

RESEARCH

Open Access



Resolving clonal evolution and selection of extrachromosomal DNA at single-cell resolution

Josephine Deleuran Hendriksen^{1,2,3}, Alessio Locallo^{1,2,3}, Balthasar Clemens Schlotmann^{1,2}, Francisco Germán Rodríguez González^{1,2}, Jane Skjøth-Rasmussen^{3,4,5}, Christina Westmose Yde⁶, Dorte Schou Nørøxe^{3,7}, Hans Skovgaard Poulsen^{3,8}, Ulrik Lassen^{3,7} and Joachim Weischenfeldt^{1,2,3*}

*Correspondence:
joachim.weischenfeldt@bric.ku.dk

¹ The Finsen Laboratory, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

² Biotech Research and Innovation Centre (BRIC), Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

³ The DCCC Brain Tumor Center, Danish Comprehensive Cancer Center, Copenhagen, Denmark

⁴ Department of Neurosurgery, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

⁵ Department of Clinical Medicine, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

⁶ Department of Genomic Medicine, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

⁷ Department of Oncology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

⁸ Department of Radiation Biology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

Abstract

Background: Genomic alterations are a hallmark of cancer, and extrachromosomal DNA (ecDNA) has emerged as a key source of oncogene selection, tumor growth, and drug resistance. The intratumor heterogeneity and clonal selection of ecDNA is, however, poorly understood.

Results: In this study, we pursue a computational approach that leverages allelic imbalance and outlier expression from standard single-cell RNA sequencing (scRNA-seq) to deconvolve the tumor heterogeneity of ecDNA at the single-cell level (ecSingle). Using this approach, we identify oncogene-carrying ecDNAs in tumor samples at the single-cell level, which we validate using genome sequencing. Moreover, we show the superiority of using single-molecule long-read sequencing in resolving ecDNA. ecDNAs displayed extensive intratumor heterogeneity, including subclonal oncogene-carrying ecDNA in primary tumor cells that segregate with distinct transcriptional cell states. Importantly, we show that a rare ecDNA+ clone in the primary tumor can expand to form dominant clones in relapse tumors.

Conclusions: Our study introduces a novel approach to studying ecDNA at the single-cell level, enabling both clonal evolution and transcription cell state analysis. We apply this approach to cancer samples to gain deeper insights into the role of ecDNA in intratumor heterogeneity and cellular plasticity.

Keywords: Extrachromosomal DNA, Cancer, Single cell, Clonal evolution, Long-read sequencing, Oncogene amplification

Background

Focal amplifications in cancer are copy number increases of a small genomic region, most often containing oncogenes, whose overexpression can drive tumor growth. One intriguing type of focal amplification is extrachromosomal DNA (ecDNA), which forms circular DNA structures outside the regular chromosomes. EcDNA amplicons can be



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

unevenly segregated during mitosis, enabling rapid up- or downregulation of the copy number level [1, 2]. This dynamic behavior allows cancer cells to respond swiftly to treatment and selective pressures [3–6]. Studies have found substantial differences in ecDNA occurrences across different cancer types, with up to 60% in glioblastoma [7, 8] and 18% in medulloblastoma [9], often associated with amplification of oncogenes such as *EGFR* (60% of ecDNA+ samples) [10] and *MYCN/MYC* (37% of ecDNA+ samples) [9], respectively. The presence of ecDNA has been linked with more aggressive disease, with ecDNA-positive medulloblastoma patients three times as likely to die within 5 years compared to patients without ecDNA [9].

Until recently, the genomics of ecDNA has been studied at the bulk tumor level despite its high intratumor heterogeneity. In recent years, advances have made it possible to perform single-cell genomic analysis of ecDNA using sophisticated molecular biology techniques, such as scEC&T-seq and scCircle-seq [9, 11, 12]. This demonstrated that the ecDNA amplicons varied in size and frequency among cells within the same tumor sample. The heterogeneity highlights that the binary assignment of ecDNA presence within a tumor oversimplifies the diversity of ecDNA molecules.

Yet, these methods require laborious analysis of precious samples, including physical isolation of single cells and DNA amplification, a process that is often not feasible due to the scarcity of the material. Here, we aimed to develop an approach, termed ecSingle, to detect and study ecDNA directly from existing scRNA-seq data (source code provided in Data availability). Single-cell RNA-sequencing (scRNA-seq) methods have become a standard technique for investigating expression at the single-cell level from normal and disease tissues, such as cancer. To achieve this, ecSingle computes the allelic imbalance of heterozygous single-nucleotide polymorphisms (SNPs), estimates copy number segments, and identifies outlier expression of oncogenes to discover focally amplified ecDNAs. Here, we apply ecSingle to scRNA-seq from patient-derived tumor samples to investigate the intratumor heterogeneity and clonal evolution of ecDNA at the single-cell resolution. We also demonstrate the utility and increased resolution of using Oxford Nanopore Technologies (ONT) long-read sequencing with ecSingle.

Results

ecSingle integrates allelic imbalance and outlier expression from single cells to identify putative ecDNA segments

To directly identify focal amplifications compatible with ecDNA from a scRNA-seq sample, we employ two key metrics: outlier expression (OE) and allelic imbalance (AI) (Fig. 1), defined as significantly higher expression and significant deviation in B-allele frequency, respectively. These metrics provide information on both the genomic and transcriptomic state of a region, which we integrate with cell type annotation per cluster.

The allelic imbalance is calculated by performing a SNP-based pileup on all cells [13]. ecSingle can either use population-based common SNPs or alternatively be supplied with heterozygous germline SNPs from the patient if available. Subsequently, the B allele frequency (BAF) is determined at each SNP position in each cell. To overcome the sparsity of scRNA-seq data, ecSingle performs cell clustering to obtain a robust BAF estimation.

