nature biotechnology

LETTERS

https://doi.org/10.1038/s41587-020-0497-0

Single-cell RNA counting at allele and isoform resolution using Smart-seq3

Michael Hagemann-Jensen[1], Christoph Ziegenhain[1], Ping Chen[2], Daniel Ramsköld[1], Gert-Jan Hendriks[1], Anton J. M. Larsson[1], Omid R. Faridani[2,3,4] and Rickard Sandberg[1,2] ✉

Large-scale sequencing of RNA from individual cells can reveal patterns of gene, isoform and allelic expression across cell types and states[1]. However, current short-read single-cell RNA-sequencing methods have limited ability to count RNAs at allele and isoform resolution, and long-read sequencing techniques lack the depth required for large-scale applications across cells[2,3]. Here we introduce Smart-seq3, which combines full-length transcriptome coverage with a 5′ unique molecular identifier RNA counting strategy that enables in silico reconstruction of thousands of RNA molecules per cell. Of the counted and reconstructed molecules, 60% could be directly assigned to allelic origin and 30–50% to specific isoforms, and we identified substantial differences in isoform usage in different mouse strains and human cell types. Smart-seq3 greatly increased sensitivity compared to Smart-seq2, typically detecting thousands more transcripts per cell. We expect that Smart-seq3 will enable large-scale characterization of cell types and states across tissues and organisms.

Most single-cell RNA-sequencing (scRNA-seq) methods count RNAs by sequencing a unique molecular identifier (UMI) together with a short part of the RNA (from either the 5′ or 3′ end)[4]. These RNA end counting strategies have been effective in estimating gene expression across large numbers of cells, while controlling for PCR amplification biases, yet RNA end sequencing provides limited coverage of transcribed genetic variation and transcript isoform expression. Moreover, many massively parallel methods suffer from rather low sensitivity (that is, capturing a small fraction of RNAs present in cells)[5]. In contrast, Smart-seq2 has combined higher sensitivity with full-length coverage[6], which enabled allele-resolved expression analyses[7], but at the cost of lower cellular throughput, higher cost and without the incorporation of UMIs. Sequencing of full-length transcripts using long-read sequencing technologies can directly quantify allele- and isoform-level expression, yet their current read depths hinder their broad application across cells, tissues and organisms[2,3]. To overcome these shortcomings, we sought to develop a sensitive short-read sequencing method that would extend the RNA counting paradigm to directly assign individual RNA molecules to isoforms and establish their allelic origin in single cells.

Results

First, we aimed to considerably improve the sensitivity of the Smart-seq2[6] protocol, which consists of oligo-dT priming, reverse transcription followed by template switching, full cDNA amplification using PCR and, finally, Tn5-based tagmentation and library construction (Fig. 1a). After assessing hundreds of different reaction conditions in HEK293FT cells, with the 96 most notable conditions sequenced (Extended Data Fig. 1 and Supplementary Table 1), we found that the highest sensitivity was obtained using Maxima H-minus reverse transcriptase (hereafter called Maxima), in line with recent work[8]. Additionally, we noted that switching the salt during reverse transcription from KCl to either NaCl or CsCl improved sensitivity in Maxima-based single-cell reactions compared to standard KCl conditions (Extended Data Fig. 2), most likely due to reduced RNA secondary structures[9]. Moreover, performing reverse transcription in 5% PEG improved yields by molecular crowding, as recently demonstrated[8], and we added GTPs[10] or dCTPs to stabilize or promote the template-switching reaction (Extended Data Fig. 2a–c). We tested several DNA polymerase enzymes, of which KAPA HiFi Hot-Start polymerase remained most compatible with the reaction chemistry and yielded the highest sensitivity. We constructed a template-switching oligo (TSO) that harbored a primer site consisting of a partial Tn5 motif[11] and a novel 11-bp tag sequence, followed by an 8-bp UMI sequence and three riboguanosines, the latter hybridizing to the non-templated nucleotide overhang at the end of the single-stranded cDNA. After sequencing, the 11-bp tag can be used to unambiguously distinguish 5′ UMI-containing reads from internal reads (Fig. 1a). Therefore, we obtained strand-specific 5′ UMI-containing reads and internal reads from either strand spanning the full transcript without UMIs in the same sequencing reaction (Fig. 1b). The proportions of 5′ to internal reads could be tuned both by altering tagmentation efficiency and during PCR, but it was also affected by sequencer-specific biases (Fig. 1c and Extended Data Fig. 3a–d). We observed that remaining TSO oligos could prime during PCR (Extended Data Fig. 3e) and therefore cause a false increase in UMIs detected (Extended Data Fig. 3f) without affecting gene detection (Extended Data Fig. 3g). Notably, increasing the concentration of forward PCR primer outcompeted TSO priming and minimized this effect (Extended Data Fig. 3f). We termed the final protocol Smart-seq3 and summarized the main improvements in Supplementary Note 1. Profiling gene expression with Smart-seq3 significantly improved the detection of polyA$^+$ protein coding (Fig. 1d) and noncoding RNAs (Extended Data Fig. 4) in HEK293FT cells. Compared to Smart-seq2, the cell-to-cell correlations in gene expression profiles improved significantly with Smart-seq3 (Fig. 1e), and we uncovered remarkable complexity with up to 150,000 unique molecules detected per HEK293FT cell (Fig. 1f). Zooming in on four genes, for which we previously generated single-molecule RNA (smRNA)-fluorescence in situ hybridization (FISH) data[12], revealed that Smart-seq3 detected

[1]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. [2]Integrated Cardio Metabolic Center, Karolinska Institutet, Stockholm, Sweden. [3]Lowy Cancer Research Centre, School of Medical Sciences, University of New South Wales, Sydney, Australia. [4]Garvan Institute of Medical Research, Sydney, Australia. ✉e-mail: rickard.sandberg@ki.se

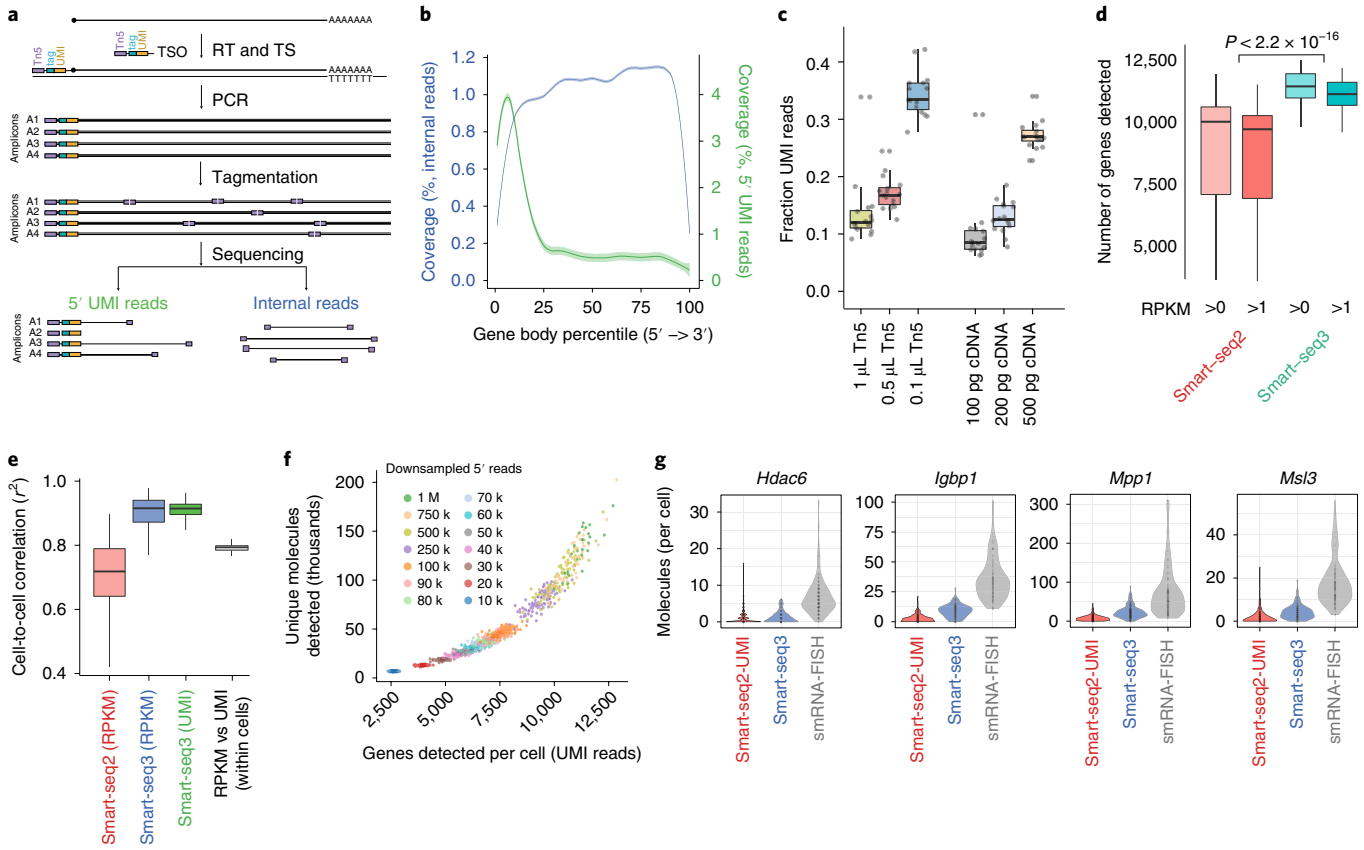


Fig. 1 | Overview of single-cell RNA sequencing in Smart-seq3. a, Library strategy for Smart-seq3. PolyA$^+$ RNA molecules are reverse transcribed, and template switching is carried out at the 5′ end. After PCR pre-amplification, tagmentation via Tn5 introduces near-random cuts in the cDNA, producing 5′ UMI-tagged fragments and internal fragments spanning the whole gene body. **b**, Gene body coverage averaged over HEK293FT ($n = 96$) cells sequenced with the Smart-seq3 protocol. Shown is the mean coverage of UMI reads (green) and internal reads (blue) shaded by the standard deviation. **c**, Effect of tagmentation conditions on the fraction of UMI-containing reads (16 HEK293FT cells per condition). Left, varying Tn5 amounts with constant 200 pg of cDNA input. Right, varying cDNA input with constant 0.5 µl Tn5. **d**, Gene detection sensitivity for Smart-seq2 ($n = 44$ HEK293FT cells) and Smart-seq3 ($n = 102$ HEK293FT cells), downsampled to 1 million raw reads per HEK293FT cell. Shown are numbers of genes detected over 0 or 1 RPKM. *P* value was computed as a two-sided *t*-test. **e**, Reproducibility in gene expression quantification across HEKF293FT cells for Smart-seq2 ($n = 44$ cells) and Smart-seq3 ($n = 102$ cells) at RPKM and UMI level. Shown are adjusted r^2 values for all pairwise cell-to-cell linear model fits in libraries downsampled to 1 million reads per cell. **f**, Sensitivity to detect RNA molecules in Smart-seq3 is shown by summarizing the number of unique error-corrected UMI sequences and genes detected per HEK293FT cell. Colors indicate the per-cell downsampling depth ranging from 10,000 ($n = 102$ cells) to 1,000,000 ($n = 26$ cells) UMI-containing sequencing reads. **g**, Violin plots summarizing the number of molecules detected per fibroblast with Smart-seq2-UMI ($n = 222$ cells) and Smart-seq3 ($n = 172$ cells) and using smRNA-FISH for the four X-chromosomal genes: *Hdac6* ($n = 148$ cells), *Igbp1* ($n = 118$ cells), *Mpp1* ($n = 120$ cells) and *Msl3* ($n = 140$ cells). The boxplots shown in **c**, **d** and **e** show the median, first and third quartiles as a box, and the whiskers indicate the most extreme data point within 1.5 lengths of the box. The HEK and fibroblast data used in Fig. 1 were generated using 0.5 µM forward primer and were not affected by TSO priming.

significantly more molecules per cell than Smart-seq2 but fewer than smRNA-FISH (Fig. 1g). We note that a matched comparison is difficult because, for example, larger cells available for smRNA-FISH are likely excluded in plate-based methods owing to stringent gating during fluorescence-activated cell sorting (FACS). Altogether, this demonstrated that Smart-seq3 has significantly increased sensitivity compared to Smart-seq2.

