D810–D815 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 25 October 2008*
doi:10.1093/nar/gkn798

VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts

Makoto K. Shimada[1,2], Ryuzou Matsumoto[3], Yosuke Hayakawa[2,3], Ryoko Sanbonmatsu[1,2], Craig Gough[1,2], Yumi Yamaguchi-Kabata[1], Chisato Yamasaki[1,2], Tadashi Imanishi[1,*] and Takashi Gojobori[1,4]

[1]Integrated Database and Systems Biology Team, Biomedicinal Information Research Center, National Institute of Advanced Industrial Science and Technology, [2]Japan Biological Informatics Consortium (JBIC), [3]Hitachi Software Engineering Co., Ltd., Tokyo and [4]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, Japan

Received August 14, 2008; Revised October 8, 2008; Accepted October 10, 2008

ABSTRACT

Creation of a vast variety of proteins is accomplished by genetic variation and a variety of alternative splicing transcripts. Currently, however, the abundant available data on genetic variation and the transcriptome are stored independently and in a dispersed fashion. In order to provide a research resource regarding the effects of human genetic polymorphism on various transcripts, we developed VarySysDB, a genetic polymorphism database based on 187 156 extensively annotated matured mRNA transcripts from 36 073 loci provided by H-InvDB. VarySysDB offers information encompassing published human genetic polymorphisms for each of these transcripts separately. This allows comparisons of effects derived from a polymorphism on different transcripts. The published information we analyzed includes single nucleotide polymorphisms and deletion–insertion polymorphisms from dbSNP, copy number variations from Database of Genomic Variants, short tandem repeats and single amino acid repeats from H-InvDB and linkage disequilibrium regions from D-HaploDB. The information can be searched and retrieved by features, functions and effects of polymorphisms, as well as by keywords. VarySysDB combines two kinds of viewers, GBrowse and Sequence View, to facilitate understanding of the positional relationship among polymorphisms, genome, transcripts, loci and functional domains. We expect that VarySysDB will yield useful information on polymorphisms affecting gene expression and phenotypes. VarySysDB is available at http://h-invitational.jp/varygene/.

INTRODUCTION

Accumulated information on human genetic polymorphisms has encouraged genome-wide association studies that use polymorphisms as markers. This approach is now commonly used in various studies (1,2), and has led to a greater understanding of the diversity in phenotypes as well as pathogenic biological processes.

Currently, several kinds of human genetic polymorphism databases aid researchers in exploring genetic information for various applications. Examples of such databases include the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/index.html) (3) containing information on single nucleotide polymorphisms (SNPs) and short deletion and insertion polymorphisms (DIPs) as submitted by the corresponding authors of the published data. The Database of Genomic Variants (http://projects.tcag.ca/variation/) (4) provides genomic regions involved in structural variations as defined by alternations in DNA segments larger than 1 kb. Short Tandem Repeats (STRs), also known as simple sequence repeats or microsatellites are a different type of major source of genomic diversity. Information on human STRs is available from public domains such as UgMicroSat*db* (http://www.veenuash.info/web1/index.htm) (5), UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgGateway) (6) and GDBS/H-GOLD (http://hinv.jp/gdbs/) (7). Accordingly, these polymorphism data are described by

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: t.imanishi@aist.go.jp
Correspondence may also be addressed to Takashi Gojobori. Tel: +81 55 981 6847; Fax: +81 55 981 6848; Email: tgojobor@genes.nig.ac.jp
Present address:
Makoto K. Shimada, Institute for Comprehensive Medical Science, Fujita Health University, Aichi 470-1192, Japan

position in the human reference genome (i.e. genome coordinate).

Recently, the accumulated knowledge on alternative splicing and the regulation of gene expression has reinforced the importance of transcript multiplicity as another source of diversity of protein function. H-InvDB catalogs a comprehensive annotation of the human transcriptome including transcript diversities and gene expression profiles (8,9). H-InvDB is concentrating on full-length cDNA annotation to overcome the limitation of conventional databases based on high-throughput EST data, such as scarce distributions around the 5′-ends of mRNAs and absence of some combination of the alternative splicing (AS) exons (10). Thus, H-DBAS, a satellite database of H-InvDB which is a specialized AS database, is the comprehensive database containing AS accurately annotated manually and automatically based on highly reliable cDNA sequences (11).

The currently available human genetic polymorphism databases described above have not yet been integrated with well-annotated AS isoforms conforming to a uniform standard. Therefore, we developed VarySysDB, a database of human genetic polymorphisms based on all of the 187 156 matured mRNA transcripts from 36 073 loci provided by H-InvDB. [Hereinafter, these matured mRNA transcripts annotated by H-InvDB and the loci defined by transcript clusters will be called H-inv transcripts (HITs) and H-Inv clusters (HIXs), respectively]. VarySysDB provides separately annotated genetic polymorphisms for each HIT, even from multiple transcripts forming a HIX. It provides information regarding SNPs, DIPs, STRs, single amino acid repeats (SARs), structural variation (or copy number variations; CNVs), linkage disequilibrium (LD) regions and their relationship with the genome, HITs, and functional domains. Moreover, we designed VarySysDB to include annotations we made, which covers intronic SNPs located on conserved dinucleotide splice sites, nonsynonymous SNPs that affect functional (InterPro) and protein structural (SCOP) domains, and polymorphic tandem repeat sequences, as well as other publicly available information. Since VarySysDB is a satellite database of H-InvDB, it is well designed to provide appropriate links for each HIT to H-InvDB, as well as to other related public databases. All of the annotation data in VarSysDB is available to all users, with no restriction to academic users only. We hope that VarySysDB will deliver an even greater understanding of the various biological processes, permit a detailed evaluation of how polymorphisms affect different phenotypes, and foster a rich research environment focused on exploring the causes of genetic variation through genome-wide association studies.

CONSTRUCTION OF THE DATABASE

Source of data

Table 1 lists the data used to construct VarySysDB. This database includes the transcript data from H-InvDB, as well as published genetic polymorphism data.

Mapping genetic polymorphism on H-Inv transcripts

We mapped all the genetic polymorphism data onto the exact transcript position using our in-house program to convert their location from genome coordinates to those of the HIT.

VarySysDB contains these polymorphism data with the following conditions as well as our own annotations. (i) SNPs and DIPs: SNP and DIP data were downloaded from dbSNP (Table 1). We eliminated SNP and DIP

Table 1. Data used in VarySysDB

	Number of data available in VarySysDB	Database: name and version (or date of download)	Provider	URL	References
H-Inv Transcripts (HITs)	187 156	H-InvDB 5.0[a]	BIRC[b]	http://h-invitational.jp/hinv/	(8,9)
SNPs & DIPs	11 817 893[c]	dbSNP build 128	NCBI[d]	http://www.ncbi.nlm.nih.gov/projects/SNP/	(3,14)
STRs	18 637	H-InvDB 5.0	BIRC	http://h-invitational.jp/hinv/	(8,9)
SARs	33 007	H-InvDB 5.0	BIRC	http://h-invitational.jp/hinv/	(8,9)
CNVs	11 966	DGV (hg18.v3)[e]	TCAG[f]	http://projects.tcag.ca/variation/	(4)
LD-bins	99 921	D-HaploDB[g]	Kyushu University	http://orca.gen.kyushu-u.ac.jp/	(13)
OMIM allelic variants	950	OMIM (1.28. 2008)[h]	NCBI	http://www.ncbi.nlm.nih.gov/omim/	(14,15)
Functional domain	–	InterPro 15.1	EBI[i]	http://www.ebi.ac.uk/interpro/	(16,17)
Structural domain (SCOP)	–	GTOP (2.4. 2008)[j]	NIG[k]	http://sybock.genes.nig.ac.jp/~hinv4/gtop.html	(18)

[a]H-Invitational Database Release 5.0 (used human genome sequence version: hg18, NCBI Build36.2).
[b]Biomedicinal Information Research Center.
[c]The number of SNP & DIP data downloaded from dbSNP and analyzed for VarySysDB. The numbers of annotated SNP and HIT pairs are as follows: 568 982 for non-synonymous, 431 433 for synonymous, 747 for synonymous at stop-codon, 7227 for termination, 1510 for stop codon to amino acid, 8945 for NMD.
[d]National Center for Biotechnology Information.
[e]Database of Genomic Variants.
[f]The Centre for Applied Genomics.
[g]Database of Definitive Haplotypes.
[h]Online Mendelian Inheritance in Man™.
[i]European Bioinformatics Institute.
[j]PDB 2007-Apr-6, Swissprot 52.1, SCOP 1.69, Pfam 21.0, ProSite 20.0, Wormpep 174, HUGE 2003-11-6(kiaa2038).
[k]National Institute of Genetics.

Figure 1. View of polymorphism search page, which is one of the search pages contained in VarySysDB. In the polymorphism search page, users can search the polymorphism data by features, classification and our analysis results such as effects on functional domains and protein 3D structures. Four boxes ('Search by postion', 'Polymorphism features', 'Polymorphism classification', 'Search for analysis result') organize the search criteria by subject. When multiple search criteria are specified 'over' these boxes, an 'and' search is conducted, offering polymorphisms matching all the specified criteria.

data if their alleles contradicted the transcript sequences (12). (ii) STRs and SARs: We searched HIT sequences for STRs, with an STR defined as a repeat of ten or more dinucleotides and a repeat of five or more tri-, tetra- and penta-nucleotide sequences. For SARs, we searched the amino acid sequences translated from HIT sequences for single amino acid repeats of five or more. (iii) OMIM allelic variant (OMIM AV): we downloaded OMIM AVs with MIM Number Prefixes of 'gene with known sequence' using the 'limit' GUI of the OMIM web page (Table 1). We filtered the OMIM AVs in the exonic region included in the dbSNP by checking each location in the HIT using our in-house programs to annotate separately.

Annotation

We classified SNPs according to their effect on translation based on each HIT sequence (Table 1). This highlighted our unique annotation regarding the effect of each SNP on different HITs within a HIX. We also classified SNPs and DIPs according to their locations in HITs, which includes the promoter (defined as the region within 2 kb upstream of first exon), the splice dinucleotide site or the exonic regions. Furthermore, we annotated SNPs and DIPs in the coding region into the following categories: (i) those that alter functional (InterPro) domain sequences so drastically that InterProcScan results change (effect on functional domain); (ii) those that are located in protein structural (SCOP) domains and change amino acid characters so as to result in harmful effects on the protein 3D structure; (iii) those that match their location in HITs and alleles to descriptions of OMIM AVs (OMIM Allelic Variants) (Figure 1). Cases in category (ii), those that have an effect on protein structure, are subdivided into three subcategories chosen according to the effect of the polymorphism: (a) 'Not Harmful'; (b) 'Recessively Harmful' due to loss or reduction of function; (c) 'Possible to be Harmful (Unclear)' because of a drastic change of a structural domain which may induce toxic aggregation.

Within STRs and SARs, we distinguished the polymorphic cases by transcript sequence alignments.

For CNVs, we downloaded from DGV (Table 1), and classified them according to the detection methods described in the downloaded data into six divisions for the convenience of users.

ACCESSING THE DATABASE

Database contents and organization

Table 2 lists the web-interfaces or GUIs in VarySysDB containing six search pages. The results of searches can be downloaded as well as easily displayed on the computer screen. VarySysDB is composed of three subsystems, including Varygene2, LD Search System and GBrowse. A menu bar of Varygene2 is designed to select search pages from 'Polymorphisms', 'Transcripts', 'STRs/SARs' and 'CNVs.'

By clicking 'Polymorphisms', users can search by feature and our aforementioned annotation regarding SNPs and DIPs (Figure 1).

