called landmark SNPs, and the indirect relationships between polymorphisms and phenotypic variations were examined to identify genomic regions where causative genes are located.

Another approach in finding pathological variants is to extract polymorphisms that alter amino acids in functional genes or affect gene expression or splicing, using a comprehensive set of functional elements of the human genome. Several studies have analyzed nonsynonymous SNPs to predict pathological variants [7,8,9, 10,11,12,13,14]. A large number of nonsynonymous SNPs also have been examined for associations with diseases[15,16].

Although many pathological mutations have been identified [17,18], the number of such variants is small compared to the number of known polymorphisms, and it is still unclear which polymorphisms have biological effects. In a study of consanguineous marriage [19], it was estimated that each person has deleterious alleles that are equivalent to a few lethal genes. Gene-centric SNP surveys have shown that the ratio of nonsynonymous to synonymous SNPs is significantly higher in the low frequency class than in the common frequency class [20,21,22]. These results suggest that a large fraction of the low frequency nonsynonymous SNPs are deleterious. To understand the molecular basis of the effects of human genetic variations on phenotypic variations, a prediction analysis of possible effects of polymorphisms on gene function in all human genes appears to be needed.

In this study, to detect polymorphisms affecting gene function, we analyzed all publicly available polymorphisms in the Single Nucleotide Polymorphism Database (dbSNP) (build 125) in the exons of all 36,712 protein-coding genes that were defined in an annotation project of all human genes and transcripts (H-InvDB ver3.8)[23,24]. In summary with representative transcripts (one transcript from one gene), we detected 53,754 nonsynonymous SNPs and 1,417 SNPs causing changes between amino acids and stop codons. Among possible point mutations in ORFs, nonsense mutations cause the most drastic changes of gene products. In fact, several reports have shown that nonsense mutations cause genetic diseases [25,26,27,28]. Truncation of a polypeptide by a premature termination codon causes a drastic change in the gene product. Furthermore, it is known that a nonsense mutation can cause decay of mRNA resulting in the absence of the gene product. This process, called 'nonsense-mediated decay (NMD)' limits the synthesis of abnormal proteins[29,30,31]. On the other hand, the loss of a termination codon in a transcript also appears to cause decay of mRNA (referred to as non-stop decay) and thus to prevent translation[32,33]. In spite of the severe effects of nonsense mutations, the distribution of nonsense SNPs in human genes is little understood. In this study, we examined the density of nonsense SNPs in human genes, and showed that nonsense SNPs exist at a lower density than nonsynonymous SNPs, possibly due to the more severe effects of premature stop codons than amino acid changes. About a half of nonsense SNPs are predicted to cause NMD. The correspondence between known pathological variants and nonsense SNPs suggests that nonsense SNPs causing NMD are more likely to be involved in phenotypic variations.

Results

Selection and classification of polymorphisms in exon regions

We analyzed 9,235,997 polymorphisms (dbSNP build 125) in the human genome with exon positions and predicted ORFs that were revealed in our annotation project of human genes (H-InvDB) (Figure 1). In all of the 36,712 protein-coding loci in the genome, we detected 252,555 SNPs and 8,479 insertions and deletions (indels) that exist in exon regions of the representative transcript (one transcript from one gene) (Table 1). The polymorphisms in the exon regions were further classified according to the predicted ORFs. We detected 96,164 SNPs within the ORFs, 51,881 SNPs in the 5′UTR regions and 104,510 SNPs in the 3′UTR regions. Among the SNPs in the ORFs, 40,484 were synonymous and 53,754 were nonsynonymous (Further analyses of nonsynonymous SNPs are described in Results S1.). Most of the indels were detected in the UTR regions. The ORF regions contained 1,258 SNPs that cause changes between amino acids and stop codons (Table S1). Of the 1,258 SNPs, 1,183 SNPs were regarded as nonsense SNPs, while 75 were found to have stop codons as ancestral alleles. We also detected 247 SNPs at termination codon sites, 88 of which were synonymous. The remaining 159 SNPs were changes between stop codons and amino acids. After checking ancestral alleles, 110 of the 159 SNPs were inferred to be read-through SNPs, while the other 49 were inferred to changes to stop codons.

Distribution of polymorphisms in exon regions

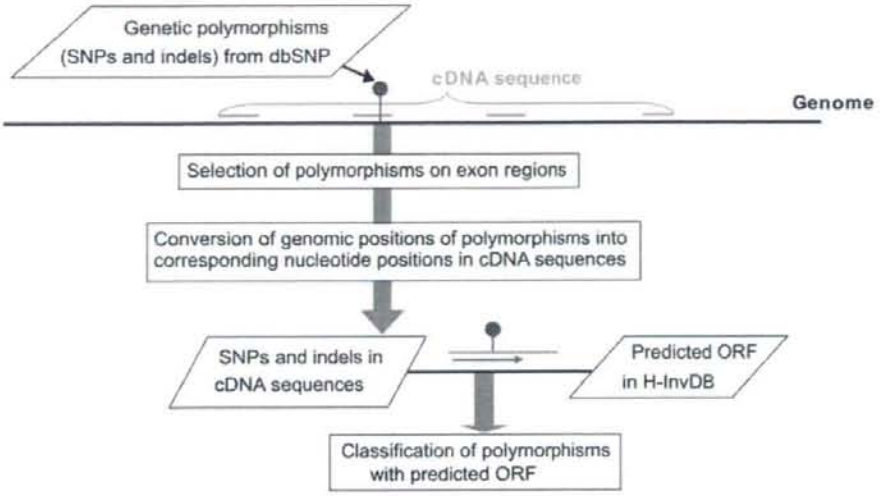
Densities of polymorphisms were estimated for 23,717 genes whose functions are clearly defined or suggested (similarity category I–III, see Materials and Methods) and genes annotated as conserved hypothetical proteins (similarity category IV). To estimate the densities of SNPs for synonymous, nonsynonymous and nonsense SNPs in the ORFs, we calculated the numbers of potential nucleotide sites for synonymous, nonsynonymous and nonsense mutations in the coding regions. The fractions of sites (%) in the coding regions for synonymous, nonsynonymous, and nonsense mutations were estimated to be 28.5%, 68.1%, and 3.4%, respectively. Of the three types of SNPs, synonymous SNPs had the highest density, 4.1×10^{-3} per synonymous site, in ORFs (Table 2). The estimated density of nonsynonymous SNP was 2.1×10^{-3} per site (Table 2). The lower density of nonsynonymous SNPs compared with synonymous SNPs (51%) is due to the functional constraint of amino acid changes, and is in agreement with previous studies [20,22,34]. However, the ratio of the numbers of nonsynonymous SNPs to synonymous SNPs per site is higher in this study compared with previous studies (32–34%) [20,21,22], which they focused on specific populations. The higher ratio of nonsynonymous SNPs in this study may be due to the fact that our study is based on pooled data from various populations world wide. This study includes many nonsynonymous SNPs that exist in relatively lower frequencies and are likely to be more population-specific in comparison to synonymous SNPs [20].

Among random nucleotide mutations in ORFs, 3.4% would be expected to be nonsense mutations; however, the distribution of nonsense SNPs has not been evaluated or reported. The density of nonsense SNPs was estimated to be 0.85×10^{-3} per site (Table 3), which is only 21% of the density of synonymous SNPs, and 40% of the density of nonsynonymous SNPs. The reason for the lowest density of nonsense SNPs may be that premature stop codons have more severe effects than amino acid changes.

In the exons of the 36,712 loci, 8479 indels were detected, and 1,532 of them were found in ORFs. Among the latter, 1,331 are expected to cause frame shifts, resulting in drastic changes of proteins. The density of indels in ORFs was much lower than in the UTR regions (Table 4). The lower density of indels in the 5′UTRs than in the 3′UTRs suggests that functional constraint for insertions and deletions is higher in the 5′UTR regions than in the 3′UTR regions.

Nonsense SNPs

We examined the patterns and the positions of the nonsense SNPs. There are 23 possible ways to change codons into stop



Single Nucleotide Polymorphism (SNP) and indel					
Location	Variation	dbSNP ID	Strand	CDS/UTR	Translation
115 .. 115	T/C	rs17103291	-	CDS	Synonymous[Asp28Asp]
136 .. 136	G/A	rs555367	-	CDS	Synonymous[Val35Val]
234 .. 234	T/C	rs693001	-	CDS	Nonsynonymous[Ile68Thr]
404 .. 404	C/T	rs497116	+	CDS	AA-STOP[Arg125*]
743 .. 743	A/G	rs647039	-	CDS	Nonsynonymous[Ser238Gly]
849 .. 849	A/G	rs513131	+	CDS	Nonsynonymous[His273Arg]

Figure 1. Analysis of polymorphisms with gene structure. Top: Scheme of analysis pipeline of polymorphisms with gene structure. Bottom: Screen shots taken from 'Transcript View' in H-InvDB that show classified SNPs and their positions (blue bars) in the *CASP12* gene.
doi:10.1371/journal.pone.0003393.g001

codons (nine, seven and seven for the first, second and third positions, respectively), and all 23 were found (Table 5). Nonsense SNPs were more frequent at the first codon position than at the second and third positions ($p<0.005$, chi-square test). The most frequent type of nonsense mutation is the change from CGA to TGA (Table 5), which is a transitional change at CpG mutation hotspots [35]. However, it is notable that there were frequent transversional mutations such as GAA to TAA and GAG to TAG. Our analyses of nonsense polymorphisms revealed that changes between hydrophilic amino acids and termination codons by nucleotide changes at the first codon positions were very frequent.

We examined the positions of 1,183 nonsense polymorphisms in the coding regions. On average, nonsense SNPs were located at 250 codons upstream of the original termination codons. To predict whether a nonsense mutation causes nonsense-mediated decay (NMD) of mRNA, we examined the locations of nonsense SNPs in the exon-intron structure of the genes (Table 6). As a result, of the 1183 nonsense SNPs, 581 were predicted to cause NMD, and thus to prevent translation. The other 602 cases of nonsense SNPs were predicted to result in truncated proteins. For the cases that truncated proteins are produced, the average truncation was estimated to be 75 amino acids.

To see which of these nonsense SNPs were known pathological mutations, we compared them with allelic variants in the Online Mendelian Inheritance in Man (OMIM) database. Only eight of 1,183 nonsense SNPs (rs17602729 in *AMPD1*, rs283413 in *ADH1C*, rs10250779 in *PGAM2*, rs17215500 in *KCNQ1*, rs497116 in *CASP12*, rs2228325 in *ACTN3*, rs3092891 in *RB1* and rs28989186 in *BUB1B*) matched the variants in the OMIM database that are known variants with phenotypic variations (Table 7). This low value suggests that the biological effects of most nonsense SNPs have not yet been reported. Interestingly, each of the eight cases that matched known pathological variants was predicted to cause NMD (Table 7).

Table 1. SNPs and indels in exon, intron and other genomic regions.

	Exon	Intron	Other genomic regions
SNPs	249,182	3,332,537	5,209,127
Indels	9,742	185,761	249,648

Polymorphisms mapped on single positions were analyzed with 36,712 protein-coding genes.
doi:10.1371/journal.pone.0003393.t001

SNPs that cause read-though of the original termination codon

Among the 247 SNPs at termination codon sites, 119 SNP-mRNA pairs were found to be read-through mutations. If the allele having the stop codon is the ancestral type, the SNP is

Table 2. Classified SNPs in exon regions.

Region	Effects on translation	Genes in category I–IV[a]	All protein-coding genes[b]
5′UTR		23454 [3.3×10^{-3}/site][c]	51881
ORF	Total	85233 [2.7×10^{-3}/site]	96164
	Synonymous	37484 [4.1×10^{-3} /site]	40484
	Nonsynonymous	46261 [2.1×10^{-3} /site]	53754
	AA↔Ter[d]	938	1258
	Unclassified[e]	398	421
Stop codon	Total	152	247
	Synonymous	63	88
	Ter↔AA[d]	89	159
3′UTR		69691 [3.3×10^{-3}/site]	104510
Total		178378	252555

[a]Representative transcripts in 23,717 genes whose function were defined or suggested (similarity category I–III) and genes annotated as conserved hypothetical proteins (similarity category IV).
[b]Representative transcripts in all protein-coding genes (36,712) including genes in similarity category I–IV plus similarity category V–VII (hypothetical protein, hypothetical short protein, and pseudogene candidate, respectively).
[c]Densities of polymorphisms are shown in brackets as average number of polymorphisms per site. The average lengths of the 5′UTR, ORF and 3′UTR regions in 23717 genes were 303.9 bp, 1343.5 bp, and 877.6 bp, respectively. The densities of SNPs for synonymous, nonsynonymous and nonsense SNPs in ORFs were calculated based on the numbers of potential nucleotide sites for synonymous, nonsynonymous and nonsense mutations in coding regions. The density of nonsense SNPs is shown in Table 3.
[d]SNPs causing changes between amino acids and stop codons.
doi:10.1371/journal.pone.0003393.t002

Table 3. SNPs causing changes between amino acids and stop codons.

