Supplementary Table S2.4 Comparison of Three Preliminary Assemblies.

Item	Description	Atlas-WGS	Celera Assembler	PCAP
I	**Read statistics**			
	Total assembled bases	2,864,035,527	2,827,563,539	3,325,382,728
	Total assembled q20 bp	2,860,153,945	2,809,362,341	3,266,879,945
	Percentage q20 bp	99.86%	99.36%	98.24%
II	**Contig statistics**			
	Total sequence length	2,921,521,432	2,874,820,792	3,381,158,713
	Total contig length in scaffolds	2,837,338,156	2,827,563,539	3,325,382,728
	Number of contigs	406,629	267,019	717,725
	N50 size of contigs	15,090	26,190	12,380
	Number of contigs greater than N50	55,228	31,262	72,233
	Average size of contigs	7,043	10,589	4,633
	Largest contig size	139,301	219,335	159,505
	Major contigs (>1k)	324,954	262,193	564,096
	Statistics			
	By tiers			
	0.0 - 2.5 GB			
	Contig number	184796	111851	183021
	Average contig length	13528	22351	13660
	Maximum contig length	139301	219335	159505
	N50 contig length	17409	29919	17352
	N50 contig number	43999	25409	44039
	2.5 - 2.7 GB			
	Contig number	66964	56612	57635
	Average contig length	2987	3533	3470
	Maximum contig length	4074	5979	4382
	N50 contig length	3076	3952	3505
	N50 contig number	28214	20337	25517
	2.7 GB - end			
	Contig number	154869	98556	477069
	Average contig length	1059	1294	1311

	Maximum contig length	2126	1877	2788
	N50 contig length	1297	1317	1402
	N50 contig number	48884	41701	162898
III	**Scaffold statistics**			
	Total	2,837,338,156	2,827,563,539	3,325,382,728
	Number of scaffolds or supercontigs	109,408	94,322	354,753
	N50 size of scaffolds	1,484,948	1,890,602	1,174,170
	Number of scaffolds greater than N50	536	402	696
	Average size of scaffolds	31,187	29,978	9,374
	Largest scaffold size	10,317,268	11,720,342	12,067,300
	Major scaffolds (contigs >1kb)	23,186	94,256	225,173
	By tiers			
	0.0 - 2.5 GB			
	Supercontig number	2093	1910	2105
	Average supercontig length	1194462	1308939	1187737
	Maximum supercontig length	10317268	11720342	12067300
	N50 supercontig length	1745418	2259030	1754951
	N50 supercontig number	423	323	408
	2.5 - 2.7 GB			
	Supercontig number	3272	7185	2096
	Average supercontig length	61126	27836	95424
	Maximum supercontig length	240592	193375	239581
	N50 supercontig length	106066	66340	110634
	N50 supercontig number	614	903	607
	2.7 GB - end			
	Supercontig number	86469	85227	350552
	Average supercontig length	922	1496	1783
	Maximum supercontig length	9266	4939	40383
	N50 supercontig length	n.a.	1395	1936
	N50 supercontig number	n.a.	31447	60134
IV	**Genome content**			
	Total GC count in genome in contigs	1,171,425,227	1,154,580,149	1,360,546,724
	Total GC count in genome in scaffolds	1,160,430,236	1,154,580,149	1,360,546,724
	Percentage	40.10%	40.83%	40.91%

