D700–D706 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 16 November 2006*
doi:10.1093/nar/gkl826

PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease

Anil G. Jegga[1], Sivakumar Gowrisankar[2], Jing Chen[2] and Bruce J. Aronow[1,2,*]

Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, [1]Department of Pediatrics, College of Medicine and [2]Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH 45229, USA

Received August 15, 2006; Revised September 29, 2006; Accepted October 2, 2006

ABSTRACT

As knowledge of human genetic polymorphisms grows, so does the opportunity and challenge of identifying those polymorphisms that may impact the health or disease risk of an individual person. A critical need is to organize large-scale polymorphism analyses and to prioritize candidate non-synonymous coding SNPs (nsSNPs) that should be tested in experimental and epidemiological studies to establish their context-specific impacts on protein function. In addition, with emerging high-resolution clinical genetics testing, new polymorphisms must be analyzed in the context of all available protein feature knowledge including other known mutations and polymorphisms. To approach this, we developed PolyDoms (http://polydoms.cchmc.org/) as a database to integrate the results of multiple algorithmic procedures and functional criteria applied to the entire Entrez dbSNP dataset. In addition to predicting structural and functional impacts of all nsSNPs, filtering functions enable group-based identification of potentially harmful nsSNPs among multiple genes associated with specific diseases, anatomies, mammalian phenotypes, gene ontologies, pathways or protein domains. PolyDoms, thus, provides a means to derive a list of candidate SNPs to be evaluated in experimental or epidemiological studies for impact on protein functions and disease risk associations. PolyDoms will continue to be curated to improve its usefulness.

INTRODUCTION

Single nucleotide polymorphisms in coding regions (cSNPs) and regulatory regions have the potential to affect gene function (1–3). Non-synonymous cSNPs (nsSNPs), which change the amino acid sequence of proteins and are likely to affect the structure and function of the proteins, are good candidates for disease-modifying alleles. However, not infrequently molecular epidemiological studies have reported little or no association between cSNPs and disease susceptibility (4–6). Thus, as much as possible, it is essential to identify nsSNPs most likely to have functional effects before undertaking large-scale association studies. Established efforts to predict whether an nsSNP can affect the protein function and structure range from tools to visualize SNPs in their three-dimensional context (7,8), and predict molecular effects and potential impact of nsSNPs (4,9–13), to the recent SNPs3D (14) which integrates a variety of relevant information sources of nsSNPs [for additional details see the recent review by Mooney (15)]. Most of these approaches and analytical methods, however, are divided across various databases and interfaces, and users typically have to go through several web sites to analyze a single nsSNP. To overcome this, we have developed the PolyDoms resource to integrate most of these resources and results for each nsSNP, collating these data along with Gene Ontology, disease and other protein functional annotations in a web-accessible query interface.

DATA SOURCES

Table 1 and Figure 1 list the various types of data and their sources used for building the PolyDoms database. PolyDoms currently houses a total of 39 325 human RefSeq proteins, representing 26 378 unique RefSeq genes of which 6567 have alternate spliced products. The public repository of SNPs, NCBI's dbSNP database Build 125 (16) is our cSNP resource. We retrieved a total of 47 267 nsSNPs from dbSNP Build 125. To maximize our coverage of potential functional cSNPs, we included all the cSNPs from dbSNP without limiting to validated cSNPs alone. Another reason for this inclusion is that there are many reports of non-validated nsSNPs in the clinical literature [e.g. G1120E

*To whom correspondence should be addressed at Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA. Tel: +1 513 636 4865; Fax: +1 513 636 2056; Email: bruce.aronow@cchmc.org

The authors wish it to be known that, the first three authors should be regarded as joint First Authors

Table 1. Data type and sources used in PolyDoms

Data type	Source	URL (Reference)
Gene/protein	NCBI Reference Sequence	http://www.ncbi.nlm.nih.gov/RefSeq/ (30)
cSNPs	NCBI dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/ (16)
Protein domains	NCBI CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml (31)
Protein structure	PDB	http://www.rcsb.org/pdb/ (32)
Protein interactions	NCBI Entrez Gene (file interactions.gz)	ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/
Gene Ontology annotations	NCBI Entrez Gene (file gene2go.gz)	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
Gene families	HGNC Gene Families/Grouping Nomenclature	http://www.gene.ucl.ac.uk/nomenclature/genefamily.html
Pathways	KEGG	http://www.genome.ad.jp/kegg/pathway.html (33)
	Biocarta	http://biocarta.com/
	BioCyc	http://www.biocyc.org/ (34)
	Reactome	http://www.genomeknowledge.org/ (35)
Mutations	OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
	SwissChange	http://www.expasy.ch/cgi-bin/lists?humpvar.txt
Disease–gene association and mammalian phenotype	OMIM	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
	GAD	http://geneticassociationdb.nih.gov/ (22)
	MGI	http://www.informatics.jax.org/searches/MP_form.shtml (23)
Links to other external resources	iHOP	http://www.ihop-net.org (36)
	MutDB	http://mutdb.org/ (7)
	UCSC Proteome	http://genome.ucsc.edu/cgi-bin/pbGateway (37)