We next developed a strategy for the in silico reconstruction of RNA molecules. The PCR pre-amplification of full-length cDNA in Smart-seq3 is followed by Tn5 tagmentation, so copies of the same cDNA molecule with the same UMI obtain variable 3′ ends that map to different parts of the specific transcript (Fig. 2a). Therefore, paired-end sequencing of these libraries results in 3′ end sequences that span different parts of the initial cDNA molecule that we can computationally link to the specific molecule based on the 5′ UMI sequence, thus enabling parallel reconstruction of the RNA molecules (Fig. 2a). To experimentally investigate the RNA molecule reconstructions, we created Smart-seq3 libraries from 369 individual primary mouse fibroblasts (F_1 offspring from CAST/EiJ and C57/Bl6J strains) that we subjected to paired-end sequencing. Aligned and UMI error-corrected read pairs[13] were investigated and linked to molecules by their UMI identity. An example of read pairs that were derived from a particular molecule transcribed from the *Cox7a2l* locus in a single fibroblast is visualized in Extended Data Fig. 5a. We then explored how often the reconstructed parts of the RNA molecules covered strain-specific single-nucleotide polymorphisms (SNPs). Notably, unambiguous identification of allelic origin by direct sequencing of SNPs in reads linked to the UMI was observed for 61% of all detected molecules (Fig. 2b) and increased with the SNP density within transcripts (Fig. 2c and Extended Data Fig. 5b). Previous single-cell studies estimated allelic expression as the product of the RNA quantification (in molecules or RPKMs)

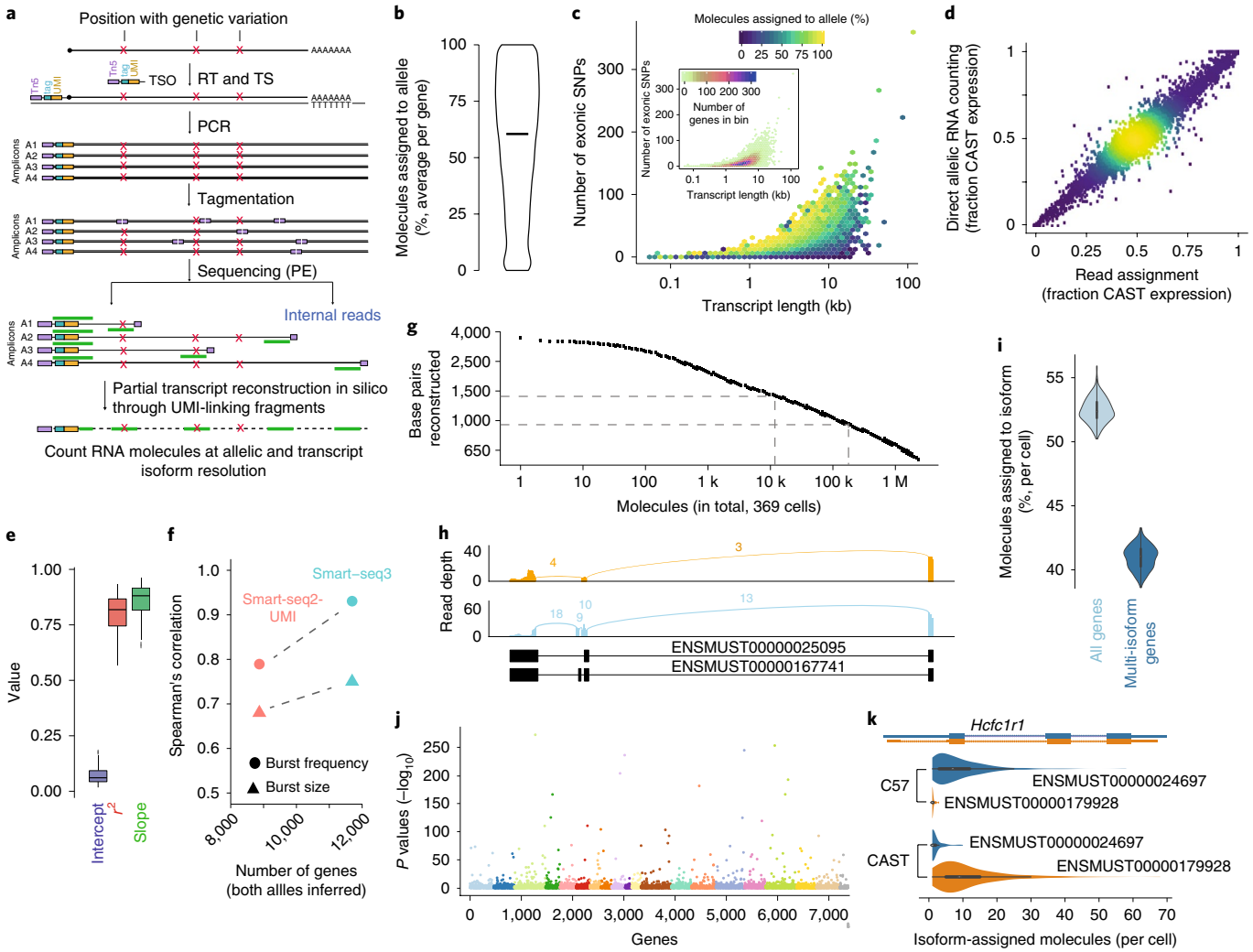


Fig. 2 | Single-cell RNA counting at allele and isoform resolution. a, Strategy for obtaining allelic and isoform resolved information using Smart-seq3. Red crosses indicate transcript positions with genetic variation between alleles. After tagmentation, UMI fragments are subjected to paired-end sequencing (indicated in green), linking molecule-counting 5′ ends with various gene-body fragments that can cover allele-informative variant positions and spanning isoform-informative splice junctions, thus allowing in silico reconstruction of isoforms and assignment to the allele of origin. **b**, Average percentage of molecules that could be assigned to an allele of origin based on covered SNPs, from 369 individual CAST/EiJ × C57/Bl6J hybrid mouse fibroblasts. Only genes detected in more than 5% of cells were considered ($n = 15{,}158$ genes), and their distribution was visualized in a violin plot with a median line. **c**, Effect of transcript length and number of exonic SNPs on allele assignment of RNA molecules. Shown are genes ($n = 15{,}158$) grouped into 50 two-dimensional bins colored by the average gene-wise percentage of molecules assigned to allele of origin. Inset shows the number of genes per visualized bin. **d**, Concordance of allele expression from RNA counting and traditional estimates based on separated expression and allele fractions from internal reads. Shown are the average CAST allele fractions for 12,664 genes with allelic information in at least 2% of the 369 mouse fibroblasts. Dot color reflects the local density of data points. **e**, Results from linear models that compared direct allelic RNA counting to previous read-based estimates of allelic expression, within each of 369 individual fibroblasts. For each cell ($n = 369$), we computed a linear model fit of CAST allele fraction between direct reconstructed molecule assignment and traditional read-based estimates. Shown are box plots of the intercept, slope and r^2 values obtained from each linear model per cell. **f**, Demonstrating the improved abilities of Smart-seq3 to infer transcriptional burst kinetics compared to Smart-seq2-UMI (i.e. Smart-seq2 chemistry combined with a UMI in the TSO). Inference was made in F_1 CAST/EiJ × C57/Bl6J mouse fibroblasts, and we show the Spearman's correlation between the CAST and C57 kinetics across genes for burst size and frequency. Additionally, the *x* axis shows the number of genes for which we could reliably infer the bursting kinetics. **g**, Summarizing the numbers of RNA molecules (*x* axis, $\log_{10}$) reconstructed to different lengths (in base pairs, *y* axis), showing only molecules additionally assigned to a unique transcript isoform. In total, the 1 million longest reconstructed RNA molecules are shown from one experiment with 369 mouse fibroblasts, with molecules shown in descending order. **h**, Sashimi plots visualizing two reconstructed RNA transcripts that supported two distinct transcript isoforms of *Cox7a2l* (ENSMUST00000167741 in orange and ENSMUST00000025095 in light blue), observed in a mouse fibroblast (cell barcode, TTCCGTTCGCGACTAA). **i**, Violin plots showing the percentage of detected molecules that could be assigned to a specific Ensembl transcript isoform, per F_1 CAST/EiJ × C57/Bl6J mouse fibroblast ($n = 369$ cells). Reported are the results on all Ensembl genes or the subset with two or more annotated isoforms ('multi-isoform genes'). The median percentages of assigned molecules per cell were 52.37% for all genes (and 41.04% for multi-isoform genes). **j**, Visualizing significant strain-specific isoform expression in mouse fibroblasts ($n = 369$ cells), colored by chromosomes. The *y* axis shows Benjamini–Hochberg-corrected *P* values ($-\log_{10}$) from individual chi-squared tests performed per gene evaluating association between allelic origin and isoforms. **k**, Visualizing the significant strain-specific isoform expression of *Hcfc1r1* in CAST/EiJ and C57/Bl6J mouse strains. Violin plots depict isoform expression in mouse fibroblasts ($n = 369$), separated by strain and isoform. Top shows the transcript isoform structures. The box plots shown in **e**, **i** and **k** show the median, first and third quartiles as a box, and the whiskers indicate the most extreme data point within 1.5 lengths of the box.

and the fraction SNP-containing reads supporting each allele[7,12,14]. We next investigated how those estimates compared to the direct allelic RNA counting made possible with Smart-seq3. Allelic expression estimates and direct allelic RNA counting showed good overall correlation when aggregated over cells (Fig. 2d). Moreover, using a linear model to quantify the agreement of the two measures across genes within cells revealed a strong correlation (Spearman's rho = 0.82 ± 0.08 and slope = 0.88 ± 0.06) without any apparent bias (intercept = 0.06 ± 0.03) (Fig. 2e). Thus, direct allelic RNA counting is feasible in single cells and validates previous efforts to estimate allelic expression from separated expression and allelic estimates in single cells[7,12,14].

We previously showed that allele-resolved scRNA-seq can be used to infer bursting kinetics of gene expression that are characteristic of transcription[12]. Strikingly, Smart-seq3-based analysis enabled kinetic inference for thousands more genes than using Smart-seq2 alone with a 5′ UMI (11,766 using Smart-seq3 and 8,464 using Smart-seq2-UMI) and with significantly improved correlation between the CAST and C57 alleles (0.94 and 0.75 for Smart-seq3 and 0.79 and 0.68 for Smart-seq2-UMI, respectively, for burst frequency and size) (Fig. 2f and Extended Data Fig. 5c,d). We conclude that Smart-seq3 enables more sensitive reconstruction of transcriptional bursting kinetics across single cells.

We investigated to which lengths RNA molecules could be reconstructed and to what extent information on transcript isoform structures was contained. In our experiment with 369 cells, we observed in total 22,196 molecules reconstructed to a length of 1.5 kb or longer and approximately 200,000 molecules reconstructed to 1 kb or longer (Fig. 2g). On average, we reconstructed 8,710 molecules to a length of 500 bp or longer per cell. To validate the RNA reconstructions, we further amplified (Smart-seq3 pre-amplified) cDNA from two fibroblasts for PacBio sequencing (Supplementary Methods). Comparing error-corrected UMIs detected with both Smart-seq3 and PacBio resulted in 54,302 matched RNA molecules and demonstrated that the Smart-seq3 reconstructions had, on average, captured 46% of the full-length sequence detected by PacBio (Extended Data Fig. 6a). Detailed inspection of one of the longest reconstructed molecules (from the Col1a2 locus) demonstrated that Smart-seq3 had accurately reconstructed 1.9 kb of this 2.3-kb transcript (Extended Data Fig. 6b).

The reconstructed molecules could often be assigned to specific transcript isoforms, here exemplified by Sashimi plots for two reconstructed molecules from the *Cox7a2l* gene (Fig. 2h). Comparing reconstructed RNAs to Ensembl transcript annotations revealed that 53% of all molecules could be assigned to a single isoform (41% when considering only multi-isoform genes) (Fig. 2i). To validate the reconstructions and assignments to transcript isoforms, we generated two sets of RNA spikes that were engineered to harbor genetic variation (on positions in the range of 5–35 nts from transcript start) in combination with unique downstream splicing patterns (Extended Data Fig. 7a). Thus, the 5′ genetic variation was covered by the 5′ UMI reads that directly revealed the splicing pattern of the molecule. First, we noted that over 95% of all 5′ UMI reads originated from the first five bases of the engineered spikes, demonstrating that the spikes were fully reverse transcribed (Extended Data Fig. 7b). The isoform assignments were highly consistent with the genetic variation (Extended Data Fig. 7c), and the power to assign molecules to unique isoforms decreased with more downstream splicing variation (Extended Data Fig. 7d,e). Moreover, comparing the molecules with unique isoform assignments in Smart-seq3 to the isoform detected in matched molecules by PacBio sequencing demonstrated consistent assignment for 99% of molecules (Extended Data Fig. 6c,d). Having validated the reconstructions and isoform assignments with PacBio sequencing and engineered RNA spikes, we sought to survey isoform regulation in mouse strains.

Strain-specific transcript isoform regulation has previously been difficult to study, because the simultaneous quantification of strain-specific SNPs and splicing outcomes on the same RNAs has been challenging. We assigned the in silico reconstructed molecules both to allelic origins and transcript isoform structures, which revealed statistically significant strain-specific (CAST or C57) expression of transcript isoforms for 2,172 genes (adjusted $P < 0.05$, chi-squared test, Benjamini–Hochberg correction; $P < 0.05$, gene-specific permutation test) (Fig. 2j and Supplementary Table 2). For example, transcripts for *Hcfc1r1* were processed into two isoforms (ENSMUST00000024697 and ENSMUST00000179928) that differed both in coding sequence (3-amino-acid deletion from a 12-bp alternative 3′ splice site usage) and in 5′ untranslated region splicing. Notably, the two isoforms had a significant mutually exclusive pattern of expression between strains (adjusted $P < 10^{-208}$, chi-squared test, Benjamini–Hochberg correction) (Fig. 2k). Thus, Smart-seq3 can simultaneously quantify genotypes and splicing outcomes, here exemplified by strain-specific splicing patterns in mice.

Next, we sought to benchmark Smart-seq3 on a more complex sample consisting of many different types of cells. To this end, we sequenced 5,376 individual cells from the Human Cell Atlas (HCA) benchmarking sample[4]— a cryopreserved and complex cell sample composed of human peripheral blood mononuclear cells (PBMCs), primary mouse colon cells and cell line spike-ins of human HEK293T, mouse NIH3T3 and dog MDCK cells. Smart-seq3 clearly separated the cells according to species (Extended Data Fig. 8a,b) and cell types (Fig. 3a), and at the same cutoffs, 77% of cells passed quality filtering, which were significantly higher percentages than the range of 29–63% for other available protocols[4], showcasing the robustness of Smart-seq3 (Extended Data Fig. 8c,d).

Gene detection sensitivity was significantly higher in all cell types when compared to Smart-seq2 (Fig. 3b). This was the case even at shallow sequencing depths and when approaching saturation (Extended Data Fig. 9a). $CD14^+$ monocytes were an exception, which might be due to them being more vulnerable to the year-long freezer storage before FACS and Smart-seq3 profiling. This improvement in the number of genes detected extended into traditionally difficult cell types with low messenger RNA content, such as T cells and B cells, for which we typically observed 1,000 more genes per cell. Notably, we detected two distinct clusters of B cells (using Louvain clustering, also separated in uniform manifold approximation and projection (UMAP) visualization) (Fig. 3a) that were not separated in single-cell data from existing methods[4] (Extended Data Fig. 9b). Differential expression between the B cell populations reported 279 genes with significant expression difference ($P < 0.05$, Wilcoxon rank-sum test, Bonferroni corrected), which included several known marker genes for naïve and memory B cells (Fig. 3c). This demonstrated the improved ability of Smart-seq3 to separate biologically meaningful clusters of cells compared to previous methods.

scRNA-seq data are mainly analyzed by first selecting a set of genes with the most biological variation for downstream clustering and low-dimensionality projections, a procedure that counteracts the typically large amount of Poisson noise in the incomplete sampling of RNA molecules. We hypothesized that the increased sensitivity with Smart-seq3 might mitigate the need to select only the most variable genes, and we therefore analyzed all genes expressed in at least 1% of cells (22,589 genes in total). Notably, both the Louvain clustering and UMAP representation revealed a larger granularity (Fig. 3d) than seen when basing the analysis on just 2,000 most variable genes (Fig. 3a). This increased granularity was not associated with donors but rather reflected the identification of additional cell types (plasma cells and plasmacytoid dendritic cells) and larger structures within T and natural killer (NK) cells (Fig. 3d and Extended Data Fig. 10a,b), further supporting the method's improved resolution of biologically meaningful clusters.

