From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing Supplementary Materials

Georgi K. Marinov[1,*], Brian A. Williams[1,*], Ken McCue[1], Gary P. Schroth[2], Jason Gertz[3], Richard M. Myers[3], and Barbara J. Wold[1]

[1]*Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA*
[2]*Illumina, Inc., Hayward, CA 94545, USA*
[3]*HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA*
[*]*These authors contributed equally to this work*

Supplementary Methods

Except where specified otherwise, all computational processing was carried out using custom-written Python scripts.

Single cell collection

Single cell harvesting from live cultures requires a micropipet with a polished glass tip with an approximate diameter of 40μm. Borosilicate glass microfiber pipettes (FHC omega dot fiber 30-30-0) were pulled on a Sutter Instruments P80/PC microcapillary puller with the following parameters: 750 heat, 150 pull, 100 velocity, 5 time. After pulling, the microcapillary tips were mounted on a glass microscope slide using modeling clay, and broken by closing a pair of #5 Dumont forceps around the glass. We used a scaled eyepiece reticle to judge the width of the break at about 40μm. After breaking, the tips were smoothed using a microforge. To prevent sticking of cells to the interior of the capillary, we treated the pipettes with Sigmacote by attaching Tygon tubing and a syringe to the blunt end of the microcapillary, and drawing the Sigmacote solution into the tip. This also provided assurance that the forged tips had not closed. The capillaries were then rinsed with distilled water twice using the same technique, and allowed to dry at room temperature overnight.

An aliquot ($5x10^6$ cells) of GM12878 cells were thawed rapidly and cultured in 10mL of medium (RPMI 1640, 15% FBS, 2mM L-glutamine, 1% penicillin-streptomycin). The cells were grown at density of $2x10^5 - 2x10^6$ cells/mL of medium for 11 days until harvest. On the day prior to harvest, the culture volume was increased to 100mL by the addition of fresh medium, bringing the density to $2x10^5$ cells/mL. At harvest time (23 hours later), cells were triturated using a 10mL pipette, and a small aliquot ($\sim$100μLs) of the culture was removed. A few μLs of the cell suspension was added to a 250μL volume of "cell picking medium" (RPMI1640 with 15% Superblock (Pierce catalog #37515) and 2mM glutamine). This diluted cell suspension was then placed in a 3cm culture dish and returned to the 37 °C incubator for 10 minutes prior to single cell harvesting.

The microcapillary pipet was mounted on a micromanipulator and attached to a 100μL glass syringe via Tygon tubing. A dish of picking medium was brought to the illuminated stage on the phase contrast scope, and the tip was submerged using the micromanipulator. Picking medium was drawn up into the microcapillary to a height of about 75mm. The tip was removed from the picking medium, resubmerged into a dish from the incubator containing the dilute cell suspension, and lowered gently to the floor of the dish. Individual cells were aspirated into the pipet by gentle vacuum applied via the glass syringe. When a single cell had been aspirated, the tip was rapidly lifted out of the picking medium, and the picking dish was removed from the illuminated area of the stage. A small sliver of silanated cover glass (Molecular Dimensions, catalog #MD406) was then placed on a glass slide on the stage, and a 4.5μL drop of cell lysis solution was placed on the sliver with a Rainin P10 micropipette. The lysis solution contains 2.5μL of reaction buffer (Clontech SMARTer Ultra Low RNA kit), 1μL of 3 SMART CDS Primer IIA (Clontech) and 1μL of spike-in quantification standards. The drop of lysis solution was visualized on the illuminated area of the stage, and the pipette tip containing the picked cell was lowered into it. Gentle pressure was applied to the syringe to ex-

Corresponding author: Barbara J. Wold. Biology Division, California Institute of Technology. Pasadena, CA 91125.

pel the cell from the pipette, and the tip was then lifted from the lysis solution. Visual confirmation was made at high power, while the cell dissolved in the lysis solution. The glass sliver was lifted from the stage using forceps, and placed in the bottom of a 200μL PCR tube. The tube was spun for 15 seconds at 10,000g, the sliver was removed, and the lysed cell was immediately frozen on dry ice. Twenty individual cells were collected in this way. We also collected two samples of ten cell pools into the same volume of lysis buffer, using the pipette picking method.

For $\sim$100 cell pools, cells were first diluted in picking buffer to a concentration of 10 cells/μL. 10μL of the dilute cell suspension were added to 90μL of picking buffer in a 200μL PCR tube, and spun at 2500g for 90 seconds to pellet the cells. The tube was then mounted sideways in modeling clay on a glass slide, and the pellet was visualized under the phase contrast scope. A drawn glass pipette tip attached to the micromanipulator was advanced into the picking medium and the excess picking medium was withdrawn using the syringe. A 4.5μL aliquot of lysis buffer was then added to the cell pellet, and the lysate was spun and frozen as for the above samples.

After picking the individual and pooled cell samples, the remainder of the culture ($\sim 2x10^7$ cells) was spun down in two aliquots for 5 minutes at 1000g at 4 °C. The culture medium was removed, the pellet was rinsed with PBS, and re-spun as above. After the removal of PBS, both pellets were lysed with 1.2mL lysis buffer from the Ambion mirVana kit (catalog #AM1560). The lysates were then processed according to the manufacturer's protocol. After eluting total RNA from the columns, we performed a DNA digestion step to remove residual contaminating genomic DNA, using the DNA-free kit from Ambion (catalog #AM1907). After quality control with Qubit and the Agilent BioAnalyzer, the bulk prep total RNA was diluted to both 10ng/μL and 100pg/μL concentrations. We then added single microliter aliquots to the lysis buffer described above, and froze the samples for processing using the single cell protocol.

First strand cDNA synthesis and amplification

The frozen samples were brought to the lab bench on dry ice. Lysis and denaturation were accomplished by heating the samples for 3 minutes at 72 °C. The samples were spun down and placed in a cooling rack at 4 °C. 5.5μL of first strand reaction buffer (Clontech) was then added (2μL of buffer, 1μl of RNAse inhibitor, 1μl of dNTPs, 0.25μl of DTT, 1μL of SMARTer IIa oligos, and 1μl of SMARTScribe reverse transcriptase). The samples were reverse transcribed at 42 °C for 90 minutes and denatured at 70 °C for 10 minutes. After denaturation, the samples were spun down and 25μL of Ampure XP SPRI beads (Beckman Coulter genomics) were added. The samples were incubated for 8 minutes at room temperature, then the beads were separated on a magnet for 5 minutes. The supernatant solution was removed with a pipette, and the beads were spun at 1000g for 1 minute to pellet. The sample was then placed back on the magnet, and excess supernatant was removed with a 10μL Rainin pipet tip. 50μL of amplification solution were then used to resuspend the beads (5μL of PCR buffer, 2μL of dNTPs, 2μL of amplification primers and 2μL of Advantage2 polymerase mix), and the samples were amplified under the following conditions: 1 minute at 95 °C, followed by cycles of 15 seconds at 95 °C, 30 seconds at 65 °C, 6 minutes at 68 °C, and final elongation for 10 minutes at 72 °C. Single cell and pool/split samples were amplified for 26 cycles, the 10 cell pools were amplified for 22 cycles, the 100 cell pools were amplified for 18 cycles, and the bulk prep RNA samples were amplified for 15 cycles. The amplified cDNA was spun down, and 90μL of Ampure XP beads were added. The beads were incubated with the amplified product for 8 minutes, then separated on a magnet for 5 minutes. The reaction solution was removed and the beads were washed twice with 200μL of freshly prepared 80% ethanol for 30 seconds. After the second ethanol wash, the beads were pelleted at 1000g for 1 minute, the residual ethanol was removed with a P10 Rainin pipette tip, and the beads were allowed to dry until the pellet showed signs of cracking. The beads were then resuspended in 20μL of 10mM Tris-HCl pH 8.5 for 10 minutes, and then separated on the magnet for 5 minutes. The supernatant containing the amplified cDNA was then withdrawn and 1μL was used for quantification with Qubit HS DNA reagents (Lifetech). An additional 1μL aliquot of the amplified sample was diluted to 3ng/μL, and then used for fragment length estimation on the Agilent BioAnalyzer using the HS cDNA kit.

Ten of the single cell samples were reverse transcribed and amplified as single cell aliquots. Ten were lysed and denatured, then pooled together and re-split to homogenize the mRNA populations in each (pool/split samples). The 10 and 100 cell pools were processed as the single cell aliquots, except they were amplified for 22 and 18 cycles each.

Tagmentation

Tagmentation (Illumina/Nextera) uses a transposase mixture to simultaneously fragment and tag the ends of fragmented cDNA with amplification primers. 50ng aliquots of the SMART amplified cDNA were combined with tagmentation reagents according to the manufacturers protocol. After tagmentation, the reaction was cleaned up using 1.5 volumes of QG buffer (Qiagen) and 1.8 volumes of Ampure XP SPRI beads, according to the protocol of Gertz et al. 2012. The tagmented cDNA was eluted from the beads in 20μL of Tris-HCL pH 8.5, and subjected to an additional 5 rounds of amplification, according to the manufacturers

protocol. The amplified and tagmented cDNA was cleaned up using 0.8 volumes of SPRI beads, washed twice with 200μL of 80% ethanol, dried and eluted with 30μL of Tris-HCl pH 8.5.

The tagmented libraries were quantified with Qubit HS DNA reagents, and 3ng from each sample were assayed on the Agilent BioAnalyzer using the HS cDNA kit. Libraries were judged to be acceptable if they showed a peak in the 300-400bp range. Library sequencing was performed on the HiSeq 2000 Illumina instrument. Single-end, 100bp reads were generated.

Preparation of quantification standards

The quantification spike-in standards are designed to test a range of copy number concentrations over 3 factors of 10. We chose two size ranges ($\sim$ 300nt and $\sim$ 1400nt) to test the effect of transcript length on counting accuracy. The following transcripts were amplified from *Arabidopsis* total RNA for use as quantitation standards: VATG (376nt), OBF5 (1444nt), Apetala2 (1405nt), PDF (348nt), EPR (1451nt), AGP (323nt). These amplified cDNAs were cloned into a modified cloning vector containing the pBluescript II promoters and multiple cloning site, flanking an elongated polyA sequence. The resulting clones were linearized downstream of the polyA sequence, so that in vitro transcription would result in the automatic inclusion of a polyA tail, without the need for polyA polymerase. In vitro transcription was performed using the EpiCentre Ampliscribe T3 in vitro transcription kit (catalog #AS3103). The reactions were cleaned up using a Qiagen RNA cleanup column (Qiagen catalog #74124). The transcribed products were quantified using Qubit RNA reagents (3 repeated measures) and then size verified on the Agilent BioAnalyzer using RNA Nano reagents. The transcripts were then diluted in diluent containing yeast tRNA as a carrier (Ambion Catalog #AM7119) and RNAse inhibitor (Clontech catalog #2313A), and then combined into a cocktail for use as 1μL aliquots. The final concentrations for tRNA was 100pg/μL. The final concentrations of the spike-in standards are listed in Supplementary Table 2.

In silico simulation of single-cell and cell pool transcriptomes

We aimed primarily to examine the effects of the levels of technical stochasticity and the amount of input, but also tried to approximate what a real population of cells might look like, with all the variation of gene expression on the single-cell level that exists in it. To this end, we used the following model.

Let $|S|$ be the number of cells pooled, and p_{E_g} be the probability that a gene g belonging to the set of polyadenylated genes G is expressed in any given cell $S_i \in S$. There likely exist a group of housekeeping genes for which $p_{E_g} \approx 1$, and then there is a continuum of genes for which $p_{E_g} < 1$. Finally, there likely exist genes that are present only in a small fraction of cells for which $p_{E_g} \ll 1$. We denote with T the total number of mRNA molecules expressed in each cell S_i, with C_{C_g} the true number of transcript copies per cell for each gene $g \in G$, where G is the set of all genes. By definition, $T = \sum_{g \in G} C_{C_g}$. For simplicity, we assume it is constant for each cell.

We derive FPKM estimates FPKM_g for each gene based on bulk RNA-seq measurements. For simplicity, and since this does not in any way affect the conclusions of the simulations, we assume that the ratios of FPKM values between genes are equal to the ratios between the their absolute number of transcript molecules in the very large cell pool from which the library was built. We then derive an estimate for the average value of C_{C_g} when a gene is expressed in a given cell as follows:

$$C_{C_g} = \frac{E_g * FPKM_g}{\sum_{g \in G} E_g * FPKM_g} * T \quad (1)$$

where we account for the fact that only a portion of cells express the gene by setting $E_g = 1$ when a gene is expressed in a given cell, and $E_g = 0$ when it is not (E_g is set based on the probability p_{E_g}, as described further below).

Finally, we define the single-molecule capture efficiency p_{smc} as the probability that any given RNA molecule in a cell will be converted into cDNA, amplified and eventually present in the sequencing library.

We use the following algorithm for generating *in silico* cell pool transcriptomes and then the FPKM values in the corresponding libraries. We denote the number of original transcript copies present in the final library (after the effects of technical stochasticity have been modeled) with C_{L_g}

Algorithm 1 Cell pool RNA-seq simulation

for $g \in G$ **do**
  $C_{L_g} \leftarrow 0$
end for
for $i = 1 \rightarrow |S|$ **do**
  for $g \in G$ **do**
    $p \leftarrow$ random number $\in [0, 1]$
    if $p \leq p_{E_g}$ **then**
      $E_g \leftarrow 1$
    else
      $E_g \leftarrow 0$
    end if
  end for
  for $g \in G$ **do**
    $C_{C_g} \leftarrow \frac{E_g * FPKM_g}{\sum_{g \in G} E_g * FPKM_g} * T$

for $i = 1 \rightarrow C_{C_g}$ **do**
$p \leftarrow$ random number $\in [0, 1]$
if $p \leq p_{smc}$ **then**
$C_{L_g} = C_{L_g} + 1$
end if
end for
end for
end for
for $g \in G$ **do**

1. $FPKM_{L_g} \leftarrow \frac{C_{L_g}}{\sum_{g \in G} C_{L_g}} \sum_{g \in G} FPKM_g$
2. $FPKM_{C_g} \leftarrow \frac{C_{C_g}}{\sum_{g \in G} C_{C_g}} \sum_{g \in G} FPKM_g$
3. compare $FPKM_{L_g}$ with $FPKM_{C_g}$

end for

In practice, we have no reliable estimates of what the distribution of p_{E_g} might be across the whole transcriptome (this in itself is a major open research question). We assigned p_{E_g} values to genes by first splitting all genes expressed at FPKM $\geq$ 1 in 10 percentile groups in order of increased expression: $PG \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. For each expression percentile group PG we modeled the distribution of the p_{E_g} values with the normalized Gaussian density $\mathcal{N}(\mu, \sigma)$ over the interval $[F, 1]$ where $\mu = PG$ and $\sigma = |0.9 - PG|$, and F is the minimal fraction of cells a gene can be expressed in (which we set to 0.01).

Sequence alignment and gene expression quantification

Reads were aligned against a combined Bowtie (Langmead et al. 2009) index of the NCBI GRCh37 (hg19) version of the human genome (downloaded from UCSC) excluding the Y chromosome (as GM12878 cells are of female origin) and random chromosomes and the spike-in sequences using TopHat version 1.4.1 (Trapnell et al. 2009; Trapnell et al. 2012) and the GENCODE V13 annotation (Harrow et al. 2012) with the `"--GTF"` option. Read mapping statistics are available in Supplementary Table 1. Gene expression was quantified using Cufflinks version 2.0.2 (Trapnell et al. 2010; Trapnell et al. 2012) for both the GENCODE V13 and refSeq annotations. FPKMs were converted to estimates for copies-per-cell numbers using spike-in sequences of known abundance (Supplementary Table 2); FPKMs were calculated for the spike-ins and used to create a calibration curve for each library (forcing the regression through 0 to avoid the assignment of positive copies per cell to genes with 0 FPKMs) on the basis of which and the Cufflinks FPKMs copies-per-cell estimates were derived for each gene.

Single-molecule capture efficiency estimation

We estimated the average single-molecule capture efficiency based on the number of libraries with 0 FPKM for each spike and the number of input molecules at which that spike was present in the reaction. The actual single-molecule capture efficiency need not be exactly the same for all libraries. It is a binomial process, but it is not possible to estimate it precisely by directly modeling the outcome with a binomial distribution as only the number of complete failures (libraries with 0 FPKM for a given spike, in which all C_s trials where C_s is the number of input copies for spike s) is known. The number of successes (and the corresponding exact number of failures) is not known because multiple copies of each spike are used as input, and as a result, in a library with a non-zero FPKM it is only known that some copies were successfully captured but not how many exactly. We derived an approximate estimate for the single-molecule capture efficiency by treating individual libraries as single trials in a binomial process, then dividing the estimated single-molecule capture efficiency by the number of input copies:

$$p_{smc} = \frac{1}{C_s} \arg\max_p \mathcal{L}(p|L_0 + L_1, L_1) \quad (2)$$

Where L_0 is the number of libraries with 0 FPKM and L_1 is the number of libraries with non-zero FPKM for the spike. This is a relatively crude way to estimate p_{smc} and it works well only when its value is small but in practice the p_{smc} value is indeed small.

For the AGP23 spike (spiked-in at 5 copies), the estimated single-molecule capture efficiency was 0.138 (95% confidence interval 0.106 to 0.164); for the EPR1 spike (10 copies), the estimated single-molecule capture efficiency was 0.053 (95% confidence interval 0.037 to 0.068), and for the PDF1 spike (20 copies), it was 0.045 (95% confidence interval 0.038 to 0.048). As these are approximate estimates, for simplicity we used an average single-molecule capture efficiency $p_{smc} = 0.10$ in subsequent calculations.

Analysis of allele-biased expression

The diploid (May 2011 release) NA12878 genome containing phased SNPs and indels based on the NCBI build 36 (hg18) version of the human genome was downloaded from `http://sv.gersteinlab.org/NA12878_diploid/`. Coordinates for the refSeq annotation for hg18 (downloaded from the UCSC genome browser) were converted into paternal and maternal coordinates. Heterozygous transcriptomes containing two copies of each transcript were built and reads were aligned using Bowtie (Langmead et al. 2009) (version 0.12.7) with the following settings: `"-v 0 -a --best --strata"`, i.e. with no mismatches allowed. Reads were assigned to an allele if they mapped

only to one of the alleles of a gene. All identical reads were collapsed into a single count in order to eliminate PCR amplification artifacts. Allele-biased expression was assessed as follows. First, for each gene using the total number of allele-specific reads for each allele (over all heterozygous positions), a binomial test with a uniform read distribution expectation, a 0.05 p-value cutoff, and a Bonferroni multiple-hypothesis testing correction where the correction factor is the number of genes with sufficiently many allele-specific reads for the binomial test to pass the specified p-value in the case of complete dominance of one of the alleles. Second, the number of copies for each gene was used to derive an estimate for the absolute number of copies per cell for each allele, i.e., for alleles A and a and a per-cell copies estimate for the gene C_E:

$$C_{E_A} = \frac{N_{reads}(A)}{N_{reads}(A) + N_{reads}(a)} C_E \tag{3}$$

Another binomial test similar to the one described above was then run using the C_{E_A} and C_{E_a} estimates. As it is possible that only a small number of reads map differentially to the two alleles of a gene (due, for example, to heterozygous positions being located in a region of poor sequencing coverage) while the gene itself is expressed highly, thus resulting in a significant binomial test using the copies-per-cell estimates that is, however, poorly supported on the read level, both tests were required to pass statistical significance for an allele bias call to be made.

Finally, due to the imperfect single-molecule capture efficiency of the single-cell RNA-seq library building process, it is possible that apparent allele biases are the result of purely stochastic differences between the capture efficiency for the two alleles in a given library. For this reason, we applied a third filter for allele-biased expression calls, which required that the probability of obtaining apparently statistically significant differences in the estimated copies per cell for the two alleles C_{E_A} and C_{E_a} by chance from two independent binomial process with the estimated single-molecule capture efficiency p_{smc} is low ($p \leq 0.05$ after applying Bonferroni multiple hypothesis testing correction):

$$p = \sum_{C_E}^{C_C} \frac{NB(C_C - C_E, p_{smc})}{\sum_{C_E}^{C_C} NB(C_C - C_E, p_{smc})} \sum_{0}^{C_{E_a}} B(C_{C_a}, p_{smc}) \sum_{C_{E_A}}^{C_A} B(C_{C_A}, p_{smc}) \tag{4}$$

Where $C_C = 2 * C_{C_a} = 2 * C_{C_A}$ refer to the actual number of copies per cell (as opposed to the estimated number of copies $C_E = C_{E_a} + C_{E_A}$), $NB(C_C - C_E, p_{smc})$ refers to the negative binomial probability that the actual number of copies is C_C given the estimated number of copies C_E:

$$NB(C_C - C_E, p_{smc}) = \binom{C_E + (C_C - C_E) - 1}{C_E - 1} p_c^{C_E} (1 - p_{smc})^{C_C - C_E}$$

and the binomial probabilities $B(C_{C_A}, p_{smc})$ and $B(C_{C_a}, p_c)$ are defined as:

$$B(C_{C_A}, p_{smc}) = \binom{C_{C_A}}{C_{E_A}} {p_{smc}}^{C_{E_A}} (1 - p_{smc})^{C_{C_A} - C_{E_A}}$$

and

$$B(C_{C_a}, p_{smc}) = \binom{C_{C_a}}{C_{E_a}} {p_{smc}}^{C_{E_a}} (1 - p_{smc})^{C_{C_a} - C_{E_a}}$$

The probability was evaluated for possible values of the actual number of copies per cell up to $C_C = min(5000, 100 * C_E)$.

Genes on the X chromosome were excluded from all analysis as the GM18278 cell line is female. The inactivation of the X chromosome leading to a corresponding allelic exclusion was observed as expected (data not shown).

Alternative splicing analysis

We mapped reads using TopHat with *de novo* junction discovery turned on; such alignments are in principle suited for the discovery and analysis of novel splice junctions, a large number of which has been recently reported by the ENCODE consortium (Djebali & Davis et al. 2012). An important step in such analysis is distinguishing between true novel splice junctions on one hand and mapping and

library-building artifacts on the other. Such artifacts certainly exist as we observe "novel junctions" in our spike-in quantification standards, which are not spliced (Supplementary Table 5). Confidence in the reality of newly discovered splice junctions in traditional RNA-seq is boosted by the number of distinct sequencing fragments supporting them, and by replication in other libraries. However, the former line of evidence is not applicable to single-cell RNA-seq due to the one-to-many relationship between original founder RNA molecules and sequencing fragments in the final library, while the latter is difficult to apply in all cases given the uniqueness of each individual single cell. For these reasons, we restricted alternative splicing analysis to known splice junctions and novel junctions, at least one end of which was annotated as splice site in GENCODE V13.

We calculated 5' and 3' splicing inclusion ψ scores as follows (Pervouchine et al. 2013):

$$\psi_5(D,A) = \frac{N_{reads}(D,A)}{\sum_{A_i \in A} N_{reads}(D,A_i)} \quad (5)$$

$$\psi_3(D,A) = \frac{N_{reads}(D,A)}{\sum_{D_i \in D} N_{reads}(D_i,A)} \quad (6)$$

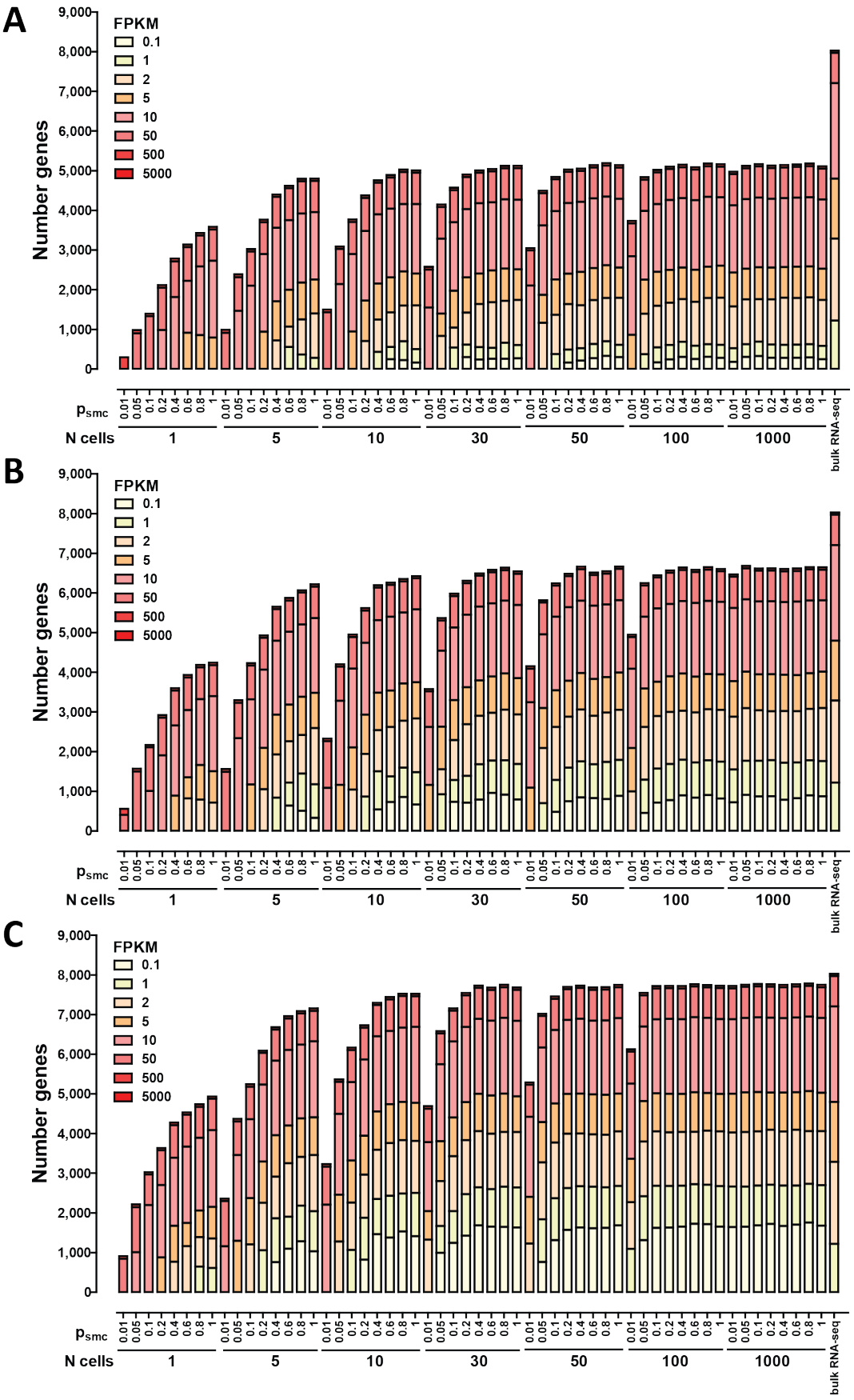
Where D and A refer to the donor and acceptor splice sites, respectively, and the number of reads N_{reads} refers to the number of spliced reads crossing a splice or donor sites after apparent PCR duplicates have been collapsed into a single count. We note that any given donor or acceptor splice site need not be included in all transcripts expressed from the gene it belongs to. Since isoform-level quantification is not a completely solved problem and it is even less clear what its relative stability is for single-cell RNA-seq compared to the bulk RNA datasets for which algorithms have been designed, we only included alternative splice sites for which the donor or acceptor site was found in all annotated transcripts for the gene (GENCODE V13 annotation) as well as novel junctions (compared to the GENCODE V13 annotation) derived from the TopHat mappings involving such splice sites. This allows us to use gene-level FPKM estimates, which are in general more reliable than isoform-level ones, and the mRNA copies-per-cell estimates based on those to derive the approximate absolute number of transcripts containing a given splice junction in each. The statistical significance of bias towards one of the sites was assessed analogously to the approach described for allele-biased expression above, with one significant modification: in cases of more than two possible A_i acceptor sites, for a donor site D or D_i sites for an acceptor site A, the major pair (the one with the most reads) was compared to the sum of reads for all other pairs as if those pairs constituted as single pair. This approach was adopted so that a maximum number of alternative splicing events are included in the analysis and with a focus on identifying cases of robust and statistically significant splice site use switches between individual single cells. When the major (D, A) pair did not have more than half of all reads, the site was excluded from further analysis.

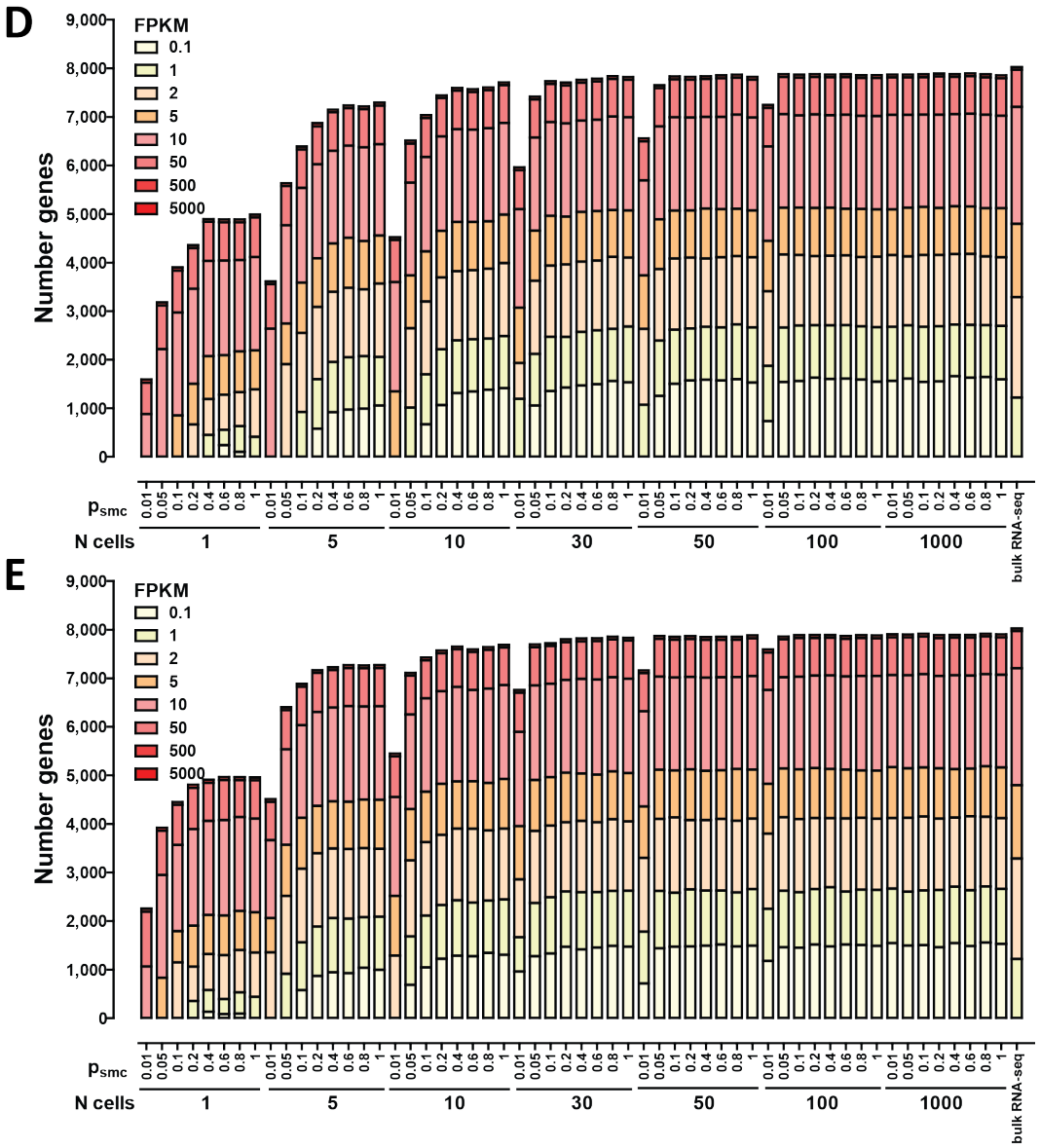
Gene expression clustering and weighted correlation network analysis

Weighted correlation networks (Zhang & Horvath 2005) were constructed from the single-cell vectors of estimated mRNA copies using the WGCNA R package (Langfelder & Horvath 2008) using the `blockwiseModules` function with $\beta = 6$ (Supplementary Figure 34) and a minimum module size of 10 genes. Input genes were filtered as follows: first, we required that genes be expressed at more than one estimated copy per cell C_E in at least one cell; second, we required that the ratio between the C_E variance in single cells and the C_E variance in pool/split libraries be more than 1.5. The latter requirement was imposed in order to minimize the identification of apparently coexpressed gene modules due to purely stochastic differences in transcript capture (see Supplementary Figure 33 for more detail).

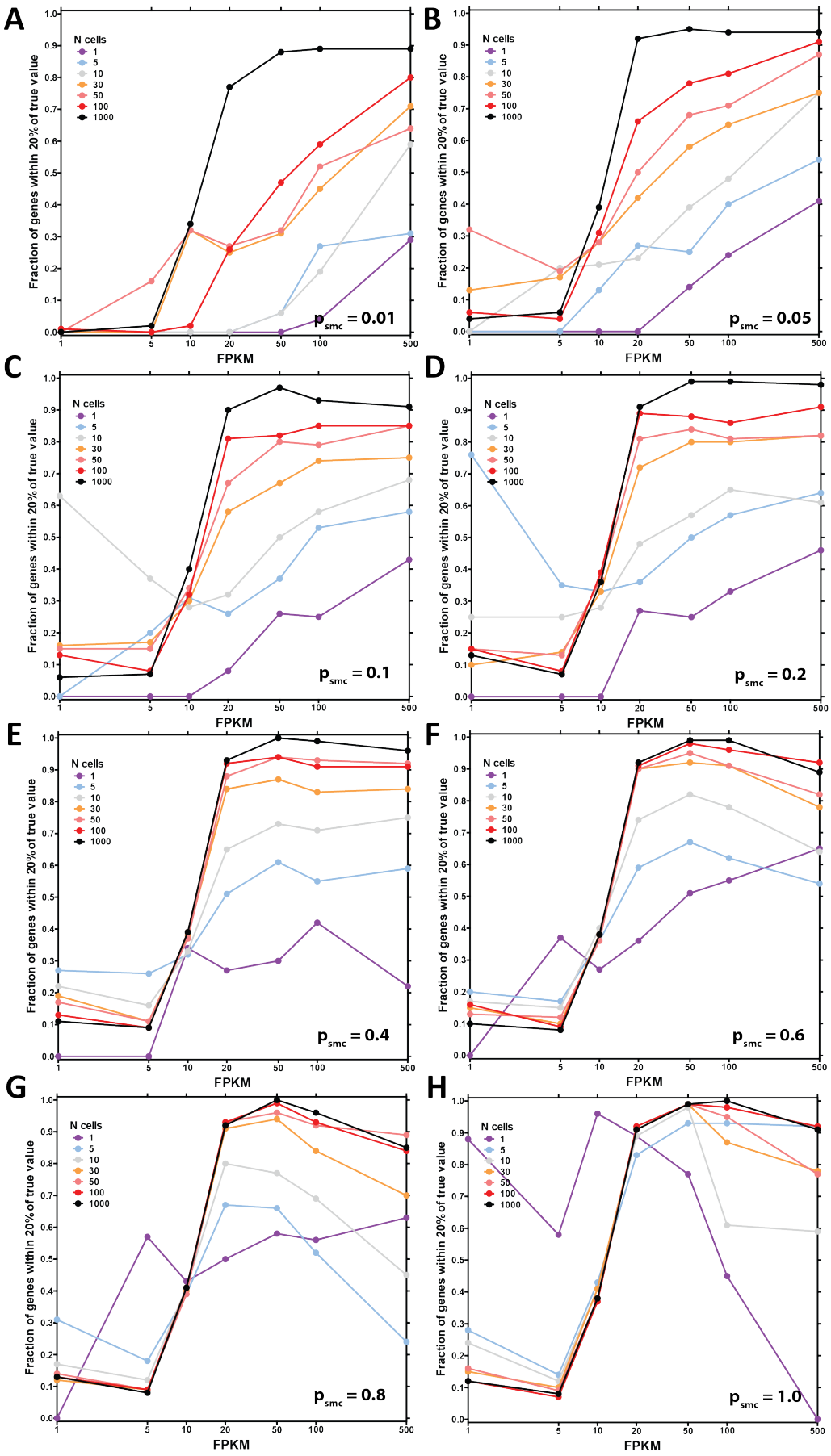
Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al. 2009). Gene expression clustering was carried out using Cluster 3.0 (de Hoon et al. 2004) and visualized using TreeView (Saldanha 2004).

