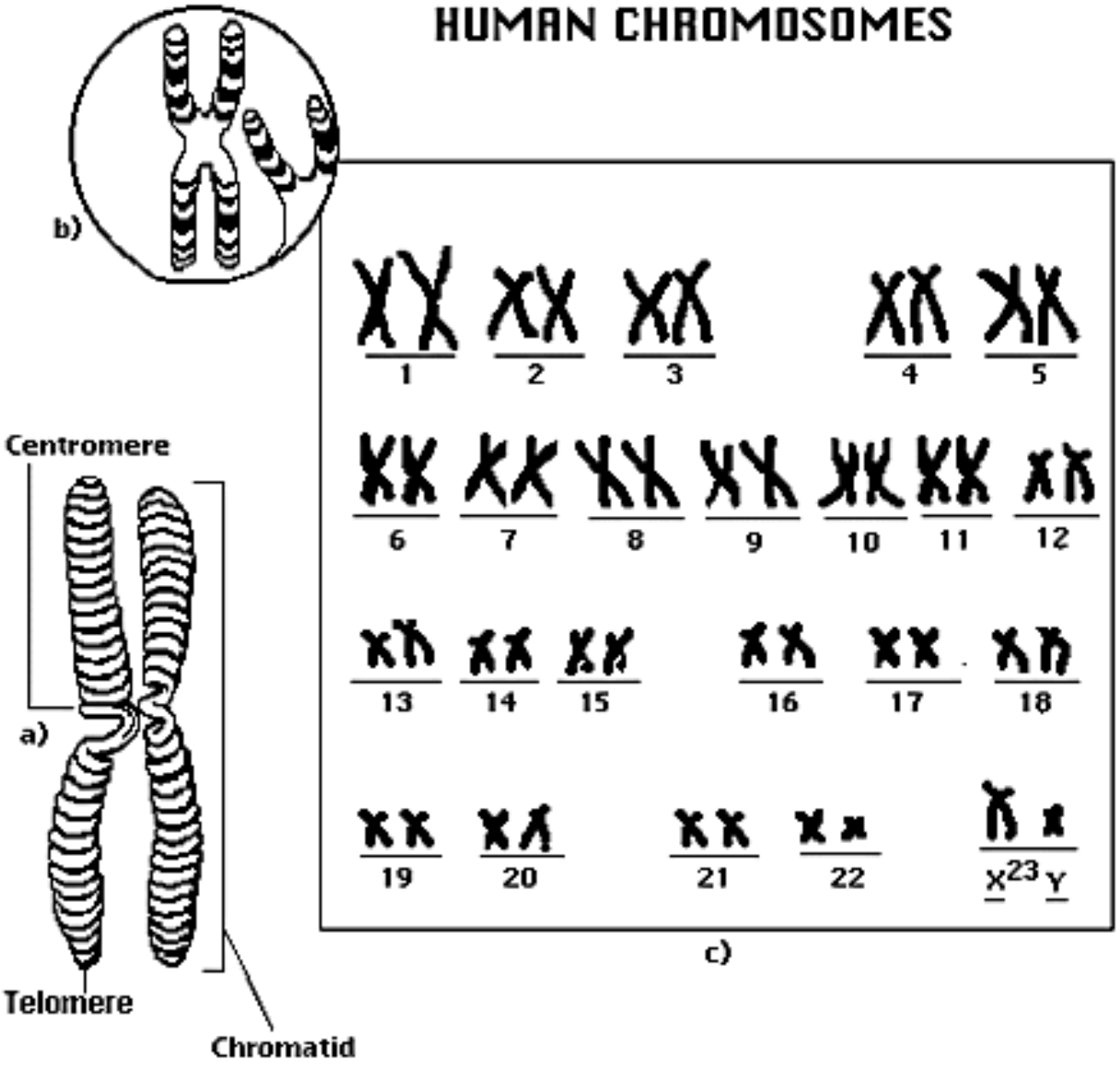
Characterization of Allele-Specific Copy Number in Tumor Genomes

Hao Chen[2] Haipeng Xing[1]
Nancy R. Zhang[2]

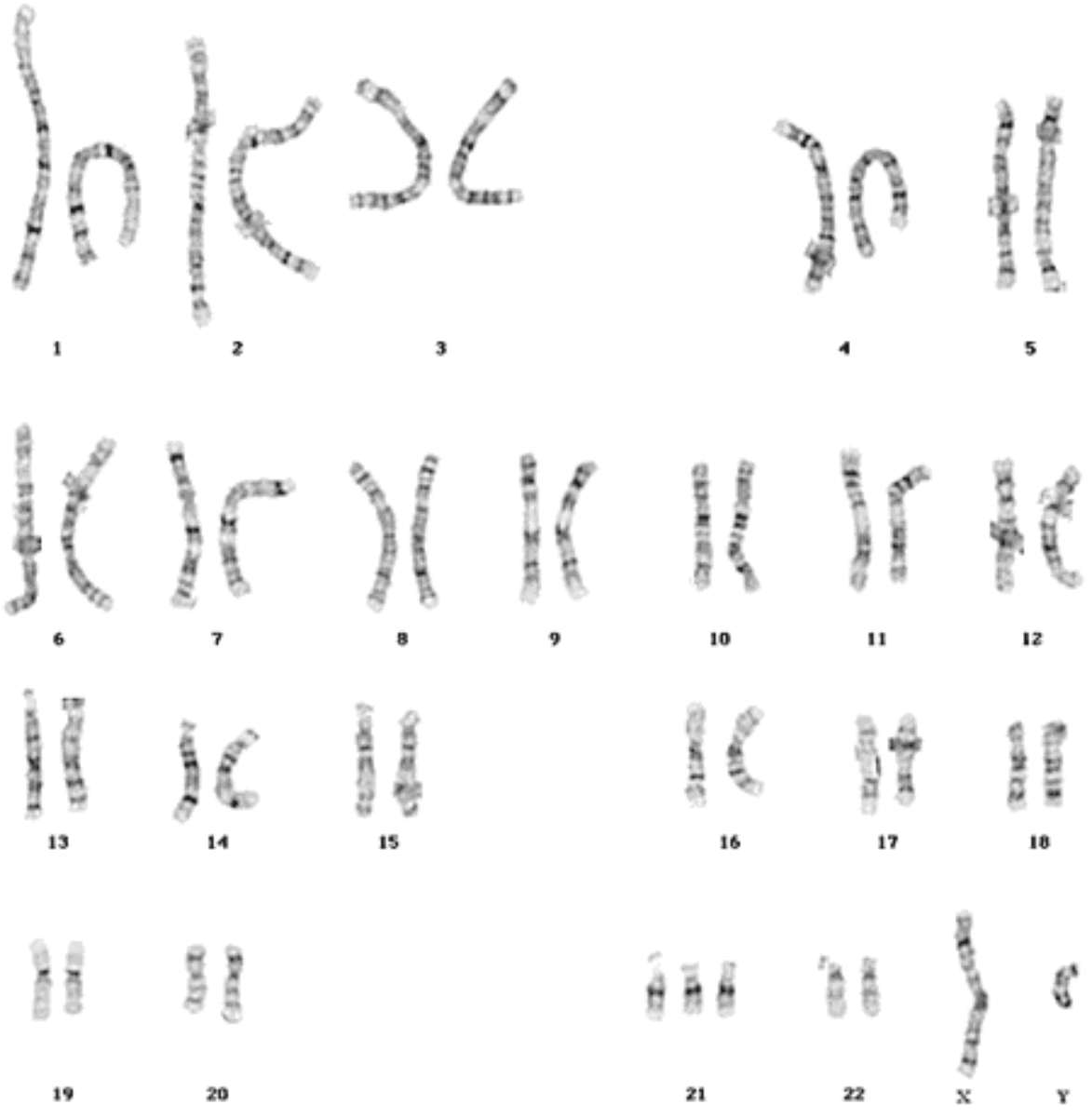
[1]Department of Statistics
Stonybrook University of New York
[2]Department of Statistics
Stanford University

Statistical Genomics Workshop, June 2009

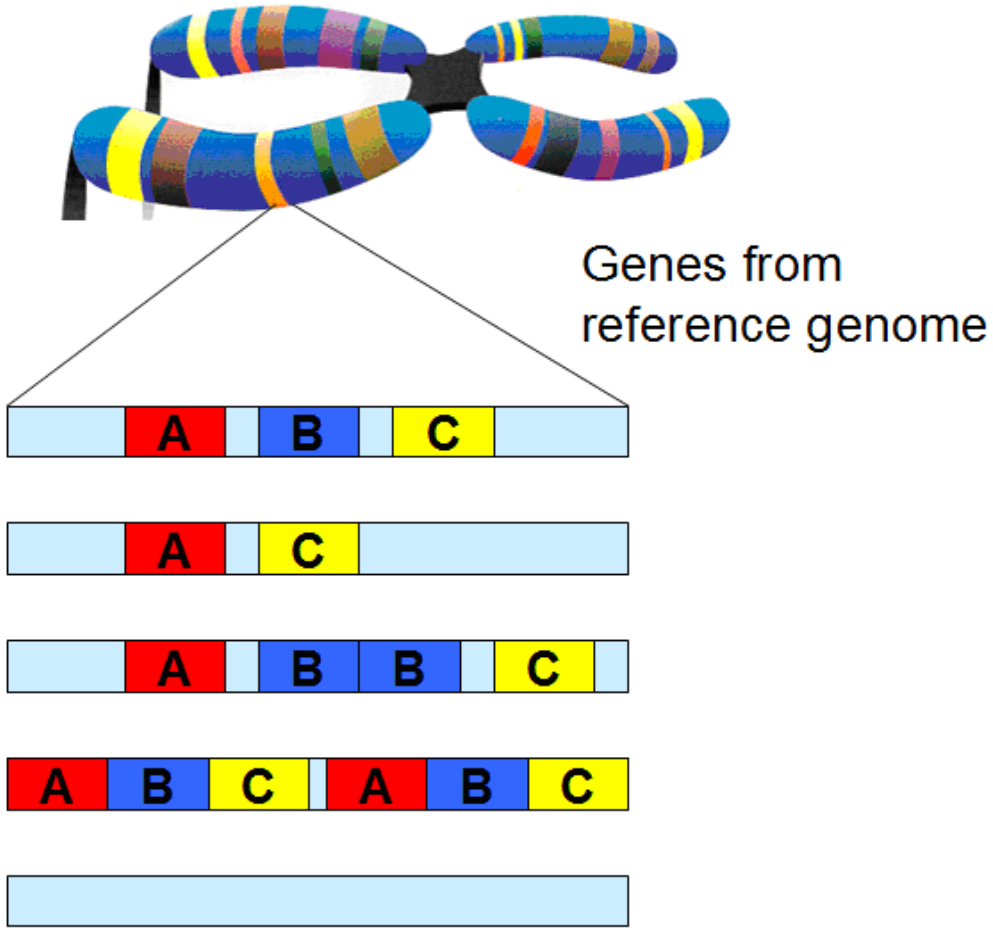
Human chromosomes come in pairs.



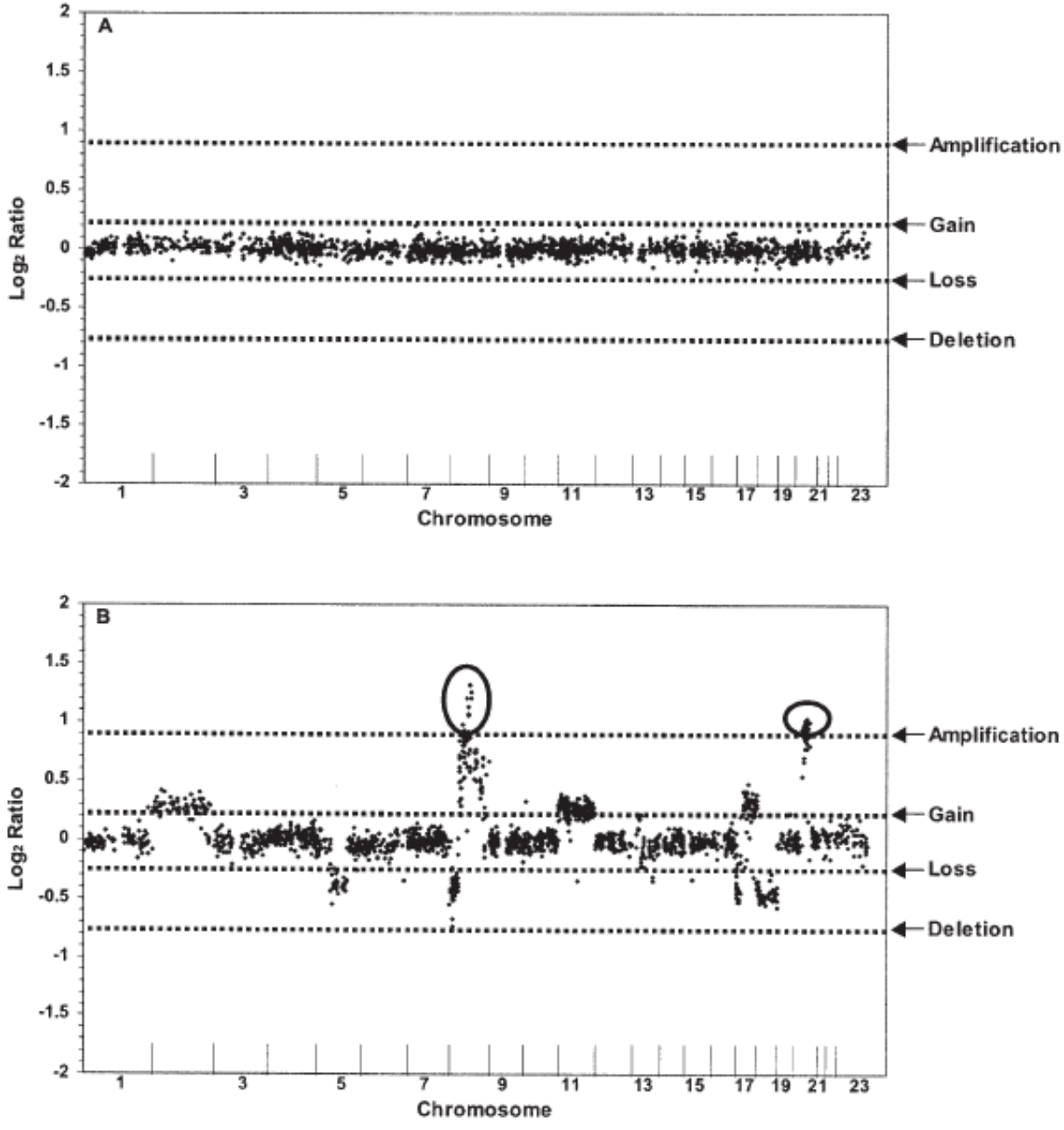
Trisomy of 21 in Downs syndrome



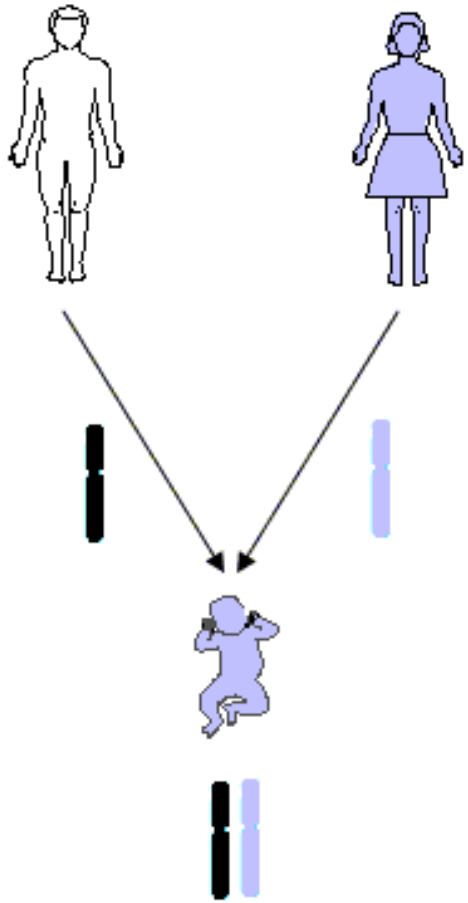
Copy Number Changes at a Smaller Scale



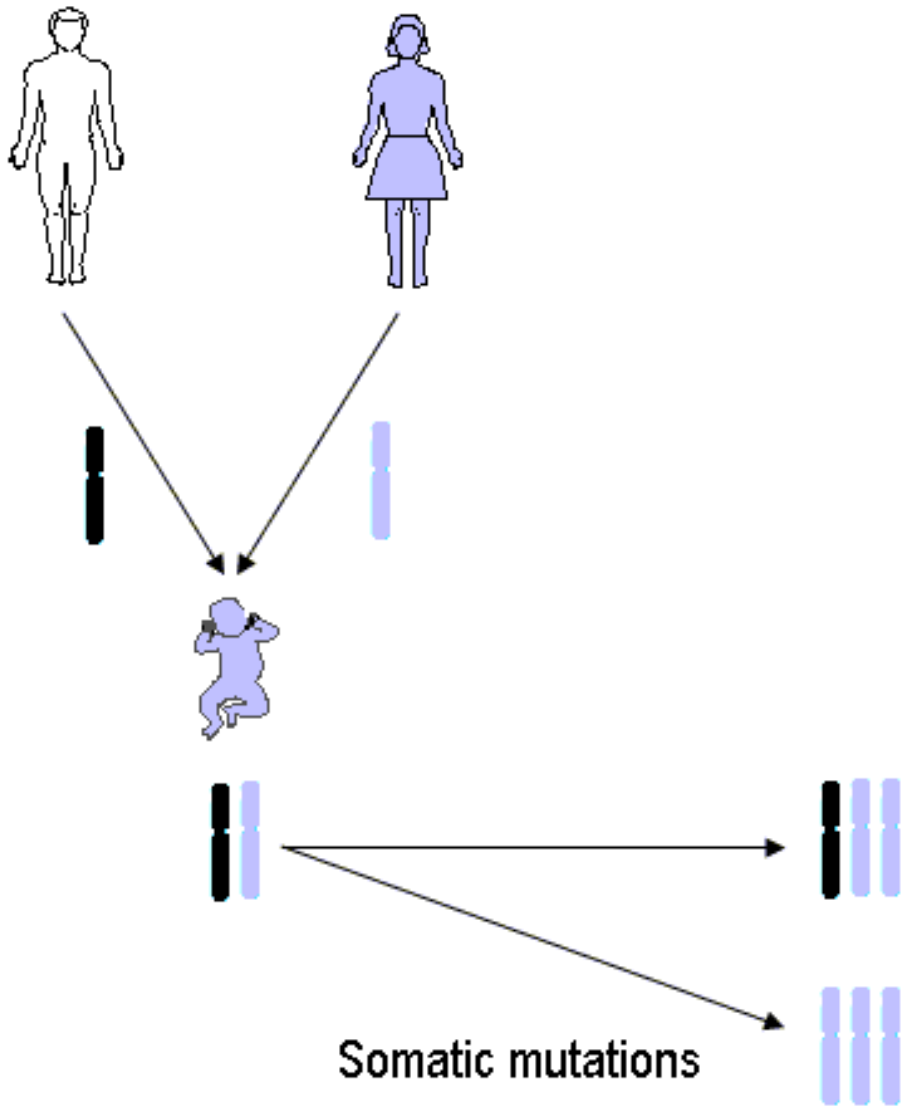
Change in total copy number in cancer



Which chromosome hasgained or lost copies?

