**Outline 1**

The main pipeline of CNACS is largely divided into 2 phases: (1) calculation of adjusted depths and B-allele frequencies (BAFs) from input BAM files, which is illustrated in Outline 2, and (2) segmentation based on adjusted depths and BAFs and identification of copy-number alterations (CNA), which is illustrated in Outline 3.

**Outline 2**

Calculation of adjusted depths and BAFs is performed in “ref\_count.sh” in control samples or in “cnacs.sh” in tumors. Both scripts include almost the same processes. However, slightly different subscripts are used, e.g., “bam2hetero\_ref.sh” in controls and “bam2hetero.sh” in tumors. In addition, “cnacs.sh” also executes segmentation and identification of CNAs in tumor samples (“cnacs\_main.sh”), which is described in Outline 3.

* Processing BAM files (“proc\_bam.sh”)

An input BAM file is processed into a filtered BAM file and a BED file containing fragment information. In this process, inappropriate sequencing reads are filtered out based on their mapping qualities, intersection with target regions, and FLAG bits in the SAM format. At the same time information on fragment-length distribution is obtained.

* Calculation of adjusted BAF (“bam2hetero.sh”)

Based on the filtered BAM file, raw BAFs for target SNPs are calculated in controls and tumors. Deviation of raw BAFs in heterozygous controls from 0.50 means biased representation of alternative alleles. BAF values in tumor samples are corrected for this bias as follows.

(: a read count of a major allele in a tumor sample, : a read count of a minor allele in a tumor sample, : a mean of BAFs in heterozygous control samples)

* Size assortment (“divide\_bed.sh”)

Based on their lengths, fragments in the BED file are categorized into 4 bins: (1) shortest, (2) 2nd shortest, (3) 2nd longest, and (4) longest 25% fragments. Depths in the 4 bins separately undergo correction for the biases due to BAFs, GC contents, and duplication rates, and are eventually merged into adjusted depths, as described below.

1. Correction for BAFs

Fragments in each bin are classified into SNP-overlapping fragments and nonoverlapping ones. As for SNP-overlapping fragments, we use a scaling factor, (“probe2scale.sh”), to correct the bias due to BAFs. Then, BAF-adjusted depths are calculated as follows (“correct\_baf.sh”).

(: one of fragments which overlaps with the probe region, : the length of overlap between fragment and the probe region, : a scaling factor for a SNP in the probe region, : the length of the probe region, :the set of SNP-overlapping fragments,: the set of non-overlapping fragments)

1. Correction for GC contents (illustrated in Outline 4)

First, setting *L* as a median fragment length in a bin, we obtain GC contents of all possible positions of fragments with length *L* around all probes. Numbers of fragment positions stratified by GC contents (*N*GC) are calculated (“bait\_gc.pl”). Second, number of the actual fragments in the bin on all positions are also counted, and GC-stratified fragment counts (*C*GC) are obtained. Fragment generation rates (*R*GC) are obtained by division of *C*GC by *N*GC (“stratify\_gc.pl”). Third, depths for all probes are predicted, as follows. Here, we simulated generation of fragments with length *L* from each position. Contribution of each fragment position to depth is predicted to be the product of the length of overlap with the probe region and the GC-matched fragment generation rate. Predicted depth is calculated as the sum of this product over the probe region.

(*B*: the length of the probe region, *L*: the median length of fragments in the bin, : the order of the current fragment position, *bi*: the length of overlap between the probe region and the *i-*th fragment, *:* the GC content of the *i-*th fragment position)

Finally, GC-adjusted depth is obtained as a ratio of BAF-adjusted depth to the predicted depth.

1. Correction for duplication rates

Proportions of duplication reads in the 4 bins (*d*1~4) are calculated (“count\_dup.sh”). Correction factor for duplication in each bin (*F1 ~ 4*) is obtained, solving the following equation.

GC-adjusted depths from 4 bins are corrected for duplication and are merged into adjusted depths (“correct\_length.sh”).

(: GC-adjusted depths in the *i*-th bin, : the correction factor for duplication in the *i*-th bin)

Finally, the adjusted depths on all probes are normalized so as the mean depth to be zero.

* Calculation of gene depth

In the analysis of WES data, depths for individual genes are also calculated. First, predicted depths and BAF-adjusted depths for probes within a single gene are combined in each length bin.

