Shotgun sequence assembly and recent segmental duplications within the human genome

Xinwei She[1], Zhaoshi Jiang[1], Royden A. Clark[2], Ge Liu[2], Ze Cheng[1], Eray Tuzun[1], Deanna M. Church[3], Granger Sutton[4], Aaron L. Halpern[5] & Evan E. Eichler[1]

[1]*Department of Genome Sciences, University of Washington School of Medicine, 1705 NE Pacific Street, Seattle, Washington 98195, USA*
[2]*Department of Genetics, Case Western Reserve University, Cleveland, Ohio 44106, USA*
[3]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA*
[4]*Applied Biosystems, 45 West Gude Drive, and* [5]*The Center for the Advancement of Genomics, 1901 Research Boulevard, Suite 600, Rockville, Maryland 20850, USA*

Complex eukaryotic genomes are now being sequenced at an accelerated pace primarily using whole-genome shotgun (WGS) sequence assembly approaches. WGS assembly was initially criticized because of its perceived inability to resolve repeat structures within genomes. Here, we quantify the effect of WGS sequence assembly on large, highly similar repeats by comparison of the segmental duplication content of two different human genome assemblies. Our analysis shows that large (>15 kilobases) and highly identical (>97%) duplications are not adequately resolved by WGS assembly. This leads to significant reduction in genome length and the loss of genes embedded within duplications. Comparable analyses of mouse genome assemblies confirm that strict WGS sequence assembly will oversimplify our understanding of mammalian genome structure and evolution; a hybrid strategy using a targeted clone-by-clone approach to resolve duplications is proposed.

The optimal method to generate assembled genomic sequence data for the scientific community has been a matter of considerable debate[1,2]. Public efforts originally advocated strict clone-order-based approaches, citing logistical limitations, computational issues and an unknown genome structure as arguments against a WGS approach. In the case of the human genome, the clone-ordered approach involved the sequencing of large insert genomic clones (>100 kilobases (kb)) derived from a physical map generated before sequencing. This effectively reduced the genome project into a collection of local projects ($n = 45{,}000$) that could be subsequently assembled into a final genome sequence. The alternative WGS sequencing approach involved random sequencing of a large collection ($n = \sim 27{,}000{,}000$) of clones of various insert size as a single project. It assembled the genome sequence 'on the fly' based on sequence overlap and paired end-sequence linker information. Private initiatives demonstrated the efficacy of WGS sequence assembly to generate rapidly draft versions of eukaryotic genomes[3,4], although the inclusion of public clone-ordered data complicated interpretation of its power as a stand-alone strategy[5]. Since that time WGS assembly (WGSA)-based approaches have become widely adopted within the sequencing community and are now the predominant component of most publicly funded genome projects[6,7]. Despite their general acceptance, the impact of such strategies on our understanding of genome biology is not well understood.

Recently, two independent assemblies of the human genome were released—one based largely on clone-ordered sequence (build34) and the other based on exclusive use of WGSA data. This landmark event provides the first opportunity to compare two distinct genome assembly approaches[8,9]. It should be pointed out that both assemblies have matured over multiple rounds of reiteration. The current finished build34 benefited from several years of hand curation and experimental validation from a large number of genome annotators. Similarly, the WGSA was generated by an assembler that was enhanced after algorithmic improvements introduced during the Celera mouse assembly[8]. We present a detailed study of the organization and structure of segmental duplications within these two assemblies. These results have important implications not only in directing and improving future genome assemblies, but, more importantly, in providing insight into how whole-genome sequence can be meaningfully interpreted by the biological community.

