GENOMIC COPY NUMBER

Rudy Guerra
Department of Statistics
Rice University
April 14, 2008

Outline

- Introduction
- Statistical methods for copy number
- On-going research with TCH

Introduction

* Genomic disorder: a disease caused by an alteration of the genome that results in complete loss, gain or disruption of the structural integrity of a dosage sensitive gene(s).

Examples: Sotos syndrome, split hand-split foot malformation

* Rearrangements often flanked by large (usually >10 kb), highly homologous low copy repeat (LCR) structures.

* Such rearrangements occur via recombination mechanisms whereas point mutations usually result from DNA replication or repair errors.

Genomic disorders: rearrangements
Mendelian disorders: (point) mutations

Cancer

Breast cancer:

- 7 cell lines
- 75 gains and 48 losses

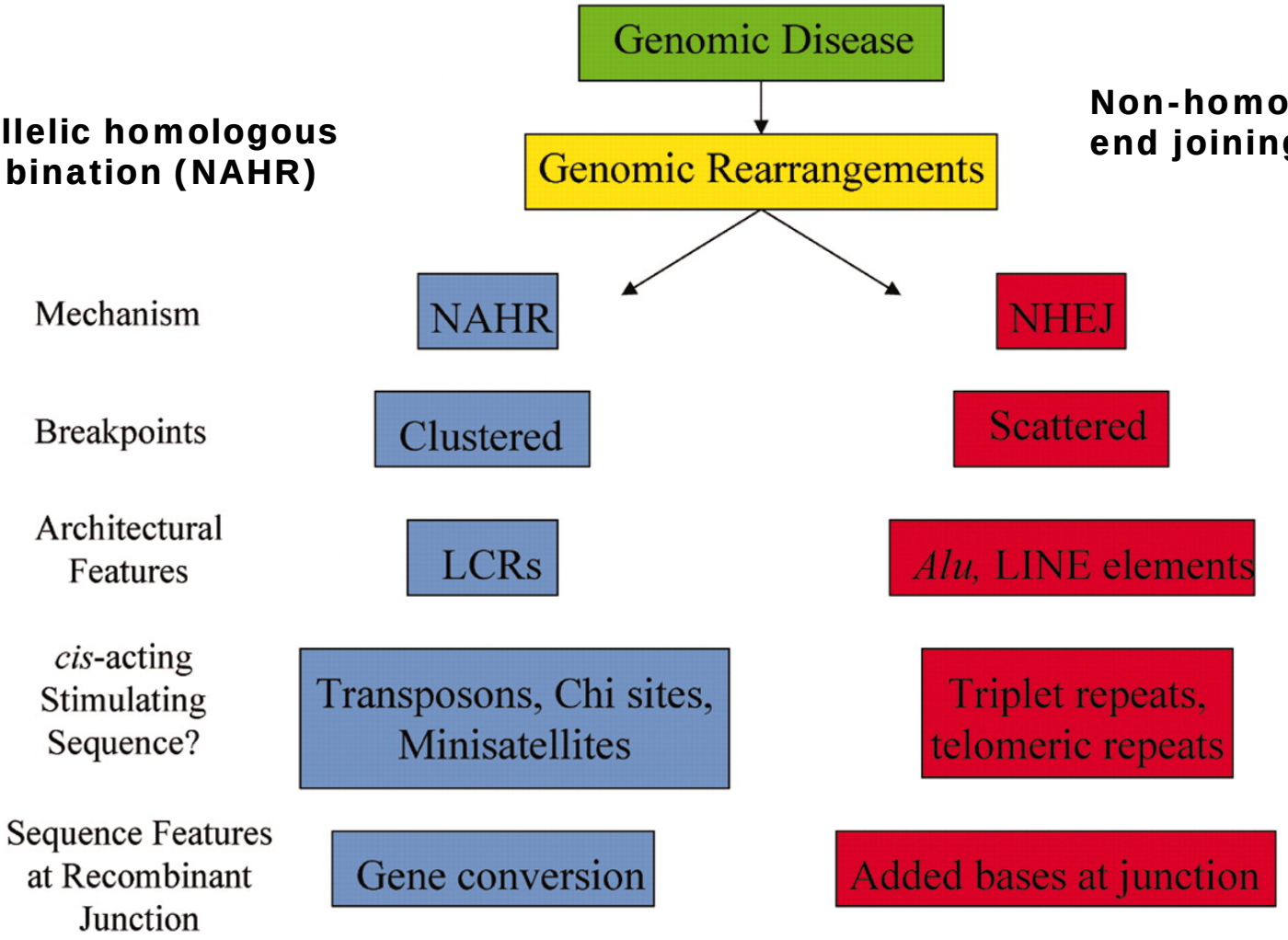
Prostate cancer:

- 4 cell lines
- 28 gains and 18 loss

Colorectal cancer:

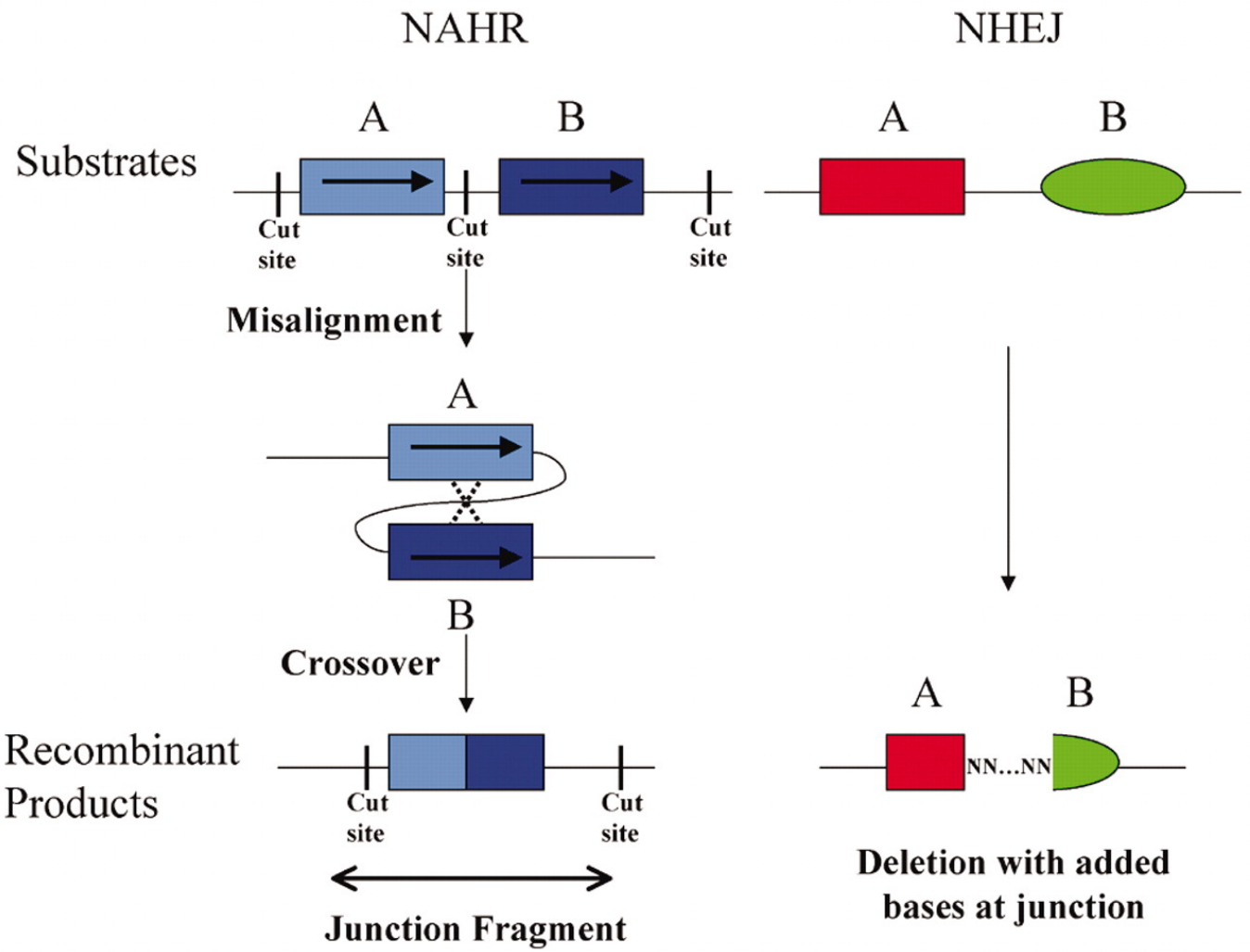
- 48 cell lines
- gain of chromosome 20

Mechanisms of Genomic Rearrangements



Shaw, C. J. et al. Hum. Mol. Genet. 2004 13:R57-64R; doi:10.1093/hmg/ddh073

Human Molecular Genetics



Shaw, C. J. et al. Hum. Mol. Genet. 2004 13:R57-64R; doi:10.1093/hmg/ddh073

Human Molecular Genetics

DNA Copy Number

Copy Number is the number of copies of a particular segment of DNA sequence.

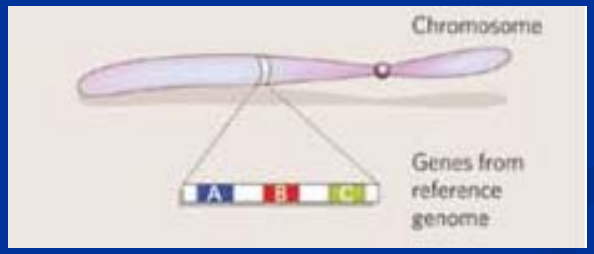
Normal: CN =2
Loss: CN = 0, 1
Gain: CN = 3
Amplification: CN = > 3

Copy Number Variation

A segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome [Feuk et al. 2006]

Copy Number Variation (CNV)

- A segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome [Feuk et al. 2006]
- Categories



Deletion

Duplication

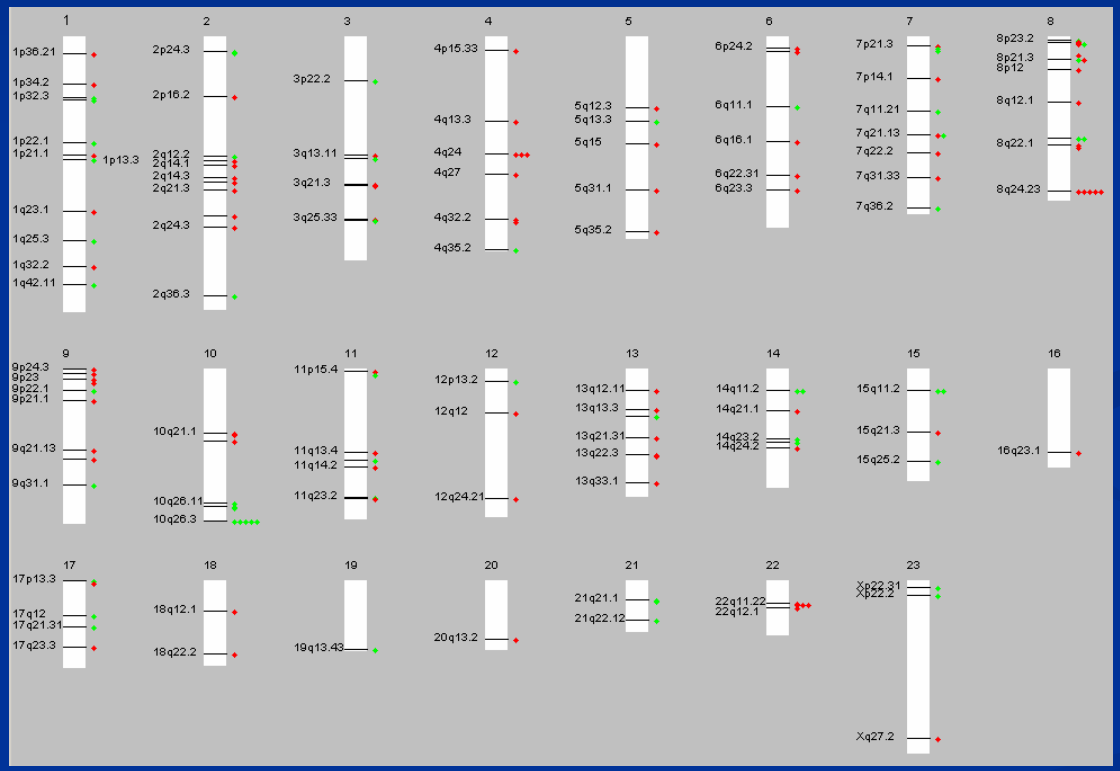
Insertion

Images are made from http://www.ghastlyfop.com/blog/uploaded_images/smaller-737680.jpg

Why are we interested in identifying CNVs?

- Another source of genomic variation
- Understand the genetic variation among the normal population (not necessarily two copies of each sequence)
- Distinguish normal copy number variation from aberrant genetic lesions in cancer

Whole-genome view of CNVs



- 143 CNV regions
  - 70% CNVs are novel
  - 37% gain
  - 63% loss
  - Min: 68 bp
  - Max: 18 Mb
  - Median: 86 Kb

Studies and Data for Copy Number Estimation

Four typical ways people get data to estimate copy number

- Karyotyping
- Comparative genomic hybridization (CGH)
- Array CGH
- SNP microarray

Down Syndrome

Cancer

Breast cancer:

7 cell lines

75 gains and 48 losses

Prostate cancer:

4 cell lines

28 gains and 18 loss

Colorectal cancer:

48 cell lines
gain of chromosome 20

First CGH: Kallioniemi et al., *Science* 1992

- Single experiment
- Comparative Genomic Hybridization: CGH
- CGH resolution: p and q arms
- Samples: green = tumor & red = normal reference
- Hybridizations: - tumor to normal metaphase
  - reference to normal metaphase
- Ratio = green/red ~ *k* CN/CN [log(green/red)]
- Reference is a control for hybridization noise (normalization)

