Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/89366

Please be advised that this information was generated on 2022-08-22 and may be subject to change.

DNA Research 17, 197–208, (2010)
Advance Access Publication: 1 April 2010

doi:10.1093/dnares/dsq010

Mutation@A Glance: An Integrative Web Application for Analysing Mutations from Human Genetic Diseases

Atsushi Hijikata[1], Rajesh Raju[2,3,4], Shivakumar Keerthikumar[2,3,4], Subhashri Ramabadran[2,3], Lavanya Balakrishnan[2,3], Suresh Kumar Ramadoss[2], Akhilesh Pandey[3,5], Sujatha Mohan[2,3], and Osamu Ohara[1,6,*]

Laboratory for Immunogenomics, RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan[1]; Research Unit for Immunoinformatics, RIKEN Research Center for Allergy and Immunology, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan[2]; Institute of Bioinformatics, International Technology Park, Bangalore 560 066, India[3]; Department of Biotechnology and Bioinformatics, Kuvempu University, Jnanasahyadri, Shimoga 577 451, India[4]; McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, 733 N. Broadway, BRB Room 527, Baltimore, MD 21205, USA[5] and Laboratory of Genome Technology, Department of Human Genome Research, Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan[6]

*To whom correspondence should be addressed. Tel. +81 45-503-9696. Fax. +81 45-503-9694. Email: oosamu@rcai.riken.jp

Edited by Katsumi Isono
(Received 17 January 2010; accepted 8 March 2010)

Abstract

Although mutation analysis serves as a key part in making a definitive diagnosis about a genetic disease, it still remains a time-consuming step to interpret their biological implications through integration of various lines of archived information about genes in question. To expedite this evaluation step of disease-causing genetic variations, here we developed Mutation@A Glance (http://rapid.rcai.riken.jp/mutation/), a highly integrated web-based analysis tool for analysing human disease mutations; it implements a user-friendly graphical interface to visualize about 40 000 known disease-associated mutations and genetic polymorphisms from more than 2600 protein-coding human disease-causing genes. Mutation@A Glance locates already known genetic variation data individually on the nucleotide and the amino acid sequences and makes it possible to cross-reference them with tertiary and/or quaternary protein structures and various functional features associated with specific amino acid residues in the proteins. We showed that the disease-associated missense mutations had a stronger tendency to reside in positions relevant to the structure/function of proteins than neutral genetic variations. From a practical viewpoint, Mutation@A Glance could certainly function as a 'one-stop' analysis platform for newly determined DNA sequences, which enables us to readily identify and evaluate new genetic variations by integrating multiple lines of information about the disease-causing candidate genes.

Key words: genetic disease; mutation; polymorphism; bioinformatics; protein structure

1. Introduction

Genetic diseases are caused by structural changes in genes and/or chromosomes. In the Online Mendelian Inheritance in Man (OMIM, http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim) database, more than 2200 genes are known to have mutations causing genetic diseases.[1] For instance, primary immunodeficiency diseases (PIDs) are caused by congenital defects in genes involved in the development and maintenance of the immune system,[2,3] and they can be diagnosed using mutation analysis that identifies pathogenic mutations in candidate PID genes. This process plays a critical role in improving

© The Author 2010. Published by Oxford University Press on behalf of Kazusa DNA Research Institute.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

the quality of life for PID patients.[4] In this regard, the recent advances in DNA sequencing technology will extremely expedite this process. Thus, the next bottleneck to be addressed is obviously how to clarify the associations between newly identified patient-specific genetic variations and disease phenotypes, even when familial disease history is absent. To eliminate the bottleneck in mutation analysis, we need a bioinformatics tool that would enable us to readily evaluate the impact of a genetic variation on the structure/function of a gene product at the molecular level. Towards this end, our first step was to develop an integrated 'one-stop' analysis platform where we could cross-reference multiple lines of information regarding known genetic variations, including a huge amount of non-synonymous (ns) single-nucleotide polymorphisms (nsSNPs) in healthy individuals,[5–7] in genes of interest.

Bioinformatics resources and methods played an indispensable role in creating this platform.[8–12] Although a number of databases regarding reported human disease mutations and SNPs have been already constructed,[13–25] these databases were launched as a static archive for genetic variation data, not necessarily an interactive tool for evaluating newly identified sequence variation data. Several computational algorithms for predicting the effects of ns substitutions on a corresponding protein have been developed using evolutionary and protein three-dimensional (3D) structure information.[26–31] However, despite public availability of these software/web servers, there are at least two hurdles, especially for clinical researchers to exploit them for the mutation analysis: (i) since these servers usually require information about the position of the genetic variation occurred in a submitted sequence as a query input, the users have to specify the variation position in the sequence before submitting the query; (ii) since these servers do not necessarily incorporate known disease-associated mutation data into their systems, the users have to manually compare their newly identified genetic variation data from patients with previously reported data. Thus, we thought it was important to integrate predictive bioinformatics tools, such as the one described above, with a comprehensive set of known genetic variation data, to create a 'one-stop' mutation analysis platform.[32]

In this context, here we present Mutation@A Glance (http://rapid.rcai.riken.jp/mutation/), a new web-based integrated bioinformatics tool for analysing mutations from human genetic diseases. The user-friendly graphical interface of Mutation@A Glance makes it possible to allocate known disease-associated mutation data on the nucleotide and amino acid sequences of a gene of interest, and to link these mutation data to the 3D structure of the gene product along with various lines of information about the mutated amino acid residues (e.g. the extent of evolutional sequence conservation, post-translational modifications and molecular interactions). Furthermore, this tool enables users to identify and evaluate newly identified sequence variations in a query DNA sequence from a gene of interest by comparing them with known disease-associated mutation data and using the SIFT program,[26] which is one of the most accurate and widely used program to specifically predict the effects of ns substitutions based on evolutionary information for each residue position.[33] Therefore, Mutation@A Glance surely serves as a 'one-stop' informational platform to identify and evaluate new genetic variations by integrating multiple lines of information about the disease-causing candidate genes.

2. Materials and methods

2.1. *Data resources for disease-associated genes and sequence variations*

Human disease-associated mutation data were obtained from the following three databases: OMIM (http://www.ncbi.nlm.nih.gov/omim/),[1] UniProt (http://www.uniprot.org/)[34] and RAPID (http://rapid.rcai.riken.jp/).[17] Sequence variations that were associated with OMIM in the dbSNP database (Build 130, http://www.ncbi.nlm.nih.gov/projects/SNP/)[18] were considered to be disease-associated mutations and other variations were considered non-disease associated. For the mutation data in the UniProt database, VARIANT features associated with diseases in the human entries were considered. RAPID is a molecular database that we have recently established for reported disease mutation data in genes causing PIDs.[17] The RAPID database is directly connected to our local server and the mutation data (as of August 2009) are retrieved using a Perl script. The human genome sequence (Build 36.3), RefSeq sequences for nucleotides and proteins of human were downloaded from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/). Information regarding residue-wise functional features (Transmembrane helix, signal peptide, nucleotide binding, disulphide bond, metal binding, active site and post-translational modification site) was extracted from the human entries in the UniProt database. Information regarding the exon–intron structures of each gene was downloaded from the NCBI ftp site.

