Sequencing thousands of single-cell genomes with combinatorial indexing

Sarah A. Vitak*, Kristof A. Torkenczy*, Jimi L. Rosenkrantz, Andrew J. Fields, Lena Christiansen, Melissa H. Wong, Lucia Carbone, Frank J. Steemers, and Andrew Adey

Supplementary Tables

Supplementary Figures

Bulk Library	Duplication Rate	Reads MQ≥10	Reads in HeLa DHS sites	Percent in DHS sites	Fold Enrichment	Estimated Library Size
xSDS	4.50%	83,507,827	2,307,825	2.76%	2.22	798,085,544
LAND	1.86%	64,353,617	2,240,466	3.48%	2.79	1,657,844,868
ATAC	27.24%	60,494,125	8,179,083	13.52%	10.84	170,409,197
SHOT*	NA	60,000,000	1,031,310	1.72%	1.38	NA

Supplementary Table 1 | Bulk library statistics. Information on bulk cell libraries constructed to evaluate nucleosome depletion. *SHOT library is a random sampling of 60M reads obtained from the HeLa dbGaP repository under accession: phs000640.v4.p1 (ref. 19). Library size estimates were generated using Picard tools function “EstimateLibraryComplexity”. For shotgun sequencing, the read used were duplicate removed, and therefore duplication rate and library size estimates were not determined.

a.

					EM Mixed Model		
Library	Sample	Nucleosome Depletion Method	PCR Wells	Nuc. / well	λ (noise, single cell)	μ (noise, single cell)	σ
GM12878.LAND1	Human (GM12878)	LAND (27.6 μM LIS)	96	25	0.872007,0.127993	1.080594,3.841161	0.66137
GM12878.LAND2	Human (GM12878)	LAND (13.8 μM LIS)	96	25	0.419749,0.580251	0.291373,1.982663	0.513341
GM12878.NSTLAND	Human (GM12878)	LAND (13.8 μM LIS + 200 mM NaCl)	96	22	0.752279,0.247721	1.177030,3.937951	0.736942
GM12878.xLAND	Human (GM12878)	x-link + LAND (13.8 μM LIS)	96	22	0.803801,0.196199	0.814446,3.409897	0.578019
GM12878.LAND3	Human (GM12878)	LAND (13.8 μM LIS)	96	22	0.842110,0.157890	1.307204,4.047124	0.680607
GM12878.LAND4	Human (GM12878)	LAND (4.6 μM LIS)	96	22	0.861427,0.138573	1.184529,3.689950	0.619864
GM12878.arrLAND	Human (GM12878)	Arrested, LAND (4.6 μM LIS)	96	22	0.970847,0.0291532	1.280424,4.3764043	0.526405
HeLa.LAND1	Human (HeLa S3)	LAND (4.6 μM LIS)	96	22	0.884456,0.115544	1.489698,4.622590	0.740663
HeLa.LAND2	Human (HeLa S3)	LAND (4.6 μM LIS)	96	22	0.849262,0.150738	0.816437,3.448150	0.496199
HeLa.LAND3	Human (HeLa S3)	LAND (4.6 μM LIS)	96	22	0.838170,0.161830	1.476571,4.135318	0.539156
HumMus.LAND1	Human (HeLa S3), Mouse (3T3)	LAND (27.6 μM LIS)	96	25	0.816623,0.183377	0.826636,2.662703	0.559918
HumMus.LAND2	Human (HeLa S3), Mouse (3T3)	LAND (113.8 μM LIS)	96	25	0.784437,0.215563	1.223024,3.960925	0.716764
HumMus.LAND3	Human (GM12878), Mouse (3T3)	LAND (4.6 μM LIS)	96	22	0.863399,0.136601	1.473206,4.590961	0.627049
HumMus.LAND4	Human (GM12878), Mouse (3T3)	LAND (4.6 μM LIS)	96	22	0.973846,0.0261538	1.448882,5.0360715	0.712699
RhesusInd1.LAND	Rhesus Individual 1 (frozen)	LAND (4.6 μM LIS)	16	22	0.823777,0.176223	1.774362,4.301835	1.09558
GM12878.xSDS	Human (GM12878)	xSDS	288	22	0.871926,0.128074	1.781897,4.291739	0.764169
HumMus.xSDS	Human (HeLa S3), Mouse (3T3)	xSDS	96	22	0.868349,0.131651	1.776006,4.209856	0.878084
CRC.xSDS	Stage 2 Colorectal Cancer (frozen)	xSDS	16	22	0.911423,0.0885767	1.423343,4.7335258	0.83885
PDAC.xSDS	Stage 3 Pancreatic Ductal Adenocarcinoma (fresh)	xSDS	288	22	0.915855,0.0841453	1.713041,4.4984682	0.872799
RhesusInd1.xSDS	Rhesus Individual 1 (frozen)	xSDS	96	22	0.953348,0.0466516	1.122175,4.4798411	0.788582
RhesusInd2.xSDS	Rhesus Individual 2 (frozen)	xSDS	96	22	0.931090,0.0689105	1.091425,4.5055530	0.763775

b.

Library	Sequenced Reads					
	Single Cell Read Cutoff	Single Cell Libraries	Median Unique MQ≥10 Reads	Mean Unique MQ≥10 Reads	Median Complexity	Cells ≥ 5e4 Reads
GM12878.LAND1	1,512	621	11,721	37,055	45.96	129
GM12878.LAND2	1,000	113	2,091	3,434	90.79	0
GM12878.NSTLAND	1,588	1,060	13,734	52,244	71.33	313
GM12878.xLAND	1,000	1,212	6,384	14,148	58.80	72
GM12878.LAND3	2,325	1,015	16,673	84,010	34.42	232
GM12878.LAND4	1,529	616	7,151	32,614	87.55	68
GM12878.arrLAND	7,079	119	33,923	94,036	36.51	54
HeLa.LAND1	7,619	573	67,077	100,016	91.01	338
HeLa.LAND2	1,000	648	4,756	18,026	37.45	29
HeLa.LAND3	3,946	1,140	18,695	25,501	97.73	120
HumMus.LAND1	1,000	263	2,699	6,174	2.90	4
HumMus.LAND2	1,754	1,346	13,876	51,952	71.31	388
HumMus.LAND3	9,202	645	61,408	74,329	96.69	378
HumMus.LAND4	21,055	115	119,428	359,175	95.84	99
RhesusInd1.LAND	5,947	340	141,449	165,453	88.21	248
GM12878.xSDS	6,051	3,123	29,550	64,986	53.08	1,056
HumMus.xSDS	5,970	1,331	44,699	64,659	87.89	605
CRC.xSDS	7,846	151	72,753	110,823	89.70	111
PDAC.xSDS	5,164	1,715	49,272	86,592	68.60	846
RhesusInd1.xSDS	4,912	171	55,142	120,769	24.36	92
RhesusInd2.xSDS	5,517	381	62,731	122,602	23.76	213
		16,698				**5,395**

Supplementary Table 2 | SCI-seq library summary. Information on library construction and statistics on the actual depth obtained for each SCI-seq library preparation. (a) Details of library construction and the mixed model used to determine the read count threshold for single cell libraries. (b) Details on libraries for the actual sequence depth obtained in this study.

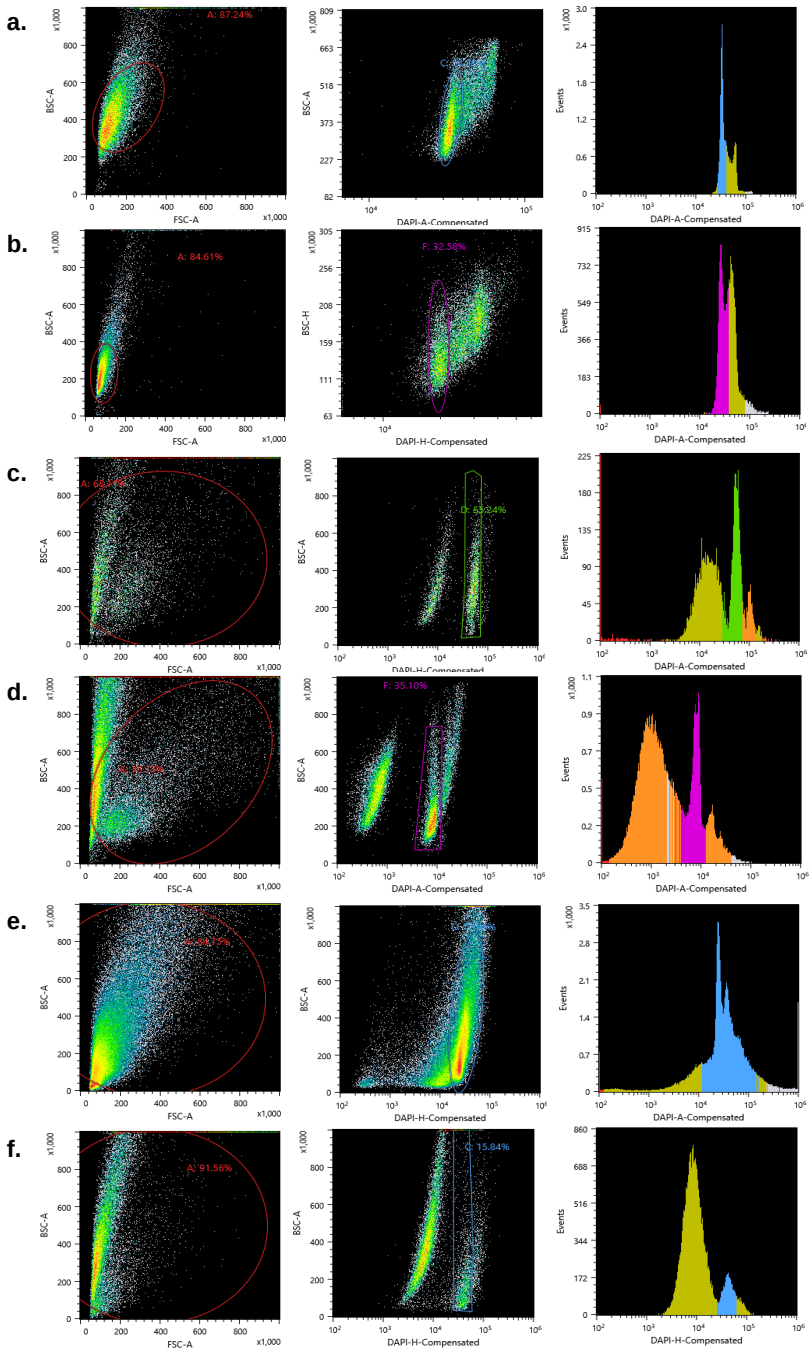
a.

Library	Library Projections								
	Projected to Median of 50% Complexity			Projected to Median of 25% Complexity			Projected to Median of 10% Complexity		
	Median	Mean	Raw Reads	Median	Mean	Raw Reads	Median	Mean	Raw Reads
GM12878.LAND1	378,305	1,176,230	1,640,000,000	554,638	1,734,782	4,850,000,000	653,064	2,115,189	14,120,000,000
GM12878.NSTLAND	44,608	155,477	350,000,000	68,418	228,101	1,030,000,000	83,359	269,951	2,980,000,000
GM12878.LAND3	218,132	1,318,212	2,430,000,000	323,878	2,133,376	7,220,000,000	399,607	2,718,203	21,050,000,000
GM12878.LAND4	135,114	746,530	810,000,000	200,422	1,239,687	2,390,000,000	246,204	2,082,902	6,940,000,000
GM12878.arrLAND	1,490,817	3,684,539	910,000,000	2,141,086	5,687,397	2,710,000,000	2,792,344	7,024,499	7,900,000,000
HeLa.LAND1	3,997,311	5,642,469	7,180,000,000	6,140,962	8,861,587	21,380,000,000	7,399,793	11,190,892	62,310,000,000
HeLa.LAND3	736,813	901,740	2,350,000,000	1,107,204	1,415,747	6,970,000,000	1,337,941	1,806,880	20,320,000,000
HumMus.LAND1	35,991	79,857	50,000,000	56,355	126,673	150,000,000	70,808	161,428	420,000,000
HumMus.LAND2	44,393	154,148	440,000,000	67,957	226,139	1,290,000,000	82,696	277,161	3,760,000,000
HumMus.LAND3	2,257,543	2,638,358	3,890,000,000	3,453,346	3,957,131	11,600,000,000	4,186,331	4,806,388	33,740,000,000
HumMus.LAND4	4,305,319	11,479,621	3,260,000,000	6,126,707	16,880,151	9,710,000,000	7,474,417	20,732,681	28,290,000,000
RhesusInd1.LAND	454,681	514,445	530,000,000	685,354	756,326	1,570,000,000	823,686	902,566	4,570,000,000
GM12878.xSDS	26,791	63,223	580,000,000	39,666	94,153	1,710,000,000	48,089	113,352	4,980,000,000
CRC.xSDS	352,168	530,978	190,000,000	532,772	790,798	560,000,000	641,639	946,770	1,620,000,000
PDAC.xSDS	71,378	129,304	590,000,000	107,615	191,011	1,750,000,000	130,444	228,852	5,110,000,000

b.

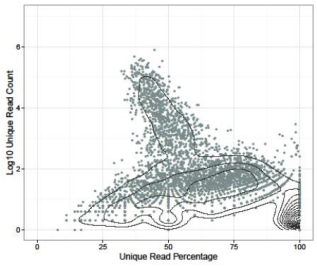
Library	Number of Cells That Can Reach N Reads from Projections						
	5.00E+04	1.00E+05	1.50E+05	2.50E+05	5.00E+05	7.50E+05	1.00E+06
GM12878.LAND1	619	604	579	504	373	308	268
GM12878.NSTLAND	662	504	439	340	183	112	78
GM12878.LAND3	990	886	810	674	470	370	310
GM12878.LAND4	574	474	403	319	211	167	137
GM12878.arrLAND	119	119	118	117	115	107	102
HeLa.LAND1	573	573	573	572	557	547	541
HeLa.LAND3	1,140	1,138	1,129	1,115	1,057	941	812
HumMus.LAND1	167	113	76	40	19	11	6
HumMus.LAND2	851	636	550	421	228	137	100
HumMus.LAND3	645	645	645	641	634	610	593
HumMus.LAND4	115	115	115	115	115	115	115
RhesusInd1.LAND	328	299	277	260	219	186	148
GM12878.xSDS	1,804	1,094	769	468	183	69	22
CRC.xSDS	151	147	144	137	107	70	43
PDAC.xSDS	1,356	1,080	874	601	242	98	54

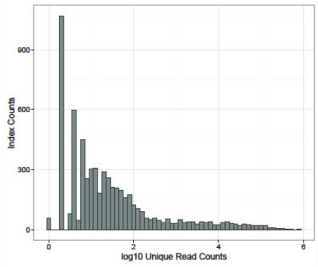
Supplementary Table 3 | SCI-seq library projection statistics. Information on projected statistics of each SCI-seq library if increased sequencing depth were obtained. Projections use the model described in the methods section. Libraries that either failed (GM12878.LAND2 and HeLa.LAND2), or were sequenced to saturation for which the projections do not apply (Rhesus.Ind1.xSDS and Rhesus.Ind2.xSDS) are not included. (**a**) Projections out to a given median complexity including the raw read count to reach that point. (**b**) The number of single cells meeting various read count thresholds are listed if libraries were sequenced to saturation (median complexity of 5%).



