

**The following resources related to this article are available online at [www.sciencemag.org](http://www.sciencemag.org) (this information is current as of December 2, 2009):**

**Updated information and services**, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/cgi/content/full/316/5822/235>

**Supporting Online Material** can be found at:

<http://www.sciencemag.org/cgi/content/full/316/5822/235/DC1>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/cgi/content/full/316/5822/235#related-content>

This article **cites 25 articles**, 14 of which can be accessed for free:

<http://www.sciencemag.org/cgi/content/full/316/5822/235#otherarticles>

This article has been **cited by** 7 article(s) on the ISI Web of Science.

This article has been **cited by** 2 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/cgi/content/full/316/5822/235#otherarticles>

This article appears in the following **subject collections**:

Evolution

<http://www.sciencemag.org/cgi/collection/evolution>

Information about obtaining **reprints** of this article or about obtaining **permission to reproduce this article** in whole or in part can be found at:

<http://www.sciencemag.org/about/permissions.dtl>

## REPORT

# Human-Specific Changes of Genome Structure Detected by Genomic Triangulation

R. A. Harris,<sup>1,2</sup> J. Rogers,<sup>3</sup> A. Milosavljevic<sup>1,2\*</sup>

Knowledge of the rhesus macaque genome sequence enables reconstruction of the ancestral state of the human genome before the divergence of chimpanzees. However, the draft quality of nonhuman primate genome assemblies challenges the ability of current methods to detect insertions, deletions, and copy-number variations between humans, chimpanzees, and rhesus macaques and hinders the identification of evolutionary changes between these species. Because of the abundance of segmental duplications, genome comparisons require the integration of genomic assemblies and data from large-insert clones, linkage maps, and radiation hybrid maps. With genomic triangulation, an integrative method that reconstructs ancestral states and the structural evolution of genomes, we identified 130 human-specific breakpoints in genome structure due to rearrangements at an intermediate scale (10 kilobases to 4 megabases), including 64 insertions affecting 58 genes. Comparison with a human structural polymorphism database indicates that many of the rearrangements are polymorphic.

The human, chimpanzee, and rhesus macaque genomes have now all been sequenced, but the draft quality of the nonhuman primate genomes hampers accurate comparisons between these species. A challenge in using draft-quality genome assemblies to analyze evolutionary changes is that relatively large numbers of differences less than 4 Mb in size among closely related organisms, due to microinsertions, deletions, and copy-number variations, occur in highly duplicated regions that are hard to assemble and compare. Lineage-specific rearrangements in primates have been detected by chromosome banding (1) and fluorescence in situ hybridization (2), but these methods are not sensitive enough to identify microinsertions and deletions. Array comparative genomic hybridization can detect microinsertions and deletions (3–6) but cannot position them on the genome. Chimpanzee fosmid end sequences (FESs) that map inconsistently, based on insert size and orientation, onto the human genome have been used to detect structural differences (7), but such studies are prone to false positives and negatives in regions where the genome assembly is incomplete or erroneous and are limited to comparisons of two species. FES mapping, along with pairwise genome assembly comparisons between human and chimpanzee

(8), could not establish the ancestral genome state nor assign structural rearrangements to specific lineages. It is increasingly evident that the reconstruction of primate genome evolution requires the integration of genomic data obtained by different methodologies (9–11), due, in part, to segmental duplications that confound both genome assembly and determination of orthology. We performed a three-way comparison of the rhesus, human, and chimpanzee genomes to reconstruct the ancestral genome at the branching point of chimpanzee and human. We were able to infer human-specific rearrangements, including insertions, deletions, and inversions, at an intermediate scale of resolution above the 10-kb retroposon insertion size and below the 4-Mb size detectable by cytogenetic methods.

Genomic triangulation reconstructs genome evolution by inputting genomes and other available genomic data from at least three species, two of which form a monophyletic group and one of which is basal to those two and serves as an outgroup. The output is a reconstruction of the ancestral genome of the monophyletic group and a map of breakpoints in specific branches of an unrooted phylogenetic tree. Genomic triangulation consists of three main steps (Fig. 1 and SOM methods): (i) blockset construction, (ii) ancestral threading, and (iii) ancestral gapset construction.

The blockset construction step (Fig. 1) takes two genomes and produces blocks consisting of pairs of orthologous chromosomal segments unbroken by large-scale rearrangements (12), each block corresponding to a segment of the ancestral genome of the monophyletic group. Blocks are inferred from collinear orthologous anchors

across two extant genomes. The orthologous anchors are markers in either comparative linkage maps or radiation hybrid maps or are inferred from genome assemblies and large-insert clones. When clone ends of one species anchor onto loci in another species at a distance consistent with insert size and in correct orientation, a clone-sized block is inferred. Overlapping consistent clone-sized blocks are merged into larger blocks. Blocks produced by different anchoring methods are similarly merged on the basis of regions where their genomic coordinates overlap. If different methods produce inconsistent blocks, the longest block is chosen as the best orthology assignment.

