GAIA: an R-Bioconductor Package for Identification of Recurrent Genetic Mutations

MORGANELLA S (1, 3), PAGNOTTA S M (2), CECCARELLI M (1, 3)
(1) Department of Biological and Environmental Studies
University of Sannio, 82100, Benevento, Italy
(2) Department PE.ME.IS. - Statistical Section
University of Sannio, 82100, Benevento, Italy
(3) Bioinformatics CORE, BIOGEM s.c.a.r.l. - Contrada Camporeale, Ariano Irpino, Italy

Motivation
DNA copy number alterations (CNAs) are genomic regions larger than 1 kb where observed copy number differs respect to a reference genome. CNAs play a fundamental role in the developing and progression of genetic diseases: deletions involving tumor suppressor genes and chromosomal amplifications involving oncogenes have been widely observed in cancer, in addition, copy-neutral loss of heterozygosity (LOH) is receiving greater attention as a mechanism of possible tumor initiation. Identification of CNAs has been facilitated by the developing of array comparative genomic hybridization (aCGH) technology that enables copy number measurement in hundreds of thousands of genomic loci (probes). As consequence of the vast amount of produced data, great efforts have been addressed to the developing of approaches aimed at the identification of CNAs. Proposed approaches differs in many aspects such as data representation and underlying model, but all of them share the same goal, that is, the distinction between driver mutations (that play a fundamental role in cancer progression) and passenger mutations (which are random alterations with no selective advantages). Despite the high resolution of aCGH and the large number of published algorithms, accurate analysis of aCGH data is yet a challenge.

Methods
Many biological studies have been designed to look for CNAs that are common to a significant number of subjects (recurrent CNAs). These studies are based on the observation that passenger mutations are not directly related with the disease under inspection but they are due by random phenomena. Therefore, passenger mutations are present only into few samples and placed in different genomic positions. In contrast, although driver mutations can be not present in all samples, they are present in most of them and they also share the same genomic position. For this reason, counting for each probe the number of CNAs occurring contemporaneously within the dataset we can get information on what genomic positions are most likely to be sites of driver mutations. Here we
present GAIA (Genomic Analysis of Important Aberrations), an R-Bioconductor package aimed at discovering of recurrent CNAs. GAIA uses a segmentation algorithm to obtain a discrete representation of the data, so that it can gain accuracy when different sources of information are used to label the genomic regions. In order to distinguish between passenger and driver mutations GAIA implements a statistical hypothesis model based on a conservative permutation test. From observed data we obtain Bonferroni-corrected p-values that are then used into an iterative procedure to identify the most significant independent regions. These regions represent CNAs having a high probability to contain driver mutations.

Results
We applied GAIA on a real aCGH dataset composed by eight Mantle Cell Lymphoma cell lines where for each sample about 32 thousands of probes spanning the entire human genome were observed. In order to perform a qualitative evaluation of GAIA a comprehensive list of confirmed cytogenetic mutations was used. Results demonstrated the ability of GAIA in detection of most well-known biological mutations with an accurate detection of CNAs: on 25 identified cytobands 9 of them were biologically confirmed. In addition, previously undescribed cytogenetic aberrations were found highly significant and so they deserve for further biological investigations.

Availability

Contact email
morganella@unisannio.it