nature genetics

Genome assembly comparison identifies structural variants in the human genome

Razi Khaja[1], Junjun Zhang[1], Jeffrey R MacDonald[1], Yongshu He[1], Ann M Joseph-George[1], John Wei[1], Muhammad A Rafiq[1,2], Cheng Qian[1], Mary Shago[1], Lorena Pantano[3], Hiroyuki Aburatani[4], Keith Jones[5], Richard Redon[6], Matthew Hurles[6], Lluis Armengol[3], Xavier Estivill[3,7], Richard J Mural[8], Charles Lee[9], Stephen W Scherer[1] & Lars Feuk[1]

Numerous types of DNA variation exist, ranging from SNPs to larger structural alterations such as copy number variants (CNVs) and inversions. Alignment of DNA sequence from different sources has been used to identify SNPs[1,2] and intermediate-sized variants (ISVs)[3]. However, only a small proportion of total heterogeneity is characterized, and little is known of the characteristics of most smaller-sized (<50 kb) variants. Here we show that genome assembly comparison is a robust approach for identification of all classes of genetic variation. Through comparison of two human assemblies (Celera's R27c compilation and the Build 35 reference sequence), we identified megabases of sequence (in the form of 13,534 putative non-SNP events) that were absent, inverted or polymorphic in one assembly. Database comparison and laboratory experimentation further demonstrated overlap or validation for 240 variable regions and confirmed >1.5 million SNPs. Some differences were simple insertions and deletions, but in regions containing CNVs, segmental duplication and repetitive DNA, they were more complex. Our results uncover substantial undescribed variation in humans, highlighting the need for comprehensive annotation strategies to fully interpret genome scanning and personalized sequencing projects.

The most sensitive method for identifying all variation existing between two DNA donors is through direct comparison of accurately completed sequence assemblies of the genomes under study. For the human genome, there are two assembly products, one from the International Human Genome Sequencing Consortium (IHGSC)[4] and another from Celera Genomics[5], which used primarily clone-based sequencing and whole-genome shotgun sequencing, respectively. Although these assemblies have been evaluated for content and quality[6–10], little effort has been made to make use of their differences to annotate new sequence variants.

As both assemblies represent mosaics of different donor DNA sources (with neither being fully completed), they are not the ideal substrate for comparison, but we show that much valuable data can be extracted. Our premise was to perform a thorough comparison between Celera's most complete assembly and the IHGSC reference sequence, herein called R27c (WGA2) and National Center for Biotechnology Information (NCBI) Build 35, respectively. R27c contains 2,830,275,312 bp in 14,071 scaffolds, and Build 35 contains 3,094,710,260 bp with an estimated 345 annotated gaps.

Although R27c contains some Build 35 sequences[8], for this study we selected it over other Celera-only whole-genome shotgun assemblies (Build 35 also contains some Celera sequence). We rationalized that larger scaffolds would increase the likelihood of finding variants that might be missed using methods more sensitive to size restrictions, such as comparative genomic hybridization using arrays spotted with BAC clones (which has a lower limit of detection of ~50 kb)[11] or fosmid-end sequencing (which, in current form, does not identify variants <8 kb or insertions >40 kb)[3]. As Build 35 is the human reference and has a higher nucleotide content than R27c, we focus our discussion on the sequences present or variable when comparing R27c with Build 35, but we also performed the reciprocal analysis.

We used MegaBLAST[12] to align R27c to Build 35 and found 2,758,752,087 bp (97.5%) of matching sequence. We also used another alignment algorithm called A2Amapper[8,13] (**Table 1** and **Supplementary Table 1**). Then using the newly developed Genome Comparison Algorithm (GCA), we extracted variants between the assembly alignments. To reduce the potential for false positives owing to alignment