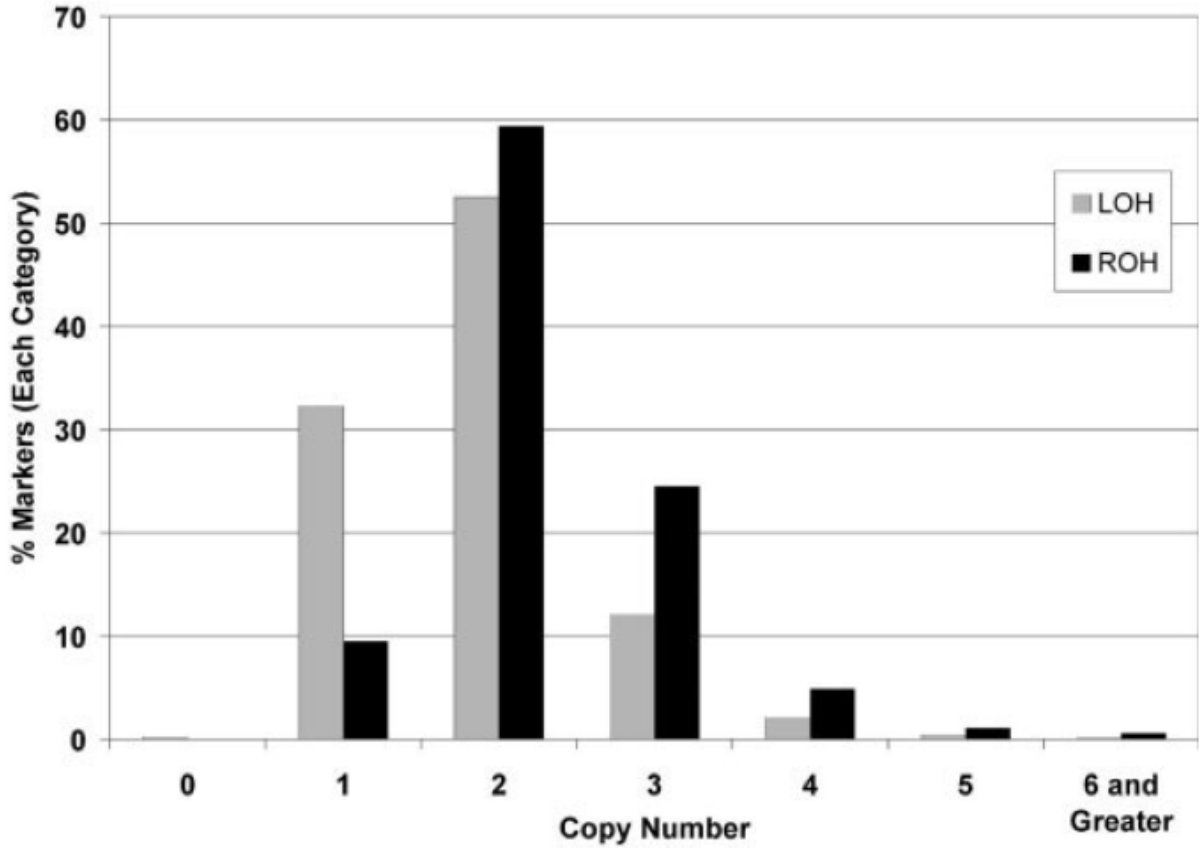


Which chromosome hasgained or lost copies?



Loss of Heterozygosity

A large fraction of LOH events don't involve change in total copy number.



Data from 24 pancreatic cancer cell lines, Calhoun et al., Genes, Chromosomes and Cancer 45:1070

Total versus Allele-specific Copy Numbers

Each individual has two copies of every chromosome, the maternal copy and the paternal copy.

Total versus Allele-specific Copy Numbers

Each individual has two copies of every chromosome, the maternal copy and the paternal copy.

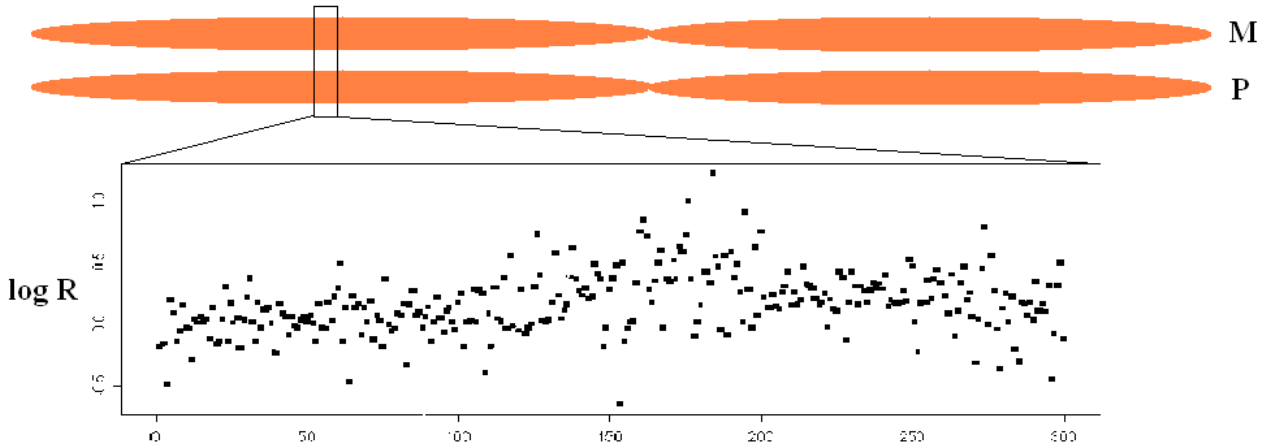
1. Certain experimental platforms, such as Agilent and Nimblegen, provides only total copy number estimates, that is, the sum of the copy numbers of both maternal and paternal chromosomes.

Total versus Allele-specific Copy Numbers

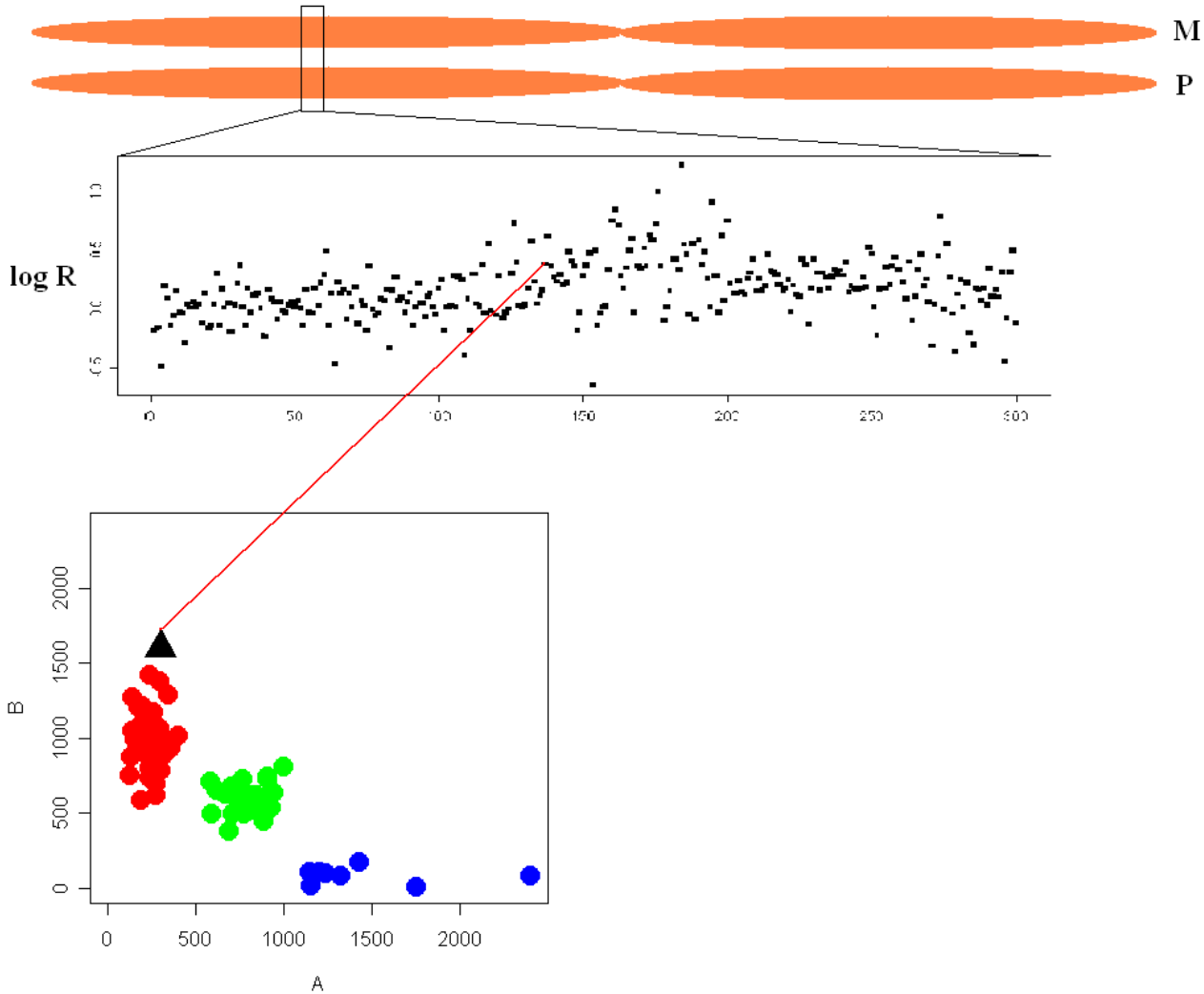
Each individual has two copies of every chromosome, the maternal copy and the paternal copy.

1. Certain experimental platforms, such as Agilent and Nimblegen, provides only total copy number estimates, that is, the sum of the copy numbers of both maternal and paternal chromosomes.
2. Other platforms, such as genotyping arrays (Illumina Beadarray, Affymetrix) can provide esimates of the copy number of each allele at selected polymorphic loci.

Data from SNP arrays (Illumina, Affymetrix)

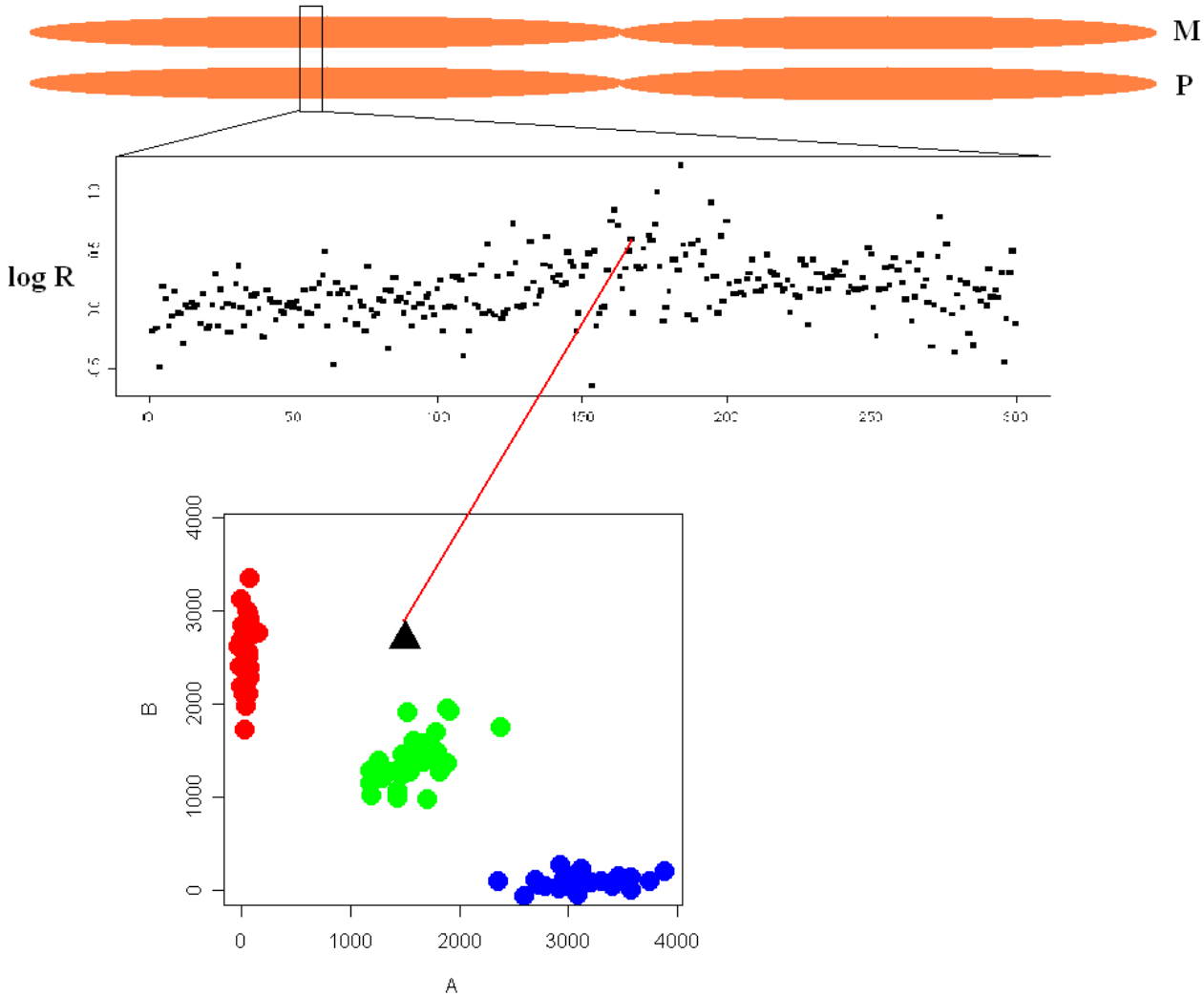


Data from SNP arrays (Illumina, Affymetrix)



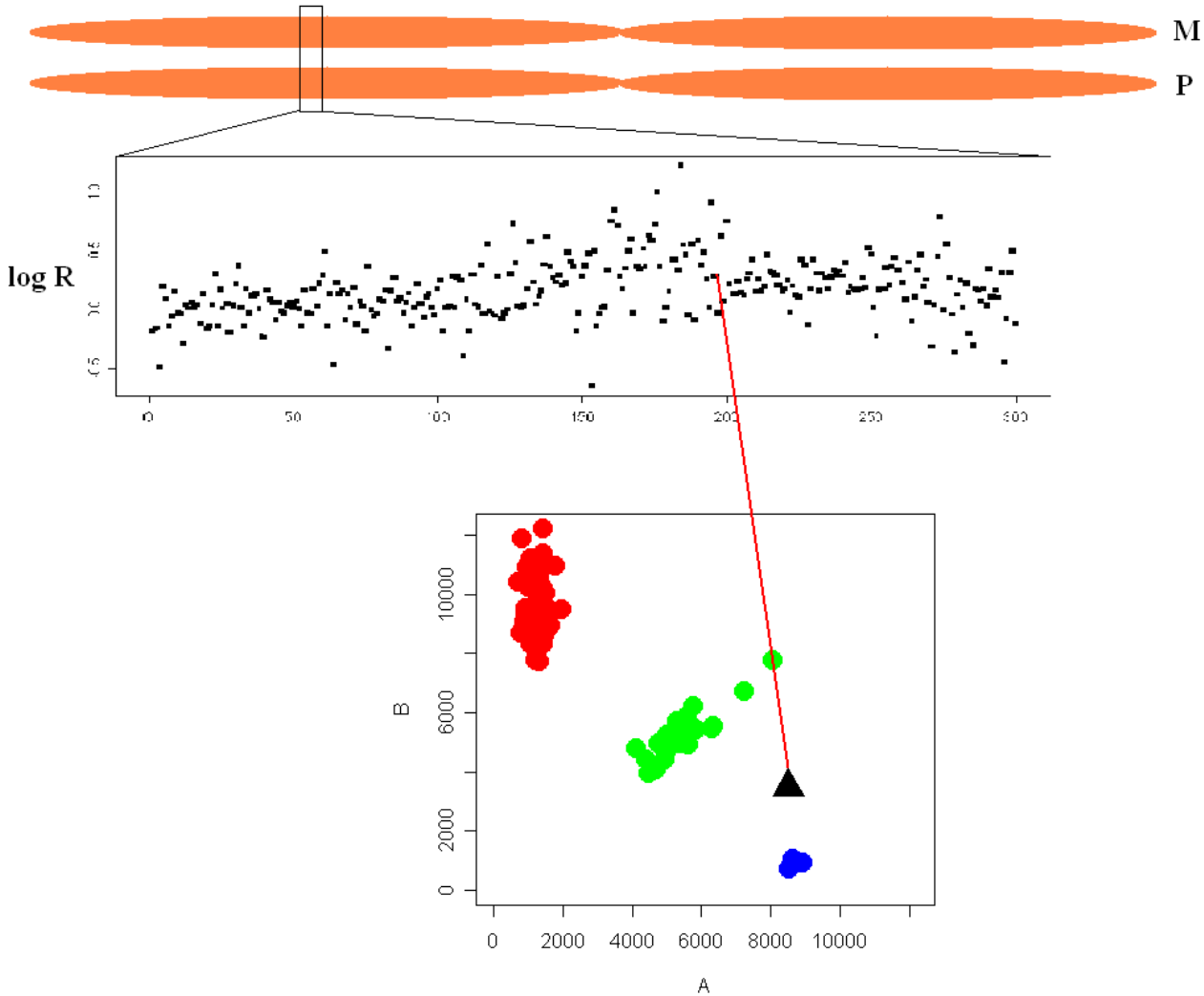
Colored points: population control, black triangle: target sample.

Data from SNP arrays (Illumina, Affymetrix)



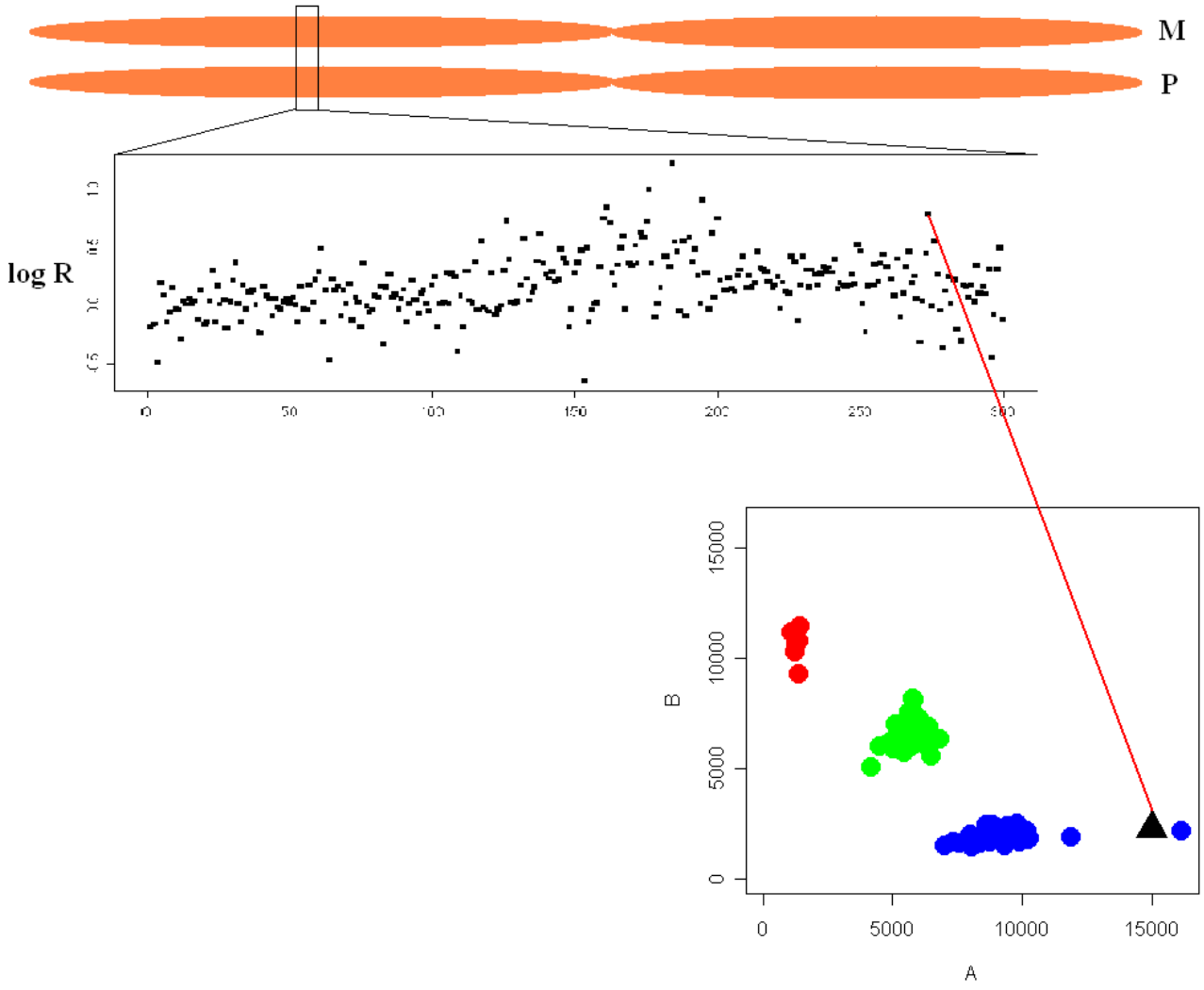
Colored points: population control, black triangle: target sample.

Data from SNP arrays (Illumina, Affymetrix)



Colored points: population control, black triangle: target sample.