STRs/SARs with length polymorphisms proven by our sequence alignment can be extracted from an STR/SAR Search page. VarySysDB can retrieve STRs and SARs according to features such as the repeat unit sequence (e.g. 'at' nucleotides for STR, 'P' amino acid for SARs) and number of repeats.

By clicking 'CNVs', users can search by features of CNVs, such as CNV class (i.e. copy number variation or inversion) and detection method.

Table 2. System and web-interface design of VarySysDB

Subsystem	Web-interface	Function
Varygene 2	Polymorphism search	Retrieving and displaying genetic polymorphisms.
	Polymorphism table	Displaying detailed information on polymorphisms.
	Transcript search	Retrieving and displaying transcript information.
	Transcript table	Displaying detailed information on transcripts.
	Sequence view	Displaying cDNA sequence with information on polymorphisms and functional domains.
	STR/SAR search	Retrieving and displaying STRs and SARs.
	CNV search	Retrieving and displaying CNVs.
	CNV table	Displaying detailed information on CNVs.
	Keyword search	Retrieving and displaying by ID, gene name or definition.
	System information	Displaying summary table showing total numbers of transcripts and polymorphisms in VaryGene2.
LD-Search	–	Retrieving and displaying LD-bins within the specified region.
GBrowse	–	Displaying genomic region specified with HITs, HIXs and polymorphisms.

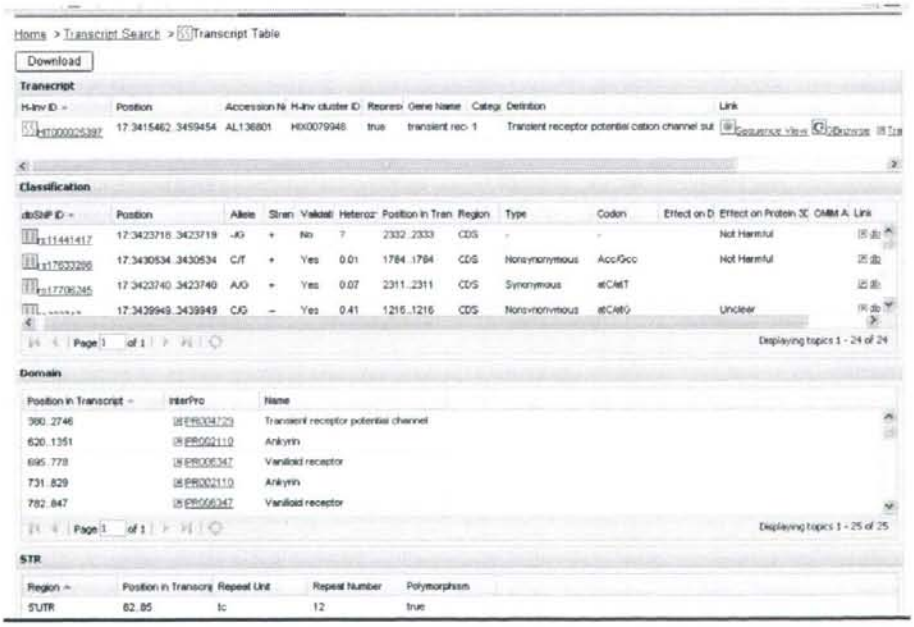

Figure 2. Transcript table containing information on HIT, polymorphism mapped on the HIT (SNP classification, STRs and SARs), and functional domain included in the HIT.

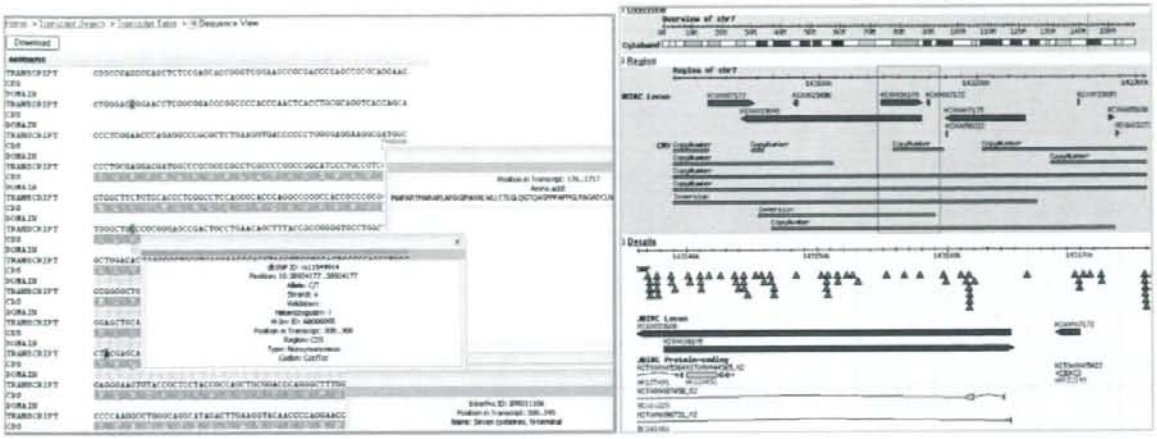

Figure 3. Two graphical viewers in VarySysDB. Left: Sequence View containing polymorphism, domain, sequence of HIT and amino acid sequence. Right: GBrowse showing position of SNP, CNV, HIT and HIX.

The genetic polymorphism data in VarySysDB are also searchable by the features of the HIT on which the polymorphism is located (Transcript Search in Table 2). The HIT features defined in H-InvDB include 'representative transcript', that is, the best HIT to represent a HIX, and 'similarity category' as determined by the level of similarity to known human proteins or InterPro domains (Figure 2).

Sequence View shows the sequence of a HIT and the corresponding amino acid sequence with positional relationship among SNPs, DIPs and functional domains (Figure 3).

VarySysDB also has an LD Search System. This is a subsystem to retrieve LD-bin data distributed within a specified region of the chromosome. The LD-bins offered here are definitive haplotypes that originate from a single sperm, indicating that they are free from errors, which are typically caused by the inference from diploid genotypes (13). This enables users to detect associations among polymorphisms.

GBrowse in VarySysDB can be used to navigate positional relationships among HITs, HIXs and polymorphisms. Since GBrowse is an open-source architecture with various functions, users can conveniently download information from the retrieved region and upload their own data to make comparisons with the information in VarySysDB (Figure 3).

These various web interfaces enable users to extract human genetic polymorphism annotations with user-friendly search systems.

Availability

VarySysDB can be downloaded and freely accessed, with no restriction to academic users only, from http://h-invitational.jp/varygene/. A help document is also available from http://www.h-invitational.jp/hinv/help/Documents/VarySysDB_help.pdf.

ACKNOWLEDGEMENTS

We thank members of the Integrated Database and Systems Biology Team from the Biomedicinal Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST) for their helpful suggestions and cooperation. Especially we thank Akihiro Matsuya, Takuya Habara and Tomohiro Endo for their technical support for constructing and publishing the database. We are also grateful to Drs Shinsei Minoshima (Hamamatsu Univ. School of Medicine), Satoshi Fukuchi (NIG) and Kenshi Hayashi and Koichiro Higasa (Kyusyu Univ.) for effective discussion on this work.

FUNDING

The Ministry of Economy, Trade and Industry of Japan; Japan Biological Informatics Consortium. Funding for open access publication charge: Japan Biological Informatics Consortium.

Conflict of interest statement. None declared.

REFERENCES

1. Maresso,K. and Broeckel,U. (2008) Genotyping platforms for mass-throughput genotyping with SNPs, including human genome-wide scans. In Rao,D.C. and Gu,C.C. (eds), *Advance in Genetics*. Vol. 60, Elsevier, Amsterdam, pp. 107–139.
2. Seng,K.C. and Seng,C.K. (2008) The success of the genome-wide association approach: a brief story of a long struggle. *Eur. J. Hum. Genet.*, **16**, 554–564.
3. 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.
4. Iafrate,A.J., Feuk,L., Rivera,M.N., Listewnik,M.L., Donahoe,P.K., Qi,Y., Scherer,S.W. and Lee,C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.
5. Aishwarya,V. and Sharma,P.C. (2008) UgMicroSatdb: database for mining microsatellites from unigenes. *Nucleic Acids Res.*, **36**, D53–D56.

6. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
7. Tamiya,G., Shinya,M., Imanishi,T., Ikuta,T., Makino,S., Okamoto,K., Furugaki,K., Matsumoto,T., Mano,S., Ando,S. *et al.* (2005) Whole genome association study of rheumatoid arthritis using 27 039 microsatellites. *Hum. Mol. Genet.*, **14**, 2305–2321.
8. Imanishi,T., Itoh,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, e162.
9. Yamasaki,C., Murakami,K., Fujii,Y., Sato,Y., Harada,E., Takeda,J., Taniya,T., Sakate,R., Kikugawa,S., Shimada,M. *et al.* (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res.*, **36**, D793–D799.
10. Takeda,J.-i., Suzuki,Y., Nakao,M., Barrero,R.A., Koyanagi,K.O., Jin,L., Motono,C., Hata,H., Isogai,T., Nagai,K. *et al.* (2006) Large-scale identification and characterization of alternative splicing variants of human gene transcripts using 56 419 completely sequenced and manually annotated full-length cDNAs. *Nucleic Acids Res.*, **34**, 3917–3928.
11. Takeda,J.-i., Suzuki,Y., Nakao,M., Kuroda,T., Sugano,S., Gojobori,T. and Imanishi,T. (2007) H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Res.*, **35**, D104–D109.
12. Yamaguchi-Kabata,Y., Shimada,M.K., Hayakawa,Y., Minoshima,S., Chakraborty,R., Gojobori,T. and Imanishi,T. (2008) Distribution and effects of nonsense polymorphisms in human genes. *PLoS ONE*, **3**, e3393.
13. Higasa,K., Miyatake,K., Kukita,Y., Tahira,T. and Hayashi,K. (2007) D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples. *Nucleic Acids Res.*, **35**, D685–D689.
14. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
15. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
16. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
17. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P., Bucher,P. *et al.* (2002) InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform.*, **3**, 225–235.
18. Kawabata,T., Fukuchi,S., Homma,K., Ota,M., Araki,J., Ito,T., Ichiyoshi,N. and Nishikawa,K. (2002) GTOP: a database of protein structures predicted from genome sequences. *Nucleic Acids Res.*, **30**, 294–298.

Gene 427 (2008) 19–31

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

The evolutionary relationship between gene duplication and alternative splicing

Lihua Jin [a,1], Kirill Kryukov [b,1], Jose C. Clemente [a], Tomoyoshi Komiyama [c], Yoshiyuki Suzuki [a], Tadashi Imanishi [d], Kazuho Ikeo [a], Takashi Gojobori [a,d,*]

[a] Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Yata 1111, Mishima, Shizuoka, 411-8540, Japan
[b] Division of Population Genetics, National Institute of Genetics, Yata 1111, Mishima, Shizuoka, 411-8540, Japan
[c] Department of Clinical Pharmacology, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa, 259-1193, Japan
[d] Integrated Database Group, Biological Information Research Center (BIRC), National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, 135-0064, Japan

ARTICLE INFO

Article history:
Received 29 July 2008
Received in revised form 3 September 2008
Accepted 3 September 2008
Available online 12 September 2008

Keywords:
Gene duplication
Alternative splicing
Multi-loci gene
Single-locus gene
Functional constraint
"Function-sharing model"
"Accelerated splicing model"

ABSTRACT

Gene duplication and alternative splicing (AS) are the two major evolutionary mechanisms that can bring the functional variation by increasing gene diversification. The purpose of this research is to understand the evolutionary relationship between these two different mechanisms, utilizing available data resources. We found the proportion of AS loci and the average number of AS isoforms per locus to be larger in duplicated genes compared to those in singleton genes. However we also found that small gene families have larger proportion of AS loci and larger average number of AS isoforms per locus than large gene families. These results suggest that gene duplication allows for more alternative splicing events to occur on newly duplicated copies than on singletons, probably due to the reduced functional constraint on the duplicates. Smaller average number of AS isoforms in the larger gene families can be explained by the decreased possibility for new useful function to be created via a new alternative splicing event.

