QDNAseqFLOW: A Computational Analysis Workflow of DNA Copy Number Aberrations from Low-Coverage Whole Genome Sequencing Reads

Christian Rausch [1,3], Beatriz Carvalho [1], Remond Fijneman [1], Gerrit Meijer [1], Mark van de Wiel [2]

[1] Department of Pathology, Netherlands Cancer Institute, Amsterdam, Netherlands; [2] Department of Epidemiology & Biostatistics and Department of Mathematics, VU University Medical Center and VU University, Amsterdam, Netherlands; [3] Contact: c.rausch@nki.nl

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

Abstract

- This workflow, written in the R programming language
  - Is useful to:
    - determine and visualize gains and losses of genomic DNA (see fig. 1)
    - determine statistical differences in the aberration pattern between 2 groups
  - Requires cost-effective low coverage (≥0.1 x) whole genome sequencing data
  - 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
  - Implements own functions for quality control and statistics of genomic aberrations
  - Can be run by users with no programming experience by invoking 3 main R scripts, on Windows, Linux or MacOSX
  - Is available from github.com/NKI-Pathology incl. further info

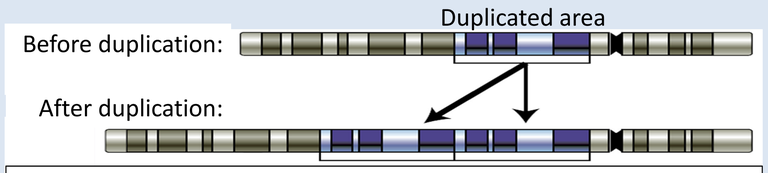

Fig. 1: Example: loss of genomic material, image source: Talking Glossary of Genetic Terms

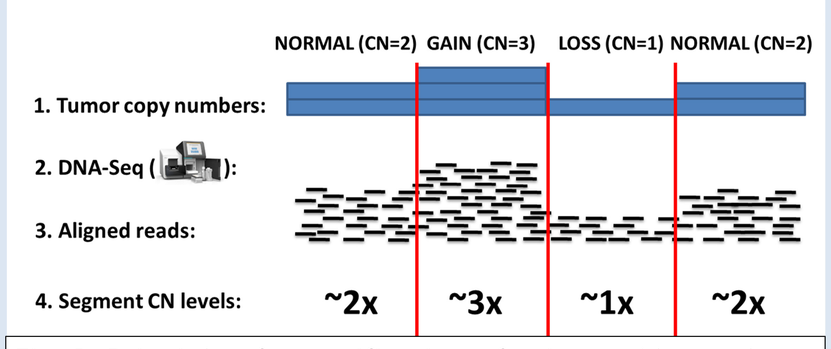

Fig. 2: Principle of quantification of gains and losses by read counts

Data Preprocessing

Input: seq. reads as fastq files → QC of fastq (FastQC & MultiQC) → If needed: Trim low quality + adapters → fastq files

Map reads to ref. genome hg19 w. unlocalized and unplaced contigs (BWA mem) → Picard Mark Duplicates → BAM files → QC of BAM (FastQC & MultiQC)

Samples passing FastQC/MultiQC quality criteria, esp. regarding
- Sequence quality
- Overrepresented sequences and
- Adaptor content

→ Output: BAM files

Workflow

QDNAseqFlow step ① → raw read counts

Input: BAM files, bin size, Output folder, Project name

Script: `QDNAseq_BAM2CopyNumbers.R`
- Select input via GUI or command line
- Available bin sizes: 1, 5, 15, 30, 50, 100, or 1000 kb.
- Reads per bin are counted
- Output as .rds, .igv, and .tsv table

QDNAseqFlow step ② → copy number calls & stats

Input: read counts .rds, Dewaving: yes/no, Inclusion list .xls, Output folder, Project name

Script: `QDNAseq_CopyNumbers2PlotsCallsSegmentFileInclDewave.R`
- Select input via GUI or command line
- Optional: Dewaving, removing a wavy pattern seen in some plots, using NoWaves algorithm (effect see fig. 3 l.+m.)
- Optional: select 'inclusionlist.xls' to continue only with a selection of samples in input rds
- Removal of ~7% known noisy bins
- Correction of read counts based on 2D LOESS function of GC-content and mappability
- Median-normalization
- Smoothing of outliers

Output:
- plots (png) of smooth, segmented and called copy number profiles
- frequency plot of aberrations found in whole group (example see fig. 3r)
- calls as file (txt & rds)
- Statistics:
  - segments, noise, read counts per plot, outliers highlighted based on 1.5xIQR rule
  - percentage and counts of aberrations per chromosome arms

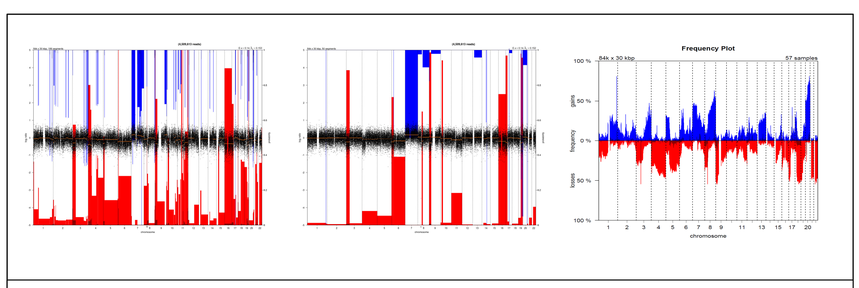

Fig. 3: Left and middle: Copy number call plot before and after dewaving with NoWaves. Right: frequency plot.

QDNAseqFlow step ③ → test differentially aberrated regions

Input:
- Called copy nb (rds file)
- List of samples and their group membership (csv table)

Script: `QDNAseq_FrequencyCGHregionsCGHtest.R`
- Select input via GUI
- Groups of samples are defined via input csv table
- Make frequency plots of aberrations of all groups
- Genomic segments are joint to regions →one file for each pair of sample groups using CGHregions
- Pairs of groups are tested for significantly differently aberrated regions (gained vs. non-gained regions & lost vs. non-lost regions), according Wilcoxon-Mann-Whitney- and χ^2-tests using CGHtest, see example output in tab. 1.

Output:
- Frequency plots for every group
- Tables of significant regions, all possible pairwise comparisons (see tab. 1)

chromosomes	bpstart	bpend	bins	pvalue	fdr
7	151800001	157080001	156	0.0152	0.10
18	28590001	29610001	33	0.0024	0.09

Tab. 1: Example output of CGHtest: tested regions have significant differences regarding losses in the mentioned regions.

References

Please follow links to software and scientific publications:
BWA mem: bio-bwa.sourceforge.net
FastQC: www.bioinformatics.babraham.ac.uk/projects/fastqc
MultiQC: multiqc.info
QDNAseq, DNAcopy, CGHcall and CGHregions: bioconductor.org
NoWaves and CGHtest: www.few.vu.nl/~mavdwiel/software.html

Acknowledgement

We thank Daoud Sie for help with QDNAseq.