**Copy-scAT: An R package for detection of large-scale and focal copy number alterations in single-cell chromatin accessibility datasets**

Ana Nikolic[1,2,3], Divya Singhal[1,2,3], Katrina Ellestad[1,2,3], Michael Johnston[1,2,3], Aaron Gillmor[1,2,3], Sorana Morrissy[1,2,3], Jennifer A Chan[2,3,4], Paola Neri[2,3,4], Nizar Bahlis[2,3,4], Marco Gallo[1,2,3]*

[1]Department of Biochemistry and Molecular Biology
[2]Arnie Charbonneau Cancer Institute
[3]Alberta Children's Hospital Research Institute
[4]Department of Oncology
Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

*Correspondence to: Marco Gallo
marco.gallo@ucalgary.ca

**ABSTRACT**

The single-cell assay for transposase accessible chromatin (scATAC) is an invaluable asset to profile the epigenomic landscape of heterogeneous cells populations in complex tissue and organ systems. However, the lack of tools that enable the use of scATAC data to discriminate between malignant and non-malignant cells has prevented the widespread application of this technique to clinical tumor samples. Here we describe Copy-scAT, a new computational tool that uses scATAC data to infer both large-scale and focal copy number alterations. Copy-scAT can call both clonal and subclonal copy number changes, allowing identification of cancer cells and cell populations that putatively constitute the tumor microenvironment. Copy-scAT therefore enables downstream chromatin accessibility studies that focus on malignant or non-malignant cell populations in clinical samples that are profiled by scATAC.

**INTRODUCTION**

Single-cell genomic tools have been invaluable in efforts to deconvolute complex and heterogeneous cellular systems, including cancer[1]. Single-cell RNA sequencing (scRNA-seq) generates data on transcriptional profiles of individual cells and has been robustly and widely implemented[2]. The sequencing-based single-cell assay for transposase-accessible chromatin (scATAC-seq) produces a snapshot of the genomic areas that are in accessible (i.e. "open") chromatin, a conformation normally associated with active regulatory regions and with gene transcription[3]. Whereas numerous computational tools have been developed over the years for scRNA-seq applications[4], downstream analysis platforms for scATAC-seq datasets are still limited, posing significant challenges to more widespread implementation of this epigenomic approach.

scATAC-seq can be used to dissect mechanisms of transcriptional regulation in discrete cell populations. These mechanisms can be inferred by mapping active putative regulatory regions - like enhancers and super enhancers - or footprints associated with occupancy of transcription factors (TFs) and TF families that are active in specific cell types. When applied to heterogeneous tissues, these techniques can deconvolute the transcriptional and chromatin states that are responsible for the emergence and maintenance of specific cell fates. When deployed in the context of heterogeneous cancer types, they can provide information on the epigenomic states that define populations with disparate functional properties, including putative cancer stem cells and more differentiated cell types with limited self-renewal.

One major challenge in the application of scATAC-seq to investigations of cancer is sample heterogeneity and inability to reliably distinguish between neoplastic and non-neoplastic cells. As most neoplastic cells have some degree of chromosomal instability, they typically will have at least some large-scale copy number variants (CNVs) or chromosomal gains or losses, which are typically absent from normal cells. Analysis of CNVs in single-cell data has been approached by multiple groups in a number of different sample types. For full-length single-cell RNA-seq data, the Suvà group used binning of transcript values over different regions of chromosomes to impute CNVs[5,6]. CONICSmat is an R package designed to perform similar analyses, with the option

to combine allele haplotypes and windowed gene transcription from scRNA-seq expression matrices[7]. CONICSmat was also used more recently to infer CNVs from ATAC-seq data, by using imputed gene activity scores generated with Snap-ATAC[8] in lieu of an expression matrix[9].

Algorithms have been developed to impute CNVs from single-cell DNA-sequencing data. Early approaches included read depth assessment with variable bins[10], and a method by Ning et al leveraging GC bias correction[11]. More recent methods include SCAN-SNV, which leverages allelic imbalance to call CNVs in scDNA-seq data[12], and CHISEL, which again leverages single-cell haplotypes to call CNVs in single cells and cell subpopulations[13]. However, to date, no dedicated method has been developed to call CNVs using scATAC-seq data, limiting its potential applications to the study of complex tumor types. scATAC-seq datasets are challenging as they are quite sparse, preferentially sample accessible chromatin only, and have a certain degree of bias due to the sequence preferences of the Tn5 transposase. We set out to develop a tool that enables reliable calling of CNVs using scATAC-seq datasets. To this end, we describe Copy-scAT (*copy* number inference using *s*ingle-*c*ell *AT*AC), an R package that uses a combination of Gaussian segmentation and changepoint analysis[14] to identify large-scale gains and losses and regions of focal loss and amplification in individual cells. We provide proof-of-principle validation of the functionalities of this tool using scATAC-seq datasets we generated from adult glioblastoma (aGBM; n = 3), pediatric GBM (pGBM; n = 6), and multiple myeloma (MM; n = 10) clinical specimens.

