LabelSeg: segment annotation for tumor copy number alteration profiles

Hangjia Zhao[1,2], Michael Baudis[1,2*]

[1]Department of Molecular Life Sciences, University of Zurich
[2]Swiss Institute of Bioinformatics

1 Abstract

Somatic copy number alterations (SCNA) are a predominant type of oncogenomic alterations that affect a large proportion of the genome in the majority of cancer samples. Current technologies allow high-throughput measurement of such copy number aberrations, generating a large number of SCNA profiles. However, annotation and integration of these profiles are challenging due to the presence of multiple sources of noise and heterogeneity in measurement platforms. In this study, we present *LabelSeg*, an algorithm for fast and accurate annotation of CNA segments, to improve the interpretation of tumor SCNA profiles. *LabelSeg* uses segment files as input to estimate calling thresholds for identifying different relative copy number states and is compatible with most CNA measurement platforms. We confirmed its performance on simulated data and sample-derived data from The Cancer Genome Atlas (TCGA) reference dataset, and we demonstrated its utility in integrating heterogeneous segment profiles from different data sources. Our comparative and integrative analysis revealed common SCNA patterns in cancer and protein-coding genes with a strong correlation between SCNA and mRNA expression, promoting the investigation of the role of SCNA in cancer development.

2 Introduction

Genomic instability is a nearly ubiquitous hallmark of cancer. Cancer cells often lose the ability to maintain genome integrity, but the molecular basis of genomic instability is not always clear [1]. One consequence of genomic instability is the occurrence of somatic copy number alterations (SCNA), which are changes in the copy number of chromosome segments from the regional allele count in somatic (i.e. post germline) tissues. SCNA represent the by extent largest contributions to genomic variation in cancer, with genetic components affected by SCNA frequently conferring selective advantages to affected cells, thereby promoting cancer initiation and progression [2].

Various methods are employed to detect SCNAs ranging from cytogenetic and locus-specific techniques such as karyotype analysis, interphase fluorescence in-situ hybridization (FISH), and spectral karyotyping (SKY) to more recent technologies such as genomic microarrays and next-generation sequencing (NGS) methods. Microarrays used for CNA calling include various types of comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) based arrays. With the advent of NGS, many tools have been developed in recent years to enable CNA detection from sequencing data. However, no single approach can unanimously capture all CNA information, as different technologies and platforms have different detection biases. Examples for such differences include upper and lower detection sensitivity for CNA events of various sizes, the need for matched reference samples, or the ability to detect allele-specific CNA events [3].

Beyond such considerations when evaluating individual platforms, the meta-analysis of large and heterogeneous SCNA datasets presents additional challenges. A major hurdle is the interpretation and comparison of segmented copy number profiles, which are often the only accessible data due to patient privacy concerns. Segmented data reflect changes in relative copy number rather than absolute copy number ratios, posing challenges in comparing results across datasets based solely on signal values. This is further complicated by the fact that signal scale and noise level can vary widely across samples due to variations in clonal sample purity, ploidy, experimental steps in bio-sample preparation, measurement platform, and other factors. To overcome this problem, some tools such as CNVkit [4] and VarScan2 [5] choose fixed, empirical thresholds and classify segments with signals beyond these thresholds into "duplication" or "deletion" CNA types. However, the choice of these thresholds can significantly impact the results of such analyses. Low cut-off values improve the sensitivity of variant detection - especially for samples with admixed non-cancer tissue - but can lead to many false positive calls. In contrast, large cut-off values may improve the calling precision but may miss true variants and thus introduce systematic bias e.g. due to cancer-specific differences in sample purity. Some studies [6, 7] optimize this process by adding purity estimation and adjusting thresholds based on the estimated purity. However, the additional estimation makes SCNA profiling more complicated, which requires manual selection of the best solutions [8]. Several complex models have been developed to improve

SCNA detection, including Gaussian mixture models [9, 10] and hidden Markov models [11, 12]. Although theoretically promising for providing more accurate CNA calls, these CNA detection models often require raw data from individual measurement platforms that may not be accessible or allele specific data which some technologies don't include. Moreover, most of these methods are designed to provide an absolute copy number quantification rather than the identification of different empirical CNA types, such as broad CNA and focal CNA, which differ in size, magnitude, and potential functions [13]. While GISTIC [14] can discriminate between CNA types or levels, its primary aims are to identify recurrent amplified or deleted genomic regions across a set of samples, rather than precisely calling CNA levels in individual samples. Thus, the fast and accurate annotation of individual CNA segment profiles remains a challenging problem in the field of SCNA profiling.

Here we developed a new method called *LabelSeg* for CNA segment classification and annotation, which estimates calling thresholds from individual segment profiles to identify different CNA levels. The one-dimensional clustering and direct cut-off approach using estimated thresholds enable rapid processing of CNA profiles. The only input required is copy number segment profiles, which can be generated by most CNA measurement platforms and processing pipelines. Therefore, *LabelSeg* is particularly suited to large-scale meta-analyses involving tens or hundreds of individual studies with a total of tens of thousands of samples and more. In the two cancer validation cohorts from TCGA [15], *LabelSeg* outperformed GISTIC and fixed thresholds. The integrative analysis spanning 4 research projects, > 2000 glioblastoma samples, and > 1200 lung squamous cell carcinoma samples demonstrated that *LabelSeg* is capable of achieving fast, accurate, and comprehensive CNA profiling of large numbers of heterogeneous cancer samples, paving the way for future comparative CNA analysis in cancer.