1. Introduction

Functional innovation and evolutionary divergence of genes and gene networks contributes to the biological evolution that is accomplished by various kinds of mechanisms, such as nucleotide substitutions, deletion/insertion, gene duplication, alternative splicing (AS), and so on (Krylov et al., 2003). Among these mechanisms, we focused upon the gene duplication and alternative splicing in this study, because they are known as two major mechanisms that can increase gene varieties.

In particular, it has become known for a long time that gene duplication is one of the major mechanisms for increasing the number of genes in the genome (Gu et al., 2002). Once gene duplication takes place, two initially identical copies of the gene immediately start to diverge by accumulating mutations (Wagner 2002). In most cases, such evolutionary diversification led to functional diversification, which eventually increases in varieties of gene function (Senetar and McCann 2005, Katju and Lynch 2006). The alternative outcomes of duplicate genes are generally considered to include the following three functionality options. (1) For non-functionalization, one copy may become silenced by degenerative mutations. (2) For neo-functionalization, one copy may acquire a beneficial new function and become preserved by natural selection with other copy retaining the original function. (3) For sub-functionalization, both copies may become partially compromised by mutation till their total capacity is equal to the ancestral gene (Lynch and Conery 2000).

The other major mechanism of increasing functional variation of genes is alternative splicing (Breitbart et al., 1987; Graveley 2001). Even though the number of genes themselves does not increase, alternative splicing can efficiently amplify the gene variation and relative functional differentiation by producing different transcripts through a splicing mechanism (Bastos et al., 2006, Bugeon et al., 2006). In particular, the full-length cDNA projects of human and mouse have proved that these organisms have a tremendous number of splicing variants (Modrek and Lee 2003, Zavolan et al., 2003). Recently, the completion of the human genome showed that the number of protein-coding genes was only 30,000–40,000 (International Human Genome Sequencing Consortium 2001). This was a kind of surprise because the number of protein-coding genes was much smaller than expected. It is particularly surprising when morphological and physiological complexity of a human body is compared with that of a worm in which the total number of genes is close to 20,000 (The *C. elegans* sequence consortium 1998). Therefore, the capability of alternative splicing in creating diversity in the proteom may hold the key to at least some of the observed complexity of the eukaryotic organisms.

Abbreviation: AS, alternative splicing.

* Corresponding author. Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Yata 1111, Mishima, Shizuoka, 411-8540, Japan. Tel.: +81 55 981 6847; fax: +81 55 981 6848.
E-mail address: tgojobor@genes.nig.ac.jp (T. Gojobori).

[1] These authors contributed equally to this work.

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.09.002

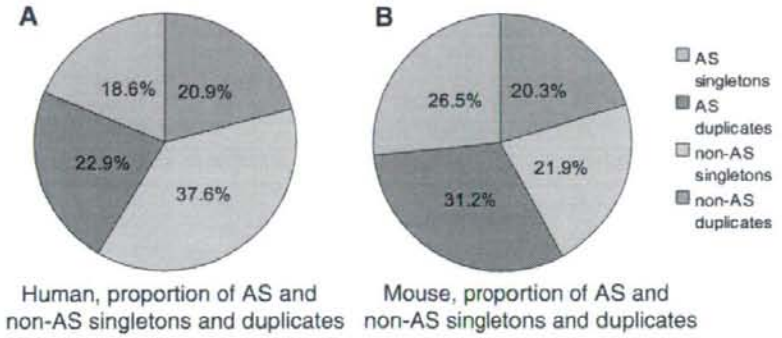

Fig. 1. Proportion of singletons, duplicates, AS and non-AS loci in human and mouse datasets.

Alternative splicing may affect the protein sequence in a variety of ways, where four main patterns can be distinguished. In the first pattern a short and long isoforms are created, with the short one missing a complete exon which is present in the longer isoform (cassette type AS or length difference AS). In the second pattern two isoforms differ by a mutually substituted segment (mutually exclusive type AS or substitutive AS). In the third and fourth patterns a part of one exon is missing on one of the isoforms. When a 5′ region of an exon is removed, it is called "Cryptic acceptor type", and if a 3′ region is removed, it is called "Cryptic donor type". More complex alternative splicing patterns can be created by combination of multiple occurrences of these basic patterns.

It is not clear if the relation exists between the occurrences of various splicing patterns and gene duplication. The cassette type of alternative splicing is more common and is primarily responsible for generation of specific regulators of protein function. On the other hand, the mutually exclusive type of AS is able to create a multitude of functionally distinct protein isoforms (Kondrashov and Koonin 2003). It was reported that some of the tandem duplicated exons could provide the materials for mutually exclusive type AS and some cassette type AS (Kondrashov and Koonin 2001).

Thus, gene duplication and alternative splicing can both contribute to the increase of functional diversity by increasing the variation of gene product. One question is which one of them would be preferred during the evolution. In this study we are interested in general relation between the occurrence of gene duplication and alternative splicing events.

Two models describing such relation were suggested in previous studies. "Independent model" describes AS and duplication as two independent processes which both contribute to the genome complexity (Su et al., 2006). This model predicts no relation between gene family size and the amount of AS. The opposing "Function-sharing model" claims that alternative splicing and duplication can perform the same function interchangeably, so a new function can be created by either one of the two mechanisms. This model predicts a negative correlation between the gene family size and the amount of

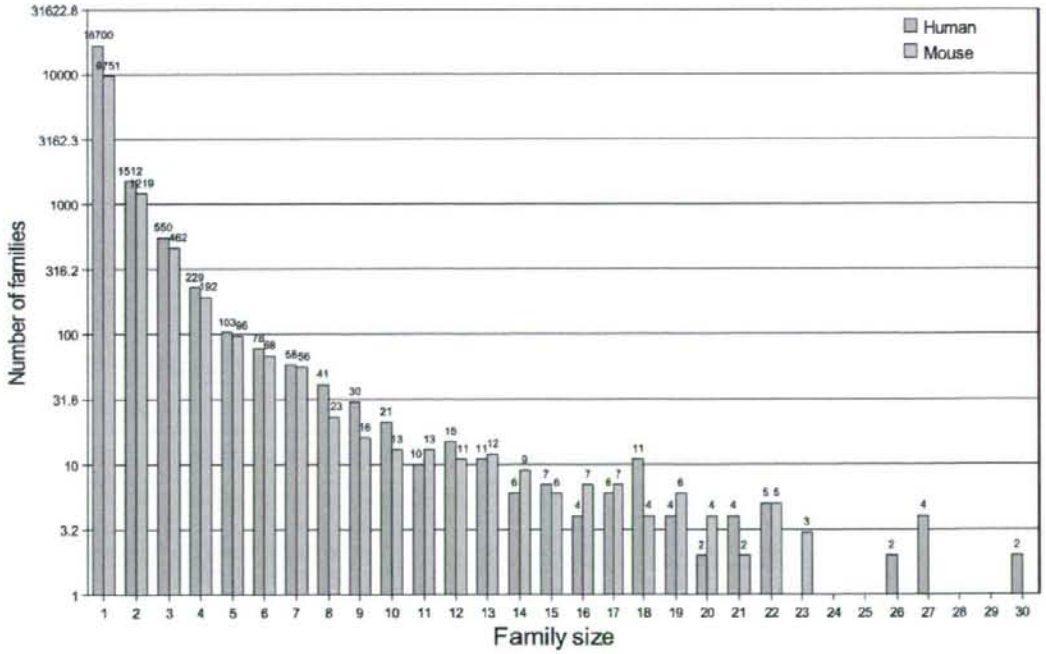

Fig. 2. Number of gene families of each size, in human and mouse. Y axis is in a logarithmic scale. Families of up to 30 members are included in this graph. Families of size 1 are singleton genes.

AS. Recent studies generally supported the "Function-sharing model", as they reported negative correlation between gene family size and number of AS loci or splicing variants (Yu et al., 2003; Kopelman et al., 2005; Su et al., 2006; Talavera et al., 2007).

The purpose of this study is to elucidate the hidden rules between gene duplication and alternative splicing and to suggest possible explanations. We did this by analyzing recent integrated genomic data of human and mouse. All genes were classified into two categories: single-locus genes (singletons — those having just one copy in the genome) and multi-loci genes (those that have closely related paralogs). We analyzed how these two categories are related to alternative splicing by comparing the average number of AS isoforms produced by the genes from those two categories. A correlation between gene duplication and number of AS isoforms was observed and discussed.

2. Materials and methods

2.1. Data resources

Human protein sequences were obtained from the H-InvDB 5.0 database (http://www.h-invitational.jp/, Imanishi et al., 2004). Since our purpose is understanding duplication, we used selection of representative ORF sequences from the database. This selection contained 35,316 sequences originally. We filtered it to remove hypothetical short proteins, which left us with 29,671 protein sequences. Human alternative splicing annotation was obtained from H-InvDB 5.0 database as well. It describes 42,384 individual isoforms for 12,495 AS loci. 6611 loci had AS type annotation.

Mouse sequences and alternative splicing data were obtained from Riken's FANTOM3 database (The FANTOM Consortium et al., 2005). We only had complete ORF database with 110,624 sequences, so we needed to extract a representative ORF for each locus. We did it by choosing the longest ORF sequence out of all ORF mapped to a particular locus, obtaining 20,119 protein sequences. Alternative splicing data contained 122,186 clones mapped to 42,177 loci. 9986 loci had AS type information.

Human gene ontology annotation was obtained from H-Inv DB 5.0 and includes 480,294 clone_id – GO_term pairs. Database of GO terms and paths was downloaded from http://www.geneontology.org/.

2.2. Detecting gene families

We conducted BLAST (Altschul et al., 1997) self-search (search with the same dataset used as query and subject database) on human and mouse representative ORF databases separately to detect homology. We used e-value of 1e−10 in the BLAST search. We combined individual BLAST hits found for the same pair of ORFs in the following way: when two ORFs have multiple BLAST hits with overlapping alignable regions, we selected a non-overlapping combination of those hits with the longest alignable regions out of all possible combinations. We then treated such combination as a single BLAST hit. Next we filtered the BLAST hits to keep only those with at least 30 identities, bitscore of at least 50, aligning at least 50% of the length of both ORFs, and excluding self-hits. Then we defined duplicates, or multi-loci genes, as pairs of ORFs having two-way hits in the filtered set of BLAST hits. Thus all ORFs were classified as singletons or duplicates.

We then identified gene families using single-linkage clustering: Step 1. Initially all genes are in their own families. Step 2. When two genes A and B are found to have a two-way hit, their whole families are merged together. Step 3. Repeat step 2 until no further merging can be done.

