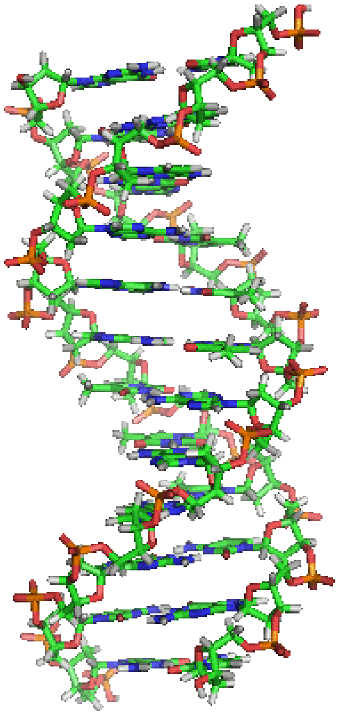
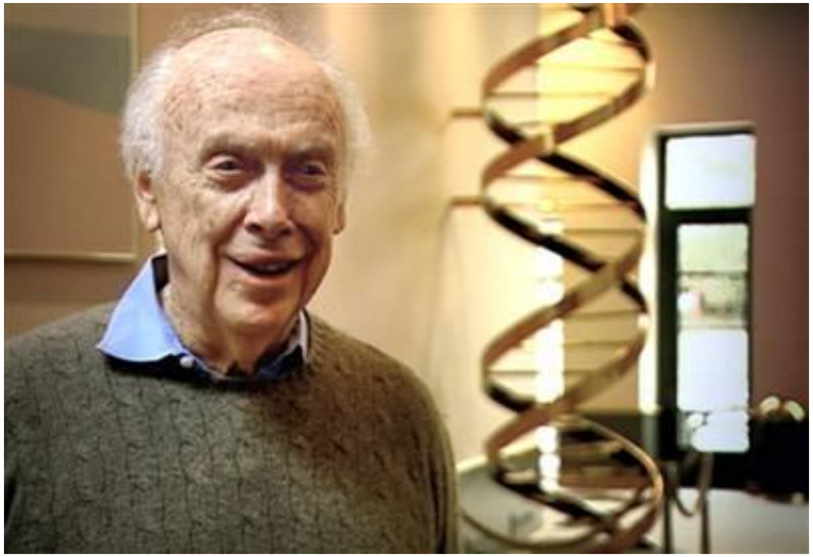
Discovering the buildup of the Human Genome



U. Menzel, Berlin 2009-09-10

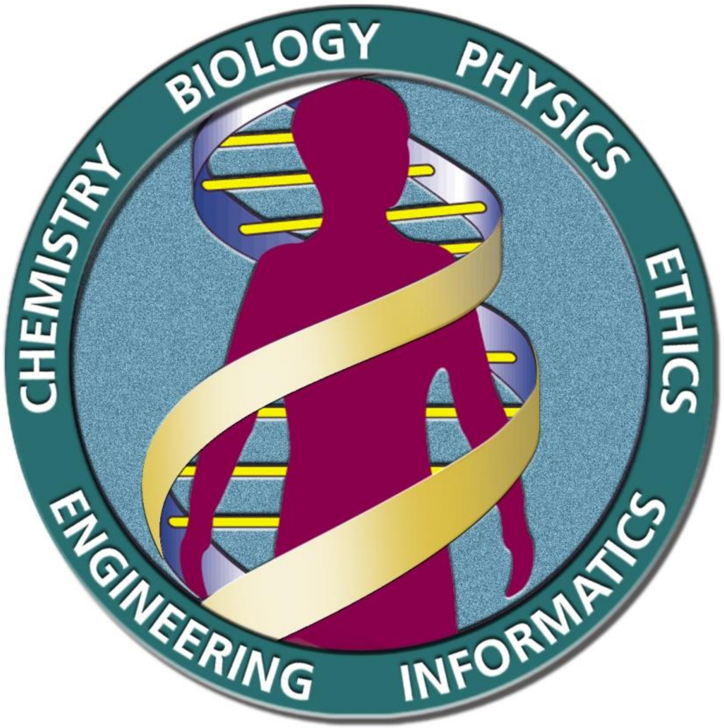
We used to think our fate was in our stars.
Now we know, in large measure, our fate is in our genes.

James Watson, 1989



HUGO

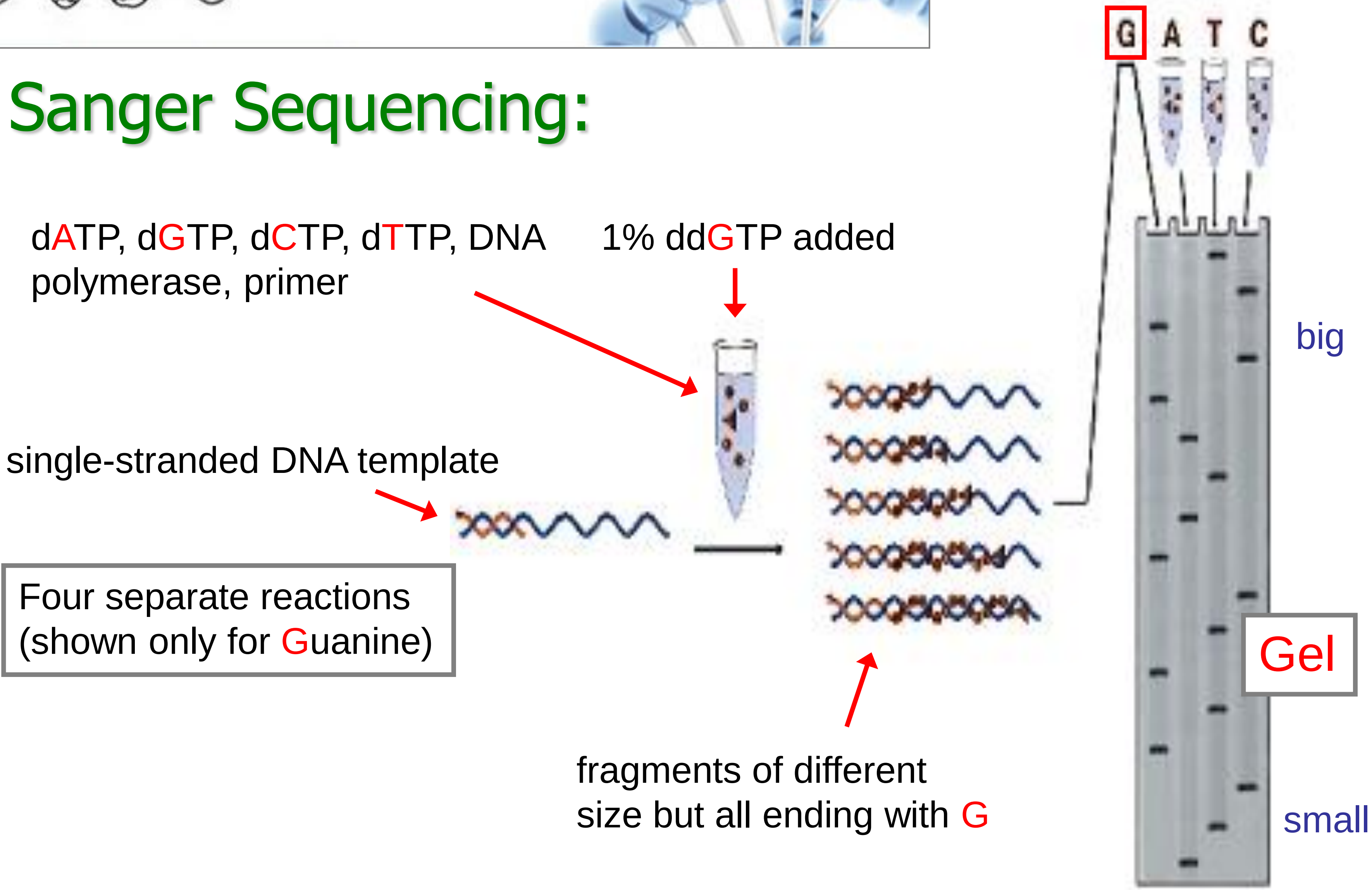
Human Genome Project



- Institute of Molecular Biotechnology (IMB) Jena
- DNA sequence of human Chr21, Chr8, ChrX (HUGO)
- Shotgun sequencing strategy
- Sanger sequencing (chain termination)

HUGO Human Genome Organisation

Sanger Sequencing:



dATP, dGTP, dCTP, dTTP, DNA polymerase, primer

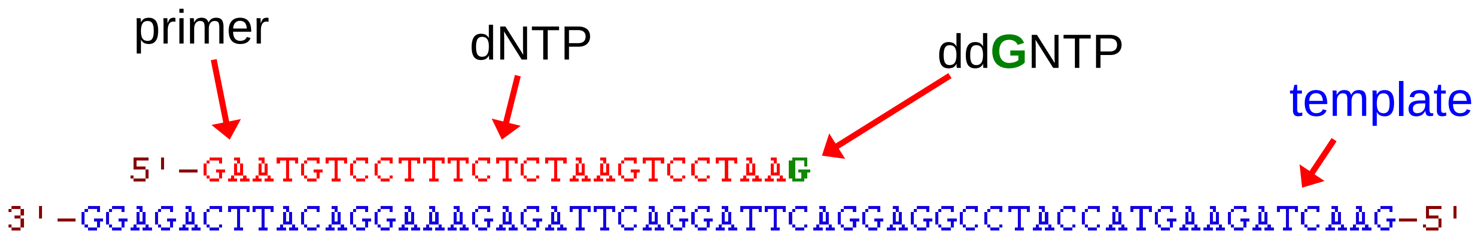
1% ddGTP added

single-stranded DNA template

Four separate reactions (shown only for Guanine)

fragments of different size but all ending with G

Chain termination



5'-GAATGTCCTTTCTCTAAGTCCTAAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

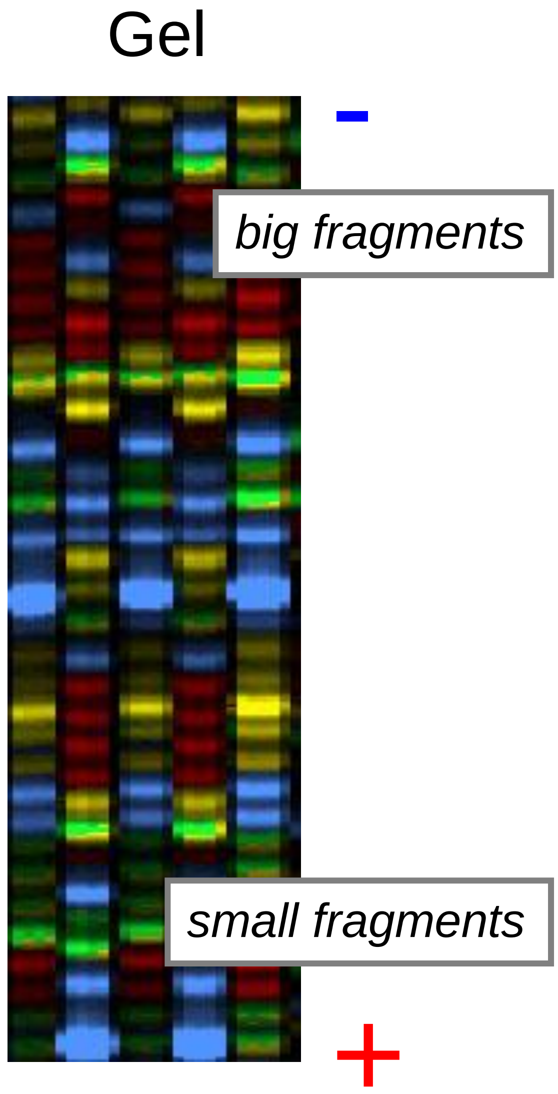
5'-GAATGTCCTTTCTCTAAGTCCTAAGTCCTCCGGATGGTACTTCTAG
3'-GGAGACTTACAGGAAAGAGATTCAGGATTCAGGAGGCCTACCATGAAGATCAAG-5'

fragments of different length all ending with Guanine

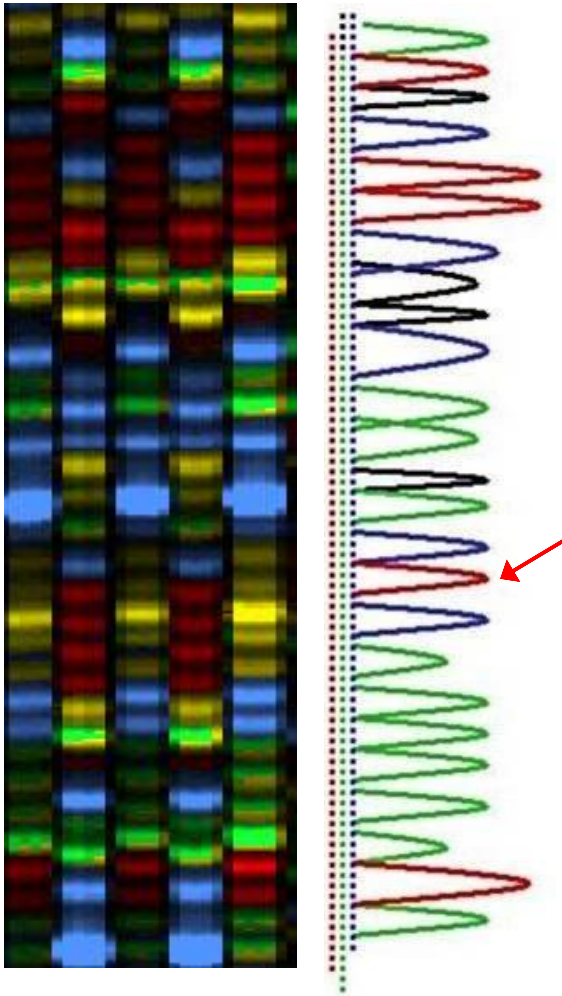
Sequencing

- Shotgun: fragment target-DNA randomly
- Synthesis: produce fractions of the fragment with different length using color-labeled ddNTPs
- Gel-electrophoresis: size-separate fractions by running them through a gel (resolution=1nt)

A T G C



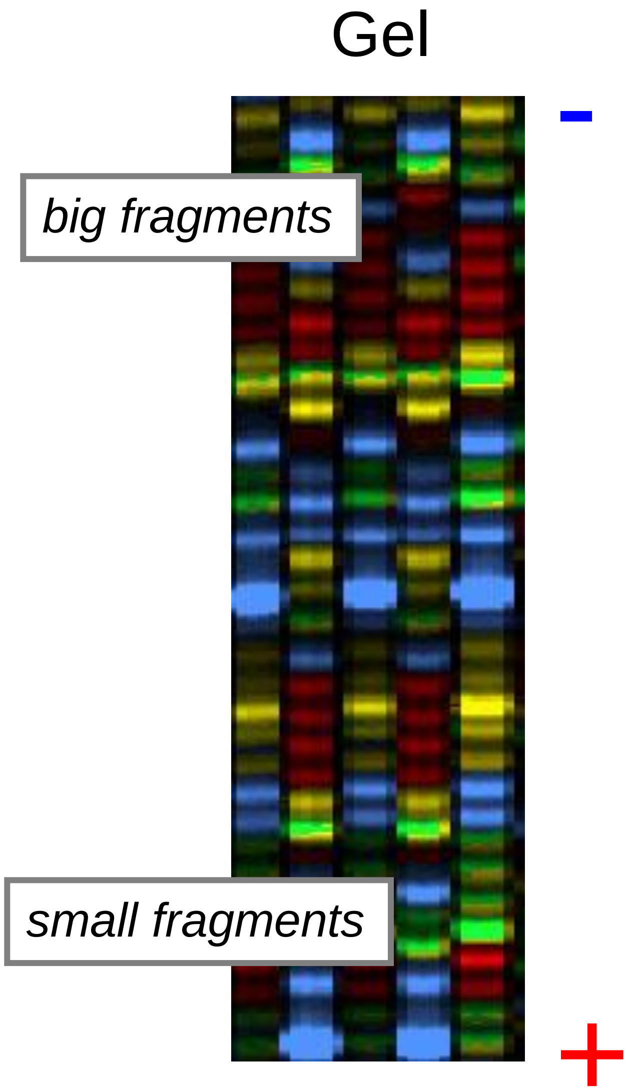
Chromatogram



Reading out the flourescent signal yields one chromatogram (trace file) for each lane.

A T G C

Dye-terminator sequencing



- labelling of the four chain terminator ddNTPs with fluorescent dyes
- permits sequencing in a single reaction, rather than four reactions

A T G C

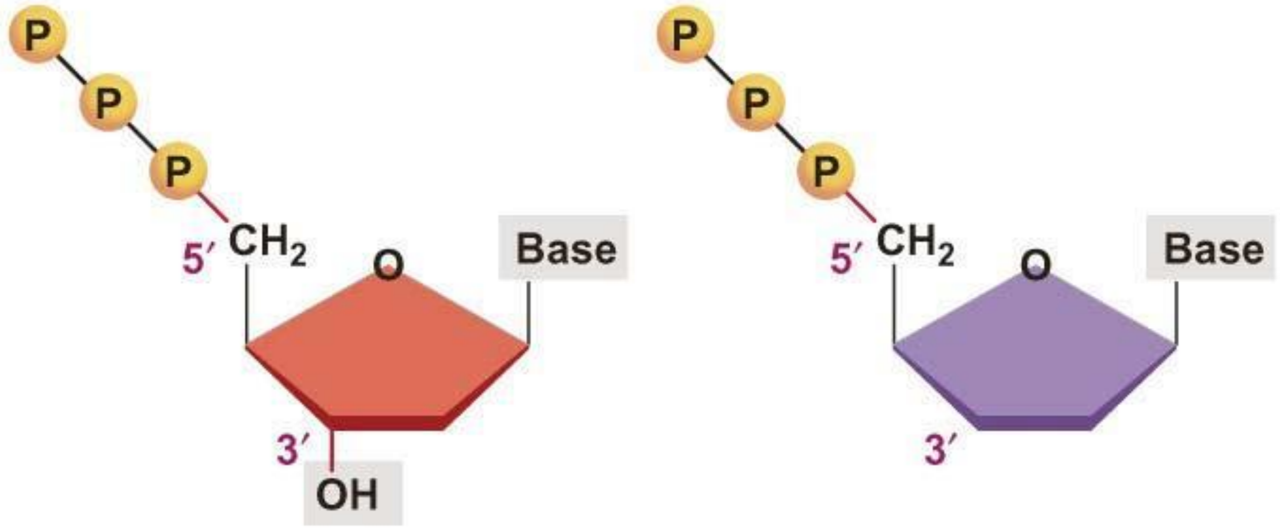
Chain termination method (Sanger)

- DNA sample is divided into 4 sequencing reactions, each containing:
  - the single-stranded DNA template
  - the 4 standard deoxynucleotides (dATP, dGTP, dCTP, dTTP),
  - DNA polymerase, DNA primer.
- *One* of the 4 ddNTPs added to each reaction (in low concentration):
  - ddNTP terminates the chain

  → one reaction ends with A, one with C, one with G, one with T
  - the fragments in each reaction are separated by size by gel electrophoresis (with a resolution of 1 bp !)

Chain termination method

(a) ddNTPs terminate DNA synthesis.

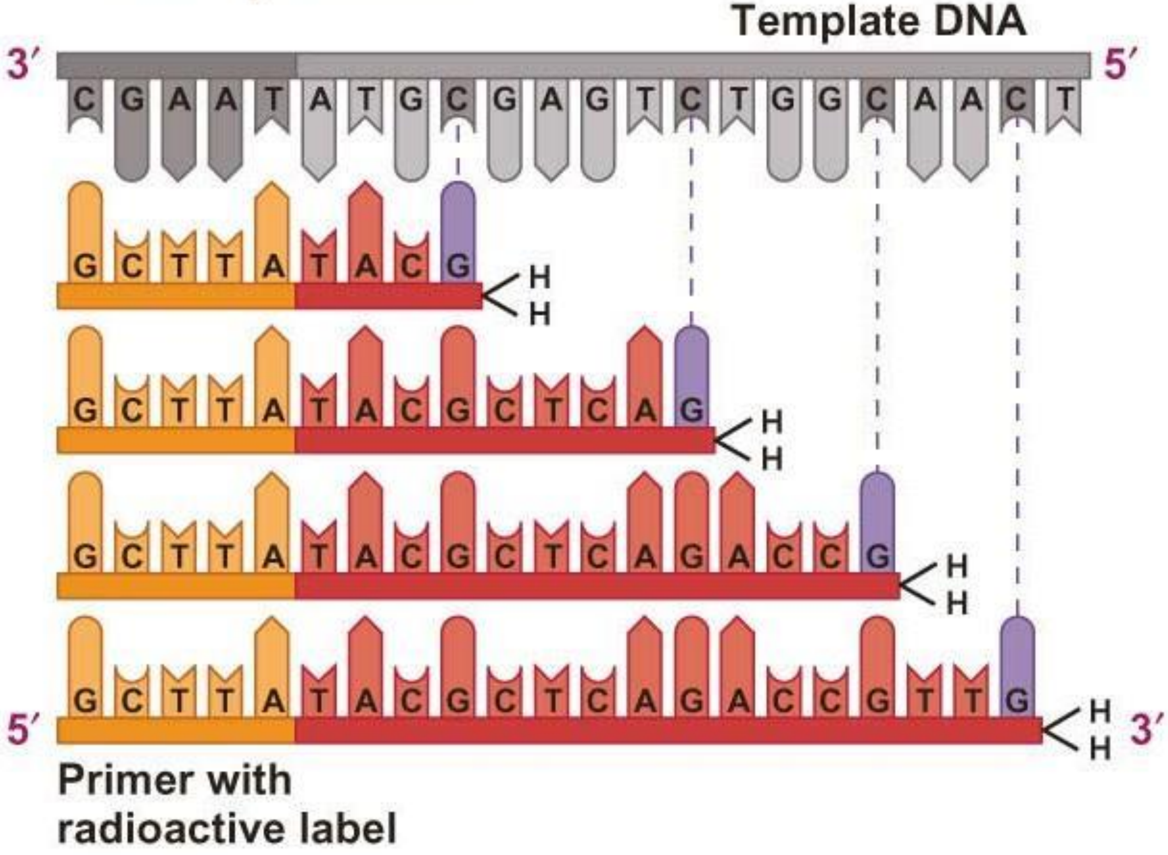


Normal dNTP
(extends DNA strand)

ddNTP
(terminates synthesis)

Chain termination method

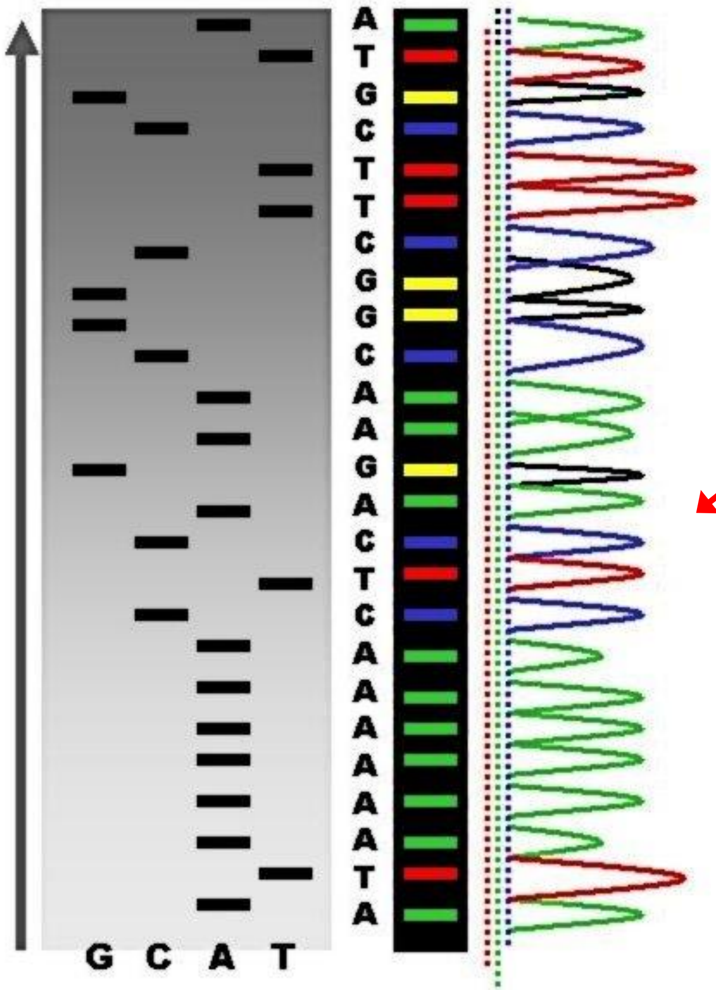
(b) Using ddNTPs, daughter strands of different length can be produced.


Template DNA
3′
5′
C G A A T A T G C G A G T C T G G C A A C T
G C T T A T A C G
H
H
G C T T A T A C G C T C A G
H
H
G C T T A T A C G C T C A G A C C G
H
H
G C T T A T A C G C T C A G A C C G T T G
H
H
5′
3′
Primer with radioactive label

Gel-electrophoresis

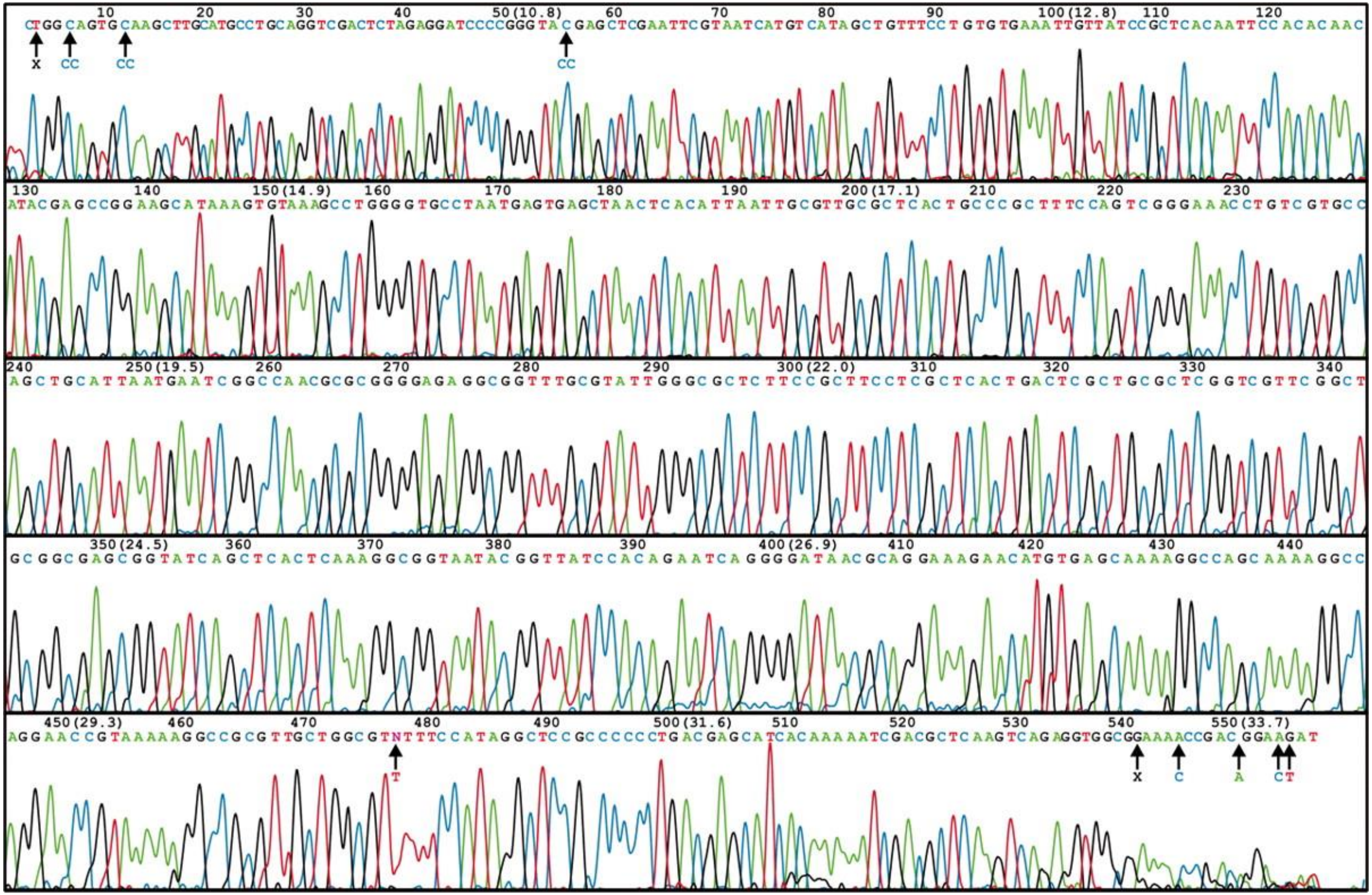
fragments labeled with dyes

read out the fluorescense



Chromatogram (trace file)

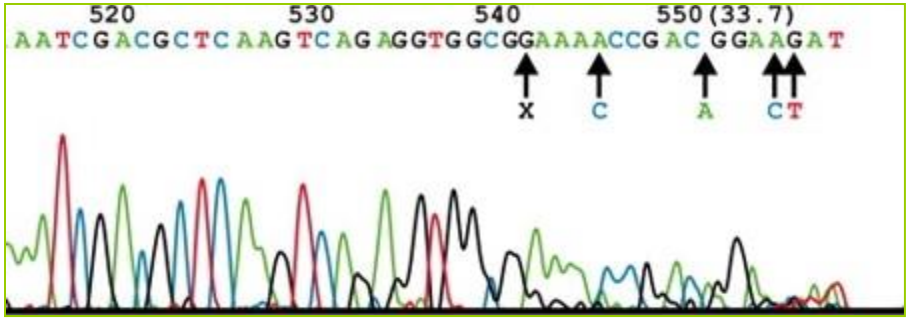
High througput

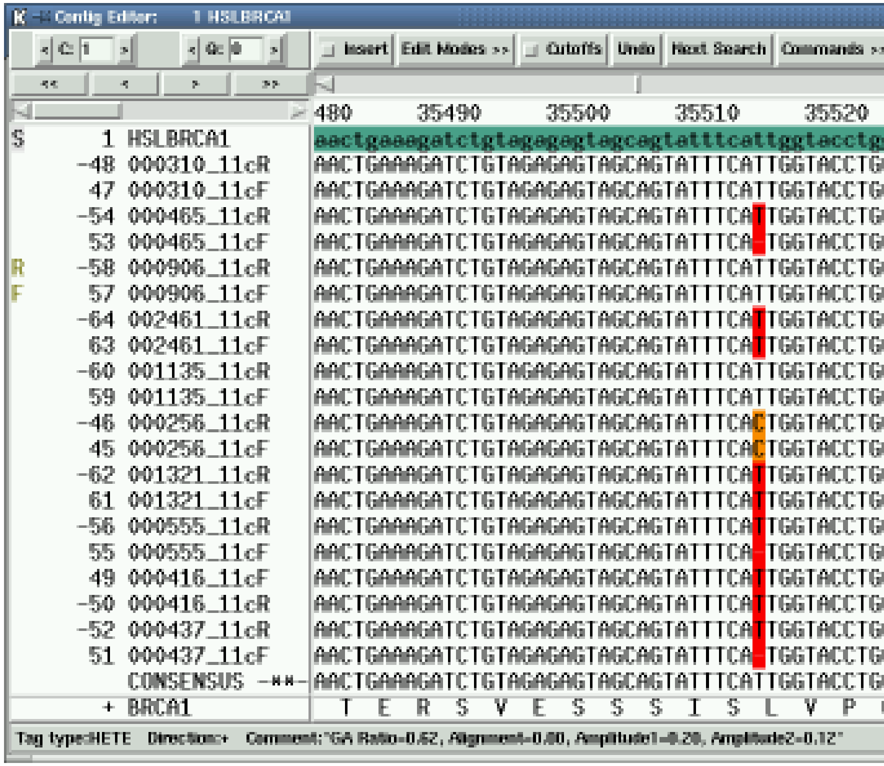


10 20 30 40 50 (10.8) 60 70 80 90 100 (12.8) 110 120
CTGGCAGTGCAAGCTTGCATGCCTGCAGGTCGACTCTAGAGGATCCCCGGGTACGAGCTCGAATTCGTAATCATGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAAC
x
CC
CC
CC
130 140 150 (14.9) 160 170 180 190 200 (17.1) 210 220 230
ATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCC
240 250 (19.5) 260 270 280 290 300 (22.0) 310 320 330 340
AGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCT
350 (24.5) 360 370 380 390 400 (26.9) 410 420 430 440
GCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCC
450 (29.3) 460 470 480 490 500 (31.6) 510 520 530 540 550 (33.7)
AGGAACCGTAAAAAGGCCGCGTTGCTGGCGTNTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGGAAAACCGACGGAAGAT
T
x
C
A
C
T

