Recent Segmental Duplications in the Human Genome

Jeffrey A. Bailey,[1] Zhiping Gu,[2] Royden A. Clark,[1] Knut Reinert,[2] Rhea V. Samonte,[1] Stuart Schwartz,[1] Mark D. Adams,[2] Eugene W. Myers,[2] Peter W. Li,[2] Evan E. Eichler[1]*

Primate-specific segmental duplications are considered important in human disease and evolution. The inability to distinguish between allelic and duplication sequence overlap has hampered their characterization as well as assembly and annotation of our genome. We developed a method whereby each public sequence is analyzed at the clone level for overrepresentation within a whole-genome shotgun sequence. This test has the ability to detect duplications larger than 15 kilobases irrespective of copy number, location, or high sequence similarity. We mapped 169 large regions flanked by highly similar duplications. Twenty-four of these hot spots of genomic instability have been associated with genetic disease. Our analysis indicates a highly nonrandom chromosomal and genic distribution of recent segmental duplications, with a likely role in expanding protein diversity.

Initial analyses of the human genome sequence have identified a large amount of interspersed as well as tandem segmental duplications (*1*–*3*). These observations raise the possibility that segmental duplications may have played a significant role in gene and genome evolution compared with whole-genome duplication models (*4*). Furthermore, segmental duplications may underlie a greater amount of human phenotypic variation and disease than was previously recognized (*5, 6*). Unfortunately, duplicated regions of the genome are marginalized within both private and public assemblies (*7*). The overarching problem stems from the inability of current assembly strategies to differentiate highly similar duplicated sequence from true overlaps that remain unassembled.

Using computational methods, we have developed a simple statistical test to determine whether a given stretch of sequence is duplicated based on its overrepresentation and average sequence identity within a random sample of genomic sequence. Comparing a unique sequence with a random sample will detect a limited number of highly identical sequence matches. In contrast, a duplicated sequence will also detect paralogous matches, increasing the overall number of sequence alignments and decreasing the average pairwise sequence identity. The power of such an approach requires that the sample be randomly distributed and as large as possible. Currently, the largest sample available for these purposes is the about fivefold coverage of whole-genome shotgun (WGS) reads generated by Celera Genomics (*3*).

To test the random nature of this data set, we initially analyzed 27 autosomal and X chromosomal loci that had been determined to be unique by experimental analysis (Table 1) (table S1) (*8*). Genomic sequence from a public GenBank accession was used as a reference and compared against the WGS sequences over 5-kb windows, sliding every 1 kb across the accession. Within the unique control set, both the average read depth and average sequence identity were tightly distributed around their respective means indicative of a random sample of WGS reads. Next, we compared these statistics with 14 known loci (Table 1) (table S1) that contain recent (<40 million years ago) segmental duplications of various sizes, copy number, and divergence (*9*). We observed a significant increase in depth of coverage and significant decrease in sequence identity (Table 1), although the latter became more insensitive as the sequence identity of the duplicates approached 100%. Moreover, graphic visualization of both statistics allowed duplicated portions within the reference clones to be easily discerned within 2 kb of previously characterized junctions (Fig. 1A) (fig. S1). For known duplications with experimentally determined copy number, we assessed the depth of coverage specifically over the duplicated segments. The number of reads within 5-kb windows correlated strongly with the copy number (Fig. 1B; $R^2 = 0.96$). These data indicate that the WGS library is sufficiently deep and random to develop a duplication metric for large, highly homologous segmental duplications.

We chose to analyze independently each genomic accession underlying the public assembly of the human genome. We compared each sequence (32,610 clones) against the random WGS read data (27.3 million reads) and constructed a multiple sequence alignment based on the recruitment of sequence reads with >94% sequence identity. We computed the average degree of sequence identity and the depth of coverage in sliding windows of 5 kb along the alignment. The distribution of random reads and test statistics is available for each clone (*10*). In our analysis, we extracted all regions exceeding defined thresholds as potential segmental duplications and analyzed the read distribution to precisely delineate the boundaries of each duplicated region (Fig. 1A). We set our thresholds of duplication detection at 81 reads per 5 kb for autosomes and 47 reads per 5 kb for the sex chromosomes (3 SD beyond the mean, based on our analysis of unique regions) (Table 1) (table S1). With such a database of duplicated sequence, other sequences or assemblies could be screened and the positions of highly similar duplications determined. A consensus sequence from the multiple sequence alignment (both the public clone and WGS reads) was constructed if the clone showed an increased read depth (*8*). The consensus is analogous to consensus sequence for common repeat elements. The resulting segmental duplication database contains 8595 regions representing 130.5 megabases (Mb) of DNA. This sequence database is available [(*10*); see also the August 2001 assembly

[1]Department of Genetics, Center for Computational Genomics, and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, OH 44106, USA. [2]Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA.

