Low-Level Copy Number Analysis

Part 1 - Background

Henrik Bengtsson

Post doc, Department of Statistics,

University of California, Berkeley, USA

CEIT Workshop on SNP arrays,

Dec 15-17, 2008, San Sebastian

Acknowledgments

UC Berkeley:

- James Bullard
- Kasper Hansen
- Elizabeth Purdom
- Terry Speed

Lawrence Berkeley National Labs:

- Amrita Ray
- Paul Spellman

John Hopkins, Baltimore:

- Benilton Carvalho
- Rafael Irizarry

WEHI, Melbourne, Australia:

- Mark Robinson
- Ken Simpson

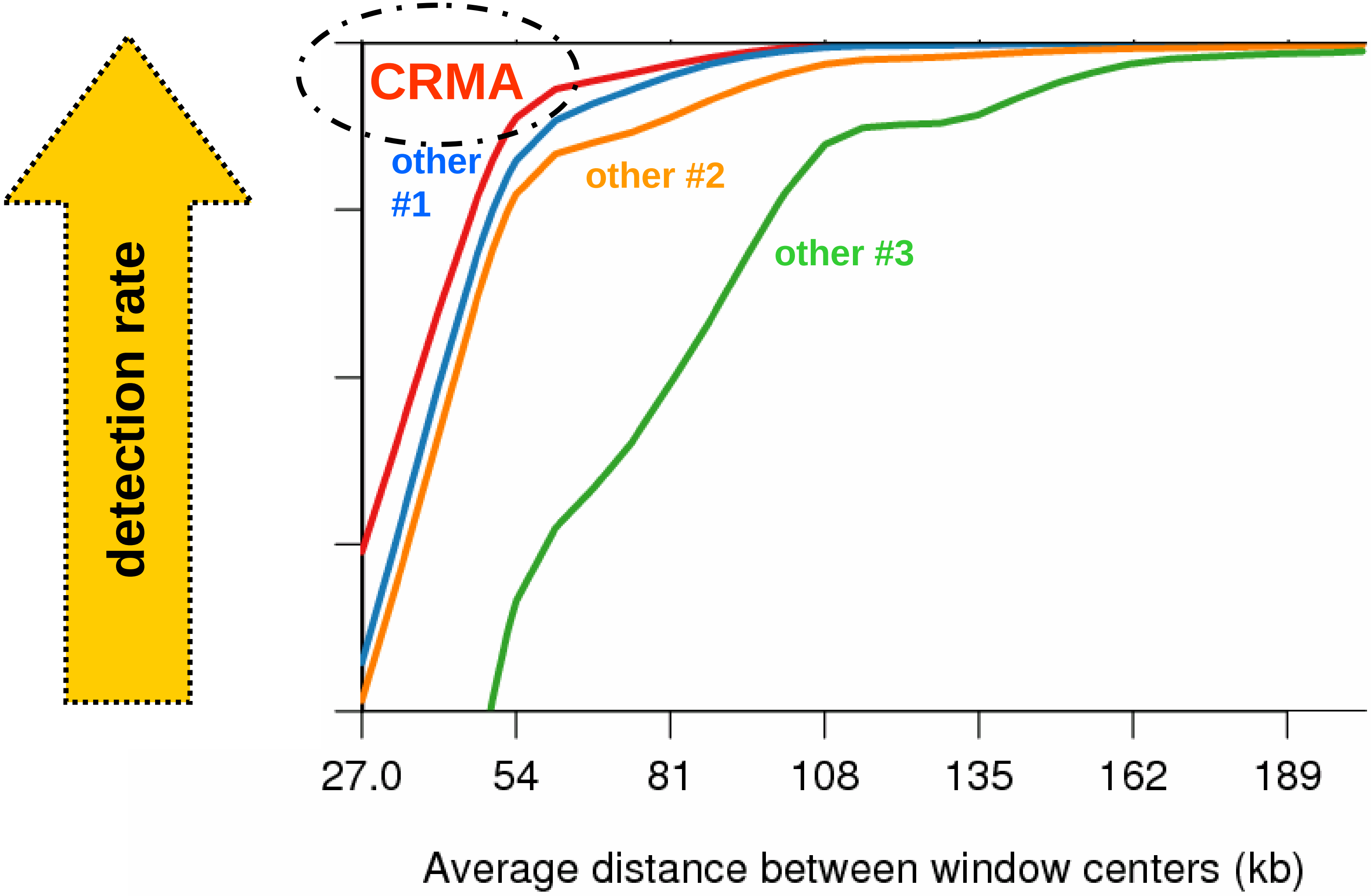
ISREC, Lausanne, Switzerland:

- Pratyaksha “Asa” Wirapati

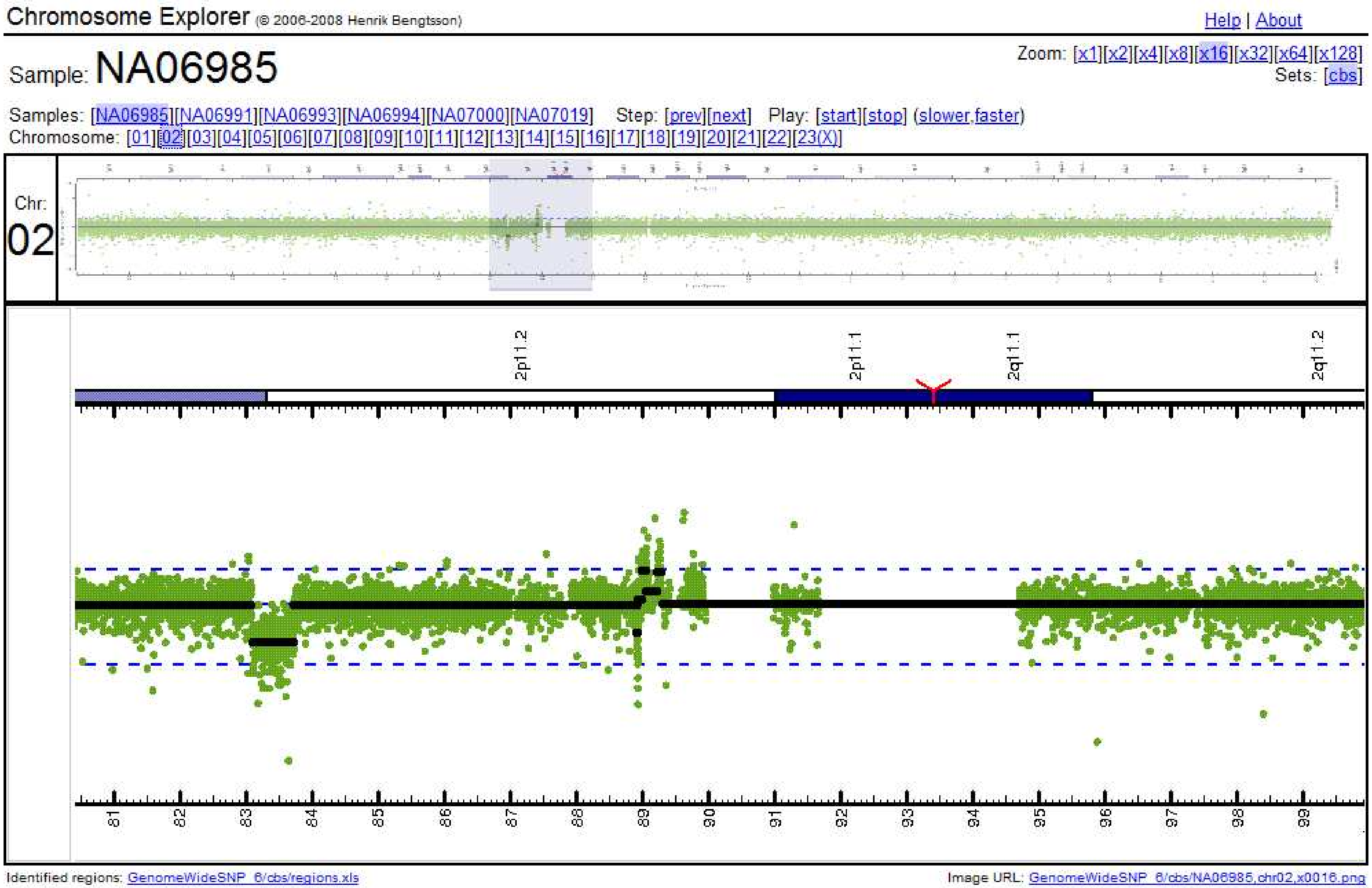
Affymetrix, California:

- Ben Bolstad
- Simon Cawley
- Steve Chervitz
- Harley Gorrell
- Earl Hubbell
- Luis Jevons
- Chuck Sugnet
- Jim Veitch
- Alan Williams

Detect more and smaller aberrations with less errors



Copy number analysis is about finding "aberrations" in one or several individuals



Chromosome Explorer (© 2006-2008 Henrik Bengtsson)

Help | About

Sample: NA06985

Zoom: [x1][x2][x4][x8][x16][x32][x64][x128]

Sets: [cbs]

Samples: [NA06985][NA06991][NA06993][NA06994][NA07000][NA07019] Step: [prev][next] Play: [start][stop] (slower,faster)

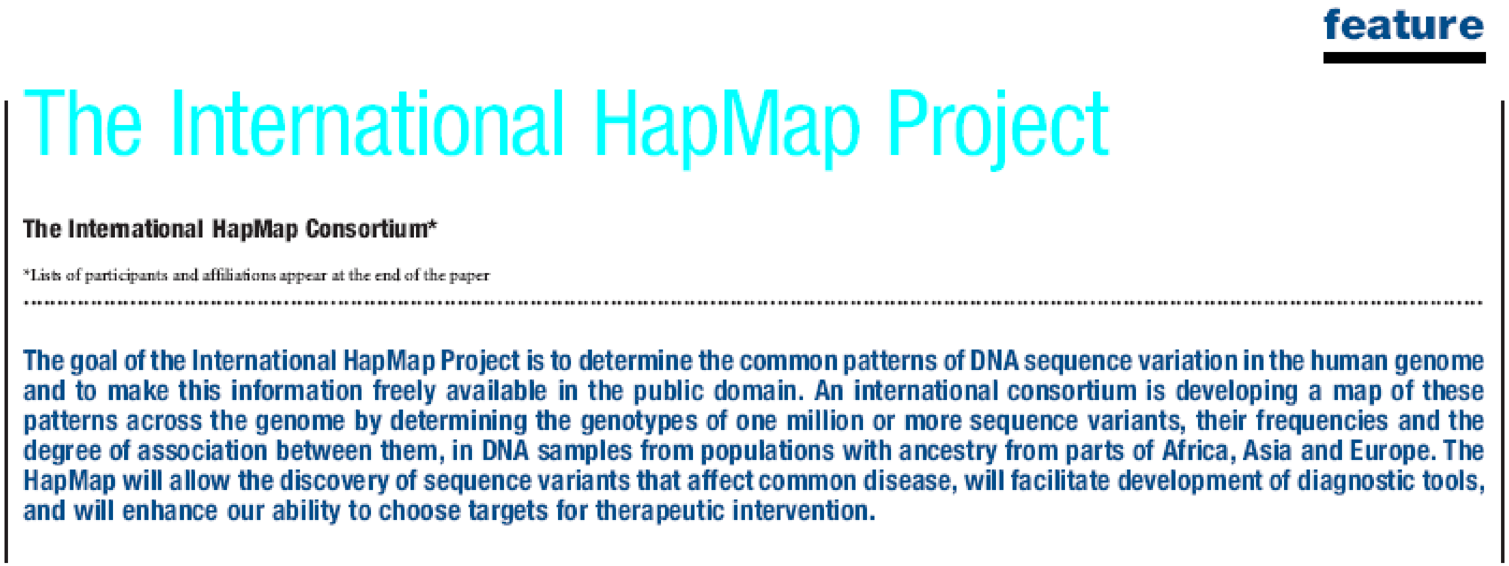
Chromosome: [01][02][03][04][05][06][07][08][09][10][11][12][13][14][15][16][17][18][19][20][21][22][23(X)]

Chr: 02

Identified regions: GenomeWideSNP_6/cbs/regions.xls

Image URL: GenomeWideSNP_6/cbs/NA06985,chr02,x0016.png

The HapMap project
- Large project to identify SNPs in Humans (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.

The HapMap is a catalog of common genetic variants (SNPs) that occur in human beings. It describes what These variants are, where they occur in our DNA, and How they are distributed among people within populations and among populations in different parts of the world.

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

The HapMap project

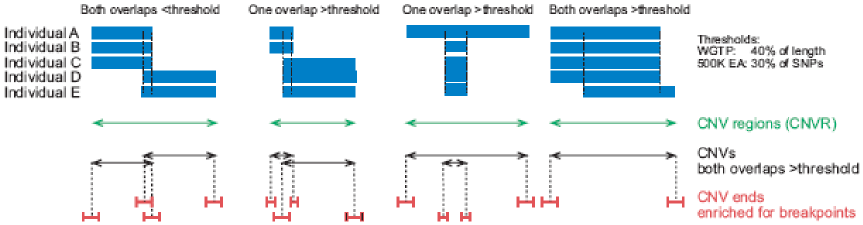
- 270 normal individuals genotyped by different labs using various technologies

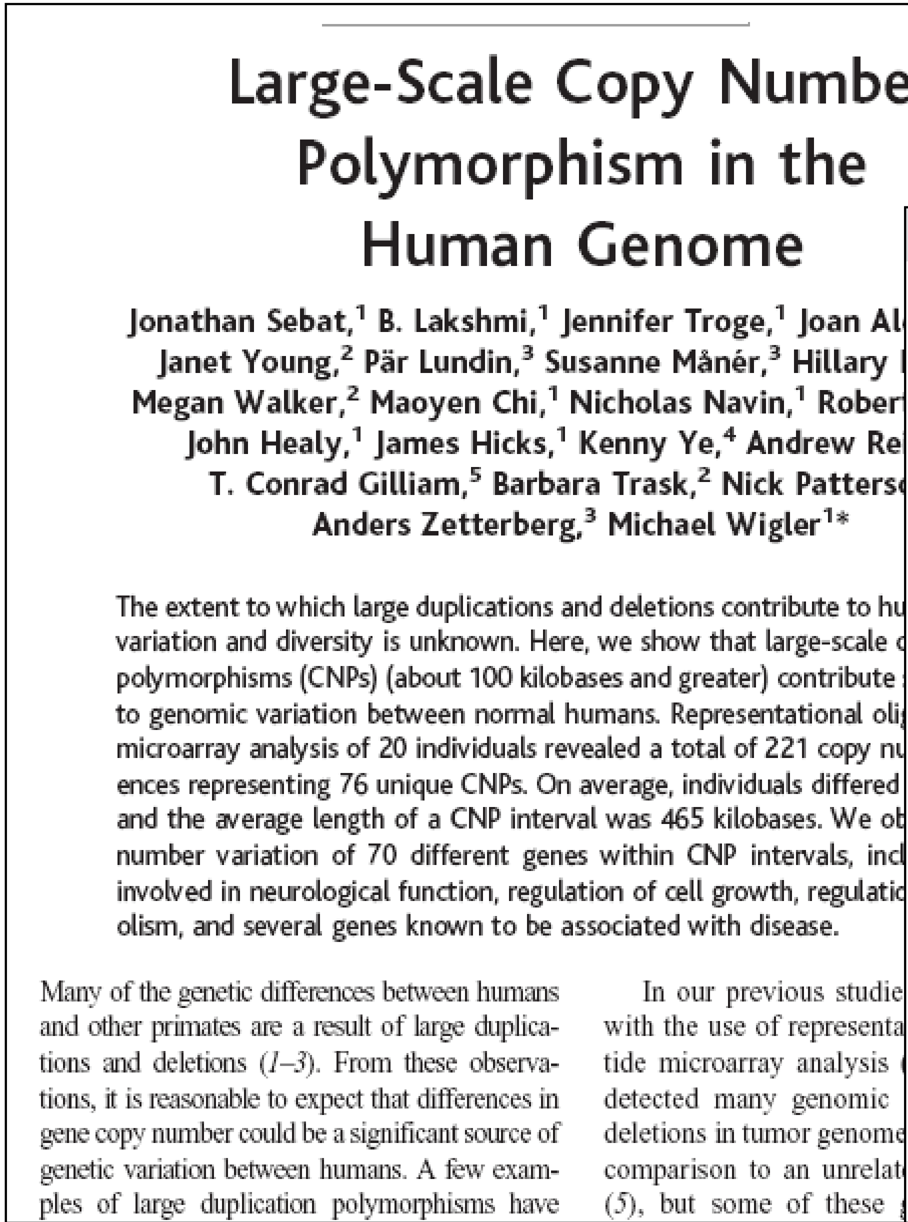
- 90 CEU individuals (Utah/Europe, 30 trio families)
- 90 YRI individuals (Nigeria; 30 trio families)
- 45 CHB (China; unrelated)
- 45 JPT (Japan; unrelated)

Publicly available:

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

Copy number polymorphism
- *People share common CN aberrations (2005-)*





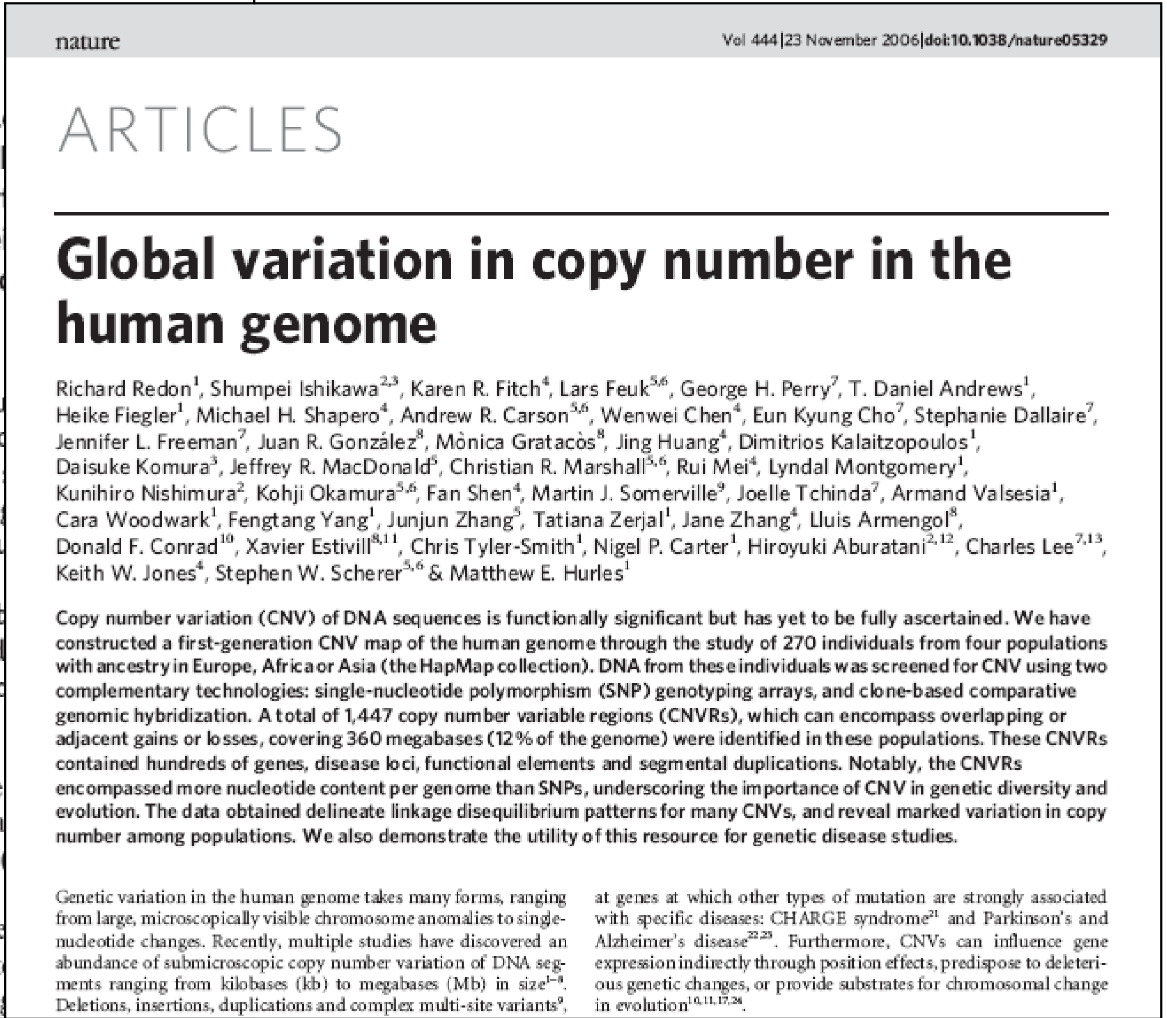
Large-Scale Copy Number Polymorphism in the Human Genome

