Supplementary information

Table S1. 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 seq. 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	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	Min. seq. length to consider	250-1000, default: 500

[1]Best handled in LINKS

Table S2. Datasets used in our study.

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

Table S3. 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 Stage	Sequences $\geq 3kbp$	NG50 (bp)	NGA50 (bp)	Number of breakpoints
contig	80,910	50,351	47,878	1,746
scaffold	4,037	4,889,645	4,377,837	2,923

Table S4. 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 onto 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 S5. 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 S6. 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 S7. Supernova (SN) assemblies of a human Chromium datasets and comparison to ARCS scaffolding of a human ABySS scaffold assembly.

Dataset	Individual	Assembly	Cut-off[1] size (kbp)	n	NG50 (bp)	N50 (bp)	Largest (bp)	Breakpoints
4	NA12878	10XG SN v1.0	10	1,231	14.66e6	16.40e6	68.87e6	3,737
4	NA12878	Local SN v1.1	10	1,341	14.74e6	16.22e6	57.01e6	3,782
4	NA12878	Local SN v1.1	0.5	21,774	14.74e6	16.10e6	57.01e6	3,782
1	NA24143	Local SN v1.1	0.5	23,693	13.47e6	15.03e6	95.16e6	3,879
3	NA24143	ARCS v1.0[2]	0.5	64,922	19.48e6	21.82e6	97.86e6	3,027
5	NA12878	ARCS v1.0[3]	0.5	64,516	18.34e6	22.16e6	111.6e6	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

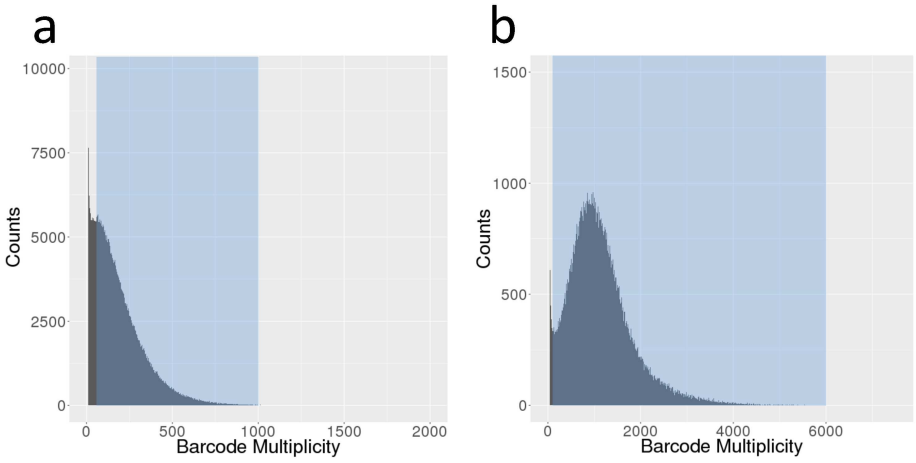

Fig. S1. Distributions of barcode-read multiplicities (read frequency per index) in human (a) NA24143 and (b) NA12878 Chromium datasets. Blue shades show the multiplicity range we set in ARCS as *-m* 50-1000 and *-m* 50-6000 for the NA24143 and NA12878 Chromium sequence data, respectively.