%

Calling a female male:
#False-positive : 5
FP rate: 5/29 = 17%

Receiver Operator Characteristic (ROC)



Single-SNP comparison
A random SNP



TP rate

(correctly calling a males male)

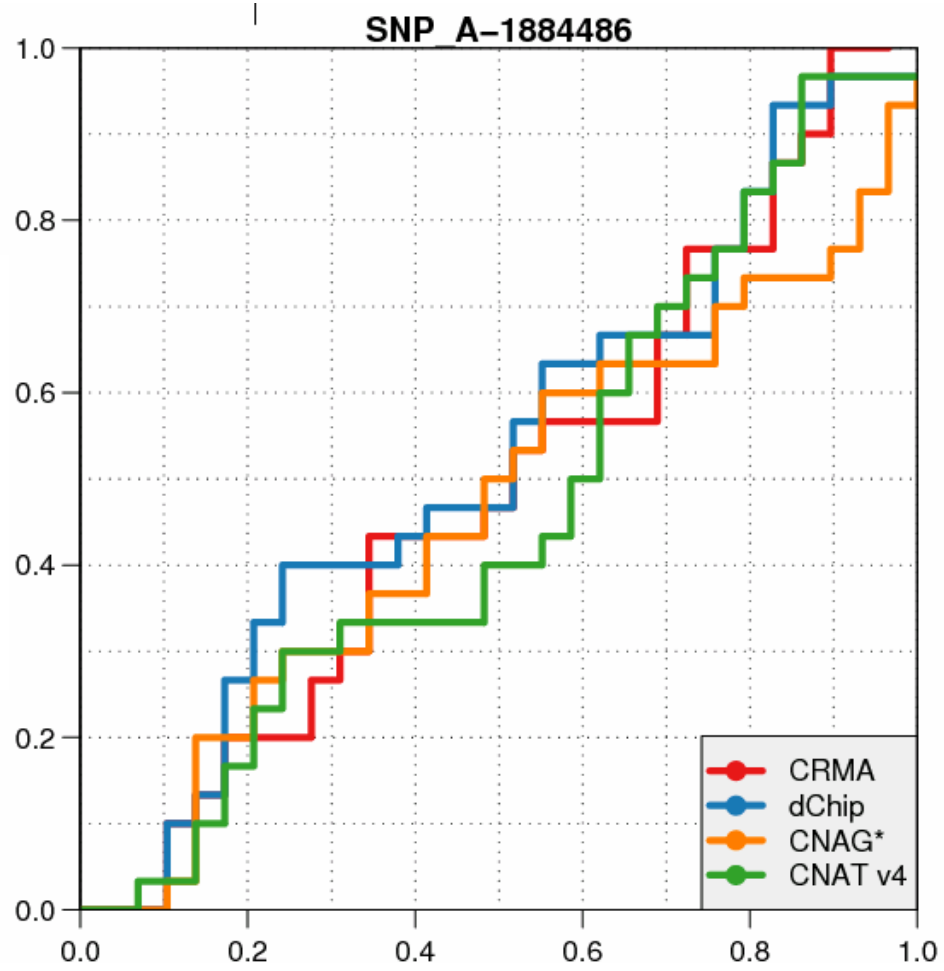
FP rate

(incorrectly calling females male)

Single-SNP comparison
A non-differentiating SNP

TP rate

(correctly calling a males male)