[1]Program in Genetics and Genomic Biology, The Hospital for Sick Children and Department of Molecular and Medical Genetics, University of Toronto and The Centre for Applied Genomics, MaRS Centre, Toronto, Ontario, M5G 1L7, Canada. [2]Department of Biosciences, Commission on Science and Technology for Sustainable Development in the South Institute of Information Technology (CIIT), Islamabad-44000, Pakistan. [3]Genes and Disease Program, Center for Genomic Regulation, Charles Darwin s/n, Barcelona Biomedical Research Park, 08003 Barcelona, Catalonia, Spain. [4]Genome Science Laboratory, Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo 153-8904 Japan. [5]Affymetrix, Inc., Santa Clara, California 95051, USA. [6]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [7]Pompeu Fabra University, Charles Darwin s/n, and National Genotyping Centre, Passeig Marítim 37-49, Barcelona Biomedical Research Park, Barcelona, Catalonia, Spain. [8]Windber Research Institute, 620 7th Street, Windber, Pennsylvania 15963-1331, USA. [9]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 20 Shattuck St., Boston, Massachusetts 02115, USA. Correspondence should be addressed to S.W.S. (steve@genet.sickkids.on.ca).

Received 20 July; accepted 13 October; published online 22 November 2006; doi:10.1038/ng1921

Table 1 Overview of alignment results comparing the Celera R27c assembly (2,830,275,312 nt) with the Build 35 assembly (3,094,710,260 nt) of the human genome sequence

	MegaBLAST	A2Amapper
Match (M)		
Total events	2,538,043	2,150,775
Total length (nt)	2,758,752,087	2,760,765,102
Shared length (nt)	2,746,748,012	2,746,748,012
Unique length (nt)	12,004,075	14,017,090
Mismatch (P)(1–10 nt)		
Total events	1,857,736	1,671,038
Total length (nt)	1,888,107	1,690,866
Shared length (nt)	1,613,458	1,613,458
Unique length (nt)	274,649	77,408
Unmatched (D)		
Total events	363,699	272,119
Total length (nt)	49,254,260	47,438,293
Shared length (nt)	35,318,386	35,318,386
Unique length (nt)	13,935,874	12,119,907
Gap (N)		
Total events	23,588	23,588
Total length (nt)	20,380,858	20,381,051
Shared length (nt)	20,380,764	20,380,764
Unique length (nt)	94	287
Total length (M+P+D+N) (in nt):	**2,830,275,312**	**2,830,275,312**

Alignment results are divided into four classes: match, mismatch, unmatched and gap. For each class, the number of events and the sequence content identified by each algorithm are shown. 'Shared length' refers to alignments identified and categorized the same way by both algorithms, whereas 'unique length' indicates what is uniquely identified by each algorithm. The total length of match, mismatch, unmatched and gap categories adds up to the total sequence content of the R27c assembly, indicating that all nucleotides have been accounted for. Results from the reciprocal analysis of Build 35 compared with R27c are shown in **Supplementary Table 1**. We note that during manuscript preparation, we compared the unmatched sequences present in R27c that are not in Build 35 against the recently released reference sequence (Build 36, March 2006) and found an additional 181 fragments totaling 829,890 bp, further supporting our data.

errors, we describe only those differences found by GCA in both the MegaBLAST and A2Amapper comparisons (**Table 2**). We grouped these differences into five classes: (i) small sequence mismatches (including SNPs), (ii) unmatched sequences (including insertions, deletions and CNVs), (iii) copy-unmatched sequences (a subset of ii), (iv) inversions and (v) internal assembly gaps (**Fig. 1** and **Supplementary Tables 2–4**). Any difference detected could represent actual difference between the DNA sources, an assembly artifact (computational or clone-induced) or alignment error.

In the first class of small sequence changes, we identified 1,613,458 nucleotides; 1,591,291 (98.6%) of these represented single-nucleotide differences, and the remainder (22,167 bp) represented other small changes ≤10 bp in size.

In the second class, we found 13,066 regions totaling 23,859,805 bp of unmatched sequence (average size was 1,826 bp; **Supplementary Table 4**). We used stringent filtering criteria to obtain this data set of putative insertion and deletion variants. In addition to removing regions lacking support from both alignment tools and all regions shorter than 50 bp, we also removed regions with a repeat content >95% and sequences that could be realigned to Build 35 using BLAT[14] (with >98% match, >50% coverage). Putative insertion points in Build 35 can be assigned for unmatched sequences when they are flanked by anchored neighboring alignments. In total, we are able to assign putative insertion coordinates for 4,536 unmatched fragments into Build 35 (corresponding to 10,469,693 bp; **Fig. 2** and **Supplementary Table 4**).