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] Robert
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 c
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, regulati
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

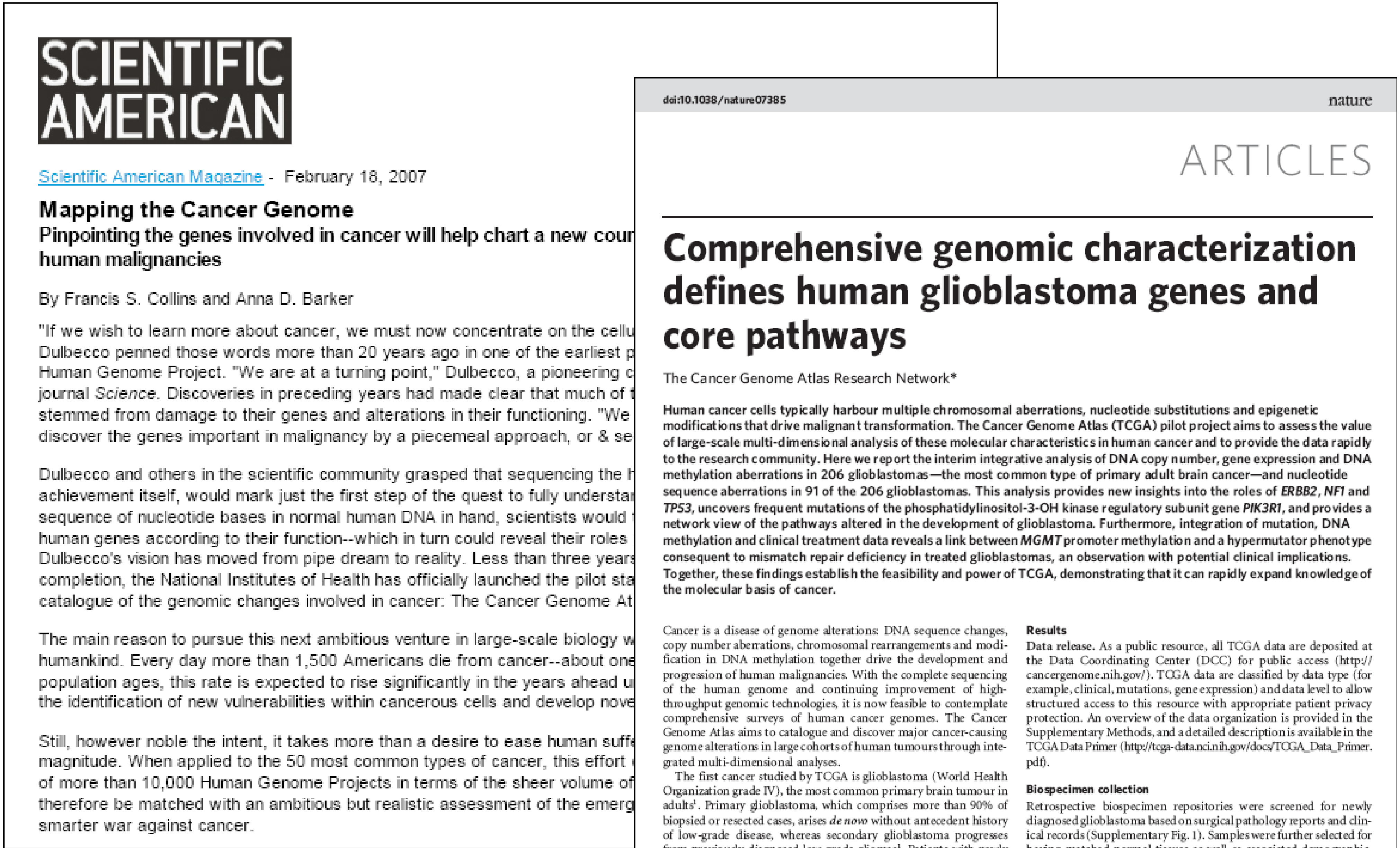
Richard Redon[1], Shumpei Ishikawa[2,3], Karen R. Fitch[4], Lars Feuk[5,6], George H. Perry[7], T. Daniel Andrews[1], Heike Fiegler[1], Michael H. Shapero[4], Andrew R. Carson[5,6], Wenwei Chen[4], Eun Kyung Cho[7], Stephanie Dallaire[7], Jennifer L. Freeman[7], Juan R. González[8], Mònica Gratacòs[8], 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[1–8]. 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[10,11,17,24].

The Cancer Genome Atlas (TCGA) project
- Large project for genetic mapping of tumors (2007-)



SCIENTIFIC AMERICAN

Scientific American Magazine - February 18, 2007

Mapping the Cancer Genome
Pinpointing the genes involved in cancer will help chart a new cour
human malignancies

By Francis S. Collins and Anna D. Barker

"If we wish to learn more about cancer, we must now concentrate on the cellu
Dulbecco penned those words more than 20 years ago in one of the earliest p
Human Genome Project. "We are at a turning point," Dulbecco, a pioneering c
journal *Science*. Discoveries in preceding years had made clear that much of t
stemmed from damage to their genes and alterations in their functioning. "We
discover the genes important in malignancy by a piecemeal approach, or & se

Dulbecco and others in the scientific community grasped that sequencing the h
achievement itself, would mark just the first step of the quest to fully understa
sequence of nucleotide bases in normal human DNA in hand, scientists would
human genes according to their function--which in turn could reveal their roles
Dulbecco's vision has moved from pipe dream to reality. Less than three years
completion, the National Institutes of Health has officially launched the pilot sta
catalogue of the genomic changes involved in cancer: The Cancer Genome At

The main reason to pursue this next ambitious venture in large-scale biology w
humankind. Every day more than 1,500 Americans die from cancer--about one
population ages, this rate is expected to rise significantly in the years ahead u
the identification of new vulnerabilities within cancerous cells and develop nove

Still, however noble the intent, it takes more than a desire to ease human suff
magnitude. When applied to the 50 most common types of cancer, this effort
of more than 10,000 Human Genome Projects in terms of the sheer volume of
therefore be matched with an ambitious but realistic assessment of the emerg
smarter war against cancer.

doi:10.1038/nature07385 nature

ARTICLES

Comprehensive genomic characterization defines human glioblastoma genes and core pathways

The Cancer Genome Atlas Research Network*

Human cancer cells typically harbour multiple chromosomal aberrations, nucleotide substitutions and epigenetic modifications that drive malignant transformation. The Cancer Genome Atlas (TCGA) pilot project aims to assess the value of large-scale multi-dimensional analysis of these molecular characteristics in human cancer and to provide the data rapidly to the research community. Here we report the interim integrative analysis of DNA copy number, gene expression and DNA methylation aberrations in 206 glioblastomas—the most common type of primary adult brain cancer—and nucleotide sequence aberrations in 91 of the 206 glioblastomas. This analysis provides new insights into the roles of *ERBB2*, *NF1* and *TP53*, uncovers frequent mutations of the phosphatidylinositol-3-OH kinase regulatory subunit gene *PIK3R1*, and provides a network view of the pathways altered in the development of glioblastoma. Furthermore, integration of mutation, DNA methylation and clinical treatment data reveals a link between *MGMT* promoter methylation and a hypermutator phenotype consequent to mismatch repair deficiency in treated glioblastomas, an observation with potential clinical implications. Together, these findings establish the feasibility and power of TCGA, demonstrating that it can rapidly expand knowledge of the molecular basis of cancer.

Cancer is a disease of genome alterations: DNA sequence changes, copy number aberrations, chromosomal rearrangements and modification in DNA methylation together drive the development and progression of human malignancies. With the complete sequencing of the human genome and continuing improvement of high-throughput genomic technologies, it is now feasible to contemplate comprehensive surveys of human cancer genomes. The Cancer Genome Atlas aims to catalogue and discover major cancer-causing genome alterations in large cohorts of human tumours through integrated multi-dimensional analyses.

The first cancer studied by TCGA is glioblastoma (World Health Organization grade IV), the most common primary brain tumour in adults[1]. Primary glioblastoma, which comprises more than 90% of biopsied or resected cases, arises *de novo* without antecedent history of low-grade disease, whereas secondary glioblastoma progresses

Results

Data release. As a public resource, all TCGA data are deposited at the Data Coordinating Center (DCC) for public access (http://cancergenome.nih.gov/). TCGA data are classified by data type (for example, clinical, mutations, gene expression) and data level to allow structured access to this resource with appropriate patient privacy protection. An overview of the data organization is provided in the Supplementary Methods, and a detailed description is available in the TCGA Data Primer (http://tcga-data.nci.nih.gov/docs/TCGA_Data_Primer.pdf).

Biospecimen collection

Retrospective biospecimen repositories were screened for newly diagnosed glioblastoma based on surgical pathology reports and clinical records (Supplementary Fig. 1). Samples were further selected for

The TCGA project

- A large number of tissues are studies with many DNA & RNA technologies

- Tumor types:
  - brain cancer (glioblastoma multiforme, or GBM),
  - lung cancer (squamous cell carcinoma of the lung), and
  - ovarian cancer (serous cystadenocarcinoma of the ovary).
- 234 tumors (of 500) characterized.
- Multiple labs in the US
  - Broad, Harvard, Stanford, LBNL, …
- High quality data.
- Platforms: Affymetrix, Illumina, Agilent, …
- Gene-, exon-, microRNA- expression, methylation, SNP & CN, sequencing…
- Raw and summarized data immediately available (publicly), e.g. Affymetrix CEL files.

Combining copy numbers across platforms & labs

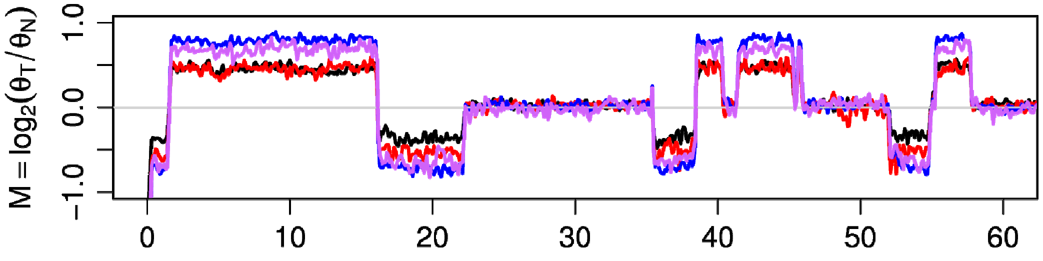
Henrik Bengtsson (UC Berkeley), Amrita Ray (LBNL), Paul Spellman (LBNL), Terry Speed (UC Berkeley)

BACKGROUND:

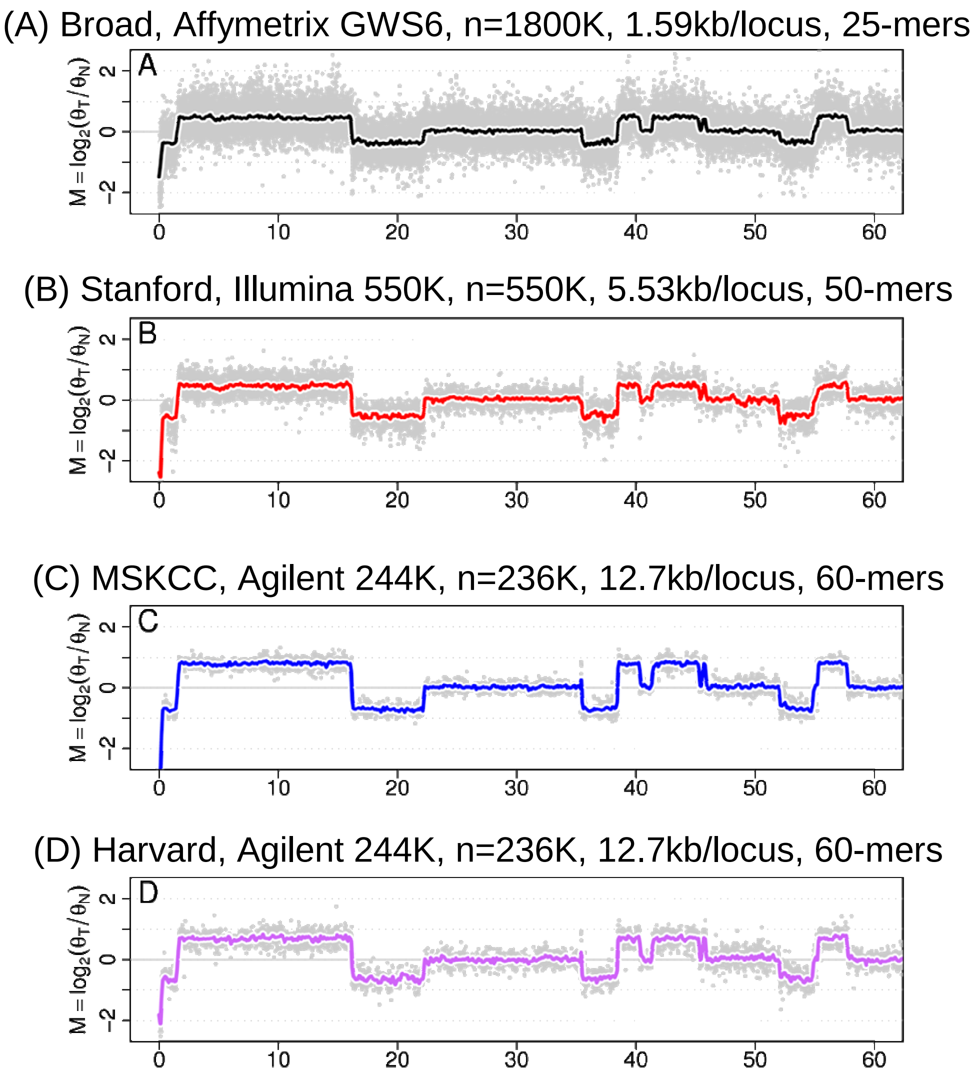
Whole-genome copy-number (CN) studies are rapidly expanding, and with this expansion comes a **demand for increased precision and resolution** of CN estimates. Several recent studies have obtained **CN estimates from more than one platform** on the same samples, and it is natural to want to **combine** the different estimates in order to meet this demand.

PROBLEM:

CN estimates from **different platforms show different degrees of attenuation** of the true CN changes. Differences can also be observed in CN estimates from the same platform run in different labs, or in the same lab, with different analytical methods. This is the reason why it is **not straightforward matter to combine CN estimates from different sources** (platforms, labs, analysis methods, etc).



The smoothed raw CNs from the four sources have similar CN profiles but different mean levels.


(A) Broad, Affymetrix GWS6, n=1800K, 1.59kb/locus, 25-mers

(B) Stanford, Illumina 550K, n=550K, 5.53kb/locus, 50-mers

(C) MSKCC, Agilent 244K, n=236K, 12.7kb/locus, 60-mers

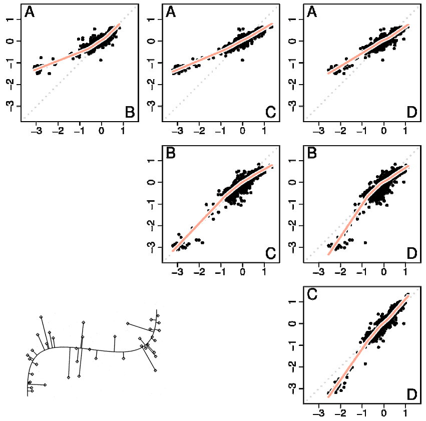
(D) Harvard, Agilent 244K, n=236K, 12.7kb/locus, 60-mers

Tumor/normal CNs by four TCGA centers ("sources") in a 60Mb region on Chr3 in sample TCGA-02-104. (The combined set would consist of 2,822K loci with 0.95kb/locus.)

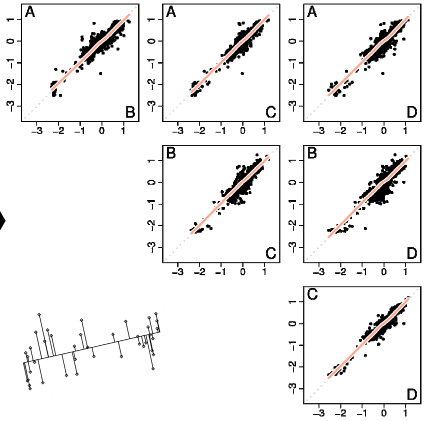
METHOD:

We have developed a single-sample multi-source normalization that **brings full-resolution CN estimates to the same scale** across sources.

Kernel estimators and principal component curves are used to estimate the **non-linear relationships** between the sources. Full-resolution data is then normalized such that these **relationships become linear**.

