# De novo detection of somatic variants in long-read single-cell RNA sequencing data

**Arthur Dondi[1,2], Nico Borgsmüller[1,2], Pedro F. Ferreira[1,2], Brian J. Haas[3], Francis Jacob[4], Viola Heinzelmann-Schwarz[4], Tumor Profiler Consortium, Niko Beerenwinkel[1,2,*]**

* Corresponding author

[1] ETH Zurich, Department of Biosystems Science and Engineering, Schanzenstrasse 44, 4056 Basel, Switzerland

[2] SIB Swiss Institute of Bioinformatics, Schanzenstrasse 44, 4056 Basel, Switzerland

[3] Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA

[4] University Hospital Basel and University of Basel, Ovarian Cancer Research, Department of Biomedicine, Hebelstrasse 20, 4031 Basel, Switzerland

# Abstract

In cancer, genetic and transcriptomic variations generate clonal heterogeneity, possibly leading to treatment resistance. Long-read single-cell RNA sequencing (LR scRNA-seq) has the potential to detect genetic and transcriptomic variations simultaneously. Here, we present LongSom, a computational workflow leveraging LR scRNA-seq data to call de novo somatic single-nucleotide variants (SNVs), copy-number alterations (CNAs), and gene fusions to reconstruct the tumor clonal heterogeneity. For SNV calling, LongSom distinguishes somatic SNVs from germline polymorphisms by reannotating marker gene expression-based cell types using called variants and applying strict filters. Applying LongSom to ovarian cancer samples, we detected clinically relevant somatic SNVs that were validated against single-cell and bulk panel DNA-seq data and could not be detected with short-read (SR) scRNA-seq. Leveraging somatic SNVs and fusions, LongSom found subclones with different predicted treatment outcomes. In summary, LongSom enables de novo SNVs, CNAs, and fusions detection, thus enabling the study of cancer evolution, clonal heterogeneity, and treatment resistance.

# Introduction

Cancer cells accumulate genomic variations, such as single-nucleotide variants (SNVs), copy number alterations (CNAs), and gene fusions during their lifetime, leading to subpopulations with distinct genotypes. Together with changes in the tumor microenvironment, genomic variations result in distinct phenotypes, such as expression patterns (Lappalainen et al. 2013). Intratumor heterogeneity, i.e., the existence of cancer subpopulations with distinct genotypes and phenotypes, is presumed to be a leading cause of therapy resistance and one of the main reasons for poor overall survival in cancer patients with metastatic disease (Jamal-Hanjani et al. 2015; Ramón Y Cajal et al. 2020). The adaptive mechanisms underlying therapy resistance are of both genetic (SNVs, CNAs, gene fusions, etc.) and non-genetic (epigenetic, transcriptomic, microenvironment, etc.) origin. The first step to identifying therapy-resistant subclones is to capture those genetic and transcriptomic variants through sequencing (Mansoori et al. 2017; Marine et al. 2020). Unraveling different subpopulations is particularly challenging with bulk techniques; however, the advent of single-cell sequencing technologies has significantly improved our ability to decipher intratumor heterogeneity within complex tissue samples (Dagogo-Jack and Shaw 2018).

In scDNA-seq data, cancer cell subpopulations are inferred from SNVs and CNAs, which are conventionally obtained from exome or whole-genome sequencing approaches (Roth et al. 2016; Duan et al. 2018). In scRNA-seq, gene expression patterns are commonly used to differentiate between cell types or cancer cell subpopulations. However, relying solely on gene-level expression may be insufficient, as cells can express different isoforms, resulting in different phenotypes (Ding et al. 2020). Isoform-specific cancer resistance can be induced, for example, through alternative splicing (Mitra et al. 2009; Chen et al. 2022), polyadenylation (Guo et al. 2022), or large genomic rearrangements leading to gene fusions (Amatu et al. 2016; Lei et al. 2018; (Cesi et al. 2018). These interlinked features need to be examined together, thus requiring complete isoform coverage (Foord et al. 2023). High-throughput

58 droplet-based scRNA-seq protocols (10X Genomics Chromium) capture reads via their 3'
59 polyA tails. In short-read (SR) scRNA-seq, this results in a heavy coverage bias towards the
60 3' ends as only a few hundred base pairs of each molecule are sequenced. Long-read (LR)
61 scRNA-seq, in contrast, sequences full-length RNA molecules, and thus can access gene
62 expression at the isoform level (Joglekar et al. 2021; Al'Khafaji et al. 2023; Dondi et al. 2023).

63 Linking genetic to transcriptomic variations is crucial to understanding treatment resistance in
64 cancer (Vasan et al. 2019). However, this is challenging with SR sequencing, as genetic
65 variations are difficult to recover from SR scRNA-seq data due to capture bias, while scDNA-
66 seq cannot assess gene expression. Recently, DNA-free de novo scRNA SNV (Muyas et al.
67 2023; Zhang et al. 2023) and CNA (Serin Harmanci et al. 2020); (Gao et al. 2021, 2023) calling
68 methods were developed for SR sequencing, compensating the 3' capture bias by pooling
69 large amounts of cells or sequencing at very high read depths per cell. However, SR
70 sequencing is unsuited to detect isoforms or gene fusions. Because it is less sensitive to
71 capture bias, we have shown in recent work that LR scRNA-seq is more suited to detect
72 genetic variations than SR scRNA-seq (Dondi et al. 2023). Furthermore, LR scRNA-seq can
73 simultaneously detect SNVs, CNAs, fusions, and gene isoform expression in the same cells
74 (Dondi et al. 2023; Shiau et al. 2023).

75 In this study, we present LongSom, a computational workflow for calling de novo somatic
76 SNVs, fusions, and CNAs in LR scRNA-seq, and integrating them to reconstruct the samples'
77 clonal heterogeneity. Applied to omentum metastasis samples obtained from three chemo-
78 naive high-grade serous ovarian cancer (HGSOC) patients, we show that LongSom can detect
79 clinically relevant somatic SNVs validated against scDNA and panel data, whereas SR
80 scRNA-seq fails to do so. We demonstrate that by leveraging somatic SNVs and fusions,
81 LongSom can detect subclones with different predicted treatment outcomes, and those
82 subclones were highly concordant with gene expression clusters and CNAs subclones.
83 Additionally, we find that tumor microenvironment cells are contaminated by cancer cell-
84 derived mitochondria.

# Results

## Overview of LongSom

We developed LongSom, a workflow for detecting genetic variants in LR scRNA-seq data without requiring matched DNA sequencing and finding cancer subclones based on these. Briefly, LongSom takes BAM files as input, calls SNVs in pseudo-bulk and fusions and CNAs in single cells with the Trinity Cancer Transcriptome Analysis Toolkit (CTAT, https://github.com/NCIP/Trinity_CTAT), and then reconstructs the clonal heterogeneity using the Bayesian non-parametric clustering method BnpC (Borgsmüller et al. 2020).

LongSom first calls candidate SNV loci in a pseudo-bulk generated by aggregating LR scRNA-seq data from all cells, using CTAT-Mutations (https://github.com/NCIP/ctat-mutations), which we enhanced here for scRNA-seq and long isoform reads (see **Methods**). Next, to distinguish between somatic and germline variants, the variant allele frequency (VAF) is calculated for each candidate locus and each cell, and cells are grouped into cancer or non-cancer cells based on marker-gene expression. SNVs detected across multiple cell types are considered germline polymorphisms. Accordingly, if cancer cells are misannotated as non-cancer cells, SNVs will wrongly be filtered out as germline variants (false negatives). To account for this, LongSom first defines a set of cancer-specific variants (SNVs and fusions). SNVs are defined as cancer-specific if their VAF is high in cancer, low in non-cancer, and, when available, zero in normal sample cells (**Methods**). Fusions are detected using CTAT-LR-fusion (https://github.com/TrinityCTAT/CTAT-LR-fusion) (Qin et al. 2024) and cancer-specific fusions are those expressed in more than 5% of cancer cells and less than 1% of non-cancer cells. LongSom reannotates cells as cancer cells if they harbor at least two cancer-specific variants (**Figure 1, Methods**).

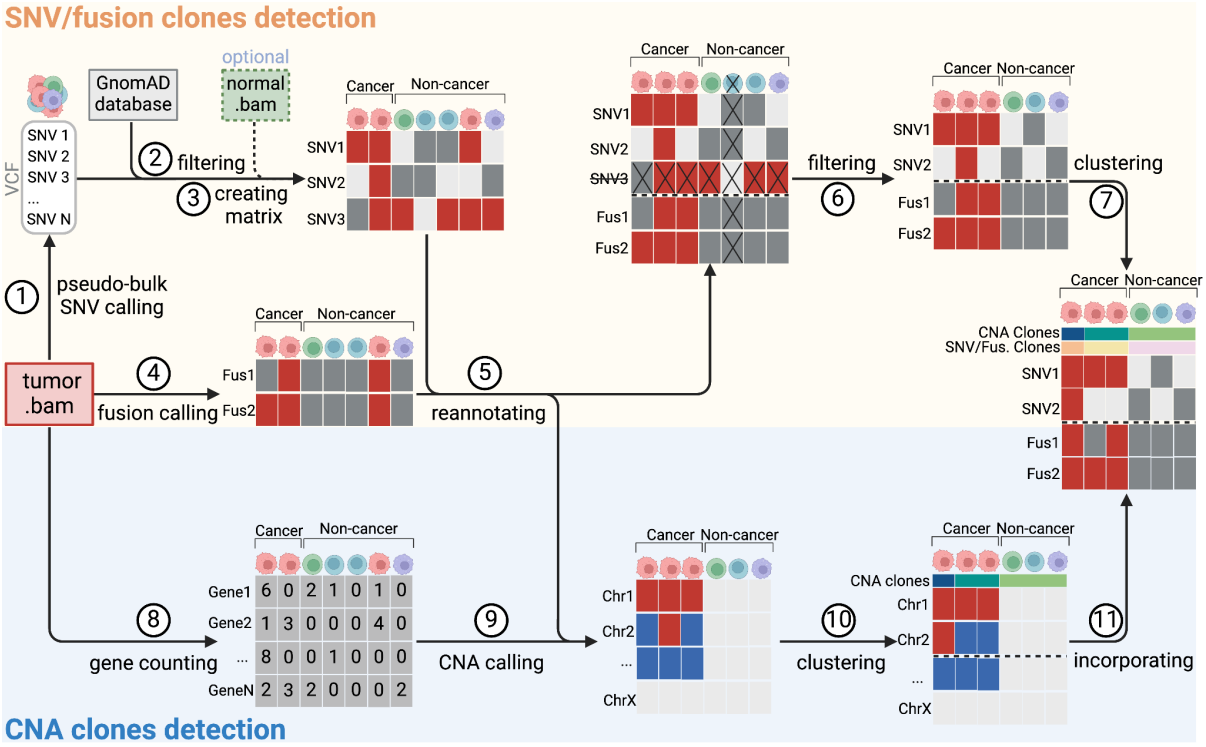


**Figure 1: Overview of LongSom.**

LongSom's methodology for detecting somatic SNVs, fusions, and CNAs and subsequently inferring cancer subclones in LR scRNA-seq individual patients data. (1) SNV candidates are detected from pseudo-bulk samples. (2) Population germline SNVs and SNVs present in normal samples (optional) are filtered out. (3) A cell-SNV matrix based on the remaining SNV candidates is computed. (4) A cell-fusion matrix is computed. (5) Using high-confidence cancer fusions and SNVs, cells are reannotated. (6) Following reannotation, SNVs present in non-cancer cells (germlines) are filtered out. (7) cells are clustered based on somatic fusions and SNVs. In parallel, (8) gene expression per cell is computed, (9) CNAs are detected, (10) cells are clustered based on CNAs, and (11) CNA clones are incorporated to the fusions and SNVs clustered matrix.

After cell reannotation, LongSom performs germline SNV filtering in five steps: (A) It filters SNV loci detected in the matched normal, when available. (B) It filters SNV loci from the gnomAD database (Chen et al. 2024) with a frequency of at least 0.01% in the total population. (C) After cell-type reannotation, it filters SNV loci that were called in more than 1% of the non-cancer cells. (D) SNV loci where less than 1% of the non-cancer cells are covered by at least one read are filtered. This step helps to filter germline SNVs not detected due to low

127 expression in non-cancer cells. (E) Finally, adjacent SNV loci within a 10,000 bp distance are
128 filtered, as these are likely to be misalignment artifacts in low-complexity regions. Of note,
129 steps (C) and (E) are not applied to mitochondrial SNVs. Finally, LongSom keeps somatic loci
130 that are mutated in a minimum of five cancer cells or 5% of cancer cells (user-defined
131 parameters) and filters loci matching known RNA-editing sites.
132
133 Finally, LongSom infers the clonal structure of the samples using two different approaches.
134 One approach leverages the detected SNVs and fusions as input for the Bayesian non-
135 parametric clustering method BnpC (Borgsmüller et al. 2020). The other approach predicts
136 CNAs based on gene expression in cancer cells and defines subclusters using inferCNV
137 (https://github.com/broadinstitute/infercnv) (**Methods**).

