Predicting copy number alterations and allelic status in cancer genomes with Control-FREEC using whole genome or exome sequencing data

Valentina Boeva, T. Popova, K. Bleakley, A. Zinovyev, J.-P. Vert, I. Janoueix-Lerosey, O. Delattre

Institut Curie, 26 rue d'Ulm, Paris, F-75248 France; INSERM, U900, Paris, F-75248 France; Mines ParisTech, Fontainebleau, F-77300 France; INSERM, U830, Paris, F-75248 France; INRIA, Saclay, France; valentina.boeva@curie.fr

Emmanuel BARILLOT

emmanuel.barillot@curie.fr

In addition to point mutations, cancer genomes often show a vast number of copy number alterations (CNAs): gains and losses of chromosomal material. Tumor suppressor genes can be completely or partially deleted, whereas oncogenes can be amplified in order to increase expression of corresponding genes. For instance, tumors of different types show deletion of chromosome arm 17p containing the most famous tumor suppressor gene *TP53*; the locus encompassing the neuroblastoma oncogene *MYCN* located at chromosome arm 2p can be present in more than fifty times in certain neuroblastoma tumors.

In some cases, instead of using the mechanism of homozygous deletion to completely disable a tumor suppressor gene, tumor cells use loss of heterozygosity (LOH): the functional allele is lost while the mutated allele is duplicated. It is thus important to be able to identify LOH regions in addition to chromosomal gains and losses in cancer studies.

With the arrival of new high-throughput sequencing technologies, our current power to detect CNA and LOH regions has significantly improved. Using high-depth whole genome sequencing or exome sequencing, we can now identify CNAs and LOH at the unprecedented resolution. In that way, we simultaneously get information about point mutations, short insertions/deletions and large chromosomal events.

We developed a method that allows detecting CNAs and LOH in whole genome sequencing data [1, 2]. Recently, we updated the method so that it can be applied to exome sequencing

data. Running time performance was also significantly improved. Within our software, we solve two important issues in the analysis of cancer genomes: contamination by normal cells and possible polyploidy. We also predict the somatic status of identified events.

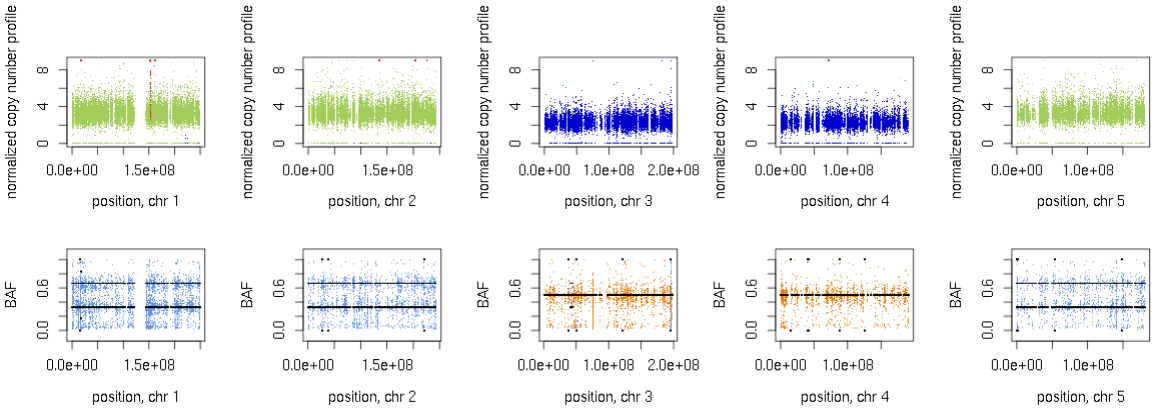

Figure 1. Visualization of Control-FREEC v6.0 output for neuroblastoma exome sequencing data (Illumina HiSeq 2000). Top panel: copy number profiles for chromosomes 1 to 5, normal copy number status is shown in green (copy number 3), losses are shown in blue; Bottom panel: allelic status, AAB/ABB (blue), AB (orange).

The algorithm normalizes read count profiles using polynomial regression; then it applies a LASSO-based segmentation procedure to predict copy number status; allelic status is inferred using nucleotide coverage of SNP positions. Only high quality positions in reads are used if the user specifies quality threshold. The user can also choose an option that will allow recheck identified copy number status using the allelic profiles. This option turns out to be extremely helpful for the analysis of noisy exome sequencing datasets.

Control-FREEC v6.0 is compatible with the SAM and pileup alignment formats and provides output files for the graphical visualization of predicted CNAs and LOH regions.

Funding: The Ligue Nationale contre le Cancer (V.B., T.P, A.Z., E.B., I.J.-L. and O.D. are members of a labeled team).

1. V. Boeva et al. (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization, *Bioinformatics*, **27**(2):268-9.

2. V. Boeva et al. (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**:423–425.