Limitations of next-generation genome sequence assembly

Can Alkan, Saba Sajjadian & Evan E Eichler

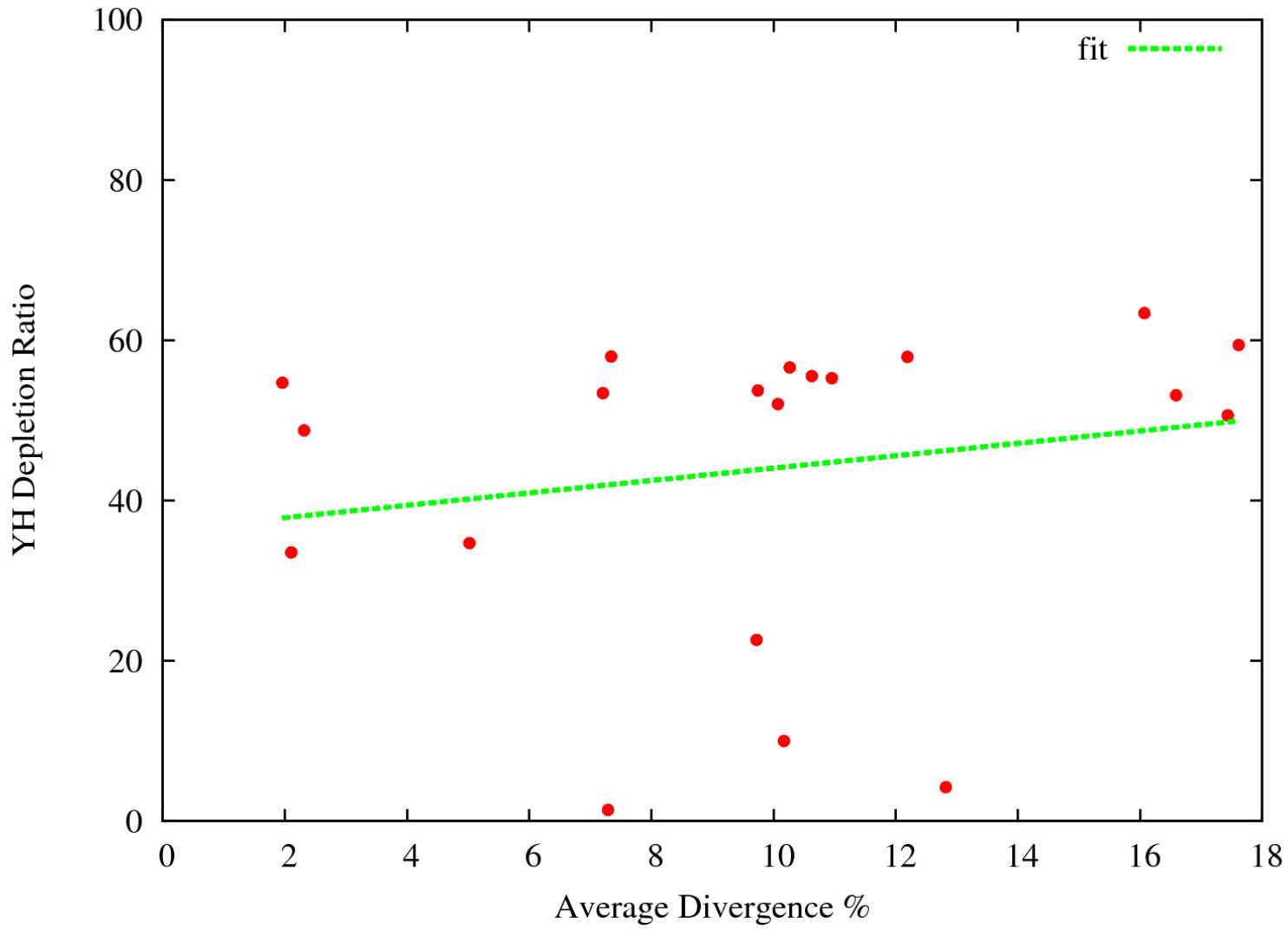
Note: Supplementary Tables 1 and 3–5 are available on the Nature Methods website.

