Supplementary Materials

Supplementary Methods

Except where otherwise indicated all analyses were carried out using custom-written Python scipts.

Short-read RNA-seq data analysis

The BAM file for the K562 Caltech PolyA+ Rep1 RNA-seq dataset was downloaded from `http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/`. The indicated number of mapped sequencing fragments were randomly subsampled from the BAM files. Subsequently, gene expression values were calculated for each number of fragments using Cufflinks [1] (version 2.0.2). For a given gene g, each subsampled value $\text{FPKM}_{g,ss}$ was considered within 10% of its final value FPKM_g if $|\text{FPKM}_{g,ss} - \text{FPKM}_g|/\text{FPKM}_g \leq 0.1$.

Direct RNA-seq simulation

The simulation of direct RNA-seq quantifications was carried out as follows. Without loss of generality, the FPKM values from Cufflinks quantifications were used as a starting point, i.e. the relative average transcripts abundances per cell were taken to be the relative abundances of genes in the set of FPKM values. Thus, $\text{TPM}_g = (\text{FPKM}_g / \sum_{g \in G} FPKM_g) \cdot 10^6$. The expected number of original transcripts for each gene/isoform was then calculated from the number of cells N and the number of transcripts per cell T_{cell}: $T_g = \text{TPM}_g \cdot N \cdot T_{\text{cell}} \cdot 10^{-6}$. The number was stochastically rounded to the next higher integer with a probability $p = T - \lfloor T \rfloor$.

Finally, the sequencing process was simulated as described above, with each transcript successfully passing the library conversion or sequencing steps with probabilities p_{lib} and p_{seq}, respectively, the product of which is the total single molecule capture probability p_{smc}.

Formally we have:

Algorithm 1 Direct RNA-seq simulation

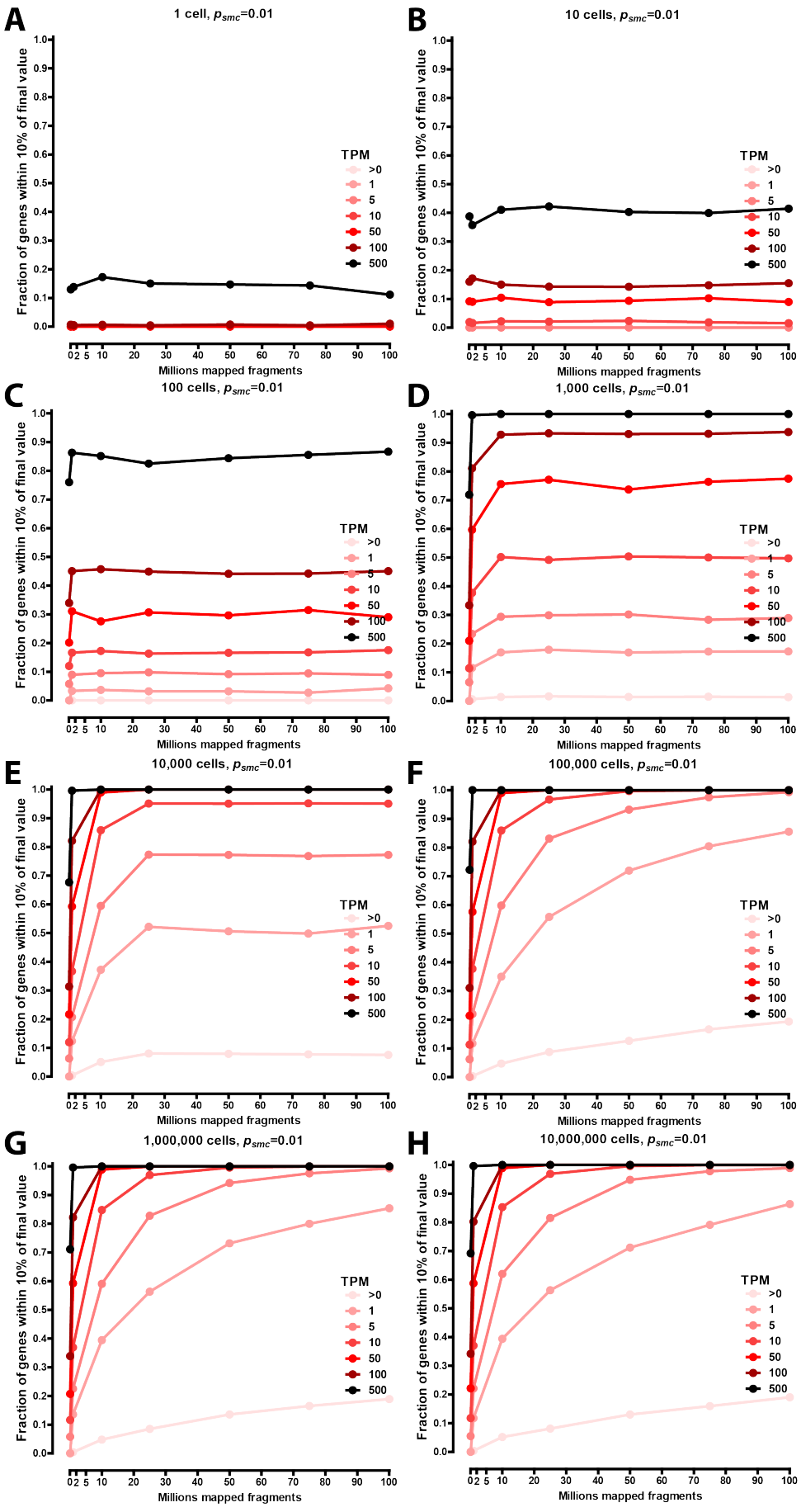
for $g \in G$ **do**
  $\text{TPM}_g \leftarrow (\text{FPKM}_g / \sum_{g \in G} FPKM_g) \cdot 10^6$
  $|T_g|_{original} \leftarrow \text{TPM}_g \cdot T_{cell} \cdot N \cdot 10^{-6}$
  $p \leftarrow$ random number $\in [0, 1]$
  if $p \leq T - \lfloor T \rfloor$ **then**
    $T = \lceil T \rceil$
  else
    $T = \lfloor T \rfloor$
  end if
  $|T_g|_{library} \leftarrow 0$
  $i \leftarrow 0$
  while $i \leq |T_g|_{original}$ **do**
    $i \leftarrow i + 1$
    $p \leftarrow$ random number $\in [0, 1]$
    if $p \leq p_{lib}$ **then**
      $|T_g|_{library} \leftarrow |T_g|_{library} + 1$
    end if
  end while
  $|T_g|_{sequenceable} \leftarrow 0$
  $i \leftarrow 0$
  while $i \leq |T_g|_{library}$ **do**
    $i \leftarrow i + 1$
    $p \leftarrow$ random number $\in [0, 1]$
    if $p \leq p_{seq}$ **then**
      $|T_g|_{sequenceable} \leftarrow |T_g|_{sequenceable} + 1$
    end if
  end while
end for
if $R > \sum_{g \in G} |T_g|_{sequenceable}$ **then**
  $p_{sampling} \leftarrow R / \sum_{g \in G} |T_g|_{sequenceable}$
else
  $p_{sampling} \leftarrow 1$
end if
for $g \in G$ **do**
  $|R_g| \leftarrow 0$
  $i \leftarrow 0$
  while $i \leq |T_g|_{sequenceable}$ **do**
    $i \leftarrow i + 1$
    $p \leftarrow$ random number $\in [0, 1]$
    if $p \leq p_{sampling}$ **then**
      $|R_g| \leftarrow |R_g| + 1$
    end if
  end while
end for

