Synonymous SNP: Rare versus frequent codon can cause Phenotypic changes in the human genome

Ashu Srivastav^{1*}

¹Centre for Cellular & Molecular Biology, Habsiguda, Uppal Road, Hyderabad- 500 007, India.

Communication Address:

***Ashu Srivastav**

Research Scholar,
Dept of chemical engineering,
Institute of Chemical Technology,
NP Marg, Matunga E, Mumbai,
Maharashtra- 400019, India.
Mobile: 09967546612
Email: aanshu.deokuliar7@gmail.com

ABSTRACT

Since the initial sequencing of the human genome, many projects are underway to understand the effects of genetic variation and phenotypic changes between individuals. Single nucleotide polymorphisms (SNPs) are an increasingly important tool for genetic and biomedical research. Synonymous Single Nucleotide Polymorphisms (sSNP) is an important source of human genome variability. It does not produce altered coding sequences therefore expected not to change the function of protein in which they occur. Examination of synonymous SNPs that change a rarely used codon into a frequently used one or vice versa may help in predicting their phenotypic effect on the individual carrying the change. Detail information of Human Synonymous Single-nucleotide-polymorphism may accelerate the research of personalized medicine since it has the crucial impact in the field of non-synonymous SNP.

Keywords: Single Nucleotide Polymorphism, Codon, Synonymous, Alleles

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the most common form of genetic variation in humans comprising nearly 1/1,000th of the average human genome (Taillon-Miller et al. 1998). Traditionally, SNPs are assumed to be biallelic, *i.e.* only two of the four common nucleotides are found in that position, with lower frequency allele being present in at least 1% of the population (Sean Mooney 2005). The distribution and function of SNPs are important areas of current research. We know that synonymous codons (those encoding the same amino acid) are not equally used. This phenomenon is called codon usage bias, which exists in a wide range of biological systems from prokaryotes to eukaryotes. Synonymous codons usage patterns vary significantly among genomes, even often among genes within a single genome. Analysis of codon usage patterns can provide a basis for understanding the mechanism of biased usage of synonymous codons and selecting appropriate host expression systems to improve the expression of target genes *in vivo* and *in vitro* (Powell and Moriyama 1997; Zheng et al. 2007).

The variation of synonymous codon preference is correlated with many factors, which are different in organisms. In unicellular organisms, such as *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, and *Dictyostelium discoideum*, the codon usage is attributable to the equilibrium between natural selection and base compositional mutation bias (Sharp et al. 1993; Shields et al. 1987; Sharp and Li 1986). Reviews are available for understanding how SNPs affect protein structure, the use of SNPs in genetics studies and identifying functional variants in candidate genes. (Marnellos 2004; Rebbeck et al. 2004) Research suggests that most SNPs fall in the 95 per cent non-coding region of the genome (Hagmann 1999).

NCBI's dbSNP currently contains 4,540,241 validated human entries. Although only 1–3% of the human genome is taken up by protein-coding regions, this small subset of coding SNPs (together with SNPs in gene regulatory regions) has the highest likelihood of being functionally relevant.

The vast majority of prokaryotic and eukaryotic species have non-random codon usage. The major factor in codon choice in many unicellular and some multicellular organisms is Darwinian selection between synonymous codons; highly expressed genes using a restricted set of codons (Gouy and Gautier 1982; Ikemura 1985; Sharp and Matassi 1994) and this selection is almost certainly for optimal translational efficiency, and is most

pronounced in highly expressed genes in species whose effective population size is large (Bulmer 1991; Li 1987).

