

# Patterns of Segmental Duplication in the Human Genome

Liqing Zhang,\*† Henry H. S. Lu,‡ Wen-yu Chung,§ Jing Yang,\* and Wen-Hsiung Li\*

\*Department of Ecology and Evolution, University of Chicago; †Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg; ‡Institute of Statistics, National Chiao-Tung University, Hsinchu, Taiwan; and §Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park

We analyzed the completed human genome for recent segmental duplications (size  $\geq 1$  kb and sequence similarity  $\geq 90\%$ ). We found that approximately 4% of the genome is covered by duplications and that the extent of segmental duplication varies from 1% to 14% among the 24 chromosomes. Intrachromosomal duplication is more frequent than interchromosomal duplication in 15 chromosomes. The duplication frequencies in pericentromeric and subtelomeric regions are greater than the genome average by approximately threefold and fourfold. We examined factors that may affect the frequency of duplication in a region. Within individual chromosomes, the duplication frequency shows little correlation with local gene density, repeat density, recombination rate, and GC content, except chromosomes 7 and Y. For the entire genome, the duplication frequency is correlated with each of the above factors. Based on known genes and Ensembl genes, the proportion of duplications containing complete genes is 3.4% and 10.7%, respectively. The proportion of duplications containing genes is higher in intrachromosomal than in interchromosomal duplications, and duplications containing genes have a higher sequence similarity and tend to be longer than duplications containing no genes. Our simulation suggests that many duplications containing genes have been selectively maintained in the genome.

## Introduction

Segmental duplication, defined as duplication of a DNA segment longer than 1 kb, has played an important role in shaping the evolution of the human genome. Studies of individual chromosomes and different versions of the nearly completed human genome all showed that the human genome has undergone numerous segmental duplications during the past 35 Myr (Hattori et al. 2000; Lander et al. 2001; Bailey et al. 2002; Samonte and Eichler 2002; Cheung et al. 2003; Hillier et al. 2003;). Locating and characterizing these segmental duplications is of great interest because these recent genomic changes might have significantly contributed to the species divergence between human and the apes or Old World monkeys (Edelmann et al. 2001; Stankiewicz et al. 2001) and because some genomic rearrangements have been found to be the causes of several genetic diseases in humans (Lupski 1998; Stankiewicz and Lupski 2002).

To date, two groups have done genome-wide analyses of segmental duplications (Bailey et al. 2002; Cheung et al. 2003). Bailey et al. (2002) estimated the proportion of duplicated segments ( $\geq 1$  kb and  $\geq 90\%$  sequence similarity) in the entire genome to be 5.2%, and Cheung et al. (2003) estimated the proportion of duplicated segments ( $\geq 5$  kb and  $\geq 90\%$  sequence similarity) to be 3.5%. Apart from different criteria of duplication size, the discrepancy between the two estimates could also be caused by different methods used to identify duplicated regions and different genome assembly versions (Cheung et al. 2003).

A more complete assembly version of the human genome became available in April 2003 but has not yet been analyzed. In this study, we identified the segmental duplications that are  $\geq 1$  kb in length and  $\geq 90\%$  in sequence similarity in the hg15 version and found great variation in the extent of segmental duplication within and

among chromosomes. To understand the causes of the observed variation, we examined a number of factors, including regional gene density, repeat sequence density, recombination rate, and GC content. Why are there so many segmental duplications in the human genome? To address this question, we contrasted duplications containing genes with duplications containing no genes in terms of duplication frequency, size, and sequence similarity.

## Materials and Methods

The hg15 assembly of the human genome was downloaded from the UCSC Web site (<http://genome.cse.ucsc.edu/index.html>); repeats were masked using RepeatMasker before downloading. All chromosomes were divided into 500-kb segments and Blasting was performed on all-against-all segments using the default parameters. In this study, we were interested in examining duplications with size  $\geq 1$  kb and sequence similarity  $\geq 90\%$ ; we did not consider older duplications, because they are more difficult to define or detect. From the Blast results, self-hits of each DNA segment and hits with less than 90% similarity were discarded. For the remaining Blast hits, we combined hits that are less than 50 kb apart on the same chromosome into one tentative duplication block.

After this step, we took out the sequences of each block pair plus the 10-kb sequences from each side of the block. We then used the GS-aligner program (Shih and Li 2003) to align the two sequences of each block pair. The GS-aligner produces HSP (high-scoring segment pairs) and non-HSP regions. HSP regions are highly similar regions without gaps, whereas non-HSP regions have a lower similarity and may contain gaps. Two HSP regions with  $\geq 90\%$  sequence similarity are combined if the non-HSP region between them also has a sequence similarity  $\geq 90\%$ . However, non-HSP regions may have a similarity lower than 90% because of random fluctuations. To be more vigorous, we applied a binomial test to each non-HSP region, and if the sequence similarity was not significantly lower ( $P < 0.05$ ) than 90%, the two flanking HSP regions and the non-HSP region were combined into

Key words: gene density, recombination rate, repetitive elements, positive selection.

