A Single-Array Preprocessing Method for Estimating Full-Resolution Raw Copy Numbers from all Affymetrix Genotyping Arrays

Henrik Bengtsson

(MSc CS, PhD Statistics)

Dept of Statistics, UC Berkeley

(joint work with Terry Speed & Pratyaksha Wirapati)

Comprehending Copy Number Variation (Tools, Applications and Results)

March 16, 2009, San Diego, CA

this-gen

A single-array CN method

Henrik Bengtsson

Dept of Statistics, UC Berkeley

(joint work with Terry Speed & Pratyaksha Wirapati)

Comprehending Copy Number Variation (Tools, Applications and Results)

March 16, 2009, San Diego, CA

Single-sample methods

There is a need for single-sample methods

World #1 – Large-scale projects:

- New platforms generate more data than previous generations.
- New studies involve more samples than even before.
- Data and knowledge is gathered incrementally over time.

World #2 – Personalized medicine:

- The era of personal diagnostics and treatment is around the corner.

Issues:

- Batch processing inconvenient / not possible.
- Data from one sample should not affect the result of another.

Our goal:

- Single-sample data processing.

Immediate and efficient processing with single-sample methods

Low latency:

- Arrays can be processed immediately after scanning.
- No need for reprocessing when new arrays arrive.
- Paired tumor-normal analysis requires only two hyb’s.

Scalable:

- Arrays can be processed in parallel on multiple hosts.
- Bounded memory (by definition).

Practical:

- In applied medical diagnostics individuals can be analyzed at once.

At UC Berkeley we have a few single-sample methods in place

1. **Single-array CN preprocessing**
   - improved total (and allele-specific) CN estimates from any Affymetrix SNP & CN chip type.

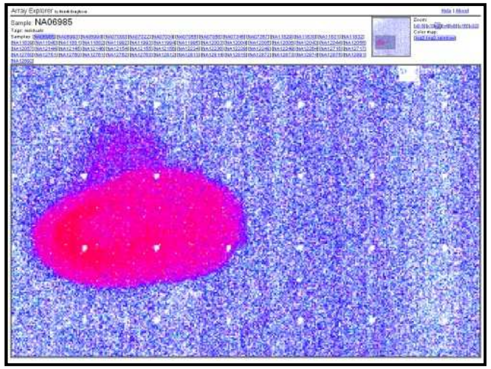
2. **Single-sample multi-platform CN normalization**
   - makes CN estimates from Affymetrix, Illumina, Agilent, qPCR, Solexa sequencing etc. comparable for downstream **integration**.
   - Facilitate **transition between technologies**.

3. **Single-sample calibration of allele-specific CNs**
   - much **cleaner ASCNs** from Affymetrix SNP chip types, maybe also Illumina (work in progress with Pierre Neuvial, UC Berkeley).

All of the above is done without using priors.

An open-source aroma.affymetrix framework for analyzing large Affymetrix data sets

- Processes **unlimited number of arrays**:
  - Bounded memory algorithms, e.g. RMA on ~5,000 HG-U133A arrays uses ~500MB of RAM.
  - Works toward file system.
  - Persistent memory: **robust** & picks up where last stopped.
- Supports most Affymetrix chip types and custom CDFs.



Array Explorer

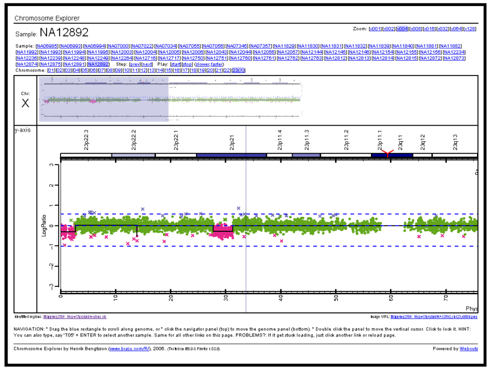
Help | About

Sample: NA06985

Zoom:

Color map:

- **Low-level analysis**: Background correction, allele crosstalk calibration, quantile normalization, nucleotide-position normalization etc. Most probe-summarization models. Post-processing: PCR fragment-length normalization, …
- **Copy-number analysis**, **alternative splicing,** and more.
- **Reproducibility.**



Chromosome Explorer

Sample: NA12892

Zoom: [x001][x002][x004][x008][x016][x032][x064][x128]

Sample: [NA06985] [NA06993] [NA06994] [NA07000] [NA07022] [NA07034] [NA07055] [NA07056] [NA07345] [NA07357] [NA11829] [NA11830] [NA11831] [NA11832] [NA11839] [NA11840] [NA11881] [NA11882] [NA11992] [NA11993] [NA11994] [NA11995] [NA12003] [NA12004] [NA12005] [NA12006] [NA12043] [NA12044] [NA12056] [NA12057] [NA12144] [NA12145] [NA12146] [NA12154] [NA12155] [NA12156] [NA12234] [NA12236] [NA12239] [NA12248] [NA12249] [NA12264] [NA12716] [NA12717] [NA12750] [NA12751] [NA12760] [NA12761] [NA12762] [NA12763] [NA12812] [NA12813] [NA12814] [NA12815] [NA12872] [NA12873] [NA12874] [NA12875] [NA12891] [NA12892] Step: [prev][next] Play: [start][stop] (slower,faster)

Chromosome: [01][02][03][04][05][06][07][08][09][10][11][12][13][14][15][16][17][18][19][20][21][22][23(X)]

Identified regions: Mapping250K_Nsp+Sty,glad/regions.xls

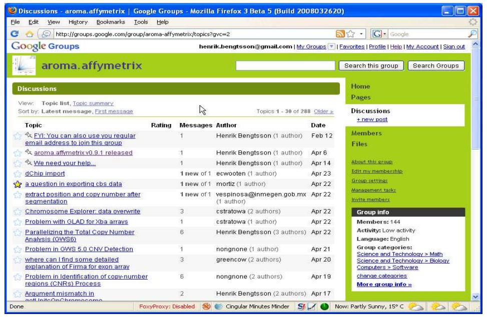
Image URL: Mapping250K_Nsp+Sty,glad/NA12892,chr23,x0004.png

NAVIGATION: * Drag the blue rectangle to scroll along genome, or * click the navigator panel (top) to move the genome panel (bottom). * Double click the panel to move the vertical cursor. Click to lock it. HINT: You can also type, say 'T05' + ENTER to select another sample. Same for all other links on this page. PROBLEMS?: If it get stuck loading, just click another link or reload page.

Chromosome Explorer by Henrik Bengtsson (www.braju.com/R/), 2006. (Tested on IE6.0 & Firefox 1.5/2.0)

Powered by Webcuts

- **Cross platform R package**: Linux/Unix, Windows, OSX.
- Large number of component and **redundancy tests**.
- **Open source** and online **user forum.**



Discussions - aroma.affymetrix | Google Groups - Mozilla Firefox 3 Beta 5 (Build 2008032620)

http://groups.google.com/group/aroma-affymetrix/topics?gvc=2

Google Groups

henrik.bengtsson@gmail.com | My Groups | Favorites | Profile | Help | My Account | Sign out

aroma.affymetrix

Discussions

View: Topic list, Topic summary

Sort by: Latest message, First message

Topics 1 - 30 of 288 Older »

Topic	Rating	Messages	Author	Date
FYI: You can also use you regular email address to join this group		1	Henrik Bengtsson (1 author)	Feb 12
aroma.affymetrix v0.9.1 released		1	Henrik Bengtsson (1 author)	Apr 6
We need your help...		1	Henrik Bengtsson (1 author)	Apr 14
dChip import		1 new of 1	ecwooten (1 author)	Apr 23
a question in exporting cbs data		1 new of 1	mortiz (1 author)	Apr 22
extract position and copy number after segmentation		1 new of 1	vespinosa@inmegen.gob.mx (1 author)	Apr 22
Chromosome Explorer: data overwrite		3	cstratowa (2 authors)	Apr 22
Problem with GLAD for Xba arrays		1	cstratowa (1 author)	Apr 22
Parallelizing the Total Copy Number Analysis (GWS6)		6	Henrik Bengtsson (3 authors)	Apr 22
Problem in GWS 5.0 CNV Detection		1	nongnone (1 author)	Apr 21
where can I find some detailed explanation of Firma for exon array		3	greencow (2 authors)	Apr 20
Problem in Identification of copy-number regions (CNRs) Process		6	nongnone (2 authors)	Apr 19
Argument mismatch in getUnitsOnChromosome		2	Henrik Bengtsson (2 authors)	Apr 17

Home

Pages

Discussions

+ new post

Members

Files

About this group

Edit my membership

Group settings

Management tasks

Invite members

Group info

Members: 144

Activity: Low activity

Language: English

Group categories:
Science and Technology > Math
Science and Technology > Biology
Computers > Software

change categories

More group info »

Done | FoxyProxy: Disabled | Cingular Minutes Minder | Now: Partly Sunny, 15° C

CRMA v2

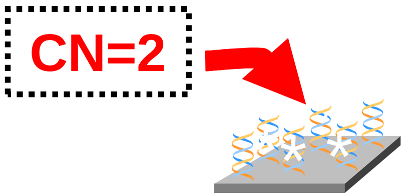
-

Single-array CN preprocessing

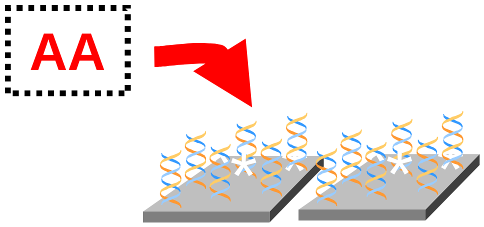
(all Affymetrix SNP & CN chips)

Affymetrix CN & SNP probes are used to quantify the amount of DNA at known loci

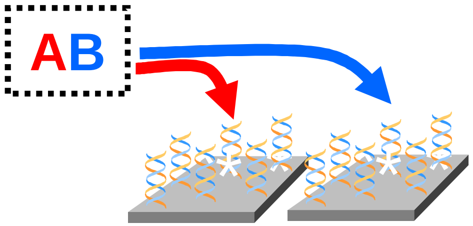
Copy neutral (CN = 2)


CN=2

PM = 2 * const


AA

$PM = PM_A + PM_B$
= 2 * const

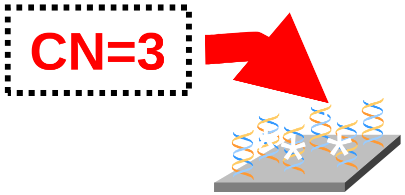

AB

$PM = PM_A + PM_B$
= 2 * const

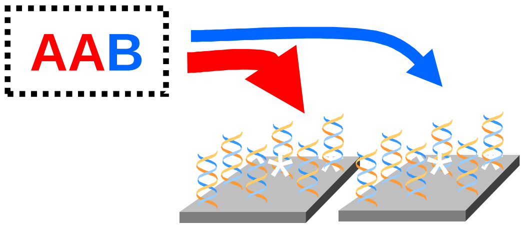
Non-polymorphic CN probes

SNP probe pairs

Amplification (e.g. CN=3)


