



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Gene 312 (2003) 207–213

GENE
AN INTERNATIONAL JOURNAL ON
GENES AND GENOMES

www.elsevier.com/locate/gene

Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution

Zhongming Zhao^a, Yun-Xin Fu^a, David Hewett-Emmett^a, Eric Boerwinkle^{a,b,*}

^aHuman Genetics Center, 1200 Herman Pressler, Suite E447, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

^bInstitute of Molecular Medicine, University of Texas Health Science Center at Houston, Houston, TX 77030, USA

Received 5 November 2002; received in revised form 16 April 2003; accepted 29 April 2003

Received by T. Gojobori

Abstract

We investigated the single nucleotide polymorphism (SNP) density across the human genome and in different genic categories using two SNP databases: Celera's CgsSNP, which includes SNPs identified by comparing genomic sequences, and Celera's RefSNP, which includes SNPs from a variety of sources and is biased toward disease-associated genes. Based on CgsSNP, the average numbers of SNPs per 10 kb was 8.33, 8.44, and 8.09 in the human genome, in intergenic regions, and in genic regions, respectively. In genic regions, the SNP density in intronic, exonic and adjoining untranslated regions was 8.21, 5.28, and 7.51 SNPs per 10 kb, respectively. The pattern of SNP density based on RefSNP was different from that based on CgsSNP, emphasizing its utility for genotype-phenotype association studies but not for most population genetic studies. The number of SNPs per chromosome was correlated with chromosome length, but the density of SNPs estimated by CgsSNP was not significantly correlated with the GC content of the chromosome. Based on CgsSNP, the ratio of nonsense to missense mutations (0.027), the ratio of missense to silent mutations (1.15), and the ratio of non-synonymous to synonymous mutations (1.18) was less than half of that expected in a human protein coding sequence under the neutral mutation theory, reflecting a role for natural selection, especially purifying selection.

© 2003 Elsevier Science B.V. All rights reserved.

Keywords: Celera; RefSNP; CgsSNP; Nonsense mutation; Missense mutation; Silent mutation; Natural selection

1. Introduction

SNPs are valuable tools for localizing and identifying disease susceptibility genes (Risch and Merikangas, 1996), understanding the molecular mechanisms of mutation (Li et al., 1984; Zhao and Boerwinkle, 2002), and deducing the origins of modern human populations (Kaessmann et al., 1999; Zhao et al., 2000; Jorde et al., 2001). Because of their mutational history and population structure, it is believed that a subset of SNPs will capture the relevant information in the full complement of SNPs across the genome (Daly et al., 2001). Discovery of SNPs has been carried out via

surveys at the genome-level (Venter et al., 2001) and concerted efforts focusing on specific genomic regions (Nickerson et al., 1998). Based on limited protein and DNA sequence information, the earliest estimates indicated that there is about 1 SNP per 1 kb sequence, or nucleotide diversity of 10×10^{-4} in the human genome (Li and Sadler, 1991; Wang et al., 1998; Cargill et al., 1999; Halushka et al., 1999). More recently, Venter et al. (2001) provided a broad survey of sequence variation across the genome by comparing two human genome consensus sequences. However, no study has systematically evaluated the distribution and density of SNPs in the human genome as well as among chromosomes with special reference to the relationship between SNP density and the presumed function of the gene region (e.g. introns vs. exons).

Under the neutral theory of molecular evolution, the majority of DNA variation observed in a population is due to random drift of neutral or nearly neutral mutations (Kimura, 1985). Natural selection, such as purifying

Abbreviations: SNPs, single nucleotide polymorphisms; RefSNP, Celera's human SNP database; CgsSNP, SNP data identified by Celera genomic sequences; HGMD, Human Gene Mutation Database; HGBASE, Human Genic Bi-Allelic Sequences; CUTG, Codon Usage Tabulated from GenBank; kb, kilobase pairs; GC, guanine–cytosine; UTR, untranslated regions.

* Corresponding author. Tel.: +1-713-500-9816; fax: +1-713-500-0900.