E-mail: whli@uchicago.edu.

*Mol. Biol. Evol.* 22(1):135–141. 2005

doi:10.1093/molbev/msh262

Advance Access publication September 15, 2004

one segment. The alignment end was extended by a dynamic programming for up to 5 kb outward from both ends of the alignment, while maintaining the requirement  $\geq 90\%$  sequence similarity during the extension process.

Noting that some regions showed a higher density of duplication than others, we examined these regions for possible causes of duplication. Specifically, we divided each chromosome into nonoverlapping regions of 500 kb, except for pericentromeric and subtelomeric regions (see the definition used in Bailey et al. [2001]). For each region, we calculated the duplication-enrichment index, which is defined as the ratio of the observed percentage of duplications in the region to the percentage of duplications in the entire genome in terms of sequence length. We also considered the duplication-enrichment index in terms of the number of duplications in a region for all our analyses.

We examined several factors that might affect the frequency of duplication. First, we examined the relationship between the gene density and the duplication-enrichment index of a region. We used two gene databases for this analysis: known genes and Ensembl genes. Second, we examined the relationship between density of repetitive elements and extent of duplication in the region. Third, because repetitive elements (such as microsatellites and transposable elements) tend to accumulate in low-recombination-rate regions (Bartolome, Maside, and Charlesworth 2002), and because recent segmental duplications are considered as low-copy repeats (Stankiewicz and Lupski 2002), it is interesting to compare how duplicated regions distribute with respect to local recombination rates. We used the deCODE recombination rates available at <http://genome.cse.ucsc.edu/index.html>. Fourth, we calculated the GC content of each chromosomal region and examined the correlation between duplication-enrichment index and GC content.

Finally, we divided all duplications into duplications containing complete genes and duplications containing no complete genes and compared their frequency, size, and sequence similarity. We simulated 10,000 samples under a neutral-duplication model to examine whether our observed frequencies and size distributions are expected under the neutral model. The neutral model assumes that duplications can occur anywhere on the chromosome and that the frequency and size distribution for each type of duplication is simply the result of the random distribution of genes on the chromosome. For example, the neutral expectations of the frequencies of the two types of duplications were obtained as follows.

Each simulated duplication is randomly sampled without replacement from the observed duplications. The size of the simulated duplication is the same as the sampled duplication. If the sampled duplication is intrachromosomal, we pick a site from a uniform distribution of all the sites on the chromosome. If the sampled duplication is interchromosomal, we pick one chromosome from the two chromosomes involved with equal probability and randomly pick a site on the chosen chromosome. We then determine the type of the duplication based on known genes and Ensembl genes. This procedure is continued until each observed duplication is simulated, so that a simulated sample is completed. The procedure was used

to obtain 10,000 simulated samples. The frequencies of the two types of duplications were then calculated for each sample and compared with the observed frequencies.

In a similar manner, the neutral size distributions for the two types of duplications were obtained, except that duplications were sampled with replacement from the observed duplications and every simulated sample contains the same number of each type of duplications as that in our observed data. Next, the two types of simulated duplications (i.e., duplications containing genes or no genes) were compared to see whether the difference in the observed duplications is expected under the neutral-duplication model.

## Results

### Segmental Duplications in the Human Genome

We estimated that approximately 4.0% of the human genome has been duplicated in recent times (table 1 [detailed results are available upon request]). The average sizes are 18,564 bp and 14,759 bp for intrachromosomal and interchromosomal duplications, respectively, so intrachromosomal duplications tend to be larger (Wilcoxon rank sum test:  $P \leq 2.2e-16$ ).

The distribution of sequence similarities for interchromosomal duplications is skewed towards the low end of the 90% to 100% range, different from that for intrachromosomal duplications (figure 1 [only the similarity of the best matched pair was used for sequence comparisons]). A nonparametric Wilcoxon rank sum test shows that the latter has a significantly higher sequence similarity (average similarities: 93.7% versus 94.8%;  $P \leq 2.2e-16$ ), suggesting that either there are more recent intrachromosomal duplications or rates of sequence homogenization tend to be higher in intrachromosomal duplications. We found that for intrachromosomal duplications, sequence similarity is negatively correlated with the distance between two duplicated regions ( $r = -0.12$ ,  $P \leq 2.2e-16$ ).

