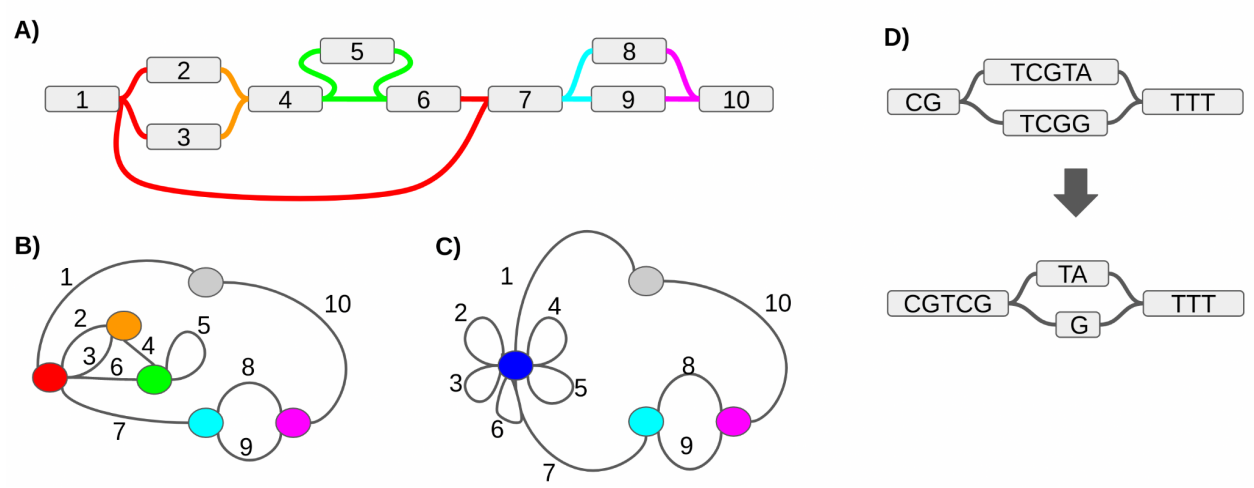
nature biotechnology

Article

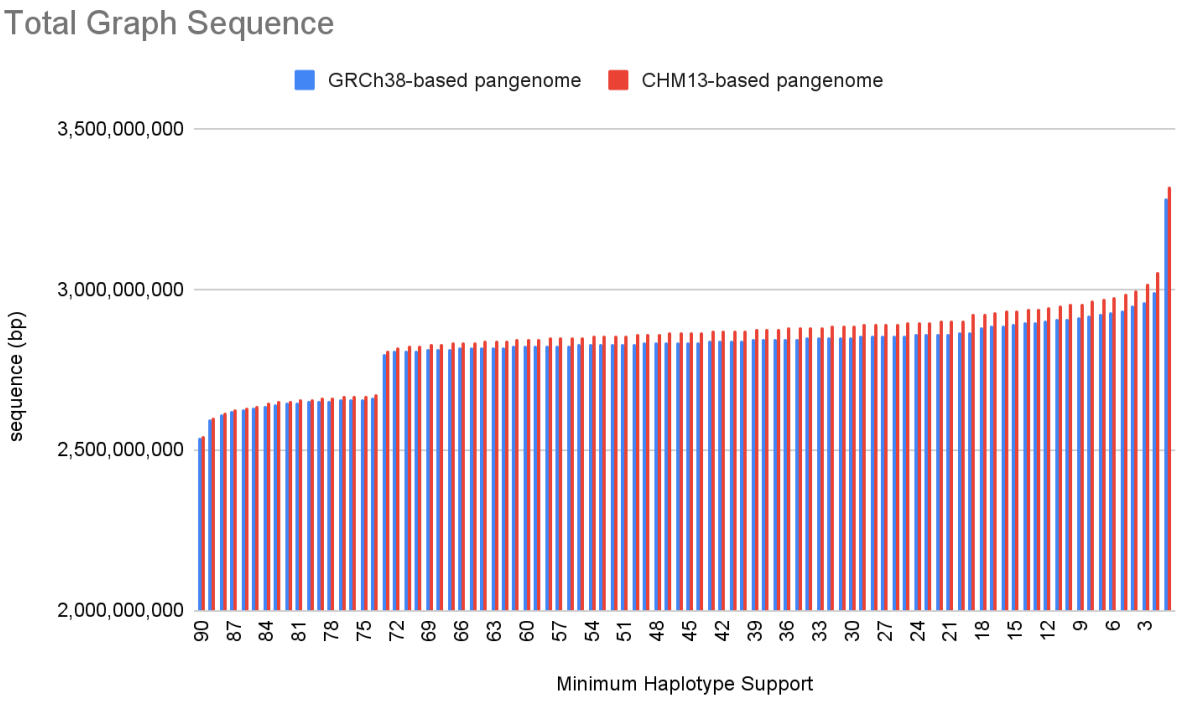
https://doi.org/10.1038/s41587-023-01793-w

Pangenome graph construction from genome alignments with Minigraph-Cactus

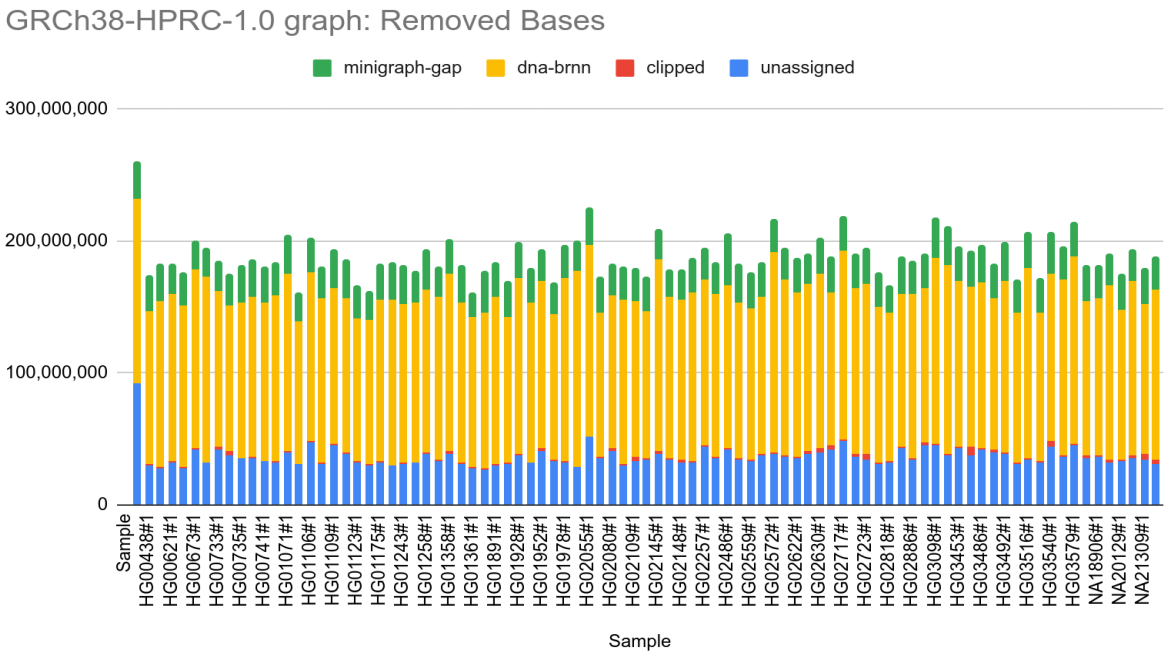
In the format provided by the authors and unedited

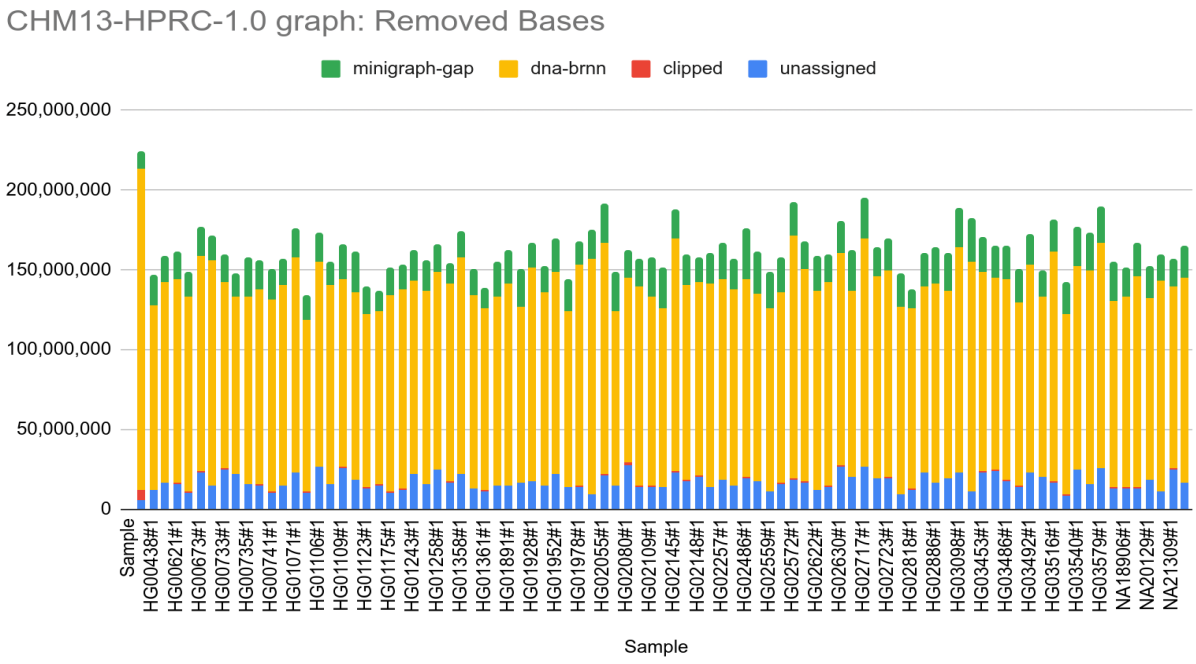


Supplementary Figure 1: A-C: Cactus graph example. **A)** A sequence graph where the ID of each node is shown but not its sequence. We consider two edges “connected” if they are incident to the same “end” of the same node. Connected components of edges under this definition are grouped together, with each component being given a separate colour. **B)** Each connected component of edges in the sequence graph is grouped together into a node. Each node in the sequence graph is transformed into an edge. A “root” node (gray) is created to connect to all node ends that have degree 0 in the sequence graph (stubs). **C)** The cactus graph is created by merging together all 3-edge-connected components of nodes in the graph from B). This graph has the property that no edge is part of more than one simple cycle. **D)** 3bp of redundant sequence, “TCG”, is removed with GFAffix. This sequence is redundant in the sense that its removal does not affect the number of possible haplotype sequences that can be represented by the graph.

