UCRL-JRNL-211171

The sequence and analysis of duplication rich human chromosome 16

J. Martin, C. Han, L. A. Gordon, A. Terry, S. Prabhakar, X. She, G. Xie, U. Hellsten, Y. M. Chan, M. Altherr, O. Couronne, A. Aerts, E. Bajorek, S. Black, H. Blumer, E. Branscomb, N. Brown, W. J. Bruno, J. Buckingham, D. F. Callen, C. S. Campbell, M. L. Campbell, E. W. Campbell, C. Caoile, J. F. Challacombe, L. A. Chasteen, O. Chertkov, H. C. Chi, M. Christensen, L. M. Clark, J. D. Cohn, M. Denys, J. C. Detter, M. Dickson, M. Dimitrijevic-Bussod, J. Escobar, J. J. Fawcett, D. Flowers, D. Fotopulos, T. Glavina, M. Gomez, E. Gonzales, D. Goodstein, L. A. Goodwin, D. L. Grady, I. Grigoriev, M. Groza, N. Hammon, T. Hawkins, L. Haydu, C. E. Hildebrand, W. Huang, S. Israni, J. Jett, P. B. Jewett, K. Kadner, H. Kimball, A. Kobayashi, M.-C. Krawczyk, et al.

April 7, 2005

Nature

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

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, Chitra F. Manohar#, 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-Gyamfị⇓, 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 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

ABSTRACT

Human chromosome 16 features one of the highest levels of segmentally duplicated sequence among the human autosomes. We report here the 78,884,754 base pairs of finished 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 among humans. While the segmental duplications of chromosome 16 are enriched in the relatively gene poor pericentromere of the p-arm, some are involved in recent gene duplication and conversion events likely to have had an impact on the evolution of primates and human disease susceptibility.

INTRODUCTION

The U.S. Department of Energy (DOE) initiated the mapping and sequencing of human chromosome 16 in 1988 with the aim of contributing to our understanding of radiation and its relationship to human biology. This particular chromosome was in part targeted for sequencing due to the localization of the DNA repair gene *ERCC4* to the p arm of chromosome 16[1], the availability of a unique flow-sorted chromosome-specific cosmid library[2], and access to a mouse-human hybrid cell panel enabling the localization of clones to discrete cytogenetic intervals[3]. Further interest in human chromosome 16 stemmed from the clustering of metallothionein genes on this chromosome which participate in heavy metal transport and detoxification coinciding with important biological interests of the DOE[4,5]. 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[6-8] with a minimal final tiling path 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 ~25 kb of missing sequence using alternative vectors[9]. The final gap is found near the telomere of the q arm in a region of subtelomeric repeats distal to the last identifiable cosmid subclone (AC137934) of a 16q telomere half-YAC as previously described[10].

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[11]. 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[12] 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, sized together at 9.8 Mb (see methods), 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[13] micro-satellite, the DeCODE[14], and the Marshfield[15] 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[16].

Gene Catalog

We manually curated gene models as previously described[17] and identified a total of 880 protein-coding gene loci (Table 1, Supplementary 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[17], 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. Additionally, we annotated 19 tRNA genes and three tRNA pseudogenes based on previous data[18]. 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[19] 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[20] with 100% coverage and >99% identity. We then eliminated loci consisting of repetitive DNA[21,22], 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[19] 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[23,24]. 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[10]. 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 with 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 (Fig. 2 and Supplementary 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. We further investigated a recently described[25] 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 among 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[25,26] of abundant copy number polymorphisms within the genomes of normal individuals, which include those described here.

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 (Supplementary Table 2). In comparison to other finished chromosomes, and to the human genomic average (5.3%), chromosome 16 is quite enriched for segmental duplications

(Supplementary Table 2 and Supplementary Fig. 1). 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 and Supplementary Fig. 2). 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[27,28]. Based on substitution rates between great apes 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[29].

Segmental duplications are particularly clustered along the p arm of the chromosome (Supplementary Fig. 1 and Supplementary Fig. 3). As described previously[30], 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 (Supplementary Table 4) 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. 3b). 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[31]. 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* (Supplementary Table 3). 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[28,32].

Comparative Genomics

We compared human chromosome 16 to the chimpanzee, dog, mouse[33], rat[34], chicken, and fish[35] (*Fugu rubripes*) draft genomes to further explore the evolution and constraint 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 (see Methods). We found no major rearrangements relative to the homologous chimpanzee chromosome 18. Comparison versus the mouse and rat genomes revealed 26 chromosomal segments unbroken in any of the three species, ranging in size from 250 kb to 10.7 Mb (Fig. 4a).

Further addition of the chicken genome to the multi-dimensional map yielded 33 segments ranging in size from 250 kb to 8.7 Mb (Fig. 4a). These segmental maps provide the substrates to precisely define the breakpoints that, in some cases, may have disrupted gene loci in the species containing the rearrangement.