*To whom correspondence should be addressed. E-mail: eee@cwru.edu

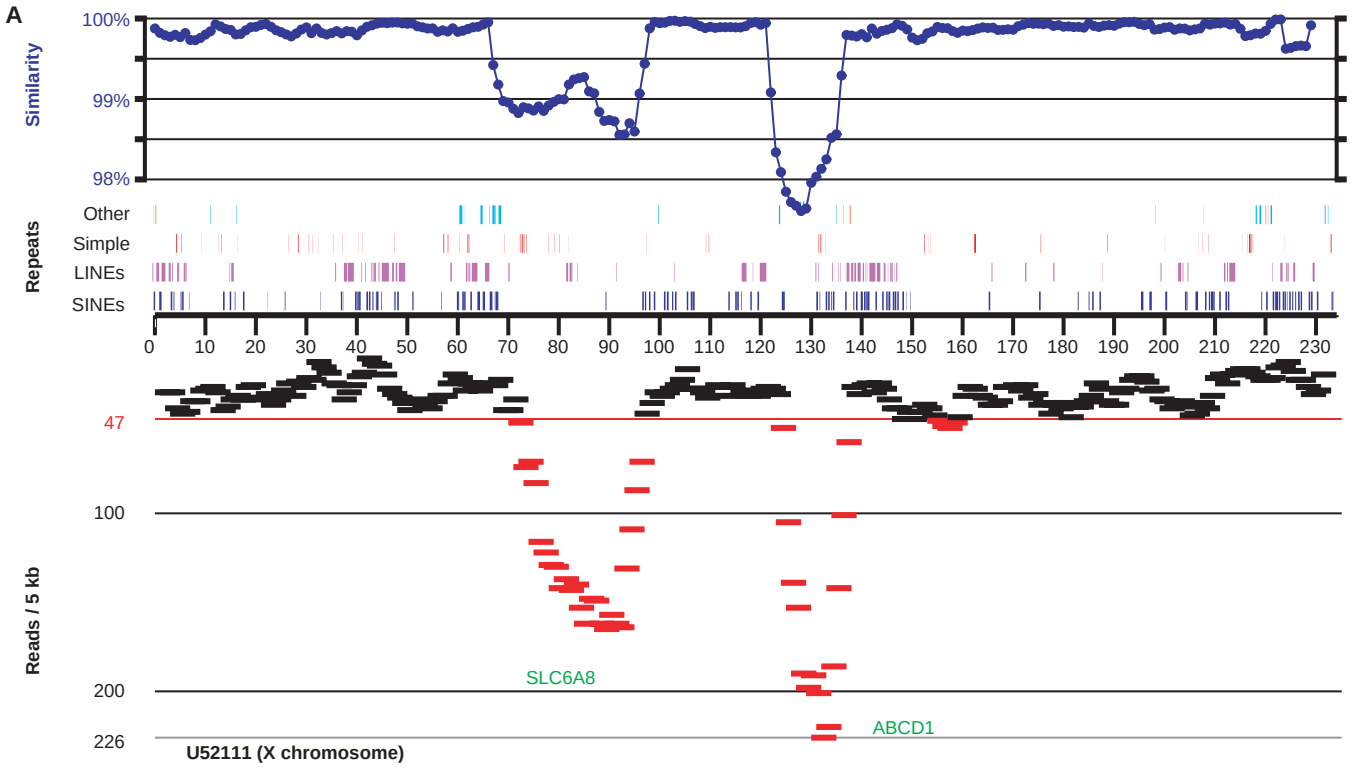
Table 1. Pilot study sequences.

