DEFOR: Depth- and Frequency-Based Somatic Copy Number Alteration Detector

Supplementary Data

He Zhang[1]*, Xiaowei Zhan[1], James Brugarolas[2], Yang Xie[1]*

1. Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.
2. Department of Internal Medicine and Kidney Cancer Program, Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.

* To whom correspondence should be addressed.

Table of Contents

1 Introduction

DEFOR is a software package that uses exome sequencing or whole genome sequencing data to identify somatic copy number alteration (SCNA) from paired tumor and normal samples. DEFOR doesn't rely on the strong assumption that there are only a small proportion of SCNAs in the genome. Our evaluation showed that DEFOR have better accuracy than other six available methods for SCNA profiling from exome-sequencing, especially in the situation that there are large-scale copy number alterations in the genome.

2 Methods

2.1 Statuses of copy number

DEFOR supports the estimation of copy number alterations in six different statuses (**Supplementary Table S1**), and these statuses can be distinguished by allele frequency and/or depth ratio between a pair of tumor and normal samples.

2.2 Estimation of allele frequency clusters

Reference allele frequency is estimated for each site. An EM-algorithm is used to estimate the cluster pattern of allele frequency for each sliding window. Then genome is segmented into blocks according to the allele frequency clusters in each overlapping window.

2.3 Estimation of depth ratio

Considering the possible target capture efficiency bias in exome-sequencing, the depth-ratio between paired tumor and normal samples are used for absolute copy number estimation. Depth ratio between paired tumor and normal samples is estimated for each sliding window by counting the depth of coverage of each site.

2.4 Normalization of depth ratio

2.4.1 Principle

Since the depth of coverage may be different for tumor and normal samples, the raw depth ratio must be normalized before copy number estimation. That means we must estimate the 'standard raw depth ratio' (DR_{st}) which represents the normal status (copy number of two), and then all raw depth-ratio can be normalized based on DR_{st}.

In most existing methods, the median or mean value of depth-ratio across genome is usually used as the DR_{st}. However, such estimation relies on the assumption that SCNAs only occupy a small proportion of the genome. If that assumption is not correct, the estimation of DR_{st} is not reliable because the mean or median depth ratio may not represent the normal status. To improve the accuracy of the estimation of DR_{st} especially when there are large –scale SCNAs in the tumor genome, allele frequency was incorporated as an important factor in DEFOR.

Although the allele frequency cannot be used to estimate the absolute copy number directly, it indeed can help select the regions without large-scale SCNAs. Allele frequency distribution has different pattern for different SCNA events (Supplementary Table. S1). If imbalanced SCNA (two alleles have different copy numbers) occurred, it's expected that the frequencies of two alleles in the heterozygosity site would not be

around 50%, and the reference allele frequency of different sites could be grouped into two clusters (Supplementary Table S1). Meanwhile, in the normal regions with copy number of two, the expected allele frequency in heterozygosity sites should be around 50%. So only for the region with AF = 50% could be used to estimate the depth ratio representing the normal statues of copy number. However, when balanced SCNAs (the copy numbers of two alleles are the same) occurred, we can still expect the reference allele frequency to be 50% in heterozygosity sites, and so balanced SCNAs cannot be distinguished with normal regions using only allele frequencies. Considering such situation, we must also consider the observed depth ratio to distinguish between normal regions and balanced SCNAs when estimate the correct normalization factor.

2.4.2 Assumption

To facilitate the estimation of the normalized depth ratio, the following assumptions were used in DEFOR:

1. L(CN = 2) > 50Mb;

2. L(CN = 2) > L(CN = 0);

3. L(CN = 2) > L(CN = 4);

L(CN = 2) : total length of regions with copy number of two

L(CN = 0) : total length of regions with copy number of zero (loss of both alleles)

L(CN = 4) : total length of regions with copy number of four (amplification of both alleles)

This set of assumption is weaker than the commonly use assumption that most of the genome don't have SCNAs. With these assumptions, we are able to use both depth and allele frequency to identify the candidate regions that represent the status of copy number of two.

2.4.3 Clustering of allele frequencies in each segment

For all positions in one segment, the allele frequencies were assigned to one of four clusters. In the initial step, we define the frequency of four clusters at 0, 0.33, 0.66 and 1. The standard deviation of each cluster is set as 0.05. For each position, the initial probabilities for four clusters are assigned as 0.25. the following algorithm is used to assign allele frequencies into four clusters:

1) Then the probability density of each cluster for each position (p) is calculated using Gaussian distribution:

$$D_{pi} = pdf\left(F_p - \mu_i, sd_i\right)$$

pdf: probability density function of normal distribution
F_p: observed allele frequency at position p
μ_i: mean frequency of cluster i
sd_i: standard deviation of cluster i

2) The probability (P_{pi}) of each cluster for each position is calculated:

$$P_{pi} = \frac{D_{pi}}{\sum_{i=1}^{4} D_{pi}}$$

3) The mean frequency and standard error of each cluster is updated:

$$\mu_i{}' = \mu_i$$

$$\mu_i = \frac{\sum_{p=1}^{n} P_{pi} F_p}{\sum_{p=1}^{n} P_{pi}}$$

$$sd_i = \sqrt{\frac{\sum_{p=1}^{n} P_{pi} (F_p - \mu_i)^2}{\sum_{p=1}^{n} P_{pi}}}$$

4) For any one of the clusters, if $|\mu_i - \mu_i'|$ is greater than 0.001, go back to step 1); otherwise go to next step.
5) Finally, we get the mean frequency of four clusters μ_1, μ_2, μ_3 and μ_4.