We next identified slowly evolving regions, presumably under evolutionary constraint, through fine-scale DNA comparison of chromosome 16 with other vertebrate genome assemblies. Four different species combinations were selected to represent the accessible range of vertebrate evolutionary divergence times: human/mouse/rat, human/mouse/rat/dog, human/mouse/dog/chicken, and human/mouse/*Fugu* (see Methods). To explore potential non-coding functional elements on chromosome 16, the results were filtered for overlap with annotated genes, spliced ESTs or mRNAs in human, mouse and rat, which resulted in the identification of 5,187 discrete conserved non-coding regions between human/mouse/rat, 6,159 between human/mouse/rat/dog, 1,862 between human/mouse/dog/chicken, and 191 between human/mouse/*Fugu* (Fig. 4b, Supplementary Table 1). Compared to genome-wide averages, the densities of human/mouse/rat and human/mouse/dog/chicken elements were only slightly higher for human chromosome 16 (Supplementary Table 1). In contrast, human/mouse/*Fugu* elements are present at ~2.4 times the genome-wide density, indicating that although chromosome 16 as a whole has had "normal" levels of non-coding constraint since the mammal-bird split, it has conserved more ancient functions to a surprising degree. Functional studies on these conserved elements are warranted to assess their possible biological activity in the ~98% of the human genome which is non-coding.

We further explored an 8.7 Mb region at 16q12 based on extreme features of evolutionary conservation. This region was first identified as the largest unbroken synteny segment between human/mouse/dog/chicken on chromosome 16 and contains 59% (112/191) of the human/mouse/*Fugu* non-coding elements. These elements are entirely clustered in a gene-poor 5 Mb sub-region, which contains four developmental transcription factors: *SALL1* and three iroquois genes (*IRX3*, *IRX5* and *IRX6*). This clustering is an example of the general bias of human-fish conserved sequences towards developmental genes[36]. Interestingly, at least 9 of these human/mouse/*Fugu* elements have significant sequence similarity to counterparts in the paralogous *IRX* gene cluster on chromosome 5, which is similarly located in a "forest" of human-fish conservation[37]. *In vivo* mouse transgenic data indicate that a significant percentage of these *IRX* conserved non-coding sequences behave as gene enhancers[38], suggesting 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 following genomic duplication events. This synteny block is an outlier even in terms of more recent non-coding conservation, with 917 (105/Mb) human/mouse/rat and 590 (67.5/Mb) human/mouse/dog/chicken elements.

The second longest chromosome 16 synteny block in human/mouse/dog/chicken neighbors the highly conserved *SALL1-IRX* segment and is similar in length (8.19 Mb) (Fig. 4c). Once again this region is gene poor, with its telomeric 7.6 Mb containing only three annotated genes, all members of the cadherin family: *CDH8*, *CDH11* and *CDH5*. Within the full 8.19 Mb interval, we identified 968 (118/Mb) human/mouse/rat conserved non-coding sequences. This is twice the genome-wide density, as was the case in the *SALL1-IRX* region. However, in stark contrast to the neighboring *SALL1-IRX* region, this synteny block has no non-coding conservation between human/mouse/*Fugu* suggesting that its non-coding functions, though just as constrained among mammals, are more diverged in distant species.

As a special category of constrained DNA, we also searched for ultra-conserved non-coding sequences, recently defined by the stringent criterion of at least 200 bp in length and 100% identity between the human, mouse and rat genomes[39]. 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. The above-mentioned bias towards developmental genes has also been noted[39] for ultra-conserved human/rodent elements. Indeed, 9 of the 15 ultra-conserved elements found on chromosome 16 lie in the same *SALL1-IRX* synteny block that contains the mammal/fish conservation cluster. This contrasts with the similarly sized cadherin synteny block that contains no human-fish non-coding conservation and only one ultra-conserved element.

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[40] (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 located within the large cadherin gene desert described above and is completely devoid of genes (Fig. 4d). Nonetheless, it harbors the same high density of human/mouse/rat and human/mouse/dog/chicken conserved non-coding elements as the rest of the cadherin synteny block, suggesting the presence of numerous 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.

Conclusions

The primary sequence of human chromosome 16, as well as the human genome as a whole, now provides a key foundation for ongoing efforts such as ENCODE to deeply annotate all types of information encoded in our genome. This represents an enormous long-term challenge since genomic signatures embedded within the sequence of DNA perform a vast number of different operations across the trillions of cells within our bodies. These features range from relatively easily identified genes, to sequences involved in gene regulation, which use a plethora of signals to determine when and where a given gene is expressed and under what conditions, to likely even more complicated features such as higher order chromosome structure and DNA involvement in replication and repair. It is inspiring to reminisce that just 50 years ago was our first glimpse into the structure of DNA which provided the foundation for our ability to generate the nearly entire human euchromatic sequence. The next 50 years will likely also bring similarly impressive gains and enable us to precisely relate our primary genomic sequence to functional genomic signatures and their relationship to human biology.

