Difficulties in finding DNA mutations and associated phenotypic data in web resources using simple, uncomplicated search terms, and a suggested solution

Elizabeth A. Webb,[1] Timothy D. Smith[1] and Richard G.H. Cotton[1,2]*

[1]Genomic Disorders Research Centre, Melbourne, Vic 3053, Australia
[2]Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Parkville, Vic 3010, Australia
**Correspondence to:* Tel: +61 3 8344 1893; Fax: +61 3 9347 6842; E-mail: lizwebb@connexus.net.au

Date received (in revised form): 6th January 2011

Abstract

DNA mutation data currently reside in many online databases, which differ markedly in the terminology used to describe or define the mutation and also in completeness of content, potentially making it difficult both to locate a mutation of interest and to find sought-after data (eg phenotypic effect). To highlight the current deficiencies in the accessibility of web-based genetic variation information, we examined the ease with which various resources could be interrogated for five model mutations, using a set of simple search terms relating to the change in amino acid or nucleotide. Fifteen databases were investigated for the time and/or number of mouse clicks; clicks required to find the mutations; availability of phenotype data; the procedure for finding information; and site layout. Google and PubMed were also examined. The three locus-specific databases (LSDBs) generally yielded positive outcomes, but the 12 genome-wide databases gave poorer results, with most proving not to be searchable and only three yielding successful outcomes. Google and PubMed searches found some mutations and provided patchy information on phenotype. The results show that many web-based resources are not currently configured for fast and easy access to comprehensive mutation data, with only the isolated LSDBs providing optimal outcomes. Centralising this information within a common repository, coupled with a simple, all-inclusive interrogation process, would improve searching for all gene variation data.

Keywords: *web, mutation search, simple search strategies*

Introduction

Data on genetic *changes* are being generated at an ever-increasing rate and mutations (variations which cause genetic diseases[1]) are now being collected and catalogued in many different web-based databases. This information is used, and will increasingly be accessed, by clinicians and healthcare workers seeking to determine the significance of a mutation found to be present in any particular patient. It therefore needs to be comprehensive, reliable and readily accessible. It is important that information in databases can be rapidly retrieved, as healthcare workers often have severe constraints on the time which can be spent on any individual patient[2] or patient's diagnostic test.[3]

The way in which gene variations are collected at present is either in general databases which contain genome-wide data (reviewed by George *et al.*[4]) or in locus-specific databases (LSDBs) (reviewed by Claustres *et al.*[5]), where the emphasis is on collecting data pertaining to one single gene.

The advantage of the latter is that they are usually initiated and administered by experts in the field and, as such, tend to contain more data and more reliable data. The potential disadvantages of both types of databases, however, are that—due to their specialist nature—they may not be amenable to simple search techniques for data mining or use by non-geneticists.[6] In order for a wide range of health practitioners easily to access all information pertaining to a particular mutation, it would be highly desirable to be able quickly to interrogate a central web-based resource using simple, uncomplicated terms.[7] The most useful terms potentially being those relating directly to the change in either amino acid or nucleotide.

To test the current situation for these types of web searches, we sought to examine the ease of use of both genome-wide databases and LSDBs. We searched for five representative mutations using a variety of straightforward search terms relating to the amino acid or nucleotide change (not requiring knowledge of special parameters such as accession or other numbers) and measuring the time and number of computer mouse clicks (CMCs) taken to find the mutation. The availability of phenotype data and general usability were also monitored. Since Google and PubMed are also used as resources for obtaining information about specific genetic changes, these were also searched using the same search criteria.

Methods

The mutations used for web searches are shown in Table 1.

