Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer

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

These authors contributed equally: Arthur Dondi, Ulrike Lischetti

* 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] Current affiliation: 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. Here, we perform long-read single-cell RNA sequencing (scRNA-seq) on clinical samples from three ovarian cancer patients presenting with omental metastasis and increase the PacBio sequencing depth to 12,000 reads per cell. Our approach captures 152,000 isoforms, of which over 52,000 are novel. Isoform-level analysis accounting for non-coding isoforms reveals 20% overestimation of protein-coding gene expression on average. We also detect cell type-specific isoform and poly-adenylation site usage in tumor and mesothelial cells, and find that mesothelial cells transition into cancer-associated fibroblasts in the metastasis, partly through the TGF-β/miR-29/Collagen axis. Furthermore, we identify gene fusions, including an experimentally validated *IGF2BP2::TESPA1* fusion, which is misclassified as high *TESPA1* expression in matched short-read data, and call mutations confirmed by targeted NGS cancer gene panel results. With these findings, we envision long-read scRNA-seq to become increasingly relevant in oncology and personalized medicine.

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 isoforms[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 sequencing field is rapidly expanding[20], with novel scRNA-seq methods being constantly developed and improved on Nanopore[21,22] and PacBio[23–27] platforms. So far, long-read RNA-seq has however only been applied on the bulk level in the field of oncology[25,28,29]. 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[10,30,31].

Here, we used high-quality, high-throughput long-read scRNA-seq to capture cell type-specific genomic and transcriptomic alterations in 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 PacBio scRNA-seq dataset with the deepest coverage to date. We were

able to identify over 150,000 isoforms, of which a third were novel, as well as novel cell type-specific isoforms. Isoform-level analysis revealed that, on average, 20% of the protein-coding gene expression was non-coding, leading to an overestimation of the protein expression. By combining differential isoform and poly-adenylation site usage analysis between cells from the metastatic TME and distal omental biopsies, we found evidence that in omental metastases, mesothelial cells transition into CAFs, partly through the TGF-β/miR-29/Collagen axis. Additionally, we discovered dysregulations in the insulin-like growth factor (IGF) network in tumor cells on the genomic and transcriptomic levels. 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 Tables 1 and 2**) 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 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 artifact removal, resulted in 30.7 Mio unique molecular identifiers (UMIs), for an average of 12k UMIs per cell. (**Extended Data Table 1**). There was a mean of four cDNA molecules concatenated per sequencing read overall, and cDNA length was similar across samples (**Extended Data Fig. 1a,b**). Artifact removal filtered 51% of the reads, and it included the removal of intra-priming (63%), non-canonical isoforms (36%), and reverse transcriptase switching (1%)[32] (**Extended Data Fig. 1c**). It must be emphasized that those artifact reads emerge from the single-cell library preparation and are also present in short-read data, where they cannot be filtered and are hence accounted for as valid reads.

The long-read dataset revealed 152,546 isoforms, each associated with at least three UMIs. We classified the isoforms according to the SQANTI classification[32] 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[33] and their 3' end to the PolyASite database[34] (**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%).

144 FSM, NIC, and NNC had overall better 3' and 5' validation than the full-length tagged isoforms
145 in the GENCODE database (**Fig. 1e**). Only the 'Matched Annotation from NCBI and EMBL-
146 EBI' (MANE[35]) containing curated representative transcripts cross-validated between the
147 GENCODE and RefSeq database had a better 3' and 5' validation of 95%. A total of 52,884
148 novel isoforms were complete (NIC+NNC), of which 40,046 were confirmed as valid novel
149 isoforms by GENCODE (corresponding to 17% of the current GENCODE v36 database), and
150 3,695 were extended versions of existing isoforms. Isoforms that were not confirmed were
151 mainly either "partially redundant with existing transcripts", or "overlapping with multiple loci".
152 Finally, we assessed the biotypes of our newly discovered isoforms, indicative of their
153 presumed functional categorization. We found that 42% are protein-coding, more than the
154 36% of protein-coding isoforms found in the GENCODE database (230k entries) (**Fig. 1f,g**).
155 This demonstrates the ability of concatenated long-read sequencing to generate high-yield,
156 high-quality data and discover novel isoforms with enhanced annotation.

157 Long-read sequencing allows for short-read-independent cell type 158 identification

159 Next, through comparison to short-read data, we assessed the ability of long-read sequencing
160 to cluster cells and to identify cell types. We generated short- and long-read gene count
161 matrices and removed non-protein-coding, ribosomal, and mitochondrial genes. After filtering,
162 we obtained 16.5 Mio unique long-reads associated with 12,757 genes, and 26.3 Mio unique
163 short-reads associated with 13,122 genes (**Extended Data Table 2**). The short- and long-
164 read datasets were of similar sequencing depth with a median of 4,930 and 2,750 UMIs per
165 cell, respectively (average 10,235 and 6,413 UMIs, **Extended Data Fig. 2a**). Long-read data
166 contained slightly fewer detected genes, and genes detected in both datasets overlapped by
167 86.4% (**Extended Data Fig. 2b,c**). Paradoxically, the genes detected were overall shorter in
168 long-reads than short-reads, likely due to the concatenation step (**Extended Data Fig. 2d**).
169
170 We first identified cell types independently per cell, using cell type marker gene lists (**Methods,**
171 **Extended Data Fig. 3a**). We compared short- and long-read data and found that both data
172 types identified cell types with very similar percentages, namely HGSOC (13,4% in short-read
173 vs 13.6% in long-read data), mesothelial cells (20.2 vs 20.5%), fibroblasts (10.7 vs 10.7%), T
174 cells (38.7 vs 38.5%), myeloid cells (14.6 vs 14.9%), B cells (1.4 vs 1.1%), and endothelial
175 cells (1.1 vs 1.4%) (**Extended Data Fig. 3b**). We then projected short-read gene, long-read
176 gene, and long-read isoform expression onto 2-dimensional embeddings using UMAP[36] (**Fig.**
177 **2a**). We manually clustered cell types based on the embeddings and calculated the Jaccard
178 distance between clusters. Cell clusters based on short- and long-reads were very similar,
179 with a Jaccard distance >94% for all cell types except B-cells, where the Jaccard distance was
180 >75% (**Fig. 2b**). Furthermore, Jaccard similarity analysis between cell type clusters and
181 attributed cell type labels were analogous between short- and long-read data, with a better

182 prediction of B cells and endothelial cells for long-reads **(Extended Data Fig. 3c**). Except for
183 tumor cells, cell types did not show patient-specific clustering (**Extended Data Fig. 3d**).

184 Long-read sequencing captures germline and somatic mutations and 185 identifies increased neojunctions in tumor cells

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

195 Cell type-specific isoform expression analysis (overall and per gene) revealed increased
196 isoform and transcript expression in HGSOC cells compared to other cell types **(Extended**
197 **Data Fig. 4a-c)**. This difference does however not translate into mean UMIs per isoform, as
198 isoforms expressed in cancer cells harbor fewer UMIs than in mesothelial cells **(Extended**
199 **Data Fig. 4d)**. This means that cancer cells express more low-abundant isoforms suggesting
200 wider isoform diversity and broader cellular functions, and controls. Isoform class distribution
201 between cell types revealed a higher fraction of novel isoforms and neojunctions (NNC) in
202 tumor cells (**Fig. 2e**). Looking into unique isoform expression in the different cell types, we
203 found that cancer cells contained more than 8% (9,476) of cell type-specific isoforms, and
204 35% of those contained neojunctions (NNC) (**Methods, Extended Data Fig. 4e**). This was an
205 increase of 2.3-10.6 times compared to the other cell types. At the cellular level, 0.5% of the
206 cancer-specific isoforms were also unique to a single cell, which is between 3-6 times the
207 percentage of unique isoforms in other cell types, and 50% contained neojunctions (**Extended**
208 **Data Fig. 4f**). Those rare isoforms were difficult to detect with previous methods, hence their
209 novelty. Taken together, cancer cells expressed at least twice as many unique isoforms than
210 other cell types, indicating an increased transcriptomic diversification and supporting previous
211 findings of cancer-specific neojunction expression in bulk data[16].

212 Tumor microenvironment shows epithelial-to-mesenchymal transition 213 through TGFβ-driven miR-29 downregulation

214 In the subsequent analysis, we compared stromal fibroblasts and mesothelial cells derived
215 from metastasis (TME) and matched tumor-free (distal) omentum. In distal samples,
216 fibroblasts and mesothelial cells formed distinct clusters in both short- and long-read data (**Fig.**
217 **3a left, Extended Data Fig. 5a**). In metastatic samples, however, TME fibroblasts and

218 mesothelial cells formed a bridge in the UMAP embedding, suggesting that these cells might
219 undergo a cell state transition. To test this hypothesis, we conducted gene set enrichment
220 analysis and found that the epithelial-to-mesenchymal transition (EMT) pathway was enriched
221 in TME compared to distal mesothelial and fibroblast cells (**Fig. 3a, right**). Similarly, collagen
222 fibril organization and extracellular matrix (ECM) pathways were enriched in TME cells,
223 indicating a reprogramming of the TME cells during metastasis formation (**Extended Data Fig.
224 5b**). Additionally, we compared the alternative poly-adenylation (APA) (**Methods**). Among the
225 2,876 genes tested for APA, the isoforms of 26 genes in TME mesothelial cells exhibited
226 significant 3'UTR lengthening compared to distal mesothelial cells, while 13 genes showed
227 shortened 3'UTRs (**Fig. 3b**). Collagen-encoding genes *COL1A2*, *COL3A1*, *COL5A2*, and
228 *COL6A1* were similarly lengthened and up-regulated in TME cells (**Fig. 3c**), with *COL1A2*
229 having the highest effect size (P_{corr}=3.42x10^{-67}, 53% change).
230
231 As 3'UTR lengthening is usually associated with a decrease in expression due to microRNA
232 (miRNA) silencing[37], the increased usage of lengthened 3'UTR in TME mesothelial cells
233 suggests that distal cells may express a distinct set of miRNAs not present in TME cells.
234 Collagen-encoding genes are known to be regulated by the miR-29 family in fibroblasts[38].
235 Thus, we used miRDB[39] to predict gene targets for silencing by miR-29a/b/c (**Supplementary
236 Table 2**). Among the 26 isoforms with lengthened 3'UTR, 9 were predicted targets of miR-
237 29a, almost all described as EMT actors (collagens[40–43], *KDM6B*[44], *TNFAIP3*[45], *FKBP5*[46] and
238 *RND3*[47]). In contrast, none of the 13 shortened 3'UTR isoforms were predicted to be miR-29
239 targets (**Fig. 3b**). Furthermore, we compared gene expression of TME and distal mesothelial
240 cells (**Supplementary Table 3)**, and genes with lengthened 3'UTR isoforms predicted to be
241 silenced by miR-29 were significantly overexpressed (P = 1.35x10^{-3}) in TME mesothelial cells
242 compared to the ones not predicted to be silenced by miR-29a (**Fig. 3d**). Mesothelial TME
243 cells also differentially expressed miR-29 targets which are major ECM genes, such as
244 collagen gene *COL1A1* (fold-change=9.0, P_{corr}=1.72x10^{-124}), *MMP2*[48] (fold-change=4.1,
245 P_{corr}=1.51x10^{-30}), and *LOX*[49] (fold-change=10.6, P_{corr}=6.33x10^{-51}), which is also lengthened
246 *(*P_{corr}=7.07x10^{-2}, 77% change). Overall, ECM-related genes known to be targeted by miR-29
247 were upregulated in TME cells compared to distal cells (**Fig. 3e**), supporting the hypothesis
248 that the observed EMT was potentially linked with the miR-29 downregulation. When
249 comparing differentially expressed isoforms in TME mesothelial cells to distal cells, *COL1A1*
250 was also the gene with the highest change in relative isoform abundance amongst all its
251 isoforms (P_{corr}=6.34x10^{-49}, 86% usage change, **Methods**) **(Fig. 3f)**. In TME mesothelial cells,
252 the *COL1A1* canonical 3' poly-adenylation site was used, whereas distal cells used a
253 premature poly-adenylation site, leading to the formation of truncated isoforms (**Fig. 3f**). When
254 incorporating only protein-coding isoforms and removing the truncated isoforms from the
255 analysis, the gene expression fold-change increased from 9 to 62-folds (P_{corr}=3.15x10^{-183}).
256 This overexpression of canonical *COL1A1* in the TME can be explained by the absence of
257 miR-29 silencing, as previously described[50].

258
259 The miR-29 family is known to be an EMT inhibitor[50]. Its silencing through the TGFβ pathway
260 correlates with the upregulation of ECM-encoding genes, including multiple collagens, as
261 reported in the present study. The main TGFβ gene *TGFB1* was found to be enriched in TME
262 mesothelial cells (fold-change=1.4, P_{corr}=2.32x10^{-2}). Furthermore, in distal mesothelial cells,
263 38% of *TGFB1* isoforms comprise an alternative 3' exon, leading to aberrant protein
264 expression (**Extended Data Fig. 5c**), while the canonical protein-coding *TGFB1* isoform
265 *ENST00000221930.6* is overexpressed in TME cells (fold-change=2.3, P_{corr}=6.59x10^{-9}). miR-
266 29 is also regulated through the expression of non-coding RNAs that act as molecular
267 sponges, directly binding to miR-29 and, therefore, leading to the overexpression of their
268 targets. The TGFβ-regulated long non-coding RNA H19, which enhances carboplatin
269 resistance in HGSOC, has been reported to promote EMT through the *H19*/miR-29b/*COL1A1*
270 axis[51–53] and was found to be overexpressed in the TME mesothelial cells (fold-change=4.6,
271 P_{corr}=3.46x10^{-34}). Circular RNAs have also been described as miR-29 sponges, notably
272 circMYLK, and circKRT7 in HGSOC. circMYLK and circKRT7 originate from *MYLK* and *KRT7*,
273 respectively, which are both significantly overexpressed in TME mesothelial cells (fold-
274 change=4.1, P_{corr}=6.97x10^{-96} and fold-change=4.6, P_{corr}=2.18x10^{-13})[54,55]. Similarly, TME
275 mesothelial cells expressed the endogenous isoform of *GSN* (*cGSN*), while distal cells only
276 expressed the secreted isoform (pGSN)[56] (**Extended Data Fig. 5d**). *cGSN* has been shown
277 to be under TGF-β control in breast cancer and to increase EMT marker expression[57]. In
278 conclusion, our findings strongly support that, in omental metastases, the mesothelial cells
279 transition into cancer-associated fibroblasts (CAFs), partly through the TGF-β/miR-
280 29/Collagen axis.