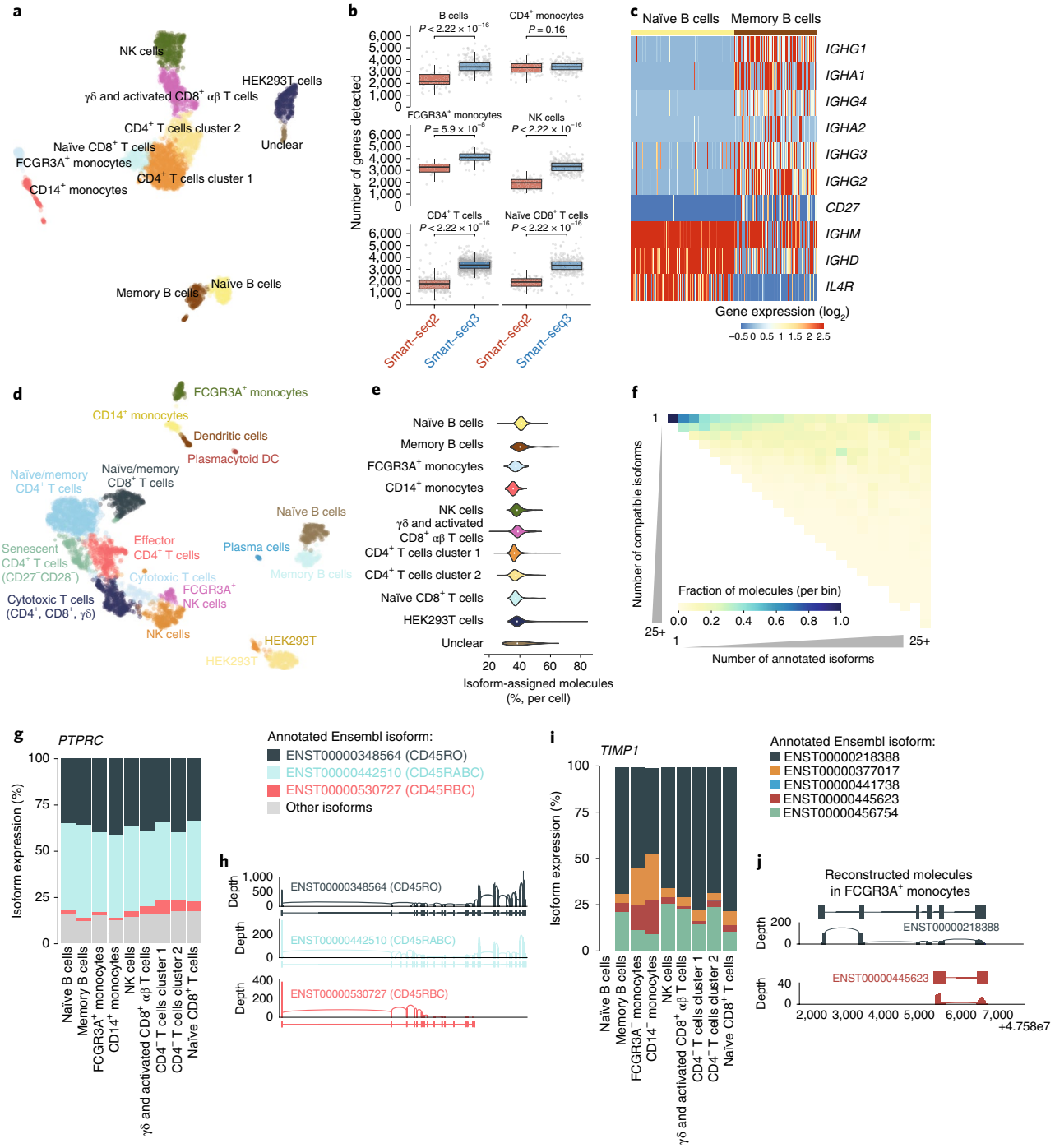


Fig. 3 | Smart-seq3 analysis of a complex human sample. a, Dimensionality reduction (UMAP) of 3,129 human cells sequenced with the Smart-seq3 protocol and colored by annotated cell type, based on the 2,000 genes with highest biological variability. **b**, Comparison of sensitivity to detect genes between Smart-seq2 and Smart-seq3 in various cell types. Cells were downsampled to 100,000 raw reads per cell, and two-sided *t*-test *P* values are annotated for each pairwise comparison. The box plots show the median, first and third quartiles as a box, and the whiskers indicate the most extreme data point within 1.5 lengths of the box. **c**, Heat map showing gene expression for selected marker genes that were expressed at statistically significantly different levels in naïve ($n = 203$) and memory ($n = 163$) B cells. Color scale represents normalized and scaled expression values. **d**, Dimensionality reduction (UMAP) of 3,129 human cells sequenced with the Smart-seq3 protocol and colored by annotated cell type as in **a** but based on the 22,589 genes with expression in at least 1% of cells. **e**, The percentage of reconstructed RNA molecules that could be assigned to a single Ensembl isoform, separated by cell types ($n = 3,089$ cells). Distribution of values is shown as a violin plot with a dot indicating the median. **f**, Matrix showing the fraction of reconstructed molecules that could be assigned to either one or *n* of isoforms (as in **e**) after we filtered the assignments to only those isoforms with detectable expression (TPM > 0) in Salmon (including internal reads without linked UMIs). **g**, Bar plots showing the fraction of molecules assigned to different PTPRC isoforms, separated by cell type and aggregating over all cells within cell types. **h**, Sashimi plots of reconstructed molecules assigned to either the RO or RABC isoform of PTPRC in γδ T cells. **i**, Bar plots showing the fraction of molecules assigned to different TIMP1 isoforms, separating by cell type and aggregating over cells within cell types. **j**, Sashimi plots of reconstructed molecules assigned to two TIMP1 isoforms in FCGR3A+ monocytes.

Investigating the performance of RNA molecule reconstruction across the human cell types revealed that 36–41% of all detected molecules could be assigned to a specific isoform across cell types (Fig. 3e). To investigate the isoform assignment in greater detail, we visualized the number of compatible isoforms for each reconstructed RNA molecule, binning genes by the number of annotated isoforms. Many additional molecules could be assigned to a small set of transcript isoforms (Extended Data Fig. 10c,d). We further reasoned that the internal reads in Smart-seq3 could provide more information on isoform expression. Therefore, we computed isoform expressions using Salmon[15] on all reads from Smart-seq3 and filtered the direct RNA reconstruction-based assignment of molecules to only those isoforms that had detectable expression (transcripts per million (TPM) > 0) in Salmon (Methods). This strategy further increased the assignment of molecules to unique isoforms (42% of all molecules) (Fig. 3f), and we used the Salmon-filtered isoform expression levels for the remainder of the study.

Next, we investigated the patterns of isoform expression across cell types, and 2,186 genes had statistically significant patterns of differential isoform expressions across cell types (adjusted P values < 0.05, Kruskal–Wallis test and Benjamini–Hochberg correction) (Supplementary Table 3). One of the significant genes was protein tyrosine phosphatase receptor type C (PTPRC) (also known as CD45), which can be post-transcriptionally processed into several different isoforms[16], including a full-length isoform (called RABC) and one that has excluded three consecutive exons (called RO). We mainly observed these two isoforms across the human immune cell types, although at significantly varying levels (Fig. 3g). Aggregating the reads supporting these two isoforms in γδ T cells (Fig. 3h) further showed how the reconstructed molecules separated the inclusion or skipping of the three consecutive exons. Other specific isoform patterns were shared by certain cell types; for example, both $CD14^+$ and $FCGR3A^+$ monocytes expressed specific isoforms of the *TIMP1* gene (Fig. 3i,j). Both monocyte populations specifically expressed a shorter isoform of the *TIMP1* gene, whereas the long, full-length isoform was dominant across other cell types (Fig. 3i), again supported by the reconstructed molecules (Fig. 3j). Altogether, these results highlight the new and improved capabilities of using Smart-seq3 to query isoform expression and regulation across cell types.

Discussion

Mammalian genes typically produce multiple transcript isoforms from each gene[17], with frequent consequences on RNA and protein functions. Analysis of transcript isoform expression using short-read sequencing technologies have often focused on individual splicing events (for example, skipped exon) or used the read coverage over shared and unique isoform regions to infer the most likely isoform expression[18,19]. This is because paired short reads seldom carry sufficient information to assess interactions between distal splicing outcomes or combined with allelic expression from transcribed genetic variation. Long-read sequencing technologies can be used to directly sequence transcript isoforms in single cells[2,3]. However, these strategies have limited cellular throughput and depth. For example, the Mandalorion approach provided comprehensive isoform data for seven cells[2], whereas scISOr-seq investigated isoform expression in thousands of cells at an average depth of 260 molecules per cell[3]. In contrast, we obtained on average 8,710 reconstructed molecules per cell (above 500 bp). Moreover, in scISOr-seq, the pre-amplified cDNA was sequenced on both short- and long-read sequencers in parallel to characterize cell types and subtypes, and the isoform-level sequencing data were mainly aggregated over cells according to clusters[3]. The use of two parallel library construction methods and sequencing technologies for the same pre-amplified cDNA from individual cells substantially increases cost and labor.

We developed Smart-seq3 to be both highly sensitive, thus improving the ability to identify cell types and states, and isoform specific, to simultaneously reconstruct millions of partial transcripts across cells. Compared to known transcript isoform annotations, these partial transcript reconstructions were sufficient to assign 40–50% of detected molecules to a specific isoform, which further revealed strain- and cell-type-specific isoform regulation. This reconstruction should improve the ability to perform mapping of trait loci subject to splicing regulation, because both splicing outcomes and transcribed SNPs can now be directly quantified. The full Smart-seq3 protocol was deposited at protocols.io (https://doi.org/10.17504/protocols.io.bcq4ivyw) and can be readily implemented by molecular biology laboratories without the need for specialized equipment.

Several large-scale projects aim to systematically construct cell atlases across human tissues and those of model organisms[20]. These efforts are increasingly relying on scRNA-seq methods that count RNAs toward annotated gene ends (for example, 10× Genomics) while providing little information on isoform expression patterns across cell types and tissues. Moreover, large-scale efforts are also emerging to use single-cell genomics for the systematic analysis of disease (for example, the LifeTime project) to identify disease mechanisms and consequences. As post-transcriptional gene regulation has been tightly linked to disease[21], it would be a missed opportunity for such efforts and atlases to disregard isoform-level expression patterns. In contrast to long-read sequencing efforts, Smart-seq3 simultaneously provides cost-effective RNA counting at isoform resolution within the same assay. This is currently achieved at a cost per sequencing-ready cell library of approximately €0.5–€1 and in 384-well plates with moderate cellular throughput. However, cell atlas efforts could benefit from the ability of plate-based methods to shallowly sequence cells at random and later select rare cells for in-depth sequencing and transcript isoform reconstruction. Altogether, we introduce a scRNA-seq method that is applicable to characterize cell types and annotate cell atlases at the level of gene, isoform and allelic expression.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0497-0.

Received: 22 October 2019; Accepted: 24 March 2020;
Published online: 04 May 2020

References

1. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11**, 22–24 (2014).
2. Byrne, A. et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**, 16027 (2017).
3. Gupta, I. et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4259 (2018).
4. Mereu, E. et al. Benchmarking single-cell RNA sequencing protocols for cell atlas projects. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0469-4 (2020).
5. Ziegenhain, C. et al. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell* **65**, 631–643 (2017).
6. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
7. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
8. Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).

9. Guo, J. U. & Bartel, D. P. RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria. *Science* **353**, aaf5371 (2016).
10. Ohtsubo, Y., Nagata, Y. & Tsuda, M. Compounds that enhance the tailing activity of Moloney murine leukemia virus reverse transcriptase. *Sci. Rep.* **7**, 6520 (2017).
11. Cole, C., Byrne, A., Beaudin, A. E., Forsberg, E. C. & Vollmers, C. Tn5Prime, a Tn5 based 5′ capture method for single cell RNA-seq. *Nucleic Acids Res.* **46**, e62 (2018).
12. Larsson, A. J. M. et al. Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
13. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. zUMIs - a fast and flexible pipeline to process RNA sequencing data with UMIs. *Gigascience* **7** https://doi.org/10.1093/gigascience/giy059 (2018).
14. Reinius, B. et al. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA-seq. *Nat. Genet.* **48**, 1430–1435 (2016).
15. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
16. Martinez, N. M. & Lynch, K. W. Control of alternative splicing in immune responses: many regulators, many predictions, much still to learn. *Immunol. Rev.* **253**, 216–236 (2013).
17. Wang, E. T. et al. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
18. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
19. Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
20. Regev, A. et al. The human cell atlas. *Elife* **6**, e27041 (2017).
21. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nat. Rev. Genet.* **17**, 19–32 (2016).