We used BLAT to compare R27c unmatched sequences to the chimpanzee assembly (NCBI Build 1) and identified 888 fragments

Table 2 Putative genetic variation detected by GCA

Variation type	GCA
Mismatch (P) (1–10 nt)	
Total events	1,602,411 (1,591,291 SNPs)
Total length (nt)	1,613,458
Unmatched	
Total events	13,066[a]
Total length (nt)	23,859,805
Copy-unmatched	
Total events	419
Total length (nt)	3,599,058
Inversions	
Total events	49
Total length (nt)	995,798

The table shows only regions identified from both MegaBLAST and A2Amapper alignments, with the exception of the copy-unmatched category, which cannot be extracted from A2Amapper data.

[a]The unmatched category has undergone further filtering of repetitive sequences as described in the main text and **Supplementary Methods**. The vast reduction in the number of unmatched events is due to exclusion of all events <50 bp in length.

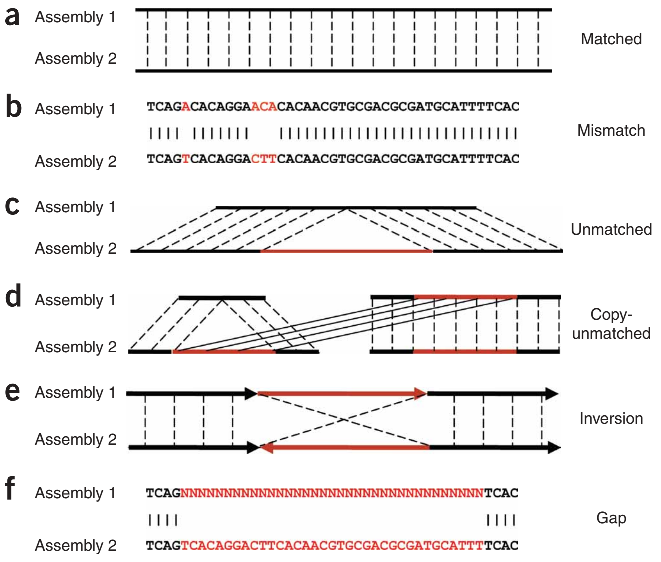

Figure 1 Overview of the different types of alignments and assembly differences extracted from the R27c and Build 35 genome assemblies. (**a**) Matched alignments account for the majority of the sequence. (**b**) Mismatches are small intra-alignment differences ≤10 bp in length. (**c**) Unmatched sequences are sequences that are present in one assembly but absent in the other. These sequences are candidates for insertion/deletion polymorphism. (**d**) Copy-unmatched sequences. This category contains sequences that are present in both assemblies but that have additional copies in one of the assemblies. Here we focus on regions >1 kb in size for which the additional copy has at least 98% identity. These sequences are candidates for copy number variation. (**e**) Inversions are sequences that appear in different orientation in the two assemblies. (**f**) Gaps are sequences represented by Ns. These can be aligned either to sequence or to gaps in the other assembly.