281 Differential isoform usage in cancer reveals changes in biotypes

282 After comparing cells from the TME with distal cells, we investigated which isoforms, biotypes,
283 and poly-adenylation sites were differentially used between cancer and all distal cells. HGSOC
284 cells expressed isoforms differentially with a change in relative isoform abundance of more
285 than 20% in 960 genes (15.1%), compared to all distal cells (6,353 genes tested in total, **Fig.
286 4a**, **Supplementary Table 4, Methods)**. In 36% of those 960 genes, the highest expressed
287 isoform biotype changed between conditions (**Fig. 4b)**. In 32% of instances, there was a
288 transition from a protein-coding to a non-protein-coding isoform, and in 17% of cases, cancer
289 cells expressed a protein-coding isoform while distal cells expressed a non-protein-coding
290 isoform. Only 39 genes (0.6%) had an isoform switch with a change in relative isoform
291 abundance of more than 50%. 59% of these switched isoforms demonstrated a biotype
292 transition (49% of protein-coding to non-protein-coding transition, **Fig. 4c)**, and in 33% of
293 cases, cancer cells expressed a protein-coding isoform while distal cells expressed a non-
294 protein-coding isoform. Additionally, in cancer, distal, and TME cells, on average 20-21% of
295 the expression in protein-coding genes was non-coding (**Extended Data Fig. 6a)**, and 13-
296 14% of protein-coding genes had more than 50% of non-coding expression. This means that,

297 on average, using only gene-level information to estimate protein expression (as done in short-
298 read data) will lead to an overestimation of 20%.
299
300 The ten genes with the statistically most significant switches were *IGF1*, *TPM2*, *NCALD*,
301 *VAMP8-VAMP5*, *EXOSC7*, *ICAM3*, *CERK*, *OBSL1*, *GSN,* and *RTN1* (**Supplementary**
302 **Table 4, Methods**). In *IGF1*, cancer cells across all patients predominantly used the second
303 exon of the gene as their transcription start site (secreted isoform, *Class II*), whereas non-
304 cancerous cells primarily used the first exon (endogenous isoform, *Class I)*[58] (**Fig. 4d**). On the
305 contrary, similarly to TME mesothelial cells, cancer cells expressed the endogenous isoform
306 of *GSN* (*cGSN*), while distal cells only expressed the secreted isoform (*pGSN*)[56] (**Extended**
307 **Data Fig. 5d**). In *RTN1*, distal cells expressed the isoform *RTN1-A* and *RTN1-B,* while Patient
308 3's cancer cells expressed *RTN1-C*, an isoform known to bind to the anti-apoptotic protein
309 Bcl-xL and reduce its activity[59,60] (**Extended Data Fig. 6b).** In the tropomyosin gene *TPM2*,
310 which is involved in TGF-β-induced EMT, cancer cells differentially expressed exon 6b
311 (isoform *TPM2.3*, expressed in epithelial cells[61]) and the alternative 3'UTR exon 9a (**Extended**
312 **Data Fig. 6c**). In *VAMP5*, the overexpressed isoform in HGSOC cells was a *VAMP8-VAMP5*
313 read-through gene, i.e., a novel gene formed of two adjacent genes, previously described in
314 human prostate adenocarcinoma[62] (**Extended Data Fig. 6d**). HGSOC cells expressed almost
315 no wild-type (wt) *VAMP5* but had a significantly higher *VAMP8* expression than distal cells
316 (P_{corr}=1.0x10^{-15}), indicating that this read-through gene was under transcriptional control of
317 *VAMP8*. With a short-read 3' capture method, this *VAMP8-VAMP5* expression cannot be
318 distinguished from the wt *VAMP5* expression. For *NCALD* and *OBSL1*, only cancer cells
319 expressed canonical protein-coding isoforms, while other cells expressed short non-coding
320 isoforms (**Extended Data Fig. 6e,f)**. For *CERK*, by contrast, Patient 2's HGSOC cells strongly
321 expressed a novel isoform leading to a shortened protein (**Extended Data Fig. 6g)**. Finally, in
322 *ICAM3,* cancer cells mainly expressed a short protein-coding isoform, while distal cells (mainly
323 T cells) expressed the canonical isoform. More characterization of those isoforms will be
324 necessary in the future to explore the biological implications linked to their expression
325 (**Extended Data Fig. 6h**).
326
327 Although isoforms differentially expressed in cancer cells were similar among patients, there
328 was one significant case of patient-specific expression. For *OAS1*, Patient 3 predominantly
329 expressed isoform *p42*, while Patients 1 and 2 exhibited a balanced distribution of isoforms
330 *p42* and *p46*. The *p42* and *p46* expressions are known to be allele-specific, caused by the
331 rs10774671 SNP, a splice acceptor A/G variation. However, this cannot explain the different
332 expression levels as all patients have both the A and G alleles (**Fig. 4e**). Given that isoform
333 *p42* is more susceptible to Nonsense-Mediated mRNA Decay (NMD) and possesses
334 diminished enzymatic activity[63], the observed differences could potentially indicate a
335 diminished OAS1 activity in Patient 3. Whether *OAS1* has an impact on ovarian cancer is still
336 to be elucidated.

When testing for differential APA between HGSOC and distal cells, we found shortened 3'UTR in 85 and lengthened 3'UTR in 203 genes (n=4758) (**Fig. 4f**). There was a notable trend toward lengthening of the 3'UTR in cancer cells ($P=5.59x10^{-20}$), with *COL1A2* emerging once more as the most prominent finding ($P_{corr}=7.48x10^{-47}$, 61% change) (**Fig. 3b-c**, **Fig. 4f**). Expression levels remained consistent between genes featuring either shortened or lengthened 3'UTRs. Furthermore, neither miRNA profiles nor canonical pathways exhibited an overlap exceeding 20% with either the lengthened or shortened gene sets (**Methods**).

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 5**). 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[64]. *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 in T cells, as well as in HGSOC cells, where its expression values were elevated. High expression was exclusive to Patient 2 HGSOC cells and colocalized with *IGF2BP2* expression (**Fig. 5e,f**). *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 re-aligned Patient 2's short-reads to a custom reference including the *IGF2BP2::TESPA1* transcriptomic breakpoint as well as wt *TESPA1* and wt *IGF2BP2* junctions (**Extended Data Fig. 7a,**

375 **Methods**). Out of the 994 reads mapping to the custom reference, 93% preferentially aligned
376 to *IGF2BP2::TESPA1* (99.8% of those were from HGSOC cells). This means that, when given
377 the option, reads previously aligning to *IGF2BP2* or *TESPA1* are preferentially mapping to the
378 fusion reference, and the reported overexpression of *TESPA1* in short reads is likely an
379 *IGF2BP2::TESPA1* expression. Furthermore, reads covering the *TESPA1* 3' UTR region
380 harbored three heterozygous single nucleotide polymorphisms (hSNPs): chr12:54.950.144
381 A>T (rs1047039), chr12:54.950.240 G>A (rs1801876), and chr12:54.950.349 C>G (
382 rs2171497). In long-reads, wt *TESPA1* was either triple-mutated or not mutated at all,
383 indicating two different alleles. All fusion long-reads, however, were triple-mutated, indicating
384 a genomic origin and monoallelic expression of the fusion (**Fig. 5g**). In short reads, the three
385 loci were mutated in nearly all reads, supporting the hypothesis that the observed *TESPA1*
386 expression represents almost completely *IGF2BP2::TESPA1* expression and that it has a
387 genomic origin.

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

389 To validate that the *IGF2BP2::TESPA1* gene fusion is the result of genomic rearrangements,
390 both bulk and single-cell DNA sequencing (scDNA-seq) data from matched omental
391 metastasis was used to query the genomic breakpoint. A putative genomic breakpoint was
392 first found in the RNA data. Two long-read fusions were mapped to intronic regions of
393 *IGF2BP2* and *TESPA1* (**Extended Data Fig. 7b**), pinpointing the location of the breakpoint at
394 position chr3:185,694,020-chr12:54,960,603. Subsequent genotyping PCR on genomic DNA
395 extracted from patient-matched tissue samples using *IGF2BP2::TESPA1*, wt *IGF2BP2*, and
396 wt *TESPA1* primer pairs flanking the genomic breakpoint confirmed the presence of
397 *IGF2BP2::TESPA1* in Patient 2 in 3 out of 4 tested samples (**Fig. 6a, Methods**). In contrast
398 and as expected, the fusion was not found in Patient 1.
399
400 To assess whether the fusion was exclusive to cancer cells, we further investigated scDNA-
401 seq data from Patient 2. For the identification of cancer cells, we inferred the scDNA-seq copy
402 number profiles of all cells. We successfully identified two distinct clones within the pool of 162
403 cells. These clones encompassed a cancer clone designated as "Subclone 0" and a
404 presumably non-cancer clone without copy number alterations labeled as "Subclone 1" (**Fig.
405 6b**). We next aligned the scDNA-seq data to a custom reference covering the breakpoint
406 (**Methods)** and only found cancer reads mapping to the fusion breakpoint (P=0.032), while a
407 mixture of reads from cancer and non-cancer cells mapped to wt *IGF2BP2* and wt *TESPA1*
408 (P=0.78 and P=1.00, respectively) (**Fig. 6c**). Thus, genotyping PCR of bulk extracted DNA
409 and scDNA-seq data confirmed the genomic fusion breakpoint in the intronic region detected
410 in long-read scRNA-seq data. scDNA-seq also confirmed that the *IGF2BP2::TESPA1* fusion
411 was cancer-cell specific, as suggested by long-read scRNA-seq data.
412

413 *IGF2BP2* was also overexpressed in cancer cells from Patient 2 compared to other patients
414 on both RNA and protein levels (**Extended Data Fig. 8a-c**). In Patient 2, there was an elevated
415 copy number observed within the genomic region encompassing *IGF2BP2* (**Fig. 6b**).
416 Therefore, the presence of a fusion allele on one allele does not seem to hinder the
417 transcription of the wt *IGF2BP2* allele. Coherent with the high IGF2BP2 protein levels, *IGF2*
418 RNA, which is bound by the wt IGF2BP2 protein, is also largely overexpressed in Patient 2
419 cancer cells compared to other patients (**Extended Data Fig. 8d**). This could indicate that the
420 fusion happened partly due to accessible chromatin. In the ovarian cancer TCGA RNA
421 dataset, the expression levels of exons surrounding breakpoints of *IGF2BP2* and *TESPA1* did
422 not change (**Extended Data Fig. 9a,b**), and the overall expression of the genes did not
423 correlate in any patient, suggesting that we detected an uncommon, patient-specific fusion
424 (**Extended Data Fig. 9c**).

425

Discussion

The detection of genomic alterations like mutations[65,66] and gene fusions[67,68], alongside isoform expression and usage[15] at the single-cell scale offers valuable insights into cancer development, metastasis formation, the role of the tumor microenvironment (TME), potential drug targets, and therapy responses[69]. Here, we applied PacBio HiFi high-throughput long-read RNA-seq on five omental metastases and matching tumor-free samples from chemo-naive HGSOC patients to detect and quantify all of these alterations.

Until now, a combination of single-cell short- and long-read sequencing was necessary to identify cell-specific isoforms: the increased depth of short-read sequencing allowed for cell typing based on gene expression, while long-read sequencing was used for isoform identification[23]. Leveraging multiple strategies to maximize the long-read sequencing output, we achieved a 50-fold increased sequencing depth compared to the first long-read PacBio scRNA-seq study[23], allowing for short read-comparable cell type identification. As a result, future studies employing comparable or higher long-read capacities will no longer depend on parallel short-read sequencing, leading to cost and labor savings.

Taking advantage of the long-read technology, our study provides evidence that cancer cells may induce EMT in TME mesothelial cells. In omental metastases, it has been shown that HGSOC-secreted TGF-β triggers EMT and converts TME cells, including mesothelial cells, into CAFs, which in turn may favor tumor cell adhesion, invasion, and proliferation[70–72]. Our findings provide further evidence of this phenomenon and reveal that the process might partly be controlled through the TGF-β/miR-29/Collagen axis: secreted and endogenous TGF-β downregulates miR-29 expression, thus increasing the expression of its targets, including collagens[73]. Coincidentally, Han and colleagues[48] recently showed that omental CAF-derived exosomes from HGSOC patients contained significantly lower miR-29c than normal fibroblasts. Furthermore, the reduced levels of miR-29 have been demonstrated to play a role in the development of cisplatin resistance by upregulating collagen expression[74].

In addition to TGF-β, HGSOC cells also promote EMT and CAF activation through the secretion of other growth factors such as insulin-like growth factors (*IGF*s)[75]. In our data, HGSOC cells revealed a profoundly modified IGF system among all patients with a drastic switch from endogenous *Class I* to secreted *Class II IGF1* isoform, *IGF2* overexpression, and a highly expressed *IGF2BP2::TESPA1* gene fusion in one patient. Secreted IGF1 (Class II) and IGF2 activate EMT through IGF1R, by triggering an uncontrolled wound healing response[76]. The *IGF* gene family also promotes cancer growth, survival, proliferation, and drug resistance through signaling via *PI3K-AKT* or *MAPK*, and is a known clinical target in ovarian cancer[77].

