Limitations of next-generation genome sequence assembly

Can Alkan, Saba Sajjadian & Evan E Eichler

High-throughput sequencing technologies promise to transform the fields of genetics and comparative biology by delivering tens of thousands of genomes in the near future. Although it is feasible to construct *de novo* genome assemblies in a few months, there has been relatively little attention to what is lost by sole application of short sequence reads. We compared the recent *de novo* assemblies using the short oligonucleotide analysis package (SOAP), generated from the genomes of a Han Chinese individual and a Yoruban individual, to experimentally validated genomic features. We found that *de novo* assemblies were 16.2% shorter than the reference genome and that 420.2 megabase pairs of common repeats and 99.1% of validated duplicated sequences were missing from the genome. Consequently, over 2,377 coding exons were completely missing. We conclude that high-quality sequencing approaches must be considered in conjunction with high-throughput sequencing for comparative genomics analyses and studies of genome evolution.

The plummeting costs and massive throughput of second-generation sequencing platforms are paving the way for *de novo* sequencing applications to characterize the genomes of thousands of species. Recently, researchers from the Beijing Genome Institute sequenced the cucumber genome using both capillary sequencing and Illumina technology[1], and the panda genome was the first mammalian genome to be assembled using sequence data generated solely using next-generation sequencing (NGS) platforms[2]. An international consortium of scientists proposes more ambitious projects, such as the Genome 10K Project to sequence the genomes of 10,000 vertebrate species[3]. The information obtained from sequencing these genomes will help us better understand genome evolution, providing rapid access to gene models from many more organisms than previously anticipated. However, a critical assessment of NGS genome assemblies should be performed in comparison to known standards, and a robust classification of what is missing from the assemblies needs to be taken into account. Such analyses are essential to correctly perform comparative genomics studies. Typical genome assembly standards, such as complete cDNA libraries or sequences from large-insert genomic clones that sample the genome do not yet exist for newly sequenced genomes such as the cucumber and panda and are unlikely to be generated to test the assembly quality. We can, however, compare the recently generated *de novo* sequence assemblies of two human individuals[4] with the human reference genome[5,6] to assess the limitations of such genomes assembled primarily with short reads. Here we present a formal analysis of the *de novo* sequence assembly generated from the genome of a Han Chinese individual (YH; **Supplementary Note**) using the Illumina platform[4] with an emphasis on the repeats and segmental duplications that cover approximately half of the human genome. In addition, we analyzed the new sequences discovered from the genome of another individual (Yoruban from Ibadan, Nigeria; NA18507) to test the utility of *de novo* assemblies in the characterization of new sequence insertions.

Sequence properties and algorithmic challenges

NGS technologies typically generate shorter sequences with higher error rates from relatively short insert libraries[7,8]. For example, one of the most commonly used technologies, Illumina's sequencing by synthesis, routinely produces read lengths of 75–100 base pairs (bp) from libraries with insert sizes of 200–500 bp. It is, therefore, expected that assembly of longer repeats and duplications will suffer from this short read length. Similar to the whole-genome shotgun sequence

Department of Genome Sciences, University of Washington School of Medicine and Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to E.E.E. (eee@gs.washington.edu).