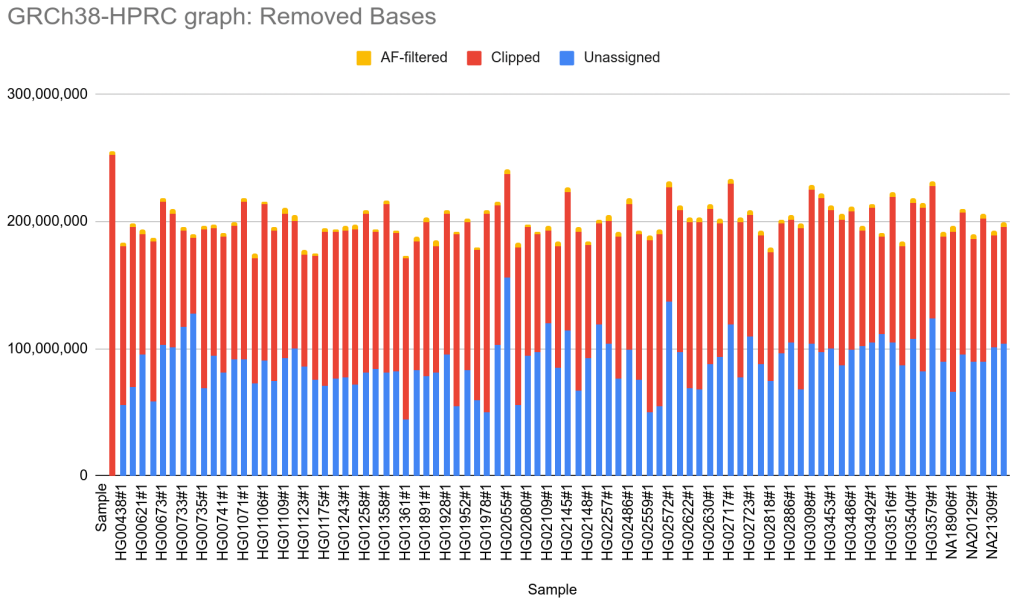


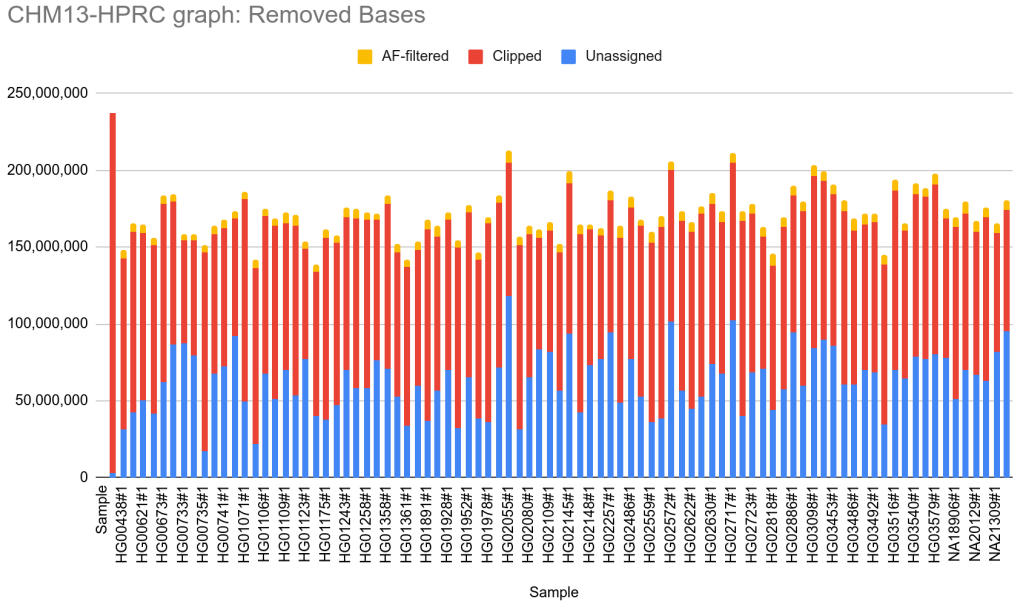
Supplementary Figure 2: The amount of sequence in the HPRC graphs by the minimum number of haplotypes that contain it. The step in the graph is due to 14 male haplotypes not possessing an X chromosome.



