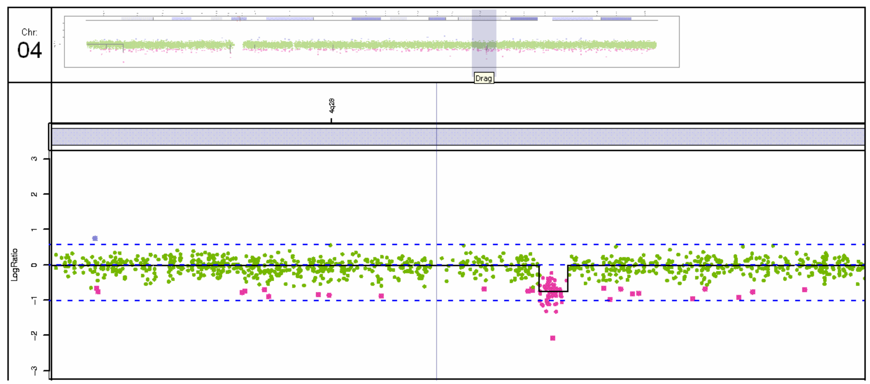
Estimating chromosomal copy numbers from Affymetrix SNP & CN chips



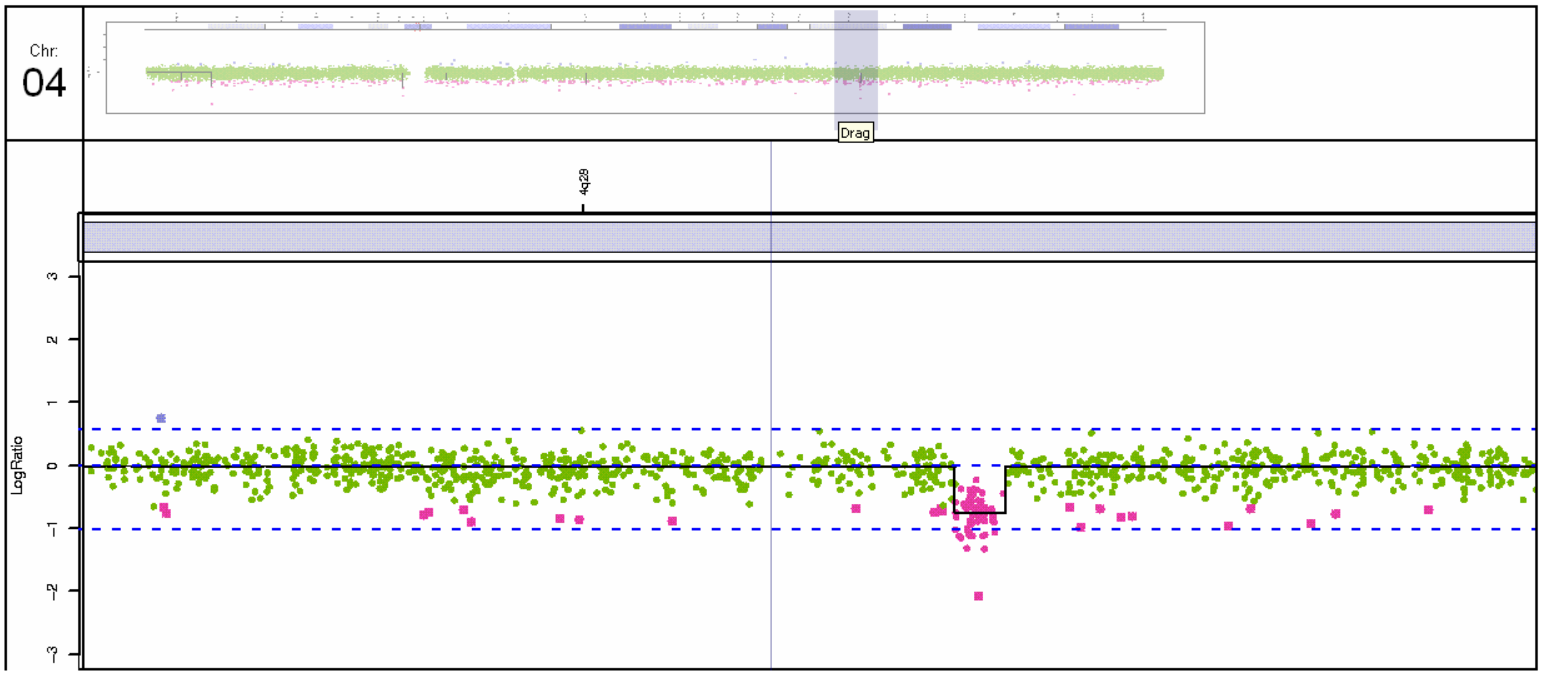
Henrik Bengtsson & Terry Speed

Dept of Statistics, UC Berkeley

September 13, 2007

"Statistics and Genomics Seminar"

What are copy numbers and segmentation?



Size = 264 kb, Number of SNPs = 72

Objectives

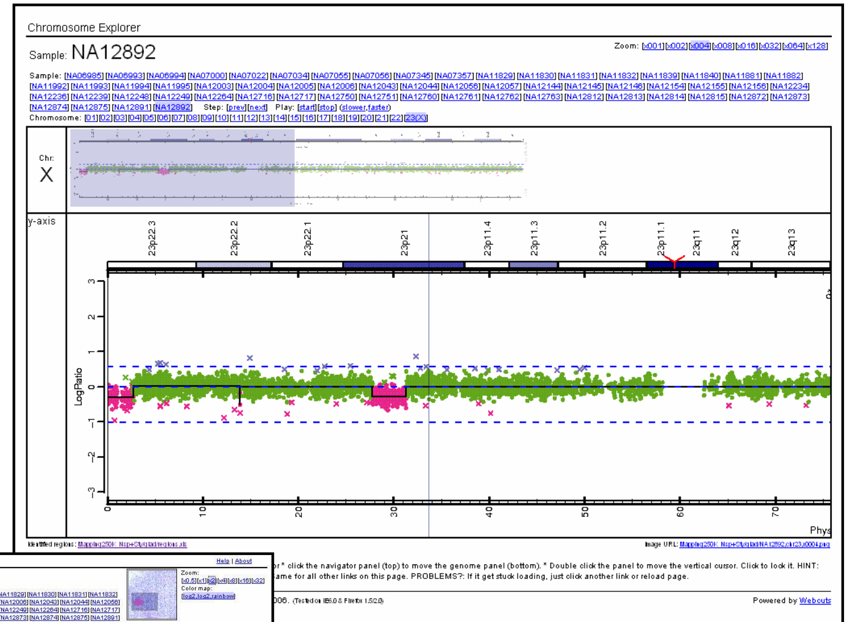
- Total copy number estimation/segmentation
- Estimate single-locus CNs well (segmentation method takes it from there)

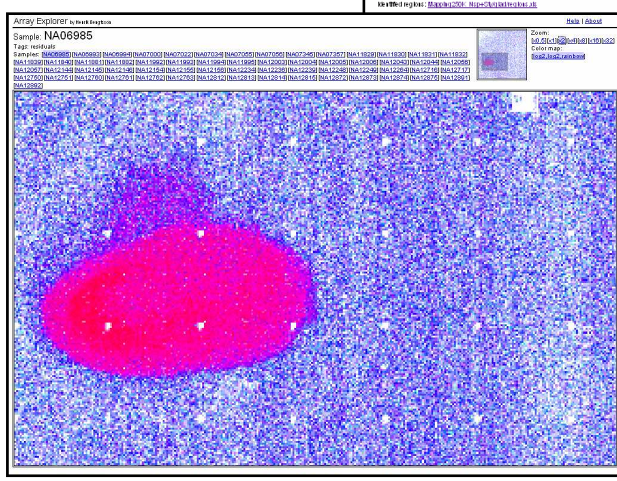
- All generations of Affymetrix SNP arrays:
  - SNP chips: 10K, 100K, 500K
  - SNP & CN chips: 5.0, 6.0

- 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





Acknowledgments

WEHI, Melbourne:
Ken Simpson

UC Berkeley:
James Bullard
Kasper Hansen
Elizabeth Purdom

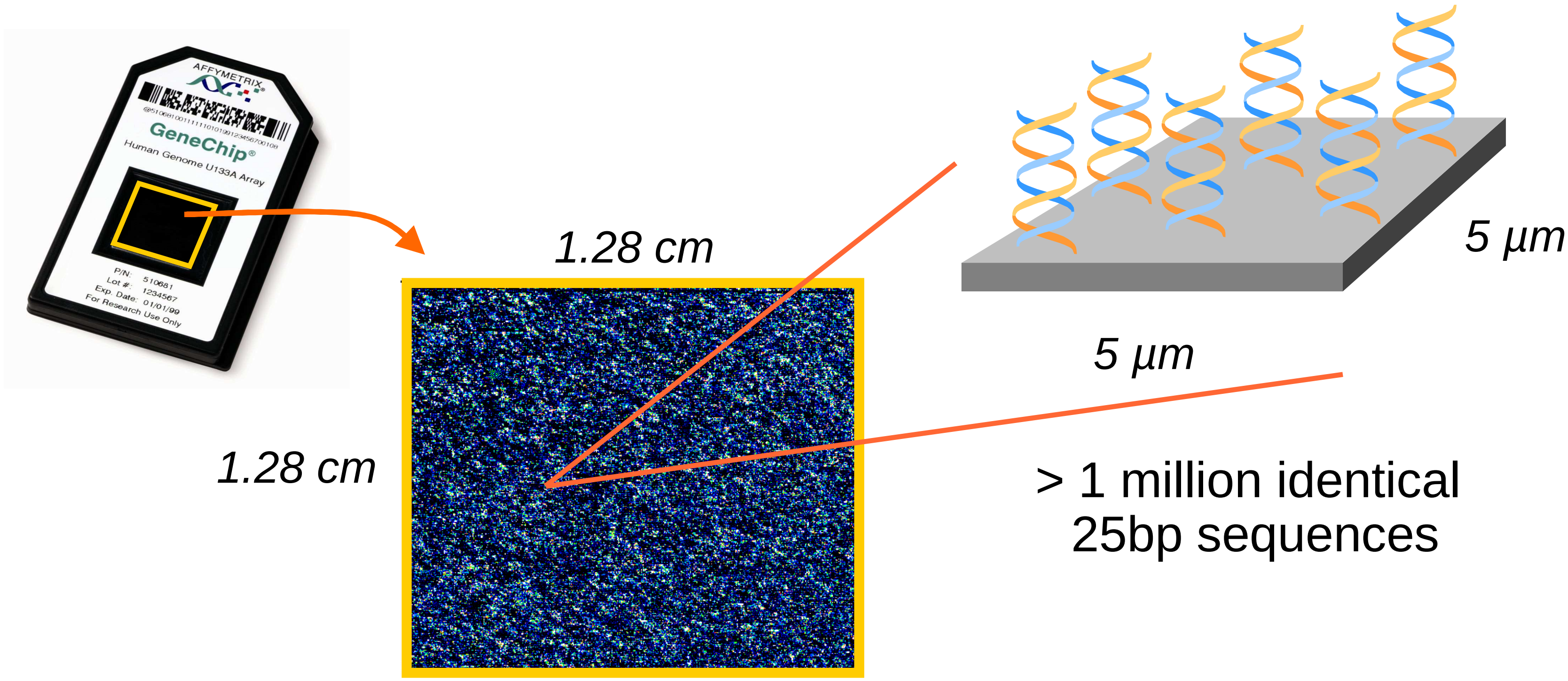
ISREC, Lausanne:
“Asa” Wirapati

John Hopkins:
Benilton Carvalho
Rafael Irizarry

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

Affymetrix chips

Generic Affymetrix chip



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**.

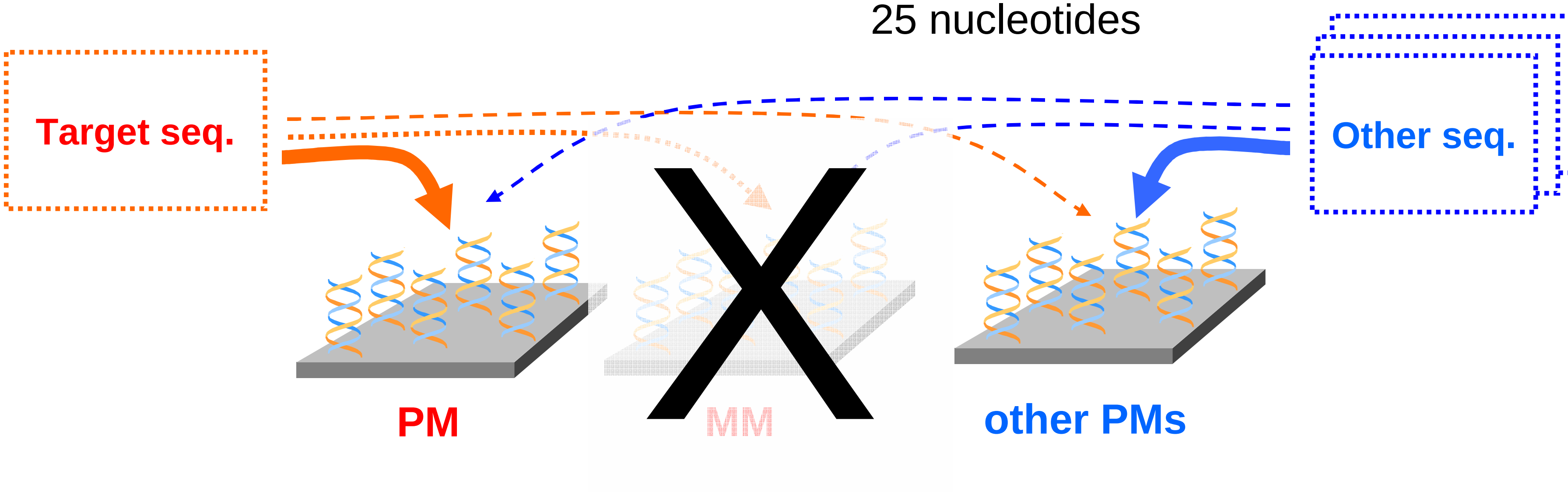
Affymetrix probe terminology

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 genomes may differ by a single nucleotide (A, T, C, or G).