CN=3

PM = 3 * const


AAB

$PM = PM_A + PM_B = 3$ * const

CRMA v2

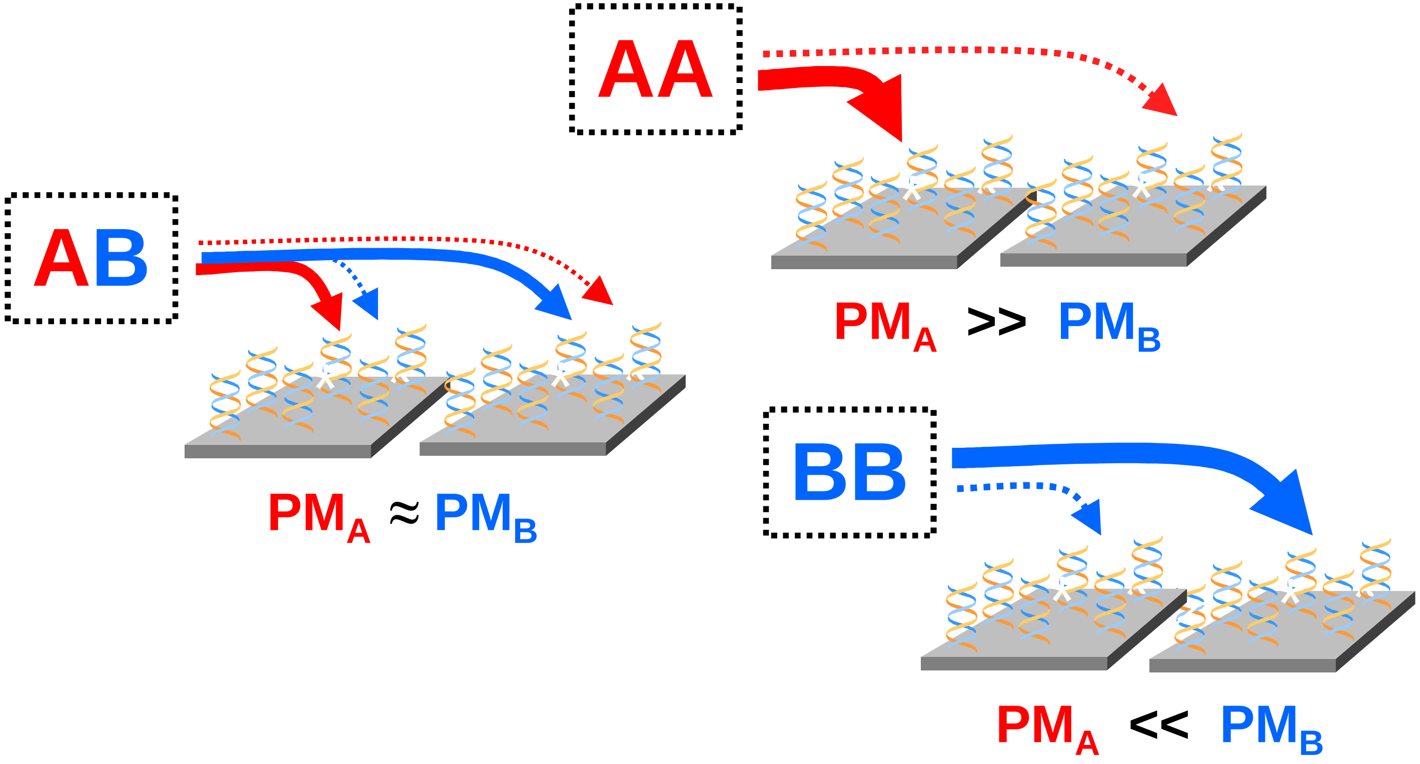
Probe signals	Allele-crosstalk calibration Probe-sequence normalization
Summarization *(allele-specific* *or total)*	Robust averaging $\theta_{ijjA} = \text{median}_k(PM_{ijAk})$ $\theta_{ijjB} = \text{median}_k(PM_{ijBk})$ $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$ array i, loci j, probe k
Summaries	PCR fragment-length normalization
Relative CNs	Log ratios $M_{ij} = \log_2(\theta_{ij} / \theta_{Rj})$ reference R

Crosstalk between alleles
- adds significant artifacts to signals

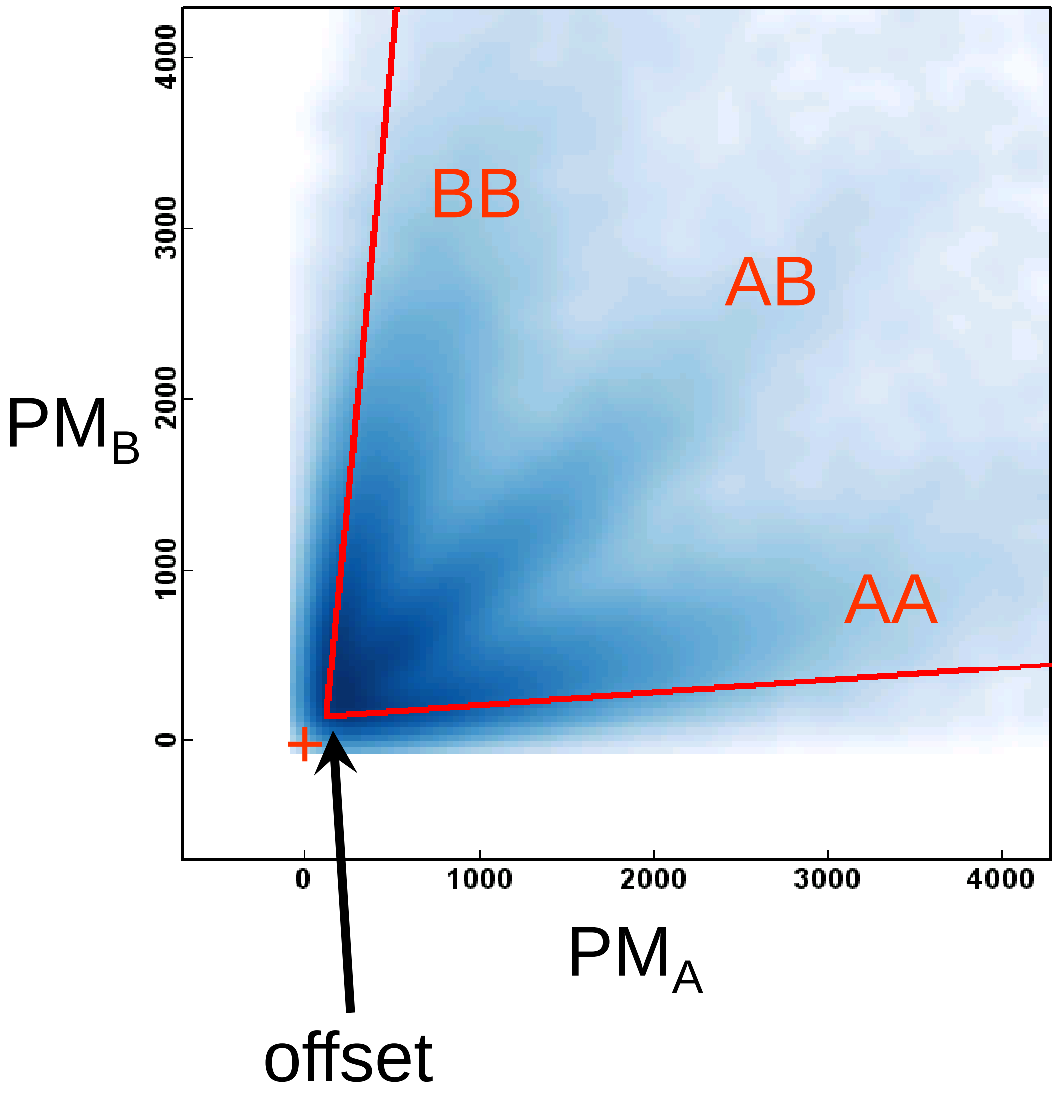
Cross-hybridization:

Allele A: ***TCGGTAAGTACTC***

Allele B: ***TCGGTATGTACTC***

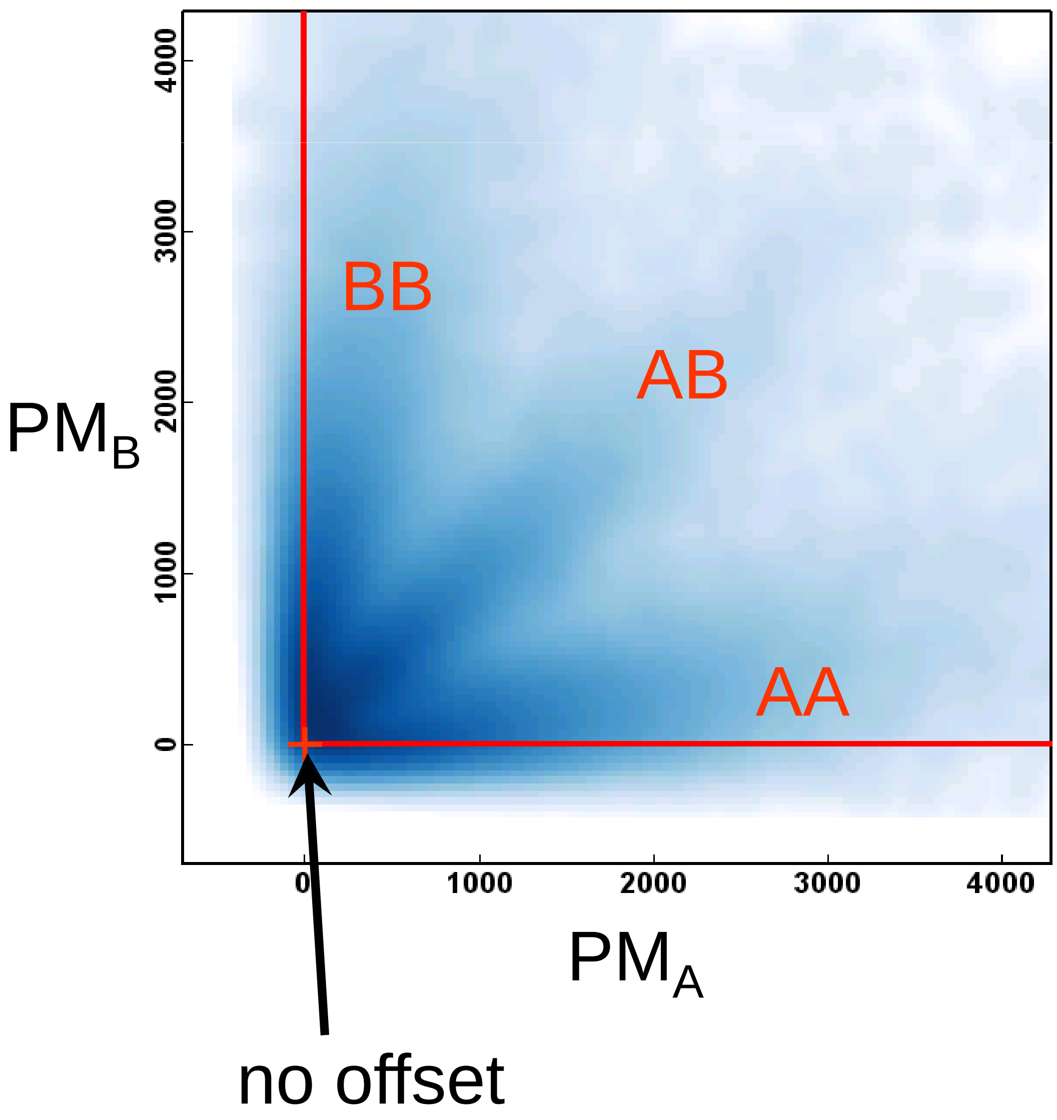


Crosstalk between alleles is easy to spot



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

Crosstalk between alleles and offset can be estimated and corrected for

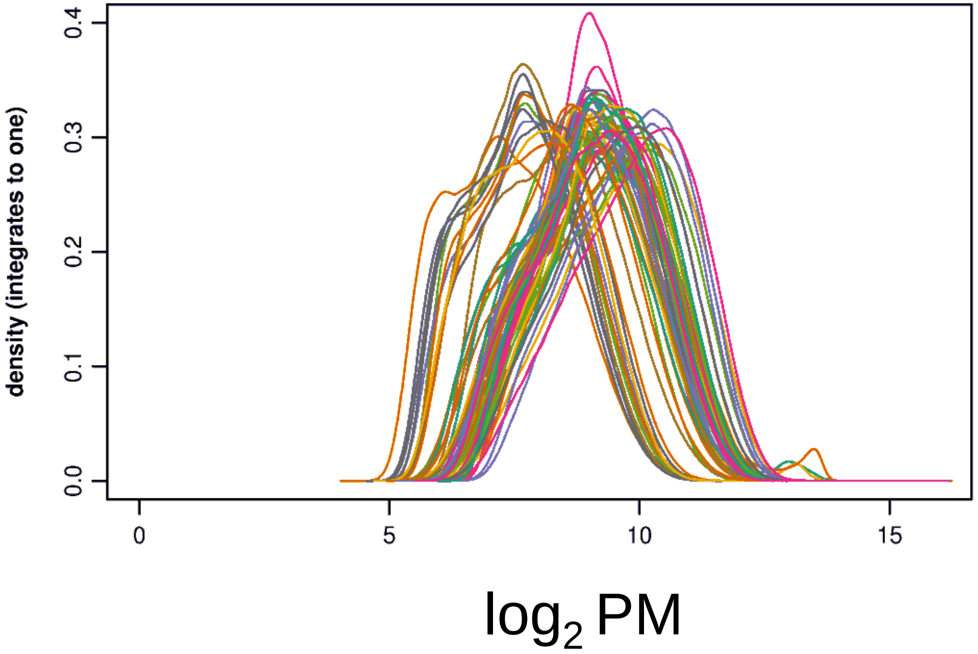


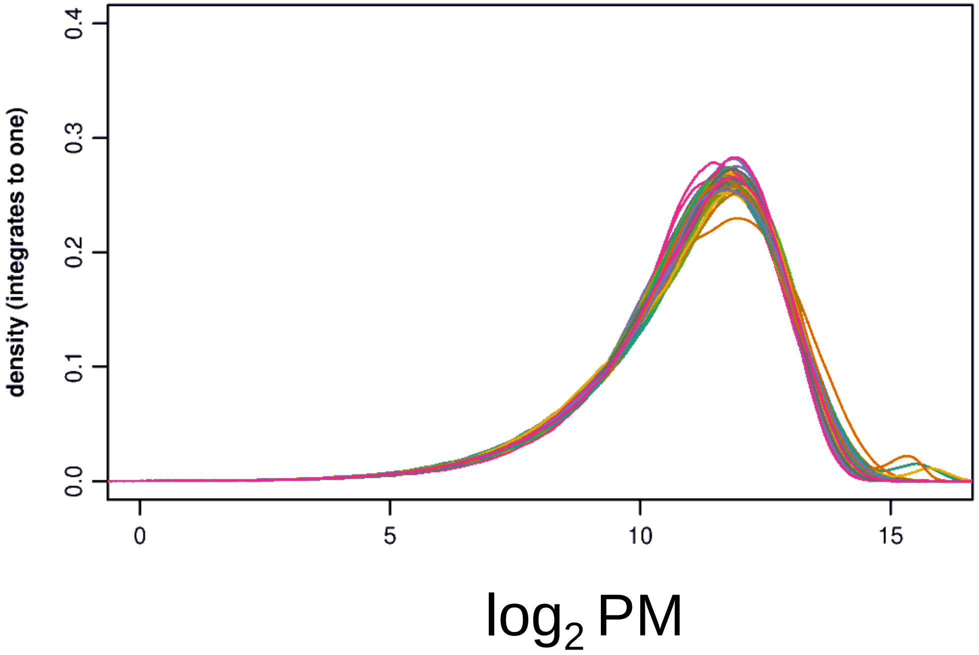
What is done:
1. Offset is removed from SNPs and CN units.
2. Crosstalk is removed from SNPs.

Crosstalk calibration corrects for differences in signal distributions too

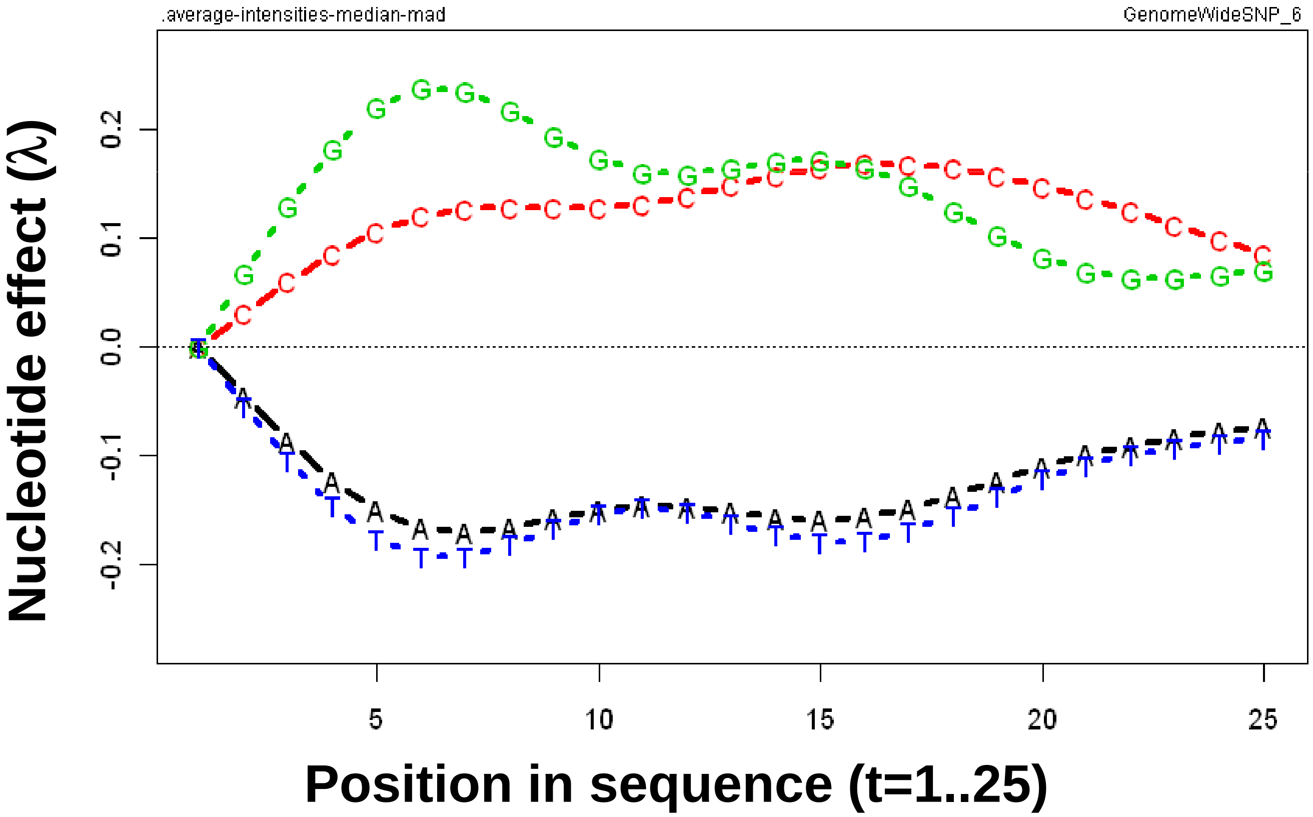
Before removing crosstalk the arrays differ significantly...

...when removing offset & crosstalk differences goes away.





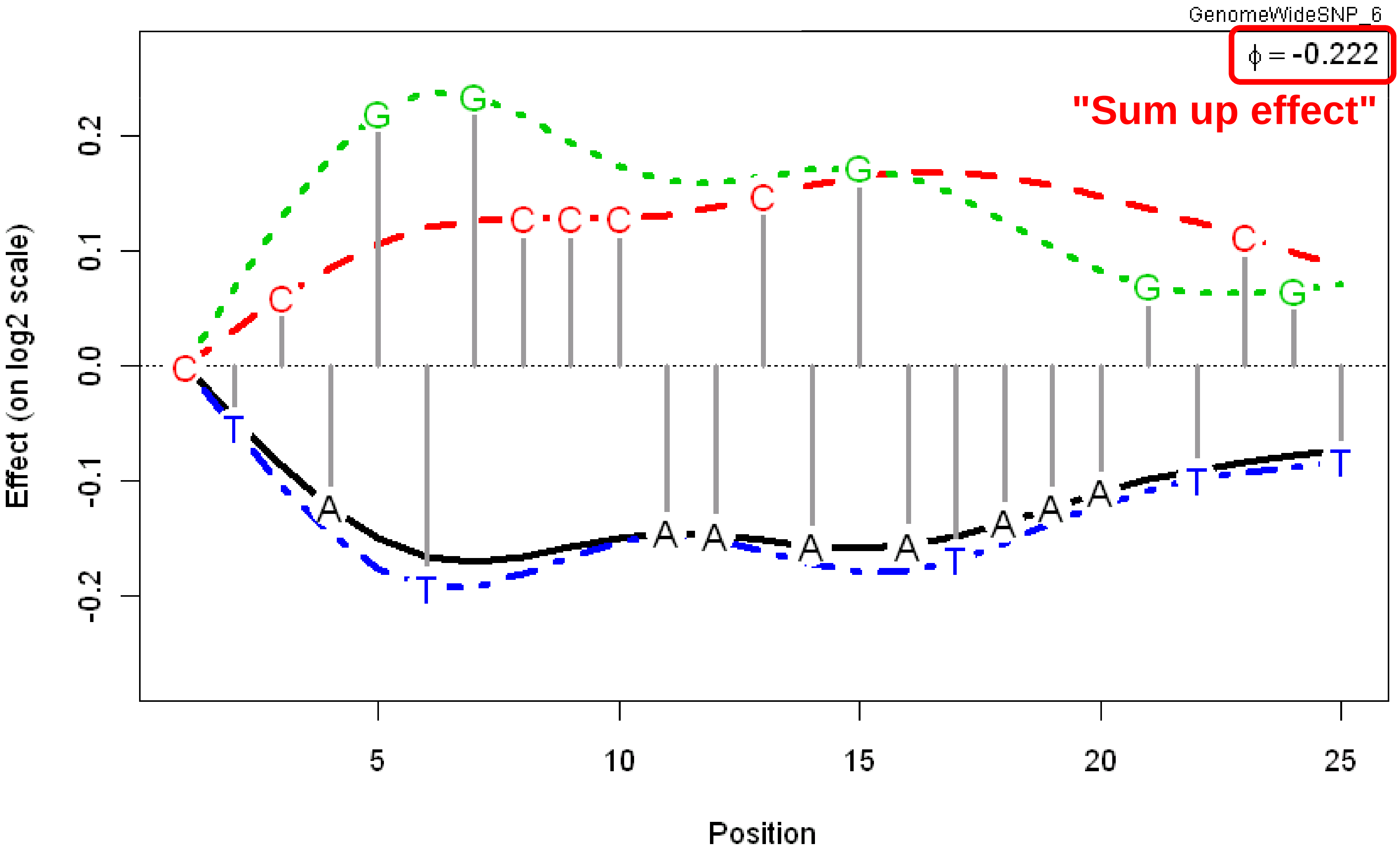
The nucleotide (A, C, G or T) and its position in the probe adds to the affinity



Probe-position affinity for probe k:

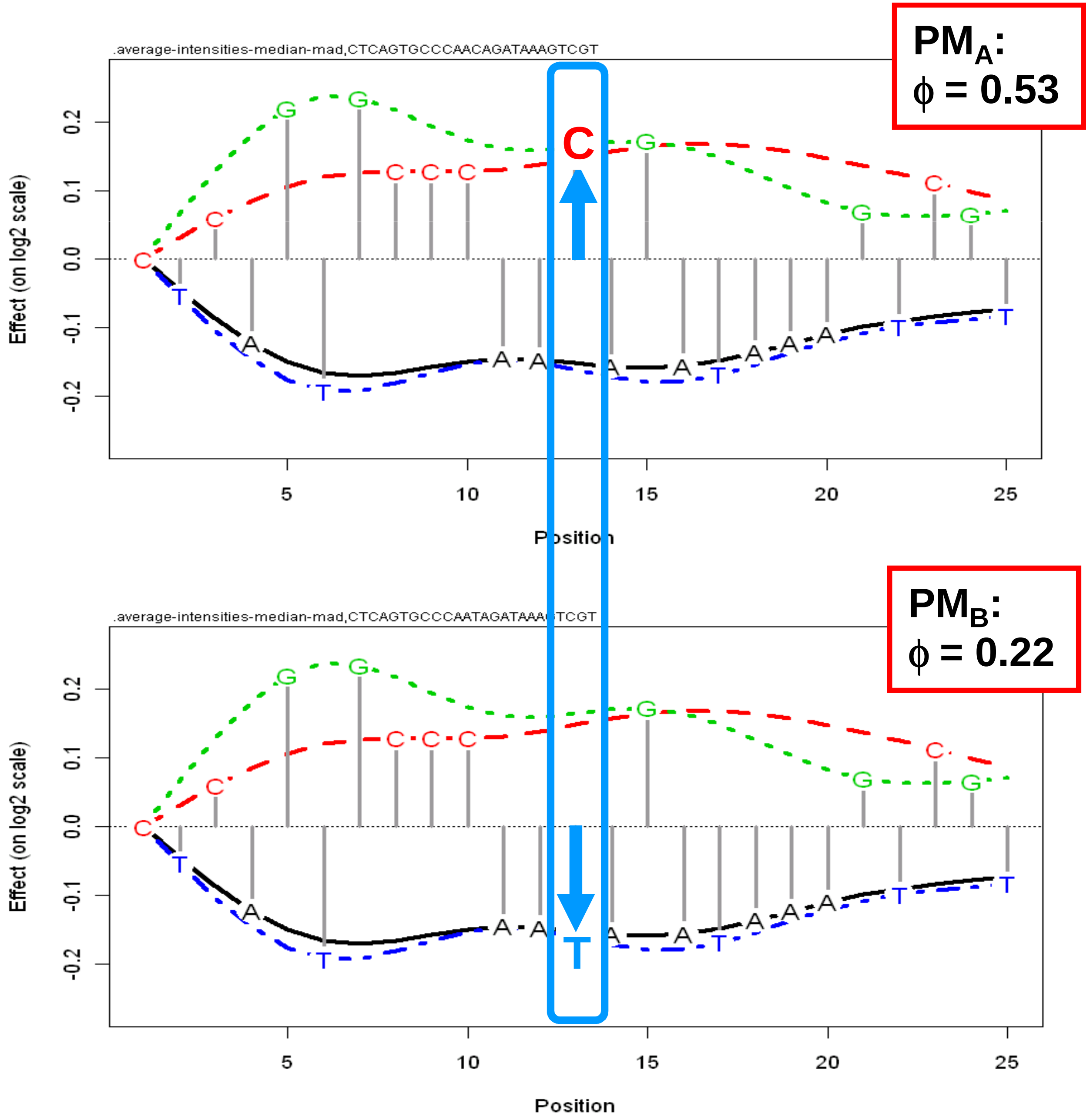
$$\phi_k = \phi((b_{k,1},b_{k,2},\ldots,b_{k,25})) = \sum_{t=1..25} \left[\sum_{b=\mathbf{A},\mathbf{C},\mathbf{G},\mathbf{T}} \mathbb{I}(b_{k,t}=b)\, \lambda_{b,t} \right]$$