## 138 Cell-type reannotation improves somatic SNV detection sensitivity

139 We applied LongSom to previously published (Dondi et al. 2023) SR and LR scRNA-seq data
140 of five omentum metastasis samples obtained from three chemo-naive HGSOC patients: P1,
141 P2, and P3. Three samples were derived from HGSOC omental metastases and two from
142 matching distal tumor-free omental tissues (normal). After cell-type reannotation (**Methods**),
143 the reannotated cells were always more similar to the expression-based clustering (Jaccard
144 similarity score in patient P1: 0.97, P2: 0.99, P3: 0.97) than the previous annotation derived
145 from marker-gene expression (Jaccard similarity score in patient P1: 0.94, P2: 0.98, P3: 0.76),
146 supporting the reannotation (**Figure 2a**). We found that 6, 2, and 21% of the cells that we
147 annotated as cancer were previously annotated as non-cancer cells in the tumor biopsy
148 samples of patients P1, P2, and P3, respectively (**Figure 2b**). The tumor biopsy of patient P3
149 had only 10% cancer cells (Dondi et al. 2023), which could explain the high level of cell
150 misannotation. In the following, cancer or non-cancer cells refer to the reannotated cell types.
151
152 After cell-type reannotation and germline polymorphism filtering, we found 32, 50, and 62
153 somatic SNVs and 4, 16, and 2 somatic fusions in patients P1, P2, and P3, respectively

154 (**Supplementary Tables 1, 2**). In patient P1, a variant at locus chr21:8455886 was manually
155 detected as a technical false positive due to mismapping in a highly repetitive region, and it
156 was excluded from further analyses. Without cell type reannotation, we only found 23/32
157 (72%), 48/50 (96%), and 19/62 (31%) of those SNVs, in P1, P2, and P3, respectively, and no
158 additional SNV was discovered (**Figure 2c)**. In patient P3, numerous cancer cells were
159 misannotated as non-cancer cells before reannotation (**Figure 2a,b**), leading to 69% of false
160 negative somatic SNVs during germline SNV filtering (**Figure 2c)**. Cells reannotated from non-
161 cancer to cancer cells showed a mean VAF across somatic SNVs significantly different from
162 cells annotated as non-cancer cells in both methods ($P<0.001$ in all patients, two-tailed two-
163 sample t-test), but not from cells annotated as cancer in both methods ($P>0.05$ in all patients),
164 thus further supporting the cell-type reannotation (**Figure 2d**). Out of the 144 somatic SNVs
165 identified, we found 32.6% of variants in or affecting coding regions (2.1% start or stop codon
166 gain (n=3), 1.4% splice region (n=2), 22.2% missense (n=32), and 6.9% synonymous variants
167 (n=10)) and 67.4% in non-coding regions (17.4% 3' or 5' UTR (n=25), 40.3% intron (n=48) and
168 9.7% intergenic variants (n=14)) (**Figure 2e**).
169

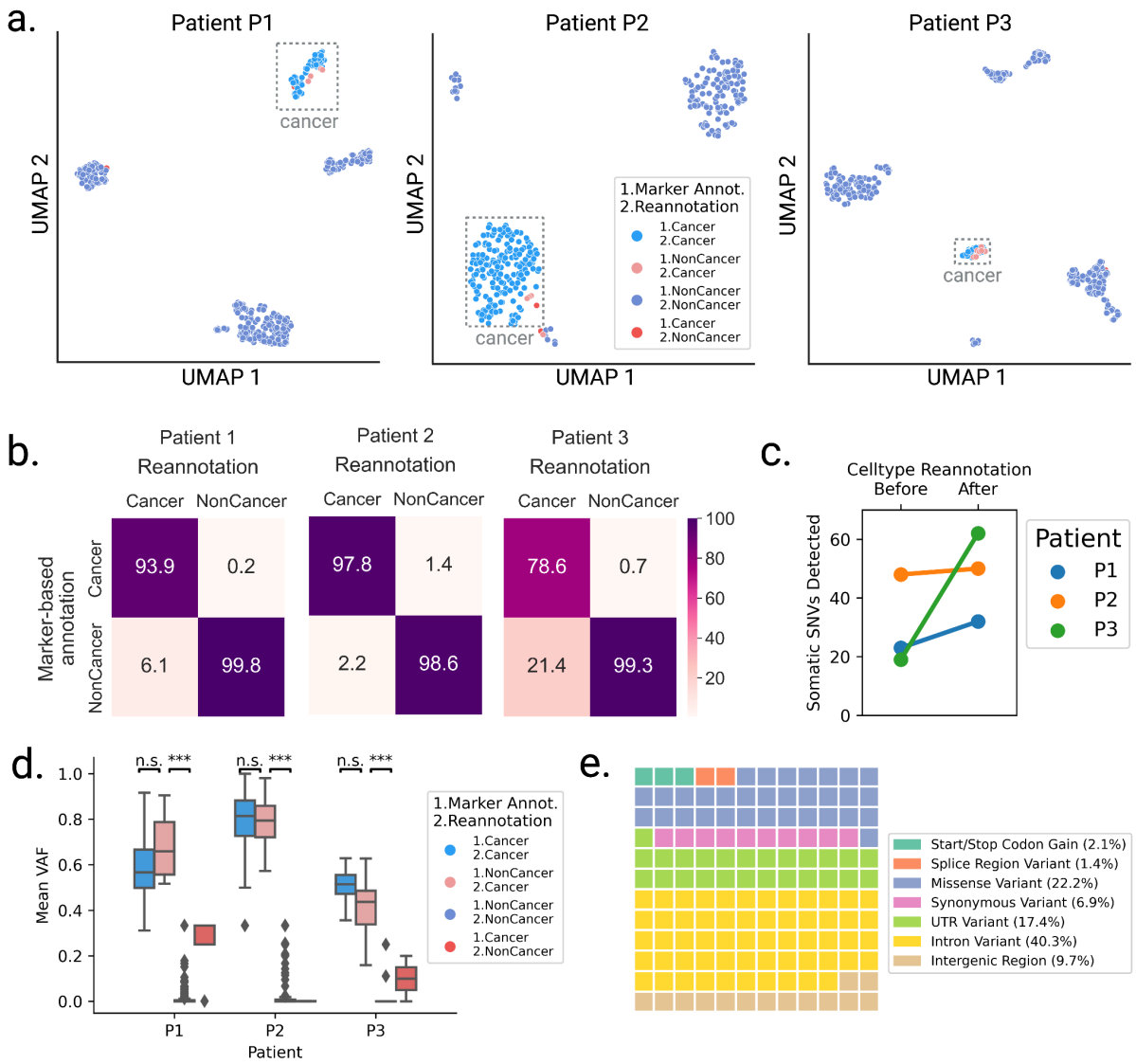


**Figure 2: Cell-type reannotation improves somatic SNVs detection sensitivity**

**a.** UMAP embeddings of LR scRNA-seq expression per patient. Cells are colored by annotation status; light-red cells show cells predicted as non-cancer using marker gene expression based annotation, and cancer using high-confidence somatic variants reannotation **b.** Confusion matrices of cells predicted as cancer or non-cancer using marker genes, and cells reannotated as cancer or non-cancer by LongSom, colored and annotated by the percentage of the total number of cells in each category. E.g. the bottom left square represents cells previously annotated as non-cancer that were reannotated as cancer (false negative cancer cells). **c.** Number of SNVs found per patient, with or without cell type reannotation before filtering germline SNPs. **d.** Boxplots of the mean VAF per covered SNV loci of each cell, per patient, colored by their annotation status. Boxes display the first to third quartile with median as horizontal line, whiskers encompass 1.5 times the interquartile range, and data beyond that threshold

182 is indicated as outliers. P values were calculated using a two-sided Student's t-test between groups

183 and are described with the following symbols: n.s : P > 0.05, *: P ≤ 0.05, **: P ≤ 0.01, ***: P ≤ 0.001. **e.**

184 Waffle plot representing each somatic SNV detected, colored by their genomic region and effect on the

185 coding sequence.

## 186 Validation of LR scRNA-seq-derived SNVs using scDNA-seq data

187 For validation of the SNVs detected using LongSom, we employed scDNA-seq data from

188 matched omental metastases for each patient. First, we inferred the cellular copy number

189 profiles based on the scDNA-seq data and identified fully diploid subclones (likely non-cancer)

190 and aneuploid subclones (likely cancer) from these (Kuipers et al. 2020) (**Methods**). We found

191 two aneuploid clones in patient P1, one in P2, and two in P3 (**Figure 3a-c**). In each somatic

192 locus detected by LongSom, we estimated the mean VAF of diploid and aneuploid scDNA

193 subclones by generating pseudo-bulks. We assumed that a scDNA subclone supported an

194 SNV if the mean VAF was greater than 10% at the respective locus.

195

196 Overall, 55% (n=79) of the somatic SNVs detected in LR scRNA were found exclusively in

197 scDNA aneuploid subclones and were therefore likely somatic (**Figure 3d-f**). In all cases

198 where SNVs were not detected in scDNA aneuploid subclones, the scDNA-seq coverage was

199 <10x (**Figure 3g**). The 10% (n=15) of SNVs detected in diploid scDNA subclones (suggesting

200 germline polymorphism) were all in patient P2 except one in patient P1 which was only

201 supported by one read (**Figure 3h**). No normal LR scRNA-seq sample was available for

202 Patient P2, and non-cancer cells were mainly T-cells with an overall low gene

203 expression(Joglekar et al. 2021), possibly explaining why germline SNVs were insufficiently

204 filtered. This finding highlights the utility of matched normal samples to filter germlines

205 sufficiently.

