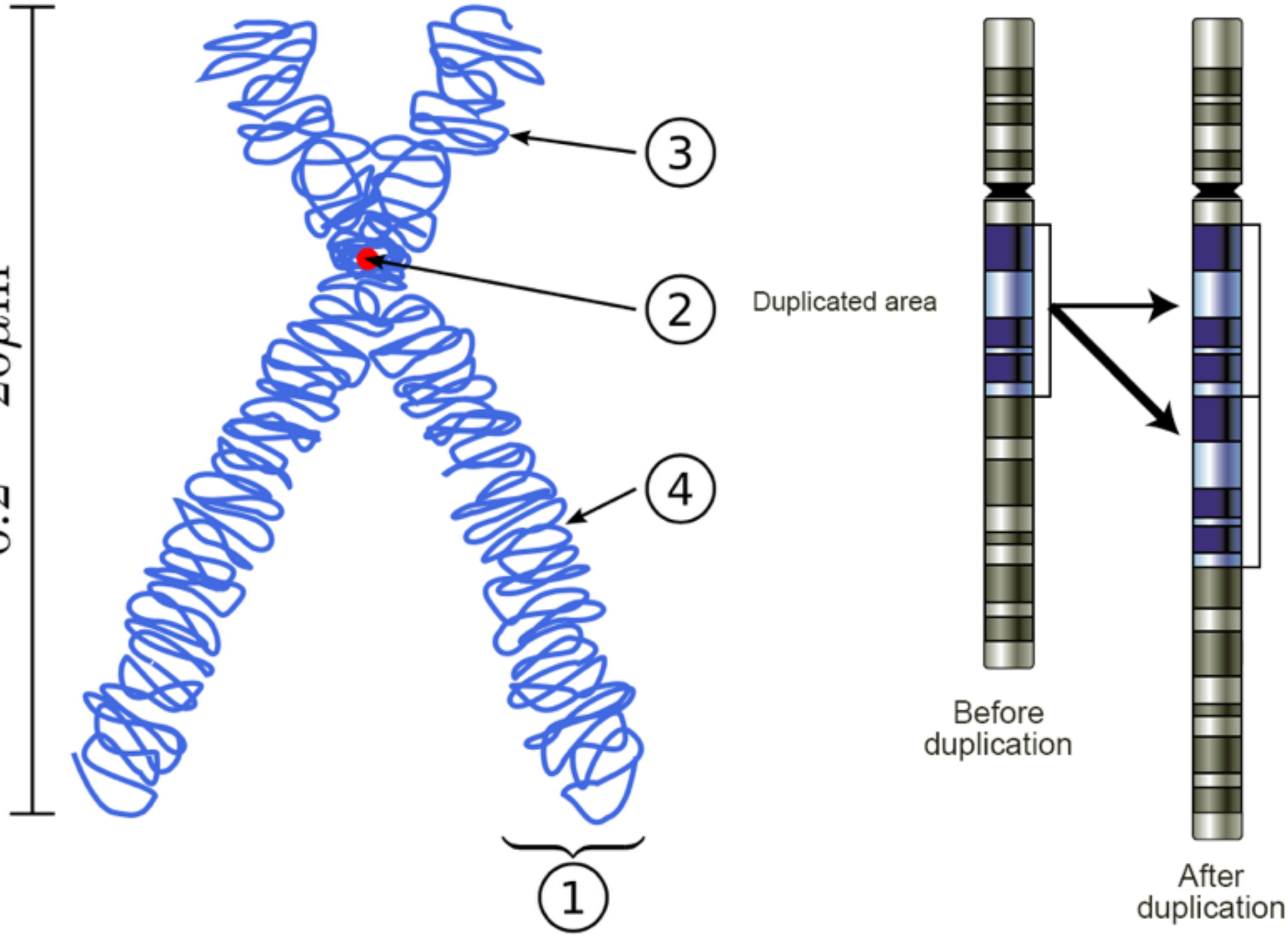
GISTIC



Somatic Copy Number Alterations

SCNAs

Individual .CEL Files

Step 1:
Accurate definition
of the copy number
profile in each sample

Array Calibration
Copy Number Estimation
Segmentation

Segmented Copy Number Profiles

Step 2:
Identification/separation
of underlying SCNAs

Elimination of arm-level SCNAs by
use of amplitude threshold

Deconstruction of segmented profile
into underlying SCNAs
Allows for modelling of background
rate of SCNAs and length-based
separation of arm-level
and focal SCNAs

Step 3:
Scoring SCNAs in each
region according to
likelihood of occuring
by chance

G = frequency x amplitude

p-values computed by random
permutation of markers
across genome

G = -log(Probability | Background)

Scores computed on markers or genes

p-values computed by random
permutation of markers or bins
across genome

Step 4:
Defining independent
genomic regions
undergoing significant
levels of SCNA

Greedy segment peel-off algorithm

Iteratively subtracts segments
covering each peak and
rescores until no significant peaks
remain on chromosome

Arbitrated peel-off algorithm

Formalizes idea that segments can
have multiple targets by allowing
segment scores to be split
among multiple potential peaks
during peel-off

Step 5:
Accurate definition
of the copy number
profile in each sample

Leave-k-out

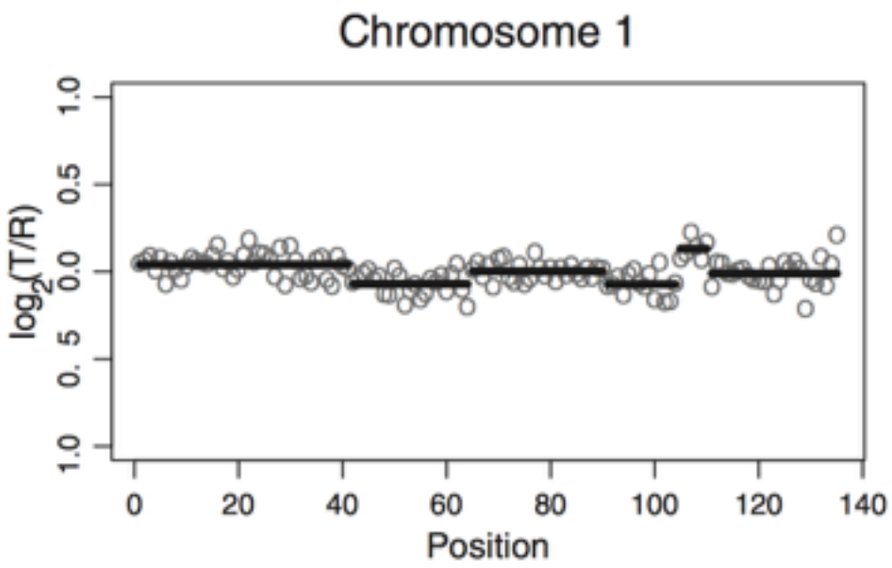
Assumes that at most ÔkÕ passenger
events aberrantly define the minimal
common region