## RESULTS

### *Copy-scAT identifies chromosome-arm level copy number variation and focal amplifications in scATAC-seq data using read depth analysis*

We designed Copy-scAT, an R package that uses scATAC-seq information to infer copy number alterations. Copy-scAT uses fragment files generated by cellranger-atac (10xGenomics) as input to generate chromatin accessibility pileups, keeping only barcodes with a minimum number of fragments (defaulting to 5,000 fragments). It then generates a pileup of total coverage (number of reads × read lengths) over bins of determined length (1 million bp as default) **(Figure 1A)**. Reads are then normalized along linear scales (i.e. they are not log-normalized) over the total signal in each cell to account for differences in read depth, and chromosomal bins which consist predominantly of zeros (at least 80% zero values) are discarded from further analysis. All parameters, including reference genome, bin size, and minimum length cut-off are user-customizable. Copy-scAT then implements different algorithms to detect focal amplifications and larger-scale copy number variation.

To call focal amplifications **(Figure 1B)**, Copy-scAT generates a linear scaled profile of density over the normalized 1 Mbp bins along each chromosome on a single-cell basis, centering on the median and scaling using the range. We use changepoint analysis (see Methods) to identify segments of abnormally high signal (Z-score > 5) along each chromosome in each single cell. These calls are then pooled together to generate consensus regions of amplification, in order to identify putative double minutes and extrachromosomal amplifications. Each cell is scored as present/absent for each region.

Segmental losses are called in a similar fashion, by calculating a quantile for each bin on a chromosome, running changepoint analysis to identify regions with abnormally low average signal, and then using Gaussian decomposition of total signal in that region to identify distinct clusters of cells.

For larger copy number alterations, Copy-scAT pools the bins further at the chromosome arm level using a trimmed mean, while normalizing the data on the basis of length of CpG islands contained in each bin **(Figure 1C)**. Data is then scaled for each chromosome arm, compared to a pseudodiploid control, and cluster assignments are generated using Gaussian decomposition. Cluster assignments are then normalized to get an estimate of copy number for each cell. These assignments are then optionally combined with clustering information to generate consensus genotypes for each cluster of cells and further filter false positives. For full details regarding the execution of Copy-scAT, see Methods.

