SNPs on human chromosomes 21 and 22 – analysis in terms of protein features and pseudogenes

Suganthi Balasubramanian, Paul Harrison, Hedi Hegyi, Paul Bertone, Nicholas Luscombe, Nathaniel Echols, Patrick McGarvey, ZhaoLei Zhang & Mark Gerstein[†]

[†]*Author for correspondence Department of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520-8114 USA Tel: +1 203 432 6105; Fax: +1 360 838 7861; E-mail: Mark.Gerstein@ yale.edu*

SNPs are useful for genome-wide mapping and the study of disease genes. Previous studies have focused on SNPs in specific genes or SNPs pooled from a variety of different sources. Here, a systematic approach to the analysis of SNPs in relation to various features on a genome-wide scale, with emphasis on protein features and pseudogenes, is presented. We have performed a comprehensive analysis of 39,408 SNPs on human chromosomes 21 and 22 from the SNP consortium (TSC) database, where SNPs are obtained by random sequencing using consistent and uniform methods. Our study indicates that the occurrence of SNPs is lowest in exons and higher in repeats, introns and pseudogenes. Moreover, in comparing genes and pseudogenes, we find that the SNP density is higher in pseudogenes and the ratio of nonsynonymous to synonymous changes is also much higher. These observations may be explained by the increased rate of SNP accumulation in pseudogenes, which presumably are not under selective pressure. We have also performed secondary structure prediction on all coding regions and found that there is no preferential distribution of SNPs in α-helices, β-sheets or coils. This could imply that protein structures, in general, can tolerate a wide degree of substitutions. Tables relating to our results are available from http://genecensus.org/pseudogene.

SNPs are single base variations between genomes within a species. SNPs are useful markers for diseases in haplotype-based association studies and in linkage disequilibrium analysis [1-3]. It is also believed that these small genomic level differences may be used to explain the differential drug response behavior of individuals towards a drug and can be used to tailor drugs based on an individual's genetic makeup [3].

The sequence variations in many important human genes have been extensively studied. In these studies, it was found that approximately half of the coding SNPs resulted in an amino acid change in the protein sequence [4-6]. Ng and Henikoff have predicted the effect of SNPs on protein function based on sequence homology methods [7]. Several groups have predicted the effect of SNPs on the structure of proteins in order to rationalize the effect of SNPs on protein function [8-11]. Wang and Moult showed that SNPs resulting in deleterious amino acid changes predominantly affect the stability of the protein. Sunyaev *et al.* estimate that ~ 20% of common nonsynonymous SNPs will have deleterious effects on protein structure based on the location of SNPs mapped onto 3D structures and comparative homology analyses [10]. Chasman and Adams estimated that 26–32% of nonsynonymous SNPs have effects on function [8]. With the release of the human genome sequence, broad overviews of its SNP landscape have been published [12,13]. The analyses by both Celera and TSC on different SNP data sets indicates that the distribution of SNPs is not uniform throughout the genome.

The publication of the human genome sequence [13,14] and a plethora of other genomic sequences has made it possible to perform large-scale surveys of different features relating to the whole genome [15-18]. Many different databases containing data on SNPs are now publicly available [19,20]. A huge repository of coding SNPs obtained by data mining of expressed sequence tags (ESTs) database using statistical methods is also available [21,22]. Our study pertains to a large-scale survey of SNPs in the TSC database [12,23]. We chose the TSC data set because the SNPs have been obtained by randomly sequencing human genomic DNA of 24 unrelated individuals and thus, represents an unbiased random sampling of the SNPs in the human genome. In addition, the TSC data are homogenous as the majority of SNPs are obtained by the application of uniform automated methods. It must be noted that the TSC data set consists of mostly high frequency SNPs i.e., SNPs that occur at frequencies > 10% of the populations surveyed. It is estimated that 77% of SNPs have a minor allele frequency of > 20% in at least one population surveyed [12].

We have performed a detailed analysis of SNPs in the TSC database in human chromosomes 21 and 22. Chromosomes 21 (Chr21) and 22 (Chr22) were chosen for this analysis because of their completeness and high quality of sequence and assembly available when this work began [24,25].

Methods

Release 10 of the TSC data was used for this analysis. The data was downloaded from [101]. Assembled sequences of Chr21 and Chr22 were obtained from NCBI and The Sanger Centre respectively. We have explicitly listed the exact files used because data and databases are constantly updated and we used these files throughout our analyses to make sure that all results obtained were consistent with each other. For Chr21, the file, hs_chr21all.fna, dated May 17, 2000, was retrieved from [102] and the May 19, 2000 release of Chr22 sequence was retrieved from [103]. The exact sequences that were used can also be found at [104].

The sequences corresponding to SNPs in Chr21 and 22 were extracted using the tables that contained SNPs mapped onto them using the 'golden path' assembly [105] in the TSC database. We remapped them onto the above mentioned versions of Chr21 and 22 using BLASTN [26] to maintain consistency with all our data analyses.