Allele A: ***A***

...CGTAGCCATCGGTA/GTACTCAATGATAG...

Allele B: ***G***

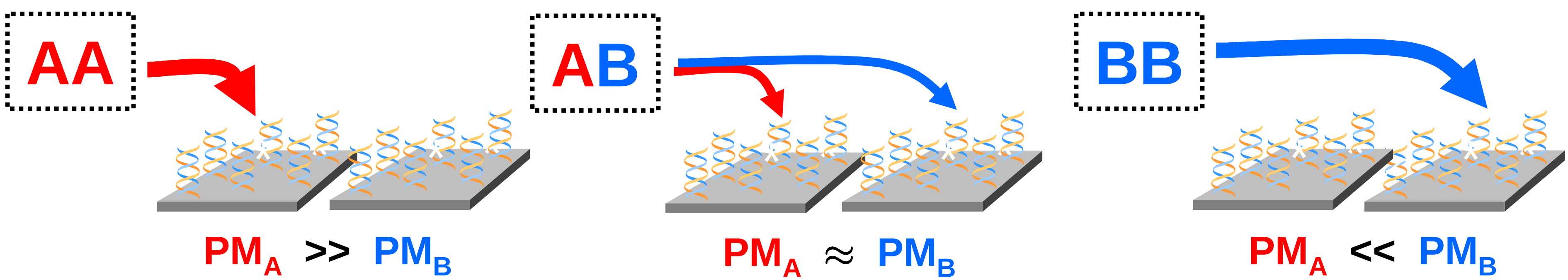
A person is either AA, AB, or BB at this SNP.

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!)



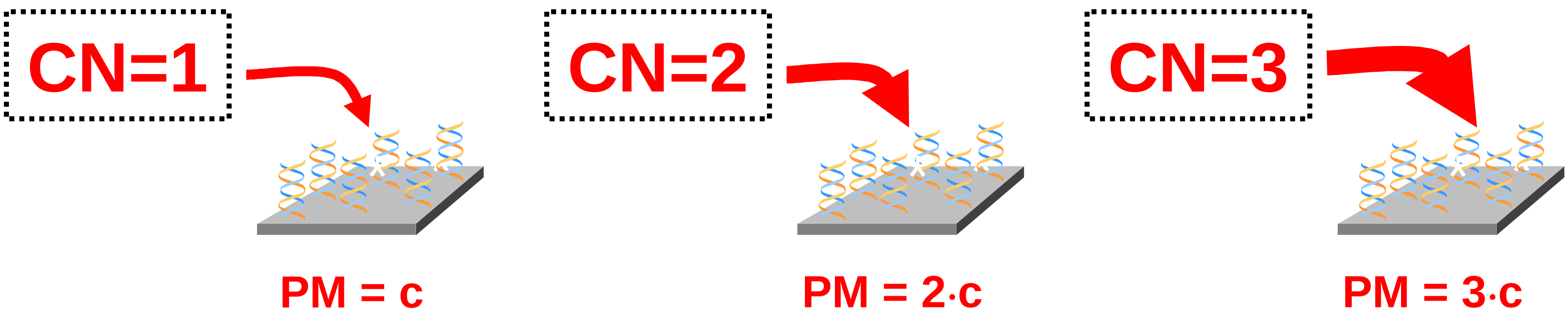
Affymetrix SNP & CN chips

(Genome-Wide Human SNP Array 5.0 & 6.0)

Copy-number/non-polymorphic probes (CNPs)

CN locus: *...CGTAGCCATCGGTAAGTACTCAATGATAG...*

PM: **ATCGGTAGCCATTCATGAGTTACTA**



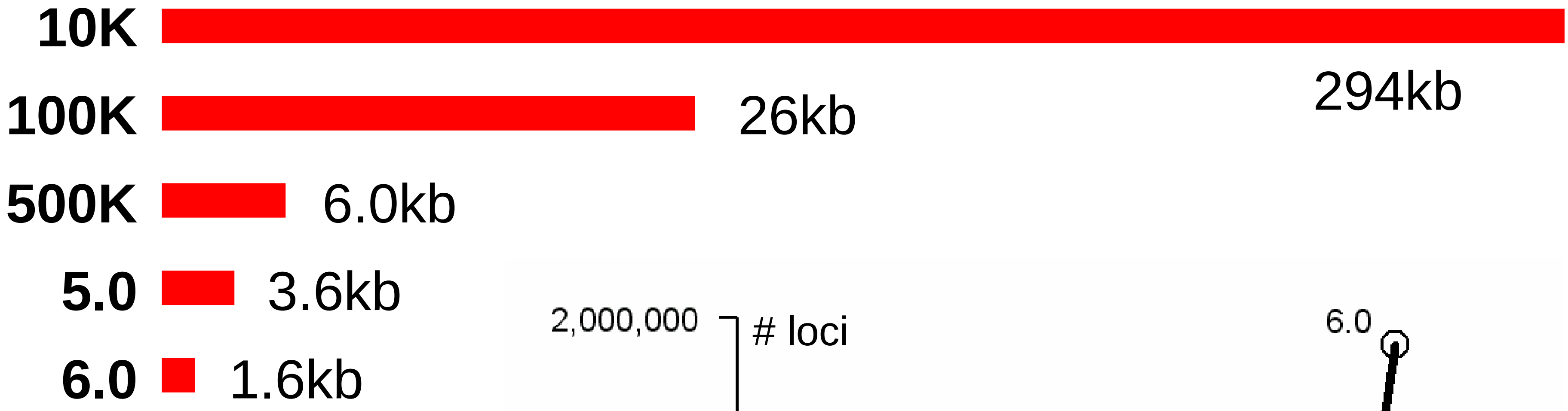
Genome-Wide Human SNP Array 6.0
includes frequently requested properties

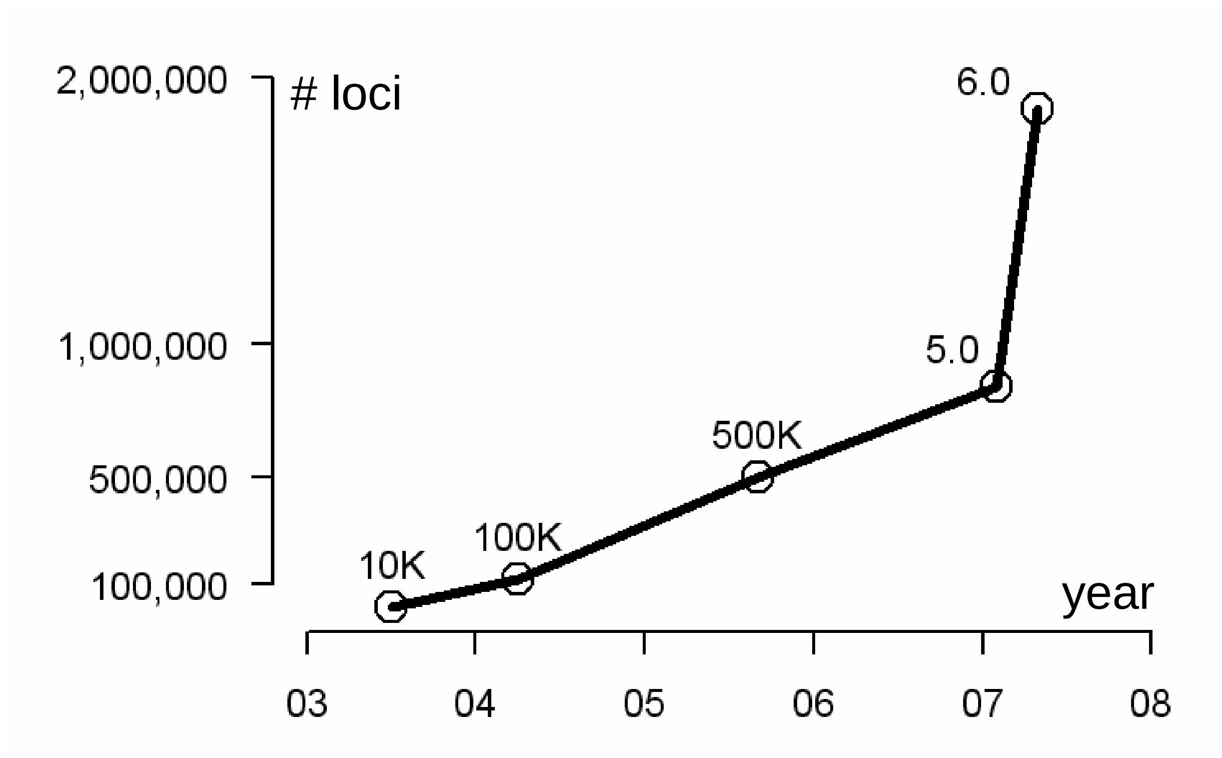
- **> 906,600 SNPs:**
  - Unbiased selection of 482,000 SNPs: historical SNPs from the SNP Array 5.0 (== 500K)
  - Selection of additional 424,000 SNPs:
    - Tag SNPs
    - SNPs from chromosomes X and Y
    - Mitochondrial SNPs
    - Recent SNPs added to the dbSNP database
    - SNPs in recombination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

- **> 946,000 copy number probes (CNPs):**
  - 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

Large increase in density

4× further out…





History of SNP & CNP chips
Affymetrix & Illumina are competing

	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/, September 2007.

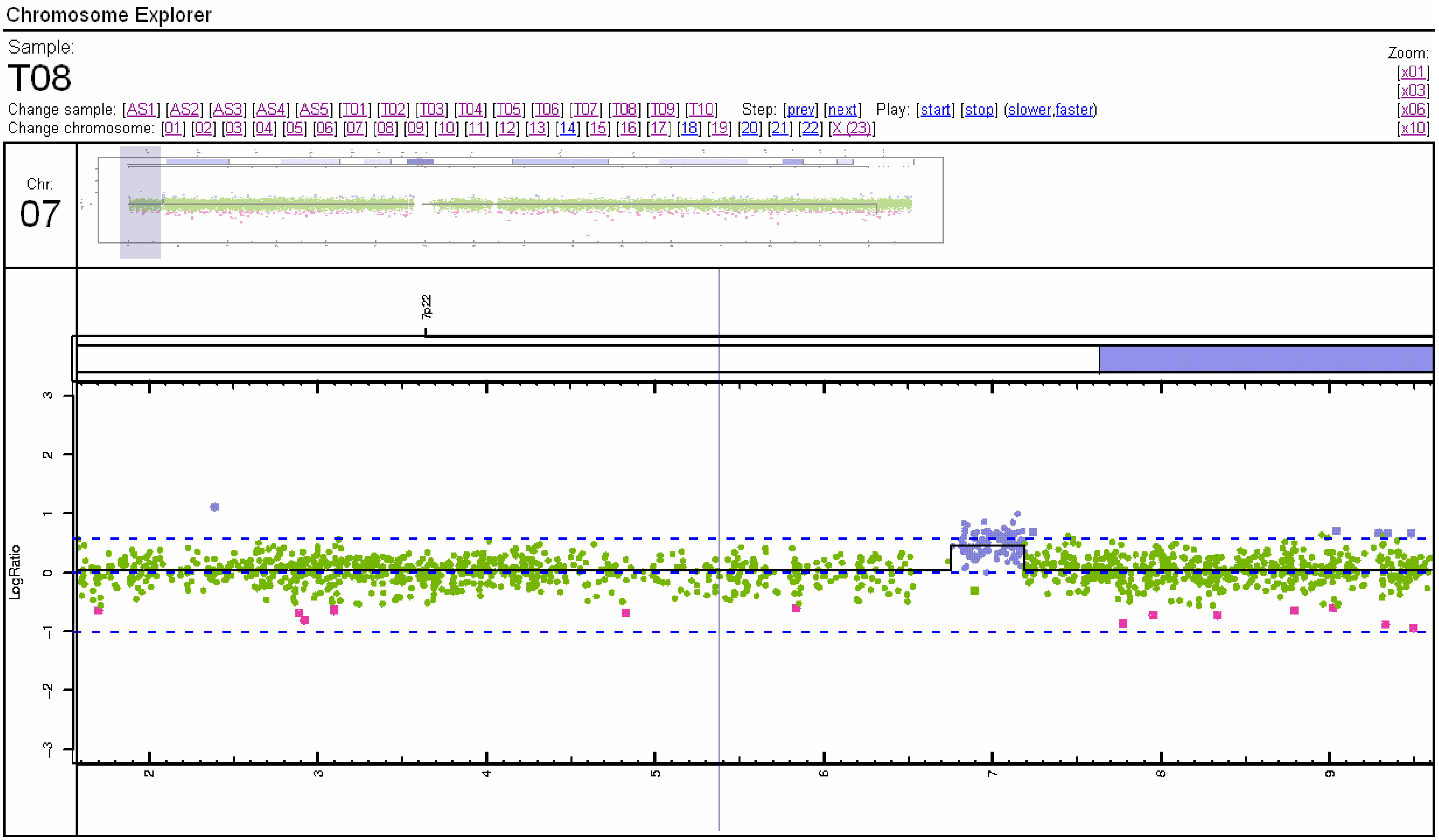
Copy-number analysis with SNP arrays

(10K, 100K, 500K)

