Systematic Identification and Repository of RNA Editing Site in Human Genome

Jui-Hung Hung[1, †], Wei-Chi Wang[1, †], Hsien-Da Huang[1, 2, 3, ¶]
[1]*Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan*
[2]*Department of Biological Science and Technology, National Chiao Tung University, Hsin-Chu 300, Taiwan*
[3]*Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan*
bryan@mail.nctu.edu.tw

1: ABSTRACT

RNA editing is the modification of RNA sequence through nucleotide deletion, insertion or substitution mechanisms. RNA editing mechanisms can affect pre-mRNA splicing, frameshifts, RNA structural changes and mRNA translation. A database comprising RNA editing sites in mammalian genomes is crucial for deciphering RNA editing mechanism. Since two types of substitution RNA editing in human such as A-to-I editing and C-to-U editing were investigated recently, this work mainly focuses on the substitution RNA editing sites in human and excludes the insertion and deletion RNA editing. In this work, we present a database, namely EdRNA, comprising the RNA editing sites which are computationally predicted based on the expressed sequences such as mRNA sequences and EST sequences. Computational methods based on comparative sequence analysis were applied for identifying RNA editing sites. The RNA editing site candidates are also cross-referenced to gene structures, repeats and SNP. Moreover, cross-species comparison between human and mouse is performed to confer the evolution meanings of the RNA editing sites. Both textual and graphical interface are provided for effectively retrieving the RNA editing sites in EdRNA. The resource is now freely available at http://EdRNA.mbc.nctu.edu.tw/.

2: INTRODUCTION

RNA editing is the modification of RNA sequence through nucleotide deletion, insertion, or substitution mechanisms. RNA editing mechanisms can affect many biological processes such as the pre-mRNA splicing, frameshfits, RNA structural changes and mRNA translation. Substitution RNA editing is an RNA processing event that alters the nucleotide sequence of a substrate. Changing in nucleotide composition is likely to vary the translation product and to trigger alternative biological processes (1). Moreover, the RNA secondary structure may be changed into more or less stable conformation, which deters or precipitates the interaction with spliceosome, polymerase II and other factors (2).

In mammals, two types of substitution RNA editing, such as A-to-I editing (3) and C-to-U editing (4), were reported recently. It reported that A-to-I editing is induced by an enzyme ADARs (Adenosine Deaminased that Acts on RNA), which catalyses adenosine (A) to inosine (I) and interacts with duplex RNA with low selectivity (2). ADAR1 knockout mice die embryonically and ADAR2 knockout mice were born at full term but die prematurely (5). In human, altered editing levels have also been observed in malignant gliomas, schizophrenic patients and suicide victims, and may be affected in patients with Alzheimer's and Huntington's disease (6). Additionally, the C-to-U editing is induced by APOBEC1 (apoB mRNA-editing cytidine deaminase subunit-1), which is an important factor that determine the LDL and VLDL metabolism (7).

Since the substitution RNA editing is a significant mechanism affecting organisms in diverse biological processes, the determination of the occurrences and locations of the RNA editing sites are essential and valuable. In plant, there are a few substitution RNA editing databases available, such as ChloroplastDB (8) and PREP-Mt (9). ChloroplastDB used BLAST to perform all-against-all sequence alignment to find the mismatches between genes, nucleotide sequences and annotated protein sequences. PREP-Mt incorporated a statistically trained classifier to identify the substitution RNA editing sites. Furthermore, Levanon et al successfully constructed a data set of A-to-I editing sites in human by a systematic computational method (5). Their work mainly focuses on A-to-I editing and is with lack of further annotations of the putative RNA editing sites.

For lack of this sort of comprehensive compilation and prediction of substitution RNA editing database for human genome (10), this work presents a database, namely EdRNA, comprising the RNA editing sites which are computationally predicted based on the expressed sequences such as mRNA sequences and EST sequences. In addition to the two major types of RNA editing sites previously discovered, such as A-to-I

editing and C-to-U editing, this work also collects the other types of substitution RNA editing sites. BLAST program were applied for sequence comparison between gene, mRNA and EST sequences. A semi-global pairwise alignment method is implemented here for identifying RNA editing sites within the matching blocks generated by BLAST. Moreover, the RNA editing site candidates are also cross-referenced to the gene structures, repeats and SNP. In addition to the genomic location of putative RNA editing sites, cross-species comparison between human and mouse is performed to confer the evolution meanings of the RNA editing sites. Both the textual and graphical interface are provided for effectively retrieving the RNA editing sites in EdRNA.

