LabelSeg: A new strategy for segment annotation of tumor copy number alteration profiles

ELIXIR All Hands , 5-8 June 2023, Dublin, Ireland

Background

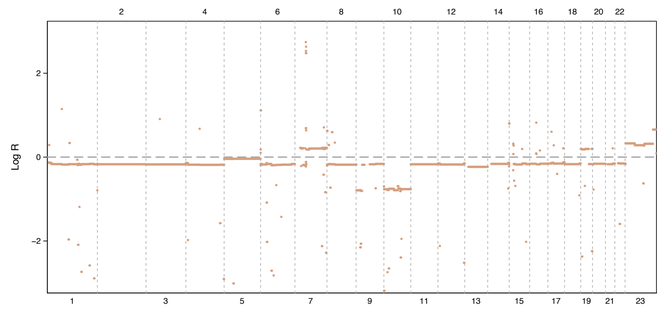
What is SCNA?

Somatic copy number alterations (SCNAs) refer to changes in the copy number of chromosome segments arising in somatic (i.e. post germline) tissues. SCNAs are prevalent in many types of cancer and overall represent the by extent largest contributions to genomic variation in cancer.

Challenges in interpretation of SCNA segment profiles

Difficult to estimate copy number ratio from measured signal because

- Normal sample contamination
- Aneuploidy and subclones of tumor cells
- Platform-based issues



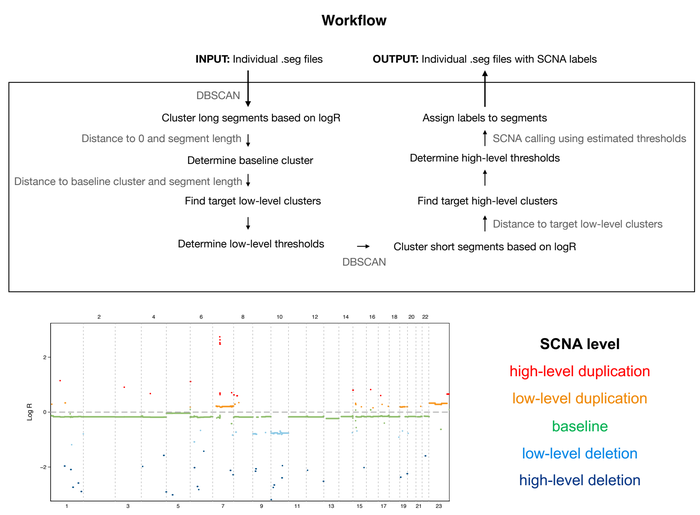
Current calling methods

- Direct cut-off: simple, fast, but fixed thresholds are flexible parameters that may not apply to all samples
- Estimation supported by alliec information: not available in some cases

Method

LabelSeg

Adaptive multi-level thresholds estimation



Validation

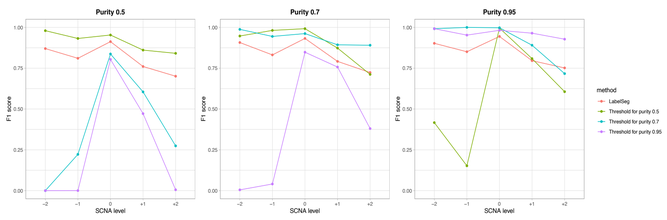


Fig. 2. Performance on simulated data. Average F1 scores across samples were calculated when calling different levels of SCNA.

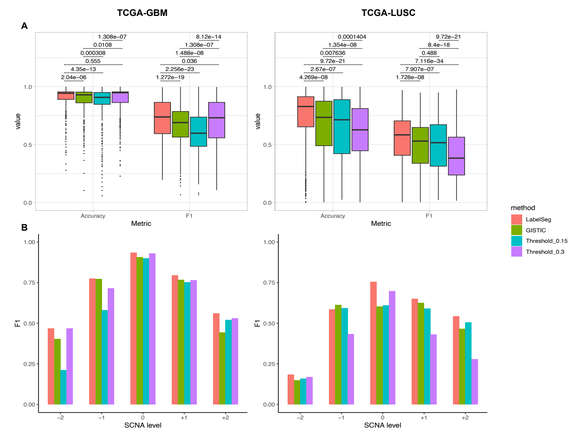


Fig. 3. Performance in TCGA datasets. (A) Balanced F1 score and accuracy for each sample. (B) The average F1 score across samples in calling different levels of SCNA.

Conclusion

***LabelSeg* exhibited robustness towards varying levels of tumor sample purity and higher performance than other methods. In addition, it demonstrated advantages in handling more challenging data owing to its robustness .**