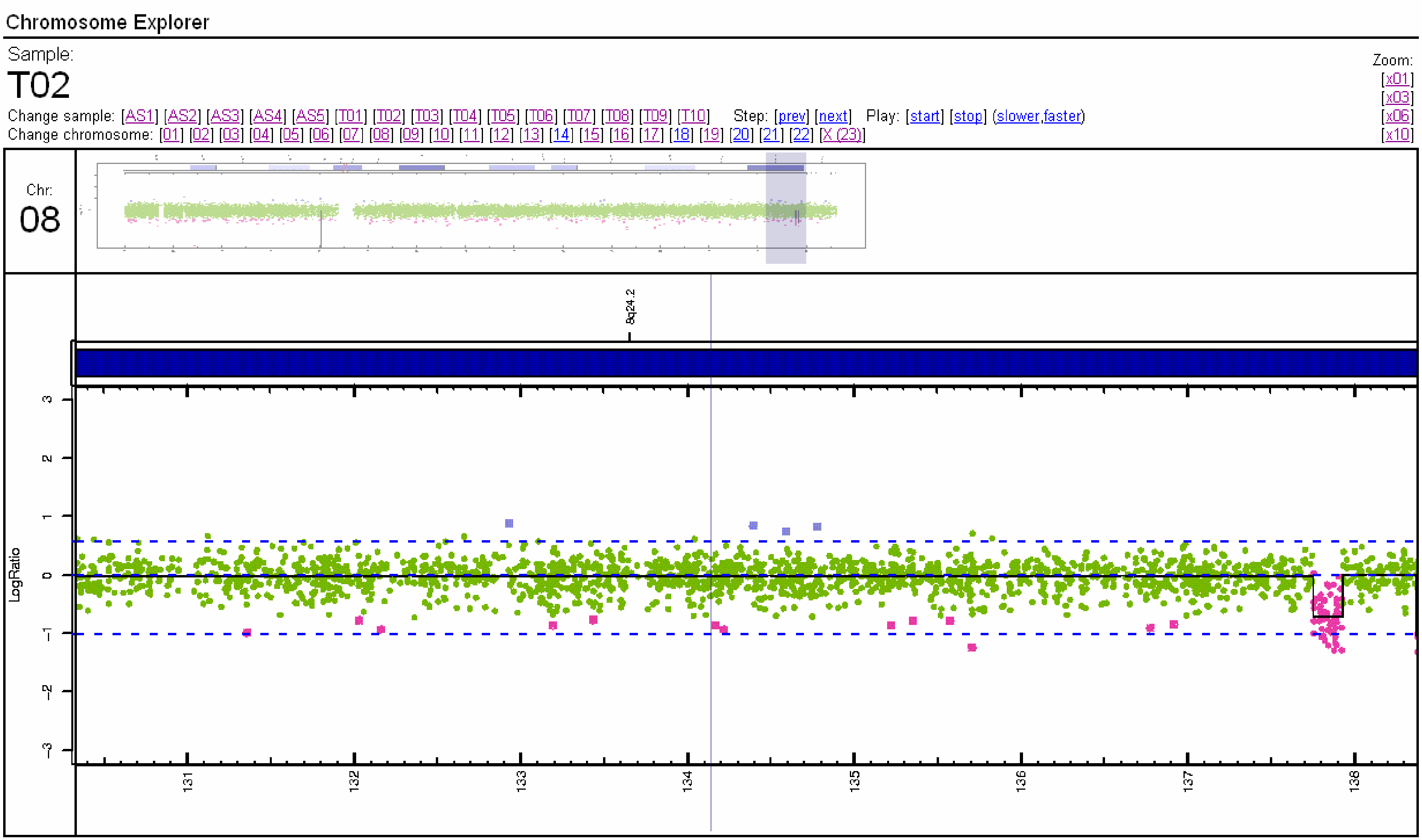
SNP chips can be used to determine copy number

Some sample figures based on a **250K SNP chip** showing deletions and amplifications

Size = 424 kb, Number of SNPs = 118
Results using of dChip and GLAD.



Size = 168 kb, Number of SNPs = 55



Our method:

CRMA

(10K, 100K, 500K)

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*

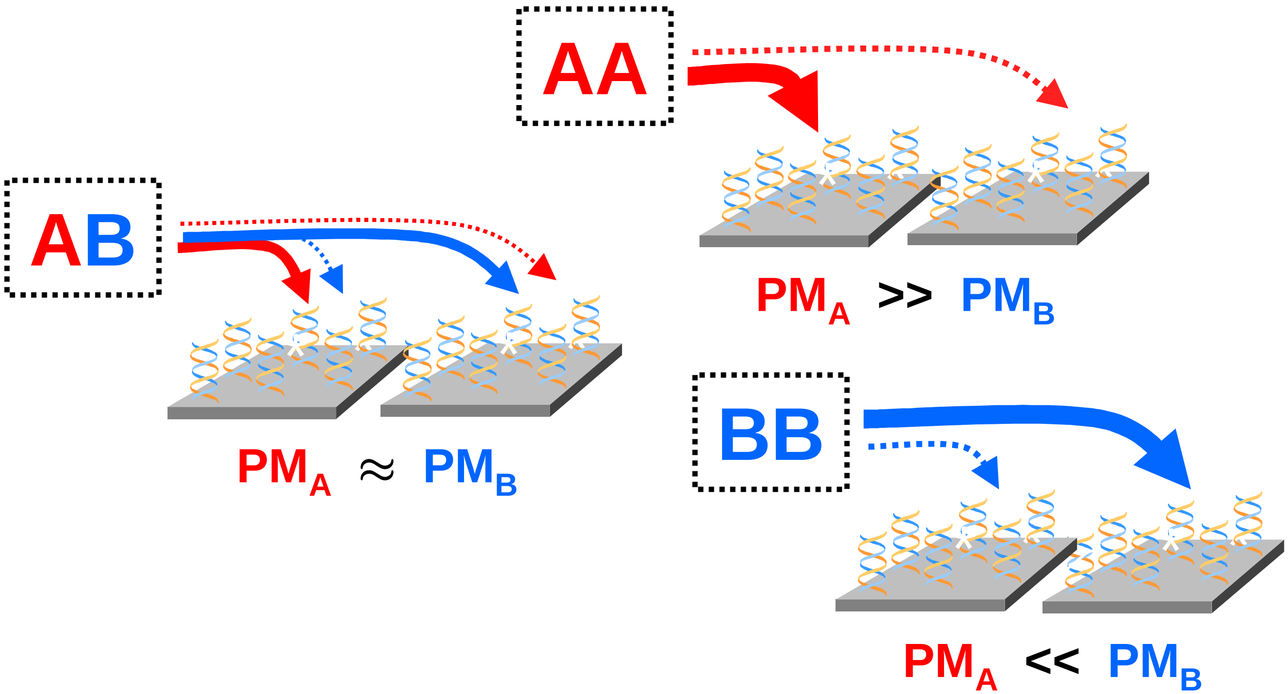
Copy-number estimation using Robust Multichip Analysis (CRMA)

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

Cross-hybridization:

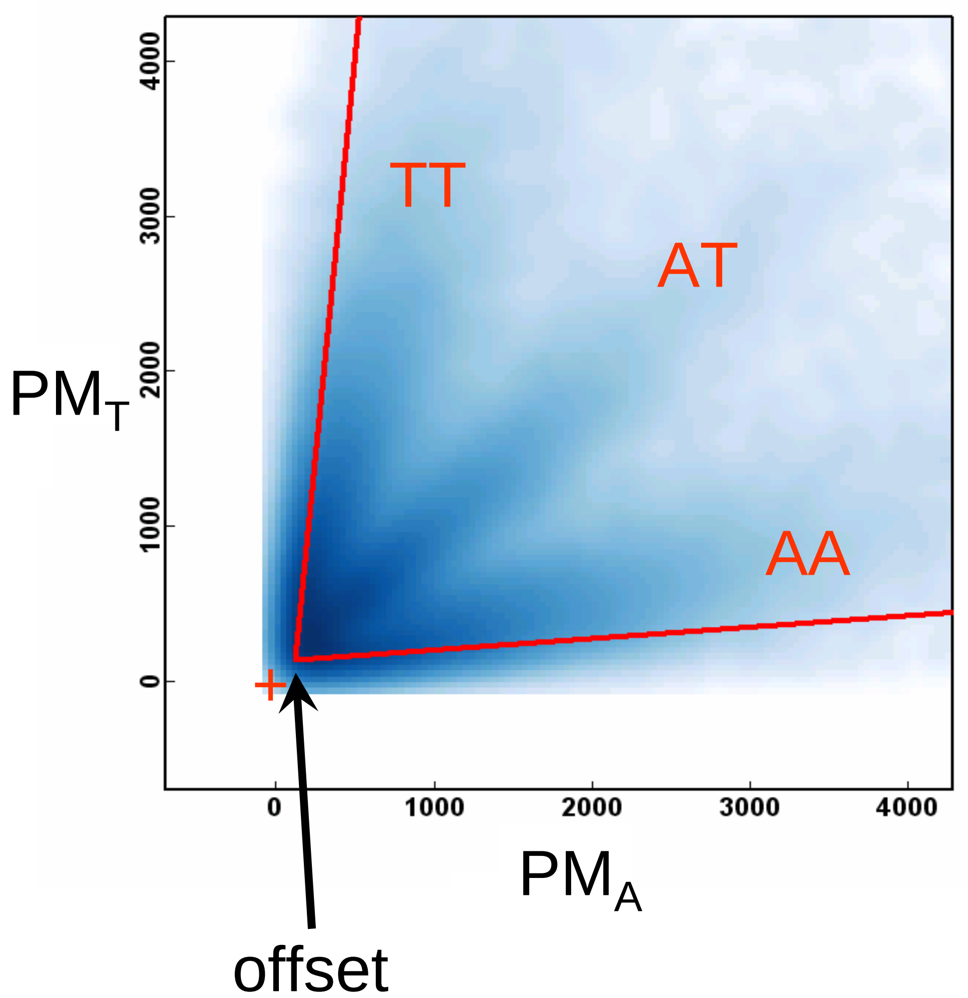
Allele A: ***TCGGTAAGTACTC***

Allele B: ***TCGGTATGTACTC***



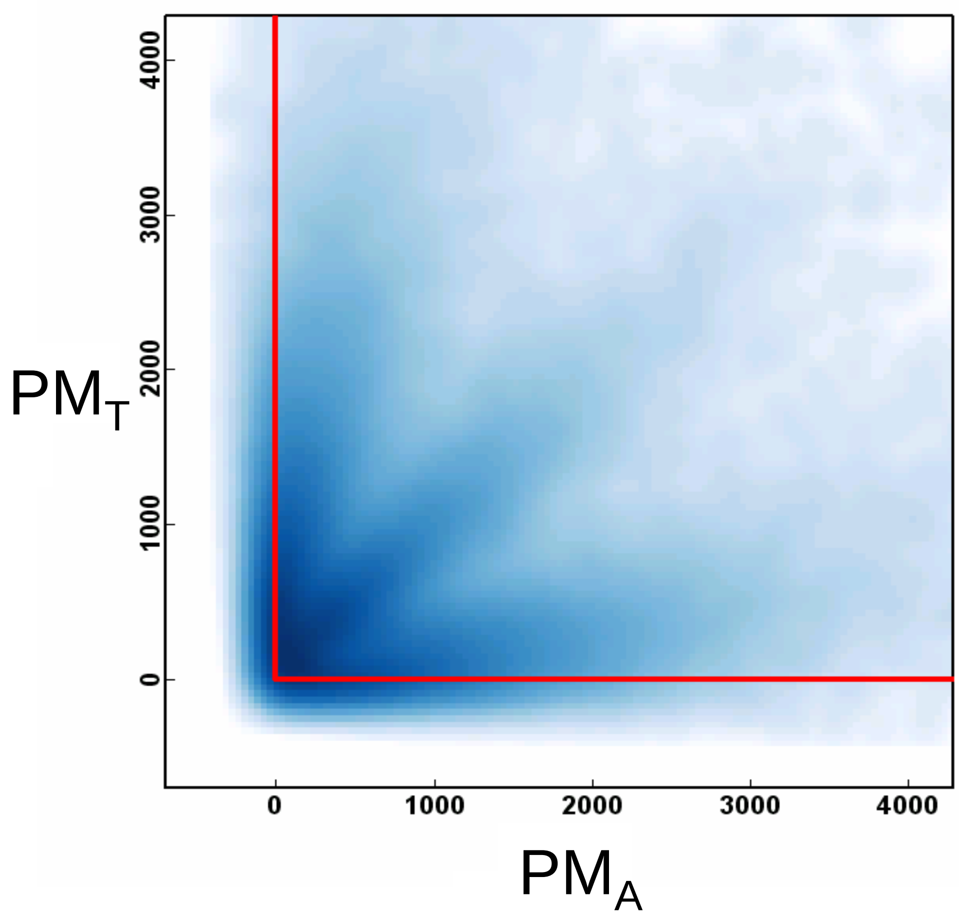
Copy-number estimation using Robust Multichip Analysis (CRMA)

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



Copy-number estimation using Robust Multichip Analysis (CRMA)

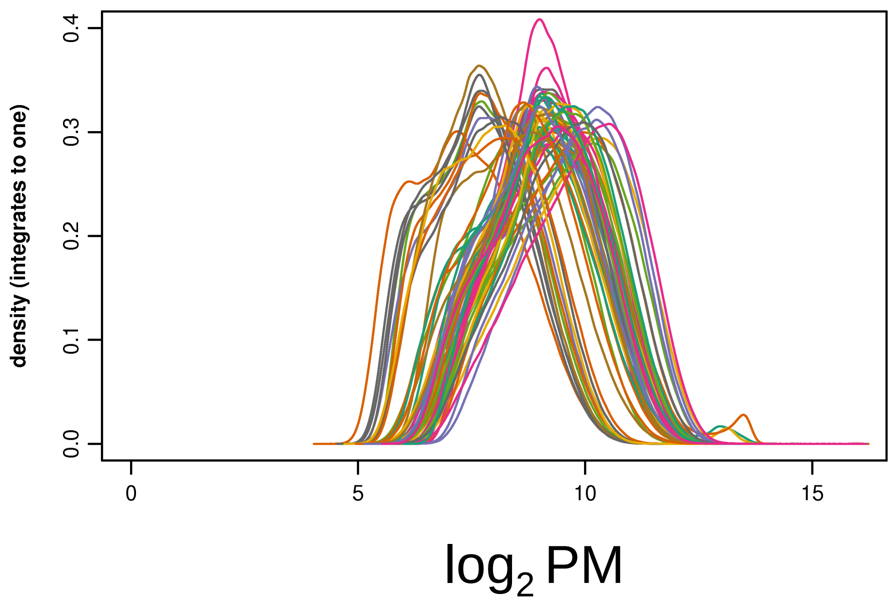
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$