	Sequence assessed		Number of reads per 5 kb†			Average percent identity		
	Number of loci	Total kb	Mean	SD	Maximum	Mean	SD	Minimum
Autosomal	19	2775	47.2	9.4	80	99.89	0.08	98.70
X chromosome	8	1243	28.2	6.47	46	99.89	0.19	98.33
Duplicated*	14	1379	228.6	256.13	1926	99.06	0.69	97.14

*Duplicated clones contained at least 50% known duplicated sequence (*9*). †Six instances of increased number of reads due to recently integrated transposable elements (including L1P and HERV elements) occurred. These were not included in the calculation of thresholds.

browser at the University of California, Santa Cruz (UCSC)].

We tested the power of this method to detect duplications in three ways. First, we analyzed the depth of coverage across human chromosome 22, whose segmental duplication pattern has been extensively characterized (fig. S2) (*11–13*). Unique regions (28 Mb of sequence) showed a narrow distribution of 50.4 ± 12.8 reads per 5 kb, which attests to the uniform nature of the WGS reads. Observed increases in read number that were false positive were almost exclusively due to the presence of high–copy number repeats, which were then filtered (*8*). Within duplicated regions, all duplications >10 kb and with >95% similarity had demonstrable increases in the number of reads per 5 kb. Second, we analyzed a set of duplicated BACs that had duplications detectable by fluorescence in situ hybridization (FISH) that also had been sequenced (table S2). We identified 36/37 of these BACs as duplicated based on our standards, which suggests a false-negative rate of 2.5%. A reciprocal experiment analyzing large-insert clones that tested positive with WGS detection (WSSD) showed 13/14 as being duplicated by metaphase and/or interphase FISH analysis (table S3). As a final test of sensitivity, we examined whether our thresholds could detect well-characterized duplications from the literature (table S4) (*6, 14, 15*). We analyzed a total of 27 genomic regions and detected all duplications of >15 kb and with >95% identity, many of which are associated with known genomic disorders. Because of our initial alignment parameters (*8*), duplications with a sequence identity of <94% were not



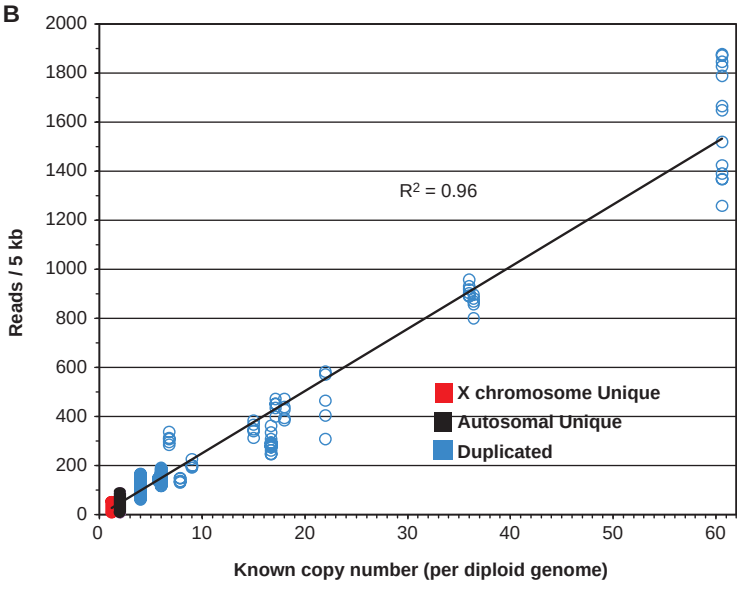


Fig. 1. WGS sequence detection of segmental duplications. (**A**) A genomic reference sequence (U52111) containing a 26.5-kb creatine transporter (SLC6A8) and 9.7-kb adrenoleukodystrophy (ABCD1) segmental duplication was used to search all WGS reads (Celera) using the combining assembler algorithm (*3*). This analysis was performed independently of the Celera assembly of the human genome. A multiple alignment (>94% sequence identity) was constructed and the number of reads and average sequence identity were calculated across 5-kb windows. The number of reads (*x* axis bottom) begins to rise and the average sequence identity (*x* axis top) drops precipitously, precisely at the known transition regions between unique and duplicated sequence (red horizontal line represents the X chromosomal threshold set at 3 SD above the mean depth coverage for unique X chromosome sequence). Both segmental duplications are readily identified. LINES and SINES are long and short interspersed repeat elements, respectively; also shown is a scale in 10-kb increments. (**B**) Correlation of number of WGS reads and known diploid copy number of genomic segment. The number of reads for each 5-kb window overlying known duplications (≥94% and ≥15 kb) was plotted against expected copy number. Segments with one copy (X chromosome) and two copies (autosome) represent unique loci used as controls (Table 1). A strong correlation between expected copy number and number of reads is found (R^2 = 0.96). Additional graphic representations of known segmental duplications are in table S1 and fig. S1.