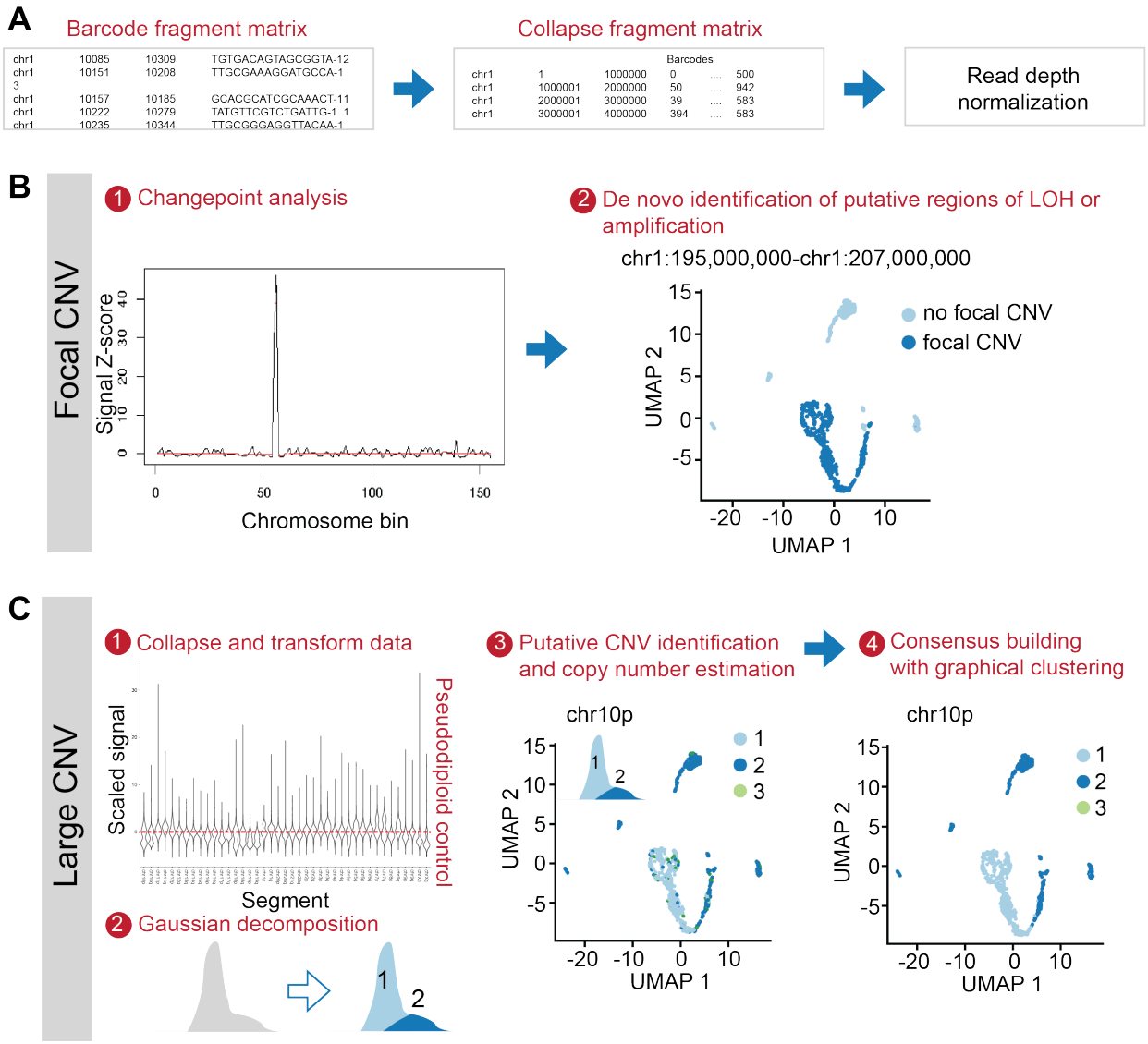


**Figure 1. Overview of the Copy-scAT workflow.**
**(A)** Initial sample processing.
**(B)** Pipeline for detection of focal amplifications and losses.
**(C)** Pipeline for detection of large-scale CNVs.

***Copy-scAT highlights segmental amplifications and losses in solid tumours***

By implementing Copy-scAT, we were able to detect both amplifications and large-scale copy number events using scATAC-seq datasets we generated from surgically resected GBM samples. Some copy number events were subclonal (see for instance chromosome 10p loss in **Figure 2A**). However, the majority of the putative copy number events that were detected appeared in nearly all tumour cells, were not detected in adjoining normal cells (inferred by the lack of called CNVs), and were therefore deemed to be clonal **(Figure 2B-2C)**. As an example, we found that amplifications of the distal region of chromosome 1q involving *MDM4* appeared to be clonal in sample 4349 **(Figure 2D)**.

However, Copy-scAT also called subclonal focal amplifications at the *PDGFR* and *EGFR* loci in this tumor. *PDGFRA* and *EGFR* focal amplifications were mutually exclusive (**Figure 2E-2F)**, a result that is in agreement with previous observations in aGBM[15]. In contrast, only one subclone of sample 4218 showed loss of chromosome 18, along with a distinct gain of chromosome 1p, the latter not being detected in whole-genome sequencing (WGS) data **(Figure 2G-I)**. This type of information would be difficult to obtain through conventional WGS, in part because of underlying sample heterogeneity and the relatively small size of some subclones. These results suggest that new information on clonal architecture can be gleaned through implementation of Copy-scAT compared to more traditional bulk sequencing technologies.