μ_1 and μ_4 represent the clusters around 0 and 1, which were composed of homozygosity sites and a part of sequencing errors. μ_2 and μ_3 provide information about the clusters around the center, and they are used to infer copy numbers in the following steps.

2.4.4 Inference of candidate normal regions

To infer the candidate normal regions (with copy number of two), only the region with allele frequency clustered around 0.5 ($|\mu_2 - \mu_3| < 0.1$) were selected. The depth ratios for these regions are grouped by each chromosome. The median value and standard deviation of depth ratios of all windows from candidate normal regions is estimated for each chromosome, and then only the chromosome with the standard deviation greater than twice of the standard deviation is excluded from candidate normal region. This step is used to exclude the chromosomes where there is large proportion of both regions with balanced SCNAs.

2.4.5 GC-content based correction

Because the strong correlation between depth ratio and GC-content is observed in both previous studies and our study, the raw depth ratio need to be adjusted according to GC-content which is termed as GC-content based correction. Only the candidate normal regions are used for GC-content based correction.

The entire interval of GC-content ([0, 1]) is split into 20 small intervals, and the length of each GC interval is 0.05, and so the first interval is [0, 0.05], the second interval is (0.05, 0.1], …, and the last one is (0.95, 1]. The middle point of each GC-content interval is used to represent the GC-content (GC_i, $1 \leq i \leq 20$) of that interval. Genomic regions in different sliding windows are assigned into different GC intervals according to their GC-content. Respectfully, the depth ratio (in logarithmic scale) can be estimated for each sliding window in candidate normal regions (identified in previous steps). The median depth ratio across all of the regions assigned to the *i*-th GC-content interval (GC_i) is denoted as depth ratio DR_i. Then for any given GC-content (GC_{gc}), the expected depth ratio (DR_{gc}') for GC_i can be calculated via linear interpolation as follows (assume $GC_i \leq GC_{gc} < GC_{i+1}$):

$$\frac{DR_{gc} - DR_i}{GC_{gc} - GC_i} = \frac{DR_{gc} - DR_{i+1}}{GC_{gc} - GC_{i+1}}$$

For any region with observed raw depth ratio DR_{obs} and GC-content GC_{gc}, the difference between the observed depth ratio (DR_{obs}) and the expected depth ratio (DR_{gc}) is a better statistic to reflect the copy number of the given region. Thus the GC-content adjusted depth ratio ($DR_{gc\text{-}adj}$) is estimated according to the following formula:

$$DR_{gc-adj} = DR_{obs} - DR_{gc}$$

2.4.6 Normalized depth ratio

After the estimation of GC adjusted depth ratio, the median value of depth ratio of the candidate normal regions is chosen as the standard depth ratio representing the copy number of two (DR_{st}).

And then all depth ratios are adjusted according to DR_{st} using the following formula:

$$DR_{adj} = DR_{gc-adj} - DR_{st}$$

After this adjustment, the normalization of depth ratio has been finished. If the copy number of a region is two, then the expected DR_{adj} of this region should be around 0.

2.5 Estimation of copy number

Based on the adjusted depth ratio and allele frequency clusters in each region, copy number status was can be estimated (Supplementary Figure S5) based on the principle proposed in Supplementary Table S1.

3 Evaluation

3.1 Data

The validation data was from a published study on kidney cancer (Pena-Llopis, et al., 2012). Evaluation was based on 9 pairs of normal-tumor samples with both SNP array (Affymetrix 6.0) data and exome-sequencing data.

3.2 Estimation of copy number from SNP array data

High density SNP array was served as the gold standard for SCNA detection for a long time considering the good resolution and coverage. To evaluate the performance of DEFOR and some of the other methods for exome-sequencing data, we used SCNAs identified based on SNP array data as the gold standard. To avoid possible artificial bias, a third party web-based pipeline, Copy Number Inference Pipeline from GenePattern (Reich, et al., 2006), was used to conduct SCNAs calling from SNP array data.

3.3 Estimation of copy number from exome sequencing data

All reads was mapped to the reference genome (hg19) using bwa-mem (Li and Durbin, 2009). SCNA detection was conducted using DEFOR and six other tools, including CNVkit, Falcon, VarScan2, cn.mops, CNVnator and CNV-seq.