High-throughput data-mining

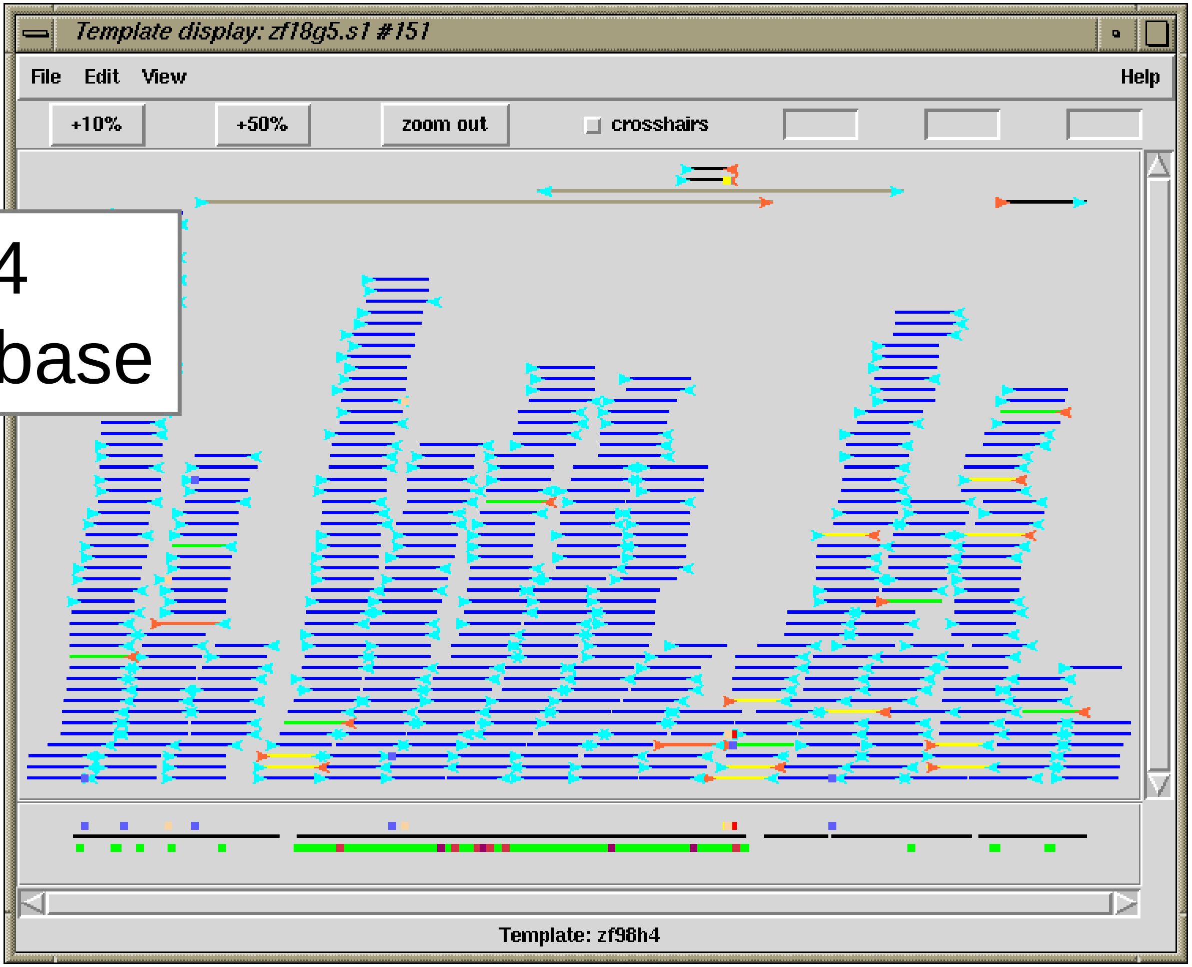
Base calling → Base accuracy calculation → Quality clipping → Vector Removal → Repeat masking → ... *next slides*





```
                          480      35490     35500     35510     35520
S      1 HSLBRCA1          aactgaaagatctgtagagagtagcagtatttcattggtacctg
     -48 000310_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
      47 000310_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
     -54 000465_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
      53 000465_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCA-TGGTACCTG
R    -58 000906_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
F     57 000906_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
     -64 002461_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
      63 002461_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
     -60 001135_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
      59 001135_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
     -46 000256_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCACTGGTACCTG
      45 000256_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCACTGGTACCTG
     -62 001321_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
      61 001321_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
     -56 000555_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
      55 000555_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCA-TGGTACCTG
      49 000416_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
     -50 000416_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
     -52 000437_11cR       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
      51 000437_11cF       AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCA-TGGTACCTG
         CONSENSUS   -**-  AACTGAAAGATCTGTAGAGAGTAGCAGTATTTCATTGGTACCTG
       + BRCA1               T  E  R  S  V  E  S  S  S  I  S  L  V  P
Tag type:HETE   Direction:+   Comment:"GA Ratio=0.62, Alignment=0.80, Amplitude1=0.20, Amplitude2=0.12"
```

Assembling the fragments

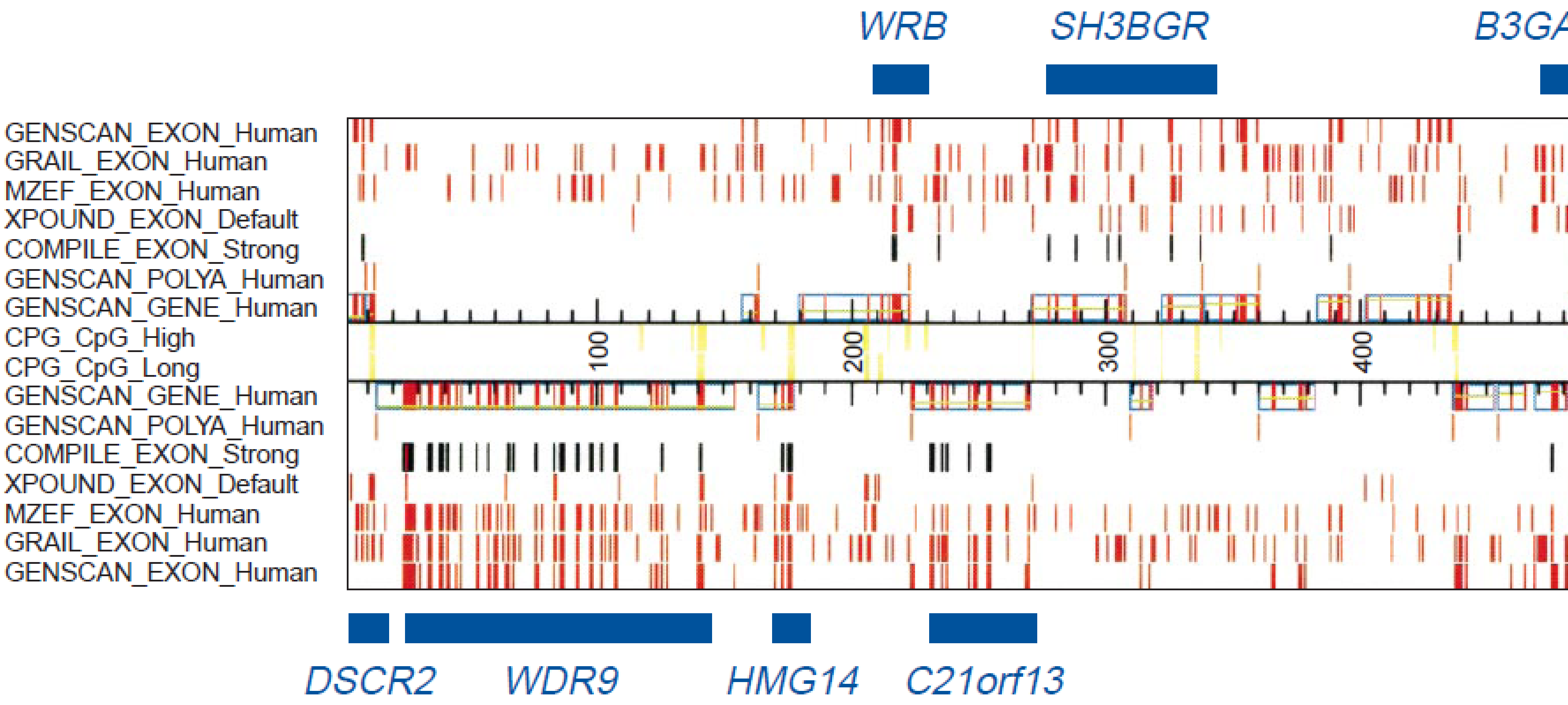


Gap4 database

the DNA sequence of the overlapping pieces must be concordant

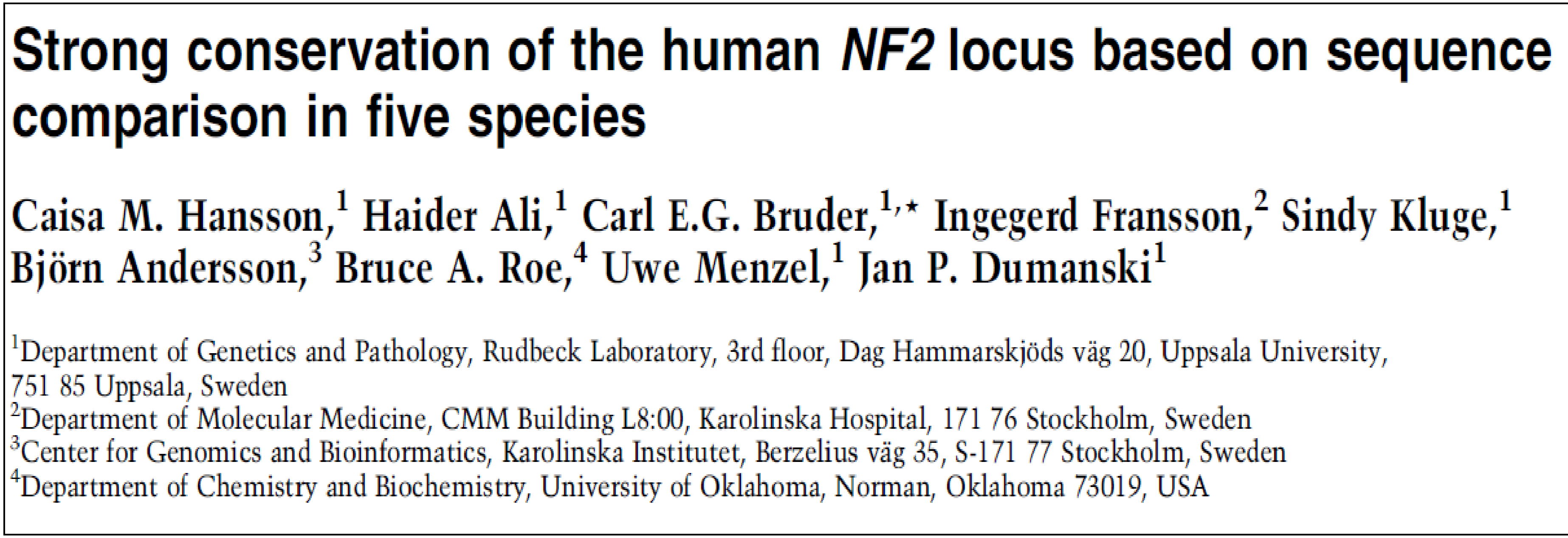
Annotation of the sequence

FIGURE 2. Graphical output of RUMMAGE



Each row represents the hits of one single program. The number of kilobases are shown in the centre. Because of the clustering of h detected easily. For each hit, detailed information is available by a mouse click. Furthermore, the user can zoom in on any desired p Gene names and blue bars are not shown by RUMMAGE and were added to emphasize the clusters of hits.

Sequencing and multiple species comparison



Strong conservation of the human *NF2* locus based on sequence comparison in five species

Caisa M. Hansson,[1] Haider Ali,[1] Carl E.G. Bruder,[1,*] Ingegerd Fransson,[2] Sindy Kluge,[1] Björn Andersson,[3] Bruce A. Roe,[4] Uwe Menzel,[1] Jan P. Dumanski[1]

[1]Department of Genetics and Pathology, Rudbeck Laboratory, 3rd floor, Dag Hammarskjöds väg 20, Uppsala University,
751 85 Uppsala, Sweden
[2]Department of Molecular Medicine, CMM Building L8:00, Karolinska Hospital, 171 76 Stockholm, Sweden
[3]Center for Genomics and Bioinformatics, Karolinska Institutet, Berzelius väg 35, S-171 77 Stockholm, Sweden
[4]Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019, USA

Incorporating *Mouse Genome*
Mammalian Genome
Genes and Phenotypes

Comparison of five species

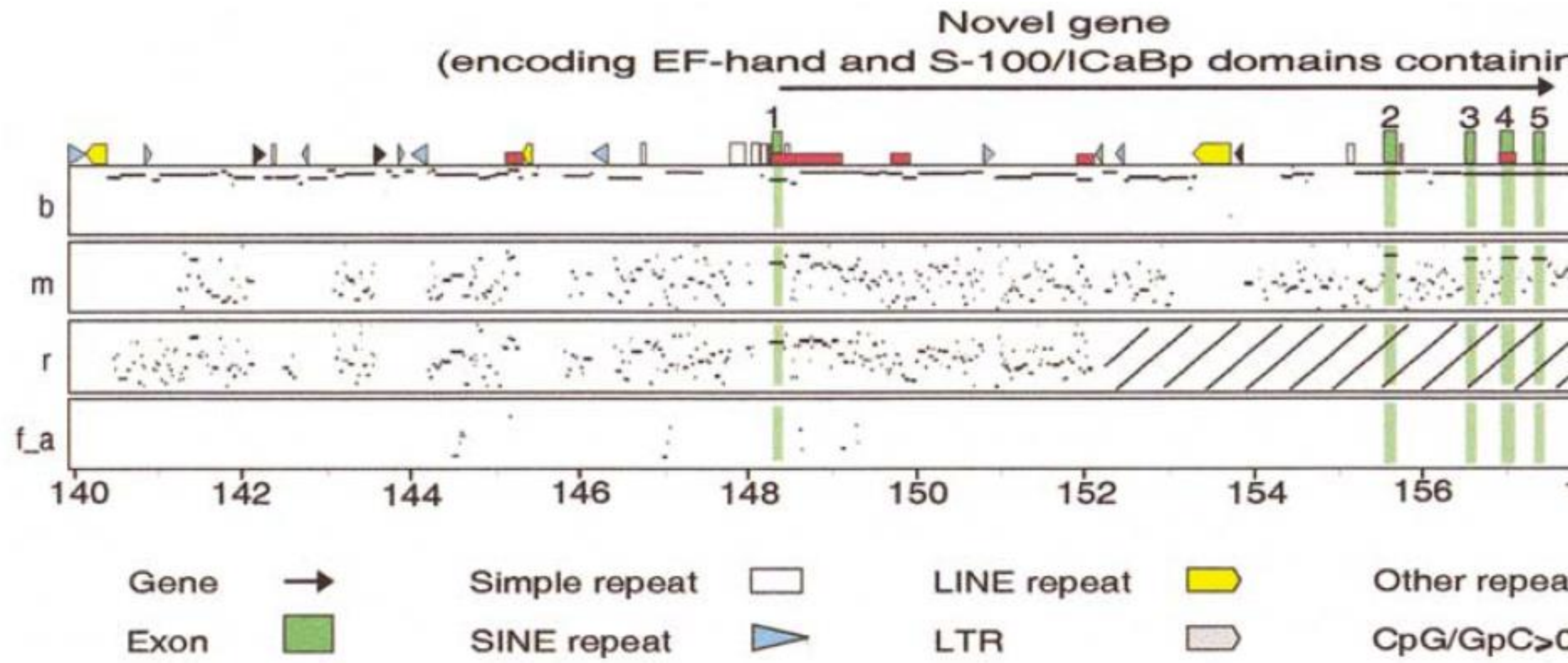
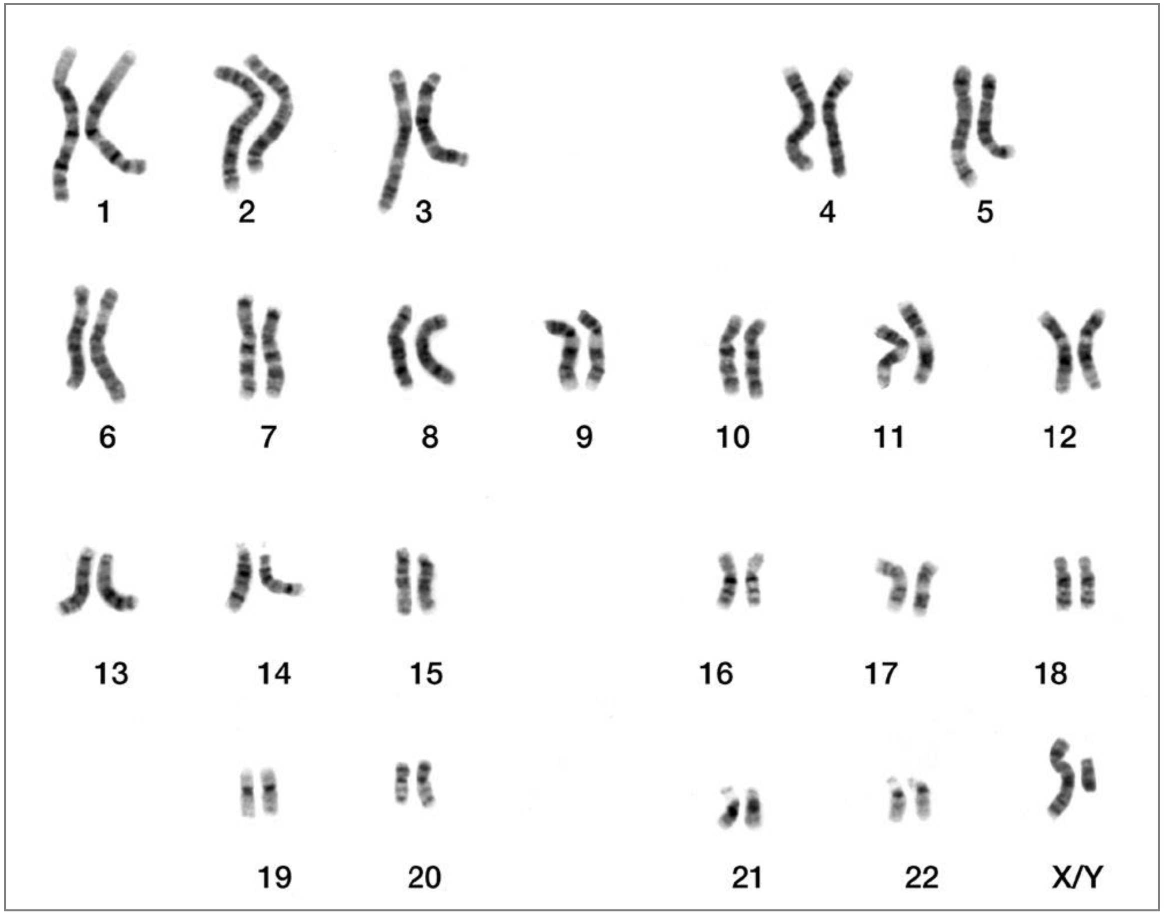


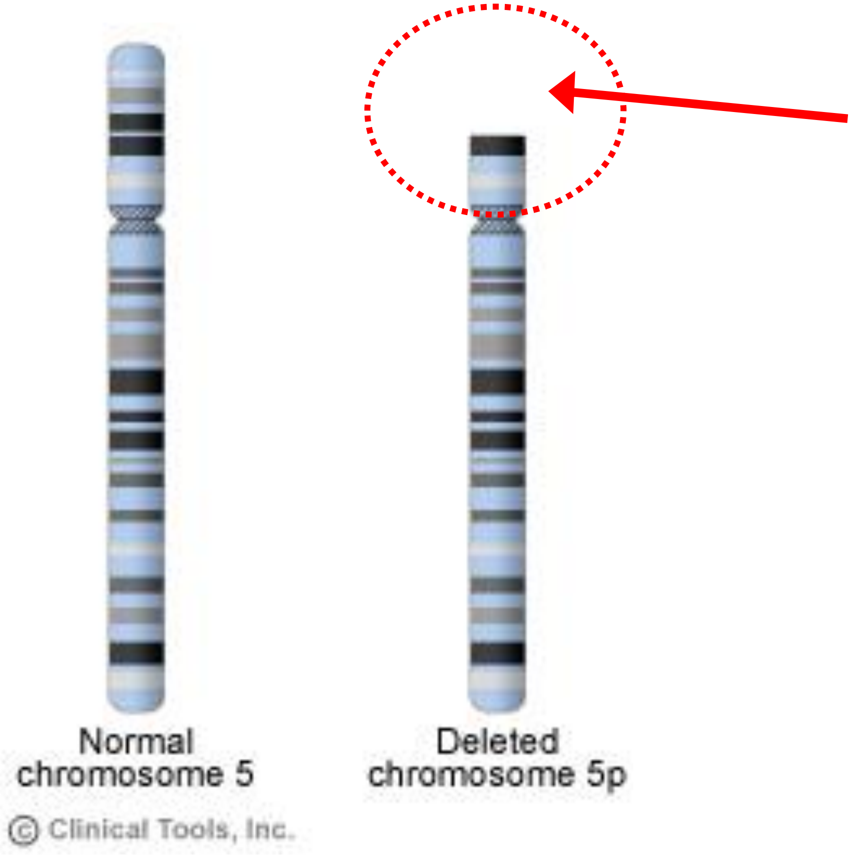
Fig. 1. A *P*ercentage *I*dentity *P*lot (PIP), produced by MultiPipMaker, shows the evolutionary c
nomic sequences from human (reference), baboon (b), mouse (m), rat (r), and pufferfish *f_neurofibr*
human *NF2* gene spans a region between positions 31.5 and 126.5 kb. Human, baboon, mouse and
locus are flanked by exon 1 of *NIPSNAP1* gene and a novel gene, encoding EF-hand and S-100/IC

Copy Number Aberrations in Genomes

two copies
in each
(somatic)
cell



Copy Number Aberrations in Genomes

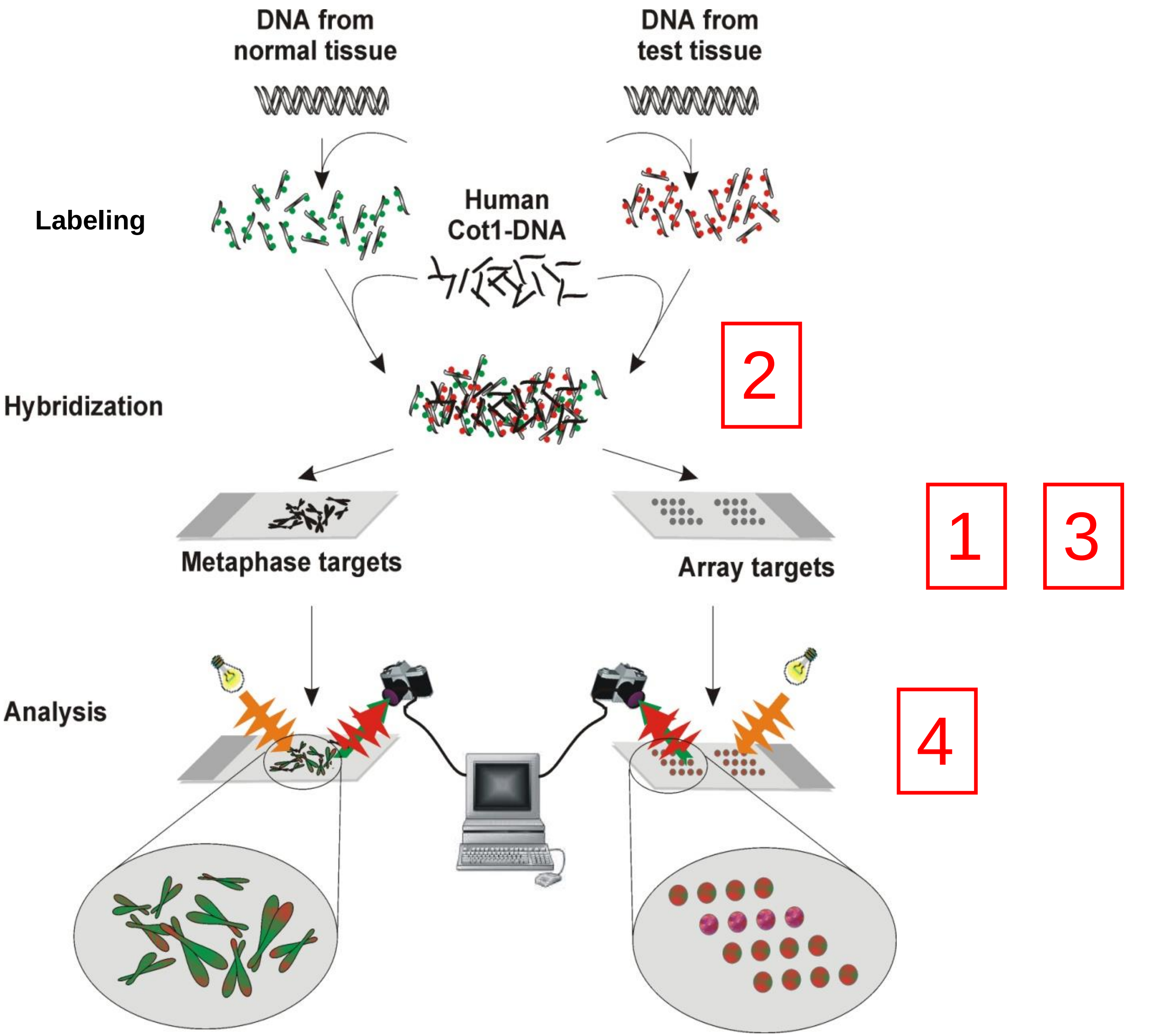


- One or both copies can get lost (deletion)
- Additional copies (gain, amplification)
- Size: 1 bp - few Mbp, (whole chromosome - trisomy 21)
- Important in cancer development (TSG, Oncogenes)

Method: Array CGH

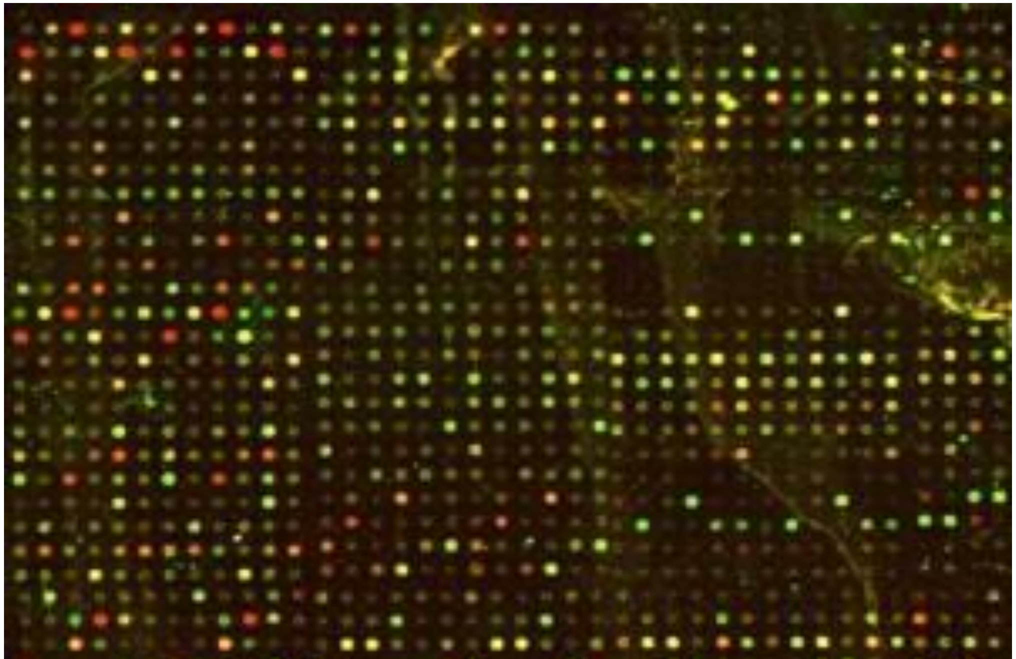
- **C**omparative **G**enomic **H**ybridization
- Compare the genomes of two individuals or of two different tissues of the same individual (e.g. normal tissue – tumor tissue)

Metaphase-CGH and Array-CGH

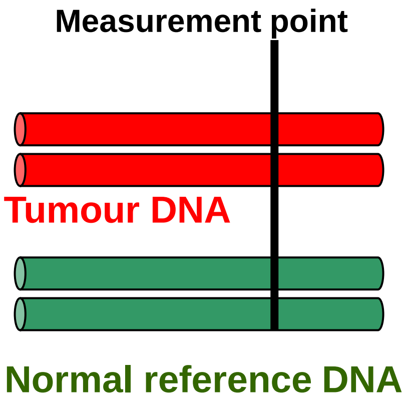


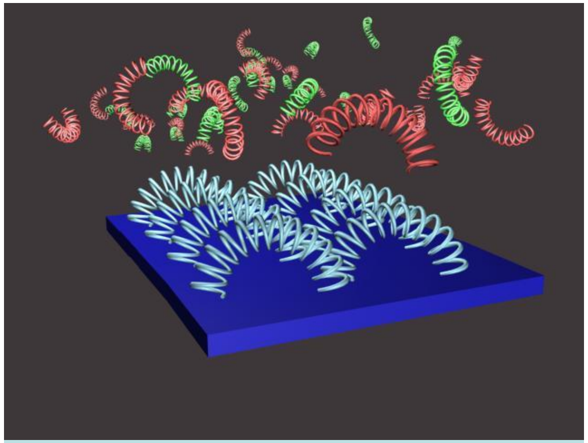
Array CGH

- The red and green intensities are measured in each spot of the array.
- The intensity ratio allows estimation of the relative copy number of test and reference DNA.
- 2 copies of both test + reference DNA → $\boxed{R=1, \log_2 R=0}$