Region	Effects on translation	Genes in category I–IV[a]	All protein-coding genes[a]
ORF	Nonsense	910 [0.85×10^{-4}/site][d]	1183
	Read-through[b]	28	75
Stop codon	Read-through	67	110
	Nonsense[c]	22	49

[a]These two gene sets are the same as Table 2.
[b]Possible read-through SNPs in which alleles coding stop codons were ancestral type. This may be due to existence of shorter ORFs in the ancestral population.
[c]Possible nonsense SNPs in which alleles coding stop codons were derived alleles. This may be due to existence of longer ORFs in the ancestral population.
[d]The densities of nonsense SNPs in ORFs were calculated based on the numbers of potential nucleotide sites for nonsense mutations in coding regions.
doi:10.1371/journal.pone.0003393.t003

Table 4. Insertions and deletions in exon regions.

	Genes in category I–IV[a]	All protein-coding genes[a]
5′UTR	785 [0.11×10^{-3}][b]	2005
ORF	1120 [0.035×10^{-3}]	1532
3′UTR	3323 [0.16×10^{-3}]	4942
Total	5225[c]	8479

[a]These two gene sets are the same as Table 2.
[b]Densities of polymorphisms are shown in brackets as average number of polymorphisms per site.
[c]Three indels were located on both of ORF and UTR.
doi:10.1371/journal.pone.0003393.t004

Table 5. Frequency of each type of codon change for nonsense SNPs.

	TAA		TAG		TGA		Total
	Aaa→Taa	33	Aag→Tag	31	Aga→Tga	20	
1st	**Caa→Taa**	**62**	**Cag→Tag**	**162**	**Cga→Tga**	**203**	748*
	Gaa→Taa	80	Gag→Tag	125	Gga→Tga	32	
	tCa→tAa	27	tCg→tAg	19	tCa→tGa	25	
2nd			**tGg→tAg**	**80**			200
	tTa→tAa	18	tTg→tAg	18	tTa→tGa	13	
	taC→taA	25	taC→taG	25	tgC→tgA	22	
3rd					**tgG→tgA**	**85**	235
	taT→taA	19	taT→taG	27	tgT→tgA	32	
Total		264		487		432	1183

Bold letters show nucleotide changes by transition.
*$P<0.005$ by chi-square test.
doi:10.1371/journal.pone.0003393.t005

regarded as a change causing elongation of the polypeptide. However, an extended polypeptide would be expected only if there is an additional termination codon downstream. For 108 SNP-mRNA pairs, an additional termination codon was found in the 3′UTR region. The average extension was estimated to be 29 amino acids. Interestingly, we found five SNP-mRNA pairs that have no stop codons in the 3′UTR at all (The remaining six SNP-mRNA pairs do not have 3′UTR regions). For example, the T-to-C substitution (rs15941) in the *DDR2* gene (X74764) is predicted to be a read-through mutation (from TAG to CGA), and the transcript has no other stop codon in the 3′UTR region. The frequency of this SNP is unknown (it is monomorphic in the four populations in HapMap project [4]). However, if this polymorphism really exists, transcripts having this read-through mutation would not produce a protein. Another example is the T-to-C substitution (rs17850833) in

Table 6. Nonsense SNPs and prediction of NMD.

	Predicted to cause NMD[a]	Not for NMD[b]	Total
Known pathological variants	8[c]	0	8
Other nonsense SNPs	573	602	1175
Total	581	602	1183

[a]This prediction is based on that mRNA would be destroyed if a stop codon occurs in the 5′ side of the boundary, which is 50–55 nucleotides upstream from the 3′ end of the second to last exon. Here, the nonsense SNPs located in the 5′ side of the boundary, which was set at 50 nucleotides upstream from the 3′ end of the second to last exon, were predicted to cause NMD.
[b]This number includes SNPs in genes consisting of only one exon.
[c]P = 0.0033 by Fisher's exact test.
doi:10.1371/journal.pone.0003393.t006

the *MFSD3* gene (CR620962), which causes a change from TGA to CGA resulting in a change to arginine.

Functional bias of genes having nonsense SNPs

To see whether there is any functional bias in genes having nonsense SNPs, we examined the frequent biological terms in the genes having nonsense SNPs. We classified the genes having nonsense SNPs into two categories: genes with nonsense SNPs that are predicted to cause NMD and genes with nonsense SNPs that are not predicted to cause NMD. For genes having nonsense SNPs that would cause NMD (Table 8), the molecular functions that are most overrepresented included phosphorylation, ATP binding, iron/calcium ion binding, nucleotide/RNA binding and transporter activity. The localization of these genes was also biased to the cell membrane and the proteinaceous extracellular matrix. On the other hand, the genes having nonsense SNPs predicted to not cause NMD showed less bias in biological function (Table 9).

Table 7. Nonsense SNPs with known pathological effects.

Acc#	Chr	Gene symbol	SNP	Variation	OMIM	Biological effects
M60092	1	*AMPD1*	rs17602729	Gln12Ter	102770	AMPD deficiency
M12272	4	*ADH1C*	rs283413	Gly78Ter	103730	Parkinson disease
BC073741	7	*PGAM2*	rs10250779	Trp78Ter	261670	Myopathy
AF000571	11	*KCNQ1*	rs17215500	Arg518Ter	607542	Long QT syndrome 1
AY358222	11	*CASP12*	rs497116	Arg125Ter	608633	Sepsis susceptibility
M86407	11	*ACTN3*	rs2228325	Arg577Ter	102574	Athletic performance
L41870	13	*RB1*	rs3092891	Arg445Ter	180200	Bilateral retinoblastoma
AF068760	15	*BUB1B*	rs28989186	Arg194Ter	602860	Premature chromatid separation trait and mosaic variegated aneuploidy syndrome

doi:10.1371/journal.pone.0003393.t007

Table 8. Functional bias of genes having nonsense SNPs causing NMD.

Top level	Gene Ontology no.	Gene Ontology	Observed gene no.[a]	Expected gene no.[b]	Ratio of enrichment	P value[c]
Biological process	0006118	electron transport	15	4.23	3.55	5.03×10^{-5}
	0006468	protein amino acid phosphorylation	16	7.28	2.20	4.98×10^{-3}
Cellular component	0016020	membrane	41	22.55	1.82	5.57×10^{-4}
	0005578	proteinaceous extracellular matrix	8	1.21	6.62	2.17×10^{-6}
Molecular function	0005524	ATP binding	35	17.15	2.04	1.79×10^{-4}
	0004713	protein tyrosine kinase activity	16	6.46	2.48	1.56×10^{-3}
	0004674	protein serine/threonine kinase activity	16	6.78	2.36	2.51×10^{-3}
	0000166	nucleotide binding	14	5.61	2.50	2.79×10^{-3}
	0004672	protein kinase activity	16	7.15	2.24	4.21×10^{-3}
	0003723	RNA binding	10	3.11	3.22	1.82×10^{-3}
	0005506	iron ion binding	8	2.00	4.00	1.32×10^{-3}
	0005509	calcium ion binding	16	7.65	2.09	7.89×10^{-3}
	0005215	transporter activity	10	3.44	2.91	3.76×10^{-3}
	0016491	oxidoreductase activity	11	4.24	2.59	5.76×10^{-3}
	0003779	actin binding	6	1.27	4.74	2.24×10^{-3}
	0004759	carboxylesterase activity	5	0.24	20.44	4.19×10^{-6}

[a]Number of genes with a molecular function in the 581 genes in which nonsense SNPs causing NMD were found.
[b]Expected number of genes that have a biological function in a sample of 581 genes, assuming a proportion of genes with a molecular function in all human genes.
[c]Enrichment of a biological term in the genes for nonsense SNPs was statistically evaluated as a upper probability in a hypergeometric distribution.
doi:10.1371/journal.pone.0003393.t008

Table 9. Functional bias of genes having nonsense SNPs not causing NMD.

Top level	Gene Ontology no.	Gene Ontology	Observed gene no.[a]	Expected gene no.[b]	Ratio of enrichment	P value[c]
Biological process	0007156	homophilic cell adhesion	6	1.42	4.23	3.05×10^{-3}
	0006310	DNA recombination	3	0.19	15.50	8.25×10^{-4}
	0006414	translational elongation	3	0.34	8.85	4.48×10^{-3}
	0042254	ribosome biogenesis and assembly	2	0.15	13.77	8.68×10^{-3}
Cellular component	0005853	eukaryotic translation elongation factor 1 complex	2	0.13	15.50	6.82×10^{-3}
Molecular function	0004194	pepsin A activity	2	0.18	11.27	1.30×10^{-2}
	0003746	translation elongation factor activity	2	0.29	6.89	3.35×10^{-2}

[a]Number of genes with a molecular function in the 602 genes in which nonsense SNPs causing NMD were found.
[b]Expected number of genes that have a biological function in a sample of 602 genes, assuming a proportion of genes with a molecular function in all human genes.
[c]Enrichment of a biological term in the genes for nonsense SNPs was statistically evaluated as a upper probability in a hypergeometric distribution.
doi:10.1371/journal.pone.0003393.t009

Discussion

In this study, we conducted an extensive analysis of human genome polymorphisms with a comprehensive catalogue of human genes, and detected more than 50,000 polymorphisms that affect proteins. The distribution of polymorphisms showed different densities of polymorphisms among the 5′UTR, ORF and 3′UTR. The density of SNPs was lower in ORFs than in the 5′UTR and 3′UTR. The density of synonymous SNPs in the ORFs was higher than the densities of SNPs in the UTR regions. The reduction in density of SNPs in the UTR regions is consistent that there are functional constraints on nucleotide changes in UTRs related to the transcriptional and translational efficiency[22]. The density of nonsynonymous SNPs was much lower than the densities of other types of SNPs, possibly due to that the nucleotide changes with alteration of amino acids changes are under strong negative selection [36]. It was not known how nonsense SNPs are distributed in protein-coding regions. Here we showed that the density of nonsense SNPs is much lower than that of nonsynonymous SNPs. Although the biological effects of nonsense mutations appear to vary widely depending on their positions and the genes, the low density of nonsense SNPs that we found suggests that nonsense mutations have more disadvantageous effects than nonsynonymous mutations.

While nonsense mutations that cause NMD result in 'loss of function', nonsense mutations that do not cause NMD produce truncated proteins which could have the dominant effects. The proportion of predicted nonsense SNPs causing NMD in this study is in agreement with a previous study which showed that dbSNP (build 125) has 1301 nonsense SNPs, about half of which were predicted to result in NMD [37]. In order to understand the biological effects of nonsense SNPs, it is important to know whether they do or do not cause NMD, because premature stop codons in a gene can have distinct disease phenotypes depending on the positions of mutations [27,38].

The molecular functions that were overrepresented in the genes having nonsense SNPs included several molecular functions that were observed in human-specific pseudogenes[39], such as ATP binding, actin binding, calcium ion binding, extracellular matrix, nucleic acid binding and oxidoreductase. This is in accord with that nonsense mutations contribute to 'pseudogenization'. It is interesting that nonsense SNPs causing NMD were frequently found in genes that encode proteins involved in phosphorylation, cell-cell interaction, signal transduction and transport. This may be because changes in the length of polypeptides caused by nonsense mutations are under strong negative selection in the genes involved in signal transduction or transportation because abnormal translation products could cause dominant effects. Therefore, inactivation of translation by nonsense mutations in those genes could have milder effects than changes of the length of polypeptides.

Our results showed a low proportion of matches of nonsense SNPs with known pathological variants in OMIM, suggesting that the effects of most nonsense polymorphisms are unknown or not reported. Furthermore, the correspondence of the nonsense SNPs to the OMIM allelic variants (Table 6, Table 7) suggests that nonsense polymorphisms that are subject to NMD are more likely to be involved in phenotypic variations.