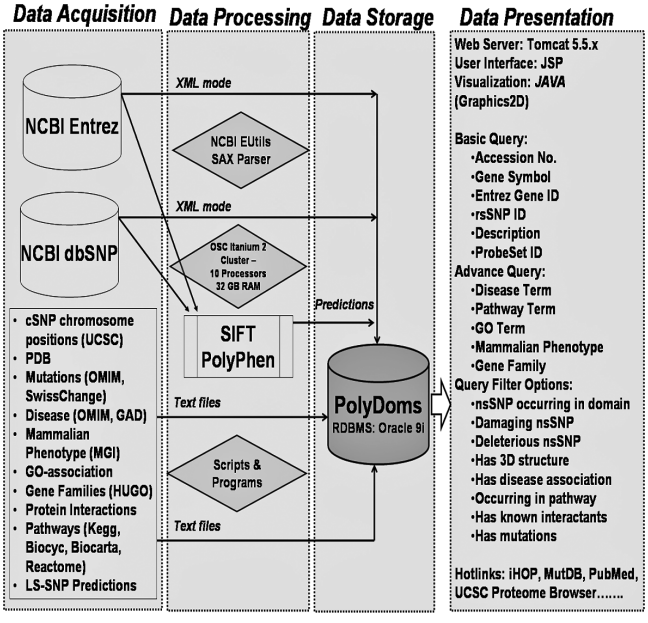


Figure 1. Schematic representation of PolyDoms data resources, work-flow and features.

in the APC protein in patients with gastric cancer (17)]. The protein sequence data and all associated annotations were extracted from NCBI's Entrez databases. Other sequence annotations and nsSNP-related information from various sources (see Figure 1) were downloaded as text files from original sources. Supplementary Data 1 summarizes the current status of PolyDoms database.

DATA PROCESSING AND STORAGE

Data processing

The NCBI's Entrez Programming Utilities (EUtils) (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) were used to download the protein (including protein domain information) and the cSNP-related data. The results fetched using EUtils XML mode were parsed using SAX parser (available as part of J2SDK 5.0). For nsSNPs in genes with more than one mRNA transcript, individual entries were recorded for each unique transcript to reflect potential differences in amino acid numbering. Individual entries were also recorded where more than one allele frequency submission was available. For example, an nsSNP with three mRNA transcripts and four different submissions resulted in a total of twelve separate entries.

JAVA programs were written to parse and normalize other downloaded text files (GO-gene associations, protein–protein interactions, OMIM/SwissChange mutations, LS-SNP predictions, mammalian phenotype gene associations) and uploaded to PolyDoms database.

Prediction of nsSNP implication

We used two sequence homology-based tools, SIFT (Sort Intolerant from Tolerant; version 2.1) (9) and PolyPhen (Polymorphism Phenotype; version 1.1) (4), to predict the potential impact of nsSNP on protein function. Additionally, when available, we have included the LS-SNP predictions (11). LS-SNP predicts positions where nsSNPs destabilize proteins, interfere with the formation of domain–domain interfaces, have an effect on protein–ligand binding or severely impact human health (11). In cases, due to data-related errors, where an amino acid residue position in the dbSNP record did not match with the amino acid residue at the same position in the corresponding protein record from RefSeq database, SIFT/PolyPhen analysis returned errors. For example, rs11557865 denotes nsSNP Ser551Pro; but the corresponding protein sequence (NP_061872; KIAA1128) has aspartic acid at position 551. Similarly, rs10891338 represents nsSNP Pro208Leu whereas the corresponding protein, BCDO2 (NP_114144), has lysine at position 208.