The proportion of duplications containing genes was decided using the databases of known genes and Ensembl genes. For known genes, the proportions of duplications containing genes are 6.2% and 1.3% in intrachromosomal and interchromosomal duplications, respectively, and for Ensembl genes, the corresponding proportions are 14.9% and 7.6%. The differences between the two types of duplication are significant (for known genes:  $\chi^2 = 183$ ,  $df = 1$ ,  $P \leq 2.2e-16$ ; for Ensembl genes:  $\chi^2 = 140$ ,  $df = 1$ ,  $P \leq 2.2e-16$ ). Note that some of the above conclusions could be misleading if there is a bias towards underrepresentation of interchromosomal duplications in the genome assembly.

### Duplication-Enrichment Indexes

The extent of duplication shows great variation among chromosomal regions. For example, for chromosome 7, there is an approximately 23-fold enrichment at one subtelomeric region, a depletion of duplication at the other subtelomeric region, and an approximately sevenfold enrichment in the pericentromeric region (fig. 2). On

**Table 1**  
**Segmental Duplications in the Human Genome**

Chr	Length (bp)	Intrachromosomal		Interchromosomal		Total	
		Length	%	Length	%	Length	%
1	245,203,898	6,431,462	2.6	3,964,057	1.6	8,678,912	3.5
2	243,315,028	6,380,301	2.6	3,757,634	1.5	8,935,221	3.7
3	199,411,731	1,646,046	0.8	1,870,056	0.9	2,671,459	1.3
4	191,610,523	2,323,764	1.2	2,547,466	1.3	3,927,792	2.0
5	180,967,295	4,066,897	2.2	2,083,920	1.2	5,208,550	2.9
6	170,740,541	2,048,892	1.2	1,123,050	0.7	2,854,222	1.7
7	158,431,299	9,629,716	6.1	3,734,503	2.4	11,722,991	7.4
8	145,908,738	1,576,863	1.1	1,694,593	1.2	2,153,612	1.5
9	134,505,819	8,451,476	6.3	4,371,262	3.2	9,403,888	7.0
10	135,480,874	6,460,047	4.8	1,919,342	1.4	7,741,228	5.7
11	134,978,784	4,223,832	3.1	2,147,666	1.6	5,382,256	4.0
12	133,464,434	1,616,743	1.2	1,134,900	0.9	2,582,114	1.9
13	114,151,656	1,451,225	1.3	1,655,399	1.5	2,700,321	2.4
14	105,311,216	282,478	0.3	849,400	0.8	1,116,676	1.1
15	100,114,055	5,520,203	5.5	3,339,498	3.3	7,091,918	7.1
16	89,995,999	7,378,691	8.2	3,456,338	3.8	8,247,312	9.2
17	81,691,216	5,505,106	6.7	1,217,149	1.5	6,432,722	7.9
18	77,753,510	230,844	0.3	1,400,896	1.8	1,627,497	2.1
19	63,790,860	1,763,189	2.8	918,571	1.4	2,531,577	4.0
20	63,644,868	772,190	1.2	1,068,246	1.7	1,369,456	2.2
21	46,976,537	431,633	0.9	1,714,574	3.6	1,734,567	3.7
22	49,476,972	2,303,175	4.7	1,633,388	3.3	3,481,523	7.0
X	152,634,166	3,579,325	2.3	4,550,908	3.0	8,047,172	5.3
Y	50,961,097	6,651,452	13.1	1,462,582	2.9	7,353,078	14.4
Total	3,070,521,116	90,725,550	3.0	53,615,398	1.7	122,996,064	4.0

average, pericentromeric and subtelomeric regions have a 2.9-fold and a 4.1-fold duplication enrichment, respectively (table 2), as compared with the 3.7-fold and 1.7-fold estimated by Bailey et al. (2001), who used an earlier assembly version (January 2001) with 21 chromosomes only.

For known genes, the correlation is positive for chromosomes 2, 5, 6, 7, 8, 10, 13, 15, 18, and Y ( $P < 0.05$ , table 3). After the Bonferroni correction, the correlation remains significant for chromosomes 5, 7, 13, and Y ( $P < 0.00033$ ). For Ensembl genes, the correlation is positive for chromosomes 1 to 8, 10, 12, 13, 15, 16, 18, 22, and Y ( $P < 0.05$  [table 3]). The correlation remains significant after the Bonferroni correction for chromosomes 1, 2, 5, 6, 8, 10, 13, 15, and Y. Ensembl genes have

more cases of significant correlation, because Ensembl has more genes than known genes. Overall, the correlation coefficient is low: for the entire genome, Kendall's  $\tau$  is 0.085 for known genes and 0.156 for Ensembl genes.

