Supplementary Information: Versatile genome assembly evaluation with QUAST-LG

Alla Mikheenko[1], Andrey Prjibelski[1], Vladislav Saveliev[1], Dmitry Antipov[1], and Alexey Gurevich[1]

[1]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia

Supplementary Tables

Supplementary Table 1: QUAST and QUAST-LG performance. The four compared modes are: *QUAST* (the default QUAST v4.5 distribution), *adj.QUAST* (QUAST v4.5 with the adjusted parameters, so the tool uses the same minimal contig, minimal alignment, and extensive misassembly thresholds as in QUAST-LG), *QLG-NUCmer* (QUAST-LG with Minimap2 [1] aligner replaced by NUCmer aligner from MUMmer v3.23 [2] which was used in all versions of QUAST before v5.0), *QUAST-LG* (the default QUAST-LG v5.0 distribution). The running time for the latter is given separately for *Old stats* (quality metrics available both in QUAST and QUAST-LG), *New stats* (novel features and quality metrics added in QUAST-LG, that is, k-mer-based statistics and upper bound assembly generation), and *BUSCO* (search for conservative single-copy orthologs using BUSCO [3]); the rest three modes include only *Old stats* for a fair comparison with conventional QUAST software. *BUSCO* is shown separately since this module is intended for reference-free evaluation and should not be normally run with referenced-based *Old* and *New stats*. *Num* stands for the number of assemblies being processed. Note, that in addition to input assemblies QUAST-LG computes and evaluates the upper bound assembly which was provided to QUAST as one more input assembly for a fair comparison. All running *times* are in hh:mm format, maximal *RAM* consumption is in GB (computed for three modes only), "—" indicates the fact that the NUCmer-based tools were not able to process the human datasets in a reasonable time. All benchmarking was done on a server with Intel Xeon X7560 2.27GHz CPUs using 8 threads.

Dataset	Genome	Num	*QUAST*	*adj.*		*QLG-*	*QUAST-LG*				
	size			*QUAST*		*NUCmer*	Old stats		New stats	BUSCO	RAM
	(Mb)		Time	Time	RAM	Time	Time	RAM			(total)
Yeast$_{PB}$	12.1	6	00:09	00:06	1.2	00:01	00:01	1.1	00:39	00:35	6.2
Yeast$_{NP}$	12.1	5	00:08	00:04	1.2	00:01	00:01	0.6	00:58	00:44	8.5
Worm$_{PB}$	100.3	6	22:31	02:51	8.4	02:40	00:08	6.3	04:28	00:58	32.3
Fly$_{MP}$	137.6	7	71:34	04:55	13.8	03:17	00:21	9.8	05:01	00:58	27.2
Human$_{MP}$	3,088.3	4	—	—	—	—	03:55	135.2	24:27	13:21	184.1
Human$_{NP}$	3,088.3	4	—	—	—	—	04:05	135.4	32:01	09:55	178.4

Supplementary Table 2: Comparison of QUAST-LG alignment metrics computed using Minimap2 and NUCmer. Each cell represents the average value of the corresponding alignment metric among all input assemblies (upper bound assembly is not counted). *GF (%)* stands for Genome fraction in percents, *Dupl.* is for Duplication ratio, *Largest al.* and *Total len.* are for the largest alignment length and the total number of aligned bases, respectively. *# mis.* (*# local mis.*) is for the number of extensive (local) misassemblies. *Mis. len.* stands for the total length of the contigs containing at least one extensive misassembly. *# TEs* is the number of misassembly events probably caused by transposable elements (not counted against *# mis.*). *MM* and *IND* stand for the number of mismatches and indels per 100 kb, respectively. *# unal.* and *Unal. len.* is the number of contigs that have no alignment to the reference sequence and their total length, respectively.

Dataset	Yeast$_{PB}$		Yeast$_{NP}$		Worm$_{PB}$	
Aligner	Minimap2	NUCmer	Minimap2	NUCmer	Minimap2	NUCmer
GF (%)	97.52	97.50	98.58	98.63	99.22	99.34
Dupl.	1.033	1.036	1.008	1.009	1.016	1.022
Largest al. (kb)	1256	1265	1162	1200	3032	3181
Total len. (Mb)	12.385	12.385	12.134	11.985	105.959	105.959
NGA50 (kb)	610	644	685	670	1201	1160
NGA75 (kb)	370	420	457	451	718	681
LGA50	8.4	8.2	8.25	7.5	28.6	30.2
LGA75	15.6	14.4	12.75	13.0	55.8	58.8
# mis.	34.6	28.4	9.5	9.25	152.6	238.0
# local mis.	48	65.6	14.25	27.5	423.2	1045.4
Mis. len. (Mb)	6263	5654	2871	2973	81315	83207
# TEs	8.8	21	3.5	2.3	80.0	136.0
MM	344	252	172	142	43	26
IND	65	64	377	361	66	58
# unal.	2.0	2.4	1.0	1.5	24.8	21.2
Unal. len. (Mb)	147	109	55	39	4872	4170

Supplementary Table 3: QUAST-LG report on the Yeast$_{PB}$ dataset. All statistics are given for scaffolds $\geq$ 3 kb. The best value for each column is indicated in bold.

Assembly	UpperBound	Canu	FALCON	Flye	MaSuRCA	Miniasm
# contigs	38	32	77	29	51	49
Largest contig	1524479	**1534530**	1527374	1084210	857809	1525027
Total length	12163350	12482519	12283154	12205282	**12571691**	12382136
Reference length	12157105	12157105	12157105	12157105	12157105	12157105
GC (%)	38.15	38.21	38.46	38.17	38.23	38.15
Reference GC (%)	38.15	38.15	38.15	38.15	38.15	38.15
N50	**776910**	776810	762979	776728	432306	737373
NG50	**776910**	776810	762979	776728	432306	737373
N75	564006	**564467**	465562	556525	229910	448848
NG75	564006	**564467**	465562	564435	271502	467749
L50	**6**	**6**	**6**	7	11	7
LG50	**6**	**6**	**6**	7	11	7
L75	**11**	**11**	**11**	12	21	12
LG75	**11**	**11**	**11**	**11**	19	**11**
# misassemblies	**0**	35	19	24	60	35
# misassembled contigs	**0**	18	11	10	33	16
Misassembled contigs length	**0**	6548228	4962794	6938326	8168042	4698174
# local misassemblies	**0**	52	38	35	72	43
# scaffold gap size mis.	**0**	**0**	**0**	**0**	1	**0**
# possible MGEs	0	8	16	6	4	10
# unaligned mis. contigs	**0**	**0**	**0**	**0**	**0**	1
# unaligned contigs	**0+0p**	1+20p	4+12p	5+12p	0+40p	0+28p
Unaligned length	**0**	118761	152915	171056	136267	154665
Genome fraction (%)	**99.923**	98.770	96.074	98.040	97.413	97.307
Duplication ratio	**1.001**	1.030	1.039	1.010	1.050	1.034
# N's per 100 kbp	**0.00**	**0.00**	**0.00**	0.82	26.93	**0.00**
# mm per 100 kbp	**0.00**	579.50	184.09	118.27	680.43	155.74
# indels per 100 kbp	**0.00**	48.25	92.09	30.23	50.04	104.48
Complete BUSCO (%)	**99.31**	**99.31**	88.28	**99.31**	**99.31**	89.31
Partial BUSCO (%)	0.00	0.00	6.90	0.00	0.00	**8.62**
Largest alignment	**1524479**	1511901	1501819	1083357	686084	1511718
Total aligned len	12163350	12339852	12117327	12017546	**12399677**	12218762
NA50	**776910**	668909	694355	676772	345836	663236
NGA50	**776910**	668909	694355	676772	345836	663236
NA75	**564006**	428050	390819	429820	179383	376178
NGA75	**564006**	428050	390819	429820	179514	419810
LA50	**6**	7	7	7	14	7
LGA50	**6**	7	7	7	14	7
LA75	**11**	13	13	13	27	14
LGA75	**11**	13	13	13	26	13
K-mer-based completeness	**99.90**	64.39	86.31	91.72	62.36	85.46
K-mer-based cor. length (%)	**99.39**	84.57	90.42	71.44	66.97	78.57
K-mer-based mis. length (%)	**0.00**	14.25	6.01	27.88	32.14	19.80
# k-mer-based misjoins	**0**	3	3	7	10	3

Supplementary Table 4: QUAST-LG report on the Yeast$_{NP}$ dataset. All statistics are given for scaffolds $\geq$ 3 kb. The best value for each column is indicated in bold.