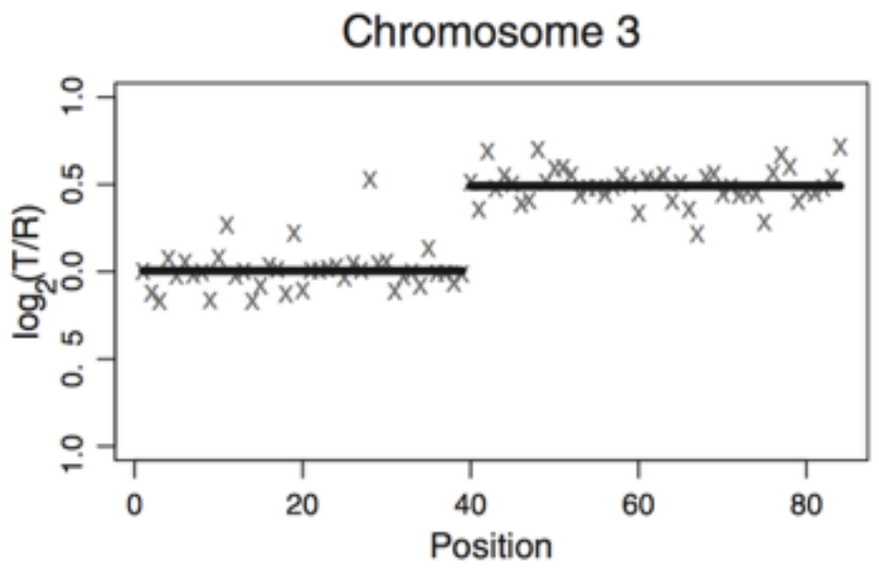
RegBounder

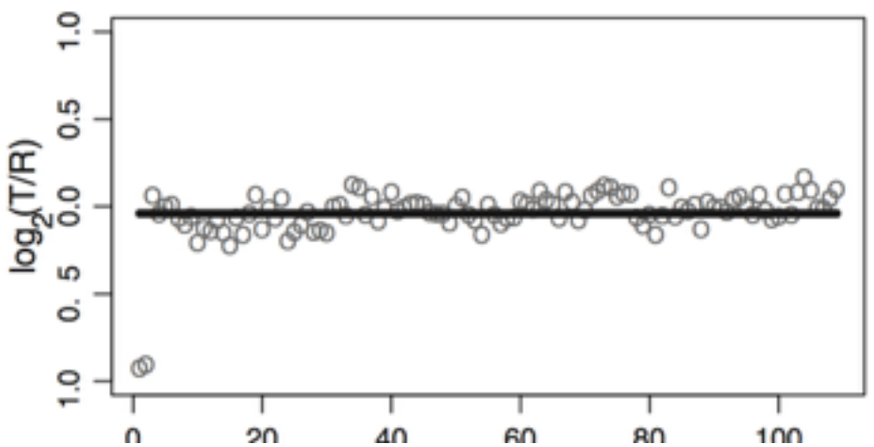
Models expected local variation in
G-score to define boundaries predicted
to contain the true target with
predetermined confidence

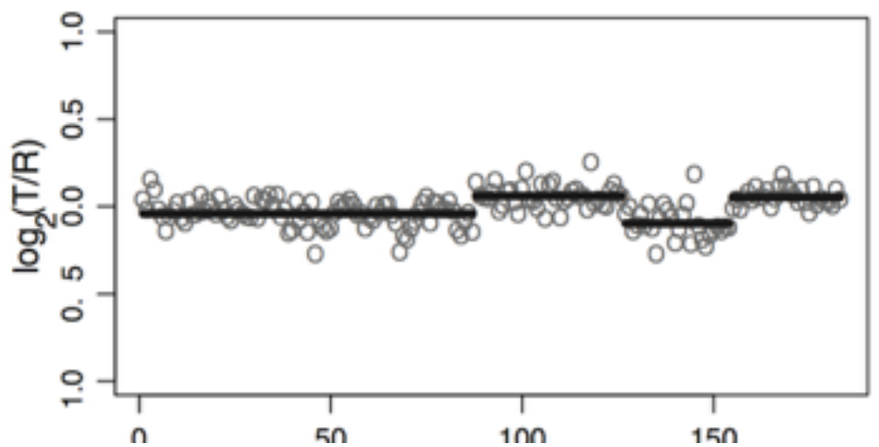
GISTIC 1.0 (Beroukhim et al, 2007)

GISTIC 2.0





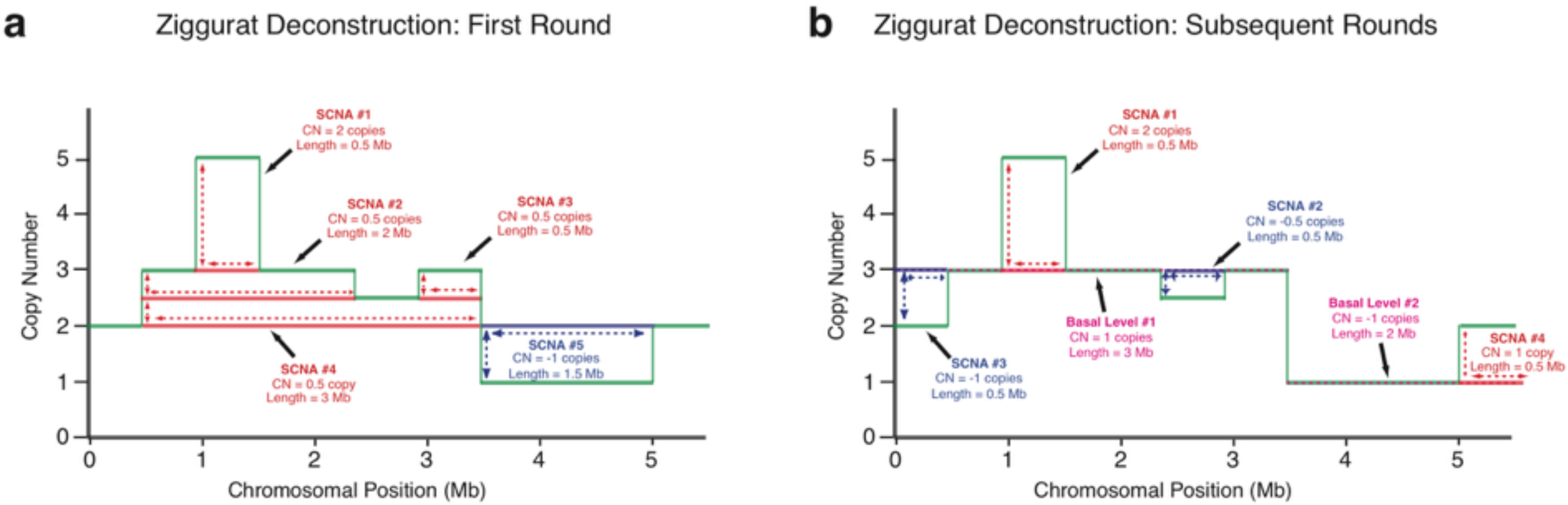




Step 1

Copy Number Segregation

Circular Binary Segregation (Olshen '04)



Step 2

Identification/Separation of Underlying SCAs

Ziggurat Deconstruction

BOX1: Segmentation Algorithm

Goal:

$$\sigma^* = \underset{\sigma}{argmax}\, Pr(x_1, x_2, ..., x_n|\sigma) + penalty(\sigma)$$

Given:

Probe level data $x_1, x_2, \ldots x_n$ and a proposed segmentation σ.

BOX2: Ziggurat Deconstruction

Goal:

$$\mathbf{h_c^*} = \underset{\mathbf{h_c}}{argmax}\, Pr(\sigma_c|\mathbf{h_c}) + penalty(\mathbf{h_c})$$

Given:

