Supplementary items for

"Smart-seq2 for sensitive full-length transcriptome profiling in single cells"

by *Simone Picelli, Åsa K. Björklund, Omid R. Faridani, Sven Sagasser, Gösta Winberg and Rickard Sandberg*

Supplementary Figure 1. Experiments with alternative LNA modified TSOs

Supplementary Figure 2. Single-cell results comparing template switching oligonucleotides

Supplementary Figure 3. Validation of single-cell RNA-seq results using another analysis pipeline

Supplementary Figure 4. Comparing single-cell transcriptomic data generated with Smart-seq2, Quartz-seq and SMARTer

Supplementary Figure 5. Mapping statistics of single-cell RNA-seq libraries

Supplementary Figure 6. Detailed analyses of GC biases in single-cell RNA-seq data

Supplementary Figure 7. Single-cell results comparing PCR preamplification enzyme

Supplementary Figure 8. Detailed analyses of read coverage across transcripts

Supplementary Figure 9. Read coverage across transcripts

Supplementary Figure 10. Characterization of artificial gene peaks in single-cell RNA-seq data

Supplementary Figure 11. Assessing the technical and biological variability in single-cell transcriptomics using Smart-seq2

Supplementary tables are attached as Excel spreadsheets:

Nature Methods: doi:10.1038/nmeth.2639

Supplementary Table 1. Listing and analyses of cDNA yield and length from purified RNA

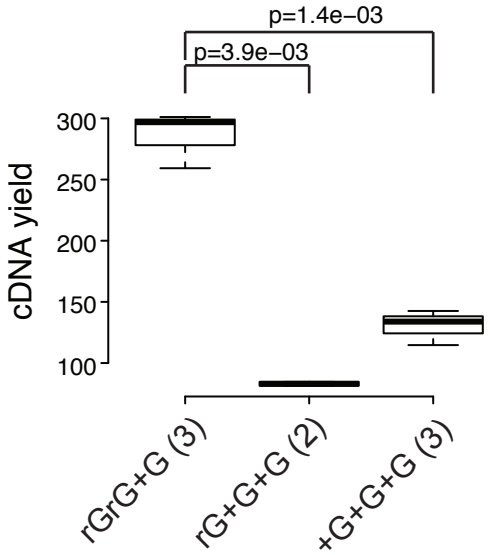
Supplementary Table 2. Listing of all TSO sequences tested

Supplementary Table 3. Listing and analyses of cDNA yield and length from individual cells

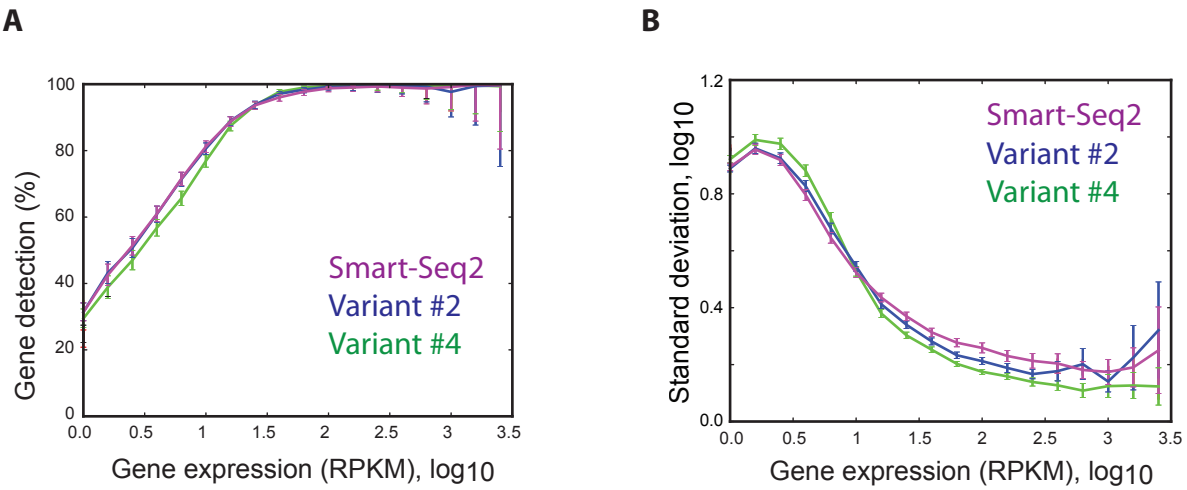
Supplementary Table 4. Variations of the Smart-seq2 protocol

