The Sequence and Analysis of Duplication Rich Human Chromosome 16

Joel Martin⇓, Cliff Hanδ, Laurie A. Gordoṇ⇓, Astrid Terry⇓, Shyam Prabhakar?, Xinwei She@, Gary Xieδ⇓, Uffe Hellsten⇓, Yee Man Chan*, Michael Altherrδ⇓, Olivier Couronne?, Andrea Aerts⇓, Eva Bajorek*, Stacey Black*, Heather Blumerδ, Elbert Branscomb#⇓, Nancy C. Brownδ, William J. Brunoδ, Judith M. Buckinghamδ, David F. Callenδ, Connie S. Campbellδ, Mary L. Campbellδ, Evelyn W. Campbellδ, Chenier Caoile*, Jean F. Challacombeδ, Leslie A. Chasteenδ, Olga Chertkovδ, Han C. Chiδ, Mari Christensen#, Lynn M. Clarkδ, Judith D. Cohnδ, Mirian Denys*, John C. Detter⇓, Mark Dickson*, Mira Dimitrijevic-Bussodδ, Julio Escobar*, Joseph J. Fawcettδ, Dave Flowers*, Dea Fotopulos*, Tijana Glavina⇓, Maria Gomez*, Eidelyn Gonzales*, David Goodstein⇓, Lynne A. Goodwinδ, Deborah L. Gradyδ, Igor Grigoriev⇓, Matthew Groza#, Nancy Hammon⇓, Trevor Hawkins⇓, Lauren Haydu*, Carl E. Hildebrandδ, Wayne Huang⇓, Sanjay Israni⇓, Jamie Jett⇓, Phillip E. Jewettδ, Kristen Kadner⇓, Heather Kimball⇓, Arthur Kobayashi⇓#, Marie-Claude Krawczykδ, Tina Leybaδ, Jonathan L. Longmireδ, Frederick Lopez*, Yunian Lou⇓, Steve Lowry⇓, Thom Ludemanδ, Graham A. Markδ, Kimberly L Mcmurrayδ, Linda J. Meinckeδ, Jenna Morgan⇓, Robert K. Moyzisδ, Mark O. Mundtδ, A. Christine Munkδ, Richard D. Nandkeshwar#, Sam Pitluck⇓, Martin Pollard⇓, Paul Predki⇓, Beverly Parson-Quintanaδ, Lucia Ramirez*, Sam Rash⇓, James Retterer*, Darryl O. Rickeδ, Donna L. Robinsonδ, Alex Rodriguez*, Asaf Salamov⇓, Elizabeth H. Saundersδ, Duncan Scott⇓, Timothy Shoughδ, Raymond L. Stallingsδ, Malinda Stalveyδ, Robert D. Sutherlandδ, Roxanne Tapiaδ, Judith G. Tesmerδ, Nina Thayerδ⇓, Linda S. Thompsonδ, Hope Tice⇓, David C. Torneyδ, Mary Tran-Gyamfi⇓, Ming Tsai*, Levy E. Ulanovskyδ, Anna

Ustaszewska⇓, Nu Vo*, P. Scott Whiteδ, Albert L. Williamsδ, Patricia L. Willsδ, Jung-Rung Wuδ, Kevin Wu*, Joan Yang*, Pieter DeJong%, David Bruceδ, Norman Doggettδ, Larry Deavenδ, Jeremy Schmutz*, Jane Grimwood*, Paul Richardson⇓, Daniel S. Rokhsar⇓, Evan E. Eichler@, Paul Gilnaδ, Susan M. Lucas⇓ Richard M. Myers*, Edward M. Rubin?⇓, and Len A. Pennacchio?⇓

⇓ DOE's Joint Genome Institute, 2800 Mitchell Avenue, Walnut Creek, California 94598, USA

δ Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, California 94550, USA

? Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, California 94720, USA

@ Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

* Stanford Human Genome Center, Department of Genetics, Stanford University School of Medicine, 975 California Ave, Palo Alto, California 94304, USA

% Children's Hospital Oakland, Oakland, California 94609, USA

To whom correspondence should be addressed: EMRubin@lbl.gov

ABSTRACT

We report here the 78,884,754 base pairs of finished human chromosome 16 sequence, representing over 99.9% of its euchromatin. Manual annotation revealed 880 protein-coding genes confirmed by 1,637 aligned transcripts, 19 tRNA genes, 341 pseudogenes and 3 RNA pseudogenes. These genes include metallothionein, cadherin and iroquois gene families, as well as the disease genes for polycystic kidney disease and acute myelomonocytic leukemia. Several large-scale structural polymorphisms spanning hundreds of kilobasepairs were identified and result in gene content differences across humans. One of the unique features of chromosome 16 is its high level of segmental duplication, ranked among the highest of the human autosomes. While the segmental duplications are enriched in the relatively gene poor pericentromere of the p-arm, some are involved in recent gene duplication and conversion events which are likely to have had an impact on the evolution of primates and human disease susceptibility.

INTRODUCTION

The mapping and sequencing of human chromosome 16 was initiated by the U.S. Department of Energy (DOE) in 1988 based on long-term interests in providing a fundamental understanding of radiation and its relationship to human biology. The localization of the DNA repair gene *ERCC4* to chromosome 16[1] coupled with the availability of a unique flow-sorted chromosome 16-specific cosmid library solidified the choice of this chromosome as a DOE sequencing target. Further interest in the role metallothioneins play in heavy metal transport, detoxification and their clustering on human chromosome 16 also coincided well with DOEs biological mission[2-4]. Here we describe the finished human chromosome 16 sequence which provides a reference for the further exploration of genomic sequence alterations and their relationship to human biology.

Mapping and Sequencing

To provide the foundation for sequencing human chromosome 16, we constructed a physical map based on previous STS content maps[5-7] consisting of 716 clones; which include 618 BACs, 79 cosmids, 7 fosmids, 5 PACs, 3 YAC subclones, 2 P1s, 2 phage vectors and 5 genomic PCR fragments. The final sequence contains four gaps, with two in each of the chromosome arms. One of the gaps is found in the highly duplicated pericentromeric region in the p arm, while two of the remaining non-pericentromeric gaps are resistant to stable cloning with conventional vectors and efforts are ongoing to close the estimated ~25kb of missing sequence using alternative vectors[8]. The final gap is found near the telomere of the q arm in a region distal to the last identifiable half YAC[9].

The high degree of segmental duplication of chromosome 16, coupled with the multiple haplotypes represented in the numerous clone libraries comprising the tiling path, hindered efforts to construct a valid clone based representation of this chromosome. To resolve this issue, we adopted a strategy of high depth clone coverage from a library constructed from a single individual[10]. This enabled the determination of both of the diploid haplotypes across the segmentally duplicated intervals. Overall, these efforts resulted in the generation of 78,884,754 base pairs of finished euchromatic sequence with an estimated accuracy[11] exceeding 99.9% and covering in excess of 99.9% of its euchromatin. Including the centromere and its adjacent heterochromatic portion of the q arm, the total size of the chromosome is estimated at 88.7 Mb.