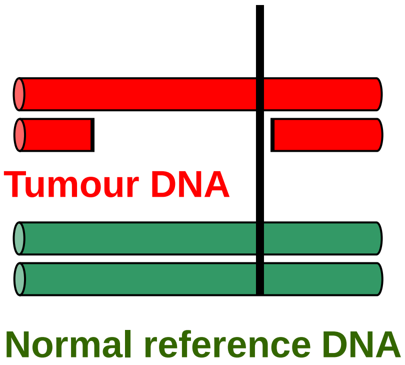
Normal Case

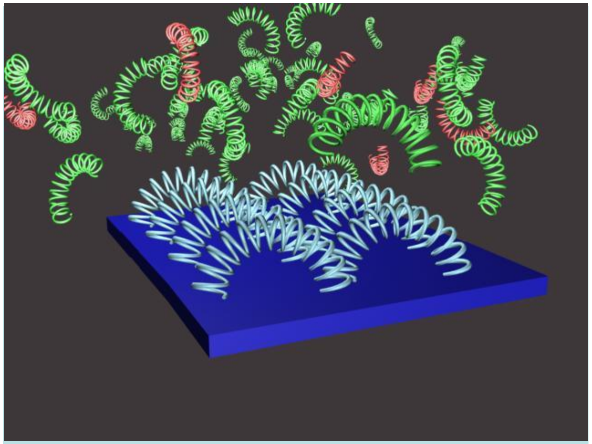




$$\text{Ratio} = \frac{\text{Red}}{\text{Green}} = \frac{n_{red}}{n_{green}} = 1$$

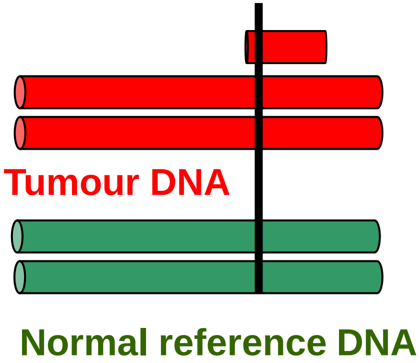
Deletion

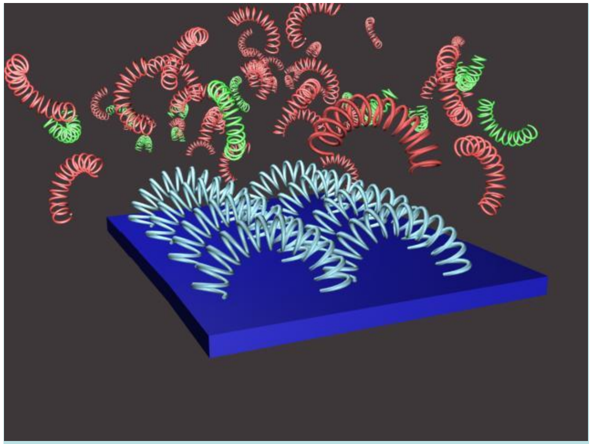




$$\text{Ratio} = \frac{\text{Red}}{\text{Green}} = \frac{n_{red}}{n_{green}} = 0.5$$

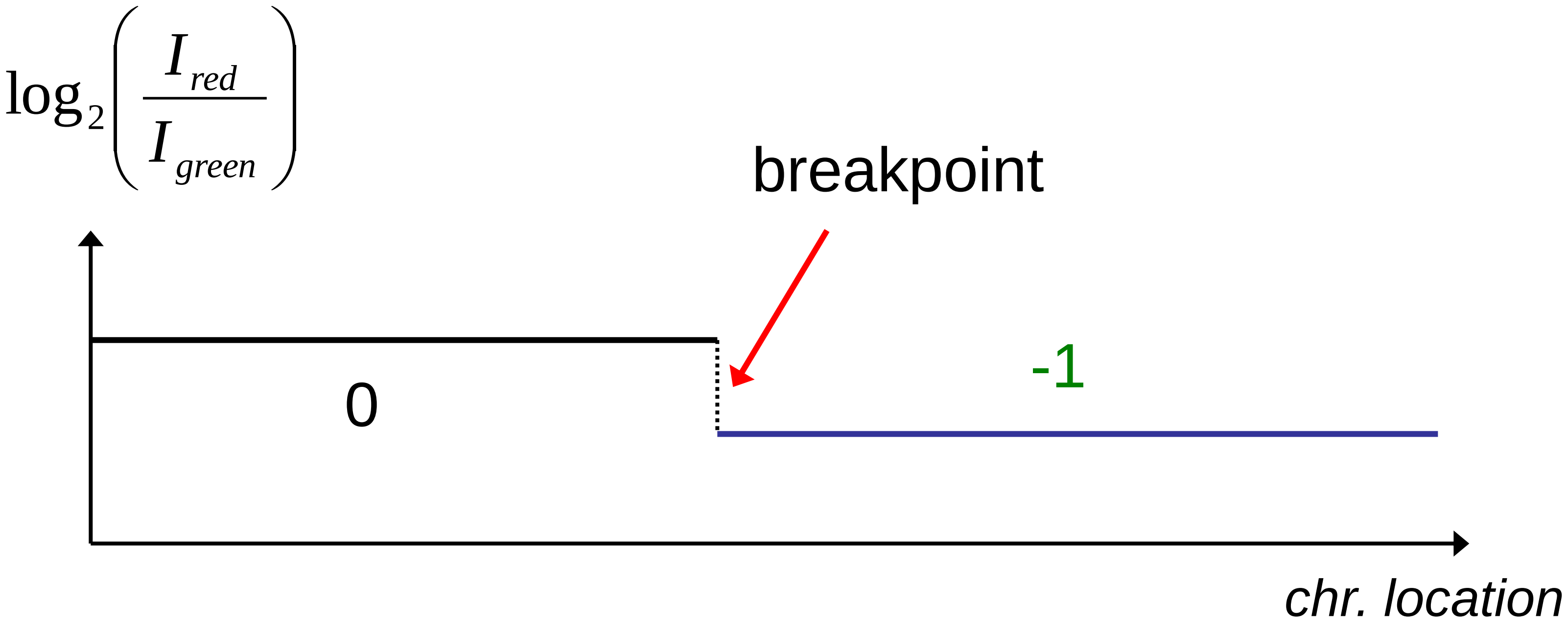
Duplication





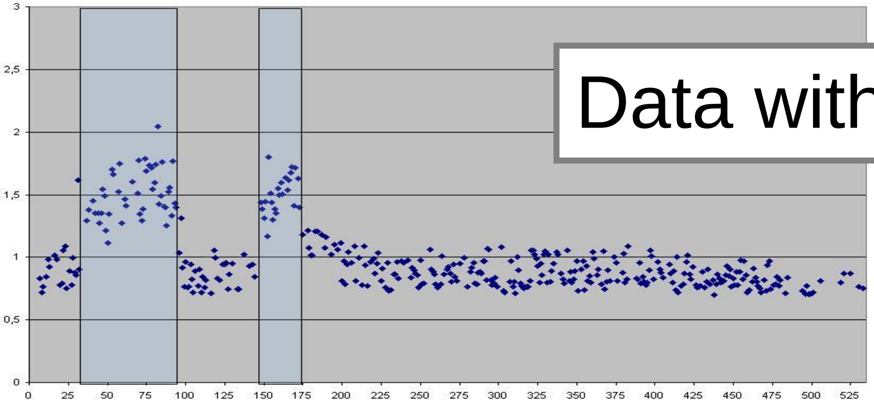
$$\text{Ratio} = \frac{\text{Red}}{\text{Green}} = \frac{n_{red}}{n_{green}} = 1.5$$

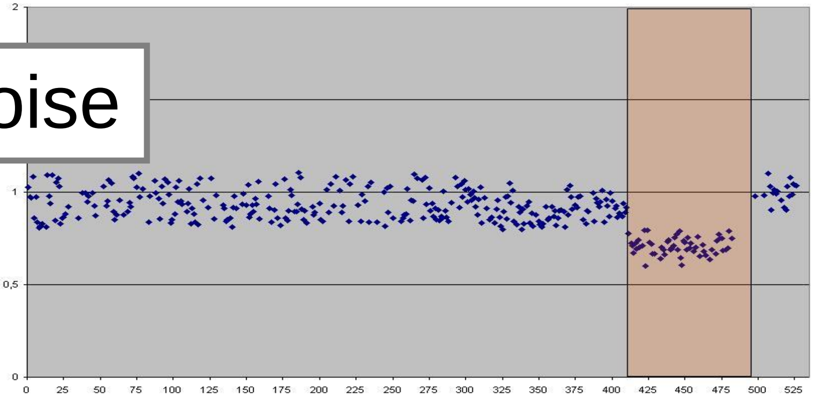
Biological model

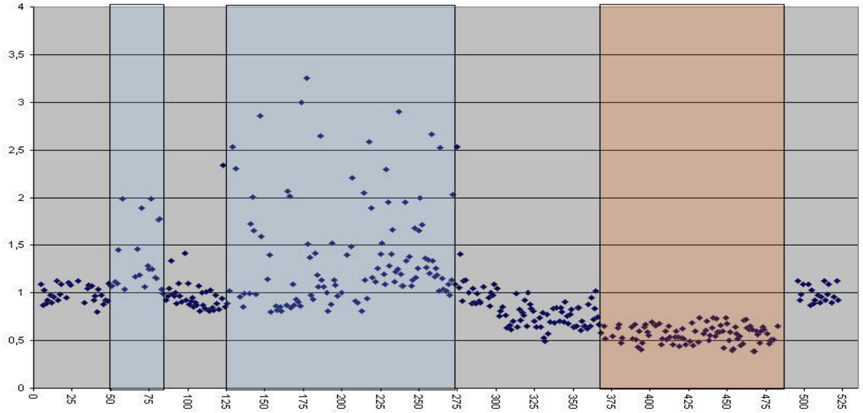


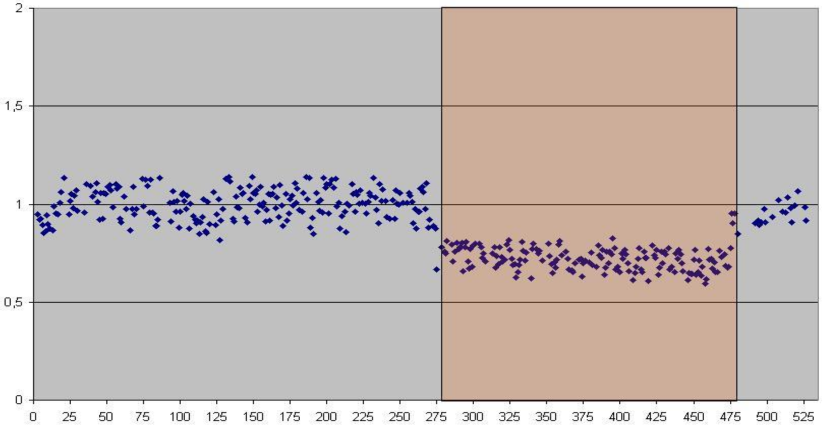
<u>Assumption:</u> Genomic rearrangements lead to gain or loss of contiguous segments of the genome.

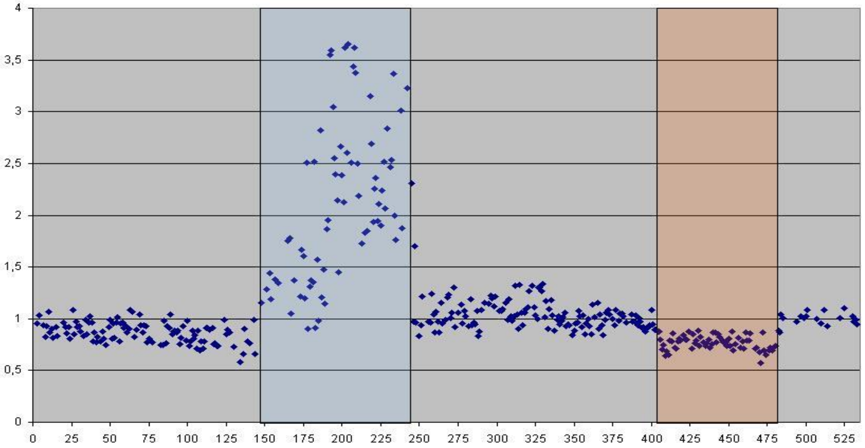
Data with noise

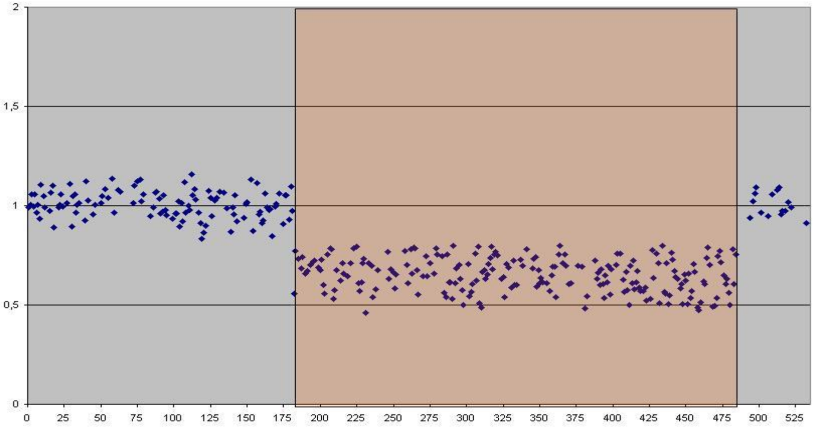




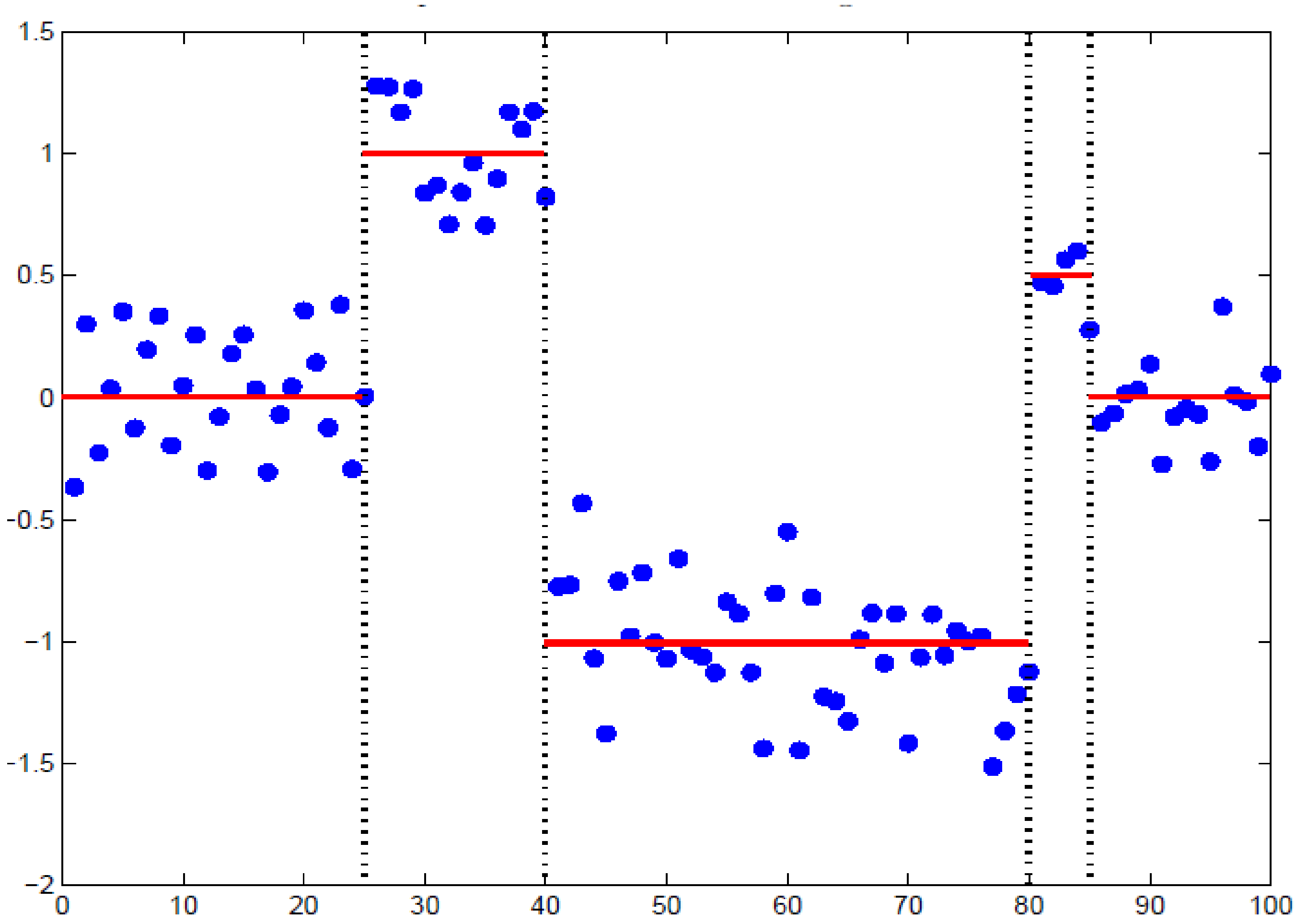




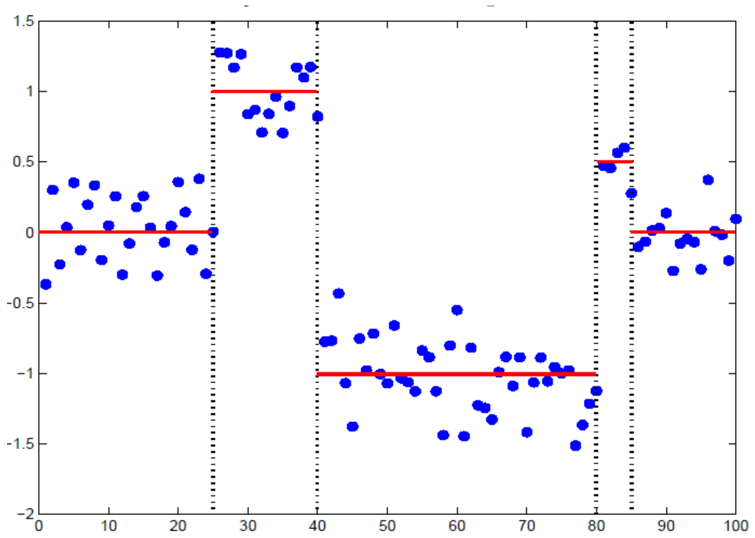




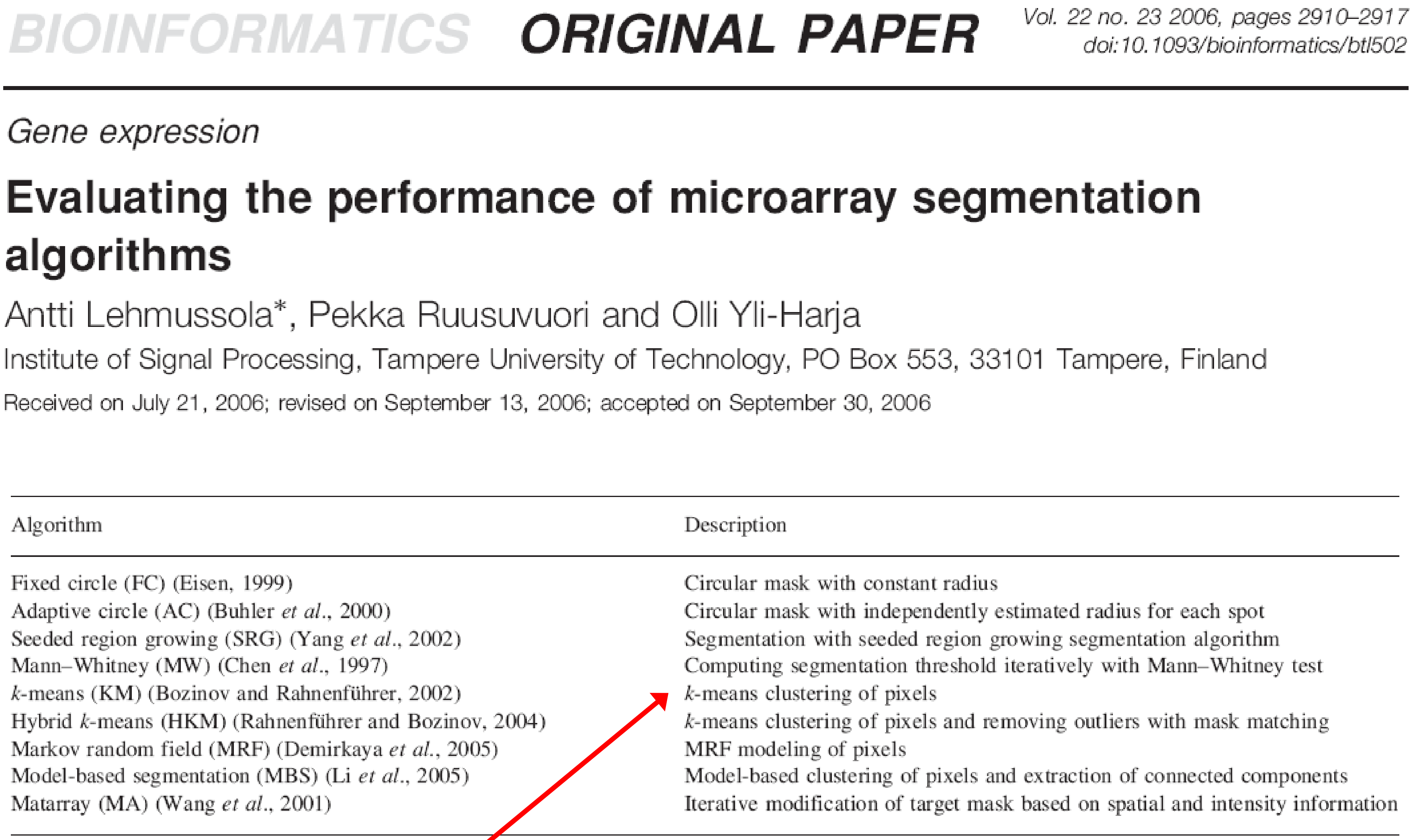
Segmentation



Segmentation:



- (Smoothing →) thresholds
  - based on the variability
  - Weiss 2003
- Normal mixture models
  - e.g. 3 Gaussian components: deletion, normal, gain
  - Hodgson 2001
- Clustering
  - Autio (2003)



BIOINFORMATICS ORIGINAL PAPER
Vol. 22 no. 23 2006, pages 2910–2917
doi:10.1093/bioinformatics/btl502

Gene expression

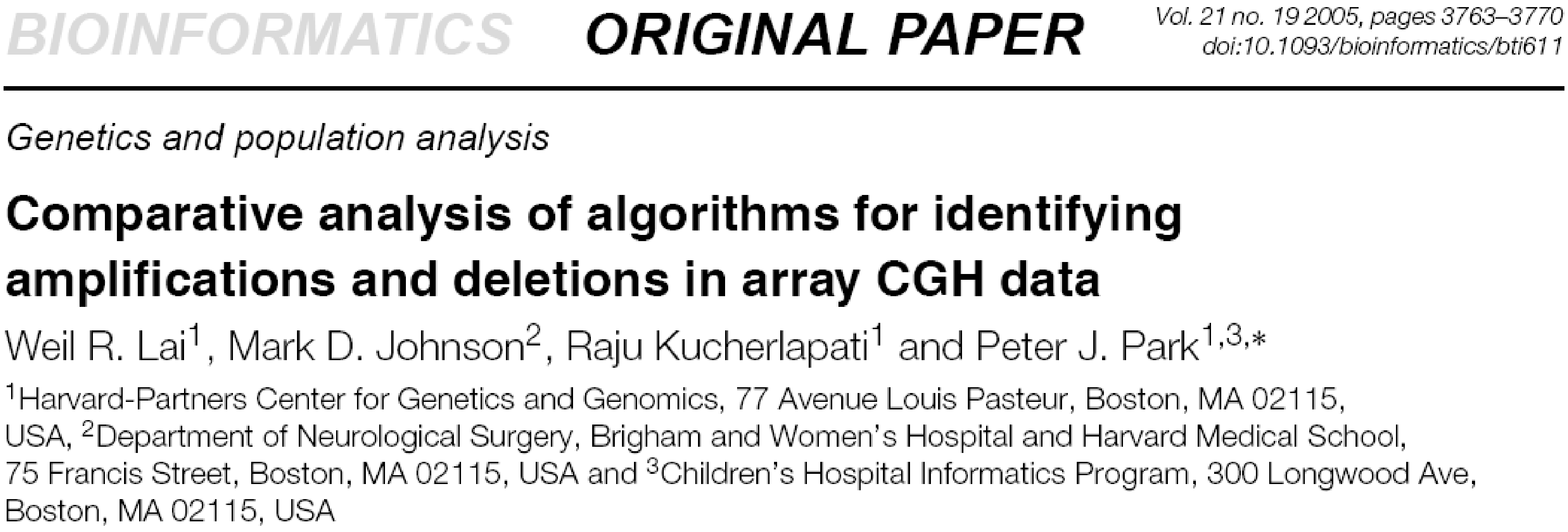
Evaluating the performance of microarray segmentation algorithms

Antti Lehmussola*, Pekka Ruusuvuori and Olli Yli-Harja
Institute of Signal Processing, Tampere University of Technology, PO Box 553, 33101 Tampere, Finland

Received on July 21, 2006; revised on September 13, 2006; accepted on September 30, 2006

Algorithm	Description
Fixed circle (FC) (Eisen, 1999)	Circular mask with constant radius
Adaptive circle (AC) (Buhler *et al*., 2000)	Circular mask with independently estimated radius for each spot
Seeded region growing (SRG) (Yang *et al*., 2002)	Segmentation with seeded region growing segmentation algorithm
Mann–Whitney (MW) (Chen *et al*., 1997)	Computing segmentation threshold iteratively with Mann–Whitney test
k-means (KM) (Bozinov and Rahnenführer, 2002)	*k*-means clustering of pixels
Hybrid *k*-means (HKM) (Rahnenführer and Bozinov, 2004)	*k*-means clustering of pixels and removing outliers with mask matching
Markov random field (MRF) (Demirkaya *et al*., 2005)	MRF modeling of pixels
Model-based segmentation (MBS) (Li *et al*., 2005)	Model-based clustering of pixels and extraction of connected components
Matarray (MA) (Wang *et al*., 2001)	Iterative modification of target mask based on spatial and intensity information

k-means algorithm best (simulated data)



BIOINFORMATICS ORIGINAL PAPER

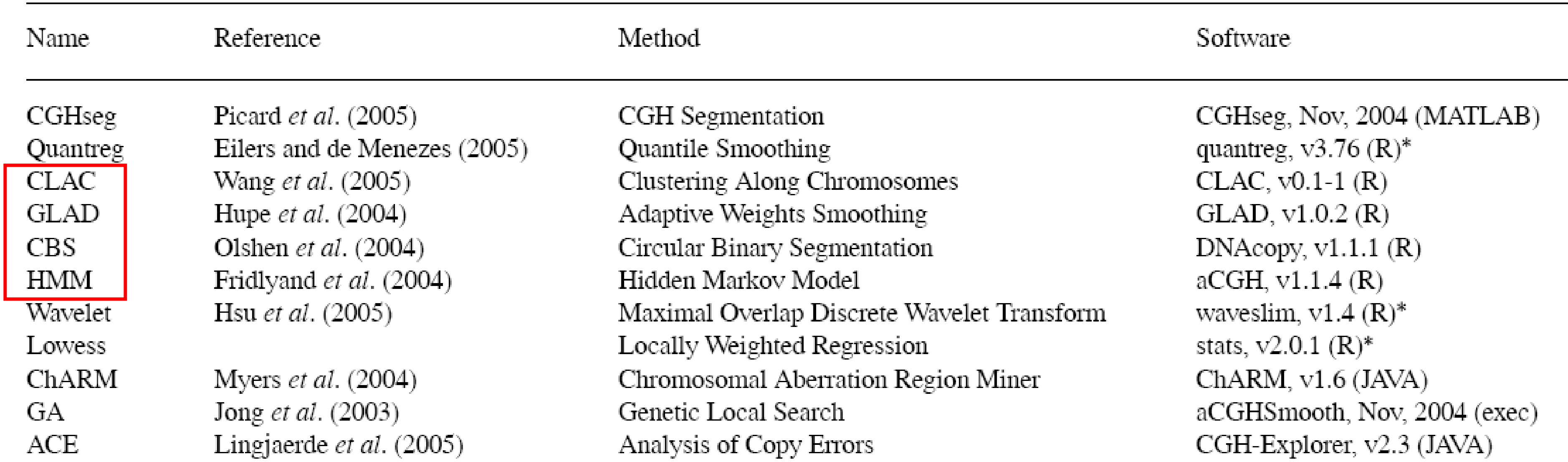
Vol. 21 no. 19 2005, pages 3763–3770
doi:10.1093/bioinformatics/bti611

