Genomics & Informatics Vol. 5(4) 188-193, December 2007

FESD II: A Revised Functional Element SNP Database of Human Ethnicities

Hyun Ju Kim[1], Il Hyun Kim[1], Ki Hoon Shin[1], Young-Kyu Park[1], Hyojin Kang[2] and Young Joo Kim[1]*

[1]Korea Research Institute of Bioscience and Biotechnology, #111 Gwahagno, Yuseong-gu, Daejeon, 305-806, Korea, [2]Department of Bio and Brain, KAIST, Daejeon 305-701, Korea

Abstract

The Functional Element SNPs Database (FESD) categorizes functional elements in human genic regions and provides a set of single nucleotide polymorphisms (SNPs) located within each area. Users may select a set of SNPs in specific functional elements with haplotype information and obtain flanking sequences for genotyping. Our previous version of FESD has been improved in several ways. We regenerated all the data in FESD II from recently updated source data such as HapMap, UCSC GoldenPath, dbSNP, OMIM, and TRANSFAC®. Users can obtain information about tagSNPs and simulate LD blocks for each gene from four ethnicities in the HapMap project on the fly. FESD II employs a Java/JSP web interface for better platform portability and higher speed than PHP in the previous version. As a result, FESD II provides its users with more powerful information about functional element SNPs of human ethnicities.

Availability: FESD II is freely available from *http://sysbio.kribb.re.kr:8080/fesd/.*

Contact: yjkim8@kribb.re.kr

Keywords: SNP, tagSNP, haplotype, functional elements, LD block, hapMap

Introduction

Extensive studies are currently being conducted to find associations of disease susceptibility with a particular form of genetic variation, namely, single nucleotide polymorphisms (SNPs), as well as with other types of common genetic variations, including microsatellites and copy number variants (Stranger, Forrest *et al*. 2007). Even though SNPs are less polymorphic and thus less informative than microsatellite markers, SNPs are abundant (more than 12 million in humans), more stably inherited, and exist in functioning regions such as exons. Thus, investigation of the genome for SNPs may be an effective way to elucidate the causes of complex phenotypes and diseases in humans.

The database of SNPs (dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/), a repository for single-base nucleotide substitutions and short deletion and insertion polymorphisms (Sherry, Ward *et al*. 2001), contains over 12 million human SNPs and an additional 22 million from a variety of other organisms, with 14 million of those added over the past year (Wheeler, Barrett *et al*. 2007).

As the number of SNPs deposited in dbSNP has been increasing exponentially, many researchers who are interested in complex genetic diseases are focusing on candidate gene approaches using gene-based functional element SNPs and haplotypes, which are collections of SNPs located throughout the functional regions of candidate genes (Ring and Kroetz 2002; Jung, Park *et al*. 2004). Linkage disequilibrium (LD), which refers to the nonrandom association of alleles at different loci in haplotypes, plays a central role in genome-wide association studies for identifying genetic variation with different phenotypes (Zhang, Qin *et al*. 2004). A tagSNP is a representative SNP in a region of the genome with high LD. Researchers can use tagSNPs to discover genes responsible for certain disorders. Most genome-wide association studies currently underway will not be well powered for rare causal SNPs. For such studies, we can impute unobserved HapMap SNPs with tag SNPs (Marchini, Howie *et al*. 2007), reducing the time and labor spent choosing tagSNPs instead of all the variations in a sequence.

Most SNP genotyping methods currently in use begin with a PCR amplification step and need PCR primers (Nothnagel, Furst *et al*. 2002). In high-throughput SNP genotyping machines, the primers often include not only PCR primers but also extension primers (Wise, Paris *et al*. 2003). For the design of PCR primers, researchers need to prepare flanking sequences of their target SNPs. Although the dbSNP database provides flanking sequences for each refSNP, the length of flanking sequences is variable and is often not long enough to design primers. There may be many different methods to obtain the flanking sequence, such as using the UCSC genome browser or by

*Corresponding author: E-mail yjkim8@kribb.re.kr
Tel +82-42-879-8118, Fax +82-42-879-8119
Accepted 3 December 2007

NCBI blast search. However, these methods require multiple steps and do not support multiple queries (Kent, Sugnet *et al*. 2002; Karolchik, Hinrichs *et al*. 2004).