The ancestral threading step (Fig. 1) deduces the ancestral genome structure from blocksets in a process similar to standard genome assembly. The overlap of blocks from different blocksets is deduced from overlaps of positional coordinates of the blocks within a single species. All overlapping blocks from the three pairwise blocksets are “threaded” by this method into scaffolds of an ancestral genome. The inferred ancestral genome and the pairwise blocksets between extant species are then used to deduce ancestral blocksets connecting the extant genomes with the ancestral genome.

Because every breakpoint is flanked by two blocks, for each pair of adjacent blocks we define an entity we call a gap. The distances between the paired blocks in the two genomes are associated with each gap. The ancestral gapset construction step (Fig. 1) infers ancestral gapsets from ancestral blocksets. The ancestral gaps localize breakpoints positionally within a genome and evolutionarily within a branch of the phylogenetic tree.

Genomic triangulation was applied to map the evolution of the genomes of the human, chimpanzee, and rhesus macaque. The results, integrated with those obtained by other methods, are summarized in (13). The input data for genomic triangulation were human, chimpanzee, and rhesus genome assemblies anchored by Pash (14) and University of California Santa Cruz (UCSC) Alignment Nets (15), mapped fosmid and bacterial artificial chromosome (BAC) end sequences, rhesus BACs mapped by pooled genomic indexing (16), and human-rhesus comparative linkage (17) and radiation hybrid (18) maps. Varying combinations of merged data sets were input to detect breakpoints at six different levels of resolution (table S4). Detected breakpoints were merged into a single set for further analysis. A total of 288 putative human-specific breakpoints were detected by genomic triangulation. Figure 1 shows the general method that allowed us to infer the existence of a human-specific breakpoint.

Visual inspection of the breakpoint loci by means of the UCSC Genome Browser revealed 158 breakpoints that could not be corroborated

<sup>1</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. <sup>2</sup>Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. <sup>3</sup>Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX 78245, USA.

\*To whom correspondence should be addressed.

# The Rhesus Macaque Genome

and were classified as possible false positives (SOM methods and table S5). Of these, 31 were due to gaps in the human assembly. The remaining 127 are disagreements with alignment nets in the UCSC Genome Browser and may represent artifacts associated with the current implementation of the genomic triangulation method. Among the remaining 130 human-specific breakpoints, there are 64 insertions, 7 deletions, 16 inversions, and 43 breakpoints that could not be unambiguously assigned to a specific type of rearrangement (Fig. 2). The average size of detected rearrangements on the human genome was 110,063 base pairs (bp), ranging from 20 to 1,365,171 bp.

The human-specific rearrangements were compared with rearrangements detected by the mapping of chimpanzee FESs onto the human genome (7). A total of 52 breakpoints detected by genomic triangulation overlapped with 68 rearrangements, out of a total of 592, detected by FES mapping. Because FES mapping does not assign rearrangements to a specific lineage, by virtue of overlap the 68 rearrangements detected by FES mapping could now be placed in the human lineage. In order to estimate the amount of overlap expected by chance, 100 sets of random genomic locations, with sizes matched to the rearrangements detected by genomic triangulation, were created and their overlap with the FES rearrangements was determined. On average, 25 random locations overlapped with FES rearrangements, indicating a 2.1-fold (52/25) enrichment for overlap in the genomic triangulation set as compared to the random set.

Comparison of detected rearrangements to the Segmental Dups track from the UCSC Genome Browser (19) revealed that 1 deletion, 7 inversions, and 32 unclassified rearrangements were within 10 kb of a segmental duplication. Of the 64 insertions, 42 (66%) overlapped segmental duplications, with 25 insertions being at least 98% identical to their paralogs, as would be expected for recent duplication events (Fig. 2). The remaining 20 insertions were mostly under 20 kb, with 17 consisting of retrotransposons flanking a small segment of putatively transposed sequence, suggesting transduction. The insertions account for roughly one-half of the 8 Mb (8) of human-specific sequence estimated from comparisons of human and chimpanzee genome assemblies, but only a small fraction of the 37 Mb of human-specific sequence estimated on the basis of whole-genome shotgun-read analysis (20).