Data from SNP arrays (Illumina, Affymetrix)



Colored points: population control, black triangle: target sample.

Why do we care about parent-specific copy number?

1. Many loss-of-heterozygosity events do not cause a change in total copy number.

Why do we care about parent-specific copy number?

1. Many loss-of-heterozygosity events do not cause a change in total copy number.
2. Allele-specific nature of amplification and deletion often has biological relevance.

Why do we care about parent-specific copy number?

1. Many loss-of-heterozygosity events do not cause a change in total copy number.
2. Allele-specific nature of amplification and deletion often has biological relevance.
3. Making use of information from both alleles can improve CNV detection accuracy.

Why do we care about parent-specific copy number?

1. Many loss-of-heterozygosity events do not cause a change in total copy number.
2. Allele-specific nature of amplification and deletion often has biological relevance.
3. Making use of information from both alleles can improve CNV detection accuracy.
4. Teasing apart the parent-specific copy numbers allow us to quantify the fraction of cells in the sample that contain the copy number aberration. This makes possible the quantitative study of tumor clonality.

Why do we care about parent-specific copy number?

1. Many loss-of-heterozygosity events do not cause a change in total copy number.
2. Allele-specific nature of amplification and deletion often has biological relevance.
3. Making use of information from both alleles can improve CNV detection accuracy.
4. Teasing apart the parent-specific copy numbers allow us to quantify the fraction of cells in the sample that contain the copy number aberration. This makes possible the quantitative study of tumor clonality.

How should this data be modeled and visualized?

More on the Clonality of Cancer Samples

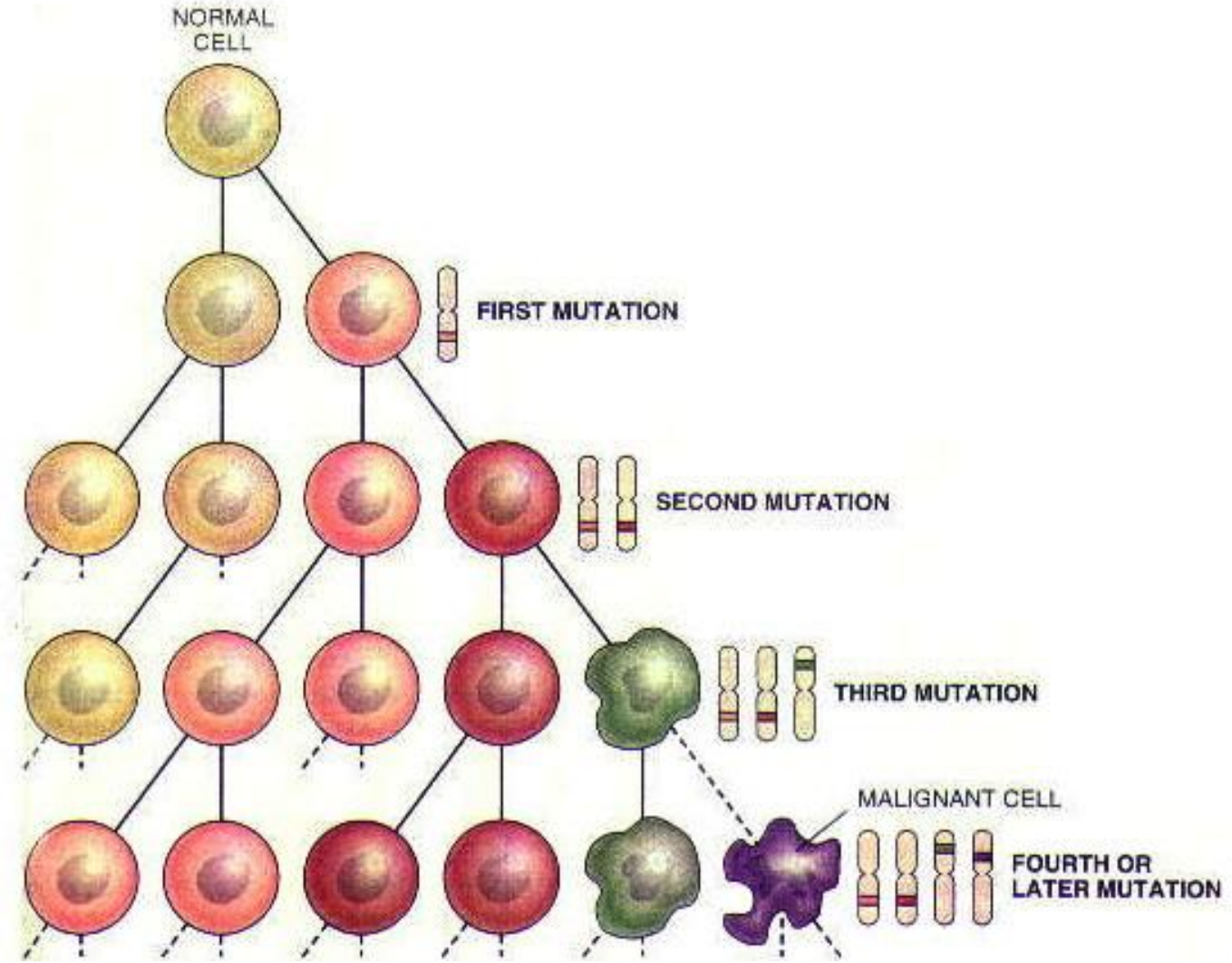


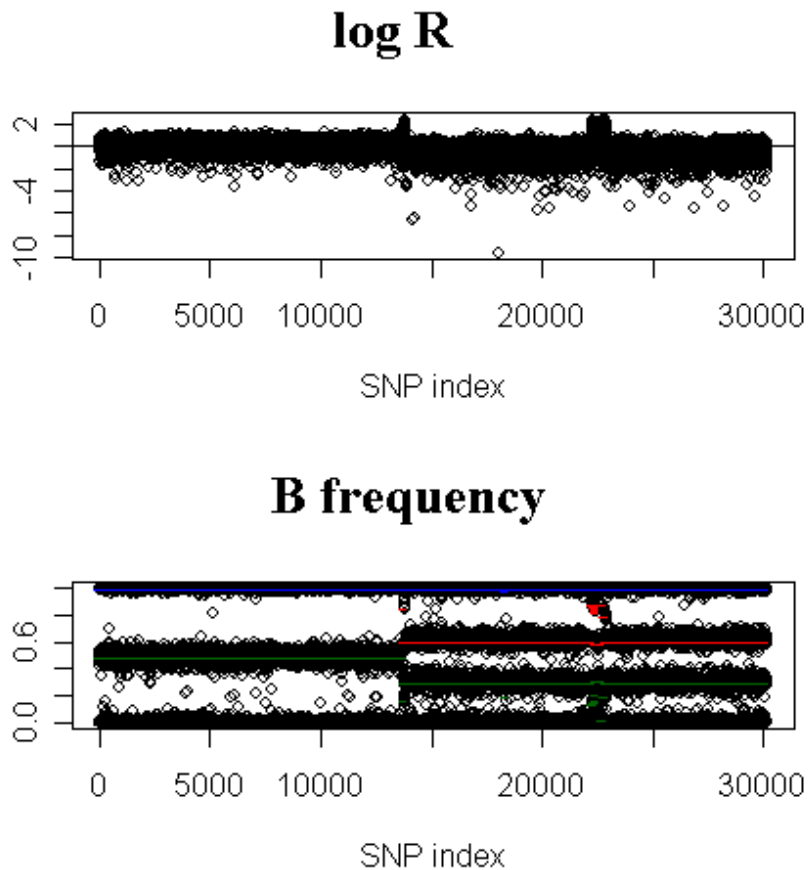
Image from: http://science.kennesaw.edu/ mhermes/cisplat/cisplat19.htm

Using information in both alleles can improve power

Illumina outputs SNP-normalized data:

$$R = (A + B)/[E(A + B)], \quad b = (2/\pi)\arctan(B/A),$$

Normalized to population controls...



Copy number estimation using SNP chip data

Existing approaches:

1. Segment sequence based on total copy number, then cluster segments based on allele information (LaFramboise et al., 2005).

Copy number estimation using SNP chip data

Existing approaches:

1. Segment sequence based on total copy number, then cluster segments based on allele information (LaFramboise et al., 2005).

   This misses copy neutral loss of heterozygosity events.

Copy number estimation using SNP chip data

Existing approaches:

1. Segment sequence based on total copy number, then cluster segments based on allele information (LaFramboise et al., 2005).

   This misses copy neutral loss of heterozygosity events.

2. Use existing segmentation tools to separately segment the log ratio and some thresholded version of B-frequency (Staaf et al. (2008), Assié et al. (2008)).

Copy number estimation using SNP chip data

Existing approaches:

1. Segment sequence based on total copy number, then cluster segments based on allele information (LaFramboise et al., 2005).

   This misses copy neutral loss of heterozygosity events.

2. Use existing segmentation tools to separately segment the log ratio and some thresholded version of B-frequency (Staaf et al. (2008), Assié et al. (2008)).
   B-frequency is a mixture residing within $[0, 1]$*...*

Copy number estimation using SNP chip data

Existing approaches:

1. Segment sequence based on total copy number, then cluster segments based on allele information (LaFramboise et al., 2005).

   This misses copy neutral loss of heterozygosity events.

2. Use existing segmentation tools to separately segment the log ratio and some thresholded version of B-frequency (Staaf et al. (2008), Assié et al. (2008)).
   B-frequency is a mixture residing within $[0, 1]$*...*

3. Hidden Markov model with hidden states representing genotypes *A*, *B*, *AA*, *AB*, *BB*, *AAB*, *ABB*, ... PennCNV, QuantiSNP.

Copy number estimation using SNP chip data

Existing approaches:

1. Segment sequence based on total copy number, then cluster segments based on allele information (LaFramboise et al., 2005).

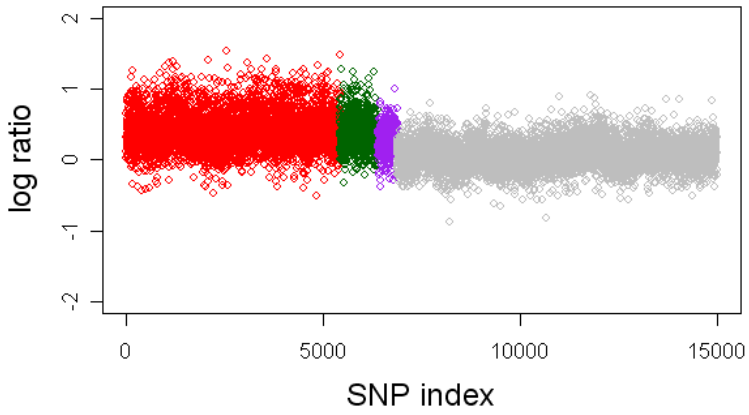
   This misses copy neutral loss of heterozygosity events.

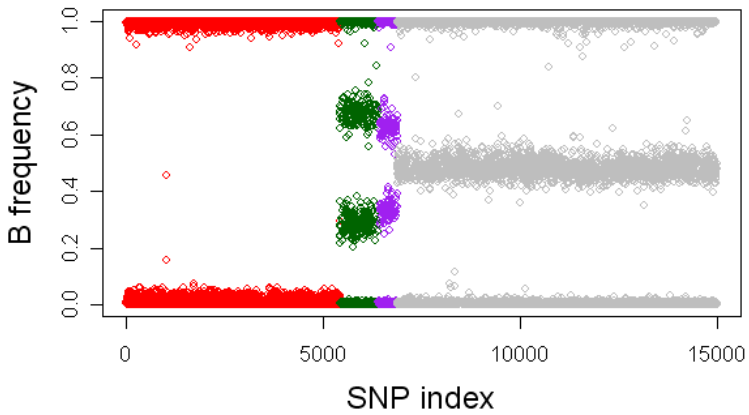
2. Use existing segmentation tools to separately segment the log ratio and some thresholded version of B-frequency (Staaf et al. (2008), Assié et al. (2008)).
   B-frequency is a mixture residing within $[0, 1]$*...*

3. Hidden Markov model with hidden states representing genotypes *A*, *B*, *AA*, *AB*, *BB*, *AAB*, *ABB*, ... PennCNV, QuantiSNP.

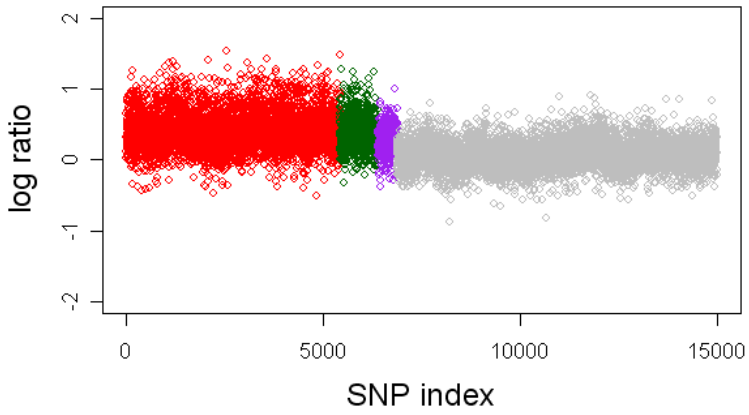
   Cancer samples are often a mixture of sub-populations with different genotypes.