Search terms

Amino acid change:
Mutation 1. PAH, G148S, GLY148SER, Gly148 Ser, 148
Mutation 2. MLH1, Q62K, GLN62LYS, Gln62Lys, 62
Mutation 3. MLH1, K618A, LYS618ALA, Lys61 8Ala, 618
Mutation 4. BRCA2, D2723H, ASP2723HIS, Asp2723His, 2723
Mutation 5. BRCA1, V772A, VAL772ALA, Val772Ala, 772
Nucleotide change:
Mutation 1. PAH, 442G > A, GGT/AGT, c.44 2G > A

Selection criteria for mutations and databases for study

Mutations for this study were chosen to cover a range of probable importance (eg frequency of disease causation or reporting and clinical outcomes). Mutation 1 has a low frequency of leading to inherited disease (0.13–2.0 per cent).[8] Mutation 2 is reported four times in the InSiGHT database, while mutation 3 is reported 48 times (for the nucleotide change c.1853A > G). Mutation 4 is reported to be pathogenic,[9] while mutation 5 is regarded as an unclassified variant (Hyland, unpublished data).

Representative LSDBs were then selected for study along with well-recognised, commonly used genome-wide databases.

Search methods

Searches of databases (listed below) were performed by one operator between October 2008 and January 2009. Each database was searched in the order mutation 1 to 5, using combinations of the search terms listed above (usually gene symbol, followed by the amino acid change or position, but this was dependent on the database set-up).

Databases were interrogated using the following criteria: the number of CMCs and time required to find the mutation; whether phenotype data were provided; and the ease with which the search was achieved (database layout, search options, information updating and links supplied). Specifically, this was achieved by accessing the database home page, starting the timer and assessing the page to identify potential search approaches by visual examination and then mousing over examination of the available fields. The database was then interrogated using the set group of search terms and the number of CMCs necessary to find the mutation by going directly through the search possibilities was

Table 1. Mutations used for web searches

Mutation number	Gene name	Gene symbol	Disorder	Amino acid number	Amino acid change
1	Phenylalanine hydroxylase	PAH	Phenylketonuria	148	Glycine/ serine
2	mutL homolog 1	MLH1	Hereditary nonpolyposis colon cancer	62	Glutamine/ lysine
3	mutL homolog 1	MLH1	Hereditary non-polyposis colon cancer	618	Lysine/ alanine
4	Breast cancer type 2 susceptibility protein	BRCA2	Breast cancer	2723	Aspartic acid/ histidine
5	Breast cancer 1, early onset	BRCA1	Breast cancer	772	Valine/ alanine

recorded. The timer was then stopped and the types of information found were recorded. The availability of phenotypic data and other information (eg date of last update, links) were documented. The general presentation of the database from the point of view of searching for a specific mutation was also assessed.

Searches were also performed directly in Google and PubMed (during January 2009) using the amino acid search terms above. The Google Advanced search option was used, initially with 'exact word or phrase' and then unlinked terms. PubMed was searched with linked terms (for PAH), then in a two-stage process first using gene symbol and then searching these results for amino acid change.

Databases studied

The databases studied are shown in Table 2.

Results

Ability to find mutations using amino acid-related search terms

As shown in Table 3, all mutations were able to be found using the specified search terms only in HGMD, MUTdb, UniProt and their respective LSDBs; however, mutation 3 was incorrectly numbered in HGMD as codon 617. This may reflect the fact that this publicly-available database has not been updated since 2006. Only one mutation was located in OMIM (not necessarily surprising, given its policy of only including variations which are the first reported or most significant changes). Three of the five mutations were found in the Ensembl database. Gene variation data in the GeneCards database was expressed in the single-letter code (eg D/H for aspartic acid/histidine). When this was used as a search term, only mutation 4 was found and this was incorrectly numbered as 2722. Most general databases examined could not be interrogated by the search terms or did not directly contain variation data (eg in GeneReviews, which provides this information through links to LSDBs and general databases).

Analysing the frequency with which a particular mutation was found in any web resource showed that the mutations of perceived lesser importance (mutations 1, 2 and 5) were found six times each for mutations 1 and 5, and five times for mutations 2, whereas mutations 3 and 4 (of potentially greater clinical interest) were found 17 and nine times, respectively.