3: DATA GENERATION

The data generation flow of the EdRNA is briefly depicted in Fig. 1. The data generation flow contains three major steps: (i) the data collection and preprocessing; (ii) the prediction of RNA editing sites by aligning gene, mRNA and EST sequences; and (iii) the cross-references to numerous gene annotations, repeats, SNP and human/mouse conserved regions for each predicted RNA editing site.

Firstly, the gene annotations, mRNA sequences and EST sequences in human were extracted from Ensembl (11) and GenBank (12). Then, for each gene group comprising gene, mRNA and EST sequences, EdRNA performed sequence alignment among gene, mRNA and EST sequences to determine the mismatch nucleotide for RNA editing sites. Finally, each predicted editing sites is further annotated by referring to the gene annotations and other genomic loci information, such as gene structures (5'UTR, coding region and 3'UTR), repeats, Single Nucleotide Polymorphism (SNP) and human/mouse conserved regions. The detailed of each component is described below.

3.1: Data preprocessing

Table 1 gives the list of the data sources used in EdRNA. The genomic sequences and gene information were obtained from Ensembl (Release Nov, 2005) and NCBI Genbank (Human Genome Build 35). The mRNA sequences were retrieved from Ensembl (Release Nov, 2005), Refseq (13) (Release 9 Dec, 2005) and EMBL (Release 9 Dec, 2005). The human gene clusters were obtained from UniGene (14) (Release 9 Dec, 2005). EST sequences and cDNA library information were obtained from dbEST (15) (Release 9 Dec, 2005). SNP data are obtained from dbSNP (16) (Build 124). The information of repeats in human genome were obtained from Ensembl. Human and mouse conserved regions and gene structure information (5'UTR, coding regions and 3'UTR) were obtained from UCSC Genome Browser Database (17) (Release 10 May, 2006). All databases were integrated into MySQL database system.

Category	Data sources	No. of entries	Ref.
Gene sequences/annotations	GenBank	23,030	(12)
	Ensembl	34,370	(11)
mRNA sequences	Ensembl	39,240	(11)
	EMBL	143,104	(18)
	RefSeq	28,102	(13)
EST sequences	dbEST	5213,323	(15)
SNP information	dbSNP	9,276,275	(16)
Repeats	Ensembl	3,243,190	(13)
Gene structures	UCSC Genome Browser Database	66,215	(19)
Human/mouse conserved regions	UCSC Genome Browser Database	19,123,554	(19)

Table 1. The list of external databases used in this work.

3.2: Predicting RNA editing sites by aligning gene, mRNA and EST sequences

Following the collection and the preprocessing of all required data in human genome, the substitution RNA editing sites were identified based on the alignment between the expressed sequences belonging to the same gene, such as mRNA and EST sequences. The prediction of RNA editing sites comprises three steps. Firstly, this work groups mRNA and EST sequences according to the annotation obtained from UniGene, whose data were constructed by performing large-scale sequence alignment among all available mRNA and EST sequences in human. Secondly, pariwise alignments between mRNA and EST sequences belonging to the same gene group were performed by BLAST (20). The matched regions with sequence identity greater than 95% were retained. In order to efficiently find the matching blocks between the query sequence and sequence database, BLAST performs the alignment by heuristic methods, which is insufficient for precisely determining the mismatch for RNA editing sites investigated in this work. Therefore, this work implemented a program which is based on semi-global alignment to determine the exact position of RNA editing site within the matched blocks generated by BLAST.

Finally, for determining the chromosomal locations of RNA editing sites, this work performed the alignment between the gene sequence and the mRNA sequence. From the alignment results generated by SIM4 (21), the exonic and intronic regions within the gene sequence corresponding to the mRNA sequence can be determined. Thus, the editing site positions discovered by previous steps can be mapped to the chromosomal positions as well as the exonic and intronic regions.