As a further assessment of the physical sequence we compared it to the existing physical and genetic maps. We were able to account for all sequence-tagged sites from the Genethon[12] micro-satellite, the DeCODE[13] and the Marshfield[14] genetic maps. We also compared the final DNA sequence with recombination distances in the DeCODE female, male and sex-averaged meiotic maps (Fig. 1). We found the female recombination distances for chromosome 16 were similar to other human chromosomes, showing a relatively linear relationship between recombination and physical distances at an average of 1.93 cM/Mb, excluding heterochromatin. However, the male meiotic map displayed substantial differences in the region from 17-72 Mb with a meiotic distance of only 22.5 cM, yielding an average of 0.50 cM/Mb. Finally, we found a marked increase in male recombination near the telomeres, exceeding 3 cM/Mb, consistent with other human chromosomes[15].

Gene Catalog

We manually curated gene models as previously described[16] and identified a total of 880 protein-coding gene loci (Table 1 and http://www.jgi.doe.gov/human_chr16) supported by 1670 full-length (or nearly full-length) transcripts. These provided an average of 1.9 annotated transcripts per locus with 450 of the loci showing strong evidence for alternative splicing with 2 or more annotated mRNA transcripts. Additionally, 208 loci have "expressed sequence tag" (EST) evidence for alternative splice forms, resulting in nearly 75% of loci displaying some evidence for alternative splice variants. Loci were further classified as either: 'known genes', 'novel genes' or 'pseudogenes', consistent with our previous definitions[16], excluding loci without unique open reading frames and *ab initio* predictions without supporting evidence. Of the 'known genes', 771 were modeled based on 2,435 Refseq transcripts as well as other cDNA sequence evidence in GenBank. Comparison of these 'known genes' with Refseq revealed 36% of transcripts were extended by more than 50 bp at the 5' end and 18% at the 3' end while maintaining their original open reading frame.

We identified thirty 'novel genes' based on cDNA sequence, spliced ESTs, and/or protein similarity to known human or mouse genes and we modeled an additional 79 putative 'novel genes' using orthologous mouse cDNA sequences and *ab initio* predictions. We also annotated 19 tRNA genes and three tRNA pseudogenes based on previous data[17]. Finally, we identified 341 pseudogenes and pseudogene fragments of which 120 appear to be non-processed since they displayed an exon structure similar to the parent locus and are therefore likely to have resulted from genomic duplication events. The remaining 221 appear to be processed pseudogenes,

presumably resulting from viral retro-transposition of spliced mRNAs or from mitochondrial genome insertion. At least one frameshift or premature stop codon (in comparison to the parent gene) was identified in 233 pseudogenes and the remaining 108 were processed pseudogenes lacking introns and displaying poly-A's in the adjacent genomic sequence. This supports the likely nonfunctional nature of these vestigial genes. To assess the quality of our pseudogene collection, we compared it to an earlier analysis[18] describing 250 processed pseudogenes on chromosome 16. Initially we were able to map 233 of these 250 pseudogenes to 429 loci on chromosome 16 using BLAT[19] with 100% coverage and >99% identity. We then eliminated loci consisting of repetitive DNA[20, 21], those covering less than 50% of the parent gene and cases where there was clearly a retained intron/exon structure. This resulted in 146 processed pseudogenes in agreement between Zhang et al[18] and our study and suggested our manual curation of the finished sequence identified 58 additional members.

Large Structural Polymorphisms

We observed several large structural polymorphisms based on the finished sequence of chromosome 16 which were often associated with segmental duplications. For instance, we further characterized a previously described stable length polymorphism within the 16p subtelomeric region[22, 23]. While the shortest and most common allele was previously finished (represented in NCBI Build 34), we isolated and sequenced the majority of a longer allele derived from a 16p telomere half YAC, located within close proximity of the TTAGGG telomere repeat as defined by Riethman et al[9]. This allele is ~137.5 kb longer than the current assembly, however this allele is not simply a truncation of the longer form; rather the telomeric 21,056 bp

of the short allele is not present in the long allele and the telomeric 158,607 bp of the long allele is not shared with the short allele. Both of these unique regions contain genes with the short allele containing a putative gene(s) represented by cDNAs MGC:75272 and MGC:52000 and the long allele containing genes encoding hypothetical protein XP_375548 (similar to septin), hypothetical protein XP_379920 (similar to capicua) and beta-tubulin 4Q (AAL32434).

We also identified one of the most extensively duplicated regions on chromosome 16 corresponding to a 500 kb interval at 16p11.2-12.1 composed of approximately 54 intrachromosomal duplications (Supp table 2). This interval includes seven full or partial gene duplicates including the eukaryotic translation initiation factor 3, subunit 8 (*EIF3S8*), sulfotransferase 1A (*SULT1A1*) and the Batten disease gene (*CLN3*). Assembly of the region was initially complicated by the fact that the duplications were long (~200 kb) and showed an extraordinary degree of homology (98.33%). During the mapping of this region, sequence for a second haplotype variant from the RPCI-11 BAC library was completed except for one gap of ~100 kb. Sequence comparison of these two haplotypes (EIFvar1 and EIFvar2) revealed a 452 kb inversion between them (Fig. 2). Analysis of the breakpoints suggests that a large duplication palindrome is responsible for this rearrangement.

Finished sequence was also generated across a recently duplicated 360 kb polymorphism of the human homolog of the hydrocephalus inducing gene (*HYDIN*) at 16q22 which is inserted in some humans at chromosome 1q21.1. We observed that the RPCI-11 BAC library appears to be heterozygous for this insertional polymorphism with the current genomic assembly for chromosome 1 containing the haplotype version lacking the insertion. In addition, we further

investigated a recently described[24] copy number polymorphism between 16p11.2 and 6p25 which contains the *DUSP22* gene. Based on extensive drafting of RPCI-11 BACs in the region and comparisons with drafted clones from monochromosomal libraries for chromosomes 6 and 16, we were able to determine that the RPCI-11 library is homozygous and lacking the *DUSP22* duplication on chromosome 16. Taken together, these recently arisen large structural polymorphisms are striking examples of variability in the human genome and support a potential mechanism that contributes to phenotypic or disease susceptibility differences in humans. It is worth noting that 91 genes on chromosome 16 are located within segmental duplications, any of which could be unstable and challenge researchers studying phenotypes linked to these gene-containing regions. These observations are particularly relevant based on the recent findings[24, 25] of abundant copy number polymorphisms with the genomes of normal individuals.