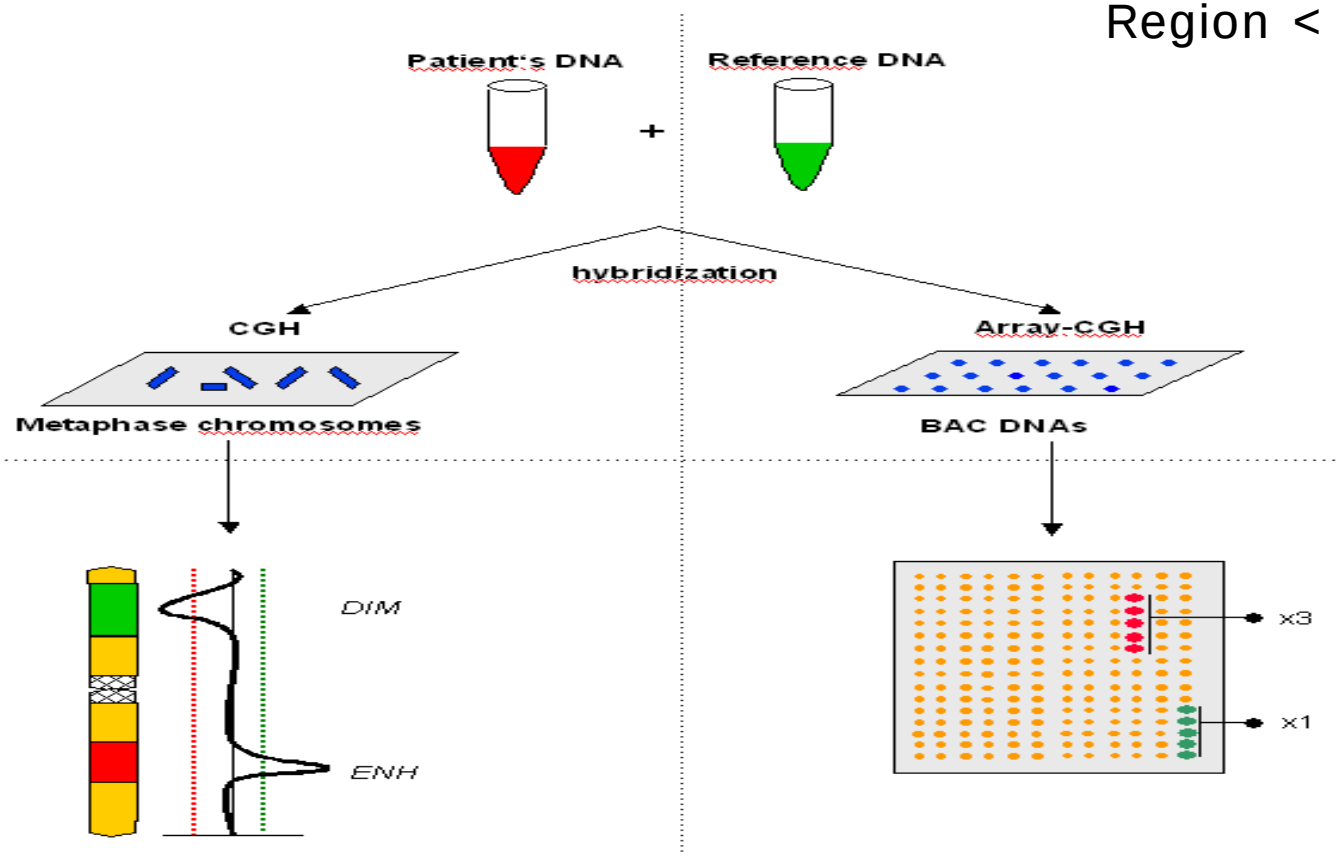
Limitations of CGH

- Regions < 20 Mb not detectable
- Closely spaced aberrations not detectable

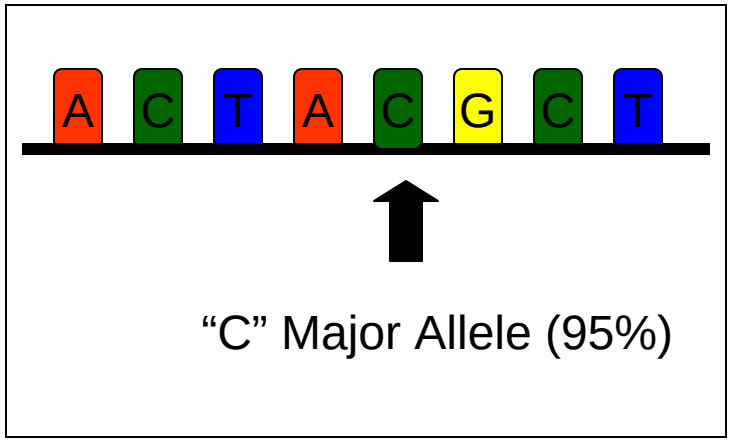
CGH and array CGH

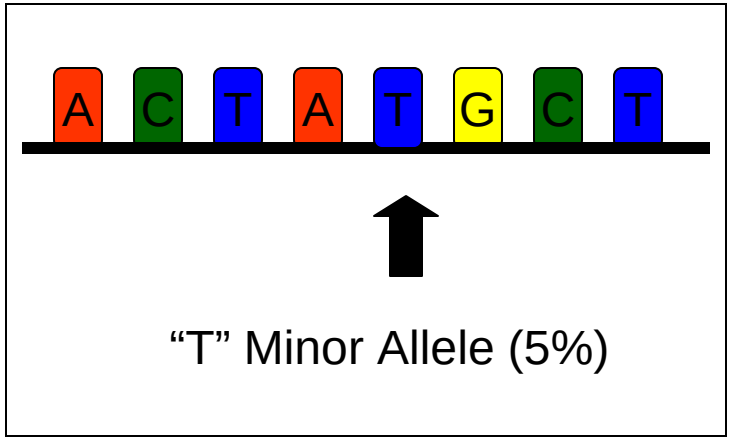
- DNA from a tumor sample and reference sample are labelled differently, the mix is hybridized to normal metaphase chromosomes for CGH or a slide containing thousands of probes for aCGH.
- The color ratio is used to evaluate regions of DNA gain or loss in the subject sample.
- aCGH gives a higher resolution level than CGH

Region < 20 Mb



Single Nucleotide Polymorphism (SNP)

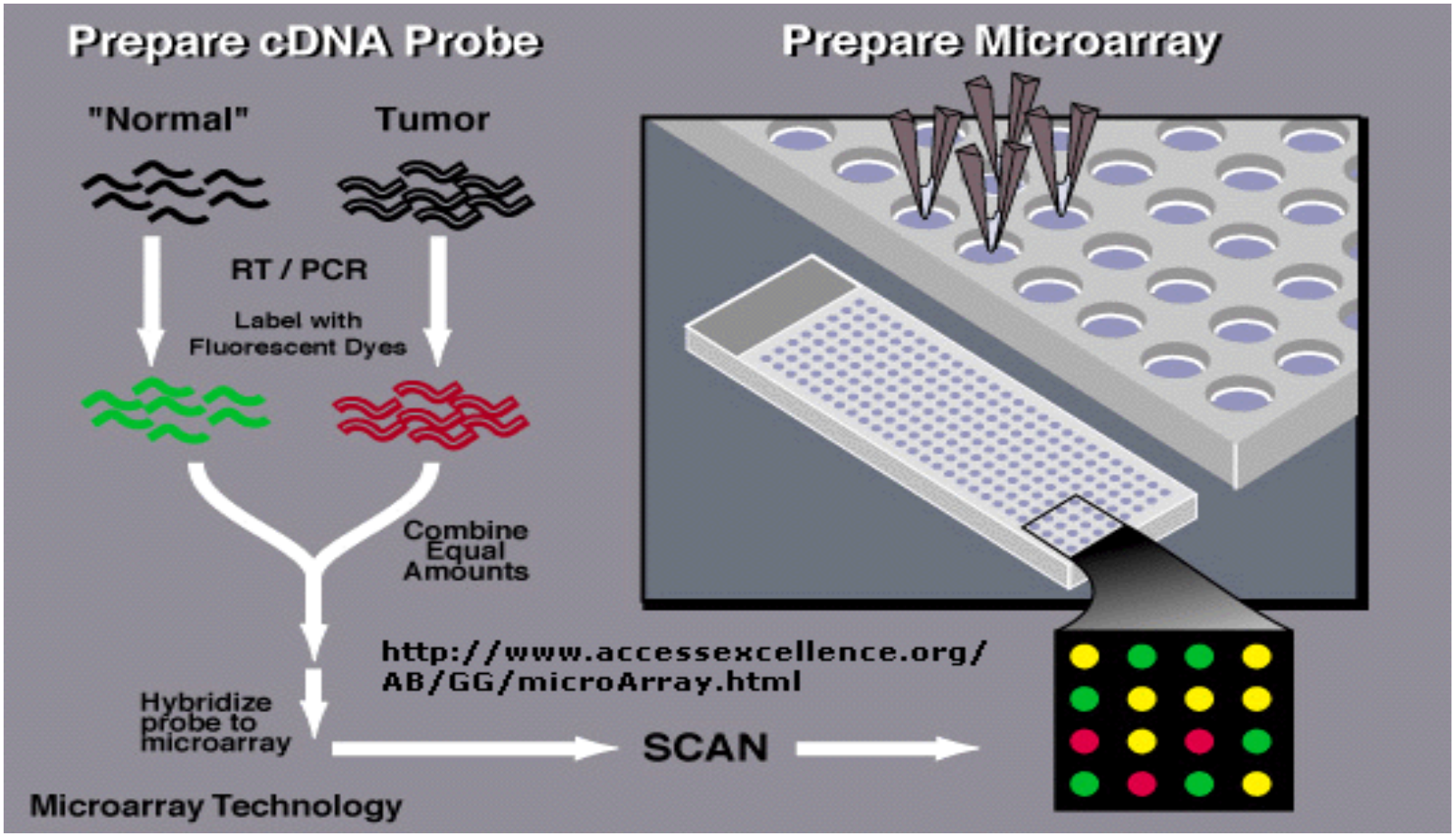




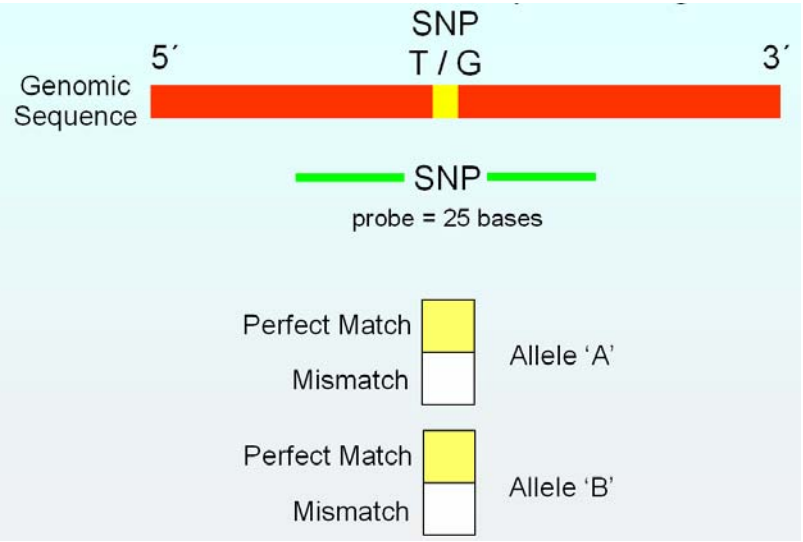
A SNP is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of a population. Almost all common SNPs have only two alleles.

Microarray gene expresion

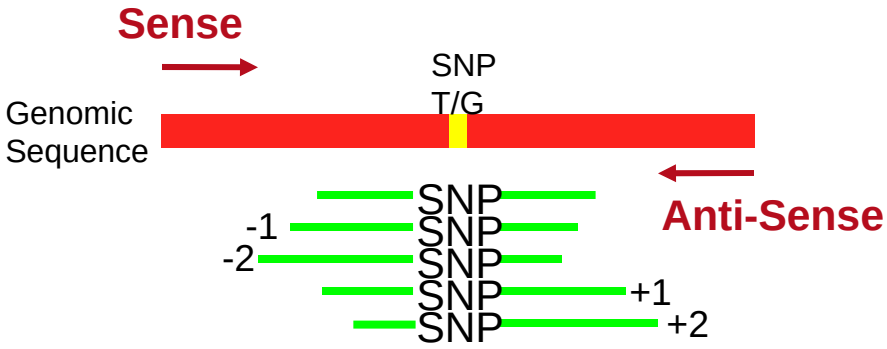
-simultaneously measure abundance of mRNA transcripts for thousands of genes