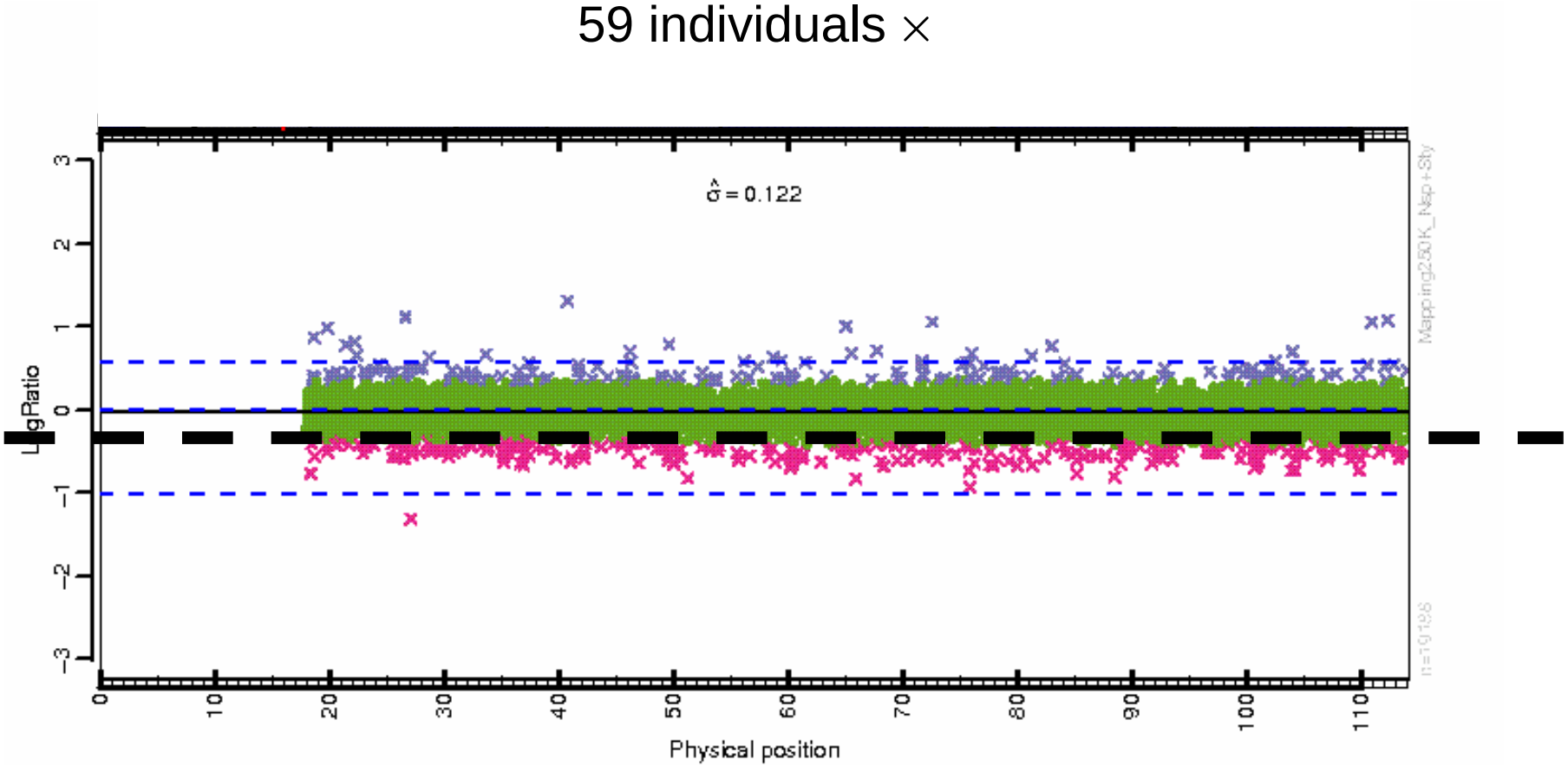


FP rate

(incorrectly calling females male)