2.3. Number of AS isoforms per locus

Number of AS isoforms per locus (NIPL) was obtained from the original AS data, for both human and mouse. We compared the

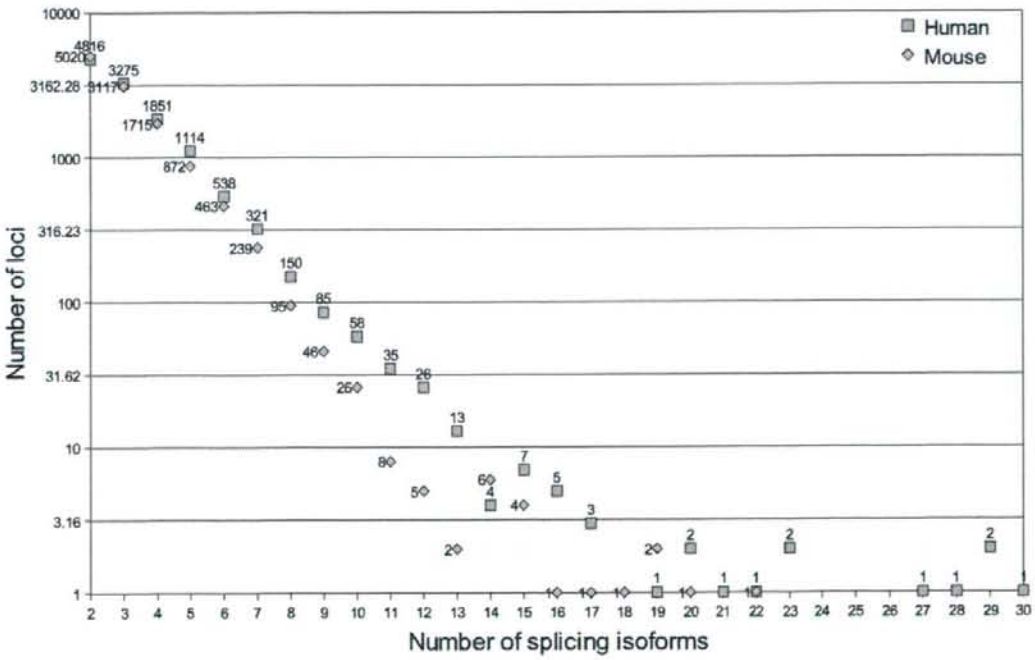

Fig. 3. Number of loci with particular number of splicing isoforms, in human and mouse. Human AS data is obtained from H-Inv DB 5.0, and mouse AS data is obtained from Riken's FANTOM3. Y axis is in a logarithmic scale.

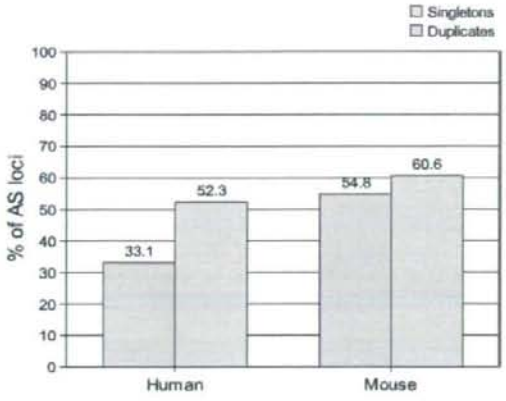

Fig. 4. Proportion of AS loci among singleton loci and among duplicate loci, in human and mouse.

average NIPL between singleton and duplicate loci, in human and mouse. We plotted the average NIPL for gene families of different sizes. We also plotted the average NIPL for different splicing types.

We performed the Mann–Whitney non-parametric statistical test to verify the significance of the difference between the average NIPL of the singleton loci and that of the duplicate loci. Since there were numerous occurrences of identical values in the tested dataset, we used the following corrected formula for standard deviation: $\sigma_u = \mathrm{sqrt}(n_1 * n_2 * (N^3 - N - \Sigma T)/(12 * N * (N-1)))$, where n_1 is the size of the first group, n_2 is the size of the second group, N is $n_1 + n_2$ and ΣT is the sum of $(t^3 - t)$ for each group of identical values. Here t is the number of identical values in such group. Two tests were done: human singleton loci vs human duplicate loci and mouse singleton loci vs mouse duplicate loci.

2.4. Gene ontology

We were interested to see if gene function has any effect on the relation between duplication and alternative splicing. H-Inv DB provides high quality manually curated GO annotation for human genes which allows us to perform this analysis. We extracted a set A of gene ontology terms up to depth 3 (if root term is depth 0), but excluding obsolete terms. Then for each pair of clone id 'c' and GO term 'x' from GO annotation data we find GO term 'y' which is a member of A nearest to 'x'. We then find locus id 'd' corresponding to clone id 'c'. We then assign locus 'd' to the GO term 'y'. We also assign

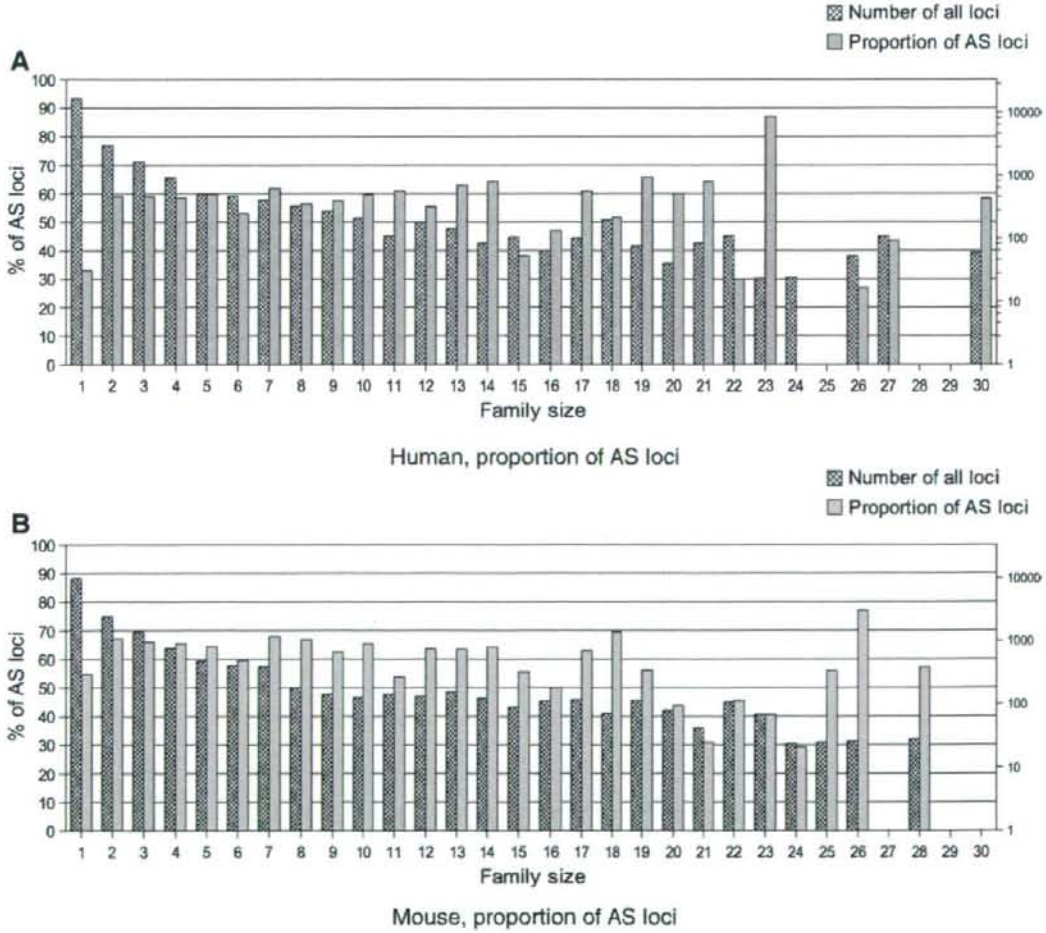

Fig. 5. Proportion of AS loci in gene families of different sizes, for human and mouse. Gene families of up to 30 members are included. Panel A represents human, panel B represents mouse.

locus 'd' to the parent term of 'y'. We then filtered the A set to remove all GO terms that have less than 100 corresponding loci. This left us with 59 GO terms, including the root term. We then counted AS and non-AS singleton and duplicate loci among the loci assigned to each GO term from A. We also counted average number of AS isoforms per locus for loci assigned to each GO term from A.

3. Results

3.1. Number of gene families of different sizes

Using our method described in Section 2.2. we marked 16,700 human ORF and 9751 mouse ORF as singletons (Fig. 1). The remaining 12,971 human ORF and 10,368 mouse ORF were found to have homologous partners, thus forming gene families. Number of gene families of each size is shown in Fig. 2. As expected an L-shape graph can be seen even in the logarithmic scale, due to the existence of a few large gene families. Human and mouse are showing very similar figures. The only big difference is in the number of singletons, which should be explained by the differences in annotation approaches and completeness. This difference however did not influence our subsequent analysis.

3.2. Alternative splicing data

Fig. 1 shows the proportion of AS and non-AS loci for human and mouse. Fig. 3 shows how many AS loci of human and mouse have particular number of splicing isoforms. The linear character of the graph shown with a logarithmic scale indicates exponential decrease of the number of loci with the increase of the number of splicing isoforms. Human and mouse numbers are very close to each other, despite originating from different projects.

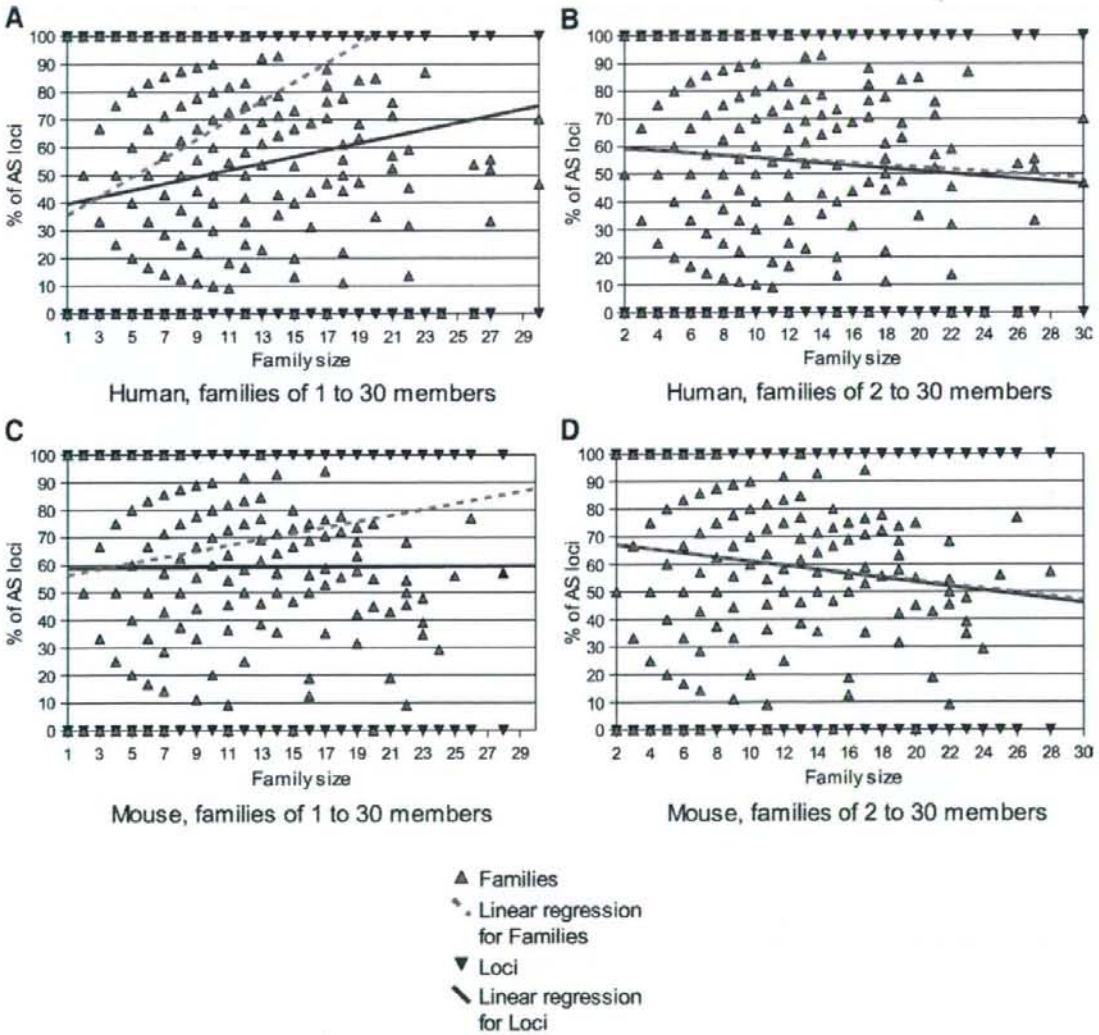

Fig. 6. Proportion of AS loci depending on family size, in human and mouse, using gene families of up to 30 members. Panels A and B display human data, panels C and D show mouse data. Panels A and C include singletons and multi-gene families, while panels B and D show only families of 2 or more genes. Each red triangle represents a family of genes, or several families if they have the same size and same proportion of AS loci. Each blue triangle represents a locus, or multiple loci if they belong to the families of the same size and are all either AS loci or non-AS loci. This sampling allows to show regression lines based on both loci and families.