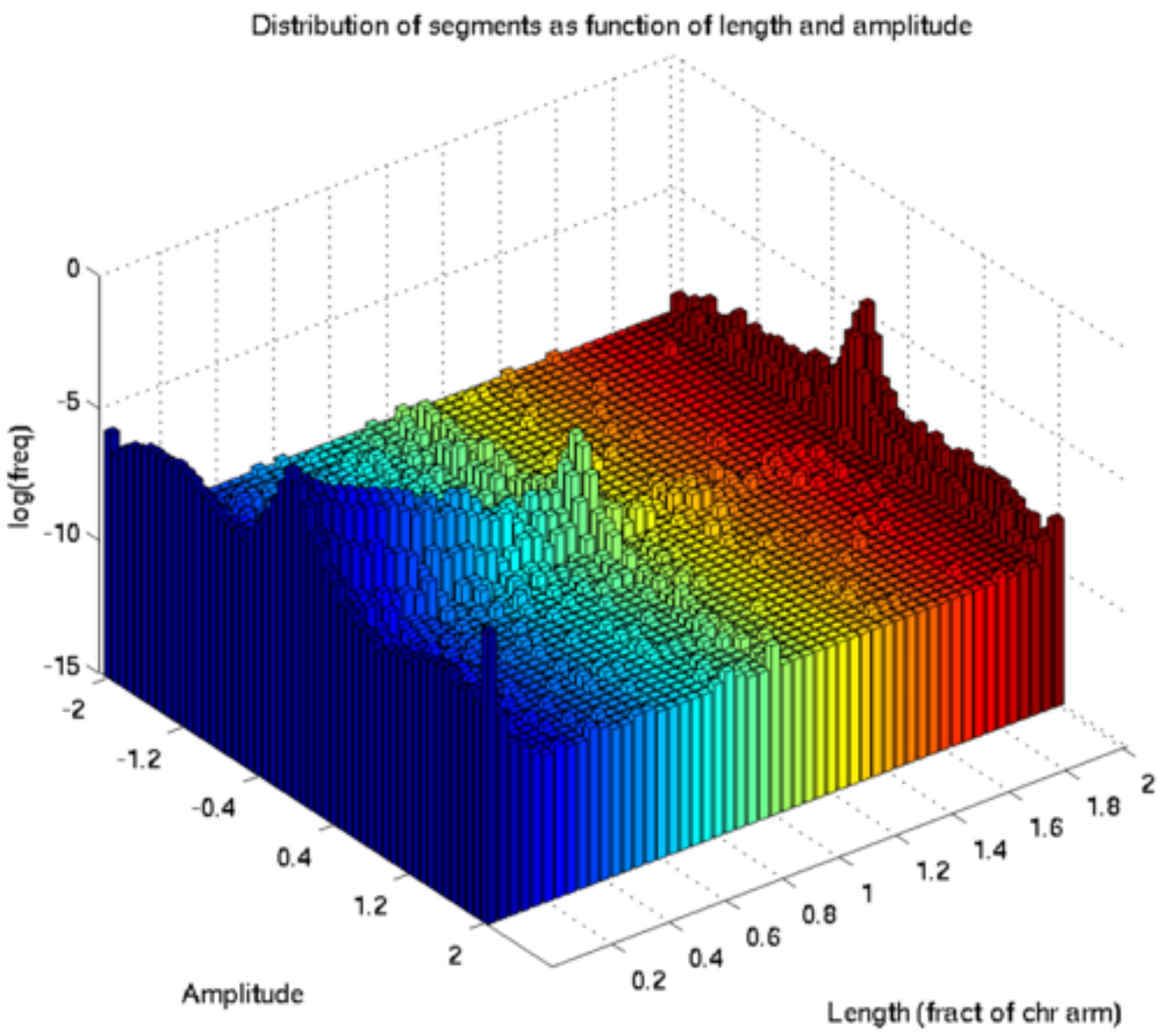
A chromosomal segmentation profile σ_c and proposed SCNA history $\mathbf{h_c}$

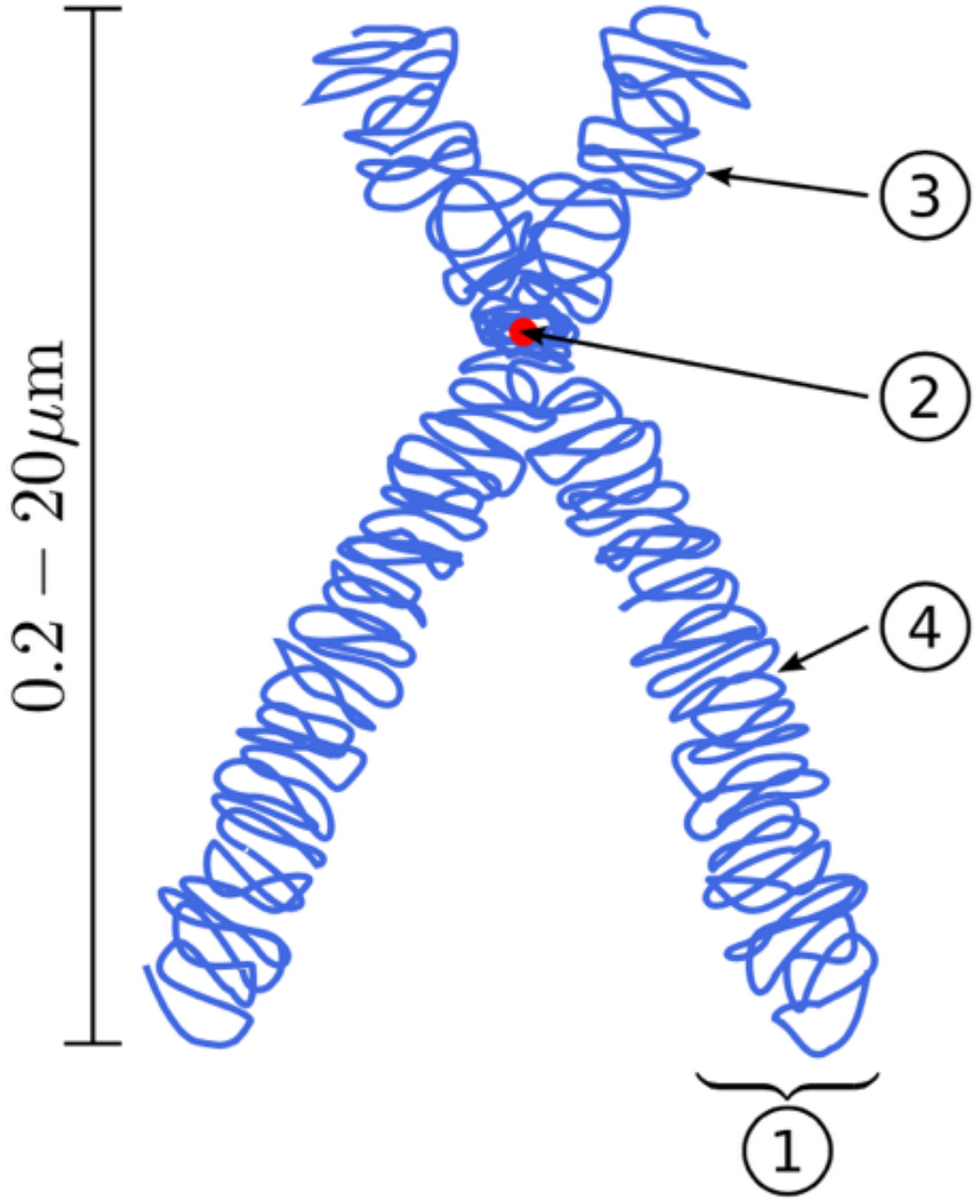
ZD performs this likelihood maximization by iterating between two complementary procedures:

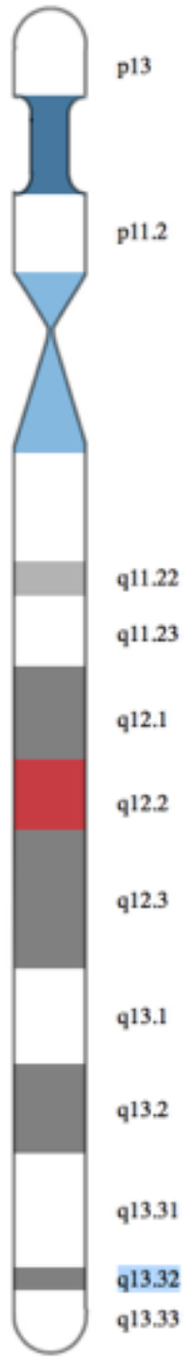
Deconstruction: Converts segmentation profiles into the most likely history of underlying SCNAs, using an estimate for the background rate of SCNAs as a function of length and amplitude (e.g. Pr(**e**) = f(*L*,*A*) for SCNA **e** of length *L* and amplitude *A*).

