Supplementary Information

Integrated genome and transcriptome sequencing from the same cell

Siddharth S. Dey[1,2,3], Lennart Kester[1,2,3], Bastiaan Spanjaard[1,2], Magda Bienko[1,2] & Alexander van Oudenaarden[1,2]

[1]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, The Netherlands.

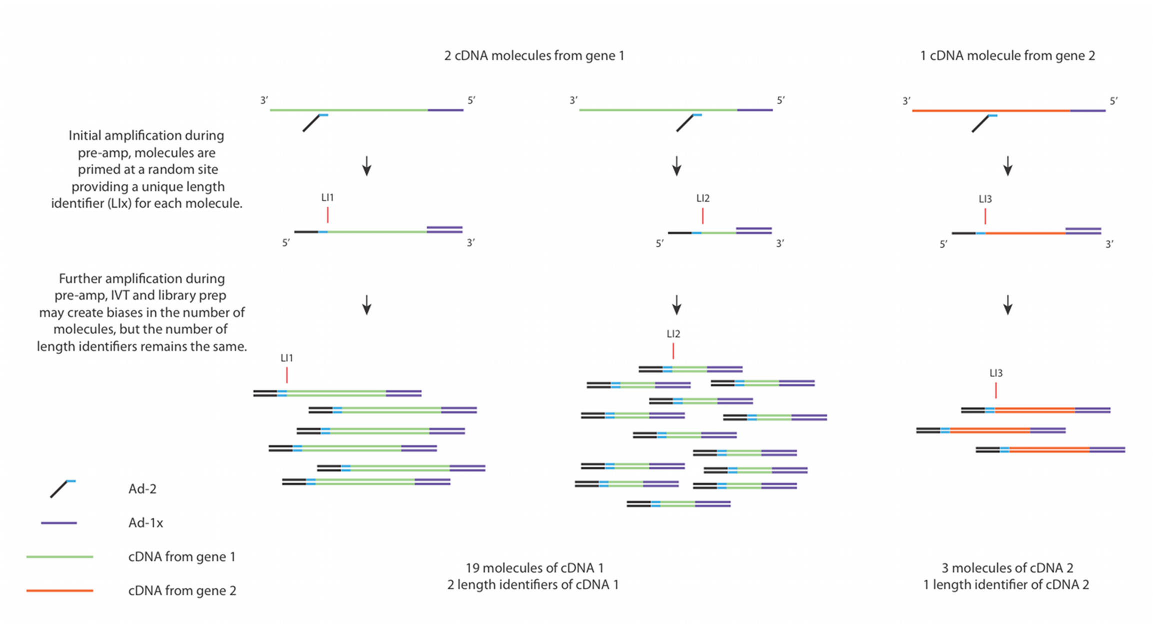
[2]University Medical Center Utrecht, Utrecht, The Netherlands.

[3]These authors contributed equally to this work.

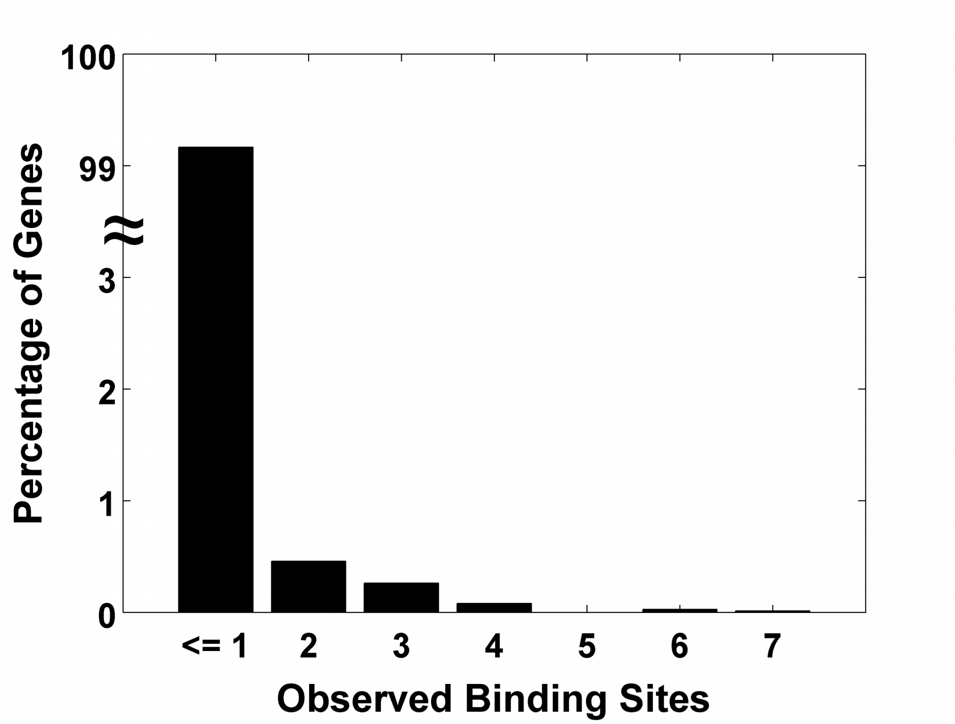
Correspondence should be addressed to A.v.O. (a.vanoudenaarden@hubrecht.eu)

Supplementary Item	Title or Caption
Supplementary Figure 1	Genomic location of random priming events during quasilinear amplification can be used to minimize amplification biases and achieve resolution close to identifying unique cDNA molecules in DR-Seq
Supplementary Figure 2	Original cDNA molecules are primed only once on average during quasilinear amplification in DR-Seq.
Supplementary Figure 3	Identification of theoretical binding sites in the mouse transcriptome
Supplementary Figure 4	Distribution of theoretical number of binding sites for each gene in the mouse transcriptome
Supplementary Figure 5	The number of original cDNA molecules can be accurately estimated by length-based identifiers in DR-Seq without reaching saturation
Supplementary Figure 6	Reduction in cell-to-cell variability of gene expression after correcting read-based data with unique molecule identifiers in CEL-Seq
Supplementary Figure 7	Unique molecule identifier correction reduces technical noise in CEL-Seq
Supplementary Figure 8	Pairwise Pearson correlations between single cells show reduction in technical noise after correcting the read-based data in CEL-Seq and DR-Seq
Supplementary Figure 9	The entire complexity of the single-cell mRNA libraries are sequenced in CEL-Seq and DR-Seq
Supplementary Figure 10	Number of genes identified by different sequencing methods
Supplementary Figure 11	Comparing ERCC spike-in detection between CEL-Seq and DR-Seq
Supplementary Figure 12	Quasilinear amplification in DR-Seq does not introduce additional biases in single-cell mRNA quantification
Supplementary Figure 13	Reads from the gDNA fraction in DR-Seq are mapped to the genome after masking out the coding regions

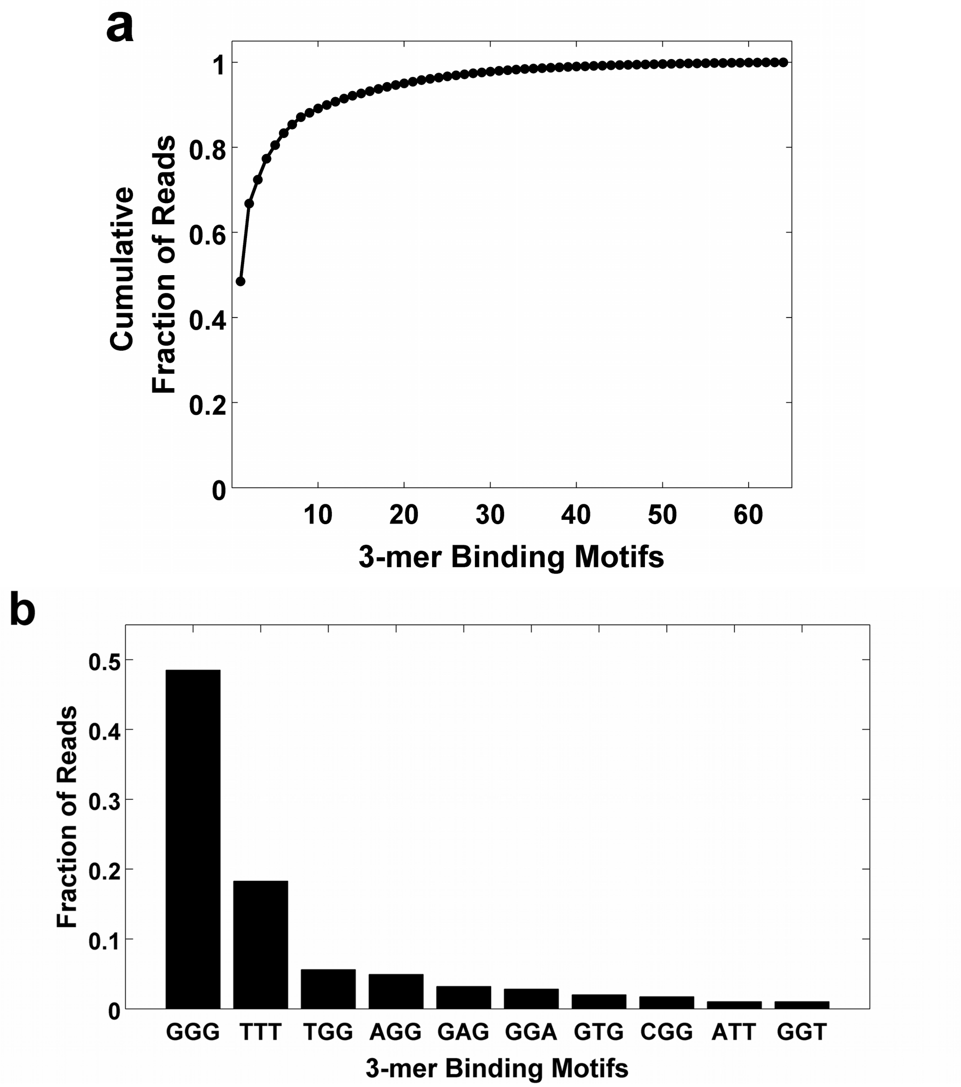
Supplementary Item	Title or Caption
Supplementary Table 1	Gene expression correlations between single cell and bulk sequencing data in E14 cells
Supplementary Table 2	Comparing mean single cell copy numbers to bulk copy numbers for read and coverage-based methods
Supplementary Table 3	Sequencing Statistics
Supplementary Table 4	Comparing cell-to-cell variability in copy numbers obtained from DR-Seq and DNA FISH in SK-BR-3 cells
Supplementary Table 5	Identifying single nucleotide variants (SNV) within coding regions of the genome for SK-BR-3 cells
Supplementary Note	Using length-based identifiers to minimize technical noise in single-cell mRNA sequencing data from DR-Seq
Supplementary Note	Quasilinear amplification does not introduce additional biases in the single-cell transcriptome data in DR-Seq
Supplementary Note	Reducing technical noise in single-cell gDNA sequencing data from DR-Seq
Supplementary Note	Copy number calling from single cell gDNA sequencing data
Supplementary Note	Error estimation in copy number calling from single cell gDNA sequencing data
Supplementary Note	Correlation between gene expression noise and copy number is not influenced by the mean expression level of genes


2 cDNA molecules from gene 1
1 cDNA molecule from gene 2
3′
5′
Initial amplification during pre-amp, molecules are primed at a random site providing a unique length identifier (LIx) for each molecule.
LI1
LI2
LI3
5′
3′
Further amplification during pre-amp, IVT and library prep may create biases in the number of molecules, but the number of length identifiers remains the same.
Ad-2
Ad-1x
cDNA from gene 1
cDNA from gene 2
19 molecules of cDNA 1
2 length identifiers of cDNA 1
3 molecules of cDNA 2
1 length identifier of cDNA 2

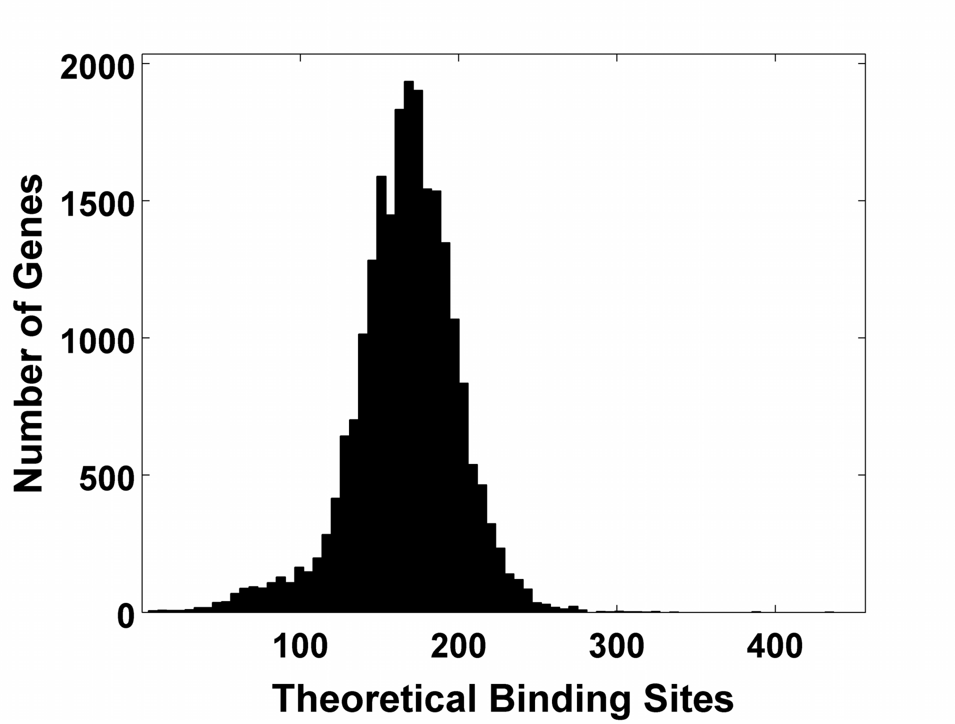
Supplementary Figure 1 | Genomic location of random priming events during quasilinear amplification can be used to minimize amplification biases and achieve resolution close to identifying unique cDNA molecules in DR-Seq. Schematic showing how the genomic location of random priming by adapter Ad-2 can be used to obtain length-based identifiers and uniquely tag original cDNA molecules. For example, the schematic shows a gene with 2 original cDNA molecules and another gene with one cDNA molecule in a cell (shown in green and red, respectively). During quasilinear amplification, these cDNA molecules are randomly primed at unique genomic locations. Depending on the quasilinear amplification round in which the original cDNA molecules were first primed, which results in PCR amplification for the remainder of the cycles (as well as potential amplification biases introduced during *in vitro* transcription and PCR amplification for Illumina library preparation), the finals reads obtained for the two genes could be quite distorted from the original ratio of 2:1. However, the unique priming location of each original cDNA molecule could be used to identify two unique length-based identifiers for gene 1 and one unique length-based identifier for gene 2. Thus, length-based identifiers could be used to remove duplicated reads, thereby significantly reducing PCR bias and achieving resolution close to identifying unique molecules before estimating transcript counts (RPM) for each gene.