Segmental duplications and human assembly comparison

Both working-draft and WGS sequence assemblies have had difficulties resolving the structure of large, highly identical duplications[10–13]. We analysed recent segmental duplications (>90% identity, >1 kb in length) using methods that were developed during the analysis of the human genome[10,14]. Both experimental and *in silico* analyses initially suggested that 5–6% of human euchromatin is composed of segmental duplications[9,15]. Precise determination of the organization and structure required a high-quality genome assembly. Within the finished build34 genome we identified 150.8 megabases (Mb; 5.3%) of segmental duplications (Table 1) of which 140.2 Mb could be confirmed using an assembly-independent strategy[14], indicating that these were not artefacts of the build34 assembly (see Fig. 4b of ref. 9). Although this assembly represents a marked improvement from previous genome assemblies, gaps still remain particularly within duplication regions. Incremental improvements and increases in duplication content are expected. A more recent assembly of the human genome (build35), for example, captured an additional 2.0 Mb of duplicated sequence (Supplementary Table 1). A total of 98.7% of the duplication structure was identical between build34 and build35. In contrast to these, an independent analysis of the WGSA revealed a significantly reduced content of segmental duplications (60.3 Mb or 2.2% of the WGSA genome). The results of these duplication analyses including an overlay of WGSA on build34 are available in UCSC-browser format at http://humanparalogy.gs.washington.edu.

It had been predicted that duplications with the greatest degree of sequence identity would be the most difficult to resolve but, until

now, the threshold for this effect has been impossible to determine. For duplications with less than 95% identity there is a good correspondence between the two methods, although the WGSA shows fewer alignments at all bin sizes (Fig. 1a). At 95.5% a marked decline in the fraction of duplicated bases becomes apparent within the WGSA. As the sequence identity of duplications increases, the largest and most highly identical duplications disappear. We calculate that only ~9% (8.0 Mb out of the expected 94 Mb) of duplications whose sequence identity exceeds 97% are represented within the WGSA as duplications. Duplications that are virtually identical (>98%) appear to be completely absent within the assembly or take the form of apparent unique sequence composed of extremely short sequence alignments. The former corresponds to ~26 Mb of sequence (1,748 regions greater than 10 kb in length).

As expected, genes embedded within these segments are also conspicuously absent. We identified 67 genes that are partially deleted and another 36 genes that are completely absent from the Celera WGSA (Supplementary Table 2 and Supplementary Fig. 1). This set included rapidly evolving gene families such as nuclear-pore interacting protein (*NPIP*), sperm protein associated with nucleus (*SPANX*) and variable charge basic protein Y (*VCY*) gene families. In addition, cancer-related antigen markers (*GAGE, NY-REN*), both survival of motor neuron genes (*SMN1* and *SMN2*) and several important immune-related genes—interleukin 27 (*IL27*), neutrophil cytosolic factor 1 (*NCF1*) and epithelial beta defensins (*DEF*)—were lost from the WGSA owing to their association with segmental duplications.

Next, we analysed the length distribution of duplication alignments between the two genome assemblies. Build34 duplications were distributed within a total of 28,728 pairwise alignments whose length averaged 9.2 kb. WGSA duplications were less frequent (20,818 alignments) and shorter (4.04 kb). An analysis of the sum of aligned bases as a function of alignment length showed a marked depletion of longer alignments (>15 kb) within the WGSA when compared with build34 (Fig. 1b). The greatest discrepancy occurred among the largest alignments and pinpoints a failure of whole-genome assembly methods to traverse through such large, complicated repeat structures.

Large blocks of highly identical duplications are enriched near human centromeres and telomeres as well as specific focal regions within euchromatin. Not surprisingly, these regions are very poorly represented within the WGSA (Fig. 2; see also Supplementary Fig. 2). Such duplication regions are also frequently associated with genomic disease owing to non-allelic homologous rearrangement between intrachromosomal duplications. We examined the WGSA for five disease breakpoint regions (spinal muscular atrophy type I, Charcot–Marie–Tooth disease, velocardiofacial/DiGeorge syndrome, Prader–Willi syndrome and William's syndrome). Between 71–97% of the sequence corresponding to these large segmental duplications was absent (Supplementary Fig. 1). It follows that strict dependence on a WGSA approach would severely oversimplify the architecture of our genome and limit an understanding of the molecular aetiology of such diseases.

As a final assessment of segmental duplications between the assemblies, we analysed 19 duplicons whose copy number and distribution within the human genome had been experimentally validated by fluorescence *in situ* hybridization and/or hybridization data. We mapped 75 sequence tags corresponding to these 19 duplicons by BLAST sequence similarity searches against the two human genome assemblies (Supplementary Table 3). Within the finished build34 genome a total of 535 copies mapped to specific chromosome positions—in good agreement with experimental data ($n = \sim 580$ copies). Eleven mapped to positions within the unplaced or random sequence contigs. By comparison, only 240 discrete loci could be identified within the WGSA. Of these, 94 mapped to specific chromosomal positions, whereas the remainder localized to an unknown chromosome. We conclude that a minor fraction (~20%) of the duplicated sequence is correctly placed within the WGSA and that more than one-half of the duplications have been collapsed or lost.