Studies on synonymous codon usage provide information about the molecular evolution of individual genes and can offer data to train methods that detect protein coding regions in uncharacterized genomic DNA (Cancilla et al. 1995b; Freirepicos et al. 1994; Gharbia et al. 1995). Knowledge of synonymous codon usage patterns can also be utilized to design DNA primers and to detect horizontal transfer events (Delorme *et al.* 1994; Groisman et al. 1992; Medigue et al. 1991). Codon-usage analyses can provide insights into the functional categories and histories of genes in a genome. More information can be gained from a codon-usage analysis than a G+C analysis, and it does not require the identification of homologous proteins from other genomes, as is the case for inferring molecular phylogenies. Recent studies have revealed that the heterogeneity in nucleotide composition and the patterns of codon usage bias within many microbial genomes are far more complex than previously imagined (Sharp and Li 1986b; Bulmer 1991). It is well established that synonymous codon usage in various organisms, particularly in unicellular ones, often reflect a balance between the genomic G+C-bias and translational selection (Sharp and Li 1986; Bulmer 1991). The strength and direction of these selection forces vary at the intra and inter-genomic level, greater is the intragenomic bias stronger is the selection for translation efficiency (Musto et al. 2003).

Selection can arise for translational elongation speed or translational accuracy. In *Drosophila*, highly conserved positions more often used preferred codons for amino acids than nonconserved positions in the same gene, suggesting the importance of accuracy (Akashi 1994). Ikemura (1981) showed that abundant proteins of *Escherichia coli* use synonymous codons corresponding to the anticodons of abundant tRNA species. Experimentally, it was found that the correlation is stronger at higher growth rates in *E. coli* in a way that optimizes translation rate (Dong et al. 1996). In the last few years, analyses on the emerging resources of several microbial genomes allowed the perception of several other trends influencing the codon usage. For instance, selection for biased codon usage has also been related with thermophily (Lynn et al. 2002) and the metabolic cost of amino acids (Akashi and Gojobori 2002).

EXPERIMENTAL SECTIONS

Tools & Software

Ensembl is a genome browser to visualize and analyze human and many other species genomes. The URL for ensembl genome browser: <https://asia.ensembl.org/index.html>. Ensembl were used to enlist the information related to sSNP. BioMart is a query-oriented data management and integration system. The URL for biomart tool: <http://asia.ensembl.org/biomart/martview/15524ff404a10c7f886a9071124cf6c7>. Using Biomart tool were used to retrieve the gene sequence. Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python programming language were used to run individual gene sequences and get the codon list. Codon bias table were used to categories the coding sequence as rare versus frequent codon.

Synonymous SNP's data collection

To collect the data related with the synonymous SNP we used the ensembl genome browser. Using the ENSEMBL genome browser with the help of Biomart all the synonymous SNPs in the human genome were obtained. Individual chromosomes were used to retrieve the sSNP by selecting following categories such as: Variant id, Chromosome name, Position on chromosome, Variant alleles, Phenotypic description, Associated gene, Phenotype name, Consequence to transcript, Consequence specific alleles, Transcript strand, Ensembl transcript id, Ensembl gene id, Variation start in CDS(in bps).

Sequence retrieval

The sequences of each genes containing the sSNP were retrieved from the ENSEMBL genome browser with the help of the Biomart tool by using ENSEMBL gene ID's. The following parameters were selected: Chromosome name, Ensembl gene id, Coding sequence to retrieve the gene.

Python programming for finding all the codon of each sequence

Individual gene sequences corresponding to the ensembl gene ID were further processed to get each codon (3 bp's). To retrieve the codons we used the programming language 'Python' to identify all the codons affected by the synonymous SNP in each of the genes. The Human codon bias table were used to select the synonymous codon. The script were written by considering each bp position of the codon. For each run time there is a need to enter the chromosome filename as a raw input. FASTA format of the sequence were used and each

individual sequence were recognized by the symbol '>'. It will pass the sequence after the run. Three file were generated as "change in first base of codon ", c1 , "second base of codon ", c2 ,and "third base of codon ", c3 after the run.

Preparing the list of rare vs frequent codon using the human codon usage table

The human codon usage table were used to enlist the frequency of each codon. Using this table the rare and frequent codons were marked. The criterion used was that a codon is marked as rare if its usage in humans is less than its expected frequency if uniform usage is assumed. For example there are 4 synonymous codons for Ala, so the expected frequency of each codon is 0.25. From the codon usage table it is clear that GCC is a frequently used codon while GCG is a rare codon. Using the above definition of rare and frequent codons the table were prepared.