(: the number of probes within the gene,: the order of the probe within the gene, : the length of the -th probe, : the predicted depth of the gene in the *i*-th bin, : the predicted depth of the -th probe in the *i*-th bin, : the BAF-adjusted depth of the gene in the *i*-th bin, : the BAF-adjusted depth of the -th probe in the *i*-th bin)

GC-adjusted gene depths in 4 bins are calculated as the ratios of the BAF-adjusted gene depths to the predicted gene depths.

(: the GC-adjusted depth of the gene in the *i*-th bin)

The GC-adjusted gene depths in 4 bins are corrected for duplication and combined. Here, the correction factor for duplication in each bin (*F1 ~ 4*) is the same as those calculated above.

(: the adjusted depth of the gene)

Finally, adjusted depths for individual genes are normalized so as the mean depth to be zero.

**Outline 3**

Segmentation and identification of CNAs in a tumor sample are achieved through repetitions of several processes, which are repeated until convergence of the result or the end of the *10*-th loop. Before beginning the loop, probes with BAF < 0.05 (including homozygous SNPs) are excluded, and BAF values are converted as follows.

(: converted BAF signals, : adjusted BAFs)

* Adjustment by synthetic control signals (“make\_control.R”)

Background depth signals in the tumor sample are virtually synthesized as a linear combination of adjusted depths in multiple control samples, by fitting a regression model below. In the *1*-st loop, we use adjusted depths in the tumor sample as target signals in this regression model.

(: adjusted depths in the tumor sample in the 1st loop, : virtually synthesized background depths, : the number of control samples, : a coefficient for the *i*-th control sample, : depth signals in the *i*-th control sample)

Then, the logarithms of ratios of the adjusted depths in the tumor sample to the synthetic background depths are calculated (Depth signal 1).

(: Depth signal 1, : adjusted depths in the tumor sample)

* Adjustment by replication timings (“adjust\_reptime.R”)

Depth signal 1 is regressed on replication timings of the positions of individual probes. Depths corrected for replication timings (Depth signal 2) are obtained as residuals in this regression model.

(: a coefficient for replication timings, : replication timings of the positions of individual probes, : Depth signal 2)

* Segmentation based on depths and BAFs (“cbs.R”)

Segmentation based on Depth signal 2 is performed by the circular binary segmentation algorism. Another segmentation is performed based on the BAF signals. The results of these segmentations are merged into Segmentation 1.

* Extraction of CNA regions (“filt\_cna.R”)

Segments in Segmentation 1 are clustered based on their depth and BAF signals using Dirichlet process (DP) and Marcov chain monte carlo (MCMC). Identified clusters other than the one closest to the diploid signal (equivalent to ploidy = 2.0 and BAF = 0.5) are considered to represent CNA regions, and are stored in Segmentation 2.

* Normalization by depth regions without CNAs (“norm\_depth\_cnacs.pl”)

All probes are clustered based on Depth signal 2 and the BAF signals using DP and MCMC. Clusters with mean BAF > 0.45 are considered to represent diploid regions without CNAs. Based on the mean depth in these regions, depths in CNA regions are calibrated. At the same time, depths and BAFs are converted into ploidies and allele-specific copy numbers (ASCNs).

(: ploidies of individual probes, : ASCNs of individual probes, : the mean depth of probes without CNAs)

Ploidies and ASCNs for CNA segments are also calculated.

(: a ploidy of a CNA segment, : an ASCN of a CNA segment, : a mean depth in a CNA segment, : a mean BAF in a CNA segment)

* Transition to the next loop (“compare\_result.pl”)

If the resulting CNA profile is different from that in the previous loop, another loop will be initiated. In the first loop, this comparison is skipped. Then, the target depths used for the synthesis of background depth signals are corrected for the current CNA profile, and are used in the next loop.

When 10 loops are completed or the CNA profile is the same as previous one, current result is regarded as a final output, which contains ploidies and ASCNs of CNA segments and all probes.

* Supplementation of additional CNAs (“add\_cna.R”, “add\_cna.pl”, “add\_upd.pl”)

Applying clustering based on DP and MCMC to the Depth signal 2 in the final loop and BAF signals, additional CNAs are explored. In addition to that, based on abnormal BAF values ranging more than consecutive 100 probes, additional UPDs are identified. Here, we also used SNP probes excluded before initiating the loop (BAF < 0.05). Newly identified CNAs are added to the final CNA profile.