Example: Probe-position affinity for CTCAGTGCCCAACAGATAAAGTCGT



"Sum up effect"

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



PM_A:
$\phi = 0.53$

PM_B:
$\phi = 0.22$

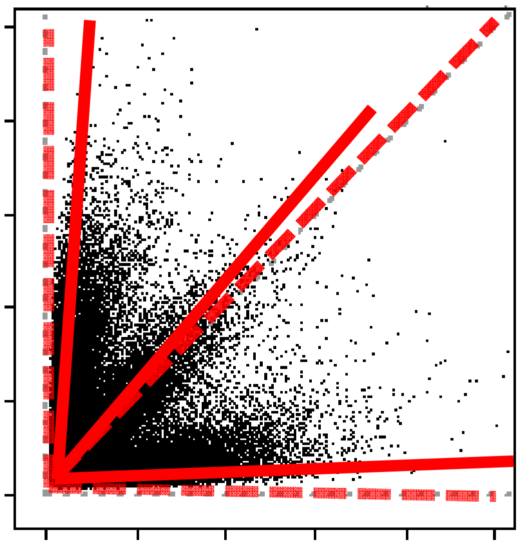
Genotypic imbalances:

A/B: nucleotides C/T

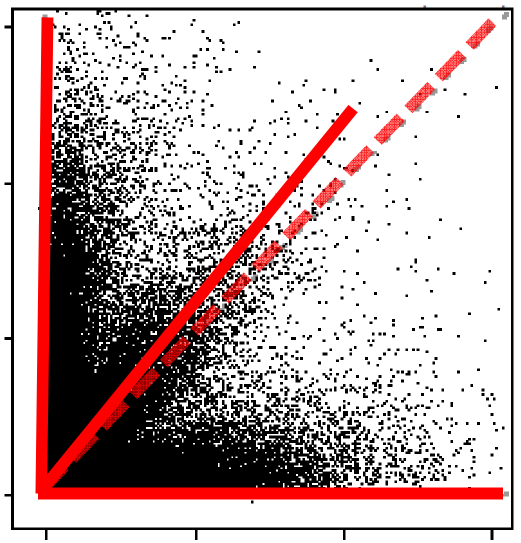
$PM=PM_A+PM_B$:
CC: 0.53+0.53 = 1.06
CT: 0.53+0.22 = 0.75
TT: 0.22+0.22 = 0.44

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

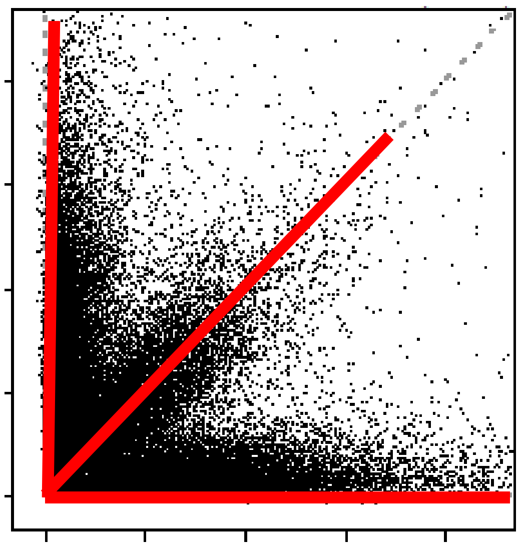
Allele-crosstalk calibration & nucleotide-position normalization work together



(raw)



ACC



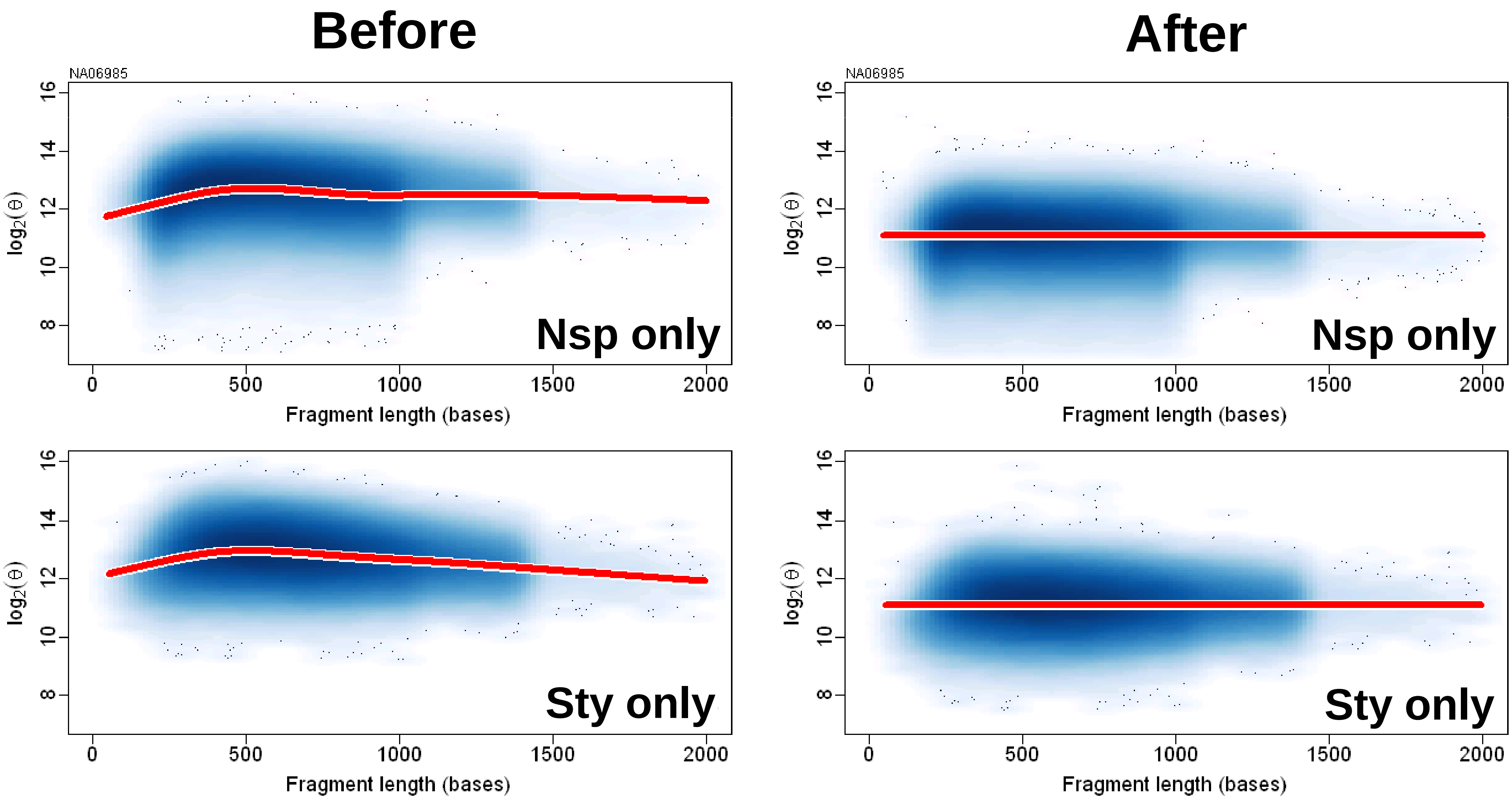
ACC + NPN

Probe summarization

- CN units: All single-probe units:
  - Non-polymorphic signal: $\theta_{ij} = PM_{ij1}$
- SNPs: Identically replicated probe pairs:
  - Probe pairs: (PM_{ijAk},PM_{ijBk}); k=1,2,3
  - Allele-specific signals: $\theta_{ijA} = \text{median}_k\{PM_{ijAk}\}$, $\theta_{ijB} = \text{median}_k\{PM_{ijBk}\}$
  - Non-polymorphic signal: $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$

Fragment-length effects

- Multi-enzyme normalization removes them



Before

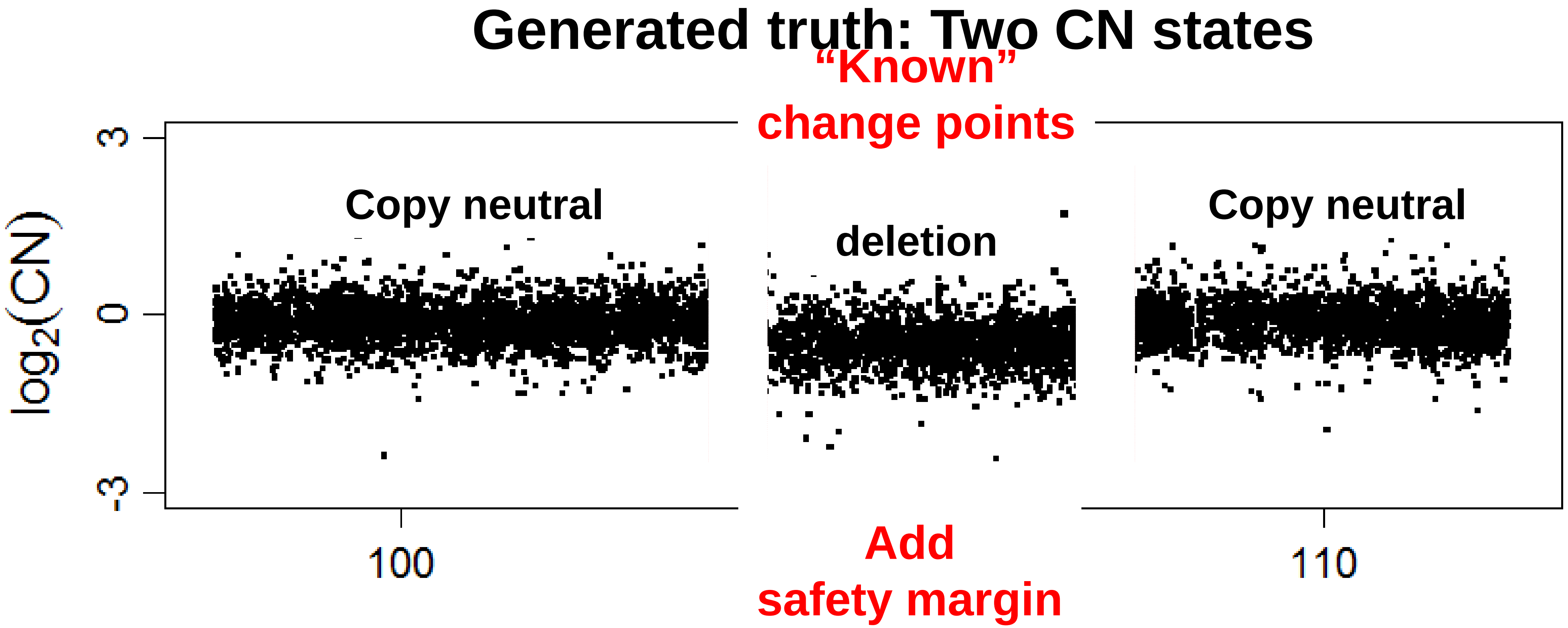
After

CRMA v2

Probe signals	Allele-crosstalk calibration Probe-sequence normalization
Summarization *(allele-specific* *or total)*	Robust averaging: $\theta_{ijjA} = \text{median}_k(PM_{ijAk})$ $\theta_{ijjB} = \text{median}_k(PM_{ijBk})$ $\theta_{ij} = \theta_{ijA} + \theta_{ijB}$ array i, loci j, probe k
Summaries	PCR fragment-length normalization
Relative CNs	Log ratios: $M_{ij} = \log_2(\theta_{ij} / \theta_{Rj})$ reference R

Evaluation

Idea: How well can be detect a known CN aberration?