Supplementary Figure 1	**Depletion pattern versus sequence divergence observed in Alu repeats in the YH genome**
Supplementary Figure 2	**Missing segmental duplications in base pairs**
Supplementary Table 2	**Analysis of repeat content in the YH genome compared to the reference human genome assembly NCBI Build 36**
Supplementary Note	**Supplementary Note**

Note: Supplementary Tables 1,3,4 and 5 are available on the Nature Methods website

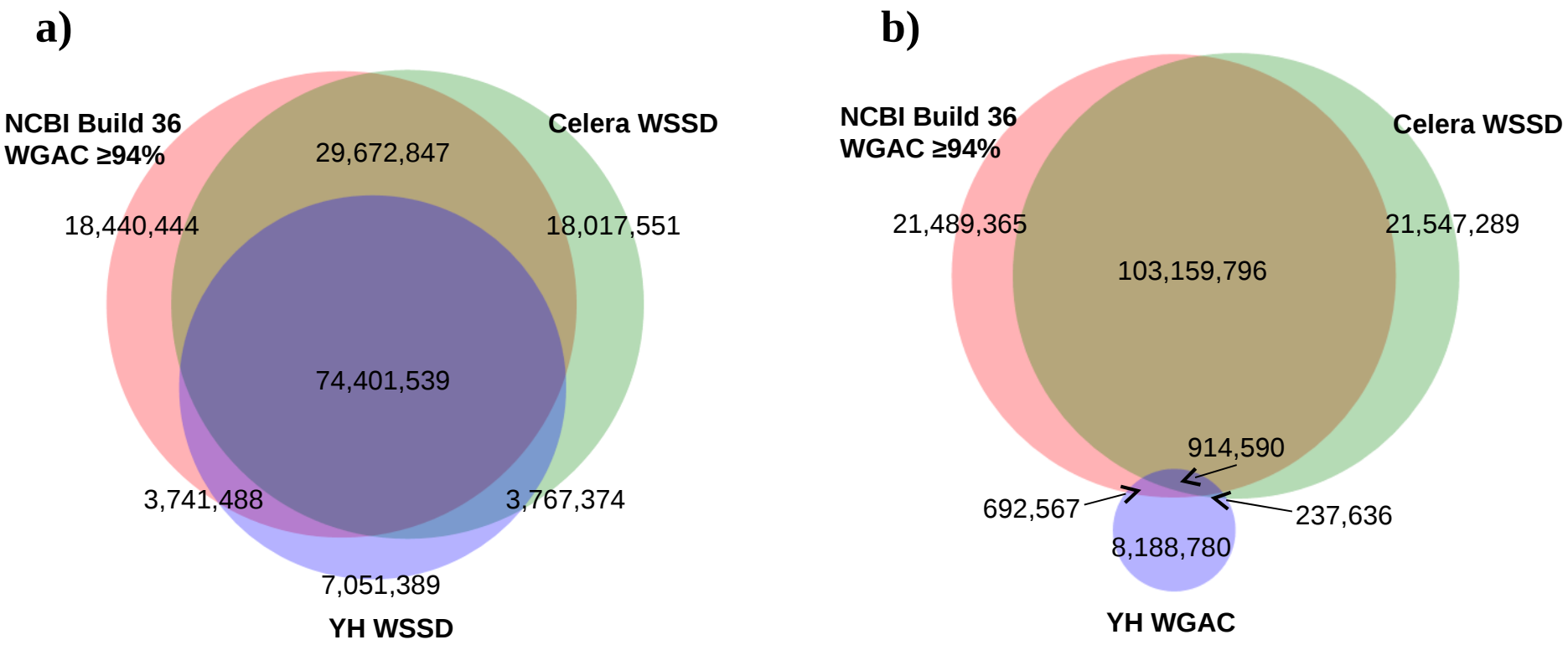
Supplementary Figure 1

Depletion pattern versus sequence divergence observed in Alu repeats in the YH genome.



Although the correlation is weak (R^2=0.02), the loss of low-divergence Alu repeats in the assembly is visible.

Supplementary Figure 2
Missing segmental duplications in base pairs.