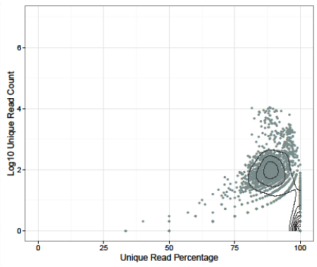
Supplementary Figure 1 | Fluorescence Activated Nuclei Sorting (FANS). Representative plots from FANS sorting of single nuclei. All plots are from sorting the second (PCR) plate unless noted otherwise. (**a**) ATAC-seq Nuclei (**b**) LAND (**c**) HeLa S3 and 3T3 (**d**) xSDS (**e**) PDAC Sort 1 Transposase Plate (**f**) PDAC Sort 2 PCR plate.

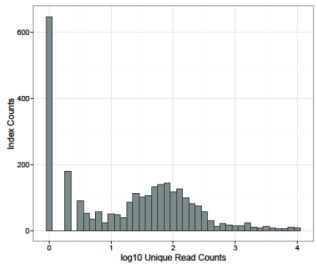
GM12878.LAND1



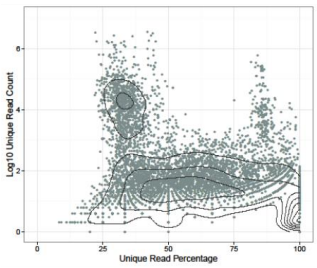


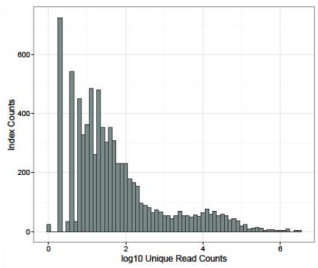
GM12878.LAND2 (failed library)



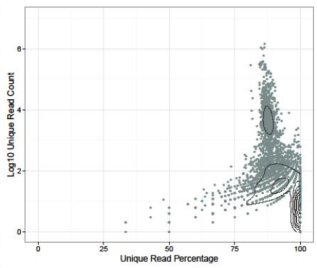


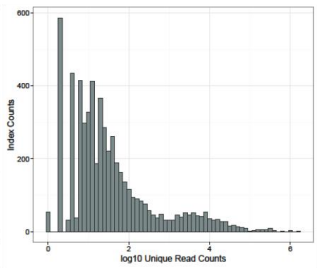
GM12878.LAND3



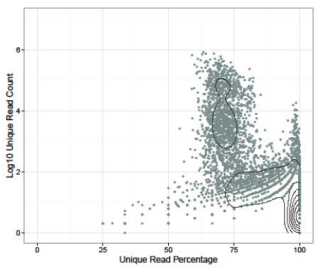


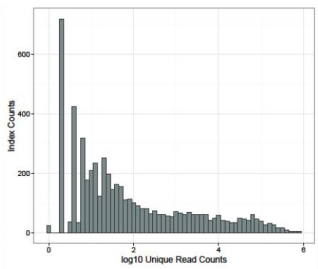
GM12878.LAND4



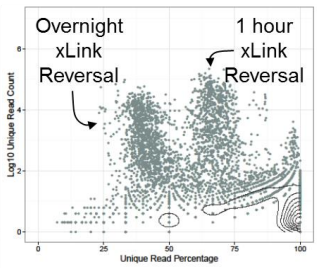


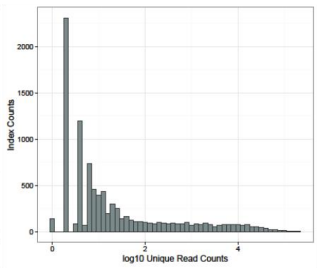
GM12878.NSTLAND



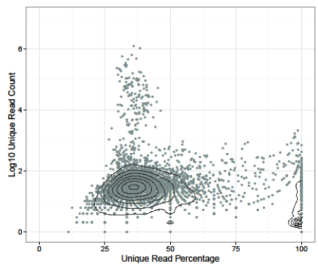


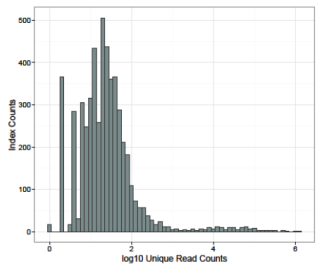
GM12878.xLAND



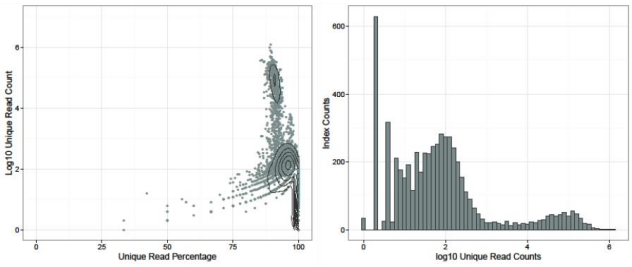


GM12878.arrestLAND

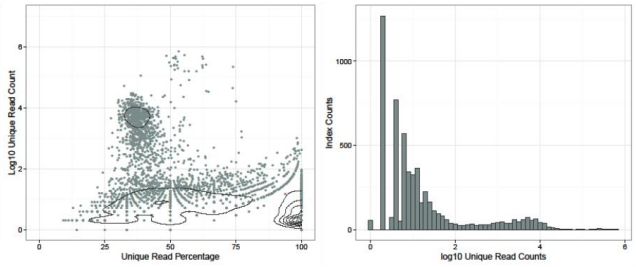




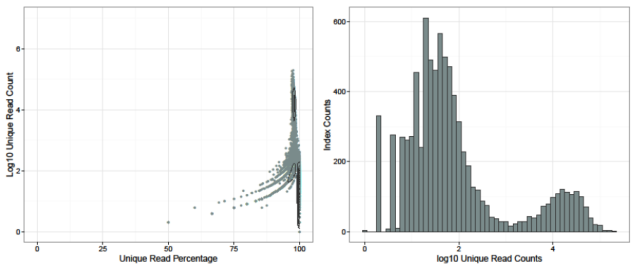
HeLa.LAND1



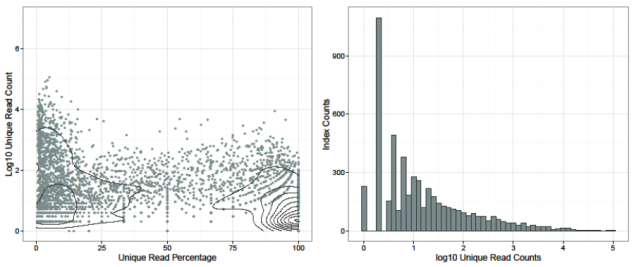
HeLa.LAND2



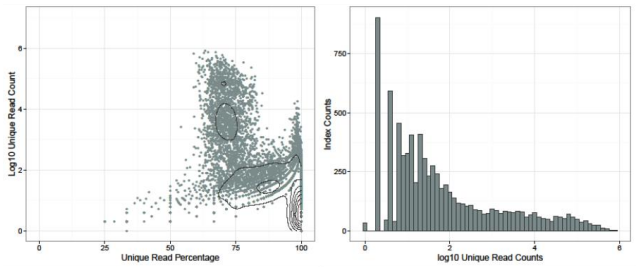
HeLa.LAND3



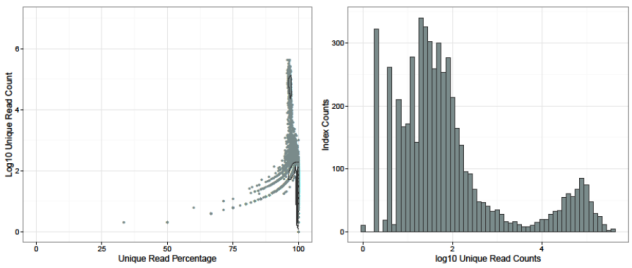
HumMus.LAND1 (failed library)