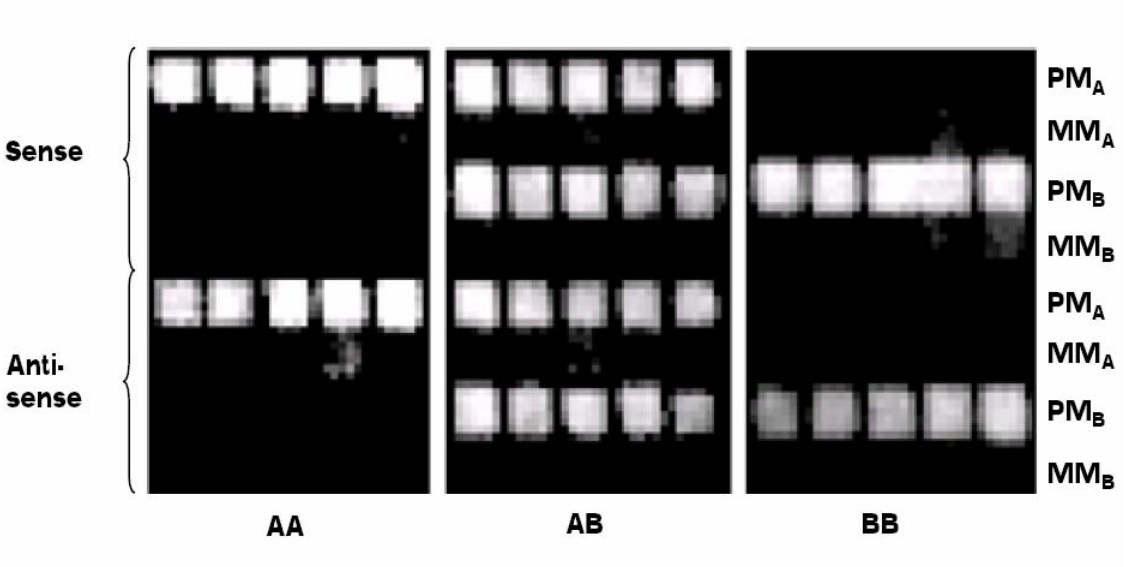
SNP Array Design



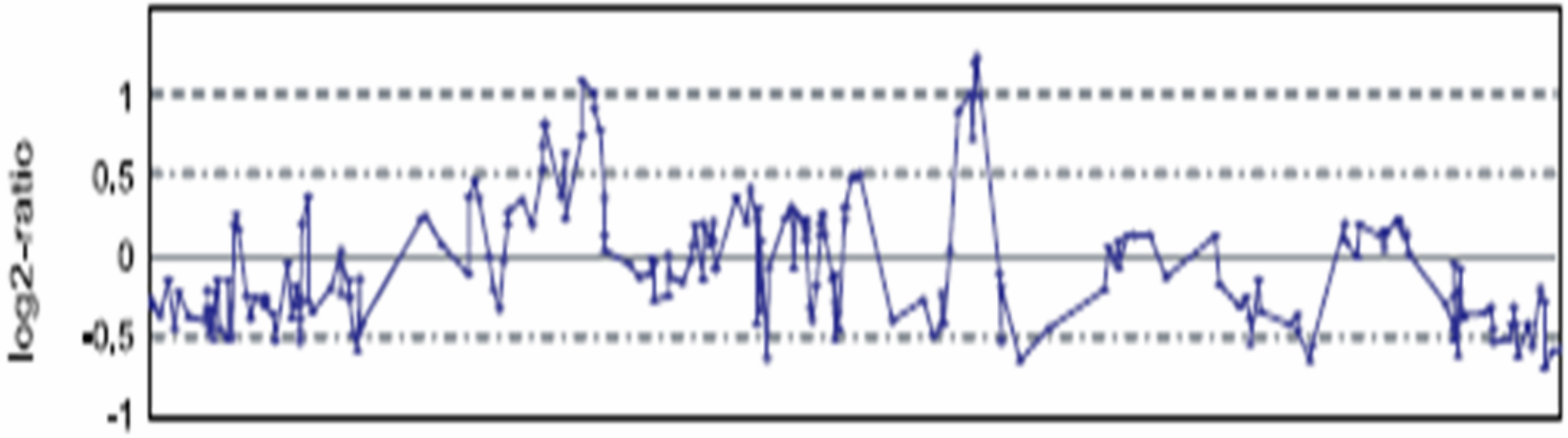
High Probe redundancy



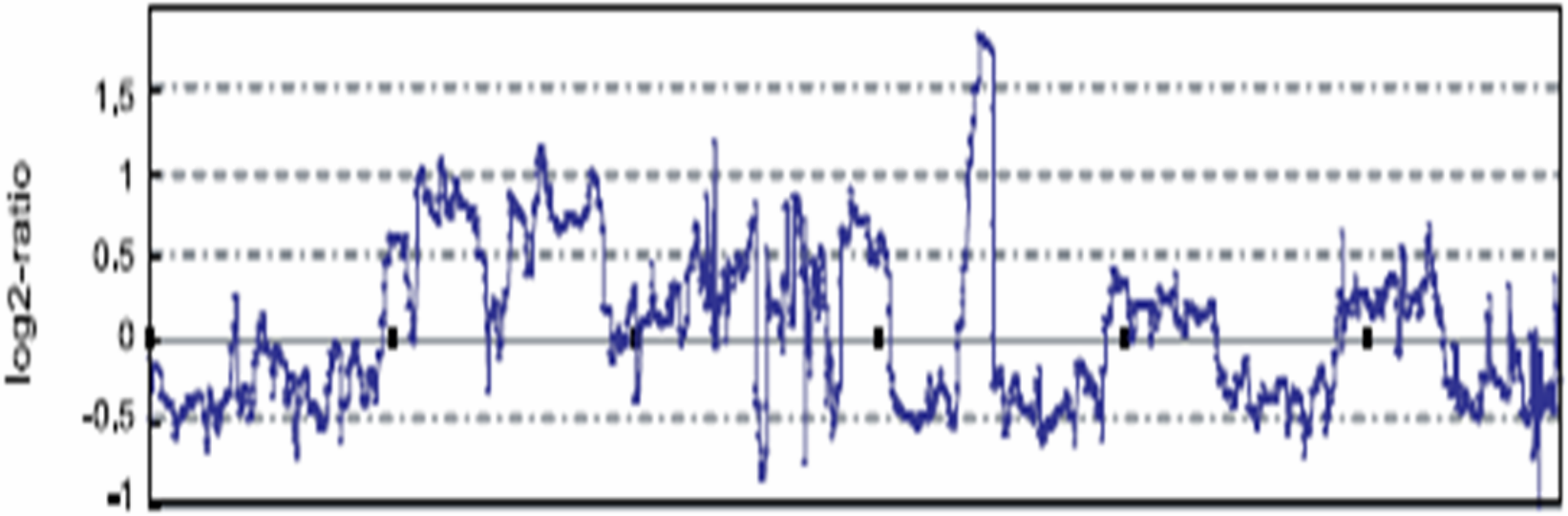
40 probes (PM and MM) per SNP



BAC-CGH



SNP array



Statistical methods to estimate copy number

Summary of Main Approaches

Methods base analysis of copy number on a ratio of test sample to normal reference

- Threshold – Weiss et al. (2003), Pollack et al. (2002)
  test=normal & reference=normal
  normal-normal plus FDR → thresholds for gain/loss
- Normal mixture model – Hodgson et al. (2001)
  3 components: decrease, normal, increase
- Clustering – Autio et al. (2003)
- Hidden Markov Models – Snijders et al. (2003)
- Change points – Linn et al. (2003)
- Circular binary segmentation – Olshen & Venkatraman (2004)

Circular binary segmentation

Olshen and Venkatraman, 2004, Biostatistics, 5:557-572

Split chromosomes into regions of equal CN that accounts for noise in the data – change point ideas.

c = change point if $X_1,\ldots,X_c$ has distribution F and $X_{c+1},\ldots,$ has Distribution G.

X_j = log-ratio of intensities, indexed by marker location

$S_i = X_1 + \ldots + X_i$, i=1,...,n on a given chromosome

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

The likelihood ratio test statistic for:

H0: no change point
H1: exactly one change point at an unknown location i

$$Z_B = \max_{1 \le i < n} |Z_i|$$

Binary segmentation: apply test recursively until no more changes are detected in any of segments obtained from change points already found

$S_i/i - (S_n - S_i)/(n - i)$ = ave of first i X's minus last n-i X's

Split the data in two and compare the two side

cf: Levine's test of homogeneity

Circular binary segmentation is an extension to this idea that detects more than one change point.

Null reference distribution: permutation test

CGHmix – Broet et al (2006)

chromosome k, BAC i, state c

c: 1=loss, 2=normal, 3=gain

L_{ik} = latent state

Z = log-ratio (normalized)

$f(Z \mid L_{ik} = c) \sim N(\mu_c, \sigma_c)$

- Marginal of Z: $f(Z_{i,k}) = \sum_{c=1}^{3} \omega_{c,i,k} \times \phi(\cdot \mid \mu_c, \sigma_c^2)$
- $\omega_{c,i,k}$: probability that BAC i of chromosome k is in state c

The weights satisfy

$$0 \leq \omega_{c,i,k} \leq 1 \quad \text{and} \quad \sum_{c=1}^{3} \omega_{c,i,k} = 1$$

Gaussian Markov random fields

conditional auto-regression

Suppose $X=(X_1,\ldots X_n)^T$ has density

$$p(x) \propto e^{-\frac{1}{2}x^T Q x}, \; x \in R^n$$

where Q is positive definite symmetric matrix

Intrinsic conditional auto-regression (improper)

Notice

$$x^T Q x \equiv \sum_i Q_{i+} x_i^2 - \sum_{i<j} Q_{ij} (x_i - x_j)^2$$

Limiting form of conditional auto-regression where Q is similar but Q1=0. (positive semi-definite)

$$p(x) \propto e^{\frac{1}{2}\sum_{i<j} Q_{ij} (x_i - x_j)^2}$$

CGHmix – Broet et al (2006)

Spatial structure

$$\omega_{c,i,k} = \frac{\exp(x_{c,i,k})}{\sum_{l=1}^{3} \exp(x_{l,i,k})}$$

where *c* = state, *i* = snp, *k* = chromosome

$x_{c,k} = \{x_{c,i,k}; 1 \le i \le n_k\}$ are three independent latent first-order Markov random fields, each distributed according to the intrinsic Gaussian conditional auto-regression model

$$x_{c,i,k} \mid x_{c,(-i),k} \sim N\left(\frac{1}{m_{i,k}} \sum_{l \in \delta_{i,k}} x_{c,l,k}; \frac{\tau^2_{c,k}}{m_{i,k}}\right)$$

Dchip – Li & Wong (2001)

Multiplicative probe effect Model

- Assume

$$MM_{ij} = \nu_j + \theta_i \alpha_j + \varepsilon$$

$$PM_{ij} = \nu_j + \theta_i \alpha_j + \theta_i \phi_j + \varepsilon$$

- Model

$$y_{ij} = PM_{ij} - MM_{ij} = \theta_i \phi_j + \varepsilon_{ij}$$

where θ_i is the copy number index of SNP i, α_j, ϕ_j are effects of jth probe, υ_j is the baseline response of jth probe pair.

Dchip – Li & Wong (2001)

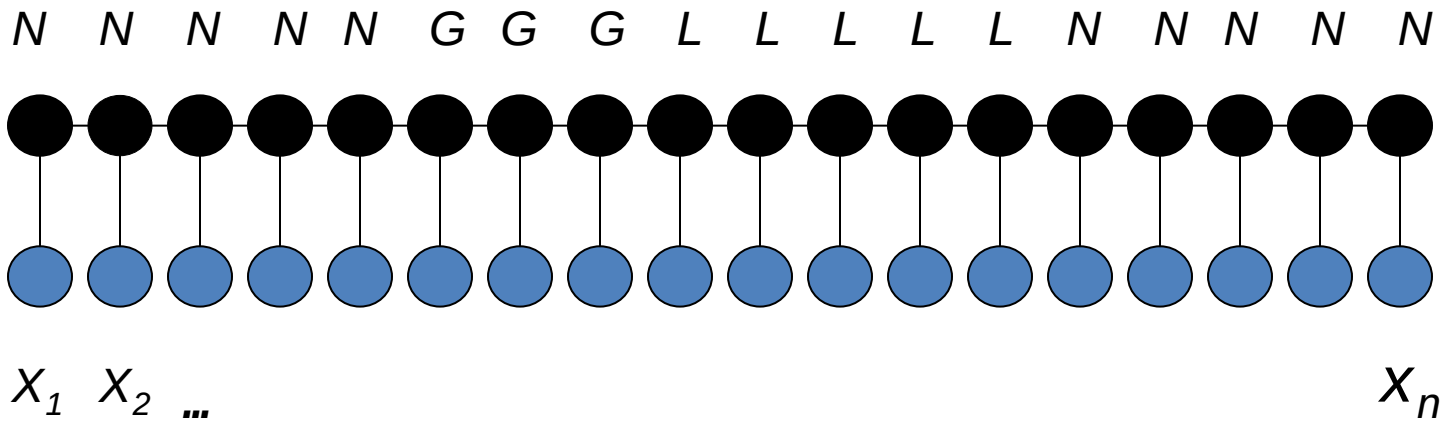
Raw copy number estimation

Use θ_i as copy number index for SNP i

- Mean signal of 2 copy for each SNP:
  - average of copy number index of all normal sample

- Raw copy number
  - (observed signal/mean signal of 2 copy) *2

Recall the HMM



- The shaded nodes represent the hidden states (L/N/G).
- X_1, …, X_n are the observed values (raw copy number).
- **Initial probabilities** : probability that the starting observation comes from state k. π_k
- **Emission probabilities** : probability that o is emitted from state j. $b_j(o)$
- **Transition probabilities** : probability from state i to j. a_{ij}
- **Viterbi algorithm** gives the most likely underlying path.

Dchip – Li & Wong (2001)

Infer copy number

- Initial probabilities:

  0.9 for 2 copy, 0.1/(N-1) for others.

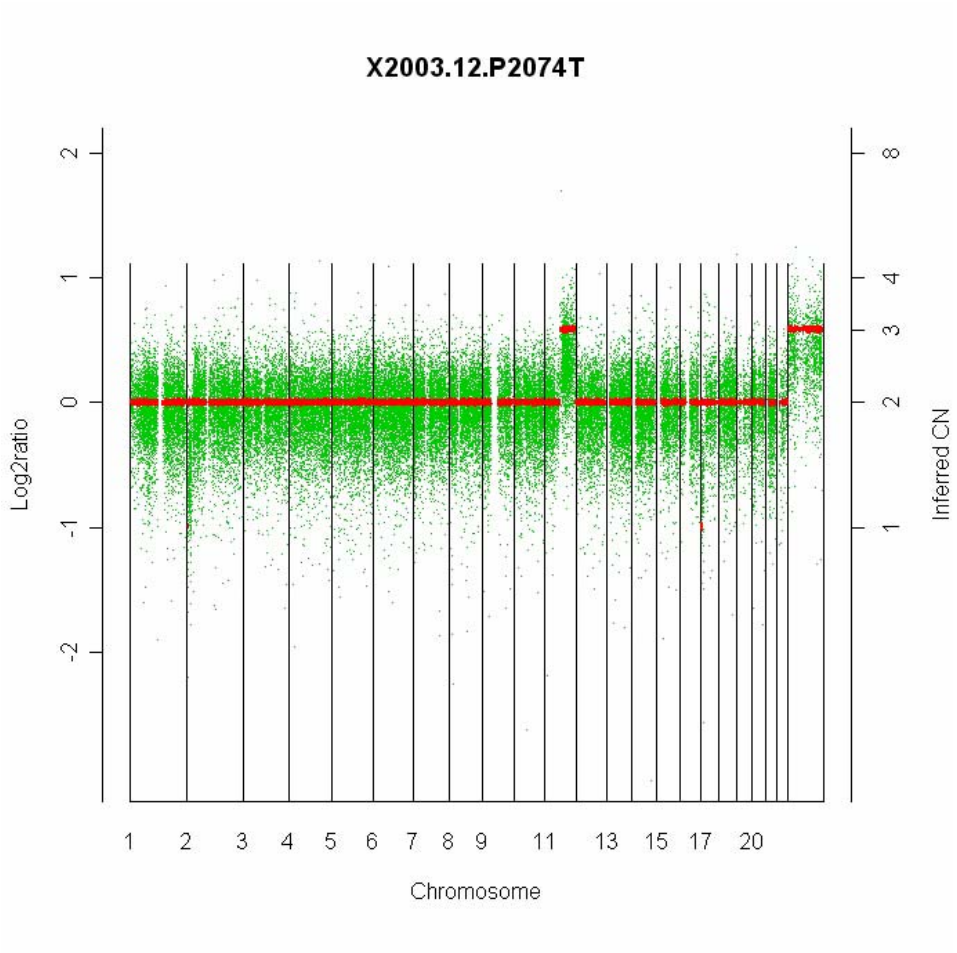
- Emission probabilities
  (signal – true cn)/sd ~ t(40)