3 Materials and Methods

LabelSeg combines segment length and log ratios (logR) to estimate appropriate thresholds for calling different levels of SCNA (Figure 1). There are several assumptions. First, the majority of detected copy number events are driven by predominant clones. It is a prevalent biological assumption that serves as the foundation for most contemporary algorithms utilized in the detection of SCNAs [16]. Under this assumption, segments of individual profiles form distinct clusters in logR values. These clusters are likely to represent different copy number states. Second, arm-level SCNAs are generally low copy number changes, whereas focal SCNAs can be of very high amplitude. This length-amplitude relationship of SCNA, which has been previously reported [13], allows reliable discrimination of different SCNA levels, including low-level duplication/deletion and high-level duplication/deletion. Currently, the magnitudes of CNA levels are inconsistently defined across different studies [17–23]. In general, high-level CNAs involve substantial changes in the copy number of chromosomal segments, with high-level duplication indicating the presence of multiple copies of certain genomic regions, and high-level deletion indicating either the complete loss or substantial reduction of specific regions. In contrast, low-level CNAs involve smaller changes in copy number compared to high-level CNAs.

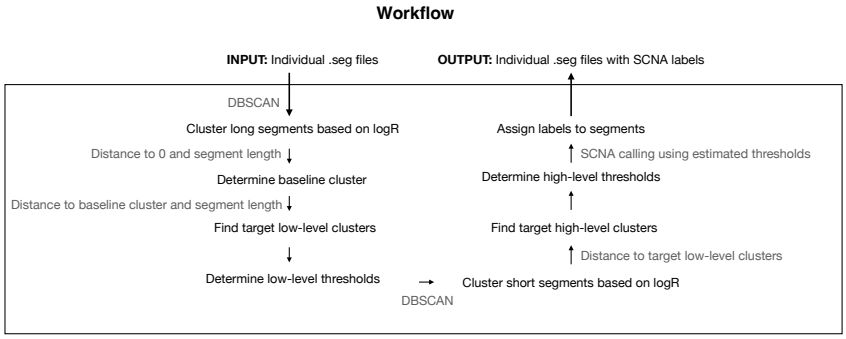

Figure 1: Workflow of *LabelSeg*

3.1 Algorithm

3.1.1 Clustering

The segments in a sample are divided into long and short segments based on their length relative to the corresponding chromosomes. The criterion for segment size separation was determined based on the empirical distribution of segments derived from a combined dataset across various tumor types (Supplementary Figure S2). Segments occupying $\geq 20\%$ of a chromosome are considered as long segments, and segments occupying $< 20\%$ of a chromosome are considered as short segments. Long segments and short segments are clustered by logR

values respectively using DBSCAN [24]. The intuition is that long segments are generally generated from more measurement markers and represent broad CNA. As a result, long segments have much lower variance and scales in logR than short segments. There are two important parameters in DBSCAN with respect to its application in our method: ϵ and $minPts$. The parameter $minPts$ is the minimum number of points required to form a dense region. It is set to 1 in this method because it is possible for a segment to form a cluster (e.g. focal amplification). The parameter ϵ in DBSCAN is the maximum distance between two points for one to be considered as in the neighborhood of the other. This parameter is flexible in each estimation, starting from a self-defined initial value (0.05 and 0.1 for long and short segment clustering respectively) and decreasing by 0.01 until the standard deviations of all clusters are smaller than a certain threshold (0.05 and 0.1 for long and short segment clustering respectively).

After clustering, we compute a feature value V_c for each cluster c. For long segment clusters, the feature value is the segment-length weighted mean. This is used to compute thresholds for calling low-level SCNAs. For short segment clusters, the feature value is the value closest to 0, e.g. the minimum value in a cluster of positive logR values or the maximum value in a cluster of negative logR values. Thresholds for calling high-level SCNAs are calculated from the feature values in short segment clusters.

3.1.2 Estimate baseline

The baseline is determined from the long segment clustering results. The cluster with the feature value closest to 0 and occupying $\geq 40\%$ of the total measured genomic region is considered as the baseline cluster c_{baseline}. If the noise in the data is high or the profile is over-segmented, it is possible that no long segment clusters occupy more than 40% of the total length. In this case, a step-wise reduction of the percentage limit, from an initial value of 40% to 20%, by a decrement of 10%, is performed until the baseline cluster is ascertained.

During this step, users can interactively adjust the baseline up or down. The algorithm will subsequently identify the cluster that is closest to the predetermined baseline cluster and satisfies the aforementioned conditions.