The duplication-enrichment index is positively correlated with the repeat density for chromosomes 7, 13, 16, 19, Y, and the entire genome data ( $P < 0.05$  [table 3]). After the Bonferroni correction, the correlation remains significant for chromosomes 7, Y, and the entire genome ( $P < 0.00033$ ). The duplication-enrichment index is negatively correlated with the recombination rate for chromosomes 9, 10, 11, 16, 17, 19, and the entire genome ( $P < 0.05$  [table 3]). After the Bonferroni correction, the correlation remains significant for chromosome 9 and the entire genome ( $P < 0.00033$ ).

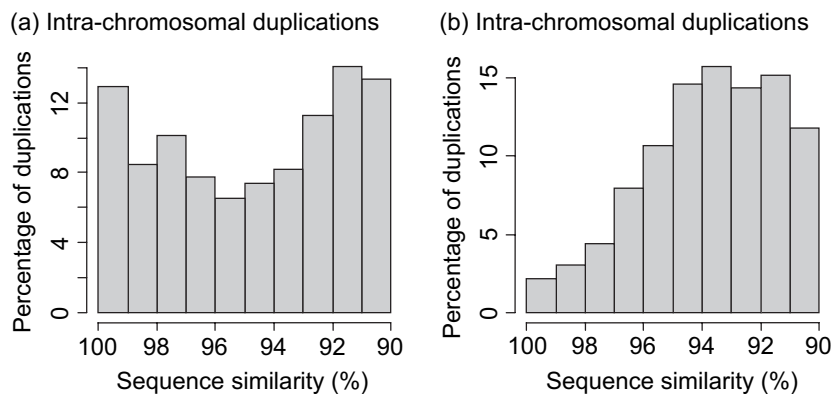


FIG. 1.—The distribution of DNA sequence similarity between segmental duplications. (a) Intrachromosomal duplications. (b) Interchromosomal duplications.

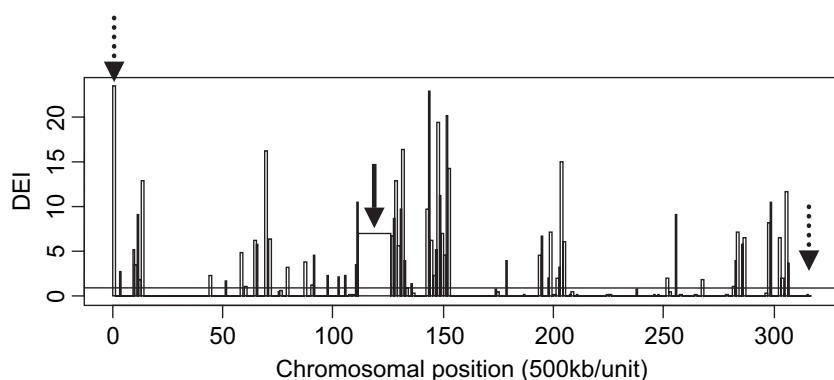


FIG. 2.—Duplication-enrichment index (DEI) along chromosome 7. DEI is the ratio of the observed percentage of duplications in the region to the percentage of duplications in the entire genome in terms of sequence length. The line represents duplication-enrichment index equal to 1; that is, no duplication enrichment compared with the genome average. The arrow points to the pericentromeric regions, and the dashed arrows point to the subtelomeric regions.

The duplication-enrichment index shows a positive correlation with GC content for chromosomes 7 and Y and a negative correlation for chromosome 10 (table 3 [ $P < 0.00033$ ]). Several points are worth mentioning. First, using the number of duplications instead of the percent coverage in the region for duplication-enrichment index does not change the results qualitatively. Second, the above significant correlations are not caused by chromosome Y, because excluding chromosome Y from the entire genome data does not affect any of the correlation analyses qualitatively (table 3). Third, excluding pericentromeric and subtelomeric regions also does not affect the results qualitatively.

**Table 2**  
Duplication-Enrichment Indexes in Pericentromeric, Subtelomeric, and Other Chromosomal Regions

Chromosome	Pericentromeric	Subtelomeric	Others
1	0.7	9.9	0.8
2	6.9	3.3	0.6
3	0.2	3.8	0.3
4	0.9	6.2	0.4
5	0.8	7.6	0.6
6	2.6	5.7	0.2
7	7.0	11.7	1.5
8	0.4	4.5	0.3
9	5.7	14.8	1.3
10	4.7	4.2	1.1
11	1.7	1.4	0.9
12	2.1	0.8	0.3
13*	2.0	0.4	0.4
14*	1.0	0.3	0.2
15*	4.4	4.2	1.4
16	4.9	4.9	1.9
17	4.8	0.0	1.6
18	3.4	2.8	0.1
19	0.0	5.2	1.0
20	2.4	1.2	0.2
21*	4.2	0.5	0.3
22*	2.8	1.5	1.3
X	0.5	2.3	1.3
Y	4.4	0.0	5.4
Total	2.9	4.1	0.8

NOTE.—Asterisks (\*) indicate only the DEI in the subtelomeric region of the q-arm is calculated for the acrocentric chromosome.