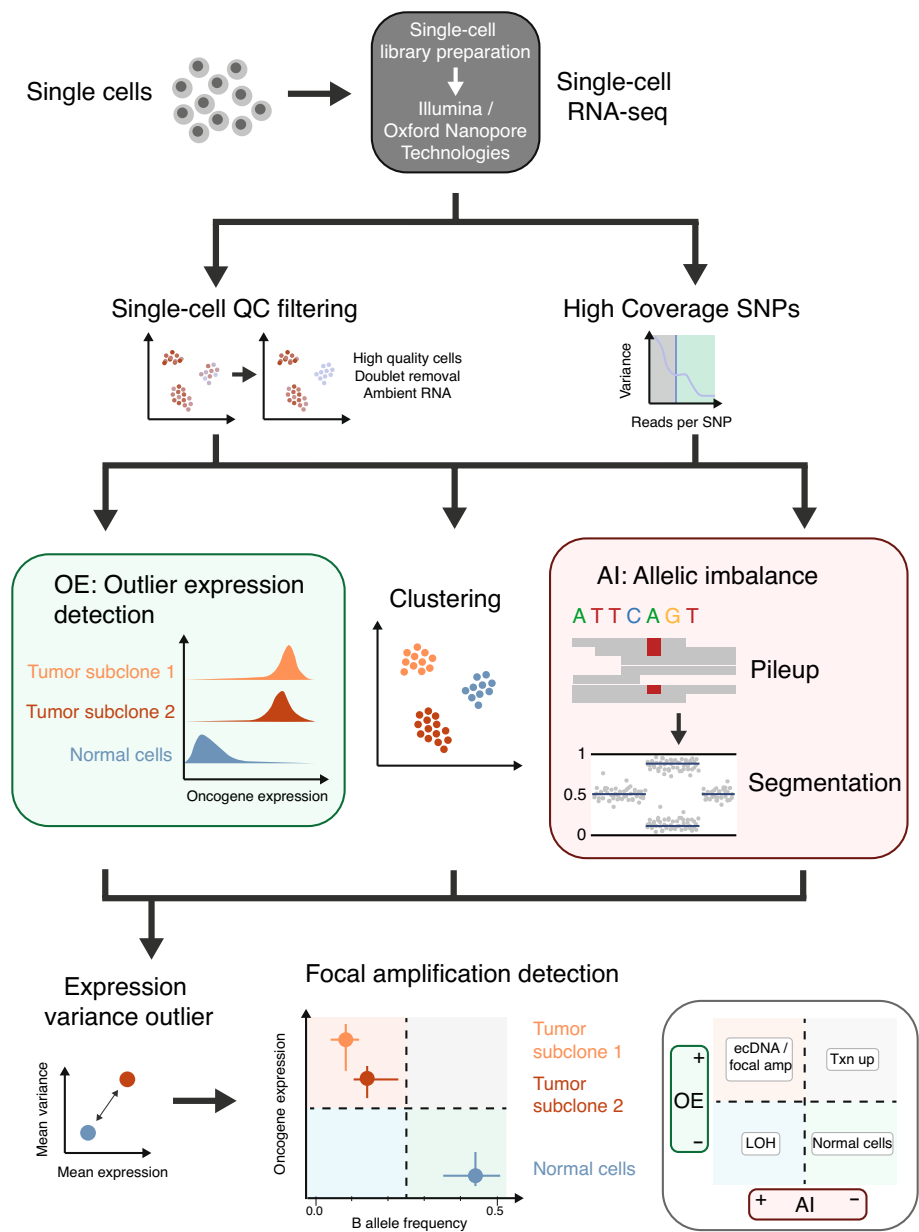


Fig. 1 ecSingle workflow to identify ecDNA from scRNA-seq. ScRNA-seq data (top) is first subjected to single-cell QC filtering strategies (including removal of low-quality cells, doublets, and ambient RNA) as well as SNP filtering (removal of homozygous SNPs and SNPs with low coverage). The resulting scRNA-seq data is next used for outlier expression (OE) detection (middle left), clustering (middle center), and SNP-based allelic imbalance (AI, middle right), to identify putative ecDNA genomic segments. High gene expression level and variance are key properties of genes amplified on ecDNA compared to genomic amplified genes, which we distinguish using an expression variance outlier metric. Genes on genomic segments from single cells are clustered based on these metrics to classify (bottom panel) normal cells (AI-, OE-), LOH (AI+, OE-), transcriptional upregulation (AI-, OE+), and focal amplified ecDNA (AI+, OE+)

A robust BAF estimate depends on a minimum number of reads supporting a SNP. ecSingle estimates the required SNP coverage cutoff to minimize the variance of the BAF in diploid regions (Additional file 1: Fig. S1A-C). BAF values are clustered to distinguish tumor cells from normal cells, and a mean BAF is calculated for each heterozygous SNP.

The tumor-specific heterozygous SNP-based BAF values are used to perform genome-wide copy number segmentation. Candidate focal amplified regions are identified from allelic imbalance segments, identified as deviations of the tumor-specific BAF segments compared to the normal cells. These segments may result from either high-level amplification or loss-of-heterozygosity through a deletion.

To identify genes upregulated on segments with allelic imbalance, a hallmark of ecDNAs, ecSingle uses differentially expressed genes to perform a tumor-specific outlier expression analysis. The distribution of tumor-specific gene expression can be skewed by ambient RNA, cell-free mRNA released during sample preparation, as highly expressed genes are more likely to be detected across the entire dataset. We find that applying ambient RNA removal using CellBender [14] increased signal-to-noise (Additional file 1: Figs. 1 and S1D).

ecDNAs can contain two or more segments from different genomic loci, often at the same copy number level, supporting one connected DNA string. Compared to high-level focal amplifications, also known as homogeneously staining region (HSR), which can also exhibit outlier expression and allelic imbalance, ecDNA typically exhibits high cell-to-cell variability due to unequal segregation between daughter cells [15]. To this end, we calculate an expression variance outlier metric, which identifies genomic segments whose gene expression levels and variances in a candidate ecDNA cell population are distinct from a comparison control cell population (Fig. S2A-B and Methods).

As ecSingle uses two different metrics, allelic imbalance and outlier expression, the method can also identify regions of LOH (loss of heterozygosity; strong allelic imbalance, no outlier expression) and heterozygous regions with gene expression upregulation (no allelic imbalance, high outlier expression). Similar to ecDNAs, these regions can be analyzed within tumor subclusters, enabling the tracing and analysis of genomic and transcriptional heterogeneity with high resolution.

Evaluating ecSingle on ecDNA- and HSR-containing COLO320 cell lines

To assess the performance of ecSingle on detecting ecDNA and distinguishing it from chromosomal amplifications, we used a pair of cell lines derived from the same patient, COLO320: COLO320DM, containing *MYC* on ecDNA, and COLO320HSR, in which *MYC* is in a high-level chromosomal amplification in the form of HSR [16]. The amplicon copy number has been shown to be similar in these cell lines, supporting their use as a controlled comparison between ecDNA and HSR [17].

As the cell lines consist solely of tumor cells without normal cells for comparative analyses, we here defined allelic imbalance as segments with significantly different BAF profiles than the remaining genome ($P < 0.01$, Wilcoxon rank sum test, one-sided). The segmentation of COLO320DM identified allelic imbalance in the expected *MYC*-containing region (chr8:126432938–127901762) (Fig. 2A). While the expression of all genes, including *MYC*, was high within this segment in the COLO320HSR cell line, the

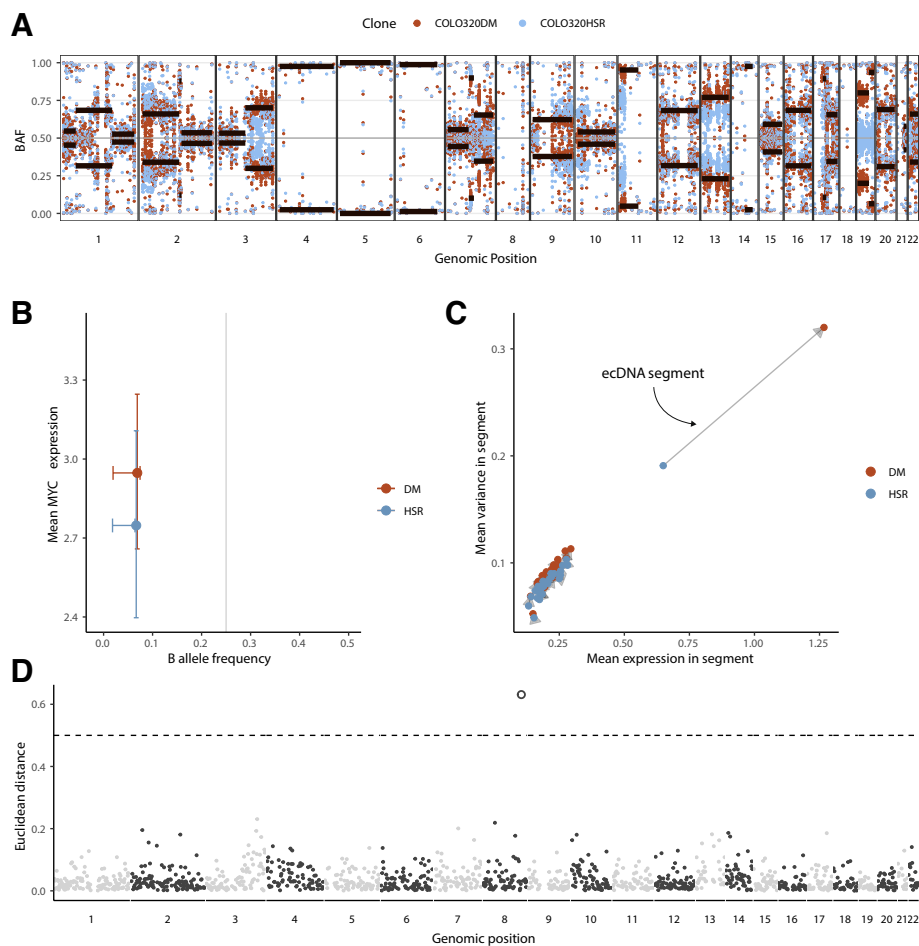


Fig. 2 Benchmarking ecSingle in COLO320 cell lines. **A** Whole genome BAF plot of colon cancer cell lines, COLO320DM and COLO320HSR. Segmentation is performed on the ecDNA+ COLO320DM cells. Allelic imbalance is seen on the majority of chromosomes, including the *MYC* region on chr8. **B** BAF vs expression plot. Mean oncogene (*MYC*) expression vs the BAF value of COLO320DM (DM) and COLO320HSR (HSR) cells. Both cell lines show both outlier expression and allelic imbalance due to the focal amplification of the region; however, the DM cells show higher outlier expression. Lines represent the 25th and 75th percentiles, while the point denotes the mean expression across cells and mean BAF across SNPs in the segment. **C** Mean vs variance plot. Segment-wise gene expression mean and variance in DM and HSR cells. Arrows connect corresponding segments between the two cell lines and point from HSR cells to DM cells, illustrating the Euclidean distance between them. **D** Genome-wide Euclidean distance rank plot (1.5 Mb segments), showing the ecDNA segment at chromosome 8 in COLO320DM vs COLO320HSR cells

expression levels were greater in COLO320DM, suggesting an ecDNA-specific effect on expression (Fig. 2B).