Duplication Analysis of Chromosome 16

We performed a detailed analysis of duplicated genomic sequence (≥90% sequence identity and ≥1 kb in length) comparing chromosome 16 against the July 2003 assembly of the human genome. 9.89% (7.8 Mb) of chromosome 16 is found to consist of segmental duplications (Supp 1). Compared to other finished chromosomes, as well as the human genomic average (5.3%), chromosome 16 is quite enriched for segmental duplications (Supp 1, Supp 2). Nearly 9% of genome-wide human duplication alignments map to this chromosome. Intrachromosomal duplications are longer and show higher sequence identity when compared to interchromosomal duplications (Fig. 3a, Supp 3). While there is a general inverse correlation between duplication length and divergence, the effect is most pronounced for intrachromosomal duplication where

the average length of duplicated DNA exceeds 16 kb. A clear bimodal distribution pattern of sequence identity is distinguishable based on the distribution pattern of the alignments. The majority of interchromosomal duplication alignments show 93-95% sequence identity while intrachromosomal duplications show greater than 97% sequence identity, consistent with a recent expansion of intrachromosomal duplications along the chromosome[26, 27]. We estimate that as much as 7% of the mass of human chromosome 16 was added by segmental duplication events within the last 10 million years of human evolution.

Segmental duplications are particularly clustered along the p arm of the chromosome (Supp 2, Supp 4). As described previously[28], the 16p11 pericentromeric region represents the largest zone of interchromosomal duplications (Fig. 3b) accounting for 44% (937/2146) of the total number of chromosome 16 alignments (Supp 6) and 55% (752/1365) of all chromosome 16 interchromosomal alignments. Most of the interchromosomal duplications in this region map to the pericentromeric regions of other chromosomes (Fig 2b). Large-tracts of interstitial alpha-satellite DNA have been finished within proximal 16p11 and it is possible that such sequences have played a role in the frequent evolutionary exchange of pericentromeric DNA among non-homologous chromosomes[29]. In stark contrast to 16p11, there is little evidence for extensive pericentromeric duplication on the q arm, despite the fact that centromeric satellite boundary sequences have been traversed.

An additional 19 blocks of extensive duplication (>100 kb and > 5 duplication alignments) were identified within the euchromatic portion of chromosome 16. These regions are composed of as many as 119 underlying duplicons (also known as low-copy repeats on 16—LCR16(n)) that have

been juxtaposed in different combinations within the duplication blocks. These contain various genes and gene fragments, such as *NPIP*, *SULT1A*, *EIF3S8* and *SMG1* (Supp 5). Most are duplicated multiple times in varying copy numbers with a high degree of sequence identity to their putative ancestral genes. Most appear to have been duplicated in concert with LCR16a, a segment which contains one of the most rapidly evolving gene families of the human genome[27, 30].

Comparative Genomics

We compared human chromosome 16 versus the available dog, mouse, rat, chicken and fish (*Fugu rubripes*) draft genomes to further explore the evolution and function of sequences found along this chromosome. By first building segmental maps from DNA alignments of all the vertebrate species described above, we were able to examine the global homologous chromosomal relationships between these vertebrate genomes and human chromosome 16. Comparison versus the mouse and rat genomes revealed 28 chromosomal segments unbroken in any of the three species, ranging in size from 250 kb to 10.7 Mb (Fig. 4). In contrast, comparisons with the dog genome yielded 20 segments ranging in size from 250 kb to 11.8 Mb, and with the chicken genome resulted in 23 segments (the largest of which is 2.5 Mb). These findings are consistent with previous descriptions of an increased number of evolutionary rearrangement events within the rodent lineage and provide the substrate for the precise definition of these break points which may have disrupted gene loci in the species containing the rearrangement [31].

We next identified slowly evolving regions (p-value < 0.01), presumably under evolutionary constraint, through fine-scale DNA comparison of chromosome 16 with its orthologs in the rodent, dog, chicken and *Fugu* genomes. This chromosome-wide analysis was filtered for spliced ESTs, mRNA or protein coding sequences and resulted in the identification of 4,654 discrete conserved non-coding regions between human/mouse/rat, 5,498 between human/mouse/rat/dog, 2,883 between human/mouse/dog/chicken and 97 between human/mouse/*Fugu* (Fig. 4, Table 1). These elements represent candidate sequences for possessing biological activity in the ~98% of the human genome which is noncoding. We also compared the density of conserved noncoding sequences across the three chromosomes sequenced and annotated by the Joint Genome Institute which spanned a wide range of gene densities and segmental duplication frequencies (Table 1). While human chromosomes 5 and 16 contain ~50 conserved noncoding regions per Mb, gene rich chromosome 19 displays only 15, well below the genome wide average of 42. This is likely explained by the large number of recent gene family expansions on chromosome 19 which hinder comparative efforts to identify orthologous conserved sequences[16]. We also confirmed that the distribution of human/mouse/*Fugu* conserved noncoding sequences on chromosome 16 is highly uneven with ~40% (38) surrounding three Iroquois developmental transcription factor genes (*IRX3*, *IRX5* and *IRX6*). We found similar results on human chromosome 5 where a paralogous set of three Iroquois genes (*IRX1*, *IRX2* and *IRX4*) contained 42 out of the 213 human/mouse/*Fugu* conserved noncoding sequences found on this chromosome[32]. Interestingly, 9 of the 38 chromosome 16 human/mouse/*Fugu* elements in the *IRX* gene cluster contain significant similarity to noncoding sequence within the chromosome 5 *IRX* gene cluster. Furthermore, *in vivo* mouse transgenic data indicate that a significant percentage of these *IRX* conserved

noncoding sequences behave as gene enhancers[33]. These data support that in addition to the well described conservation of the protein encoding portions of genomic duplications, evolutionarily constraint is also observable in adjacent gene regulatory sequences follow genomic duplication events.

As an additional category of constrained DNA, we also searched for ultra-conserved noncoding sequences, recently defined by the stringent criterion of at least 200 bp in length and 100% identity between the human, mouse and rat genomes[34]. Of the 482 ultra-conserved elements found in the entire human genome, 15 (3.1%) were found on chromosome 16 with 11 having some evidence of being transcribed and processed into mature mRNAs. We found that only 2 of the 15 ultra-conserved elements on chromosome 16 are conserved with the *Fugu* genome, despite the extreme level of observed sequence identity between the human genome and each type of comparison. This supports that both human-rodent ultra-conserved and human-*Fugu* conserved sequences are complementary comparative strategies to identify highly constrained genomic sequence. Interestingly, similar to observations made between the human and *Fugu* conserved noncoding sequences, these ultra-conserved elements are biased towards development genes with 6 of the 15 being found near the embryonic transcription factors *SALL1* and the *IRX* gene cluster[35].