Supplementary Figures

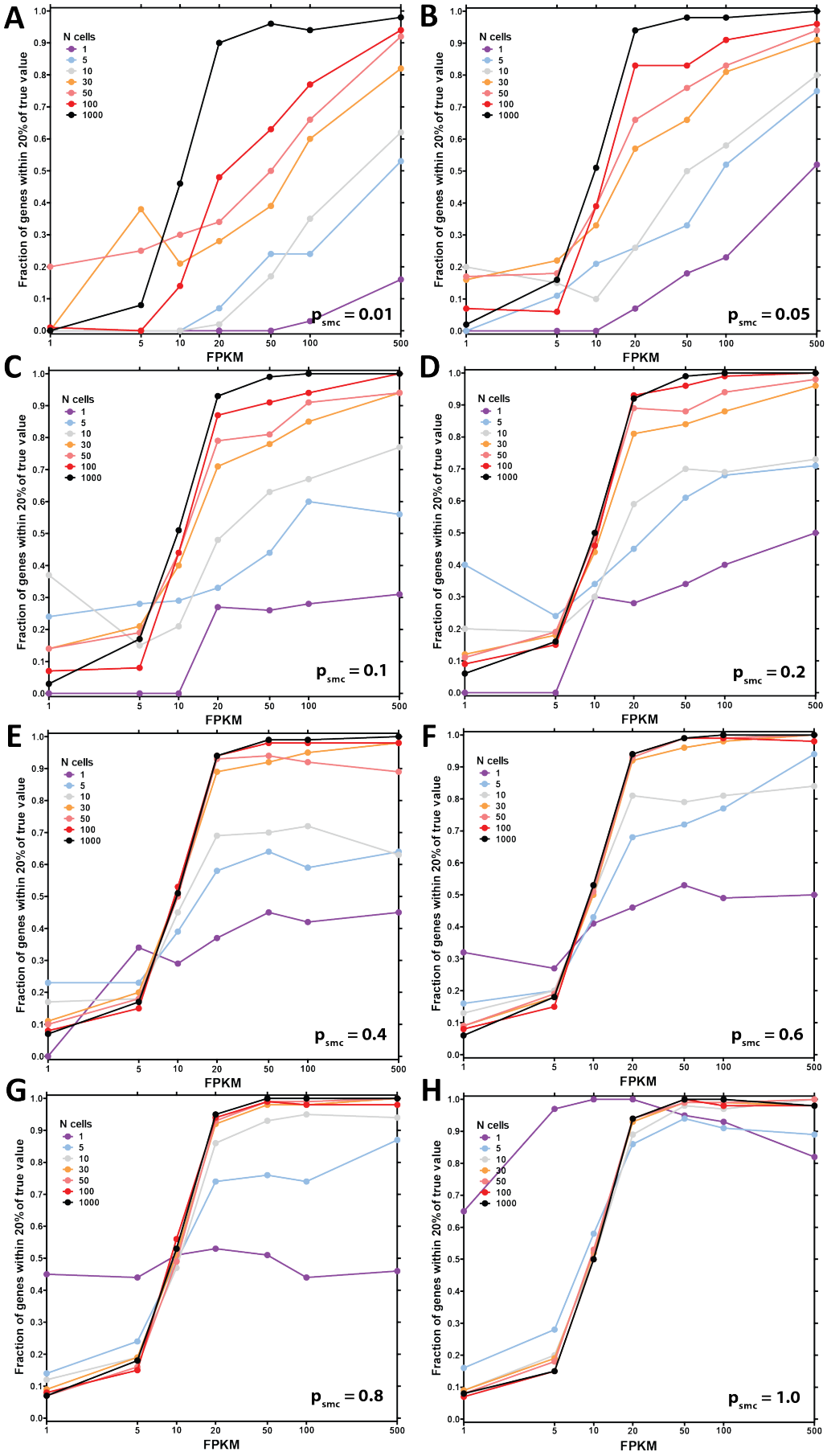




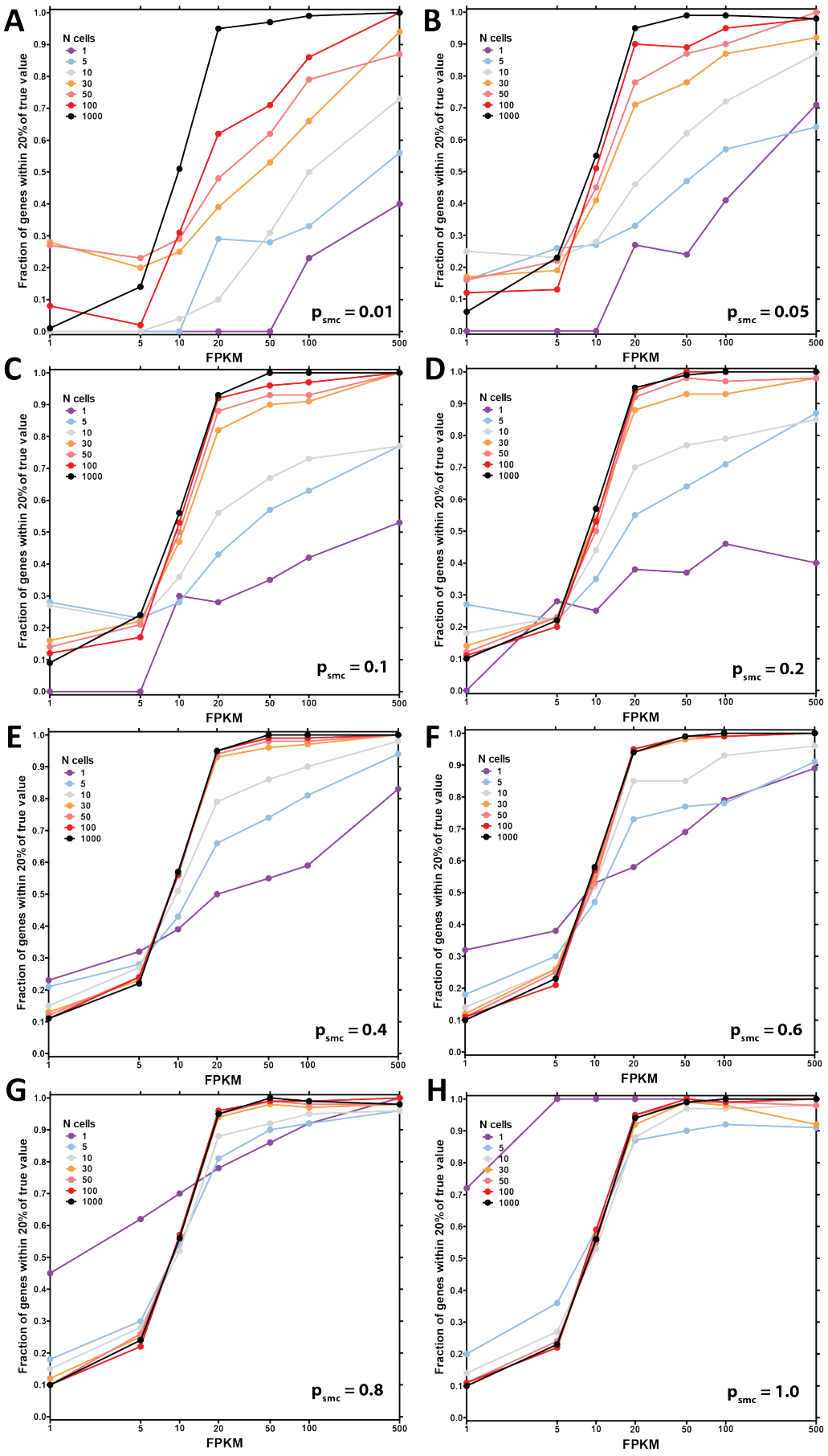
Supplementary Figure 1: Detection of expressed genes in simulated datasets as a function of the single molecule capture efficiency, the number of cells and the average number of transcripts per cell. (A) Average of 50,000 mRNAs per cell. (B) Average of 100,000 mRNAs per cell. (C) Average of 200,000 mRNAs per cell. (D) Average of 500,000 mRNAs per cell. (E) Average of 1,000,000 mRNAs per cell. See the Methods section for full details on how the simulation was carried out.



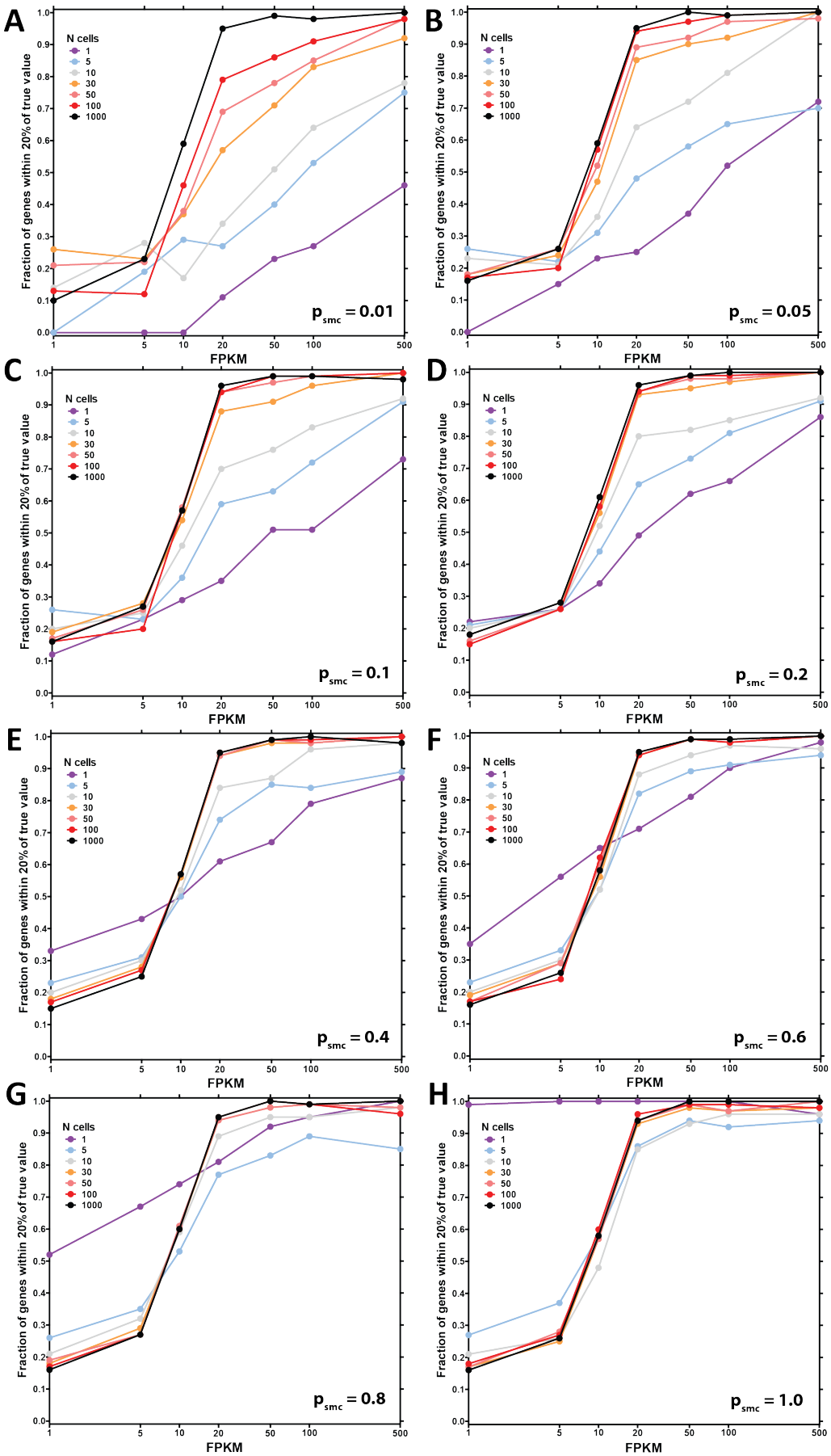
Supplementary Figure 2: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability. Average of 50,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



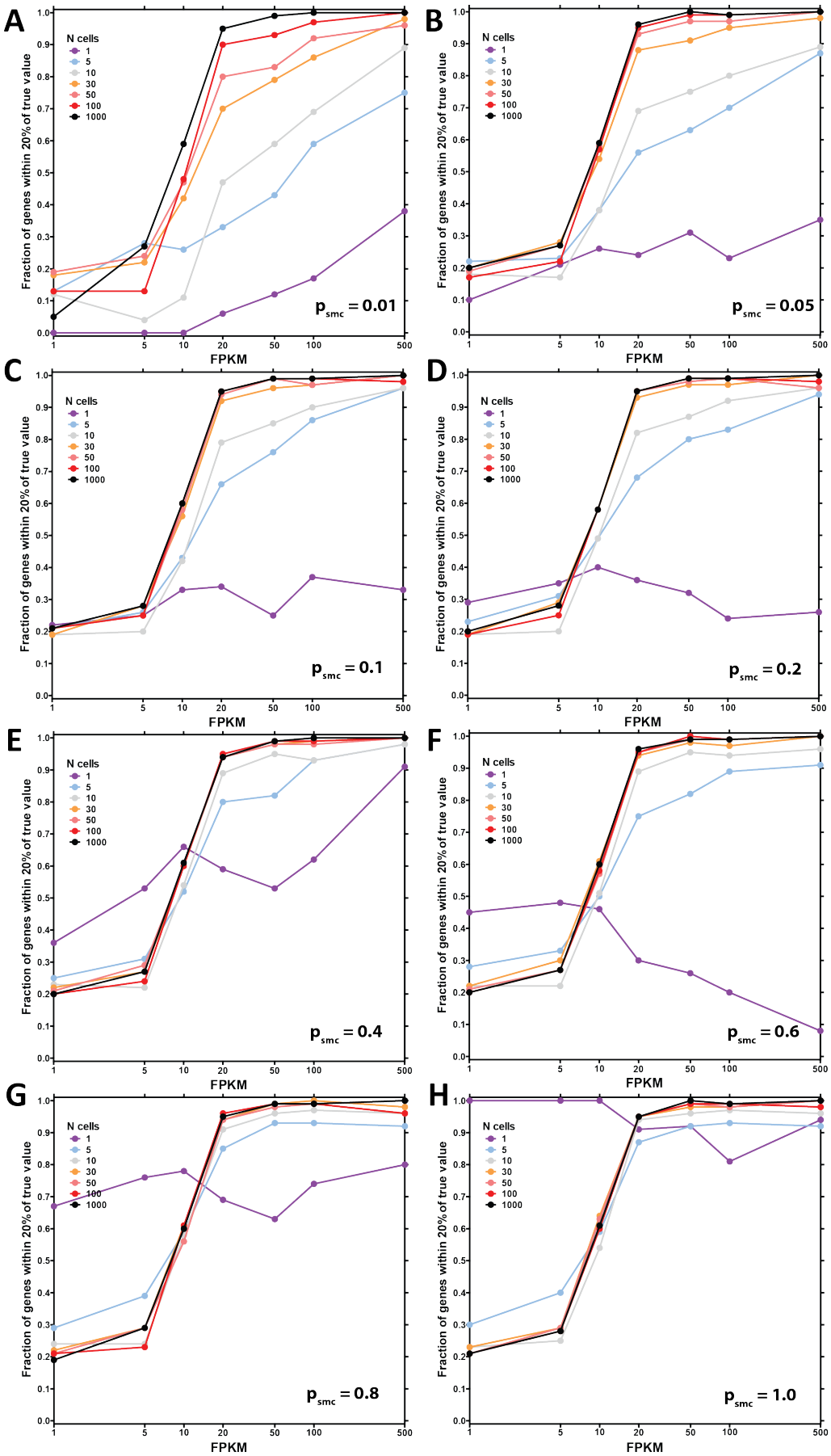
Supplementary Figure 3: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability. Average of 100,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



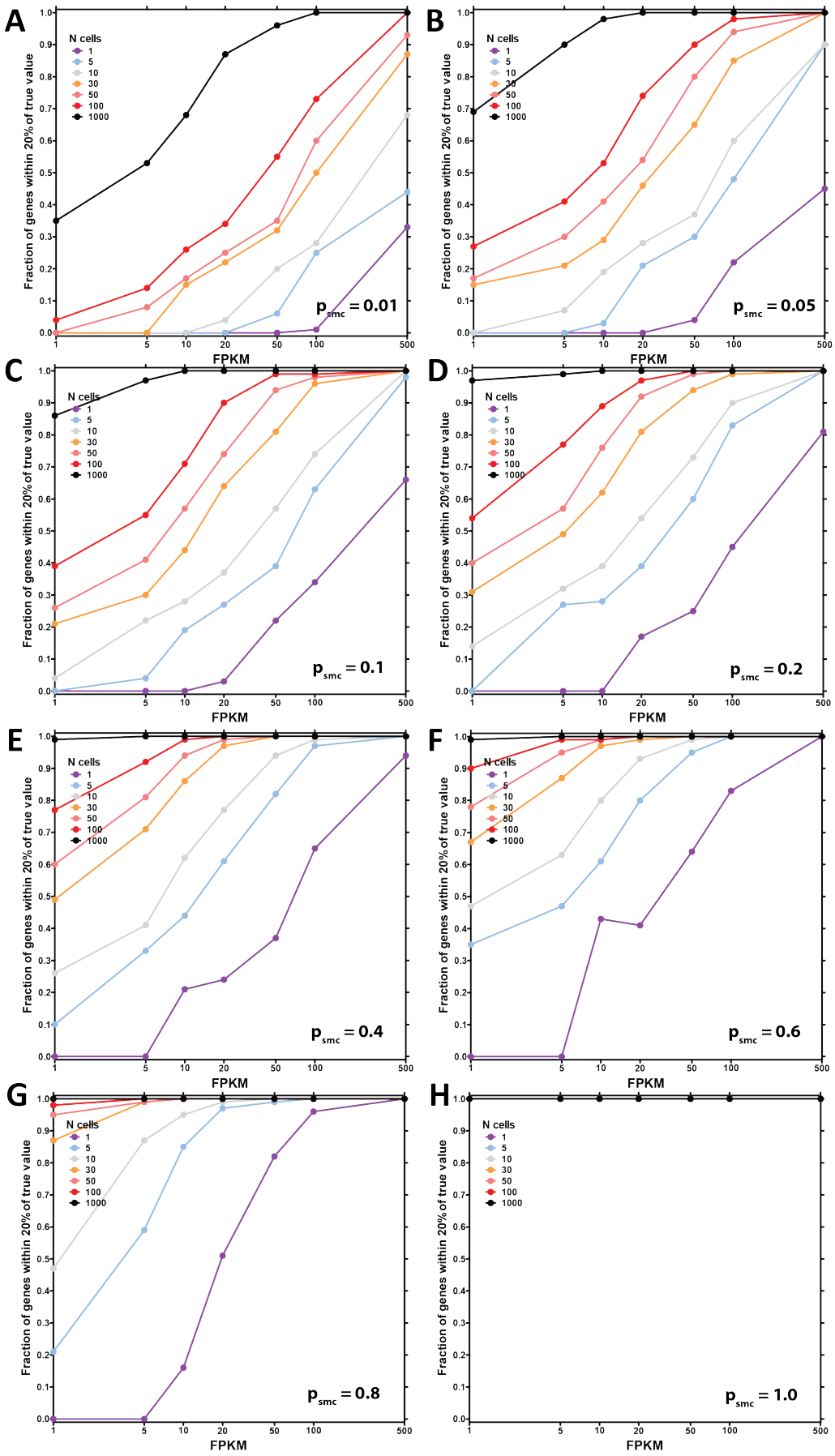
Supplementary Figure 4: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability. Average of 200,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



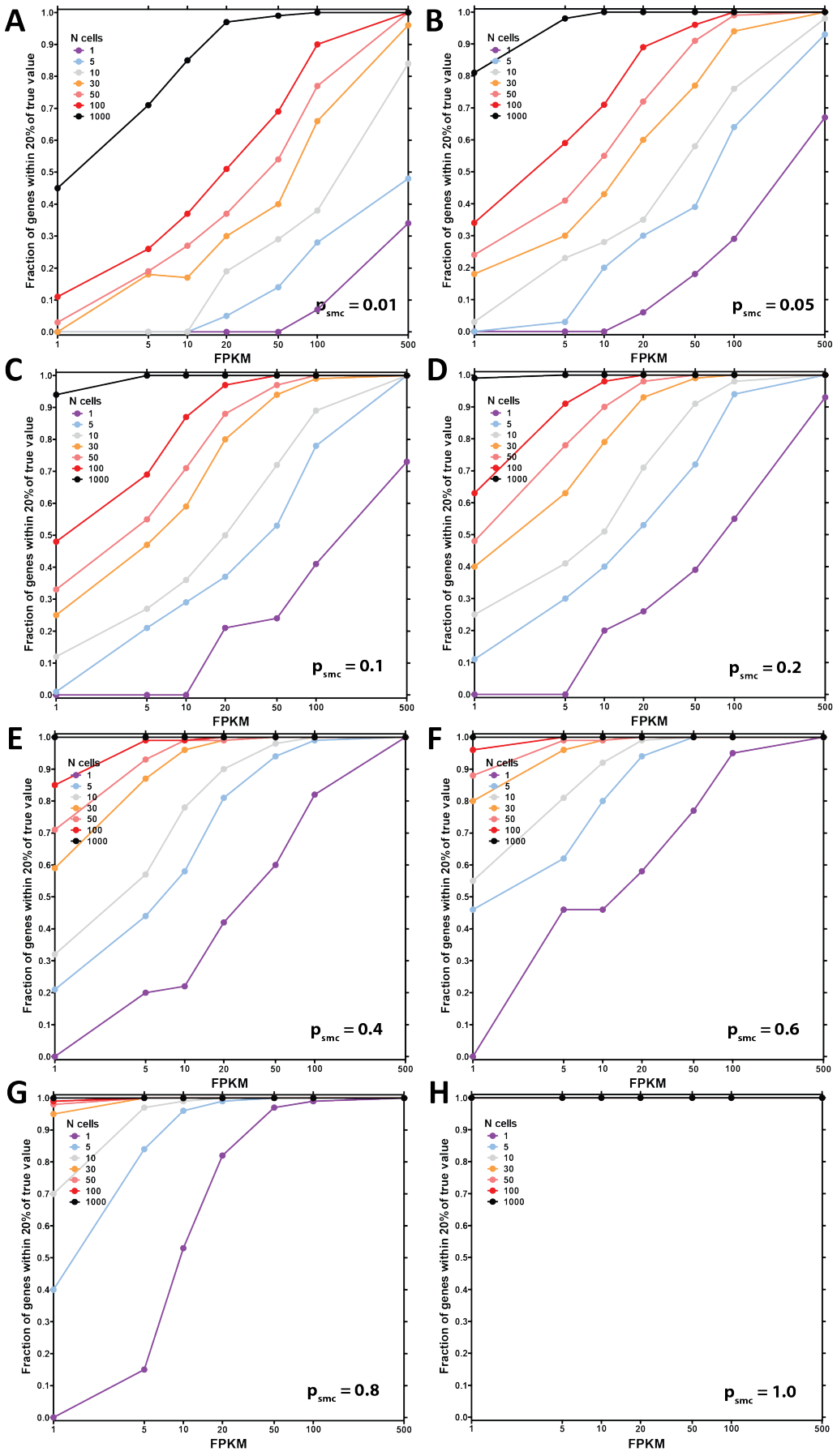
Supplementary Figure 5: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability. Average of 500,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



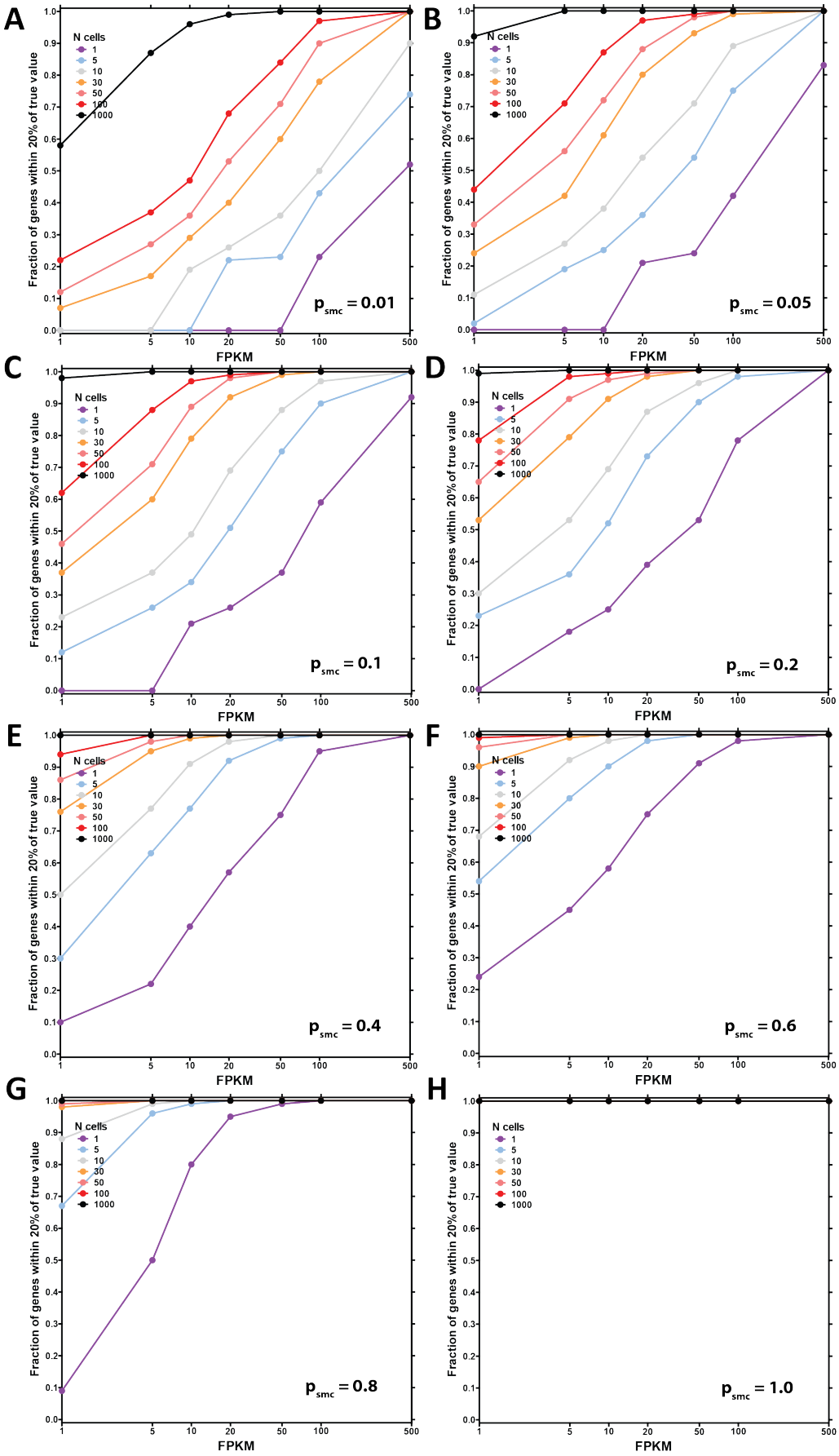
Supplementary Figure 6: Accuracy of estimation of population-level gene abundance as a function of the number of cells pooled and the single molecule capture probability. Average of 1,000,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



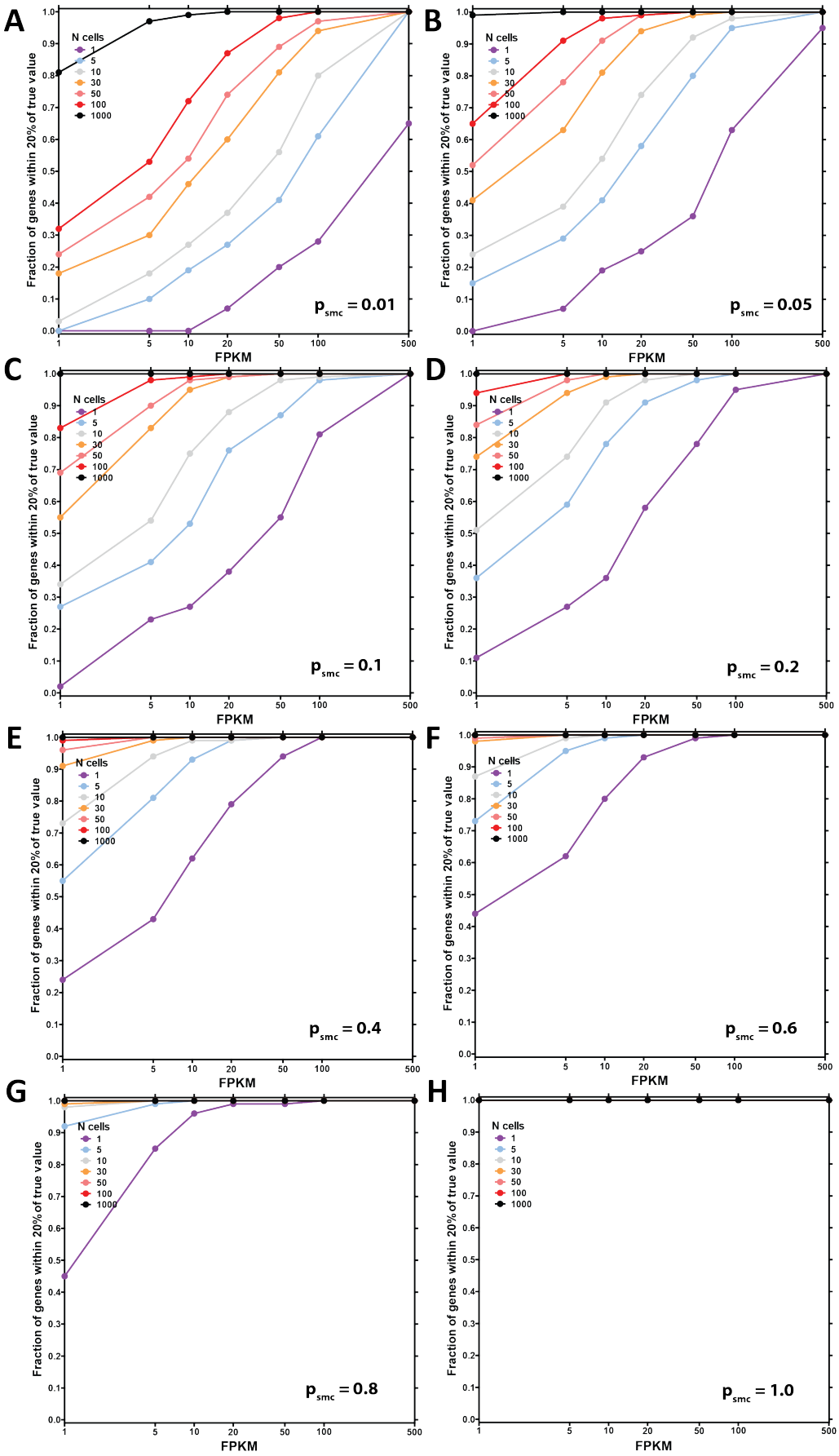
Supplementary Figure 7: Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 50,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



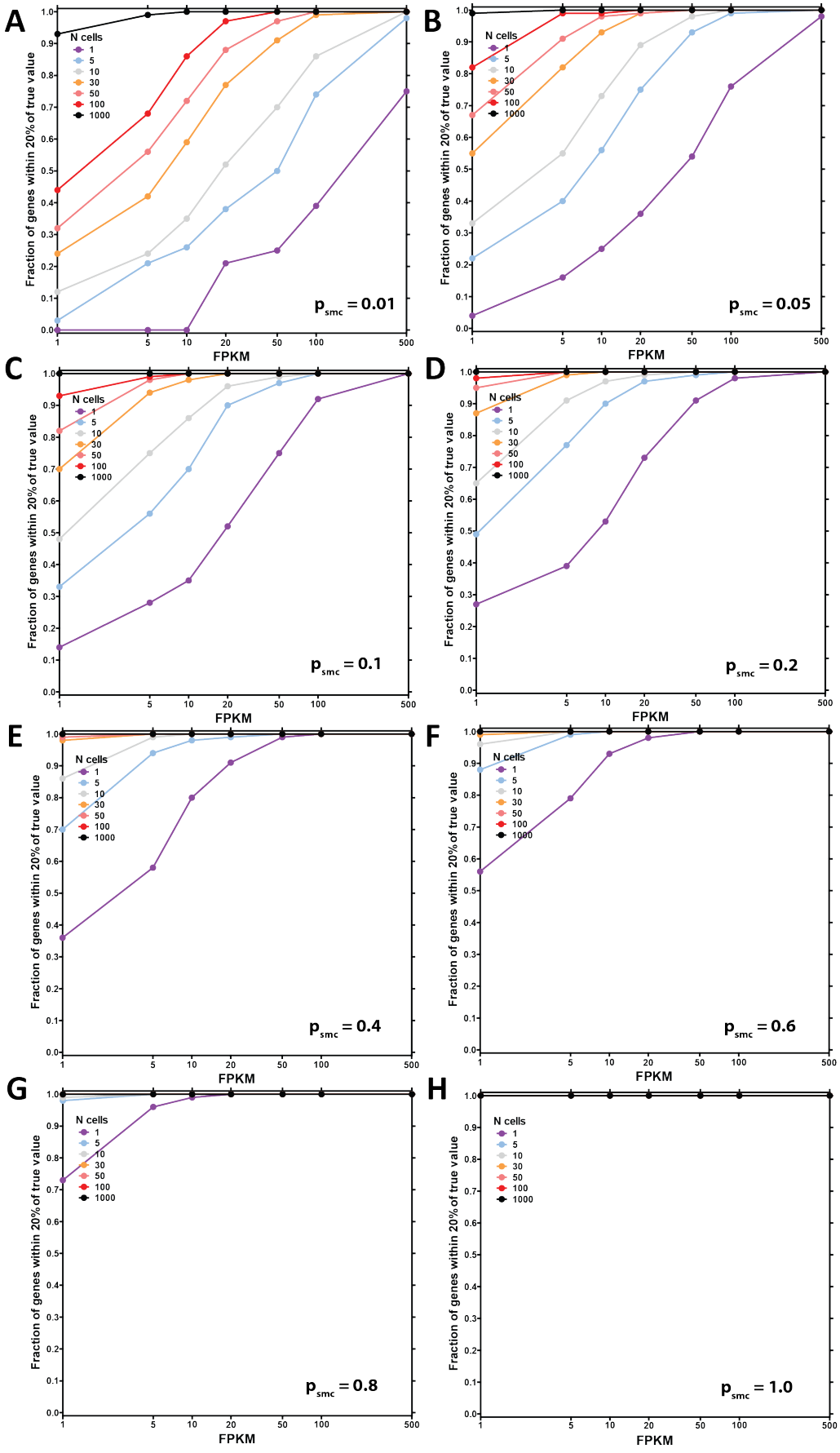
Supplementary Figure 8: Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 100,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