Supplementary Table 5. Cost estimate of commercial Smart-seq and Smart-seq2

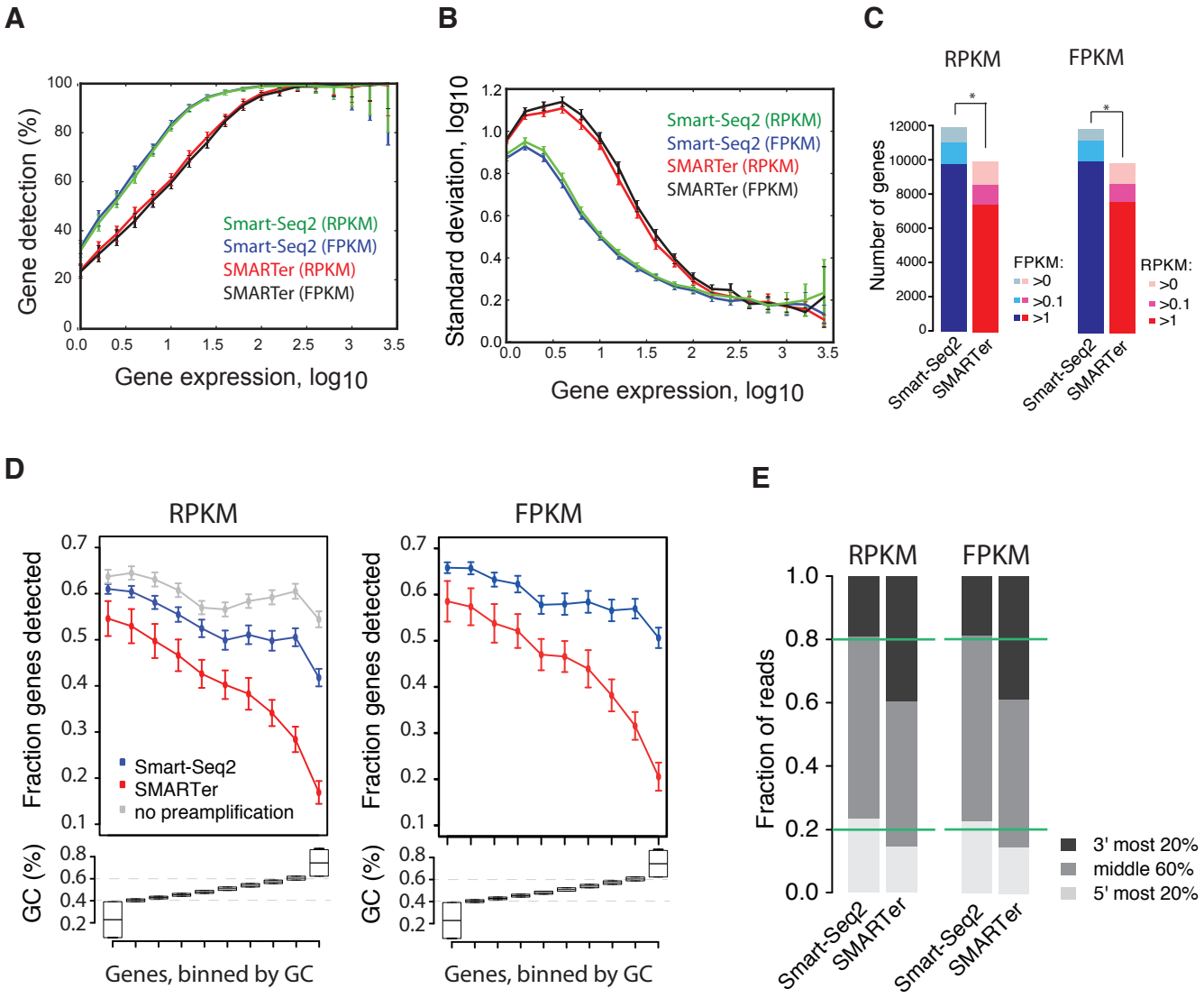
Nature Methods: doi:10.1038/nmeth.2639



Supplementary Figure 1. cDNA yield using LNA bases in template switching oligonucleotide. cDNA yields obtained with different LNA bases in the template switching oligonucleotide, where rG and +G denotes ribo- and LNA-guanylates, respectively. All reactions were performed on 1 ng of total brain RNA and cDNA yields were estimated on an Agilent BioAnalyzer.