As expected from the known unequal segregation and intratumor heterogeneity of ecDNA-containing cells, we observed greater variability in COLO320DM, suggesting increased heterogeneity within genes located on ecDNA molecules compared to the same genes on the HSR. This likely reflects differences in copy numbers across individual cells, consistent with the dynamic and uneven segregation of ecDNA. Genome-wide, our expression variance outlier metric identified a Euclidean distance of 0.63 between COLO320DM and COLO320HSR for this segment, with the second-highest genome-wide Euclidean distance at 0.06 (Fig. 2C). Across nine ecDNA-containing samples, we established Euclidean distance and expression variance thresholds of 0.5 and 0.25,

respectively, as robust thresholds for identifying segments likely to be located extrachromosomally (Additional File 1: Fig. S2A-B).

To further corroborate the distance metric, we partitioned the genome into 1.5 Mb segments and calculated the variance and Euclidean distance between COLO320DM and COLO320HSR for each segment. We found that the ecDNA segment showed a variance of 0.32 in COLO320DM cells, resulting in a threefold higher Euclidean distance than the segment with the second-highest distance score, supporting its distinct transcriptional profile (Fig. 2D, Additional File 1: Fig. S2C-D). These results support the ability of ecSingle to detect ecDNA-specific amplifications within single cells and that expression variance is distinct between ecDNA and chromosomal amplifications.

Identification of oncogene-containing ecDNA from cancer single-cell data

To assess the performance on cancer patient material, we applied ecSingle to a cohort of patients with available scRNA-seq data (see Additional File 1: Table S1). For one glioblastoma sample, CE34, we identified whole-chromosome imbalance of chromosomes 7, 10, 19, and 20 (Fig. 3A).

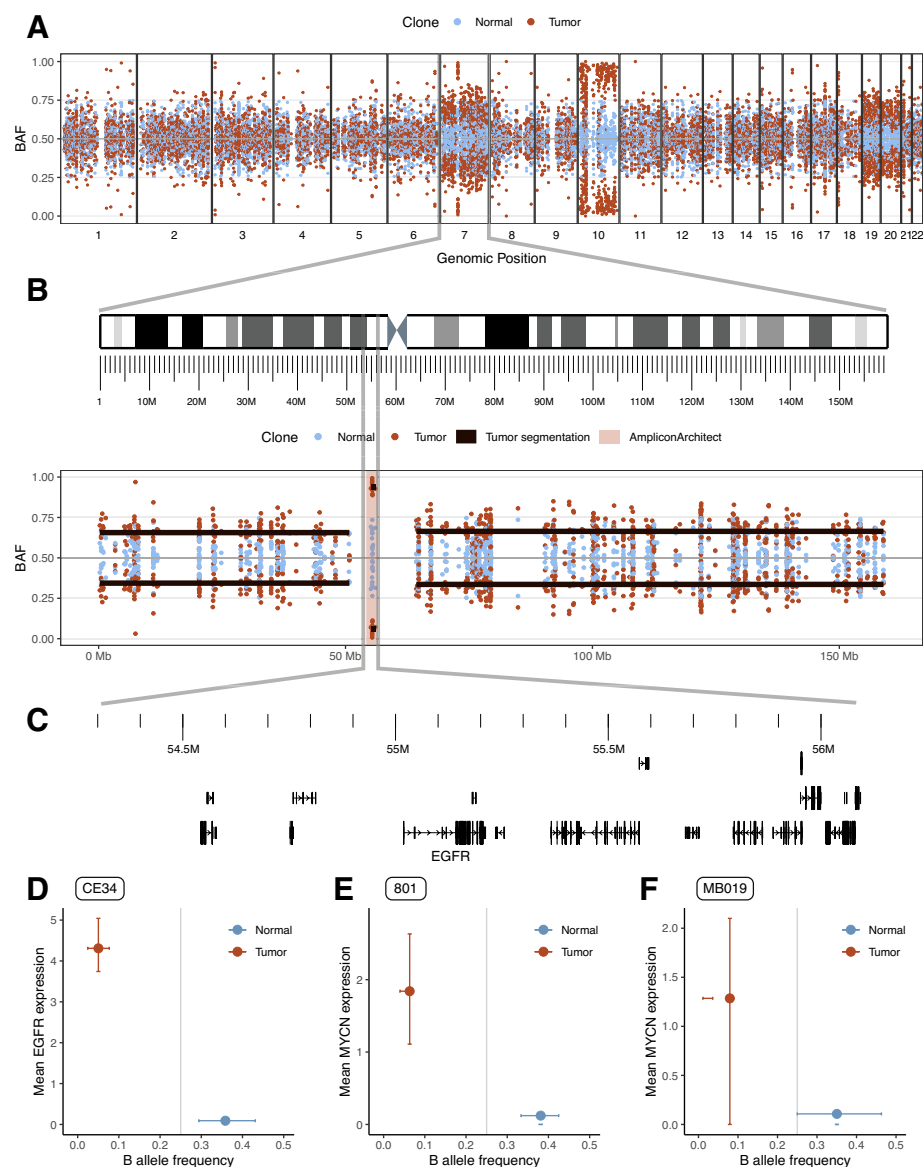
ecSingle identified two segments in CE34 with different allelic imbalances on chromosome 7, a large whole chromosome region displaying a BAF around 0.33/0.66, suggesting triploidy, and a focal segment at chr7:54295003–56099902. The latter displayed a substantial allelic imbalance shift, with BAF values close to 0 (0.05/0.95), suggesting either a deletion or a high-level amplification of this region (Fig. 3B).

The region contained a total of 22 genes, from the lipid regulator *VSTM2A* at 54.5 Mb to the protein kinase *PHKG1* at 56.1 Mb (Fig. 3C). Across these 22 genes, 10 were sufficiently expressed in the sample to be detected by ecSingle, including *EGFR*, a gene known to be highly amplified in around 50 % of glioblastoma tumors (Fig. 3D) [18]. All 10 genes were significantly upregulated in the tumor cells compared to normal cells (binomial test, $p < 0.01$). These findings strongly support the presence of high-level amplification, likely on ecDNA, due to the high copy number and expression.

To assess the accuracy of the segmentation, we performed WGS ($84 \times$ genome coverage) on the same tumor sample and used the state-of-the-art tool AmpliconArchitect to detect ecDNA segments [19]. This approach found the region chr7:54212672–55359360 as cyclic ecDNA, with both breakpoints differing by less than 800 Kb from the scRNA-seq-based region identified by ecSingle. This demonstrates that ecSingle successfully identifies the same ecDNA regions as detected by WGS (Fig. 3B-D).

While *EGFR* is the most frequently amplified oncogene in glioblastomas [10], we also identified *PDGFRA* and *NFIB* amplified on ecDNAs in another glioblastoma sample, CE65, both associated with outlier expression (Additional File 1: Fig. S3).

The allelic imbalance and outlier expression approach of ecSingle can also identify other transcriptional deregulation patterns in addition to ecDNA segments. For example, we identified a region showing (biallelic) transcriptional upregulation (*GPM6A* in CE34, Additional File 1: Fig. S4A), another region showing allelic imbalance due to loss-of-heterozygosity (*PTEN* in CE34, Additional File 1: Fig. S4B), and regions with minor allelic imbalance and outlier expression, which we classify as low copy-number chromosomal gains (*GNAS* in CE34 and *EGFR* in CE65, Additional File 1: Fig. S4C-D).



To evaluate the performance of ecSingle on other cancer types, we tested the approach on medulloblastoma samples from Riemondy et al. [20] and Gold et al. [21]. We identified *MYCN* on ecDNAs in two medulloblastoma samples (801 and MB019), both belonging to the SHH subgroup (available copy number data for 801 supports *MYCN* amplification) (Fig. 3E-F). ecSingle did not detect ecDNA in scRNA-seq data from three other medulloblastoma samples from Gold et al., with sufficient scRNA-seq coverage but unknown ecDNA status (Additional File 1: Fig. S4E). In support, we found no evidence for gene expression upregulation of either *MYC* or *MYCN* in these samples.

