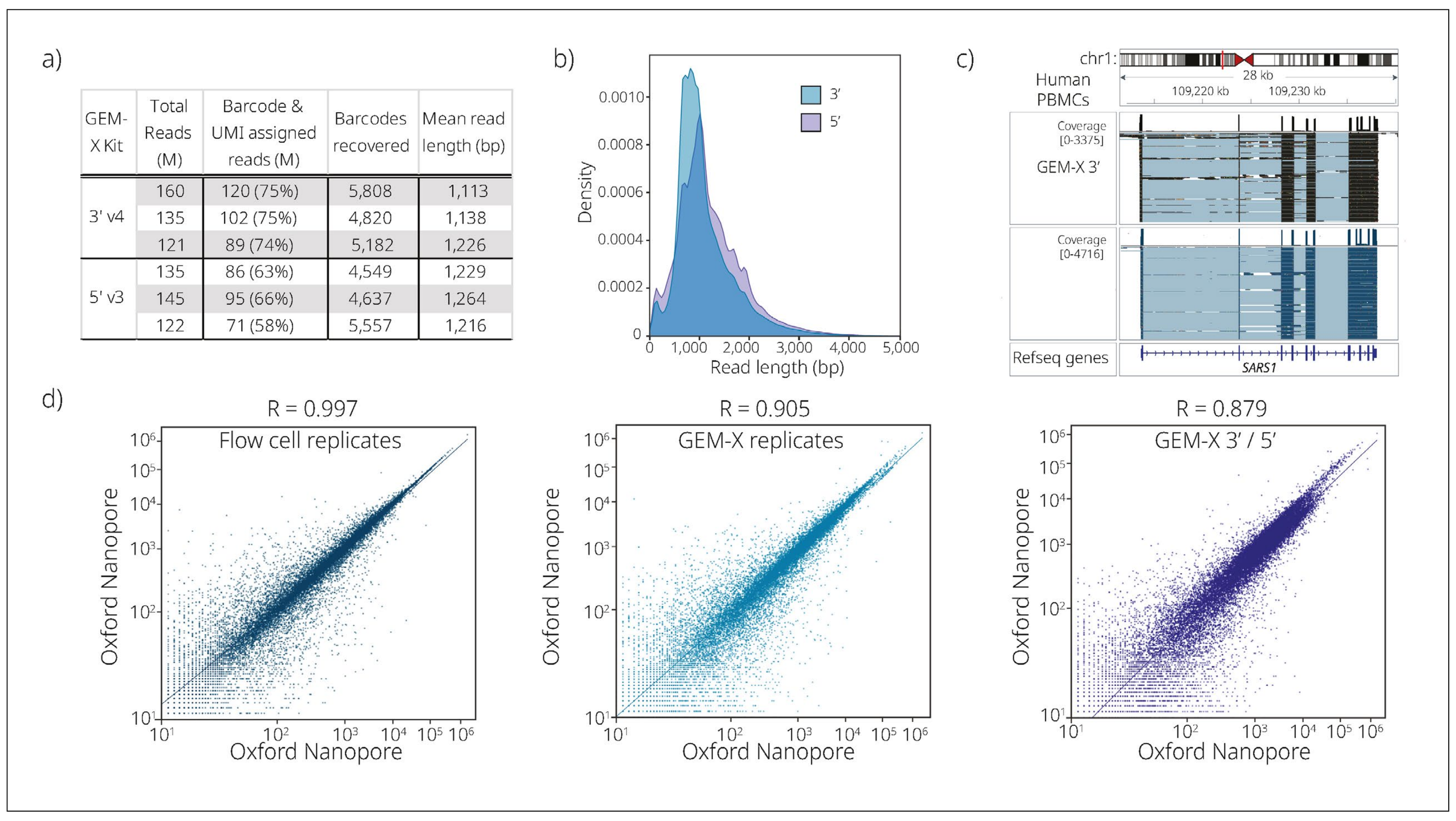
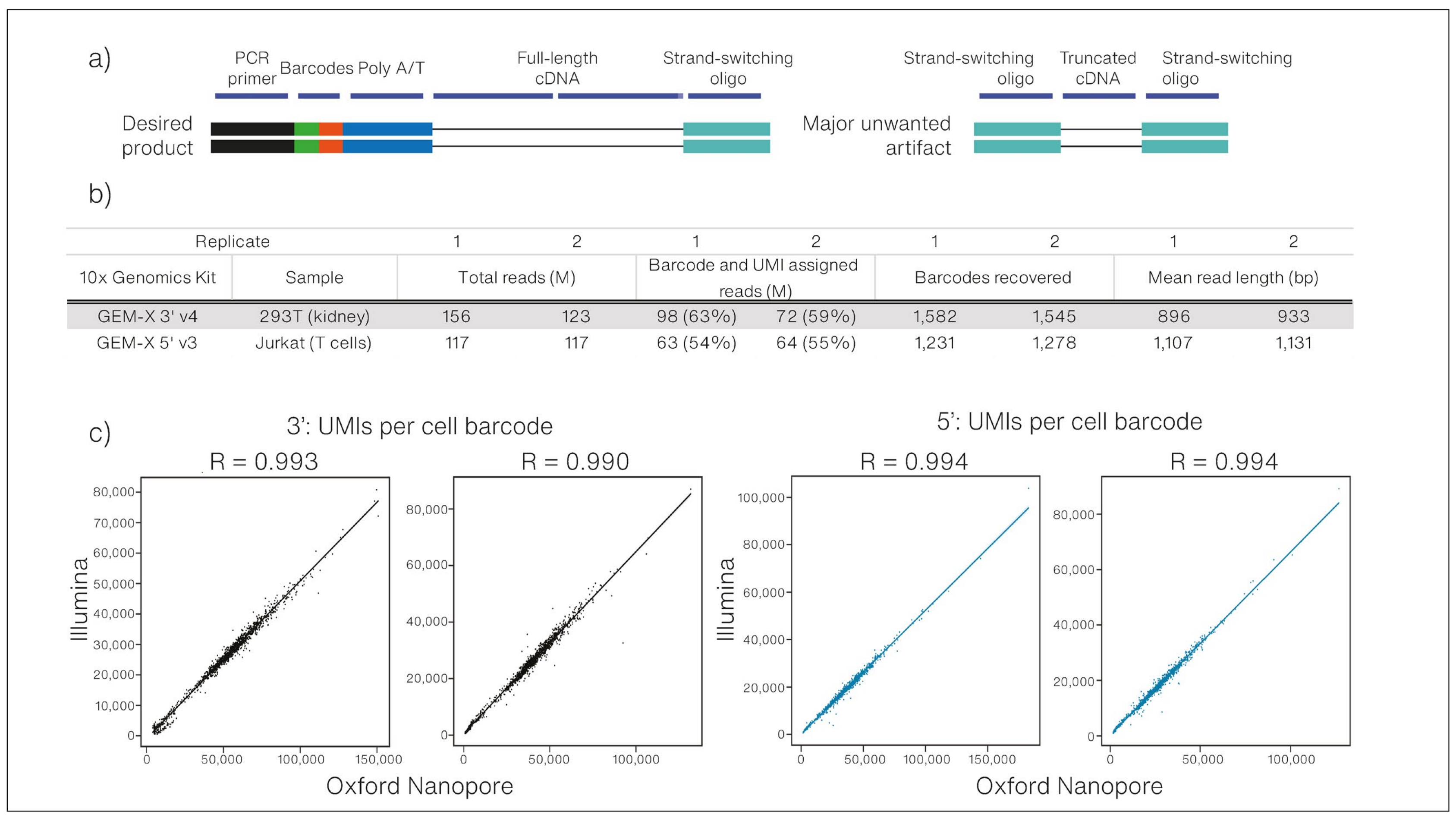
Single-cell transcriptomics with full-length cDNA sequencing provides unique insights into cellular diversity

