ARCS: Scaffolding Genome Drafts with Linked Reads

Supplementary Information

Supplemental Tables

Table S1. Sequencing data source. The Illumina paired-end and mate pair sequencing data were used to generate contig and scaffold baseline assemblies with ABySS. The 10x Genomics (10XG) Chromium linked reads were used for assembly with Supernova, or processed for scaffolding the baseline contig and scaffold assemblies with ARCS, Architect and fragScaff.

Individual	Data type	URL
NA24143	Illumina paired-end 2x250bp	`https://github.com/genome-in-a-bottle/giab_data_indexes/blob/master/AshkenazimTrio/sequence.index.AJtrio_Illumina_2x250bps_06012016`
NA24143	Illumina mate-pair 6kbp	`https://github.com/genome-in-a-bottle/giab_data_indexes/blob/master/AshkenazimTrio/sequence.index.AJtrio_Illumina_6kb_matepair_wgs_08032015`
NA24143	10XG Chromium 2x151bp (raw[1])	`ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG004_NA24143_mother/`
NA24143	10XG Chromium 128/151bp (bam[2])	`https://github.com/genome-in-a-bottle/giab_data_indexes/blob/master/AshkenazimTrio/alignment.index.AJtrio_10Xgenomics_ChromiumGenome_GRCh37_GRCh38_06202016`
NA12878	10XG Chromium 2x151bp (raw[3])	`http://support.10xgenomics.com/de-novo-assembly/datasets/msNA12878`

[1] Corresponds to Table S3 dataset 1
[2] Corresponds to Table S3 dataset 2
[3] Corresponds to Table S3 dataset 4

Table S2. Baseline assembly specs. Contiguity length metrics and number of sequence alignment breakpoints for the baseline ABySS v2.0 contigs and scaffolds obtained from assembling GIAB NA24143 Illumina WGS 2x250 bp paired-end and 6 kbp mate-pair sequence data.

Assembly	Sequences	NG50	NGA50	Number of
Stage	$\geq 3kbp$	(bp)	(bp)	breakpoints
contig	80,910	50,351	47,878	1,746
scaffold	4,037	4,889,645	4,377,837	2,923

Table S3. 10x Genomics Chromium datasets used in our study. Chromium reads from individuals NA24143 and NA12878 were downloaded (datasets 1, 2 and 4). The sequencing data was converted from a container BAM file (dataset 2) to FASTQ format (dataset 3) or processed with 10XG longranger (Weisenfeld et al., 2016; Zheng et al., 2016) to generate barcode-containing interleaved FASTQ files (datset 5).

Dataset	Individual	Processing	Number of	Read length	Fold
		step	read pairs	(bp)	coverage
1	NA24143	Raw reads sequenced[1]	523,746,206	151	51.2
2	NA24143	Reads from BAM	420,496,741	128/151	34.9
3	NA24143	Filtered from BAM[2]	305,846,648	128/151	25.3
4	NA12878	Raw reads sequenced[1]	1,598,106,419	151	156.3
5	NA12878	Post Long Ranger[2]	1,514,291,941	128/151	136.8

[1] Used with Supernova

[2] Used with ARCS, Architect and fragScaff

Table S4. ARCS parameters and pertinent LINKS parameters for building the scaffold layout.

Module	Parameter	Description	Recommended range/value
ARCS	-f	Genome seq. assembly draft file (Multi-FASTA)	NA
ARCS	-a	File of file names listing BAM alignment files	NA
ARCS	-s	Min. percent sequence identity to consider reads	90-100, default: 98
ARCS	-c	Min. number of mapping read pairs/barcode and seq.	3-5, default: 5
ARCS	-l	Min. number of barcode links to create graph edge[1]	0-5, default: 0
ARCS	-z	Minimum sequence length to consider	250-5000, default: 500
ARCS	-m	Barcode read frequency range (min-max)	25-100000, default: 50-10000
ARCS	-d	Max. degree of nodes in graph	typically set to 0
ARCS	-e	Max. length to consider in 5' and 3' of seq.	10000-60000, default: 30000
ARCS	-r	Max. p-val. head/tail and orientation assignments	0.05-0.1, default: 0.05
LINKS	-l	Min. number of links to consider an edge	3-5, default: 5
LINKS	-a	Max. barcode link ratio between two edges at fork	0.3-0.9, default: 0.3
LINKS	-z	Minimum sequence length to consider	250-1000, default: 500