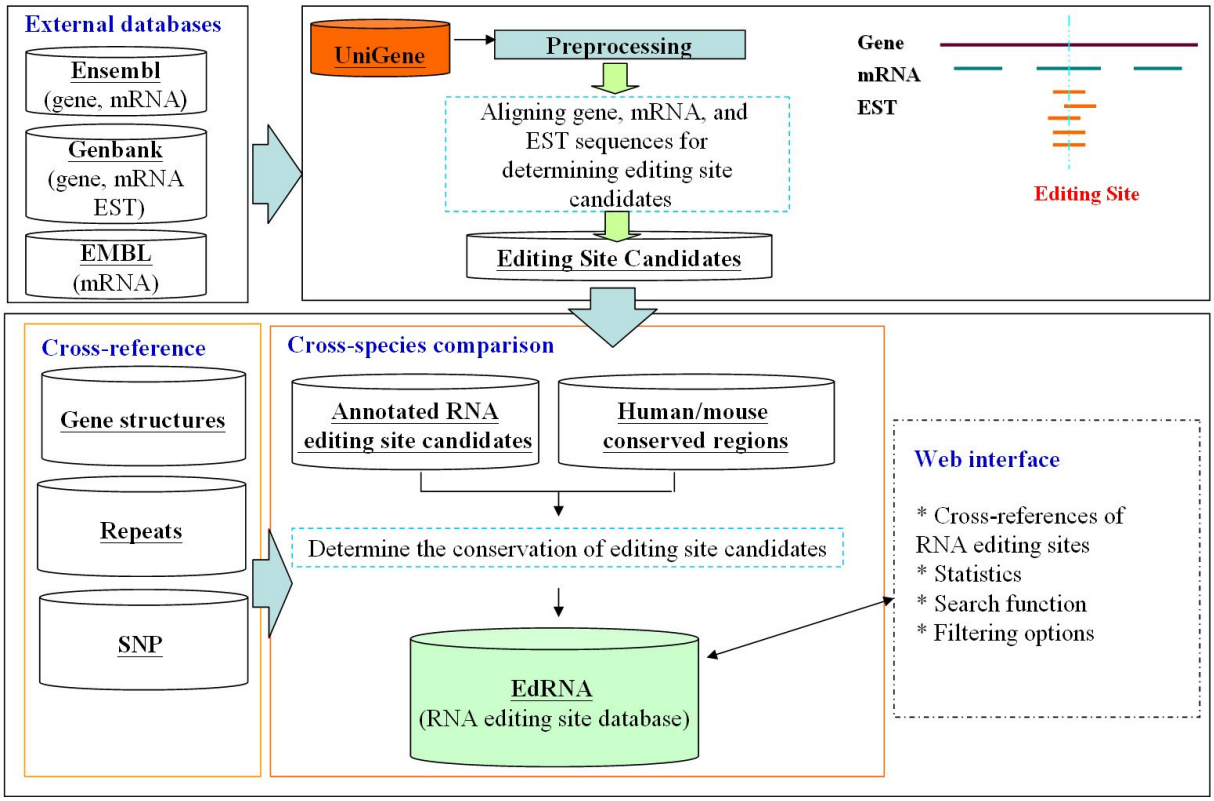

Figure 1. The data generation flow of the EdRNA database.

3.3 Cross-references to gene annotations, SNP, repeats and conserved regions

For each RNA editing site detected in previous step and supported by at least ten EST sequences, it is further annotated by referring the occurrence of the editing sites to other genomic regions with other biological meanings, such gene structures (5'UTR, coding regions and 3'UTR), repeats, SNP and human/mouse conserved regions.

Referring to gene structures, EdRNA determines whether the editing site is located in 5'UTR, coding region or 3'UTR of genes. RNA editing sites occurring in different regions leads to diverse influence with the regulatory mechanism in mRNA and protein level. Generally speaking, an editing site locating in 5'UTR may interfere with the initiation of translation and an editing site within the coding region may change the residue of protein, which possibly affects the protein satiability and function.

According to the chromosomal locations of the predicted RNA editing sites and the chromosomal locations of the SNP obtained from dbSNP, this work examines whether these editing sites are annotated as SNP. Eisenberg et al presented the evidence for hundreds of dbSNP records that are actually editing sites (22). Therefore, even an editing site is annotated as SNP, biologists should ruminate its reliability with the consideration to other effective information provided in EdRNA.