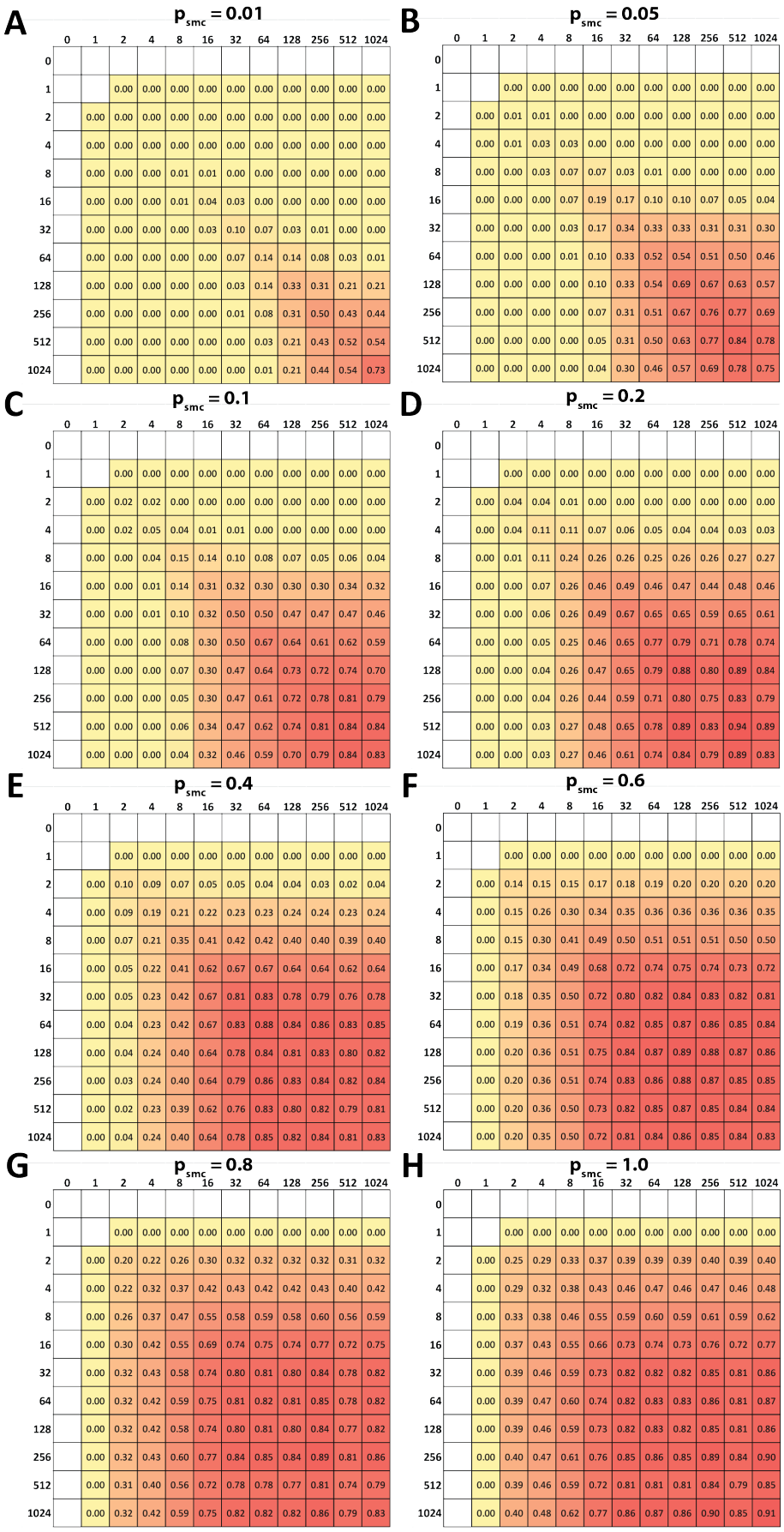
Supplementary Figure 9: Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 200,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



Supplementary Figure 10: Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 500,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



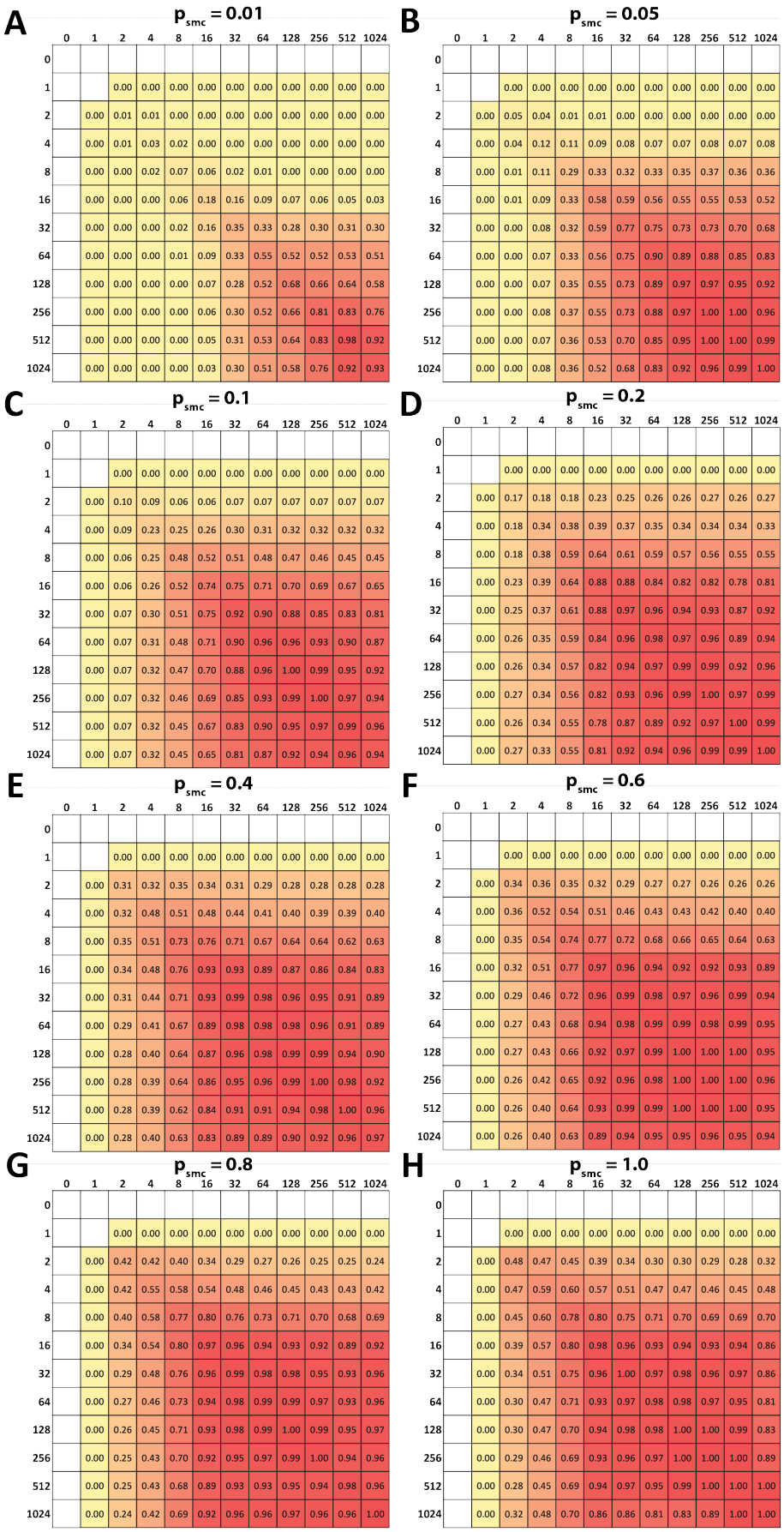
Supplementary Figure 11: Accuracy of estimation of gene abundance within a cell pool as a function of the number of cells pooled and the single molecule capture probability. Average of 1,000,000 mRNAs per cell. Shown is the fraction of genes at the indicated expression levels in FPKM in a bulk RNA-seq dataset, whose estimated expression level in FPKM in simulated libraries was within 20% of their true value after stochasticity due to the probability of capture of cells that express them and the single-molecule capture efficiency of the library-building protocol have been modeled. See the Methods section for full details on how the simulation was carried out.



Supplementary Figure 12: Estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A single cell, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}} \Big/ \frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

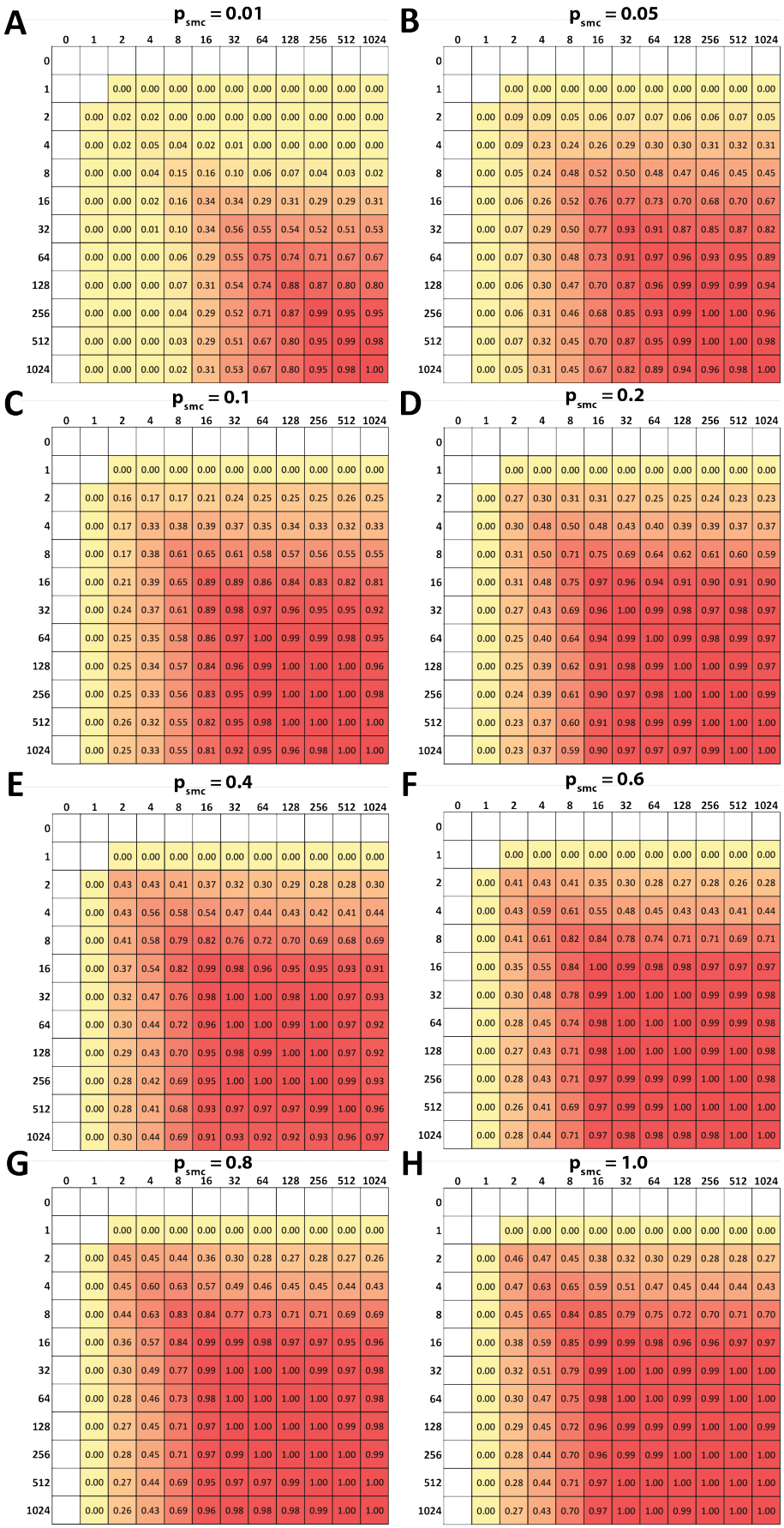
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 13: Estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 5 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\frac{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}{\mathrm{FPKM}_A^{bulk}}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

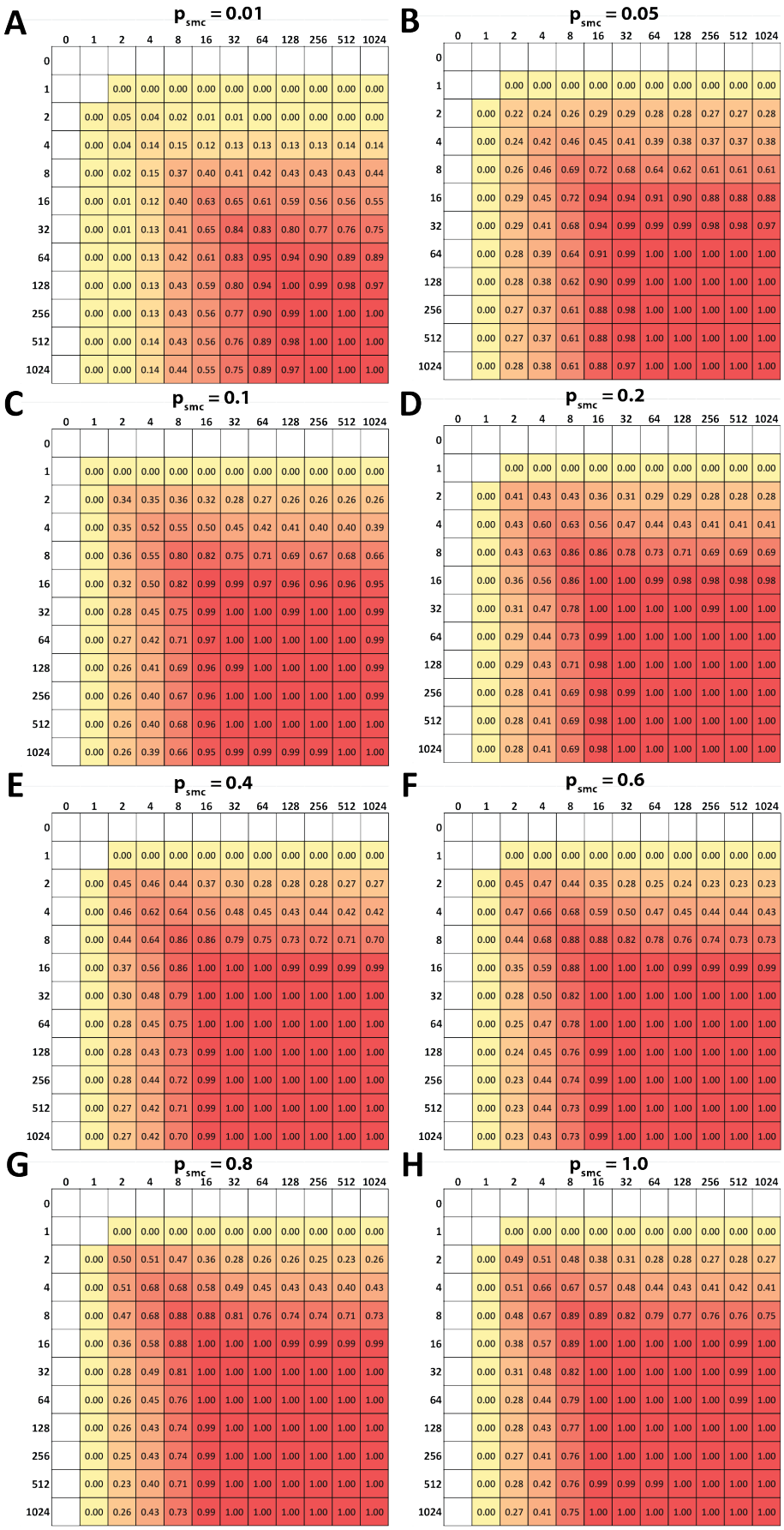
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 14: Estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 10 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}} \Big/ \frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

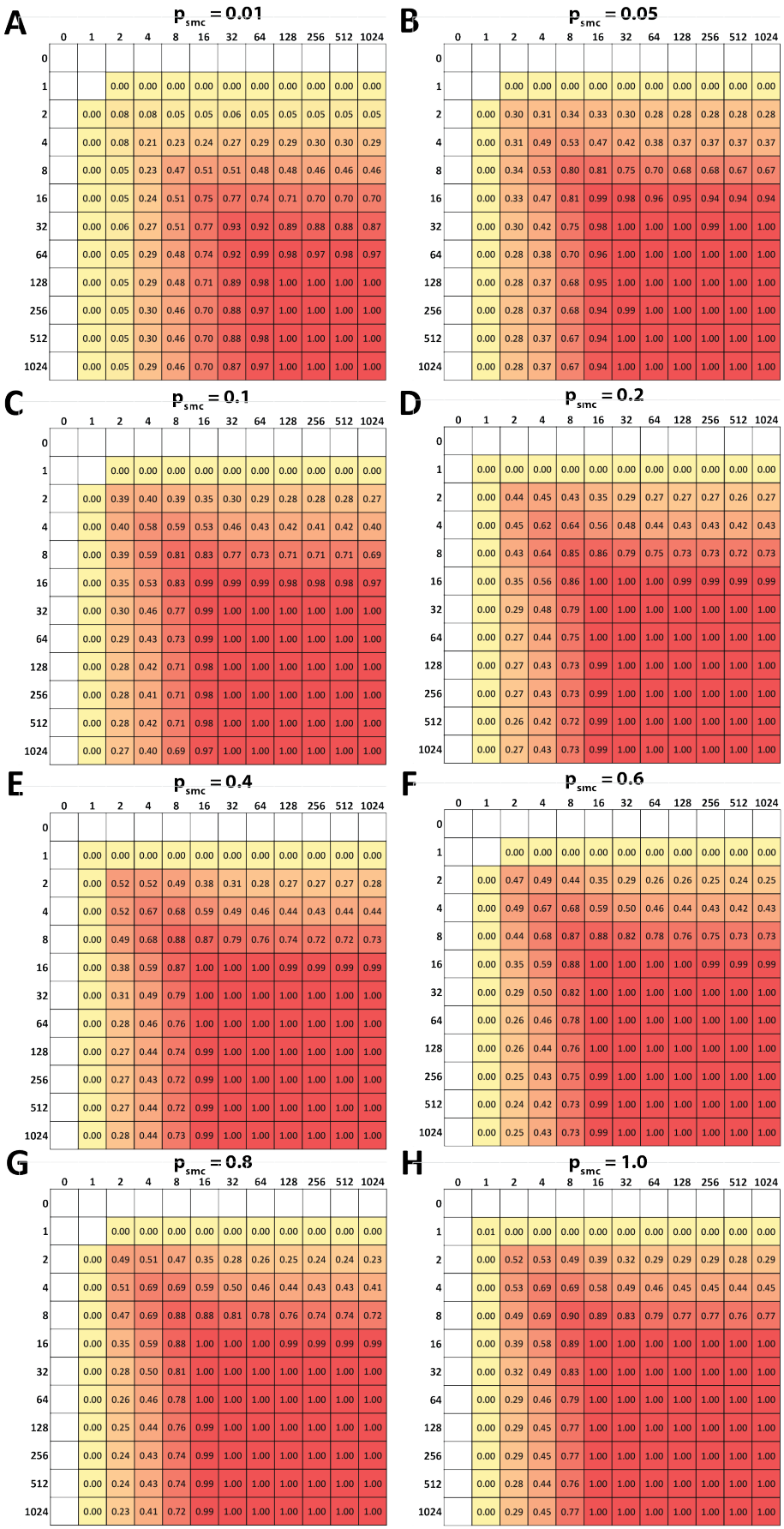
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 15: Estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 30 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}} \Big/ \frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

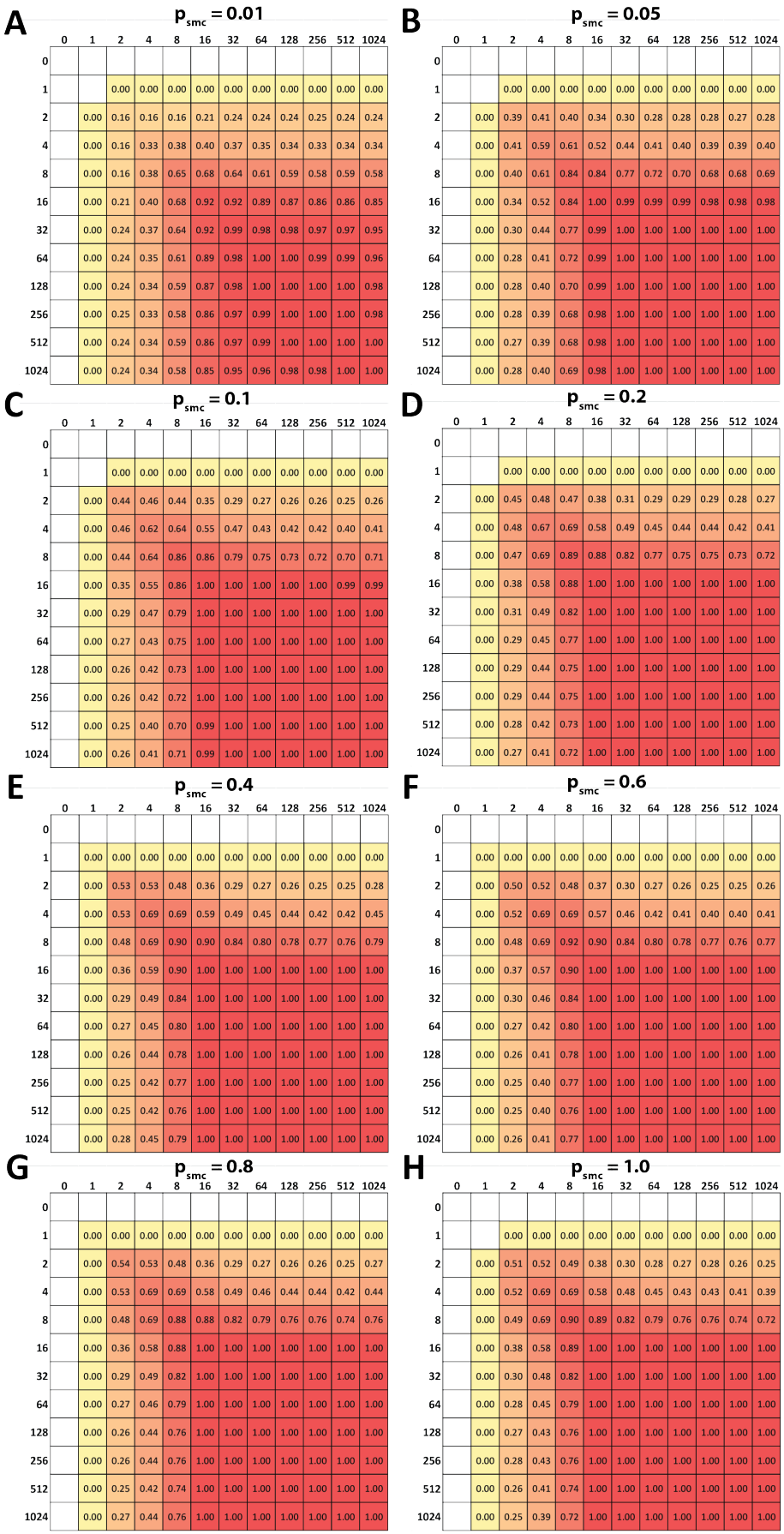
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 16: Estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 50 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bluk}}}$$

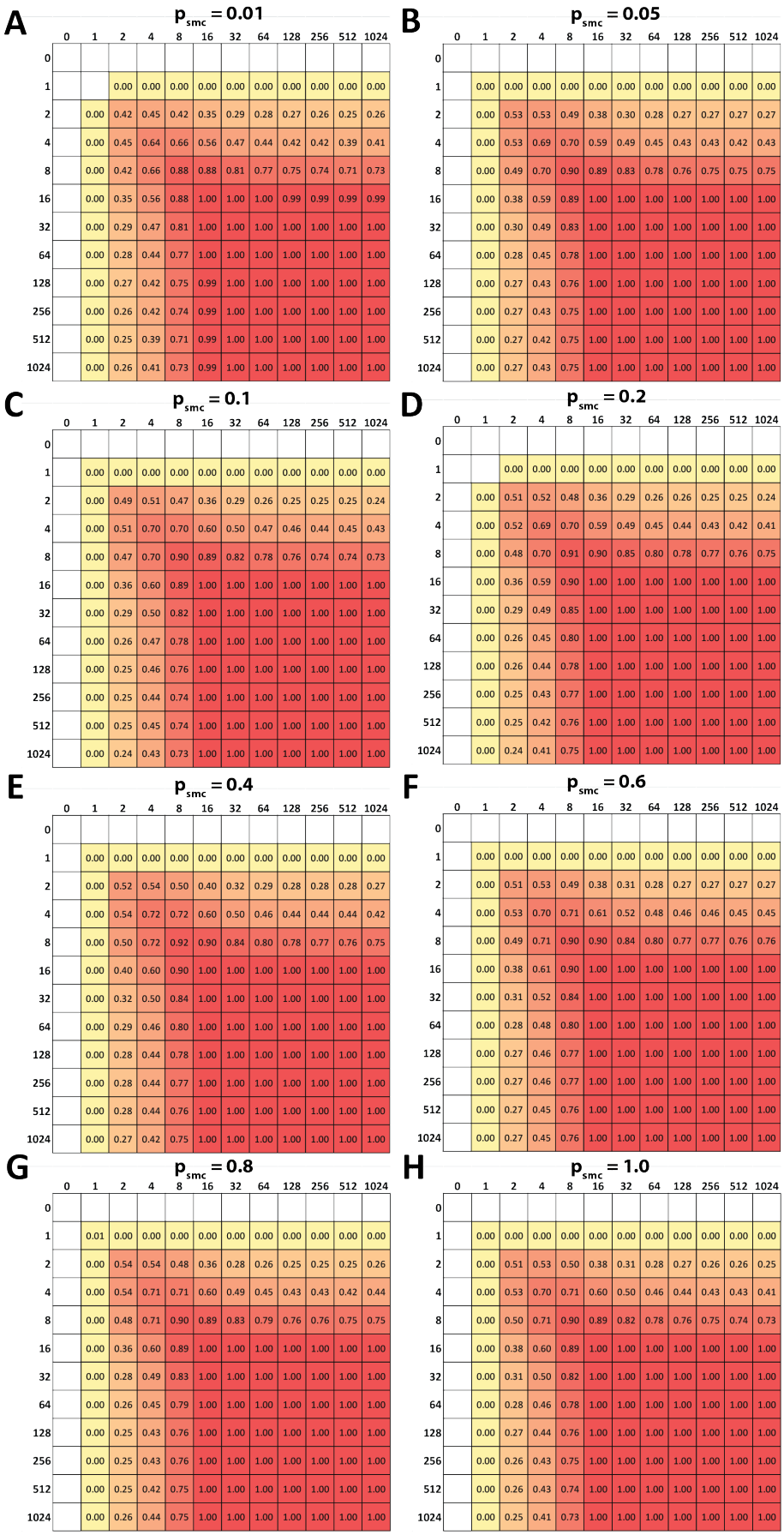
and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 17: Estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 100 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}}$$

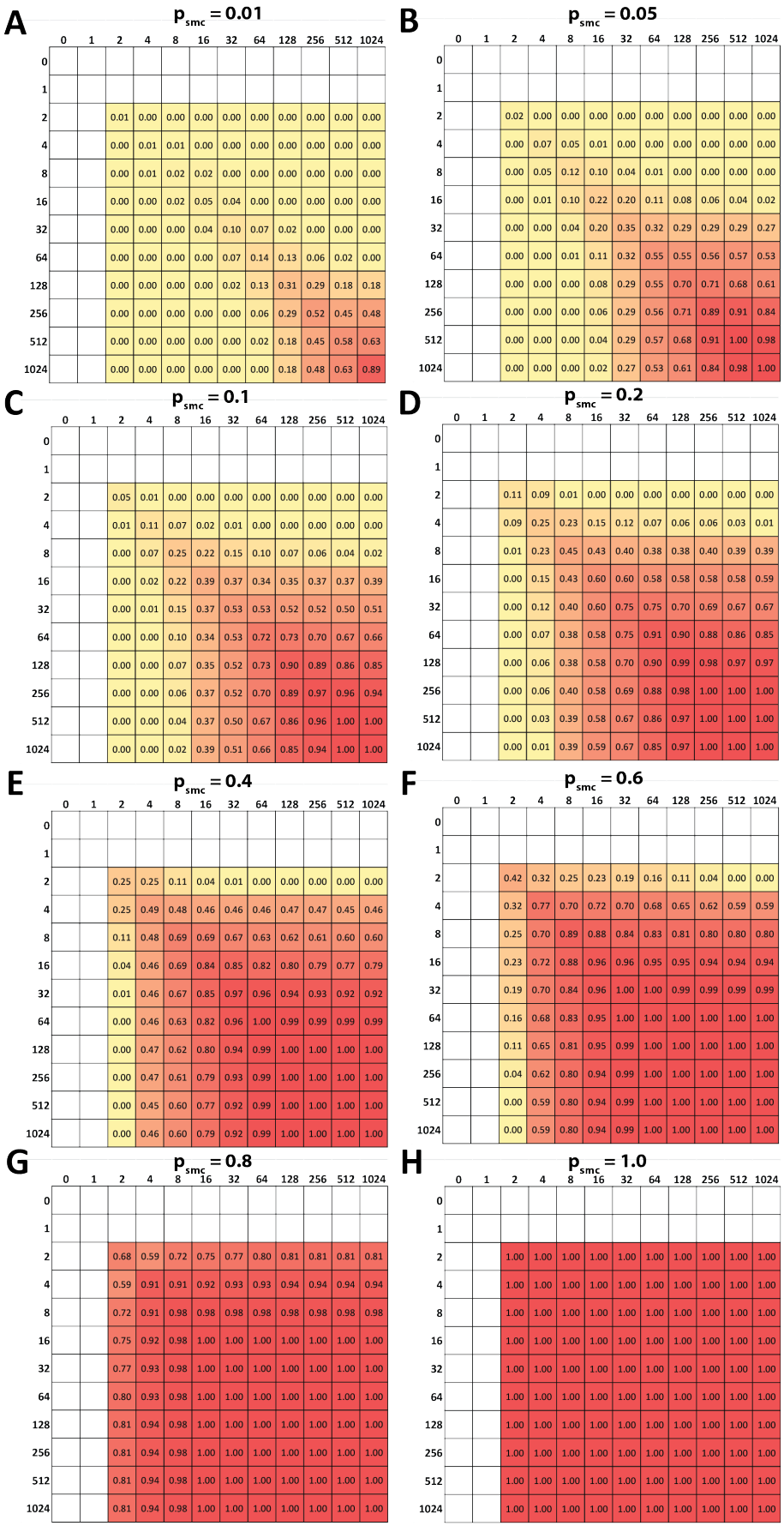
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 18: Estimation of the ratio between the expression values of two genes in bulk RNA-seq as a function of the single molecule capture probability and the size of the cell pool in simulated transcriptomes. A pool of 1000 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\text{FPKM}_A^{pool}}{\text{FPKM}_A^{pool} + \text{FPKM}_B^{pool}}}{\frac{\text{FPKM}_A^{bulk}}{\text{FPKM}_A^{bulk} + \text{FPKM}_B^{bluk}}}$$