Methods

Cell cultures. HEK293FT cells (Invitrogen) were cultured in complete DMEM medium containing 4.5 g l^{-1} glucose and 6 mM L-glutamine (Gibco), supplemented with 10% fetal bovine serum (FBS) (Sigma-Aldrich), 0.1 mM MEM nonessential amino acids (Gibco), 1 mM sodium pyruvate (Gibco) and 100 µg ml^{-1} pencillin-streptomycin (Gibco). Cells were dissociated using TrypLE Express (Gibco) and stained with propidium iodide, to exclude dead cells, before distribution into 96- or 384-well plates containing 3 µl of lysis buffer using a BD FACSMelody 100-µm nozzle (BD Biosciences). The Smart-seq3 lysis buffer consisted of 0.5 U $µl^{-1}$ of recombinant RNase inhibitor (RRI) (Takara), 0.15% Triton X-100 (Sigma), 0.5 mM dNTP (Thermo Scientific), 1 µM Smart-seq3 oligo-dT primer (5′-biotin-ACGAGCATCAGCAGCATACGA T_{30}VN-3′; IDT), 5% PEG (Sigma) and 0.05 µl of 1:40,000 diluted ERCC spike-in mix 1 (for HEK293FT cells). For HEK293FT cell experiments with the do-it-yourself (DIY) spike-ins, the lysis buffer contained 0.01 pg DIY spikes per reaction. The plates were spun down immediately after sorting and stored at −80 °C.

Mice handling and derivation of primary fibroblasts. Primary mouse fibroblasts were obtained from tail explants of CAST/EiJ × C57/Bl6J mice (>10 weeks old) and of both sexes. All mouse experiments were performed in accordance with Swedish legislation and approved by the Stockholm North Animal Ethics Committee. Cells were cultured and passaged twice in DMEM high glucose (Invitrogen), 10% embryonic stem cell FBS (Gibco), 1% penicillin-streptomycin (Invitrogen), 1% nonessential amino acids (Invitrogen), 1% sodium pyruvate (Invitrogen) and 0.1 mM β-mercaptoethanol (Sigma), before being stained with propidium iodide and sorted into 384-well plates containing 3 µl of Smart-seq3 lysis buffer. Again, plates were spun down and stored at −80 °C immediately after sorting.

HCA reference sample. The HCA reference sample, consisting of a mix of human PBMCs, mouse colon and fluorescent-labeled cell lines HEK-293T-RFP, NiH3T3-GFP and MDCK-Turbo650, was thawed according to specified instructions[4]. Cells were stained with LIVE/DEAD Fixable Green Dead Cell Stain Kit (Invitrogen), facilitating the exclusion of dead cells as well as NIH3T3-GFP cells. Additionally, both debris and doublets were excluded in the gating. Cells were index sorted into 384-well plates, containing 3 µl of Smart-seq3 lysis buffer, using a BD FACSMelody sorter with 100-µm nozzle (BD Biosciences).

Generation of Smart-seq2 libraries. Smart-seq2 cDNA libraries were generated according to the published protocol[22]. For Smart-seq2-UMI, cDNA libraries were generated as previously described[12]. Recipes for other 'intermediate' Smart-seq2 reactions can be found in Supplementary Table 1. Tagmentation was performed with similar cDNA input and volumes as for Smart-seq3 described below.

Generation of Smart-seq3 libraries. To facilitate cell lysis and denaturation of the RNA, plates were incubated at 72 °C for 10 min and immediately placed on ice afterwards. Next, 1 µl of reverse transcription mix, containing 25 mM Tris-HCL, pH 8.3 (Sigma), 30 mM NaCl (Ambion), 1 mM GTP (Thermo Scientific), 2.5 mM $MgCl_2$ (Ambion), 8 mM DTT (Thermo Scientific), 0.5 U $µl^{-1}$ RRI (Takara), 2 µM of different Smart-seq3 TSOs (see additional table for a list of evaluated TSOs; 5′-biotin-AGAGACAGATTGCGCAATGNNNNNNNNrGrGrG-3′; IDT) and 2 U $µl^{-1}$ of Maxima H-minus reverse transcriptase enzyme (Thermo Scientific), was added to each sample. Reverse transcription and template switching were carried out at 42 °C for 90 min followed by 10 cycles of 50 °C for 2 min and 42 °C for 2 min. The reaction was terminated by incubating at 85 °C for 5 min. PCR pre-amplification was performed directly after reverse transcription by adding 6 µl of PCR mix, bringing reaction concentrations to 1× KAPA HiFi PCR buffer (containing 2 mM $MgCl_2$ at 1× (Roche), 0.02 U $µl^{-1}$ DNA polymerase (Roche), 0.3 mM dNTPs, 0.1 µM Smartseq3 forward PCR primer (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTGCGCAATG-3′; IDT) and 0.1 µM Smartseq3 reverse PCR primer (5′-ACGAGCATCAGCAGCA TACGA-3′; IDT)). PCR was cycled as follows: 3 min at 98 °C for initial denaturation, 20–24 cycles of 20 s at 98 °C, 30 s at 65 °C and 6 min at 72 °C. Final elongation was performed for 5 min at 72 °C. For various iterations and optimization conditions, see Supplementary Table 1 for information about specific conditional changes to library preparation.

Generation of Smart-seq3 libraries to investigate effect of TSO in PCR. Cell lysis, RNA denaturation and reverse transcription was carried out as described above for Smart-seq3 for HEK293FT cells. After reverse transcription, each well was purified using homemade 22% PEG beads (see step 27 in protocol https://doi.org/10.17504/protocols.io.p9kdr4w at protocols.io) at a ratio of 3 µl of beads to 4 µl of cDNA sample. Each well was eluted in 5 µl 10 mM Tris, pH 8. After bead cleanup, PCR was performed in 10 µl with or without additional new fresh TSO added to the reaction. Reaction concentrations were as follows: 1× KAPA HiFi PCR buffer (containing 2 mM $MgCl_2$ at 1× (Roche), 0.02 U $µl^{-1}$ DNA polymerase (Roche) and 0.3 mM dNTPs, with 0 µM, 0.1 µM, 0.5 µM or 1 µM of Smartseq3 forward PCR primer, 0.1 µM Smartseq3 reverse PCR primer and 0.8 µM TSO (when added). PCR was cycled for 24 cycles as described above.

Sequence library preparation. After PCR pre-amplification, all samples, regardless of protocol used, were purified with either AMpure XP beads (Beckman Coulter) or homemade 22% PEG beads (see step 27 in protocol https://doi.org/10.17504/protocols.io.p9kdr4w at protocols.io). Library size distributions were checked on a high-sensitivity DNA chip (Agilent Bioanalyzer), and all cDNA concentrations were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Scientific). cDNA was subsequently diluted to 100–200 pg $µl^{-1}$. Tagmentation was carried out in 2 µl, consisting of 1× tagmentation buffer (10 mM Tris, pH 7.5, 5 mM $MgCl_2$, 5% DMF), 0.08–0.1 µl ATM (Illumina XT DNA sample preparation kit) or TDE1 (Illumina DNA sample preparation kit), 1 µl cDNA and water. Plates were incubated at 55 °C for 10 min, followed by the addition of 0.5 µl 0.2% SDS to release Tn5 from the DNA. Library amplification of the tagmented samples was performed using either 1.5-µl Nextera XT index primers (Illumina) or 1.5 µl custom-designed Nextera index primers containing either 8- or 10-bp indexes (0.1 µM each), differing with a minimal Levenshtein distance of 2 between any two indices. Three µl of PCR mix (1× Phusion Buffer (Thermo Scientific), 0.01 U $µl^{-1}$ Phusion DNA polymerase (Thermo Scientific) and 0.2 mM dNTP) was added to each well and incubated at 3 min 72 °C; 30 s at 95 °C; 12 cycles of (10 s at 95 °C; 30 s at 55 °C; 30 s at 72 °C); and 5 min at 72 °C, in a thermal cycler. For the experiments optimizing the UMI fragment conditions, the following changes to the tagmentation procedure (cDNA input, amount of ATM and time at 55 °C) are shown in Fig. 1c. After tagmentation, samples were pooled, and the pool was purified with Ampure XP beads or 22% homemade PEG beads at a 1:0.6 ratio. Libraries were sequenced at the 75-bp single end, 50-bp single end or 150-bp paired end on a high-output flow cell using an Illumina NextSeq500 instrument, an Illumina HiSeq3000 instrument or a NovaSeq S4 flow cell 150-bp paired end.

Generation of PacBio sequencing libraries of Smart-seq3 primary fibroblast cDNA. Two ng from each selected primary fibroblast Smart-seq3 cDNA library was diluted to 10 µl in water and split into five separate PCR reactions containing 2 µl of diluted cDNA. PCR amplification was performed in 1× KAPA HiFi HotStart ReadyMix buffer (2.5 mM $MgCl_2$, 0.02 U $µl^{-1}$ polymerase and 0.3 mM dNTP), 0.1 µM Smart-seq3 forward PCR primer (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTGCGCAATG-3′; IDT) and 0.1 µM Smart-seq3 reverse PCR primer (5′-ACGAGCATCAGCAG CATACGA-3′; IDT) at 25-µl reaction volume. PCR was cycled as follows: 3 min at 98 °C for initial denaturation, 12 cycles of 20 s at 98 °C, 15 s at 67 °C and 6 min at 72 °C. Final elongation was performed for 5 min at 72 °C. After PCR, the five separate reactions were pooled and purified with 22% homemade PEG beads at a 1:0.8 ratio and eluted in 30 µl of water to achieve less than 1 µg of cDNA per cell. Sequencing libraries were generated from amplified cDNA using the SMRTbell Template Prep Kit 1.0-SPv3 (500–2,000 bp). Each library was sequenced on a PacBio Sequel SMRT Cell 1 M v3 using the Sequel Sequencing Kit v3. SMRTbell library prep and PacBio sequencing were performed at the National Genomics Infrastructure / Uppsala Genome Center.

Generation of DIY spike-ins. Synthetic sequences (1 kb long) were designed to have minimal alignment to the human and mouse genomes and transcriptomes. Different 100-bp stretches were removed from these designs as shown in Extended Data Fig. 7 to mimic alternative isoforms. A 60-bp overlap to pUC19 as well as a T7 promoter were added to the 5′ end of these synthetic sequences and were ordered as gBlocks from IDT. Together with an overlapping oligonucleotide hardcoding a poly(A) tail, these sequences were cloned into the pUC19 vector (linearized by digestion with EcoRI and XbaI (NEB)) using a Gibson Assembly (NEB). The plasmids encoding DIY spike-ins are available from Addgene (136948–136957). Plasmids encoding DIY spike-ins were linearized using HindIII (NEB) and in vitro transcribed individually using the MAXIscript T7 kit (Thermo Fisher) according to the manufacturer's protocol.

Read alignments and gene expression estimation. Raw non-demultiplexed fastq files were processed using zUMIs (version 2.4.1 or newer) with STAR (v2.5.4b) to generate expression profiles for both the 5′ ends containing UMIs as well as combined full-length and UMI data. To extract and identify the UMI-containing reads in zUMIs, find_pattern: ATTGCGCAATG was specified for file1 as well as base_definition: cDNA (23–75; single end), (23–150-bp, paired end) and UMI (12–19) in the YAML file. UMIs were collapsed using a Hamming distance of 1. Human cells were mapped to hg38 genome, and mouse fibroblast cells were mapped against mm10 genome with CAST SNPs masked with N to avoid mapping bias, both supplemented with additional STAR parameters '--limitSjdbInsertNsj 2000000 --outFilterIntronMotifs --RemoveNoncanonicalUnannotated --clip3pAdapterSeq CTGTCTCTTATACACATCT'. Experiments containing HEK293FT cells were quantified with gene annotations from Ensembl GRCh38.91. Mouse primary fibroblast data were quantified with gene annotations from Ensembl GRCm38.91.

Allele calling of F_1 mouse molecules. CAST/EiJ strain-specific SNPs were obtained from the Mouse Genomes Project[23] dbSNP 142 and filtered for variants clearly observed in existing CAST/EiJ × C57/Bl6J F_1 data, yielding 1,882,860 high-quality SNP positions. Uniquely mapped read pairs were extracted using samtools and CIGAR values parsed using the GenomicAlignments package[24] to

match aligned read bases to genomic positions. Reads with coverage over known high-quality SNPs were retained. Basecalls over known SNP positions were extracted and grouped by UMI sequence. Molecules with more than 33% of bases at SNP positions showing neither the CAST nor the C57 allele were discarded, and we required more than 66% of observed SNP bases within molecules to show one of the two alleles to make an assignment.

Inference of transcriptional burst kinetics. Allele-resolved UMI counts were used to generate maximum likelihood inference of bursting kinetics from scRNA-seq data as described previously[12]. Inference scripts are available at https://github.com/sandberg-lab/txburst. To ensure a fair comparison with the data generated in this study, we reprocessed the Smart-seq2 data deposited at the European Nucleotide Archive (E-MTAB-7098) using zUMIs and the same SNP set as described above.

Primary data processing for mixed-species benchmarking sample. The complete data set was mapped against a combined reference genome for human (hg38), mouse (mm10) and dog (CanFam3.1). Cells mapping clearly (>75% of reads) to the mouse or dog were removed. The remaining cells representing HEK293T, PBMCs and potential low-quality libraries were processed using zUMIs (version 2.5.5) and mapped against the human genome only.

Analysis of HCA benchmark samples. First, cells were filtered for low-quality libraries requiring more than 100,000 raw reads, more than 75% of reads mapped to the genome and more than 25% exonic fractions. Further analysis was done within v3.1 of Seurat[25] retaining cells with more than 500 genes detected (intron + exon quantification). Data were normalized ('LogNormalize') and scaled to 10,000 as well as regressing out the total number of counts per cell. The top 2,000 variable genes were found using the 'vst' method and used for principal component analysis dimensionality reduction. The first 20 principal components were used for both SNN neighborhood construction as well as UMAP dimensionality reduction. Lastly, Louvain clustering was applied (resolution = 0.7) to find cell groupings. Major cell types were readily identifiable by common marker genes: $CD4^+$ T cells (CD4, IL7R, CD3D, CD3E and CD3G), $CD8^+$ T cells (CD8A and CD8B), $CD14^+$ monocytes (CD4, CD14 and S100A12), $FCGR3A^+$ monocytes (FCGR3A), B cells (MS4A1, CD19 and CD79A), NK cells (NKG7, LYZ and NCAM1) and HEK cells (high number of genes detected). Naïve T cells were separated from activated by CCR7, SELL, CD27 and IL7R and lack of FAS, TIGIT and CD69. γδ T cells were separated from other T cells by TRGC1, TRGC2 and TRDC and lack of TRAC, TRBC1 and TRBC2.