3.1.3 Estimate low-level calling threshold

The next step is to find clusters that represent low-level SCNA events. The remaining long segment clusters are ranked by the sum of segment length from largest to smallest, denoted by $\{c_1, c_2, ..., c_k\}$. The target low-level clusters are determined as follows.

$$c_{\text{low-dup target}} = c_a$$
$$\text{where } a = \underset{i \in \{1,\ldots,k\}}{\arg\min}\ V_{c_{\text{baseline}}} + 0.15 \leq V_{c_i} < V_{c_{\text{baseline}}} + 0.7$$
$$c_{\text{low-del target}} = c_b$$
$$\text{where } b = \underset{i \in \{1,\ldots,k\}}{\arg\min}\ V_{c_{\text{baseline}}} - 1.5 < V_{c_i} \leq V_{c_{\text{baseline}}} - 0.15$$

'Target cluster' is the cluster used to estimate calling thresholds. The lower bound ± 0.15 is to filter out noise, and the higher bounds 0.7 and -1.5 for duplication and deletion are to avoid calling of high-level SCNA. Length-based ranking increases tolerance to sub-clone effects.

To ensure calling sensitivity, the standard deviation of the target cluster is included in the calculation of calling thresholds. Suppose T is the calling threshold, and σ_c is the standard deviation of cluster c. The calling threshold for low-level SCNA is calculated as follows:

$$\mathrm{T}_{\text{low-level duplication}} = V_{c_{\text{low-dup target}}} - 2 \times \sigma_{c_{\text{low-dup target}}}$$
$$\mathrm{T}_{\text{low-level deletion}} = V_{c_{\text{low-del target}}} + 2 \times \sigma_{c_{\text{low-del target}}}$$

It is noted that the cluster standard deviation is usually underestimated since a variance control is applied in the initial clustering. Therefore, the cluster standard deviation is modified if it is anomalous (see the Supplementary Data).

3.1.4 Estimate high-level calling threshold

Estimation of high-level calling thresholds uses clustering of short segments, target low-level clusters determined in previous steps, and the numerical relationship of logR values from different copy number states. This numerical relationship is robust to variable tumor sample purity and ploidy (see the Supplementary Data). Therefore, this method could find a reliable calling threshold in heterogeneous samples and is able to discriminate high-level duplication/deletion from low-level duplication/deletion.

Short segment clusters are ranked by feature values from smallest to largest, denoted as $\{c_1, c_2, ..., c_m\}$. The target high-level clusters are determined as follows:

$$c_{\text{high-dup target}} = c_d \text{ where } V_{c_d} > V_{c_{\text{low-dup target}}} \ \&$$
$$d = \underset{i \in \{1,\ldots,m\}}{\arg\min}\ \frac{V_{c_i} - V_{c_{\text{baseline}}}}{V_{c_{\text{low-dup target}}} - V_{c_{\text{baseline}}}}$$
$$\geq 2.2 - 0.6 \times (V_{c_{\text{low-dup target}}} - V_{c_{\text{baseline}}})$$
$$c_{\text{high-del target}} = c_e \text{ where } V_{c_e} < V_{c_{\text{low-del target}}} \ \&$$
$$e = \underset{i \in \{1,\ldots,m\}}{\arg\max}\ \frac{V_{c_{\text{baseline}}} - V_{c_i}}{V_{c_{\text{baseline}}} - V_{c_{\text{low-del target}}}} \geq 2$$

To better separate high-amplitude focal SCNA from low-amplitude broad SCNA, only focal SCNAs with an amplitude greater than broad SCNAs are called high-level SCNAs. The calling threshold for high-level SCNAs is calculated as follows. Let $\text{logR}_{\mathbf{L}}$ be the set of logR values of all previously defined long segments,

$$\text{T}_{\text{high-level duplication}} = max(V_{c_{\text{high-dup target}}}, max(\log R_{\mathbf{L}}) + 0.01)$$
$$\text{T}_{\text{high-level deletion}} = min(V_{c_{\text{high-del target}}}, min(\log R_{\mathbf{L}}) - 0.01)$$

3.1.5 SCNA Calling

Once calling thresholds have been estimated, the segments are assigned labels that represent relative copy number states or SCNA levels. The segments with $\text{logR} \leq \text{T}_{\text{high-level deletion}}$ are labeled as "-2", indicating high-level deletion. If the sample is diploid, the high-level deletion is a homozygous deletion. The segments with $\text{logR} > \text{T}_{\text{high-level deletion}}$ and $\leq \text{T}_{\text{low-level deletion}}$ are labeled as "-1", meaning low-level deletion. The segments with $\text{logR} \geq \text{T}_{\text{low-level duplication}}$ and $< \text{T}_{\text{high-level duplication}}$ are labeled as "+1", meaning low-level duplication. The segments with $\text{logR} \geq \text{T}_{\text{high-level duplication}}$ are labeled as "+2", meaning high-level duplication. The segments with $\text{logR} > \text{T}_{\text{low-level deletion}}$ and $< \text{T}_{\text{low-level duplication}}$ are labeled as "0", meaning no SCNA. There are other exceptions. For example, the segment profile is over-segmented and there are no long segments, or only high-level focal SCNAs occurred in a sample. Details of the strategy for dealing with these exceptions are given in the Supplementary Data.

3.2 Implementation

LabelSeg was written under R and the software is available at `https://github.com/baudisgroup/LabelSeg`. The average runtime was approximately 0.00546 seconds per sample with 200 segments on a MacBookPro18,4 (10 core, Apple M1 Max, 64 GB RAM).