Based on the extreme features of having conserved synteny and minimal small rearrangement events in comparison to mouse, rat, dog and chicken, we explored a large 8.12 Mb region on chromosome 16 (located from 16q21 to 16q22.1) (Fig. 4). Remarkably, the telomeric 7.6 Mb of this segment contains only three annotated genes, all members of the cadherin family: *CDH8*,

CDH11 and *CDH5*. Within the full 8.12 Mb interval, we identified 621 human/mouse/rat and 278 human/mouse/rat/chicken conserved noncoding sequences, resulting in 76 and 34 respective elements per Mb (Fig. 4). This is 50% higher than the average density found on chromosome 16, and 75% higher than the overall genomic average, suggesting the enrichment of functional non-coding elements in this gene poor interval of chromosome 16. This observation of a large gene-poor region displaying a "forest" of conserved noncoding sequences parallels that found in gene deserts on human chromosome 5[32].

Finally, three regions on chromosome 16 have been selected by the National Human Genome Research Institute as part of the **ENC**yclopedia **O**f **D**NA **E**lements (ENCODE) project, an effort aimed at rigorously analyzing 1% of the human genome sequence (http://www.genome.gov/10005107). These three ENCODE regions include the well-studied alpha-globin containing interval (Enm008) and two randomly chosen regions (Enr211 on 16p12.1 and Enr313 on 16q21). Interestingly, Enr313 is a 0.5 Mb region located within the large 8.12 Mb gene desert describe above and is completely devoid of genes (Fig. 4). Nonetheless, it contains 43 human/mouse/rat and 16 human/mouse/chicken conserved noncoding sequences; again well above the chromosome-wide average, suggesting the presence of unassigned functional sequences within this region. Ongoing studies by ENCODE will better define the overlap of functionality and comparative sequence data such as that presented here.

Human Disease/Conclusions

As shown by substantial number of intra- and inter- chromosomal segmental duplications, chromosome 16 is extremely dynamic. This high plasticity suggests a mechanism whereby recent duplications on chromosome 16 can result in the production of morbid alleles, as extreme examples of large structural polymorphisms. For instance, there is ample evidence that thalassemias can result from unequal crossing over between the highly similar *HBA1* and *HBA2* loci[36, 37]. Furthermore, in the case of polycystic kidney disease, there are at least four closely-related loci on the chromosome with transcriptional evidence, suggesting a more complex relationship to disease than a single gene to a morbidity phenotype. Currently, nine protein-coding loci associated with morbid phenotypes are also associated with intrachromosomal segmental duplications (Table 2). However, there are still twenty morbid loci currently mapped on chromosome 16 for which coding sequence explanations have not been defined. It is anticipated that the completion of this chromosomal sequence will significantly lessen the challenge of uncovering the genetic basis of these disorders and in some cases their potential relationship to segmental duplications.

Methods

Segmental Duplication Analysis

We used a BLAST-based detection scheme[38] to identify all pair-wise similarities representing duplicated regions (≥1 kb and ≥90% identity) within the finished sequence of chromosome 16 and compared it to all other chromosomes in the NCBI genome assembly (build 34). A total of 2146 pair-wise alignments representing 26.12 Mb of aligned basepairs and 7.8 Mb of non-redundant duplicated bases were analyzed on chromosome 16. The program Parasight (http://humanparalogy.gene.cwru.edu/parasight/) was used to generate images of pair-wise alignments. Divergence of duplication, the number of substitutions per site between the two sequences, were calculated using Kimura's two-parameter method, which corrects for multiple events and transversion/transition mutational biases[39]. Analysis of haplotype structural variation was performed using the program *Miropeats* (threshold =3000)[40]. Gene content of each 1% duplicated regions of 90%-100% identity was analyzed using a non-redundant/non-overlapping set of known genes. A gene feature (exon) was considered duplicated if >50 bp of the feature overlapped duplication. Thus, exons less than 50 bp were lost in this analysis.

Pseudogene identification

Pseudogenes were defined as gene models built by homology to known human genes where the alignment between the model and the homolog shows at least one stop codon or frameshift. We identified homologies[41] of human IPI proteins on repeatmasked[20, 21] genomic chromosome 16 sequence. For each such fragment of genomic sequence we built gene models using the

GeneWise[42] program. Overlapping models were then clustered and the top-scoring model was analyzed for the presence of premature stop codons and frameshifts. Remaining models were then manually checked to confirm their pseudogene status.

Comparative Analysis

Multi-species segmental homology maps were computed using PARAGON (v2.2; Couronne, unpublished work), which is based on BLASTZ[42] pairwise alignments of all genomes to human. MLAGAN[43] alignments of homologous segments were scanned for evolutionarily conserved regions (p-value < 0.01) using GUMBY (Prabhakar, unpublished work). These were visualized using Rank-VISTA (Prabhakar, unpublished work). GUMBY goes through a 3-step process to identify statistically significant conservation. First, noncoding regions in the alignment are used to estimate the local neutral mutation rates[44] between all pairs of aligned sequences. The rates are used to derive a log-likelihood score for slow versus neutral evolution at each aligned position[45]. Conserved regions show up as high-scoring segments, which are assigned p values relative to random permutations of the alignment columns[46].

References

1. Siciliano, M. J. Chromosomal assignment of human genes coding for DNA repair functions. Isozymes Curr Top Biol Med Res 15, 217-23 (1987).
2. Griffith, B. B., Walters, R. A., Enger, M. D., Hildebrand, C. E. & Griffith, J. K. cDNA cloning and nucleotide sequence comparison of Chinese hamster metallothionein I and II mRNAs. Nucleic Acids Res 11, 901-10 (1983).
3. Hildebrand, C. E. & Enger, M. D. Regulation of Cd2+/Zn2+-stimulated metallothionein synthesis during induction, deinduction, and superinduction. Biochemistry 19, 5850-7 (1980).
4. Stallings, R. L., Munk, A. C., Longmire, J. L., Hildebrand, C. E. & Crawford, B. D. Assignment of genes encoding metallothioneins I and II to Chinese hamster chromosome 3: evidence for the role of chromosome rearrangement in gene amplification. Mol Cell Biol 4, 2932-6 (1984).
5. Han, C. S. et al. Construction of a BAC contig map of chromosome 16q by two-dimensional overgo hybridization. Genome Res 10, 714-21 (2000).
6. Doggett, N. A. et al. An integrated physical map of human chromosome 16. Nature 377, 335-65 (1995).
7. Cao, Y. et al. A 12-Mb complete coverage BAC contig map in human chromosome 16p13.1-p11.2. Genome Res 9, 763-74 (1999).
8. Kouprina, N. et al. Construction of human chromosome 16- and 5-specific circular YAC/BAC libraries by in vivo recombination in yeast (TAR cloning). Genomics 53, 21-8 (1998).
9. Riethman, H. C. et al. Integration of telomere sequences with the draft human genome sequence. Nature 409, 948-51 (2001).
10. Osoegawa, K. et al. A bacterial artificial chromosome library for sequencing the complete human genome. Genome Res 11, 483-96 (2001).
11. Schmutz, J. et al. Quality assessment of the human genome sequence. Nature 429, 365-8 (2004).
12. Dib, C. et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380, 152-4 (1996).
13. Kong, A. et al. A high-resolution recombination map of the human genome. Nat Genet 31, 241-7 (2002).
14. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. Am J Hum Genet 63, 861-9 (1998).
15. Yu, A. et al. Comparison of human genetic and sequence-based physical maps. Nature 409, 951-3 (2001).
16. Grimwood, J. et al. The DNA sequence and biology of human chromosome 19. Nature 428, 529-35 (2004).
17. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955-64 (1997).

18. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res 13, 2541-58 (2003).
19. Kent, W. J. BLAT--the BLAST-like alignment tool. Genome Res 12, 656-64 (2002).
20. Smit, A. & Green, P. (1999).
21. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 16, 418-20 (2000).
22. Wilkie, A. O. et al. Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. Cell 64, 595-606 (1991).
23. Flint, J. et al. The relationship between chromosome structure and function at a human telomeric region. Nat Genet 15, 252-7 (1997).
24. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. Science 305, 525-8 (2004).
25. Iafrate, A. J. et al. Detection of large-scale variation in the human genome. Nat Genet 36, 949-51 (2004).
26. Loftus, B. et al. Genome duplications and other features in 12 Mbp of DNA sequence from human chromosome 16p and 16q. Genomics 60, 295-308 (1999).
27. Johnson, M. E. et al. Positive selection of a gene family during the emergence of humans and African apes. Nature 413, 514-9. (2001).
28. She, X. et al. The structure and evolution of centromeric transition regions within the human genome. Nature Accepted (2004).
29. Guy, J. et al. Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. Genome Res 13, 159-72 (2003).
30. Eichler, E. E. et al. Divergent origins and concerted expansion of two segmental duplications on chromosome 16. J Hered 92, 462-8 (2001).
31. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520-62 (2002).
32. Schmutz, J. The DNA sequence and comparative analysis of human chromosome 5. Nature In Press (2004).
33. Pennacchio, L. A. unpublished observation.
34. Bejerano, G. et al. Ultraconserved elements in the human genome. Science 304, 1321-5 (2004).
35. Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. Nat Rev Genet 5, 456-65 (2004).
36. Orkin, S. H. et al. The molecular basis of alpha-thalassemias: frequent occurrence of dysfunctional alpha loci among non-Asians with Hb H disease. Cell 17, 33-42 (1979).
37. Lauer, J., Shen, C. K. & Maniatis, T. The chromosomal arrangement of human alpha-like globin genes: sequence homology and alpha-globin gene deletions. Cell 20, 119-30 (1980).
38. 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, 1005-17 (2001).
39. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16, 111-20 (1980).

40. Parsons, J. D. Miropeats: graphical DNA sequence comparisons. Comput Appl Biosci 11, 615-9 (1995).
41. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402 (1997).
42. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. Genome Res 14, 988-95 (2004).

Acknowledgements: We thank the International Chimpanzee Sequencing Consortium for pre-publication access to and permission to analyze the relevant portions of the chimpanzee genomic sequence and the Washington University Genome Sequencing Center for pre-publication access to the chicken genomic assembly. We also thank David Gordon of the University of Washington for his assistance in developing and customizing finishing tools, and Terry Furey and Greg Schuler for their efforts toward assessing the quality and completeness of our assemblyes. This work was performed under the auspices of the US DOE's Office of Science, Biological, and Environmental Research Program, by the University of California, Los Alamos National Laboratory, Joint Genome Institute and Stanford University.

This sequence is deposited in GenBank under accessions

Table 1: A comparison of selected properties of the three human chromosomes annotated at the JGI. PCG=Protein Coding genes; PCT=Protein Coding Transcripts; CNS=Conserved Noncoding Sequence.

	Chr 5	**Chr 16**	**Chr 19**
Gap-free size (finished bp)	177702766	78884754	55779685
Protein coding genes	923	880	1461
Pseudogenes	577	341	321
Avg. # Genes/Mb	5.2	11.2	26.2
Avg. % GC content	39.5	44.7	48.3
Protein cdng. Transcripts	1598	1670	2338
Ann. transcripts pr. Gene	1.7	1.9	1.6
%Alu coverage	8.4	16.4	25.8
%L1 coverage	18.5	11.8	10.0
%L2 coverage	2.7	2.6	2.2
Total % repeat masked	46.3	47.8	55.8
(Ensembl PCG)	1008	946	1377
(Ensembl PCT)	1320	1300	1972
# Genes	766	710	1133
# Genes/Mb	4.2	7.9	17.8
# Human/Rodent CNSes	10105	4654	962
#CNSes/Gene	13.2	6.6	0.85
#CNSes/Mb	55.8	51.7	15.1
# Human/Rodent/Dog/Chicken CNSes	4526	2883	558
#CNSes/Gene	5.9	4.1	0.49
#CNSes/Mb	25	32	8.7

Table 2: Disease genes located in segmental duplications on chromosome 16.