[1] Best handled in LINKS

Table S5. ARCS contiguity length metrics and breakpoints obtained from scaffolding contigs and scaffolds greater than 3kbp with various parameterizations. The NG50 and NGA50 lengths were calculated for scaffolds 500 bp and longer. Values in bold are plotted in the manuscript, Fig. 2. In ARCS, We consider (-*c* or more) reads that align to the 5' and 3' end (-*e* or less) bases of each sequences. The number of read pairs of the same barcode aligning to the head or the tail of a scaffold is tallied, and a binomial test is used to calculate whether the observed distribution is significantly different from a uniform distribution (threshold p=0.05, parameter -*r*). Once oriented relative to each other, pairs of sequence IDs are passed on to LINKS for generating the scaffold layout. Edges in the graph are considered with sufficient (-*l* or more) barcode links. Forks in the graph are resolved by choosing the edge with the most support, and when the ratio of barcode links of the second most supported edge relative to it is equal or below a threshold (-*a*).

Baseline assembly	*e*	*r*	*c*	*l*	*a*	NG50 (bp)	NGA50 (bp)	Breakpoints
contig[1]	30,000	0.05	5	5	0.3	**82,979**	**72,782**	**1,851**
contig	30,000	0.05	5	5	0.5	142,140	127,239	1,915
contig	30,000	0.05	5	5	0.7	**207,455**	**184,753**	**1,972**
contig	30,000	0.05	5	5	0.9	**303,034**	**268,962**	**2,030**
scaffold[1,2]	30,000	0.05	5	5	0.3	**11.74e6**	**7.87e6**	**2,985**
scaffold	30,000	0.05	5	5	0.5	13.81e6	9.05e6	2,999
scaffold	30,000	0.05	5	5	0.7	**15.13e6**	**10.22e6**	**3,003**
scaffold	30,000	0.05	5	5	0.9	**19.48e6**	**11.00e6**	**3,027**
scaffold	60,000	0.05	5	5	0.3	13.24e6	8.07e6	3,016
scaffold	60,000	0.05	5	5	0.5	15.69e6	9.38e6	3,033

[1] Benchmarking results for the corresponding assemblies are reported in the manuscript, Table 1

[2] The corresponding assembly is depicted in the manuscript, Fig. 3b

Table S6. fragScaff contiguity length metrics and breakpoints obtained from scaffolding contigs and scaffolds greater than 3kbp with various parameterizations. The NG50 and NGA50 lengths were calculated for scaffolds 500 bp and longer. Values in bold are plotted in the manuscript, Fig. 2. The parameters -*E*, -*C*, -*j*, and -*u* respectively control the sequence end node size, the minimum number of reads required to align to a node, the mean number of passing links across nodes and link validity.