4 Results

4.1 Performance in simulated data

We evaluated the performance of *LabelSeg* on simulated data by simulating various scenarios with different sample purities and comparing it with optimal thresholds for specific sample purities (Figure 2). The generation of simulated data was based on Willenbrock and Fridlyand's work with a slight modification [25], and details are provided in the Supplementary Data. As expected, the optimal threshold consistently demonstrated superior overall performance when applied to samples with the true tumor purity for which the threshold was optimized. However, the efficacy of the threshold was compromised when applied to samples with purity values substantially different from the presumed purity. In contrast, *LabelSeg* exhibited high F1 scores across all scenarios and demonstrated robustness towards varying levels of tumor sample purity.

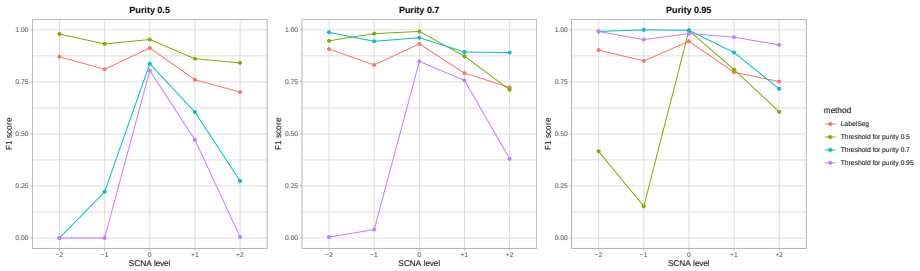

Figure 2: Performance on simulated data. Average F1 scores across samples were calculated when calling different levels of SCNA.

4.2 Performance in real data

To assess the performance of *LabelSeg* on real biological data, we analyzed two TCGA datasets of glioblastoma and lung squamous cell carcinoma samples, and compared it with other methods, including GISTIC, a set of more stringent thresholds $\{\pm 0.3, \pm 1\}$, and a set of more relaxed thresholds $\{\pm 0.15, \pm 0.7\}$ (Figure 3 A). ASCAT calls derived from the same datasets were used as references. In the GBM cohort, *LabelSeg* had the best performance in terms of balanced F1 score and accuracy while the stringent threshold (Threshold_0.3) also outperformed in terms of accuracy (adjusted P-value < 0.05 by Benjamini–Hochberg method (BH), paired Wilcoxon rank sum test). The relaxed threshold (Threshold_0.15) had the worst performance. In the LUSC cohort, *LabelSeg* performed the best again in terms of balanced F1 score and accuracy (BH-adjusted P-value < 0.05, paired Wilcoxon rank sum test), followed by the relaxed threshold and GISTIC. In contrast to the GBM dataset, the stringent threshold didn't work well in the LUSC dataset, indicating the limitation that fixed thresholds are difficult to adapt to various tumor purities, as previous research has shown

that the average sample purity is higher in GBM samples than in LUSC samples [26]. In both datasets, *LabelSeg* exhibited the highest average F1 score for calling all SCNA levels, with the exception of calling low-level deletions in the LUSC cohort where GISTIC demonstrated better performance with higher recall and lower precision (Figure 3 B), despite *LabelSeg*'s high precision and good recall as illustrated in the Supplementary Figure S3. Furthermore, the performance of all compared methods was better in the GBM dataset compared to the LUSC data, potentially due to differences in tumor sample purity, noise level, and segmentation quality between the two cohorts. Notably, *LabelSeg* exhibited advantages in handling more challenging data owing to its robustness.

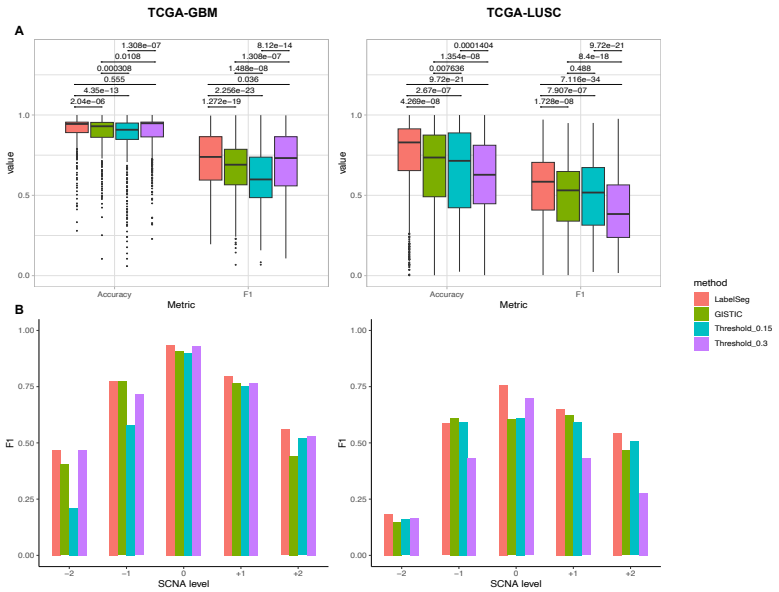

Figure 3: Performance in TCGA datasets. (A) Balanced F1 score and accuracy for each sample. (B) The average F1 score across samples in calling different levels of SCNA.

