Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing for personalized oncology

Arthur Dondi[1,2,*], Ulrike Lischetti[3,*,], Francis Jacob[3], Franziska Singer[2,4], Nico Borgsmüller[1,2], Tumor Profiler Consortium, Viola Heinzelmann-Schwarz[3], Christian Beisel[1,**], Niko Beerenwinkel[1,2,**]**

* Equal contributions
** Corresponding authors

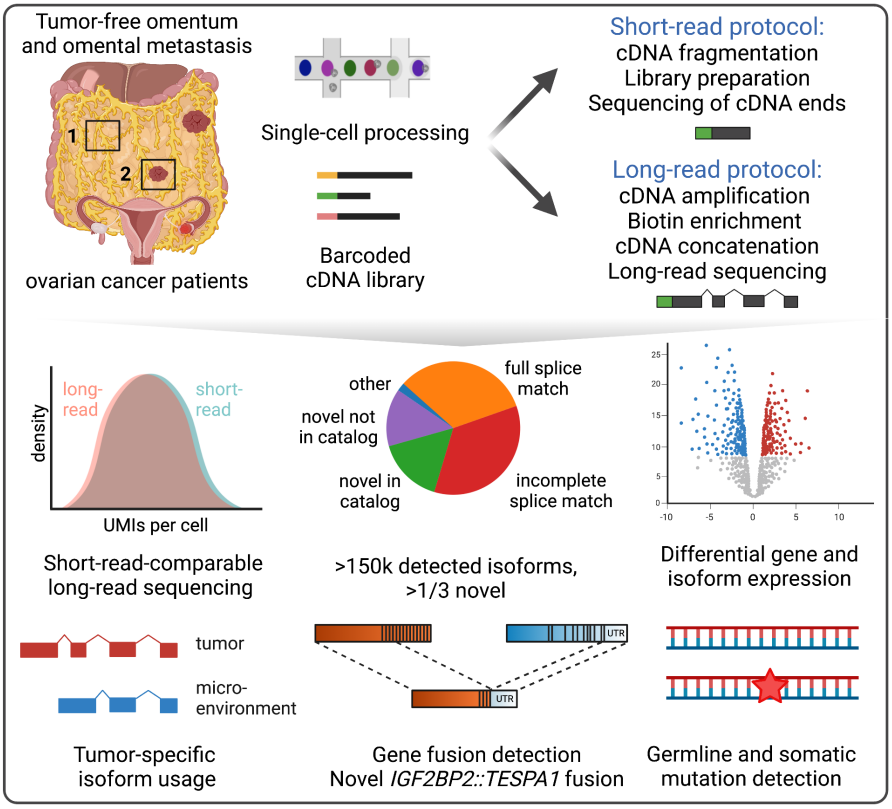
[1] ETH Zurich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Mattenstrasse 26, 4058 Basel, Switzerland
[3] University Hospital Basel and University of Basel, Ovarian Cancer Research, Department of Biomedicine, Hebelstrasse 20, 4031 Basel, Switzerland
[4] ETH Zurich, NEXUS Personalized Health Technologies, Wagistrasse 18, 8952 Schlieren, Switzerland

Keywords: full-length single-cell RNA sequencing, long-read PacBio sequencing, transcript concatenation, isoforms, mutations, gene fusions, ovarian cancer

Abstract

Understanding the complex background of cancer requires genotype-phenotype information in single-cell resolution. Long-read single-cell RNA sequencing (scRNA-seq), capturing full-length transcripts, lacked the depth to provide this information so far. Here, we increased the PacBio sequencing depth to 12,000 reads per cell, leveraging multiple strategies, including artifact removal and transcript concatenation, and applied the technology to samples from three human ovarian cancer patients. Our approach captured 152,000 isoforms, of which over 52,000 were novel, detected cell type- and cell-specific isoform usage, and revealed differential isoform expression in tumor and mesothelial cells. Furthermore, we identified gene fusions, including a novel scDNA sequencing-validated *IGF2BP2::TESPA1* fusion, which was misclassified as high *TESPA1* expression in matched short-read data, and called somatic and germline mutations, confirming targeted NGS cancer gene panel results. With multiple new opportunities, especially for cancer biology, we envision long-read scRNA-seq to become increasingly relevant in oncology and personalized medicine.

Graphical Abstract:



Introduction

Cancer is a complex disease characterized by genomic and transcriptomic alterations[1] that drive multiple tumor-promoting capabilities or hallmarks[2]. Among others, these alterations include point mutations, insertions and deletions (indels), and gene fusions on the genomic level, and splice isoforms on the transcriptomic level. Their detection offers great potential for personalized oncology as they can serve as direct therapeutic targets[3,4] or potential neoantigens informing on the immunogenicity of the tumor[5]. Gene fusions arising from large-scale genomic rearrangements, for example, play an oncogenic role in a variety of tumor types[6], and are successfully used as therapeutic targets[7,8]. Like mutations[9] and copy number variations[10], fusion rates can vary widely across cancer types, and gene fusions are thought to be drivers in 16.5% of cancer cases, and even the only driver in more than 1%[11]. Furthermore, out-of-frame gene fusions are more immunogenic than mutations and indels, making them an ideal target for immunotherapies and cancer vaccines[12,13]. On the transcriptomic level, alternative splicing is a major mechanism for the diversification of a cell's transcriptome and proteome[14] and can impact all hallmarks of tumorigenesis. It also presents a fairly novel non-genomic source of potential neoantigens[15]. In breast and ovarian cancer, 68% of samples had at least one isoform with novel exon-exon junction (neojunction) detected in proteomic data[16].

The complexity of cancer further extends to intra-tumor heterogeneity[17] and its intricate interplay with the tumor microenvironment (TME)[18]. Ultimately, to fully decipher functional tumor heterogeneity and its effect on the TME, single-cell resolution providing both phenotype and genotype information is required. Single-cell RNA sequencing (scRNA-seq) is now widely used for the phenotypic dissection of heterogeneous tissues. It can be divided into short-read, high-throughput technologies allowing for gene expression quantification and long-read, low-throughput technologies that cover full-length transcripts[19]. Up to now, short- and long-read methods had to be used in parallel to combine the advantages of each technology. The long-read scRNA-seq field is rapidly expanding, with methods being constantly developed and improved on Nanopore[20,21] and PacBio[22–26] long-read platforms. So far, long-read RNA-seq has however only been applied on the bulk level in the field of oncology[24,27,28]. High-quality, high-throughput, long-read scRNA-seq has the potential to provide isoform-level cell type-specific readouts and capture tumor-specific genomic alterations. With near ubiquitous p53 mutations and defective DNA repair pathways causing frequent non-recurrent gene fusions, high-grade serous ovarian cancer (HGSOC) is an ideal candidate to investigate these alterations[10,29,30].

Here, for the first time, we used high-quality, high-throughput long-read scRNA-seq to capture cell type-specific genomic and transcriptomic alterations on clinical cancer patients. We applied both short-read and long-read scRNA-seq to five samples from three HGSOC patients, comprising 2,571 cells, and generated the largest PacBio scRNA-seq dataset to date. We were able to identify over 150,000 isoforms, of which a third were novel, as well as novel cell type- and cell-specific isoforms. We detected differential isoform usage in tumor cells and cells of the TME. Additionally, we discovered dysregulations in the insulin like growth factor (IGF) network in tumor cells on the genomic and transcriptomic level. Thereby, we demonstrated that scRNA-seq can capture genomic alterations accurately, including cancer- and patient-specific germline and somatic mutations in genes such as *TP53*, as well as gene fusions, including a novel *IGF2BP2::TESPA1* fusion.

Results

Long-read scRNA-seq creates a catalog of isoforms in ovarian cancer patient-derived tissue samples

We generated short-read and long-read scRNA-seq data from five omentum biopsy samples (**Extended Data Table 1**) from three HGSOC patients. Three samples were derived from HGSOC omental metastases and two from matching distal tumor-free omental tissues (**Fig. 1a**). To generate long reads, we opted for the PacBio platform for its generation of high-fidelity (HiFi) reads through circular consensus sequencing (CCS). To overcome its limitations in sequencing output and optimize for longer library length, we 1) removed template-switch oligo artifacts that can account for up to 50% of reads through biotin enrichment, 2) concatenated transcripts to sequence multiple cDNA molecules per CCS read, and 3) sequenced on the PacBio Sequel II platform (2-4 SMRT 8M cells per sample, **Methods**). This allowed the generation of a total of 212 Mio HiFi reads in 2,571 cells, which, after demultiplexing, deduplication, and intrapriming removal, resulted in 30.7 Mio unique molecular identifiers (UMIs) (**Extended Data Table 1**). On average, 12k UMIs were detected per cell.

The long-read dataset revealed 152,546 isoforms, each associated with at least three UMIs. We classified the isoforms according to the SQANTI classification[31] and calculated their proportions (**Methods**, **Fig. 1b**,**c**): full splice match (FSM) - isoforms already in the GENCODE database (32.8%), incomplete splice matches (ISM) - isoforms corresponding to shorter versions of the FSM (35.1%), novel in catalog (NIC) - isoforms presenting combinations of known splice donors and acceptors (15.9%), and novel not in catalog (NNC) - isoforms harboring at least one unknown splice site, or neojunction (14.4%). Novel isoforms (classes NIC and NNC) accounted for 30% of the isoforms, and 11% of the total reads in all samples, while FSM accounted for 33% of the isoforms and 80% of the reads (**Fig. 1c,d**), indicating that high coverage is required for the reliable detection of new, low abundant, transcripts.

To evaluate the structural integrity of all isoforms, we compared their 5' end to the FANTOM5 CAGE database[32] and their 3' end to the PolyASite database[33] (**Fig. 1e**). More than 82% of the NIC and 74% of NNC isoforms could be validated on 3' and 5' ends, similarly to FSM. As expected, fewer ISM isoforms were found to be complete (42%): they are either incompletely sequenced isoforms missing their 5' end (30%) or the result of early 3' termination (55%).

FSM, NIC, and NNC had overall better 3' and 5' validation than the full-length tagged isoforms in the GENCODE database (**Fig. 1e**). Only the 'Matched Annotation from NCBI and EMBL-EBI' (MANE[34]) containing curated representative transcripts cross-validated between the GENCODE and RefSeq database had a better 3' and 5' validation of 95%. A total of 52,884 novel isoforms were complete (NIC+NNC), of which 40,046 were confirmed as valid novel isoforms by GENCODE, corresponding to 17% of the current GENCODE v36 database. Isoforms that were not confirmed were mainly either "partially redundant with existing transcripts", or "overlapping with multiple loci". Finally, we assessed the biotypes of our newly discovered isoforms, indicative of their presumed functional categorization. We found that 42% are protein coding, more than the 36% of protein coding isoforms found in the GENCODE database (230k entries) (**Fig. 1f-g**). This demonstrates the ability of concatenated long-read

154 sequencing to generate high yield, high-quality data and discover novel isoforms with
155 enhanced annotation.

156 Long-read sequencing allows for short-read-independent cell type 157 identification

158 Next, through comparison to short-read data, we assessed the ability of long-read sequencing
159 to cluster cells and to identify cell types. We generated short- and long-read gene count
160 matrices and removed non-protein-coding, ribosomal, and mitochondrial genes. After filtering,
161 we obtained 16.5 Mio unique long reads associated with 12,757 genes, and 26.3 Mio unique
162 short reads associated with 13,122 genes (**Extended Data Table 1**). The short- and long-read
163 datasets were of similar sequencing depth with a median of 4,930 and 2,750 UMIs per cell,
164 respectively (average 10,235 and 6,413 UMIs, **Extended Data Fig. 1a**). Also, the genes
165 detected in both datasets overlapped by 86.4% (**Extended Data Fig. 1b,c**).
166
167 We first identified cell types independently per cell, using cell type marker gene lists
168 (**Methods**). We compared short- and long-read data and found that both data types identified
169 cell types with similar percentages, namely HGSOC (13% in short-read vs 15% in long-read
170 data), mesothelial cells (22 vs 23%), fibroblasts (9 vs 8%), T cells (38 vs 37%), myeloid cells
171 (both 14%), B cells (3 vs 1%), and endothelial cells (both 1%). Those cell populations
172 expressed cell type specific marker genes (**Extended Data Fig. 1c**). We then projected short-
173 read gene, long-read gene, and long-read isoform expression onto 2-dimensional embeddings
174 using UMAP[35] (**Fig. 2a**). We manually clustered cell types based on the embeddings and
175 calculated the Jaccard distance between clusters. Cell clusters based on short- and long-
176 reads were very similar, with a Jaccard distance >94% for all cell types except B-cells, where
177 the Jaccard distance was >75% (**Fig. 2b**). Furthermore, Jaccard similarity analysis between
178 cell type clusters and attributed cell type labels were analogous between short- and long-read
179 data, with a better prediction of B cells and endothelial cells for long reads **(Extended Data**
180 **Fig. 1b**). These findings show that long-read gene and isoform expression data can be used
181 to identify cell types reliably and independently from short-read data.