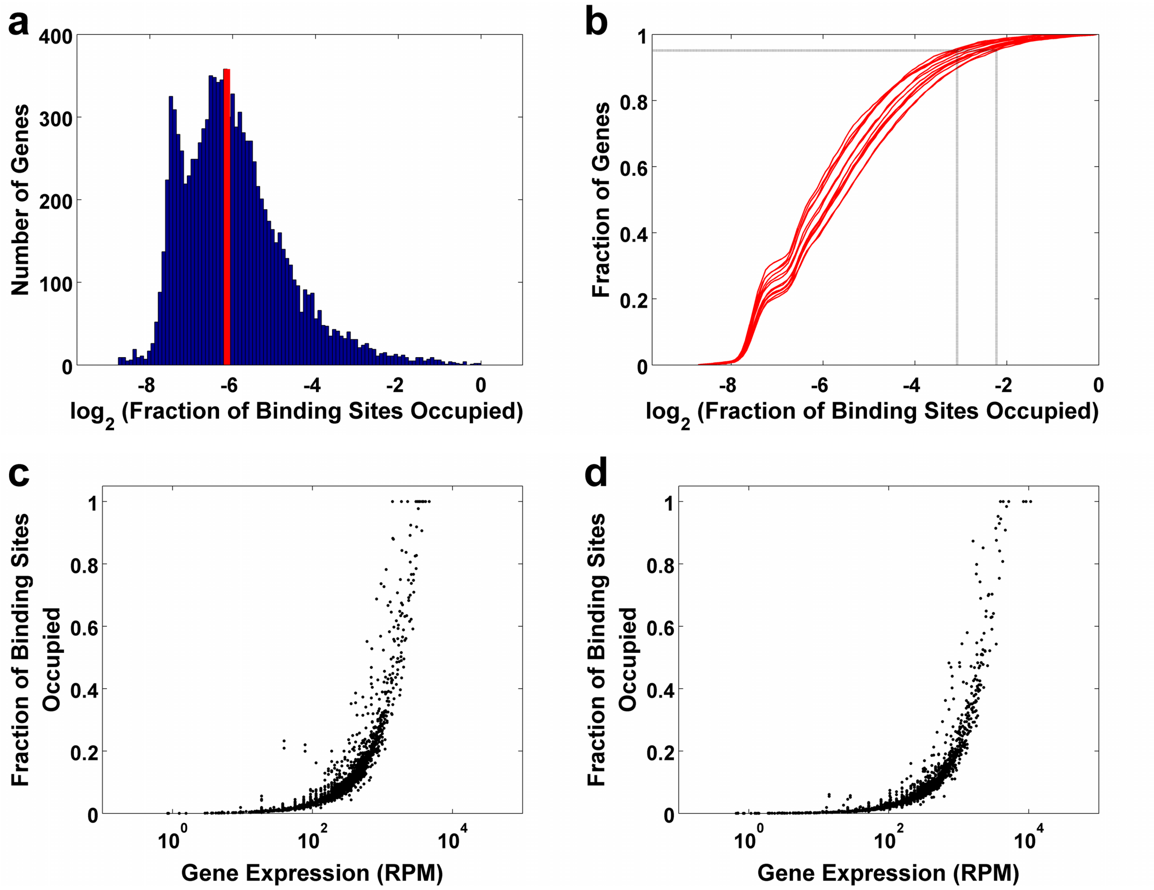
Supplementary Figure 2 | Original cDNA molecules are primed only once on average during quasilinear amplification in DR-Seq. Original cDNA molecules could theoretically be primed seven times during quasilinear amplification. However, length-based identifiers would be most accurate only if the original cDNA molecules are primed once on average during quasilinear amplification. Using a set of lowly expressed genes obtained from CEL-Seq (that are expressed at the level of one or zero transcripts per cell), the data shows that 99.2% of such genes are not primed or primed only once during quasilinear amplification in DR-Seq. Thus, length-based identifiers can be used to accurately count the original number of cDNA molecules in DR-Seq.



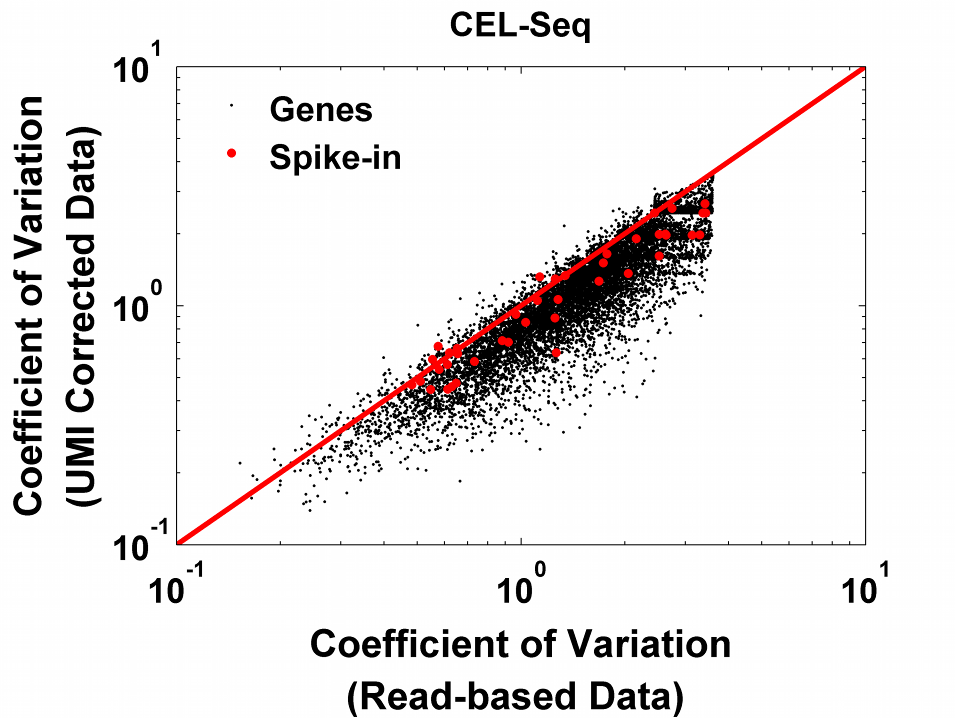
Supplementary Figure 3 | Identification of theoretical binding sites in the mouse transcriptome. **(a)** The 3' end of adapter Ad-2 has the sequence GGG or TTT. Since the adapter Ad-2 randomly primes cDNA molecules, potential 3-mer binding sites were identified using the DR-Seq data of single E14 cells. From the 64 (=4^3) possible 3-mer binding sites, the top 10 most prevalent 3-mer binding sites were found in approximately 90% of all reads. These 10 top binding sites were used as a stringent cutoff, rather than all positions along the cDNA, to estimate the theoretical maximum number of binding sites for each gene in the transcriptome (**Supplementary Fig. 4**). **(b)** As expected, GGG and TTT were the most prevalent 3-mer binding sites for adapter Ad-2, approximately comprising 70% of all reads. The bar graph shows the prevalence of the top 10 3-mer binding sites.



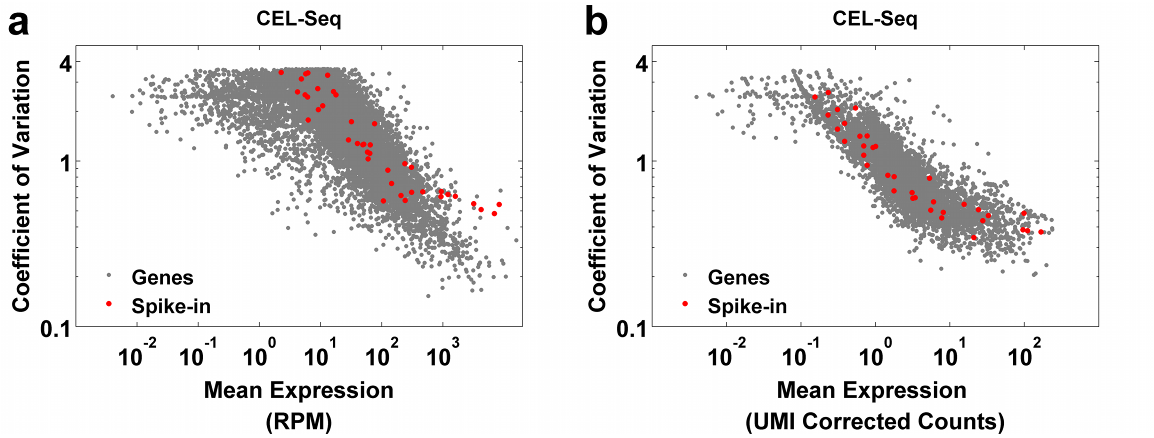
Supplementary Figure 4 | Distribution of theoretical number of binding sites for each gene in the mouse transcriptome. The histogram shows that a majority of genes contain 50-250 theoretical binding sites. Estimation of the theoretical number of binding sites for each gene is important to understand the number of original cDNA molecules that can be counted accurately by length-based identifiers before reaching saturation. Recently, 4-bp random barcodes were used as unique molecule identifiers (similar to the range of 50-250 theoretical binding sites) to accurately quantify the original number of cDNA molecules for a majority of genes without reaching saturation[1,2]. Since DR-Seq sequences the 3' end of transcripts, with the length of the sequencing library up to 1500 bp, the last 1000 nucleotides were used as a stringent estimator of the theoretical number of binding sites. We also found that our results estimating the extent of saturation of binding sites and therefore the accuracy in quantifying the original number of cDNA molecules was not sensitive to our choice of using 1000 nucleotides (**Supplementary Fig. 5**). Genes encoding for miRNA and snoRNA were not included in the histogram.



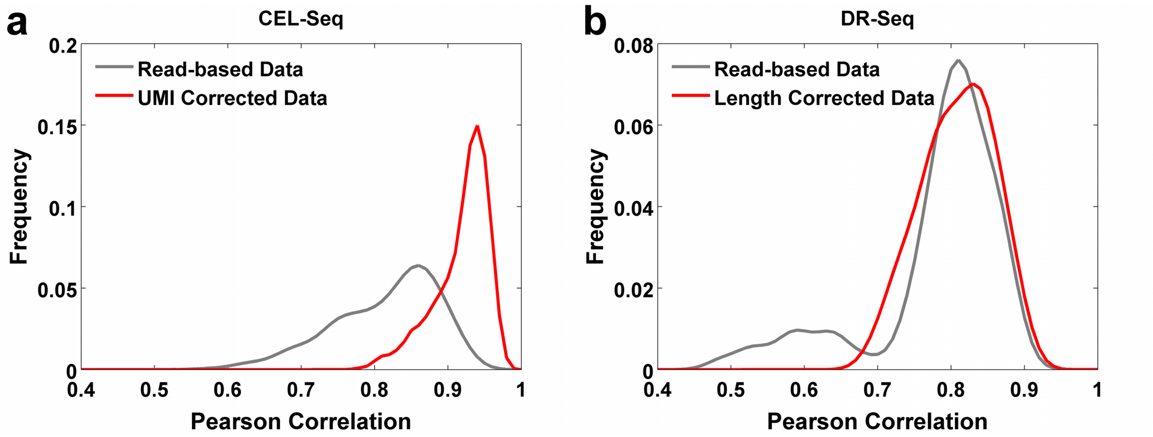
Supplementary Figure 5 | The number of original cDNA molecules can be accurately estimated by length-based identifiers in DR-Seq without reaching saturation. **(a)** The histogram shows the average fraction of binding sites that are occupied (i.e., the number of length-based identifiers found for a gene divided by the theoretical number of binding sites for that gene) for each gene in the single-cell E14 dataset processed by DR-Seq. The red line shows the median of the distribution. **(b)** For each E14 single cell, the figure shows the cumulative distribution of the fraction of binding sites that are occupied. For 95% of the genes in all single cells, the percentage of binding sites that are occupied is less than 15%. Thus, for a majority of the genes, the number of observed length-based identifiers are much smaller than the theoretical maximum, implying that the original number of cDNA molecules can be counted accurately without underestimation. **(c)** A plot of the fraction of binding sites occupied versus gene expression for one single cell shows that expect for a few highly expressed genes, length-based identifiers are not close to saturation and can therefore be used to accurately estimate the original number of cDNA molecules. **(d)** The plot shows data similar to (c) for another single cell.



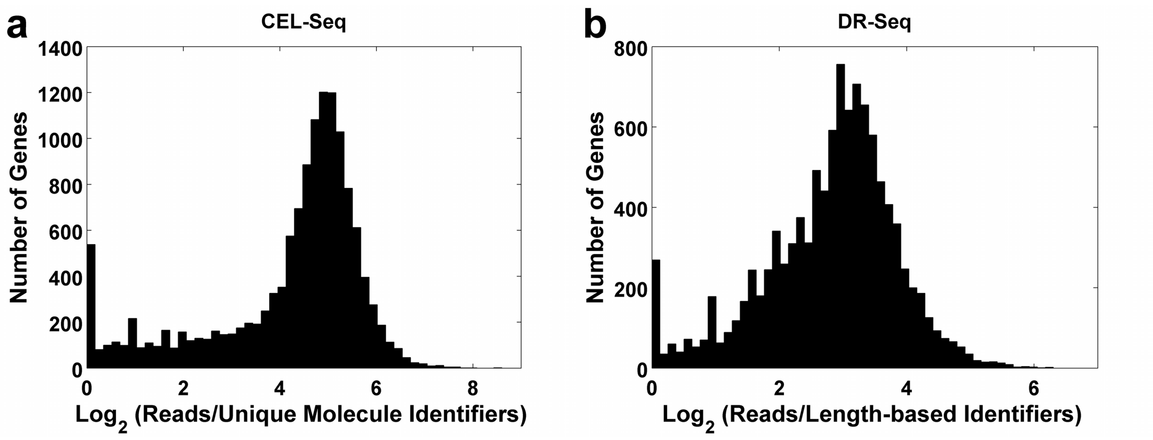
Supplementary Figure 6 | Reduction in cell-to-cell variability of gene expression after correcting read-based data with unique molecule identifiers in CEL-Seq. As shown previously[1], the coefficient of variation in gene expression reduces after correcting the expression data with unique molecule identifiers (UMI). Gene expression noise reduces for approximately 91% of the genes after correcting the read-based data with UMIs in CEL-Seq. This reduction in gene expression noise in similar to that observed in DR-Seq after correcting the read-based data with length-based identifiers (**Fig. 2a**). Thus, UMI correction in CEL-Seq and length-based correction in DR-Seq allows quantification of the original number of cDNA molecules in a cell and thereby reduces amplification biases and technical noise, resulting in the reduction of coefficient of variation in gene expression. All genes that are expressed (≥ 1 transcript) in at least 2 single cells are considered for this analysis. Endogenous genes and spike-ins are shown using black and red dots, respectively.