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

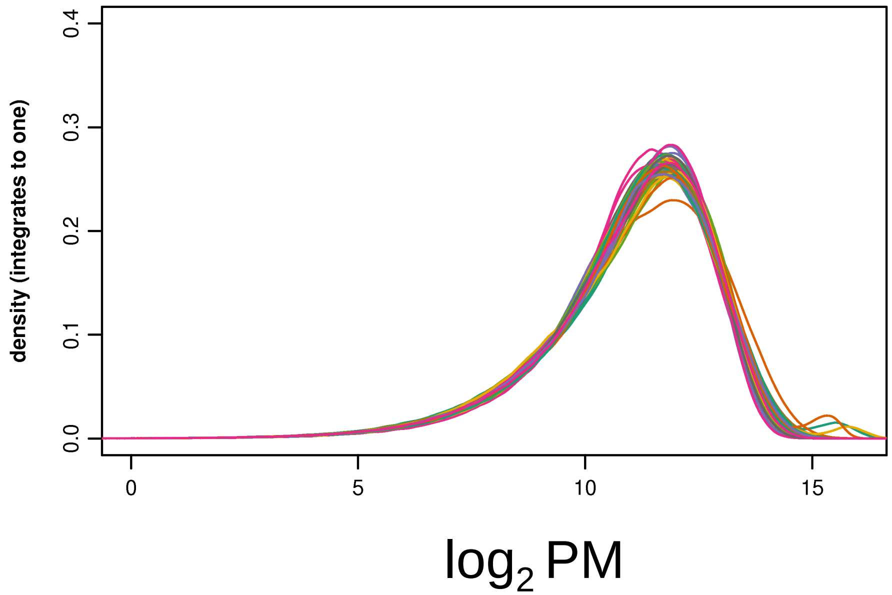
Crosstalk calibration corrects for differences in distributions too



Copy-number estimation using Robust Multichip Analysis (CRMA)

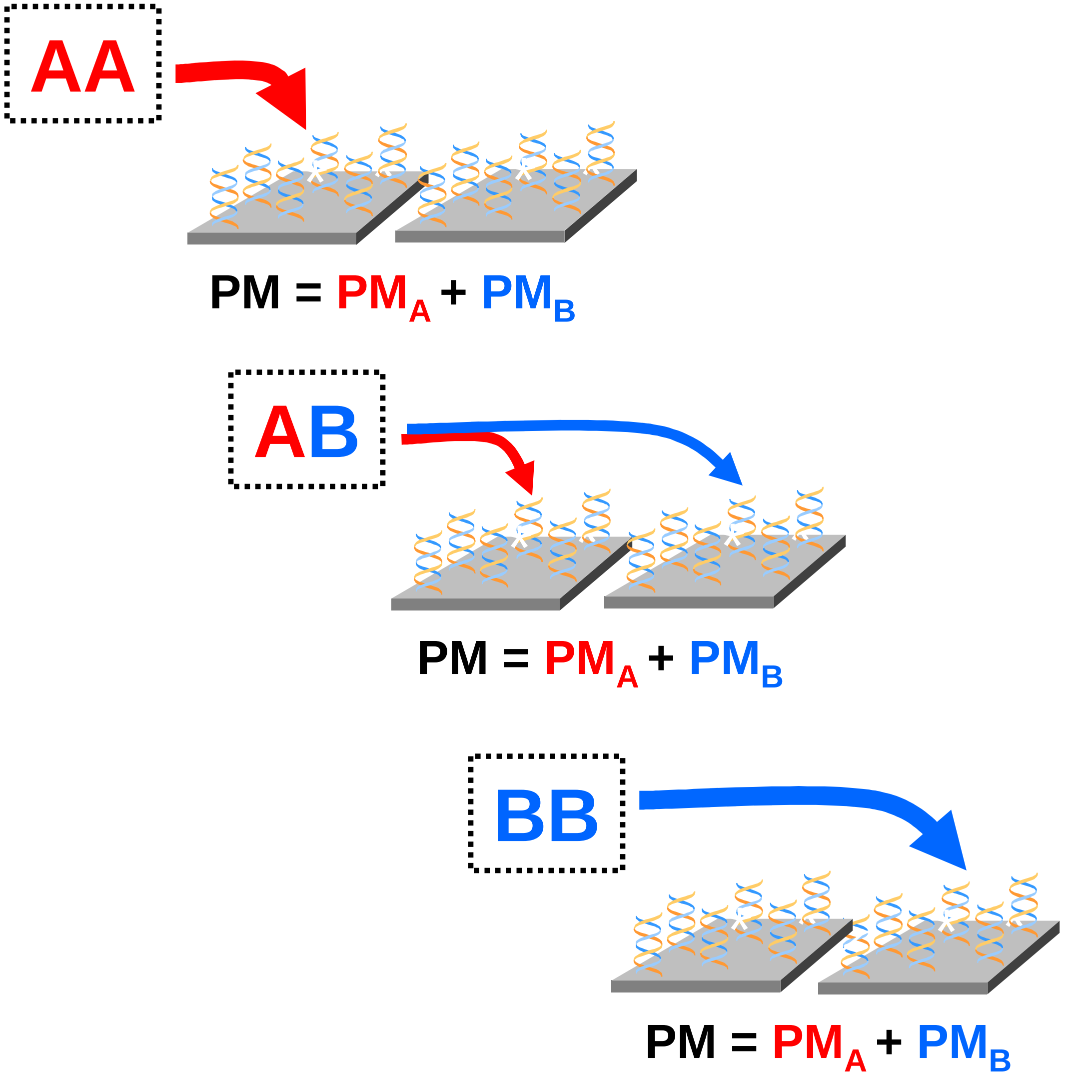
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM = PM_A + PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

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



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:

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

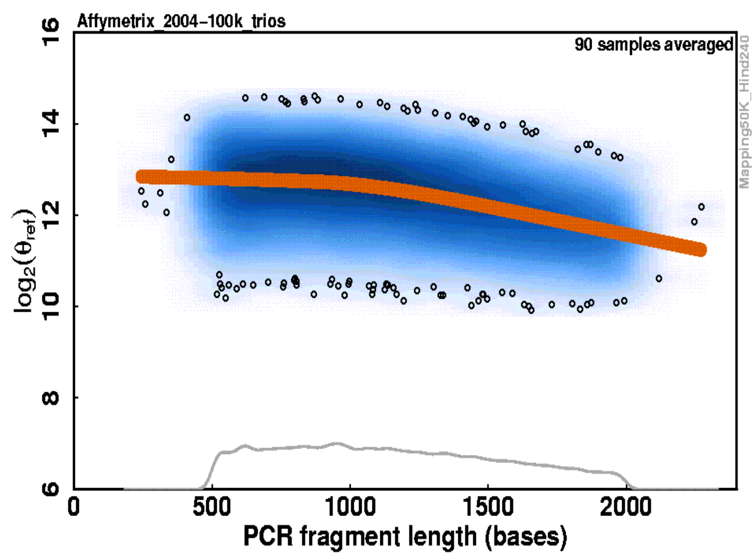
sample *i*, SNP *j*, probe *k*.

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

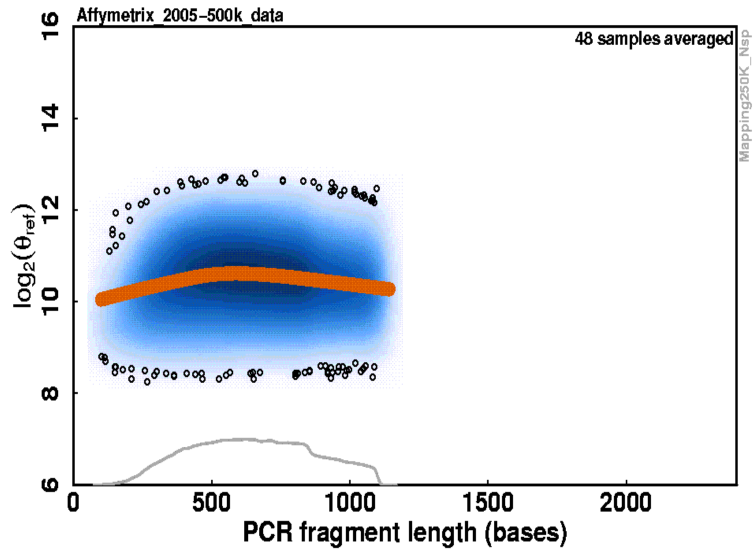


100K

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

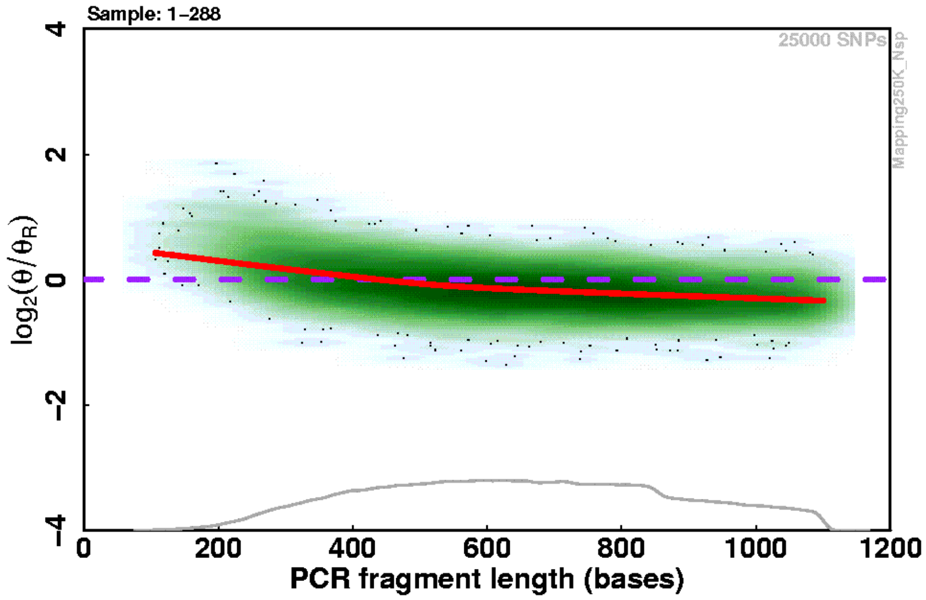


500K

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

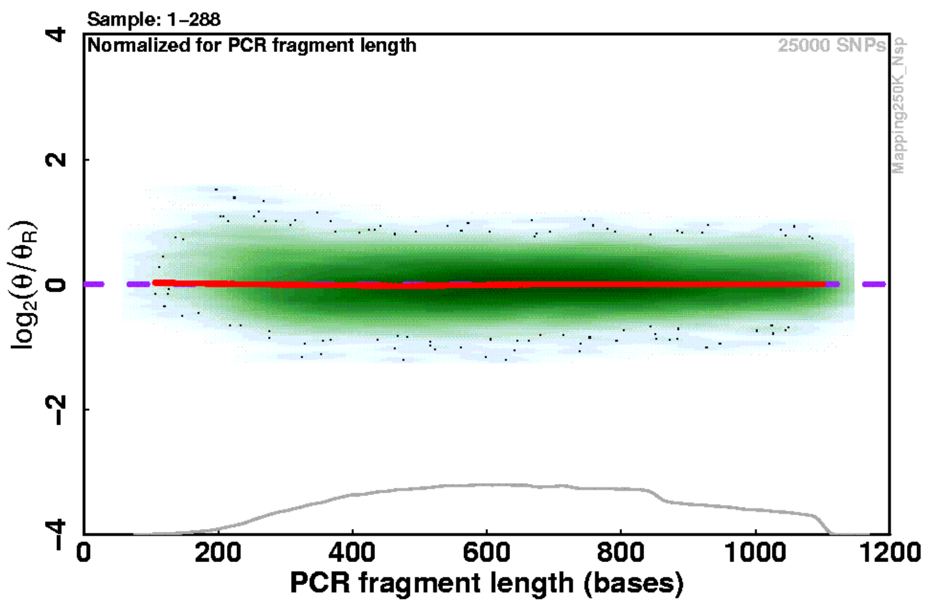
Normalize to get same fragment-length effect for all hybridizations



Copy-number estimation using Robust Multichip Analysis (CRMA)

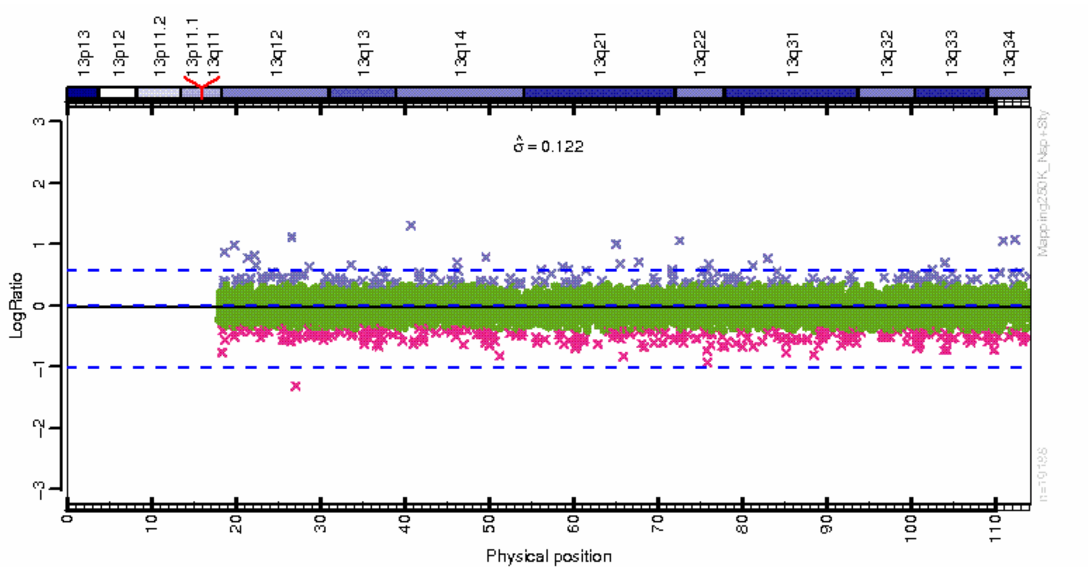
	CRMA
Preprocessing *(probe signals)*	allelic crosstalk (quantile)
Total CNs	$PM=PM_A+PM_B$
Summarization *(SNP signals θ)*	log-additive (PM-only)
Post-processing	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