182 Long-read sequencing captures germline and somatic mutations and 183 identifies increased neojunctions in tumor cells

184 Next, we assessed the potential of long-read data for mutation detection, and used somatic
185 mutations to further validate the cell type annotation. Germline mutations are expected in all
186 cell types, whereas somatic mutations should be present only in tumor cells. As reference, we
187 used mutations called from a panel covering 324 genes on patient-matched bulk DNA samples
188 (**Methods**). We identified germline variants in 48 cells belonging to all cell types from distal
189 omentum and tumor sites (**Fig. 2c, Supplementary Table 1**). Somatic mutations were called
190 in 34 cells, all in the cell cluster annotated as tumor cells (**Fig. 2d**). In 20 of those cells, *TP53*
191 was found mutated (**Supplementary Table 1**). Thus, high-fidelity long-read data can be
192 leveraged for both germline and somatic mutation calling.

193 We analyzed the expression of cell type-specific isoforms. HGSOC cells expressed more
194 genes, transcript isoforms, and RNA molecules than other cell types **(Extended Data Fig. 3a-**
195 **c)**. This difference does however not translate into mean UMIs per isoform, as isoforms
196 expressed in cancer cells harbor fewer UMIs than in mesothelial cells, for example. This

means that cancer cells express more low-abundant isoforms **(Extended Data Fig. 3d)** suggesting wider isoform diversity and broader cellular functions and controls. Isoform class distribution between cell types revealed a higher fraction of novel isoforms and neojunctions (NNC) in tumor cells (**Fig. 2e**).

We then looked into isoforms uniquely expressed in the different cell types. At the cell type level, cancer cells contained more than 8% (9,476) of cell type-specific isoforms, between 2.3-10.6 times more than the most frequent other cell types (myeloids, T/NK cells, fibroblasts and mesothelial cells) (**Methods, Extended Data Fig. 3e**). At the cellular level, 0.5% of the cancer-specific isoforms were also unique to a single cell, which is between 3-6 times the percentage of unique isoforms in other cell types (**Extended Data Fig. 3e**). In all cell types, cell type-specific isoforms (**Extended Data Fig. 3f)** had a higher percentage of novel isoforms than non-specific isoforms distributed across cells **(Fig. 2e)**. This phenomenon was even stronger in cell-specific isoforms: in cancer, more than 75% of isoforms unique to cells were novel, and 50% of these were neojunctions (NNC) (**Extended Data Fig. 3e**). Those rare isoforms were difficult to detect for previous methods, hence their novelty. Taken together, cancer cells expressed at least twice as many unique isoforms than other cell types, indicating an increased transcriptomic diversification and support previous findings of cancer-specific neojunction expression in bulk data[16].

Differential isoform expression in the tumor microenvironment reveals epithelial-to-mesenchymal transition

Comparing cells from metastatic and tumor-free samples, we found that mesothelial and fibroblast cells showed distinct clustering, in both short- and long-read embeddings (**Fig. 3a**). We observed a bridge between TME fibroblasts and mesothelial cells on the UMAPs, suggesting that TME cells might undergo a form of transdifferentiation. To understand this phenomenon, we analyzed differential isoform and gene expression in TME vs. distal mesothelial and fibroblast cells. For mesothelial cells, the gene with the highest change in relative isoform abundance amongst all its transcripts was the collagen type 1 alpha chain (*COL1A1)* (P_{corr}=6.34x10^{-49}, |ΔΠ|=0.86, **Methods**) **(Fig. 3b)**. TME mesothelial cells used the canonical 3' transcription termination site, while distal cells had a premature transcription termination, resulting in a truncated protein (**Fig. 3c**). *COL1A1* was also the top differentially expressed gene (P = 2x10^{-3}) between TME and distal mesothelial cells, and the fifth most differentially expressed gene between TME and distal fibroblasts (P = 0.015), with TME cells overexpressing it in both cases compared to their distal counterparts. *COL1A2*, was also found to be differentially spliced in TME mesothelial cells (P_{corr}=6.85x10^{-91}, |ΔΠ|=0.37) and fibroblasts (P_{corr}=2.02x10^{-77}, |ΔΠ|=0.36). HGSOC cells showed the same *COL1A2* splicing pattern as TME cells when compared to all non-tumor cells (P_{corr}=6.54x10^{-79}, |ΔΠ|=0.42). Both expressed transcripts with a canonical 3'UTR, longer than the 3'UTR expressed in distal cells (**Fig. 3d**). Thus, in two cases, tumor-associated stromal cells overexpressed and used longer collagen matrix isoforms than their distal counterparts. Another top differentially expressed isoform in TME vs. distal mesothelial cells was gelsolin (*GSN*), which exists in two main protein variants: one residing in the cytoplasm (*cGSN*), the other in the extracellular (plasma) environments (*pGSN*)[36]. At the gene level, *GSN* was not significantly overexpressed in TME vs. distal or in HGSOC vs. non-HGSOC cells. However, TME mesothelial cells had a significantly higher *cGSN/pGSN* isoform ratio than distal ones (P_{corr}=2.49x10^{-18}, |ΔΠ|=0.34) (**Fig. 3e**). Similarly, cancer cells had a significantly higher *cGSN/pGSN* ratio than non-cancer

242 cells (P_{corr}=3.4x10^{-127}, |ΔΠ|=0.28), and consistent with findings for *COL1A*, TME cells displayed
243 a cancer-like isoform expression profile compared to cells from distal sites, suggesting tissue
244 mimicry. To test if the differential expression of those structural isoforms in TME cells could
245 be linked to epithelial-to-mesenchymal transition (EMT), we performed gene set enrichment,
246 which revealed the EMT pathway as enriched in TME mesothelial and fibroblasts cells (**Fig.**
247 **3f**) supporting the idea of a tumor-transformed stroma.

248 Differential isoform expression in cancer reveals isoform-specific *IGF1* 249 usage

250 HGSOC cells significantly expressed different isoforms in 17% of the genes, compared to
251 all distal cells, but only 0.6% were switched with |ΔΠ|>0.5 (6,841 genes tested, **Methods**)
252 (**Extended Data Fig. 4a).** One of the most significant switches was found in the insulin-like
253 growth factor gene *IGF1* (P_{corr}=1.1x10^{-130}, |ΔΠ|=0.68), a gene coding for a hormone linked to
254 the development, progression, survival, and chemoresistance of many cancer types including
255 ovarian cancer[37]. Cancer cells from all patients almost exclusively used the second exon of
256 the gene as their transcription start site (Class II isoform), whereas other cells mainly used the
257 first exon (Class I isoform)[38] (**Fig. 4a,b**). The Class II isoform was highly expressed in HGSOC,
258 with 95% of cancer cells expressing it (**Fig. 4c,d**). Reflecting the findings of the DIE analysis
259 in mesothelial cells, fibroblasts and mesothelial cells in the TME also expressed a higher
260 fraction of class II isoforms than cells derived from distal biopsies (**Fig. 4d**). *IGF1* was found
261 to be significantly higher expressed in cancer cells (P_{corr}=4.8x10^{-32}) as well as in TME
262 mesothelial cells and fibroblasts compared to distal mesothelial cells and fibroblasts
263 (P_{corr}=4.05x10^{-32}).
264
265 Similarly, cancer and TME cells differentially expressed multiple isoforms in the two actin-
266 associated tropomyosin genes *TPM1* and *TPM2*. Cancer cells expressed terminal exon 9a
267 and exon 6b of TPM2 (P_{corr}<10^{-293}, |ΔΠ|=0.28), and TME cells also expressed those exons
268 more than distal ones (**Extended Data Fig. 4a-d**). Cancer cells also preferentially expressed
269 exon 1b and 6a of *TPM1* (**Extended Data Fig. 4e**). Another strongly switched gene in cancer
270 cells is vesicle-associated *VAMP5* (P_{corr}=4.59x10^{-17}, |ΔΠ|=0.70). Indeed, the overexpressed
271 isoforms in HGSOC cells were a (predicted protein-coding) *VAMP8-VAMP5* read-through
272 gene, i.e., a novel gene formed of two adjacent genes (**Extended Data Fig. 4f**). HGSOC cells
273 expressed almost no wild-type (wt) *VAMP5* but had a significantly higher *VAMP8* expression
274 than other cells (P_{corr}=1.0x10^{-15}t), indicating that this read-through gene was under
275 transcriptional control of *VAMP8*. Amongst others, HGSOC cells also differentially expressed
276 isoforms in the Golgi vesicle-associated *AP1S2* gene (P_{corr}=6.52x10^{-97}, |ΔΠ|=0.60).
277 Fibroblasts, mesothelial, and myeloid cells expressed the canonical isoform (Uniprot: P56377-
278 1), whereas HGSOC cells used another terminal 3' exon (Uniprot: A0A5F9ZHW1) (**Extended**
279 **Data Fig. 4g**). Last, patient 2 cancer cells highly expressed a novel shortened isoform of
280 ceramide kinase gene *CERK*, (P_{corr}=1.38x10^{-39}, |ΔΠ|=0.78) (**Extended Data Fig. 4h**). In
281 summary, tumor cells showed differential isoform usage in genes associated with hormonal
282 (*IGF1*), actin (*TPM1, TPM2, GSN*), vesicle (*VAMP8-VAMP5, APS1A*), and sphingolipid
283 (*CERK*) functions.

Long-read sequencing captures gene fusions and identifies an *IGF2BP2::TESPA1* fusion that was misidentified in short-read data

To detect fusion transcripts, we aligned long reads to the reference genome and filtered for reads split-aligned across multiple genes. We then ranked fusion transcripts with counts across all cells of more than 10 UMIs (**Supplementary Table 2**). Out of the 34 detected fusion entries, 21 were genes fused with mitochondrial ribosomal RNA (*mt-rRNA1-2*) and ubiquitous among all cell types, 11 isoforms were *IGF2BP2*::*TESPA1* fusions specific to patient 2, one was a cancer cell-specific *CBLC* (chr8:43.064.215) fusion to a long non-coding RNA (lncRNA) expressed in patient 3, and one was a cancer cell-specific fusion of *FNTA* with a lncRNA expressed in patient 1. The ubiquitous *mt-rRNA* fusions were likely template-switching artifacts from the library preparation, as *rRNA* makes up to 80% of RNA in cells[39]. *IGF2BP2::TESPA1* was a highly expressed fusion event in patient 2: 2,174 long-reads mapped to both *IGF2BP2* (Chr3) and *TESPA1* (Chr12). The gene fusion consisted of 5' located exons 1-4 of *IGF2BP2,* corresponding to 112 amino acids (aa) and including the RNA recognition motif 1 (RRM1) and half of the RRM2 domain*,* linked to the terminal *TESPA1* 3' untranslated region (UTR) exon, encoding 69 aa as in-frame fusion and including no known domains (**Fig. 5a**). In total, the gene fusion encoded 181 aa, compared to 599 aa of wt *IGF2BP2* and 521 aa of wt *TESPA1* **(Fig. 5b)**. 98.9% of fusion reads were found in HGSOC cells and the fusion was detected in 86.8% of patient 2's cancer cells, making it a highly cancer cell- and patient-specific fusion event (**Fig. 5c**). Cancer cells lacking the gene fusion had lower overall UMI counts, suggesting low coverage as a possible reason for the absence of the gene fusion (**Fig. 5d**).

We next investigated the footprint of the gene fusion in the short-read data. The *TESPA1* gene was expressed uniquely in T cells and highly expressed only in patient 2, almost exclusively in HGSOC cells, and colocalized with *IGF2BP2* expression (**Fig. 5e,f**). In short-read data, *TESPA1* was the highest differentially expressed gene in cancer cells compared to non-cancer cells in patient 2 ($P_{corr}=1.17x10^{-14}$). Next, we designed a custom reference including the *IGF2BP2::TESPA1* transcriptomic breakpoint as well as wt *TESPA1* and wt *IGF2BP2* junctions and re-aligned Patient 2's short-reads (**Extended Data Fig. 5, Methods**). Out of the 989 reads mapping to the custom reference, 94% preferentially aligned to *IGF2BP2::TESPA1* (99.8% of those in HGSOC cells). This implies that the reported overexpression of *TESPA1* in short-reads is false, as nearly all junction reads map to the fusion and not the wt gene. Reads covering the *TESPA1* 3' UTR region harbored three heterozygous single nucleotide polymorphisms (hSNPs): chr12:54.950.144 A>T (rs1047039), chr12:54.950.240 G>A (rs1801876), and chr12:54.950.349 C>G (rs2171497). In long reads, wt *TESPA1* was either triple-mutated or not mutated at all, indicating two different alleles. All fusion long reads, however, were triple-mutated, indicating a genomic origin and monoallelic expression of the fusion (**Fig. 5g**). In short reads, the three loci were mutated in nearly all reads, supporting the hypothesis that the observed *TESPA1* expression represents almost completely *IGF2BP2::TESPA1* expression and that it has a genomic origin.