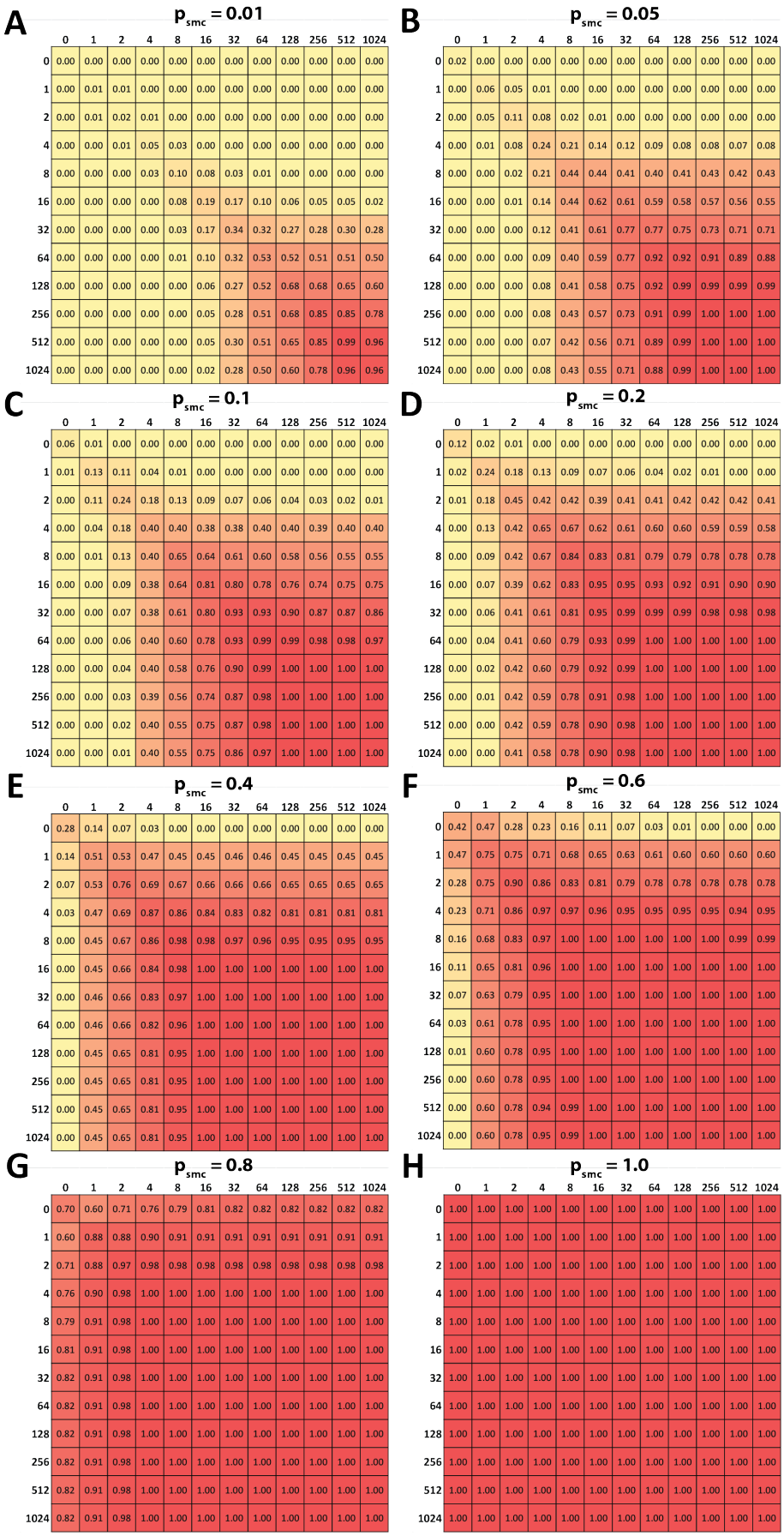
and $\text{FPKM}_A^{bulk} < \text{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 19: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A single cell, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}}$$

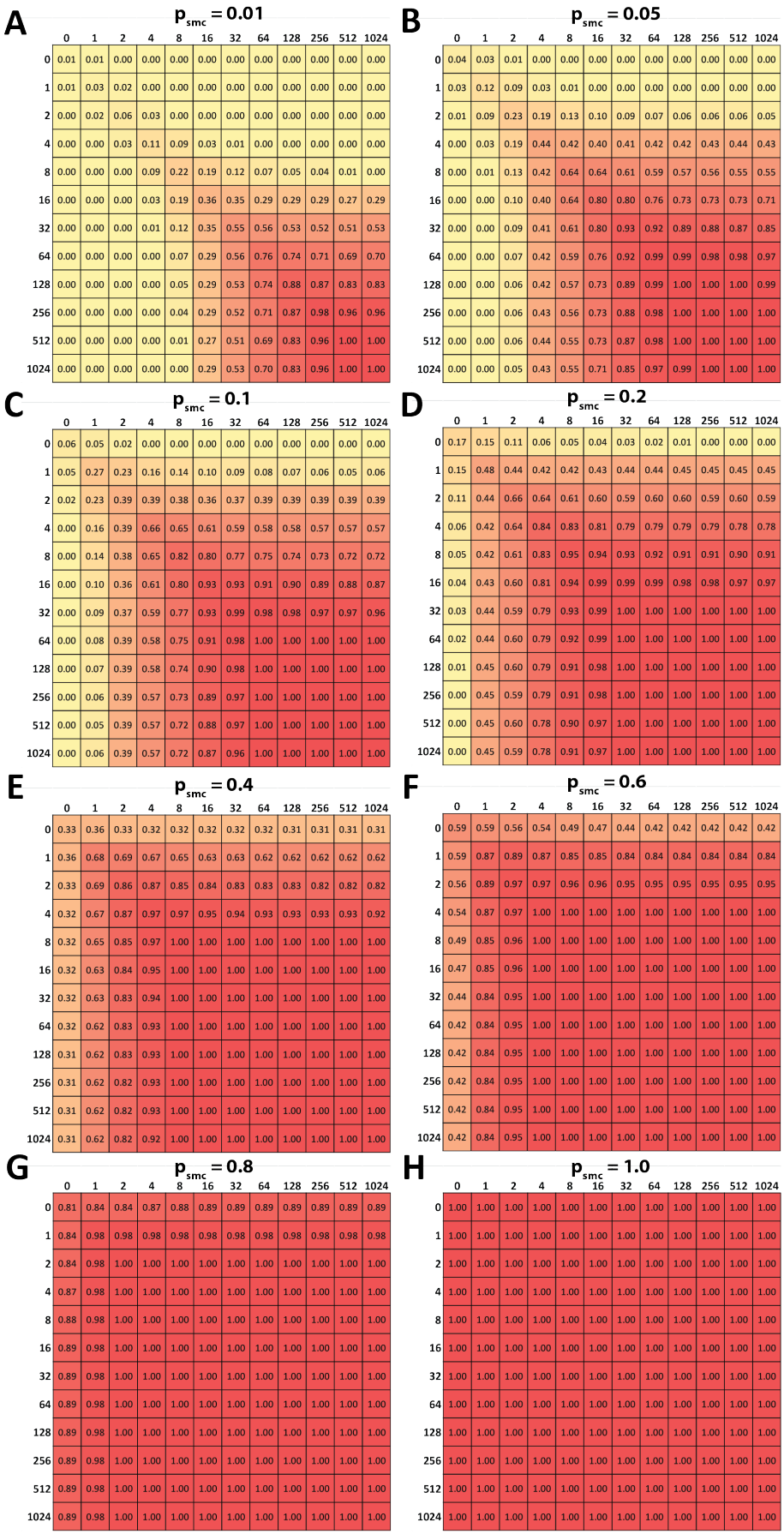
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 20: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 5 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}} \Big/ \frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

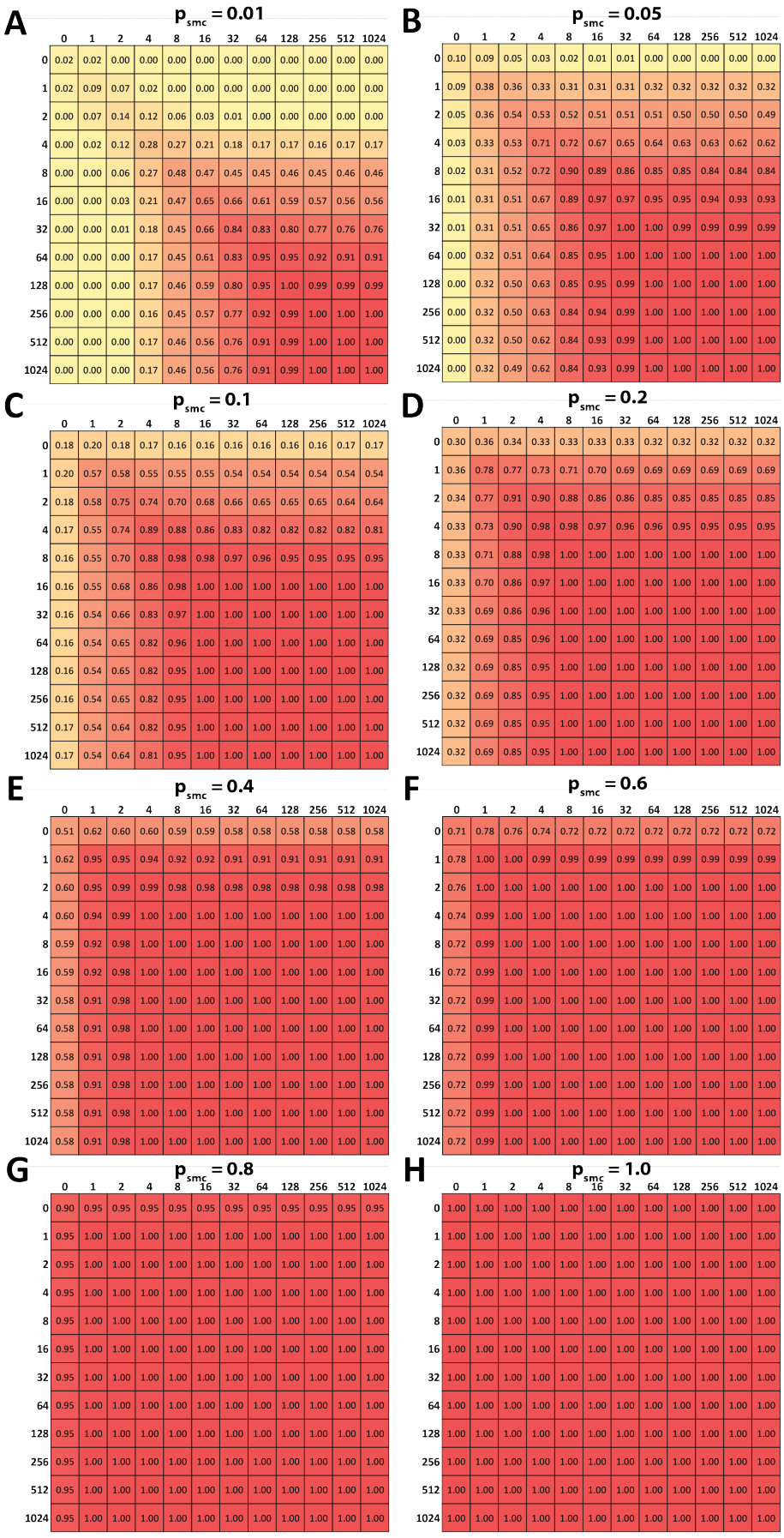
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 21: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 10 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\dfrac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}}$$

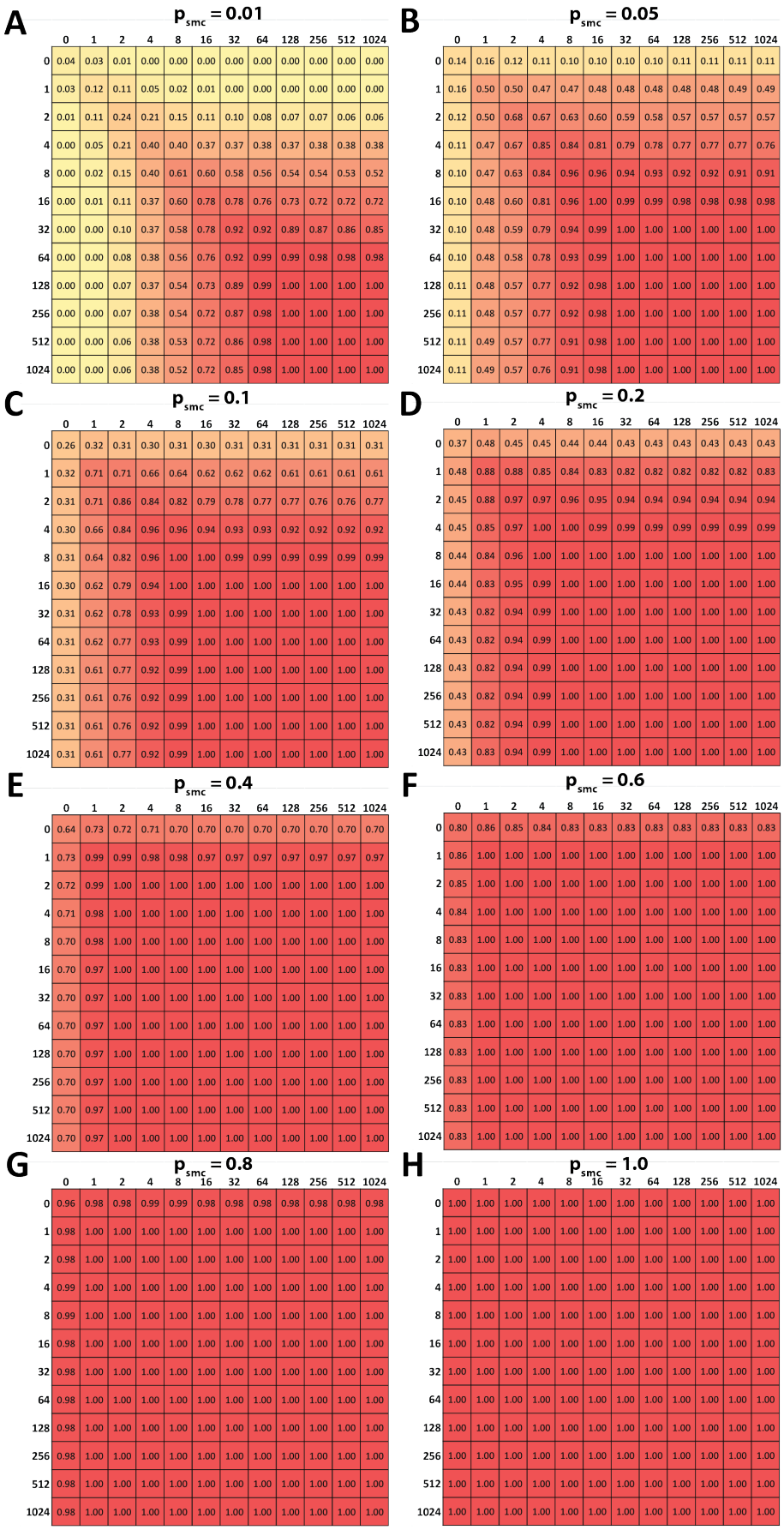
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 22: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 30 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}} \Big/ \frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

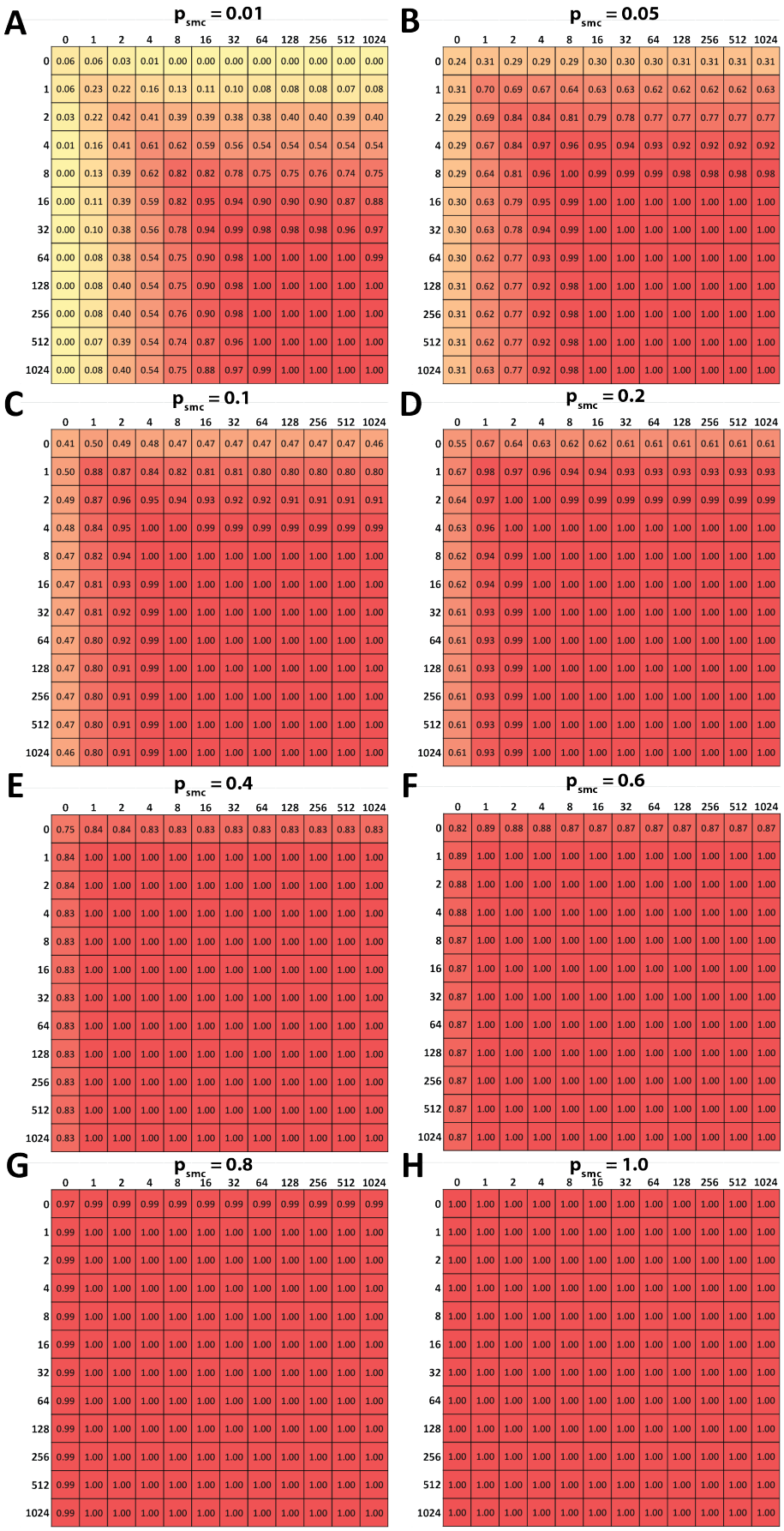
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 23: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 50 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}} \Bigg/ \frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

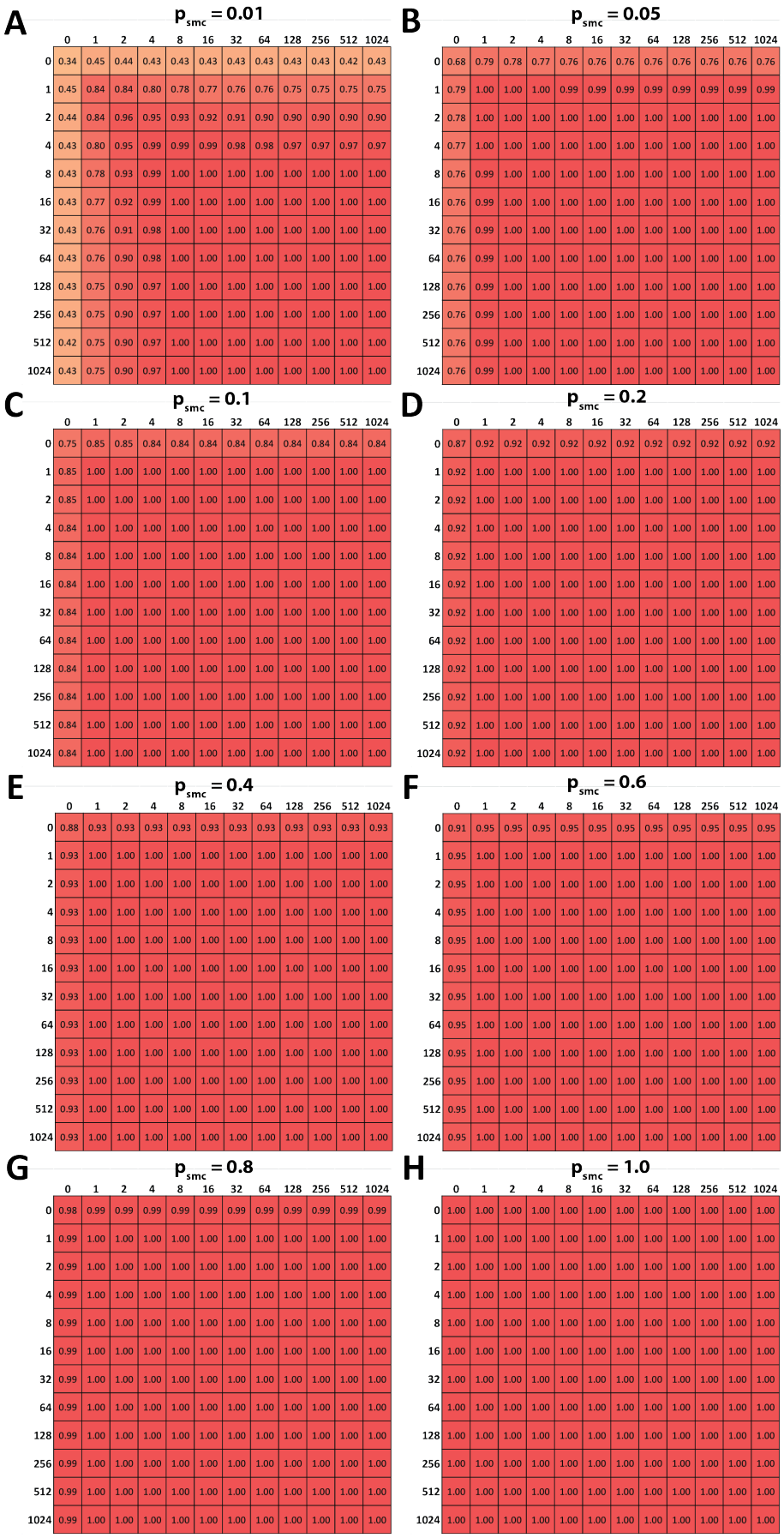
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



Supplementary Figure 24: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 100 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\dfrac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}}}{\dfrac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}}$$

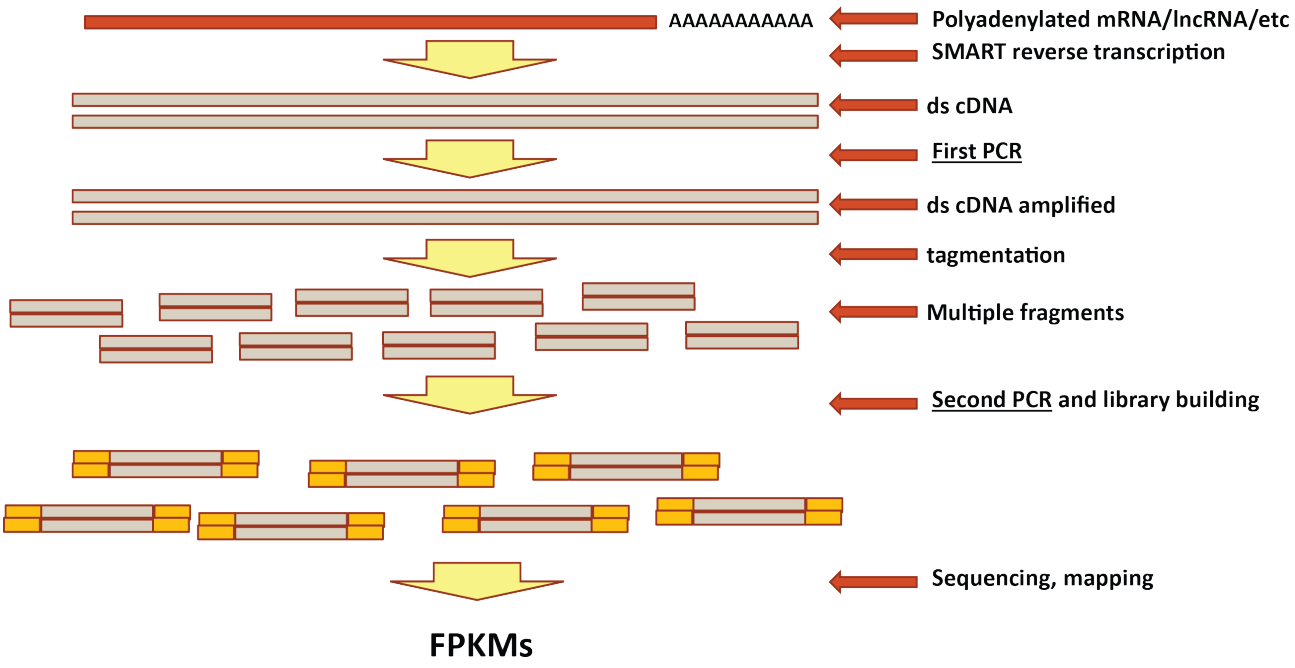
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



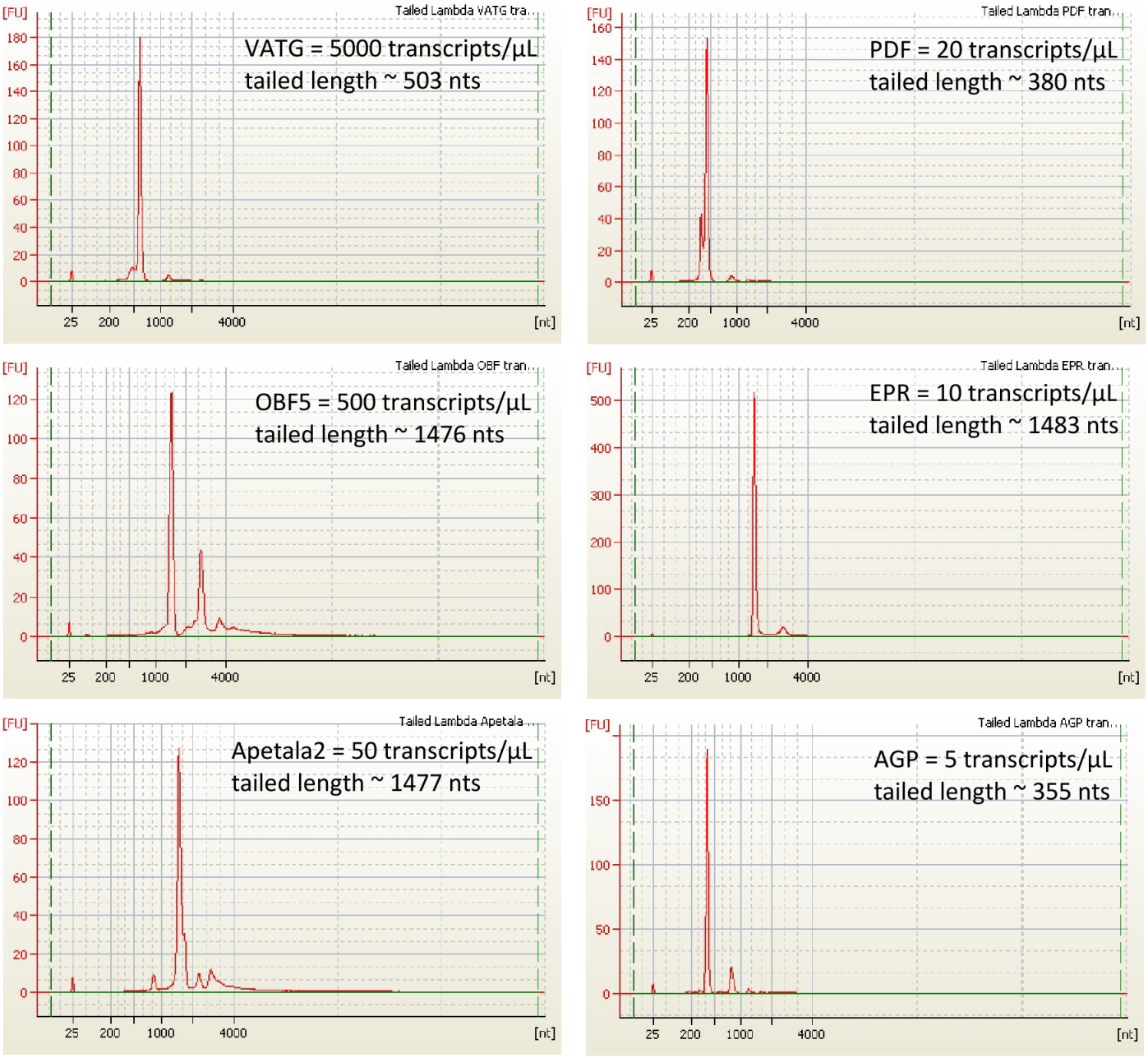
Supplementary Figure 25: Accuracy of estimation of the ratio between the expression values of two genes in a cell pool as a function of the single molecule capture probability. A pool of 1000 cells, average of 100,000 mRNAs per cell. Genes were split into groups according to their expression levels (step size of 1 on a log_2 scale, shown on each axis) and the fraction of gene pairs $\{A, B\}$ for which $R_{AB} < 0.5$ was calculated, where

$$R_{AB} = \frac{\mathrm{FPKM}_A^{pool}}{\mathrm{FPKM}_A^{pool} + \mathrm{FPKM}_B^{pool}} \Big/ \frac{\mathrm{FPKM}_A^{bulk}}{\mathrm{FPKM}_A^{bulk} + \mathrm{FPKM}_B^{bluk}}$$

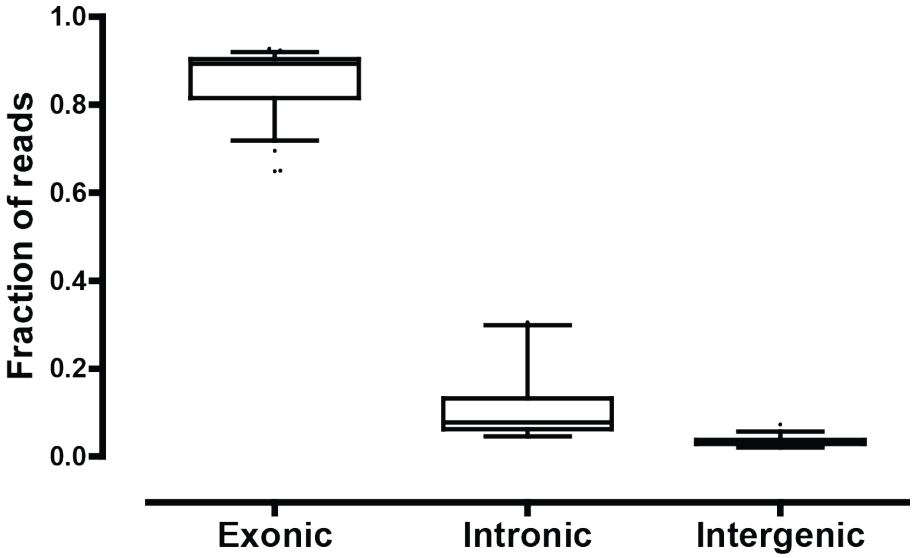
and $\mathrm{FPKM}_A^{bulk} < \mathrm{FPKM}_B^{bulk}$. Empty cells contain no gene pairs with the indicated expression values.



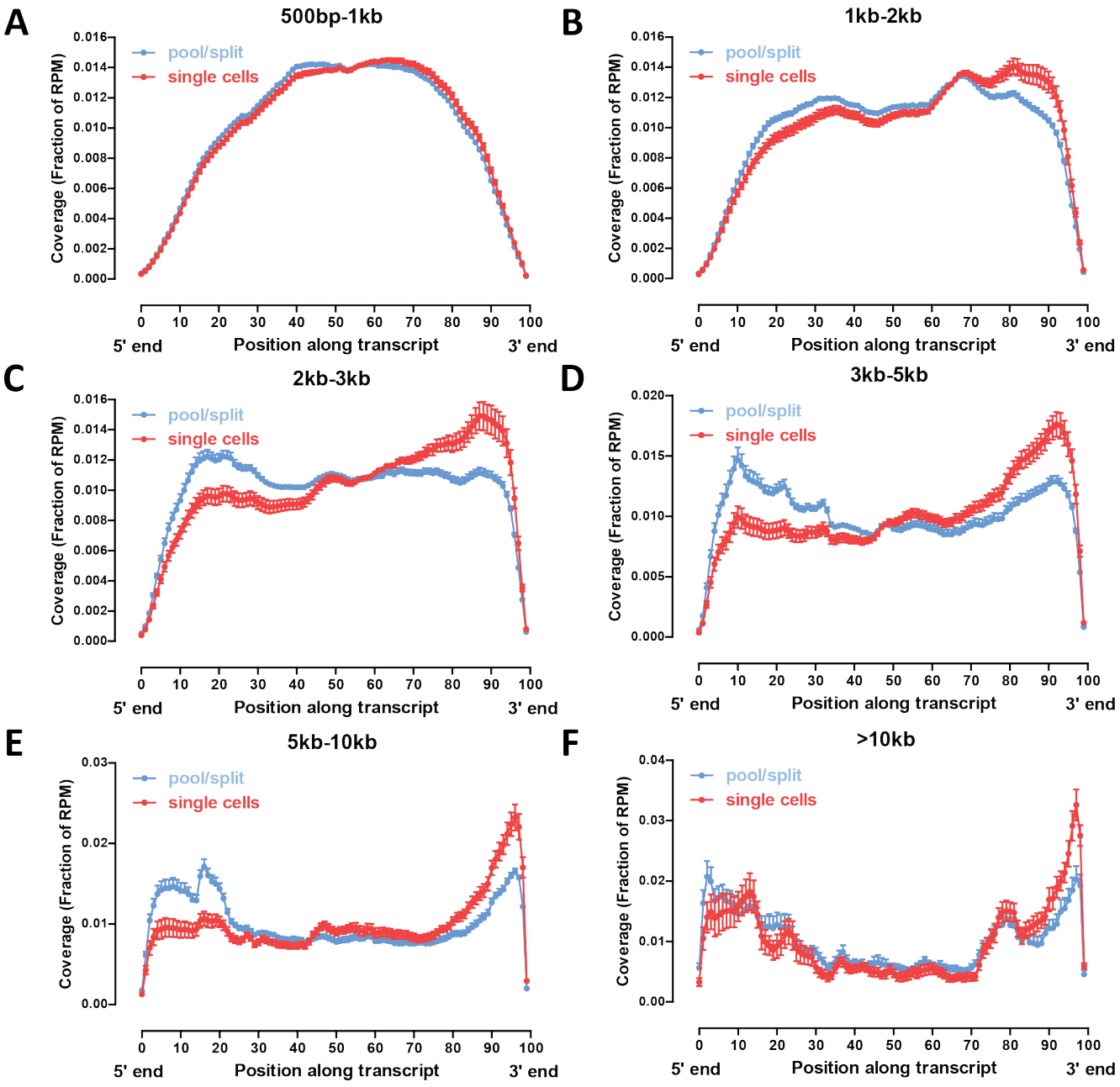
Supplementary Figure 26: Outline of the single-cell SMART-seq RNA-seq library generation workflow.



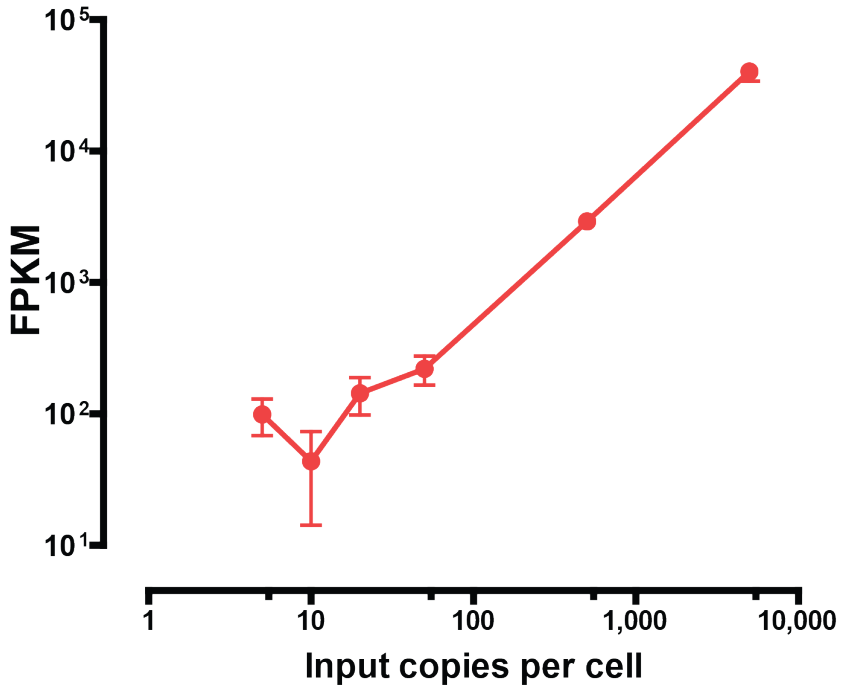
Supplementary Figure 27: Bioanalyzer plots for the transcript quantitation spike-in controls. The nominal transcript length (cloned sequence plus the polyA region) and the number of transcripts included in the cocktail of standards are indicated. The plots indicate that the majority of transcription product is the expected length.