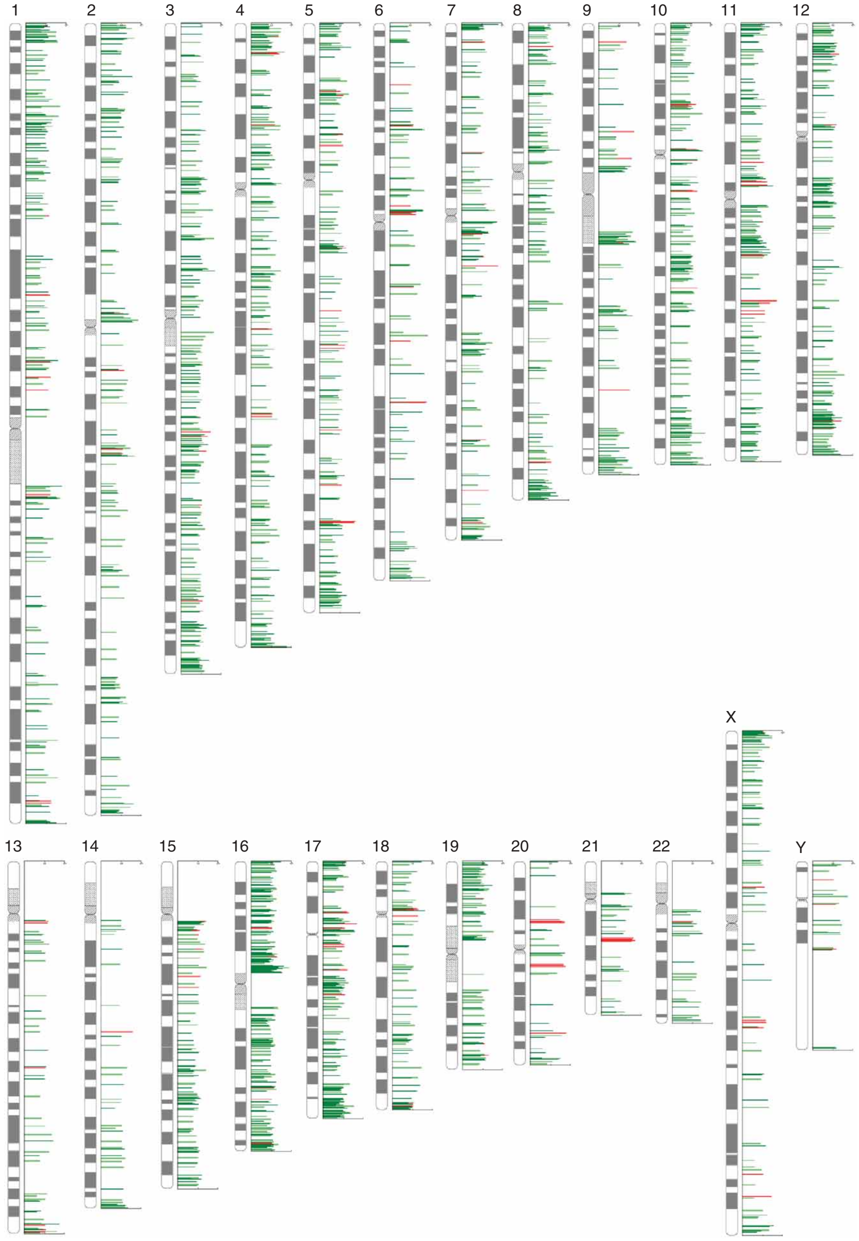

Figure 2 Genome-wide overview of insertion points of unmatched and copy-unmatched sequences present in R27c with no corresponding match to Build 35. Each bar represents an insertion point, and the length of each bar indicates the size of the unmatched fragment (log scale). Green and red bars represent unmatched and copy-unmatched sequences, respectively. The data shown in this figure are based on anchored unmatched data and copy-unmatched data encompassing 13,837,593 bp (**Supplementary Tables 2** and **4**).

covering 1,713,610 bp with a high identity match (>96% identity over 50% of the query). As these sequences have been identified both in humans and chimpanzees, they should represent either insertion or deletion polymorphisms or sequences missing in the reference genome. Next, we analyzed unmatched sequences in comparison with known genes. We found 903 RefSeq[15] genes that contained insertion points for the R27c unmatched sequence. In a separate analysis, we aligned all RefSeq mRNAs to both assemblies and identified 26 human mRNAs with >50 bp of coding sequence present in R27c but missing in Build 35 (some of these mRNAs spanned or extended into gaps, whereas other sequences were simply not present) (**Supplementary Table 5**). For example, *DOCK3* has an exon mapping within a sequence inverted in Build 35. To verify that the coding sequences missing in Build 35 are indeed represented as mRNA, we amplified and sequenced the cDNA from 14 different genes in five tissues (**Supplementary Table 6**) and obtained the expected results.

Copy-unmatched sequences are defined as fragments >1 kb that have two or more copies in R27c with >98% identity but have fewer copies present in Build 35. Thus, these regions represent putative CNVs but could also be explained by ubiquitous segmental duplications for which only one copy is annotated in Build 35. Celera shotgun reads have been used previously to identify regions of segmental duplications[16], but this approach does not assign an insertion point in the assembly for the additional copies. We identified 419 copy-unmatched fragments, which had an average size of 8.6 kb. Of these, we were able to assign an insertion point for 287 fragments. We also compared the copy-unmatched fragments with the regions previously detected by shotgun read depth analysis, and 63% overlapped.