There is a possibility that the nonsense SNPs detected here have pathological effects, in particular, if non-dispensable genes have nonsense mutations. First, a defect in one gene by a nonsense mutation or a frame-shifting indels causing a premature termination codon could be a cause of genetic diseases including complex diseases[40]. Second, there is a possibility that nonsense mutations cause recessive lethal alleles that would not be detected as causative variant of diseases. Probably, focusing on nonsense polymorphisms observed in specific populations would be a good way of selection for finding variants with deleterious effects.

The effect of single nonsense SNPs can be compensated by the products of other genes having similar functions[41] and the other splicing isoforms of the gene [42]. Thus, single nonsense SNPs may not always cause severe phenotypic effects. In fact, some nonsense SNPs with high allele frequencies were found across populations[43]. There is a report of fixation of an inactive form of caspase 12 by a nonsense mutation (rs497116) in non-African populations[43], and this is an example supporting the 'less is more hypothesis'[44]. This example suggests that some of nonsense mutations are not disadvantageous and that the increase of frequency of a nonsense allele could be driven by positive selection.

Elongation of polypeptides by read-through mutations can affect protein folding and aggregation of proteins, which could affect phenotypic variations. Furthermore, a read-through mutation can cause more severe effects on translation when no additional stop codon follows. Such mutations are subject to 'non-stop decay' [32,33], and would result in no gene product. It has been suggested that non-stop decay and NMD serve to remove toxic, aberrant proteins [29]. It is unclear how frequently such mutations prevent mRNA from producing proteins. Therefore, it would be quite useful to be able to predict the effects of various types of genetic changes on mRNA.

Although the present results are based on representative transcripts (one transcript for one gene), the total number of

SNPs causing changes between amino acids and stop codons in all the splicing isoforms was much larger (2,234). These variations, which cause changes in the length of a polypeptide or which determine whether a protein is translated, may include pathological variants that have yet not been reported. Therefore, it is important to examine their presence in human populations.

Materials and Methods

Data of human genetic polymorphisms

As data of genetic polymorphisms of human genome, single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) in dbSNP [1] were used in this study. The whole data of human SNPs and indels were downloaded from dbSNP (build 125). We used all SNPs and indels that were mapped on single position in the genome, except for 'large insertions' in dbSNP.

Data of human genes

The data of human gene structure were obtained from H-InvDB ver3.8 (http://www.h-invitational.jp/), created by the annotation project of human genes (H-Invitational project) [23,45]. Our analysis of all human genes that corresponds to H-InvDB (ver 3.8) predicted 36,712 protein coding loci. All protein-coding genes were annotated and classified based on similarity to known genes as follows; Category I, Identical to known human protein; Category II, Similar to known protein; Category III, IPR domain containing protein; Category IV, Conserved hypothetical protein; Category V, Hypothetical protein; Category VI, Hypothetical short protein; Category VII, pseudogene candidate. We used the following three kinds of data of the gene structure: 1) genomic location of exons to the human genome (build 35), 2) predicted ORF regions in transcripts, and 3) original and curated cDNA sequences.

Analysis

1. Analysis of polymorphism with exons and predicted ORFs. Selection of polymorphisms on exon regions. We selected polymorphisms in exon regions by comparing the genomic positions of polymorphisms and the start and end positions of exons that were obtained from mapping cDNA sequences to the human genome (Figure 1). Polymorphisms in introns were also selected in a same way.

Conversion of genomic position of polymorphism into nucleotide position in cDNA sequence. To analyze polymorphisms with a predicted ORF, nucleotide positions of polymorphisms in the human genome sequences were converted into the nucleotide positions in cDNA sequences. Because there could be gaps in the alignment of cDNA sequence and the human genome sequence, the nucleotide position was converted considering possible gaps in the alignment. When the cDNA sequence was corrected in ORF prediction because of frame-shifting and remaining intron, the nucleotide position of SNP was modified based on addition or deletion of nucleotides. For a quality control of polymorphism data used for classification, we conformed that one of the nucleotides in each pair of SNP alleles was the same nucleotide at the corresponding position in the cDNA sequence.

Classification of polymorphisms with predicted ORF. Polymorphisms within ORF were classified according to their effect on ORF. For SNPs with two alleles, alleles in nucleotide were converted into 'alleles in codon' by adding two other nucleotides in the codon from cDNA sequence. When a cDNA sequence was corrected in the annotation process by removing a remaining intron or by correcting a frameshift error, the corrected cDNA sequence was used. If these alleles in codon do not contain any stop codon, the alleles were classified into synonymous and nonsynonymous. In case a stop codon is included in the alleles in codon, they were classified into 1) premature termination (nonsense) codon, 2) read-through of original stop codon, and 3) synonymous at stop codon site, by assuming that the cDNA sequence has an ancestral allele. Indels were classified based on whether they are located in ORF. The indels within ORF were further classified by whether the insertion or deletion causes frame shifting in translation.

Inference of direction of nonsense and read-through mutations. Ancestral alleles were obtained from dbSNP (build 128) to check direction of mutations for SNPs causing changes between amino acids and stop codons. For nonsense SNPs in protein-coding regions, we checked whether the ancestral allele codes amino acids. In case that the ancestral allele codes stop codon, we do not regard this SNP as nonsense SNP, but is a read-through mutation assuming that there was a variant having a shorter ORF. For read-though SNPs at termination codon site, we checked whether the ancestral allele codes stop codon. In case that the ancestral allele codes amino acids, we regard this SNP not as a read-through mutation, but as a nonsense mutation in a variant having a longer ORF.

Number of sites for synonymous, nonsynonymous and nonsense mutations. To estimate densities of synonymous, nonsynonymous and nonsense SNPs, the numbers of potential synonymous, nonsynonymous and nonsense sites by single nucleotide changes were estimated for the ORF sequences. This is an extension of estimation of the numbers of synonymous and nonsynonymous sites[46]; the number of synonymous sites is calculated as the number of four-fold degenerate sites plus one-third of the number of two-fold degenerate sites. For 61 codons encoding amino acids, the numbers of nucleotide sites that would cause synonymous, nonsynonymous and nonsense mutations by a single nucleotide change were estimated with a model of nucleotide change. Here, the relative occurrence of a transitional mutation versus a transversional mutation (r) was set to be 4.0 (the expected ratio in the numbers of transitional and transversional mutations was 2.0). For example of the TTA codon for leucine, the number of nonsense sites was estimated to be $2.0/(r+2.0)$, because two types of transversional mutations at the second position cause nonsense mutations.

2. Correspondence to known pathological variants. To check whether the polymorphisms that alter proteins are known pathological variants with phenotypic effect, we examined correspondence of SNPs with data of known pathological variants. We used data of 'allelic variant' in the Online Mendelian Inheritance in Man (OMIM) database [18] as information of variants with phenotypic effect. For nonsynonymous and nonsense SNPs, their effects on translation and positions in ORF were compared with the 'list of alleles' in OMIM (e.g. described as "TRP324TER" or "ALA279THR" for the *NGAS* gene).

3. Prediction of nonsense SNPs causing NMD. Some of nonsense mutations cause nonsense-mediated decay (NMD), resulting in prevention of translation. It has been reported that mRNA would be destroyed if a stop codon occurs in the 5′ side of the boundary, which is 50–55 nucleotides upstream from the end of the second to last exon [30,31]. To predict whether a nonsense SNP causes NMD, we examined whether a nonsense SNP is located in the 3′ side of the boundary, which was set at 50 nucleotides upstream from the end of the second to last exon, in the exon-intron structure. This method is the same as the method in SNP2NMD [37] when 'NMD distance' is 50 nucleotides.

5. Functional bias of genes with nonsense SNPs. For each biological term from Gene Ontology (www.geneontology.org), a proportion of genes with the biological function in the genes having nonsense SNPs was compared with that in all human genes (representative transcripts in all human genes in H-InvDB ver 5.0), and the significance of over representation of a molecular function

in the genes having nonsense SNPs was evaluated as the upper probability of the hypergeometric distribution.

Supporting Information

Results S1 Supplementary results and a table for analyses of nonsynonymous SNPs.
Found at: doi:10.1371/journal.pone.0003393.s001 (0.70 MB DOC)

Table S1 Nonsense SNPs and read-through SNPs on representative transcripts.
Found at: doi:10.1371/journal.pone.0003393.s002 (4.24 MB DOC)

Acknowledgments

We thank Drs. Craig Gough, Norikazu Yasuda, Shuhei Mano, Naoki Nagata, Yoshiyuki Suzuki and Hisakazu Iwama for helpful discussion. We also thank Ryuzou Matsumoto, Seigo Hosono, and all the members in the Integrated Database Team of BIRC, AIST for their technical assistance and providing data of gene structure and annotation.

Author Contributions

Conceived and designed the experiments: YYK SM RC TG TI. Analyzed the data: YYK MS YH. Wrote the paper: YYK MS TI.

References

1. Sherry ST, Ward M, Sirotkin K (1999) dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677–679.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
3. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389.
4. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320.
5. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, et al. (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet 32: 650–654.
6. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
7. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, et al. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 33: D527–532.
8. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genome Res 12: 436–446.
9. Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33: W480–482.
10. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591–597.
11. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, et al. (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21: 3176–3178.
12. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7: 166.
13. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32: D520–522.
14. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, et al. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21: 2814–2820.
15. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, et al. (2007) Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39: 1329–1337.
16. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, et al. (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet 39: 207–211.
17. Minoshima S, Mitsuyama S, Ohtsubo M, Kawamura T, Ito S, et al. (2001) The KMDB/MutationView: a mutation database for human disease genes. Nucleic Acids Res 29: 327–328.
18. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30: 52–55.
19. Morton NE, Crow JF, Muller HJ (1956) An Estimate of the Mutational Damage in Man from Data on Consanguineous Marriages. Proc Natl Acad Sci U S A 42: 855–863.
20. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231–238.
21. Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227–1234.
22. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, et al. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22: 239–247.
23. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, et al. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2: e162.
24. Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, et al. (2008) The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. Nucleic Acids Res 36: D793–799.
25. Chang JC, Kan YW (1979) beta 0 thalassemia, a nonsense mutation in man. Proc Natl Acad Sci U S A 76: 2886–2889.
26. Rosenfeld PJ, Cowley GS, McGee TL, Sandberg MA, Berson EL, et al. (1992) A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa. Nat Genet 1: 209–213.
27. Inoue K, Khajavi M, Ohyama T, Hirabayashi S, Wilson J, et al. (2004) Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. Nat Genet 36: 361–369.
28. Mimori A, Hidaka Y, Wu VC, Tarle SA, Kamatani N, et al. (1991) A mutant allele common to the type I adenine phosphoribosyltransferase deficiency in Japanese subjects. Am J Hum Genet 48: 103–107.
29. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE (2004) Nonsense-mediated decay approaches the clinic. Nat Genet 36: 801–808.
30. Thermann R, Neu-Yilik G, Deters A, Frede U, Wehr K, et al. (1998) Binary specification of nonsense codons by splicing and cytoplasmic translation. Embo J 17: 3484–3494.
31. Zhang J, Sun X, Qian Y, LaDuca JP, Maquat LE (1998) At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. Mol Cell Biol 18: 5272–5283.
32. Frischmeyer PA, van Hoof A, O'Donnell K, Guerrerio AL, Parker R, et al. (2002) An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science 295: 2258–2261.
33. van Hoof A, Frischmeyer PA, Dietz HC, Parker R (2002) Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science 295: 2262–2264.
34. Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, et al. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A 100: 15754–15757.
35. Ehrlich M, Wang RY (1981) 5-Methylcytosine in eukaryotic DNA. Science 212: 1350–1357.
36. Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624–626.
37. Han A, Kim WY, Park SM (2007) SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics 23: 397–399.
38. Thein SL, Hesketh C, Taylor P, Temperley IJ, Hutchinson RM, et al. (1990) Molecular basis for dominantly inherited inclusion body beta-thalassemia. Proc Natl Acad Sci U S A 87: 3924–3928.
39. Wang X, Grus WE, Zhang J (2006) Gene losses during human origins. PLoS Biol 4: e52.
40. Senee V, Chelala C, Duchatelet S, Feng D, Blanc H, et al. (2006) Mutations in GLIS3 are responsible for a rare syndrome with neonatal diabetes mellitus and congenital hypothyroidism. Nat Genet 38: 682–687.
41. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
42. Takeda J, Suzuki Y, Nakao M, Barrero RA, Koyanagi KO, 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.
43. Xue Y, Daly A, Yngvadottir B, Liu M, Coop G, et al. (2006) Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet 78: 659–670.
44. Olson MV (1999) When less is more: gene loss as an engine of evolutionary change. Am J Hum Genet 64: 18–23.
45. Yamasaki C, Koyanagi KO, Fujii Y, Itoh T, Barrero R, et al. (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). Gene 364: 99–107.
46. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418–426.