461 Additionally, we demonstrated the potential of the technology in terms of coverage and
462 sequencing accuracy to detect mutations and gene fusions. In particular, in one patient, the
463 novel fusion *IGF2BP2::TESPA1* was highly overexpressed compared to wt *IGF2BP2* (~10x
464 more) and *TESPA1* (~150x more). *IGF2BP2* is known to be regulated via 3'UTR miRNA
465 silencing[78], however, the *IGF2BP2::TESPA1* fusion has the unregulated 3'UTR of *TESPA1*,
466 which could explain its overexpression. *TESPA1* is normally expressed in T cells[79] and long-
467 read data confirmed T cell-specific wt *TESPA1* expression. Short read data however
468 erroneously reported *TESPA1* as the most differentially expressed gene in cancer cells,
469 resulting from 3' end capture of the fusion transcripts. This highlights that short-read scRNA-
470 seq data fails to distinguish between gene and fusion expression, potentially leading to wrong
471 biological conclusions.

472 In addition to their canonical mRNA isoforms, numerous protein-coding genes express non-
473 coding RNA isoforms, which are wrongly accounted for as protein-coding on the gene
474 expression level and can serve as miRNA sponges and competing endogenous RNAs,
475 especially in EMT and metastasis[80,81]. In our data, we found that 20% of the protein-coding
476 gene expression was non-coding and, in the genes exhibiting a significant isoform switch
477 between cancer and healthy cells, nearly half (49%) transitioned from a protein-coding to a
478 non-coding isoform. Furthermore, 51% of the UMIs were composed of intra-priming and non-
479 canonical artifacts before filtering, and their detection was only possible through isoform
480 classification. Overall, our findings highlight the need for an isoform-specific quantification to
481 accurately assess protein-coding gene expression[82] and narrow the RNA-protein gap[83–86].
482 Additionally, a better isoform characterization is needed to understand their biological
483 implications, as we partly demonstrated with *IGF1 Class I/II*, *cGSN, TPM2.3, RTN1-C,* and
484 *OAS1 p42*.

485 Although the achieved sequencing depth allowed for short-read-independent cell typing and
486 clustering, a further increased depth is needed to capture low-abundance transcripts. For
487 example, we did not obtain sufficient reads to retrieve and characterize the T cell receptor
488 repertoire. This is consistent with a long-read scRNA-seq study in blood lymphocytes that
489 reported a 3.6-fold lower pairing rate for T cell receptors than the higher abundant B cell
490 receptors from plasmablasts[87]. Using a single-cell long-read targeted sequencing approach,
491 Byrne and colleagues were able to achieve a T cell receptor alpha and beta pairing rate of
492 50% in an ovarian cancer patient sample, showing the potential to extract immune repertoire
493 information from long-read data with increased sequencing depth[88]. Enrichment for low
494 abundant transcripts for long-read sequencing or depletion of mitochondrial and ribosomal
495 RNA[89] represent interesting avenues forward.

496 Taken together, we demonstrate that long-read sequencing provides a more complete picture
497 of cancer-specific changes. These findings highlight the manifold advantages and new

498 opportunities that this technology provides to the field of precision oncology, opening the
499 premise of personalized drug prediction and neoantigen detection for cancer vaccines[90,91].

500 # Materials and Methods

501 ## Omentum patient cohort

502 The use of material for research purposes was approved by the corresponding cantonal ethic
503 commissions (EKNZ: 2017–01900, to V.H.S.), and informed consent was obtained for all
504 human primary material. Tissue samples were immediately collected from the theater and
505 transferred on ice to the Department of Biomedicine of the University Hospital Basel for tissue
506 dissociation.

507 ## Sample processing

508 Fresh omentum and omental HGSOC tumor metastasis biopsy samples were cut into small
509 pieces and dissociated in digestion solution (1 mg/mL collagenase/Dispase [Sigma cat. no.
510 10269638001], 1 unit/mL DNase I [NEB, cat. no. M0303] and 10% fetal bovine serum (FBS)
511 in DMEM [Sigma, cat. no. D8437-500mL]) for 30 min at 37°C. To focus on the non-adipose
512 cell fraction, adipocytes were separated by centrifugation and the cell pellet was collected.
513 Red blood cell lysis (RBC) was performed using MACS red blood lysis solution (cat. no. 130-
514 094-183). Then, the cell pellet was resuspended into MACS dead cell removal microbeads
515 (cat. no. 130-090-101) and loaded into the AutoMACS separator to remove dead cells. After
516 counting cell numbers, cells were resuspended in PBS with 1% BSA and transferred to the
517 Genomics Facility Basel. The cell suspension was again filtered and cell number and viability
518 were assessed on a Cellometer K2 Image Cytometer (Nexcelom Bioscience, cat. no.
519 Cellometer K2) using ViaStain AOPI Staining Solution (Nexcelom Bioscience, cat. no. CS2-
520 0106-5mL) and PD100 cell counting slides (Nexcelom Bioscience, cat. no. CHT4-PD100-003).
521 For samples with viability below 70% and when cell numbers allowed (>10^5 cells total),
522 apoptotic and dead cells were removed by immunomagnetic cell separation using the Annexin
523 Dead Cell Removal Kit (StemCell Technologies, cat. no. 17899) and EasySep Magnet
524 (StemCell Technologies, cat. no. 18000). If the cell pellet appeared still red, additional RBC
525 lysis was performed. Cells were washed with a resuspension buffer (PBS with 0.04% BSA),
526 spun down, and resuspended in a resuspension buffer. Finally, cells were again counted and
527 their viability was determined. The cell concentration was set according to 10x Genomics
528 protocols (700-1,200 cells/µL).

529 ## 10x Genomics single-cell capture and short-read sequencing

530 Cell suspensions were loaded and processed using the 10x Genomics Chromium platform
531 with the 3P v3.1 kit on the 10x Genomics Chromium Single Cell Controller (10x Genomics,
532 cat. no. PN-120263) according to the manufacturer's instructions. 500 or 1,000 cells were
533 targeted per lane. The quality of cDNA traces and GEX libraries were profiled on a 5200
534 Fragment Analyzer (Agilent Technologies).

535 Paired-end sequencing was performed on the Illumina NovaSeq platform (100 cycles, 380pm
536 loading concentration with 1% addition of PhiX) at recommended sequencing depth (20,000-
537 50,000 reads/cell).

538 Long-read library preparation and PacBio sequencing

539 To increase long-read PacBio sequencing throughput, we followed the strategy of cDNA
540 concatenation of the HIT-scISOseq protocol[24] with the modification of two rounds of biotin-
541 PCR in order to further reduce template-switch oligo (TSO) artifacts from the data.
542 Full protocol details:

543 cDNA amplification and biotin-enrichment

544 15 ng of each patient's cDNA library were amplified using the KAPA HiFi HotStart Uracil+
545 ReadyMix 2x (Kapa Biosystems, cat. no. KK2801) with 0.5 µM final concentration of custom-
546 primers (Integrated DNA Technologies, HPLC purified). Primers contained overhang
547 sequences adapted from Hebelstrup *et al.*[92] with a single deocxyuredine (dU) residue at a 10
548 nt distance from the 5' terminus enabling USER enzyme digestion and creating single-
549 stranded overhangs. Generated PCR fragments thus contain a single dU residue per DNA
550 strand. The forward primer was specific to the 10x Genomics partial Read 1 sequence and
551 contained a biotin modification allowing for biotin enrichment of amplified full-length cDNA
552 molecules. The reverse primer was specific to the 10x Genomics partial TSO sequence.
553 Forward Primer: /5Biosg/AGGTCTTAA/ideoxyU/CTACACGACGCCTTCCGATCT
554 Reverse Primer: ATTAAGACC/ideoxyU/AAGCAGTGGTATCAACGCAGAG
555 The PCR was run according to the manufacturer's instruction with two cycles at an annealing
556 temperature of 63°C followed by 7 cycles at an annealing temperature of 67°C; annealing time
557 was 30 seconds. Extension was performed at 72°C for 90 seconds. PCR products were
558 purified at 0.6X SPRIselect bead cleanup (Beckman Coulter, cat. no. B23318) according to
559 the manufacturer's instructions and eluted in 22 µL EB buffer (Qiagen, cat. no. 19086). DNA
560 concentrations were measured using the Qubit dsDNA HS Assay Kit (Thermo Fisher
561 Scientific, cat. no. Q32854), which were in the range of 1.5 µg per sample. cDNA traces were
562 additionally evaluated on a 5200 Fragment Analyzer System (Agilent Technologies) using the
563 HS NGS Fragment Kit, 1-6000 bp (Agilent, cat. no. DNF-474-0500). Full-length cDNAs were
564 enriched through capture on 5 µL streptavidin-coated M-280 dynabeads using the
565 Dynabeads™ kilobaseBINDER™ Kit (Invitrogen, cat. no. 60101), thus depleting TSO-TSO
566 artifacts. Washed Dynabeads containing the DNA-complexes were directly resuspended in 20
567 µL USER reaction buffer containing 10 µL StickTogether DNA Ligase Buffer 2x (NEB, cat. no.
568 B0535S), 1.5 µL USER Enzyme (NEB, cat. no. M5505S) and 8.5 µL Nuclease-free water
569 (Invitrogen, AM9939) and incubated in a thermocycler at 37°C for 20 min and held at 10°C (no
570 annealing). This created a nick at the deoxyuracil site forming palindrome overhangs and
571 releasing the biotin-bound DNA molecules from the beads. Beads were removed by magnetic

572 separation and the supernatant with the biotin-released cleaved PCR products was subjected
573 to a 0.6X SPRIselect cleanup step. Approximately 100 ng of purified product per sample were
574 split into two aliquots and subjected to a second PCR amplification step with 6 cycles using
575 an annealing temperature of 67°C. Reactions were pooled, purified by 0.6X SPRIselect
576 cleanup, and quality checked on both Qubit and Fragment Analyzer. Total DNA yield was
577 between 5-8 µg, which were subjected to a second round of streptavidin-purification using 10
578 µL of beads.

579 Transcript ligation

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

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

590 To enable PCR amplification of the ligated construct, the NEBNext Ultra II DNA Library Prep
591 Kit for Illumina was followed (NEB, cat. no. E7645S) using total DNA yield as input material.
592 2.5 µL of 5 µM dT overhang adapter (Roche, cat. no. KK8727) was used for the End Prep
593 reaction. Adapter-ligated libraries were purified by 0.39X SPRIselect cleanup, eluted in 22 µL
594 EB buffer and products were evaluated by HS Large Fragment kit. Total yield of around 40 ng
595 was split in two and PCR amplified using 2X KAPA HiFi Hot-Start ReadyMix (Roche, cat. no.
596 KK2602) and KAPA Library Amplification Primer Mix (10X concentration, Roche, cat. no.
597 KK2623), 10 µL library input each with 11 cycles and 9 min extension time. Following a 0.38X
598 SPRIselect cleanup and elution in 48 µL EB buffer, products were evaluated on a large
599 fragment gel revealing an average fragment length of libraries of 4.6 kb and an average total
600 of 1.1 µg DNA. To increase the total yield to 2 µg DNA required for SMRTbell library
601 preparation of a product with 5 kb amplicon size, the PCR was repeated with three additional
602 cycles and 5 min extension time. After 0.4X SPRI cleanup and Fragment Analyzer inspection,
603 the final yield was 2 µg per library.

604 PacBio SMRTbell library preparation

605 The SMRTbell Express Template Kit (PacBio, cat. no. 100-938-900) was used following
606 manufacturer's instructions for DNA damage repair, end repair/dA-tailing and ligation of a
607 hairpin adapter (double amount used). Final purification of the SMRTbell template was

608 performed by 0.42X SPRIselect cleanup and elution in 43 µL EB buffer. Exonuclease
609 treatment was performed by addition of 5 µL of NEBbuffer1 (NEB, cat. no. B7001S) and 1 µL
610 of each Exonuclease I (NEB, cat. no. M0293S) and Exonuclease III (NEB, cat. no. M0206S)
611 bringing the total volume to 50 µL per reaction. Enzyme treatment was performed at 37°C for
612 60 min. After SPRIselect cleanup, products were quantified on a large fragment gel at 1:30
613 dilution. Final yield was approximately 650 ng per sample, a sufficient amount for long-read
614 sequencing.

615 PacBio Sequel II sequencing

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

618 Genotyping PCR on genomic DNA

619 Genomic DNA was extracted from homogenized tumor tissue samples (n=8 samples matching
620 sampling time, Basel Ovarian Biobank) derived from patients using the DNeasy Blood &
621 Tissue kit (QIAGEN, cat. no. 69504). Isolated DNA underwent QC using Nanodrop and Qubit
622 measurements. Genotyping PCR on genomic DNA was performed using the MyTaq DNA
623 Polymerase system (Bioline, cat. no. BIO-21105). Briefly, MyTaq reaction buffer and MyTaq
624 DNA polymerase were pipetted together with 200 nM primer (Sigma Aldrich) pairs
625 (gPCR_IGF2BP2-TESPA1_Bp_F 5'-CCT GCT TTG AGG AGG GGA GGG A-3' &
626 gPCR_IGF2BP2-TESPA1_Bp_R 5'-ACT GAG GAC AAT GCT ACG CAA GA-3';
627 gPCR_TESPA1_F 5'-CCT GCT TTG AGG AGG GGA GGG A-3' & gPCR_TESPA1_R 5'-TGA
628 GAA CTG CTG TTC CAG GAG ACA-3'; gPCR_IGF2BP2_F 5'-ACA CTG GAC CCA TGC
629 TTG AGC T-3' & gPCR_IGF2BP2_R 5'-GCG TGC TAT GAA CAC TCC AGG CC-3'), and 50
630 ng genomic DNA (gDNA). PCR conditions were 1 cycle at 94°C for 5 min followed by 35 cycles
631 at 95°C for 20 sec, 58°C for 15 sec, 72°C for 1 min, and finished with 1 cycle at 72°C for 5
632 min. Amplicons were visualized on a 1.2 % agarose gel together with DNA ladder.
633

634 Immunofluorescence

635 Formalin-fixed and paraffin-embedded tissue samples were obtained from the Basel Ovarian
636 Biobank matching with patients 1 and 2 on sampling time and site. Briefly, samples were
637 deparaffinized and immersed for 10 min in a 10 mM sodium citrate buffer at pH 6.0 (Sigma
638 Aldrich, cat. no. C9999) at 95°C for antigen retrieval. Samples were permeabilized in 0.25%
639 (v/v) Triton™ X-100 (Roth, cat. no. 3051.3) in PBS (Sigma Aldrich, cat. no. D8537-500M) for
640 5 min and blocked in 5% FBS (Sigma Aldrich, cat. no. F7524-500ml), 0.1% Triton™ X-100,
641 1% BSA (Roth, Fraction V, cat. no. 8076.4) in PBS for 1 hour. The following antibodies were
642 used for this study: rabbit IGF2BP2 (C-terminal-specific, ThermoFisher Scientific, cat. no.
643 MA5-42874), EpCAM (Cell Signaling Technologies, cat. no. 5488S,) and goat anti-rabbit Alexa
644 Fluor® 647 (Cell Signaling Technology, cat. no. 4414). Slides were mounted using ProLong®

645 Gold Antifade Reagent with DAPI (Cell Signaling Technology, cat. no. 8961). Images were
646 acquired using a Nikon spinning-disk confocal microscope (Nikon CSU-W1 spinning-disk
647 confocal microscope) and processed with Fiji. Cell quantification was performed using an in-
648 house developed QuPath script for cell detection and annotations.