	Percentage of bases in contigs	40.09%	40.16%	40.23%
	Percentage of bases in scaffolds	40.89%	40.83%	40.91%
	Total AT count in genome	1,692,607,221	1,672,983,390	1,964,823,238
	Percentage	57.90%	59.20%	59.10%
a	Total N count in genome	57,488,984	0	12,766
	Total N count in genome in contigs	3,079	0	12,766
	Total N count in genome in scaffolds	2,819	0	12,766
	Percentage	2.00%	0.00%	0.00%
V	**Mate Pair analysis**			
	Total Reads	20,399,141	19,810,790	22,260,797
	Total mate pairs in contigs	6,260,956	7,065,394	6,180,279
	Number of satisfied mate pairs in a contig	6,039,469	6,880,959	5,995,925
	Percentage satisfied mate pairs in a contig	96.46%	97.39%	97.02%
	Satisfied is insert size within 30% of mean	6,042,537	6,881,025	5,996,187
	Percentage satisified insert size	96.51%	97.39%	97.02%
	Satisfied is expected orientation	6,183,782	7,058,001	6,151,196
	Percentage satisfied orientation	98.77%	99.90%	99.53%
	Unsatisfied pairs within a contig	211,487	184,435	184,354
	Distance between pairs outside of acceptable range	144,313	177,042	155,271
	Insertion	89,731	105,210	77,425
	Deletion	54,582	71,832	77,846
	Orientation wrong	77,174	7,393	29,083
	Same direction	52,123	6,370	22,432
	Cross direction	25,051	1,023	6,651
	Singlet - one mate is not in any contig	2,997,359	2,358,717	1,935,070
b	No library data	5,148	725,681	6,437
	Total mate pairs between contigs	2,439,935	1,297,802	3,543,838
	Total mate pairs between scaffolds	331,817	146,421	1,181,935
	Total mate pairs between contigs and within scaffolds	2,108,118	1,151,381	2,361,903
	Number of satisfied mate pairs in a scaffold	1,790,577	1,119,131	2,080,850
	Percentage satisfied mate pairs in a scaffold	84.94%	97.20%	88.10%
	Percentage satisfied mate pairs between contigs	73.39%	86.23%	58.72%

Satisfied is insert size within 30% of mean	1,793,198	1,119,269	2,086,367
Percentage satisfied insert distance between contigs	73.49%	86.24%	58.87%
Percentage satisfied insert distance in scaffolds	85.06%	97.21%	88.33%
Satisfied is expected orientation	2,085,456	1,150,305	2,341,374
Percentage satisfied orientation between contigs	85.47%	88.63%	66.07%
Percentage satisfied orientation in scaffolds	98.81%	99.90%	99.42%
Unsatisfied pairs between contigs	649,358	178,671	1,462,988
Unsatisfied pairs in scaffolds	316,411	32,250	281,053
Distance between pairs outside of acceptable range	294,878	31,174	260,524
Insertions	166,738	25,102	222,238
Deletions	128,140	6,072	38,286
Orientation wrong	21,533	1,076	20,529
Same direction	14,784	560	2,537
Cross direction	6,749	516	17,992
Total pairs in and between contigs within scaffold	8,369,074	8,216,775	8,542,182
Number of satisfied mate pairs in a scaffold	7,830,046	8,000,090	8,076,775
Percentage sastisfied mate pairs in a scaffold	93.56%	97.36%	94.55%
Satisfied is insert size within 30% of mean	7,835,735	8,000,294	8,082,554
Percentage satisfied distance	93.63%	97.37%	94.62%
Satisfied is expected orientation	8,269,238	8,208,306	8,492,570
Percentage sastified orientation	98.81%	99.90%	99.42%
Unsatisfied pairs in scaffolds	537,898	218,685	465,407
Distance between pairs outside of acceptable range	439,191	210,216	415,795
Insertion	256,496	130,312	299,663
Deletion	182,722	79,904	116,132
orientation wrong	98,707	8,469	49,612
same direction	66,907	6,930	24,969
cross direction	31,800	1,539	24,643