Where:

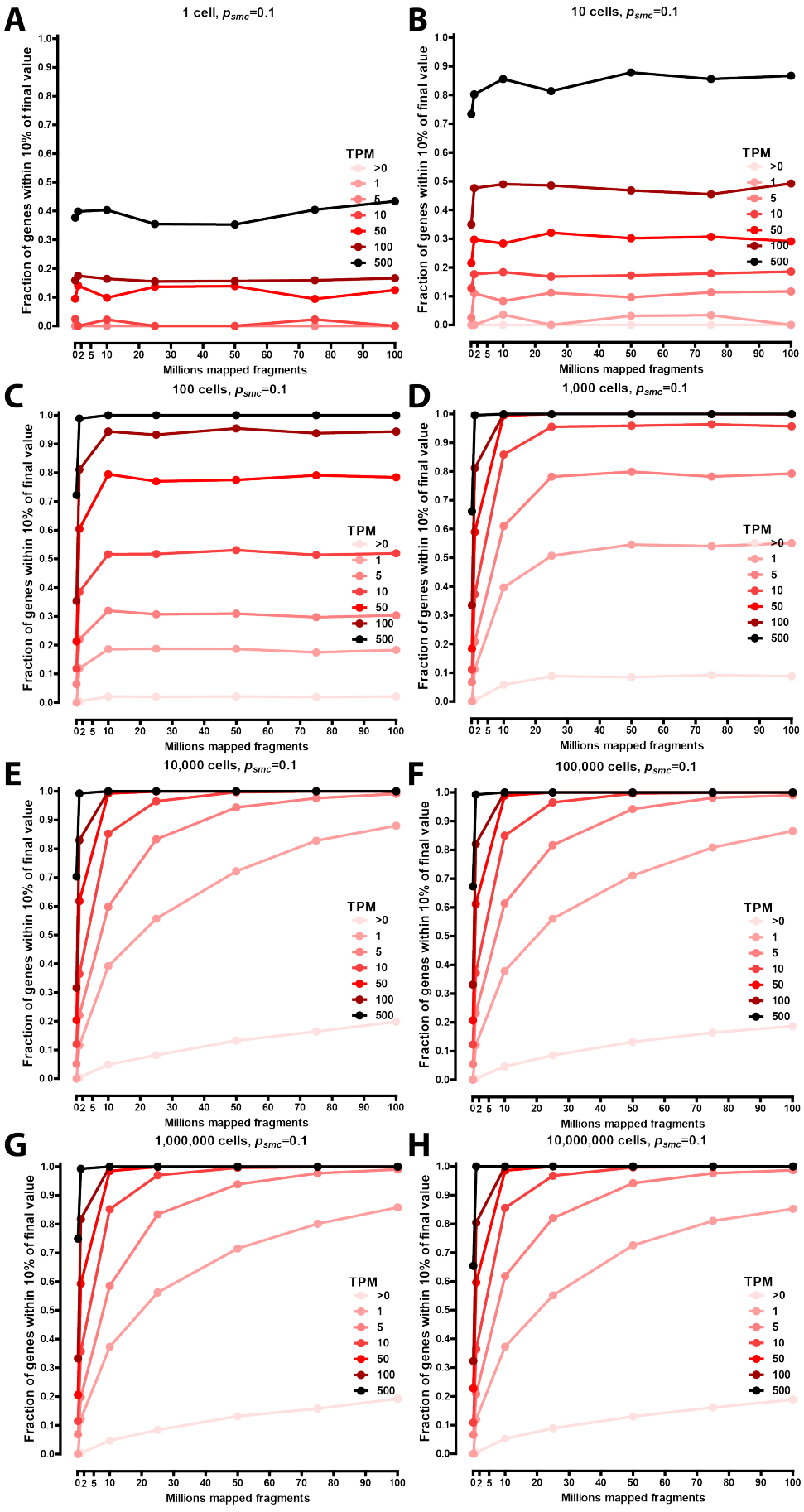
- FPKM_g: FPKM values for each gene g in the set of all genes G
- N: number of cells
- T_{cell}: average number of transcripts per cell
- p_{lib}: probability of successful library conversion for any given individual RNA molecule
- p_{seq}: probability of successful sequencing for any given RNA molecule successfully converted into the library
- R: the number of sequencing reads