Exonic regions were identified from homology matches to all the SWISS-PROT proteins. This procedure was used because the number of genes, their locations and gene annotations vary depending on the gene prediction program or method of analysis used. In addition, many pseudogenes have been erroneously annotated as genes. We used the entire SWISS-PROT because a complete set of human proteins is not yet available. Significant matches to exonic regions of nonhuman proteins indicate a high likelihood of a corresponding human complement. In addition, homology to nonhuman proteins is relevant to our analysis as they may represent pseudogenes arising either from laterally transferred or extinct genes.

Homology matches to proteins were obtained by performing a six frame translational BLAST search of the genomic sequences of Chr21 and 22 against the SWISS-PROT database [27]. All the sequence coordinates were translated to absolute coordinates with respect to the chromosomal assembly.

For finding the homology matches to the SWISS-PROT database, all matches that corresponded to repeat regions were eliminated which were identified by the RepeatMasker2 program [28] and included:

- known human repeat sequences
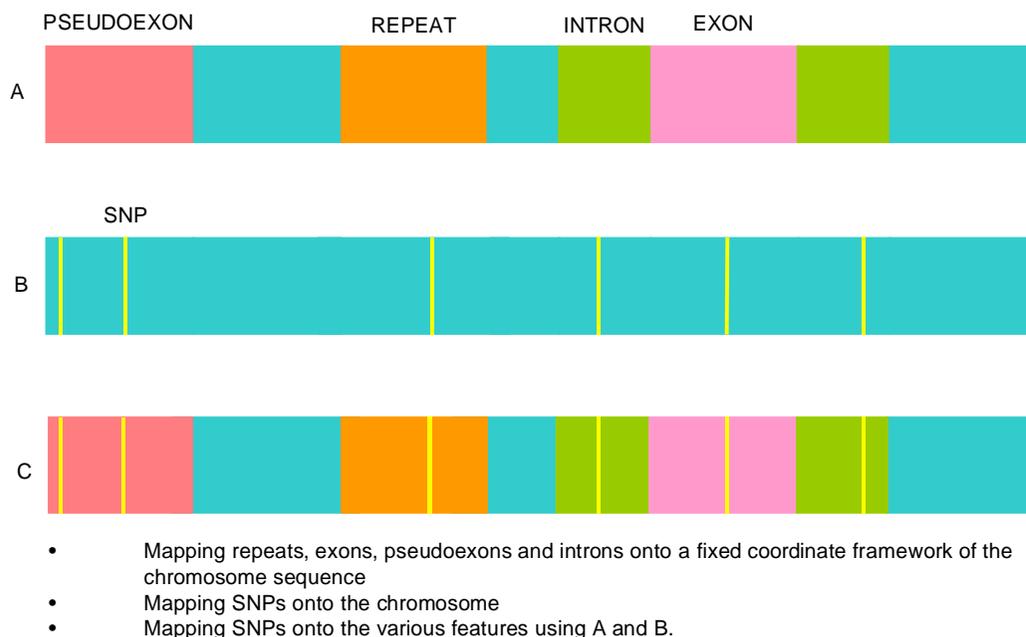- retroviral elements
- low complexity regions

Although some coding regions may contain repeats, we had to remove such matches because the genome is riddled with numerous repeats, such as Alu elements. Often, a lot of good matches are obtained to a wide variety of different sequences primarily because of the presence of low complexity repeat sequences. The only automated way to remove these spurious matches in such large-scale analyses is by elimination. Thus, this set of matches represents a conservative lower estimate of exonic regions. This set was further reduced to get a nonredundant set. First, only matches with e-values $< 10^{-4}$ were considered as significant matches. These matches were then sorted in decreasing order of significance. This set was reduced for mutual overlap by deleting matches that overlap substantially with a picked match (if two matches overlapped by more than 30 nucleotides, they were appropriately merged or removed if one was a subset of the other). Pseudogenic matches were removed from this set as described below and the remaining matches constituted the exonic matches to SWISS-PROT proteins.

Pseudogenes were obtained from the above set of matches by identifying matches that contained premature stop codons. A nonredundant set of pseudogene matches was obtained after discarding potential pseudogene sequences that overlapped artifactually with known genes [29,30].

Amino acid changes were classified as conservative and nonconservative based on Gonnet Pam250 matrix [31]. All amino acid changes within groups with a score > 0.5 were considered to be conservative. For our analysis, we considered amino acid changes within the following groups to be conservative: STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY and FYW.

Secondary structure prediction of the SNP sequences pertaining to matches to exons and pseudogenes were performed using the GORIV program [32]. The accuracy of secondary structure predictions varies depending on the method used [33,34]. Multiple sequence alignments at best have an accuracy of ~ 72% [33]. Being a single sequence prediction method, GORIV is less accurate than methods based on multiple sequence alignments. The

Figure 1. Mapping of SNPs onto the various genomic features.

- Mapping repeats, exons, pseudoexons and introns onto a fixed coordinate framework of the chromosome sequence
- Mapping SNPs onto the chromosome
- Mapping SNPs onto the various features using A and B.