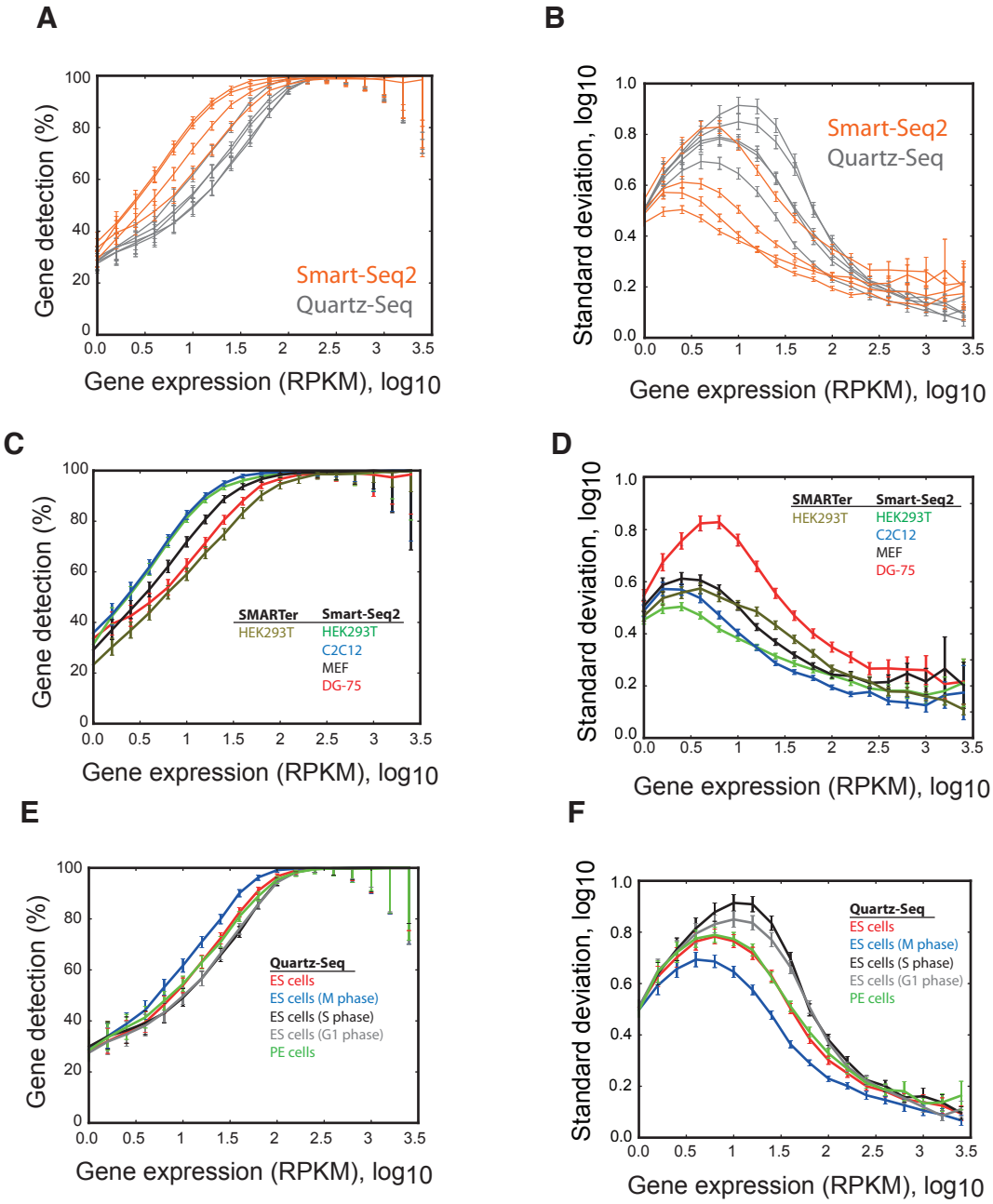


Supplementary Figure 2. Single-cell RNA-Seq sensitivity and variability.
(**A**) Percentage of genes reproducibly detected in replicate cells, binned according to expression level. We performed all pair-wise comparisons within replicates and report the mean and 90% confidence interval. Single HEK293T cell libraries were generated using variants to the template switching oligonucleotide (variant #2 uses rGrG+N and variant #4 uses rGrGrG), see Supplementary Table 4 for detailed information on protocol variations. (**B**) Standard deviation in gene expression estimates within replicates in bins of genes sorted according to expression levels. Error bars, s.e.m. (n≥4).



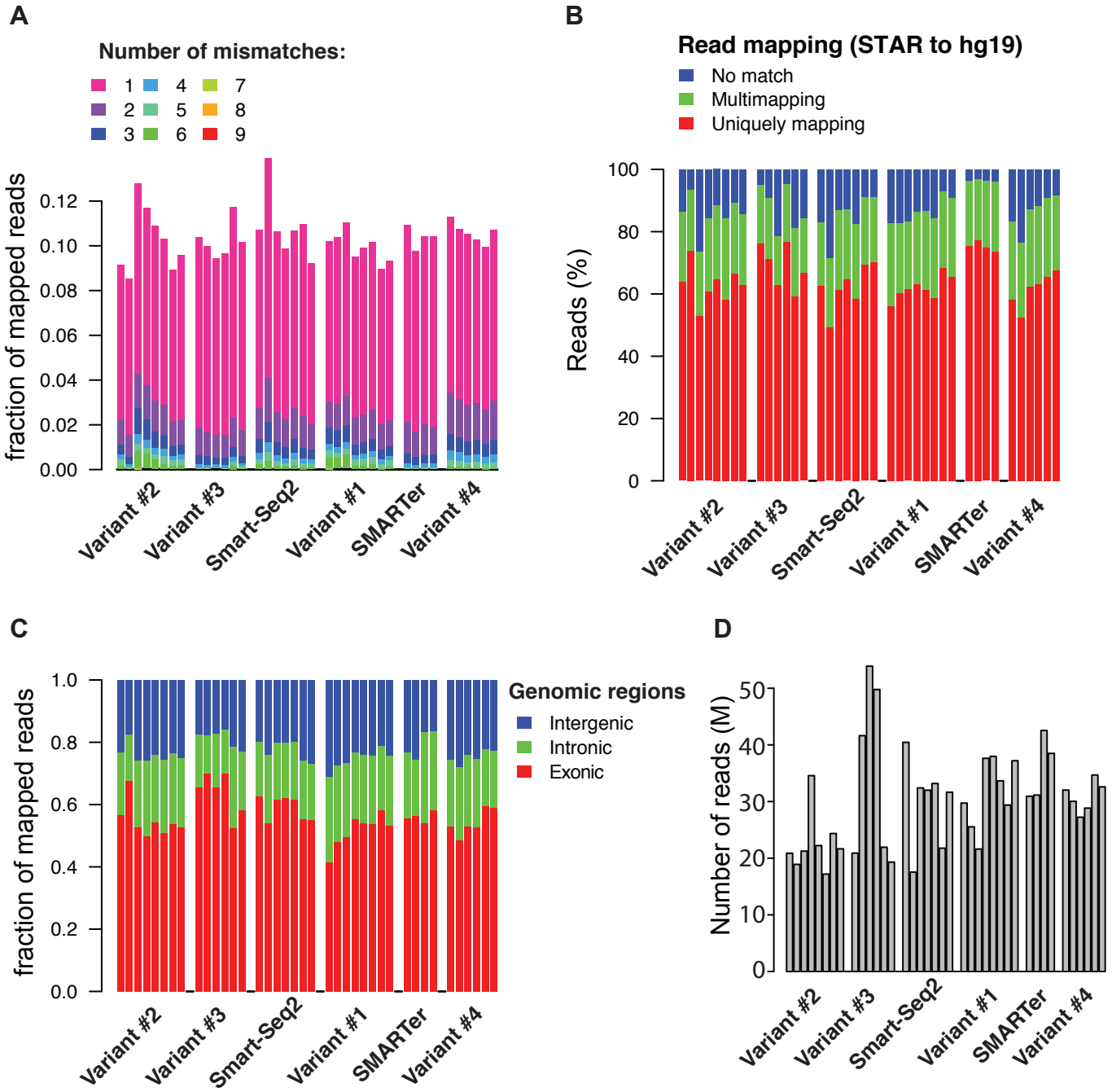
Supplementary Figure 3. Validation of single-cell RNA-Seq results using another analysis pipeline.