We developed and reported the Functional Element SNPs Database (FESD), which categorizes functional elements in human genic regions and provides a set of SNPs located within each area (Kang, Choi *et al*. 2005). In the FESD, human genic regions were divided into 10 different functional elements: promoter regions, CpG islands, 5'-untranslated regions (UTRs), translation start sites, splice sites, coding exons, introns, translation stop sites, polyadenylation signals (PASes), and 3'-UTRs. Subsequently, all known SNPs were assigned to each functional element at its respective position. With the FESD web interface, users can select a set of SNPs in specific functional elements and obtain flanking sequences for further genotyping experiments.

In this study, we have revised and updated the previous version of FESD by incorporating the International HapMap Project data with 270 samples. Users can obtain not only updated functional element SNPs but also haplotype information with tagSNPs. FESD II will serve as a much more versatile tool that aids users in finding mutations that contribute to common and polygenic diseases.

Updated Functional Element SNPs Database II (FESD II)

FESD categorizes functional elements in human genic regions and provides a set of SNPs located within each functional element as well as flanking sequences required for genotyping experiments (Kang, Choi *et al*. 2005). Through the FESD web interface, users can identify the functional position of SNPs and select a set of SNPs for their haplotype-based study.

FESD II was improved in many aspects. First, we regenerated all data in FESD II from recently updated source data such as HapMap, UCSC GoldenPath, dbSNP, OMIM, and TRANSFAC®. The reference sequences of the human genome build 36.2 (April 2007) were downloaded from the NCBI database, and functional element sequences were extracted from sequence boundary information derived from UCSC database hg 18 version (Karolchik, Baertsch *et al*. 2003). Among the 16 tracks within Genes and Gene Prediction Tracks in the UCSC database, the RefSeq Gene track was produced from mRNA sequence data (Maglott, Ostell *et al*. 2005). We used the 25,313 refSeq IDs in the refGene track. In total, 11,811,594 SNPs from the dbSNP database build 127 were positioned with the flat file downloaded from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/snp/human/). In particular, the matrix data of transcription factor estimation were updated from TRANSFAC® professional ver. 7.4.1 to ver. 11.1.

Second, the dbSNP data were included, as were the International HapMap Project phase I and phase II data with 270 samples (Nigerian Yoruba African, Han Chinese, Northern and Western European, and Japanese populations), allowing users to gain information about LD blocks and tagSNPs from each population. The LD block of each gene is simulated at on time in FESD II, helping users obtain haplotype information in order to select a set of SNPs in specific functional elements and to have tagSNPs with certain coverages. The LD blocks were generated by the Haploview program (Barrett, Fry *et al*. 2005). Moreover, the tagSNPs of each gene from these four populations were also provided. We adopted algorithms from the Tagger program (de Bakker, Yelensky *et al*. 2005) to select and evaluate tagSNPs.

Third, we transformed the computer program into Java/JSP, which supports better platform portability and higher speeds than did PHP in the previous version.

The human genome has two different regions—genic and intergenic. The genic region can be divided into 10 different functional elements; promoter regions, CpG islands, 5'-untranslated regions (UTRs), translation start sites, splice sites, coding exons, introns, translation stop sites, polyadenylation signal sites (PASes), and 3'UTRs. Promoter regions were estimated in the 2-kb region upstream of the start codon using a transcription factor matrix, and CpG islands found in the promoter regions were reported. Detailed information on the 10 functional regions can be gleaned from our previous paper (Kang, Choi *et al*. 2005); the 10 functional regions are depicted in the Fig. 1.

SNPs whose locations are known were assigned to the corresponding functional element. To estimate the promoter region, we used the Match program (TRANSFAC Professional 11.1), which is a weight matrix-based tool for searching putative transcription factor binding sites in DNA sequences (Kel, Gossling *et al*. 2003). The SNPflank database consists of the complete genome sequence from RefSeq genome contigs and the positional information of refSNPs from the dbSNP database (Benson, Karsch-Mizrachi *et al*. 2004). Both were downloaded from the NCBI ftp site, parsed with