Genomic breakpoint validation of the *IGF2BP2::TESPA1* fusion

To validate that the *IGF2BP2::TESPA1* gene fusion is the result of genomic rearrangements, we looked for a breakpoint in single-cell DNA sequencing (scDNA-seq) data from a patient 2-matched metastatic sample. Two RNA fusion long reads mapped to intronic regions of *IGF2BP2* and *TESPA1* (**Extended Data Fig. 5**) indicating the location of the breakpoint at

329 chr3:185.604.020-chr12:54.960.603. We then estimated the scDNA-seq copy number
330 profiles of all cells and identified two clones among the 162 cells of the scDNA sample: a
331 cancer clone (Subclone 0) and a copy number-neutral non-cancer clone (Subclone 1)
332 (**Fig. 6a**). We next aligned the scDNA data to a custom reference covering the breakpoint
333 (**Methods, Supplementary Dataset 1**), including the wt *TESPA1*, wt *IGF2BP2,* and
334 *IGF2BP2::TESPA1* fusion sequences. We found nine reads mapping to the breakpoint (nine
335 in subclone 0 cancer cells, zero in sublone 1 cells, P=0.0321) **(Fig. 6b)**. We also found 14
336 reads mapping to wt *IGF2BP2* (ten in subclone 0 cells, four in sublone 1 cells, P=0,78) (**Fig.
337 6c**), and eight reads mapping to wt *TESPA1* (five subclone 0 cells, three subclone 1 cells,
338 P=1.0) (**Fig. 6d**). Thus, scDNA-seq data confirmed the breakpoint in the intronic region
339 detected by the long-read scRNA-seq. The scDNA-seq data also confirmed that the
340 *IGF2BP2::TESPA1* fusion was cancer-cell specific, as suggested by long-read scRNA-seq
341 data. *IGF2* RNA, which is bound by the wt IGF2BP2 protein, is also largely overexpressed in
342 patient 2 cancer cells compared to other patients ($P_{corr}<2.54x10^{-15}$). The genomic region
343 containing *IGF2BP2* has an increased copy number (**Fig. 6a**) in patient 2, so the fact that
344 one allele is a fusion allele does not impair the wt *IGF2BP2* transcription.

345 Discussion

346 Detecting genomic alterations such as mutations[40,41] and gene fusions[42,43] in combination with
347 isoform-level[15] transcriptomic readouts on the single-cell level can provide valuable
348 information on cancer formation, progression, the role of the TME, drug targets, and therapy
349 response[44]. Here, we applied PacBio HiFi high-throughput long-read RNA-seq on five omental
350 metastases and tumor-free samples from chemo-naive HGSOC patients to detect and quantify
351 all of these alterations.

352 Until now, a combination of single-cell short- and long-read sequencing was necessary to
353 identify cell-specific isoforms: the higher depth of short-read sequencing allowed for cell typing
354 based on gene expression, while long-read sequencing was used to identify isoforms[22].
355 Leveraging multiple strategies to generate high PacBio sequencing output, we achieved a 50-
356 fold increased sequencing depth compared to the first long-read PacBio scRNA-seq study[22]
357 allowing for short read-comparable cell type identification. Consequently, future studies with
358 similar or increased long-read throughput will not have to rely on parallel short-read
359 sequencing, thereby saving cost and labor.

360 Our analysis revealed a differential isoform usage between distal tumor-free and TME
361 mesothelial cells in extracellular matrix associated genes (*COL1A1, COL1A2, GSN*). A
362 geneset enrichment analysis between the two sites revealed higher EMT pathway enrichment
363 in TME-derived mesothelial cells and fibroblasts. These findings are consistent with increasing
364 evidence that EMT in the TME is induced by cancer cells, leading to cancer-associated
365 phenotypes[45] including TGFβ1-induced mesenchymal states of mesothelial cells in ovarian
366 cancer[46]. Notably, in *IGF1, TPM2, GSN* and *COL1A2* genes, we found overlap in isoform
367 usage between cancer and TME cells (fibroblasts and mesothelial cells). Whether this cancer
368 mimicry of the TME is caused by signaling or the result of mRNA exchange via tumor-secreted
369 extracellular vesicles[47], as it was shown for *GSN*[48], requires further investigation.

370 Additionally, we demonstrated the potential of the technology in terms of coverage and
371 sequencing accuracy to detect mutations and gene fusions. In particular, in one patient, the

372 novel fusion *IGF2BP2::TESPA1* was highly overexpressed compared to wt *IGF2BP2* (~10x
373 more) and *TESPA1* (~150x more). *IGF2BP2* is known to be regulated via 3'UTR miRNA
374 silencing[49], however the *IGF2BP2::TESPA1* fusion has the unregulated 3'UTR of *TESPA1*,
375 which could explain its overexpression. *TESPA1* is normally expressed in T cells[50] and long-
376 read data confirmed T cell-specific wt *TESPA1* expression. Short read data however
377 erroneously reported *TESPA1* as the most differentially expressed gene in cancer cells,
378 resulting from 3' end capture of the fusion transcripts. This highlights that short-read scRNA-
379 seq data fails to distinguish between gene and fusion expression, potentially leading to wrong
380 biological conclusions.
381
382 Overall, HGSOC cells revealed a profoundly modified IGF system in all patients, with a drastic
383 switch from *IGF1* Class I to Class II isoform, *IGF2* overexpression, and a highly expressed
384 *IGF2BP2* gene fusion in one patient. The *IGF* protein family promotes cancer growth, survival,
385 proliferation, and drug resistance through signaling via *PI3K-AKT* or *MAPK*, and is a known
386 clinical target in ovarian cancer[37]. Secreted (Class II) *IGF1* is associated with the progression
387 of ovarian cancer[51] and the observed overexpression of Class II IGF1 in HGSOC cells could
388 mediate uncontrolled cell proliferation in the tumor.

389 Although the achieved sequencing depth allowed for short-read independent cell typing and
390 clustering, a further increased depth is needed to capture low abundance transcripts. For
391 example, we did not obtain sufficient reads to retrieve and characterize the T cell receptor
392 repertoire. This is consistent with a long-read scRNA-seq study in blood lymphocytes that
393 reported a 3.6-fold lower pairing rate for T cell receptors than the higher abundant B cell
394 receptors from plasmablasts[52]. With further technological advances and decreased
395 sequencing costs, however, we expect that these limitations can and will be overcome.
396 Enrichment for low abundant transcripts for long-read sequencing or depletion of mitochondrial
397 and ribosomal RNA[53] represent interesting avenues forward.

398 Altogether, we demonstrate that long-read sequencing provides a more complete picture of
399 cancer-specific changes. These findings highlight the manifold advantages and new
400 opportunities that this technology provides to the field of precision oncology, opening the
401 premise of personalized drug prediction and neoantigen detection for cancer vaccines[54,55].

402 Materials and Methods

403 Omentum patient cohort

404 The use of material for research purposes was approved by the corresponding cantonal ethic
405 commissions (EKNZ: 2017–01900, to V.H.S.) and informed consent was obtained for all
406 human primary material. Tissue samples were immediately collected from the theater and
407 transferred on ice to the department of biomedicine of the University Hospital Basel for tissue
408 dissociation.

409 Sample processing

410 Fresh omentum and omental HGSOC tumor metastasis biopsy samples were cut into small
411 pieces and dissociated in digestion solution (1 mg/mL collagenase/Dispase [Sigma cat. no.
412 10269638001], 1 unit/mL DNase I [NEB, cat. no. M0303] and 10% FBS in DMEM [Sigma, cat.

413 no. D8437-500mL]) for 30 min at 37°C. To focus on the non-adipose cell fraction, adipocytes
414 were separated by centrifugation and the cell pellet was collected. Red blood cell lysis (RBC)
415 was performed using MACS red blood lysis solution (cat. no. 130-094-183). Then, the cell
416 pellet was resuspended into MACS dead cell removal microbeads (cat. no. 130-090-101) and
417 was loaded into the AutoMACS separator to remove dead cells. After counting cell number,
418 cells were resuspended in PBS with 1% BSA and transferred to the Genomics Facility Basel.
419 The cell suspension was again filtered and cell number and viability was assessed on a
420 Cellometer K2 Image Cytometer (Nexcelom Bioscience, cat. no. Cellometer K2) using
421 ViaStain AOPI Staining Solution (Nexcelom Bioscience, cat. no. CS2-0106-5mL) and PD100
422 cell counting slides (Nexcelom Bioscience, cat. no. CHT4-PD100-003). For samples with
423 viability below 70% and when cell numbers allowed (>10^5 cells total), apoptotic and dead cells
424 were removed by immunomagnetic cell separation using the Annexin Dead Cell Removal Kit
425 (StemCell Technologies, cat. no. 17899) and EasySep Magnet (StemCell Technologies, cat.
426 no. 18000). If the cell pellet appeared still red, additional RBC lysis was performed. Cells were
427 washed with a resuspension buffer (PBS with 0.04% BSA), spun down and resuspended in a
428 resuspension buffer. Finally, cells were again counted and their viability determined. The cell
429 concentration was set according to 10x Genomics protocols (700-1,200 cells/µL).

430 10x Genomics single-cell capture and short-read sequencing

431 Cell suspensions were loaded and processed using the 10x Genomics Chromium platform
432 with the 3P v3.1 kit on the 10x Genomics Chromium Single Cell Controller (10x Genomics,
433 PN-120263) according to the manufacturer's instructions. 500 or 1,000 cells were targeted per
434 lane. The quality of cDNA traces and GEX libraries were profiled on a 5200 Fragment Analyzer
435 (Agilent Technologies).
436 Paired-end sequencing was performed on the Illumina NovaSeq platform (100 cycles, 380pm
437 loading concentration with 1% addition of PhiX) at recommended sequencing depth (20,000-
438 50,000 reads/cell).

439 Long-read library preparation and PacBio sequencing

440 To increase long-read PacBio sequencing throughput, we followed the strategy of cDNA
441 concatenation of the HIT-scISOseq protocol[23] with the modification of two rounds of biotin-
442 PCR in order to further reduce template-switch oligo (TSO) artifacts from the data.
443 Full protocol details:

444 cDNA amplification and biotin-enrichment

445 15 ng of each patient's cDNA library were amplified using the KAPA HiFi HotStart Uracil+
446 ReadyMix 2x (Kapa Biosystems, cat. no. KK2801) with 0.5 µM final concentration of custom-
447 primers (Integrated DNA Technologies, HPLC purified). Primers contained overhang
448 sequences adapted from Hebelstrup *et al.*[56] with a single deocxyuredine (dU) residue at a 10
449 nt distance from the 5' terminus enabling USER enzyme digestion and creating single-
450 stranded overhangs. Generated PCR fragments thus contain a single dU residue per DNA
451 strand. The forward primer was specific to the 10x Genomics partial Read 1 sequence and
452 contained a biotin modification allowing for biotin enrichment of amplified full-length cDNA
453 molecules. The reverse primer was specific to the 10x Genomics partial TSO sequence.
454 Forward Primer: /5Biosg/AGGTCTTAA/ideoxyU/CTACACGACGCCTTCCGATCT
455 Reverse Primer: ATTAAGACC/ideoxyU/AAGCAGTGGTATCAACGCAGAG

456 The PCR was run according to the manufacturer's instruction with two cycles at an annealing
457 temperature of 63°C followed by 7 cycles at an annealing temperature of 67°C; annealing time
458 was 30 seconds. Extension was performed at 72°C for 90 seconds. PCR products were
459 purified at 0.6X SPRIselect bead cleanup (Beckman Coulter, cat. no. B23318) according to
460 the manufacturer's instructions and eluted in 22 µL EB buffer (Qiagen, cat. no. 19086). DNA
461 concentrations were measured using the Qubit dsDNA HS Assay Kit (Thermo Fisher
462 Scientific, cat. no. Q32854), which were in the range of 1.5 µg per sample. cDNA traces were
463 additionally evaluated on a 5200 Fragment Analyzer System (Agilent Technologies) using the
464 HS NGS Fragment Kit, 1-6000 bp (Agilent, cat. no. DNF-474-0500). Full-length cDNAs were
465 enriched through capture on 5 µL streptavidin-coated M-280 dynabeads using the
466 Dynabeads™ kilobaseBINDER™ Kit (Invitrogen, cat. no. 60101), thus depleting TSO-TSO
467 artifacts. Washed Dynabeads containing the DNA-complexes were directly resuspended in 20
468 µL USER reaction buffer containing 10 µL StickTogether DNA Ligase Buffer 2x (NEB, cat. no.
469 B0535S), 1.5 µL USER Enzyme (NEB, cat. no. M5505S) and 8.5 µL Nuclease-free water
470 (Invitrogen, AM9939) and incubated in a thermocycler at 37°C for 20 min and held at 10°C (no
471 annealing). This created a nick at the deoxyuracil site forming palindrome overhangs and
472 releasing the biotin-bound DNA molecules from the beads. Beads were removed by magnetic
473 separation and the supernatant with the biotin-released cleaved PCR products was subjected
474 to a 0.6X SPRIselect cleanup step. Approximately 100 ng of purified product per sample were
475 split into two aliquots and subjected to a second PCR amplification step with 6 cycles using
476 an annealing temperature of 67°C. Reactions were pooled, purified by 0.6X SPRIselect
477 cleanup and quality checked on both Qubit and Fragment Analyzer. Total DNA yield was
478 between 5-8 µg, which were subjected to a second round of streptavidin-purification using 10
479 µL of beads.