Table 2. The databases studied

Database	URL	Description
HGMD—Human Gene Mutation Database (open access content)	http://www.hgmd.org	Established for the study of mutational mechanisms in human genes and provides information of practical diagnostic importance
OMIM—Online Mendelian Inheritance in Man	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim	Comprehensive collection of human genes and genetic phenotypes with information on all known Mendelian disorders
dbSNP—NCBI Single Nucleotide Polymorphism[10]	http://www.ncbi.nlm.nih.gov/projects/SNP/	Established to serve as a central repository both for single base nucleotide substitutions and for short deletion and insertion polymorphisms
MutDB—Structurally Annotated Mutation Data	http://mutdb.org/	Annotation of human variation data with protein structural information and other functionally relevant information
MutView—Mutation View	http://mutview.dmb.med.keio.ac.jp/MutationView/jsp/mutview/index.jsp	Developed by the Keio University School of Medicine in collaboration with Chi Co., Ltd.
dbGaP—NCBI Genotypes and Phenotypes[11]	http://www.ncbi.nlm.nih.gov/gap	Developed to archive/distribute results of studies investigating the interaction of genotype/phenotype
GeneRev—GeneReviews	http://www.genereviews.org	Expert-authored, peer-reviewed, current genetic disease descriptions
GeneCards	http://www.genecards.org/	Searchable, integrated database of human genes
UniProt—Universal Protein Resource	http://www.uniprot.org/	A comprehensive, high-quality and freely accessible resource of protein sequence and functional information
GDB—Human Genome Database	http://www.gdb.org	A community-curated collection of human genomic data. No longer operational
Ensembl	http://www.ensembl.org/index.html	Software system which produces and maintains automatic annotation on selected eukaryotic genomes
DGV—Database of Genomic Variants	http://projects.tcag.ca/variation/	Comprehensive summary of structural variation in the human genome
PAHdb—Phenylalanine Hydroxylase Locus Knowledgebase	http://www.pahdb.mcgill.ca	Maintains and centralises mutation data on the *PAH* gene
InSiGHT—LOVD (Leiden Open Variation Database)	http://www.insight-group.org/lovd.html	International organisation aiming to improve quality of care for patients with hereditary gastrointestinal tumours
BIC—Open Access On-Line Breast Cancer Mutation Data Base	http://research.nhgri.nih.gov/bic/	Maintains a central repository for information regarding mutations and polymorphisms in breast cancer susceptibility genes

Table 3. Searches of general and locus-specific databases

Database	Mutation found					CMC to find or [not find]					Time to find or [not find] (min)					Phenotype found					Detailed phenotype[b]				
	1[d]	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
HGMD	+	+	+[a]	+	+	5	5	5[a]	5	5	6.15	3.27	8.95[a]	1.55	1.35	+	+	+[a]	+	+	−	−	−	−	−
OMIM	−	−	+	−	−	[15]	[17]	8	[19]	[15]	[7.73]	[7.03]	1.50	[6.03]	[5.88]	−	−	+	−	−	−	−	−	−	−
dbSNP	−	−	−	−	−	[18]	[18]	[25]	[15]	[12]	[10.80]	[11.57]	[11.57]	[11.56]	[4.48]	−	−	−	−	−	−	−	−	−	−
MutDB	+	+	+	+	+	4	4	4	4	4	6.50	2.90	1.58	1.15	0.98	+	+	+	+	+	+	+	+	+	+
MutView	−	−	−	−	−	[9]	[5]	[5]	[9]	[9]	[4.48]	[1.32]	[1.32]	[4.78]	[4.90]	−	−	−	−	−	−	−	−	−	−
GeneCards	−	−	−	−[c]	−	[19]	[16]	[15]	[15]	[14]	[9.57]	[9.62]	[8.68]	[8.02]	[8.07]	−	−	−	−	−	−	−	−	−	−
GeneRev	−	−	−	−	−	[15]	[4]	[4]	[4]	[4]	[13.00]	[3.48]	[3.48]	[2.70]	[2.70]	−	−	−	−	−	−	−	−	−	−
dbGaP	−	−	−	−	−	[3]	[3]	[3]	[2]	[2]	[1.11]	[1.10]	[1.10]	[0.98]	[0.98]	−	−	−		−	−	−	−	−	−
UniProt	+	+	+	+	+	3	3	3	3	3	1.18	1.00	0.75	1.03	0.80	+	+	+	+	+	+	+	+	+	+
Ensembl	−	+	+	+	−	[27]	6	6	6	[24]	[15.50]	5.95	4.75	2.75	[12.50]	−	−	−	−	−	−	−	−	−	−
GDB (NO)																									
DGV	−	−	−	−	−	[3]	[3]	[3]	[3]	[3]	[2.75]	[1.07]	[1.07]	[1.07]	[1.82]	−	−	−	−	−	−	−	−	−	−
PAHdb—Table 2	+					2					1.15					−					−				
InSiGHT		+	+				10	4				10.85	1.98				+	+				+	+		
BIC				+	+				5	5				1.00	0.92				+	+				+	+