Molecule reconstruction and transcript isoform assignment. The genomic alignments of 5′ UMI-containing reads and their paired reads from the same fragments were generated by zUMIs (version 2.4.1 or newer) with UMI and cell barcode error correction. Read pairs coming from the same RNA molecule (having the same error-corrected UMI) was merged into reconstructed fragments. The reconstructed sequences were mapped to annotated exonic regions of Ensembl transcript isoforms (Ensembl GRCm38.91 for mouse fibroblast data and Ensembl GRCh38.95 for human HCA data). Reconstructed sequences were compared to annotated transcript structures and represented as a Boolean string indicating which exons were supported by exonic coverage or splice junction coverage ('1') and which regions were without support ('0'). For exons not covered with reads (without any information), 'N' was used to signify lacking data. The Boolean string from the reconstructed molecule was matched to the string corresponding to each reference isoform of the same gene to return compatible isoform(s) for each molecule. Molecule isoform assignments were further corrected based on reads aligning to alternative 5′ and 3′ splice sites of overlapping exons from different isoforms. Furthermore, we developed a script that writes the reconstructed sequences to a bam file (stitcher.py, available at Smart-seq3 github repository).

Isoform assignments by integrating non-UMI reads. Transcriptome bam files generated using zUMIs were demultiplexed per cell, and isoform abundances were quantified using Salmon[15] (v0.14.0) quant command and using the following settings '--fldMean 700 --fldSD 100 --fldMax 2000 --minAssignedFrags 1 --dumpEqWeights'. We corrected the Salmon output for cases where all reads were assigned to one of many possible isoforms belonging to the same equivalent classes. For each cell, isoforms with TPM > 0 from Salmon were considered expressed and used to filter compatible isoforms of the reconstructed molecules. If more than one isoform was compatible with a reconstructed molecule (after Salmon filtering), each compatible isoform obtained a partial molecule count (1/N compatible isoforms).

Strain-specific isoform expression in mouse fibroblasts. To investigate mouse strain-specific isoform expression, we used all molecules with both an allele assigned and only a unique isoform assigned. We considered only genes for which we detected two or more isoforms and expression from both alleles. For each gene, we constructed a contingency table based on the counts of molecules assigned to each allele and isoform. Significance was tested by using the chi-squared test, and the resulting P values were corrected for the multiple testing using the Benjamini–Hochberg procedure. We further scrutinized the significant strain–isoform interactions (with an adjusted $P < 0.05$). For each significant gene, we performed 1,000 independent randomizations of allele and isoform labels of all molecules, and we computed the chi-squared test on each permutation. We further required that the real P values obtained were below the 5% lowest P values from the randomizations.

PacBio data processing. Circular consensus sequencing (CCS) reads were generated from raw reads using the SMRTlink pipeline (v8.0.0.79519). Next, CCS reads were pre-processed by detecting the Smart-seq3 5′ UMI tag sequence ATTGCGCAATG. We used the tag sequence to bring all CCS reads to the same strand. Next, we extracted UMI sequences and cDNA sequences up to the poly(A) tail (≥ 15× A) in an unmapped SAM file. We used STARlong (v2.5.4b) to map the reads to the mouse genome (mm10). Gene assignment, UMI error correction and expression quantification were performed as for Illumina data using the zUMIs pipeline.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All sequencing data have been deposited under ArrayExpress E-MTAB-8735 at the European Bioinformatics Institute.

Code availability

Capacity to process Smart-seq3 libraries has been incorporated in zUMIs (https://github.com/sdparekh/zUMIs). Code for molecule reconstruction and allele- and isoform-resolution assignments are available at Github (https://github.com/sandberg-lab/Smart-seq3).

References

22. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
23. Keane, T. M. et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
24. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
25. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902(2019).

Acknowledgements

We would like to thank H. Heyn for providing us with the HCA sample, K. Annusver and P. Johnsson for help with image analysis, S. Parekh for help with the zUMIs pipeline and B. Reinius for discussions. C.Z. is supported by an EMBO long-term fellowship (ALTF 673–2017). G.-J.H. is funded by Human Frontier Science Program long-term fellowship LT000155/2017-L. This work was supported by grants to R.S. from the European Research Council (648842), the Swedish Research Council (2017–01062), the Knut and Alice Wallenberg Foundation (2017.0110), the Bert L. and N. Kuggie Vallee Foundation, the Göran Gustafsson Foundation and the National Institutes of Health. We would also like to acknowledge UPPMAX, National Genomics Infrastructure, Uppsala Genome Center funded by RFI and VR and Science for Life Laboratory, Sweden, National Genomics Infrastructure in Stockholm, funded by Science for Life Laboratory, and the Knut and Alice Wallenberg Foundation.

Author contributions

M.H.-J. developed Smart-seq3 chemistry, generated scRNA-seq libraries, performed computational analysis, prepared figures and wrote the manuscript text. C.Z. provided input to Smart-seq3 chemistry, developed the reconstruction procedure, performed computational analysis, prepared figures and wrote the manuscript text. P.C. developed the reconstruction procedure, performed computational analyses and prepared figures. D.R. performed computational analysis. G.-J.H. and A.J.M.L. developed the reconstruction procedure. O.R.F. developed Smart-seq3 chemistry. R.S. planned and supervised work and wrote the manuscript.

Competing interests

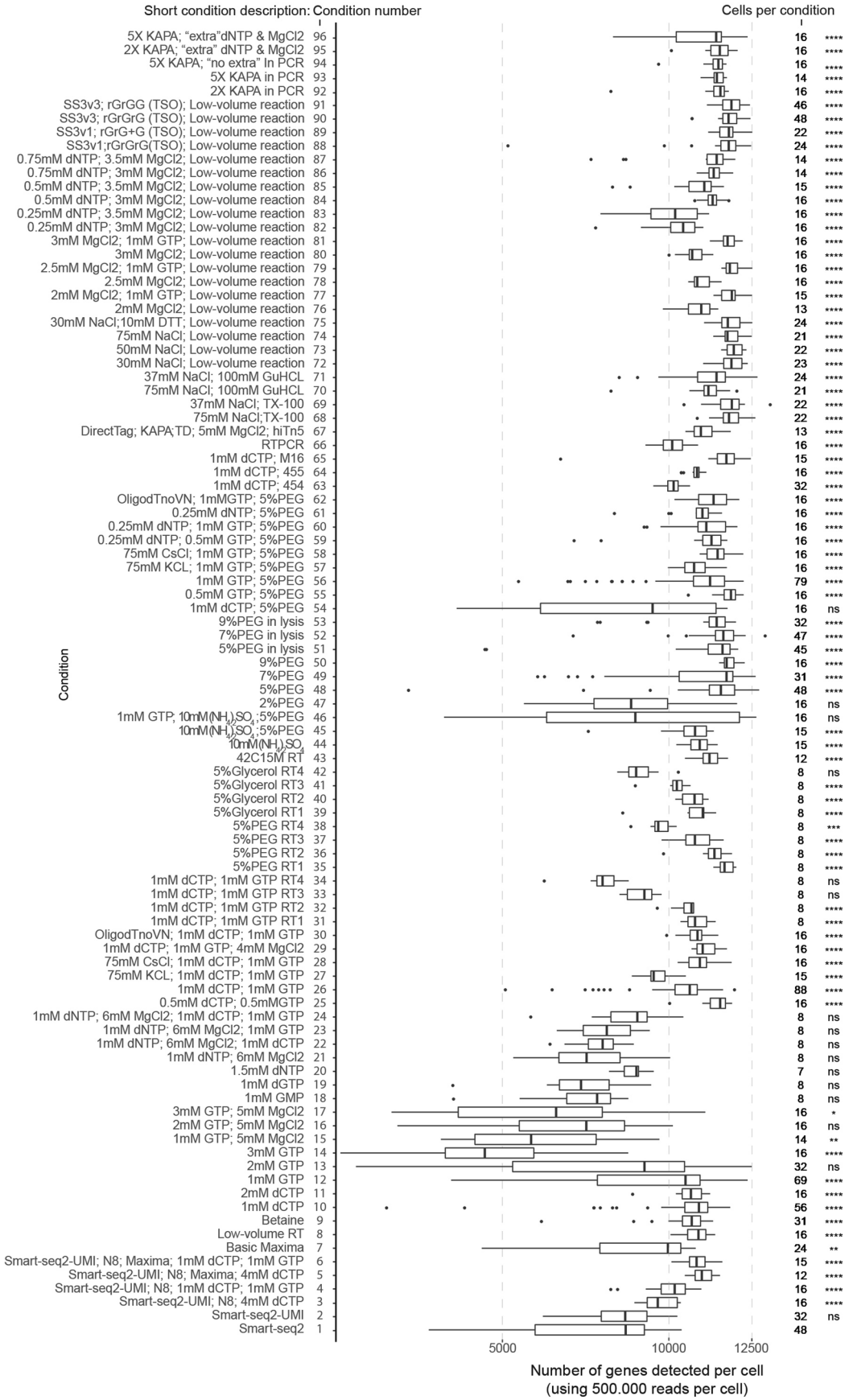
R.S., M.H.-J. and O.R.F have filed a patent application on Smart-seq3.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-020-0497-0.

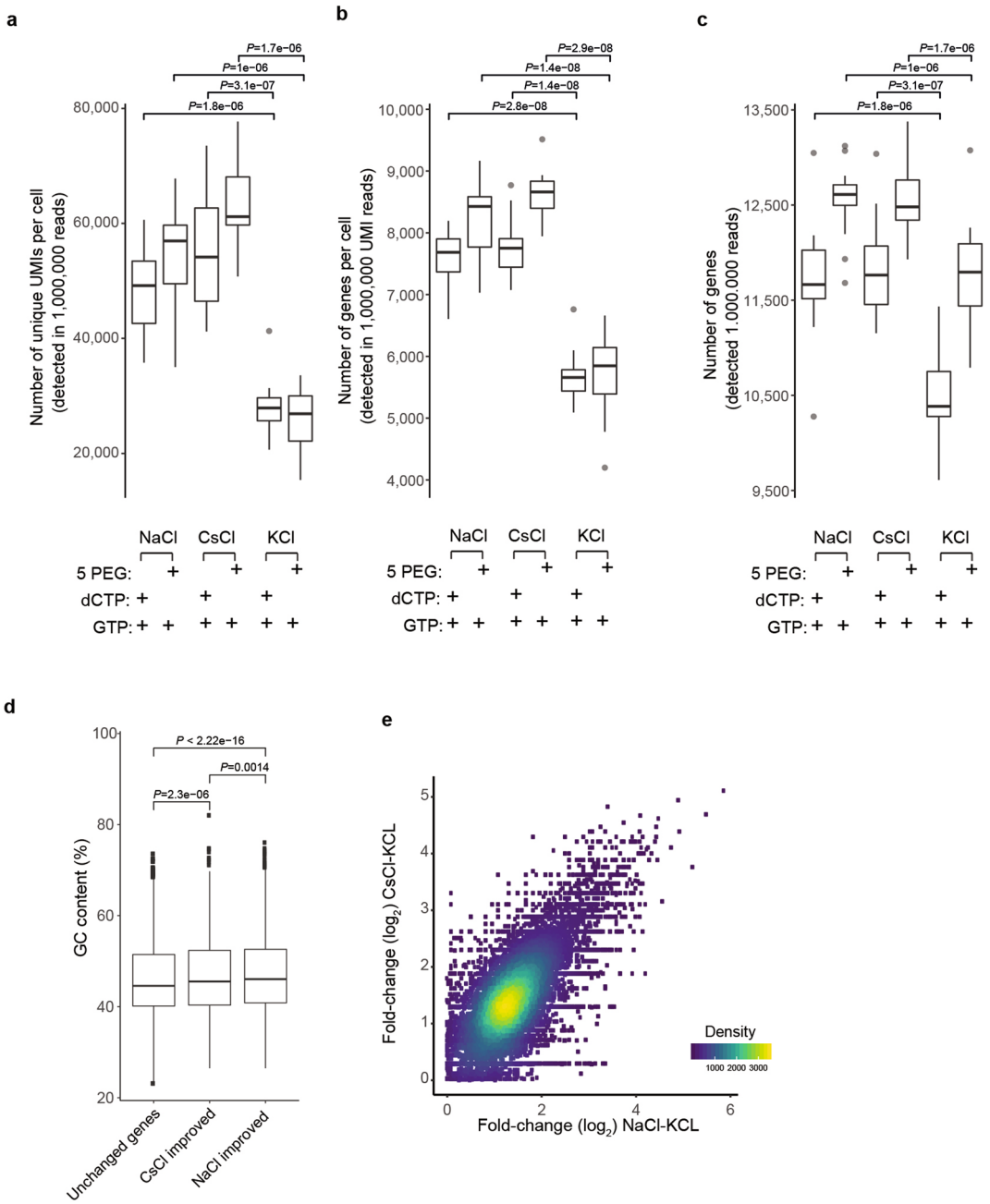
Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0497-0.

Correspondence and requests for materials should be addressed to R.S.**Reprints and permissions information** is available at www.nature.com/reprints.