TPM values were calculated for each gene/isoform based on the abundances of transcripts in the resulting dataset. As before, the fraction of genes/transcripts for each $|\text{TPM}_{g,ss} - \text{TPM}_g|/\text{TPM}_g \leq 0.1$ was plotted.

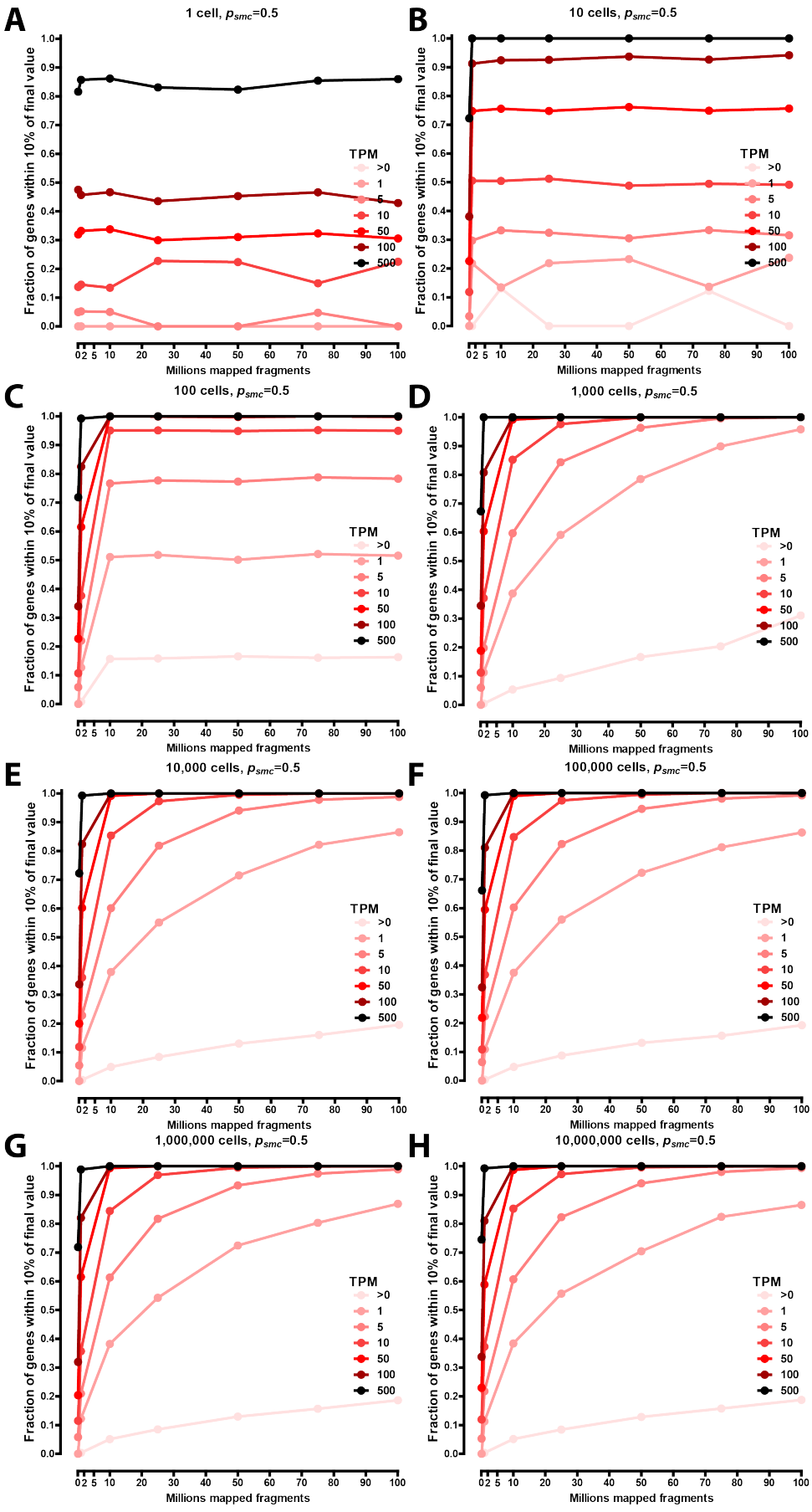