[a]Wrongly numbered as 617.
[b]Defined as being appropriate additional information (eg impact on severity of disease).
[c]Mutation found using D/H format. Wrongly numbered as 2722. No phenotype data.
[d]Mutations 1–5.
Abbreviation: NO, not operational.

Time and number of CMCs required to find mutations

Where mutations were found (HGMD, OMIM, MUTdb, UniProt, Ensembl and the appropriate LSDBs), this was accomplished in time-frames varying from 45 seconds to 10.85 minutes using one to ten CMCs (see Table 3). As shown in Figure 1, the longest search times were generally for mutation 1 (the first search performed for all general databases), with decreasing times required for subsequent searches (mutations 2 to 5), indicating that all sites required a period of time to become familiar with layout and to select appropriate search strategies. This was also evident from the two InSiGHT searches, where the initial search for mutation 2 took much longer than the second search for mutation 3. The search for mutation 3 in HGMD took longer, as the time to confirm that codon 618 was wrongly numbered was included. Although the search times for each mutation decreased in HGMD, MUTdb, UniProt and Ensembl, the number of CMCs required to find data remained the same (five, four, three and six, respectively). Since none of the chosen mutations was found in dbSNP, and it does contain a 'clinical/LSDB submissions' category, some time (52 minutes) was spent unsuccessfully trying to locate (using the specific search terms) a mutation known to be present in the database.

Availability of phenotype data

For databases where mutations were found using the chosen search terms, phenotype data were supplied in all except Ensembl and the PAH KnowledgeBase (Table 3), although how informative was this reporting varied considerably between databases. For example, for mutation 3, the phenotype was reported as follows: in HGMD as 'colorectal cancer, non-polyposis'; in OMIM as 'colorectal cancer, hereditary non-polyposis type 2' and 'associated with colorectal cancer'; in MUTdb as 'in HNPCC2'; requires 2 nucleotide substitutions. SIFT Score – 0.02, SIFT prediction - DELETERIOUS'; in InSiGHT as 'reported pathogenicity' and 'concluded pathogenicity'. The 48 entries of separate instances of the same mutation 'reported pathogenicity' showed seven entries as 'pathogenic', one as 'probably pathogenic', 21 as 'non-pathogenic' and 19 as 'unknown', whereas all entries were recorded as 'unknown' in the 'concluded pathogenicity' category (meaning that pathogenicity status is still undetermined); in UniProt, the phenotype was reported as 'K → A common polymorphism; requires 2 nucleotide substitutions'. In an attempt to differentiate between these different database reports, MUTdb, InSiGHT and UniProt are listed in Table 3 as having detailed phenotype information, since some measure of the impact on pathogenicity is provided.