Venn diagrams comparing duplication contents in base pairs. **(a)** Duplications in the reference genome (≥94% identity), Celera and YH WGS libraries share most of the duplicated base pairs. **(b)** Only 25% of the duplications detected in the YH assembly overlap with either the reference genome or Celera WGS. We converted duplications previously detected in the YH WGS library[23] from Build 35 coordinates to Build 36 using the UCSC liftOver tool. WGAC: whole-genome assembly comparison, WSSD: whole-genome shotgun sequence detection.

Nature Methods: doi:10.1038/nmeth.1527

Supplementary Table 2. Analysis of repeat content in the YH genome compared to the reference human genome assembly NCBI Build 36.

	NCBI Build 36					YH						
Repeat class	**Avg_div**	**Avg_del**	**Avg_ins**	**Total_bp**	**count**	**Avg_div**	**Avg_del**	**Avg_ins**	**Total_bp**	**count**	**Enrichment**	**Difference_BP**
DNA	25.578762	4.361067	2.285396	591,156	4,459	25.621592	4.320534	2.231502	590,216	4,470	NCBI	940
DNA/hAT	21.526989	5.757693	3.603512	1,983,418	13,815	21.59539	5.704643	3.018746	1,901,668	13,227	NCBI	81,750
DNA/hAT-Blackjack	24.314577	4.851852	2.356327	3,478,222	20,434	24.314377	4.804063	2.058326	3,384,902	19,971	NCBI	93,320
DNA/hAT-Charlie	23.455845	5.643371	3.184175	45,249,734	254,397	23.859253	5.551682	2.787262	42,969,478	243,605	NCBI	2,280,256
DNA/hAT-Tip100	24.952152	6.354397	2.99509	6,930,670	32,608	24.995386	6.272807	2.584993	6,727,299	31,559	NCBI	203,371
DNA/Merlin	15.022222	3.216667	3.003704	17,576	54	15.341665	3.508334	2.448333	16,451	60	NCBI	1,125
DNA/MuDR	16.119011	3.451966	2.81492	695,039	2,011	17.039833	3.532235	2.467925	654,466	1,886	NCBI	40,573
DNA/PiggyBac	13.197611	5.44331	2.006533	543,343	2,373	16.300703	5.143344	1.734046	474,882	2,150	NCBI	68,461
DNA/TcMar	27.251108	6.011425	3.050665	712,972	3,851	27.225538	5.965585	2.7242	684,141	3,682	NCBI	28,831
DNA/TcMar-Mariner	16.172951	3.363606	2.538522	2,857,068	16,470	17.530828	3.4256	2.037282	2,632,663	15,034	NCBI	224,405
DNA/TcMar-Pogo	26.122227	2.7	1.525	4,128	36	26.211115	2.455556	1.652778	4,075	36	NCBI	53
DNA/TcMar-Tc2	24.274256	6.753686	3.177402	1,670,808	8,072	24.235708	6.639272	2.768624	1,598,568	7,673	NCBI	72,240
DNA/TcMar-Tigger	19.30619	5.122066	2.685673	34,336,032	105,884	21.709957	4.951272	2.155558	31,744,172	103,893	NCBI	2,591,860
LINE/CR1	30.777571	6.545103	2.942751	10,140,812	58,235	30.712835	6.486411	2.816154	9,860,795	56,864	NCBI	280,017
LINE/Dong-R4	30.302248	5.588061	2.420897	115,259	536	30.309149	5.558792	2.313921	115,263	546	YH	4
LINE/L1	20.640287	5.233125	3.00023	510,191,638	958,128	23.386259	5.260641	2.542391	337,032,466	913,782	NCBI	173,159,172
LINE/L2	30.222355	6.791392	3.528672	99,428,971	458,766	30.196173	6.730637	3.376416	96,576,404	445,001	NCBI	2,852,567
LINE/Penelope	27.998148	4.896297	2.901852	10,812	54	28.013205	4.696227	2.875472	10,683	53	NCBI	129
LINE/RTE	31.99925	6.360174	3.08243	3,391,171	16,859	31.964676	6.320067	2.923941	3,305,552	16,451	NCBI	85,619