Selection of rare vs frequent codon

As per the list of rare and frequent codons, codons for both the alleles were marked manually as rare or frequent codon.

RESULTS

Synonymous snp's data collection

The result shows that all the selected parameters were obtained in a tabulated form in excel sheet. The respective column shows the related information which is necessary to process further.

Microsoft Excel											
File Home Insert Formulas Data Review Layout References											
Formulas											
Name Box: A1 Formula Bar: =A1											
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA AB AC AD AE AF AG AH AI AJ AK AL AM AN AO AP AQ AR AS AT AU AV AW AX AY AZ BA BB BC BD BE BF BG BH BI BJ BK BL BM BN BO BP BQ BR BS BT BU BV BW BX BY BZ CA CB CC CD CE CF CG CH CI CJ CK CL CM CN CO CP CQ CR CS CT CU CV CW CX CY CZ DA DB DC DD DE DF DG DH DI DJ DK DL DM DN DO DP DQ DR DS DT DU DV DW DX DY DZ EA EB EC ED EE EF EG EH EI EJ EK EL EM EN EO EP EQ ER ES ET EU EV EW EX EY EZ FA FB FC FD FE FF FG FH FI FJ FK FL FM FN FO FP FQ FR FS FT FU FV FW FX FY FZ GA GB GC GD GE GF GG GH GI GJ GK GL GM GN GO GP GQ GR GS GT GU GV GW GX GY GZ HA HB HC HD HE HF HG HH HI HJ HK HL HM HN HO HP HQ HR HS HT HU HV HW HX HY HZ IA IB IC ID IE IF IG IH II IJ IK IL IM IN IO IP IQ IR IS IT IU IV IW IX IY IZ JA JB JC JD JE JF JG JH JI JJ JK JL JM JN JO JP JQ JR JS JT JU JV JW JX JY JZ KA KB KC KD KE KF KG KH KI KJ KL KM KN KO KP KQ KR KS KT KU KV KW KX KY KZ LA LB LC LD LE LF LG LH LI LJ LK LM LN LO LP LQ LR LS LT LU LV LW LX LY LZ MA MB MC MD ME MF MG MH MI MJ MK ML MN MO MP MQ MR MS MT MU MV MW MX MY MZ NA NB NC ND NE NF NG NH NI NJ NK NL NO NP NQ NR NS NT NU NV NW NX NY NZ OA OB OC OD OE OF OG OH OI OJ OK OL OM ON OO OP OQ OR OS OT OU OV OW OX OY OZ PA PB PC PD PE PF PG PH PI PJ PK PL PM PN PO PP PQ PR PS PT PU PV PW PX PY PZ QA QB QC QD QE QF QG QH QI QJ QK QL QM QN QO QQ QR QS QT QU QV QW QX QY QZ RA RB RC RD RE RF RG RH RI RJ RK RL RM RN RO RP RQ RR RS RT RU RV RW RX RY RZ SA SB SC SD SE SF SG SH SI SJ SK SL SM SN SO SP SQ SR SS ST SU SV SW SX SY SZ TA TB TC TD TE TF TG TH TI TJ TK TL TM TN TO TP TQ TR TS TT TU TV TW TX TY TZ UA UB UC UD UE UF UG UH UI UJ UK UL UM UN UO UP UQ UR US UT UU UV UW UX UY UZ VA VB VC VD VE VF VG VH VI VJ VK VL VM VN VO VP VQ VR VS VT VU VW VX VY VZ WA WB WC WD WE WF WG WH WI WJ WK WL WM WN WO WP WQ WR WS WT WU WV WW WX WY WZ XA XB XC XD XE XF XG XH XI XJ XK XL XM XN XO XP XQ XR XS XT XU XV XW XX XY XZ YA YB YC YD YE YF YG YH YI YJ YK YL YM YN YO YP YQ YR YS YT YU YV YW YX YY YZ ZA ZB ZC ZD ZE ZF ZG ZH ZI ZJ ZK ZL ZM ZN ZO ZP ZQ ZR ZS ZT ZU ZV ZW ZX ZY ZZ											
Variant ID											
Chromosome Position on Chromosome Variant ID Phenotype Associated with Variant ID											
Chromosome Position on Chromosome Variant ID Phenotype Associated with Variant ID											
rs11220	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11221	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11222	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11223	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11224	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11225	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11226	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11227	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11228	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11229	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11230	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11231	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11232	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11233	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11234	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11235	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11236	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11237	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11238	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11239	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11240	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11241	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11242	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11243	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11244	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11245	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11246	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11247	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11248	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11249	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11250	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11251	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11252	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11253	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11254	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11255	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11256	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11257	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11258	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11259	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11260	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11261	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11262	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11263	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11264	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11265	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11266	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11267	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11268	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11269	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11270	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11271	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11272	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11273	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11274	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11275	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11276	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11277	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11278	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11279	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11280	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11281	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11282	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11283	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11284	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11285	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11286	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11287	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11288	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11289	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11290	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11291	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11292	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11293	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11294	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11295	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11296	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11297	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11298	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11299	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11300	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11301	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11302	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11303	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11304	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11305	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11306	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11307	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11308	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11309	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11310	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11311	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11312	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11313	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11314	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11315	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11316	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11317	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11318	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11319	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11320	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11321	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11322	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11323	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11324	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11325	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11326	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11327	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11328	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11329	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11330	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11331	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11332	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11333	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11334	1	4892627	A	G	SNP	SPONCHP	SPONCHP	1	1	+	NS
rs11335	1	4892627	A	G	SNP						

