scientific reports

Check for updates

OPEN

Transcriptome profiling for precision cancer medicine using shallow nanopore cDNA sequencing

Andreas Mock[1,2,4], Melissa Braun[1], Claudia Scholl[3], Stefan Fröhling[1,2] & Cihan Erkut[3]✉

Transcriptome profiling is a mainstay of translational cancer research and is increasingly finding its way into precision oncology. While bulk RNA sequencing (RNA-seq) is widely available, high investment costs and long data return time are limiting factors for clinical applications. We investigated a portable nanopore long-read sequencing device (MinION, Oxford Nanopore Technologies) for transcriptome profiling of tumors. In particular, we investigated the impact of lower coverage than that of larger sequencing devices by comparing shallow nanopore RNA-seq data with short-read RNA-seq data generated using reversible dye terminator technology (Illumina) for ten samples representing four cancer types. Coupled with ShaNTi (Shallow Nanopore sequencing for Transcriptomics), a newly developed data processing pipeline, a turnaround time of five days was achieved. The correlation of normalized gene-level counts between nanopore and Illumina RNA-seq was high for MinION but not for very low-throughput Flongle flow cells (r = 0.89 and r = 0.24, respectively). A cost-saving approach based on multiplexing of four samples per MinION flow cell maintained a high correlation with Illumina data (r = 0.56–0.86). In addition, we compared the utility of nanopore and Illumina RNA-seq data for analysis tools commonly applied in translational oncology: (1) Shallow nanopore and Illumina RNA-seq were equally useful for inferring signaling pathway activities with PROGENy. (2) Highly expressed genes encoding kinases targeted by clinically approved small-molecule inhibitors were reliably identified by shallow nanopore RNA-seq. (3) In tumor microenvironment composition analysis, quanTIseq performed better than CIBERSORT, likely due to higher average expression of the gene set used for deconvolution. (4) Shallow nanopore RNA-seq was successfully applied to detect fusion genes using the JAFFAL pipeline. These findings suggest that shallow nanopore RNA-seq enables rapid and biologically meaningful transcriptome profiling of tumors, and warrants further exploration in precision cancer medicine studies.

Nanopore sequencing is an emerging third-generation DNA and RNA sequencing (RNA-seq) technology. It is based on the phenomenon that a single DNA or RNA molecule in an electrophysiological solution passes through a nanometer-scale protein pore, accompanied by ions in varying concentrations depending on the nucleotide composition. This causes quantifiable patterns of current fluctuations attributable to the nucleotide sequence[1]. Using the digitized current-level information, pretrained artificial neural networks can predict the sequence of very long DNA fragments or full-length transcripts with high accuracy. Because this principle does not require imaging, unlike methods based on reversible dye terminator technology (Illumina), sequencing devices could be substantially downsized[2]. Of particular interest to many investigators is the small, portable MinION sequencer (Oxford Nanopore Technologies), which enables rapid and decentralized sequencing with low investment costs but also has lower throughput than previous methods. In cancer research, MinION-based nanopore sequencing has been successfully employed for mutation detection[3–5], DNA methylome analysis[6–8], DNA copy number profiling[7,9], and the identification of gene fusions[10–12]. In addition, full-length cDNA sequencing has been used to detect aberrant splicing in cancer and to perform differential expression analysis[13,14].

[1]Division of Translational Medical Oncology, National Center for Tumor Diseases (NCT) Heidelberg, German Cancer Research Center (DKFZ), Heidelberg, Germany. [2]German Cancer Consortium (DKTK), Heidelberg, Germany. [3]Division of Applied Functional Genomics, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT) Heidelberg, Heidelberg, Germany. [4]Present address: Institute of Pathology, Ludwig Maximilians University Munich, Munich, Germany. ✉email: cihan.erkut@nct-heidelberg.de

Despite these developments, the full utility of nanopore RNA-seq for transcriptome analysis of human cancers remains elusive. This could be highly relevant for the clinical implementation of precision oncology approaches. Compared to DNA-based stratification approaches alone, the detection of aberrantly expressed genes has the potential to substantially increase the proportion of patients whose care can be individualized based on molecular information[15]. For example, mRNA expression analysis of the receptor tyrosine kinase genes *FGFR1-3* identified a larger patient population eligible for treatment with the pan-FGFR inhibitor rogaratinib[16]. More broadly, the increasing understanding of the target spectrum of clinically available kinase inhibitors[17] allows the predictive value of kinase expression for response to these drugs to be systematically studied[18–20]. In addition, inferring pathway activities from RNA-seq data is becoming increasingly important as they have been shown to outperform the expression of single genes as biomarkers[21,22]. However, the functional taxonomy of tumor ecosystems based on their transcriptomes is not limited to the cancer cell compartment. For example, deconvolution algorithms are widely applied to estimate the composition of the immune cell microenvironment[23–25].

MinION-based nanopore RNA-seq can be considered "shallow" RNA-seq since its yield is considerably lower than that of standard Illumina RNA-seq. However, predictive in silico modeling suggests that biomarkers for precision cancer medicine can be developed with drastically reduced sequencing depth[26]. We investigated the feasibility of using a MinION sequencer for rapid transcriptome profiling of human tumors. In addition to a comparison with matching Illumina RNA-seq data, we explored four applications for precision cancer medicine, i.e., (1) pathway activity inference, (2) expression quantification of kinases targeted by approved inhibitors, (3) immune cell deconvolution, and (4) fusion gene detection. Our results show that shallow nanopore RNA-seq enables biologically meaningful transcriptome profiling of tumors and warrants further development as a stratification tool in the clinic.

Results

Estimation of gene expression. To benchmark shallow nanopore RNA-seq for tumor transcriptome profiling, we analyzed tissue samples from ten patients for whom Illumina RNA-seq data had been generated within the MASTER precision oncology program[27] (Fig. 1a). The workflow from extracted RNA to processed and normalized RNA-seq data took five days for each sequencing run. We automated the standard data preproc-