Patterns of unsatisfied mates that indicate assembly issues

	Insertions	256,496	130,312	299,663
	Deletions	182,722	77,904	116,132
	Orientation wrong	98,707	8,469	49,612
VI	**Template analysis**			
	Library insert size distribution			
	837320595 average size	3822	3727	3800
	994688578 average size	39000		
	837320585 average size		3716	
	969080626 average size		3049	3171
	837320587 average size		3723	3788
	LAWEP average size	2262	2143	2215
	LAWFP average size	2815	2672	2763
	LAWNE average size	3278	3094	3248
	LAWNP average size	3265	3085	3233
	LAWNQ average size	3266	3060	3224
	LAWRP average size	3769	3574	3730
	MACAQUE_T14211 average size	9225	9068	9117
	MACAQUE_T14212 average size		11163	
	MACAQUE_T14218 average size	2032	1973	2022
	MACAQUE_T14219 average size	2450	2389	2437
	RHESUS_MACAQUE_T23563-RT-1P3KB average size	1934	1854	1912
	837320595 Number of templates in contigs	103796	118710	103958
	994688578 Number of templates in contigs	7888		
	837320585 Number of templates in contigs		102769	
	969080626 Number of templates in contigs		97067	96322
	837320587 Number of templates in contigs		108786	94724
	LAWEP Number of templates in contigs	457061	488645	451801
	LAWFP Number of templates in contigs	141723	155834	140150
	LAWNE Number of templates in contigs	532394	597750	526785
	LAWNP Number of templates in contigs	411023	462463	409216
	LAWNQ Number of templates in contigs	457805	513943	455978
	LAWRP Number of templates in contigs	121079	142751	118583
	MACAQUE_T14211 Number of templates in	425511	630908	437269

	contigs			
	MACAQUE_T14212 Number of templates in contigs		85673	
	MACAQUE_T14218 Number of templates in contigs	123292	135067	126753
	MACAQUE_T14219 Number of templates in contigs	346226	388240	354362
	RHESUS_MACAQUE_T23563-RT-1P3KB Number of templates in contigs	173343	192262	180998
	Template coverage - average	9.2	10.1	8.8
	Bases used in calculation	2,525,990,148	2,620,294,223	2,592,224,255
IX	**Comparison to finished sequence**			
	Sequences to compare to			
	Finished BACs – Encode region DP000005			
	BCM Megablast alignments of scaffolds			
	Total number of matched regions	174	131	170
	Total Coverage	1,621,316	1,623,897	1,643,175
	Percentage Coverage	96.59%	96.74%	97.89%
	Total Overlap	1,259	785	12,690
	Percentage Overlap	0.0777%	0.0483%	0.7723%
	Maximum gap between matches	5,626	17,262	5,257
	Location of maximum gap	1267009-1281155	391807-399900	1215665-1232064
	Unaligned bases	5,347	3,695	23,121
	Percent of total unmatched bases (ATGC/ATGCN)	8.00%	8.00%	54.00%
X	**Comparison to expressed sequences**			
	Macaque			
	Macaque Proteins pmatch Number	1,524	1,516	1,541
	Macaque Proteins pmatch Percent matched	86.59	86.14	87.56
	Macaque Proteins pmatch Percent missed	13.41	13.86	12.44
	Macaque Proteins genewise Number	1,433	1,436	1,499
	Macaque Proteins genewise Percent matched	81.42	81.59	85.17
	Macaque Proteins genewise Percent missed	18.58	18.41	14.83
	Macaque Proteins exonerate Number	1,730	1,728	1,731