Methods

Sizing of Heterochromatic Gaps

Te estimate the size of the alpha satellite bands (16p11.1-16q11.1) encompassing the centromere and the satellite II heterochromatin in band 16q11.2 we used CHEF pulsed-field gel electrophoresis at various pulse times to resolve macrorestriction fragments between 100 Kb and > 7000 Kb. DNA from CY18 (a mouse-human hybrid containing a single human chromosome 16) was digested with several different rare cutting restriction enzymes and separated on CHEF gels. Hybridization to blots of these gels with 16-1 (16 specific alpha satellite) and pHuR 195 (16 specific satellite II) probes revealed a single band of alpha satellite (in three different enzyme digests) that did not overlap with any satellite II bands (data not shown). The smallest of these bands was an 1800 Kb Xho I fragment which provided an upper size limit for the alpha satellite array, encompassing the centromere on chromosome 16. Sal I fragmented the satellite II heterochromatin into well resolved large restriction fragments without cutting within the alpha satellite array. The sum of the Sal I satellite II fragments was estimated at ~7800 Kb providing a upper size limit of the 16q11.2 satellite II heterochromatin at nominally 8 Mb. Together these account for 9.8 Mb of unsequenced heterochromatin encompassing cytogenetic bands 16p11.1-16q11.2, although it is likely that we did sequence partially into the boundaries of these regions in the adjacent tiling set clones.

Segmental Duplication Analysis

We used a BLAST-based detection scheme[41] 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[42]. Analysis of haplotype structural variation was performed using the program *Miropeats* (threshold =3000)[43]. 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[44] of human IPI proteins on repeatmasked[21,22] genomic chromosome 16 sequence. For each such fragment of genomic sequence we built gene models using the GeneWise[45] 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[46] pairwise alignments of all genomes to human. After filtering out segments shorter than 250 kb in humans, MLAGAN[47] alignments of homologous blocks were scanned for evolutionarily conserved regions using GUMBY (v1.5; Prabhakar, unpublished work). These were visualized using Rank-VISTA (Prabhakar, unpublished work). GUMBY goes through a 3-step process to identify statistically significant conservation in the input global alignment: 1) First, non-coding regions in the alignment are used to estimate the local neutral mutation rates[48] between all pairs of aligned sequences. The rates are used to derive a log-likelihood scoring scheme for slow versus neutral evolution[49], where the slow rate is set at half the neutral rate. 2) Each alignment position is then assigned a conservation score using a phylogenetically weighted sum-of-pairs scheme. 3) Finally, a dynamic programming step scans the alignment for high-scoring segments (conserved regions) of any length. Conserved regions detected in this manner are assigned p-values using the same statistical formalism[50] as the BLAST algorithm[44]. Whereas BLAST assigns p-values relative to random permutations of the query and target sequences, GUMBY p-values relate to random permutations of the columns in the input alignment. Here, all the results were generated using a GUMBY p-value threshold of 0.01 and a baseline human sequence length of 100 kb. Conserved non-coding regions were defined as conserved segments that overlap annotated exons, spliced ESTs or mRNAs from human, mouse or rat over no more than 25% of their length. At a GUMBY p-value threshold of 0.01, 2.2% of the ungapped positions in the human genome were assigned to human/mouse/rat conserved non-coding segments.

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. Deaven, L. L. et al. Construction of human chromosome-specific DNA libraries from flow-sorted chromosomes. *Cold Spring Harb Symp Quant Biol* 51 Pt 1, 159-67 (1986).
3. Callen, D. F. et al. High-resolution cytogenetic-based physical map of human chromosome 16. *Genomics* 13, 1178-85 (1992).
4. Hildebrand, C. E. & Enger, M. D. Regulation of Cd2+/Zn2+-stimulated metallothionein synthesis during induction, deinduction, and superinduction. *Biochemistry* 19, 5850-7 (1980).
5. 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).
6. 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).
7. Doggett, N. A. et al. An integrated physical map of human chromosome 16. *Nature* 377, 335-65 (1995).
8. 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).
9. 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).
10. Riethman, H. C. et al. Integration of telomere sequences with the draft human genome sequence. *Nature* 409, 948-51 (2001).
11. Osoegawa, K. et al. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res* 11, 483-96 (2001).
12. Schmutz, J. et al. Quality assessment of the human genome sequence. *Nature* 429, 365-8 (2004).
13. Dib, C. et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380, 152-4 (1996).
14. Kong, A. et al. A high-resolution recombination map of the human genome. *Nat Genet* 31, 241-7 (2002).
15. 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).
16. Yu, A. et al. Comparison of human genetic and sequence-based physical maps. *Nature* 409, 951-3 (2001).
17. Grimwood, J. et al. The DNA sequence and biology of human chromosome 19. *Nature* 428, 529-35 (2004).
18. 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).

