Detecting Copy Number Variations

(using Illumina arrays)

Uwe Menzel
Uppsala University

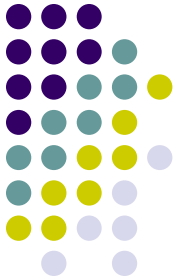
human610-quad beadchip

Genotyping & CNV analysis

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

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

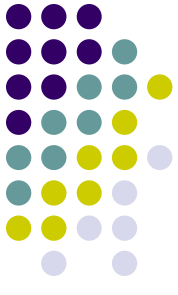
human610-quad beadchip



- 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*

Illumina GenomeStudio
(replaces BeadStudio)



Software: GenomeStudio - Genotyping Module

The GenomeStudio Genotyping (GT) Module supports the analysis of Infinium and GoldenGate Genotyping Assay data collected by the iScan System and BeadXpress Reader. This module enables efficient genotyping data normalization, genotype calling, clustering, data intensity analysis, loss of heterozygosity (LOH) calculation, and copy number variation (CNV) analysis. Fully integrated with Infinium LIMS server, the GT Module allows you to directly access data and manage projects from within GenomeStudio.

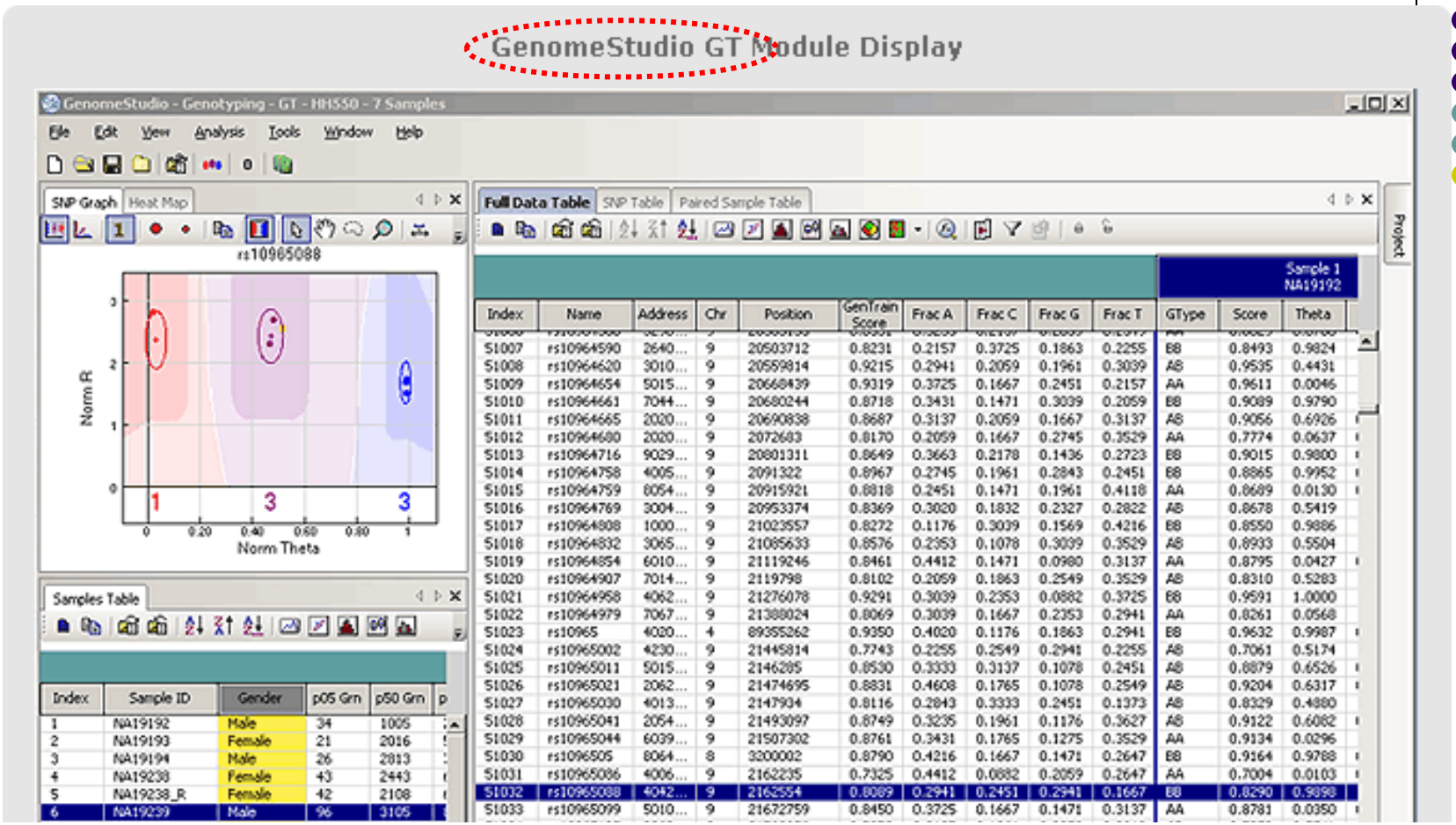
As in all GenomeStudio modules, the GenomeStudio Framework displays data output in tabular form and enables you to quickly and easily visualize your results using the Illumina Genome Viewer and Illumina Chromosome Browser graphical tools.

http://www.illumina.com/pages.ilmn?ID=309

GenomeStudio – Genotyping Module

Structural variation is identified using the same markers as genotyping and intensity-only probes with algorithms to calculate loss of heterozygosity (LOH), and abnormal copy numbers (CNVs). Identified structural variants can be bookmarked

see also: GenomeStudioTM Data Analysis Software.pdf page 3



"Export genotype data to various third party applications; access multiple CNV algorithms and copy number variation analysis tools"

From GenomeStudio:

	A	B	C	D	E	F
1	**Name**	**Chr**	**Position**	**X.GType**	**X.Log R Ratio**	**X.B Allele Freq**
2	rs12354060	1	10004	BB	0.05767579	1
3	rs2691310	1	46844	NC	-0.1525835	0.5361379
4	rs2531266	1	59415	NC	0.2049716	0.3973282
5	rs4124251	1	97215	NC	-0.09452707	0.09473621
6	rs8179466	1	224176	NC	-0.02189994	1
7	rs6603779	1	227744	NC	0.0836625	0.6220879
8	cnvi0007379	1	311662	NC	-0.1188801	1
9	cnvi0019140	1	314893	NC	-0.1406044	0
10	cnvi0007389	1	318309	NC	-0.1599025	0
11						

Tomas Axelsson, ETJ1 data

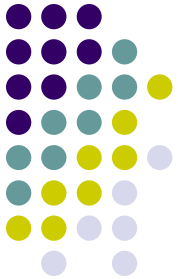
CBS script on anaconda:

Usage: run_CBS OPTIONS

-f S1.txt : Illumina SNP-CGH file for one sample
-lcol 5 : Number of the column containing the log2 intensity ratios
-minsnp 10 : Minimum number of markers in an aberration
-minlength 50k : Minimum length of an aberration

EXAMPLE: run_CBS -f Sample1.txt -lcol 5 -minsnp 10 -minlength 50k

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

Biostatistics (2004), **5**, 4, *pp.* 557–572
doi: 10.1093/biostatistics/kxh008

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

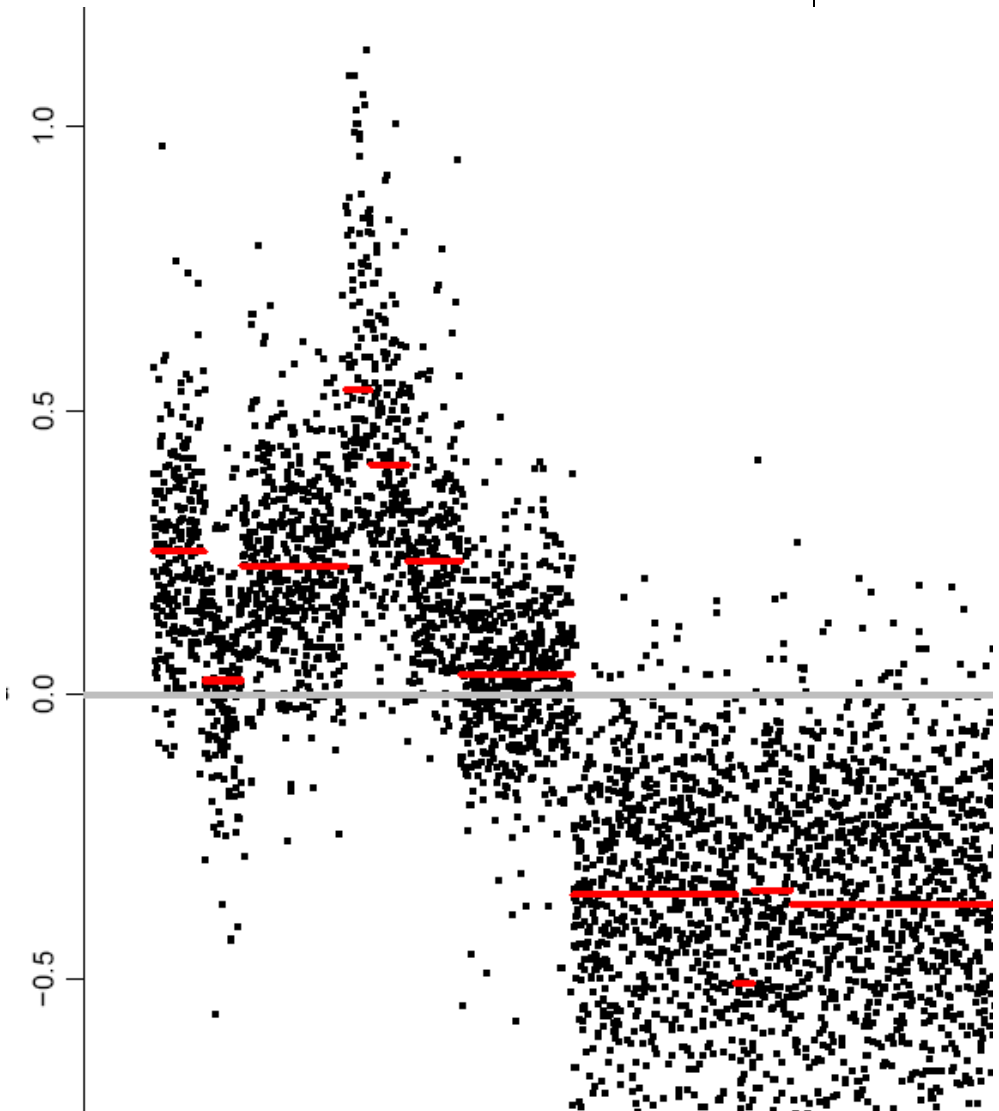
ROBERT LUCITO, MICHAEL WIGLER
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

SUMMARY

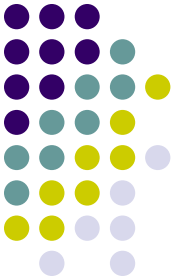
DNA sequence copy number is the number of copies of DNA at a region of a genome. Cancer progression often involves alterations in DNA copy number. Newly developed microarray technologies enable simultaneous measurement of copy number at thousands of sites in a genome. We have developed a modification 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.

Change-point method

- looks for changes in the distribution of data
- uses log-ratios of normalized intensities



Binary Segmentation
(Sen & Srivastava, 1975)



Likelihood ratio statistics

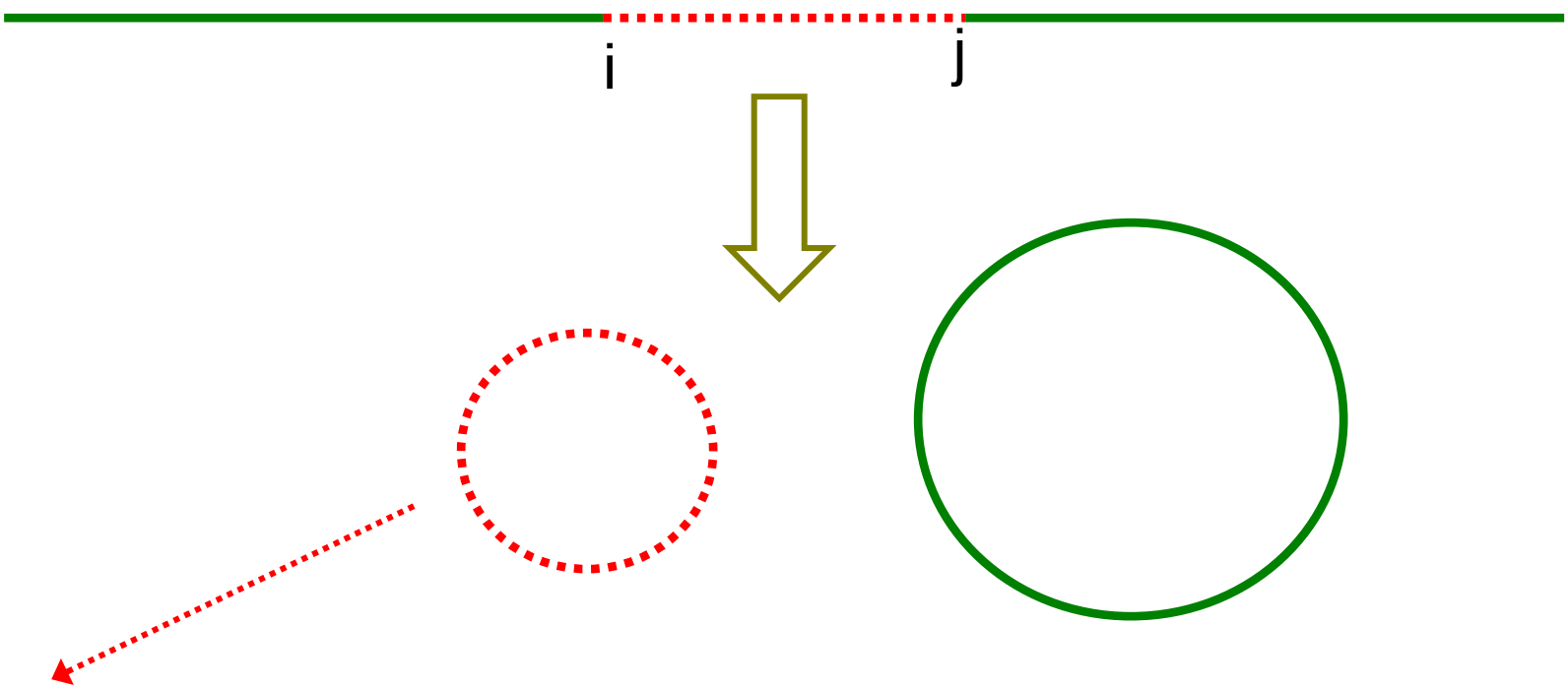
H_0: no change of the (normal-) distribution

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

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

H_0 is rejected if Z_i gets too large → change point at i (break point)