Supplementary Figures



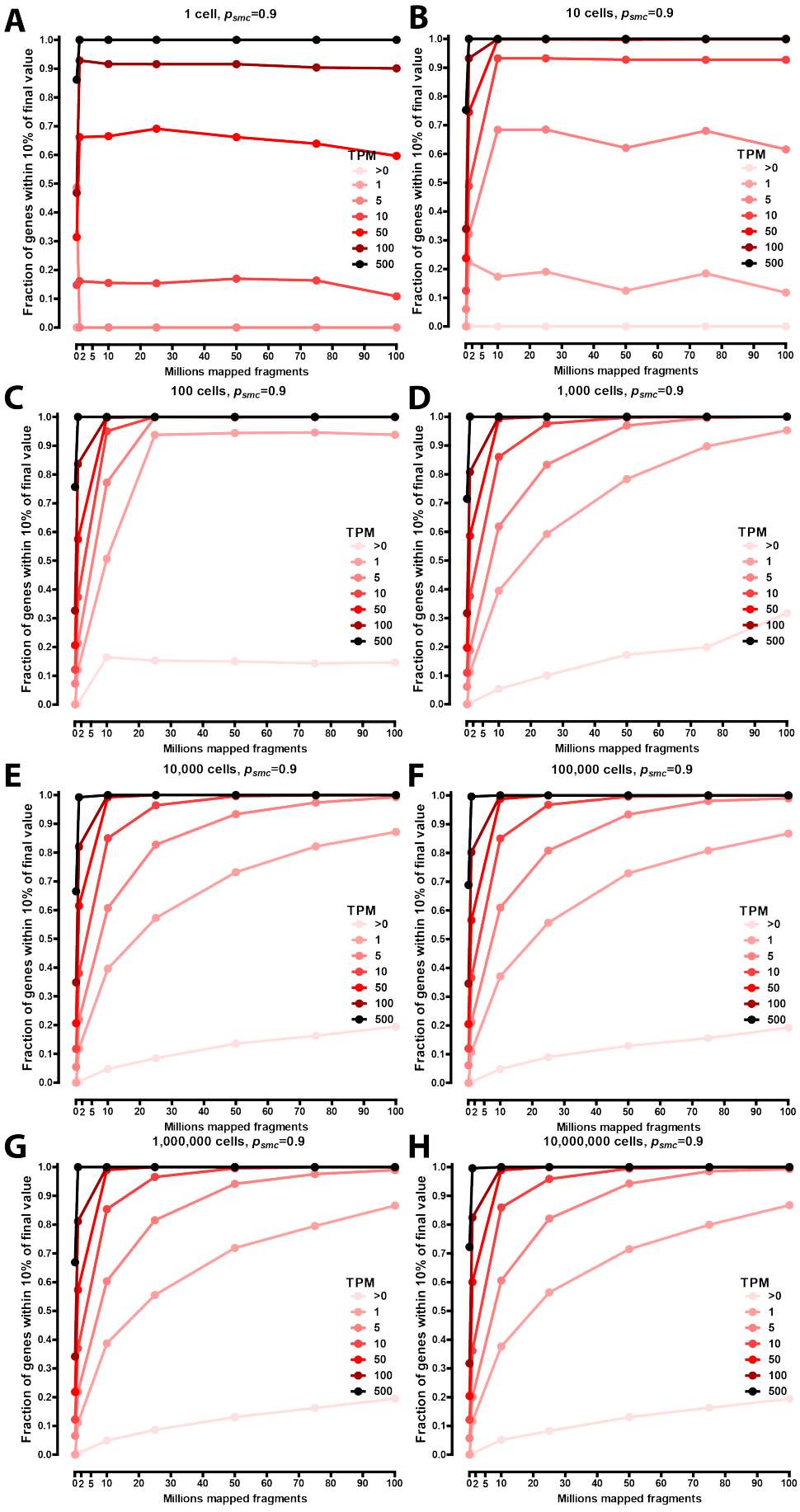
Supplementary Figure 1: Accuracy of direct RNA-seq quantification at the gene level as a function of input cell number at $p_{smc} = 0.01$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