19. 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).
20. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-64 (2002).
21. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16, 418-20 (2000).
22. Smit, A. & Green, P. (1999).
23. Flint, J. et al. The relationship between chromosome structure and function at a human telomeric region. *Nat Genet* 15, 252-7 (1997).
24. 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).
25. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* 305, 525-8 (2004).
26. Iafrate, A. J. et al. Detection of large-scale variation in the human genome. *Nat Genet* 36, 949-51 (2004).
27. 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).
28. Johnson, M. E. et al. Positive selection of a gene family during the emergence of humans and African apes. *Nature* 413, 514-9. (2001).
29. Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* 68, 444-56 (2001).
30. She, X. et al. The structure and evolution of centromeric transition regions within the human genome. *Nature* Accepted (2004).
31. 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).
32. Eichler, E. E. et al. Divergent origins and concerted expansion of two segmental duplications on chromosome 16. *J Hered* 92, 462-8 (2001).
33. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-62 (2002).
34. Gibbs, R. A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428, 493-521 (2004).
35. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297, 1301-10 (2002).
36. Boffelli, D., Nobrega, M. A. & Rubin, E. M. Comparative genomics at the vertebrate extremes. *Nat Rev Genet* 5, 456-65 (2004).
37. Schmutz, J. The DNA sequence and comparative analysis of human chromosome 5. *Nature* In Press (2004).
38. Pennacchio, L. A. Unpublished observation.
39. Bejerano, G. et al. Ultraconserved elements in the human genome. *Science* 304, 1321-5 (2004).
40. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636-40 (2004).

41. 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).

42. 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).

43. Parsons, J. D. Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* 11, 615-9 (1995).

44. 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).

45. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* 14, 988-95 (2004).

46. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* 13, 103-7 (2003).

47. Brudno, M. et al. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* 13, 721-31 (2003).

48. Cooper, G. M. et al. Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res* 14, 539-48 (2004).

49. Boffelli, D. et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299, 1391-4 (2003).

50. Karlin, S. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Advances in Applied Probability*, 113-10 (1992).

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

Correspondence and requests for materials should be addressed to Edward M. Rubin (emrubin@lbl.gov) or Len A. Pennacchio (LAPennacchio@lbl.gov).

The sequence described here has been deposited in public databases with accession number NC_000016.

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, the Broad Institute for pre-publication access to the dog genome assembly, 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 assemblies. 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 and UC/LLNL (W-7405-Eng-48).

Table 1: Chromosome 16 sequence features. PCG=Protein Coding genes; PCT=Protein Coding Transcripts; CNS=Conserved Noncoding Sequence.