SIFT uses sequence homology among related genes and domains across species to predict the impact of all 20 possible

amino acids at a given position, allowing users to determine which nsSNPs would be of most interest to study. The SIFT algorithm has been shown to predict a phenotype for an nsSNP more accurately than previously used substitution scoring matrices, such as BLOSUM62, as these matrices do not incorporate information specific to the protein of interest (18,19). Another advantage of using SIFT is the potential to analyze a larger number of nsSNPs than methods that are dependent on the availability of protein structure alone (19,20). The PolyPhen algorithm, such as SIFT, takes an evolutionary approach in distinguishing deleterious nsSNPs from functionally neutral ones. However, it also takes into account the data from protein structure databases, such as PDB (Protein Data Bank) and PQS (Protein Quarternary Structure), DSSP (Dictionary of Secondary Structure in Proteins), and three-dimensional structure databases to determine if a variant may have an effect on the secondary structure of the protein, interchain contacts, functional sites and binding sites (4).

SIFT and PolyPhen analyses were performed on Ohio Supercomputer Center's (OSC) Itanium 2 Cluster (http://www.osc.edu/hpc/computing/it2/), configured in shared memory parallel running mode with a maximum of 10 processors and 32 GB RAM. Under this configuration, ~50 SIFT or ~600 PolyPhen jobs can be processed in an hour. The LS-SNP predictions were downloaded from the original source, parsed and uploaded to PolyDoms database.

Data storage

The PolyDoms database is implemented in Oracle 9i. The central table is 'Gene' that has an up-to-date list of all human RefSeq genes. The Gene table is linked to several other master tables. The cSNP table, apart from annotations, contains the SIFT and PolyPhen predictions. Other tables linked to the Gene table are as follows: the Transcript table (RefSeq mRNAs); the Protein table (RefSeq proteins); the ProbeSets; the Mutation table (OMIM and SwissChange) the Disease tables (OMIM and GAD); Mammalian Phenotype; Pathway (KEGG, Biocarta, BioCyc and Reactome), Protein–protein interactions (BIND, HPRD and Reactome); and Protein function (GO).

ACCESS AND INTERFACE

The main access to PolyDoms is through its web interface at http://polydoms.cchmc.org, by querying with sequence accession numbers, gene symbols, Entrez Gene IDs, rsSNP IDs, description or probeset IDs (Illumina; Affymetrix). Additionally, it is possible to retrieve a list of genes and associated cSNPs using a GO term, disease term (OMIM or GAD), pathway term (KEGG, Biocarta BioCyc or Reactome), mammalian phenotype or gene family (Figure 1). The output of a search presents the user with an option to view synonymous SNPs or nsSNPs. cSNPs are represented graphically in the context of protein sequence and domains (Figure 2). The results of the SIFT and PolyPhen predictions along with the LS-SNP extracted predictions for all nsSNPs of a protein are provided as a table below the image. Where available the mutant allele information from OMIM and SwissChange, and the protein–protein interactions are also provided. Up-to-date literature references implicating polymorphisms in disease are also provided. An expandable list provides links to various GO terms, pathways, diseases and phenotypes associated with the queried protein. Apart from these, the resource page is supplemented with cross-references to PDB, iHOP, MutDB and the UCSC Proteome Browser. All cross-references to data sources are hyperlinked enabling the original data to be viewed.

UTILITY

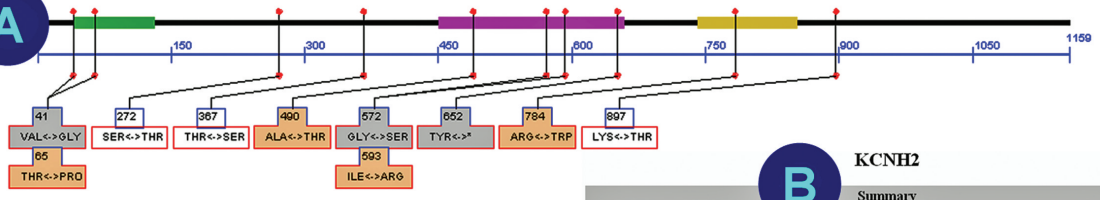
We present the utility and various features of PolyDoms through one case study using mammalian phenotype as an example. Since knowledge of complex diseases is limited, a comprehensive list of candidate genes and a method of ranking those genes by their disease-relevance is important in designing a good association study (14). Using NCBI's OMIM (21) and NIA's GAD database (22) and the mammalian phenotype (23), we provide a query interface through which a user can select any disease/phenotype term and create an nsSNP list based on the candidate genes associated with that particular disease/phenotype term. Although, we can assume the complexity of phenotype based on the number of genes associated with it, it may also partly reflect the current state of knowledge for that particular phenotype. Additional examples illustrating the utility and contents of database can be accessed through various case studies from Supplementary Data 8.

Case study: using the mammalian phenotype to investigate SNP–phenotype relationships

Aim: To obtain a list of human orthologous genes based on mouse genes associated with the phenotype 'abnormal podocytes'.