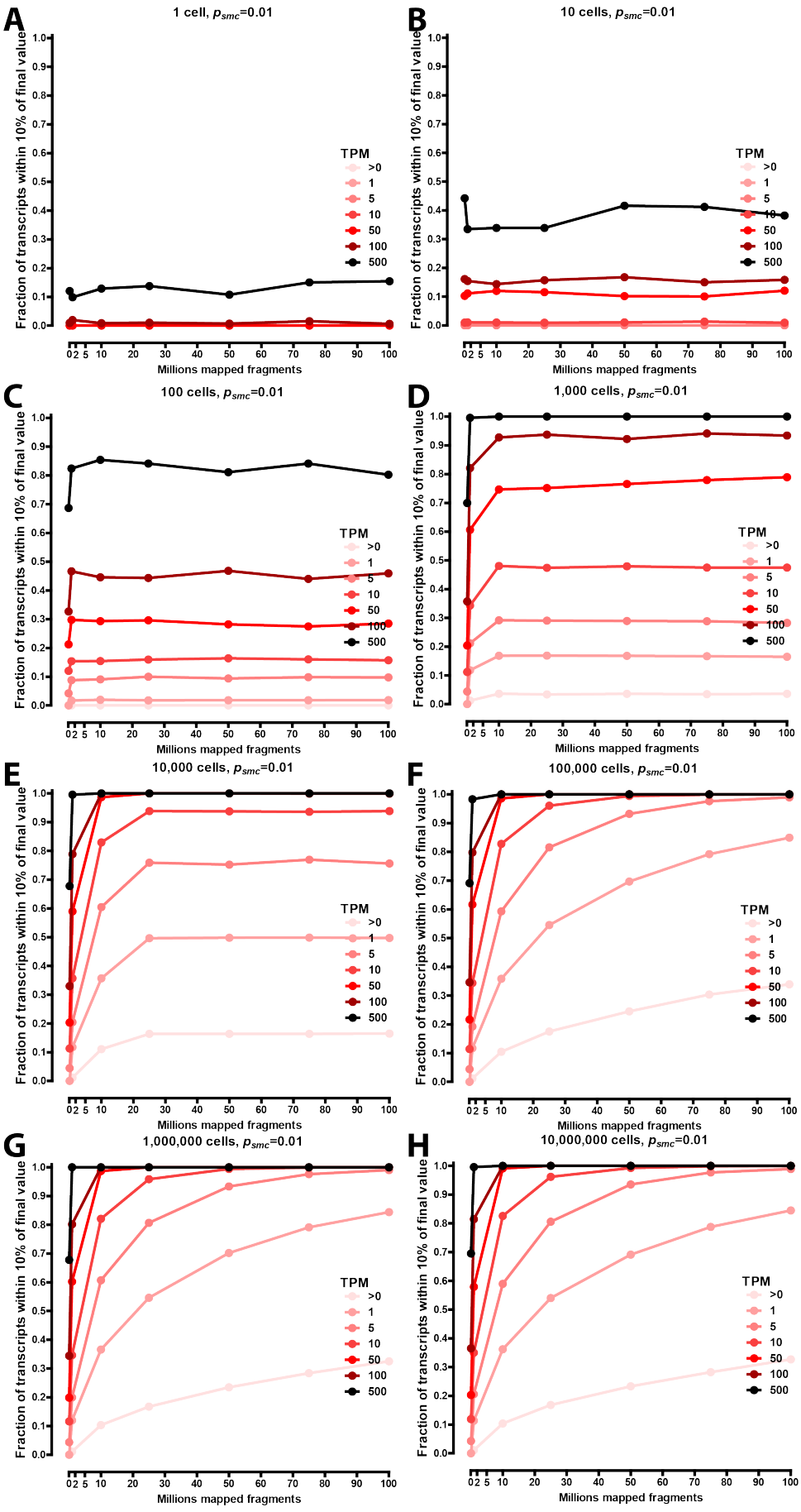
Supplementary Figure 2: Accuracy of direct RNA-seq quantification at the gene level as a function of input cell number at $p_{smc} = 0.1$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