The last two classes included inversions and gap sequences. We detected 47 intrascaffold inversions, and two entire scaffolds were in inverse orientation in R27c. Gaps are regions that contain Ns in the query sequence. Defining this class was necessary for sequence accounting but was not relevant for variation studies.

To validate computational predictions and test for polymorphism, we performed PCR analysis, quantitative real-time PCR or FISH. Initially, we selected 49 regions (38 unmatched regions, six inversions and five copy-unmatched regions; see Methods and **Supplementary Table 6**). We performed PCR for unmatched and inversion regions on a panel of 12 controls from CEPH pedigrees. We tested copy-unmatched regions by quantitative PCR on a panel of 48 controls. We found that 17 of 38 (45%) unmatched regions, one of six (17%) inversions and two of five (40%) copy-unmatched regions were polymorphic (a total of 20 of 49, or 41%), with one allele supporting each assembly. For 19 of 38 (50%) unmatched regions, the unmatched sequence was found in each sample tested. Of these, three extend into gaps in the reference assembly. For the two remaining unmatched fragments, we detected only the Build 35 sequence, indicating that these represent rare variants, R27c assembly errors or alignment artifacts. For six regions where the unmatched sequence was present in all individuals tested, we examined the genomic clone used by the IHGSC to generate the reference sequence. In three of six cases, we detected the unmatched sequence, indicating that absence in Build 35 was likely to be due to cloning or assembly problems.