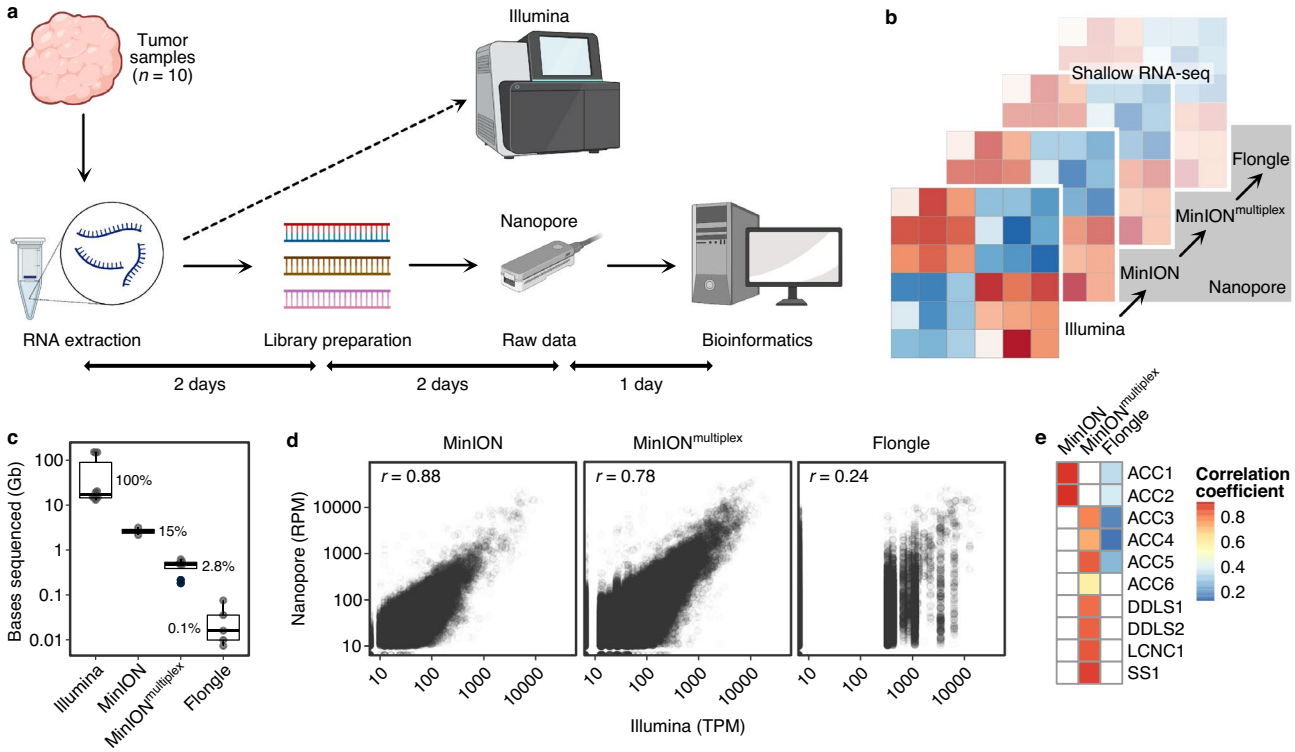

Figure 1. Comparison of shallow nanopore and Illumina RNA-seq. (**a**) Overview of sample and data processing for nanopore RNA-seq. Created with BioRender.com. (**b**) Schematic of the different sequencing depths used to perform nanopore RNA-seq (MinION flow cell with one sample, MinION flow cell with four samples [MinION^{multiplex}], Flongle flow cell with one sample). For illustration, transcriptomic profiling is depicted by a heatmap with increasing transparency representing decreasing sequencing depth. (**c**) Comparison of sequencing depth, i.e., bases sequenced, between technologies. Nanopore RNA-seq covered a median of 0.1–15% of the bases sequenced with Illumina RNA-seq. (**d**) Correlation of normalized gene-level counts obtained by nanopore (RPM values) and Illumina (TPM values) RNA-seq across all samples and protein-coding genes (Spearman rank correlation coefficient). Each dot represents a gene. (**e**) Correlation of normalized gene-level counts obtained by nanopore and Illumina RNA-seq for individual samples (Spearman rank correlation coefficient).

essing steps, i.e., basecalling, read filtering, alignment to the reference genome, and read counting, with a custom bioinformatics pipeline, which we termed ShaNTi (Shallow Nanopore sequencing for Transcriptomics) (Figure S1). In addition to processed data, ShaNTi generates detailed statistics of sequencing depth per run, mean read length per run, percent basecall identity per run, number of aligned reads per sample, alignment accuracy per sample, etc. (Table S3). Read length distributions are shown in Figure S2 for both MinION and Flongle runs.

Before we thoroughly compared the two technologies, we performed a series of technical benchmarks. First, we discarded forward-oriented direct cDNA reads. Direct cDNA sequencing has no PCR amplification step. In a theoretically perfect case where all mRNA molecules are reverse transcribed into double-stranded cDNA, each cDNA strand will be sequenced only once, and we should obtain twice as many reads as target mRNA molecules. The read originating from the first cDNA strand (i.e., the strand synthesized using the mRNA as the template) should align to the reference genome in the opposite direction (antiparallel) as the gene model because it is the reverse complement of the coding sequence. Such alignments are called reverse-oriented. The second cDNA strand should then align in the same direction (parallel) as the gene model. Such alignments are called forward-oriented. However, direct cDNA sequencing deviates from this theoretically perfect case. Empirically, we observed that reverse-oriented alignments were twice as many as forward-oriented alignments (Figure S3a). This is mainly due to known reverse transcription artifacts, which include the full-length first strand fused to the second strand through a hairpin loop[28]. When cDNA molecules are denatured, they result in a single strand almost twice as long as the template mRNA molecule. The first half of the read originating from such a molecule maps to the reference genome as a reverse-oriented primary alignment, whereas the rest of the read registers as a forward-oriented supplementary alignment. These artifacts disrupt the balance between strand-specific read counts. However, by considering only reverse-oriented primary reads, we count each mRNA template at most once and prevent overestimation of gene expression levels at the expense of losing sequencing depth.

Another reason we exclude forward-oriented reads is overlapping gene pairs in the genome (Figure S3b). These are encoded on opposite strands of the DNA and therefore have opposite directions. However, they share 5' and 3' untranslated regions (UTRs). Since long-read sequencing can capture UTRs, these regions also map to the reference genome. However, read summarization software may not always accurately assign these reads to the correct gene. In such cases, we obtain a similar number of reads for both genes, but in opposite orientations. By counting only reverse-oriented reads, we can quantify the expression levels of genes in such paired configurations more accurately.

Second, we decided to use reads per million (RPM) and transcripts per million (TPM) normalizations to estimate gene expression levels based on nanopore and Illumina data, respectively. The expression level of a gene is estimated based on the total number of reads aligned to its exonic regions in the reference genome. The total length of all exons of a gene approximates its effective transcript length. The total read counts of all genes in a sample define its library size. While both RPM and TPM normalize read counts to library size, the TPM metric also normalizes them to effective transcript length.

In short-read sequencing, reads are often shorter than transcripts. Therefore, multiple reads tandemly align to the gene locus in the reference genome. This creates a bias in gene expression levels measured as RPM because shorter genes appear to be expressed less (Figure S4a). The TPM metric normalizes read counts to total exon length to alleviate this problem to some extent (Figure S4b). Long-read sequencing, however, is fairly unbiased in this respect because a single long read often spans the majority of exons (Figure S4c). TPM normalization of long reads, however, introduces a monotone inverse relation between gene expression levels and gene length (Figure S4d).

Comparison of shallow nanopore and Illumina RNA-seq for tumor transcriptome profiling. We examined different sequencing depths (Fig. 1b) and observed that running a single sample on a MinION or Flongle flow cell yielded a median of 15% or 0.1% of the bases sequenced, respectively, compared to Illumina RNA-seq (Fig. 1c). Multiplexing four samples on a MinION flow cell yielded a median of 2.8% of the bases sequenced compared to Illumina RNA-seq (Fig. 1c). The correlation between normalized gene-level counts determined by nanopore or Illumina RNA-seq across samples was high for MinION flow cells (single sample, $r = 0.88$; four multiplexed samples, $r = 0.78$) (Fig. 1d, Table S4–S7) but low when we used a Flongle flow cell ($r = 0.24$). The correlation between MinION-based nanopore and Illumina RNA-seq was also high for the majority of individual samples (single samples, $r = 0.88$–0.89; multiplexed samples $r = 0.56$–0.86) (Fig. 1e). As expected, samples analyzed by nanopore or Illumina RNA-seq formed two distinct clusters in an unsupervised comparison (Figure S5). In summary, these data showed that the transcriptomic measurements obtained by shallow nanopore and Illumina RNA-seq were consistent with each other.

Inference of signaling pathway activities and kinase expression analysis. An emerging application of RNA-seq in precision oncology is the inference of signaling pathway activities from tumor transcriptional profiles. Aberrantly-activated signaling cascades provide insight into upstream mutant drivers and associated molecular dependencies that can be exploited therapeutically. To test the use of nanopore RNA-seq data in such an analysis, we applied the PROGENy method[29]. In the original PROGENy framework, 100 so-called footprint genes per pathway were used to compute the activity of a pathway. To simulate the impact of shallow nanopore RNA-seq on the performance of PROGENy, we compared the accuracy of determining an active pathway based on nanopore or Illumina RNA-seq with different coverages. Using the default number of 100 footprint genes per pathway resulted in a drop in nanopore performance relative to Illumina with decreasing coverage (area under the receiver operating characteristic curve [AUROC], 0.74–0.63; Fig. 2a). The same pattern was observed for 200, 300, and 500 footprint genes per pathway. This dependence of PROGENy performance on coverage was no longer present when the number of footprint genes per pathway was set to 1000 or more (Fig. 2a).