206

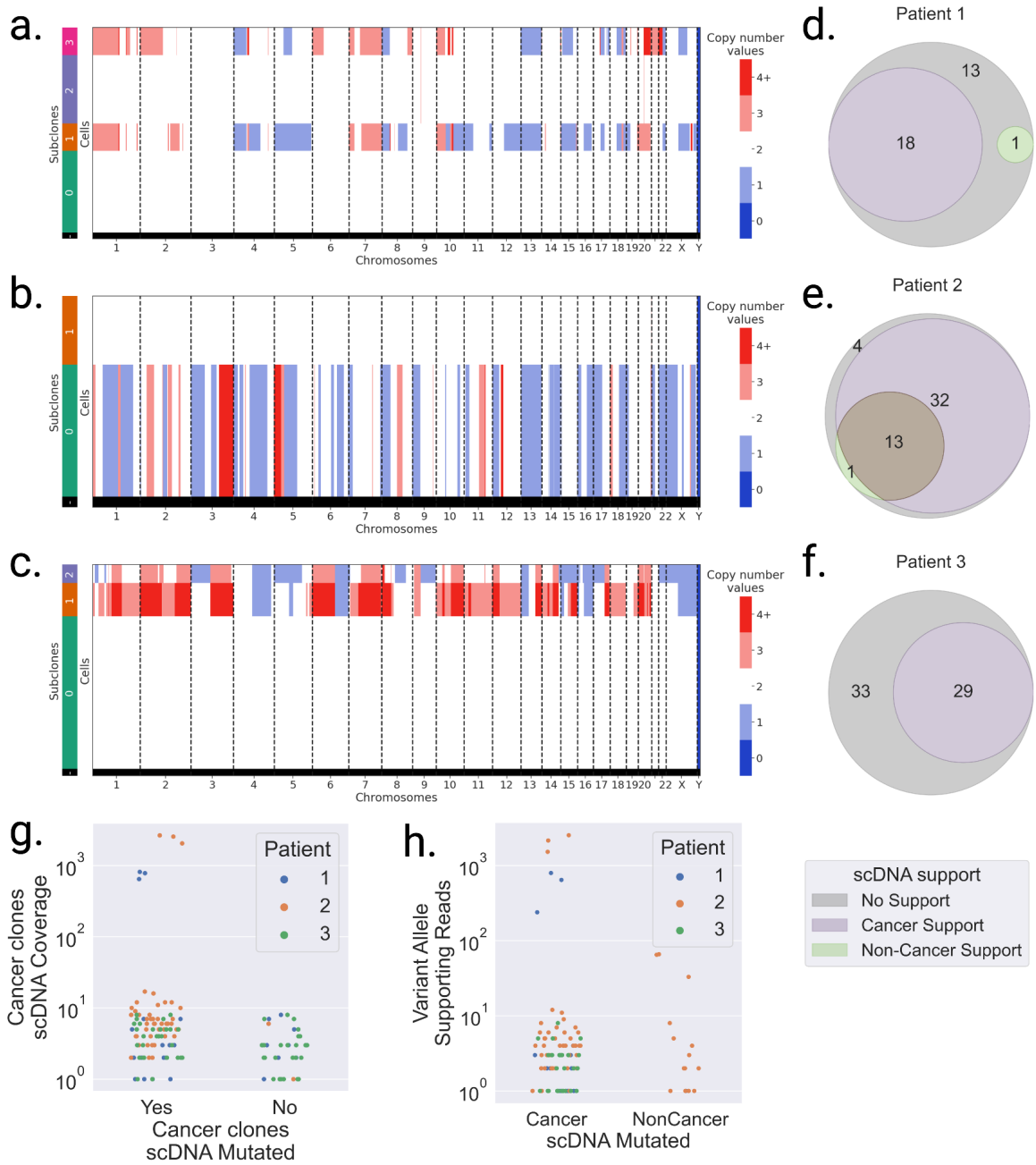


**Figure 3: Somatic SNVs detected in scRNA are validated as somatic in scDNA**.

**a,b,c.** scDNA-seq copy number values per subclone in **a.** patient 1, **b.** patient 2, and **c.** patient 3 data. Subclones with multiple copy number alterations are aneuploid (likely cancer), while copy number neutral subclones are diploid (likely non-cancer). **d,e,f** Venn diagrams of somatic SNVs supported (VAF>10%) in scDNA cancer subclones (purple), scDNA non-cancer subclones (green), and both (brown). **g.** scDNA cancer subclones coverage per somatic locus identified in scRNA, categorized by whether they are found mutated in cancer subclones (Yes) or not (No). **h.** Number of mutated reads in

215 scDNA subclones per somatic loci identified in LR scRNA, categorized by cancer and non-cancer
216 scDNA subclones.

## 217 Somatic mitochondrial reads contaminate tumor microenvironment cells
## 218 in scRNA-seq and scDNA-seq data

219 As somatic mitochondrial SNVs can also be used for genotype and clonal reconstruction
220 (Miller et al. 2022), LongSom also detects them. Due to the high mitochondrial RNA
221 expression in cells(Osorio and Cai 2021), somatic mitochondrial SNVs (mtSNVs) were
222 amongst the most covered across all cell types and patients in the HGSOC dataset. We found
223 three somatic mitochondrial SNVs in patient P1 (chrM:3092:T>C, chrM:5179:T>C,
224 chrM:16192:C>T), three in patient P2 (chrM:2573:G>A, chrM:4308:G>A, chrM:16065:G>A),
225 and none in patient P3 (**Supplementary Table 1**).
226
227 In patient P1, at locus chrM:3092, all covered cancer cells exhibited a >99% VAF in scRNA
228 data, while non-cancer cells showed heteroplasmy (VAF ranging between 0-40%, with 28%
229 of cells mutated, median VAF when mutated 4%) (**Figure 4a)**. However, all cells from distal
230 samples exhibited a VAF <1% (>99% cells covered), ruling out a germline SNV. We detected
231 the same mutational profile in matching scDNA-seq data: amongst the diploid subclones, the
232 average VAF was 5%, while the average VAF in aneuploid subclones was >99% (**Figure 3a**).
233 At locus chrM:5179, cancer cells were either mutated (n = 44, median VAF = 49.7%) or not (n
234 = 30, median VAF <0.1%) in scRNA-seq data, suggesting the presence of two subclones. In
235 non-cancer cells, the VAF ranged from 0 to 33% (11% cells mutated, median VAF when
236 mutated 3%), and all cells from distal samples exhibited again a VAF <1% (>99% cells
237 covered, **Figure 4b**). In the matched scDNA-seq data, at locus chrM:5179, the VAF of
238 aneuploid subclone C3 (**Figure 3a**) was 42%, while it was <1% in the second aneuploid
239 subclone C1 and 2% in the diploid subclones, confirming the subclone specificity of this
240 somatic SNV (**Figure 4c**).

241 In Patient P2, locus chrM:2573 showed the same pattern with a mean VAF of 75% in cancer
242 cells compared to a mean VAF of 2% in non-cancer cells. This SNV was observed in 20% of
243 the non-cancer cells, with a mean VAF of 5% (SD=3.6)) when mutated (**Figure 4d**). No
244 matching normal sample was available for this patient. In scDNA, the aneuploid subclone had
245 a VAF of 19% while the diploid subclone had a VAF of 81%. At locus chrM:4308, cancer cells
246 had a mean VAF of 97%, and non-cancer cells had a VAF ranging between 0 and 100% (19%
247 cells mutated, mean VAF when mutated 88% (SD=25)). All cells mutated at this locus had
248 only one variant allele read (**Figure 4e**).
249
250 In summary, we observed mitochondrial SNVs with high VAF in cancer cells and lower (but
251 non-zero) VAF in non-cancer cells from the same scRNA-seq and scDNA-seq samples.
252 Remarkably, cells from distal (normal) scRNA samples had a VAF of zero in those loci,
253 suggesting that the mutated mitochondrial reads found in non-cancer cells originate from
254 cancer cells. This phenomenon could be explained via biological mechanisms such as
255 intercellular mitochondrial transfer (Liu et al. 2021), or via technical contaminations such as
256 mitochondria from dead cancer cells being captured together with non-cancer cells during
257 single-cell encapsulation. The correlation between the amount of mutated mitochondrial SNV
258 reads found in cancer and in non-cancer supports the contamination hypothesis (**Figure 4f**).
259 To account for those contaminations, LongSom does not apply step (E) of germline filtering
260 (filtering of loci that were called in more than 1% of the non-cancer cells).
261

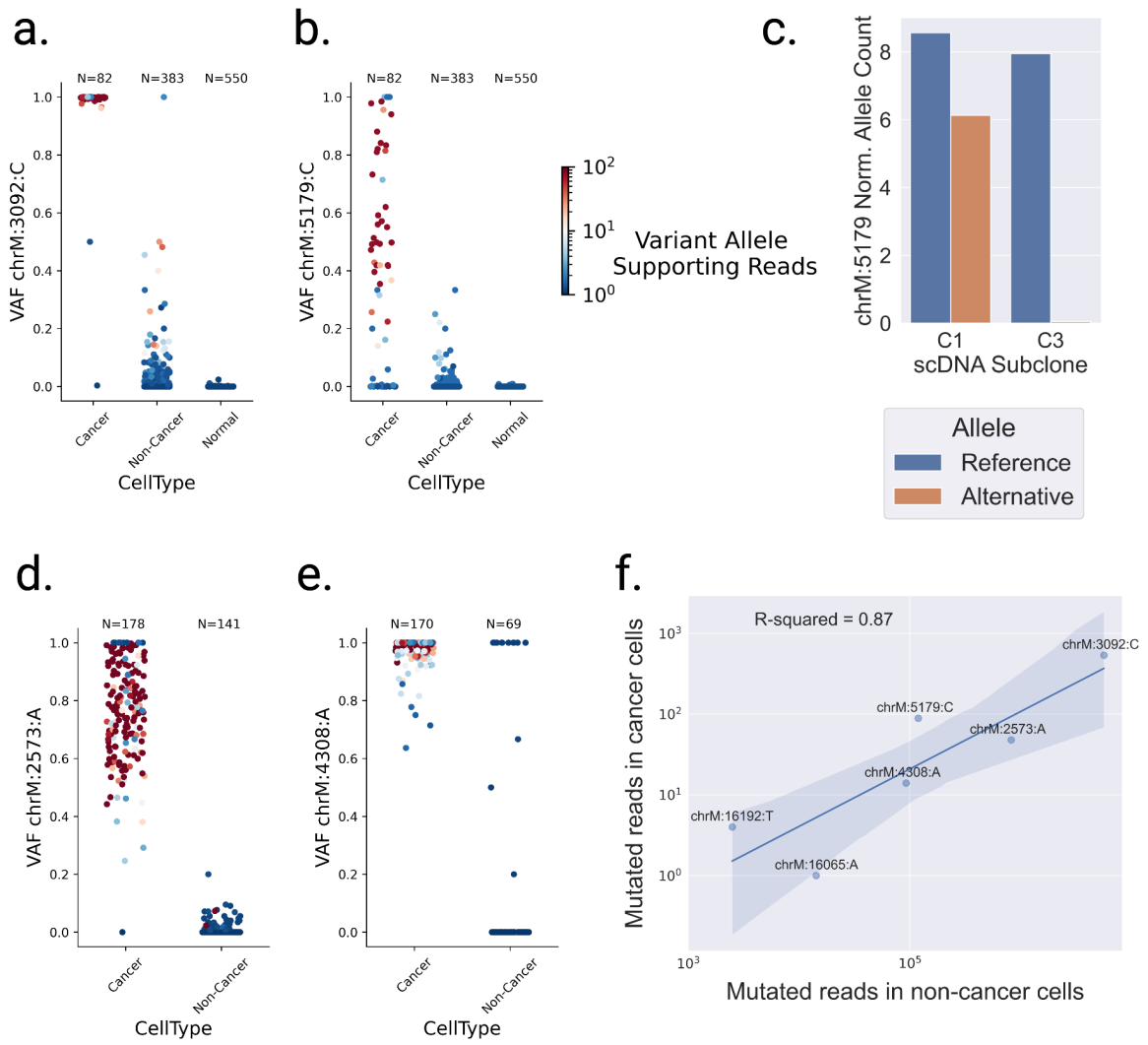


**Figure 4: Non-cancer cells are contaminated by cancer cells mitochondria.**

**a,b,** VAF of cells in patient P1 at loci **a.** chrM:3092:C and **b.** chrM:5179:C, categorized by reannotated cell types. Color gradient represents the number of variant allele reads per cell. Cells with more than 100 mutated reads are represented with 100 mutated reads. N refers to the number of cells with at least one read covering the locus. **c.** Number of reads supporting the reference or alternative allele in patient P1's scDNA aneuploid (cancer) subclones C1 and C3 at locus chrM:5179, normalized by the number of cells per subclone. **d,e,** VAF of cells in patient P2 at loci **d.** chrM:2573:C and **e.** chrM:4308:C. **f.** Log aggregated mutated reads in non-cancer cells, as a function of log aggregated mutated reads in cancer cells for loci chrM:3092:C, chrM:5179:C and chrM:16192:T in patient P1, and chrM:2573:C, chrM:4308:C and chrM:16065:A in patient 2.

## LR scRNA-seq enables somatic SNV detection with higher sensitivity than SR scRNA-seq

Next, we aimed to compare LR to SR scRNA-seq to detect SNVs. The HGSOC study had LR and SR scRNA-seq data from the same cells available, and while the SR dataset had 4.3 times more sequenced reads compared to LR (mean 117.4k vs. 26.9k reads per cell), it had 3.5 times fewer mapped bases (mean 16Gb mapped vs. 33Gb mapped) due to shorter read length (**Supplementary Figure 1a,b**). When we applied LongSom to SR scRNA-seq, we found only 4/32 (13%), 9/50 (18%), and 9/62 (15%) somatic SNVs in patients P1, P2, and P3 respectively, and no new SNV was detected (**Supplementary Figure 1c**). Additionally, only 1/4 (25%), 9/17 (53%), and 1/2 (50%) fusions were detected in SR scRNA-seq data from patients P1, P2, and P3, respectively (Qin et al. 2024).

We computed cell-variants sparse matrices from each patient's LR and SR data, using cells as columns and somatic SNVs and fusions as rows (**Methods, Figure 5a-c**). For comparison, we computed the same matrix in each patient using SR scRNA-seq data (**Figure 5d-f**). On average, 13.7% (standard deviation (SD) = 1.2) of the matrix positions had at least one LR coverage, whereas only 4.7% (SD = 1.2) had at least one SR coverage (**Supplementary Figure 1d-f**). However, the coverage depends on the cell type expression, and certain cell types, for example, T cells, rarely express mutated genes (**Figure 5a-f**). In cancer cells, on average 27.9% (SD = 7.5) of the matrix positions were covered by at least one LR, whereas only 8.1% (SD = 0.8) had at least one SR coverage (**Figure 5g-i**). On average, LR covered 94.8% (SD=3) of the positions covered by SR and covered an additional 3.4 times more positions (SD = 0.6), in line with the 3.5 times more bases mapped in LR.