Sequence retrieval

Biomart tool were used to get the home sapiens gene which gives the sequence in FASTA format. Each generated notepad shows a series of genes sequences in FASTA.

Table:2 a. Biomart tool selecting parameters, b. Extraction of gene sequences using Biomart.

Synonymous codon list

As a output of the compiled python program, a separate file for each codon position (1,2 or 3) affected by the SNP were generated. All three different files will be named according to their Ensembl gene id. The excel file shows the ensembl gene id, ensembl transcript id and the coding amino acid name, transcript strand as plus or minus strand. The synonymous codon of each variant were also enlisted as triplet bases.

Figure 2b is a screenshot of an Excel spreadsheet. The spreadsheet has columns for 'ENSEMBL TRANSCRIPT ID', 'ENSEMBL GENE ID', 'AMINO ACID SEQUENCE', and 'CODON'. The data is organized into rows, with each row representing a specific gene and its corresponding amino acid sequence. The 'AMINO ACID SEQUENCE' column contains the sequence of amino acids, and the 'CODON' column contains the corresponding triplet bases.

Table:3 Codon listing of each Ensembl gene ID.

Rare or frequent codon list

The prepared list of codon frequency (rare *vs.* frequent codon) were used to categorised the ensembl gene sequences. Each synonymous codons were categorised as previous codon frequency with the changed one. The changed codon frequency were enlisted as rare to frequent or vice-versa based on their allelic code. All the data were prepared and stored for further analysis associated with this codon frequency changes.

FREQUENT CODON	RARE CODON	TWIF CODON	CODAT CODON
DGC	GCT	GGG	GGA
DGG	GGA	GGG	GCG
NCA	TTC	TGG	ATG
NAG	TCC	TGG	TGG
ITC	AAT		
ICT	ATA		
NAC	CTA		
ETC	CTA		
ETG	GAT		
ETG	GAT		
ICF	GAA		
DIC	CCC		
DCA	AAT		
NAC	ACA		
NCA	GGA		
SAT	GCG		
SAC	GAT		
SAC	AAA		
NBC	TTC		
NCT	TGT		
ITC	GCT		
ITAC	GAA		
EAC	AAA		
EAG	GCT		
NAC	GGA		
NAG	TGT		
SAC			
SAG			
ITC			