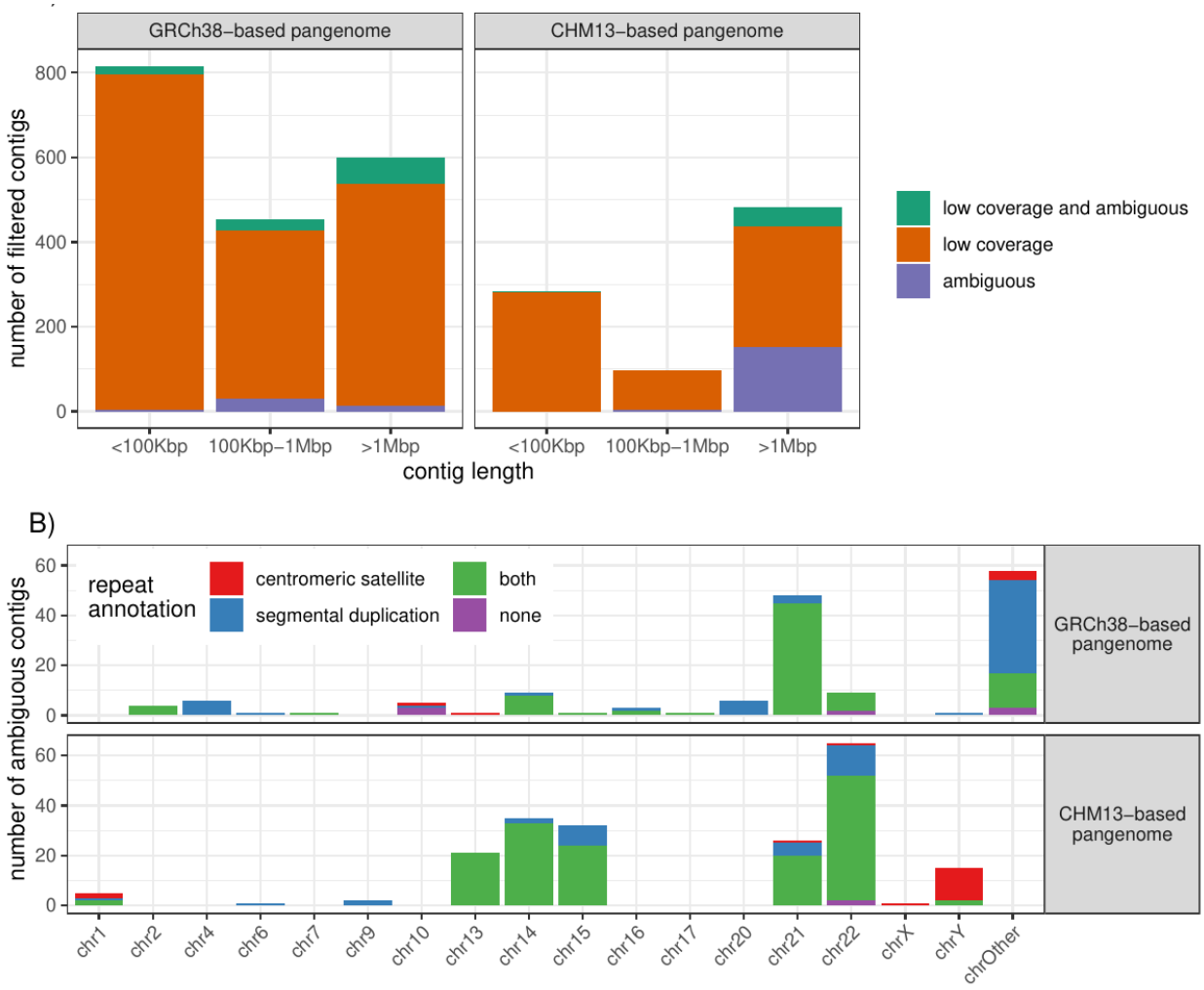


Supplementary Figure 3: Sequence excluded from the HPRC pangenome

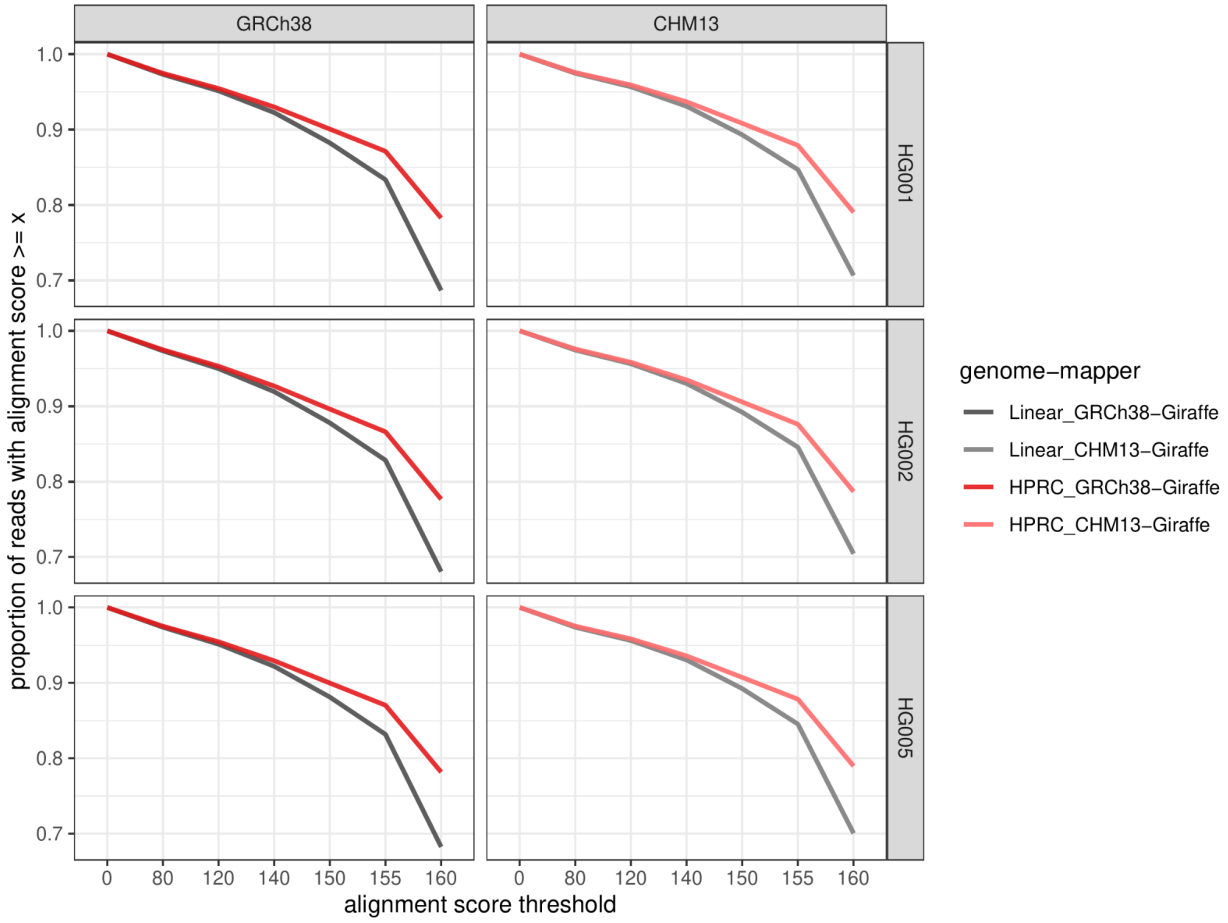




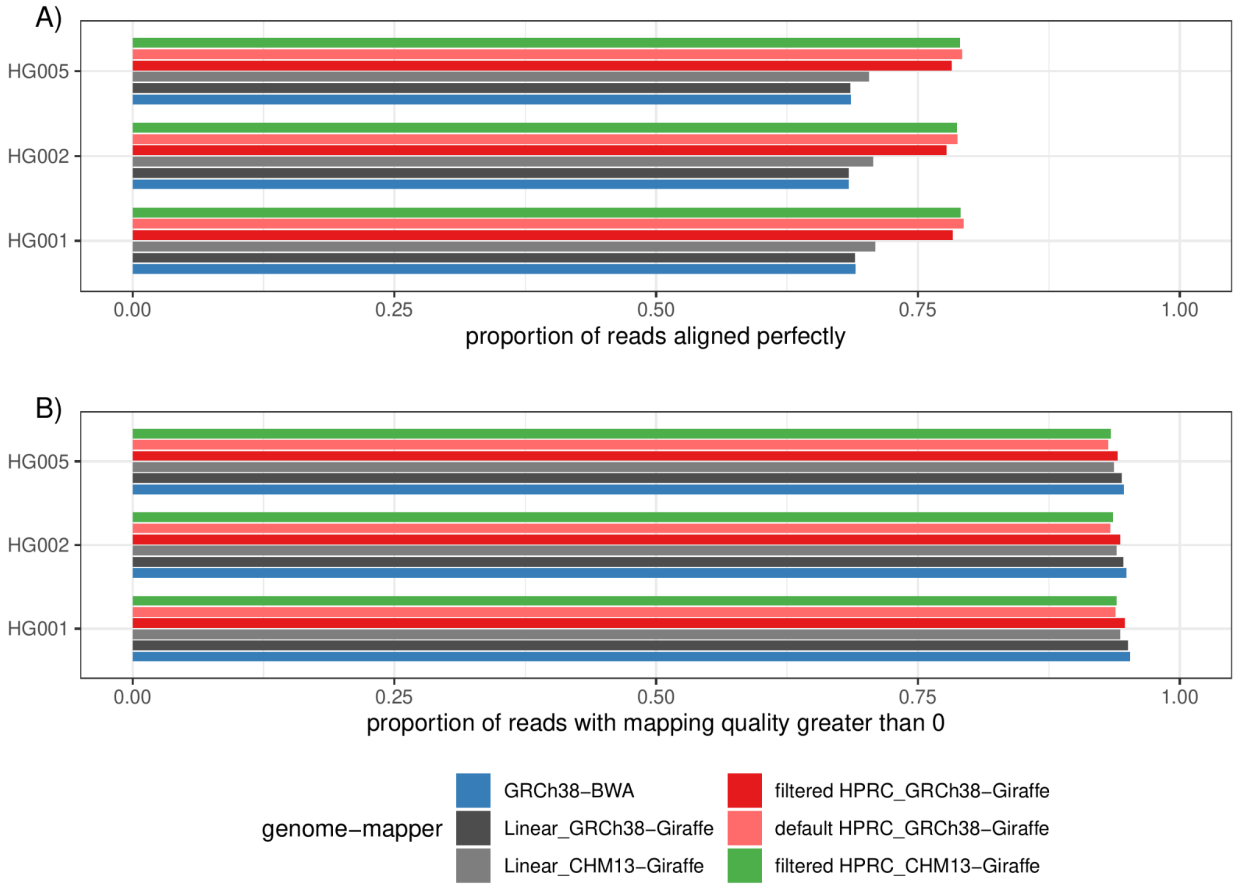
Supplementary Figure 4: Sequence excluded from the HPRC pangenomes when using the current pipeline (without dna-brnn preprocessing).



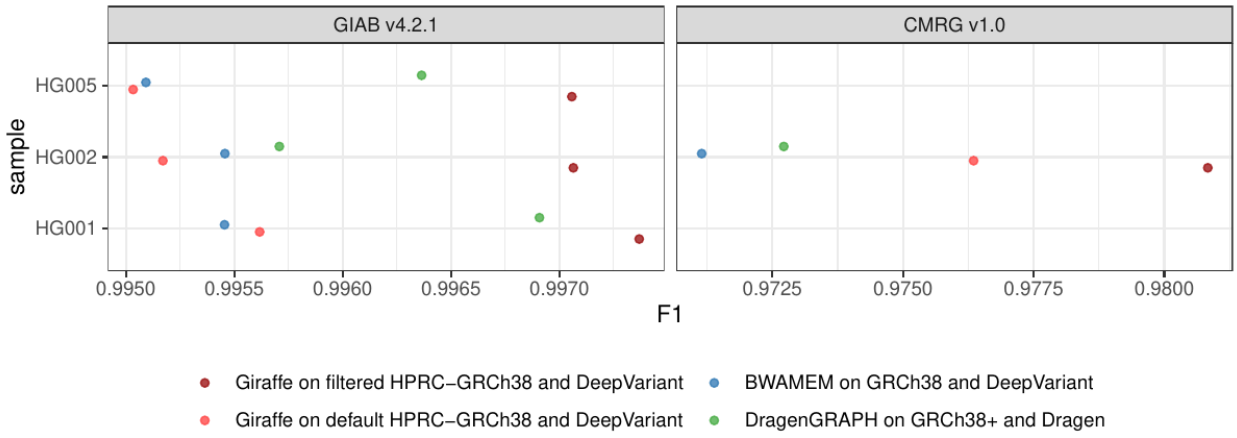
Supplementary Figure 5: Characterization of contigs filtered out as low coverage or ambiguous from the HPRC pangenomes (corresponding to the blue bars in **Supplementary Figure 4**). “chrOther” refers to any unlocalized or unplaced scaffold from GRCh38 (ex. chr1_KI270709v1_random). **A)** Distribution of contigs by size as filtered out by coverage and/or ambiguity. **B)** The distribution of chromosomes with the highest coverage for each contig that couldn’t be unambiguously assigned to a chromosome. The colors indicate contigs with at least 10% of sequence annotated as segmental duplication (blue), centromeric satellite (red), or both (green). The annotation tracks for each assembly, released as part of Liao et al., 2022, were produced with SEDEF and dna-brnn, for segmental duplications and centromeric satellites respectively.



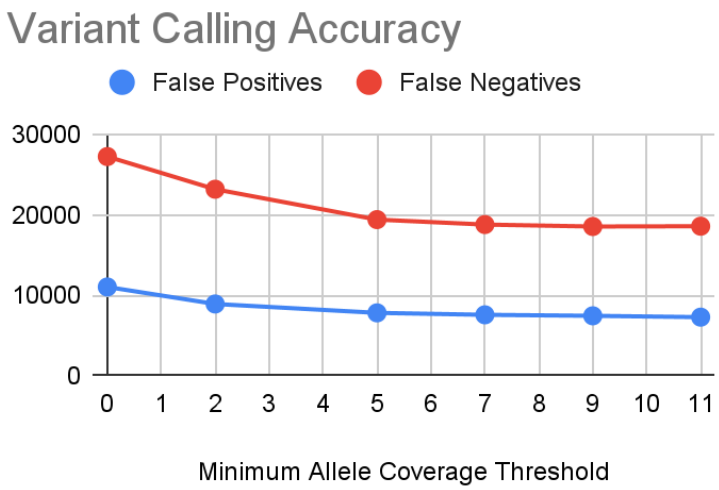
Supplementary Figure 6: ~30x Illumina short-reads for three GIAB samples (horizontal panels) were mapped using two approaches: vg Giraffe on the linear pangenomes with just the reference genome (greys), and vg Giraffe on the HPRC pangenome (reds). The left panels compare GRCh38-referenced pangenomes, the right panels compare CHM13-referenced pangenomes. The curves show the proportion of reads (y-axis) with an alignment score greater or equal to the threshold defined by the x-axis.