High-resolution single-cell transcriptomics with full-length cDNA sequencing of 10x Genomics libraries enables detailed analysis of differential gene and isoform expression, fusion transcripts, and genetic variants



GEM-X Kit	Total Reads (M)	Barcode & UMI assigned reads (M)	Barcodes recovered	Mean read length (bp)
3' v4	160	120 (75%)	5,808	1,113
	135	102 (75%)	4,820	1,138
	121	89 (74%)	5,182	1,226
5' v3	135	86 (63%)	4,549	1,229
	145	95 (66%)	4,637	1,264
	122	71 (58%)	5,557	1,216



Replicate		1	2	1	2	1	2	1	2
10x Genomics Kit	Sample	Total reads (M)		Barcode and UMI assigned reads (M)		Barcodes recovered		Mean read length (bp)	
GEM-X 3' v4	293T (kidney)	156	123	98 (63%)	72 (59%)	1,582	1,545	896	933
GEM-X 5' v3	Jurkat (T cells)	117	117	63 (54%)	64 (55%)	1,231	1,278	1,107	1,131

Fig. 1 a) GEM-X 10x Genomics performance b) read lengths c) alignments and d) gene counts to UMIs.

Fig. 2 a) Schematic of full-length transcript and artifact b) GEM-X performance and c) UMI correlation.

Full-length single-cell transcriptome sequencing shows strong reproducibility with 10x Genomics libraries

Single-cell sequencing is critical for resolving transcriptomes in heterogeneous cell populations. 10x Genomics offers two approaches, 3' or 5', that capture different ends of the polyadenylated transcript, but both provide full-length cDNAs. Using human PBMCs, Oxford Nanopore PromethION™ sequencing of libraries produced from the 3' and 5' 10x Genomics GEM-X kits generated >120 M reads, of which >70 M reads can be successfully assigned a barcode and UMI (**Fig. 1a**). All read lengths are sequenced from 3' and 5' 10x single-cell libraries, capturing short and long transcripts (**Fig. 1b**). Using either library prep, transcripts accurately map to the human GRCh38 transcriptome reference with high coverage at annotated exons, as shown for the *SARS1* gene (**Fig. 1c**). The correlation of gene counts between flow cells (left), 10x library replicates (centre) and a 3' and 5' 10x library (right), demonstrate the high reproducibility of 10x libraries sequenced with Oxford Nanopore (**Fig. 1d**).

Barcode and UMI assignment is highly correlated to short reads across GEM-X 3' and 5' 10x libraries

PCR artifacts produced during the amplification of barcoded single-cell cDNAs can limit the proportion of full-length transcript reads in libraries, for example, truncated cRNAs with strand-switching dimers (**Fig. 2a**). Prior to Oxford Nanopore sequencing, artifacts are depleted from 3' libraries using biotin enrichment, while 5' libraries are directly sequenced with ligation chemistry. Nanopore sequencing with PromethION produced from either 10x Genomics GEM-X kits generated >115 M reads, of which >60 M reads can be successfully assigned a barcode and UMI (**Fig. 2b**). Oxford Nanopore sequencing of the GEM-X 3' and 5' 10x libraries yielded mean read lengths of 914 bp and 1,119 bp, respectively. We see a high correlation of gene counts per cell barcode between the single-cell expression levels and those from matching Illumina datasets across all supported 10x Genomics GEM-X kits (**Fig. 2c**).