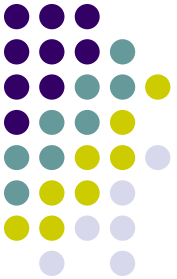
Circular Binary Segmentation



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

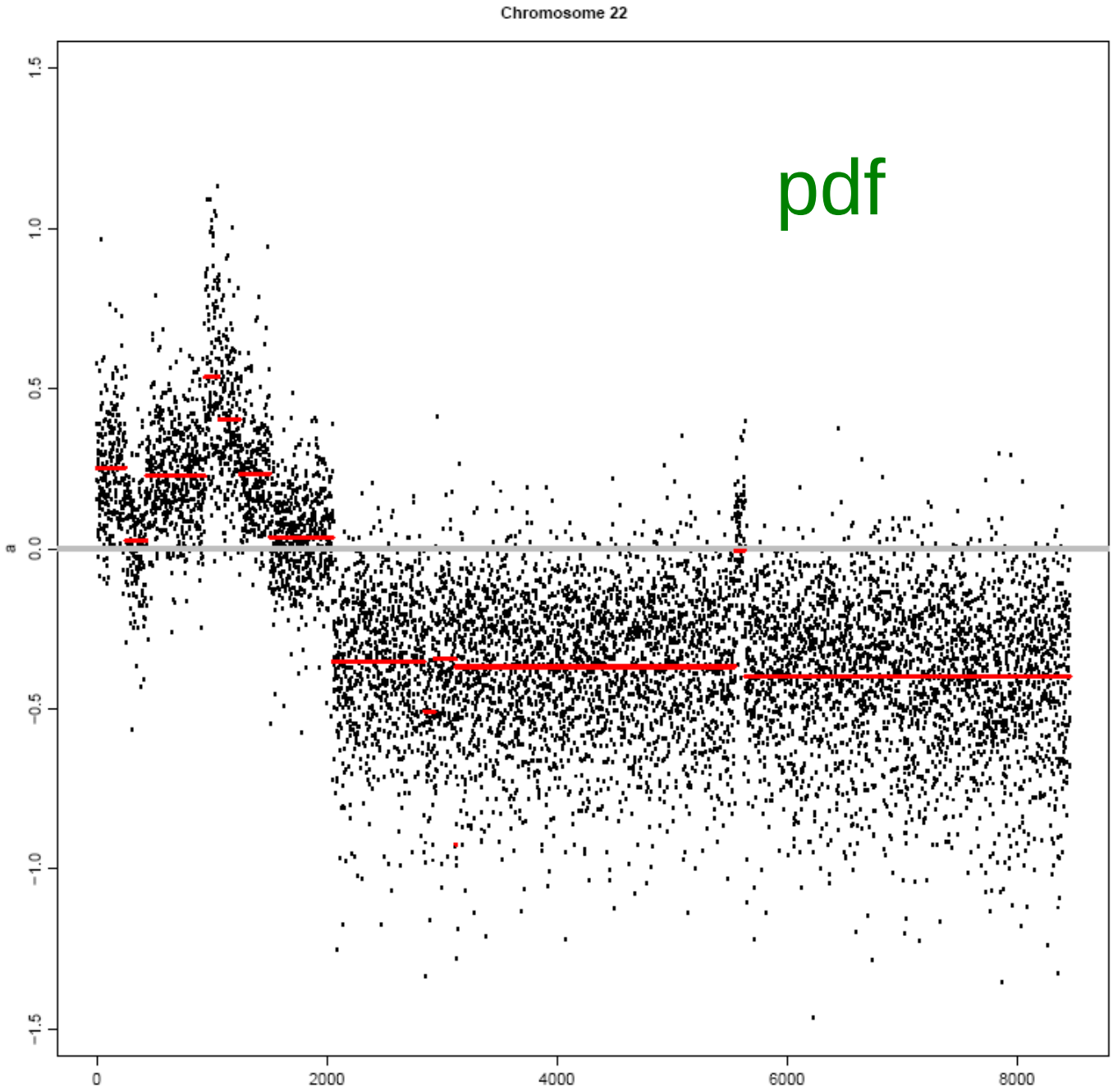
Tests if the arc from i to j has a mean which is different from the mean of the **complement.**

Practical example

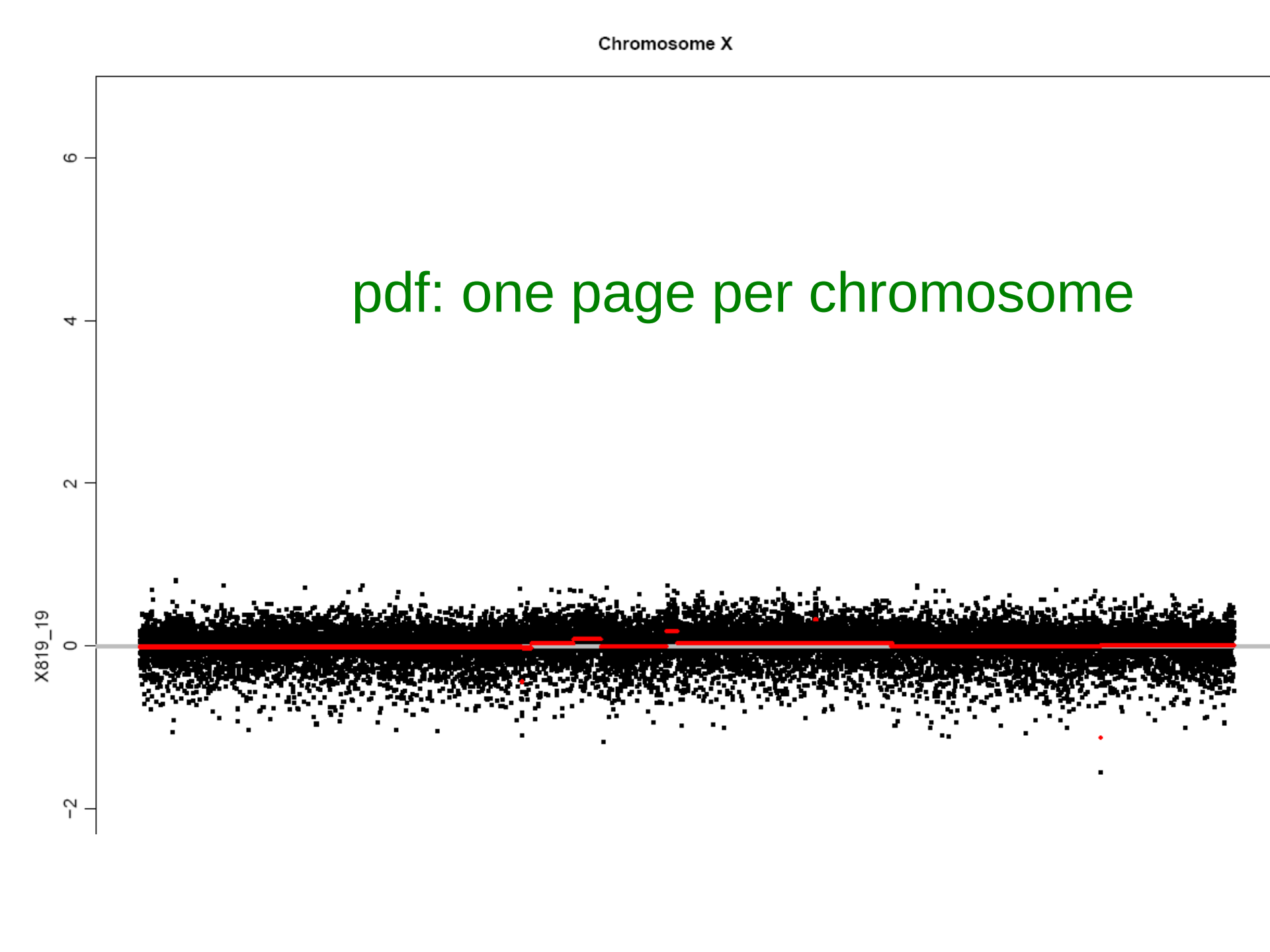


- anaconda:/nfs/1d/menzel/RUN_CBS >
  - run_CBS -f *Sample1.txt* -lcol *5* -minsnp *10* -minlength *50k*
  - 620312 rows in Sample1.txt
  - ≤ 45 minutes
- Output:
  - Sample1_CBS.pdf
  - Sample1_CBS.txt
  - Sample1_CBS.gff

} next 3 slides

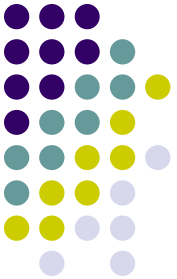


from Devins data:
chr22, Tumor 5
run_DNAcopy.R
21/10/08
anaconda


Chromosome X
pdf: one page per chromosome
X819_19
6
4
2
0
−2

Textual output of DNAcopy

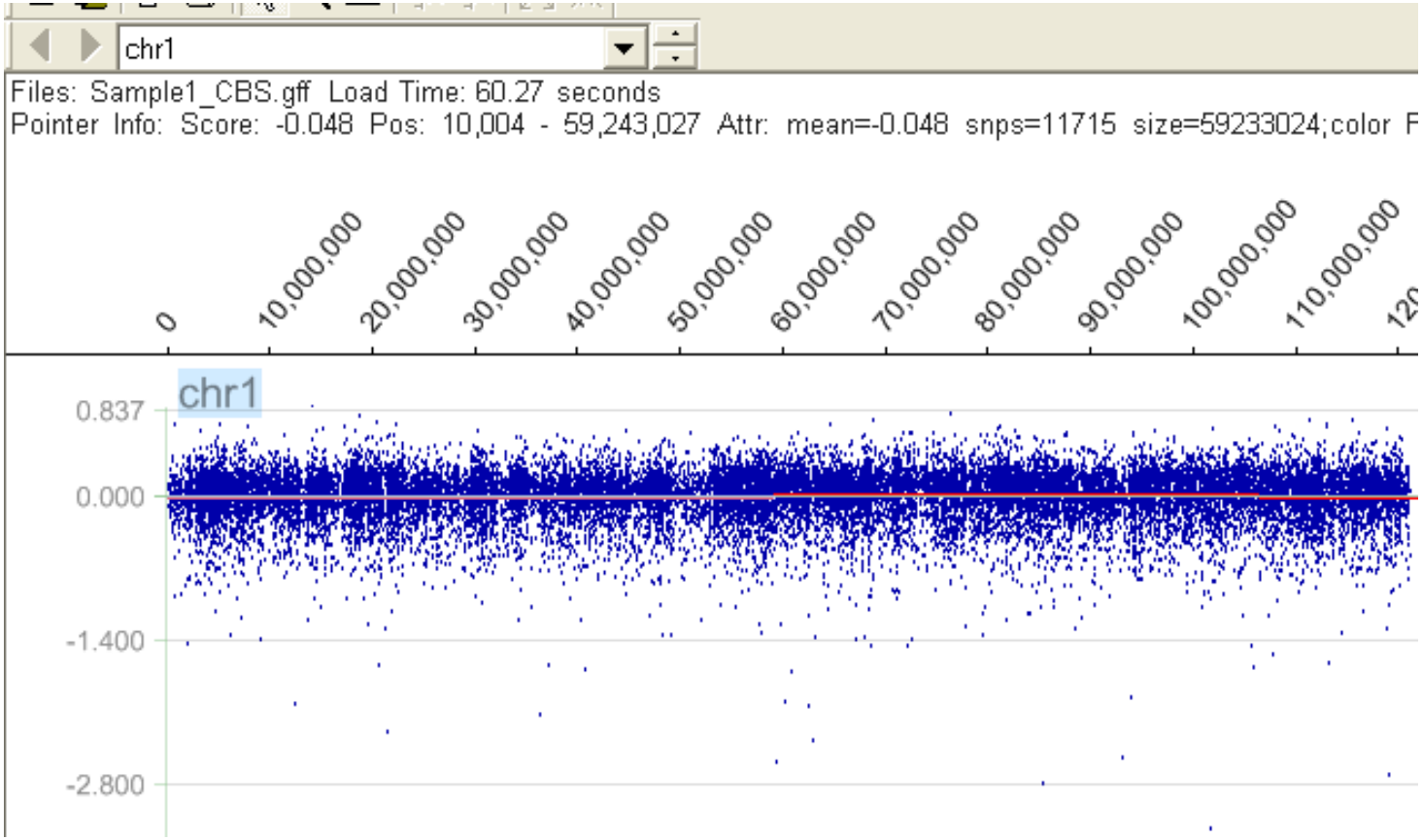
txt



ID	chrom	loc.start	loc.end	num.mark	seg.mean
X819_19	1	10004	59243027	11715	-0.048
X819_19	1	59244985	106543804	9983	-0.0292
X819_19	1	106552132	167468191	8082	-0.0514
X819_19	1	167495768	167505182	3	-5.8199
X819_19	1	167524930	247185943	17298	-0.0439
X819_19	10	59083	135372744	32059	-0.0385
X819_19	11	123495	36481331	9235	-0.0429
X819_19	11	36481704	42432739	1057	0.0072
X819_19	11	42445071	55124465	1596	-0.0446
X819_19	11	55127597	55165276	21	-0.5282
X819_19	11	55171592	89354301	6805	-0.0473
X819_19	11	89362241	111723307	4856	-0.2925
X819_19	11	111725592	125065390	3311	-0.2358
X819_19	11	125067925	134445626	2801	-0.0465
X819_19	12	21054	44185663	10725	-0.0424
X819_19	12	44194223	44211231	3	-0.9389
X819_19	12	44217885	132288869	18582	-0.0417
X819_19	13	17922259	56654503	9313	-0.0374
X819_19	13	56659471	56680301	6	-0.8908
X819_19	13	56690719	114123122	13367	-0.0235
X819_19	14	18070422	40502100	5283	-0.0314
X819_19	14	40503056	48154747	1449	0.0102