Assembly	UpperBound	Canu	Flye	MaSuRCA	Miniasm
# contigs	42	35	29	24	31
Largest contig	1524479	1090297	1080635	**1546094**	1056583
Total length	12170018	12264084	11963777	**12324198**	11985552
Reference length	12157105	12157105	12157105	12157105	12157105
GC (%)	38.15	38.31	38.33	38.14	38.61
Reference GC (%)	38.15	38.15	38.15	38.15	38.15
N50	776910	783642	782882	**813521**	659164
NG50	776910	783642	782882	**813521**	659164
N75	**564006**	449401	463994	541850	451115
NG75	**564006**	449401	463994	541850	451115
L50	**6**	7	7	**6**	8
LG50	**6**	7	7	**6**	8
L75	11	12	12	**10**	13
LG75	11	12	12	**10**	13
# misassemblies	**0**	12	5	14	7
# misassembled contigs	**0**	10	3	8	5
Misassembled contigs len	**0**	2037150	1834995	5269173	2344082
# local misassemblies	**0**	22	9	10	16
# scaffold gap size mis.	0	0	0	0	0
# possible MGEs	0	4	2	4	4
# unaligned mis. contigs	**0**	**0**	1	**0**	1
# unaligned contigs	**0+0p**	1+14p	1+14p	2+2p	0+12p
Unaligned length	**0**	56584	66963	60723	34934
Genome fraction (%)	**99.869**	98.843	97.708	99.518	98.261
Duplication ratio	1.002	1.016	1.002	1.014	**1.000**
# N's per 100 kbp	**0.00**	**0.00**	1.67	**0.00**	**0.00**
# mm per 100 kbp	**0.00**	565.83	56.14	12.39	52.60
# indels per 100 kbp	**0.00**	101.41	649.09	2.87	754.08
Complete BUSCO (%)	**99.31**	97.93	34.14	**99.31**	34.48
Partial BUSCO (%)	0.00	1.03	**33.79**	0.00	32.41
Largest alignment	**1524479**	1089592	1080623	1521997	1056509
Total aligned len	12170018	12183658	11879194	**12244032**	11936970
NA50	**776910**	657536	663200	740331	638568
NGA50	776910	657536	663200	**782448**	638568
NA75	**564006**	447675	461922	459637	441933
NGA75	**564006**	447675	461922	478274	441933
LA50	**6**	7	8	7	8
LGA50	**6**	7	8	**6**	8
LA75	**11**	13	13	12	14
LGA75	**11**	13	13	**11**	14
K-mer-based compl.	**99.94**	62.04	52.09	99.12	48.08
K-mer-based cor. len (%)	**99.26**	96.44	85.41	74.93	94.49
K-mer-based mis. len (%)	**0.00**	2.73	14.56	24.61	5.09
# k-mer-based misjoins	**0**	1	3	5	1

Supplementary Table 5: QUAST-LG report on the Worm$_{PB}$ dataset. All statistics are given for scaffolds $\geq$ 3 kb. The best value for each column is indicated in bold.

Assembly	UpperBound	Canu	FALCON	Flye	MaSuRCA	Miniasm
# contigs	62	104	96	93	189	222
Largest contig	**12666685**	6800719	5092131	10667848	3938208	5310534
Total length	100290985	107035688	100867711	102947220	107273906	**111671537**
Reference length	100286401	100286401	100286401	100286401	100286401	100286401
GC (%)	35.44	35.92	35.45	35.55	36.09	36.11
Reference GC (%)	35.44	35.44	35.44	35.44	35.44	35.44
N50	**3507402**	3187530	2013998	2275506	1393052	2056353
NG50	3507402	**3634244**	2013998	2321891	1435395	2105818
N75	**1884483**	1807374	1201702	1598883	836666	1368386
NG75	1884483	**1931153**	1201702	1629564	946610	1629523
L50	**8**	12	17	15	26	19
LG50	**8**	11	17	14	24	16
L75	**18**	24	32	28	51	35
LG75	**18**	21	32	27	45	29
# misassemblies	**0**	147	94	122	138	262
# misassembled contigs	**0**	46	40	43	71	94
Misassembled contigs len	**0**	98959366	71628774	79289587	61136699	95559931
# local misassemblies	**0**	610	321	358	328	499
# scaffold gap size mis.	**0**	**0**	**0**	1	3	**0**
# possible MGEs	0	78	80	76	78	88
# unaligned mis. contigs	**0**	9	**0**	3	4	32
# unaligned contigs	**0+0p**	29+74p	12+74p	19+58p	16+146p	48+148p
Unaligned length	**0**	6035396	1430131	2157004	5467532	9272460
Genome fraction (%)	**99.951**	99.541	98.670	99.312	99.179	99.413
Duplication ratio	**1.001**	1.012	1.005	1.012	1.024	1.027
# N's per 100 kbp	**0.00**	**0.00**	**0.00**	1.65	141.18	**0.00**
# mm per 100 kbp	**0.00**	41.18	65.11	19.77	33.28	54.47
# indels per 100 kbp	**0.00**	7.29	126.13	43.50	7.79	143.88
Complete BUSCO (%)	**96.37**	**96.37**	88.78	**96.37**	**96.37**	95.05
Partial BUSCO (%)	0.00	0.00	**5.61**	0.00	0.00	0.99
Largest alignment	**12666685**	3373829	3052157	3354145	2541679	2839481
Total aligned len	100290985	100835739	99369894	100735702	101477301	**102229046**
NA50	**3507402**	1226858	1176205	1297968	972567	979636
NGA50	**3507402**	1292248	1176205	1305538	1016420	1214817
NA75	**1884483**	669558	679770	707741	534593	528888
NGA75	**1884483**	766486	697542	789608	642715	692975
LA50	**8**	30	29	27	36	34
LGA50	**8**	27	29	26	32	29
LA75	**18**	58	57	53	73	72
LGA75	**18**	51	56	50	64	58
K-mer-based compl.	**99.96**	99.09	88.94	95.23	97.45	87.41
K-mer-based cor. len (%)	**99.99**	91.91	85.86	91.12	73.63	83.13
K-mer-based mis. len (%)	**0.00**	4.72	13.95	8.12	21.87	9.95
# k-mer-based misjoins	**0**	1	8	6	25	5

Supplementary Table 6: QUAST-LG report on the Fly$_{MP}$ dataset. All statistics are given for scaffolds $\geq$ 3 kb. The best value for each column is indicated in bold. This report was adjusted to fit the page: *Total, Reference, and Total aligned lengths* were converted from bp to Mb, few assembler and metric names were abbreviated.