Baseline assembly	*E*	*C*	*j*	*u*	NG50 (bp)	NGA50 (bp)	Breakpoints
contig	5,000	5	1	2	313,774	253,880	5,651
contig	5,000	5	1	3	193,266	171,334	3,263
contig	5,000	5	1	4	148,482	112,428	2,270
contig	5,000	5	1	5	85,861	76,430	2,017
contig	5,000	10	1	2	176,775	144,624	7,592
contig	5,000	10	1	3	141,559	122,812	5,180
contig	5,000	10	1	4	102,642	92,638	3,149
contig	5,000	10	1	5	77,145	70,639	2,301
contig	30,000	5	1	2	**304,926**	**231,937**	**6,345**
contig	30,000	5	1	3	**182,369**	**160,833**	**3,393**
contig[1]	30,000	5	1	4	**145,539**	**130,710**	**2,622**
contig	30,000	5	1	5	145,539	130,710	2,622
contig	30,000	5	2	2	673,216	314,033	13,191
contig	30,000	5	3	2	1.23e6	330,317	17,376
scaffold	5,000	1	1.25	2	14.13e6	6.41e6	3,575
scaffold	5,000	3	1.25	2	13.98e6	6.44e6	3,492
scaffold	5,000	5	1.25	2	11.85e6	6.10e6	3,495
scaffold	5,000	5	1	2	11.85e6	6.10e6	3,495
scaffold	5,000	5	1	3	11.20e6	6.10e6	3,435
scaffold	5,000	5	1	4	9.57e6	5.87e6	3,331
scaffold	5,000	5	2	2	11.85e6	6.10e6	3,495
scaffold	30,000	5	1	2	**13.13e6**	**6.41e6**	**3,438**
scaffold	30,000	5	1	3	**13.01e6**	**6.62e6**	**3,355**
scaffold[1,2]	30,000	5	1	4	**11.74e6**	**6.52e6**	**3,231**
scaffold	30,000	5	1	5	10.55e6	6.30e6	3,151
scaffold	30,000	5	2	2	16.93e6	6.52e6	3,813
scaffold	30,000	5	3	2	16.93e6	6.52e6	3,813

[1] Benchmarking results for the corresponding assemblies are reported in the manuscript, Table 1

[2] The corresponding assembly is depicted in the manuscript, Fig. 3a

Table S7. Architect contiguity length metrics and breakpoints obtained from scaffolding contigs and scaffolds greater than 3kbp with various parameterizations. The NG50 and NGA50 lengths were calculated for scaffolds 500 bp and longer. Values in bold are plotted in the manuscript, Fig. 2. The parameters *-t*, *-abs*, *-rel* and *-prun* in Architect control the minimum number of aligned reads from a barcode required for a sequence hit, the minimum number of aligning reads from a given barcode required to create a graph edge, the relative barcode support needed for creating edges and the relative barcode support needed for pruning edges, respectively.

Baseline assembly	*t*	*abs*	*rel*	*prun*	NG50 (bp)	NGA50 (bp)	Breakpoints
contig	5	3	0.2	0.2	59,442	**48,048**	**10,922**
contig	5	3	0.3	0.2	52,502	**47,887**	**4,035**
contig	5	3	0.4	0.2	50,689	47,876	2,113
contig	5	5	0.2	0.2	59,428	48,044	10,900
contig	5	5	0.3	0.2	52,499	47,887	4,030
contig	5	5	0.4	0.2	50,689	47,876	2,110
contig	10	3	0.2	0.2	58,171	48,026	9,105
contig	10	3	0.3	0.2	51,951	47,880	3,297
contig	10	3	0.4	0.2	50,577	47,876	1,995
contig	10	5	0.2	0.2	58,170	48,026	9,083
contig	10	5	0.3	0.2	51,948	47,880	3,292
contig[1]	10	5	0.4	0.2	50,570	47,876	1,991
scaffold	5	3	0.1	0.1	5.48e6	4.38e6	3,293
scaffold	5	3	0.2	0.1	5.01e6	4.38e6	3,076
scaffold	5	3	0.2	0.2	**5.01e6**	**4.38e6**	**3,076**
scaffold	5	3	0.3	0.2	**4.93e6**	**4.38e6**	**2,991**
scaffold	5	3	0.4	0.2	4.93e6	4.38e6	2,974
scaffold	5	5	0.2	0.2	5.01e6	4.38e6	3,076
scaffold	5	5	0.3	0.2	4.93e6	4.38e6	2,991
scaffold[1]	5	5	0.4	0.2	4.93e6	4.38e6	2,974
scaffold	10	3	0.1	0.1	5.36e6	4.38e6	3,216
scaffold	10	3	0.2	0.1	5.01e6	4.38e6	3,060
scaffold	10	3	0.2	0.2	5.01e6	4.38e6	3,060
scaffold	10	3	0.3	0.2	4.93e6	4.38e6	2,981
scaffold	10	3	0.4	0.2	4.89e6	4.38e6	2,973
scaffold	10	5	0.2	0.2	5.01e6	4.38e6	3,056
scaffold	10	5	0.3	0.2	4.93e6	4.38e6	2,981
scaffold	10	5	0.4	0.2	4.89e6	4.38e6	2,973

