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

A single-array CN method

this-gen

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.

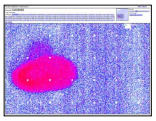
At UC Berkeley we have a few single-sample methods in place

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).

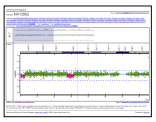
All of the above is done without using priors.

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.



- **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.**



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

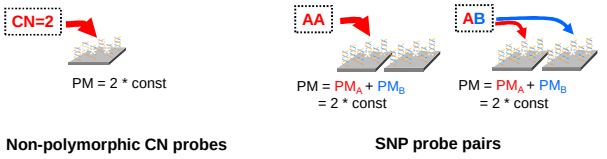
CRMA v2

-

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)



Non-polymorphic CN probes **SNP probe pairs**

Amplification (e.g. CN=3)

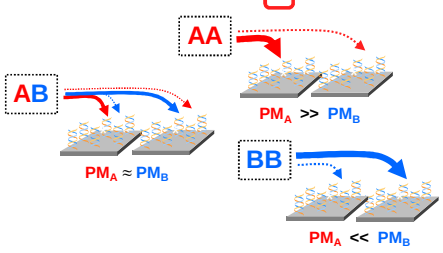
CRMA v2

Probe signals	Allele-crosstalk calibration Probe-sequence normalization
Summarization *(allele-specific or total)*	Robust averaging θ_{ijjA}= median$_k$(PM_{ijAk}) θ_{ijjB}= 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

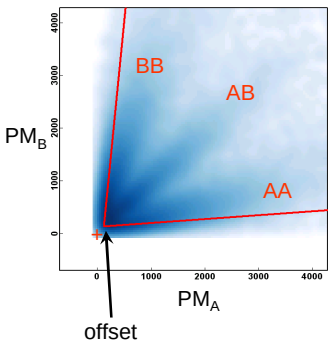
Crosstalk between alleles *- adds significant artifacts to signals*

Cross-hybridization:

Allele A: ***TCGGTAAGTACTC***
Allele B: ***TCGGTATGTACTC***

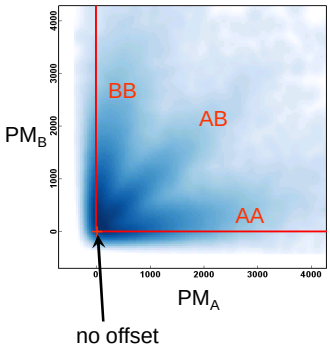


Crosstalk between alleles is easy to spot



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

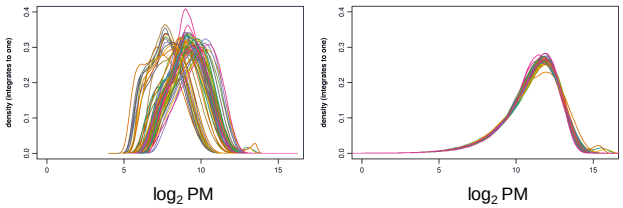


What is done:
1. Offset is removed from SNPs and CN units.
2. Crosstalk is removed from SNPs.

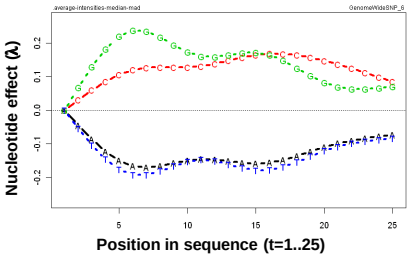
Crosstalk calibration corrects for differences in signal distributions too

Before removing crosstalk the arrays differ significantly...

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



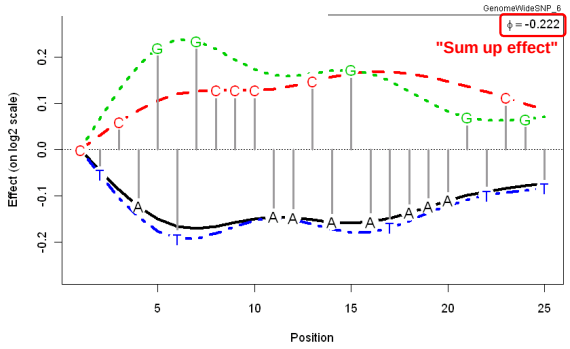
The nucleotide (A, C, G or T) and its position in the probe adds to the affinity



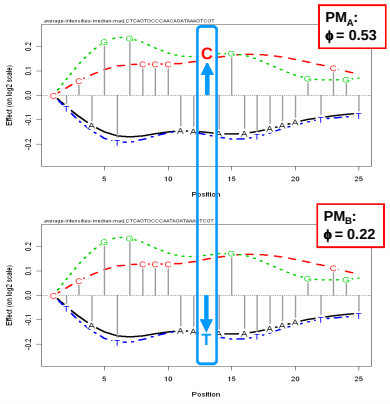
Probe-position affinity for probe k:

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

Example: Probe-position affinity for CTCAGTGCCCAACAGATAAAGTCGT



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



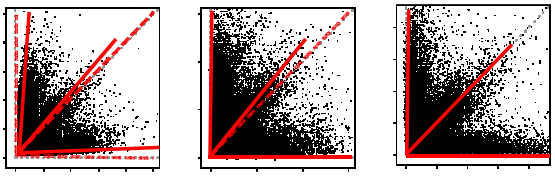
Genotypic imbalances:

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



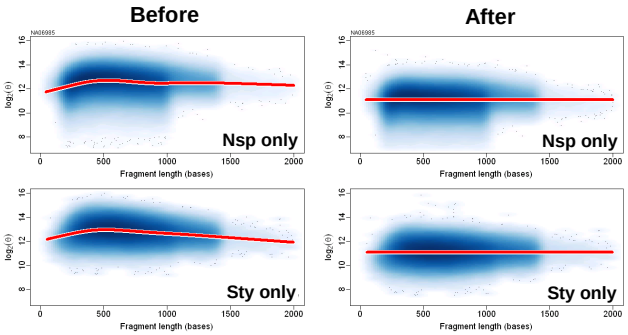
(raw) ACC ACC + NPN

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

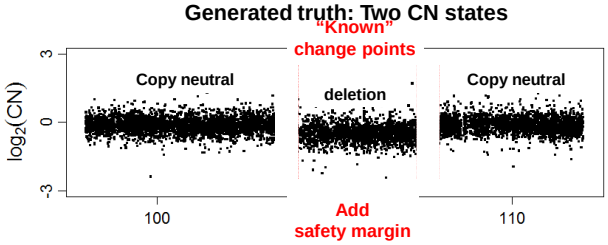


CRMA v2

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

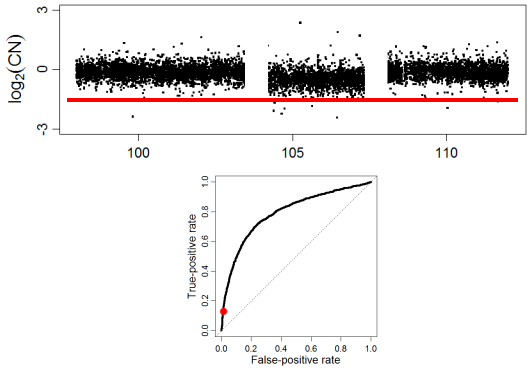
Idea: How well can be detect a known CN aberration?



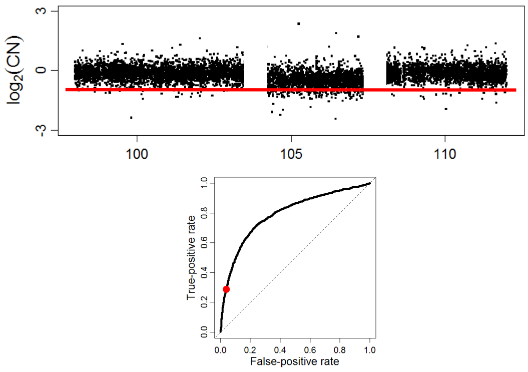
Example: 3.9 Mb deletion on Chromosome 1 in tumor GSM337641.
Data set: Chiang et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nature Methods, 2009.

How well does the two states separate?

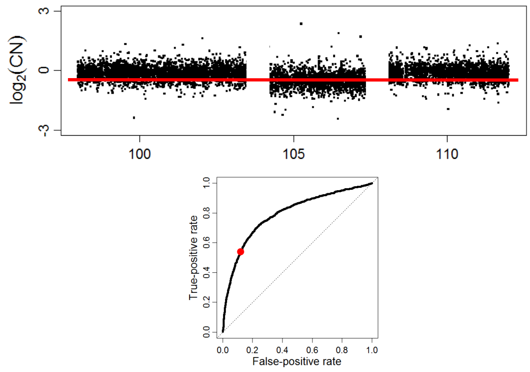
Calling CN deletion with common threshold



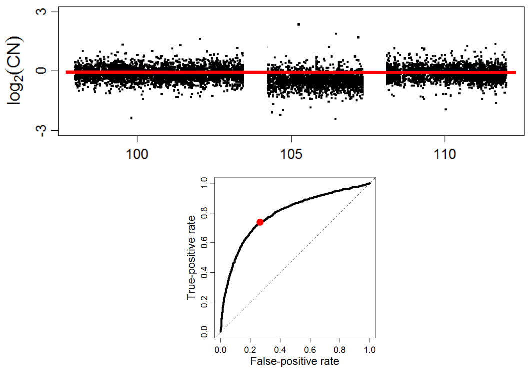
How well does the two states separate?
Calling CN deletion with common threshold