The gff-file can be loaded to SignalMap (NimbleGen)

gff



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

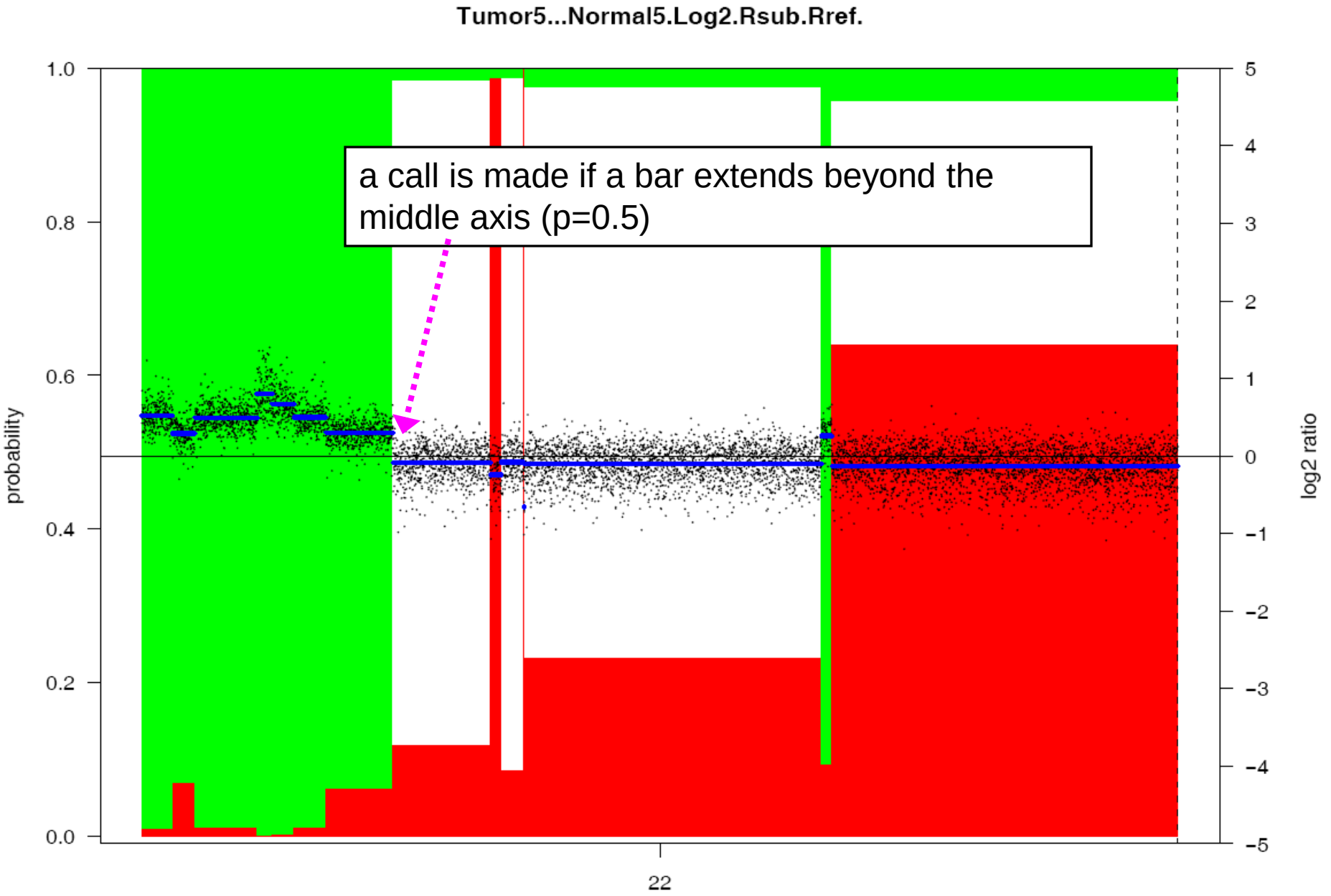
Received on December 15, 2006; revised on January 23, 2007; accepted on January 23, 2007

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



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

Textual output of CGHcall

Name	Chr	Position	Tumor1	T2	T3	T5	T6
rs4911642	22	14884399	0	0	0	1	0
rs2027653	22	15298335	0	0	0	1	0
rs5747620	22	15412698	0	0	0	1	0
rs9605903	22	15434720	0	0	0	1	0
rs5747968	22	15447504	0	0	0	1	0
rs2236639	22	15452483	0	0	0	1	0
rs5747999	22	15455353	0	0	0	1	0
rs11089263	22	15467656	0	0	0	1	0
rs2096537	22	15474749	0	0	0	1	0
rs9604959	22	15479107	0	0	0	1	0
rs9604967	22	15492342	0	0	0	1	0
rs4819849	22	15532611	0	0	0	1	0
rs9605028	22	15534984	0	0	0	1	0
rs1892844	22	15535383	0	0	0	1	0

anaconda:/nfs/1d/menzel/TEST_DNAcopy/CGHcall_results.txt

0,1,-1

Comparison Studies

BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 22 2005, pages 4084–4091
doi:10.1093/bioinformatics/bti677

Genome analysis

A comparison study: applying segmentation to array CGH data for downstream analyses

Hanni Willenbrock[1] and Jane Fridlyand[2,*]

[1]Center for Biological Sequence Analysis, Department of Biotechnology, Building 208, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark and [2]Department of Epidemiology and Biostatistics, University of California at San Francisco, 2340 Sutter Street, N224, San Francisco, CA 94143, USA

Received on July 5, 2005; revised on September 8, 2005; accepted on September 12, 2005
Advance Access publication September 13, 2005

frequently cited by BioDiscovery (Nexus)

Willenbrock, Fridlyand:

Our results have indicated that segmentation by any of the three methods aids downstream analyses of array CGH data. Of the methods under comparison, DNAcopy has the best operational characteristics in terms of its sensitivity and FDR for breakpoint detection. However, it should be noted that it is not able to identify single clone aberrations. While our comparison was limited to only three methods, albeit widely used, our study sets an example as a

DNAcopy
HMM (Fridlyand)
GLAD (Hupe)

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

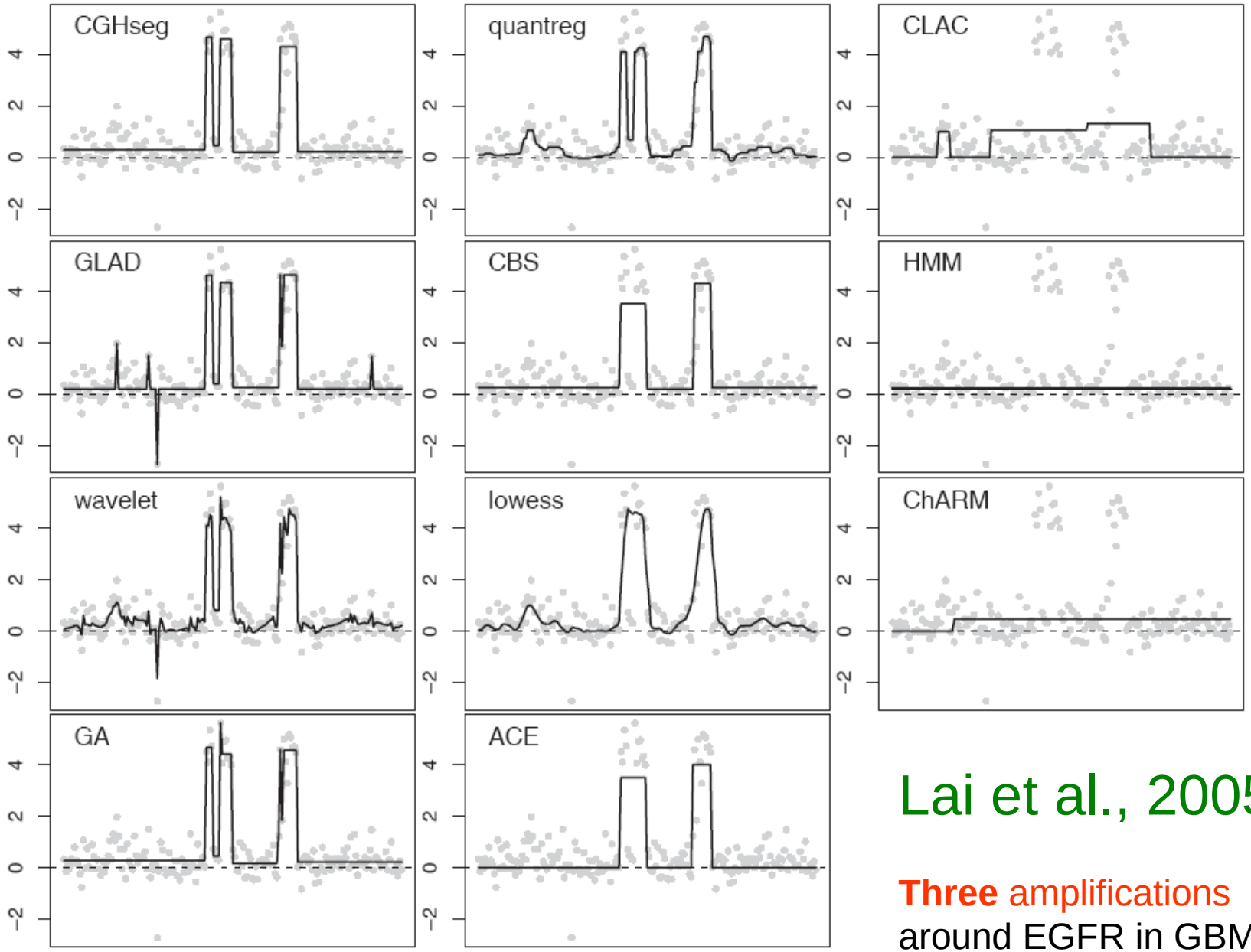
Genetics and population analysis

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

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

”SNP-CGH”

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

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

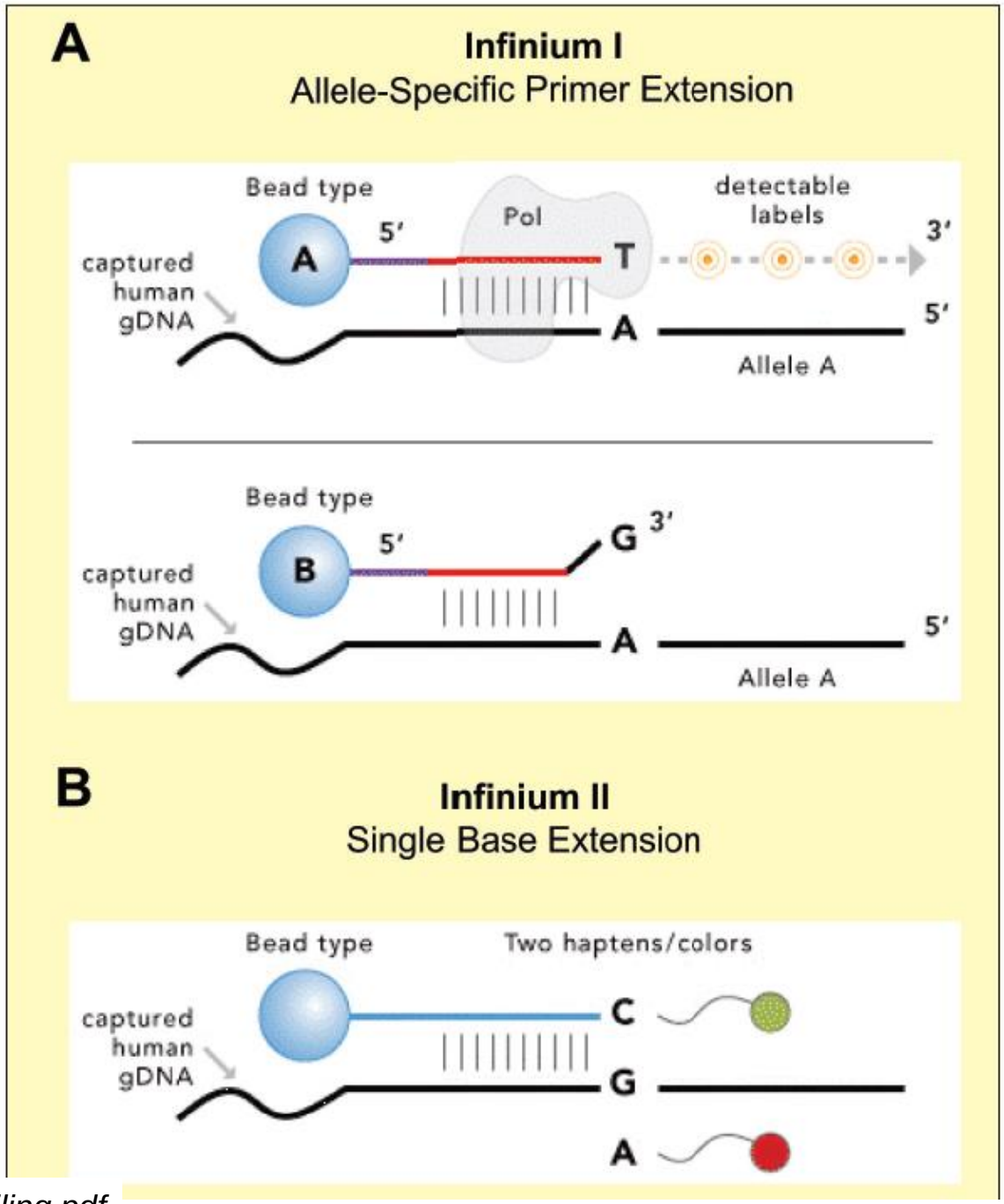
Genome Res. 2006 16: 1136-1148; originally published online Aug 9, 2006;
Access the most recent version at doi:10.1101/gr.5402306

High-resolution genomic profiling of chromosomal aberrations.pdf

SNP-CGH

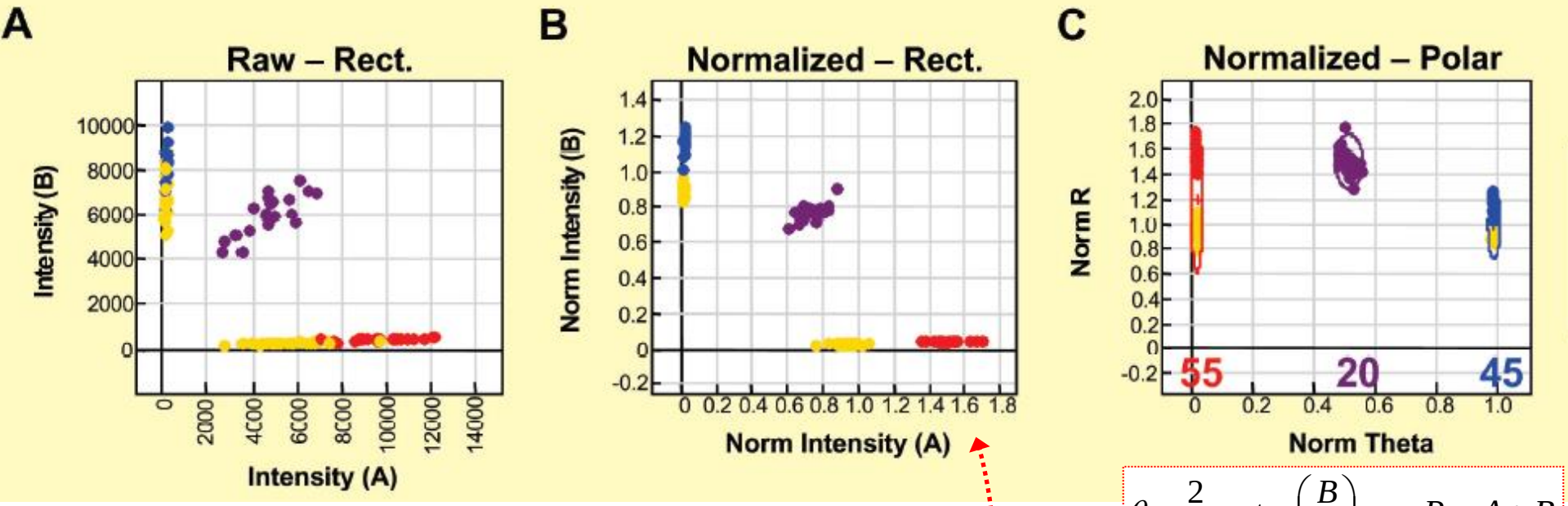
- simultaneous measurement of both signal intensity and allelic composition
  - **(BAF = B-Allele Frequency)**
- detect both copy number changes and copy-neutral loss-of-heterozygosity (LOH)
- Infinium whole-genome genotyping (WGG) BeadChips