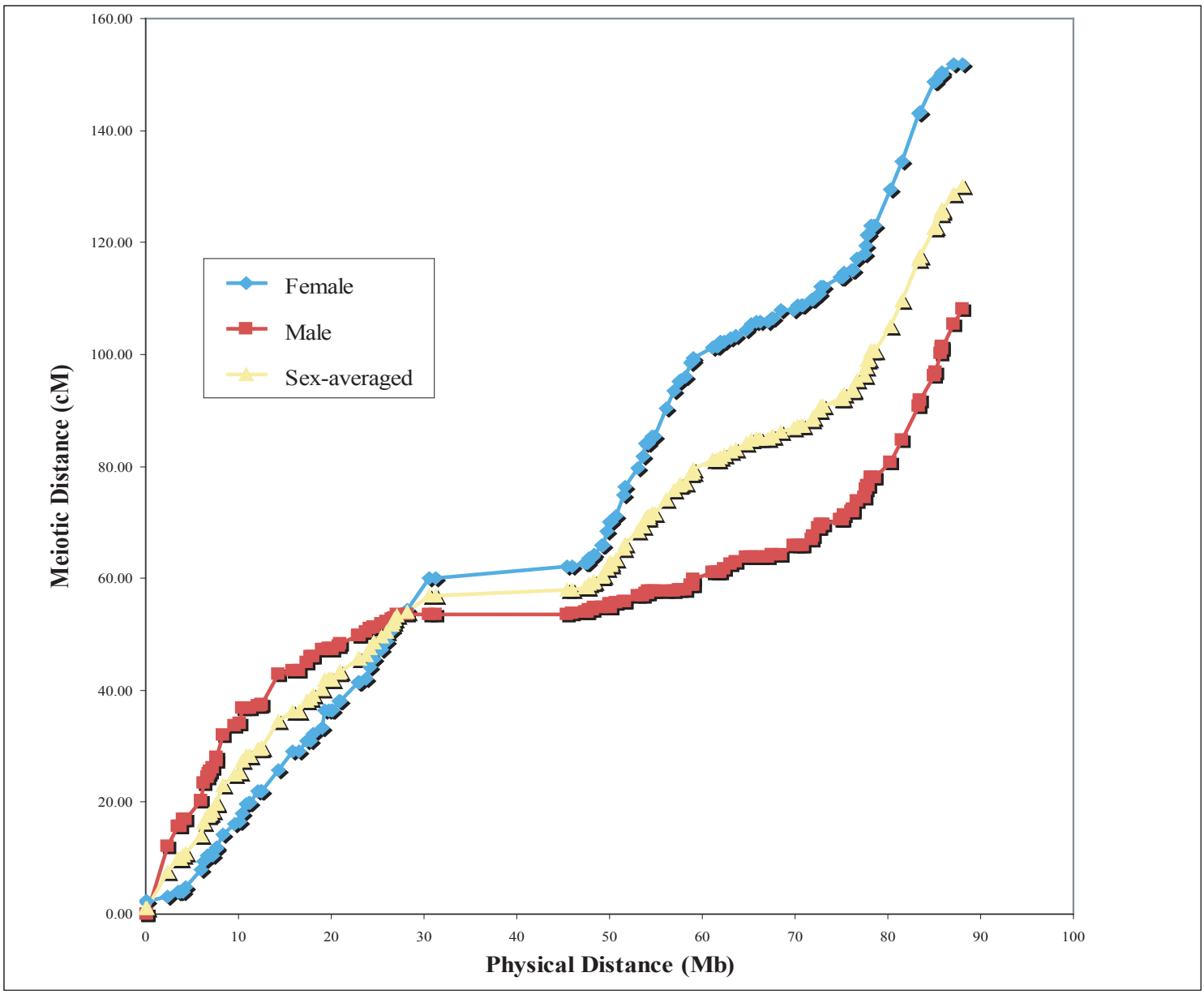
Gene Name	Disease
HBA1	Alpha-thalassemia
HBA2	Alpha-thalassemia
ABCC6	Multidrug resistance in cancer cells
HAGH	Deficiency of glyoxalase II
OTOA	Deafness
CLN3	Batten's disease (ceroid lipofuscinosis)
ALDOA	Aldolase A deficiency
CDH1	Multiple cancers
PKD1	Polycystic kidney disease

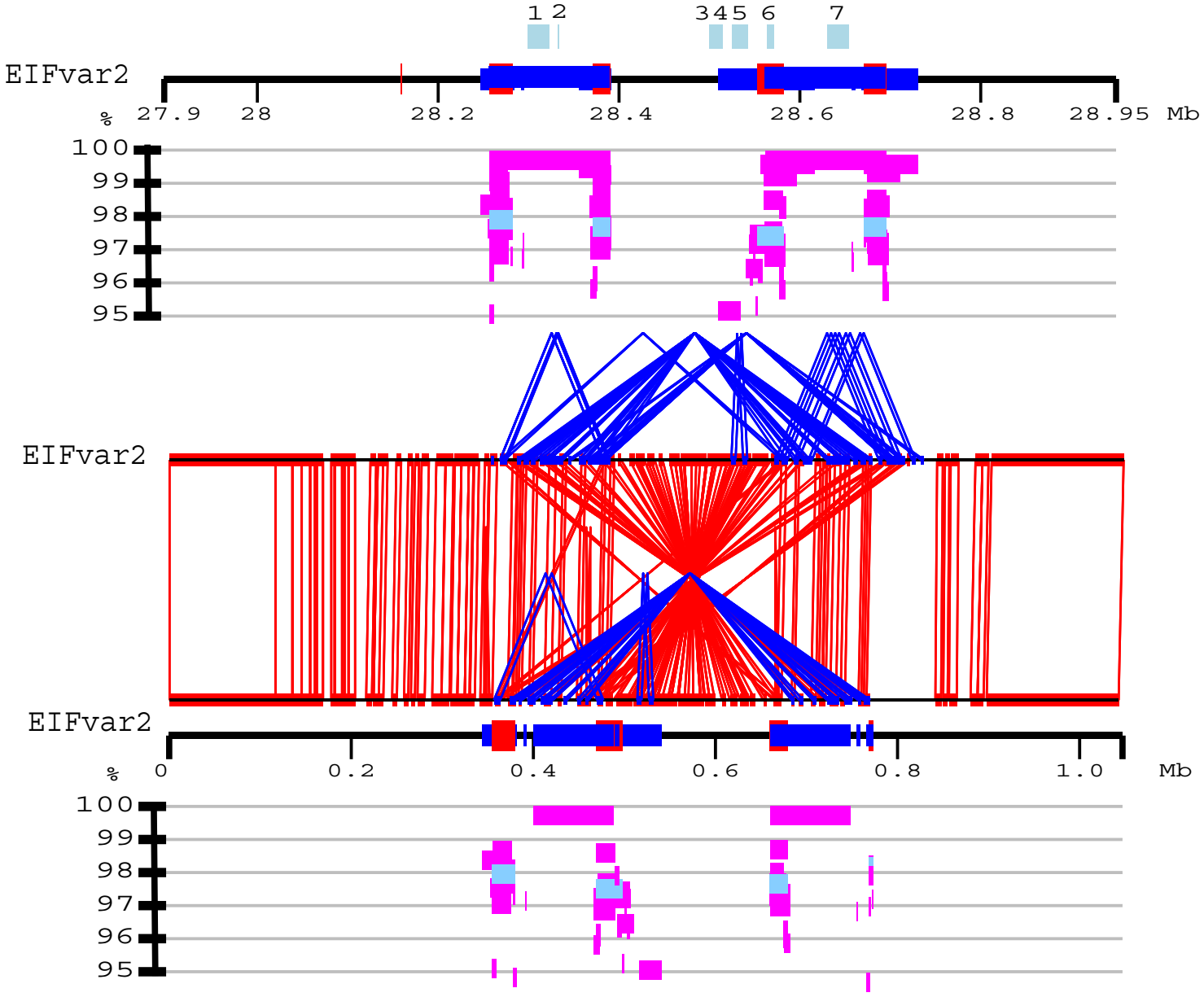
Figure 1: Comparison of meiotic distance to the physical map of chromosome 16, from the telomere of the short arm to the telomere of the long arm and reading left to right.