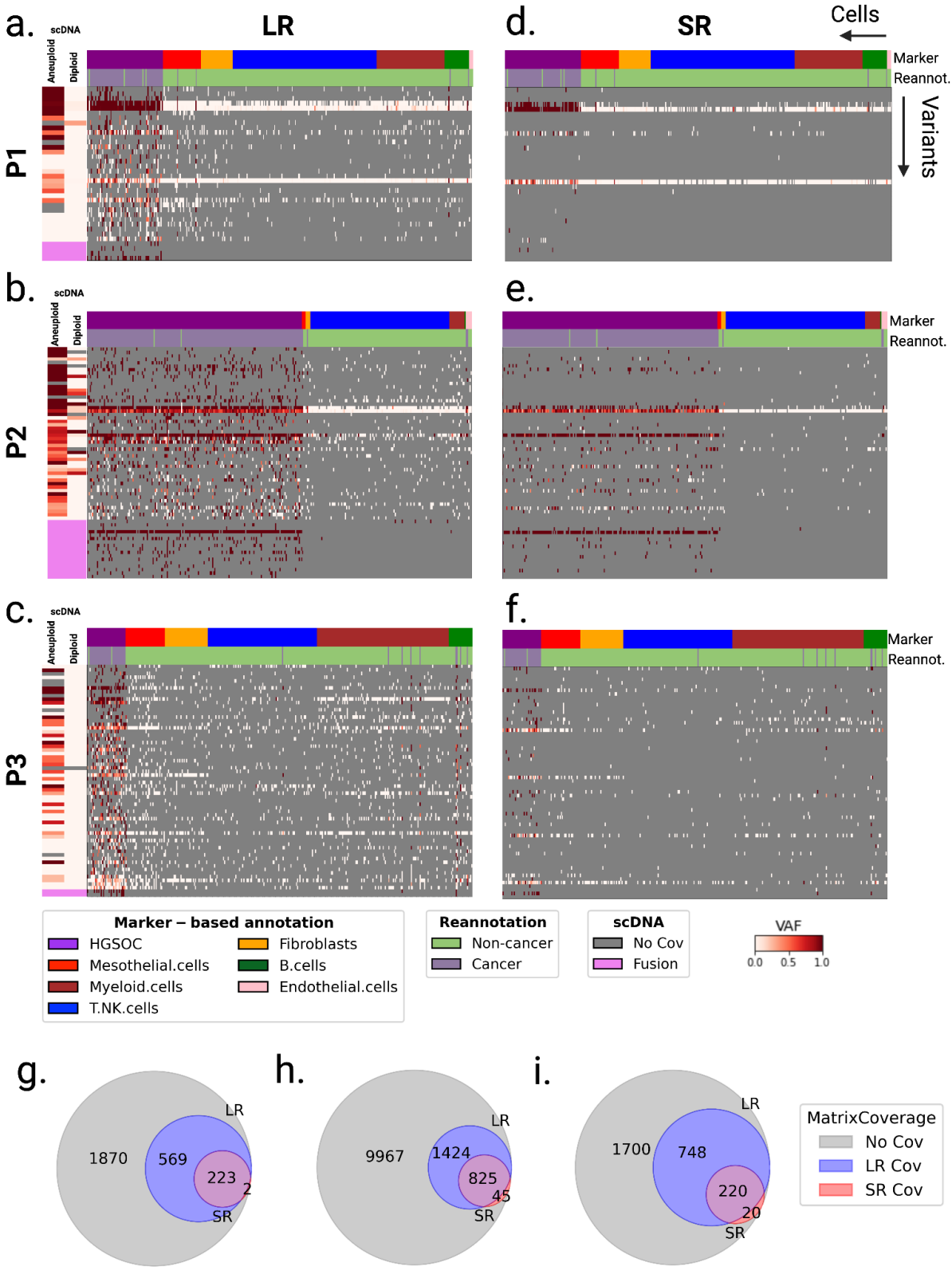


**Figure 5: Patient-specific cell-variant matrices created from LR and SR scRNA-seq.**

**a-f.** Matrices of somatic SNVs and fusions (rows) by single cells (columns) computed using LR sc-RNA-seq from the tumor biopsy of **(a)** patient P1, **(b)** P2 and **(c)** P3, and using SR sc-RNA-seq of **(d)** patient P1, **(e)** P2, and **(f)** P3, ordered by gene expression-derived cell types. VAF is depicted as a gradient

302 from white (no mutated reads, VAF=0) to red (only mutated reads, VAF=1). Grey indicates no coverage
303 in the cell at a given locus. Rows are colored by the scDNA VAF of aggregated diploid and aneuploid
304 cells at the loci: SNVs with high aneuploid VAF and low diploid VAF are somatic in scDNA data. RNA
305 fusions do not give a direct indication of the DNA breakpoint, thus we could not assess their presence
306 in scDNA data, and they appear in pink. Columns are colored by marker-expression-derived cell-types
307 (top row) and cell-types reannotated by LongSom (bottom row) **d,e,f.** Venn diagram of matrices'
308 positions covered in the cancer cells in **(h)** patient P1, (**i)** patient P2, and (**i)** patient P3, colored by
309 sequencing data modality. Total positions equal n variants x m cancer cells. Blue indicates positions
310 with coverage >0 in LR and 0 in SR. Red indicates positions with coverage 0 in LR and coverage >0 in
311 SR. Purple indicates positions with coverage >0 in both LR and SR. Grey indicates positions with
312 coverage 0 in both LR and SR.

## 313 LongSom detects panel-validated resistance-associated variants

314 The three patients also underwent bulk panel DNA sequencing (Methods), where 29 SNVs
315 were found (**Supplementary Table 3**). All three patients had at least one somatic SNV called
316 in *TP53* (including a variant introducing a stop codon in patient P3) with a VAF >30%, and
317 patient P1 had a second *TP53* SNV detected with VAF 1%. LongSom detected all *TP53*
318 somatic SNVs with VAF >30% in LR scRNA-seq. The remaining SNVs detected in the panel
319 were not retained with our method for the following reasons: they were either germline variants
320 found in normal and non-cancer cells (n=5), detected in cancer but with insufficient coverage
321 in non-cancer cells (n=3), detected but not in enough cancer cells (n=7), not detected despite
322 sufficient coverage (n=3), or not covered (n=8) (**Supplementary Table 3**). Overall, 62% of the
323 SNVs detected in the panel also found support in scRNA data. Of note, two deletions were
324 found in panel sequencing, and they were detected manually in the LR scRNA-seq data.
325 LongSom detected none of the panel SNVs in SR scRNA-seq data.
326
327 In addition to *TP53*, LongSom was able to detect SNVs predicted as clinically relevant in genes
328 not included in the bulk panel (Methods). In patient P1, we found missense variants predicted
329 as pathogenic in the apoptosis regulator genes *CCAR2* (Arg722Trp) and *FAM129B*

330 *(*Leu508Pro*)*, and another missense variant in the ferroptosis regulator *ALDH3A2* (Val321Leu)
331 (**Supplementary Table 4**). *ALDH3A2* is a tumor suppressor in multiple cancers (Xia et al.
332 2023) and *ALDH3A2, CCAR2,* and *FAM129B* are all associated with treatment resistance in
333 ovarian cancer (Cheng et al. 2019; Iyer et al. 2022; Dong et al. 2023). In patient P2, the
334 chrM:4308 G>A variant was predicted as likely pathogenic. In patient P3, we detected a
335 missense variant in *AHCY*, as well as a pathogenic *NIF3L1* variant and a missense variant in
336 the resistance-associated gene *KDM6B (He et al. 2019)*. In SR scRNA-seq data, LongSom
337 found no clinically relevant variants in patients P1 and P3, and only found chrM:4308 G>A in
338 patient P2.

## 339 LongSom identifies subclones in LR scRNA-seq data

340 Next, LongSom inferred the clonal structure of the tumors based on the SNVs and fusions it
341 detected using BnpC. LongSom also inferred the clonal structure from CNA profiles in the
342 same cells, using inferCNV (**Supplementary Figure 2a-c, Methods**). We also clustered the
343 cells based on their gene expression, manually annotated the cancer clusters, and used those
344 clusters as transcriptomic validation. Finally, we used the subclones inferred from scDNA as
345 external validation (**Figure 3)**.
346
347 In patient P2, LongSom found one cancer clone using mutations and fusions, and this clone
348 coincided very well with the aneuploid CNA clone found in scRNA (Jaccard similarity = 98%)
349 and the gene-expression-based cancer cluster (Jaccard similarity = 97%, **Supplementary**
350 **Figures 2b, 3a**). Similarly, in scDNA-seq data we only found one aneuploid CNA clone
351 (**Figure 3b**). Therefore, all available data modalities point toward a monoclonal cancer
352 population in this patient. Using SR scRNA-seq, LongSom reconstructed the clonal structure
353 with lower accuracy than LR due to low coverage (Jaccard similarity BnpC clone - cancer
354 cluster = 85%, **Supplementary Figure 3b**).
355

356 In patient P3, LongSom found one clone, coinciding with the scRNA gene expression-based
357 cancer cluster (Jaccard similarity = 93%, **Supplementary Figure 4a**), however, two aneuploid
358 subclones were detected in both scDNA-seq and scRNA-seq data using CNA analysis (**Figure**
359 **2c, Supplementary Figures 2c,4a**). This difference could be due to the difficulty of calling
360 subclones in a low number of cancer cells (n=42 after reannotation) or due to inter-sample
361 heterogeneity. Using SR data, the clustering resulted in a subclone only partially covering the
362 expression-based cancer cluster, and many individual cells formed singleton subclones
363 (Jaccard similarity BnpC subclone - cancer cluster = 36%, **Supplementary Figure 4b**).
364
365 In patient P1, LongSom found two cancer subclones, referred to as A and B, as well as a
366 subclone composed of one cell that we assigned to subclone B (**Figure 6a**). The larger
367 subclone A (n = 40 cells) was predominantly defined by an SNV at locus chrM:5179 and the
368 smaller subclone B (n = 34 cells) was mainly defined by the fusion SMG7--CH507-513H4.1.
369 In expression-based UMAP embedding, cancer cells formed two distinct expression clusters
370 that highly overlapped the genotypic cancer subclones found based on SNVs and fusions
371 (Jaccard similarity BnpC subclone A - Expression cluster 1 = 79%, BnpC subclone B -
372 expression cluster 2 = 72%, **Figure 6 a-d**). CNAs subclones and expression clusters were
373 also very similar (Jaccard similarity inferCNV subclone A - expression cluster 2 = 89%,
374 inferCNV subclone B - expression cluster 2 = 85%), likely because they are both derived from
375 gene expression (**Figure 6a,e, Supplementary Figure 3a).** Clonal assignments based on
376 SNVs and fusions and on CNA data were also similar (Jaccard similarity subclone A = 72%,
377 subclone B = 68%. In patient P1's matched scDNA data, we also found two aneuploid (cancer)
378 subclones based on CNA profiles (**Figure 3a**), and only one of the subclones harbored the
379 SNV chrM:5179:T>C (**Figure 4c**), concordantly with LR scRNA-seq data. Unfortunately, the
380 three other subclone-defining SNVs had insufficient scDNA-seq coverage, and we could not
381 detect any reads supporting the variant allele at those loci, therefore we could not confirm their
382 subclonality. Using SR scRNA-seq, LongSom also identified cancer subclones in patient P1,
383 mainly based on chrM:5179 status. However, as the fusion SMG7--CH507-513H4.1 was not

384 detected in SR, multiple cancer cells clustered with non-cancer cells (Jaccard similarity BnpC
385 SR subclone A - Expression cluster 1 = 70%, BnpC SR subclone B - expression cluster 2 =
386 57%, **Supplementary Figure 5**).