What B-allele frequency



SNP-CGH technologies for genomic profiling.pdf

Calculation of the BAF



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

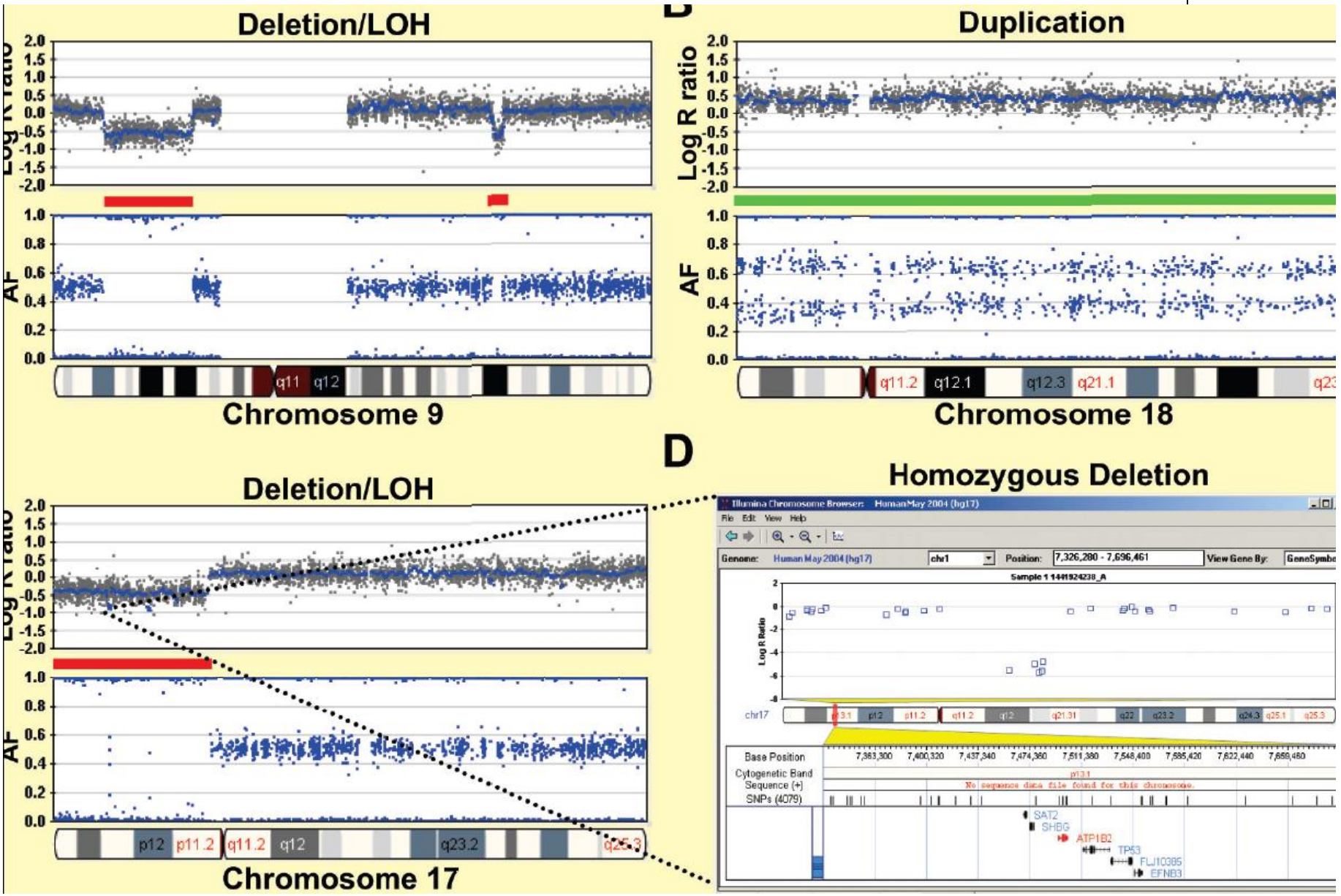
chrX, 120 normal individuals, one particular SNP
raw intensities: males in yellow, the others females
normalization is done using a ”proprietary” algorithm (B)
conversion to ”*polar coordinates*” (C)

B/A	θ
0	0
1	0.5
inf	1

”*canonical clusters*”

SNP-CGH technologies for genomic profiling.pdf

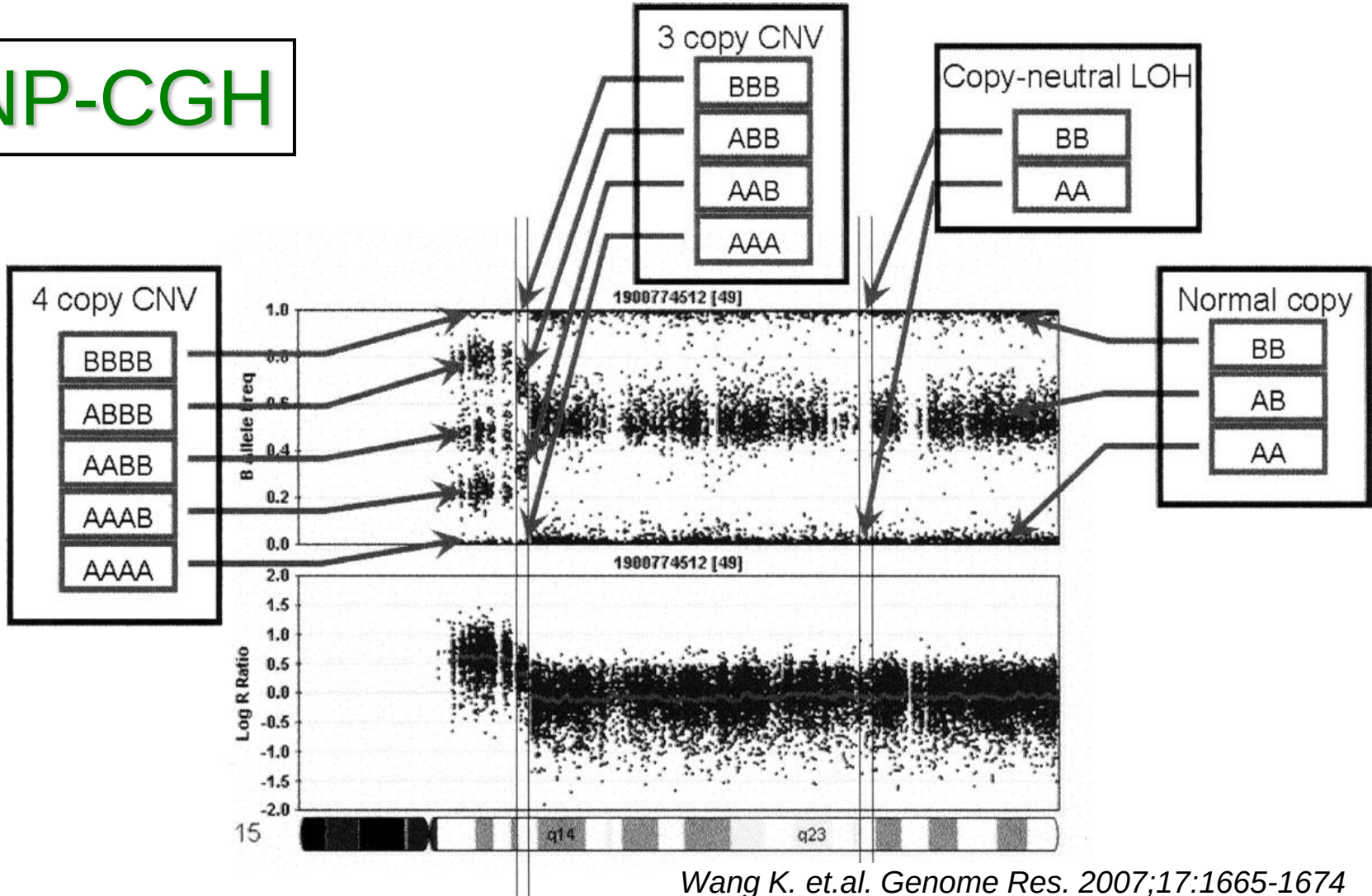
LRR and BAF illustrated



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



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

Illumina academic partners

University of Pennsylvania
Department of Genetics
University of Pennsylvania
Philadelphia, PA 19104
Phone: 215.898.0021
Contact: Kai Wang
Email: kai@mail.med.upenn.edu
Web: www.neurogenome.org/cnv/penncnv/
Application Areas Supported: Copy Number Variation Analysis (CNV)

PennCNV

The Wellcome Trust Centre for Human Genetics
Roosevelt Drive
Oxford
OX3 7BN
United Kingdom
Phone: +44 (0)1865 287500
Contact: Iaonnis Ragoussis
Email: ioannisr@well.ox.ac.uk
Web: www.well.ox.ac.uk
Application Area Supported: Copy Number Variation (CNV) analysis

QuantiSNP

plink...

PLINK
Whole genome association analysis toolset

The link below describes how Illumina's WG arrays data from BeadStudio can be converted into a file input format for WGAS using PLINK.

PLINK Support for BeadStudio data output

dChip

dChip
Analysis and visualization of gene expression and SNP microarrays
Contact: Dr Cheng Li
Department of Biostatistics, Harvard School of Public Health and
Department of Biostatistics and Computational Biology
Dana-Farber Cancer Institute
375 Longwood Ave, 6th Floor
Boston, MA, 02215 Email: cli@hsph.harvard.edu
Web: http://biosun1.harvard.edu/complab/dchip/
Application Areas Supported: Copy Number Variation (CNV) analysis, Gene Expression (GEX)

dChip

PennCNV script on anaconda:

Usage: run_PennCNV OPTIONS

-f S1.txt : Illumina SNP-CGH file for one sample
-minsnp 10 : Minimum number of markers in an aberration
-minlength 50k : Minimum length of an aberration
-gcmodel : Try reducing fluctuations caused by GC-content variation

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

PennCNV Input:

	A	B	C	D	E	F
1	**Name**	**Chr**	**Position**	**X.GType**	**X.Log R Ratio**	**X.B Allele Freq**
2	rs12354060	1	10004	BB	0.05767579	1
3	rs2691310	1	46844	NC	-0.1525835	0.5361379
4	rs2531266	1	15	NC	0.2049716	0.3973282
5	rs4124251	1		NC	-0.09452707	0.09473621
6	rs8179466	1	22	NC	-0.02189994	1
7	rs6603779	1	2277	NC	0.0836625	0.6220879
8	cnvi0007379	1	311662		-0.1188801	1
9	cnvi0019140	1	314893		-0.1406044	0
10	cnvi0007389	1	318309	N	0.1599025	0
11						

same as for run_CBS

Tomas Axelsson, ETJ1 data

PennCNV Paper

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

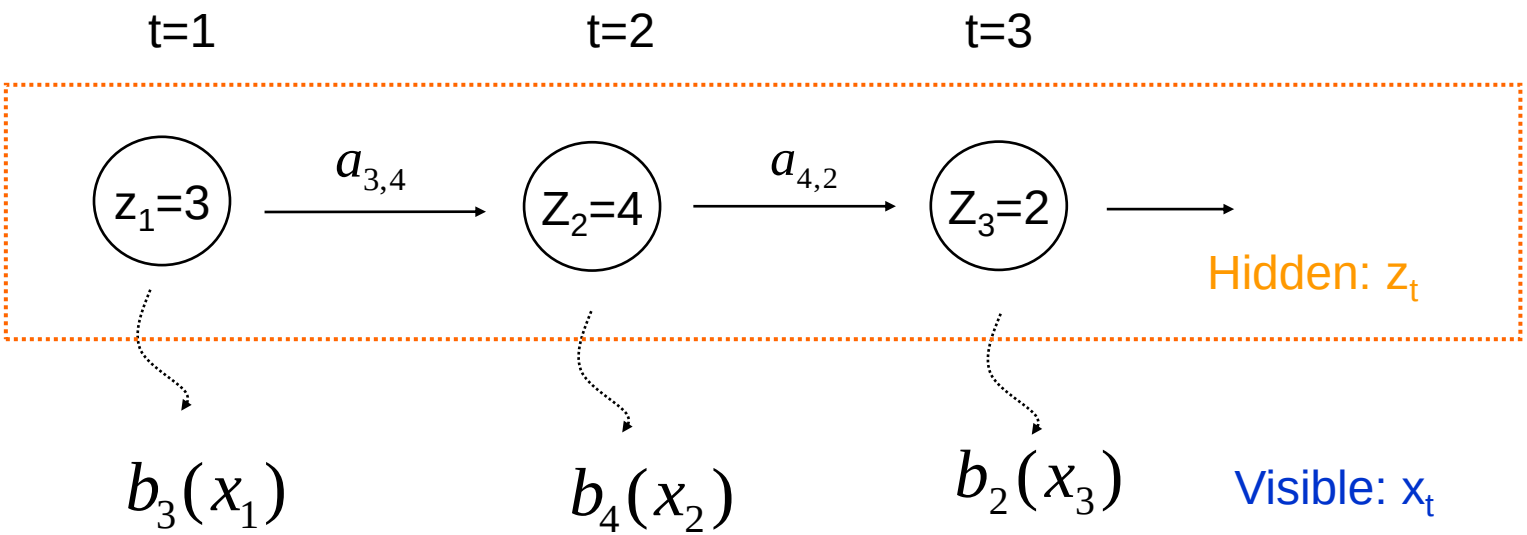
Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F.A. Grant, Hakon Hakonarson and Maja Bucan

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)