(i) From the homepage, click on the 'Phenotype Selector' (under section 'Search by disease, gene ontology, pathway, or gene family').
(ii) A new window ('Search for Mammalian Phenotype') opens up. Enter the search term 'abnormal podocytes' (or 'podocyte') and hit 'Search'. Select the term 'abnormal podocytes' from the search results window and hit 'use this phenotype for search' button to populate the 'Phenotype already selected' window. Click 'Done' to return to the PolyDoms query page.
(iii) Hitting the 'Search' button without selecting any of the 'Filter Options' will return the human orthologous genes (37 proteins, 19 unique in the current version) of mouse genes associated with the phenotype 'abnormal podocytes'. At this stage, users can either download the results as a spreadsheet by clicking on the link 'Download the results' or proceed to view the non-synonymous or synonymous model of each of the protein (see Figure 2 for a description of the output). Selecting the download option presents the user with a list of fields to select from and add to the spreadsheet.
(iv) Alternatively, use the 'Filter options' to refine the query. For example, from the 'Filter options' select 'Occurring in domain', 'Deleterious nsSNP' and 'Damaging nsSNP' and hit 'Search'. This will return 4 proteins (*ARHGDIA*,

A



CDD: 29035 PAS (41 - 132)
CDD: 40607 Ion_trans (451 - 659)
CDD: 28920 CAP_ED (742 - 853)

Deleterious (and/or) Damaging (SIFT/PolyPhen).
Mutation (OMIM/SWISSCHANGE).
Both

B

KCNH2

Summary

mRNA acc. no. : NM_000238
Protein acc. no. : NP_000229
Description : voltage-gated potassium channel, subfamily H, member 2 isoform a
Entrez Gene ID : 3757

Related Links

1. 3D Structure (Chime required)
2. Allelic Variants
3. Interactions
4. Pubmed Reference
5. View Alignments
6. UCSC Proteome Browser
7. iHOP
8. MutDB

Related Annotations

Disease(s) (OMIM)
1. Short QT syndrome-1, 609620 (3)
2. {Long QT syndrome, acquired, susceptibility to} (3)
3. Long QT syndrome-2 (3)

Gene Ontology
1. cation transport
2. delayed rectifier potassium channel activity (7736582)
3. integral to membrane
4. membrane
5. membrane fraction
6. muscle contraction (7736582)
7. potassium ion binding
8. potassium ion transport (7736582)
9. regulation of heart contraction (7736582)
10. regulation of transcription, DNA-dependent
11. sensory perception of sound
12. two-component sensor activity
13. two-component signal transduction system (phosphorelay)
14. voltage-gated potassium channel complex (7736582)

Mammalian Phenotype(s)
1. **Brugada syndrome** (16155735)
2. **EKG, abnormal** (15840476, 16132053, 16155735, 15851119, 15534720, 15051636, 15176425, 15746444, 15367556, , 12690509)
3. **Long QT syndrome** (14998624)
4. **QT interval** (12142119)
5. **cardiac repolarization.** (12829173)
6. **cardiovascular disease** (12829173)
7. **decreased heart rate** (12612061)
8. **long QT syndrome** (11997281, 12402336, 11743032, 11136691)
9. **long-QT syndrome** (11289718)

Pathway(s)
1. Cadmium induces DNA synthesis and [illegible]ion in macrophages (BioCarta)

C

Residue1	Pos	Residue2	refSNP ID	SNP UCSC Map	5'Sequence	Alleles	3'Sequence	Implication: Polyphen results	Implication: SIFT results	Implication: LS-SNP: Destabilizing SNP	Implication: LS-SNP: Destabilizing SNP Near Interface	Implication: LS-SNP: Destabili SNP Ne Ligan
VAL	41	GLY	rs731506	UCSC	Fasta	G/T	Fasta	Possibly damaging	Deleterious	-	-	-
THR	65	PRO	rs28933095	UCSC	Fasta	A/C	Fasta	Possibly damaging	Deleterious	-	-	-
SER	272	THR	rs13229961	UCSC	Fasta	A/T	Fasta	Benign	Tolerated	-	-	-
THR	367	SER	rs2228160	UCSC	Fasta	A/T	Fasta	Benign	Tolerated	-	-	-
ALA	490	THR	rs28928905	UCSC	Fasta	A/G	Fasta	Possibly damaging	Deleterious	destabilizing	-	-
GLY	572	SER	rs9333649	UCSC	Fasta	A/G	Fasta	Possibly damaging	Deleterious	-	-	-
ILE	593	ARG	rs28928904	UCSC	Fasta	G/T	Fasta	Probably damaging	Deleterious	buried charge change, destabilizing	-	-
TYR	652	*	rs8179010	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	-	-	-
TYR	652	*	rs17424631	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	-	-	-
TYR	652	*	rs1137617	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	-	-	-
TYR	652	*	rs17221791	UCSC	Fasta	C/G/T	Fasta	Probably damaging	Deleterious	-	-	-
ARG	784	TRP	rs12720441	UCSC	Fasta	C/T	Fasta	Probably damaging	Deleterious	buried charge change	-	-
LYS	897	THR	rs1805123	UCSC	Fasta	A/C	Fasta	Benign	Tolerated	-	-	