Background Estimation: Updates the background rate of SCNA formation (e.g. Pr(**e**) = f(*L*,*A*)) given the sequence of SCNAs inferred from the current deconstruction.

$$Pr(\sigma_c|\mathbf{h_c}) = \prod_{e_i \in \mathbf{h_c}} Pr(e_i) = \prod_{e_i \in \mathbf{h_c}} f(l_i, a_i)$$


Distribution of segments as function of length and amplitude
log(freq)
0
-5
-10
-15
Amplitude
-2
-1.2
-0.4
0.4
1.2
2
Length (fract of chr arm)
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

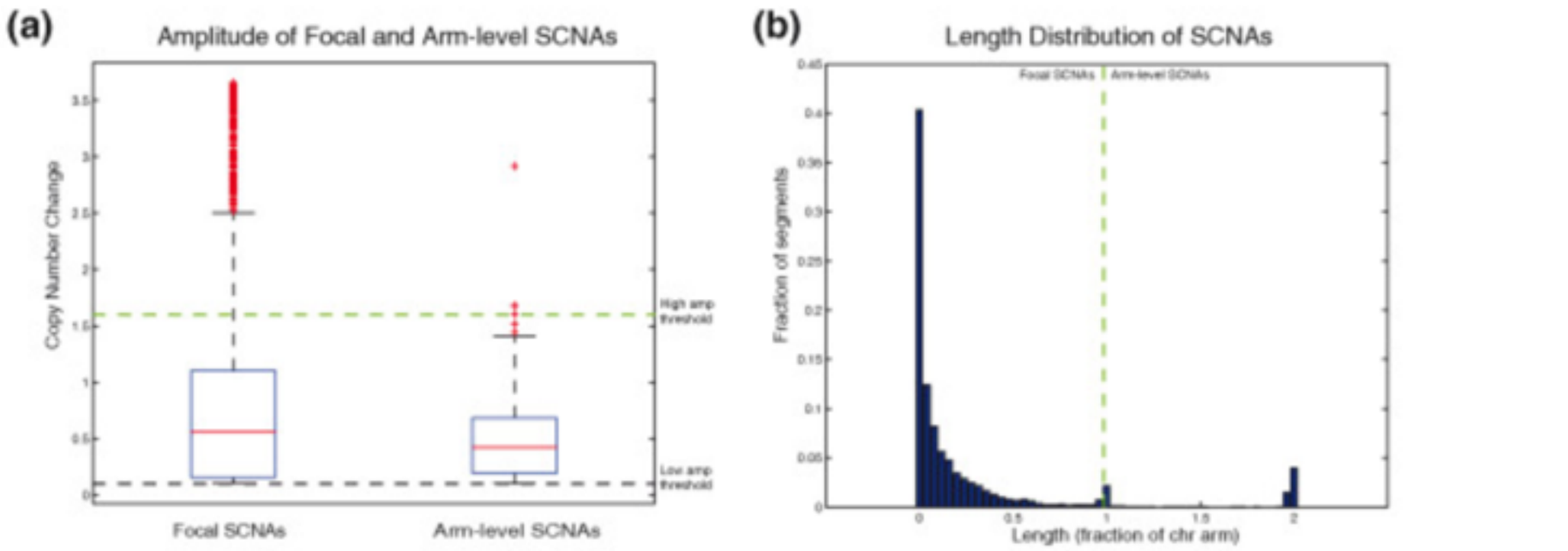




Length-based Separation of Focal and Arm-Level SCNAs

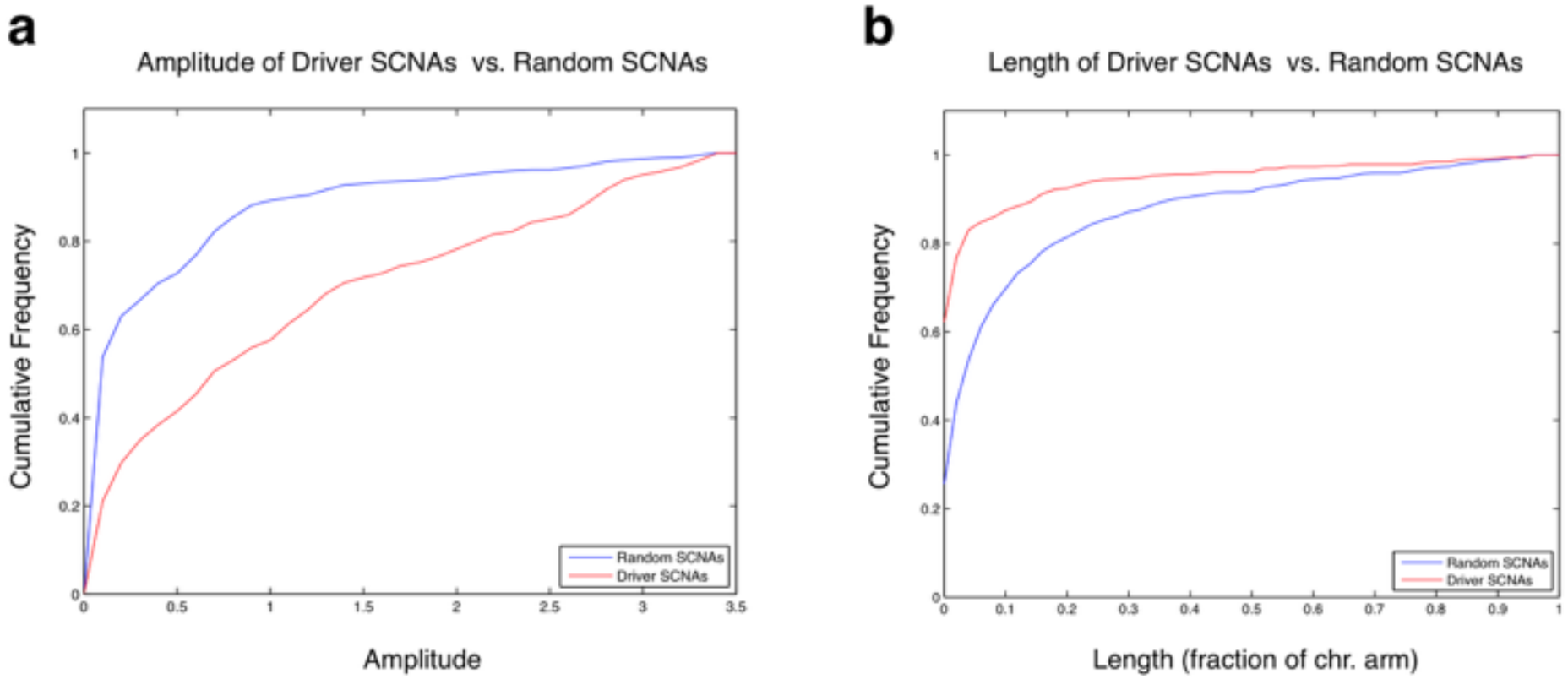
Figure 2.

Resolution: **standard** / high



Computational separation of arm-level and focal SCNAs. **(a)** Boxplot showing the distribution of copy-number changes for amplified focal (length < 98% of a chromosome arm) and arm-level (length > 98% of a chromosome arm) SCNAs across 178 GBM profiles from TCGA. The black dotted line denotes a typical low-level amplitude threshold used to eliminate artifactual SCNAs, while the green dotted line denotes a typical high-level amplitude threshold used in previous version of GISTIC to eliminate arm-level SCNAs. **(b)** Histogram showing the frequency of observing SCNAs of a given length across 178 GBM samples. The high frequency of events occupying exactly one chromosome arm led us to distinguish between focal and arm-level SCNAs. **(c)** Heatmaps showing the total segmented copy-number profile of the TCGA GBM set (leftmost panel), and the results of computationally separating these samples into arm-level profiles (middle panel) and focal profiles (rightmost panel) by summing arm-level and focal SCNAs. In each heatmap, the chromosomes are arranged vertically from top to bottom and samples are arranged from left to right. Red and blue represent gain and loss, respectively.