### Segmental Duplications with and Without Genes

Duplications containing genes and duplications containing no genes were compared for their frequency, size, and sequence similarity. First, the proportion of duplications containing complete genes is 3.4% for known genes and 10.7% for Ensembl genes. Both values are significantly higher than expected under the neutral-duplication model (for known genes:  $P$  value = 0.02; for Ensembl genes:  $P$  value  $\leq 2.2e-16$ . [see, e.g., figure 3]).

Second, the two types of duplications show a striking difference in size distribution (fig. 4): duplications containing genes have a much longer tail of distribution and tend to be larger (Wilcoxon rank sum test:  $P \leq 2.2e-16$ ). However, a similar pattern is also observed for the simulated duplications under the neutral-duplication model, suggesting that the observed difference could be simply the result of the distribution of genes on the chromosomes (fig. 4). Note that for a duplication to contain one or more complete genes, the duplication must be considerably long.

Third, duplications containing genes have a distribution skewed towards the high end of the 90% to 100% similarity range, which is the opposite of duplications containing no genes (fig. 5a). The difference between the two distributions is highly significant (Wilcoxon rank sum test:  $P \leq 2.2e-16$ ). The result still holds when repetitive sequences are excluded in both types of duplications. Furthermore, the proportion of duplications containing genes tends to decrease with the decrease of sequence similarity (fig. 5b).

### Discussion

#### Factors Affecting the Frequency of Segmental Duplication

Why does the frequency of segmental duplication vary greatly among chromosomes and, at a finer scale, among regions of a chromosome (table 1 and fig. 2)? In an attempt to answer this question, we examined some properties that have been documented to vary along chromosomes to see whether they might affect the duplication frequency.

**Table 3**  
**Correlation Between Duplication Enrichment Index and Gene Density, Repeat Density, Recombination Rate, and GC Content**

Chr	Known Genes		Ensembl Genes		Repeat Density		Recombination Rate		GC Content	
	$\tau^a$	<i>P</i> value	$\tau^a$	<i>P</i> value	$\tau^a$	<i>P</i> value	$\tau^a$	<i>P</i> value	$\tau^a$	<i>P</i> value
1	0.044	0.341	0.169	0.000	-0.044	0.337	-0.078	0.102	-0.016	0.723
2	0.117	0.011	0.211	0.000	-0.011	0.810	-0.084	0.072	0.020	0.663
3	0.025	0.619	0.128	0.012	0.095	0.063	-0.024	0.640	0.118	0.021
4	0.102	0.050	0.184	0.000	0.069	0.185	-0.085	0.104	0.055	0.296
5	0.194	0.000	0.294	0.000	0.050	0.349	-0.058	0.283	0.051	0.345
6	0.169	0.002	0.199	0.000	-0.027	0.633	-0.020	0.716	0.082	0.137
7	0.241	0.000	0.190	0.001	0.255	0.000	-0.054	0.355	0.210	0.000
8	0.164	0.006	0.329	0.000	-0.055	0.361	0.093	0.129	0.023	0.709
9	-0.016	0.799	0.053	0.401	0.066	0.294	-0.301	0.000	-0.008	0.898
10	0.146	0.019	0.309	0.000	-0.090	0.152	-0.154	0.014	-0.269	0.000
11	-0.048	0.443	0.053	0.398	0.085	0.176	-0.181	0.004	-0.068	0.284
12	0.113	0.073	0.173	0.006	0.005	0.937	-0.007	0.916	0.091	0.150
13	0.301	0.000	0.295	0.000	0.139	0.043	0.042	0.571	0.127	0.065
14	0.006	0.929	0.100	0.163	0.101	0.157	-0.038	0.628	0.096	0.180
15	0.183	0.012	0.350	0.000	0.075	0.309	-0.074	0.350	0.131	0.074
16	0.061	0.434	0.173	0.025	0.177	0.021	-0.250	0.002	0.136	0.077
17	-0.047	0.570	0.037	0.650	0.007	0.928	-0.280	0.001	-0.166	0.043
18	0.195	0.020	0.269	0.001	-0.010	0.906	0.045	0.598	0.017	0.839
19	-0.111	0.238	-0.128	0.173	0.190	0.042	-0.305	0.001	-0.080	0.393
20	0.170	0.071	0.106	0.264	0.055	0.563	-0.077	0.441	0.049	0.603
21	0.085	0.445	0.084	0.455	0.116	0.301	-0.140	0.283	0.051	0.646
22	0.162	0.150	0.271	0.015	0.158	0.162	0.070	0.592	0.230	0.040
X	0.054	0.359	0.064	0.278	0.112	0.057	0.025	0.677	-0.049	0.405
Y	0.475	0.000	0.562	0.000	0.646	0.000	NA	NA	0.708	0.000
Total	0.071	0.000	0.142	0.000	0.087	0.000	-0.074	0.000	0.044	0.0008
Without						0.000				
Y	0.085	0.000	0.156	0.000	0.060		NA	NA	0.044	0.0009

<sup>a</sup> Kendall's  $\tau$  is a nonparametric measure similar to the correlation coefficient.