Table 1 **Comparison of segmental duplication within two human genome assemblies**

Chromosome	Build34 assembly			WGSA		
	Length (bp)	Duplication (bp)	Fraction	Length (bp)	Duplication (bp)	Fraction
1	221,562,941	11,553,369	0.0521	209,662,503	2,629,537	0.0125
2	237,541,603	10,000,492	0.0421	223,960,456	2,034,342	0.0091
3	194,473,779	3,299,552	0.0170	189,481,828	2,626,848	0.0139
4	186,841,959	4,287,299	0.0229	180,981,699	2,899,296	0.0160
5	177,552,822	5,956,951	0.0336	170,281,266	1,351,471	0.0079
6	167,256,575	3,600,793	0.0215	161,428,330	1,660,626	0.0103
7	154,676,518	13,096,209	0.0847	144,247,908	5,496,523	0.0381
8	142,347,919	3,250,852	0.0228	136,878,554	1,013,574	0.0074
9	115,624,042	11,096,428	0.0960	104,630,165	1,826,794	0.0175
10	131,173,206	8,937,553	0.0681	122,948,635	3,379,280	0.0275
11	130,908,854	5,535,297	0.0423	126,253,176	4,007,704	0.0317
12	129,826,277	2,922,438	0.0225	125,900,476	2,050,906	0.0163
13	95,559,980	3,212,091	0.0336	92,484,206	1,953,930	0.0211
14	87,191,216	1,587,527	0.0182	84,198,821	951,062	0.0113
15	81,259,656	8,577,567	0.1056	74,059,970	2,599,187	0.0351
16	79,932,429	9,124,179	0.1141	66,369,068	994,790	0.0150
17	77,677,744	7,746,457	0.0997	73,627,628	3,657,946	0.0497
18	74,654,041	1,898,132	0.0254	71,253,215	370,517	0.0052
19	55,785,651	4,051,295	0.0726	51,679,110	2,835,683	0.0549
20	59,424,990	1,479,847	0.0249	57,238,069	963,199	0.0168
21	33,924,307	1,791,042	0.0528	31,584,736	410,918	0.0130
22	34,352,051	3,982,963	0.1159	31,357,605	1,590,197	0.0507
X	149,215,391	10,057,692	0.0674	121,809,144	2,105,297	0.0173
Y	24,649,555	12,745,541	0.5171	7,151,840	728,694	0.1019
Unplaced	2,592,022	980,700	0.3784	36,146,472	10,186,469	0.2818
Total	2,865,069,170	150,772,266	0.0530	2,695,614,880	60,324,790	0.0224

Segmental duplications (>90% sequence identity and >1 kb length) were calculated using the whole-genome assembly comparison method[10] for the finished human genome assembly (July 2003) and the whole-genome shotgun sequence assembly (WGSA)[8]. Due to the fragmentation of duplications within the WGSA, duplicated bases were calculated without welding across gaps in the assembly. Totals do not include gaps or centromeric/acrocentric regions of chromosomes. Both assemblies were compared using exactly the same parameters. The unplaced chromosome contains the largest proportion of WGSA duplicated sequence—28.2% (10.2 Mb based on the analysis of WGSA). Of the 21.8 Mb that could be mapped back to build34, we found that 9.2 Mb (42.3%) corresponded to duplications within our segmental duplication database.