Assembly	UpperB.	ABYSS	MaSuRCA	Meracul.	Platanus	SOAP	SPAdes
# contigs	1148	1296	2467	971	949	1746	1304
Largest contig	3557620	**6942246**	5434256	5161855	5833330	2635402	5300823
Total length (Mb)	136.0	120.9	**142.2**	128.1	116.5	139.6	123.3
Reference length (Mb)	137.6	137.6	137.6	137.6	137.6	137.6	137.6
GC (%)	42.12	42.72	42.45	42.55	42.57	42.31	42.64
Reference GC (%)	42.08	42.08	42.08	42.08	42.08	42.08	42.08
N50	1031903	**1531186**	340080	816401	1308396	669069	950929
NG50	1014905	**1195395**	357539	755527	987018	670273	827856
N75	346631	529440	57131	401491	**654705**	227940	414317
NG75	**327413**	255679	70042	268560	298738	243593	267284
L50	43	25	77	45	**24**	61	37
LG50	44	**32**	70	51	33	60	45
L75	98	60	355	100	**55**	151	84
LG75	102	94	299	122	**91**	144	116
# misassemblies	**0**	266	922	305	280	713	388
# misassembled contigs	**0**	120	462	168	131	360	188
Misassembled contigs len	**0**	91265295	79411861	89481955	96460407	104322404	94586889
# local misassemblies	**0**	3337	5261	3656	2973	6106	3453
# scaffold gap size mis.	**0**	60	198	113	18	79	88
# possible MGEs	0	22	222	34	12	300	64
# unaligned mis. contigs	**0**	4	21	5	3	19	2
# unaligned contigs	**0+0p**	7+185p	78+655p	28+304p	25+166p	209+565p	14+219p
Unaligned length	**0**	2669127	5566280	3557185	3282998	5813267	3032850
Genome fraction (%)	**99.160**	79.453	84.608	82.583	81.071	84.639	80.405
Duplication ratio	**1.001**	1.086	1.178	1.100	1.019	1.153	1.091
# N's per 100 kbp	**426.96**	8450.06	13111.33	9572.59	2383.77	10522.38	8553.88
# mm per 100 kbp	**0.00**	1166.59	1316.66	1241.33	1288.45	1308.22	1173.67
# indels per 100 kbp	**0.00**	92.10	90.11	91.06	91.18	91.12	93.08
Complete BUSCO (%)	99.67	99.01	**100.00**	98.68	99.01	99.67	99.01
Partial BUSCO (%)	0.00	0.00	0.00	**0.33**	0.00	0.00	0.00
Largest alignment	**3557620**	2693915	1806503	1586080	2811460	1631359	1656356
Total aligned len (Mb)	**136.0**	110.8	120.8	115.7	111.3	121.6	112.5
NA50	**1031903**	433183	144621	375066	454454	235632	336200
NGA50	**1014905**	330827	156571	316385	370701	237681	287132
NA75	**346631**	155112	24633	157501	222678	57501	132523
NGA75	**327413**	31044	31580	87353	69240	64413	42800
LA50	**43**	73	201	97	71	159	100
LGA50	**44**	94	186	111	97	155	123
LA75	**98**	189	815	234	162	443	240
LGA75	**102**	351	691	295	274	418	379
K-mer-based compl.	**97.28**	60.50	63.35	63.51	62.36	63.50	61.39
K-mer-based cor. len (%)	**99.22**	97.69	82.89	97.93	89.40	94.28	54.62
K-mer-based mis. len (%)	**0.00**	1.71	11.70	0.58	10.18	0.89	44.69
# k-mer-based misjoins	**0**	3	66	6	24	12	108

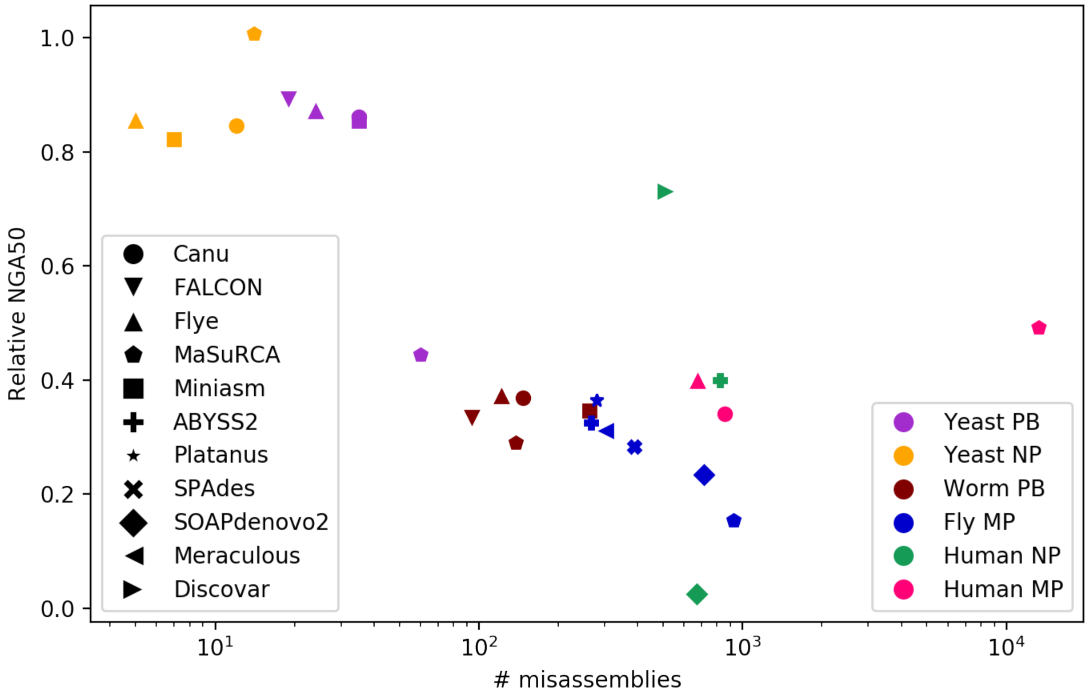
Supplementary Table 7: QUAST-LG report on the Human$_{MP}$ dataset. All statistics are given for scaffolds $\geq$ 3 kb. The best value for each column is indicated in bold.

Assembly	UpperBound	ABYSS	DISCOVAR	SOAPdenovo
# contigs	4958	5014	5802	55725
Largest contig	35878198	21476592	**51394569**	2364922
Total length	2916500502	2814649061	2814592927	**3199541566**
Reference length	3088286401	3088286401	3088286401	3088286401
Reference GC (%)	40.87	40.87	40.87	40.87
N50	**8821063**	4179768	8212463	258443
NG50	**8309069**	3781103	7007197	271290
N75	**4204302**	2060311	3949423	110550
NG75	**3315684**	1534153	2836629	125424
L50	102	197	**93**	3296
LG50	112	231	**111**	3085
L75	223	433	**217**	7945
LG75	**258**	547	280	7239
# misassemblies	**0**	820	508	670
# misassembled contigs	**0**	580	315	586
Misassembled contigs length	**0**	922999235	1078600471	164310662
# local misassemblies	**0**	12192	4881	124705
# scaffold gap size mis.	**0**	25	6	886
# structural variations	0	29	52	85
# possible MGEs	0	168	110	140
# unaligned mis. contigs	**0**	58	100	746
# unaligned contigs	**62+0p**	299+822p	677+795p	5829+19689p
Unaligned length	**13997**	10874314	13388107	104828616
Genome fraction (%)	**99.062**	93.558	94.805	85.100
Duplication ratio	**1.002**	1.020	1.006	1.238
# N's per 100 kbp	**105.30**	2508.11	881.87	20430.97
# mm per 100 kbp	**0.00**	100.49	106.24	129.15
# indels per 100 kbp	**0.00**	27.44	25.87	50.41
Complete BUSCO (%)	**90.10**	89.44	**90.10**	79.54
Partial BUSCO (%)	2.64	4.29	3.30	**10.89**
Largest alignment	**35878198**	20391533	31628681	2193378
Total aligned len	**2915987106**	2766839333	2791316622	2720361441
NA50	**8821063**	3707375	6712668	197114
NGA50	**8309069**	3325956	6093924	210075
NA75	**4204302**	1824769	3159015	48712
NGA75	**3315684**	1305025	2225855	63787
LA50	**102**	224	116	3999
LGA50	**112**	263	138	3725
LA75	**223**	494	263	11628
LGA75	**258**	629	339	10133
K-mer-based compl.	**99.24**	86.92	88.15	77.73
K-mer-based cor. len (%)	**99.22**	77.44	82.75	95.48
K-mer-based mis. len (%)	**0.0**	22.10	16.74	0.77
# k-mer-based misjoins	**0**	572	535	93