The future of biocuration

To thrive, the field that links biologists and their data urgently needs structure, recognition and support.

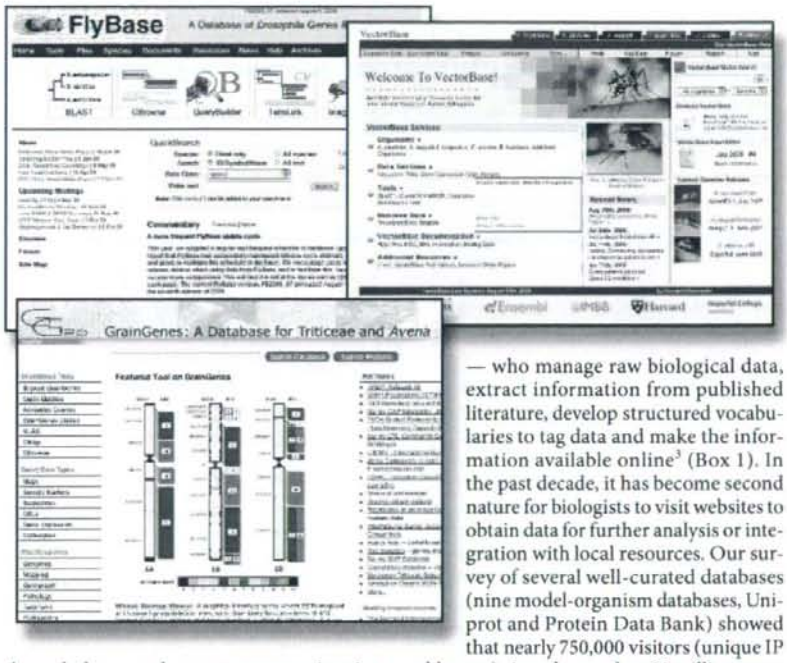
Doug Howe, Seung Yon Rhee *et al.*

The exponential growth in the amount of biological data means that revolutionary measures are needed for data management, analysis and accessibility. Online databases have become important avenues for publishing biological data. Biocuration, the activity of organizing, representing and making biological information accessible to both humans and computers, has become an essential part of biological discovery and biomedical research. But curation increasingly lags behind data generation in funding, development and recognition.

We propose three urgent actions to advance this key field. First, authors, journals and curators should immediately begin to work together to facilitate the exchange of data between journal publications and databases. Second, in the next five years, curators, researchers and university administrations should develop an accepted recognition structure to facilitate community-based curation efforts. Third, curators, researchers, academic institutions and funding agencies should, in the next ten years, increase the visibility and support of scientific curation as a professional career.

Failure to address these three issues will cause the available curated data to lag farther behind current biological knowledge. Researchers will observe an increasing occurrence of obvious gaps in knowledge. As these gaps expand, resources will become less effective for generating and testing hypotheses, and the usefulness of curated data will be seriously compromised.

When all the data produced or published are curated to a high standard and made accessible as soon as they become available, biological research will be conducted in a manner that is quite unlike the way it is done now. Researchers will be able to process massive amounts of complex data much more quickly. They will garner insight about the areas of their interest rapidly with the help of inference programs. Digesting information and generating hypotheses at the computer screen will be so much faster that researchers will get back to the bench quickly for more experiments. Experiments will be designed with more insight; this increased specificity will cause an exponential growth in knowledge, much as we are experiencing exponential growth in data today.



Data avalanche

Biology, like most scientific disciplines, is in an era of accelerated information accrual and scientists increasingly depend on the availability of each others' data. Large-scale sequencing centres, high-throughput analytical facilities and individual laboratories produce vast amounts of data such as nucleotide and protein sequences, protein crystal structures, gene-expression measurements, protein and genetic interactions and phenotype studies. By July 2008, more than 18 million articles had been indexed in PubMed and nucleotide sequences from more than 260,000 organisms had been submitted to GenBank[1,2]. The recently announced project to sequence 1,000 human genomes in three years to reveal DNA polymorphisms (www.1000genomes.org) is a tip of the data iceberg.

Such data, produced at great effort and expense, are only as useful as researchers' ability to locate, integrate and access them. In recent years, this challenge has been met by a growing cadre of biologists — 'biocurators' — who manage raw biological data, extract information from published literature, develop structured vocabularies to tag data and make the information available online[3] (Box 1). In the past decade, it has become second nature for biologists to visit websites to obtain data for further analysis or integration with local resources. Our survey of several well-curated databases (nine model-organism databases, Uniprot and Protein Data Bank) showed that nearly 750,000 visitors (unique IP addresses) viewed more than 20 million pages in just one month (March 2008, Eva Huala, Peter Rose, Rolf Apweiler, personal communications).

Despite the essential part that it plays in today's research, biocuration has been slow to develop. To provide a forum for the exchange of ideas and methods, and to facilitate collaborations and training, more than 150 biocurators met at two international conferences and created a mailing list and a website (www.biocurator.org). These meetings and discussions have honed in on the three actions, outlined above and elaborated on below, that must now be addressed to ensure scientists' continued access to the high-quality data on which their research depends.

Come together

Extracting, tagging with controlled vocabularies, and representing data from the literature, are some of the most important and time-consuming tasks in biocuration. Curated information from the literature serves as the gold-standard data set for computational analysis, quality assessment of high-throughput data and benchmarking of data-mining

algorithms. Meanwhile, the boundaries of the biological domain that researchers study are widening rapidly, so researchers need faster and more reliable ways to understand unfamiliar domains. This too is facilitated by literature curation.

Typically, biocurators read the full text of articles and transfer the essence into a database. For a paper about the molecular biology of a particular gene, process or pathway, such information might include gene-expression patterns, mutant phenotypes, results of biochemical assays, protein-complex membership and the authors' inferences about the functions and roles of the gene products studied. As each paper uses different experimental and analysis methods, capturing this information in a consistent fashion requires intensive thought and effort. Limited resources and staff mean that most curation groups can't keep up with all the relevant literature.

How information is presented in the literature greatly affects how fast biocurators can identify and curate it. Papers still often report newly cloned genes without providing GenBank IDs or the species from which the genes were cloned. The entities discussed in a paper, including species, genes, proteins, genotypes and phenotypes must be unambiguously identified during curation. For example, using the HUGO Gene Nomenclature Committee resource (www.genenames.org), we find that the human gene *CDKN2A* has ten literature-based synonyms. One of those, *p14*, is also a synonym for five other genes: *CDK2AP2*, *CTNNBL1*, *RPP14*, *S100A9* and *SUB1*. To confirm the identity of the gene described, curators make inferences from synonyms, reported sequences, biological context and bibliographic citations. This time-consuming and error-prone step could be eliminated by compliance with data-reporting standards[4–9].

Most recent efforts in this direction have been developed by the communities that produce large-scale genomics data. The vast majority of the peer-reviewed literature does not yet have a reporting-structure standard. As publication has become a mainly digital endeavour, however, publications and biological databases are becoming increasingly similar. Properly cross-referenced and indexed, each could serve as an access point to the other[10]. Such collaboration between databases and journals would improve researchers' access to data and make their work more visible.

We recommend that all journals and reviewers require that a distinct section of the Methods (or a supplemental document) of all published articles includes approved gene symbols (which are inherently unstable) and model-organism database IDs (which do not change) for genes discussed; nucleotide or protein accession numbers (GenBank or UniProt ID) for isoforms of each gene or protein discussed; and descriptions of species, strains, cell types and genotypes used. Examples of sources for this information are listed in Table 1. This would accelerate literature curation, uphold information integrity, facilitate the proper linkage of data to other resources and support automated mining of data from papers. Another model is for authors to provide a 'structured digital abstract' — a machine-readable XML summary of pertinent facts in the article[11] — along with a manuscript. This approach is in an experimental phase at the journal *FEBS Letters*[12].

Box 1 | The role of biocurators

- To extract knowledge from published papers
- To connect information from different sources in a coherent and comprehensible way
- To inspect and correct automatically predicted gene structures and protein sequences to provide high-quality proteomes
- To develop and manage structured controlled vocabularies that are crucial for data relations and the logical retrieval of large data sets
- To integrate knowledge bases to represent complex systems such as metabolic pathways and protein-interaction networks.
- To correct inconsistencies and errors in data representation
- To help data users to render their research more productive in a timely manner
- To steer the design of web-based resources
- To interact with researchers to facilitate direct data submissions to databases

Journals should also mandate direct submission of data into appropriate databases as a part of publication. This has been implemented by the journal *Plant Physiology* and curators of The *Arabidopsis* Information Resource (TAIR) database[13]. On acceptance of a manuscript, the corresponding author must fill out a simple web-based form to provide appropriate genetic and molecular information about the *Arabidopsis* genes in the publication. The information is sent to TAIR for integration by biocurators, who work with the authors to ensure that the data reported are of high quality and accurate.

"To date, not much of the research community is rolling up its sleeves to annotate."

As this infrastructure develops, we would like to see authors routinely tagging all aspects of the data in their publication semantically using universally agreed tag standards. Examples of such tags include the National Center for Biotechnology Information (NCBI) Taxon IDs, the Gene Ontology (GO) IDs and Enzyme Commission (EC) numbers. This information should be embedded in the electronic versions of publications or provided in a supplemental file similar to the crystallographic information file (CIF) currently required for publication of a crystal structure. The CIF file is submitted to the Protein Data Bank (www.pdb.org), which offers software to assist in preparation and validation of such crystallographic data[14]. An analogous system to help authors identify, tag and validate the crucial basic information in their research reports before publication would accelerate the automated linkage of literature to key records in existing databases and improve the accuracy of the published data.

In short, authors and publishers must use the existing publication infrastructure to facilitate literature curation much more to the benefit of all parties.

Community curation

Curation of large-scale genomics and post-genomics data enjoys no such luxury of 'an existing publication infrastructure' to leverage, although emerging standards of data reporting are promising[4–9]. Sooner or later, the research community will need to be involved in the annotation effort to scale up to the rate of data generation. This transition will require annotation tools, standardized methods, oversight by expert curators and a combination of social infrastructure, tool development, training and feedback. Biocurators are especially important for establishing such an infrastructure and training to maintain consistency and accuracy.

To date, not much of the research community is rolling up its sleeves to annotate. What will be the tipping point? The main limitation in community annotation is the perceived lack of incentive. For example, several model-organism databases have requested that authors annotate the genes they publish. This has historically failed for one main reason: contributions by experts consist of information they already know, and do not increase the value of the resource to themselves. A mechanism tied to career or research advancement may be required before community curation can be established as a broadly accepted and productive scientific endeavour[15]. Incentives for researchers to curate data should include new information or insight for their research interests, improvement in academic reputation or impact, career advancement and better funding chances. Academic departments and funding agencies should consider community annotation as a productive contribution to the scientific research corpus and a natural extension of the publication process.

For example, in the *Daphnia* Genomics Consortium (http://daphnia.cgb.indiana.edu) collaboration wiki, a community of more than 300 contributors took ownership of annotation of the genome while it was being sequenced at the Joint Genome Institute in Walnut Creek, California, and shared publication authorship as a consortium. Similarly, the International *Glossina* Genomics Initiative (http://iggi.sanbi.ac.za) hosted an annotation jamboree for field workers, population geneticists and molecular biologists to annotate tsetse fly molecular data as the sequence information became available. This

consortium-based publication mechanism is analogous to that used by other large-scale scientific projects such as the Sloan Digital Sky Survey (www.sdss.org). This is a viable course for communities that lack funding for dedicated curators, and offers a reward structure through consortium publication for participation and subsequent satellite papers.

The recently launched WikiProfessional Life Sciences (www.wikiprofessional.org) project links community curation with research and reputation gains. WikiProfessional indexed more than one million authors from PubMed and comparable numbers of biological concepts from authoritative databases and generated a simple way for researchers to update the information[16]. Because new potential 'facts' are mined from the network of associated concepts, the more accurate and comprehensive a particular concept is, the more chance it will have of being associated with other relevant ones, which in turn will lead to more potential new facts. All the updates researchers make are immediately publicly visible under their own name. Similarly, the Gene Wiki project generated thousands of wiki stubs in Wikipedia for human genes in an attempt to make it easier for the community to update the gene pages[17]. Although these wiki-based approaches provide an infrastructure for contributors to be recognized, there is not yet a standard practice for these contributions to be cited like a publication. It is imperative that the researchers, journal publishers and database curators start building a standard mechanism for citing annotation data sets.