PUBLISHED ONLINE 21 NOVEMBER 2010; DOI:10.1038/NMETH.1527

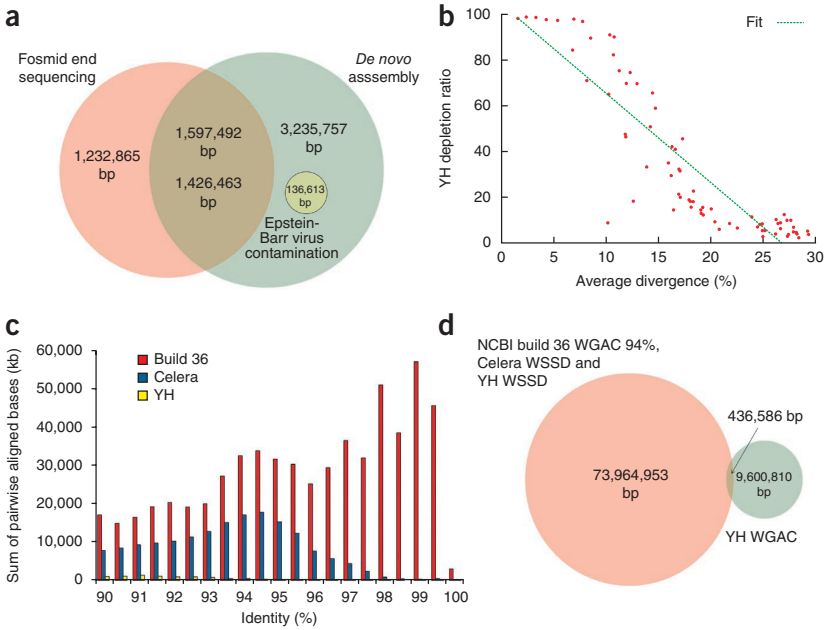

Figure 1 | Summary of *de novo* genome assembly and new sequence analysis. (**a**) Venn diagram comparing insertion sequences (total base pairs that do not exist in the reference genome build 36) detected by fosmid end sequencing[27] and *de novo* assembly[18] for the same genome (NA18507). The number of base pairs of Epstein-Barr virus contamination is also shown. Approximately 1.6 Mbp of new insertion sequence aligns with 1.42 Mbp detected by *de novo* assembly with NGS. (**b**) Average sequence identity of L1 common repeat sequences and depletion ratio in the YH genome assembly. (**c**) The pairwise sequence identity distribution of duplicated sequences in the YH genome compared to the human reference genome and a WGS assembly based on capillary sequence[24] (Celera). (**d**) Number of base pairs in segmental duplications detected in the YH assembly (YH WGAC) compared with duplications common to NCBI build 36 WGAC analysis (≥94% sequence identity) and read-depth analyses of the capillary-based (Celera) and YH (intersection of three datasets).

(WGS) assembly algorithms that use capillary-based data such as the Celera assembler[9], the predominant assembly methods for short reads are based on de Bruijn graph and Eulerian path approaches[10], which have difficulty in assembling complex regions of the genome. As argued by groups that presented various implementations of this approach (for example, the algorithms named EULER-USR[11], ABySS[12] and SOAPdenovo[4]), paired-end sequence libraries with long inserts help to ameliorate this bias. However, even the longest currently available inserts (<17 kilobases with Roche 454 sequencing[13]) are insufficient to bridge across regions that harbor the majority of recently duplicated human genes. Criticisms of WGS assembly algorithms and characterization of various types of errors associated with them as well as requirements for better assemblies have been previously discussed[14–16].

Contamination or new insertions?

An important consideration of any sequencing project, including those that use Sanger sequencing, is DNA contamination from other organisms. Before analyzing the genomes, we searched for possible contaminants by comparing the repeat-masked YH genome against the US National Center for Biotechnology Information (NCBI) nucleotide (nt) database[17] (**Supplementary Note**). We identified 1,033 contigs and 166 scaffolds (361 kbp) with high-identity matches to other species (**Supplementary Note**). Although this represents a small fraction of the total genome, nearly half the sequence (152 kbp) (**Supplementary Table 1**) was classified as new human sequences corresponding to ~15% of all reported insertion polymorphisms for YH (1,079 of 7,211 sequences or 3% of 5.12 Mb)[18] (**Supplementary Note**). Similarly, 2.8% (136.6 kbp of 4.8 Mbp) of the new sequences reported using the genome of a Yoruban individual (NA18507), likely represent contamination. The majority of these contaminations had high sequence identity to Epstein-Barr virus, an agent commonly used to immortalize cell lines (**Fig. 1a** and **Supplementary Table 1**). (Note that the NA18507 genome was sequenced using DNA from a cell line, whereas the YH genome sequence was generated from blood DNA.) Thus, although *de novo* sequence assemblies may be an important source for the discovery of insertion polymorphisms and are complementary to clone-based methods (**Fig. 1a**), such sequences require particular scrutiny and additional validation because of their tendency to enrich for contamination artifacts. Discriminating such sequences before sequence assembly becomes particularly problematic when the underlying sequence read data are short.