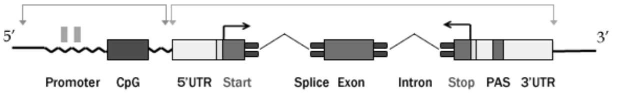
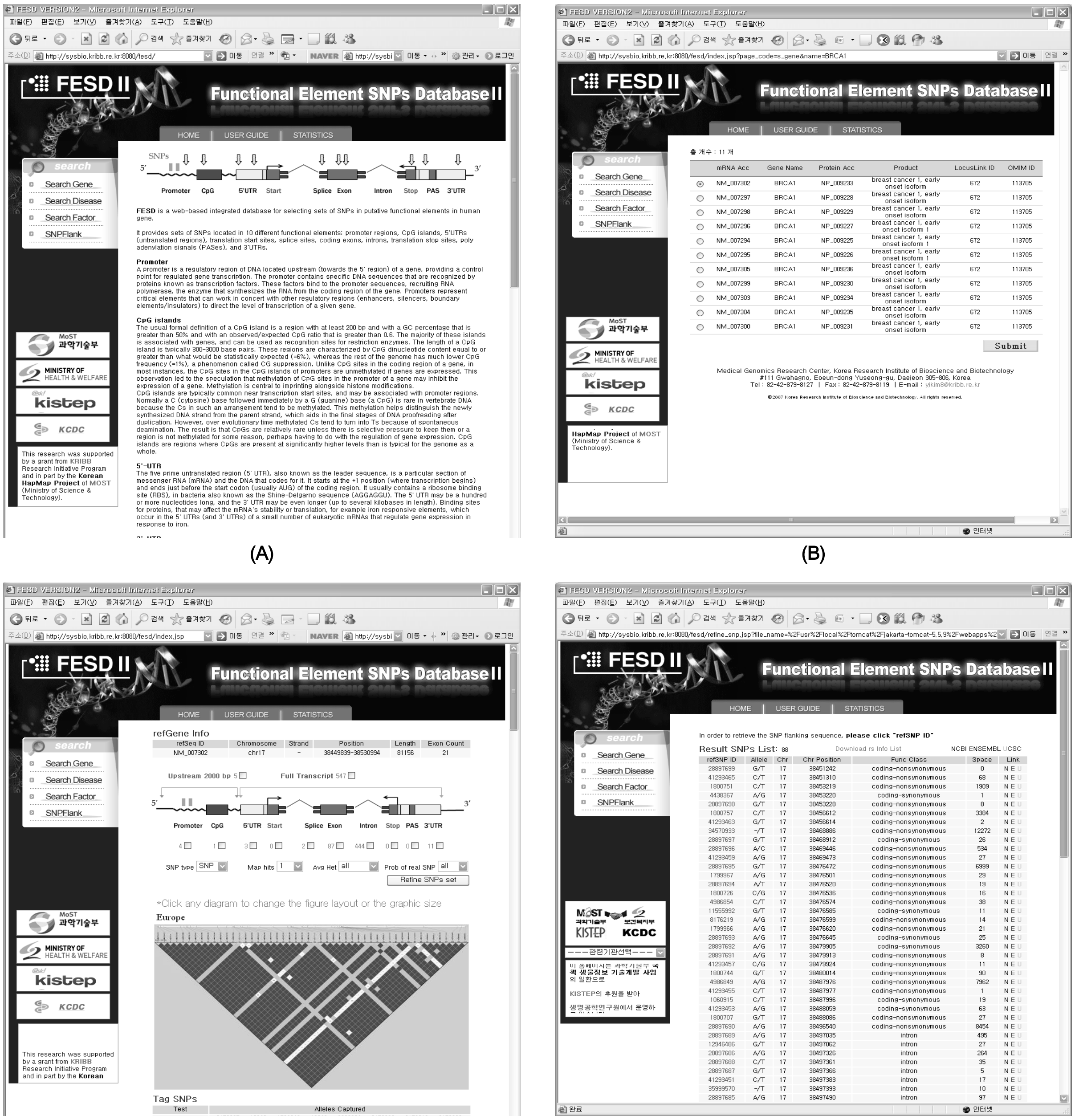


Fig. 1. Categorization of functional elements in human genes. Functional elements are divided into 10 different functional regions: promoter regions, CpG islands, 5'-untranslated regions (UTRs), translation start sites, splice sites, coding exons, introns, translation stop sites, polyadenylation signal sites (PASes) ,and 3'UTRs.



This research was supported by a grant from KRIBB Research Initiative Program and in part by the **Korean HapMap Project** of MOST (Ministry of Science & Technology).

FESD is a web-based integrated database for selecting sets of SNPs in putative functional elements in human gene.

It provides sets of SNPs located in 10 different functional elements; promoter regions, CpG islands, 5'UTRs (untranslated regions), translation start sites, splice sites, coding exons, introns, translation stop sites, poly adenylation signals (PASes), and 3'UTRs.