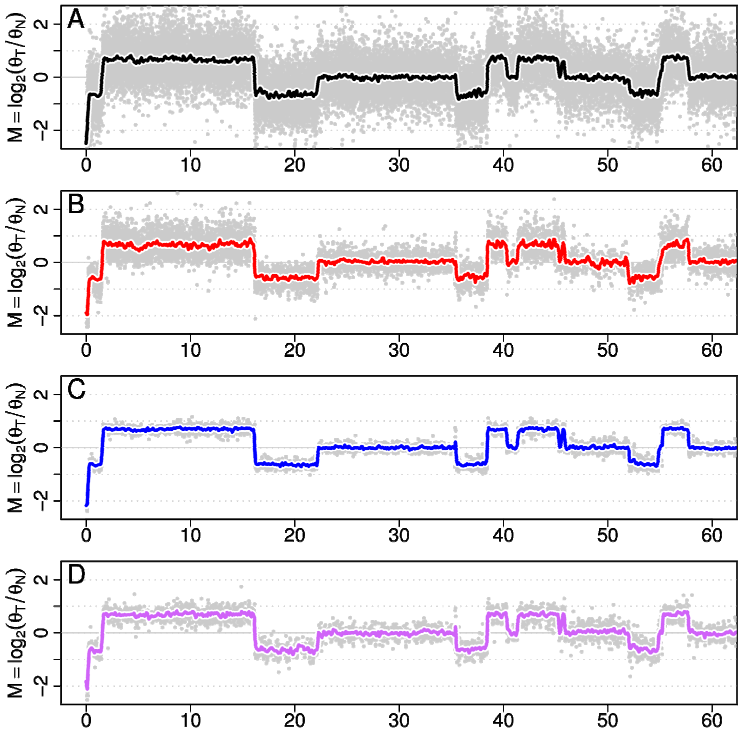


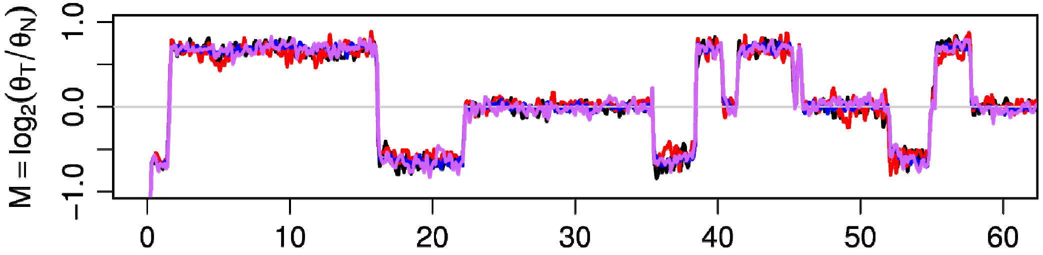
Before normalization:
Non-linearity between pairs



After normalization:
Linearity between pairs

The normalized estimates are such that for any underlying CN level, **the mean level of the CN estimates is the same regardless of source**. CNs with consistent mean levels are **better suited for being combined across sources**, e.g. existing segmentation methods may be used to identify aberrant regions.





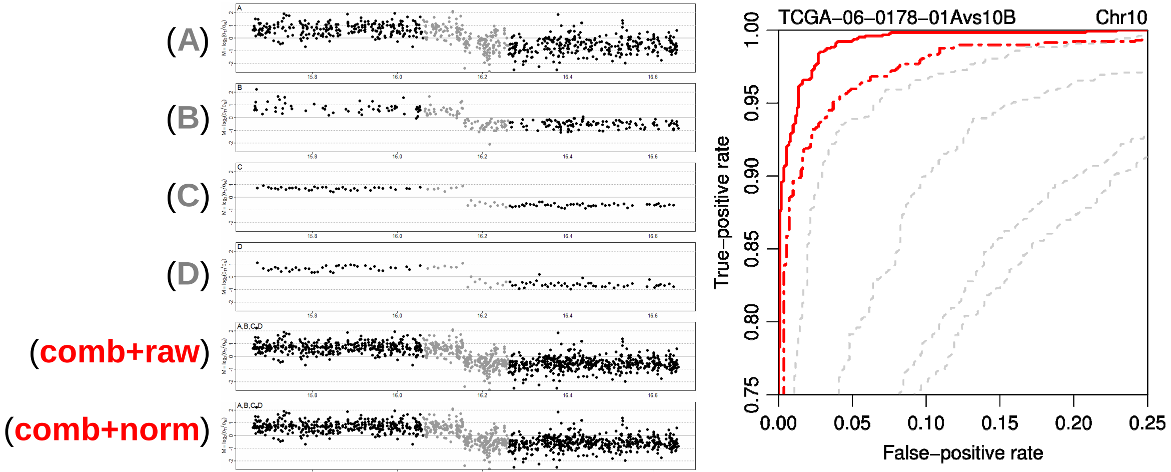
The smoothed ***normalized*** *CNs from the four sources have similar CN profiles* ***and same mean levels****.*

Normalized ***full-resolution*** *CNs for the four sources.*

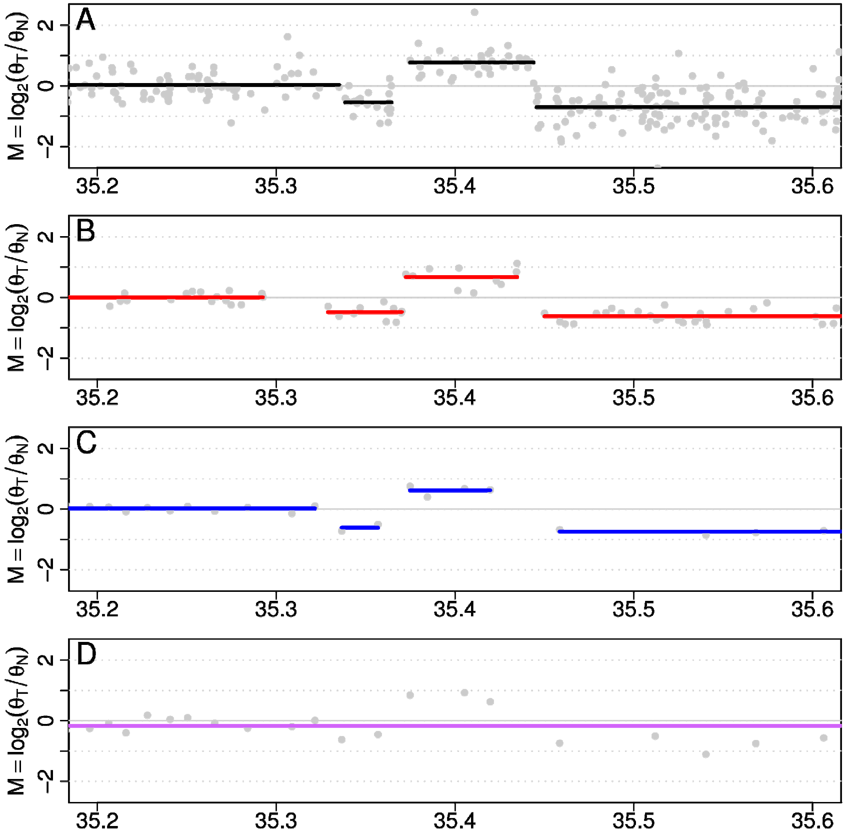
RESULTS:

We use microarray-based CN estimates from **The Cancer Genome Atlas (TCGA)** project to illustrate the method. We show that **after normalization the mean levels** of randomly selected CN aberrations **are the same across platforms**, and that the normalized and combined data better separate two CN states at a given resolution.

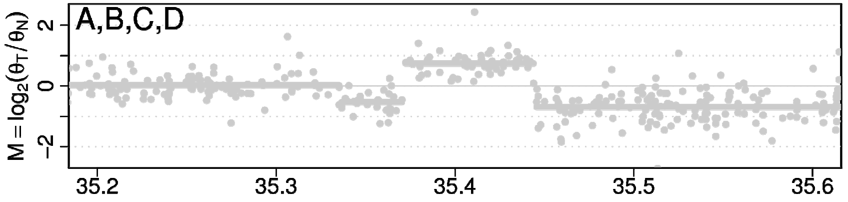
We conclude that it is possible to **combine CNs** from multiple sources such that the **resolution becomes effectively larger**, and when multiple platforms are combined, they also **enhance the genome coverage** by complementing each other in different regions.



*At any given resolution (amount of smoothing), with combined normalized CNs (**solid red**) one can separate two CN states **better** than with combined raw CNs (**dot-dashed red**), and with each of the individuals sources (**gray dotted**).*

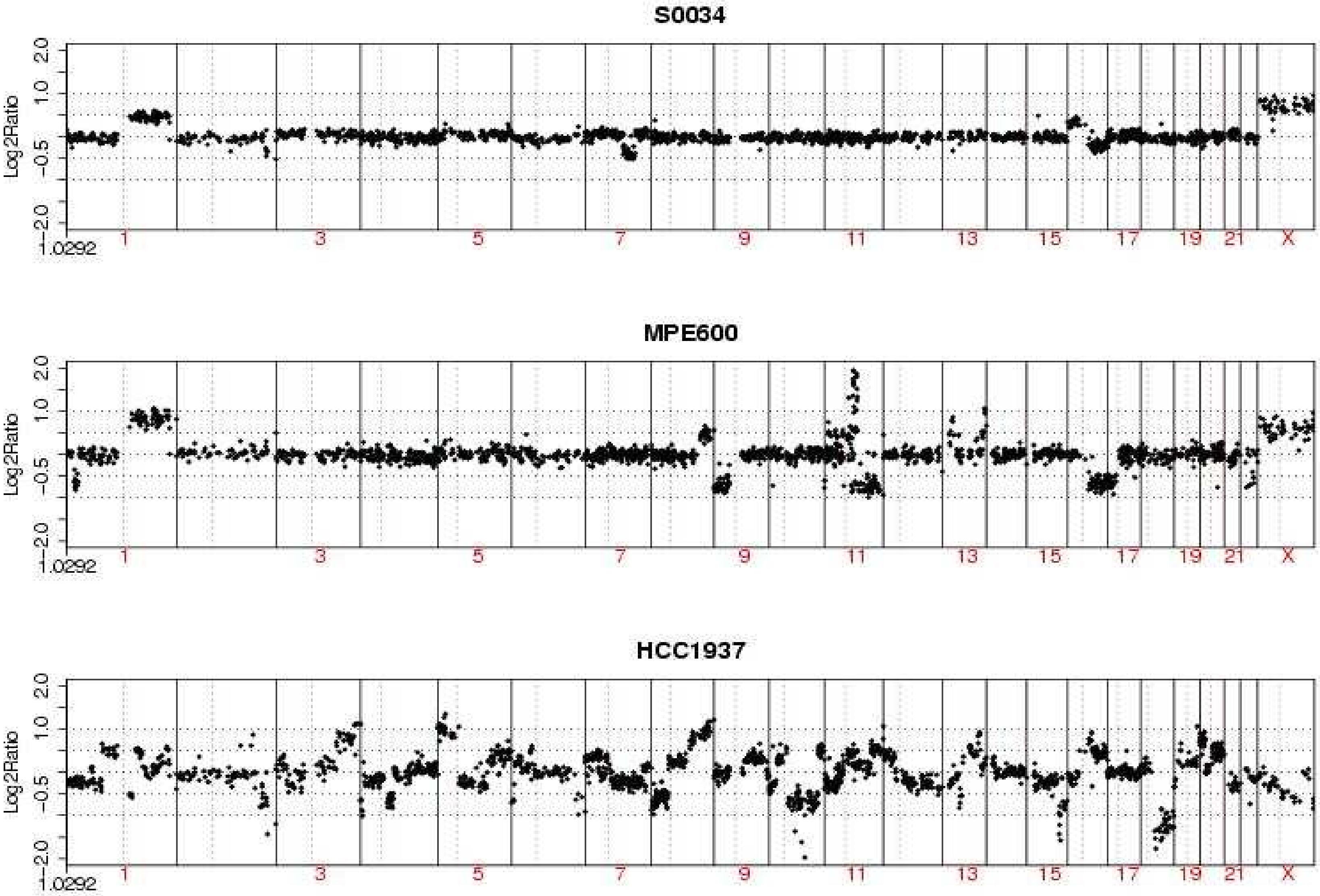


A 400kb region in TCGA-02-104 on Chr 3: CNs from different sources give different segmenting results at different precisions.



With combined normalized CNs, there is more power to detect change points (CPs) and their locations are more precise.

Examples of genomic profiles



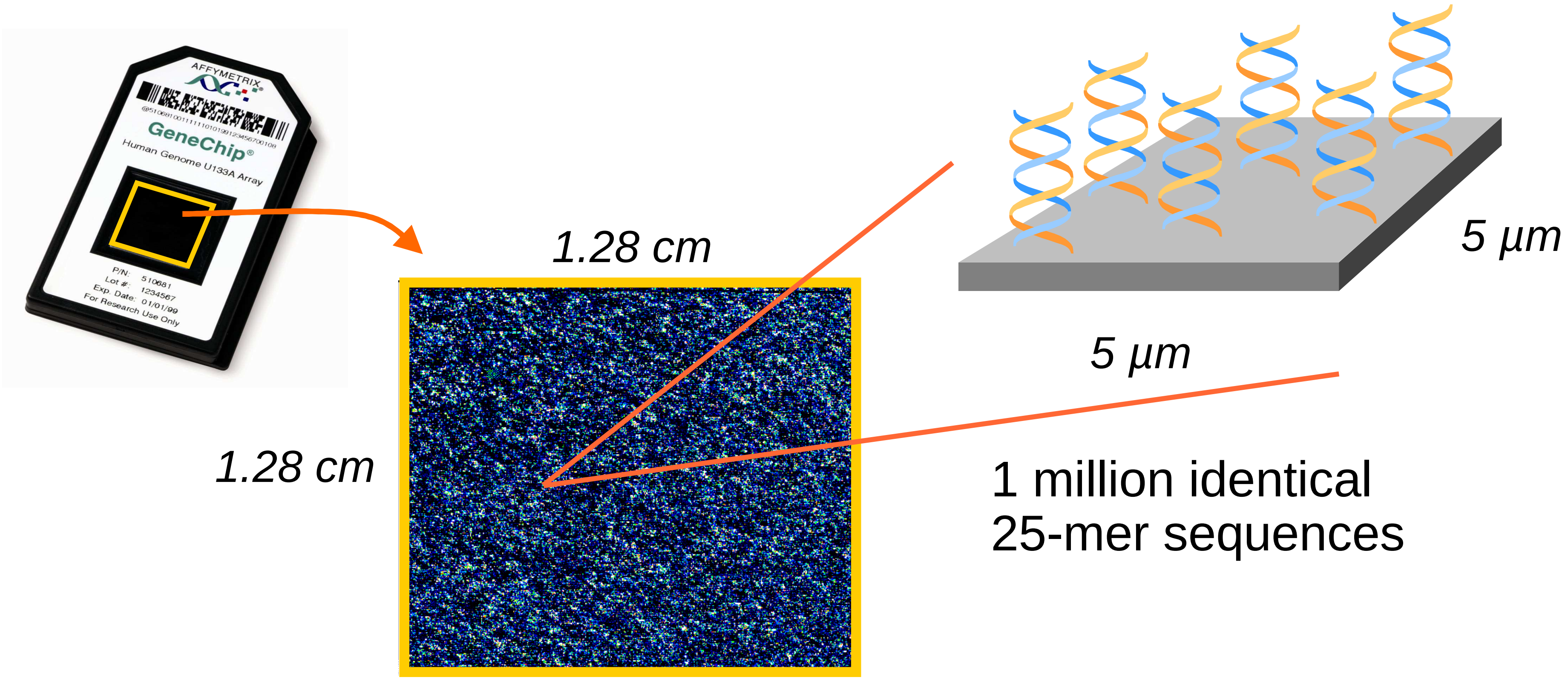
S0034

MPE600

HCC1937

The Affymetrix platform

The Affymetrix GeneChip
is a synthesized high-density (single-array) microarray



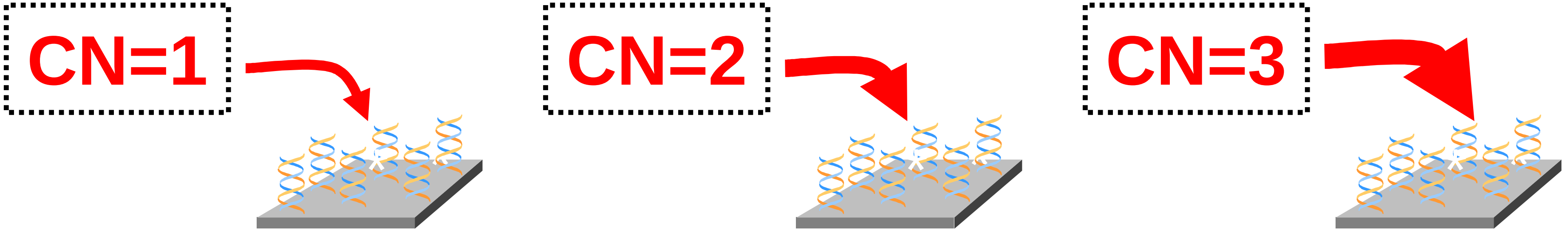
1 million identical
25-mer sequences

6.5 million probes/chip

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

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

PM: **ATCGGTAGCCATTCATGAGTTACTA**



CN=1

PM = c

CN=2

PM = 2c

CN=3

PM = 3c

Single Nucleotide Polymorphism (SNP)

Definition:
A sequence variation such that two chromosomes 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

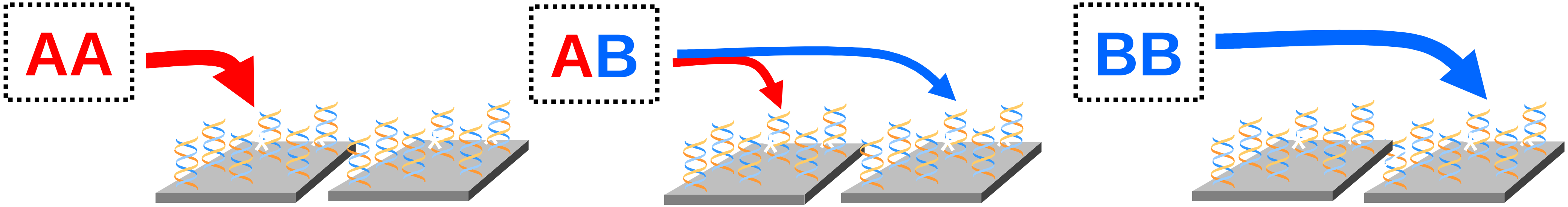
PM_A: ATCGGTAGCCATTCATGAGTTACTA

Allele A: *...CGTAGCCATCGGTAAGTACTCAATGATAG...*

Allele B: *...CGTAGCCATCGGTAGGTACTCAATGATAG...*

PM_B: ATCGGTAGCCATCCATGAGTTACTA

(Also MMs, but not in the newer chips, so we will not use these!)



