MetaCNV - a consensus approach to infer accurate copy numbers from low coverage data

CURRENT STATUS: UNDER REVISION

BMC Medical Genomics BMC Series

Stefanie Friedrich
Stockholms Universitet

✉ Stefanie.Friedrich@scilifelab.se *Corresponding Author*
ORCID: <https://orcid.org/0000-0002-3889-5589>

Remus Barbulescu
Stockholms Universitet

Thomas Helleday
Karolinska Institutet

Erik LL Sonnhammer
Stockholms Universitet

DOI:

10.21203/rs.2.15757/v1

SUBJECT AREAS

Epigenetics & Genomics

KEYWORDS

Human genome analysis, copy number calling, low coverage data

Abstract

Background : The majority of copy number callers requires high read coverage data that is often achieved with elevated material input, which increases the heterogeneity of tissue samples. However, to gain insights into smaller areas within a tissue sample, e.g a cancerous area in a heterogeneous tissue sample, less material is used for sequencing, which results in lower read coverage. Therefore, more focus needs to be put on copy number calling that is sensitive enough for low coverage data.

Results : We present MetaCNV, a copy number caller that infers reliable copy numbers for human genomes with a consensus approach. MetaCNV specializes in low coverage data, but also performs well on normal and high coverage data. MetaCNV integrates the results of multiple copy number callers and infers absolute and unbiased copy numbers for the entire genome. MetaCNV is based on a meta-model that bypasses the weaknesses of current calling models while combining the strengths of existing approaches. Here we apply MetaCNV based on ReadDepth, SVDetect, and CNVnator to five datasets in order to demonstrate how the approach improves copy number calling. **Conclusions :** MetaCNV, available at <https://bitbucket.org/sonnhammergroup/metacnv> , provides accurate copy number prediction on low coverage data and performs well on high coverage data.

Full-text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed.

However, the manuscript can be downloaded and accessed as a PDF.

Tables

Table 1. Cancer cell lines used for accurateness testing. Coverage and number of deleted and amplified genes extracted from COSMIC are presented. Copy numbers were called from sequenced genomes and experimentally confirmed with PICNIC [34].

Cancer cell line	Coverage	# of amplified genes	# of deleted genes	Total # of genes with CNVs	Ratio amplified/deleted genes
HCC1187	104x	154	46	200	3.48
HCC2218	93x	200	36	236	5.55
MCF7	62x	49	30	79	1.63
PC3	76x	179	222	401	0.81
Single-cell SKBR3	6 stepwise merged single cells with 1x to 6x	161	29	190	5.55

Figures

Figure 1

MetaCNV workflow to call copy numbers. MetaCNV requires as input an SVDetect prediction using a simulated null alignment (simNull) as matched sample (Figure S14).

Figure 2

Genome segmentation by MetaCNV. A. Segmentation of the genome according to the bins and breakpoints from the input callers. B. Consensus segment prediction by MetaCNV is marked as a thick line. Bins c is predicted as amplified and bin d as deleted. Here there is no conflict between the input callers, hence the consensus is that c is amplified and d is deleted. Bin f is not predicted by SVDetect, but because ReadDepth predicts it as an amplification, this becomes the consensus. Bins b, e and g have conflicting ReadDepth and SVDetect predictions. CNVnator judges that e and g are amplifications, hence this becomes the consensus. Bin b is set to CN 2 because only ReadDepth predicts it as an amplification and CNVnator makes no prediction (Table S14).

Figure 3

Distribution of absolute copy numbers. Copy numbers were called with ReadDepth of cancer cell line HCC2218 after segmentation (6,867,679 segments with an average segment length of 450 bp). A. The bias occurs around each integer copy number and is lower for CNRD ~ 1 , but increases for CNRD ~ 2 . The factor for normalisation is a linear function (equation 1b) which serves to adjust the bias for different CNRD. B. Distribution of absolute copy numbers called with ReadDepth after normalisation

Figure 4

Prediction accurateness of MetaCNV on low coverage data, compared to CNVnator, Control-FREEC, ReadDepth, and SVDetect. A and B present the benchmark results with MLRE and MCC per merged alignment. C and D show the overall results averaged across all benchmarked alignments.

Figure 5

Prediction accurateness of MetaCNV and other copy number callers for high coverage data.

A and B present the average benchmark results per cancer cell line. SVDetect using matchedNormal could only be performed for HCC1187 and HCC2218, for which matched blood samples were available. C and D show the overall results averaged across all benchmarked cell lines.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

metaCNV_manuscript_supplement.docx
 CCL_knownCNVs.xlsx
 MetaCNV_README.txt