Promoter
A promoter is a regulatory region of DNA located upstream (towards the 5' region) of a gene, providing a control point for regulated gene transcription. The promoter contains specific DNA sequences that are recognized by proteins known as transcription factors. These factors bind to the promoter sequences, recruiting RNA polymerase, the enzyme that synthesizes the RNA from the coding region of the gene. Promoters represent critical elements that can work in concert with other regulatory regions (enhancers, silencers, boundary elements/insulators) to direct the level of transcription of a given gene.

CpG islands
The usual formal definition of a CpG island is a region with at least 200 bp and with a GC percentage that is greater than 50% and with an observed/expected CpG ratio that is greater than 0.6. The majority of these islands is associated with genes, and can be used as recognition sites for restriction enzymes. The length of a CpG island is typically 300-3000 base pairs. These regions are characterized by CpG dinucleotide content equal to or greater than what would be statistically expected (=6%), whereas the rest of the genome has much lower CpG frequency (=1%), a phenomenon called CG suppression. Unlike CpG sites in the coding region of a gene, in most instances, the CpG sites in the CpG islands of promoters are unmethylated if genes are expressed. This observation led to the speculation that methylation of CpG sites in the promoter of a gene may inhibit the expression of a gene. Methylation is central to imprinting alongside histone modifications.
CpG islands are typically common near transcription start sites, and may be associated with promoter regions. Normally a C (cytosine) base followed immediately by a G (guanine) base (a CpG) is rare in vertebrate DNA because the Cs in such an arrangement tend to be methylated. This methylation helps distinguish the newly synthesized DNA strand from the parent strand, which aids in the final stages of DNA proofreading after duplication. However, over evolutionary time methylated Cs tend to turn into Ts because of spontaneous deamination. The result is that CpGs are relatively rare unless there is selective pressure to keep them or a region is not methylated for some reason, perhaps having to do with the regulation of gene expression. CpG islands are regions where CpGs are present at significantly higher levels than is typical for the genome as a whole.

5'-UTR
The five prime untranslated region (5' UTR), also known as the leader sequence, is a particular section of messenger RNA (mRNA) and the DNA that codes for it. It starts at the +1 position (where transcription begins) and ends just before the start codon (usually AUG) of the coding region. It usually contains a ribosome binding site (RBS), in bacteria also known as the Shine-Delgarno sequence (AGGAGGU). The 5' UTR may be a hundred or more nucleotides long, and the 3' UTR may be even longer (up to several kilobases in length). Binding sites for proteins, that may affect the mRNA's stability or translation, for example iron responsive elements, which occur in the 5' UTRs (and 3' UTRs) of a small number of eukaryotic mRNAs that regulate gene expression in response to iron.

(A)

http://sysbio.kribb.re.kr:8080/fesd/index.jsp?page_code=s_gene&name=BRCA1

총 개수 : 11 개

	mRNA Acc	Gene Name	Protein Acc	Product	LocusLink ID	OMIM ID
●	NM_007302	BRCA1	NP_009233	breast cancer 1, early onset isoform	672	113705
○	NM_007297	BRCA1	NP_009228	breast cancer 1, early onset isoform	672	113705
○	NM_007298	BRCA1	NP_009229	breast cancer 1, early onset isoform	672	113705
○	NM_007296	BRCA1	NP_009227	breast cancer 1, early onset isoform 1	672	113705
○	NM_007294	BRCA1	NP_009225	breast cancer 1, early onset isoform 1	672	113705
○	NM_007295	BRCA1	NP_009226	breast cancer 1, early onset isoform 1	672	113705
○	NM_007305	BRCA1	NP_009236	breast cancer 1, early onset isoform	672	113705
○	NM_007299	BRCA1	NP_009230	breast cancer 1, early onset isoform	672	113705
○	NM_007303	BRCA1	NP_009234	breast cancer 1, early onset isoform	672	113705
○	NM_007304	BRCA1	NP_009235	breast cancer 1, early onset isoform	672	113705
○	NM_007300	BRCA1	NP_009231	breast cancer 1, early onset isoform	672	113705

Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology
#111 Gwahagno, Eoeun-dong Yuseong-gu, Daejeon 305-806, Korea
Tel : 82-42-879-8127 | Fax : 82-42-879-8119 | E-mail : yjkim8@kribb.re.kr

©2007 Korea Research Institute of Bioscience and Biotechnology. All rights reserved.