[1] Benchmarking results for the corresponding assemblies are reported in the manuscript, Table 1. The parameters were abbreviated to fit the table: *abs, rel* and *prun* correspond to *–rc-abs-thr, –rc-rel-edge-thr* and *–rc-rel-prun-thr*, respectively

Table S8. Total wall-clock time and peak memory usage for ARCS (-*c* 5 -*e* 30000 -*r* 0.05 -*l* 5 -*a* 0.3,0.7,0.9), Architect (-*t* 5 -*rc-abs-thr* 3 -*rc-rel-prun-thr* 0.2 -*rc-rel-edge-thr* 0.2,0.3, abbreviated to rel) and fragScaff (-*C* 5 -*E* 30000 -*j* 1 -*u* 2,3,4) scaffolding applied to the baseline contig assembly.

Scaffolder	ARCS	ARCS	ARCS	fragScaff	fragScaff	fragScaff	Architect	Architect
Parameters	a=0.3	a=0.7	a=0.9	u=2	u=3	u=4	rel=0.2	rel=0.3
Number of threads	1	1	1	64	64	64	1	1
Wall-clock time (h:mm)	1:12	1:12	1:11	6:40	6:35	6:37	190:44	187:25
Peak memory (GB)	9.4	9.4	9.4	8.1	8.1	8.1	13.3	13.2

Table S9. Supernova (SN) assemblies of a human Chromium datasets and comparison to ARCS scaffolding of a human ABySS scaffold assembly. Values in bold are plotted in the manuscript, Fig. 2b.

Data-	Individual	Assembly	Cut-off[1]	n	NG50	NGA50	N50	Largest	Break-
set			size (kbp)		(Mbp)	(Mbp)	(Mbp)	(Mbp)	points
4	NA12878	10XG SN v1.0	10	1,231	14.66	5.27	16.40	68.87	3,737
4	NA12878	Local SN v1.1	10	1,341	14.74	5.12	16.22	57.01	3,782
4	**NA12878**	**Local SN v1.1**	0.5	21,774	**14.74**	**5.12**	16.10	57.01	**3,782**
1	**NA24143**	**Local SN v1.1**	0.5	23,693	**13.47**	**5.38**	15.03	95.16	**3,879**
3	NA24143	ARCS v1.0[2]	0.5	64,922	19.48	11.00	21.82	97.86	3,027
5	NA12878	ARCS v1.0[3]	0.5	64,516	18.34	8.95	22.16	111.6	3,225

[1] Cut-off size for reporting the assembly length metrics

[2] Parameters: -*m* 50-1000 -*s* 98 -*z* 3000 -*e* 30,000 -*r* 0.05 -*c* 5 -*l* 5 -*a* 0.9

[3] Parameters: -*m* 50-6000 -*s* 98 -*z* 3000 -*e* 30,000 -*r* 0.05 -*c* 5 -*l* 5 -*a* 0.9

Table S10. Average contiguity and breakpoints analysis of ARCS assemblies. In triplicate experiments, we sub-sampled Chromium read data and ran ARCS (-*c* 5 -*r* 0.05 -*e* 30000 -*z* 3000 -*m* 50-6000 for NA12878, -*m* 50-1000 for NA24143) with LINKS (-*l* 5 -*a* 0.9) on the baseline scaffold assembly and report the average NG50, NGA50 length metrics, breakpoints and standard deviation (S.D.).