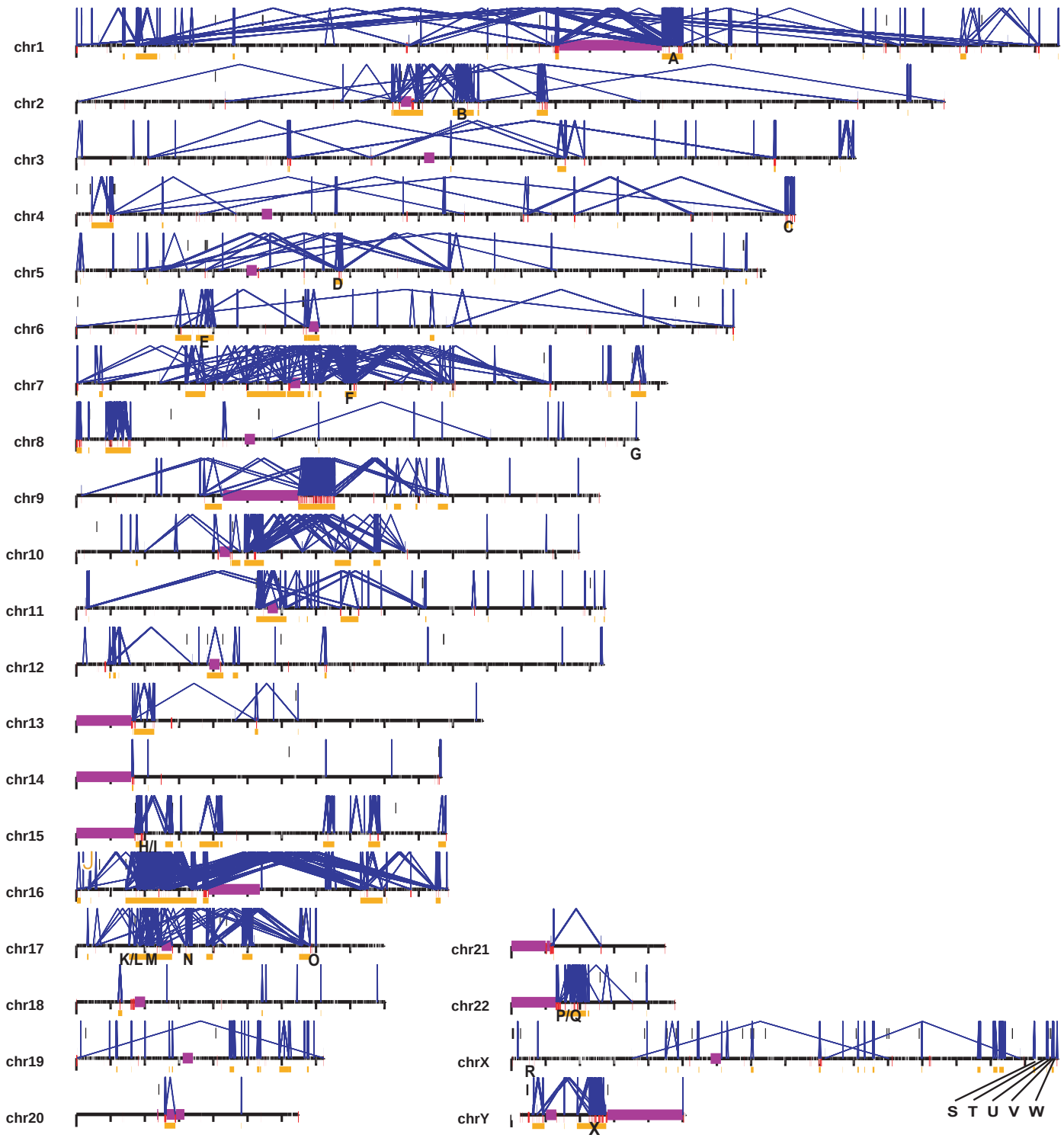


Fig. 2. Patterns of intrachromosomal and interchromosomal duplication (≥10 kb; ≥95%). The graphic shows a genome-wide view of intrachromosomal (blue, with connecting lines) and interchromosomal (red bars) segmental duplications. Purple bars represent areas (acrocentric chromosomal arms, heterochromatin satellite DNA, and centromeres) not targeted as part of the Human Genome Project. Unique regions (≥50 kb and <10 Mb) of the genome encompassed by intrachromosomal duplications (≥95% sequence identity and ≥10 kb) are shown as gold bars. Such regions are typically associated with recurrent chromosomal structural rearrangements associated with genetic disease. A total of 169 regions (~298 Mb of sequence) were identified as potential hot spots for genomic rearrangement. Twenty-four of these regions (labeled A to X) correspond to known genomic disorders: (A) Gaucher disease, (B) familial juvenile nephronophthisis, (C) fascioscapulohumeral muscular dystrophy, (D) spinal muscular atrophy, (E) congenital adrenal hyperplasia III, (F) Williams-Beuren syndrome, (G) glucocorticoid-remediable aldosteronism, (H) Prader-Willi syndrome, (I) Angelman syndrome, (J) polycystic kidney disease, (K) Charcot-Marie-Tooth disease type 1A, (L) hereditary neuropathy with liability to pressure palsies, (M) Smith-Magenis syndrome, (N) neurofibromatosis, (O) pituitary dwarfism, (P) cat eye syndrome, (Q) DiGeorge/velocardiofacial syndrome, (R) ichthyosis, (S) Hunter syndrome (mucopolysaccharidosis type II), (T) red-green color blindness, (U) Emery-Dreifuss muscular dystrophy, (V) incontinentia pigmenti, (W) hemophilia A, and (X) azoospermia (AZFc region). Specific regions corresponding to each region can be found at http://humanparalogy.cwru.edu/SDD/hotspots.htm. For details about patterns of interchromosomal duplications, see fig. S4.

reliably predicted within this set. Such duplications, however, are easily identified by genome assembly comparisons (see below).