When one chromosome is examined at a time, often some factors are more important than others for some but not all chromosomes. Combining all chromosomes, we did find that regional duplication frequency is positively correlated with regional gene density, repeat density, recombination rates, and GC content. Nevertheless, the overall pattern emerging from our genome-wide analysis is that none of the above properties has a strong effect on the

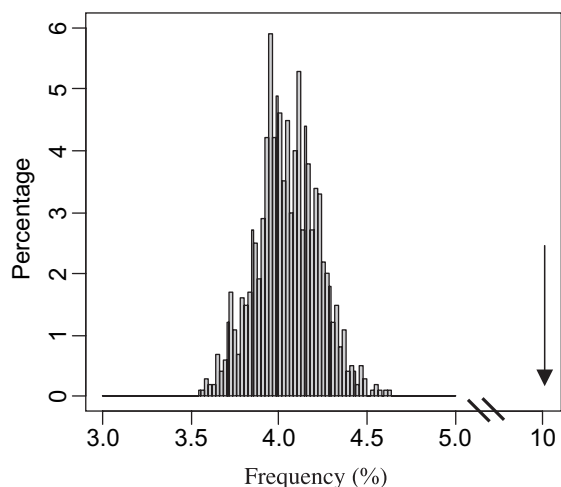


FIG. 3.—The distribution of the frequency of duplications containing complete genes under the neutral-duplication model (e.g., based on Ensembl genes). The arrow marks the observed frequency.

extent of segmental duplication, because our multiple-regression analysis shows that these factors account for only approximately 4% of the total variation in duplication frequency. Among these factors, gene density seems to be the most important in influencing duplication frequency because it alone accounts for 3.4% of the variation in duplication frequency.

Some repetitive elements, such as small ribonucleo-protein RNAs (srpRNAs), satellite DNAs, long-terminal repeats (LTRs), and, especially, *Alu* repeats have been found to be enriched in duplication borders (Bailey et al. 2001; Bailey, Liu, and Eichler. 2003; Cheung et al. 2003). However, repeat density does not seem to have a strong influence on the duplication frequency of a region (table 3), suggesting that although segmental duplication may be facilitated by repetitive elements, how often a region is involved in duplication does not significantly depend on the density of repetitive elements in the region.

Repetitive DNAs, such as microsatellites and transposable elements, tend to accumulate in low-recombination-rate regions (Bartolome, Maside, and Charlesworth 2002). This has been thought to be caused by the possibility that insertion and expansion of these repetitive elements are slightly deleterious, and selection is not efficient in removing them in low-recombination-rate regions. However, although recent segmental duplications are one type of repetitive DNAs, there is only a weak negative correlation between duplication frequency and local recombination rate (table 3). In an earlier study, Zhang and Gaut (2003) found that the frequency of