D

Allelic Variants

Variant	Source	Description
1-BP DEL	OMIM	LONG QT SYNDROME 2
27-BP DEL	OMIM	LONG QT SYNDROME 2
ALA490THR	OMIM	LONG QT SYNDROME, BRADYCARDIA-INDUCED
ALA561VAL	OMIM	LONG QT SYNDROME 2
ARG582CYS	OMIM	LONG QT SYNDROME 2
ARG752GLN	OMIM	LONG QT SYNDROME 2
ARG784TRP	OMIM	ACQUIRED LONG QT SYNDROME, SUSCEPTIBILITY TO
ASN470ASP	OMIM	LONG QT SYNDROME 2
ASN588LYS	OMIM	SHORT QT SYNDROME 1
ASN588LYS	OMIM	SHORT QT SYNDROME 1
GLY572ARG	OMIM	LONG QT SYNDROME 2
GLY628SER	OMIM	LONG QT SYNDROME 2
ILE593ARG	OMIM	LONG QT SYNDROME 2
IVS3, G-C, +1	OMIM	LONG QT SYNDROME 2
SER818LEU	OMIM	LONG QT SYNDROME 2
THR65PRO	OMIM	LONG QT SYNDROME 2
TRP1001TER	OMIM	LONG QT SYNDROME 2
VAL822MET	OMIM	LONG QT SYNDROME 2
F->L 29	SWISSCHANGE	Long QT syndrome type 2 (LQT2)
N->T 33	SWISSCHANGE	Long QT syndrome type 2 (LQT2)
G->V 47	SWISSCHANGE	Long QT syndrome type 2 (LQT2)
G->R 53	SWISSCHANGE	Long QT syndrome type 2 (LQT2)

E

PubMed References

1. Tester DJ, Cronk LB, Carr JL, Schulz V, Salisbury BA, Judson RS, Ackerman MJ.
Allelic dropout in long QT syndrome genetic testing: a possible mechanism und results.
Heart Rhythm. 2006 Jul;3(7):815-21. Epub 2006 Mar 16.
PMID: 16818214 [PubMed - in process]

2. Novotny T, Kadlecova J, Papousek I, Chroust K, Bittnerova A, Florianova A, Ceskova E, Weisla M, Toman O, Gaillyova R, Spinar J.
[Mutational analysis of LQT genes in individuals with drug induced QT interval
Vnitr Lek. 2006 Feb;52(2):116-8. Czech.
PMID: 16623272 [PubMed - indexed for MEDLINE]

3. Linna EH, Perkiomaki JS, Karsikas M, Seppanen T, Savolainen M, Kesaniemi YA, Makikallio T,
Functional significance of KCNH2 (HERG) K897T polymorphism for cardiac r analysis of T-wave morphology.
Ann Noninvasive Electrocardiol. 2006 Jan;11(1):57-62.
PMID: 16472284 [PubMed - indexed for MEDLINE]

4. Fitzgerald PT, Ackerman MJ.
Drug-induced torsades de pointes:
Heart Rhythm. 2005 Nov;2(2 Suppl):S30-
PMID: 16253929 [PubMed - indexed for

5. Lai LP, Su YN, Hsieh FJ, Chiang FT, Ju MH, Tsao HM, Chen SA, Lin TK, Wu MH,
Denaturing high-performance liquid sodium and potassium channel gen polymorphisms.
J Hum Genet. 2005;50(9):490-6. Epub 20
PMID: 16155735 [PubMed - indexed for

F

Interactions

Interactant1	Interactant2	Source	Description	Pubmed Reference
KCNH2(NP_000229)	ALG10	HPRD	-	14525949
KCNH2(NP_000229)	ALG10B (NP_001013642)	HPRD	-	14525949
KCNH2(NP_000229)	KCNE2	HPRD	-	11278781
KCNH2(NP_000229)	YWHAE	HPRD	-	11953308
KCNH2(NP_000229)	YWHAH	HPRD	-	11953308
KCNH2	ALG10 (CAC41349)	HPRD	-	14525949
KCNH2	KCNE2 (NP_751951)	HPRD	-	11278781
KCNH2	YWHAE (NP_006752)	HPRD	-	11953308
KCNH2	YWHAH (NP_003396)	HPRD	-	11953308

LAMA5, *NCK1* and *NPHS2*), each of which has at least one nsSNP that occurs in a conserved domain and has been predicted as 'Deleterious/Damaging' by SIFT/PolyPhen.