- Transition probabilities

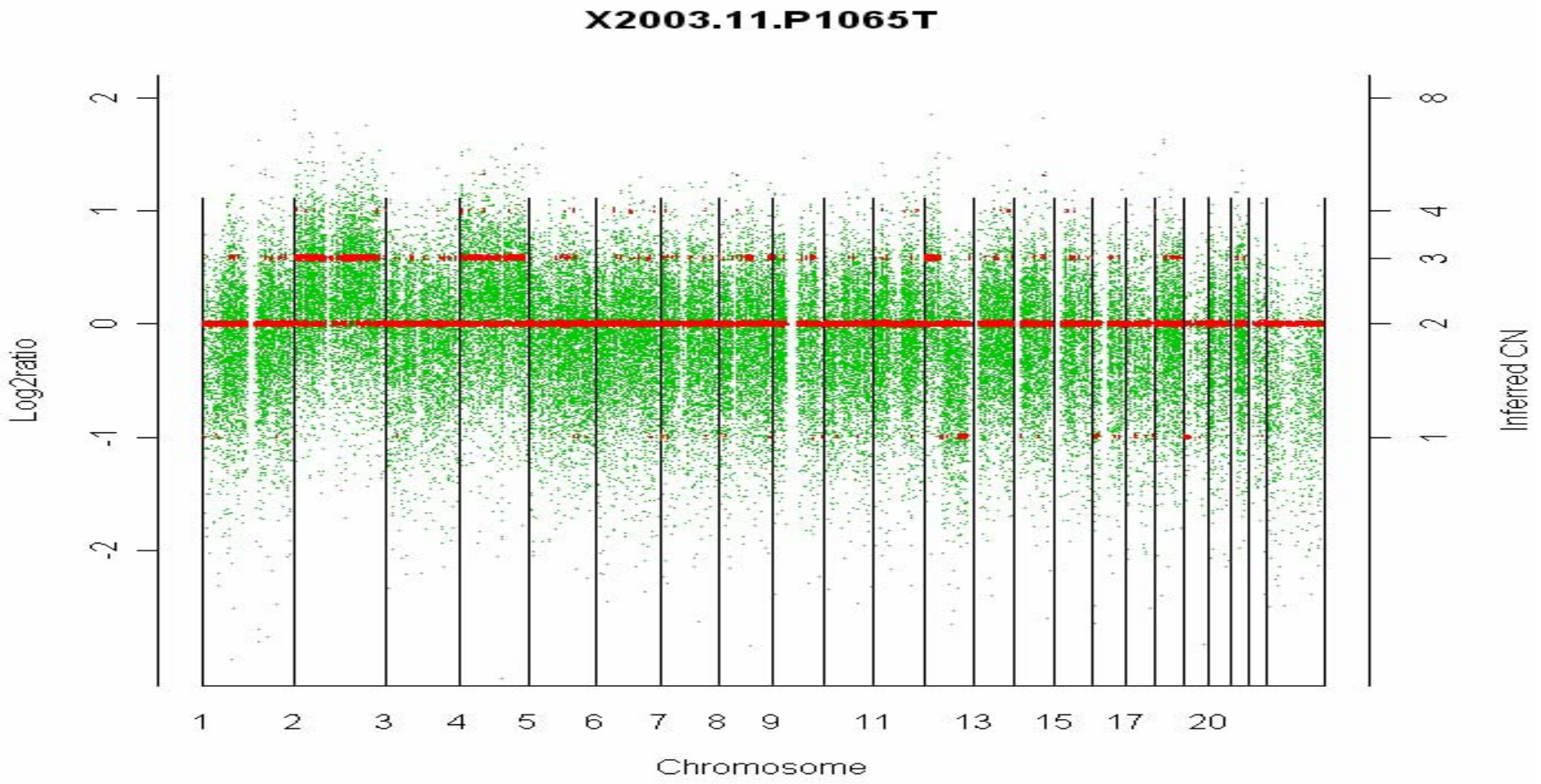
  Use Haldane's map function $\theta = \frac{1}{2}(1 - e^{-2d})$

  convert the genetic distance d between two SNPs to the probability that the copy number will return to the background distribution.

A good case from dchip



A not good case from dchip



CNAG – Nannya et al (2005)

$$\theta_i = {}^c\theta_i + \sum_{j=1}^{2} (a_j + b_j x_j + c_j x_j^2)$$

- θ_i is the logratio of ith SNP in sample 1 and 2
- ${}^c\theta_i$ represents the corrected logratio
- x_1 and x_2 represent length and GC content of the fragment that contains the SNP.

Determine ${}^c\theta_i$ by a series of linear regressions.

${}^c\theta_i$ shows a lower SD than θ_i

CNAG – Nannya et al (2005)

- For a given sample , the averaged best-fit m references is calculated as

$$S_{i,m}^{REF}$$

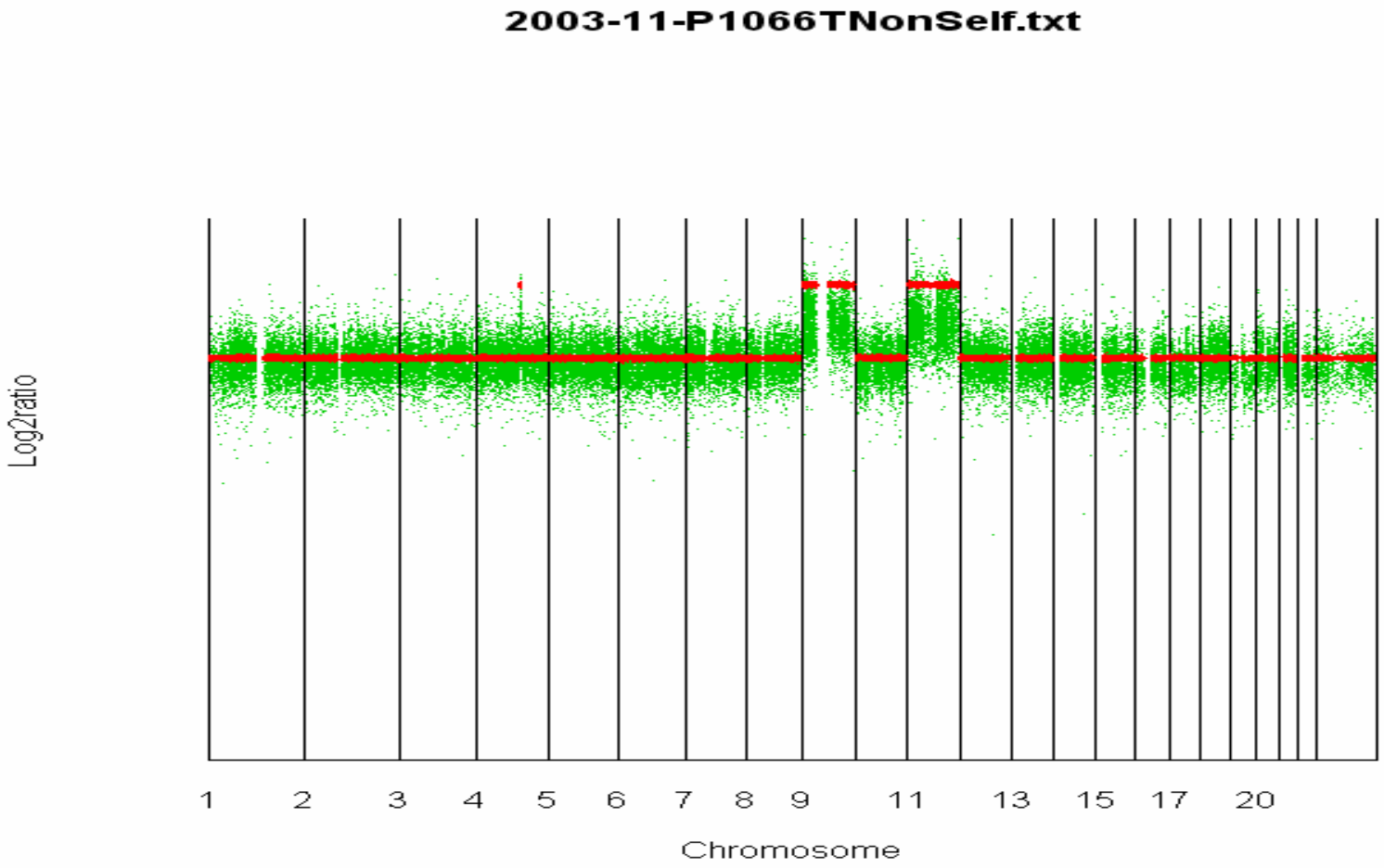
$$\frac{1}{m}\sum_{j} 2^{{}^{c}\theta_{i}^{j,\zeta}} \times S_{i}^{\zeta}$$

where j(1,2,…m) represents the m reference samples in which the SD of takes the lowest values.

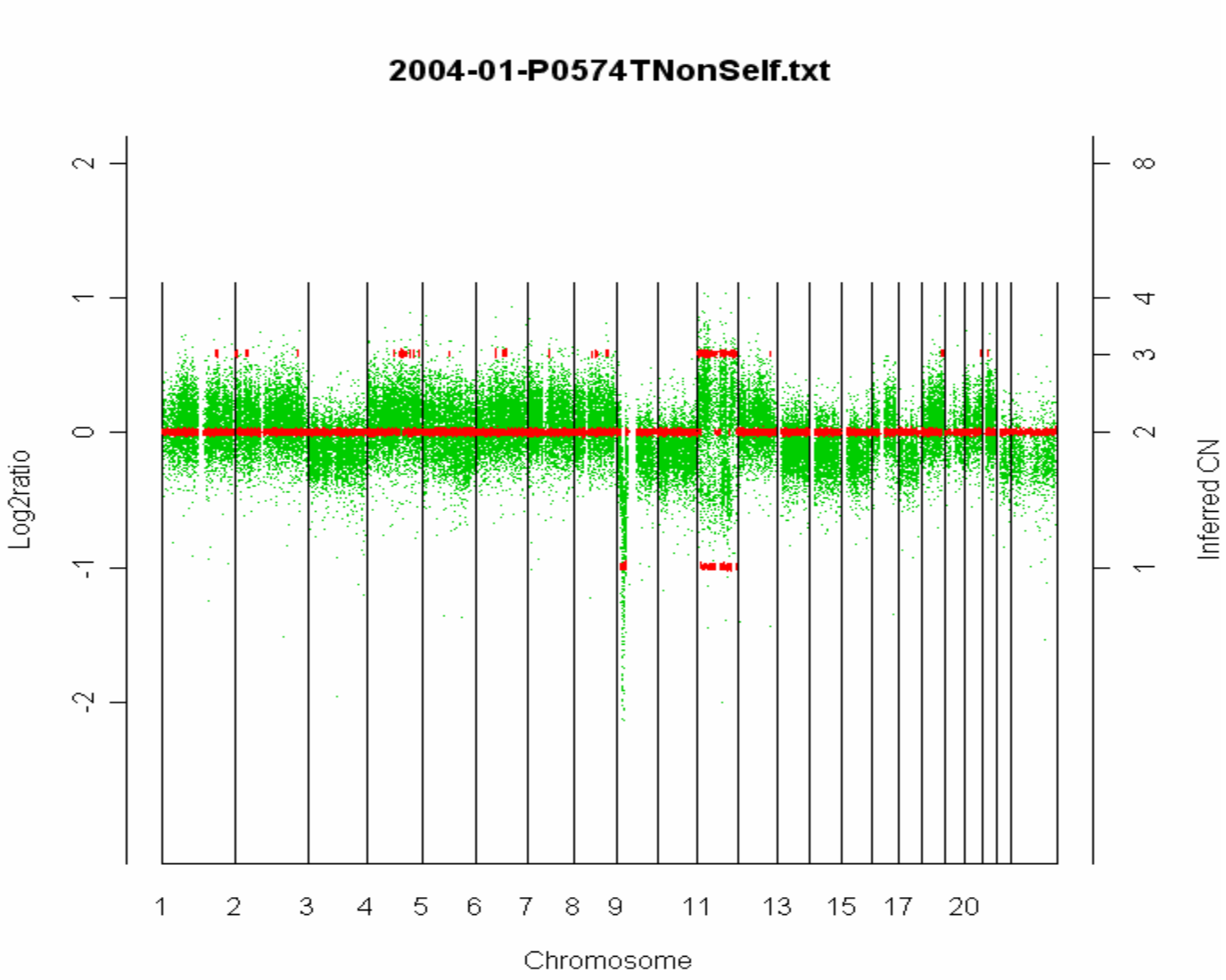
$${}^{c}\theta_{i}^{\zeta,j}$$

- Use HMM to infer integer copy numbers.

A good case from CNAG



A not good case from CNAG



Comparison

Advantages

- CGHmix: incorporate spatial dependence in mixture; use information of the whole genome
- Dchip: consider probe effect
- CNAG: improve the logratios by accounting for PCR products and use an optimized reference.