480 Transcript ligation

481 Beads were incubated in 19 µL USER reaction buffer at 37°C for 20 min for USER digestion
482 and 25°C for 17 min for overhang annealing. Beads were then removed by magnetic
483 separation and the supernatant was transferred to a new PCR tube. 1 µL of T4 DNA ligase
484 high-concentration (2,000,000, units/mL, NEB, cat. no. M0202T) was added, mixed and
485 incubated at 10°C for >24hrs and heat inactivated at 65°C for 10 min. To efficiently deplete
486 any non-ligated transcripts, 0.38X SPRIselect cleanup was performed, eluted in 20 µL EB
487 buffer and traces were evaluated on the Fragment Analyzer using the HS Large Fragment kit
488 (Agilent Technologies, cat. no. DNF-492-0500) at 1:5 dilutions. Ligation products were 8-11kb
489 long; average yield was 100 ng per sample.

490 End repair/dA tailing, adapter ligation and PCR amplification

491 To enable PCR-amplification of the ligated construct, the NEBNext Ultra II DNA Library Prep
492 Kit for Illumina was followed (NEB, cat. no. E7645S) using total DNA yield as input material.
493 2.5 µL of 5 µM dT overhang adapter (Roche, cat. no. KK8727) were used for the End Prep
494 reaction. Adapter-ligated libraries were purified by 0.39X SPRIselect cleanup, eluted in 22 µL
495 EB buffer and products were evaluated by HS Large Fragment kit. Total yield of around 40 ng
496 was split in two and PCR amplified using 2X KAPA HiFi Hot-Start ReadyMix (Roche, cat. no.
497 KK2602) and KAPA Library Amplification Primer Mix (10X concentration, Roche, cat. no.
498 KK2623), 10 µL library input each with 11 cycles and 9 min extension time. Following a 0.38X
499 SPRIselect cleanup and elution in 48 µL EB buffer, products were evaluated on a large
500 fragment gel revealing an average fragment length of libraries of 4.6 kb and average total of

501 1.1 µg DNA. To increase total yield to 2 µg DNA required for SMRTbell library preparation of
502 a product with 5 kb amplicon size, the PCR was repeated with three additional cycles and
503 5 min extension time. After 0.4X SPRI cleanup and Fragment Analyzer inspection, the final
504 yield was 2 µg per library.

505 PacBio SMRTbell library preparation

506 The SMRTbell Express Template Kit (PacBio, cat. no. 100-938-900) was used following
507 manufacturer's instructions for DNA damage repair, end repair/dA-tailing and ligation of a
508 hairpin adapter (double amount used). Final purification of the SMRTbell template was
509 performed by 0.42X SPRIselect cleanup and elution in 43 µL EB buffer. Exonuclease
510 treatment was performed by addition of 5 µL of NEBbuffer1 (NEB, cat. no. B7001S) and 1 µL
511 of each Exonuclease I (NEB, cat. no. M0293S) and Exonuclease III (NEB, cat. no. M0206S)
512 bringing the total volume to 50 µL per reaction. Enzyme treatment was performed at 37°C for
513 60 min. After SPRIselect cleanup, products were quantified on a large fragment gel at 1:30
514 dilution. Final yield was approximately 650 ng per sample, a sufficient amount for long-read
515 sequencing.

516 PacBio Sequel II sequencing

517 Libraries were sequenced on the PacBio Sequel II platform with the SMRT cell 8M. Omentum
518 metastasis and tumor-free omentum were run on three and two 8M cells, respectively.

519 Single-cell DNA-sequencing

520 Cell suspensions were loaded and processed using the 10x Genomics Chromium platform
521 with the single-cell CNV kit on the 10x Genomics Chromium Single Cell Controller (10x
522 Genomics, PN-120263) according to the manufacturer's instructions. Paired-end sequencing
523 was performed on the Illumina NovaSeq platform (100 cycles, 380pm loading concentration
524 with 1% addition of PhiX) at recommended sequencing depth.

525 Data Analysis

526 Short-read data analysis

527 Preprocessing

528 Raw reads were mapped to the GRCh38 reference genome using 10x Genomics Cell Ranger
529 3.1.0 to infer read counts per gene per cell. We performed index-hopping removal using a
530 method developed by Griffiths *et al.*[57].

531 10x Genomics short-read analysis

532 GEX data of each sample was analyzed using the scAmpi workflow[58]. In brief, UMI counts
533 were quality controlled and cells and genes filtered to remove known contaminants. Cells
534 where over 50% of the reads mapped to mitochondrial genes and cells with fewer than 400
535 different expressed genes were removed, as well as non protein-coding genes and genes that
536 were expressed in less than 20 cells. Doublet detection was performed using scDblFinder[59].
537 Subsequently, counts were normalized and corrected for cell cycle effects, library size, and
538 sample effect using sctransform[60]. Similar cells were grouped based on unsupervised

539 clustering using Phenograph[61] and an automated cell type classification was performed
540 independently for each cell[62] using gene lists defining highly expressed genes in different cell
541 types from previous publications. Major cell type marker lists were developed in-house based
542 on unpublished datasets (manuscripts in preparation) including the Tumor Profiler Study[63]
543 using the Seurat FindMarkers method[64]. Immune subtype marker gene lists were obtained
544 from Newman *et al.*[65], enriched with T cell subtypes from Sade-Feldman *et al.*[66]

545 Long-read data analysis

546 Generating CCS

547 Using SMRT-Link (version 9.0.0.92188), we performed circular consensus sequencing (CCS)
548 with the following modified parameters: maximum subread length 50,000 bp, minimum
549 subread length 10 bp, and minimum number of passes 3.

550 Unconcatenating long reads

551 We used NCBI BLAST (version 2.5.0+) to map the 5' and 3' primers to CCS constructs, with
552 parameters: "-outfmt 7 -word_size 5" as described previously[23]. Sequences between two
553 successive primers were used as input for primer trimming using IsoSeq3 Lima (parameters:
554 --isoseq --dump-clips --min-passes 3). Cell barcodes and UMIs were then demultiplexed using
555 IsoSeq3 tag with parameter --design T-12U-16B. Finally, we used IsoSeq3 refine with option
556 --require-polya to remove concatemers and trim polyA tails. Only reads with a correct 5'-3'
557 primer pair, a barcode also found in the short-read data, a UMI, and a polyA tail were retained.

558 Isoform classification

559 Demultiplexing UMIs with IsoSeq3 dedup and calling isoforms on the cohort level with
560 collapse_isoforms_by_sam.py resulted in unfeasible runtimes. Therefore, we called isoforms
561 first on the cell level as a pre-filtering step. Long-reads were split according to their cell
562 barcodes, and UMI deduplication was performed using IsoSeq3 dedup. Next, reads were
563 mapped and aligned to the reference genome (hg38) using minimap2 with parameters: -ax
564 splice -uf --secondary=no -C5. Identical isoforms were merged based on their aligned exonic
565 structure using collapse_isoforms_by_sam.py with parameters: -c 0.99 -i 0.95 --
566 gen_mol_count. We then classified isoforms using SQANTI3 [31] with arguments: --skipORF --
567 fl_count --skip_report. We finally filtered artifacts including intrapriming (accidental priming of
568 pre-mRNA 'A's), reverse-transcriptase template switching artifacts, and mismapping to non-
569 canonical junctions. In order to have a unique isoform catalog for all our samples, we then
570 retained only reads associated to isoforms passing the SQANTI3 filter, and we ran
571 collapse_isoforms_by_sam.py, SQANTI3 classification and filtering again on all cells together.
572 The described pipeline is available here and was implemented in Snakemake, a reproducible
573 and scalable workflow management system[67].

574 3' and 5' isoform filtering

575 For SQANTI3-defined isoforms, incomplete splice match, novel in catalog and novel not in
576 catalog, we only retained isoforms falling within 50 bp of a CAGE-validated transcription start
577 site (FANTOM5 CAGE database), and 50 bp of a polyA site form the PolyASite database[33].
578 The GENCODE database was used as a comparison, all protein-coding isoforms were
579 grouped under the GENCODE.full label, a subset including only full-length isoforms was

580 labeled as GENCOD.FL, and the Matched Annotation from NCBI and EMBL-EBI (MANE[34])
581 was named GENCODE.MANE.

582 Isoforms biotypes

583 Novel isoform biotypes were assessed internally by the GENCODE team with biotypes
584 matching those described by Frankish *et al*.[68].

585 Cell type-specific isoforms

586 Considering only the SQANTI3-defined 'full splice match', 'novel not in catalog' and 'novel in
587 catalog' isoforms with at least 3 reads, we established the following classification: "Cell-
588 specific" isoforms are present in only 1 cell and "cell type specific" isoforms are present in >=3
589 cells of an unique cell type.
590
591 Cell type annotation
592
593 Cells were annotated with long-reads the same way as short-reads, using scROSHI. The
594 major cell types were modified according to gene expression in long-reads. Immune subtype
595 marker gene lists were unchanged.

596 Mutation detection

597 Positions of mutations from Foundation Medicine's targeted NGS panel (Foundation One CDx)
598 mutations described in Table 1 were used as reference. One mutation not present in the list,
599 TP53_P151H, was visually detected in Patient 1 and added to the list. If a position was
600 mutated at least in one cell belonging to a distal biopsy sample, the mutation was classified
601 as a germline variant. Cells with one mutated read in one of the positions were considered
602 mutated.

603 Differential isoform tests

604 Differential isoform testing was performed using a $\chi 2$ test as previously described in
605 Scisorseqr[25]. Briefly, counts for each isoform ID were assigned to individual cell types, and
606 genes were discarded if they did not reach sufficient depth per condition (25 reads per
607 condition per gene). P-values from a $\chi 2$ test for differential isoform usage were computed per
608 gene where a sufficient depth was reached, and we corrected for multiple testing using
609 Benjamini Hochberg correction with a 5% false discovery rate. If the corrected p-value was
610 ≤ 0.05 and the sum of change in the relative percent of isoform ($\Delta\Pi$) of the top two isoforms in
611 either positive or negative direction was more than 10%, then the gene was called differentially
612 spliced. To classify the top differentially spliced genes, we took the rank of genes by $\Delta\Pi$ and
613 corrected p-values, and summed those two ranks. The smallest sum of ranks were considered
614 as the top differentially expressed genes. Differentially used isoforms were visualized using
615 ScisorWiz[69].

Pathway enrichment analysis

We used GSVA to perform pathway enrichment analysis. Gene sets were obtained from the default scAmpi workflow[70], with the addition of the EPITHELIAL_MESENCHYMAL_TRANSITION pathway from GSEA.

Fusion Discovery

Mapped reads from isoform classification were pooled. We called reads mapping to two separate genes at a distance of more than 100,000 bp or to different chromosomes using fusion_finder.py (cDNA_Cupcake package, https://github.com/Magdoll/cDNA_Cupcake) with parameters --min_locus_coverage_bp 200 -d 1000000. Fusion isoforms with sufficient depth (min. 10 reads) were kept, and their breakpoint, expression per cell type and number of cells in which they are expressed was assessed.

Short-reads re-alignment to *IGF2BP2::TESPA1*

We designed a custom reference including *IGF2BP2::TESPA1* transcriptomic breakpoint as well as the wild-type *IGF2BP2* and *TESPA1* exon junction covering the breakpoint. The reference was composed of 5 sequences of 80 nucleotides (40 bases upstream and downstream of the breakpoint), sequences XXX_1 and XXX_2 represent the breakpoints of the two main isoforms seen in each gene:

```
>TESPA1_wt_1
TTCTGTCAGACCACATGCTGTTGTGGTGGTGGAGAAAGCAATTCTGGAGGCTGGCAAATCCAAG
GTCAAAAGCCTGCA

>TESPA1_wt_2
TTCTGTCAGACCACATGCTGTTGTGGTGGTGGAGAAAGCTTCACGAGTCTTTGCCAGCAAAAGTC
TGGTGGTGGTGGG

>IGF2BP2_wt_1
ATGTGACGTTGACAACGGCGGTTTCTGTGTCTGTGTTGACTTGTTCCACATTCTCCACTGTCCCA
TATTGAGCCAAAA

>IGF2BP2_wt_2
ATCACTGGATTGTGTGTTCTTCTGAATTACTTCTTTAGGCTTGTTCCACATTCTCCACTGTCCCAT
ATTGAGCCAAAA

>TESPA1_IGF2BP2_fusion_1
TTCTGTCAGACCACATGCTGTTGTGGTGGTGGAGAAAGCCTTGTTCCACATTCTCCACTGTCCCA
TATTGAGCCAAAA

>TESPA1_IGF2BP2_fusion_2
CAAATCCAAGGTCAAAAGCCTGCATCTGGTGAGGGCCTCCTTGTTCCACATTCTCCACTGTCCCA
TATTGAGCCAAAA
```

Patient 2 reads were aligned to this reference using minimap2 with parameters: -ax sr --secondary=no. Reads mapping unambiguously to one of those reference sequences were then attributed to the cell type to which their cell barcode belonged.

scDNA analysis

Cell Ranger DNA was used to demultiplex and align Chromium-prepared sequencing samples. We used the cellranger-dna mkfastq command to generate FASTQ files from the Illumina raw BCL files, and we ran the command cellranger-dna cnv to align FASTQ files to the hg38 reference genome, call cells, and estimate copy numbers. We obtained the copy number profiles and detected the main clonal structure of samples using SCICoNE[71].