Macaque Proteins exonerate Percent matched	98.30	98.18	98.35
Macaque Proteins exonerate Percent missed	1.70	1.82	1.65
Macaque cDNAs Exonerate 90/97 Number	871	844	910
Macaque cDNAs Exonerate 90/97 Percent matched	54.30	52.62	56.73
Macaque cDNAs Exonerate 90/97 Percent missed	45.70	47.38	43.27
Macaque cDNAs Exonerate unfiltered Number	1,365	1,366	1,365
Macaque cDNAs Exonerate unfiltered Percent matched	85.10	85.16	85.10
Macaque cDNAs Exonerate unfiltered Percent missed	14.90	14.84	14.90
Macaque ESTs Exonerate 90/97 Number	21,138	22,488	23,914
Macaque ESTs Exonerate 90/97 Percent matched	40.09	42.65	45.36
Macaque ESTs Exonerate 90/97 Percent missed	59.91	57.35	54.64
Macaque ESTs Exonerate unfiltered Number	51,199	51,307	51,502
Macaque ESTs Exonerate unfiltered Percent matched	97.10	97.31	97.68
Macaque ESTs Exonerate unfiltered Percent missed	2.90	2.69	2.32
Human and Chimp			
Human Proteins pmatch Number	57,999	57,717	58,109
Human Proteins pmatch Percent matched	87.82	87.39	87.99
Human Proteins pmatch Percent missed	12.18	12.61	12.01
Human Proteins genewise Number	52,124	52,351	52,833
Human Proteins genewise Percent matched	78.92	79.27	80.00
Human Proteins genewise Percent missed	21.08	20.73	20.00
Human Proteins exonerate Number	65,364	65,295	65,399
Human Proteins exonerate Percent matched	98.97	98.87	99.02
Human Proteins exonerate Percent missed	1.03	1.13	0.98
Human cDNAs Exonerate 90/97 Number	84,122	84,122	89,986
Human cDNAs Exonerate 90/97 Percent matched	42.67	42.67	45.65
Human cDNAs Exonerate 90/97 Percent missed	57.33	57.33	54.35

	Human cDNAs Exonerate unfiltered Number	192,697	192,407	192,663
	Human cDNAs Exonerate unfiltered Percent matched	97.75	97.61	97.73
	Human cDNAs Exonerate unfiltered Percent missed	2.25	2.39	2.27
XII	**Fragment incorporation**			
	Alignment to other assemblies			BAC read data missing for VI, this will alter aligned and overlapping sequence values, if uniformly distributed in contigs, overall picture remains.
	Aligned to BCM (Gb)	n/a	2.69	2.73
	Aligned to BCM - Overlapping sequences (Mb)	n/a	9.6	163.2
	Aligned to BCM number of regions	n/a	5670	85750
	Aligned to VI (Gb)	2.58	n/a	2.64
	Aligned to VI - Overlapping sequences (Mb)	26.7	n/a	119.9
	Aligned to VI number of regions	9868	n/a	64451
	Aligned to WU (Gb)	2.57	2.84	n/a
	Aligned to WU - Overlapping sequences (Mb)	11.1	5.2	n/a
	Aligned to WU number of regions	4611	3361	n/a
XIII	**Comparison to other genomes**			
	Human NCBI 35			
	Total bp in ungapped matches	2,175,116,580	2,167,520,892	2,159,510,782
	Bp in matches in clumps	2,141,623,881	2,147,709,290	2,135,462,800
	Bp in matches near clumps, same orientation	8,395,708	3,826,013	5,830,551
	Bp in matches near clumps, reverse orientation	9,097,889	3,791,444	5,543,130
	Bp in matches remote from clumps	15,999,102	12,194,145	12,674,301
	Total number of clumps	36,476	17,622	60,770
	smallest clump needed to cover 1Gb of matches	1,292,875	1,811,956	1,300,856

smallest clump needed to cover 2Gb of matches	139,145	167,070	56,209
Blastz to NCBI 35			
Total human bp covered by assembly	2,551,073,721	2,536,641,486	2,572,302,415
Percentage Coverage	89.00%	88.50%	89.75%
Scaffold statistics for Large Clumps - mapping			
Total Scaffolds with > 1 Clump	469	140	89
Scaffods with clumps to >1 human chromosome	181	62	6
Scaffolds with >2 clumps	51	18	14
Scaffolds with clumps in different orientations on 1 chr	193	48	50
Other	98	33	31
Scaffold statistics for Large Clumps - shared breakpoints			
Confirmed by one or more other assemblies	189	182	172
Spanned by 1	131	37	15
Spanned by 2	678	90	9
Confirmed and Spanned	7	8	9
Unexplained	67	21	21
Total Scaffold breakpoints	1058	322	208

Notes

a. Different methods were used to count Ns in the gaps in the different assemblies

b. BAC libraries were not labeled in Celera assembly, and IMBGA libraries not labeled for Atlas-WGS and PCAP assemblies, so counts for library data are not complete.