AA

PM_A >> PM_B

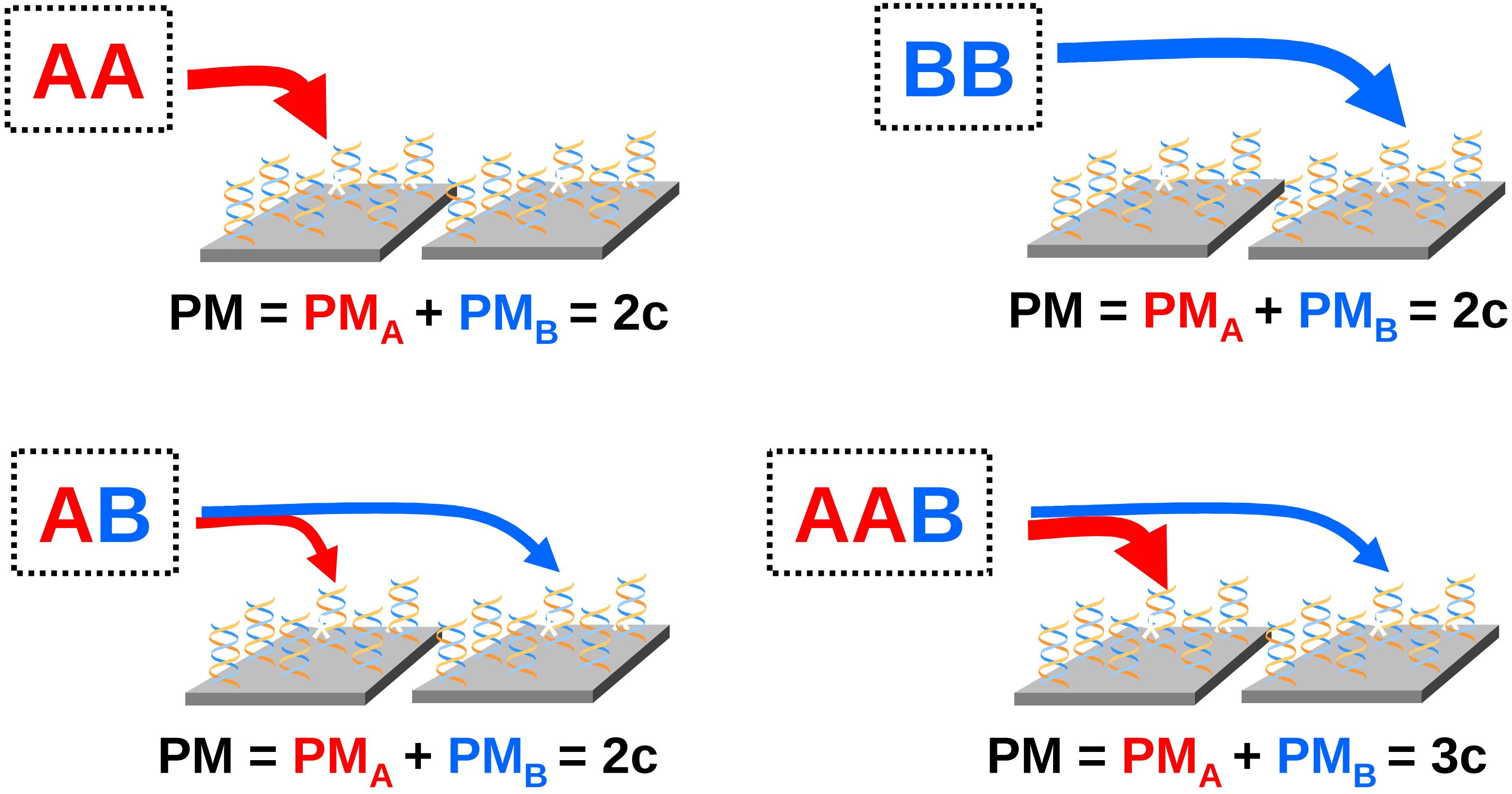
AB

PM_A ¼ PM_B

BB

PM_A << PM_B

SNP probes can also be used to estimate total copy numbers



AA

$\mathrm{PM} = \mathrm{PM}_A + \mathrm{PM}_B = 2c$

BB

$\mathrm{PM} = \mathrm{PM}_A + \mathrm{PM}_B = 2c$

AB

$\mathrm{PM} = \mathrm{PM}_A + \mathrm{PM}_B = 2c$

AAB

$\mathrm{PM} = \mathrm{PM}_A + \mathrm{PM}_B = 3c$

The Affymetrix assay
- takes 4-5 working days to complete

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

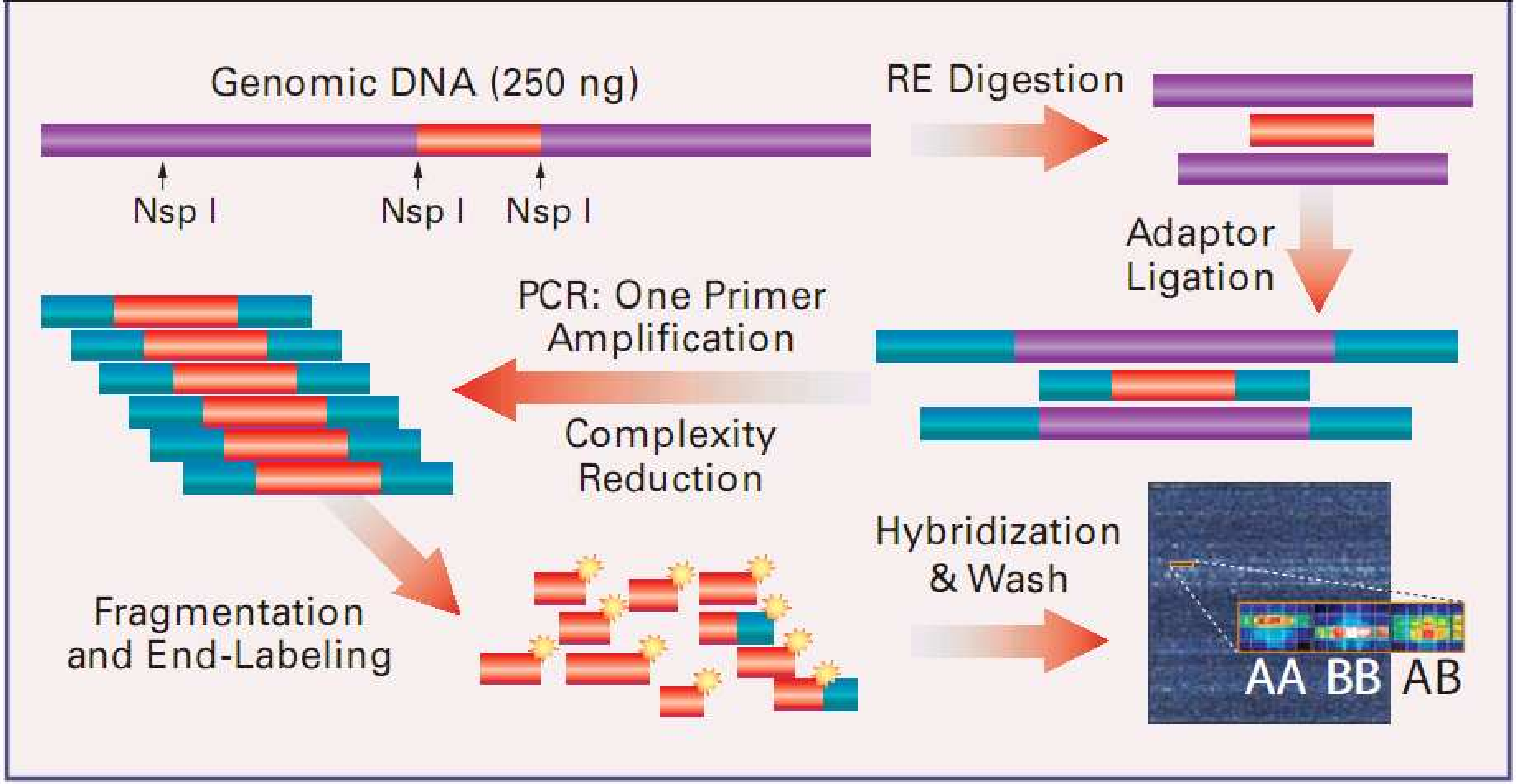
2. Obtain ***labeled single-stranded*** target DNA fragments for hybridization to the probes on the chip.

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

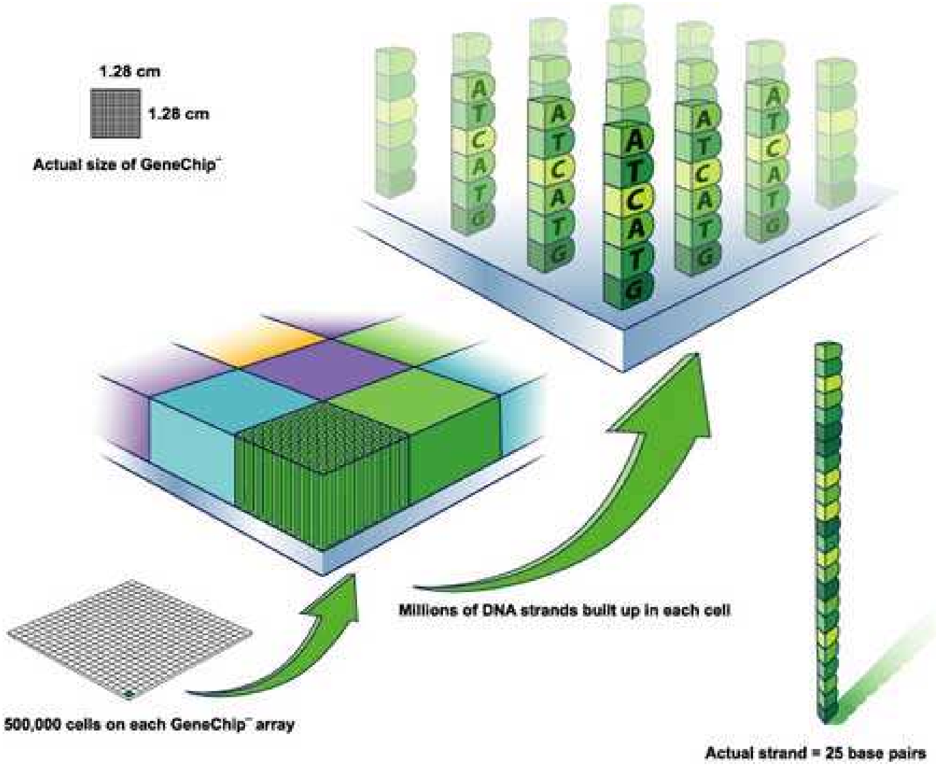
4. Image summarized across pixels to ***probe-level intensities*** before we begin. This is our "**raw data**".

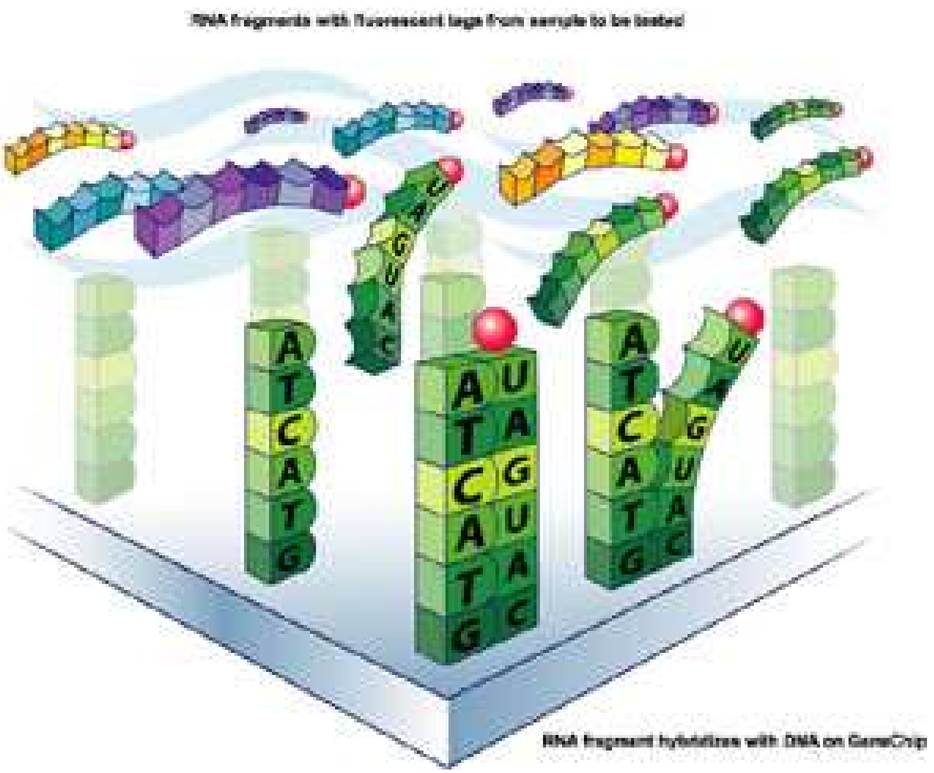
Restriction enzymes digest the DNA, which is then amplified and hybridized

Figure 1: GeneChip® Mapping Assay Overview.



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


1.28 cm
1.28 cm
Actual size of GeneChip™
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs


RNA fragments with fluorescent tags from sample to be tested
RNA fragment hybridizes with DNA on GeneChip

Scanning

Example array: 1600x1600 cells; 65536 intensity levels (16 bits).

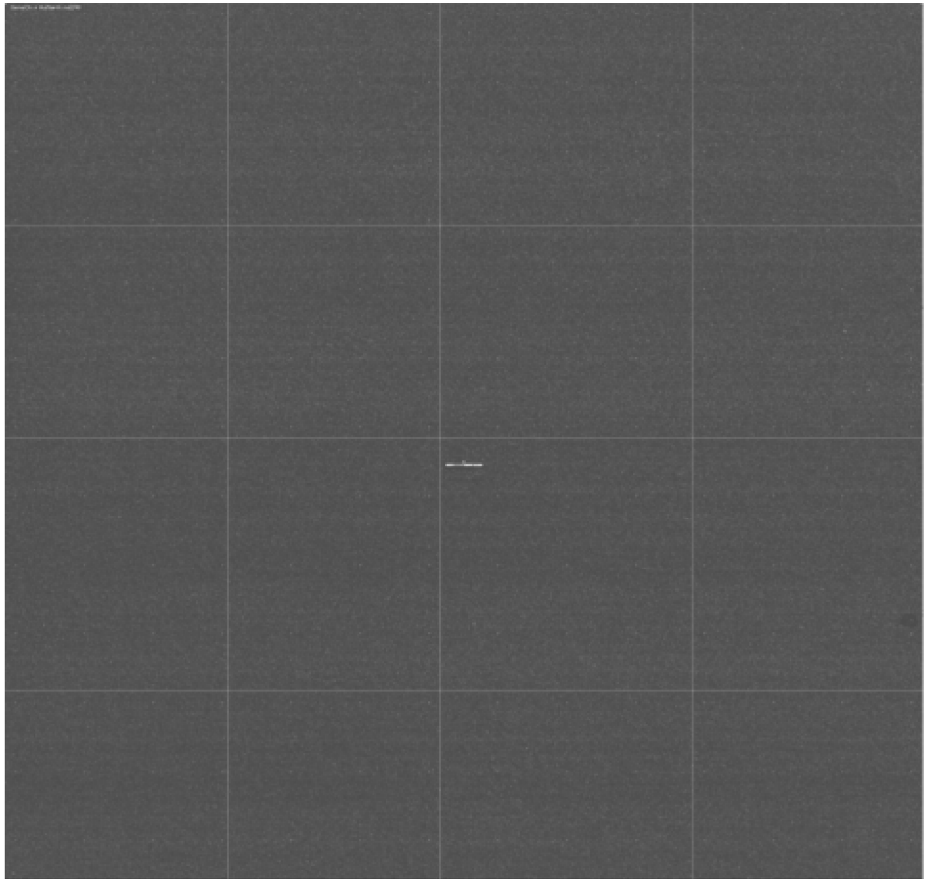


Image Analysis

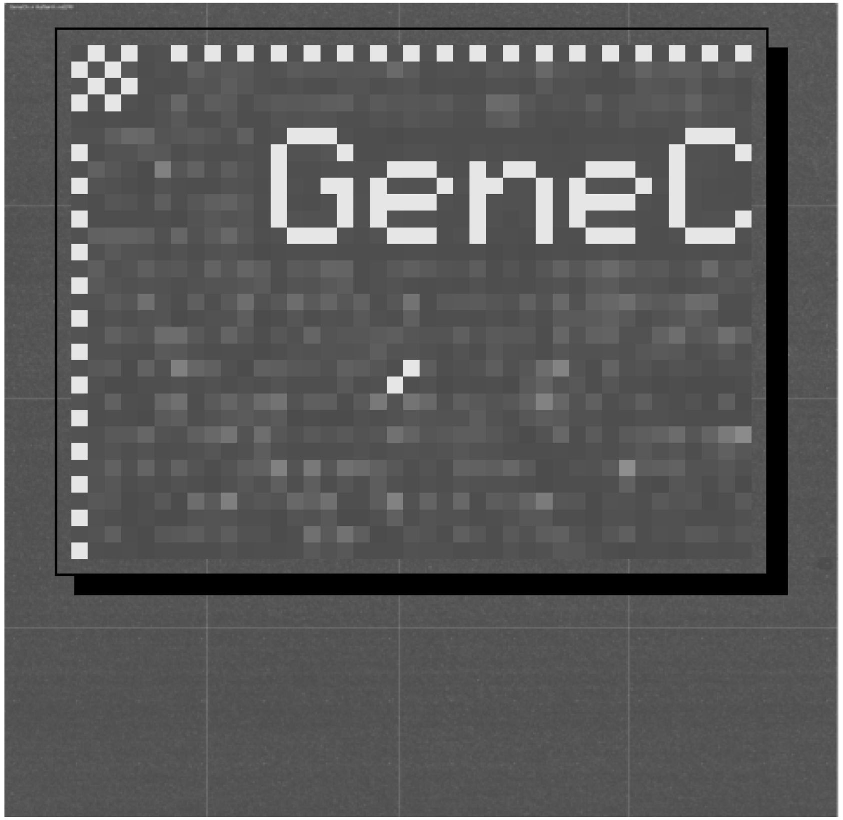
Example array:

Dimensions: 1600x1600 cells

Each cell: 3x3 pixels

Dynamic range: 65536 (16-bits) intensity levels

Cell summaries: (mean pixel, stddev pixel, #pixels)



Preparation
+ Hybridization
+ Scanning

DAT File(s)
[Image, pixel intensities]

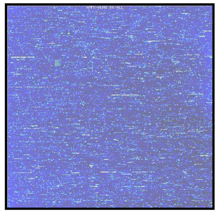


Image analysis

workable raw data

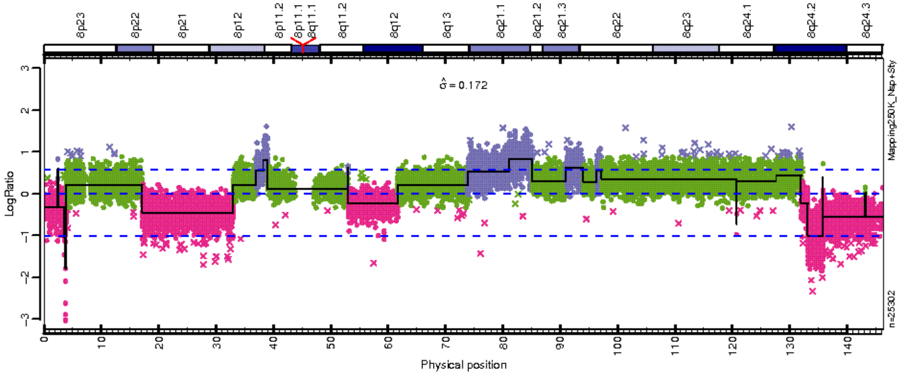
CEL File(s)
[Probe Cell Intensity]

+

CDF
[Chip Description File]