DNA breakpoint validation

To validate in scDNA data breakpoints found in scRNA data, we used the putative scRNA breakpoint reads as a reference to re-align scDNA reads using BWA with options: -pt8 -CH. For the *IGF2BP2::TESPA1* fusion, the reference was composed of 3 sequences of 184 nucleotides (92 bases upstream and downstream of the breakpoint):

```
>IGF2BP2_WT
CAAACTTGTAGAAATGTGAATTTTTCTTGTTATTTTACAAGATTTGCAAAGGGACCTGAGACCCCG
AAAAGCTTAAGGACTACTGTTAAAAATACTGTTTGTTAAATAACTTTAAAGCAGCTGCAGCCTTTAT
GGGTTGCAGGGAGTTGTATGTAATGCTCAGAAAGAGCTGCCACTGAGAAT

>TESPA1_WT
TTCAATGATGTGGGCTGATTAGAACATAGCTGAAAGCAGGTGTTGGGATATTGATTTCCATGGCT
GGTCCTCACCTGTTACAAAACTTCTACTACAATGAGTTTCAAACTTCAATATGCAATCAATTATCTA
ACCTAAAGATCTTGGTAAAACTGTGATTCATTAGGTCTGGGGTGGGGGCTG

>IGF2BP2_TESPA1_Fusion
TTCAATGATGTGGGCTGATTAGAACATAGCTGAAAGCAGGTGTTGGGATATTGATTTCCATGGCT
GGTCCTCACCTGTTACAAAACTTCTACTACTGTTTGTTAAATAACTTTAAAGCAGCTGCAGCCTTT
ATGGGTTGCAGGGAGTTGTATGTAATGCTCAGAAAGAGCTGCCACTGAGAAT
```

Reads mapping unambiguously to one of those reference sequences were then attributed to the clone to which their cell barcode belonged.

Data and code availability

The raw sequencing files reported in this study have been deposited in the European Genome-phenome Archive (EGA) under the accession number EGAS00001006807. The software used to analyze the data of this study has been deposited at the GitHub repository: https://github.com/cbg-ethz/scIsoPrep

Acknowledgements

We thank Ina Nissen (Genomics Facility Basel) for technical support with PacBio sequencing, Ching-Yeu Liang (University Hospital Basel) for assistance with sample dissociation, and Anne Bertolini for her help with the scAmpi R package. We also wish to thank Elisabeth Tseng (Pacific Bioscience) for her help with cDNA_Cupcake, SQANTI3 and long-read sequencing analysis, Anoushka Joglekar (Weill Cornell Medicine) for her help with ScisorSeq and ScisorWiz as well as Lucia Csepregi for her help with immune repertoire analysis. Furthermore, we are grateful to Adam Frankish (Wellcome Sanger Institute) and the GENCODE Project team for their help with the manual annotation of transcripts. Graphical

illustrations were created with BioRender.com. Illumina sequencing was carried out in the Genomics Facility Basel of the University of Basel and the Department of Biosystems Science and Engineering, ETHZl. PacBio sequencing was done in the Functional Genomics Center Zurich of the University of Zurich and ETHZ, and the Lausanne Genomic Technologies Facility, University of Lausanne.

Funding Information

Part of this work was funded by the SNSF SPARK grant #190413, the grant #2017-510 of the Strategic Focal Area “Personalized Health and Related Technologies (PHRT)” of the ETH Domain, and the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement #766030.

Author contributions

UL and CB acquired funding and conceived and designed the experiments. VHS provided patient material. UL, FJ, AD, and CB selected the clinical cohort. UL performed 10x Genomics sample processing as well as short-read and long-read sequencing library preparation. The Tumor Profiler Consortium provided scDNA-seq data and NGS panel results. AD designed the analysis pipeline, and implemented it with the help of NBo. AD conducted all computational analyses. FS assisted in short-read scRNA-seq analysis. AD, UL, FJ, NBe, and CB interpreted the data. AD and UL wrote the manuscript with contributions of all authors. All authors read and approved the final manuscript.

Tumor Profiler Consortium authors list

Rudolf Aebersold[2], Melike Ak[28], Faisal S Al-Quaddoomi[9,17], Silvana I Albert[7], Jonas Albinus[7], Ilaria Alborelli[24], Sonali Andani[6,17,26,31], Per-Olof Attinger[11], Marina Bacac[16], Daniel Baumhoer[24], Beatrice Beck-Schimmer[39], Niko Beerenwinkel[4,17], Christian Beisel[4], Lara Bernasconi[27], Anne Bertolini[9,17], Bernd Bodenmiller[8,35], Ximena Bonilla[6,17,26], Lars Bosshard[9,17], Byron Calgua[24], Ruben Casanova[35], Stéphane Chevrier[35], Natalia Chicherova[9,17], Ricardo Coelho[18], Maya D'Costa[10], Esther Danenberg[37], Natalie Davidson[6,17,26], Monica-Andreea Drăgan[4], Reinhard Dummer[28], Stefanie Engler[35], Martin Erkens[14], Katja Eschbach[4], Cinzia Esposito[37], André Fedier[18], Pedro Ferreira[4], Joanna Ficek[6,17,26], Anja L Frei[31], Bruno Frey[13], Sandra Goetze[7], Linda Grob[9,17], Gabriele Gut[37], Detlef Günther[5], Martina Haberecker[31], Pirmin Haeuptle[1], Viola Heinzelmann-Schwarz[18,23], Sylvia Herter[16], Rene Holtackers[37], Tamara Huesser[16], Alexander Immer[6,12], Anja Irmisch[28], Francis Jacob[18], Andrea Jacobs[35], Tim M Jaeger[11], Katharina Jahn[4], Alva R James[6,17,26], Philip M Jermann[24], André Kahles[6,17,26], Abdullah Kahraman[17,31], Viktor H Koelzer[31], Werner Kuebler[25], Jack Kuipers[4,17], Christian P Kunze[22], Christian Kurzeder[21], Kjong-Van Lehmann[6,17,26], Mitchell Levesque[28], Ulrike Lischetti[18], Sebastian Lugert[10], Gerd Maass[13], Markus G Manz[30], Philipp Markolin[6,17,26], Martin Mehnert[7], Julien Mena[2], Julian M Metzler[29], Nicola Miglino[1], Emanuela S Milani[7], Holger Moch[31], Simone Muenst[24], Riccardo Murri[38], Charlotte KY Ng[24,34], Stefan Nicolet[24], Marta Nowak[31], Monica Nunez Lopez[18], Patrick GA Pedrioli[3], Lucas Pelkmans[37], Salvatore Piscuoglio[18,24], Michael Prummer[9,17], Natalie Rimmer[18], Mathilde Ritter[18], Christian Rommel[14], María L Rosano-González[9,17], Gunnar Rätsch[3,6,17,26], Natascha Santacroce[4], Jacobo Sarabia del Castillo[37], Ramona Schlenker[15],

Petra C Schwalie[14], Severin Schwan[11], Tobias Schär[4], Gabriela Senti[27], Wenguang Shao[7], Franziska Singer[9,17], Sujana Sivapatham[35], Berend Snijder[2,17], Bettina Sobottka[31], Vipin T Sreedharan[9,17], Stefan Stark[6,17,26], Daniel J Stekhoven[9,17], Tanmay Tanna[4,6], Alexandre PA Theocharides[30], Tinu M Thomas[6,17,26], Markus Tolnay[24], Vinko Tosevski[16], Nora C Toussaint[9,17], Mustafa A Tuncel[4,17], Marina Tusup[28], Audrey Van Drogen[7], Marcus Vetter[20], Tatjana Vlajnic[24], Sandra Weber[27], Walter P Weber[19], Rebekka Wegmann[2], Michael Weller[33], Fabian Wendt[7], Norbert Wey[31], Andreas Wicki[30,36], Mattheus HE Wildschut[2,30], Bernd Wollscheid[7], Shuqing Yu[9,17], Johanna Ziegler[28], Marc Zimmermann[6,17,26], Martin Zoche[31], Gregor Zuend[32]

[1]Cantonal Hospital Baselland, Medical University Clinic, Rheinstrasse 26, 4410 Liestal, Switzerland, [2]ETH Zurich, Department of Biology, Institute of Molecular Systems Biology, Otto-Stern-Weg 3, 8093 Zurich, Switzerland, [3]ETH Zurich, Department of Biology, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland, [4]ETH Zurich, Department of Biosystems Science and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland, [5]ETH Zurich, Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 1-5/10, 8093 Zurich, Switzerland, [6]ETH Zurich, Department of Computer Science, Institute of Machine Learning, Universitätstrasse 6, 8092 Zurich, Switzerland, [7]ETH Zurich, Department of Health Sciences and Technology, Otto-Stern-Weg 3, 8093 Zurich, Switzerland, [8]ETH Zurich, Institute of Molecular Health Sciences, Otto-Stern-Weg 7, 8093 Zurich, Switzerland, [9]ETH Zurich, NEXUS Personalized Health Technologies, John-von-Neumann-Weg 9, 8093 Zurich, Switzerland, [10]F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland, [11]F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland, , [12]Max Planck ETH Center for Learning Systems, , [13]Roche Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany, [14]Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, Grenzacherstrasse 124, 4070 Basel, Switzerland, [15]Roche Pharmaceutical Research and Early Development, Roche Innovation Center Munich, Roche Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany, [16]Roche Pharmaceutical Research and Early Development, Roche Innovation Center Zurich, Wagistrasse 10, 8952 Schlieren, Switzerland, [17]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland, [18]University Hospital Basel and University of Basel, Department of Biomedicine, Hebelstrasse 20, 4031 Basel, Switzerland, [19]University Hospital Basel and University of Basel, Department of Surgery, Brustzentrum, Spitalstrasse 21, 4031 Basel, Switzerland, [20]University Hospital Basel, Brustzentrum & Tumorzentrum, Petersgraben 4, 4031 Basel, Switzerland, [21]University Hospital Basel, Brustzentrum, Spitalstrasse 21, 4031 Basel, Switzerland, [22]University Hospital Basel, Department of Information- and Communication Technology, Spitalstrasse 26, 4031 Basel, Switzerland, [23]University Hospital Basel, Gynecological Cancer Center, Spitalstrasse 21, 4031 Basel, Switzerland, [24]University Hospital Basel, Institute of Medical Genetics and Pathology, Schönbeinstrasse 40, 4031 Basel, Switzerland, [25]University Hospital Basel, Spitalstrasse 21/Petersgraben 4, 4031 Basel, Switzerland, [26]University Hospital Zurich, Biomedical Informatics, Schmelzbergstrasse 26, 8006 Zurich, Switzerland, [27]University Hospital Zurich, Clinical Trials Center, Rämistrasse 100, 8091 Zurich, Switzerland, [28]University Hospital Zurich, Department of Dermatology, Gloriastrasse 31, 8091 Zurich, Switzerland, [29]University Hospital Zurich, Department of Gynecology, Frauenklinikstrasse 10, 8091 Zurich, Switzerland, [30]University Hospital Zurich, Department of Medical Oncology and Hematology, Rämistrasse 100, 8091 Zurich, Switzerland, [31]University Hospital Zurich, Department of Pathology and Molecular Pathology, Schmelzbergstrasse 12, 8091 Zurich, Switzerland,

[32]University Hospital Zurich, Rämistrasse 100, 8091 Zurich, Switzerland, [33]University Hospital and University of Zurich, Department of Neurology, Frauenklinikstrasse 26, 8091 Zurich, Switzerland, [34]University of Bern, Department of BioMedical Research, Murtenstrasse 35, 3008 Bern, Switzerland, [35]University of Zurich, Department of Quantitative Biomedicine, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [36]University of Zurich, Faculty of Medicine, Zurich, Switzerland, [37]University of Zurich, Institute of Molecular Life Sciences, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [38]University of Zurich, Services and Support for Science IT, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [39]University of Zurich, VP Medicine, Künstlergasse 15, 8001 Zurich, Switzerland

Conflict of interest

The authors declare no competing interests.