3.3.1 Comparison between results from exome sequencing and SNP array data

SCNAs identified from exome-sequencing data and array data were compared for each base. If a SCNA was detected on both exome-sequencing data and SNP data, and the changing direction (gain or loss) also matched with each other, we considered this result as true positive (TP). If a SCNA was identified only in exome-sequencing data but not in array data, we considered this SCNA as false positive (FP). If a SCNA was identified only in array data but not in exome-sequencing data, this SCNA was considered as false negative (FN). Total length of TP, FP and FN were calculated, and then the recall, precision and F-score were estimated (Supplementary Table S2 and S3) to evaluate the accuracy of different methods. The proportion of SCNAs in each tumor sample was estimated based on the array data.

3.3.2 Samples with SCNAs occupied < 30% of the genome

Based on the evaluation result (Supplementary Table S2 and S3), when the SCNAs occupy a small proportion of the genome (< 30%), DEFOR and CNVkit outperformed the other methods. For these five samples with relative low proportion of SCNAs, DEFOR and CNVkit had a good concordance with array data, and DEFOR (precision = 97.3%, recall = 98.2%) performed better in both precision and recall than CNVkit (precision = 96.9%, recall = 97.4%).

3.3.3 Samples with SCNAs occupied >30% of the genome

When the SCNAs occupy a large proportion of tumor genomes (> 30%), the concordance between the results from exome-sequencing and arrays seems not as good as those samples with less proportion of SCNAs. Then we inspected the results in much details. To facilitate interpreting the results, depth ratio, reference allele frequency and SCNAs calling results from different methods were plotted (Supplementary Fig. S1 – S4).

For sample T164T (Supplementary Fig. S1), the relative copy number in the results from array, DEFOR and CNVKit are highly consistent with each other, but the estimated absolute copy numbers are different significantly. The key point that caused the observed difference between these results is that the positions of 'central' line (red line) representing the normal status are different. That means the genomic regions that represent the copy number of two were different for different methods.

Based on the pattern of allele frequency distribution and depth ratio, we think the result from DEFOR is better. As mentioned before, for the regions with copy number of two (around red line), the allele frequency in heterozygosity sites were distributed around 50%, while for the regions where the regions with imbalanced SCNAs, the allele frequency of heterozygosity sites should be departed from 50%. Using sample T164 as an example, based on the array result, the copy number of chromosome 6, 8, 9, 11, 13, 14, 15, 17, 18 and 22 are thought slightly less than two, and the copy number of chromosome 7, 10, 16, 19, 20 were greater than two. The result of CNVkit indicates the copy number of chromosome 6, 8, 9, 11, 13, 14, 15, 17, 18 and 22 are nearly two, and the copy number of chromosome 7, 10, 16, 19, 20 are greater than two. But the result of DEFOR shows that the copy number of chromosome 6, 8, 9, 11, 13, 14, 15, 17, 18 and 22 greater than two, while chromosome 7, 10, 16, 19, 20 are around two. Based on the pattern of allele frequency distribution, the AF of heterozygosity sites in chromosome 7, 10, 16, 19, 20 are distributed around 0.5, but the AF of heterozygosity sites in chromosome 6, 8, 9, 11, 13, 14, 15, 17, 18 and 22 are not 0.5 but very close to 0 or 1. Considering these observations, the results from DEFOR represent a reasonable solutionof correct copy number status across the genome. Meanwhile, if another whole genome duplication happened after the copy number alterations estimated from DEFOR, the result of CNVKit also represent another reasonable solution. Without other type of data, it's difficult to tell which one is more reasonable using only exome-sequencing data. But based on the parsimony principle, we think the result from DEFOR is better, because one more step (whole genome duplication) is needed to interpret the result from CNVKit.

Although it seems that the concordance between DEFOR and array results are not high, we think the results of DEFOR make sense in this example. Both array results and CNVkit result may have some problem in estimating the position of 'center line'. Based on the results for T166T, T144T and T108M, DEFOR also performed well in these samples (Supplementary Fig. S2 – S4).

4 Reference

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754-1760.
Pena-Llopis, S., *et al.* BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet* 2012;44(7):751-759.
Reich, M., *et al.* GenePattern 2.0. *Nat Genet* 2006;38(5):500-501.