## 387 Subclones identified in patient P1 have differing predicted treatment 388 outcomes

389 To explore the potential therapeutic resistance of the subclones identified in Patient P1 by
390 LongSom, we investigated the genomic and transcriptomic variations between them. The
391 *ALDH3A2* pathogenic variant identified earlier was exclusively expressed in subclone A, while
392 the *CCAR2* pathogenic variant was exclusive to subclone B (**Supplementary Table 4**).
393 Remarkably, *ALDH3A2* is a ferroptosis inhibitor and its loss of function would lower cisplatin
394 resistance (Dong et al. 2023), while *CCAR2* is a suppressor of homologous recombination,
395 and its loss would lead to resistance against PARP inhibitors (Iyer et al. 2022). Therefore,
396 based on SNVs, subclone A is more likely to be treatment-sensitive, while subclone B is more
397 likely to be treatment-resistant. Fusions SMG7--CH507-513H4.1 and GS1-279B7.2--GNG4
398 were exclusively expressed in subclone B (**Figure 6a**), however, their pathogenicity is difficult
399 to predict. On the transcriptomic level, Subclone B had notably downregulated expression of
400 keratin genes *KRT8* and *KRT18*, two epithelial markers used to differentiate HGSOC cells
401 from non-cancer cells (**Figure 6f**). When compared to cancer subclones in patients P2 and
402 P3, *KRT8* and *KRT18* were downregulated in subclone B, but not upregulated in subclone A,
403 thus confirming a downregulation in subclone B (**Supplementary Figure 6a,b**). It has been
404 shown in vitro that *KRT8* and *KRT18* have a protective effect against cell death (Bozza et al.
405 2018), and loss of *KRT8* and *KRT18* leads to increased invasiveness but also cisplatin
406 sensitivity (Fortier et al. 2013). Subclone B is therefore more likely to be chemosensitive than
407 subclone A. We additionally investigated differential isoform usage, and while both subclones
408 were mostly similar, we found a significant difference in *CHPF* (**Figure 6g**), *MYL6*, the tumor
409 suppressor *BTG2,* and *NUTM2B-AS1* (**Supplementary Figure 6c-e**), however, we could not

410 predict their pathogenicity. None of the subclone-exclusive variants or isoforms were detected

411 in Patient P1 SR scRNA-seq data.

412

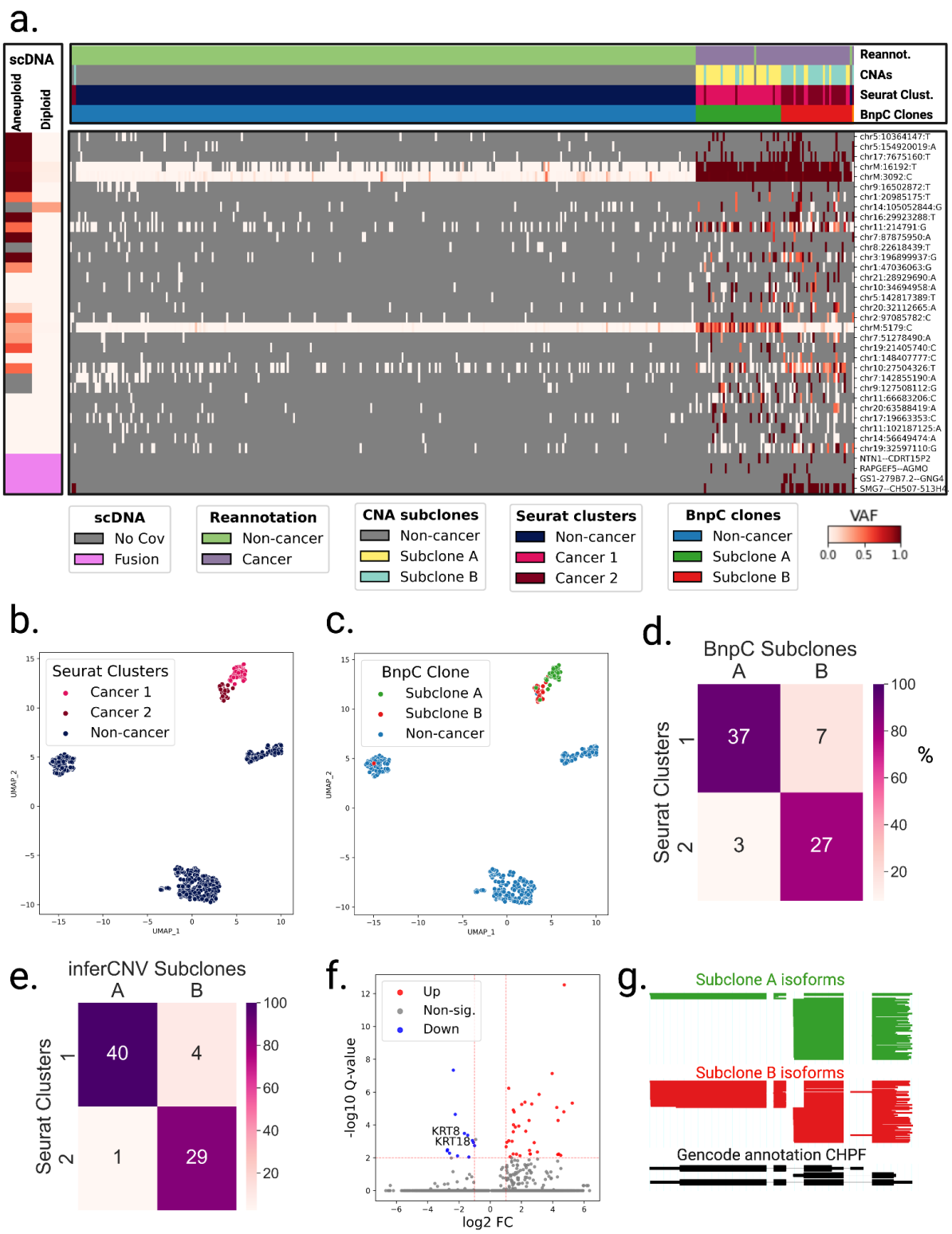


**Figure 6: Analysis of intra-tumor heterogeneity using somatic variants detected in LR scRNA-seq in Patient P1.**

**a.** BnpC clustering of single cells from the tumor biopsy of patient P1 (columns) by somatic SNVs and fusions (rows). VAF is depicted as a gradient from white (no mutated reads, VAF=0) to red (only mutated

418 reads, VAF=1). Grey indicates no coverage in the cell at a given locus. Rows are colored by the scDNA
419 VAF of aggregated diploid and aneuploid cells at the loci: SNVs with high aneuploid VAF and low diploid
420 VAF are somatic in scDNA data. Fusions appear in pink. Columns are colored from top to bottom by
421 cell types reannotated by LongSom, CNAs subclones, expression clusters, and BnpC clusters **b,c.**
422 UMAP embedding of patient P1 gene expression data, colored by (**b**) cell types reannotated by
423 LongSom and (**c**) subclones inferred from somatic SNVs and fusions. The dashed line indicates the
424 manual separation between cancer clusters 1 and 2. **d,e.** Confusion matrix of cells in each expression-
425 derived cancer cluster (rows) and **(d)** cells in the subclones inferred from BnpC, and **(e)** cells in the
426 subclones inferred from inferCNV (columns), colored by the percentage of the total number of cells in
427 each subclone and annotated by the absolute numbers. **f**. Volcano plot of differentially expressed genes
428 identified between subclones B and A. Keratin genes downregulated in subclone B are annotated. **g.**
429 ScisorWiz representation of CHPF isoforms in subclones A and B. Colored areas are exons, whitespace
430 areas are intronic space, not drawn to scale, and each horizontal line represents a single read colored
431 according to subclones.

# 432 Discussion

433 SNVs, CNAs, fusions, gene expression, isoforms expression, and the micro-environment
434 composition can all affect cancer treatment outcomes (Marine et al. 2020). Assessing all of
435 these variations simultaneously from a single patient sample is particularly relevant in a clinical
436 setting, where biological material is limited. Here, we show for the first time that this is possible
437 using LR scRNA-seq data and we introduce LongSom, a workflow for detecting de novo
438 somatic SNVs, fusions, and CNAs in LR scRNA-seq. When applied to data from three HGSOC
439 patients, it detected panel- and scDNA-seq-validated SNVs, including clinically relevant *TP53,*
440 *ALDH3A2,* and *CCAR2* SNVs. By integrating SNVs and fusions, LongSom successfully
441 reconstructed the clonal heterogeneity and linked variants-defined subclones to CNA-defined
442 subclones and gene expression clusters. Finally, in each subclone, we identified differentially
443 expressed genes as well as subclone-specific SNVs with different implications for treatment

444 resistance. Thus, we demonstrated that LR scRNA-seq is suitable for predicting treatment
445 outcomes.
446
447 The cell-type reannotation step implemented in LongSom, based on the somatic variation
448 profile of cells, led to the detection of up to 2.4 times more somatic SNVs (patient P3) and
449 significantly reduced the false-negative rate without sacrificing sensitivity for precision. In
450 general, cell-type annotation is an open challenge due to overlapping, poorly expressed, or
451 incomplete marker genes, e.g., in ovarian cancer omentum metastases (Lähnemann et al.
452 2020; Van Egeren et al. 2021, 2022). Our proposed reannotation can improve existing
453 methods and shows the potential of combining genomic variants with transcriptomic cell
454 typing.
455
456 To our knowledge, LongSom is the first method combining de novo detection of SNVs and
457 fusions from the same cell to reconstruct clonal heterogeneity. Besides nuclear SNVs, which
458 are commonly obtained from RNA (Muyas et al. 2023; Zhang et al. 2023) and DNA seq (Zafar
459 et al. 2016), LongSom also calls mitochondrial SNVs. In the analyzed HGSOC dataset, the
460 mitochondrial SNVs were called in most cancer and non-cancer cells, and some high-
461 confidence fusion calls were expressed in most clones or subclones (P2: IGF2BP2::TESPA1,
462 P1: SMG7::CH507-513H4.1, etc.), making them ideal variations for cell-type reannotation and
463 clustering. Furthermore, both can be clinically relevant (Amatu et al. 2016; Lei et al. 2018; Cesi
464 et al. 2018; Dentro et al. 2021), as we demonstrated in Patient P2. Mitochondrial RNA is
465 particularly abundant in cancer cells, especially HGSOC (Yuan et al. 2023), and an increasing
466 number of methods are leveraging them for clonal reconstruction or validation (Kwok et al.
467 2022; Miller et al. 2022; Gao et al. 2023). However, we demonstrated that mitochondrial SNVs
468 require special filtering thresholds, as non-cancer cells were frequently contaminated by
469 cancer mitochondrial reads. We assume that entire cancer mitochondria might contaminate
470 non-cancer cells, as we observed mitochondrial SNVs in both scRNA and scDNA-seq data.
471 Whether these mitochondria originate from a biological mechanism, e.g. intercellular transfer

472 from cancer cells into non-cancer cells for microenvironment revitalization (Liu et al. 2021;
473 Zampieri et al. 2021), or technical contaminations, e.g. cancer mitochondria encapsulated
474 jointly with non-cancer cells during single-cell preparation, requires further investigation.
475
476 One limitation of this study is the lack of isoform and fusion annotation in the literature,
477 resulting from the difficulty of detecting them in SR scRNA-seq (Dentro et al. 2021), making it
478 challenging to explore the biological implications of subclone-specific isoforms or fusions. To
479 fully exhaust the possibilities of LR scRNA-seq, characterizing more isoforms and fusions will
480 be necessary in the future. Furthermore, a population-level database dedicated to fusions,
481 similar to gnomAD (Chen et al. 2024) for SNVs, would be beneficial to filter germline fusions.
482 We believe that the reliable detection of isoforms and fusions with LR scRNA-seq is the first
483 step toward this goal.
484
485 Despite rapid progress in the LR scRNA-seq field (Al'Khafaji et al. 2023; Dondi et al. 2023;
486 Joglekar et al. 2023; Marx 2023), multiple technical limitations remain unsolved, limiting the
487 potential of downstream analysis. First, variant detection remains challenging due to the
488 sparsity and low coverage of scRNA-seq assays. LongSom excludes SNVs called in non-
489 cancer cells to filter germline SNVs, possibly leading to false negative somatic SNVs, as
490 shown by the matched panel-seq. Second, read coverage is also uneven within a transcript,
491 as transcripts produced by 10X Genomics Chromium remain incomplete on the 5' end (Hsu
492 et al. 2022). Third, RNA-seq is inherently limited to detecting only expressed SNVs and
493 fusions. Nevertheless, LongSom detected a large fraction of variants in intronic or even
494 intergenic regions. Last, indels are the most common source of errors in LR scRNA-seq data,
495 whereby they are frequently excluded from the analyses, ours included (Shiau et al. 2023). To
496 further improve the genomic analyses of scRNA-seq data, algorithms for filtering technical
497 indels while detecting somatic indels are required, especially as technical indels can lead to
498 false positive somatic SNVs (Ahsan et al. 2021).
499

In summary, we proposed a workflow for detecting multiple genetic variants (SNVs, CNAs, fusions) in LR scRNA-seq, enabling clonal heterogeneity reconstruction and clonal genotypes to treatment-resistance phenotype linkage. LR scRNA-seq provides a unique snapshot of the cellular mechanisms by capturing multiple genomic and transcriptomic readouts from the same cell, including expressed isoforms, fusion transcripts, SNVs, and CNAs, more effectively than with any other sequencing technique. With decreasing costs and increasing data size in LR scRNA-seq, we envision that LR scRNA-seq will become more common, potentially facilitating a better understanding of the processes underlying cancer treatment resistance. LongSom can be a valuable first step in guiding these analyses.