10XG	Read pairs	Fold	NG50	S.D.	NGA50	S.D.	Breakpoints	S.D.
Dataset	(M)	coverage	(Mbp)	(Mbp)	(Mbp)	(Mbp)		
NA12878	45.7	4.1	8.0	0.3	6.1	0.1	2,956.0	11.5
NA12878	200.0	18.1	14.2	0.6	8.1	0.0	3,066.0	0.0
NA12878	400.0	36.3	15.4	1.8	8.7	0.2	3,139.3	10.5
NA12878	600.0	54.4	16.4	1.4	8.7	0.2	3,170.5	2.5
NA12878	800.0	72.5	15.8	2.2	8.9	0.1	3,192.0	18.5
NA12878	1,000.0	90.7	17.0	1.6	8.9	0.0	3,212.0	10.4
NA12878	1,200.0	108.8	17.3	0.3	8.9	0.0	3,192.0	17.2
NA12878	1,400.0	126.9	16.4	1.4	8.9	0.0	3,212.0	4.0
NA24143	100.0	8.5	10.8	0.1	7.5	0.4	2,963.7	10.8
NA24143	200.0	17.1	18.8	1.5	11.1	0.1	3,003.3	15.7
NA24143	300.0	25.6	19.4	0.0	11.0	0.0	3,031.0	0.0

Table S11. Average contiguity and breakpoints analysis of Architect assemblies. In triplicate experiments, we sub-sampled Chromium read data and ran Architect (-*t* 5 –*rc-abs-thr* 3 –*rc-rel-edge-thr* 0.2 –*rc-rel-prun-thr* 0.2) on the baseline scaffold assembly and report the average NG50, NGA50 length metrics, breakpoints and standard deviation (S.D.).

10XG	Read pairs	Fold	NG50	S.D.	NGA50	S.D.	Breakpoints	S.D.
Dataset	(M)	coverage	(Mbp)	(Mbp)	(Mbp)	(Mbp)		
NA12878	45.7	4.1	5.0	0.0	4.4	0.0	3027.7	2.1
NA12878	200.0	18.1	5.5	0.0	4.4	0.0	3352.3	8.5
NA12878	400.0	36.3	5.5	0.0	4.4	0.0	3357.0	14.5
NA12878	600.0	54.4	5.5	0.0	4.4	0.0	4.6	4.6
NA12878	800.0	72.5	5.5	0.0	4.4	0.0	4.6	4.6
NA12878	1000.0	90.7	5.5	0.0	4.4	0.0	5.7	5.7
NA12878	1200.0	108.8	5.5	0.0	4.4	0.0	7.1	7.1
NA12878	1400.0	126.9	5.5	0.0	4.4	0.0	0.6	0.6
NA24143	100.0	8.5	4.9	0.0	4.4	0.0	2977.3	0.6
NA24143	200.0	17.1	5.0	0.0	4.4	0.0	3014.7	2.3
NA24143	300.0	25.6	5.0	0.0	4.4	0.0	3078.0	0.0

Table S12. Average contiguity and breakpoints analysis of fragScaff assemblies. In triplicate experiments, we sub-sampled Chromium read data and ran fragScaff (-*E* 30000 -*C* 5 -*j* 1 -*u* 2) on the baseline scaffold assembly and report the average NG50, NGA50 length metrics, breakpoints and standard deviation (S.D.).