All single-cell sequencing libraries were reanalysed by first aligning reads with TopHat and computing expression levels (as FPKMs) using Cufflinks. We required transcripts to be longer than 500 nts as short transcripts are not reliably estimated with Cufflinks. Results were compared to expression levels estimated using rpkmforgenes (as RPKMs) on reads aligned with STAR . Very similar (or identical) results were obtained with the two strategies (RPKM for STAR/rpkmforgenes; FPKM for tophat2/cuflinks) when assessing **(A)** sensitivity as a function of expression levels for gene expression (as in Fig. 2A), (**B**) Variability in expression level estimates (as in Fig. 2B) (**C**) Gene detection at different thresholds (as in Fig. 2C), (**D**) GC content effects on gene detection (as in Fig 2D) and (**E**) in coverage across genes (as in Fig 2E).

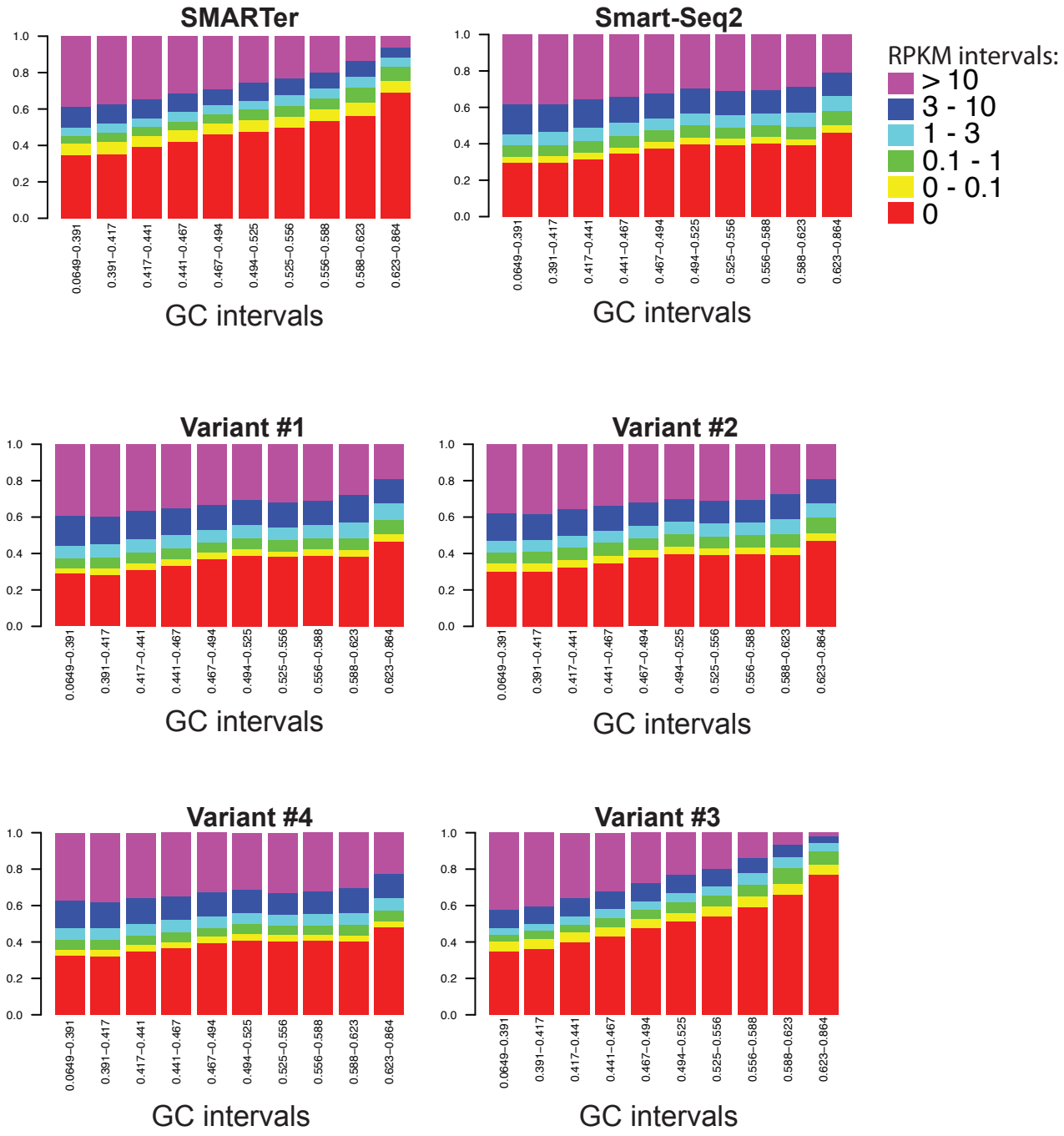


Supplementary Figure 4. Comparing single-cell transcriptomic data generated with Smart-Seq2, Quartz-Seq and SMARTer.

We compared the sensitivity and variability in single cell transcriptomic profiles generated using these protocol for all available cell types. Although different cell types were analyzed with Smart-Seq2 and Quartz-Seq, all but one cell type analyzed with Smart-Seq2 had higher sensitivity and lower technical noise than all cell types analyzed with Quartz-Seq. (**A**) Percentage of genes reproducibly detected in replicate cells, binned according to expression level. We performed all pair-wise comparisons between individual cells per cell type analyzed with Smart-Seq2 (orange) and Quartz-Seq (gray), and reported the mean and 90% confidence interval. (**B**) Standard deviation in gene expression estimates within replicates in bins of genes sorted according to expression levels. Samples as in (A) and error bars denote s.e.m. (n≥4). **(C)** Sensitivty as in (A) for cell types analyzed with Smart-Seq2 and SMARTer. (**D**) Variability in expression levels estimates as in (B) for all cell types analyzed with Smart-Seq2 and SMARTer. (**E**) Sensitivity as in (A) for all cell types analyzed with Quartz-Seq. (**F**) Variability in expression level estimates as in (B) for all cell types analyzed with Quartz-Seq.