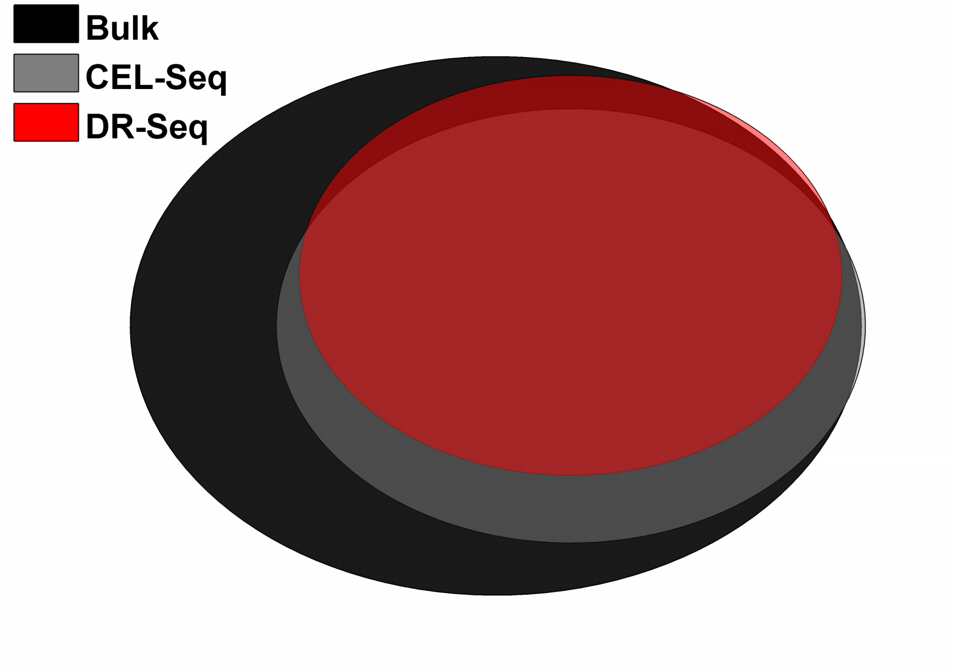
Supplementary Figure 7 | Unique molecule identifier correction reduces technical noise in CEL-Seq. **(a)** Coefficient of variation versus mean expression of genes for read-based data in CEL-Seq. Since all cells contain the same number of spike-in molecules (shown in red), these synthetic spike-ins should in general have low gene expression noise compared to endogenous genes for a given mean expression. Read-based data shows considerable noise in the detection of spike-ins, suggesting that the raw data contains significant amount of technical noise. **(b)** After correcting the data with UMIs in CEL-Seq, spike-in molecules in general have lower noise than most endogenous genes for a given mean level of expression. Thus, technical noise in reduced after using UMIs in CEL-Seq, similar to reduction in technical noise after correcting the data in DR-Seq using length-based identifiers (**Fig. 2b,c**). Endogenous genes are shown in gray.



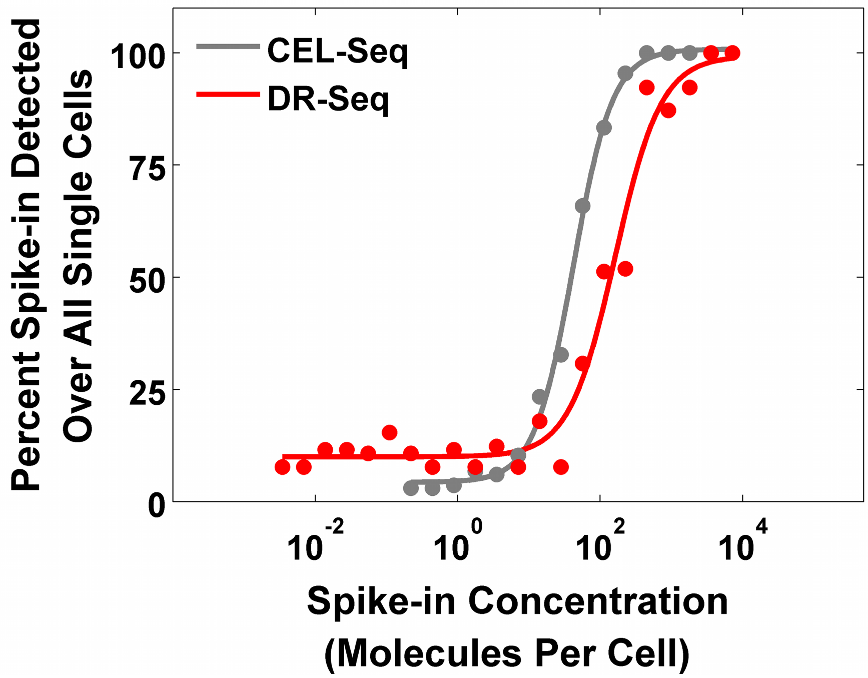
Supplementary Figure 8 | Pairwise Pearson correlations between single cells show reduction in technical noise after correcting the read-based data in CEL-Seq and DR-Seq. The figure show the distribution of all pairwise correlations in the expression of endogenous genes between single cells. **(a)** Correcting the read-based data in CEL-Seq using UMIs shows that the peak with the lower correlations in the read-based data disappears. **(b)** Similarly, the peak with the lower correlations in the read-based data in DR-Seq disappears after correcting the data using length-based identifiers. The peak with the lower correlations in the read-based data possibly arises from amplification and PCR biases in certain single cells, which make them appear more variable than most other single cells. After correcting the data in CEL-Seq and DR-Seq to more accurately reflect the original distribution of cDNA molecules within cells, the pairwise correlations between single cells improve and the peak with the lower correlations disappear due to the removal of PCR induced artifacts between single cells.



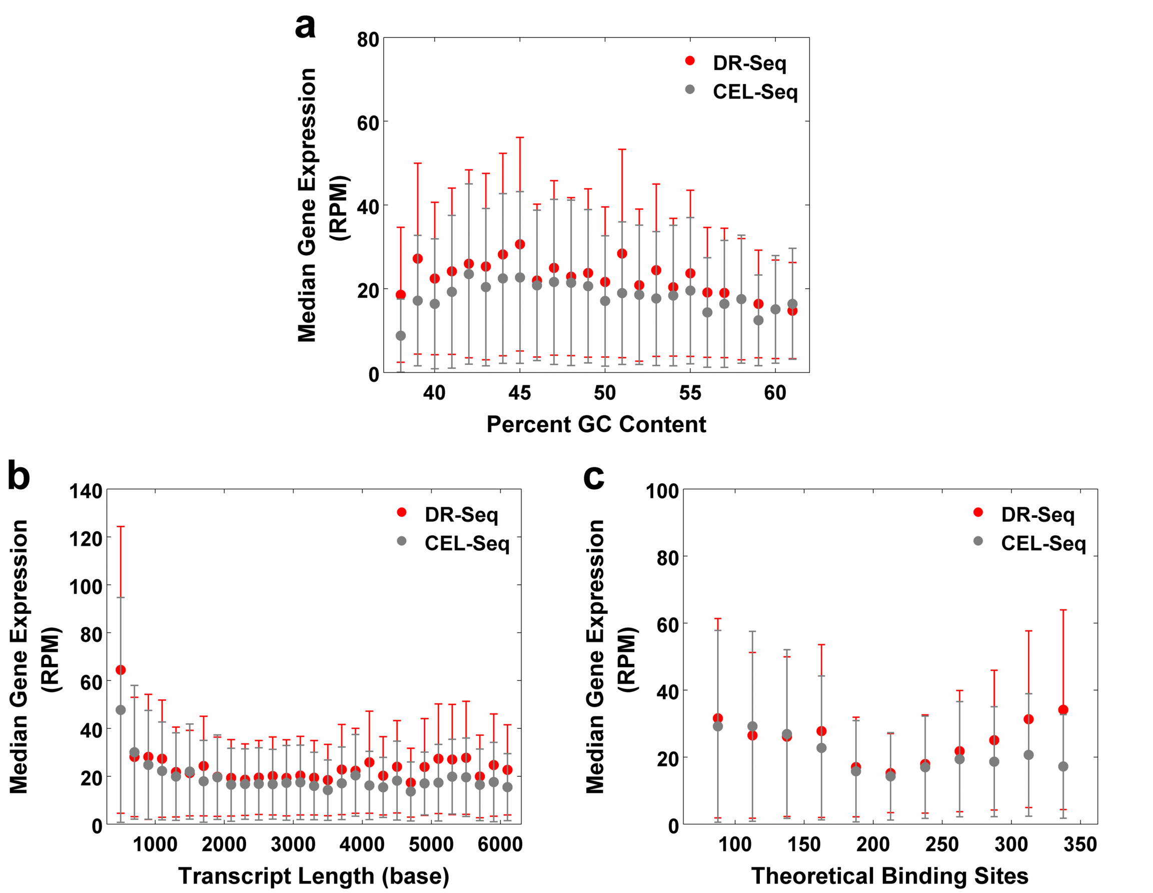
Supplementary Figure 9 | The entire complexity of the single-cell mRNA libraries are sequenced in CEL-Seq and DR-Seq. The figures show that for the given sequencing depth and mappability of reads (**Supplementary Table 3a**) for the single E14 cells, the entire complexity of the single-cell mRNA libraries are sequenced in CEL-Seq and DR-Seq. **(a)** The distribution of the ratio of reads to the number of UMIs for each gene in CEL-Seq shows that reads from a majority of genes have been over-sequenced 16-64 (=2^4-2^6) times. **(b)** Similarly, the distribution of the ratio of reads to the number of length-based identifiers for each gene in DR-Seq shows that reads from a majority of genes have been over-sequenced 4-16 (=2^2-2^4) times.



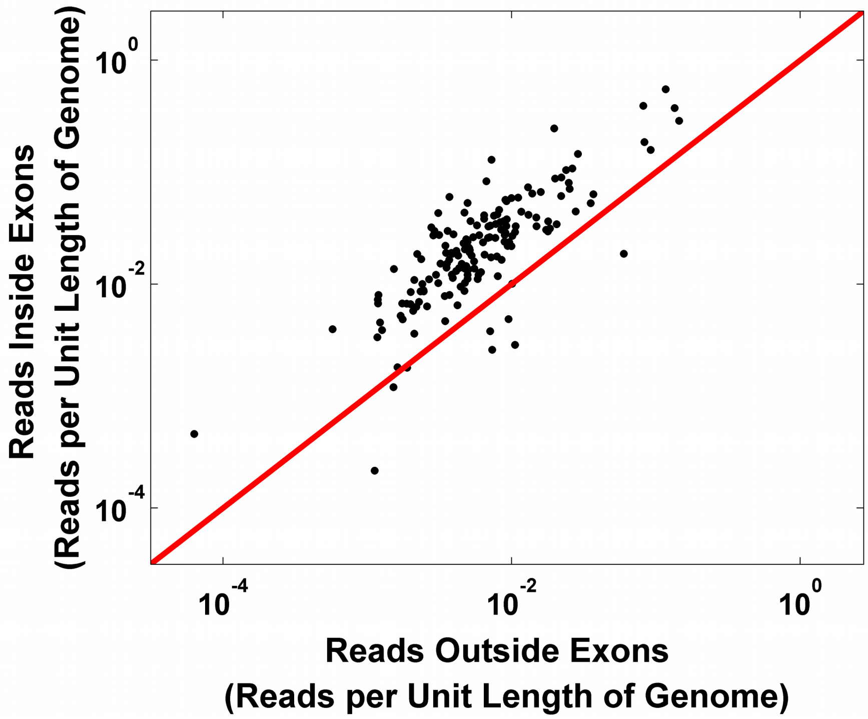
Supplementary Figure 10 | Number of genes identified by different sequencing methods. Bulk mRNA sequencing (black) identified 19,378 genes in E14 cells while CEL-Seq (gray) and DR-Seq (red) identified 12,573 and 10,674 genes, respectively. The venn diagram shows that most of the genes identified by CEL-Seq and DR-Seq are the same with 9,735 genes common between bulk sequencing, CEL-Seq and DR-Seq. For the overlap between CEL-Seq and DR-Seq, the p-value ($p < 1x10^{-16}$, hypergeometric test) demonstrates that the probability that this overlap in gene set occurs by random chance is extremely small. Similarly, the p-values for the overlap of genes identified from bulk mRNA sequencing and CEL-Seq or DR-Seq are also very small ($p < 1x10^{-16}$). When compared to bulk mRNA sequencing, the rates of false positive detection in DR-Seq and CEL-Seq are 0.27% and 0.14%, respectively. These low and similar false positive rates in both DR-Seq and CEL-Seq demonstrate that DR-Seq broadly displays the same performance as CEL-Seq in capturing the transcriptome of single cells.



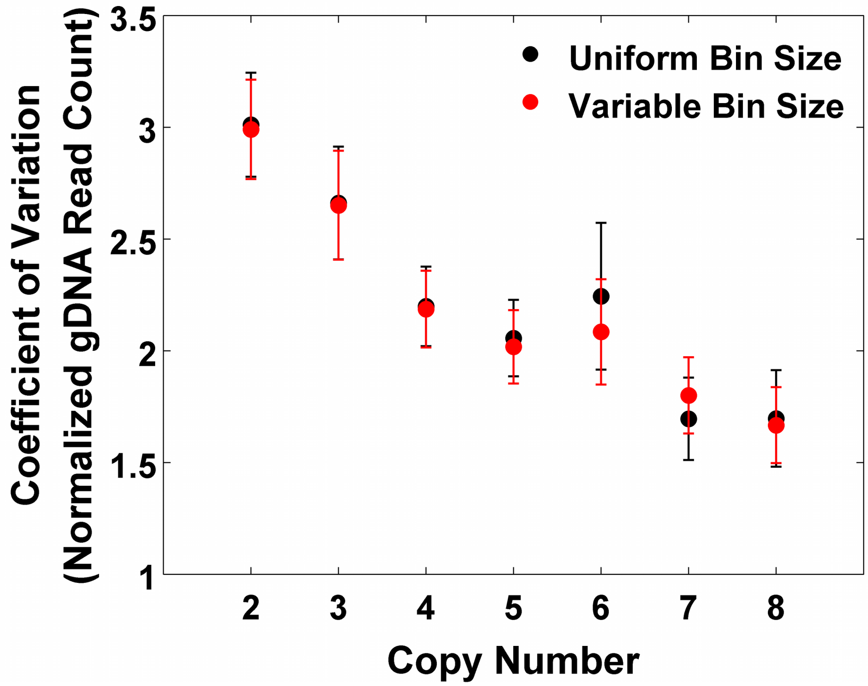
Supplementary Figure 11 | Comparing ERCC spike-in detection between CEL-Seq and DR-Seq. The ERCC spike-in mix contains 92 synthetic mRNA species over a large concentration range with a few spike-in species for each concentration. The figure shows the average percentage of spike-ins that are detected for different concentrations. Both CEL-Seq and DR-Seq perform similarly at higher concentrations. However, CEL-Seq fails to pick up the lowest concentration spike-ins that are identified in DR-Seq. The data shown is averaged over the 13 single E14 cells in DR-Seq and the 33 single cells in CEL-Seq. Experimental data points are shown as dots with the smooth curves serving as a guide to the eye.