(v) The Supplementary Data 2 lists all the mammalian phenotypes along with the associated genes and the number of deleterious and damaging nsSNPs.

Prioritizing candidate nsSNPs

We screened a total of 44 641 (94%) nsSNPs associated with 14 967 protein sequences using SIFT and PolyPhen. Of these, 14 819 (33%) were predicted as 'deleterious' by SIFT and 14 622 (33%) as 'damaging' by PolyPhen. About 9021 nsSNPs (representing 5436 unique genes) were predicted as both deleterious and damaging indicating a concordance of ~62% between SIFT and PolyPhen predictions (see Supplementary Data 1 for additional details). Three studies (24–26) thus far have combined both the SIFT and PolyPhen algorithms to screen for deleterious nsSNPs. Xi *et al*. (26) and Johnson *et al*. (24) reported a concordance of 62 and 73% between these two programs analyzing the nsSNPs of genes involved in DNA repair and steroid hormone metabolism, respectively. In an earlier analysis of nsSNPs involved in DNA repair, cell cycle regulation, apoptosis and drug metabolism we used both SIFT and PolyPhen and identified 57 potentially deleterious nsSNPs (25). The Supplementary Data 3 lists all the nsSNPs that have been predicted as deleterious and damaging by both SIFT and PolyPhen. The Supplementary Data 4 and 5 list the disease/phenotype-associated genes that have at least one nsSNP predicted as damaging and deleterious by both PolyPhen and SIFT, respectively.

Although useful and widely used, both SIFT and PolyPhen have certain limitations. First, both of these require homologous sequences. Second, both of these algorithms disregard the impacts of a combination of variants (24,27). Third, SIFT and PolyPhen predict the impact of cSNPs only whereas non-coding SNPs (SNPs occurring in promoter or enhancer regions or splicing junctions) can also affect protein levels or protein function (24).

cSNPs resulting in premature stop codons and protein truncation

cSNPs introducing premature termination codons (nonsense SNPs) can alter the stability and function of transcripts and proteins and thus are considered to be biologically important. We retrieved a total of 965 nonsense SNPs (from 830 genes) from dbSNP Build 125 and 416 out of 965 nonsense SNPs affect an amino acid residue that is part of a functional protein domain. This led us to hypothesize that these cSNPs are likely to affect gene/protein function, although their biological relevance needs to be further investigated. However, we have noticed that some of the nonsense SNPs in dbSNP build 125 are either changed or removed from the dbSNP build 126. For instance, in the dbSNP build the number of nonsense SNPs affecting an amino acid residue which is part of a functional domain is 367. These changes will be reflected in our database when it is updated. Supplementary Data 6 lists all the cSNPs (based on dbSNP build 125) resulting in premature stop codons and also includes a comparison with the current dbSNP build 126.

KNOWN MUTATIONS VERSUS nsSNP FUNCTIONAL PREDICTION

To assess the potential for functional consequences of the PolyDoms defined intolerant nsSNPs, we downloaded 1338 SNPs from 611 candidate genes with known disease mutations (ftp://ftp.ncbi.nih.gov/snp/Entrez/snp_omimvar.txt) and subjected them to SIFT and PolyPhen analysis. Of the 1008 nsSNPs analyzed (330 out of 1338 nsSNPs were ignored because some of them were either non-coding SNPs or had erroneous annotations with mismatch of the residues), 568 (56%) nsSNPs showed concordance between SIFT and PolyPhen predictions and were classified both as 'deleterious' and 'damaging' (Supplementary Data 7). A total of 782 out of 1008 (78%) nsSNPs were either predicted as deleterious or damaging or both. Apart from confirming the utility of these prediction tools in prioritizing the candidate nsSNPs, it also suggests that nsSNPs predicted as damaging and deleterious and already associated with a phenotype/disease (Supplementary Data 4 and 5) represent a pool of candidate loci that should be interrogated further in association studies. We also noticed that only 34 out of the 568 nsSNPs predicted as damaging and deleterious are validated (by frequency) nsSNPs.