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



Comparison

(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

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

Chromosome X is known:

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

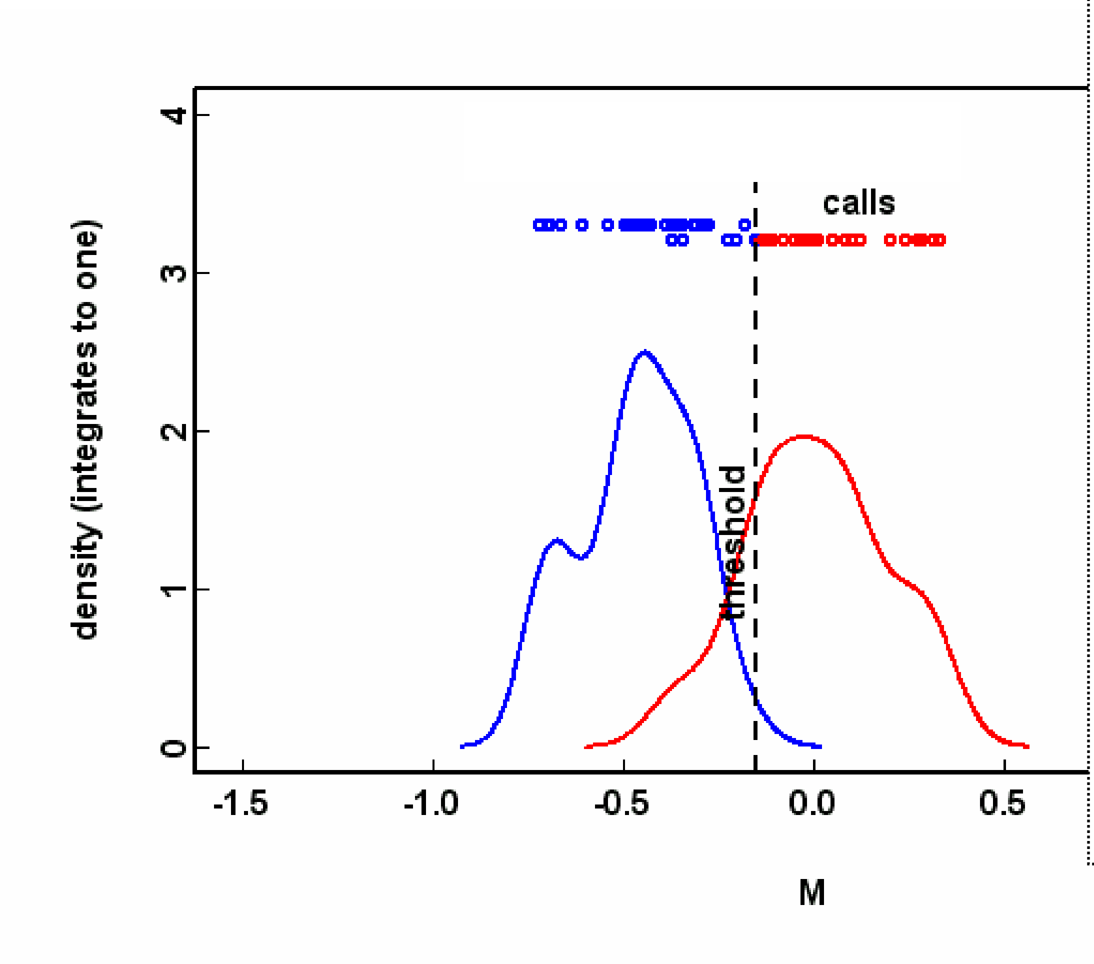
5,608 SNPs

Classification rule:

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

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

Calling samples for SNP_A-1920774



males: 30
females: 29

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

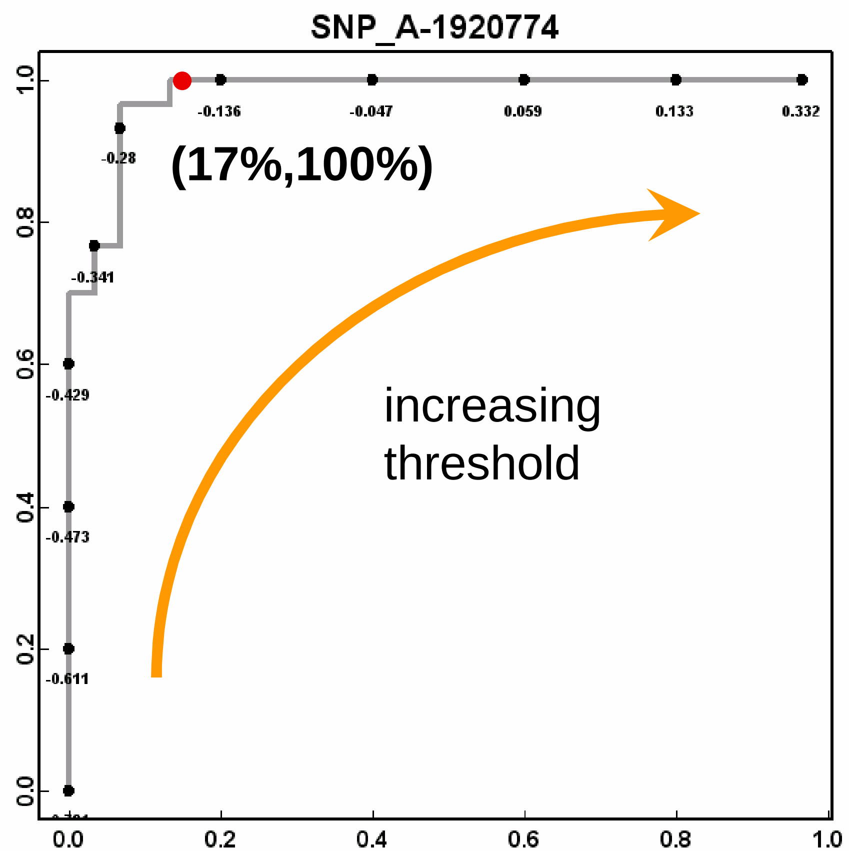
Calling a male male:
#True-positives: 30
TP rate: 30/30 = 100%

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

Receiver Operator Characteristic (ROC)

TP rate

(correctly calling a males male)



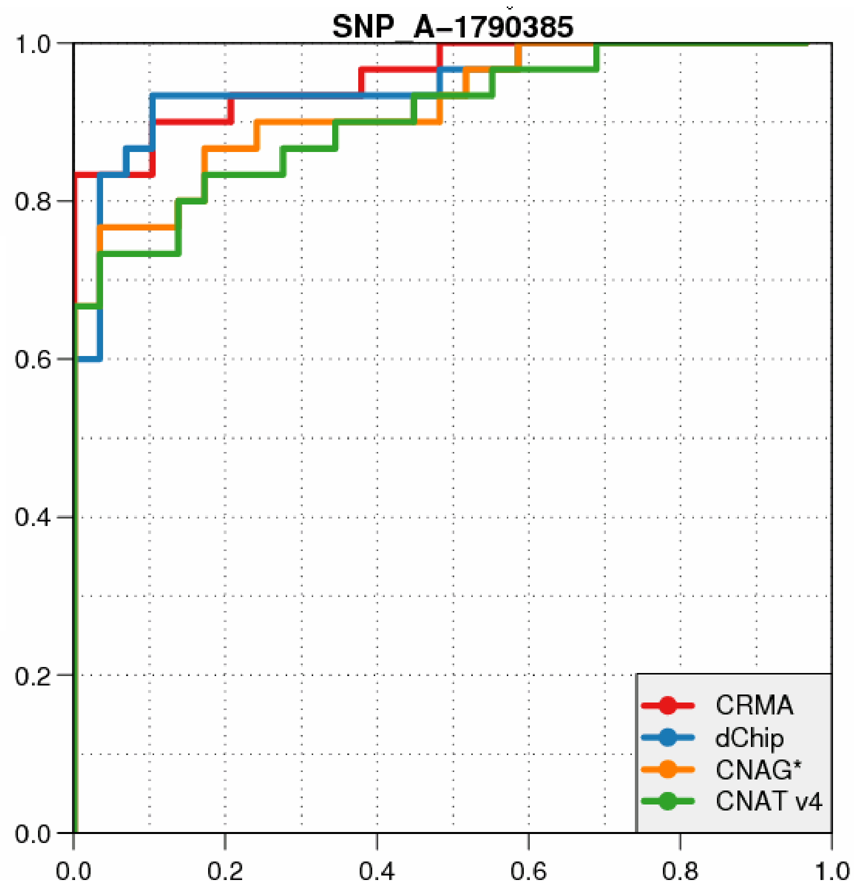
FP rate

(incorrectly calling females male)

Single-SNP comparison
A random SNP

TP rate

(correctly calling a males male)



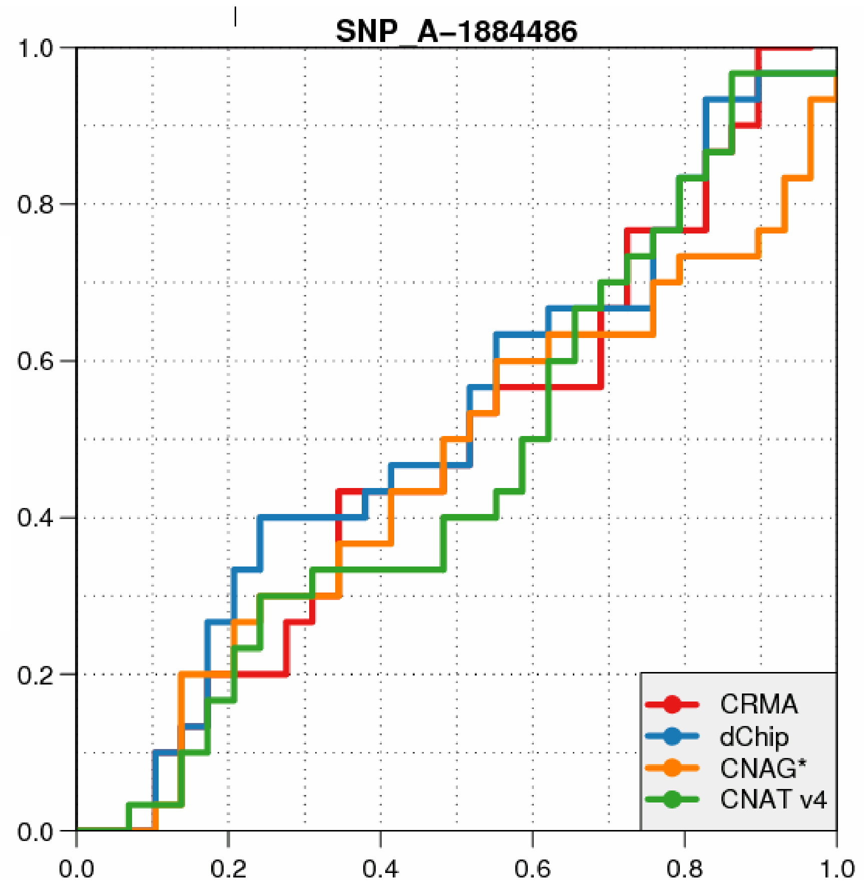
FP rate

(incorrectly calling females male)

Single-SNP comparison
A non-differentiating SNP

TP rate

(correctly calling a males male)

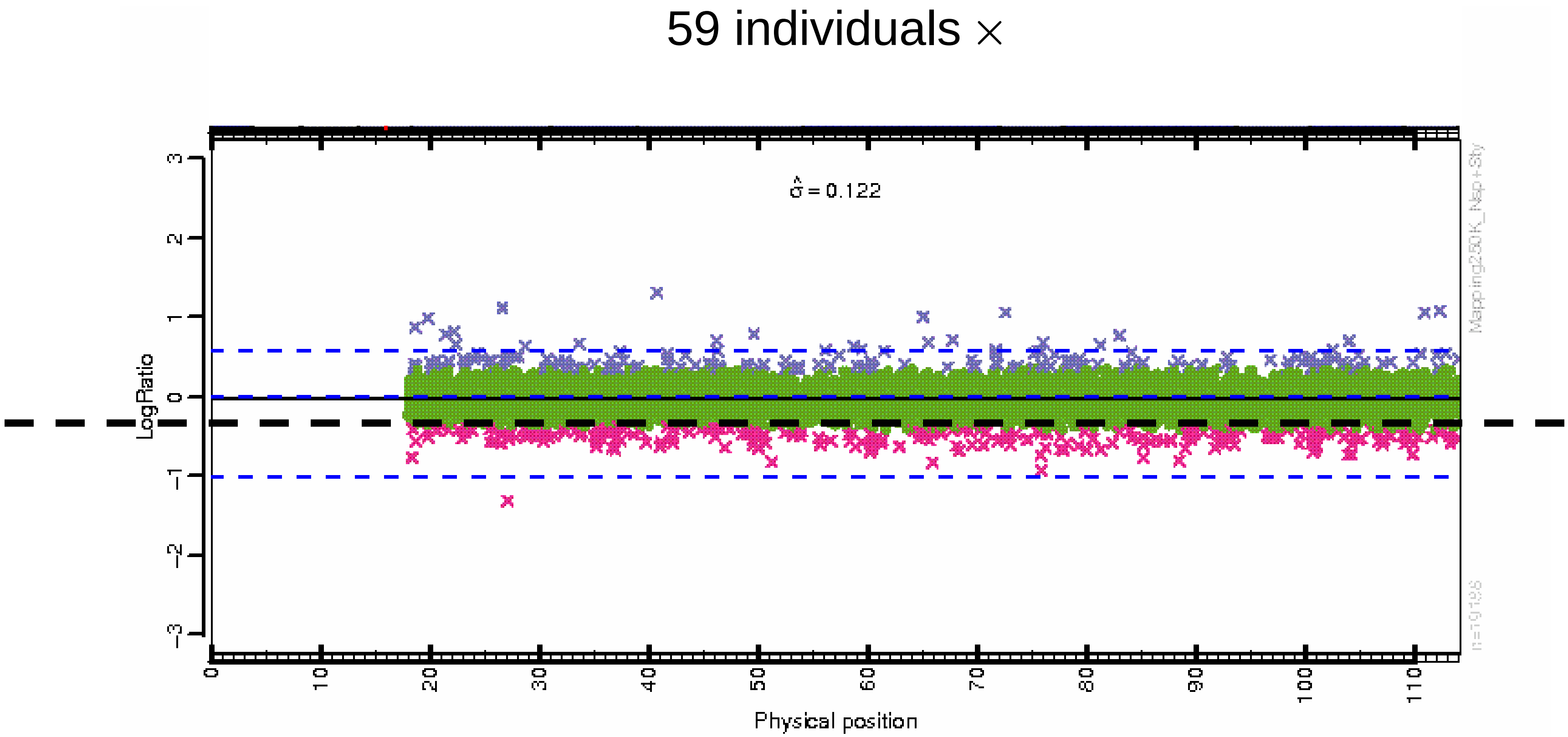


FP rate

(incorrectly calling females male)