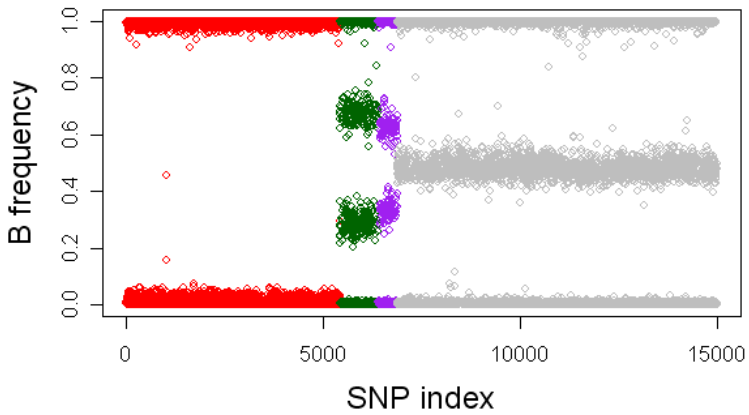
Fractional changes

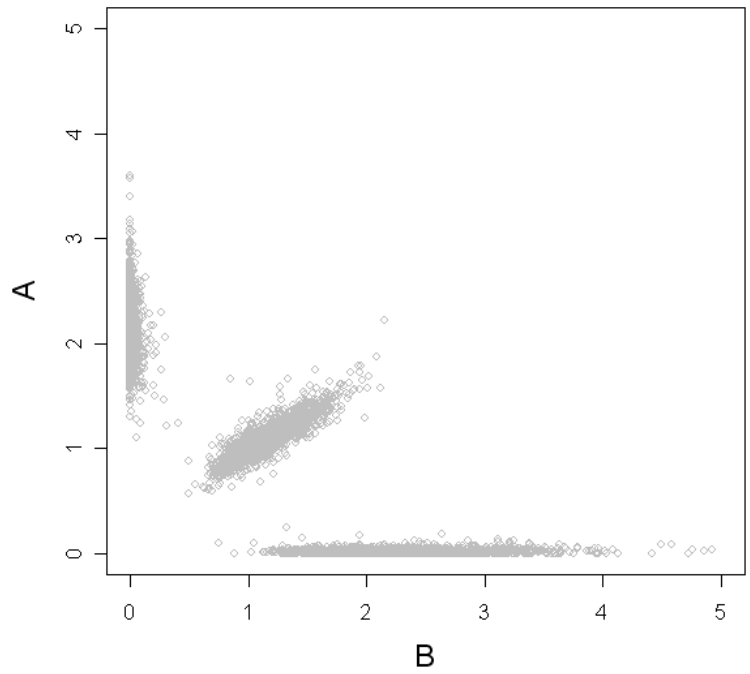




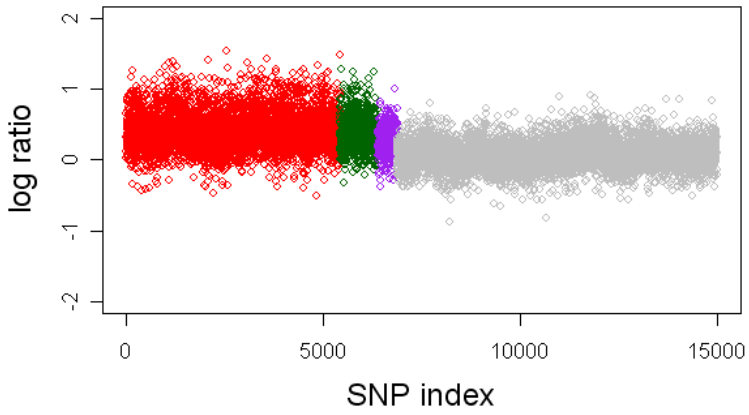
Fractional changes

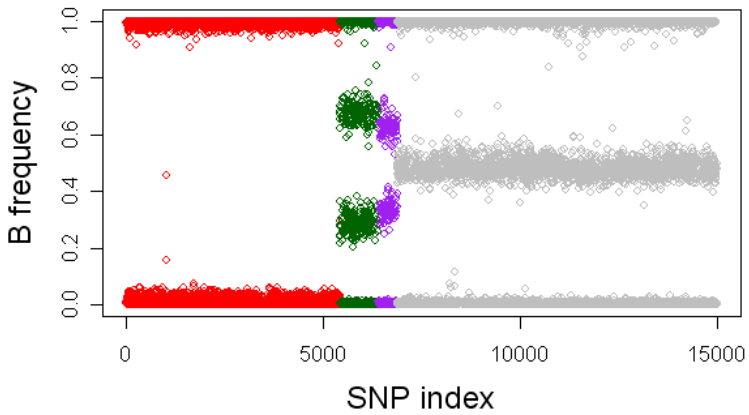


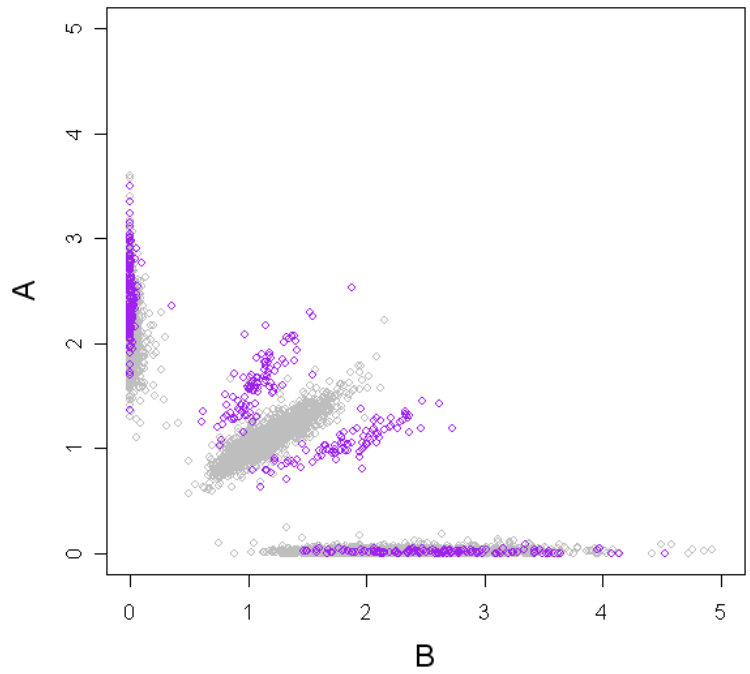




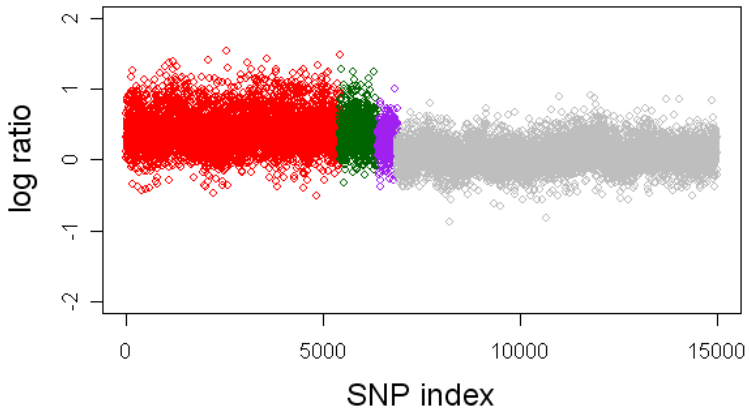
Fractional changes

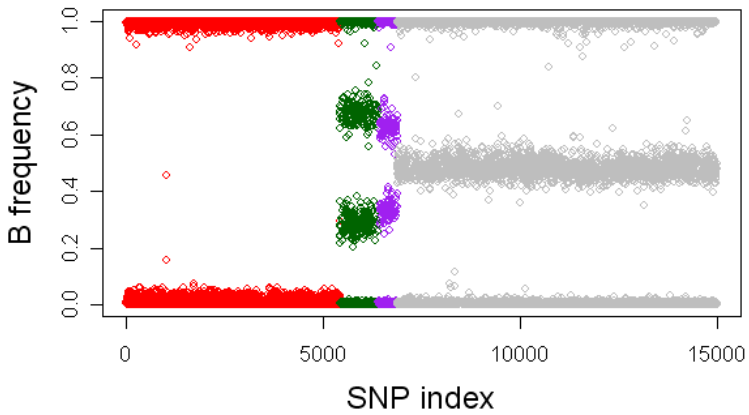


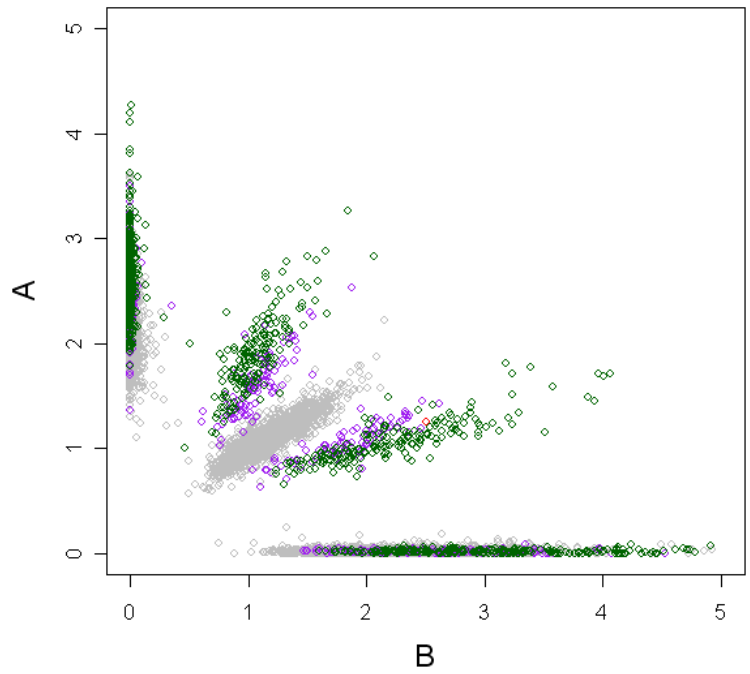




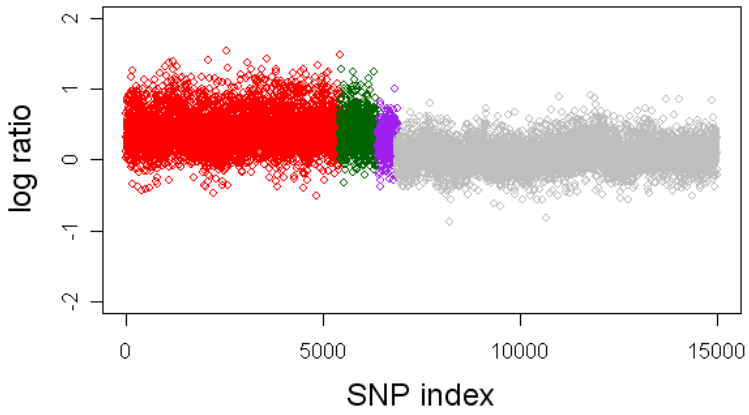
Fractional changes

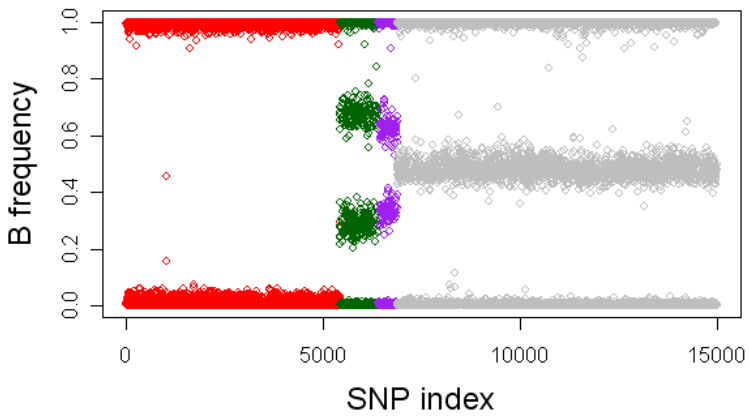


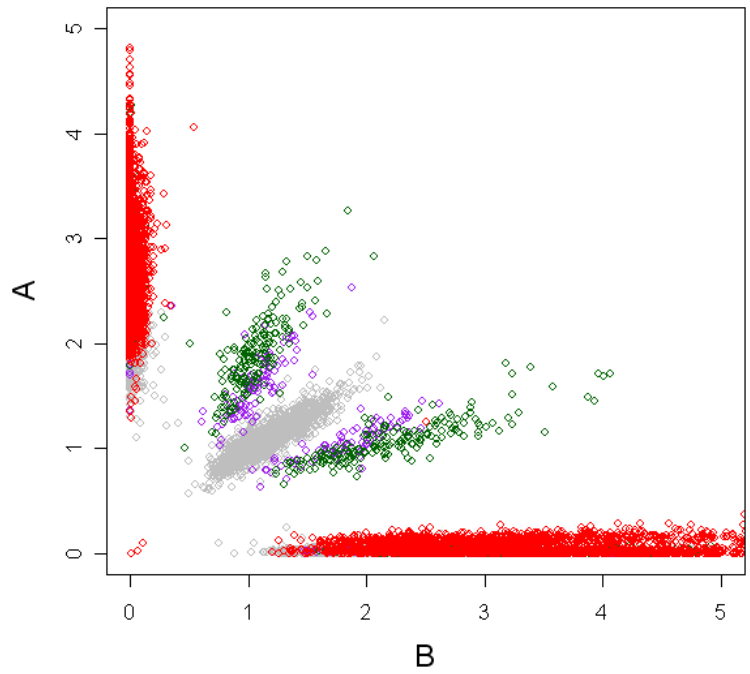




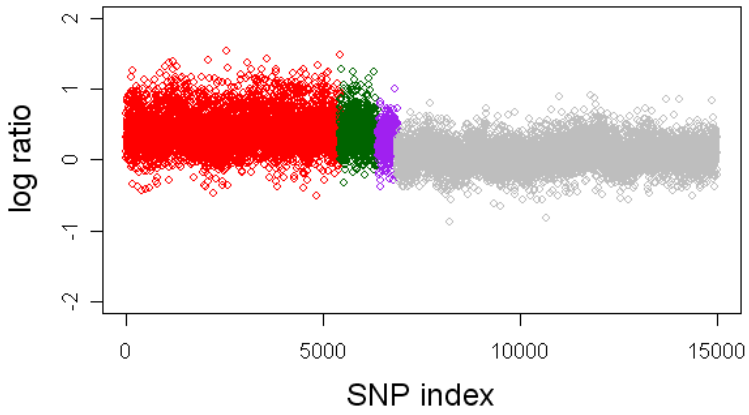
Fractional changes

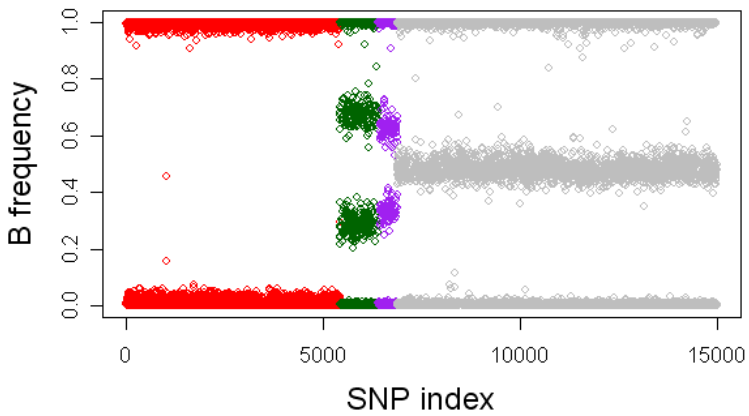


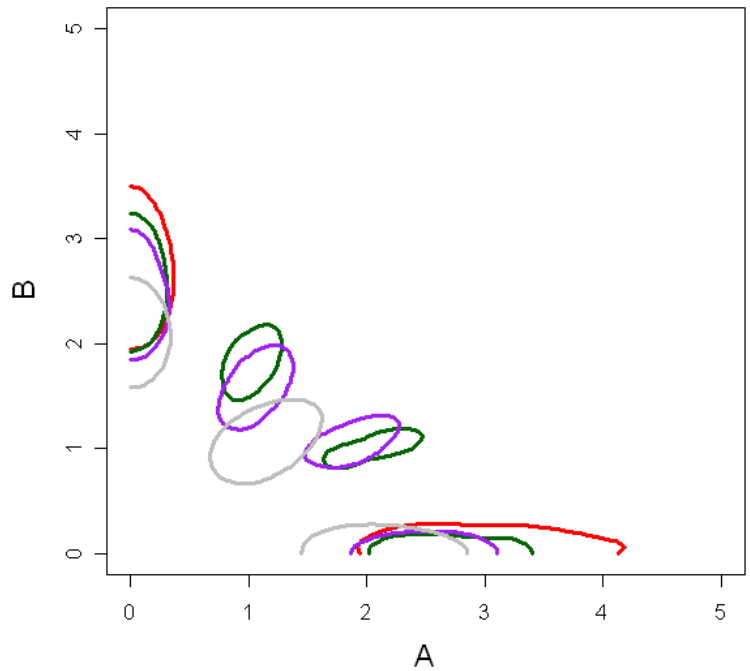




Fractional changes







A helpful observation

For any homogeneous segment, the (A, B) allele intensities must take on either 2, 3, or 4 clusters. The locations of these clusters depend on the quantities of the maternal and paternal chromosome. The cluster membership of any specific SNP depends on the genotype.

Two-chromosome Markov jump model

1. Model description.
2. Estimation Procedure.
3. Algorithmic details.
4. Results on data.

Parental allele configuration

s_t	Interpretation
AA	Both parental chromosomes carry A.
AB	Maternal carries A, paternal carries B.
BA	Maternal carries B, paternal carries A.
BB	Both parental chromosomes carry B.

Parental allele configuration

s_t	Interpretation
AA	Both parental chromosomes carry *A*.
AB	Maternal carries *A*, paternal carries *B*.
BA	Maternal carries *B*, paternal carries *A*.
BB	Both parental chromosomes carry *B*.

Parent-specific copy numbers at location t: $\theta_{t,1}, \theta_{t,2}$.

Parental allele configuration

s_t	Interpretation
AA	Both parental chromosomes carry A.
AB	Maternal carries A, paternal carries B.
BA	Maternal carries B, paternal carries A.
BB	Both parental chromosomes carry B.

Parent-specific copy numbers at location t: $\theta_{t,1}, \theta_{t,2}$.

Allele specific copy numbers: $x_{t,1}, x_{t,2}$.

s_t	$x_{t,1}$	$x_{t,2}$
AA	$\theta_{t,1} + \theta_{t,2}$	0
AB	$\theta_{t,1}$	$\theta_{t,2}$
BA	$\theta_{t,2}$	$\theta_{t,1}$
BB	0	$\theta_{t,1} + \theta_{t,2}$