4.3 Integrative analysis of heterogeneous SCNA profiles

By design, *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). Therefore, we collected SCNA segment profiles from different data resources and performed an integrative analysis. There are four data resources from which the data was derived: TCGA, Progenetix [27], Clinical Proteomic Tumor Analysis Consortium

(CPTAC) [28], and Cancer Cell Line Encyclopedia (CCLE) [29]. The details of the analyzed datasets are described in Table 1 and the Supplementary Data.

In glioblastoma samples, the SCNA patterns from different datasets were relatively consistent (Figure 4). Low-level SCNA frequency was calculated from segments with labels "+1" and "-1". Chromosome 7 duplication and chromosome 10 deletion were the most prominent low-level SCNA features in glioblastoma samples with occurrence greater than 50% in all datasets and reaching 75% in the TCGA and CPTAC datasets. Chromosomes 9p, 13, 14 deletions and chromosomes 19, 20 duplications occurred in more than 25% of samples in most datasets. The UMAP plots based on the called low-level SCNA coverage of individual samples (Supplementary Figure S4) reveal the association of SCNAs across the genome. The co-occurrence of chromosome 7 duplication and chromosome 10 deletion was frequent in the analyzed samples. Duplications of chromosomes 19 and 20, and deletions of chromosome 9p were more likely to occur in samples with simultaneous chromosome 7 duplication and chromosome 10 deletion (P-value $<$ 3e-12 for chr19, P-value $<$ 9e-06 for chr20, P-value $<$ 0.003 for chr9p, Pearson's Chi-squared test). The SCNA pattern observed in the CCLE dataset was characterized by greater heterogeneity and noise compared to other datasets. This could possibly be explained by the difference between tumor samples and cell line models [30].

High-level SCNA frequency was calculated from segments with labels "+2" and "-2". Multiple consensus high-level SCNA focal peaks were observed across various analyzed projects, indicating the recurrent SCNAs' robustness and reliability. Similar to low-level SCNA pattern, the CCLE samples were also more heterogeneous in high-level SCNAs. Supplementary Tables S1-S2 provide further information regarding congruent high-level duplication peaks from at least two projects and high-level deletion peaks from at least three projects. The most frequent amplification cross datasets happened in chr7: 54.1-56.1 MB with a frequency of around 46% in TCGA and CPTAC, 21% in Progenetix, and 12% in CCLE. The most frequent high-level deletion (probably homozygous deletion) occurred in chr9: 21-23 MB with a frequency around 62% in TCGA and CPTAC, 28% in Progenetix, and 70% in CCLE. The reduced high-level frequency detected in the Progenetix samples could potentially be attributed to the heterogeneity of the microarrays employed, which vary in their detection sensitivity of small CNAs, in conjunction with the diversity of sample types analyzed (Supplementary Figure S5). Amplification peaks were identified at chr6: 31.8-32.8 MB and chr7: 92.1-93.1 MB exclusively in the CCLE and Progenetix cohorts. Since all analyzed CCLE samples are cell lines, we hypothesized that amplification in these peaks is more likely to occur in glioblastoma cell lines. To test this hypothesis, we examined the composition of the Progenetix samples that exhibited such amplification. We found that a considerable proportion of the samples (37.6%) with the interesting amplification were cell lines, which is significantly higher than the overall proportion of cell lines in the Progenetix samples (6.3%). This observation was confirmed by a Pearson's Chi-squared test with a P-value less than 2.2e-16, indicating a potential association between cell line samples and amplification in those loci.

We also analyzed the lung squamous cell carcinoma datasets from these data resources (Supplementary Figure S6). Low-level and high-level SCNA patterns in lung squamous cell carcinoma samples were generally accordant across projects. Duplications of chromosomes 1q, 2p, 3q, 5p, 7, 8q and deletions of chromosomes 3p, 5q, 8p were characteristic low-level SCNA patterns in these samples with frequencies between 25% and 50%. Apart from focal peaks, high-level SCNAs spanned chromosomes 3q and 5p with amplification.

Table 1: Summary of the glioblastoma segment datasets

Dataset	Sample	Sample type	Platform
TCGA	566	patient tumor samples	Affymetrix SNP 6.0
Progenetix	1390	patient tumor samples and cancer cell lines	CGH array and SNP array
CPTAC	97	patient tumor samples	WGS
CCLE	64	cancer cell lines	WES and WGS

4.4 Relationship between copy-number dosage and mRNA expression

To illustrate the practical utility of *LabelSeg* and the utility of CNA levels, we examined the correlation between copy-number dosage and mRNA expression of protein-coding genes with recurrent high-level CNAs. Specifically, we analyzed paired mRNA expression data and CNA profiles of glioblastoma samples from the TCGA-GBM project. We only considered protein-coding genes that were amplified or highly deleted with a frequency of $> 5\%$ in TCGA-GBM samples, and that were located in the consensus focal regions mentioned in section 4.3 and Supplementary Tables S1-S2.

Our results showed that 62 out of 68 frequently amplified genes had significantly increased mRNA expression when the SCNA level was "+2" compared to those without SCNAs (BH-adjusted P-value < 0.05, Kruskal-Wallis test). Similarly, 21 out of 24 frequently high-level deleted genes had significantly decreased mRNA expression when the SCNA level was "-2" compared to those without SCNAs (BH-adjusted P-value < 0.05, Kruskal-Wallis test). The most frequently amplified and homozygous deleted genes in this cohort were EGFR and CDKN2A, respectively, and we observed a strong association between SCNA levels and mRNA expression for both genes (Figure 5 A). Supplementary Figures S7-S8 provide box plots illustrating the correlation between SCNA and mRNA expression for other genes.