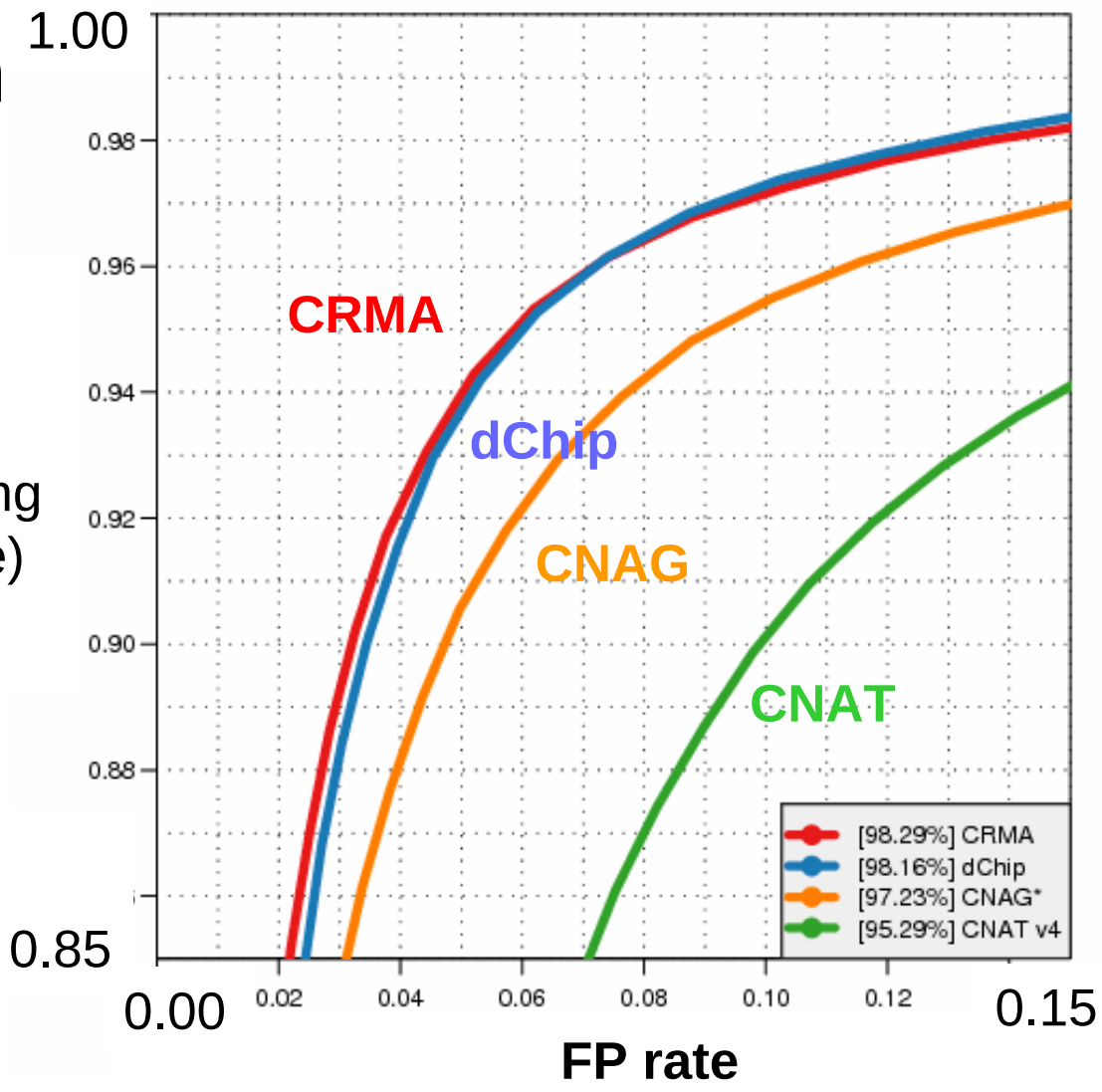
Performance of an average SNP with a common threshold