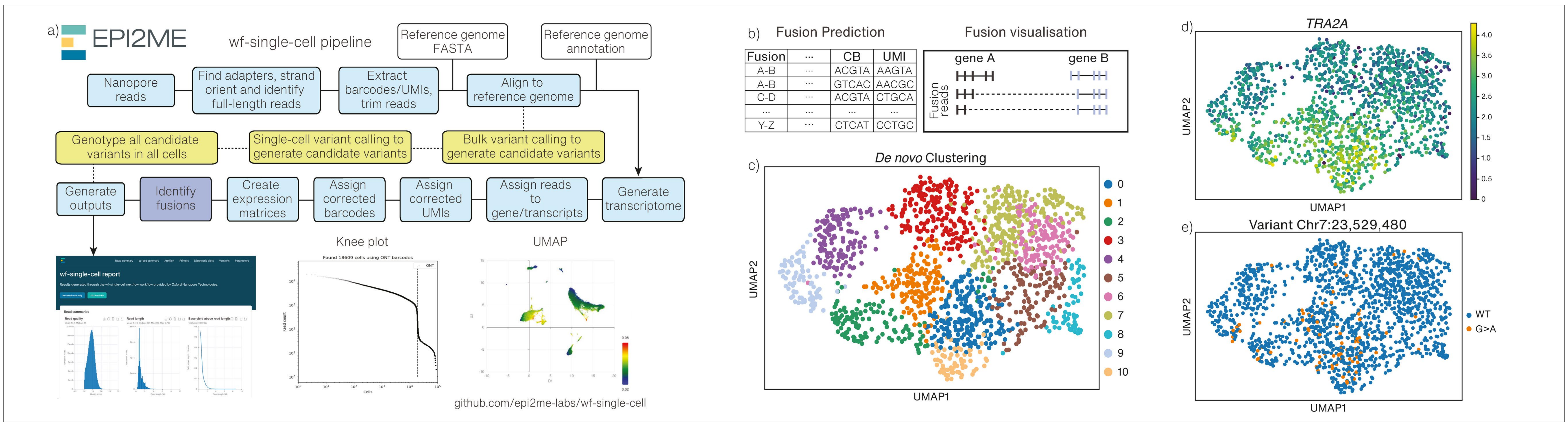


Fig. 3 a) EPI2ME™ single-cell analysis workflow and outputs b) fusion transcript detection c) *de novo* clustering by gene expression d) clustering by *TRA2A* expression e) clustering by *TRA2A* variant expression.

EPI2ME wf-single-cell is an end-to-end solution that enables advanced secondary analysis of complex single-cell data accessible to all bioinformatic skill levels and supports the detection of fusion transcripts and genetic variants

EPI2ME wf-single-cell takes a reference genome and FASTQ or BAM inputs to perform cell barcode (CB)/unique molecular identifier (UMI) assignment, alignment, and expression analysis (**Fig. 3a**). The workflow will output a summary report of basic sequencing and single-cell statistics. Diagnostic plots are generated to assess data quality, including the number of high-quality cells with a knee plot; UMAPs are generated to visualise the gene and isoform expression differences between cells. For fusion transcript detection (purple), we use a state-of-the-art external tool, CTAT-LR-fusion, which uses community detection methods to find fusion gene transcripts, exon-level breakpoints, and links predicted fusions to specific cells. Identified fusions are reported in a machine-readable tabular format on a per-fusion/CB/UMI level, providing a path for further downstream analysis in both pseudo-bulk and single-cell contexts (**Fig. 3b**). The output includes a visualisation of the fusion gene pseudocontigs and fusion-supporting transcripts, highlighting the involved transcripts, exons, and breakpoints. To detect single-cell genomic variants (yellow), variants are called, both in bulk and on individual cells, using the LongShot variant caller with parameters customised to maximise recall and account for allelic expression skew. A second pass is performed to force-genotype all discovered variants in all cells, and all genotypes are merged into a final output VCF. Example UMAP outputs are shown from a malignant melanoma cell line, COLO829BL. First, the Leiden algorithm is used to *de novo* cluster the tumour single-cell expression data into 11 clusters based on gene expression patterns (**Fig. 3c**). Next, we searched for somatic variants enriched in particular cell clusters, which revealed a G to A substitution on chromosome 7 in the *TRA2A* gene, encoding an RNA splicing protein linked to cancer progression (**Fig. 3d**). This gene is expressed in most cells, but the substitution is predominantly observed in a set of neighboring clusters: 0, 1, 2, and 10 (**Fig. 3e**).