2.2. *Calculation of sequence conservation in ns substitution sites*

Homologous protein sequences in other organisms to the human proteins encoded by disease-causing genes were identified using the BLAST program[35] against the RefSeq database (6 691 817 amino acid

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

sequences) with a cut-off *E*-value of 10^{-4}. If the sequence identity and the coverage between a sequence hit and the human were higher than 40% and 80%, respectively, the sequence was selected as a homologous sequence. When two or more sequences from an organism were found as homologous sequences, the sequence with the highest sequence identity was only considered. The homologous protein sequences from various organisms were aligned using the CLUSTAL W program.[36] A degree of sequence conservation at each amino acid position in the multiple sequence alignment (simply designated as 'residue conservation' in Fig. 1) was defined as the ratio of (the number of the homologous protein sequences which carried an identical amino acid residue to that in the human sequence) to (the number of the aligned homologous protein sequences) at the specified position in the multiple sequence alignment. For example, if Ala appears in an aligned position in the human sequence and the corresponding positions in all of the other homologous sequences are also Ala, the residue conservation in this position is defined as 1.0. The frequency distribution of the residue conservations in disease-associated missense mutation or nsSNP positions for proteins analysed in this study was represented using bins of the interval of 0.2. The value in each bin was normalized by the frequency of the total number of residues in each bin.

2.3. *Protein 3D structure information*

Protein 3D structure data were downloaded from the Protein Data Bank (PDB, http://www.rcsb.org/pdb/).[37] In cases where the 3D structure of a human protein had not yet been determined, we searched the available sequences in the PDB entries for a template structure for homology modelling using the BLAST program as described above. When the alignment of the human protein sequence and a known 3D structure showed >30% identity and >90% coverage, a homology model was built using the MODELLER package.[38] For each target, 20 model structures were generated and their reliabilities were assessed with the Discrete Optimized Protein Energy (DOPE) method.[39] Eventually, the model with the best DOPE score was selected as the final model for each protein. Information about protein quaternary structures was also extracted from the PDB database. Entries from the PDB that contained information about the biological unit structure and entries with polypeptide chains showed >85% identities with a human protein sequence were considered. When a distance of one atom in a residue in a given polypeptide chain was <5.0 Å from that of another residue in the other polypeptide or nucleotide chain, the residue was considered to be located at a molecular interaction interface.

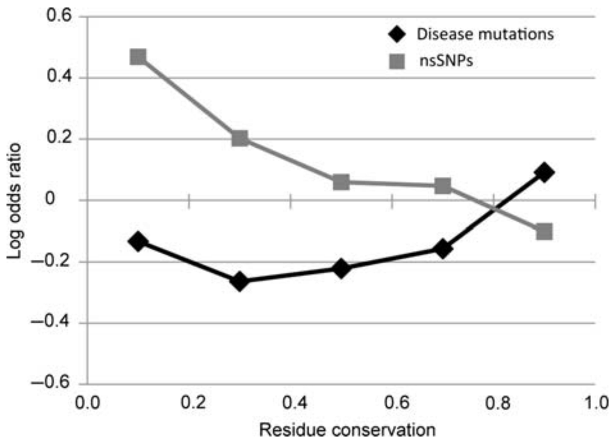


Figure 1. Comparison of frequency distributions of residue conservations in disease-associated missense mutations and nsSNPs. The vertical axis depicts the log-odds ratio of the frequency of ns substitution residue positions (disease-associated mutations or nsSNPs) to those of total number of residues in each residue conservation bin.

2.4. *Solvent accessibility calculations*

The solvent accessibilities of the amino acid residues in a 3D modelled structure were calculated using a modification of the Shrake and Rupley method,[40] with a water molecule represented by a 1.4 Å radius sphere. The solvent accessibility is represented by values ranging from 0 to 1. The residue was considered as an exposed residue on the protein surface, if the solvent accessibility was >0.25 and buried otherwise.

2.5. *Disorder prediction*

We used the DISOPRED2 program[41] to analyse each amino acid sequence of a gene product and predict intrinsically unstructured (disordered) regions in the protein sequence. If the program predicted a region consisting of more than three amino acid residues in a sequence to be 'disordered', we assigned this region as an intrinsically unstructured one.

2.6. *Predicting the effect of ns substitutions on proteins*

The effects of ns substitutions on a given protein were evaluated on a local server using the SIFT program[26] which predicts the effects of missense substitutions on a protein based on evolutionary information from homologous protein sequences.

2.7. *System implementation*

At the server end, a set of common gateway interface programs was written in Perl and is running on an Apache web server. The information regarding the disease-associated genes and the sequence variations

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

described above was integrated into a MySQL database implemented in the server. At the client end, JavaScript frameworks such as prototype.js (http://www.prototypejs.org/) and scriptaculous.js (http://script.aculo.us/) were used to make the user interface more interactive. Jmol, a Java applet (http://www.jmol.org/), was implemented for visualizing protein 3D structures in a web browser.

3. Results and discussion

3.1. *Statistics of the sequence variation data on Mutation@A Glance*

From three data resources for human disease mutations, OMIM, UniProt and RAPID, we obtained 25 616 disease-associated mutations and 21 199 nsSNPs in 2656 human genes (Table 1) and integrated into the local database. Functional classification of the proteins encoded by the disease-associated genes showed a wide variety of molecular functions such as metabolic enzymes, protein kinases, transcription factor/regulators and structural proteins (Table 1 and Supplementary Table S1). Because we have been actively analysing mutations found in patients of PIDs with paediatricians in Japan, we constructed RAPID and used it as our original data resource for genetic variations in genes responsible for PIDs.[17] RAPID contains manually curated mutation data from published literature, including nonsense (582 sites in 96 genes), frameshift (851 sites in 101 genes) and insertion/deletion (85 sites in 42 genes) mutations as well as missense mutations (1564 sites in 116 genes) in the protein-coding regions of 155 PID genes (as of August 2009). For non-PID genes, we used two publicly available data sets from UniProt and OMIM. The UniProt database contains only missense mutation data (22 258 entries in 2614 genes). On the other hand, the OMIM database contains a large number of missense mutation (1899 entries in 556 genes) and a relatively small number of the other types of mutations (99 entries in 13 genes). The RAPID and the OMIM databases also contain 699 disease-associated mutation data in intronic regions of 147 genes that cause splice anomaly effects. Thus, the most frequent mutation type in our data sets was missense mutation (89% of the total entry) as reported in the previous study.[13]

Table 1. Functional classification of disease-associated gene products

Molecular class	No. of genes	No. of mutations[a]	No. of nsSNPs
Enzymes	410	5406 (5003)	2476
Protein kinases	258	1947 (1340)	2452
Transcription factor/ regulator	239	2889 (2743)	1502
Structural proteins	132	1588 (1377)	1800
Cell surface receptors	123	1271 (1165)	838
Transport/cargo protein	116	1617 (1411)	1139
DNA/RNA binding proteins	97	429 (369)	580
Integral membrane protein	87	446 (434)	698
Channels	79	958 (948)	639
GTPase/GTPase regulators	71	371 (351)	450
Membrane transport protein	67	755 (751)	523
Immunity proteins	58	496 (183)	423
Extracellular matrix protein	53	886 (884)	939
Proteases	53	345 (284)	368
Cell adhesion molecules	52	390 (363)	428
Others	430	4647 (4091)	3863
Unclassified	331	1175 (987)	2081
Total	2656	25 616 (22 684)	21 199

[a] The numbers in parentheses indicate the number of disease-associated missense mutations.