Mermel *et al. Genome Biology* 2011 **12**:R41 doi:10.1186/gb-2011-12-4-r41

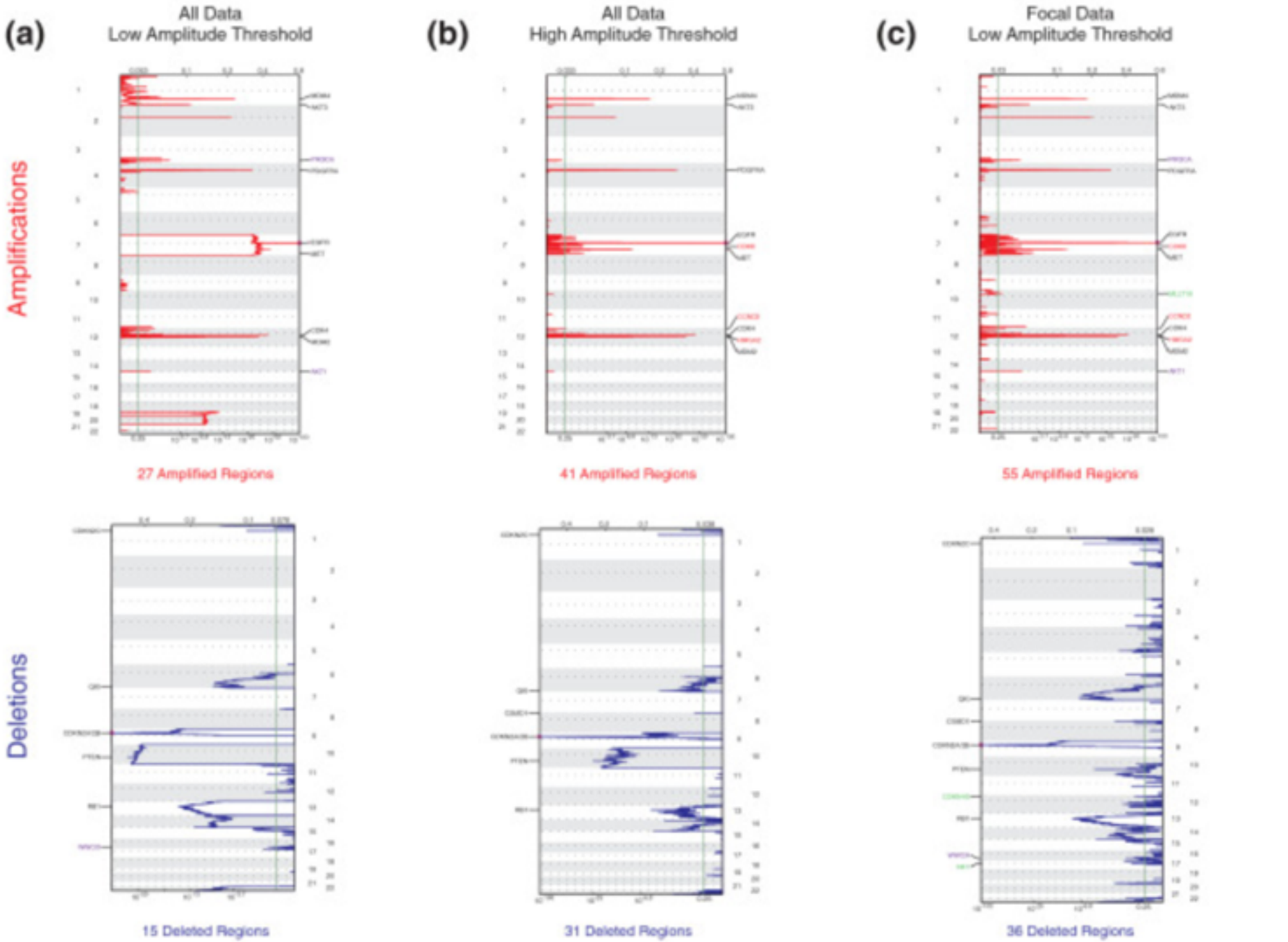


Step 3

Scoring SCNAs in Each Region According to Likelihood of Occurring by Chance

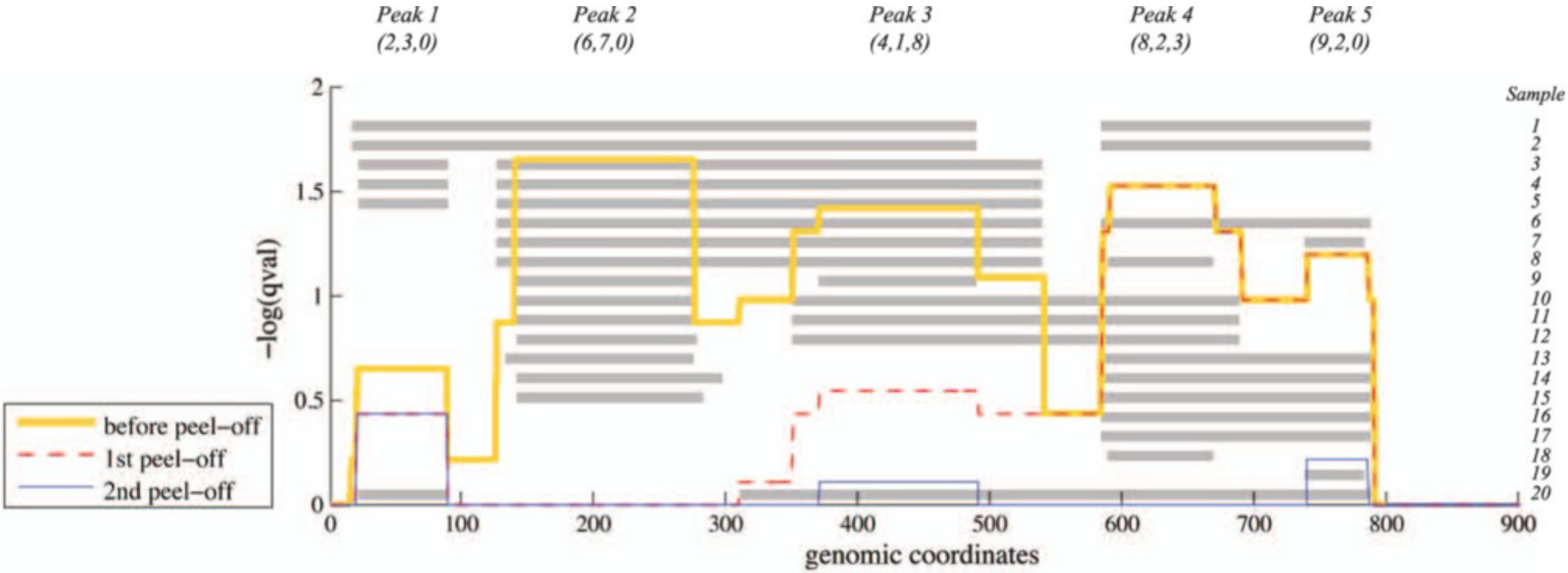
Figure 3.

Resolution: **standard** / high



Effects of amplitude-based or length-based filtering of arm-level events on GISTIC results. (a-c) GISTIC amplification (top) and deletion (bottom) plots using all data and a low amplitude threshold (a), using all data and a high amplitude threshold (b), and using the focal data and a low amplitude threshold (c). The genome is oriented vertically from top to bottom, and GISTIC q-values at each locus are plotted from left to right on a log scale. The green line represents the significance threshold (q-value = 0.25). For each plot, known or interesting candidate genes are highlighted in black when identified by all three analyses, in red when identified by the high amplitude or focal length analyses, in purple when identified by the low amplitude or focal length analyses, and in green when identified only in the focal length analysis.