E-mail address: eric.boerwinkle@uth.tmc.edu (E. Boerwinkle).

selection, may eliminate those mutations that have deleterious effects on function. This will reduce the ratio of those mutations over neutral mutations observed in the present population. In addition, it may lead to a high proportion of mutations with rare allele frequencies under rapid population expansion (Zhao et al., 2000). In protein coding sequences, non-synonymous mutations (i.e. amino acid-altering mutations, nonsense or missense mutations) are possibly deleterious while synonymous mutations (i.e. mutations without altering the amino acids, silent mutations) are neutral or nearly neutral with respect to fitness. Selection pressure on non-synonymous mutations can be detected or measured by comparing the patterns of non-synonymous and synonymous mutations in the human genome. One straight-forward measurement is to compare the ratio of nonsense mutations over missense mutations, missense mutations over silent mutations, and non-synonymous mutations over synonymous mutations using all the available data in the human genome.

In this study, we investigate the SNP density in genic and intergenic regions. In genic regions, we evaluate the SNP density in the intronic, exonic and adjoining untranslated regions. We also examine variation in SNP density among chromosomes. Finally, we compare the patterns of non-synonymous and synonymous variations, and investigate the distribution of nonsense mutations in the human genome. Two sets of data were used in these analyses: Celera's RefSNP (version 3.2) because of its comprehensive nature, and Celera's CgsSNP (SNPs identified by comparing Celera genomic sequences among two to ten chromosomes) because it represents an unbiased survey of SNPs across the genome.

2. Materials and methods

2.1. Data

SNPs from the Celera's Human RefSNP database were retrieved at <http://www.celera.com/> released on November 30, 2001 (release version 3.2). A total of 3,632,212 SNPs were recorded, including 2,462,046 SNPs identified by Celera genomic sequence assembly (CgsSNP), 1,496,047 from NCBI's dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>), 14,918 from HGMD (Human Gene Mutation Database, <http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html>), and 486,590 from HGBASE (Human Genic Bi-Allelic Sequences, <http://hgbase.interactiva.de/>). Among the HGMD SNPs, 98% were missense mutations or nonsense mutations. We analyzed 3,580,926 SNPs in RefSNP and 2,438,592 SNPs in CgsSNP with chromosomal and gene structure annotation. Gene data were retrieved from Celera's Genes database released on December 20, 2001. There were a total of 33,418 genes with annotated chromosome map locations. We downloaded the NCBI's RefSeq at <http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html>

(September 06, 2001 release) on January 9, 2002. The UCSC human working draft sequence was downloaded from <http://genome.ucsc.edu> (December 22, 2001 freeze). Frequencies of codon usage in humans were obtained from 19,312,280 codons deposited in database CUTG (Codon Usage Tabulated from GenBank) at <http://www.kazusa.or.jp/codon/> (Nakamura et al., 2000).

2.2. Data analysis

Computer programs were written in Perl to obtain and analyze the necessary information from the downloaded data. Annotation information was obtained from the first record of a SNP. Chromosome length and genome size was estimated by the positions of the first and last SNPs on a chromosome. This estimated genome size (2.93×10^9 bp) was between the sizes from the NCBI's RefSeq (2.86×10^9 bp) and from the UCSC working draft sequences (3.15×10^9 bp). SNP density was expressed as the averaged number of SNPs in a 10 kb sequence.

Gene length was calculated by the difference of the beginning position of the 5' UTR and ending position of the 3' UTR in Celera's Genes database. The total gene length was 9.33×10^8 bp, 31.87% of the whole genome. The exon length for each gene was calculated based on three times the protein length. The total exon length was 3.50×10^7 bp; 1.20% of the whole genome. The length for the UTR regions was 1.60×10^7 bp, calculated by subtracting the exon length from the transcript length.

The expected number of mutations occurring in a protein coding sequence was computed based on the universal genetic code and one of the following three assumptions: (1) nucleotide substitutions occur randomly in a random protein coding sequence (Li, 1997); (2) nucleotide substitutions occur randomly in a protein coding sequence in which codon frequencies are used based on CUTG database; (3) nucleotide substitutions occur on the mutation direction ratios observed in pseudogenes (see Li, 1997) in a protein coding sequence in which codon frequencies are used based on CUTG database. The expected number of silent (synonymous) mutations, missense mutations, and nonsense mutations was used to obtain the ratios of nonsense mutations to missense mutations, missense mutations to silent mutations, and non-synonymous mutations to synonymous mutations. We compared the observed ratios from Celera's CgsSNP to the ratios expected under the neutral theory.