ORCiDs

Ulrike Lischetti: 0000-0002-9956-3043

Arthur Dondi: 0000-0003-3234-2550

Francis Jacob: 0000-0002-0446-1942

Christian Beisel: 0000-0001-5360-2193

Niko Beerenwinkel: 0000-0002-0573-6119

Franziska Singer: 0000-0002-6017-1595

Nico Borgsmüller: 0000-0003-4073-3877

Viola Heizelmann: 0000-0002-4056-3225

References

1. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
2. Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer Discov.* **12**, 31–46 (2022).
3. Hertzman Johansson, C. & Egyhazi Brage, S. BRAF inhibitors in cancer therapy. *Pharmacol. Ther.* **142**, 176–182 (2014).
4. Li, J. *et al.* A functional genomic approach to actionable gene fusions for precision oncology. *Sci. Adv.* **8**, eabm2382 (2022).
5. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69–74 (2015).
6. Yu, Y.-P. *et al.* Identification of recurrent fusion genes across multiple cancer types. *Sci.*

826 *Rep.* **9**, 1074 (2019).
827 7. Bower, H. *et al.* Life expectancy of patients with chronic myeloid leukemia approaches
828 the life expectancy of the general population. *J. Clin. Oncol.* **34**, 2851–2857 (2016).
829 8. Khan, M. *et al.* ALK inhibitors in the treatment of ALK positive NSCLC. *Front. Oncol.* **8**,
830 557 (2018).
831 9. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**,
832 415–421 (2013).
833 10. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**,
834 1134–1140 (2013).
835 11. Gao, Q. *et al.* Driver fusions and their implications in the development and treatment of
836 human cancers. *Cell Rep.* **23**, 227-238.e3 (2018).
837 12. Wang, Y., Shi, T., Song, X., Liu, B. & Wei, J. Gene fusion neoantigens: Emerging targets
838 for cancer immunotherapy. *Cancer Lett.* **506**, 45–54 (2021).
839 13. Wei, Z. *et al.* The Landscape of Tumor Fusion Neoantigens: A Pan-Cancer Analysis.
840 *iScience* **21**, 249–260 (2019).
841 14. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126**, 37–47
842 (2006).
843 15. Bonnal, S. C., López-Oreja, I. & Valcárcel, J. Roles and mechanisms of alternative
844 splicing in cancer - implications for care. *Nat. Rev. Clin. Oncol.* **17**, 457–474 (2020).
845 16. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from
846 8,705 Patients. *Cancer Cell* **34**, 211-224.e6 (2018).
847 17. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim.*
848 *Biophys. Acta* **1805**, 105–117 (2010).
849 18. Whiteside, T. L. The tumor microenvironment and its role in promoting tumor growth.
850 *Oncogene* **27**, 5904–5912 (2008).
851 19. Hedlund, E. & Deng, Q. Single-cell RNA sequencing: Technical advancements and
852 biological applications. *Mol. Aspects Med.* **59**, 36–46 (2018).
853 20. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read

854 sequencing. *Nature* **608**, 353–359 (2022).
855 21. Philpott, M. *et al.* Nanopore sequencing of single-cell transcriptomes with scCOLOR-seq.
856 *Nat. Biotechnol.* **39**, 1517–1520 (2021).
857 22. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands
858 of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).
859 23. Zheng, Y.-F. *et al.* HIT-scISOseq: High-throughput and High-accuracy Single-cell Full-
860 length Isoform Sequencing for Corneal Epithelium. *BioRxiv* (2020)
861 doi:10.1101/2020.07.27.222349.
862 24. Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmable
863 cDNA concatenation. *BioRxiv* (2021) doi:10.1101/2021.10.01.462818.
864 25. Joglekar, A. *et al.* A spatially resolved brain region- and cell type-specific isoform atlas of
865 the postnatal mouse brain. *Nat. Commun.* **12**, 463 (2021).
866 26. Hardwick, S. A. *et al.* Single-nuclei isoform RNA sequencing unlocks barcoded exon
867 connectivity in frozen brain tissue. *Nat. Biotechnol.* **40**, 1082–1092 (2022).
868 27. Veiga, D. F. T. *et al.* A comprehensive long-read isoform analysis platform and
869 sequencing resource for breast cancer. *Sci. Adv.* **8**, eabg6711 (2022).
870 28. Namba, S. *et al.* Transcript-targeted analysis reveals isoform alterations and double-hop
871 fusions in breast cancer. *Commun. Biol.* **4**, 1320 (2021).
872 29. Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian
873 carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
874 30. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian
875 carcinoma. *Nature* **474**, 609–615 (2011).
876 31. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript
877 sequences for quality control in full-length transcriptome identification and quantification.
878 *Genome Res.* (2018) doi:10.1101/gr.222976.117.
879 32. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for
880 GRCh38 and GRCm38 genome assemblies. *Sci. Data* **4**, 170107 (2017).
881 33. Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3'

882 end sequencing. *Nucleic Acids Res.* **48**, D174–D179 (2020).
883 34. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and
884 research. *Nature* **604**, 310–315 (2022).
885 35. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat.*
886 *Biotechnol.* **37**, 38–44 (2018).
887 36. García-Bartolomé, A. *et al.* Altered Expression Ratio of Actin-Binding Gelsolin Isoforms
888 Is a Novel Hallmark of Mitochondrial OXPHOS Dysfunction. *Cells* **9**, (2020).
889 37. Liefers-Visser, J. A. L., Meijering, R. A. M., Reyners, A. K. L., van der Zee, A. G. J. & de
890 Jong, S. IGF system targeted therapy: Therapeutic opportunities for ovarian cancer.
891 *Cancer Treat. Rev.* **60**, 90–99 (2017).
892 38. Philippou, A., Maridaki, M., Pneumaticos, S. & Koutsilieris, M. The complexity of the IGF1
893 gene splicing, posttranslational modification and bioactivity. *Mol. Med.* **20**, 202–214
894 (2014).
895 39. Choy, J. Y. H., Boon, P. L. S., Bertin, N. & Fullwood, M. J. A resource of ribosomal RNA-
896 depleted RNA-Seq data from different normal adult and fetal human tissues. *Sci. Data* **2**,
897 150063 (2015).
898 40. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev.*
899 *Cancer* **20**, 555–572 (2020).
900 41. Tate, J. G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids*
901 *Res.* **47**, D941–D947 (2019).
902 42. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of gene
903 fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
904 43. Yang, W. *et al.* Immunogenic neoantigens derived from gene fusions stimulate T cell
905 responses. *Nat. Med.* **25**, 767–775 (2019).
906 44. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer
907 medicine. *Nat. Rev. Clin. Oncol.* **15**, 353–365 (2018).
908 45. Jing, Y., Han, Z., Zhang, S., Liu, Y. & Wei, L. Epithelial-Mesenchymal Transition in tumor
909 microenvironment. *Cell Biosci.* **1**, 29 (2011).

910 46. Kenny, H. A. *et al.* Mesothelial cells promote early ovarian cancer metastasis through
911 fibronectin secretion. *J. Clin. Invest.* **124**, 4614–4628 (2014).
912 47. Hoshino, A. *et al.* Tumour exosome integrins determine organotropic metastasis. *Nature*
913 **527**, 329–335 (2015).
914 48. Asare-Werehene, M. *et al.* The exosome-mediated autocrine and paracrine actions of
915 plasma gelsolin in ovarian cancer chemoresistance. *Oncogene* **39**, 1600–1616 (2020).
916 49. Yang, Y. *et al.* Tumor Suppressor microRNA-138 Suppresses Low-Grade Glioma
917 Development and Metastasis via Regulating IGF2BP2. *Onco Targets Ther* **13**, 2247–
918 2260 (2020).
919 50. Wang, D. *et al.* Tespa1 is involved in late thymocyte development through the regulation
920 of TCR-mediated signaling. *Nat. Immunol.* **13**, 560–568 (2012).
921 51. Brokaw, J. *et al.* IGF-I in epithelial ovarian cancer and its role in disease progression.
922 *Growth Factors* **25**, 346–354 (2007).
923 52. Volden, R. & Vollmers, C. Single-cell isoform analysis in human immune cells. *Genome*
924 *Biol.* **23**, 47 (2022).
925 53. Dutton, G. CRISPR-Cas9 Technology Cuts Clutter from Sequencing Libraries. *Genetic*
926 *Engineering & Biotechnology News* **41**, 24–25 (2021).
927 54. Lang, F., Schrörs, B., Löwer, M., Türeci, Ö. & Sahin, U. Identification of neoantigens for
928 individualized therapeutic cancer vaccines. *Nat. Rev. Drug Discov.* **21**, 261–282 (2022).
929 55. Lin, M. J. *et al.* Cancer vaccines: the next immunotherapy frontier. *Nat. Cancer* **3**, 911–
930 926 (2022).
931 56. Hebelstrup, K. H. *et al.* UCE: A uracil excision (USER)-based toolbox for transformation
932 of cereals. *Plant Methods* **6**, 15 (2010).
933 57. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and
934 removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 2667 (2018).
935 58. Bertolini, A. *et al.* scAmpi-A versatile pipeline for single-cell RNA-seq analysis from basics
936 to clinics. *PLoS Comput. Biol.* **18**, e1010097 (2022).
937 59. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet

938 identification in single-cell sequencing data using scDblFinder. *F1000Res.* **10**, 979
939 (2021).
940 60. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-
941 seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
942 61. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like
943 Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
944 62. Prummer, M. *et al.* scROSHI - robust supervised hierarchical identification of single cells.
945 *BioRxiv* (2022) doi:10.1101/2022.04.05.487176.
946 63. Irmisch, A. *et al.* The Tumor Profiler Study: integrated, multi-omic, functional tumor
947 profiling for clinical decision support. *Cancer Cell* **39**, 288–293 (2021).
948 64. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29
949 (2021).
950 65. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles.
951 *Nat. Methods* **12**, 453–457 (2015).
952 66. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint
953 Immunotherapy in Melanoma. *Cell* **175**, 998-1013.e20 (2018).
954 67. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine.
955 *Bioinformatics* **28**, 2520–2522 (2012).
956 68. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes.
957 *Nucleic Acids Res.* **47**, D766–D773 (2019).
958 69. Stein, A. N., Joglekar, A., Poon, C.-L. & Tilgner, H. U. ScisorWiz: Visualizing Differential
959 Isoform Expression in Single-Cell Long-Read Data. *BioRxiv* (2022)
960 doi:10.1101/2022.04.14.488347.
961 70. Bertolini, A. *et al.* scAmpi - A versatile pipeline for single-cell RNA-seq analysis from
962 basics to clinics. *BioRxiv* (2021) doi:10.1101/2021.03.25.437054.
963 71. Kuipers, J., Tuncel, M. A., Ferreira, P., Jahn, K. & Beerenwinkel, N. Single-cell copy
964 number calling and event history reconstruction. *BioRxiv* (2020)
965 doi:10.1101/2020.04.28.065755.