HapMap Project of MOST (Ministry of Science & Technology).

(B)

refGene Info

refSeq ID	Chromosome	Strand	Position	Length	Exon Count
NM_007302	chr17	-	38449839-38530994	81156	21

Upstream 2000 bp 5 □ Full Transcript 547 □

Promoter 4 □ CpG 1 □ 5'UTR 3 □ Start 0 □ Splice 2 □ Exon 87 □ Intron 444 □ Stop 0 □ PAS 0 □ 3'UTR 11 □

SNP type: SNP; Map hits: 1; Avg Het: all; Prob of real SNP: all

*Click any diagram to change the figure layout or the graphic size

Europe

Tag SNPs

Test	Alleles Captured

This research was supported by a grant from KRIBB Research Initiative Program and in part by the **Korean**

(C)

http://sysbio.kribb.re.kr:8080/fesd/refine_snp.jsp?file_name=%2Fusr%2Flocal%2Ftomcat%2Fjakarta-tomcat-5.5.9%2Fwebapps%2
MOST 과학기술부
보건복지부
KISTEP
KCDC
----관련기관선택----
완료
인터넷

이 홈페이지는 과학기술부 국책 생물정보 기술개발 사업의 일환으로

KISTEP의 후원을 받아

생명공학연구원에서 운영하

In order to retrieve the SNP flanking sequence, **please click "refSNP ID"**

Result SNPs List: 88 Download rs Info List NCBI ENSEMBL UCSC

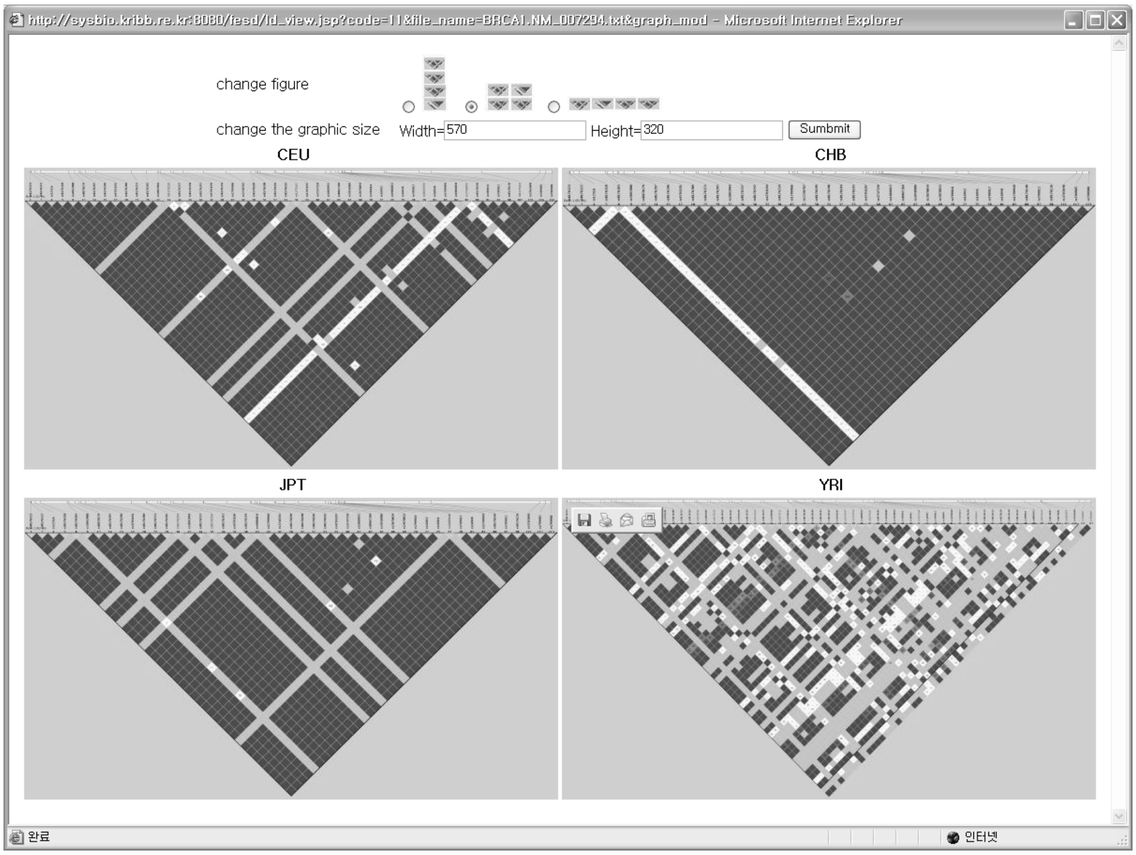
refSNP ID	Allele	Chr	Chr Position	Func Class	Space	Link
28897699	G/T	17	38451242	coding-nonsynonymous	0	N E U
41293465	C/T	17	38451310	coding-nonsynonymous	68	N E U
1800751	C/T	17	38453219	coding-nonsynonymous	1909	N E U
4438367	A/G	17	38453220	coding-synonymous	1	N E U
28897698	G/T	17	38453228	coding-nonsynonymous	8	N E U
1800757	C/T	17	38456612	coding-nonsynonymous	3384	N E U
41293463	G/T	17	38456614	coding-nonsynonymous	2	N E U
34570933	-/T	17	38468886	coding-nonsynonymous	12272	N E U
28897697	G/T	17	38468912	coding-synonymous	26	N E U
28897696	A/C	17	38469446	coding-nonsynonymous	534	N E U
41293459	A/G	17	38469473	coding-nonsynonymous	27	N E U
28897695	G/T	17	38476472	coding-nonsynonymous	6999	N E U
1799967	A/G	17	38476501	coding-nonsynonymous	29	N E U
28897694	A/T	17	38476520	coding-nonsynonymous	19	N E U
1800726	C/G	17	38476536	coding-nonsynonymous	16	N E U
4986854	C/T	17	38476574	coding-nonsynonymous	38	N E U
11555992	G/T	17	38476585	coding-synonymous	11	N E U
8176219	A/G	17	38476599	coding-nonsynonymous	14	N E U
1799966	A/G	17	38476620	coding-nonsynonymous	21	N E U
28897693	A/G	17	38476645	coding-synonymous	25	N E U
28897692	A/G	17	38479905	coding-synonymous	3260	N E U
28897691	A/G	17	38479913	coding-nonsynonymous	8	N E U
41293457	C/G	17	38479924	coding-nonsynonymous	11	N E U
1800744	G/T	17	38480014	coding-nonsynonymous	90	N E U
4986849	A/G	17	38487976	coding-nonsynonymous	7962	N E U
