Control-FREEC: a tool for assessing copy number and allelic content using next generation sequencing data

Valentina Boeva, Tatiana Popova, Kevin Bleakley, Pierre Chiche, Julie Cappo, Gudrun Schleiermacher, Isabelle Janoueix-Lerosey, Olivier Delattre and Emmanuel Barillot.

Supplementary Methods

Workflow

1. Copy Number evaluation: GC correction, smoothing, segmentation
2. Estimation of normal contamination
3. BAF evaluation: filtering homozygous SNPs, smoothing, segmentation
4. Genotype status annotation

1. Copy Number evaluation

Calculation of copy number profiles is mainly done as described in (Boeva *et al.*, 2010). Control-FREEC accepts data produced by Illumina and SOLiD platforms. If a control sample is unavailable, Control-FREEC uses GC-content to normalize read count. The user can choose to use mappability profiles as well. This will produce a smoother copy number profile. For example, compare the profile of a normal genome (NA18507, a male Yoruba) calculated using mappability - http://bioinfo.curie.fr/projects/freec/images/NA18507_3kbW1kbS_wMapp35m2hg19.tiff with the profile for the same genome calculated without mappability information - http://bioinfo.curie.fr/projects/freec/images/NA18507_3kbW1kbS_woMapp.tiff.

2. Estimation of normal contamination

Contamination of a tumor sample by normal constitutional DNA is a frequent event seen in tumor genomic profiles. Normal contamination in sequencing data arises due to reads coming from two copies of normal DNA, and manifests itself through a shift of altered regions towards the normal baseline in the normalized copy number profile.

Control-FREEC provides two possible ways to deal with normal contamination: first, the user can apply a correction for normal contamination estimated *a priori* (for example, from the biological analysis of the sample); second, contamination can be evaluated using predicted regions of gain and loss.

In the first case, the tolerance of the algorithm to estimation of contamination is rather good (the correction will work well if the estimated contamination is within 0.2 of the real value).

In the second case, we estimate the level of contamination *c* as follows:

$$c = median\left(\frac{O_j - E_j}{ploidy - E_j}\right),$$

where O_j and E_j are the observed and predicted copy number values in window j, and *ploidy* is the *a proiri* known tumor genome ploidy. After estimation of the contamination level, the normalized read counts are recalculated using the inverse formula:

$$New\ NormReadCount_j = \frac{NormReadCount_j - c}{1 - c}.$$

If the estimated contamination is less than 0.5, the algorithm produces meaningful results. If the contamination accounts for more than 50% of normal cells or in the case of significant presence of a sub-clonal cell population in the sample (cells with different alterations), the algorithm may fail to deliver the correct prediction.

3. BAF evaluation

3.1. Preprocessing; filtering homozygous SNPs

We characterize allelic content by B allele frequency (BAF), introduced previously for SNP arrays (Popova *et al.*, 2009). We limit the list of genomic positions which we consider for allelic content evaluation to known SNPs (Sherry *et al.*, 2001). We do so because unknown (not in the current version of dbSNP) variation is thought to amount to the equivalent of only about 5% of all detected SNPs (1000 Genomes Project Consortium, 2010). We also do not consider reads with other nucleotides at SNP positions specified in the SNP annotation because we want to limit noise coming from sequencing errors and misalignment. For example, for the SNP "chr1; 10492; C/T; +; C; rs55998931", we count only reads that have C and T in the corresponding positions.

SNPs that are homozygous in the genome being considered give no information about allelic content (in SNP arrays they are denoted uninformative) and therefore putatively homozygous positions need to be discarded. To detect homozygous positions we use the probability to have variation due to sequencing error under the condition of actual homozygosity. We fix the probability to have a sequencing error equal to 0.01. Then, we calculate the binomial distribution probability to have at least $\#\text{reads}_{\{\text{Nucleotide at } i = \text{known SNP at } i\}}$ by chance under the condition of a homozygous reference allele; if this probability is greater than 0.05 we call this SNP homozygous. For the alternative allele, we call a SNP homozygous if the binomial distribution probability to have at least $\#\text{reads}_{\{\text{Nucleotide at } i = \text{reference at } i\}}$ by chance, if the position is homozygous and bears only the SNP allele, is greater than 0.01.