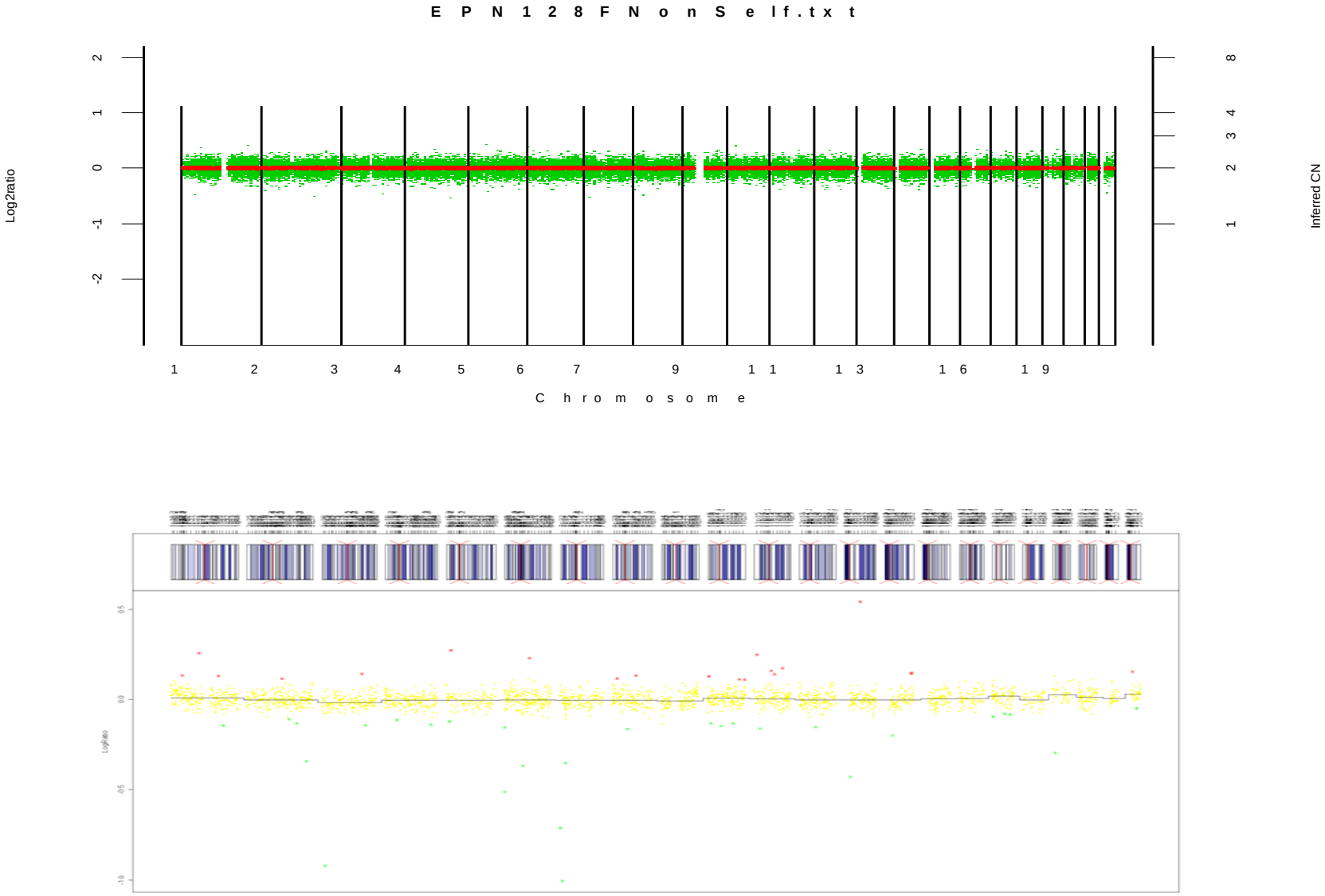
Disadvantages

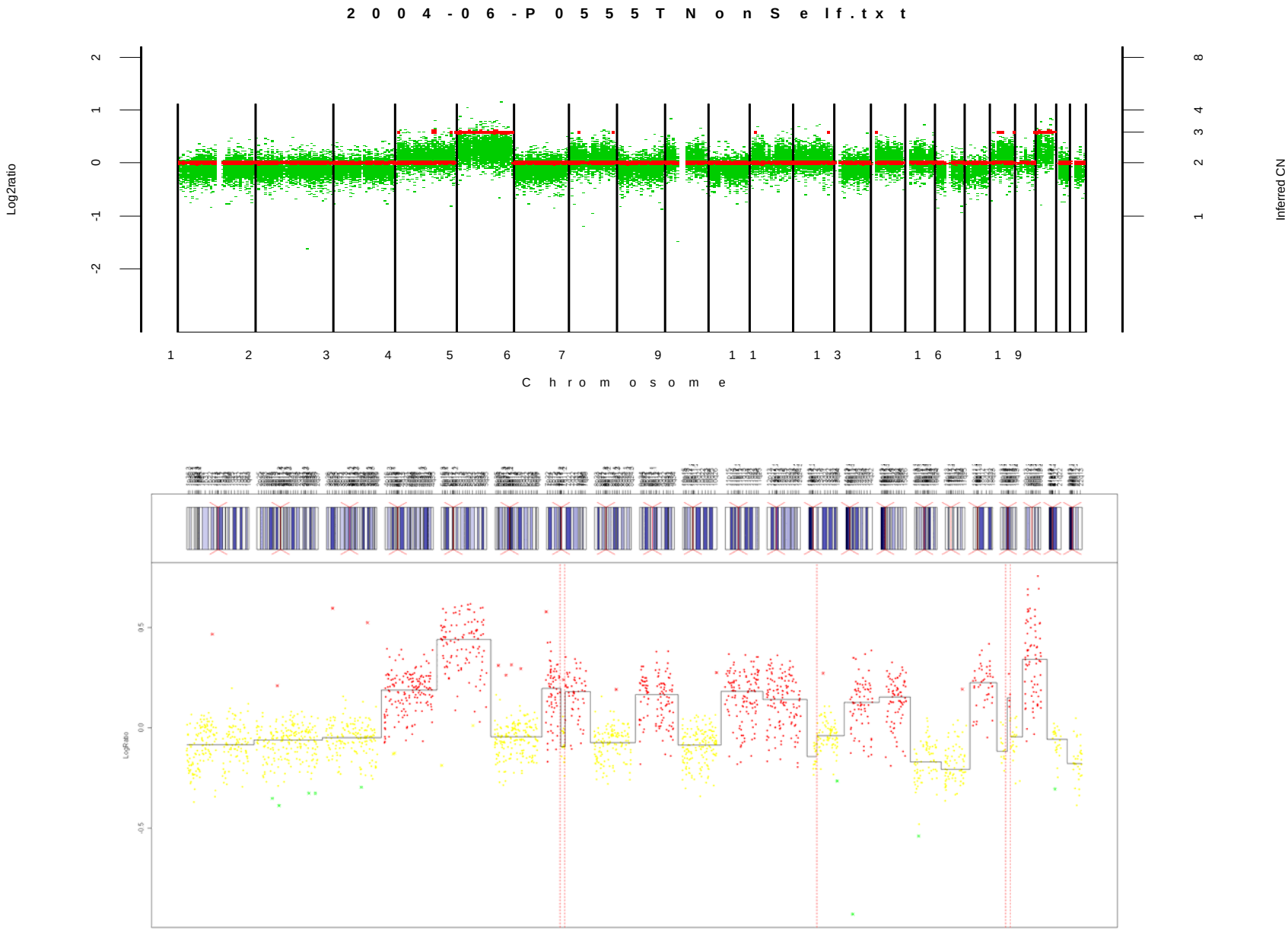
- CGHmix: cannot deal with cases where there is only one group.
- Dchip: does not work well for high density arrays.
- Dchip and CNAG: apply the model to one chromosome at a time, thus doesn't consider neighboring information; Viterbi for HMM also has some drawbacks.

Preliminary work

Copy number estimation
- from CGH to SNP

- CGH is the gold standard to estimate copy numbers.
- SNP arrays have higher resolution than CGH so hopefully they can detect smaller copy number segments.
- Comparing copy number estimates from the two platforms: Mostly concordant, but some discrepancies. The main reason is that CGH algorithm consider whole genome.


EPN128FNonSelf.txt
Log2ratio
2
1
0
-1
-2
Inferred CN
8
4
3
2
1
1
2
3
4
5
6
7
9
11
13
16
19
Chromosome
LogRatio
0.5
0.0
-0.5
-1.0


2004-06-P05555TNonSelf.txt
Log2ratio
Inferred CN
Chromosome
LogRatio

Concordance between SNP and BAC copy number

	583	602	555	574	648	695	1065	1066	2074	128	208	210	409	522	549	565
chr1	1	1	0.994	0.984	1	1	1	1	1	1	0.543	0	0	1	1	1
chr2	1	1	1	0.885	1	1	1	1	0.875	1	1	1	1	1	1	1
chr3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
chr4	1	1	0	0.785	1	1	0.185	0.95	0.95	1	1	1	0.981	1	1	1
chr5	1	1	0.915	1	1	1	0.996	1	1	1	0.921	1	1	1	1	1
chr6	1	1	1	0.889	1	1	1	1	1	1	1	1	1	1	1	1
chr7	1	1	0.062	0	1	1	0	1	1	1	1	1	0	1	1	1
chr8	1	1	1	0.956	1	1	0.939	1	1	1	1	1	1	1	1	1
chr9	1	1	0	0.782	1	1	0.993	0.993	1	1	0.995	1	0	0.659	1	1
chr10	1	1	1	0.003	1	1	1	1	1	1	0.996	1	1	1	1	1
chr11	1	0.052	0.047	0.07	1	1	1	0.979	0.42	1	1	1	1	1	1	1
chr12	1	1	0	1	1	1	0.02	1	1	1	1	1	0.99	1	1	1
chr13	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
chr14	1	1	1	1	1	1	1	1	1	1	1	1	0.982	1	1	1
chr15	0.995	0.996	0	0.008	0.996	0.996	0.996	0.996	0.996	1	0.996	0.996	0.996	0.996	1	0.996
chr16	1	1	1	1	1	1	0	0	1	1	1	0	0	1	0.95	1
chr17	1	1	1	0.068	1	1	1	0	0.507	1	1	0	0.082	1	1	1
chr18	1	1	0.143	0	1	1	1	0.994	1	1	1	1	1	1	1	1
chr19	1	1	0.692	0	1	0.923	0	0	1	1	1	0.077	0.077	0	1	1
chr20	1	1	1	0.977	1	1	0	1	1	1	1	0	1	1	1	1
chr21	1	1	1	0.969	1	1	1	1	1	1	1	1	1	1	1	1
chr22	0.974	0.974	0.974	0.026	0.974	0.974	0.026	0.026	0.974	0.97	0.974	0.026	0.026	0.974	0.97	0.974

Proposed work

SNP based mixture models for copy number

Basic mixture model for iid data

$$f(y_j) = \sum_{i=1}^{3} \pi_i \phi(y_j; \mu_i, \Sigma_i)$$

- Act as marginal distribution for models involving discrete latent variables, e.g, clustering.
- Capture many properties of real data such as multimodality, skewness and unobserved heterogeneity.

Mixture model for copy number

$$f(x) = p_L N(x; \mu_L, \sigma_L^2) + p_N N(x; \mu_N, \sigma_N^2) + p_G N(x; \mu_G, \sigma_G^2)$$

p_L, p_N, p_G are proportions of loss, normal and gain, and they sum to one.

mixture model for iid sample

Standard fittings:

- EM algorithm
  - E step
  - M step
- Bayesian approach
  - Prior
  - Gibbs sampling

Mixture model for SNP data

- Mixture models have been used to CGH, but have not been applied to SNP data yet.

- Cannot just apply the exact model for CGH since there exist differences in the two platforms:
  - resolution
  - probes used to interrogate the same SNP

SNP mixture model development

$$f(Z_{i,k}) = \sum_{c=1}^{3} \omega_{c,i,k} \times \phi(\cdot \mid \mu_c, \sigma_c^2)$$

Introduce spatial correlation via Markov random field.

where

$$\omega_{c,i,k} = \frac{\exp(x_{c,i,k})}{\sum_{l=1}^{3} \exp(x_{l,i,k})}$$

$$x_{c,i,k} \mid x_{c,(-i),k} \sim N\left(\frac{1}{m_{i,k}} \sum_{l \in \delta_{i,k}} x_{c,l,k}; \frac{\tau_{c,k}^2}{m_{i,k}}\right)$$

SNP mixture model development

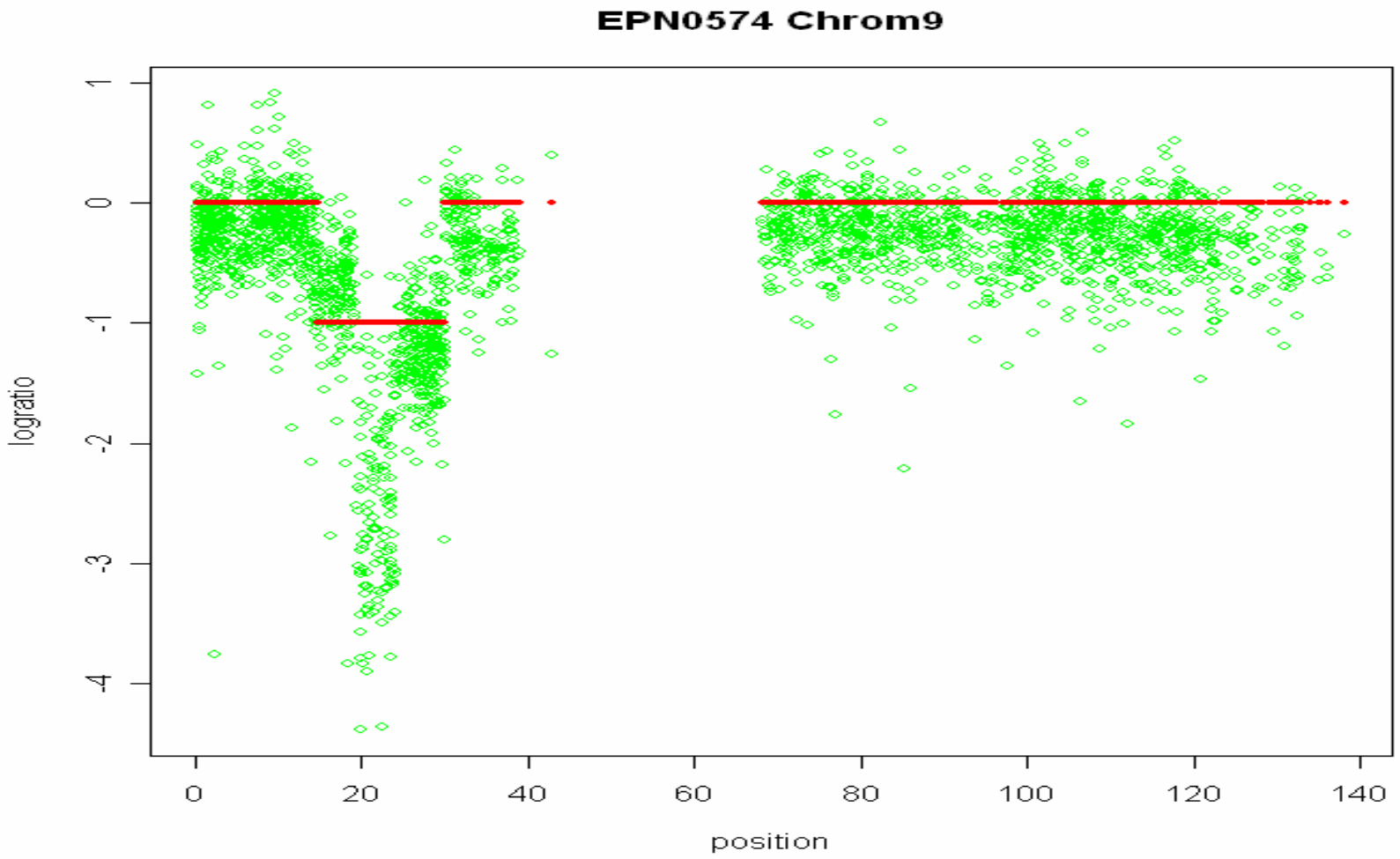
- Model computation: MCMC in WinBugs

- Allocation rule: assign a SNP to a modified state if its corresponding posterior probability is above a threshold value p_{cut}, otherwise it is assigned to the normal copy state.