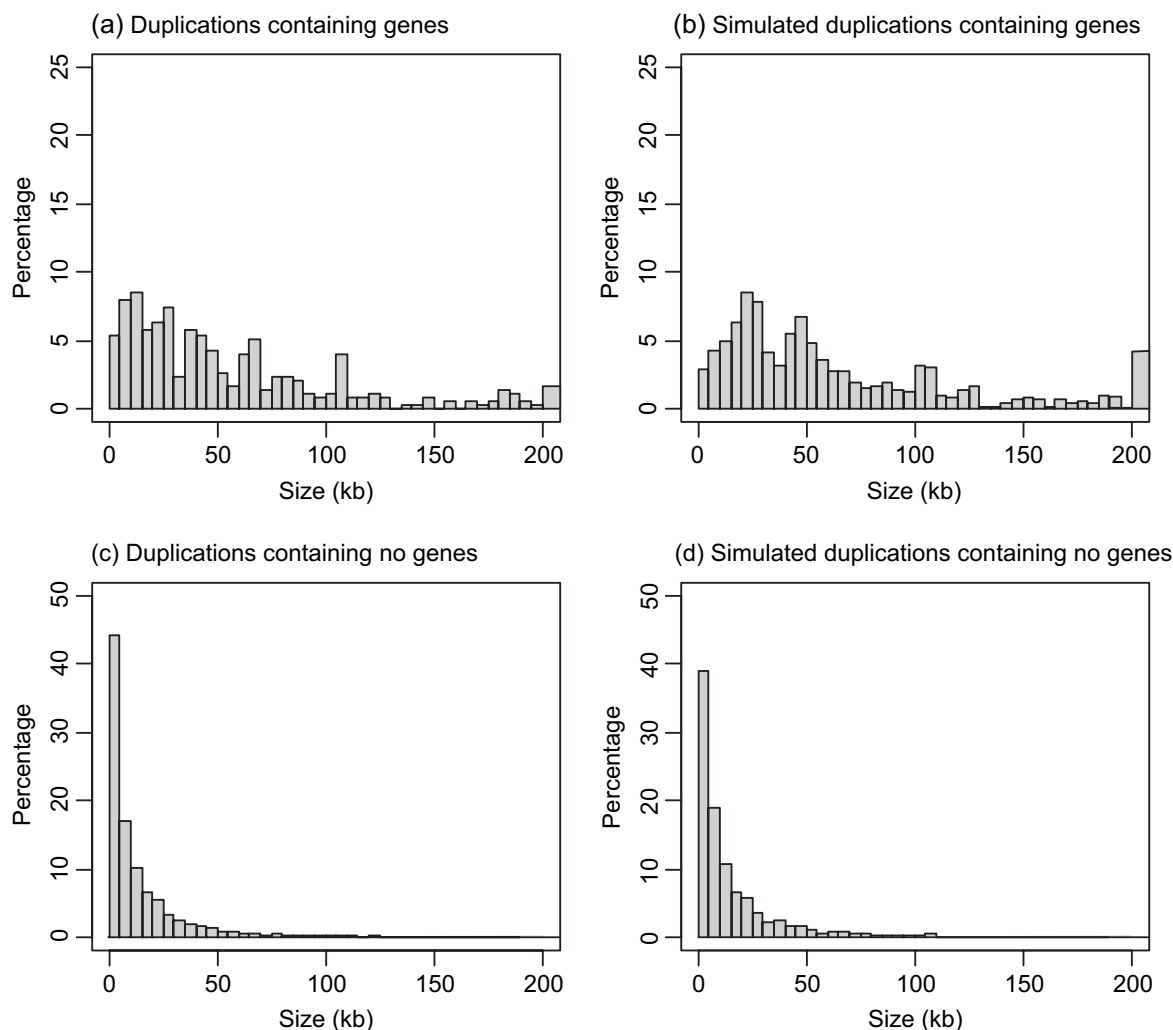


FIG. 4.—Size distributions for the observed and simulated duplications. (a and b) The observed and simulated duplications containing genes. (c and d) The observed and simulated duplications containing no genes. The last bar in the figure represents all the duplications that are longer than 200 kb.

tandemly arrayed genes is positively, rather than negatively, correlated with local recombination rate for three of the five chromosomes and has no significant correlation for the other two chromosomes in the

*Arabidopsis thaliana* genome. Taken together, it suggests that low-copy repeats may have different dynamics and distributions from high-copy repeats such as microsatellites and transposable elements.

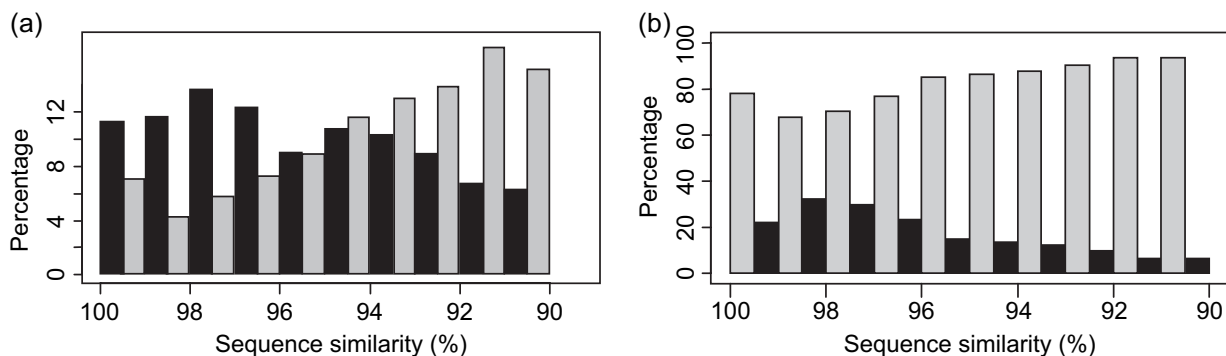


FIG. 5.—(a) Distributions for both duplications containing genes (black bars) and duplications containing no genes (gray bars). (b) All duplications: the black and gray bars represent, respectively, the proportions of duplications containing genes and no genes within each bin of sequence similarity.