3.2. Smoothing and segmentation of BAF data

For a known SNP at position i, we calculate its B allele frequency:

$$\text{BAF}_i = \frac{\#\text{reads}_{\{\text{Nucleotide at } i = \text{known SNP at } i\}}}{\#\text{reads}_{\{\text{Nucleotide at } i = \text{reference at } i\}} + \#\text{reads}_{\{\text{Nucleotide at } i = \text{known SNP at } i\}}}$$

Then, for each window *j* that has not been calculated to be entirely homozygous, we evaluate:

$$\text{Med}_j = \text{median}\left(\left|\text{x}_{j,i} - 0.5\right|\right),$$

where $\{\text{x}_{j,i}\}$ are BAF values of possibly heterozygous SNP positions in window *j*. The reader may imagine that if we discard "homozygous" windows, we discard entirely homozygous segments. In reality, this does not happen. The noise is strong enough to leave most windows of such regions in play (with BAF values centered around 0.1).

We then segment the Med_j profiles using the same lasso-based algorithm as used for copy number (Boeva *et al.*, 2010; Harchaoui and Lévy-Leduc, 2008).

4. Genotype status annotation

We predict genotype status for each genomic segment independently by choosing the allelic content that corresponds to the maximal log-likelihood, given the copy number detected previously. The technique we use is similar to the GAP method described for SNP arrays (Popova *et al.*, 2009).

We combine breakpoints issued from both copy number and median B-allele frequency segmentations to get genomic segments with presumably one status. For each fragment *F*, its predicted copy number C_F is the closest integer to the median value of the normalized reads count M_F in *F*: C_F = round(M_F). We say that there is no ambiguity in C_F if abs(C_F - M_F)<0.4. Otherwise, we define the copy number call as ambiguous and consider two possible copy numbers for further genotype inference (C_F^1 and C_F^2, where abs(C_F^1 - C_F^2)=1, i.e., two consecutive integers around M_F). Note that ambiguity in copy number attribution usually means that there is significant sub-clonal cell population which has different alteration compared to the main clone.

Each genomic copy number *N* has a set of possible allelic contents denoted as $(A)_{N-k}(B)_k$, where $k \in 0..\lfloor N/2 \rfloor$, (*N*≥2), and alleles A and B are set arbitrarily. For example, possible allelic contents at the 2-copy level (*N*=2) are "AA" (2 identical, corresponding to uniparental disomy) and "AB" (2 different, corresponding to the normal genome); 3-copy level (*N*=3) has possible allelic contents "AAA" (3 identical) and "AAB" (2 identical one different); 4-copy level (*N*=4) has one of "AAAA", "AAAB" and "AABB" allelic contents, etc. Copy number and allelic content define the genotype of the genomic fragment. It is worth mentioning, that the allelic content refers to the status of a *region* not a single SNP; genomic fragment with genotype "AB" would have individual SNPs with "AA", "AB" and "BB" (homozygous and heterozygous) genotypes, while genomic fragment with genotype "AA" would have individual SNPs with "AA" and "BB" (homozygous) genotypes only. Even though homozygous SNPs have been filtered out in the presented realization of Control-FREEC, fragments with any allelic content would have residual homozygous bands (AA, AAA, etc.); however, it is the heterozygous band that determines the genotype status of a genomic fragment. The absence of a heterozygous component implies homozygous status of the fragment (LOH).

Given a possible copy number value for *F*, we fit Gaussian mixture models (GMMs) with fixed means corresponding to the possible allelic contents of the observed BAF values, and select the model that provides the maximum log-likelihood. More precisely, we fit GMMs for $(A)_{N-k}(B)_k$, where $N=C_F$ is the predicted copy number and $k \in 0..\lfloor N/2 \rfloor$, (*N*≥2). So, for *N*=2, we test genotypes "AA" and "AB"; for *N*=3, "AAA" and "AAB", for *N*=4, "AAAA", "AAAB" and "AABB". When we fit data to a genotype $(A)_{N-k}(B)_k$, we assume that we observe BAFs corresponding to $(A)_{N-k}(B)_k$, $(A)_k(B)_{N-k}$ plus some residual homozygous SNPs at $(A)_N$ and $(B)_N$. We do not fit a model if the predicted copy number is 0 or 1, because only one genotype ("0" and "A", respectively) is possible in each case.