LINE/RTE-BovB	28.95499	5.067747	3.40113	69,497	620	28.974882	5.000823	3.393434	68,352	609	NCBI	1,145
Low_complexity	53.156448	0.265181	0.432638	15,794,261	326,026	54.353218	0.203444	0.31058	12,982,552	305,306	NCBI	2,811,709
LTR	28.92474	6.705166	2.92381	506,226	2,419	28.960777	6.577746	2.757112	492,246	2,341	NCBI	13,980
LTR/ERV	22.854744	5.653713	3.746978	192,071	579	22.922972	5.599472	3.315019	184,691	566	NCBI	7,380
LTR/ERV1	17.122652	4.9924	3.232221	84,094,558	176,564	20.40588	4.763459	2.706346	62,013,301	165,076	NCBI	22,081,257
LTR/ERVK	9.849643	2.620327	1.502536	8,889,569	10,597	19.048605	2.37136	1.13399	3,380,780	10,121	NCBI	5,508,789
LTR/ERVL	24.222309	6.632281	3.282236	57,943,642	160,933	24.96604	6.501039	2.967106	52,983,879	156,316	NCBI	4,959,763
LTR/ERVL-MaLR	21.143831	6.235567	3.273872	111,108,272	345,803	22.895668	6.064951	2.931131	90,995,341	325,937	NCBI	20,112,931
LTR/Gypsy	29.239372	6.661669	3.892048	3,812,553	18,755	29.248384	6.636292	3.778308	3,733,933	18,256	NCBI	78,620
Other	6.902519	2.86463	13.251629	3,698,327	3,896	12.13868	1.619812	3.499057	47,644	212	NCBI	3,650,683
RC/Helitron	27.213715	7.354219	2.46403	464,961	2,274	27.226831	7.319725	2.305311	451,870	2,185	NCBI	13,091
RNA	24.477697	3.913464	1.754359	128,873	780	25.032867	3.986139	1.667652	124,285	779	NCBI	4,588
rRNA	15.977732	2.070267	0.87818	176,370	1,769	16.946005	2.279471	0.804455	171,391	1,661	NCBI	4,979
Satellite	23.20871	2.834661	2.706113	3,158,639	3,958	24.18578	2.272773	2.248773	1,832,062	4,363	NCBI	1,326,577
Satellite/acro	18.440912	0.413636	0.184091	30,997	44	20.584618	0.761539	0.330769	2,927	13	NCBI	28,070
Satellite/centr	20.455177	2.280156	1.56904	8,603,107	2,283	21.348553	1.372787	0.666609	3,051,577	7,415	NCBI	5,551,530
Satellite/telo	21.400312	7.888328	3.896529	225,755	317	24.219797	5.069307	3.730197	73,143	202	NCBI	152,612
scRNA	10.639685	2.261153	1.257254	120,576	1,282	11.113765	2.323656	0.80078	105,694	1,154	NCBI	14,882
Simple_repeat	8.834008	0.643018	0.510548	24,247,492	406,084	9.405177	0.524413	0.40923	13,950,397	277,568	NCBI	10,297,095
SINE	29.499653	4.852304	2.822056	301,584	2,589	29.481789	4.880548	2.817913	296,556	2,529	NCBI	5,028
SINE/Alu	12.350141	1.368106	1.250857	307,690,702	1,179,962	13.677691	1.34242	1.088535	148,446,712	645,290	NCBI	159,243,990
SINE/Deu	28.362974	4.506756	3.099184	319,605	3,065	28.515886	4.481086	3.095412	317,821	3,009	NCBI	1,784
SINE/MIR	28.710352	6.457099	2.527495	82,229,854	585,225	28.691162	6.446703	2.387307	80,018,577	569,055	NCBI	2,211,277
SINE/tRNA	30.797533	4.593553	2.680593	254,456	1,829	30.867697	4.526232	2.722283	255,326	1,849	YH	870
snRNA	12.612516	1.372303	0.793474	342,867	4,337	13.068939	1.47209	0.681944	294,791	3,827	NCBI	48,076
srpRNA	16.540611	2.272955	1.584034	264,504	1,453	17.817299	1.889776	1.496473	262,404	1,956	NCBI	2,100
tRNA	12.23175	0.509411	0.80768	107,771	1,849	12.659088	0.555061	0.904456	88,003	1,571	NCBI	19,768
Unknown	28.203812	6.317936	2.658277	1,312,906	7,267	28.175659	6.289093	2.557417	1,282,866	7,113	NCBI	30,040