## Possible Adaptive Significance of Recent Segmental Duplications

What are the forces that maintain the recent segmental duplications in the human genome? Evidence to date suggests that some segmental duplications are maintained by selection. PMCHL1 and PMCHL2, which arose from a recent segmental duplication on chromosome 5, show different expression patterns (Courseaux and Nahon 2001). The duplicated DGCR6 genes, which arose from a segmental duplication in the past 35 Myr, have been selectively maintained in the genome (Edelmann et al. 2001). The morpheus gene family, produced by recent segmental duplications on chromosome 16, shows molecular signatures of positive selection (Johnson et al. 2001).

In this study, we constructed a neutral-duplication model to examine whether the relative frequencies of duplications containing genes and duplications containing no genes are simply a result of regional variation in gene lengths and gene densities. Based on the model, the fixation of any duplication in the population does not depend on where it occurs on the chromosome; that is, whether the duplication includes genes or not has no fitness effect on the organism. Therefore, if a duplication containing genes and a duplication containing no genes have the same probability of fixation, the simulated duplications should have similar relative frequencies for the two types of duplication as the observed relative frequencies. However, we found that the observed frequency of duplications containing genes is much higher than the simulated values, suggesting that many duplications containing genes were selectively advantageous and, thus, have been maintained by selection after duplication (fig. 3).

It is puzzling that the proportion of duplications containing genes increases, whereas that containing no genes decreases, as sequence similarity increases (fig. 5b). Here, we present several possible explanations: First, the observations suggest that the rate of segmental duplication has not been constant over time. It is possible that duplications containing no genes had occurred more frequently in the past than in recent times, whereas the opposite trend is true for duplications containing genes. Second, duplications containing genes have, on average, been subject to stronger purifying selection than duplications containing no genes, so that their sequence similarity has been better maintained. Third, gene conversion might have contributed to some extent to the differences between the two distributions: if the rate of gene conversion increases with sequence similarity, duplications containing genes would have better chances of being homogenized than duplications containing no genes because sequence similarity in coding regions would tend to be better conserved than noncoding regions. Whether any of these speculations are plausible remain to be examined in the future.

## Acknowledgments

We thank Anton Nekrutenko, Wanli Min, and Shina Tan for help and suggestions. This study was supported by NIH grants. H.H.S.L. is partially supported by a grant from National Science Council, Taiwan.

## Literature Cited

- Bailey, J. A., Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297**:1003–1007.
- Bailey, J. A., G. Liu, and E. E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am. J. Hum. Genet.* **73**:823–834.
- Bailey, J. A., A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.* **11**:1005–1017.
- Bartolome, C., X. Maside, and B. Charlesworth. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol. Biol. Evol.* **19**:926–937.
- Cheung, J., X. Estivill, R. Khaja, J. R. MacDonald, K. Lau, L. C. Tsui, and S. W. Scherer. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4**:R25.
- Courseaux, A., and J.-L. Nahon. 2001. Birth of two chimeric genes in the hominidae lineage. *Science* **291**:1293–1297.
- Edelmann, L., P. Stankiewicz, E. Spiteri, R. K. Pandita, L. Shaffer, J. R. Lupski, and B. E. Morrow. 2001. Two functional copies of the DGCR6 gene are present on human chromosome 22q11 due to a duplication of an ancestral locus. *Genome Res.* **11**:208–217.
- Hattori, M., A. Fujiyama, T. D. Taylor et al. (34 co-authors). 2000. The DNA sequence of human chromosome 21. *Nature* **405**:311–319.
- Hillier, L. W., R. S. Fulton, L. A. Fulton et al. (107 co-authors). 2003. The DNA sequence of human chromosome 7. *Nature* **424**:157–164.
- Johnson, M. E., L. Viggiano, J. A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi, and E. E. Eichler. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**:514–519.
- Lander, E. S., L. M. Linton, B. Birren et al. (255 co-authors). 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.
- Lupski, J. R. 1998. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**:417–422.
- Samonte, R. V., and E. E. Eichler. 2002. Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.* **3**:65–72.
- Shih, A. C., and W. H. Li. 2003. GS-Aligner: a novel tool for aligning genomic sequences using bit-level operations. *Mol. Biol. Evol.* **20**:1299–1309.
- Stankiewicz, P., and J. R. Lupski. 2002. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**:74–82.
- Stankiewicz, P., S. S. Park, K. Inoue, and J. R. Lupski. 2001. The evolutionary chromosome translocation 4;19 in *Gorilla gorilla* is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP. *Genome Res.* **11**:1205–1210.
- Zhang, L., and B. S. Gaut. 2003. Does recombination shape the distribution and evolution of tandemly arrayed genes (TAGs) in the *Arabidopsis thaliana* genome? *Genome Res.* **13**:2533–2540.

Kenneth Wolfe, Associate Editor

Accepted September 8, 2004