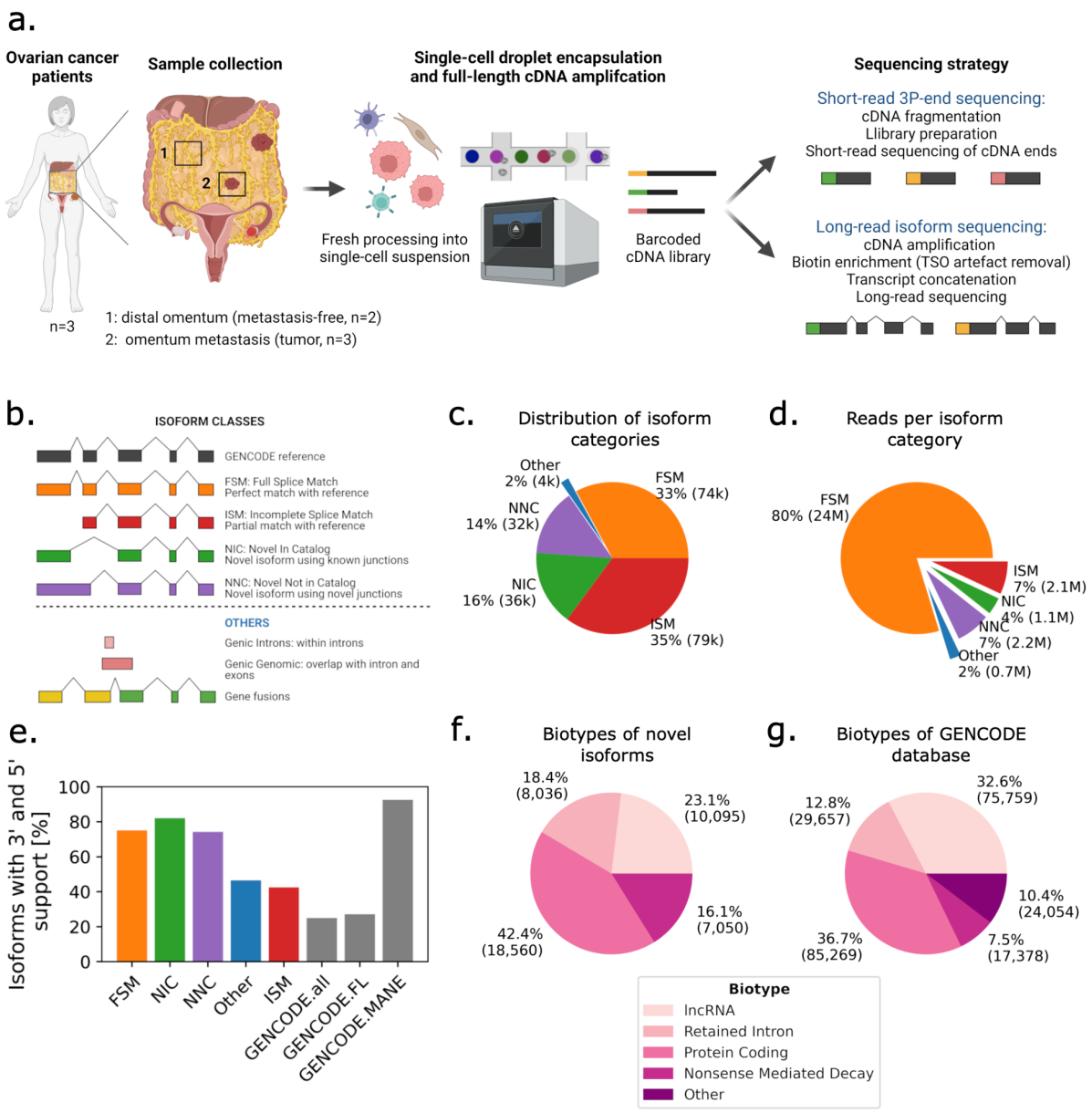


Figure 1: Study design and long read data overview. (a) Schematic of freshly processed HGSOC metastasis and patient-matched tumor-free omentum tissue biopsies, scRNA-seq. **(b)** Definition of SQANTI-defined isoform structural categories. **(c)** Proportions of isoform structural categories detected in merged metastasis and healthy omentum samples. Percentage and total number of isoforms per category are indicated. **(d)** Proportions of unique reads attributed to isoforms detected in **(c)**. Percentage and total number of UMIs per category are indicated. **(e)** Percentage of isoforms for which transcription start site is supported by CAGE (FANTOM5) data and transcription termination site is supported by polyA (PolyASite) data, per isoform structural categories. GENCODE.all indicates all protein-coding isoforms in the GENCODE database, GENCODE.FL is a subset of GENCODE.full containing only isoforms tagged as full-length, and GENCODE.MANE is a subset of canonical transcripts, one per human protein coding locus. **(f)** GENCODE defined biotypes composition of novel isoforms. **(g)** Biotypes composition of the GENCODE database.

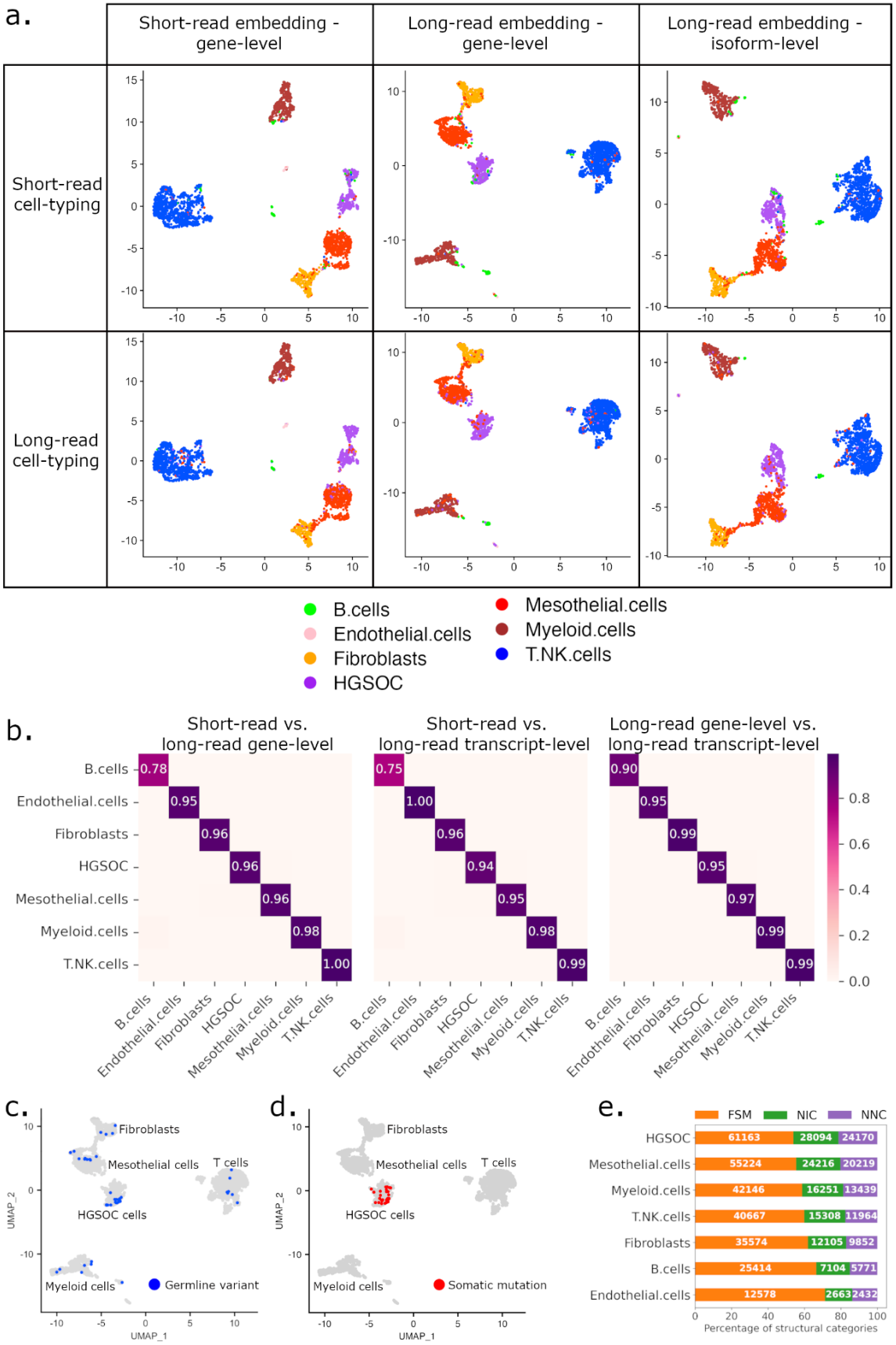


Figure 2: Clustering and cell type specific isoform distribution.

(a) Cohort UMAP embeddings by data types and automatic cell type annotation. Top and bottom rows: cell type labels based on short- and long-read data, respectively. Left column: embedding on short-read data - gene level, middle column: embedding on long-read data - gene level, right column: embedding on long-read data - isoform level. **(b)** Jaccard distance of cell populations in different UMAP embeddings: short-reads - gene level versus long-reads - gene level (left), short-reads - gene level versus long-reads - isoform level (middle), long-reads - gene level versus long-reads - isoform level (right). **(c)** Long-reads - gene level UMAP cohort visualizations of cells with at least one somatic mutation also found in bulk DNA. **(d)** Long-reads - gene level UMAP cohort visualization of cells with at least one germline variant. Germline variants are variants detected in healthy omentum distal samples. **(e)** SQANTI-defined structural category normalized distribution of isoforms detected per cell type (number of isoforms displayed in white).

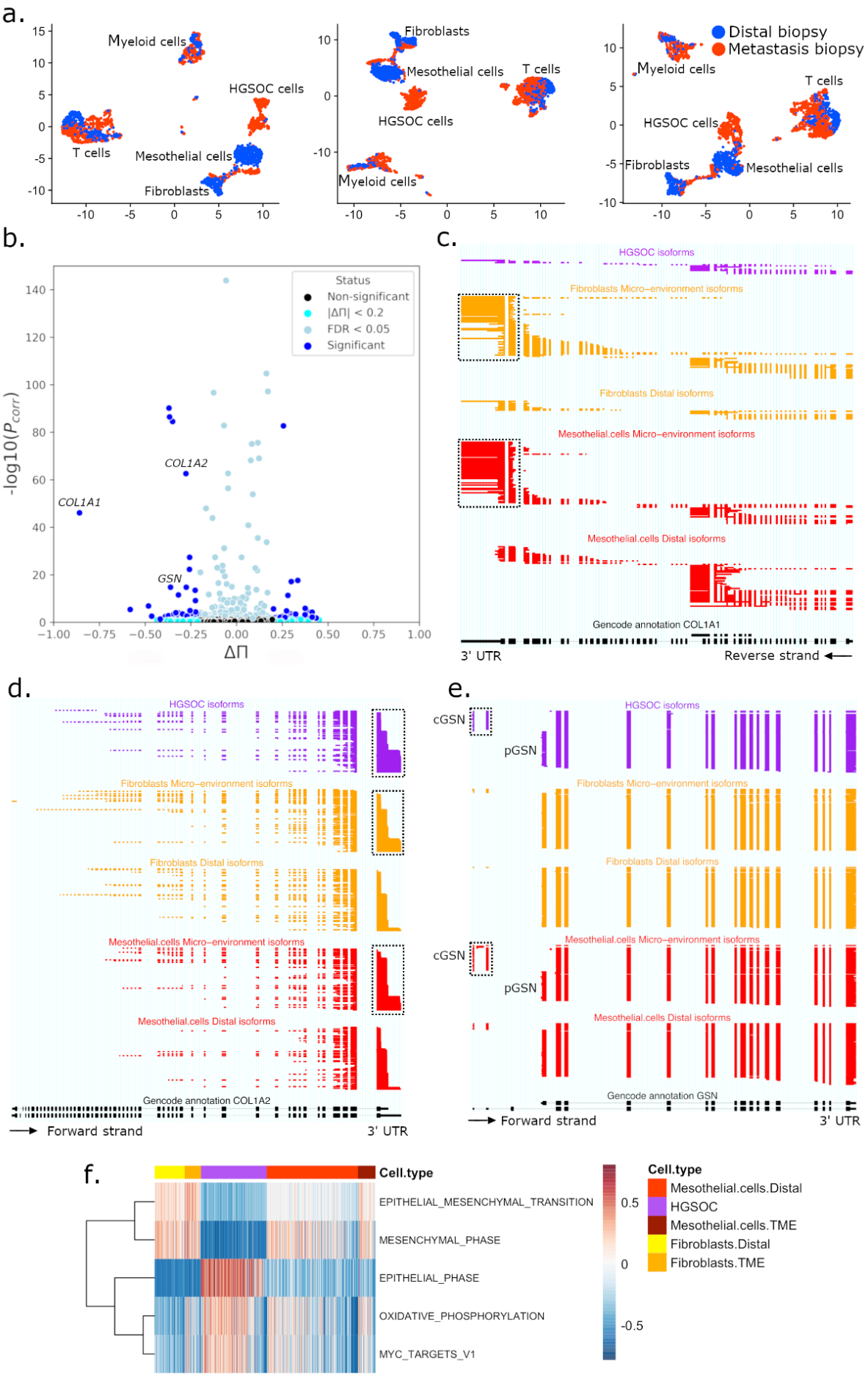

a.
Myeloid cells
HGSOC cells
T cells
Mesothelial cells
Fibroblasts
Distal biopsy
Metastasis biopsy
b.
Status
Non-significant
|ΔΠ| < 0.2
FDR < 0.05
Significant
COL1A2
COL1A1
GSN
-log10(Pcorr)
ΔΠ
c.
HGSOC isoforms
Fibroblasts Micro−environment isoforms
Fibroblasts Distal isoforms
Mesothelial.cells Micro−environment isoforms
Mesothelial.cells Distal isoforms
Gencode annotation COL1A1
3' UTR
Reverse strand
d.
HGSOC isoforms
Fibroblasts Micro−environment isoforms
Fibroblasts Distal isoforms
Mesothelial.cells Micro−environment isoforms
Mesothelial.cells Distal isoforms
Gencode annotation COL1A2
Forward strand
3' UTR
e.
cGSN
pGSN
HGSOC isoforms
Fibroblasts Micro−environment isoforms
Fibroblasts Distal isoforms
Mesothelial.cells Micro−environment isoforms
Mesothelial.cells Distal isoforms
Gencode annotation GSN
Forward strand
3' UTR
f.
Cell.type
EPITHELIAL_MESENCHYMAL_TRANSITION
MESENCHYMAL_PHASE
EPITHELIAL_PHASE
OXIDATIVE_PHOSPHORYLATION
MYC_TARGETS_V1
Mesothelial.cells.Distal
HGSOC
Mesothelial.cells.TME
Fibroblasts.Distal
Fibroblasts.TME

Figure 3: Differential isoform expression in tumor microenvironment reveals epithelial-to-mesenchymal transition.

(a) Cohort UMAPs embedding of short-read data - gene level (left), long-read data - gene level (middle), long-read data - isoform level (right), colored by tissue type. **(b)** Volcano plot of mesothelial TME vs. distal cells differential isoform usage. The X-axis represents the effect size in the gene, the Y-axis is the p-value derived from a $\chi 2$ test corrected for multiple testing using the Benjamini–Hochberg method. **(c)** ScisorWiz representation of isoforms in *COL1A1*, each horizontal line represents a single read colored according to cell types. Dashed boxes highlight the use of the canonical 3' UTR in TME fibroblasts and mesothelial cells, while distal mesothelial cells use an earlier 3' exon termination. **(d)** ScisorWiz representation of isoforms in *COL1A2*. Dashed boxes highlight the 3'UTR, where TME and HGSOC cells differentially express a longer 3'UTR than distal cells. **(e)** ScisorWiz representation of isoforms in *GSN*. Dashed boxes highlight the TSS, where mesothelial TME and HGSOC cells differentially express the *cGSN* isoform, while mesothelial distal cells and fibroblasts use *pGSN*. **(f)** Gene set variation analysis (GSVA) scores for different cell types. Heatmap colors from blue to red represent low to high enrichment.