3. Results

3.1. SNP distribution in Celera's CgsSNP

The analysis used 2,438,592 SNPs from Celera's CgsSNP database and associated genome annotation information, and, therefore, represents an analysis of

Table 1
Number and distribution of SNPs in CgsSNP^a

Chromosome	GC ^b (%)	Total	Intergenic (%) ^c	Genic (%)	Intron	Exonic (%)	Silent mutation	Missense mutation	Nonsense mutation	5' UTR	3' UTR
1	41.72	162573	109466 (67.33)	53107 (32.67)	50597	1427 (0.88)	678	735	14	195	888
2	40.14	205378	144535 (70.38)	60843 (29.62)	58862	1293 (0.63)	603	676	14	125	563
3	39.86	172237	120539 (69.98)	51698 (30.02)	49934	971 (0.56)	472	485	14	161	632
4	38.26	169293	126966 (75.00)	42327 (25.00)	40922	884 (0.52)	371	499	14	95	426
5	39.72	154154	114675 (74.39)	39479 (25.61)	38178	812 (0.53)	363	439	10	79	410
6	39.62	156229	112292 (71.88)	43937 (28.12)	42230	1014 (0.65)	432	572	10	114	579
7	40.65	140067	96536 (68.92)	43531 (31.08)	42002	898 (0.64)	410	477	11	89	542
8	40.07	136207	98559 (72.36)	37648 (27.64)	36465	708 (0.52)	320	374	14	96	379
9	41.38	105955	70257 (66.31)	35698 (33.69)	34542	701 (0.66)	331	359	11	70	385
10	41.56	128404	83468 (65.00)	44936 (35.00)	43491	894 (0.70)	377	508	9	104	447
11	41.65	118972	83352 (70.06)	35620 (29.94)	33902	1146 (0.96)	541	595	10	106	466
12	40.77	115669	76282 (65.95)	39387 (34.05)	37864	854 (0.74)	408	438	8	124	545
13	38.63	89268	64927 (72.73)	24341 (27.27)	23697	433 (0.49)	184	240	9	35	176
14	40.76	74023	51627 (69.74)	22396 (30.26)	21395	594 (0.80)	271	318	5	95	312
15	42.12	70787	44417 (62.75)	26370 (37.25)	25371	642 (0.91)	280	347	15	60	297
16	44.83	76638	48847 (63.74)	27791 (36.26)	26586	681 (0.89)	347	327	7	104	420
17	45.02	69272	41054 (59.26)	28218 (40.74)	26523	1028 (1.48)	485	528	15	101	566
18	39.84	68604	50538 (73.67)	18066 (26.33)	17595	285 (0.42)	123	159	3	18	168
19	48.33	56554	30603 (54.11)	25951 (45.89)	23885	1437 (2.54)	660	766	11	109	520
20	44.11	55918	35200 (62.95)	20718 (37.05)	19141	856 (1.53)	381	425	50	59	662
21	40.89	35228	26442 (75.06)	8786 (24.94)	8428	212 (0.60)	95	115	2	21	125
22	47.64	35907	20570 (57.29)	15337 (42.71)	14494	485 (1.35)	242	239	4	67	291
X	39.39	40859	32246 (78.92)	8613 (21.08)	8228	234 (0.57)	104	125	5	29	122
Y	39.11	396	336 (84.85)	60 (15.15)	53	2 (0.51)	1	1	0	0	5
Genome	40.90	2438592	1683734 (69.05)	754858 (30.95)	724385	18491 (0.76)	8479	9747	265	2056	9926

^a Results were based on Celera CgsSNP (version 3.2). Annotation information was obtained from the first record of a SNP.

^b GC content was estimated from NCBI RefSeq (September 06, 2001 release).