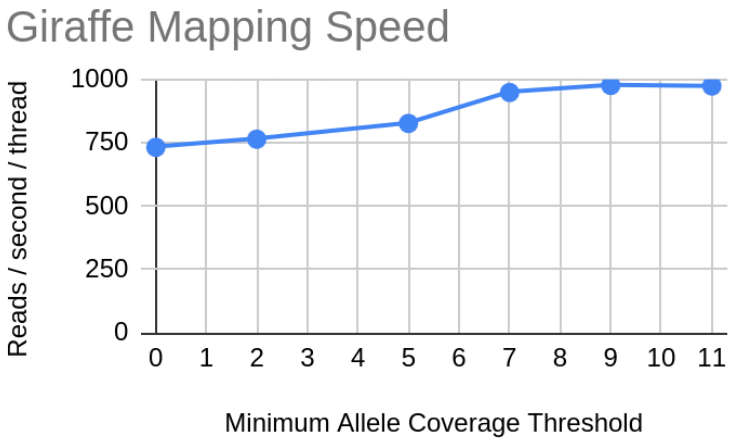
Supplementary Figure 7*:* ~30x Illumina short-reads for three GIAB samples were mapped using three approaches: BWAMEM on GRCh38 (blue), vg Giraffe on the linear pangenomes with GRCh38 or CHM13 (grey), vg Giraffe on the GRCh38-referenced or CHM13-referenced HPRC pangenomes (red and green). The darker redbar corresponds to the default GRCh38-based HPRC pangenome, while the lighter redto the frequency-filtered pangenome used in practice for read mapping and variant calling. A) Proportion of the reads aligning perfectly to the (pan-)genome for each sample (y-axis). B) Proportion of reads with a mapping quality greater than 0.



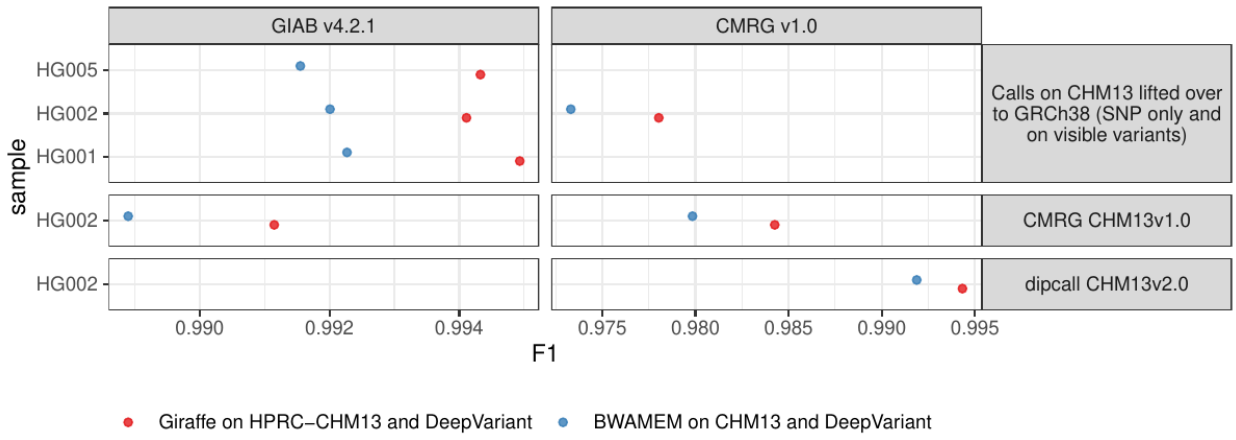
Supplementary Figure 8*:* Evaluation of calls made on both the default pangenome (light red) and the frequency-filtered pangenome (dark red). The results when aligning reads with BWAMEM (blue) or using the Dragen pipeline (green) are also shown. The F1 score is shown on the x-axis across samples from the Genome in a Bottle (y-axis). Left: Genome in a Bottle v4.2.2 truth set. Right: Challenging Medically Relevant Genes v1.0 truth set.



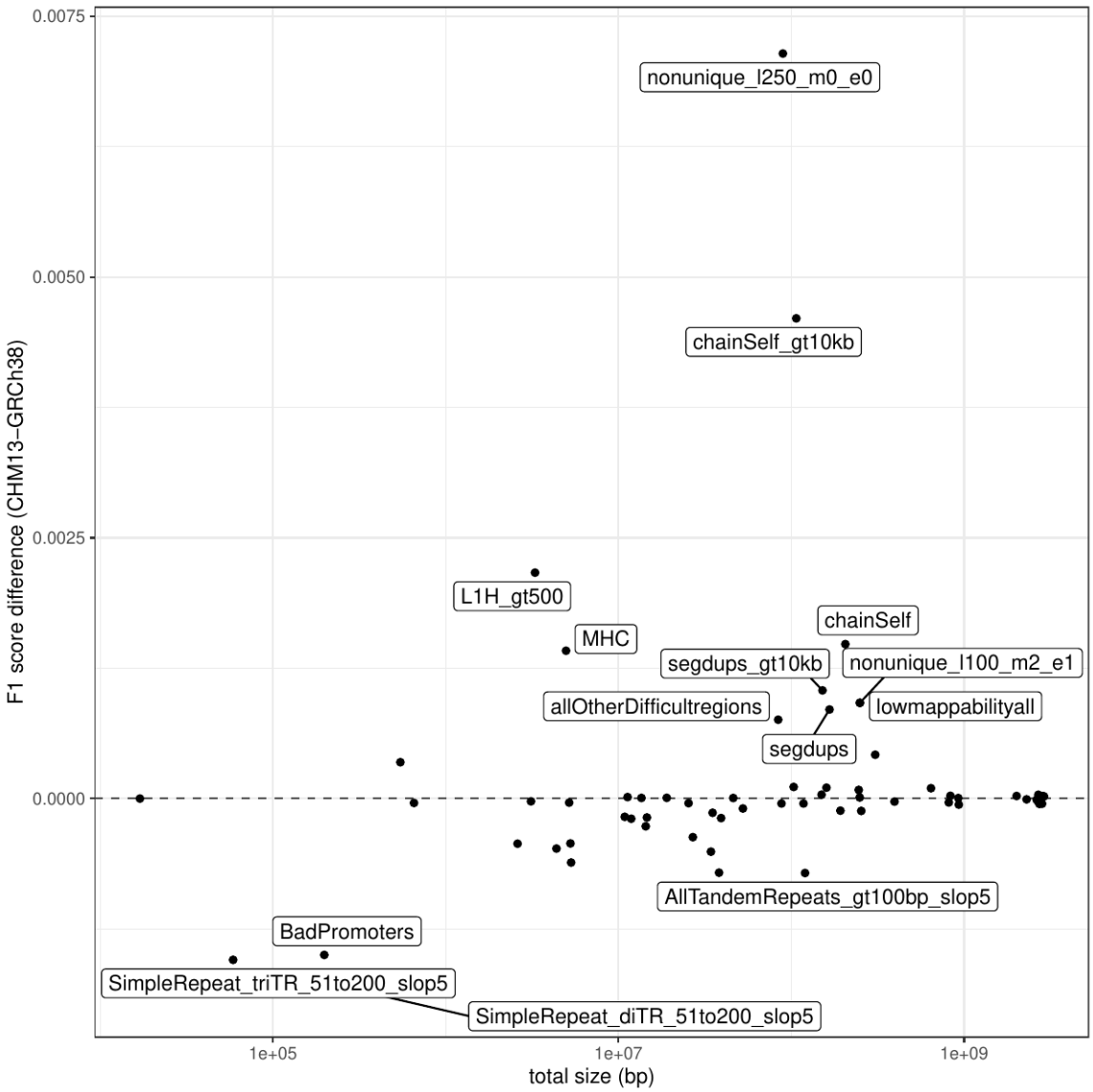
Supplementary Figure 9*:* Variant calling false positives and false negatives on 30X Genome in a Bottle v4.2.1 Illumina reads for HG003 for the CHM13-based pangenome as a function of the allele frequency filtering threshold used. The 0 column is the unfiltered graph and the 9 column is the 10% (9/10) filter used for all other short-read mapping experiments. The accuracy was measured using rtg vcfeval v3.91 on the evaluation regions provided by GIAB for this sample. The truth set has 3,831,915 calls total.



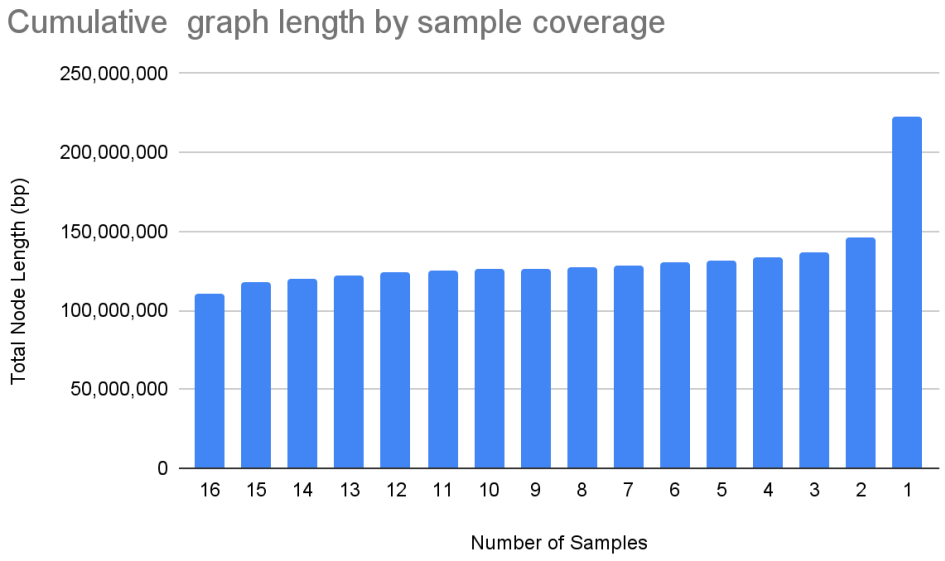
Supplementary Figure 10*:* Short read mapping speed on 30X Genome in a Bottle v4.2.1 Illumina reads for HG003 for the CHM13-based pangenome as a function of the allele frequency filtering threshold used, as reported by vg giraffe. The 0 column is the unfiltered graph and the 9 column is the 10% (9/90) filter used for all other short-read mapping experiments.