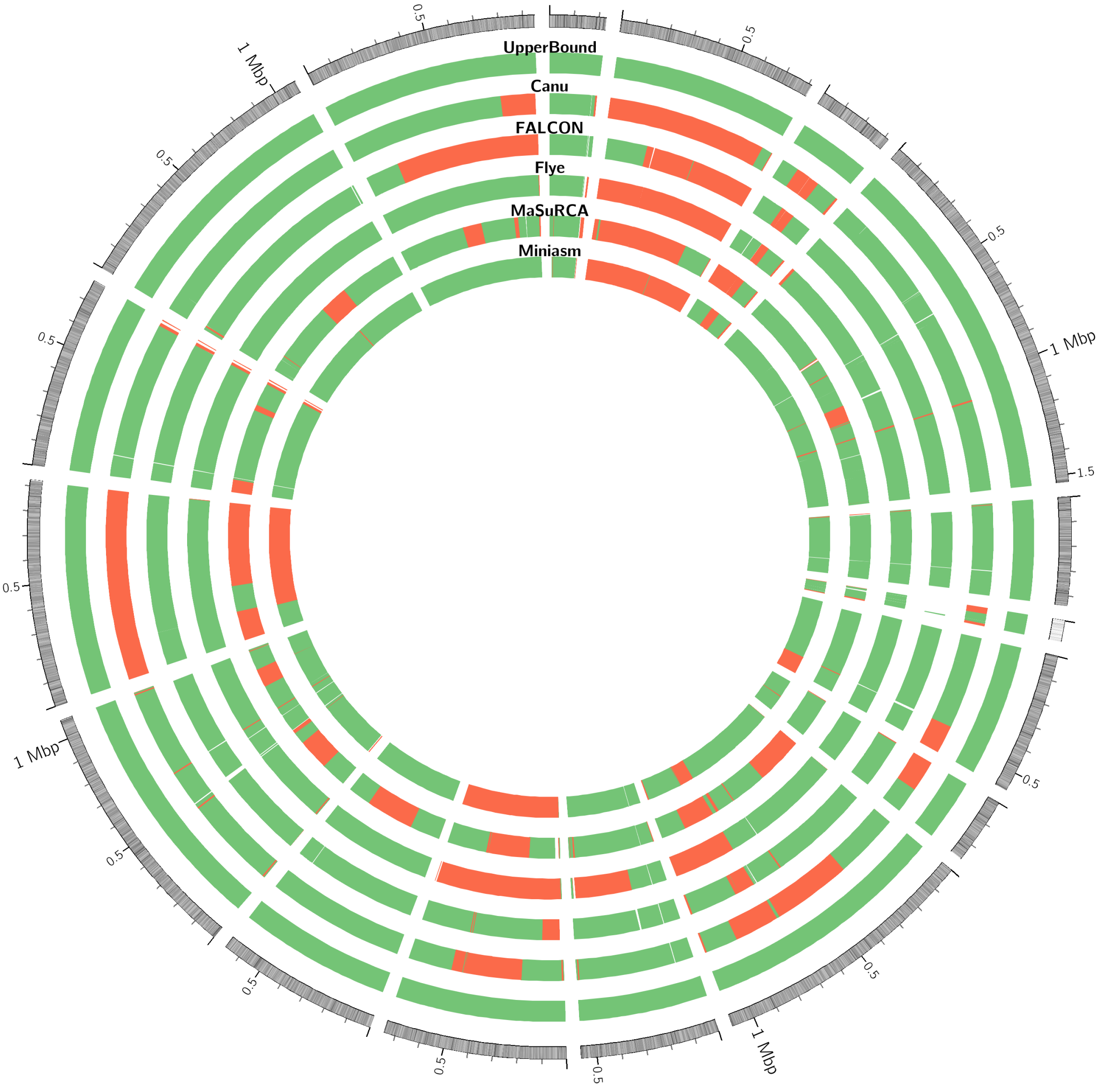
Supplementary Table 8: QUAST-LG report on the Human$_{NP}$ dataset. All statistics are given for scaffolds $\geq$ 3 kb. The best value for each column is indicated in bold.

Assembly	UpperBound	Canu	Flye	MaSuRCA
# contigs	2768	2879	3338	10211
Largest contig	**75724015**	28413671	21995043	22430362
Total length	**2917182483**	2763064770	2803317233	2882560136
Reference length	3088286401	3088286401	3088286401	3088286401
Reference GC (%)	40.87	40.87	40.87	40.87
N50	**8389762**	3763377	4316080	5288590
NG50	**7862149**	3241232	3767461	4968454
N75	**3936041**	1667697	2073525	2576159
NG75	**3530087**	1036013	1439662	2036226
L50	**95**	197	191	162
LG50	**105**	244	227	182
L75	**218**	467	427	353
LG75	**252**	649	550	419
# misassemblies	**0**	853	673	13227
# misassembled contigs	**0**	435	423	4213
Misassembled contigs length	**0**	858930906	746258721	1368133947
# local misassemblies	**0**	54331	24403	12317
# scaffold gap size mis.	**0**	**0**	**0**	1
# structural variations	0	624	411	589
# possible MGEs	0	278	68	656
# unaligned mis. contigs	**0**	112	237	1141
# unaligned contigs	**3+0p**	90+2600p	832+2193p	2710+6632p
Unaligned length	**603**	59016136	129987016	80295329
Genome fraction (%)	**99.074**	92.249	91.909	93.707
Duplication ratio	1.002	0.998	**0.990**	1.018
# N's per 100 kbp	38.36	**0.00**	**0.00**	11.87
# mm per 100 kbp	**0.00**	258.95	580.26	184.06
# indels per 100 kbp	**0.00**	68.04	1125.37	31.94
Complete BUSCO (%)	**90.10**	86.47	51.49	83.50
Partial BUSCO (%)	2.64	5.61	**18.15**	4.29
Largest alignment	**75724015**	25750637	21734865	22412626
Total aligned len	**2916136310**	2703373535	2672642008	2799134474
NA50	**8389762**	3144867	3610146	4214832
NGA50	**7862149**	2744681	3172168	3931830
NA75	**3936041**	1408424	1618830	2064920
NGA75	**3530087**	776072	1042595	1547559
LA50	**95**	241	223	201
LGA50	**105**	296	266	226
LA75	**218**	567	510	439
LGA75	**252**	795	672	525
K-mer-based compl.	**99.51**	83.93	26.59	85.72
K-mer-based cor. len (%)	**99.31**	84.62	92.36	62.97
K-mer-based mis. len (%)	**0.00**	14.93	4.43	32.77
# k-mer-based misjoins	**0**	523	97	892