Segmental duplications and WGSA chromosome length

The size of human euchromatin within build34 (2,865 Mb) is significantly larger than that predicted by the WGSA (2,696 Mb). This size difference is not uniformly distributed among chromosomes (Table 1; see also Supplementary Fig. 3). Although some of this lost euchromatin, 170 Mb, has been attributed to reduced coverage of the sex chromosomes as a result of the male donor in the WGSA[8], differences in the length of the sex chromosomes would only account for 44.9 Mb of sequence. As human chromosomes vary considerably in duplication content, we sought to determine whether there was a correlation between chromosomes that carry large blocks of duplications with high sequence identity and reduced chromosomal length (Supplementary Fig. 3). There is a strong correlation ($r = 0.9$) between the reduction in chromosome length and the number of highly identical duplication bases (Table 1). Chromosome 16 is most notable in this regard. This chromosome is reduced by 17% in the WGSA. It is also the autosome that has the largest fraction of highly identical segmental duplications (Table 1). We estimate that missing segmental duplications contribute more than 50% to the reduced size of the WGSA when compared with build34 (90 Mb out of the 170 Mb reduced size).

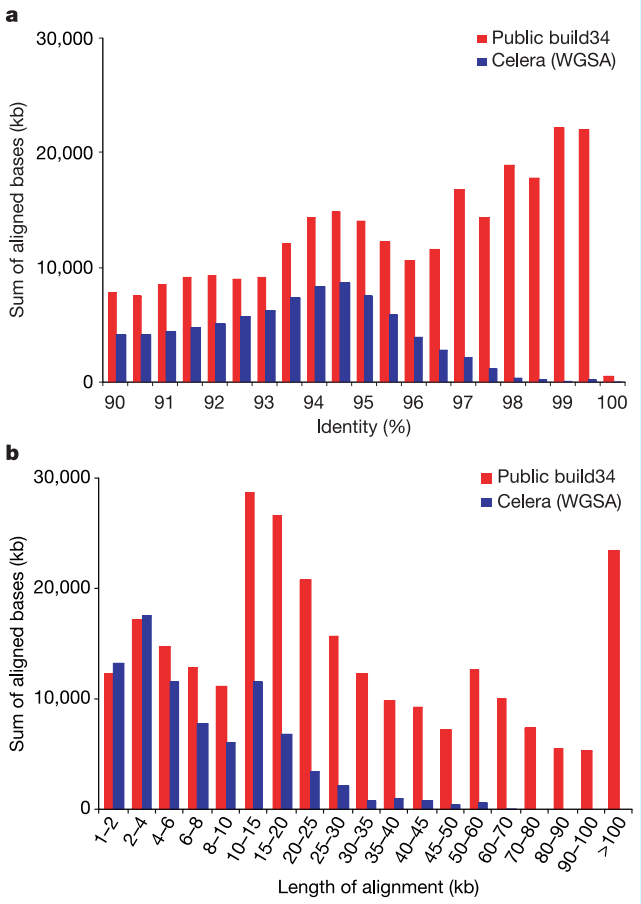

Figure 1 Sequence identity and alignment length of segmental duplications. **a**, **b**, All duplication alignments between 90–100% were categorized based on sequence identity (**a**) (0.5% bins) and the alignment length (**b**). The sum of aligned base pairs for each bin is compared between WGSA and build34 human genome sequence assemblies. The proportion of WGSA aligned base pairs begins to decline most rapidly as the sequence identity exceeds 96–97% and the length of the alignments exceeds 15 kb. Note that the reduction in WGSA alignments below 96% is probably due to the fact that divergent duplications are frequently part of larger alignments where the degree of sequence identity is higher. As highly identical alignments are lost, the embedded, more divergent pairwise alignments are also eliminated from further consideration.

Implications

Our analysis clearly shows that strict WGSA has limited capacity to resolve the structure of duplicated regions within genomes. Most of this effect, however, occurs among duplications that exceed >15 kb in length and show greater than 97% sequence identity. We predict that the largest, nearly identical duplications will be absent from WGS sequence assemblies. Clearly, different assembly algorithms[16,17] may perform better or worse than the Celera assembler—with a trade-off between ability to separate repeats and robustness in the face of polymorphism or sequencing error—but these thresholds provide a useful benchmark for future genome assembly comparison. This study has several important ramifications.