Detection of ecDNA is enhanced through long-read single-cell RNA-sequencing

Segmentation relies strongly on the number of covered SNPs, which in turn depends on the sequencing depth across cells as well as the read length. The vast majority of single-cell datasets, however, consist of short-read sequencing, limiting the number of candidate SNPs.

To understand the influence of read length on the number of informative SNPs for segmentation and test to what extent higher coverage can compensate for short reads, we performed long-read sequencing of a 10xGenomics library of sample CE26. To evaluate the practicality of ONT long-read sequencing for this use case, we additionally performed WGS on the sample and analyzed the data using AmpliconArchitect. The AmpliconArchitect segments showed consistent agreement with the segments identified by ecSingle using the long-read scRNA-seq dataset, despite the sample's high complexity in some regions, including chromosome 7 (Fig. 4A-B). The ecSingle approach detected two ecDNA segments in this patient, one containing outlier expression of *EGFR* in tumor cells, supporting the ecDNA status of this sample (Fig. 4C).

Besides the two ecDNA regions, the ecSingle segmentation also identified allelic imbalance in two other segments on chromosome 7. The strong allelic imbalance combined with the low expression in these segments suggests LOH, a finding supported by analysis of the WGS (Additional File 1: Fig. S5). AmpliconArchitect found an additional small region at 10.5–10.8 Mb included in the ecDNA amplicon (362 Kb), a region too small to be detected in scRNA-seq data.

To evaluate the effect of the long-read sequencing, we compared three read lengths of the same sample. These included: 1) short-read; regular Illumina short-read 91 bp (as recommended by 10xGenomics), 2) medium-read; Illumina short-read sequenced with average read length 293 bp, 3) long-read; long-read ONT sequencing with a median read length of 1,144 bp.

We defined informative SNPs as the number of SNPs that are heterozygous and have sufficient coverage (min 300 cells) across the sample. Each of the three datasets was downsampled, and the number of informative SNPs was determined, allowing us to fit a polynomial regression.

We found that the amount of informative SNPs was highly dependent on the read length, as expected. At the highest sequencing depth, the short-read dataset recovered 34% (3645/10572) of the SNPs identified using the medium-read dataset (Fig. 4D, Additional File 1: Fig. S5A), demonstrating the added value of longer sequencing reads. Furthermore, we estimated that the short-read dataset would reach saturation around 3,700 SNPs (sigmoid function, Fig. 4D). In comparison, we found the estimated sequencing

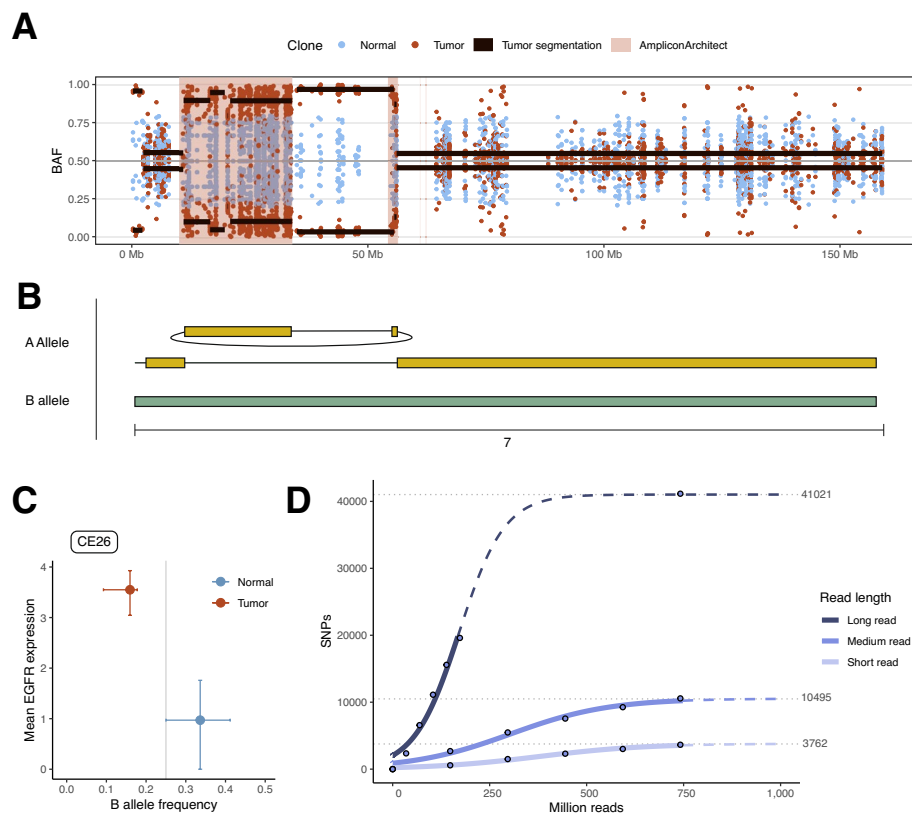


Fig. 4 Improved allelic imbalance detection using long-read sequencing to identify ecDNA. **A** Long-read scRNA-seq segmentation of chromosome 7 in CE26, similar to Fig. 2A. **B** Schematic of the expected ecDNA construct and A and B alleles of chromosome 7 in CE26. **C** BAF vs expression plot. Mean oncogene (*EGFR*) expression vs the BAF value of tumor and normal cells in CE26. The tumor cells show both outlier expression and allelic imbalance due to the ecDNA in the region. **D** The effect of read length on the amount of informative SNPs in sample CE26. Three read lengths of the same sample are compared: short-read (91 bp), longer-sequenced short-read (293 bp), and long read (median 1,144 bp). The coverage of each dataset is downsampled, and a polynomial regression is fitted. The dotted line represents extrapolation from a polynomial fit

depth of the long-read dataset to reach 41,000 informative SNPs at the same sequencing coverage as short and medium, providing a 10.9-fold increased transcript coverage compared to the standard short-read set.

Resolving intra-tumor heterogeneity of ecDNA at the single-cell level

A major advantage of single-cell data is the ability to resolve intratumor heterogeneity compared to bulk RNA-seq or WGS, which relies on subclonal inference. Similar tumor-subclones will tend to cluster together at the gene expression and genomic level, and we utilize this to group the cells through genomically aware single-cell clustering techniques (e.g., InferCNV) as demonstrated here. Alternatively, gene expression-based clustering (e.g., Leiden [22]) can be used as input, for example, in cases where distinct transcriptional clusters are present.

To explore the intratumor heterogeneity in an external single-cell dataset, we took advantage of a 10x Genomics scRNA-seq dataset of glioblastoma samples from Bhaduri et al. [23]. In one representative sample with adequate sequencing coverage (>2,000