Allowing anyone with a web browser, including the general public, to annotate entries would increase the number of potential annotators substantially, as pioneered in several astronomy projects. At Galaxy Zoo (www.galaxyzoo.org), 80,000 astronomers and members of the public manually classified the morphology of one million galaxies in less than three weeks. An analogous system to allow the public to contribute to biological annotation could be just as powerful if presented properly. For example, one could show a user an image of an *in situ* hybridization experiment and ask them to grade it as 'not expressed', 'restricted expression' or 'ubiquitous expression'. Even such basic information, if available for many thousands of genes, would be useful as first-pass annotation.

In sum, researchers (and even the general public) can be mobilized to provide the substantial resources needed to address the immense volume of data, if participation is appropriately rewarded. In the next five years, curators, funding agencies and academic institutions alike must find ways to consider substantial contributions to community curation efforts, much like a peer-reviewed publication, when it comes to issues of promotion, salary, hiring and funding.

Table 1 | Examples of knowledge-sharing databases

Species	Database	URL
Model organism databases		
Aedes aegypti	VectorBase	www.vectorbase.org
Anopheles gambiae	VectorBase	www.vectorbase.org
Arabidopsis thaliana	The *Arabidopsis* Information Resource	www.arabidopsis.org
Caenorhabditis elegans	WormBase	www.wormbase.org
Candida albicans	*Candida* Genome Database	www.candidagenome.org
Culex pipiens	VectorBase	www.vectorbase.org
Danio rerio	Zebrafish Information Network	http://zfin.org
Dictyostelium discoideum	dictyBase	http://dictybase.org
Drosophila sp.	FlyBase	http://flybase.org
Glycine max	SoyBase	www.soybase.org
Homo sapiens	HUGO Gene Nomenclature Committee	www.genenames.org
Hordeum vulgare	Barley Genetic Stocks Database	http://ace.untamo.net/bgs
Ixodes scapularis	VectorBase	www.vectorbase.org
Leishmania sp.	GeneDB	www.genedb.org
Mus musculus	Mouse Genome Informatics	www.informatics.jax.org
Oryza sp.	Gramene	http://gramene.org
Paramecium tetraurelia	ParameciumDB	http://paramecium.cgm.cnrs-gif.fr
Pediculus humanus	VectorBase	www.vectorbase.org
Rattus norvegicus	Rat Genome Database	http://rgd.mcw.edu
Saccharomyces cerevisiae	Saccharomyces Genome Database	www.yeastgenome.org
Schizosaccharomyces pombe	GeneDB	www.genedb.org
Solanaceae sp.	Sol Genomics Network	http://sgn.cornell.edu
Strongylocentrotus purpuratus	SpBase	http://sugp.caltech.edu/SpBase
Triticum sp.	GrainGenes	http://wheat.pw.usda.gov
Trypanosoma sp.	GeneDB	www.genedb.org
Xenopus laevis	Xenbase	www.xenbase.org
Xenopus tropicalis	Xenbase	www.xenbase.org
Zea mays	Maize Genetics and Genomics Database	www.maizegdb.org
Nucleotide, protein and structure databases		
All Species	GenBank	www.ncbi.nlm.nih.gov/Genbank
All Species	UniProt	www.pir.uniprot.org
All Species	Protein Data Bank	http://rcsb.org/pdb/home/home.do
Taxonomy		
All Species	NCBI Entrez Taxonomy	www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

Biological databases contain unique identifiers for the unambiguous identification of biological entities (such as genes, proteins, species and chemicals). These identifiers do not change as common biological names do. Authors should consult these databases for stable identifiers to cite in their publications.

Career path

How can biocuration mature faster as a career? Biocurators currently streamline submission to databases, automate curation, standardize data and facilitate contributions to annotation by research communities interested in the annotation process. To handle the increasing volume and types of data, journal publishers and researchers who generate data will need to be involved in the curation process and the roles of biocurators will expand to include editing and teaching. As biology moves towards more precise, quantitative science, biologists also need to adapt to thinking more quantitatively, systematically and objectively about their data; biocuration will need to become an inherent part of research and education in biology.

Biocuration requires a blend of skills and experience, including advanced scientific research and competence in database management systems, multiple operating systems and scripting languages. This type of background has typically been garnered through a combination of self-teaching and on-the-job experience, which can be narrow and spotty. Happily, formal education is becoming available. For example, the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign offers a biological information specialist master's degree and a specialization in data curation[18]. Experienced biocurators must lead the way in establishing more and better formal training programmes. In the next 5–10 years, biology curricula should include courses in biocuration as this becomes an increasingly common activity for all biological researchers. And interdisciplinary programmes that include courses in

biology, computer science and information science will be vital.

Attracting highly qualified individuals into this field has been challenging. The whole community must promote scientific curation as a professional career option. Funding agencies must assess the impact of curated data and support the development of innovative curation methods. To improve the profession, curators need a forum to share their experiences and publish their works. Oxford University Press plans to begin publishing a new journal in 2009 called *Database: The Journal of Biological Databases and Curation*. This may provide one such venue for publication of noteworthy advances in biocuration (www.database.oxfordjournals.org). Meanwhile, a committee of 20 biocurators and researchers is forming an International Society for Biocuration (www.biocurator.org/BiocuratorSociety.html) to make the discipline more visible and to promote it as an attractive career path. The official launch of the society is planned for the third International Biocuration Meeting next April in Berlin (http://projects.eml.org/Meeting2009).

Biology today needs more robust, expressive, computable, quantitative, accurate and precise ways to handle data. It is time to recognize that biocuration and biocurators are central to the future of the field. ■

1. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. *Nucl. Acid. Res.* **36,** D25-D30 (2008).
2. Wheeler, D. L. *et al. Nucl. Acid. Res.* **36,** D13-D21 (2008).
3. Salimi, N. & Vita, R. *PLoS Comput. Biol.* **2,** e125 (2006).
4. Brazma, A. *et al. Nature Genet.* **29,** 365-371 (2001).
5. Deutsch, E. W. *et al. Nature Biotechnol.* **26,** 305-312 (2008).
6. Field, D. *et al. Nature Biotechnol.* **26,** 541-547 (2008).
7. Jenkins, H. *et al. Nature Biotechnol.* **22,** 1601-1606 (2004).
8. Orchard, S. *et al. Nature Biotechnol.* **25,** 894-898 (2007).
9. Taylor, C. F. *et al. Nature Biotechnol.* **25,** 887-893 (2007).
10. Bourne, P. *PLoS Comput. Biol.* **1,** 179-181 (2005).
11. Seringhaus, M. R. & Gerstein, M. B. *BMC Bioinformatics* **8,** 17 (2007).
12. Seringhaus, M. & Gerstein, M. *FEBS Lett.* **582,** 1170 (2008).
13. Ort, D. R. & Grennan, A. K. *Plant Physiol.* **146,** 1022-1023 (2008).
14. Burkhardt, K., Schneider, B. & Ory, J. *PLoS Comput. Biol.* **2,** e99 (2006).
15. Rhee, S. Y. *Plant Physiol.* **134,** 543-547 (2004).
16. Mons, B. *et al. Genome Biol.* **9,** R89 (2008).
17. Huss, J. W. *et al. PLoS Biol.* **6,** e175 (2008).
18. Palmer, C. L., Heidorn, P. B., Wright, D. & Cragin, M. H. *Int. J. Dig. Curation* **2,** 31-40 (2007).

Author information Correspondence and requests for materials should be addressed to D.H. (e-mail: dhowe@cs.uoregon.edu) and S.Y.R. (e-mail: rhee@acoma.stanford.edu).

See Editorial, page 1.

Authorship

Doug Howe[1], Maria Costanzo[2], Petra Fey[3], Takashi Gojobori[4], Linda Hannick[5], Winston Hide[6,7], David P. Hill[8], Renate Kania[9], Mary Schaeffer[10,11], Susan St Pierre[12], Simon Twigger[13], Owen White[14] and Seung Yon Rhee[15]

[1]The Zebrafish Information Network, 5291 University of Oregon, Eugene, Oregon 97403-5291, USA. [2]*Saccharomyces* and *Candida* Genome Databases, Stanford University, Stanford, California 94305-5120, USA. [3]dictyBase, Northwestern University Biomedical Informatics Center, 750 N. Lake Shore Drive, 11-175, Chicago, Illinois 60611, USA. [4]Centre for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Yata, Mishima 411-8540, Japan. [5]J. Craig Venter Institute, Applied Bioinformatics, Rockville, Maryland 20850, USA. [6]South African National Bioinformatics Institute, University of the Western Cape, Private Bag X17, Bellville 7535, South Africa. [7]Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, USA. [8]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine 04609, USA. [9]Scientific Databases and Visualization, EML Research GmbH, Villa Bosch, Schloss-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany. [10]Division of Plant Sciences, University of Missouri, Columbia, Missouri, USA. [11]Plant Genetics Research Unit, Agricultural Research Service, United States Department of Agriculture, Columbia, Missouri 65211-7020, USA. [12]FlyBase, Harvard University, Cambridge, Massachusetts 02138, USA. [13]Rat Genome Database, Bioinformatics Research Center, Medical College of Wisconsin, 8701 Watertown Plank Rd, Milwaukee, Wisconsin 53226, USA. [14]Department of Epidemiology and Preventative Medicine, Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA. [15]The *Arabidopsis* Information Resource, Carnegie Institution for Science, Department of Plant Biology, 260 Panama Street, Stanford, California 94305, USA.

Published online 18 December 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D793–D799***
doi:10.1093/nar/gkm999

The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts*

Received September 16, 2007; Revised October 20, 2007; Accepted October 22, 2007

ABSTRACT

Here we report the new features and improvements in our latest release of the H-Invitational Database (H-InvDB; http://www.h-invitational.jp/), a comprehensive annotation resource for human genes and transcripts. H-InvDB, originally developed as an integrated database of the human transcriptome based on extensive annotation of large sets of full-length cDNA (FLcDNA) clones, now provides annotation for 120 558 human mRNAs extracted from the International Nucleotide Sequence Databases (INSD), in addition to 54 978 human FLcDNAs, in the latest release H-InvDB_4.6. We mapped those human transcripts onto the human genome sequences (NCBI build 36.1) and determined 34 699 human gene clusters, which could define 34 057 (98.1%) protein-coding and 642 (1.9%) non-protein-coding loci; 858 (2.5%) transcribed loci overlapped with predicted pseudogenes. For all these transcripts and genes, we provide comprehensive annotation including gene structures, gene functions, alternative splicing variants, functional non-protein-coding RNAs, functional domains, predicted sub cellular localizations, metabolic pathways, predictions of protein 3D structure, mapping of SNPs and microsatellite repeat motifs, co-localization with orphan diseases, gene expression profiles, orthologous genes, protein–protein interactions (PPI) and annotation for gene families. The current H-InvDB annotation resources consist of two main views: Transcript view and Locus view and eight sub-databases: the DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group.

INTRODUCTION

Human transcripts represent a biologically and functionally rich format for examining the structure of human genes and alternative splicing isoforms. In particular, cloning and sequencing of full-length cDNAs (FLcDNAs) that cover all exons but no introns can facilitate the precise determination of human gene structure (1). Studies on human transcripts have thus been systematically and extensively carried out to draw the outline of the human transcriptome (2–6). The human transcriptome consists of protein-coding mRNAs and non-coding functional RNAs. Analysis of these sequences will provide insights into how genomic information is transformed into higher order biological phenomena. By comparative analysis of the transcriptome with the human genome, we will be able to determine the transcribed regions of the genome and better understand the regulatory machinery of transcription (7, 8). It is therefore of great significance to collect information about human transcripts as well as their annotations. We thus held the first international workshop entitled 'Human Full-length cDNA Annotation Invitational' (abbreviated as H-Invitational or H-Inv) in Tokyo, Japan from 25th August to 3rd September 2002, and constructed a novel, integrative database of the human transcriptome, called H-InvDB (9,10). This consists of the annotation of 42 421 human FLcDNAs, collected from six high-throughput producers of human FLcDNAs in the world human gene collections.