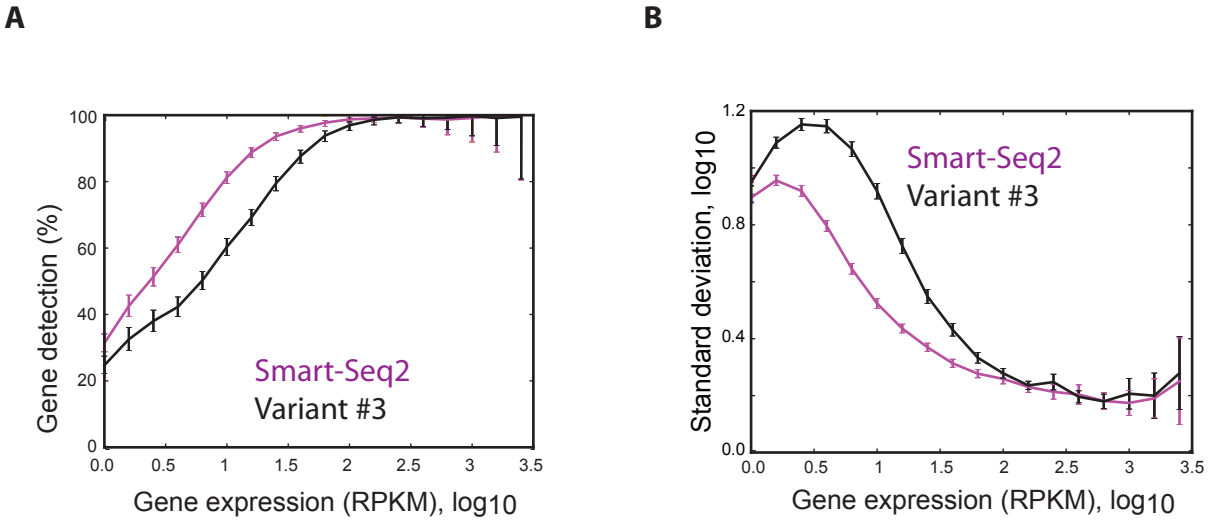


Supplementary Figure 5. Mapping statistics for single-cell libraries generated using SMARTer, optimized Smart-Seq and variants of the optimized protocol.
(**A**) The fraction of uniquely aligned reads with 1 to 9 mismatches for each single-cell RNA-Seq library. (**B**) Percentage of reads that could be aligned uniquely, aligned to multiple genomic coordinates (multimapping) or did not align for all single-cell RNA-Seq libraries. (**C**) The fraction of uniquely aligned reads that mapped to exonic, intronic or intergenic regions (annotations based on RefSeq gene models). (**D**) Number of sequenced reads per cell and library preparation protocol. (A-D) Variant protocols differ in volume of TSO used (variant #1 use 2 ul instead of 1ul), template switching oligo (variant #2 uses rGrG+N, variant #4 uses rGrGrG) or preamplification enzyme (variant #3 uses Advantage 2), see Supplementary