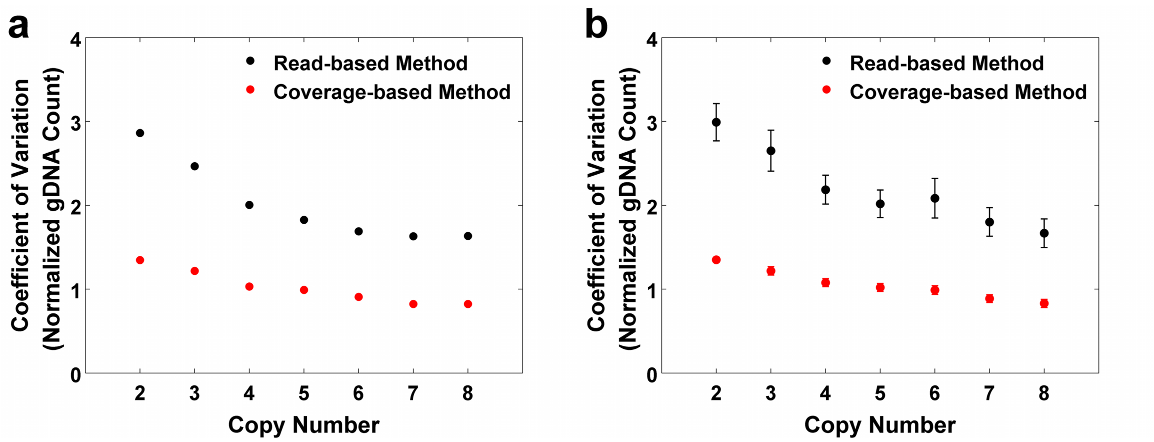
Supplementary Figure 12 | Quasilinear amplification in DR-Seq does not introduce additional biases in single-cell mRNA quantification. The panels use different metrics to show that quasilinear amplification in DR-Seq does not introduce additional biases in quantifying the transcriptome of single cells compared to CEL-Seq. **(a)** Since cDNA molecules are amplified by PCR during the quasilinear amplification steps, the median expression of genes was compared between DR-Seq and CEL-Seq over the entire range of GC content of transcripts. Both methods show similar median expression for all GC content. **(b)** Comparison of the median expression of genes for different transcript lengths between DR-Seq and CEL-Seq shows no statistically significant difference between the two methods. **(c)** Similarly, comparison of the median expression of genes over the whole range of theoretical binding sites shows that the number of binding sites within genes does not introduce bias in quantifying gene expression in DR-Seq over CEL-Seq.



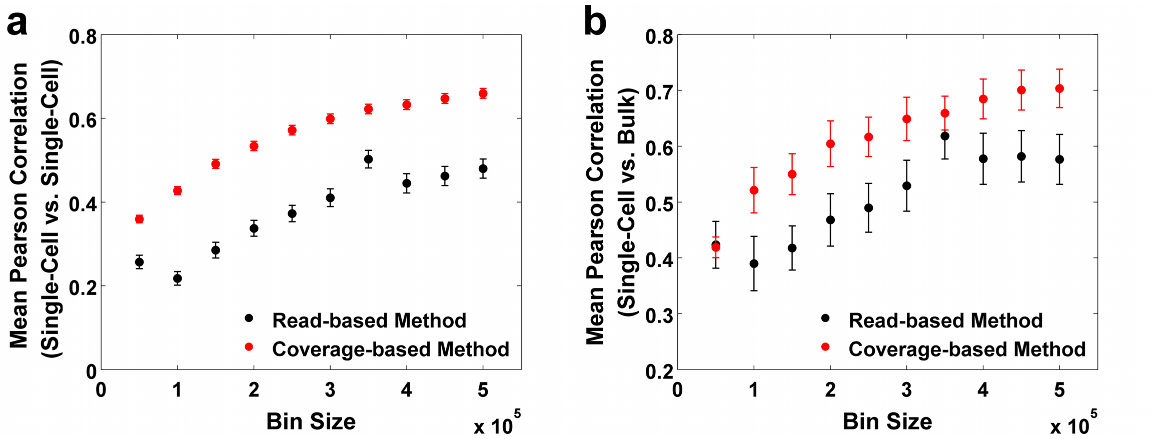
Supplementary Figure 13 | Reads from the gDNA fraction in DR-Seq are mapped to the genome after masking out the coding regions. Reads from the gDNA fraction in DR-Seq also contain reads that originate from cDNA molecules within coding regions. The figure shows data for each copy number segment within one single SK-BR-3 cell. For most of the segments, the normalized number of reads inside exons exceed reads outside exons. Thus, masking out coding regions of the genome is important to discard excess ambiguous reads within these regions that might originate from cDNA or gDNA. Since the coding region makes up approximately 2% of the genome, such as masking strategy does not influence copy number calling over large genomic regions. Finally, a strategy was developed for binning gDNA reads that accounted for the masking of the coding regions of the genome (**Supplementary Fig. 14**).



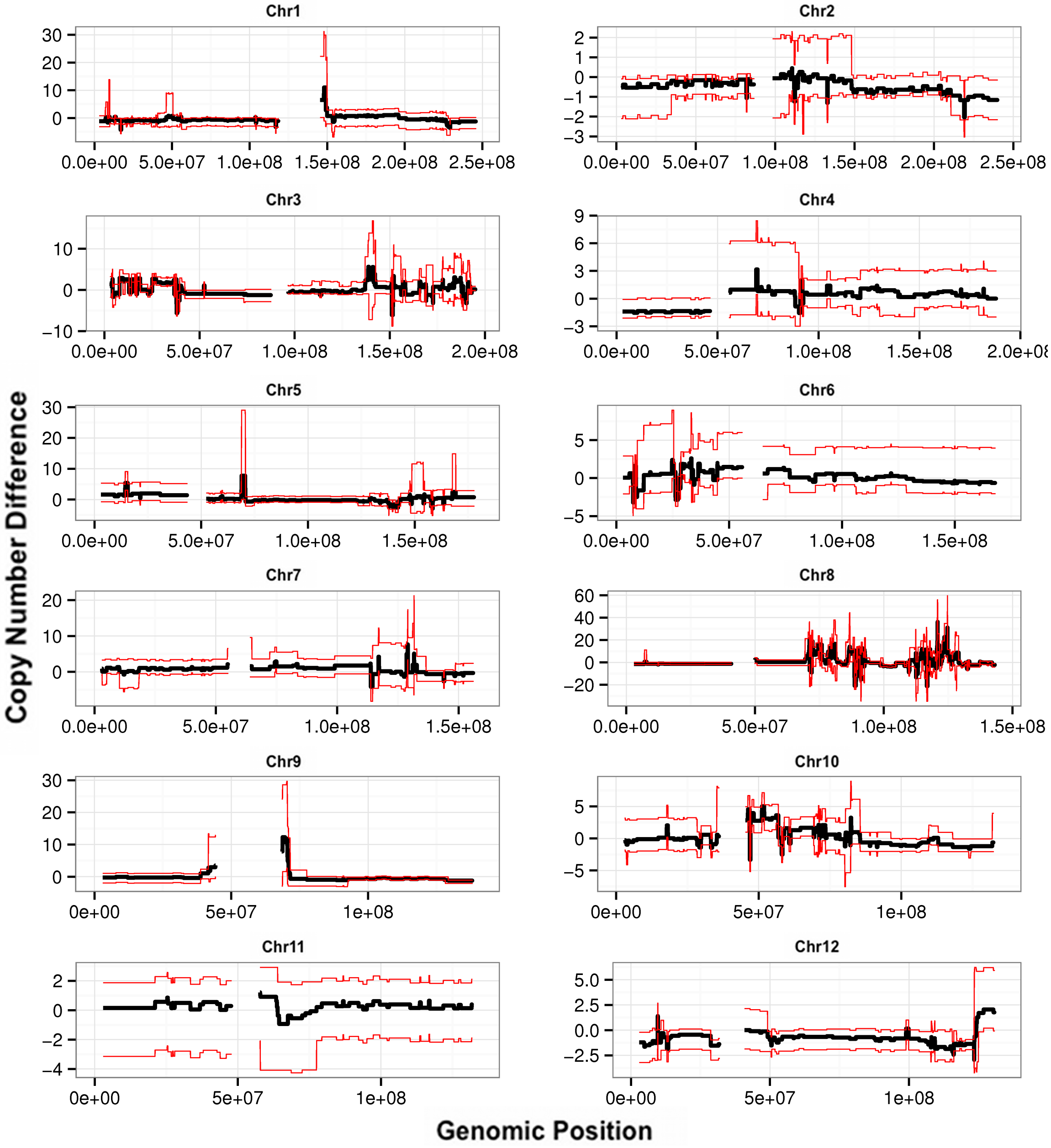
Supplementary Figure 14 | Strategy for binning gDNA reads in DR-Seq. To account for the masking of the coding regions, the genome was divided into unequal bins such that the length of each bin after excluding the coding regions was the same as that obtained when binning the genome uniformly. Such a variable binning strategy provides a more accurate description of the distribution of reads within bins since the coding regions have been masked from the genome. The figure shows bin-to-bin technical noise in read counts using a uniform binning versus variable binning strategy for different copy number regions (as determined from bulk gDNA sequencing) within single SK-BR-3 cells. The figure shows data for a uniform bin size of 50 kb or a variable bin size such that the length of each bin is 50 kb after excluding coding regions within the bins.

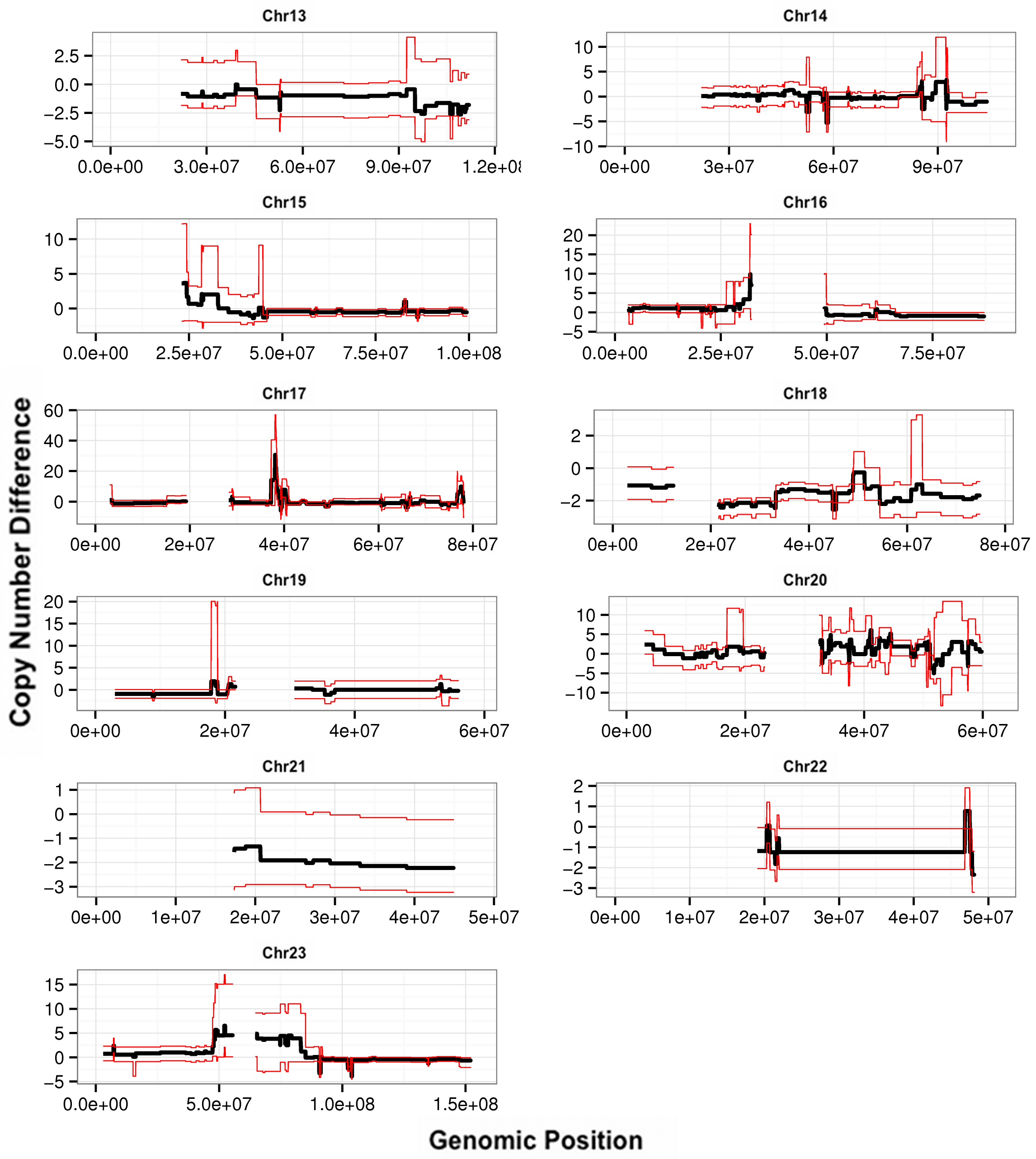


Supplementary Figure 15 | Coverage-based method reduces technical noise in single-cell gDNA sequencing data from DR-Seq. **(a)** Regions with the same copy number, as estimated from bulk sequencing, should contain bins with similar number of reads. Estimating bin-to-bin variability (as quantified by coefficient of variation) in read counts within different copy number regions for a single SK-BR-3 cell (SC13) shows that the coverage-based method displays much lower technical noise than the read-based method. **(b)** Data from all the single SK-BR-3 cells show that the coverage-based method reduces amplification biases and technical noise nearly two-fold compared to the read-based method. Error bars indicate the standard error of the mean over all the single cells.