Genetics and population analysis

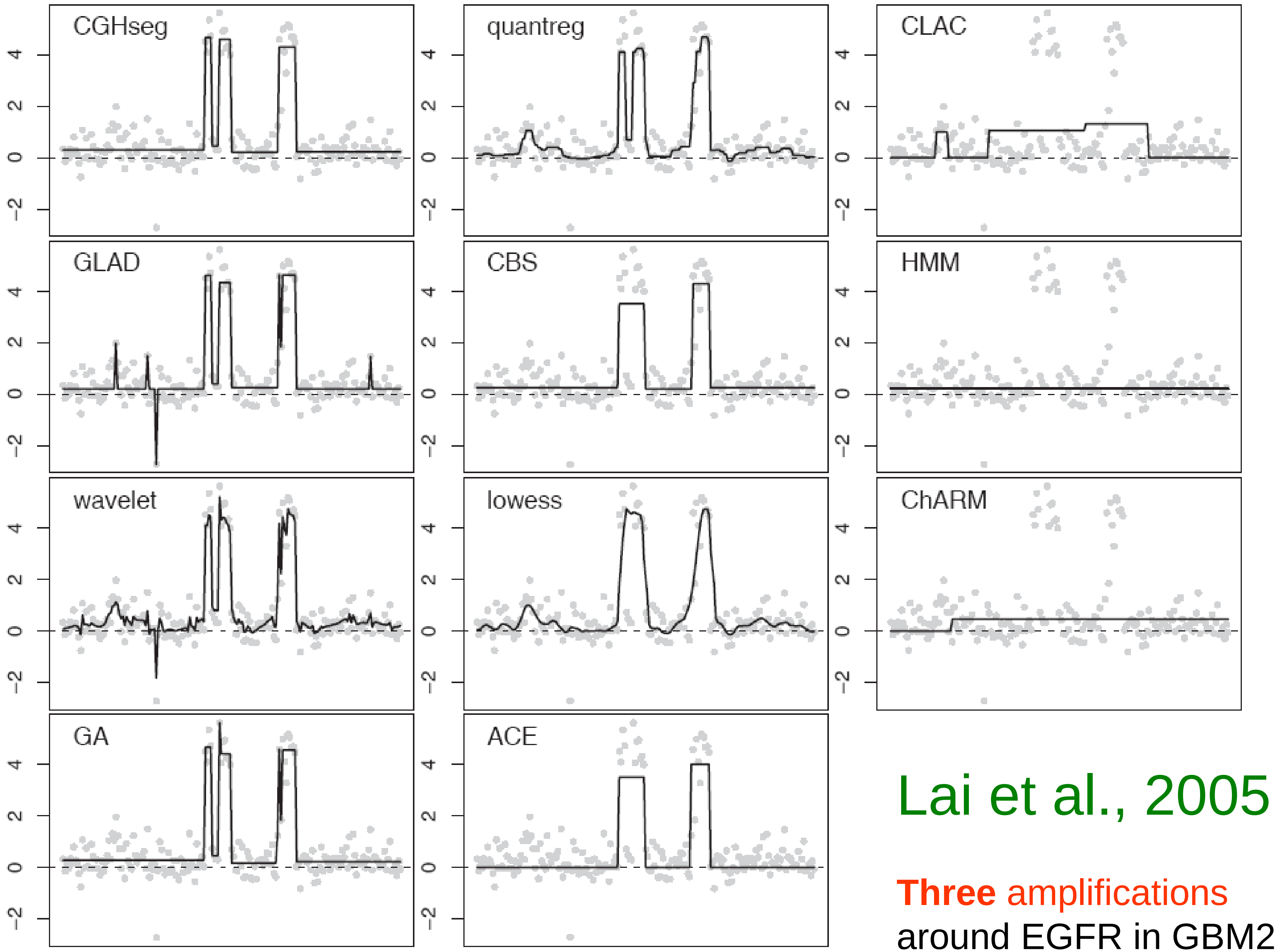
Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data

Weil R. Lai[1], Mark D. Johnson[2], Raju Kucherlapati[1] and Peter J. Park[1,3,*]

[1]Harvard-Partners Center for Genetics and Genomics, 77 Avenue Louis Pasteur, Boston, MA 02115, USA, [2]Department of Neurological Surgery, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA and [3]Children's Hospital Informatics Program, 300 Longwood Ave, Boston, MA 02115, USA

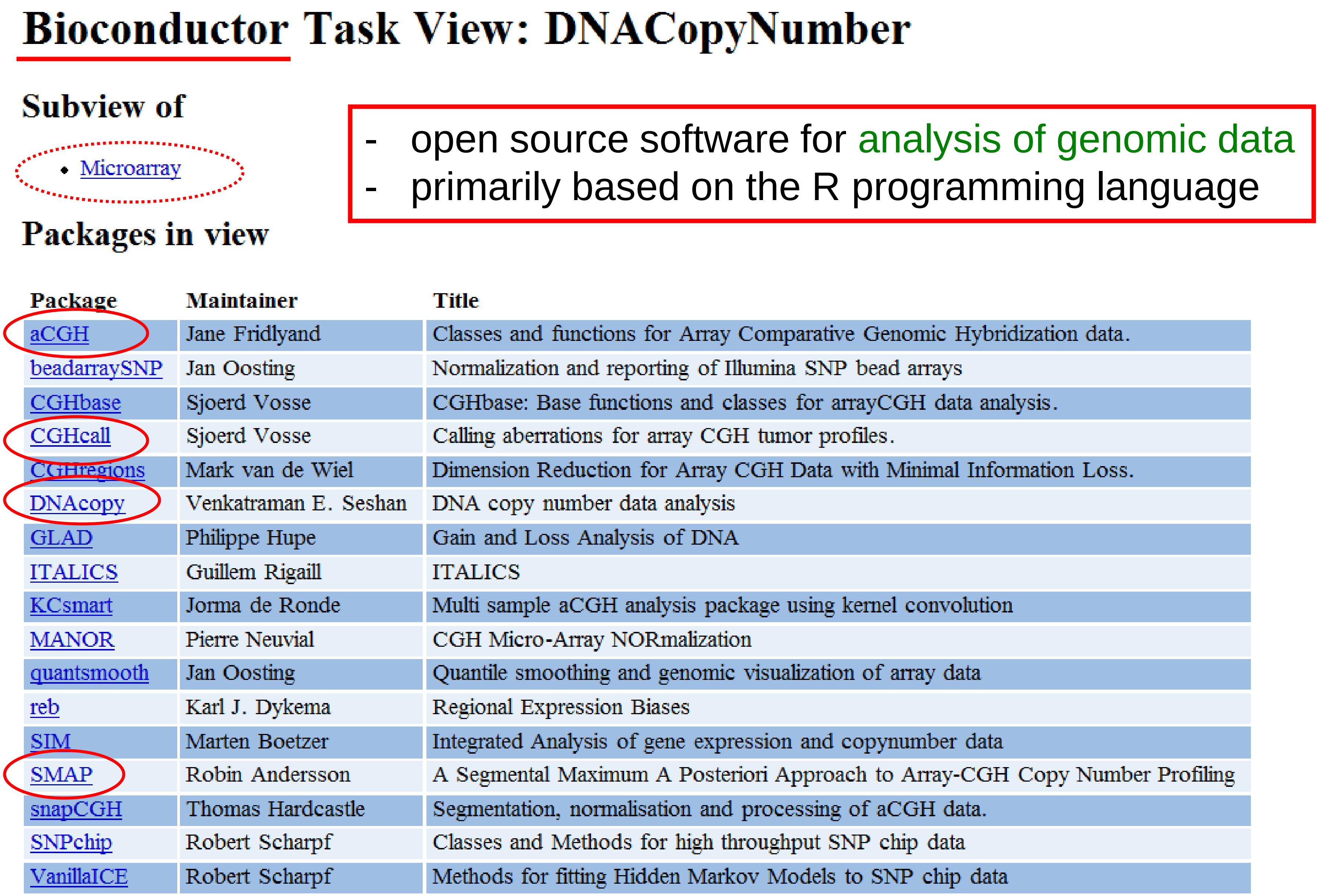


Name	Reference	Method	Software
CGHseg	Picard *et al*. (2005)	CGH Segmentation	CGHseg, Nov, 2004 (MATLAB)
Quantreg	Eilers and de Menezes (2005)	Quantile Smoothing	quantreg, v3.76 (R)*
CLAC	Wang *et al*. (2005)	Clustering Along Chromosomes	CLAC, v0.1-1 (R)
GLAD	Hupe *et al*. (2004)	Adaptive Weights Smoothing	GLAD, v1.0.2 (R)
CBS	Olshen *et al*. (2004)	Circular Binary Segmentation	DNAcopy, v1.1.1 (R)
HMM	Fridlyand *et al*. (2004)	Hidden Markov Model	aCGH, v1.1.4 (R)
Wavelet	Hsu *et al*. (2005)	Maximal Overlap Discrete Wavelet Transform	waveslim, v1.4 (R)*
Lowess		Locally Weighted Regression	stats, v2.0.1 (R)*
ChARM	Myers *et al*. (2004)	Chromosomal Aberration Region Miner	ChARM, v1.6 (JAVA)
GA	Jong *et al*. (2003)	Genetic Local Search	aCGHSmooth, Nov, 2004 (exec)
ACE	Lingjaerde *et al*. (2005)	Analysis of Copy Errors	CGH-Explorer, v2.3 (JAVA)



Lai et al., 2005

Three amplifications
around EGFR in GBM29



Bioconductor Task View: DNACopyNumber

Subview of

- Microarray

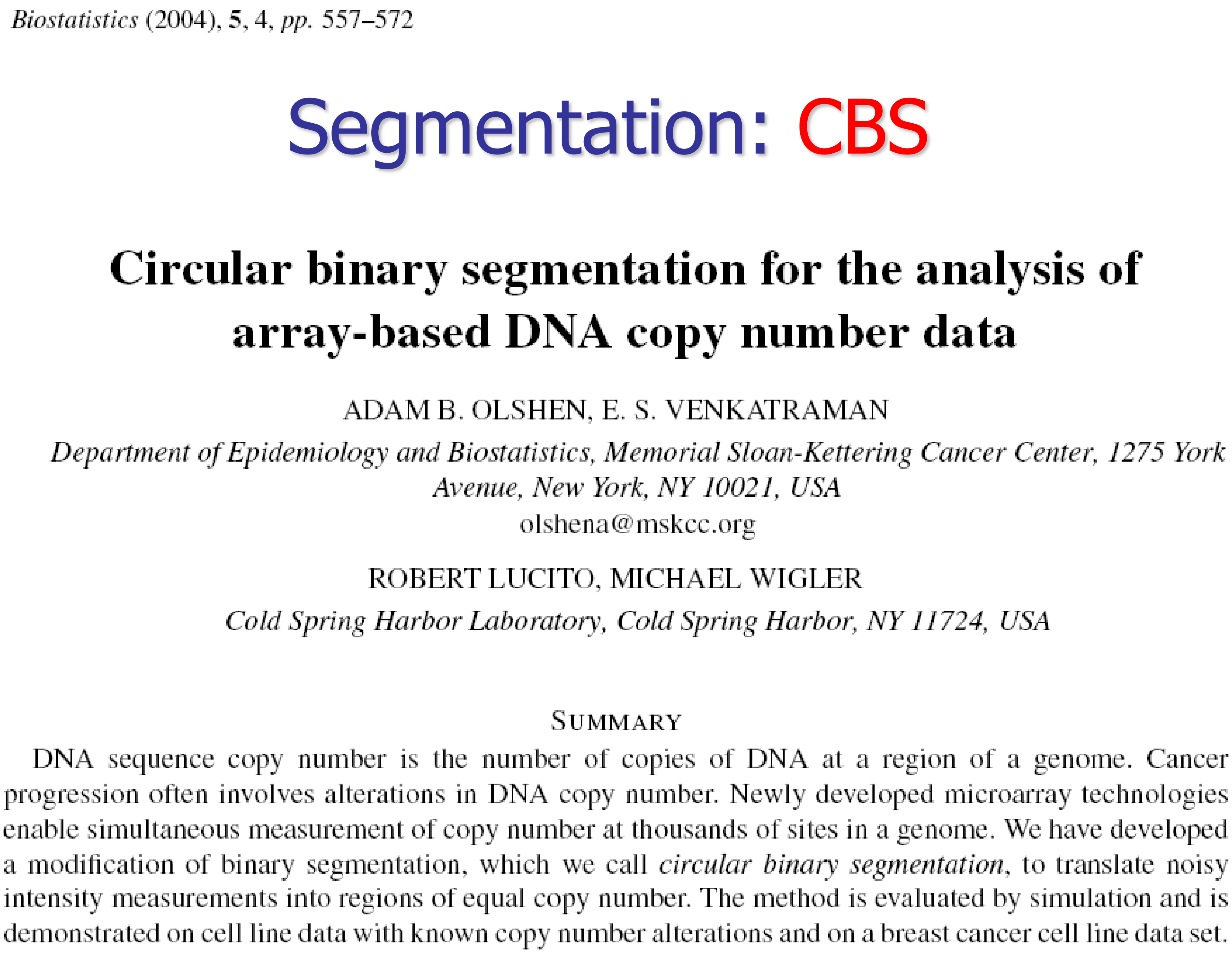
- open source software for analysis of genomic data
- primarily based on the R programming language

Packages in view

Package	Maintainer	Title
aCGH	Jane Fridlyand	Classes and functions for Array Comparative Genomic Hybridization data.
beadarraySNP	Jan Oosting	Normalization and reporting of Illumina SNP bead arrays
CGHbase	Sjoerd Vosse	CGHbase: Base functions and classes for arrayCGH data analysis.
CGHcall	Sjoerd Vosse	Calling aberrations for array CGH tumor profiles.
CGHregions	Mark van de Wiel	Dimension Reduction for Array CGH Data with Minimal Information Loss.
DNAcopy	Venkatraman E. Seshan	DNA copy number data analysis
GLAD	Philippe Hupe	Gain and Loss Analysis of DNA
ITALICS	Guillem Rigaill	ITALICS
KCsmart	Jorma de Ronde	Multi sample aCGH analysis package using kernel convolution
MANOR	Pierre Neuvial	CGH Micro-Array NORmalization
quantsmooth	Jan Oosting	Quantile smoothing and genomic visualization of array data
reb	Karl J. Dykema	Regional Expression Biases
SIM	Marten Boetzer	Integrated Analysis of gene expression and copynumber data
SMAP	Robin Andersson	A Segmental Maximum A Posteriori Approach to Array-CGH Copy Number Profiling
snapCGH	Thomas Hardcastle	Segmentation, normalisation and processing of aCGH data.
SNPchip	Robert Scharpf	Classes and Methods for high throughput SNP chip data
VanillaICE	Robert Scharpf	Methods for fitting Hidden Markov Models to SNP chip data

http://bioconductor.org/packages/2.3/DNACopyNumber.html

Segmentation: CBS



Biostatistics (2004), **5**, 4, *pp.* 557–572

Circular binary segmentation for the analysis of array-based DNA copy number data

ADAM B. OLSHEN, E. S. VENKATRAMAN

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, NY 10021, USA

olshena@mskcc.org

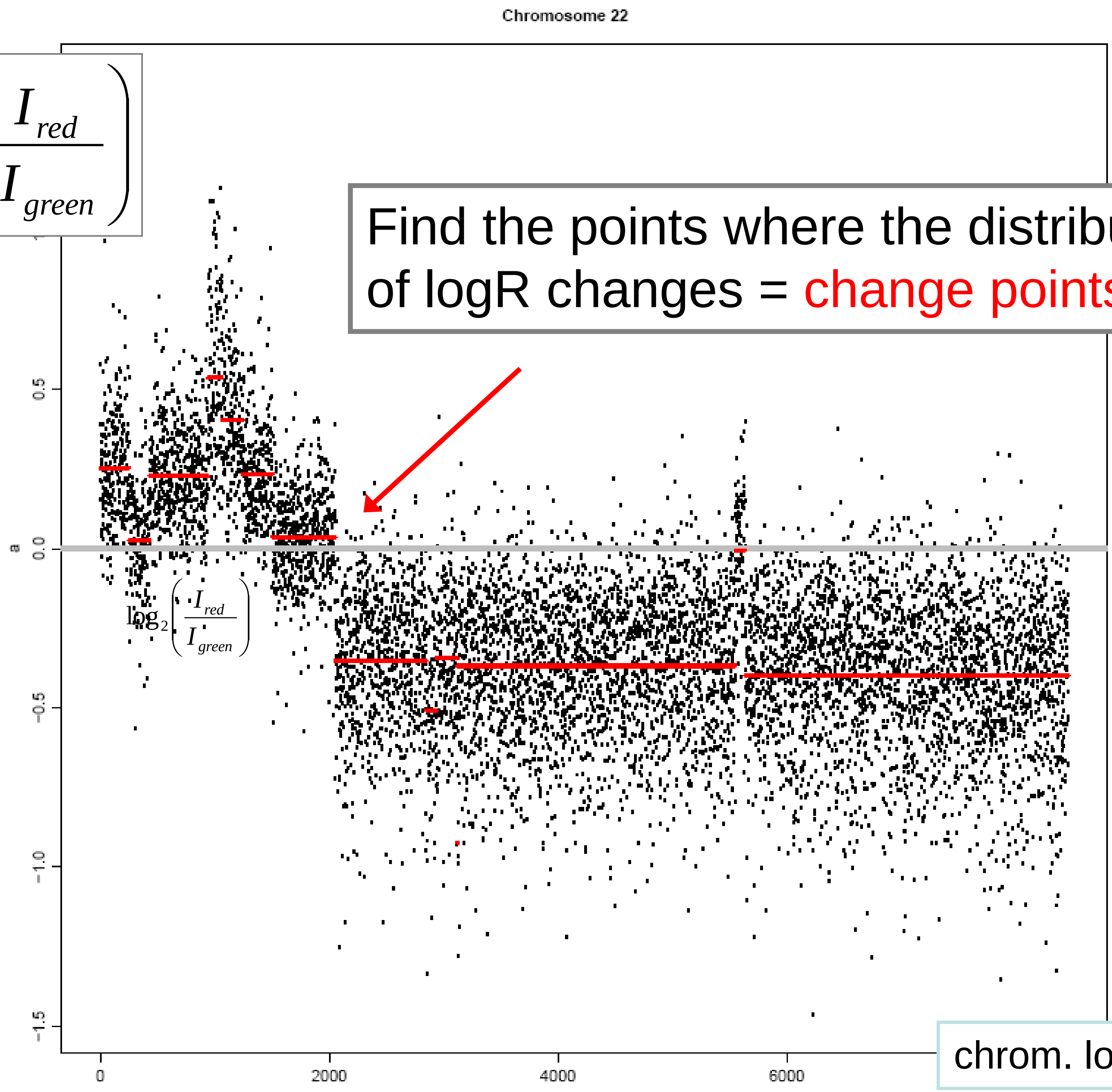
ROBERT LUCITO, MICHAEL WIGLER

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Summary

DNA sequence copy number is the number of copies of DNA at a region of a genome. Cancer progression often involves alterations in DNA copy number. Newly developed microarray technologies enable simultaneous measurement of copy number at thousands of sites in a genome. We have developed a modification of binary segmentation, which we call *circular binary segmentation*, to translate noisy intensity measurements into regions of equal copy number. The method is evaluated by simulation and is demonstrated on cell line data with known copy number alterations and on a breast cancer cell line data set.

$$\log_2\left(\frac{I_{red}}{I_{green}}\right)$$



$$\log_2\left(\frac{I_{red}}{I_{green}}\right)$$

Find the points where the distribution of logR changes = change points

chrom. location

Binary Segmentation
(Sen & Srivastava, 1975)

Likelihood ratio statistics

H_0: no change of the (normal-) distribution

$$S_i = X_1 + \cdots + X_i \,,\, 1 \leqslant i \leqslant n$$

partial sums of logR

$$Z_i = \{1/i + 1/(n-i)\}^{-1/2}\{S_i/i - (S_n - S_i)/(n-i)\}$$

H_0 is rejected if max$|Z_i|$ gets too large → change point at i

statistics used

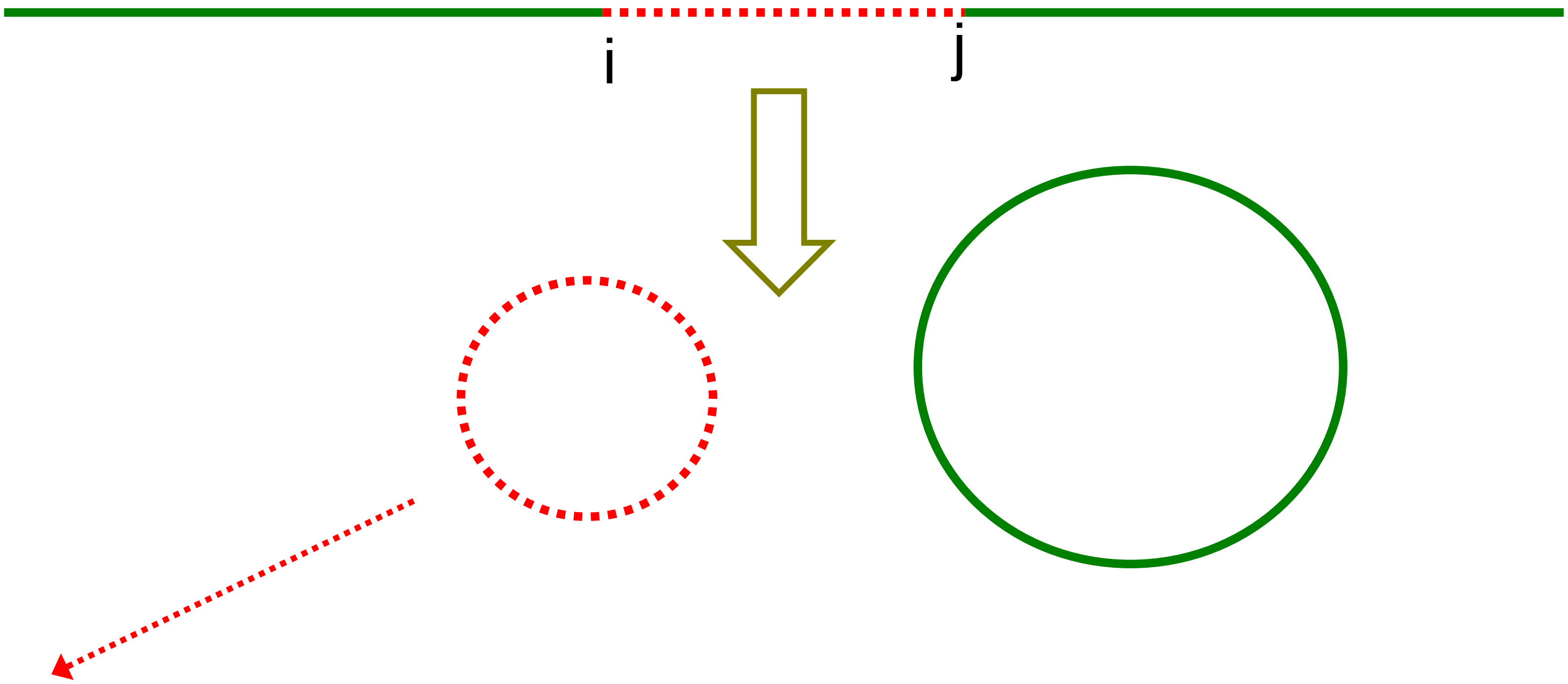
Likelihood ratio statistics

$$H_0: \quad \theta = \theta_0,$$
$$H_1: \quad \theta = \theta_1.$$

$$\Lambda = \frac{L(x \mid \theta_0)}{L(x \mid \theta_1)} \qquad \textit{statistic}$$

If $\Lambda > c$, do not reject H_0
If $\Lambda < c$, reject H_0

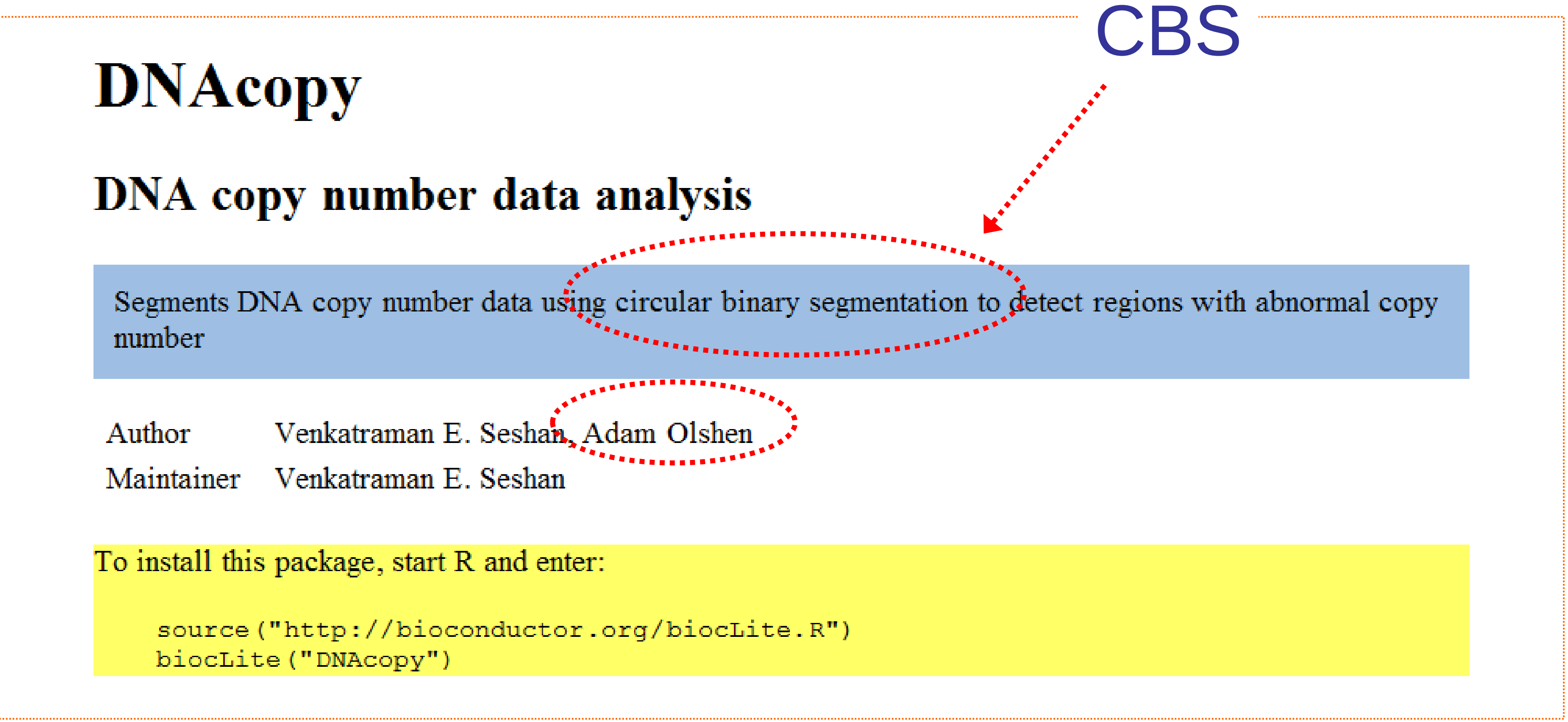
Circular Binary Segmentation



$$Z_{ij} = \{1/(j-i) + 1/(n-j+i)\}^{-1/2}\{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)\}$$

Tests if the arc from i to j has a mean which is different from the mean of the **complement** (reject H_0 if $\max|Z_{ij}|$ too big).

run_CBS runs *DNAcopy* from the Bioconductor package (R)



CBS

DNAcopy

DNA copy number data analysis

Segments DNA copy number data using circular binary segmentation to detect regions with abnormal copy number

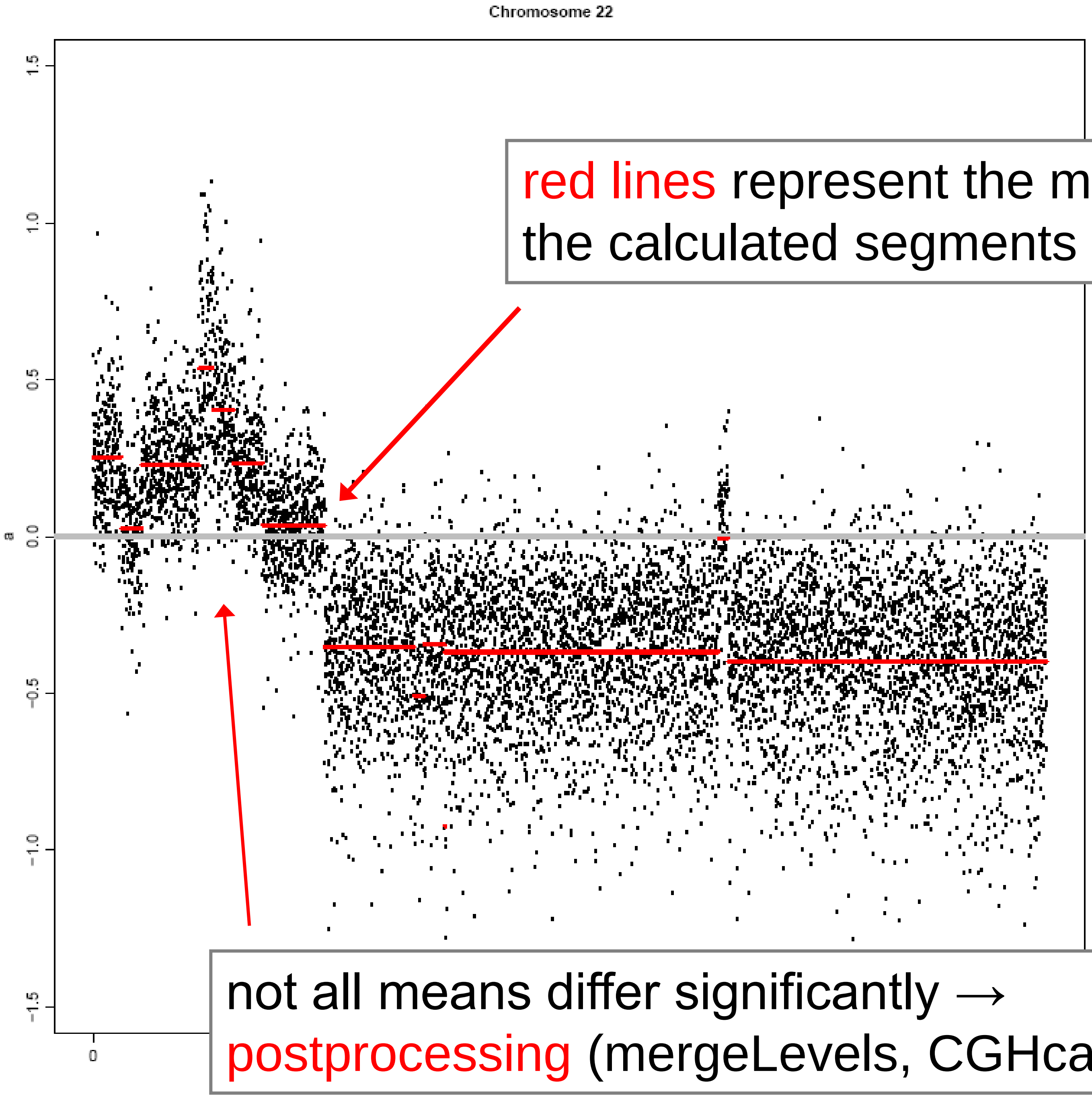
Author Venkatraman E. Seshan, Adam Olshen