Performance of an average SNP with a common threshold

59 individuals ×

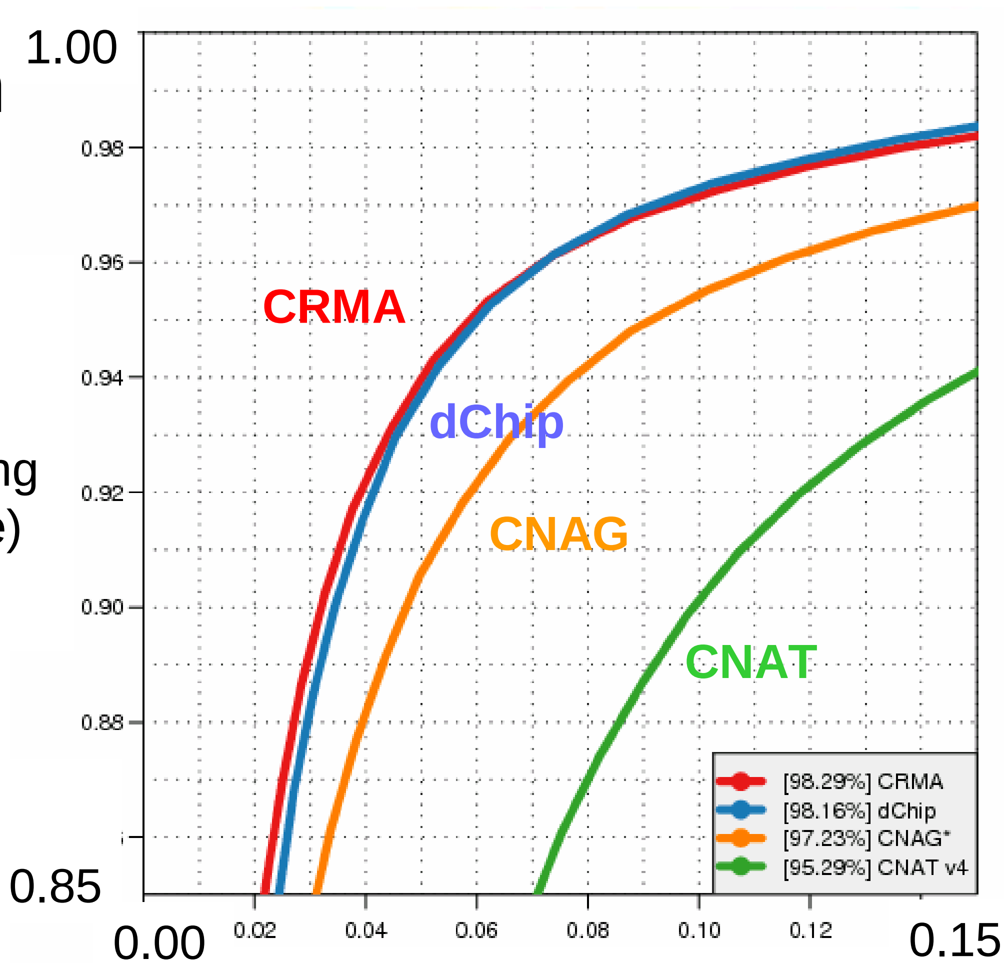


CRMA & dChip perform better for an average SNP *(common threshold)*

Zoom in

TP rate

(correctly calling a males male)



FP rate

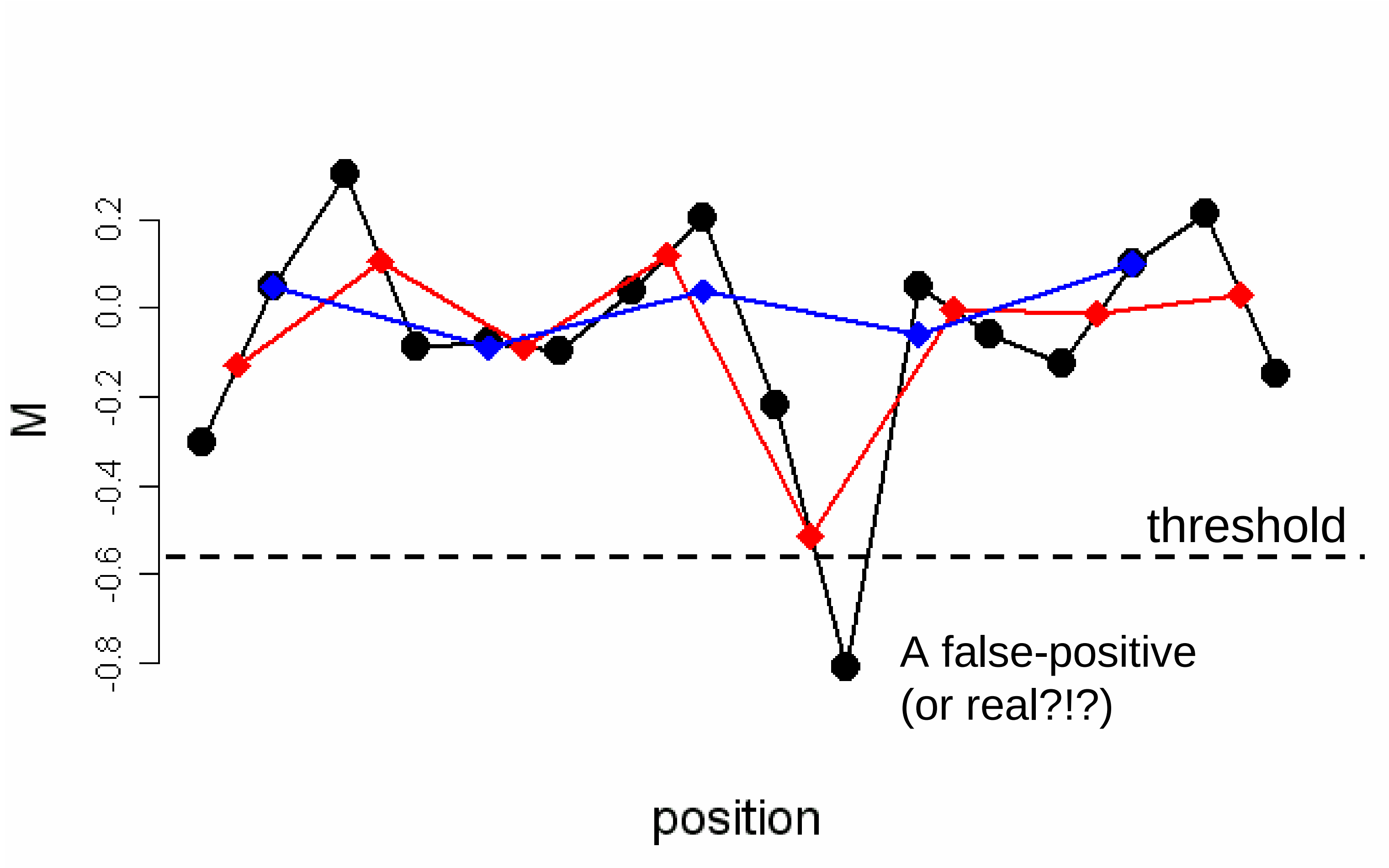
(incorrectly calling females male)

Number of calls:
59×5,608 = 330,872

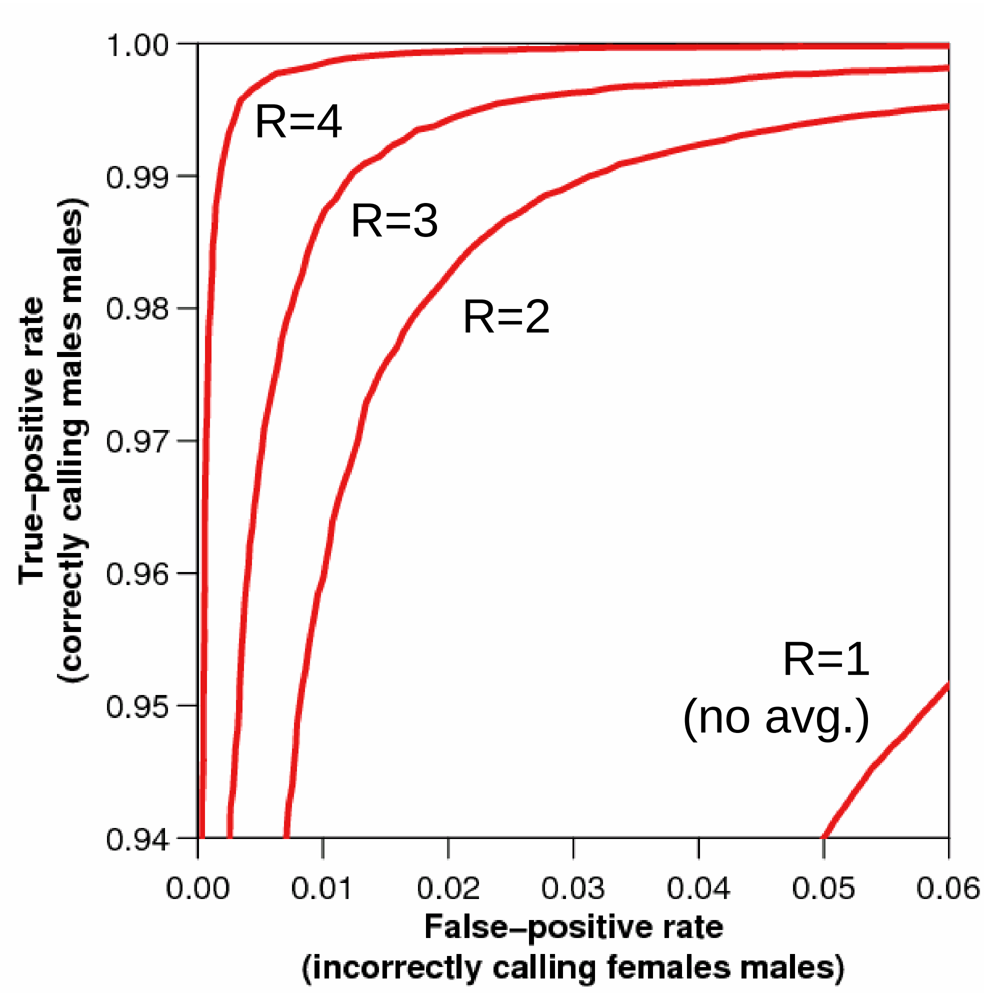
"Smoothing"