# Methods

## scRNA expression analysis

### Gene expression counts

LR gene expression counts were generated as described in (Dondi et al. 2023). Briefly, we preprocessed the BAM files using scIsoPrep (https://github.com/cbg-ethz/scIsoPrep/tree/master) and generated a gene expression-cell matrix. UMI counts were quality-controlled and cells and genes were filtered to remove mitochondrial and ribosomal contaminants. Cells for which over 50% of the reads mapped to mitochondrial genes and cells with fewer than 400 genes expressed were removed. By default, all non-protein-coding genes, genes coding for ribosomal and mitochondrial proteins, and genes that were expressed in less than 20 cells were removed. Subsequently, counts were normalized with sctransform (Hafemeister and Satija 2019), regressing out cell cycle effects and library size as non-regularized dependent variables.

523 ### Marker gene expression-based annotation

524 Cells were annotated with scROSHI (Prummer et al. 2023) using ovarian cancer marker gene
525 lists. Marker genes are available at https://github.com/ETH-
526 NEXUS/scAmpi_single_cell_RNA/blob/master/required_files/ovarian/celltype_list_ovarian.g
527 m). We used "HGSOC" labels as cancer cells, and "Mesothelial.cells", "Fibroblast",
528 "T.NK.cells", "B.cells", "Myeloid.cells", "Endothelial.cells" labels as non-cancer cells.

529 ### Clustering and visualization

530 Similar cells were grouped using Seurat FindClusters (Hao et al. 2024), and clusters with a
531 majority (>90%) of non-cancer cells were grouped together as "non-cancer". The results of the
532 clustering and cell typing are visualized in a low-dimensional representation using Uniform
533 Manifold Approximation and Projection (UMAP).

534 ### Differential gene expression analysis

535 Differential expression was computed using Seurat FindMarkers (Hao et al. 2024),
536 which uses a Wilcoxon test, corrected for multiple testing using the Bonferroni
537 correction. A threshold of corrected P-value <0.01 and abs(log2(fold change)) >1 was
538 used for significance.

539 ### Differential isoform usage analysis

540 Isoform classification and quantification were performed using scIsoPrep. Differential isoform
541 testing was performed using a $\chi 2$ test as previously described in Scisorseqr (Joglekar et al.
542 2021). Differentially used isoforms were visualized using ScisorWiz (Stein et al. 2022).

## Somatic variants calling in LR scRNA-seq data with LongSom

To call somatic variants in LR scRNA-seq, we developed LongSom, a workflow implemented in python3 using Snakemake (Köster and Rahmann 2012) and available at https://github.com/cbg-ethz/LongSom.

### Preprocessing

Long reads with minimal quality Q20 were de-concatenated, adapters were trimmed, demultiplexed, polyA tails were trimmed and finally, UMIs were deduplicated using scIsoPrep (https://github.com/cbg-ethz/scIsoPrep/tree/master) as described in (Dondi et al. 2023). All deduplicated reads belonging to a cell passing filter (cells for which under 50% of the reads mapped to mitochondrial genes and cells with more than 400 genes expressed, see (Dondi et al. 2023), were then pooled together in a pseudo bulk fashion. Gene expression-based cell types were derived from the same work (Dondi et al. 2023).

### SNV calling in LR scRNA-seq data using CTAT-Mutations

First, LongSom calls somatic SNVs in the tumor and (when available) normal biopsy pseudo bulks, using the CTAT mutations pipeline v4.0.0 (https://github.com/NCIP/ctat-mutations/releases/tag/CTAT-Mutations-v4.0.0), which we enhanced to enable compatibility with long reads and report variants according to single cell barcodes. When executed with option --is_long_reads, minimap2 (Li 2018) is used to align long isoform reads to the reference genome hg38 (instead of the STAR aligner used with shorter Illumina RAN-seq), followed by our implementation of the GATK best practices for variant calling using RNA-seq (https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels). Loci flagged as RNA-editing sites or with less than 5 reads mutated are filtered out. For generating variant reports at single-cell resolution, allele-supporting reads annotated with cell barcodes and UMIs were captured from the aligned reads, tallied, and reported for downstream integration with cell typing and related metadata.

568 ## Fusion calling in LR scRNA-seq data using CTAT-LR-Fusion

569 LongSom detects fusions on the single cell level using CTAT-LR-fusion v0.13.0 (
570 https://github.com/TrinityCTAT/CTAT-LR-fusion/releases/tag/ctat-LR-fusion-v0.13.0) with
571 standard options (Qin et al. 2024).

572 ## Cell-variant matrices construction

573 LongSom defines three groups based on the marker-expression-based cell types: cancer cells
574 in the tumor biopsy (in this study, HGSOC cells), non-cancer cells in the tumor biopsy (in this
575 study: mesothelial cells, fibroblasts, T cells, myeloid cells, B cells, and endothelial cells) and,
576 if available, normal cells from the normal biopsy. For each of those groups, LongSom builds a
577 cell-variant matrix with n cells (columns) and m SNVs + p fusions (rows). For SNV rows, the
578 matrices are filled as follows: if at least one read is covering the locus in a cell, a VAF is
579 computed for this cell (with a value ranging from 0 to 1), otherwise, the position is a missing
580 value. A cell is defined as "mutated" at an SNV locus if it has a VAF >= 0.3. For fusion rows,
581 the matrices are filled as follows: a cell with at least one fusion read is considered "mutated"
582 for this fusion (value = 1), otherwise, it is a missing value.

583 ## Cell type reannotation

584 To improve the cell type annotation, LongSom defines a set of "high-confidence cancer
585 variants". To be a "high-confidence cancer variant", an SNV needs to (1) be mutated in more
586 than 5% of cancer cells, (2) be mutated in >20% of the cancer cells covering the locus, (3)
587 have >1% of non-cancer cells covering the locus, (4) be mutated in less than 5% of the non-
588 cancer cells covering the locus, and (5) be mutated in 0 normal cells (optional). For
589 mitochondrial SNVs, due to the contaminations observed, LongSom does not follow those
590 rules. Instead, a mitochondrial SNV is a "high-confidence cancer variant" if:

591 $\%\, of\ cancer\ cells\ \ mutated\ -\ \ \%\, of\ noncancer\ cells\ mutated > 20\%\,.$

592 To be a "high-confidence cancer variant", a fusion needs to be found in more than 5 cancer
593 cells and less than 5% of the non-cancer cells. We then reannotated the cell types by defining

as “cancer” any cell mutated in more than two of the "high-confidence cancer variants", and as “non-cancer” all the other cells in the tumor biopsy.

### Final somatic variants call set and matrix

After cell reannotation, LongSom rebuilds two cell-variants matrices using the annotated cancer and non-cancer labels. Longsom then filters germline polymorphisms (rows) from the variant matrices in five steps: (A) It filters SNV loci detected in the matched normal, when available. (B) It filters SNV loci from the gnomAD database (Chen et al. 2024) with a frequency of at least 0.01% in the total population. (C) After cell-type reannotation, it filters SNV loci that were called in more than 1% of the non-cancer cells. (D) SNV loci where less than 1% of the non-cancer cells are covered by at least one read are filtered. This step helps to filter germline SNVs not detected due to low expression in non-cancer cells. (E) Finally, adjacent SNV loci within a 10,000 bp distance are filtered, as these are likely to be misalignment artifacts in low-complexity regions. Of note, steps (C) and (E) are not applied to mitochondrial SNVs. Finally, LongSom keeps somatic loci that are mutated in a minimum of five cancer cells or 5% of cancer cells (user-defined parameters).

Cancer and non-cancer cell-variant matrices containing only somatic SNVs and fusions are then concatenated to create the final cell-variant matrix. SNVs are sorted in decreasing order by:

$$Diff = mean(\%\ of\ covered\ cancer\ cells\ mutated) - mean(\%\ of\ covered\ noncancer\ cells\ mutated)$$

## Clonal detection based on SNVs and fusions

LongSom uses the cell-variant matrices as input for Bayesian non-parametric clustering (BnpC) (Borgsmüller et al. 2020) to detect subclones in cancer samples, with arguments -n 16 --steps 1000 --DPa_prior [1,1] --conc_update_prob 0 --param_prior [1,1].

## Clonal detection based on CNAs

LongSom first computes cell-gene matrices using featureCounts from Subread v2.0.6 (https://subread.sourceforge.net/) with parameters -L, using hg38 and gencode v36 as reference. It then uses those matrices as input for inferCNV to detect CNA subclones (https://github.com/broadinstitute/infercnv). For running CreateInfercnvObject, reannotated non-cancer cells are used as a reference, and the parameter min_max_counts_per_cell=c(1e3,1e7) is used. For running inferCNV, the parameters cutoff=0.1 and leiden_resolution=0.01 are used. The CNA profiles displayed in this study are the ones obtained from the Hidden Markov Model learned by inferCNV.

# scDNA analysis

## Preprocessing and clonal reconstruction

Using annotated cell types, we re-computed the cell-variant matrices as well as the percentage of cells mutated, the percentage of cells covered, and the percentage of covered cells mutated, for each locus. We then called the final somatic SNVs set at all loci mutated in more than 5% of cancer cells, mutated in less than 1% of the non-cancer cells (min. 1% non-cancer cells covered), and mutated in no normal cells. We obtained copy number profiles and detected the main clonal structure of samples using SCICoNE (Kuipers et al. 2020). Subclones were considered as cancer subclones if they had an aneuploid CNA profile, and as non-cancer subclones if they had a fully diploid CNA profile.

638 ### Variant allele calling in scDNA subclones

639 We investigated all loci from the final somatic SNV set in scDNA subclones in a pseudobulk
640 manner. Cancer subclones were pooled together as well as non-cancer subclones because
641 the coverage was low (<10x per subclone). scDNA subclones with a mean VAF>10% in an
642 SNV locus were considered as supporting the SNV.

643 ### Clinically relevant SNVs

644 Clinically relevant SNVs were detected using the CTAT-Mutations pipeline
645 (https://github.com/NCIP/ctat-mutations/releases/tag/CTAT-Mutations-v4.0.0). Briefly, an
646 SNV was considered clinically relevant if it completed one of these conditions: it was flagged
647 as pathogenic by ClinVar (Landrum et al. 2014), the CHASMplus (Tokheim and Karchin 2019)
648 P-value was <0.05, the VEST (Carter et al. 2013) P-value was <0.05, or FATHMM (Rogers et
649 al. 2018) flagged it as "CANCER", or "PATHOGENIC".

650 ### Panel validation

651 To investigate LongSom somatic SNV calls, we used the FoundationOne®CDx targeted
652 NGS panel (Milbury et al. 2022) in matched bulk DNA samples. SNVs detected in the bulk
653 DNA panel but not by LongSom were independently investigated in scRNA-seq data to
654 detect variant allele read support.

655 ## Data availability

656 The raw sequencing files, as well as the associated analysis files reported in this study are
657 available in the European Genome-phenome Archive (EGA) under the accession number
658 EGAS00001006807. Gencode v36 gene annotation used in this study is available at
659 https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_36/gencode.v36.anno
660 tation.gtf.gz. All additional information will be made available upon reasonable request to the

661 authors. Marker genes for cancer and non-cancer cells are available at
662 https://github.com/ETH-
663 NEXUS/scAmpi_single_cell_RNA/blob/master/required_files/ovarian/celltype_list_ovarian.g
664 mx.

# 665 Code availability

666 LongSom is available at https://github.com/cbg-ethz/LongSom.
667

# 668 Bibliography


669 Ahsan MU, Liu Q, Fang L, Wang K. 2021. NanoCaller for accurate detection of SNPs
670 and indels in difficult-to-map regions from long-read sequencing by haplotype-aware
671 deep neural networks. *Genome Biol* **22**: 261.