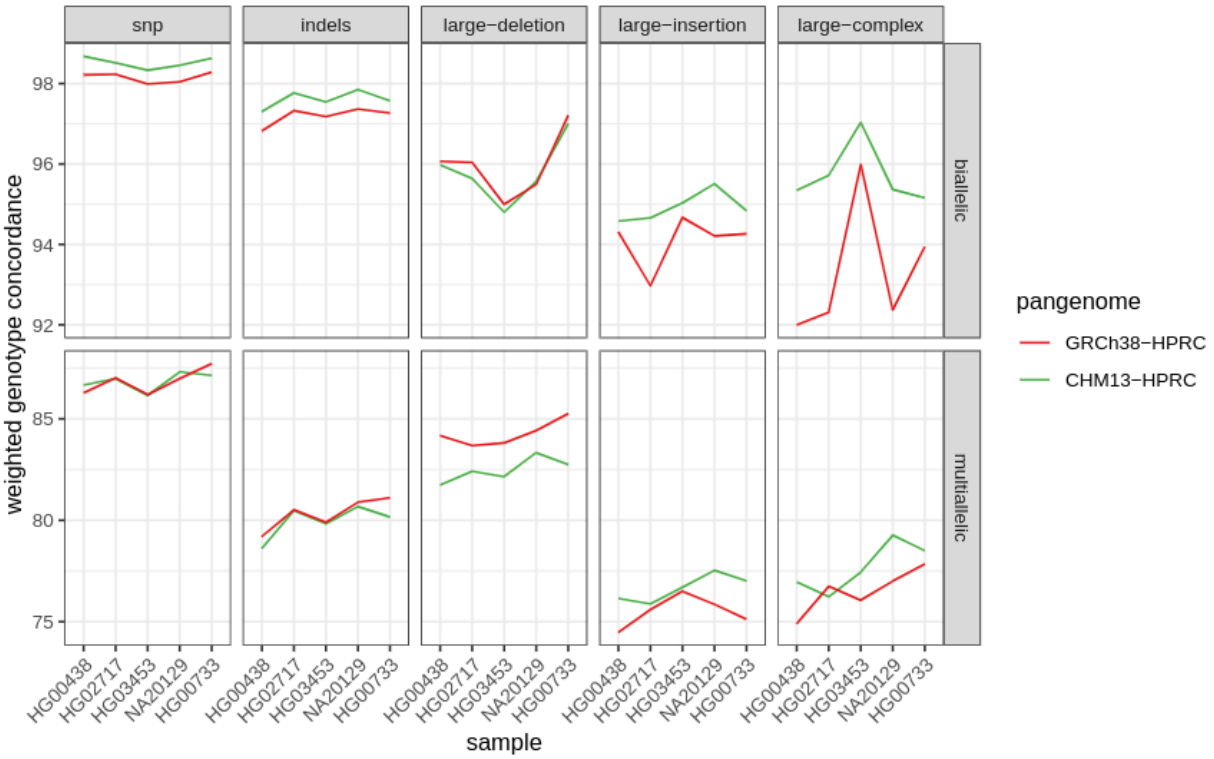
Supplementary Figure 11: Evaluation of calls made on CHM13: aligning reads with BWAMEM (blue), or to the CHM13-based HPRC pangenome and projecting them to CHM13 (red). The F1 score is shown on the x-axis across samples from the Genome in a Bottle (y-axis). Left: Genome in a Bottle v4.2.2 truth set. Right: Challenging Medically Relevant Genes v1.0 truth set. Three approaches are shown as horizontal panels. Top: variants called on CHM13 were lifted over to be evaluated against the GRCh38 truth sets. Only SNPs and variant that are visible (not homozygous for the reference allele) on both reference genomes were used. Middle: the CMRG truth set for CHM13 v1.0 was lifted to CHM13 v2.0. The whole genome evaluation (left) was limited to the GIAB v4.2.1 confident regions lifted from GRCh38 to CHM13. Bottom: Preliminary draft truth set for CHM13 v2.0 based on HiFi assemblies analyzed with dipcall.



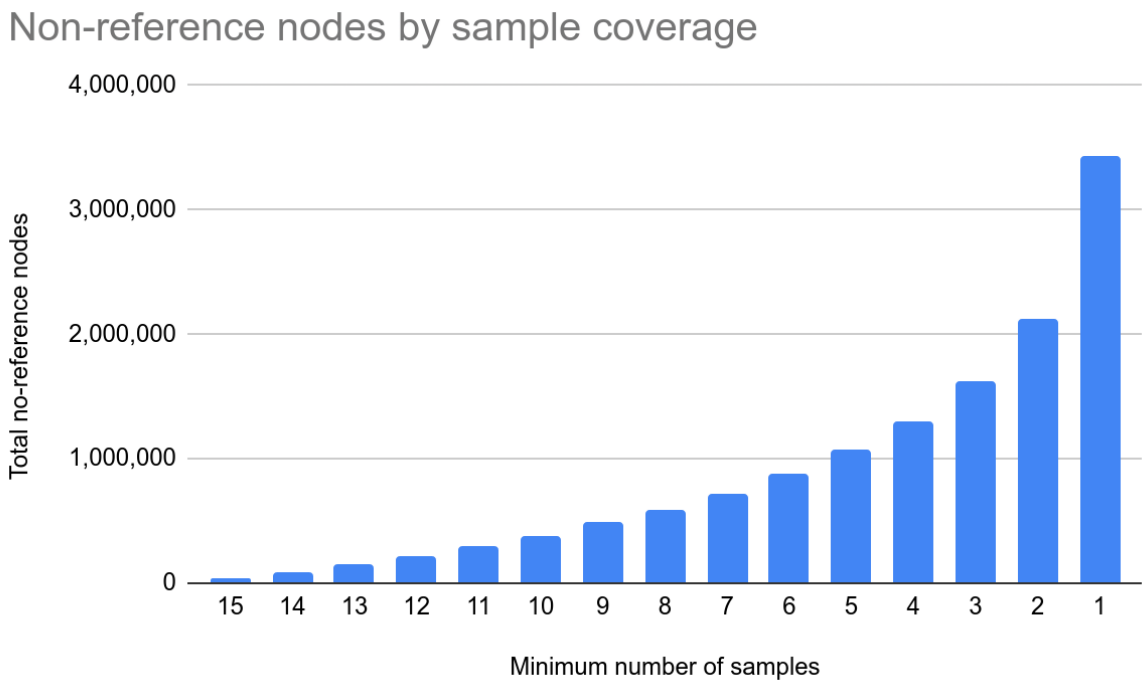
Supplementary Figure 12: Difference between the F1 score obtained when using the CHM13-based pangenome compared to the GRCh38-based pangenome (y-axis), stratified by region sets from the GIAB (points). The total amount of sequence that represents each region set is shown on the x-axis. The top 10 most regions with the largest differences are labeled.



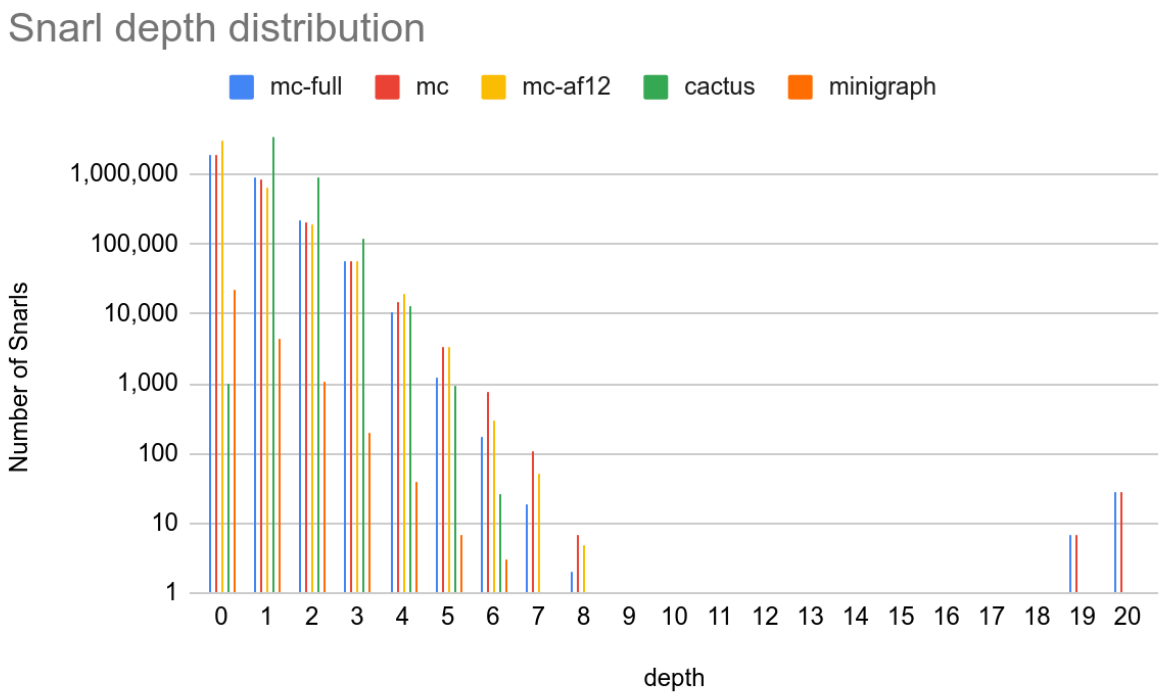
Supplementary Figure 13: The amount of sequence in the *D. melanogaster* graph by the minimum number of haplotypes that contain it.