First, estimates of genome-wide duplication content between species should be tempered by the underlying method of assembly. Apparent differences in content may be a consequence of differences in genome assembly rather than a true biological effect. This may explain why other complex genomes that have been sequenced primarily by WGS sequencing show reduced recent duplication content when compared with human[13,18]. In such first-pass, whole-genome assemblies it is likely that duplications will be even more grossly underestimated. For example, we compared two mouse genome assemblies: one was assembled almost strictly by WGS sequencing (MGSCv.3.0) and the other was the most recent composite assembly (build33) where 57% of the mouse genome was assembled from large insert bacterial artificial chromosome (BAC) clones. The proportion of segmental duplication (>20 kb) increased by more than one order of magnitude between the two assemblies (Supplementary Table 4), where almost all of the increase (96%) was attributed to the incorporation of BAC-based sequence into the assembly.

Second, we can expect that euchromatin length will be underestimated on the basis of the misassembly of large, highly identical duplications. For the human genome sequence this effect accounts for more than 50% of the reduced genome length. It follows that organisms with greater duplication content will show greater reductions in size if WGSA is the only method applied.

Third, genes embedded within segmental duplications will be concomitantly lost. A surprising finding was that 37 duplicated gene segments were not represented even once within the assembly (Supplementary Table 2). This may be because of fracturing of the assembly due to conflicting overlap patterns and mate-pair conflicts, leading to complete rejection of these regions. In the case of the sequenced human genome, this included the loss of rapidly evolving lineage-specific gene families and genes associated with immune response and germline development.

Most importantly, an oversimplified view (Fig. 2) of the human genome structure emerges with strict WGSA. Regions enriched for duplications, such as pericentromeric and subtelomeric areas of chromosomes, are particularly under-represented (Supplementary Fig. 2). In addition, sites of recurrent chromosomal structural rearrangement associated with disease[19] and breakpoints in conserved synteny essentially disappear as a result of WGSA[20,21]. In the absence of BAC-based sequence we will forfeit an understanding of heterochromatic–euchromatic transition regions, potential mechanisms of chromosomal evolution and the molecular aetiology/origin of human genomic disease.

Hybrid strategy to sequence complex genomes

Although it is clear that the detailed clone-ordered approach is superior in the resolution of segmental duplications, it would be unrealistic to propose that the sequencing community should abandon WGSA-based approaches. These are the most efficient

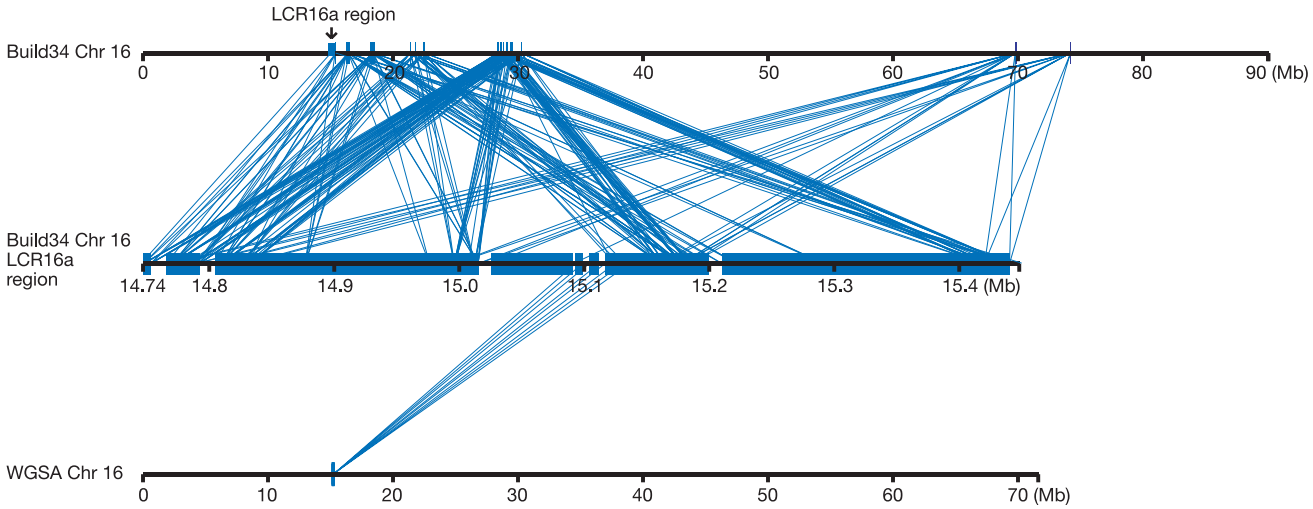

Figure 2 Distribution of LCR16a duplications in two assemblies. The pattern of duplication alignments for one 690-kb region of low-copy-repeat duplications on chromosome 16 is shown between the build34 and WGSA human genome assemblies. The entire region is duplicated to 28 distinct regions within build34 (locations have been experimentally verified) whereas only a small portion (46 kb) maps to a single location on WGSA chromosome 16.