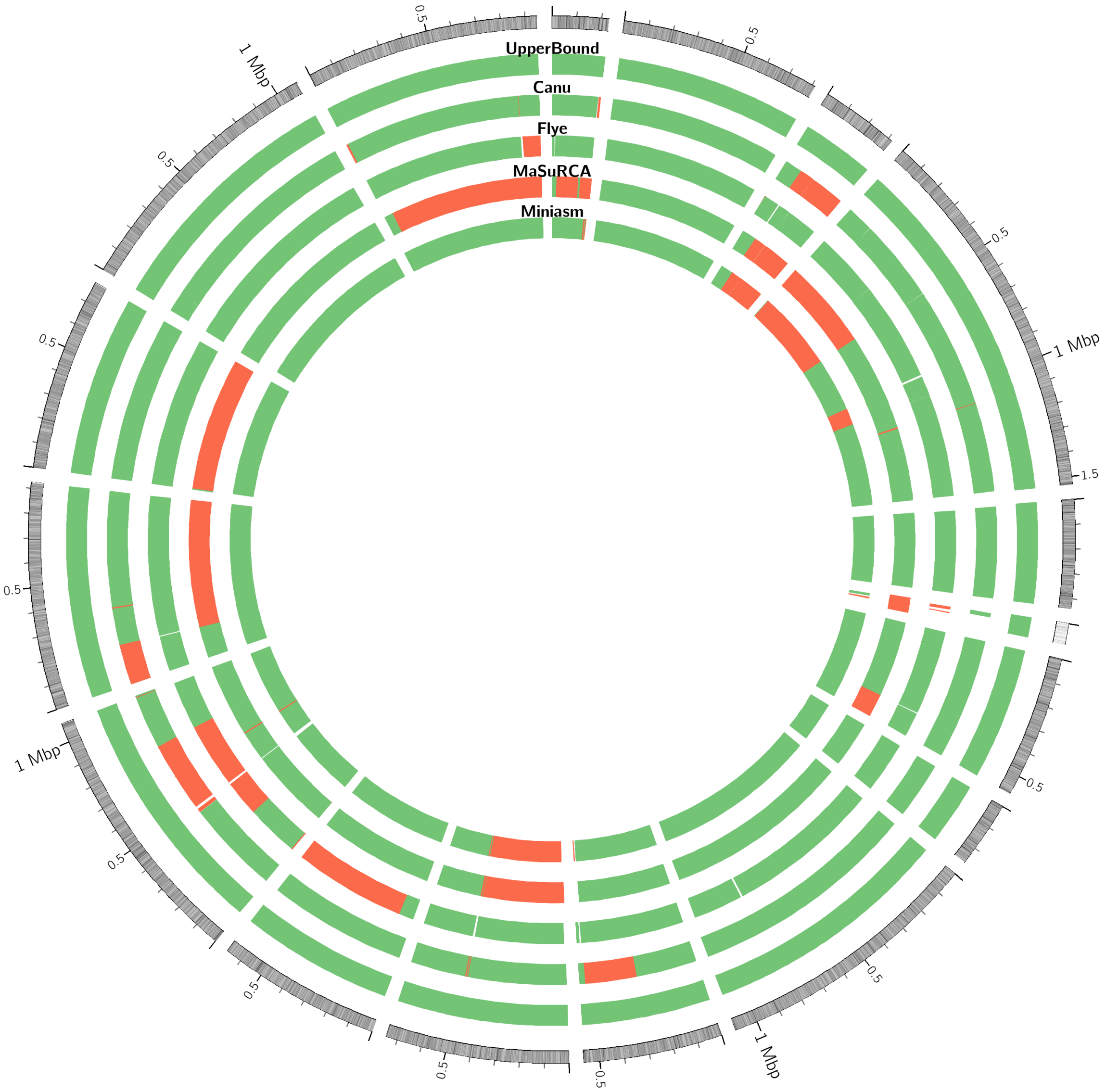
Supplementary Figures



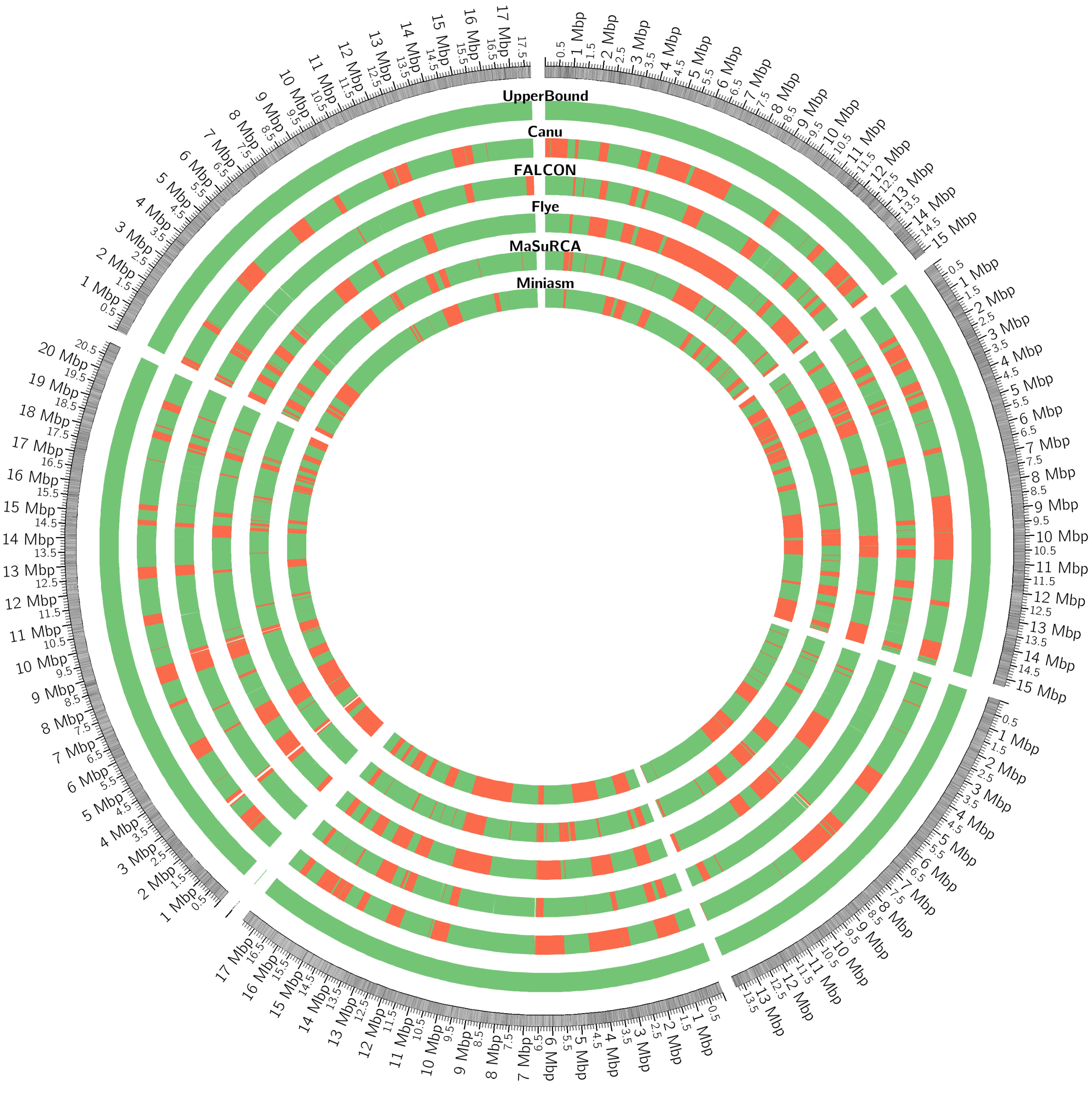
Supplementary Figure 1: **Assemblers performance on six benchmark datasets**. Datasets are indicated by color, assemblers are shown by shape. *Relative NGA50* is the scaffold NGA50 divided by UpperBound NGA50 for a given dataset. The x-axis (*# misassemblies*) is in a logarithmic scale.



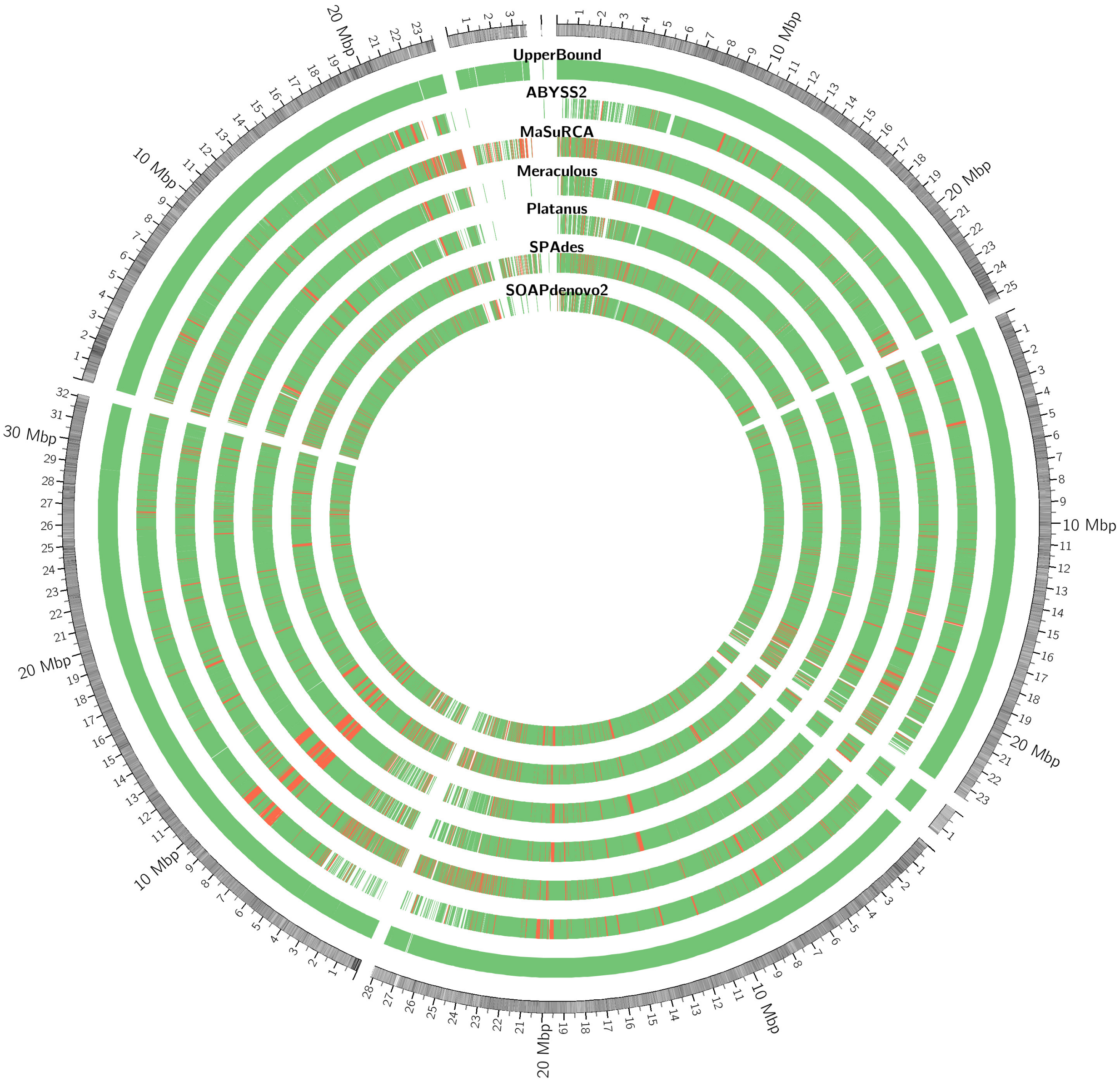
Supplementary Figure 2: **Circular alignment viewer for the Yeast$_{PB}$ dataset**. The outer circle represents reference chromosomes with GC (%) heatmap (white for GC-poor and black for GC-rich regions). The inner circles are assemblies with green for correct contigs and red for contigs containing at least one misassembly breakpoint. The figure is generated using Icarus [4] and Circos [5] software.