	GISTIC	GISTIC2.0
G Score	G = frequency x amplitude	G = -log(Probability | Background)
Focus	Markers	Markers or Genes
P-val	computed by random permutation of markers across genome	computed by random permutation of markers or bins across genome

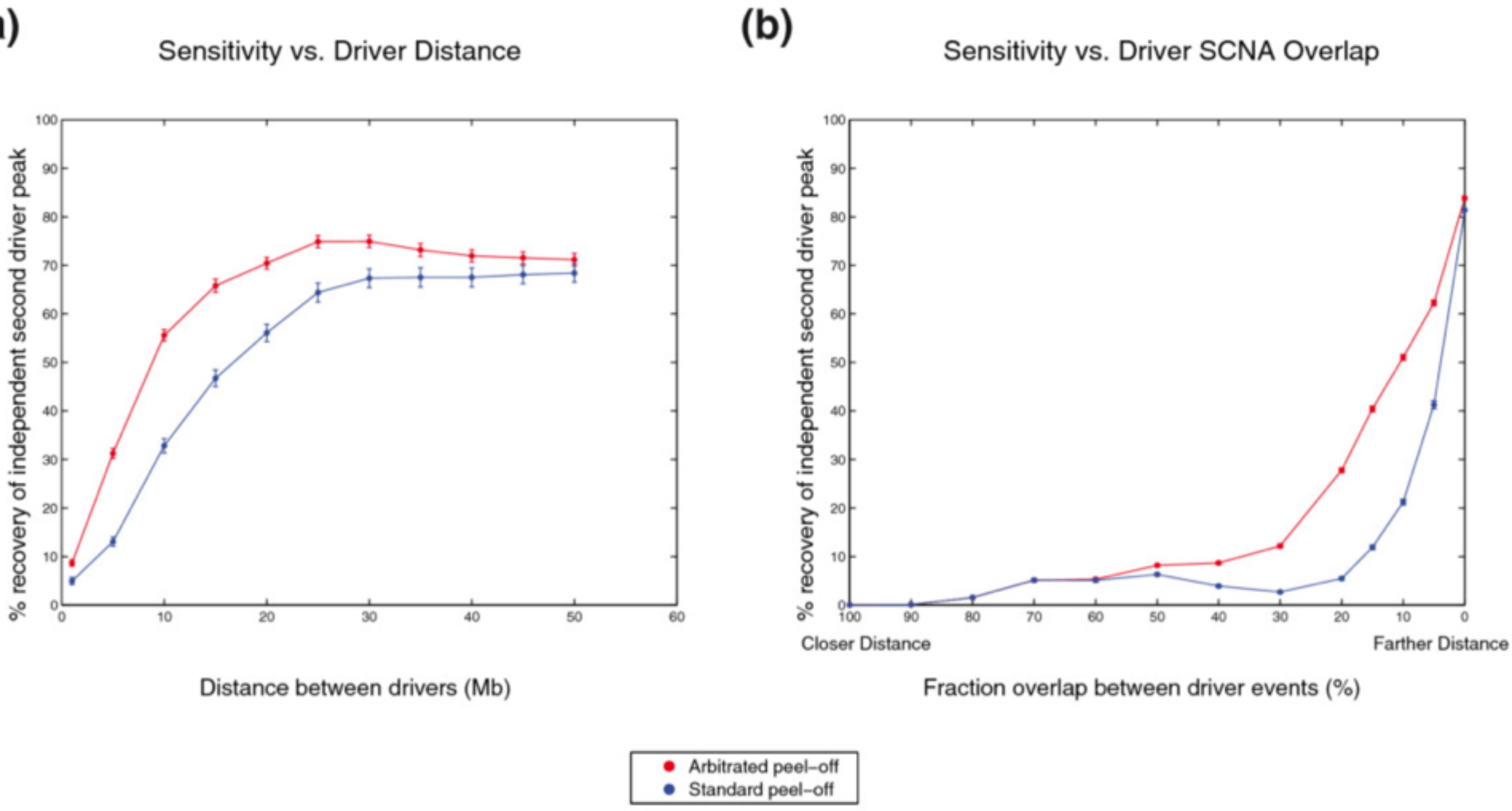


Step 4

Defining Independent Genomic Regions Undergoing Significant Levels of SCNA

	GISTIC	GISTIC2.0
Peel Off	Greedy	Arbritrated
Wij	Wij: Si if segment i cov. by j 0 else	$\mathbf{w_j} = \sum_i \mathbf{w_{ij}}$

Figure 4.

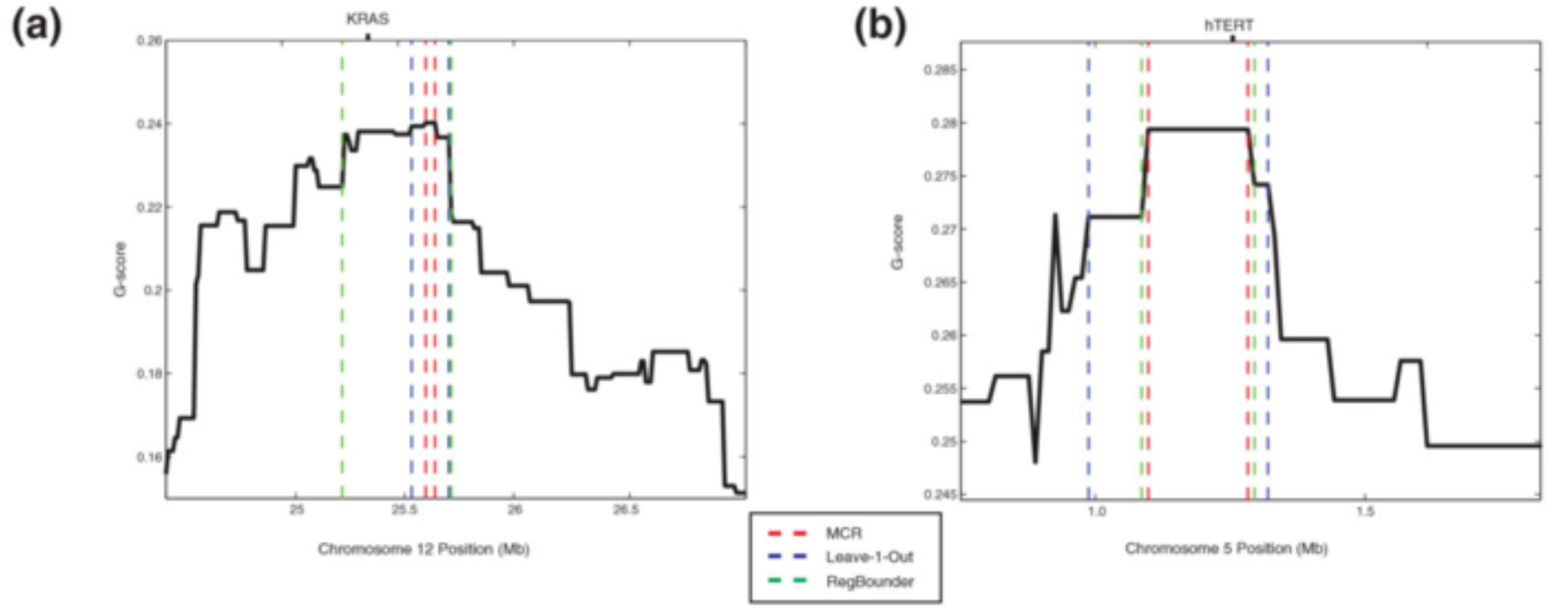


Sensitivity of peel-off to detect secondary driver events. The average fraction of secondary driver events recovered in independent (not containing the primary driver) peaks by GISTIC using the standard peel-off method (blue line) or arbitrated peel-off (red line) is shown for two simulated datasets. **(a)** The data are derived from 1,000 simulated chromosomes across 300 samples with a primary driver event present in 10% of samples and a secondary driver event a fixed distance away that is present in 5% of samples. **(b)** Data are derived from 10,000 simulated chromosomes across 300 samples with a primary driver event present in 10% of samples and a secondary driver event present in 5% of samples, where the fraction of the secondary driver events that overlapped with the primary driver event was varied between 100% (complete dependence; far left) and 0% (complete independence; far right). Error bars represent the mean ± standard error of the mean (some are too small to be visible).

Mermel *et al. Genome Biology* 2011 **12**:R41 doi:10.1186/gb-2011-12-4-r41

Figure 6.