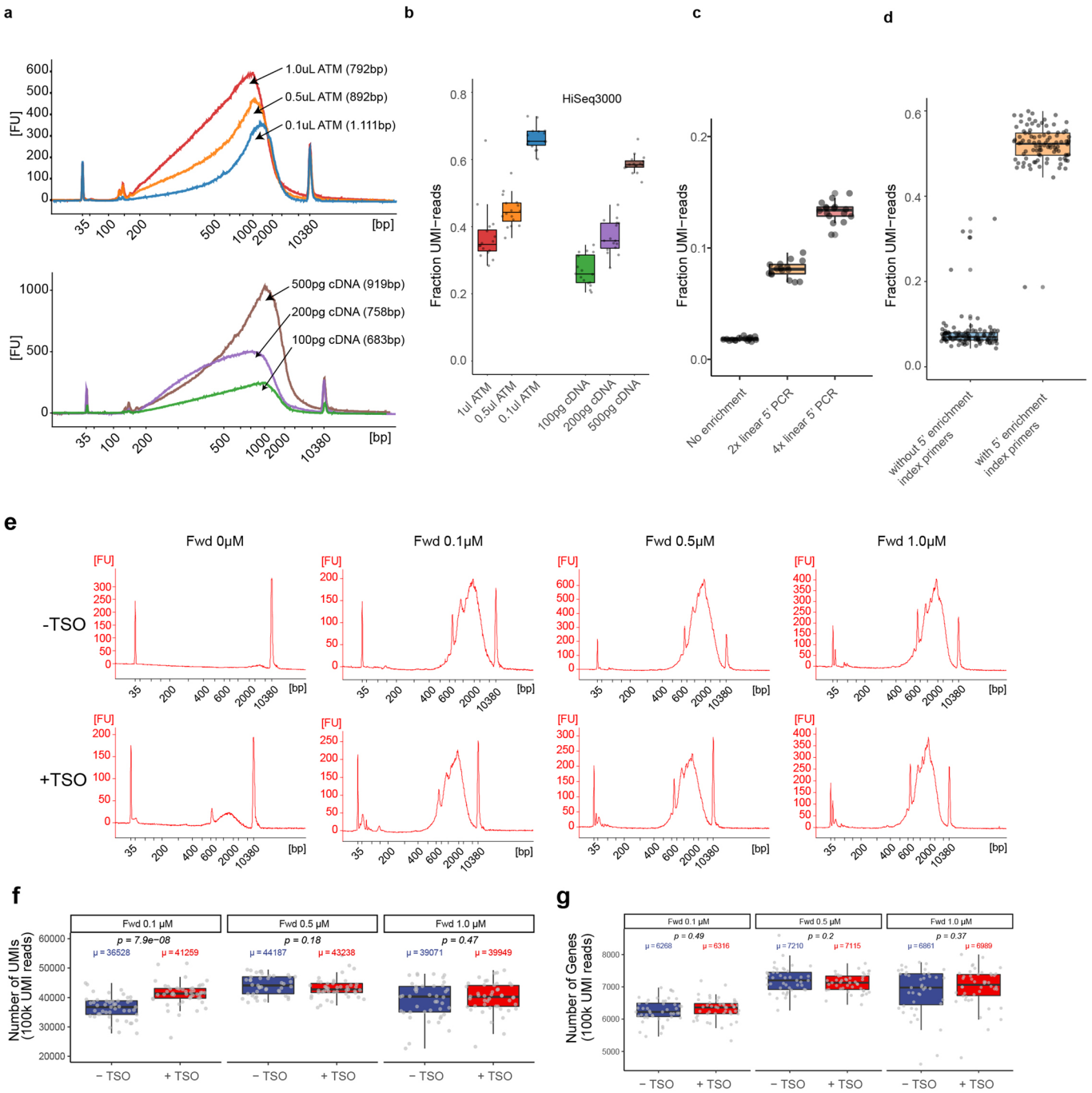


Extended Data Fig. 1 | See next page for caption.

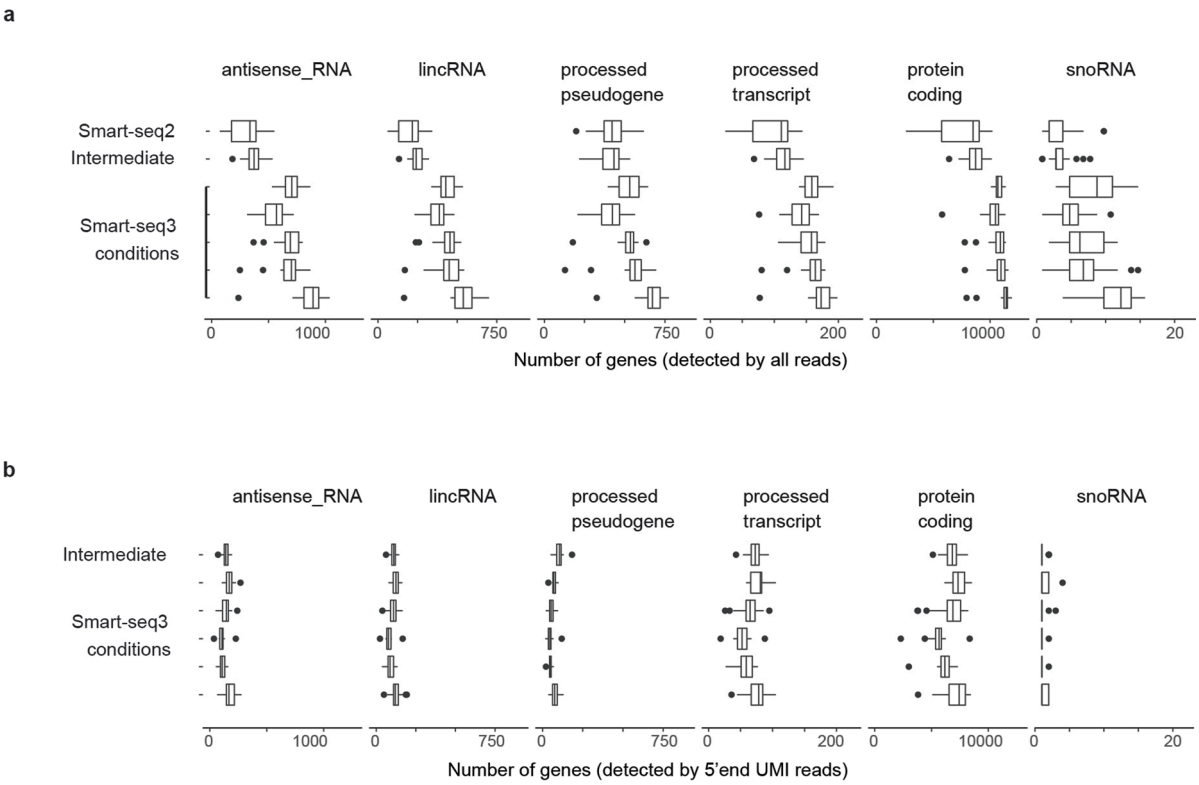
Extended Data Fig. 1 | Overview of sequenced conditions and iterations of Smart-seq3. Each row shows a tested reaction condition and the number of genes detected in individual HEK293FT cells at 0.5 M raw fastq reads. The numbers of individual cells that contained at least one million sequenced reads per condition are listed on the right. Several earlier versions of Smart-seq2 with elements of Smart-seq3 chemistry are included as "Smart-seq-UMI" in this figure. The exact reaction conditions per row are listed in Supplementary Table 1. The asterisks in the right shows the significance level (*0.05, **0.01, ***0.001, ****< 0.001 and ns for non-significance) when comparing the numbers of genes detected in each condition against the gene detection of Smart-seq2 (bottom row) using Wilcoxon rank sum (two-sided). test. Boxplots denote median and first and third quartiles. Whiskers indicate the most extreme data point within 1.5 lengths of the box.



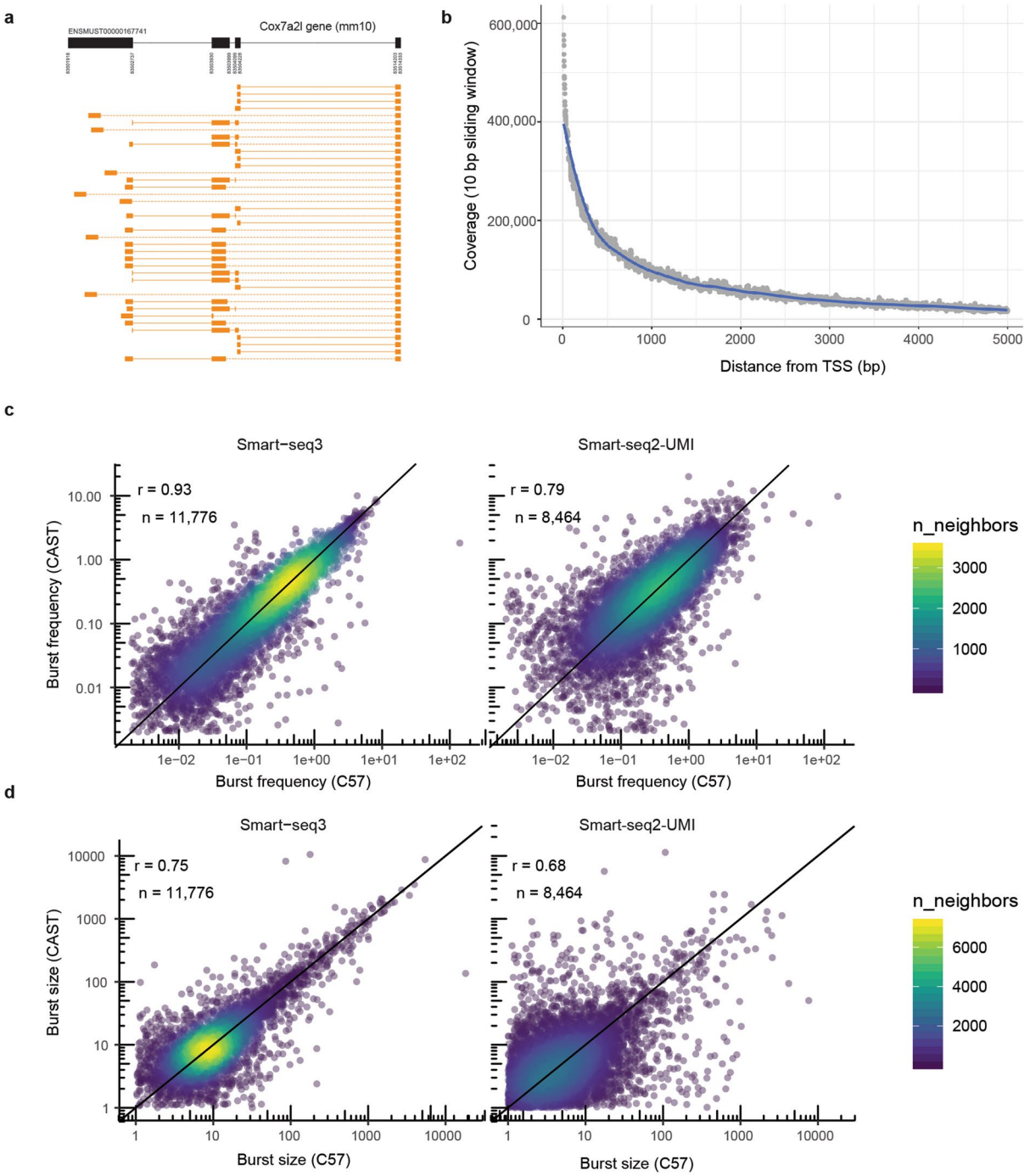
Extended Data Fig. 2 | Effects of salts, PEG and additives on Smart-seq3 reverse transcription. (**a**) Testing the performance of Maxima H-minus reverse transcription reactions on different reaction conditions. For each condition, we summarized boxplots with the number of unique UMIs detected in individual HEK293FT cells at 1 M raw fastq reads. We tested reverse transcription in either the standard KCl based buffer or using NaCl or CsCl. Moreover, we evaluated the effects of adding of 5% PEG or 1 mM dCTP (n = 16 cells per condition). (**b**) Reaction conditions as in (a) summarized against the number of genes identified from 1 million raw UMI-reads per cell (n = 16 cells per condition). (**c**) Reaction conditions as in (a) summarized against the number of genes identified from 1 million raw reads (sub-sampling from both 5′ UMI and internal reads) per cell (n = 16 cells per condition). (**d**) Genes were classified as having improved detection in Na or Cs salt (either going from undetected to detected, or positive log2 FC > 2 in UMI counts) versus detection in K salt buffer (n = 16 cells per condition). Boxplots show GC content in unchanged genes (n = 9,686), and genes with improved detection in Na (n = 8,477) and Cs salts (n = 6,261). Significance in GC content of gene sets were evaluated using two-sided t-tests, as indicated in figure. Boxplots denote median and first and third quartiles. Whiskers indicate the most extreme data point within 1.5 lengths of the box. (**e**) Genes with improved detection in Na and Cs are highly consistent. Shown are log2 fold-changes for genes when compared UMI counts in Na and Cs salt buffer compared to K salt buffer.