To further substantiate the insertions, chimpanzee and rhesus BAC and fosmid clone end sequences with mappings that span the insertions were identified. Chimpanzee or rhesus clone ends that map onto the human genome at a distance significantly greater than expected based on the distribution of the clone insert sizes suggest that an insert has occurred in the human

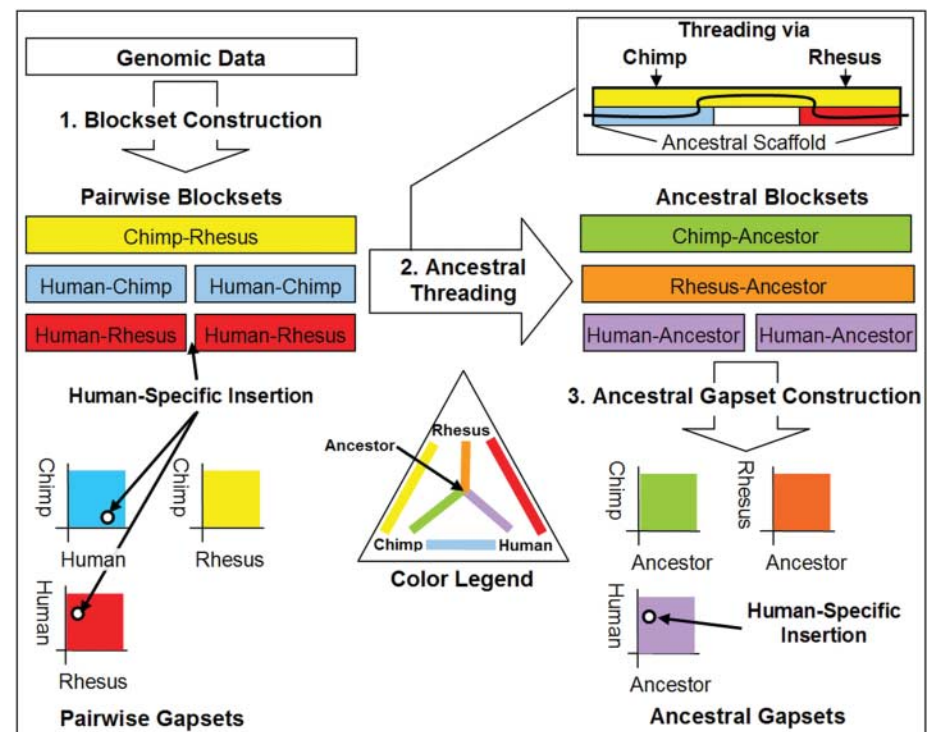
relative to the other species. Of 64 insertions, 22 were spanned by both chimpanzee and rhesus clones and 24 by clones from only one of the species, with a mapping distance significantly longer than expected.

A total of 22 of 130 (17%) breakpoints occur on the X chromosome, which comprises only 5% of the genome. This highly significant enrichment ( $z$  score  $> 6$ ) is consistent with a threefold enrichment in breakpoints in X detected in rhesus-human comparisons (13) and contrasts both with the significantly lower base pair-level divergence of the primate X chromosome as compared to the autosomes (21–23) and with the conservation of the cytogenetically detectable structure of the X chromosome in primates.

We identified 58 genes affected by insertions (table S8), including 36 gene copies fully contained within insertions. An additional 22 genes were either partially duplicated or contained an insertion. Of the 36 fully duplicated genes, 7 were previously identified as human-specific (20).