59 individuals ×



CRMA & dChip perform better for an average SNP *(common threshold)*

Zoom in



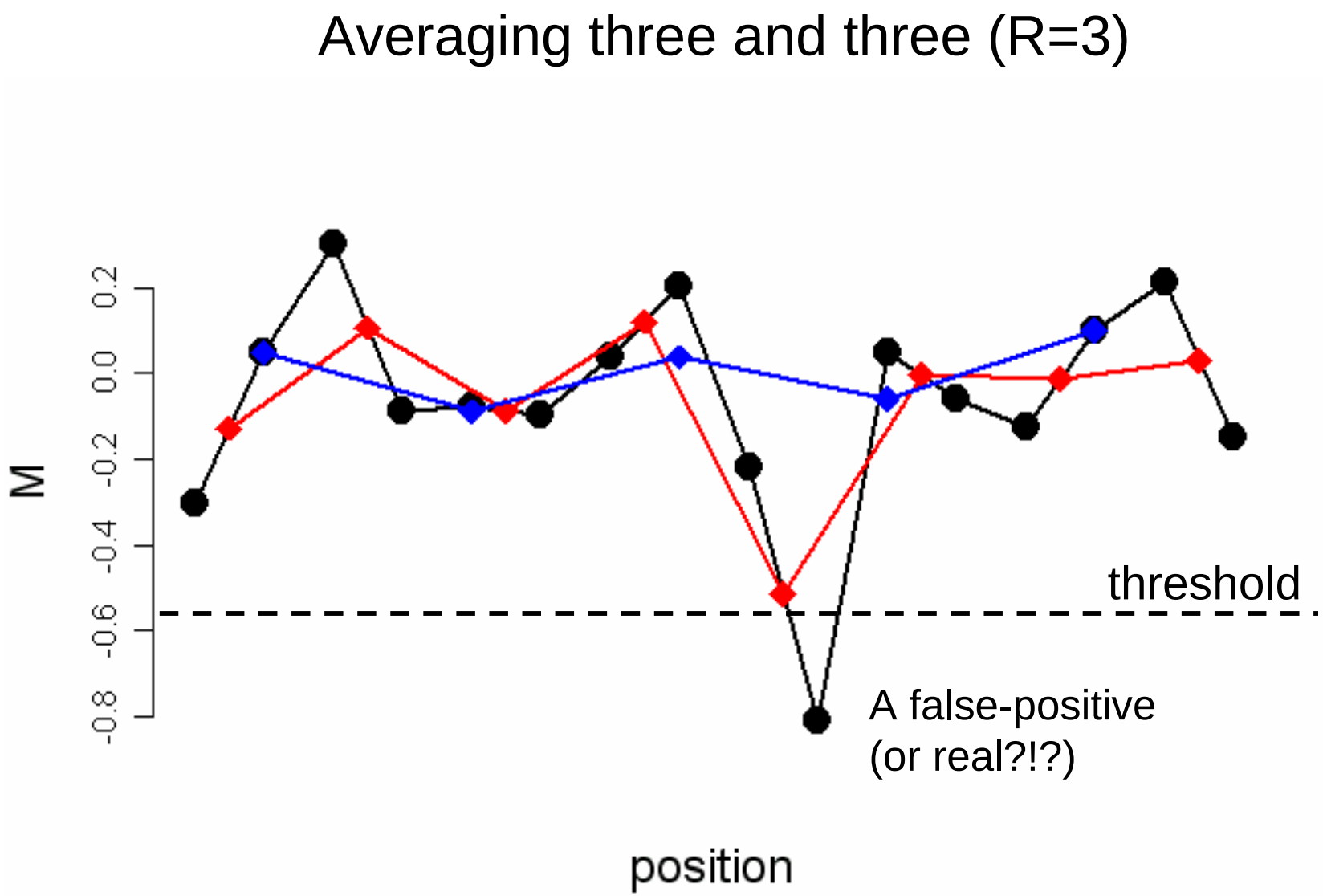
TP rate

(correctly calling a males male)