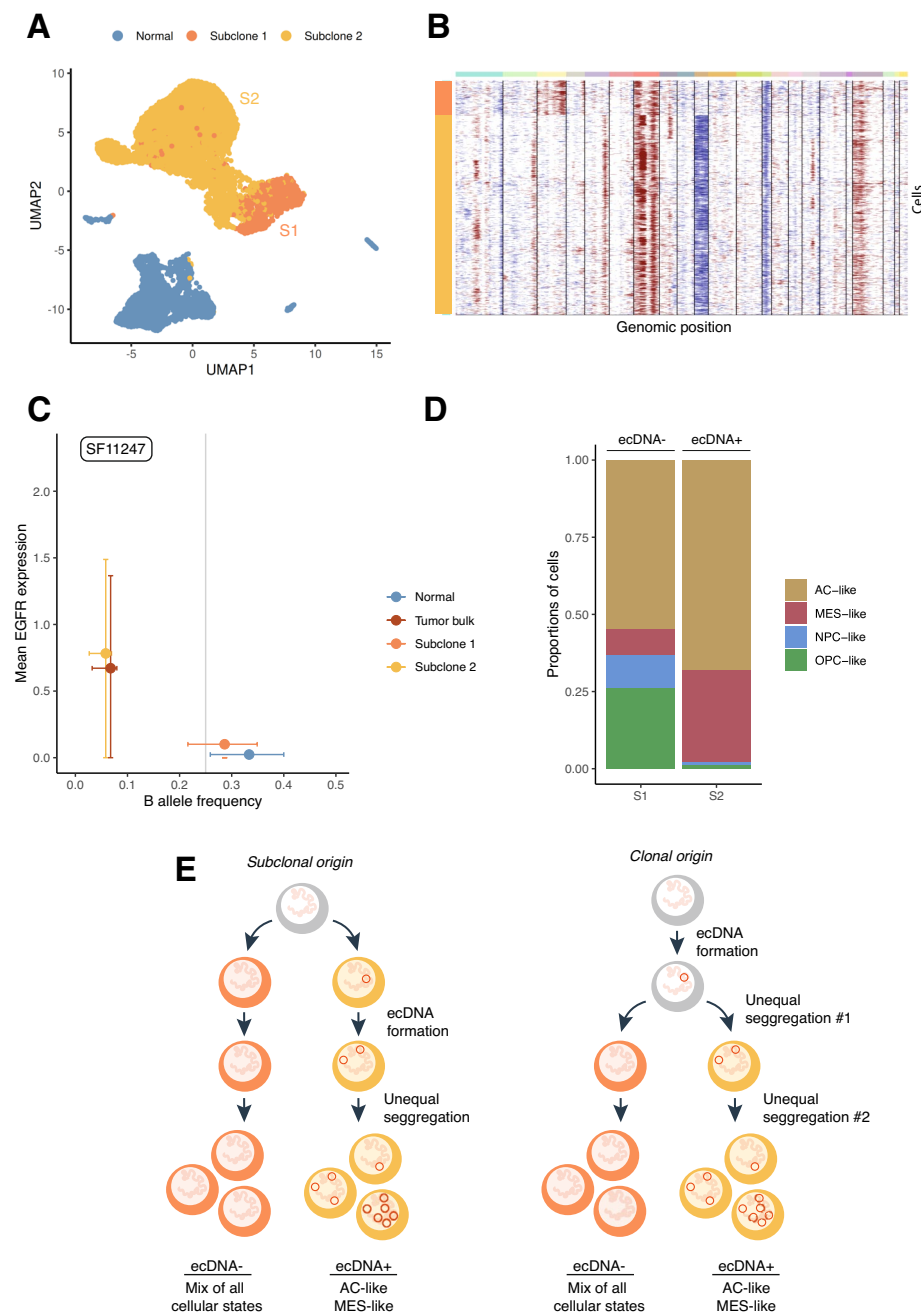


Fig. 5 ecSingle identifies subclonal variation in ecDNA levels. **A** UMAP of sample SF11247, colored by tumor subclone S1 and S2. **B** Copy number alterations (InferCNV) of SF11247 with chromosomes horizontally, and cells vertically grouped into subclones (red and blue colors represent gains and losses, respectively). The two subclones are colored in the column. **C** BAF vs expression plot of the ecDNA region. Mean oncogene (*EGFR*) expression (y-axis) vs the BAF value (x-axis) of the tumor subclones and normal cells, as well as tumor bulk level (pseudo-bulk of all tumor cells). **D** Barplot of the glioblastoma cellular states across subclones in SF11247 with the ecDNA identity shown above. **E** Schematic of hypothetical evolution by unequal segregation during cell division driving intratumor heterogeneity of ecDNAs in SF11247. Subclonal and clonal origin of the ecDNA can both be plausible explanations for the ecDNA presence in subclone 1

UMIs/cell), SF11247, we initially performed an unbiased pseudo-bulk tumor analysis. This analysis revealed a region around *EGFR* with both allelic imbalance and outlier expression, suggesting an ecDNA, but with a wide distribution in both allelic imbalance and outlier expression, indicating a heterogeneous tumor population. Through k-means subclonal clustering, we identified two major tumor subclones (Fig. 5A). The main copy number events were shared across clones, including a deletion on chromosome 13, a whole-chromosome amplification of chromosome 7, and amplification of the p-arm of chromosome 19. Subclone 1 differed from subclone 2 with an amplification of chromosome 3 and no deletion of chromosome 10, while subclone 2 exhibited smaller CNVs (Fig. 5B and Additional File 1: Fig. S6). Upon separating the tumor cells into subclones, we found that subclone 1 showed no signs of ecDNA, while subclone 2 exhibited both allelic imbalance and outlier expression, suggesting ecDNA presence.

We investigated the clonal evolution of ecDNA in NVB33 for which both scRNA-seq and WGS data were available at the primary tumor and relapse timepoints [24] (Additional File 1: Fig. S7A). Here we found that the smaller subclone 1 exhibited high levels of ecDNA containing *EGFR* (strong allelic imbalance and outlier expression in a region surrounding *EGFR*), while the larger subclone 2 displayed no signs of focal amplification in this region (Additional File 1: Fig. S7B-C). Strikingly, we found the same ecDNA-containing cells to be clonal in the relapse sample from the same patient (Additional File 1: Fig. S7D). We noted a lower level of ecDNA in the relapse (reduced allelic imbalance and outlier expression), suggesting clonal selection of the ecDNA-positive clone, followed by copy-number downregulation due to treatment. Similar dynamics have previously been shown in ecDNA+ cell lines, in which treatment caused a rapid decline in the levels of ecDNA [2, 6].

In addition to genomic heterogeneity, glioblastoma tumors are characterized by extensive transcriptional plasticity and heterogeneity, including four cellular states: astrocyte-like (AC-like), mesenchymal-like (MES-like), oligodendrocyte precursor cell-like (OPC-like), and neural progenitor cell-like (NPC-like) [25]. To this end, we next investigated the extent to which cellular states associate with ecDNA at the subclonal level. In the majority of glioblastomas, all four states are found, but at varying proportions. In agreement, we identified all four states to be present in SF11247. This was recapitulated in clone S1 (ecDNA-), dominated by OPC- and AC-like cell states, with smaller contributions from MES-like and NPC-like states (28%, 56%, 5%, 11%, respectively). Interestingly, we found a striking shift in cell states in the ecDNA+ clone (S2), dominated by AC-like and MES-like cell states (Fig. 5D).

A similar pattern was observed in NVB33, in which the ecDNA+ subclone in the primary tumor was dominated by AC-like, while the relapse sample contained mainly AC-like and MES-like cells (Additional File 1: Fig. S7E). We note that *EGFR*, present on the ecDNA in all these subclones, is known to be enriched in AC-like tumors [25]. In the paired primary and relapse samples from patient NVB33, we identified no change in ecDNA segment size; however, previous studies have suggested that ecDNA can evolve structurally over time [26].

These findings strongly suggest that both clonal selection and transcriptional plasticity contributed to the evolution in these samples. A cell state-specific ecDNA presence can favor certain cell states, for example, AC-like and MES-like states in glioblastoma,

likely driven by *EGFR*. This is supported by the previous finding on a bulk-tumor basis that the AC-like cell state is associated with a high level of *EGFR* amplification, whereas OPC-like tumors exhibit low levels. This phenomenon may arise from either a subclonal origin of ecDNA, where ecDNA formation occurs within a specific tumor subclone, or a clonal origin followed by uneven segregation, resulting in ecDNA loss in certain subclones (Fig. 5E). Further investigation is needed to clarify these dynamics, which will be critical for understanding ecDNA behavior and its role in tumor evolution. In summary, we demonstrate that intratumor heterogeneity can be disentangled at the single-cell level, uncovering joint genomic and transcriptional programs in glioblastoma.

Discussion

It is becoming increasingly clear that ecDNAs are highly relevant for driving intratumor heterogeneity and cancer progression across a wide range of cancers [10, 27]. Current approaches are primarily based on bulk whole-genome DNA sequencing, which does not allow a sensitive analysis and detection of ecDNA-containing tumor cells at the subclonal level. This is especially important with the recent identification of ecDNA-containing tumor cells frequently driving relapse [9]. Such cells can be present below the detection level in the primary tumor and thereby complicate and potentially misinterpret the tumor evolution in cancer patients. Here, we demonstrate a computational methodology, ecSingle, to identify and explore ecDNAs at the single-cell level using standard scRNA-seq methodologies. We identify striking examples of subclonal ecDNA-containing tumor cells, in one case with a minor ecDNA+subclone in the primary tumor expanding to become the dominant clone in the tumor relapse. ecDNA molecules have been found to segregate in a non-Mendelian random manner between daughter cells [1, 2]. In support, we find several lines of evidence for differences in ecDNA content between tumor cells. This is supported by complementary and more laborious methods such as scEC&T-seq [12] and scCircle-seq [11]. While both methods use sophisticated molecular biology methods on both DNA and RNA, these cannot be applied to existing available scRNA-seq, a major bottleneck in most setups, particularly for limited patient-derived tumor material.