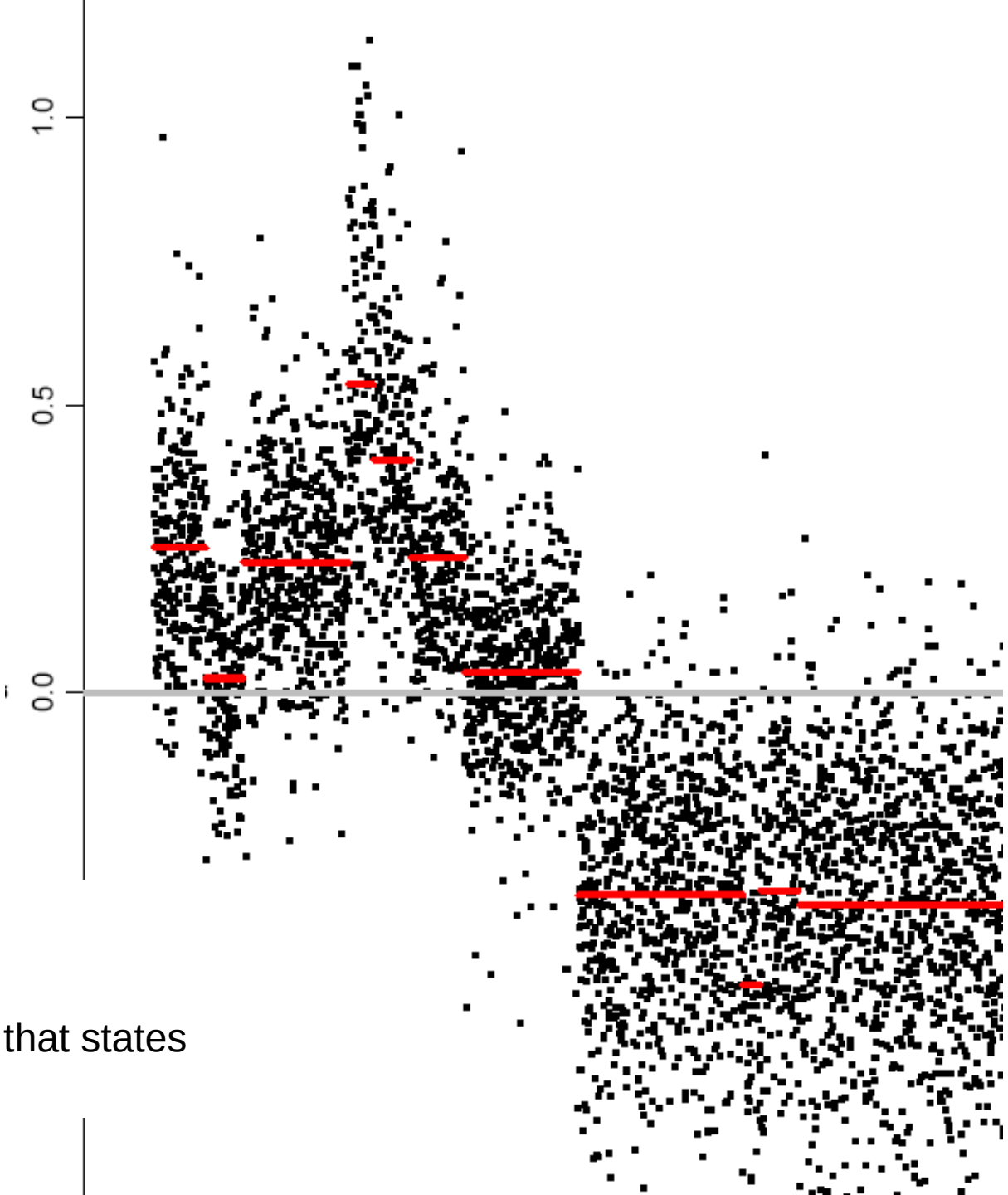
Hidden Markov Model



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

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

HMM for CNV detection



states

. . . . emissions from that states

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

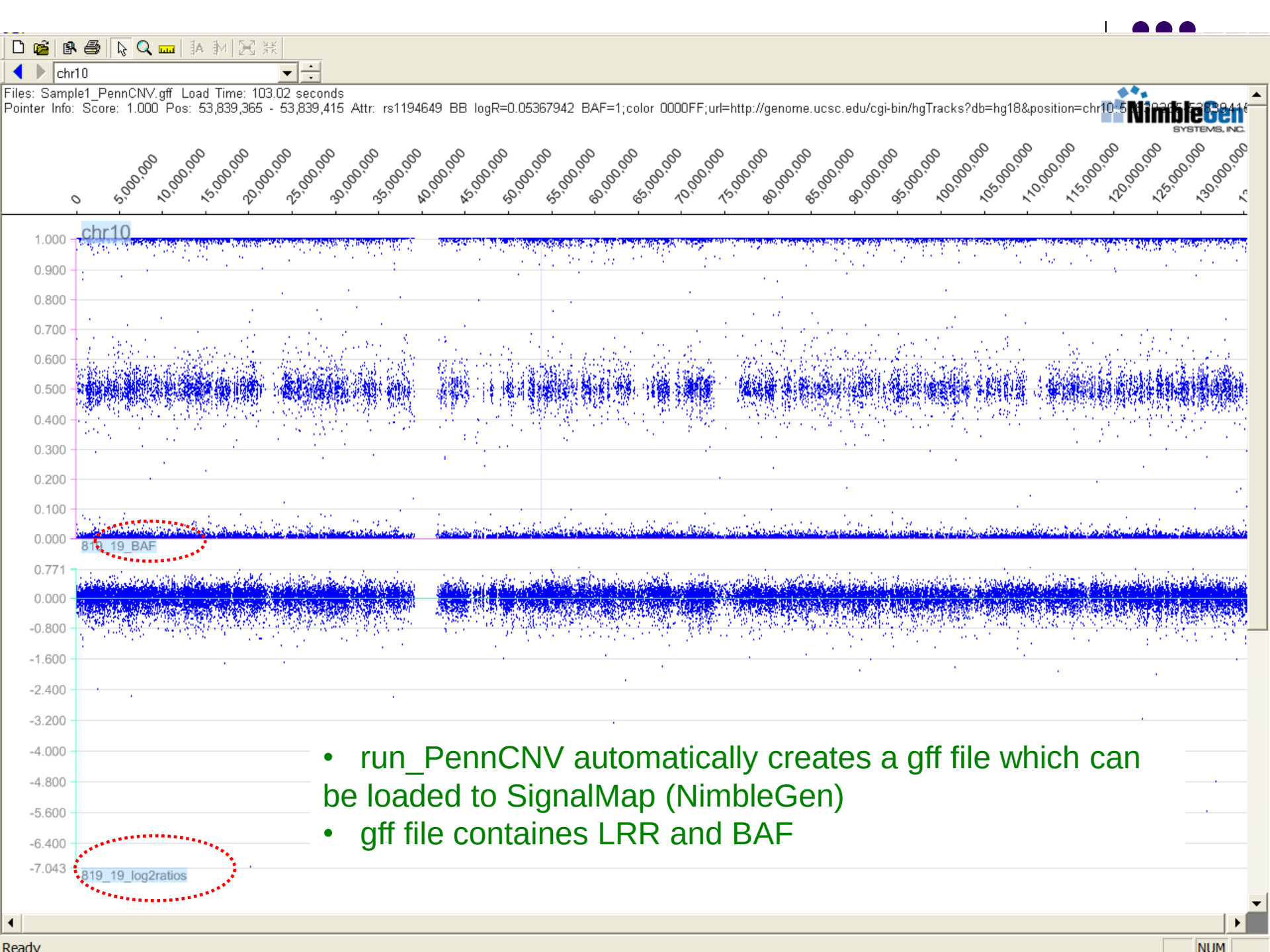
Allele frequency information included in the states of the HMM

PennCNV implementation

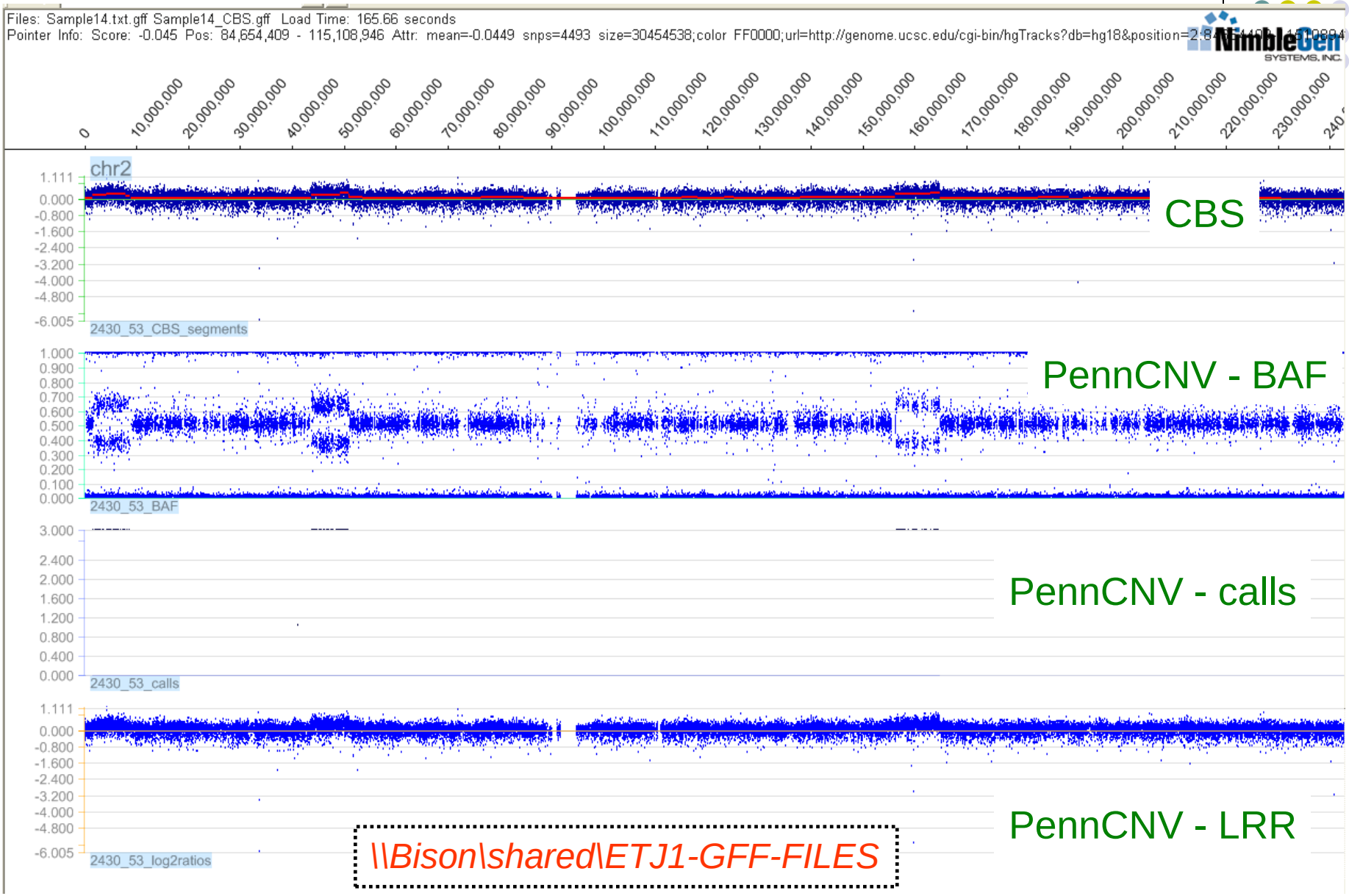
run_PennCNV -f Sample1.txt -minsnp 10 -minlength 50k -gcmodel

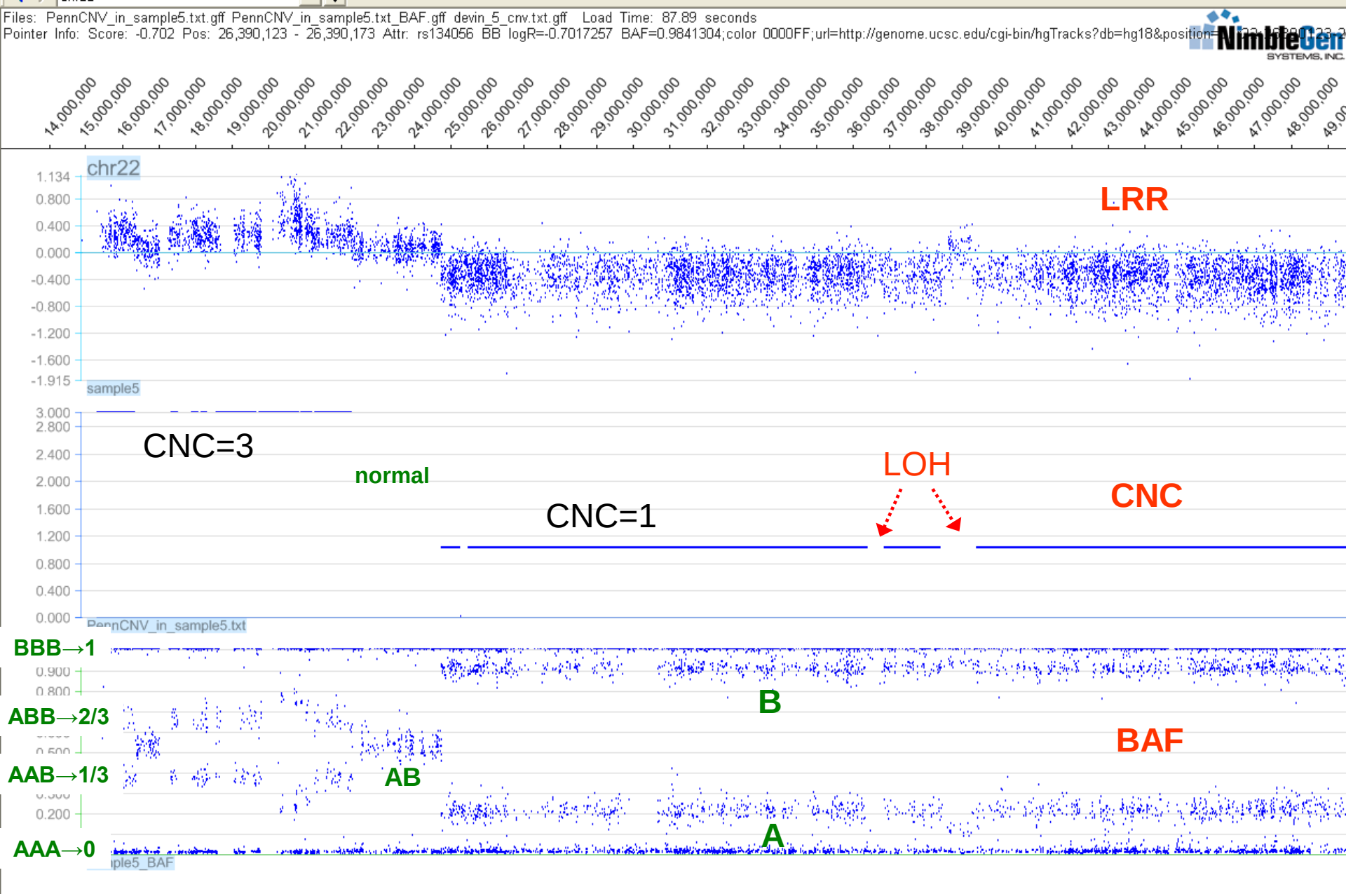
same inputfile as for run_CBS

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


chr10
Files: Sample1_PennCNV.gff Load Time: 103.02 seconds
Pointer Info: Score: 1.000 Pos: 53,839,365 - 53,839,415 Attr: rs1194649 BB logR=0.05367942 BAF=1;color 0000FF;url=http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg18&position=chr10:5
NimbleGen SYSTEMS, INC.
0 5,000,000 10,000,000 15,000,000 20,000,000 25,000,000 30,000,000 35,000,000 40,000,000 45,000,000 50,000,000 55,000,000 60,000,000 65,000,000 70,000,000 75,000,000 80,000,000 85,000,000 90,000,000 95,000,000 100,000,000 105,000,000 110,000,000 115,000,000 120,000,000 125,000,000 130,000,000
chr10
1.000 0.900 0.800 0.700 0.600 0.500 0.400 0.300 0.200 0.100 0.000
819_19_BAF
0.771 0.000 -0.800 -1.600 -2.400 -3.200 -4.000 -4.800 -5.600 -6.400 -7.043
819_19_log2ratios
• run_PennCNV automatically creates a gff file which can be loaded to SignalMap (NimbleGen)
• gff file containes LRR and BAF
Ready
NUM