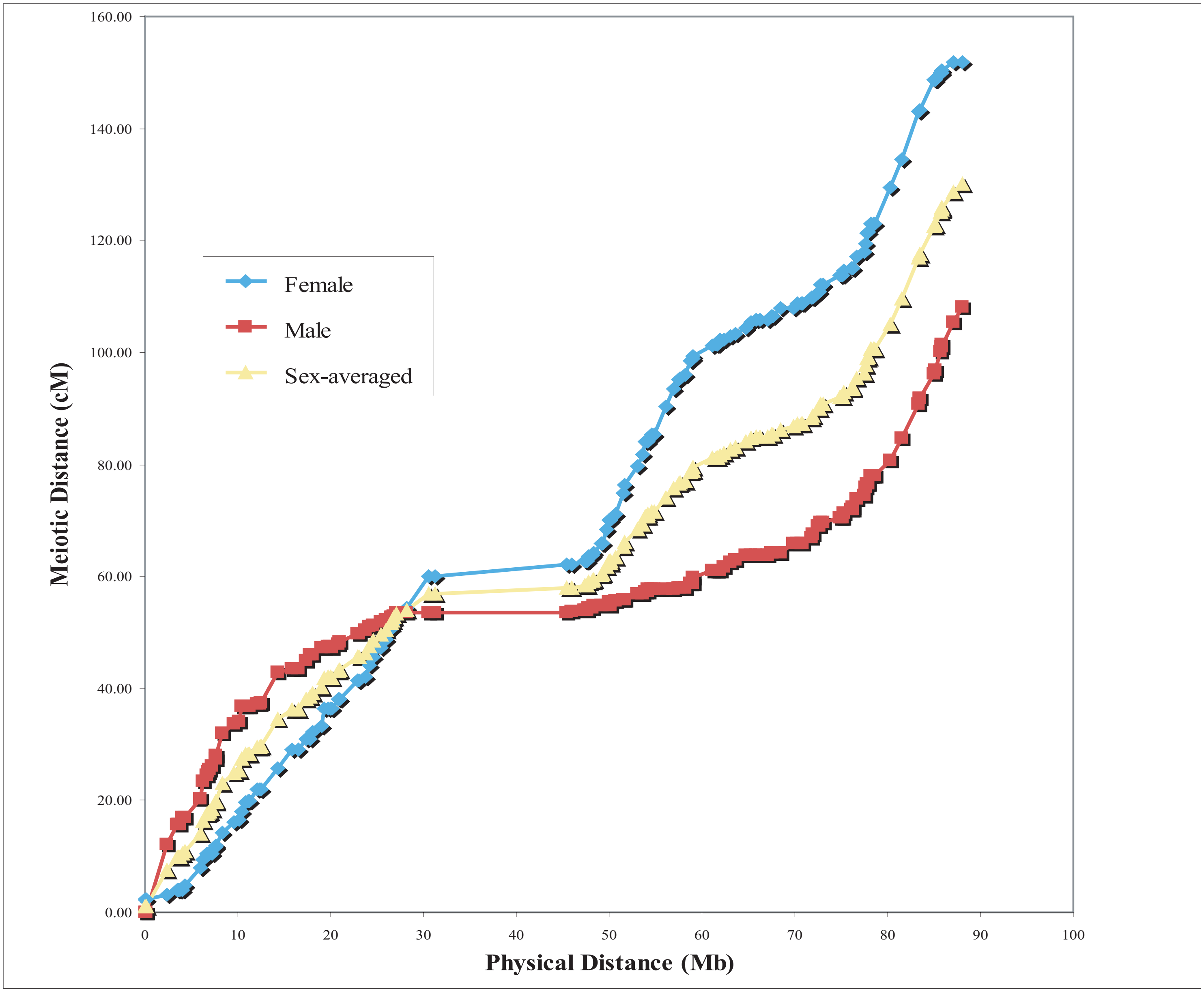
	Chr 16
Gap-free size (finished bp)	78,884,754
Protein coding genes	880
Processed Pseudogenes	221
Non-Processed	120
Protein coding Genes/Mb	11.2
Avg. % GC content	44.7
Protein coding Transcripts	1670
Average transcripts per Gene	1.9
%Alu coverage	16.4
%L1 coverage	11.8
%L2 coverage	2.6
Total % repeat masked	47.8
(Ensembl PCG)	960
(Ensembl PCT)	1441
Ensembl Genes/Mb	12.16
Ensembl Transcripts/Gene	1.5
Human/Rodent CNSes	5,187
CNSes/Gene	5.9
CNSes/Mb	65.8
Human/Mouse/Dog/Chicken CNSes	1,862
CNSes/Gene	2.1
CNSes/Mb	23.6

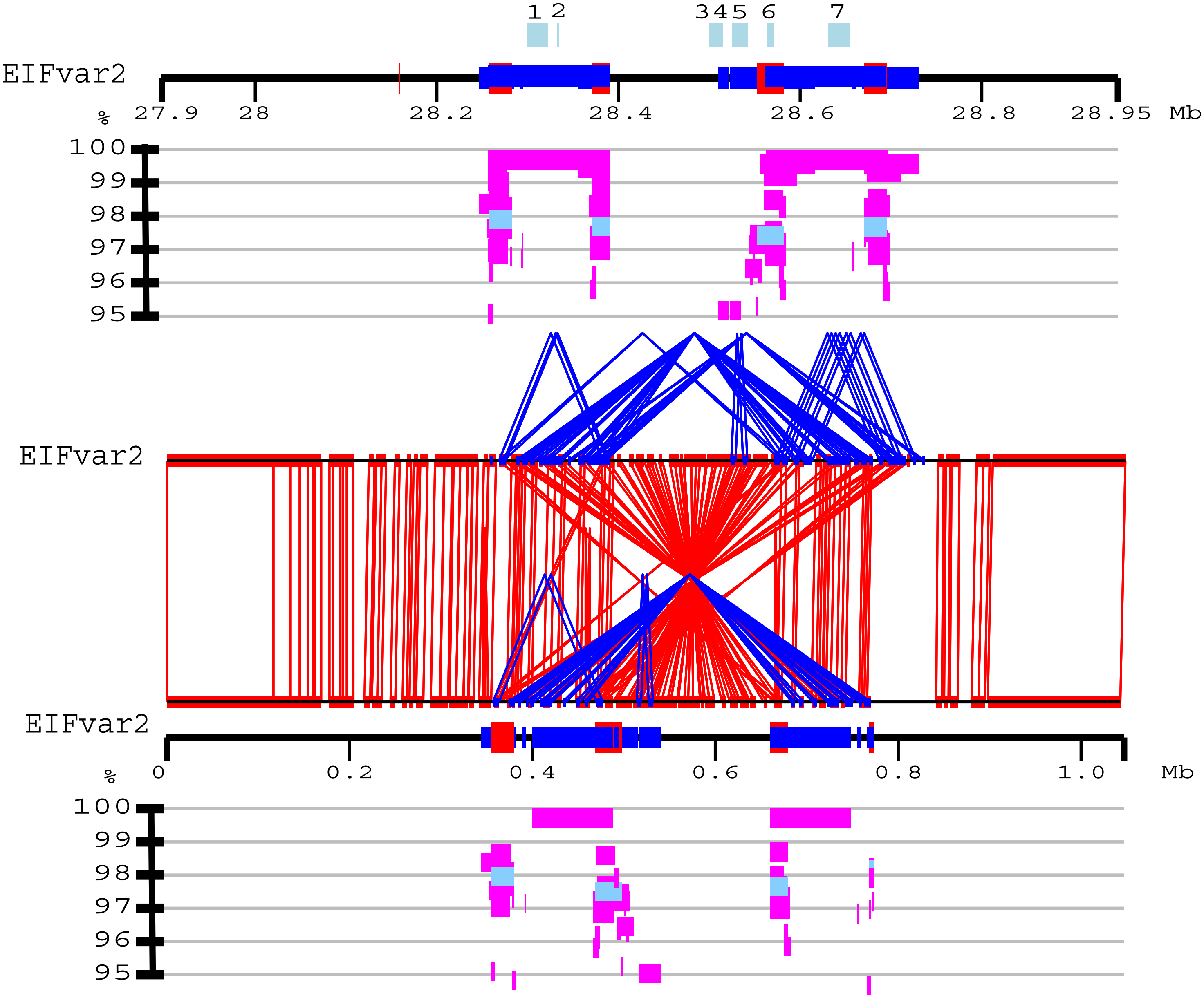
Figure 1: Comparison of meiotic distance to the physical map of chromosome 16, from the telomere of the short arm to the telomere of the long arm and reading left to right.