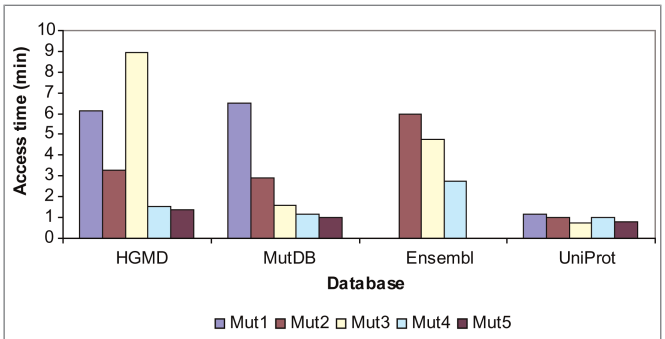

Figure 1. Time taken to access mutations (Mut) 1 to 5 on the Human Gene Mutation Database (HGMD), Structurally Annotated Mutation Data (MutDB), Ensembl and Universal Protein Resource (UniProt) websites.

Database rating

Databases were rated, against a number of criteria with regard to their ease of use for finding mutations and information about mutations using the specified search terms (Table 4). These results show the UniProt, InSiGHT and MUTdb databases to be the best-performers. The first two comply in all fields, with the only negative aspect being the complexity of the InSiGHT site, due mainly to the large number of data fields and amount of information contained within them. This made the initial InSiGHT search longer, but once one was acquainted with the set-up, the second search took around a fifth of the time of the first search. MutDB only failed on the criterion of recent updated – the site indicated that the system itself was last updated on 2nd June, 2007. OMIM, while complying in all fields, failed to display all the mutations, supplying

Table 4. Database comparisons for ease of use characteristics for finding variations causing inherited disease (mutations)

Database	Mutations found	Time to find[a] (<5)	CMCs to find (<10)	Phenotype found	Password registration not required	Database aim or explanation	Clear options for searching	Clear site layout	Recent DB update (2008)
HGMD	*	*	*	*			*	*	[e]
OMIM	1[b]	*	*	*	*	*	*	*	*
dbSNP[c]	0[b]	NA	NA	NA	*				*
MutDB	*	*	*	*	*	*	*	*	
MutView	0[b]	NA	NA	NA	*		*	*	
GeneCards	0[b]	NA	NA	NA	*	*	*	[d]	*
GeneRev	0[b]	NA	NA	NA	*		*	*	
dbGap	0[b]	NA	NA	NA	*	*			
UniProt	*	*	*	*	*	*	*	*	*
Ensembl	3[b]	*	*		*	*	*	[d]	*
DGV	0[b]	NA	NA	NA	*	*	*	*	
PAHdb	*	*	*		*	*	*	*	*
InSiGHT	*	*	*	*	*	*	*	[d]	*
BIC	*	*	*	*		*	*	*	

*Indicates compliance with column heading criteria.
[a]For the last mutation searched (thus allowing for an 'experience' factor).
[b]Actual number of mutations found.
[c]Included for comparison, not a mutation database.
[d]Although layout is relatively clear, many fields are included in these databases, making them complex to navigate initially.
[e]Public version only.
NA, Not applicable (mutation not found).

only one result where the mutation fell within its policy of only including the first reported or most significant genetic changes. The weakest performing databases were dbSNP[10] and dbGaP,[11] which failed not only in their inability to find mutations, but also on site layout and ease of searching—presumably reflecting the fact that collection and display of mutations are not the main aims of these databases. As noted in Table 4, dbSNP[10] (not set up to be a mutation database) was included for comparative purposes; however, now, it does have some mutation data in the category of information listed as 'Clinical/LSDB submissions'.

The links for various databases were also documented (Table 5) and showed a wide range of results, with some databases having extensive links (eg GeneCards), while others had very few.

Ability to find mutations using nucleotide-related search terms

This study was only performed for mutation 1 using five databases (HGMD, OMIM, dbSNP,[10] MUTdb and PAH). PAHdb was the only database that was able to locate the mutation using nucleotide terminology. It was easily found, in a time of 1.8 minutes, using only four CMCs, although phenotype data were not available.