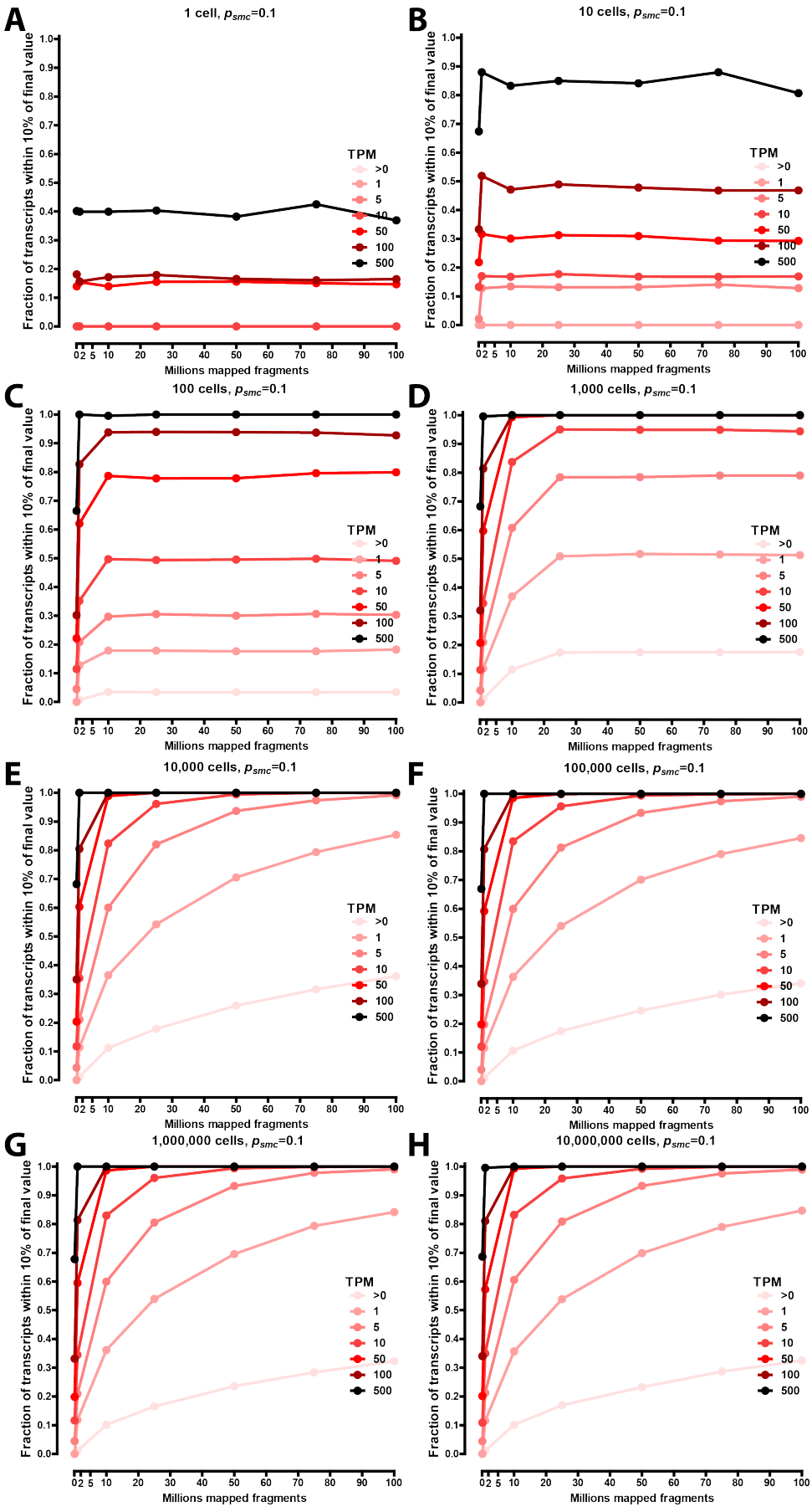
Supplementary Figure 3: Accuracy of direct RNA-seq quantification at the gene level as a function of input cell number at $p_{smc} = 0.5$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



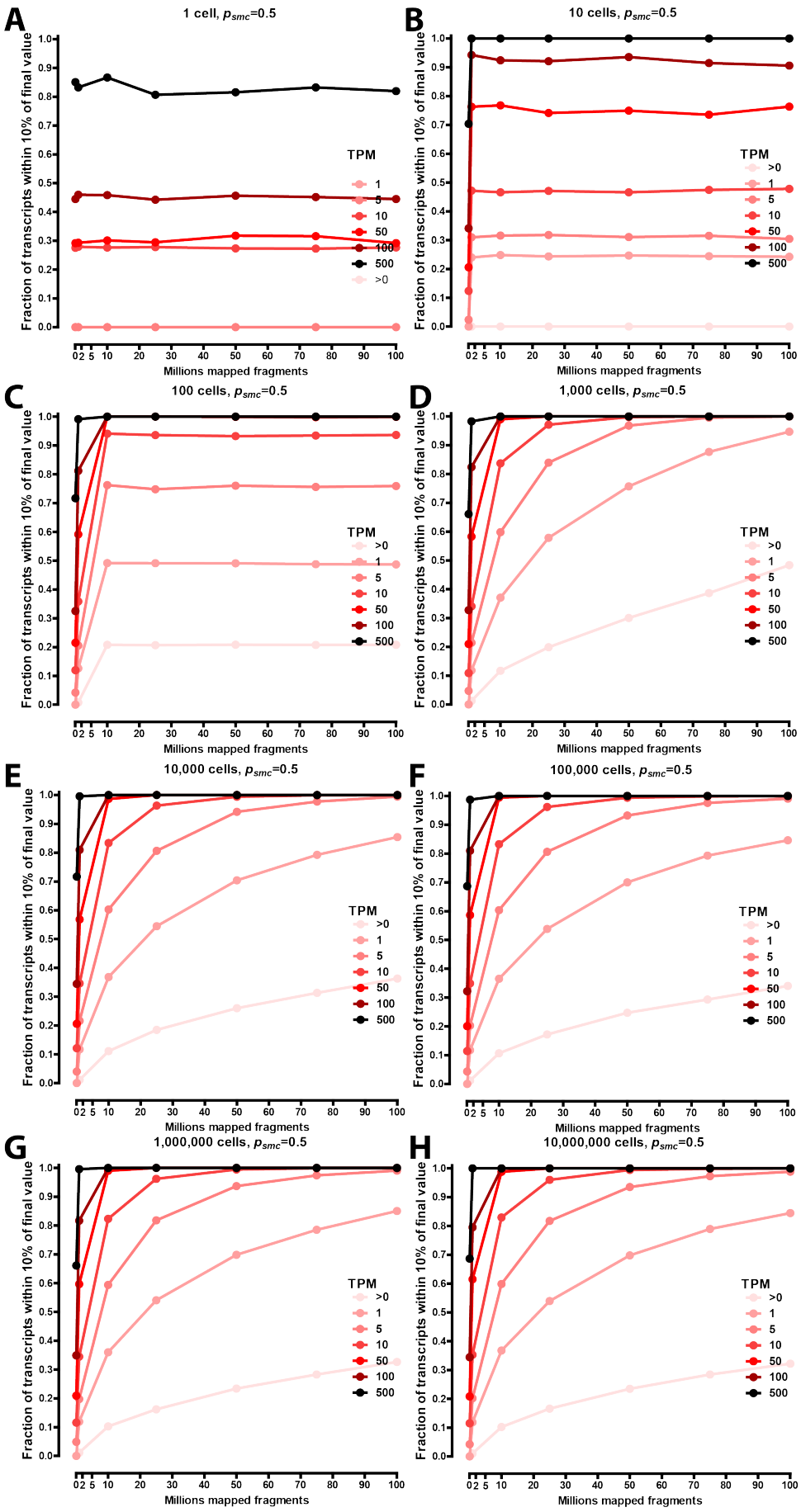
Supplementary Figure 4: Accuracy of direct RNA-seq quantification at the gene level as a function of input cell number at $p_{smc} = 0.9$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