Resolution: standard / **high**



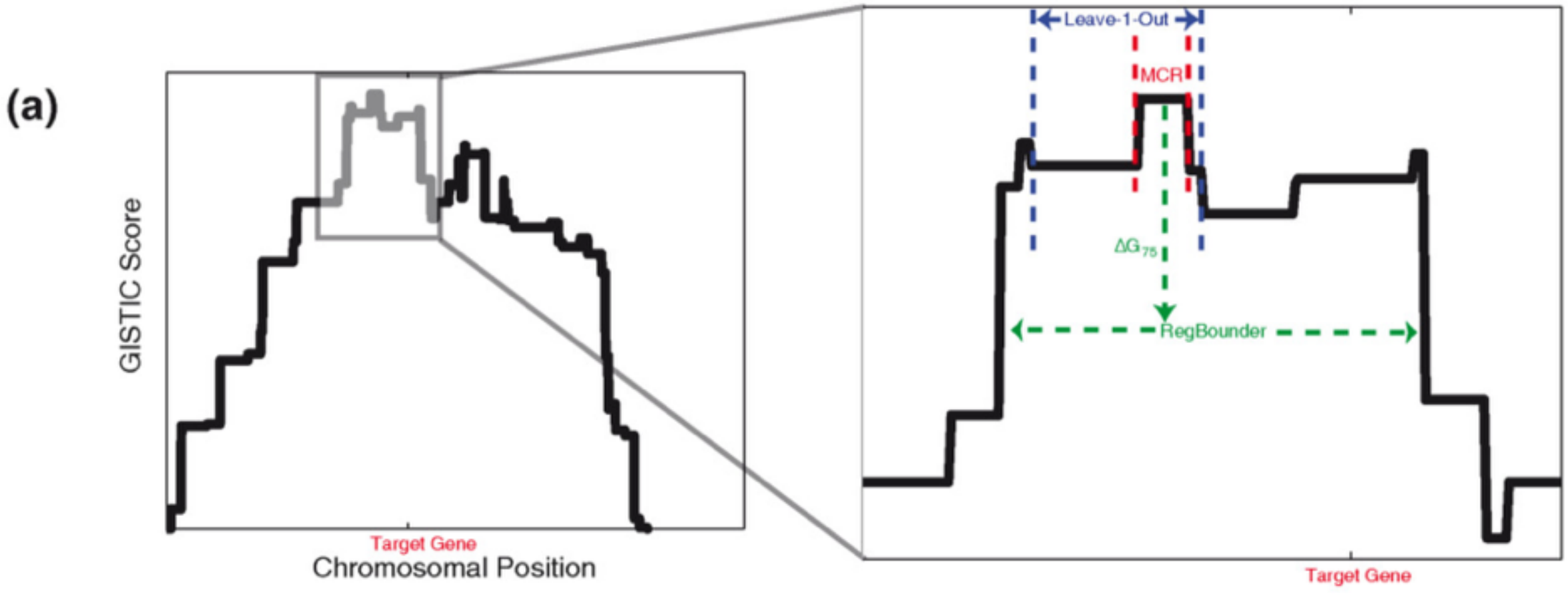
Comparison of RegBounder to MCR and leave-1-out procedures applied to primary lung adenocarcinomas. The advantages of RegBounder over previous peak-finding procedures are illustrated for two well-described oncogene peaks identified in GISTIC analysis of 371 lung adenocarcinoma samples characterized on the Affymetrix 250K StyI SNP array (as published in [16]). **(a)** A well-described amplification peak is identified on chromosome 12p12.1 with MCR (red dotted lines) near to but not containing the known lung cancer oncogene *KRAS*. Because there are more than two apparent passenger events in this region, the leave-1-out peak (blue dotted lines) also does not contain *KRAS*. However, RegBounder (green dotted lines) produces a wider peak that captures *KRAS*. **(b)** An amplification peak on chromosome 5p15.33 contains *hTERT*, the catalytic subunit of the human telomerase holoenzyme, within the MCR (red dotted lines). In this case, RegBounder (green dotted lines) produces a narrower peak region than the corresponding leave-1-out peak (blue dotted lines), demonstrating the ability of RegBounder to achieve a greater balance between peak region size and accuracy. In both (a) and (b), the y-axis depicts the amplification G-score and the x-axis denotes position along the corresponding chromosome.

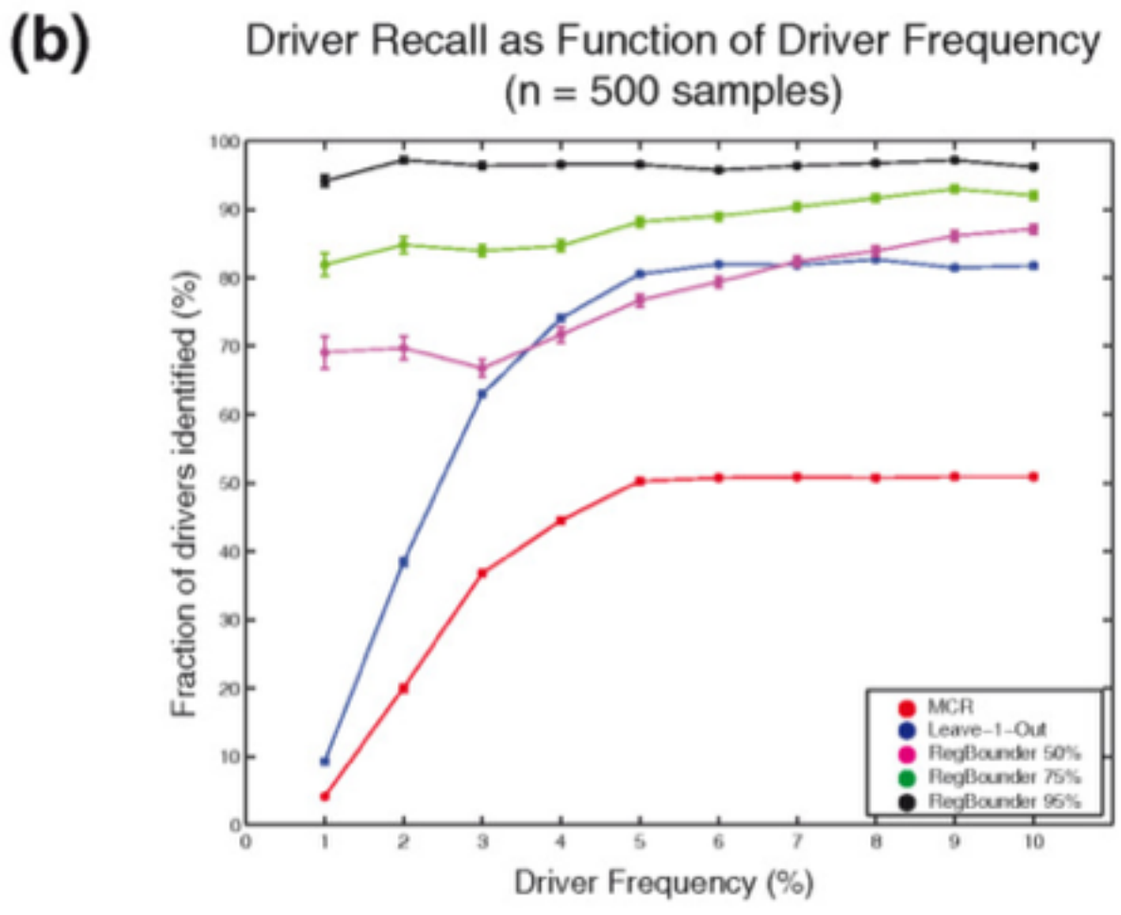
Mermel *et al. Genome Biology* 2011 **12**:R41 doi:10.1186/gb-2011-12-4-r41

Step 5

Accurate Definition of the Copy Number Profile In Each Sample

Figure 5.