Maintainer Venkatraman E. Seshan

To install this package, start R and enter:

```
    source("http://bioconductor.org/biocLite.R")
    biocLite("DNAcopy")
```

http://www.bioconductor.org/packages/2.3/bioc/html/DNAcopy.html

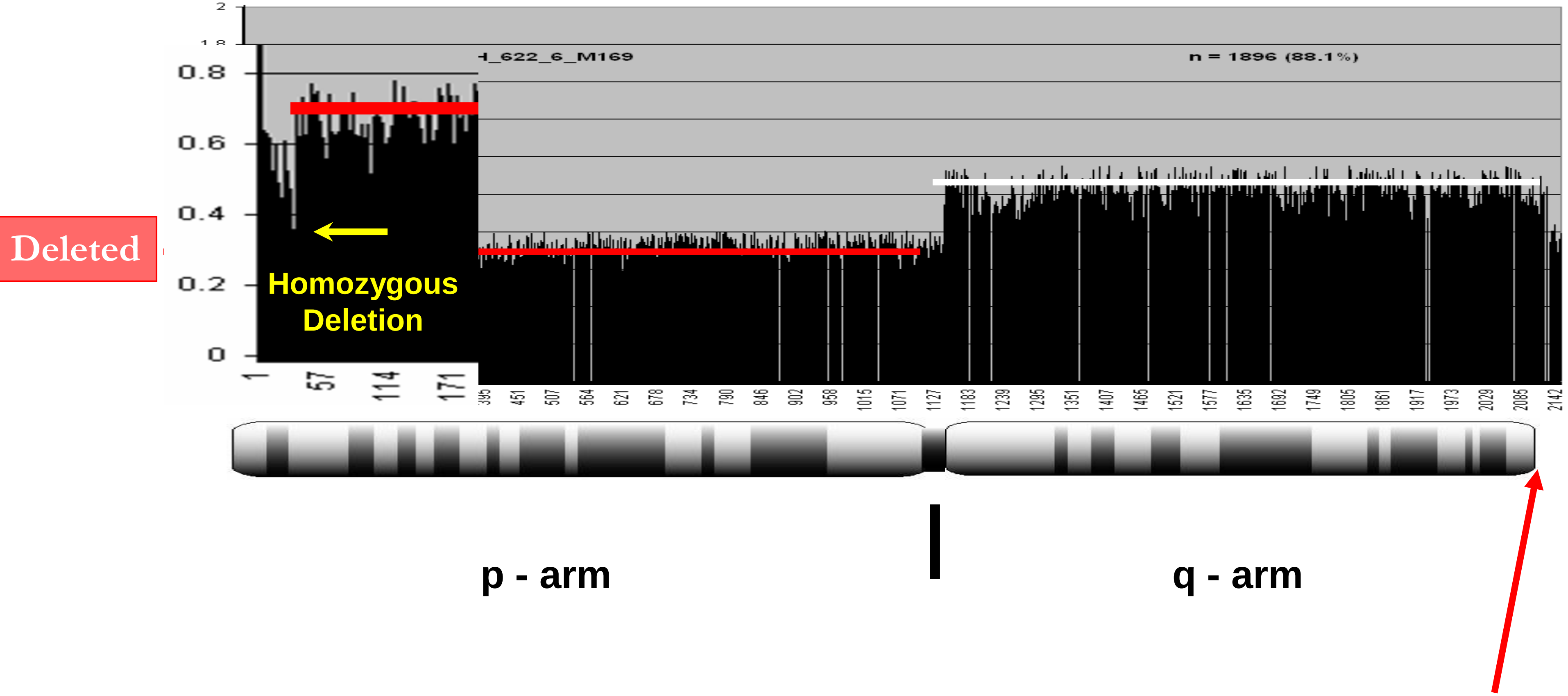


red lines represent the means of the calculated segments

not all means differ significantly → postprocessing (mergeLevels, CGHcall)

from Devins data:
chr22, Tumor 5
run_DNAcopy.R
21/10/08
anaconda

Chromosome 1 array analysis of one tumor (meningioma from a male patient)



Deleted

Homozygous Deletion

p - arm

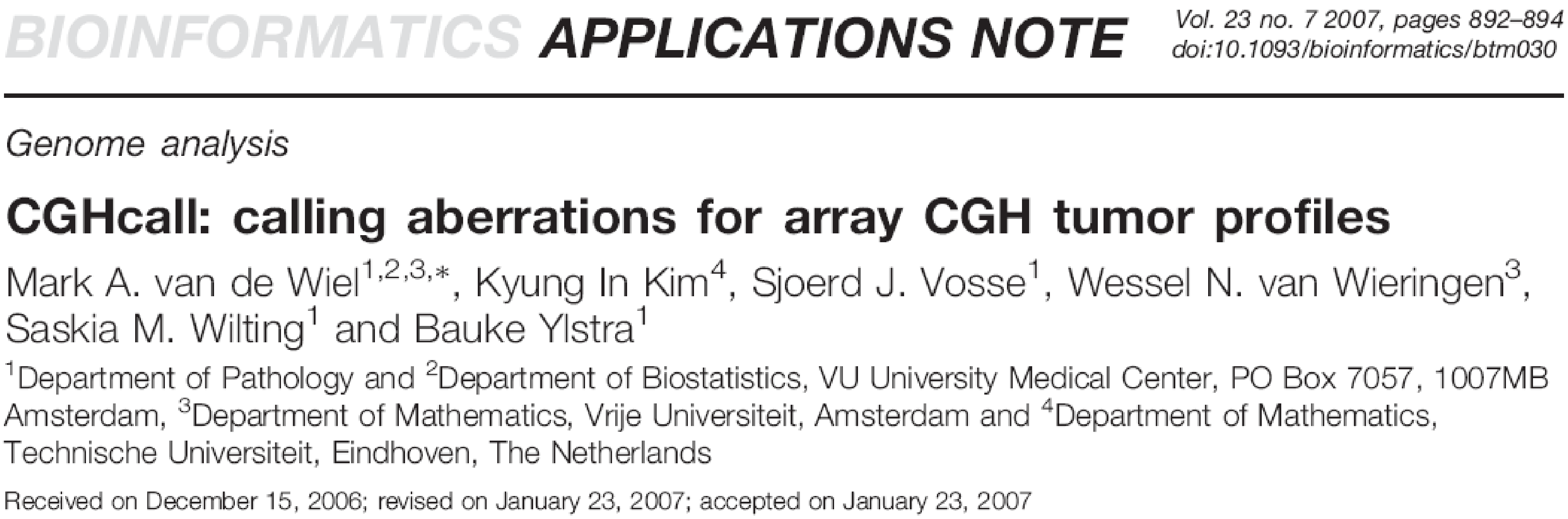
q - arm

Chromosome X - controls

CBS (DNAcopy)

- ... does not make ”calls”

CGHcall



BIOINFORMATICS APPLICATIONS NOTE

Vol. 23 no. 7 2007, pages 892–894
doi:10.1093/bioinformatics/btm030

Genome analysis

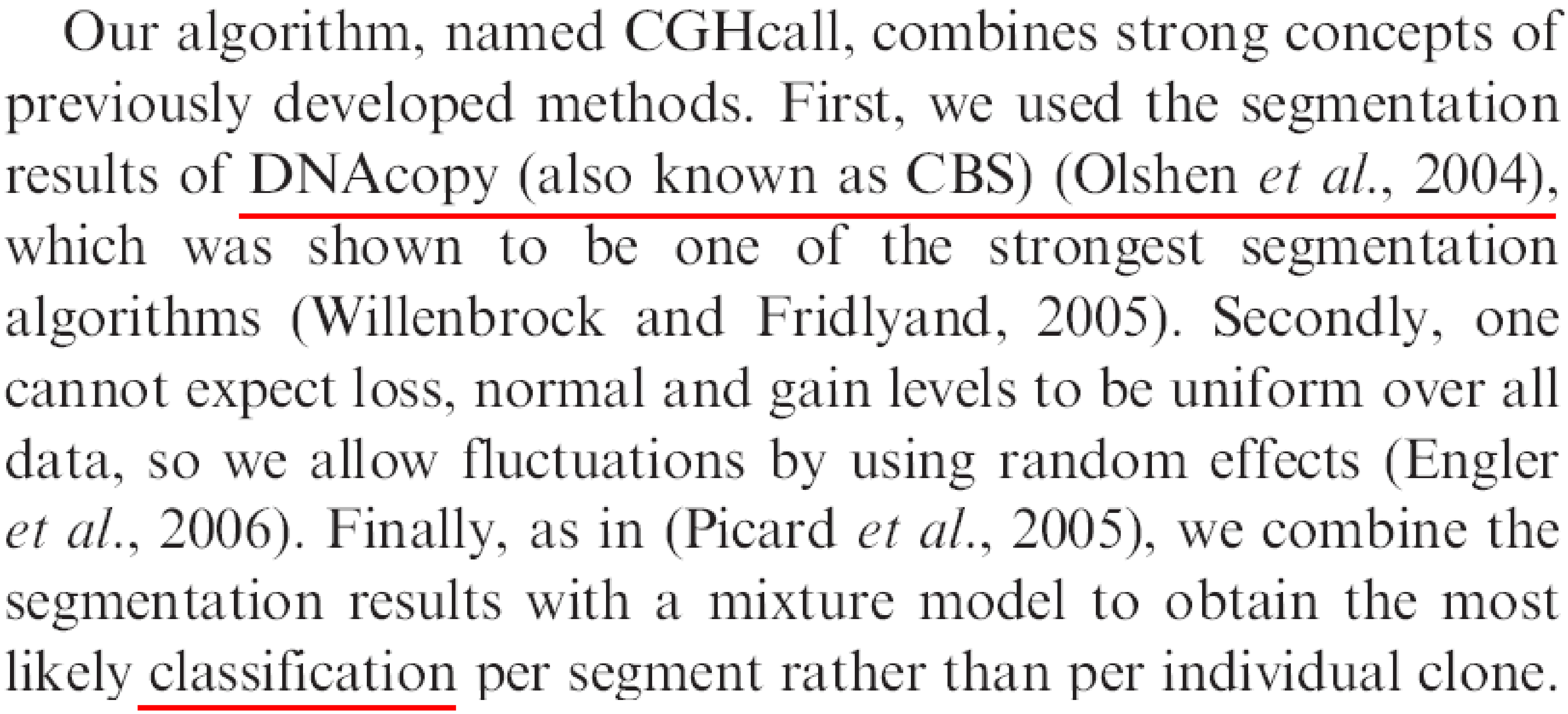
CGHcall: calling aberrations for array CGH tumor profiles

Mark A. van de Wiel[1,2,3,*], Kyung In Kim[4], Sjoerd J. Vosse[1], Wessel N. van Wieringen[3], Saskia M. Wilting[1] and Bauke Ylstra[1]

[1]Department of Pathology and [2]Department of Biostatistics, VU University Medical Center, PO Box 7057, 1007MB Amsterdam, [3]Department of Mathematics, Vrije Universiteit, Amsterdam and [4]Department of Mathematics, Technische Universiteit, Eindhoven, The Netherlands

Received on December 15, 2006; revised on January 23, 2007; accepted on January 23, 2007

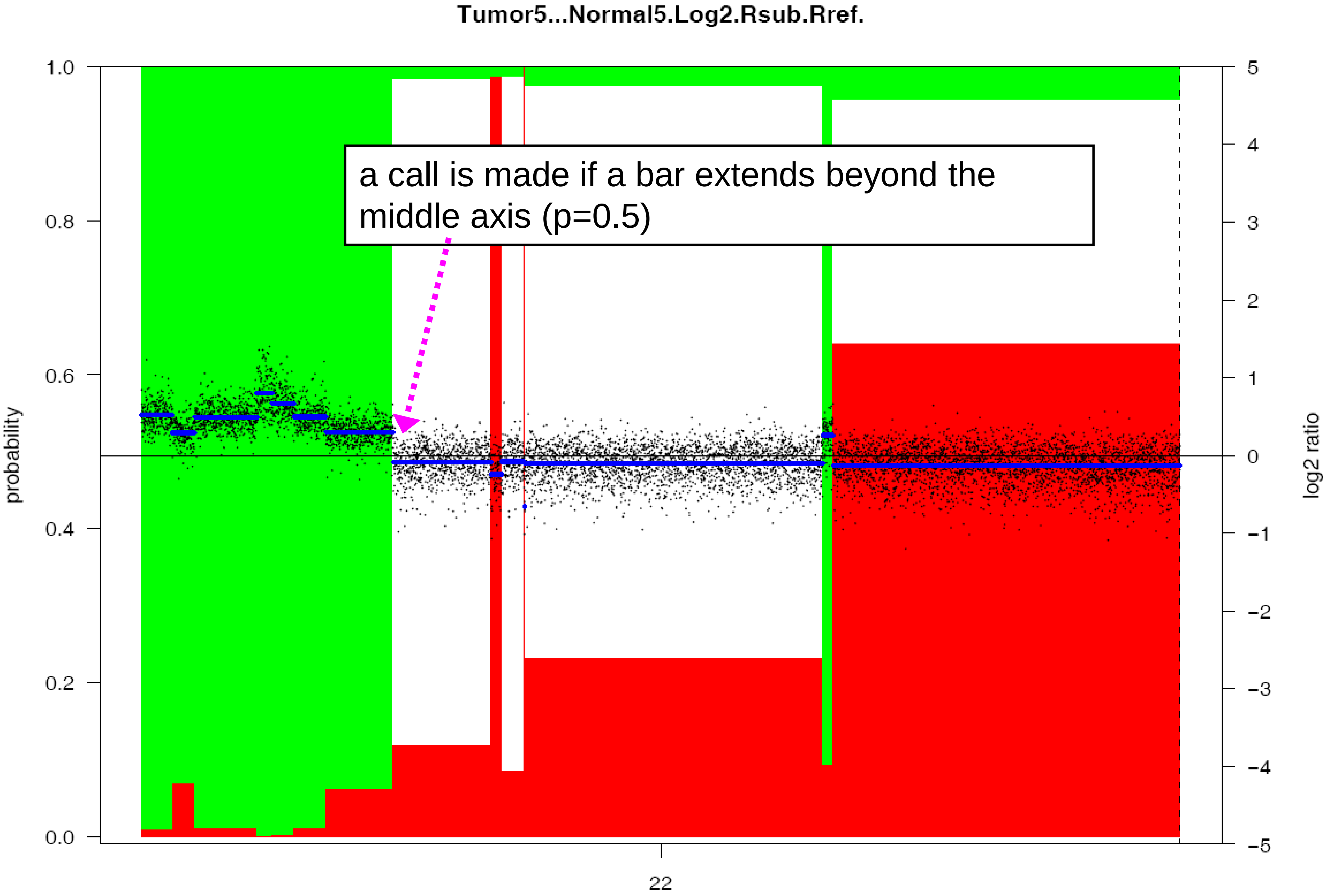
CGHcall



Our algorithm, named CGHcall, combines strong concepts of previously developed methods. First, we used the segmentation results of DNAcopy (also known as CBS) (Olshen *et al.*, 2004), which was shown to be one of the strongest segmentation algorithms (Willenbrock and Fridlyand, 2005). Secondly, one cannot expect loss, normal and gain levels to be uniform over all data, so we allow fluctuations by using random effects (Engler *et al.*, 2006). Finally, as in (Picard *et al.*, 2005), we combine the segmentation results with a mixture model to obtain the most likely classification per segment rather than per individual clone.

”calls”

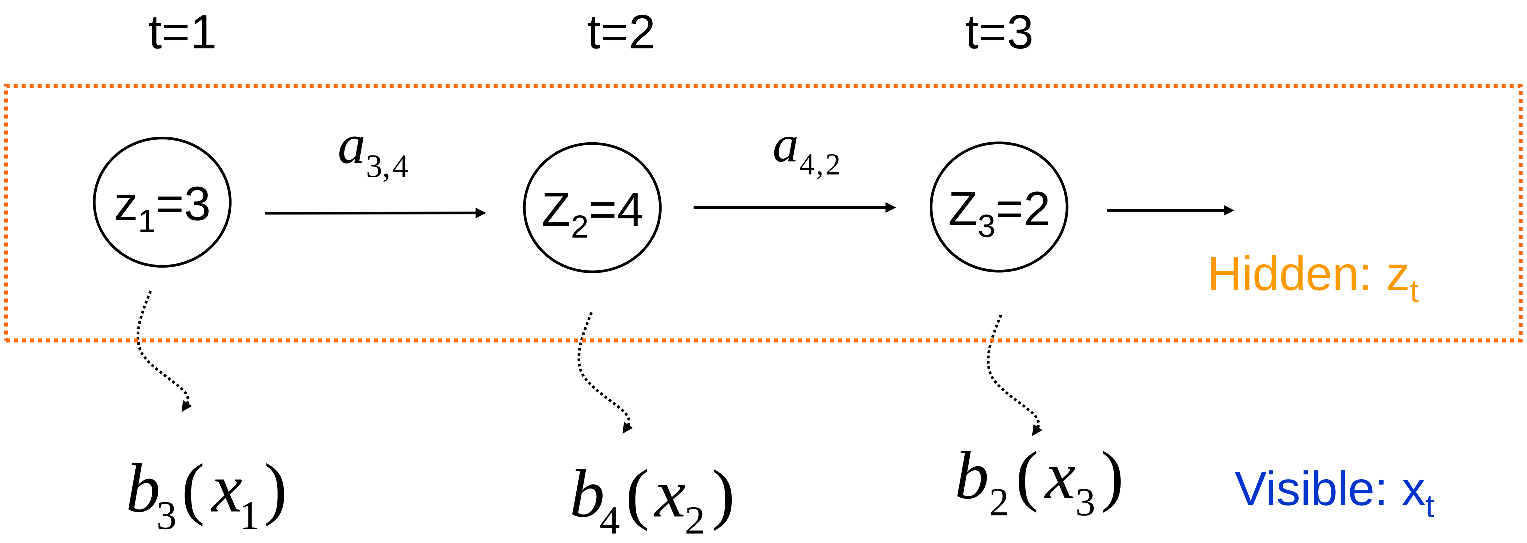
run_CGHcall.R



a call is made if a bar extends beyond the middle axis (p=0.5)

anaconda:/nfs/1d/menzel/TEST_DNAcopy/Sample_550K_Paired_LogR_chr22_nos4.txt.sample4.pdf

Hidden Markov Model

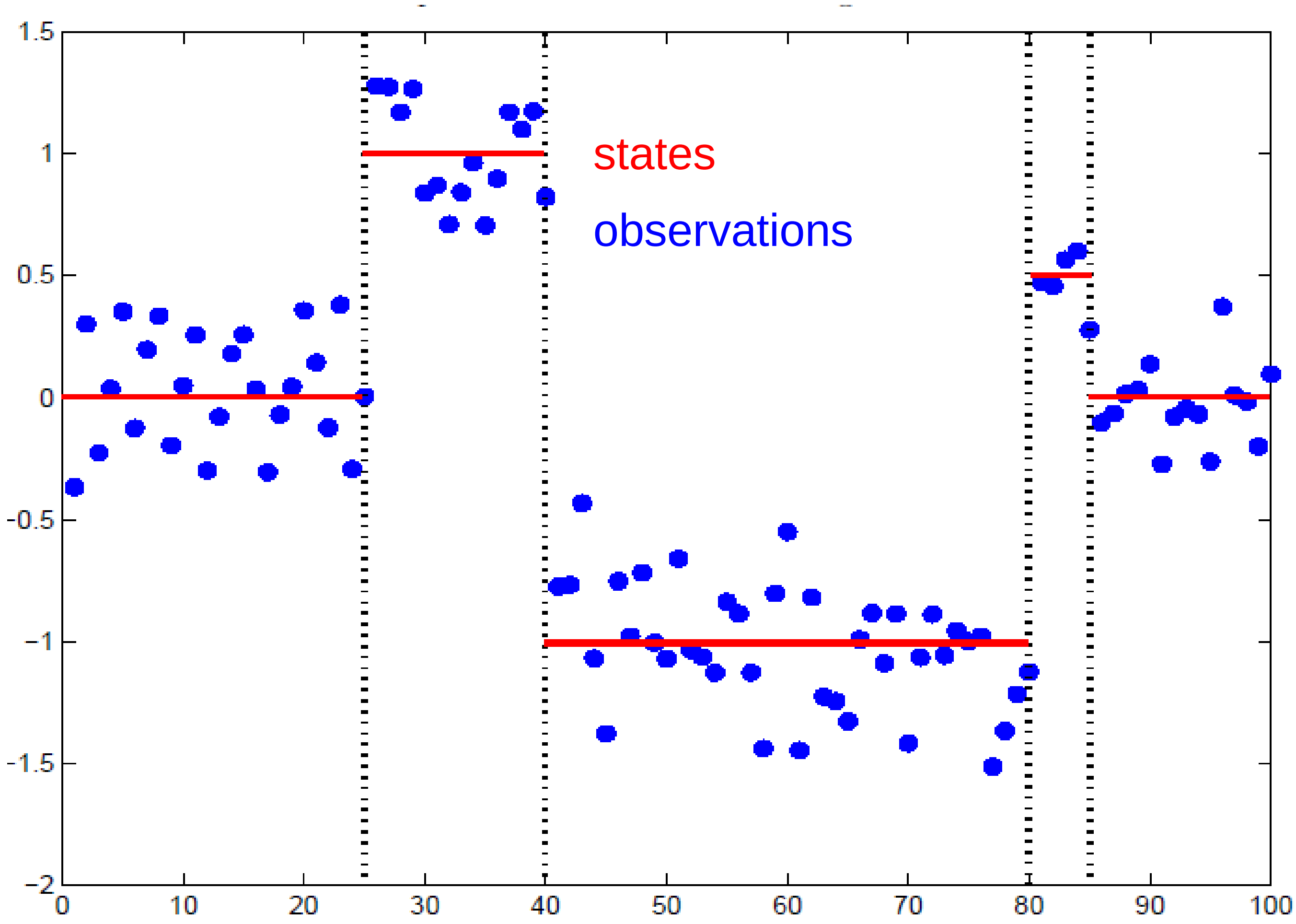


$$P(x_1, x_2, \ldots, z_1, z_2, \ldots \mid \theta) = p(z_1) \cdot \quad b_{z_1}(x_1) \cdot a_{z_1, z_2} \cdot \quad b_{z_2}(x_2) \cdot a_{z_2, z_3} \cdot \quad \ldots$$

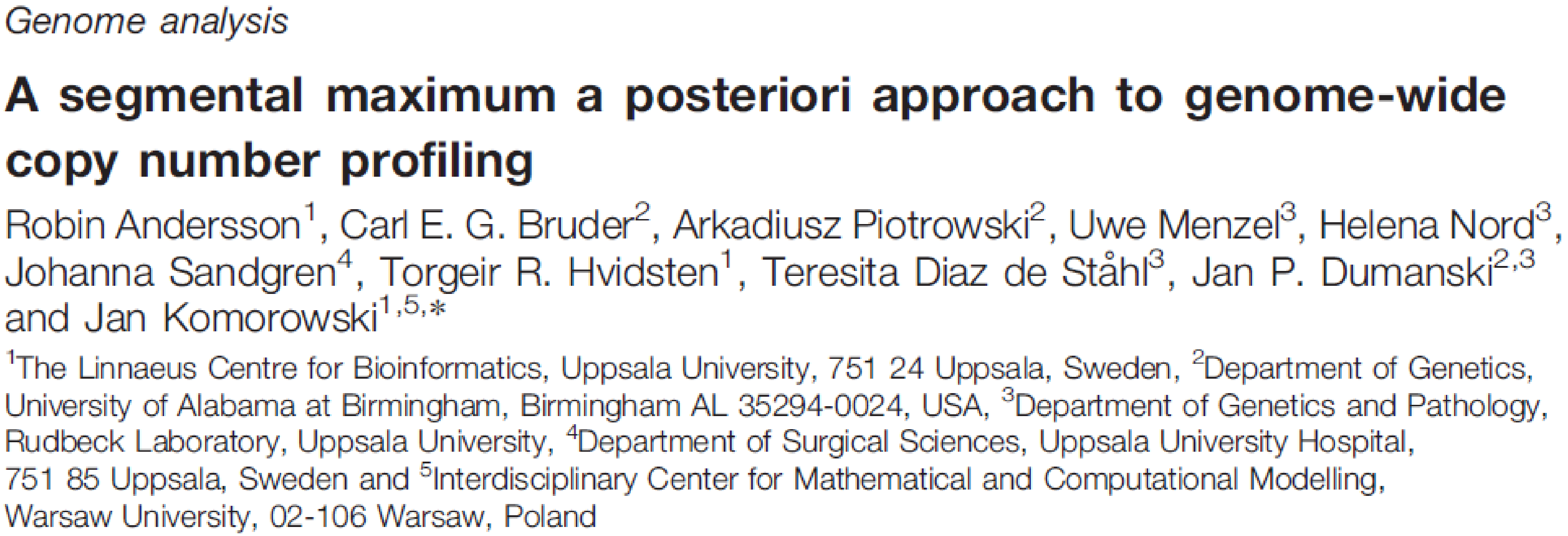
$$P(x, z \mid \theta) = p(z_1) \cdot \prod_{i=1}^{T} b_{z_i}(x_i) \cdot a_{z_i, z_{i+1}}$$

CDHMM

Continuous Density Hidden Markov Model



SMAP

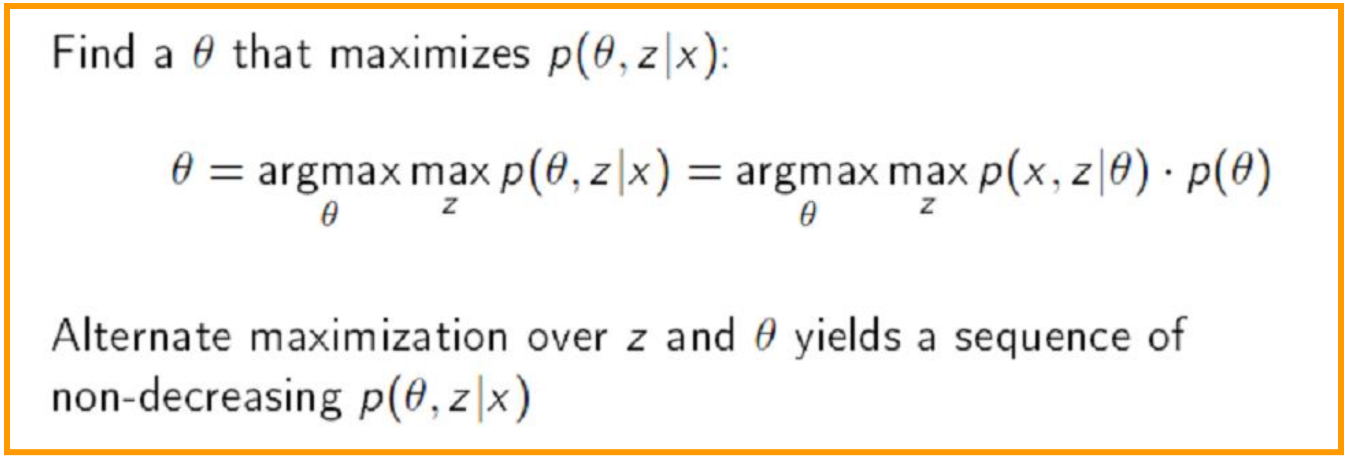


Genome analysis

A segmental maximum a posteriori approach to genome-wide copy number profiling

Robin Andersson[1], Carl E. G. Bruder[2], Arkadiusz Piotrowski[2], Uwe Menzel[3], Helena Nord[3], Johanna Sandgren[4], Torgeir R. Hvidsten[1], Teresita Diaz de Ståhl[3], Jan P. Dumanski[2,3] and Jan Komorowski[1,5,*]

[1]The Linnaeus Centre for Bioinformatics, Uppsala University, 751 24 Uppsala, Sweden, [2]Department of Genetics, University of Alabama at Birmingham, Birmingham AL 35294-0024, USA, [3]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, [4]Department of Surgical Sciences, Uppsala University Hospital, 751 85 Uppsala, Sweden and [5]Interdisciplinary Center for Mathematical and Computational Modelling, Warsaw University, 02-106 Warsaw, Poland



Find a θ that maximizes $p(\theta, z|x)$:

$$\theta = \underset{\theta}{\operatorname{argmax}} \max_{z} p(\theta, z|x) = \underset{\theta}{\operatorname{argmax}} \max_{z} p(x, z|\theta) \cdot p(\theta)$$

Alternate maximization over z and θ yields a sequence of non-decreasing $p(\theta, z|x)$

Maximum likelihood estimation – MAP

f(*x* | θ) be the probability of *x* when the underlying population parameter is θ.

$$\theta \mapsto f(x|\theta)$$ ML function

$$\hat{\theta}_{\mathrm{ML}}(x) = \arg\max_{\theta} f(x|\theta)$$ ML estimation of ⍰

MAP estimation of ⍰:

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} \frac{f(x|\theta)\, g(\theta)}{\displaystyle\int_{\Theta} f(x|\theta')\, g(\theta')\, d\theta'} = \arg\max_{\theta} f(x|\theta)\, g(\theta).$$

If the prior is flat, i.e. g(⍰)=C → MAP estimate is the same as the ML estimation

Segmental MAP

x – observation, known
z – states
⍰ - parameters

$$p(\theta, z|x) = \frac{p(z,\theta,x)}{p(x)} = \frac{p(z,x|\theta)\cdot p(\theta)}{p(x)}$$

Find a θ that maximizes $p(\theta, z|x)$:

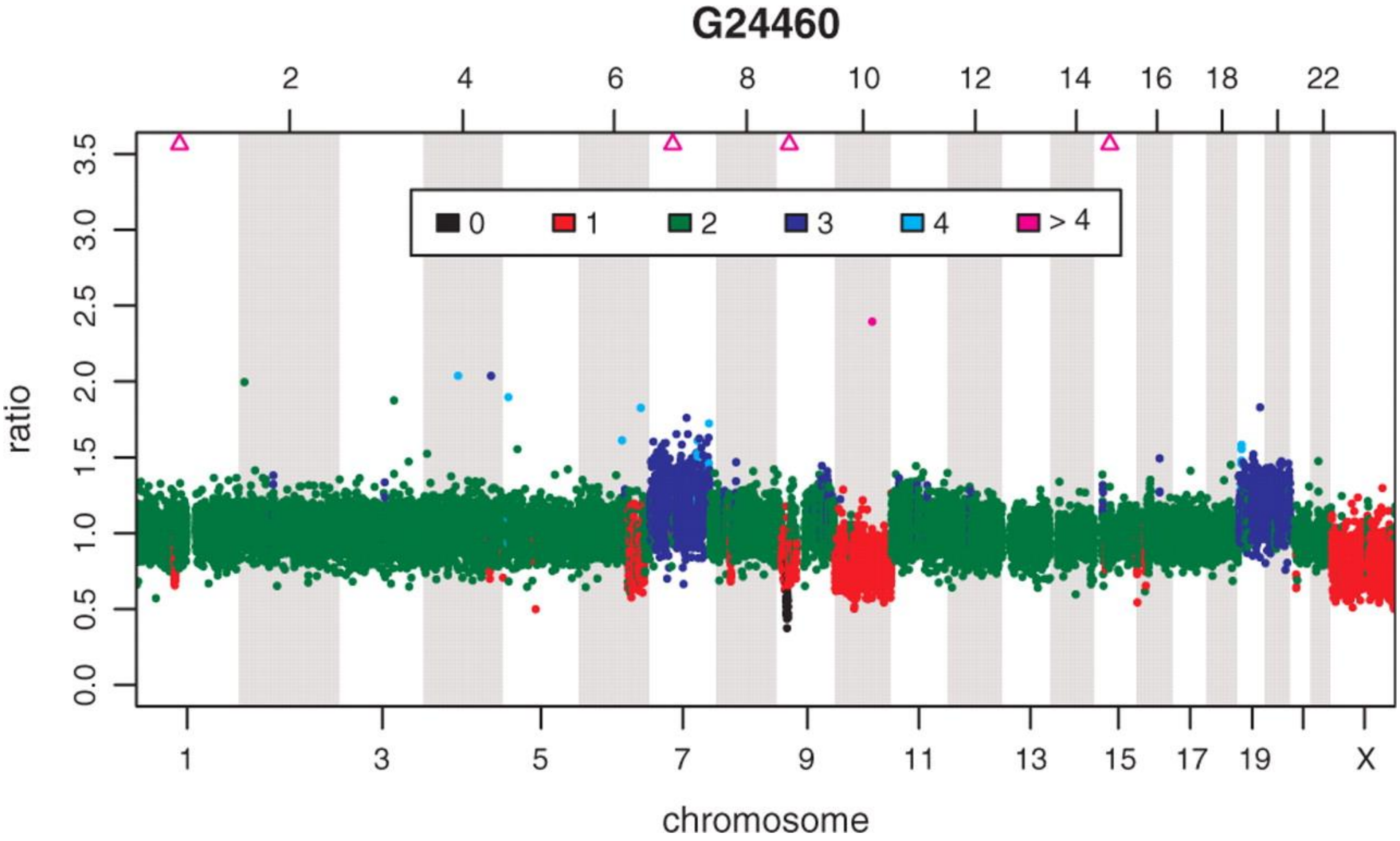
$$\theta = \underset{\theta}{\operatorname{argmax}} \max_{z} p(\theta, z|x) = \underset{\theta}{\operatorname{argmax}} \max_{z} p(x, z|\theta) \cdot p(\theta)$$

Alternate maximization over z and θ yields a sequence of non-decreasing $p(\theta, z|x)$:

proof!