RELATED WORK

Although it is beyond the scope of the current article to compare PolyDoms with other resources of similar nature (see Introduction), some of the features that are unique to PolyDoms are related to the management of sets of nsSNPs—the ability to refine, export nsSNP sets as a whole and to create sets of cSNPs through complex queries (such as using pathways or Gene Ontology or mammalian phenotype classes described earlier and in the Supplementary Data 8). The goals of the recently published SNPs3D (14) are similar to ours: to integrate all of the available data relevant for assessing the likely role of particular genes and nsSNPs in a disease and help the researchers in making informed judgments. Additionally, the PolyDoms filter options make the data-mining process and compiling a 'hit-list' of nsSNPs relatively easy.

Figure 2. PolyDoms feature displays. (**A**) PolyDoms image of a nsSNP model of the protein *KCNH2*. Numbers in the image indicate the amino acid residue positions from the corresponding RefSeq protein sequence. The pink, yellow and green blocks over the protein sequence represent the three known domains derived from the NCBI's CDD. Vertical lines represent nsSNPs. The color codes indicate the predictions—gray represents an nsSNP predicted as deleterious and/or damaging; yellow indicates mutation (based on OMIM/SwissChange); orange indicates an nsSNP that has been predicted as deleterious and/or damaging and also reported as a mutation. (**B**) The summary view gives the basic sequence annotations along with an expandable list of diseases, GO terms, mammalian phenotypes and Pathways associated with the queried gene. (**C**) Tabular description of nsSNP predictions based on PolyPhen and SIFT analysis and LS-SNP annotations (refer to B above for descriptions of color codes). (**D**) Tabular list of allelic variants derived from OMIM and SwissChange. (**E**) Top five relevant abstracts, when available, related to queried gene polymorphisms and disease association. The list is generated dynamically and therefore is up-to-date with current literature. (**F**) List of protein–protein interactions (from NCBI Entrez Gene).

CONCLUSION

We have classified and catalogued the predicted functionality of nsSNPs in human genes to facilitate sequence-based association studies. The current version of PolyDoms however has some limitations. First, the current version of PolyDoms does not contain information on SNP co-occurences, complex haplotype or other relationships among the SNPs. Therefore, one of our future goals is to incorporate the SNP haplotype data (28). This will facilitate retrieving genotype and frequency data, picking tag-SNPs for use in association studies, viewing haplotypes graphically and examining marker-to-marker LD patterns. Second, since PolyDoms is built using multiple sources, keeping it up-to-date and synchronized with external resources, taking into account the different data formats, or the changes in their formats is tedious. However, we will strive to automate this process as much as possible. Third, PolyDoms does not still provide the complete range of analysis tools that can be useful in evaluating and characterizing cSNPs in terms of their potential effects (e.g. relative solvent accessibility of the variant residue). We are in the process of filling this gap using the SABLE server (29). Finally, PolyDoms does not include information about other SNPs (human non-coding SNPs or SNPs from other species). In conclusion, the use of PolyDoms and other resources similar to select functional nsSNPs for epidemiology studies can be an efficient way to explore the role of genetic variation in disease risk or altered response to therapeutic regimens, and to contain cost. However, it should be noted that deleterious effects on protein stability alone may not be sufficient conditions for disease predisposition.

AVAILABILITY

The PolyDoms database can be accessed freely at http://polydoms.cchmc.org.

SUPPLEMENTARY DATA

Supplementary Data are available at http://polydoms.cchmc.org/polydoms/supplementary/

DISCLAIMER

The purpose of this resource is to distribute functional annotations of human cSNP data. These cSNPs and their annotations are meant to be used as guidelines for basic research. Do not use these results to make clinical decisions.

ACKNOWLEDGEMENTS

The authors would like to thank Drs Deb Nickerson, Robert Livingston and Robert Weiss for super discussions and the Ohio Supercomputer Center for the assistance in using their supercomputing clusters to run whole genome SIFT and PolyPhen analyses. This work was supported by grants NCI UO1 CA84291-07 (Mouse Models of Human Cancer Consortium), NIH R24 DK 064403 (Digestive Diseases Research Development Center—DDRDC), NIEHS ES-00-005 (Comparative Mouse Genome Centers Consortium) and NIEHS P30-ES06096 (Center for Environmental Genetics). Funding to pay the Open Access publication charges for this article was provided by CCHMC, Cincinnati, OH, USA.

Conflict of interest statement. None declared.