649 Single-cell DNA-sequencing

650 Cell suspensions were loaded and processed using the 10x Genomics Chromium platform
651 with the single-cell CNV kit on the 10x Genomics Chromium Single Cell Controller (10x
652 Genomics, cat. no. PN-120263) according to the manufacturer's instructions. Paired-end
653 sequencing was performed on the Illumina NovaSeq platform (100 cycles, 380pm loading
654 concentration with 1% addition of PhiX) at recommended sequencing depth.

655 Data Analysis

656 Short-read data analysis

657 Preprocessing

658 Raw reads were mapped to the GRCh38 reference genome using 10x Genomics Cell Ranger
659 3.1.0 to infer read counts per gene per cell. We performed index-hopping removal using a
660 method developed by Griffiths *et al.*[93].

661 10x Genomics read data processing, normalization, and visualization.

662 Expression data of each sample was analyzed using the scAmpi workflow[94]. In brief, UMI
663 counts were quality-controlled and cells and genes were filtered to remove mitochondrial and
664 ribosomal contaminants. Cells for which over 50% of the reads mapped to mitochondrial genes
665 and cells with fewer than 400 genes expressed were removed. By default, all non-protein-
666 coding genes, genes coding for ribosomal and mitochondrial proteins, and genes that were
667 expressed in less than 20 cells were removed. Doublet detection was performed using
668 scDblFinder[95]. Subsequently, counts were normalized with sctransform[96], regressing out cell
669 cycle effects, library size, and sample effects as non-regularized dependent variables. Similar
670 cells were grouped based on unsupervised clustering using Phenograph[97], and automated
671 cell type classification was performed independently for each cell[98] using gene lists defining
672 highly expressed genes in different cell types. Major cell type marker lists were developed in-
673 house based on unpublished datasets (manuscripts in preparation), including the Tumor
674 Profiler Study[99], using the Seurat FindMarkers method[100]. Immune subtype marker gene lists
675 were obtained from Newman *et al.*[101] and enriched with T cell subtypes from Sade-Feldman
676 *et al.*[102] The results of the unsupervised clustering and cell typing are visualized in a low-
677 dimensional representation using Uniform Manifold Approximation and Projection (UMAP).

Long-read data analysis

Generating CCS

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

Unconcatenating long-reads

We used NCBI BLAST (version 2.5.0+) to map the 5' and 3' primers to CCS constructs, with parameters: "-outfmt 7 -word_size 5" as described previously[24]. Sequences between two successive primers were used as input for primer trimming using IsoSeq3 Lima (parameters: --isoseq --dump-clips --min-passes 3). Cell barcodes and UMIs were then demultiplexed using IsoSeq3 tag with parameter --design T-12U-16B. Finally, we used IsoSeq3 refine with option --require-polya to remove concatemers and trim polyA tails. Only reads with a correct 5'-3' primer pair, a barcode also found in the short-read data, a UMI, and a polyA tail were retained.

Isoform classification

Demultiplexing UMIs with IsoSeq3 dedup and calling isoforms on the cohort level with collapse_isoforms_by_sam.py resulted in unfeasible runtimes. Therefore, we called isoforms first on the cell level as a pre-filtering step. Long reads were split according to their cell barcodes, and UMI deduplication was performed using IsoSeq3 dedup. Next, reads were mapped and aligned to the reference genome (hg38) using minimap2 with parameters: -ax splice -uf --secondary=no -C5. Identical isoforms were merged based on their aligned exonic structure using collapse_isoforms_by_sam.py with parameters: -c 0.99 -i 0.95 --gen_mol_count. We then classified isoforms using SQANTI3[32] with arguments: --skipORF --fl_count --skip_report. We finally filtered artifacts including intrapriming (accidental priming of pre-mRNA 'A's), reverse-transcriptase template switching artifacts, and mismapping to non-canonical junctions. To create a unique isoform catalog for all our samples, we then retained only reads associated with isoforms passing the SQANTI3 filter, and we ran collapse_isoforms_by_sam.py, SQANTI3 classification and filtering again on all cells together. The described pipeline is available here and was implemented in Snakemake, a reproducible and scalable workflow management system[103].

3' and 5' isoform filtering

For SQANTI3-defined isoforms, incomplete splice match, novel in catalog, and novel not in catalog, we only retained isoforms falling within 50 bp of a CAGE-validated transcription start

site (FANTOM5 CAGE database), and 50 bp of a polyA site form the PolyASite database[34]. The GENCODE database was used as a comparison, all protein-coding isoforms were grouped under the GENCODE.full label, a subset including only full-length isoforms was labeled as GENCOD.FL, and the Matched Annotation from NCBI and EMBL-EBI (MANE[35]) was named GENCODE.MANE.

Cell type-specific isoforms

Considering only the SQANTI3-defined 'full splice match', 'novel not in catalog' and 'novel in catalog' isoforms with at least 3 reads, we established the following classification: "Cell-specific" isoforms are present in only 1 cell and "cell type-specific" isoforms are present in >=3 cells of a unique cell type.

Cell type annotation

Cells were annotated with long-reads the same way as short-reads, using scROSHI. The major cell types were modified according to gene expression in long-reads. Immune subtype marker gene lists were unchanged.

Normalization and visualization

Long-read gene expression counts were normalized and visualized as described above for short reads. Long-read gene expression counts were normalized using 10,000 features instead of the default 3,000 in sctransform[96].

Mutation detection

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

Alternative polyadenylation (APA) analysis

To analyze differences between 3'UTR lengths, we used a modified version of DaPars2[104] (https://github.com/ArthurDondi/DaPars2_LR), with an APA site detection adapted to long-read coverage.

Briefly, we identified the 3'UTR exon of each isoform, and overlapping 3'UTR exons with different 5' start positions were discarded from analysis, as they create false positive APAs. Then, for each remaining 3'UTR, we computed the coverage for each cell type. The distal site position was defined as the most 3' position with coverage superior to 10 in any cell type:

$L^* = max(\{k: w_k^c > 10, 1 < k < L, 1 < c < m\})$
where L is the length of the annotated 3′ UTR region and $L*$ is the defined distal site position. w_k^c is the coverage in cell type c at position k, and m is the total number of cell types.

We inferred the exact location of APA sites by maximizing the coverage gap between the 50 positions before and after the possible APA sites, based on the two-polyA-site model, the most common model of APA regulation:

$$(C^*, P^*) = argmax_{1<c<m, 1<P<L^*} \left(\frac{1}{50} \sum_{i=1}^{50} w_{P+i}^c - \frac{1}{50} \sum_{i=1}^{50} w_{P-i}^c \right)^2$$

where P is the estimated length of alternative proximal 3′ UTR, and the optimal proximal site $P*$ in cell type $C*$ is the one with the maximal objective function value. w_k^c is the coverage at the position k and cell type c, and m is the total number of cell types.

Then distal and proximal site coverages W_d and W_p in condition C were defined as:
$W_d = \frac{1}{50} \sum_{i=1}^{50} w_{L^*-i}^c$, $W_p = \frac{1}{50} \sum_{i=1}^{50} w_{P^*-i}^c$
The fraction of distal polyA site usage is then defined as $F = \frac{w_d}{w_d + w_p}$

The degree of difference in APA usage in cell types C_1 and C_2 can be quantified as a $Fraction\ Change$, which is capable of identifying 3′ UTR lengthening (positive index) or shortening (negative index):

$Fraction\ Change = F_{C_1} - F_{C_2}$.

Differential isoform tests

Differential isoform testing was performed using a χ2 test as previously described in Scisorseq[26]. Briefly, counts for each isoform ID were assigned to individual cell types, and genes were discarded if they were mitochondrial, ribosomal, or if they did not reach sufficient depth per condition (25 reads per condition per gene). P-values from a χ2 test for differential isoform usage were computed per gene where a sufficient depth was reached, and we corrected for multiple testing using Benjamini Hochberg correction with a 5% false discovery rate. If the corrected p-value was ≤0.05 and the sum of change in the relative percent of isoform (ΔΠ) of the top two isoforms in either positive or negative direction was more than 20%, then the gene was called differentially spliced. To classify the top differentially spliced genes, we took the rank of genes by ΔΠ and corrected p-values, and took the square root of the multiplication of those two ranks. The smallest ranks obtained this way were considered

781 as the top differentially expressed genes. Differentially used isoforms were visualized using
782 ScisorWiz[105].

783 Isoform biotypes

784 Novel isoforms biotypes were assessed internally by the GENCODE team with biotypes
785 matching those described by Frankish *et al.*[106]. Known isoform biotypes were retrieved from
786 the GENCODE v36 database.

787 Biotype change analysis

788 In genes with differential isoform usage between conditions, we compared the biotypes of the
789 most expressed isoform of each condition, and if they were not identical this was considered
790 a change in biotype.

791 Protein-coding gene expression

792 Considering only genes with protein-coding biotype (GENCODE) with a minimal expression
793 of 20 UMIs in a condition, we assessed the ratio of non-coding isoforms (NMD, intron retention,
794 etc.) expression to total gene expression (non-coding and coding).

795 Pathway enrichment analysis

796 We used GSVA to perform pathway enrichment analysis. When comparing TME and distal
797 cells, gene sets were obtained from MSigDB[107], except for the miR-29 targets that were
798 obtained from Cushing and colleagues[108]. When we searched for pathways enriched in
799 lengthened or shortened 3'UTRs in cancer cells, we used the investigate function of the GSEA
800 webpage (https://www.gsea-msigdb.org/gsea/msigdb/human/annotate.jsp), and compared
801 against all MSigDB gene sets[107].

802 Fusion Discovery

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

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

810 We designed a custom reference including *IGF2BP2::TESPA1* transcriptomic breakpoint as
811 well as the wild-type *IGF2BP2* and *TESPA1* exon junction covering the breakpoint. The
812 reference was composed of 6 sequences of 80 nucleotides (40 bases upstream and

813 downstream of the breakpoint), sequences XXX_1 and XXX_2 represent the breakpoints of
814 the two main isoforms seen in each gene:
815
816 >TESPA1_wt_1
817 TTCTGTCAGACCACATGCTGTTGTGGTGGTGGAGAAAGCAATTCTGGAGGCTGGCAAATCCAAG
818 GTCAAAAGCCTGCA
819
820 >TESPA1_wt_2
821 TTCTGTCAGACCACATGCTGTTGTGGTGGTGGAGAAAGCTTCACGAGTCTTTGCCAGCAAAAGTC
822 TGGTGGTGGTGGG
823
824 >IGF2BP2_wt_1
825 ATGTGACGTTGACAACGGCGGTTTCTGTGTCTGTGTTGACTTGTTCCACATTCTCCACTGTCCCA
826 TATTGAGCCAAAA
827
828 >IGF2BP2_wt_2
829 ATCACTGGATTGTGTGTTCTTCTGAATTACTTCTTTAGGCTTGTTCCACATTCTCCACTGTCCCAT
830 ATTGAGCCAAAA
831
832 >TESPA1_IGF2BP2_fusion_1
833 TTCTGTCAGACCACATGCTGTTGTGGTGGTGGAGAAAGCCTTGTTCCACATTCTCCACTGTCCCA
834 TATTGAGCCAAAA
835
836 >TESPA1_IGF2BP2_fusion_2
837 CAAATCCAAGGTCAAAAGCCTGCATCTGGTGAGGGCCTCCTTGTTCCACATTCTCCACTGTCCCA
838 TATTGAGCCAAAA
839
840 Patient 2 reads were aligned to this reference using minimap2 with parameters: -ax sr --
841 secondary=no. Reads mapping unambiguously to one of those reference sequences were
842 then attributed to the cell type to which their cell barcode belonged.

843 scDNA analysis

844 Cell Ranger DNA was used to demultiplex and align Chromium-prepared sequencing
845 samples. We used the cellranger-dna mkfastq command to generate FASTQ files from the
846 Illumina raw BCL files, and we ran the command cellranger-dna cnv to align FASTQ files to
847 the hg38 reference genome, call cells, and estimate copy numbers. We obtained the copy
848 number profiles and detected the main clonal structure of samples using SCICoNE[109].

849 DNA breakpoint validation

850 To validate in scDNA data breakpoints found in scRNA data, we used the putative scRNA
851 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.

868 # Data availability

869 The raw sequencing files reported in this study have been deposited in the European Genome-
870 phenome Archive (EGA) under the accession number EGAS00001006807. Patient 2 scDNA
871 reads mapping the *IGF2BP2::TESPA1* fusion breakpoint are available at https://eth-
872 nexus.github.io/tu-pro_website/publications/dondi_et_al_2022/. TCGA data was obtained at
873 https://portal.gdc.cancer.gov/repository, with the following specifications: Project=TCGA-OV,
874 DataType=Gene Expression Quantification.
875

876 # Code availability

877 The code used to pre-process and analyze the data of this study has been deposited in the
878 GitHub repository: https://github.com/cbg-ethz/scIsoPrep. The code to analyze differential
879 poly-adenylation site in long-reads has been deposited in the GitHub repository:
880 https://github.com/ArthurDondi/DaPars2_LR.