41293455	C/T	17	38487977	coding-nonsynonymous	1	N E U
1060915	C/T	17	38487996	coding-synonymous	19	N E U
41293453	A/G	17	38488059	coding-synonymous	63	N E U
1800707	G/T	17	38488086	coding-nonsynonymous	27	N E U
28897690	A/G	17	38496540	coding-nonsynonymous	8454	N E U
28897689	A/G	17	38497035	intron	495	N E U
12946486	G/T	17	38497062	intron	27	N E U
28897686	A/G	17	38497326	intron	264	N E U
28897688	C/T	17	38497361	intron	35	N E U
28897687	G/T	17	38497366	intron	5	N E U
41293451	C/T	17	38497383	intron	17	N E U
35999570	-/T	17	38497393	intron	10	N E U
28897685	A/G	17	38497490	intron	97	N E U

(D)

Fig. 2. (A) Web interface of FESD II. Users can search genes by gene name, mRNA accession #, protein accession #, LocusLink ID, OMIM ID, chromosome position, clinical disorder, clinical synopsis, cytogenetic band position, and transcription factor name. (B) List of mRNAs selected using BRCA1 as an example. (C) RefGene information of a given mRNA along with each functional element SNPs, their LD blocks, and tagSNPs. (D) Refined SNP information. One can obtain further information with the flanking sequence tool.

Perl scripts, and then imported into a MySQL relational database.

FESD II provides capabilities in searching for genes in several ways: gene name, mRNA accession number, protein accession number, LocusLink ID, OMIM ID, position on chromosome, clinical disorder, clinical synopsis, cytogenetic band position, and transcription factor name. Fig. 2 shows the main page of FESD II (Fig. 2 - A) and the steps taken to search for detailed information on SNPs, using BRCA1 as an example (Fig. 2 - B). The result shows the refGene information of 11 BRCA1 mRNAs, each with functional element SNPs (Fig. 2 - C). The LD blocks of four ethnicities



http://sysbio.kribb.re.kr:8080/fesd/ld_view.jsp?code=11&file_name=BRCA1.NM_007294.txt&graph_mod - Microsoft Internet Explorer

change figure ○ ● ○

change the graphic size Width=570 Height=320

Fig. 3. Example of LD blocks simulated in FESD II for the BRCA1 gene. One can change the width and height of the picture for better resolution. Four LD blocks of four ethnic groups are available. CEU: 60 unrelated European-descent individuals (from 30 parent-parent-offspring trios) from Utah. CHB: 45 Han Chinese individuals from Beijing. JPT: 45 Japanese individuals from Tokyo. YRI: 60 unrelated individuals (from 30 trios) from Ibadan, Nigeria.