^c The proportion of SNPs in intergenic, genic and exonic regions was given in parentheses in the corresponding column.

sequence variation systematically identified across the human genome (Kerlavage et al., 2002). Table 1 shows the number of SNPs in each genome structural category for each chromosome. Genic regions contained 30.95% of the SNPs, while intergenic regions contained 69.05%. Within the genic regions, the proportion of SNPs occurring in intronic, exonic, 5' UTR, and 3' UTR regions was 95.96, 2.45, 0.27, and 1.31%, respectively. SNPs in the 3' UTR were observed 4.83 times of those in the 5' UTR regions. Although the exact reason for the lower number of SNPs in the 5' UTR versus the 3' UTR is unknown, it is more likely due to a longer definition of the 5' UTR itself relative to the 3' UTR rather than a large difference in SNP density. Among the SNPs in the exonic regions, 45.85% were silent mutations, 52.71% were missense mutations, and 1.43% was nonsense mutations.

Table 2 shows the expected and observed ratios of nonsense mutations to missense mutations, missense mutations to silent mutations, and non-synonymous mutations to synonymous mutations in the protein coding sequences (see Section 2.2). In the absence of natural selection, these three ratios were 0.056, 3.17, and 3.35, respectively, in a protein coding sequence having the averaged codon usage frequencies in humans (second column, Table 2). In CgsSNP, the observed ratios were

0.027, 1.15, and 1.18, respectively, less than half of that under neutral expectations.

Among the 22 autosomal chromosomes, the highest number of SNPs occurred on chromosome 2 (8.42% of the total) and the lowest number occurred on chromosome 21 (1.44% of the total). The number of SNPs was correlated with the length of the chromosomes (rank correlation coefficient = 0.98). The distribution of SNPs among gene structure categories depended on the GC content of the chromosomes. On chromosome 19 which has the highest

Table 2
Ratios for mutations in protein coding sequences

	Expected ^a	Expected ^b	Expected ^c	CgsSNP
Nonsense / missense	0.059	0.056	0.063	0.027
Missense / silent	2.93	3.17	2.37	1.15
Non-synonymous/ synonymous	3.10	3.35	2.52	1.18

^a Ratios were obtained assuming random substitutions in a random protein coding sequence.

^b Ratios were obtained assuming random substitutions in a protein coding sequence having the averaged codon usage frequencies based on CUTG database.

^c Ratios were obtained assuming substitutions occur as those observed in mammalian pseudogenes (Li 1997) in a protein coding sequence as in b.

GC content among the chromosomes, the proportion of SNPs in genic and exonic regions was 45.89 and 2.54%, respectively. On the other hand, these proportions were 25.00 and 0.52% on chromosome 4, which has the lowest GC content.

3.2. SNP density in Celera's CgsSNP

Table 3 shows the density of SNP per 10 kb of intergenic, genic and exonic sequence on each chromosome. Overall and for comparison, there was on average 8.33 SNPs/10 kb across the human genome. There were 8.44 SNPs/10 kb in the intergenic regions and 8.09 SNPs/10 kb in the genic regions. In the genic regions, there was an average of 8.21 SNPs/10 kb for introns, 5.28 SNPs/10 kb for exons, and 7.51 SNPs/10 kb for the UTR regions. On average, there were 22.59 SNPs per gene and 50.38 SNPs per intergenic regions. Nonsense mutations occurred at a density of 0.003 /10 kb in the genic regions or 0.008 per gene, comparing to that of 0.104 /10 kb in genic regions or 0.29 per gene for missense mutations. Unlike the distribution of SNPs among the gene structure categories, the overall density of SNPs for each chromosome was not significantly correlated with the GC content of the chromosome (rank correlation coefficient = 0.25).

3.3. SNP distribution and density in Celera's RefSNP

The analysis of Celera's RefSNP data and associated genome annotation information represents an analysis of all the sequence variation available from a variety of sources and collected for a variety of purposes (Kerlavage et al., 2002). In RefSNP, 3,580,926 SNPs and associated annotation information were used for further analyses (Table 4). Genic regions contained 35.91% of the SNPs, while intergenic regions contained 64.09%. Within the genic regions, the proportion of SNPs occurring in intronic, exonic, 5' UTR, and 3' UTR regions was 92.85, 4.94, 0.27, and 1.94%, respectively. Among the SNPs in the exonic regions, 32.02% were silent mutations, 61.70% were missense mutations, and 6.28% were nonsense mutations. Based on RefSNP, non-synonymous mutations were observed 2.12 times of synonymous mutations, and SNPs in the 3' UTR were observed 7.25 times of those in the 5' UTR regions.