As to the cross-references to repeats in genomic sequence, EdRNA annotates each editing site if it occurs within the genomic regions of repeats. Since several previous studies suggest that substitution RNA editing may be associated with the repeats, for instance, A-to-I editing sites are reported to frequently occur in Alu repeat regions. Furthermore, this work also examines whether the RNA editing sites are conserved both in human and mouse. The human and mouse conserved regions were obtained from UCSC Genome Browser Database (17), and the conserved regions with the sequence identity greater than 70% were selected. The substitution RNA editing sites are more believable if they are within the conserved regions between human and mouse.

4. DATA STATISTICS

In EdRNA, there are totally 575,320 editing sites and each editing site is supported averagely by 14.819 EST sequences. The number of editing sites supported by at least ten EST sequences is 318,126. Table 2 gives the distributions of different editing types, where each editing site is supported by at least ten ESTs. For instance, there are 79934 A/G (A-to-G and G-to-A) editing sites, which are discovered in 5185 genes. Among these 318,126 editing sites, there are 212819, 10759 and 94548 editing sites located in coding region, 5'UTR and 3'UTR, respectively (Table 4).

RNA editing types	No. of editing sites	No. of genes
A/G (A-to-G, G-to-A)	79,934	5,185
C/T (C-to-T, T-to-C)	68,486	5,053
C/G (C-to-G, G-to-C)	44,219	4,237
A/T (A-to-T, T-to-A)	30,612	3,283
G/T (G-to-T, T-to-G)	50,527	4,538
A/C (A-to-C, C-to-A)	44,438	4,418

Table 2. The number of editing sites of each editing type.

Furthermore, this work compared the occurrences of the RNA editing sites to the existence of repeats, such as SINE, LINE, LTR and Alu. Table 3 gives the number of editing sites corresponding to each type of repeat. For instance, there are 1007 editing sites found in the region of Alu repeats. It shows that about 97% of editing sites do not occur in any repeats. Table 4 gives the number of editing sites of different editing types in coding region, 5'UTR and 3'UTR. For instance, there are 55920, 2534, 21480 A/G editing sites located in coding region, 5'UTR and 3'UTR, respectively. The number of editing sites for each editing type in human/mouse conserved regions is given in Table 5. For instance, 37,838 C/T editing sites are found in human/mouse conserved regions.

Repeat type	No. of editing sites in repeat regions
SINE/ALU	1,007
SINE/MIR	407
LTR/MaLR	201
LTR/ERV1	157
Trf	1,280
LINE/L1	211
LINE/L2	383
Dust	3,776
Simple_repeat	629
DNA/MER1_type	244
None (not in repeat region)	303,106

Table 3. The number of editing sites within different types of repeats

RNA editing type	No. of editing sites in coding region	No. of editing sites in 5'UTR	No. of editing sites in 3'UTR
A/G	55,920	2,534	21,480
C/T	45,687	2,781	20,018
C/G	30,823	1,600	11,796
A/T	18,829	784	10,999
G/T	31,580	2,187	16,760
A/C	29,980	873	13,495
Total	212,819	10,759	94,548

Table 4. The number of editing sites of each RNA editing type within coding region, 5'UTR and 3'UTR.

The average number of editing sites for every Mbps in chromosome one and nineteen are 249 and 781, respectively. This suggests that the occurrences of RNA editing sites are not likely to be expected at random.

RNA editing type	No. of editing sites in conserved region	Total No. of editing sites
A/G	44,789	79,934
C/T	37,838	68,486
C/G	24,300	44,219
A/T	15,923	30,612
G/T	27,047	50,527
A/C	24,987	44,348

Table 5. The number of editing sites of each RNA editing type within human/mouse conserved regions

5. INTERFACE

This work designs and develops a web interface to facilitate the access to the content of EdRNA. The search function allows users to input the gene accession number (GenBank, Ensembl), mRNA accession number (RefSeq, Ensembl, EMBL) and gene symbol to find the gene or mRNA, which are interested by users. Users can browse the database by selecting a variety of filtering options, such as editing site types, editing site occurrence location (in 5'UTR, coding region or 3'UTR) and the amount of EST sequences supporting an editing site. After clicking on "Get result" button, the system returns all possible editing sites that satisfy the filtering conditions.