To cover the increased number of human FLcDNAs since the initial release of H-InvDB, we held the second international annotation meeting entitled 'H-Invitational 2 Functional Annotation Jamboree' (abbreviated as H-Invitational 2 or H-Inv2) in Tokyo, Japan from 15th to 20th November 2003. The second major release of H-InvDB (release 2.0) was based on the annotation carried out at the H-Inv2 annotation jamboree. After H-Inv2, we initiated the Genome Information Integration Project (GIIP) and held the third and fourth annotation meetings in October 2005 and October 2006. The products of those two annotation meetings comprised releases 3.0 and 4.0 of H-InvDB. The increases in the number of entries in H-InvDB are summarized in Table 1.

THE ANNOTATION IN OUR LATEST UPDATE, H-InvDB 2007

In our latest release H-InvDB_4.6, we annotated 120 558 human mRNAs extracted from the International Nucleotide Sequence Databases (INSD) in addition to 54 978 human FLcDNAs that were available on 15th June 2006. We mapped those human transcripts onto the human genome sequences (NCBI build 36.1) and determined 34 699 human gene clusters, which could define 34 057

*A complete list of authors appears at the end of this article.

Table 1. Statistics of H-InvDB entries

H-InvDB release	Date of release	Number of transcripts (HIT)	Number of gene clusters (HIX)	Number of proteins (HIP)	Human genome	Date of sequence data-fix
1.0	2004/4/20	41 118	21 037	–	NCBI build 34.1	2002/7/15
2.0	2005/8/31	56 419	25 585	–	NCBI build 34.1	2003/9/1
3.0	2006/3/31	167 992	35 005	–	NCBI build 35.1	2005/3/1
4.0	2007/3/30	175 542	34 701	116 228	NCBI build 36.1	2006/6/15
4.6	2007/9/27	175 536	34 699	116 142	NCBI build 36.1	2006/6/15

Table 2. Statistics of manually curated representative H-Inv proteins

Category	Definition	Number of representative HITs	%
I	Identical to known[a] human protein (≥98% identity, =100% coverage)	12 404	36.42
II	Similar to known[a] protein (≥50% identity, ≥50% coverage)	3165	9.29
III	InterPro domain containing protein	3056	8.97
IV	Conserved hypothetical protein	4210	12.33
V	Hypothetical protein	5124	15.05
VI	Hypothetical short protein (20–79 amino acids)	5250	15.42
VII	Pseudogene candidates	858	2.52
Total		34 057	100

[a]'Known' proteins are experimentally validated proteins in literatures.

(98.1%) protein-coding and 643 (1.9%) non-protein-coding loci, while 858 (2.5%) transcribed loci overlapped with predicted pseudogenes. We basically followed the mapping technique we described previously (9,10). We updated annotation for the mitochondrial transcripts since the previous major release, H-InvDB_4.0, which resulted in a slightly decreased number for the transcripts and clusters. Then we assigned a standardized functional annotation to each H-Inv transcript by human curation, based on the results of similarity searches and InterPro-Scan (11). The numbers of manually curated human proteins in each category are summarized in Table 2.

For these transcripts and genes, we provide comprehensive annotation including descriptions of their gene structures, alternative splicing isoforms, functional non-protein-coding RNAs, functional domains of proteins, predicted sub cellular localizations, metabolic pathways, predictions of protein 3D structure, mapping of SNPs and microsatellite repeat motifs, co-localization with orphan diseases, gene-expression profiles, orthologous genes and evolutionary features in model animals, protein–protein interaction (PPI) and annotation for gene families. We have also annotated several new features related to transcript quality.

NEW ANNOTATED FEATURES IN H-InvDB

Classification of ncRNA

We annotated the transcripts that do not have homology to known protein-coding genes or InterPro-domain-containing genes as non-protein-coding transcript candidates. We classified 1216 non-protein-coding transcripts into 'Identical to known ncRNA' (124), 'Similar to known ncRNA' (74) and 'Putative ncRNA' (1018) by homology with known ncRNA databases and discrimination analysis

Sequence quality features: nonsense-mediated decay (NMD), read-through, reverse orientation

A total of 269 transcripts were annotated as candidates of read-through and 2731 as targets of NMD by the extended sequence quality annotation.

Category VII: pseudogene candidates

To annotate transcribed pseudogene candidates, we did the following: First, we filtered out the functional protein-coding genes by only targeting representative category II transcripts and those identified to have frame shifts and/or nonsense mutations; Second, we predicted transcribed pseudogene candidates based on a support vector machine (SVM) method. In the current release, we annotated 1112 transcribed pseudogene candidates (Category VII).

Annotation of gene families/groups

We annotated four selected gene families/groups: T-cell receptor (TCR), Immunoglobulin (Ig), Major Histocompatibility Complex (MHC) or Human Leukocyte Antigen (HLA) and Olfactory receptor (OR) using the original pipeline based on sequence analysis against genome and protein databases complemented by a text-mining approach. In the current release, we identified 15 TCR, 21 Ig, 72 MHC and 122 OR gene clusters.

All the annotation items and features of H-Inv transcript sequences are stored and shown in the main views or sub-databases in H-InvDB.

COMPREHENSIVE ANNOTATION RESOURCES IN H-InvDB

The current H-InvDB annotation resources consist of two main views, Transcript view and Locus view, and eight sub-databases: the DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group view with the appropriate cross-links. An overview of the comprehensive annotation resources of the human gene and transcripts in H-InvDB is shown in Figure 1.

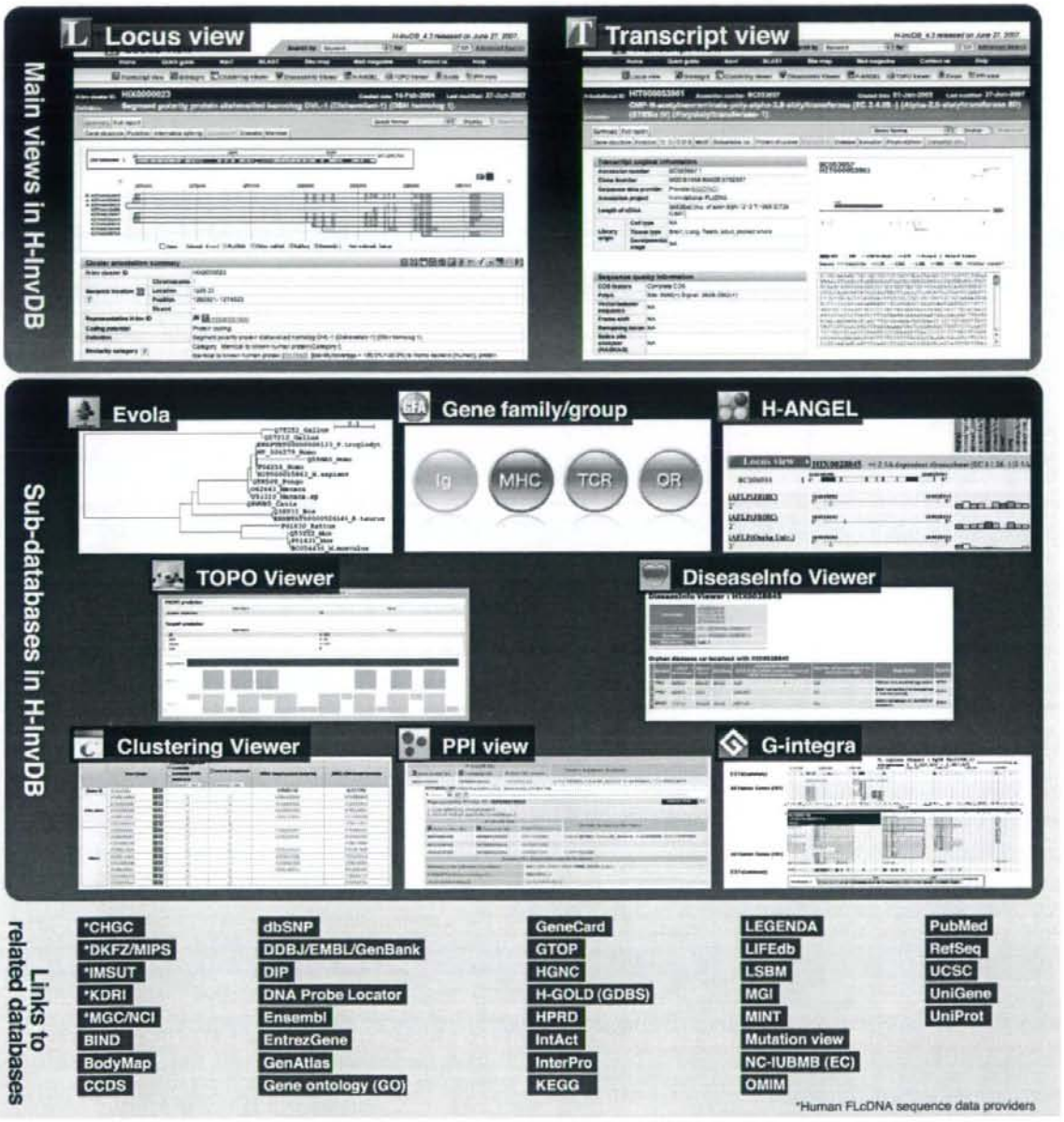


Figure 1. H-InvDB: overview of the comprehensive annotation resource for the human genes and transcripts. The current H-InvDB annotation resources consist of two main views, Transcript view and Locus view, and eight sub-databases: the DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group view. The Transcript view and the Locus view are the main viewers to display the annotation of each H-Invitational transcript (HIT) and H-Invitational cluster (HIX). The DiseaseInfo Viewer, H-ANGEL, the Clustering Viewer, G-integra, the TOPO Viewer, Evola, the PPI view and the Gene family/group view are sub-databases to provide detailed annotation for each annotation feature. The links to related databases are provided from the appropriate viewers.

Transcript view

The transcript view shows all the annotation of the H-Inv transcript in 12 section tabs: (i) gene structure, (ii) gene function, (iii) gene ontology, (iv) predicted CDS, (v) functional motif, (vi) sub cellular localization, (vii) protein structure information, (viii) gene expression, (ix) disease/pathology, (x) evolutionary information, (xi) polymorphism (SNP, indel and microsatellite) and

interspersed repeat information and (xii) transcript and sequence quality information. As seen in the example of a transcript view shown in Figure 1, this view also has links to many external public databases including DDBJ/EMBL/GenBank, RefSeq, UniProtKB, HGNC, InterPro, Ensembl, EntrezGene, PubMed, dbSNP, GO and GTOP and to web sites of the original data producers of the FLcDNA clones and sequences including the Chinese National Human Genome Center (CHGC), German cDNA Consortium (DKFZ/MIPS), Helix Research Institute, Inc. (HRI), the Institute of Medical Science in the University of Tokyo (IMSUT), the Kazusa DNA Research Institute (KDRI), the Mammalian Gene Collection (MGC/NCI) and NEDO. This view was previously known as the cDNA view (mRNA view).

Locus view

The Locus view shows all the annotation of a locus in six section tabs: (i) gene structure and location in the human genome, (ii) gene function, (iii) alternative splicing pattern, (iv) gene expression, (v) disease/pathology and (vi) cluster member information. As seen in the example of a Locus view shown in Figure 1, it shows links to external public databases including DDBJ/EMBL/GenBank, RefSeq, EntrezGene, GeneCards, HGNC and OMIM.

DiseaseInfo Viewer

The DiseaseInfo Viewer is a database of known and orphan genetic diseases and their relation to H-Inv clusters with EntrezGene and OMIM cross-links. The DiseaseInfo Viewer provides two kinds of disease information related to H-Inv clusters: known disease-related genes and co-localized orphan diseases. An orphan disease is defined as a disease mapped on a chromosomal region, but for which the responsible gene has not been identified yet. Co-localization does not necessarily mean a direct relationship between gene and disease; however, genes that are cytogenetically co-localized with a disease could be possible candidate genes for that disease. The co-localized H-Inv clusters are chosen by computing the physical range of each cytogenetic band with a 1 Mbp margin.

Human anatomic gene expression library (H-ANGEL)

H-ANGEL is a database of expression patterns that we constructed to obtain a broad outline of such patterns for human genes (12). We collected gene-expression data in normal and adult human tissues that were generated by three types of methods and in seven different platforms, including: iAFLP, a PCR-based quantitative expression profiling method; DNA arrays (long oligomers, short oligomers and cDNA microarrays); and cDNA sequence tags (SAGE, EST, BodyMap and MPSS). The H-ANGEL database comprises the largest and most comprehensive collection of gene expression patterns so far, which also provides a classification of human genes in terms of their expression.