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%, >1 kb) 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). Interhaplotype (red) and intrahaplotype (blue) sequence alignments are shown based on chromosome assembly for EIFvar1. A palindromic duplication structure (200 kb) 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*.

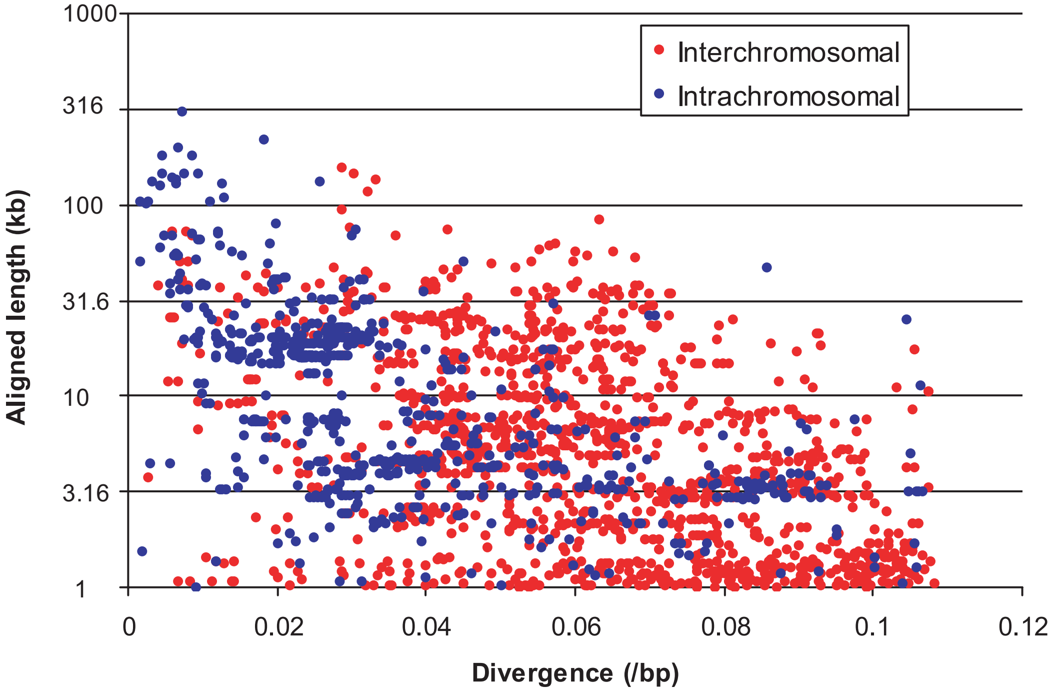
Figure 3: Chromosome 16 Segmental Duplications. **a**. The scatter plot depicts the length (log 10) and divergence of inter- (red) and intra- (blue) chromosomal segmental duplication. Divergence (K) is calculated as the number of substitutions per site between the two sequences. **b.** The parasight 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 chimpanzee, mouse, rat, dog and chicken genomes. Syntenic segments are color-coded by chromosome, with arrowheads indicating the strand. **b**, Gene density (blue) and non-coding conservation density (magenta) over the entire chromosome. Densities are normalized so that the darkest shade in each track denotes 3.5 times the genomic average. **c**, Conservation in the second-largest human/mouse/dog/chicken synteny block on human chromosome 16, which spans 8.19 Mb at 16q21 (*NCBI34*-chr16:58,626,110-66,811,606), 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/dog/chicken comparison. **d**, Similar plots of ENCODE Region ENr313 (*NCBI34*-chr16:62,051,662-62,551,661), which lies near the center of the gene-poor region in **c**. **e**, ENCODE Region ENr211 (*NCBI34*-chr16:25,839,478-26,339,477), another gene-poor region on 16p12.1. Rat is excluded because of a large sequencing gap. In **c**, **d** and **e**, the height of the bars is proportional to –log (conservation p-value) (GUMBY and Rank-VISTA, see Methods).