Repeat content

Any WGS-based *de novo* sequence assembly algorithm will collapse identical repeats, resulting in reduced or lost genomic complexity[14]. We compared the repeat content of the YH genome and the human reference genome (build 36) generating summary statistics for various repeat classes[19] (**Supplementary Table 2**). Although the repeat structure may vary between individuals, most retrotransposons are fixed in the human lineage[20], thus we would expect to observe a similar number of base pairs corresponding to retrotransposon-derived repeats in the genome of any human individual and the reference genome assembly. We identified 420.2 Mbp of missing common repeat sequence from the YH assembly corresponding to 173.6 Mbp of missing LINE1 (L1) and 159.2 Mbp of missing Alu repeats. As highly identical sequences will be more problematic, we quantified this effect by comparing this depletion as a function of sequence divergence. The depletion of repeat sequences was enriched in L1 classes with lower sequence divergence ($R^2 = 0.86$; **Fig. 1b**). We found that the depletion rose rapidly (>50%) for L1 repeat subfamilies where sequence identity exceeded 85%.

In general, most Alu subfamilies were underrepresented, but evolutionarily younger Alu repeats with higher identity to consensus sequences had high depletion rates although this trend was weak ($R^2 = 0.02$, **Supplementary Fig. 1**), likely because of the shorter sequence length of the Alu repeat class. Most common repeat classes showed reduced representation in the YH genome (**Supplementary Table 2**).

Table 1 | Summary of segmental duplication statistics

		NCBI build 36	Celera WGS assembly	YH genome assembly
Genome size (bp)		3,107,677,273	2,695,614,880	2,874,204,399
Nonredundant duplication space (bp)	Intrachromosomal	114,538,257	36,232,042	5,178,588
	Interchromosomal	74,560,372	32,383,828	4,891,680
	Total	159,204,446	58,887,898	10,034,278
Pairwise alignments	Intrachromosomal	9,245	7,080	1,652
	Interchromosomal	16,699	13,308	1,754
	Total	25,944	20,388	3,406

The YH genome assembly includes 496 Mb of scaffold gaps.

Segmental duplications

We used the whole-genome assembly comparison (WGAC) method[21] to analyze the segmental duplication content in the YH genome. Despite the fact that genomes typically contain 140.2 Mbp to 159.6 Mbp (25,914 pairwise alignments) of euchromatic segmental duplication[22], we detected only 10 Mb of segmental duplications (1,652 pairwise alignments) in the YH assembly (**Table 1**). Although the depletion becomes more pronounced with increasing sequence identity, the number of pairwise alignments was dramatically reduced (>90%) for all classes of duplication (**Fig. 1c**). This is in stark contrast to capillary sequencing–based WGS assembly, which recovered a substantial fraction of duplications with less than 95% sequence identity[22]. We previously constructed a duplication map of the YH genome using read-depth methods and validated copy-number differences using array comparative genomic hybridization[23]. We discovered 92 Mb of segmental duplications (>94% sequence identity) and found that the duplication content was similar to that of other human genomes (**Supplementary Fig. 2a**). We did not observe the common human duplication pattern within the YH genome *de novo* assembly (**Fig. 1d** and **Supplementary Fig. 2b**). If we limit our analysis to those duplications commonly present in the human reference genome and duplications we detected through read-depth analysis of a capillary sequencing–based WGS dataset[24] (Celera) and YH (total of 72 Mbp common duplications), we conclude that 99.4% of true pairwise segmental duplications were absent. We predict that 95.6% of the duplications in the YH *de novo* assembly are likely false because they did not correspond to duplications predicted by read depth or were not detected by array comparative genomic hybridization analysis[23] of the YH genome.

Missing and fragmented genes

Finally, we analyzed the impact of this genomic reduction on both gene coverage and fragmentation of genes into multiple scaffolds. We examined a nonredundant autosomal gene set (17,601 genes; **Supplementary Note**) and required ≥98% sequence identity between the assembly and the reference gene set. (At the exon level, we found that 93% of all coding exons (159,621 of 171,746 exons) were completely represented in the YH assembly. At the gene level, however, only 56.3% of the genes (9,909 of 17,601 genes) had sufficient representation in the assembly (≥95% of the gene). Not surprisingly, among the 2,377 protein-coding exons that were completely absent, 47.7% (1,133 of 2,377 exons) mapped to segmental duplications (**Supplementary Tables 3** and **4**), representing a tenfold enrichment of duplicated sequence. Although these losses would prevent appropriate annotation of at least 1,112 genes, we also noted 83 genes for which all exons were completely missing or had less than 1% of their protein-coding sequence represented. Of these genes, 81.9% (68 of 83 genes) corresponded to members of duplicated gene families, many of which are high in copy number in the YH genome, as we previously characterized[23] (**Supplementary Tables 3** and **4**).