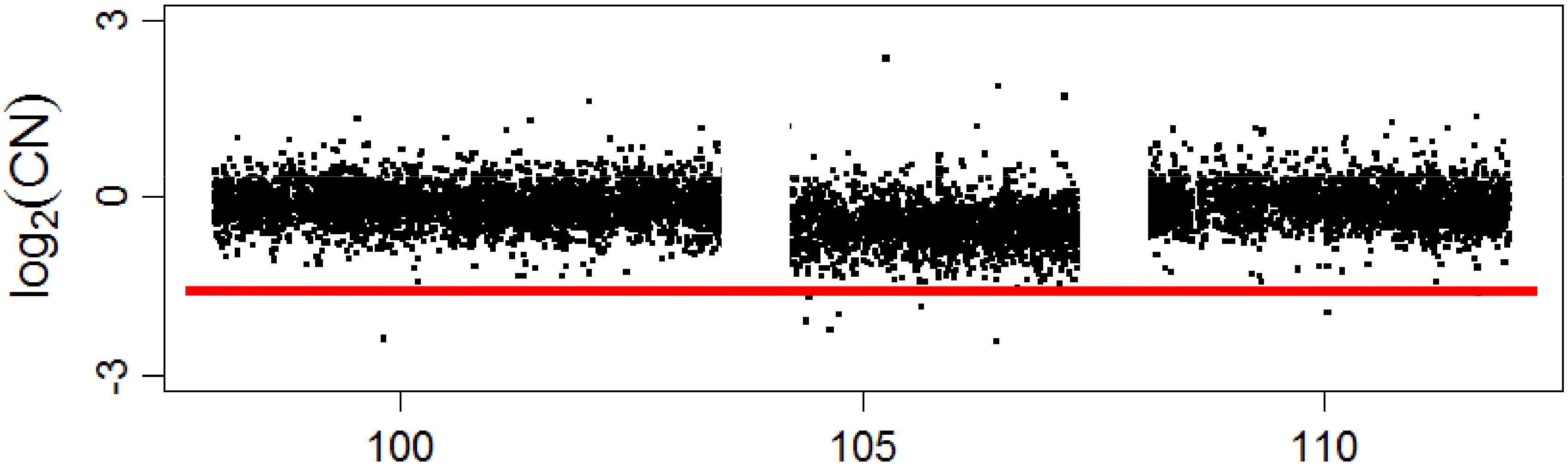


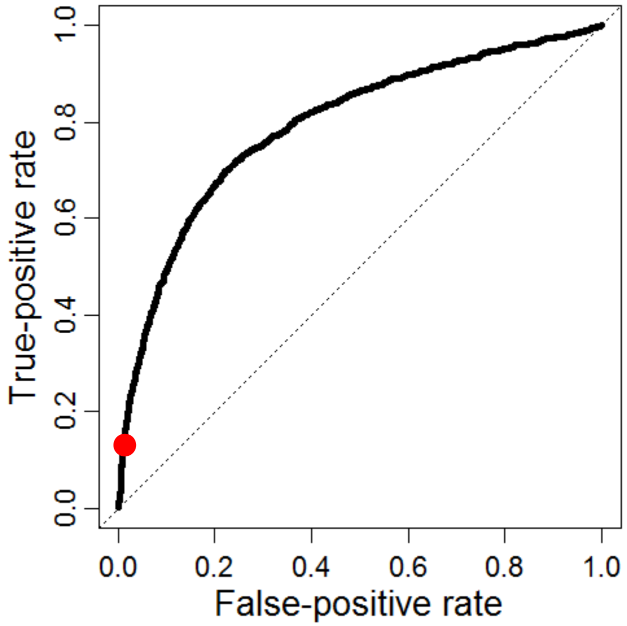
Example: 3.9 Mb deletion on Chromosome 1 in tumor GSM337641.

Data set: Chiang et al. High-resolution mapping of copy-number alterations
with massively parallel sequencing. Nature Methods, 2009.

How well does the two states separate?

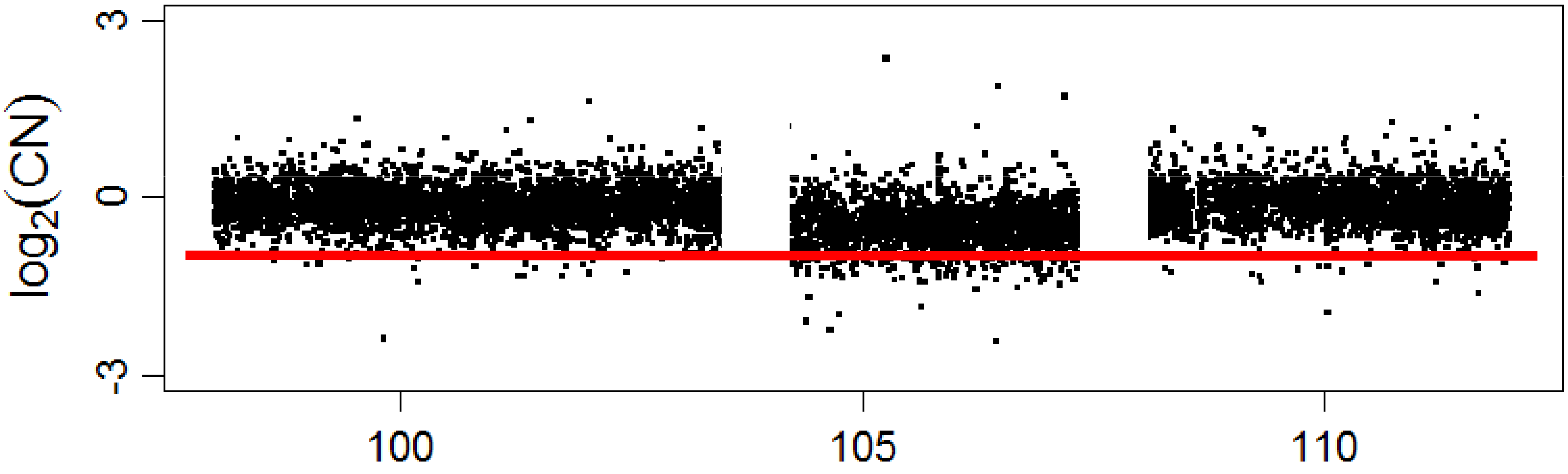
Calling CN deletion with common threshold

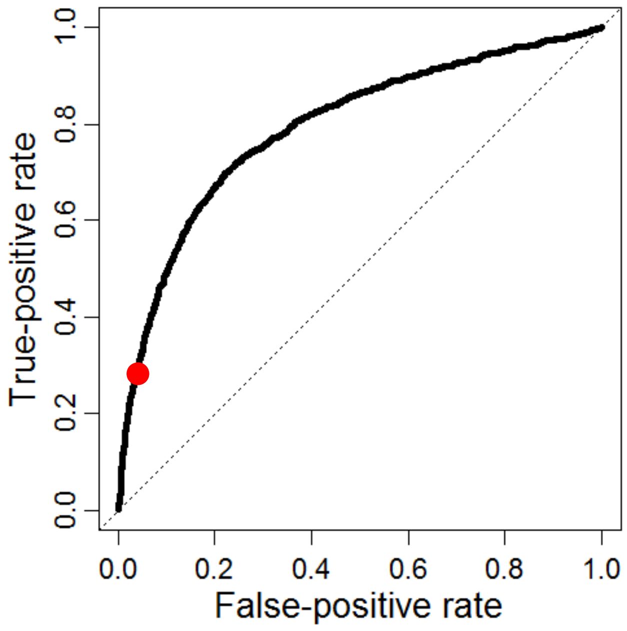




How well does the two states separate?

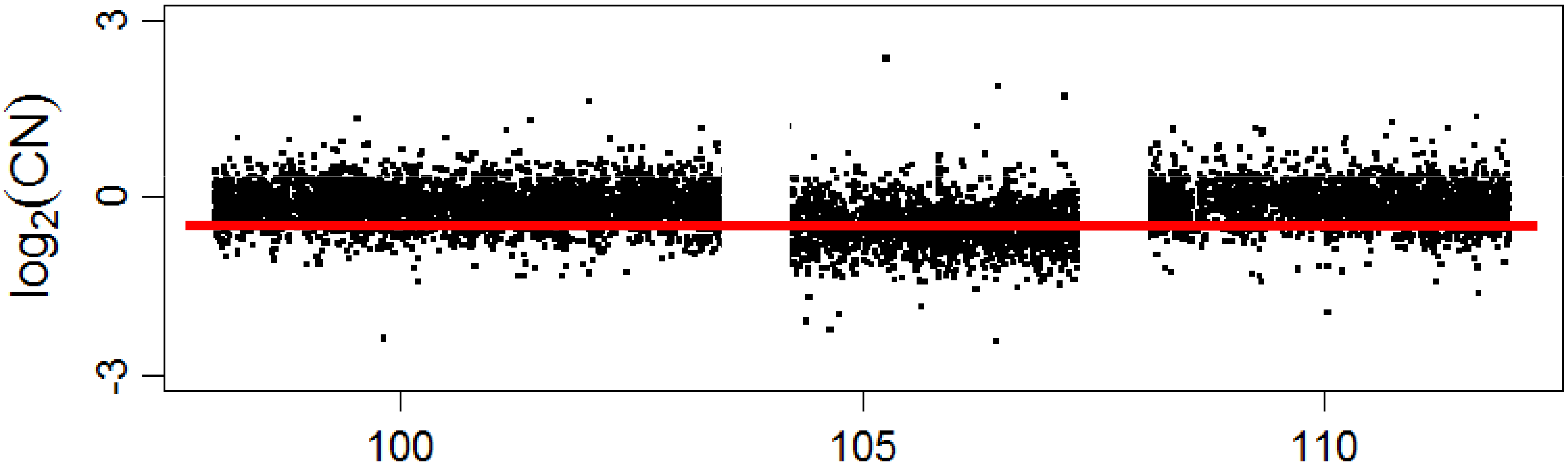
Calling CN deletion with common threshold

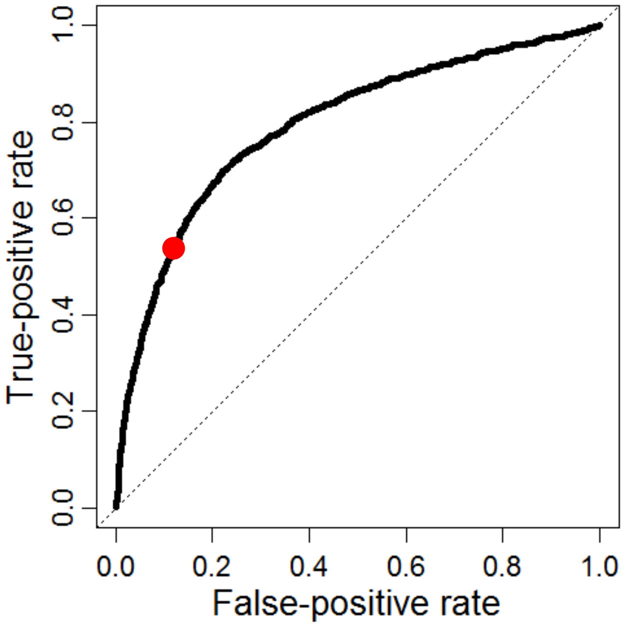




How well does the two states separate?