Mapping of SNPs onto the various genomic features is done by mapping all features onto a fixed chromosomal coordinate system and merging the data.

version IV of GOR has a mean accuracy of 64.4% for a three state prediction (Q3) [106]. However, GORIV is a useful prediction program for large-scale analyses where multiple sequence alignments are often not possible or very time consuming. Exon predictions were obtained from GenomeScan [35]. This data were used to extract predicted intron coordinates based on the predicted exon coordinates. SNPs were modeled to fit a Poisson distribution using the following equation (where m = mean value of y and y = number of SNPs observed):

$$P(y) = (m^y e^{-m}) /y!$$

Optimized runs of the translational BLAST search of Chr22 against SWISS-PROT took ~ 8 days of CPU time (on a 600 MHz processor). The repeat masking of Chr21 and 22 took ~ 10 days each on a 600 MHz processor. Chr21 and 22 are the smallest chromosomes in the human genome. When this work began, human genome sequence data was in a continuous state of flux. Clearly, the remapping of all the features onto a dynamically changing human genome sequence data would be computationally very intensive.

Results

Distribution of SNPs in the various features of Chr21 and 22

Our analysis basically incorporates three steps:

- A. Mapping of the different genomic features onto genomic DNA sequences of Chr21 and 22.
- B. Mapping of SNPs onto the two chromosomes.
- C. Mapping SNPs onto repeats, exons, pseudogenic regions and introns by combining the results obtained from A and B.

This is illustrated in **Figure 1**. A prerequisite for this kind of systematic analysis is the availability of high quality sequence information, both in terms of completeness and assembly of the chromosomes because all the genomic features are annotated with reference to a stable coordinate system of the chromosomal sequence. Step A is computationally time intensive. Step B is a straightforward BLAST search and therefore, this approach enabled us to map the SNPs onto chromosomes quickly as the TSC database changed fairly rapidly. The mapping of SNPs directly onto the various features (for example, mapping onto exons by a translational BLAST search against SWISS-PROT proteins) would
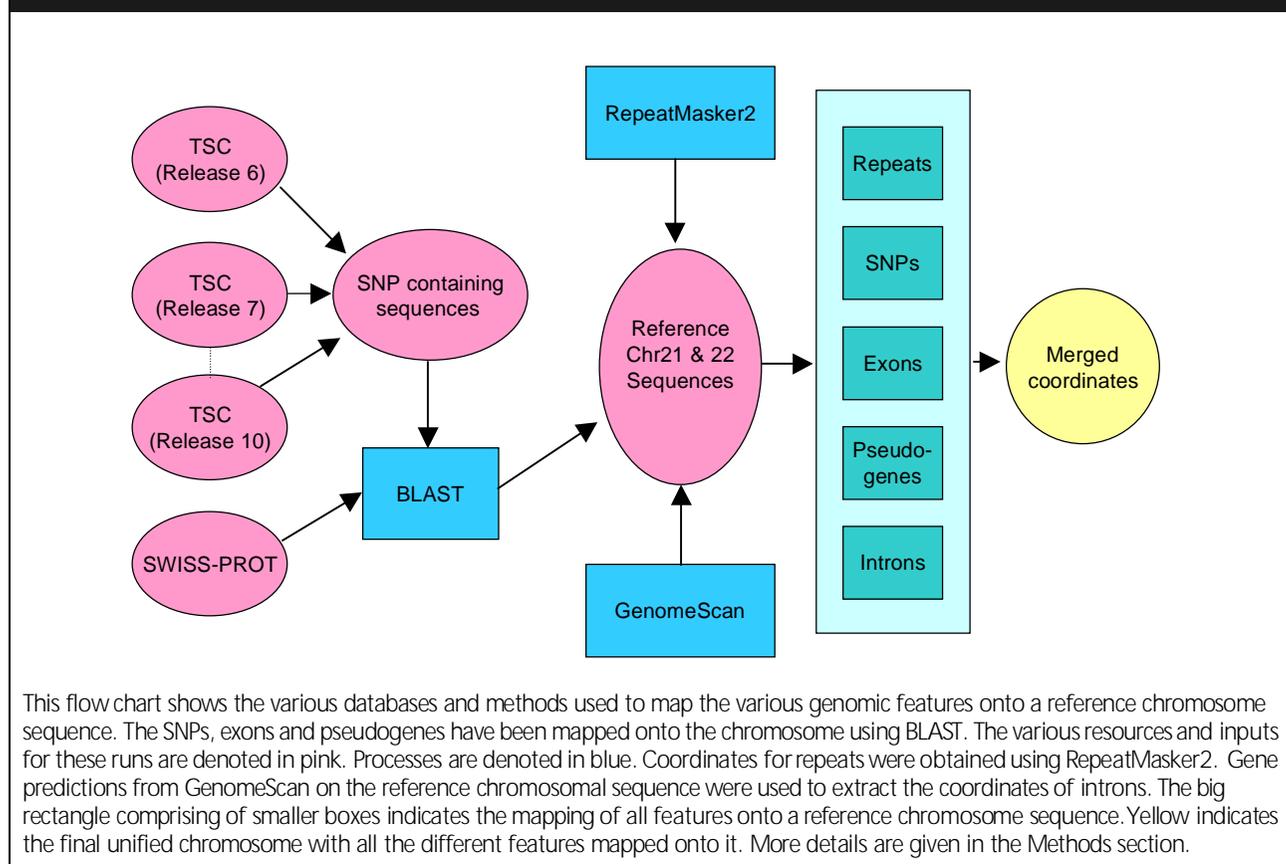
Table 1. An overview of SNP distribution.