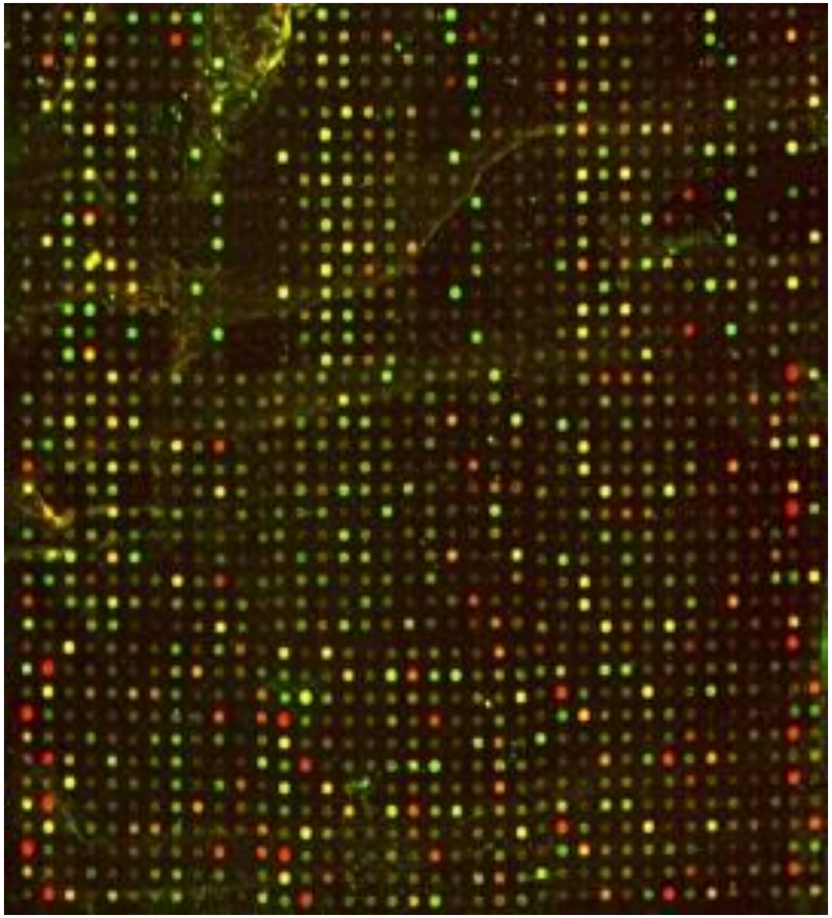
$$\begin{aligned} z_{t+1} &= \underset{z}{\operatorname{argmax}}\, p(x, z|\theta_t) && \text{Viterbi} \\ \theta_{t+1} &= \underset{\theta}{\operatorname{argmax}}\, p(x, z_{t+1}|\theta) \cdot p(\theta) \end{aligned}$$

SMAP - Result



G24460
2
4
6
8
10
12
14
16
18
22
■ 0
■ 1
■ 2
■ 3
■ 4
■ > 4
ratio
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1
3
5
7
9
11
13
15
17
19
X
chromosome

PCR-based arrays



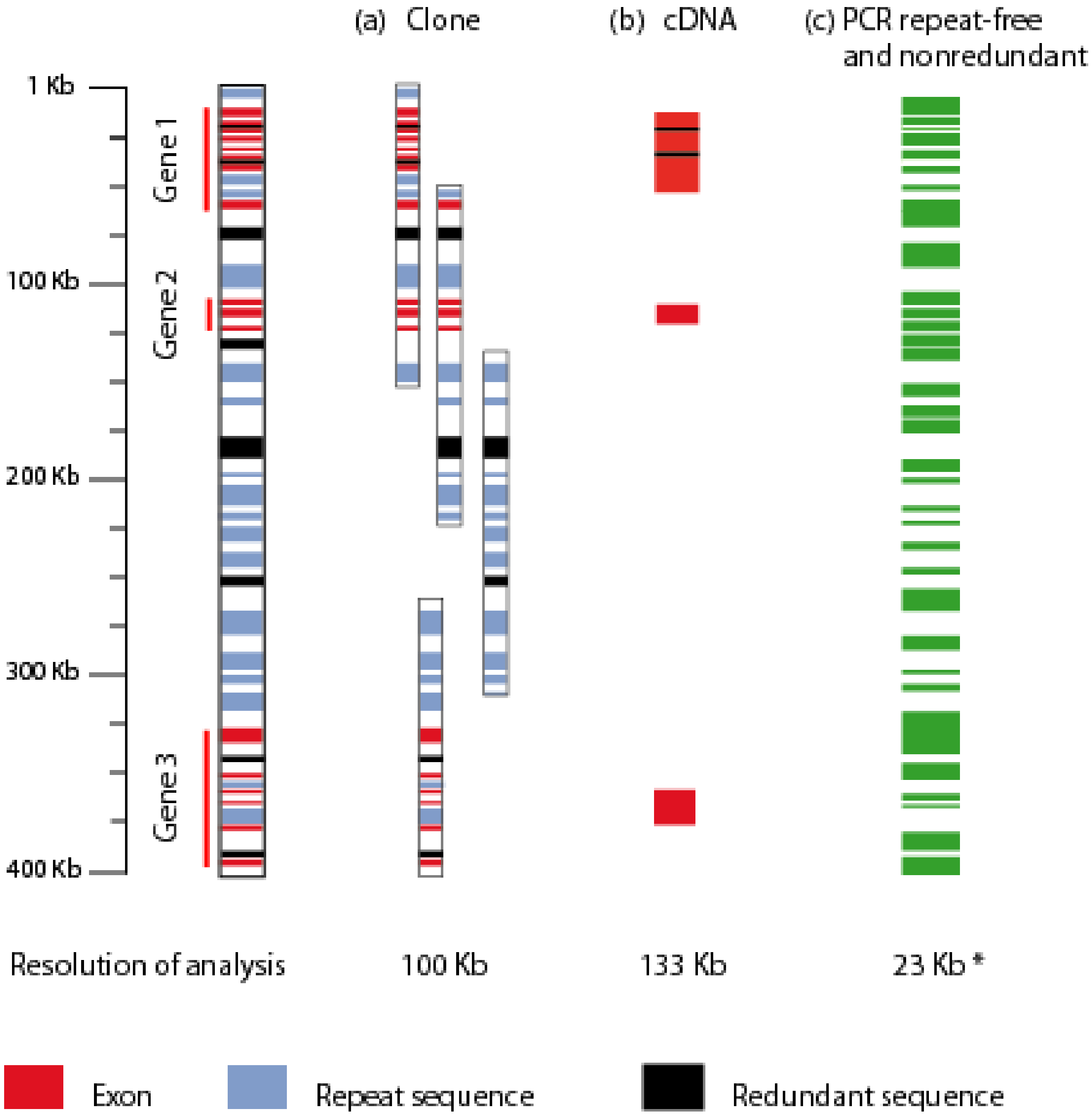
(Pools of) PCR products

Comparison between 3 approaches to construct genomic arrays:

(a) genomic clone-based;

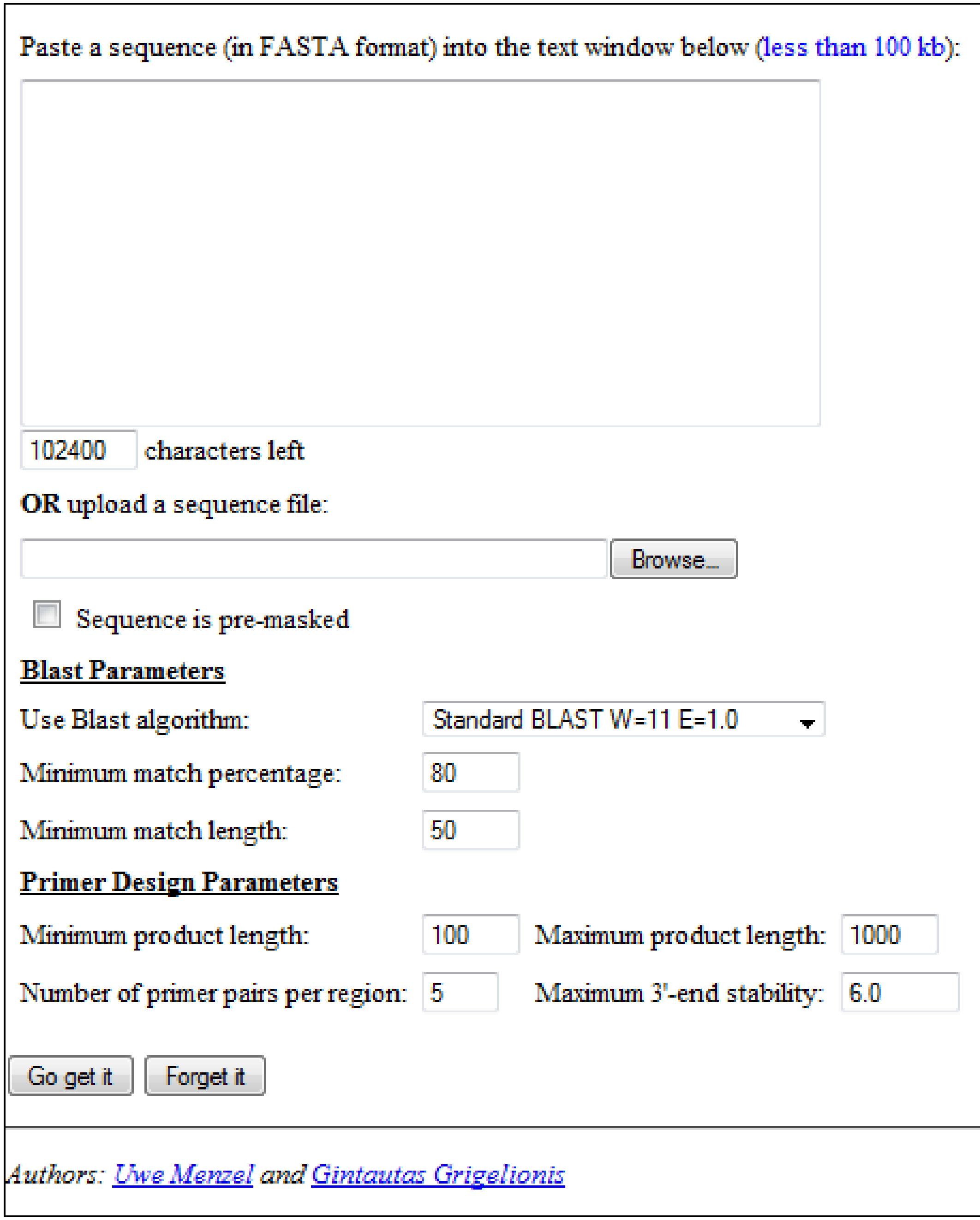
(b) cDNA-based;

(c) repeat-free and non redundant



"Allocator"

automaticly, find repeat-free and non-redundant regions in a certain chromosomal region and define unique primer pairs on it



Paste a sequence (in FASTA format) into the text window below (less than 100 kb):

102400 characters left

OR upload a sequence file:

□ Sequence is pre-masked

Blast Parameters

Use Blast algorithm: Standard BLAST W=11 E=1.0

Minimum match percentage: 80

Minimum match length: 50

Primer Design Parameters

Minimum product length: 100

Maximum product length: 1000

Number of primer pairs per region: 5

Maximum 3'-end stability: 6.0

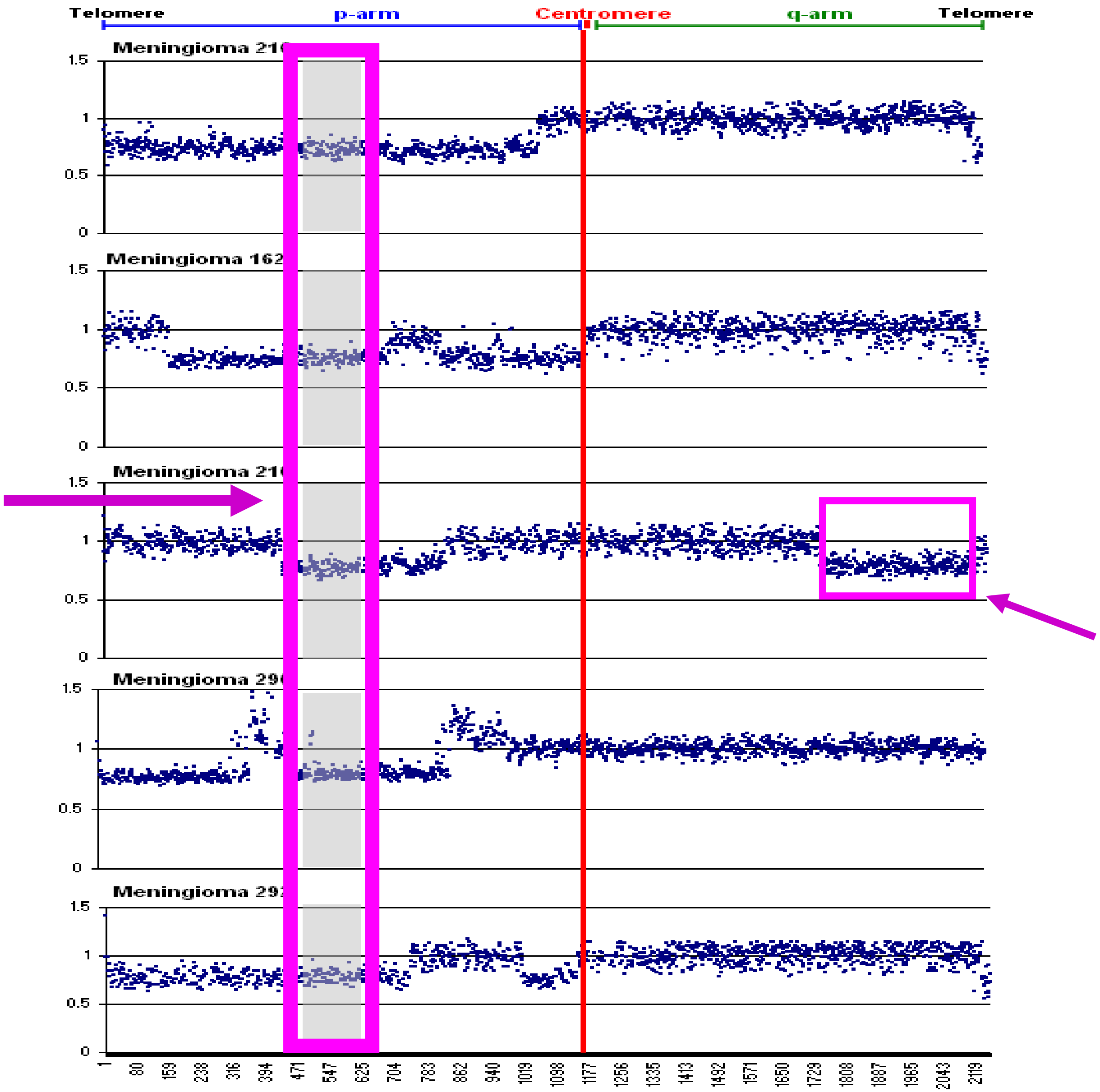
Authors: Uwe Menzel and Gintautas Grigelionis

Clinically relevant findings

- find changes that are characteristic for a certain kind of tumor
- phenotype ⍰ genotype
- identify genes in this deleted regions: TSG/Oncogene
- pathway analysis (GO, KEGG)

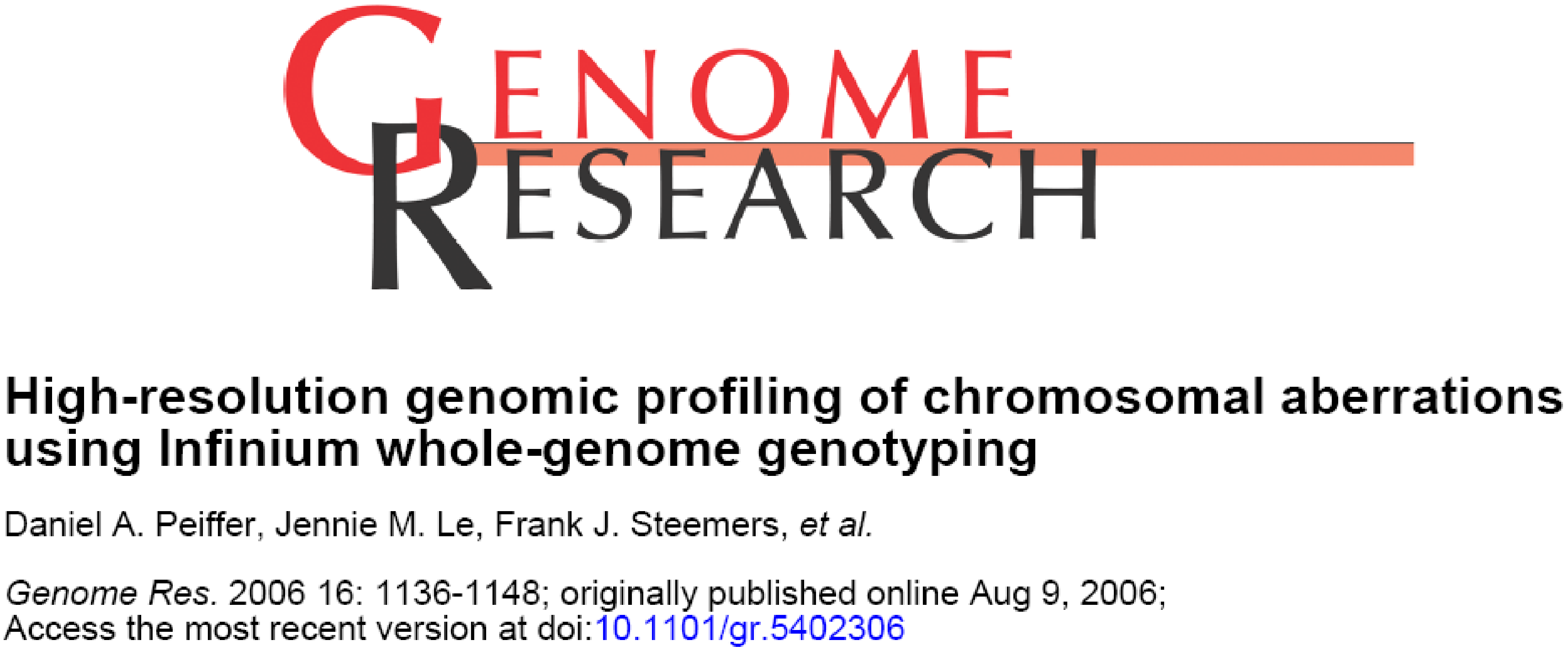
Six meningiomas analyzed on chr. 1 array

Analysis of 1p will allow to define a small overlapping region of deletions



Deletions on 1q have not been described so far in meningioma

”SNP-CGH”



GENOME RESEARCH

High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping

Daniel A. Peiffer, Jennie M. Le, Frank J. Steemers, *et al.*

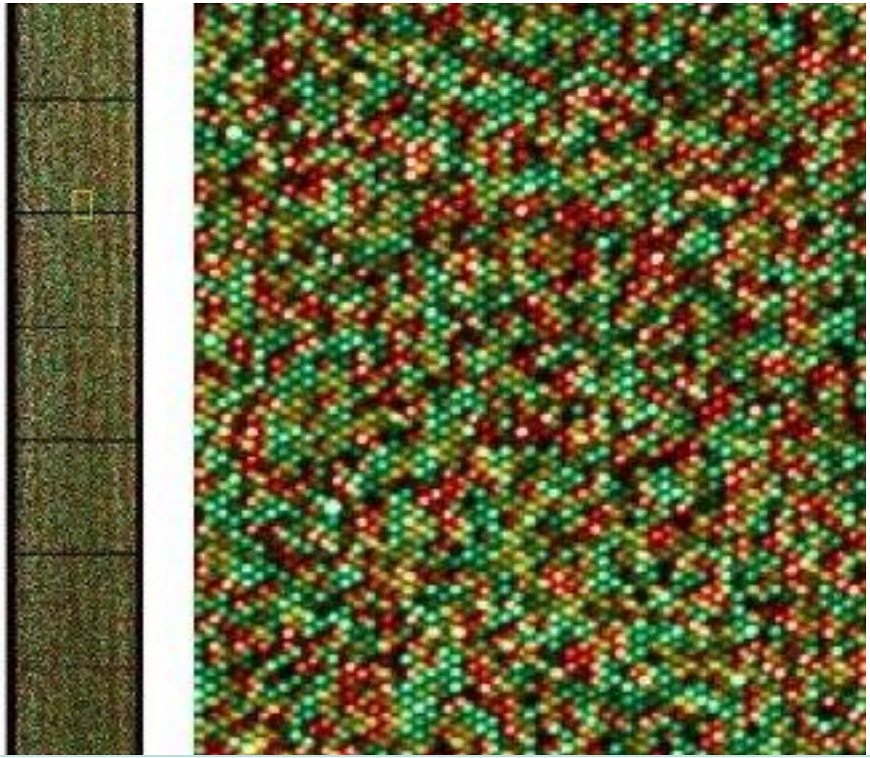
Genome Res. 2006 16: 1136-1148; originally published online Aug 9, 2006;
Access the most recent version at doi:10.1101/gr.5402306

High-resolution genomic profiling of chromosomal aberrations.pdf

”SNP-CGH”

- simultaneous measurement of both signal intensity and allelic composition
- detect both copy number changes and copy-neutral loss-of-heterozygosity (LOH)
- Infinium whole-genome genotyping (WGG) BeadChips (Illumina)

SNP arrays



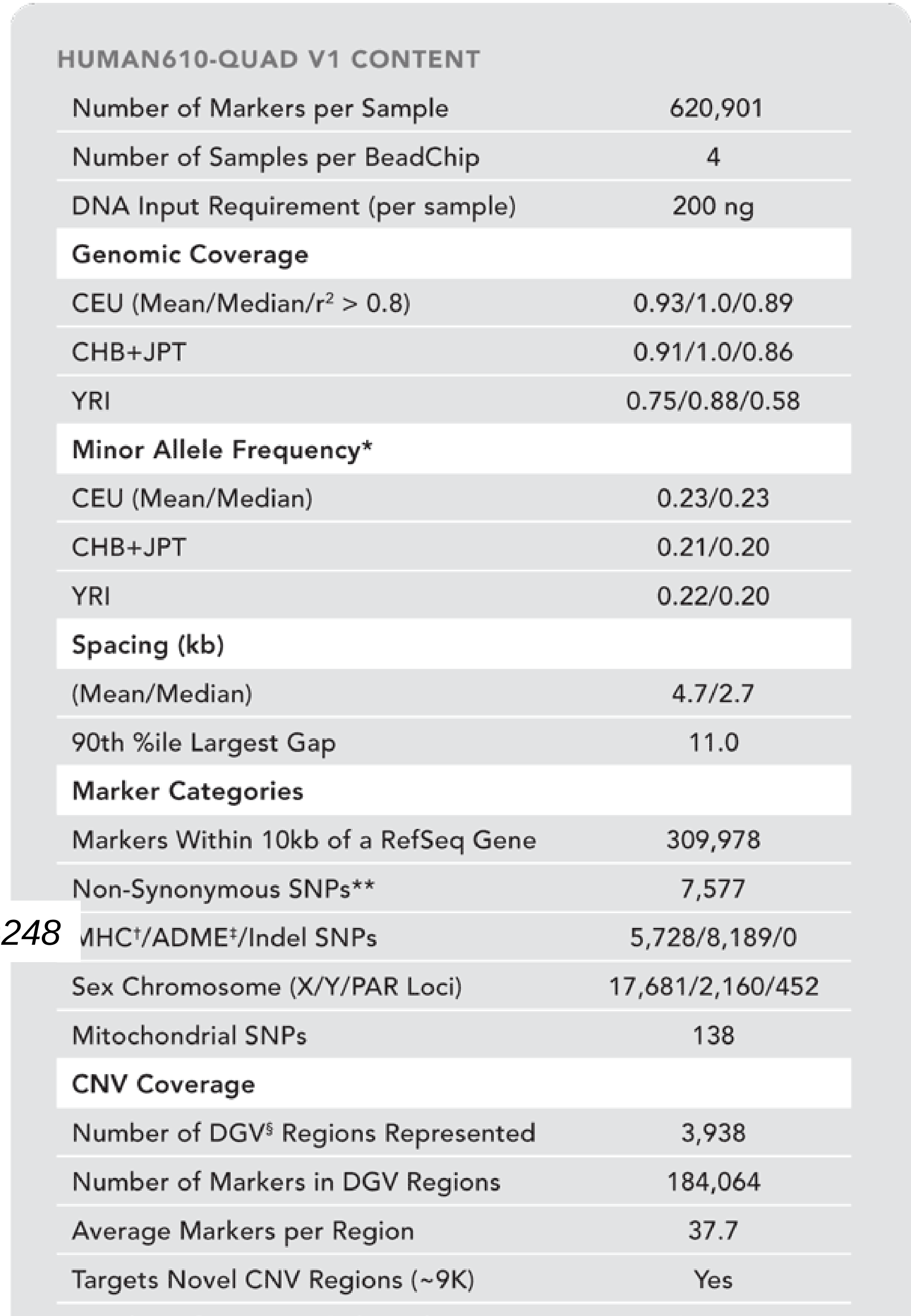
Oligonucleotides representing SNPs

- 610,000 rationally selected tag SNPs per sample
- captures the majority of known variations (haplotypes) (based on HapMap[1] release 23)

human610-quad beadchip

Genotyping & CNV analysis

http://www.illumina.com/pages.ilmn?ID=248

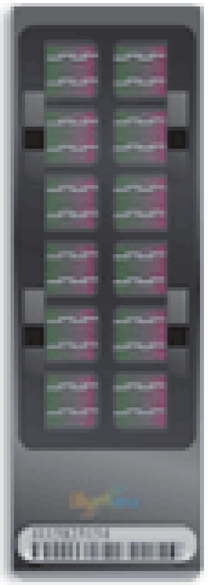


HUMAN610-QUAD V1 CONTENT

Number of Markers per Sample	620,901
Number of Samples per BeadChip	4
DNA Input Requirement (per sample)	200 ng
Genomic Coverage	
CEU (Mean/Median/$r^2 > 0.8$)	0.93/1.0/0.89
CHB+JPT	0.91/1.0/0.86
YRI	0.75/0.88/0.58
Minor Allele Frequency*	
CEU (Mean/Median)	0.23/0.23
CHB+JPT	0.21/0.20
YRI	0.22/0.20
Spacing (kb)	
(Mean/Median)	4.7/2.7
90th %ile Largest Gap	11.0
Marker Categories	
Markers Within 10kb of a RefSeq Gene	309,978
Non-Synonymous SNPs**	7,577
MHC[†]/ADME[‡]/Indel SNPs	5,728/8,189/0
Sex Chromosome (X/Y/PAR Loci)	17,681/2,160/452
Mitochondrial SNPs	138
CNV Coverage	
Number of DGV[§] Regions Represented	3,938
Number of Markers in DGV Regions	184,064
Average Markers per Region	37.7
Targets Novel CNV Regions (~9K)	Yes

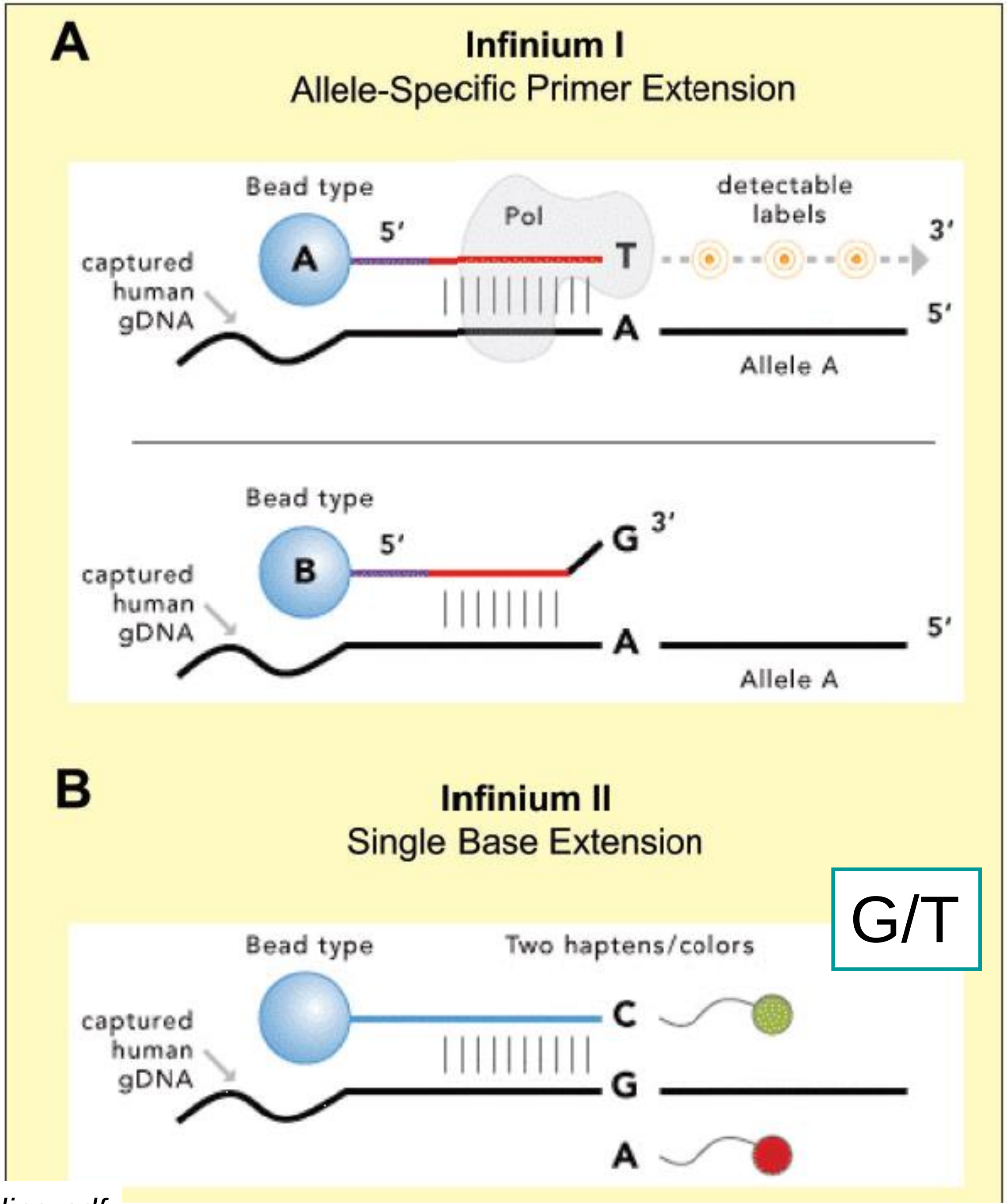
human610-quad beadchip

- 610,000 rationally selected tag SNPs and markers per sample
- captures the majority of known variations (haplotypes) (based on HapMap[1] release 23)
- detection of both known and novel CNV regions