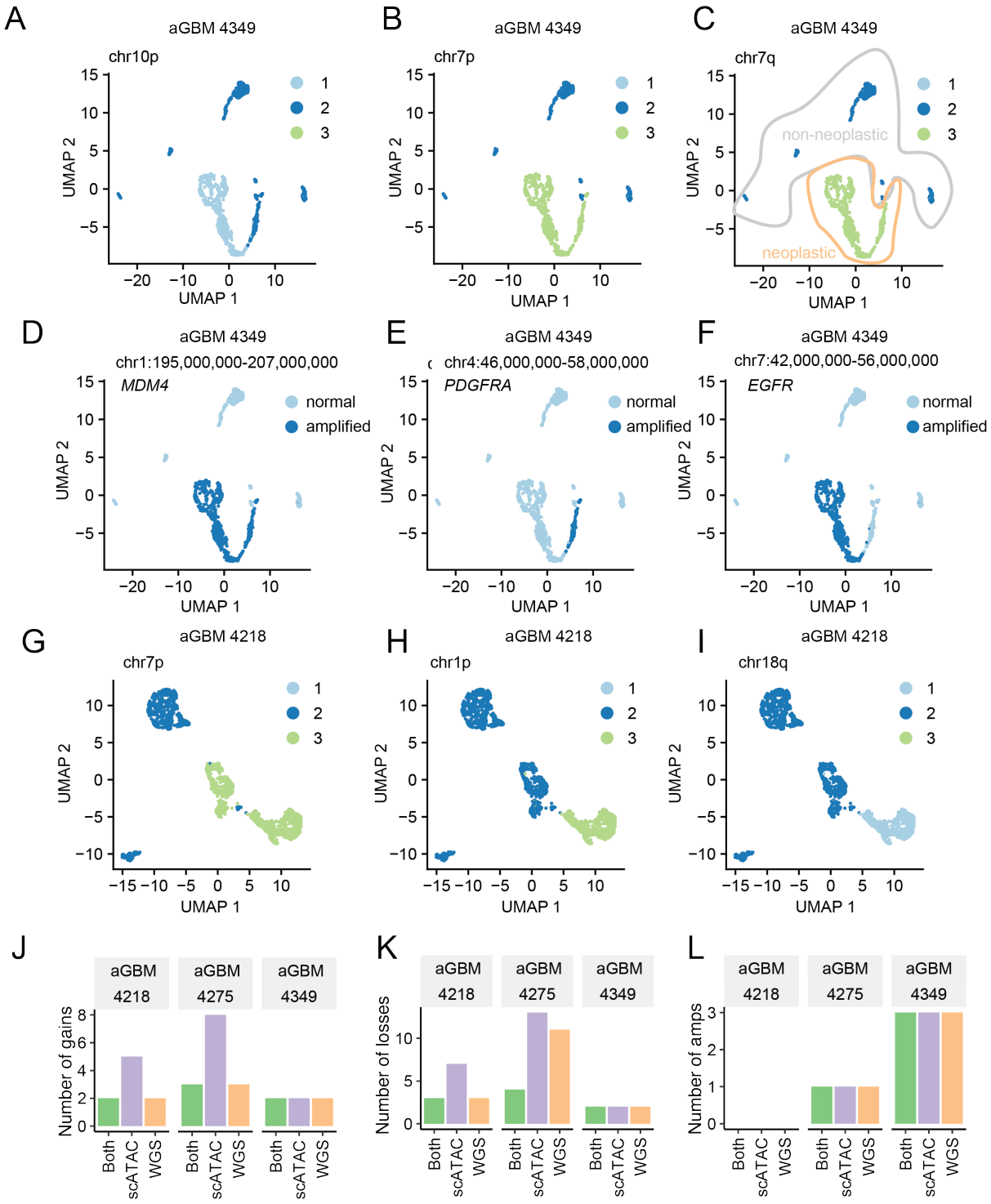


**Figure 2. Identification of amplifications and copy number alterations on single-cell data with Copy-scAT.**

**(A-C)** Identification of GBM-specific copy number variants in neoplastic cells in a primary aGBM sample.

**(D-F)** Clonal amplification at *MDM4*, and subclonal amplifications of *PDGFRA* and *EGFR* detected in a GBM sample.
**(G-I)** Subclonal gains in 1p and loss of 18q in aGBM 4218.
**(J-L)** Performance of Copy-scAT with scATAC-seq datasets generated from aGBM samples compared to WGS for gains, losses and amplification events, respectively.

---

***Copy-scAT effectively calls clonal and subclonal CNVs***

Overall, in 3 aGBM samples and 6 pGBM samples (diagnostic/relapse pairs from 3 patients), Copy-scAT detected the majority of copy number alterations detected by WGS, with a sensitivity of 1.0, 0.79 and 1.0 for gains, losses and focal amplifications in adult samples, and lower sensitivity of 0.73 and 0.73 for gains and losses in pediatric samples **(Figure 2J-K; Figure 3A-B; Table 1).** Specificity of the calls was greater than 0.89 in all cases. Some of the differences in sensitivity may have to do with input sample characteristics, as the pGBM samples had both a lower average read count per filtered cell (average pGBM count: 36,062; average aGBM count: 54,797) and were generated from snap-frozen archival tissue rather than cryopreserved cells, possibly leading to a lower quality of transposition **(Table 1)**. Examples of copy number variants detected in a paired pGBM sample (pGBM-1) showed some variants identified in both matched samples, with others only being detected in one of the two samples **(Figure 3C-F)**.
Our tool was also tested in a cohort of 10 MM samples, which had been profiled with single-cell copy number (scCN) assays (10xGenomics). There was reasonable concordance between gains and losses determined by Copy-scAT and the scCN assay **(Figure 4A, 3B)** in all 10 samples, and similar numbers of gains and losses were detected by both methods **(Figure 4C, 4D)**. Sensitivity in the 10 MM samples was slightly lower than for GBMs, with 0.51 and 0.67 for gains and losses, respectively **(Table 1)**. These samples had a much lower average read counts and more cells per dataset than the GBM samples. Concordance between methods was not associated with overall read counts per cell and number of cells per sample **(Figure 4E-H)**, although there was a trend towards improved detection of CNVs in samples with larger numbers of cells. Concordance was reasonable even in samples with 10,000-20,000 reads per cell. Examples of unfiltered CNV calls from one sample (MM1555) show detection of multiple variants also detected in the scCN experiments. There was clear distinction between one large cluster and two smaller clusters, with a few small clusters lacking the CNVs, consistent with non-neoplastic hematopoietic cells **(Figure 4I-K)**.