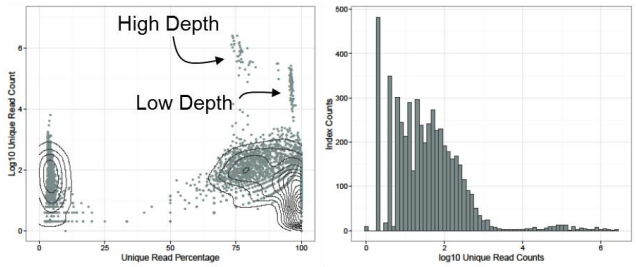
HumMus.LAND2

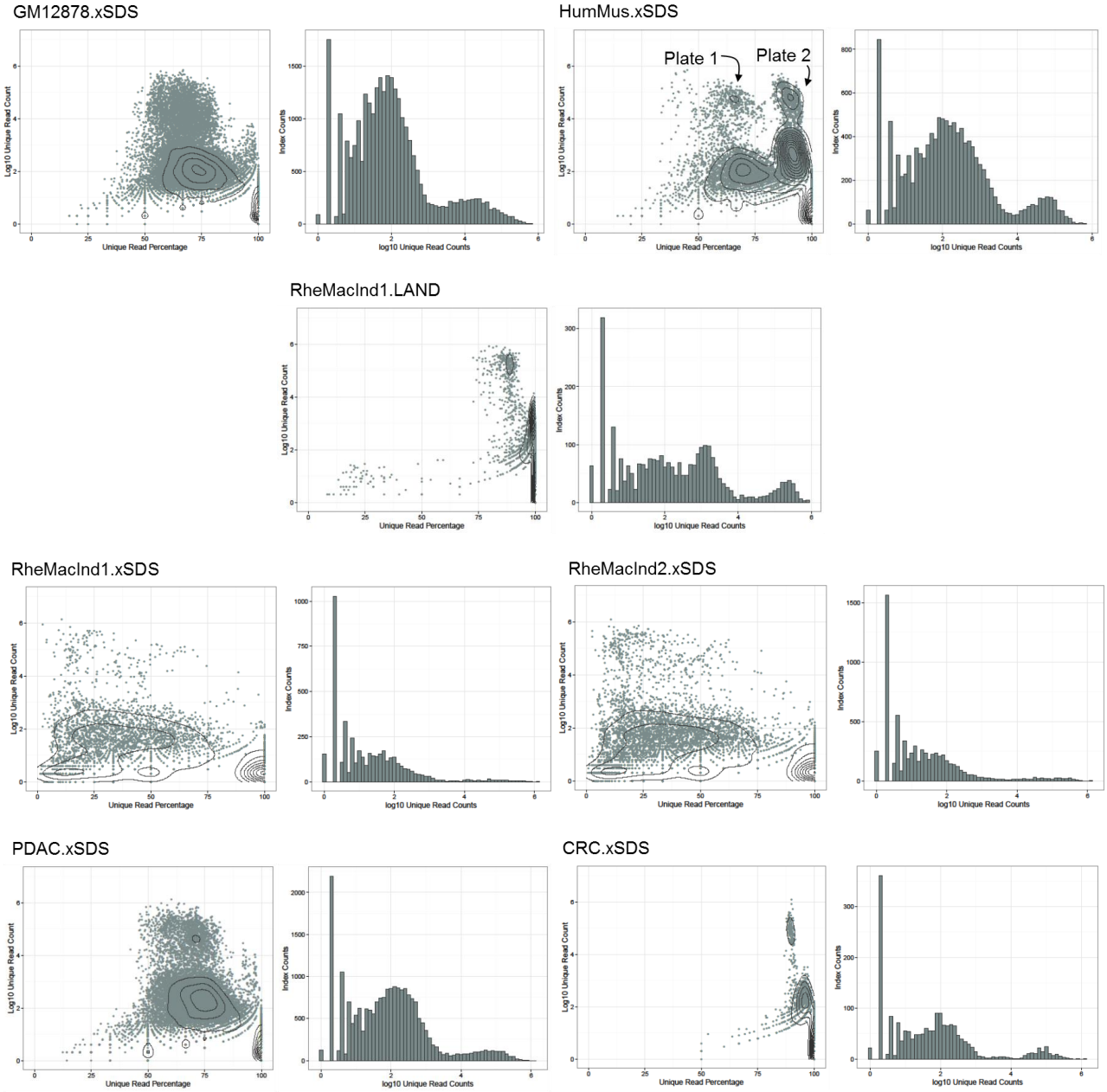


HumMus.LAND3

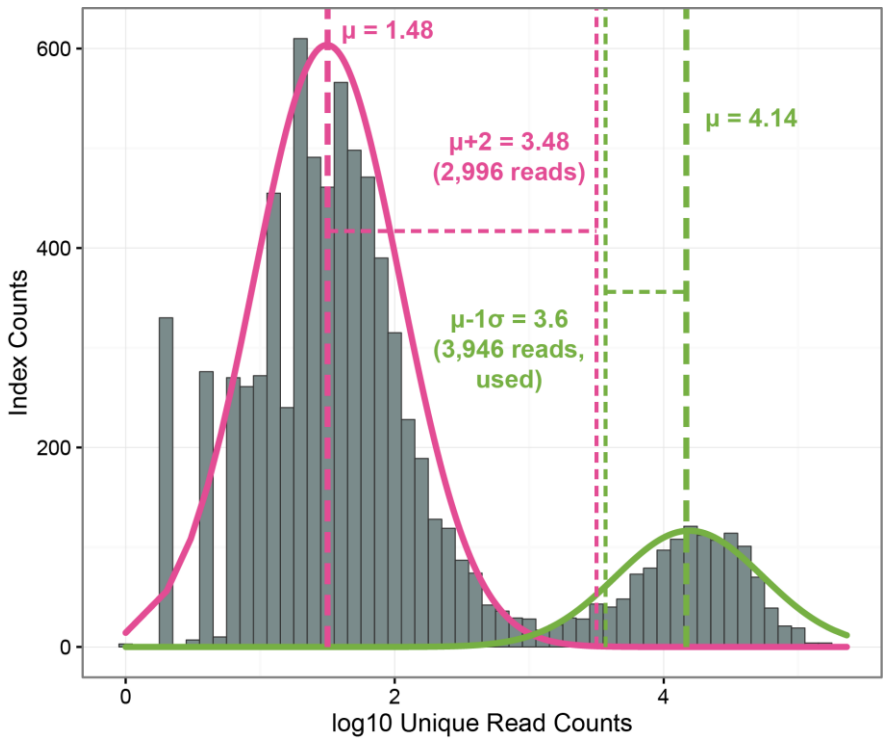


HumMus.LAND4

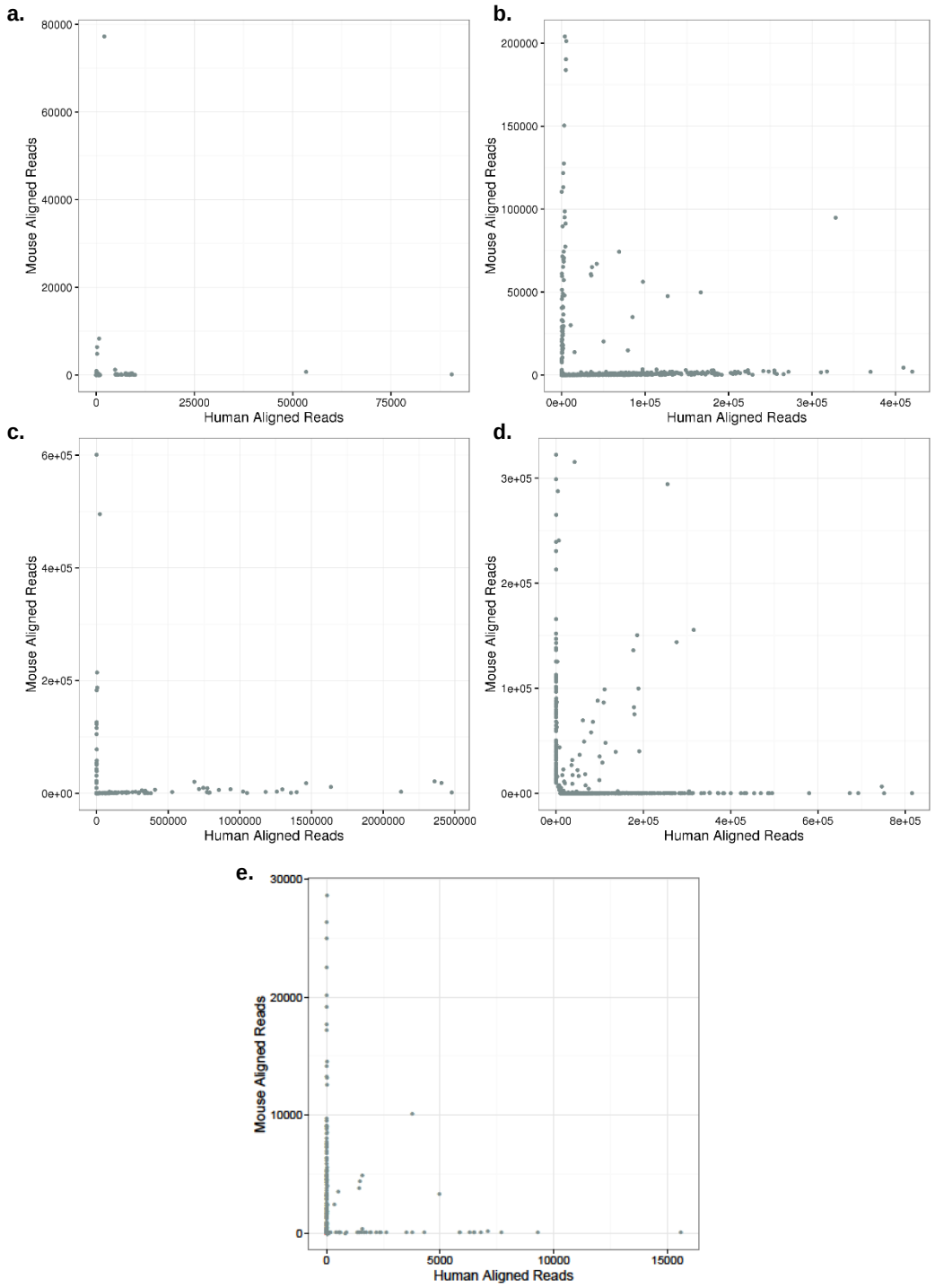




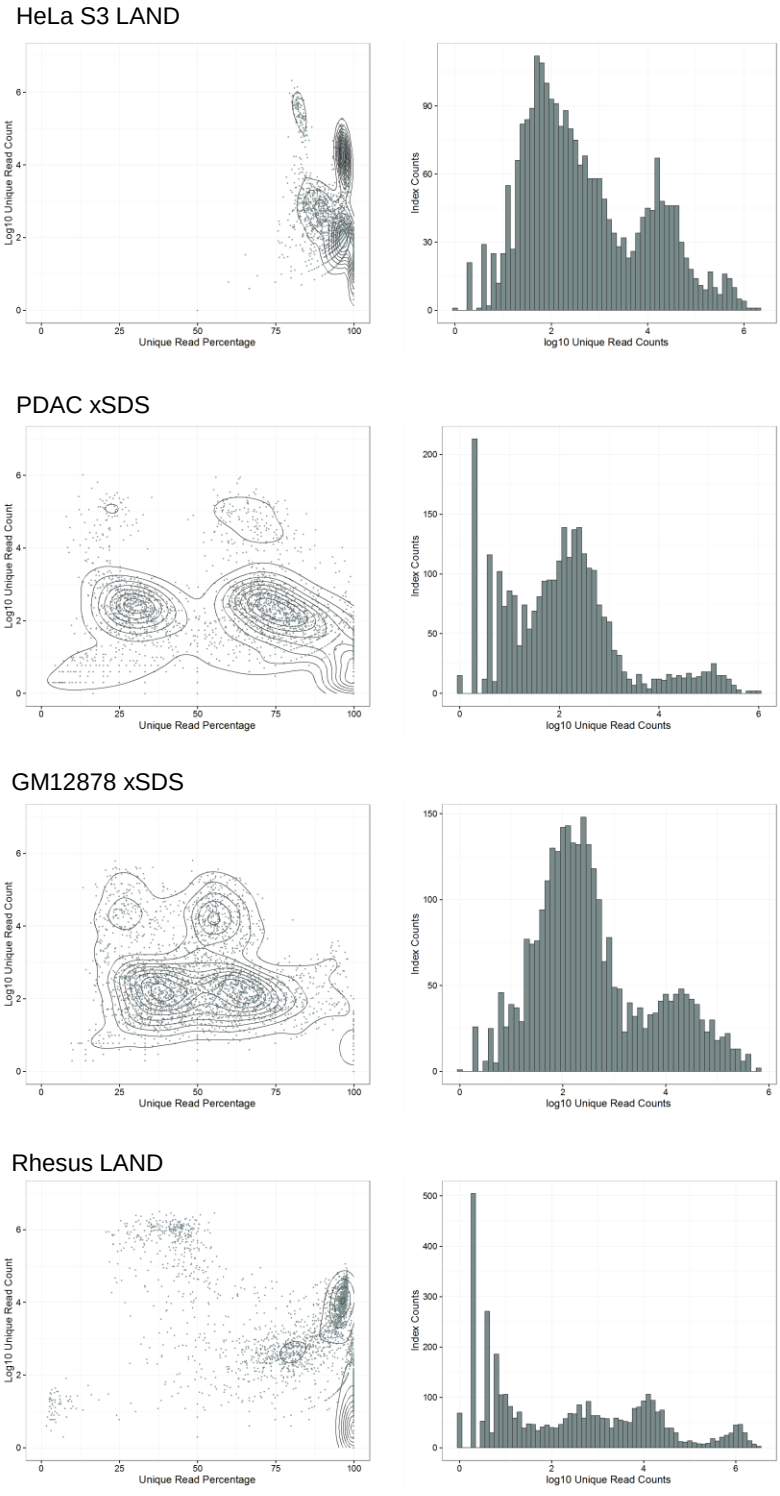
Supplementary Figure 2 | SCI-seq library complexity and index read count distributions for all preparations. For each preparation two plots are shown. Left: each point represents a unique index combination, x-axis is the fraction of unique reads assigned to that index combination, y-axis is the log10 unique read count for the index combination. Contour lines represent point density. Right: A histogram of the log10 unique read counts for each of the index combinations. We expect the majority of potential index combinations not to represent a single cell library and therefore containing very few unique reads (leftmost distribution), with the single cell libraries having far greater read counts (right distribution, or tail in lower performance libraries). Since the plot is on a log10 scale, the noise distribution actually only takes up a minority of the total read counts.



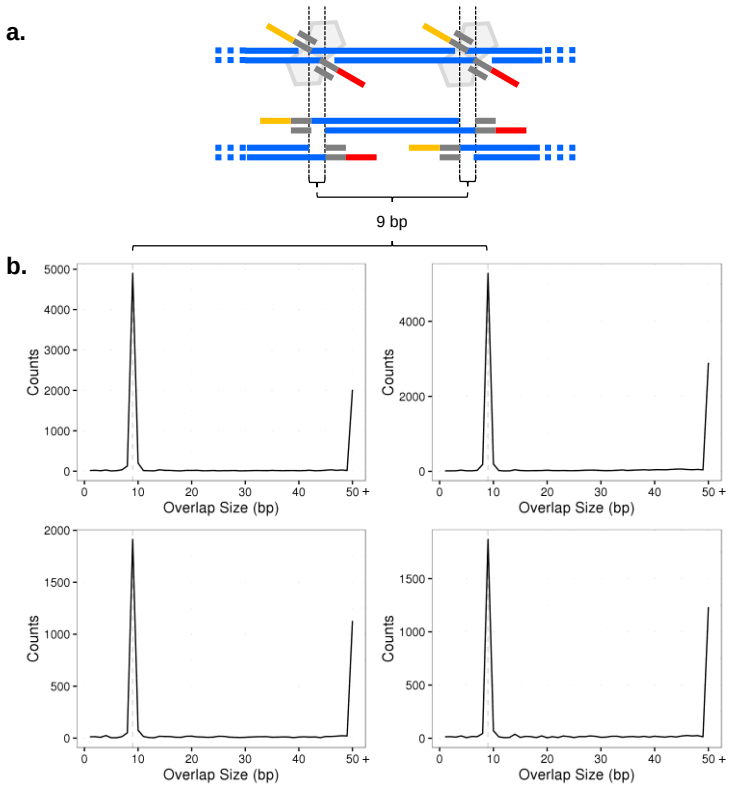
Supplementary Figure 3 | SCI-seq single cell determination using a mixed model. HeLa.LAND3 shown. normalmixEM of the R package *mixtools* is used to identify each distribution: noise index combinations (pink) and single cell libraries (green). The read count threshold to consider an index combination as a single cell library is the greater of either one standard deviation (in log10 space) below the mean of the single cell distribution, or 2 greater (in log 10 space, thus 100 fold greater) than the mean of the noise distribution and at a minimum of 1,000. For the library shown, one standard deviation below the mean of the single cell component is greater and therefore used as the read count threshold.



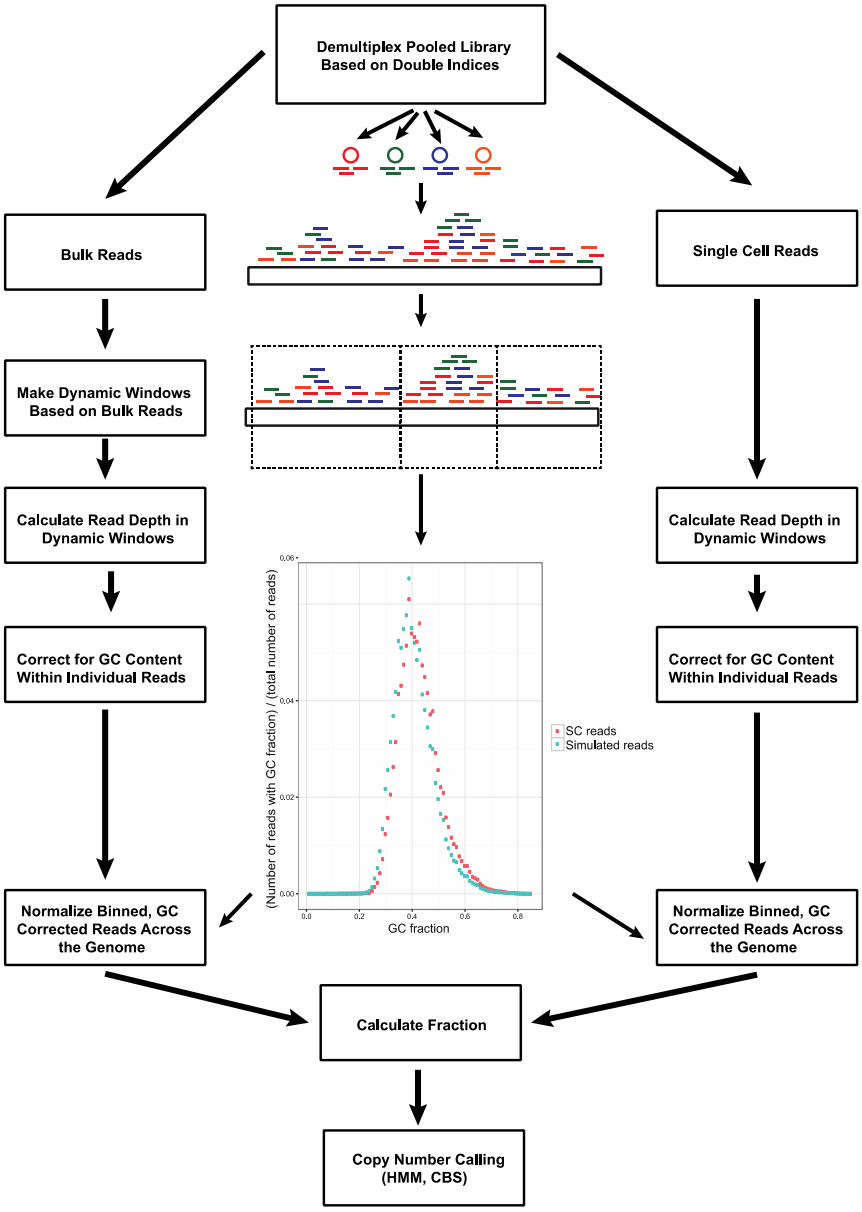
Supplementary Figure 4 | SCI-seq on a mix of human and mouse cells. For all panels the number of reads for each index component are plotted based on the count aligning to the human reference genome, or the mouse reference genome. (**a**,**b**) LAND nucleosome depletion on Human (GM12878) and Mouse (3T3), (**c**,**d**) LAND nucleosome depletion on Human (HeLa S3) and Mouse (3T3), (**e**) xSDS nucleosome depletion on Human (HeLa S3) and Mouse (3T3).