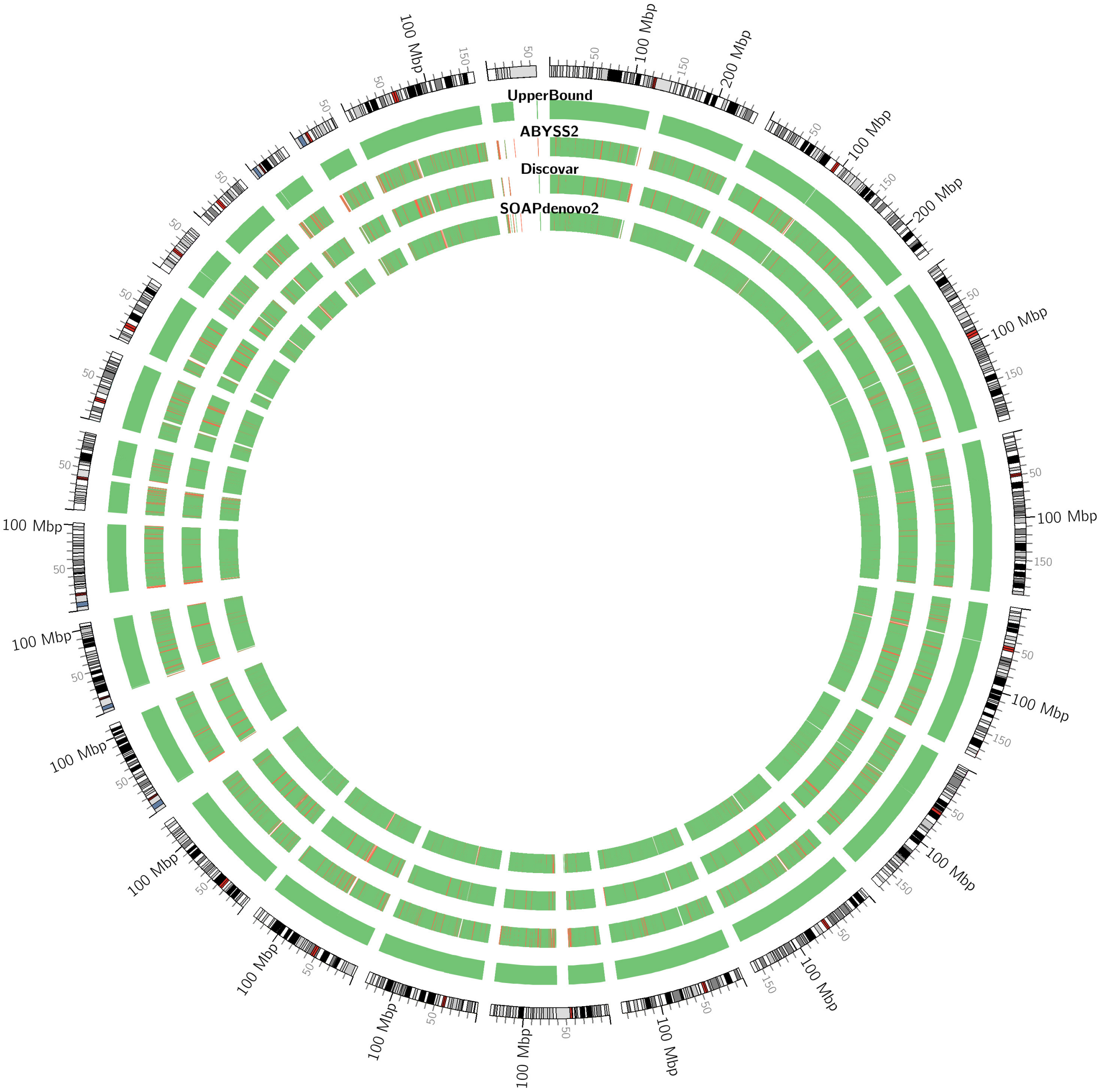
Supplementary Figure 3: **Circular alignment viewer for the Yeast$_{NP}$ dataset**. The outer circle represents reference chromosomes with GC (%) heatmap (white for GC-poor and black for GC-rich regions). The inner circles are assemblies with green for correct contigs and red for contigs containing at least one misassembly breakpoint. The figure is generated using Icarus [4] and Circos [5] software.



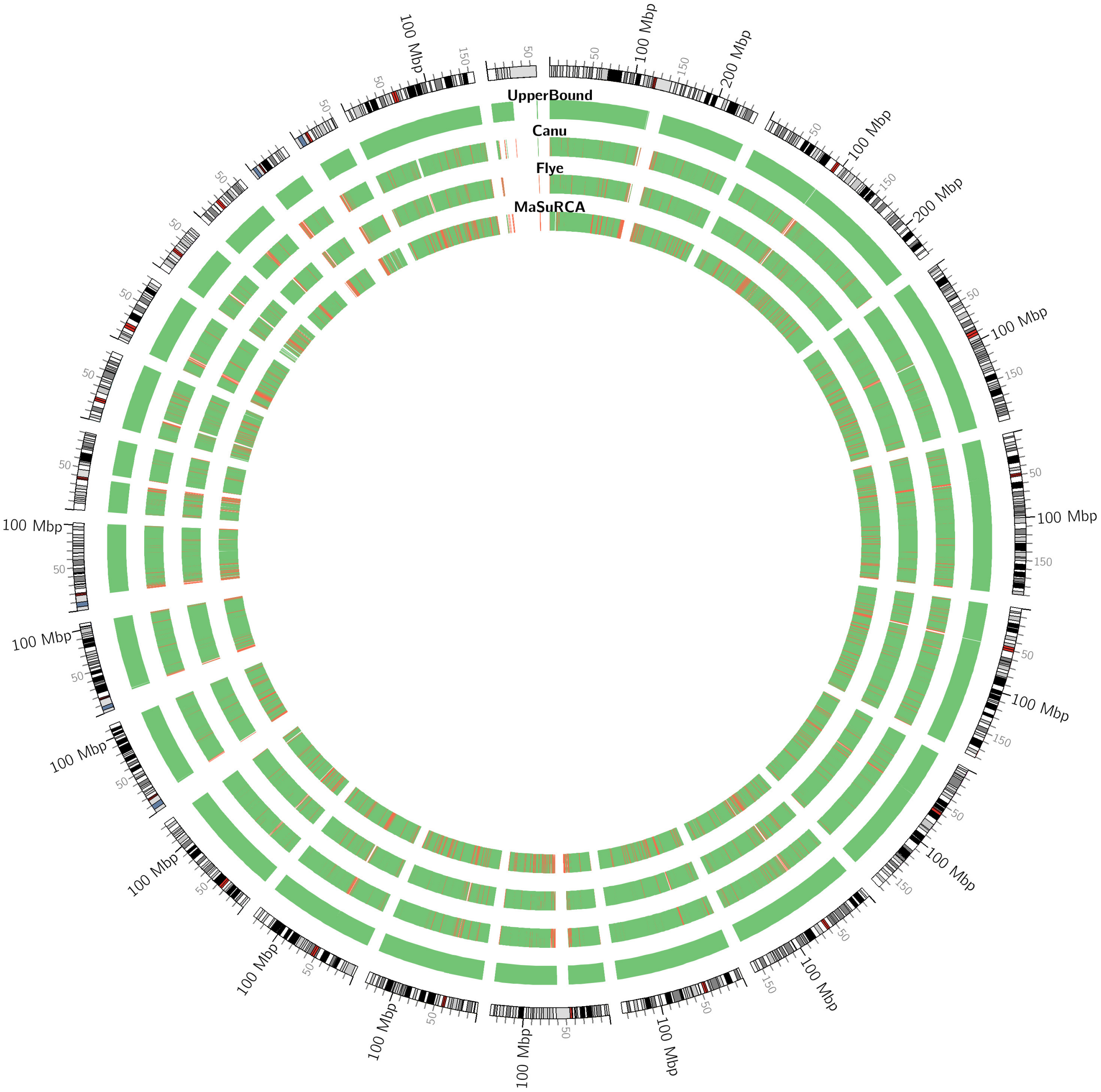
Supplementary Figure 4: **Circular alignment viewer for the Worm$_{PB}$ dataset**. The outer circle represents reference chromosomes with GC (%) heatmap (white for GC-poor and black for GC-rich regions). The inner circles are assemblies with green for correct contigs and red for contigs containing at least one misassembly breakpoint. The figure is generated using Icarus [4] and Circos [5] software.