If *k*>0, we set a minimal weight of 60% to $(A)_{N-k}(B)_k$ + $(A)_k(B)_{N-k}$ components. We also fix the maximal standard deviation of these components equal to 0.1. If there is no contamination by normal cells, the mean values of the GMM are fixed to be equal to *k*/*N*, (*N*-*k*)/*N*, 0.11 and 0.89

for $(A)_{N-k}(B)_k$, $(A)_k(B)_{N-k}$, $(A)_N$ and $(B)_N$, respectively (Suppl. Fig. 1). When tumor tissue contamination by normal cells is equal to c, the mean values are corrected as follows:

$$\text{mean for } (\mathrm{A})_{N-k}(\mathrm{B})_k = \frac{k(1-c)+c}{N(1-c)+2c}$$

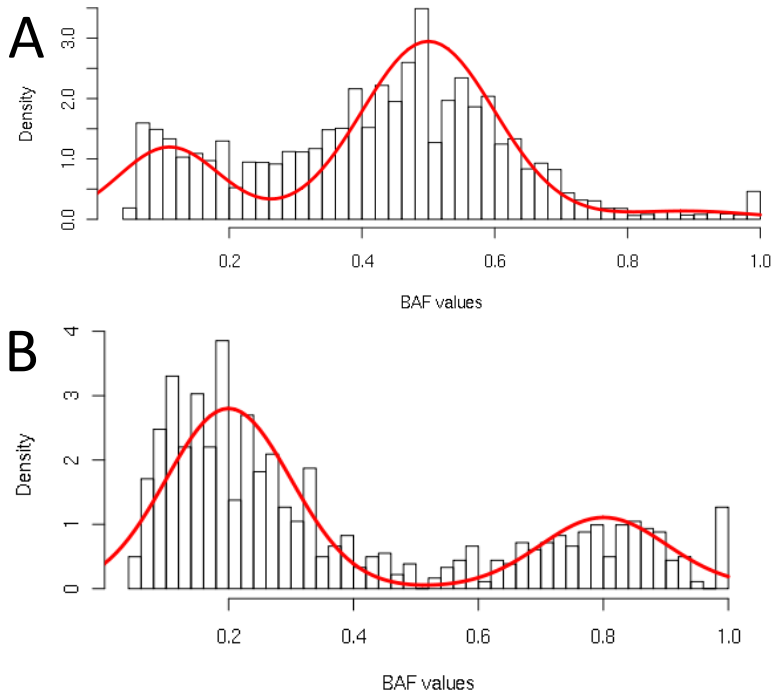
$$\text{mean for } (\mathrm{A})_k(\mathrm{B})_{N-k} = 1 - \frac{k(1-c)+c}{N(1-c)+2c}$$

$$\text{means for } (\mathrm{A})_N = \begin{cases} 0.11, & \text{when testing for k} > 0 \\ 0.11 \text{ and } \max\left(0.11, \frac{c}{N(1-c)+2c}\right), & \text{when testing for k} = 0 \end{cases}$$

$$\text{means for } (\mathrm{B})_N = \begin{cases} 0.89, & \text{when testing for k} > 0 \\ 0.89 \text{ and } \min\left(0.89, 1 - \frac{c}{N(1-c)+2c}\right), & \text{when testing for k} = 0 \end{cases}$$

After the GMM parameters have been evaluated, we calculate the log-likelihood of the data under the models. We annotate the genomic fragment with the genotype corresponding to the model that has the maximum log-likelihood. In the case of initial ambiguity in the copy number, the fragment is ascribed with the copy number corresponding to the best fit. We also output a measure m of uncertainty in the genotype prediction:

$$m = 100/(Best\ LogLikelihood - Second\ Best\ LogLikelihood).$$



Supplementary Figure 1. GMM fit with fixed means of observed BAF values. (A) The best GMM fit suggests an "AB" genotype. Fixed means are {0.11, 0.5, 0.89}. (B) The best GMM fit suggests "AA" genotype with 40% contamination by normal ("AB") cells. Fixed means are {0.11, 0.2, 0.8, 0.89}.