672 Al'Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, Gatzen
673 M, Sarkizova S, Schwartz MA, Blaum EM, et al. 2023. High-throughput RNA isoform
674 sequencing using programmed cDNA concatenation. *Nat Biotechnol*.

675 Amatu A, Sartore-Bianchi A, Siena S. 2016. NTRK gene fusions as novel targets of
676 cancer therapy across multiple tumour types. *ESMO Open* **1**: e000023.

677 Borgsmüller N, Bonet J, Marass F, Gonzalez-Perez A, Lopez-Bigas N, Beerenwinkel N.
678 2020. BnpC: Bayesian non-parametric clustering of single-cell mutation profiles.
679 *Bioinformatics* **36**: 4854–4859.

680 Bozza WP, Zhang Y, Zhang B. 2018. Cytokeratin 8/18 protects breast cancer cell lines
681 from TRAIL-induced apoptosis. *Oncotarget* **9**: 23264–23273.

682 Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian
683 disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**: S3.

684 Cesi G, Philippidou D, Kozar I, Kim YJ, Bernardin F, Van Niel G, Wienecke-Baldacchino
685 A, Felten P, Letellier E, Dengler S, et al. 2018. A new ALK isoform transported by
686 extracellular vesicles confers drug resistance to melanoma cells. *Mol Cancer* **17**: 145.

687 Cheng K-C, Lin R-J, Cheng J-Y, Wang S-H, Yu J-C, Wu J-C, Liang Y-J, Hsu H-M, Yu J,
688 Yu AL. 2019. FAM129B, an antioxidative protein, reduces chemosensitivity by
689 competing with Nrf2 for Keap1 binding. *EBioMedicine* **45**: 25–38.

690 Chen C, Zhao S, Zhao X, Cao L, Karnad A, Kumar AP, Freeman JW. 2022.
691 Gemcitabine resistance of pancreatic cancer cells is mediated by IGF1R dependent
692 upregulation of CD44 expression and isoform switching. *Cell Death Dis* **13**: 682.

693 Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA,
694 Vittal C, Gauthier LD, et al. 2024. A genomic mutational constraint map using variation
695 in 76,156 human genomes. *Nature* **625**: 92–100.

696 Dagogo-Jack I, Shaw AT. 2018. Tumour heterogeneity and resistance to cancer
697 therapies. *Nat Rev Clin Oncol* **15**: 81–94.

698 Dentro SC, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar AG, Yu K,
699 Rubanova Y, Macintyre G, Demeulemeester J, et al. 2021. Characterizing genetic intra-
700 tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**: 2239-2254.e39.

701 Ding S, Chen X, Shen K. 2020. Single-cell RNA sequencing in breast cancer:
702 Understanding tumor heterogeneity and paving roads to individualized therapy. *Cancer*
703 *Commun (Lond)* **40**: 329–344.

704 Dondi A, Lischetti U, Jacob F, Singer F, Borgsmüller N, Coelho R, Tumor Profiler
705 Consortium, Heinzelmann-Schwarz V, Beisel C, Beerenwinkel N. 2023. Detection of
706 isoforms and genomic alterations by high-throughput full-length single-cell RNA
707 sequencing in ovarian cancer. *Nat Commun* **14**: 7780.

708 Dong H, He L, Sun Q, Zhan J, Li J, Xiong X, Zhuang L, Wu S, Li Y, Yin C, et al. 2023.
709 Inhibit ALDH3A2 reduce ovarian cancer cells survival via elevating ferroptosis

710 sensitivity. *Gene* **876**: 147515.

711 Duan M, Hao J, Cui S, Worthley DL, Zhang S, Wang Z, Shi J, Liu L, Wang X, Ke A, et
712 al. 2018. Diverse modes of clonal evolution in HBV-related hepatocellular carcinoma
713 revealed by single-cell genome sequencing. *Cell Res* **28**: 359–373.

714 Foord C, Hsu J, Jarroux J, Hu W, Belchikov N, Pollard S, He Y, Joglekar A, Tilgner HU.
715 2023. The variables on RNA molecules: concert or cacophony? Answers in long-read
716 sequencing. *Nat Methods* **20**: 20–24.

717 Fortier A-M, Asselin E, Cadrin M. 2013. Keratin 8 and 18 loss in epithelial cancer cells
718 increases collective cell migration and cisplatin sensitivity through claudin1 up-
719 regulation. *J Biol Chem* **288**: 11555–11571.

720 Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, Kumar T, Hu M, Sei E, Davis A,
721 et al. 2021. Delineating copy number and clonal substructure in human tumors from
722 single-cell transcriptomes. *Nat Biotechnol* **39**: 599–608.

723 Gao T, Soldatov R, Sarkar H, Kurkiewicz A, Biederstedt E, Loh P-R, Kharchenko PV.
724 2023. Haplotype-aware analysis of somatic copy number variations from single-cell
725 transcriptomes. *Nat Biotechnol* **41**: 417–426.

726 Guo Q, Wang H, Duan J, Luo W, Zhao R, Shen Y, Wang B, Tao S, Sun Y, Ye Q, et al.
727 2022. An alternatively spliced p62 isoform confers resistance to chemotherapy in breast
728 cancer. *Cancer Res* **82**: 4001–4015.

729 Hafemeister C, Satija R. 2019. Normalization and variance stabilization of single-cell
730 RNA-seq data using regularized negative binomial regression. *Genome Biol* **20**: 296.

731 Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, Srivastava A,
732 Molla G, Madad S, Fernandez-Granda C, et al. 2024. Dictionary learning for integrative,
733 multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**: 293–304.

734 He C, Sun J, Liu C, Jiang Y, Hao Y. 2019. Elevated H3K27me3 levels sensitize
735 osteosarcoma to cisplatin. *Clin Epigenetics* **11**: 8.

736 Hsu J, Jarroux J, Joglekar A, Romero JP, Nemec C, Reyes D, Royall A, He Y,
737 Belchikov N, Leo K, et al. 2022. Comparing 10x Genomics single-cell 3' and 5' assay in
738 short-and long-read sequencing. *BioRxiv*.

739 Iyer DR, Harada N, Clairmont C, Jiang L, Martignetti D, Nguyen H, He YJ, Chowdhury
740 D, D'Andrea AD. 2022. CCAR2 functions downstream of the Shieldin complex to
741 promote double-strand break end-joining. *Proc Natl Acad Sci USA* **119**: e2214935119.

742 Jamal-Hanjani M, Quezada SA, Larkin J, Swanton C. 2015. Translational implications of
743 tumor heterogeneity. *Clin Cancer Res* **21**: 1258–1266.

744 Joglekar A, Foord C, Jarroux J, Pollard S, Tilgner HU. 2023. From words to complete
745 phrases: insight into single-cell isoforms using short and long reads. *Transcription* 1–13.

746 Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, Marrocco J,
747 Williams SR, Haase B, Hayes A, et al. 2021. A spatially resolved brain region- and cell
748 type-specific isoform atlas of the postnatal mouse brain. *Nat Commun* **12**: 463.

749 Köster J, Rahmann S. 2012. Snakemake--a scalable bioinformatics workflow engine.
750 *Bioinformatics* **28**: 2520–2522.

751 Kuipers J, Tuncel MA, Ferreira P, Jahn K, Beerenwinkel N. 2020. Single-cell copy
752 number calling and event history reconstruction. *BioRxiv*.

753 Kwok AWC, Qiao C, Huang R, Sham M-H, Ho JWK, Huang Y. 2022. MQuad enables
754 clonal substructure discovery using single cell mitochondrial variants. *Nat Commun* **13**:
755 1205.

756 Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos
757 CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. 2020. Eleven grand challenges in
758 single-cell data science. *Genome Biol* **21**: 31.

759 Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR.
760 2014. ClinVar: public archive of relationships among sequence variation and human
761 phenotype. *Nucleic Acids Res* **42**: D980-5.

762 Lappalainen T, Sammeth M, Friedländer MR, ’t Hoen PAC, Monlong J, Rivas MA,
763 Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. 2013. Transcriptome and
764 genome sequencing uncovers functional variation in humans. *Nature* **501**: 506–511.

765 Lei JT, Shao J, Zhang J, Iglesia M, Chan DW, Cao J, Anurag M, Singh P, He X, Kosaka
766 Y, et al. 2018. Functional Annotation of ESR1 Gene Fusions in Estrogen Receptor-
767 Positive Breast Cancer. *Cell Rep* **24**: 1434-1444.e7.

768 Liu D, Gao Y, Liu J, Huang Y, Yin J, Feng Y, Shi L, Meloni BP, Zhang C, Zheng M, et al.
769 2021. Intercellular mitochondrial transfer as a means of tissue revitalization. *Signal*
770 *Transduct Target Ther* **6**: 65.

771 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:
772 3094–3100.

773 Mansoori B, Mohammadi A, Davudian S, Shirjang S, Baradaran B. 2017. The different
774 mechanisms of cancer drug resistance: A brief review. *Adv Pharm Bull* **7**: 339–348.

775 Marine J-C, Dawson S-J, Dawson MA. 2020. Non-genetic mechanisms of therapeutic
776 resistance in cancer. *Nat Rev Cancer* **20**: 743–756.

777 Marx V. 2023. Method of the year: long-read sequencing. *Nat Methods* **20**: 6–11.

778 Milbury CA, Creeden J, Yip W-K, Smith DL, Pattani V, Maxwell K, Sawchyn B, Gjoerup
779 O, Meng W, Skoletsky J, et al. 2022. Clinical and analytical validation of
780 FoundationOne®CDx, a comprehensive genomic profiling assay for solid tumors. *PLoS*
781 *ONE* **17**: e0264138.

782 Miller TE, Lareau CA, Verga JA, DePasquale EAK, Liu V, Ssozi D, Sandor K, Yin Y,
783 Ludwig LS, El Farran CA, et al. 2022. Mitochondrial variant enrichment from high-
784 throughput single-cell RNA sequencing resolves clonal populations. *Nat Biotechnol* **40**:
785 1030–1034.

786 Mitra D, Brumlik MJ, Okamgba SU, Zhu Y, Duplessis TT, Parvani JG, Lesko SM, Brogi
787 E, Jones FE. 2009. An oncogenic isoform of HER2 associated with locally disseminated

788 breast cancer and trastuzumab resistance. *Mol Cancer Ther* **8**: 2152–2162.

789 Muyas F, Sauer CM, Valle-Inclán JE, Li R, Rahbari R, Mitchell TJ, Hormoz S, Cortés-
790 Ciriano I. 2023. De novo detection of somatic mutations in high-throughput single-cell
791 profiling data sets. *Nat Biotechnol*.

792 Osorio D, Cai JJ. 2021. Systematic determination of the mitochondrial proportion in
793 human and mice tissues for single-cell RNA-sequencing data quality control.
794 *Bioinformatics* **37**: 963–967.

795 Prummer M, Bertolini A, Bosshard L, Barkmann F, Yates J, Boeva V, Tumor Profiler
796 Consortium, Stekhoven D, Singer F. 2023. scROSHI: robust supervised hierarchical
797 identification of single cells. *NAR Genom Bioinform* **5**: lqad058.

798 Qin Q, Popic V, Yu H, White E, Khorgade A, Shin A, Wienand K, Dondi A, Beerenwinkel
799 N, Vazquez F, et al. 2024. CTAT-LR-fusion: accurate fusion transcript identification from
800 long and short read isoform sequencing at bulk or single cell resolution. *BioRxiv*.

801 Ramón Y Cajal S, Sesé M, Capdevila C, Aasen T, De Mattos-Arruda L, Diaz-Cano SJ,
802 Hernández-Losa J, Castellví J. 2020. Clinical implications of intratumor heterogeneity:
803 challenges and opportunities. *J Mol Med* **98**: 161–177.

804 Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. 2018. FATHMM-
805 XF: accurate prediction of pathogenic point mutations via extended features.
806 *Bioinformatics* **34**: 511–513.

807 Roth A, McPherson A, Laks E, Biele J, Yap D, Wan A, Smith MA, Nielsen CB, McAlpine
808 JN, Aparicio S, et al. 2016. Clonal genotype and population structure inference from
809 single-cell tumor sequencing. *Nat Methods* **13**: 573–576.

810 Serin Harmanci A, Harmanci AO, Zhou X. 2020. CaSpER identifies and visualizes CNV
811 events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat Commun*
812 **11**: 89.

813 Shiau C-K, Lu L, Kieser R, Fukumura K, Pan T, Lin H-Y, Yang J, Tong EL, Lee G, Yan

814 Y, et al. 2023. High throughput single cell long-read sequencing analyses of same-cell
815 genotypes and phenotypes in human tumors. *Nat Commun* **14**: 4124.