Supplementary Note

The YH genome and data acquisition

The human genome (NCBI build 36 arguably represents one of the most intensively studied and most well-assembled complex genomes produced to date. Its high quality provides a gold standard for comparison. The *de novo* sequence assemblies of the genomes of two human individuals (Han Chinese YH and Yoruba African NA18507) were recently completed using massively parallel next-generation sequencing[1]. We primarily focused our analysis on the YH genome because of the higher sequence coverage when compared to NA18507 (71X and 44X respectively), better assembly statistics (446 Kb vs. 62 Kb N50 scaffold size), and because we had experimentally validated its pattern of segmental duplications (YH). The YH genome assembly contains 2.8 Gbp of sequence assembled into 48,160 scaffolds and 136,926 unconnected contigs (2.37 Gbp without scaffold gaps).

We downloaded the *de novo* sequence assembly generated from the Han Chinese individual (YH)[1] from the main project website (http://yh.genomics.org.cn; retrieved on January 27, 2010). This version of the assembly is listed as "Scaffold (+9.6 kb PE)" in the related publication describing this data[1]. The aforementioned paper also lists another version that reads "Contig after gap closure", however we note that many of the contigs within this next-generation sequence assembly are not actually defined but are represented by N's.

In addition, to further test for contamination, we downloaded the human novel insertion sequences (both YH and NA18507) reported in[2] from the same site.

Repeatmasking

We repeat masked both the human reference genome (NCBI build36) from the UCSC Genome Browser (http://genome.ucsc.edu) and YH genome assembly using the RepeatMasker tool[3] (version 3.2.9) with sensitive masking option enabled and species parameter set to "human" (RepeatMasker –s –species 'human'). The tabular output files of RepeatMasker were also used to analyze the repeat content difference between the YH genome and the reference human genome assembly (Supplementary Table 2).

Contamination discovery

Using MegaBLAST[4] with default options, we searched the repeat masked contigs and scaffolds of the YH genome assembly in the NCBI nucleotide (*nt*) database. Since the shortest contig length is 100 bp, we required ≥80 bp alignment with ≥90% sequence identity. We found evidence for contamination from: *Oryza sativa*, *Zea mays*, *E. coli*, *S. pombe*, *Fusarium oxysporum*, *Penicillium sp.*, *Sorghum bicolor*, *Triticum aestivum*, *Artemisia annua*, *Aegilops tauschii*, *Agave ghiesbreghtii*, *Avena sativa*, *Salmonella enterica*, zebrafish, and various cloning vectors, fungi, and bacteria. We repeated the same analysis on the reported human novel insertion sequences from the same genome and the genome of a Yoruba African individual

NA18507[2] and discovered 152 Kb of the same contaminants in the YH set and 136.6 Kb of Epstein-Barr virus contamination in the NA18507 sequence set.

Contamination in the panda genome

To verify the effect of contamination in other sequencing projects, we repeated this experiment with the giant panda genome assembly[5] (GenBank ACTA00000000, downloaded from http://panda.genomics.org.cn/download.jsp on May 1, 2010). We used the repeat coordinates provided with the assembly to mask the repeats in the panda genome. We found 74 contigs and 2,224 scaffolds that contain 235 Kb of contamination. The main source for contamination for the panda genome assembly was zebrafish, and the other contributors were oryza sativa, zea mays, and various cloning vectors.

Analysis of segmental duplications

We used the whole-genome alignment comparison (WGAC) method to discover segmental duplications and their pairwise relationship in the YH genome[6]. The WGAC analysis initially excludes repeats from the scaffolds, and then defines all pairwise alignments using a modified version of MegaBLAST[4]. Repeats are inserted back into the candidate regions (≥1 Kb, ≥90% identity), and the realignments are performed through the Needleman-Wunsch algorithm[7].

A Golden Path (AGP) creation

We assigned locations to the duplicated scaffolds to generate a corresponding AGP. We first mapped the repeat masked scaffolds to the human reference assembly using MegaBLAST with default options. Due to both masked repeats and the scaffold gaps, many local alignments of the scaffolds were generated. These alignments are "stitched" together using the BEDtools[8] allowing for at most 10-Kb alignment breaks between ordered pieces of scaffold sub-alignments (mergeBed –d 10000). In the case of a scaffold mapping to multiple loci (because of duplication), we selected the longest alignments as the anchoring location. The map locations of these contigs and scaffolds are provided in Table S5.