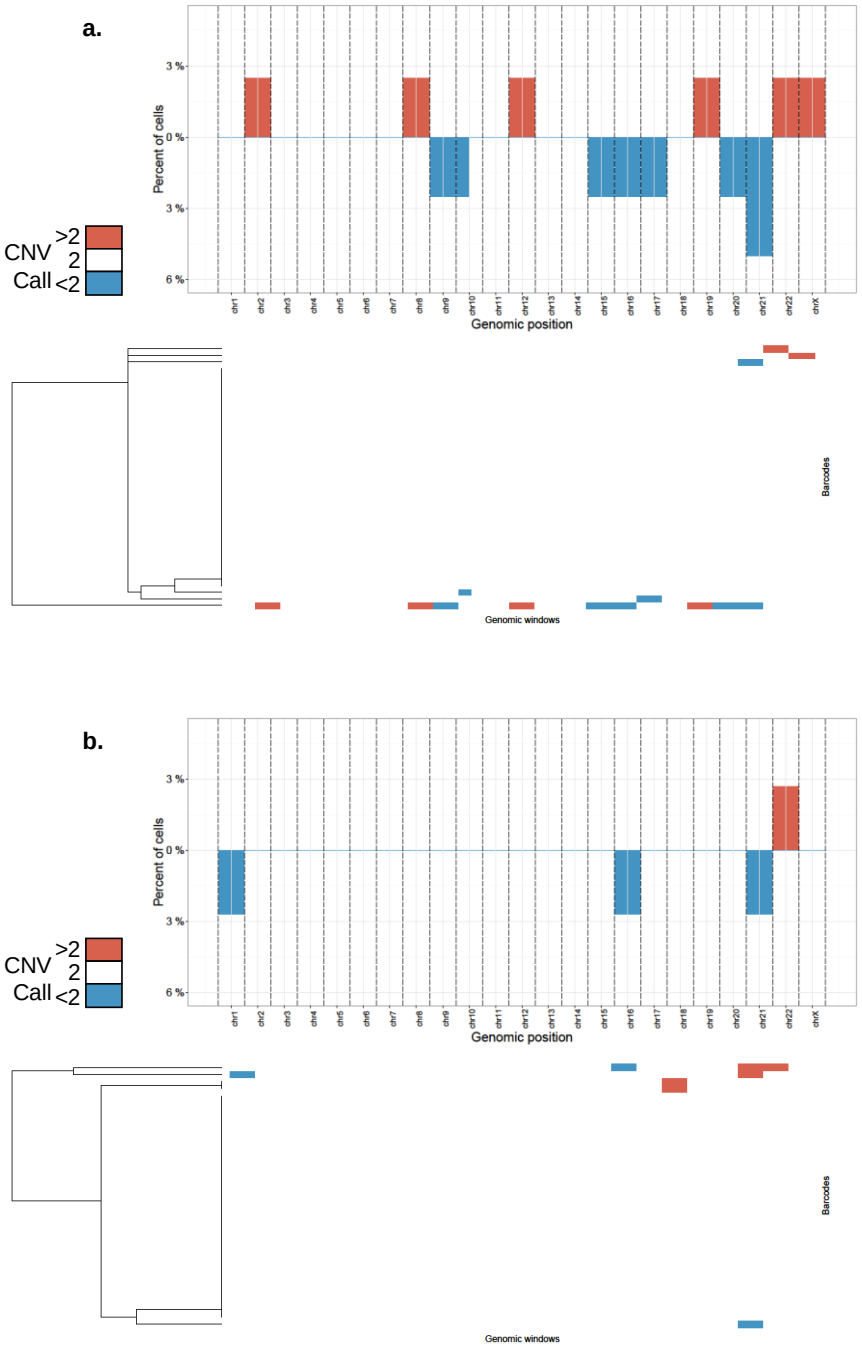
Supplementary Figure 5 | SCI-seq library complexity and index read count distributions after deeper sequencing. For each preparation two plots are shown as in S2 the left plot shows fraction of unique reads versus unique read count for each index combination. While the right plot shows a histogram of read counts for each index combination. Cells from wells sequenced more deeply are shown along with the rest of the plate that those wells belong to. The population of cells with lower complexity (more to the left) is the population that has been sequenced more deeply.



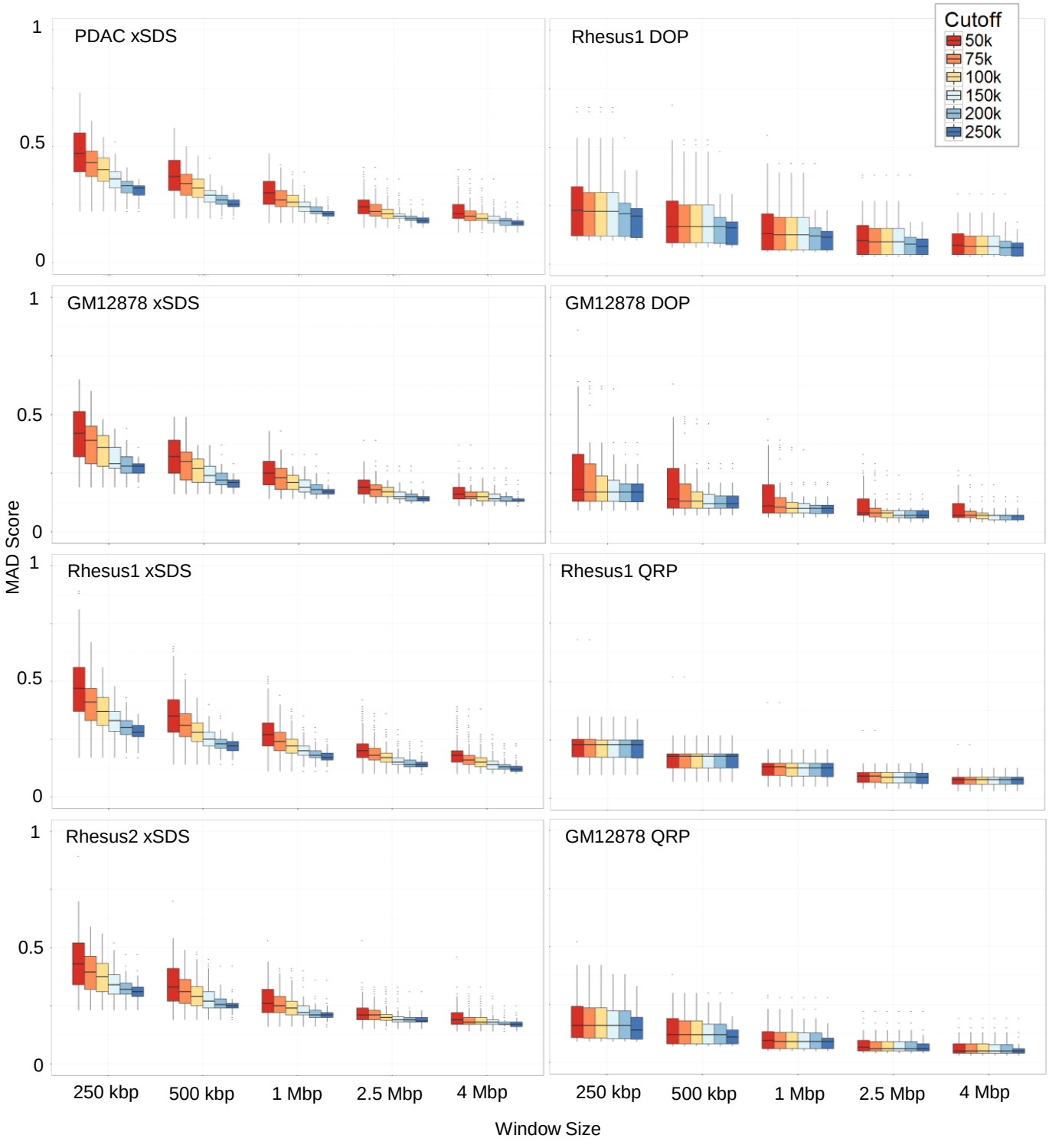
Supplementary Figure 6 | 9bp read overlaps observed from sequencing adjacent transposition events in the same single cell. (**a**) Diagram of how the 9bp copying occurs from the transposition event. (**b**) Representative single cells showing the size of all amplicon overlaps with a dashed line at 9bp.

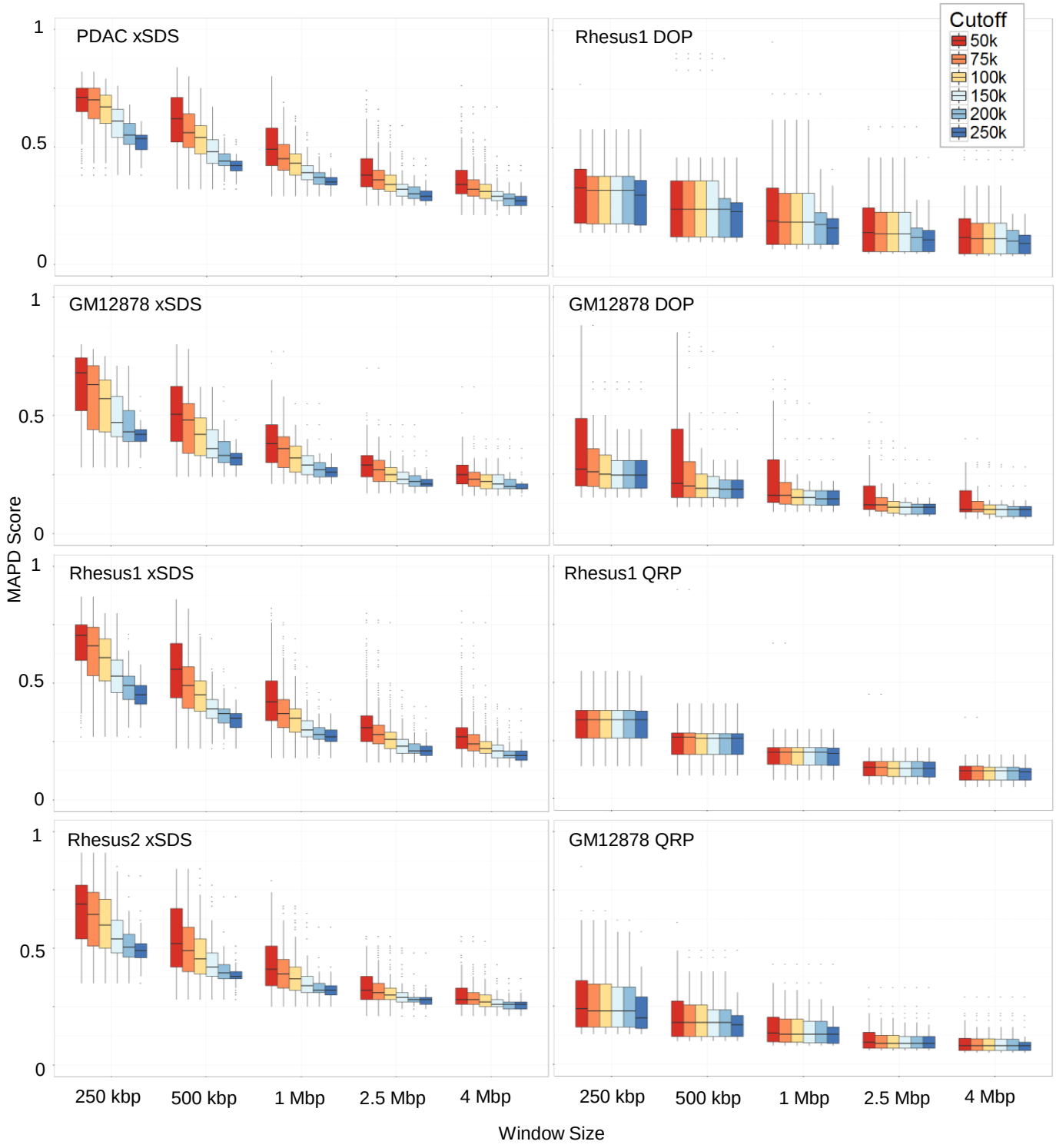


Supplementary Figure 7 | Copy number calling computational workflow for HMM and CBS. After calling, call sets for CBS and HMM were intersected together with Ginkgo and only calls present in all three sets were retained as the final call set.



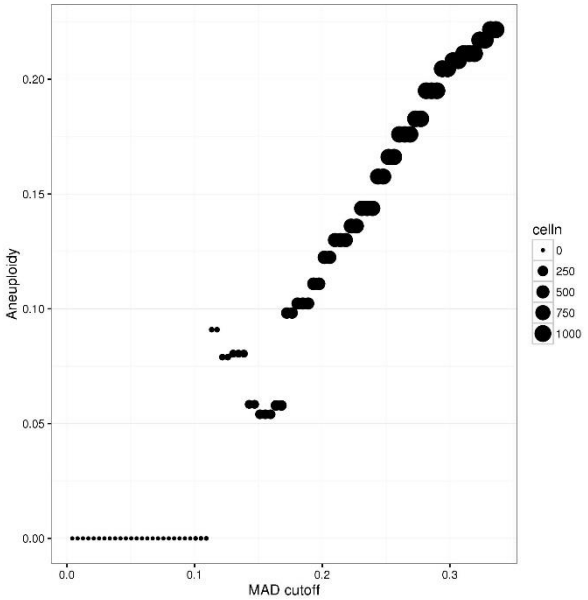
Supplementary Figure 8 | CNV assessment using standard methods of single cell sequencing on GM12878. Top: Summary of chromosome arm amplifications and deletions, Bottom: hierarchical clustering of cells.

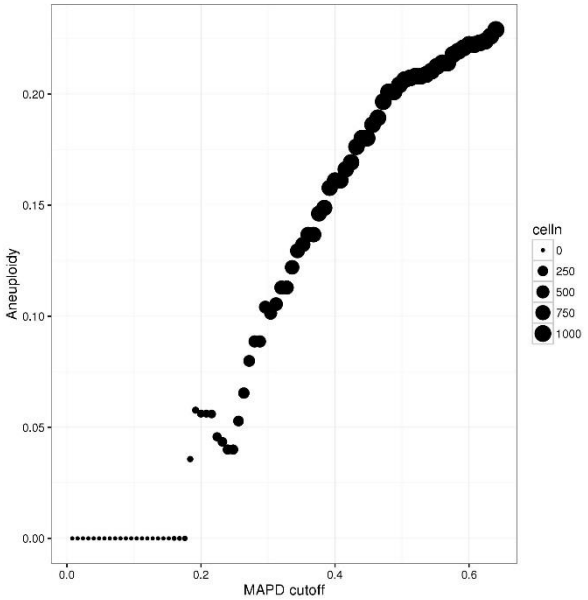

PDAC xSDS
Rhesus1 DOP
GM12878 xSDS
GM12878 DOP
Rhesus1 xSDS
Rhesus1 QRP
Rhesus2 xSDS
GM12878 QRP
Cutoff
50k
75k
100k
150k
200k
250k
MAD Score
1
0.5
0
250 kbp
500 kbp
1 Mbp
2.5 Mbp
4 Mbp
Window Size



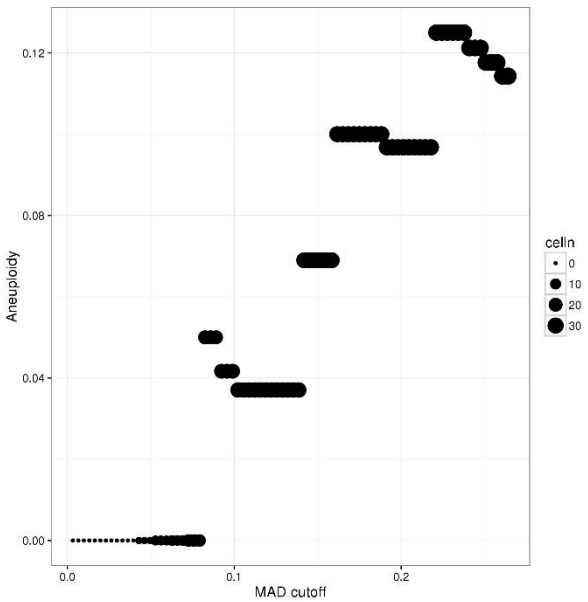
Supplementary Figure 9 | Variance by window size and read count cutoff across all methods. Plots showing the change in MAD or MAPD score as a function of window size and read counts per cell.