816 Stein AN, Joglekar A, Poon C-L, Tilgner HU. 2022. ScisorWiz: visualizing differential
817 isoform expression in single-cell long-read data. *Bioinformatics* **38**: 3474–3476.

818 Tokheim C, Karchin R. 2019. Chasmplus reveals the scope of somatic missense
819 mutations driving human cancers. *Cell Syst* **9**: 9-23.e8.

820 Van Egeren D, Escabi J, Nguyen M, Liu S, Reilly CR, Patel S, Kamaz B, Kalyva M,
821 DeAngelo DJ, Galinsky I, et al. 2021. Reconstructing the lineage histories and
822 differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell*
823 *Stem Cell* **28**: 514-523.e9.

824 Van Egeren D, Kamaz B, Liu S, Nguyen M, Reilly CR, Kalyva M, DeAngelo DJ,
825 Galinsky I, Wadleigh M, Winer ES, et al. 2022. Transcriptional differences between
826 JAK2-V617F and wild-type bone marrow cells in patients with myeloproliferative
827 neoplasms. *Exp Hematol* **107**: 14–19.

828 Vasan N, Baselga J, Hyman DM. 2019. A view on drug resistance in cancer. *Nature*
829 **575**: 299–309.

830 Xia J, Li S, Liu S, Zhang L. 2023. Aldehyde dehydrogenase in solid tumors and other
831 diseases: Potential biomarkers and therapeutic targets. *MedComm* **4**: e195.

832 Yuan Y, Ju YS, Kim Y, Li J, Wang Y, Yoon CJ, Yang Y, Martincorena I, Creighton CJ,
833 Weinstein JN, et al. 2023. Author Correction: Comprehensive molecular characterization
834 of mitochondrial genomes in human cancers. *Nat Genet* **55**: 1078.

835 Zafar H, Wang Y, Nakhleh L, Navin N, Chen K. 2016. Monovar: single-nucleotide
836 variant detection in single cells. *Nat Methods* **13**: 505–507.

837 Zampieri LX, Silva-Almeida C, Rondeau JD, Sonveaux P. 2021. Mitochondrial transfer
838 in cancer: A comprehensive review. *Int J Mol Sci* **22**.

Zhang T, Jia H, Song T, Lv L, Gulhan DC, Wang H, Guo W, Xi R, Guo H, Shen N. 2023. De novo identification of expressed cancer somatic mutations from single-cell RNA sequencing data. *Genome Med* **15**: 115.

# Acknowledgments

We thank Joanna Hård for her help with mitochondrial mutation contamination. We thank Ivan Topolsky for his support with cloud computing. We thank Lara Fuhrmann for naming LongSom. A.D. and N.Bo were supported by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement (#766030 to N.Be). B.J.H. was supported by the National Cancer Institute grant U24CA180922.

# Author information

## Corresponding author

Correspondence to Niko Beerenwinkel: <niko.beerenwinkel@bsse.ethz.ch>

## Author contributions

N.Be. acquired the funding for this project. A.D. conceptualized the idea of this project. A.D. designed the LongSom pipeline with the help of N.Bo. The Tumor Profiler Consortium provided scDNA-seq data and NGS panel results. A.D. conducted all computational analyses except scDNA subclones inference (performed by P.F.). A.D. did all the visualizations and figures. A.D., N.Bo., F.J., and N.Be interpreted the results. A.D. and N.Bo. wrote the manuscript with contributions from all authors. B.J.H integrated single cell and long read isoform support into CTAT-Mutations and assisted with fusion transcript identification via CTAT-LR-fusion. All authors read and approved the final manuscript.

861 # Tumor Profiler Consortium authors list

862 Rudolf Aebersold[2], Melike Ak[28], Faisal S Al-Quaddoomi[9,17], Silvana I Albert[7], Jonas Albinus[7],
863 Ilaria Alborelli[24], Sonali Andani[6,17,26,31], Per-Olof Attinger[11], Marina Bacac[16], Daniel
864 Baumhoer[24], Beatrice Beck-Schimmer[39], Niko Beerenwinkel[4,17], Christian Beisel[4], Lara
865 Bernasconi[27], Anne Bertolini[9,17], Bernd Bodenmiller[8,35], Ximena Bonilla[6,17,26], Lars
866 Bosshard[9,17], Byron Calgua[24], Ruben Casanova[35], Stéphane Chevrier[35], Natalia
867 Chicherova[9,17], Ricardo Coelho[18], Maya D'Costa[10], Esther Danenberg[37], Natalie
868 Davidson[6,17,26], Monica-Andreea Drăgan[4], Reinhard Dummer[28], Stefanie Engler[35], Martin
869 Erkens[14], Katja Eschbach[4], Cinzia Esposito[37], André Fedier[18], Pedro Ferreira[4], Joanna
870 Ficek[6,17,26], Anja L Frei[31], Bruno Frey[13], Sandra Goetze[7], Linda Grob[9,17], Gabriele Gut[37],
871 Detlef Günther[5], Martina Haberecker[31], Pirmin Haeuptle[1], Viola Heinzelmann-Schwarz[18,23],
872 Sylvia Herter[16], Rene Holtackers[37], Tamara Huesser[16], Alexander Immer[6,12], Anja Irmisch[28],
873 Francis Jacob[18], Andrea Jacobs[35], Tim M Jaeger[11], Katharina Jahn[4], Alva R James[6,17,26],
874 Philip M Jermann[24], André Kahles[6,17,26], Abdullah Kahraman[17,31], Viktor H Koelzer[31], Werner
875 Kuebler[25], Jack Kuipers[4,17], Christian P Kunze[22], Christian Kurzeder[21], Kjong-Van
876 Lehmann[6,17,26], Mitchell Levesque[28], Ulrike Lischetti[18], Sebastian Lugert[10], Gerd Maass[13],
877 Markus G Manz[30], Philipp Markolin[6,17,26], Martin Mehnert[7], Julien Mena[2], Julian M Metzler[29],
878 Nicola Miglino[1], Emanuela S Milani[7], Holger Moch[31], Simone Muenst[24], Riccardo Murri[38],
879 Charlotte KY Ng[24,34], Stefan Nicolet[24], Marta Nowak[31], Monica Nunez Lopez[18], Patrick GA
880 Pedrioli[3], Lucas Pelkmans[37], Salvatore Piscuoglio[18,24], Michael Prummer[9,17], Natalie
881 Rimmer[18], Mathilde Ritter[18], Christian Rommel[14], María L Rosano-González[9,17], Gunnar
882 Rätsch[3,6,17,26], Natascha Santacroce[4], Jacobo Sarabia del Castillo[37], Ramona Schlenker[15],
883 Petra C Schwalie[14], Severin Schwan[11], Tobias Schär[4], Gabriela Senti[27], Wenguang Shao[7],
884 Franziska Singer[9,17], Sujana Sivapatham[35], Berend Snijder[2,17], Bettina Sobottka[31], Vipin T
885 Sreedharan[9,17], Stefan Stark[6,17,26], Daniel J Stekhoven[9,17], Tanmay Tanna[4,6], Alexandre PA
886 Theocharides[30], Tinu M Thomas[6,17,26], Markus Tolnay[24], Vinko Tosevski[16], Nora C
887 Toussaint[9,17], Mustafa A Tuncel[4,17], Marina Tusup[28], Audrey Van Drogen[7], Marcus Vetter[20],

888 Tatjana Vlajnic[24], Sandra Weber[27], Walter P Weber[19], Rebekka Wegmann[2], Michael
889 Weller[33], Fabian Wendt[7], Norbert Wey[31], Andreas Wicki[30,36], Mattheus HE Wildschut[2,30],
890 Bernd Wollscheid[7], Shuqing Yu[9,17], Johanna Ziegler[28], Marc Zimmermann[6,17,26], Martin
891 Zoche[31], Gregor Zuend[32]

892 [1]Cantonal Hospital Baselland, Medical University Clinic, Rheinstrasse 26, 4410 Liestal,
893 Switzerland, [2]ETH Zurich, Department of Biology, Institute of Molecular Systems Biology,
894 Otto-Stern-Weg 3, 8093 Zurich, Switzerland, [3]ETH Zurich, Department of Biology, Wolfgang-
895 Pauli-Strasse 27, 8093 Zurich, Switzerland, [4]ETH Zurich, Department of Biosystems Science
896 and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland, [5]ETH Zurich, Department of
897 Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 1-5/10, 8093 Zurich, Switzerland,
898 [6]ETH Zurich, Department of Computer Science, Institute of Machine Learning,
899 Universitätstrasse 6, 8092 Zurich, Switzerland, [7]ETH Zurich, Department of Health Sciences
900 and Technology, Otto-Stern-Weg 3, 8093 Zurich, Switzerland, [8]ETH Zurich, Institute of
901 Molecular Health Sciences, Otto-Stern-Weg 7, 8093 Zurich, Switzerland, [9]ETH Zurich,
902 NEXUS Personalized Health Technologies, John-von-Neumann-Weg 9, 8093 Zurich,
903 Switzerland, [10]F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland,
904 [11]F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland, , [12]Max
905 Planck ETH Center for Learning Systems, , [13]Roche Diagnostics GmbH, Nonnenwald 2,
906 82377 Penzberg, Germany, [14]Roche Pharmaceutical Research and Early Development,
907 Roche Innovation Center Basel, Grenzacherstrasse 124, 4070 Basel, Switzerland, [15]Roche
908 Pharmaceutical Research and Early Development, Roche Innovation Center Munich, Roche
909 Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany, [16]Roche Pharmaceutical
910 Research and Early Development, Roche Innovation Center Zurich, Wagistrasse 10, 8952
911 Schlieren, Switzerland, [17]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland,
912 [18]University Hospital Basel and University of Basel, Department of Biomedicine,
913 Hebelstrasse 20, 4031 Basel, Switzerland, [19]University Hospital Basel and University of
914 Basel, Department of Surgery, Brustzentrum, Spitalstrasse 21, 4031 Basel, Switzerland,

915 [20]University Hospital Basel, Brustzentrum & Tumorzentrum, Petersgraben 4, 4031 Basel,
916 Switzerland, [21]University Hospital Basel, Brustzentrum, Spitalstrasse 21, 4031 Basel,
917 Switzerland, [22]University Hospital Basel, Department of Information- and Communication
918 Technology, Spitalstrasse 26, 4031 Basel, Switzerland, [23]University Hospital Basel,
919 Gynecological Cancer Center, Spitalstrasse 21, 4031 Basel, Switzerland, [24]University
920 Hospital Basel, Institute of Medical Genetics and Pathology, Schönbeinstrasse 40, 4031
921 Basel, Switzerland, [25]University Hospital Basel, Spitalstrasse 21/Petersgraben 4, 4031
922 Basel, Switzerland, [26]University Hospital Zurich, Biomedical Informatics,
923 Schmelzbergstrasse 26, 8006 Zurich, Switzerland, [27]University Hospital Zurich, Clinical
924 Trials Center, Rämistrasse 100, 8091 Zurich, Switzerland, [28]University Hospital Zurich,
925 Department of Dermatology, Gloriastrasse 31, 8091 Zurich, Switzerland, [29]University
926 Hospital Zurich, Department of Gynecology, Frauenklinikstrasse 10, 8091 Zurich,
927 Switzerland, [30]University Hospital Zurich, Department of Medical Oncology and Hematology,
928 Rämistrasse 100, 8091 Zurich, Switzerland, [31]University Hospital Zurich, Department of
929 Pathology and Molecular Pathology, Schmelzbergstrasse 12, 8091 Zurich, Switzerland,
930 [32]University Hospital Zurich, Rämistrasse 100, 8091 Zurich, Switzerland, [33]University
931 Hospital and University of Zurich, Department of Neurology, Frauenklinikstrasse 26, 8091
932 Zurich, Switzerland, [34]University of Bern, Department of BioMedical Research,
933 Murtenstrasse 35, 3008 Bern, Switzerland, [35]University of Zurich, Department of Quantitative
934 Biomedicine, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [36]University of Zurich,
935 Faculty of Medicine, Zurich, Switzerland, [37]University of Zurich, Institute of Molecular Life
936 Sciences, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [38]University of Zurich,
937 Services and Support for Science IT, Winterthurerstrasse 190, 8057 Zurich, Switzerland,
938 [39]University of Zurich, VP Medicine, Künstlergasse 15, 8001 Zurich, Switzerland

# 939 Conflict of interest

940 The authors declare no competing interests.