[1] *http://www.hapmap.org/whatishapmap.html.en*

What B-allele frequency



A

Infinium I

Allele-Specific Primer Extension

B

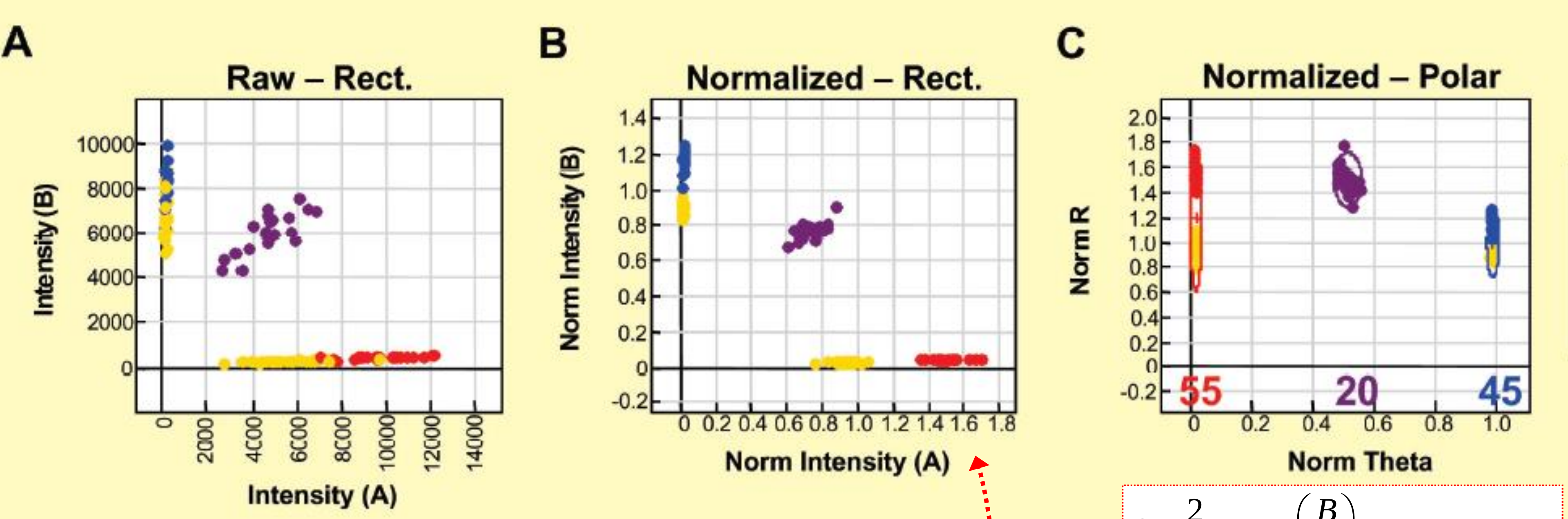
Infinium II

Single Base Extension

G/T

SNP-CGH technologies for genomic profiling.pdf

Calculation of the BAF



$$\theta = \frac{2}{\pi}\arctan\left(\frac{B}{A}\right) \qquad R = A + B$$

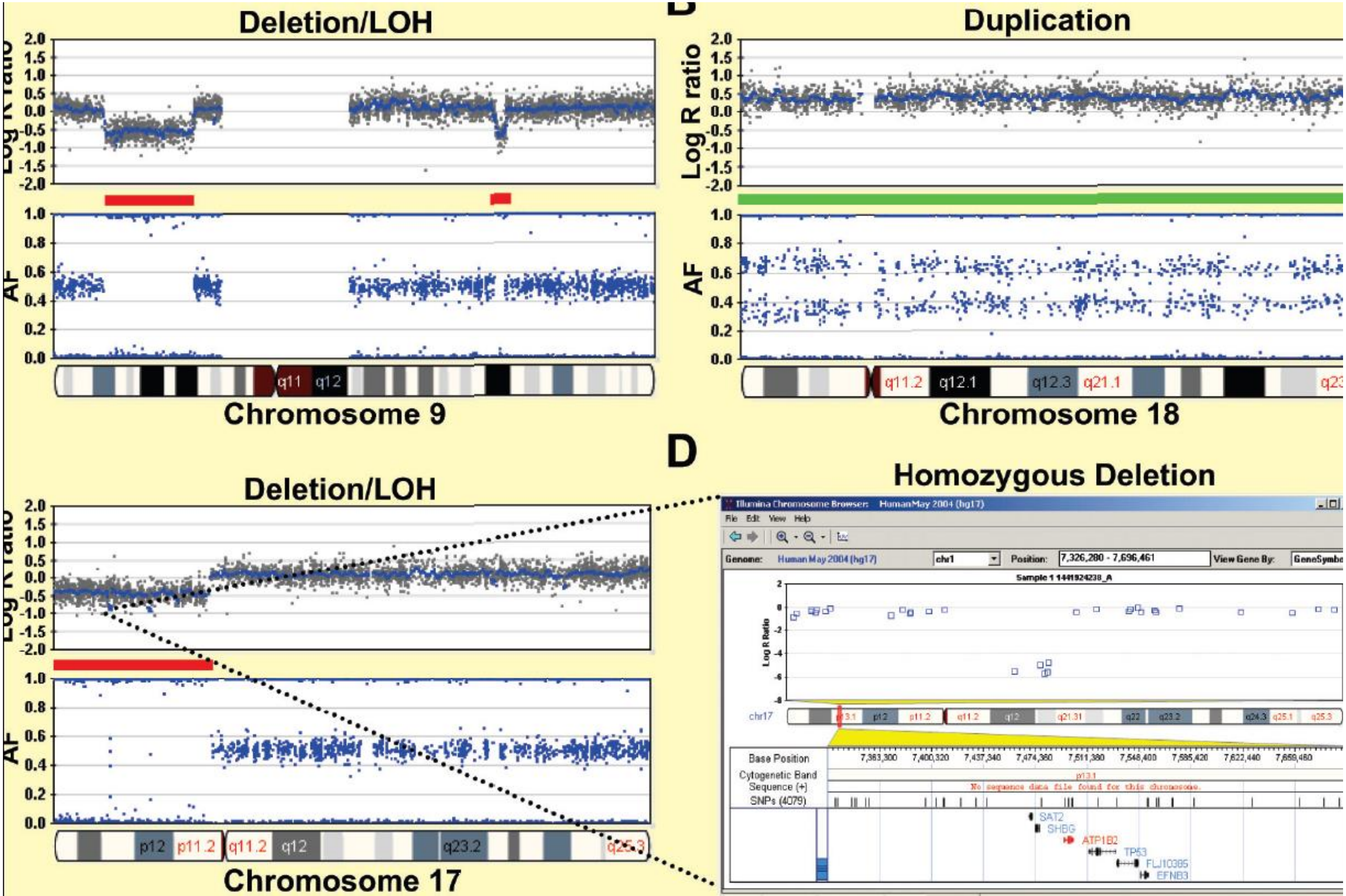
chrX, 120 normal individuals, one particular SNP
raw intensities: males in yellow, the others females
normalization is done using a ”proprietary” algorithm (B)
conversion to ”*polar coordinates*” (C)

B/A	θ
0	0
1	0.5
inf	1

”*canonical clusters*”

SNP-CGH technologies for genomic profiling.pdf

Both LRR and BAF used to interprete data



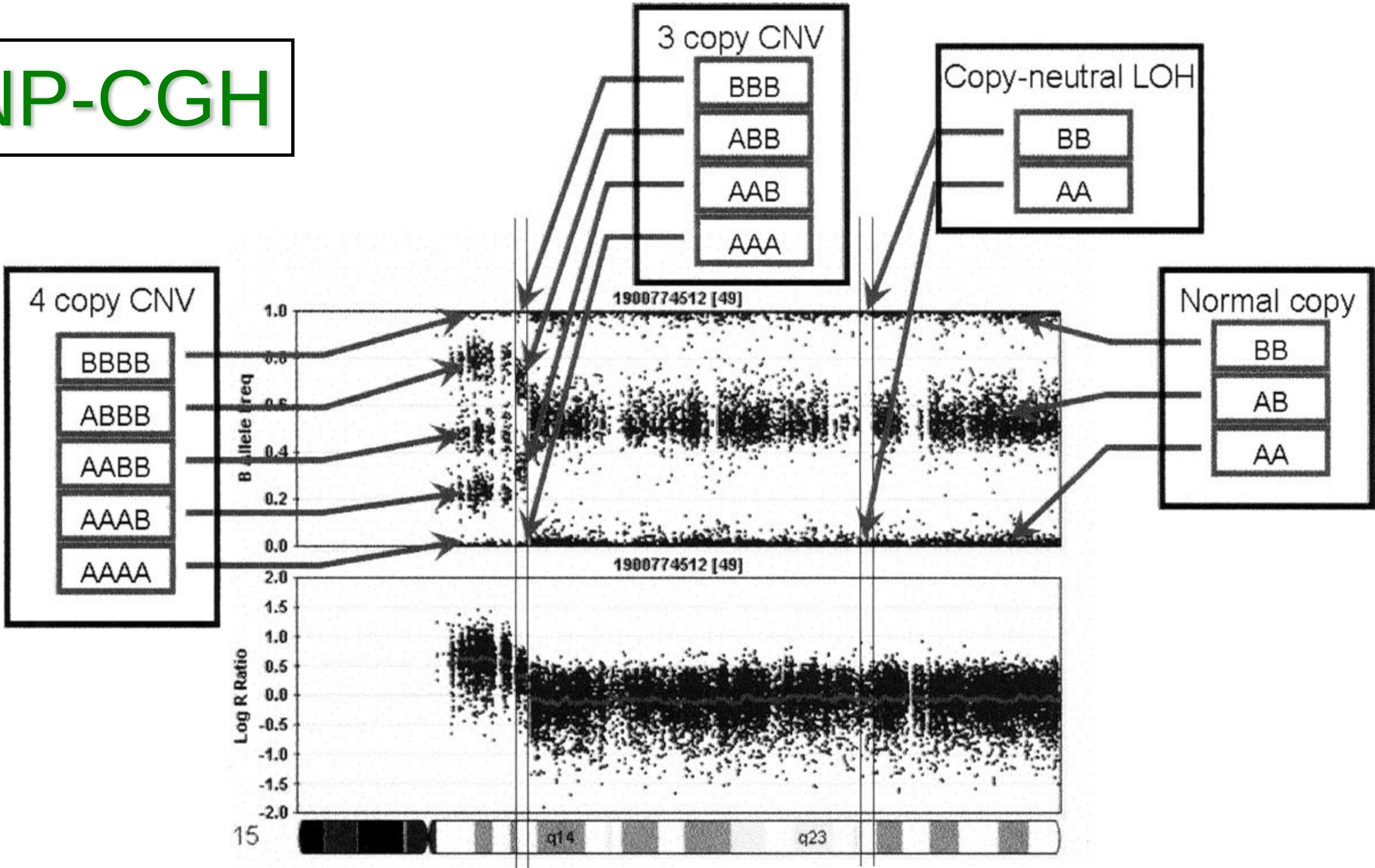
Genome: Human May 2004 (hg17) chr1 Position: 7,326,280 - 7,696,461 View Gene By: GeneSymbo

No sequence data file found for this chromosome.

Both LRR[1] and BAF[2] can be used to determine copy number

[1]LRR = Log R ratio
[2]BAF = B-allele frequency

SNP-CGH



4 copy CNV
- BBBB
- ABBB
- AABB
- AAAB
- AAAA

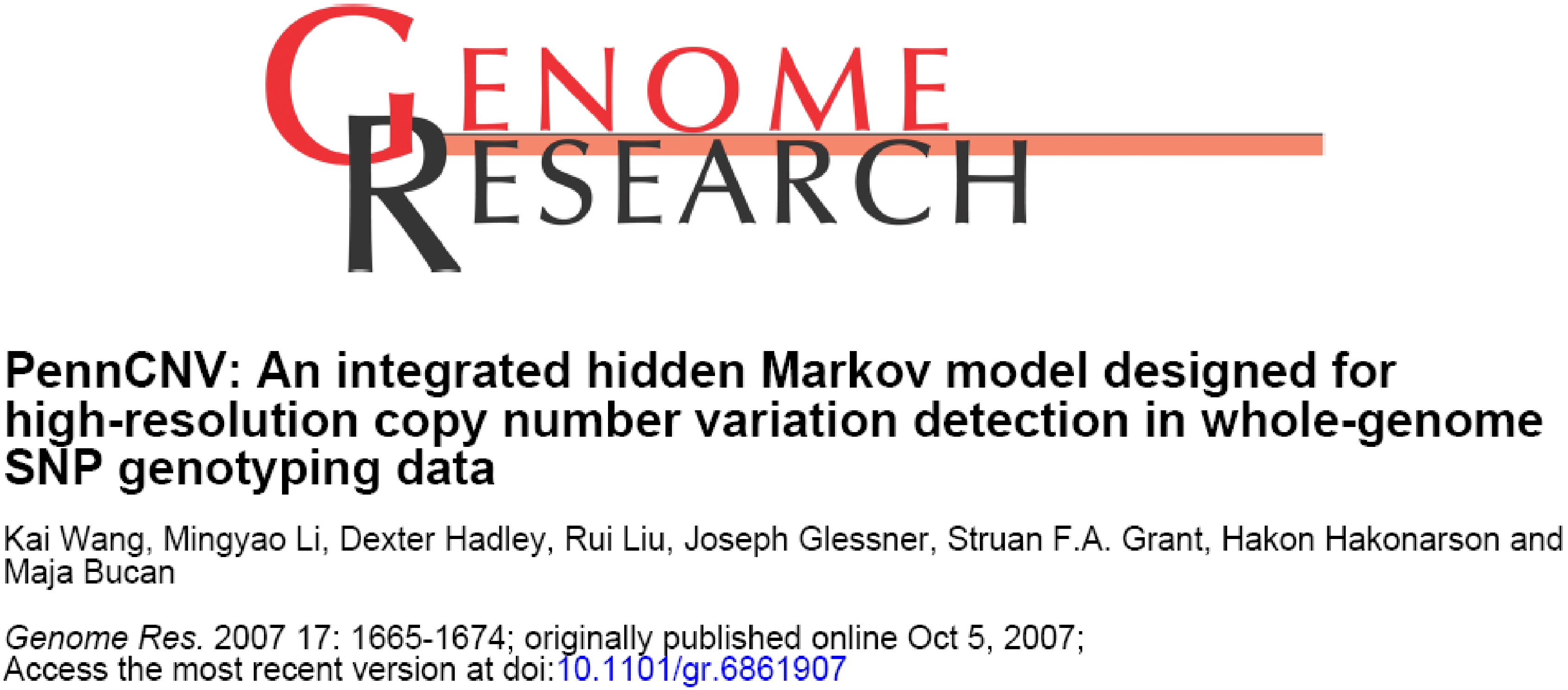
3 copy CNV
- BBB
- ABB
- AAB
- AAA

Copy-neutral LOH
- BB
- AA

Normal copy
- BB
- AB
- AA

Wang K. et.al. Genome Res. 2007;17:1665-1674

PennCNV Paper



PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data

Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F.A. Grant, Hakon Hakonarson and Maja Bucan

Genome Res. 2007 17: 1665-1674; originally published online Oct 5, 2007;
Access the most recent version at doi:10.1101/gr.6861907

PennCNV

- Detection of CNVs from Illumina (*Infinium*) high-density SNP genotyping data using:
  - total signal intensity
  - allelic intensity ratio at each SNP marker (BAF)
  - pedigree information if available
- kilobase-resolution (~10 Kb)

PennCNV – states of the HMM

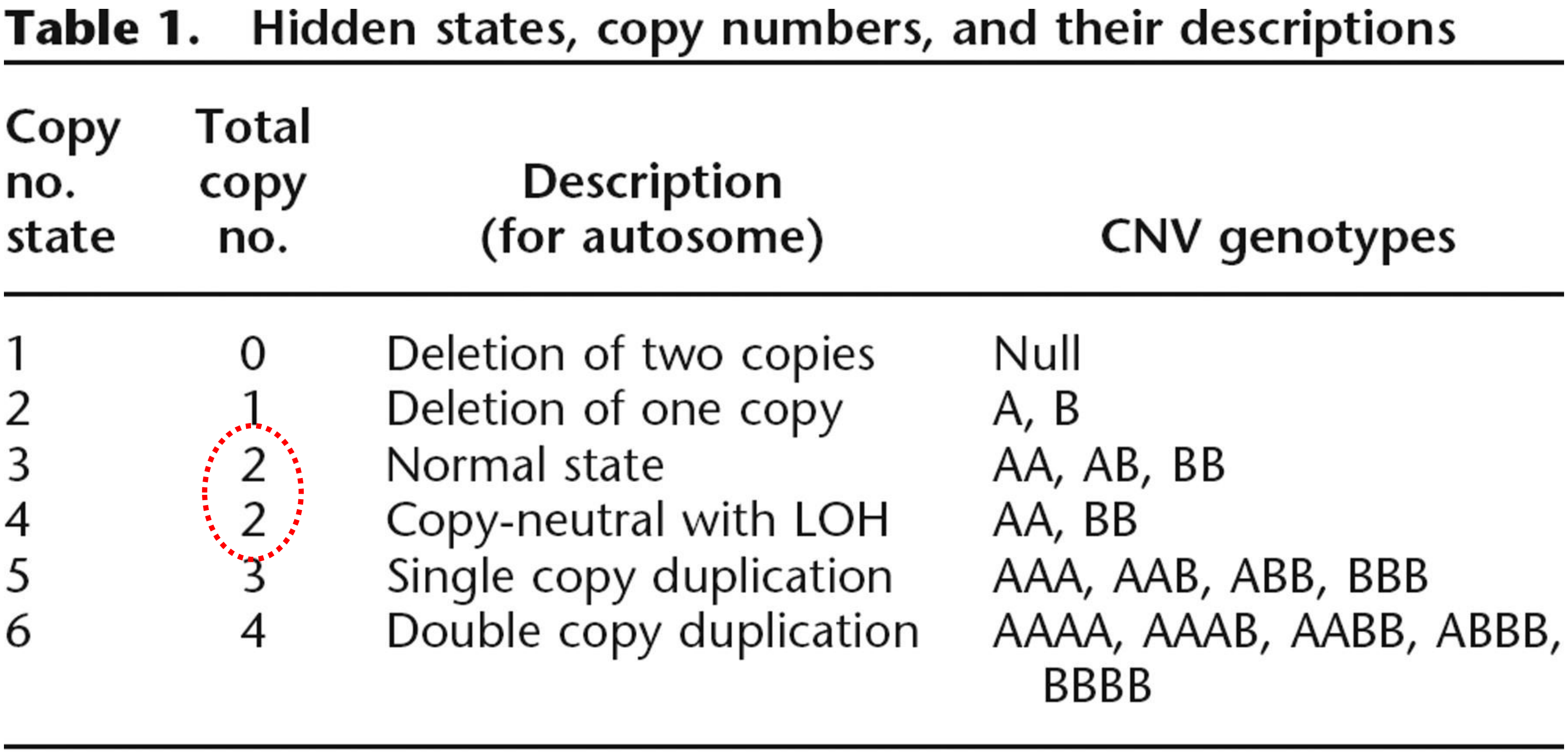


Table 1. Hidden states, copy numbers, and their descriptions

Copy no. state	Total copy no.	Description (for autosome)	CNV genotypes
1	0	Deletion of two copies	Null
2	1	Deletion of one copy	A, B
3	2	Normal state	AA, AB, BB
4	2	Copy-neutral with LOH	AA, BB
5	3	Single copy duplication	AAA, AAB, ABB, BBB
6	4	Double copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

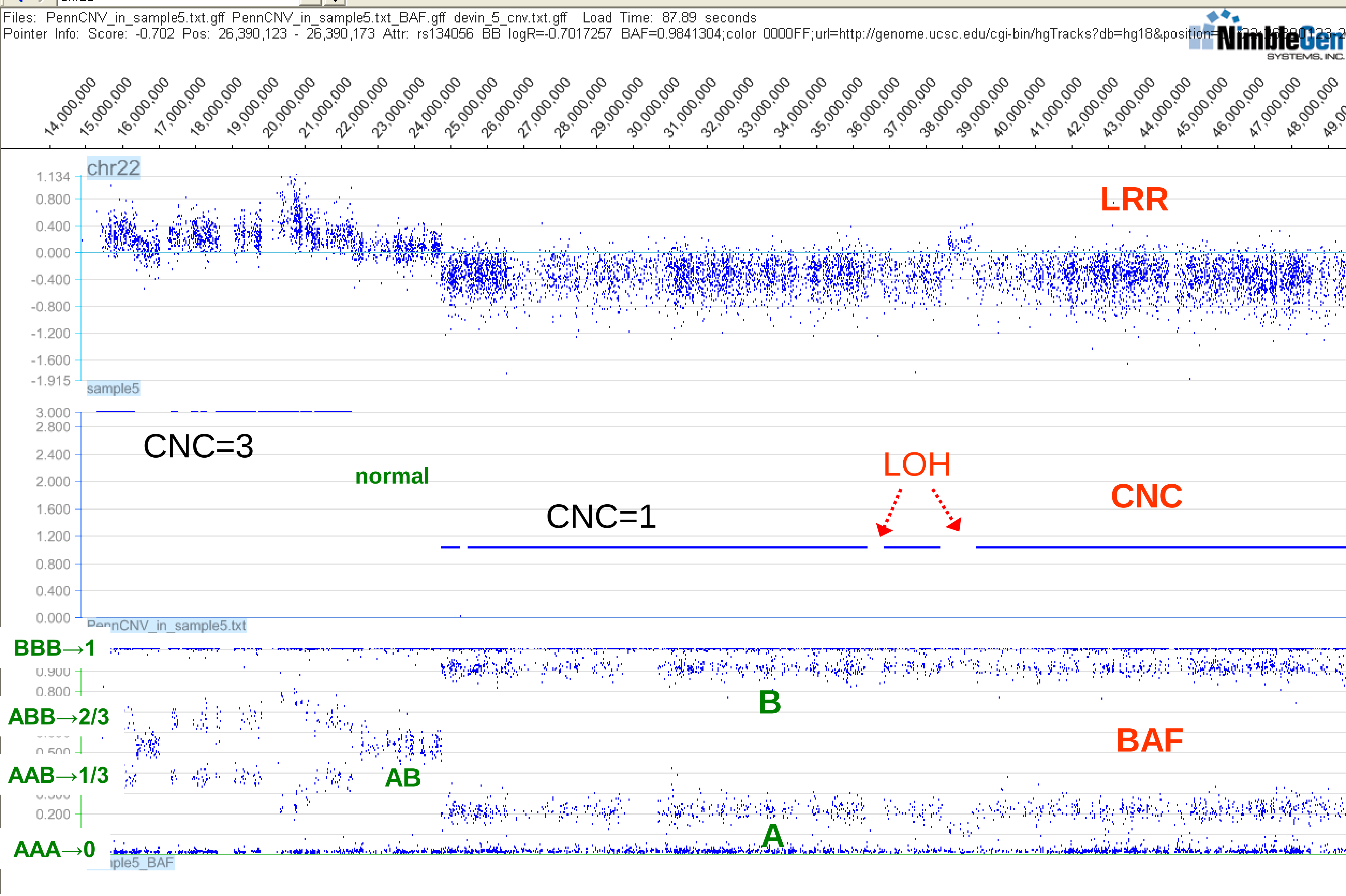
Allele frequency information included in the states of the HMM

PennCNV implementation

```
run_PennCNV   -f   Sample1.txt   -minsnp 10   -minlength 50k    -gcmodel
```

same inputfile as for run_CBS

Perl-script: /usr/local/share/BIOSW/run_PennCNV.pl
runtime: a few minutes
output: Sample1.txt.log   Sample1.txt.calls   Sample1_PennCNV.gff

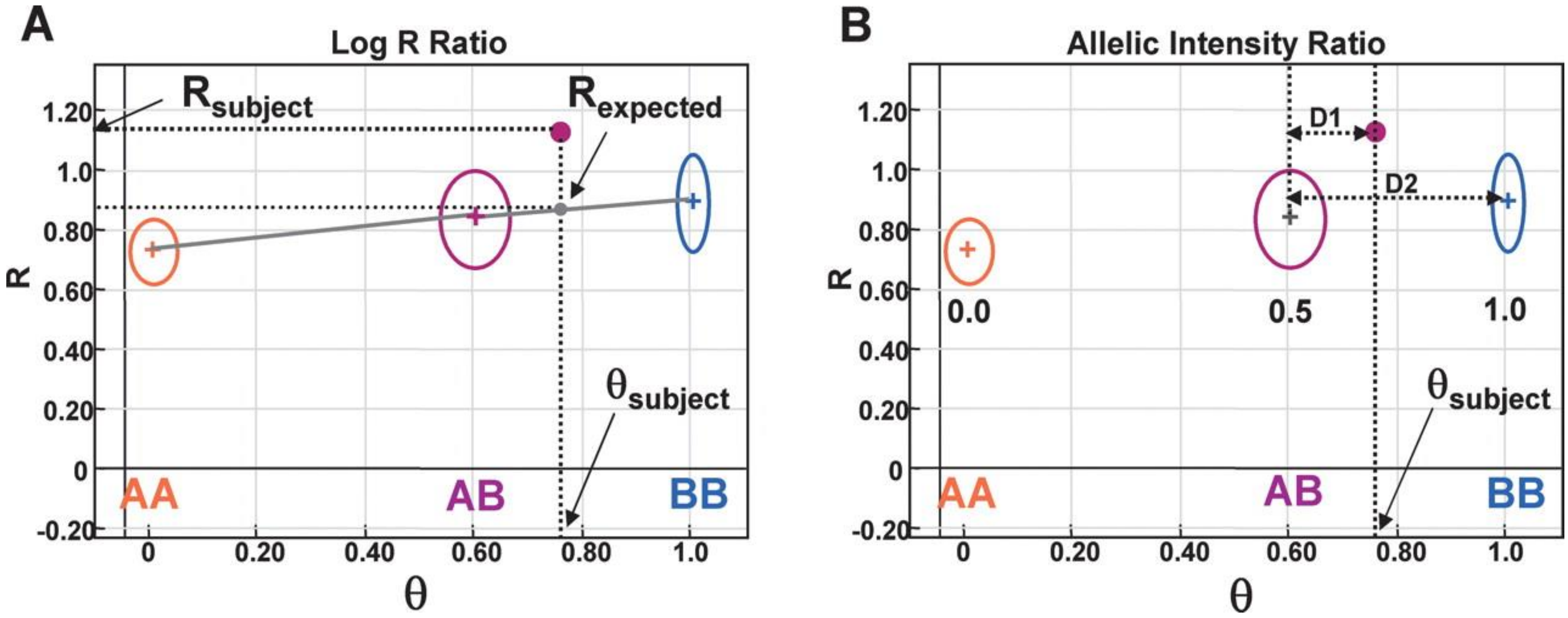


PennCNV quality assessment

- is done automatically
- identifies low-quality samples from a genotyping experiment
- several types of bad quality, see below ....

see "Illumina.ppt"

Canonical clusters



The canonical clusters are **not specific enough**

- clusters have to be defined for each machine
- or paired comparisons must be made

Peiffer D. A. et.al. Genome Res. 2006;16:1136-1148

PennCNV parameters

```
Optional arguments:
        -v, --verbose               use verbose output
        -h, --help                  print help message
        -m, --man                   print complete documentation

           --train                  train optimized HMM model (not recommended)
           --test                   test HMM model to identify CNV
           --trio                   posterior CNV calls for father-mother-offspring trio
           --quartet                posterior CNV calls for quartet
           --joint                  joint CNV calls for trio (available soon)
           --summary                                generate descriptive summary for signal quality

           --listfile <file>            a list file containing path to files to be processed
           --output <file>              specify output root filename
           --exclude_heterosomic        empirically exclude CNVs in heterosomic chromosomes
           --hmmfile <file>                         HMM model file
           --pfbfile <file>         population frequency for B allelel file
           --cnvfile <file>         specify CNV call file for use in family-based CNV calling
           --wavemodelfile <file>      a file containing regression model for wave adjustment
           --sample_index <int>        index of sample in input file (default=1) (obselete argument)
           --minsnp <int>                           minimum number of SNPs within CNV (default=3)
           --minlength <int>                        minimum length of bp within CNV
           --minconf <float>                        minimum confidence score of CNV (experimental feature)
           --loh                    display copy-neutral LOH information (obselete option)
           --chrx                   use chrX-specific treatment
           --chry                   use chrY-specific treatment (not implemented yet!)
           --fmprior <numbers>         prior belief on CN state for regions with CNV calls
           --denovo_rate <float>       prior belief on genome-wide de novo event rate
           --logfile <file>         write notification/warningn messages to this file
           --confidence             calculate confidence for each CNV (experimental feature)
           --tabout                 use tab-delimited output
           --coordinate_from_input      get marker coorindate information from signal file (rather than PFB file)
```

Function: generate CNV calls from high-density SNP genotyping data that
contains Log R Ratio and B Allele Frequency for each SNP

Other Programs: QuantiSNP

- similar to PennCNV
- several advantages of PennCNV:
  - state-specific and distance-dependent transition probabilities
  - better adapted to Illumina BAF calculation procedure
  - population frequency of the B allele considered
  - family information can be included (CNV-NDPs)

Other Programs: Birdsuite

The Birdsuite is a fully open-source set of tools to detect and report SNP genotypes, common Copy-Number Polymorphisms (CNPs), and novel, rare, or de novo CNVs in samples processed with the Affymetrix platform. While most of the components of the suite can be run individually (for instance, to only do SNP genotyping), the Birdsuite is especially intended for integrated analysis of SNPs and CNVs. Support for chips and platforms other than the Affymetrix SNP 6.0 is currently limited, but we are currently working on creating the supporting files for other common genotyping platforms.