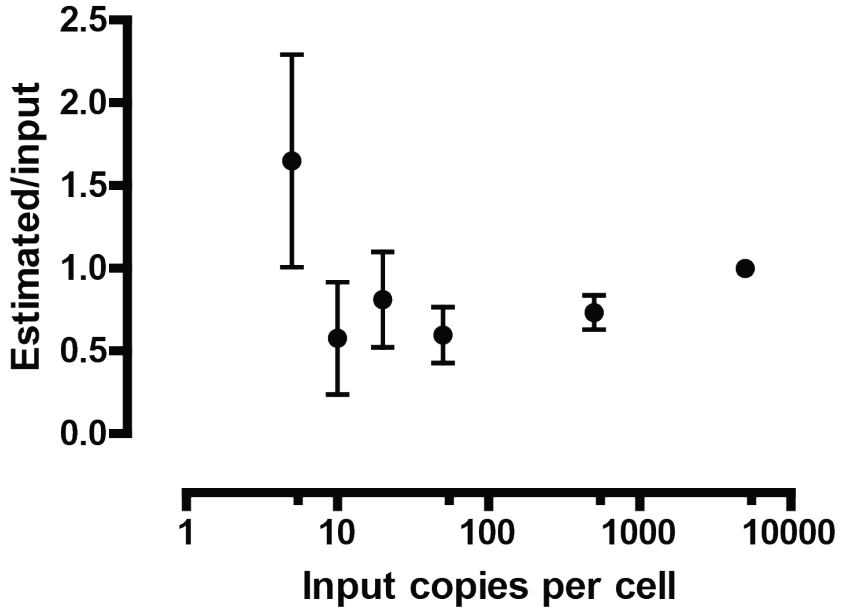
Supplementary Figure 28: Efficiency of enrichment for polydenylated messages. Shown is the fraction of reads mapping to exons, introns or integenic space (GENCODE V13 annotation).



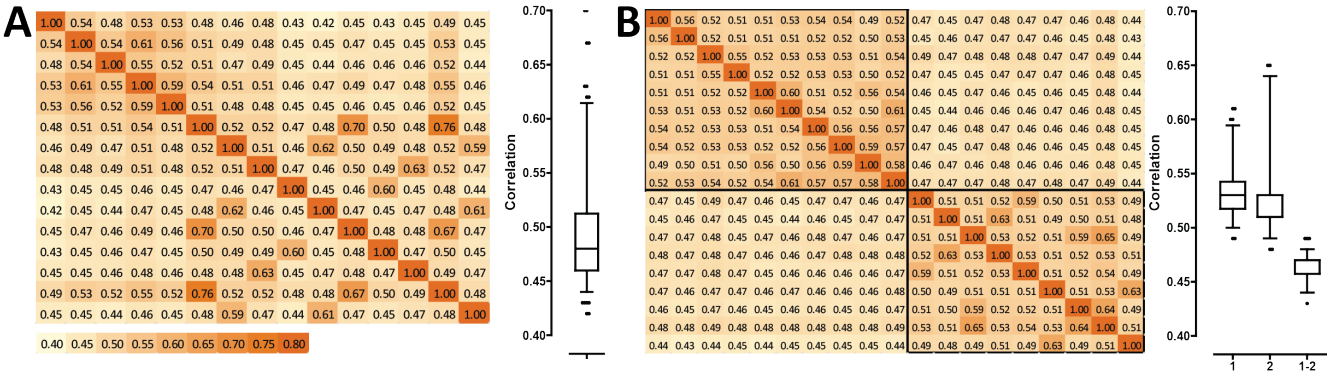
Supplementary Figure 29: Uniformity of transcript coverage as a function of transcript length. Shown is the average coverage along the length of an mRNA for single cells and pool/split experiments. Only mRNAs with a single annotated isoform in the refSeq annotation set and within the indicated length limits were included.



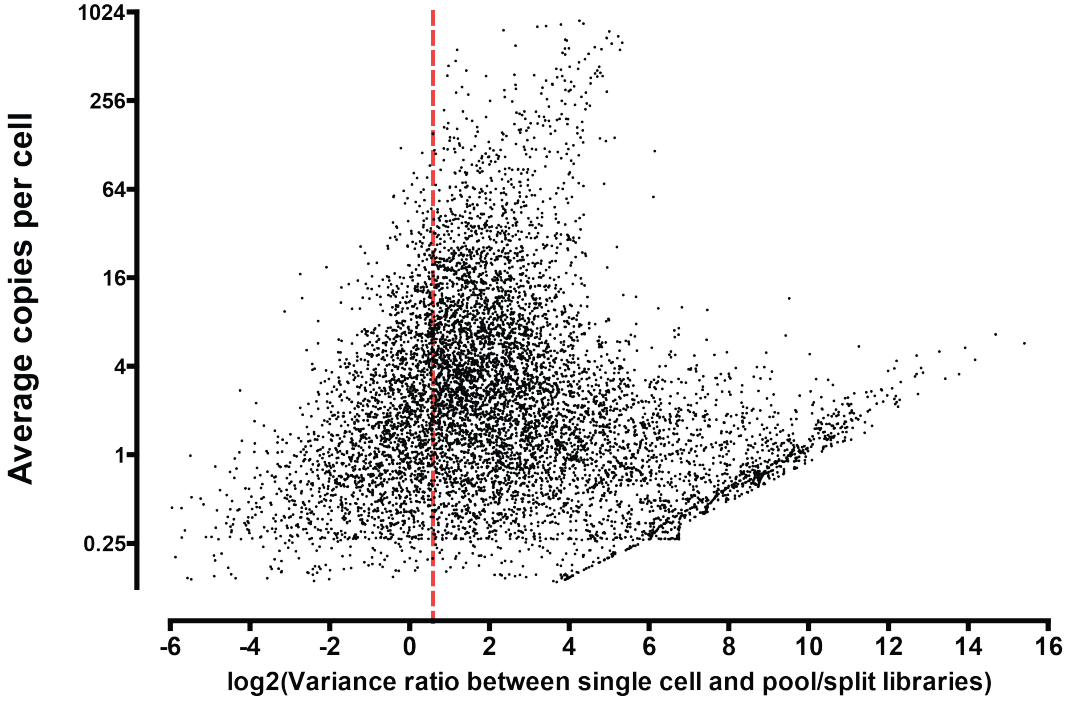
Supplementary Figure 30: Correspondence between initial spike-in amounts and spike abundance in sequenced libraries as measured in FPKMs. Error bars represent the standard error of the mean.



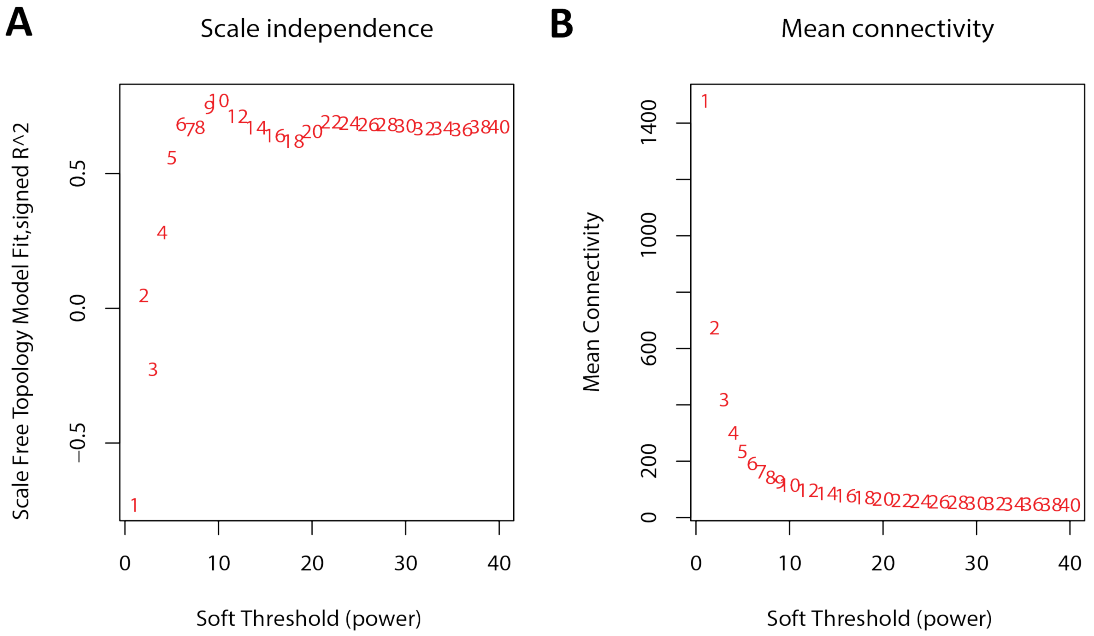
Supplementary Figure 31: Stability of copies per cell estimation. Spike-in sequences of known abundance (Supplementary Table 2) were added to each reaction prior to library building. A linear regression calibration was derived based on RPKM/FPKM values calculated for each. Shown is the average ratio of estimated copies per cell and the actual spiked in copies per cell for these spike sequences. Error bars represent the standard error of the mean.



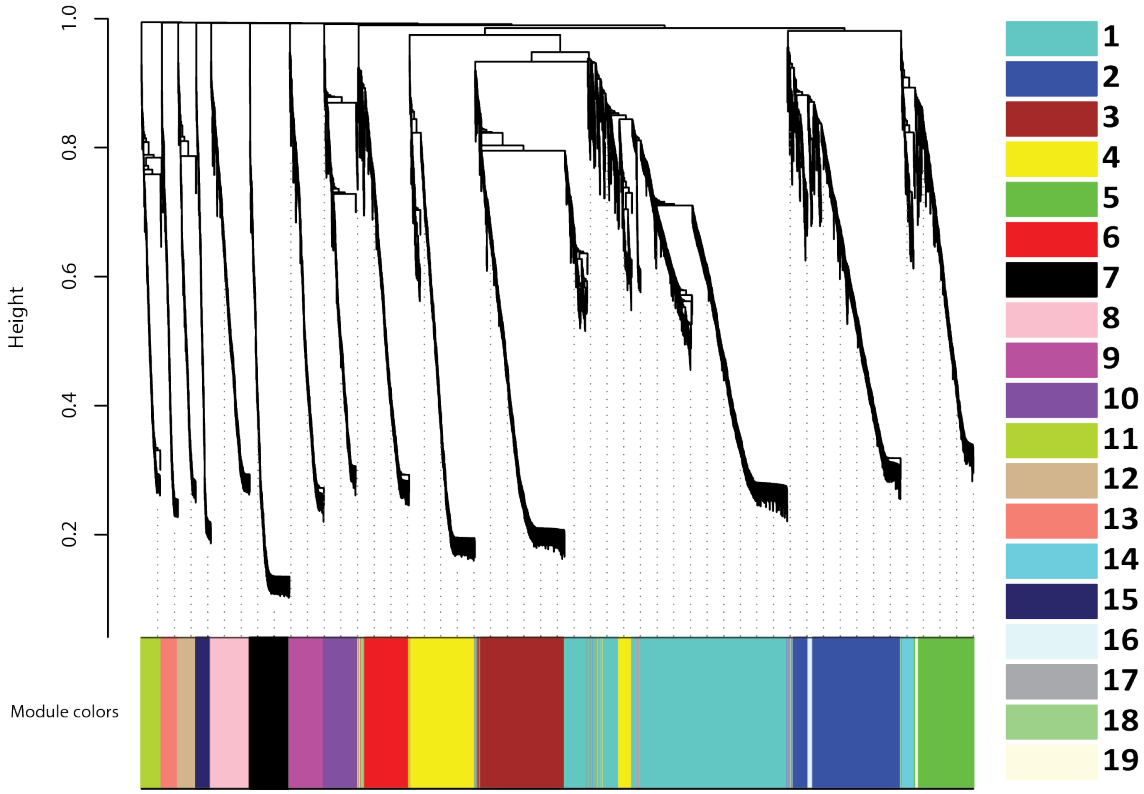
Supplementary Figure 32: Spearman correlations between single cell and pool/split libraries. (A) Single cells; (B) Pool/splits.



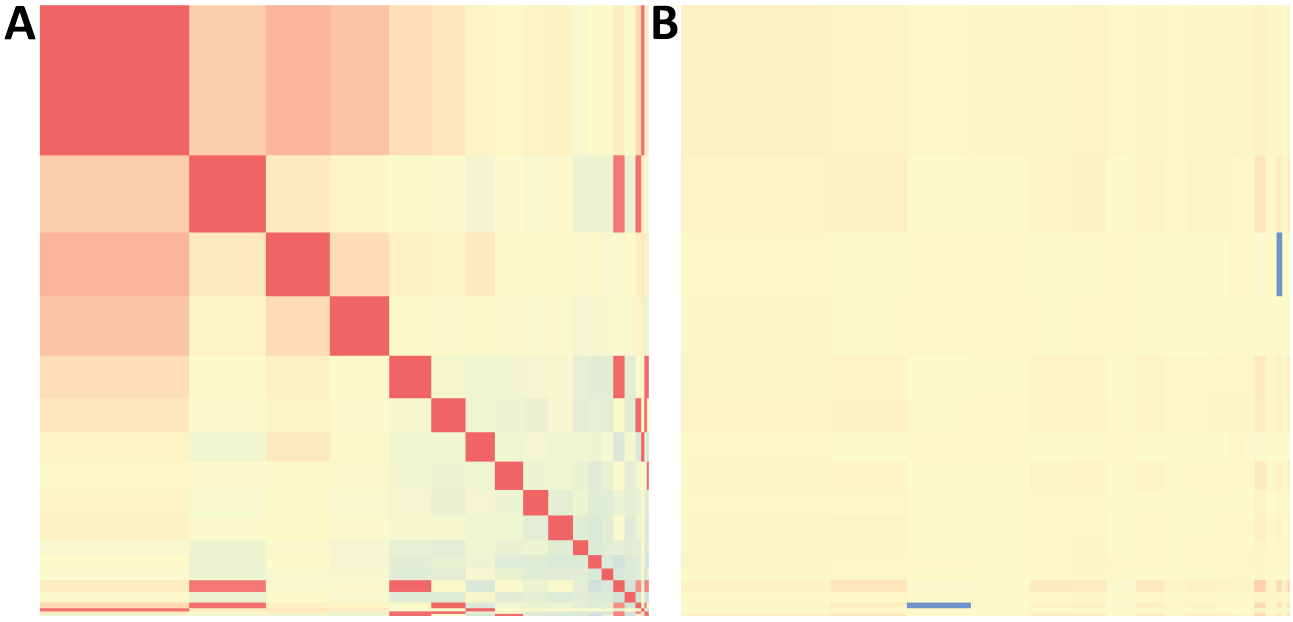
Supplementary Figure 33: Ratio of the variance of single cell and pool/split libraries vs. average estimated number of mRNA molecules. The vertical line corresponds to a variance ratio of 1.5. Genes with a variance ratio higher than 1.5 were retained for network construction. Most genes with a lower ratio (and correspondingly high variance in pool/split libraries) have a relatively low average estimated number of mRNA molecules per cell.



Supplementary Figure 34: Optimization of the soft threshold parameter for constructing weighted correlation gene expression network. (A) Scale independence (B) Mean connectivity. A value of $\beta = 6$ was used for network construction.

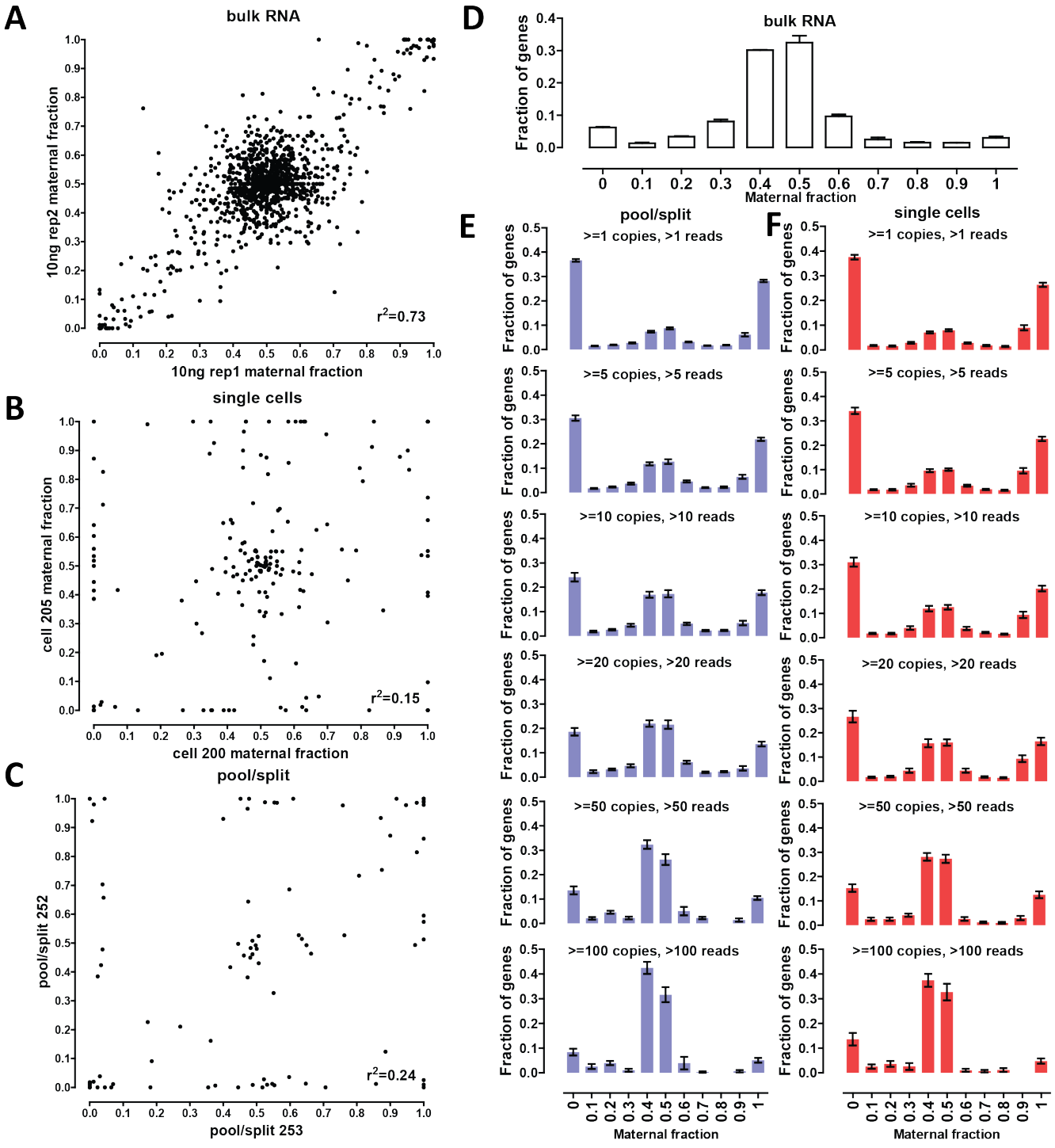


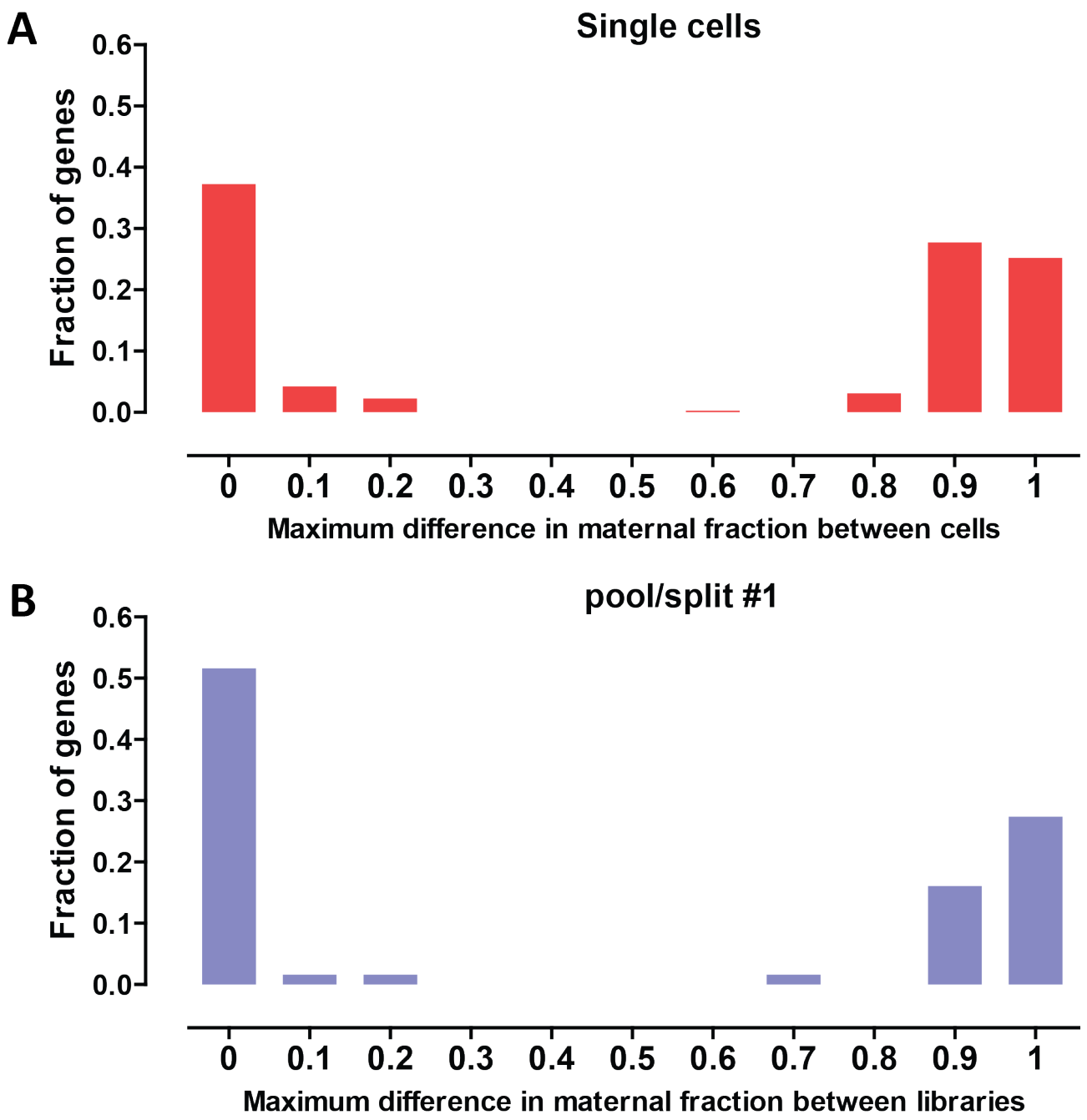
Supplementary Figure 35: Cluster dendrogram of gene coexpression modules derived from single GM12878 cells..



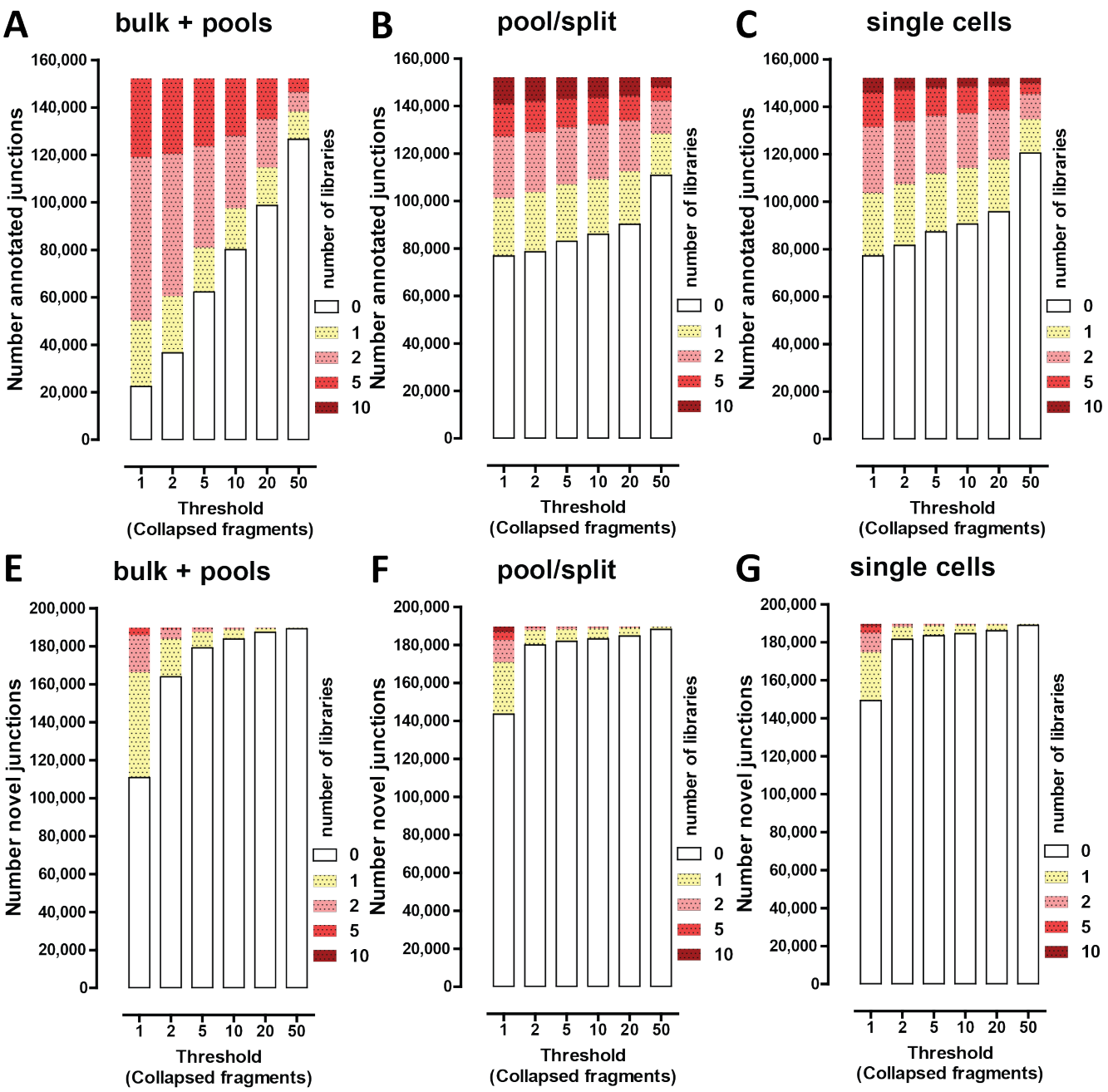
Supplementary Figure 36: Average correlation within and between coexpression modules in single cells and pool/splits. Modules are sorted by decreasing size. (A) Single cells. (B) Pool/splits
.

Supplementary Figure 37 *(following page)*: Allele-biased expression at the single-cell level. (A,B and C) Correlation between allele bias between 10ng bulk RNA replicates (A), between two individual single cells (B) and between two pool/split libraries (C). Shown is the maternal fraction of reads for genes with at least 15 reads covering heterozygous positions for 10ng libraries and for genes with at least 10 reads covering heterozygous positions and expressed at more than 10 copies per cell for single cells and pool/splits. (D). Distribution of allele bias in bulk RNA samples ($\geq$15 reads covering positions). (E and F). Distribution of allele bias as a function of the read and copies threshold in single cell (E) and pool/split (F) libraries.

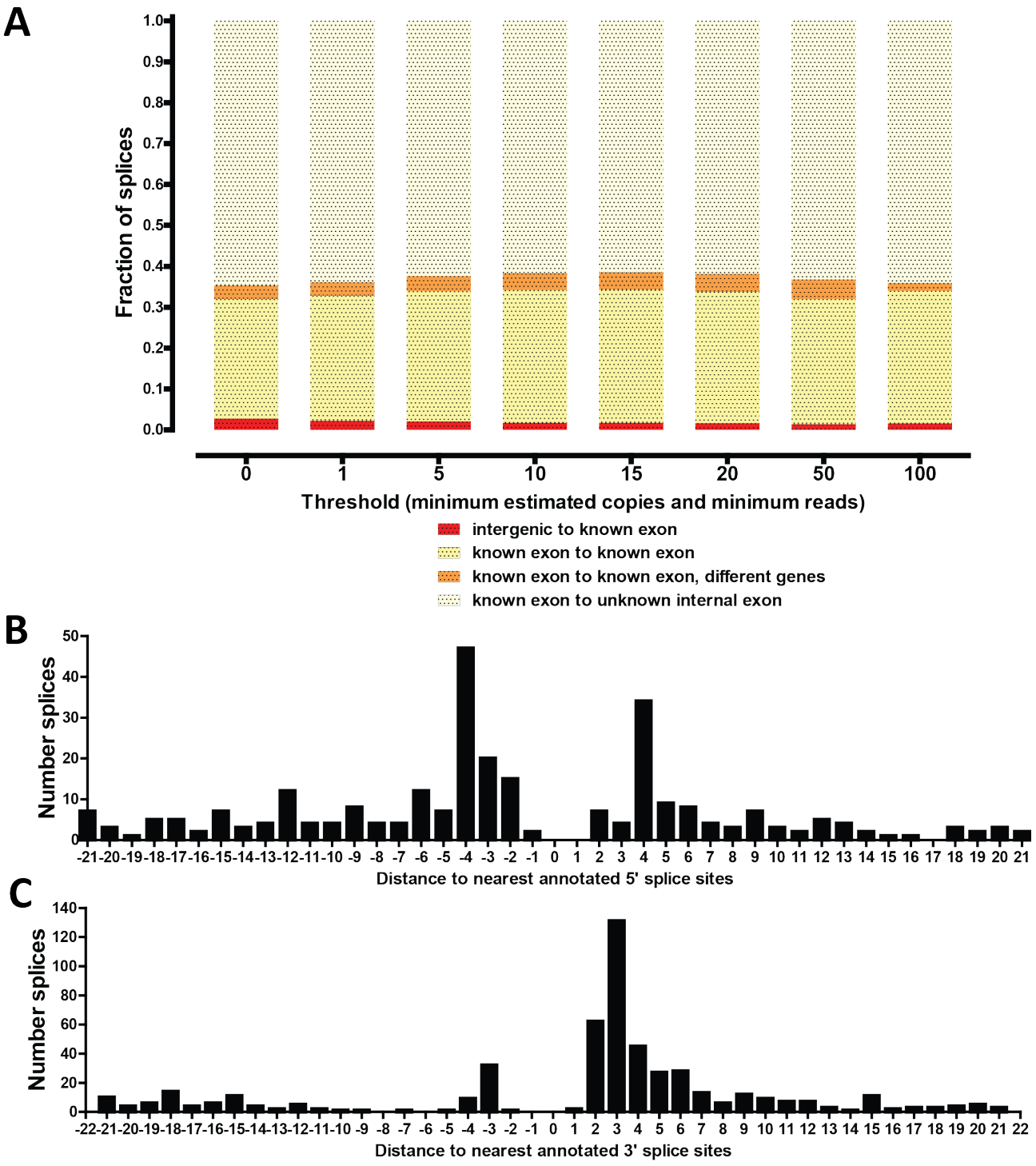

A
bulk RNA
10ng rep2 maternal fraction
10ng rep1 maternal fraction
r2=0.73
B
single cells
cell 205 maternal fraction
cell 200 maternal fraction
r2=0.15
C
pool/split
pool/split 252
pool/split 253
r2=0.24
D
bulk RNA
Fraction of genes
Maternal fraction
E
pool/split
>=1 copies, >1 reads
>=5 copies, >5 reads
>=10 copies, >10 reads
>=20 copies, >20 reads
>=50 copies, >50 reads
>=100 copies, >100 reads
Fraction of genes
Maternal fraction
F
single cells
>=1 copies, >1 reads
>=5 copies, >5 reads
>=10 copies, >10 reads
>=20 copies, >20 reads
>=50 copies, >50 reads
>=100 copies, >100 reads
Fraction of genes
Maternal fraction