➡

ENSEMBL TRANSCRIPT	ENSEMBL GENE ID	CONS. ALLELE	TRANS. STRAND	CODON	AMINO ACID	PREVIOUS %	CHANGED
ENST00000296795	* ENSG00000199918	A/G	*	* AGT	* ARGinine	* FREQUENT	* FREQUENT
ENST00000336289	ENSG00000336793	T/C		-1 GGA	GLYCINE	RARE	FREQUENT
ENST00000317889	ENSG00000317912	G/A		-1 ACC	THREONINE	RARE	FREQUENT
ENST00000366988	ENSG00000317891	T/C		1 GAT	ASPARTATE	RARE	FREQUENT
ENST00000368336	ENSG00000312676	T/C		1 CTT	LEUCINE	RARE	FREQUENT
ENST000003368336	ENSG00000312676	T/G		1 CTT	LEUCINE	RARE	FREQUENT
ENST00000343043	ENSG00000312676	T/C		1 CTT	LEUCINE	RARE	FREQUENT
ENST00000343043	ENSG00000312676	T/G		1 CTT	LEUCINE	RARE	FREQUENT
ENST00000421487	ENSG00000312676	T/C		1 CTT	LEUCINE	RARE	FREQUENT
ENST00000421487	ENSG00000312676	T/G		1 CTT	LEUCINE	RARE	FREQUENT
ENST00000351183	ENSG00000312676	T/C		1 CTT	LEUCINE	RARE	FREQUENT
ENST00000351183	ENSG00000312676	T/G		1 CTT	LEUCINE	RARE	FREQUENT
ENST00000350652	ENSG00000317193	T/C		-1 AGA	SERINE	RARE	FREQUENT
ENST00000373454	ENSG00000317193	T/C		-1 AGA	SERINE	RARE	FREQUENT
ENST00000446375	ENSG00000317193	T/C		-1 AGA	SERINE	RARE	FREQUENT
ENST00000354412	ENSG00000317193	T/C		-1 AGA	SERINE	RARE	FREQUENT
ENST00000347547	ENSG00000317193	T/C		-1 AGA	SERINE	RARE	FREQUENT
ENST00000375692	ENSG00000318610	G/A		1 CCG	PROLINE	RARE	FREQUENT
ENST00000331433	ENSG00000318610	G/A		1 CCG	PROLINE	RARE	FREQUENT
ENST00000420078	ENSG00000318610	G/A		1 CCG	PROLINE	RARE	FREQUENT
ENST00000439316	ENSG00000318610	G/A		1 CCG	PROLINE	RARE	FREQUENT
ENST00000375679	ENSG00000314908	G/A		1 GCG	ALANINE	RARE	FREQUENT
ENST00000375687	ENSG00000314908	G/A		1 GCG	ALANINE	RARE	FREQUENT
ENST00000367468	ENSG00000377356	C/G		-1 GTG	VALINE	RARE	FREQUENT
ENST00000367468	ENSG00000377356	A/G		-1 GAT	GLUTAMATE	RARE	FREQUENT

Table:4 a. Codon listing as per their frequency, b. Assigning the codon frequency changes as per the allele.

DISCUSSIONS

Synonymous SNP which don't change the amino acid sequence plays a very crucial role in personalized medicine. Synonymous variation can have effects on potential patho-physiological and pharmacogenetics. Studies shows that the presence of a rare codon, as sSNP, affects the timing of co-translational folding and even the insertion of P-gp into the membrane, resulting the alteration in the structure of substrate and inhibitor interaction sites. Even it was also observed that in some cases synonymous changes in humans have been shown to cause genetic disorders by exon skipping. While in the case of GC content at third position of codons (GC3) classically used to investigate whether natural selection influences synonymous codon usage the departure between the GC3 of a gene and the GC frequency

observed in adjacent non-coding regions which can be used as a measure of gene specific forces (e.g. selection) on synonymous mutations beyond isochoric mutational tendencies. Those genes showing the strongest departure from isochoric GC content are the best candidates for the detection of functional changes resulting from synonymous mutations. In case of protein folding it has been observed that the synonymous codon usage can affect the conformational state of the protein. It also shows that codon specific translation rate may influence the in-vivo protein folding. 3 dimensional structure of the protein were affected by the rare codon as the presence of rare codon in turns, loops and domains linkers has been observed. Further studies are required to find out more phenotypic changes caused by synonymous SNP's.