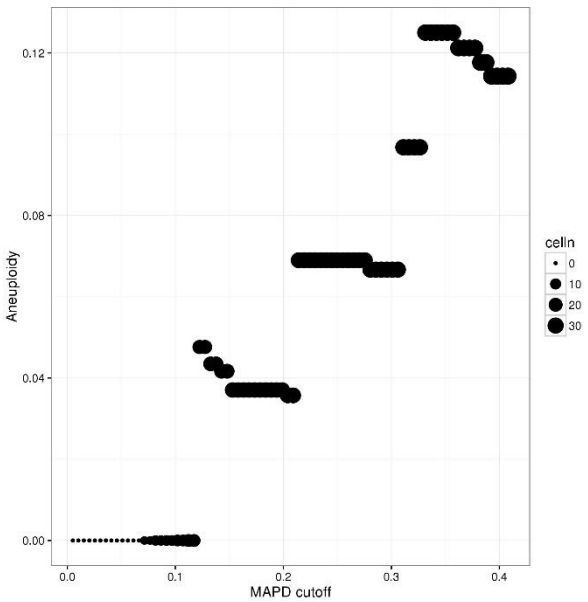
GM12878 xSDS



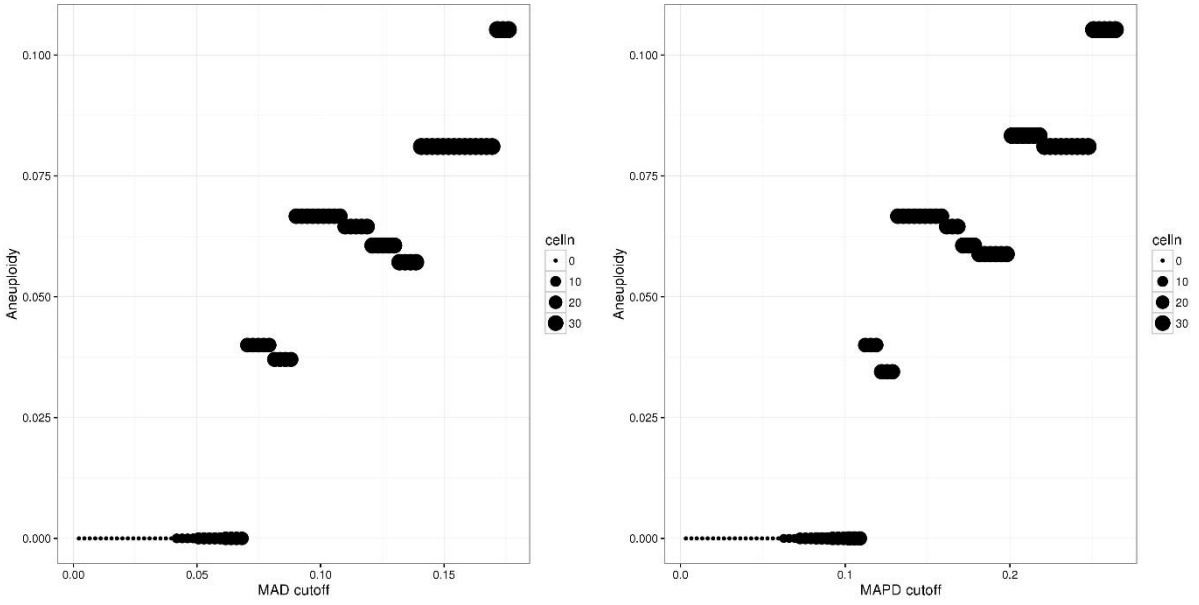


GM12878 DOP

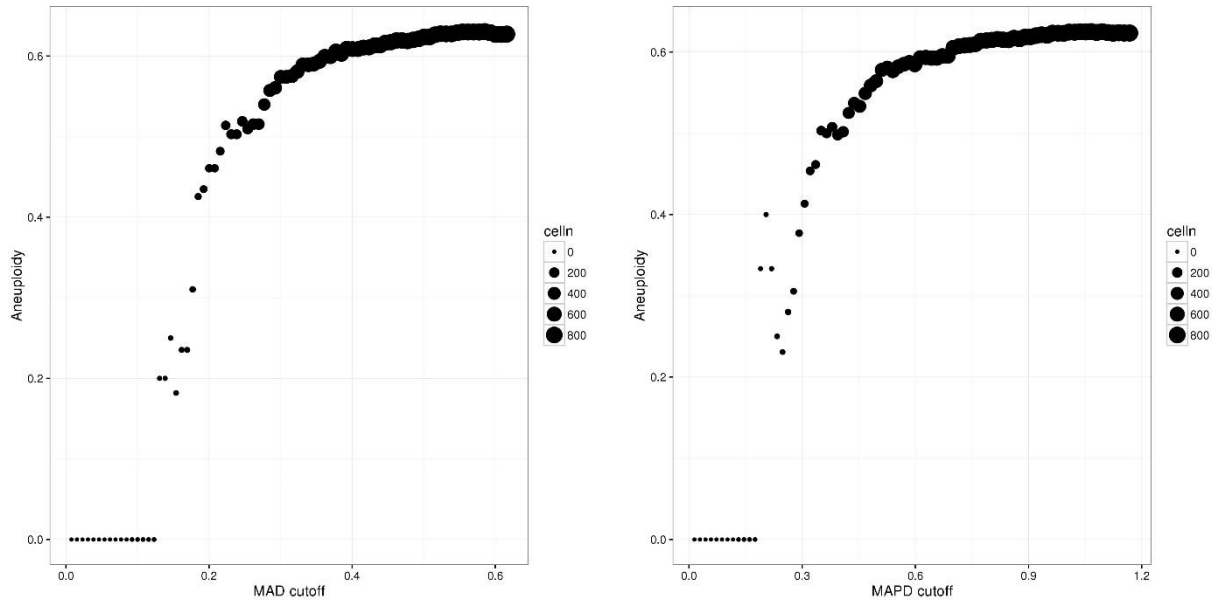




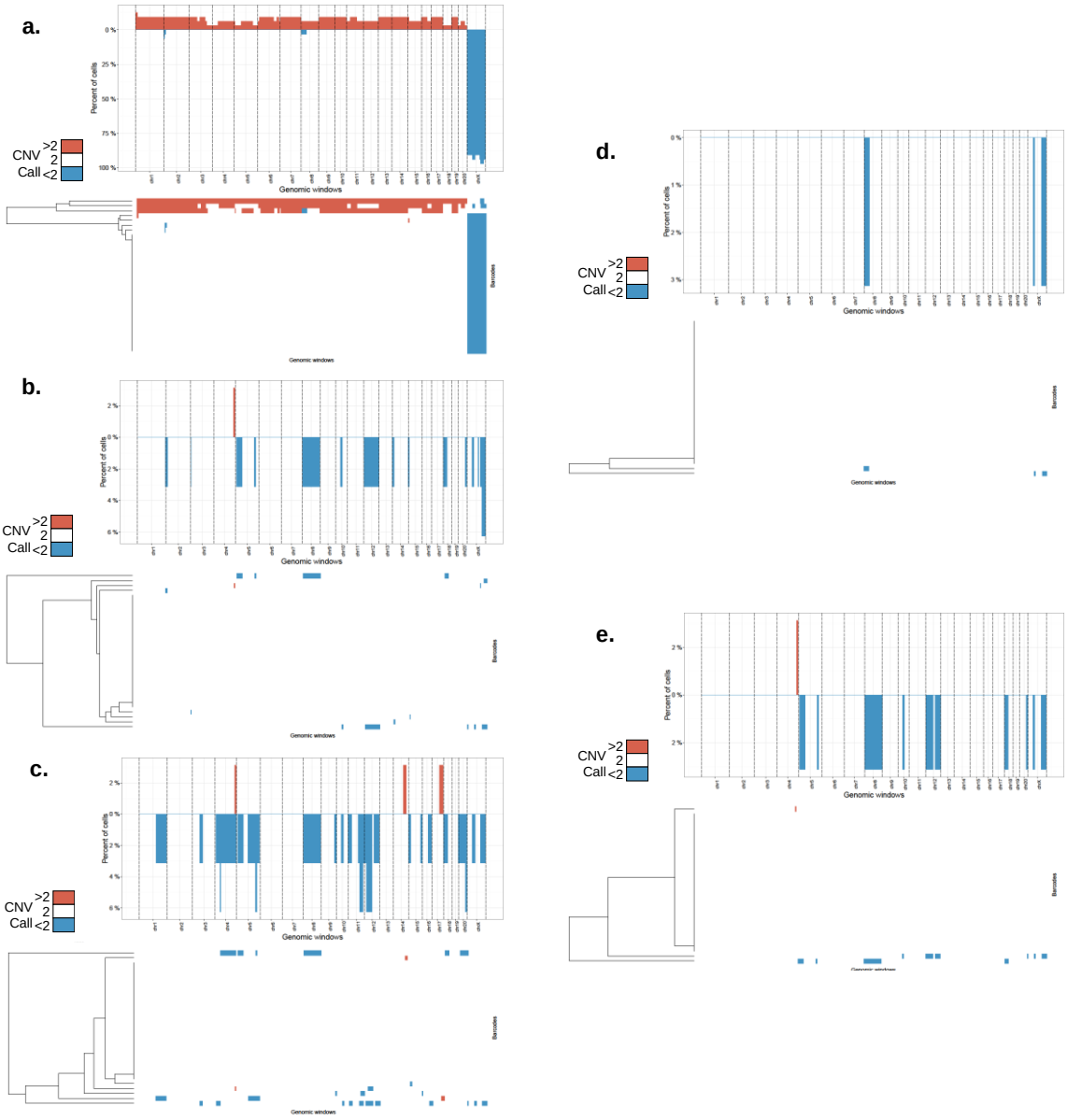
GM12878 QRP



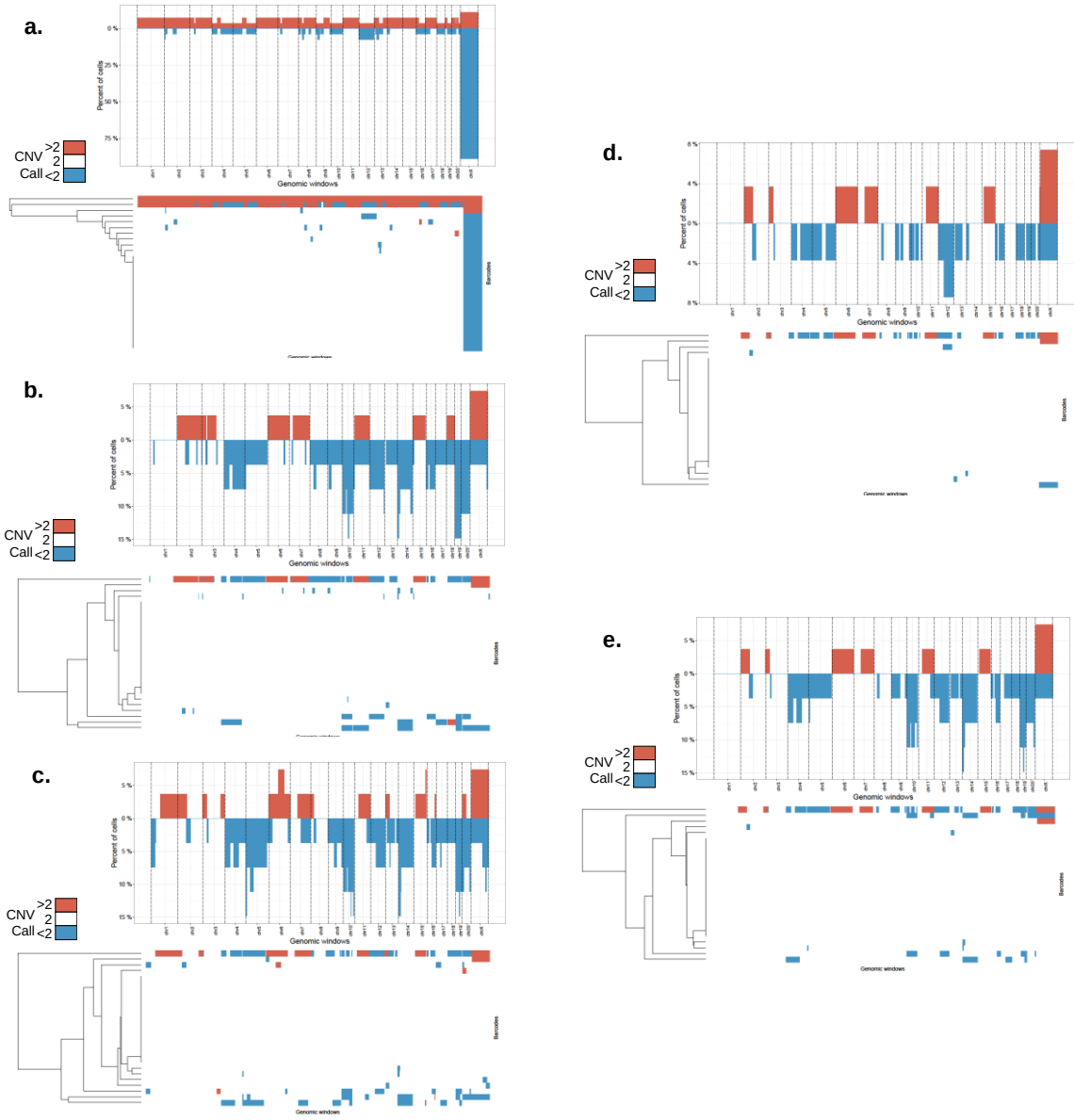
GM12878 LIS



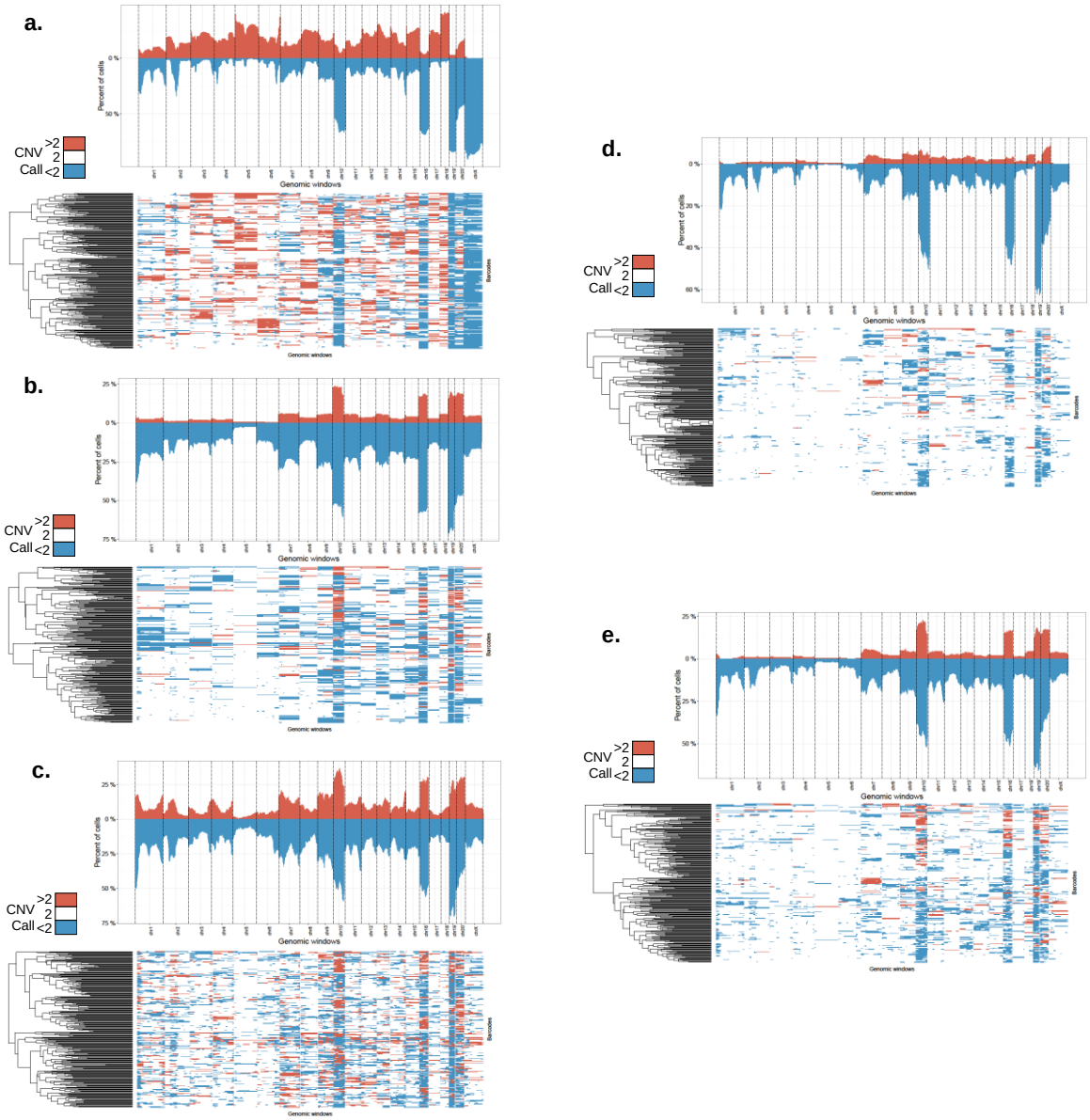
Supplementary Figure 10 | GM12878 aneuploidy rates across variance score cutoffs. Each point is the aneuploidy rate for the population of cells (y-axis), scaled by the number of cells included at a given score cutoff (x-axis).