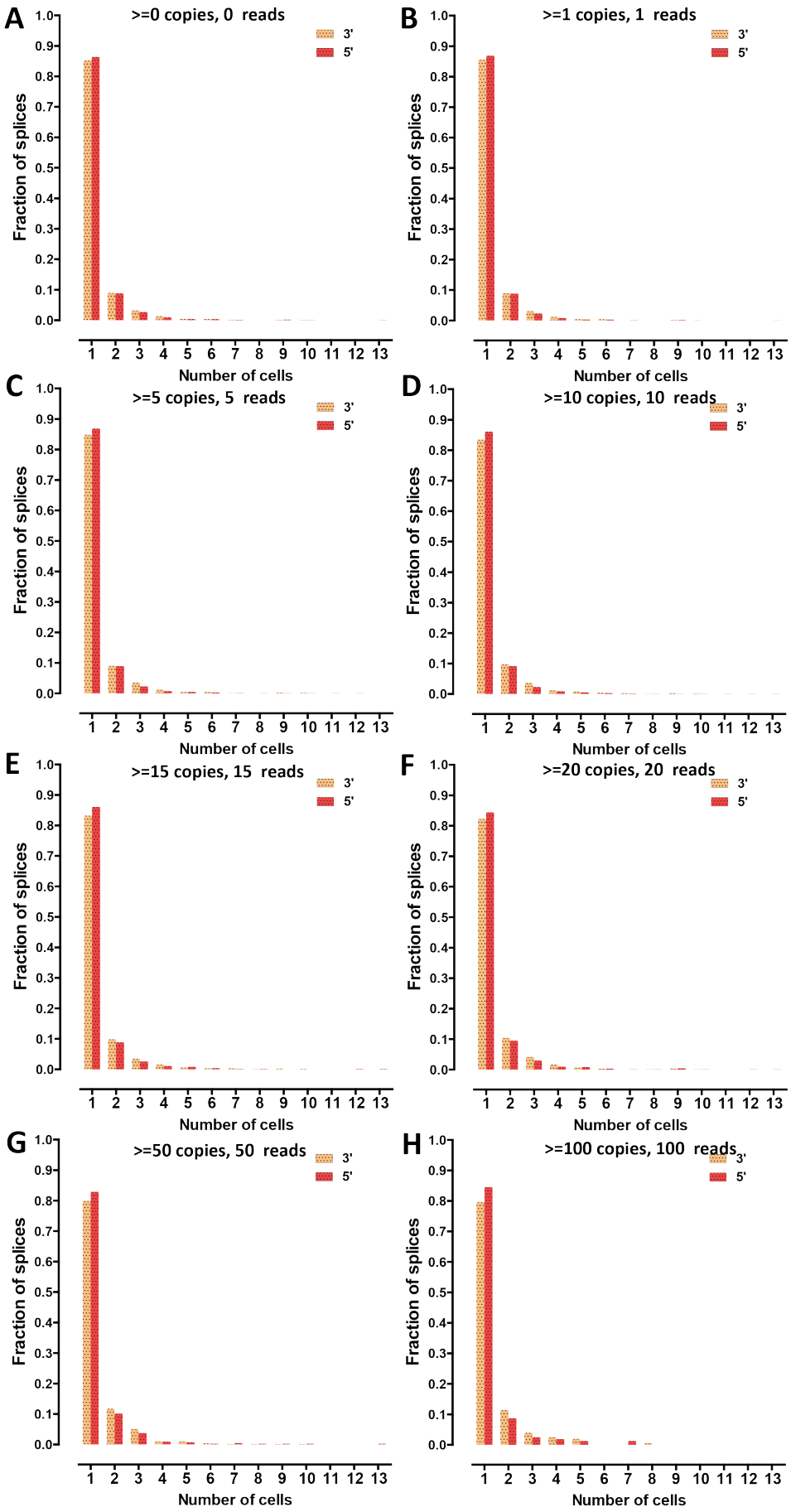
Supplementary Figure 38: Changes in allele expression bias between individual cells and between individual libraries in pool/split experiment 1. Shown is the maximum difference between the maternal fraction of reads in single-cells (A) and the pool/split (B). Only gene/library pairs for which the ψ score passed all three tests for statistical significance of bias towards one splice (described in Methods) were included



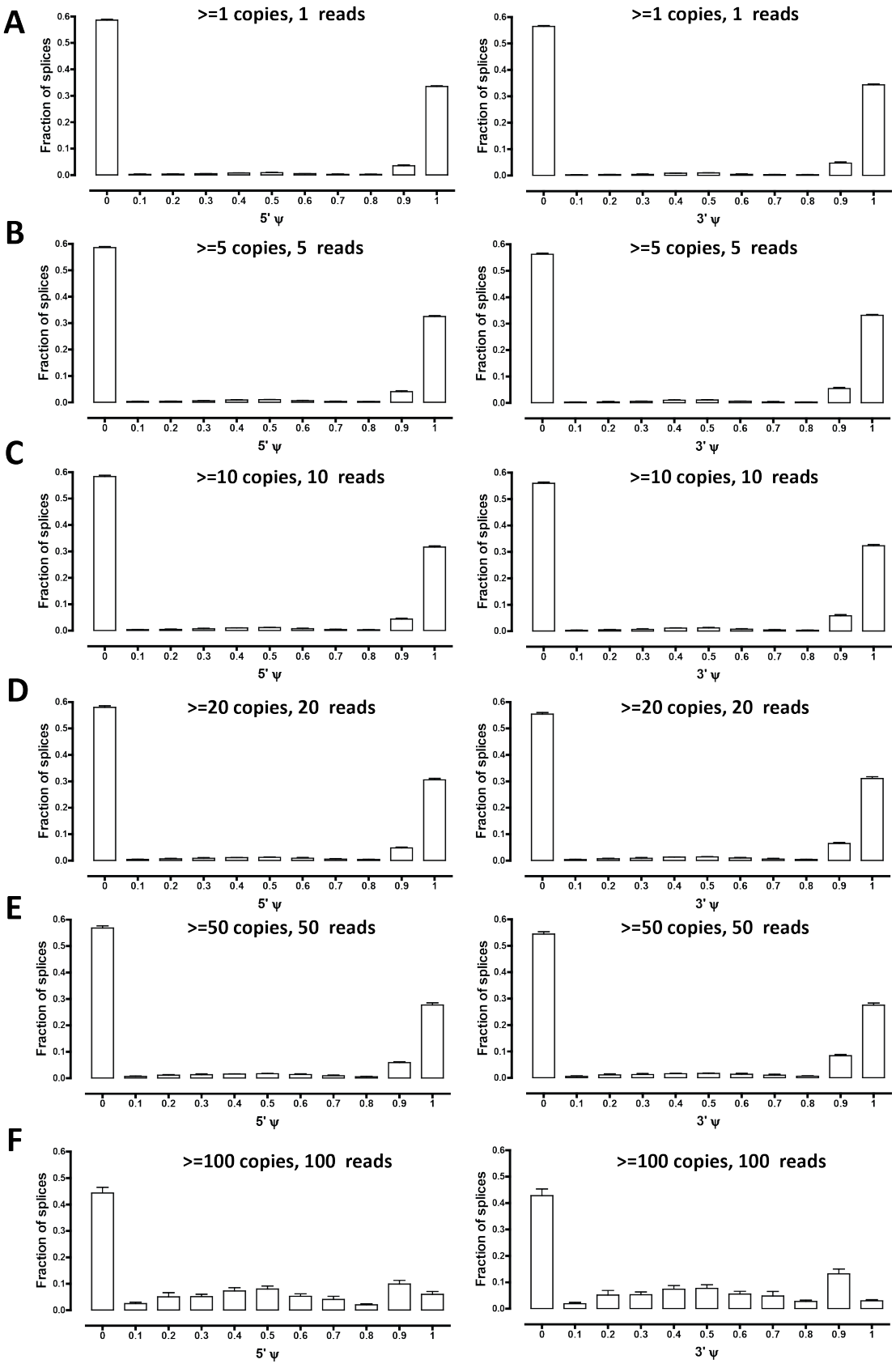
Supplementary Figure 39: Splice junctions detection. The total number of annotated or novel junctions in all libraries is included in each plot and junctions that are not detected in each group of experiments are represented by a white bar. (A, B and C) Annotated junctions in bulk and pool libraries (A), pool/split experiments (B) and single cells (C). (D, E and F) Novel junctions in bulk and pool libraries (D), pool/split experiments (E) and single cells (F). Shown are all junctions detected in pools, pool/splits or single cells; when a junction is detected in 0 libraries, only the libraries in the indicated group are referred to.



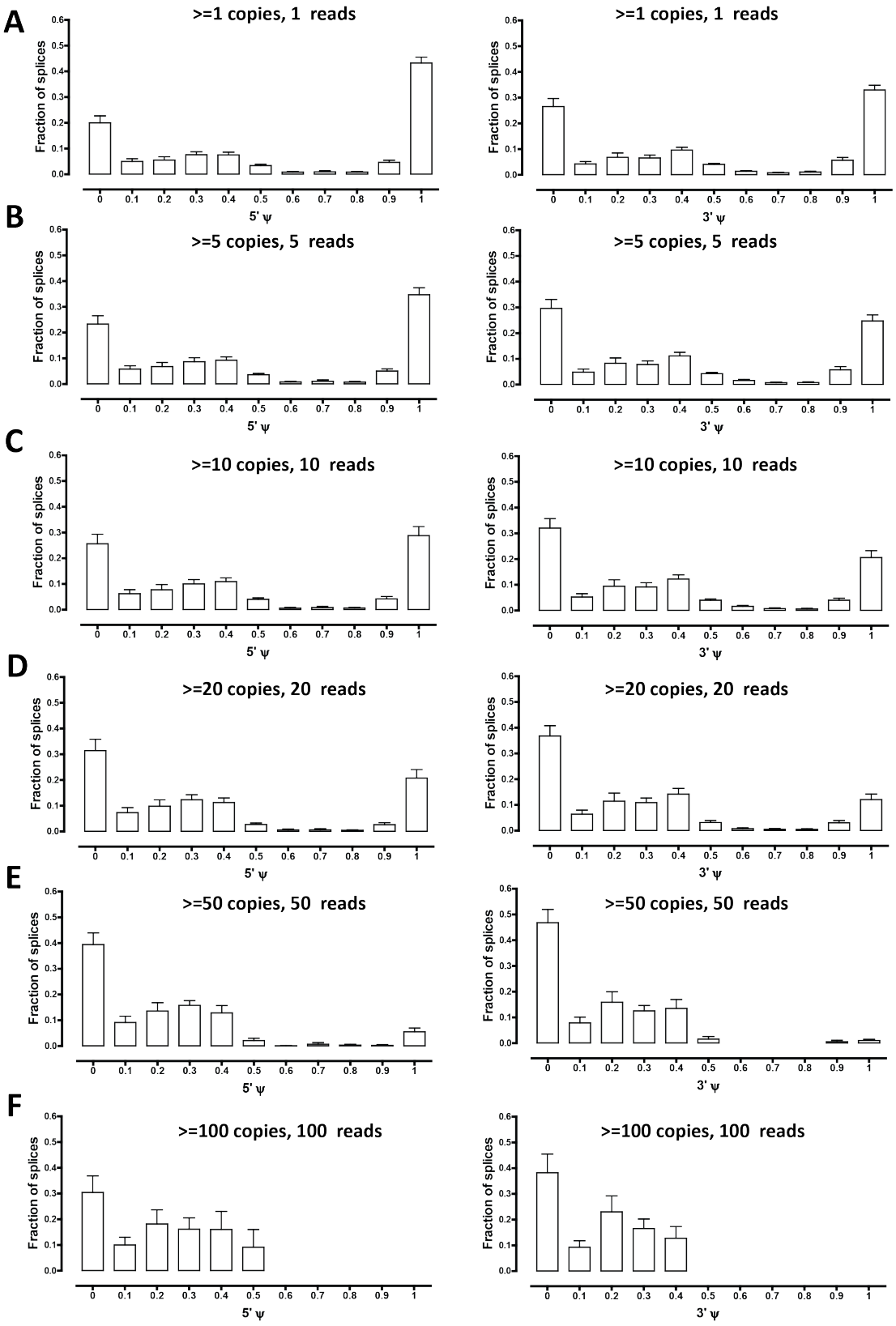
Supplementary Figure 40: Relationship of novel splice junctions to annotation. (A) Relation to annotated exons. The detection threshold (in both estimated number of copies and reads mapping to heterozygous positions) was varied as shown and the fraction of junctions belonging to each class was calculated. (B) Distance of the donor site to the nearest annotated 5' splice site. (C) Distance of the acceptor site to the nearest annotated 3' splice sites. All detected junctions were included in (B) and (C).



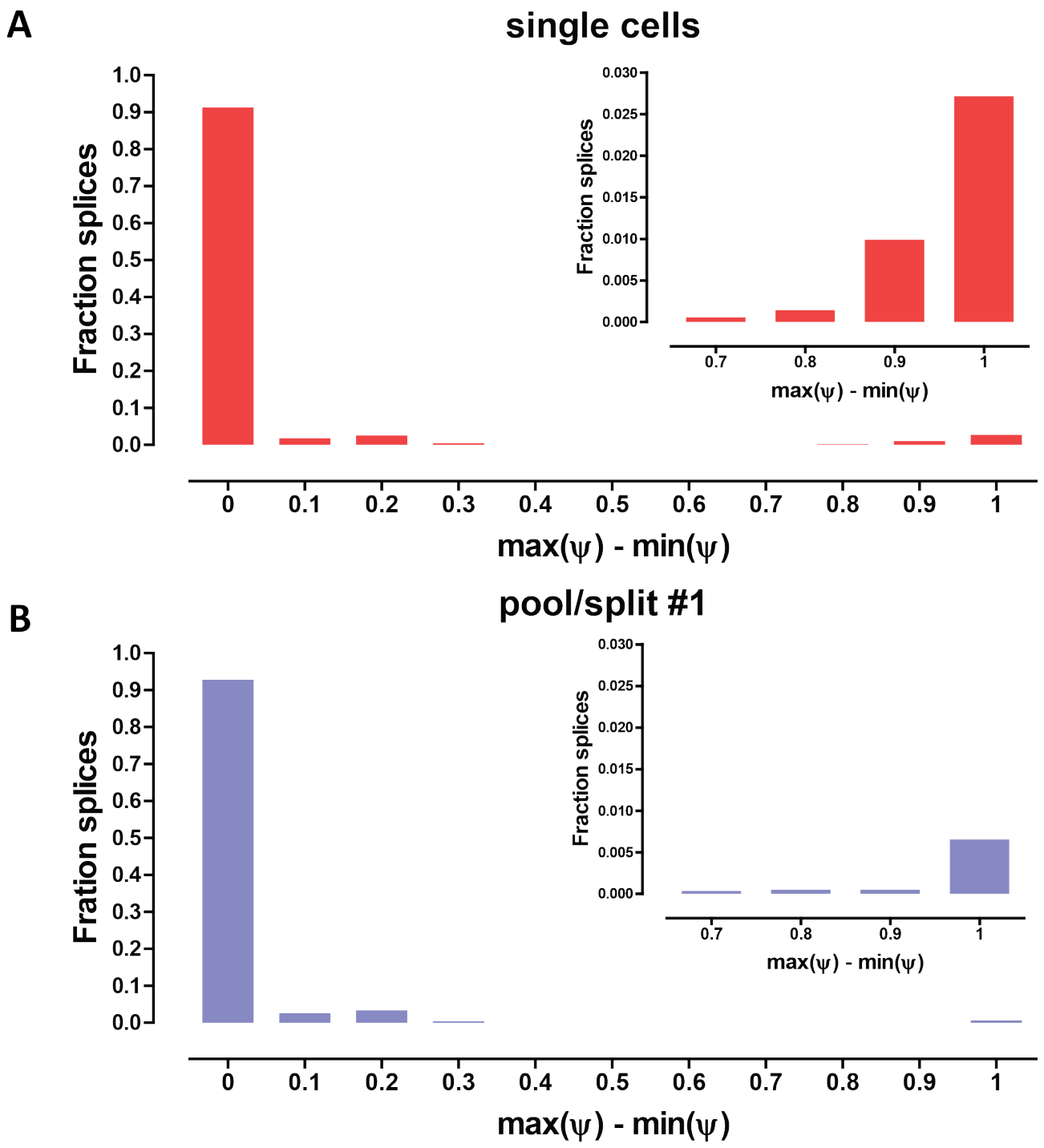
Supplementary Figure 41: Number of cells in which a novel junctions is detected. The detection threshold (in both estimated number of copies and reads mapping to heterozygous positions) was varied as shown and the fraction of splices detected in a give number of cells plotted.



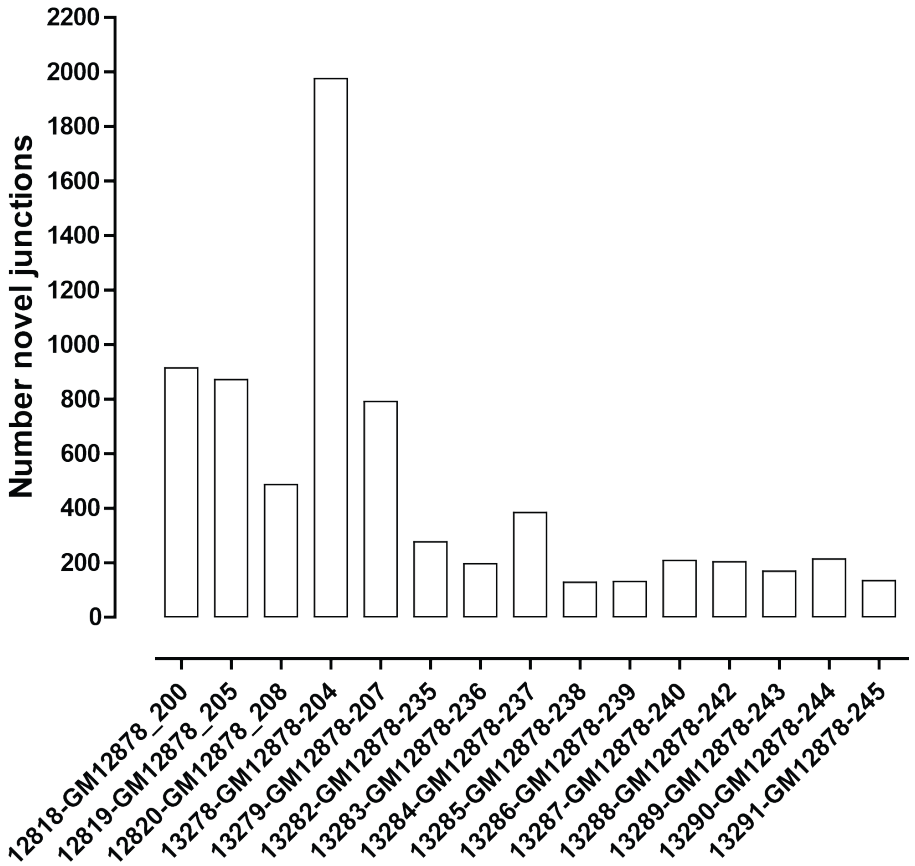
Supplementary Figure 42: Distribution of 5' and 3' ψ scores as a function of the expression and splice junction spanning reads threshold.



Supplementary Figure 43: Distribution of 5' and 3' ψ scores as a function of the expression and splice junction spanning reads threshold for novel splice junctions. Only novel splice junctions connecting at least one of the donor or acceptor site for which is annotated are included.

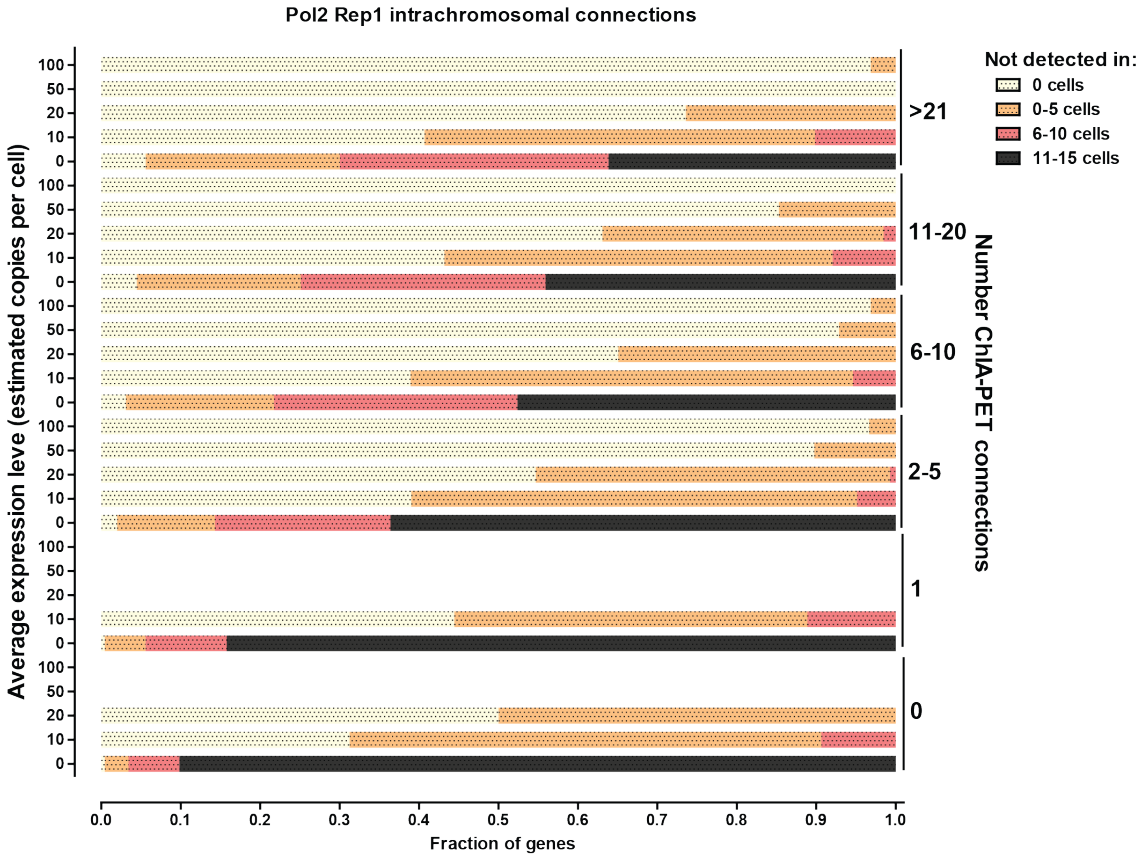


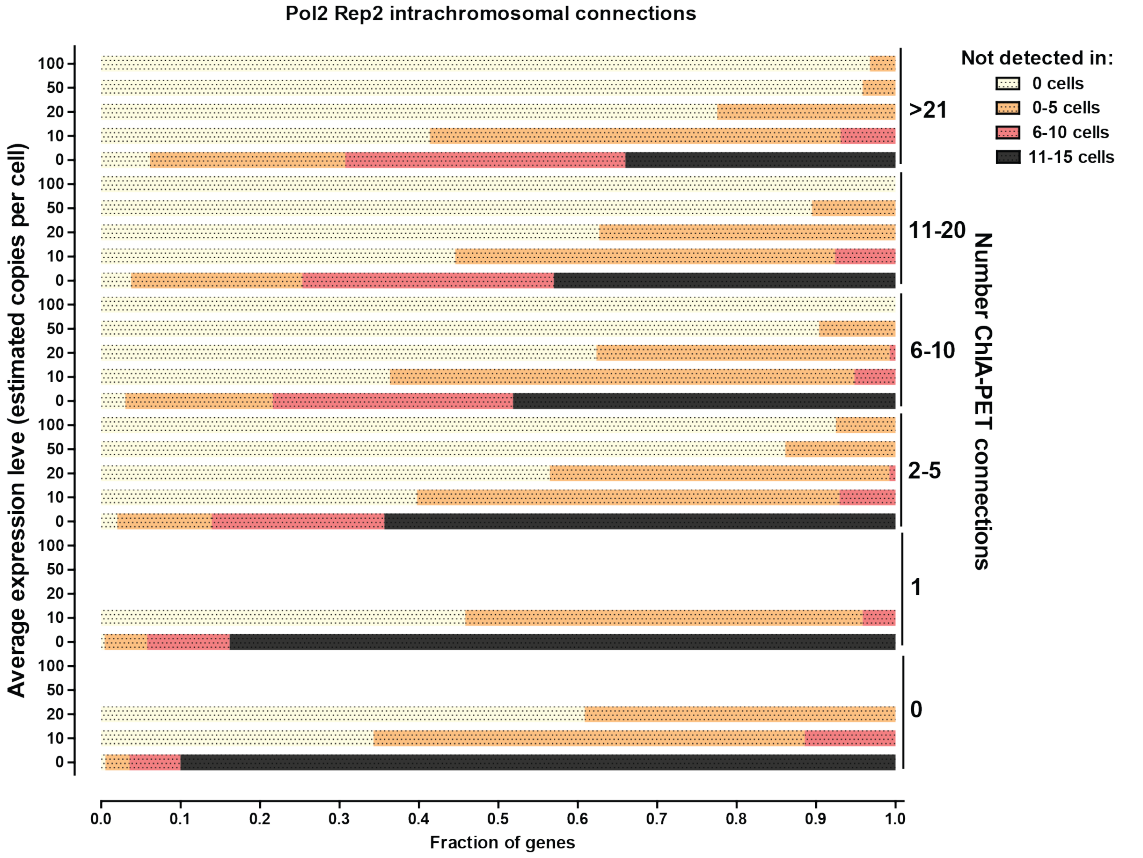
Supplementary Figure 44: Major splice site switches between individual cells. Shown is the maximum difference between ψ scores in single-cells (A) and individual libraries in pool/split experiment 1 (B). Only gene/library pairs for which the ψ score passed all three tests for statistical significance of bias towards one splice (described in Methods) were included

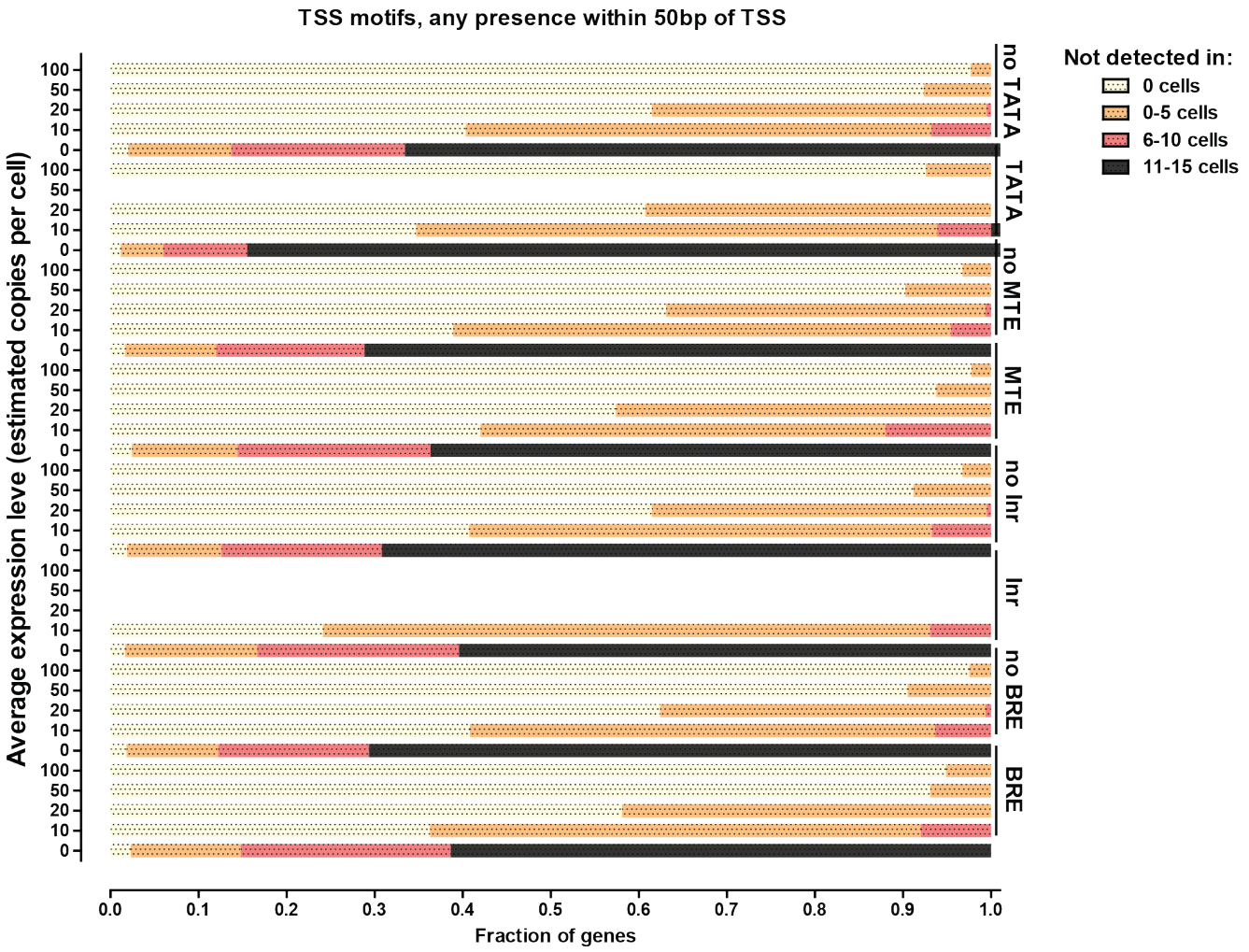


Supplementary Figure 45: Number of novel splice junctions (connecting to annotated donor and/or acceptor sites) detected in individual cells.

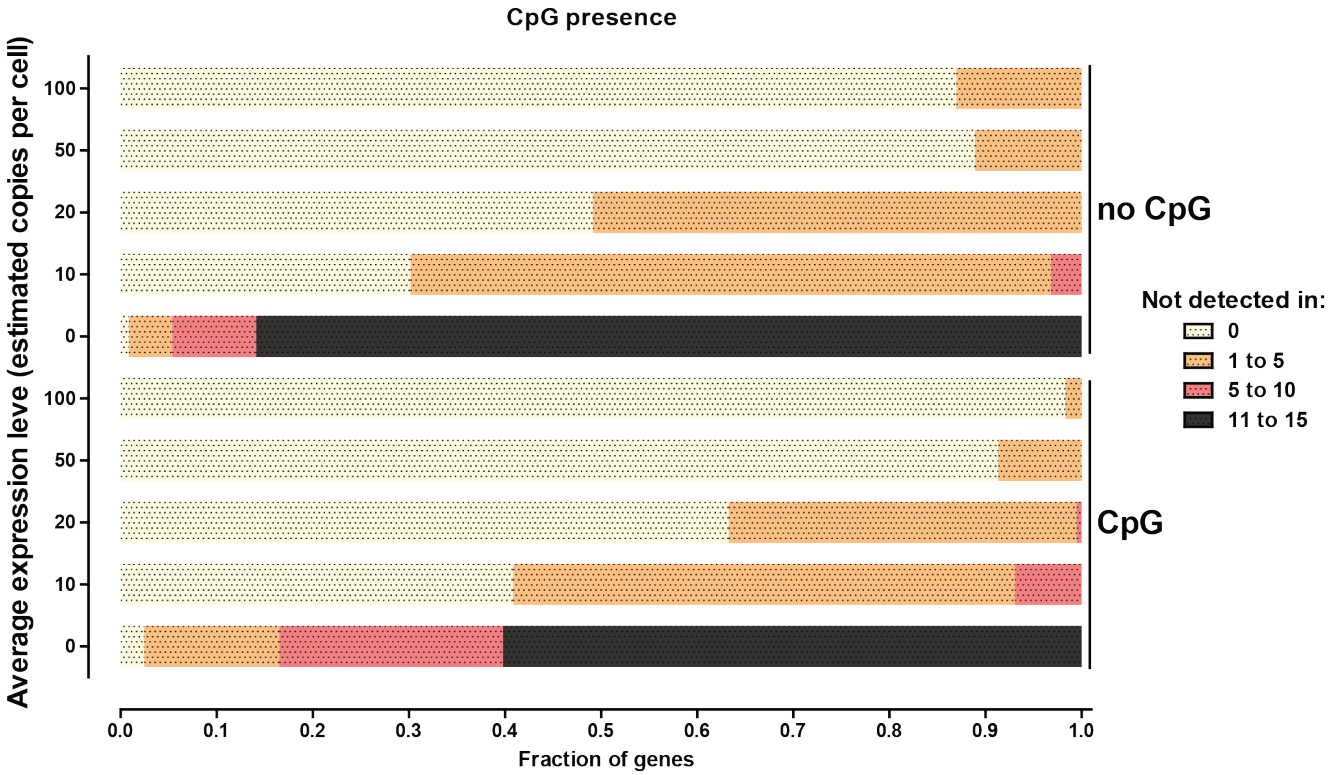
Supplementary Figure 46 ***(following page)*****: Relation between the long-range chromosomal element connectivity of TSSs and gene expression stochasticity**. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the number of ENCODE ChIA-PET connections to TSSs in K562 cells (replicates 1 and 2). K562 was used as the closest cell line to GM12878 for which such data is currently available; ChIA-PET connections were downloaded from the UCSC Genome Browser. Within each group of genes defined by the number of ChIA-PET connections, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.