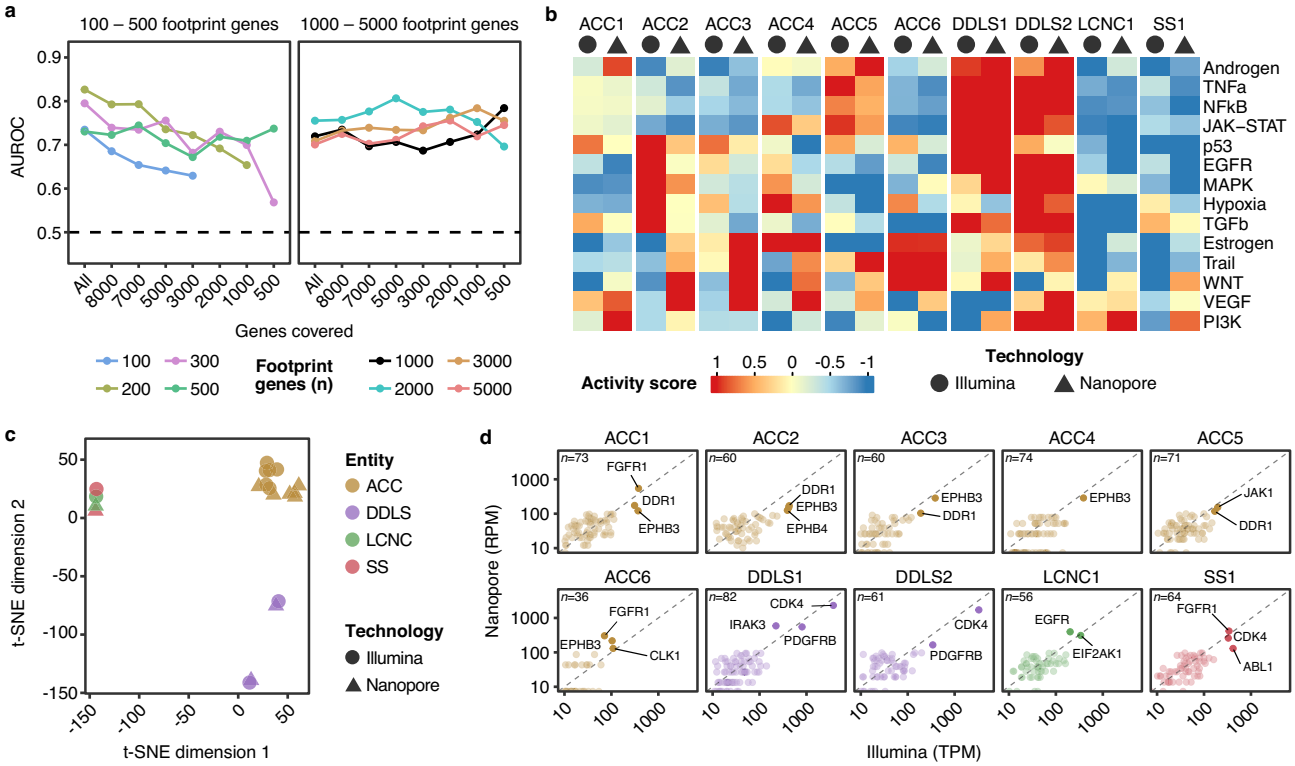

Figure 2. Analysis of pathway activities and kinase expression. (**a**) Performance of PROGENy on nanopore RNA-seq data using different numbers of footprint genes per pathway according to gene coverage. (**b**) Activities of 14 signaling pathways (rows) in ten tumor samples (columns) determined by nanopore (triangle) and Illumina (circle) RNA-seq. (**c**) Clustering of samples based on pathway activation scores shown in panel b after dimensionality reduction with t-distributed stochastic neighbor embedding (t-SNE). (**d**) Correlation of expression of kinases in the target spectrum of clinically approved inhibitors determined by nanopore and Illumina RNA-seq. For each sample, the most highly expressed kinases are labeled.

Running PROGENy with these optimized parameters, i.e., 1,000 footprint genes, resulted in an agreement between pathway activity inferences based on nanopore and Illumina RNA-seq data (AUROC, 0.72; Fig. 2b). The two dedifferentiated liposarcoma (DDLS) samples showed activity of more pathways than the adenoid cystic carcinoma (ACC), large-cell neuroendocrine carcinoma (LCNC), and synovial sarcoma (SS) samples. The results were consistent with published results on recurrent mutations or aberrant gene expression driving activation of specific pathways in the respective entities, e.g., "PI3K" in LCNC[30,31], "Androgen" in ACC[32,33], and "NFkB" in DDLS[34]. Unsupervised dimensionality reduction of pathway activation scores showed high similarity of ACC transcriptomes with minor differences attributable to sequencing technology (Fig. 2c), recapitulating the concordance of pathway activity inferences based on nanopore and Illumina RNA-seq data. The other entities clustered separately but independent of sequencing technology, indicating that ACC, DDLS, LCNC, and SS are characterized by distinct transcriptional networks.

In addition to studying entire signaling pathways, we also investigated whether nanopore RNA-seq can accurately detect highly-expressed individual kinase genes, which could be potential oncogenic drivers and exploited for kinase-targeted therapies. We, therefore, performed comparative expression analyses of 144 kinase genes that fall within the target spectrum of the 33 clinical kinase inhibitors approved by the United States Food and Drug Administration[17]. We found that in all samples, the kinase genes with the highest expression levels could be reliably identified by nanopore RNA-seq (Fig. 2d). In ACC samples, *FGFR1*, *EPHB3*, and *DDR1* were recurrently expressed at high levels (Fig. 2d). FGFR family members are commonly upregulated in ACC, and treatment with FGFR inhibitors has shown clinical efficacy in this disease[35–37]. The two DDLS samples were characterized by high *CDK4* and *PDGFRB* expression (Fig. 2d). *CDK4* is amplified in more than 90% of DDLS cases, and clinical trials have shown that the CDK4 inhibitors palbociclib and abemaciclib had a favorable effect on progression-free survival[38,39]. Together, these results suggest that nanopore RNA-seq can be used to reliably determine signaling pathway activity and kinase gene expression.

Immune cell deconvolution. Understanding the tumor microenvironment is becoming increasingly important in precision cancer medicine. For example, the prediction of response to immune checkpoint inhibitors is aided by knowledge of immune cell fractions, whose abundance can be estimated by deconvolution algorithms from bulk RNA-seq data. We employed two of the most widely used algorithms, CIBERSORT[24] and quanTIseq[23]. In CIBERSORT, deconvolution is based on 547 genes covering the profiles of 22 immune cell types. In comparison, quanTIseq uses 170 genes to detect ten immune cell types. To assess the applicability of the two

methods to shallow RNA-seq data, we first compared the average expression of CIBERSORT and quanTIseq gene sets in the Illumina and nanopore RNA-seq data (Fig. 3a). In both Illumina and nanopore RNA-seq data, the CIBERSORT reference gene set was significantly lower expressed compared to all protein-coding genes (Illumina: $t = 10.2$, $df = 594{,}95$, $p < 2.2e{-}16$; nanopore: $t = 11.572$, $df = 590.01$, $p < 2.2e{-}16$), whereas no difference was observed for the quanTIseq gene set (Illumina: $t = 0.23$, $df = 168.37$, $p = 0.82$; nanopore: $t = 0.27$, $df = 168.15$, $p = 0.79$). Accordingly, the fraction of genes with zero counts in more than 50% of samples was higher in the CIBERSORT than in the quanTIseq gene set (Fig. 3b; Illumina, 5.5% vs. 1.8%; nanopore, 64.7% vs. 44.8%). In contrast to Illumina RNA-seq, shallow nanopore RNA-seq cannot detect transcripts of very low expressed genes. Using such genes for immune cell deconvolution based on shallow RNA-seq data may therefore lead to incorrect estimates. Indeed, the average abundance values based on Illumina and shallow nanopore RNA-seq data were considerably more similar when using quanTIseq instead of CIBERSORT (Figs. 3c,d, S6).