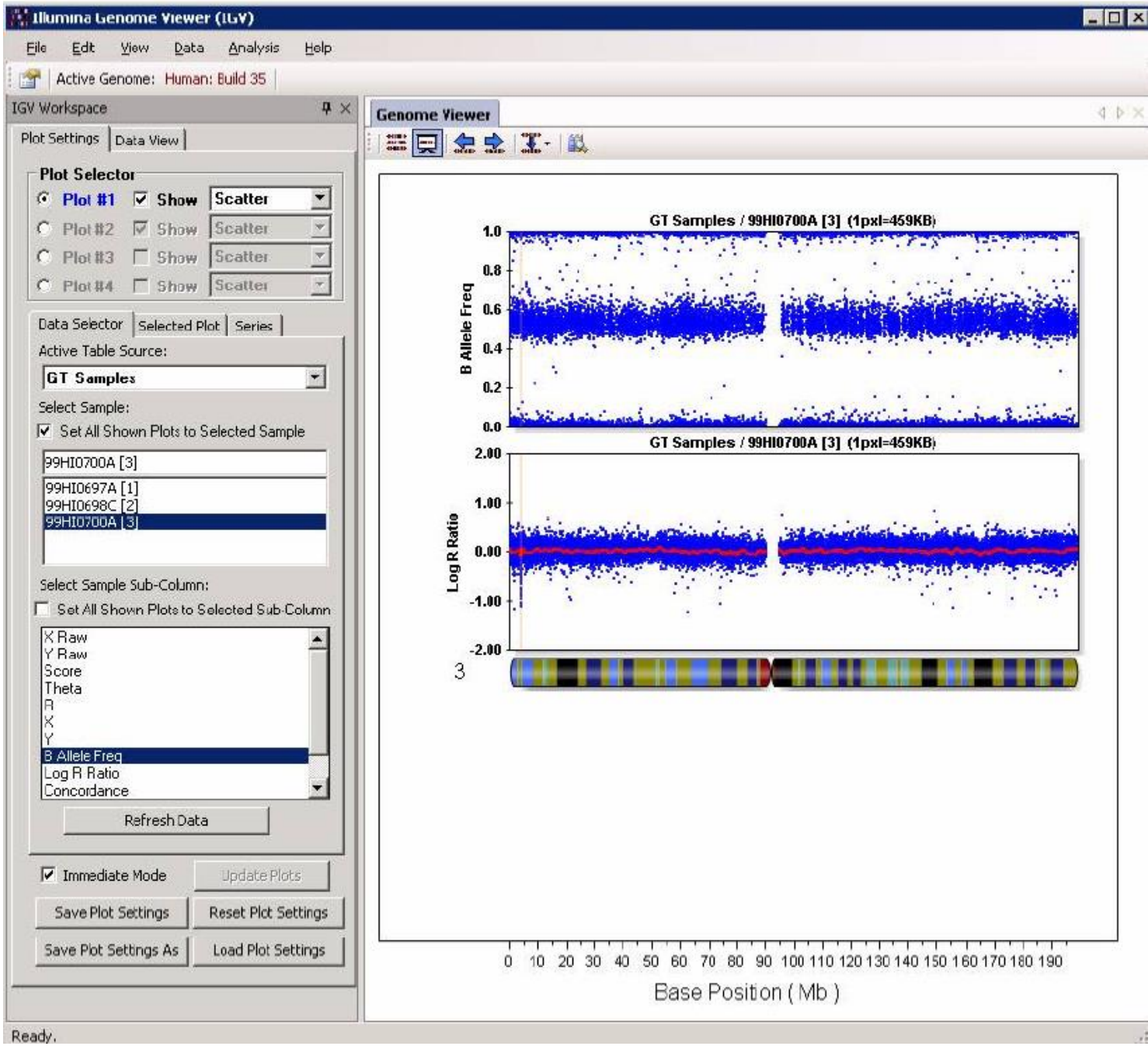
Comparison with CBS





Devin, Tumor 5, chr22 PennCNV results

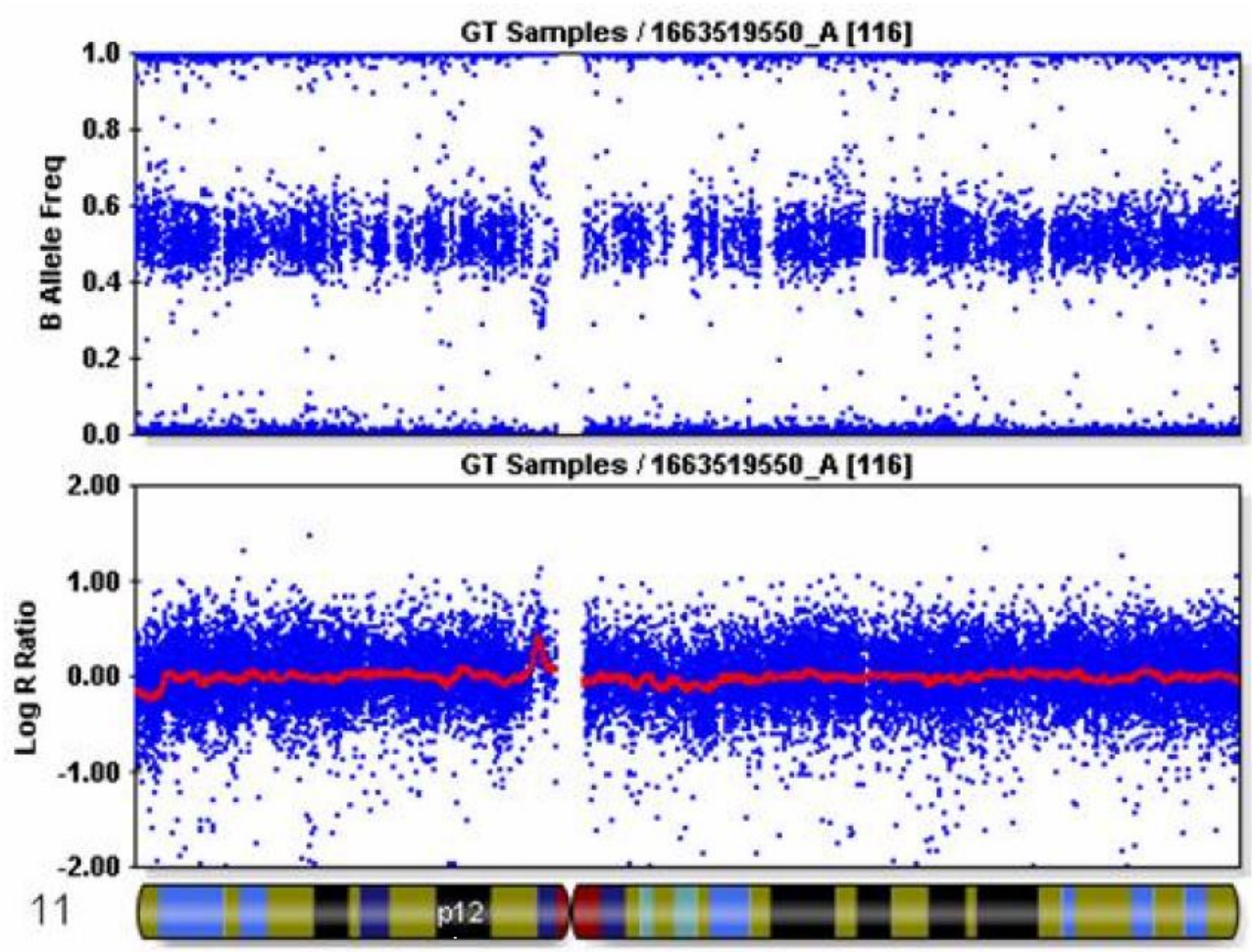
Back to Genome Studio



PennCNV quality assessment

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

Large variance of LRR values

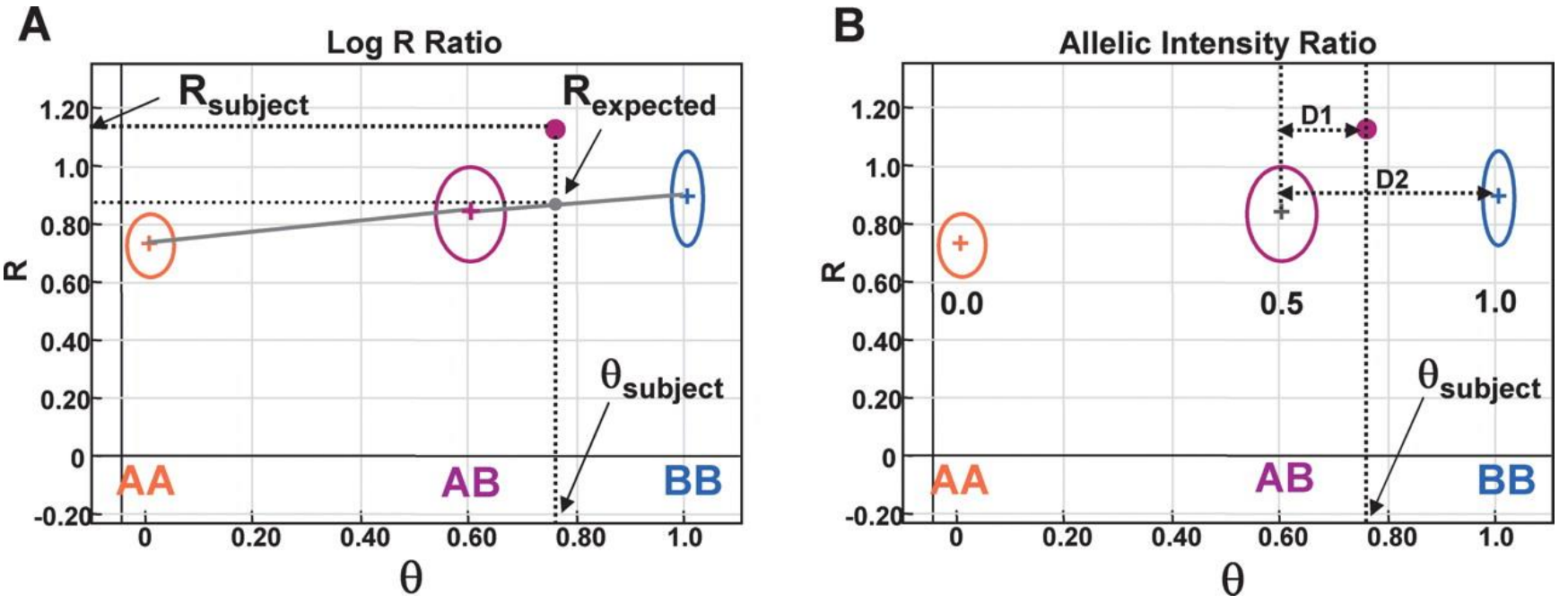


Logfile:
LRR_SD=0.2184

Threshold:
LRR_SD < 0.2

possible cause: use of non-optimal **canonical clustering** files

Canonical clusters

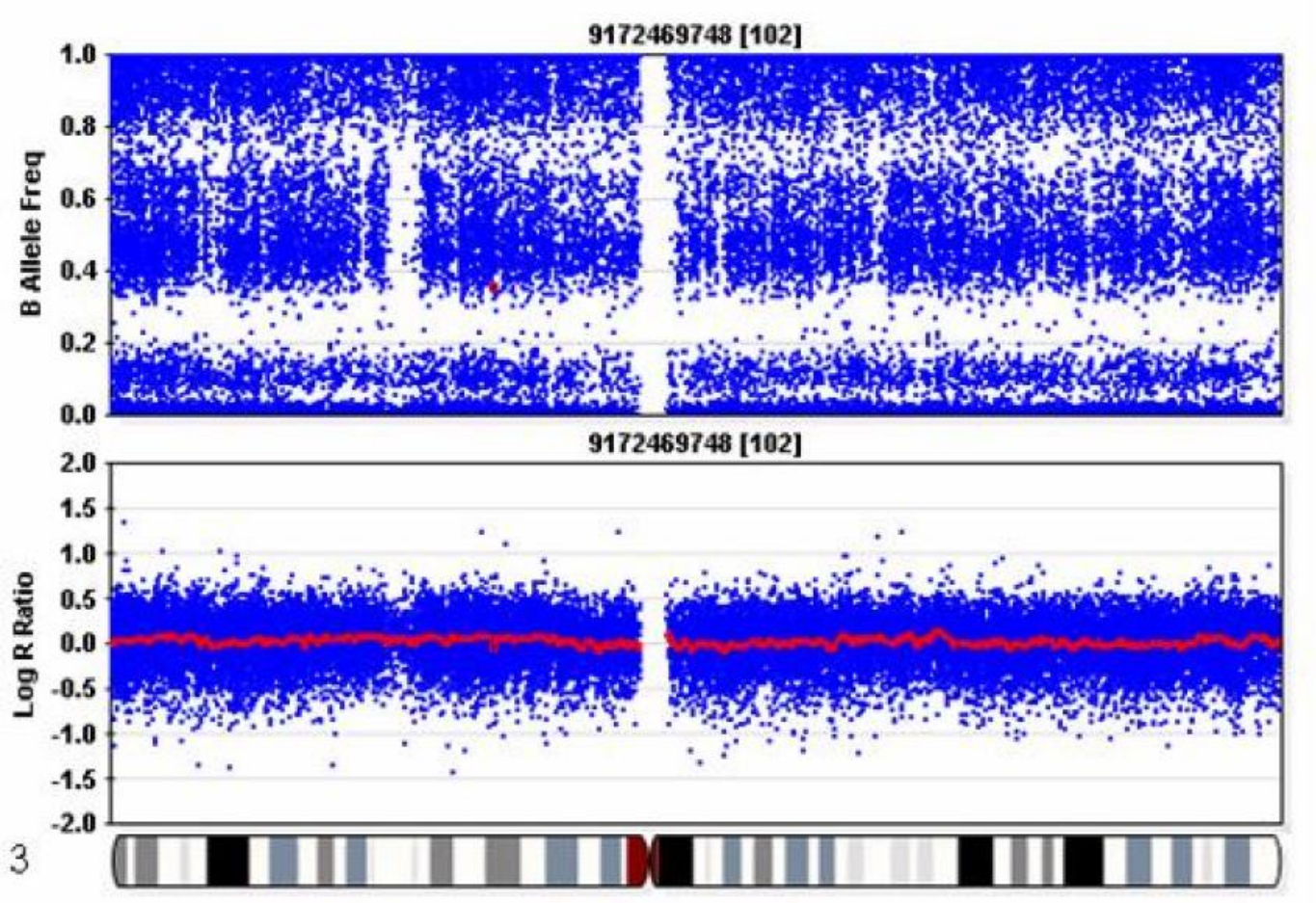


The canonical clusters are **not specific enough**

- 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