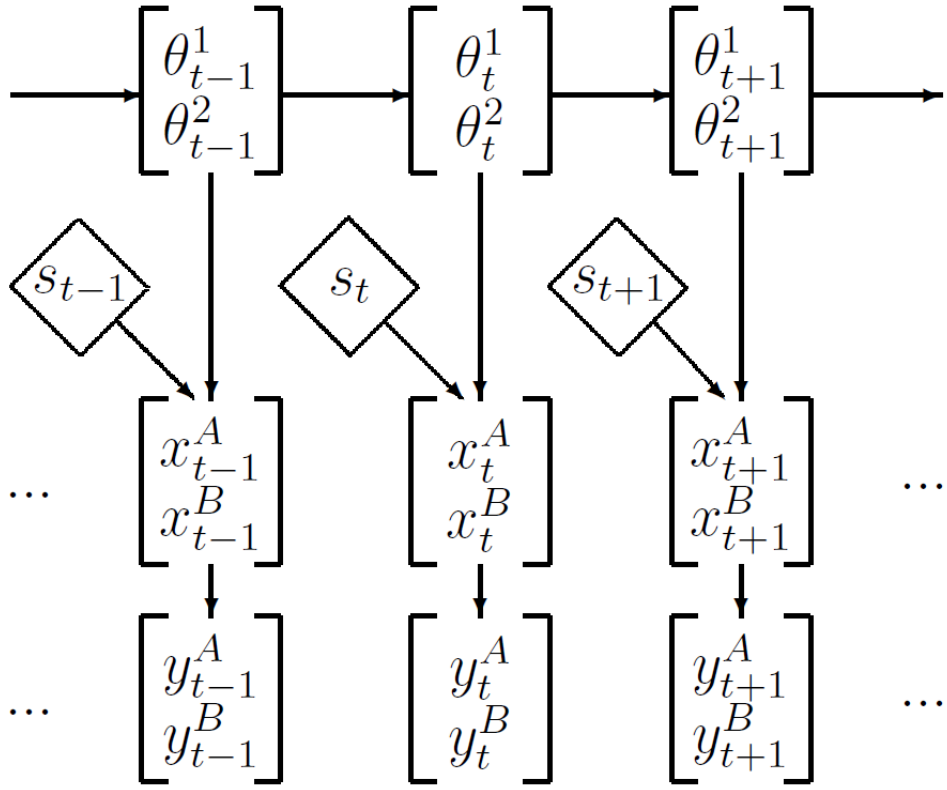
Model Overview

Parent specific Copy Number

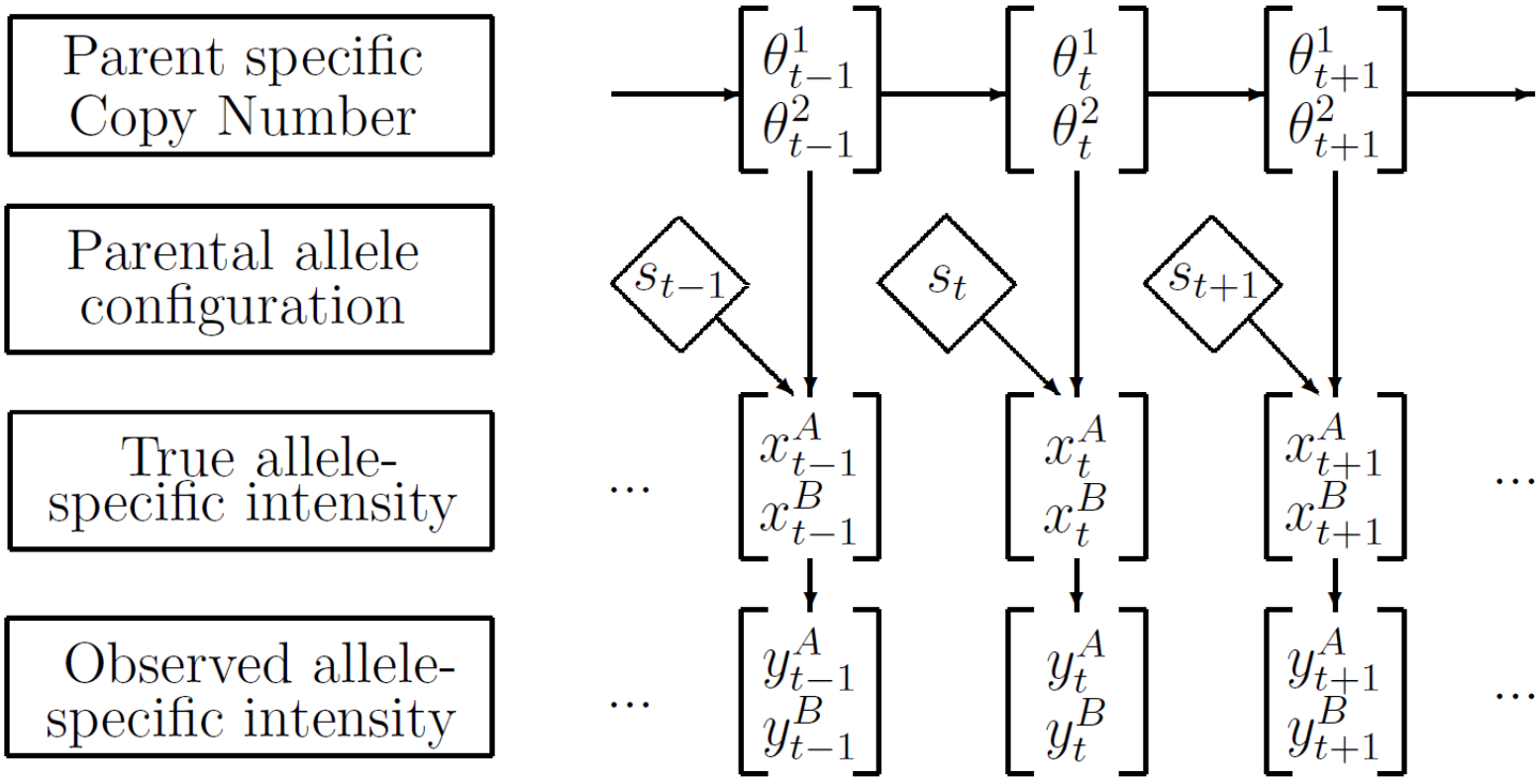
Parental allele configuration

True allele-specific intensity

Observed allele-specific intensity

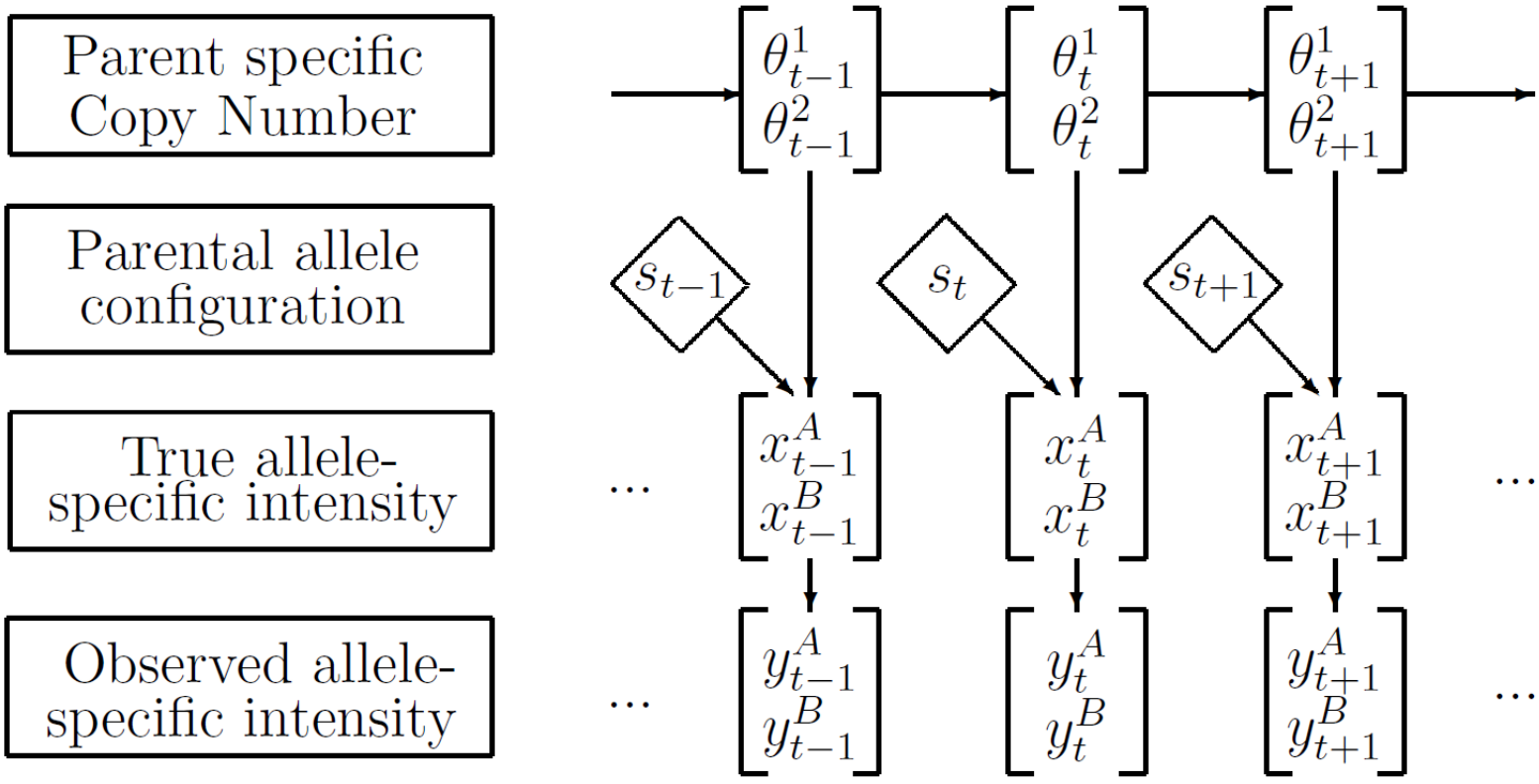


Model Overview



θ_t : Markov jump process,

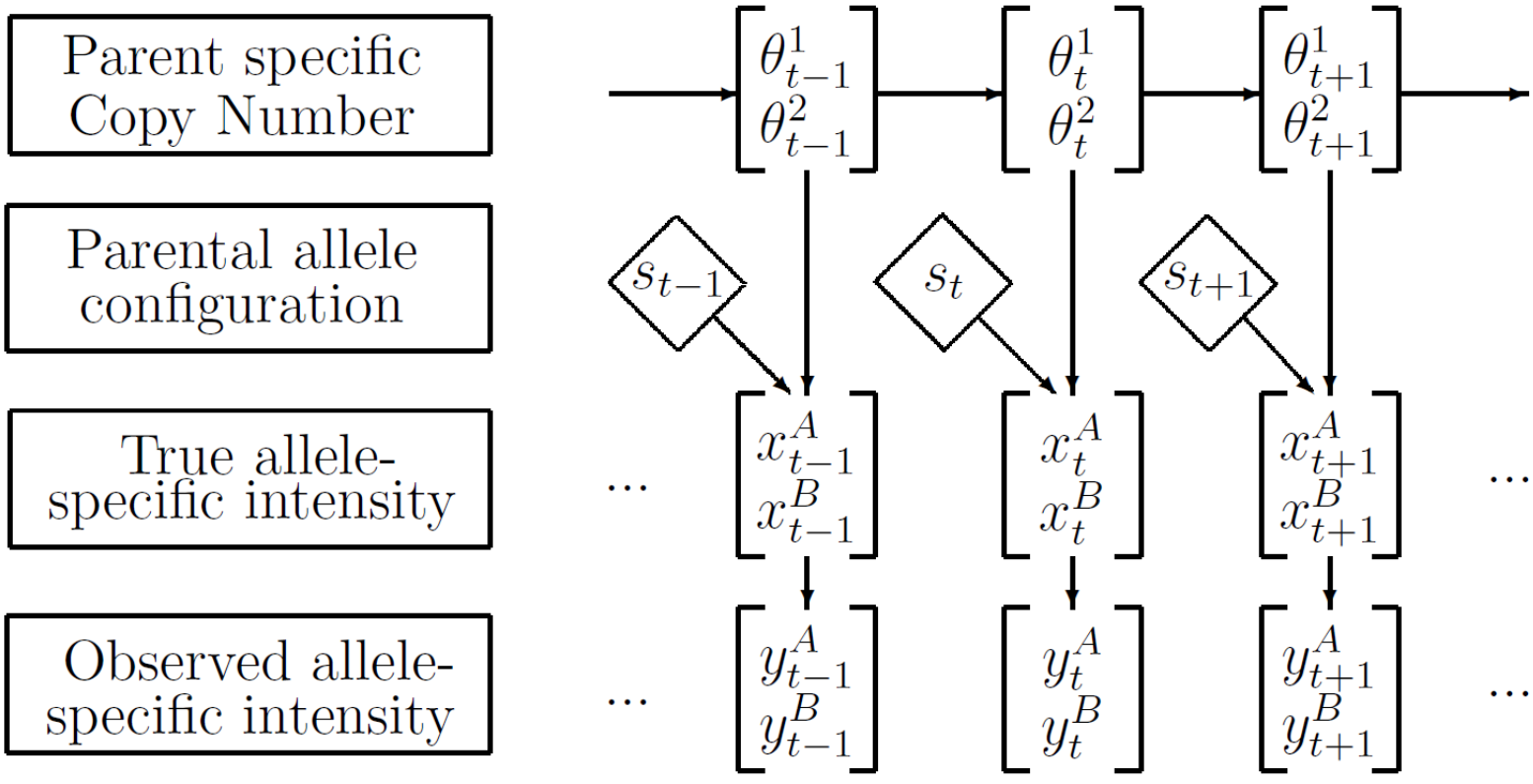
Model Overview



$$\theta_t \quad : \quad \text{Markov jump process,}$$

$$y_t \quad = \quad x_t + \epsilon, \quad \epsilon \sim N(0, \Sigma_{s_t}).$$

Model Overview



$$\theta_t \quad : \quad \text{Markov jump process,}$$

$$y_t \quad = \quad x_t + \epsilon, \quad \epsilon \sim N(0, \Sigma_{s_t}).$$

When somatic changes in copy number occur, s_t remains fixed.

Parental copy numbers θ_t

θ_t can belong to two states:

Baseline state: $\theta_t = (\frac{B}{2}, \frac{B}{2})$, changed state: $\theta_t \sim N(\mu, V)$.

Parental copy numbers θ_t

θ_t can belong to two states:

Baseline state: $\theta_t = \left(\frac{B}{2}, \frac{B}{2}\right)$, changed state: $\theta_t \sim N(\mu, V)$.

At each position, θ_t can:

1. remain at the same value as θ_{t-1},
2. jump to a changed state (if at baseline), or a new changed state (if already at changed state),
3. jump to baseline (if at changed state).

Parental copy numbers θ_t

θ_t can belong to two states:

Baseline state: $\theta_t = \left(\frac{B}{2}, \frac{B}{2}\right)$, changed state: $\theta_t \sim N(\mu, V)$.

At each position, θ_t can:

1. remain at the same value as θ_{t-1},
2. jump to a changed state (if at baseline), or a new changed state (if already at changed state),
3. jump to baseline (if at changed state).

This can be modeled with a 3-state Markov model with transition matrix:

$$P = \begin{pmatrix} 1-p & \frac{1}{2}p & \frac{1}{2}p \\ c & a & b \\ c & b & a \end{pmatrix}.$$

Parental allele configuration s_t

s_t can be modeled as:

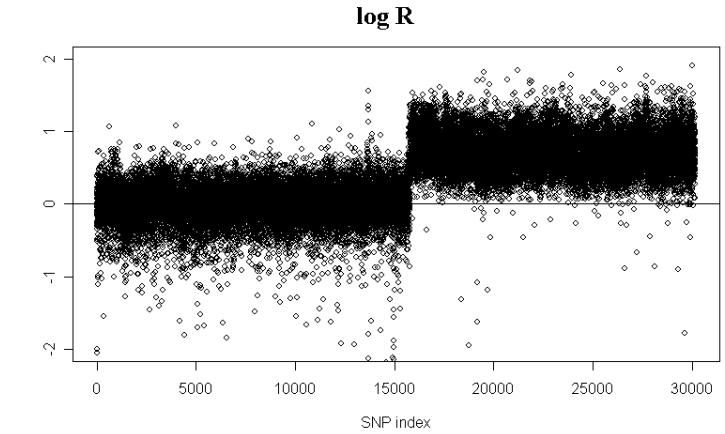
- i.i.d. multinomial distribution with known parameters

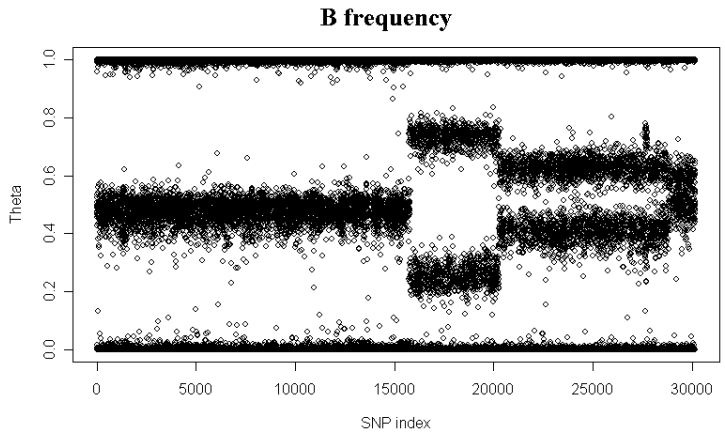
$$(p_t^{AA}, p_t^{BA}, p_t^{AB}, p_t^{BB}).$$

- Markov with transition probabilities between two SNPs estimated using Haplotype data.

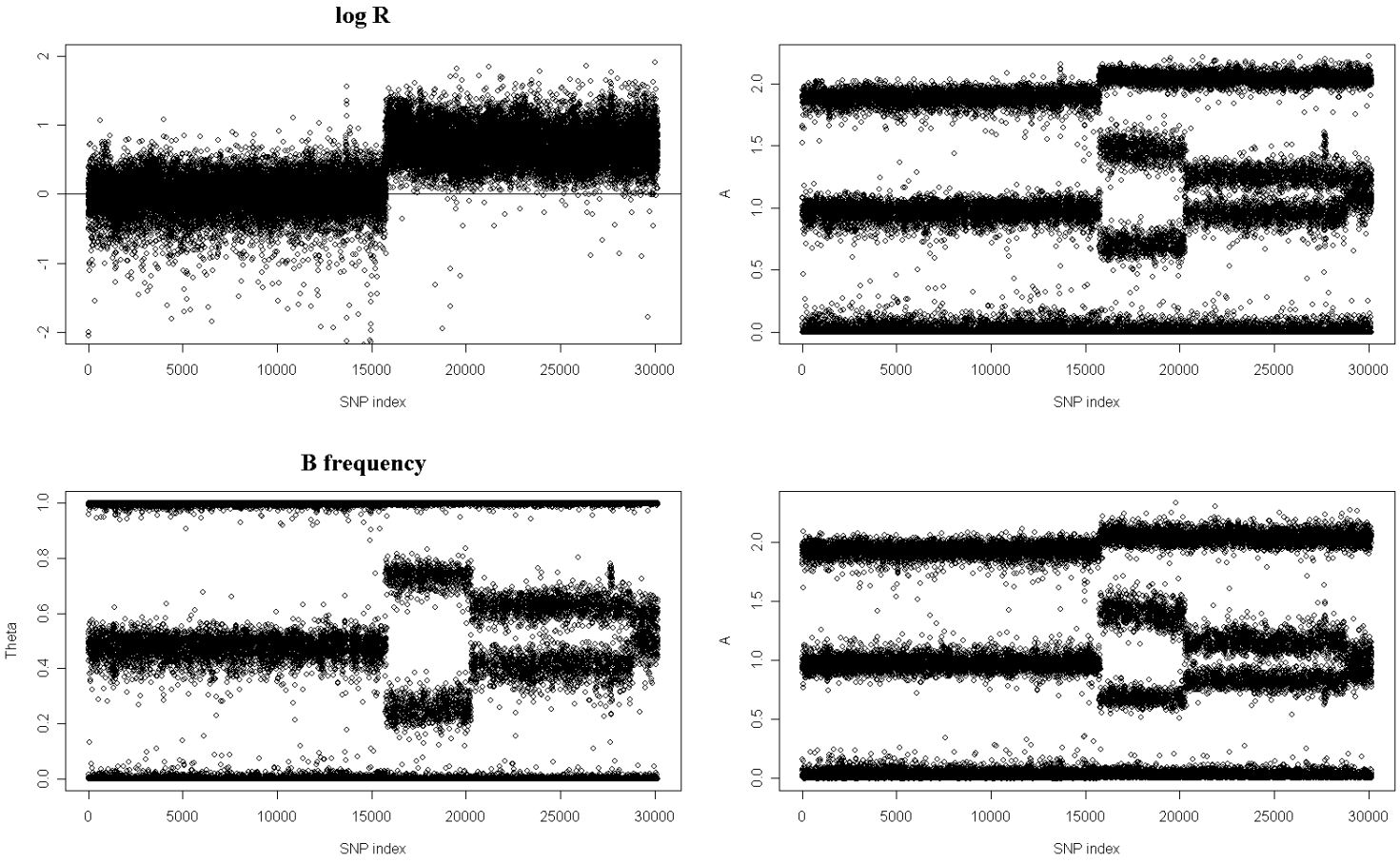
These parameters can be estimated from the appropriate population controls.

Data transformation





Data transformation



$X = (\log R + C)b, \quad Y = (\log R + C)(1 - b)$, plus some necessary symmetrization, variance stabilization work well.

Smoothing equations

If s_t were known, the posterior distribution of θ_t given the complete data sequence is a mixture of normals:

$$\theta_t|(\mathcal{Y}_{1,n}, \mathbf{s}_{1,n}) \sim \alpha_t \delta_B + \sum_{1 \le i \le t \le j \le n} \beta_{ijt} N(\mu_{ij}, V_{ij}).$$

where $\alpha_t, \beta_{ijt}, \mu_{ij}, V_{ij}$ can be explicitly computed via recursion.

Smoothing equations

If s_t were known, the posterior distribution of θ_t given the complete data sequence is a mixture of normals:

$$\theta_t|(\mathcal{Y}_{1,n}, \boldsymbol{s}_{1,n}) \sim \alpha_t \delta_B + \sum_{1 \le i \le t \le j \le n} \beta_{ijt} N(\mu_{ij}, V_{ij}).$$

where $\alpha_t, \beta_{ijt}, \mu_{ij}, V_{ij}$ can be explicitly computed via recursion.

Then we can estimate θ_t by its posterior mean.

$$E(\theta_t|\mathcal{Y}_{1,n}, \boldsymbol{s}_{1,n}) = \alpha_t B + \sum_{1 \le i \le t \le j \le n} \beta_{ijt} \mu_{ij}.$$

Smoothing equations

If s_t were known, the posterior distribution of θ_t given the complete data sequence is a mixture of normals:

$$\theta_t|(\mathcal{Y}_{1,n}, s_{1,n}) \sim \alpha_t \delta_B + \sum_{1 \le i \le t \le j \le n} \beta_{ijt} N(\mu_{ij}, V_{ij}).$$

where $\alpha_t, \beta_{ijt}, \mu_{ij}, V_{ij}$ can be explicitly computed via recursion.

Then we can estimate θ_t by its posterior mean.

$$E(\theta_t|\mathcal{Y}_{1,n}, s_{1,n}) = \alpha_t B + \sum_{1 \le i \le t \le j \le n} \beta_{ijt} \mu_{ij}.$$

Using BCMIX approximations (Lai et al., 2005), this smoothing step can be done in $O(n)$ computations, where n is number of SNPs. This translates to $\sim$ 15 seconds for 30000 SNPs (4 minutes for 500K chip).

A Twist on the EM Algorithm

Set s^1 to some initial value. Let $i = 1$.

A Twist on the EM Algorithm

Set s^1 to some initial value. Let $i = 1$. Repeat:

1. Expectation step: Given s^i, set θ_t^i to its posterior mean.

A Twist on the EM Algorithm

Set s^1 to some initial value. Let $i = 1$. Repeat:

1. Expectation step: Given s^i, set θ_t^i to its posterior mean.
2. Maximization step: Given θ^i, set s^{i+1} to its maximum a posteriori value

$$s^{i+1} = \arg\max_{s \in \mathcal{S}} P(s|\theta^i, y). \tag{3.1}$$

This can be done easily because given θ^i, y_t is a mixture of Gaussians at each t, and s_t^{i+1} is simply the identifier for each mixture component.

A Twist on the EM Algorithm

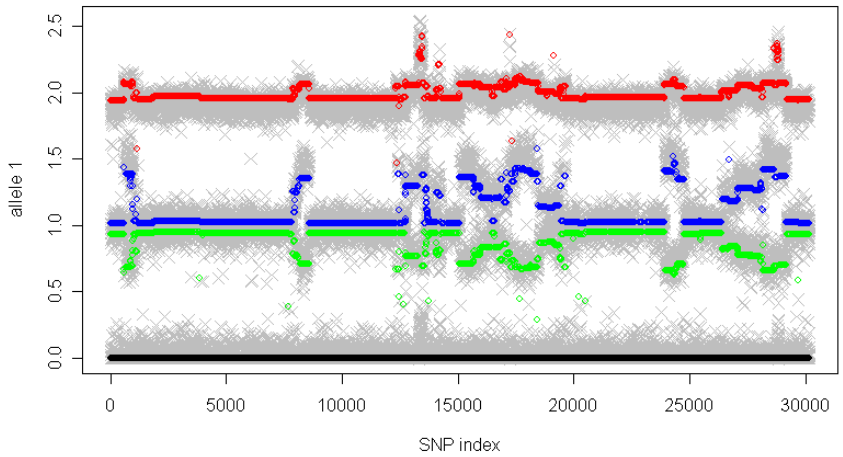
Set s^1 to some initial value. Let $i = 1$. Repeat:

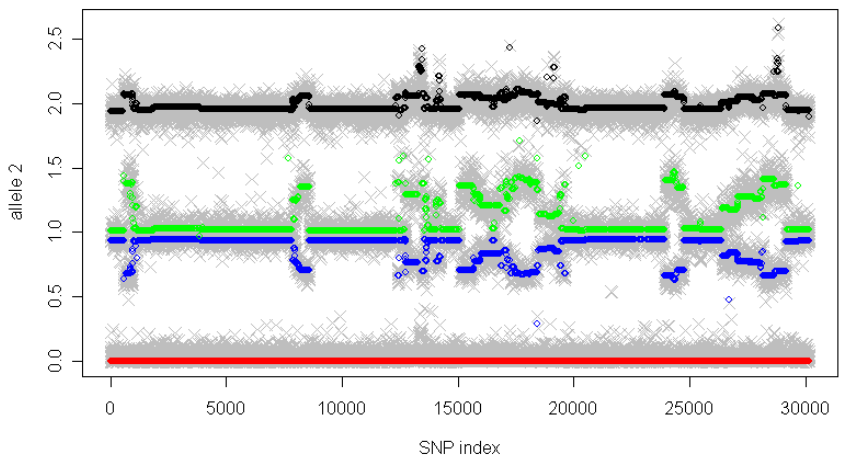
1. Expectation step: Given s^i, set θ_t^i to its posterior mean.
2. Maximization step: Given θ^i, set s^{i+1} to its maximum a posteriori value

$$s^{i+1} = \arg\max_{s \in \mathcal{S}} P(s|\theta^i, y). \tag{3.1}$$

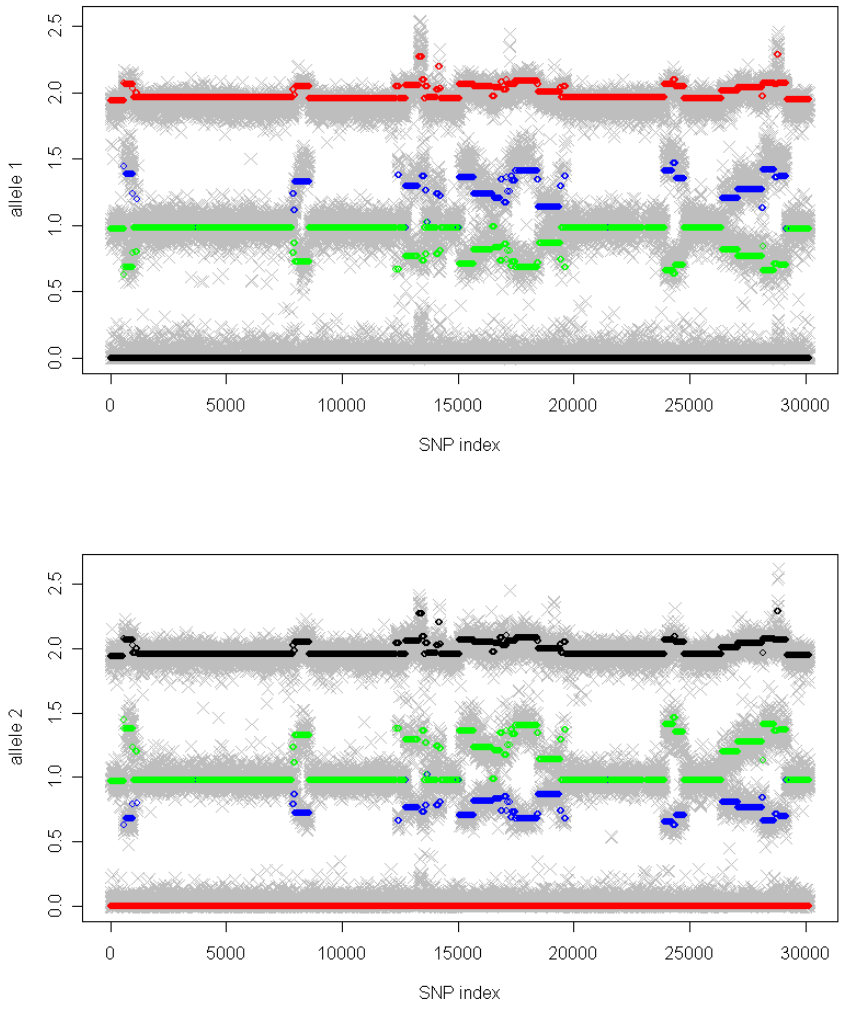
   This can be done easily because given θ^i, y_t is a mixture of Gaussians at each t, and s_t^{i+1} is simply the identifier for each mixture component.
3. If $\|\theta^{i+1} - \theta^i\| < \delta$, stop and report θ^{i+1}, s^{i+1}. Otherwise, set $i \leftarrow i + 1$ and go back to step 1.