Direct searching using the Google search engine

As shown in Table 6, most of the searches generated only a small number of results. Searches for MLH1 K618A, MLH1 Lys618Ala and BCRA2 D2723H gave larger numbers of hits, although these included repeated material, with Google

Table 5. Links available from the various databases

Links	HGMD	OMIM	dbSNP	MUTdb[a]	MutView	GeneCards[b]	GeneRev	dbGaP	UniProt	Ensembl	DGV	PAHdb[c]	InSiGHT	BIC
OMIM	*					*	*		*	*		*	*	
GDB	*											*		
NCBI			*	*		*		*		*		*		*
EntrezGene	*	*				*	*			*			*	*
GeneCards	*											*		
GenAtlas	*					*								
JSNP	*													
GAD	*	*				*								
FINDbase	*													
GeneClinics	*											*		
SwissProt	*			*		*	*			*		*		
TrEMBL						*				*				
LSDB	*	*			*	*	*							
Coriell		*												
HGVS		*												
HGMD		*				*	*					*	*	
Genage		*												
HGNC	*	*				*	*			*				
dbGaP				*										
PharmGKB				*		*								
SeattleSNPs				*										
UCSC				*		*						*		
Ensembl				*		*								
GeneTests						*							*	

Continued

Table 5. Continued

Links	HGMD	OMIM	dbSNP	MUTdb[a]	MutView	GeneCards[b]	GeneRev	dbGaP	UniProt	Ensembl	DGV	PAHdb[c]	InSiGHT	BIC
UniGene						*				*				
WikiGene						*				*				
RCSB												*		
BIOPKU												*		
THBdb												*		
WoodsMMR													*	
MMRUV													*	
HapMap						*					*			
DECIPHER											*			
dbRIP														
HSVD														
CAC														
PDB										*				
IPI										*				

*Indicates link available.
[a]Additional links specific to MUTdb: SIFT; PolyPhen; LS-SNP; SNPs3D; PolyDoms; Panther; PMut; SNPEffect; FASTSNP.
[b]Additional links specific to GeneCards: GeneLoc; dbSNP; AKS; HuGE; AceView; euGenes; miRbase; ECGene; H-InvDB; ATLAS; HORDE; IMGT; Leiden; GeneRev; Navigator; BCGD; TGDB; Pupasuite; Homologene; Pseudogene; SGD; MGI; Flybase; Wormbase; GeneDecks; GeneNote; GNF SynAtlas; GeneAnnotation; GeneTide; SAGE tags; CGAP; Source; GNF BioGPS; ExpoldeB; RNAdb; ASD; BioMol; MINT; String; Kegg; IntAct; Phosphosite; Proteopedia; OCA; ProtoNet; BLOCKS; InterPro.
[c]Additional links specific to PAHdb: Cell bank; The Waystation; PHEXdb; CASRdb; HEXdb; CYSdb; Human Genome Variation Society (nomenclature guidelines).

Table 6. Google searches

Search term	No. entries found	No. databases	Entry no./name of database	Phenotype found
Mutation 1				
'PAH G148S'[a]	0	–		
PAH G148S[b]	19	2	3/ PAH LK[f]	No
			5/ FINDbase	No
'PAH Gly148Ser'	0	–	–	
PAH Gly148Ser	0	–	–	
'PAH GLY148SER'	0	–	–	
PAH GLY148SER	0	–	–	
Mutation 2			–	
'MLH1 Q62K'	2	0	–	
MLH1 Q62K	10	0	–	
'MLH1 Gln62Lys'	0	–	–	
MLH1 Gln62Lys[c]	8	0	–	
Mutation 3			–	
'MLH1 K618A'	7	0	–	
MLH1 K618A	159	2	55;56/LOVD[d]	Yes
	(72)[e]			
'MLH1 Lys618Ala'	4	0	–	
MLH1 Lys618Ala[c]	283	2	37;49/LOVD[d]	Yes
	(80)[e]	1	53/GeneCards	No
Mutation 4				
'BRCA2 D2723H'	6	0		
BRCA2 D2723H	87	1	11/kConFab Consortium	Yes
	(33)[e]			
'BRCA2Asp2723His'	0	–		
BRCA2Asp2723His[c]	5	1	4/kConFab Consortium	Yes
Mutation 5				
'BRCA1 V772A'	0	–		
BRCA1 V772A	12	1	12/kConFab Consortium	Yes