Other Programs: SNPRank (Nexus)

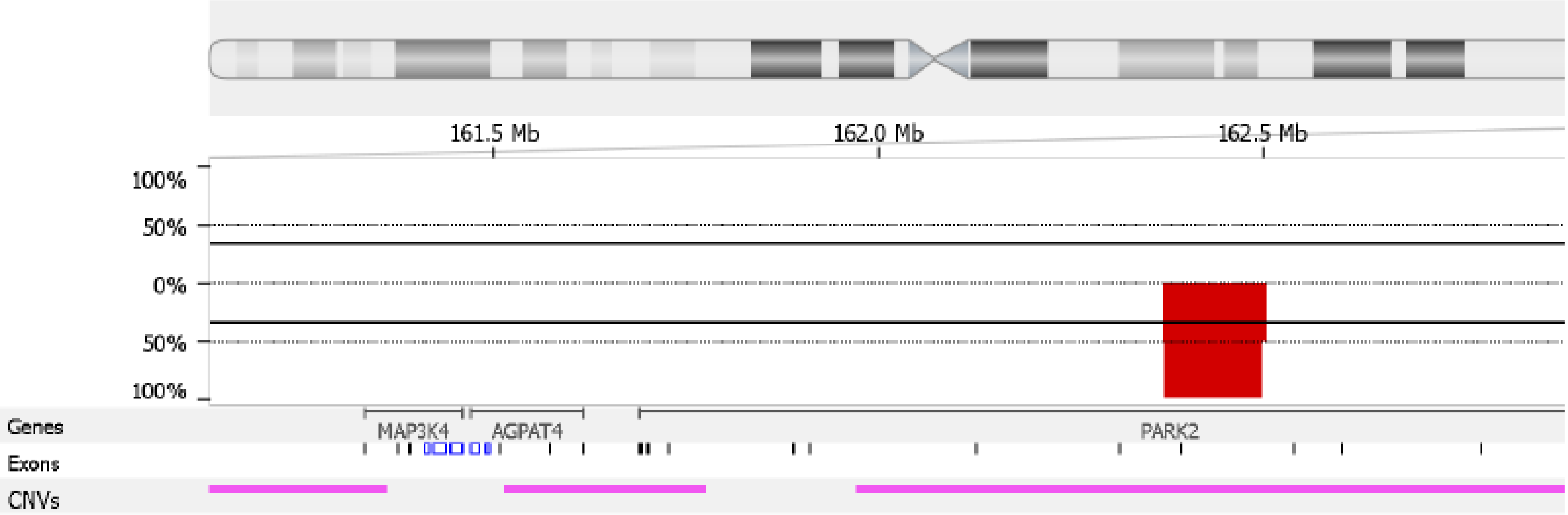
Dear Uwe,

The algorithm is new and we have developed it ourselves. It is called SNPRank. Are you working with Hanna Göransson at Uppsala?

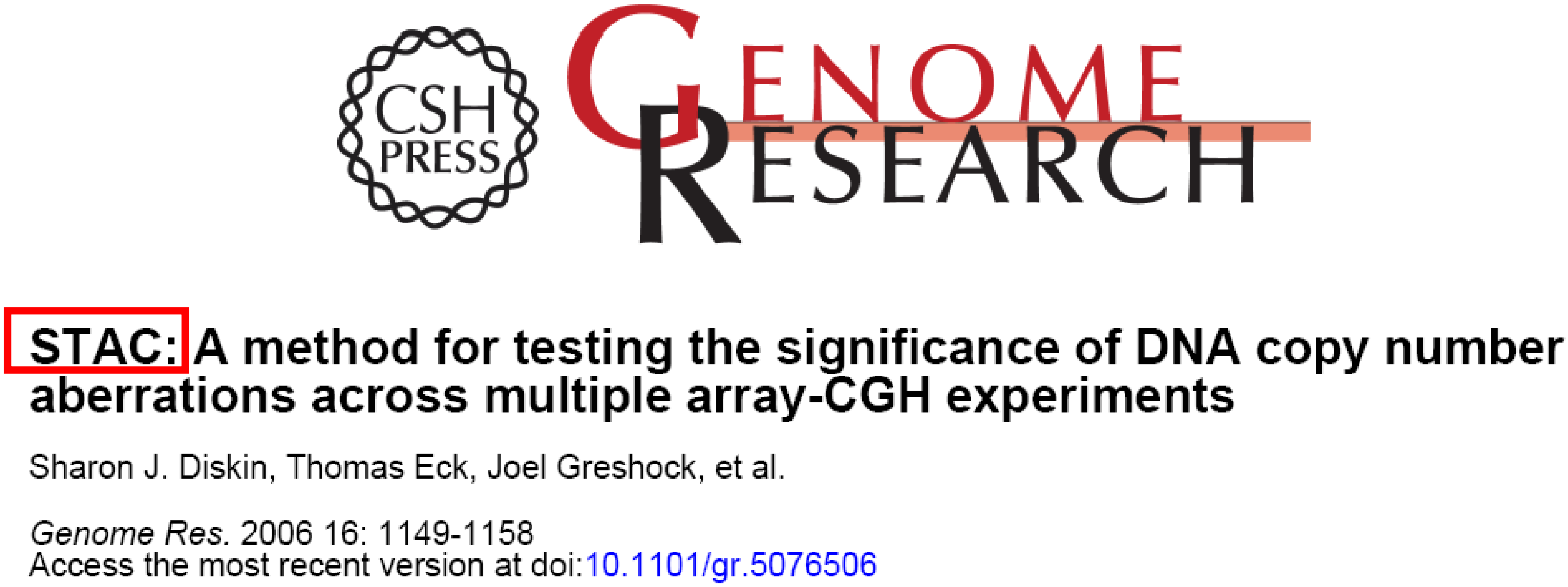
-Soheil

Comparison of samples

- Frequency plots (Nexus):



Comparison of samples



CSH PRESS

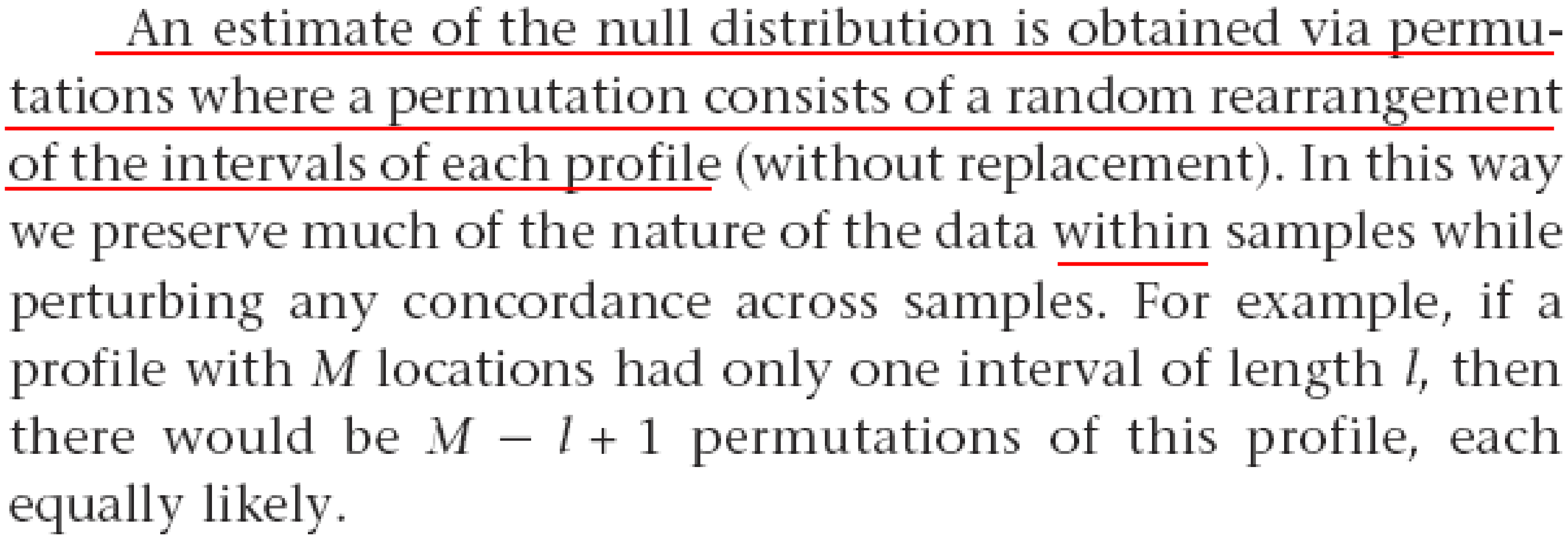
GENOME RESEARCH

STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments

Sharon J. Diskin, Thomas Eck, Joel Greshock, et al.

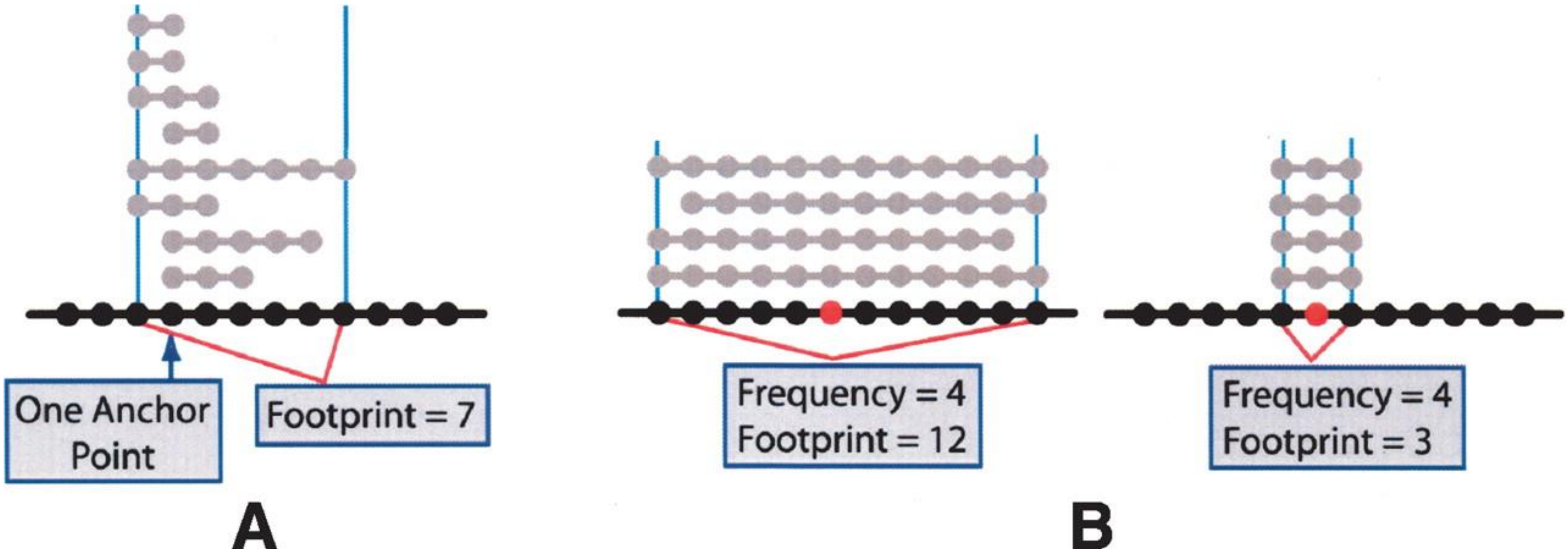
Genome Res. 2006 16: 1149-1158
Access the most recent version at doi:10.1101/gr.5076506

STAC - permutation



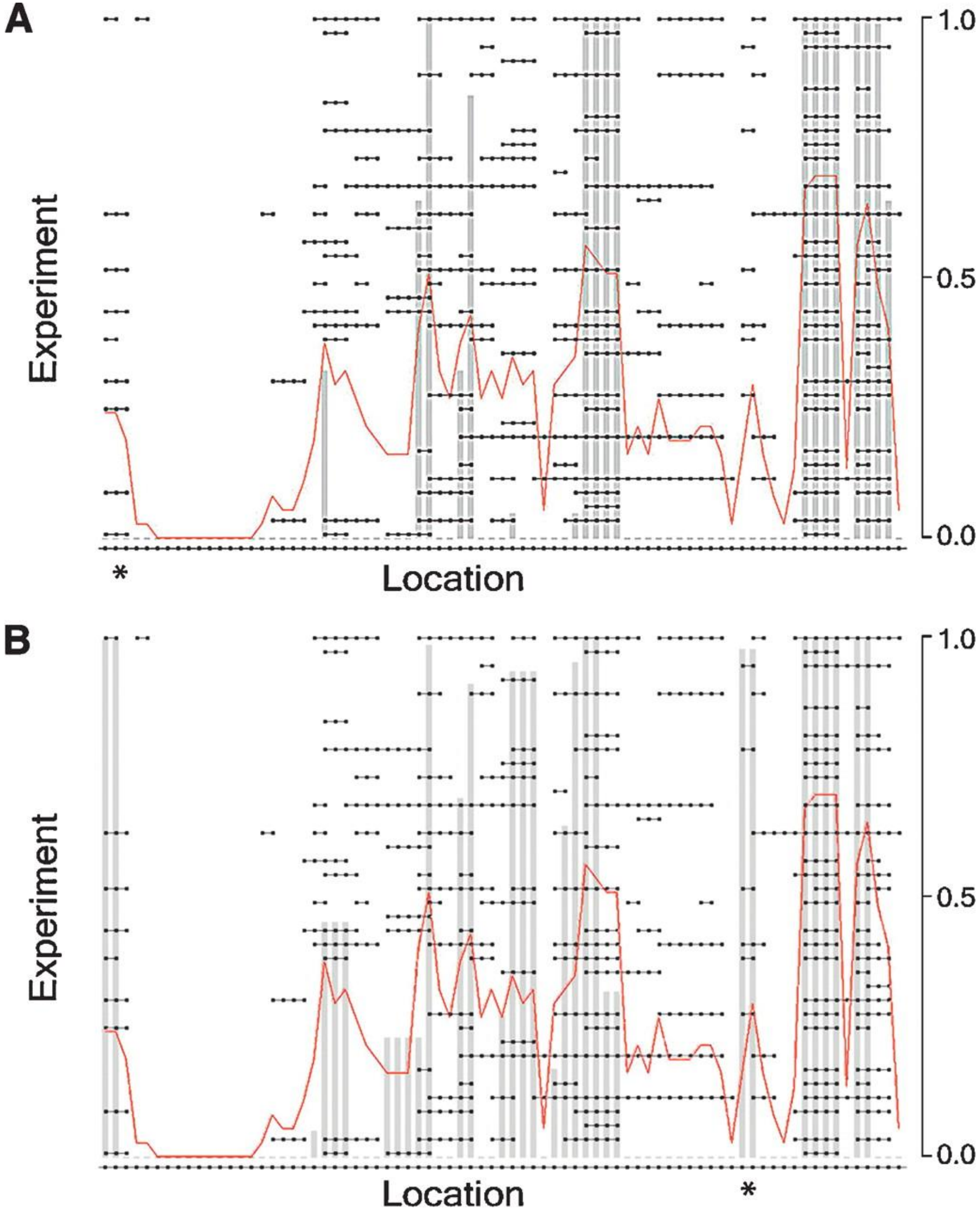
An estimate of the null distribution is obtained via permutations where a permutation consists of a random rearrangement of the intervals of each profile (without replacement). In this way we preserve much of the nature of the data within samples while perturbing any concordance across samples. For example, if a profile with M locations had only one interval of length l, then there would be $M - l + 1$ permutations of this profile, each equally likely.

STAC-results



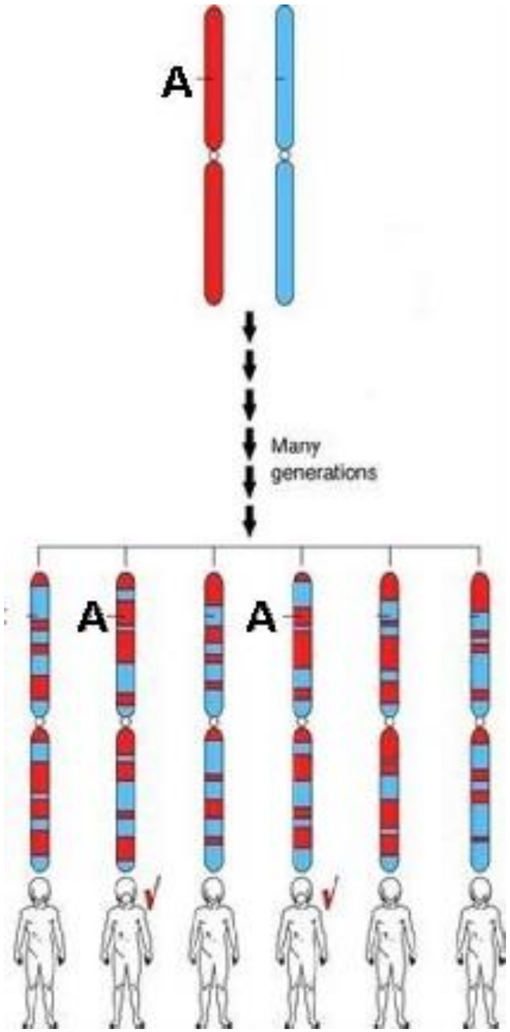
Thanks !

STAC-results



http://www.hapmap.org/whatishapmap.html.en

Haplotypes and tag SNPs



Over the course of many generations, segments of the ancestral chromosomes in an interbreeding population are shuffled through repeated recombination events. Some of the segments of the ancestral chromosomes occur as regions of DNA sequences that are shared by multiple individuals (Figure 1). These segments are regions of chromosomes that have not been broken up by recombination, and they are separated by places where recombination has occurred. These segments are the haplotypes that enable geneticists to search for genes involved in diseases and other medically important traits.

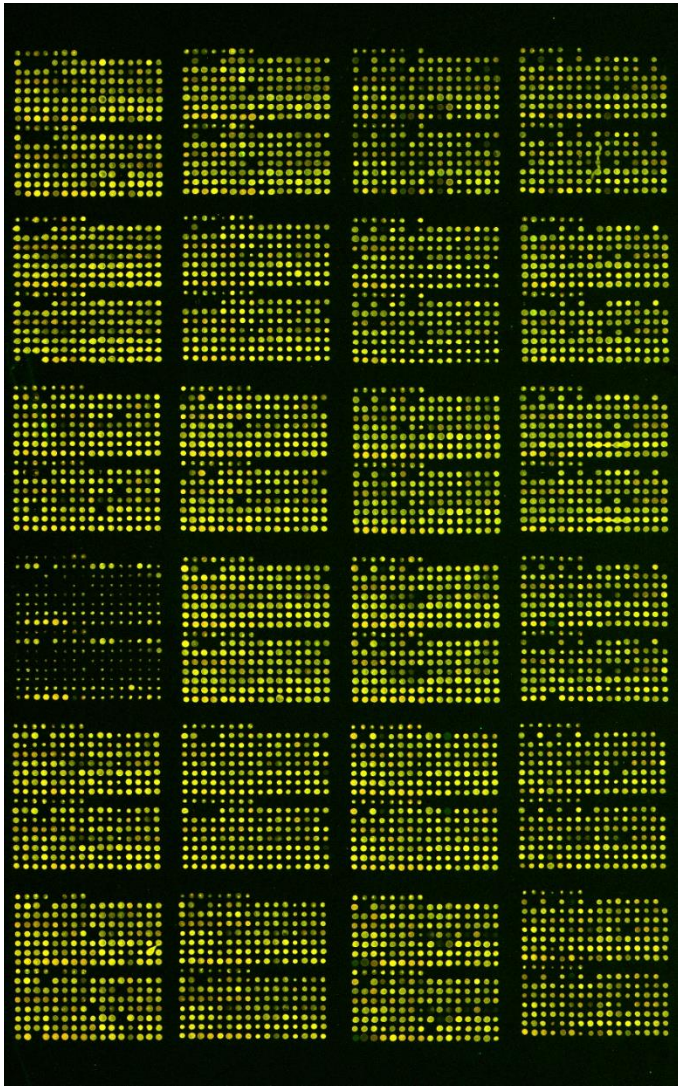
A given haplotype can occur at different frequencies in different populations.

Haplotypes and tag SNPs

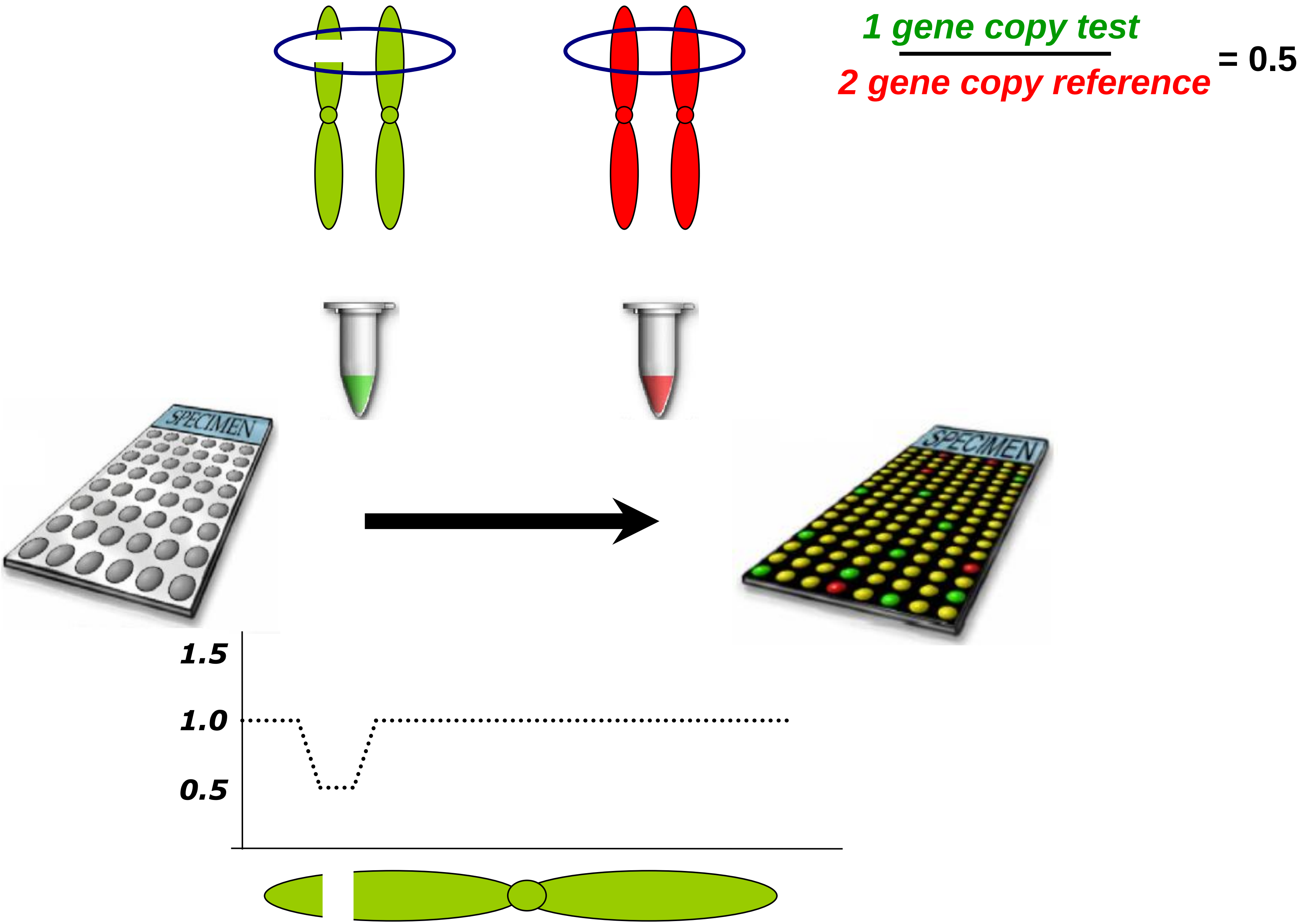
- In many parts of our chromosomes, just a handful of haplotypes are found in humans.
- In a given population, 55 % of people may have one version of a haplotype, 30 % may have another, 8 % may have a third, and the rest may have a variety of less common haplotypes.
- The HapMap Project is identifying these common haplotypes in four populations from different parts of the world.
- It also is identifying **"tag" SNPs** that uniquely identify these haplotypes:
  - testing an individual's tag SNPs (‟ genotyping”) → identification of the collection of haplotypes in that person's DNA
  - The number of tag SNPs that contain most of the information about the patterns of genetic variation is estimated to be about 300,000 to 600,000, which is far fewer than the 10 million common SNPs

http://www.hapmap.org/index.html

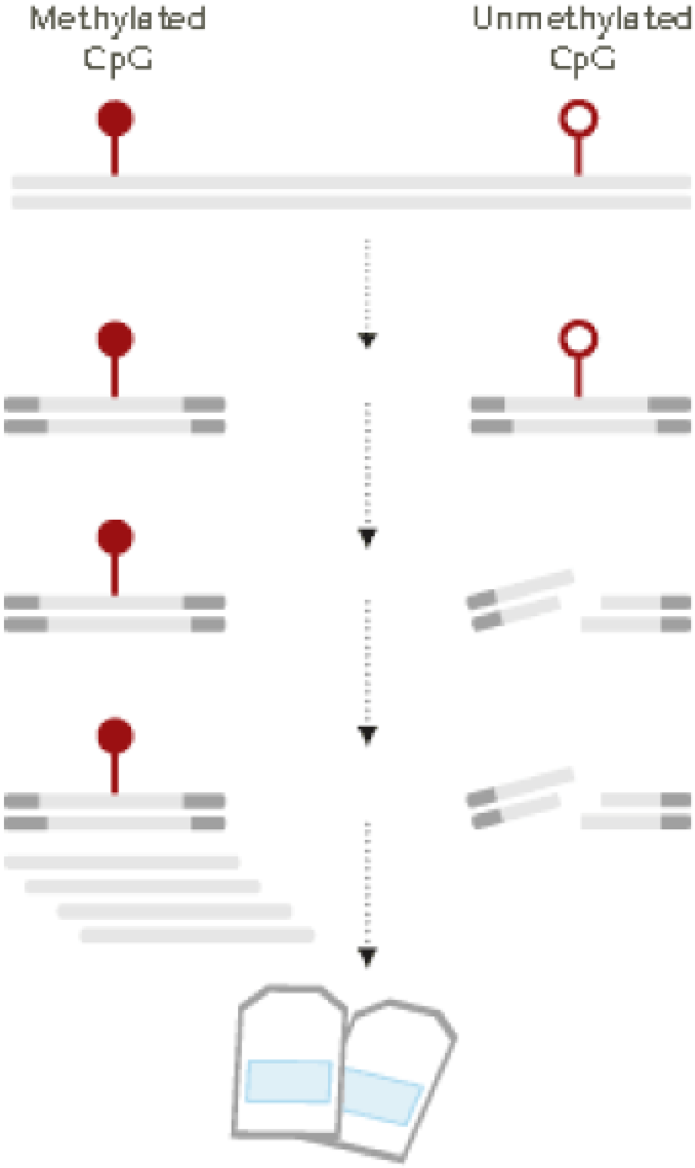
Full-coverage human chromosome 1 array, with ~2 200 data points (from Sanger Centre, UK) – application to analysis of meningioma



Test DNA Reference DNA



$$\frac{\textit{1 gene copy test}}{\textit{2 gene copy reference}} = 0.5$$



Description

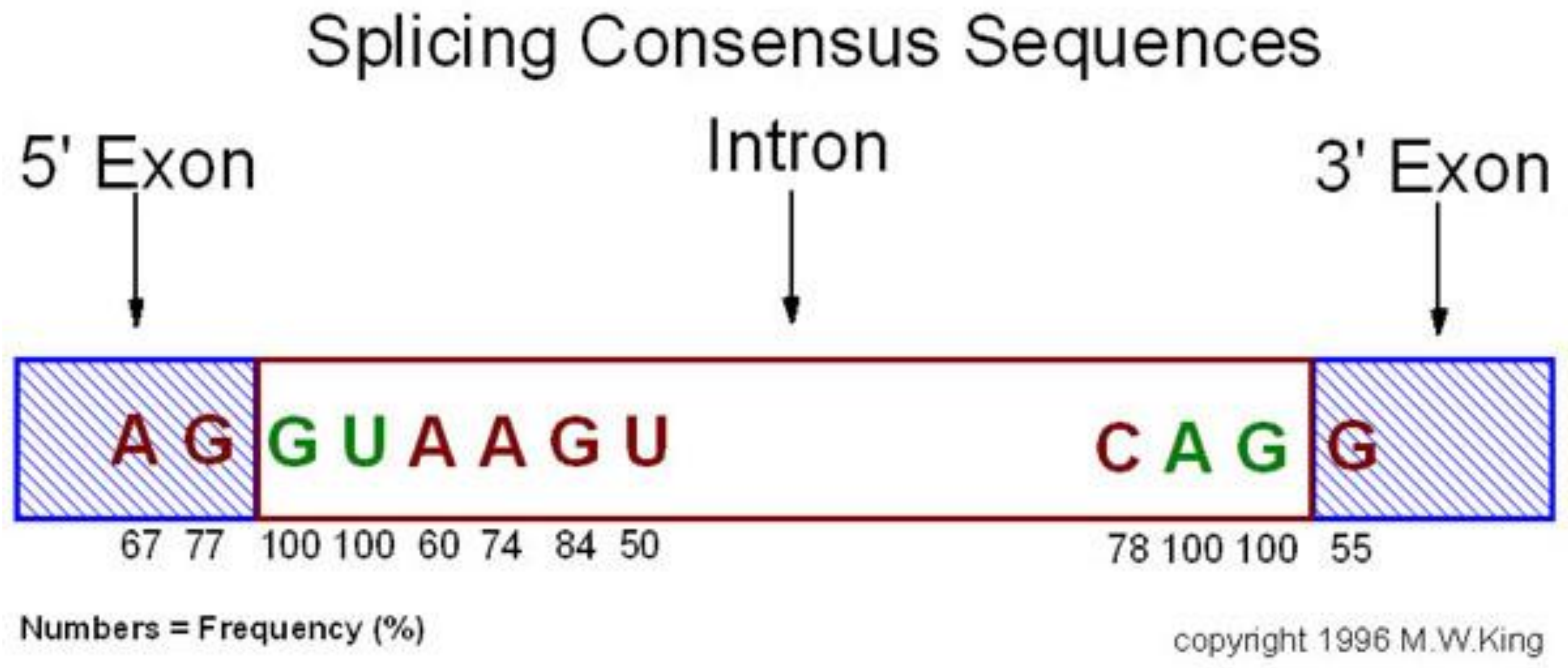
1. Genomic DNA is isolated from fresh-frozen human samples.

2. DNA is cut with methylation-insensitive restriction enzymes followed by ligation of linkers.

3. Resulting fragments are cut with methylation- sensitive restriction enzymes.

4. Un-cut (i.e. methylated) fragments are PCR amplified using linker-specific primers.

5. Amplified fragments are labeled and hybridized to Epigenomics' proprietary microarray covering 50,000 CpG-rich human genomic regions (designed by Epigenomics).



Splicing Consensus Sequences

Numbers = Frequency (%)

copyright 1996 M.W.King

GT .... AG