**Table 1. Sensitivity and Specificity of Copy-scAT in aGBM, pGBM and MM samples**

| | Gains | | Losses | | Amplifications | |
|---|---|---|---|---|---|---|
| Samples | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity |
| aGBM (n = 3) | 1.0 | 0.94 | 0.79 | 0.89 | 1.0 | 1.0 |
| pGBM (n= 6) | 0.73 | 0.93 | 0.73 | 0.95 | N/A | 0.975 |
| MM (n = 10) | 0.51 | 0.94 | 0.67 | 0.89 | N/A | N/A |

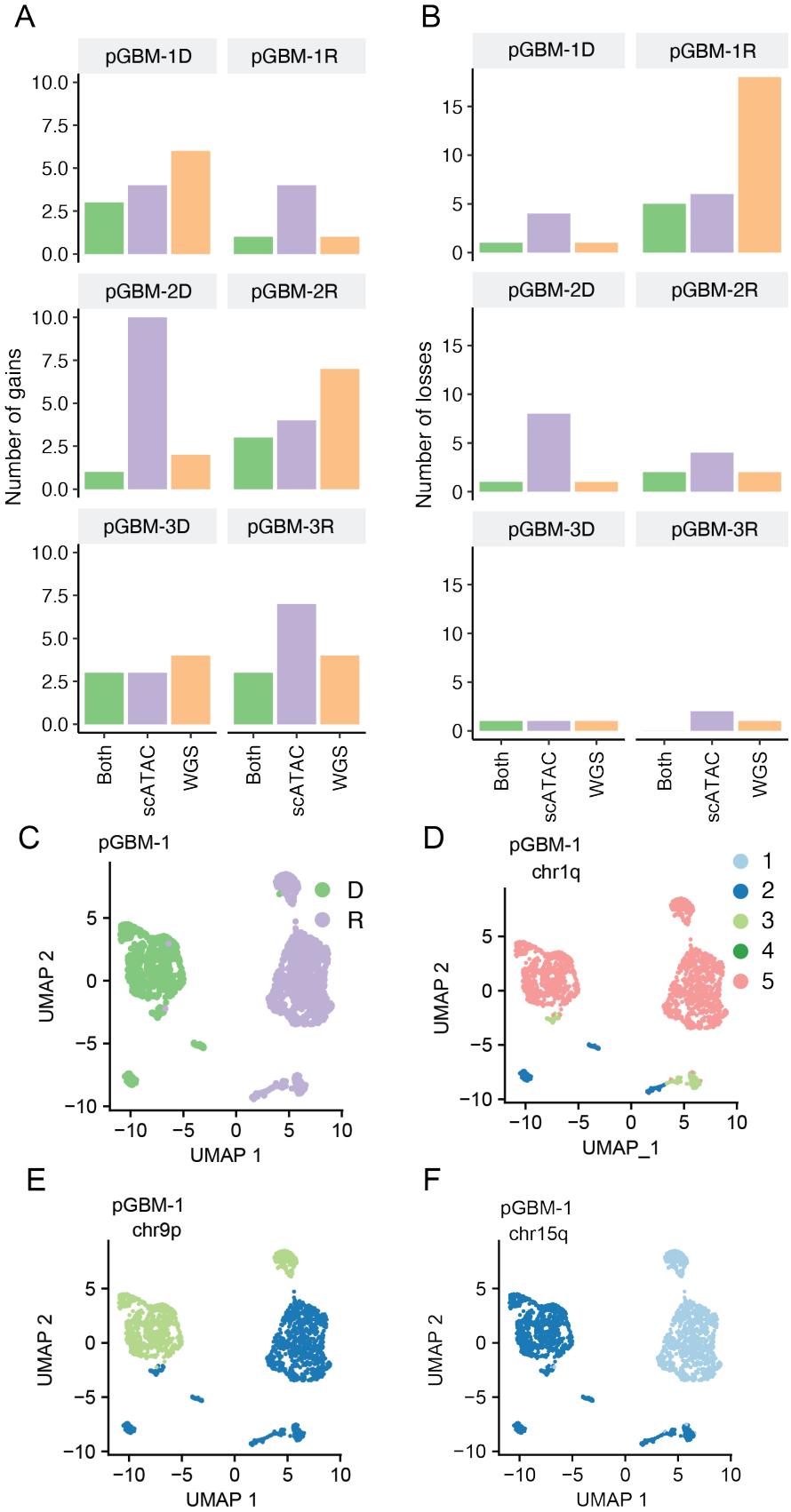
A
pGBM-1D
pGBM-1R
pGBM-2D
pGBM-2R
pGBM-3D
pGBM-3R
Number of gains
Both
scATAC
WGS
B
Number of losses
C
pGBM-1
D
R
UMAP 2
UMAP 1
D
pGBM-1
chr1q
1
2
3
4
5
UMAP_1
E
pGBM-1
chr9p
F
pGBM-1
chr15q

**Figure 3. Performance of Copy-scAT with pGBM samples.**
**(A-B)** Comparison of scATAC versus WGS datasets in calling large-scale copy number gains and losses in patient-matched diagnostic (D) and relapse (R) samples.
**(C)** UMAP plot of scATAC data for patient-matched diagnostic (D) and relapse (R) pGBM specimens.
**(D-F)** Examples of copy number alterations detected in both samples **(D,E)** or in one sample only **(F)**.