We, therefore, used quanTIseq for subsequent analyses. In the six ACC cases, both sequencing technologies identified immune-excluded microenvironments, as indicated by the low abundance of all immune cell types, consistent with a recent study[40] (Fig. 3d). The LCNC case displayed a high frequency of CD4 + T cells, again based on both Illumina and shallow nanopore RNA-seq data, which has been associated with unfavorable recurrence-free survival in LCNC of the lung[41]. In conclusion, these data suggest that quanTIseq can most likely be used to deconvolute immune cell fractions from shallow nanopore RNA-seq data; however, larger sample size is required to confirm this assumption.

Gene fusion detection. Many cancers are associated with recurrent gene fusions that drive tumorigenesis and may also be diagnostic biomarkers, such as *SS18::SSX* in synovial sarcoma[42]. Predicting and visualizing gene fusions based on short-read transcriptome data is challenging because split reads spanning fusion breakpoints carry limited information about the corresponding genomic coordinates. Since the information content of long reads is much higher in this regard, technologies such as nanopore RNA-seq are better suited for these tasks[43]. JAFFAL is a tool that uses long-read transcriptome sequencing data to accurately detect fusion genes[11]. We applied the JAFFAL pipeline to all samples independently and combined the results (Table S8). Only in sample ACC4, the pipeline failed due to an error. We then visualized the fusion events classified by JAFFAL as "high confidence" (Fig. 4). In all ACC samples, the disease-defining *MYB::NFIB* fusion was detected, even with only a few reads[44–46]. We also observed various intrachromosomal rearrangements affecting chromosomes 5 and 6 in ACC samples. Similarly, the *SS18::SSX2* fusion characteristic of SS could also be detected in sample SS1 in

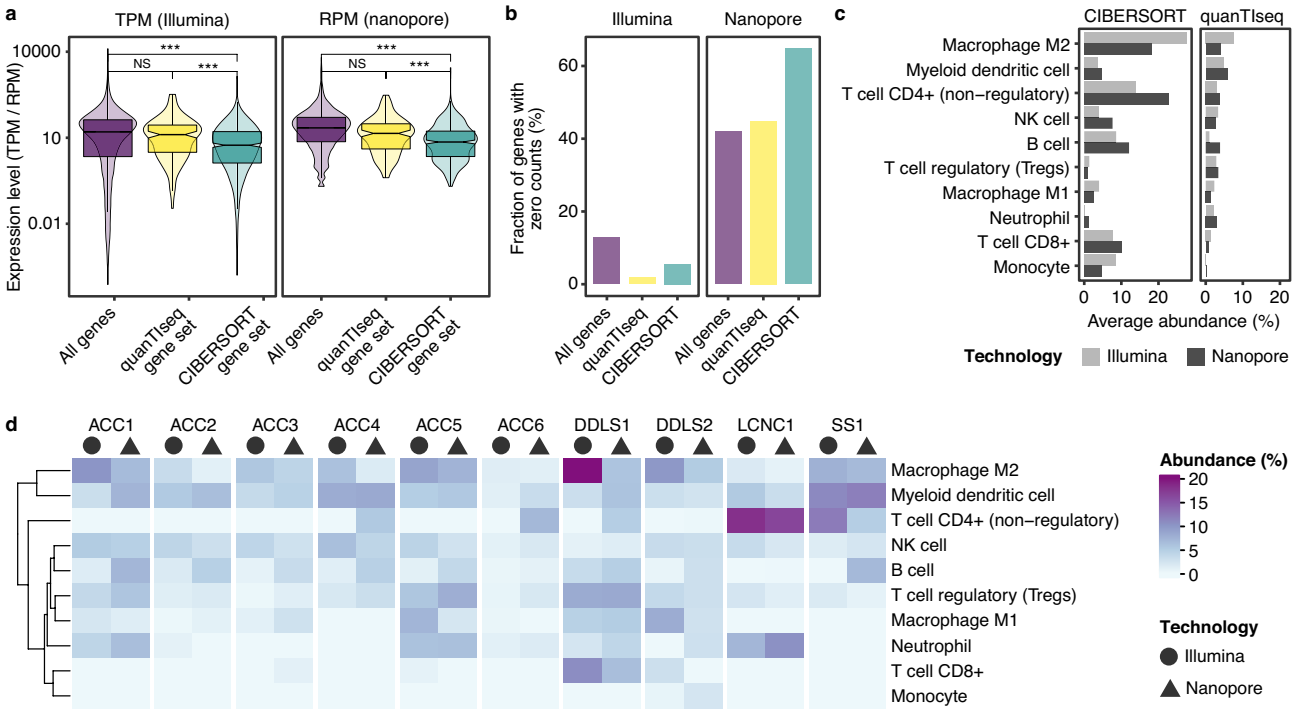

Figure 3. Estimation of immune cell abundance from shallow nanopore and Illumina RNA-seq data. (**a**) Expression levels, determined by Illumina (as TPM) and nanopore RNA-seq (as RPM), of all protein-coding genes and gene sets used in quanTIseq (n = 170) and CIBERSORT (n = 547). Groups were compared using two-sided Welch's t-test. ***, $p < 0.001$; NS, not significant. (**b**) Number of genes with zero counts in more than 50% of samples, as determined by Illumina and nanopore RNA-seq, of all protein-coding genes and gene sets used in quanTIseq and CIBERSORT. (**c**) Estimated average abundance of immune cell fractions based on Illumina and shallow nanopore RNA-seq data as a function of the deconvolution method used. (**d**) Immune cell abundance of the ten tumor samples calculated with quanTIseq based on Illumina and nanopore RNA-seq data.

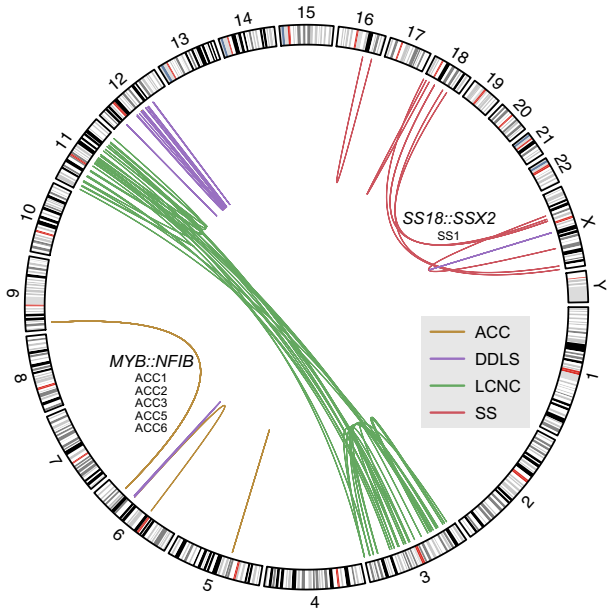

Figure 4. Visualization of structural variations. The segments in the circle represent the cytogenetic ideograms of chromosomes, oriented clockwise from start to end position and numbered consecutively. Intrachromosomal and interchromosomal rearrangements detected by the JAFFAL pipeline are represented by curves connecting the estimated breakpoints. In ACC samples (brown), the rearrangement between chromosomes 6 and 9 giving rise to the disease-defining *MYB::NFIB* fusion was observed. In sample SS1 (red), the characteristic *SS18::SSX2* fusion was observed as a result of a rearrangement between chromosomes 18 and X. Several other rearrangements accompanied these expected ones. In addition, many rearrangements were observed in samples LCNC1 (green) and DDLS (purple).

addition to several other interchromosomal (between chromosomes 18 and X) and intrachromosomal (affecting chromosomes 16, 18, and X) rearrangements. In samples DDLS1 and DDLS2, we mainly observed intrachromosomal rearrangements affecting chromosomes 6 and 12. Finally, in LCNC1, we found a large number of rearrangements between chromosomes 3 and 11 in addition to rearrangements within either of these chromosomes.