Next, we performed a whole-genome assembly comparison (WGAC) to detect duplications (pairwise alignments ≥90% and ≥1 kb, as previously described) (*2*). WGAC is a BLAST-based strategy optimized to detect segmental duplications with intervening high–copy number repeats and large insertion deletions. This method is much more sensitive than the WSSD alone, as smaller alignments with lower sequence identity may be identified. However, it requires proper assembly of unique and duplicated sequences. The WGAC detected 16.5% of sequence as being putatively duplicated (fig. S3, red bars; table S5). A similar estimate was obtained from the National Center for Biotechnology Information assembly of the human genome (11.3%). Previous analyses suggested that four of five alignments with >98% identity are false positives due to a failure to merge allelic overlaps in the working draft sequence (*1, 2*). To remove these artifactual duplications, we filtered the WGAC alignments of >98% identity with the WSSD database (*8*). This removed 85% of the alignments with >98% identity and reduced the overall amount of duplicated genomic sequence to 5.2%, which agrees well with experimental and extrapolated estimates based on the finished sequence (*1, 2*). Using the UCSC Human Genome browser, we constructed an interactive site (http://humanparalogy.cwru.edu/SDD) to allow researchers to compare the details of various duplication detection strategies. The end result of this analysis is a highly curated set of segmental duplications that have been validated by at least two independent computational methods.

We also examined the impact of duplications on single nucleotide polymorphism (SNP) discovery by analyzing the content of the the public SNP database (dbSNP) as placed on the UCSC assembly (*16*). We hypothesized that when duplications remain unrecognized, paralogous sequence variants may be falsely identified as SNPs. This would increase the apparent density of "SNPs" within duplicated regions. The average SNP density was indeed increased in duplicated regions compared with unique regions (1.33 versus 0.69 SNP per kb, respectively; table S6). Because there is no reason to expect that polymorphic variation is increased within duplicated regions, the approximate doubling of SNP density suggests that roughly one of two SNPs is, in fact, a paralogous sequence variant rather than an allele. Current in silico methods examining sequence overlaps account for most of these false positives (table S6). We estimate that about 100,000 paralogous sequence variants currently contaminate dbSNP.