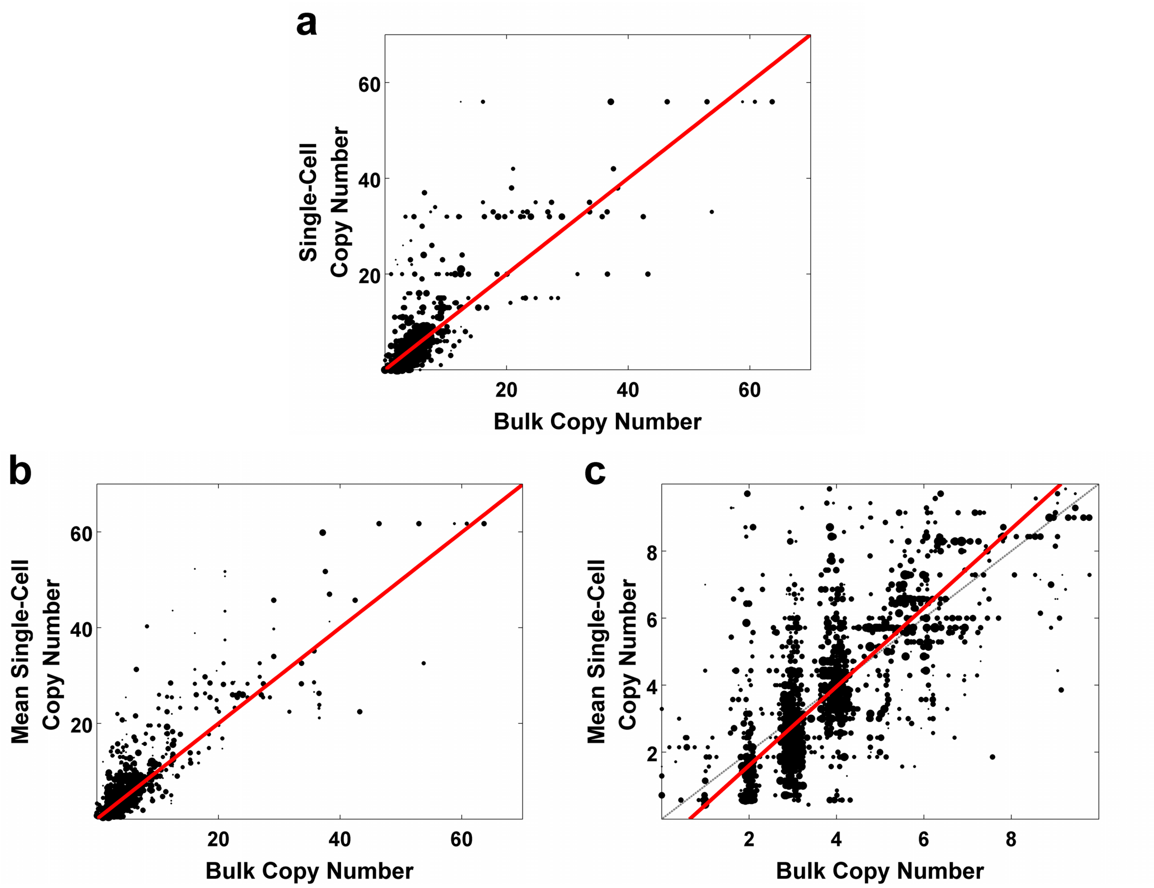


Supplementary Figure 16 | Coverage-based method improves Pearson correlation between single cells. **(a)** For different bin sizes, the mean Pearson correlation between all pairs of single cells increase in the coverage-based method compared to the read-based method. The higher Pearson correlations suggest that the coverage-based method reduces technical variability arising from amplification biases between single cells. The error bars indicate the standard error of the mean over all pairwise Pearson correlations between single SK-BR-3 cells. **(b)** The mean Pearson correlation between single SK-BR-3 cells and bulk gDNA sequencing is higher for the coverage-based method compared to the read-based method over all bin sizes. The error bars indicate the standard error in estimating the mean Pearson correlations.

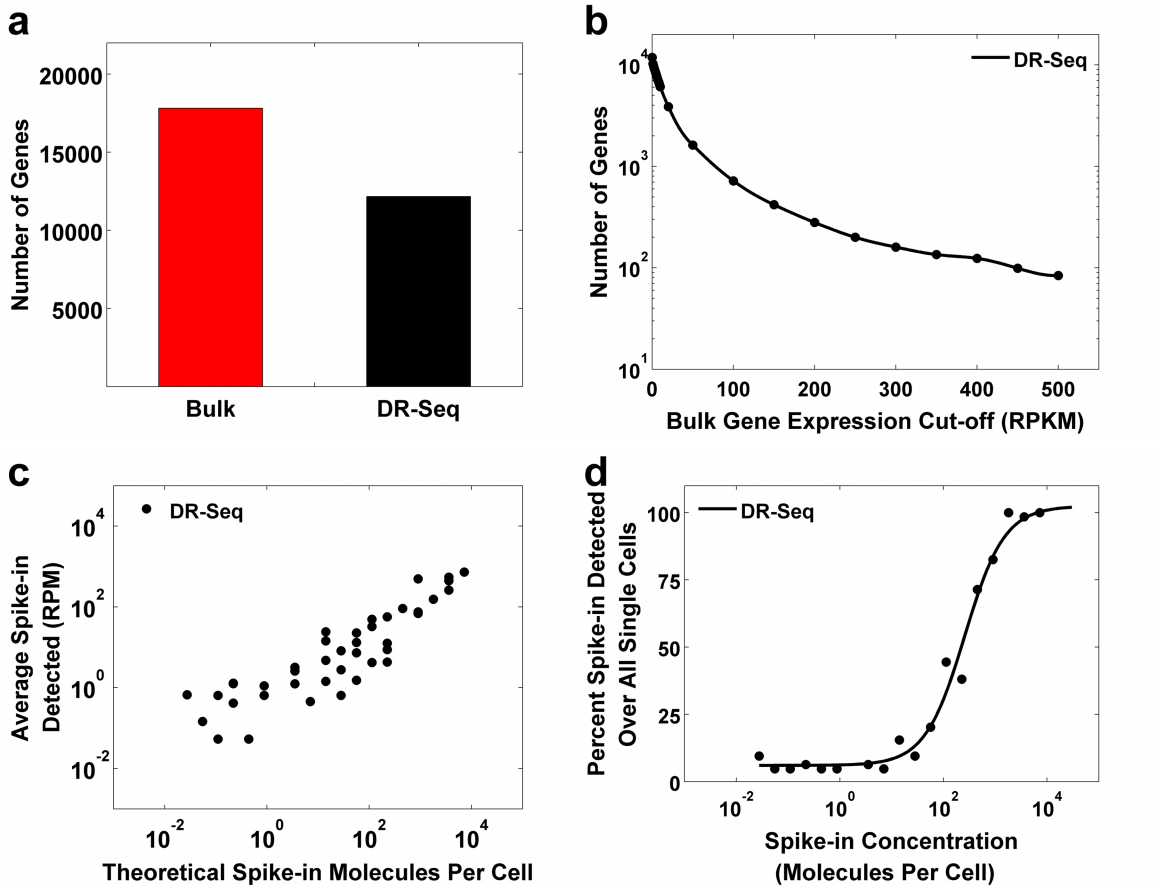

Copy Number Difference
Chr1
30
20
10
0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
Chr2
2
1
0
−1
−2
−3
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
Chr3
10
0
−10
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Chr4
9
6
3
0
−3
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+0
Chr5
30
20
10
0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Chr6
5
0
−5
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Chr7
20
10
0
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Chr8
60
40
20
0
−20
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Chr9
30
20
10
0
0e+00
5e+07
1e+08
Chr10
5
0
−5
0e+00
5e+07
1e+08
Chr11
2
0
−2
−4
0e+00
5e+07
1e+08
Chr12
5.0
2.5
0.0
−2.5
0e+00
5e+07
1e+08
Genomic Position



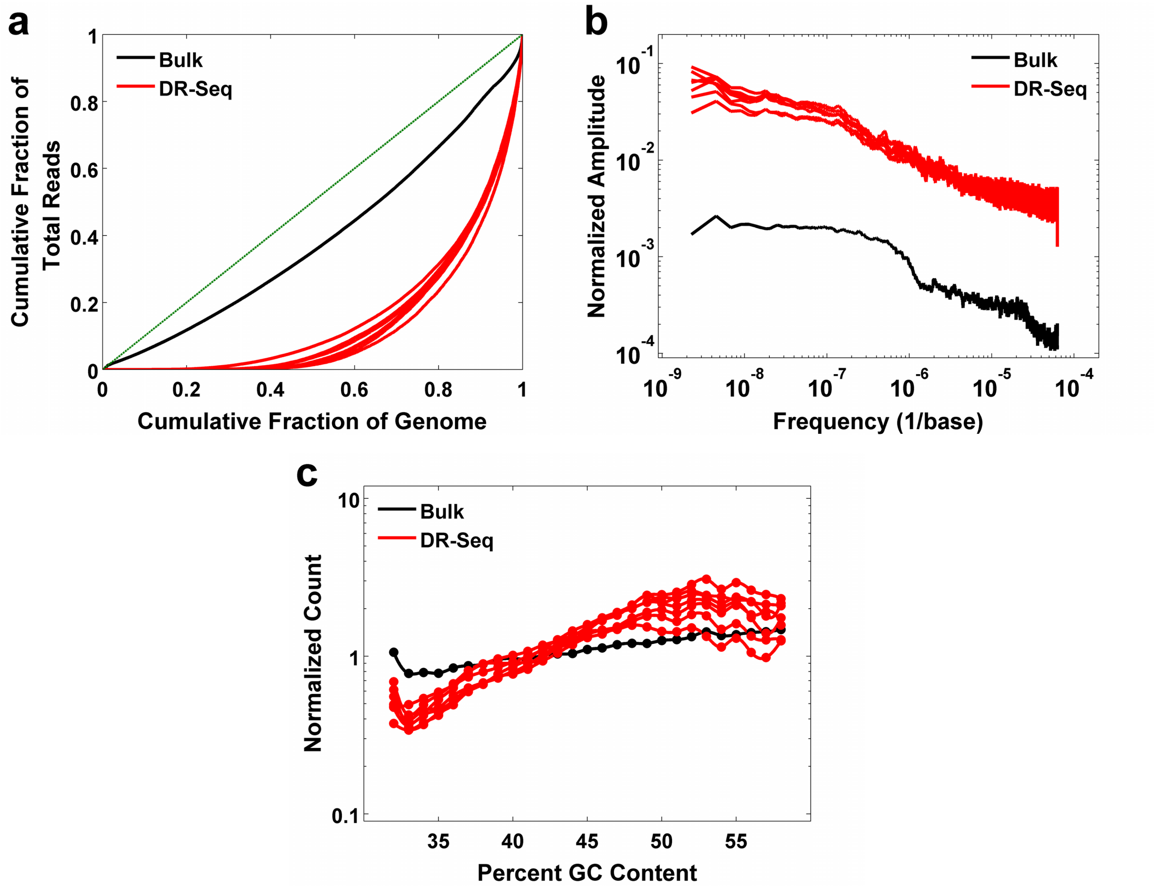
Supplementary Figure 17 | Overview of copy number calling in single SK-BR-3 cells and bulk gDNA sequencing. The figure provides an overview of the cell-to-cell variability in copy numbers and regions of the genome with consistent bias in copy number calling compared to bulk gDNA sequencing. For each chromosome, the bulk copy numbers were subtracted from the mean copy numbers obtained from single cells and plotted against genomic coordinates to compare bulk data to single-cell copy number calls. Within each panel, 3 Mb from the telomeres and from either side of the centromere are not shown. For a majority of the chromosomes, this difference in copy number (black line) was generally close to zero. Also, as expected from previous single cell gDNA sequencing studies, deviations from zero were found in centromeric and telomeric regions[3]. These artifacts are a consequence of sequencing reads not mapping accurately within these highly repetitive regions with unreliable genome assemblies. Further, deviations from zero observed for a few chromosomal regions possibly arise from the small sample size of single cells analyzed or from biases in regions with high copy numbers (**Supplementary Fig. 22**). Taken together, these results visually show that single-cell gDNA sequencing using DR-Seq broadly captures the same copy number variations as that observed in bulk sequencing (also see **Supplementary Fig. 18**). To visually assess regions of high cell-to-cell variability in copy numbers, the bulk copy number was subtracted from the highest and lowest copy number for a region across the single cells. The figure shows these as red lines over the entire genome. Several regions, such as those within chromosome 3,5,6,14 and others, display deviations from zero. This cell-to-cell variability in copy numbers was validated for four genomic loci using DNA fluorescence *in situ* hybridization (FISH). There was close agreement between the DR-Seq and DNA FISH results, further demonstrating that DR-Seq is sensitive enough to quantify cell-to-cell variability in copy numbers (**Supplementary Fig. 24** and **Supplementary Table 4**).



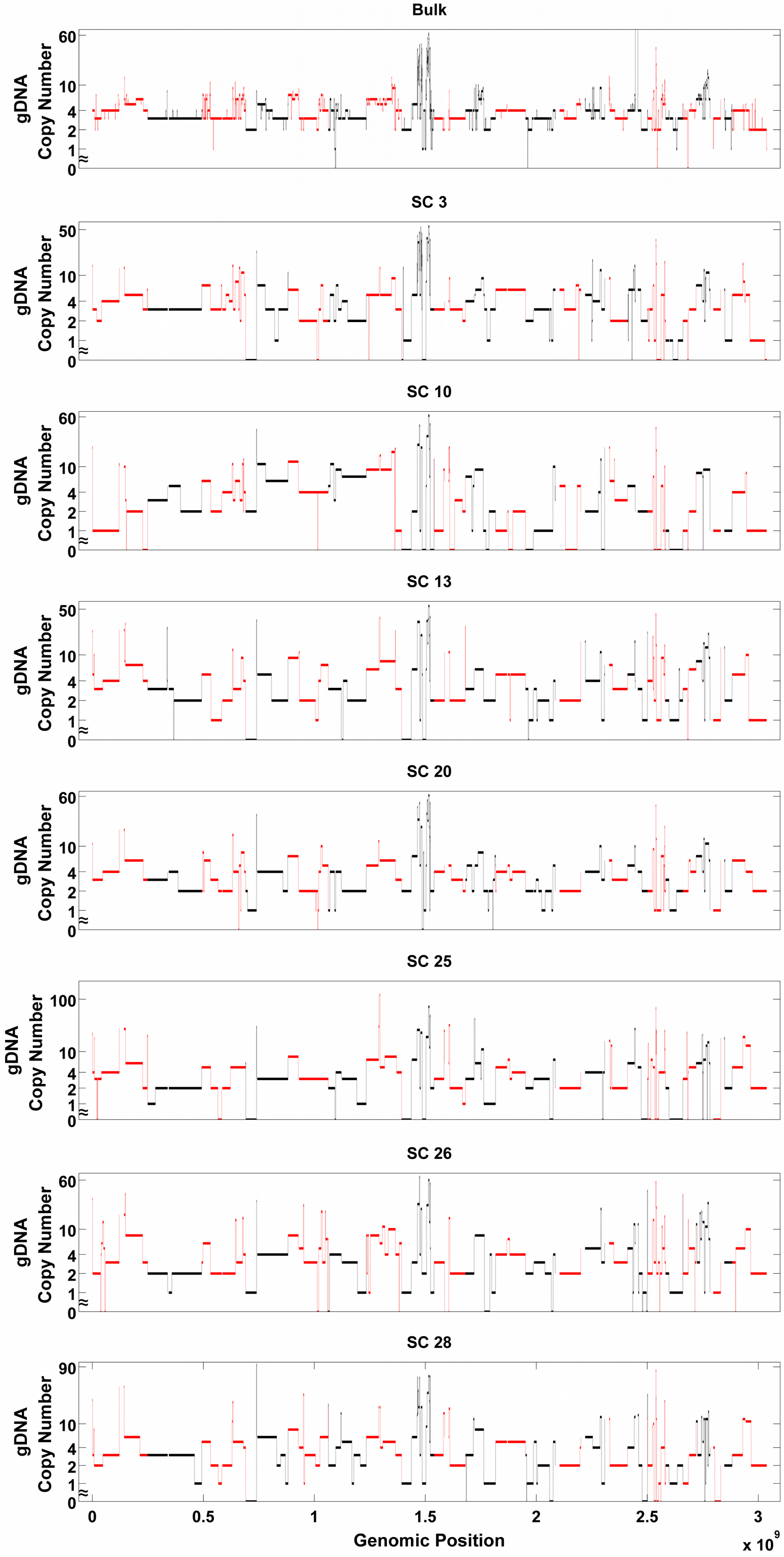
Supplementary Figure 18 | Comparison of copy number variations in single SK-BR-3 cells to bulk gDNA sequencing. The GC corrected reads from the coverage-based method of a single cell are first used to identify breakpoints using the circular binary segmentation algorithm. The median counts of each segment are then used to estimate copy numbers in single cells (see **Supplementary Note**). **(a)** Copy numbers from a single cell (SC13) are plotted against the bulk copy numbers. The weighed best-fit line is shown in red. The slope is close to 1, suggesting that copy number variations observed in bulk sequencing can be captured in single-cell gDNA sequencing using DR-Seq. **(b)** While single cells show variability in copy numbers, the mean copy numbers of single cells plotted against the bulk copy numbers are expected to be distributed around the diagonal. The slope of the weighed best-fit line (red) is 0.98, suggesting that DR-Seq can broadly capture the same copy number variations in single cells as compared to bulk sequencing. **(c)** Since the first 10 copy numbers are the most prevalent states in the SK-BR-3 cell line, focusing on these copy numbers show that the slope of the weighed best-fit line is 1.17, close to the expected value of 1. The diagonal is shown in gray and the weighed best-fit line is shown in red. The diameter of each point in the figure is proportional to the size of the genomic segment.