Smoothed data: birds eye view



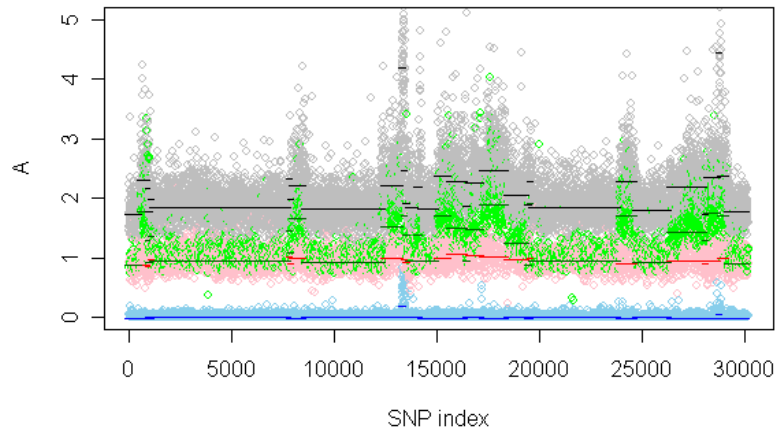


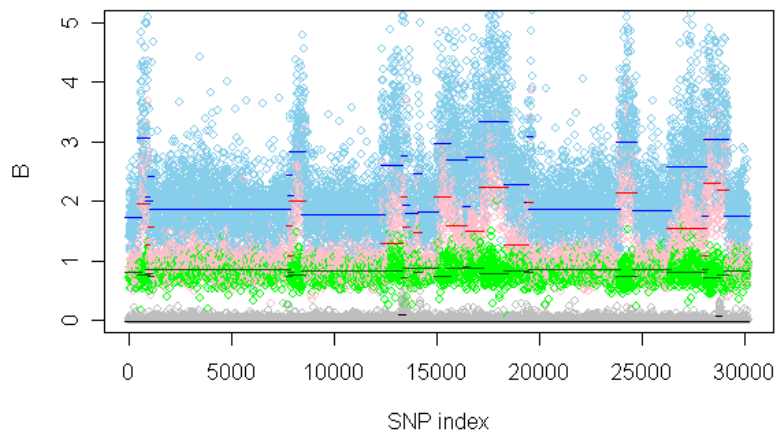
Segmented data: birds eye view



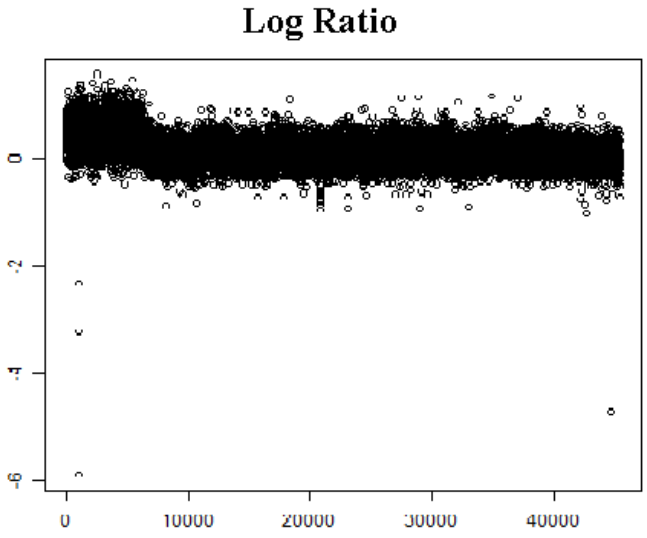
Segmentation by thresholding: if difference $> \delta$ then segment.
Plus other rules for min SNPs, min heterozygosity, etc.

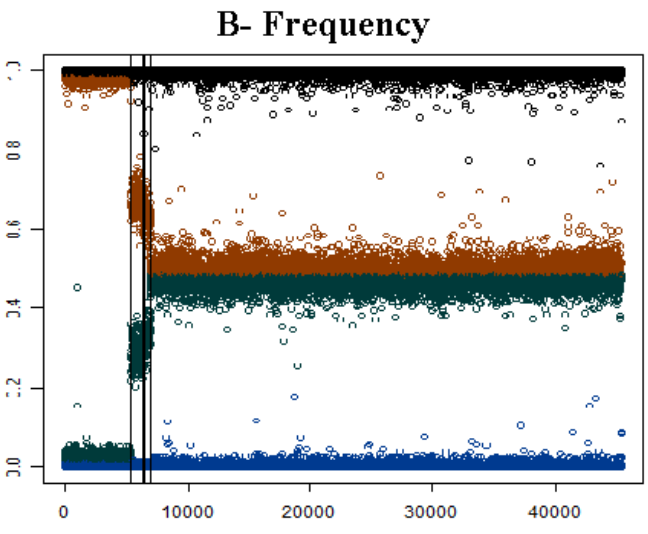
Transform back to Illumina AB Coordinates

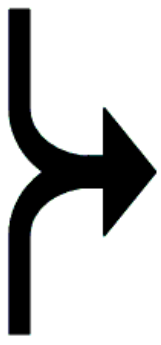


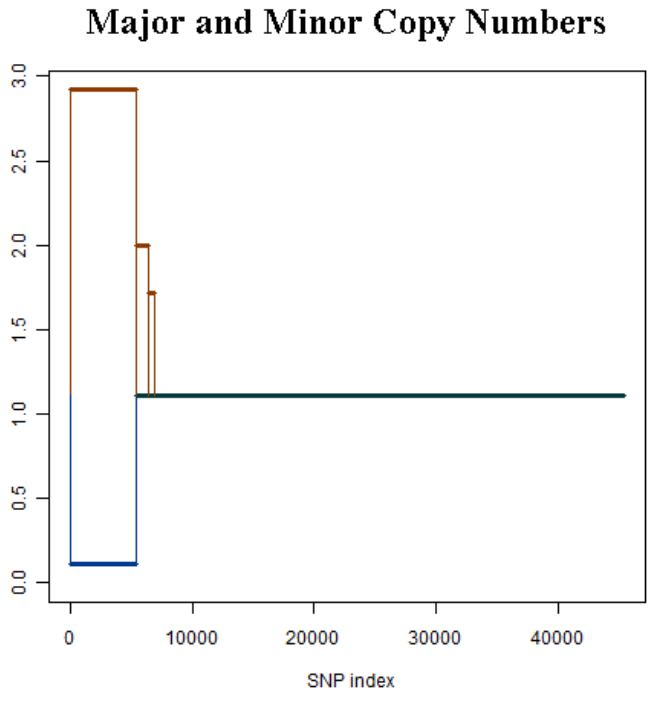


End Result: Estimates of Major and Minor Copy Numbers

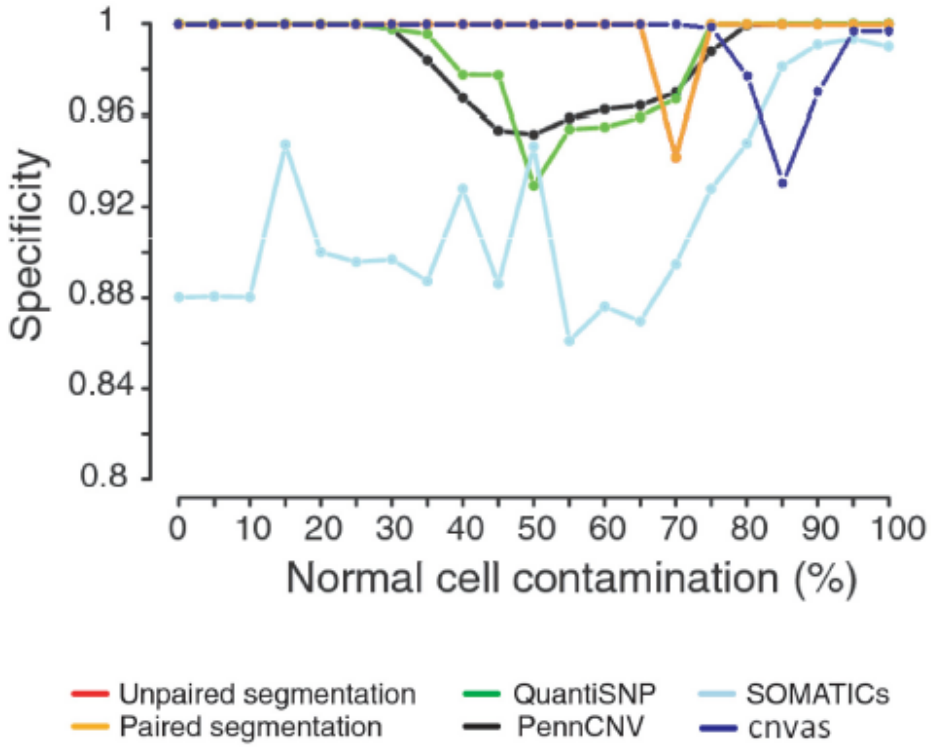




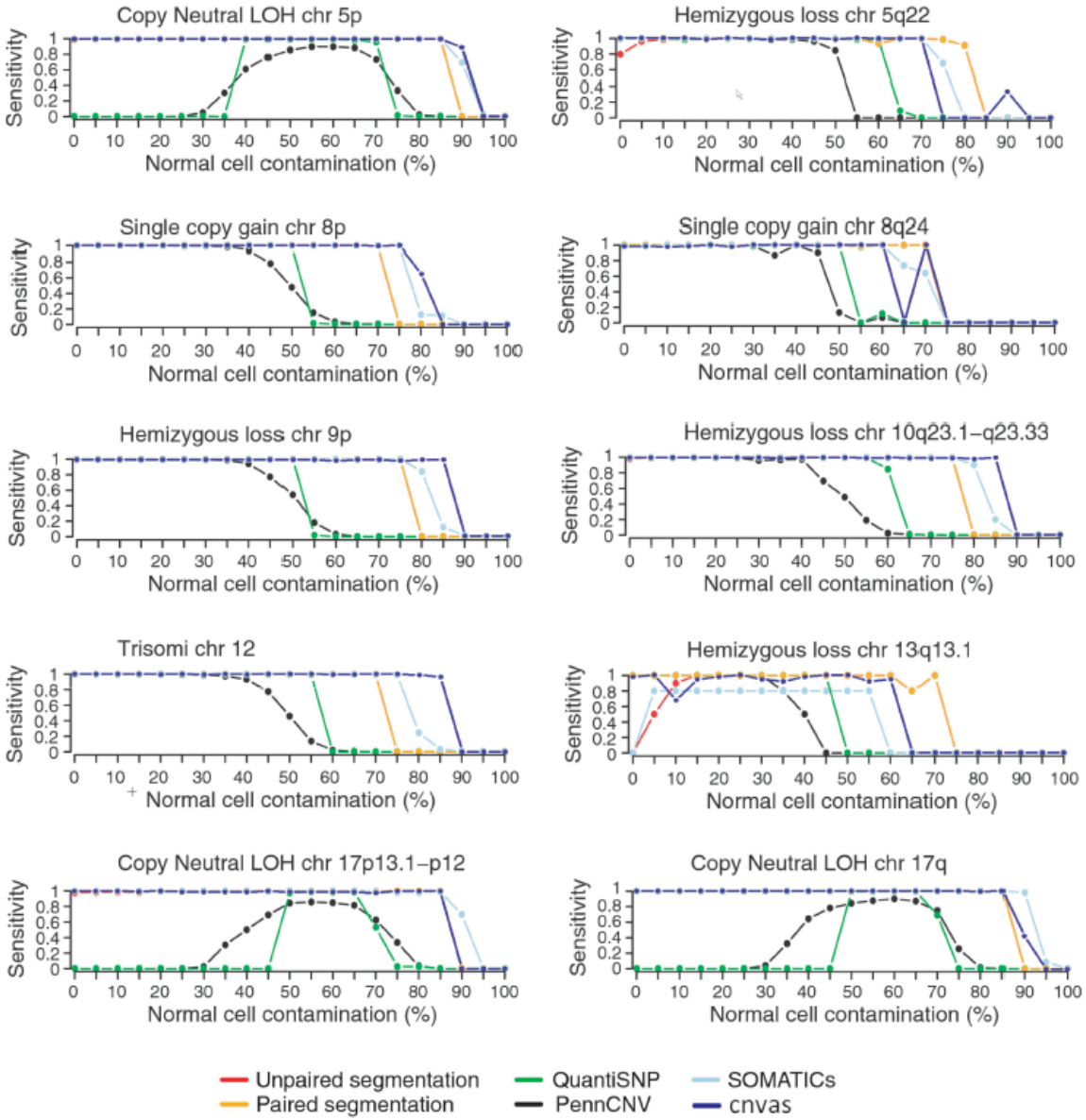




Specificity using simulated titration data of Staaf et al. (2008)



Sensitivity using simulated titration data of Staaf et al. (2008)

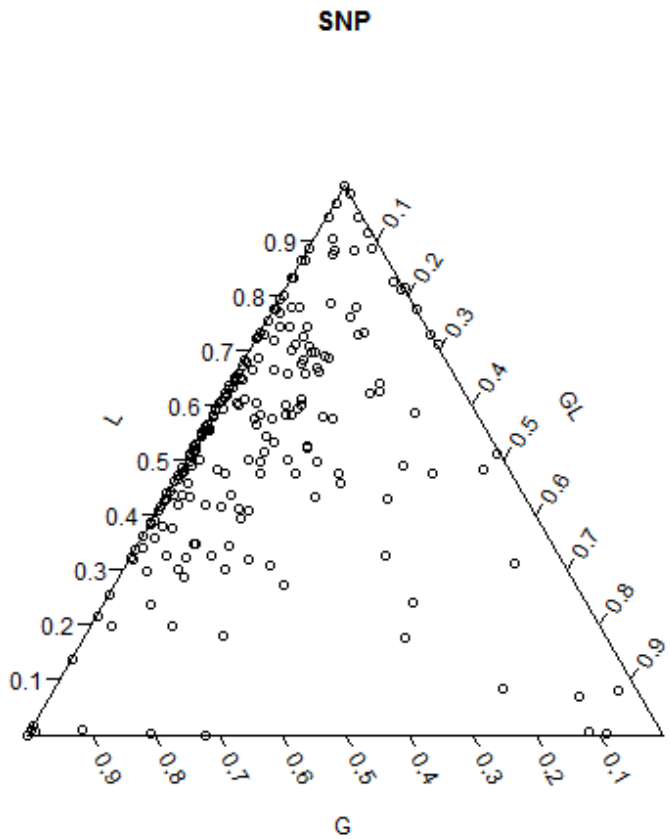


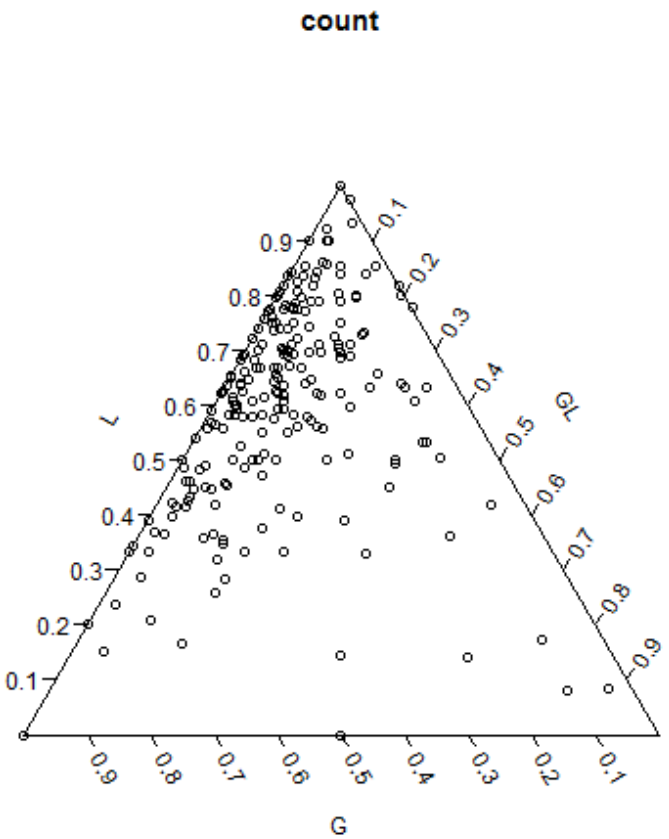
Distribution of types of aberrations across patients

1. 223 Glioblastoma samples studied by the TCGA Research Network.
2. Stringent cut offs were used to identify CNAs.
3. Percentage of bases in each of the aberration categories:

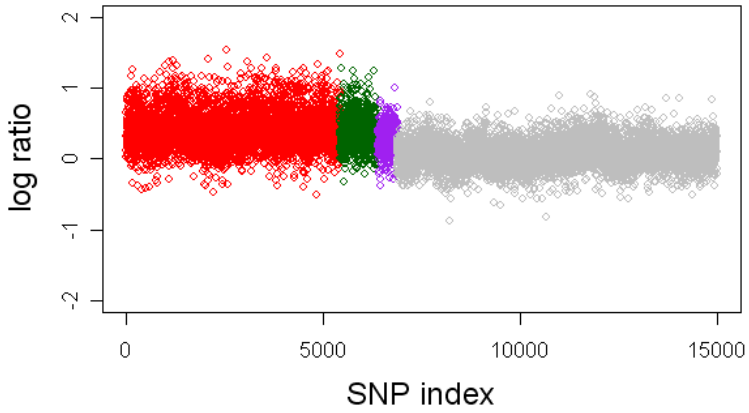
Event type	%
Gain/Gain	3
Gain/Normal	28
Gain/Loss	15
Loss/Normal	51
Loss/Loss	3

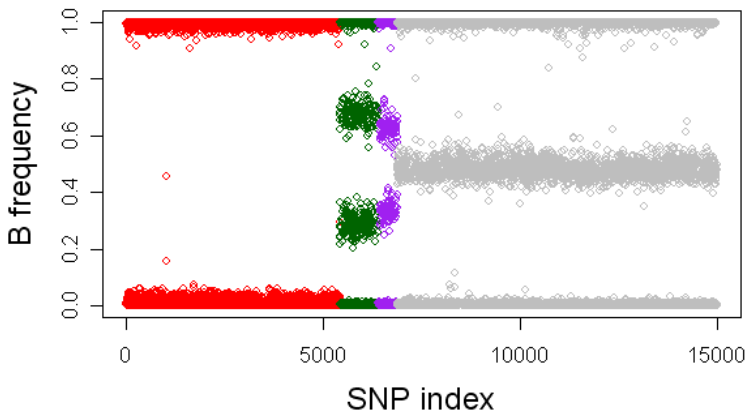
Variation of Aberration Profiles Across Patients