5 Supplementary Tables

Supplementary Table S1. The allele frequency, depth ratio and purity of tumor cells for each copy number status

	If there is no heterogeneity (purity = 1)		If the purity of tumor cells is less than 1
SCNA Status	Allele Frequency (Minor, Major)	Normalized Depth Ratio	Relationship between allele frequency (f), depth ratio (d) and purity (p)
Loss of 2 allele	NA	0	d = 1 – p f = 0.5
Loss of 1 allele	0, 1	0.5	d = 1 – 1/2 * p f = (1 – p) / (2 – p) d = 1 / (2 * (1 – f))
Loss of 1 allele then followed by amplification	0, 1	1	d = 1 f = (1 – p) / 2
Normal	0.5, 0.5	1	d = 1 f = 0.5
Gain of 1 allele	0.33, 0.66	1.5	d = 1 + 1/2 * p f = 1 / (2 + p) f = 1 / (2 * d)
Gain of 2 alleles	0.5, 0.5	2	d = 1 + p f = 1/2

Supplementary Table S2. Accuracy of the SCNAs estimated from seven methods (samples with SCNAs < 30% of the genome)

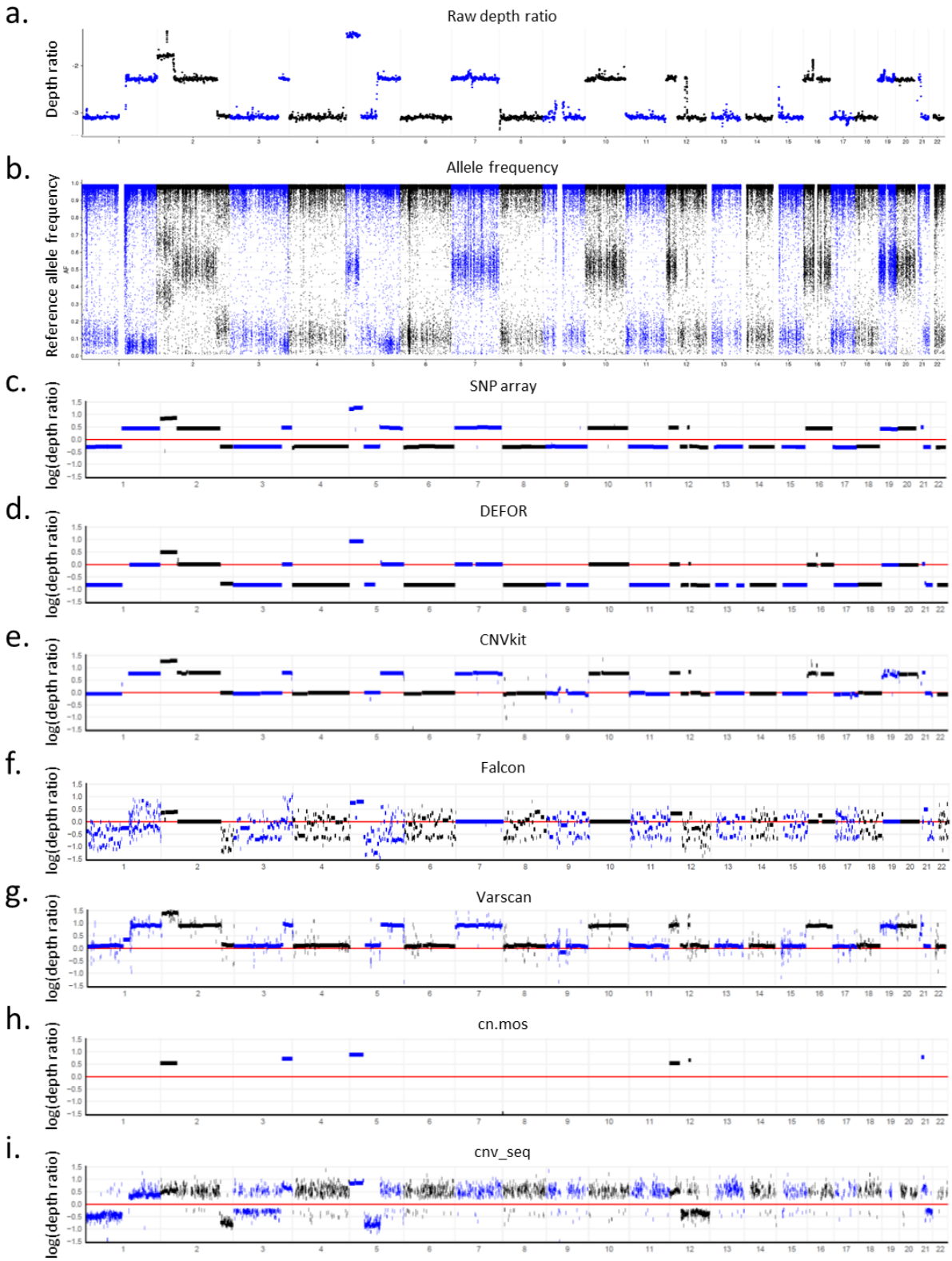
sample	Proportion of genome with SCNAs	DEFOR			CNVkit			falcon			VarScan			cnv-seq			cn.mos		
		Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
T127T	12.4%	97.0%	99.5%	0.982	96.3%	97.2%	0.967	76.5%	50.8%	0.610	6.1%	45.5%	0.107	34.4%	51.1%	0.412	2.4%	0.0%	0.001
T163T	13.8%	95.1%	92.1%	0.936	94.8%	93.1%	0.939	74.5%	73.5%	0.740	6.3%	41.2%	0.109	32.9%	88.3%	0.480	3.0%	0.0%	0.001
T142T	16.3%	98.3%	98.1%	0.982	97.7%	97.9%	0.978	80.8%	27.3%	0.408	93.4%	99.3%	0.963	33.2%	47.4%	0.390	64.8%	0.9%	0.018
T108T	22.0%	96.8%	99.8%	0.983	96.9%	98.3%	0.976	14.4%	10.8%	0.124	4.4%	16.1%	0.069	35.7%	29.3%	0.322	92.3%	6.0%	0.113
T183T	28.7%	98.2%	99.6%	0.989	97.7%	98.5%	0.981	52.2%	4.1%	0.077	17.0%	49.1%	0.252	23.8%	14.9%	0.183	20.7%	0.1%	0.002