3.2. *Evolutionary, structural and functional features of the ns substitution positions*

In general, disease-associated missense mutations tend to occur at evolutionarily conserved positions, because these positions are usually essential for the structure and/or function of a protein.[26,42,43] To verify this using the up-dated data set, we compared the frequency of disease-associated missense mutation sites (19 128 unique positions in 2622 genes) in each residue conservation bin with that of nsSNP sites (20 605 positions in 2494 genes) (Fig. 1). The results indicated that the previously reported tendency was still true for the 2622 genes in our data set; the disease-associated mutation sites were preferably appeared in the highest residue conservation bin, while nsSNP sites showed the opposite trend (Fig. 1). Next, we cross-referenced amino acid positions of the disease-associated missense mutations and nsSNPs to the functional features and 3D structures of the protein data in Mutation@A Glance. We classified these positions in terms of their functional features in a protein (annotated in the UniProt databases; Table 2). More disease-associated missense mutations were found in the positions annotated to have some functional features, except in the 'signal peptides' and 'post-translational modification sites', than nsSNPs. Using a homology modelling technique, we mapped 10 939 out of 19 128

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

Table 2. Structural and functional loci of mutation/nsSNP sites

Property	Disease mutations	nsSNPs
Transmembrane helix	1283	648
Nucleotide binding	670	102
Disulfide bond	385	39
Metal binding site	226	147
Signal peptide	101	262
Post translational modification site	97	104
Binding site[a]	48	12
Active site[a]	25	6

[a] As defined in UniProt database (described in the text).

disease-associated mutation sites (57.2%) to protein 3D structures (Fig. 2). Of these sites, 6616 sites (60.4%) were located in regions buried in protein structures (solvent accessibility <0.25). In the same way, 7106 out of 20 605 nsSNP sites (34.4%) were mapped to 3D structures, and 4258 sites (59.9%) were located on the surfaces of proteins (Fig. 2A). This observation is basically consistent with the previous findings from structural analysis.[44–46] Interestingly, nsSNP sites were located in regions predicted as intrinsically disordered at a three times higher frequency than disease-associated mutation sites (Fig. 2A). This might be ascribed to the observation that conservation in the intrinsically disordered regions is relatively lower than that in ordered regions.[47]

Proteins function with other molecules in molecular networks (e.g. signalling pathways) in many cases. Hence, the effects of mutations on molecular interactions must be intriguing in mutation

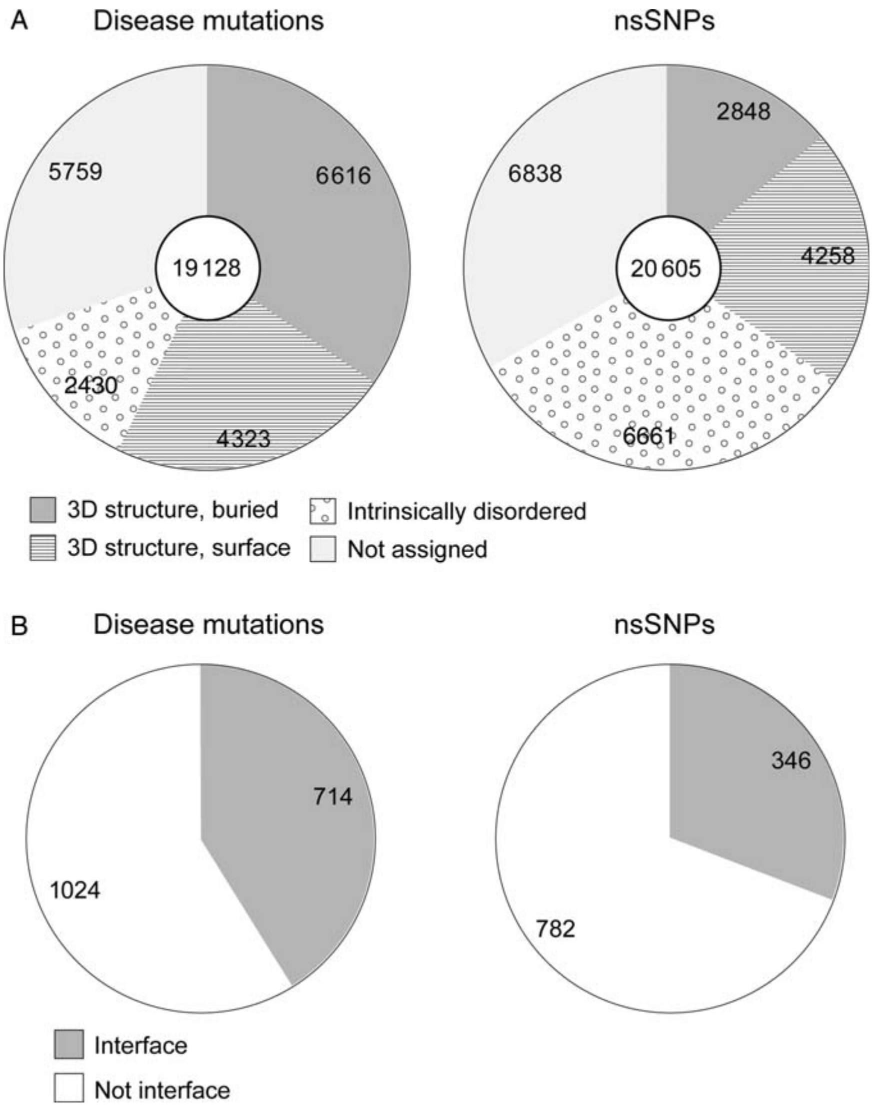


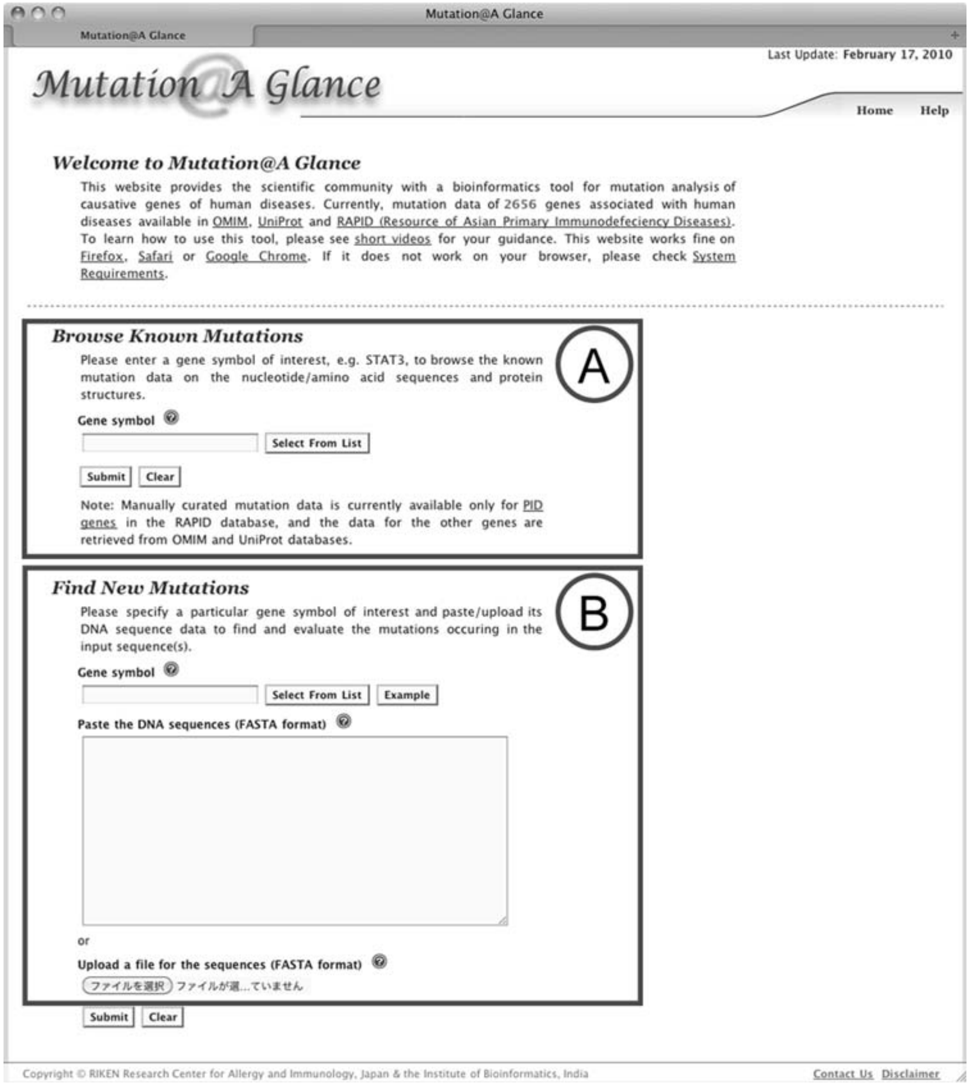
Figure 2. Classification of disease-associated mutations and nsSNPs according to their location on protein 3D structure. (A) The numbers in the pie charts depict those of ns substitution positions. (B) Proportion of ns substitution positions in the disease-associated mutations or nsSNPs that were located on the interface of the experimentally determined quaternary structures.