REFERENCES

1. Chakravarti,A. (1998) It's raining SNPs, hallelujah? *Nature Genet*., **19**, 216–217.
2. Collins,F.S., Guyer,M.S. and Charkravarti,A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
3. Syvanen,A.C., Landegren,U., Isaksson,A., Gyllensten,U. and Brookes,A. (1999) First International SNP Meeting at Skokloster, Sweden, August 1998. Enthusiasm mixed with scepticism about single-nucleotide polymorphism markers for dissecting complex disorders. *Eur. J. Hum. Genet*., **7**, 98–101.
4. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.
5. Savas,S., Kim,D.Y., Ahmad,M.F., Shariff,M. and Ozcelik,H. (2004) Identifying functional genetic variants in DNA repair pathway using protein conservation analysis. *Cancer Epidemiol. Biomarkers Prev*., **13**, 801–807.
6. Zhu,Y., Spitz,M.R., Amos,C.I., Lin,J., Schabath,M.B. and Wu,X. (2004) An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. *Cancer Res*., **64**, 2251–2257.
7. Mooney,S.D. and Altman,R.B. (2003) MutDB: annotating human variation with functionally relevant data. *Bioinformatics*, **19**, 1858–1860.
8. Stitziel,N.O., Binkowski,T.A., Tseng,Y.Y., Kasif,S. and Liang,J. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res*., **32**, D520–D522.
9. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*., **31**, 3812–3814.
10. Reumers,J., Maurer-Stroh,S., Schymkowitz,J. and Rousseau,F. (2006) SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non synonymous SNPs. *Bioinformatics*, **22**, 2183–2185.
11. Karchin,R., Diekhans,M., Kelly,L., Thomas,D.J., Pieper,U., Eswar,N., Haussler,D. and Sali,A. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.
12. 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**, 3176–3178.
13. Stone,E.A. and Sidow,A. (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res*., **15**, 978–986.
14. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
15. Mooney,S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief Bioinformatics*, **6**, 44–56.
16. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
17. Horii,A., Nakatsuru,S., Miyoshi,Y., Ichii,S., Nagase,H., Kato,Y., Yanagisawa,A. and Nakamura,Y. (1992) The APC gene, responsible for familial adenomatous polyposis, is mutated in human gastric cancer. *Cancer Res*., **52**, 3231–3233.
18. Henikoff,S. and Henikoff,J.G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res*., **19**, 6565–6572.
19. Wang,Z. and Moult,J. (2001) SNPs, protein structure, and disease. *Hum. Mutat*., **17**, 263–270.
20. Sunyaev,S., Ramensky,V. and Bork,P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*., **16**, 198–200.
21. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM),

a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **30**, 52–55.
22. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nature Genet*., **36**, 431–432.
23. Smith,C.L., Goldsmith,C.A. and Eppig,J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*., **6**, R7.
24. Johnson,M.M., Houck,J. and Chen,C. (2005) Screening for deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response. *Cancer Epidemiol. Biomarkers Prev*., **14**, 1326–1329.
25. Livingston,R.J., von Niederhausern,A., Jegga,A.G., Crawford,D.C., Carlson,C.S., Rieder,M.J., Gowrisankar,S., Aronow,B.J., Weiss,R.B. and Nickerson,D.A. (2004) Pattern of sequence variation across 213 environmental response genes. *Genome Res*., **14**, 1821–1831.
26. Xi,T., Jones,I.M. and Mohrenweiser,H.W. (2004) Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics*, **83**, 970–979.
27. Rebbeck,T.R., Spitz,M. and Wu,X. (2004) Assessing the function of genetic variants in candidate gene association studies. *Nature Rev. Genet*., **5**, 589–597.
28. Thorisson,G.A., Smith,A.V., Krishnan,L. and Stein,L.D. (2005) The International HapMap Project Web site. *Genome Res*., **15**, 1592–1593.
29. Adamczak,R., Porollo,A. and Meller,J. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753–767.
30. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*., **33**, D501–D504.
31. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D., Ke,Z. *et al*. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*., **33**, D192–D196.
32. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
33. Kanehisa,M. (2002) The KEGG database. *Novartis Found. Symp*., **247**, 91–101; discussion 101–103, 119–128, 244–152.
34. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*., **33**, 6083–6089.
35. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G.R., Wu,G.R., Matthews,L. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.
36. Hoffmann,R. and Valencia,A. (2004) A gene network for navigating the literature. *Nature Genet*., **36**, 664.
37. Hsu,F., Pringle,T.H., Kuhn,R.M., Karolchik,D., Diekhans,M., Haussler,D. and Kent,W.J. (2005) The UCSC Proteome Browser. *Nucleic Acids Res*., **33**, D454–D458.