(a)
GISTIC Score
Target Gene
Chromosomal Position
Leave-1-Out
MCR
ΔG75
RegBounder
Target Gene


(b)
Driver Recall as Function of Driver Frequency
(n = 500 samples)
Fraction of drivers identified (%)
Driver Frequency (%)
MCR
Leave-1-Out
RegBounder 50%
RegBounder 75%
RegBounder 95%

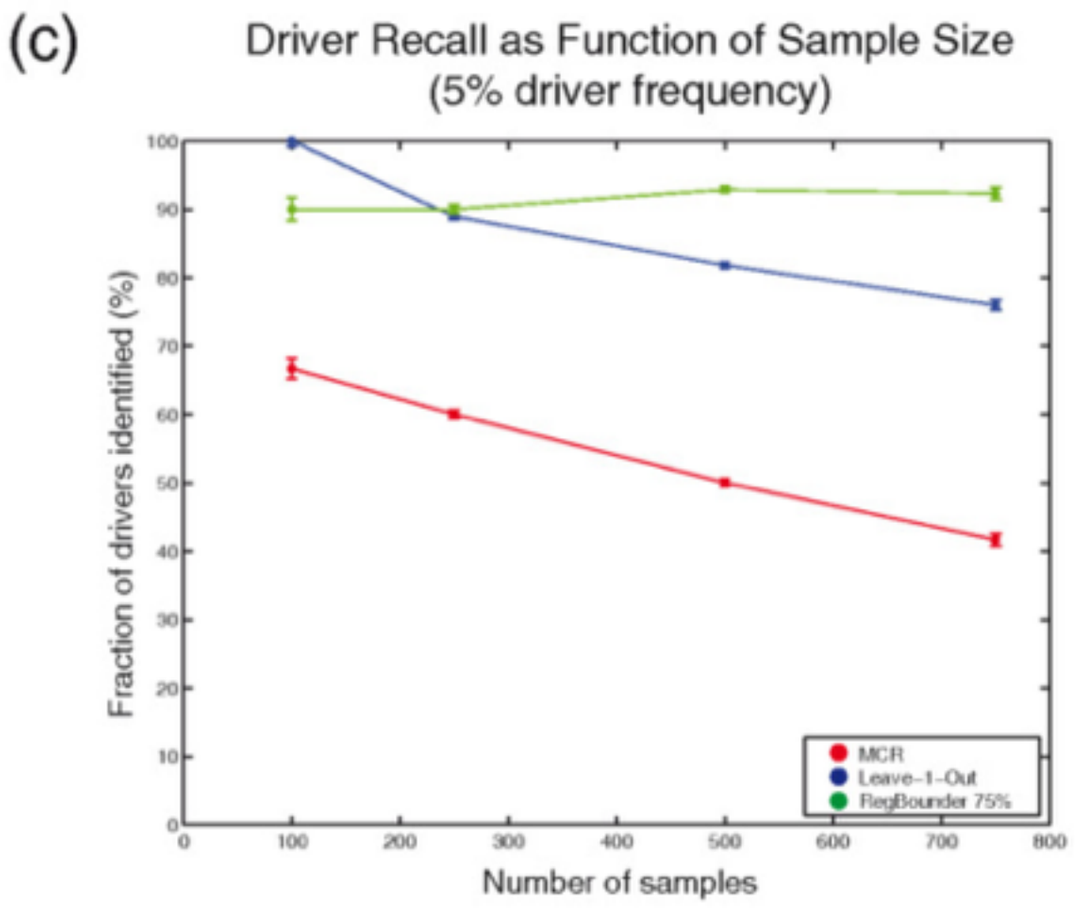
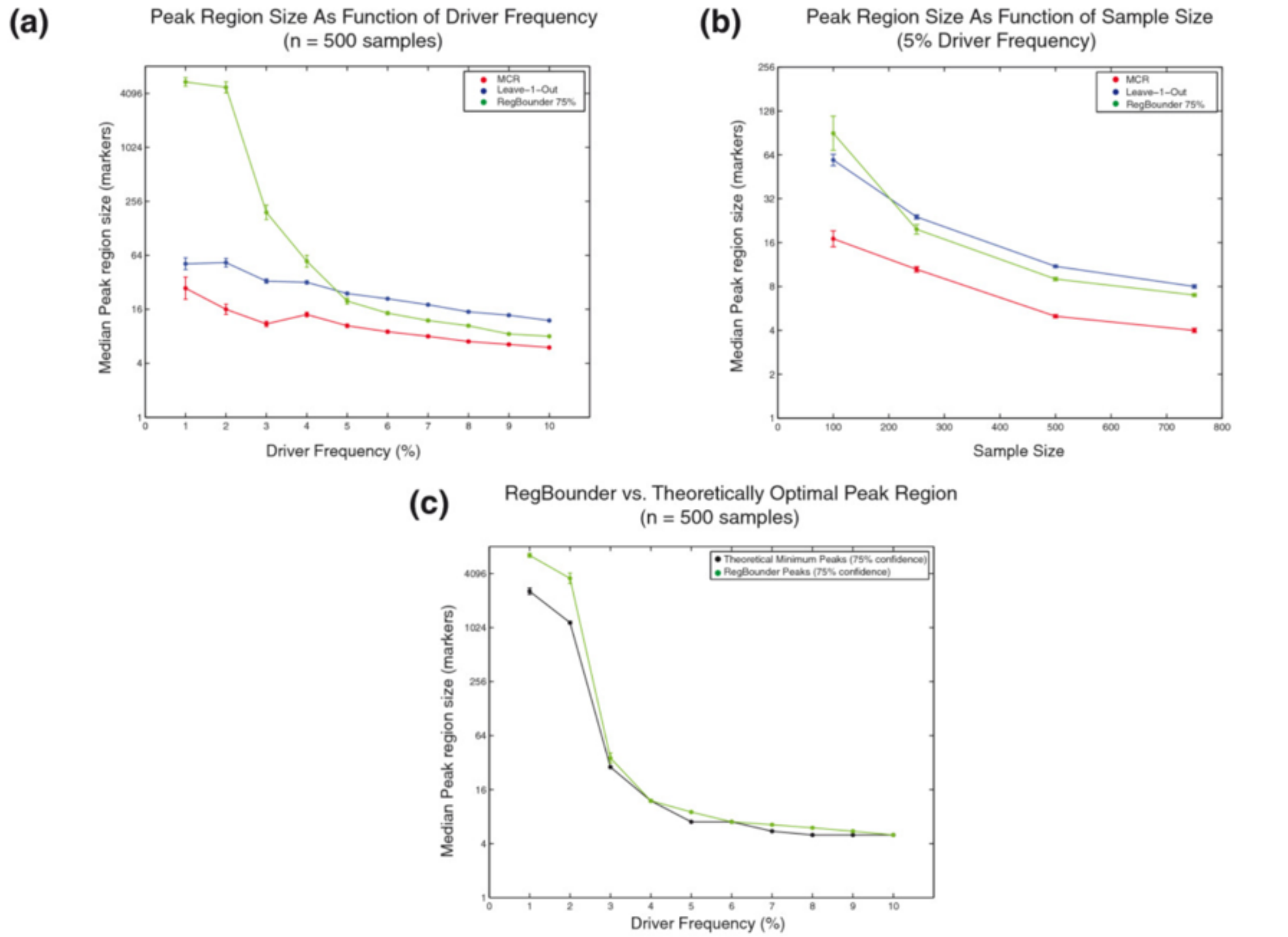

(c)
Driver Recall as Function of Sample Size
(5% driver frequency)
Fraction of drivers identified (%)
Number of samples
MCR
Leave-1-Out
RegBounder 75%

Figure 7.



Specificity of peak finding algorithms. **(a,b)** The median size of the peak regions produced by the MCR (red), leave-1-out (blue), and RegBounder (green, 75% confidence) are shown as a function of driver frequency (a) and sample size (b). In (a), data are derived from 10,000 simulated chromosomes across 500 samples in which the driver frequency varied from 1 to 10%. In (b), data are derived from 10,000 simulated chromosomes across a variable number of samples in which the driver frequency was fixed at 5%. **(c)** Comparison of the peak region sizes obtained by RegBounder (green line) with the theoretically minimal peak region sizes (black line) that could be obtained by any peak finding algorithm with a similar confidence level (Supplementary Methods in Additional file 1). Error-bars represent the mean ± standard error of the mean (some are too small to be visible).

Mermel *et al. Genome Biology* 2011 **12**:R41 doi:10.1186/gb-2011-12-4-r41