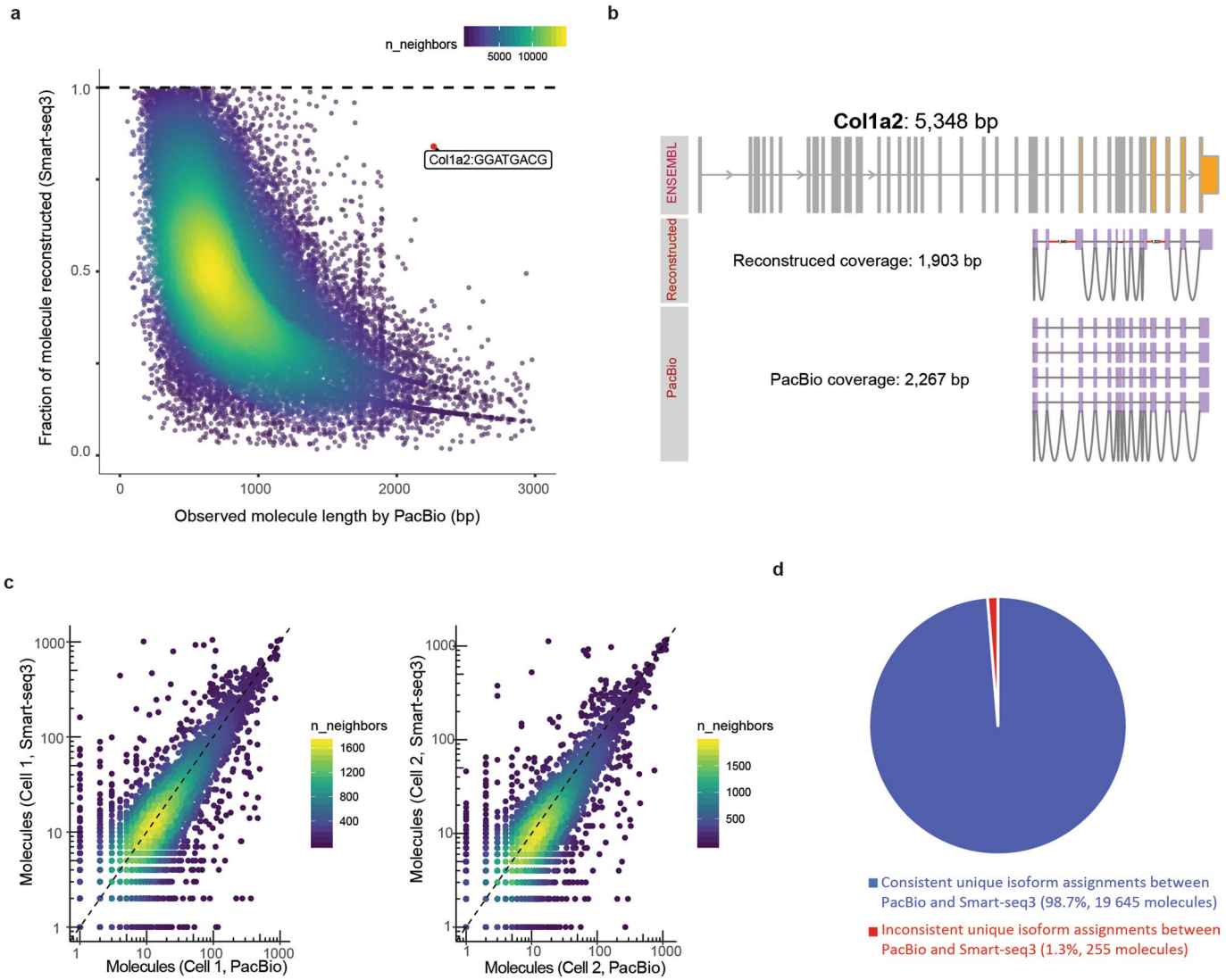
Extended Data Fig. 3 | Tuning 5′ to internal read proportions and template switching oligo PCR priming. (**a**) Bioanalyzer traces of libraries shown in Fig. 1c, demonstrating their different length distributions. (**b**) Sequencing the libraries shown in (a) on an Illumina HiSeq3000 results in higher fractions of 5′ UMI reads than when the same libraries are sequenced on the Illumina NextSeq500 (shown in Fig. 1c) (n = 16 HEK293FT cells per condition). Sequence machine biases are likely fragment length related. (**c**) Enrichment of 5′ UMI containing reads after tagmentation with a linear PCR step (Forward pre-amplification PCR primer) of either 2 of 4 cycles, before adding index primers and index PCR (n = 16 HEK293FT cells per condition). (**d**) Increased UMI containing reads with addition of custom i5 Illumina Index oligos targeting the 5′ tag during index PCR (n = 96 Fibroblasts per condition). **(e)** HEKF293FT cell bioanalyzer traces showing the effect and ability of the template switching oligo priming in PCR in absence and presence of varying amount of forward PCR primer. **(f)** UMIs detected at 100.000 UMI-reads at varying forward PCR primer concentrations with and without the presence of template switching oligo in PCR reaction (n = 48 HEK293FT cells per condition). **(g)** Number of genes detected from 100.000 UMI containing reads with increasing amount of forward PCR primer with or without the presence of the template switching oligo in PCR reaction (n = 48 HEK293FT cells per condition). Significance in (f and g) was evaluated by two-sided t-tests, indicated on the figures. Boxplots denote median and first and third quartiles. Whiskers indicate the most extreme data point within 1.5 lengths of the box.



Extended Data Fig. 4 | Improved detection of protein-coding and non-coding RNAs with Smart-seq3. (**a**) Variants of Smart-seq3 reactions (n = 16 HEK293FT cells per condition) show improved detection of protein coding RNAs and also other classes of RNAs, including poly-A + lincRNAs, antisense RNAs, processed pseudogenes, processed transcripts and snoRNAs, compared to Smart-seq2 (n = 48 HEK293FT cells) and earlier experimental versions of Smart-seq2 with UMIs (here called "intermediate") (n = 32 HEK293FT cells). (**b**) Shows genes of similar RNA classes detected by UMI containing reads in Smart-seq2 protocols using UMIs (here called "intermediate") (n = 32 cells) and Smart-seq3 variants (n = 16 cells per condition).

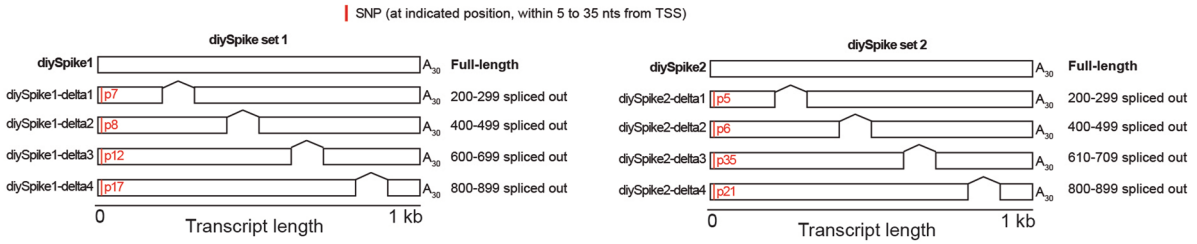


Extended Data Fig. 5 | Visualization of read pairs from a single transcribed molecule and detailed comparison of burst kinetics inference based on Smart-seq2-UMI and Smart-seq3 data. (a) Visualization of read pairs sequenced from one molecule from the Cox7a2l locus. Top show the exons and introns in the Cox7a2l locus, with genomic coordinates (mm10). Each row shows a unique read pair, where orange boxes show the mapping of sequences onto the genomic loci, dotted lines indicate that the sequences are connected by the read pairs and solid lines indicate that the exon-intron junction was captured in the sequenced reads. Note, all read pairs combined span essentially the full transcript, meaning that for this molecule we could reconstruct the full transcript. (**b**) Coverage over CAST SNPs in 5′ UMI-containing read pairs in 369 mouse fibroblasts. Shown is the coverage in number of read pairs in a 10 bp sliding window of SNP distance relative to the TSS of their gene. Blue line indicates a loess fit of the data points. (**c**) Scatter plots showing the burst frequencies inferred for the C57 (x-axis) and CAST (y-axis) alleles for genes in mouse fibroblasts. The left plot shows the results based on Smart-seq3 data and the right panel shows the results from using Smart-seq2-UMI data. (**d**) Scatter plots show the burst sizes inferred for the C57 (x-axis) and CAST (y-axis) alleles for genes in mouse fibroblasts. The left plot shows the results based on Smart-seq3 data and the right panel shows the results from using Smart-seq2-UMI data.

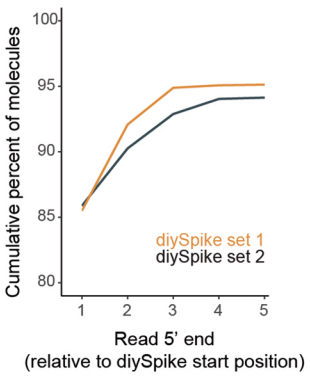


Extended Data Fig. 6 | PacBio sequencing of cDNA from two mouse fibroblasts. (**a**) Scatter plot showing the observed RNA molecule length with PacBio (x-axis) against the reconstructed RNA molecule with Smart-seq3 (y-axis). Molecules were matched by the sequencing of the same UMI per gene in the respective sequencing method. (**b**) Detailed information on the Col1a2 gene, for which Smart-seq3 had reconstructed 1,903 bp. The reconstruction was consistent with the 2,267 bp transcript sequenced by PacBio and both shorter than the 5,348 bp full-length Ensemble annotation. The two gaps in the Smart-seq3 reconstructions are shown as red lines. (**c**) Scatterplots showing the correspondence between observed molecules from PacBio sequencing (x-axis) and the observed molecules from Illumina sequencing from Smartseq3 cDNA libraries from two mouse fibroblast cells. (**d**) Pie-chart showing the consistency in isoform assignment between the PacBio and Smartseq3 assigned isoforms.

a



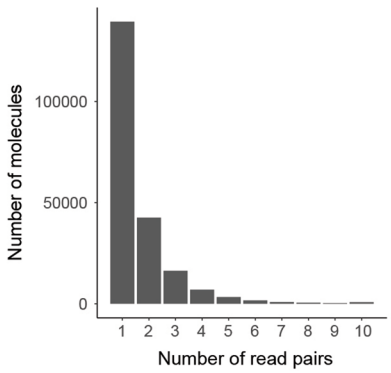
b



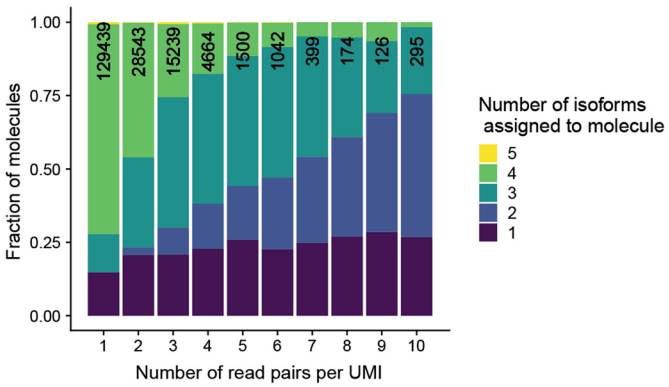
c

diySpike	Read pairs with SNP	Consistent isoform assignment	Inconsistent isoform assignment	Percent false assignments
diySpike1d1	17 666	7 970	18	0,2%
diySpike1d2	17 727	16 369	417	2,5%
diySpike1d3	15 372	2 501	24	1,0%
diySpike1d4	16 496	236	7	3,0%
diySpike2d1	15 071	11 191	22	0,2%
diySpike2d2	18 386	11 089	17	0,2%
diySpike2d3	19 947	3 521	20	0,6%
diySpike2d4	17 941	309	4	1,3%
Total	138 606	53 186	529	1,0%