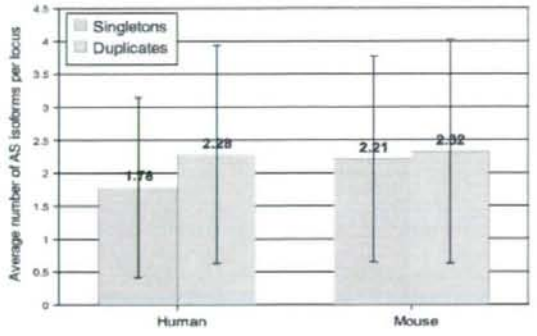

Fig. 7. Average number of AS isoforms per locus for singletons and duplicates, in human and mouse. Non-AS loci are counted as those with 1 isoform. The difference between singletons and duplicates was tested using Mann–Whitney test, which indicated significant difference in both human and mouse (p-value less than 0.001).

3.3. Proportion of AS loci among singletons and duplicates

Fig. 1 shows proportion of AS and non-AS singletons and duplicates in human and mouse datasets. Fig. 4 shows the proportion of AS loci among duplicate and singleton genes, for human and mouse. As can be seen, duplicate genes tend to have more AS loci, in case of both human and mouse. Note that human data shows larger difference.

Fig. 5 shows proportion of AS loci in gene families of different sizes, for human and mouse, up to size 30. Total numbers of loci in families of each size are shown on Fig. 5 too, in logarithmic scale. You can notice that singletons (families of size 1) have noticeably fewer AS loci compared to small families (2–10 members). As the family size increases, proportion of AS loci slightly decreases, though the results get apparently noisy due to the small number of larger families (Fig. 2). This makes it difficult to interpret this graph, so we decided to sample individual loci and families and compute regression to see the overall tendency.

Fig. 6 shows linear regression computed for human and mouse databases. We sampled all individual genes and gene families and arranged them on the X axis according to the family size. The Y axis corresponds to the percentage of the AS loci. This is why all individual loci have the Y coordinate as either 0 or 100. You can see that when both singleton and duplicate genes are included (Fig. 6, panels A and C) linear regression shows positive correlation between the gene family size and proportion of AS loci. On the other hand when we exclude singleton genes and plot only duplicate genes (Fig. 6, panels B and D) a negative correlation can be observed. This means that when singletons are removed from the picture, the proportion of AS loci will decline with the increase of gene family size.

3.4. Average number of AS isoforms per locus

We compared the average number of AS isoforms per locus between singleton and duplicate genes in human and mouse (Fig. 7). The difference appears to be larger in human than in mouse, which we have to attribute to the quality of AS annotation in mouse and human. We tested the significance of the difference in average number of AS isoforms in singletons versus duplicates using Mann–Whitney test, separately in human and mouse. It appears that the difference is significant: we obtained p-value of less than 0.001 in both human and mouse ($z = 32.8$ in human and 6.61 in mouse).

We further investigated how the average number of AS isoforms changes depending on the family size (Fig. 8). A pattern similar to that of Fig. 5 can be observed. Note that the larger families contribute more noisy results due to the smaller sample. We evaluated the tendency using linear regression, with sampling of both individual loci and families (Figs. 9 and 10). We again found slight positive relation when the singletons are included, and negative relation without singletons. This confirms our earlier observation that small families exhibit most abundant alternative splicing.

3.5. Gene ontology

We took general GO terms (up to depth 3) and allocated human loci to the corresponding terms as described in the Materials and methods section. Fig. 11 shows the number of loci corresponding to each of the general GO terms. Fig. 12 compares the proportion of duplicate genes with the proportion of AS loci among all human loci corresponding to each of the general GO terms. Fig. 13 shows the proportion of AS and non-AS singletons and duplicates among human loci corresponding to each of the general GO terms. Fig. 14 shows the average number of AS isoforms per locus and per AS locus for human loci corresponding to each of the general GO terms. The "all" category in Figs. 12, 13 and 14 represents the average result for the whole set of genes with GO annotation.

Interestingly, large variation in relative abundance of AS and gene duplication can be seen in different functional categories. The

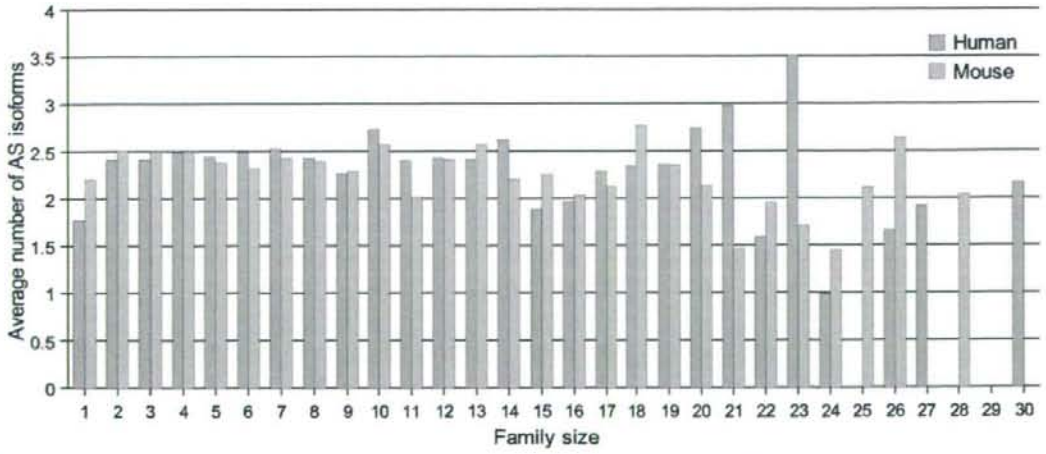

Fig. 8. Average number of AS isoforms per locus for gene families of different sizes, up to size 30, for human and mouse. Singletons are shown as families of size 1. Non-AS loci are counted as those with 1 isoform.

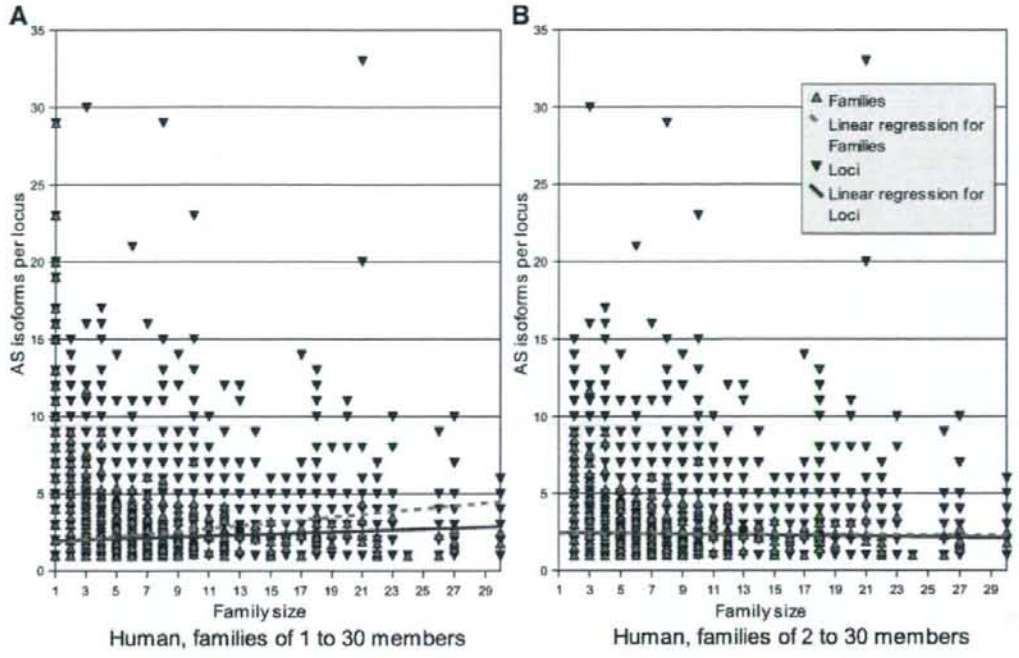

Fig. 9. Number of AS isoforms per locus and average number of AS isoforms per locus for gene families of different sizes, up to size 30, for human. Panel A shows both singletons and duplicates, and panel B shows only duplicates. Each triangle may represent multiple loci or multiple families, if they all have same family size and same average number of AS isoforms per locus.

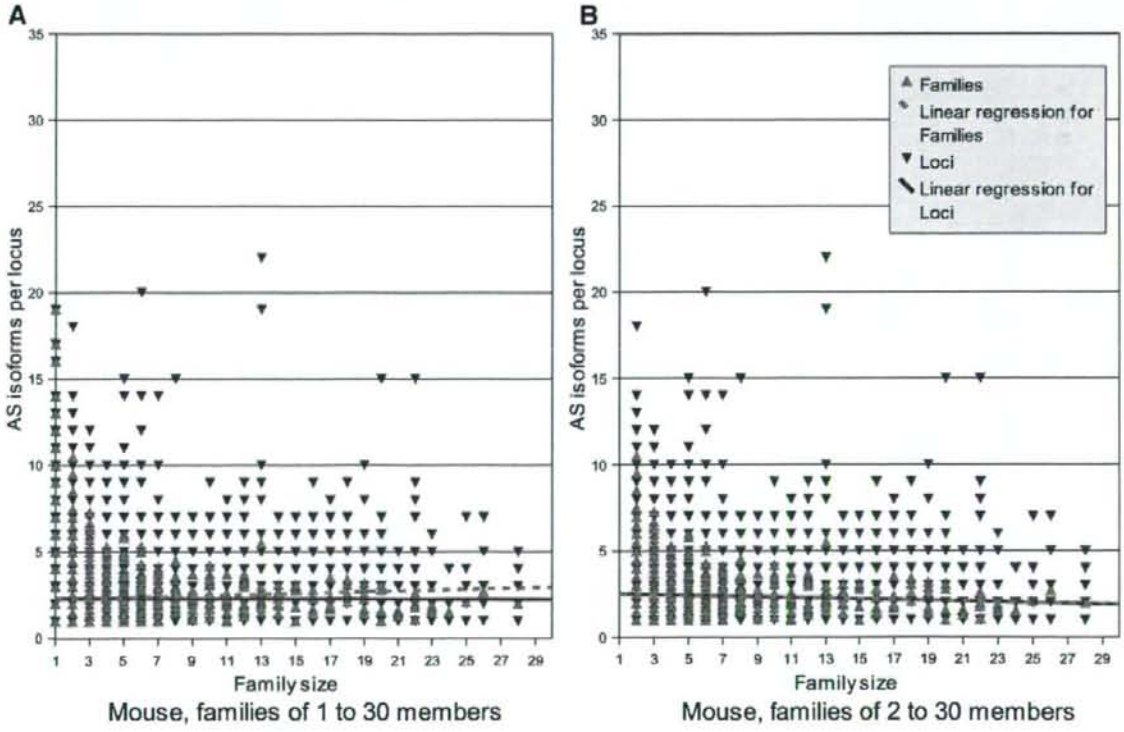

Fig. 10. Number of AS isoforms per locus and average number of AS isoforms per locus for gene families of different sizes, up to size 30, for mouse. Panel A shows both singletons and duplicates, and panel B shows only duplicates. Each triangle may represent multiple loci or multiple families, if they all have same family size and same average number of AS isoforms per locus.

proportion of AS loci and duplicate genes in each category can be compared with that of the root "all" category which represents the average proportion over the whole dataset of GO-annotated genes. In particular, following functional groups are strongly favoring using gene duplication over alternative splicing to provide the necessary diversity: "membrane part" (cellular component), "signal transducer activity" (molecular function), "cell communication" (biological process). Proportion of duplicated genes is about 2 times larger than that of AS loci in these groups. On the other hand the following functional groups are strongly depending on alternative splicing: "GTPase regulator activity", "lipid binding" (molecular function), "ligase activity", "biosynthetic process", "response to stimulus", "regulation of biological process" (biological process) (Figs. 12 and 13).