The analysis described above did not consider gene fragmentation (that is, parts of the same gene represented in different scaffolds). The presence of duplicated and repetitive sequences in introns complicates complete gene assembly and annotation, leading to genes being broken among multiple sequence scaffolds. To test for this effect of gene fragmentation, we calculated the minimum number of scaffolds in the YH *de novo* assembly required to reconstruct every human gene according to the reference genome (**Supplementary Note**). We found that 69.7% of the genes (12,268 of 17,601 genes) are contained in a single scaffold. Among the fragmented genes (those mapping to two or more scaffolds), we found that 42% intersect with segmental duplications (1,779 of 5,291 genes) or map to regions in which repeat content exceeded >50% (1,582 of 5,291 genes) (**Supplementary Table 3**). Of 11,766 nonfragmented genes with all protein-coding exons present (**Supplementary Table 3**), 255 were shuffled in their respective scaffolds (that is, the exons were out of order). We observed that 29 genes were fragmented into >100 scaffolds and most (93%) corresponded to duplicated genes (**Table 2** and **Supplementary Table 3**). Among the most shattered genes with more than 200 scaffolds were two genes (*HYDIN2* and *PRIM2*) that have high-identity segmental duplications in YH[23,25]. Although *HYDIN2* was not present in the NCBI build 36 assembly, it is now partially represented in GRCh37 human genome assembly but not assigned to a chromosomal location.

Outlook

This is a watershed moment in genomics. Although data production capabilities are substantially improved, accurately building genome assemblies and correctly annotating them remains challenging, especially among complex genomes with higher repeat and duplication content. The *de novo* assembly of the YH genome coupled with experimental validation of its duplication and repeat content allow us to quantify this effect. Other than contaminating sequence, the most noticeable casualties of a *de novo* NGS assembly are segmental duplications and larger common repeats. We found that this depletion became acute when the sequence identity exceeded 85% resulting in the loss of ~16% of the genome. This is a more considerable bias when compared to capillary sequencing–based WGS assembly of the human genome in which sequence misassembly and collapse occurred for only ~8% of the genome when duplications or repeats exceeded 95% sequence identity. In the absence of alternative NGS-based human genome assemblies with different algorithms, we cannot test the effects of assembly method, but we believe that the limitations we present in this work are due to the properties of the data and whole-genome shotgun sequencing approach in general, rather than algorithmic inefficiency.

Without complementary efforts to fully sequence complex genomes, the field of comparative genomics may face a crisis.

Table 2 | Top 20 most-fragmented or missing genes in the YH genome assembly

Chromosome	Start	End	Length (bp)	Gene symbol	Transcript name	Copy number	Type	Fragments	Coverage (%)	Duplicated[a] (%)
16	69398790	69822070	423,280	*HYDIN*	NM_032821	3.47	Fragmented	215	95.82	94.12
6	57290381	57621334	330,953	*PRIM2*	NM_000947	3.87	Fragmented	213	82.30	98.27
9	39062766	39278300	215,534	*CNTNAP3*	NM_033655	4.65	Fragmented	208	84.92	100
5	21786731	22889488	1,102,757	*CDH12*	NM_004061	1.87	Fragmented	184	95.86	50.91
11	87877393	88438782	561,389	*GRM5*	NM_001143831	2.11	Fragmented	162	90.40	67.86
7	66099252	66341931	242,679	*TYW1*	NM_018264	3.27	Fragmented	155	82.94	99.61
10	50696330	51041337	345,007	*PARG*	NM_003631	4.17	Fragmented	154	57.14	80.96
1	143663118	143787436	124,318	*PDE4DIP*	NM_022359	7.4	Fragmented	147	93.31	100
7	153380709	154316928	936,219	*DPP6*	NM_130797	1.93	Fragmented	146	80.46	38.63
1	120255701	120413799	158,098	*NOTCH2*	NM_024408	2.97	Fragmented	142	95.26	68.52
8	7408721	7427585	18,864	*FAM90A7*	NM_001136572	36.03	Missing	0	0	100
16	14938801	14953432	14,631	*NPIP*	NM_006985	30.73	Missing	0	0	100
7	76506733	76520291	13,558	*LOC100132832*	NM_001129851	19.82	Missing	0	0	100
8	12327497	12338223	10,726	*FAM86B2*	NM_001137610	20.82	Missing	0	0	100
15	80895765	80905166	9,401	*LOC440295*	NM_198181	27.21	Missing	0	0	100
7	74962235	74971564	9,329	*LOC442590*	NM_001099435	32.2	Missing	0	0	100
7	44007014	44016247	9,233	*WBSCR19*	NM_175064	32.98	Missing	0	0	100
10	135333669	135341873	8,204	*DUX4*	NM_033178	195.66	Missing	0	0	100
22	22706139	22714284	8,145	*GSTT1*	NM_000853	0.44	Missing	0	0	44.20
8	86755947	86762978	7,031	*REXO1L1*	NM_172239	134.92	Missing	0	0	100