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

analysis.[48] We thus analysed whether or not the missense mutation positions were located in the molecular interaction sites based on the quaternary protein structures available from the PDB. Consequently, 714 out of 1738 disease-associated mutation sites (41.1%) were found to locate at the interfaces of 474 distinct proteins known to be involved in protein complex structures (Fig. 2B; see Section 2.3). In contrast, the same was true for only 346 out of 1128 nsSNP sites (30.7%) in 447 genes. We confirmed that the frequency of disease-associated mutation sites located at the molecular interaction interface was significantly higher than that of nsSNP sites by χ^2 test ($P < 0.01$). These results implicated that ns substitutions at positions involved in the molecular interaction tend to be disease-related as we expected.

3.3. *The user interface for visualizing sequence variations*

Figure 3 shows the front page of the Mutation@A Glance website. It has two types of query forms, for visualizing known disease mutation data (Fig. 3A) and for evaluating novel genetic variations in query DNA sequences (Fig. 3B). For the visualization, a user inputs a given gene symbol of interest in the form. When the user enters some characters in the form, a list of gene names containing the input character string is shown to assist the user input. In addition, a user can also search for the gene name of interest from an entire list of genes available in Mutation@A Glance, which is displayed by clicking 'Select from List' button (Fig. 3A). Just as information for users, the mutation data set used for each gene is noted near the 'Select from List' button. Figure 4 shows



Last Update: February 17, 2010

Welcome to Mutation@A Glance

This website provides the scientific community with a bioinformatics tool for mutation analysis of causative genes of human diseases. Currently, mutation data of 2656 genes associated with human diseases available in OMIM, UniProt and RAPID (Resource of Asian Primary Immunodefeciency Diseases). To learn how to use this tool, please see short videos for your guidance. This website works fine on Firefox, Safari or Google Chrome. If it does not work on your browser, please check System Requirements.

Browse Known Mutations

Please enter a gene symbol of interest, e.g. STAT3, to browse the known mutation data on the nucleotide/amino acid sequences and protein structures.

Gene symbol

Select From List

Submit Clear

Note: Manually curated mutation data is currently available only for PID genes in the RAPID database, and the data for the other genes are retrieved from OMIM and UniProt databases.

Find New Mutations

Please specify a particular gene symbol of interest and paste/upload its DNA sequence data to find and evaluate the mutations occuring in the input sequence(s).

Gene symbol

Select From List Example

Paste the DNA sequences (FASTA format)

or

Upload a file for the sequences (FASTA format)

ファイルを選択 ファイルが選...ていません

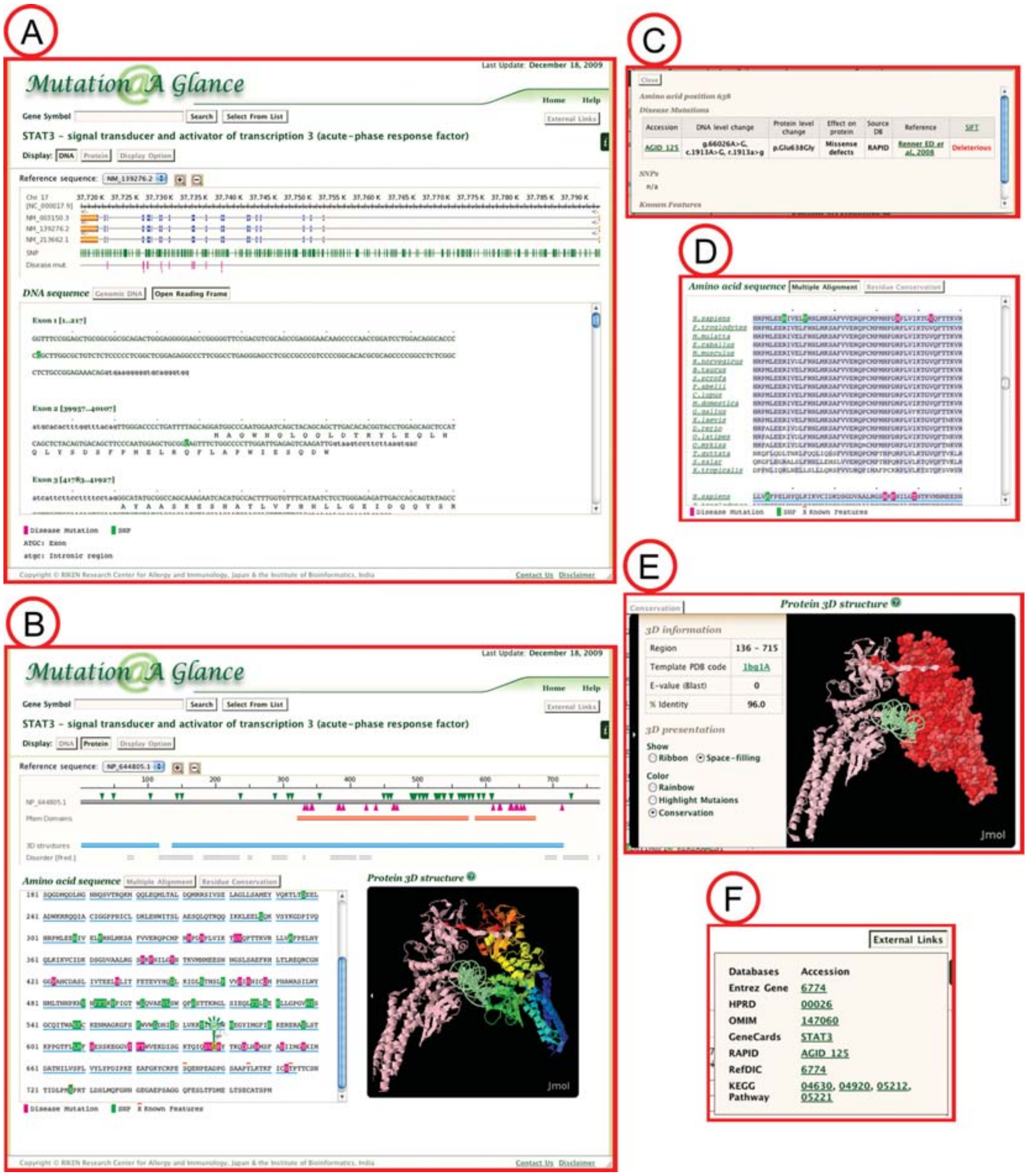
Submit Clear

Copyright © RIKEN Research Center for Allergy and Immunology, Japan & the Institute of Bioinformatics, India

Contact Us Disclaimer

Figure 3. The front page of Mutation@A Glance. There are two types of query interface for (A) browsing known mutation data and (B) evaluating novel sequence variations in DNA sequences of interest. See the main text for details of the mutation data available in Mutation@A Glance.