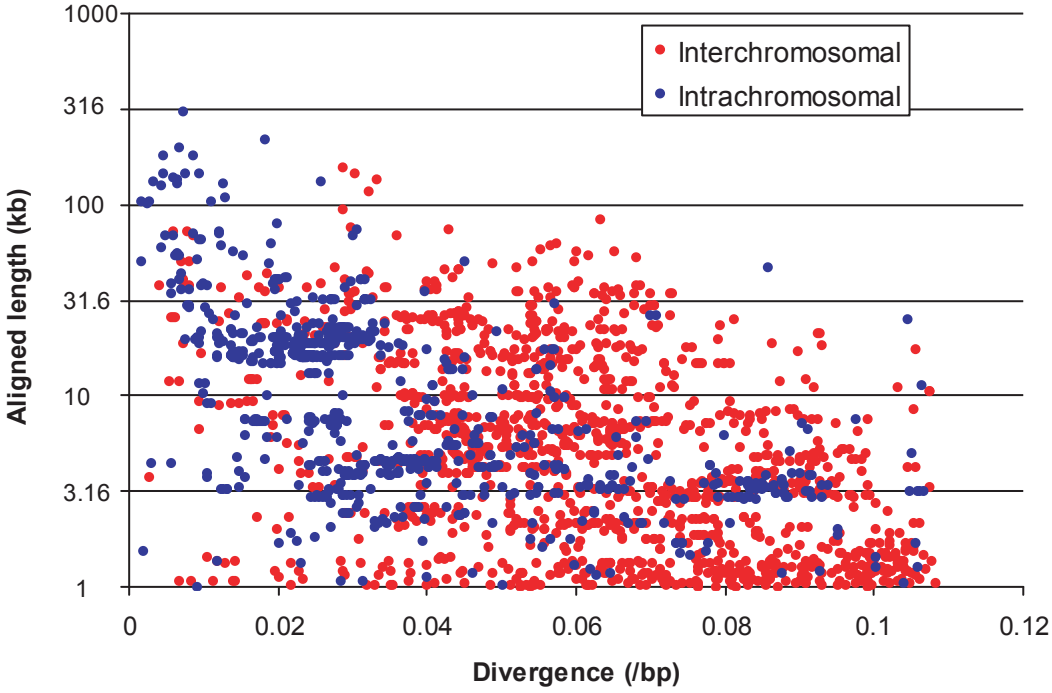
Figure 2: A 450 kb Inversion Haplotype on Chromosome 16. The duplication and inverted structure for two chromosome 16 haplotypes (EIFvar1 and EIFvar2) are compared. Top panel: Interchromosomal (red) and intrachromosomal duplications (blue) alignments (>90% >1kb) are depicted as a function of % identity below the horizontal line with different colors corresponding to the location of the pairwise alignment on different human chromosomes (i.e. chromosome 16 is shown as magenta, chromosome 18 as sky blue). The middle panel shows a 450 kb inversion between EIFvar1 and EIFvar2, using Miropeats (threshold=3000) [40] Interhaplotype (red) and intrahaplotype (blue) sequence alignments are shown based on chromosome assembly for EIFvar1. A palindromic duplication structure (200kb) demarcates the breakpoint region. Genes are depicted as light blue bars above the horizontal line in the top panel. These include: 1) eukaryotic translation initiation factor 3, subunit 8 (*EIF3S8*), 2) LOC39068, 3) LOC11286, 4) sulfotransferase 1A (*SULT1A2*), 5) sulfotransferase 1A (*SULT1A1*), 6) JGI-495, 7) *EIF3S8*.

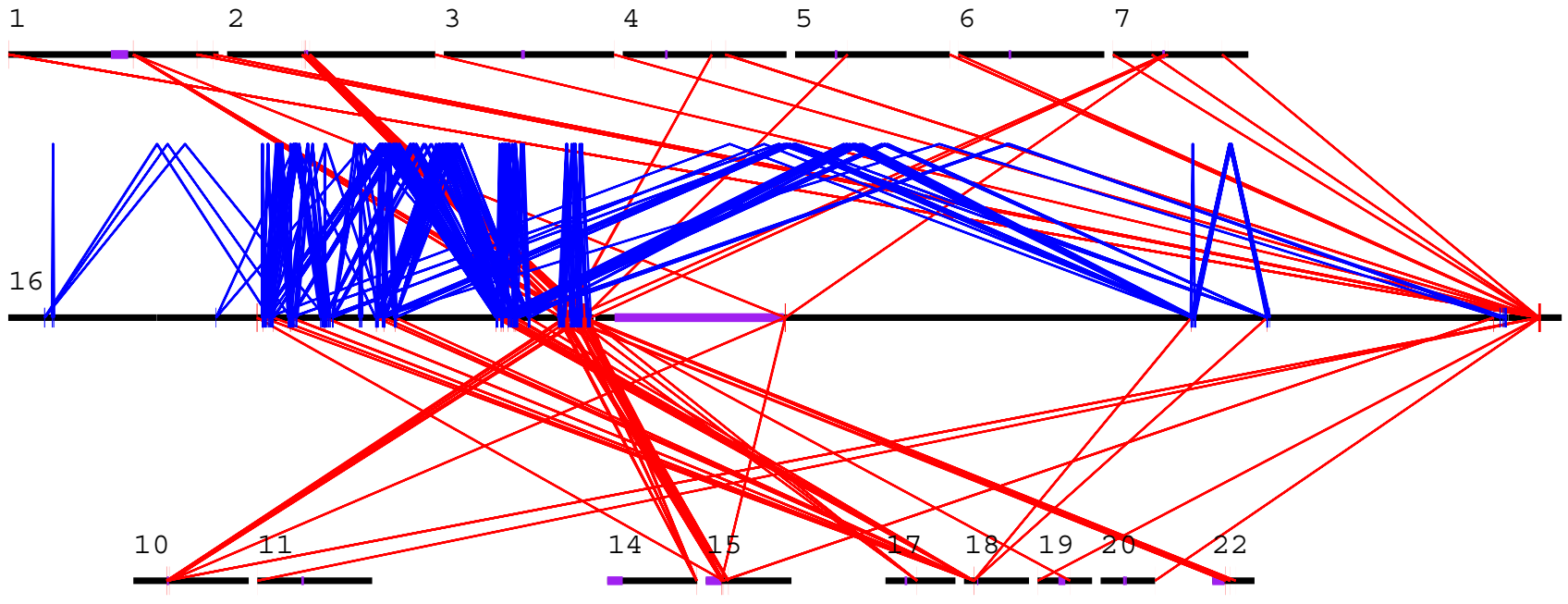
Figure 3: Chromosome 16 Segmental Duplications. (A) The scatter plot depicts the length (log 10) and divergence of inter- (red) and intra- (blue) chromosomal segmental duplication. Divergence (K) is calculated as the number of substitutions per site between the two sequences. (B) The parasight view depicts the pattern of interchromosomal (red) and intrachromosomal duplications (>20 kb, >95%) for chromosome 16. Chromosome 16 is drawn at 20X greater scale of the other chromosomes. Centromeres are shown as purple bars.