Clustering Viewer

The Clustering Viewer facilitates the comparisons of different clustering. It allows users to see whether H-Inv transcripts are consistently clustered by different clustering methods. It also displays multiple alignments of transcripts by using CLUSTALW (13). The Clustering Viewer shows all the member transcripts of an H-Inv cluster to which a query sequence belongs.

G-integra

G-integra is an integrated genome browser, in which we can examine the genomic structures of the transcripts. As seen in an example view in Figure 1, the location in the human genome and gene structure of H-Inv transcript (green), and the corresponding RefSeq and Ensembl entries are shown. The structures of the genes and transcripts for 11 non-human species, *Pan troglodytes* (chimpanzee), *Macaca* sp. (macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (opossum), *Gallus gallus* (chicken), *Danio rerio* (zebrafish), *Tetraodon nigroviridis* (tetraodon) and *Takifugu rubripes* (fugu) can be optionally displayed for comparison. Other options allow the, the results of gene prediction programs such as GenScan (14), HMMgene (15), FGENESH (16) and JIGSAW (17) to be displayed.

TOPO Viewer

The TOPO Viewer is a tool for viewing subcellular targeting signals predicted by TargetP (18) and the presence of transmembrane helices predicted by SOSUI (19) and TMHMM(20). The probabilities that a protein may be delivered to up to nine distinct sub cellular locations are predicted by WoLF PSORT (21). TargetP predicts whether a protein contains a signal peptide, a mitochondrial targeting signal or any other type of signal. The TOPO Viewer consists of four tab pages: TABLE, MAP, FILE and GFP. The TABLE tab page displays the prediction results for all the programs used.

Evola

Evola is a database of evolutionary annotation of human genes (22). It provides sequence alignments and phylogenetic trees of manually curated orthologous genes among human and 11 model organisms, *Pan troglodytes* (chimpanzee), *Macaca sp.* (macaque), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Canis familiaris* (dog), *Bos taurus* (cow), *Monodelphis domestica* (opossum), *Gallus gallus* (chicken), *Danio rerio* (zebra fish), *Tetraodon nigroviridis* (tetraodon) and *Takifugu rubripes* (fugu). Sequence alignments and phylogenetic trees of the orthologous genes and homologous genes are shown in Evola.

PPI view

The PPI view displays H-InvDB human PPI information at http://www.jbirc.aist.go.jp/hinv/ppi/. We collected PPI data from five databases; BIND, DIP, MINT, HPRD and IntAct, removed redundancies of the PPI data among the

databases based on their sequence similarities and integrated them with the H-Invitational proteins.

Gene family/Group view

The Gene family/Group view provides human-curated annotation datasets for the selected gene families/groups at http://www.jbirc.aist.go.jp/hinv/ahg-db/geneFamily Index.jsp. For H-InvDB release 4.0, we provided detailed annotations for four selected gene families/groups: TCR, Ig, MHC and OR. Each page provides the list of genes, gene names, definitions and links for the appropriate H-InvDB views.

H-InvDB New Identifier

We defined and assigned a unique identifier for each annotation unit, transcript, protein or cluster (7,8). The identifier for H-Invitational transcript is 'HIT', prefix HIT plus nine digit numbers (e.g. HIT000000001) and for H-Invitational cluster is 'HIX', prefix HIX plus seven digit numbers (e.g. HIX0000001). In order to identify the modification in sequence or annotation of an H-Inv entry, a version is assigned to each ID and always stated with the ID. Additionally, we now provide a new identifier for each H-Invitational protein, 'HIP', prefix HIP with nine digit numbers (e.g. HIP000000001).

H-InvDB Data Availability

H-InvDB is freely available for both academic and commercial use and can be accessed online at http://www.h-invitational.jp/(or hinv.jp). Annotated data can also be downloaded in FASTA sequence files, the original-format flat files or XML files at HTTP and FTP servers. The mirror database is also available at http://hinvdb.ddbj.nig.ac.jp/. Minor updates are released every three months and major updates are released once a year.

ACKNOWLEDGEMENTS

We acknowledge all the members of the H-Invitational 2 consortium and Genome Information Integration Project (GIIP), especially the staffs of JBIRC for construction of H-InvDB, Ryo Aono, Tomohiro Endo, Yukie Makita, Hiromi Kubooka, Yuji Shinso, Harutoshi Maekawa, Yasuhiro Fukunaga, Hajime Nakaoka, Yoshito Ueki, Yoshihide Mimiura, Ryuzou Matsumoto, Seigo Hosoda, Yo Takahashi, Taichirou Sugisaki, Hiroki Hokari, Hiroaki Kawashima, Yasuhiro Imamizu, Makoto Ogawa for their technical assistance. 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). Also, this work is partly supported by the Research Grant for the RIKEN Genome Exploration Research Project from MEXT to Y.H. and the Grant for the RIKEN Frontier Research System, Functional RNA research program. Funding to pay the Open Access publication charges for this article was provided by JBIC.

Conflict of interest statement. None declared.

REFERENCES

1. Ota,T. *et al.* (1997) Full-length cDNA project toward a high throughput functional analysis. *Microb. Comp. Genomics*, **2**, 204–205.
2. Yudate,H.T. *et al.* (2001) HUNT: launch of a full-length cDNA database from the helix research institute. *Nucleic Acids Res.*, **29**, 185–188.
3. Wiemann,S. *et al.* (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422–435.
4. Strausberg,R.L. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
5. Kikuno,R. *et al.* (2002) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.*, **30**, 166–168.
6. Carninci,P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
7. Frith,M.C. *et al.* (2006) Pseudo-messenger RNA: phantoms of the transcriptome. *PLoS Genet*, **2**, p. e23.
8. Gingeras,T.R. *et al.* (2007) Origin of phenotypes: genes and transcripts. *Genome Res.*, **17**, 682–690.
9. Imanishi,T. *et al.* (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, 856–875.
10. Yamasaki,C. *et al.* (2005) Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene*, **364**, 99–107.
11. Mulder,N.J. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**(Database issue), D224–D228.
12. Tanino,M. *et al.* (2005) The human anatomic gene expression library (H-ANGEL), the H-Inv integrative display of human gene expression across disparate technologies and platforms. *Nucleic Acids Res.*, **33**(Database Issue), D567–D572.
13. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
14. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
15. Krogh,A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 179–186.
16. Salamov,A.A. and Solovyev,V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res.*, **10**, 516–522.
17. Allen,J.E. and Salzberg,S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.
18. Emanuelsson,O. *et al.* (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
19. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
20. Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
21. Horton,P. *et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**(Web Server issue), W585–W587.
22. Matsuya,A. *et al.* (2008) Evola: ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees. *Nucleic Acids Res.* (in press).

LIST OF AUTHORS FOR THE GENOME INFORMATION INTEGRATION PROJECT AND H-INVITATIONAL 2 CONSORTIUM

Chisato Yamasaki[1,2], Katsuhiko Murakami[1,2], Yasuyuki Fujii[3], Yoshiharu Sato[1,2], Erimi Harada[1,2], Jun-ichi Takeda[1,2], Takayuki Taniya[1,2], Ryuichi Sakate[1,2], Shingo Kikugawa[1,2], Makoto Shimada[1,2], Motohiko Tanino[4], Kanako O. Koyanagi[5], Roberto A. Barrero[6], Craig Gough[1,2], Hong-Woo Chun[1,2], Takuya Habara[1], Hideki Hanaoka[7], Yosuke Hayakawa[1,8], Phillip B. Hilton[1,2], Yayoi Kaneko[9], Masako Kanno[1,2], Yoshihiro Kawahara[1,2], Toshiyuki Kawamura[10], Akihiro Matsuya[1,11], Naoki Nagata[12], Kensaku Nishikata[1,13], Akiko Ogura Noda[1,2], Shin Nurimoto[14], Naomi Saichi[1,2], Hiroaki Sakai[15], Ryoko Sanbonmatsu[1,2], Rie Shiba[1,2], Mami Suzuki[1,2], Kazuhiko Takabayashi[8], Aiko Takahashi[1,2], Takuro Tamura[16], Masayuki Tanaka[1,2], Susumu Tanaka[17], Fusano Todokoro[1,18], Kaori Yamaguchi[1], Naoyuki Yamamoto[1,19], Toshihisa Okido[20], Jun Mashima[20], Aki Hashizume[20], Lihua Jin[20], Kyung-Bum Lee[20], Yi-Chueh Lin[20], Asami Nozaki[20], Katsunaga Sakai[20], Masahito Tada[20], Satoru Miyazaki[21], Takashi Makino[22], Hajime Ohyanagi[20,23], Naoki Osato[20], Nobuhiko Tanaka[20], Yoshiyuki Suzuki[20], Kazuho Ikeo[20], Naruya Saitou[24], Hideaki Sugawara[20], Claire O'Donovan[25], Tamara Kulikova[25], Eleanor Whitfield[25], Brian Halligan[26], Mary Shimoyama[26], Simon Twigger[26], Kei Yura[27], Kouichi Kimura[28], Tomohiro Yasuda[28], Tetsuo Nishikawa[28,29], Yutaka Akiyama[30], Chie Motono[30], Yuri Mukai[30], Hideki Nagasaki[15,30], Makiko Suwa[30], Paul Horton[30], Reiko Kikuno[31], Osamu Ohara[31], Doron Lancet[32], Eric Eveno[33,34], Esther Graudens[33,34], Sandrine Imbeaud[33,34,35], Marie Anne Debily[33,34,36], Yoshihide Hayashizaki[37,38], Clara Amid[39], Michael Han[39], Andreas Osanger[39], Toshinori Endo[5], Michael A. Thomas[40], Mika Hirakawa[41], Wojciech Makalowski[42], Mitsuteru Nakao[43], Nam-Soon Kim[44], Hyang-Sook Yoo[44], Sandro J. De Souza[45], Maria de Fatima Bonaldo[46], Yoshihito Niimura[47], Vladimir Kuryshev[48], Ingo Schupp[48], Stefan Wiemann[48], Matthew Bellgard[6], Masafumi Shionyu[49], Libin Jia[50], Danielle Thierry-Mieg[51], Jean Thierry-Mieg[51], Lukas Wagner[51], Qinghua Zhang[34,52], Mitiko Go[53], Shinsei Minoshima[54], Masafumi Ohtsubo[54], Kousuke Hanada[55], Peter Tonellato[56], Takao Isogai[29], Ji Zhang[34,57], Boris Lenhard[58], Sangsoo Kim[59], Zhu Chen[34,60,61], Ursula Hinz[62], Anne Estreicher[62], Kenta Nakai[63], Izabela Makalowska[64], Winston Hide[65], Nicola Tiffin[65], Laurens Wilming[66], Ranajit Chakraborty[67], Marcelo Bento Soares[68], Maria Luisa Chiusano[69], Yutaka Suzuki[70], Charles Auffray[33,34], Yumi Yamaguchi-Kabata[2], Takeshi Itoh[2,15], Teruyoshi Hishiki[2], Satoshi Fukuchi[20], Ken Nishikawa[20], Sumio Sugano[2,70], Nobuo Nomura[2], Yoshio Tateno[20], Tadashi Imanishi[2,5,†] and Takashi Gojobori[2,20]