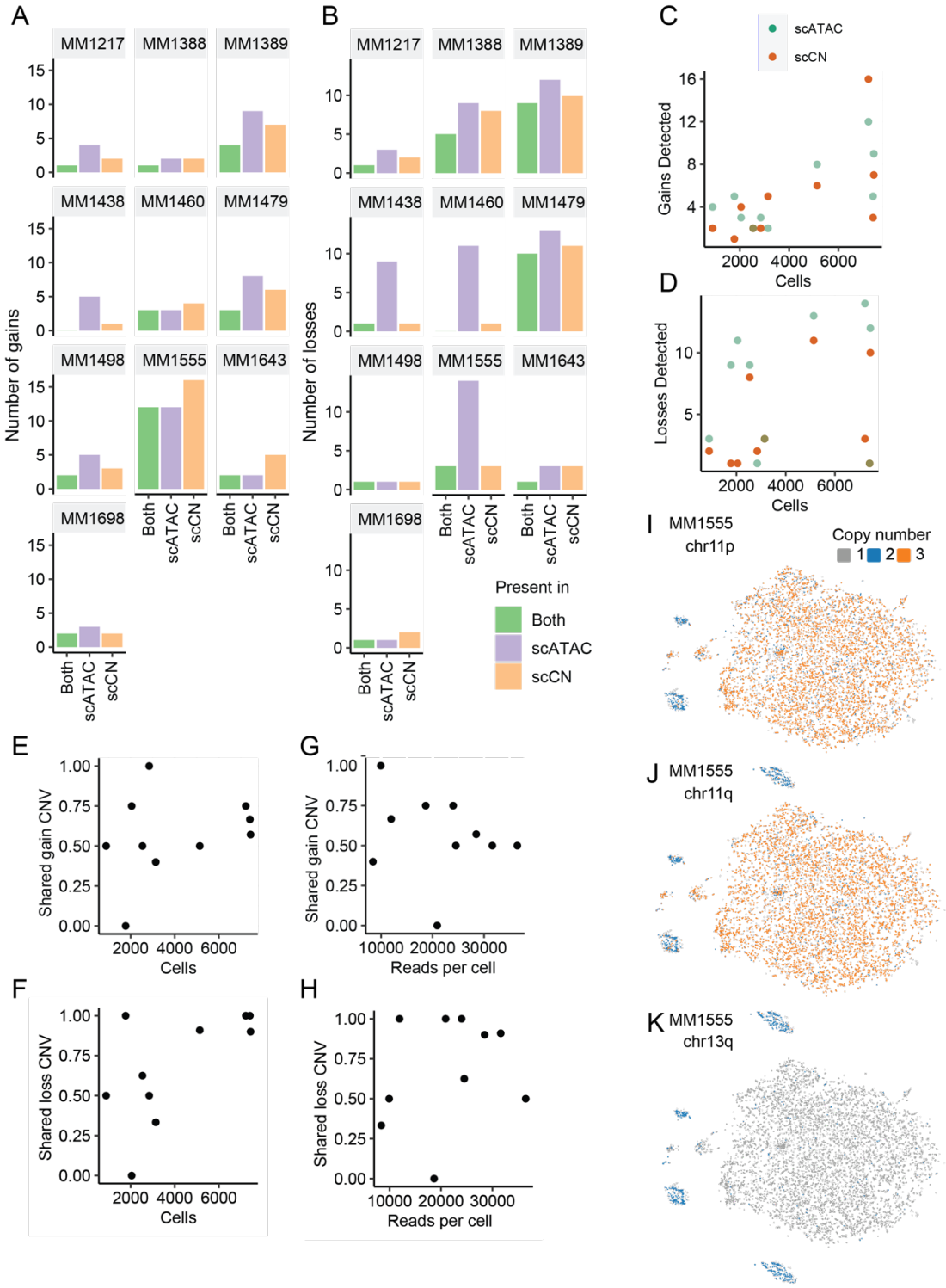


**Figure 4. Performance of Copy-scAT in calling copy number alterations in MM samples compared to the scCN assay.**

**(A-B)** Number of events detected in each specimen by Copy-scAT, scCN or both, for copy number gains and losses.

**(C-D)** Comparison of total number of events detected for copy number gain and loss events by number of cells in sample.
**(E-F)** CNVs detected by both Copy-scAT and scCN analysis versus sample size.
**(G-H)** CNVs detected by both Copy-scAT and scCN analysis versus average reads per cell in scATAC sample.
**(I-K)** CNVs detected in sample MM1555, all of which were also identified by scCN.

---

## DISCUSSION

We have developed Copy-scAT, a novel tool which leverages Gaussian decomposition and changepoint analysis to call focal amplifications and large-scale CNVs using scATAC-seq datasets. This is accomplished without the need for a non-neoplastic control. We benchmarked Copy-scAT using scATAC-seq datasets we generated from clinical samples, but it should be widely applicable to the study of cell lines and other systems characterized by inherent epigenomic and genomic heterogeneity. Our data demonstrate that Copy-scAT can identify clonal and subclonal CNVs in primary cancer samples, enabling the discrimination between malignant and non-malignant cells in clinical specimens.