Large variance of BAF-values

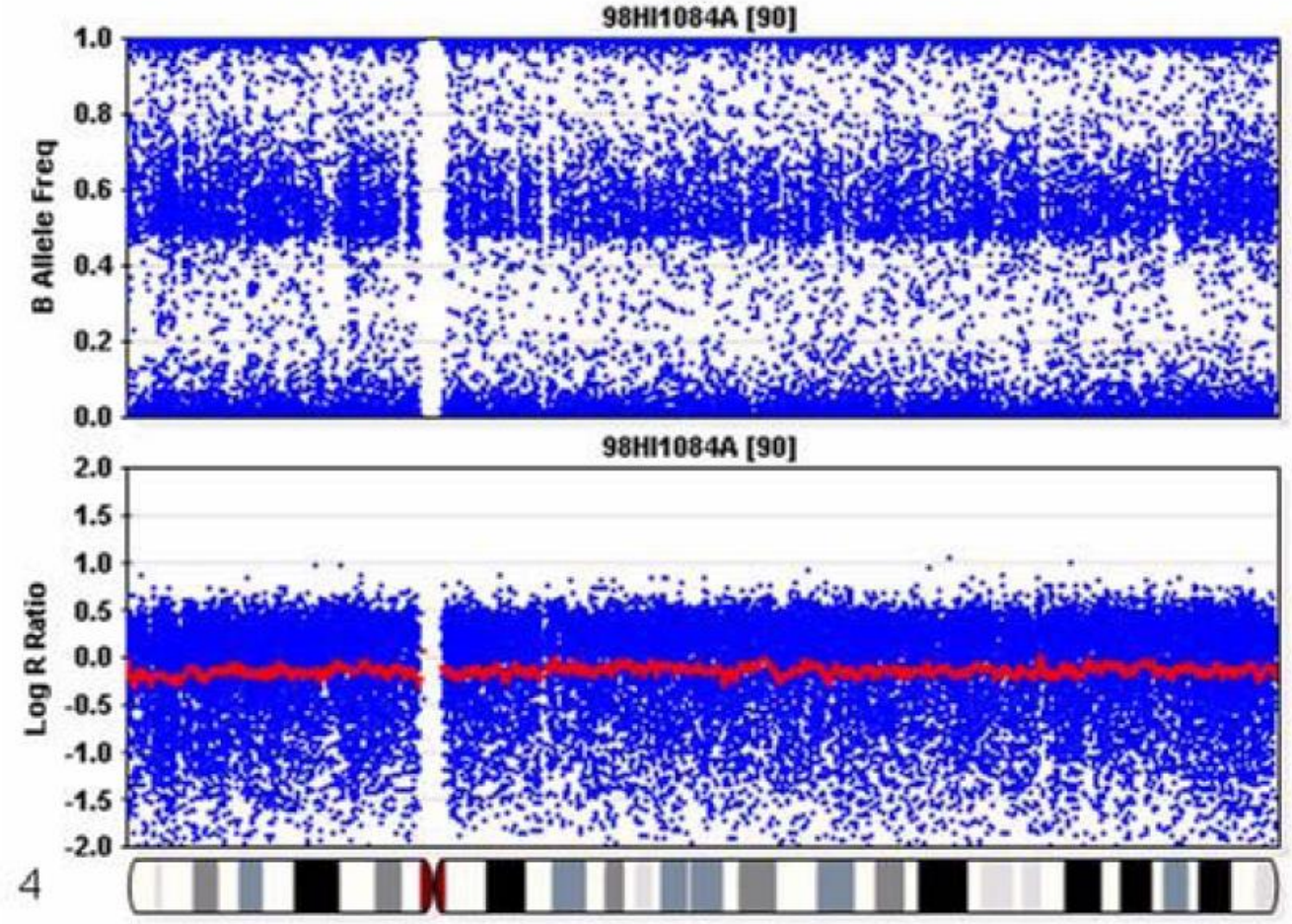


BAF_SD: stdev. of all autosome BAF values between 0.25 and 0.75

possible cause: mixing two different genomes (?)

PennCNV_quality.pdf

Failure on one allele

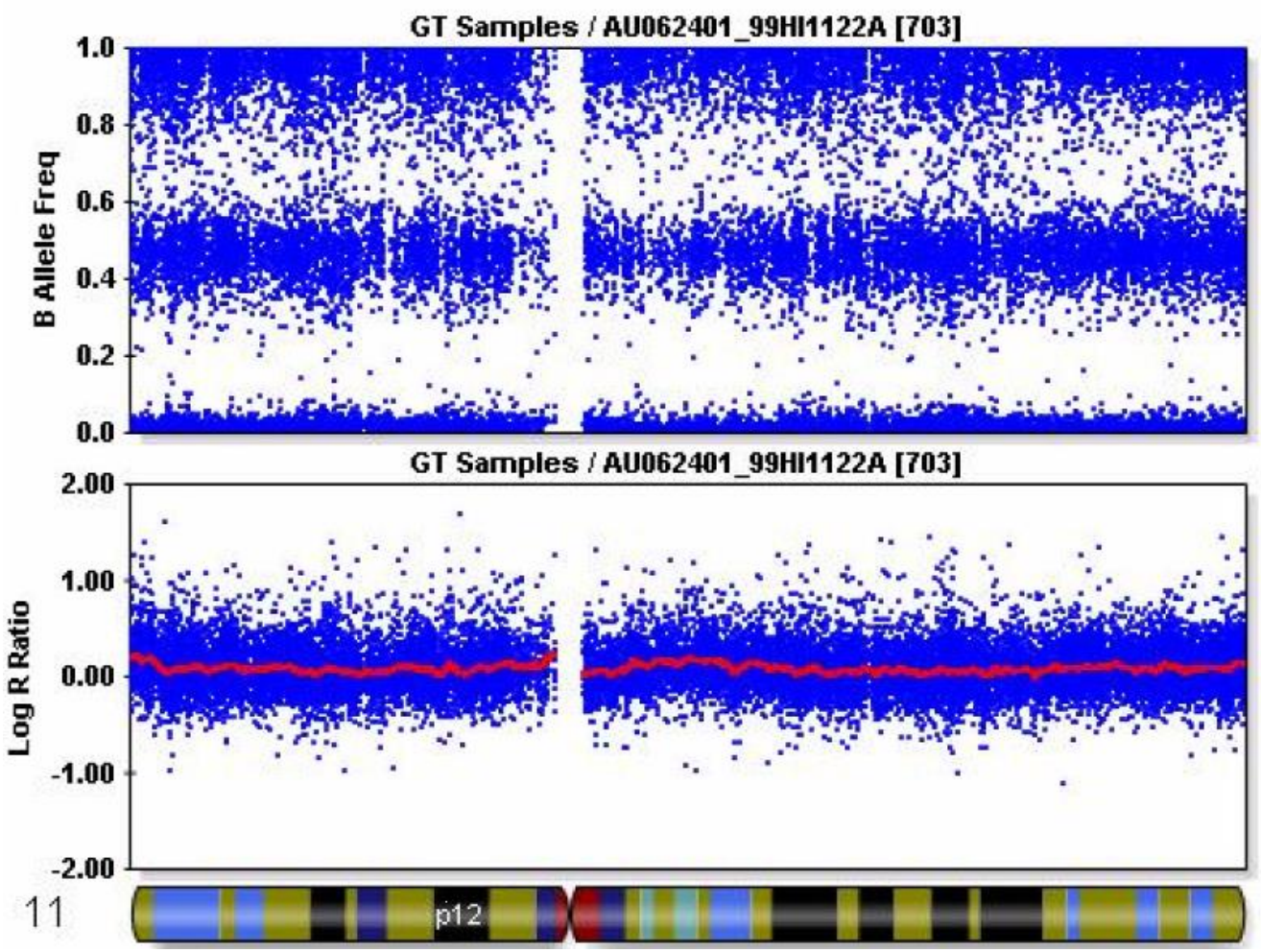


Allele A generally not measured well

cause: unknown

PennCNV_quality.pdf

Upshift or downshift of BAF values

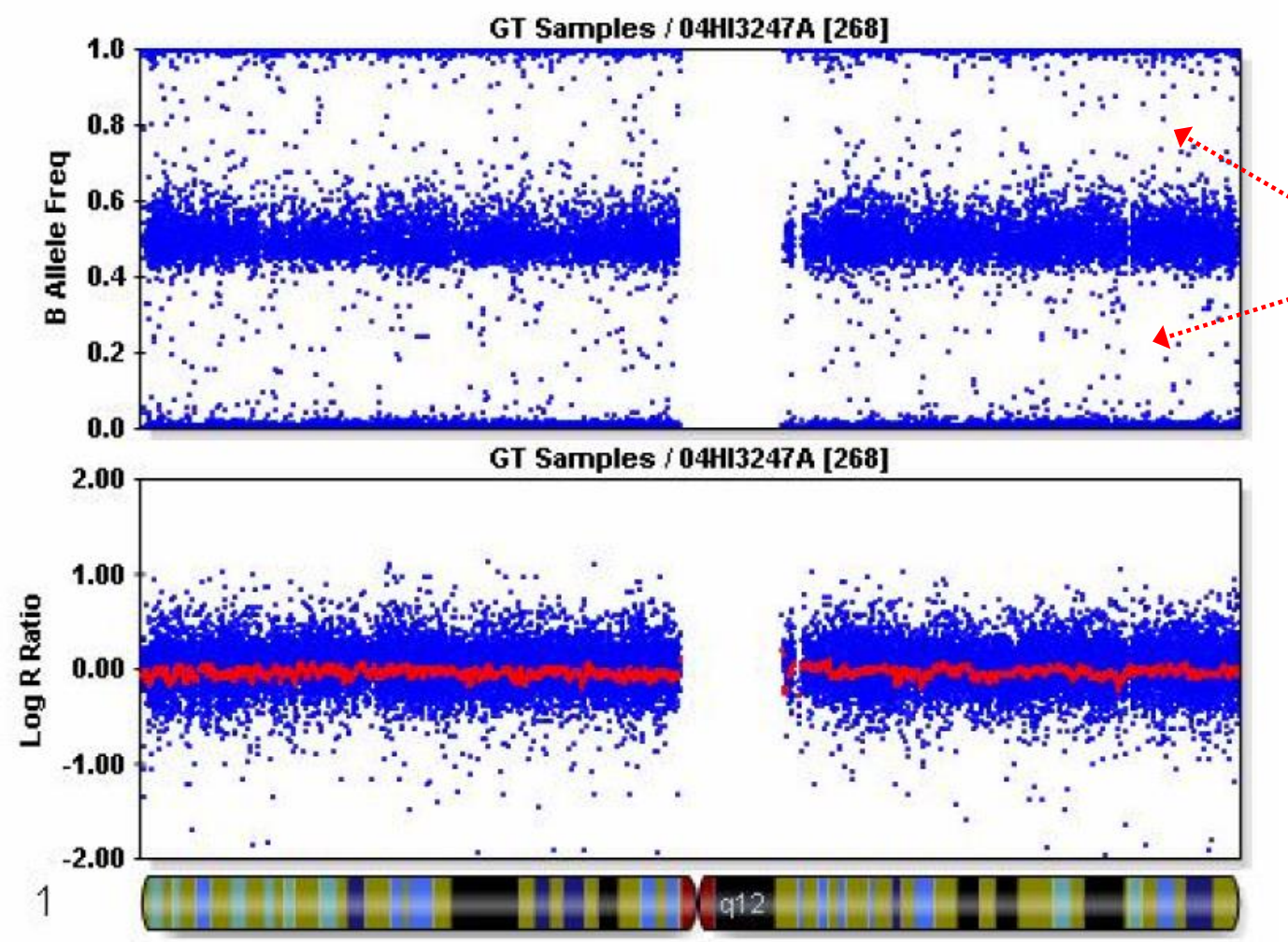


downshift

causes false positive CNV duplication calls

cause: unknown

Random failures

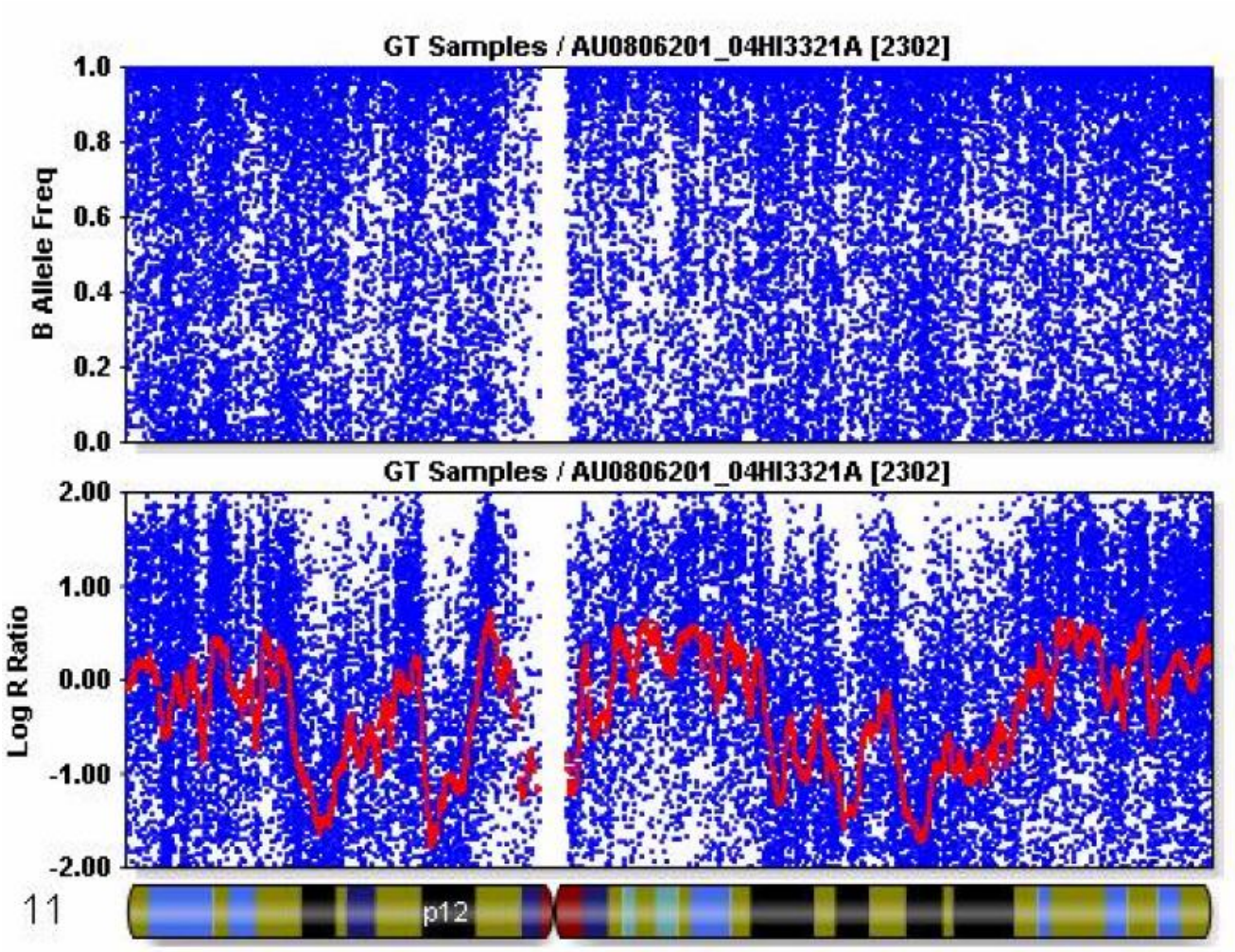


here should be nothing!