Application

Comparative analysis

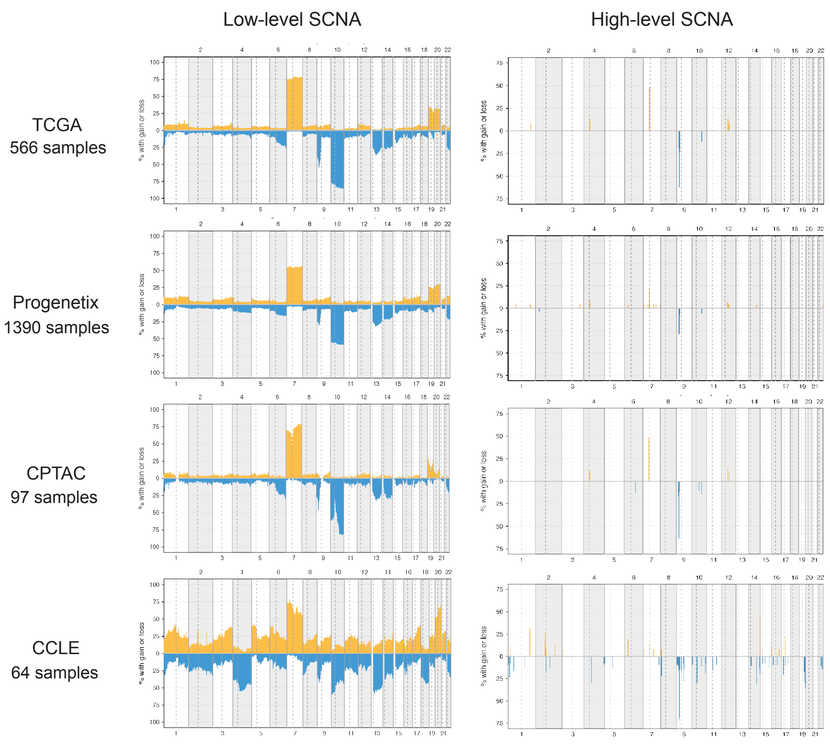


Fig. 4. Frequency of SCNA called in different datasets of glioblastoma samples. Yellow represents duplication. Blue represents deletion. The Y-axis is the percentage of samples with SCNA overlapping with the genomic bin of 1MB size. The numbers on the X-axis represent chromosomes. Low-level SCNAs are called SCNAs with labels "+1" and "-1". High-level SCNAs are called SCNAs with labels "+2" and "-2". Background noise peaks were filtered in the frequency plots of the high-level SCNAs.

Conclusion

***LabelSeg* could provide a reliable calling no matter how heterogeneous the data are in terms of measurement platforms (input is segment file), tumor sample purities, and noise (clustering-based estimation).**

Correlation between copy number dosage and mRNA

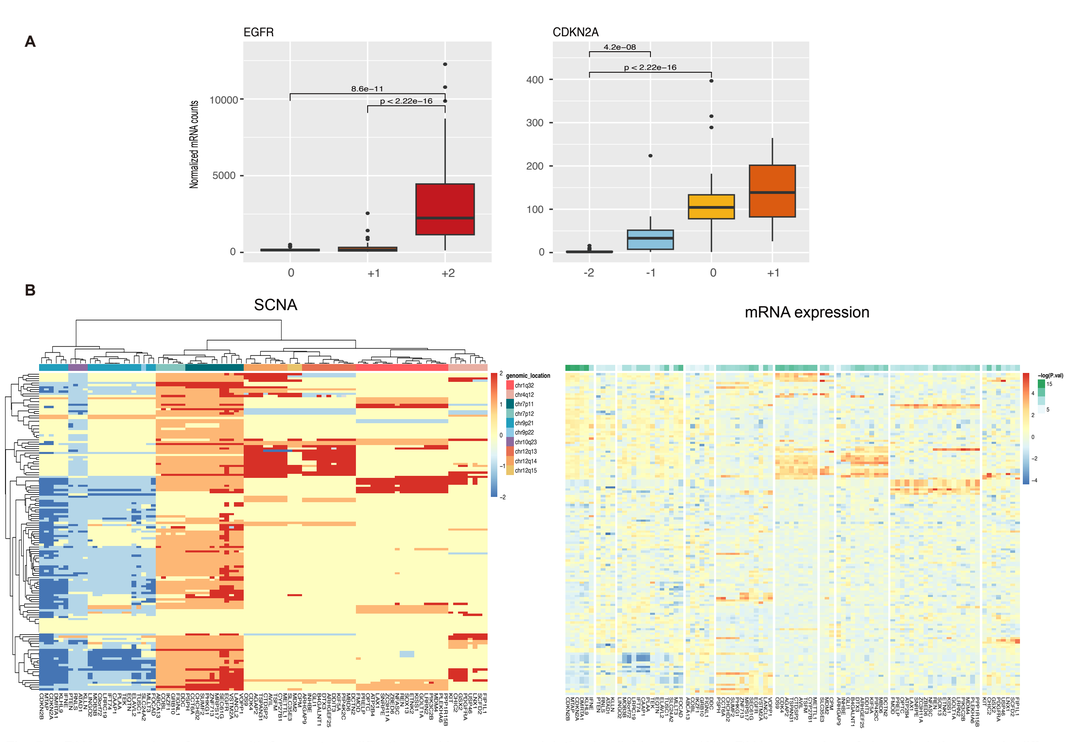


Fig. 5. mRNA expression of genes with frequent amplification and homozygous deletion in glioblastoma (A) mRNA expression of characteristic genes in different SCNA levels. (B) Heatmap of SCNA level and mRNA expression in TCGA glioblastoma samples. Rows represent samples and columns represent genes. In the left SCNA heatmap, the values are SCNA labels. In the right mRNA heatmap, the row and column order is the same as in the left plot. TMM-normalized (31) mRNA counts were log-transformed and standardized across samples per gene for visualization. BH-adjusted P values were calculated using the Kruskal-Wallis rank sum test on normalised mRNA counts.

Conclusion

We observed strong correlation between matched mRNA expression and SCNA levels assigned by *LabelSeg*, indicating the important role of (high-level) SCNA in cancer development.

Contact

Hangjia Zhao, Michael Baudis

University of Zürich, Switzerland

Swiss Institute of Bioinformatics

hangjia.zhao@uzh.ch

University of Zurich[UZH]

SIB Swiss Institute of Bioinformatics

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

@ELIXIREurope

/company/elixir-europe