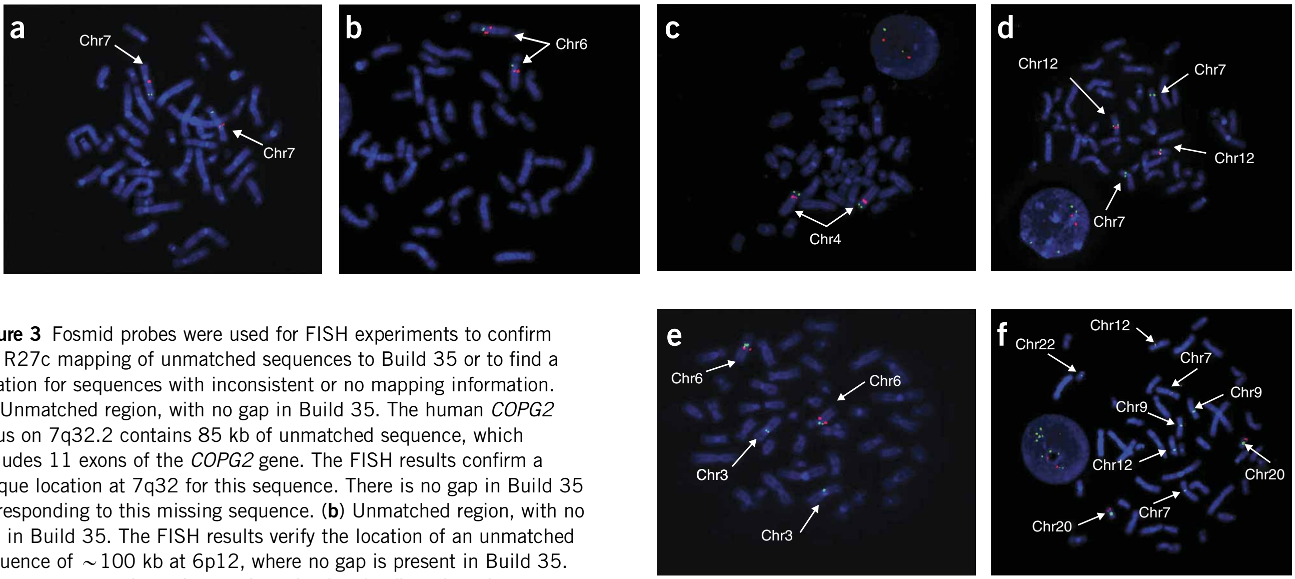

Figure 3 Fosmid probes were used for FISH experiments to confirm the R27c mapping of unmatched sequences to Build 35 or to find a location for sequences with inconsistent or no mapping information. (**a**) Unmatched region, with no gap in Build 35. The human *COPG2* locus on 7q32.2 contains 85 kb of unmatched sequence, which includes 11 exons of the *COPG2* gene. The FISH results confirm a unique location at 7q32 for this sequence. There is no gap in Build 35 corresponding to this missing sequence. (**b**) Unmatched region, with no gap in Build 35. The FISH results verify the location of an unmatched sequence of ~100 kb at 6p12, where no gap is present in Build 35. (**c**) Unmatched region, with gap in Build 35. Confirmation of an unmatched sequence mapping to a Build 35 assembly gap at 4p16. (**d**) FISH results for a sequence mapped to chromosome 7 in R27c and chromosome 12 in Build 35. The sequence does not correspond to an annotated segmental duplication. The results indicate that the sequence is present on both chromosomes in each tested individual. (**e**) An unanchored scaffold assigned to chromosome 6 in R27c, with no match in Build 35. The results show localization to centromeric regions on chromosomes 3 and 6. (**f**) An unanchored scaffold mapping to chromosome 12 in R27c and chromosome 20 in Build 35, with segmental duplications mapping to chromosomes 7, 12, 15 and 20. The result confirms multiple mapping locations. Only one homolog of chromosome 22 consistently showed a signal, indicating that this sequence may be polymorphic.

We performed FISH on three individuals using fosmid clones whose ends mapped within unmatched regions. We tested four types of regions: (i) 11 unmatched fragments with a location assigned in R27c, (ii) 21 fragments mapping to different chromosome locations in R27c and Build 35, (iii) six unanchored scaffolds with no coordinates assigned in either assembly and (iv) three scaffolds of uncertain orientation in R27c. Representative results for the first three categories are shown in **Figure 3**, and detailed results are summarized in **Supplementary Table 7**. The FISH analysis confirmed the expected mapping for seven fragments corresponding to Build 35 assembly gaps and two fragments corresponding to regions in which no gap is currently present in Build 35 (**Fig. 3a–c**). All FISH results for sequences assigned to different chromosomes in the two assemblies and for those with no coordinates assigned showed hybridization to multiple locations (**Fig. 3d–f**), often including centromere regions. The majority of these also demonstrated differences either in intensity or localization of hybridization signals between individuals. We experimentally verified that three scaffolds of uncertain orientation in R27c supported the orientation in Build 35 (**Supplementary Table 7**).

We further assessed putative variants between assemblies by comparison with other data sources (**Supplementary Tables 8** and **9**). First, we found 1,521,291 of 1,591,291 (95.6%) single-nucleotide mismatches to be present in dbSNP; 840,802 of these were Celera-based SNPs, whereas the others were from different projects. We compared the unmatched and copy-unmatched categories with entries in the Database of Genomic Variants[17]. We found 331 CNVs to contain insertion points for unmatched regions and 55 CNVs with insertion points for copy-unmatched regions. Limiting the analysis to unmatched and copy-unmatched fragments >10 kb yielded support for 53 CNVs. Using data from ref. 18, we correlated the 913 CNVs detected by the whole-genome tile path clone array with unmatched and copy-unmatched sequences. We found a significant correlation, with 254 CNVs overlapping unmatched insertion points and 74 CNVs overlapping copy-unmatched sequences ($P < 0.0001$ for both). Of these, 88 unmatched and 13 copy-unmatched regions were >10 kb, indicating that they may explain the CNV detected. We also assessed the overlap with variants identified by the fosmid end-pair mapping approach and found support for 23 insertions identified in ref. 3. Comparison of the entire unmatched data set (including those with repeat content >95%) with the dbRIP retrotransposon polymorphism database[19] yielded support for another 54 polymorphic regions, all of which corresponded to single short interspersed nuclear elements (SINE) or long interspersed nuclear elements (LINE). We compared the 49 inversions with entries in the Database of Genomic Variants[17]; 12 corresponded to previously identified inversion polymorphisms.

A total of 3,246,015 bp of R27c sequence extended into Build 35 gaps and 1,110/4,536 (24.5%) unmatched fragments and 174/287 (60.6%) copy-unmatched sequences had insertion points in annotated segmental duplications. We also noted a strong association of insertions with segmental duplications for regions detected by the fosmid-end mapping approach[3].