Nonallelic homologous recombination between blocks of duplicated sequence leads to microdeletion, microduplication, and inversion of genomic segments. If genes flanked by these duplications are rearranged, disease may result (*17–20*). To identify such potential regions of genomic instability, we assessed the pattern of intrachromosomal duplication (Fig. 2). The most prevalent disorders usually involve duplications that are >95% similar and >10 kb, separated by 50 kb to 10 Mb of DNA (*6*). Compiling the regions encompassed by duplications meeting these criteria creates a genome map of likely rearrangement hot spots (Fig. 2; gold bars below sequence). We identified a total of 169 regions constituting roughly one-tenth of the genome (298 Mb). Twenty-four of these regions have already been associated with genomic disorders.

Different human chromosomes appear to show distinct landscapes for segmental duplication (Fig. 2). Although interchromosomal duplications within pericentromeric and subtelomeric regions are well documented (*5, 21*), these biases have not been observed for all chromosomes. It appears that many pericentromeric regions such as 3p, 3q, 4p, 4q, 5p, 6q, 8p, 8q, 12p, 18q, 20q, Xp, and Xq are quiescent, showing no sign of recent duplication between chromosomes (Fig. 2) (fig. S4). Subtelomeric regions also show variability in duplication content. Final assessment must await further completion of the reference sequence because duplicated pericentromeric and subtelomeric regions are underrepresented relative to the rest of the genome.

To assess the duplication distribution more directly, we developed a random genome model of segmental duplication. The genome was partitioned into 2881 segments of 100 kb (fig. S3 and table S5), genome sequence was randomly assigned to each bin, and the duplication content for each chromosome was calculated (n = 10,000 replicates). Human chromosomes 7, 9, 15, 16, 17, 19, 22, and Y were significantly enriched for both inter- and intrachromosomal duplications, whereas chromosomes 2, 3, 4, 5, 8, 14, and 20 appeared to be significantly reduced for segmental duplication content ($P < 0.0001$). Such variation was not due to the finished state of the chromosomes with which there is no correlation ($R^2 = 0.04$) (Fig. 2) (fig. S4).

It has been argued that duplications may occur simply as a result of relaxed negative selection in gene-poor regions that have no function; thus, a negative correlation between gene density and duplication content would be expected for chromosomes (*22*). In fact, a significant positive, rather than negative, correlation is seen when the relative gene density is

Table 2. Protein domain enrichment within segmental duplications. The number of duplicated and unique genes belonging to each INTERPRO domain was determined on the basis of analysis of the RefSeq set of mRNAs for INTERPRO numbers with five or more domains and enrichment by a factor of 2 or more. We excluded genes that showed no evidence of intron-exon splicing to avoid potential contamination from processed pseudogenes, thereby removing both olfactory receptor and histone INTERPRO domains. The immunoglobulin genes were not ascertained, as they are currently not contained in the RefSeq mRNA set.

INTERPRO (entry number)	Description	Number		Enrichment*
		Duplicated	Unique	
003006	Immunoglobulin and major histocompatibility complex	38	280	4.0
001400	Somatotropin hormone family	17	1	31.8
001254	Serine proteases, trypsin family	11	75	4.3
001909	KRAB box	10	87	3.5
001128	Cytochrome P450 enzyme	8	41	5.5
002999	Tudor domain	6	21	7.5
001870	Domain in various γ-carboxylases	5	35	4.2
003877	SPla and the ryanodine receptor (SPRY)	5	42	3.6
001664	Intermediate filament proteins	5	42	3.6
000566	Lipocalin-related protein and Bos/Can/Equ allergen	5	21	6.5
000359	Cystine-knot domain	5	17	7.7
001039	Major histocompatibility complex protein, class I	5	9	12.0
001811	Small cytokines, interleukin 8–like	4	40	3.1
000436	Sushi domain/SCR repeat/CCP module	4	39	3.1
001545	Glycoprotein hormone β chain	4	2	22.5
001271	Mammalian defensin	4	2	22.5
000340	Dual-specificity protein phosphatase	3	39	2.4
003575	Small GTPase, Ras subfamily	3	24	3.7
004045	Glutathione S-transferase NH_2 terminus	3	18	4.8
000863	Sulfotransferase	3	16	5.3
001079	Galectins (previously S-lectins)	3	10	7.8
000971	Globin	3	8	9.2
000461	Glycoside hydrolase family 13	3	3	16.8
000353	Class II histocompatibility antigen, β chain, β_1	3	2	20.2