881 # Acknowledgements

882 We thank Ina Nissen (Genomics Facility Basel) for technical support with PacBio sequencing,
883 Ching-Yeu Liang (University Hospital Basel) for assistance with sample dissociation, and Anne
884 Bertolini for her help with the scAmpi R package. We also wish to thank Elisabeth Tseng
885 (Pacific Bioscience) for her help with cDNA_Cupcake, SQANTI3, and long-read sequencing
886 analysis, Anoushka Joglekar (Weill Cornell Medicine) for her help with ScisorSeq and
887 ScisorWiz as well as Lucia Csepregi for her help with immune repertoire analysis. We thank
888 Olivier Belli for his help with 3'UTR shortening mechanisms. Furthermore, we are grateful to
889 Adam Frankish (Wellcome Sanger Institute) and the GENCODE Project team for their help
890 with the annotation of isoforms. Graphical illustrations were created with BioRender.com.
891 Illumina sequencing was carried out in the Genomics Facility Basel of the University of Basel
892 and the Department of Biosystems Science and Engineering, ETH Zurich. PacBio sequencing
893 was done in the Functional Genomics Center Zurich of the University of Zurich and ETH
894 Zurich, and the Lausanne Genomic Technologies Facility, University of Lausanne.

895 # Funding Information

896 Part of this work was funded by the SNSF SPARK grant #190413, the grants #2017-510 of
897 the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH
898 Domain, the "Personalized Health and Related Technologies (PHRT)" for Pioneer Project
899 #715 and the European Union's Horizon 2020 research and innovation program under the
900 Marie Sklodowska-Curie grant agreement #766030.

Author contributions

U. L, and C. B. acquired funding and conceived and designed the experiments. V. H. S provided patient material. U. L., F .J., A. D., and C. B. selected the clinical cohort. U. L. 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. A. D. designed the pre-processing pipeline and implemented it with the help of N. Borgsmueller. A. D. conducted all computational analyses. F. J. and R. C. performed experimental validation. F. S. assisted in short-read scRNA-seq analysis. A. D., U. L., F. J., N. Beerenwinkel, and C. B. interpreted the data. A. D. and U. L. wrote the manuscript with contributions from 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],

938 Walter P Weber[19], Rebekka Wegmann[2], Michael Weller[33], Fabian Wendt[7], Norbert Wey[31],
939 Andreas Wicki[30,36], Mattheus HE Wildschut[2,30], Bernd Wollscheid[7], Shuqing Yu[9,17], Johanna
940 Ziegler[28], Marc Zimmermann[6,17,26], Martin Zoche[31], Gregor Zuend[32]

941 [1]Cantonal Hospital Baselland, Medical University Clinic, Rheinstrasse 26, 4410 Liestal,
942 Switzerland, [2]ETH Zurich, Department of Biology, Institute of Molecular Systems Biology,
943 Otto-Stern-Weg 3, 8093 Zurich, Switzerland, [3]ETH Zurich, Department of Biology, Wolfgang-
944 Pauli-Strasse 27, 8093 Zurich, Switzerland, [4]ETH Zurich, Department of Biosystems Science
945 and Engineering, Mattenstrasse 26, 4058 Basel, Switzerland, [5]ETH Zurich, Department of
946 Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 1-5/10, 8093 Zurich, Switzerland,
947 [6]ETH Zurich, Department of Computer Science, Institute of Machine Learning,
948 Universitätstrasse 6, 8092 Zurich, Switzerland, [7]ETH Zurich, Department of Health Sciences
949 and Technology, Otto-Stern-Weg 3, 8093 Zurich, Switzerland, [8]ETH Zurich, Institute of
950 Molecular Health Sciences, Otto-Stern-Weg 7, 8093 Zurich, Switzerland, [9]ETH Zurich,
951 NEXUS Personalized Health Technologies, John-von-Neumann-Weg 9, 8093 Zurich,
952 Switzerland, [10]F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland,
953 [11]F. Hoffmann-La Roche Ltd, Grenzacherstrasse 124, 4070 Basel, Switzerland, [12]Max Planck
954 ETH Center for Learning Systems, [13]Roche Diagnostics GmbH, Nonnenwald 2, 82377
955 Penzberg, Germany, [14]Roche Pharmaceutical Research and Early Development, Roche
956 Innovation Center Basel, Grenzacherstrasse 124, 4070 Basel, Switzerland, [15]Roche
957 Pharmaceutical Research and Early Development, Roche Innovation Center Munich, Roche
958 Diagnostics GmbH, Nonnenwald 2, 82377 Penzberg, Germany, [16]Roche Pharmaceutical
959 Research and Early Development, Roche Innovation Center Zurich, Wagistrasse 10, 8952
960 Schlieren, Switzerland, [17]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland,
961 [18]University Hospital Basel and University of Basel, Department of Biomedicine, Hebelstrasse
962 20, 4031 Basel, Switzerland, [19]University Hospital Basel and University of Basel, Department
963 of Surgery, Brustzentrum, Spitalstrasse 21, 4031 Basel, Switzerland, [20]University Hospital
964 Basel, Brustzentrum & Tumorzentrum, Petersgraben 4, 4031 Basel, Switzerland, [21]University
965 Hospital Basel, Brustzentrum, Spitalstrasse 21, 4031 Basel, Switzerland, [22]University Hospital
966 Basel, Department of Information- and Communication Technology, Spitalstrasse 26, 4031
967 Basel, Switzerland, [23]University Hospital Basel, Gynecological Cancer Center, Spitalstrasse
968 21, 4031 Basel, Switzerland, [24]University Hospital Basel, Institute of Medical Genetics and
969 Pathology, Schönbeinstrasse 40, 4031 Basel, Switzerland, [25]University Hospital Basel,
970 Spitalstrasse 21/Petersgraben 4, 4031 Basel, Switzerland, [26]University Hospital Zurich,
971 Biomedical Informatics, Schmelzbergstrasse 26, 8006 Zurich, Switzerland, [27]University
972 Hospital Zurich, Clinical Trials Center, Rämistrasse 100, 8091 Zurich, Switzerland, [28]University
973 Hospital Zurich, Department of Dermatology, Gloriastrasse 31, 8091 Zurich, Switzerland,
974 [29]University Hospital Zurich, Department of Gynecology, Frauenklinikstrasse 10, 8091 Zurich,
975 Switzerland, [30]University Hospital Zurich, Department of Medical Oncology and Hematology,
976 Rämistrasse 100, 8091 Zurich, Switzerland, [31]University Hospital Zurich, Department of

977 Pathology and Molecular Pathology, Schmelzbergstrasse 12, 8091 Zurich, Switzerland,
978 [32]University Hospital Zurich, Rämistrasse 100, 8091 Zurich, Switzerland, [33]University Hospital
979 and University of Zurich, Department of Neurology, Frauenklinikstrasse 26, 8091 Zurich,
980 Switzerland, [34]University of Bern, Department of BioMedical Research, Murtenstrasse 35,
981 3008 Bern, Switzerland, [35]University of Zurich, Department of Quantitative Biomedicine,
982 Winterthurerstrasse 190, 8057 Zurich, Switzerland, [36]University of Zurich, Faculty of Medicine,
983 Zurich, Switzerland, [37]University of Zurich, Institute of Molecular Life Sciences,
984 Winterthurerstrasse 190, 8057 Zurich, Switzerland, [38]University of Zurich, Services and
985 Support for Science IT, Winterthurerstrasse 190, 8057 Zurich, Switzerland, [39]University of
986 Zurich, VP Medicine, Künstlergasse 15, 8001 Zurich, Switzerland

987 Conflict of interest

988 The authors declare no competing interests.

989 # References

990 1. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17–37
991 (2013).
992 2. Hanahan, D. Hallmarks of cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
993 3. Hertzman Johansson, C. & Egyhazi Brage, S. BRAF inhibitors in cancer therapy.
994 *Pharmacol. Ther.* **142**, 176–182 (2014).
995 4. Li, J. *et al.* A functional genomic approach to actionable gene fusions for precision
996 oncology. *Sci. Adv.* **8**, eabm2382 (2022).
997 5. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science*
998 **348**, 69–74 (2015).
999 6. Yu, Y.-P. *et al.* Identification of recurrent fusion genes across multiple cancer types. *Sci.*
1000 *Rep.* **9**, 1074 (2019).
1001 7. Bower, H. *et al.* Life expectancy of patients with chronic myeloid leukemia approaches
1002 the life expectancy of the general population. *J. Clin. Oncol.* **34**, 2851–2857 (2016).
1003 8. Khan, M. *et al.* ALK inhibitors in the treatment of ALK positive NSCLC. *Front. Oncol.* **8**,
1004 557 (2018).
1005 9. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature*
1006 **500**, 415–421 (2013).
1007 10. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* **45**,
1008 1134–1140 (2013).
1009 11. Gao, Q. *et al.* Driver fusions and their implications in the development and treatment of
1010 human cancers. *Cell Rep.* **23**, 227-238.e3 (2018).
1011 12. Wang, Y., Shi, T., Song, X., Liu, B. & Wei, J. Gene fusion neoantigens: Emerging targets
1012 for cancer immunotherapy. *Cancer Lett.* **506**, 45–54 (2021).
1013 13. Wei, Z. *et al.* The Landscape of Tumor Fusion Neoantigens: A Pan-Cancer Analysis.
1014 *iScience* **21**, 249–260 (2019).
1015 14. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126**, 37–47
1016 (2006).
1017 15. Bonnal, S. C., López-Oreja, I. & Valcárcel, J. Roles and mechanisms of alternative
1018 splicing in cancer - implications for care. *Nat. Rev. Clin. Oncol.* **17**, 457–474 (2020).
1019 16. Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from
1020 8,705 Patients. *Cancer Cell* **34**, 211-224.e6 (2018).
1021 17. Marusyk, A. & Polyak, K. Tumor heterogeneity: causes and consequences. *Biochim.*
1022 *Biophys. Acta* **1805**, 105–117 (2010).
1023 18. Whiteside, T. L. The tumor microenvironment and its role in promoting tumor growth.
1024 *Oncogene* **27**, 5904–5912 (2008).

1025 19. Hedlund, E. & Deng, Q. Single-cell RNA sequencing: Technical advancements and
1026 biological applications. *Mol. Aspects Med.* **59**, 36–46 (2018).
1027 20. Marx, V. Method of the year: long-read sequencing. *Nat. Methods* **20**, 6–11 (2023).
1028 21. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read
1029 sequencing. *Nature* **608**, 353–359 (2022).
1030 22. Philpott, M. *et al.* Nanopore sequencing of single-cell transcriptomes with scCOLOR-
1031 seq. *Nat. Biotechnol.* **39**, 1517–1520 (2021).
1032 23. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands
1033 of cerebellar cells. *Nat. Biotechnol.* **36**, 1197–1202 (2018).
1034 24. Shi, Z.-X. *et al.* High-throughput and high-accuracy single-cell RNA isoform analysis
1035 using PacBio circular consensus sequencing. *Nat. Commun.* **14**, 2631 (2023).
1036 25. Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmed
1037 cDNA concatenation. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01815-7.
1038 26. Joglekar, A. *et al.* A spatially resolved brain region- and cell type-specific isoform atlas
1039 of the postnatal mouse brain. *Nat. Commun.* **12**, 463 (2021).
1040 27. Hardwick, S. A. *et al.* Single-nuclei isoform RNA sequencing unlocks barcoded exon
1041 connectivity in frozen brain tissue. *Nat. Biotechnol.* **40**, 1082–1092 (2022).
1042 28. Veiga, D. F. T. *et al.* A comprehensive long-read isoform analysis platform and
1043 sequencing resource for breast cancer. *Sci. Adv.* **8**, eabg6711 (2022).
1044 29. Namba, S. *et al.* Transcript-targeted analysis reveals isoform alterations and double-
1045 hop fusions in breast cancer. *Commun. Biol.* **4**, 1320 (2021).
1046 30. Macintyre, G. *et al.* Copy number signatures and mutational processes in ovarian
1047 carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
1048 31. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian
1049 carcinoma. *Nature* **474**, 609–615 (2011).
1050 32. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript
1051 sequences for quality control in full-length transcriptome identification and quantification.
1052 *Genome Res.* **28**, 396–411 (2018).
1053 33. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for
1054 GRCh38 and GRCm38 genome assemblies. *Sci. Data* **4**, 170107 (2017).
1055 34. Herrmann, C. J. *et al.* PolyASite 2.0: a consolidated atlas of polyadenylation sites from
1056 3' end sequencing. *Nucleic Acids Res.* **48**, D174–D179 (2020).
1057 35. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and
1058 research. *Nature* **604**, 310–315 (2022).
1059 36. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP.
1060 *Nat. Biotechnol.* **37**, 38–44 (2018).
1061 37. Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and