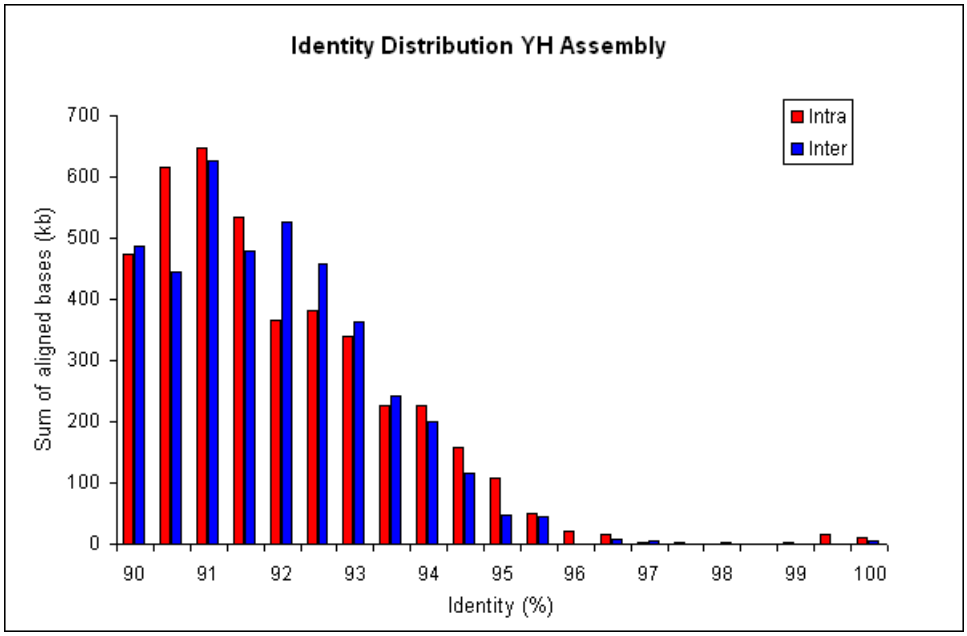
Inter vs. intrachromosomal duplications

To interrogate whether there is a bias in interchromosomal vs. intrachromosomal duplications, we assigned the duplicated YH scaffold chromosome locations in the reference genome assembly as described above and observed that 4.9/10 Mb (827 alignments) duplications were interchromosomal, where 5.18 Mb (825 alignments) were intrachromosomal.