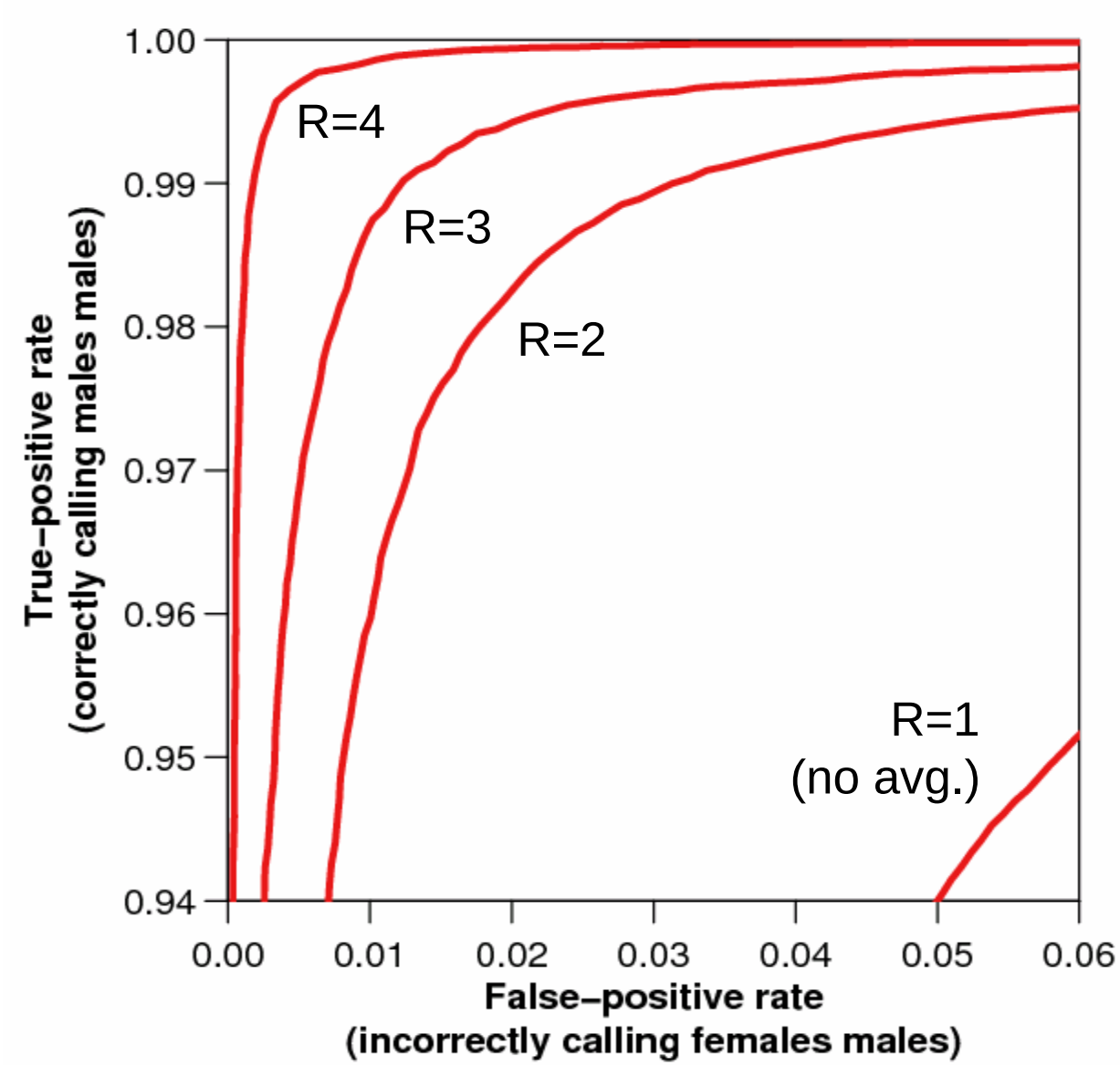
Number of calls:
59×5,608 = 330,872

"Smoothing"

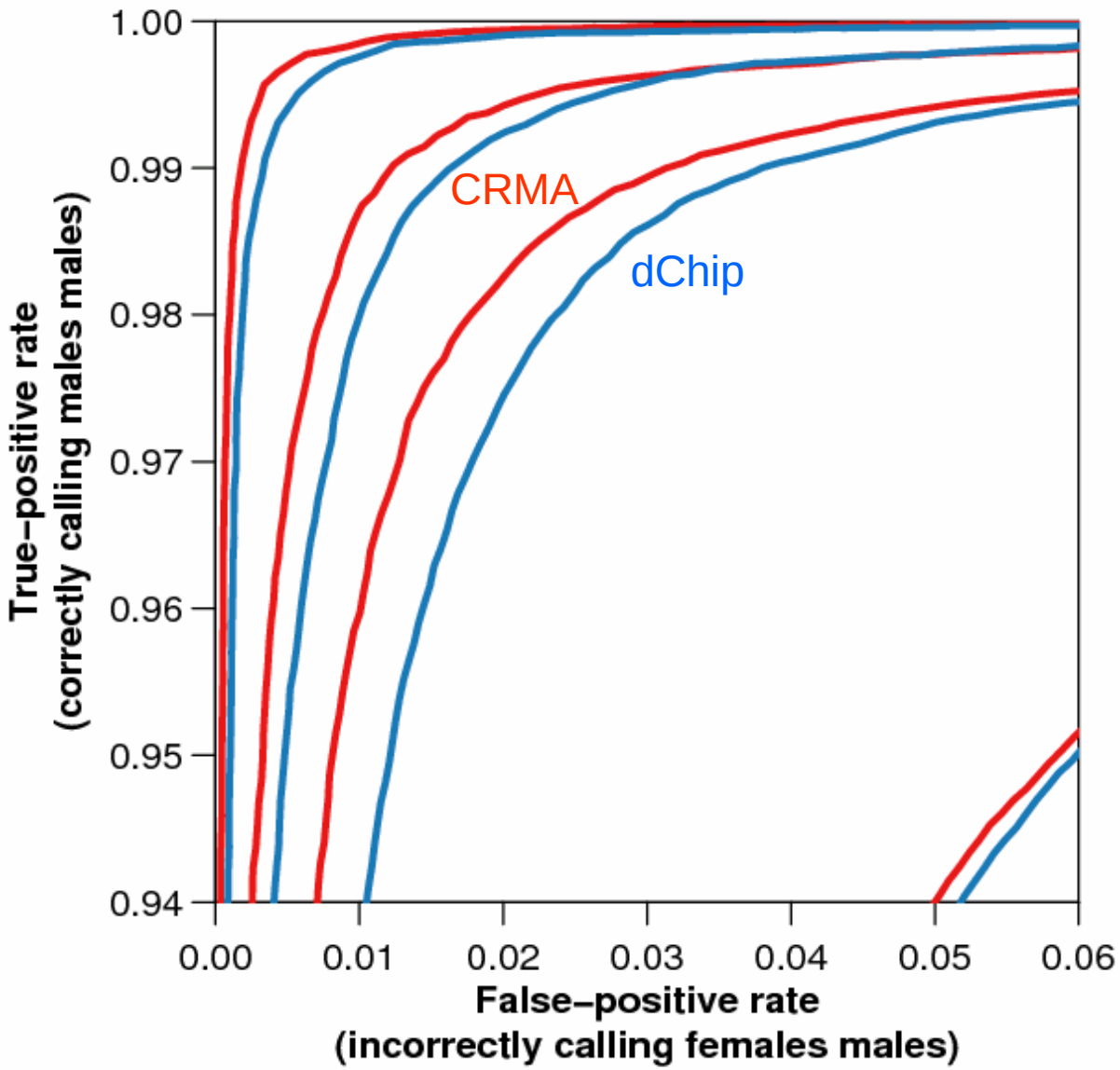
Average across SNPs
non-overlapping windows