1062 polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673–684 (2009).
1063 38. Maurer, B. *et al.* MicroRNA-29, a key regulator of collagen expression in systemic
1064 sclerosis. *Arthritis Rheum.* **62**, 1733–1743 (2010).
1065 39. Chen, Y. & Wang, X. miRDB: an online database for prediction of functional microRNA
1066 targets. *Nucleic Acids Res.* **48**, D127–D131 (2020).
1067 40. Jin, Y., Song, X., Sun, X. & Ding, Y. Up-regulation of collagen type V alpha 2 (COL5A2)
1068 promotes malignant phenotypes in gastric cancer cell via inducing epithelial-
1069 mesenchymal transition (EMT). *Open Med (Wars)* **18**, 20220593 (2023).
1070 41. Shintani, Y., Maeda, M., Chaika, N., Johnson, K. R. & Wheelock, M. J. Collagen I
1071 promotes epithelial-to-mesenchymal transition in lung cancer cells via transforming
1072 growth factor-beta signaling. *Am. J. Respir. Cell Mol. Biol.* **38**, 95–104 (2008).
1073 42. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set
1074 collection. *Cell Syst.* **1**, 417–425 (2015).
1075 43. Owusu-Ansah, K. G. *et al.* COL6A1 promotes metastasis and predicts poor prognosis
1076 in patients with pancreatic cancer. *Int. J. Oncol.* **55**, 391–404 (2019).
1077 44. Ramadoss, S., Chen, X. & Wang, C.-Y. Histone demethylase KDM6B promotes
1078 epithelial-mesenchymal transition. *J. Biol. Chem.* **287**, 44508–44517 (2012).
1079 45. Du, B. *et al.* The potential role of TNFAIP3 in malignant transformation of gastric
1080 carcinoma. *Pathol. Res. Pract.* **215**, 152471 (2019).
1081 46. D'Arrigo, P. *et al.* The splicing FK506-binding protein-51 isoform plays a role in
1082 glioblastoma resistance through programmed cell death ligand-1 expression regulation.
1083 *Cell Death Discov.* **5**, 137 (2019).
1084 47. Jie, W. *et al.* Pathophysiological functions of rnd3/rhoe. *Compr. Physiol.* **6**, 169–186
1085 (2015).
1086 48. Han, Q. *et al.* Omental cancer-associated fibroblast-derived exosomes with low
1087 microRNA-29c-3p promote ovarian cancer peritoneal metastasis. *Cancer Sci.* **114**,
1088 1929–1942 (2023).
1089 49. Chou, J. *et al.* GATA3 suppresses metastasis and modulates the tumour
1090 microenvironment by regulating microRNA-29b expression. *Nat. Cell Biol.* **15**, 201–213
1091 (2013).
1092 50. Smyth, A., Callaghan, B., Willoughby, C. E. & O'Brien, C. The Role of miR-29 Family in
1093 TGF-β Driven Fibrosis in Glaucomatous Optic Neuropathy. *Int. J. Mol. Sci.* **23**, (2022).
1094 51. Yu, C.-C., Liao, Y.-W., Hsieh, P.-L. & Chang, Y.-C. Targeting lncRNA H19/miR-
1095 29b/COL1A1 Axis Impedes Myofibroblast Activities of Precancerous Oral Submucous
1096 Fibrosis. *Int. J. Mol. Sci.* **22**, (2021).
1097 52. Tian, X., Zuo, X., Hou, M., Li, C. & Teng, Y. LncRNA-H19 regulates chemoresistance to
1098 carboplatin in epithelial ovarian cancer through microRNA-29b-3p and STAT3. *J.*

1099 *Cancer* **12**, 5712–5722 (2021).
1100 53. Lv, M. *et al.* lncRNA H19 regulates epithelial-mesenchymal transition and metastasis of
1101 bladder cancer by miR-29b-3p as competing endogenous RNA. *Biochim. Biophys. Acta*
1102 *Mol. Cell Res.* **1864**, 1887–1899 (2017).
1103 54. An, Q. *et al.* circKRT7-miR-29a-3p-COL1A1 Axis Promotes Ovarian Cancer Cell
1104 Progression. *Onco Targets Ther* **13**, 8963–8976 (2020).
1105 55. Panda, A. C. Circular RNAs Act as miRNA Sponges. *Adv. Exp. Med. Biol.* **1087**, 67–79
1106 (2018).
1107 56. García-Bartolomé, A. *et al.* Altered Expression Ratio of Actin-Binding Gelsolin Isoforms
1108 Is a Novel Hallmark of Mitochondrial OXPHOS Dysfunction. *Cells* **9**, (2020).
1109 57. Chen, Z.-Y., Wang, P.-W., Shieh, D.-B., Chiu, K.-Y. & Liou, Y.-M. Involvement of gelsolin
1110 in TGF-beta 1 induced epithelial to mesenchymal transition in breast cancer cells. *J.*
1111 *Biomed. Sci.* **22**, 90 (2015).
1112 58. Philippou, A., Maridaki, M., Pneumaticos, S. & Koutsilieris, M. The complexity of the
1113 IGF1 gene splicing, posttranslational modification and bioactivity. *Mol. Med.* **20**, 202–
1114 214 (2014).
1115 59. Tagami, S., Eguchi, Y., Kinoshita, M., Takeda, M. & Tsujimoto, Y. A novel protein, RTN-
1116 XS, interacts with both Bcl-XL and Bcl-2 on endoplasmic reticulum and reduces their
1117 anti-apoptotic activity. *Oncogene* **19**, 5736–5746 (2000).
1118 60. Gong, L. *et al.* RTN1-C mediates cerebral ischemia/reperfusion injury via ER stress and
1119 mitochondria-associated apoptosis pathways. *Cell Death Dis.* **8**, e3080 (2017).
1120 61. Parreno, J., Amadeo, M. B., Kwon, E. H. & Fowler, V. M. Tropomyosin 3.1 association
1121 with actin stress fibers is required for lens epithelial to mesenchymal transition. *Invest.*
1122 *Ophthalmol. Vis. Sci.* **61**, 2 (2020).
1123 62. Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human
1124 prostate adenocarcinoma and reference samples. *BMC Med. Genomics* **4**, 11 (2011).
1125 63. Banday, A. R. *et al.* Genetic regulation of OAS1 nonsense-mediated decay underlies
1126 association with COVID-19 hospitalization in patients of European and African
1127 ancestries. *Nat. Genet.* **54**, 1103–1116 (2022).
1128 64. Choy, J. Y. H., Boon, P. L. S., Bertin, N. & Fullwood, M. J. A resource of ribosomal RNA-
1129 depleted RNA-Seq data from different normal adult and fetal human tissues. *Sci. Data*
1130 **2**, 150063 (2015).
1131 65. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev.*
1132 *Cancer* **20**, 555–572 (2020).
1133 66. Tate, J. G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids*
1134 *Res.* **47**, D941–D947 (2019).
1135 67. Mertens, F., Johansson, B., Fioretos, T. & Mitelman, F. The emerging complexity of

1136 gene fusions in cancer. *Nat. Rev. Cancer* **15**, 371–381 (2015).
1137 68. Yang, W. *et al.* Immunogenic neoantigens derived from gene fusions stimulate T cell
1138 responses. *Nat. Med.* **25**, 767–775 (2019).
1139 69. Berger, M. F. & Mardis, E. R. The emerging clinical relevance of genomics in cancer
1140 medicine. *Nat. Rev. Clin. Oncol.* **15**, 353–365 (2018).
1141 70. Rynne-Vidal, A. *et al.* Mesothelial-to-mesenchymal transition as a possible therapeutic
1142 target in peritoneal metastasis of ovarian cancer. *J. Pathol.* **242**, 140–151 (2017).
1143 71. Cai, J. *et al.* Fibroblasts in omentum activated by tumor cells promote ovarian cancer
1144 growth, adhesion and invasiveness. *Carcinogenesis* **33**, 20–29 (2012).
1145 72. Kenny, H. A. *et al.* Mesothelial cells promote early ovarian cancer metastasis through
1146 fibronectin secretion. *J. Clin. Invest.* **124**, 4614–4628 (2014).
1147 73. Qin, W. *et al.* TGF-β/Smad3 signaling promotes renal fibrosis by inhibiting miR-29. *J.*
1148 *Am. Soc. Nephrol.* **22**, 1462–1474 (2011).
1149 74. Yu, P.-N. *et al.* Downregulation of miR-29 contributes to cisplatin resistance of ovarian
1150 cancer cells. *Int. J. Cancer* **134**, 542–551 (2014).
1151 75. Dasari, S., Fang, Y. & Mitra, A. K. Cancer associated fibroblasts: naughty neighbors that
1152 drive ovarian cancer progression. *Cancers (Basel)* **10**, (2018).
1153 76. Li, H. *et al.* IGF-IR signaling in epithelial to mesenchymal transition and targeting IGF-
1154 IR therapy: overview and new insights. *Mol. Cancer* **16**, 6 (2017).
1155 77. Liefers-Visser, J. A. L., Meijering, R. A. M., Reyners, A. K. L., van der Zee, A. G. J. & de
1156 Jong, S. IGF system targeted therapy: Therapeutic opportunities for ovarian cancer.
1157 *Cancer Treat. Rev.* **60**, 90–99 (2017).
1158 78. Yang, Y. *et al.* Tumor Suppressor microRNA-138 Suppresses Low-Grade Glioma
1159 Development and Metastasis via Regulating IGF2BP2. *Onco Targets Ther* **13**, 2247–
1160 2260 (2020).
1161 79. Wang, D. *et al.* Tespa1 is involved in late thymocyte development through the regulation
1162 of TCR-mediated signaling. *Nat. Immunol.* **13**, 560–568 (2012).
1163 80. Dhamija, S. & Menon, M. B. Non-coding transcript variants of protein-coding genes -
1164 what are they good for? *RNA Biol.* **15**, 1025–1031 (2018).
1165 81. Dhamija, S. & Diederichs, S. From junk to master regulators of invasion: lncRNA
1166 functions in migration, EMT and metastasis. *Int. J. Cancer* **139**, 269–280 (2016).
1167 82. Joglekar, A., Foord, C., Jarroux, J., Pollard, S. & Tilgner, H. U. From words to complete
1168 phrases: insight into single-cell isoforms using short and long reads. *Transcription* 1–13
1169 (2023) doi:10.1080/21541264.2023.2213514.
1170 83. Lischetti, U. *et al.* Dynamic thresholding and tissue dissociation optimization for CITE-
1171 seq identifies differential surface protein abundance in metastatic melanoma. *Commun.*
1172 *Biol.* **6**, 830 (2023).

1173 84. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from
1174 mRNA levels? *Nature* **547**, E19–E20 (2017).
1175 85. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from
1176 proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
1177 86. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on
1178 mRNA Abundance. *Cell* **165**, 535–550 (2016).
1179 87. Volden, R. & Vollmers, C. Single-cell isoform analysis in human immune cells. *Genome*
1180 *Biol.* **23**, 47 (2022).
1181 88. Byrne, A. *et al.* Single-cell long-read targeted sequencing reveals transcriptional
1182 variation in ovarian cancer. *BioRxiv* (2023) doi:10.1101/2023.07.17.549422.
1183 89. Dutton, G. CRISPR-Cas9 Technology Cuts Clutter from Sequencing Libraries. *Genetic*
1184 *Engineering & Biotechnology News* **41**, 24–25 (2021).
1185 90. Lang, F., Schrörs, B., Löwer, M., Türeci, Ö. & Sahin, U. Identification of neoantigens for
1186 individualized therapeutic cancer vaccines. *Nat. Rev. Drug Discov.* **21**, 261–282 (2022).
1187 91. Lin, M. J. *et al.* Cancer vaccines: the next immunotherapy frontier. *Nat. Cancer* **3**, 911–
1188 926 (2022).
1189 92. Hebelstrup, K. H. *et al.* UCE: A uracil excision (USER)-based toolbox for transformation
1190 of cereals. *Plant Methods* **6**, 15 (2010).
1191 93. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and
1192 removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 2667
1193 (2018).
1194 94. Bertolini, A. *et al.* scAmpi-A versatile pipeline for single-cell RNA-seq analysis from
1195 basics to clinics. *PLoS Comput. Biol.* **18**, e1010097 (2022).
1196 95. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D. Doublet
1197 identification in single-cell sequencing data using scDblFinder. *F1000Res.* **10**, 979
1198 (2021).
1199 96. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-
1200 seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
1201 97. Levine, J. H. *et al.* Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like
1202 Cells that Correlate with Prognosis. *Cell* **162**, 184–197 (2015).
1203 98. Prummer, M. *et al.* scROSHI: robust supervised hierarchical identification of single cells.
1204 *NAR Genom. Bioinform.* **5**, lqad058 (2023).
1205 99. Irmisch, A. *et al.* The Tumor Profiler Study: integrated, multi-omic, functional tumor
1206 profiling for clinical decision support. *Cancer Cell* **39**, 288–293 (2021).
1207 100. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.
1208 (2021).
1209 101. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles.

1210 *Nat. Methods* **12**, 453–457 (2015).
1211 102. Sade-Feldman, M. *et al.* Defining T Cell States Associated with Response to Checkpoint
1212 Immunotherapy in Melanoma. *Cell* **175**, 998-1013.e20 (2018).
1213 103. Köster, J. & Rahmann, S. Snakemake--a scalable bioinformatics workflow engine.
1214 *Bioinformatics* **28**, 2520–2522 (2012).
1215 104. Feng, X. *et al.* TC3A: The Cancer 3′ UTR Atlas, *Nucleic Acids Research*. **46**, 1027–1030
1216 (2018).
1217 105. Stein, A. N., Joglekar, A., Poon, C.-L. & Tilgner, H. U. ScisorWiz: visualizing differential
1218 isoform expression in single-cell long-read data. *Bioinformatics* **38**, 3474–3476 (2022).
1219 106. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes.
1220 *Nucleic Acids Res.* **47**, D766–D773 (2019).
1221 107. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**,
1222 1739–1740 (2011).
1223 108. Cushing, L. *et al.* miR-29 is a major regulator of genes associated with pulmonary
1224 fibrosis. *Am. J. Respir. Cell Mol. Biol.* **45**, 287–294 (2011).
1225 109. Kuipers, J., Tuncel, M. A., Ferreira, P., Jahn, K. & Beerenwinkel, N. Single-cell copy
1226 number calling and event history reconstruction. *BioRxiv* (2020)
1227 doi:10.1101/2020.04.28.065755.