Figure 5 B further shows the high correlation between SCNA levels and mRNA expression of the frequently altered genes in copy number dosage. These genes were clustered based on their genomic location in SCNA levels, which is not surprising since SCNA is a large-scale genomic variation affecting multiple

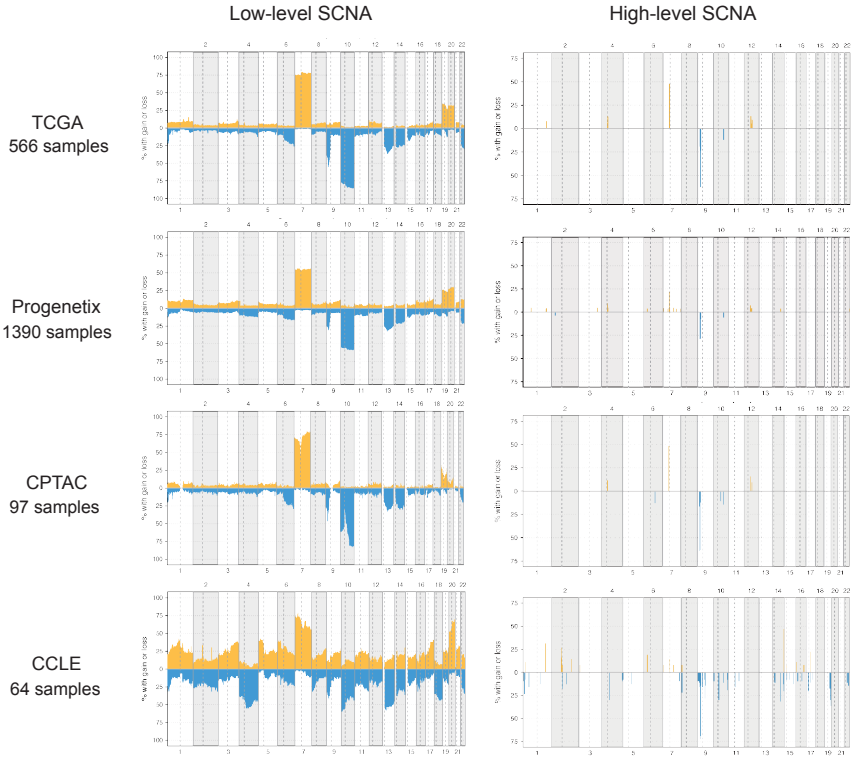

Figure 4: Frequency of SCNA called in different datasets of glioblastoma samples. Yellow represents duplication. Blue represents deletion. The Y-axis is the percentage of samples with SCNA overlapping with the genomic bin of 1MB size. The numbers on the X-axis represent chromosomes. Low-level SCNAs are 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.

genes. Genes in cytobands chr7p11, Chr9p21, chr12q13, and chr12q14 were strongly regulated in mRNA expression by the relative copy number states. Interestingly, VSTM2A and ARHGAP9 were not influenced by SCNA levels, although mRNA expression of nearby genes was strongly impacted by SCNA. Enrichment analysis was performed separately for these frequently high-level duplicated and deleted genes (Supplementary Figures S9-S10). Both sets of genes were enriched in the glioma signaling pathway, indicating the important role of high-level SCNA in cancer development.

A similar analysis was conducted on the TCGA-LUSC cohort. Of the 1383

frequently amplified genes, 938 had an associated mRNA expression with their SCNA levels, while 20 of the 25 frequently high-level deleted genes had an associated mRNA expression with their SCNA levels (BH-adjusted P-value $<$ 0.05, Kruskal-Wallis test). Notably, the high-level deleted genes with mRNA expression responsible for copy number changes were also located in cytoband chr9p21, which contains CDKN2A and CDKN2B (Supplementary Figures S12, S14). Several frequently amplified genes that did not show mRNA association were enriched in epidermal keratinocyte differentiation (Supplementary Figures S11, S13).