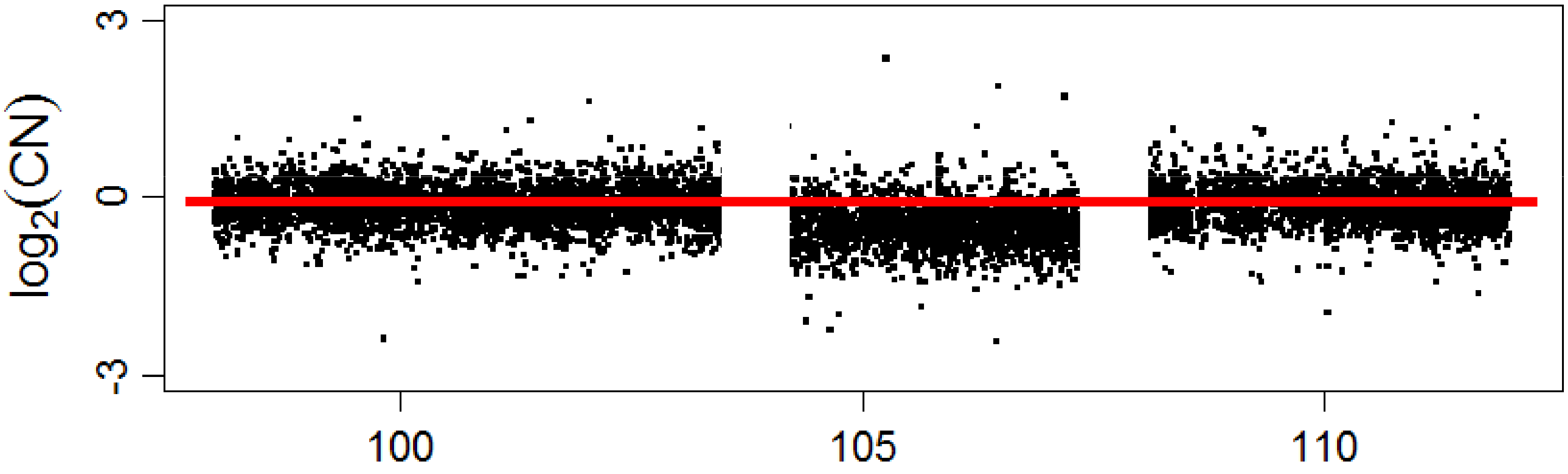
Calling CN deletion with common threshold

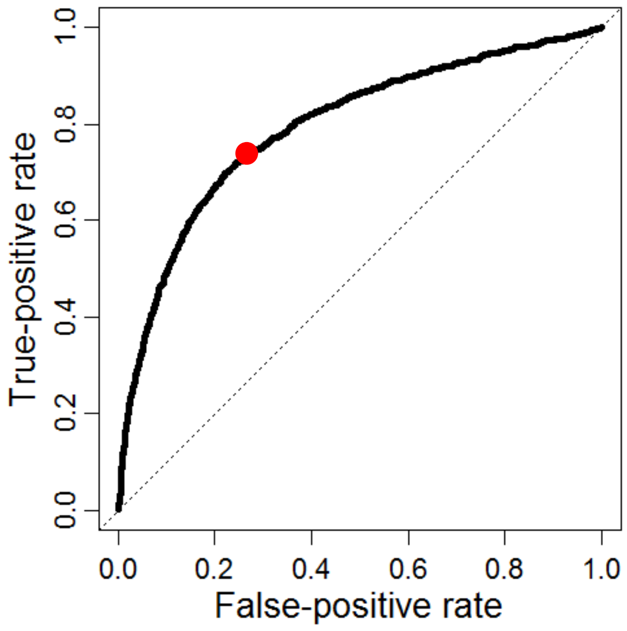




How well does the two states separate?

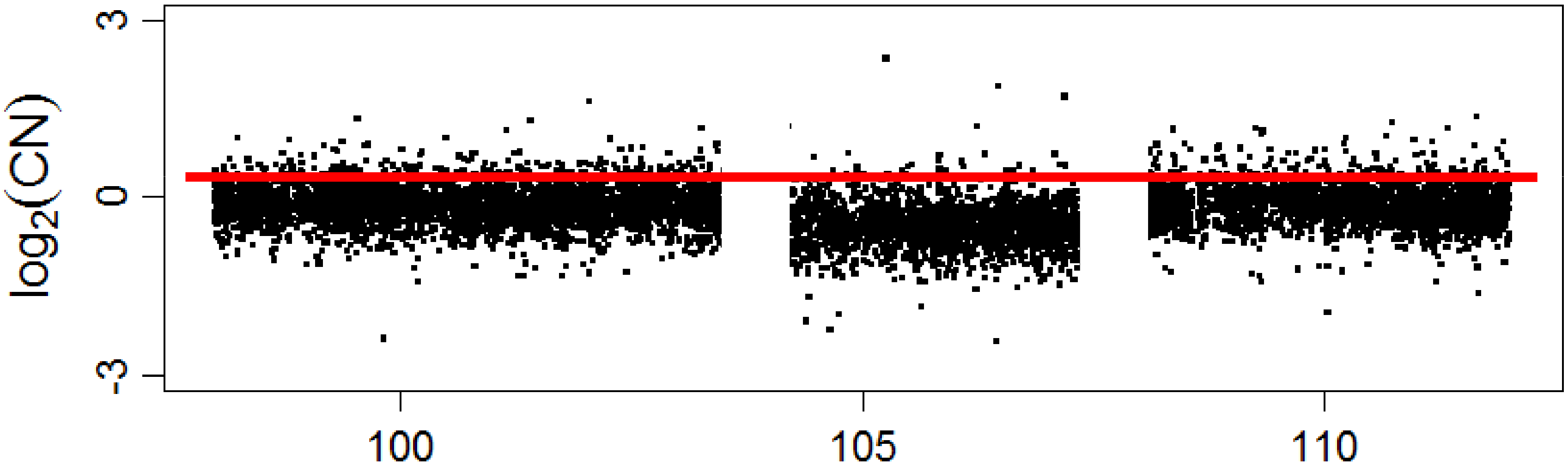
Calling CN deletion with common threshold

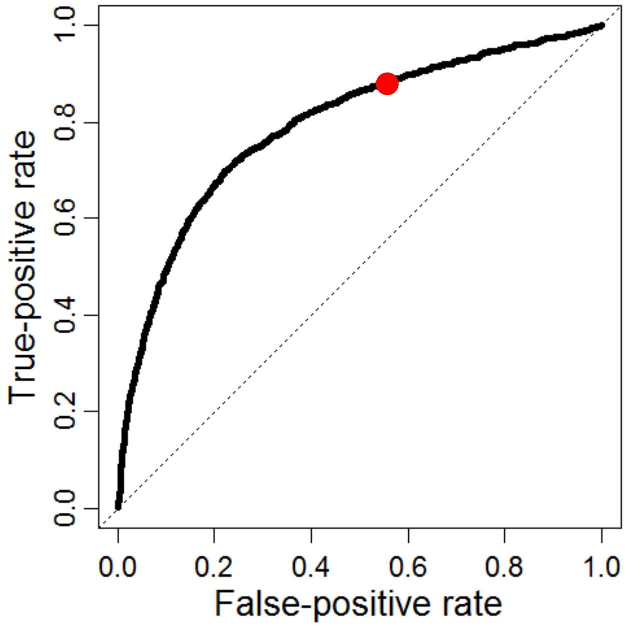




How well does the two states separate?

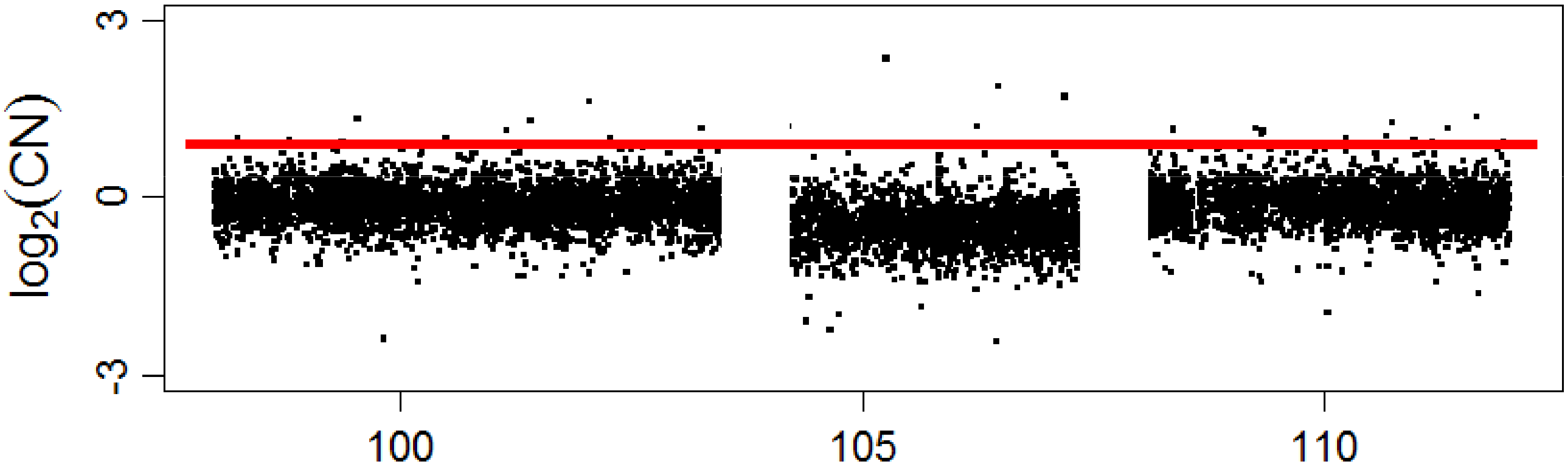
Calling CN deletion with common threshold

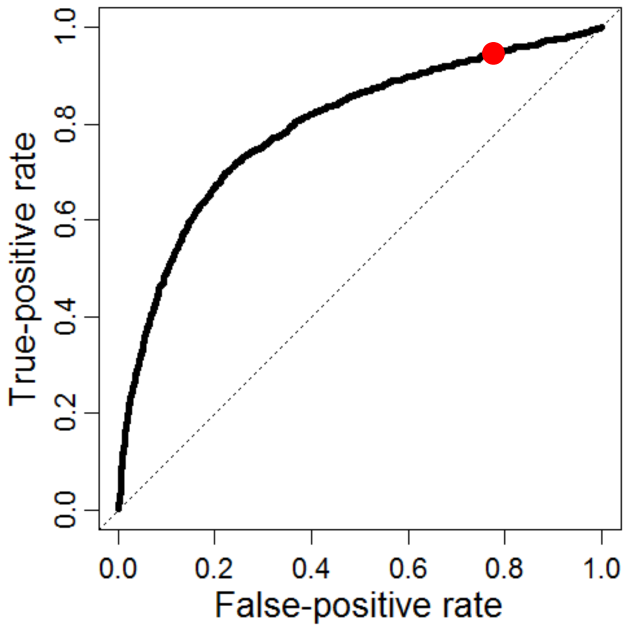




How well does the two states separate?