Supplementary Figure 5: **Circular alignment viewer for the Fly$_{MP}$ dataset**. The outer circle represents reference chromosomes with GC (%) heatmap (white for GC-poor and black for GC-rich regions). The inner circles are assemblies with green for correct contigs and red for contigs containing at least one misassembly breakpoint. The figure is generated using Icarus [4] and Circos [5] software.



Supplementary Figure 6: **Circular alignment viewer for the Human$_{MP}$ dataset**. The outer circle represents annotated reference chromosomes where centromeres are indicated with red color and all other regions are in gray scale. The inner circles are assemblies with green for correct contigs and red for contigs containing at least one misassembly breakpoint. The figure is generated using Icarus [4] and Circos [5] software.

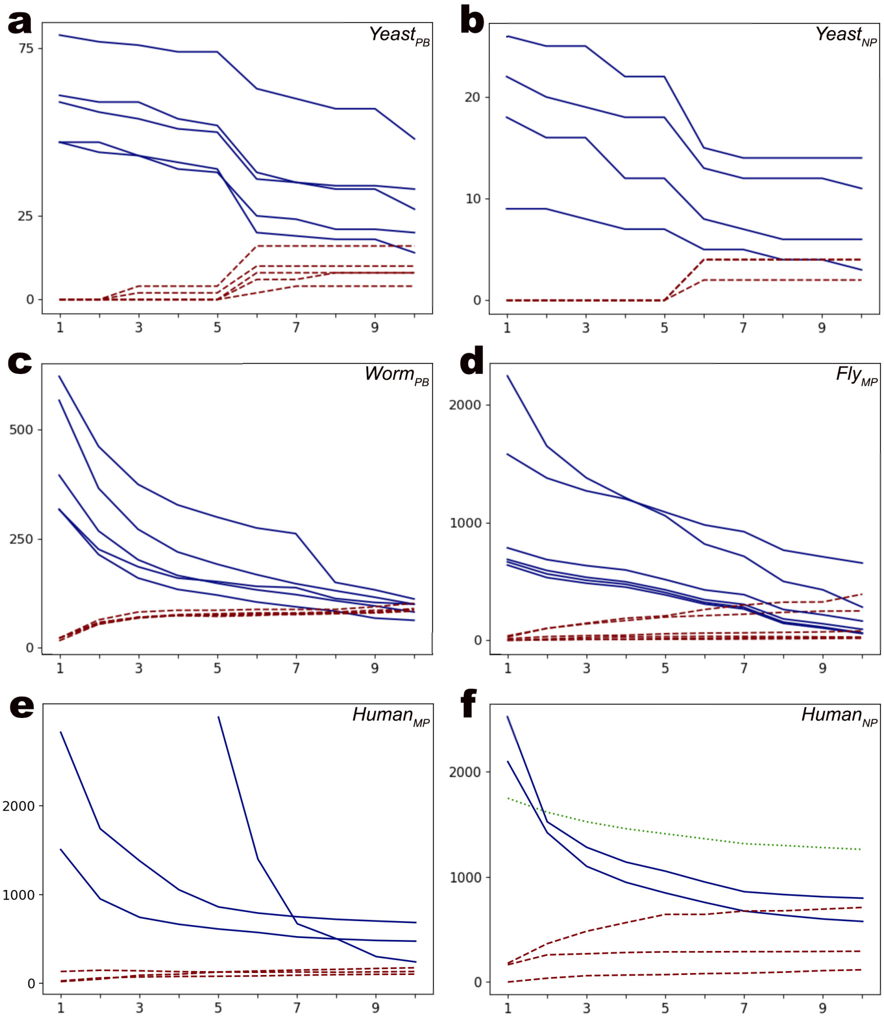


Supplementary Figure 7: **Circular alignment viewer for the Human$_{NP}$ dataset**. The outer circle represents annotated reference chromosomes where centromeres are indicated with red color and all other regions are in gray scale. The inner circles are assemblies with green for correct contigs and red for contigs containing at least one misassembly breakpoint. Note that the vast majority of the misassembled contigs are located in the centromeres regions which is visible especially in the MaSuRCA track. The figure is generated using Icarus [4] and Circos [5] software.

Supplementary Methods

Choice of the breakpoint threshold

QUAST-LG procedure for transposable elements (TEs) identification critically depends on the size of the breakpoint threshold X. This threshold should ideally fit the length of the largest TE in the genome. To find the optimal value for X, we measured the number of misassemblies and the number of possible TEs identified by our procedure in all assemblies of the six benchmark datasets for various X in a range of 1-10 kb with a step of 1 kb (Supplementary Figure 8). The figure demonstrates that the number of misassemblies drops significantly with increase of X until a certain point $X = X_c$ where the curve nearly flattens. Likewise, the number of possible TEs quickly goes up from an almost zero value at $X = 1$ kb and eventually reaches plateau. We can see that the value of X_c varies among genomes: for the two yeast datasets it is clearly close to $X_c = 6$ kb; the worm dataset reaches it at approximately 2-3 kb; and the fruit fly, as well as human datasets' critical points seem to be around 6-7 kb. In fact, X_c strongly correlates with the largest TE among the most common TEs in the corresponding organisms. Indeed, Ty1-like retrotransposons in the yeast genome are up to 5.9 kb long [6], Tc5 transposon in the worm genome is 3.2 kb long [7], TEs of LTR families in the fruit fly genome are up to 7.5 kb [8], and the most widespread active TEs in the human genome, LINE-1, are approximately 6 kb long [9, 10]. For the sake of consistency, we used the same $X = 7$ kb in all benchmark experiments and it is the default value in QUAST-LG. A user may use "–extensive-mis-size" option to force QUAST-LG to use a different X value.



Supplementary Figure 8: **The number of misassemblies and possible TEs at different breakpoint thresholds** X. The x-axis indicate X value in kb. The y-axis displays the number of misassemblies (blue solid lines) and possible TEs (red dashed lines). Each line corresponds to an assembly of a corresponding dataset; upper bound assemblies are not shown. Note that SOAPdenovo assembly of Human$_{MP}$ (e) and MaSuRCA assembly of Human$_{NP}$ (f) have too many misassemblies to fit these scaled plots, so SOAPdenovo values are shown only with $X \geq 6$ kb, and MaSuRCA values are divided by ten and plotted using a green dotted line.

Best set of alignments selection

Long contigs are rarely mapped to the reference genome perfectly as a single unambiguous alignment. An alignment software typically reports multiple alignment fragments $A = (a_1, \ldots, a_n)$ mapped to the differentps of the genome. Note, that some of the alignments may correspond to the same contig fragment mapped to distinct genomic positions. This may happen due to the presence of genomic repeats and transposable elements (TEs) in the reference genome and in some cases — algorithmic issues in the assembly and alignment software. QUAST-LG attempts to accurately assess each contig and select the set of non-overlapping alignments $A' \subseteq A$, which maximizes the total alignment score $Score(A')$ (described below). To solve this problem, we implement a dynamic programming algorithm called *BestSetSelection* (see Algorithm 1). This algorithm is conceptually similar to the algorithm described in ref. [11] for selecting the best spliced alignment for a mapped transcript but includes a significant speed up which allows to apply it for evaluation of large genome assemblies. The *BestSetSelection* algorithm takes as input the list of contig alignments A sorted according to the position in the contig of right-most base of each aligned fragment.