We were able to use Copy-scAT to effectively identify CNVs in both solid tumors (GBM samples) and liquid malignancies (MM samples) using scATAC data. CNV calls by Copy-scAT were benchmarked against (i) CNV calls made with WGS data for GBM samples and (ii) CNV calls made by scCN technology (10xGenomics) for MM samples. In general, Copy-scAT performed well on samples with at least 20,000 reads per cell and was able to detect some CNVs with samples having as few as 10,000 reads per cell. In samples with high read counts, Copy-scAT identified more putative CNVs than were identified using WGS, some confined to small clusters of cells. Some of these calls represented smaller changes in signal, which may represent either true segmental changes, or may simply reflect significant changes in accessibility (such as compartment status) in particular regions of the genome and may not be associated with an underlying copy number alteration. These observations suggest that higher per-cell sequencing depth (approximately 30,000 reads per cell) enable more reliable detection of alterations such as CNVs and amplifications. Importantly, Copy-scAT can detect both clonal and subclonal copy number events, thereby significantly expanding the downstream applications of scATAC experiments compared to what is currently possible.

There are some limitations to our approach. First, in our GBM samples, as we used a whole-genome reference, we were not able to validate the CNV status of individual cells. In addition, our benchmarking results could be impacted by our inability to perform WGS and scATAC on the same cell populations in pGBM samples. It is therefore possible that some of the discrepancies in calling CNVs between WGS and scATAC-seq/Copy-scAT may reflect true differences in subclonal compositions of the cells that were assayed with these two technologies. This is supported by the better performance of our tool with aGBM samples, where the WGS reference sequencing was performed on residual nuclei from the same tissue fragment that was used for scATAC, which resulted in sampling the same underlying cell populations with both methods.

Moreover, normalization of CpG content in ATAC samples is imperfect and may be affected by the overall enrichment in the sample and the sample type. This may lead to an irregular baseline signal for different chromosomes and may lead to inaccurate assignments of some chromosomes as lost or gained. The optimal settings for our algorithm are also variable, especially with regard to the pseudodiploid control cells, and may need to be adjusted for different sample types and sample cellularity. In general, the algorithm works best with samples that are heterogenous, with at least 5-10% non-tumour cells. In addition, as we do not leverage haplotype data, our approach is unable to detect copy-neutral loss of heterozygosity. And lastly, while CNVs are typically associated with neoplastic cells, there are tissue types, such as the developing and normal brain, where mosaicism is common, and thus populations of cells with CNVs may be seen that are not cancer cells[16–18].

Overall, Copy-scAT is a tool that will enable new types of epigenomic investigations of complex tissues and model systems. It will expand the downstream applications of scATAC, especially in the context of cancer studies, by allowing the identification of malignant and non-malignant cells in clinical samples.

## METHODS

### *Ethics and consent statement*

All samples were collected and used for research with appropriate informed consent and with approval by the Health Research Ethics Board of Alberta.

### *scATAC-seq sample processing*

scATAC libraries were prepared from GBM and MM samples using a Chromium controller (10xGenomics), as per protocol. Samples were sequenced on NextSeq 500 or Novaseq 6000 instruments (Illumina) at the Centre for Health Genomics and Informatics (CHGI; University of Calgary) using the recommended settings.

### *scATAC-seq initial data analysis*

The raw sequencing data was demultiplexed using cellranger-atac mkfastq (Cell Ranger ATAC, version 1.1.0, 10x Genomics). Single cell ATAC-seq reads were aligned to the hg38 reference genome (GRCh38, version 1.1.0, 10x Genomics) and quantified using cellranger-atac count function with default parameters (Cell Ranger ATAC, version 1.1.0, 10x Genomics).

### *Single-cell CNV analysis*

*Fragment pileup and normalization*

The fragment file was processed and signal was binned into bins of a preset size (default 1 Mb) across the hg38 chromosomes to generate a genome-wide read-depth map. Only barcodes with a minimum of 5000 reads were retained, in order to remove spurious barcodes. This flattened barcode-fragment matrix pileup was cleaned by removal of genomic intervals which were uninformative (greater than 80% zeros) and barcodes with greater than a certain number of zero

intervals. Cells passing this first filter were normalized with counts-per-million normalization using *cpm* in the *edgeR* package[19].

*Chromosome arm CNV analysis*

The normalized barcode-fragment matrix was collapsed to the chromosome arm level, using chromosome arm information from the UCSC (UCSC table: cytoBand), centromeres were removed, and signal in each bin was normalized using the number of basepairs in CpG islands in the interval using the UCSC CpG islands table (UCSC table: cpgIslandExtUnmasked). The signal was then summarized using a quantile-trimmed-mean (between the $50^{th}$ and $80^{th}$ quantiles). Only chromosome arms with a minimum trimmed mean signal were kept for analysis.