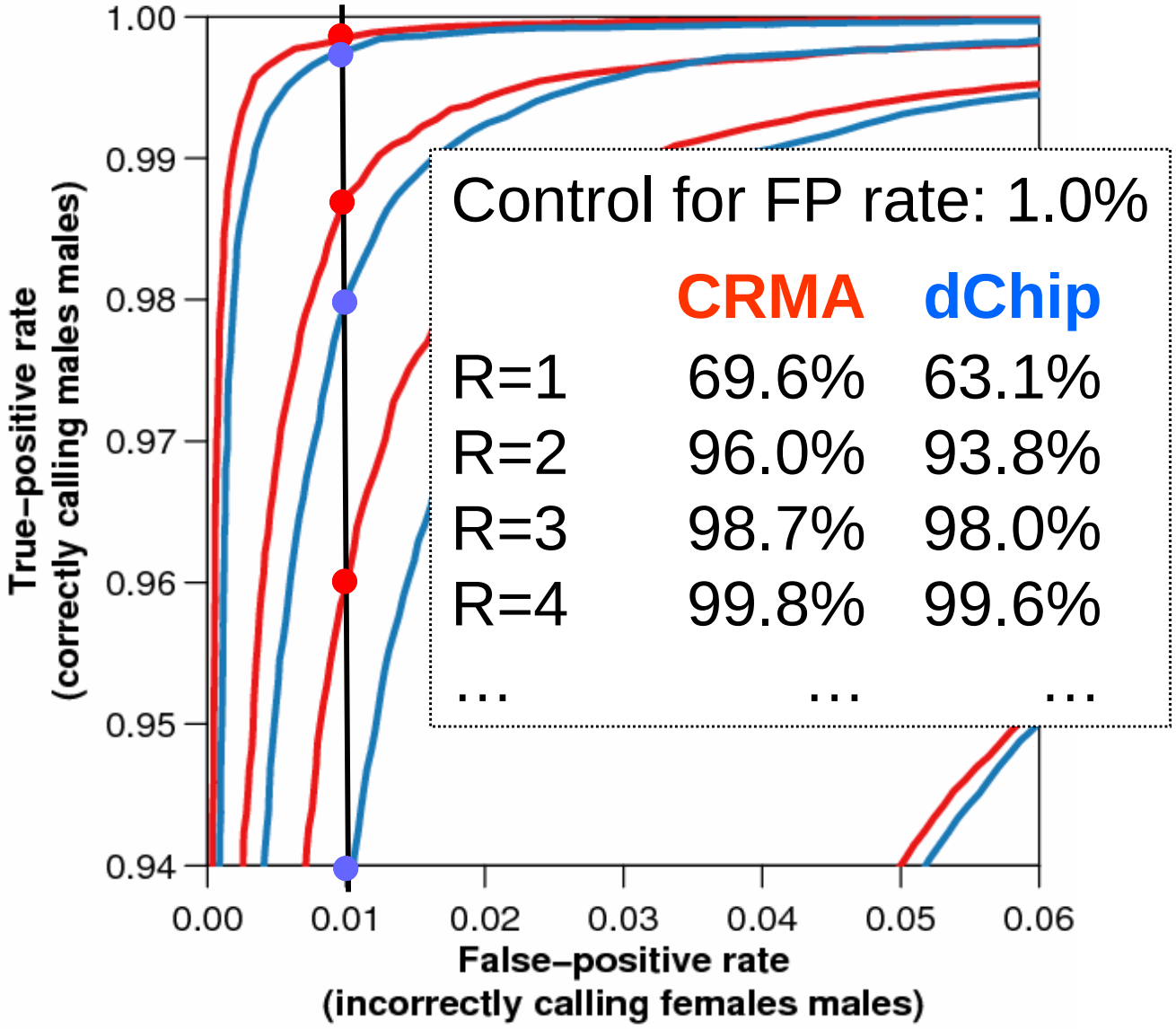
Better detection rate when averaging *(with risk of missing short regions)*