3.6. AS types

Fig. 15 shows how different AS patterns are distributed between singletons (left column in each pair) and duplicates (right column in each pair). First three pairs of columns represent different splicing site positions. The remaining four pairs of columns represent four AS patterns. It appears that none of these AS features has particular bias towards either singletons or duplicates. This suggests that alternative splicing events of different patterns happen regardless of whether the gene has paralogs or not.

4. Discussion

In this study we used the latest available alternative splicing annotation for human and mouse to investigate the relation between the AS and gene duplication. Both mechanisms contribute to the increase of the genome complexity, therefore it is reasonable to assume that the relation exists between them.

We can say that duplication is a first order mechanism because it is acting directly on the genome sequence. Alternative splicing can be called a "second order" mechanism, because it operates on the different level (transcription control). Our theory is that relative efficiency of these two mechanisms depends on the general complexity of the genome. In the beginning of the genome evolution the genome organization is still relatively simple, in terms of number of genes and complexity of regulatory mechanisms. At that stage the first order mechanisms (such as duplication) might be preferred to effectively increase genome complexity. As genome complexity

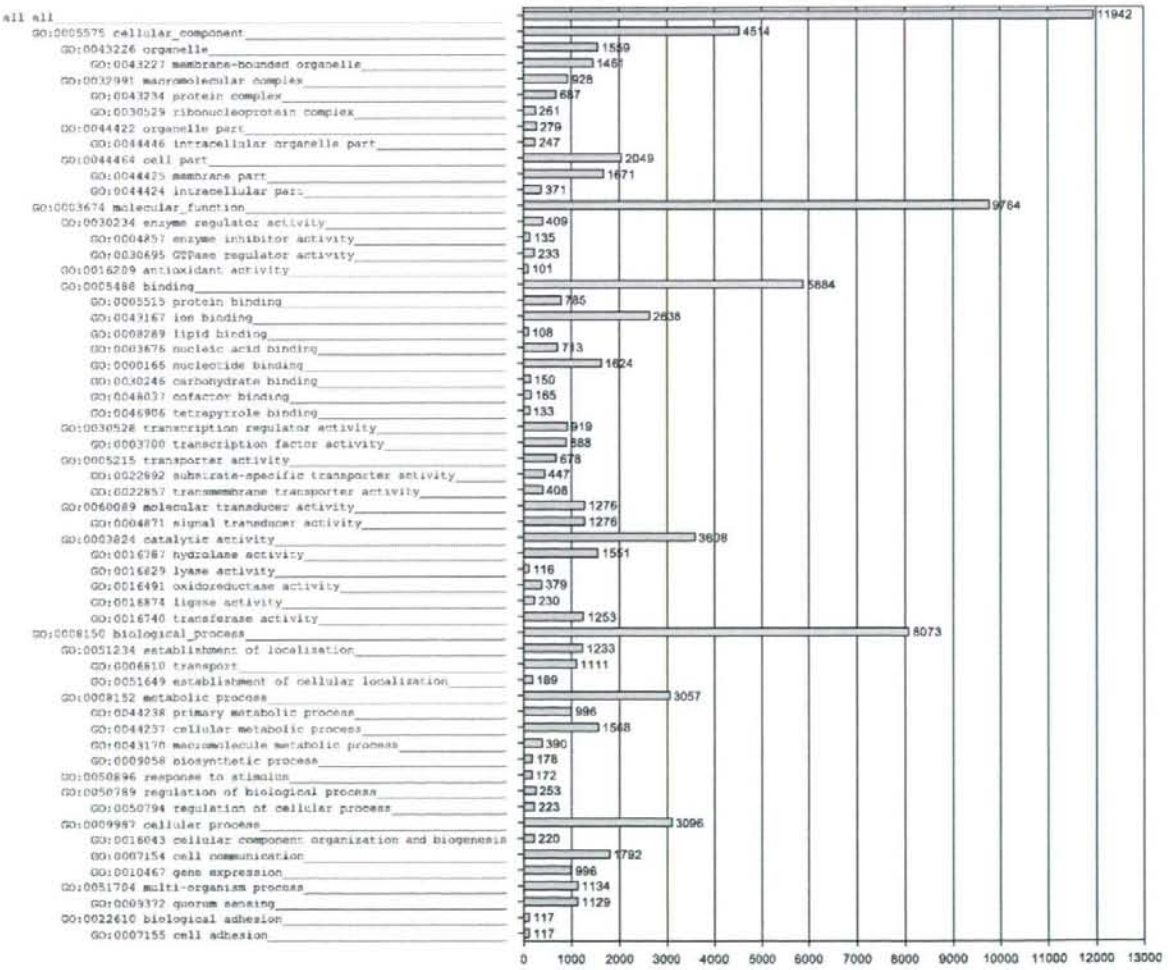

Fig. 11. Number of human loci corresponding to general GO terms. Only GO terms up to depth 3 and with at least 100 genes are shown.

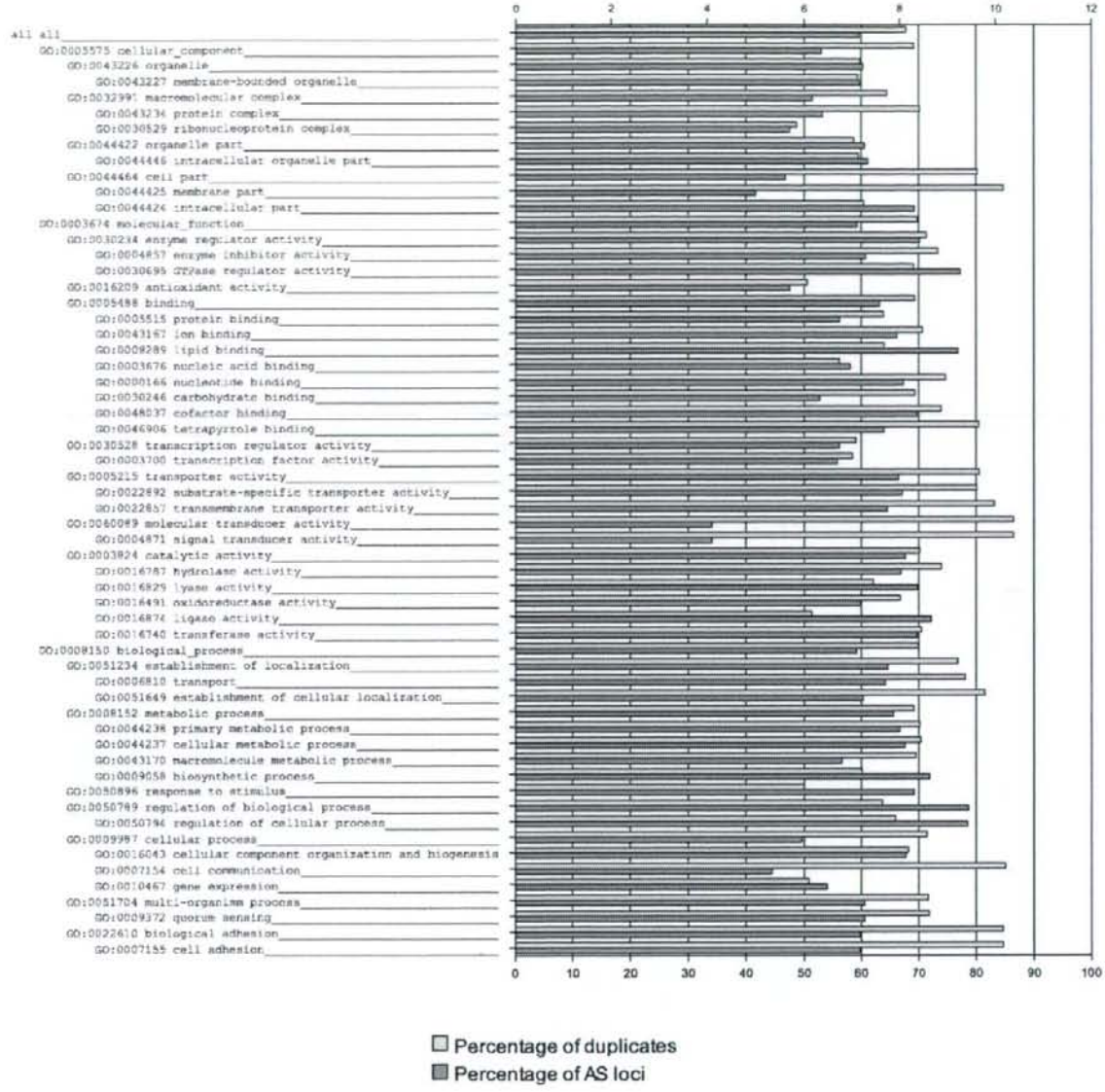

Fig. 12. Comparing percentage of AS loci and percentage of duplicates among human loci corresponding to each general GO term. Only GO terms up to depth 3 and with at least 100 genes are shown.

continued to increase, the transcription/translation regulation network was getting too complex to allow further increase of genome complexity by the first order mechanisms. So the second order regulation mechanisms (such as alternative splicing) were starting to play more important role. Alternative splicing utilizes all complexity created through duplication and further multiplies it by producing variety of products from the same source sequence. Duplication may not only provide the ingredients for the alternative splicing, but also confer some systematical influence to the occurrence and evolution of alternative splicing.

Previous reports tended to support the "Function-sharing model" over the "Independent model" of the relation between the AS and GD (Kopelman et al., 2005; Su et al., 2006; Talavera et al., 2007). Consistently with these reports, we observed a significant decrease of the AS loci proportion and average number of AS variants per locus with the increase of family size. However we also registered increase of alternative splicing in small families compared to singleton genes, which is contrary to the prediction of the "Function-sharing model".