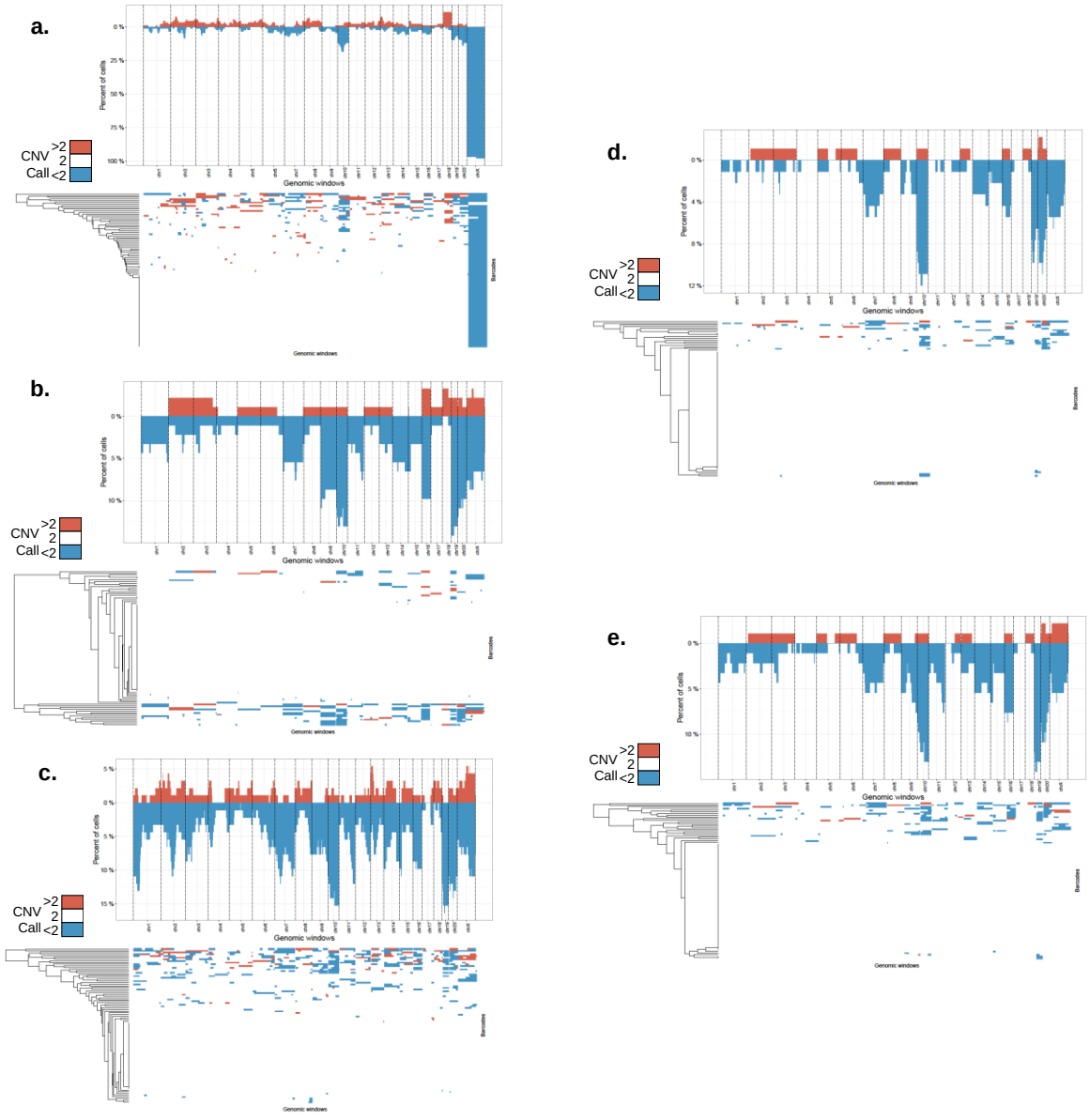
Supplementary Figure 11 | CNV profiles for Rhesus frontal cortex, Individual 1 using quasi-random priming (QRP). (**a**) Ginkgo Calls, (**b**) CBS calls, (**c**) HMM calls, (**d**) Intersection of all three, and (**e**) Intersection of just CBS and HMM.



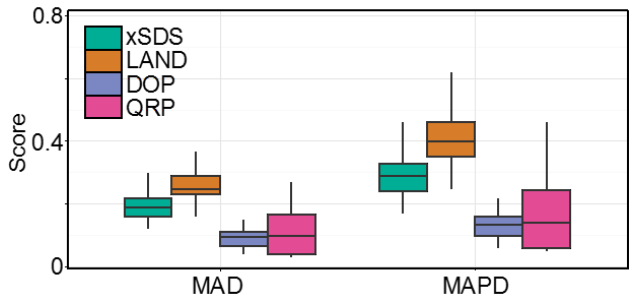
Supplementary Figure 12 | CNV profiles for Rhesus frontal cortex, Individual 1 using degenerate oligonucleotide primed PCR (DOP). (**a**) Ginkgo Calls, (**b**) CBS calls, (**c**) HMM calls, (**d**) Intersection of all three, and (**e**) Intersection of just CBS and HMM.



Supplementary Figure 13 | CNV profiles for Rhesus frontal cortex, Individual 1 using SCI-seq with LAND nucleosome depletion. (**a**) Ginkgo Calls, (**b**) CBS calls, (**c**) HMM calls, (**d**) Intersection of all three, and (**e**) Intersection of just CBS and HMM.

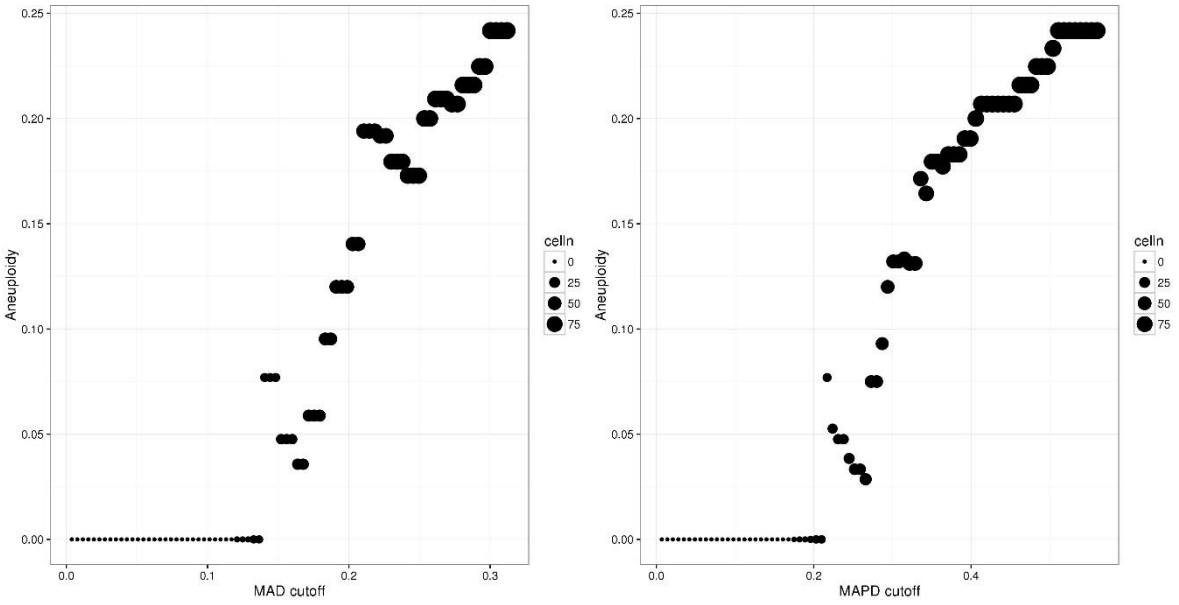


Supplementary Figure 14 | CNV profiles for Rhesus frontal cortex, Individual 1 using SCI-seq with xSDS nucleosome depletion. (**a**) Ginkgo Calls, (**b**) CBS calls, (**c**) HMM calls, (**d**) Intersection of all three, and (**e**) Intersection of just CBS and HMM.

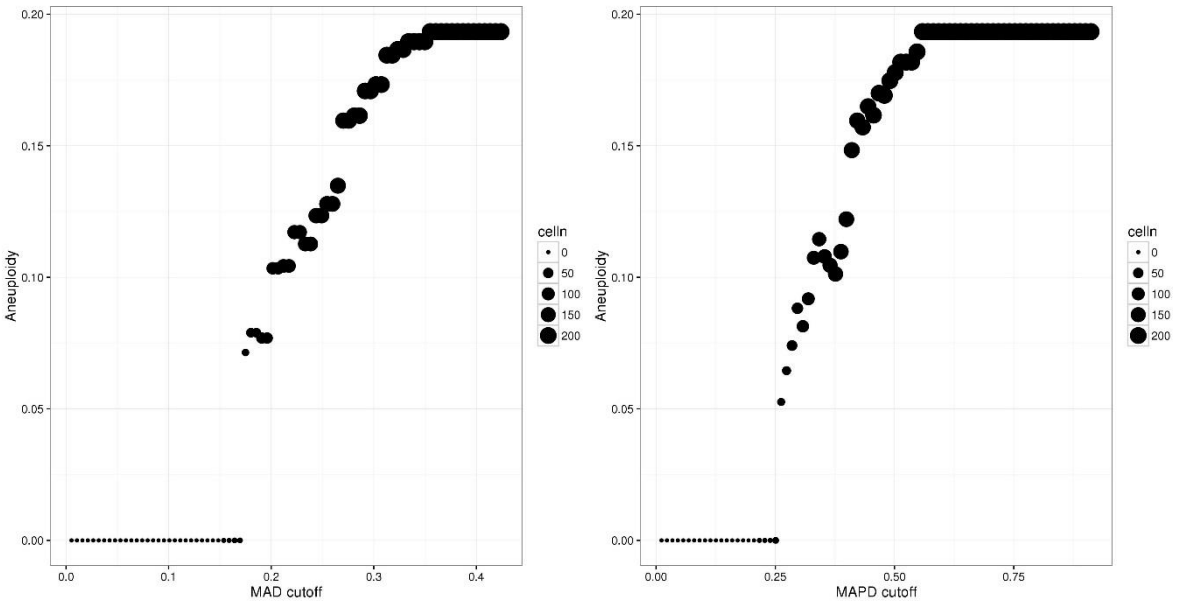


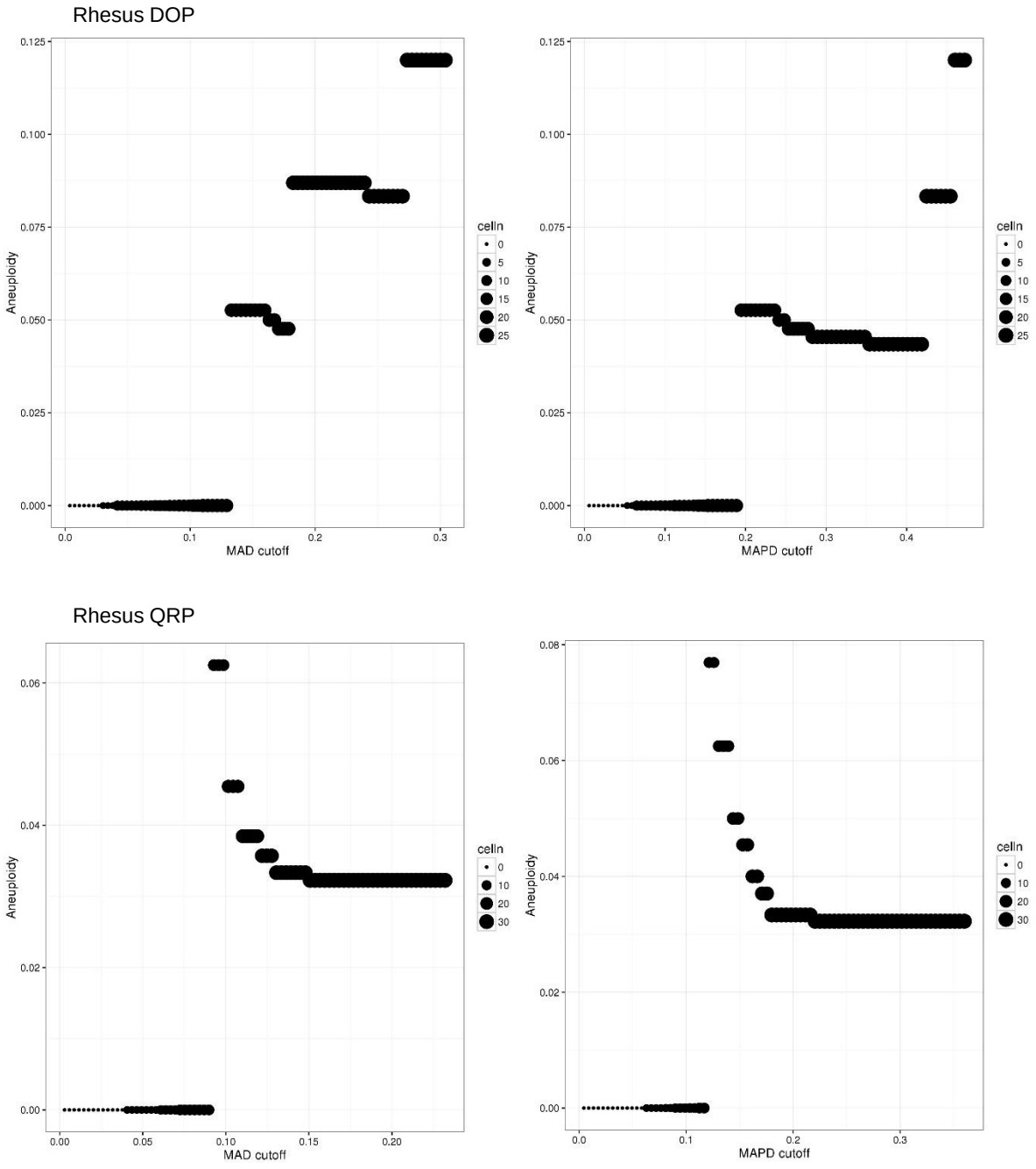
Supplementary Figure 15 | Comparison of coverage uniformity for Rhesus frontal cortex individual 1. Uniformity measures are very similar to those of GM12878 preparations (**Fig. 2b**).