10XG Dataset	Read pairs (M)	Fold coverage	NG50 (Mbp)	S.D. (Mbp)	NGA50 (Mbp)	S.D. (Mbp)	Breakpoints	S.D.
NA12878	45.7	4.1	4.9	0.1	4.4	0.0	2921.0	3.0
NA12878	200.0	18.1	5.6	0.1	4.6	0.1	3236.0	8.7
NA12878	400.0	36.3	6.5	0.3	5.1	0.1	3329.3	50.8
NA12878	600.0	54.4	7.0	0.2	5.3	0.0	3352.3	25.0
NA12878	800.0	72.5	8.3	0.3	5.6	0.1	3468.0	94.0
NA12878	1000.0	90.7	9.4	0.7	5.9	0.2	3525.3	30.9
NA12878	1200.0	108.8	10.1	0.2	6.0	0.1	3525.7	37.1
NA12878	1400.0	126.9	10.5	0.3	6.1	0.1	3539.3	34.3
NA24143	100.0	8.5	5.0	0.1	4.4	0.0	2919.7	3.2
NA24143	200.0	17.1	11.5	0.1	5.8	0.1	3293.7	18.6
NA24143	300.0	25.6	13.1	0.0	6.4	0.0	3456.0	0.0

Supplemental Figures

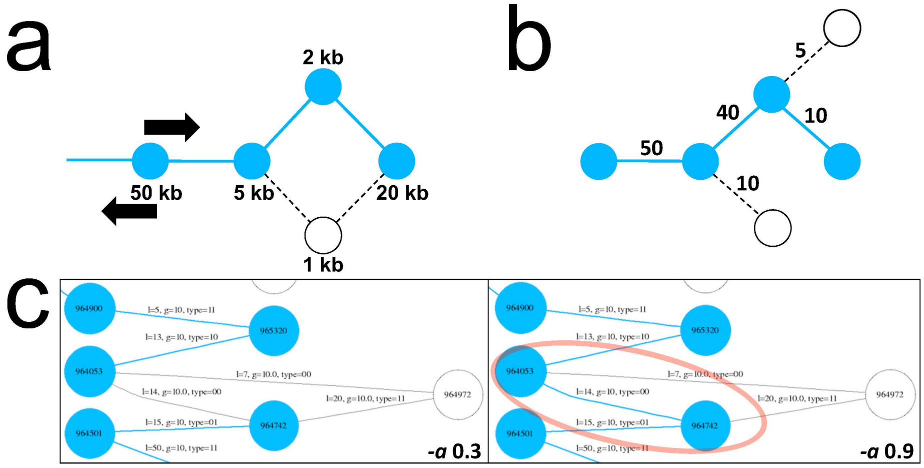

Fig. S1. LINKS scaffolding. (a) In the initial phase, the layout is progressively built starting with the largest sequence (left-most vertex), first looking at ARCS pairs in 3' (black arrow, pointing right). Sequences equal or larger than a min. length -*z* are considered until possibilities are exhausted (lengths in kbp, 2 kbp cutoff shown). The layout is then extended on the 5' end (black arrow, facing left). The likely path between sequences is shown in blue with the excluded sequence, shorter than the min. length cutoff, shown in the black outline. (b) The parameters -*l* (min. number of barcode support required, example numbers above each edge shown) and -*a* (max. links ratio) both control extension. In the example, -*l* is set to 10 and -*a* 0.3. At the left-most fork, both vertices are considered. The edge with the highest barcode support is favored, only if the links ratio is below -*a*. In the example, the links ratio, calculated as the number of barcodes of the second-most (10) compared to the most (40) supported linkage is 10/40 = 0.25 below 0.3 and thus the path with highest support is chosen. At the right-most fork, only the vertex with 10 barcode supports satisfies the -*l* cutoff. (c) Scaffold graph sections from scaffolding a draft human genome GIAB HG004 assembly with ARCS while imposing a stringent (left, -*a* 0.3) or more relaxed (right, -*a* 0.9) max. links ratio cutoff. Unconnected and connected components are shown in black outline and blue, respectively. The sequence identifier is shown within each vertex. The number of barcodes (l=), est. gap size (g=) and relative orientation of sequence pairs (type= with 1,0 for forward and reverse orientations). The ellipse highlights a new linkage.