Average across SNPs
non-overlapping windows

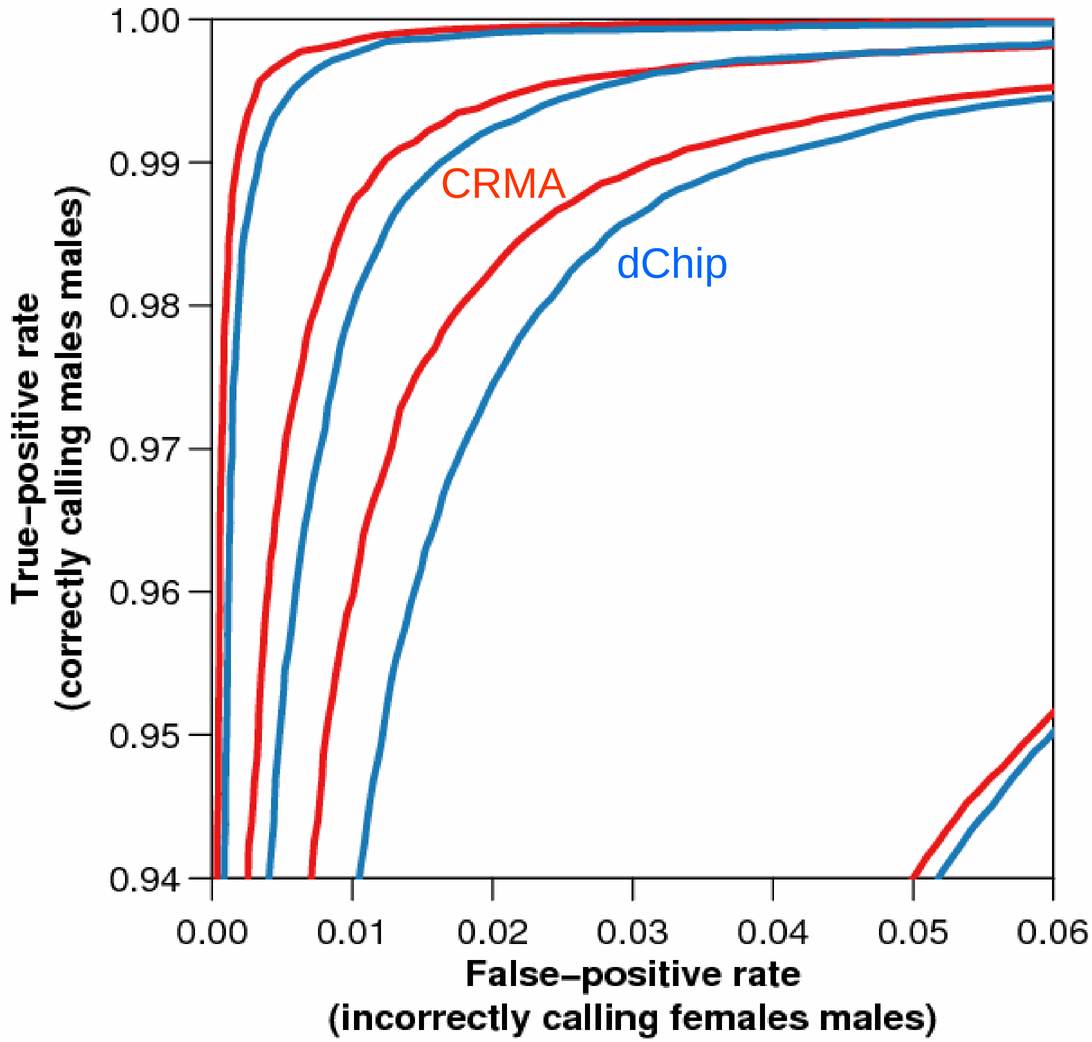
Averaging three and three (R=3)



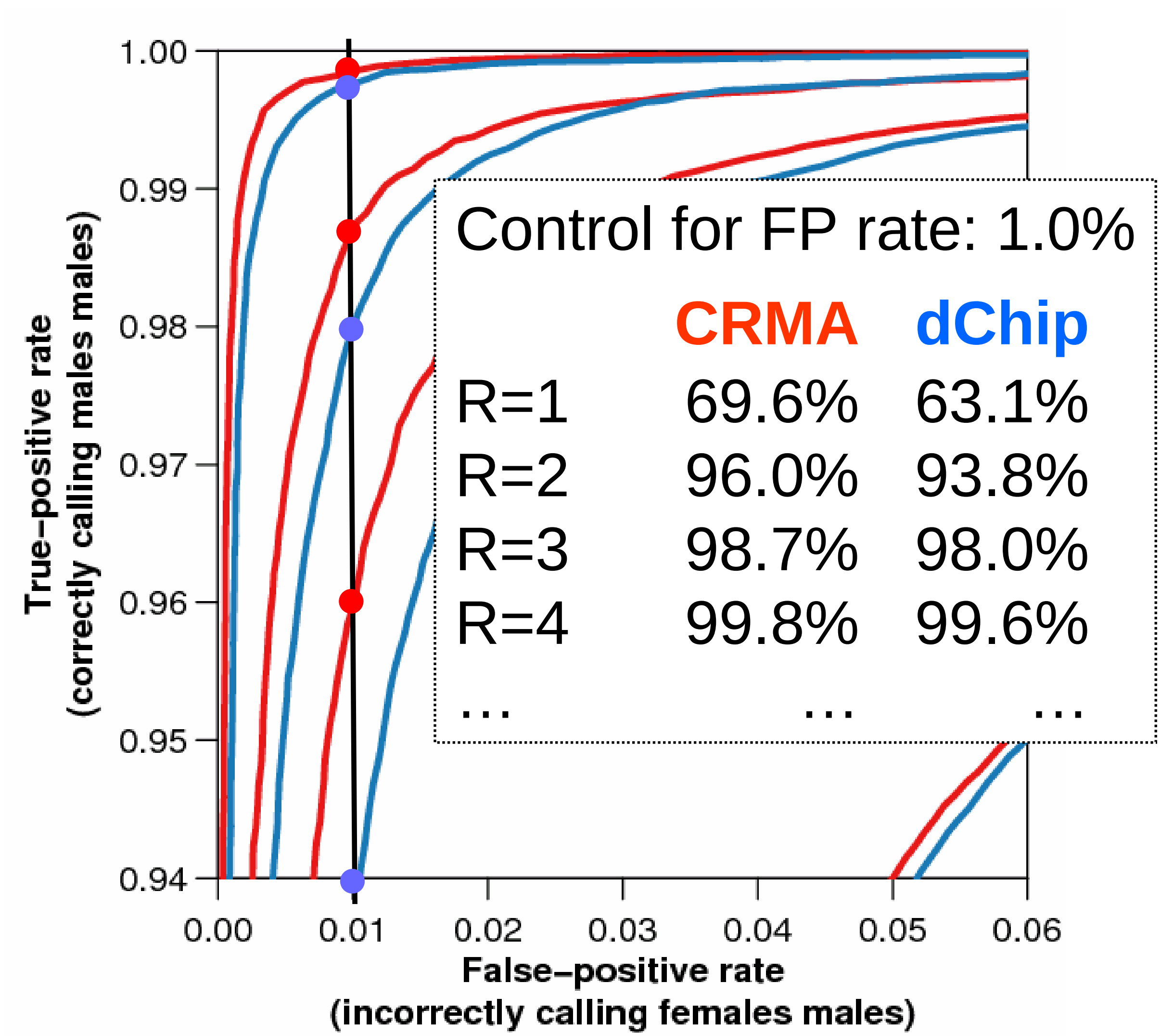
Better detection rate when averaging
(with risk of missing short regions)



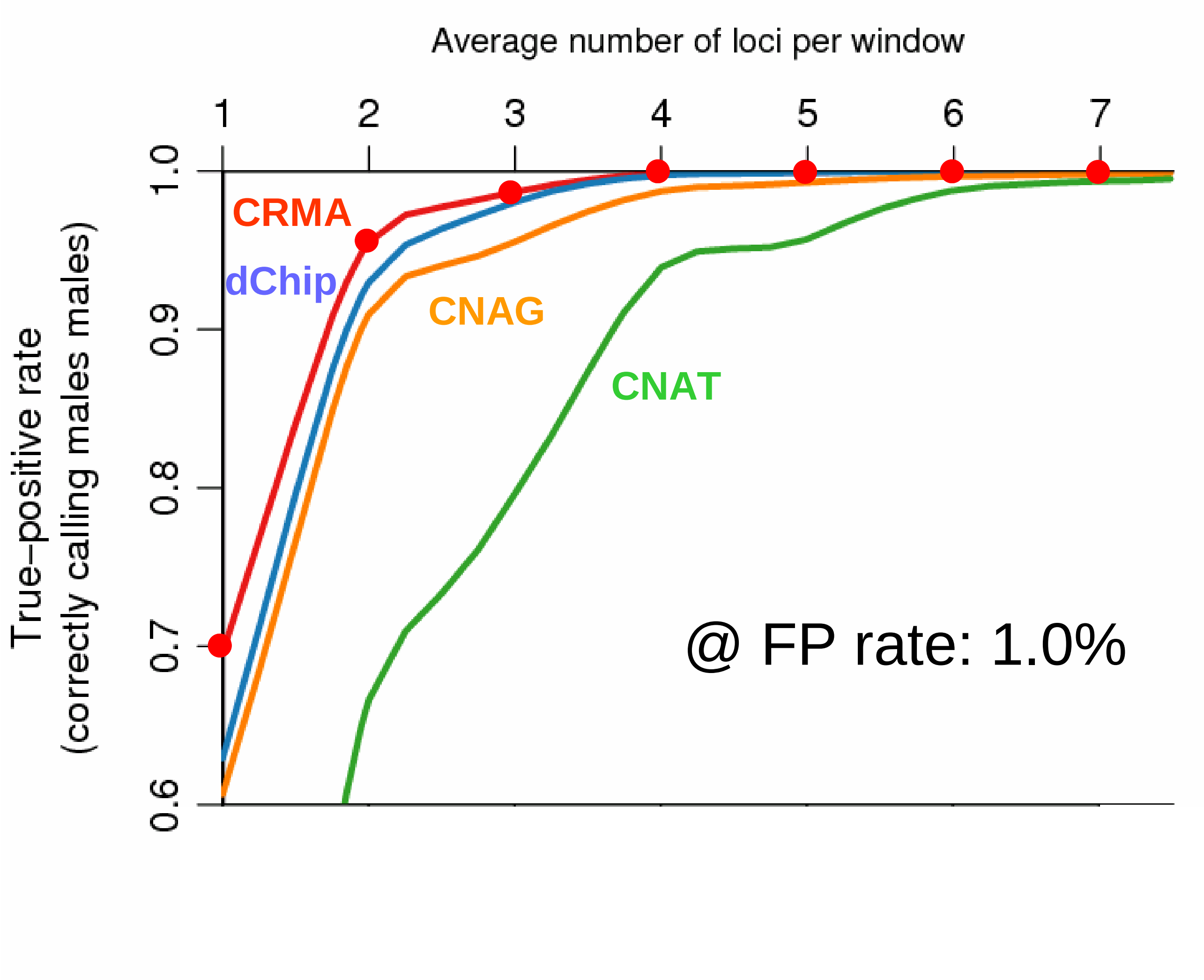
CRMA does better than dChip



CRMA does better than dChip



Comparing methods by "resolution" *controlling for FP rate*



Early work:

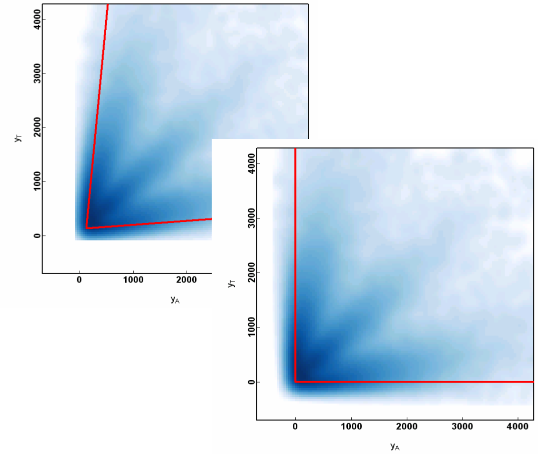
CRMA 6

(SNP 5.0 & 6.0 chips)

CRMA with CN probes

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

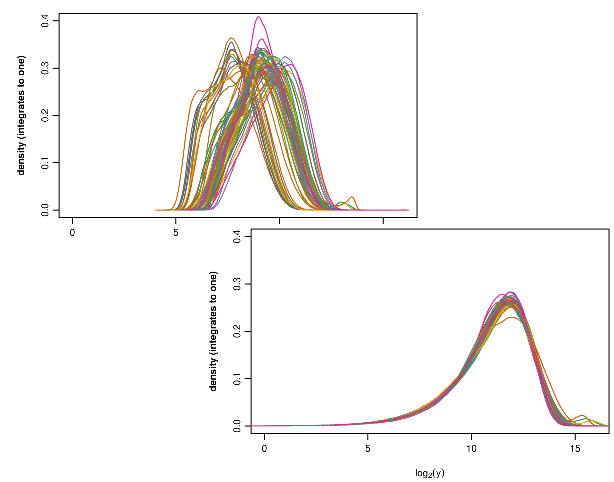
Allelic crosstalk calibration -incorporating CN probes



SNPs:

For each allele pair in {AC, AG, AT, CG, CT, GT}:

1) Estimate crosstalk model:
   offset: $\mathbf{a}_{SNP} = (a_A, a_B)$
   crosstalk matrix: $\mathbf{S} = [S_{AA}, S_{AB}; S_{BA}, S_{BB}]$
2) Calibrate probe pairs $\{\mathbf{PM}\} = \{(PM_A, PM_B)\}$:
   $\mathbf{PM'} \leftarrow \mathbf{S}^{-1} (\mathbf{PM} - \mathbf{a}_{SNP})$
3) Rescale $\{PM'_A\}$ and $\{PM'_B\}$ to have average 2200.



CN probes:

1) Calculate the average offset across all alleles:
   offset: $a_{CN} = 1/6 * \sum_k \{w_k*(a_A+a_B)/2\}$,
   with weights w_k corresponding to n:s (above).
2) Calibrate CN probes {y}:
   $PM' \leftarrow PM - a_{CN}$
3) Rescale {PM'} to have average 2200.

Probe-level modelling (PLM)

SNPs:

* Technical replicates:

$\mathbf{PM}_A = (PM_{A1}, PM_{A2}, PM_{A3})$ and $\mathbf{PM}_B = (PM_{B1}, PM_{B2}, PM_{B3})$

All should have the same probe affinities => No probe-affinity model(!)

* Suggestion:

$PM_A = \text{median}\ \{PM_{Ak}\}$

$PM_B = \text{median}\ \{PM_{Bk}\}$

$PM = PM_A + PM_B$ (compare to CN probes!)