Differences between FREEC and Control-FREEC

The major novelty of Control-FREEC is its ability to annotate genotypes and, consequently, predict loss of heterozygosity (LOH) regions. Also, in situations when the median value of normalized read counts is far from expected integer values (e.g., observed median=3.5, so the region can have three or four copies with equal probability), the knowledge of BAF values of the region can help to better identify the copy number.

There are several other important improvements in Control-FREEC with respect to FREEC (Boeva *et al*., 2011).

- A new feature of Control-FREEC is its ability to discern somatic variants from germline ones when analyzing both a tumor and a control sample.

- Adding of an option to change minimal and maximal expected GC-content is allowed when using Control-FREEC on non-mammalian genomes.

- Also, users of Control-FREEC have more flexibility in selecting/filtering out predictions occurring in low-mappability regions or involving small DNA fragments (parameters "breakPointType", "minCNAlength", see http://bioinfo.curie.fr/projects/freec/tutorial.html).

- Working with mammalian genomes, Control-FREEC users can now set the sex of the sequenced individual. Setting "sex=XY" will result in correct annotation of chrY present in one copy in male individuals as normal.

- In Control-FREEC, we also implemented a better way to identify the final number of breakpoints per chromosome. Instead of simply using a threshold on the slope of normalized Residual Sum of Squares (see Suppl. Materials of Boeva *et al.*, 2011), we use a threshold on the slope of the slope.

Comparing Control-FREEC predictions with predictions based on SNP-array data

Control-FREEC predictions (based on whole genome sequencing data) were compared with the copy number and genotype profiles inferred using the validated GAP tool (Popova *et al*., 2009) based on SNP array data generated for the same samples on the Affymetrix SNP 6.0 platform.

We used the following parameters:

FREEC parameters	window = 3000, step = 1000, breakPointType = 4, sex = XY, minCNAlength = 3, minimalCoveragePerPosition = 5, SNPfile = hg19_snp131.SingleDiNucl.1based.txt
GAP parameters	Minimal region of variation = 15 SNPs, minimal region of detected LOH = 100 SNPs

We compared Control-FREEC and GAP results for tumor and normal blood samples of the same individual (neuroblastoma patient, ~30x-coverage, unpublished data). We obtained very similar profiles using NGS and SNP arrays in these two cases. The tumor sample presented 11 large scale copy number alterations (copy number varies from 1 to 4); all of them got identical genotype statuses using the two techniques. The normal blood sample was correctly recognized as essentially normal by both methods, while a number of short homozygosity (presumably constitutional) regions were detected by Control-FREEC and missed by SNP arrays because of lack of resolution.

More precisely, for each genomic window (3kb-windows with a step of 1kb), we compared its genotype status predicted by Control-FREEC with its genotype status predicted by GAP. We did not consider windows for which Control-FREEC or GAP were not able to make predictions, i.e., poly(N) telo-/centromeric regions or repetitive low mappability regions.