Alternate sequence assemblies have been created previously for specific subregions of the human genome[20–22], facilitating the understanding of chromosome architecture. The results presented here confirm that whole-genome assembly comparison is the most sensitive way of identifying all types of genetic variation and that there is no limit to the size of the variants found. We provide experimental evidence that >40% of predicted unmatched regions can be confirmed experimentally and that many others correspond to known variable regions. As the current study is limited to two genome assemblies, most genetic variants will be presented by the major allele. Even the most conservative extrapolations, therefore, suggest that significantly more variation exists between humans than was previously estimated[4,5]. Moreover, we show that alternate assemblies can be used to contribute to the generation of a more complete reference sequence.

As an era of personalized sequencing approaches[23–25], our results emphasize that developing effective strategies for extracting the most relevant data will rely on a comprehensive understanding of the content of both test and comparator sequences.

METHODS

Assemblies and alignment algorithm. Build 35 sequences were downloaded from NCBI. Sequences for the R27c assembly were obtained from Celera but are also publicly available from NCBI with accession number AADB02000000. For detailed information on the Genome Comparison Algorithm, see **Supplementary Methods.** Briefly, chromosome sequence assemblies from NCBI Build 35 were compared with all scaffold sequences from R27c using MegaBLAST[12]. The resulting alignments were converted to GFF3 format, recording detailed alignment information. GFF3 records were sorted by raw score. A greedy algorithm applied to the GFF3 records preferentially selected optimal alignments (that is, alignments with the highest raw score), eliminated suboptimal alignments and created a nonredundant set of nonoverlapping alignments by cutting GFF3 alignment records. Unmatched sequence was determined by identifying intervening sequence between alignment records. Copy-unmatched sequence was determined by searching among suboptimal alignments for sequence already matched in one assembly, but not the other, using a cutoff of 1 kb in length and >98% sequence identity. Inversions were determined by identifying alignments whose orientation was different than adjacent alignments. The complete data set is available upon request.

Correlation with genomic features. Analyses of correlations with genomic features were performed using standard data sets. The RefSeq gene set and RefSeq mRNAs[15] were downloaded from NCBI. Information about CNVs was retrieved from the Database of Genomic Variants (http://projects.tcag.ca/variation/). Coordinates for segmental duplications were extracted from Human Genome Segmental Duplication Database (http://projects.tcag.ca/humandup/)[26], and whole-genome shotgun sequence detection (WSSD) regions and gap coordinates were downloaded from the UCSC Human Genome Browser[27]. The repeat content of unmatched sequences was determined using RepeatMasker (A.F.A. Smit, R. Hubley & P. Green, Institute for Systems Biology, Seattle; see http://www.repeatmasker.org). Detailed information regarding the genomic feature overlap analyses is given in **Supplementary Methods**.

PCR reactions. PCR experiments were designed as previously described for unmatched regions[3] and inversions[28], with reagents and optimization criteria as in ref. 28. In order to enrich for potential polymorphisms and simplify experimental design, a number of selection criteria were applied for identification of candidate regions. Regions chosen for experimental validation were all intrascaffold sequences, with <50% repeat content and <50% repeat content in the 1 kb on each side flanking the insertion point. Regions where other assembly differences mapped immediately adjacent to the insertion point in Build 35 were also avoided. See **Supplementary Table 6** for primer sequences.

Quantitative PCR. Variability in copy number in copy-unmatched regions was tested by quantitative real-time PCR, using probes from the Human Universal Probe Library (Roche Diagnostics). The change in copy number was calculated as previously described[29]. DNA from a total of 48 unrelated HapMap individuals of European ancestry was used to assess the existence of variability in copy number. We used 20 ng of total DNA in each of two replicates. Replicates with a variation coefficient over 4% were discarded. The myoglobin B IX gene was used as the reference for the relative quantifications. See **Supplementary Table 6** for primer sequences.