→ false positive CNV calls on homozygous deletions

BAF_DRIFT

Waviness



cause: too much or too little DNA (or something else)

Quality ETJ1

Sample1	large SD for LRR 0.241
Sample2	large SD for LRR 0.225
Sample3	large SD for LRR 0.327
Sample4	large SD for LRR 0.349
Sample4	drifting BAF values drift=0.0020
Sample5	large SD for LRR 0.2296
Sample6	large SD for LRR 0.2320
Sample7	large SD for LRR 0.4999
Sample7	drifting BAF values drift=0.002147
Sample8	large SD for LRR 0.3747
Sample8	drifting BAF values drift=0.004070
Sample9	large SD for LRR 0.216
Sample9	waviness factor values wf=-0.0822
Sample10	large SD for LRR 0.2011
Sample11	large SD for LRR 0.3709
Sample12	large SD for LRR 0.467
Sample12	drifting BAF values drift=0.00236
Sample13	large SD for LRR 0.2119
Sample14	large SD for LRR 0.2155
Sample15	large SD for LRR 0.2576
Sample16	large SD for LRR 0.4398

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)

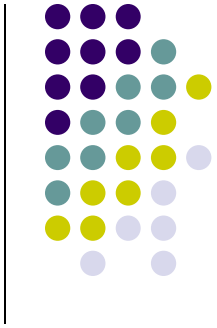
Dear Uwe,

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

-Soheil

dChip

Windows (Java), free



Research

People

Publications

Software

Links

dChip Software: Analysis and visualization of gene expression and SNP microarrays

Introduction	Manual	References	Tutorials
Download	User Group	Updates	User Help
Workshops			

New: dChip course in Boston, Jan 26 – 28, 2009, offered by bioinformatics.org.

Search manual, site & user group

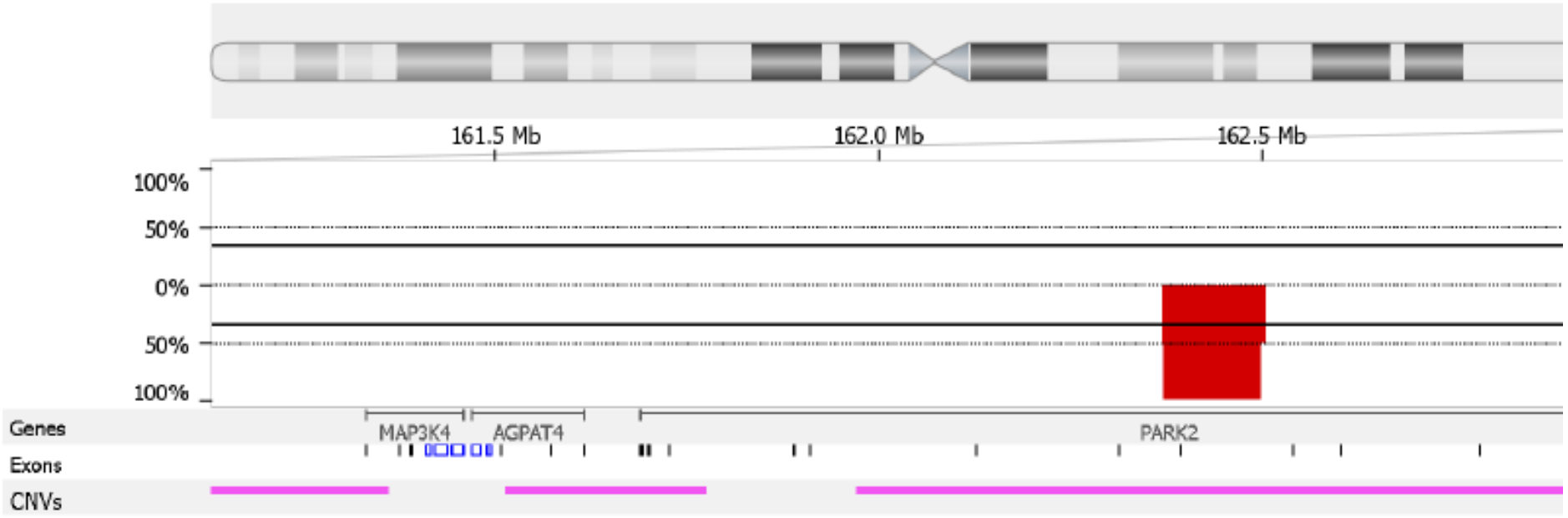
Google™ Custom Search Search

HMM to infer copy number:

Zhao et al. "*An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays*" [Cancer Research 64, 3060-3071, May 1, 2004]

Comparison of samples

- Frequency plots (Nexus):



Comparison of samples

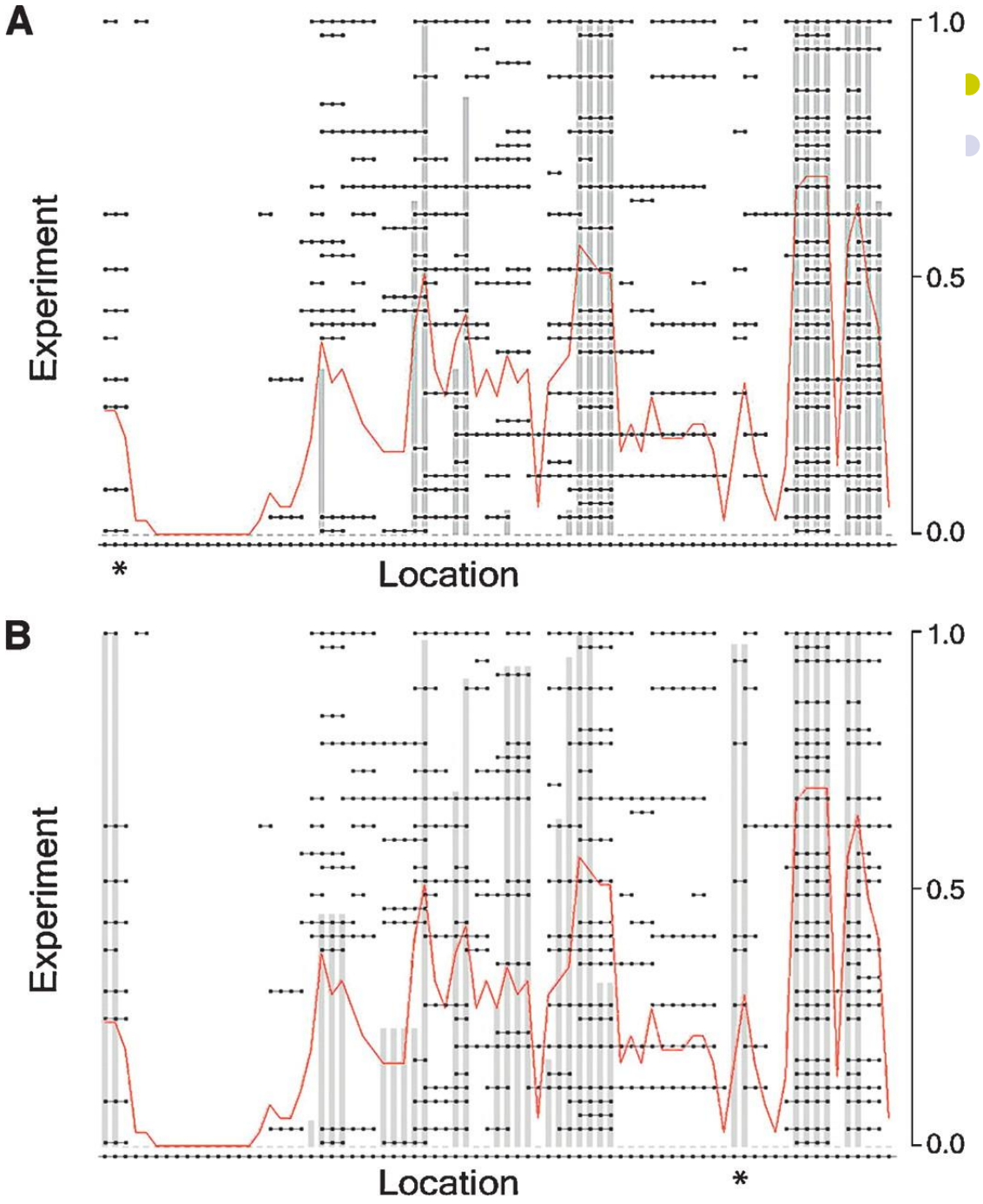
- STAC:

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

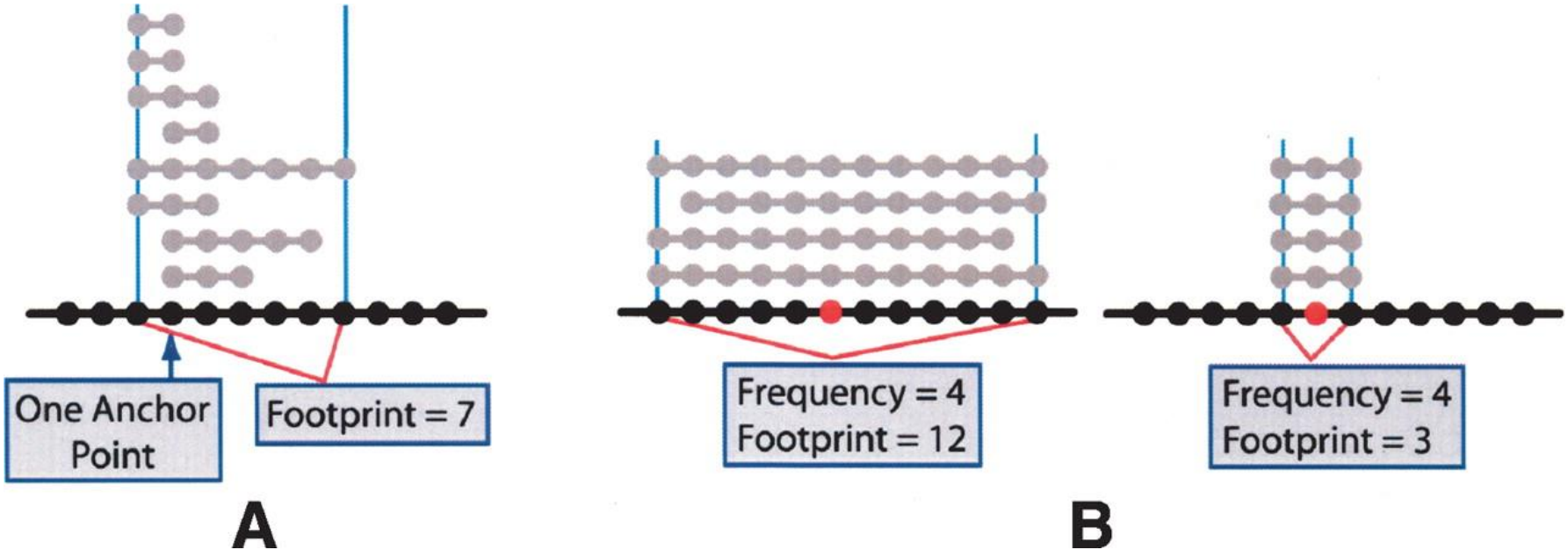
Sharon J. Diskin, Thomas Eck, Joel Greshock, et al.

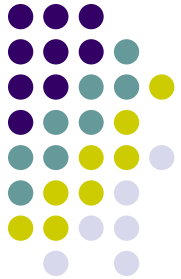
Genome Res. 2006 16: 1149-1158
Access the most recent version at doi:10.1101/gr.5076506

STAC-results



STAC-results

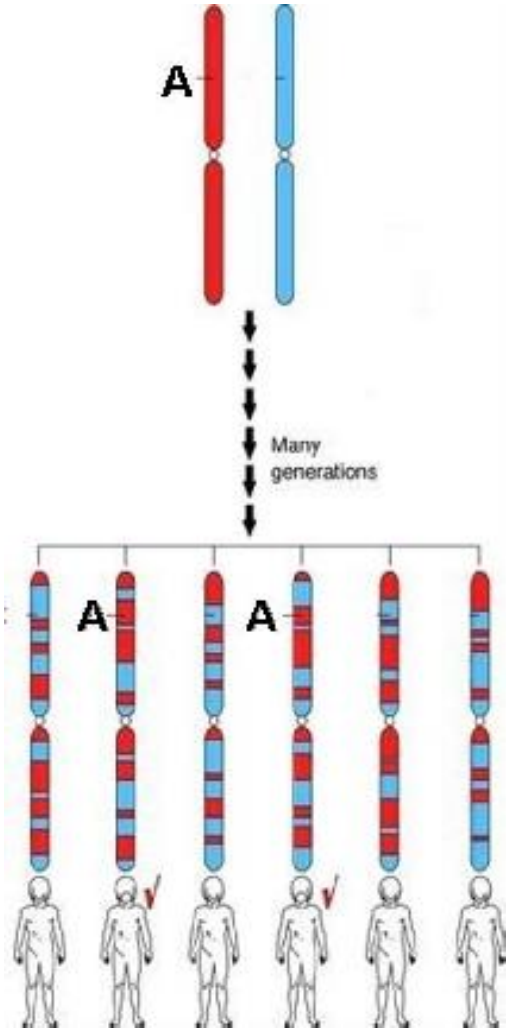




Thanks!

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

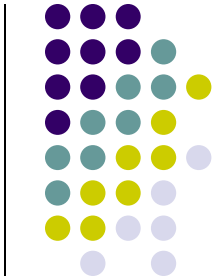
Haplotypes and tag SNPs



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