Supplementary Table S3. Accuracy of the SCNAs estimated from seven methods (samples with SCNAs > 30% of the genome)

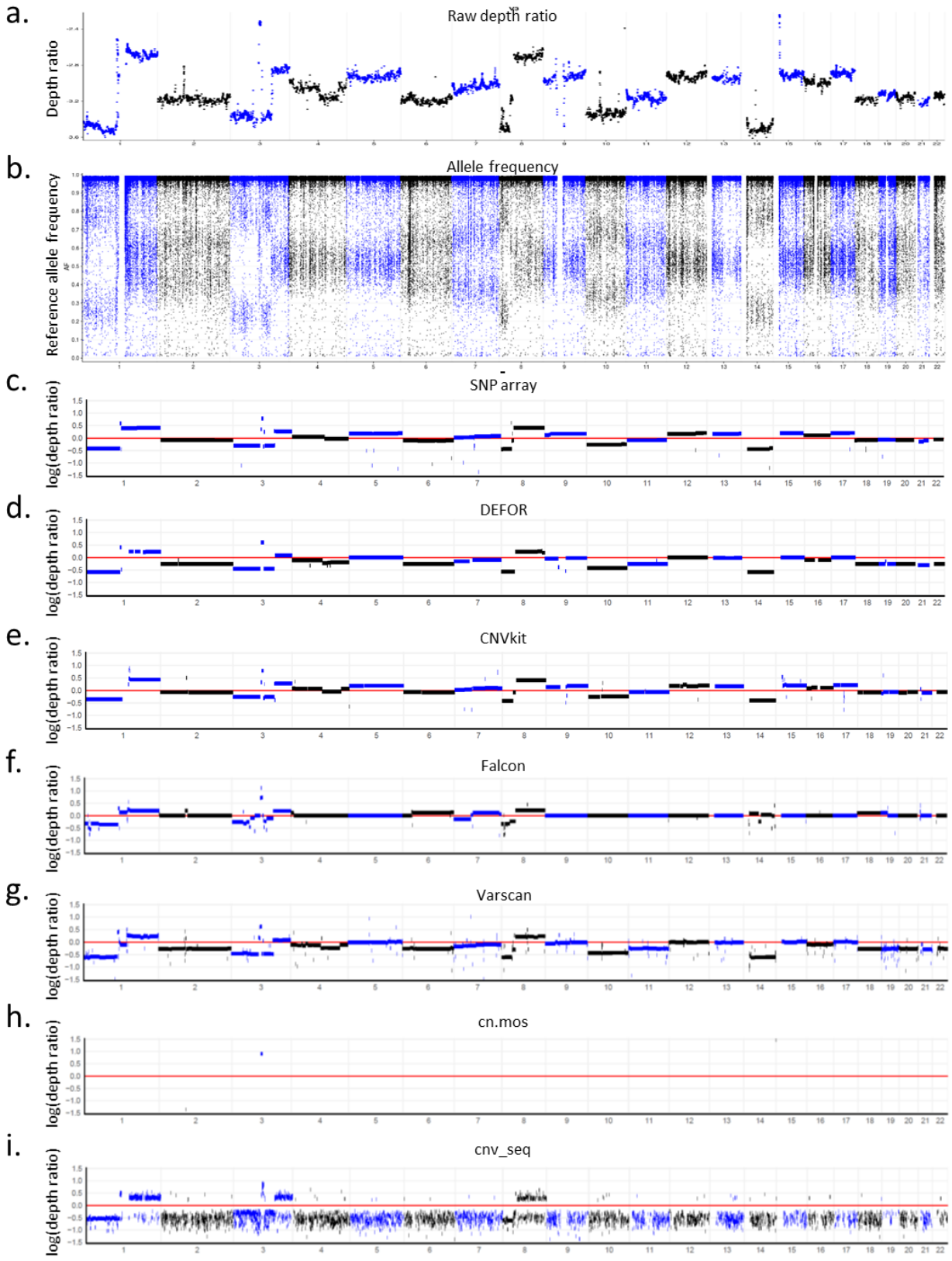
sample	Proportion of genome with SCNAs	DEFOR			CNVkit			falcon			VarScan			cnv-seq			cn.mos		
		Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
T166T	44.6%	46.7%	62.1%	0.533	91.5%	93.4%	0.925	93.7%	35.2%	0.512	42.3%	68.3%	0.523	44.2%	74.2%	0.554	62.6%	0.6%	0.011
T164T	40.8%	98.4%	59.1%	0.738	98.2%	36.1%	0.527	78.2%	39.7%	0.527	95.8%	37.8%	0.542	73.3%	44.4%	0.553	94.6%	6.6%	0.123
T144T	30.6%	96.9%	85.6%	0.909	97.8%	92.0%	0.948	78.3%	39.3%	0.524	9.6%	19.7%	0.129	65.2%	56.2%	0.603	0.3%	0.0%	0.000
T108M	57.1%	80.0%	46.6%	0.589	90.3%	53.5%	0.672	41.9%	21.6%	0.285	57.4%	46.3%	0.513	52.1%	32.6%	0.401	20.3%	0.1%	0.002