Structural variations in the genomes of these samples had already been detected by whole-genome sequencing in the MASTER program, and the exact locations of the genomic breakpoints were calculated. When we visualized the alignments to the *MYB* and *NFIB* loci in ACC samples and to *SSX2* and *SS18* in SS1, we observed reads that did not align beyond the breakpoint (Figures S7–S13). This observation supports the results of JAFFAL pipeline.

Together, results indicate that the shallow RNA-seq data generated by our workflow can be used to identify gene fusions.

Discussion

The low cost of setting up and running a MinION sequencer democratizes the use of long-read sequencing in academic laboratories. To assess the performance of this new technology in translational oncology, we generated shallow nanopore RNA-seq data using tumor RNA previously subjected to Illumina RNA-seq within a precision oncology program. Our comparative analysis of long-read nanopore and short-read Illumina RNA-seq data demonstrates the feasibility of implementing shallow RNA-seq using the MinION sequencer for transcriptome profiling of human cancers. To address the challenge of a dedicated, easy-to-use bioinformatics workflow for this application, we developed ShaNTi, an automated pipeline that provides a turnaround time from extracted RNA to processed and normalized expression data of only five days. The data preprocessing workflow starts with raw current-level data and creates alignments, transcript quantifications, and quality controls. ShaNTi is publicly available and uses exclusively open-source software and can thus be easily adapted to any workstation equipped with the necessary hardware.

To determine the optimum sequencing depth required to achieve a high correlation with Illumina RNA-seq data, we considered different flow cells and multiplexing of samples. Specifically, the comparison was made for a single sample per MinION flow cell, multiplexing four samples on a MinION flow cell, and a single sample on the smaller and lower-throughput Flongle flow cell. The average sequencing depth upon multiplexing four samples on a MinION flow cell, which was 36-fold less than Illumina (i.e., 2.8% of the base yield of Illumina), was already sufficient for the downstream applications that we performed. This result is consistent with a comprehensive computational simulation of RNA-seq data generated within The Cancer Genome Atlas initiative, which showed that a 10- to 100-fold reduction in sequencing depth resulted in no change in the performance of predicting patient outcomes based on transcriptome profiling[26].

Following this technical benchmarking, we pursued a first translational application and derived signaling pathway activities from nanopore RNA-seq data and compared them with the corresponding Illumina RNA-seq results. To this end, we applied PROGENy, an emerging methodology used to better understand oncogenic signaling and, of particular relevance to precision oncology, to relate pathway activities to drug responses[29,47,48]. Based on single-cell RNA-seq data, it has been shown that the number of footprint genes used to infer the activity of a given pathway needs to be tuned[49]. We found that such parameter tuning is also crucial for our shallow RNA-seq approach. We observed that when we used a smaller number of footpring genes per pathway (100–500), the performance of PROGENy dropped as the number of covered genes decreased. However, it remained stable when we included 1,000 genes or more. This suggests that at least 1,000 footprint genes should be used to reliably infer pathway activity based on shallow RNA-seq data. Using this optimized parameter, we observed high similarity between the pathway activities predicted by nanopore and Illumina RNA-seq. These results are consistent with previous data on the activity of specific pathways in the entities studied, e.g., androgen signaling in ACC[32,33]. However, because our study is the first application of PROGENy in ACC, LCNC, DDLS, and SS, a more comprehensive comparison is not possible.

In addition to pathway activity, we determined in each sample the most highly expressed kinase genes whose protein products are within the target range of clinically approved inhibitors. We observed that findings based on Illumina RNA-seq were confirmed in all cases by nanopore RNA-seq. For example, FGFR family members were commonly overexpressed in ACC, consistent with the clinical efficacy of FGFR blockade in this entity[35–37]. Similarly, the DDLS sample showed *CDK4* overexpression, which, due to amplification of the *CDK4* locus, occurs in more than 90% of DDLS cases and provides a rationale for using CDK4 inhibitors, which have been associated with prolonged progression-free survival in clinical trials[38,39].

Next, we explored the application of nanopore RNA-seq to predict the tumor microenvironment composition using the CIBERSORT and quanTIseq algorithms. Because the shallow RNA-seq approach cannot capture sparsely-expressed genes, we first investigated the average expression of the gene sets used in CIBERSORT and quanTIseq in our Illumina RNA-seq data. We observed that the reference genes used in CIBERSORT were significantly less expressed and more often undetectable than those used in quanTIseq. This is a potential reason why the concordance between the immune cell estimates derived from Illumina and nanopore RNA-seq data was inferior when using CIBERSORT. In line with a recent study[40], the six ACC cases displayed an immune-excluded microenvironment, with an overall low fraction of immune cells. The high proportion of CD4 + T cells in the LCNC case may be an unfavorable prognostic factor for recurrence-free survival, as previously described[41]. Further studies correlating immune cell fractions estimated by quanTIseq with immunohistochemical staining of the same tumor tissue are needed to validate these findings.

In a third application of diagnostic and therapeutic relevance, we demonstrated that shallow nanopore RNA-seq data can be used to detect gene fusions using tools such as JAFFAL[11]. In the interest of rapid turnaround, which can be critical in certain clinical situations, automatic detection of gene fusions is desirable.

In summary, we showed that shallow nanopore RNA-seq might enable biologically meaningful transcriptome profiling of human cancers and thus has the potential to complement short-read-based sequencing workflows, especially in applications where rapid processing is required. Our next steps will be to develop our workflow further and, in particular, increase the cohort size via testing it prospectively alongside the diagnostic workup of the MASTER precision oncology trial.

Methods

Patient samples. Tumor samples from ten patients with rare cancers, i.e., adenoid cystic carcinoma (ACC), dedifferentiated liposarcoma (DDLS), large-cell neuroendocrine carcinoma (LCNC), and synovial sarcoma (SS), were studied. All patients provided written informed consent for banking of tumor tissue, molecular analysis and the collection of clinical data under a protocol (S-206/2011) approved by the Ethics Committee of the Medical Faculty of Heidelberg University. This study was conducted in accordance with the Declaration of Helsinki. All samples were subjected to quality control, verification of the respective entity, and estimation of tumor cell content by experienced pathologists. RNA was extracted in the Sample Processing Laboratory of the German Cancer Research Center (DKFZ). Illumina RNA-seq had been performed within the MASTER (Molecularly Aided Stratification for Tumor Eradication Research) trial of the National Center for Tumor Diseases (NCT), DKFZ, and the German Cancer Consortium (DKTK)[27]. Library preparation for nanopore RNA-seq was performed with the same analyte previously used for Illumina RNA-seq. Samples were selected based on RNA availability and known entity-defining gene fusions (Table 1).