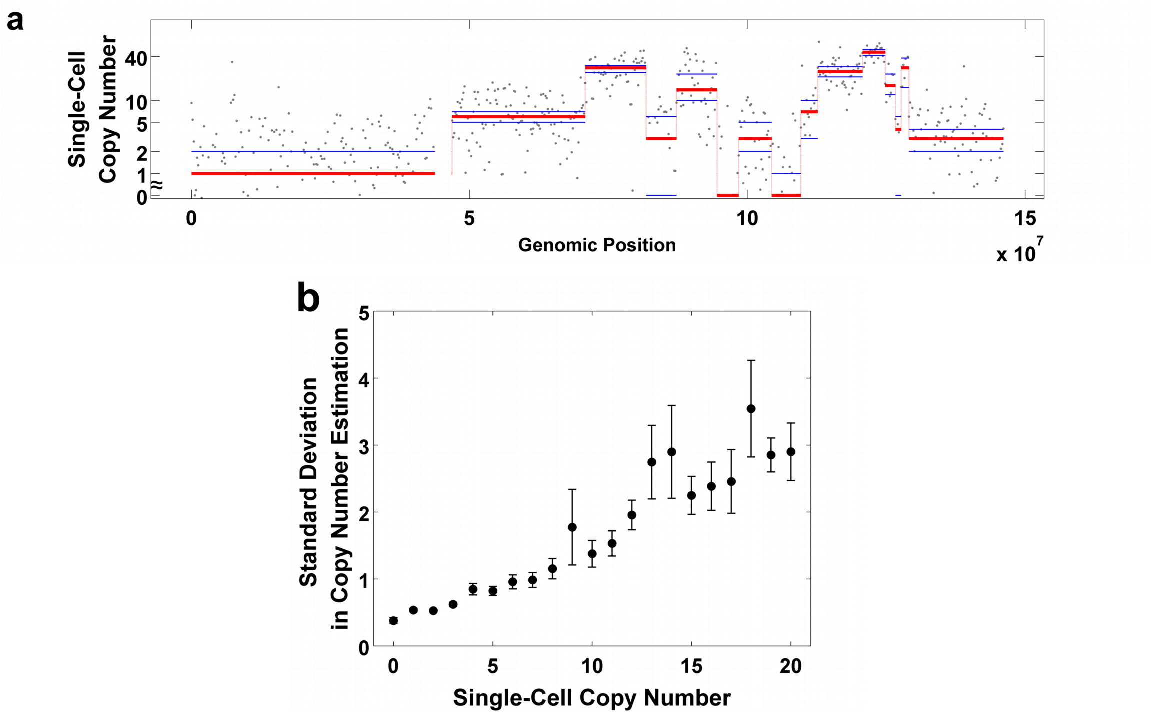
Supplementary Figure 19 | Single cell mRNA sequencing results for SK-BR-3 cells using DR-Seq. **(a)** 12,205 genes were identified in 21 single SK-BR-3 cells, similar to the number of genes identified in E14 cells (**Fig. 2d (Inset)**). **(b)** Figure shows the number of genes identified in the single cell mRNA sequencing data from DR-Seq above different expression thresholds of bulk mRNA sequencing. **(c)** The average number of spike-in molecules detected per cell correlates well with the expected concentration of each spike-in. **(d)** As with the E14 single cells (**Supplementary Fig. 11**), spike-ins introduced with the SK-BR-3 single cells are detected over the entire range of concentrations. Experimental data points are shown as black dots with the smooth curve serving as a guide to the eye. This suggests that single cell mRNA sequencing using DR-Seq performs well for both the mouse E14 and human SK-BR-3 cell lines.



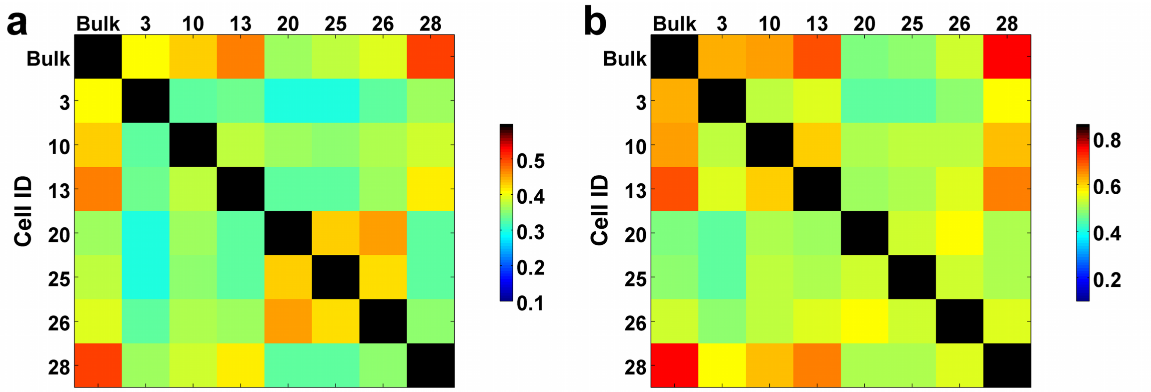
Supplementary Figure 20 | Single cell gDNA sequencing results for SK-BR-3 cells using DR-Seq. **(a)** Lorenz plots for the 7 single SK-BR-3 cells show similar biases in genome coverage as the E14 single cells (**Fig. 2g**). The green line indicates the theoretical limit with reads uniformly distributed over the entire genome. **(b)** Power spectrum analysis for the 7 single cells show that biases in read distribution over different genomic length scales are similar to those observed in single E14 cells (**Fig. 2h**). **(c)** Normalized read distributions deviate from 1 at low and high GC content. This bias in sequencing low and high GC content regions is similar to that observed in E14 cells (**Fig. 2i**). These results suggest that single cell gDNA sequencing using DR-Seq shows consistently reproducible results across different cell types.


Bulk
gDNA
Copy Number
60
10
4
2
1
0
SC 3
50
10
4
2
1
0
SC 10
60
10
4
2
1
0
SC 13
50
10
4
2
1
0
SC 20
60
10
4
2
1
0
SC 25
100
10
4
2
1
0
SC 26
60
10
4
2
1
0
SC 28
90
10
4
2
1
0
0
0.5
1
1.5
2
2.5
3
Genomic Position
x 10^9

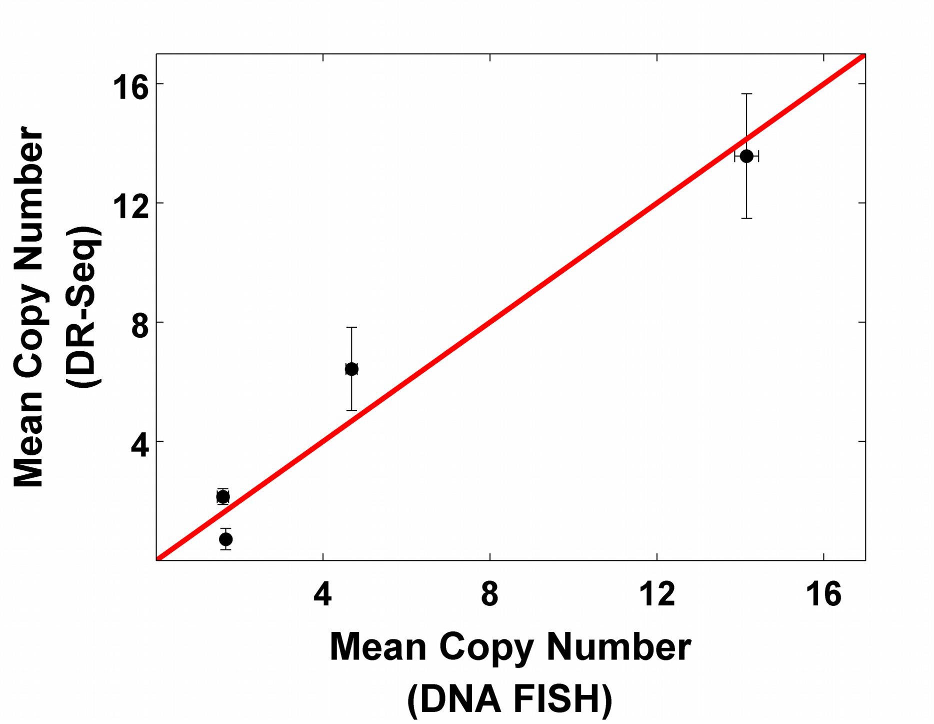
Supplementary Figure 21 | Copy number variations in single SK-BR-3 cells. The top panel shows copy numbers over the entire genome for the bulk sequencing data. Alternate chromosomes are shown in red and black. The estimated copy numbers for each of the seven single cells are shown in the lower panels. The figure shows that most of the large chromosomal changes observed in the bulk sequencing data are also well captured in the single cell data.



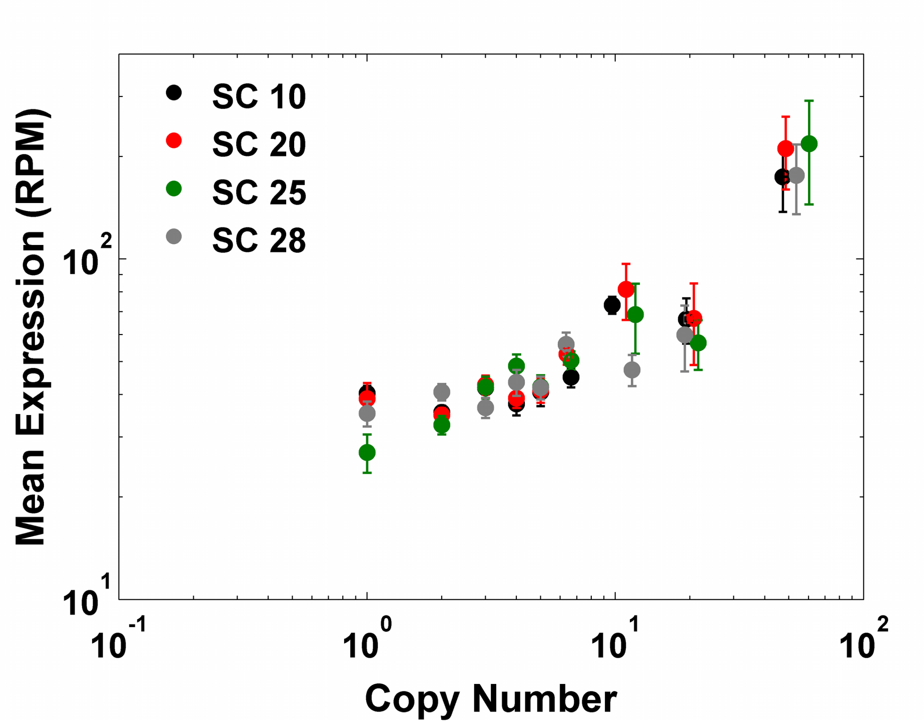
Supplementary Figure 22 | Error estimation in copy number calling from single cell gDNA sequencing data in DR-Seq. As single cell gDNA sequencing data is noisier than bulk gDNA sequencing data, a model was developed to estimate errors involved in calling copy numbers in single cells (**Supplementary Note**). **(a)** Estimated copy numbers over chromosome 8 are shown in red for a single SK-BR-3 cell (SC13). The upper and lower confidence intervals of the copy numbers are shown in blue for all the genomic regions in chromosome 8. **(b)** The figure shows systematic analysis of errors in estimating copy numbers over all the single cells. The mean standard deviations for different copy number loci in the SK-BR-3 genome are shown. For the first 8 copy number loci, the mean standard deviations in estimating copy numbers approach 1. These errors show that copy numbers can be reliably estimated in single cells using DR-Seq. Since the number of data points with high copy numbers are sparse, the x-axis is truncated to the first 20 copy numbers. Error bars represent the standard error in estimating the mean standard deviations.



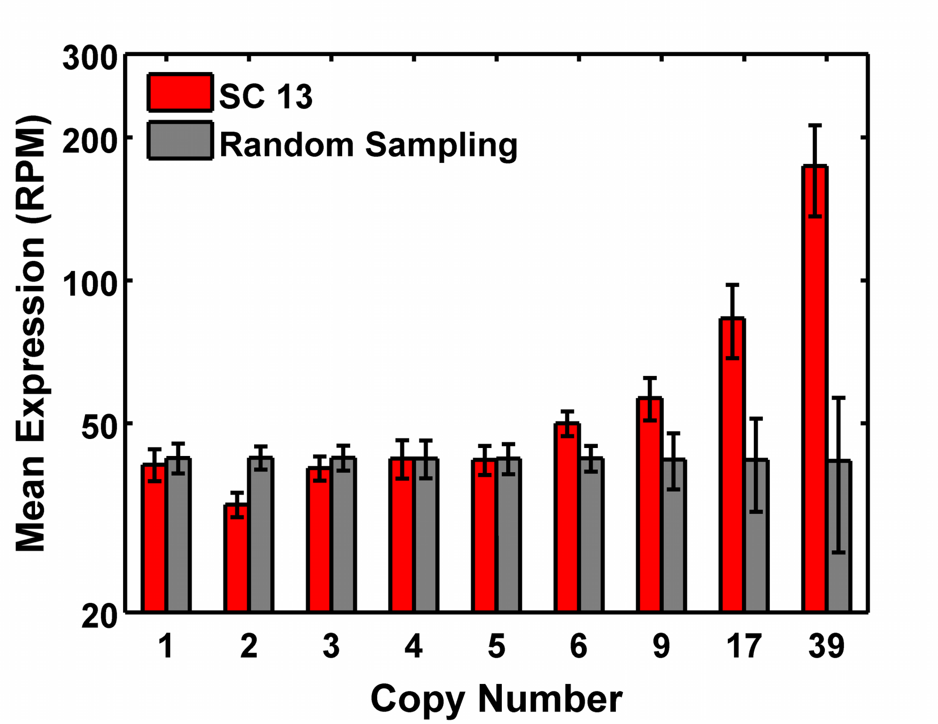
Supplementary Figure 23 | Pearson correlations of gDNA read counts between bulk and single SK-BR-3 cells. The figure shows Pearson correlations between single cells and single cells and bulk gDNA sequencing for **(a)** 50 kb and **(b)** 200 kb bin sizes. For different bin sizes, the cell-to-cell pairwise correlations are similar to each other and to bulk sequencing.