Chromosome	Average SNP density	SNP density in repeats	SNP density in exons*	SNP density in pseudogenes*	SNP density in introns‡
21	0.56/kb (18977)	0.55/kb (8406)	0.35/kb (67)	0.40/kb (38)	0.52/kb (6371)
22	0.61/kb (20431)	0.53/kb (8365)	0.45/kb (136)	0.70/kb (94)	0.53/kb (8934)

The numbers in parentheses indicate the number of SNPs. The SNP density is the number of SNPs per kb of sequence.

*Exons and pseudogenes obtained based on homology matches to SWISS-PROT.

‡Coordinates for introns obtained from predicted gene assignments by GenomeScan.

Figure 2. Flow chart of the various databases used to map genomic features.

This flow chart shows the various databases and methods used to map the various genomic features onto a reference chromosome sequence. The SNPs, exons and pseudogenes have been mapped onto the chromosome using BLAST. The various resources and inputs for these runs are denoted in pink. Processes are denoted in blue. Coordinates for repeats were obtained using RepeatMasker2. Gene predictions from GenomeScan on the reference chromosomal sequence were used to extract the coordinates of introns. The big rectangle comprising of smaller boxes indicates the mapping of all features onto a reference chromosome sequence. Yellow indicates the final unified chromosome with all the different features mapped onto it. More details are given in the Methods section.
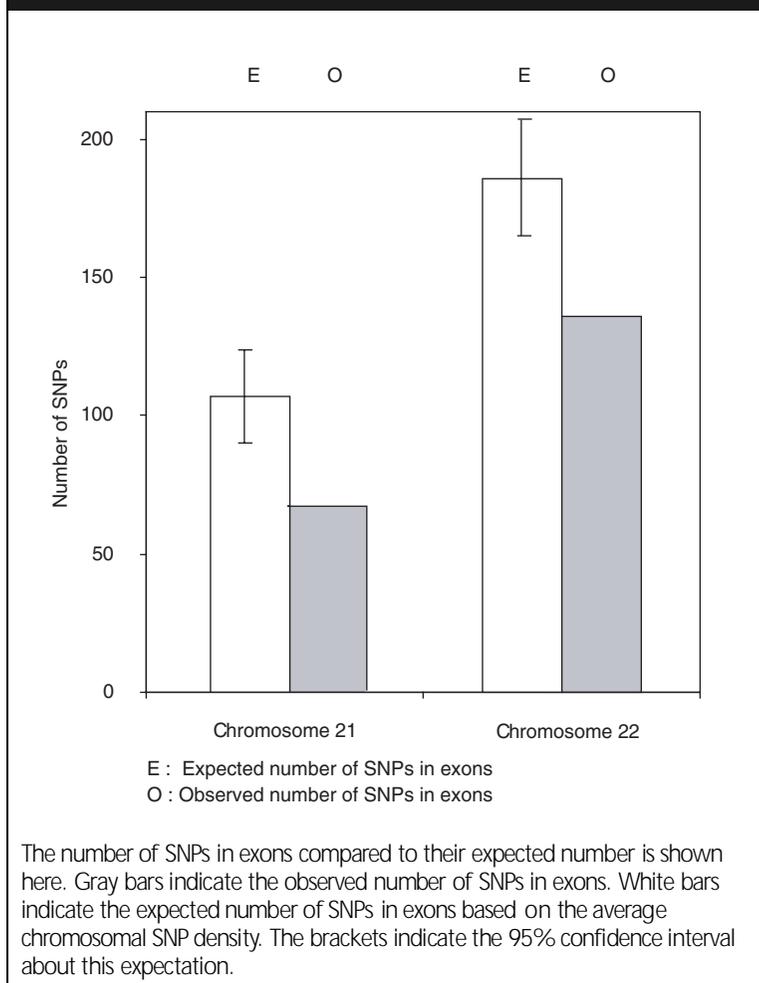
take longer computational times and would have to be performed with every new release of TSC database.

Most studies show that SNPs are found at the rate of 1–2 per kb of the human genome [36]. In the TSC data set, an average of one SNP per 2 kb is observed for both Chr21 and 22, with Chr22 having a slightly higher SNP density. An overview of the SNP distribution is shown in Table 1. While nucleotide diversity and heterozygosity measures are normally used to measure polymorphism density [4-6], it was not possible to perform such an analysis from the TSC data

because the sequence reads are not available. Instead, we define SNP density as the total number of SNPs normalized to the total number of bases in any given genomic feature. This kind of analysis is useful for looking at the TSC data, as our aim was to glean information from randomly sampled SNPs obtained by consistent and uniform methods.

As detailed in the methods section, exons and pseudogenes were identified by homology matches to proteins in SWISS-PROT, repeats were identified using RepeatMasker and intron coordinates were derived from GenomeScan pre-

Figure 3. Graph to show the number of SNP in exons compared to that expected.

The number of SNPs in exons compared to their expected number is shown here. Gray bars indicate the observed number of SNPs in exons. White bars indicate the expected number of SNPs in exons based on the average chromosomal SNP density. The brackets indicate the 95% confidence interval about this expectation.