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015



Amino acid position 638

Disease Mutations

Accession	DNA level change	Protein level change	Effect on protein	Source DB	Reference	SIFT
AGID 125	g.66026A>G, c.1913A>G, r.1913a>g	p.Glu638Gly	Missense defects	RAPID	Renner ED et al, 2008	Deleterious

SNPs

n/a

Known Features

3D information

Region	136 - 715
Template PDB code	1bg1A
E-value (Blast)	0
% Identity	96.0

3D presentation

Show
○ Ribbon ● Space-filling

Color
○ Rainbow
○ Highlight Mutaions
● Conservation

Databases	Accession
Entrez Gene	6774
HPRD	00026
OMIM	147060
GeneCards	STAT3
RAPID	AGID 125
RefDIC	6774
KEGG Pathway	04630, 04920, 05212, 05221

Figure 4. Screenshots of Mutation@A Glance. An example of visualizing mutation data for *STAT3* is shown at the DNA (A) and the protein levels (B). The nucleotide and amino acid positions of disease-associated mutations and SNPs are coloured magenta and green, respectively. At the protein level, various types of information for the mutated amino acid residues can be viewed. (C) The detailed information about the position of nucleotide or amino acid residues selected. (D) A multiple sequence alignment of human and the other organisms STAT3 protein sequences. (E) Detailed information about the 3D structure displayed with Jmol and the representation option menu for 3D structure information. (F) External links to other website for various types of information about the gene, e.g. gene expression and signalling pathway.

sample screenshots for the *STAT3* gene, which is known to be causative to hyper-IgE syndrome (HIES).[49,50] At the DNA level, positions of the disease-associated mutations, including substitution, insertion and deletion, as well as SNPs are shown on a set of exon sequences or genomic DNA sequence with/without the open-reading frame for the gene of interest (Fig. 4A). If two or more alternative transcripts exist in the RefSeq database, the genetic variation data are allocated on the reference sequence that encodes the longest amino acid sequence among the alternative transcripts whereas all the alternative transcripts are indicated in the top panel of the genomic structure. At the protein level, the disease-associated mutation and SNP sites are highlighted in the primary structure of the gene products along with available functional annotation information of the amino acid residues from the UniProt database (e.g. enzymatic active sites and post-translational modification sites, etc.) (Fig. 4B). Information regarding

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

conserved domain from Pfam (http://pfam.sanger.ac.uk/)[51] and predicted intrinsically disordered regions are also displayed. When 3D structure information for the protein is available, the positions of mutation and SNP data can be viewed on the monomer or complex 3D structures with the Jmol applet (Fig. 4B). Detailed information about nucleotide or amino acid residues of interest is displayed in another window after clicking on a residue (Fig. 4C). In particular, at the protein level, an amino acid residue becomes highlighted in the 3D structure when clicking on it (Fig. 4B). The amino acid sequence of human can be compared with those of other organisms by clicking 'Multiple Alignment' button (Fig. 4D). The representation of the 3D structure can be selected from two model types (ribbon or space-filling models) and three colouring types (by rainbow, highlighting mutation positions or residue conservation) (Fig. 4E). The 'External Links' button provides links to NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene)[52] for general information regarding the gene, Human Protein Reference Database (http://www.hprd.org/)[53] for information about the gene product, GeneCards (http://www.genecards.org/),[54] the Reference Database of Immune Cells (http://refdic.rcai.riken.jp/)[55] for gene expression profiling data and the KEGG pathway (http://www.genome.jp/)[56] for pathways involving this gene (Fig. 4F). By using this visualization facility, mapping amino acid positions of known ns substitutions on the crystal structure of the STAT3–DNA complex (PDB code: 1bg1)[57] revealed that the disease-associated missense mutation residue positions were spatially located at the interface of the homodimer or at the DNA binding site, whereas the nsSNP residue positions were located on a surface outside of the molecular interaction sites (Fig. 5). This suggests that disease-causative missense mutations in *STAT3* directly affect the protein–protein and/or protein–DNA interaction as reported previously.[49,50] This is a good demonstration how Mutation@A Glance could help us interpret mutation effects at the molecular level.

3.4. *Evaluating the sequence variations in query sequences*

One of the issues of diagnosis of genetic diseases is how to evaluate the pathogenicity of newly identified sequence variations. To address this issue, Mutation@A Glance has an interface that allows clinical researchers to assess the impact of an observed sequence variation in a given DNA sequence for a candidate disease-causing gene as the second query form (Fig. 3B). When a user submits DNA sequences of a candidate gene in question, this tool returns a list of sequence variations found in the input DNA sequences at both the DNA and the protein levels (Fig. 6). To identify genetic variations that occur in input DNA sequences of a given gene, the BLAT program[58] is implemented to align the input DNA sequences with the reference genomic DNA sequence for the corresponding gene. Figure 6A represents the alignment status of the query sequence to the reference sequence. If a sequence variation is found, multiple lines of detailed information about the variation, such as the variation types (e.g. substitution, insertion and deletion), the mutated region (e.g. exon, intron and 5′- or 3′-splice sites constituting the GT-AG rule), the amino acid changes (e.g. missense, nonsense, insertion/deletion and frame-shift), the known variation data (disease-associated mutation and SNP) and structure/function features of the position at the protein level, are displayed based on the reference human genome sequence in the public database (Fig. 6B). Sequence alignments between the query and reference sequences are also displayed (Fig. 6C). If a ns substitution is found in the query DNA sequence, it was evaluated by the SIFT program[26] (incorporated in the local system), which predicts whether amino acid substitutions in a protein will be 'Deleterious' or 'Tolerated' using evolutionary information from the homologous proteins (Fig. 6B). We tested the prediction accuracy of SIFT with our data sets of disease-associated mutations and non-disease-associated nsSNPs, and found that the