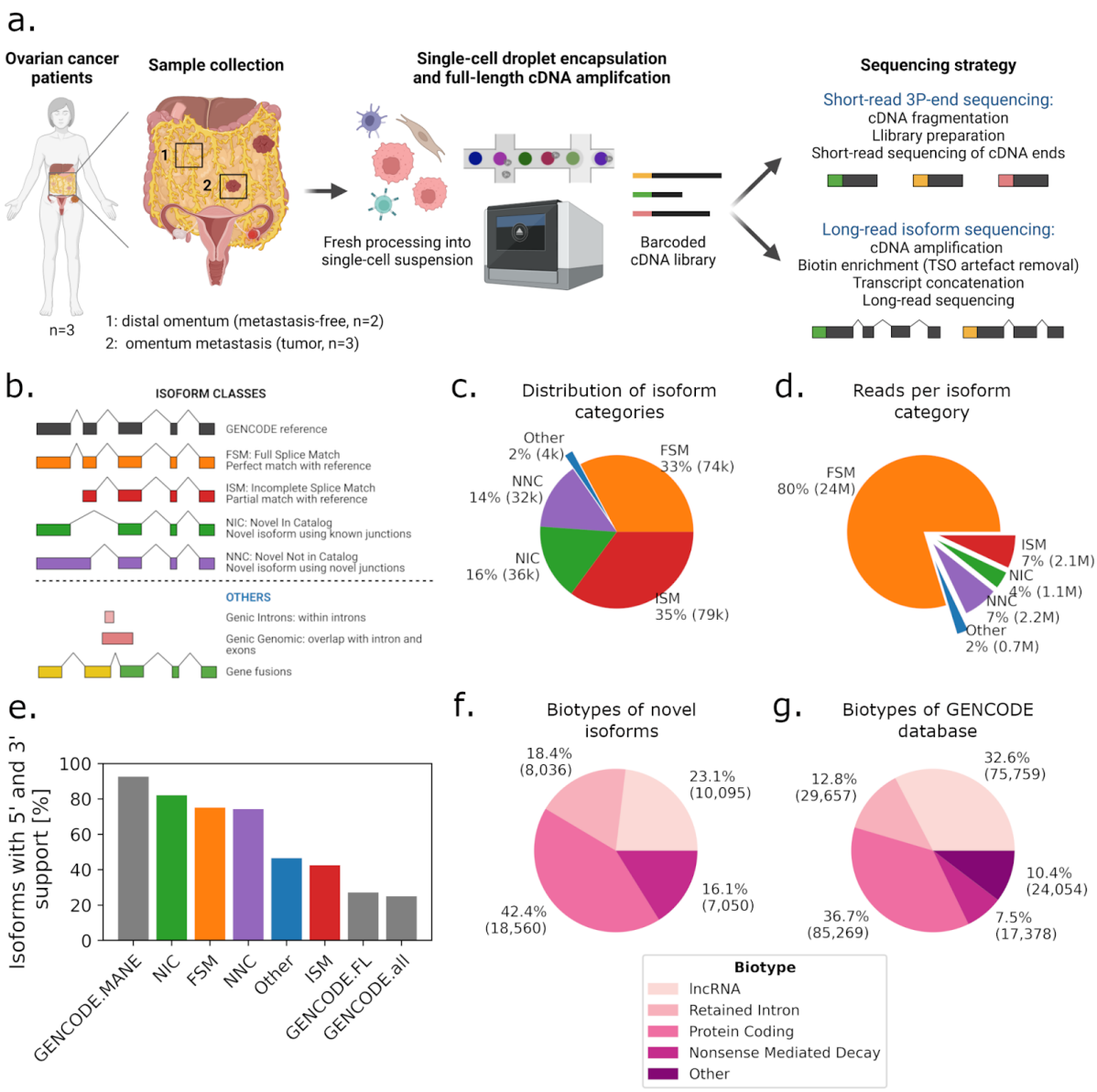

Figure 1: Study design and long-read data overview.
(a) Schematic of freshly processed HGSOC omentum metastases and patient-matched tumor-free distal omentum tissue biopsies, scRNA-seq. **(b)** Definition of SQANTI-defined isoform structural categories. **(c)** Proportions of isoform structural categories detected in merged metastasis and distal 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 poly-adenylation (PolyASite) data, per isoform structural categories. 'GENCODE.all' indicates all protein-coding isoforms in the GENCODE database, 'GENCODE.FL' is a subset of 'GENCODE.all' containing only isoforms tagged as full-length, and 'GENCODE.MANE' is a hand-curated subset of canonical transcripts, one per human protein-coding locus. **(f)** GENCODE-defined biotype composition of novel isoforms. **(g)** Biotype composition of the GENCODE database.

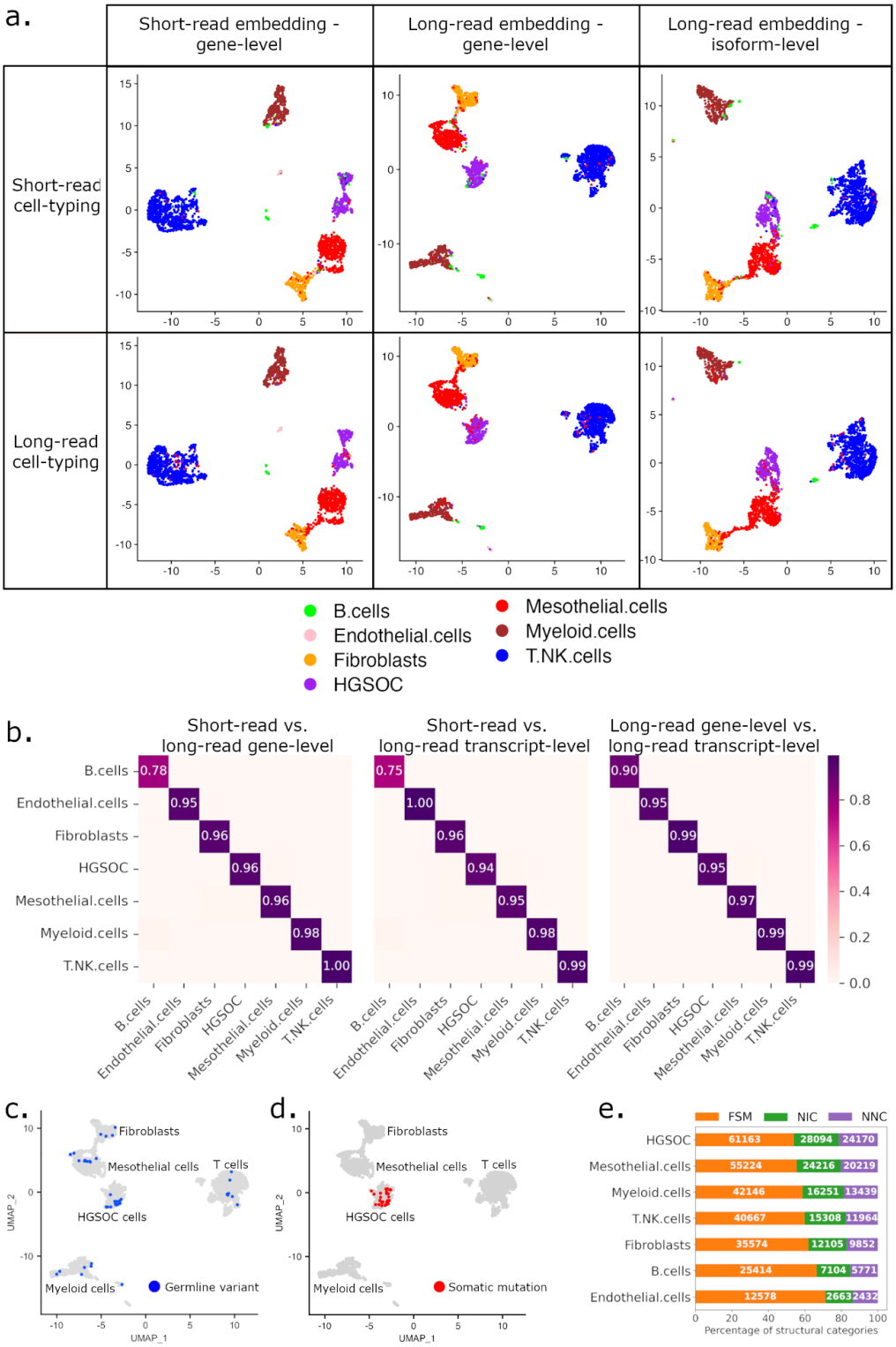

Figure 2: Clustering and cell type-specific isoform distribution.

(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). Long-reads - gene level UMAP cohort visualizations of cells with at least one germline **(c)** or somatic **(d)** mutation also found in targeted NGS panel data of matched patient samples. 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).

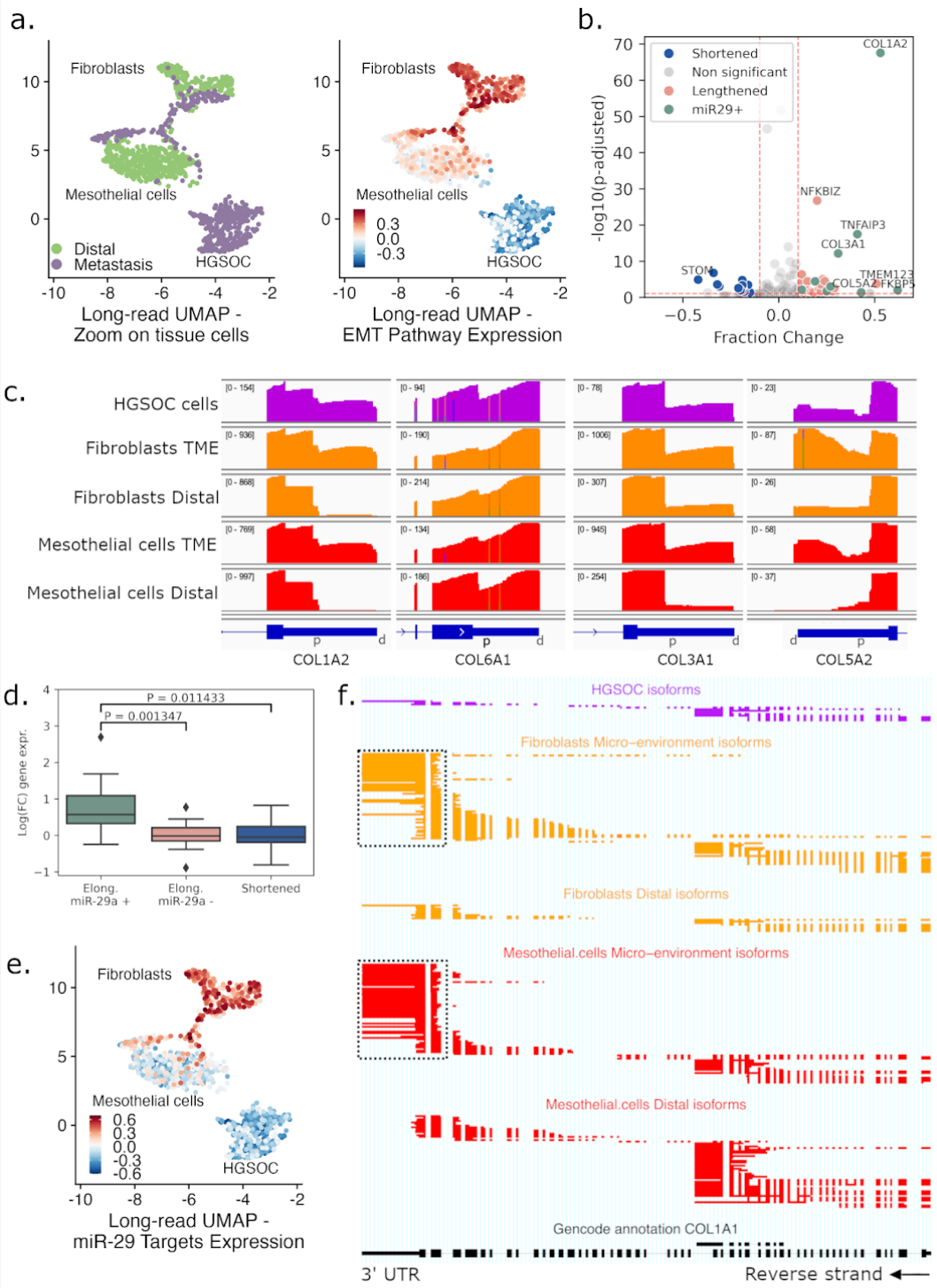

Figure 3: Epithelial-to-mesenchymal transition in the tumor microenvironment. **(a)** Zoom of UMAP embeddings of the cohorts' long-read - gene level data (Fig. 2a, middle column) highlighting tumor and stromal (mesothelial and fibroblast) cells, colored by biopsy tissue type (left) and EMT gene set signal (right). **(b)** Volcano plot of genes with APA in mesothelial cells. Genes have either a lengthened (red) or shortened (blue) 3'UTR in TME compared to distal mesothelial cells. Differentially lengthened or shortened genes targeted by miR-29 are colored in green. Genes with -log10(p-adjusted) > 10 and |Fraction Change| > 0.4 are annotated. **(c)** IGV view of 3'UTR raw coverage of *COL1A2*, *COL6A1*, *COL3A1*, and *COL5A2* in tissue cell types. On the top left between brackets, the coverage

range is displayed throughout each condition. In blue, Ensembl canonical 3'UTR, and for each gene, distal (d) and proximal (p) APA sites are annotated. **(d)** Log fold-change expression between TME and distal mesothelial cells of lengthened genes targeted (+, green) or not targeted (-, red) by miR-29, and shortened genes (blue). **(e)** Cohort UMAP embedding long-read data - gene level, colored by gene set signal of ECM-related genes targeted by miR-29. **(f)** ScisorWiz representation of *COL1A1* isoforms. Colored areas are exons, whitespace areas are intronic space, not drawn to scale, and 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.