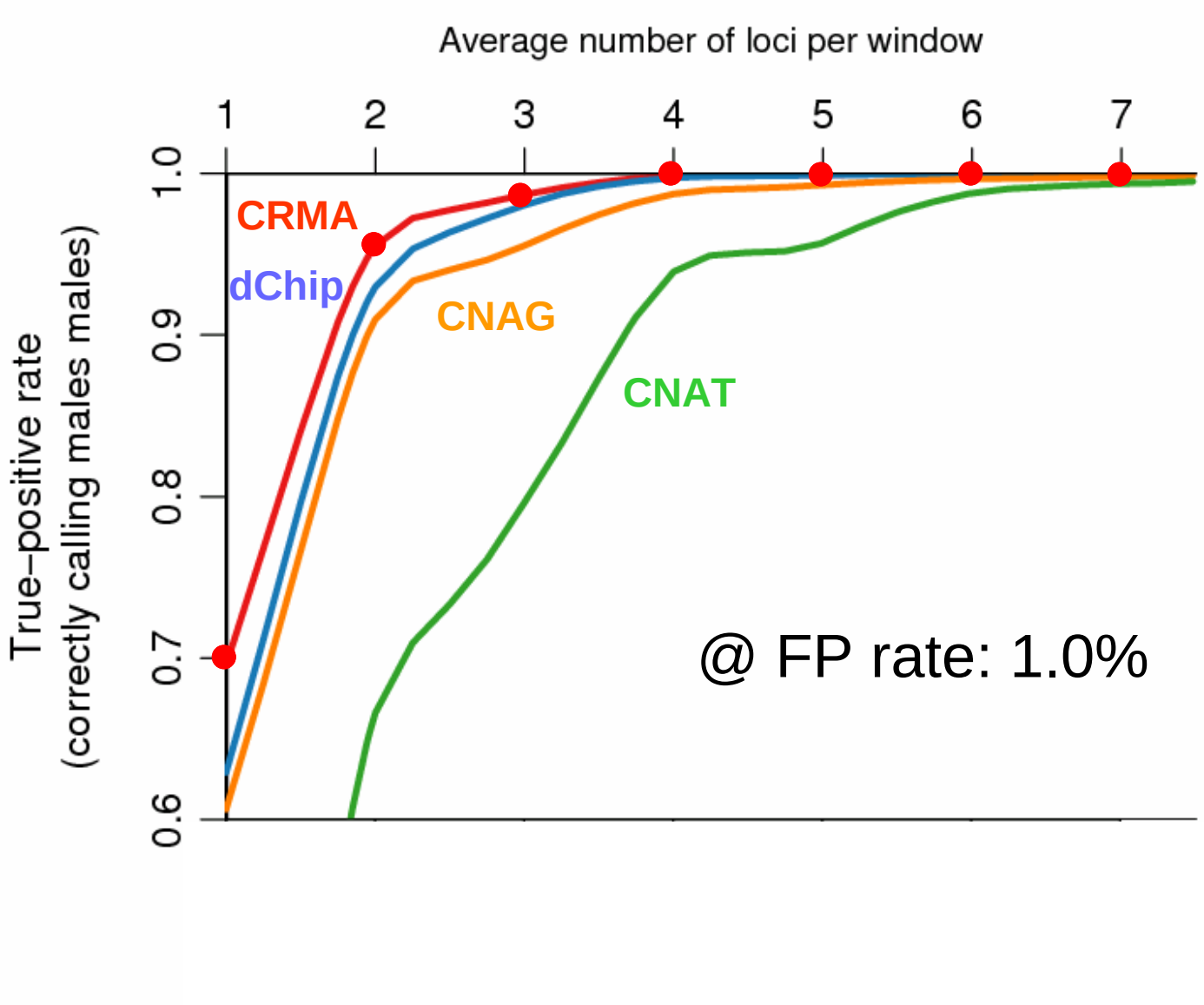
CRMA does better than dChip



CRMA does better than dChip



Comparing methods by “resolution” *controlling for FP rate*



Comparison across generations

(100K - 500K - 6.0)

We have HapMap data for several generations of platforms

HapMap (CEU):

30 males and 29 females (no children; one excl. female)

Chromosome X is known:

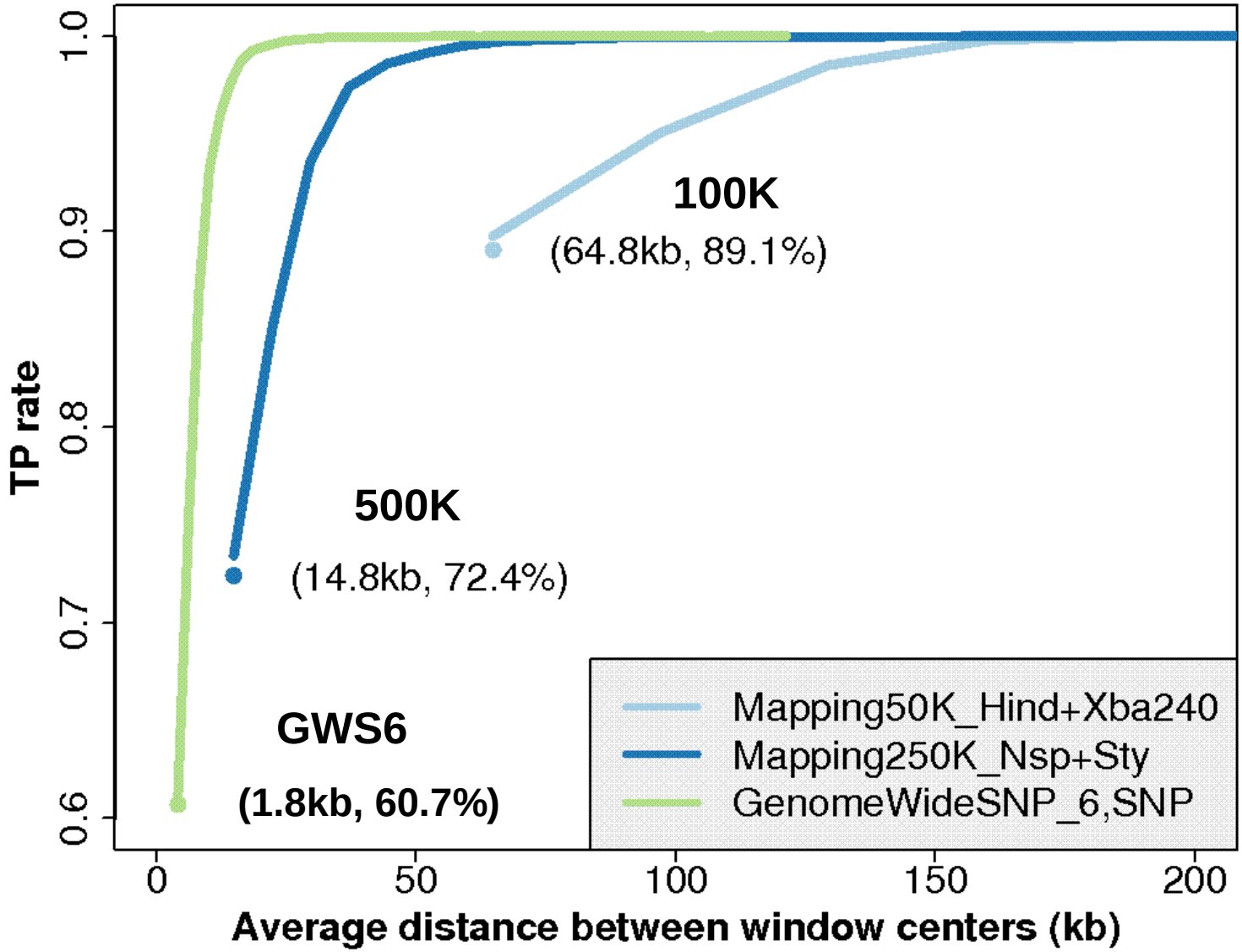
Males (CN=1) & females (CN=2)

5,608 SNPs

Platforms:

100K, 500K, 6.0.

Resolution comparison
- at 1.0% FP



Summary

Conclusions

- It helps to:
  - Control for allelic crosstalk.
  - Sum alleles at PM level: $PM = PM_A + PM_B$.
  - Control for fragment-length effects.
- Resolution: 6.0 (SNPs) > 500K > 100K (or lab effects).
- Currently estimates from CN probes are poor. Not unexpected. Better preprocessing might help.

2008: >30,000,000 loci >x3000?

On January 10, 2008:

Dr Stephen Fodor, CEO of Affymetrix, outlined new products:

Affymetrix has been focusing on new chemistry techniques, such as a new higher yield synthesis technique.

The first product that will be launched - around the first half of 2008 - is **an ultra-high resolution copy number tool**.

*"This product will allow us to analyze the genome at **around 30 times the resolution** of the current state-of-the-art technology in the marketplace,"* claimed Fodor.

Source: http://www.labtechnologist.com/

Segmentation algorithms are the bottlenecks - *we need fast algorithms/implementation*

Some methods ***Need! (…or better)***

Chip type	# loci	n	O(n^2)	time / sample	O(n)	time / sample
250K	250,000	1×	1×	0.5h	1×	5.5min
500K	500,000	2×	4×	2h	2×	12min
5.0	1,000,000	4×	16×	**8h**	4×	27min
6.0	2,000,000	8×	64×	**32h**	8×	1.0h
?	32,000,000	128×	16,384×	**341 days!**	128×	12h