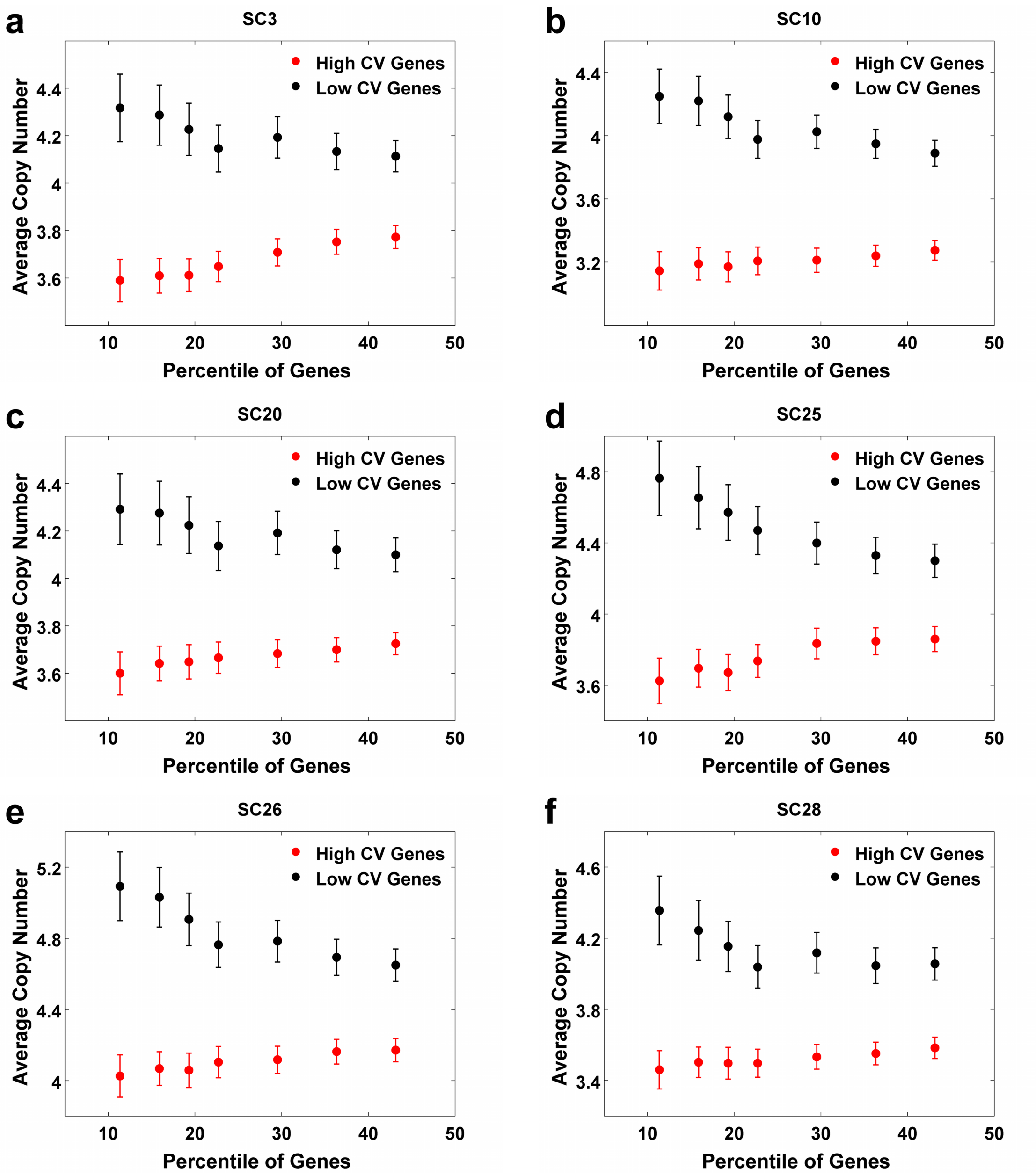
Supplementary Figure 24 | Comparing copy number estimation between DR-Seq and DNA FISH in single SK-BR-3 cells. Comparison of the mean copy number between DR-Seq and DNA FISH shows good agreement. Four genomic loci, spanning a large range of copy numbers were compared between DR-Seq and DNA FISH. Two regions had low copy numbers (*FHIT* and *HTT*), one region had intermediate copy number (*ZMIZ1*) and one region had high copy number (*CCDC40*). Further, cell-to-cell variability in copy numbers were also not statistically different between DR-Seq and DNA FISH for these 4 loci (**Supplementary Table 4**).The red line indicates the diagonal.



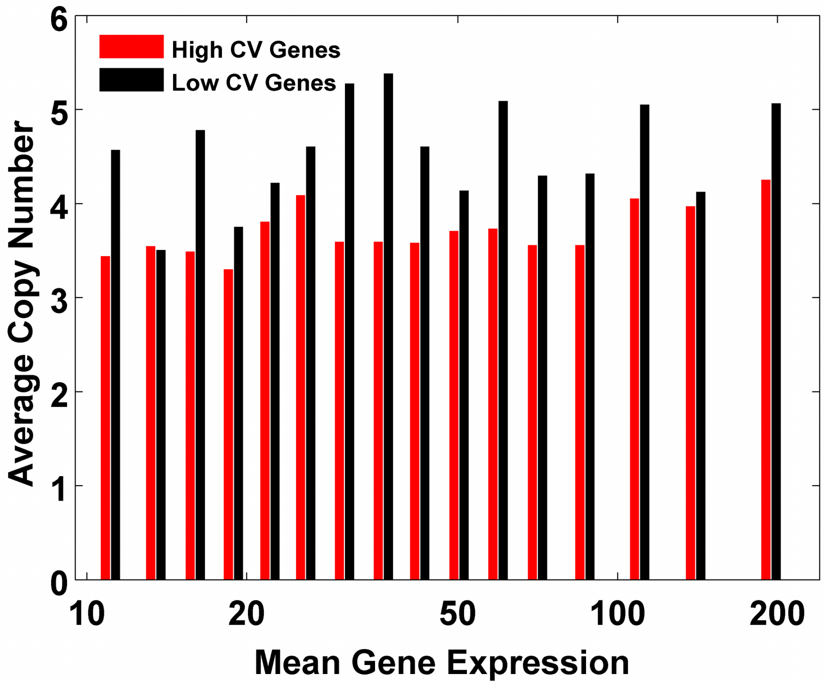
Supplementary Figure 25 | Genome-wide quantification of mean expression of genes within different copy number regions in single SK-BR-3 cells. Data from the four other single cells also show that the average expression of genes increases monotonically with increasing copy numbers (also see **Fig. 3b**). This demonstrates that copy number has a strong influence on gene expression and that DR-Seq can reliably detect both changes in copy number and transcript count from the same cell. Error bars represent standard error in estimating the mean obtained by bootstrapping the data.



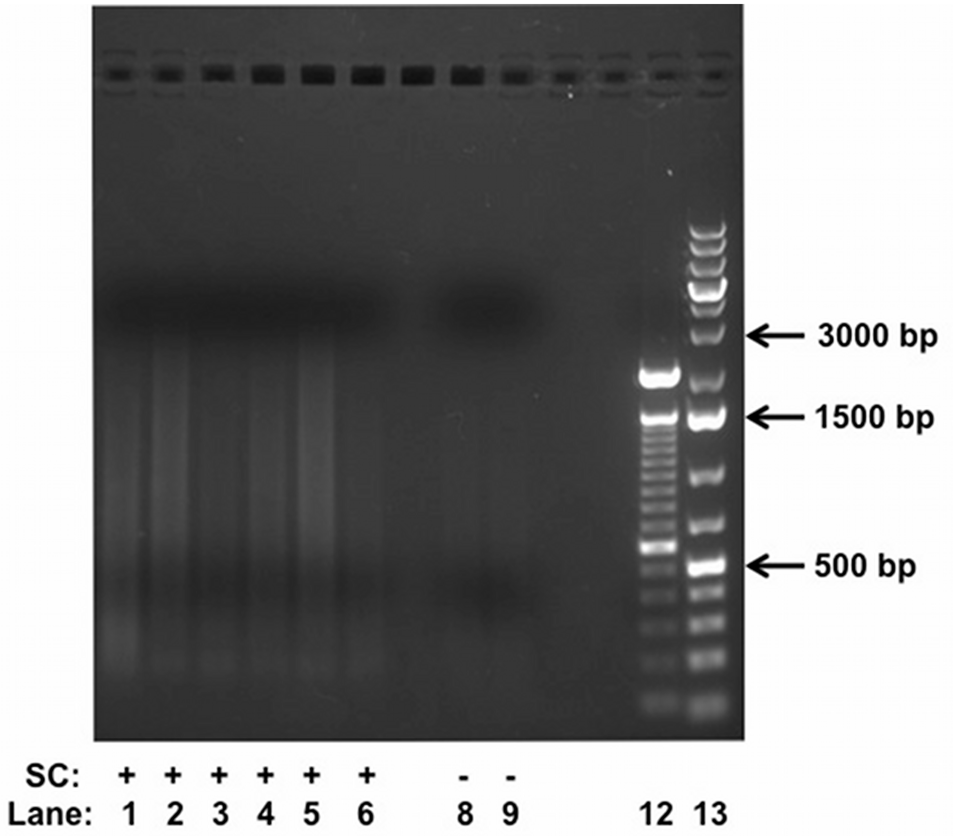
Supplementary Figure 26 | Comparing mean expression of genes within different copy number regions to random sampling of genes. Red bars show the mean expression of genes from different copy number regions for one single cell (SC13) (same as **Fig. 3b**, red dots). The same number of genes for each copy number in SC13 were then randomly sampled from the entire genome to estimate the mean (gray bars). The red bars show a significant increase in mean expression with copy number compared to the random sampling of genes. Similar trends were seen for all the other single cells. The mean and the standard error for the random sampling were obtained by bootstrapping the data.



Supplementary Figure 27 | Influence of copy number on gene expression noise. (**a-f**) Figure panels show the affect of copy number on gene expression noise for the other single cells (also see **Fig. 3c**). As observed in **Fig. 3c**, we found that for all the single cells, reduced copy numbers were associated with increased gene expression noise (quantified as coefficient of variation or CV) and vice versa.



Supplementary Figure 28 | Correlation between gene expression noise and copy number is not influenced by the mean expression level of genes. The figure shows data for SC13. To ensure that the correlation between gene expression noise and copy number is not influenced by the mean expression level, genes were binned based their mean expression levels. Within each bin, the copy number of the highest and lowest CV genes were determined. The figure shown here is based on analyzing the top ~20% noisiest and least noisiest genes within each bin. Within each expression bin, high noise genes were associated with a lower copy number and vice versa. Figure 3c shows the data combined over all the expression bins and for various percentages of high and low CV genes.



Supplementary Figure 29 | Agarose gel electrophoresis was used to identify single cells successfully amplified by DR-Seq. A typical agarose gel shows amplified gDNA from six single cells (Lanes 1-6). The gel shows that single cells in lanes 1,2,4 and 5 were successfully amplified. Lanes 8 and 9 represent negative controls where DR-Seq was performed in tubes not containing a single cell. Negative controls were used to control against contamination. Lanes 12 and 13 show DNA ladders with a majority of amplicons having a size distribution between 500-2500 bp in DR-Seq.

Correlation with Bulk mRNA Sequencing	**CEL-Seq (Best 13 cells)**	**CEL-Seq (Random Sampling)**	**DR-Seq**
Pearson r	0.77	0.71 +/- 0.02	0.69
Spearman r	0.79	0.72 +/- 0.02	0.69

Supplementary Table 1 | Gene expression correlations between single cell and bulk sequencing data in E14 cells. In column 1 (CEL-Seq: Best 13 cells), 13 single cells with the highest read count out of 33 cells that were processed by CEL-Seq were chosen. The average expression of genes from these 13 single cells were correlated to bulk mRNA sequencing. In column 2 (CEL-Seq: Random Sampling), 13 single cells that were processed by CEL-Seq were chosen at random and the average expression of genes from those 13 single cells were compared to bulk mRNA sequencing data. In column 3 (DR-Seq), the average expression of genes from 13 single cells were compared to bulk mRNA sequencing. The table shows that the correlation between bulk and average single cell mRNA expression is similar for both CEL-Seq and DR-Seq.

	Bin Size (kb)	Read-based Method	Coverage-based Method
Slope	50	0.520	0.978
	200	0.818	0.931

Supplementary Table 2 | Comparing mean single cell copy numbers to bulk copy numbers for read and coverage-based methods. The slope of the best-fit line for the plot of mean single cell copy numbers against the bulk copy numbers (**Supplementary Fig. 18b**) showed that the coverage-based method gives values closer to the expected value of 1. Thus, the coverage-based method reduced technical noise and amplification biases to give copy numbers in single cells that are in closer agreement to bulk sequencing data.

(a)

	E14			**SK-BR-3**	
	Bulk	**CEL-Seq**	**DR-Seq**	**Bulk**	**DR-Seq**
Total Reads Sequenced	64,498,306	34,561,325	26,464,966	48,254,512	29,792,816
Total Mapped Reads	50,358,766	19,782,105	4,974,543	37,908,793	7,228,515
Average Number of Mapped Reads per Cell	-	599,458	382,657	-	344,215
Percentage of Mapped Reads	78.08 %	57.24 %	18.80 %	78.56 %	24.26 %

(b)

Cell Line	Single Cell ID	Single Cell gDNA Sequencing Method	Average Sequencing Depth	Percentage Genome Coverage
E14	Bulk	-	5.41x	94.18 %
E14	OMSC 2	MALBAC	2.54x	11.42 %
E14	OMSC 3	MALBAC	2.24x	31.90 %
E14	OMSC 14	MALBAC	0.66x	2.83 %
E14	SC 2	DR-Seq	1.83x	5.11 %
E14	SC 5	DR-Seq	2.14x	4.50 %
E14	SC 11	DR-Seq	1.26x	4.49 %
SK-BR-3	Bulk	-	7.31x	90.10 %
SK-BR-3	SC 3	DR-Seq	1.63x	6.10 %
SK-BR-3	SC 10	DR-Seq	0.93x	3.44 %
SK-BR-3	SC 13	DR-Seq	1.48x	3.30 %
SK-BR-3	SC 20	DR-Seq	0.60x	2.55 %
SK-BR-3	SC 25	DR-Seq	1.34x	2.36 %
SK-BR-3	SC 26	DR-Seq	1.25x	2.55 %
SK-BR-3	SC 28	DR-Seq	1.10x	2.22 %

Supplementary Table 3 | Sequencing Statistics. **(a)** Sequencing statistics for bulk mRNA sequencing and single cell mRNA sequencing using CEL-Seq or DR-Seq for the E14 and SK-BR-3 cell lines. **(b)** Sequencing statistics for bulk gDNA sequencing and single cell gDNA sequencing using MALBAC or DR-Seq for the E14 and SK-BR-3 cell lines. DR-Seq shows slightly lower genome coverages than MALBAC for a given average sequencing depth, possibly due to the effective number of quasilinear amplifications cycles in DR-Seq being lower than in MALBAC.