Algorithm 1 Best Set Selection

```
1: procedure BestSetSelection(Sorted list of alignments $A$)
2:     $BestSets \leftarrow \{(EmptyAlignment, 0)\}$
3:     for all $a_i \in A$ do
4:         $best_i \leftarrow argmax_{B \in BestSets} Score(B \cup a_i)$
5:         $BestSets \leftarrow BestSets \cup (best_i, Score(best_i))$
       return $argmax_{B \in BestSets} Score(B)$
```

The scoring function depends on the *alignment score* of every individual alignment and *locality score* between adjacent alignments. Score of a single alignment a is defined as $AlignmentScore(a) = Length(a) * Identity(a)$, where $Identity(a)$ is reported by the alignment software (100% for a perfect match and decreases with the number of short indels and mismatches). Given a scored set of alignments $A = (a_1, \ldots, a_{k-1})$, the score of $A \cup a_k$ is computed as following:

$$Score(A \cup a_k) = Score(A) + AlignmentScore(a_k) - Penalty(a_{k-1}, a_k) - OverlapLength(a_{k-1}, a_k) * Identity(a_k), \tag{1}$$

where $Penalty(a_{k-1}, a_k)$ depends on the inconsistency between a_{k-1} and a_k in the genome. Higher penalty values are given for the extensive misassembly events and smaller coefficient for long indels or short local errors. The last term in equation (1) guarantees that the extension of the set with an alignment fully overlapping in contig with another alignment from the set is unprofitable. The described function satisfies the conditions for the scoring functions compatible with the *BestSetSelection* algorithm deduced in ref. [11].

The algorithm *BestSetSelection* takes $O(n^2)$ time, where n is the total number of alignments in A. This number is usually small since the short alignments are filtered out prior to running the algorithm. However, contigs in large eukaryotic assemblies may contain up to dozens of thousands alignments. To prevent the performance drop in this case, we implemented a speed up heuristic if the size of A exceeds threshold N (the default value is 100).

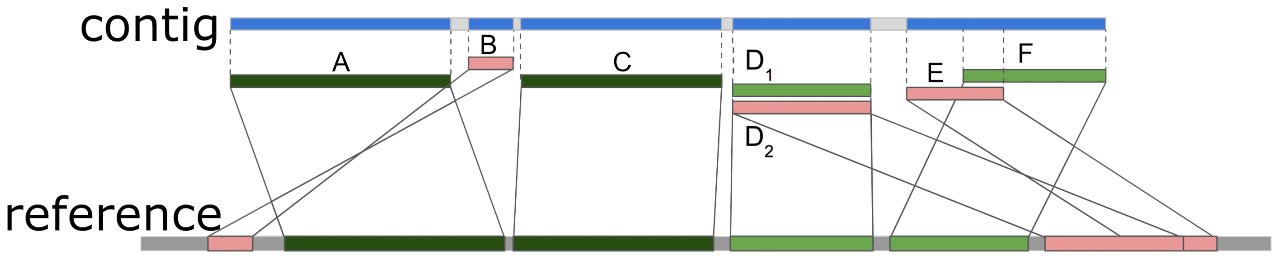
To construct $best_i$ (the best alignment set ending at a_i) the algorithm iterates through all already computed alignment sets $B \in BestSets$ and chooses the one that gives the best score for $B \cup a_i$ (line 4 of the algorithm 1). However, it appears that the distance between the majority of the alignment sets $B \in BestSets$ and a_i is large, which makes $Score(B \cup a_i)$ too small to add $B \cup a_i$ to $BestSets$. Below we suggest an heuristics that allows to iterate only through a small subset of $BestSets$ and thus reduce the running time.

We define a as a *solid* alignment if $Score(A' \cup a) > Score(A')$ for any $A' \subset A \setminus a$, i.e. if it its addition to any subset of alignments $A \setminus a$ improves its $Score$. By the definition, all solid alignments from A are included in the resulting best set of alignments (it would be possible to create a set with a higher score otherwise). Inpicular, $best_i$ includes all solid alignments located to the left of a_i in the contig. Thus, the algorithm can only iterate though those sets $B \in BestSets$ that include the right-most solid alignment before a_i. Since the $BestSets$ is constructed iteratively, these sets will include all solid alignments to the left of a_i. This speed-up resulted in up to 10x drop of the running time on the fruit fly assemblies evaluation and allowed us to complete quality assessment for the human assemblies (Supplementary Table 1).

The criteria for choosing solid alignments can be deduced from equation (1). An alignment a is guaranteed to be solid if

$$UniqueLength(a) * Identity(a) > m * MaxPenalty, \tag{2}$$

where $UniqueLength(a)$ is the length of a without overlaps with all other alignments, $MaxPenalty$ is the maximal penalty value, and $m = 1$ if a is located on the start/end of the contig and $m = 2$ otherwise. Supplementary Figure 9 shows an example of the best alignment set and solid alignments.



Supplementary Figure 9: **Solid alignments detection**. A contig (gray line at the top) has multiple alignments (dark and light green and pink bars in the middle) to the reference genome (gray line at the bottom), the alignments positions are visualized (with blue color in the contig and green/pink in the reference). Alignments A and C (dark green) are solid since they are sufficiently long and do not have large overlaps with other alignments. Alignments B, E and F have significant overlaps, so their $UniqueLength$ is not enough to mark them solid. Alignments D_1 and D_2 are ambiguous and thus cannot be solid (their $UniqueLength$ is equal to 0). Alignments for the best set are colored green (dark green for solid alignments and light green for the rest), the unused alignments are colored light pink.

The described heuristic always identifies the alignment set with the maximal alignment score by design. However, it could in theory produce zero speed up or even small slowdown if the input set contains no or very few solid alignments or, on the contrary, if all or almost all alignments in the set are solid. Nevertheless, our benchmark experiments on various real datasets demonstrate that such cases are almost impossible in practice and the heuristic always works sufficiently well. Note that there are alternative theoretical sub-quadratic algorithms for the best set selection problem [12] but they are always associated with a large constant which makes them impractical comparing to our heuristic approach.

Supplementary References

[1] H. Li, "Minimap2: fast pairwise alignment for long nucleotide sequences," *arXiv:1708.01492*, 2017.

[2] S. Kurtz *et al.*, "Versatile and open software for comparing large genomes," *Genome Biol.*, vol. 5, no. 2, p. R12, 2004.

[3] F. Simao *et al.*, "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs," *Bioinformatics*, Jun 2015.

[4] A. Mikheenko *et al.*, "Icarus: visualizer for de novo assembly evaluation," *Bioinformatics*, vol. 32, pp. 3321–3323, Nov 2016.

[5] M. Krzywinski *et al.*, "Circos: an information aesthetic for comparative genomics," *Genome Res.*, vol. 19, pp. 1639–1645, Sep 2009.

[6] C. Neuveglise, H. Feldmann, E. Bon, C. Gaillardin, and S. Casaregola, "Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts," *Genome Res.*, vol. 12, pp. 930–943, Jun 2002.

[7] J. J. Collins and P. Anderson, "The Tc5 family of transposable elements in Caenorhabditis elegans," *Genetics*, vol. 137, pp. 771–781, Jul 1994.

[8] J. S. Kaminker *et al.*, "The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective," *Genome Biol.*, vol. 3, no. 12, p. RESEARCH0084, 2002.

[9] E. T. Prak and H. H. Kazazian, "Mobile elements and the human genome," *Nat. Rev. Genet.*, vol. 1, pp. 134–144, Nov 2000.

[10] J. S. Myers, B. J. Vincent, H. Udall, W. S. Watkins, T. A. Morrish, G. E. Kilroy, G. D. Swergold, J. Henke, L. Henke, J. V. Moran, L. B. Jorde, and M. A. Batzer, "A comprehensive analysis of recently integrated human Ta L1 elements," *Am. J. Hum. Genet.*, vol. 71, pp. 312–326, Aug 2002.

[11] E. Bushmanova, D. Antipov, A. Lapidus, V. Suvorov, and A. D. Prjibelski, "rnaQUAST: a quality assessment tool for de novo transcriptome assemblies," *Bioinformatics*, vol. 32, pp. 2210–2212, Jul 2016.

[12] M. I. Abouelhoda and E. Ohlebusch, "Chaining algorithms for multiple genome comparison," *Journal of Discrete Algorithms*, vol. 3, no. 2, pp. 321 – 341, 2005. Combinatorial Pattern Matching (CPM) Special Issue.