*Biophysics, Vol. 48, Suppl. 1, 2003, pp. S81–S84.*


# Single Nucleotide Polymorphism Density Distribution in the Human Genome as Seen in Public Databases: Possibility of Gene Set Comparisons but not Comparisons of Genomic Regions

**A. V. Lobashev[1], M. Yu. Skoblov[1], M. Stepanova[1,2], I. Moiseev[1], L. Dickerson[3], A. V. Baranova[1,3], A. A. Mironov[4], N. K. Yankovsky[1]**

*[1]Institute of General Genetics, Russian Academy of Sciences , Moscow, Russia; E-mail: lobashev@vigg.ru*
*[2]Moscow Institute of Physics and Technology, Moscow, Russia*
*[3]George Mason University, Mol. and Microbiol. Dept., MSN 3E1, Fairfax, VA 22030, USA*
*[4]Moscow State University, Moscow, Russia*

Received December 11, 2003; in final form, December 19, 2003

**Abstract**—Genes containing extremely large or extremely small amounts of Single Nucleotide Polymorphisms (SNPs) are of specific interest for evolutionary biology. The genome areas with extreme SNP density may possess specific functions, for example, may have specific gene or structural content. We have emulated the data gaining process corresponding to the Celera sequencing scheme and found that the excess of low SNP density regions in the human genome found by Celera is probably a natural phenomenon rather than an artifact of the scheme. We have performed an analysis of the publicly available databases on human genome sequences to catalogue the regions of extreme SNP densities in the human genome. Less than half of the extremely SNP-rich or SNP-poor regions have been found to retain their positions in consequent GenBank releases. Another set of regions with extremely low or high SNP density are inconsistent between releases in their positions, i.e., the regions of each type are lost or gained after database revision. The whole genome content analysis of the regions with extremely high or extremely low SNP content will become reasonable when the positions of the regions with abnormal SNP content stop to fluctuate because of the process of database revision. A subset of SNPs revealed by BLAST-based alignments of human ESTs was analyzed separately. The SNP density in human genome regions covered by ESTs (putatively expressed regions) is more than twice the average over the genome, as the EST-based SNPs represent 7% of all SNPs mapped to the human genome, and EST clusters cover 3% of the genome length. Half of the total length of the putatively transcribed regions is covered by non-overlapping unique ESTs (singletons). We checked the EST/SNP distribution, and the curve appeared smooth even in the interval between 2 and 1 EST per SNP. This means that even the singleton-based SNPs are reliable enough to be included into the general SNP pool for further analysis. Normalization of the SNP number to the number of cDNA libraries for each of the EST clusters opens the way to comparative studies of genes and gene families according to the intragenic SNP distributions.

*Key words:* human genome, SNP, SNP density distribution, EST, genomic DNA

## INTRODUCTION

Single Nucleotide Polymorphisms (SNPs) reflect past mutations of single base pairs, which account for most of human variation [1]. Shared variations not only predict common ancestry; they can be used to predict or explain inheritance of particular traits, such as the risk of developing a particular disease [2]. Owing to their frequency and distribution over the genome, SNPs are superior genetic and physical markers occurring approximately every 1000 bp or less. As of June 2003, more than 3.2 million SNPs are present in publicly available databases.

As a result of re-alignment assembly of the IHGSC shotgun-sequenced BAC clones [3], small variations of individual base pairs were noted throughout the sequences. These single-base changes caused by insertions, deletions, or most frequently substitutions were classified as SNPs. The common DNA sequence variations among individuals were quickly destined to gain great significance not only for biomedical and pharmaceutical scientists, but also for developmental geneticists who compare the conservation and development of important gene sequences through phylogeny [4].