Copy Number	DNA FISH			DR-Seq							Kolmogorov-Smirnov test
	Percentiles			Cell ID							p
	5	**50**	**95**	**3**	**10**	**13**	**20**	**25**	**26**	**28**	
FHIT	0	1	4	3	2	1	3	2	2	2	0.122
HTT	0	2	4	0	2	0	2	0	1	0	0.119
ZMIZ1	2	4	8	9	1	6	2	11	8	8	0.042
CCDC40	8	14	22	16	10	8	10	11	16	24	0.431

Supplementary Table 4 | Comparing cell-to-cell variability in copy numbers obtained from DR-Seq and DNA FISH in SK-BR-3 cells. The spread in copy numbers between 5 and 95 percentiles are shown for DNA FISH, along with the median of the distribution for four genomic loci. The copy numbers obtained for these loci for the seven single cells processed by DR-Seq are also shown. The copy numbers obtained from DR-Seq are not statistically different from the DNA FISH distributions for these four loci ($p > 0.01$, Kolmogorov-Smirnov test).

	Single Cell ID						
	SC3	SC10	SC13	SC20	SC25	SC26	SC28
SNV (Bulk Sequencing)	5919	5785	5850	3275	3041	3116	3736
SNV (Shared with Single Cell)	750	690	686	381	240	247	238

Supplementary Table 5 | Identifying single nucleotide variants (SNV) within coding regions of the genome for SK-BR-3 cells. Reads from the gDNA fraction in DR-Seq were mapped to the coding regions of the genome (which could arise from either the gDNA or cDNA) and all SNVs in both the bulk sequencing data and the single-cell DR-Seq data were identified using the Genome Analysis Toolkit (GATK) HaplotypeCaller. Next, SNVs were filtered using the GATK VariantFiltration tool. To compare each single SK-BR-3 cell to the bulk sequencing data, only those genomic positions that had high coverage (>20X) in that particular single cell were considered. On average about 10% of the SNVs found in bulk sequencing were also identified in the single cells, suggesting that while there might be considerable cell-to-cell variability in SNVs for this cell line, part of this variability might also arise from false negatives in calling SNVs in single cells.

Supplementary Note

Using length-based identifiers to minimize technical noise in single-cell mRNA sequencing data from DR-Seq

During quasilinear amplification, cDNA molecules are randomly primed by adapter Ad-2 which could extend to the end of the cDNA molecule to generate amplicons that have Ad-2 sequence at the 5' end and Ad-1x sequence at the 3' end. All PCR duplicates that are generated from this amplicon (using adapters Ad-2 and Ad-1x) during further quasilinear amplification rounds have the same genomic coordinates. Thus, amplification biases could be minimized by removing such duplicate reads. Further, given the random nature of priming, the unique genomic coordinates of reads (or length-based identifiers) could be used to achieve resolution close to identifying original cDNA molecules before estimating transcript counts (RPM) for each gene.

To demonstrate that such an approach could be used in DR-Seq to remove PCR bias and identify the original number of cDNA molecules, we showed that the original cDNA molecules were primed only once on average during the quasilinear amplification rounds. Next, we demonstrated that the unique number of binding sites available for adapter Ad-2 for each gene in the transcriptome was large enough to ensure that the original number of cDNA molecules were not undercounted due to saturation of binding sites.

To understand if length-based identifiers could be used to accurately quantify the original number of cDNA molecules after reverse transcription, we estimated the number of times cDNA molecules are primed by adapter Ad-2 during quasilinear amplification. Since 7 cycles of quasilinear amplification are performed in DR-Seq, each original cDNA molecule could theoretically be primed 7 times, resulting in potential overestimation or inaccurate estimation of the original number of cDNA molecules. Thus, length-based identifiers would be accurate only if the original cDNA molecules were primed on average once during the 7 cycles of quasilinear amplification. To identify the number of priming events for each original cDNA molecule, we identified genes from the CEL-Seq data that were expressed at the level of zero or one transcript in all single E14 cells (Exact transcript counts for these genes were estimated using a 4-bp random barcode that enabled unique molecule identification, as recently described[1]). Within this set of 589 lowly expressed genes, zero or one length-based identifier for each gene in the single-cell transcriptome data from DR-Seq would imply that the length-based identifiers could be used to tag the original cDNA molecules. Two or more length-based identifiers for such genes would most likely arise from adapter Ad-2 priming the original cDNA molecule multiple times. We observed either no reads or only one length-based identifier for 99.2% of the genes in the DR-Seq dataset (**Supplementary Fig. 2**). This implied that cDNA molecules are primed only once on average by the adapter Ad-2, thereby enabling length-based identifiers to accurately estimate the original number of cDNA molecules.

Next, we estimated the unique number of theoretical binding sites for adapter Ad-2 (or the number of unique length-based identifiers) for all the genes in the transcriptome to ensure that the observed number of length-based identifiers accurately estimate the original number of cDNA molecules without resulting in undercounting of molecules due to saturation of binding sites. Since the 3' end of Adapter Ad-2 has the sequence GGG or TTT[4], the single-cell E14 transcriptome data was used to identify the potential 3-mer binding sites for Ad-2 within coding regions of genes. As expected, adapter Ad-2 preferentially bound sequence complementary to either GGG or TTT in the cDNA (**Supplementary Fig. 3b**). Further, the top 10 most prevalent 3-mer binding sites were found in approximately 90% of all reads (**Supplementary Fig. 3a**). Since adapter Ad-2 displayed preference for certain 3-mer binding sites, we applied a stringent cutoff by considering only the top 10 binding sites as potential theoretical binding sites rather than all positions along the cDNA molecule. Furthermore, since the 3' end of transcripts are sequenced in DR-Seq, only the last 1000 nucleotides were considered in identifying theoretical binding sites for a gene. With these stringent cutoffs, we found that a majority of genes within the mouse transcriptome have between 50 and 250 theoretical binding sites (**Supplementary Fig. 4**). This suggested that for most genes approximately 50-250 cDNA molecules could be uniquely counted before reaching saturation. Recently, it was shown using 4-bp random barcodes (that were used as unique molecule identifiers), that a majority of the genes can be accurately quantified without reaching saturation[1,2]. Thus, this implied that length-based identifiers could potentially be used to accurately count unique cDNA molecules and dramatically reduce amplification related biases.

After establishing that length-based identifiers could be used to estimate the original number of cDNA molecules, we next quantified the extent of saturation of the length-based identifiers. For genes where the observed number of length-based identifiers are similar to the theoretical number of binding sites, expression would be underestimated using this method. For all the 13 single E14 cells amplified by DR-Seq, we estimated the fraction of binding sites that are occupied for all genes (**Supplementary Fig. 5a**). The distribution showed that for a majority of the genes, the number of observed length-based identifiers are much smaller than the theoretical number of binding sites, implying that the number of length-based identifiers accurately estimate the original number of cDNA molecules. The cumulative distribution for each single cell showed that 95% of the genes had less than approximately 15% of the binding sites occupied (**Supplementary Fig. 5b**). Finally, a plot of the fraction of binding sites occupied against the expression level of the gene for two single cells showed that a very small fraction of highly expressed genes are close to saturation (**Supplementary Fig. 5c,d**). Besides this small set of highly expressed genes for which the length-based identifiers potentially underestimate the original number of cDNA molecules, length-based correction provides an accurate estimate of the original number of DNA molecules in a single cell.

Quasilinear amplification does not introduce additional biases in the single-cell transcriptome data in DR-Seq

In addition to showing that the length-based identifiers introduced during the quasilinear amplification step provide resolution close to identifying the original cDNA molecules, we provide more data to demonstrate that this step in DR-Seq does not introduce biases in the single-cell transcriptome data.

Since the quasilinear amplification step in DR-Seq results in PCR amplification of cDNA molecules, we tested if the GC content of cDNA molecules potentially biases transcript counts. Over the entire range of GC content, the median expression of genes in CEL-Seq was compared to DR-Seq (**Supplementary Fig. 12a**). The median expression of genes in DR-Seq was very similar to that in CEL-Seq over all GC content, suggesting that the quasilinear amplification steps in DR-Seq does not introduce additional biases.

Since PCR favors the amplification of shorter template molecules, we tested if there was a bias in DR-Seq towards shorter transcripts. Over the entire range of transcript lengths, the median expression of genes in DR-Seq and CEL-Seq was similar (**Supplementary Fig. 12b**). Only for the shortest transcripts (~100-200 genes), the median expression of genes was higher in both DR-Seq and CEL-Seq compared to other transcript lengths. However, there was no statistically significant difference in gene expression between DR-Seq and CEL-Seq for these short transcripts. Since both methods show increased expression for these genes, this is possibly a consequence of both methods sequencing the 3' end of transcripts rather than a bias introduced during quasilinear amplification. Thus, it appears that the quasilinear amplification step does not bias the quantification of transcripts of different lengths.

Finally, since quasilinear amplification in DR-Seq first depends on adapter Ad-2 randomly priming cDNA molecules, we tested if genes with differences in the number of theoretical binding sites introduces biases in amplification. For the entire range of binding sites, there was no statistically significant difference in the expression of genes between DR-Seq and CEL-Seq (**Supplementary Fig. 12c**).

Reducing technical noise in single-cell gDNA sequencing data from DR-Seq

We developed a technique to reduce amplification biases in single-cell gDNA sequencing data from DR-Seq to enable greater accuracy in calling copy number variations. One of the sources of bias during quasilinear amplification is the non-linear but sub-exponential amplification of gDNA. This implies that while the biases are less severe than exponential PCR biases, technical noise could be further reduced computationally. During the generation of the first amplicons in quasilinear amplification, adapter Ad-2 randomly primes different loci within the genome. These amplicons, with Ad-2 at only one end, do not loop out of the reaction and therefore remain templates in subsequent rounds of the quasilinear amplification. This can introduce biases between different loci within the genome, depending on the cycle in which these regions were first primed by Ad-2. Therefore, pileup of reads in certain regions of the genome can be corrected if the original amplicons that are generated from the gDNA template are accurately estimated, rather than the amplicons that were repeatedly amplified from the quasilinear amplification generated templates. To correct for this amplification bias, we developed a coverage-based model and tested if genome coverage would be a more accurate predictor of the original number of quasilinear amplicons rather than a read-based model. We quantified the bin-to-bin variability in counts for regions with the same copy number, as estimated from bulk gDNA sequencing. For example, when the two methods were applied to one single SK-BR-3 cell, we found that within regions that are expected to have the same copy number (and therefore the same counts) based on bulk sequencing, the bin-to-bin variability was significantly reduced in the new coverage-based method (**Supplementary Fig. 15a**). This suggested that the coverage-based method reduces technical noise and thereby minimizes amplification biases that are generated during the quasilinear amplification steps. Next, systematic comparison of the two methods over all the seven single SK-BR-3 cells showed that across copy numbers, bin-to-bin technical noise is reduced up to two-fold in the coverage-based method over the read-based method (**Supplementary Fig. 15b**). Further, cell-to-cell pairwise correlations between all the single SK-BR-3 cells and single cell versus bulk sequencing correlations improved significantly with the coverage-based method suggesting that this method reduces technical noise and bin-to-bin variability arising from amplification biases (**Supplementary Fig. 16a,b**).

Finally, we quantified if the reduced bin-to-bin technical noise in counts for the coverage-based method also improved copy number calling. Plotting mean copy numbers from the single cell data against the bulk copy numbers showed that the slope of the best-fit regression line was closer to the expected values of 1 for the coverage-based method compared to the read-based method (**Supplementary Fig. 18b** and **Supplementary Table 2**). Thus, the coverage-based method significantly reduces technical noise and amplification biases and results in copy number calls that are in close agreement with the bulk gDNA sequencing data.

Copy number calling from single cell gDNA sequencing data

The first step in calling copy numbers from single-cell gDNA sequencing data was to identify breakpoints using the Circular Binary Segmentation (CBS) algorithm[5]. A p-value of 0.01 was used to detect breakpoints using CBS. Further, the results presented in this study were robust over a range of p-values. The bulk copy numbers were then used to calibrate the median counts for each segment to call copy numbers in single cells. To do this, bulk copy numbers for every bin in the genome was compared to single cell median counts for that bin. For each single cell, a regression line was then fitted to the data to infer the copy number that corresponds to the median read counts. Finally, the estimated copy numbers in single cells were rounded off to the nearest integer. While single cells display variability in copy numbers (**Supplementary Table 4**), we expected differences in copy number between single cells and bulk sequencing to be evenly distributed around the diagonal line. For all the single cells, we founds fits with slopes close to 1. Supplementary Fig. 18a shows this fit for one single cell (SC13).

Error estimation in copy number calling from single cell gDNA sequencing data

We developed a model to estimate confidence intervals for the copy numbers called by our algorithm. First, breakpoints are identified in the single-cell gDNA data using the coverage-based method and the CBS algorithm[5]. Next, data points within each genomic segment are sampled with replacement to estimate the new bootstrapped median values for each segment. These bootstrapped median values are then used to estimate the errors in calling copy number for each genomic segment. **Supplementary Fig. 22a** shows the upper and lower confidence intervals in the copy number estimation for regions within chromosome 8 for one single SK-BR-3 cell. Finally, the mean standard deviations for different copy number loci in the genome were estimated for all the single SK-BR-3 cells (**Supplementary Fig. 22b**). We found that for the first 8 copy numbers, the most abundant copy number states in SK-BR-3 cells, the mean standard deviations in estimating copy numbers approach 1.

Correlation between gene expression noise and copy number is not influenced by the mean expression level of genes

To ensure that the link between copy number and gene expression noise (**Fig. 3c**) is not influenced by the mean level of expression, we compared genes that have similar levels of mean expression but varied noise and identified the copy numbers that such genes (i.e., those with the highest and lowest noise for a given mean expression) were associated with. To do so, we first divided the entire range of mean expressions into bins. Within each bin, we identified the genes with the highest and lowest noise (**Supplementary Fig. 28**). We then found that genes with the highest noise (over all the bins combined and therefore over the entire range of mean expressions) are associated with low copy numbers and vice versa (**Fig. 3c**). Further, these results were robust to the initial number of bins used to divide the range of mean expressions. Thus, this correlation between copy number and gene expression noise is conditioned over a large range of mean expressions and is therefore not influenced by it.

References

1. Grün, D., Kester, L. & van Oudenaarden, A. *Nat. Methods* **11,** 637–640 (2014).
2. Jaitin, D. A. *et al. Science* **343,** 776–779 (2014).
3. Baslan, T. *et al. Nature Protocols* **7,** 1024–1041 (2012).
4. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. *Science* **338,** 1622–1626 (2012).
5. Venkatraman, E. S. & Olshen, A. B. *Bioinformatics* **23,** 657–663 (2007).