Low-level analysis

Segmentation

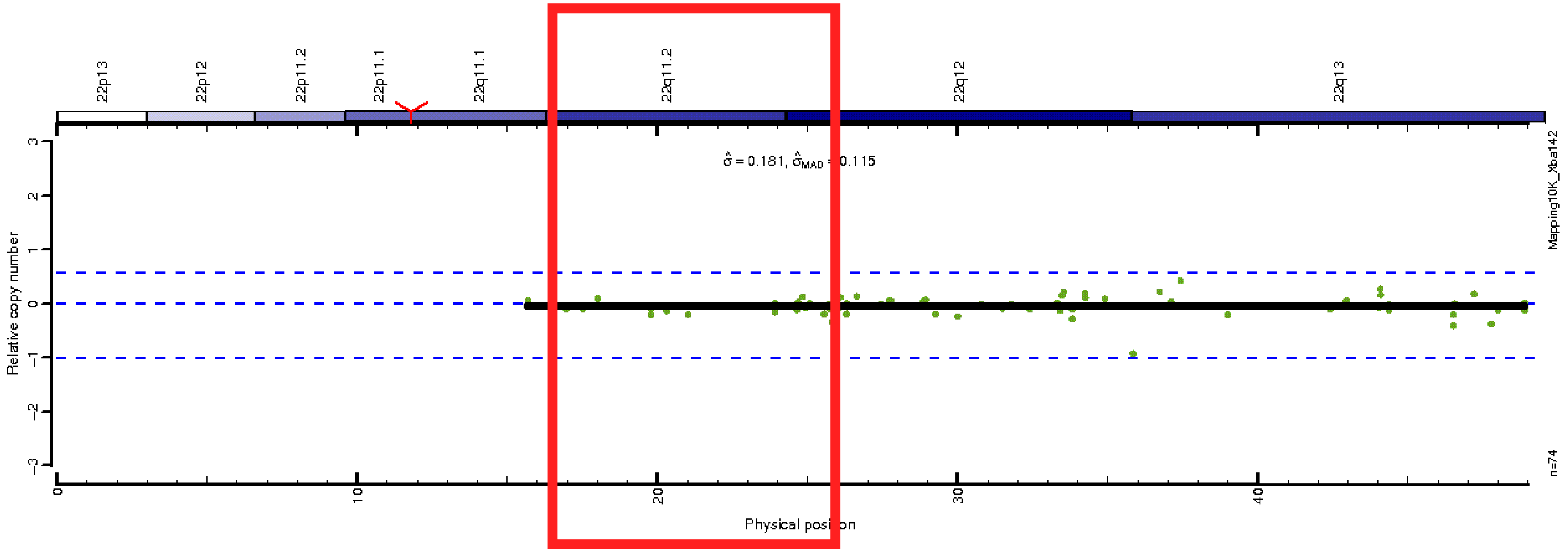


A brief history

of Affymetrix SNP & CN arrays

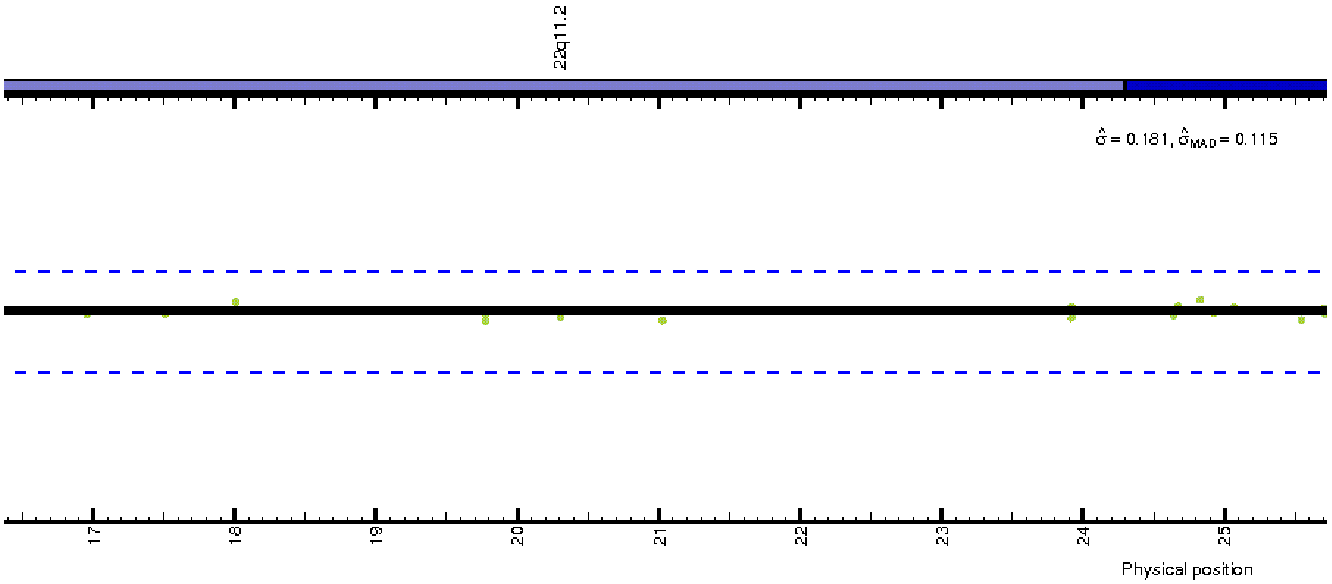
How did we get here?

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

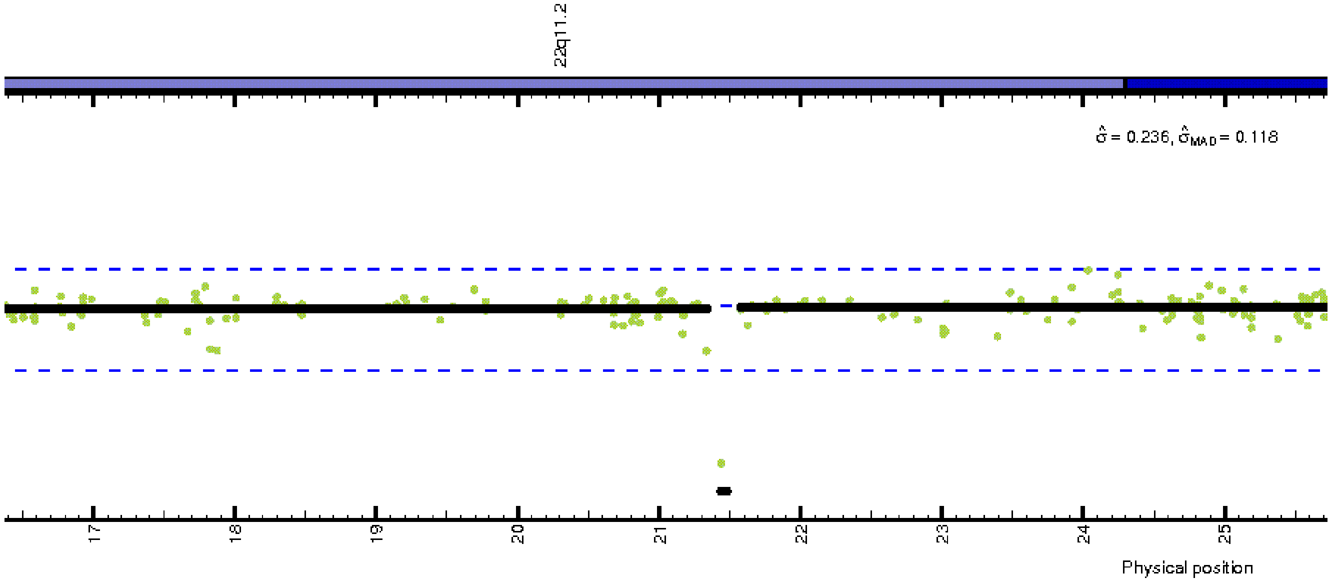


zoom in

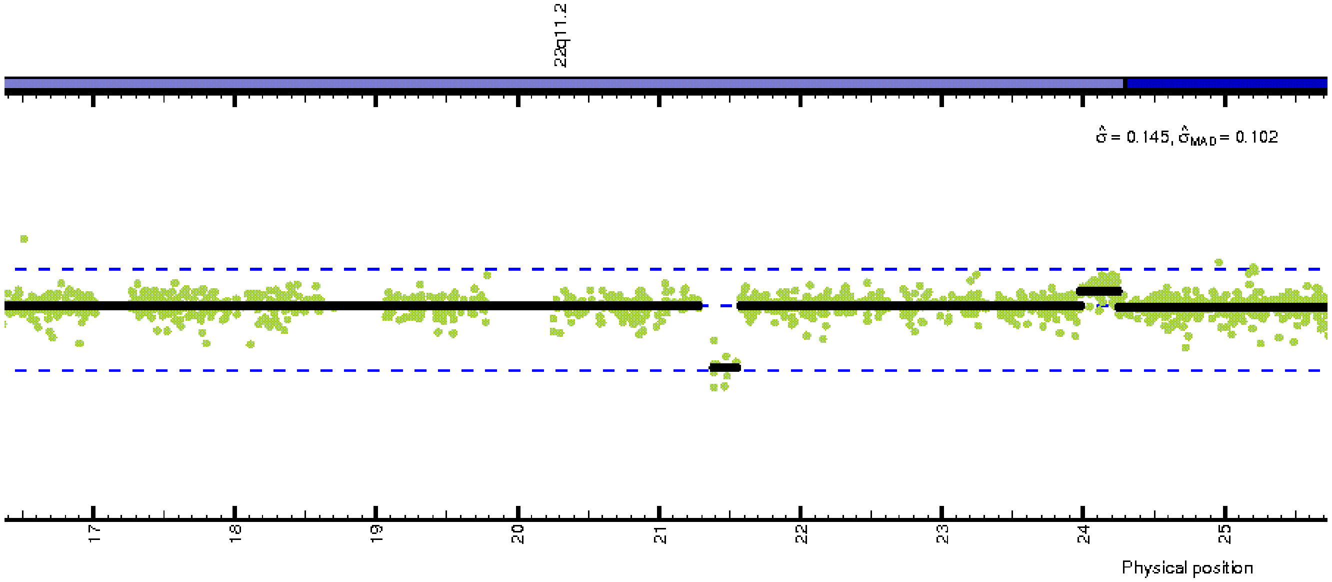
2003: 10,000 loci x1



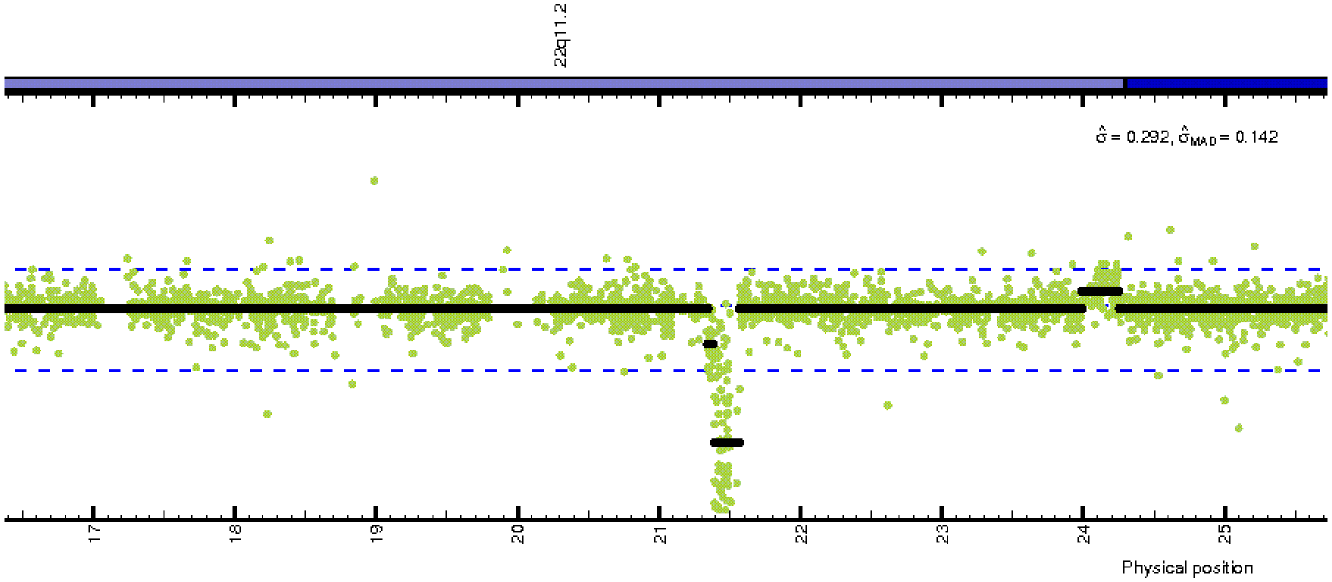
2004: 100,000 loci x10



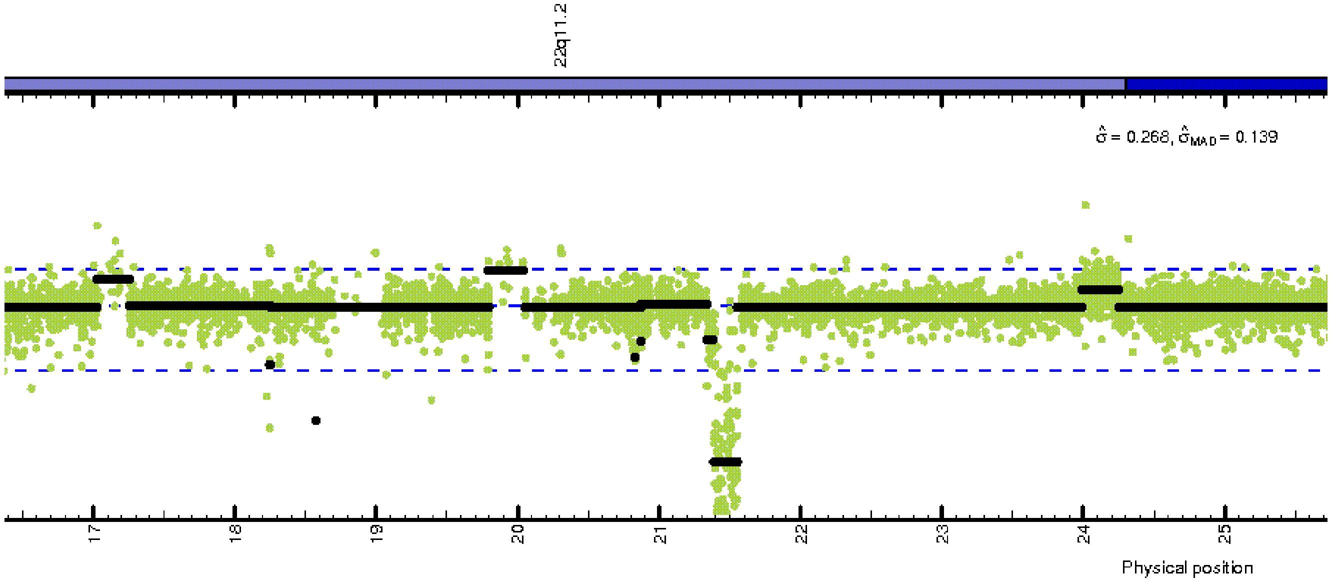
2005: 500,000 loci x50



2006: 900,000 loci   x90



2007: 1,800,000 loci x180

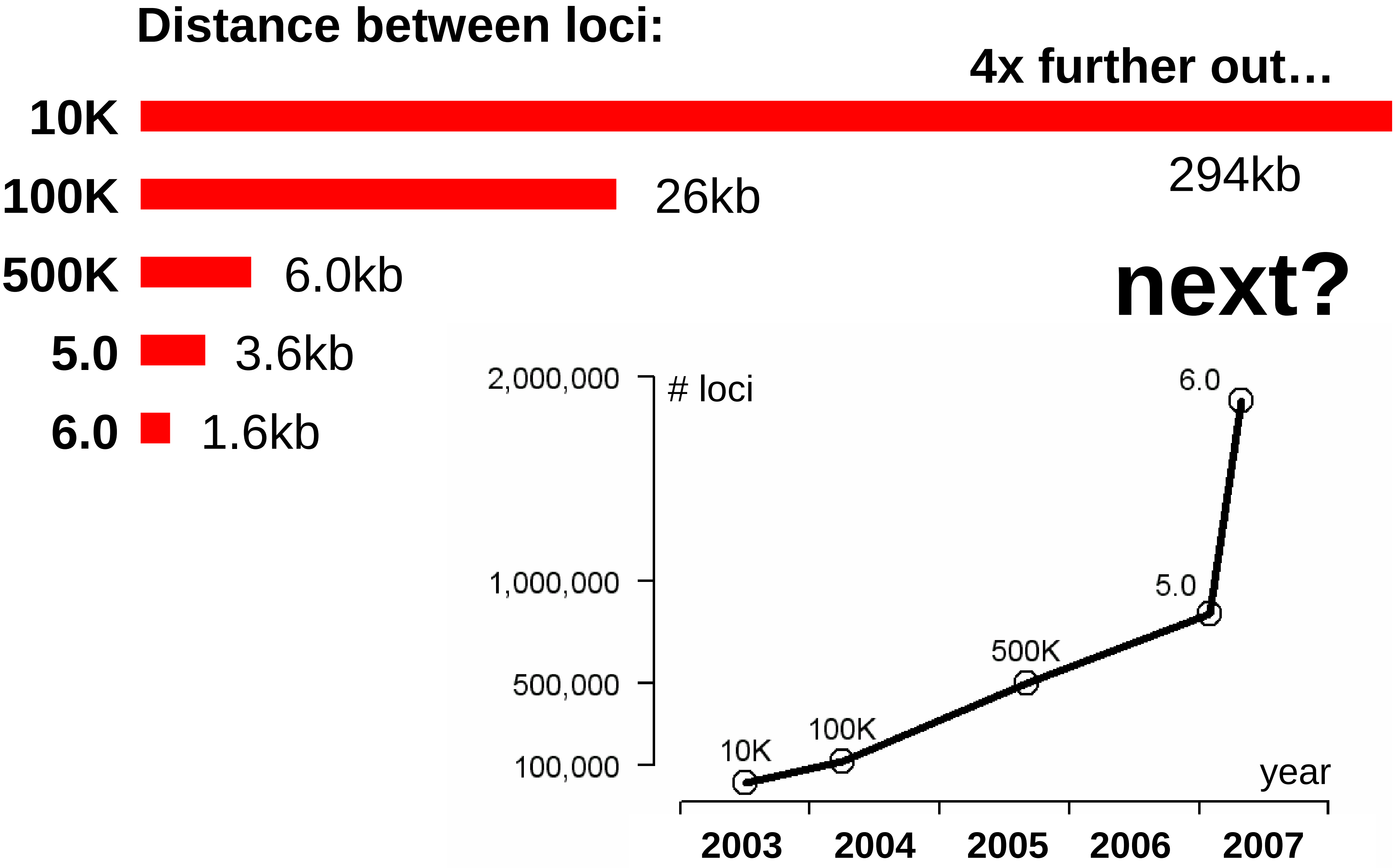


Genome-Wide Human SNP Array 6.0
- state-of-the-art array

- **> 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

Rapid increase in density



Distance between loci:

4x 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**	**300 USD**	**175 USD**	**300 USD**
# loci / cup of espresso ($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 are moving away from MM probes
- therefore we don't utilize them