Illumina RNA-seq data processing. Reads were processed with the RNA-seq workflow 1.3.0 developed by the DKFZ Omics IT and Data Management Core Facility (https://github.com/DKFZ-ODCF/RNAseqWorkflow). First, FASTQ reads were aligned via two-pass alignment using STAR 2.5.3a[50]. The STAR index was generated from the 1000 Genomes assembly and GENCODE Version 19 gene models with a sjdbOverhang of 200. Alignment call parameters are listed in Table S1. Other parameters were set as default or only pertinent for particular samples. Duplicate marking of the resultant main alignment file was done with sambamba 0.6.5[51] using eight threads. The chimeric file was sorted using samtools 1.6[52], and duplicates were marked using sambamba. BAM indexes were also generated using sambamba. Quality control was performed using samtools flagstat and the rnaseqc tool version 1.1.8[53] with the 1000 Genomes assembly and GENCODE Version 19 gene models. Depth-of-coverage analysis for rnaseqc was turned off. Gene-specific read counting was performed using featureCounts (from Subread 1.5.1)[54] over exon features based on GENCODE Version 19 gene models. Both reads of a paired fragment were used for counting, and the quality threshold was set to 255, indicating that STAR found a unique alignment. Strand-specific counting was also used. For RPKM and TPM calculations, all genes

Sample	Illumina	MinION-single	MinION-multiplex	Flongle	Fusion
ACC1	+	+		+	*MYB::NFIB*
ACC2	+	+		+	*MYB::NFIB*
ACC3	+		+	+	*MYB::NFIB*
ACC4	+		+	+	*MYB::NFIB*
ACC5	+		+	+	*MYB::NFIB*
ACC6	+		+		*MYB::NFIB*
DDLS1	+		+		
LCNC1	+		+		
DDLS2	+		+		
SS1	+		+		*SS18::SSX2*

Table 1. Tumor samples. Listed are the availability of RNA-seq data (indicated by "+") for the Illumina and various Oxford Nanopore platforms and characteristic gene fusions identified by Illumina-based sequencing.

on chromosomes X and Y, the mitochondrial genome, as well as rRNA and tRNA genes were omitted as they are likely to introduce library size estimation biases. All computations were performed on a high-performance compute cluster.

Nanopore RNA-seq. Direct cDNA sequencing was performed using the SQK-DCS109 kit (Oxford Nanopore Technologies). For analysis of a single sample on a MinION flow cell (version R9.4.1), 5 µg RNA was used as input. For multiplexing on a MinION flow cell, 2.5 µg total RNA per sample was used as input, and the native barcoding expansion kit EXP-NBD104 was employed in conjunction with SQK-DCS109. After reverse transcription with Maxima H Minus Reverse Transcriptase (Thermo Scientific), second-strand synthesis was performed using the 2 × LongAmp Taq Master Mix (New England Biolabs). The resulting double-stranded cDNA was subjected to end-repair and dA-tailing using the NEBNext Ultra End Repair/dA-Tailing Module (New England Biolabs). For multiplexed libraries, this step was followed by barcode ligation and library pooling. Next, libraries were quantified with a Qubit Fluorometer 3.0 (Life Technologies). Finally, sequencing adapters were added to the library preparations and ligated with Blunt/TA Ligase Master Mix (New England Biolabs), followed by further quality control using a Qubit. Samples ACC1 and ACC2 were analyzed on individual MinION flow cells, while the remaining eight samples were sequenced as multiplexed libraries on two MinION flow cells by pooling four samples for each run. Five ACC samples were also analyzed individually on Flongle flow cells (Table 1). The run time was between 72 and 96 h, depending on library and flow cell quality.

Nanopore RNA-seq data processing. We developed a custom pipeline for processing nanopore RNA-seq data, available at https://github.com/cihanerkut/shanti. The workflow includes basecalling, read filtering, demultiplexing (optionally), alignment, and read summarization (Figure S1). Custom parameters for alignment are listed in Table S2. All other parameters were either kept as default or adjusted to the respective sample, e.g., according to the sequencing and barcoding kits used. All computations were performed on a local workstation with 32 threads, 256 GB RAM, and an NVIDIA Tesla V100 16 GB GPU. First, GPU basecalling was performed from FAST5 files with Guppy basecaller 5.0.14 using the super-accuracy model. Adapter trimming was turned off during basecalling of multiplexed data. MinION reads with an average Phred score of less than 7 were filtered out with NanoFilt 2.6.0[55], as such reads cannot be used for accurate demultiplexing. For Flongle data, a Phred score cutoff of 4 was used due to overall lower basecalling quality. Next, high-quality reads from multiplexed experiments were demultiplexed using Guppy barcoder 5.0.14 applying adapter and barcode trimming. Demultiplexed, filtered, and trimmed reads were aligned to the same 1000 Genomes assembly reference (hs37d5) used for Illumina RNA-seq data preprocessing. Alignment was performed with minimap2 2.22[56]. Alignment parameters were optimized to achieve the highest rate of reverse-oriented protein-coding genes with the lowest alignment error. SAM-to-BAM conversion, BAM sorting, indexing, and extraction of basic alignment statistics were performed with SAMtools 1.13. Read summarization was performed with featureCounts (from Subread 2.0.3), similar to Illumina data preprocessing. Strand-specific (forward/reverse) and -unspecific count tables were merged for each sample. Quality reports from basecalled, untrimmed reads before demultiplexing, filtered and demultiplexed reads, and aligned reads were generated using NanoPlot 1.28.2[55].

Integration of mRNA abundance data from Illumina and nanopore sequencing. Raw transcript counts from all Illumina and nanopore sequencing runs were combined into one table for downstream analysis, and only the counts of reverse-oriented alignments were used. Illumina sequencing was performed with the TruSeq Stranded mRNA Library Prep system (Illumina), which effectively estimates the abundance of coding cDNA strands. Therefore, only the counts of reverse-oriented alignments were relevant in these data. A similar stranded sequencing approach is not available for nanopore cDNA sequencing, although it is possible with direct RNA sequencing, which, however, requires a large amount of starting material and multiplexing is not officially supported by Oxford Nanopore Technologies. To estimate the abundance of coding cDNA strands, we used only the counts of reverse-oriented alignments. Similarly, we used the transcripts per million (TPM) values derived from reverse-oriented alignment counts, which normalize raw read counts to the total exon length and library

size for each gene and accurately represent mRNA abundance[57]. As a comparable metric for long-read sequencing, we calculated reads per million (RPM) by normalizing reverse-oriented long-read alignment counts to the corresponding library size. Briefly, the total library size was estimated after excluding all genes on chromosomes X and Y, the mitochondrial genome, as well as rRNA and tRNA genes. Gene-wise read counts were divided by the total library size and multiplied by 1,000,000 to calculate RPM values, whereas normalization to total exon length as in TPM calculation was not desired for long-read direct cDNA sequencing.

Inference of pathway activities. To infer signaling pathway activities from RNA-seq data, the PROGENy algorithm implemented in the *progeny* R package was used[29]. PROGENy is based on a list of pathway response genes collected from publicly available perturbation experiments. Version 1.10 of the R package contains data on 14 pathways involved in cancer biology (Androgen, EGFR, Estrogen, Hypoxia, JAK-STAT, MAPK, NFkB, p53, PI3K, TGFb, TNFa, Trail, VEGF, and WNT). The input for PROGENy were the TPM and RPM matrices for Illumina and nanopore sequencing, respectively. Pathway scores were scaled to have a mean of 0 and a standard deviation of 1, the default for the function. Different numbers of footprint genes per pathway (default is 100) were investigated.