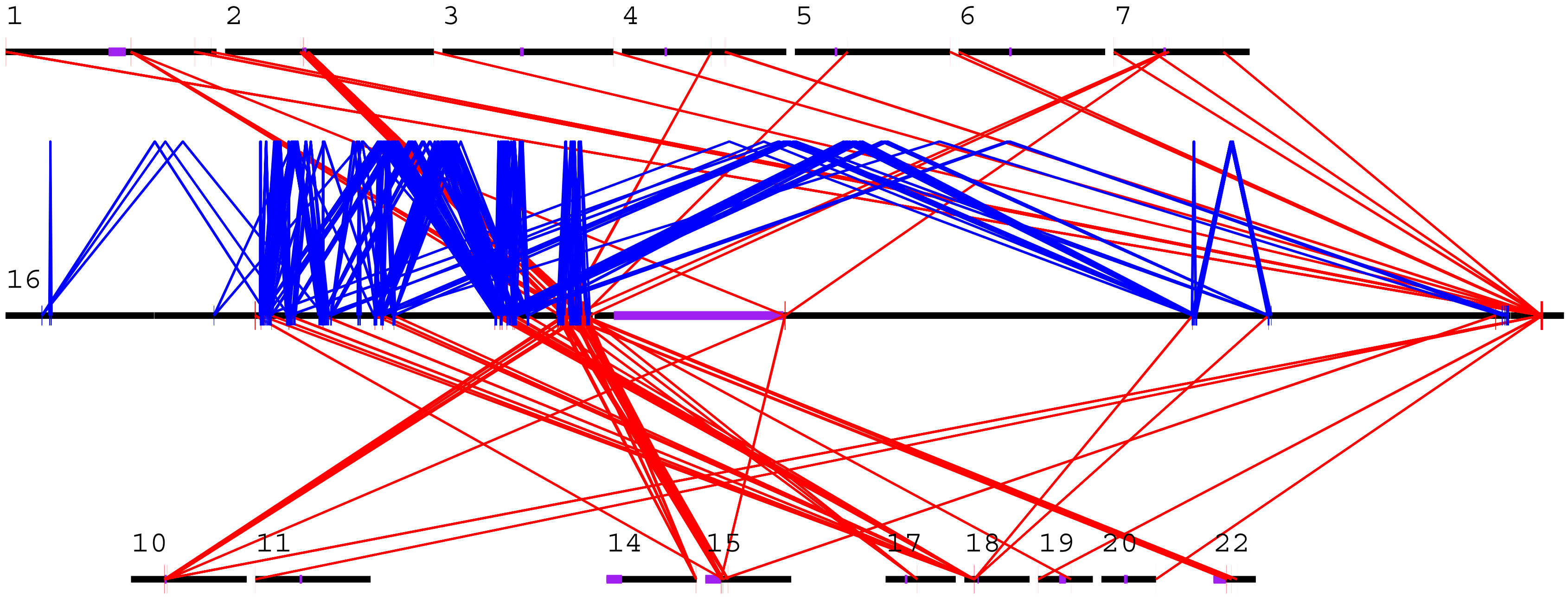

Meiotic Distance (cM)
160.00
140.00
120.00
100.00
80.00
60.00
40.00
20.00
0.00
0
10
20
30
40
50
60
70
80
90
100
Physical Distance (Mb)
Female
Male
Sex-averaged