$\theta = PM$

CN probes:

* (Mostly) single probe units, i.e. nothing much to do;

$\theta = PM$

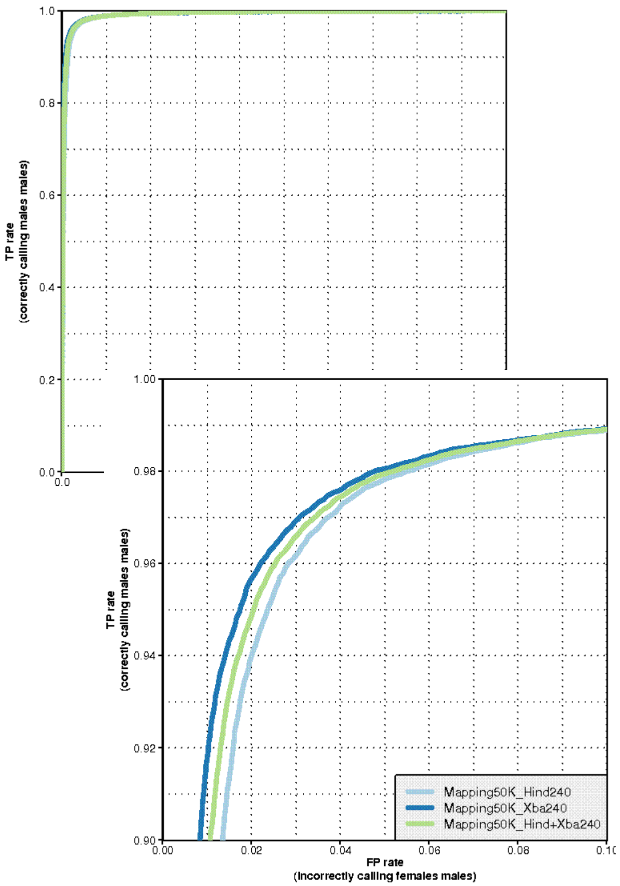
CRMA with and without CN probes

	CRMA (SNP)	***CRMA6 (SNP & CN)***
Preprocessing *(probe signals)*	allelic crosstalk (quantile)	allelic crosstalk (quantile)
Summarization *(locus signals θ)*	Total CN: $PM = PM_A + PM_B$ log-additive (PM-only) θ = "chip effects"	Averaging SNPs: $PM_A = \text{median}\{PM_A\}$ $PM_B = \text{median}\{PM_B\}$ Total CN: $PM = PM_A + PM_B$ $\theta = PM$
Post-processing	fragment-length (GC-content)	fragment-length (GC-content)
Raw total CNs	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$	$M_{ij} = \log_2(\theta_{ij}/\theta_{Rj})$

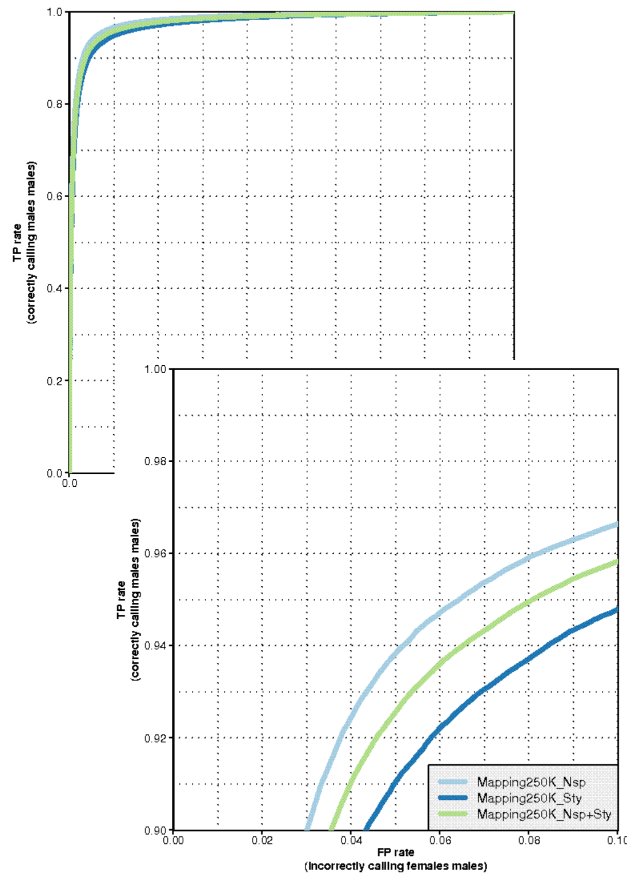
Comparison across generations

(100K - 500K - 6.0)

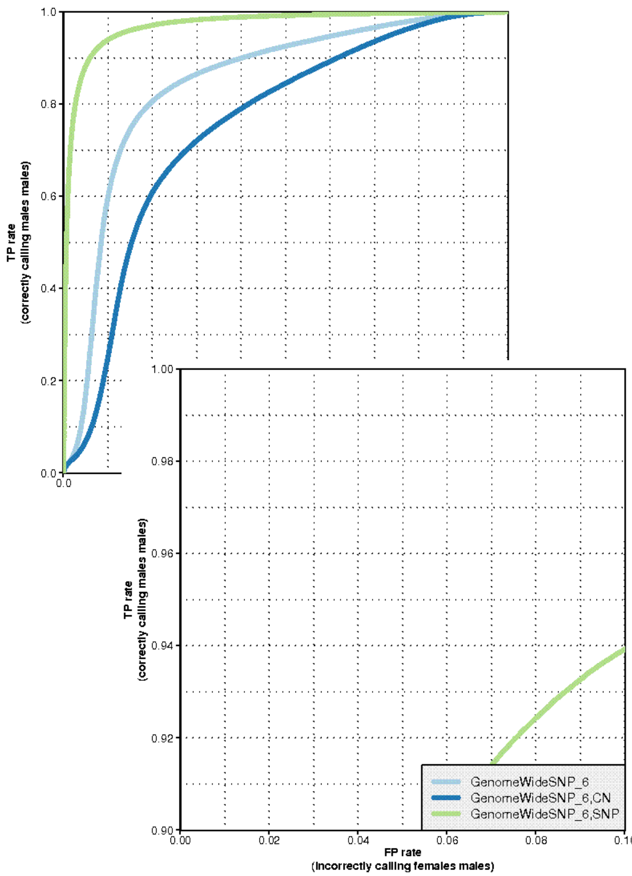
Average-locus ROC
100K → 500K → 6.0



100K:
Hind240, Xba240 & both

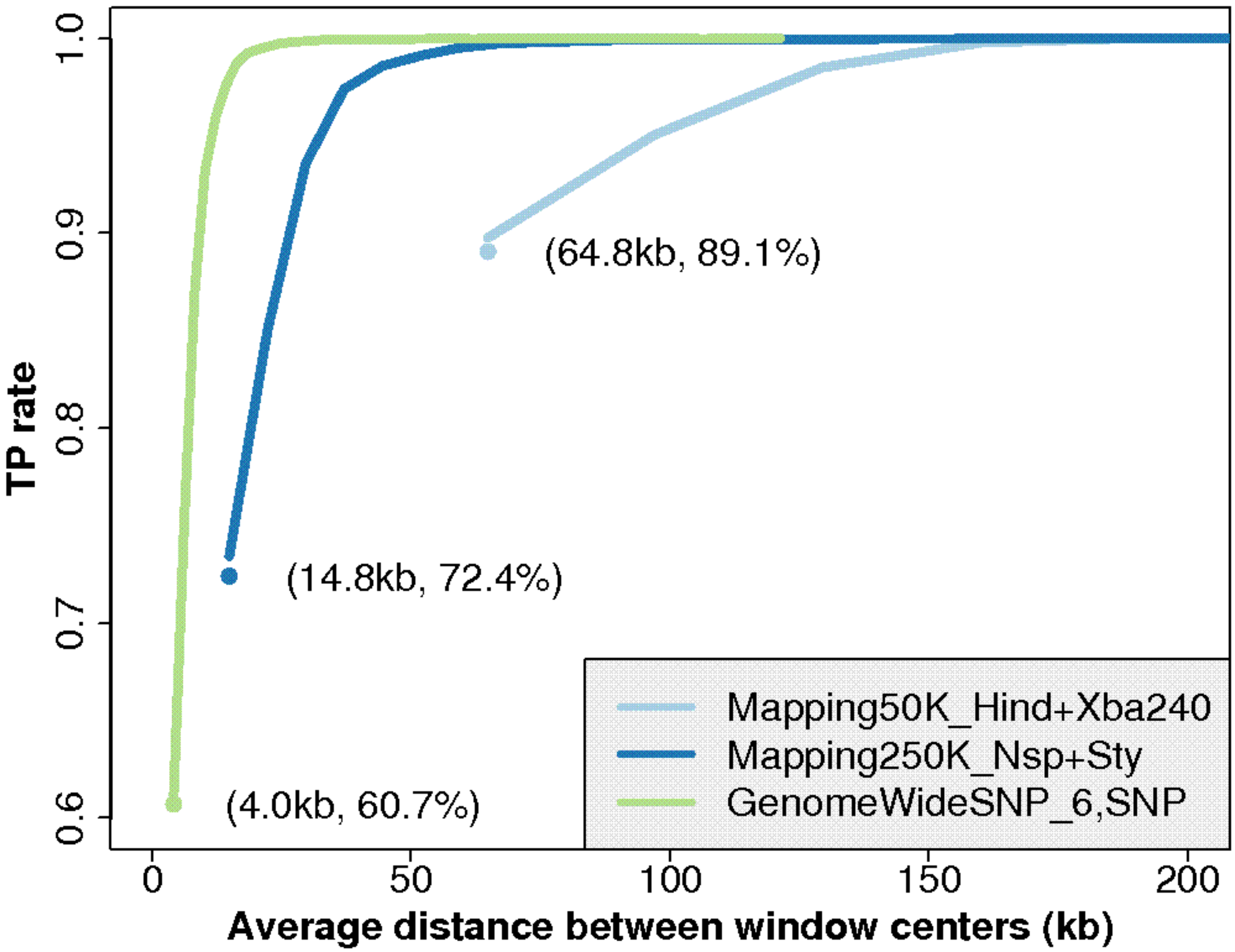


500K:
Nsp, Sty & both



6.0:
SNP, CN probes & all

Resolution comparison
- at 1.0% FP



Resolution comparison
- at 1.0% FP

At any given resolution (kb), we have:

$$TP_{6.0,SNP} > TP_{500K} > TP_{100K}$$

Note, the differences may be due to lab effects
(the HapMap 100K, 500K & 6.0 hybridization
were done in different years/labs).
In either case, the trend is in the right direction.

Summary

Conclusions

- It helps to:
  - Control for allelic crosstalk.
  - Sum alleles at PM level: $PM = PM_A + PM_B$.
  - Control for fragment-length effects.

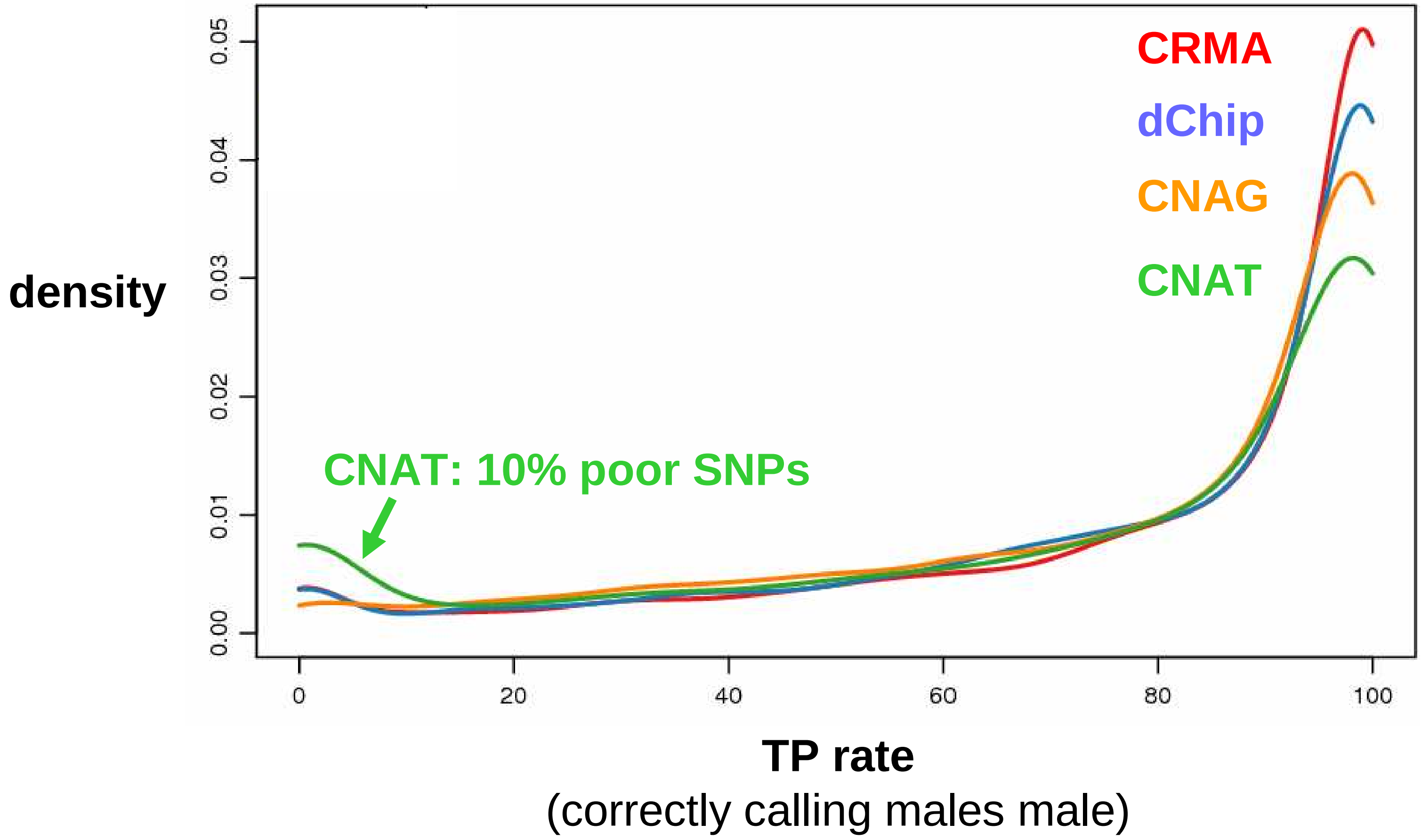
- Resolution: 6.0 (SNPs) > 500K > 100K (or lab effects).

- Currently estimates from CN probes are poor. Not unexpected. Better preprocessing might help.

Appendix

Density of TP rates when controlling for FP rate *(5,608 SNPs)*

FP rate: 1.0% (incorrectly calling females male)



Effect of different normalization steps