Future work

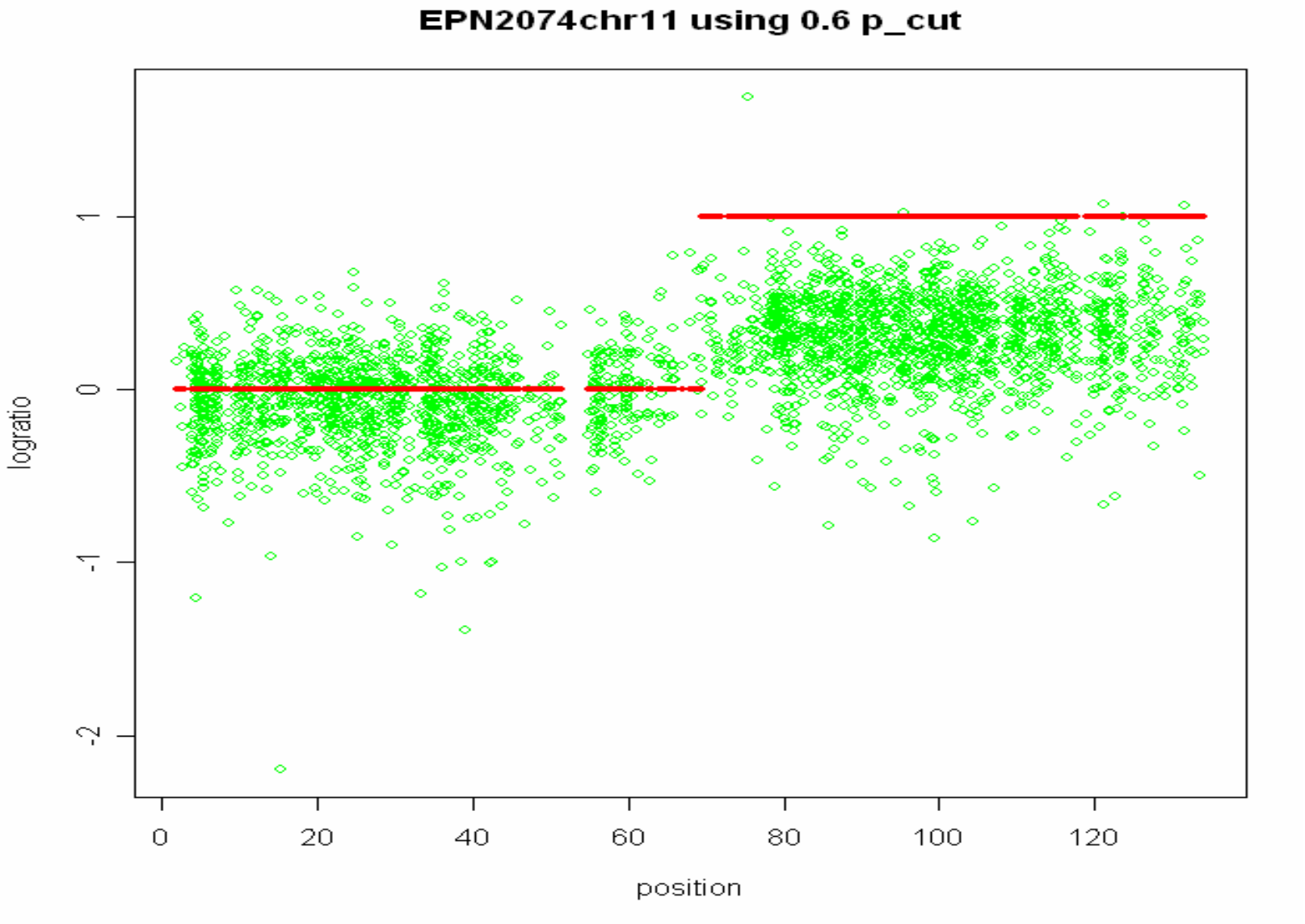
- In Markov random field, consider a window of more than two neighbors.
- Instead of assigning equal weights to all neighbors, incorporate the distance between SNPs to determine weights.
  - Physical distance
  - Genetic distance
- Determine the best window size

Mixture with 5 nbrs each side



Fit the mixture model to the raw logratios from dchip

Another case



Mixture with other than five neighbors each side

- One neighbor each side: MCMC converges slow.
- 10 or 20 neighbors each side: over-smoothing, need to reduce p_{cut} to get reasonable results.

Future work

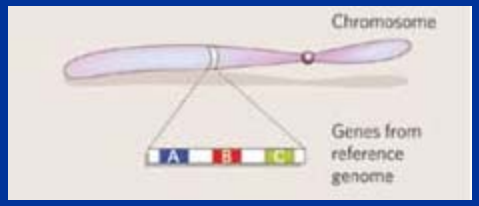
Instead of use standard MCMC in which the object of inference has a fixed dimension, use reversible jump MCMC which can jump between parameter subspaces of differing dimensionality.

Other possible work

- Model the whole genome instead of chromosomewise.
- Do simulation study to determine p_cut
- Model probe intensities directly in mixture model

What is CNV?

- A segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome [Feuk et al. 2006]
- Categories



Deletion

Duplication

Insertion

Images are made from http://www.ghastlyfop.com/blog/uploaded_images/smaller-737680.jpg

Why are we interested in identifying CNVs?

- Understand the genetic variation among the normal population
- Help to identify genetic predisposition towards cancer
- Distinguish normal copy number variation from aberrant genetic lesions in cancer

How did we identify CNVs?

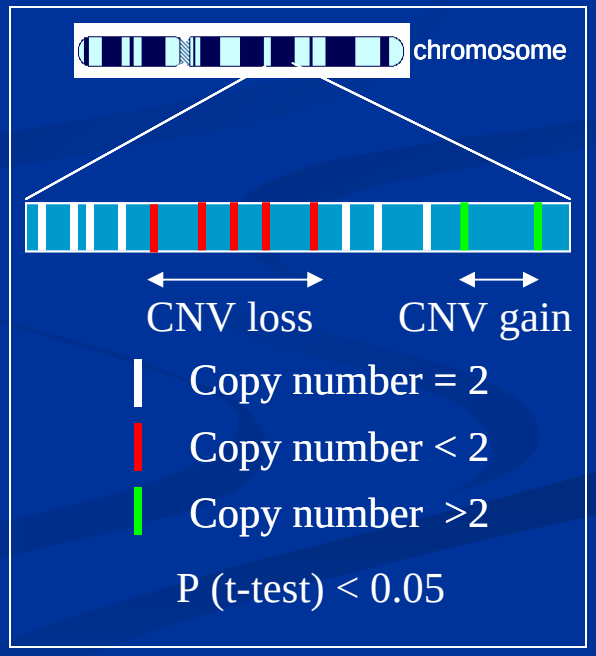
50K XbaI array data from 104 normal individuals (Asian, European, African American)

Copy number analysis by dChip
[Lin et al. 2004]

XbaI SNP CN

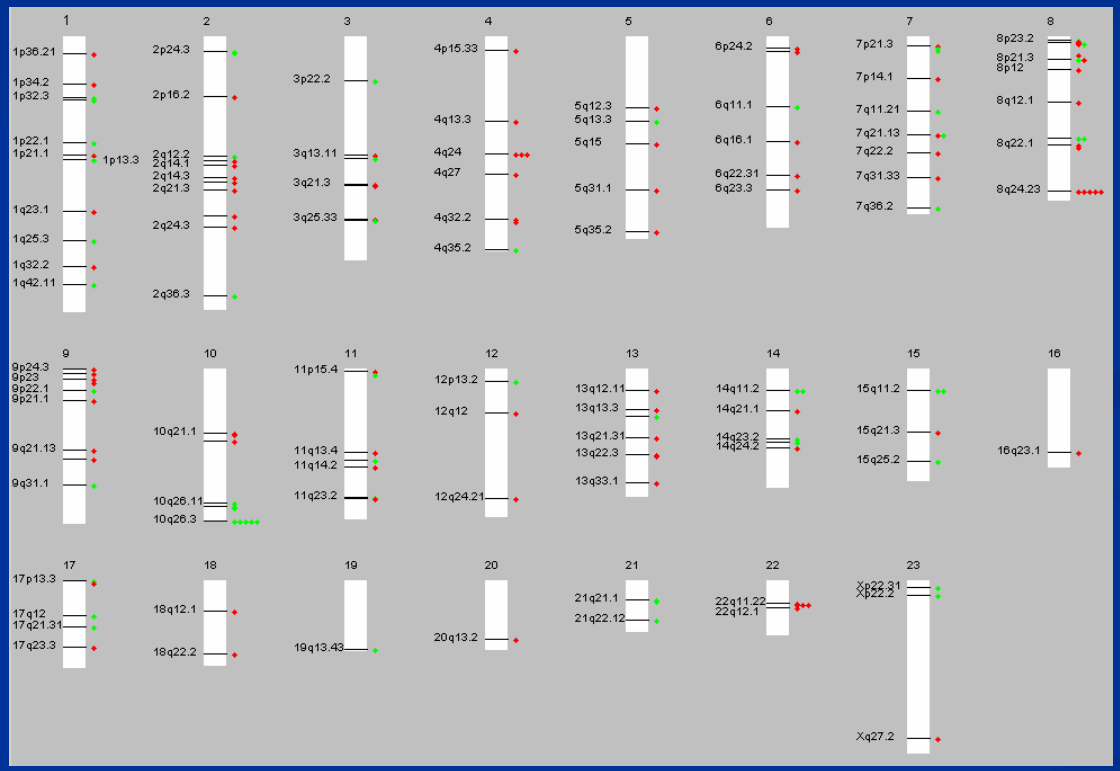
Identification of CNVs

Validation by q-PCR



Chromosome image is from: http://wimp.nsm.uh.edu/chromo8.gif

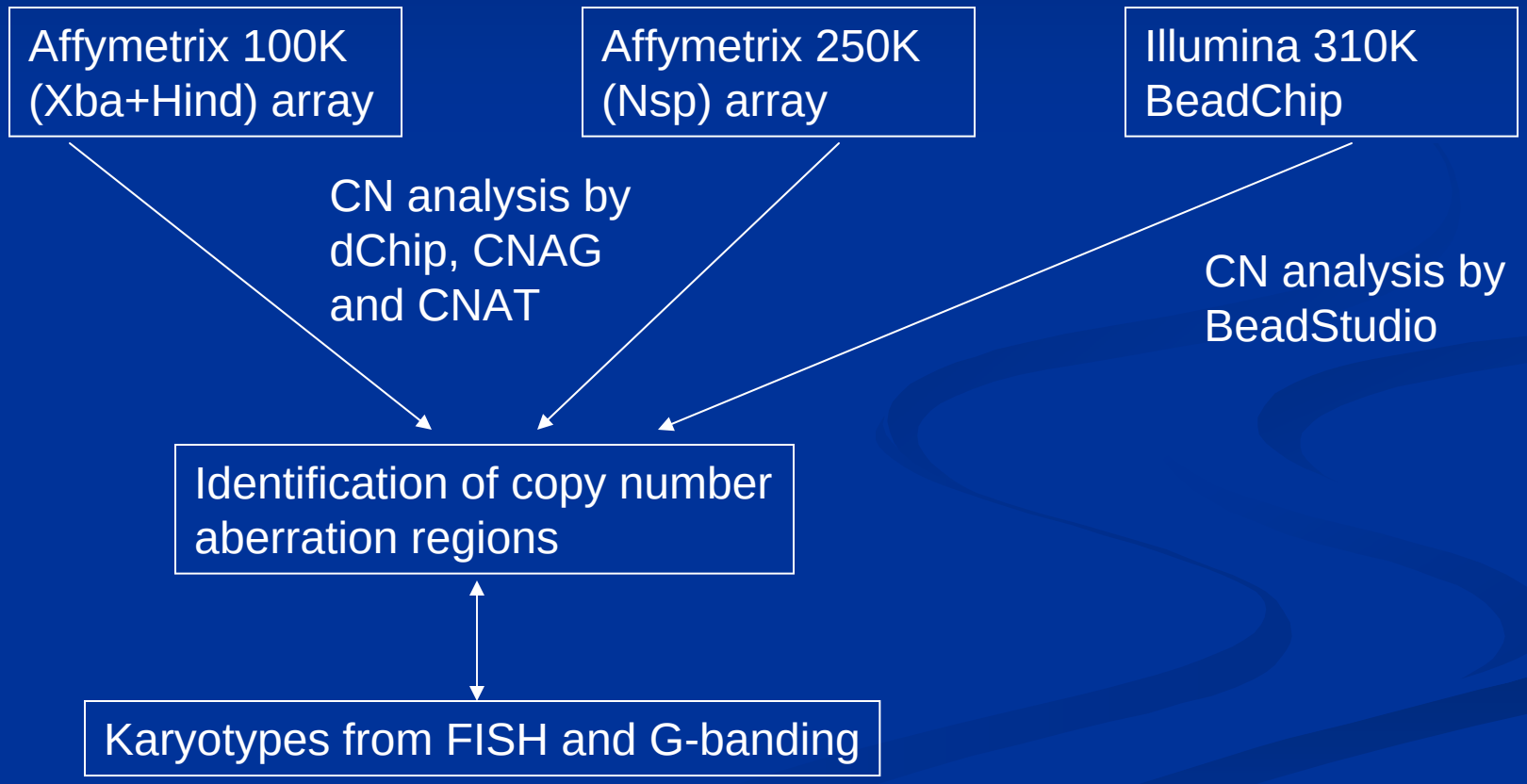
Whole-genome view of CNVs



- 143 CNV regions
  - 70% CNVs are novel
  - 37% gain
  - 63% loss
  - Min: 68 bp
  - Max: 18 Mb
  - Median: 86 Kb
  - Covers > 60 Mb of the genome

Aneuploidy project

Copy number analysis on fixed aneuploidy samples by SNP array



FISH: fluorescent in-situ hybridization

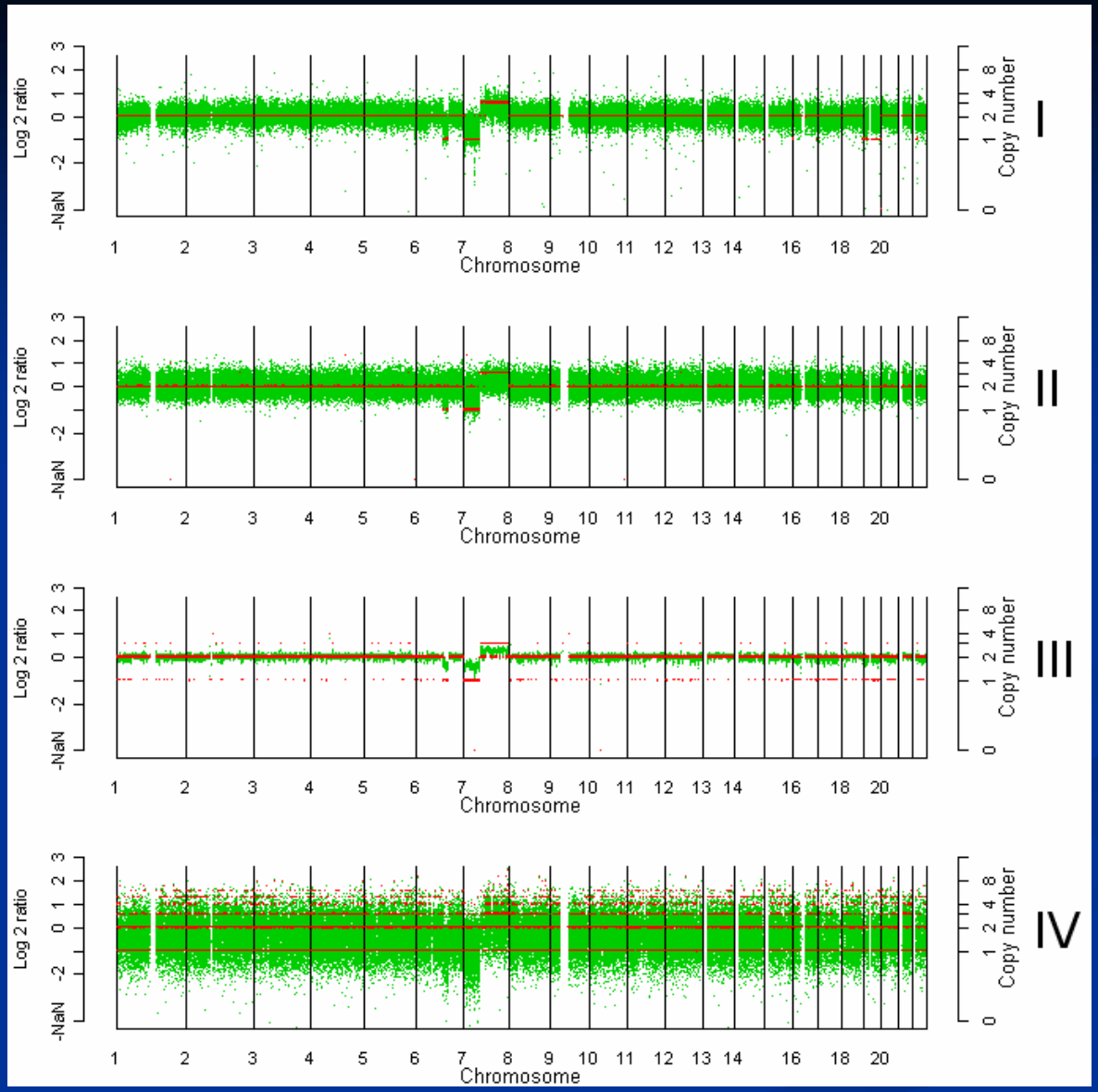


Fig. 1A: Sample 688 (47, XY: +X, del(6)(q12q21), i(7)(q10),del(9)(p21)) copy number of SNPs on 22 autosomes.
I: Illumina 310K data analyzed by BeadStudio; II: Affymetrix 250K Nsp data analyzed by CNAG2.0
III: Affymetrix 250K Nsp data analyzed by CNAT4.0.1; IV: Affymetrix 250K Nsp data analyzed by dChip