Calling CN deletion with common threshold





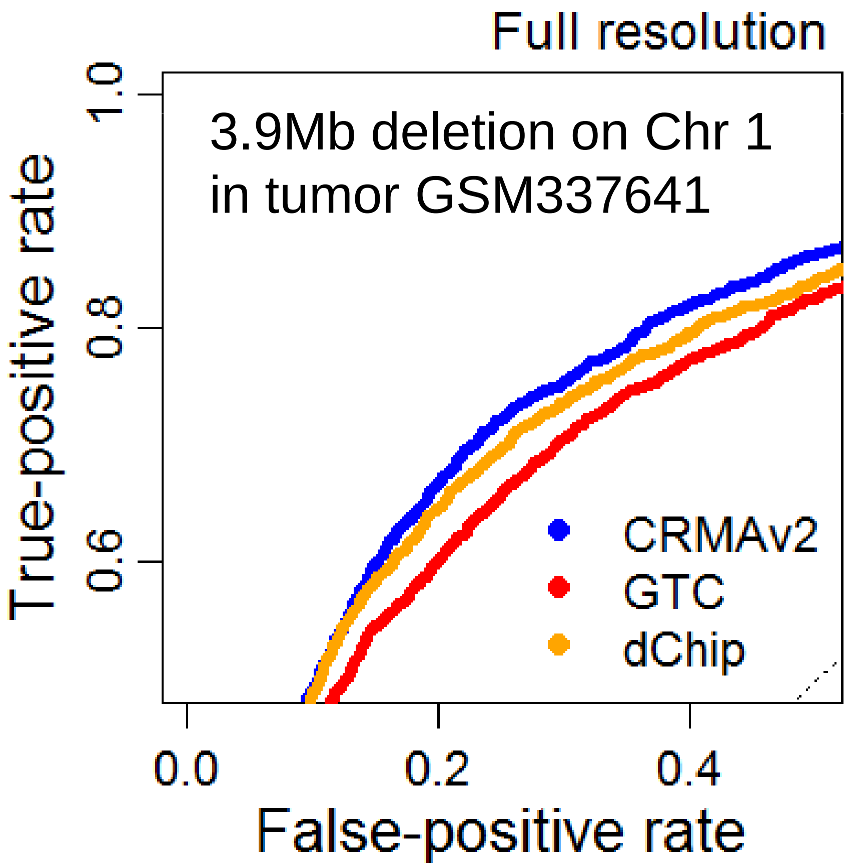
Single-array CRMAv2 performs well compared with Affymetrix GTC and dChip

Data set:

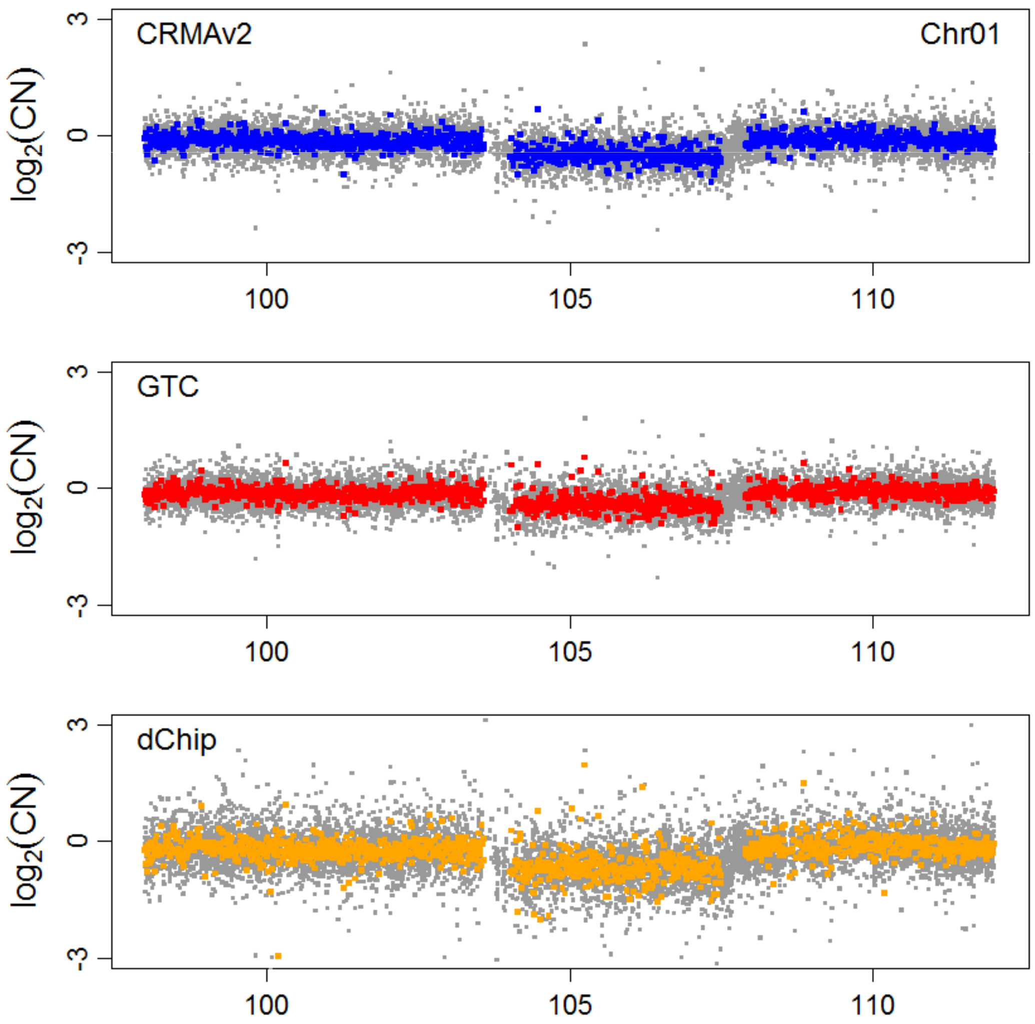
- Tumor-normal pairs.
- 68 hybridizations.
- GenomeWideSNP_6.
- Broad Institute, Chiang et al. (2009)

Preprocessing:

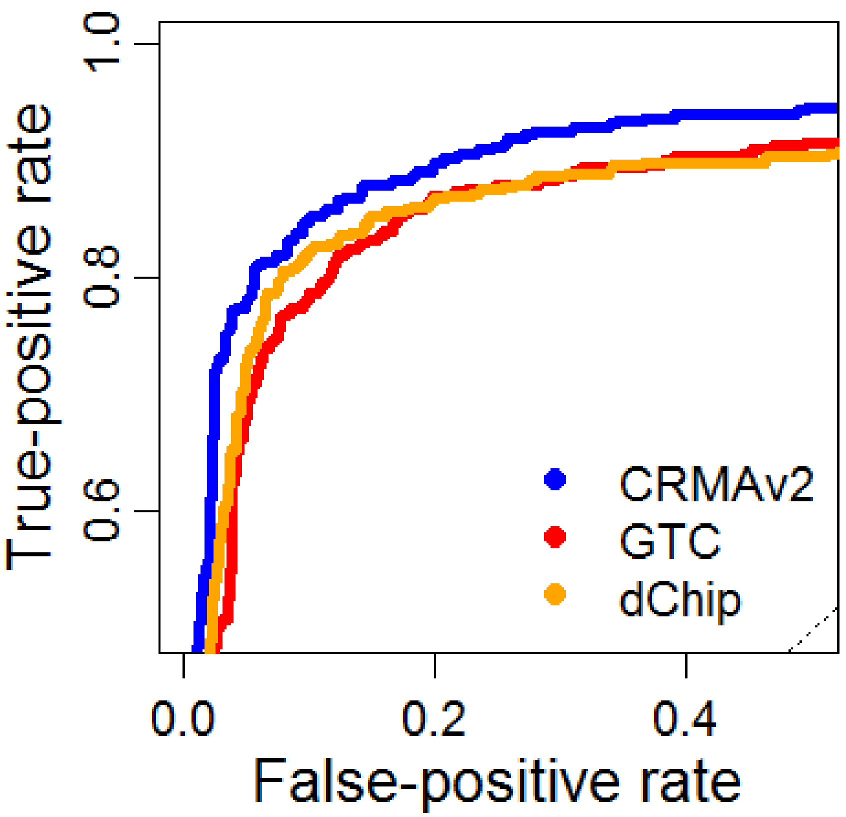
- GTC (CN5) and dChip were allowed to use all 68 arrays in their processing.



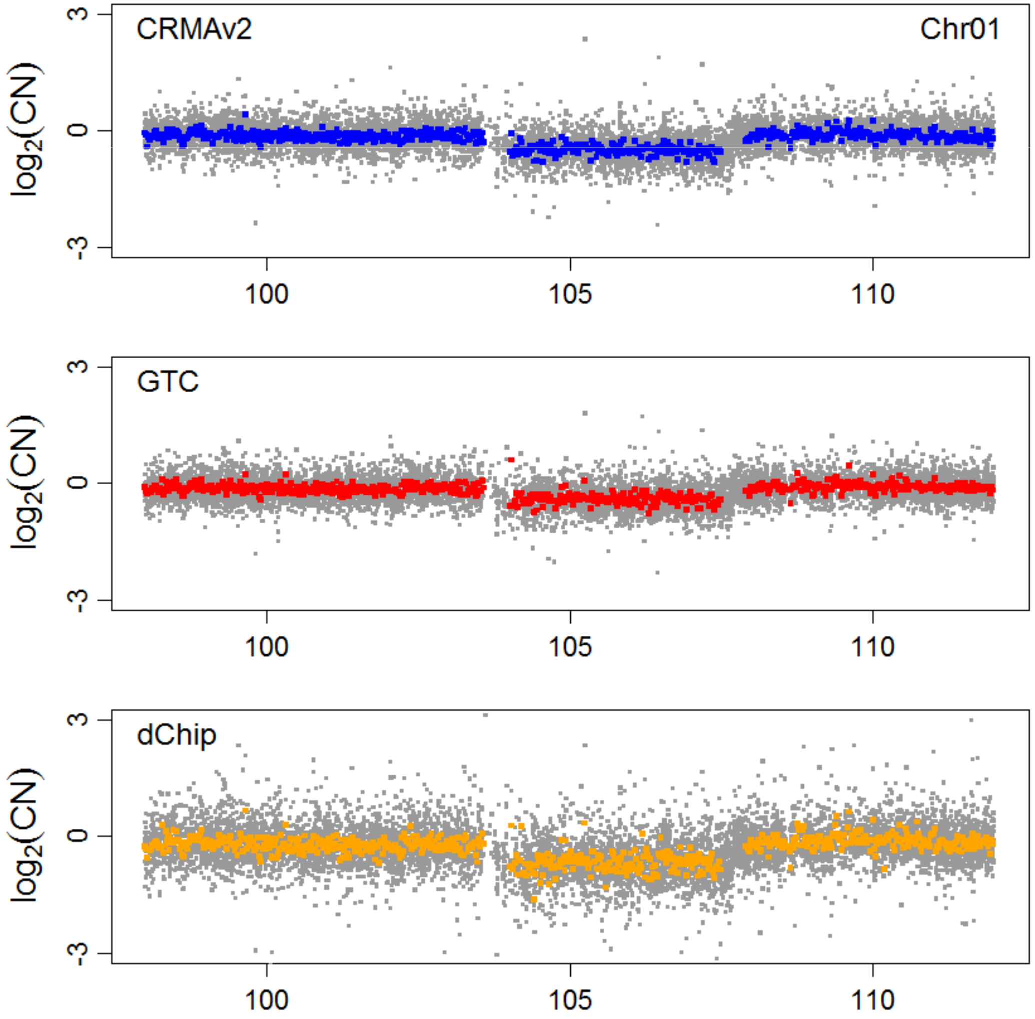
CRMAv2 performs well also at various amount of smoothing (“resolution”)