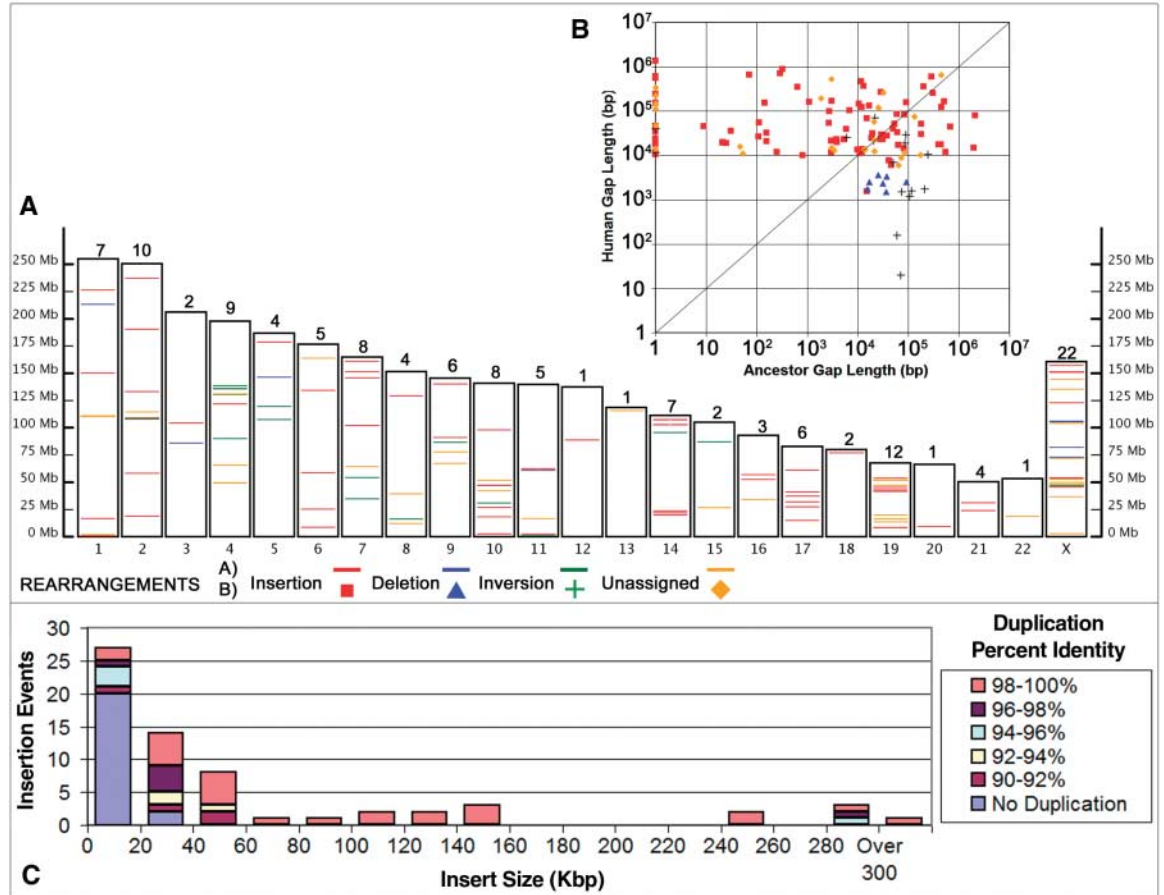
To determine whether the detected human-specific rearrangements are fixed in all humans, we identified overlaps between the rearrangements and structural variants from The Centre for Applied Genomics (TCAG) Database of Genomic Variants (24). Of the 130 rearrangements, 53 overlapped with at least one polymorphism on the basis of coordinates. We also randomly selected 100 sets of 130 genomic locations, matching the genomic triangulation–detected rearrangement sizes, and identified overlap with the TCAG database. On average, only 28 random locations overlapped with polymorphisms. This 1.9-fold (53/28) enrichment in polymorphisms in the detected set as compared to the random set suggests that a significant fraction of rearrangements are structural alleles that are either polymorphic, and hence not fixed, in all humans or else are sites of recurrent rearrangements.

Genomic triangulation clearly mimics the overlap and layout stages of the overlap–layout–consensus genome assembly method (25), but the two methods differently employ abductive (hypothesis-generating) and deductive inference.



**Fig. 1.** Detection by genomic triangulation of a human-specific breakpoint induced by an insertion. In the example shown (left), blockset construction revealed one long conserved block for a chimp-rhesus comparison. However, blocks from human-chimp and human-rhesus blocksets align except for a gap, which suggests that there has been a human-specific insertion. This is further illustrated by the pairwise gapsets. Pairwise blocksets and gapsets between extant species are represented by primary colors (red, blue, and yellow) and those between an extant species and the ancestor are represented by complementary colors (purple, green, and orange). Gapsets are represented as gap maps, with white circles representing gaps and circle coordinates indicating sizes of the gaps in specific genomes. Gapsets on the left indicate breakpoints in human-chimp and human-rhesus pairwise comparisons due to the human breakpoint. Ancestral gapsets on the right indicate the breakpoint between human and the ancestor.

**Fig. 2.** Human-specific breakpoints. **(A)** Chromosome ideograms of 130 human-specific breakpoints detected by genomic triangulation, with breakpoint counts by chromosome. **(B)** Gap map with ancestor gap lengths on the x axis and human gap lengths on the y axis. Only gaps greater than 10 kb in at least one of the genomes and smaller than 4 Mb were considered. Some insertions occur to the right of the  $y = x$  axis because visual inspection revealed inaccuracies in gap sizing, providing evidence for human insertion. **(C)** Size distribution of detected human-specific insertions. The percent identity of segmental duplications can indicate the approximate age of duplication events.