Exonic SNPs (cSNPs)

About 0.35% and 0.66% of the SNPs in Chr21 and 22 respectively, are found in exons determined by homology matches to proteins in SWISS-PROT (**Table 1**). The occurrence of SNPs is lowest in exons compared to pseudogenes, repeats and introns. The occurrence of SNPs can be modeled by a Poisson distribution because the number of SNPs per kilobase of sequence is relatively small. Based on the average chromosomal SNP density, the number of SNPs in exons is significantly lower than would be expected based on a random distribution of SNPs in the genome as seen from the P-values. The P-value for the occurrence of 67 or fewer SNPs in exons for Chr21 is 2.0e-5 and that of finding 136 or fewer SNPs in exons in Chr22 is 5.3e-5. This is clearly illustrated in **Figure 3**.

Previous reports on SNPs indicate that the number of SNPs that result in a change of amino acid (nonsynonymous) are lower or about the same as substitutions which result in silent changes (synonymous) [4,6]. The ratio of nonsynonymous to synonymous changes due to SNPs varies between 0.3–1.0 [13] depending on the data set used. In our analysis of SNPs in exons, we see that there are more nonsynonymous changes than synonymous changes in Chr21 and 22 (**Table 2**). Of the 38 nonsynonymous SNPs in Chr21, 20 changes are conservative. Of the 85 nonsynonymous SNPs, 39 changes are conservative amino acid substitutions. For both Chr21 and 22, the ratio of nonsynonymous to synonymous changes is less than one when corrected for the frequency of synonymous and nonsynonymous sites. This corrected ratio is slightly higher in Chr22 than in Chr21 (**Table 2**). The fact that SNP density is higher for synonymous sites underscores the fact that natural selection pressure in genes operates presumably to maintain structural/functional integrity.

Four nonsynonymous SNPs in Chr21 and two nonsynonymous SNPs in Chr22 result in termination codons. Of the total six nonsynonymous SNPs that result in stop codons, three of them contain glutamine as the other variant. None of these codons are close to the 3′-end of the gene and therefore the premature truncation of proteins due to such SNPs could potentially affect their structure and/or function.

dictions. This is clearly depicted in the flow chart labeled as **Figure 2**.

The occurrence of SNPs is about the same in repeats and introns and higher than in exons. The SNP density in exons is significantly lower than that of the entire chromosome. In contrast, the SNP density is higher in pseudogenes than in exons. In Chr22, the SNP density in pseudogenes is even higher than the chromosome at large. This is surprising, as we would have expected the SNP density in pseudogenes to be similar to other intergenic regions. With the exception of SNPs in pseudogenes, the distribution of SNPs amongst the other regions of the chromosome is remarkably similar in both Chr21 and 22, as seen in **Table 1**. A detailed analysis of exonic or coding SNPs (cSNPs) and the SNPs in pseudogenes is given below.

Table 2. Comaparison of nonsynonymous and synonymous changes in chromosomes 21 and 22					
Chromosome	Category	Synonymous	Nonsynonymous	Ratio (NSyn/Syn)	Corrected ratio *(Nsyn/Syn)
21	Exons	29 *(0.59/kb)	38 *(0.27/kb)	1.31	0.46
	Pseudogenes	10 *(0.40/kb)	28 *(0.39/kb)	2.80	0.98
22	Exons	51 *(0.74/kb)	85 *(0.38/kb)	1.66	0.51
	Pseudogenes	25 *(0.72/kb)	69 *(0.69/kb)	2.76	0.96

The numbers in parentheses indicates the SNP density, the number of SNPs per kb of sequence
*The number of synonymous sites was calculated as the sum of fourfold degenerate sites and half the number of twofold degenerate sites. The number of nonsynonymous sites was calculated as the sum of nondegenerate sites and half the number of twofold degenerate sites [4].
Chr: Chromosome; Nsyn: Nonsynonymous changes; Syn: Synonymous changes.

Correlation with predicted secondary structure

Secondary structure predictions of the exon sequences containing the SNPs show that SNPs are found in all the secondary structural elements: helical, β-sheet and coil regions. While the absolute number of SNPs in coils is generally more than the number of SNPs in helices and sheets put together, the SNP densities of the exonic SNPs (normalized to the total number of residues in the corresponding secondary structural class) are not significantly different from each other (**Table 3**). In general, coil regions tend to be more variable in protein structures and we may have expected to see more SNPs in coils. We do not see a preponderance of SNPs in coils. This may indicate that protein structures may have evolved to accommodate amino acid changes [37].

Proline and glycine residues are generally not favored residues in helices and sheets. Therefore, SNPs that result in amino acid substitutions to proline or glycine could affect the structure of the protein. SNPs pertaining to prolines and glycines are discussed below:

- In Chr21, there are four nonsynonymous SNPs that result in an amino acid variation involving proline residues. In all four cases, leucine is a second variant. Of the four, three occur in coil regions and presumably do not affect the structure of the protein. However, the other proline/leucine variation is in an extended sheet region and could be deleterious to the protein fold.
- Of the three nonsynonymous SNPs involving glycines in Chr21, two occur in coils and presumably do not affect the structure of the protein. The third SNP, involving a glycine,

occupies a β-sheet and may affect the structure of the protein.
- In Chr22, five SNPs lead to variations involving proline residues. Of the five, three of them occur in helices and the other two in coils. The three SNPs in helices could affect the structure of the protein, as proline is known to disrupt helices.
- Of the twelve nonsynomous SNPs that code for glycine in Chr22, there are four potentially disruptive variants: three in helices and one in a β strand.