Log2 ratio Copy number

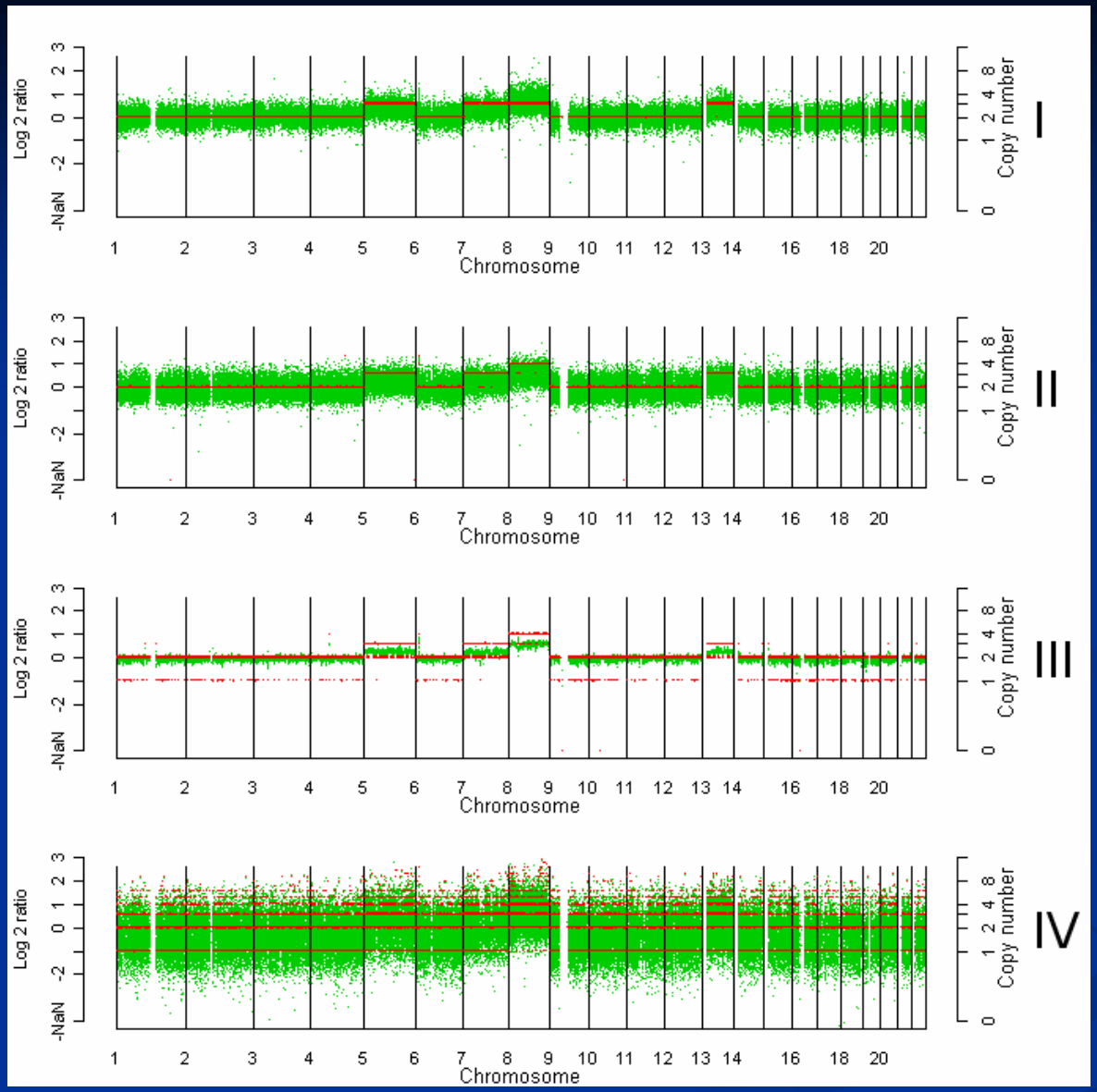


Fig. 1B: Sample 52 (51 XY, +5, +7, +8, +8, +13) copy number of SNPs on 22 autosomes.
I: Illumina 310K data analyzed by BeadStudio; II: Affymetrix 250K Nsp data analyzed by CNAG2.0
III: Affymetrix 250K Nsp data analyzed by CNAT4.0.1; IV: Affymetrix 250K Nsp data analyzed by dChip

Log2 ratio Copy number

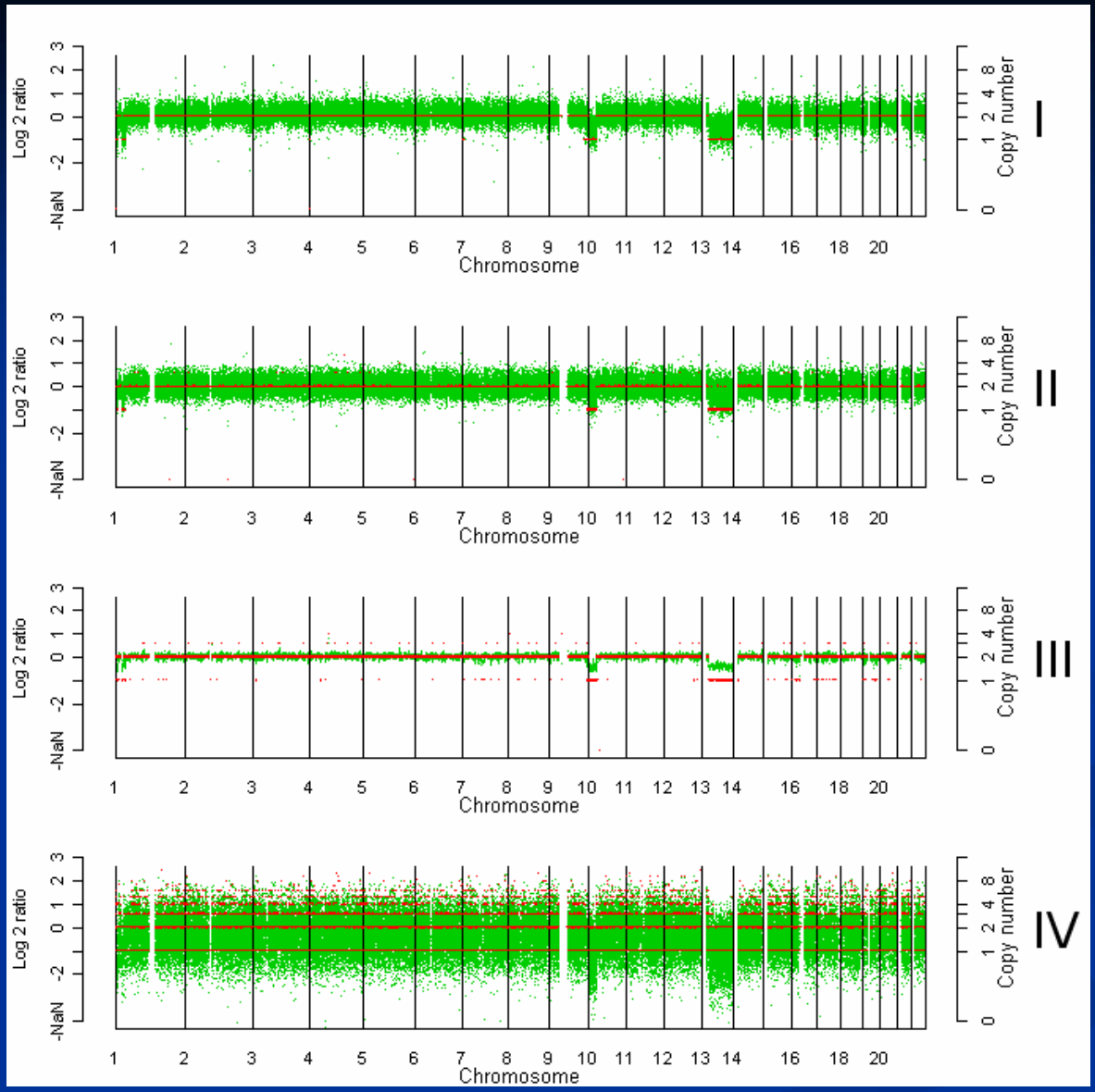


Fig. 1C: Sample 406 (45 XX, -13) copy number of SNPs on 22 autosomes.
I: Illumina 310K data analyzed by BeadStudio; II: Affymetrix 250K Nsp data analyzed by CNAG2.0
III: Affymetrix 250K Nsp data analyzed by CNAT4.0.1; IV: Affymetrix 250K Nsp data analyzed by dChip

Log2 ratio Copy number

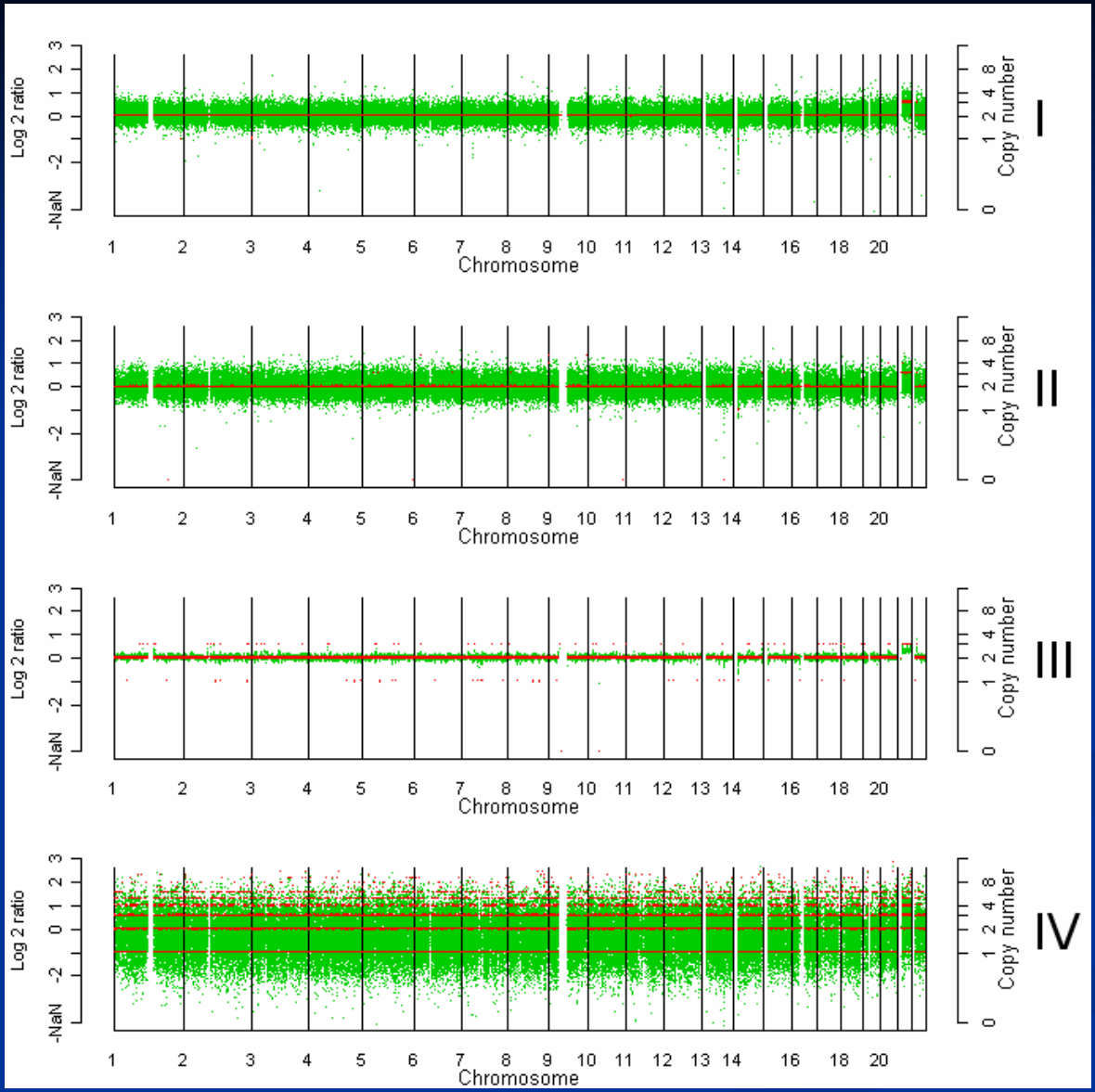


Fig. 1D: Sample 282 (47 XY, +21) copy number of SNPs on 22 autosomes.
I: Illumina 310K data analyzed by BeadStudio; II: Affymetrix 250K Nsp data analyzed by CNAG2.0
III: Affymetrix 250K Nsp data analyzed by CNAT4.0.1; IV: Affymetrix 250K Nsp data analyzed by dChip