As in CgsSNP, among the autosomal chromosomes, we observed the highest number of SNPs on chromosome 2 and the lowest number of SNPs on chromosome 21, and the number of SNPs was correlated with the length of the chromosomes (rank correlation coefficient = 0.98). There were 73,024 SNPs on the X chromosome and 4,838 SNPs on the Y chromosome. There was a higher proportion of SNPs

Table 3
Density and distribution of SNPs in CgsSNP

Chromosome	Genes	Number / genic region	Number / intergenic interval	Density (10 kb) ^a					Total
				Intergenic	Genic	Exon	Intron	UTR ^b	
1	2967	17.90	36.89	7.53	6.59	4.44	6.68	6.82	7.20
2	2420	25.14	59.73	8.55	8.21	5.14	8.34	6.70	8.45
3	1869	27.66	64.49	8.62	8.02	4.84	8.12	7.93	8.43
4	1738	24.35	73.05	9.04	8.17	5.48	8.27	7.27	8.81
5	1543	25.59	74.32	8.40	7.89	5.11	8.00	6.57	8.26
6	1716	25.60	65.44	9.17	8.61	5.68	8.73	7.92	9.00
7	1641	26.53	58.83	9.21	8.37	5.38	8.48	7.91	8.93
8	1338	28.14	73.66	9.26	9.07	5.64	9.19	7.89	9.20
9	1498	23.83	46.90	9.61	8.49	4.76	8.65	7.26	9.20
10	1608	27.95	51.91	9.63	8.93	5.83	9.05	7.77	9.37
11	1896	18.79	43.96	8.94	8.37	5.67	8.53	7.22	8.76
12	1743	22.60	43.76	8.52	8.20	4.73	8.35	7.36	8.41
13	842	28.91	77.11	9.12	8.56	5.58	8.67	6.45	8.96
14	1163	19.26	44.39	8.80	7.64	5.03	7.76	7.49	8.41
15	1141	23.11	38.93	8.69	8.10	5.30	8.24	6.32	8.46
16	1154	24.08	42.33	9.23	9.89	5.39	10.12	9.42	9.46
17	1670	16.90	24.58	8.16	8.32	5.45	8.51	7.78	8.22
18	528	34.22	95.72	8.60	8.29	4.89	8.41	6.51	8.52
19	1724	15.05	17.75	9.11	9.76	6.64	10.07	9.12	9.40
20	653	31.73	53.91	8.43	10.47	10.65	10.33	15.86	9.08
21	290	30.30	91.18	10.75	9.73	6.27	9.89	8.81	10.48
22	724	21.18	28.41	10.57	8.90	5.86	9.06	8.65	9.79
X	1358	6.34	23.75	2.96	2.51	1.74	2.55	2.31	2.85
Y	194	0.31	1.73	0.21	0.16	0.15	0.15	0.94	0.20
Genome	33418	22.59	50.38	8.44	8.09	5.28	8.21	7.51	8.33

^a Density was measured by the average number of SNPs per 10 kb sequence for gene structure categories.

^b Density in UTR regions combined the data from 5' UTR and 3' UTR.

Table 4
Distribution and density of SNPs in RefSNP

Genic categories	Number	%	Density (10 kb) ^a
Total	3580926	100	12.23
Intergenic	2294922	64.09	11.50
Genic	1286004	35.91	13.78
Intron	1194102	33.35	13.53
Exon	63519	1.77	18.15
5' UTR	3440	0.10	17.79 ^b
3' UTR	24943	0.70	

^a Density was measured by the average number of SNPs per 10 kb sequence for gene structure categories.

^b Density in UTR regions combined the data from 5' UTR and 3' UTR.

occurring in exons on the X chromosome compared to the other chromosomes. The number of nonsense and missense mutations on the X chromosome were 766 and 3,061, respectively, the highest among the chromosomes.