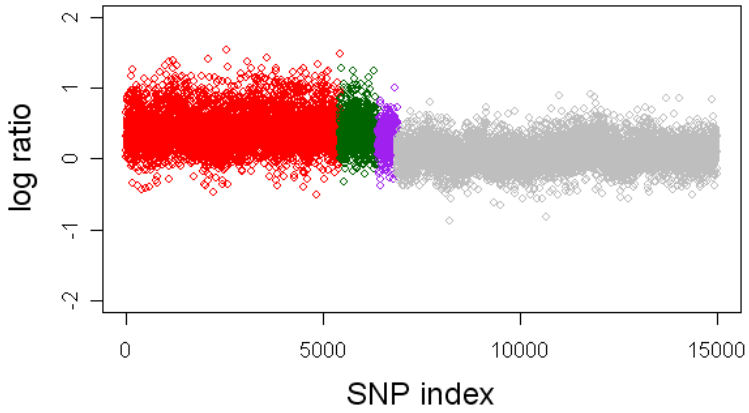


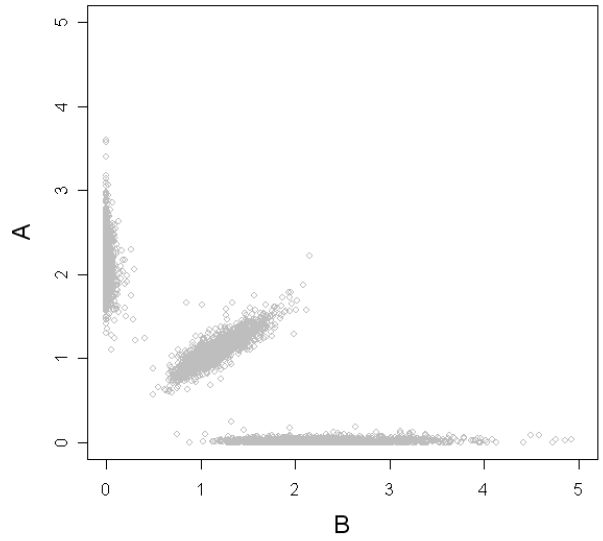
Example Regions

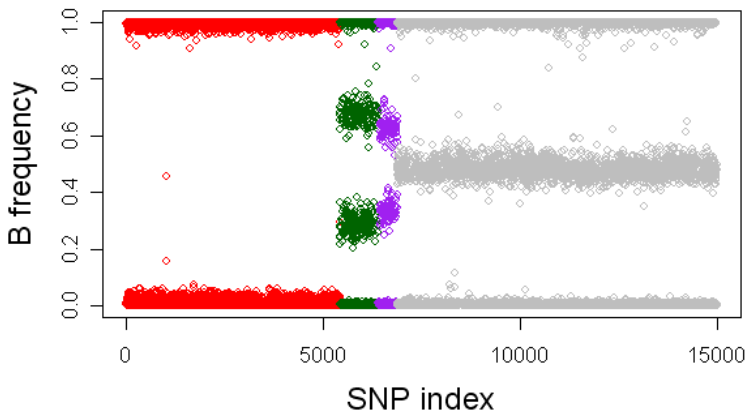




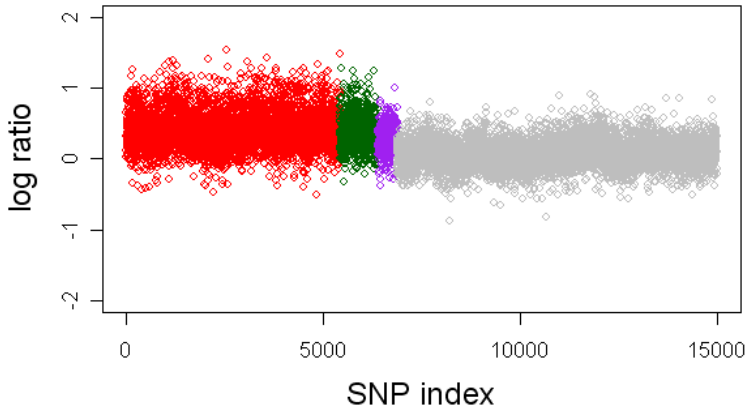
Example Regions

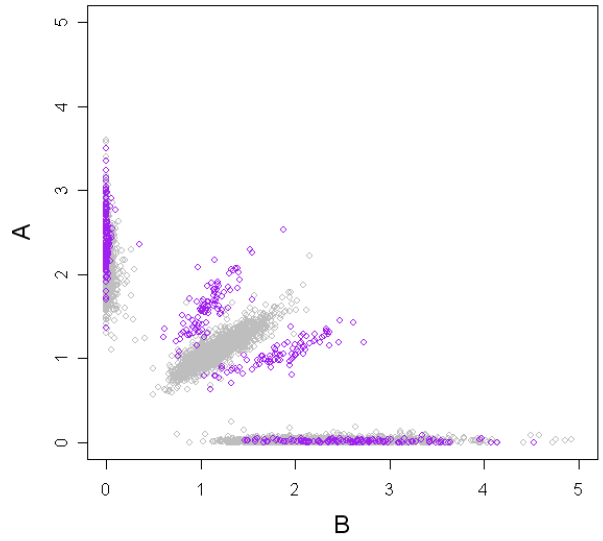


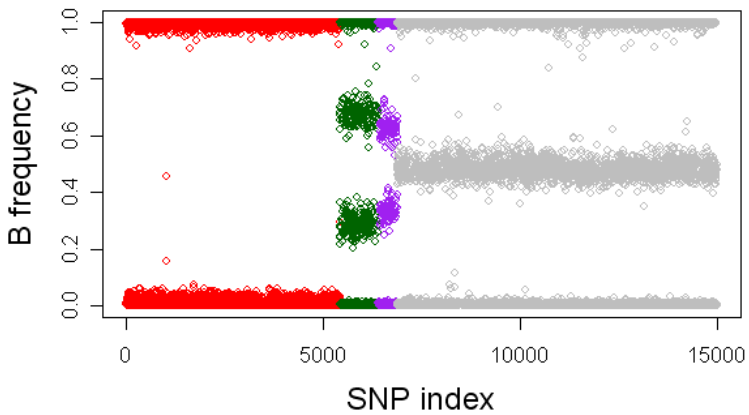




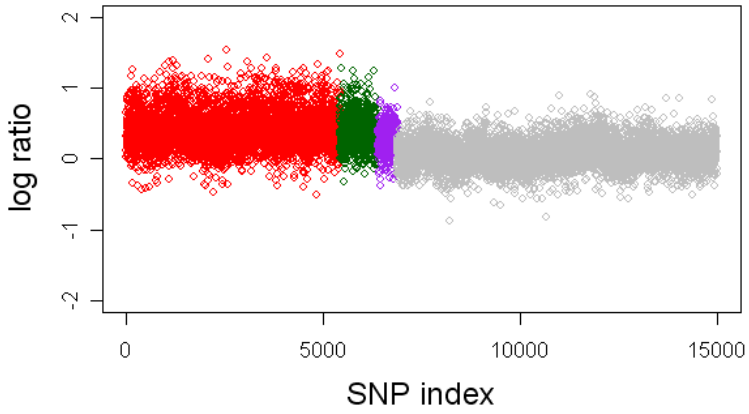
Example Regions

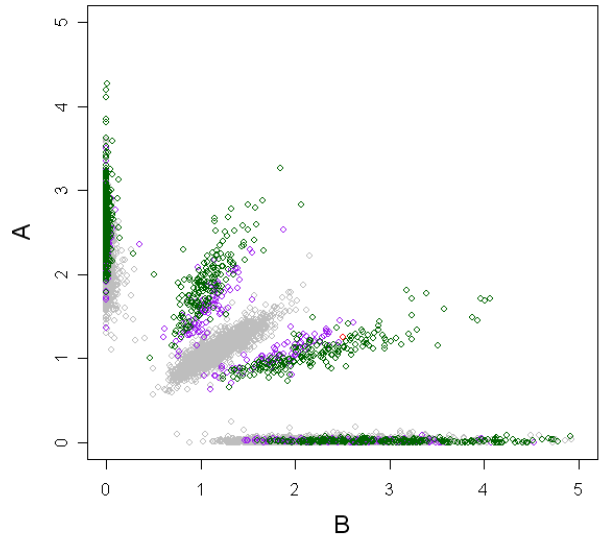


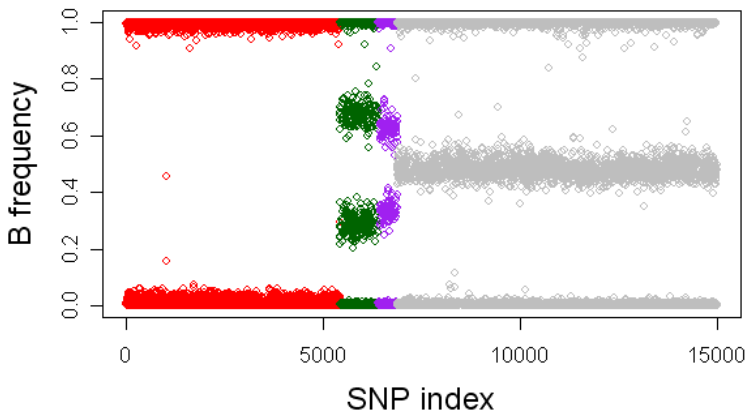




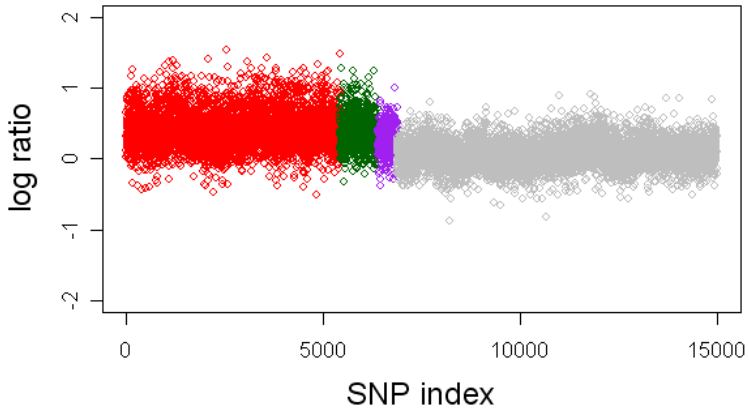
Example Regions

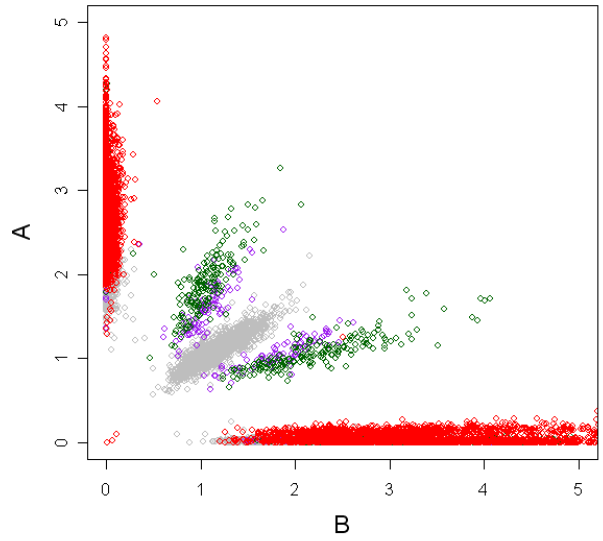


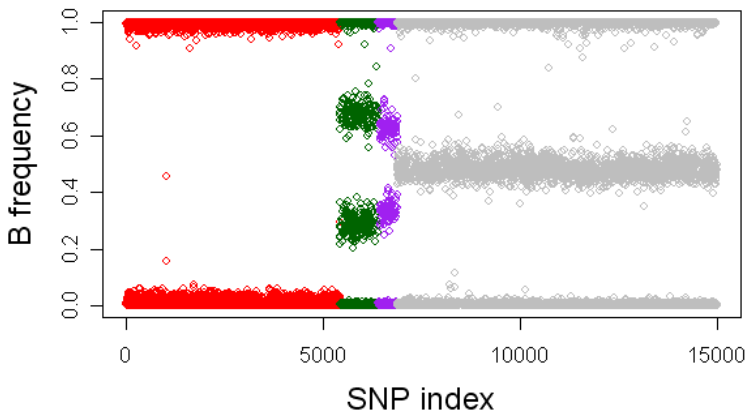




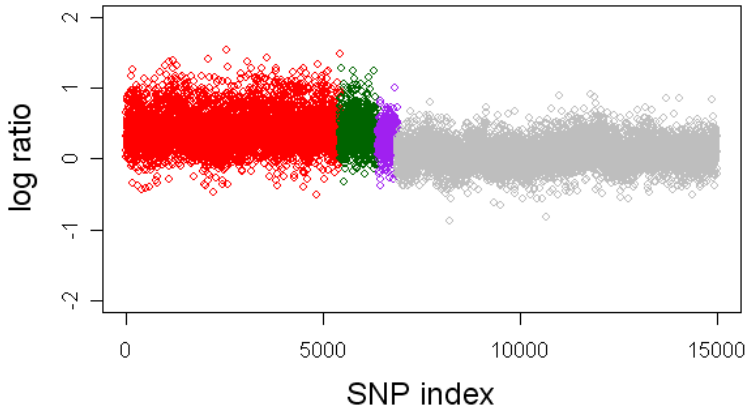
Example Regions

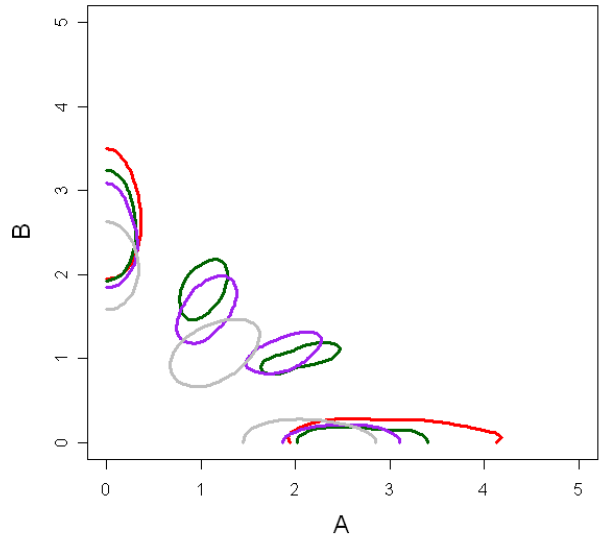


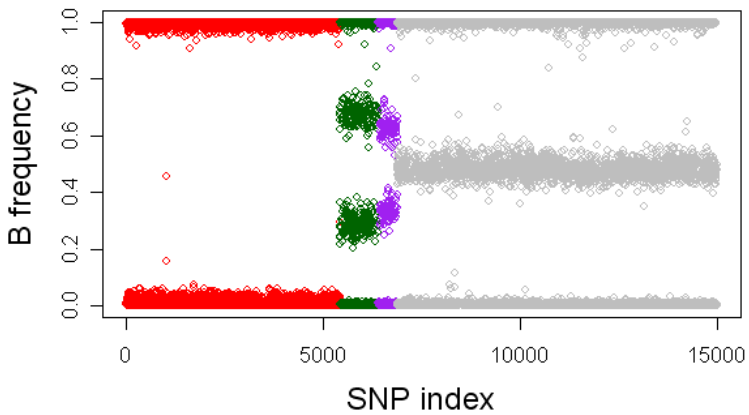




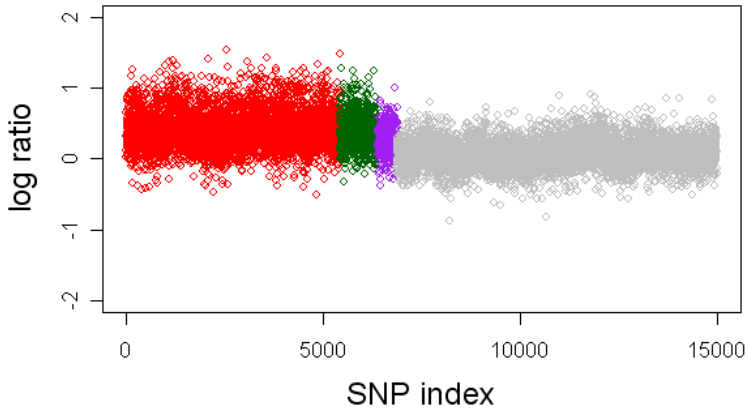
Example Regions

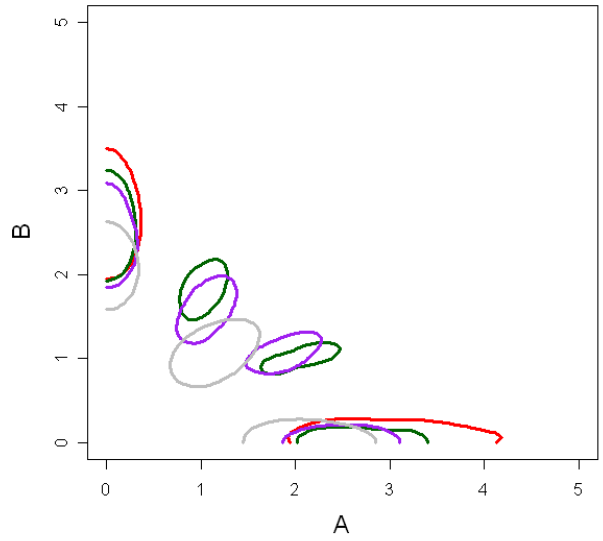


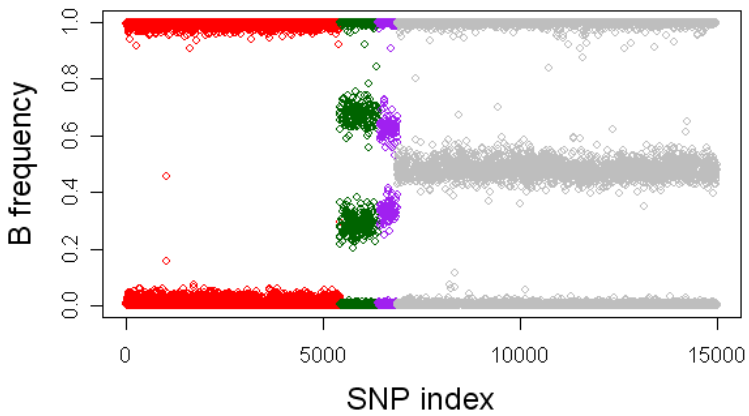


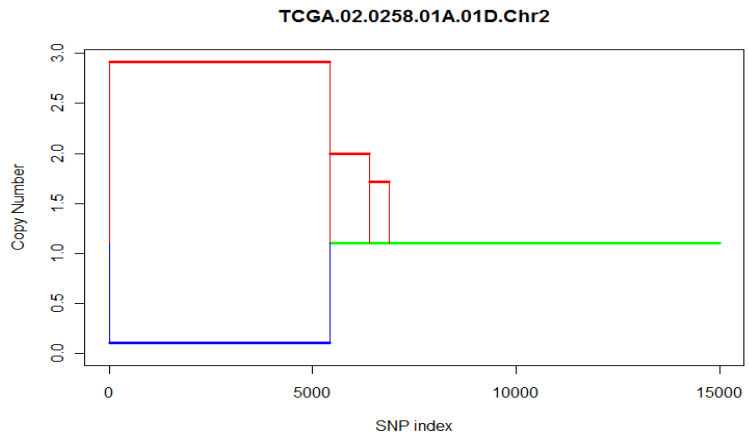


Example Regions



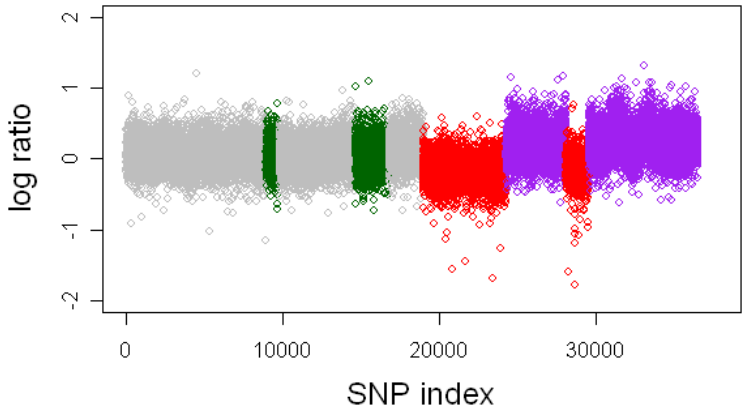


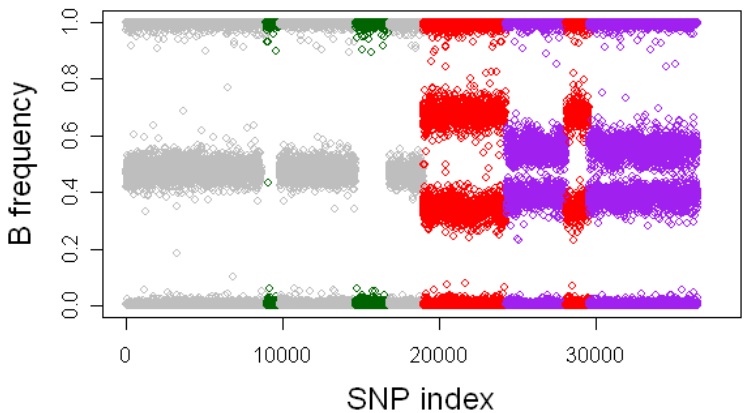




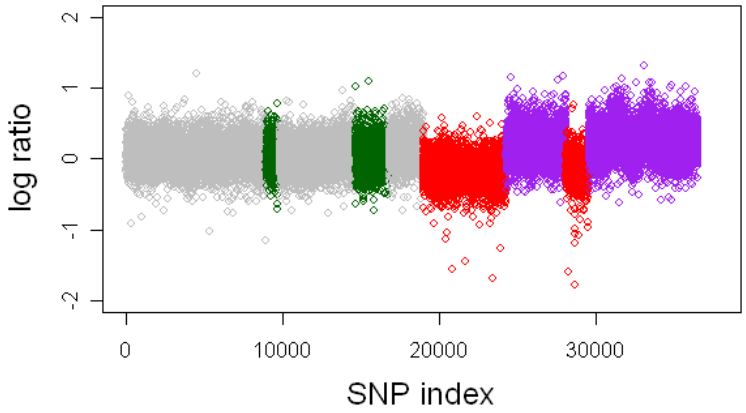
Unbalanced gain/loss, followed by fractional gain.