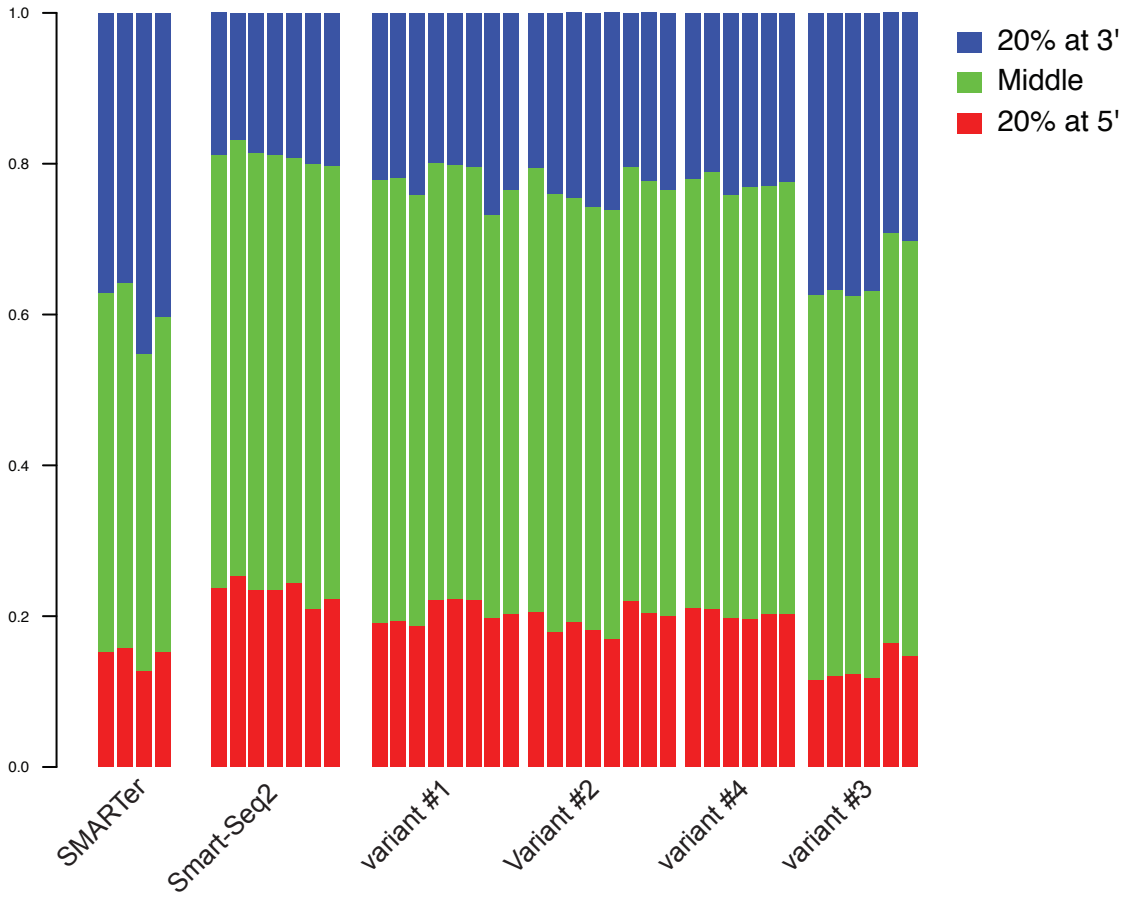


Supplementary Figure 6. Gene expression and GC levels in single-cell RNA-Seq protocols.

The fraction of genes with no reads (red) or at increasing expression level intervals, binned by the GC content of the genes. Single-cell RNA-Seq libraries were generated using Smart-Seq2, SMARTer and variants of the Smart-Seq2 protocol with differences in volume of TSO used (variant #1 use 2 ul instead of 1ul), template switching oligo (variant #2 uses rGrG+N, variant #4 uses rGrGrG) or preamplification enzyme (variant #3 uses Advantage 2), see Supplementary Table 4 for detailed information on protocol variations.

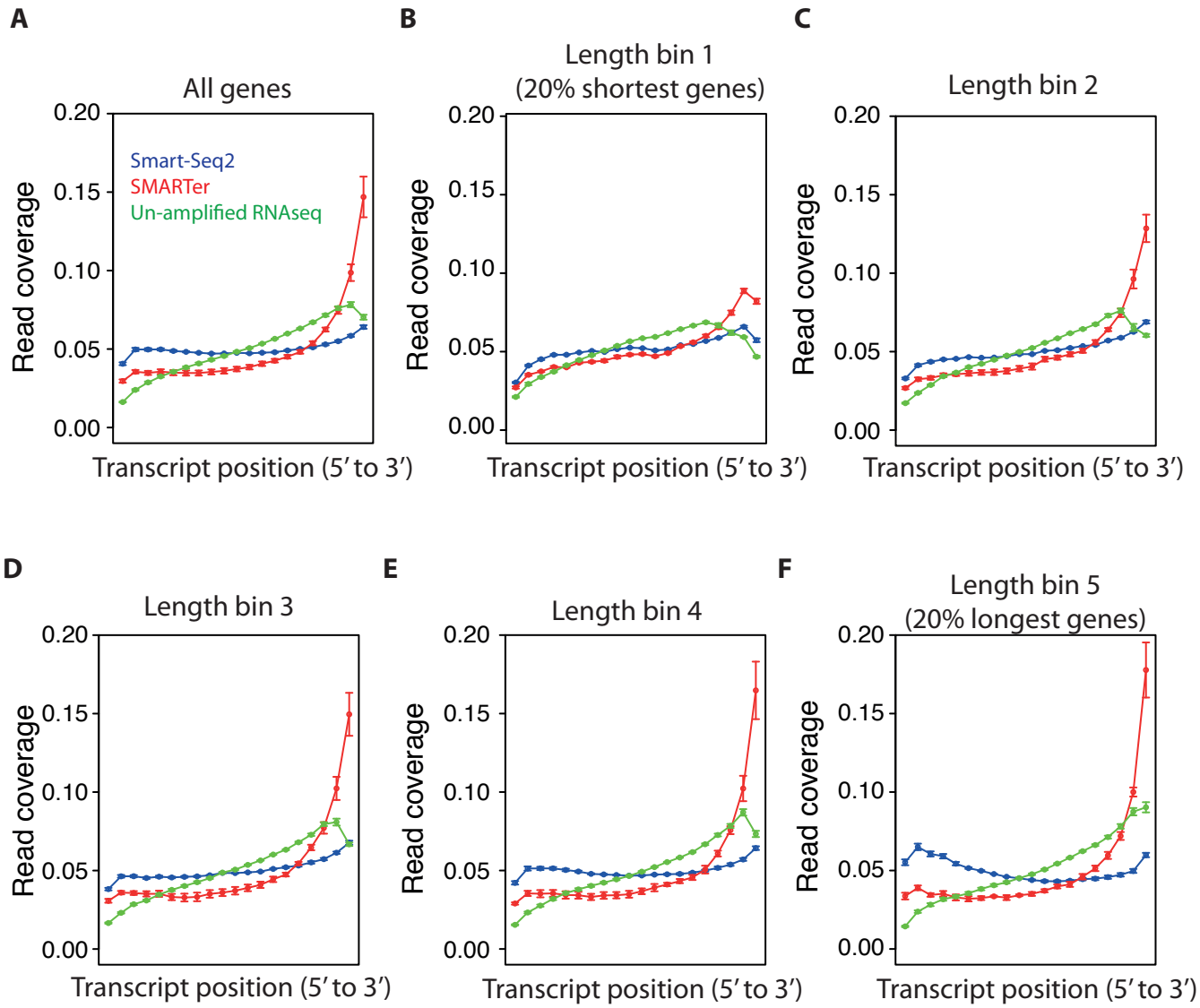


Supplementary Figure 7. Single-cell RNA-Seq sensitivity and variability.
(**A**) Percentage of genes reproducibly detected in replicate cells, binned according to expression level. We performed all pair-wise comparisons within replicates and report the mean and 90% confidence interval. (**B**) Standard deviation in gene expression estimates within replicates in bins of genes sorted according to expression levels. Single HEK293T cell libraries were generated using the Smart-Seq2 protocol and a variant of the Smart-Seq2 protocol (Variant #3 with Advantage 2 preamplification), see Supplementary Table 4 for detailed information on protocol variations. Error bars, s.e.m. (n≥4).