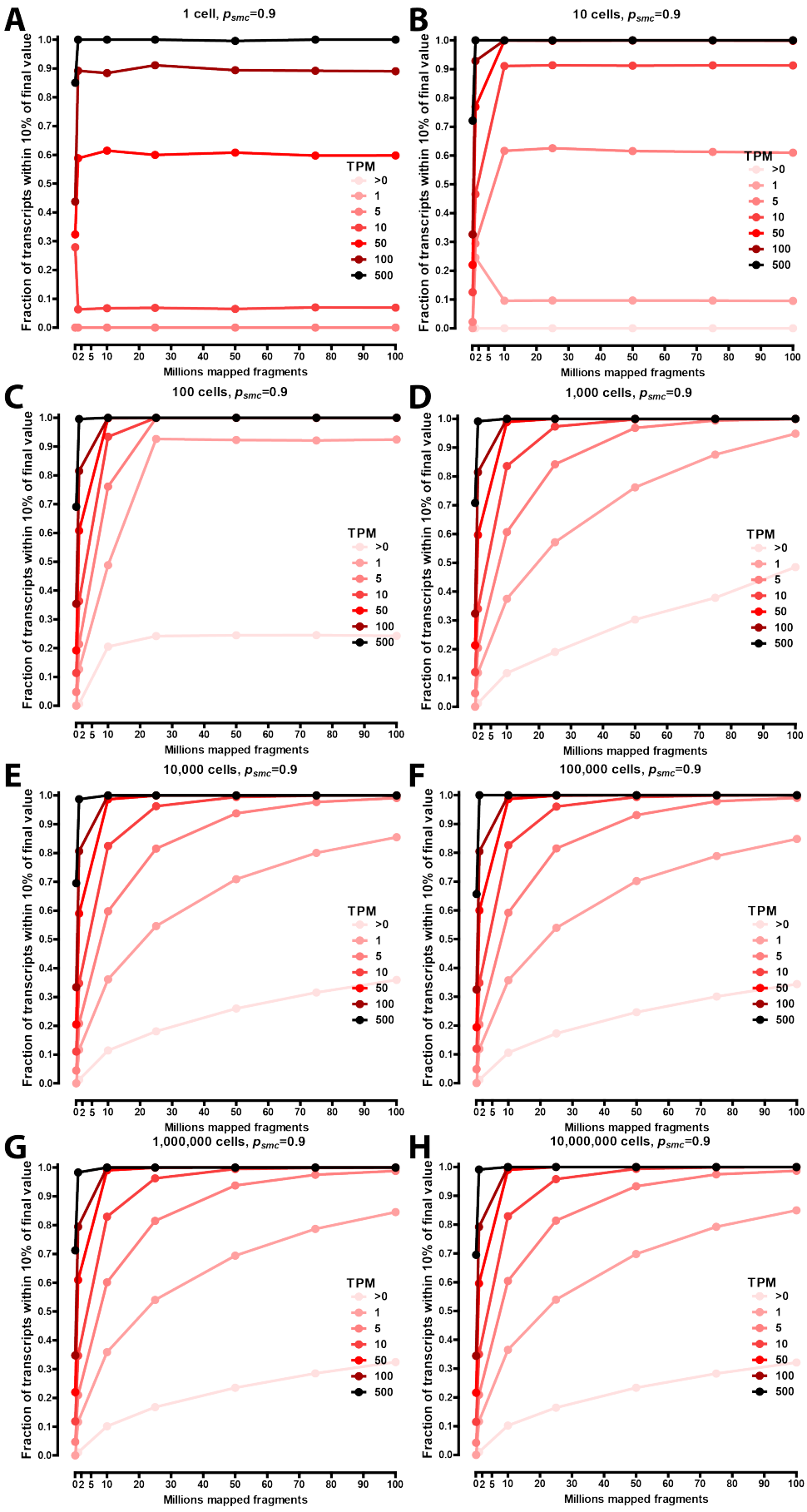
Supplementary Figure 5: Accuracy of direct RNA-seq quantification at the transcript level as a function of input cell number at $p_{smc} = 0.01$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