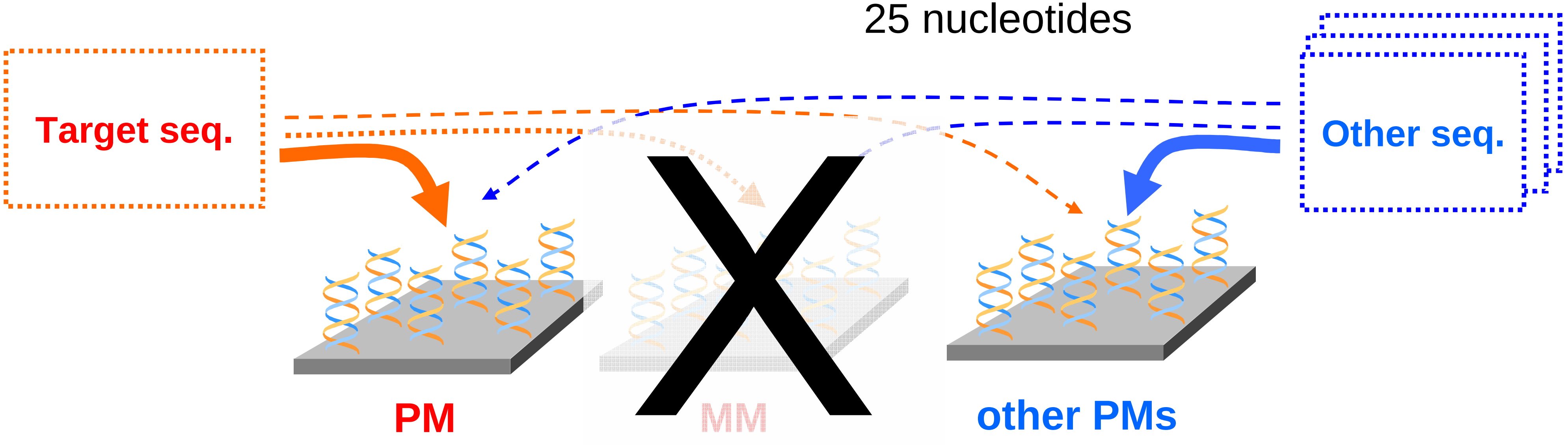
Target DNA: ***...CGTAGCCATCGGTAAGTACTCAATGATAG...***

|||||||||||||||||||||||||

Perfect match (PM): **ATCGGTAGCCATTCATGAGTTACTA**

Mis-match (MM): **ATCGGTAGCCATACATGAGTTACTA**

25 nucleotides



Low-Level
Copy Number Analysis

Part 2 – Simple preprocessing

Henrik Bengtsson

Post doc, Department of Statistics,

University of California, Berkeley, USA

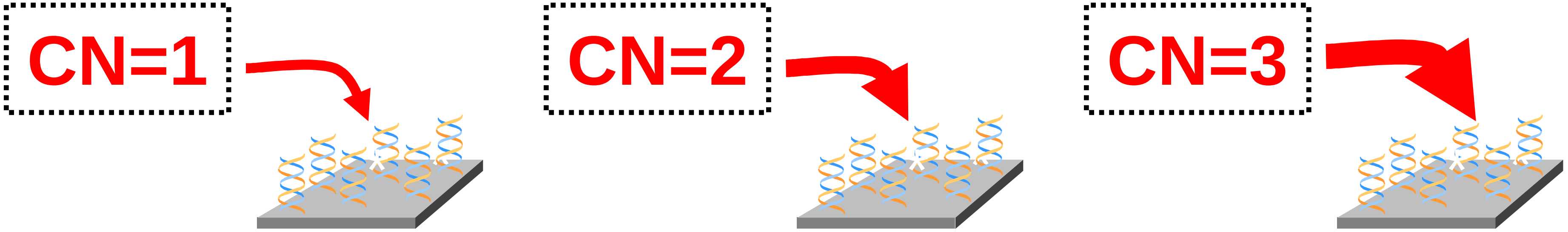
CEIT Workshop on SNP arrays,

Dec 15-17, 2008, San Sebastian

Recap: Copy-number probes

CN locus: ***...CGTAGCCATCGGTA<u>A</u>GTACTCAATGATAG...***

PM: **ATCGGTAGCCAT<u>T</u>CATGAGTTACTA**



CN=1

CN=2

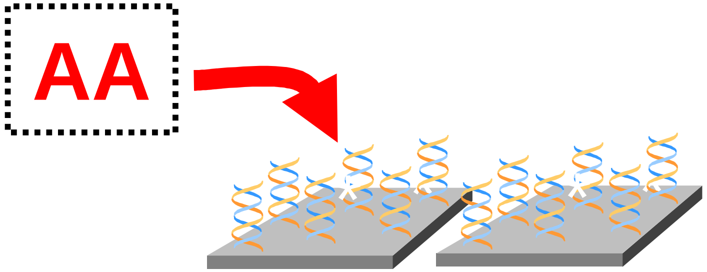
CN=3

PM = c

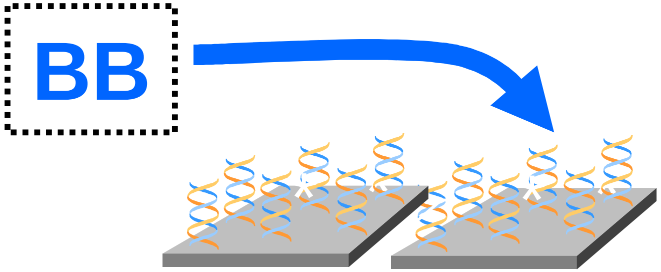
PM = 2c

PM = 3c

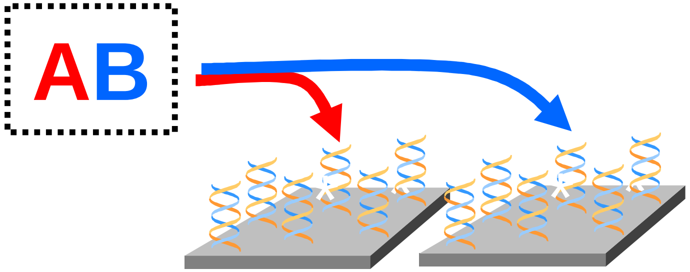
Recap: Adding SNP probes gives total CN signal



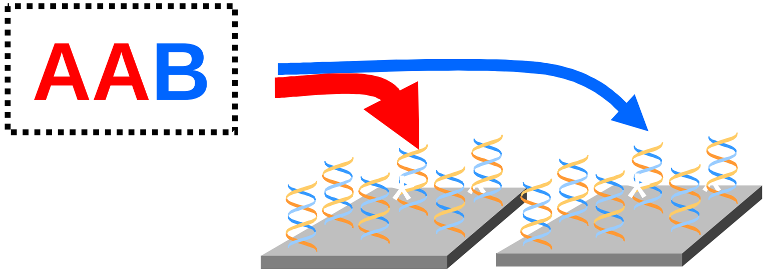
$PM = PM_A + PM_B = 2c$



$PM = PM_A + PM_B = 2c$



$PM = PM_A + PM_B = 2c$



$PM = PM_A + PM_B = 3c$

Notation
- here and in our papers

Indices:

Arrays/samples: $i = 1, 2, \ldots, I$

Loci/SNPs/CN units: $j = 1, 2, \ldots, J$

Replicated probes for SNP: $k = 1, 2, \ldots, K$

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

A simple way to 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:
    - $M_{ij} = \log_2 (\theta_{ij} / \theta_{Rj})$

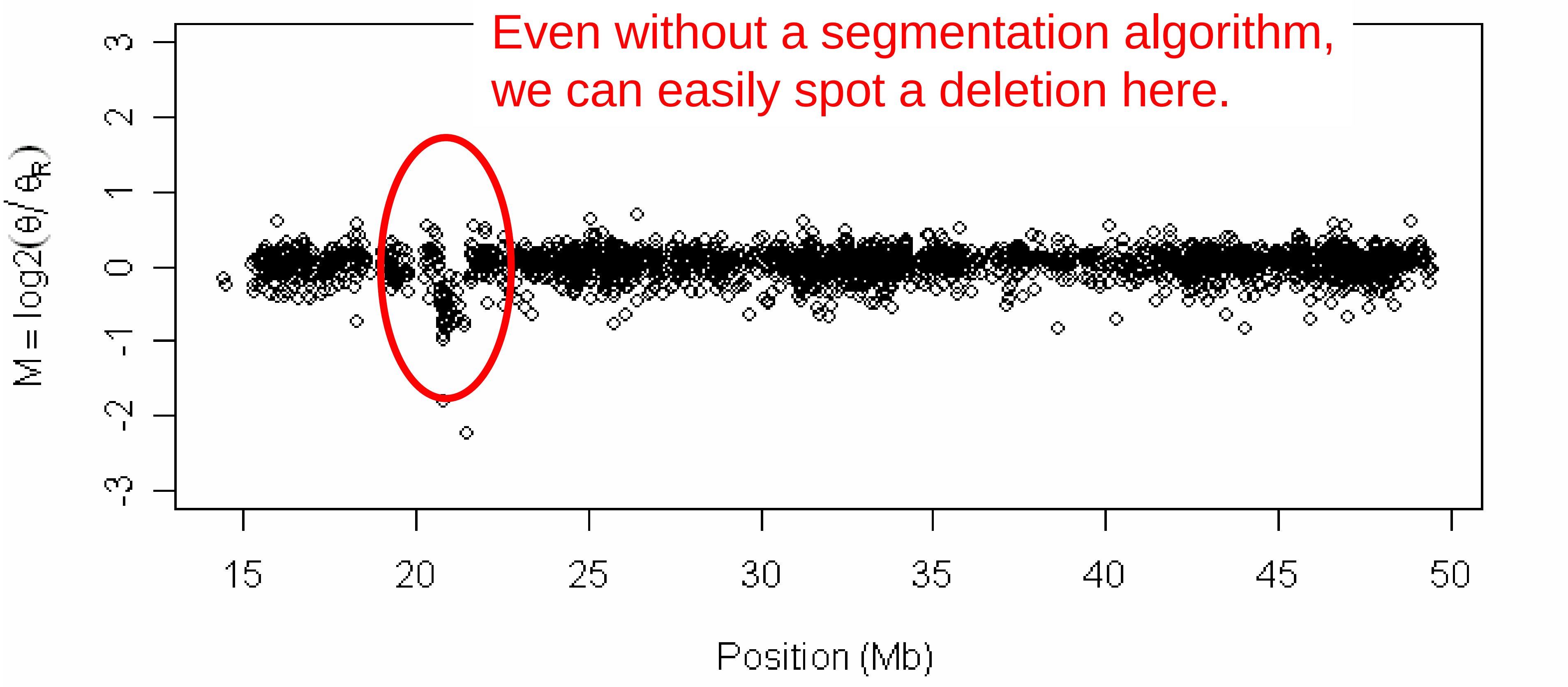
The software tools make this easy for you

- using aroma.affymetrix package

```
cs <- AffymetrixCelSet$byName(“GSE8605”,
           chipType=“Mapping10K_Xba142”);

plm <- AvgCnPlm(cs, combineAlleles=TRUE);
fit(plm);
ces <- getChipEffectSet(plm);

theta <- extractTheta(ces);
thetaR <- rowMedians(theta);
M <- log2(theta / thetaR);
```

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

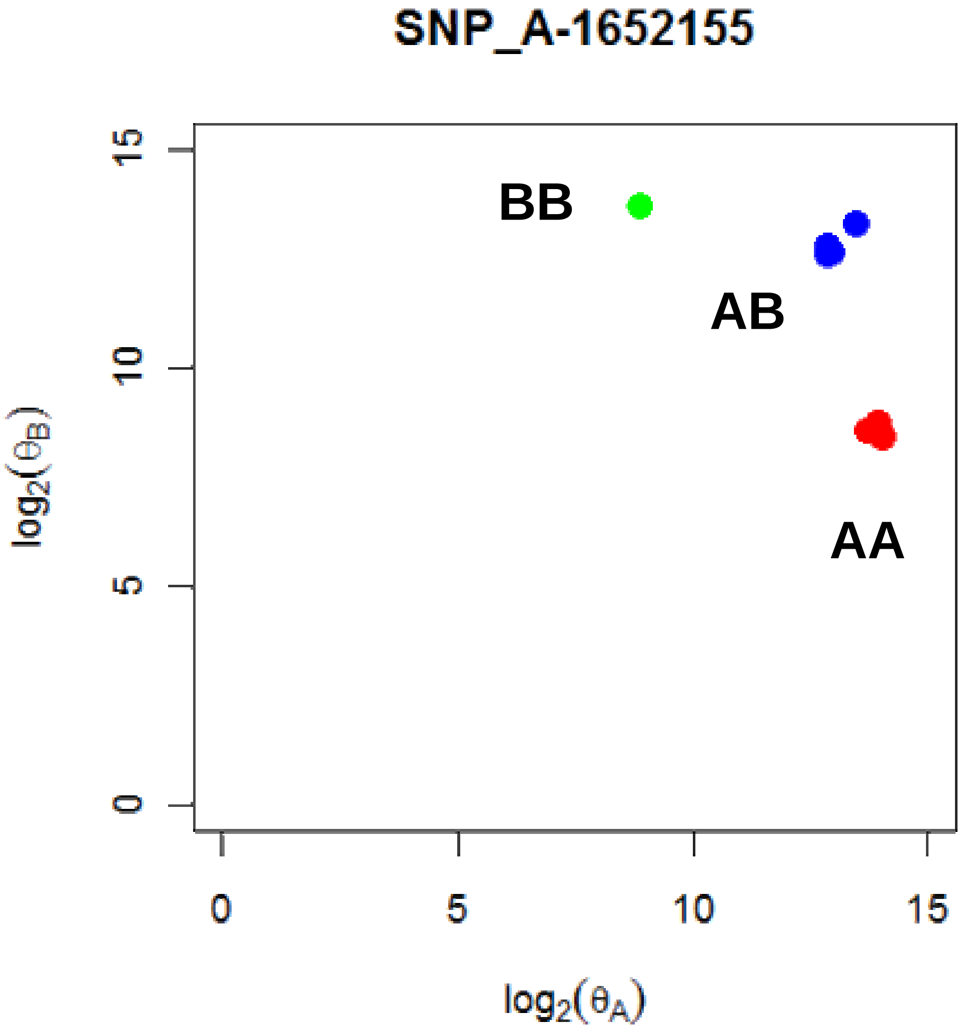
Example: Log-ratios for one sample on Chromosome 22.



Even without a segmentation algorithm, we can easily spot a deletion here.

If we don’t add up the alleles, we get allele-specific estimates from which we can get genotypes

Example: $(\theta_{ijA},\theta_{ijB})$ for one SNP across all samples

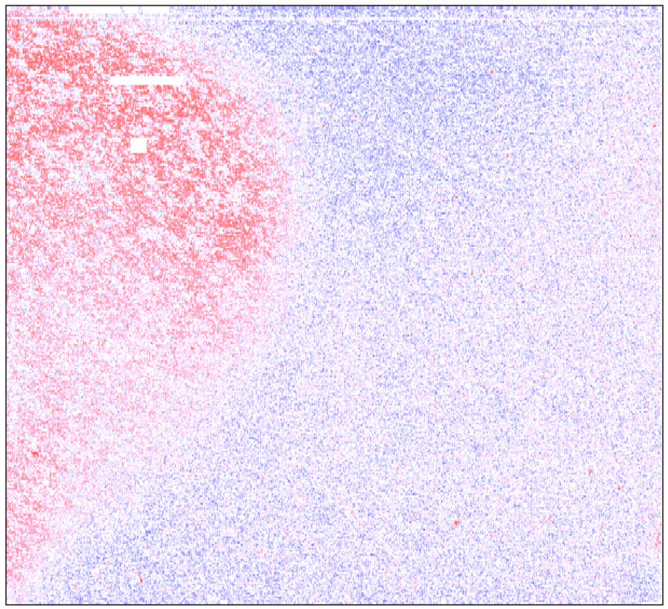


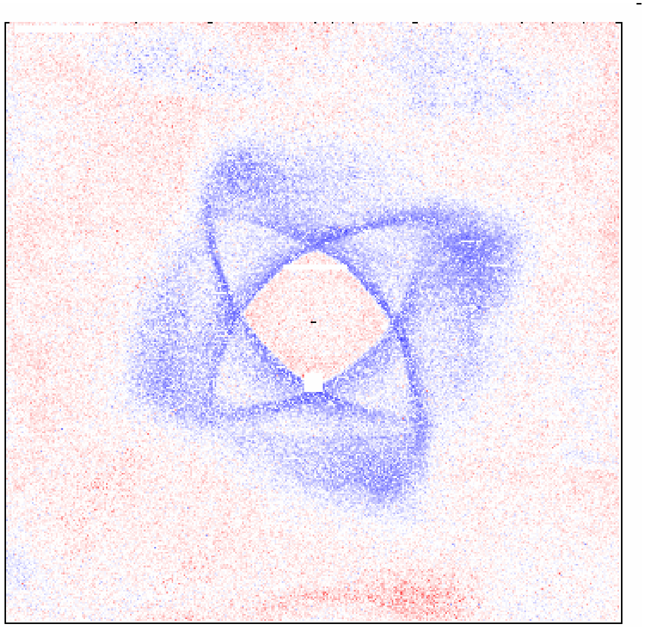
There are a lot of artifacts in microarray data
- can we do better?

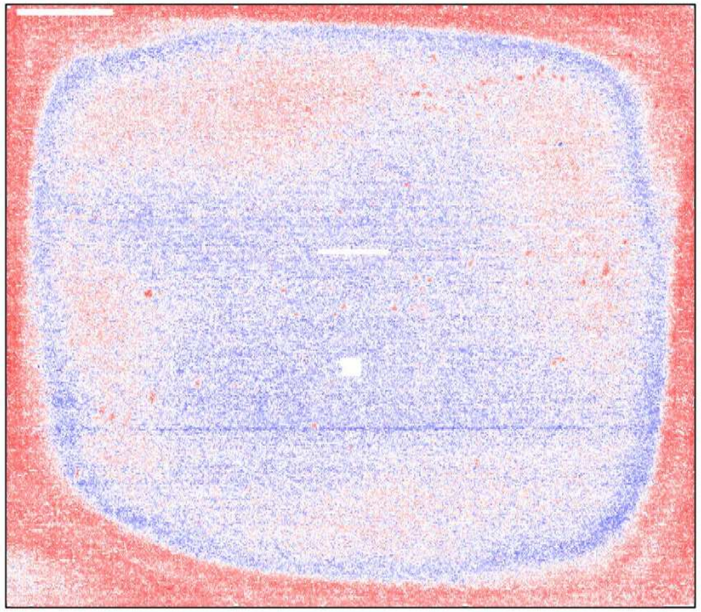
Systematic variation can be added due to:

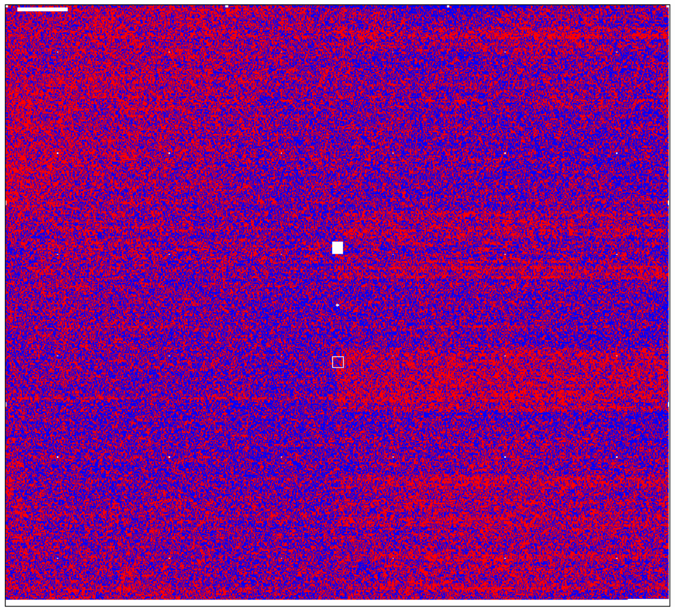
- Spatial artifacts
- Intensity dependent effects
- Probe-sequence dependent effects
- GC-content effects
- PCR effects
- Lab & people effects
- Non-calibrated scanners
- ...?