Under the direction of Glaxo Wellcome, UK, a project was instituted in 1998 to identify and map enough SNPs to make a high-quality representative map of the human genome. By 1999, collaboration between research teams, private corporations, and charities became known as the SNP Consortium (TSC) [5]. The Consortium's SNP sequencing plan was to shotgun-sequence specific subsets of pooled DNA of 24 donors representing the 'ethnic diversity of humankind' [6]. Following 3–5 fold coverage, each SNP sequence was to be classified in an SNP database and mapped before being released to the public. The current SNP database and chromosome SNP maps are available through the Consortium's managing center at Cold Spring Harbor Laboratory by public access of http://snp.cshl.org.

Five donors of different origin (Hispanic, Asian, African-American, and Caucasian) were selected for genomic DNA sequencing by Celera Genomics [7]. The comparison of the sequences from the five individuals allowed one to obtain the distribution of SNPs along the human genome. An excess of low SNP density regions in the human genome was revealed by using a 100-kb sliding window. The gene content of the regions has not been reported yet and is of great interest both from evolutionary and functional points of view.

The goal of this study is to validate snip data in publicly available databases in order to describe snip density distribution in human genome. As the main source of SNP information for the study, we have used the NCBI SNP database (http://www.ncbi.nlm.nih.gov/SNP).

## RESULTS AND DISCUSSION

One of the incentives to start the study of the whole human genome SNP density distribution was an observation that it does not correspond either to the Poisson model or to the coalescent model of human evolution. The regions of low SNP density were found in excess in the human genome [7]. This may mean that the regions need to be kept genetically monomorphic on the DNA level, and the monomorphism even on the DNA sequence level may have some adaptive value.

We assumed that the cause for such misbalance could be hidden in the experimental model used to gain sequencing data by Celera [7]. Approximately 70% of the Celera sequencing data were gained on one individual with the sequencing coverage equal to 3.6 human genome equivalents. The rest (30%) of the data were gained from 4 individuals (0.3–0.5x coverage for each) [7]. One cannot exclude that some of the regions of the finished sequence were gained from several clones derived from only one of the two homologous chromosomes of one individual. In such a case the monomorphism would result from the way the data were collected and would not reflect a natural phenomenon. The original work does not address the problem [7], and we decided to check the problem by emulating the Celera scheme of gaining data.

The main assumptions of our emulation were: (1) there are 5 sources of DNA, and 71% of sequences are based on one individual's genome sequences, the other 4 are distributed almost evenly; (2) each SNP can be revealed with high probability if there are two or more genome sequences derived from different individuals; (3) the average length of a sequencing run is about 500 bp; (4) the mean density of SNPs is 1 per 500 bp, either we observe each SNP or not.

The SNP distribution that we have found from our emulation did not differ from random Poisson distribution. We have made a conclusion that the presence of SNP-poor regions is a genuine feature of the human genome but not a consequence of incorrectly gained data. The positions of the SNP-poor regions are neither listed in the article [7] nor publicly available from electronic databases. That is why the study of the content of the SNP-poor regions in the whole human genome is not possible for us from Celera data.

Because the Celera data are not publicly available, we analyzed the SNP density distribution along the human genome using only publicly available DB. To perform this study, we downloaded all human

chromosome sequences available in GenBank releases of August 2002 as well as of June (NCBI ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/). This information has been transferred onto a local computer in the form of nucleotide contigs covering more than 90% of human genome. All nucleotide contigs were downloaded with corresponding descriptions of annotated genome elements including SNPs, genes and their exons, as well as different types of repeats. To calculate the density of SNPs in floating windows of given width, we created software SNPchart that allows one to analyze distribution of SNPs in the human genome in windows equal to 10, 30, 100, 300, and 1000 kb. We have found that each human chromosome contains a number of narrow areas with SNP density 10 or more times above average (local SNP maxima), and wide areas (>100 kb in length) with no SNP at all (local SNP minima). The regions with low SNP density (25 and more times lower than average) occupy 0.62% of the whole human genome length. The regions with high SNP density (40 and more times higher than average) occupy 0.0045% of the genome.