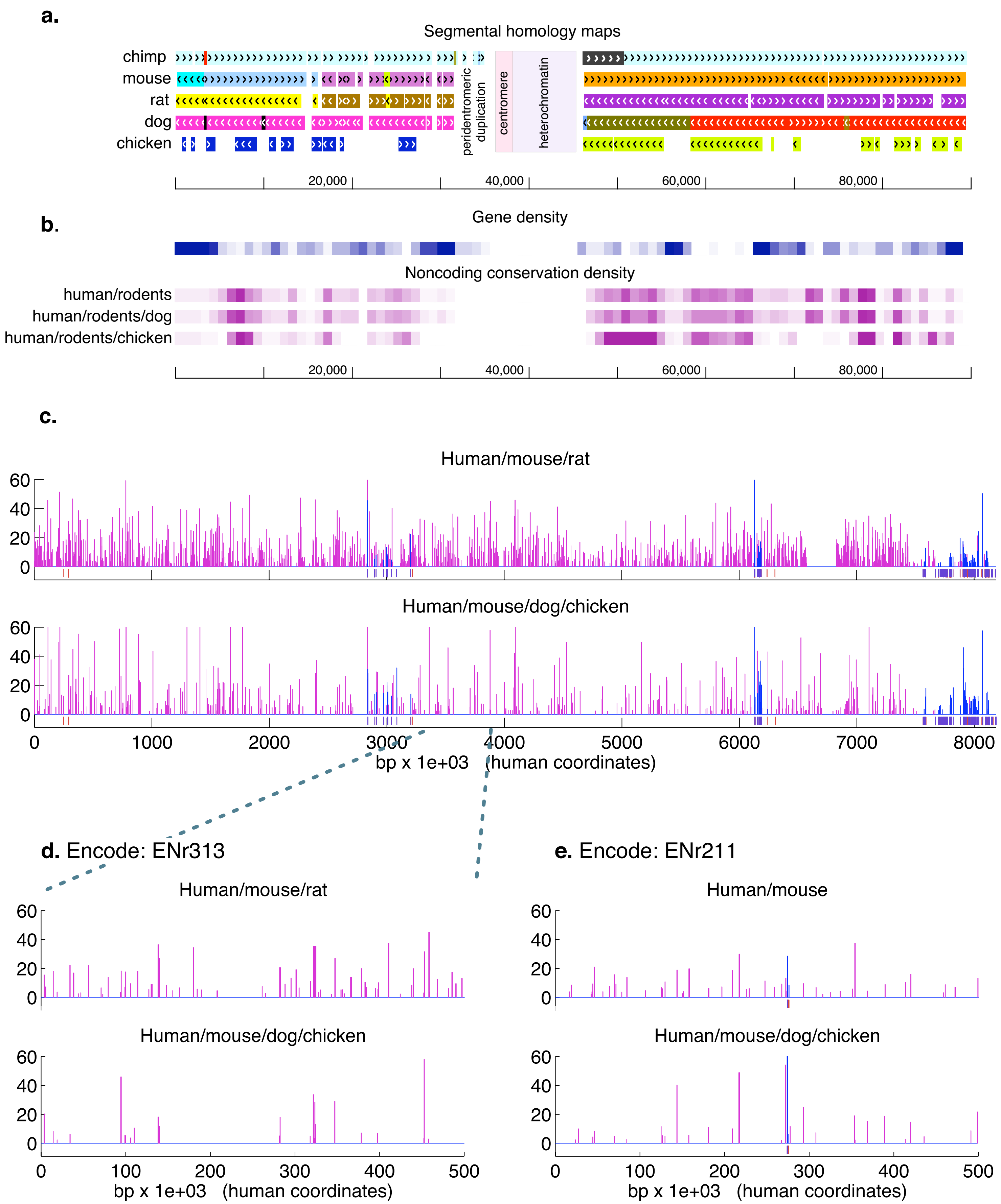


1 2
3 4 5 6
7
EIFvar2
27.9
28
28.2
28.4
28.6
28.8
28.95
Mb
%
100
99
98
97
96
95
EIFvar2
EIFvar2
0
0.2
0.4
0.6
0.8
1.0
Mb
%
100
99
98
97
96
95



Interchromosomal
Intrachromosomal
1000
316
100
31.6
10
3.16
1
Aligned length (kb)
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
peridentromeric duplication
centromere
heterochromatin
20,000
40,000
60,000
80,000
b.
Gene density
Noncoding conservation density
human/rodents
human/rodents/dog
human/rodents/chicken
20,000
40,000
60,000
80,000
c.
Human/mouse/rat
60
40
20
0
Human/mouse/dog/chicken
60
40
20
0
0
1000
2000
3000
4000
5000
6000
7000
8000
bp x 1e+03 (human coordinates)
d. Encode: ENr313
Human/mouse/rat
60
40
20
0
Human/mouse/dog/chicken
60
40
20
0
0
100
200
300
400
500
bp x 1e+03 (human coordinates)
e. Encode: ENr211
Human/mouse
60
40
20
0
Human/mouse/dog/chicken
60
40
20
0
0
100
200
300
400
500
bp x 1e+03 (human coordinates)