Spatial artifacts ("extreme")



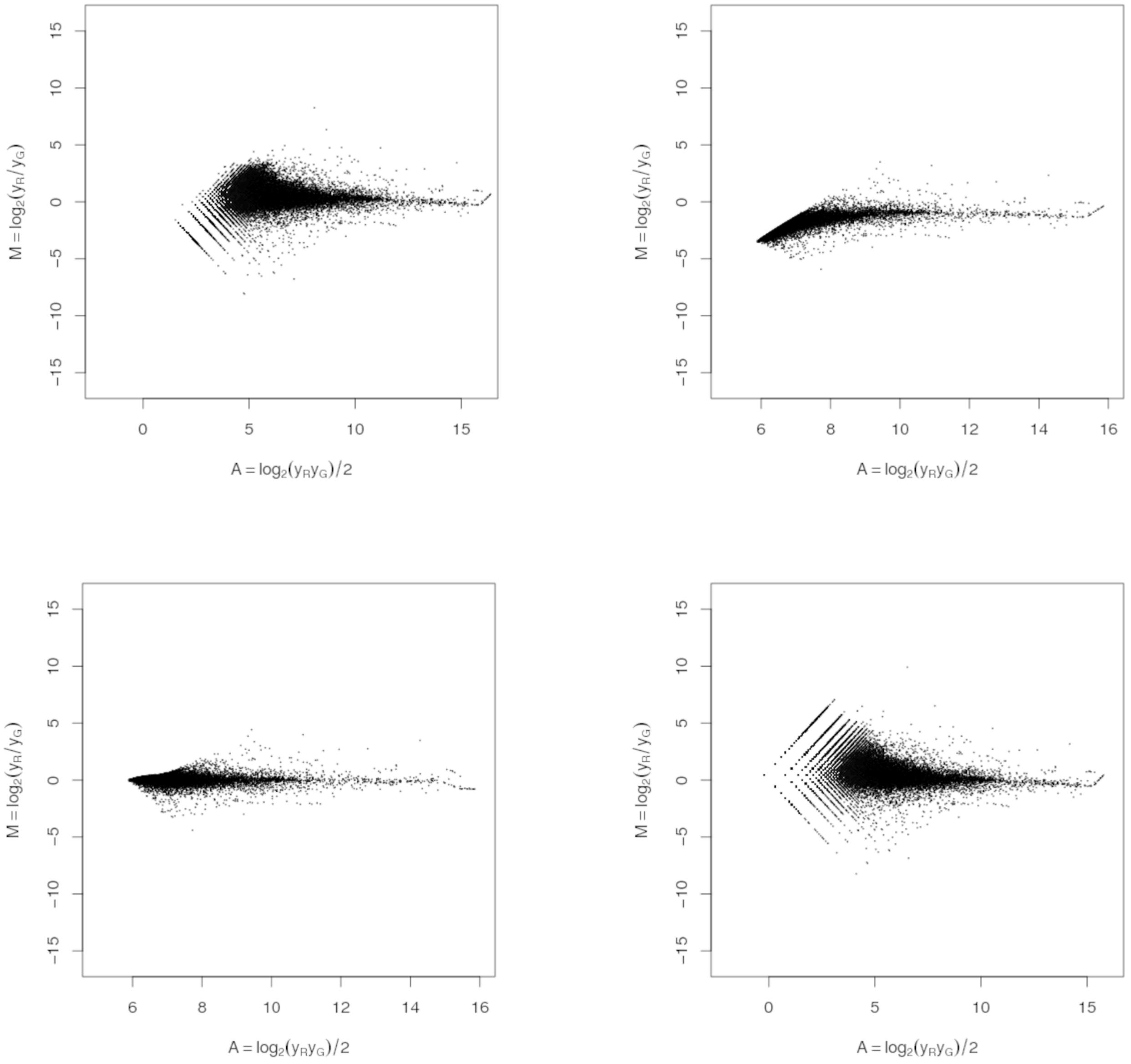




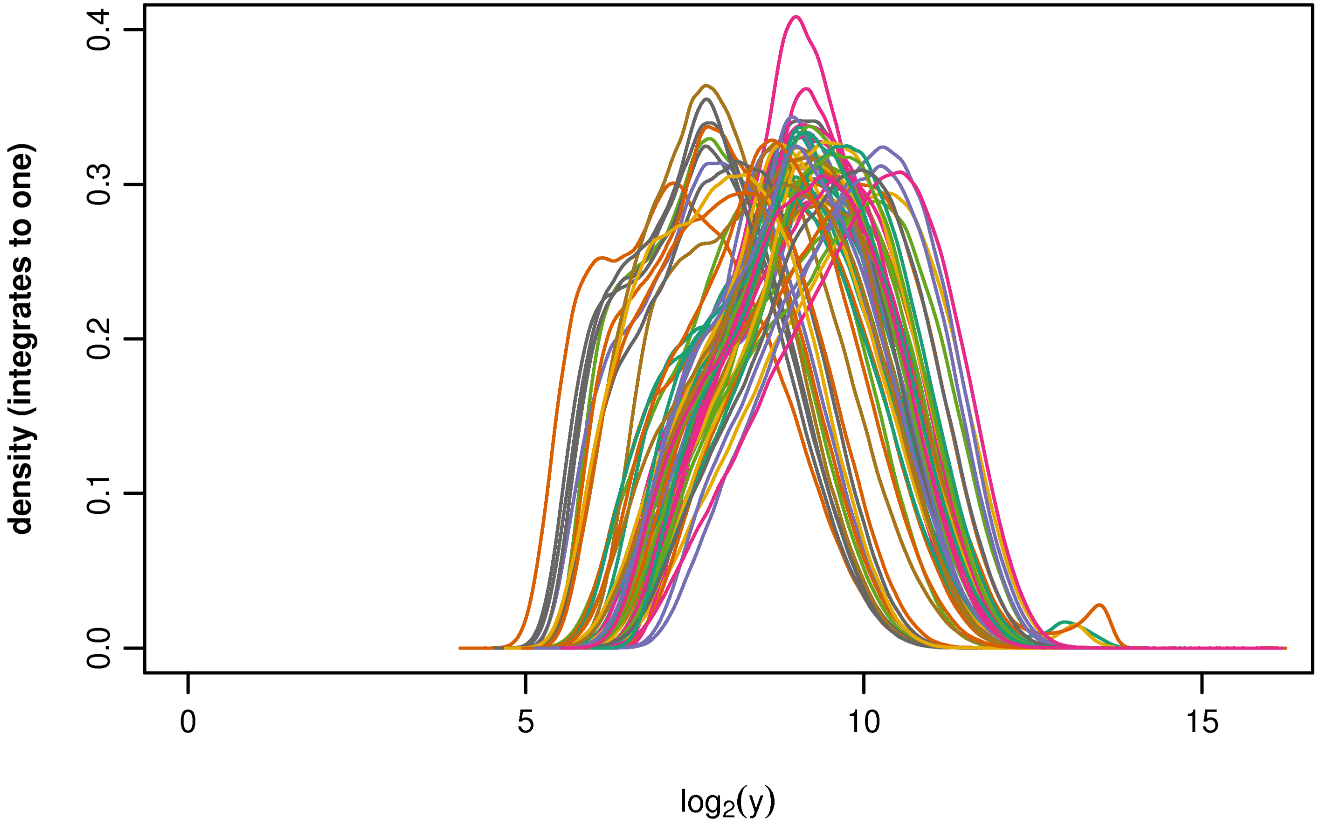


http://plmimagegallery.bmbolstad.com/

Intensity dependent artifacts/variation



Lab and people effects/variation



PCR fragment length effects/variation

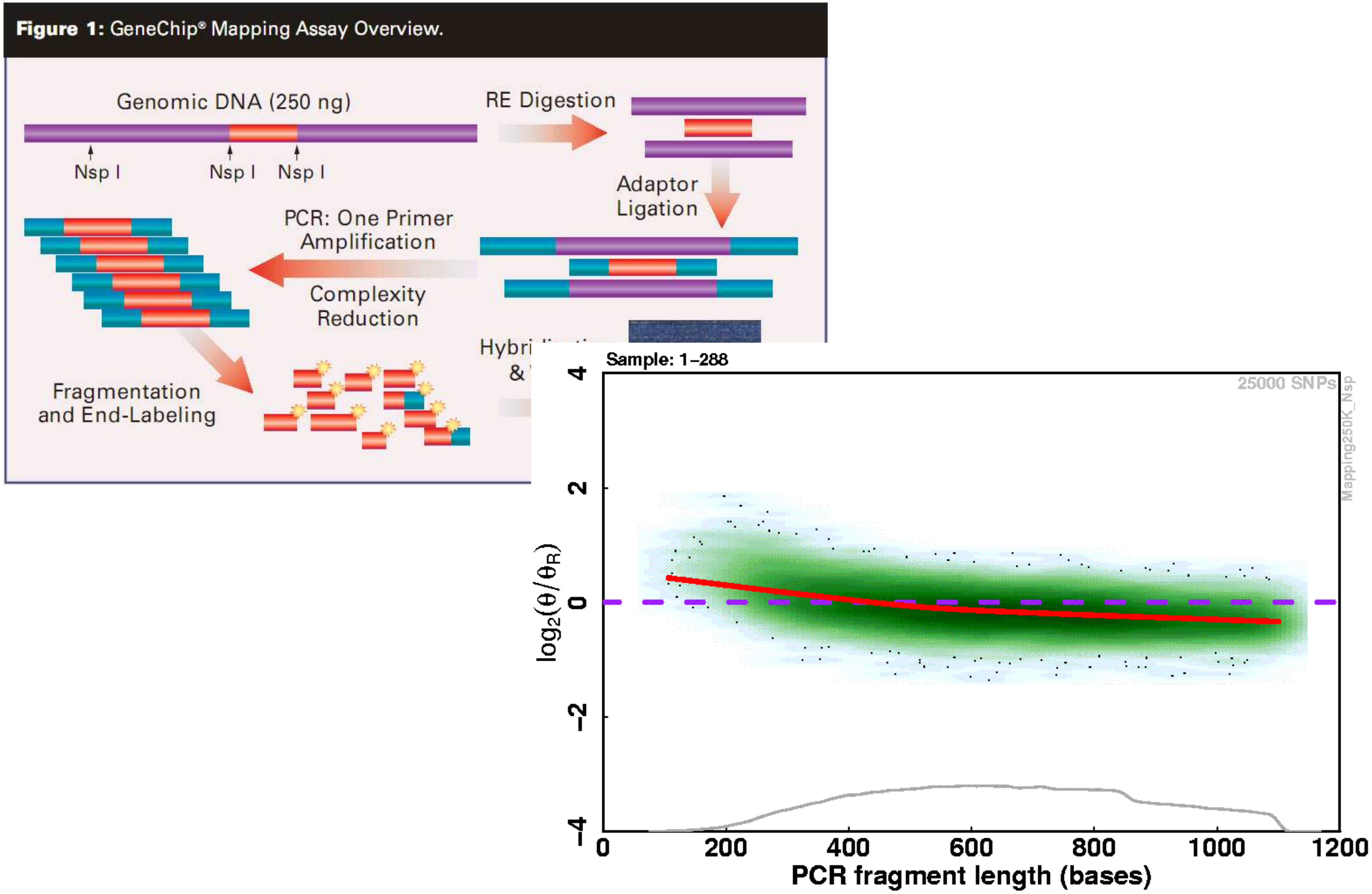
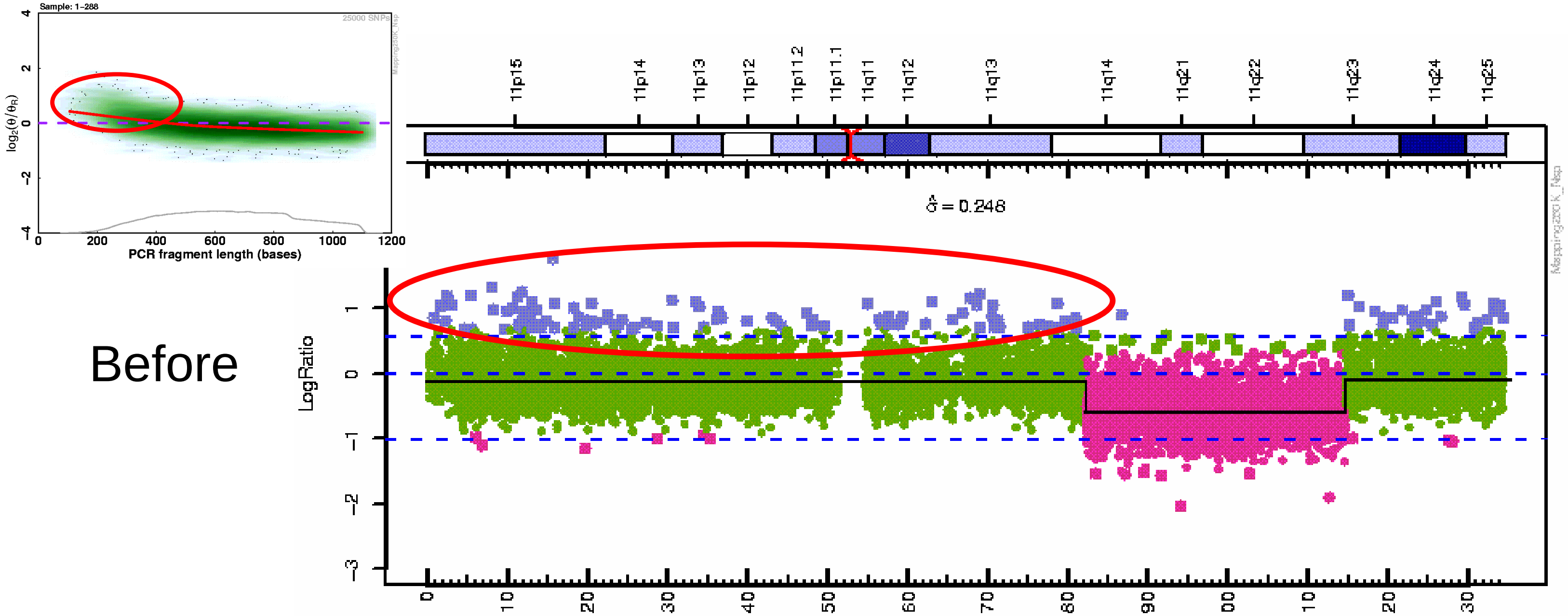


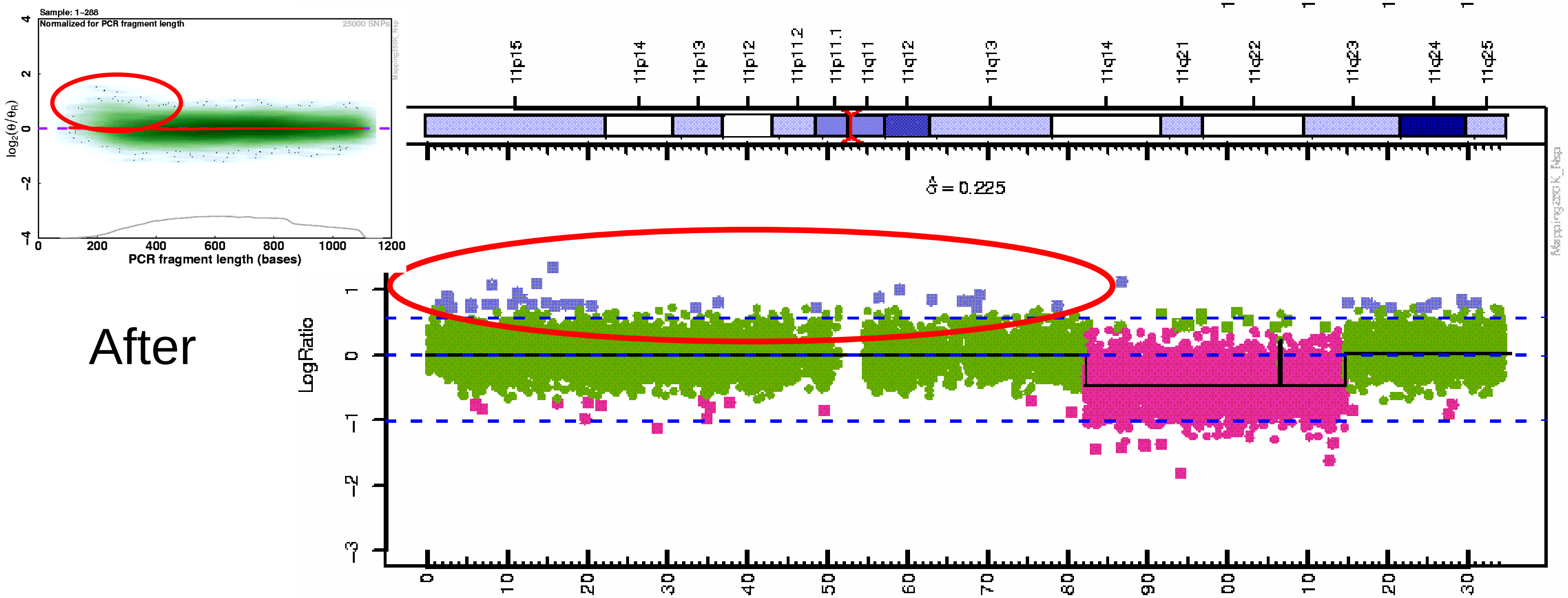
Figure 1: GeneChip® Mapping Assay Overview.