Table 1. Distribution of human SNPs throughout functional elements

	Mean density (SNPs per kb)	*Mean spacing (bps per SNP)*	*SNP count (SNPs)*	*Total length (bps)*
Promoter	4.253	235.138	79,706	18,741,942
CpG	4.389	227.850	32,153	7,326,046
5'-UTR	4.464	224.029	23,088	5,172,392
Start Codon	1.330	751.871	101	75,939
Splice Site	2.345	426.485	2,279	971,960
Coding Exon	4.111	243.244	174,494	42,444,545
Intron	3.944	253.572	5,599,449	1,419,864,483
Stop Codon	1.962	509.658	149	75,939
PAS	1.969	507.957	141	71,622
3'-UTR	5.048	198.085	131,507	26,049,588

in HapMap data and tagSNPs are also shown. If one checks the functional elements and clicks 'refine SNPs set,' one can obtain additional SNP information such as allele, chromosomal position, functional class, and links to other databases (Fig. 2 - D). If one clicks the LD block image, pictured in Fig. 2 - D, the LD block window pops up (Fig. 3). In this pop-up window, one can see the four LD blocks of each population from the International HapMap Project data. For better resolution, the width and height of the picture are adjustable

Statistics

The genic region contains 5,952,641 SNPs (48.5%), while

Table 2. Density of SNPs in genic regions by chromosome *(SNPs per kb)*

	Promoter	*CpG*	*5' UTR*	*Start*	*Splice*	*Exon*	*Intron*	*Stop*	*PAS*	*3' UTR*
Chr1	4.264	4.411	4.835	0.905	2.303	4.308	3.943	2.199	2.659	5.199
Chr2	3.787	4.083	3.883	0.643	1.532	3.484	3.606	1.286	2.994	4.708
Chr3	3.911	3.790	4.348	2.763	2.035	3.687	3.739	1.507	2.096	4.402
Chr4	4.167	4.279	4.266	0.381	1.842	3.602	3.984	2.283	2.190	4.855
Chr5	4.078	4.668	4.460	1.514	1.777	3.851	3.699	3.330	1.117	4.622
Chr6	5.119	5.110	5.239	2.113	2.707	5.175	4.154	2.905	2.928	5.627
Chr7	4.493	4.350	4.678	1.147	2.409	4.178	4.139	0.574	1.253	5.402
Chr8	4.201	4.516	4.791	0.389	2.369	4.132	4.019	0.778	1.174	5.166
Chr9	4.392	4.387	4.159	1.418	2.208	4.003	4.303	1.418	1.160	5.119
Chr10	4.440	5.071	5.317	3.581	2.194	4.245	4.100	1.302	1.550	5.304
Chr11	4.678	4.975	4.782	1.288	2.567	4.471	4.104	2.361	3.623	5.086
Chr12	4.145	4.236	4.049	1.327	1.957	3.650	3.902	2.919	2.693	4.872
Chr13	4.316	4.270	4.511	1.792	2.029	3.653	4.070	2.688	1.032	4.622
Chr14	4.260	4.404	4.362	0.425	1.480	3.863	3.728	0	0	4.578
Chr15	4.010	4.464	3.647	0.447	1.488	3.635	3.866	2.681	1.462	4.711
Chr16	4.370	4.352	4.295	3.591	2.437	4.477	4.460	1.959	0.678	5.025
Chr17	4.159	3.891	4.202	1.404	2.532	4.123	3.757	2.107	0.969	5.302
Chr18	4.357	4.410	4.231	0	2.277	3.923	4.079	0.983	2.778	4.899
Chr19	4.712	4.262	4.815	1.220	2.551	4.633	4.261	3.254	0.600	5.656
Chr20	5.930	4.914	5.066	0.433	8.996	5.540	5.245	5.202	4.781	6.763
Chr21	5.096	4.813	3.981	0	2.234	3.991	5.162	1.955	2.283	5.948
Chr22	6.197	5.269	5.588	1.712	4.006	5.271	5.113	0	0	6.301
ChrX	2.579	2.438	3.016	1.122	2.281	2.841	2.606	0.842	1.440	3.661
ChrY	0.839	1.220	1.235	0	0	1.401	1.156	0	0	1.177
Average	4.271	4.274	4.323	1.234	2.425	4.006	3.966	1.856	1.728	4.959