Immune cell deconvolution. For immune cell deconvolution from RNA-seq data, CIBERSORT[24] and quanTIseq[23], implemented in the *immunedeconv* R package[58], were employed. CIBERSORT estimates the abundance of 22 immune cell types using 547 genes, whereas in quanTIseq, ten immune cell types are deconvoluted using a set of 170 genes. The input for both algorithms were the TPM and RPM matrices. As recommended by the CIBERSORT authors, quantile normalization was disabled for RNA-seq data, and fractions were calculated in relative mode. QuanTIseq was run with default parameters.

Gene fusion detection. The JAFFAL pipeline was used with standard settings to detect gene fusions[11]. High- and medium-confidence fusions were defined as hits, although no medium-confidence fusions were reported.

Gene fusions were first confirmed using Integrative Genomics Viewer (IGV)[59]. Subsequently, detailed visualizations with coverage and alignment data were generated using the *Gviz* package in R[60]. Gene models for the GRCh37 human genome assembly were downloaded from ENSEMBL using the BioMart service. Only protein-coding transcripts with the GENCODE Basic tag were used to create a metagene model including all possible exons. Fusion breakpoints were determined by whole-genome sequencing and acquired from the clinical bioinformatics workflow of the MASTER cohort[27].

Statistical methods. All statistical analyses were performed on R statistical environment version 4.2.0. All heatmaps were created using the *ComplexHeatmap* package in R[61]. No repeated measurements were performed. Comparison of TPM and RPMs across all protein coding, quanTIseq and CIBERSORT gene sets were done after logarithmic transformation of TPM/RPM values to the normality assumption of t-test.

Data availability

Sequencing data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and CRG, under Accession Number EGAS00001006317 (https://ega-archive.org/studies/EGAS00001006317). Raw (gene-level read counts) and normalized (RPM/TPM) gene expression data are provided in Supplementary Tables S4–S6.

Code availability

Source code for the ShaNTi pipeline is available at https://github.com/cihanerkut/shanti. Source code for the RNA-Seq workflow of the ODCF is available at https://github.com/DKFZ-ODCF/RNAseqWorkflow.

Received: 20 July 2022; Accepted: 6 February 2023
Published online: 09 February 2023

References

1. Lin, B., Hui, J. & Mao, H. Nanopore technology and its applications in gene sequencing. *Biosensors* https://doi.org/10.3390/bios11070214 (2021).
2. Kono, N. & Arakawa, K. Nanopore sequencing: Review of potential applications in functional genomics. *Dev. Growth Differ.* **61**, 316–326. https://doi.org/10.1111/dgd.12608 (2019).
3. Burck, N. *et al.* Nanopore identification of single nucleotide mutations in circulating tumor DNA by multiplexed ligation. *Clin. Chem.* **67**, 753–762. https://doi.org/10.1093/clinchem/hvaa328 (2021).
4. Suzuki, A. *et al.* Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. *DNA Res.* **24**, 585–596. https://doi.org/10.1093/dnares/dsx027 (2017).
5. Thirunavukarasu, D. *et al.* Oncogene concatenated enriched amplicon nanopore sequencing for rapid, accurate, and affordable somatic mutation detection. *Genome Biol.* **22**, 227. https://doi.org/10.1186/s13059-021-02449-1 (2021).
6. Davenport, C. F. *et al.* Genome-wide methylation mapping using nanopore sequencing technology identifies novel tumor suppressor genes in hepatocellular carcinoma. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms22083937 (2021).
7. Euskirchen, P. *et al.* Same-day genomic and epigenomic diagnosis of brain tumors using real-time nanopore sequencing. *Acta Neuropathol.* **134**, 691–703. https://doi.org/10.1007/s00401-017-1743-5 (2017).
8. Kuschel, L. P. *et al.* Robust methylation-based classification of brain tumours using nanopore sequencing. *Neuropathol. Appl. Neurobiol.* https://doi.org/10.1111/nan.12856 (2022).
9. Baslan, T. *et al.* High resolution copy number inference in cancer using short-molecule nanopore sequencing. *Nucleic Acids Res.* **49**, e124. https://doi.org/10.1093/nar/gkab812 (2021).