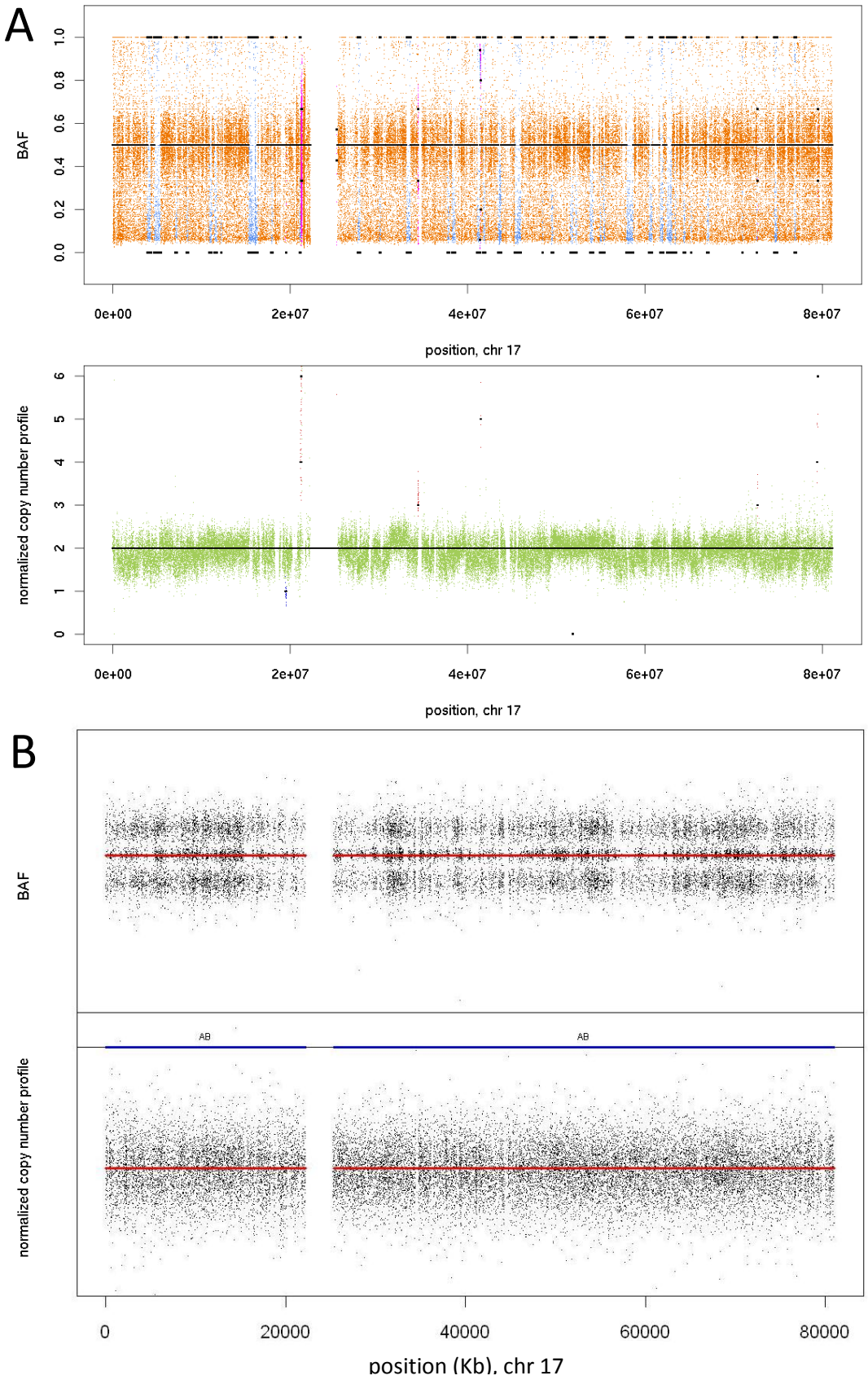
We calculated the consistency measure as

$$C = \frac{\sum_{i=0}^{n} I\left\{G^i_{Control-FREEC}=G^i_{GAP}\right\}}{n},$$

where n is the total number of considered windows, $G^i_{Control-FREEC}$ and $G^i_{GAP}$are the genotype statuses predicted in window i by Control-FREEC and GAP, respectively, and $I\{\}$ is the indicator function.

We obtained $n = 2{,}802{,}198$ valid windows (covering 2,802,198,000 bp) for the calculation of consistency of the two methods.

The consistency was high both for the blood sample ($C_{norm} = 92.3\%$) and for the tumor sample ($C_{tumor} = 95.4\%$). The difference in predictions of 7.6% for the blood sample is due to the fact that Control-FREEC predicts many small homozygous regions (716 “AA” regions, mean length 290,461 bp; Suppl. Fig. 2). These homozygous regions were missed by GAP because SNP arrays have a lower resolution.



Supplementary Figure 2. Visualization of FREEC and GAP predictions for chromosome 17 (normal sample). A. Control-FREEC predictions. Gains, losses (bottom panel) and LOH regions corresponding to "AA" status (top panel) are shown in red, blue and light blue, respectively. B. GAP predictions.

References

1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing, *Nature*, **467**, 1061-1073.

Boeva, V., et al. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization, Bioinformatics, **27**(2):268-269.

Harchaoui, Z. and Lévy-Leduc, C. (2008) Catching change-points with lasso. *Adv. Neural Inform. Process. Syst. 22*.

Popova, T., et al. (2009) Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays, *Genome Biol*, **10**, R128.

Sherry, S.T., et al. (2001) dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res*, **29**, 308-311.