Rhesus1 xSDS

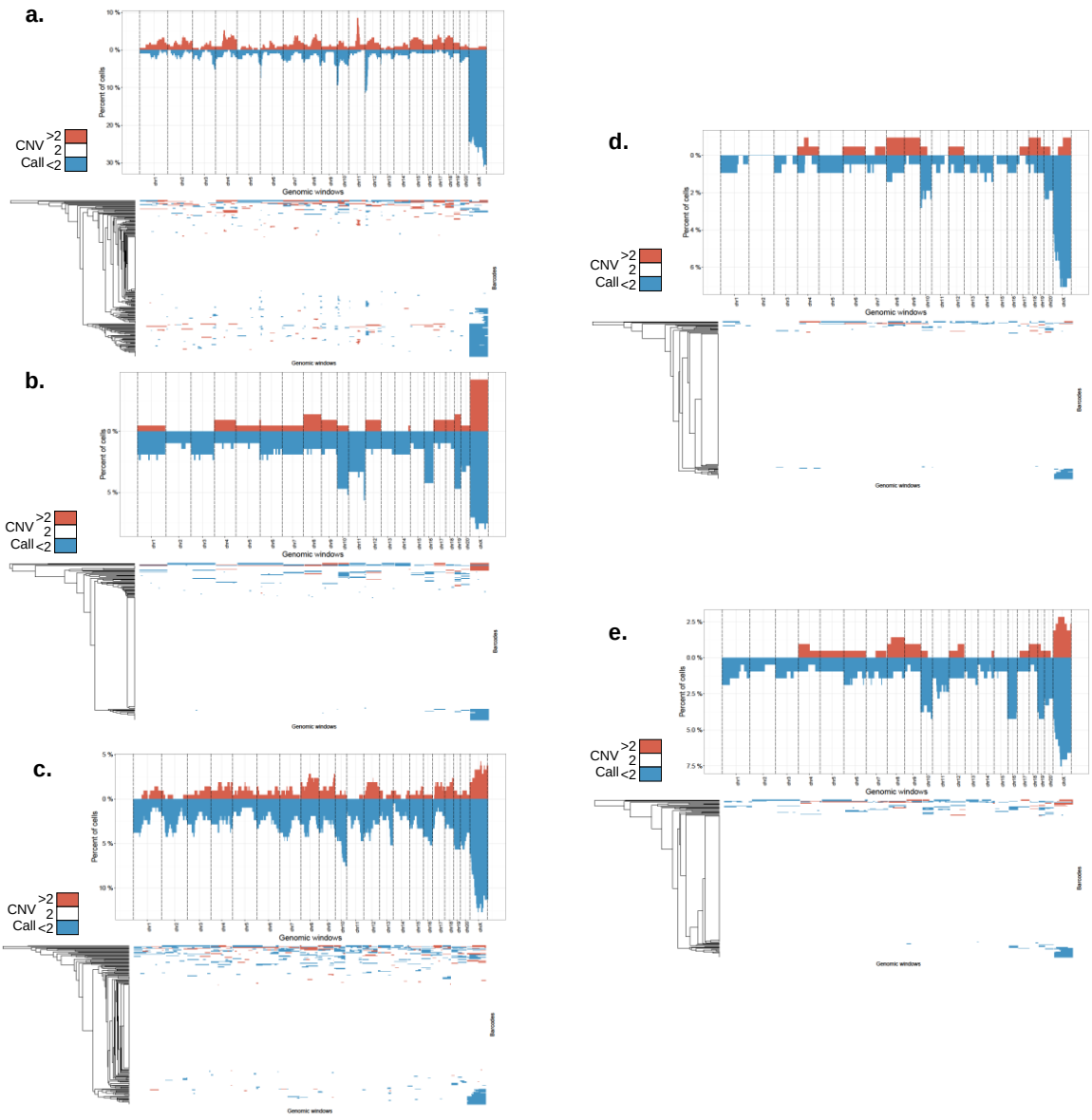


Rhesus2 xSDS

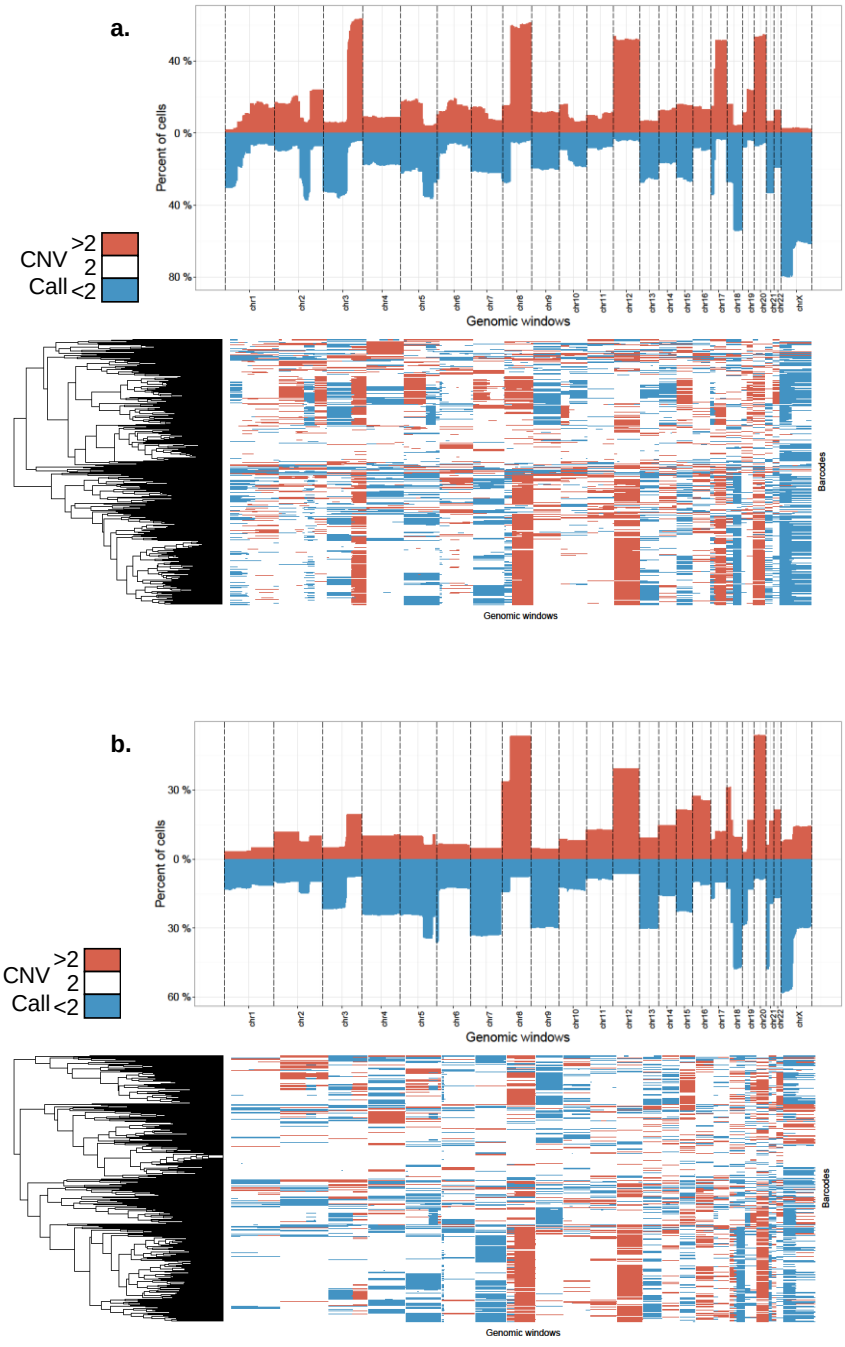


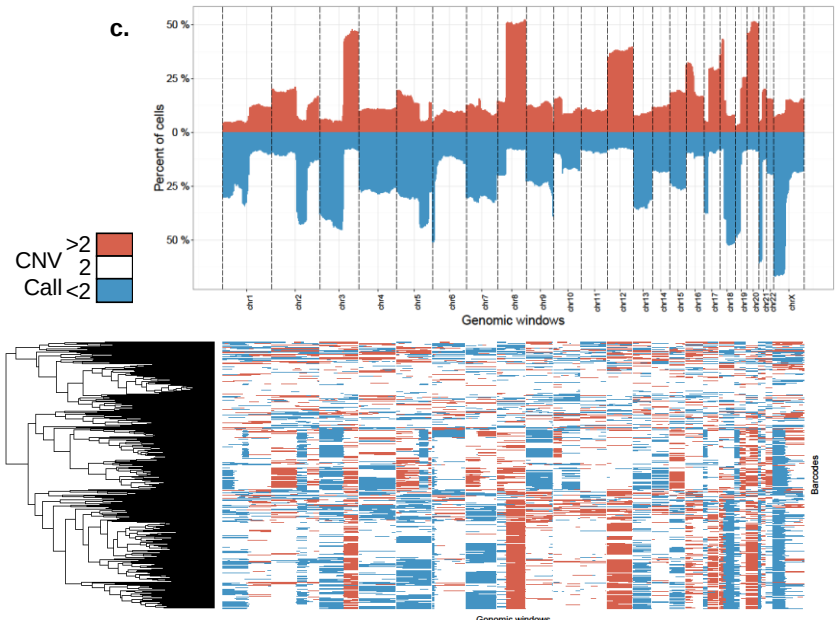


Supplementary Figure 16 | Rhesus aneuploidy rates across variance score cutoffs. Each point is the aneuploidy rate for the population of cells (y-axis), scaled by the number of cells included at a given score cutoff (x-axis).

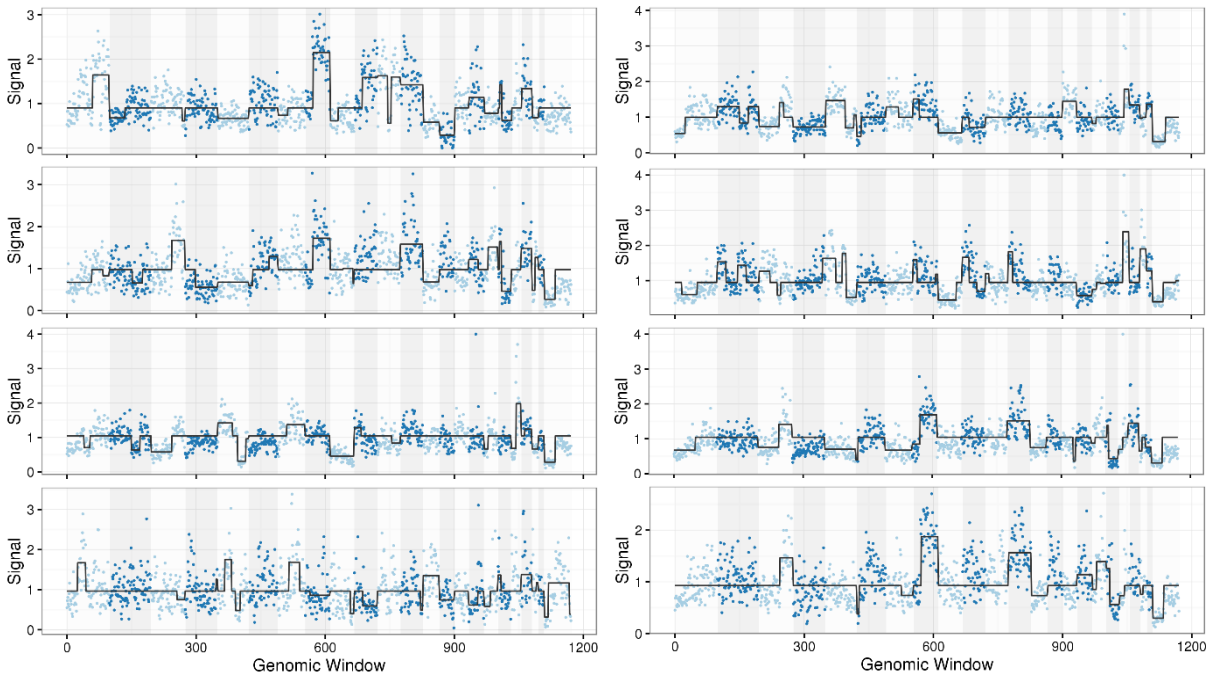


Supplementary Figure 17 | CNV profiles for Rhesus frontal cortex, Individual 2 using SCI-seq with xSDS nucleosome depletion. (**a**) Ginkgo Calls, (**b**) CBS calls, (**c**) HMM calls, (**d**) Intersection of all three, and (**e**) Intersection of just CBS and HMM.