Table 4 shows the density of SNP per 10 kb of sequence in each genic category in RefSNP. Overall, there was an average of 12.23 SNPs/10 kb. There were 11.50 SNPs/10 kb in the intergenic regions and 13.78 SNPs/10 kb in the genic regions. In the genic regions, there was an average of 13.53 SNPs/10 kb for introns, 18.15 SNPs/10 kb for exons, and 17.79 SNPs/10 kb for the UTR regions. On average, there were 38.48 SNPs per gene. Nonsense mutations occurred at a density of 0.043 /10 kb in the genic regions or 0.12 per gene, one tenth of that observed for missense mutations (0.42 /10 kb or 1.17 per gene).

3.4. Exonic SNPs in Celera's RefSNP

There were 63,519 exonic SNPs in RefSNP compared to 18,491 in CgsSNP. Table 5 shows that about half of the exonic SNPs in RefSNP were from HGMD (14,832) and HGBASE (17,574). In RefSNP, silent mutations, missense mutations, and nonsense mutations from HGMD contributed 0.57, 29.43, and 79.79%, respectively. The corresponding proportions for each exonic category from HGBASE were 26.90, 30.26, and 6.04%, respectively. As a comparison, 23.35% of the exonic SNPs in RefSNP were from HGMD although only 0.42% of the SNPs in RefSNP were from HGMD. The proportion of exonic SNPs obtained from HGBASE was 27.67% while this proportion for the total collection of SNPs was 13.59%.

Table 5
Exonic SNPs from HGMD and HGBASE

Category	RefSNP	HGMD (% ^a)	HGBASE (%)
Silent mutations	20337	115 (0.57)	5471 (26.90)
Missense mutations	39194	11535 (29.43)	11862 (30.26)
Nonsense mutations	3988	3182 (79.79)	241 (6.04)
Exonic SNPs	63519	14832 (23.35)	17574 (27.67)
Total	3580926	14918 (0.42)	486590 (13.59)

^a The proportion of SNPs from HGMD or HGBASE in RefSNP.

We further investigated the distribution of 3,885 nonsense mutations across the human genome. Table 6 shows that these nonsense mutations were identified from 1,218 genes, an average of 3.19 nonsense mutations per gene. There were 77 genes which had at least ten nonsense mutations. The averaged number of nonsense mutations per gene was in a limited range (i.e. 1.31–4.59) among the autosomal chromosomes; however, it was much larger (8.03) on the X chromosome. On the X chromosome, there were 771 nonsense mutations found in 96 genes of which 26 had at least ten nonsense mutations.

4. Discussion

4.1. Data validation and bias

Analyses restricted to genome databases have inherent limitations. In the present case, both RefSNP and CgsSNP contain variations identified through *in silico* comparison of DNA sequences. Therefore, a proportion of the variations may not represent true polymorphisms among individuals. Marth et al. (2001) reported that approximately 80% of the SNPs found by the SNP Consortium were observed to be variable in a sample of 30 individuals sampled from three ethnic groups. In a similar study, 91% of 9,000 SNPs from CgsSNP were observed to be valid (R. Genuario, Celera, Inc., personal communication). Invalid SNPs in the present

Table 6
Distribution of nonsense mutations in the human genome

Chromosome	Nonsense	Genes (≥ 1 nonsense)	Genes (≥ 10 nonsense)	Nonsense / gene
1	254	110	4	2.31
2	163	81	2	2.01
3	156	61	3	2.56
4	92	40	2	2.30
5	182	49	1	3.71
6	118	73	1	1.62
7	244	65	3	3.75
8	108	45	3	2.40
9	153	48	4	3.19
10	102	49	0	2.08
11	322	88	9	3.66
12	123	48	3	2.56
13	112	31	4	3.61
14	53	35	0	1.51
15	76	38	1	2.00
16	171	51	3	3.35
17	266	58	3	4.59
18	21	16	0	1.31
19	167	62	2	2.69
20	118	26	1	4.54
21	21	14	0	1.50
22	81	33	1	2.45
X	771	96	26	8.03
Y	11	1	1	11.00
Genome	3885	1218	77	3.19