the intergenic region contains 5,609,193 (51.5%). While there are slightly more SNPs in the intergenic region, the SNP density of the genic region is higher than that of the intergenic region. Total genome length was calculated to be 3,903,120,360 bp (http://www.ensembl.org). Table 1 shows the distribution of SNPs throughout 10 functional elements. Most SNPs in the genic region were distributed in introns (94.1% of total genic SNPs). The sum total of SNPs from 5' UTR to 3' UTR was 5,931,208, which is slightly less than the SNP count from the genic region. This slight difference is due to the overlapping physical positions of promoter regions and CpG islands. We estimated promoters in the 2 kb upstream from the start codon, and searched CpG islands in the estimated promoter regions. Therefore, the promoter region may overlap with an upstream gene, and for same reason, CpG islands can overlap.

The density of SNPs in each chromosome is shown in Table 2. In the promoter regions, CpG islands, 3'UTRs, 5'UTRs, coding exons, and introns, the average densities of SNPs were relatively higher than in other regions such as translation start sites, splice sites, translation stop sites, and PASes.

Conclusion

We have revised and updated our previous FESD in three ways. First, we have regenerated all the source data with recently updated databases. FESD II provides up-to-date information on functional element SNPs. Second, we simulated the LD blocks of each gene independently for the four ethnicities, based on International HapMap data (http://www.hapmap.org/). The LD blocks of each gene can be simulated at on time, and users can confirm the ethnic variation of SNP haplotypes with the data. Third, we ported the FESD from PHP to Java/JSP for higher performance.

Acknowledgments

This research was supported by the Korean HapMap Project and in part by the Cerebrovascular Disease Project of the Ministry of Science & Technology (MOST).

References

Barrett, J.C., Fry, B., *et al*. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21(2). 263-265.

Benson, D.A., Karsch-Mizrachi, I., *et al*. (2004). GenBank: update. *Nucleic Acids Res* 32(Database issue). D23-D26.

de Bakker, P.I., Yelensky, R., *et al*. (2005). Efficiency and power in genetic association studies. *Nat Genet.* 37(11). 1217-1223.

Jung, H.Y., Park, J.S., *et al*. (2004). HapAnalyzer: Minimum Haplotype Analysis System for Association Studies. *Genomics &Informatics* 2(2). 107-109.

Kang, H.J., Choi, K.O., *et al*. (2005). FESD: a Functional Element SNPs Database in human. *Nucleic Acids Res.* 33(Database issue). D518- D522.

Karolchik, D., Baertsch, R., *et al*. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* 31(1). 51-54.

Karolchik, D., Hinrichs, A.S., *et al*. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32(Database issue). D493-D496.

Kel, A.E., Gossling, E., *et al*. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31(13). 3576-3579.

Kent, W.J., Sugnet, C.W., *et al*. (2002). The human genome browser at UCSC. *Genome Res.* 12(6). 996-1006.

Maglott, D., Ostell, J., *et al*. (2005). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* 33(Database issue). D54-D58.

Marchini, J., Howie, B., *et al*. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 39(7). 906-913.

Nothnagel, M., Furst, R., *et al*. (2002). Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered.* 54(4). 186-198.

Ring, H.Z., and Kroetz, D.L. (2002). Candidate gene approach for pharmacogenetic studies. *Pharmacogenomics* 3(1). 47-56.

Sherry, S.T., Ward, M.H., *et al*. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29(1). 308-311.

Stranger, B.E., Forrest, M.S., *et al*. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315(5813). 848-853.

Wheeler, D.L., Barrett, T., *et al*. (2007). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 35(Database issue). D5-D12.

Wise, C.A., Paris, M., *et al*. (2003). A standard protocol for single nucleotide primer extension in the human genome using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom.* 17(11). 1195-1202.

Zhang, K., Qin, Z.S., *et al*. (2004). Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies. *Genome Res.* 14(5). 908-916.