While prolines and glycines in coils may not affect the structure of the protein, it is quite possible that such changes may affect the function of the protein. In particular, prolines and glycines are known to be conserved critical residues important for the function of some proteins [38-40].

Pseudogenes

We observed that the SNP density is higher in pseudogenes than in exons in both Chr21 and 22 (**Table 1**). The exonic matches and pseudogenes are derived from homology matches to SWISS-PROT proteins. Therefore, we modeled the distribution of SNPs in pseudogenes as a Poisson assuming that SNP density in exons reflects the probability of SNP occurrence. We see that the number of SNPs in pseudogenes in Chr22 is far greater than that expected from the SNP density in exons. The P-value for the occurrence of 94 or more SNPs in pseudogenes for Chr22 is 4.7e-5. However, in Chr21, the increased SNP density of pseudogenes over exons is not statistically significant (P > 0.1). The occurrence of more SNPs in pseudogenes than exons in Chr22 may be due to higher sub-

Table 3. Comparision of chromosomes 21 and 22 in secondary structure				
Chromosome	Category	SNPs in helices	SNPs in β-strands	SNPs in coils
21	Exons	10 (0.63)	17 (1.26)	40 (1.16)
	Pseudogenes	5 (0.95)	11 (1.68)	22 (1.09)
22	Exons	38 (1.43)	27 (1.30)	71 (1.31)
	Pseudogenes	12 (1.40)	23 (2.37)	59 (2.21)

The number in parentheses indicates the SNP density normalized to the number of amino acid residues in the corresponding secondary structural class i.e. helices, strands and coils. The SNP density is reported as the number of SNPs per 1000 amino acid residues in the corresponding secondary structural class.

stitution rates seen in pseudogenes [41] presumably due to the lack of natural selection pressure, thus allowing organisms to accumulate SNPs in nonfunctional pseudogenes.

The ratios of nonsynonymous to synonymous amino acid changes are 2.80 and 2.76 for Chr21 and 22 respectively, both higher than the corresponding numbers for exons (**Table 2**). Also, the ratio of nonsynonymous to synonymous changes when corrected for frequency of synonymous and nonsynonymous sites is about one for both Chr21 and 22. In both cases, this corrected ratio is much higher than the corresponding ratio seen in exons. This could mean that nonsynonymous SNPs are more prevalent in pseudogenes because they may not have any deleterious functional consequences.

It is of interest to see whether pseudogenes have a preponderance of SNPs that results in an amino acid change to a termination codon. Interestingly, there are no SNPs that result in such a change in Chr21. Of the 94 SNPs in pseudogenes in Chr22, seven result in premature truncation of proteins due to a codon change to a stop codon.

Discussion and expert opinion

We have performed a systematic analysis of SNPs in human Chr21 and 22 from the TSC database. The SNP density in repeats and introns is about the same and much higher than exons. The density of SNPs in exons is significantly lower than the average chromosomal SNP density. This is not surprising because genes are under selective pressure to maintain their biological functions. Pseudogenes have a higher SNP density than exons. It is possible that SNPs have a greater propensity to accumulate in pseudogenes, as they are not under any selective pressure to maintain functional integrity. The higher SNP density in pseudogenes relative to repeats and introns for Chr22 suggests that repeat regions and other noncoding DNA regions have some selection against sequence changes. It has been shown that a large fraction of conserved elements in Chr21 comprise noncoding DNA [42]. Several noncoding DNA regions have regulatory roles and presumably have other hitherto unknown important functions that may explain the lower rate of polymorphism in other intergenic regions relative to pseudogenes [43-45].

SNPs are found in all secondary structural elements of the proteins (α-helices, β-sheet and coil regions) and do not seem to preferentially populate any particular secondary structure. This implies that proteins have evolved to maintain their structural integrity and are able to accommodate a variety of substitutions. This kind of structural plasticity has been experimentally shown in several lysozyme mutants as well as other examples [37,46-49]. Perhaps proteins maintain or adapt their 3D structures to fulfill their biological roles. While there would certainly be some SNPs that would have deleterious structural consequences, it is possible that the majority of them are fairly benign to protein structure [8,10].

The number of SNPs in exons and pseudogenes is fairly small in this study. Nevertheless, the difference in the SNP rates in exons and pseudogenes is clearly significant. This approach is being extended to the study of all the SNPs in TSC database with the primary objective of understanding the impact of these SNPs on protein structure and function and the prevalence of SNPs in pseudogenes. The rate of SNP accumulation in pseudogenes could potentially be used to estimate the age of genes and provide insights into evolution and divergence.