6 Supplementary Figures

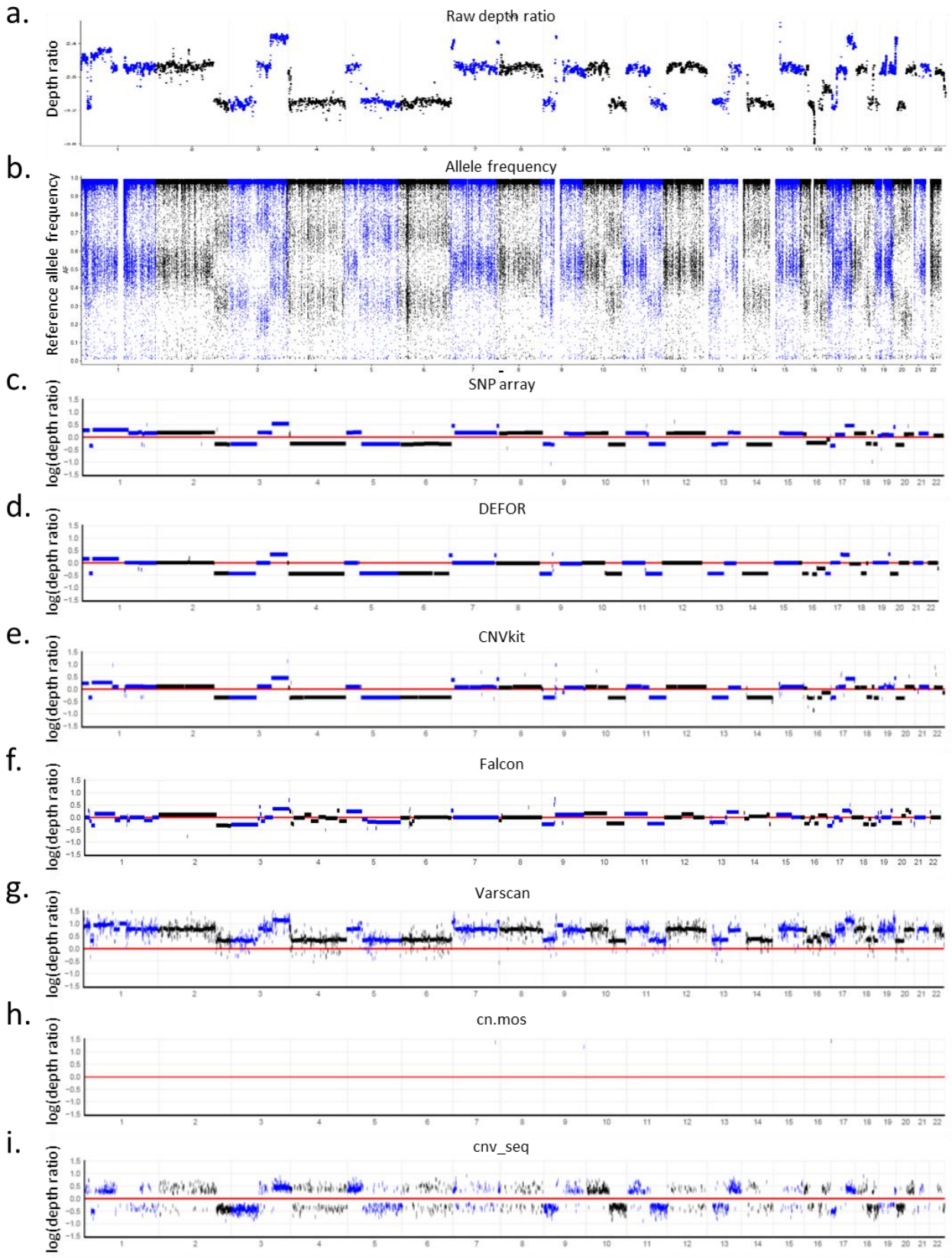
Supplementary Figure S1. Depth ratio, allele frequency and copy number for each chromosome of sample T164T. a) Raw depth ratio (from exome-sequencing data) between tumor and normal samples. b) Reference allele frequency estimated from exome-sequencing data. c) Copy numbers estimated from SNP array data. Copy numbers estimated from d) DEFOR, e) CNVkit, f) Falcon, g) Varscan, h) cn.mos, i) cnv_seq.