We acknowledge that our analyses are limited to cases in which ecDNA presence is accompanied by transcriptional upregulation, which is not universally observed. Even among ecDNA+ cases with elevated transcription, expression levels may vary independently of copy number, a phenomenon previously reported by Nathanson et al. [6].

An advantage of ecSingle is the ability to identify transcriptional upregulation versus allele-specific upregulation, such as in ecDNA formation, since transcriptional upregulation is generally found to be without major allelic imbalance. Also, our methodology is able to identify genes subjected to heterozygosity and accompanied loss in expression.

Compared to genome sequencing approaches, an additional advantage of using scRNA-seq is the ability to identify and integrate transcriptional cell states and programs with ecDNA discovery. We find a prominent co-occurrence between the oncogene-carrying ecDNA+ clone and transcriptional cell states in glioblastoma, marked by increased MES-like and AC-like states in ecDNA+ tumor cells. This is in agreement with prior studies showing the clustering of *EGFR*-amplified glioblastomas with “Classical” and AC-like cell states [25, 28]. While these findings suggest an *EGFR*-driven mechanism, it

is perceivable that the presence of ecDNA molecules can also impact transcriptional cell states. We note that our approach to identifying subclonal clusters was based on a limited number of tumor samples, and that larger cohorts will be needed to validate these findings.

We utilized four properties of ecDNAs: outlier expression, allelic imbalance, outlier expression variance, and clonal status. The allelic imbalance-based segmentation of copy number alteration regions is an important step in identifying the ecDNA sequence. Interestingly, we found that a simple metric for expression level and variance was able to distinguish between ecDNA and HSR, likely due to the unequal segregation—and therefore higher intratumor heterogeneity—of ecDNA-containing tumors. We note, however, that larger well-annotated cohorts will be needed to further validate the sensitivity and specificity of ecSingle and our empirical thresholds in detecting ecDNA and distinguishing between ecDNA and HSRs. As expected, we found the read length to significantly impact the number of informative SNPs, a key metric for allelic imbalance assessment. Current scRNA-seq methods depend on 3'-end capture, which results in limited coverage of SNPs in the non-capture region of the transcript. We also found the sequencing depth to be important, which impacts the ability to estimate the BAF at each SNP position. Moreover, the number of informative SNPs is influenced by the level of splicing in the samples, which varies depending on whether nuclei or whole-cell data is used. Since introns cover 4–5 times more of the human genome than exons, unspliced RNA offers a significantly higher number of potential SNPs for detection [29].

We note that an initial separation between tumor and normal stromal cells improves the sensitivity to identify ecDNA. The identity of normal cells can be a source of challenge, in particular, if the transcriptional differences between normal and tumor cells are minor due to similar gene expression and few CNVs. A limitation of our method is the requirement for larger segments and the presence of genes with detectable gene expression and ascertainable germline SNPs within these segments. Moreover, while ecSingle leverages several key features to identify ecDNA from scRNA-seq, follow-up DNA-based validation experiments should be performed to further support the presence of ecDNA molecules.

Tumor cells have previously been observed to transfer DNA via extracellular vesicles [30, 31]. This phenomenon suggests that ecDNA molecules could also be transferred from ecDNA-containing tumor cells to other tumor subclones, and potentially even to normal cells in the tumor microenvironment. Given the small and extrachromosomal nature of ecDNA, these amplicons could hypothetically easily integrate and influence gene expression in the recipient cell. In future studies, ecSingle could be used to test this hypothesis, for example, by utilizing plate-based scRNA-seq data to eliminate potential contamination from other cells.

Conclusions

We develop and apply a novel methodology, ecSingle, that enables the detection of onco-gene-containing ecDNA from standard scRNA-seq data using two key properties: SNP-based allelic imbalance and outlier expression. Notably, we demonstrate the superiority of using long-read single-cell sequencing for increased resolution of informative SNPs. Subclonal ecDNA was observed in multiple samples, linking the ecDNA presence with

specific transcriptional states, highlighting the interplay between genomics and cellular consequences. Furthermore, ecSingle enabled us to identify examples of clonal evolution and selection of oncogene-containing ecDNAs from primary to relapse in brain tumors. These findings demonstrate the importance of resolving intratumor heterogeneity to identify the subclonal presence of ecDNA, and that this is an important characteristic of tumors to understand the clonal expansion and therapy resistance.

Methods

Patient inclusion

Patients were 18 years of age or older and diagnosed with glioblastoma, IDH wild-type WHO grade 4. All patients were treated at Rigshospitalet, Copenhagen, Denmark. All ethical and data management approvals were obtained through the Neurogenome protocol (Danish National Medical Research Ethics Committee approval number H-21023801) [32].

Patient material and nuclei isolation

Fresh tumor biopsy samples were collected directly at the neurosurgical theatre and cryopreserved in nitrogen. Tissue was either flash-frozen in liquid nitrogen or frozen in RNeasy lysis buffer and stored until library preparation. Frozen tissue was homogenized using BioMasher III (Nippi Inc.), and the nuclei suspension was prepared and purified using the protocol by Batiuk et al. [33]. After purification using ultracentrifugation, nuclei were immediately used for scRNA-seq generation.

Sequencing

The quality and concentration of each single-nuclei suspension were estimated using Trypan Blue staining and light microscopy. From each sample, approximately 16,000 nuclei were loaded onto a well on the Chromium Controller (10xGenomics) to aim for an output of 10,000 nuclei. cDNA construction and library generation were achieved using Chromium Next GEM Single Cell 3' Gel Bead and Library kit, v3.1 single index or v3 (10xGenomics, Pleasanton, CA, USA) using the manufacturer's instructions. Finished libraries were quantified using Qubit (Thermo Fisher Scientific) and profiled using Bioanalyzer High Sensitivity DNA kit (Agilent Technologies). The sequencing was performed using paired-end 150 base-pair mode on a NovaSeq S4 flow cell (Illumina) on an Illumina NovaSeq 6000 machine at the Department of Genomic Medicine at Rigshospitalet.

scRNA-seq processing

Initial processing of short-read sequencing output was performed using Cell Ranger software version 5.0.0. The raw count matrices were corrected for ambient RNA using CellBender version 0.1.0 [14]. The corrected count matrices were imported into R (version 4.1.0) and analyzed using Seurat version 4.1.1 [34]. The data underwent extensive quality control (QC) and filtering based on the number of genes detected and the mitochondrial content.

Cell typing was performed, as described in Hendriksen et al. [24], through a combination of three complementary approaches: 1) CopyKat version 1.0.4 [35], 2) SingleR

version 1.0.5 [36] and 3) literature-curated marker genes. As marker genes, we used the following: malignant cells (*EGFR*), myeloid (*PTPRC*, *MRC1*), T cells (*PTPRC*, *CD3G*, *CD8A*), oligodendrocytes (*MOG*), fibroblasts (*DCN*), endothelial cells (*DCN*, *PECAM1*), neurons (*STMN2*), and astrocytes (*GFAP*). Genomically aware cell clustering was performed using InferCNV (inferCNV of the Trinity CTAT Project, <https://github.com/broadinstitute/inferCNV>).

The long-read scRNA-sequencing data were processed using the wf-single-cell v2.3.0 pipeline from EPI2ME Labs. Briefly, the cellular barcodes and UMIs were extracted, followed by alignment using minimap2 [37] to the cellranger reference genome refdata-gex-GRCh38-2020-A. Stringtie [38] was used to generate a transcriptome of transcripts called at least 2 times, and the reads are mapped to this transcriptome to generate the expression matrix output.

Pileup

To achieve a pileup of 10xGenomics scRNA-seq data, cellsnp-lite was run on the Cell Ranger BAM files [13] using mode 1a. As the input of SNPs, 1000genomes was used. Parameters were set to require a minimum BAF of 10% as well as an overall coverage of 100 UMIs to limit the number of output SNPs.

BAF calculation

From the pileup output, the matrix of alternative alleles (AD) as well as read depth (DP), which contains the reference and alternative alleles summed, was used for downstream analysis. For each non-zero position in the read-depth matrix, the ratio of alternative allele (BAF) was calculated.

$$BAF = \frac{AD}{DP}$$

Following the BAF calculation, the SNPs undergo several filtering steps to ensure heterozygosity and adequate coverage. Initially, SNPs with overall low coverage are excluded (see *Coverage cutoff calculation*). Subsequently, SNPs that fail to meet a per-group cutoff are discarded, ensuring sufficient coverage across both the normal and tumor cells. After the coverage filtering, the heterozygosity of SNPs is measured through their mean BAF across all normal cells. Positions that fall outside the range of 0.25–0.75 BAF are labelled as homozygous and discarded, as these are not informative for this analysis.