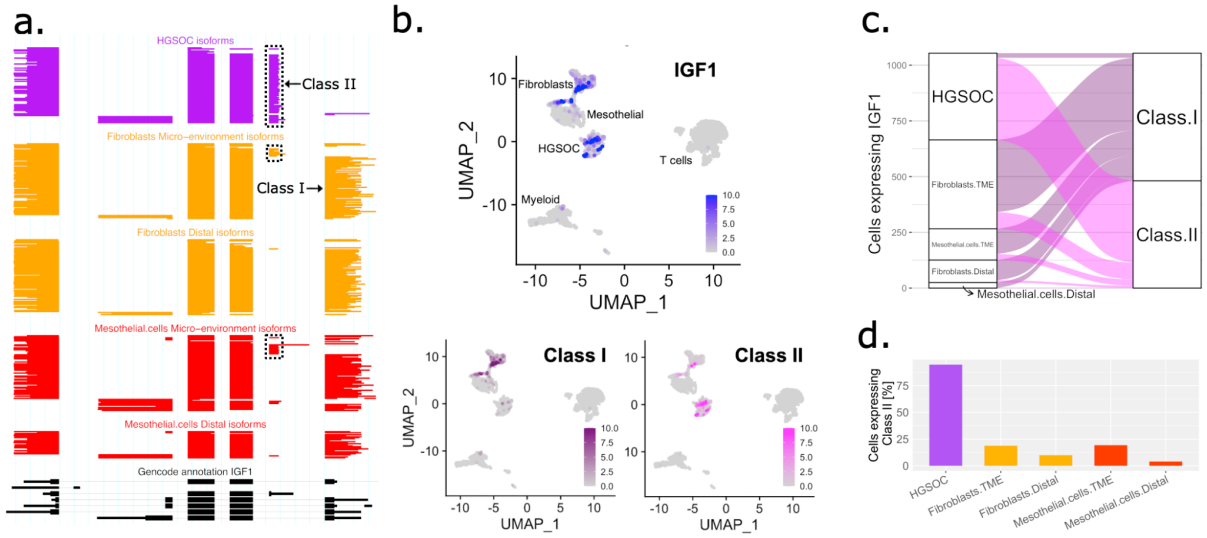


Figure 4: Differential isoform expression of *IGF1* in tumor vs non-tumor cells.
(a) ScisorWiz representation of isoforms in *IGF1*, each horizontal line represents a single isoform colored according to cell types. Colored areas are exons, and whitespace are intronic space, not drawn to scale. Exons are numbered according to the Gencode reference, Class I and II isoforms are isoforms with starting exons 1 and 2, respectively. Boxes highlight Class II expression in cancer and TME cells. **(b)** Projection of *IGF1* gene (top) and ClassI/II isoform (bottom) expression on UMAP obtained from clustering on long-reads transcripts. **(c)** Alluvial plot of cells expressing *IGF1* in different cell types (left), divided between cells expressing Class I or II (right). **(d)** Barplot of percentage of cells expressing Class II isoform in different cell types and locations colored by cell type.

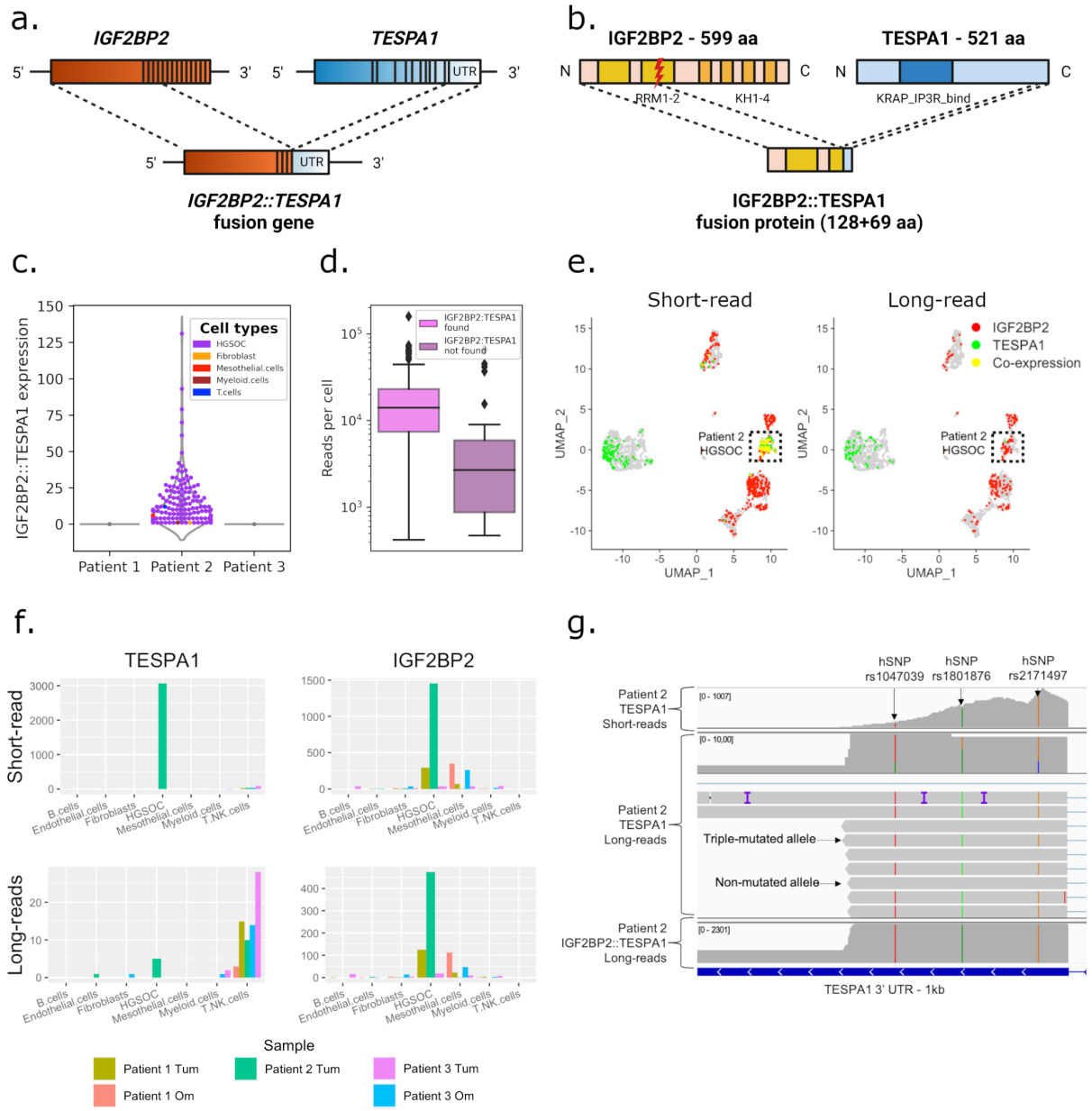


Figure 5: Tumor and patient-specific detection of novel *IGF2BP2::TESPA1* gene fusion. (a) Overview of wt *IGF2BP2*, wt *TESPA1* and gene fusions with exon structure. **(b)** Overview of wt IGF2BP2, wt TESPA1 and fusion proteins and protein domains. RRM: RNA-recognition motif, KH: hnRNP K-homology domain, KRAP_IP3R_bind: Ki-ras-induced actin-interacting protein-IP3R-interacting domain. **(c)** Violin plot showing patient and tumor specific *IGF2BP2::TESPA1* fusion transcript detection in patient 2. **(d)** UMI count in fusion-containing vs -lacking patient 2 tumor cells. **(e)** scDNA copy-number profile clustering of the matched patient 2 sample. Subclone 0 (121 cells) exhibited multiple copy number alterations along its genome representing a single tumor clone, while subclone 1 (62 cells) had a diploid genome representing non-HGSOC cells. **(f)** patient 2 scDNA reads aligning to custom *IGF2BP2::TESPA1* gene fusion breakpoint reference. Only tumor subclone reads were found to align to it.

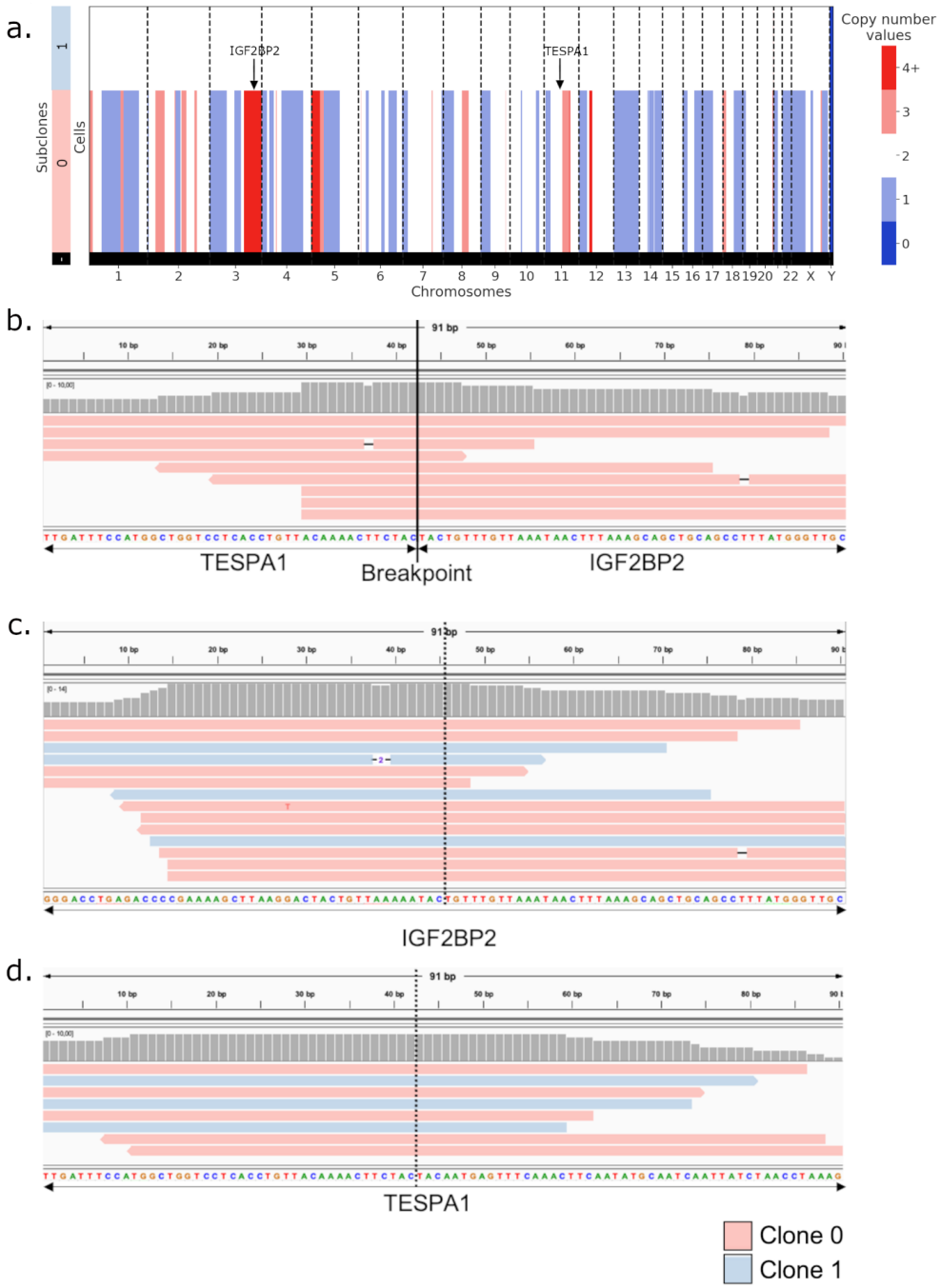


Figure 6: *IGF2BP2::TESPA1* fusion breakpoint validation in scDNA.
(a) Copy number values per subclone in Patient 2 scDNA. Sublone 0 has multiple copy number alterations, indicative of cancer, while Subclone 1 is copy-number neutral, non-cancer. **(b)** IGV view of scDNA reads aligning unambiguously to the *TESPA1::IGF2BP2* genomic breakpoint. In red, reads from Subclone 0 cells, in blue, reads from Subclone 1 cells. **(c)** IGV view of scDNA reads aligning unambiguously to wt *IGF2BP2*. The dashed line indicates the location of the (putative) breakpoint. **(d)** IGV view of scDNA reads aligning unambiguously to wt *TESPA1*. The dashed line indicates the location of the breakpoint.