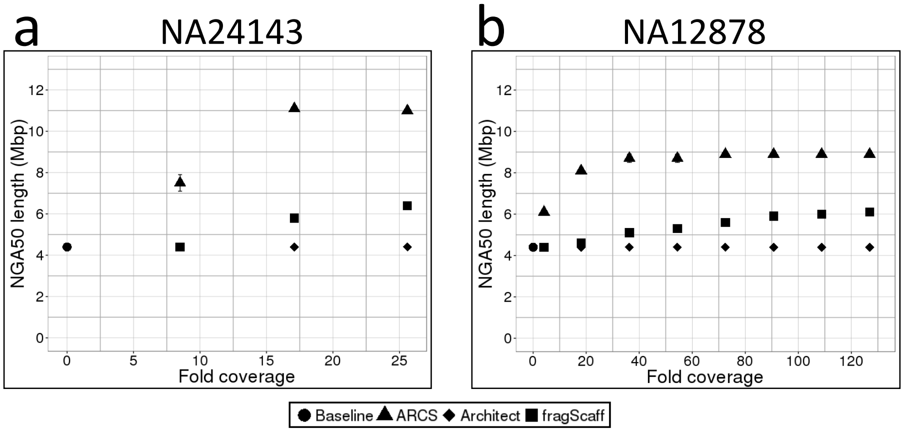

Fig. S2. ARCS, fragScaff and Architect scaffolding results on sub-sampled 10XG Chromium reads from three independent runs. In separate, triplicate experiments, we sub-sampled (a) 100, 200, 300M NA24143 and (b) 46, 200-1400M NA12878 10XG read pairs to test the effect of coverage on scaffolding of the baseline scaffold assembly draft using the three scaffolding tools. The pipeline ran on each file subset with ARCS (*-c* 5 *-r* 0.05 *-e* 30000 *-z* 3000 *-m* 50-6000 for NA12878, *-m* 50-1000 for NA24143) and LINKS (*-l* 5 *-a* 0.9), fragScaff (*-E* 30000 *-C* 5 *-j* 1 *-u* 2) and Architect (*-t* 5 *–rc-abs-thr* 3 *–rc-rel-edge-thr* 0.2 *–rc-rel-prun-thr* 0.2).

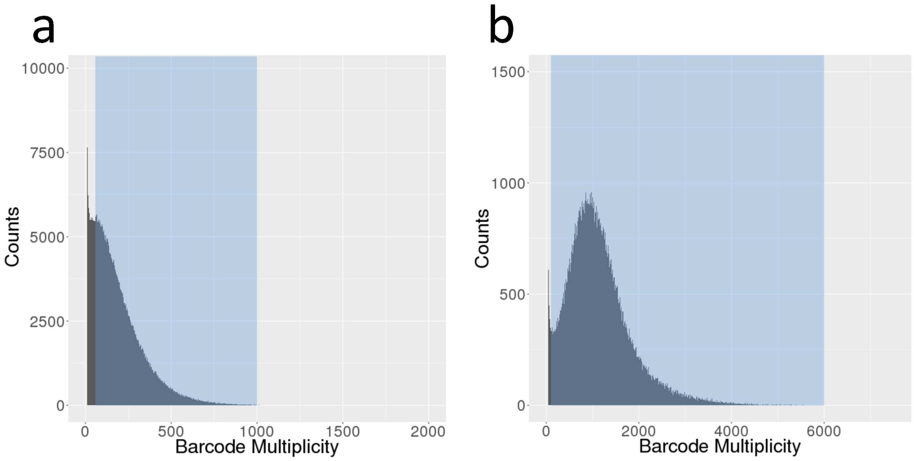

Fig. S3. Distributions of barcode-read multiplicities (read frequency per index) in human (a) NA24143 and (b) NA12878 Chromium datasets. Blue shades show the multiplicity range we set in ARCS as *-m* 50-1000 and *-m* 50-6000 for the NA24143 and NA12878 Chromium sequence data, respectively.