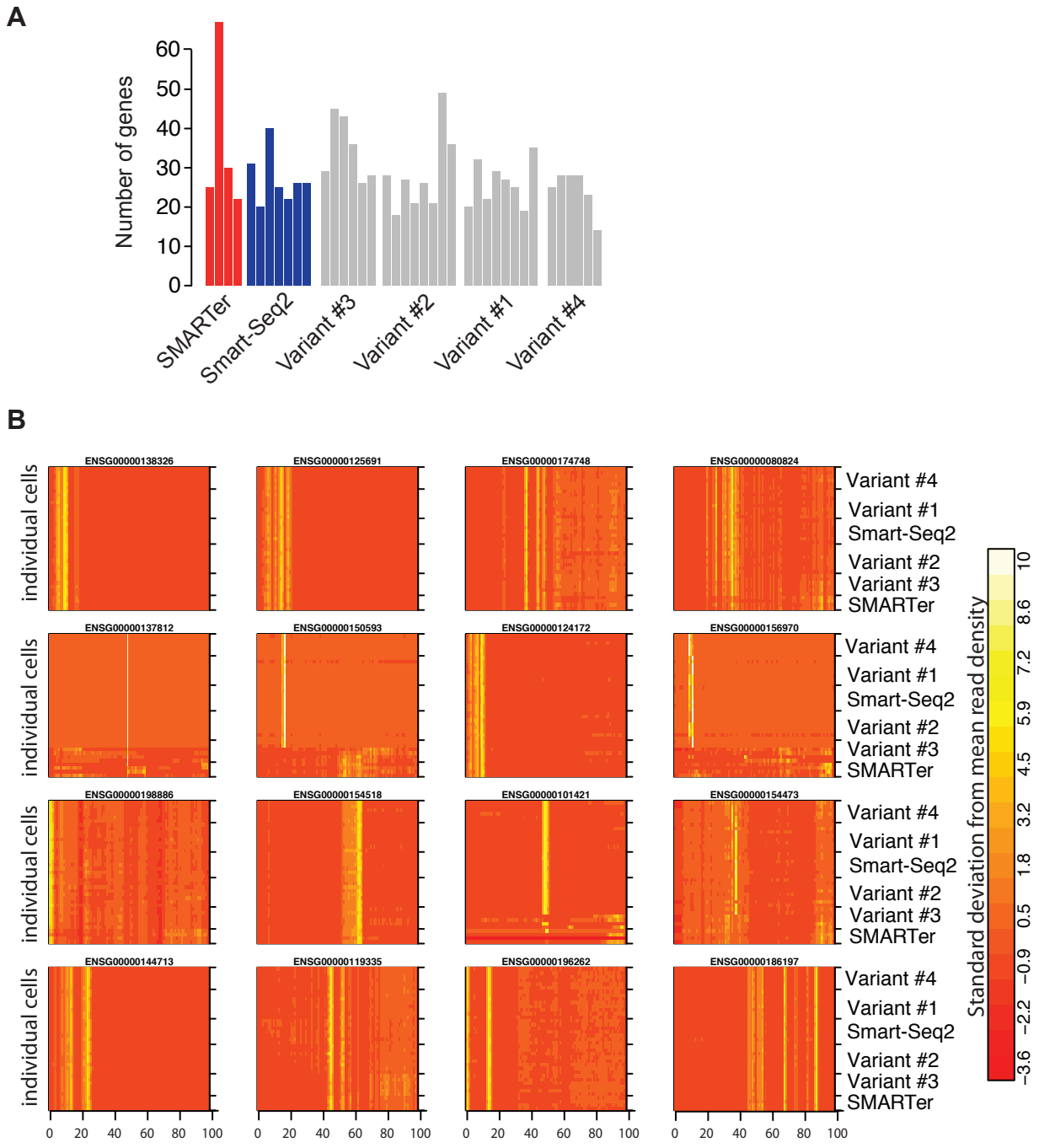


Supplementary Figure 8. Read coverage across genes in single-cell RNA-Seq data.