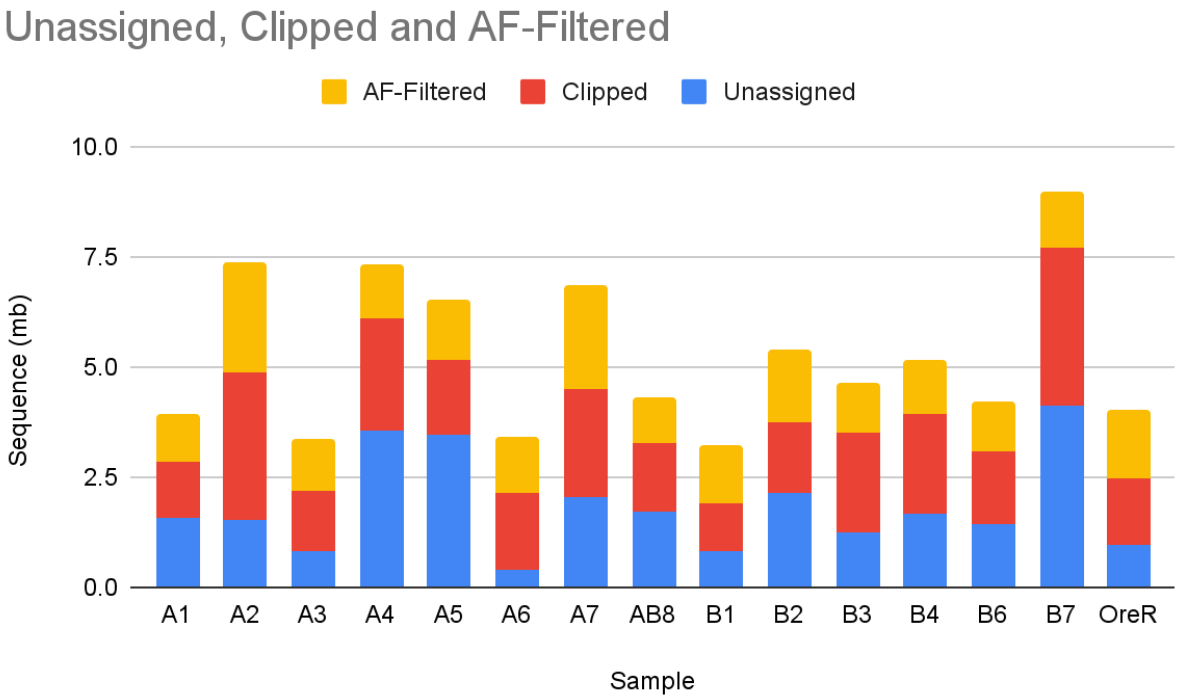
Supplementary Figure 14: Leave-one-out PanGenie experiments comparing the GRCh38 and CHM13-based HPRC graphs across the same five samples. Each of the samples was, in turn, removed along with all its private variants from the input VCF to PanGenie then genotyped from short reads. The Weighted Genotype Concordance was computed between the computed and original genotypes for each sample, excluding variants that were private (and therefore could not be re-genotyped). This is thus a comparison of the haplotypes as computed by Pangenie with those from the original assemblies within the context of the pangenome graph. Measurements are separated by variation category, as well as between bi-allelic and multiallelic sites in the graph.



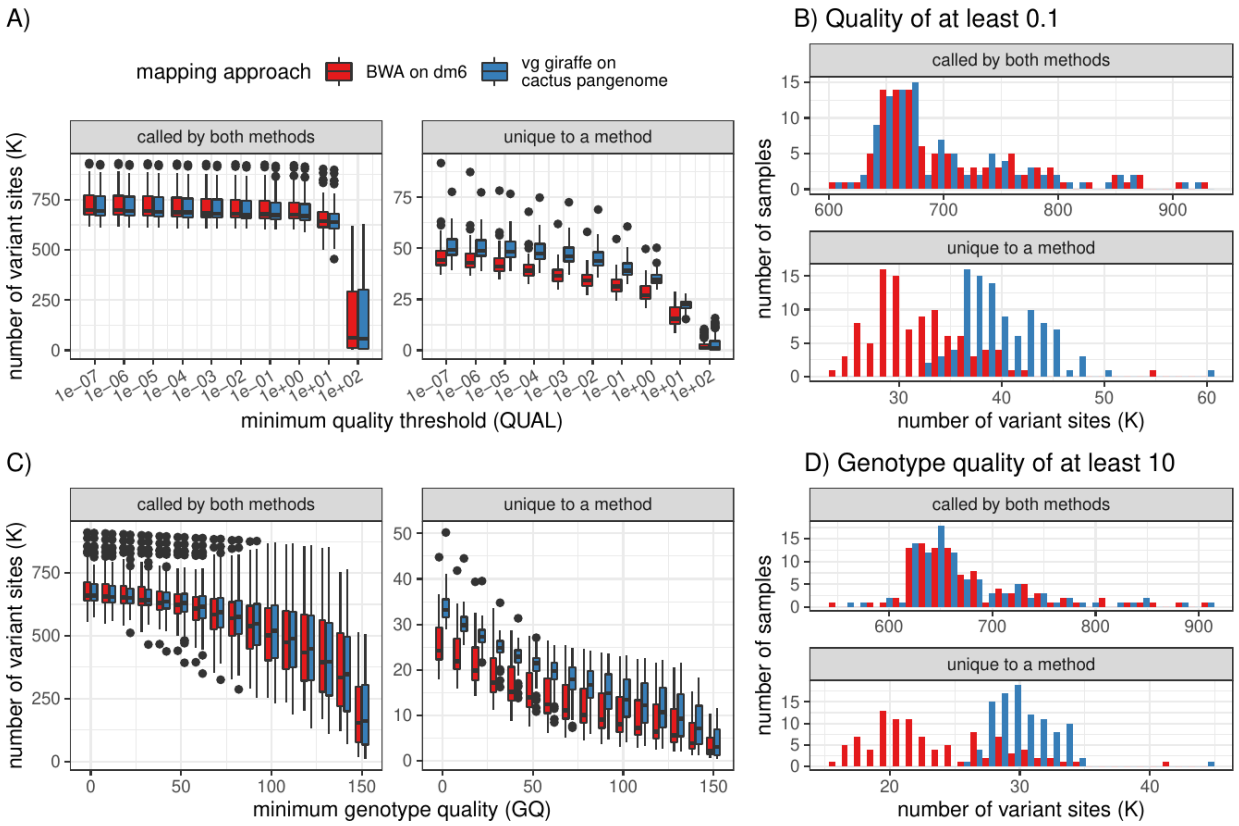
Supplementary Figure 15: The number of nodes not present in dm6 covered by at least the given number of samples.



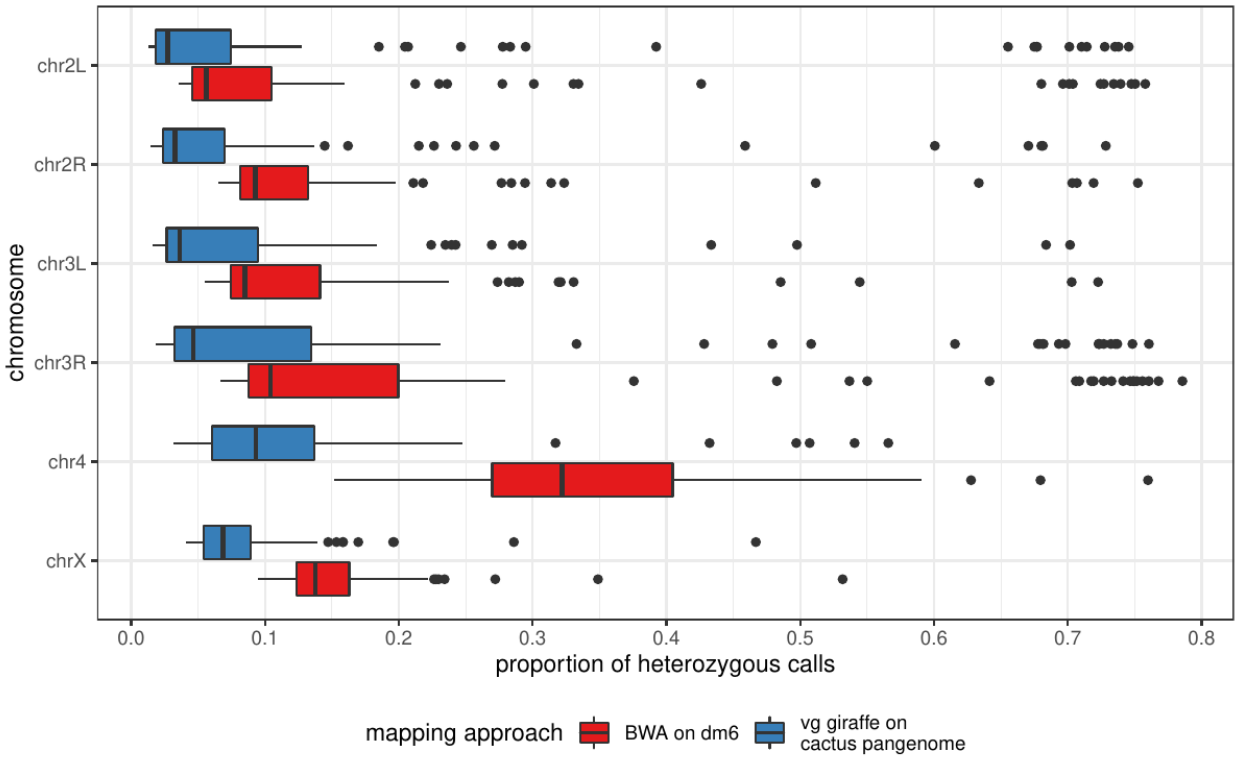
Supplementary Figure 16: Snarl depth distribution.