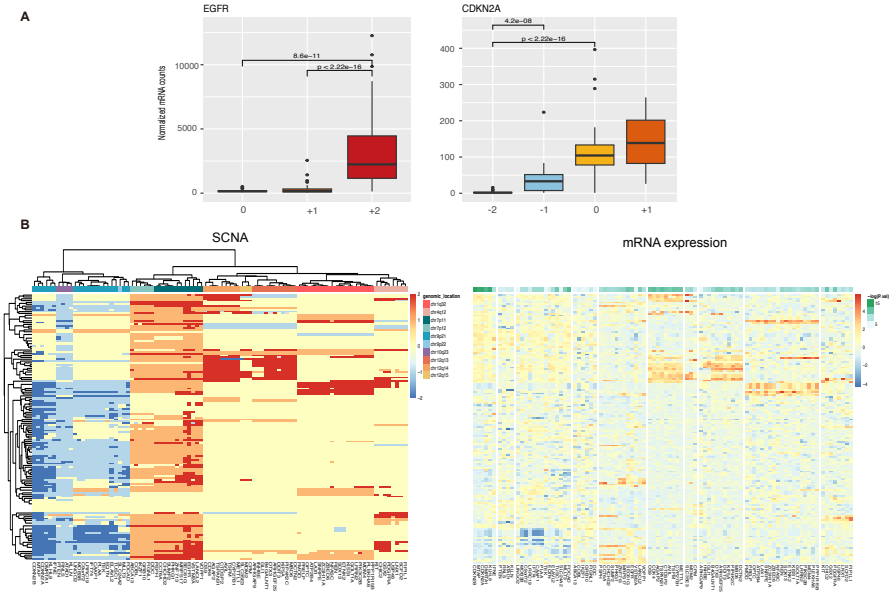

Figure 5: mRNA expression of genes with frequent amplification and homozygous deletion in glioblastoma (A) mRNA expression of characteristic genes in different SCNA levels. (B) Heatmap of SCNA level and mRNA expression in TCGA glioblastoma samples. Rows represent samples and columns represent genes. 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.

5 Discussion

Comprehensive profiling of somatic copy number alterations in cancer samples is valuable for advancing our understanding of cancer development and improving precision medicine applications. Here, we provide a novel method *LabelSeg* for fast and accurate annotation of CNA segments. Our method estimates thresholds for calling different CNA levels from individual segment profiles, allowing more complete and robust identification of SCNAs in heterogeneous samples. The use of separate clustering by segment length not only adapts to the biological and technical variance in both focal and broad segments but also increases the tolerance to sub-clone effects and noise. Compared to the use of fixed cut-off values, *LabelSeg* achieves a similar calling speed but increased accuracy. Furthermore, it does not require additional estimation such as tumor sample purity or prior knowledge, making it a more convenient and powerful tool for large-scale comparative and integrative analysis to overcome bias from individual studies or platforms.

Our study demonstrated that *LabelSeg* outperformed previous methods in SCNA profiling, as evidenced by its higher accuracy and F1 score. The robustness of *LabelSeg* was demonstrated by the consistent patterns of SCNAs detected across heterogeneous datasets, making it a suitable tool for integrative analyses. Our analysis further confirmed simultaneous chromosome 7 duplication and chromosome 10 deletion in glioblastoma samples, which has been previously reported in other studies, thus highlighting the detection accuracy of genome-wide low-level CNAs. Furthermore, we identified several consensus high-level SCNA focal peaks enriched in protein-coding genes, which were observed in at least two of the four datasets. For most of these genes, mRNA expression strongly correlated with the called SCNA status, providing insights into the impact of SCNA of driver genes on tumor evolution.

Because *LabelSeg* solely requires the logR values of segments as input, it is compatible with any technology that delivers segment data from its processing pipeline, regardless of the original data type (i.e. count or intensity-based). However, this general compatibility to a wide range of e.g. microarray and NGS platforms arrives with certain limitations of the method, particularly in estimating absolute copy number and ploidy which require some information about allelic composition. Although this is not a problem when relative copy number states are targeted, it renders the method unsuitable for some investigations that require accurate absolute quantification of copy numbers, such as assessing the degree of aneuploidy in tumor cells. Also, while sensitivity to probe-level noise affects all calling methods to various degrees, in *LabelSeg* such noise can diminish the clustering in the logR values, which can hinder the ability to set accurate calling levels.

To conclude, our study presents a new strategy for segment classification and annotation, which enhances the interpretation of heterogeneous segment profiles with respect to calling efficiency, accuracy, and granularity. The categorization of SCNAs based on distinct levels highlights their varying sizes, amplitudes, and potential functional roles in cancer pathogenesis. As CNA information becomes

increasingly used in cancer genomics applications, including clinical diagnostics and tumor classification [32–34], such dedicated profiling is helpful in investigating the relationship between SCNA and tumors diagnoses, oncogenomic subtypes, as well as for the correlation between SCNA and clinical outcomes.

Acknowledgements

We thank all members of the Baudis group for contributions to Progenetix resource. Some of the results published here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

(1) Negrini, S.; Gorgoulis, V. G.; Halazonetis, T. D. *Nature reviews. Molecular cell biology* **2010**, *11*, 220–228.

(2) Mustjoki, S.; Young, N. S. *New England Journal of Medicine* **2021**, *384*, 2039–2052.

(3) Zarrei, M.; MacDonald, J. R.; Merico, D.; Scherer, S. W. *Nature Reviews Genetics* **2015**, *16*, 172–183.

(4) Talevich, E.; Shain, A. H.; Botton, T.; Bastian, B. C. *PLOS Computational Biology* **2016**, *12*, e1004873.

(5) Koboldt, D. C.; Zhang, Q.; Larson, D. E.; Shen, D.; McLellan, M. D.; Lin, L.; Miller, C. A.; Mardis, E. R.; Ding, L.; Wilson, R. K. *Genome Research* **2012**, *22*, 568.

(6) Zack, T. I. et al. *Nature Genetics 2013 45:10* **2013**, *45*, 1134–1140.

(7) Davoli, T.; Uno, H.; Wooten, E. C.; Elledge, S. J. *Science* **2017**, *355*, DOI: `10.1126/SCIENCE.AAF8399/SUPPL_FILE/AAF8399-DAVOLI-SM.PDF`.