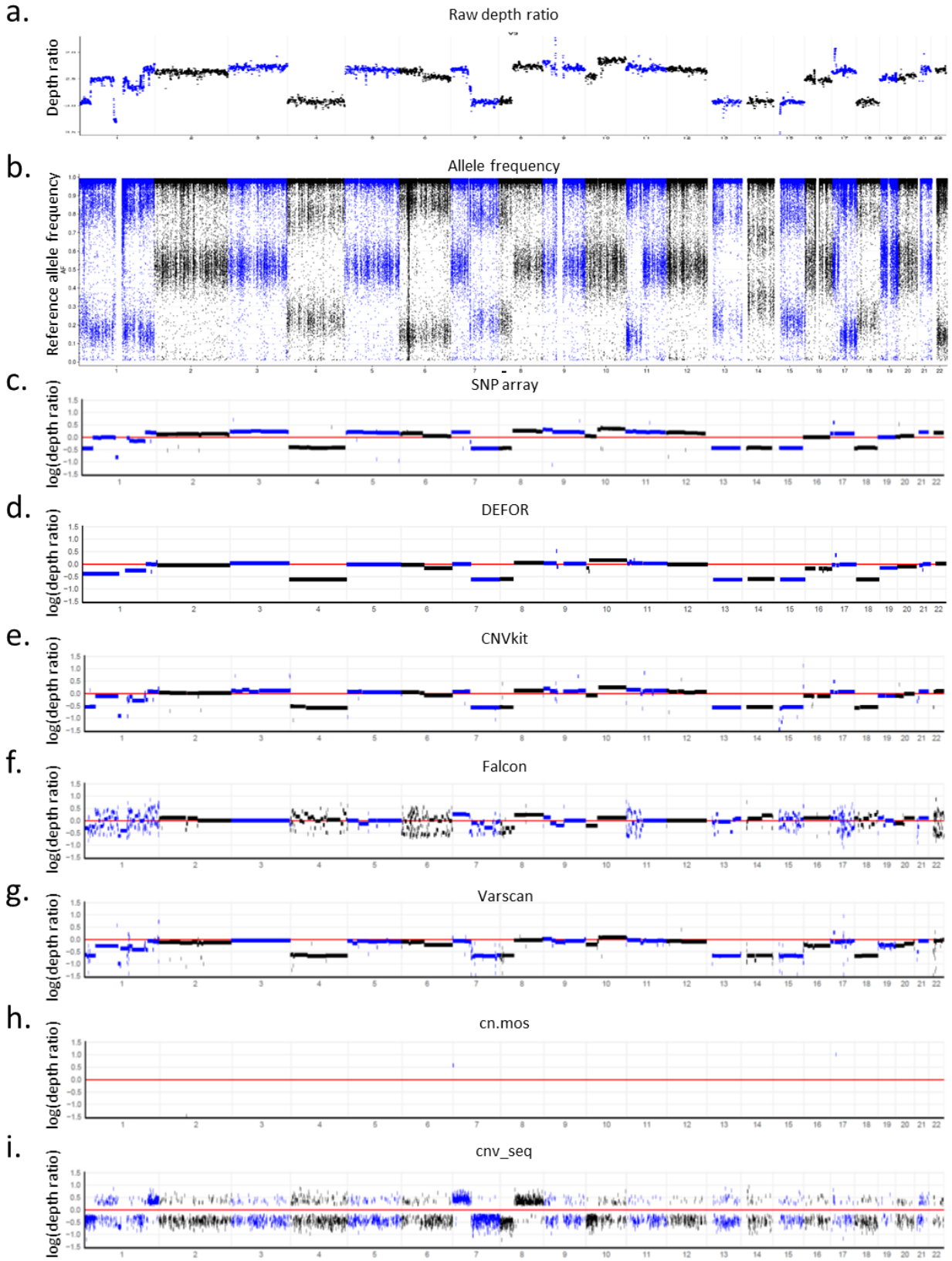
Supplementary Figure S2. Depth ratio, allele frequency and copy number for each chromosome of sample T166T. a) Raw depth ratio (from exome-sequencing data) between tumor and normal samples. b) Reference allele frequency estimated from exome-sequencing data. c) Copy numbers estimated from SNP array data. Copy numbers estimated from d) DEFOR, e) CNVkit, f) Falcon, g) Varscan, h) cn.mos, i) cnv_seq.