analysis would change the exact number and density estimates but would not change the relative comparisons among genic categories or chromosomes. In addition, the RefSNP database contains an opportunistic collection of variant sites, and is, therefore, biased in favor of coding sequence and missense and nonsense variation (Table 4). As a result, the SNPs in RefSNP should not be used for most population genetic calculations, but rather these results serve as an indication of the variation available for studies of genotype-phenotype association. Finally, the number of genes in the human genome is a topic of interest and controversy (Das et al., 2001). The number of genes used in this analysis (33,418) was based on Celera's Gene database (December 20, 2001). More genes, as some have proposed (Liang et al., 2000) and fewer genes, as others have proposed (Crollius et al., 2000), would not likely change the number of SNPs per gene or the density of SNPs among genic categories or chromosomes.

4.2. Natural selection in the human genome

The distribution of SNPs in CgsSNP can be used to investigate the influences of natural selection and other population genetic forces. For such discussions, the usual measure of nucleotide variation is π , the average number of nucleotide differences per site between two sequences (Nei and Li, 1979). Unfortunately, π cannot be readily estimated from the CgsSNP database because the number of different sequences covering each variable site is not available. Variation in CgsSNP came from the alignment of sequences from up to ten chromosomes, and most of the information may have been derived from two chromosomes from only one individual (Kerlavage et al., 2002). One may estimate nucleotide diversity or its range assuming the data were from an averaged number of chromosomes. The SNP density was 8.33 per 10 kb across the human genome based on CgsSNP. Given that the averaged number of sequences could be 2, 4, or 10, the nucleotide diversity would be close to 8.33, 4.54, or 2.94, respectively, after being normalized by Watterson's constant (Watterson, 1975). Note that the nucleotide diversity estimated by Watterson's method is usually larger than that estimated by pair-wise comparison because of the excess of rare variants often found in human population samples (e.g. Kaessmann et al., 1999; Zhao et al., 2000). If we assume that the averaged number of chromosomes used to identify the sequence differences is equal to 4 in CgsSNP, then the estimated nucleotide diversity of 4–5 per 10 kb is less than previously estimated (Li and Sadler, 1991; Cargill et al., 1999; Halushka et al., 1999; Venter et al., 2001). The lower diversity may be due to limitations of the methods used to identify SNPs in CgsSNP. The higher diversity in the autosomal regions (i.e. 8–9 per 10 kb) in Venter et al. (2001) was estimated by comparing two human consensus sequences after filtering out approximately 24% of available SNPs.

In the absence of allele frequency data, the density of

variation can be compared among structural categories. Without the effects of natural selection, the density of SNPs should be equal among categories. Therefore, the reduced density in exons (5.28 SNPs per 10 kb) relative to introns and intergenic regions (8.21 and 8.44, respectively) can be attributed to the effects of natural selection in limiting changes to the amino acid sequence of proteins. Similarly, the lowest SNP density in exons (5.29 SNPs per 10 kb) was observed in a study of the comparison of two human genome consensus sequences Celera and PFP, although the highest SNP density was observed in introns (9.21) instead of intergenic regions (7.07, Venter et al., 2001).

In CgsSNP, the observed ratio of nonsense mutations to missense mutations, missense mutations to silent mutations, and non-synonymous mutations to synonymous mutations was less than half of that under neutral expectation (Table 2). In protein coding sequences, nonsense mutations will cause the early termination of translation and lead to functional deficiencies. Missense mutations generally have more deleterious effects than silent mutations. Therefore, the observed low proportion of nonsense and missense mutations in the human genome may reflect a role for natural selection, especially purifying selection. Based on an analysis of 46 genes, Eyre-Walker and Keightley (1999) estimated that at least 38% of amino acid-altering mutations would be eliminated by natural selection. Fay et al. (2001) estimated that 80% of all amino acid substitutions were deleterious. Clearly, more information and data analyses related to this question are necessary.