(8) Carter, S. L. et al. *Nature Biotechnology* **2012**, *30*, 413–421.

(9) Van De Wiel, M. A.; Kim, K. I.; Vosse, S. J.; Van Wieringen, W. N.; Wilting, S. M.; Ylstra, B. *Bioinformatics* **2007**, *23*, 892–894.

(10) Boeva, V.; Popova, T.; Bleakley, K.; Chiche, P.; Cappo, J.; Schleiermacher, G.; Janoueix-Lerosey, I.; Delattre, O.; Barillot, E. *Bioinformatics* **2012**, *28*, 423–425.

(11) Wang, K.; Li, M.; Hadley, D.; Liu, R.; Glessner, J.; Grant, S. F.; Hakonarson, H.; Bucan, M. *Genome Research* **2007**, *17*, 1665.

(12) Ha, G. et al. *Genome Research* **2014**, *24*, 1881.

(13) Beroukhim, R.; Mermel, C. H.; Porter, D.; Wei, G.; Raychaudhuri, S.; Donovan, J.; Barretina, J.; Boehm, J. S.; Dobson, J.; Urashima, M., et al. *Nature* **2010**, *463*, 899–905.

(14) Mermel, C. H.; Schumacher, S. E.; Hill, B.; Meyerson, M. L.; Beroukhim, R.; Getz, G. *Genome biology* **2011**, *12*, 1–14.

(15) Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M. *Nature genetics* **2013**, *45*, 1113–1120.

(16) Tarabichi, M.; Salcedo, A.; Deshwar, A. G.; Ni Leathlobhair, M.; Wintersinger, J.; Wedge, D. C.; Van Loo, P.; Morris, Q. D.; Boutros, P. C. *Nature methods* **2021**, *18*, 144–155.

(17) Valsesia, A.; Rimoldi, D.; Martinet, D.; Ibberson, M.; Benaglio, P.; Quadroni, M.; Waridel, P.; Gaillard, M.; Pidoux, M.; Rapin, B., et al. *PLoS One* **2011**, *6*, e18369.

(18) Myllykangas, S.; Böhling, T.; Knuutila, S. In *Seminars in cancer biology*, 2007; Vol. 17, pp 42–55.

(19) Zhang, Y.; Chen, F.; Fonseca, N. A.; He, Y.; Fujita, M.; Nakagawa, H.; Zhang, Z.; Brazma, A., et al. *Nature communications* **2020**, *11*, 736.

(20) Waddell, N.; Pajic, M.; Patch, A.-M.; Chang, D. K.; Kassahn, K. S.; Bailey, P.; Johns, A. L.; Miller, D.; Nones, K.; Quek, K., et al. *Nature* **2015**, *518*, 495–501.

(21) Hogarty, M.; Brodeur, G. *The genetic basis of human cancer* **2001**, *2*, 115–28.

(22) Krijgsman, O.; Carvalho, B.; Meijer, G. A.; Steenbergen, R. D.; Ylstra, B. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **2014**, *1843*, 2698–2704.

(23) Jamal-Hanjani, M.; Wilson, G. A.; McGranahan, N.; Birkbak, N. J.; Watkins, T. B.; Veeriah, S.; Shafi, S.; Johnson, D. H.; Mitter, R.; Rosenthal, R., et al. *New England Journal of Medicine* **2017**, *376*, 2109–2121.

(24) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X., et al. In *kdd*, 1996; Vol. 96, pp 226–231.

(25) Willenbrock, H.; Fridlyand, J. *Bioinformatics* **2005**, *21*, 4084–4091.

(26) Aran, D.; Sirota, M.; Butte, A. J. *Nature communications* **2015**, *6*, 1–12.

(27) Huang, Q.; Carrio-Cordo, P.; Gao, B.; Paloots, R.; Baudis, M. *Database* **2021**, *2021*.

(28) Ellis, M. J.; Gillette, M.; Carr, S. A.; Paulovich, A. G.; Smith, R. D.; Rodland, K. K.; Townsend, R. R.; Kinsinger, C.; Mesri, M.; Rodriguez, H., et al. *Cancer discovery* **2013**, *3*, 1108–1112.

(29) Nusinow, D. P.; Szpyt, J.; Ghandi, M.; Rose, C. M.; McDonald III, E. R.; Kalocsay, M.; Jané-Valbuena, J.; Gelfand, E.; Schweppe, D. K.; Jedrychowski, M., et al. *Cell* **2020**, *180*, 387–402.

(30) Domcke, S.; Sinha, R.; Levine, D. A.; Sander, C.; Schultz, N. *Nature communications* **2013**, *4*, 1–10.

(31) Robinson, M. D.; Oshlack, A. *Genome biology* **2010**, *11*, 1–9.

(32) Louis, D. N.; Perry, A.; Wesseling, P.; Brat, D. J.; Cree, I. A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.; Pfister, S. M.; Reifenberger, G., et al. *Neuro-oncology* **2021**, *23*, 1231–1251.

(33) Zhang, N.; Wang, M.; Zhang, P.; Huang, T. *Biochimica et Biophysica Acta (BBA)-General Subjects* **2016**, *1860*, 2750–2755.

(34) Elsadek, S. F. A.; Makhlouf, M. A. A.; Aldeen, M. A. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018 4*, 2019, pp 198–207.