Pol2 Rep1 intrachromosomal connections
Average expression leve (estimated copies per cell)
Fraction of genes
Number ChIA-PET connections
>21
11-20
6-10
2-5
1
0
Not detected in:
0 cells
0-5 cells
6-10 cells
11-15 cells

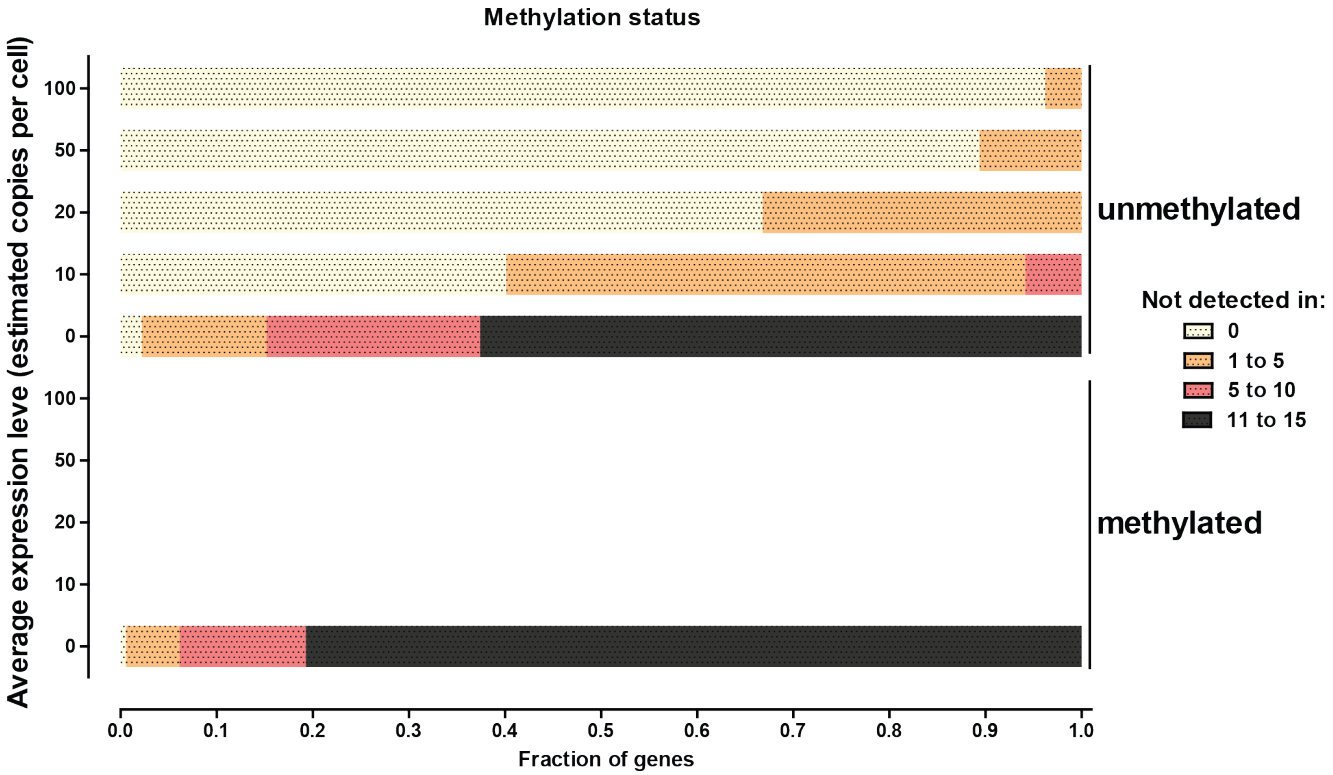

Pol2 Rep2 intrachromosomal connections
Average expression leve (estimated copies per cell)
Fraction of genes
Number ChIA-PET connections
>21
11-20
6-10
2-5
1
0
Not detected in:
0 cells
0-5 cells
6-10 cells
11-15 cells



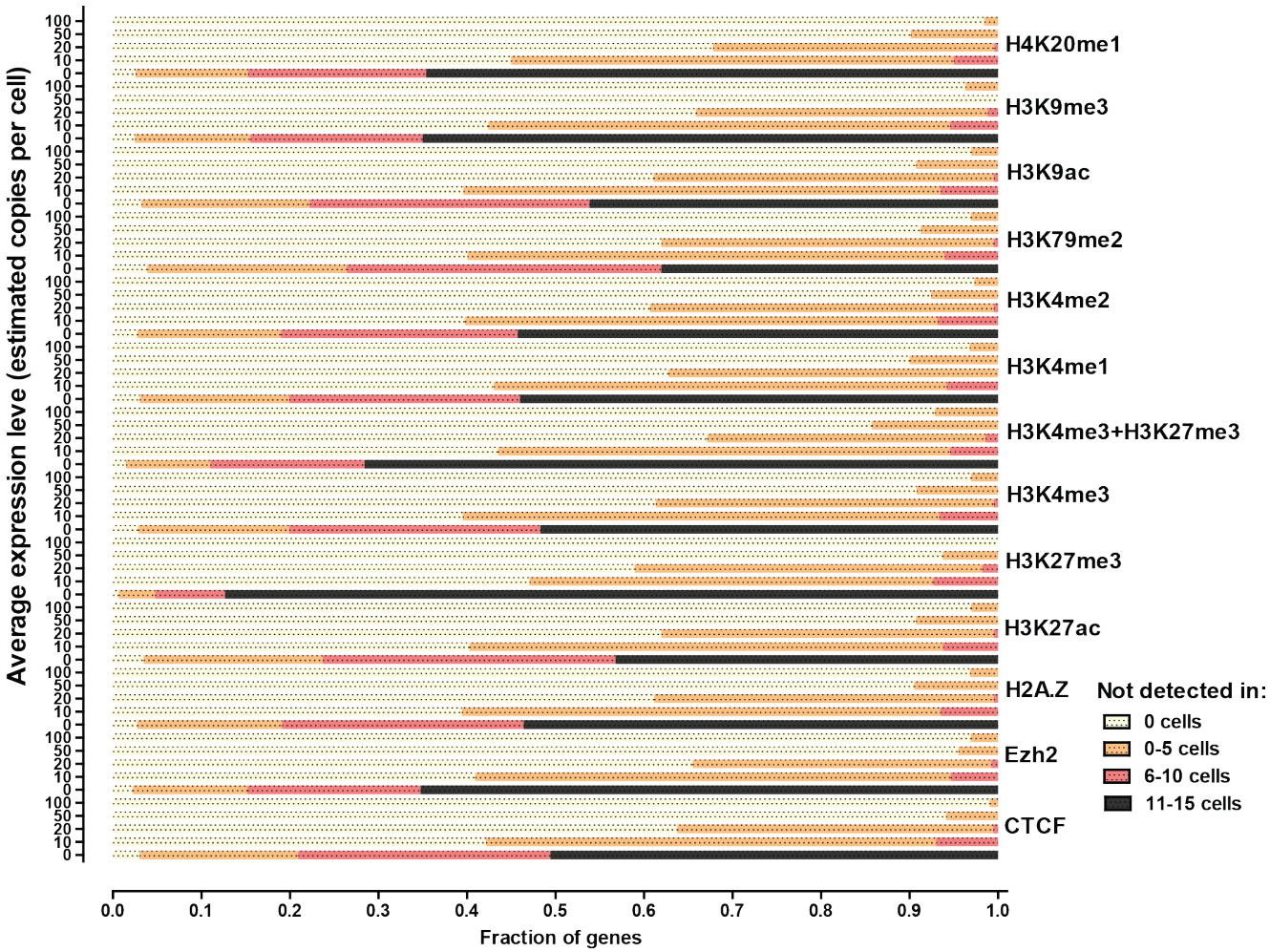
Supplementary Figure 47: Relation between the presence of TSS-associated sequence elements and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the presence or absence of sequence motifs at TSSs (defined by FIMO using position weight matrices obtained from Jin et al., 2006). Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.



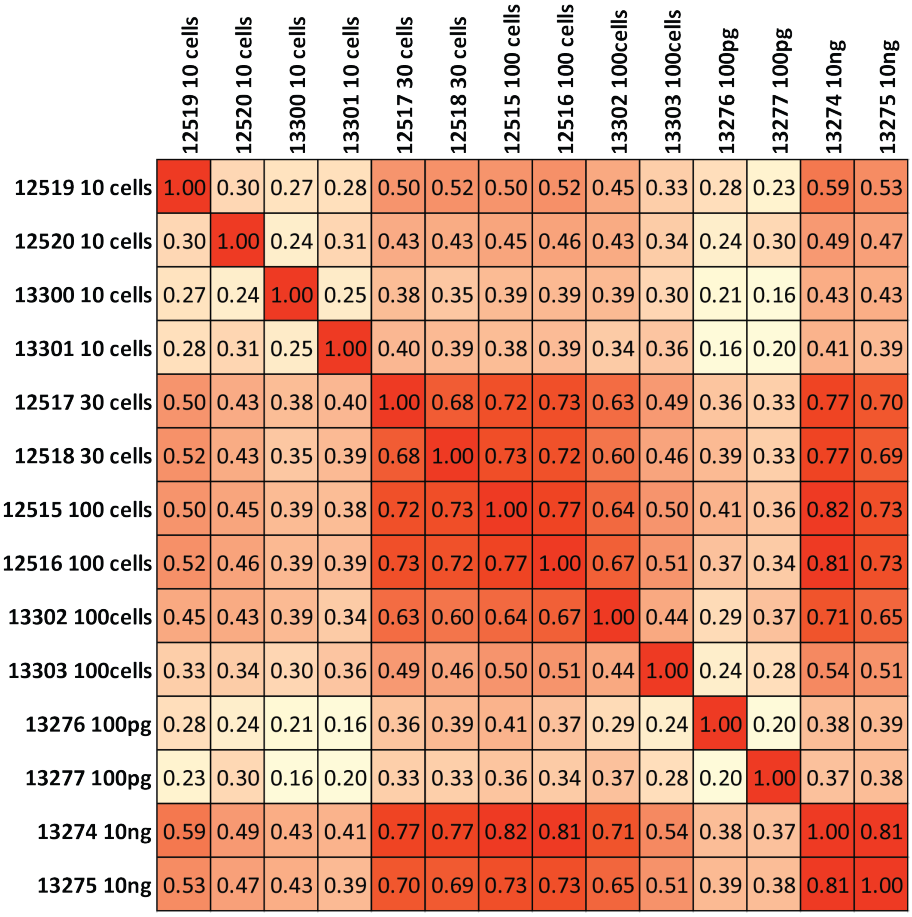
Supplementary Figure 48: Relation between the presence of CpG islands near TSSs and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the presence or absence of CpG islands within 1kb of the TSS. Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.



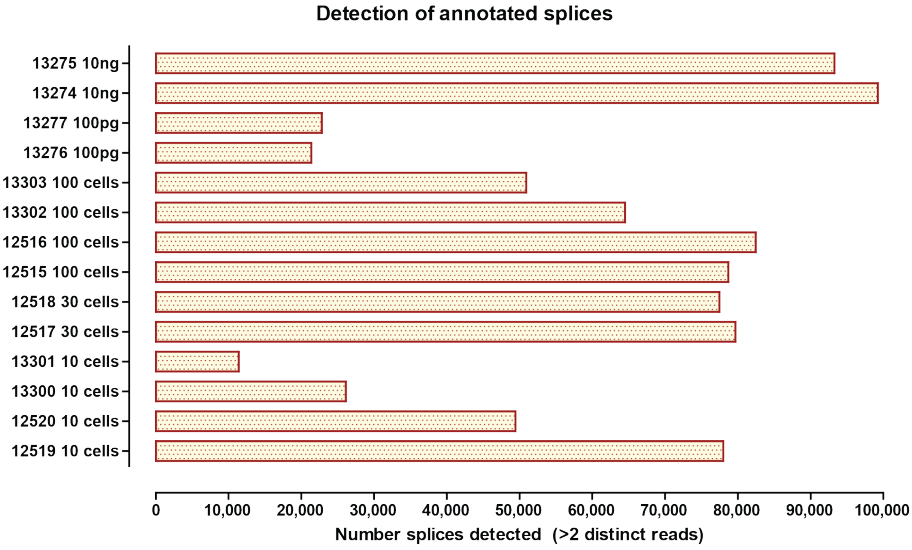
Supplementary Figure 49: Relation between the methylation status of promoters and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the methylation status of their promoters as defined using ENCODE reduced-representation bisulfite sequencing data (RRBS) for the GM12878 cell line from Varley et al., 2013, downloaded from the UCSC Genome Browser. Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.



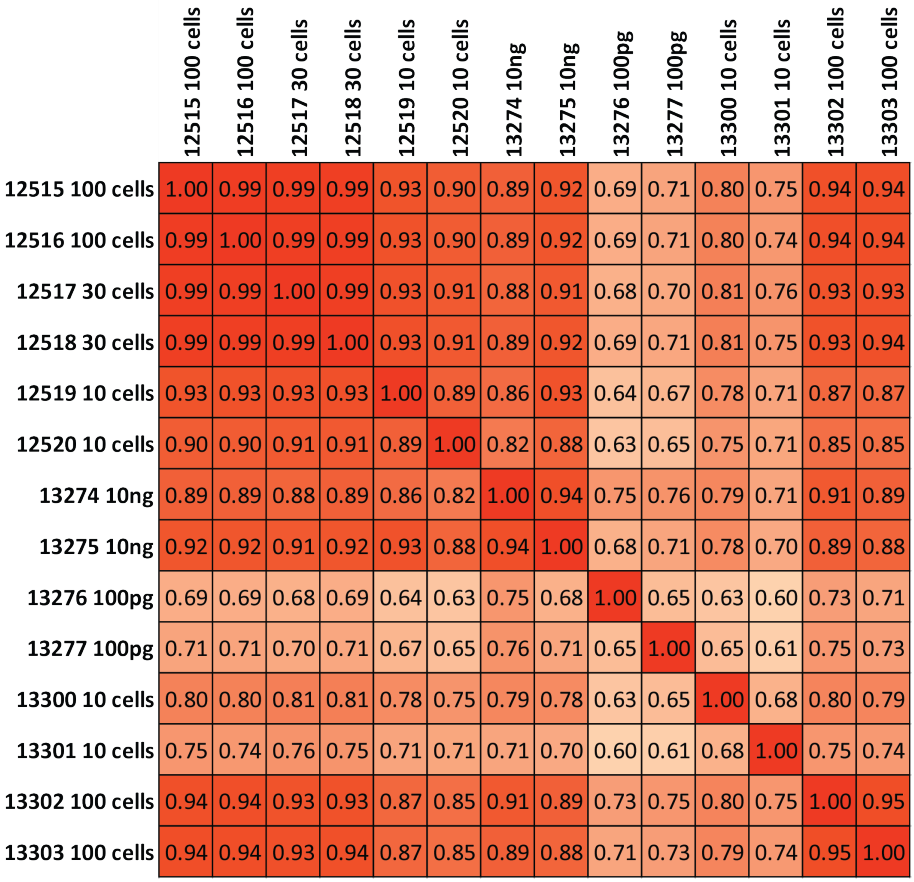
Supplementary Figure 50: Relation between the histone modification status of promoters and expression stochasticity. Shown is the number of genes not detected in 0-5, 6-10 and 11-15 cells as a function of the presence or absence of the various histone marks, the bivalent H3K4me3 + H3K27me3 combination of marks, CTCF and Ezh2 as defined from ENCODE data for the GM12878 cell line using the peak calls available from the UCSC Genome Browser. Within each such group, genes were further split by their average number of estimated copies per cell (where the average was calculated excluding libraries in which the genes were not detected) in order to define directly comparable groups of genes. Subgroups with less than 20 genes were not plotted.



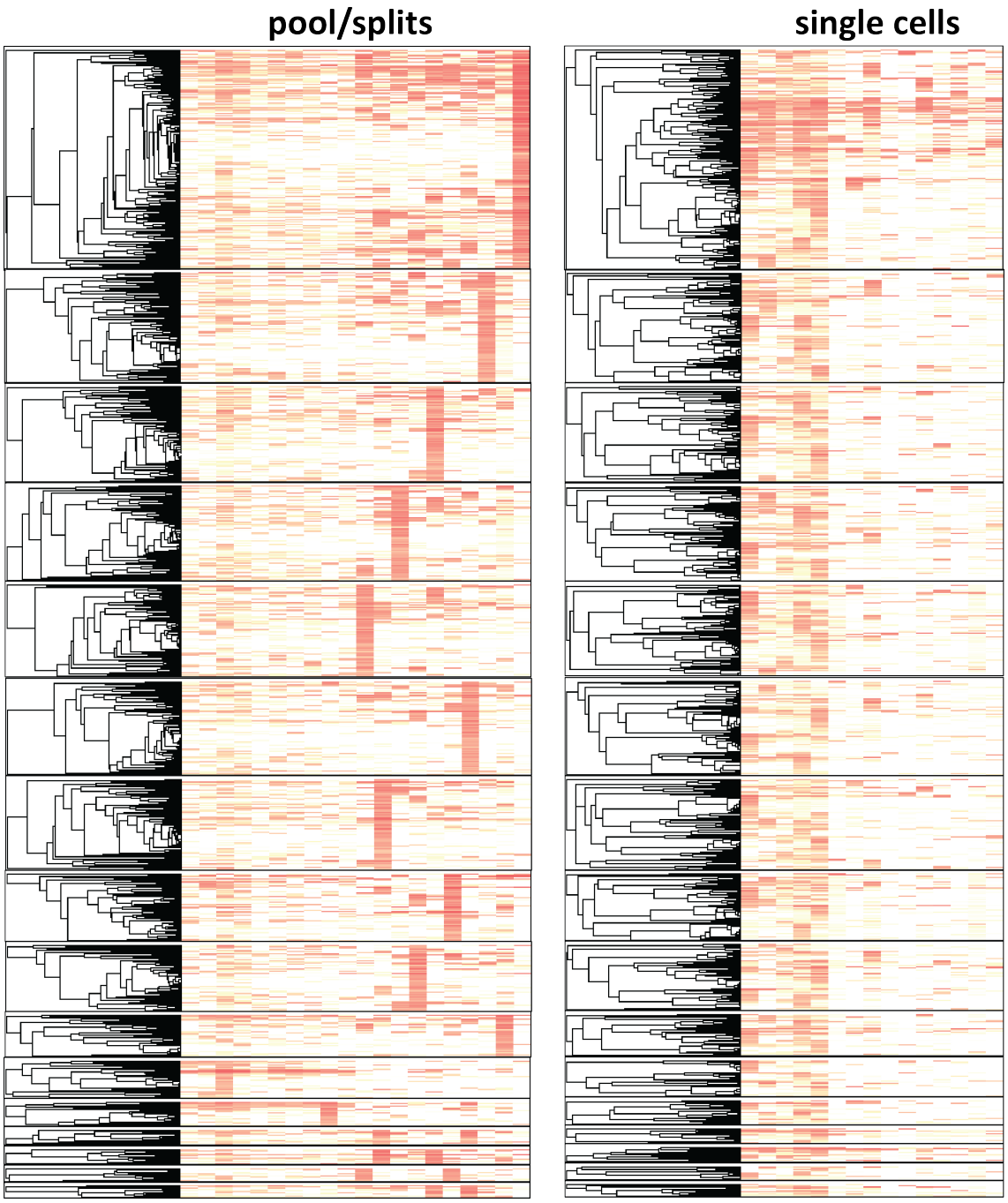
Supplementary Figure 51: Correlation between allelic bias in cell pools of different sizes.



Supplementary Figure 52: Detection of annotated splice junctions in cell pools of different sizes.



Supplementary Figure 53: Correlation between expression estimates based on different cell pools sizes and different amounts of input bulk RNA. Correlation coefficients were calculated on the $log_2(FPKM + 1)$ transform of the FPKM estimates for the refSeq annotation, with only protein coding genes present at ≥ 1 FPKM in at least one library included.



Supplementary Figure 54: Mirrored coexpression analysis of pool/split and single cell datasets. The same analysis presented in Figure 4 was carried out treating pool/splits as single cells and vice versa.

Supplementary Tables

Supplementary Table 1: Read mapping statistics. Note that libraries with numbers lower than 12543 used a different spike-in cocktail than other libraries and the correspondence between initial spike-in amounts and final FPKM scores in the sequenced libraries was poor. For this reason, we excluded those libraries from analyses based on estimating absolute transcript abundances in copies per cell; we are however making the data publicly available as we think the measurements of relative transcript abundances are of good quality.

Library	Read Length	Unique	UniqueSplices	Multi	MultiSplices
12515 100-cell pool A	1x100	17,687,845	3,209,817	2,324,217	87,366
12516 100-cell pool B	1x100	19,196,833	3,613,603	2,472,612	116,124
12517 30-cell pool A	1x100	19,656,269	3,836,281	2,747,606	112,715
12518 30-cell pool B	1x100	15,906,819	3,105,647	2,209,219	107,243
12519 10-cell pool A	1x100	25,589,985	7,716,359	3,942,315	264,713
12520 10-cell pool B	1x100	14,033,035	3,831,207	2,172,320	92,664
12522 cell 183	1x100	13,444,432	4,123,615	1,991,506	151,473
12523 cell 184	1x100	18,553,787	6,282,207	2,753,213	162,393
12524 cell 185	1x100	15,306,477	4,973,962	2,375,825	123,920
12534 cell 186	1x100	9,412,759	1,792,734	1,104,103	92,146
12535 cell 187	1x100	12,021,473	2,593,517	1,762,078	122,152
12536 cell 188	1x100	6,173,793	1,609,714	751,818	35,935
12537 cell 189	1x100	8,900,605	2,552,063	1,195,651	71,165
12538 cell 190	1x100	11,976,901	3,061,070	1,578,373	114,265
12539 cell 191	1x100	4,894,790	990,183	687,469	55,952
12540 cell 192	1x100	8,586,601	2,191,767	1,312,434	70,208
12541 cell 193	1x100	11,615,819	2,810,842	1,636,014	75,938
12542 cell 194	1x100	9,299,741	2,543,984	1,388,630	61,370
12543 cell 195	1x100	9,051,228	1,583,943	1,172,717	52,683
12818 cell 200	1x100	9,465,272	2,903,793	1,338,444	87,282
12819 cell 205	1x100	11,895,334	3,486,064	1,184,543	59,413
12820 cell 208	1x100	13,034,342	2,346,996	1,418,030	120,778
12821 pool/split 5	1x100	9,152,130	2,394,362	1,520,080	76,965
12822 pool/split 6	1x100	13,938,165	3,517,926	2,286,058	113,187
12823 pool/split 7	1x100	11,217,362	1,872,905	1,154,032	73,843
12824 pool/split 8	1x100	11,822,904	2,135,005	1,364,032	70,389
13270 pool/split 3 219	1x100	7,416,424	4,799,669	1,463,631	457,029
13271 pool/split 4 220	1x100	8,421,706	5,262,489	1,644,668	496,781
13272 pool/split 9 225	1x100	12,782,172	4,509,292	1,480,617	427,615
13273 pool/split 10 226	1x100	10,325,385	6,582,179	2,100,134	641,196
13274 10ng RNA rep1	1x100	33,234,882	4,315,401	1,629,950	267,868
13275 10ng RNA rep2	1x100	36,981,036	5,266,981	1,704,651	301,449
13276 100pg RNA rep1	1x100	11,363,854	4,904,470	1,008,244	258,637
13277 100pg RNA rep2	1x100	34,939,583	6,750,980	2,212,161	442,062
13278 cell 204	1x100	20,631,514	11,290,949	3,418,238	921,764
13279 cell 207	1x100	10,926,463	4,949,640	1,664,688	490,150
13280 pool/split 232	1x100	21,240,282	9,537,592	2,722,244	726,943
13281 pool/split 233	1x100	25,425,429	9,576,495	2,510,136	703,065
13282 cell 235	1x100	10,167,950	3,677,729	966,782	191,523
13283 cell 236	1x100	18,782,295	7,784,497	2,210,674	572,837
13284 cell 237	1x100	25,766,827	8,914,958	2,235,457	594,889
13285 cell 238	1x100	16,334,009	6,842,776	2,351,813	602,952
13286 cell 239	1x100	19,717,157	5,801,008	2,473,230	595,738
13287 cell 240	1x100	21,881,195	8,373,245	2,386,125	645,571

Continued on next page

Supplementary Table 1 – *Continued from previous page*

Library	Read Length	Unique	UniqueSplices	Multi	MultiSplices
13288 cell 242	1x100	19,165,078	6,146,306	1,338,990	330,167
13289 cell 243	1x100	24,802,270	9,575,191	2,885,175	744,245
13290 cell 244	1x100	7,400,266	3,086,583	741,408	223,657
13291 cell 245	1x100	21,024,295	7,093,623	2,111,549	519,415
13292 pool/split 246	1x100	17,296,143	8,394,643	2,223,819	572,943
13294 pool/split 248	1x100	14,399,162	6,272,094	1,784,982	459,195
13295 pool/split 249	1x100	22,428,103	10,454,916	2,898,266	815,093
13296 pool/split 250	1x100	19,745,007	8,825,294	2,468,779	697,549
13297 pool/split 251	1x100	21,239,455	9,833,749	2,936,743	724,006
13298 pool/split 252	1x100	4,674,393	2,237,759	591,303	145,488
13299 pool/split 253	1x100	20,948,852	9,729,505	2,726,042	709,672
13300 10-cell pool 254	1x100	29,113,485	8,790,470	2,600,560	702,142
13301 11-cell pool 255	1x100	34,836,093	11,643,761	3,802,039	1,080,518
13302 100-cell pool 256	1x100	18,477,603	4,084,659	1,618,554	329,602
13303 100-cell pool 257	1x100	43,640,710	11,315,061	4,819,940	1,036,363

Supplementary Table 2: Initial amounts of spiked-in sequences in absolute number of RNA copies. Note that two more spikes, "Lambda 9786 clone F" (9786bp) and "Lambda 11300 clone G" (11290bp) were included in libraries, however, they exhibit highly non-uniform read coverage leading to unreliable quantification estimates and were thus excluded.

Spike-in	Libraries 12515-12543	Libraries 12818-13303
AGP23	100	5
AP2	5	50
EPR1	20	10
OBF5	10	500
PDF1	40	20
VATG3	5000	5000

Supplementary Table 3: Full list of Gene Ontology categories enriched in coexpressed gene modules. Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al., 2009).

N	X	LOD	P	P adj	attrib ID	attrib name
Module 1						
80	85	2.22750847233495	2.4288542157057e-80	<0.001	GO:0019083	viral transcription
35	38	2.0543096800611	9.94569501487657e-35	<0.001	GO:0022625	cytosolic large ribosomal subunit
79	88	1.98429369667081	3.81815501340532e-75	<0.001	GO:0006415	translational termination
82	92	1.95783499865395	2.25183855262354e-77	<0.001	GO:0019058	viral infectious cycle
33	37	1.91934326185776	1.0606448232178e-31	<0.001	GO:0022627	cytosolic small ribosomal subunit
84	99	1.79959130536831	2.83136121010728e-75	<0.001	GO:0006414	translational elongation
79	98	1.67170885189082	2.54715717317477e-67	<0.001	GO:0043624	cellular protein complex disassembly
79	99	1.64996220882795	1.16143543924047e-66	<0.001	GO:0043241	protein complex disassembly
5	6	1.60339903304627	2.31518111579124e-05	0.042	GO:0042719	mitochondrial intermembrane space protein transporter complex
12	16	1.48484395467934	1.55655372018477e-10	<0.001	GO:0005753	mitochondrial proton-transporting ATP synthase complex
12	16	1.48484395467934	1.55655372018477e-10	<0.001	GO:0045259	proton-transporting ATP synthase complex
8	11	1.42532291837978	3.17730434970578e-07	<0.001	GO:0042274	ribosomal small subunit biogenesis
42	60	1.41110133078985	7.7723344218279e-32	<0.001	GO:0015935	small ribosomal subunit
10	14	1.40852193295129	1.23601718399618e-08	<0.001	GO:0042776	mitochondrial ATP synthesis coupled proton transport
82	119	1.40425322218248	2.90687191098935e-60	<0.001	GO:0034623	cellular macromolecular complex disassembly
43	62	1.39861873942686	2.70656133560673e-32	<0.001	GO:0015934	large ribosomal subunit
82	120	1.39279633589538	8.45636479922126e-60	<0.001	GO:0032984	macromolecular complex disassembly
14	20	1.39014850585389	1.95202402281072e-11	<0.001	GO:0015985	energy coupled proton transport, down electrochemical gradient
14	20	1.39014850585389	1.95202402281072e-11	<0.001	GO:0015986	ATP synthesis coupled proton transport
9	13	1.36475589719287	1.06059816949165e-07	<0.001	GO:0045263	proton-transporting ATP synthase complex, coupling factor F(o)
102	154	1.35835395745315	3.83509033779718e-72	<0.001	GO:0003735	structural constituent of ribosome
80	121	1.34883668562943	1.30325679955416e-56	<0.001	GO:0031018	endocrine pancreas development
6	9	1.3082177020168	2.3318724760727e-05	0.049	GO:0042273	ribosomal large subunit biogenesis
92	156	1.22084501073779	1.79769896957827e-58	<0.001	GO:0022415	viral reproductive process
9	15	1.20500031764798	6.3592221573507e-07	<0.001	GO:0042613	MHC class II protein complex
90	159	1.17815283094918	4.63835969773452e-55	<0.001	GO:0005840	ribosome
11	20	1.12368307791708	1.16201671228436e-07	<0.001	GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
126	243	1.1058240710311	8.64889403866865e-71	<0.001	GO:0006412	translation
33	66	1.04671661076047	1.04054778845837e-18	<0.001	GO:0070469	respiratory chain
88	180	1.04302248261017	8.36713797967258e-47	<0.001	GO:0071845	cellular component disassembly at cellular level
13	26	1.04120105244121	3.59500355461033e-08	<0.001	GO:0004129	cytochrome-c oxidase activity
13	26	1.04120105244121	3.59500355461033e-08	<0.001	GO:0015002	heme-copper terminal oxidase activity
13	26	1.04120105244121	3.59500355461033e-08	<0.001	GO:0016676	oxidoreductase activity, acting on a heme group of donors, oxygen as acceptor
88	181	1.03832508439038	1.50023036663094e-46	<0.001	GO:0022411	cellular component disassembly
50	105	1.01033284389759	2.12881566827084e-26	<0.001	GO:0022904	respiratory electron transport chain
13	27	1.01013943569481	6.3992281019869e-08	<0.001	GO:0016675	oxidoreductase activity, acting on a heme group of donors
13	27	1.01013943569481	6.3992281019869e-08	<0.001	GO:0019843	rRNA binding
22	46	1.00663430266062	2.01941178786577e-12	<0.001	GO:0005747	mitochondrial respiratory chain complex I
22	46	1.00663430266062	2.01941178786577e-12	<0.001	GO:0030964	NADH dehydrogenase complex
22	46	1.00663430266062	2.01941178786577e-12	<0.001	GO:0045271	respiratory chain complex I
13	28	0.981148355026407	1.10254306563049e-07	<0.001	GO:0016469	proton-transporting two-sector ATPase complex
20	44	0.965600146054987	6.60764042765157e-11	<0.001	GO:0003954	NADH dehydrogenase activity
20	44	0.965600146054987	6.60764042765157e-11	<0.001	GO:0008137	NADH dehydrogenase (ubiquinone) activity
20	44	0.965600146054987	6.60764042765157e-11	<0.001	GO:0050136	NADH dehydrogenase (quinone) activity
10	22	0.964605445604187	4.26260478570593e-06	0.009	GO:0033177	proton-transporting two-sector ATPase complex, proton-transporting domain
19	43	0.94357846286631	3.63592920641863e-10	<0.001	GO:0006120	mitochondrial electron transport, NADH to ubiquinone
56	140	0.877587213657932	1.12610252541087e-24	<0.001	GO:0022900	electron transport chain
20	50	0.870301953062918	1.08177287181172e-09	<0.001	GO:0016655	oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor
126	329	0.864920624044305	7.76265565058296e-52	<0.001	GO:0016032	viral reproduction
19	48	0.862785503901917	3.4843140169824e-09	<0.001	GO:0022613	ribonucleoprotein complex biogenesis
51	133	0.846254921603025	1.32115278867509e-21	<0.001	GO:0044455	mitochondrial membrane part
13	36	0.80019305087457	3.59382041803685e-06	0.008	GO:0042611	MHC protein complex
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072401	signal transduction involved in DNA integrity checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072413	signal transduction involved in mitotic cell cycle checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072422	signal transduction involved in DNA damage checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072431	signal transduction involved in mitotic cell cycle G1/S transition DNA damage checkpoint
22	62	0.787906724016196	2.35759581410622e-09	<0.001	GO:0072474	signal transduction involved in mitotic cell cycle G1/S checkpoint
12	34	0.785381071713187	1.12064061460209e-05	0.019	GO:0006626	protein targeting to mitochondrion
31	88	0.784090641205952	1.6208634459227e-12	<0.001	GO:0048002	antigen processing and presentation of peptide antigen
138	408	0.782753905363391	3.18149793786453e-49	<0.001	GO:0010467	gene expression
22	63	0.777286220816123	3.33709849401727e-09	<0.001	GO:0072395	signal transduction involved in cell cycle checkpoint
22	63	0.777286220816123	3.33709849401727e-09	<0.001	GO:0072404	signal transduction involved in G1/S transition checkpoint
19	55	0.770122744469575	4.78383728527817e-08	<0.001	GO:0071843	cellular component biogenesis at cellular level
173	530	0.768759824961379	2.97828415371199e-59	<0.001	GO:0030529	ribonucleoprotein complex
28	82	0.763064993637299	4.54473457242155e-11	<0.001	GO:0002474	antigen processing and presentation of peptide antigen via MHC class I
25	75	0.747063491558582	8.85423589248866e-10	<0.001	GO:0006839	mitochondrial transport

Continued on next page

Supplementary Table 3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
23	69	0.746932600935863	4.15847287902749e-09	<0.001	GO:0044085	cellular component biogenesis
32	97	0.741175253071487	5.45910901839781e-12	<0.001	GO:0015078	hydrogen ion transmembrane transporter activity
16	49	0.73399635379054	1.30379324163326e-06	0.003	GO:0051258	protein polymerization
15	47	0.719731376252704	3.8551671841249e-06	0.008	GO:0051084	'de novo' posttranslational protein folding
40	126	0.71784240561212	5.25251868522275e-14	<0.001	GO:0019882	antigen processing and presentation
23	73	0.710984404898217	1.41497713185833e-08	<0.001	GO:0071158	positive regulation of cell cycle arrest
21	67	0.707697276637143	6.62504934644734e-08	<0.001	GO:0051436	negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
95	311	0.705569345151931	4.85030590692832e-30	<0.001	GO:0048610	cellular process involved in reproduction
16	52	0.696666038434829	3.16984759196741e-06	0.007	GO:0006458	'de novo' protein folding
22	72	0.691795994094768	5.42170060581127e-08	<0.001	GO:0051352	negative regulation of ligase activity
22	72	0.691795994094768	5.42170060581127e-08	<0.001	GO:0051444	negative regulation of ubiquitin-protein ligase activity
150	525	0.677439350068986	3.32539653823493e-43	<0.001	GO:0034621	cellular macromolecular complex subunit organization
164	584	0.670332348124673	3.54836301881911e-46	<0.001	GO:0016071	mRNA metabolic process
23	78	0.669845547354692	5.69629303883152e-08	<0.001	GO:0000216	M/G1 transition of mitotic cell cycle
21	72	0.663205782638981	2.6219034657749e-07	<0.001	GO:0051437	positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
16	55	0.662279533745671	7.14593997684204e-06	0.014	GO:0000502	proteasome complex
16	55	0.662279533745671	7.14593997684204e-06	0.014	GO:0006521	regulation of cellular amino acid metabolic process
25	86	0.66117774103715	2.09408969093921e-08	<0.001	GO:0030330	DNA damage response, signal transduction by p53 class mediator
19	67	0.64634468268949	1.55090946941351e-06	0.003	GO:0033238	regulation of cellular amine metabolic process
26	92	0.644103259424952	2.09762058586999e-08	<0.001	GO:0072331	signal transduction by p53 class mediator
22	78	0.642874216140607	2.62393000307691e-07	<0.001	GO:0051439	regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
19	68	0.637453778789019	1.9842690458222e-06	0.005	GO:0060333	interferon-gamma-mediated signaling pathway
22	79	0.635227361820706	3.35087879773373e-07	<0.001	GO:0051443	positive regulation of ubiquitin-protein ligase activity
23	83	0.63224585989498	2.00949230376557e-07	<0.001	GO:0051351	positive regulation of ligase activity
27	98	0.62887220631022	2.04834672373096e-08	<0.001	GO:0006364	rRNA processing
22	80	0.627711881242392	4.25874069739404e-07	<0.001	GO:0031145	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
81	300	0.62621754092227	6.64252358742553e-22	<0.001	GO:0005743	mitochondrial inner membrane
17	63	0.617087551763504	1.13566498370941e-05	0.021	GO:0006200	ATP catabolic process
24	89	0.616024095582501	1.924740199118e-07	<0.001	GO:0031397	negative regulation of protein ubiquitination
23	86	0.61114511085871	4.04971652012907e-07	<0.001	GO:0016651	oxidoreductase activity, acting on NADH or NADPH
27	103	0.599379377993909	6.42976552229558e-08	<0.001	GO:0016072	rRNA metabolic process
84	327	0.597134802233134	3.86637763010753e-21	<0.001	GO:0019866	organelle inner membrane
24	93	0.590170688226641	4.68334780721795e-07	<0.001	GO:0051438	regulation of ubiquitin-protein ligase activity
25	97	0.589412627207871	2.79036812385036e-07	<0.001	GO:0051340	regulation of ligase activity
28	109	0.587562019917053	5.91212587104247e-08	<0.001	GO:0000084	S phase of mitotic cell cycle
20	79	0.579289164068354	5.71262604161276e-06	0.012	GO:0071346	cellular response to interferon-gamma
141	582	0.575501427143487	5.37209201940123e-32	<0.001	GO:0071822	protein complex subunit organization
35	140	0.572325037803156	2.96216165946153e-09	<0.001	GO:0000082	G1/S transition of mitotic cell cycle
27	110	0.561161855618649	2.76881959100789e-07	<0.001	GO:0042770	signal transduction in response to DNA damage
28	116	0.551583814959451	2.4384223848345e-07	<0.001	GO:0051320	S phase
19	79	0.550001609297223	2.1441603106869e-05	0.031	GO:0009206	purine ribonucleoside triphosphate biosynthetic process
100	434	0.537107458537325	3.65671802133964e-21	<0.001	GO:0031966	mitochondrial membrane
31	132	0.536082594042894	1.10062427133423e-07	<0.001	GO:0090068	positive regulation of cell cycle process
22	95	0.528168311977894	9.49580519250266e-06	0.019	GO:0034341	response to interferon-gamma
84	371	0.523771651861378	1.85277099409276e-17	<0.001	GO:0006091	generation of precursor metabolites and energy
130	595	0.513071456988699	5.70494731538591e-25	<0.001	GO:0005198	structural molecule activity
26	116	0.50961644963599	2.91400334562788e-06	0.007	GO:0031398	positive regulation of protein ubiquitination
148	687	0.508962926945213	1.09821996775926e-27	<0.001	GO:0044429	mitochondrial part
37	168	0.500349876744172	4.12597162657402e-08	<0.001	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
37	168	0.500349876744172	4.12597162657402e-08	<0.001	GO:0000398	nuclear mRNA splicing, via spliceosome
36	164	0.498410796185536	6.84854979223404e-08	<0.001	GO:0043623	cellular protein complex assembly
38	175	0.492543109861863	4.05723893001672e-08	<0.001	GO:0000375	RNA splicing, via transesterification reactions
27	126	0.484584970787029	4.69193568014889e-06	0.009	GO:0046034	ATP metabolic process
175	867	0.477792444174725	4.31717391375335e-29	<0.001	GO:0043933	macromolecular complex subunit organization
32	157	0.457121358455519	2.03558979738233e-06	0.005	GO:0000209	protein polyubiquitination
57	282	0.455331541986413	3.35174043037384e-10	<0.001	GO:0008380	RNA splicing
87	437	0.45206487497538	1.75634248577746e-14	<0.001	GO:0055114	oxidation-reduction process
31	154	0.450275487651611	3.84142188009438e-06	0.008	GO:0031396	regulation of protein ubiquitination
32	159	0.450199742241769	2.70953251758628e-06	0.006	GO:0043161	proteasomal ubiquitin-dependent protein catabolic process
155	803	0.44828162002615	1.17127238506759e-23	<0.001	GO:0003723	RNA binding
32	160	0.446779218086837	3.11868915151585e-06	0.007	GO:0010498	proteasomal protein catabolic process
27	135	0.44673045980076	1.79130278352513e-05	0.028	GO:0005774	vacuolar membrane
239	1315	0.43236771896813	1.93513347789966e-32	<0.001	GO:0005739	mitochondrion
39	201	0.43074480120704	6.14144121470793e-07	<0.001	GO:0000075	cell cycle checkpoint
43	226	0.420501515091121	2.92681541907647e-07	<0.001	GO:0071156	regulation of cell cycle arrest
30	158	0.418585946292244	1.82741185723924e-05	0.028	GO:0044437	vacuolar part
57	304	0.414291886997772	6.22443660977404e-09	<0.001	GO:0000278	mitotic cell cycle
73	401	0.400691520310638	1.88837546257743e-10	<0.001	GO:0034622	cellular macromolecular complex assembly
222	1296	0.39523148469997	2.51837746792813e-26	<0.001	GO:0034645	cellular macromolecule biosynthetic process
226	1327	0.392809170439032	1.86450696466837e-26	<0.001	GO:0009059	macromolecule biosynthetic process
334	2084	0.38137706335828	1.33257319443321e-34	<0.001	GO:0044267	cellular protein metabolic process
334	2111	0.373898530461539	1.84879744335921e-33	<0.001	GO:0005829	cytosol
161	961	0.370923240921755	4.01275183856543e-18	<0.001	GO:0032774	RNA biosynthetic process
504	3431	0.370107849871159	4.0461986238786e-44	<0.001	GO:0032991	macromolecular complex