ACKNOWLEDGEMENTS

I would like to thank Mr. Ankit Agarwal for helping me in coding language, Dr. Raja G. and Dr. Manish Kesharwani for their contribution in literature searching and necessary dealing with files. The work was performed under the guidance of Dr. Shrish Tiwari, Centre for Cellular and Molecular Biology and successfully completed.

REFERENCES

1. Anna MC, Suresh V, Michael M. G., Marnellos G. (2003). High-throughput SNP analysis for genetic association studies. *Curr. Opin. Drug Discov. Devel* 6(3):317–321.
2. Clempson AM, Pollott GE, Brickell JS, Bourne NE, Munce N, Wathes DC. (2011). Evidence that leptin genotype is associated with fertility, growth, and milk production in Holstein cows. *J Dairy Sci.* 94(7):3618-28.
3. Hagmann M. (1999). A good SNP may be hard to find. *Science.* 285:21–22.
4. Javed R, Mukesh. (2010). Current research status, databases and application of single nucleotide polymorphism. *Pak J Biol Sci.* 13(13):657-63.
5. Kimchi-Sarfaty C. Oh JM, In-Wha K, Sauna Z E A. (2007). Silent Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science.* 315: 525.
6. Ma K, Cui Q, Liu G, Wu F, Xu S, Shao Y. (2011). DNA abasic site-directed formation of fluorescent silver nanoclusters for selective nucleobase recognition. *Nanotechnology.* 22(30):305502.
7. Ng MC, Hester JM, Wing MR, Li J, Xu J, Hicks PJ, Roh BH, Lu L, Divers J, Langefeld CD, Freedman BI, Palmer ND, Bowden DW. (2011). Genome-Wide Association of BMI in African Americans. *Obesity (Silver Spring).* doi: 10.1038/oby.2011.154.

8. Piva F, Giulietti M, Baldelli L, Nardi B, Bellantuono C, Armeni T, Saccucci F, Principato G. (2011). Bioinformatic analyses to select phenotype affecting polymorphisms in HTR2C gene. *Hum Psychopharmacol*. doi: 10.1002/hup.1214.
9. Powell JR, Moriyama EN. (1997). Evolution of codon usage bias in *Drosophila*. *Proc Natl Acad Sci USA* 94(15):7784–7790.
10. Rebbeck TR, Spitz M, Wu X. (2004). Assessing the function of genetic variants in candidate gene association studies. *Nat. Rev. Genet.* 5(8): 589–597.
11. Sean Mooney (2005). Bioinformatics approaches and resources for singlenucleotide polymorphism functional analysis. *Briefings in bioinformatics*. 6:44–56.
12. Sharp PM, Li WH. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 24(1–2): 28–38.
13. Sharp PM, Stenico M, Peden JF et al. (1993). Codon usage: mutational bias, translational selection, or both? *Biochem Soc Trans* 21(4):835–841.
14. Shields DC, Sharp PM. (1987). Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucleic Acids Res* 15(19):8023–8040. doi:10.1093/nar/15.19.8023
15. Taillon-Miller P, Gu Z, Li Q. (1998). Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. *Genome Res* 8(7):748–754.
16. Zheng Y, Zhao WM, Wang H et al. (2007). Codon usage bias in *Chlamydia trachomatis* and the effect of codon modification in the MOMP gene on immune responses to vaccination. *Biochem Cell Biol* 85(2):218–226.