When the cells are divided into smaller subclones, SNPs are still filtered based on their coverage across all tumor cells. However, the BAF of a SNP is only recorded if a subclone has at least 3 cells covering the position. Following the filtering, the mean BAF of each group is calculated per position.

Cellular clustering

To cluster the single cells into subclones, we use k-means clustering with the InferCNV HMM predictions as input. The optimal number of clusters is determined as the elbow point in the within-cluster sum of squares (WSS) plot.

Coverage cutoff calculation

To balance the increased information from a higher number of informative SNPs with the noise introduced by low coverage SNPs, ecSingle determines an optimal cutoff to exclude low coverage SNPs. This is accomplished by evaluating the BAF variance across various coverage cutoffs.

First, SNPs underwent standard filtering procedures (see *BAF calculation*), using a starting coverage cutoff requiring 100 cells with SNP coverage per SNP. Next, the chromosomes that predominantly contain heterozygous regions in tumor cells are selected, requiring at least tenfold more SNPs with BAF above 0.25 than below 0.25. For each of the selected chromosomes, the variance of BAF values of positions with a BAF above 0.25 was measured. Lastly, the mean across the selected chromosomes was computed, providing a measure of the BAF variance in the regions unlikely to exhibit LOH.

This measure was calculated at different coverage levels, ranging from coverage in 100 to 500 cells. The optimal cutoff was defined as the first point where the variance increased at two consecutive points. Simultaneously, the number of informative SNPs was measured at each coverage level, ensuring that the coverage across the genome remained adequate.

Segmentation

Following the calculation of the BAF values per group, segmentation is performed on the tumor BAF values using Quickseg, a dynamic programming algorithm, to identify the genomic segments.

For each of the segments identified in the tumor cells, the BAF values of the SNPs in the segment are compared to the corresponding BAF values in the normal cells. A Wilcoxon test is used to identify segments with a significant difference in BAF values between tumor and normal, and these are selected as potentially altered segments for downstream analysis.

The minimal detectable segment size is dependent on the number of genes and SNPs. We found the minimal ascertainable ecDNA segment size to be 400 Kb, which contained 17 informative SNPs across 6 detected genes.

Differential gene expression

Differential gene expression between tumor and normal cells was calculated using the Wilcoxon-rank-sum test in the FindMarkers function in Seurat. Input was log-normalized gene counts. Genes were annotated as oncogenes according to the annotation file from AmpliconArchitect [19].

A binomial test was used to identify segments with a significant number of genes upregulated in tumor cells compared to the overall level of gene upregulation.

Expression variance outlier detection

Gene expression counts for each segment were extracted and stratified by cell group. For each gene within a segment, we calculated the mean and variance of expression. To

quantify differences in expression variance between two populations, here denoted A and B , we computed the following metrics:

X_A, X_B : Mean expression levels

Y_A, Y_B : Mean expression variances

ED_{AB} : Euclidean distance between the two populations in the expression-variance space

The Euclidean distance was calculated using the formula:

$$ED_{A,B} = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}$$

We consider population A to exhibit an *expression variance outlier* property relative to B if the following conditions are met:

1. $X_A > X_B$
2. $Y_A > \max(Y_B, 0.25)$
3. $ED_{AB} > 0.5$

A Euclidean distance threshold of 0.5 combined with a variance threshold of 0.25 was found to reliably distinguish putative ecDNA segments from chromosomal segments.

Allelic imbalance and outlier expression plot

For the final plot, the identified segments were used to calculate the mean BAF as well as the 25th and 75th quartiles of each cell group individually. Similarly, the gene expression was summarized through the identification of oncogenes in the segment or, alternatively, the gene with the highest log fold change, suggesting oncogene-like selection. The expression of the chosen oncogene is used to calculate the mean, 25th, and 75th quartiles for each cell group used for the plot.

Whole genome sequencing

DNA was extracted from fresh-frozen tumor tissue, as well as matched blood, and underwent paired-end sequencing on the NovaSeq platform.

The resulting FastQ files underwent QC using “FastQC 0.11.8” (Andrews, [39]) and were aligned to the human reference genome (GRCh38) using “BWA MEM 0.7.15” [40]. Data preprocessing was performed following the GATK Best Practices [41], using “GATK 4.1.9.0”: “MarkDuplicates” was used to identify PCR and optical duplicates; base quality score recalibration was performed using “BaseRecalibrator” followed by “ApplyBQSR”. Coverage statistics were gathered using “CollectWgsMetrics” and “mosdepth 0.3.1” [42]. QC of the final BAM files was performed by running “CollectAlignmentSummaryMetrics”, “CollectBaseDistributionByCycle”, “CollectGcBiasMetrics”, “CollectInsertSizeMetrics”, and “QualityScoreDistribution”.

QC results were inspected using “MultiQC 1.9” [43].

“Somalier 0.2.11” [44], a genetic distance-based tool, was applied to verify the correct pairing of the samples. Stromal cell contamination and ploidy were estimated using “Sequenza 3.0.0” [45].

Amplicon architect

Substructures of genomic focal amplifications were extracted from WGS samples using AmpliconArchitect [19], which was run using default parameters on tumor BAM files downsampled to 10 × and considering only genomic segments with integer copy-number greater than 5. AmpliconArchitect inferred structures were subsequently classified according to their generative mechanisms using AmpliconClassifier [19].

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-026-03933-2>.

Additional file 1: Fig. S1. Preprocessing of scRNA-seq data. Fig. S2. Variance outlier detection across samples. Fig. S3. Identifying ecDNA-containing PDGFRA and NFIB. Fig. S4. Allelic imbalance and outlier expression in non-ecDNA regions. Fig. S5. Comparison of SNP coverage across read lengths. Fig. S6. Segmentation of SF11247. Fig. S7. Subclonality in NVB33. Table S1. Clinical information on the samples, including age, sex, treatment, and scRNA-seq specific QC parameters.

Acknowledgements

The authors would like to thank all the patients who participated. We would also like to thank everyone at the DCCC Brain Tumor Centre and the Centre for Genomic Medicine, Rigshospitalet.

Peer review information

David Posada and Claudia Feng were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

Conceptualization, J.D.H. and J.W.; Software, J.D.H., A.L., and B.C.S.; Formal Analysis, J.D.H. and A.L.; Investigation, J.D.H.; Resources, J.S., C.W.Y., D.S.N., H.S.P., and U.L.; Writing – Original Draft, J.D.H. and J.W.; Writing – Review & Editing, J.D.H. and J.W.; Visualization, J.D.H.; Supervision, J.W.; Project Administration, H.S.P., U.L., and J.W. All authors read and approved the final manuscript.

Funding

Open access funding provided by Copenhagen University This work was supported by the Danish Cancer Society (#R295-A16770) and Independent Research Fund Denmark (#3101–00177). J.W. and J.D.H. were supported by Novo Nordisk Foundation (#NNF200C0060141).

Data availability

The code used in this study is available in our Bitbucket repository (<https://bitbucket.org/weischenfeldt/ecsingel>) [46] and archived on Zenodo (<https://doi.org/10.5281/zenodo.17790701>) [47]. The raw sequencing data are deposited on the European Genome Phenome Archive EGAD00001010843 [48]. The following publicly available data were also used in this study: COLO320DM and COLO320HSR (SRA: PRJNA672109) [16]; medulloblastoma sample 801 (SRA: PRJNA649773) [20]; medulloblastoma MB019, MB084, MB4113, and MB0595 (SRA: PRJNA885474) [21]. The raw sequencing data are deposited on the European Genome Phenome Archive EGAD00001010843 [48]. The following publicly available data were also used in this study: COLO320DM and COLO320HSR (SRA: PRJNA672109) [16]; medulloblastoma sample 801 (SRA: PRJNA649773) [20]; medulloblastoma MB019, MB084, MB4113, and MB0595 (SRA: PRJNA885474) [21].