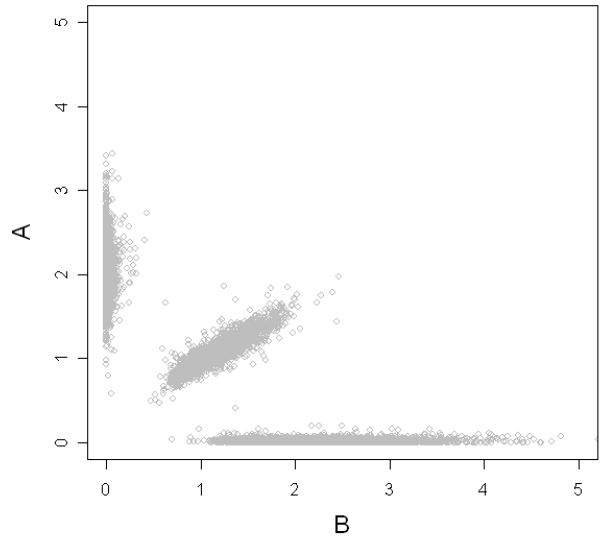
Example Regions

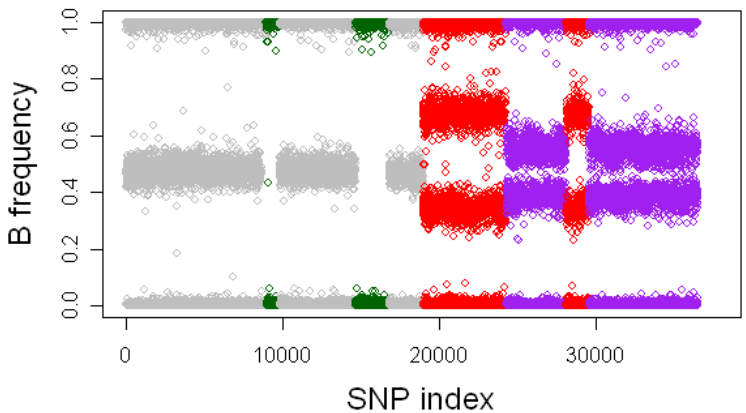




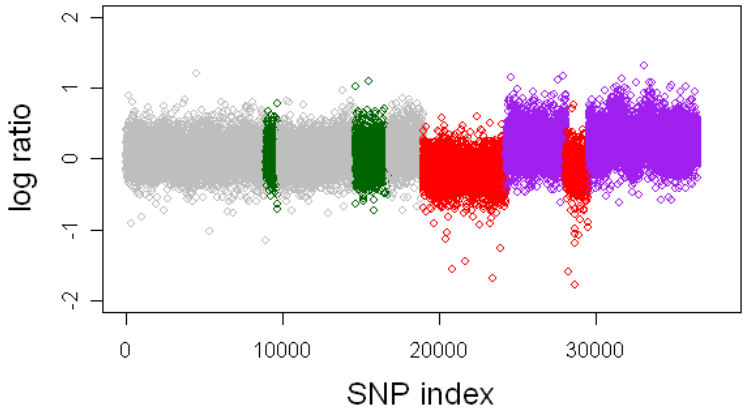
Example Regions

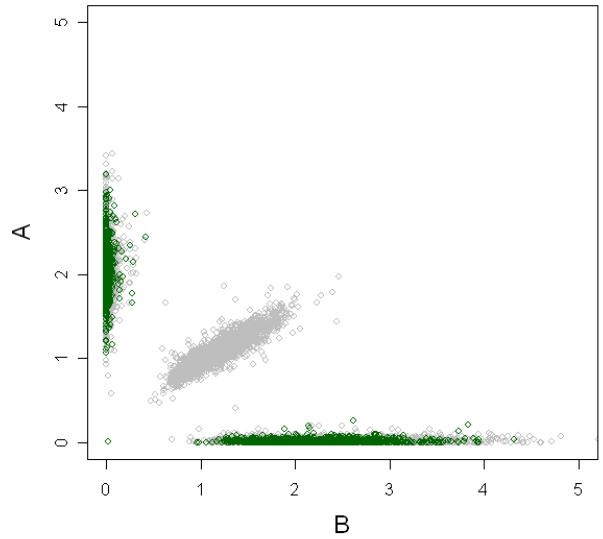


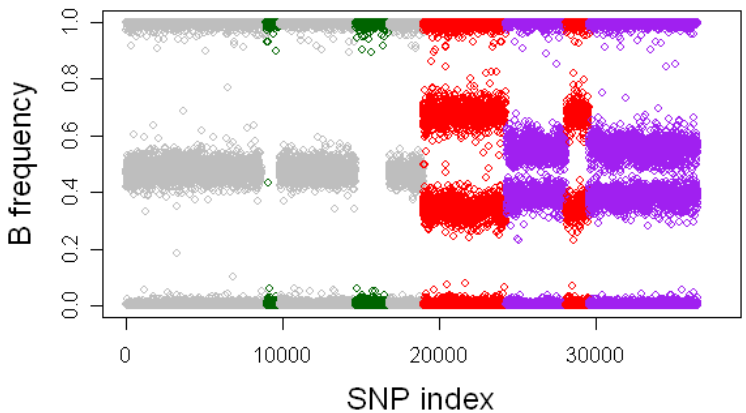




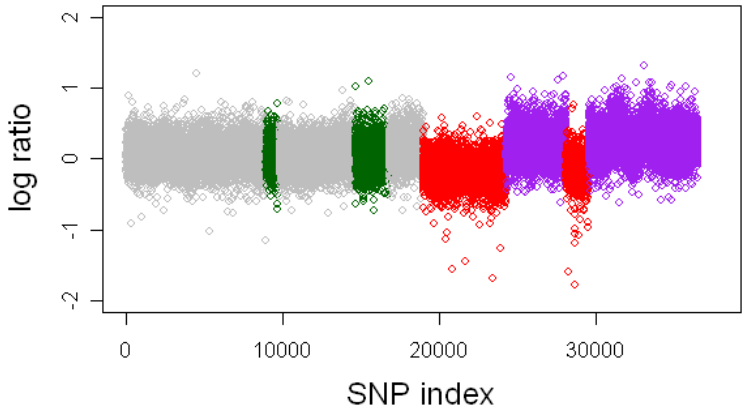
Example Regions

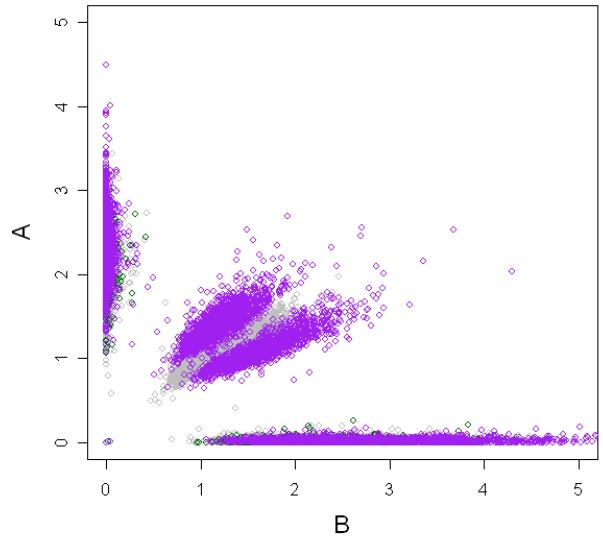


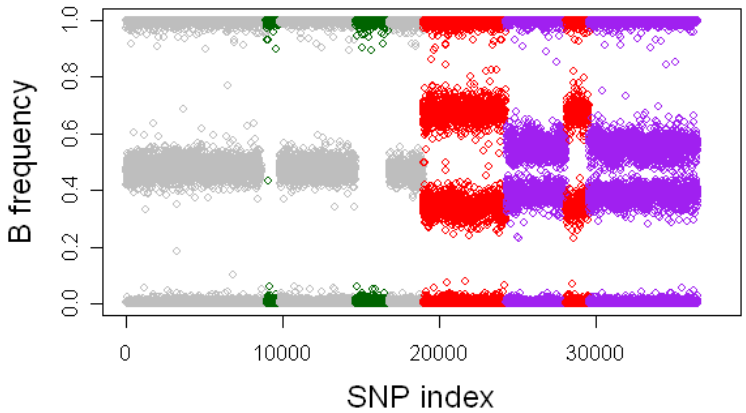




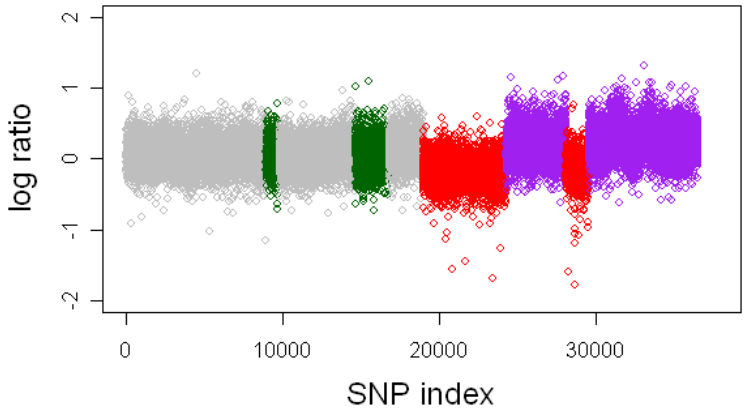
Example Regions

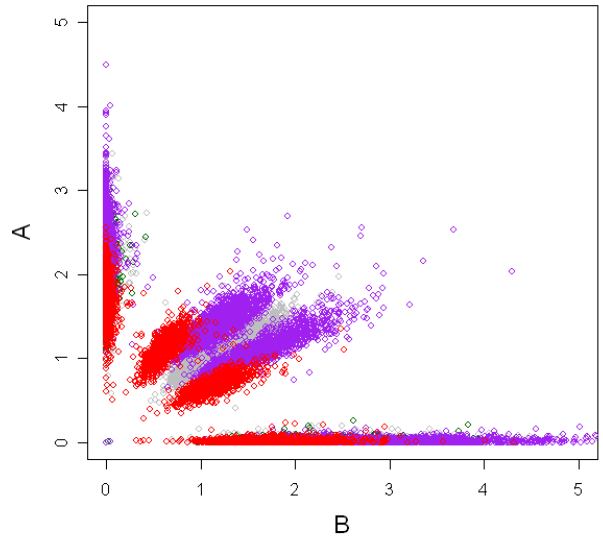


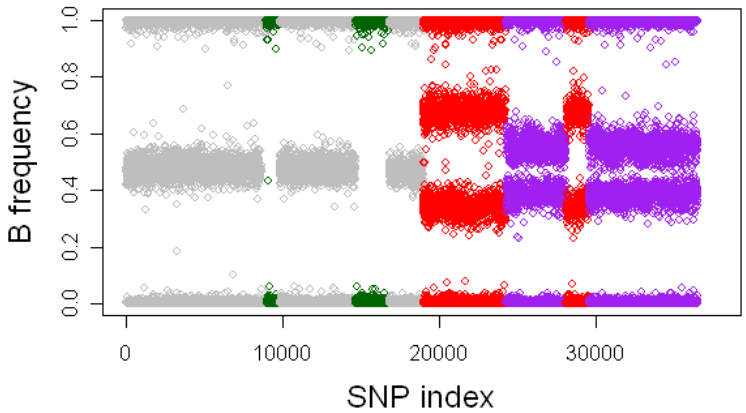




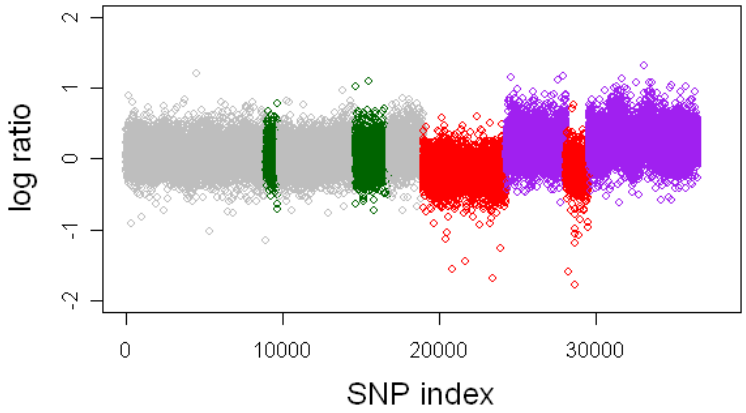
Example Regions

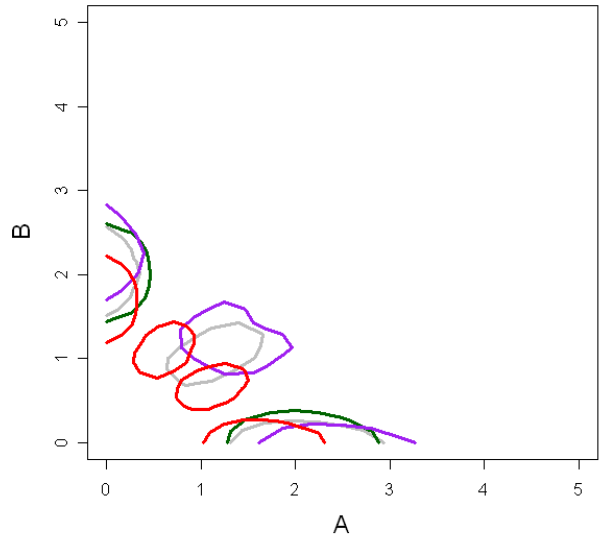


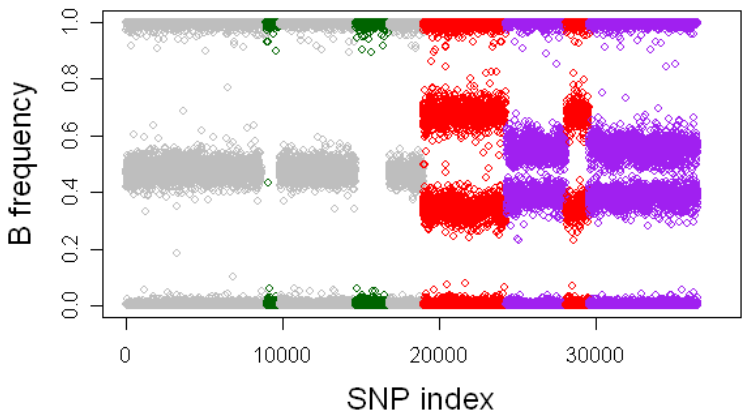




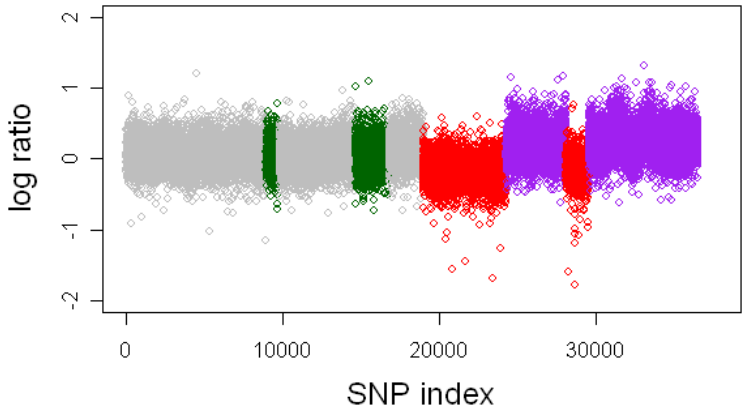
Example Regions

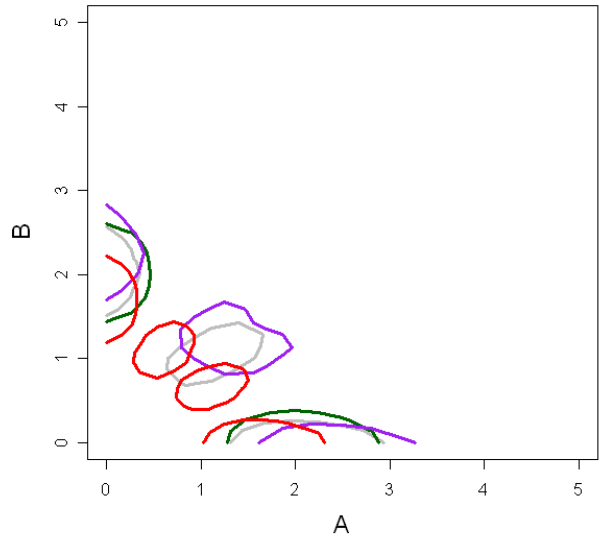


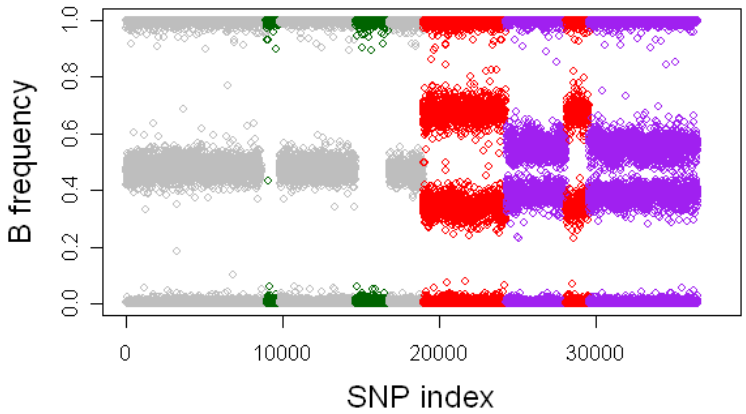


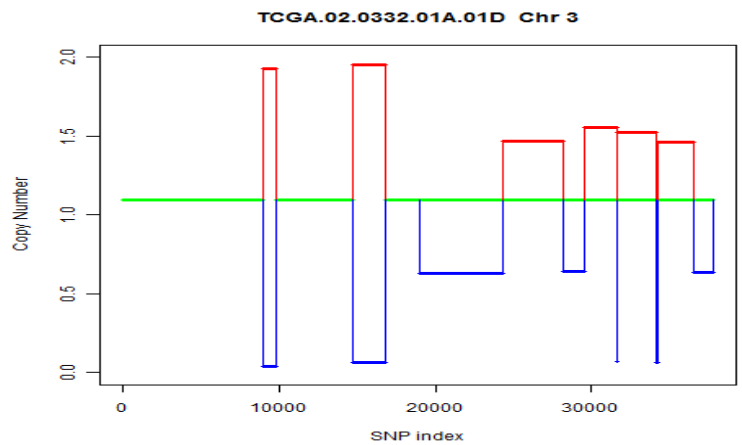


Example Regions









Copy neutral loss of heterozygosity, followed by alternating fractional gain loss.

What's next?

We can now segment a genome into regions of homogeneous parent-specific copy number. For each region, this gives estimates of the relative quantities of the major and minor chromosome.

What's next?

We can now segment a genome into regions of homogeneous parent-specific copy number. For each region, this gives estimates of the relative quantities of the major and minor chromosome.

This makes possible many new types of analyses:

1. Estimate normal cell contamination?
2. Quantify clonality?
3. Cross-sample analysis to identify allele-specific gains/losses to distinguish between passenger and driver mutations.