4.3. Exonic SNPs

The proportion of genic SNPs was 35.91% in RefSNP and 30.95% in CgsSNP compared to 31.87% of the entire genome being genic sequences. Within genic regions, the proportion of exonic SNPs was 4.94% in RefSNP and 2.45% in CgsSNP compared to 3.75% of genic sequences as coding for proteins. In exonic regions, 6.28% of SNPs were nonsense mutations in RefSNP compared to 1.43% in CgsSNP. The higher proportion of genic and exonic SNPs observed in RefSNP is likely because of the inclusion of data from HGMD and HGBASE, which have a strong bias toward disease-associated SNPs (Table 4). There were 3,988 nonsense mutations in RefSNP, 80% (3,182) were from the HGMD database. In RefSNP, only 0.42% of the data were from HGMD, sharply contrast to 23.35% of exonic SNPs from HGMD, in which 98% of the SNPs were non-synonymous mutations. Similarly, the proportion of data from HGBASE was 13.59% for all the categories but 27.67% in the exonic regions.

4.4. Conclusions

In conclusion, these analyses provide a summary of the observed *in silico* SNP density across the human genome and in various gene structure categories. Second, the results

presented here underscore the utility of Celera's CgsSNP database for a wide range of population genetic analyses, and suggest that RefSNP may be inappropriate for such analyses. Finally, analysis of the distribution of SNPs in exonic regions provides broad evidence for the effects of natural selection influencing patterns of genome variation in modern humans.

Acknowledgements

We thank Yixi Zhong for his technical assistance. This work was supported by grant from the National Heart Lung and Blood Institute and the National Institute of General Medical Sciences. Z.Z. is supported by a training fellowship from the W.M. Keck Foundation to the Gulf Coast Consortia through the Keck Center for Computational and Structural Biology.

References

- Cargill, M., Altshuler, D., Ireland, J., et al., 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238.
- Crollius, H.R., Jaillon, O., Bernot, A., et al., 2000. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nat. Genet.* 25, 235–238.
- Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., Lander, E.S., 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229–232.
- Das, M., Burge, C.B., Park, E., Colinas, J., Pelletier, J., 2001. Assessment of the total number of human transcript units. *Genomics* 77, 71–77.
- Eyre-Walker, A., Keightley, P.D., 1999. High genome deleterious mutation rates in hominids. *Nature* 397, 344–347.
- Fay, J.C., Wyckoff, G.J., Wu, C.-I., 2001. Positive and negative selection on the human genome. *Genetics* 158, 1227–1234.
- Halushka, M.K., Fan, J.-B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., Chakravarti, A., 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22, 239–247.
- Jorde, L.B., Watkins, W.S., Bamshad, M.J., 2001. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.* 10, 2199–2207.
- Kaessmann, H., Heißig, F., von Haeseler, A., Pääbo, S., 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.* 22, 78–81.
- Kerlavage, A., Bonazzi, V., di Tommaso, M., et al., 2002. The Celera Discovery System. *Nucleic Acids Res.* 30, 129–136.
- Kimura, M., 1985. *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK.
- Li, W.-H., 1997. *Molecular Evolution*. Chapter 1., Sinauer Associates, Sunderland, MA.
- Li, W.-H., Sadler, L.A., 1991. Low nucleotide diversity in man. *Genetics* 129, 513–523.
- Li, W.-H., Wu, C.-I., Luo, C.-C., 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* 21, 58–71.
- Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., Quackenbush, J., 2000. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* 25, 239–240.
- Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, Q., Duan, S., Davenport, R., Miller, R.D., Kwok, P.-Y., 2001. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat. Genet.* 27, 371–372.
- Nakamura, Y., Gojobori, T., Ikemura, T., 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* 28, 292.
- Nei, M., Li, W.-H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* 76, 5269–5273.
- Nickerson, D.A., Taylor, S.L., Weiss, K.M., et al., 1998. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat. Genet.* 19, 233–240.
- Risch, N., Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.
- Venter, J.C., Adams, M.D., Myers, E.W., et al., 2001. The sequence of the human genome. *Science* 291, 1304–1351.
- Wang, D.G., Fan, J.-B., Siao, C.-J., et al., 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.
- Watterson, G.A., 1975. On the number of segregation sites. *Theor. Popul. Biol.* 7, 256–276.
- Zhao, Z., Boerwinkle, E., 2002. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res.* 12, 1679–1686.
- Zhao, Z., Li, J., Fu, Y.-X., et al., 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc. Natl. Acad. Sci. USA* 97, 11354–11358.