It has been reported that evolutionary rate may increase after gene duplication (Lynch and Force 2000; Kondrashov et al., 2002). Therefore we propose the following "Accelerated AS model". When a singleton gene is first duplicated, a functional constraint on each copy is relaxed, so it allows accelerated evolution of one or both copies. This results in the increased possibility for sub-functionalization or neo-functionalization via creating splicing variants. The effect of the decreased functional constraint is strongest when a gene is undergoing its first duplication event, and is decreasing drastically with

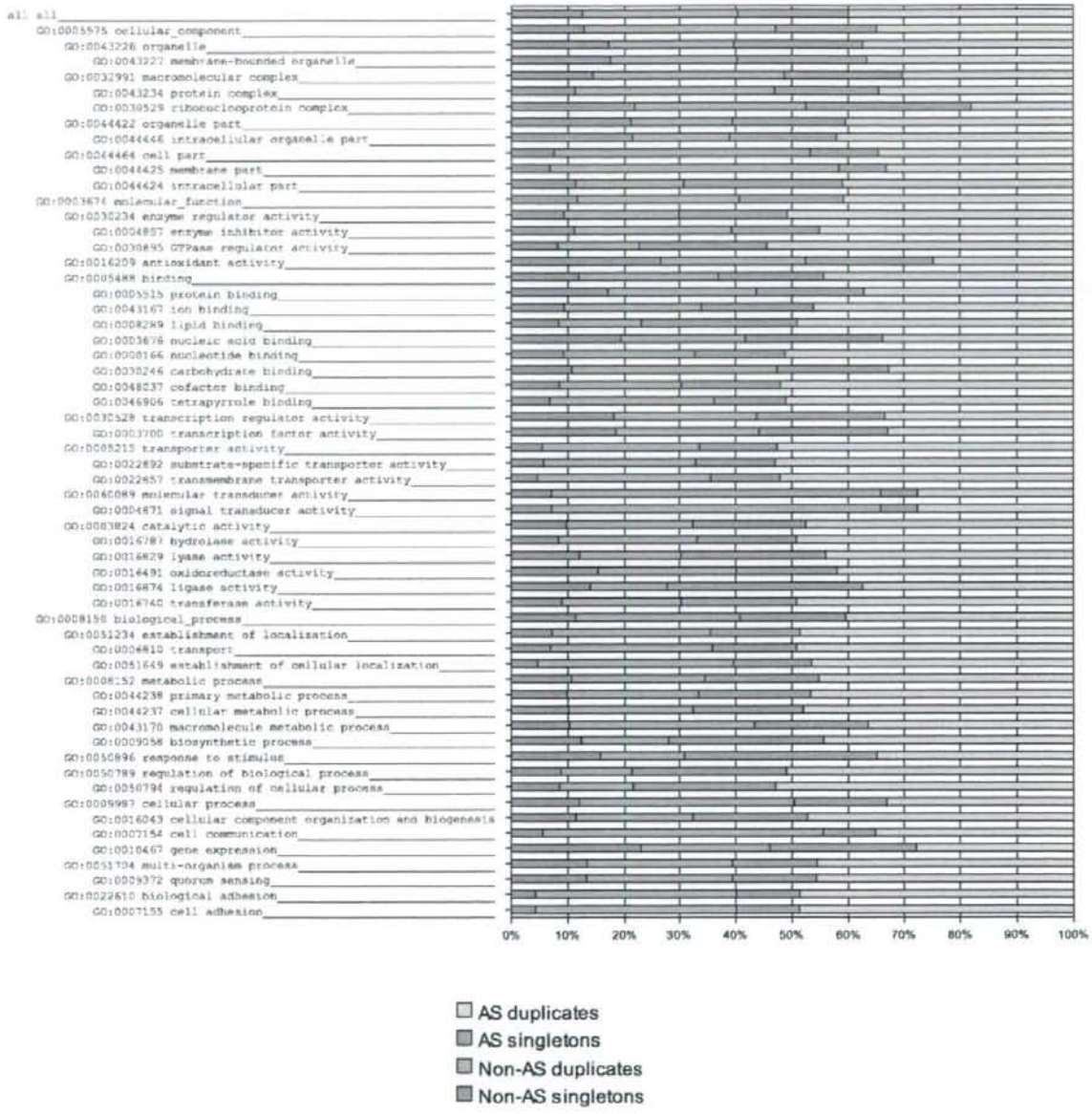

Fig. 13. Proportion of AS singletons, AS duplicates, non-AS singletons and non-AS duplicates among human genes categorized by major GO terms. Only GO terms up to depth 3 and with at least 100 genes are shown.

subsequent duplications. On the other hand, additional duplications may contribute to creating a new or sub-function, reducing the possibility of creating a new AS variant to achieve the same function. So the possibility for new useful function to be obtained by an alternative splicing event is highest when two or three paralogous copies of the gene exist in the genome. It is lower when only one copy of the gene exists in the genome, due to the strong functional constraint. It also gets lower when additional copies of the gene appear in the genome, due to the decreased possibility for new useful function to be created via a new alternative splicing event.

Our model provides more detailed insight into the evolutionary process that involves both gene duplication and alternative splicing events. During the evolution generally the more simple and efficient mechanisms of achieving a certain function are preferred over the complex mechanisms. This is why relatively simple organisms don't exhibit as much alternative splicing as more complex ones. The reason is that new functions happen to be more easily achieved by a simpler means of gene duplication. Gene duplication does not require complex transcriptional regulation to happen, so it can occur abundantly even in the most simple organisms.

As an organism gets more complex, the gene regulation network also becomes more complex. This increased complexity means that it may be more difficult for a newly duplicated gene copy to produce a new meaningful function, because the new function may require

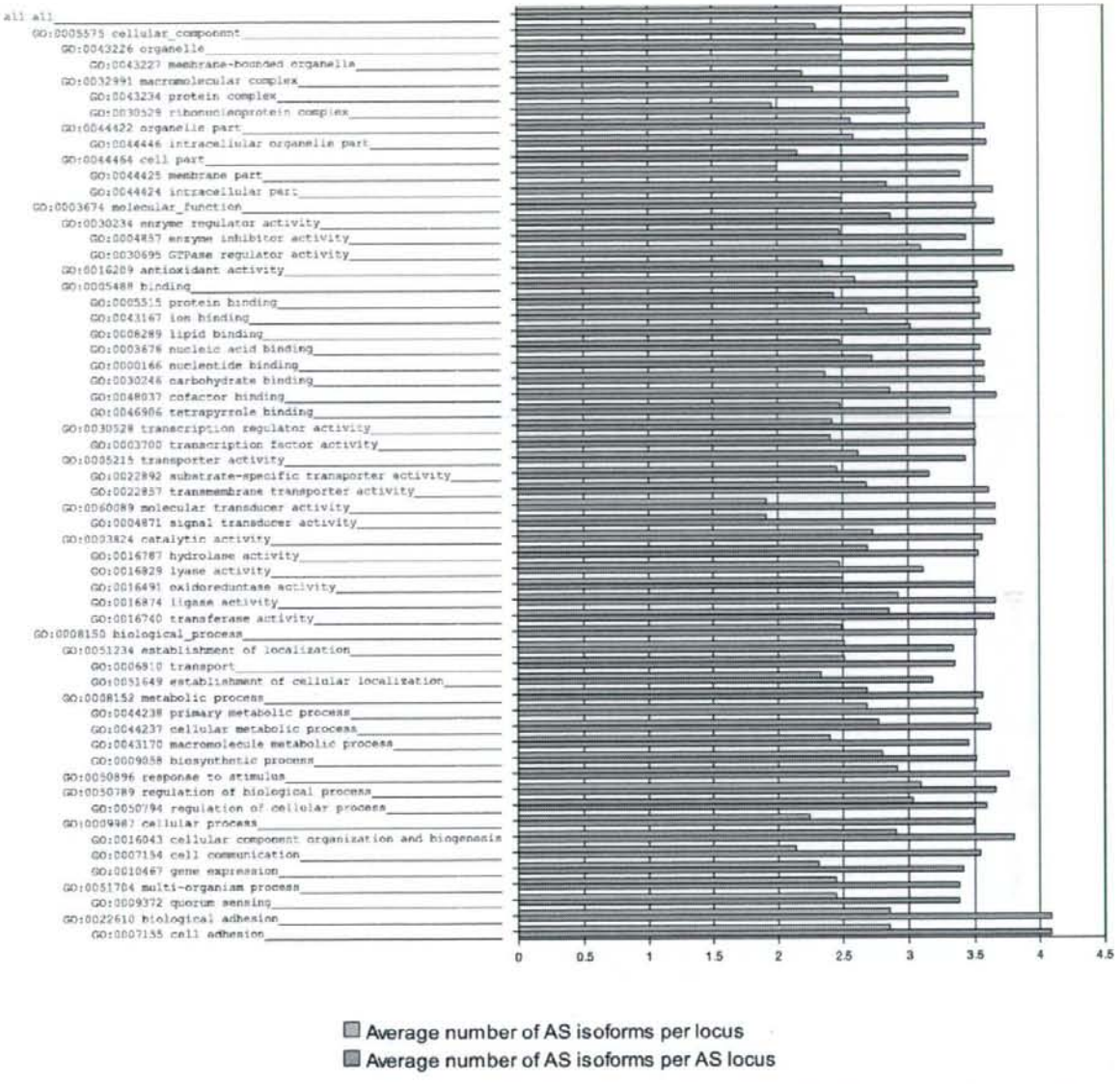

Fig. 14. Average number of AS isoforms per locus and per AS locus for human genes categorized by major GO terms. Only GO terms up to depth 3 and with at least 100 genes are shown.

totally differently regulated transcription. So the alternative splicing as a mean of achieving a novel function becomes more and more useful, as it involves different regulation mechanisms. However creation of alternative splicing isoform does not equally easily happen to any gene. This even may require some changes in the gene itself and/or in the regulatory region. When a single copy of a gene exists in a genome, the functional constraint may be strong enough to prevent much of the mutations, reducing the probability of creating an alternative splicing. Then at some point gene duplication happens, leaving two copies of the same gene. Only one copy is required for the original function, so the other one (or even both copies) begins to accumulate mutations much more quickly than a single gene copy did. This leads to the increased possibility to form an alternative splicing. This is why gene families with two members show the larger proportion of alternatively spliced sites compared to the singleton genes.