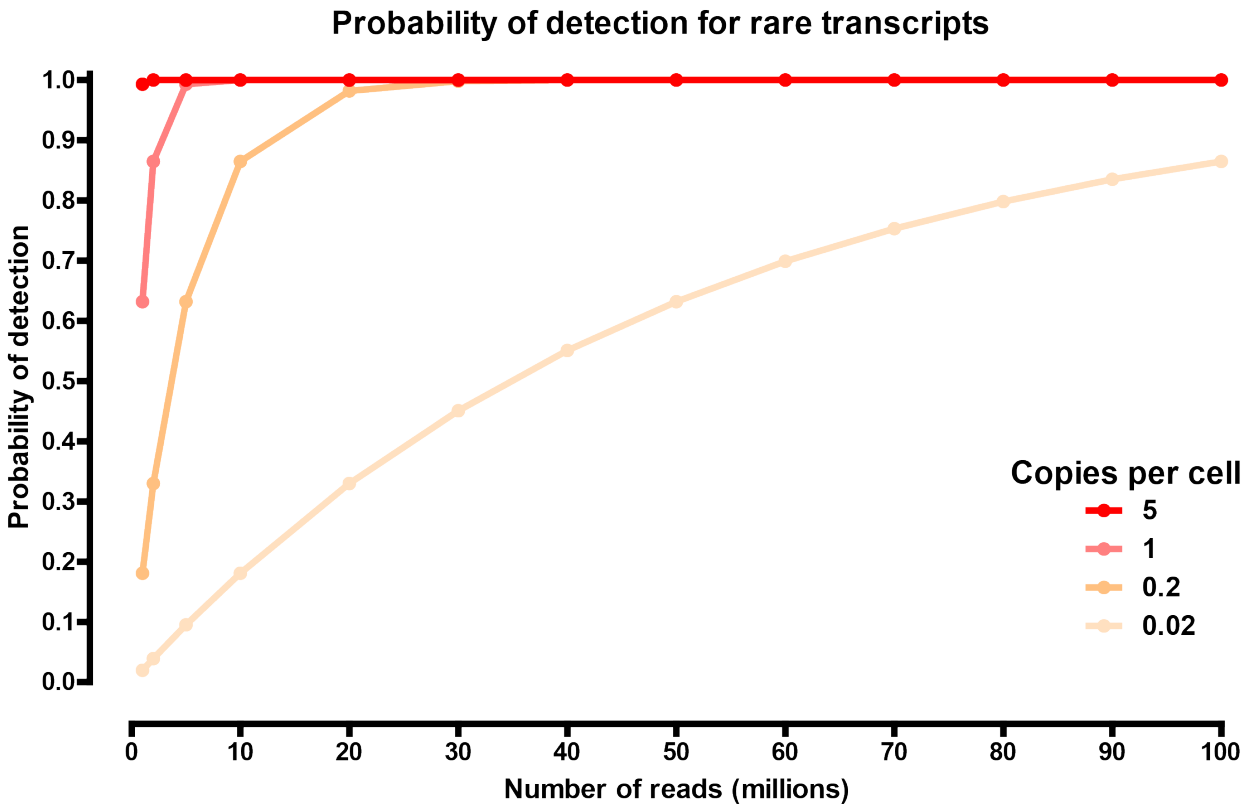
Supplementary Figure 6: Accuracy of direct RNA-seq quantification at the transcript level as a function of input cell number at $p_{smc} = 0.1$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



Supplementary Figure 7: Accuracy of direct RNA-seq quantification at the transcript level as a function of input cell number at $p_{smc} = 0.5$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



Supplementary Figure 8: Accuracy of direct RNA-seq quantification at the transcript level as a function of input cell number at $p_{smc} = 0.9$. Gene expression in human K562 cells was simulated using the GENCODE V16 annotation as described in the Methods section. (A) 1 cell; (B) 10 cells; (C) 100 cells; (D) 1000 cells; (E) 10,000 cells; (F) 100,000 cells; (G) 1,000,000 cells; (H) 10,000,000 cells. Note that because of the absence of amplification, the total number of reads cannot be higher than $\sim T_{\text{cell}} \times N \times p_{smc}$ (the number of transcripts per cell times the number of cells times the probability of capture and sequencing), a value exceeded for a number of the conditions simulated. Also note that a slightly different original transcriptome and a different sequencing process were simulated for each sequencing run, which introduces some stochasticity in the curves at low cell numbers.



Supplementary Figure 9: Detection of rare transcripts as a function of the sequencing depth of direct RNA-seq. The probability of detection given sufficiently many input cells and a sufficiently high p_{smc} is given by $1 - (1 - C_c/T_{all})^R$, where R is the number of reads, C_c is the average number of copies per cell for the transcript of interest, and T_{all} is the total number of transcripts (coding and non-coding, excluding the fraction of rRNAs and tRNAs that is removed prior to sequencing) in the cell. For simplicity, a $T_{all} = 10^6$ was used here (the number is not well constrained by existing data); note that an increase in T_{all} is effectively equivalent to a decrease in C_c.

References

1. Trapnell C, Williams BA, Pertea G, Mortazavi A, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**(5):511–515.