Log2 ratio Copy number

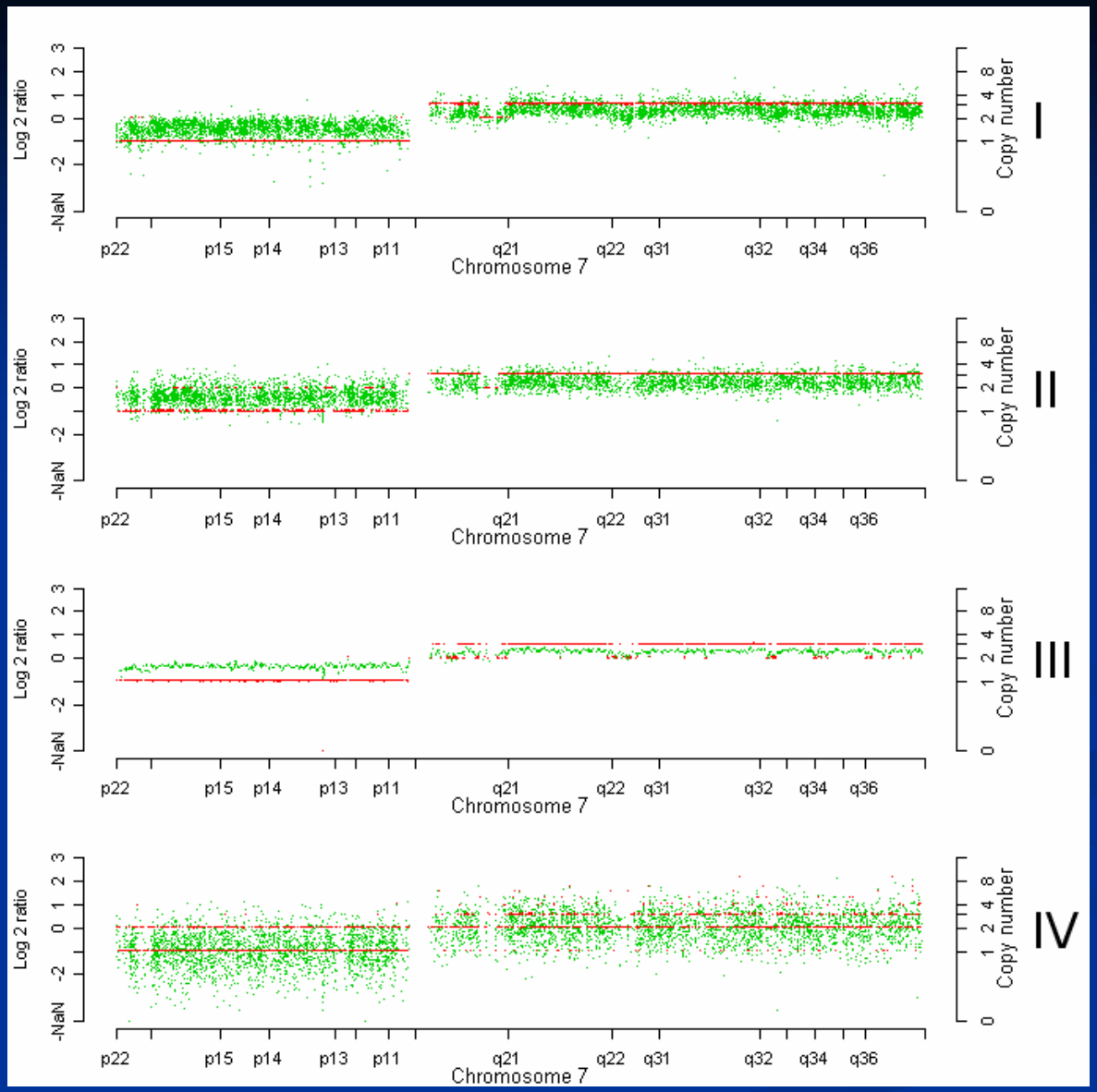


Fig. 2: Sample 688 copy number of SNPs on chromosome 7.
I: Illumina 310K data analyzed by BeadStudio; II: Affymetrix 250K Nsp data analyzed by CNAG2.0
III: Affymetrix 250K Nsp data analyzed by CNAT4.0.1; IV: Affymetrix 250K Nsp data analyzed by dChip

Log2 ratio Copy number

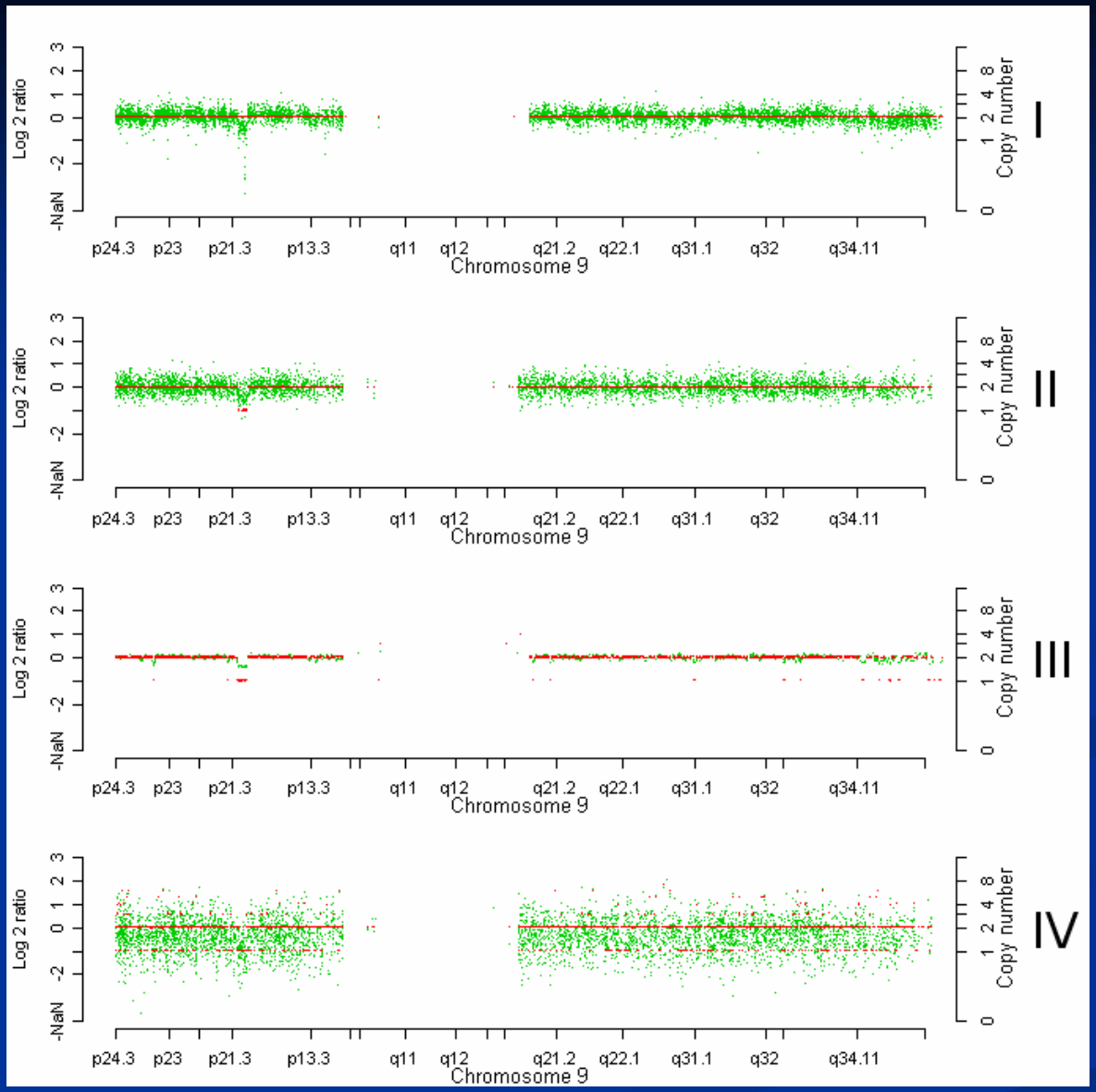
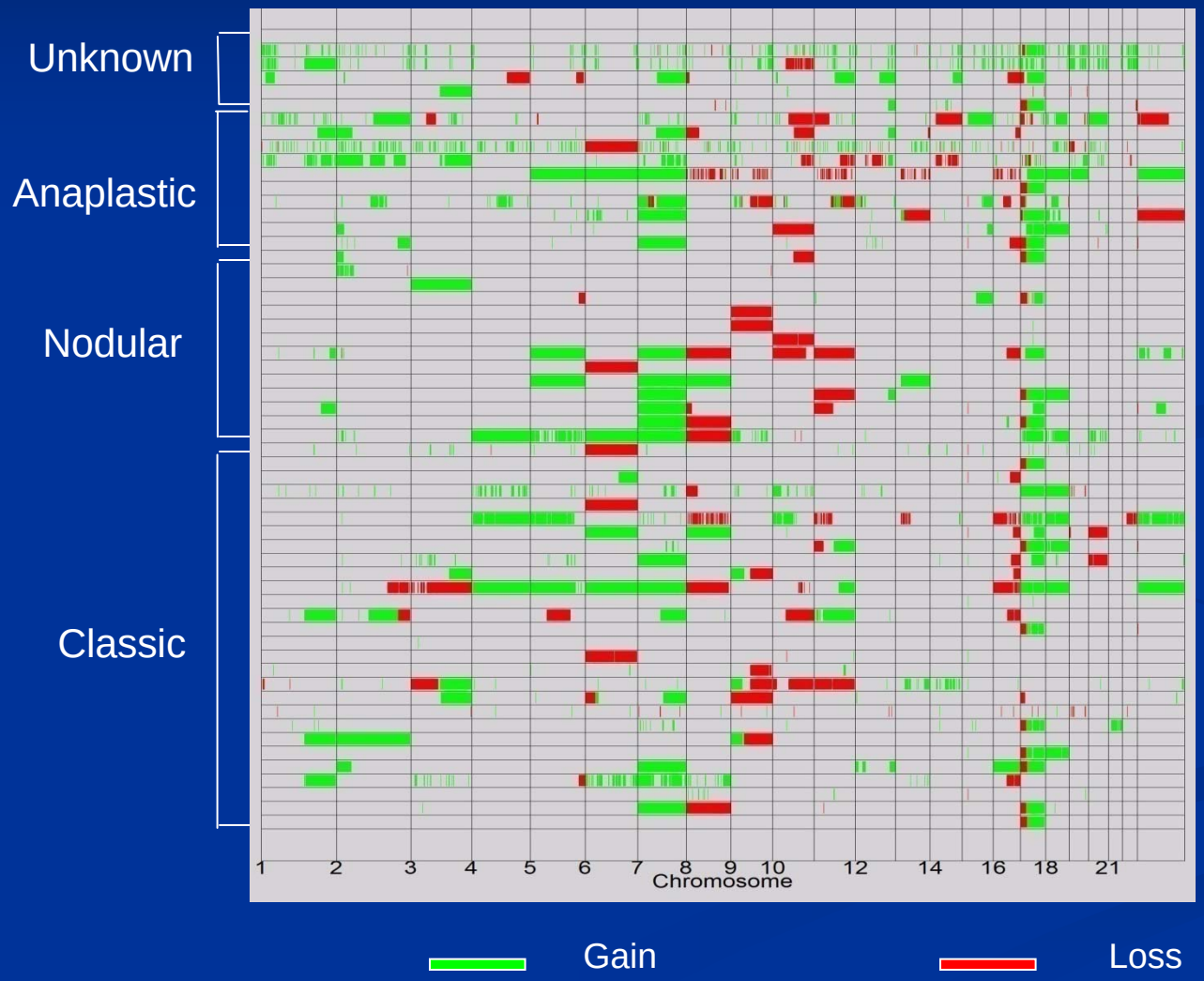


Fig. 3: Sample 688 copy number of SNPs on chromosome 9.
I: Illumina 310K data analyzed by BeadStudio; II: Affymetrix 250K Nsp data analyzed by CNAG2.0
III: Affymetrix 250K Nsp data analyzed by CNAT4.0.1; IV: Affymetrix 250K Nsp data analyzed by dChip

Log2 ratio Copy number

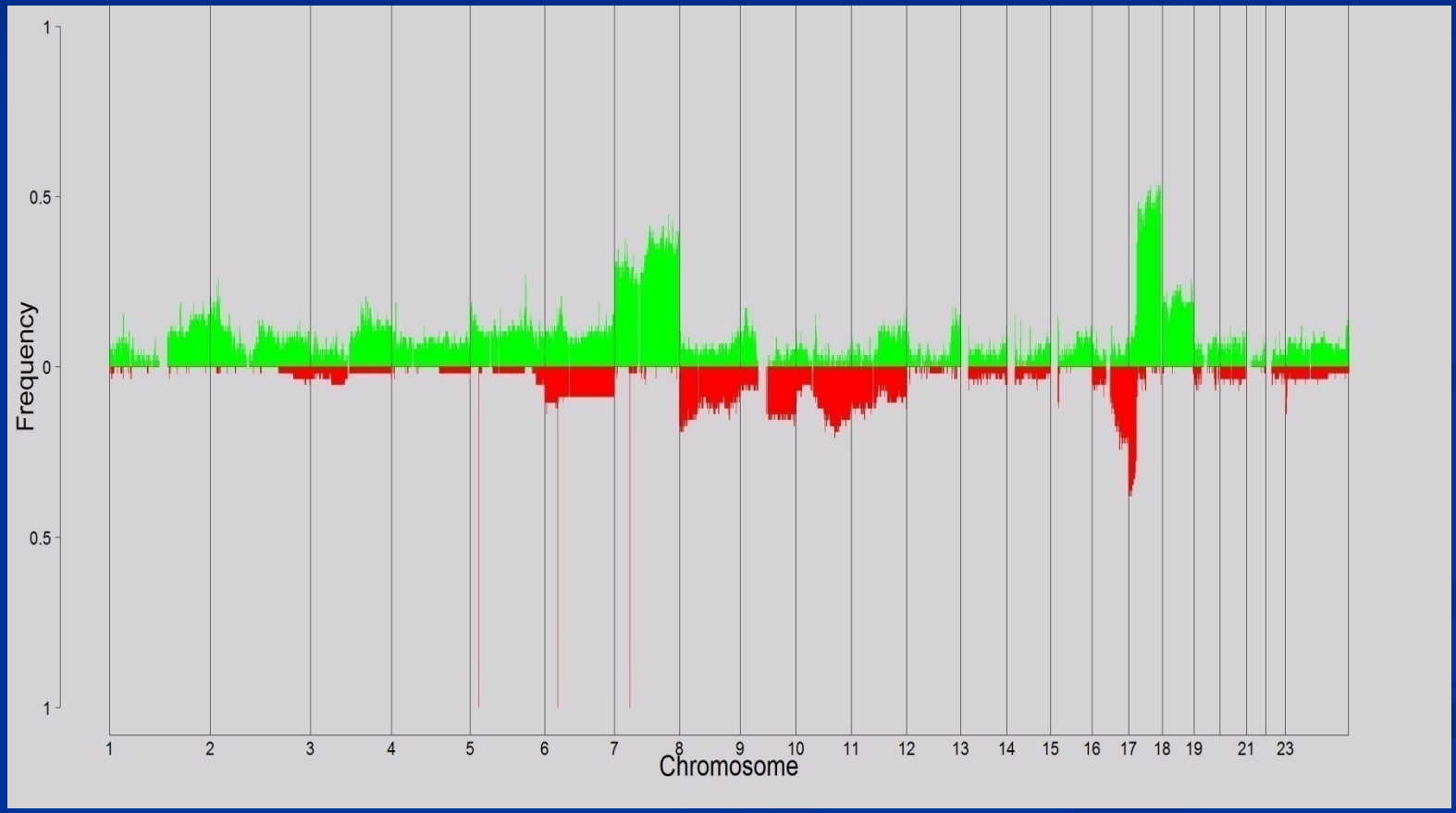
Medulloblastoma project

CNAs in 58 MBs



- 3062 CNAs
  - 53 per patient
- 754 (24.6%) losses
- 2308 (75.4%) gains

Frequency of SNP copy number changes for 58 MBs



Gain

Loss