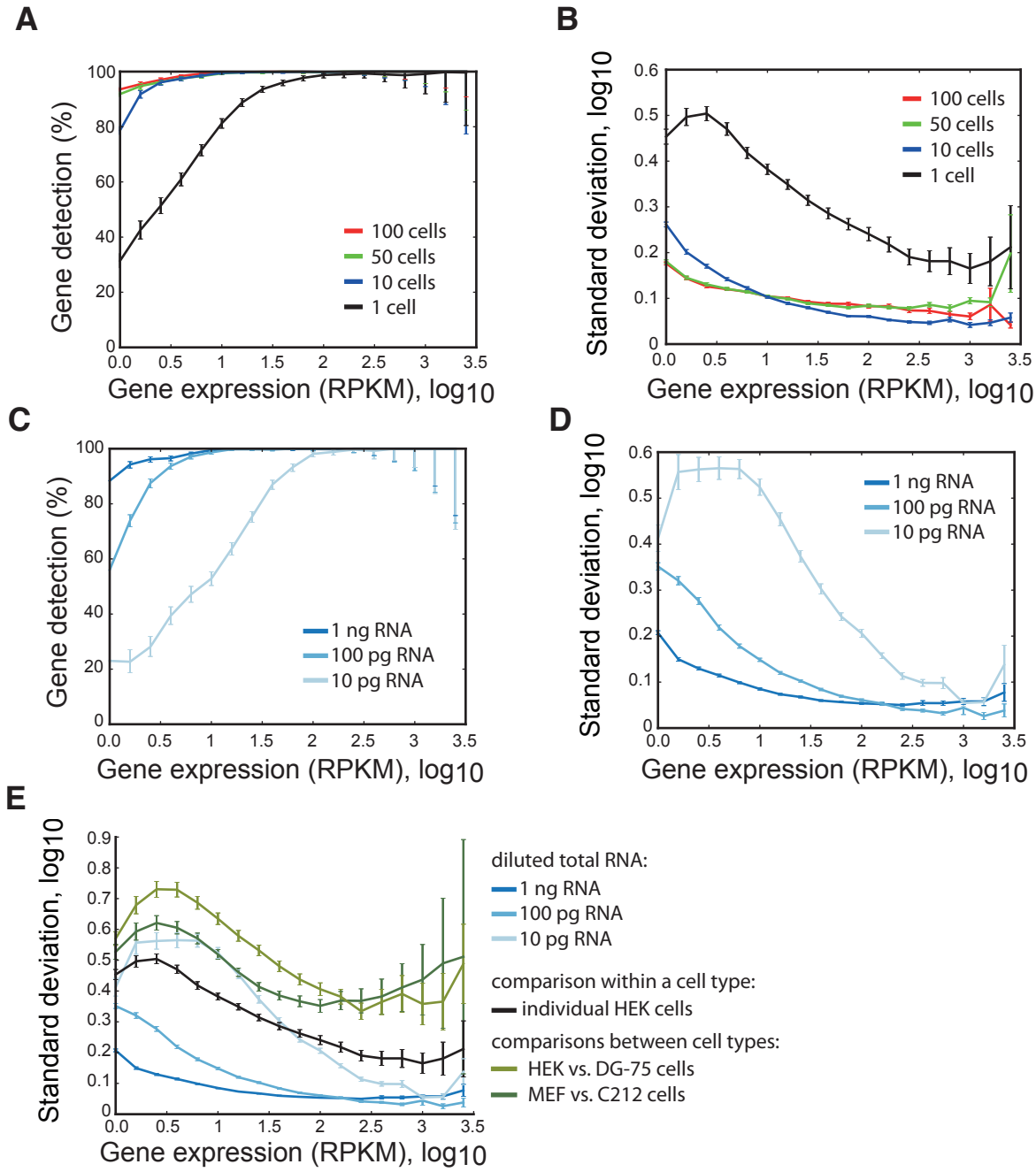
Fraction of reads mapping to the 20% 5' most, the 20% 3' most, and the 60% in the middle region for all individual single-cell transcriptome data from HEK293T cells. Variant protocols are as Smart-Seq2 except for differences in volume of TSO used (variant #1 use 2 ul instead of 1ul), template switching oligo (variant #2 uses rGrG+N, variant #4 uses rGrGrG) or preamplification enzyme (variant #3 uses Advantage 2), see Supplementary Table 4 for details on protocol variations.



Supplementary Figure 9. Read coverage across transcripts.
Single-cell RNA-Seq libraries generated with Smart-Seq2, commercial Smart-Seq (SMARTer) and unamplified RNA-Seq libraries from whole tissues were analyzed for their read coverage across genes. The longest transcript for each gene was divided into 20 bins along its length, the number of reads in each bin was normalized by number of isoforms spanning that bin, and the fraction coverage was divided by total number of reads for that transcript. The mean fraction coverage was calculated for all genes (**A**) or for transcripts grouped by their length into 5 equal sized bins (**B-F**).



Supplementary Figure 10. Read peaks in single-cell RNA-Seq data
In certain genes, we detected small regions with exceptional read densities. To systematically scan for such regions, the gene body of each gene was divided into 101 equal sized bins. Then we identified genes with one or more bin with read densities higher than 5 standard deviations over the mean across all bins. (**A**) The number of genes with one or more high density peak(s) per single-cell RNA-Seq library. (**B**) Heatmaps of read densities across the genes with peaks in the highest number of libraries (5' to 3' along x-axis), which illustrates that peaks were systematically identified in single-cell RNA-Seq data generated with SMARTer, Smart-Seq2 or variants of the latter as detailed in Supplementary Table 4. Each row represents transcriptome data from an individual cell, grouped according to the protocol used.



Supplementary Figure 11. Assessing the technical and biological variability in single-cell transcriptomics using Smart-Seq2

We assessed the technical variability in gene detection and expression level estimation through dilutions of HEK cells (A and B) and HEK total RNA (C and D) and through comparisons with biological signals in expression levels between cells of different origins (E). (**A,C**) Percentage of genes reproducibly detected in replicate dilutions, binned according to expression level, reporting the mean and 90% confidence interval. (**B,D**) Standard deviation in gene expression estimates within dilution replicates in bins of genes sorted according to expression levels. (**E**) Standard deviation in gene expression estimates within total RNA dilution replicates in bins of genes sorted according to expression levels (as in D) with additional pair-wise comparisons of cells to estimate biological variability present in single-cell transcriptome data. Pair-wise comparisons of individual cells of different origins (HEK vs. DG-75 and MEF vs. C2C12) demonstrates the magnitude of biological variability associated with differences in cell types, whereas pair-wise comparisons of individual HEK cells demonstrates the magnitude of biological variability within a homogenous cell population. All error bars denote s.e.m (n≥4).