First, because the fragments of an ancestral genome cannot be sequenced directly, they are inferred abductively in the blockset construction step. In other words, each block of similarity is explained by hypothesizing that a corresponding fragment was present in the ancestral genome. Second, the fragments are not assembled into ancestral scaffolds abductively as in genome assembly, where read fragments with similar sequences at their ends are assembled into a contig under the hypothetical assumption that their similarity is due to their origin from overlapping genomic loci. Instead, the ancestral threading step infers the ancestral genome structure deductively from block overlaps in assembled genomes of extant species. In contrast to the contiguous ancestral region (CAR)-building method (26), which relies solely on alignment nets between assembled genomes, genomic triangulation integrates assembly anchoring data with clone end sequencing and physical map information, thereby reducing problems associated with draft-quality assemblies. CAR can detect inversions, translocations, fissions, and fusions larger than 50 kb in length but is insensitive to insertions and deletions detectable by genomic triangulation.

The genomic triangulation method introduces the gapset, a key concept complementary to that of a blockset (12). Gapsets provide a means of

tracking breakpoints, which cannot be adequately expressed using blocksets. A gapset operationally defines breakpoints under minimal assumptions, thus helping to integrate genome information from multiple sources. Ancestral gapsets localize breakpoints both by position, by flanking blocks within a genome, and phylogenetically, by assigning them to a specific branch of a phylogenetic tree, thus providing maximal information for the reconstruction of evolutionary changes in genome structure.

#### References and Notes

- J. J. Yunis, O. Prakash, *Science* **215**, 1525 (1982).
- J. Wienberg, *Cytogenet. Genome Res.* **108**, 139 (2005).
- M. C. Popesco *et al.*, *Science* **313**, 1304 (2006).
- A. Fortna *et al.*, *PLoS Biol.* **2**, E207 (2004).
- G. M. Wilson *et al.*, *Genome Res.* **16**, 173 (2006).
- V. Goidts *et al.*, *Hum. Genet.* **120**, 270 (2006).
- T. L. Newman *et al.*, *Genome Res.* **15**, 1344 (2005).
- The Chimpanzee Sequencing and Analysis Consortium, *Nature* **437**, 69 (2005).
- M. Rocchi, N. Archidiacono, R. Stanyon, *Genome Res.* **16**, 1441 (2006).
- L. Froenicke *et al.*, *Genome Res.* **16**, 306 (2006).
- G. Bourque, G. Tesler, P. A. Pevzner, *Genome Res.* **16**, 311 (2006).
- M. Blanchette *et al.*, *Genome Res.* **14**, 708 (2004).
- The Rhesus Macaque Genome Sequencing and Analysis Consortium, *Science* **316**, 222 (2007).
- K. J. Kalafus, A. R. Jackson, A. Milosavljevic, *Genome Res.* **14**, 672 (2004).

- W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, D. Haussler, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 11484 (2003).
- A. Milosavljevic *et al.*, *Genome Res.* **15**, 292 (2005).
- J. Rogers *et al.*, *Genomics* **87**, 30 (2006).
- W. J. Murphy *et al.*, *Genomics* **86**, 383 (2005).
- J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res.* **11**, 1005 (2001).
- Z. Cheng *et al.*, *Nature* **437**, 88 (2005).
- M. T. Ross *et al.*, *Nature* **434**, 325 (2005).
- N. Patterson, D. J. Richter, S. Gnerre, E. S. Lander, D. Reich, *Nature* **441**, 1103 (2006).
- N. H. Barton, *Curr. Biol.* **16**, R647 (2006).
- The Centre for Applied Genomics, <http://projects.tcag.ca/variation/>.
- G. Myers, *Comput. Sci. Eng.* **1**, 33 (1999).
- J. Ma *et al.*, *Genome Res.* **16**, 1557 (2006).
- We thank A. R. Jackson for discussion and feedback and R. A. Gibbs for including us in the Rhesus Genome Sequencing Consortium. A.M. acknowledges support from NIH-National Human Genome Research Institute grants R01 02583-01 and R01 004009-1, and the NIH-National Center for Research Resources (NCRR) grant U01 RR 18464. J.R. acknowledges support from NIH-NCRR grant R24-RR015383. Additional supporting information is available at [www.genboree.org](http://www.genboree.org).

#### Supporting Online Material

[www.sciencemag.org/cgi/content/full/316/5822/235/DC1](http://www.sciencemag.org/cgi/content/full/316/5822/235/DC1)  
 Materials and Methods  
 Figs. S1 to S3  
 Tables S1 to S8  
 References and Notes

3 January 2007; accepted 16 March 2007  
 10.1126/science.1139477