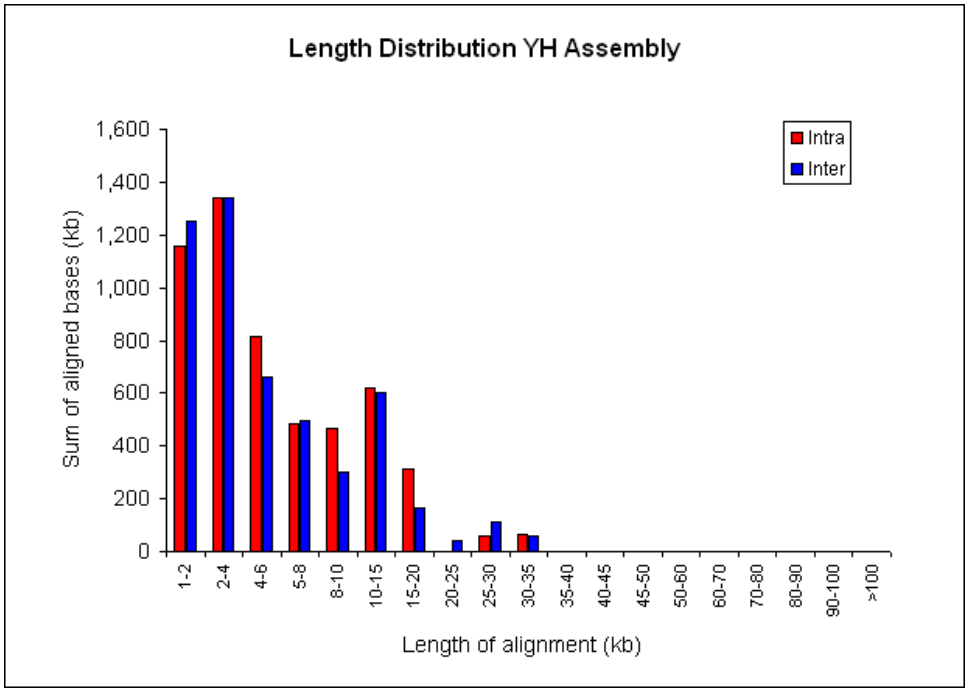
"Missing" duplications

In order to put a perspective of what is missing in terms of duplication, we analyzed the sequence identities of the pairwise alignments of duplication blocks. As expected, most of the missing

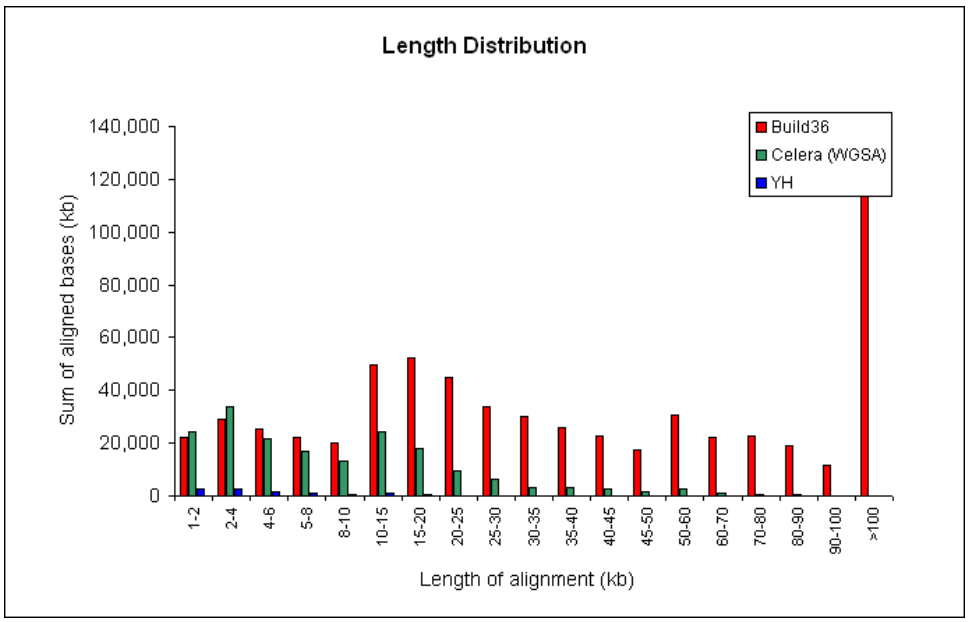
duplications (both intra and interchromosomal) are of high pairwise alignment identity (>93%) (Supplementary Note Figure 1). Similar calculation in NCBI Build36 and Celera WGSA confirms the depletion of high-identity duplications in whole-genome shotgun sequencing based approach, where the effect is more dramatic in the YH genome than Celera due to shorter read and insert sizes (Figure 1C). In addition, the segmental duplications detected in the YH genome tend to be smaller in size, where no duplication blocks > 31.8 Kb are detected (Supplementary Note Figure 2). Comparison with the segmental duplication lengths in NCBI build36 confirms the loss of longer duplication blocks in both Celera and YH assemblies (Supplementary Note Figure 3).



Supplementary Note Figure 1. The pairwise sequence identity distribution in the YH WGAC analysis shows the assembly bias against segmental duplications is more severe in higher sequence identity (≥96) and the distributions are similar in both intra and interchromosomal duplications.



Supplementary Note Figure 2. The effect of block length in discovering segmental duplications in the YH genome.



Supplementary Note Figure 3. Although the NCBI build36 contains duplications >100 Kb, the largest duplication in the Celera WGSA is <90 Kb where YH genome shows no duplication block >31.8 Kb.

Simple gene table analysis

Our gene analysis began with coding RefSeq transcripts, located on autosomes. We previously constructed a nonredundant set of genes (n=17,601) in human reference genome build35[9]. For this study, we converted the genomic locations to build36 using the liftOver tool.