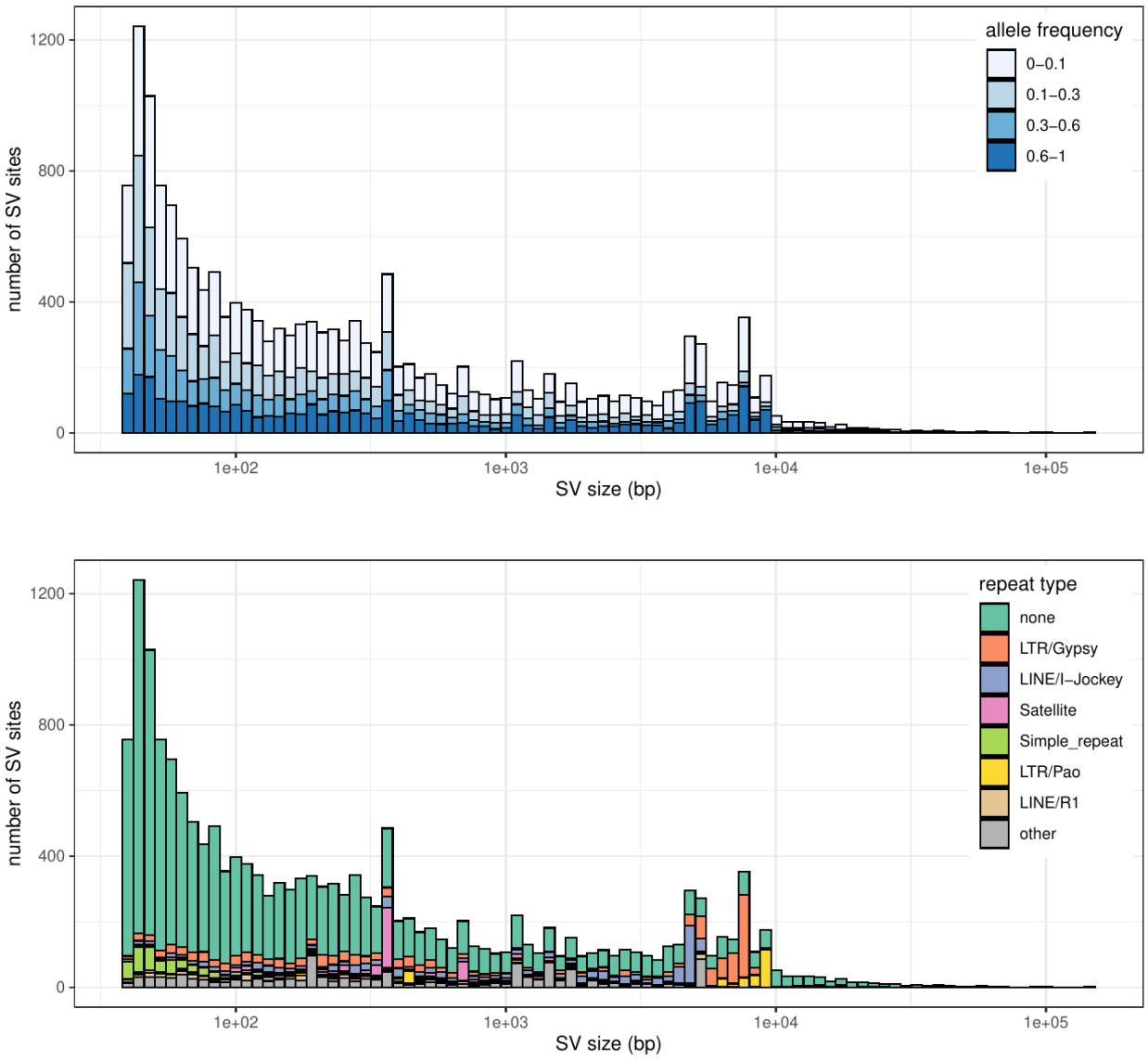
Supplementary Figure 17: Sequence excluded from the *D. melanogaster* pangenome.



Supplementary Figure 18: Number of variant sites with an alternate allele called in each of the 100 samples with FreeBayes. Two mapping approaches are compared: short-reads mapped to dm6 using BWA-MEM (red); short-reads mapped to the pangenome using vg Giraffe (blue). The variant sites were split into sites found by both approaches and sites found only by one. The boxplots in **A)** and **C)** show the median (center line), upper and lower quartiles (box limits), up to 1.5x interquartile range (whiskers), and outliers (points). **A)** Distribution of the number of variant sites for different minimum quality (QUAL field) (x-axis). **B)** Only variant sites with a quality of at least 0.1 were counted. This corresponds to x=0.1 in A). **C)** Distribution of the number of variant sites for different minimum genotype quality (GQ field) (x-axis). **D)** Only variant sites with a genotype quality of at least 10 were counted. This corresponds to x=10 in **C)**.



Supplementary Figure 19: Proportion of heterozygous small variants called by FreeBayes in each of the 100 fly samples (point). Reads were either aligned to the pangenome and projected to dm6 (blue), or mapped to dm6 with BWA-MEM (red). Due to the inbreeding of these lines, we expect low heterozygosity. The boxplots show the median (center line), upper and lower quartiles (box limits), up to 1.5x interquartile range (whiskers), and outliers (points).



Supplementary Figure 20: Distribution of the size of the SVs genotyped across 100 fly samples. The x-axis is log-scaled. **Top:** The SVs are colored by their allele frequencies. **Bottom:** The SVs are colored by the repeat class as annotated by Repeat Masker.

Graph	Nodes	Edges	Total Node Length	Total Non-ref Node Length	Total Path Length
SV Graph (GRCh38)	424,643	637,628	3,239,764,787	140,014,069	3,239,764,787
SV Graph (CHM13)	493,631	738,529	3,365,688,482	253,629,026	3,365,688,482
GRCh38-based Pangenome	81,751,614	113,258,931	3,287,932,785	188,182,067	254,821,009,311
GRCh38-based Filtered Pangenome (AF>=10%)	59,960,908	72,408,601	3,153,443,019	53,692,301	254,415,272,646
CHM13-based Pangenome	85,591,995	118,409,526	3,324,657,754	212,598,298	257,143,252,360
CHM13-based Filtered Pangenome (AF>=10%)	62,335,399	75,270,997	3,166,744,316	54,684,860	256,673,009,341

Supplementary Table 1: HPRC graph sizes. The total node length is the sum of the lengths of all nodes in the graph. The total “non-ref” node length is the sum of the lengths of all nodes that are not present in any reference path, i.e. excluding CHM13 paths from the CHM13-based graph and GRCh38 paths from the GRCh38-based graph. The total path length is the sum of all paths in the graph, which will correspond exactly to the total length of all contigs input into the construction procedure, minus any sequence clipped out or unassignable to a chromosome.

Graph	Nodes	Edges	Total Node Length	Total Non-ref Node Length	Total Path Length
SV Graph	80,853	112,742	214,547,800	71,326,590	214,547,800
Unclipped Pangenome	9,042,502	12,364,039	251,857,504	251,857,504	2,182,961,082
Pangenome	8,978,195	12,276,452	223,071,144	223,071,144	2,152,888,069
Filtered Pangenome (AF12.5%)	7,686,219	9,788,690	202,497,872	202,497,872	2,131,677,729
Progressive Cactus Graph	12,974,720	17,684,675	470,148,493	470,148,493	2,216,588,031

Supplementary Table 2: *D. Melegonaster* graph sizes. The total node length is the sum of the lengths of all nodes in the graph. The total “non-ref” node length is the sum of the lengths of all nodes that are not present in any dm6 path. The total path length is the sum of all paths in the graph, which will correspond exactly to the total length of all contigs input into the construction procedure, minus any sequence clipped out or unassignable to a chromosome.

Variant type	HPRC-CHM13			HPRC-GRCh38		
	# alleles in biallelic regions	**# alleles in multialleli c regions**	**# alleles total**	**# alleles in biallelic regions**	**# alleles in multialleli c regions**	**# alleles total**
SNPs	18,489,918	1,997,181	**20,487,099**	18,392,518	1,801,599	**20,194,117**
Indels	2,248,578	4,731,607	**6,980,185**	2,250,931	4,597,184	**6,848,115**
SV-DEL	4,270	66,440	**70,710**	4,232	52,969	**57,201**
SV-INS	12,384	187,443	**199,827**	12,192	242,420	**254,612**
SV-Others	1,382	90,766	**92,148**	1,296	100,700	**101,996**

Variant type	HPRC-CHM13		
	# alleles in biallelic regions	**# alleles in multialleli c regions**	**# alleles total**
SNPs	18,489,918	1,997,181	**20,487,099**
Indels	2,248,578	4,731,607	**6,980,185**
SV-DEL	4,270	66,440	**70,710**
SV-INS	12,384	187,443	**199,827**
SV-Others	1,382	90,766	**92,148**

Supplementary Table 3: Number of variants in graphs genotyped by PanGenie. These statistics are obtained from the VCF after preprocessing by PanGenie. Indels are small events

Phase	Dm6 (hours)	hprc-v1.0 grch38 (hours)	hprc-v1.0 chm13 (hours)	hprc grch38 (new pipeline) (hours)	hprc chm13 (new pipeline) (hours)
Minigraph construction	3.02	45.43	39.82	45.82	21.4
Minigraph mapping	1.01	3.14	4.52	1.93	1.58
Split by chromosome	0.11	1.6	2.1	1.67	1.43
Cactus alignment	0.43	11.56	7.66	5.04	5.27
Indexing and clipping (full graph)	0.13	N/A	N/A	4.37	4.68
Indexing and clipping (clipped graph)	0.26	10.02	10.98	3.17	4.32
Indexing and clipping (AF12.5% graph)	0.16	10.81*	10.75*	3.42	4.23
Total	5.12	71.75	75.83	65.42	42.91

Supplementary Table 4: Minigraph-Cactus running times (wall-times). The “new pipeline” columns refer to graphs made using the method described here which does not rely on dna-brnn for clipping. The dm6 graphs were made using up to 32 cores and 16Gb RAM. The HPRC graphs were made on an AWS cluster using up to 25 32 core 256Gb RAM machines, except for the indexing stages which were done on up to 2 64 core 512Gb RAM machines. The disk usage of each step is bounded by the total size of the input and output (plus uncompressed versions of the same if they are gzipped).