Figure 2 demonstrates the graphical view of the mapping between gene and mRNA to reveal the position of RNA editing sites in mRNA. The partial result of the alignment between the gene and mRNA sequence is also provided, where the nucleotide of editing sites are highlighted. Additionally, the alignment of mRNA and EST sequences, tissue distributions of the EST sequences and human/mouse conserved regions are also provided.

6. CONCLUSIONS

EdRNA allows biologists to retrieve the information of the substitution RNA editing sites with sufficient length of flanking sequences for designing primers for experiments.

Moreover, recent research only discovered two types of substitution RNA editing sites in human such as A-to-I editing and C-to-U editing. Although there is no previous study reporting other types of RNA editing, this work discovered a large amount of evidences to support the existence of other types of substitution RNA editing. Therefore, it is reasonable to hypothesize that other types of RNA editing, for instance C-to-A editing and T-to-C editing, might exist in human.

The hypothesis can be validated by experimentally confirming the predicted RNA editing sites in EdRNA.

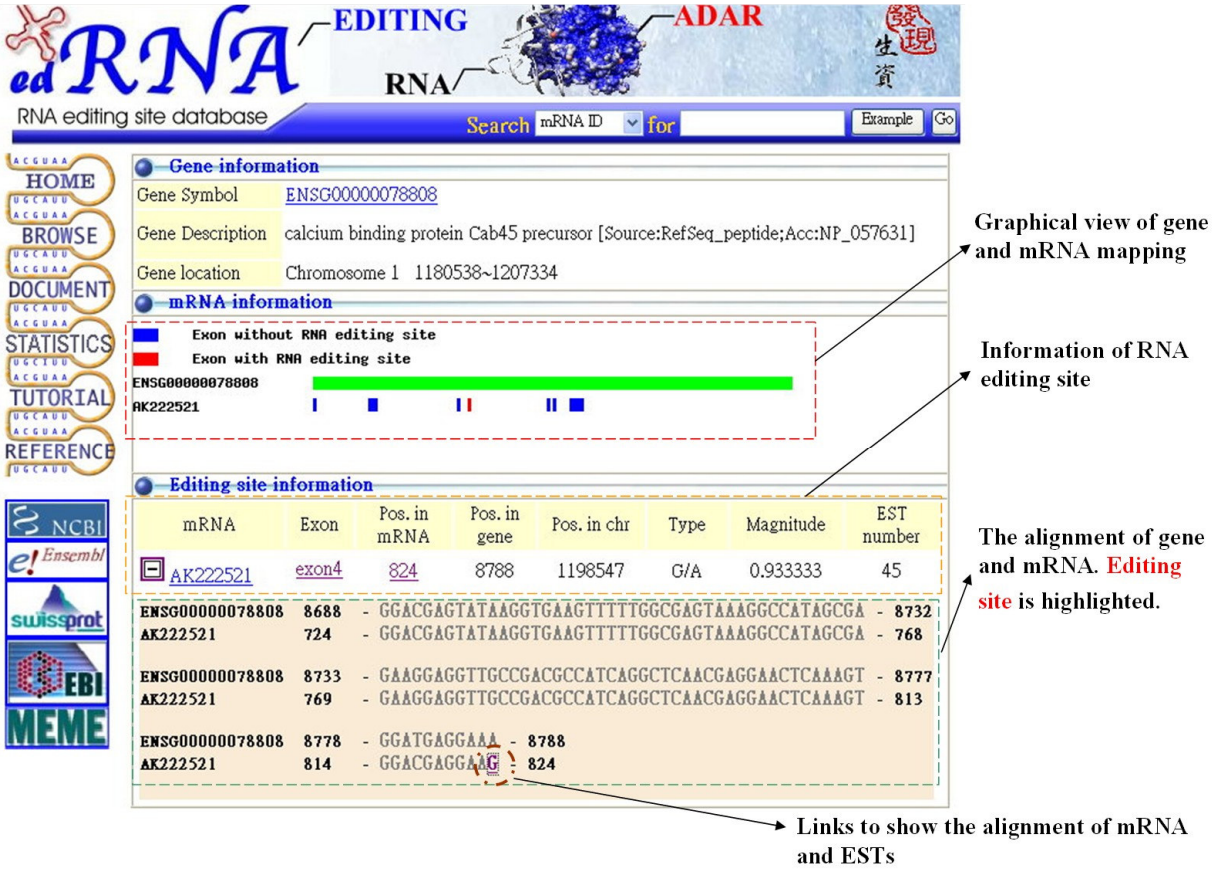

Figure 2. The mapping between gene and mRNA to show the RNA editing site.

In summary, EdRNA is a novel and comprehensive database to curate all possible types of substitution RNA editing sites in human genome. The proposed web interface is convenient and effective for retrieving the annotated RNA editing sites. The main contribution of this work is successfully providing a variety of valuable experimental candidates for the investigation of RNA editing mechanisms.

7. AVAILABILITY

The EdRNA resource will be regularly maintained and updated. The resource is now freely available at http://EdRNA.mbc.nctu.edu.tw/.

8. ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 95-3112-E-009-002. Special thanks for the financially supports from National Research Program For Genomic Medicine (NRPGM), Taiwan. This work was also partially supported by MOE ATU.

9. REFERENCES

1. Smith, H.C., Gott, J.M. and Hanson, M.R. (1997) A guide to RNA editing. *Rna*, **3**, 1105-1123.
2. Laurencikiene, J., Kallman, A.M., Fong, N., Bentley, D.L. and Ohman, M. (2006) RNA editing and alternative splicing: the importance of co-transcriptional coordination. *EMBO Rep*, **7**, 303-307.