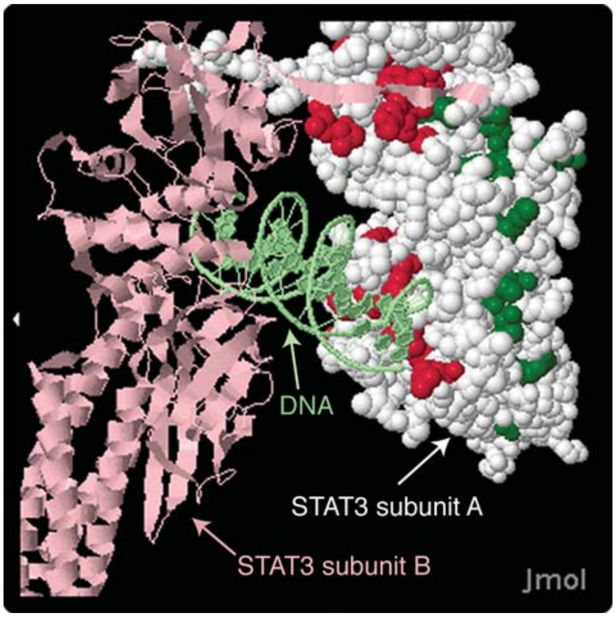
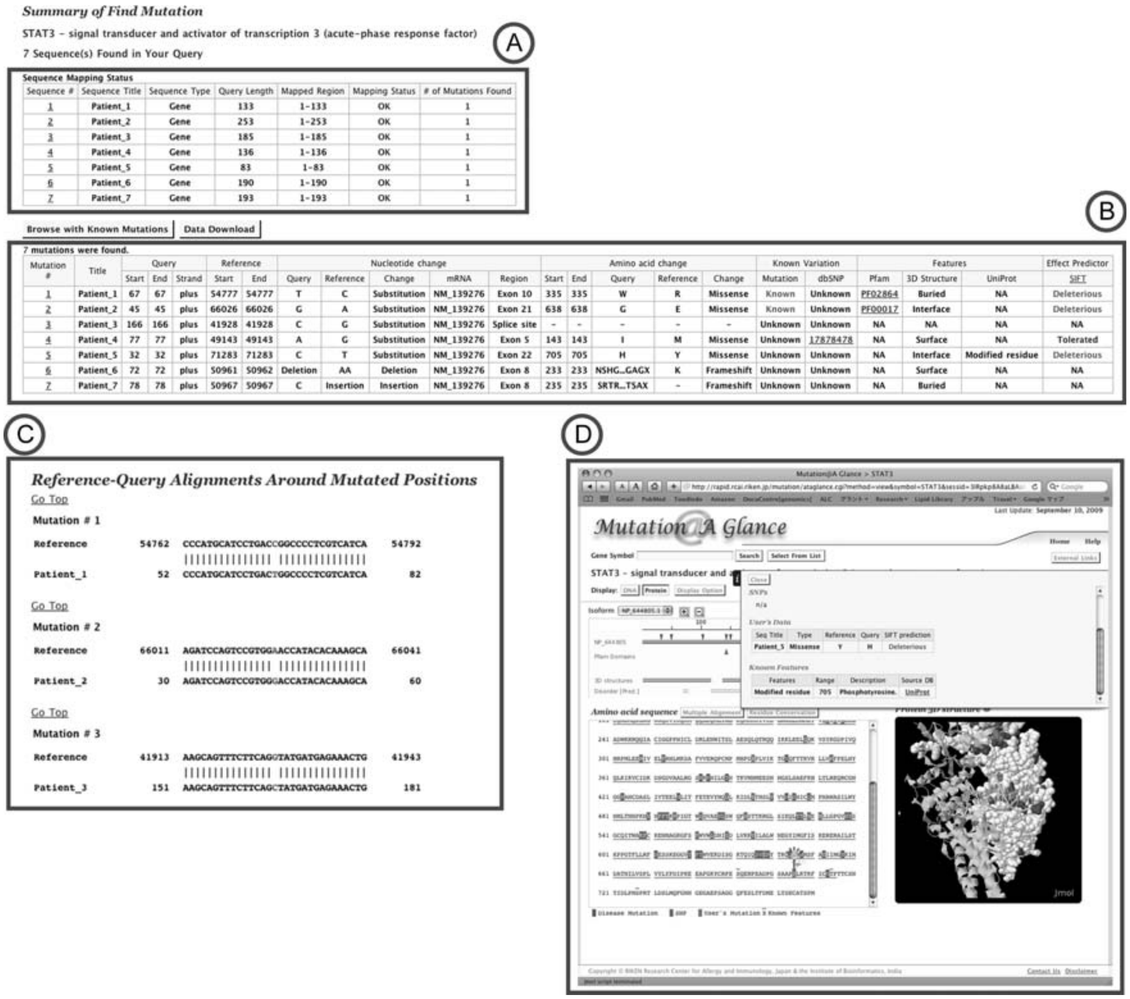


Figure 5. Spatial localization of disease-associated missense mutation sites on the STAT3 protein structure. Two STAT3 subunits are represented as a space-filling model coloured white (subunit A) and a ribbon model coloured pink (subunit B), respectively. A double-stranded DNA is represented as a ribbon model coloured light green. The disease-related missense mutations and nsSNPs of STAT3 (subunit A) are coloured magenta and green, respectively.

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015



Summary of Find Mutation

STAT3 – signal transducer and activator of transcription 3 (acute-phase response factor)

7 Sequence(s) Found in Your Query

Sequence Mapping Status

Sequence #	Sequence Title	Sequence Type	Query Length	Mapped Region	Mapping Status	# of Mutations Found
1	Patient_1	Gene	133	1-133	OK	1
2	Patient_2	Gene	253	1-253	OK	1
3	Patient_3	Gene	185	1-185	OK	1
4	Patient_4	Gene	136	1-136	OK	1
5	Patient_5	Gene	83	1-83	OK	1
6	Patient_6	Gene	190	1-190	OK	1
7	Patient_7	Gene	193	1-193	OK	1

7 mutations were found.

Mutation #	Title	Query			Reference		Nucleotide change					Amino acid change					Known Variation		Features			Effect Predictor
		Start	End	Strand	Start	End	Query	Reference	Change	mRNA	Region	Start	End	Query	Reference	Change	Mutation	dbSNP	Pfam	3D Structure	UniProt	SIFT
1	Patient_1	67	67	plus	54777	54777	T	C	Substitution	NM_139276	Exon 10	335	335	W	R	Missense	Known	Unknown	PF02864	Buried	NA	Deleterious
2	Patient_2	45	45	plus	66026	66026	G	A	Substitution	NM_139276	Exon 21	638	638	G	E	Missense	Known	Unknown	PF00017	Interface	NA	Deleterious
3	Patient_3	166	166	plus	41928	41928	C	G	Substitution	NM_139276	Splice site	-	-	-	-	-	Unknown	Unknown	NA	NA	NA	NA
4	Patient_4	77	77	plus	49143	49143	A	G	Substitution	NM_139276	Exon 5	143	143	I	M	Missense	Unknown	17878478	NA	Surface	NA	Tolerated
5	Patient_5	32	32	plus	71283	71283	C	T	Substitution	NM_139276	Exon 22	705	705	H	Y	Missense	Unknown	Unknown	NA	Interface	Modified residue	Deleterious
6	Patient_6	72	72	plus	50961	50962	Deletion	AA	Deletion	NM_139276	Exon 8	233	233	NSHG...GAGX	K	Frameshift	Unknown	Unknown	NA	Surface	NA	NA
7	Patient_7	78	78	plus	50967	50967	C	Insertion	Insertion	NM_139276	Exon 8	235	235	SRTR...TSAX	-	Frameshift	Unknown	Unknown	NA	Buried	NA	NA

Reference-Query Alignments Around Mutated Positions

Go Top

Mutation # 1

```
Reference     54762  CCCATGCATCCTGACCGGCCCCTCGTCATCA  54792
                     ||||||||||||||| |||||||||||||||
Patient_1        52  CCCATGCATCCTGACTGGCCCCTCGTCATCA     82
```

Go Top

Mutation # 2

```
Reference     66011  AGATCCAGTCCGTGGAACCATACACAAAGCA  66041
                     ||||||||||||||| |||||||||||||||
Patient_2        30  AGATCCAGTCCGTGGGACCATACACAAAGCA     60
```