** These values were not kept in the logs and were estimated using the ratios in the neighboring columns (ex 10.81 = 3.42/3.17 * 10.02).*

Phase	Dm6 (hours)
Lastz repeatmasking	0.38
All-to-all lastz alignment	17.97
Cactus alignment	0.83
Total	19.18

Supplementary Table 5: Progressive Cactus running times (wall times) using single 32-core machine with up to 64Gb RAM.

DGRP Line	Sequencing Technology	Freeze	Mapped Coverage	Raw Read Length:Read Number	NCBI SRA	NCBI SRR
DGRP_21	Illumina	F1	15.8	95bp:37046984	SRX021040	SRR834526
DGRP_31	Illumina	F2	49.2	125bp:76894692	SRX155996	SRR834509
DGRP_32	Illumina	F2	56.2	125bp:88154526	SRX155997	SRR834512
DGRP_38	Illumina	F1	28.0	95bp:56154204	SRX025317	SRR834541
DGRP_40	Illumina	F1	33.3	95bp:69063428	SRX021235	SRR835025
DGRP_42	Illumina	F1	20.2	95bp:37186556	SRX021255	SRR835027
DGRP_48	Illumina	F2	32.7	125bp:58419132	SRX155989	SRR835034
DGRP_49	Illumina	F1	15.2	75bp:37870818	SRX021267	SRR835037
DGRP_57	Illumina	F1	32.6	100bp:64966990	SRX021296	SRR933581
DGRP_75	Illumina	F1	18.5	110bp:38161744	SRX021384	SRR835087
DGRP_83	Illumina	F1	16.3	75bp:41070470	SRX023456	SRR835058
DGRP_100	Illumina	F2	52.3	125bp:87340978	SRX156026	SRR833244
DGRP_138	Illumina	F1	30.1	100bp:61689820	SRX021008	SRR932121
DGRP_142	Illumina	F1	19.7	110bp:41167794	SRX020759	SRR834551
DGRP_177	Illumina	F1	24.6	95bp:49114764	SRX021026	SRR834547
DGRP_181	Illumina	F1	24.7	75bp:64093862	SRX020912	SRR933563
DGRP_189	Illumina	F2	37.8	125bp:63289120	SRX155979	SRR834523
DGRP_223	Illumina	F2	40.8	125bp:71152512	SRX155994	SRR834527
DGRP_235	Illumina	F1	18.4	95bp:38296004	SRX021053	SRR834531
DGRP_318	Illumina	F1	15.2	75bp:39068236	SRX021082	SRR834507
DGRP_319	Illumina	F2	37.6	125bp:70621686	SRX155981	SRR834508
DGRP_320	Illumina	F1	24.2	95bp:51875680	SRX021063	SRR834510
DGRP_321	Illumina	F1	33.5	95bp:67314152	SRX021094	SRR834511
DGRP_332	Illumina	F1	25.7	75bp:65583082	SRX021095	SRR933569
DGRP_348	Illumina	F2	48.3	125bp:78515972	SRX156029	SRR834514
DGRP_352	Illumina	F1	15.6	75bp:44982388	SRX021101	SRR834516
DGRP_354	Illumina	F2	57.2	101bp:106369344	SRX156027	SRR834517
DGRP_355	Illumina	F2	44.9	101bp:84541222	SRX156028	SRR834545
DGRP_356	Illumina	F1	15.5	75bp:42903612	SRX023833	SRR834537
DGRP_359	Illumina	F1	20.2	95bp:37271884	SRX023424	SRR834546

DGRP_361	Illumina	F2	40.6	125bp:68254340	SRX155984	SRR834553
DGRP_370	Illumina	F1	20.9	95bp:43793604	SRX021104	SRR834539
DGRP_377	Illumina	F1	21.8	95bp:43796182	SRX023834	SRR834543
DGRP_381	Illumina	F1	20.9	75bp:54335852	SRX021112	SRR933573
DGRP_382	Illumina	F2	41.1	125bp:73812254	SRX156013	SRR834552
DGRP_383	Illumina	F1	19.1	95bp:39897030	SRX021113	SRR834554
DGRP_390	Illumina	F2	26.2	125bp:42709922	SRX156014	SRR834519
DGRP_392	Illumina	F1	23.2	95bp:51156860	SRX021157	SRR834520
DGRP_395	Illumina	F2	47.1	101bp:87233368	SRX156015	SRR834521
DGRP_397	Illumina	F2	30.0	125bp:48910026	SRX156017	SRR834522
DGRP_405	Illumina	F1	22.9	95bp:50080536	SRX021242	SRR835023
DGRP_406	Illumina	F1	25.0	95bp:51821248	SRX021254	SRR835024
DGRP_426	Illumina	F1	21.1	95bp:43746634	SRX021245	SRR835026
DGRP_427	Illumina	F1	16.3	45bp:64106936	SRX006155	SRR933577
DGRP_439	Illumina	F1	20.4	95bp:44762436	SRX021244	SRR835028
DGRP_440	Illumina	F1	17.2	95bp:43161850	SRX021246	SRR835029
DGRP_441	Illumina	F1	18.7	95bp:42278010	SRX023835	SRR835030
DGRP_443	Illumina	F1	28.5	95bp:57567568	SRX021260	SRR835031
DGRP_461	Illumina	F1	21.9	95bp:49324528	SRX021262	SRR835033
DGRP_491	Illumina	F1	15.1	75bp:40944392	SRX021268	SRR835035
DGRP_492	Illumina	F1	22.1	95bp:44580310	SRX021270	SRR835036
DGRP_502	Illumina	F1	21.7	95bp:44336646	SRX021271	SRR835038
DGRP_505	Illumina	F2	43.7	125bp:71295212	SRX156002	SRR835039
DGRP_508	Illumina	F1	21.2	95bp:42338556	SRX021272	SRR835040
DGRP_509	Illumina	F1	15.3	75bp:38095912	SRX021273	SRR835041
DGRP_513	Illumina	F1	19.6	95bp:42640722	SRX021282	SRR835042
DGRP_528	Illumina	F2	36.2	125bp:57697778	SRX155985	SRR835043
DGRP_530	Illumina	F2	20.7	125bp:34726088	SRX156031	SRR835044
DGRP_531	Illumina	F1	17.9	95bp:41560152	SRX021290	SRR835045
DGRP_535	Illumina	F1	15.2	75bp:40234802	SRX021293	SRR835046
DGRP_551	Illumina	F2	21.4	125bp:35225968	SRX156034	SRR835047
DGRP_555	Illumina	F1	19.2	75bp:50103810	SRX006159	SRR933580
DGRP_559	Illumina	F2	24.2	125bp:36482062	SRX156032	SRR835048
DGRP_566	Illumina	F2	48.8	101bp:89414580	SRX156033	SRR835050

DGRP_596	Illumina	F2	41.1	101bp:73915046	SRX156004	SRR835096
DGRP_627	Illumina	F2	36.7	125bp:82297368	SRX155988	SRR835097
DGRP_630	Illumina	F2	21.7	125bp:36162916	SRX156003	SRR835098
DGRP_634	Illumina	F2	19.4	125bp:32632568	SRX156018	SRR835086
DGRP_705	Illumina	F1	16.7	75bp:47006608	SRX006162	SRR933585
DGRP_707	Illumina	F1	17.8	75bp:46657404	SRX006163	SRR933586
DGRP_712	Illumina	F1	16.3	75bp:44687868	SRX006164	SRR933587
DGRP_727	Illumina	F1	27.5	75bp:73781476	SRX021382	SRR933589
DGRP_732	Illumina	F1	16.3	75bp:42170344	SRX006167	SRR933591
DGRP_737	Illumina	F1	25.1	75bp:74740132	SRX023451	SRR933592
DGRP_738	Illumina	F1	27.1	75bp:75804508	SRX021383	SRR933593
DGRP_757	Illumina	F1	28.4	75bp:74326240	SRX021385	SRR933594
DGRP_761	Illumina	F1	15.2	75bp:40867250	SRX021386	SRR835088
DGRP_776	Illumina	F1	15.6	75bp:39890986	SRX021387	SRR835089
DGRP_787	Illumina	F1	15.4	75bp:39795416	SRX021388	SRR835091
DGRP_790	Illumina	F1	17.0	95bp:35620658	SRX021389	SRR835092
DGRP_805	Illumina	F1	16.1	75bp:43182102	SRX021400	SRR835095
DGRP_810	Illumina	F1	15.5	75bp:36972402	SRX021418	SRR835051
DGRP_812	Illumina	F1	16.1	75bp:38719004	SRX021419	SRR835052
DGRP_819	Illumina	F2	73.0	100bp:150745358	SRX156006	SRR835054
DGRP_822	Illumina	F1	17.7	110bp:41079524	SRX021476	SRR835055
DGRP_837	Illumina	F1	20.7	95bp:46411538	SRX021479	SRR933599
DGRP_843	Illumina	F2	42.3	125bp:68658714	SRX156036	SRR835059
DGRP_849	Illumina	F2	39.9	125bp:61687178	SRX156035	SRR835060
DGRP_850	Illumina	F2	43.6	125bp:69699750	SRX155993	SRR835061
DGRP_855	Illumina	F1	19.2	110bp:42348166	SRX021563	SRR835062
DGRP_857	Illumina	F1	20.8	110bp:42340250	SRX021492	SRR835063
DGRP_882	Illumina	F1	17.4	75bp:44722234	SRX021496	SRR835067
DGRP_887	Illumina	F1	19.5	95bp:43595728	SRX021527	SRR835069
DGRP_890	Illumina	F1	15.9	75bp:41954706	SRX021499	SRR835071
DGRP_892	Illumina	F1	20.5	95bp:45702226	SRX023838	SRR835072
DGRP_894	Illumina	F1	16.8	95bp:35128536	SRX021528	SRR835073
DGRP_897	Illumina	F1	27.0	75bp:70892788	SRX023457	SRR933601
DGRP_907	Illumina	F1	17.5	95bp:36385056	SRX021500	SRR835074

DGRP_908	Illumina	F1	19.9	95bp:39111536	SRX021501	SRR835075
DGRP_913	Illumina	F2	43.7	125bp:69250292	SRX156024	SRR835077

Supplementary Table 6: DGRP sequencing data used for *D. Melanogaster* mapping and variant calling experiments