Highlights

- SNPs are useful as markers for tracking disease genes and may lead to a better understanding of variable drug-responses among individuals.
- The occurrence of SNPs is lowest in exons in chromosomes 21 and 22.
- The SNP density is about the same in repeats and introns.
- The SNP density in pseudogenes is higher than in exons.
- In chromosome 22, the SNP density in pseudogenes is higher than in repeats and introns.
- The ratio of nonsynonymous amino acid changes to synonymous changes is about 0.5 for exons when corrected for the frequency of synonymous and nonsynonymous sites. This indicates the strong selection against such changes in functional genes.
- The ratio of nonsynonymous amino acid changes to synonymous changes is about 1.0 for pseudogenes when corrected for the frequency of synonymous and nonsynonymous sites. This may be due to accumulation of SNPs in pseudogenes that are not under selective pressure.
- The coding SNPs are found in all secondary structural elements: helices, β-sheets and coils. This may indicate that proteins in general can accommodate mutations.

Outlook

It is amply evident from many studies that SNPs will be invaluable in medical genetics. They have been shown to be useful as markers for locating disease genes and susceptibility regions and for providing a genetic basis for differential drug-response behavior between different individuals. A large amount of SNPs have been deposited in public databases. The systematic mapping of SNPs as described in this paper will enable us to extract and integrate useful information from various heterogeneous databases (such as disease databases, clinical results etc.) in one unified format. For SNPs to be useful, it is essential to identify and pursue studies on true SNPs. This will be attainable within the next few years given the rapid advances in sequencing and genotyping technology. While it is prohibitively expensive to genotype all the candidate SNPs using current technology, the in silico analysis of SNPs will enable us to zoom in on polymorphisms that may potentially be related to the phenotype. This has already been demonstrated in studies where effects of coding SNPs on the structure and function of proteins have been predicted. Improvements in detecting regulatory regions and promoters by bioinformatic methods will enable us to predict the genotype-phenotype response of the majority of SNPs that are in noncoding regions. Identification of pseudogenes and SNPs in pseudogenes will also be useful in genetic studies because they can be used as markers to distinguish between the functional gene and its nonfunctional counterpart. In short, SNPs, their patterns and haplotype information hold great promise in unraveling new disease loci, predicting adverse side-effects of drugs based on genetic profiles and fine-tuning patient drug-dosages. In addition, analysis of SNPs among different populations and species will provide insights into human diversity and evolution.

Acknowledgements

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. Reich DE, Cargill M, Bolk S *et al.*: Linkage disequilibrium in the human genome. *Nature* 411, 199-204 (2001).
2. Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 273,1516-1517 (1996).
3. Drysdale CM, McGraw DW, Stack CB *et al.*: Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict *in vivo* responsiveness. *Proc. Natl. Acad. Sci. USA* 97, 10483-10488 (2000).
•   **Demonstrates differential drug-response to asthma medication depending on haplotype.**
4. Cargill M, Altshuler D, Ireland J *et al.*: Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231-238 (1999).
5. Halushka MK, Fan JB, Bentley K *et al.*: Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22, 239-247 (1999).
6. Stephens JC, Schneider JA, Tanguay DA *et al.*: Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293, 489-493 (2001).
7. Ng PC, Henikoff S: Predicting deleterious amino acid substitutions. *Genome Res.* 11, 863-874 (2001).
•   **Demonstrates the use of multiple sequence alignments to predict the effect of coding SNPs on protein function.**
8. Chasman D, Adams RM: Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* 307, 683-706 (2001).
•   **The effect of SNPs on protein function has been predicted using both sequence and structure information.**
9. Sunyaev S, Ramensky V, Bork P: Towards a structural basis of human non-

synonymous single nucleotide polymorphisms. *Trends Genet.* 16, 198-200 (2000).

10. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P: Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10, 591-597 (2001).

• **Prediction of SNPs that may affect the phenotype has been done based on physicochemical properties of amino acid changes, comparative homology and tertiary structure of proteins.**

11. Wang Z, Moult J: SNPs, protein structure, and disease. *Hum. Mutat.* 17, 263-270 (2001).

•• **A very elegant study that models the effect of nonsynonymous SNPs based on protein structure. The effect of SNPs is rationalized using a set of rules derived based on the results of in vitro studies of site directed mutagenesis experiments in conjunction with data of known disease-causing mutations in the context of the structure of the protein.**

12. Sachidanandam R, Weissman D, Schmidt SC *et al.*: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928-933 (2001).

• **A database containing around 1.5 million SNPs obtained by random sequencing of human DNA. The majority of the SNPs in this database have been obtained using uniform automated methods.**

13. Venter JC, Adams MD, Myers EW *et al.*: The sequence of the human genome. *Science* 291, 1304-1351 (2001).

• **The publication of the human genome sequence and discussion of SNPs identified by Celera.**

14. Lander ES, Linton LM, Birren B *et al.*: Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001).

15. Gerstein M, Hegyi H: Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.* 22, 277-304 (1998).

16. Gerstein M: How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold Des.* 3, 497-512 (1998).

17. Karlin S, Campbell AM, Mrazek J: Comparative DNA analysis across diverse genomes. *Ann. Rev. Genet.* 32, 185-225 (1998).

18. Karlin S, Mrazek J: Predicted highly expressed genes of diverse prokaryotic genomes. *J. Bacteriol.* 182, 5238-5250 (2000).