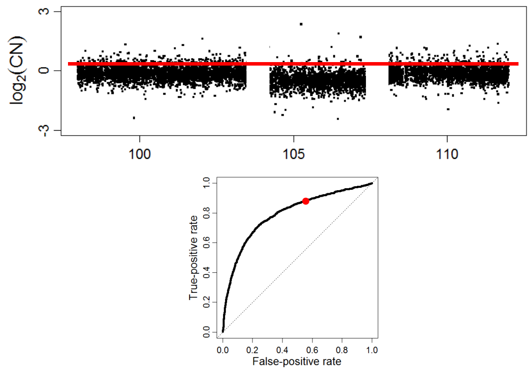
How well does the two states separate?
Calling CN deletion with common threshold



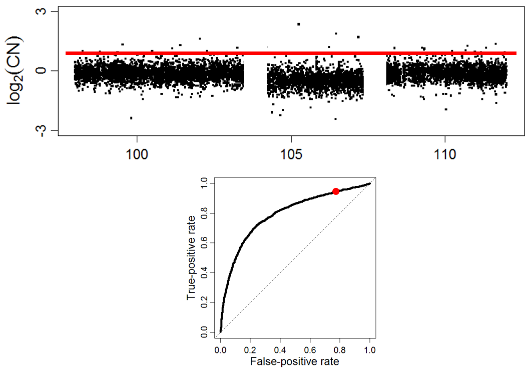
How well does the two states separate?
Calling CN deletion with common threshold



How well does the two states separate?
Calling CN deletion with common threshold



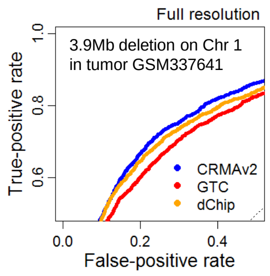
How well does the two states separate?
Calling CN deletion with common threshold



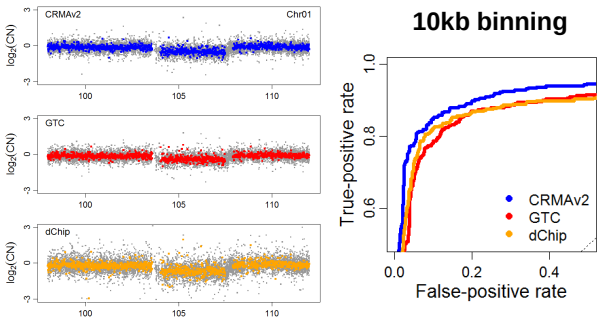
Single-array CRMAv2 performs well compared with Affymetrix GTC and dChip

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

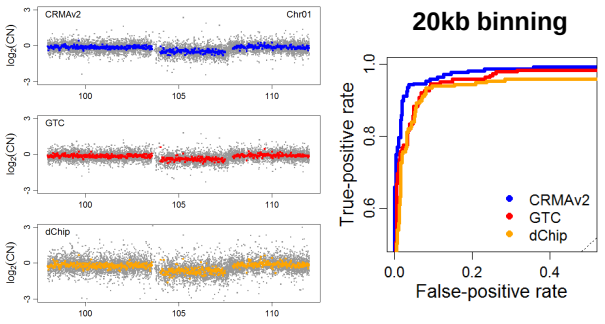
Preprocessing:
- GTC (CN5) and dChip were allowed to use all 68 arrays in their processing.



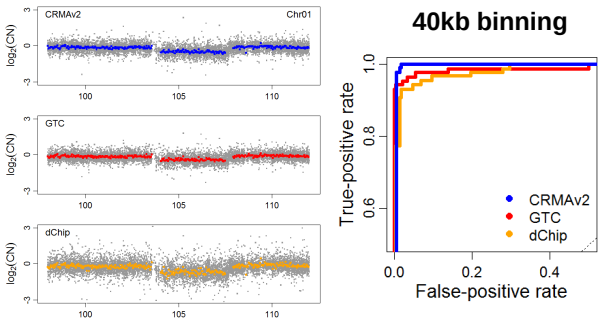
CRMAv2 performs well also at various amount of smoothing ("resolution")



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



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



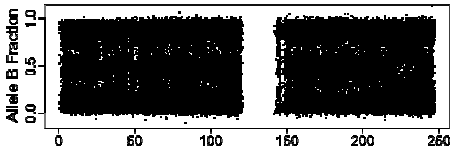
Summary

Conclusions

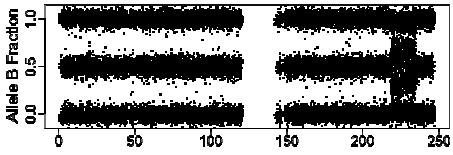
- CRMA v2:
  - a single-array preprocessing method.
  - can detect CN changes as well or better than existing multi-array methods.
  - applies to all Affymetrix chip types.
- Single-array methods are useful for:
  - large-scale projects.
  - personalized diagnostics.

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)
```