Go Top

Mutation # 3

```
Reference     41913  AAGCAGTTTCTTCAGGTATGATGAGAAACTG  41943
                     ||||||||||||||| |||||||||||||||
Patient_3       151  AAGCAGTTTCTTCAGCTATGATGAGAAACTG    181
```

STAT3 – signal transducer and a

SNPs

n/a

User's Data

Seq Title	Type	Reference	Query	SIFT prediction
Patient_5	Missense	Y	H	Deleterious

Known Features

Features	Range	Description	Source DB
Modified residue	705	Phosphotyrosine.	UniProt

Figure 6. An example of evaluating sequence variations in query *STAT3* DNA sequences. (A) The mapping status of each query sequence to the reference sequence is shown. (B) If a variation is found in the query sequence, the detailed information is shown for each variation (e.g. the positions on the DNA/protein sequences, the type of variation and the description as to whether or not it is known as a disease-associated mutation or SNP). Results from the SIFT program ('Tolerated' or 'Deleterious') are also shown if the variation caused ns substitutions. (C) The query-reference sequence alignment around the altered nucleotides is depicted. (D) The variations can be visualized in the viewer, represented by different colours for known disease mutations or SNPs as 'User's Data'.

false-negative rate (falsely predicted as 'Tolerated' for disease-associated mutations) and the false-positive rate (falsely predicted as 'Deleterious' for nsSNPs) were 25% and 39%, respectively. These accuracy values were comparable to those evaluated in previous study.[33] The current version of Mutation@A Glance does not implement a method for quantitative evaluation of mutation effects on RNA splicing, mainly because we considered the evaluation method is not matured enough yet. However, because the evaluation of mutation effects on RNA splicing/stability is very intriguing, we will place a high priority on the implementation of the evaluation tool for genetic variations affecting RNA splicing/stability in the future development.

There are several advantages of Mutation@A Glance over other existing web servers for evaluating the effects of mutations. First, users are only required to have DNA sequences from a particular gene as their input and thus do not need to pre-process their submission data; other websites for evaluating the mutation effects require a list of genetic variations as a query, not raw sequence data.[26–31] Secondly, Mutation@A Glance identifies and addresses multiple types of sequence variations (e.g. insertion/deletion, frame-shifts) from input query DNA sequences whereas the other web servers do not. Thirdly, newly identified genetic variations can be easily compared with known mutation and SNP data using the graphical visualization interface of Mutation@A Glance (Fig. 6D).

From a viewpoint of clinical use, it is obvious that any mutation analysis platform cannot serve as a useful one without reliable mutation data sets. However, whereas large amounts of disease-associated mutation

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

data for various genetic diseases have been reported, most of them are dispersed and stored locally. Only a few websites, e.g. OMIM and UniProt, integrate disease-associated mutation data and allow us to download their contents. However, the mutation data in such databases have a relatively low integrity in terms of updating and coverage. Thus, we have begun to comprehensively collect and manually curate the disease-associated mutation data from published literature focusing on PIDs and established a resource of PID research for clinical use, named RAPID.[17] Mutation@A Glance thus uses these manually curated data sets for over 150 PID genes in the RAPID database, which is solid enough for clinical use at least for PID analysis. To make Mutation@A Glance a reliable and general mutation analysis platform for other various genetic diseases in the future, we consider that data sharing with experts in particular diseases will be highly important as in the case of PID; otherwise it would take a long time to accumulate extensive mutation data of all human disease genes to an acceptable level for clinical use. In fact, similar efforts along this direction have been being made by the research community.[19]

As new technologies for determining genetic variation in humans have rapidly and continuously emerged (such as next generation DNA sequencing), amounts of genetic variation data of human are exponentially growing.[6,7,59] Therefore, we will continue to update and improve the Mutation@A Glance system, in order to cope with the larger-scale data analysis for more comprehensive identification of disease-causative candidate genes. Implementing API programs into Mutation@A Glance for query submissions and a retrieval system through command line scripts would be more convenient for this purpose.

In summary, Mutation@A Glance provides a highly integrated bioinformatics tool for mutation analysis not only for facilitating visualization of sequence variation data along with various types of information, including primary and tertiary structures of the gene products, but also for evaluating the effects of novel sequence variations in a query input DNA sequence. This tool works solely on a web browser through Internet and is open to the public. Hence, Mutation@A Glance can be used as a ‘one-stop’ integrated bioinformatics platform for analysing genotype–phenotype relationships of genetic diseases from molecular as well as clinical perspectives.

Acknowledgements: The authors would like to thank Drs Shigeaki Nonoyama, Kohsuke Imai, Hirokazu Kanegane, Toshio Miyawaki, Koichi Oshima, Fumihiko Ishikawa and Reiko Kikuno-Fukaya for their critical suggestions about this work.

Supplementary Data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work is supported by a Grant-in-aid for Special Coordination Funds for Promoting Science and Technology, from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Amberger, J., Bocchini, C.A., Scott, A.F. and Hamosh, A. 2009, McKusick’s Online Mendelian Inheritance in Man (OMIM), *Nucleic Acids Res.*, **37**, D793–6.
2. Notarangelo, L., Casanova, J.L., Fischer, A., et al. 2004, Primary immunodeficiency diseases: an update, *J. Allergy Clin. Immunol.*, **114**, 677–87.
3. Geha, R.S., Notarangelo, L.D., Casanova, J.L., et al. 2007, Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee, *J. Allergy Clin. Immunol.*, **120**, 776–94.
4. Notarangelo, L.D. and Sorensen, R. 2008, Is it necessary to identify molecular defects in primary immunodeficiency disease?, *J. Allergy Clin. Immunol*, **122**, 1069–73.
5. Frazer, K.A., Ballinger, D.G., Cox, D.R., et al. 2007, A second generation human haplotype map of over 3.1 million SNPs, *Nature*, **449**, 851–61.
6. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. and Sunyaev, S.R. 2009, Power of deep, all-exon resequencing for discovery of human trait genes, *Proc. Natl Acad. Sci. USA*, **106**, 3871–6.
7. Chun, S. and Fay, J.C. 2009, Identification of deleterious mutations within three human genomes, *Genome Res.*, **19**, 1553–61.
8. Sunyaev, S., Ramensky, V., Koch, I., Lathe, W. III, Kondrashov, A.S. and Bork, P. 2001, Prediction of deleterious human alleles, *Hum. Mol. Genet.*, **10**, 591–7.
9. Tavtigian, S.V., Byrnes, G.B., Goldgar, D.E. and Thomas, A. 2008, Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications, *Hum. Mutat.*, **29**, 1342–54.
10. Tavtigian, S.V., Greenblatt, M.S., Goldgar, D.E. and Boffetta, P. 2008, Assessing pathogenicity: overview of results from the IARC Unclassified Genetic Variants Working Group, *Hum. Mutat.*, **29**, 1261–4.
11. Tavtigian, S.V., Greenblatt, M.S., Lesueur, F. and Byrnes, G.B. 2008, In silico analysis of missense substitutions using sequence-alignment based methods, *Hum. Mutat.*, **29**, 1327–36.
12. Houdayer, C., Dehainault, C., Mattler, C., et al. 2008, Evaluation of in silico splice tools for decision-making in molecular diagnosis, *Hum. Mutat.*, **29**, 975–82.
13. Stenson, P.D., Mort, M., Ball, E.V., et al. 2009, The Human Gene Mutation Database: 2008 update, *Genome Med.*, **1**, 13.