Continued on next page

Supplementary Table 3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
101	602	0.361160597240294	9.47442144063175e-12	<0.001	GO:0006396	RNA processing
766	6058	0.346864258263512	2.31238377684638e-47	<0.001	GO:0044444	cytoplasmic part
58	357	0.337821335867861	7.933880428251e-07	0.001	GO:0010564	regulation of cell cycle process
59	364	0.336711582120088	6.91212093849837e-07	0.001	GO:0006397	mRNA processing
144	911	0.336185023040326	3.85435400725073e-14	<0.001	GO:0005654	nucleoplasm
251	1662	0.330293674534452	1.10940941066799e-21	<0.001	GO:0016070	RNA metabolic process
313	2121	0.329217998390912	1.18114081249402e-25	<0.001	GO:0044249	cellular biosynthetic process
316	2160	0.324822918675263	2.91009865940903e-25	<0.001	GO:0090304	nucleic acid metabolic process
320	2232	0.313971532133542	4.41002855691382e-24	<0.001	GO:0009058	biosynthetic process
1145	11330	0.313948898323752	4.59005189412599e-31	<0.001	GO:0044424	intracellular part
75	485	0.313905177878569	1.5748417548981e-07	<0.001	GO:0022403	cell cycle phase
377	2686	0.312982629840953	4.90694955531477e-27	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
686	5522	0.308771441881537	4.1087928647367e-37	<0.001	GO:0044446	intracellular organelle part
367	2644	0.304397651421782	2.77819227696323e-25	<0.001	GO:0019538	protein metabolic process
529	4046	0.302681908190103	8.82320727696583e-32	<0.001	GO:0044260	cellular macromolecule metabolic process
689	5601	0.302589536132723	8.22904448857063e-36	<0.001	GO:0044422	organelle part
712	5867	0.299367116917043	1.88823085526176e-35	<0.001	GO:0044237	cellular metabolic process
135	945	0.280373762273502	3.82813590337554e-10	<0.001	GO:0022414	reproductive process
74	511	0.279376163393885	2.44628059184637e-06	0.006	GO:0006974	response to DNA damage stimulus
969	9102	0.277956512781225	5.11225805494553e-30	<0.001	GO:0043229	intracellular organelle
969	9117	0.276313895789093	1.11162671743124e-29	<0.001	GO:0043226	organelle
403	3093	0.273819161695768	1.68049396494688e-22	<0.001	GO:0034641	cellular nitrogen compound metabolic process
108	765	0.270319977503456	4.55991397471596e-08	<0.001	GO:0022402	cell cycle process
99	701	0.269327927902864	1.66323414881123e-07	<0.001	GO:0007049	cell cycle
410	3199	0.265219890086891	1.5194034851802e-21	<0.001	GO:0006807	nitrogen compound metabolic process
556	4601	0.25742146969207	1.8068455687858e-24	<0.001	GO:0043170	macromolecule metabolic process
285	2171	0.257317984548616	7.98160838856529e-16	<0.001	GO:0043228	non-membrane-bounded organelle
285	2171	0.257317984548616	7.98160838856529e-16	<0.001	GO:0043232	intracellular non-membrane-bounded organelle
69	497	0.256916076214485	2.21805085904211e-05	0.031	GO:0005730	nucleolus
85	622	0.2504348652183	4.78113517987455e-06	0.01	GO:0006259	DNA metabolic process
744	6632	0.249054934285073	1.58723972093304e-25	<0.001	GO:0008152	metabolic process
293	2272	0.248502561040477	3.01319156248601e-15	<0.001	GO:0071842	cellular component organization at cellular level
295	2299	0.245967084736522	4.57315463821616e-15	<0.001	GO:0071841	cellular component organization or biogenesis at cellular level
98	729	0.243379241039335	1.87437352876668e-06	0.004	GO:0065003	macromolecular complex assembly
93	694	0.241149602052021	3.93342779914606e-06	0.008	GO:0016491	oxidoreductase activity
274	2153	0.238023679234234	1.58478501985632e-13	<0.001	GO:0044428	nuclear part
871	8242	0.235473646680484	5.82916610504252e-23	<0.001	GO:0043227	membrane-bounded organelle
870	8238	0.234562637103022	8.38290144947865e-23	<0.001	GO:0043231	intracellular membrane-bounded organelle
99	751	0.233280218653404	3.94595885765465e-06	0.008	GO:0046907	intracellular transport
665	5992	0.221831084147262	3.46559295245285e-20	<0.001	GO:0044238	primary metabolic process
102	803	0.214295076083374	1.41199427931906e-05	0.022	GO:0033554	cellular response to stress
99	783	0.211573308398039	2.28182618638875e-05	0.031	GO:0071844	cellular component assembly at cellular level
344	2896	0.208867137566652	8.16942218480433e-13	<0.001	GO:0016043	cellular component organization
346	2923	0.20711459632924	1.09506306063704e-12	<0.001	GO:0071840	cellular component organization or biogenesis
329	2821	0.195993854311548	3.24914520945555e-11	<0.001	GO:0043234	protein complex
237	2006	0.191345208630818	1.2625912051491e-08	<0.001	GO:0031090	organelle membrane
1026	10840	0.168964395995212	1.74282758664301e-11	<0.001	GO:0009987	cellular process
485	4933	0.108855314264424	1.27790612421947e-05	0.021	GO:0005737	cytoplasm
Module 2						
5	9	1.4353697859032	1.61126889116162e-05	0.024	GO:0008139	nuclear localization sequence binding
7	20	1.09388232798139	1.28535977541801e-05	0.018	GO:0051983	regulation of chromosome segregation
29	168	0.684180963922458	1.54004341495467e-10	<0.001	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
29	168	0.684180963922458	1.54004341495467e-10	<0.001	GO:0000398	nuclear mRNA splicing, via spliceosome
30	175	0.680816140112517	8.87046264555359e-11	<0.001	GO:0000375	RNA splicing, via transesterification reactions
38	282	0.558803234506848	4.52064927189642e-10	<0.001	GO:0008380	RNA splicing
19	141	0.555278310629005	1.0380427172764e-05	0.017	GO:0005681	spliceosomal complex
44	364	0.505813785001266	6.61426439925337e-10	<0.001	GO:0006397	mRNA processing
66	602	0.464516339086547	2.77728663822659e-12	<0.001	GO:0006396	RNA processing
32	288	0.460440562314741	9.75520192342185e-07	0.002	GO:0051301	cell division
62	584	0.446895852037615	5.12932632405468e-11	<0.001	GO:0016071	mRNA metabolic process
52	497	0.436204344115675	3.27009886602057e-09	<0.001	GO:0005730	nucleolus
33	351	0.378532533073338	2.40744755831632e-05	0.041	GO:0044419	interspecies interaction between organisms
172	2153	0.343432744866356	1.75600448691817e-16	<0.001	GO:0044428	nuclear part
42	485	0.340242072283576	1.48503818757099e-05	0.022	GO:0022403	cell cycle phase
40	462	0.339711208044223	2.3766092104457e-05	0.04	GO:0044265	cellular macromolecule catabolic process
60	701	0.339160869975101	3.23571054989503e-07	0.001	GO:0007049	cell cycle
41	475	0.338432406345785	2.02295400933921e-05	0.034	GO:0044427	chromosomal part
68	803	0.336220984744933	7.65671927087538e-08	<0.001	GO:0003723	RNA binding
76	911	0.331218188924209	2.33893486630229e-08	<0.001	GO:0005654	nucleoplasm
43	511	0.326468203963555	2.35721317636639e-05	0.04	GO:0006974	response to DNA damage stimulus
44	530	0.320096499039325	2.63517822318679e-05	0.043	GO:0030529	ribonucleoprotein complex
590	11330	0.307725883035438	1.0960786202461e-16	<0.001	GO:0044424	intracellular part
128	1662	0.307694952226285	4.08803837662614e-11	<0.001	GO:0016070	RNA metabolic process
51	644	0.298647909108822	2.19483688586676e-05	0.035	GO:0051726	regulation of cell cycle
58	765	0.278965516407217	2.16486041316539e-05	0.035	GO:0022402	cell cycle process
152	2160	0.26839415907082	3.64000656801177e-10	<0.001	GO:0090304	nucleic acid metabolic process

Continued on next page

Supplementary Table 3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
151	2171	0.261862930845385	9.73365119250831e-10	<0.001	GO:0043228	non-membrane-bounded organelle
151	2171	0.261862930845385	9.73365119250831e-10	<0.001	GO:0043232	intracellular non-membrane-bounded organelle
340	5601	0.253862278441637	1.31291950589379e-14	<0.001	GO:0044422	organelle part
492	9102	0.250272137858337	4.30276700995675e-14	<0.001	GO:0043229	intracellular organelle
492	9117	0.248695876930274	6.1868252491176e-14	<0.001	GO:0043226	organelle
178	2686	0.244422165955063	9.13870908091538e-10	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
332	5522	0.243714592261868	1.60539211670902e-13	<0.001	GO:0044446	intracellular organelle part
200	3093	0.237932293242852	4.47608150862857e-10	<0.001	GO:0034641	cellular nitrogen compound metabolic process
204	3199	0.231396211719754	9.30860336734763e-10	<0.001	GO:0006807	nitrogen compound metabolic process
251	4046	0.231319523511239	5.73866479299139e-11	<0.001	GO:0044260	cellular macromolecule metabolic process
449	8238	0.229250611380594	1.64011741005672e-12	<0.001	GO:0043231	intracellular membrane-bounded organelle
449	8242	0.228829514942579	1.79770699314044e-12	<0.001	GO:0043227	membrane-bounded organelle
338	5867	0.217174729554922	3.40468652787424e-11	<0.001	GO:0044237	cellular metabolic process
290	4933	0.212021859434929	3.43579358808612e-10	<0.001	GO:0005737	cytoplasm
302	5211	0.207147691965801	5.67529432661151e-10	<0.001	GO:0005634	nucleus
111	1743	0.20458125474222	1.63777858635897e-05	0.024	GO:0043412	macromolecule modification
266	4601	0.19325730820235	1.836736540134e-08	<0.001	GO:0043170	macromolecule metabolic process
363	6632	0.19076943584687	3.42220660572212e-09	<0.001	GO:0008152	metabolic process
328	5992	0.177849130815706	4.97783632875206e-08	<0.001	GO:0044238	primary metabolic process
174	3011	0.167727728807245	1.68644378475076e-05	0.031	GO:0010468	regulation of gene expression
315	6058	0.137883228026376	2.01099934629971e-05	0.034	GO:0044444	cytoplasmic part
			Module 3			
8	25	1.12174413109538	1.6195735819603e-06	0.002	GO:0051539	4 iron, 4 sulfur cluster binding
10	50	0.84991907137197	9.0161990136549e-06	0.021	GO:0051536	iron-sulfur cluster binding
10	50	0.84991907137197	9.0161990136549e-06	0.021	GO:0051540	metal cluster binding
25	229	0.538541159788376	7.47720629054225e-07	<0.001	GO:0051186	cofactor metabolic process
271	6058	0.166074173009115	2.78569620584793e-06	0.005	GO:0044444	cytoplasmic part
453	11330	0.161739187745443	1.86883086496289e-05	0.037	GO:0044424	intracellular part
			Module 4			
13	97	0.672080308928874	2.01477245308462e-05	0.047	GO:0005741	mitochondrial outer membrane
32	434	0.38218093135728	2.59169277455588e-05	0.05	GO:0031966	mitochondrial membrane
267	6058	0.209635231286387	1.1954907403548e-08	<0.001	GO:0044444	cytoplasmic part
430	11330	0.189685955412112	1.81000117796974e-06	0.003	GO:0044424	intracellular part
266	6401	0.167667578490431	3.97412665691163e-06	0.008	GO:0005515	protein binding
211	4933	0.164971820301435	1.42128334658729e-05	0.024	GO:0005737	cytoplasm
			Module 5			
7	33	1.07143232855034	1.02220662682094e-05	0.013	GO:0006695	cholesterol biosynthetic process
34	602	0.418505790853493	2.98488408837491e-06	0.005	GO:0006396	RNA processing
42	803	0.386267510008865	1.45554203385518e-06	0.003	GO:0003723	RNA binding
35	678	0.376246720272173	1.50758252204483e-05	0.044	GO:0044451	nucleoplasm part
74	1662	0.32717492580596	8.08519738698485e-08	<0.001	GO:0016070	RNA metabolic process
92	2160	0.316438423646489	1.22530842680858e-08	<0.001	GO:0090304	nucleic acid metabolic process
109	2686	0.302719794834586	5.16088110052742e-09	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
204	5867	0.290215289462742	2.42264491073003e-11	<0.001	GO:0044237	cellular metabolic process
86	2153	0.279472007299701	7.06167821789315e-07	0.001	GO:0044428	nuclear part
258	8238	0.273394775085421	3.30641000342457e-10	<0.001	GO:0043231	intracellular membrane-bounded organelle
258	8242	0.272981776660256	3.51091215759553e-10	<0.001	GO:0043227	membrane-bounded organelle
320	11330	0.270756828023873	2.59329735790549e-08	<0.001	GO:0044424	intracellular part
117	3093	0.27018336330746	6.56451317805923e-08	<0.001	GO:0034641	cellular nitrogen compound metabolic process
120	3199	0.267574242908932	6.5838967556536e-08	<0.001	GO:0006807	nitrogen compound metabolic process
274	9102	0.259905768442695	3.97303688891244e-09	<0.001	GO:0043229	intracellular organelle
145	4046	0.258957848074746	2.58310912976049e-08	<0.001	GO:0044260	cellular macromolecule metabolic process
274	9117	0.258358056899225	4.89474559380795e-09	<0.001	GO:0043226	organelle
186	5522	0.252387490202399	8.07568575286686e-09	<0.001	GO:0044446	intracellular organelle part
198	5992	0.249866826361352	8.08174185764168e-09	<0.001	GO:0044238	primary metabolic process
187	5601	0.24739223761693	1.47932676168125e-08	<0.001	GO:0044422	organelle part
212	6632	0.239108907629557	2.58709518947142e-08	<0.001	GO:0008152	metabolic process
155	4601	0.229008499074375	4.23085771940228e-07	<0.001	GO:0043170	macromolecule metabolic process
169	5211	0.214851860778109	1.14557794256476e-06	0.002	GO:0005634	nucleus
188	6058	0.198889582472837	3.92901841156595e-06	0.007	GO:0044444	cytoplasmic part
			Module 6			
5	17	1.35877762485371	1.27620339057008e-05	0.027	GO:0005680	anaphase-promoting complex
7	27	1.28109582819072	5.64148602718057e-07	0.001	GO:0007094	mitotic cell cycle spindle assembly checkpoint
7	28	1.26038559665177	7.3990905561597e-07	0.004	GO:0045841	negative regulation of mitotic metaphase/anaphase transition
7	28	1.26038559665177	7.3990905561597e-07	0.004	GO:0071173	spindle assembly checkpoint
7	28	1.26038559665177	7.3990905561597e-07	0.004	GO:0071174	mitotic cell cycle spindle checkpoint
7	30	1.22170492439836	1.23095671972835e-06	0.005	GO:0031577	spindle checkpoint
7	33	1.16944999653699	2.45864548304647e-06	0.007	GO:0030071	regulation of mitotic metaphase/anaphase transition
7	33	1.16944999653699	2.45864548304647e-06	0.007	GO:0045839	negative regulation of mitosis
7	33	1.16944999653699	2.45864548304647e-06	0.007	GO:0051784	negative regulation of nuclear division
18	288	0.561314315261963	1.18040781186384e-05	0.026	GO:0051301	cell division
204	8238	0.251051414092351	1.67494198073935e-07	<0.001	GO:0043231	intracellular membrane-bounded organelle
204	8242	0.250640665799907	1.75182061275321e-07	<0.001	GO:0043227	membrane-bounded organelle
216	9102	0.231422137926905	1.80780431583472e-06	0.006	GO:0043229	intracellular organelle
216	9117	0.229882447927919	2.1050968158218e-06	0.007	GO:0043226	organelle

Continued on next page

Supplementary Table 3 – *Continued from previous page*

N	X	LOD	P	P adj	attrib ID	attrib name
135	5211	0.208102879924293	2.05983907876377e-05	0.037	GO:0005634	nucleus
Module 8						
3	4	2.15674334421143	1.64587987532083e-05	0.032	GO:0048280	vesicle fusion with Golgi apparatus
5	21	1.31441473215979	1.74263922633301e-05	0.037	GO:0032201	telomere maintenance via semi-conservative replication
8	35	1.28611338854416	6.73739625327887e-08	0.001	GO:0006261	DNA-dependent DNA replication
5	23	1.26467580787443	2.80631885844604e-05	0.047	GO:0000722	telomere maintenance via recombination
6	34	1.15088257068433	1.54366414227154e-05	0.024	GO:0010833	telomere maintenance via telomere lengthening
8	69	0.935700200428791	1.48897081315905e-05	0.024	GO:0009411	response to UV
15	170	0.802685314737047	1.12157969185014e-07	0.001	GO:0006260	DNA replication
23	511	0.490955055548057	9.92989139304257e-06	0.019	GO:0006974	response to DNA damage stimulus
42	1327	0.340009071984013	1.98581541366046e-05	0.04	GO:0009059	macromolecule biosynthetic process
41	1296	0.338927033802861	2.55286781292076e-05	0.043	GO:0034645	cellular macromolecule biosynthetic process
63	2160	0.318931266474648	1.75657887022238e-06	0.002	GO:0090304	nucleic acid metabolic process
128	5211	0.298549267749992	1.8238869432011e-08	<0.001	GO:0005634	nucleus
60	2153	0.292881834504383	1.38556689832231e-05	0.023	GO:0044428	nuclear part
100	4046	0.267683010803659	1.63739514172819e-06	0.002	GO:0044260	cellular macromolecule metabolic process
174	8238	0.263181580222692	4.66457450693039e-07	0.002	GO:0043231	intracellular membrane-bounded organelle
174	8242	0.262771810066496	4.85332067450148e-07	0.002	GO:0043227	membrane-bounded organelle
107	4601	0.239108491634204	1.07913454880254e-05	0.02	GO:0043170	macromolecule metabolic process
182	9102	0.22936103280645	1.25445382233209e-05	0.02	GO:0043229	intracellular organelle
182	9117	0.227824532081016	1.42589555343329e-05	0.023	GO:0043226	organelle
Module 10						
190	11330	0.26726174432135	1.71596077751879e-05	0.032	GO:0044424	intracellular part
Module 14						
20	1065	0.528808096440223	1.73407898205248e-05	0.032	GO:0044248	cellular catabolic process
33	2111	0.479914638806195	1.06712795272747e-06	<0.001	GO:0005829	cytosol
39	3093	0.388858388574349	1.71524549827148e-05	0.032	GO:0034641	cellular nitrogen compound metabolic process
42	3431	0.382865420385352	1.42566606441697e-05	0.023	GO:0032991	macromolecular complex
60	5867	0.34446562764228	2.14508227397237e-05	0.032	GO:0044237	cellular metabolic process
61	6058	0.338468119322826	2.84679848054408e-05	0.041	GO:0044444	cytoplasmic part

Supplementary Table 4: Full list of Gene Ontology categories enriched in coexpressed gene modules derived from a mirrored analysis of pool/split datasets. Gene Ontology enrichment in modules was assessed using FuncAssociate2.0 (Berriz et al., 2009). A total of 16 modules were detected

N	X	LOD	P	P adj	attrib ID	attrib name
Module 1						
13	272	0.693095067094612	1.02007791233691e-05	0.011	GO:0006511	ubiquitin-dependent protein catabolic process
13	279	0.68135455023532	1.33744095433394e-05	0.018	GO:0019941	modification-dependent protein catabolic process
13	282	0.676415542913035	1.49811431033085e-05	0.019	GO:0043632	modification-dependent macromolecule catabolic process
13	289	0.665098403683509	1.94061972015811e-05	0.024	GO:0051603	proteolysis involved in cellular protein catabolic process
17	462	0.57616161468926	1.43352606692572e-05	0.018	GO:0044265	cellular macromolecule catabolic process
19	545	0.553524223505169	9.47597917000467e-06	0.009	GO:0009057	macromolecule catabolic process
29	1065	0.451649915259144	5.44374777913331e-06	0.007	GO:0044248	cellular catabolic process
50	2153	0.408372755551646	1.54662372619939e-07	¡0.001	GO:0044428	nuclear part
38	1647	0.385284888523943	8.15733567738168e-06	0.007	GO:0006464	protein modification process
39	1743	0.371699071795165	1.23788499774497e-05	0.013	GO:0043412	macromolecule modification
75	4046	0.33429458014633	5.14482381026089e-07	¡0.001	GO:0044260	cellular macromolecule metabolic process
93	5522	0.313764330996431	7.30843785397851e-07	¡0.001	GO:0044446	intracellular organelle part
93	5601	0.304538452915935	1.46302793919708e-06	0.002	GO:0044422	organelle part
79	4601	0.296962579640552	5.07939215329675e-06	0.005	GO:0043170	macromolecule metabolic process
95	5867	0.292386561578076	3.34586773855207e-06	0.003	GO:0044237	cellular metabolic process
104	6632	0.29101237790736	3.17018865849522e-06	0.003	GO:0008152	metabolic process
96	5992	0.287467105557634	4.64261918086881e-06	0.005	GO:0044238	primary metabolic process
Module 3						
5	16	2.01970458115929	9.51160735183834e-09	¡0.001	GO:0016254	preassembly of GPI anchor in ER membrane
6	30	1.76902612001682	5.55484070660725e-09	¡0.001	GO:0018410	C-terminal protein amino acid modification
5	25	1.76842092954302	1.11841352190966e-07	0.001	GO:0006501	C-terminal protein lipidation
5	49	1.43120717990547	3.66720733305651e-06	0.006	GO:0006497	protein lipidation
7	131	1.12827314119116	3.16442237797319e-06	0.005	GO:0043687	post-translational protein modification
22	1743	0.516764340612836	1.71498823237458e-05	0.023	GO:0043412	macromolecule modification
Module 9						
19	2084	0.56988964370998	2.20527274140205e-05	0.031	GO:0044267	cellular protein metabolic process

Supplementary Table 5: Novel splice junctions discovered in spikes. No novel junctions are expected to be discovered in spike-in sequences, therefore such junctions are almost certainly experimental artifacts. Shown is the number of collapsed fragments supporting each junctions in a library

chr	left	right	strand	12517 GM12878 30 cells	12520 10 cells	12820 cell 208	12821 pool/split 5	12823 pool/split 7	12824 pool/split 8	13268 217 pool/split 1	13269 218 pool/split 2	13270 219 pool/split 3	13271 220 pool/split 4	13273 226 pool/split 10	13278 cell 204	13280 232 pool/split
AP2	308	1191	−	0	0	0	0	24	0	0	0	0	0	0	0	0
AP2	357	1205	−	1	0	0	0	0	0	0	0	0	0	0	0	0
AP2	365	1213	+	2	0	0	0	0	0	0	0	0	0	0	0	0
AP2	380	476	+	0	44	0	0	0	0	0	0	0	0	0	0	0
AP2	380	520	+	0	0	0	28	0	0	0	0	0	0	0	0	0
Lambda 9786 clone F	5092	5172	−	0	0	0	0	0	0	0	38	0	0	0	0	0
OBF5	334	466	+	0	0	0	39	0	0	0	0	0	0	0	0	0
OBF5	376	607	+	0	0	0	0	0	0	0	0	0	0	1	0	0
OBF5	381	612	+	0	0	0	0	0	0	0	0	0	0	47	0	0
OBF5	394	893	+	0	0	0	0	0	0	0	0	3	0	0	0	0
OBF5	411	860	+	0	0	1	0	0	0	0	0	0	0	0	0	0
OBF5	413	862	+	0	0	39	0	0	0	0	0	0	0	0	0	0
OBF5	417	526	−	0	0	0	28	0	0	0	0	0	0	0	0	0
OBF5	558	724	−	0	0	0	0	0	0	45	0	0	0	0	0	0
OBF5	558	988	−	0	0	0	0	0	0	0	0	0	44	0	0	0
OBF5	563	1000	+	0	0	0	0	0	0	0	0	0	0	0	0	2
OBF5	572	806	+	0	0	0	0	0	17	0	0	0	0	0	0	0
OBF5	572	848	+	0	0	18	0	0	0	0	0	0	0	0	0	0
OBF5	708	893	+	0	0	0	0	0	31	0	0	0	0	0	0	0
OBF5	708	992	+	0	0	0	0	0	0	0	0	0	0	0	36	0
VATG3	28	277	+	0	0	0	6	0	0	0	0	0	0	0	0	0

Supplementary Table 6: Gene Ontology categories enriched in genes displaying splice site switching between individual cells. Gene Ontology enrichment was assessed using FuncAssociate2.0 (Berriz et al., 2009).

N	X	LOD	P	P adj	attrib ID	attrib name
7	18	1.60434572330591	7.81336494766272e-09	<0.001	GO:0030530	heterogeneous nuclear ribonucleoprotein complex
6	22	1.38371032623861	1.0731443930779e-06	<0.001	GO:0000313	organellar ribosome
6	22	1.38371032623861	1.0731443930779e-06	<0.001	GO:0005761	mitochondrial ribosome
9	62	1.04144125075565	6.88184640563181e-07	<0.001	GO:0015934	large ribosomal subunit
10	79	0.972456465728994	6.10664091150019e-07	<0.001	GO:0071013	catalytic step 2 spliceosome
19	154	0.964236659283489	7.8714230213414e-12	<0.001	GO:0003735	structural constituent of ribosome
20	175	0.927300715975638	9.17354662099906e-12	<0.001	GO:0000375	RNA splicing, via transesterification reactions
19	168	0.92117404141009	3.70476457645963e-11	<0.001	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
19	168	0.92117404141009	3.70476457645963e-11	<0.001	GO:0000398	nuclear mRNA splicing, via spliceosome
15	141	0.888071082569232	1.03966880671155e-08	<0.001	GO:0005681	spliceosomal complex
16	159	0.861650703193511	7.48121263599197e-09	<0.001	GO:0005840	ribosome
23	243	0.838249153771732	1.25842265057125e-11	<0.001	GO:0006412	translation
46	530	0.826167157208287	9.02684713922854e-21	<0.001	GO:0030529	ribonucleoprotein complex
23	282	0.766509674540248	2.54718803100937e-10	<0.001	GO:0008380	RNA splicing
26	364	0.706261794294909	3.04918522987691e-10	<0.001	GO:0006397	mRNA processing
28	408	0.689368099675319	1.53848128981224e-10	<0.001	GO:0010467	gene expression
38	584	0.67579720156425	3.18781787872945e-13	<0.001	GO:0016071	mRNA metabolic process
47	803	0.636519902260068	1.87953754853827e-14	<0.001	GO:0003723	RNA binding
36	602	0.632953237795702	1.62794718956973e-11	<0.001	GO:0006396	RNA processing
16	277	0.597979679500433	1.35166171675457e-05	0.022	GO:0034660	ncRNA metabolic process
22	462	0.511423641446678	7.66202769258217e-06	0.015	GO:0044265	cellular macromolecule catabolic process
39	911	0.476845884038679	3.34218894462279e-08	<0.001	GO:0005654	nucleoplasm
24	545	0.47631618686561	1.09283168115262e-05	0.018	GO:0009057	macromolecule catabolic process
118	3431	0.473494407052489	1.56368496887015e-17	<0.001	GO:0032991	macromolecular complex
29	687	0.460527035699242	2.93531833594463e-06	0.005	GO:0044429	mitochondrial part
80	2160	0.454899179204351	5.73798891785767e-13	<0.001	GO:0090304	nucleic acid metabolic process
64	1662	0.453902252858483	4.77393199090586e-11	<0.001	GO:0016070	RNA metabolic process
51	1296	0.449261990101858	3.27437399406931e-09	<0.001	GO:0034645	cellular macromolecule biosynthetic process
79	2153	0.448759965851525	1.38228172140963e-12	<0.001	GO:0044428	nuclear part
52	1327	0.448329514135066	2.57197941829179e-09	<0.001	GO:0009059	macromolecule biosynthetic process
94	2686	0.444930025696524	6.62602821448192e-14	<0.001	GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
77	2111	0.443065370078769	4.00922406264811e-12	<0.001	GO:0005829	cytosol
127	4046	0.437107059343321	7.36675726742068e-16	<0.001	GO:0044260	cellular macromolecule metabolic process
102	3093	0.424687277579968	1.53499538750493e-13	<0.001	GO:0034641	cellular nitrogen compound metabolic process
234	11330	0.410177502418009	4.26279474281887e-11	<0.001	GO:0044424	intracellular part
73	2121	0.408294011011211	2.6476698540441e-10	<0.001	GO:0044249	cellular biosynthetic process
102	3199	0.40626576617879	1.33115703600882e-12	<0.001	GO:0006807	nitrogen compound metabolic process
47	1315	0.398127144349952	2.71558632079032e-07	<0.001	GO:0005739	mitochondrion
157	5867	0.394560459760076	4.17387500074186e-14	<0.001	GO:0044237	cellular metabolic process
150	5522	0.390555837986407	8.85159181992679e-14	<0.001	GO:0044446	intracellular organelle part
74	2232	0.390317428579321	1.04923257025558e-09	<0.001	GO:0009058	biosynthetic process
131	4601	0.386021432185656	5.00037984912871e-13	<0.001	GO:0043170	macromolecule metabolic process
150	5601	0.381254965130809	3.18071089534335e-13	<0.001	GO:0044422	organelle part
157	6058	0.372860438339392	8.46004601459964e-13	<0.001	GO:0044444	cytoplasmic part
151	5992	0.342490468755501	4.80405598147904e-11	<0.001	GO:0044238	primary metabolic process
161	6632	0.334875288021427	1.12251899150943e-10	<0.001	GO:0008152	metabolic process
63	2084	0.331075785878351	7.59282101998268e-07	<0.001	GO:0044267	cellular protein metabolic process
184	8242	0.317619158465355	1.64375270407462e-09	<0.001	GO:0043227	membrane-bounded organelle
183	8238	0.311179434676407	3.27953523717327e-09	<0.001	GO:0043231	intracellular membrane-bounded organelle
194	9102	0.300093429975042	2.09541238702377e-08	<0.001	GO:0043229	intracellular organelle
194	9117	0.298556745854624	2.46813702747217e-08	<0.001	GO:0043226	organelle
70	2644	0.269207945316679	1.84621206416052e-05	0.04	GO:0019538	protein metabolic process
74	2821	0.267942614823057	1.34448891236763e-05	0.021	GO:0043234	protein complex
213	10840	0.265155481178129	2.70868933554443e-06	0.005	GO:0009987	cellular process

Supplementary References

Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. 2009. Next generation software for functional trend analysis. *Bioinformatics* **25**(22):3043-3044.

de Hoon MJ, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. *Bioinformatics* **20**(9):1453-1454.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**(7414):101–108.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **22**(9):1760-1774.

Jin VX, Singer GA, Agosto-Pérez FJ, Liyanarachchi S, Davuluri RV. 2006. Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics* **7**:114.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**:559

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**(3):R25.

Saldanha AJ. 2004. Java Treeview–extensible visualization of microarray data. *Bioinformatics* **20**(17):3246-3248.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* **25**(9):1105-1111.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2012. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* **31**(1):46-53.

Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE, et al. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* **23**(3):555-567.