*Enrichment was calculated as the fraction of duplicated domains for an INTERPRO number over the average fraction for all INTERPRO domains detected in the genome (647 duplicated/21,147 total). Table S7 provides a complete list of all INTERPRO domains examined by this analysis.

compared with chromosomal duplication content (R^2 = 0.16). The correlation was due to intrachromosomal duplications (fig. S5; R^2 = 0.20; P = 0.04; F test) and was absent for interchromosomal duplications (R^2 = 0.002). The three most gene-rich chromosomes showed high levels of duplication, and the seven most gene-poor chromosomes were among the least duplicated chromosomes.

To determine what role recent segmental duplications have played in current gene evolution, we characterized the gene content in our filtered set of duplicated genomic sequence. We analyzed a highly curated set of 13,351 mRNAs assigned to the human genome assembly (RefSeq, www.ncbi.nlm.nih.gov/LocusLink/refseq.html). We partitioned exons from each gene into a unique or duplicated sequence on the basis of their map position (>90% sequence identity). We identified a total of 7777 exons as being transcribed from recently duplicated sequence, corresponding to 6.1% of all RefSeq exons (128,467). This is slightly greater than the genomic representation of segmental duplication (5.2%), which confirms that gene-poor regions have not been preferentially duplicated. In many cases, a complete complement of exons was not duplicated. These incomplete duplicated genes were often found adjacent to other duplicated cassettes that originated from elsewhere in the genome. By comparing our data with human expressed sequence tag databases, we found evidence for "chimeric" or fusion transcripts that emerged from the physical juxtaposition of incomplete segmental duplications. Although the mechanism for recent segmental duplications is not understood, the existing data suggest the process may play a role in exon shuffling associated with expanding protein diversity. A complete list of all genes with one more exons within duplicated genomic sequence is available (*8*).

To further assess whether specific kinds of genes or biological processes have been preferentially duplicated, we compared all RefSeq mRNAs on the basis of their INTERPRO protein domain classification (Table 2) (table S7) (*23*). In this analysis, we considered a gene duplicated only if all its exons were contained within a duplicated genomic region. Our analysis suggests a nonrandom distribution of segmental duplications within the proteome. Genes associated with immunity and defense (natural killer receptors, defensins, interferons, serine proteases, cytokines), membrane surface interactions (galectins, HLA, lipocalins, carcinoembryonic antigens), drug detoxification (cytochrome P450), and growth/development (somatotropins, chorionic gonadotropins, pregnancy-specific glycoproteins) were particularly enriched. It should be emphasized that our gene analysis is restricted to genomic segments that show ≥90% sequence identity. On the basis of neutral expectation of divergence, this corresponds to duplications that have emerged over the last ~40 million years of human evolution (*24*). Gene duplication followed by functional specialization has long been considered a major evolutionary force for gene innovation (*25*). Therefore, these genes embedded within recent genomic duplications may be considered excellent candidates for adaptations specific to primate evolution.