and cost-effective means of capturing the bulk of euchromatic sequence. Segmental duplications, however, should not be considered as an acceptable casualty of this process. In humans, duplicated regions show high transcriptional content[22], are associated with disease[19] and large-scale copy number polymorphisms[23], and have played an important role in the chromosomal evolution of mammalian genomes[18,20]. Although the precise balance of clone-ordered sequencing and WGS sequencing during the assembly process has yet to be determined, the availability of two methods of genome assembly provides an important insight into this issue by refining the precise limitations of the WGS approach.

We propose a two-tier plan to ensure the resolution of such regions. During the first phase, WGS-based assemblies would be generated at sufficient depth (5–7-fold coverage) to provide an initial draft assembly of a genome. The same sequence reads could then be remapped to the assembly and analysed for regions of excessive divergence and excessive read depth as a means to detect sites of potential duplication[14]. Caution must be exercised to ensure that short sequence contigs are not completely excluded during this process, as those that do not originate from bacterial contamination often map to repetitive or duplicated portions of the genome. During the second phase, BACs corresponding to these regions of excess divergence and read depth would be selected based on BAC end sequence placement and submitted for further mapping/sequencing to establish long-range continuity across these regions. Sequence from these large-insert clones would be preferentially integrated into the WGSA. Retrospectively, on the basis of the known human genome structure, we estimate that 94 Mb of genomic sequence within ~380 regions of the human genome would require BAC-based sequence. This would entail the pre-selection of approximately 3,000 BACs (2,300 BACs at fourfold coverage plus an additional 700 BACs that span transition regions). Low-level draft sequence and/or high-density fingerprinting would reduce this set to a minimal tiling path for higher quality sequence. In theory, WGS sequencing (~sixfold genome coverage) coupled with final clone-order-based sequencing of ~1,500 BAC clones would be sufficient to represent accurately the true architecture of the human genome. □

Received 21 February; accepted 27 September 2004; doi:10.1038/nature03062.

1. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7,** 401–409 (1997).
2. Green, P. Against a whole-genome shotgun. *Genome Res.* **7,** 410–417 (1997).
3. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).
4. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
5. Waterston, R. H., Lander, E. S. & Sulston, J. E. More on the sequencing of the human genome. *Proc. Natl Acad. Sci. USA* **100,** 3022–3024 (2003) author reply 3025–3026.
6. Rat Genome Sequencing Project Consortium, Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
7. Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
8. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101,** 1916–1921 (2004).
9. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature* doi:10.1038/nature03001 (this issue).
10. 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).
11. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4,** R25 (2003).
12. International Human Genome sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
13. Tuzun, E., Bailey, J. A. & Eichler, E. E. Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res.* **14,** 493–506 (2004).
14. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
15. Cheung, V. G. *et al.* Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).
16. Huang, X., Wang, J., Aluru, S., Yang, S. P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13,** 2164–2170 (2003).
17. Batzoglou, S. *et al.* ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12,** 177–189 (2002).
18. Bailey, J. A., Church, D. M., Ventura, M., Rocchi, M. & Eichler, E. E. Analysis of segmental duplications and genome assembly in the mouse. *Genome Res.* **14,** 789–801 (2004).
19. Stankiewicz, P. & Lupski, J. R. Genomic architecture, rearrangements and genomic disorders. *Trends Genet.* **18,** 74–82 (2002).
20. Armengol, L., Pujana, M. A., Cheung, J., Scherer, S. W. & Estivill, X. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum. Mol. Genet.* **12,** 2201–2208 (2003).
21. Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. & Eichler, E. E. Hotspots of mammalian chromosomal evolution. *Genome Biol.* **5,** R23 (2004).
22. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424,** 157–164 (2003).
23. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.E.E. (eee@gs.washington.edu).