FISH and probe selection. Fosmid clones were used as probes for FISH experiments. Fosmid end-sequences from NCBI (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/FOSMIDS/) were downloaded and aligned to the R27c and Build 35 genome assemblies using BLAT. Best unique matches were retrieved, and fosmids that mapped only to the Celera genome, fosmids with discrepancies in span size between assemblies and fosmids with best reciprocal matches found on different chromosomes were all recorded and compared with the GCA output. All FISH experiments were performed on three samples using standard protocols as described previously[17,28,30]. Five to ten selected metaphases were examined using fluorescence microscopy, analyzed and imaged. Three-color interphase FISH for inversion testing was performed as previously described[30].

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

ACKNOWLEDGMENTS

We thank T. Tang, L. Wong, J. Wittnam, C.-F. Chu and W. Hwang of The Centre for Applied Genomics for technical assistance. Computational analyses were supported by the Shared Hierarchical Academic Research Computing Network (SHARCNET) and the Centre for Computational Biology at the Hospital for Sick Children. The work was supported by Genome Canada/Ontario Genomics Institute, the Canadian Institutes of Health Research (CIHR), the Canada Foundation for Innovation and the McLaughlin Centre for Molecular Medicine (all to S.W.S). L.A. and X.E. are supported by Genoma España and Genome Canada joint R+D+I projects and by the Generalitat de Catalunya (Departament d'Universitats, 2005SGR00008, and Departament de Salut). L.F. is supported by CIHR. S.W.S. is an Investigator of CIHR and International Scholar of Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

The study was designed by R.K., S.W.S. and L.F. The GCA algorithm was created by R.K. Sequence alignment and computational analysis was performed by R.K., J.Z., J.R.M, J.W., C.Q., L.A. and R.J.M. FISH analysis was performed by Y.H., A.M.J.G., M.S. and C.L. PCR analysis was performed by M.A.R., L.P., L.A. and L.F. J.Z., J.R.M, J.W., C.Q., H.A., K.J., R.R., M.H., L.A., X.E., C.L., S.W.S. and L.F contributed to the analysis of overlap with genomic features, creation of data sets for such analysis and interpretation of the data. S.W.S. and L.F conceptualized, designed and coordinated the experiments. The paper was written by S.W.S and L.F.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

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

1. Marth, G.T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**, 452–456 (1999).
2. Tsui, C. *et al.* Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res.* **31**, 4910–4916 (2003).
3. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
4. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
5. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
6. Myers, E.W., Sutton, G.G., Smith, H.O., Adams, M.D. & Venter, J.C. On the sequencing and assembly of the human genome. *Proc. Natl. Acad. Sci. USA* **99**, 4145–4146 (2002).
7. Adams, M.D., Sutton, G.G., Smith, H.O., Myers, E.W. & Venter, J.C. The independence of our genome assemblies. *Proc. Natl. Acad. Sci. USA* **100**, 3025–3026 (2003).
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. Waterston, R.H., Lander, E.S. & Sulston, J.E. On the sequencing of the human genome. *Proc. Natl. Acad. Sci. USA* **99**, 3712–3716 (2002).
10. 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).
11. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
12. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7**, 203–214 (2000).
13. Mobarry, C. & Sutton, G. An assembly-to-assembly comparison tool. in *Proceedings of the Third Annual RECOMB Satellite Meeting on DNA Sequencing Technologies and Computation* (2003).
14. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
15. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
16. Bailey, J.A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).

17. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
18. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* (in the press).
19. Wang, J. *et al.* dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.* **27**, 323–329 (2006).
20. Hillier, L.W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424**, 157–164 (2003).
21. Scherer, S.W. *et al.* Human chromosome 7: DNA sequence and biology. *Science* **300**, 767–772 (2003).
22. Schmutz, J. *et al.* The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, 268–274 (2004).
23. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* **5**, 335–344 (2004).
24. Bennett, S.T., Barnes, C., Cox, A., Davies, L. & Brown, C. Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**, 373–382 (2005).
25. Service, R.F. Gene sequencing. The race for the $1000 genome. *Science* **311**, 1544–1546 (2006).
26. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**, R25 (2003).
27. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
28. Feuk, L. *et al.* Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.* **1**, e56 (2005).
29. Pfaffl, M.W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, e45 (2001).
30. Osborne, L.R. *et al.* A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat. Genet.* **29**, 321–325 (2001).