3. Tonkin, L.A., Saccomanno, L., Morse, D.P., Brodigan, T., Krause, M. and Bass, B.L. (2002) RNA editing by ADARs is important for normal behavior in Caenorhabditis elegans. *Embo J*, **21**, 6025-6035.
4. Blanc, V. and Davidson, N.O. (2003) C-to-U RNA editing: mechanisms leading to genetic diversity. *J Biol Chem*, **278**, 1395-1398.
5. Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D. *et al.* (2004) Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotechnol*, **22**, 1001-1005.
6. Gott, J.M. (2003) Expanding genome capacity via RNA editing. *C R Biol*, **326**, 901-908.

7. Maris, C., Masse, J., Chester, A., Navaratnam, N. and Allain, F.H. (2005) NMR structure of the apoB mRNA stem-loop and its interaction with the C to U editing APOBEC1 complementary factor. *Rna*, **11**, 173-186.
8. Cui, L., Veeraraghavan, N., Richter, A., Wall, K., Jansen, R.K., Leebens-Mack, J., Makalowska, I. and dePamphilis, C.W. (2006) ChloroplastDB: the Chloroplast Genome Database. *Nucleic Acids Res*, **34**, D692-696.
9. Mower, J.P. (2005) PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinformatics*, **6**, 96.
10. Blow, M., Futreal, P.A., Wooster, R. and Stratton, M.R. (2004) A survey of RNA editing in human brain. *Genome Res*, **14**, 2379-2387.
11. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res*, **30**, 38-41.
12. Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J. and Ouellette, B.F. (1998) GenBank. *Nucleic Acids Res*, **26**, 1-7.
13. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**, D501-504.
14. Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E. *et al.* (1996) A gene map of the human genome. *Science*, **274**, 540-546.
15. Boguski, M.S., Lowe, T.M. and Tolstoshev, C.M. (1993) dbEST--database for "expressed sequence tags". *Nat Genet*, **4**, 332-333.
16. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308-311.
17. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F. *et al.* (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*, **34**, D590-598.
18. Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R. *et al.* (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*, **32**, D27-30.
19. Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res*, **31**, 51-54.
20. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
21. Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res*, **8**, 967-974.
22. Eisenberg, E., Adamsky, K., Cohen, L., Amariglio, N., Hirshberg, A., Rechavi, G. and Levanon, E.Y. (2005) Identification of RNA editing sites in the SNP database. *Nucleic Acids Res*, **33**, 4612-4617.