Some of the gene copies may become non-functional and degrade completely, but the surviving functional copies may duplicate again producing a gene family of three or more members. When this happens, naturally the functional constraint is reduced even more than after the first duplication. However this effect is not nearly as dramatic as the first duplication event. If an alternatively spliced gene gets duplicated, the resulting copies are initially alternatively spliced too. This creates an increased redundancy, which is larger than when a non-AS gene is duplicated. Therefore some of the copies begin to accumulate mutations that may result in some of the AS variants of this gene copy disappearing or becoming non-functional. It happens easily because the other copies of this gene carry on function lost by this

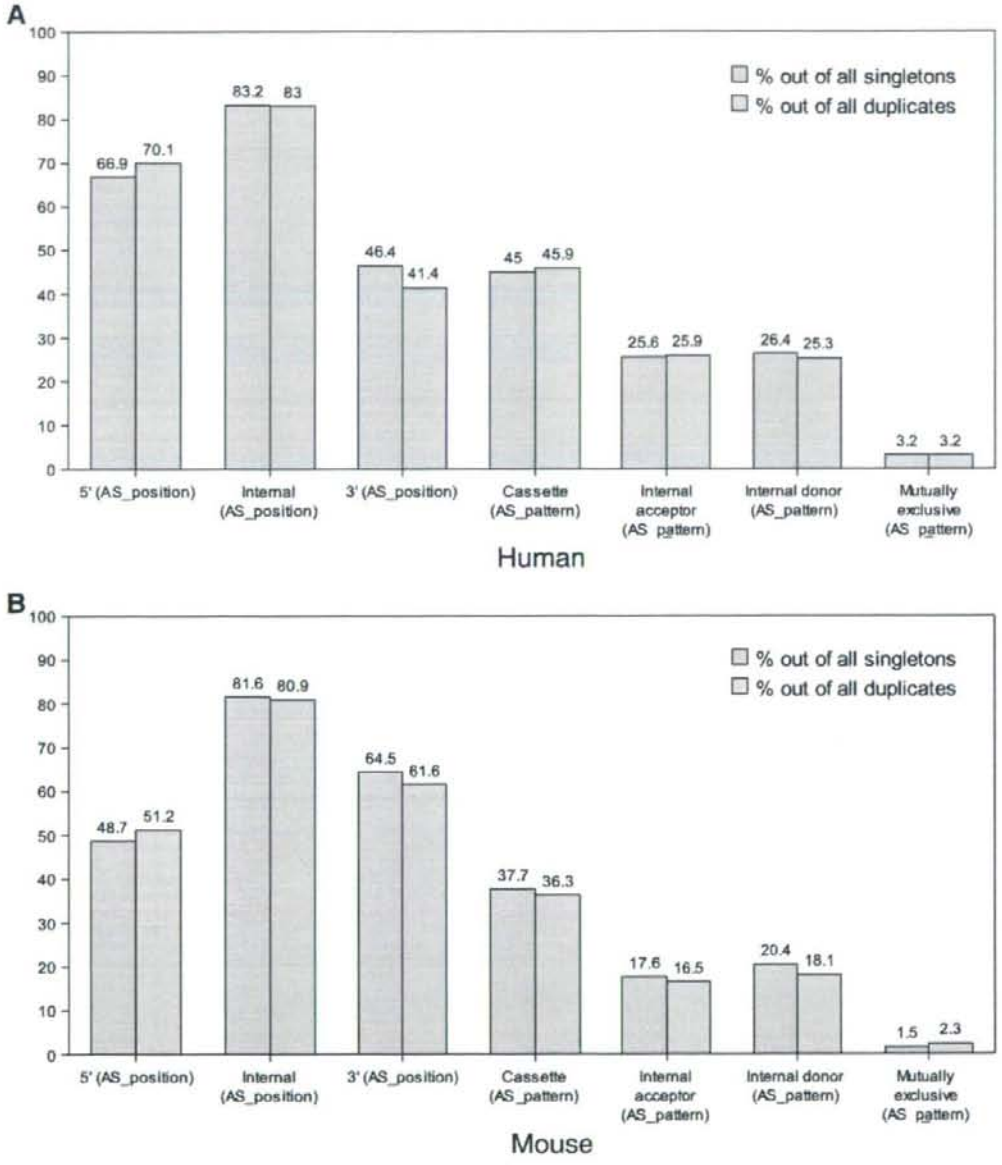

Fig. 15. Percentage of AS loci with particular splicing site position or splicing pattern among all singletons and duplicate genes, for human and mouse.

copy. Thus the total number of alternatively spliced copies in this gene family will decrease, as well as the average number of AS isoforms per gene copy. This is again supported by our observed results.

In recent years there have been remarkable progress in developing whole genome integrated databases which include AS annotation. We used AS data from the H-Inv integrated database for human and from FANTOM3 database for mouse. In the H-Inv jamboree both manual and computational methods were employed. These two approaches have different advantages: Computational methods help to reduce human error, and manual annotation protects from detecting spurious AS due to the mapping errors or ambiguities. For these reasons both strategies were utilized to maximize the accuracy of the analysis.

In this study the relationship between gene duplication and alternative spicing was analyzed using the latest available genome annotation databases. The evidences for the effect of gene duplication on the occurrence of alternative splicing events provide a novel viewpoint for understanding the mechanisms of functional innovation and genome divergence along the biological evolution.

Acknowledgements

We greatly appreciate Jun-ichi Takeda, Chisato Yamazaki and Tanino Motohiko for the technical assistance. We give special thanks to Roberto Barrero and Yoshio Tateno for their technical advices. This

research was, in part, supported by Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan and by BIRC (Biological Information Research Center) at AIST (National Institute for Advanced Industrial Studies and Technology).

References

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bastos, E., Avila, S., Cravador, A., Renaville, R., Guedes-Pinto, H., Castrillo, J.L., 2006. Identification and characterization of four splicing variants of ovine POU1F1 gene. Gene 382, 12–19.

Breitbart, R.E., Andreadis, A., Nadal-Ginard, B., 1987. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu. Rev. Biochem. 56, 467–495.

Bugeon, L., Wong, K.K., Rankin, A.M., Hargreaves, R.E., Dallman, M.J., 2006. A negative regulatory role in mouse cardiac transplantation for a splice variant of CD80. Transplantation 82 (10), 1334–1341.

Graveley, B.R., 2001. Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17, 100–107.

Gu, Z., Cavalcanti, A., Chen, F., Bouman, P., Li, W., 2002. Extent of gene duplication in the genomes o *Drosophila*, nematode, and yeast. Mol. Biol. Evol. 19 (3), 256–262.

Imanishi, et al., 2004. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLOS biology 2, 1–21.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Katju, V., Lynch, M., 2006. On the formation of novel genes by duplication in the Caenorhabditis elegans genome. Mol. Biol. Evol. 23 (5), 1056–1067.

Kondrashov, F.A., Koonin, E.V., 2001. Origin of alternative splicing by tandem exon duplication. Hum. Mol. Genet. 10 (23), 2661–2669.

Kondrashov, F.A., Koonin, E.V., 2003. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. Trends Genet. 19 (3), 115–119.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Selection in the evolution of gene duplication. Genome Biol. 3 research0008.

Kopelman, N.M., Lancet, D., Yanai, I., 2005. Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat. Genet. 37 (6), 588–589.

Krylov, D.M., Wolf, Y.I., Rogozin, I.B., Koonin, E.V., 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13, 2229–2235.

Lynch, M., Force, A., 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154, 459–473.

Lynch, M., Conery, J.S., 2000. The evolutionary fate and consequence of duplicate genes. Science 290, 1151–1155.

Modrek, B., Lee, C.J., 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat. Genet. 34 (2), 177–180.

Senetar, M.A., McCann, R.O., 2005. Gene duplication and functional divergence during evolution of the cytoskeletal linker protein talin. Gene 362, 141–152.

Su, Z., Wang, J., Yu, J., Huang, X., Gu, X., 2006. Evolution of alternative splicing after gene duplication. Genome Res. 16, 182–189.

Talavera, D., Vogel, C., Orozco, M., Teichmann, S.A., de la Cruz, X., 2007. The (in) dependence of alternative splicing and gene duplication. PLoS Comput. Biol. 3 (3), e33.

The C. elegans sequence consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012–2018.

The FANTOM Consortium, RIKEN Genome Research Exploration Group and Genome Science Group, 2005. The transcriptional landscape of the mammalian genome. Science 309, 1559–1563.

Wagner, A., 2002. Asymmetric functional divergence of duplicate genes in yeast. Mol. Biol. Evol. 19 (10), 1760–1768.

Yu, W.P., Brenner, S., Venkatesh, B., 2003. Duplication, degeneration and sub-functionalization of the nested synapsin–Timp genes in Fugu. Trends Genet. 19, 180–183.

Zavolan, M., et al., 2003. Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. Genome Res. 13, 1290–1300.

Distribution and Effects of Nonsense Polymorphisms in Human Genes

Yumi Yamaguchi-Kabata[1¤a], Makoto K. Shimada[1,2¤b], Yosuke Hayakawa[1,2], Shinsei Minoshima[3], Ranajit Chakraborty[4], Takashi Gojobori[1,5], Tadashi Imanishi[1]*

1 Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan, **2** Japan Biological Information Research Center, Japan Biological Informatics Consortium, Tokyo, Japan, **3** Hamamatsu University School of Medicine, Hamamatsu, Shizuoka, Japan, **4** Center for Genome Information, University of Cincinnati, Cincinnati, Ohio, United States of America, **5** Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Shizuoka, Japan

Abstract

Background: A great amount of data has been accumulated on genetic variations in the human genome, but we still do not know much about how the genetic variations affect gene function. In particular, little is known about the distribution of nonsense polymorphisms in human genes despite their drastic effects on gene products.

Methodology/Principal Findings: To detect polymorphisms affecting gene function, we analyzed all publicly available polymorphisms in a database for single nucleotide polymorphisms (dbSNP build 125) located in the exons of 36,712 known and predicted protein-coding genes that were defined in an annotation project of all human genes and transcripts (H-InvDB ver3.8). We found a total of 252,555 single nucleotide polymorphisms (SNPs) and 8,479 insertion and deletions in the representative transcripts in these genes. The SNPs located in ORFs include 40,484 synonymous and 53,754 nonsynonymous SNPs, and 1,258 SNPs that were predicted to be nonsense SNPs or read-through SNPs. We estimated the density of nonsense SNPs to be 0.85×10^{-3} per site, which is lower than that of nonsynonymous SNPs (2.1×10^{-3} per site). On average, nonsense SNPs were located 250 codons upstream of the original termination codon, with the substitution occurring most frequently at the first codon position. Of the nonsense SNPs, 581 were predicted to cause nonsense-mediated decay (NMD) of transcripts that would prevent translation. We found that nonsense SNPs causing NMD were more common in genes involving kinase activity and transport. The remaining 602 nonsense SNPs are predicted to produce truncated polypeptides, with an average truncation of 75 amino acids. In addition, 110 read-through SNPs at termination codons were detected.

Conclusion/Significance: Our comprehensive exploration of nonsense polymorphisms showed that nonsense SNPs exist at a lower density than nonsynonymous SNPs, suggesting that nonsense mutations have more severe effects than amino acid changes. The correspondence of nonsense SNPs to known pathological variants suggests that phenotypic effects of nonsense SNPs have been reported for only a small fraction of nonsense SNPs, and that nonsense SNPs causing NMD are more likely to be involved in phenotypic variations. These nonsense SNPs may include pathological variants that have not yet been reported. These data are available from Transcript View of H-InvDB and VarySysDB (http://h-invitational.jp/varygene/).

Citation: Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, et al. (2008) Distribution and Effects of Nonsense Polymorphisms in Human Genes. PLoS ONE 3(10): e3393. doi:10.1371/journal.pone.0003393

Editor: Alan Christoffels, University of Western Cape, South Africa

Received March 6, 2008; **Accepted** September 3, 2008; **Published** October 14, 2008

Funding: This research is financially supported by the Ministry of Economy, Trade and Industry of Japan (METI), the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) and the Japan Biological Informatics Consortium (JBIC).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: t.imanishi@aist.go.jp

¤a Current address: Center for Genomic Medicine, RIKEN, Yokohama City, Kanagawa, Japan
¤b Current address: Institute for Comprehensive Medical Science, Fujita Health University, Toyoake, Aichi, Japan

Introduction

Genetic variations in the human genome are maintained by a balance of mutation, selection and random genetic drift. Some of the polymorphisms cause phenotypic variations and diseases. Therefore, many studies have attempted to identify causative variants of genetic diseases and the relationships between genetic variations and phenotypic effects. Genetic variations within linked loci are inherited to the same gamete. Based on the linkage of genetic variations, loci that contain disease-causing genes have been mapped by using polymorphic markers. At present, about 14 million clusters of genetic polymorphisms have been identified in the human genome [1]. On average, two haploid genomes are estimated to differ by one single nucleotide polymorphism (SNP) in every 1200–1500 bp [2]. SNPs have been recently used to conduct genome-wide association studies to find genomic regions that are susceptible to diseases and phenotypic variations [3,4,5,6]. In this approach, usually, causative polymorphisms for diseases or phenotypic variations are identified after the identification of susceptible genomic regions by using SNP markers. Such SNPs are