[1]Japan Biological Information Research Center, Japan Biological Informatics Consortium, [2]Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, [3]Graduate School Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, [4]DNA Chip Research Inc., Kanagawa, [5]Hokkaido University, Hokkaido, Japan, [6]Centre for Comparative Genomics, Murdoch University, WA, Australia, [7]Biotechnology Research Center, The University of Tokyo, [8]Hitachi Software Engineering Co., Ltd., [9]Mitsubishi Kagaku Institute of Life Sciences, [10]Fujitsu Limited, Tokyo, [11]Hitachi, Co., Ltd., Saitama, [12]Japan Science and Technology Agency, [13]NEC Soft, Ltd., [14]Mitsui Knowledge Industry Co., Ltd, Tokyo, [15]National Institute of Agrobiological Sciences, Ibaraki, [16]BITS Co., Ltd., Shizuoka, [17]Tokyo Institute of Psychiatry, Tokyo, [18]DYNACOM Co., Ltd., Chiba, [19]C's Lab Co., Ltd., Hokkaido, [20]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, [21]Tokyo University of Science, Chiba, Japan, [22]University of Dublin, Trinity College, Dublin, Ireland, [23]Mitsubishi Space Software Co., Ltd., Ibaraki, [24]Division of Population Genetics, National Institute of Genetics, Shizuoka, Japan, [25]EMBL Outstation-Hinxton, European Bioinformatics Institute, Cambridge, UK, [26]Bioinformatics Research Center, Medical College of Wisconsin, WI, USA, [27]Center for Computational Science and Engineering, Japan Atomic Energy Agency, Kyoto, [28]Central Research Laboratory, Hitachi Ltd., [29]Reverse Proteomics Research Institute, CO., Ltd., [30]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, [31]Department of Human Gene, Kazusa DNA Research Institute, Chiba, Japan, [32]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel, [33]Genexpres, Functional Genomics and Systems Biology for Health (CNRS and Pierre & Marie Curie University - Paris VI), Villejuif, France, [34]Sino-French Laboratory in Life Sciences and Genomics, Shanghai, China, [35]Centre de Génétique Moléculaire, CNRS and Gif/Orsay DNA Microarray Platform, Gifs/Yvette, [36]Laboratory of Genomes Functional Exploration, CEA, DSV, IRCM, Evry, France, [37]Genomic Sciences Center, RIKEN Yokohama Institute, Kanagawa, [38]Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, Saitama, Japan, [39]GSF - National Research Center for Environment and Health, Institute for Bioinformatics,

Neuherberg, Germany, [40]Idaho State University, ID, USA, [41]Institute for Chemical Research, Kyoto University, Kyoto, Japan, [42]Institute of Bioinformatics, University of Muenster, Muenster, Germany, [43]Kazusa DNA Research Institute, Chiba, Japan, [44]Korea Research Institute of Bioscience & Biotechnology, Taejeon, Korea, [45]Ludwig Institute for Cancer Research, Sao Paulo, Brazil, [46]Medical Education and Biomedical Research Facility, University of Iowa, IA, USA, [47]Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan, [48]Molecular Genome Analysis, German Cancer Research Center, Heidelberg, Germany, [49]Nagahama Institute of Bio-Science and Technology, Shiga, Japan, [50]National Cancer Institute, National Institutes of Health, MD, [51]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, MD, USA, [52]National Engineering Center for Biochips at Shanghai, Shanghai, China, [53]Ochanomizu University, Tokyo, [54]Photon Medical Research Center, Hamamatsu University School of Medicine, Shizuoka, [55]Plant Science Center, RIKEN Yokohama Institute, Kanagawa, [56]Harvard Medical School, MA, USA, [57]Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [58]Center for Genomics and Bioinformatics, Karolinska Institute, Stockholm, Sweden, [59]Soongsil University, Seoul, Korea, [60]State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, Rui Jin Hospital, Shanghai Jiao Tong University School of Medicine, [61]Chinese National Human Genome Center at Shanghai, Shanghai, China, [62]Swiss Institute of Bioinformatics, Geneva, Switzerland, [63]The Institute of Medical Science, The University of Tokyo, Tokyo, Japan, [64]The Pennsylvania State University, PA, USA, [65]The South African National Bioinformatics Institute, University of Western Cape, Cape Town, South Africa, [66]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK, [67]University of Cincinnati, OH, [68]Children's Memorial Research Center, Northwestern University, Feinberg School of Medicine, USA, [69]University of Naples "Federico II", Naples, Italy and [70]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

[†]To whom correspondence should be addressed. Tel: +81-3-3599-8800; Fax: +81-3-3599-8801; E-mail: 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

Published online 3 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* **D787–D792**
doi:10.1093/nar/gkm878

Evola: Ortholog database of all human genes in H-InvDB with manual curation of phylogenetic trees

Akihiro Matsuya[1,2], Ryuichi Sakate[1,3,*], Yoshihiro Kawahara[1,3], Kanako O. Koyanagi[4], Yoshiharu Sato[1,3], Yasuyuki Fujii[1,3], Chisato Yamasaki[1,3], Takuya Habara[1,3], Hajime Nakaoka[5], Fusano Todokoro[1,6], Kaori Yamaguchi[1,3], Toshinori Endo[4], Satoshi OOta[7], Wojciech Makalowski[8], Kazuho Ikeo[9], Yoshiyuki Suzuki[9], Kousuke Hanada[9], Katsuyuki Hashimoto[10], Momoki Hirai[11], Hisakazu Iwama[12], Naruya Saitou[13], Aiko T. Hiraki[1,3], Lihua Jin[9], Yayoi Kaneko[1,3], Masako Kanno[1,3], Katsuhiko Murakami[1,3], Akiko Ogura Noda[1,3], Naomi Saichi[1,3], Ryoko Sanbonmatsu[1,3], Mami Suzuki[1,3], Jun-ichi Takeda[1,3], Masayuki Tanaka[1,3], Takashi Gojobori[3,9], Tadashi Imanishi[3] and Takeshi Itoh[3,14]

[1]Integrated Database Group, Japan Biological Information Research Center, Japan Biological Informatics Consortium, [2]Government & Public Corporation Information Systems, Hitachi, Co., Ltd., [3]Integrated Database Group, Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, [4]Graduate School of Information Science and Technology, Hokkaido University, Hokkaido, [5]C's Lab Co., Ltd., Hokkaido, [6]DYNACOM Co., Ltd, Chiba, [7]BioResource Center, RIKEN, Ibaraki, Japan, [8]Institute of Bioinformatics, University of Muenster, Muenster, Germany, [9]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Shizuoka, [10]Department of Biomedical Resources, National Institute of Biomedical Innovation, Osaka, [11]International Research and Educational Institute for Integrated Medical Sciences, Tokyo Women's Medical University, Tokyo, [12]Kagawa University, Kagawa, [13]Department of Population Genetics, National Institute of Genetics, Shizuoka and [14]Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, Ibaraki, Japan

Received August 15, 2007; Revised September 27, 2007; Accepted October 1, 2007

ABSTRACT

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Currently, with the rapid growth of transcriptome data of various species, more reliable orthology information is prerequisite for further studies. However, detection of orthologs could be erroneous if pairwise distance-based methods, such as reciprocal BLAST searches, are utilized. Thus, as a sub-database of H-InvDB, an integrated database of annotated human genes (http://h-invitational.jp/), we constructed a fully curated database of evolutionary features of human genes, called 'Evola'. In the process of the ortholog detection, computational analysis based on conserved genome synteny and transcript sequence similarity was followed by manual curation by researchers examining phylogenetic trees. In total, 18968 human genes have orthologs among 11 vertebrates (chimpanzee, mouse, cow, chicken, zebrafish, etc.), either computationally detected or manually curated orthologs. Evola provides amino acid sequence alignments and phylogenetic trees of orthologs and homologs. In 'd_N/d_S view', natural selection on genes can be analyzed between human and other species. In 'Locus maps', all transcript variants and their exon/intron structures can be compared among orthologous gene loci. We expect the Evola to serve as a comprehensive and reliable database to be utilized in comparative analyses for obtaining new knowledge about human genes. Evola is available at http://www.h-invitational.jp/evola/.

INTRODUCTION

A large number of genome and transcript sequences accumulated in the last decade give us an opportunity

*To whom correspondence should be addressed. Tel: +81 3 3599 8800; Fax: +81 3 3599 8801; Email: rsakate@ni.aist.go.jp

for large-scale comparative analyses. In particular, detection of orthologs, groups of genes in different species that evolved by speciation, accelerates functional and evolutionary studies. Despite the past efforts to develop bioinformatics methods for analyzing a large number of sequences, it is still a challenge to comprehensively identify orthologs between species. A number of automated pairwise distance-based methods for ortholog detection have been proposed, as represented by the reciprocal best BLAST hits (RBH) method (1) and the reciprocal smallest distance (RSD) method (2). However, as genes might have frequently undergone duplications and losses in evolutionary lineages leading to human (3), pairwise distance-based methods might lead to erroneous inferences of phylogenetic relationships and thus of orthologs. Thus, phylogenetic tree-based detection can be the most plausible solution to provide more reliable orthologs.

Here this database 'Evola', a sub-database complementary to the H-Invitational database (H-InvDB), was developed to provide orthology information for the originally annotated human genes in H-InvDB. Evola features its ortholog detection in which genome synteny-based computational analysis was followed by manual curation of molecular phylogenetic trees. Evola differs in this way from other ortholog databases such as Inparanoid (4), Ensembl-Compara (5), Homologene (6), HOGENOM (7) and TreeFam (8). These databases are based on BLAST hits (Inparanoid), BLAST hits and synteny (Ensembl-Compara and Homologene) and phylogenetic trees (HOGENOM and TreeFam). The concept of Evola is that genomic region (gene locus) is a unit of genes that are duplicated or lost. In collaboration with H-InvDB, Evola enables users to compare gene structure, transcript variants, upstream/downstream region of the genome among species.

H-InvDB is an integrated database of annotated human genes providing annotation of human full-length enriched cDNAs (9,10,11). At the meetings of the Human Full-Length cDNA Annotation Invitational held in Japan (2002 and 2003), Evola started with H-InvDB to annotate evolutionary features of the human genes. With several updates afterwards and a subsequent All Human Genes Evolutionary Annotation (AHG-EV) meeting in 2006, the current strategy of evolutionary annotation (computational analysis and manual curation) in Evola has been established. Orthology information for human and other 11 vertebrates is currently included in the Evola: human, chimpanzee, macaque, mouse, rat, dog, cow, opossum, chicken, zebrafish, Tetraodon and Fugu. Several visualization tools are incorporated into the database, including sequence alignment viewer, natural selection plot and graphical representation of orthologous gene loci among different species. Evola is now one of the databases listed in the Comparison of Orthology Predictions project of the HUGO Gene Nomenclature Committee (HGNC, http://www.genenames.org/).

ORTHOLOG DETECTION

Computational analysis: Ortholog detection based on conserved genomic synteny and pairwise distance

Species for ortholog detection were selected with consideration of completeness of their genome assemblies (chromosome level), abundance of transcript sequences (~20 000) and importance in biology (intensively studied or a representative of a phylogenetic clade). Whole genome sequence assemblies of human (hg18), chimpanzee (panTro2), macaque (rheMac2), mouse (mm8), rat (rn4), dog (canFam2), cow (rn4), opossum (monDom4), chicken (galGal3), zebrafish (danRer4), Tetraodon (tetNig1) and Fugu (fr1) were downloaded from UCSC (http://genome.ucsc.edu/). Conserved syntenic regions were detected by a modified pairwise genome alignment method (12) using BLASTZ (13) with the options of C = 2, T = 4, Y = 3400 between human and other primates (between more similar genome sequences), and C = 2 between human and non-primate vertebrates (between less similar genome sequences).

For human transcripts, H-InvDB representative transcripts (HITs) were used. Other vertebrates' transcripts (mRNAs) were downloaded from DDBJ (http://www.ddbj.nig.ac.jp/) release 66, Ensembl (http://www.ensembl.org/) release 38 and RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) release 17, and their genomic locations (one location per transcript) were detected on cognate genomes by a hybrid method using BLAT (14), BLAST (15) and est2genome (16) as they were used to detect genomic locations of human transcripts in H-InvDB. Representative transcripts (one transcript per gene locus) were determined in consideration of percent identity and coverage to the genome, number of exons, etc. of all transcripts in each locus (9,10,11). Thus, in Evola, representative transcripts were defined as genes.

Lengths of overlapping exons of each gene pair between human and other species were calculated in the genome alignment. A gene pair with the maximum length was selected as the best assignment (not a minimum length was defined). Every gene in a species was assigned to a gene in the other species. If two human genes were assigned to one mouse gene, this was defined as a two-to-one ortholog. As a result, Evola contains not only one-to-one orthologs but also many-to-many orthologs. For all the assignment pairs, coding sequences (CDSs) and amino acid (a.a.) sequences of other species were predicted by FASTY (17). They were predicted by comparing with the amino acid sequences of the corresponding human genes. Finally, if the length of the alignable region between human and other species ortholog candidates was $\geq$80 a.a., they were defined as computationally detected orthologs.

Manual curation: Examination of phylogenetic trees by experts

Homologs of human genes (amino acid sequences) were obtained from UniProt (http://www.uniprot.org/) and human RefSeq (NP) by FASTY similarity searches with the option of E-value of <1e−5. For each human gene, a