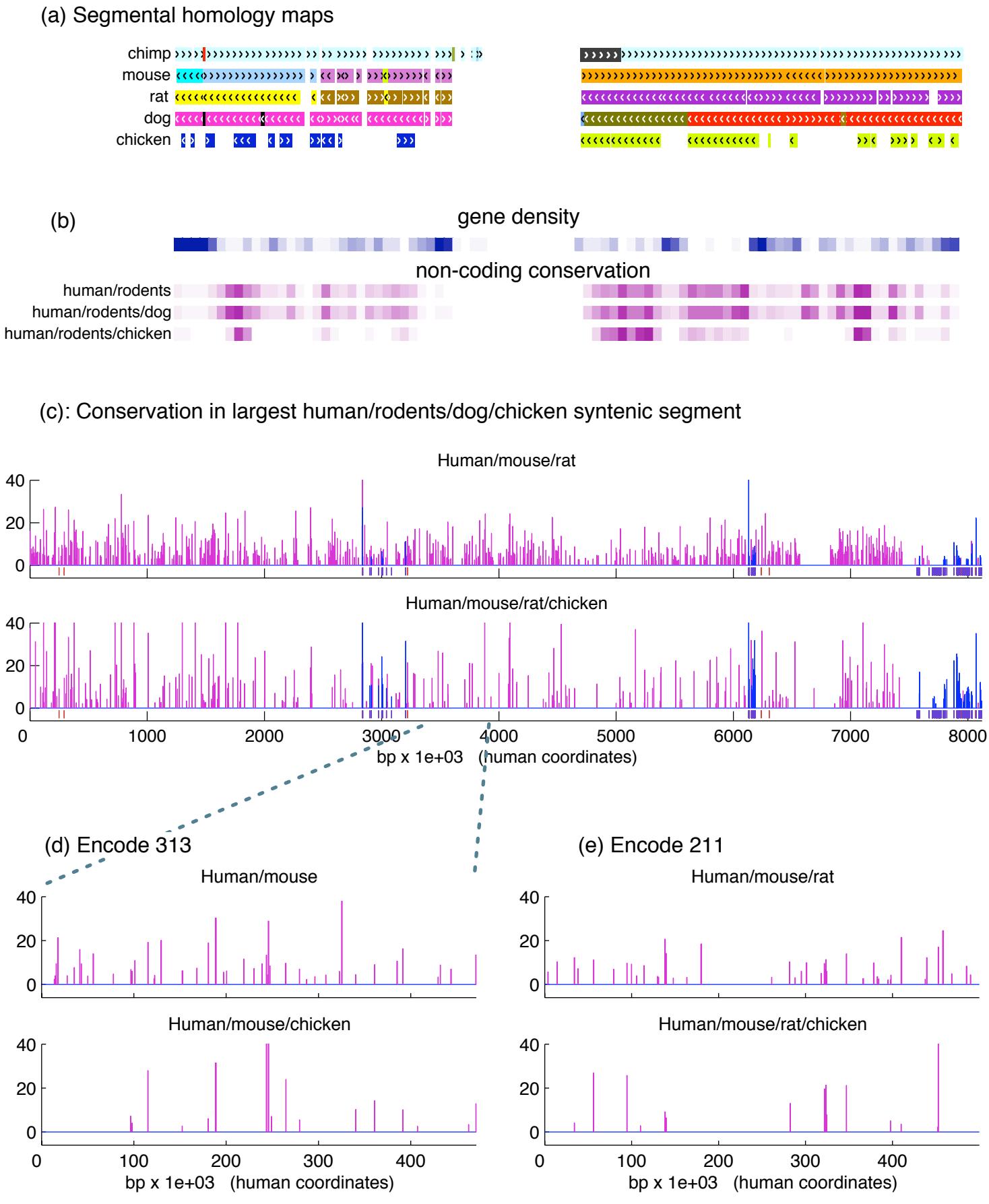
Figure 4: Comparative Analysis of Human Chromosome 16. **(A)** Segmental homology maps between human chromosome 16 and the mouse, rat, dog and chicken genomes (see Methods). (**B)** Normalized gene density (blue) and non-coding conservation density (magenta) over the entire chromosome. (**C)** Conservation in the largest human/mouse/rat/dog/chicken synteny block on human chromosome 16, which spans 8.12 Mb at 16q21 (*hg16*-chr16:58,625,483-66,746,256), and contains four cadherin genes. The upper plot shows coding (blue) and non-coding (magenta) conservation p-values in the human/mouse/rat comparison. The lower plot shows the human/mouse/rat/chicken comparison. (**D)** Similar plot of ENCODE Region 313 (*hg16*-chr16:62,053,179-62,549,053), which lies near the center of the gene-poor region in the previous subfigure. (**E)** ENCODE Region 211 (*hg16*-chr16:25,868,011-26,338,951), another gene poor region on 16p12.1. In Subfigures c, d and e, conserved elements are ranked by their statistical significance relative to the local neutral mutation rate. The height of the bars is proportional to – log (p-value) (GUMBY and Rank-VISTA, see Methods).


160.00
140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
Meiotic Distance (cM)
Female
Male
Sex-averaged
0
10
20
30
40
50
60
70
80
90
100
Physical Distance (Mb)


1 2
3 4 5 6
7
EIFvar2
%
27.9
28
28.2
28.4
28.6
28.8
28.95
Mb
100
99
98
97
96
95
EIFvar2
EIFvar2
%
0
0.2
0.4
0.6
0.8
1.0
Mb
100
99
98
97
96
95


Interchromosomal
Intrachromosomal
Aligned length (kb)
1000
316
100
31.6
10
3.16
1
0
0.02
0.04
0.06
0.08
0.1
0.12
Divergence (/bp)


1
2
3
4
5
6
7
16
10
11
14
15
17
18
19
20
22


(a) Segmental homology maps
chimp
mouse
rat
dog
chicken
(b)
gene density
non-coding conservation
human/rodents
human/rodents/dog
human/rodents/chicken
(c): Conservation in largest human/rodents/dog/chicken syntenic segment
Human/mouse/rat
Human/mouse/rat/chicken
40
20
0
0
1000
2000
3000
4000
5000
6000
7000
8000
bp x 1e+03 (human coordinates)
(d) Encode 313
Human/mouse
Human/mouse/chicken
100
200
300
400
(e) Encode 211
Human/mouse/rat
Human/mouse/rat/chicken