Gene coverage calculation

To estimate the sequence coverage of the genes in the YH assembly, we mapped the repeat masked contigs and scaffolds to the human reference genome using MegaBLAST with default options. We then filtered the map output to remove alignment with less than 98% sequence identity and merged the resulting map locations in the reference genome using BEDtools[8]. Finally, we intersect the "mapping intervals" with the genomic locations of the nonredundant gene set. 9/17,601 genes (*HES*, *HLA-C*, *CTAGE6*, *CTAGE4*, *CHCHD9*, *LRRC26*, *KRTAP9-4*, *KRTAP10-5*, *LOC402057*) were *covered* at < 80 bp (≥ 24 bp), which we removed from intersection considering as spurious hits. Note that, in this calculation, the fragmentation of a gene to multiple contigs/scaffolds is not considered; all partial alignments are merged. We calculate the *coverage percentage* of the genes over the unmasked base pairs, i.e.:

$$Coverage\ percentage\ (gene_i) = covered\_bp(gene_i) / (\ length\ (gene_i) - repeat\_length(gene_i)\ )$$

We found that 35 full-length genes were not represented in the YH genome, where 23/35 of these genes correspond to the segmental duplications (Table S3). Among these genes, *LCE3B* and *LCE3C* are known to be common CNVs and were previously found to be deleted in the YH genome[9]. In addition, *GSTT1* is reported as a possible partial deletion[9]. 10/35 unrepresented genes were predicted to have > 30 copies in the YH genome (*NPIP*, *DUX4*, *DUB3*, *REXO1L1*, *FAM90A7*, *WBSCR19*, *LOC442590*, *MGC119295*, *LOC650293*, *PPIAL4)*. Furthermore, 48 genes could be mapped to the YH assembly with only ≤ 1% of their length (> 80 bp, 45/48 within segmental duplications). We observed that 9,909/17,601 (56.3%) genes had sufficiently high coverage (≥ 95%).

Gene fragmentation estimation

In order to assess the extent of gene fragmentation, we first extracted RefSeq genes (n=17,601) from the repeat masked human reference genome using BEDtools[8]. We then mapped the gene sequences to the YH genome assembly through MegaBLAST and discarded alignments with <80 bp and ≤ 98% sequence identity. Note that alignments could be fragmented due to the masked repeats within the same scaffold, and there might be short repeats missed by RepeatMasker that cause subsequence of some genes to be mapped to multiple scaffolds. To prevent such bias in our fragmentation estimation, we computed the minimum number of "containing scaffolds" by a method similar to the approximate solution for the *set-cover*[10,11] problem. For a gene G_i that has MegaBLAST hits to a scaffold S_j, we denote the number of basepairs in G_i that align to S_j by $Cover_{Gi}(S_j)$. First, for each gene G_i, we sort the MegaBLAST output in descending order of the $Cover_{Gi}(S_j)$ values. Then we greedily pick the scaffolds in this order, and we mark a scaffold hit "necessary" for a gene if and only if the genic interval contained by the new scaffold was not

previously covered by any of the scaffolds picked before that scaffold. In this way, we approximately minimize the number of "covering scaffolds" for a gene. Finally, we count the number of "necessary scaffolds" for each gene to calculate the gene fragmentation.

1. **Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* 20, 265-272 (2009).**
2. **Li, R. *et al.* Building the sequence map of the human pan-genome. *Nature biotechnology* 28, 57-63 (2009).**
3. **Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker Open-3.0. (1996-2004).**
4. **Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7, 203-214 (2000).**
5. **Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311-317 (2009).**
6. **Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. & Eichler, E.E. Segmental duplications: organization and impact within the current human genome project assembly. *Genome research* 11, 1005-1017. (2001).**
7. **Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48, 443-453 (1970).**
8. **Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* 26, 841-842 (2010).**
9. **Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature genetics* 41, 1061-1067 (2009).**
10. **Karp, R.M., in *Complexity of Computer Computations*, edited by J.W.T. R. E. Miller (Plenum, New York, 1972), pp. 85-103.**
11. **Vazirani, V.V., *Approximation algorithms*. (Springer, 2001).**