The chromosome arm signal matrix is mixed with a generated set proportion of pseudodiploid control cells, defined using the mean of chromosome segment medians with a defined standard deviation. This cell-signal matrix is then scaled across each chromosome arm and centered on the median signal of all chromosomes. Each chromosome arm segment is then analyzed using Gaussian decomposition with Mclust[20]. The subsequent clusters are filtered based on Z scores and mixing proportions, and redundant clusters are combined. These Z scores are then translated into estimated copy numbers for each segment for each barcode. The barcode CNV assignments can be optionally used to assign consensus CNVs to clusters generated in other software packages such as Loupe or Seurat/Signac.

*De novo amplification detection*

The normalized barcode-fragment matrix was scaled and mean-variance changepoint analysis using the Changepoint package was performed for each cell and each chromosome to identify areas of abnormally high signal (Z score greater than 5)[14]. The consensus coordinates of each amplification region were generated across all cells and only abnormalities affecting a minimum number of cells were kept for analysis.

*De novo loss of heterozygosity detection*

The normalized barcode-fragment matrix was scaled as above. As overall coverage levels in these samples are quite sparse, a chromosome-wide coverage profile was generated for the entire sample in bulk, using the 30% quantile as a cut-off, and then changepoint analysis was used to find inflection points. This was followed by Gaussian decomposition of the values using Mclust to identify putative areas of loss or gain, thresholded by a minimum difference in signal between the clusters identified by Mclust.

***Whole genome sequencing***

DNA was extracted from residual nuclei from the same samples and tissue fragments used for scATAC-seq of adult GBM samples, using the Qiagen DNEasy Blood and Tissue DNA extraction kit (Qiagen # 69504). Libraries were prepared using the NEBNext Ultra II DNA Library Prep Kit (#E7645) and sequenced on the Novaseq 6000 (Illumina) at the CHGI (University of Calgary), in paired-end mode.

***Whole genome data processing***

Genome data was aligned to the hg38 assembly using bwa mem (bwa 0.7.17)[21]. Samtools was used to extract high-quality reads (Q > 30) and picard tools (Broad Institute) was used to remove duplicates[22].

***Whole genome SNV and CNV detection***

Gatk mutect2 (Broad Institute) was run on the filtered data to detect SNVs with low stringency using the following settings: *--disable-read-filter MateOnSameContigOrNoMappedMateReadFilter*. CNVkit was subsequently used to call copy number variants using the following parameters: *--filter cn -m clonal –purity 0.7*[23]. Adjacent segments were further combined and averaged using bedtools[24].

***Data visualization and clustering***

Data was visualized and UMAP plots were generated using Seurat 3.0.0 and Signac 1.0.0 (Github: https://github.com/timoast/signac)) and Cell Loupe version 4.0.0[25].

**Data and code availability:**

The package and a sample tutorial for Copy-scAT is available on Github at http://github.com/spcdot/CopyscAT . Processed single-cell ATAC-seq data files will be made available upon publication.

**REFERENCES**


1. Lim, B., Lin, Y. & Navin, N. Advancing Cancer Research and Medicine with Single-Cell Genomics. *Cancer Cell* **37**, 456–470 (2020).
2. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
3. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
4. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14**, (2018).
5. Tirosh, I. *et al.* Large-scale single-cell RNA-seq reveals a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
6. Venteicher, A. S. *et al.* Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science (80-. ).* **355**, (2017).
7. Müller, S., Cho, A., Liu, S. J., Lim, D. A. & Diaz, A. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor sub-clones. *Bioinformatics* **34**, 3217–3219 (2018).
8. Fang, R. *et al.* Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv* 615179 (2019). doi:10.1101/615179
9. Wang, L. *et al.* The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. *Cancer Discov.* **9**, 1708–1719 (2019).
10. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* 90–94

(2011). doi:10.1038/nature09807
11. Ning, L. *et al.* Quantitative assessment of single-cell whole genome amplification methods for detecting copy number variation using hippocampal neurons. *Sci. Rep.* **5**, (2015).
12. Luquette, L. J., Bohrson, C. L., Sherman, M. A. & Park, P. J. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. *Nat. Commun.* **10**, 3908 (2019).
13. Zaccaria, S. & Raphael, B. J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat. Biotechnol.* (2020). doi:10.1038/s41587-020-0661-6
14. Killick, R. & Eckley, I. A. Changepoint: An R package for changepoint analysis. *J. Stat. Softw.* **58**, (2014).
15. Snuderl, M. *et al.* Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma. *Cancer Cell* **20**, 810–7 (2011).
16. Rehen, S. K. *et al.* Constitutional aneuploidy in the normal human brain. *J. Neurosci.* **25**, 2176–2180 (2005).
17. Rohrback, S., Siddoway, B., Liu, C. S. & Chun, J. Genomic mosaicism in the developing and adult brain. *Dev. Neurobiol.* **78**, 1026–1048 (2018).
18. McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science (80-. ).* (2013). doi:10.1126/science.1243472
19. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
20. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016).
21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
22. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
23. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, (2016).
24. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–2 (2010).
25. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).