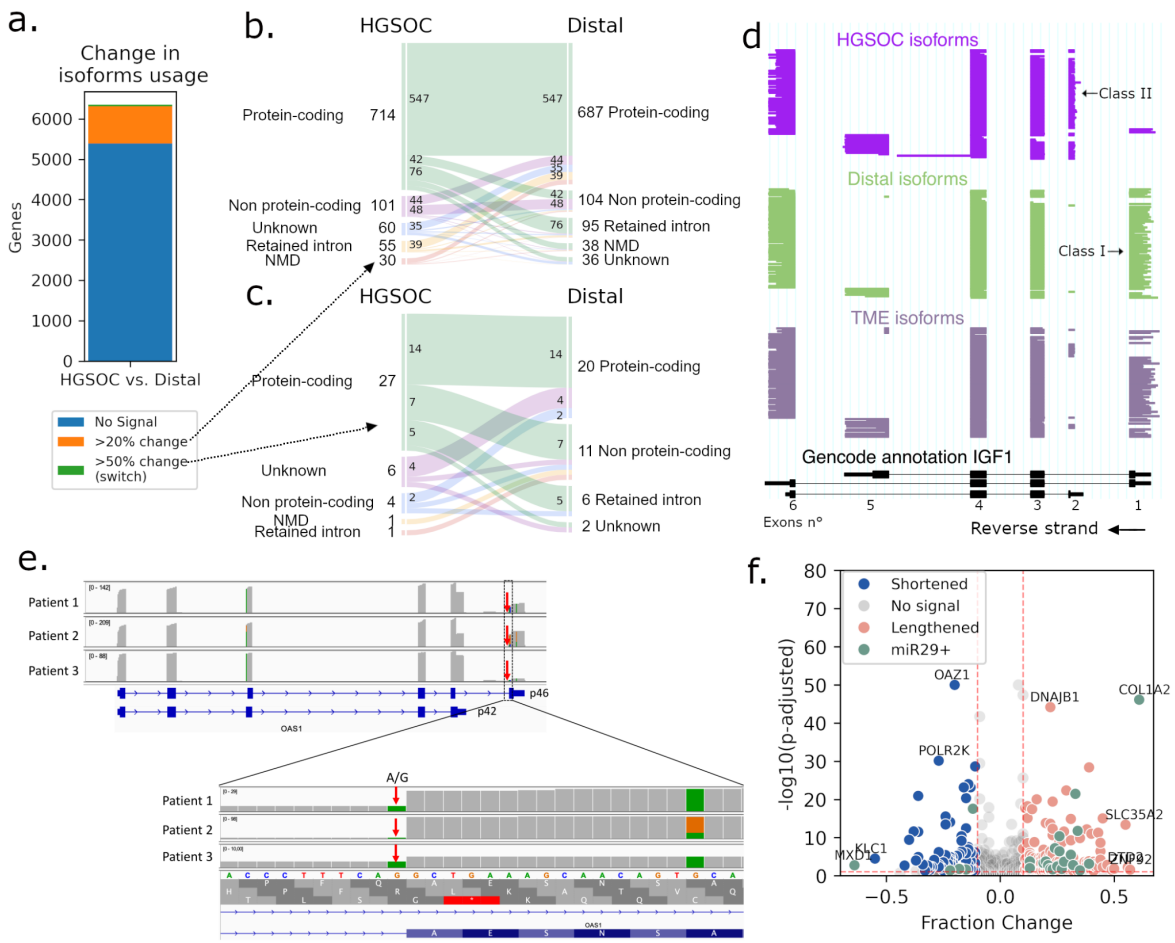

Figure 4: Differential isoforms and 3'UTR lengths in cancer.

(a) Number of genes with change in isoform usage between HGSOC and all distal cells. In orange, genes with differentially expressed isoforms and a change in relative isoform abundance >20% (>50% in green). In blue, genes with no differentially expressed isoforms or change in relative isoform abundance <20%. **(b)** Alluvial plot of biotypes of most expressed isoforms in HGSOC and distal cells in genes containing an isoform change >20% (n=960). Each vein represents the conversion of one biotype to another. For example, in 7 genes, the most expressed isoform in HGSOC cells is protein-coding, while distal cells' one is non-protein-coding. **(c)** Alluvial plot of biotypes of most expressed isoforms in HGSOC and distal cells in genes containing an isoform switch (>50% change, n=39). **(d)** ScisorWiz representation of isoforms in *IGF1*, each horizontal line represents a single isoform colored according to cell types. Exons are numbered according to the Gencode reference, *Class I* and *II* isoforms are isoforms with starting exons 1 and 2, respectively. **(e)** Top: IGV view of *OAS1* expression in patients. Patient 3 has low p46 expression compared to others. Bottom: zoom on the last exon of isoform p46, where all patients have at least one mutated A allele in the splice acceptor site. **(f)** Volcano plot of genes with APA in cancer versus distal cells. Genes have either a lengthened (red) or shortened (blue) 3'UTR in cancer cells compared to all distal cells. Differentially lengthened or shortened genes targeted by miR-29 are colored in green. Genes with -log10(p-adjusted) > 10 and |Fraction Change| > 0.5 are annotated.

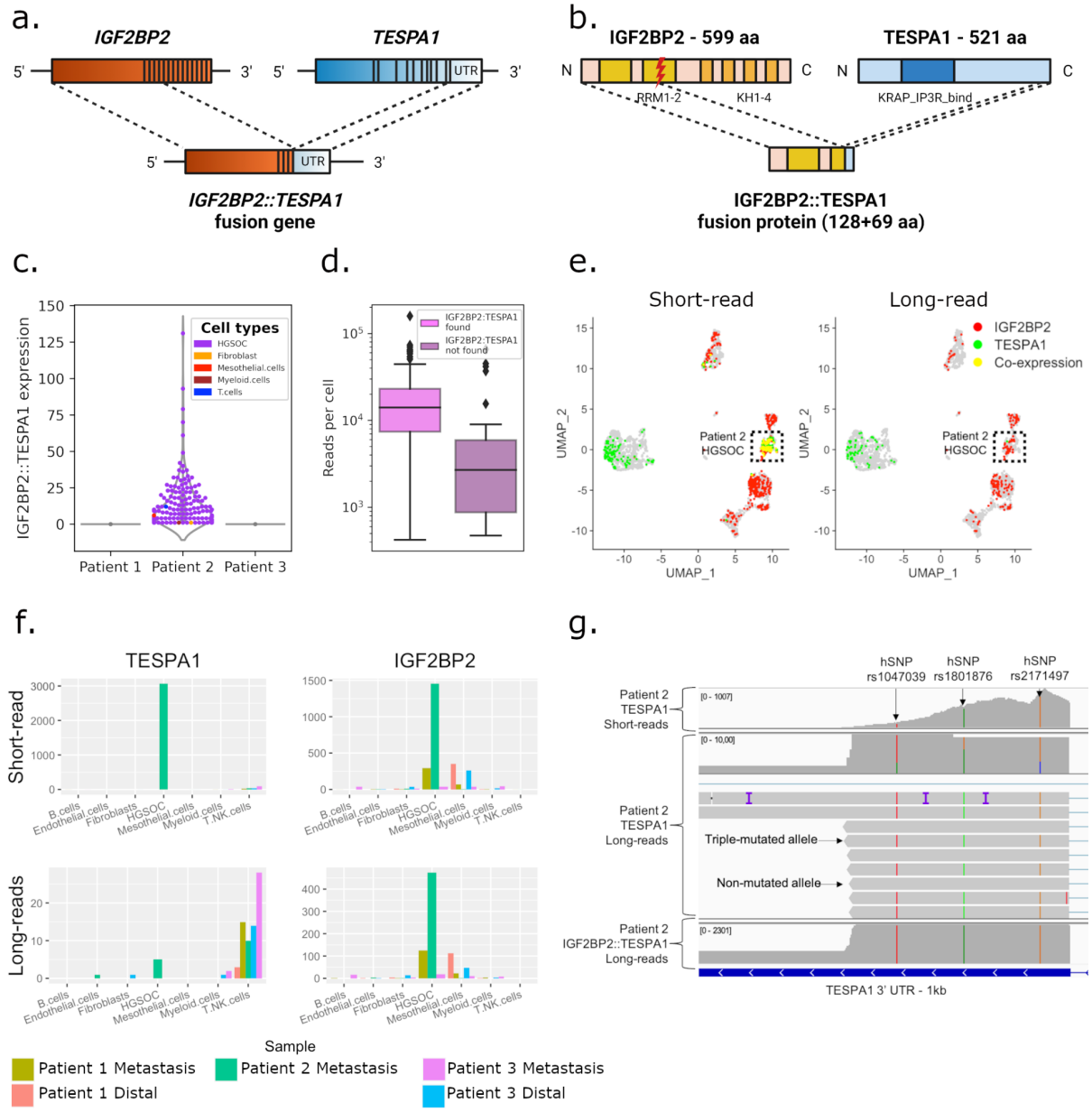

Figure 5: Tumor and patient-specific detection of novel *IGF2BP2::TESPA1* gene fusion. **(a)** Overview of wt *IGF2BP2*, wt *TESPA1,* and *IGF2BP2::TESPA1* gene fusion with exon structure. **(b)** Overview of wt IGF2BP2, wt TESPA1, and fusion protein with protein domains. RRM: RNA-recognition motif, KH: heterogeneous nuclear ribonucleoprotein 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)** UMAP embeddings of the cohorts' short-read data. Cells are colored if they express *IGF2BP2* (red), *TESPA1* (green), or both (yellow) in short- (left panel) or long-reads (right panel). **(f)** Raw expression of *TESPA1* (left) and *IGF2BP2* (right) in short- (top) or long-reads (bottom), by sample and cell type. **(g)** IGV view of short-reads (top), non-fusion long-reads (middle), and fusion long-reads (bottom) mapping to the 3'UTR of *TESPA1*. Non-fusion reads are either triple hSNP-mutated or non-mutated, while fusion and short-reads are only triple hSNP-mutated.

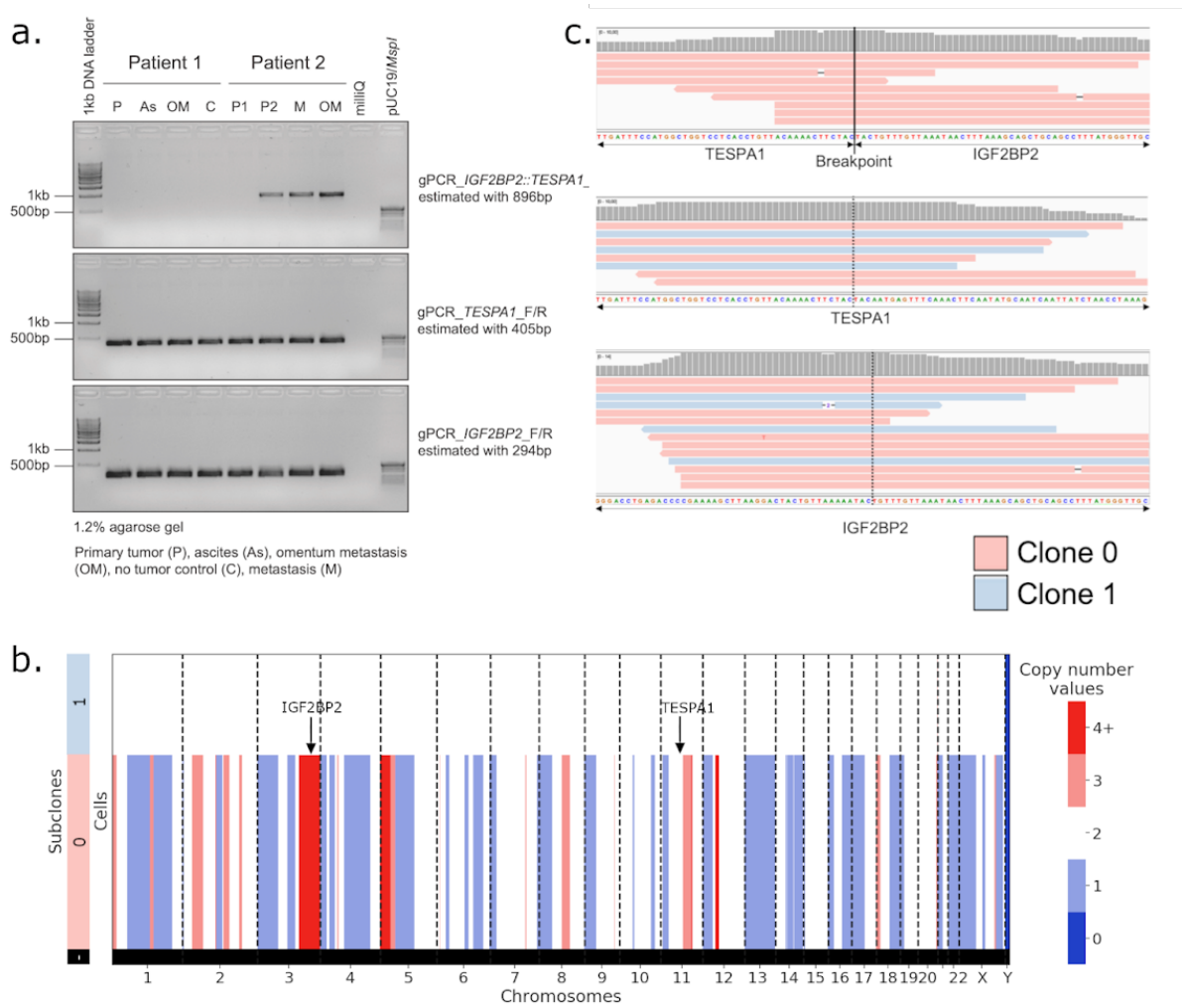

Figure 6: *IGF2BP2::TESPA1* fusion breakpoint validation in bulk and scDNA.
(a) Genotyping PCR on genomic DNA isolated from matched patient samples using gene-specific primers for *IGF2BP2::TESPA1* genomic breakpoint (top), wt *TESPA1* (middle) and wt *IGF2BP2* (bottom). **(b)** Copy number values per subclone in Patient 2 scDNA-seq data. Sublone 0 has multiple copy number alterations, indicative of cancer, while Subclone 1 is copy-number neutral, presumably non-cancer. **(c)** IGV view of scDNA reads aligning unambiguously to the *IGF2BP2::TESPA1* genomic breakpoint (top), wt *TESPA1* (middle), or wt *IGF2BP2* (bottom). In red, reads from Subclone 0 cells (cancer); in blue, reads from Subclone 1 cells (non-cancer).