References and Notes

1. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
2. J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, E. E. Eichler, *Genome Res.* **11**, 1005 (2001).
3. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
4. S. Ohno, U. Wolf, N. Atkin, *Hereditas* **59**, 169 (1968).
5. E. E. Eichler, *Trends Genet.* **17**, 661 (2001).
6. P. Stankiewicz, J. R. Lupski, *Trends Genet.* **18**, 74 (2002).
7. E. E. Eichler, *Genome Res.* **11**, 653 (2001).
8. See supporting data on *Science* Online.
9. V. E. Cheung *et al.*, *Nature* **409**, 953 (2001).
10. The sequence and underlying test statistics for all duplicated regions of the genome are available at http://humanparalogy.cwru.edu/SDD. WGAC comparisons of the human genome assembly (UCSC, August freeze, 2001; http://genome.ucsc.edu) were done as described (*2*). The WSSD-filtered set of WGAC duplications can be interactively searched (http://humanparalogy.cwru.edu/SDD). This includes extracted sequence files, the actual alignments, the location of the alignments within the assembly, and whole-chromosomal views comparing WGAC and WSSD duplication patterns. An updated WSSD based on the analysis of 39,298 clones from April 2002, detecting an additional 36 Mb of duplicated sequence, is also available.
11. T. H. Shaikh *et al.*, *Hum. Mol. Genet.* **9**, 489 (2000).
12. J. A. Bailey *et al.*, *Am. J. Hum. Genet.* **70**, 83 (2002).
13. L. Edelmann, R. K. Pandita, B. E. Morrow, *Am. J. Hum. Genet.* **64**, 1076 (1999).
14. R. Mazzarella, D. Schlessinger, *Genome Res.* **8**, 1007 (1998).
15. J. R. Lupski, *Trends Genet.* **14**, 417 (1998).
16. S. T. Sherry *et al.*, *Nucleic Acids Res.* **29**, 308 (2001).
17. K. Chen *et al.*, *Nature Genet.* **17**, 154 (1997).
18. S. L. Christian, J. A. Fantes, S. K. Mewborn, B. Huang, D. H. Ledbetter, *Hum. Mol. Genet.* **8**, 1025 (1999).
19. D. E. Jenne *et al.*, *Am. J. Hum. Genet.* **69**, 516 (2001).
20. T. Kuroda-Kawaguchi *et al.*, *Nature Genet.* **29**, 279 (2001).
21. H. C. Mefford, B. J. Trask, *Nature Rev. Genet.* **3**, 91 (2002).
22. J. Guy *et al.*, *Hum. Mol. Genet.* **9**, 2029 (2000).
23. M. Ashburner *et al.*, *Nature Genet.* **25**, 25 (2000).
24. W. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).
25. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, Berlin, 1970).

26. We thank L. Christ, M. Eichler, and U. Neuss for technical assistance, and H. Willard, J. Nadeau, T. Hassold, D. Locke, and J. Horvath for helpful comments. Supported by NIH grants GM58815 and HG002318 and U.S. Department of Energy grant ER62862 (E.E.E.), NIH Career Development Program in Genomic Epidemiology of Cancer (CA094816) and Medical Scientist Training Grant (J.A.B.), the W. M. Keck Foundation, and the Charles B. Wang Foundation.

Supporting Online Material

www.sciencemag.org/cgi/content/full/297/5583/1003/DC1
Materials and Methods
Tables S1 to S7
Figs. S1 to S5

20 March 2002; accepted 7 June 2002

Predictive Identification of Exonic Splicing Enhancers in Human Genes

William G. Fairbrother,[1,2]* Ru-Fang Yeh,[1]* Phillip A. Sharp,[1,2] Christopher B. Burge[1]†

Specific short oligonucleotide sequences that enhance pre-mRNA splicing when present in exons, termed exonic splicing enhancers (ESEs), play important roles in constitutive and alternative splicing. A computational method, RESCUE-ESE, was developed that predicts which sequences have ESE activity by statistical analysis of exon-intron and splice site composition. When large data sets of human gene sequences were used, this method identified 10 predicted ESE motifs. Representatives of all 10 motifs were found to display enhancer activity in vivo, whereas point mutants of these sequences exhibited sharply reduced activity. The motifs identified enable prediction of the splicing phenotypes of exonic mutations in human genes.

Human genes are generally transcribed as much longer precursors, typically tens of kilobases in length, from which large introns must be precisely removed and flanking exons precisely ligated to create the mRNA that will direct protein synthesis. Sequences around the splice junctions—the 5′ and 3′ splice sites (5′ss and 3′ss)—are clearly important for splice site recognition. However, these signals appear to contain only about half of the information required for exon and intron recognition in human transcripts (*1*). The sequence or structure context in the vi-

[1]Department of Biology, [2]Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: cburge@mit.edu