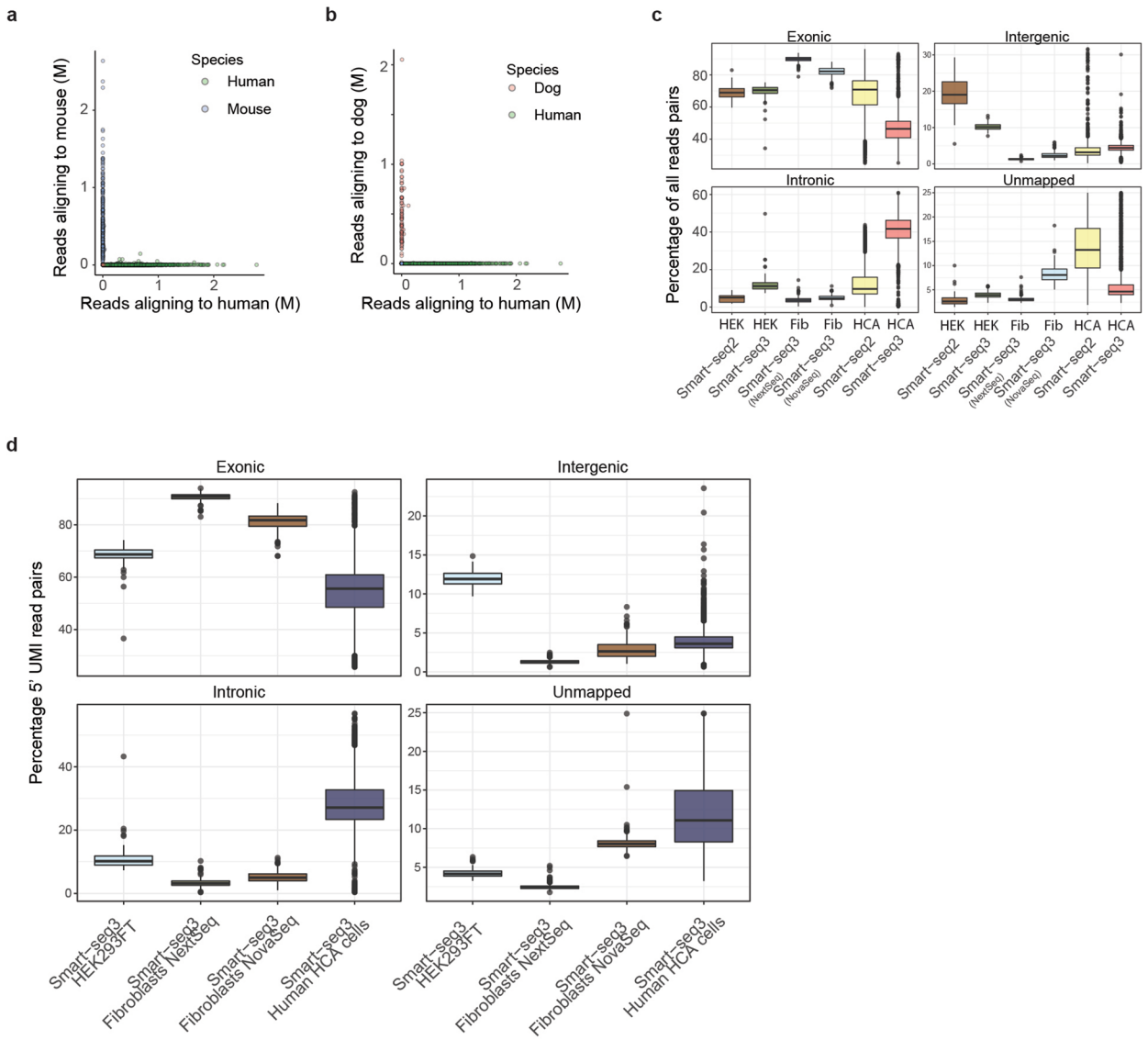
d



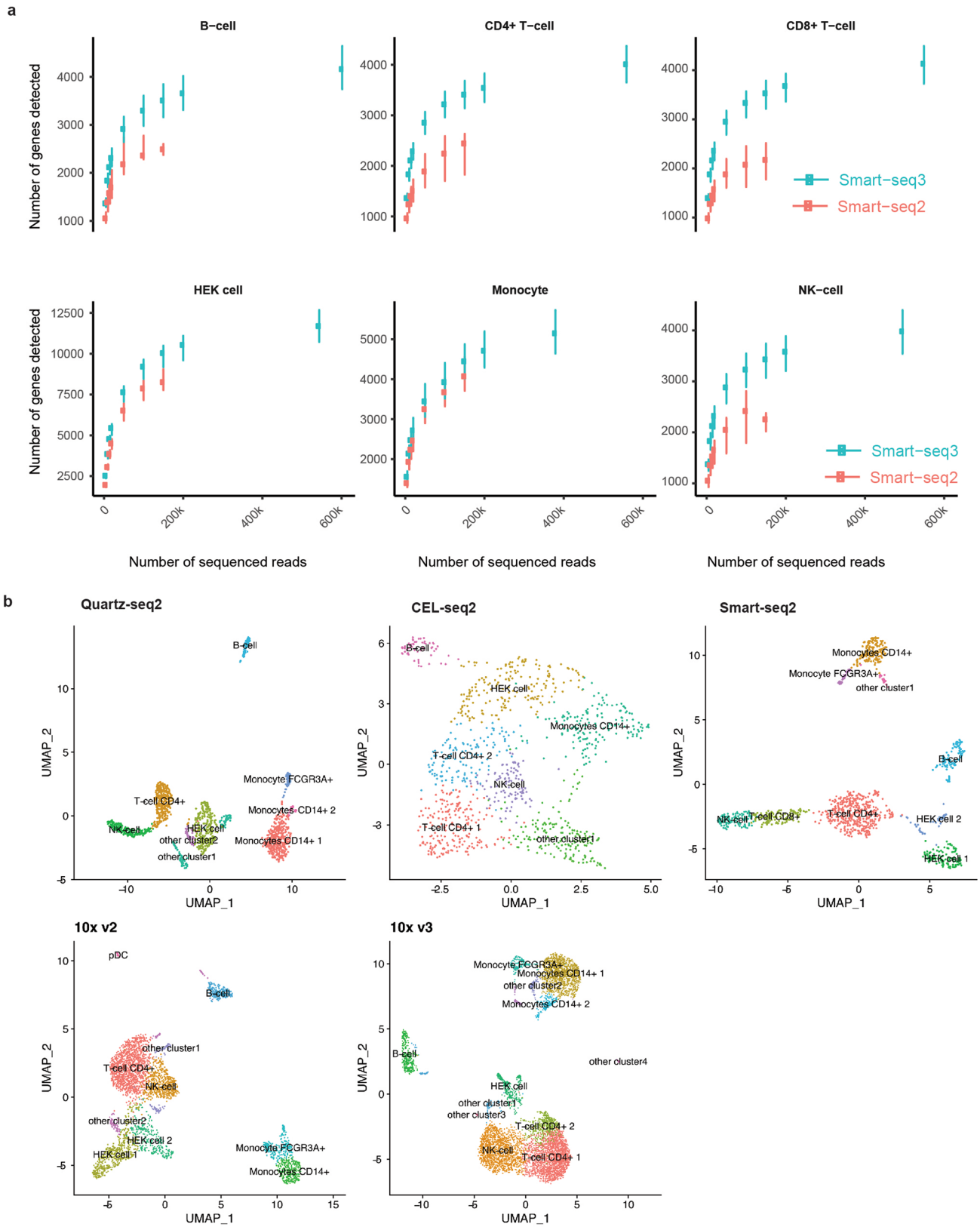
e



Extended Data Fig. 7 | Engineered DIY spikes for reconstruction validation. (**a**) Overview illustrations of the two sets of DIY spikes designed (set 1 and 2, respectively) and each set contains a full-length isoform and four shorter isoforms that harbor connected genetic variation in positions 5 to 35 (marked in red) and 100 bp downstream exclusions. (**b**) Cumulative percentage of spiked-in molecules that had its most 5′ base within the five first bases on the designed spike-in sequence. (**c**) Table summarizing correct and incorrect assignment of read pairs towards each diySpike isoform. (**d**) Histogram showing the read depth over the diySpikes, that is the number of read pairs per observed molecule. (**e**) Fraction of molecules assigned to a unique isoform or compatible with two or more isoforms as a function of the unique number of read pairs from each molecule.

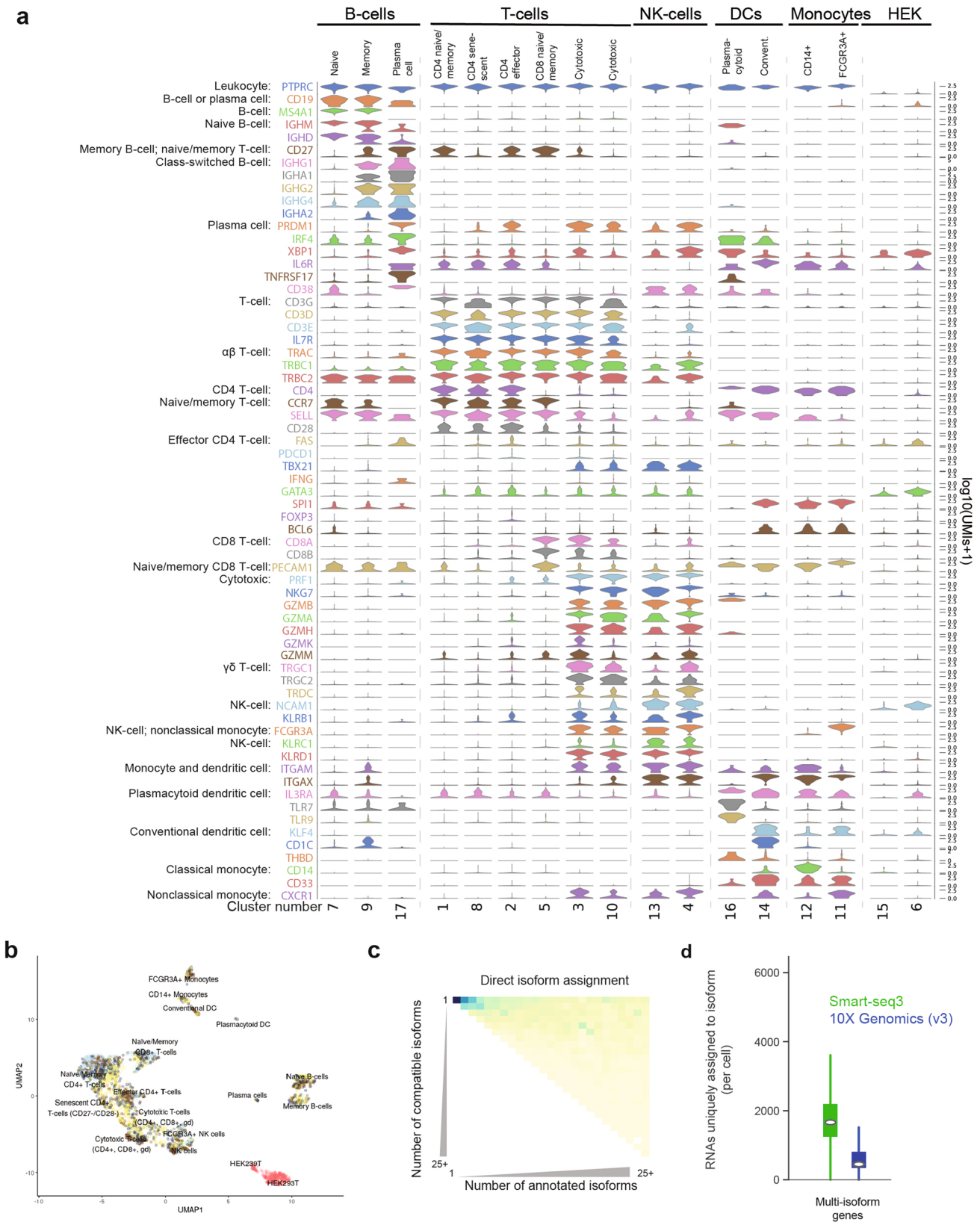


Extended Data Fig. 8 | Species-mixing, doublets and mappings statistics in Smart-seq3. (**a**) Scatter plot showing the number of reads that aligned to human (x-axis) and mouse (y-axis) for the complex HCA sample that contained both human, mouse and dog cells. (**b**) Scatter plot showing the number of reads that aligned to human (x-axis) and dog (y-axis) for the complex HCA sample that contained both human, mouse and dog cells. Few cells show any signal towards more than one genome, demonstrating a very low doublet rate. (**c**) Percentage of unmapped read pairs, and read pairs that aligned to exonic, intronic and intergenic regions. Separated per protocol (Smart-seq2 and Smart-seq3) and experiment (HEK293FT, Mouse Fibroblasts, HCA cells). (**d**) Mapping statistics for 5'UMI-containing read pairs in Smart-seq3. Percentage of unmapped read pairs, and read pairs that aligned to exonic, intronic and intergenic regions. Separated per experiment (HEK293FT, Mouse Fibroblasts, HCA cells). The boxplots shown in (c and d) show the median, first and third quartiles as a box, and the whiskers indicate the most extreme data point within 1.5 lengths of the box.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Gene detection of Smart-seq3 and HCA comparison against other scRNA-seq methods. (a) Number of genes detected in Smart-seq3 and Smart-seq2 as a function of sequence depth. The median detection of genes across cells were represented as dots and the lines indicate the first and third quartiles. Separate plots were generated for cells of different cell types, as indicated on top of each figure item. The respective sample-sizes for Smart-seq3 and Smart-seq2 are: B-cell (n = 366, n = 112 cells), CD4+ T-cell (n = 1,270, n = 356 cells), CD8+ T-cell (n = 665, n = 222 cells), HEK cell (n = 236, n = 62 cells), Monocyte (n = 200, n = 302 cells), NK-cell (n = 352, n = 152 cells). **(b)** UMAP visualizations of sequenced HCA sample cells on different scRNA-seq protocols (data from Mereu et al. 2019), colored according to the Louvain clustering performed independently on cells from each protocol. The same computational pipeline and parameters was used for these analyses as in Figure 3a, except requiring a depth of just 10,000 reads per cell. Please note, this analysis is not intended to be a thorough benchmarking of methods as the data has merely been scaled and not sub-sampled to account for differences in sequencing depths or cell numbers between protocols. Instead the full data per protocol (Quartz-seq2: n = 1,422 cells, CEL-seq2: n = 750, Smart-seq2: n = 1,160, 10x v2 n = 3,592, 10x v3 n = 6,175 cells) was analyzed and run through a standardized scRNA-seq analysis pipelines, revealing that the B-cells do not easily separate with these other methods.



Extended Data Fig. 10 | HCA cluster markers, donor details and additional information on isoform assignments. (**a**) Violin plots showing the distribution of expression for genes across all cell type clusters revealed in Figure 3d (total number of cells n = 3,129). Genes were selected to inform on overall types of cells and known sub-type markers. (**b**) UMAP of HCA sample cells (n = 3,129) with Smart-seq3 (as in Figure 3d) but colored according to donors. (**c**) Matrix showing the fraction of reconstructed molecules that could be assigned to either one or N number of isoforms, where molecules were first grouped by the number of annotated isoform available for its genes. (**d**) Boxplots showing the number of molecules per cell (Smart-seq3 n = 3,129 cells, 10x v3 n = 6,175 cells) with unique assignments in Smart-seq3 and 10x Genomics (v3) for genes with more than one annotated isoform in Ensembl. Boxes indicate first and third quartiles with the median marked in white. Whiskers indicate the most extreme data point within 1.5 lengths of the box.

nature research

Corresponding author(s): Rickard Sandberg

Last updated by author(s): Mar 18, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used the following software to collect data: FACS software (BC FACSChorus 1.3), sequencer software (Illumina NextSeq Control Software 2.2.0), bioanalzyer software (Agilent Bioanalyzer 2100 Expert B.02.10.SI764), bcl2fastq v2.20.0.422. .
Data analysis	Analysis in R and Python used the following packages: R (v3.6.3); ggplot2 (v3.3.0); data.table (v1.12.8); Seurat (v3.1.2); zUMIs (v2.5.1 & v2.7.0); STAR (v2.5.4b); STARlong (v2.5.4b); samtools (v1.9); pigz (v2.4); python (v3.6.9); scipy (v.1.2.0); pandas (v0.24.1); numpy (v1.16.1); pysam (v0.15.2); pyranges (v0.0.76); Salmon (v0.14.0); SMRTlink pipeline (v8.0.0.79519). Custom code and scripts generated within this project is available at Github (https://github.com/sandberg-lab/Smart-seq3).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data has been deposited in the ArrayExpress and European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EBI) with accession: E-MTAB-8735

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were not predetermined using statistical analysis, but cell type comparisons or methods comparisons typically contained hundreds of cells per group. For the methods development part (e.g. Extended Data Figure 1) we used 15 to 50 cells per conditions.
Data exclusions	Single-cell RNA-seq data were filtered according to established criteria for removing technically failed cells. Cutoffs are listed where appropriate, and involved reads per cell, fraction reads mapping uniquely to the genome and exons.
Replication	All experiments were performed across hundreds of individual cells. We validated reconstructions by sequencing matched cDNA libraries on PacBio system.
Randomization	Not relevant because FACS sorting of individual cells into random wells of microplates.
Blinding	Investigators were not blinded to groups of samples.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK293FT: Thermo Fisher. NiH3T3-GFP and MDCK-Turbo650: provided in the HCA reference samples (Mereu et al. Nature Biotechnology 2020)
Authentication	HEK293FT were authenticated by PCR-single-locus-technology (Eurofins Forensik). NiH3T3-GFP and MDCK-Turbo650 were not authenticated as they were already present in the reference sample prepared in Mereu et al. 2020.
Mycoplasma contamination	HEKF293FT were confirmed free of mycoplasma contamination (Eurofins). NiH3T3-GFP and MDCK-Turbo650 were not authenticated as they were already present in the reference sample prepared in Mereu et al. 2020.
Commonly misidentified lines (See ICLAC register)	HEK293FT were used but authenticity was confirmed (see above).

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mouse F1 offspring of CAST/EiJ X C57/Bl6J crosses were used in this study. All F1 mice were 10 weeks or older and we used both male and female mice.
Wild animals	This study did not involve wild animals.

Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	All mouse experiments were performed in accordance to Swedish legislation and approved by the Stockholm North Animal Ethics Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.