19. Brookes AJ: HGBASE-a unified human SNP database. *Trends Genet.* 17, 229 (2001).

20. Sherry ST, Ward MH, Kholodov M *et al.*: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308-311 (2001).

21. Irizarry K, Kustanovich V, Li C *et al.*: Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* 26, 233-236 (2000).

22. Buetow KH, Edmonson MN, Cassidy AB: Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* 21, 323-325 (1999).

23. Mullikin JC, Hunt SE, Cole CG *et al.*: An SNP map of human chromosome 22. *Nature* 407, 516-520 (2000).

24. Hattori M, Fujiyama A, Taylor TD *et al.*: The DNA sequence of human chromosome 21. *Nature* 405, 311-319 (2000).

25. Dunham I, Shimizu N, Roe BA *et al.*: The DNA Sequence of human chromosome 22. *Nature* 402, 489-495 (1999).

26. Altschul SF, Madden TL, Schaffer AA *et al.*: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402 (1997).

27. Bairoch A, Apweiler R: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45-48 (2000).

28. Smit AF: Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657-663 (1999).

29. Harrison PM, Echols N, Gerstein MB: Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res.* 29, 818-830 (2001).

•• **Large-scale computational analysis of *C. elegans* genome to identify pseudogenes.**

30. Harrison PM, Hegyi H, Balasubramanian S *et al.*: Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* 12, 272-280 (2002).

•• **Identification and analysis of pseudogenes in human chromosomes 21 and 22 by computational methods.**

31. Gonnet GH, Cohen MA, Benner SA: Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix. *Biochem. Biophys. Res. Commun.* 199, 489-496 (1994).

32. Garnier J, Gibrat JF, Robson B: GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* 266, 540-553 (1996).

33. Frishman D, Argos P: Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 27, 329-335 (1997).

34. King RD, Sternberg MJ: Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* 5, 2298-2310 (1996).

35. Yeh RF, Lim LP, Burge CB: Computational inference of homologous gene structures in the human genome. *Genome Res.* (2001) 11, 803-816.

36. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY: Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.* 8, 748-754 (1998).

37. Taverna DM, Goldstein RA: Why are proteins so robust to site mutations? *J. Mol. Biol.* 315, 479-484 (2002).

38. Komatsu K, Driscoll WJ, Koh YC, Strott CA: A P-loop related motif (GxxGxxK) highly conserved in sulfotransferases is required for binding the activated sulfate donor. *Biochem. Biophys. Res. Commun.* 204, 1178-1185 (1994).

39. Koonin EV: Multidomain organization of eukaryotic guanine nucleotide exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs. *Protein Sci.* 4, 1608-1617.

40. Venkatachalam KV, Fuda H, Koonin EV, Strott CA: Site-selected mutagenesis of a conserved nucleotide binding HXGH motif located in the ATP sulfurylase domain of human bifunctional 3′-phosphoadenosine 5′-phosphosulfate synthase. *J. Biol. Chem.* 274, 2601-2604 (1999).

41. Li W-H, Graur D: Fundamentals of molecular evolution. Sinauer Associates, Inc., Sunderland, Massachusets, USA (1991).

42. Frazer KA, Sheehan JB, Stokowski RP *et al.*: Evolutionarily conserved sequences on human chromosome 21. *Genome Res.* 11, 1651-1659 (2001).

43. Hamdi HK, Nishio H, Tavis J, Zielinski R, Dugaiczyk A: Alu-mediated phylogenetic novelties in gene regulation and development. *J. Mol. Biol.* 299, 931-939 (2000).

44. Santamarina-Fojo S, Peterson K, Knapper C *et al.*: Complete genomic sequence of the human ABCA1 gene: analysis of the human and mouse ATP-binding cassette A promoter. *Proc. Natl. Acad. Sci. USA* 97, 7987-7992 (2000).

45. Willoughby DA, Vilalta A, Oshima RG: An Alu element from the K18 gene confers position-independent expression in transgenic mice. *J Biol. Chem.* 275, 759-768 (2000).

46. Vetter IR, Baase WA, Heinz DW, Xiong JP, Snow S, Matthews BW: Protein structural plasticity exemplified by insertion and deletion mutants in T4 lysozyme. *Protein Sci.* 5, 2399-2415 (1996).

47. Lommer BS, Luo M: Structural plasticity in influenza virus protein NS2 (NEP). *J. Biol. Chem.* 20, 20 (2001).

48. Radaev S, Rostro B, Brooks AG, Colonna M, Sun PD: Conformational plasticity revealed by the cocrystal structure of NKG2D and its class I MHC-like ligand ULBP3. *Immunity* 15, 1039-1049 (2001).

49. Zhang XJ, Wozniak JA, Matthews BW: Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. *J. Mol. Biol.* 250, 527-552 (1995).

Websites

101. http://snp.cshl.org
     Single nucleotide polymorphisms for biomedical research.

102. ftp://ncbi.nlm.nih.gov/genbank/genomes/H_sapiens

103. ftp://ftp.sanger.ac.uk/pub/human/chr22/sequences/Chr_22/complete_sequence/Chr_22_19-05-2000.fa
     Wellcome Trust Sanger Institute anonymous FTP site.

104. http://bioinfo.mbb.yale.edu/genome/snps
     Index of /genome/snps.

105. http://genome.ucsc.edu
     Human genome project working draft.

106. http://genamics.com/expression/strucpred.htm
     Predicting the secondary structure of your protein.

107. http://genecensus.org/pseudogene
     Analyses of eukaryotic pseudogenes.