QDNAseqFlow: A Computational Analysis Workflow of DNA Copy Number Aberrations from low-coverage whole genome sequencing reads.

Authors:

Christian Rausch (corresponding author)

Beatriz Carvalho

Remond Fijneman

Gerrit Meijer

Mark van de Wiel

BACKGROUND: Gains and losses of genetic material, also known as DNA copy number alterations are aberrations that are involved in the development of cancer. Their analysis is therefore critical for research and diagnostics in oncology. DNA sequencing based determination of copy number aberrations is becoming the most cost effective way as compared to microarray based techniques at equal resolution. To obtain copy number calls from low coverage whole genome sequencing reads requires the combined usage of several programs with various steps, followed by more statistical analysis tools for pairwise comparisons, etc. A complete workflow would therefore be useful for many other researchers in the field.

RESULTS: Here we present QDNAseqFlow, a computational workflow that produces DNA copy number plots along with various summaries and statistics, including the aberration differences found between groups of input samples. Written in the R programming language, it relies on Bioconductor packages QDNAseq, DNAcopy, CGHcall and CGHregions as well as the open-source R packages NoWaves and CGHtest, all of them described in peer-reviewed journal articles.

USAGE: The program is written in the R programming language and can be run without programming skills on Windows, MacOSX and Linux through provided wrapper scripts. The user is guided by simple graphical pop-ups to enter parameters or select file locations, while access to the program code allows users with R programming skills to change advanced parameters. FEATURES and WORKFLOW: (1) Reads obtained from low-coverage (= "shallow") whole genome sequencing of DNA samples need to be provided as BAM files obtained by alignment to the human reference genome hg19. (2) Copy number plots and -files are created using Bioconductor package QDNAseq. (3) 'Waves' in the profiles are smoothed with the R package NoWaves (van de Wiel et al., 2009) and subsequently, aberrated regions are combined with the circular binary segmentation (CBS) algorithm implemented in Bioconductor package DNAcopy and the copy numbers of obtained segments are called using Bioconductor package CGHcall. (3) Summarizing frequency plots and quality statistics for all plots are created. Plots are flagged if their noise and/or number of segments is higher than expected, based on the inter-quartile range of values observed for all samples, and can then be checked and removed by the user from subsequent

analysis. (4) If the user provides a grouping for the samples, individual frequency plots, aberration summaries (per chromosome arm) and a differential aberration analysis will be produced. To obtain the latter, Bioconductor package CGHregions is used to slightly adjust the segments in all samples in a way to obtain regions with start and end positions identical in all samples with minimal information loss. Then, with the help of R package CGHtest (van de Wiel et al., 2005), a Wilcoxon-Mann-Whitney two-sample test or Kruskal-Wallis k-sample test is applied to all aberrated regions to calculate which aberration is significantly different between the groups.

CONCLUSIONS: QDNAseqFlow is a comprehensive workflow for the analysis of copy number aberrations. It is available at github.com/NKI-Pathology.