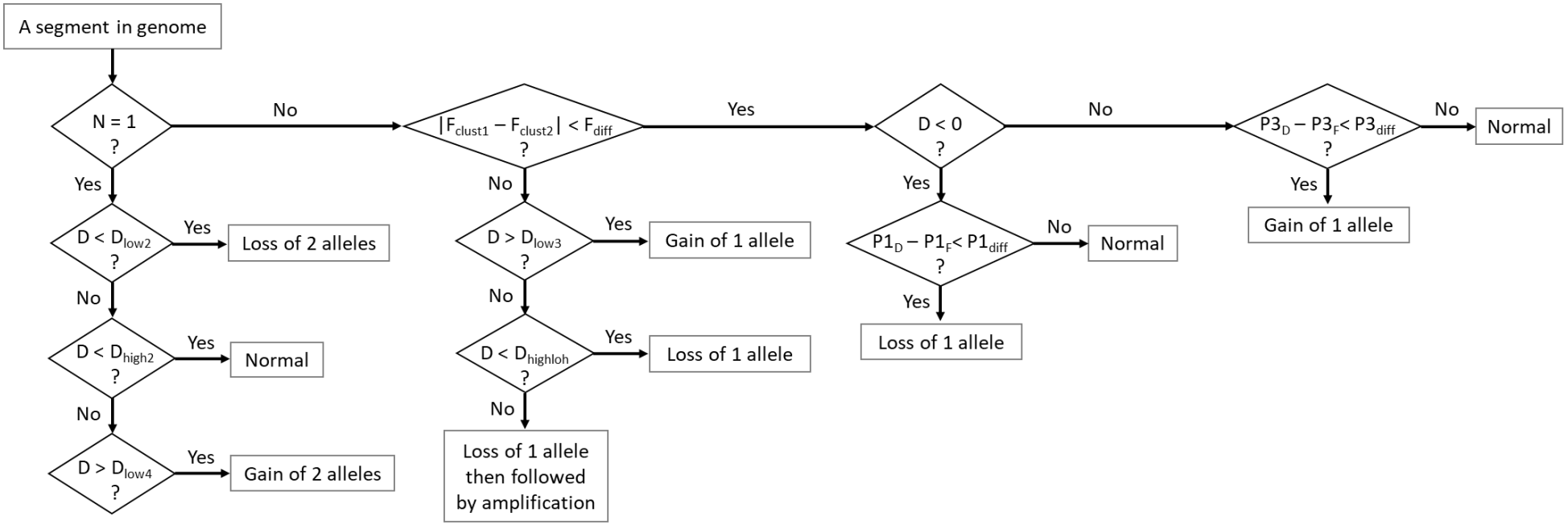
Supplementary Figure S3. Depth ratio, allele frequency and copy number for each chromosome of sample T144T. a) Raw depth ratio (from exome-sequencing data) between tumor and normal samples. b) Reference allele frequency estimated from exome-sequencing data. c) Copy numbers estimated from SNP array data. Copy numbers estimated from d) DEFOR, e) CNVkit, f) Falcon, g) Varscan, h) cn.mos, i) cnv_seq.



Supplementary Figure S4. Depth ratio, allele frequency and copy number for each chromosome of sample T108M. a) Raw depth ratio (from exome-sequencing data) between tumor and normal samples. b) Reference allele frequency estimated from exome-sequencing data. c) Copy numbers estimated from SNP array data. Copy numbers estimated from d) DEFOR, e) CNVkit, f) Falcon, g) Varscan, h) cn.mos, i) cnv_seq.



Supplementary Figure S5. Float chart of copy number status assignment



N: number of allele frequency clusters
D: $\log_2$(normalized depth ratio)
SD_{normal}: standard deviation of D in potential normal regions
D_{low2}: lower cutoff for D in normal status; the value is -3xSD_{normal}
D_{high2}: upper cutoff for D in normal status; the value is 3xSD_{normal}
D_{low3}: lower cutoff for D in 'gain of 1 alleles' status; the value is SD_{normal}
D_{low4}: lower cutoff for D in 'gain of 2 alleles' status; the value is 4xSD_{normal}
D_{high1}: upper cutoff for D in 'loss of 1 alleles'; the value is -3xSD_{normal}
F_{clust1}: median frequency of allele frequency cluster 1;
F_{clust2}: median frequency of allele frequency cluster 2;
F_{diff}: cutoff for the difference between the median frequencies of two clusters; the default value is 0.15
$P1_D$: purity estimated from depth ratio based on 'loss of 1 allele' status; the value is $2-2x2^D$
$P1_F$: purity estimated from allele frequency based on 'loss of 1 allele' status; the value is $2-1/(0.5+|F_{clust1}–F_{clust2}|/2)$
$P1_{diff}$: cutoff for the difference between the $P1_D$ and $P1_F$; the default value is 0.05
$P3_D$: purity estimated from depth ratio based on 'gain of 1 allele' status; the value is $2x(2^D-1)$
$P3_F$: purity estimated from allele frequency based on 'gain of 1 allele' status; the value is $1/(0.5-|F_{clust1}–F_{clust2}|/2)-2$
$P3_{diff}$: cutoff for the difference between the $P1_D$ and $P1_F$; the default value is 0.05