Limitations of next-generation genome sequence assembly

Can Alkan, Saba Sajjadian & Evan E Eichler

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 YH genome compared to NCBI Build 36
Supplementary Note

Note: Supplementary Tables 1 and 3–5 are available on the Nature Methods website.
<table>
<thead>
<tr>
<th>Supplementary Figure 1</th>
<th>Depletion pattern versus sequence divergence observed in Alu repeats in the YH genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplementary Figure 2</td>
<td>Missing segmental duplications in base pairs</td>
</tr>
<tr>
<td>Supplementary Table 2</td>
<td>Analysis of repeat content in the YH genome compared to the reference human genome assembly NCBI Build 36</td>
</tr>
<tr>
<td>Supplementary Note</td>
<td>Supplementary Note</td>
</tr>
</tbody>
</table>

*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 ($R^2 \approx 0.02$), the loss of low- divergence Alu repeats in the assembly is visible.
Supplementary Figure 2

Missing segmental duplications in base pairs.

(a) 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\textsuperscript{23} from Build 35 coordinates to Build 36 using the UCSC liftOver tool. WGAC: whole-genome assembly comparison, WSSD: whole-genome shotgun sequence detection.
**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 ghiubreghtii*, *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
NA18507² 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 assembly⁵ (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 genome⁶. The WGAC analysis initially excludes repeats from the scaffolds, and then defines all pairwise alignments using a modified version of MegaBLAST⁴. Repeats are inserted back into the candidate regions (≥1 Kb, ≥90% identity), and the realignments are performed through the Needleman-Wunsch algorithm⁷.

**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 BEDtools⁸ 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 WGS A 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 WGA C 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. 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. 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.:

\[
\text{Coverage percentage (gene)} = \frac{\text{covered bp(gene)}}{\text{(length (gene)} - \text{repeat_length(gene))}}
\]

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. In addition, GSTT1 is reported as a possible partial deletion. 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. 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 problem. For a gene Gi that has MegaBLAST hits to a scaffold Si, we denote the number of basepairs in Gi that align to Si by CoverGi(Si). First, for each gene Gi, we sort the MegaBLAST output in descending order of the CoverGi(Si) 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.