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

14. Singh, A., Olowoyeye, A., Baenziger, P.H., et al. 2008, MutDB: update on development of tools for the biochemical analysis of genetic variation, *Nucleic Acids Res.*, **36**, D815–9.
15. Mailman, M.D., Feolo, M., Jin, Y., et al. 2007, The NCBI dbGaP database of genotypes and phenotypes, *Nat. Genet.*, **39**, 1181–6.
16. Piirila, H., Valiaho, J. and Vihinen, M. 2006, Immunodeficiency mutation databases (IDbases), *Hum. Mutat.*, **27**, 1200–8.
17. Keerthikumar, S., Raju, R., Kandasamy, K., et al. 2009, RAPID: Resource of Asian Primary Immunodeficiency Diseases, *Nucleic Acids Res.*, **37**, D863–7.
18. Sherry, S.T., Ward, M.H., Kholodov, M., et al. 2001, dbSNP: the NCBI database of genetic variation, *Nucleic Acids Res.*, **29**, 308–11.
19. Kaput, J., Cotton, R.G., Hardman, L., et al. 2009, Planning the human variome project: the Spain report, *Hum. Mutat.*, **30**, 496–510.
20. Lee, P.H. and Shatkay, H. 2008, F-SNP: computationally predicted functional SNPs for disease association studies, *Nucleic Acids Res.*, **36**, D820–4.
21. Yuan, H.Y., Chiou, J.J., Tseng, W.H., et al. 2006, FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization, *Nucleic Acids Res.*, **34**, W635–41.
22. Yue, P., Melamud, E. and Moult, J. 2006, SNPs3D: candidate gene and SNP selection for association studies, *BMC Bioinformatics*, **7**, 166.
23. Chelala, C., Khan, A. and Lemoine, N.R. 2009, SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms, *Bioinformatics*, **25**, 655–61.
24. Yang, J.O., Hwang, S., Oh, J., Bhak, J. and Sohn, T.K. 2008, An integrated database-pipeline system for studying single nucleotide polymorphisms and diseases, *BMC Bioinformatics*, **9** (Suppl 12), S19.
25. Frezal, J. 1998, Genatlas database, genes and development defects, *C. R. Acad. Sci. III*, **321**, 805–17.
26. Ng, P.C. and Henikoff, S. 2003, SIFT: predicting amino acid changes that affect protein function, *Nucleic Acids Res.*, **31**, 3812–4.
27. Ramensky, V., Bork, P. and Sunyaev, S. 2002, Human non-synonymous SNPs: server and survey, *Nucleic Acids Res.*, **30**, 3894–900.
28. Thomas, P.D., Campbell, M.J., Kejariwal, A., et al. 2003, PANTHER: a library of protein families and subfamilies indexed by function, *Genome Res.*, **13**, 2129–41.
29. Bromberg, Y. and Rost, B. 2007, SNAP: predict effect of non-synonymous polymorphisms on function, *Nucleic Acids Res.*, **35**, 3823–35.
30. Ferrer-Costa, C., Gelpi, J.L., Zamakola, L., Parraga, I., de la Cruz, X. and Orozco, M. 2005, PMUT: a web-based tool for the annotation of pathological mutations on proteins, *Bioinformatics*, **21**, 3176–8.
31. Bao, L., Zhou, M. and Cui, Y. 2005, nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms, *Nucleic Acids Res.*, **33**, W480–2.
32. Thusberg, J. and Vihinen, M. 2009, Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods, *Hum. Mutat.*, **30**, 703–14.
33. Ng, P.C. and Henikoff, S. 2006, Predicting the effects of amino acid substitutions on protein function, *Annu. Rev. Genomics Hum. Genet.*, **7**, 61–80.
34. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. 2007, UniProtKB/Swiss-Prot, *Methods Mol. Biol.*, **406**, 89–112.
35. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
36. 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**, 4673–80.
37. Berman, H.M., Battistuz, T., Bhat, T.N., et al. 2002, The Protein Data Bank, *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
38. Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. 2000, Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
39. Shen, M.Y. and Sali, A. 2006, Statistical potential for assessment and prediction of protein structures, *Protein Sci.*, **15**, 2507–24.
40. Shrake, A. and Rupley, J.A. 1973, Environment and exposure to solvent of protein atoms. Lysozyme and insulin, *J. Mol. Biol.*, **79**, 351–71.
41. Ward, J.J., McGuffin, L.J., Bryson, K., Buxton, B.F. and Jones, D.T. 2004, The DISOPRED server for the prediction of protein disorder, *Bioinformatics*, **20**, 2138–9.
42. Sunyaev, S., Ramensky, V. and Bork, P. 2000, Towards a structural basis of human non-synonymous single nucleotide polymorphisms, *Trends Genet.*, **16**, 198–200.
43. Vitkup, D., Sander, C. and Church, G.M. 2003, The amino-acid mutational spectrum of human genetic disease, *Genome Biol.*, **4**, R72.
44. Wang, Z. and Moult, J. 2001, SNPs, protein structure, and disease, *Hum. Mutat.*, **17**, 263–70.
45. Ferrer-Costa, C., Orozco, M. and de la Cruz, X. 2002, Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties, *J. Mol. Biol.*, **315**, 771–86.
46. Kono, H., Yuasa, T., Nishiue, S. and Yura, K. 2008, coliSNP database server mapping nsSNPs on protein structures, *Nucleic Acids Res.*, **36**, D409–13.
47. Chen, J.W., Romero, P., Uversky, V.N. and Dunker, A.K. 2006, Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions, *J. Proteome Res.*, **5**, 879–87.
48. Schuster-Bockler, B. and Bateman, A. 2008, Protein interactions in human genetic diseases, *Genome Biol.*, **9**, R9.
49. Holland, S.M., DeLeo, F.R., Elloumi, H.Z., et al. 2007, STAT3 mutations in the hyper-IgE syndrome, *N. Engl. J. Med.*, **357**, 1608–19.
50. Minegishi, Y., Saito, M., Tsuchiya, S., et al. 2007, Dominant-negative mutations in the DNA-binding

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015

domain of STAT3 cause hyper-IgE syndrome, *Nature*, **448**, 1058–62.

51. Finn, R.D., Tate, J., Mistry, J., et al. 2008, The Pfam protein families database, *Nucleic Acids Res.*, **36**, D281–8.
52. Sayers, E.W., Barrett, T., Benson, D.A., et al. 2009, Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.*, **37**, D5–15.
53. Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al. 2009, Human Protein Reference Database—2009 update, *Nucleic Acids Res.*, **37**, D767–72.
54. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. 1998, GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support, *Bioinformatics*, **14**, 656–64.
55. Hijikata, A., Kitamura, H., Kimura, Y., et al. 2007, Construction of an open-access database that integrates cross-reference information from the transcriptome and proteome of immune cells, *Bioinformatics*, **23**, 2934–41.
56. Kanehisa, M. and Goto, S. 2000, KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res.*, **28**, 27–30.
57. Becker, S., Groner, B. and Muller, C.W. 1998, Three-dimensional structure of the Stat3beta homodimer bound to DNA, *Nature*, **394**, 145–51.
58. Kent, W.J. 2002, BLAT—the BLAST-like alignment tool, *Genome Res.*, **12**, 656–64.
59. Ahn, S.M., Kim, T.H., Lee, S., et al. 2009, The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group, *Genome Res.*, **19**, 1622–9.

Downloaded from http://dnaresearch.oxfordjournals.org/ at :: on October 12, 2015