32.5Mb deletion on chr 11



Before

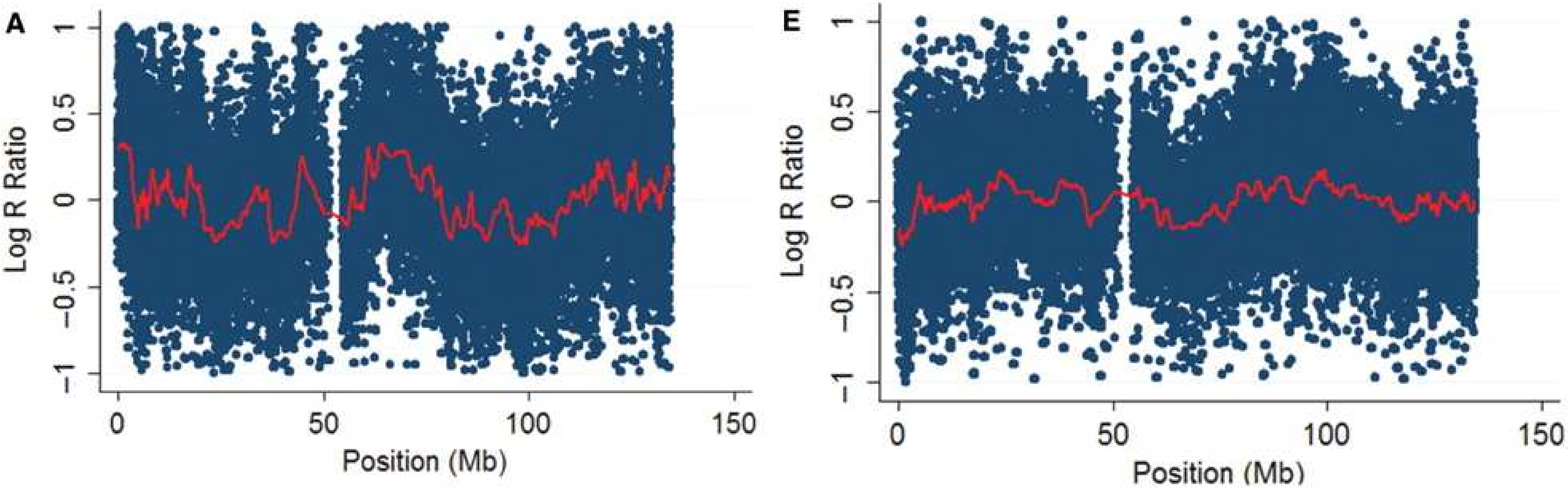
σ = 0.246



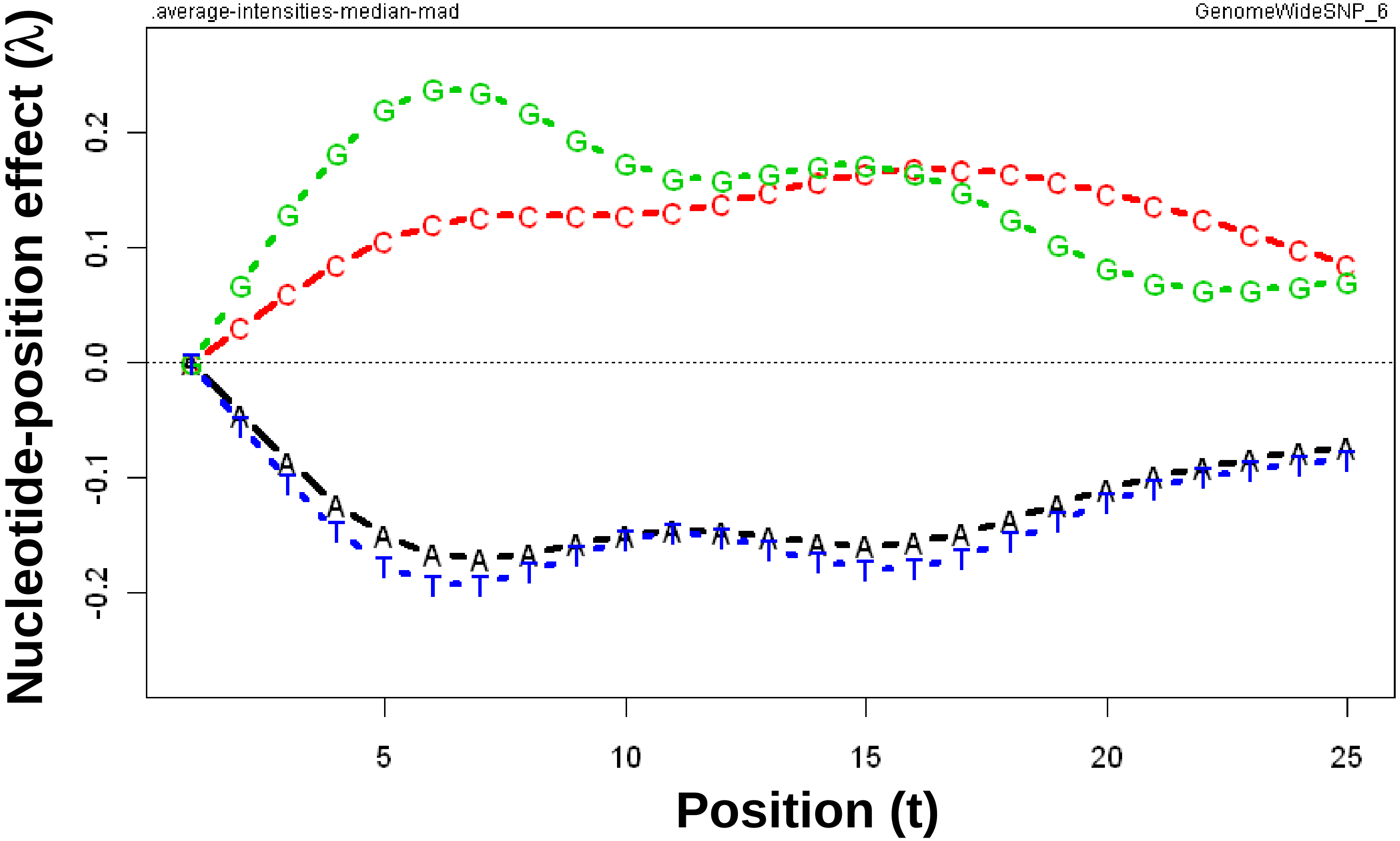
After

σ = 0.225

“Wave” patterns along genome



Probe-sequence effects/variation
- probes respond differently



Low-Level Copy Number Analysis

Part 3 – aroma.affymetrix

Henrik Bengtsson

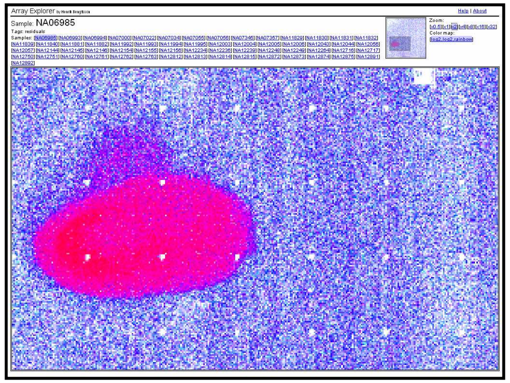
Post doc, Department of Statistics,

University of California, Berkeley, USA

CEIT Workshop on SNP arrays,

Dec 15-17, 2008, San Sebastian

aroma.affymetrix processes unlimited number of arrays



Array Explorer by Henrik Bengtsson

Help | About

Sample: NA06985

Tags: residuals

Samples: [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]

Zoom: [x0.5] [x1] [x2] [x4] [x8] [x16] [x32]

Color map: [log2,log2,rainbow]

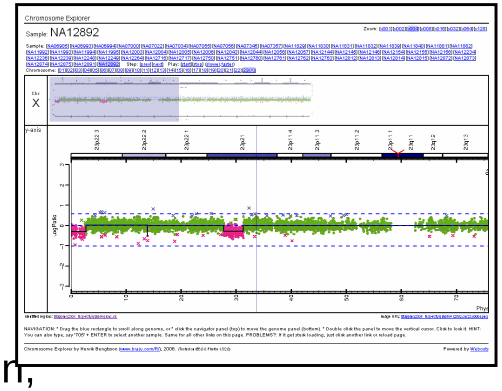
- Processes **unlimited number of arrays**:
  - Bounded memory algorithms.
  - Works toward file system.
  - Persistent memory: robust & picks up where last stopped.

- Memory requirements: **1.0-2.0GB RAM**.
  - Example: RMA on 4500 HG-U133A arrays uses ~500MB of RAM.
  - Example: CRMA on 300 SNP6.0 arrays uses ~1.5GB of RAM.
  - Example: FIRMA on 200 HuEx-1.0 arrays uses ~1.5GB of RAM.

- **Cross platform**: Linux/Unix, Windows, OSX.

- Supports **most Affymetrix chip types**:
  - All chip types with a CDF (and some more).
  - Custom CDFs.

aroma.affymetrix "implements" several existing methods



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

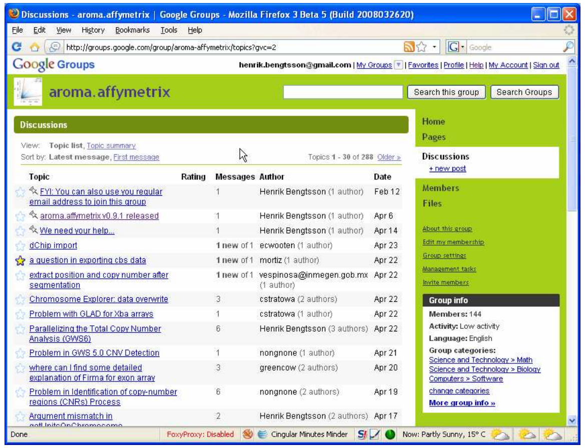
NAVIGATION: * Drag the blue rectangle to scroll along genome, or * click the navigator panel (top) to move the genome panel (bottom). * Double click the panel to move the vertical cursor. Click to lock it. HINT: You can also type, say 'T05' + ENTER to select another sample. Same for all other links on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

- **Calibration and normalization**:
  - Background correction methods: RMA, gcRMA, ...
  - Allelic cross-talk calibration, quantile normalization, spatial normalization, probe-sequence normalization, …
  - PCR fragment-length normalization, GC-content normalization.
- **Probe-level summarization**:
  - multiplicative (dChip), affine, and log-additive (RMA) models. Easy to add new.
- **Quality assessment**:
  - RLE (*Relative Log Expressions)*, NUSE (Normalized Unscaled Standard Error)
  - Spatial plots: probe signals, PLM residuals, chip effects, CDF annotations, ...
- Paired & non-paired **copy-number analysis**:
  - All SNP & CN platforms. Multiple chip types.
  - CRMA (our methods for estimating raw CNs).
  - Allele-specific and/or total CN estimates
  - Genotyping via CRLMM
  - Segmentation method: CBS & GLAD. Easy to add more.
- **Miscellaneous**:
  - ***Alternative splicing (exon arrays)***: Finding Isoforms using RMA (FIRMA)
  - ***Tiling-array analysis****:* MAT processing
  - ***Resequencing arrays***
  - ***Gene expression arrays*** (of course)

aroma.affymetrix is an open-source solution


Discussions - aroma.affymetrix | Google Groups - Mozilla Firefox 3 Beta 5 (Build 2008032620)
File Edit View History Bookmarks Tools Help
http://groups.google.com/group/aroma-affymetrix/topics?gvc=2
Google
Google Groups
henrik.bengtsson@gmail.com | My Groups | Favorites | Profile | Help | My Account | Sign out
aroma.affymetrix
Search this group
Search Groups
Home
Pages
Discussions
+ new post
Members
Files
About this group
Edit my membership
Group settings
Management tasks
Invite members
Done
FoxyProxy: Disabled
Cingular Minutes Minder
Now: Partly Sunny, 15° C

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		2	Henrik Bengtsson (2 authors)	Apr 17

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 »

- Community:
  - **250+ users** worldwide (approx 10 installation per day)
  - Active mailing list (Google Groups; 150+ message per month)
  - **Collaborative documentation** and vignettes (Google Groups).
- Development:
  - **~3 years** since start:
  - Jan 2006-Oct 2006: Phase I: **Identifying API** (1-3 person project).
  - Oct 2006-Feb 2007: Phase II: **Maturing API** and testing (10-15 users & developers).
  - Feb 2007-Aug 2007: Phase III: **Extended real-world testing** (30-50 users & developers).
  - Aug 2007-Fall 2008: Phase IV: **Public release** and heaps of CPU mileage (more "wild" use cases).
  - Fall 2008-… Phase V: Third party extensions are coming in. More chip types supported.

  - 3-5 active developers. One main maintainer / code coordinator.
  - Some external code snippet contributors.

  - Lots of **validation code** - catches existing and future bugs.
  - 1,000+ pages code / Rd pages.
- Standards:
  - **Standard file formats**, e.g. reads/writes CEL (via **affxparser**/Fusion SDK).
  - Imports and exports to: APT, GTC, CNAG, CNAT, dChip, Bioconductor etc.
  - **Strict directory structures** and relative pathnames
=> portable scripts, robustness, more automated validations, easier to troubleshoot, simplified support.

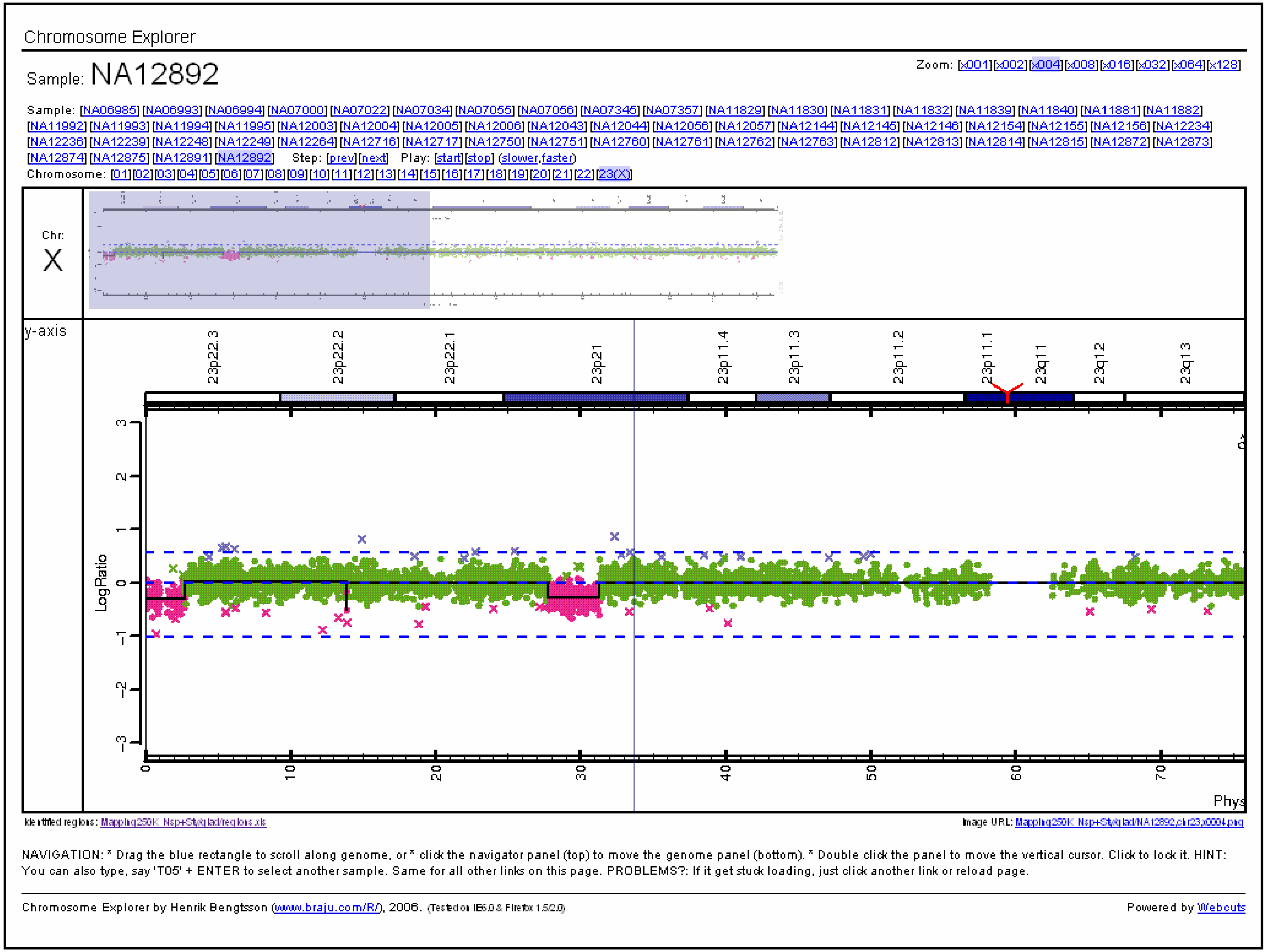
  - Utilizes **existing packages**, e.g. preprocessCore, gcRMA, DNAcopy...

Walk-through example

Complete aroma.affymetrix script for copy-number analysis of 270 SNP6.0 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, combineAlleles=TRUE)
fit(plm)

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

seg <- CbsModel(cesN)

ce <- ChromosomeExplorer(seg)
process(ce)
```

Offline & online dynamic HTML reports
Example: ChromosomeExplorer



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

Setup is as simple as placing the files in
a strict & standardized directory structure

```
annotationData/
  chipTypes/
    GenomeWideSNP_6/
      GenomeWideSNP_6.CDF
      GenomeWideSNP_6.UGP
      GenomeWideSNP_6.UFL

rawData/
  HapMap270,CEU/
     GenomeWideSNP_6/
       *.CEL
```

No (absolute) pathnames are used
- maximizes portability

```
annotationData/
  chipTypes/
    GenomeWideSNP_6/
      GenomeWideSNP_6.CDF  GenomeWideSNP_6.UGP  ...

cdf <- AffymetrixCdfFile$byChipType("GenomeWideSNP_6")
print(cdf)

AffymetrixCdfFile:
Path: annotationData/chipTypes/GenomeWideSNP_6
Filename: GenomeWideSNP_6.cdf
Filesize: 470.44MB
File format: v4 (binary; XDA)
Chip type: GenomeWideSNP_6
Dimension: 2572x2680
Number of cells: 6892960
Number of units: 1881415
...
```

The file system is the memory
- data is loaded only when needed

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

AffymetrixCelSet:
Name: HapMap270
Tags: CEU
Path: rawData/HapMap270,CEU/GenomeWideSNP_6
Chip type: GenomeWideSNP_6
Number of arrays: 270
Names: NA06985, NA06991, ..., NA07019
Total file size: 17.7GB
RAM: 0.01MB
```

Normalized data is stored as CEL files
- import to any software

```
acc <- AllelicCrosstalkCalibration(csR)
csC <- process(acc)
print(csC)
AffymetrixCelSet:
Name: HapMap270
Tags: CEU,ACC,ra,-XY
Path: probeData/HapMap270,CEU,ACC,ra,-XY/GenomeWideSNP_6
Chip type: GenomeWideSNP_6
Number of arrays: 270
Names: NA06985, NA06991, ..., NA07019
Total file size: 17.7GB
RAM: 0.01MB

files <- getPathnames(csC)
print(files[1])
[1] "probeData/HapMap270,CEU,ACC,ra,-XY/
             GenomeWideSNP_6/NA06985.CEL"
```

Data sets (directories) are marked with unique tags

```
qn <- QuantileNormalization(csC)
csN <- process(qn)
print(csN)
```

```
AffymetrixCelSet:
Name: HapMap270
Tags: CEU,ACC,ra,-XY,ACC,QN
Path: probeData/HapMap270,CEU,ACC,ra,-XY,QN/GenomeWideSNP_6
Chip type: GenomeWideSNP_6
Number of arrays: 270
Names: NA06985, NA06991, ..., NA07019
Total file size: 17.7GB
RAM: 0.01MB
```