10. Au, C. H. *et al.* Rapid detection of chromosomal translocation and precise breakpoint characterization in acute myeloid leukemia by nanopore long-read sequencing. *Cancer Genet.* **239**, 22–25. https://doi.org/10.1016/j.cancergen.2019.08.005 (2019).
11. Davidson, N. M. *et al.* JAFFAL: Detecting fusion genes with long-read transcriptome sequencing. *Genome Biol.* **23**, 10. https://doi.org/10.1186/s13059-021-02588-5 (2022).
12. Stangl, C. *et al.* Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. *Nat. Commun.* **11**, 2861. https://doi.org/10.1038/s41467-020-16641-7 (2020).
13. Oka, M. *et al.* Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer. *Genome Biol.* **22**, 9. https://doi.org/10.1186/s13059-020-02240-8 (2021).
14. Yu, T. *et al.* Receptor-tyrosine kinase inhibitor ponatinib inhibits meningioma growth in vitro and in vivo. *Cancers* https://doi.org/10.3390/cancers13235898 (2021).
15. Rodon, J. *et al.* Genomic and transcriptomic profiling expands precision cancer medicine: The WINTHER trial. *Nat. Med.* **25**, 751–758. https://doi.org/10.1038/s41591-019-0424-4 (2019).
16. Schuler, M. *et al.* Rogaratinib in patients with advanced cancers selected by FGFR mRNA expression: A phase 1 dose-escalation and dose-expansion study. *Lancet Oncol.* **20**, 1454–1466. https://doi.org/10.1016/S1470-2045(19)30412-7 (2019).
17. Klaeger, S. *et al.* The target landscape of clinical kinase drugs. *Science* https://doi.org/10.1126/science.aan4368 (2017).
18. Bello, T. & Gujral, T. S. KInhibition: A kinase inhibitor selection portal. *iScience* **8**, 49–53. https://doi.org/10.1016/j.isci.2018.09.009 (2018).
19. Essegian, D., Khurana, R., Stathias, V. & Schurer, S. C. The clinical kinase index: A method to prioritize understudied kinases as drug targets for the treatment of cancer. *Cell Rep. Med.* **1**, 100128. https://doi.org/10.1016/j.xcrm.2020.100128 (2020).
20. Spanheimer, P. M. *et al.* Receptor tyrosine kinase expression predicts response to sunitinib in breast cancer. *Ann. Surg. Oncol.* **22**, 4287–4294. https://doi.org/10.1245/s10434-015-4597-x (2015).
21. Ben-Hamo, R. *et al.* Predicting and affecting response to cancer therapy based on pathway-level biomarkers. *Nat. Commun.* **11**, 3296. https://doi.org/10.1038/s41467-020-17090-y (2020).
22. Wang, X., Sun, Z., Zimmermann, M. T., Bugrim, A. & Kocher, J. P. Predict drug sensitivity of cancer cells with pathway activity inference. *BMC Med. Genom.* **12**, 15. https://doi.org/10.1186/s12920-018-0449-4 (2019).
23. Finotello, F. *et al.* Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34. https://doi.org/10.1186/s13073-019-0638-6 (2019).
24. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457. https://doi.org/10.1038/nmeth.3337 (2015).
25. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782. https://doi.org/10.1038/s41587-019-0114-2 (2019).
26. Milanez-Almeida, P., Martins, A. J., Germain, R. N. & Tsang, J. S. Cancer prognosis with shallow tumor RNA sequencing. *Nat. Med.* **26**, 188–192. https://doi.org/10.1038/s41591-019-0729-3 (2020).
27. Horak, P. *et al.* Comprehensive genomic and transcriptomic analysis for guiding therapeutic decisions in patients with rare cancers. *Cancer Discov.* **11**, 2780–2795. https://doi.org/10.1158/2159-8290.CD-21-0126 (2021).
28. Grunberger, F., Ferreira-Cerca, S. & Grohmann, D. Nanopore sequencing of RNA and cDNA molecules in *Escherichia coli*. *RNA* **28**, 400–417. https://doi.org/10.1261/rna.078937.121 (2022).
29. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat. Commun.* **9**, 20. https://doi.org/10.1038/s41467-017-02391-6 (2018).
30. Miyoshi, T. *et al.* Genomic profiling of large-cell neuroendocrine carcinoma of the lung. *Clin. Cancer Res.* **23**, 757–765. https://doi.org/10.1158/1078-0432.CCR-16-0355 (2017).
31. Zhang, Z. & Wang, M. PI3K/AKT/mTOR pathway in pulmonary carcinoid tumours. *Oncol. Lett.* **14**, 1373–1378. https://doi.org/10.3892/ol.2017.6331 (2017).
32. Viscuse, P. V., Price, K. A., Garcia, J. J., Schembri-Wismayer, D. J. & Chintakuntlawar, A. V. First line androgen deprivation therapy vs. chemotherapy for patients with androgen receptor positive recurrent or metastatic salivary gland carcinoma: A retrospective study. *Front. Oncol.* **9**, 701. https://doi.org/10.3389/fonc.2019.00701 (2019).
33. Yigit, S., Etit, D., Hayrullah, L. & Atahan, M. K. Androgen receptor expression in adenoid cystic carcinoma of breast: A subset of seven cases. *Eur. J. Breast Health* **16**, 44–47. https://doi.org/10.5152/ejbh.2019.5068 (2020).
34. Loria, R. *et al.* HMGA1/E2F1 axis and NFkB pathways regulate LPS progression and trabectedin resistance. *Oncogene* **37**, 5926–5938. https://doi.org/10.1038/s41388-018-0394-x (2018).
35. Doddapaneni, R. *et al.* Fibroblast growth factor receptor 1 (FGFR1) as a therapeutic target in adenoid cystic carcinoma of the lacrimal gland. *Oncotarget* **10**, 480–493. https://doi.org/10.18632/oncotarget.26558 (2019).
36. Humtsoe, J. O. *et al.* Newly identified members of FGFR1 splice variants engage in cross-talk with AXL/AKT axis in salivary adenoid cystic carcinoma. *Cancer Res.* **81**, 1001–1013. https://doi.org/10.1158/0008-5472.CAN-20-1780 (2021).
37. Tchekmedyian, V. *et al.* Phase II study of lenvatinib in patients with progressive, recurrent or metastatic adenoid cystic carcinoma. *J. Clin. Oncol.* **37**, 1529–1537. https://doi.org/10.1200/JCO.18.01859 (2019).
38. Dickson, M. A. *et al.* Phase 2 study of the CDK4 inhibitor abemaciclib in dedifferentiated liposarcoma. *J. Clin. Oncol.* **37**, 11004–11004. https://doi.org/10.1200/JCO.2019.37.15_suppl.11004 (2019).
39. Dickson, M. A. *et al.* Progression-free survival among patients with well-differentiated or dedifferentiated liposarcoma treated with CDK4 inhibitor palbociclib: A phase 2 clinical trial. *JAMA Oncol.* **2**, 937–940. https://doi.org/10.1001/jamaoncol.2016.0264 (2016).
40. Linxweiler, M. *et al.* The immune microenvironment and neoantigen landscape of aggressive salivary gland carcinomas differ by subtype. *Clin. Cancer Res.* **26**, 2859–2870. https://doi.org/10.1158/1078-0432.CCR-19-3758 (2020).
41. Ohtaki, Y. *et al.* Prognostic significance of PD-L1 expression and tumor infiltrating lymphocytes in large cell neuroendocrine carcinoma of lung. *Am. J. Transl. Res.* **10**, 3243–3253 (2018).
42. McBride, M. J. *et al.* The SS18-SSX fusion oncoprotein hijacks BAF complex targeting and function to drive synovial sarcoma. *Cancer Cell* **33**, 1128–1141. https://doi.org/10.1016/j.ccell.2018.05.002 (2018).
43. Sakamoto, Y., Sereewattanawoot, S. & Suzuki, A. A new era of long-read sequencing for cancer genomics. *J. Hum. Genet.* **65**, 3–10. https://doi.org/10.1038/s10038-019-0658-5 (2020).
44. Ho, A. S. *et al.* Genetic hallmarks of recurrent/metastatic adenoid cystic carcinoma. *J. Clin. Investig.* **129**, 4276–4289. https://doi.org/10.1172/JCI128227 (2019).
45. Mitani, Y. *et al.* Comprehensive analysis of the MYB-NFIB gene fusion in salivary adenoid cystic carcinoma: Incidence, variability, and clinicopathologic significance. *Clin. Cancer Res.* **16**, 4722–4731. https://doi.org/10.1158/1078-0432.CCR-10-0463 (2010).
46. Persson, M. *et al.* Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck. *Proc. Natl. Acad. Sci. USA* **106**, 18740–18744. https://doi.org/10.1073/pnas.0909114106 (2009).
47. Huang, C. *et al.* Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell* **39**, 361–379. https://doi.org/10.1016/j.ccell.2020.12.007 (2021).
48. Mock, A. *et al.* EGFR and PI3K pathway activities might guide drug repurposing in HPV-negative head and neck cancers. *Front. Oncol.* **11**, 678966. https://doi.org/10.3389/fonc.2021.678966 (2021).
49. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* **21**, 36. https://doi.org/10.1186/s13059-020-1949-z (2020).

50. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21. https://doi.org/10.1093/bioinformatics/bts635 (2013).
51. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034. https://doi.org/10.1093/bioinformatics/btv098 (2015).
52. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
53. DeLuca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532. https://doi.org/10.1093/bioinformatics/bts196 (2012).
54. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930. https://doi.org/10.1093/bioinformatics/btt656 (2014).
55. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669. https://doi.org/10.1093/bioinformatics/bty149 (2018).
56. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191 (2018).
57. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285. https://doi.org/10.1007/s12064-012-0162-3 (2012).
58. Sturm, G. *et al.* Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445. https://doi.org/10.1093/bioinformatics/btz363 (2019).
59. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. https://doi.org/10.1038/nbt.1754 (2011).
60. Hahne, F. & Ivanek, R. Visualizing genomic data using gviz and bioconductor. *Methods Mol. Biol.* **1418**, 335–351. https://doi.org/10.1007/978-1-4939-3578-9_16 (2016).
61. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849. https://doi.org/10.1093/bioinformatics/btw313 (2016).

Acknowledgements

We thank Stefanie Reinhart, Tatjana Walther, and Lena Figur for technical assistance; Roman Doll for his contribution to the development of the ShaNTi pipeline; the NCT/DKFZ Sample Processing Lab for providing RNA samples; the DKFZ Omics IT and Data Management Facility for clinical bioinformatics workflows; and the NCT/DKFZ/DKTK MASTER team and all members of Scholl and Fröhling groups for valuable discussions. This project was supported by endowment funds (Stiftung für Krebs- und Scharlachforschung) of the Medical Faculty of Heidelberg University to A.M. A.M. was the recipient of a fellowship of the Physician-Scientist Program of the Medical Faculty of Heidelberg University.

Author contributions

A.M. and C.E. conceptualized the study and wrote the paper. A.M., M.B., and C.E. performed nanopore sequencing and collected and analyzed the data. C.E. developed the bioinformatics pipeline. A.M., C.S., S.F., and C.E. interpreted the data. All authors reviewed and commented on the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-29550-8.

Correspondence and requests for materials should be addressed to C.E.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023