The number of SNPs in the four consequent releases of the databases (issued between 01.01.2002 and 01.07.2003) grows by more than 30% between releases, and the growth may not be uniform along the genome. Before entering into the content analysis, we have checked the consistence of the SNP density extreme region positions in the human genome between different database releases. DB releases of 11.03.2003 (2.2 mln SNP) and 05.06.2003 (3.2 mln SNP) were used for comparison. The regions of extreme SNP density were checked for chromosome 19.

There were 7 extremely high SNP density regions in the old release and only 4 such regions in the new release. Only 2 of these were in the same positions in the human genome. Five of the regions of the old release lost the SNP-high status in the new release because the overall number of SNPs grew but not enough new SNPs were added to previously dense regions to maintain the status. Two new SNP-dense regions appeared in the new release, because a relatively large amount of SNPs was added to the regions of the old release.

There were 4 regions with extremely low SNP density in the old release and 6 such regions in the new release. Only three of the regions were in the same positions in the human genome. Three new SNP-low density regions appeared in the new release, because the overall number of SNPs grew but not enough new SNPs were added to the three regions. One of the regions of the old release lost the SNP-low status in the new release because a relatively large number of SNPs was added to the region of the old release.

Hence when the SNP number increased by half between releases, less then 50% of the SNP extreme regions stayed constant. It is possible but not guaranteed that the positions of the extreme regions that stay constant between the two releases will still stay constant in the next releases. That is why we believe that the whole genome content analysis of the regions with extremely high or extremely low SNP content will become reasonable only when the positions of the regions with abnormal SNP content stop to fluctuate because of the process of database revision.

One of the SNP sources in public databases is provided by alignments of human EST sequences versus other EST or chromosomal sequences. It is necessary to consider that EST SNP data are far from random, because their distribution depends on: (1) the presence of a transcribed region in the genome region; (2) the level of the gene expression influencing the amount of ESTs that could be sequenced from the given library; (3) how often the region is sequenced by independent researchers (this reflects the attention paid to the region or gene by the world research community).

A subset of SNPs revealed by BLAST-based alignments of human ESTs was analyzed separately. The SNP density in human genome regions covered by ESTs (putatively expressed regions) is more than two times higher than average density of SNPs in genome, as the EST-based SNPs represent 7% of all SNPs mapped to human genome, and EST clusters cover 3% of the human genome length. Half of the total length of the putatively transcribed regions is covered by non-overlapping unique ESTs (singletons). To answer the question whether single EST-based SNPs are reliable enough to count on, we have analyzed how evenly the resulting EST sequences are distributed among the DNA material sources (EST libraries). We have found that the curve (number of snips as a function of the number of ESTs covering it) looks smooth even in the interval between 2 and 1 EST per SNP. If single EST-based SNPs were not reliable, a somewhat higher rate of SNP/EST might be

expected. As this is not the case, we can say that even the singleton-based SNPs are reliable enough to be included into the general SNP pool for further analysis. Normalization of SNP number per number of cDNA libraries for each of the EST clusters opens the way to comparative studies of genes and gene families according to the intragenic SNP distributions.

## ACKNOWLEDGMENT

This work was supported by THE Russian Ministry of Science grant FCNTP "Whole Human Genome Analysis of Polymorphisms."

## REFERENCES

1. Chakravarti, A., *Nature*, 2001, vol. 409, pp. 822–823.
2. Dennis, C. and Gallagher, R., *The Human Genome*, 2001, Palgrave, Houndsmills, UK.
3. Colin, A., Semple, M., Evans, K.L., and Porteous, D.J., *Genome Biol*., 2001, vol. 2, no. 3, pp. 1–5.
4. Shastry, B.S., *J. Hum. Genet*., 2002, vol. 47, no. 11, pp. 561–566.
5. Masood, E., *Nature*, 1999, vol. 398, no. 6728, pp. 545–546.
6. Smaglik, P., *Nature*, 2000, vol. 404, p. 912.
7. Venter, C., *et al*., *Science*, 2001, vol. 291, p. 1334.