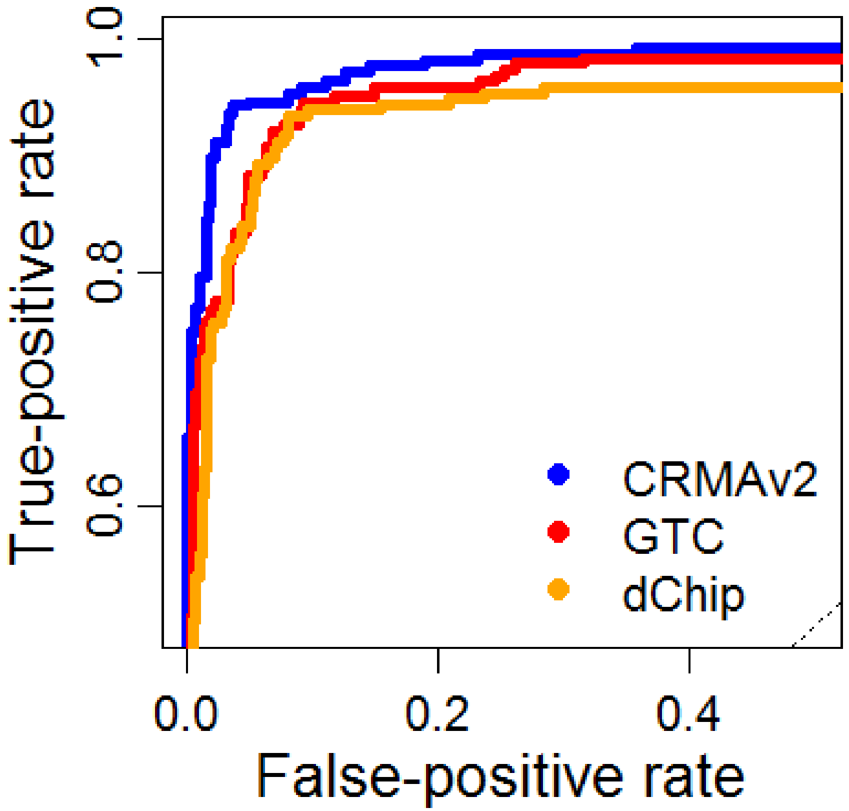
10kb binning



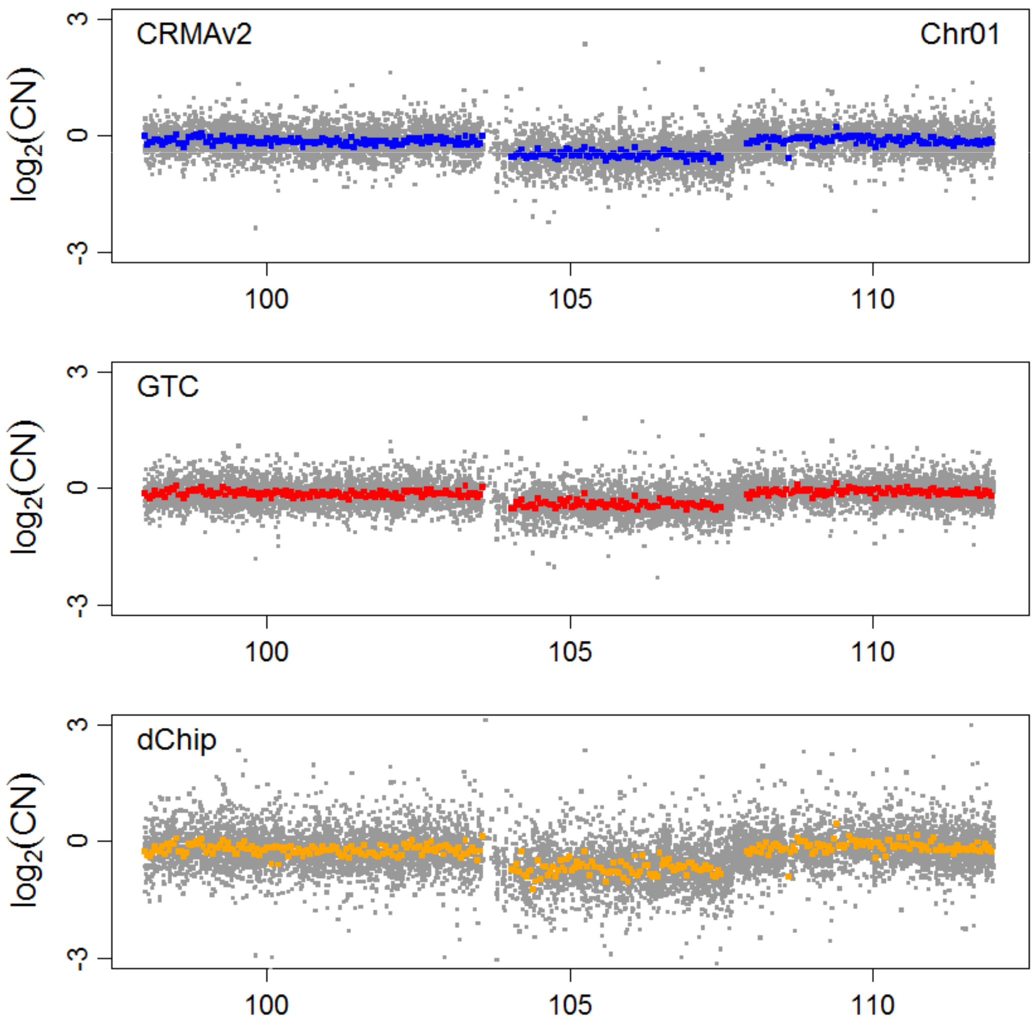
CRMAv2 performs well also at various amount of smoothing (“resolution”)



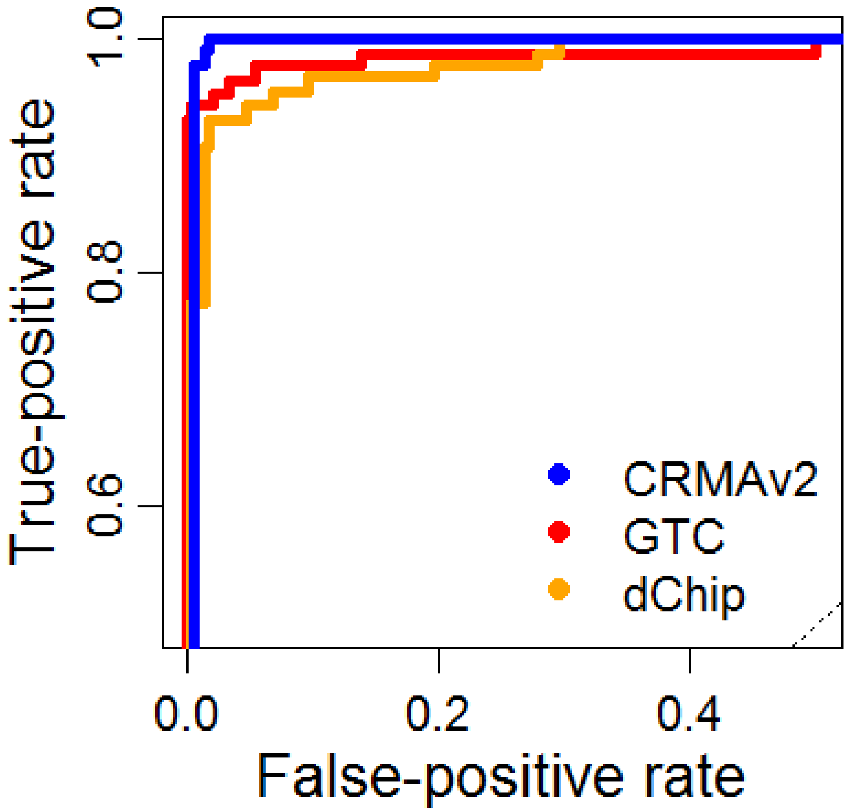
20kb binning



CRMAv2 performs well also at various amount of smoothing (“resolution”)



40kb binning



Summary

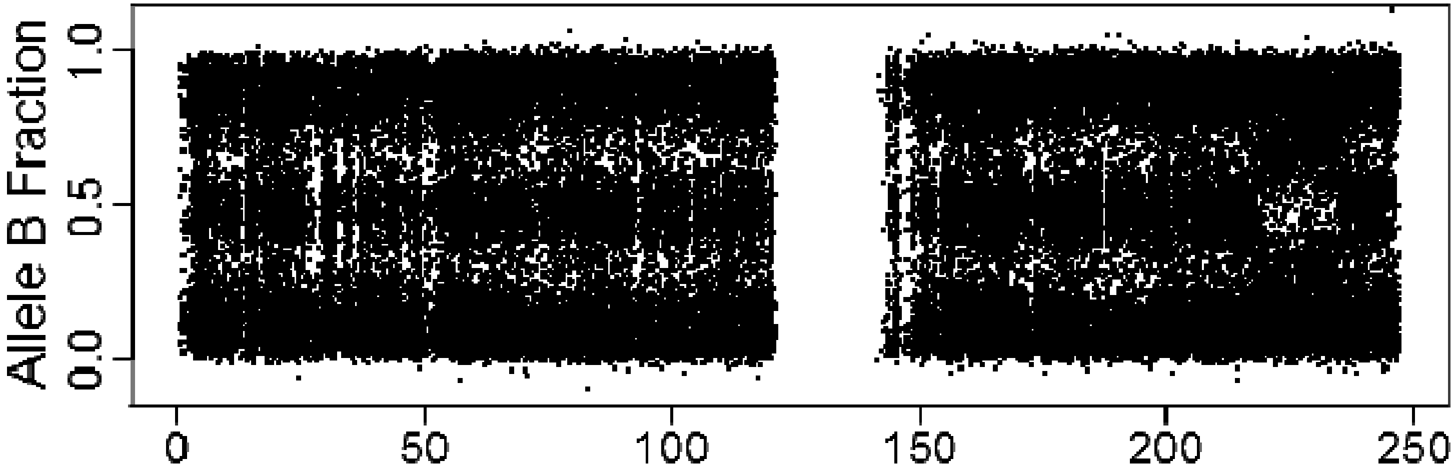
Conclusions

- CRMA v2:
  - a single-array preprocessing method.
  - can detect CN changes as well or better than existing multi-array methods.
  - applies to all Affymetrix chip types.

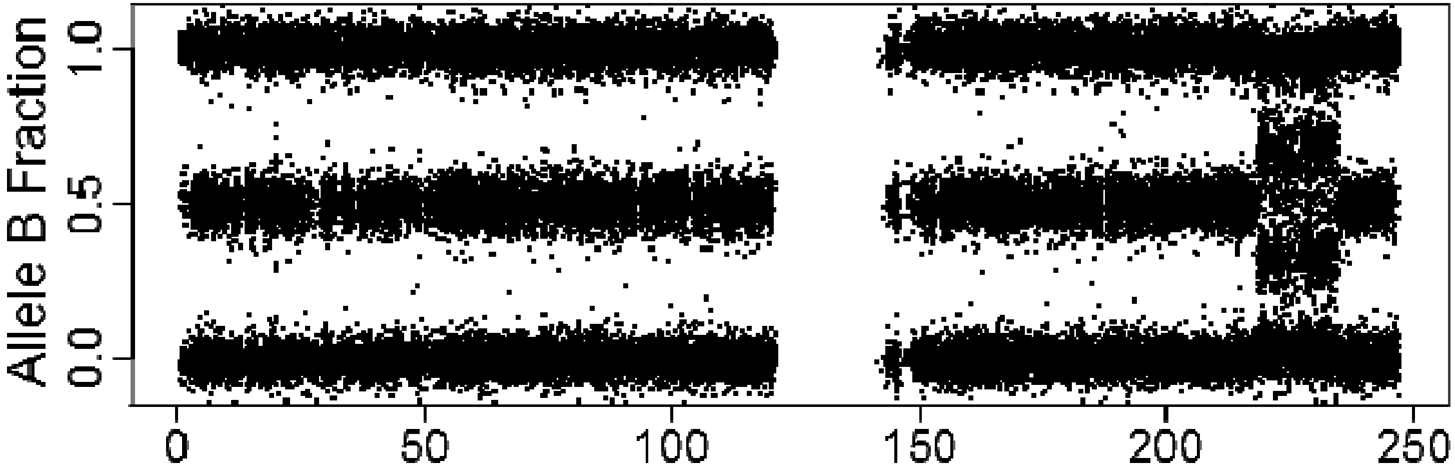
- Single-array methods are useful for:
  - large-scale projects.
  - personalized diagnostics.

Near future: Single-sample calibration of allele-specific CN estimates

Now:



Next:



Acknowledgments

UC Berkeley:

- James Bullard
- Kasper Hansen
- Pierre Neuvial
- Terry Speed

Lawrence Berkeley National Labs:

- Amrita Ray
- Paul Spellman

WEHI, Melbourne, Australia:

- Mark Robinson
- Ken Simpson

John Hopkins, Baltimore:
Benilton Carvalho
Rafael Irizarry

ISREC, Lausanne, Switzerland:

- Pratyaksha “Asa” Wirapati

Affymetrix, California:

- Ben Bolstad
- Simon Cawley
- Jim Veitch

Appendix

Complete aroma.affymetrix script for copy-number analysis of 270 SNP6.0 HapMap samples

```
cdf <- AffymetrixCdfFile$byChipType("GenomeWideSNP_6")
csR <- AffymetrixCelSet$byName("HapMap270", cdf=cdf)

acc <- AllelicCrosstalkCalibration(csR)
csC <- process(acc)

bpn <- BasePositionNormalization(csC)
csN <- process(bpn)

plm <- AvgCnPlm(csN)
fit(plm)

ces <- getChipEffectSet(plm)
fln <- FragmentLengthNormalization(ces)
cesN <- process(fln)

seg <- CbsModel(cesN)
regions <- fit(seg)
```