[a]Includes RepeatMasker data, whole-genome assembly comparison, Celera whole-genome shotgun sequence detection and YH whole-genome shotgun sequence detection.

There is the problem that although the genomes of many more species are now accessible, the portion of each genome that can be reliably accessed has diminished substantially (<80%). Such biases are ironically transforming our definition of what it means to sequence a genome and threaten to skew our understanding of organismal biology and genome evolution. Third-generation technologies, which increase read length or library insert sizes, promise to alleviate this deficit, but the issue is fundamentally greater than a technological gap. The expertise and motivation to sequence genomes to a high quality are disappearing. Large-insert clone library resources such as bacterial artificial chromosomes, required for accurate assembly of the human genome, were once a mainstay of genome sequencing projects but are now considered too costly to create or maintain for many organisms. Moreover, if 'genome manuscripts' can now be published without accounting for the 20% that is missing, what incentive remains to spend the additional effort and cost to sequence these genomes well? Such biases can be minimized when the genome of a closely related species finished with high-quality, clone-based sequencing is available (such as closely related nonhuman primate genomes compared against the reference human genome assembly). The problem is exacerbated when analyzing genomes without a reference index genome. In these cases, the portions that are missing or misassembled cannot be readily inferred and are invisible to the biologist. Biases against duplications and repeats, as well as fragmentation, raise questions related to the accuracy and completeness of similarly assembled genomes such as the panda genome[2], as recently discussed[26]. It is the responsibility of the scientific community to enforce standards of quality that can be measured and assessed. In our opinion, it is critical to develop new hybrid sequencing approaches, such as multiplatform strategies including the third-generation long-read technologies, high-quality finished long-insert clones and new assembly algorithms that can accommodate these heterogeneous datasets. The genome assemblies themselves must be experimentally validated. Large-molecule, high-quality sequencing should not be abandoned until the balance between quantity and quality of genomes has been reestablished.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank E. Karakoc and P. Sudmant for helpful discussions, T. Marques-Bonet and J.M. Kidd for providing the nonredundant gene table, and T. Brown for proofreading the manuscript. This work was partly supported by US National Institutes of Health grant HG002385 to E.E.E. E.E.E. receives funds as an Investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

C.A. and E.E.E. conceived the study and wrote the manuscript. C.A. and S.S. analyzed the data.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Huang, S. *et al.* The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **41**, 1275–1281 (2009).
2. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
3. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.* **100**, 659–674 (2009).
4. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
5. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
6. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
7. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
8. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
9. Myers, E.W. *et al.* A whole-genome assembly of *Drosophila*. *Science* **287**, 2196–2204 (2000).
10. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**, 9748–9753 (2001).
11. Chaisson, M.J., Brinza, D. & Pevzner, P.A. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res.* **19**, 336–346 (2009).

12. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
13. Schuster, S.C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
14. Green, P. Whole-genome disassembly. *Proc. Natl. Acad. Sci. USA* **99**, 4143–4144 (2002).
15. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
16. Meader, S., Hillier, L.W., Locke, D., Ponting, C.P. & Lunter, G. Genome assembly quality: assessment and improvement using the neutral indel model. *Genome Res.* **20**, 675–684 (2010).
17. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
18. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
19. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
20. Mills, R.E., Bennett, E.A., Iskow, R.C. & Devine, S.E. Which transposable elements are active in the human genome? *Trends Genet.* **23**, 183–191 (2007).
21. 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 Res.* **11**, 1005–1017 (2001).
22. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
23. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
24. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
25. Doggett, N.A. *et al.* A 360-kb interchromosomal duplication of the human *HYDIN* locus. *Genomics* **88**, 762–771 (2006).
26. Worley, K.C. & Gibbs, R.A. Genetics: decoding a national treasure. *Nature* **463**, 303–304 (2010).
27. Kidd, J.M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods* **7**, 365–371 (2010).