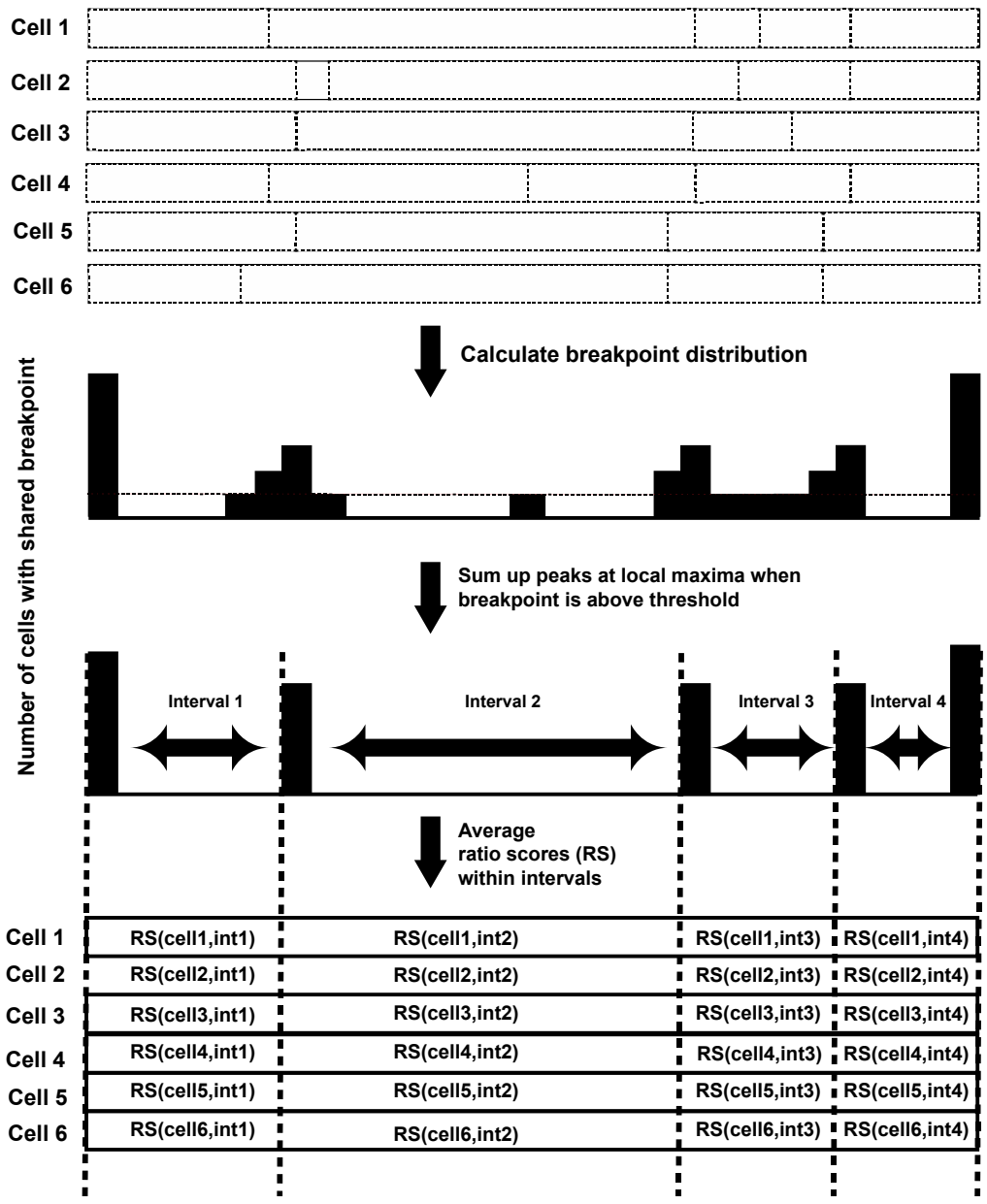

a.
Percent of cells
40 %
0 %
40 %
80 %
CNV
>2
2
Call
<2
Genomic windows
Barcodes
Genomic windows
b.
Percent of cells
30 %
0 %
30 %
60 %
CNV
>2
2
Call
<2
Genomic windows
Barcodes
Genomic windows



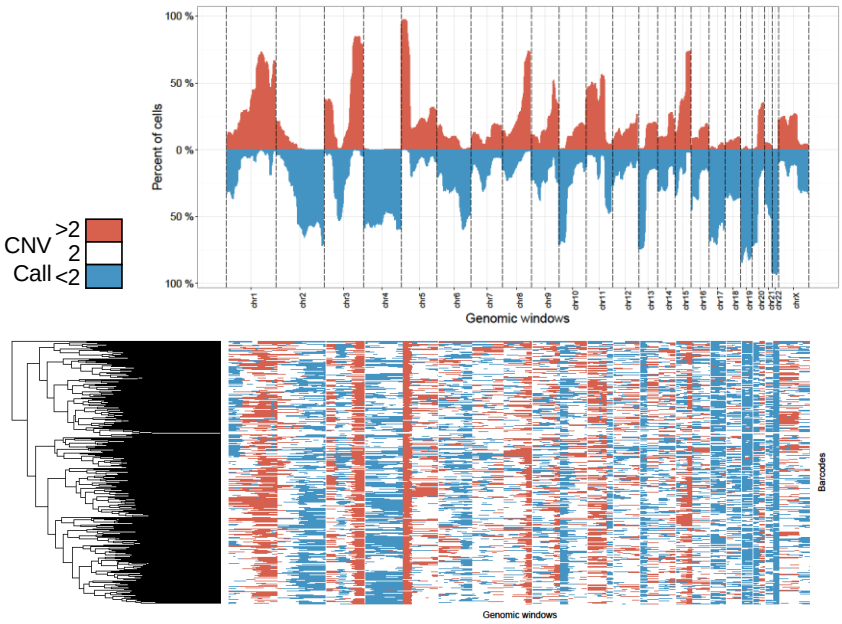
Supplementary Figure 18 | SCI-seq using xSDS-based nucleosome depletion on pancreatic ductal adenocarcinoma. Copy number call summary for 2.5 Mbp windows for the three methods of copy number calling used in the analysis: (**a**) Ginkgo, (**b**) CBS, and (**c**) HMM.



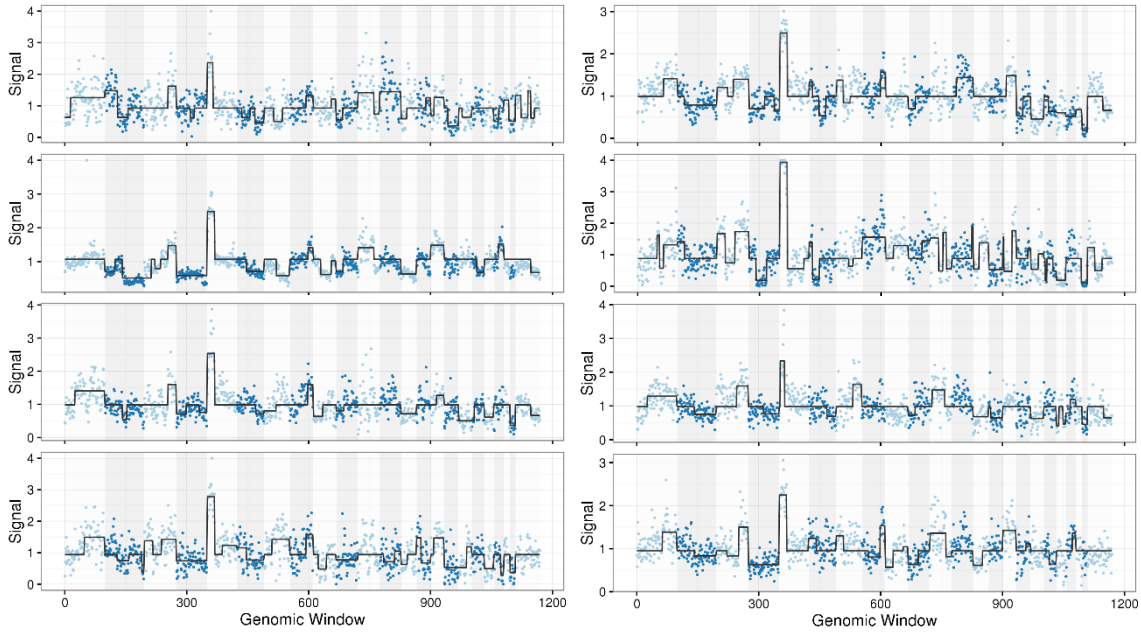
Supplementary Figure 19 | Single cell CNV calls on primary PDAC using xSDS SCI-seq. Representative single cell signal plots.



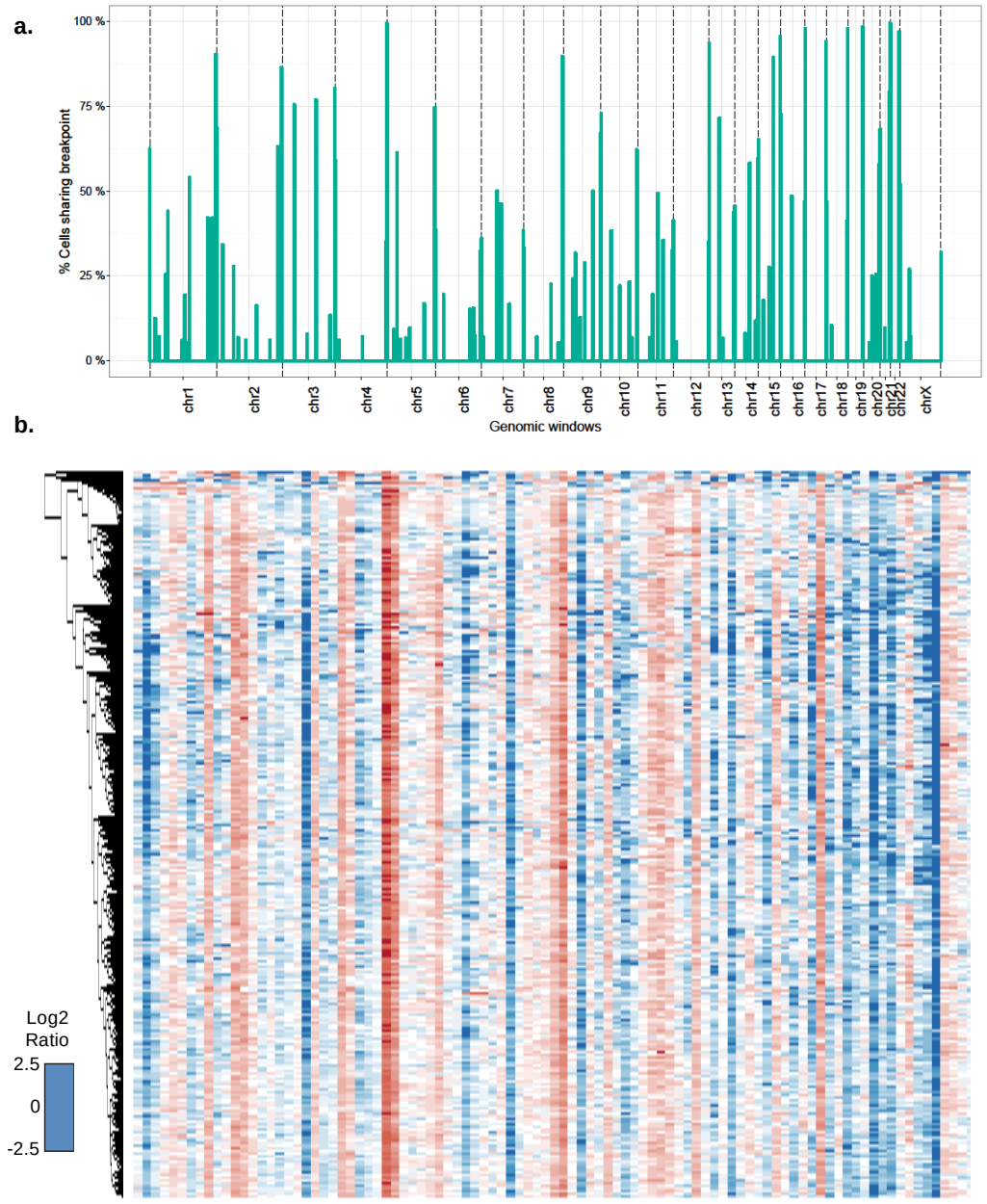
Supplementary Figure 20 | Schematic of breakpoint analysis workflow. First, individual cells are analyzed for breakpoints. Breakpoints from all cells are merged and locally summed when above threshold. Intervals are defined between local shared breakpoints and average ratio scores are found within each interval.



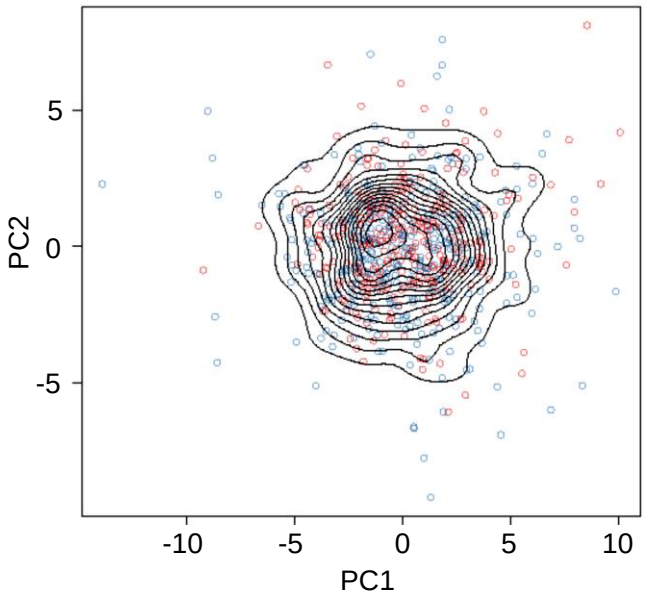
Supplementary Figure 21 | SCI-seq using LAND-based nucleosome depletion on HeLa S3 using the Hidden Markov Model method for copy number variant calling. Summary of windowed (2.5 Mbp) calls and hierarchical clustering of cells. CBC copy number calling resulted in a heavy bias against subchromosomal calls and Ginkgo failed to properly identify the ploidy in a number of cells resulting in a majority of cells called as entirely amplified.



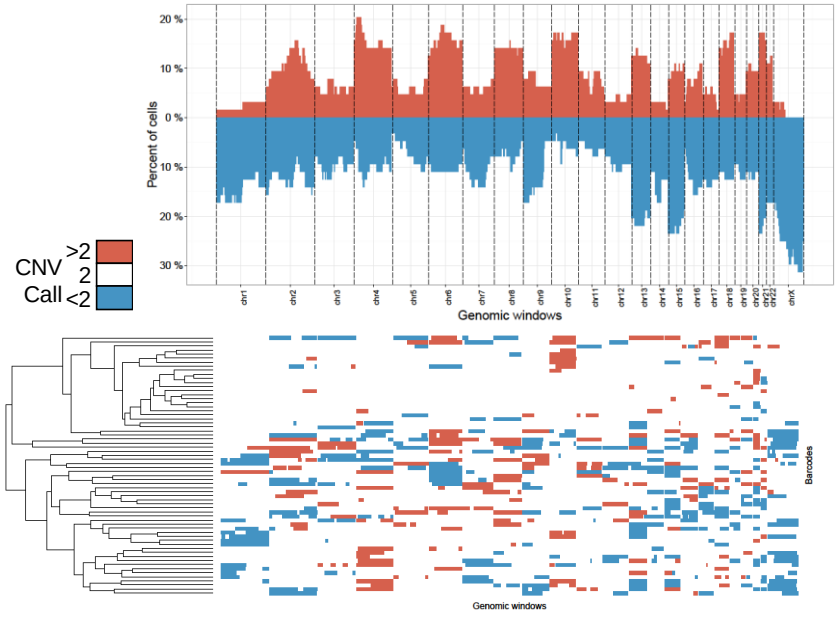
Supplementary Figure 22 | SCI-seq using LAND-based nucleosome depletion on HeLa S3 copy number variant calling in single cells using the Hidden Markov Model method. Representative single cell signal plots. A signal of 1 corresponds to the mean ploidy of 2.98.



Supplementary Figure 23 | Breakpoint analysis of HeLa. (**a**) Breakpoints identified in the HeLa cell line from an HMM analysis using 2.5 Mbp windows. (**b**) Log2 matrix of HeLa breakpoint windows for cells normalized to GM12878.



Supplementary Figure 24 | PCA on HeLa breakpoint windows. HeLa produces a single population as expected based on the stability of the cell line. Red and blue points indicate different preparations.



Supplementary Figure 25 | SCI-seq using xSDS-based nucleosome depletion on a banked stage II rectal cancer sample. Intersected copy number call summary for 2.5 Mbp windows.