Declarations

Ethics approval and consent to participate

The study is approved by Videnskabs Etisk Råd (VEK) (case number 1707335). The Declaration of Helsinki is the accepted basis for clinical study ethics, and the study is performed in accordance with the international standards of Good Clinical Practice and according to all local laws and regulations concerning clinical studies.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 5 March 2025 Accepted: 5 January 2026

Published online: 29 January 2026

References

- deCarvalho AC, Kim H, Poisson LM, Winn ME, Mueller C, Cherba D, et al. Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat Genet.* 2018;50:708–17.
- Lange JT, Rose JC, Chen CY, Pichugin Y, Xie L, Tang J, et al. The evolutionary dynamics of extrachromosomal DNA in human cancers. *Nat Genet.* 2022;54:1527–33.
- Verhaak RGW, Bafna V, Mischel PS. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat Rev Cancer.* 2019;19:283–8.
- Wu S, Turner KM, Nguyen N, Raviram R, Erb M, Santini J, et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature.* 2019. <https://doi.org/10.1038/s41586-019-1763-5>.
- Morton AR, Dogan-Artun N, Faber ZJ, MacLeod G, Bartels CF, Piazza MS, et al. Functional Enhancers Shape Extrachromosomal Oncogene Amplifications. *Cell.* 2019;179:1–12.
- Nathanson DA, Gini B, Mottahedeh J, Visnyei K, Koga T, Gomez G, et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science.* 2014;343:72–6.
- Kim H, Nguyen N-P, Turner K, Wu S, Gujar AD, Luebeck J, et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet.* 2020;52:891–7.
- Noorani I, Haughey M, Luebeck J, Rowan A, Grönroos E, Terenzi F, et al. Extrachromosomal DNA driven oncogene spatial heterogeneity and evolution in glioblastoma. *bioRxiv.* 2024. Available from: <https://doi.org/10.1101/2024.10.22.619657>.
- Chapman OS, Luebeck J, Sridhar S, Wong IT-L, Dixit D, Wang S, et al. Circular extrachromosomal DNA promotes tumor heterogeneity in high-risk medulloblastoma. *Nat Genet.* 2023;55:2189–99.
- Bailey C, Pich O, Thol K, Watkins TBK, Luebeck J, Rowan A, et al. Origins and impact of extrachromosomal DNA. *Nature.* 2024;635:193–200.
- Chen JP, Diekmann C, Wu H, Chen C, Della Chiara G, Berrino E, et al. Scircle-seq unveils the diversity and complexity of extrachromosomal circular DNAs in single cells. *Nat Commun.* 2024;15:1768.
- Chamorro González R, Conrad T, Stöber MC, Xu R, Giurgiu M, Rodriguez-Fos E, et al. Parallel sequencing of extrachromosomal circular DNAs and transcriptomes in single cancer cells. *Nat Genet.* 2023;55:880–90.
- Huang X, Huang Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics.* 2021;37:4569–71.
- Fleming SJ, Chaffin MD, Arduini A, Akkad A-D, Banks E, Marioni JC, et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using cell bender. *Nat Methods.* 2023;20:1323–35.
- Yi E, Gujar AD, Guthrie M, Kim H, Zhao D, Johnson KC, et al. Live-cell imaging shows uneven segregation of extrachromosomal DNA elements and transcriptionally active extrachromosomal DNA hubs in cancer. *Cancer Discov.* 2022;12:468–83.
- Hung KL, Yost KE, Xie L, Shi Q, Helmsauer K, Luebeck J, et al. Ecdna hubs drive cooperative intermolecular oncogene expression. *Nature.* 2021;600:731–6.
- Tang J, Weiser NE, Wang G, Chowdhry S, Curtis EJ, Zhao Y, et al. Enhancing transcription-replication conflict targets ecDNA-positive cancers. *Nature.* 2024;635:210–8.
- Barthel FP, Johnson KC, Varn FS, Moskalik AD, Tanner G, Kocakavuk E, et al. Longitudinal molecular trajectories of diffuse glioma in adults. *Nature.* 2019. <https://doi.org/10.1038/s41586-019-1775-1>.
- Deshpande V, Luebeck J, Nguyen N-PD, Bakhtiari M, Turner KM, Schwab R, et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat Commun.* 2019;10:392.
- Riemondy KA, Venkataraman S, Willard N, Nellan A, Sanford B, Griesinger AM, et al. Neoplastic and immune single-cell transcriptomics define subgroup-specific intra-tumoral heterogeneity of childhood medulloblastoma. *Neuro Oncol.* 2022;24:273–86.
- Gold MP, Ong W, Masteller AM, Ghasemi DR, Galindo JA, Park NR, et al. Developmental basis of SHH medulloblastoma heterogeneity. *Nat Commun.* 2024;15:1–20.
- Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep.* 2019;9:5233.
- Bhaduri A, Di Lullo E, Jung D, Müller S, Crouch EE, Espinosa CS, et al. Outer radial glia-like cancer stem cells contribute to heterogeneity of glioblastoma. *Cell Stem Cell.* 2020;26:48–63.e6.
- Hendriksen JD, Locallo A, Maarup S, Debnath O, Ishaque N, Hasselbach B, et al. Immunotherapy drives mesenchymal tumor cell state shift and TME immune response in glioblastoma patients. *Neuro Oncol.* 2024;26:1453–66.
- Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell.* 2019;178:835–49.
- Luebeck J, Ng AWT, Galipeau PC, Li X, Sanchez CA, Katz-Sommercorn AC, et al. Extrachromosomal DNA in the cancerous transformation of Barrett's oesophagus. *Nature.* 2023;616:798–805.
- Kim H, Kim S, Wade T, Yeo E, Lipsa A, Golebiewska A, et al. Mapping extrachromosomal DNA amplifications during cancer progression. *Nat Genet.* 2024;56:2447–54.
- Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell.* 2010;17:98–110.
- Sakharkar MK, Chow VTK, Kanguene P. Distributions of exons and introns in the human genome. *In Silico Biol.* 2004;4:387–93.
- Friend C, Marovitz W, Henie G, Henie W, Tsuei D, Hirschhorn K, et al. Observations on cell lines derived from a patient with Hodgkin's disease. *Cancer Res.* 1978;38:2581–91.
- Kucharzewska P, Belting M. Emerging roles of extracellular vesicles in the adaptive response of tumour cells to microenvironmental stress. *J Extracell Vesicles.* 2013;2:20304.
- Nørøxe DS, Maarup S, Fougner V, Muhic A, Møller S, Urup T, et al. The neurogenome study: comprehensive molecular profiling to optimize treatment for Danish glioblastoma patients. *Neuro-Oncol Adv.* 2023;5:vdad137.

33. Batiuk MY, Tyler T, Mei S, Rydbirk R, Petukhov V, Sedmak D, et al. Selective vulnerability of supragranular layer neurons in schizophrenia. *bioRxiv*. 2021 [cited 2022 Jun 13]. p. 2020.11.17.386458. Available from: <https://www.biorxiv.org/content/10.1101/2020.11.17.386458>.
34. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184:3573–87.e29.
35. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat Biotechnol*. 2021. <https://doi.org/10.1038/s41587-020-00795-2>.
36. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019;20:163–72.
37. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.
38. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. Stringtie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015;33:290–5.
39. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
40. Li H, Durbin R. Fast and accurate short read alignment with Burrows-wheeler transform. *Bioinformatics*. 2009;25:1754–60.
41. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, Inc.; 2020.
42. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*. 2018;34:867–8.
43. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32:3047–8.
44. Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, et al. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med*. 2020;12:1–9.
45. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol*. 2015;26:64–70.
46. Hendriksen JD, Locallo A, Schlotmann BC, Rodríguez González FG, Skjøth-Rasmussen J, Yde CW, Nørøxe DS, Skovgaard Poulsen H, Weischenfeldt J. ecSingle. Bitbucket. 2025. Available from: <https://bitbucket.org/weischenfeldt/ecsingl>.
47. Hendriksen JD, Locallo A, Schlotmann BC, Rodríguez González FG, Skjøth-Rasmussen J, Yde CW, Nørøxe DS, Skovgaard Poulsen H, Weischenfeldt J. ecSingle. Zenodo. 2025. Available from: <https://doi.org/10.5281/zenodo.17790701>.
48. Hendriksen JD, Locallo A, Maarup S, Debnath O, Ishaque N, Hasselbach B, Skjøth-Rasmussen J, Yde CW, Poulsen HS, Lassen U, Weischenfeldt J. EGAD00001010843. Datasets. European Genome Phenome Archive. 2024. Available from: <https://ega-archive.org/datasets/EGAD00001010843>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.