Continued

Table 6. Continued

Search term	No. entries found	No. databases	Entry no./name of database	Phenotype found
'BRCA1 Val772Ala'	0	–		
BRCA1 Val772 Ala[c]	1	1	1/kConFab Consortium	Yes

[a]Quote marks indicate Google Advanced exact wording or phrase search.
[b]Lack of quote marks indicates Google Advanced search with unlinked terms.
[c]Upper case letters gave the same result as lower case.
[d]Leiden Open Variation Database.
[e]Google estimate of unique entries.
[f]PAH Locus Knowledgebase.

estimating the number of unique entries to be much lower.

For mutation 1, using PAH G148S in the 'exact wording or phrase' search produced no result, but when were used unlinked terms Google found references to two databases. The PAH Locus Knowledgebase was included in this study, so it was already known that the mutation could be found here, but that no phenotype information is available. FINDbase provides information about the frequency of different mutations leading to inherited disorders in various world populations and does not provide phenotype data. No other combinations of search terms was successful.

None of the searches for mutation 2 supplied results. The mutation 3 unlinked search terms, MLH1 and Lys618Ala, found references to data from LOVD and GeneCards. The mutation was found in the LOVD entries with mention of phenotype, but the GeneCards entry linked to the general MLH1 page in this database (the recognition of amino acid change coming from a journal reference). MLH1 K618A also found references to LOVD and the mutation was found along with phenotype data. This mutation topped the number of hits generated, with 159 for MLH1 K618A and 283 for MLH1 Lys618Ala (used as unlinked terms). Importantly, a comparison of the unique entries for each of these searches (72 and 80, respectively) showed only five overlapping results.

As shown also in Table 6, a number of different searches for both mutations 4 and 5 detected the kConFab Consortium site. Investigating this site showed both mutations represented, with mutation 4 classed as 'pathogenic' and mutation 5 shown as an 'unclassified variant'.

Most of the entries generated in each search related to journal articles. As an example, for mutation 4, it took 10.68 minutes to examine ten entries. Only one had potentially useful information which was available in an open-access journal.

PubMed searches

Since the initial searches with combinations of PAH search terms were unsuccessful, subsequent searches were performed in two stages (Table 7). The first, searching for PAH, MLH1, BRCA2 and BRCA1 alone, generated large numbers of entries. Subsequent searches (within the initial search results) for the amino acid change, however, were only successful for mutations 3 and 4, yielding five articles for K618A, one for Lys618Ala and one each for D2723H and Asp2723His. On examination of these, only one of the former[14] contained data relating to phenotype, and this cast doubt on the clinical importance of this often-reported mutation. Reference 9 strongly suggested that mutation 4 is deleterious in BRCA2.

Discussion

Web-based databases documenting human gene variation have developed over recent decades, generally to fulfil a particular need within laboratories

Table 7. PubMed searches

First search term	Second search term	No. entries produced	Reference
Mutation 1			
'PAH G148S'		0	
PAH G148S		0	
PAH		8420	
PAH	G148S	0	
PAH	Gly148Ser[a]	0	
Mutation 2			
MLH1		2353	
MLH1	Q62K	0	
MLH1	Gln62Lys[a]	0	
Mutation 3			
MLH1	K618A	5	12–16
MLH1	Lys618Ala[a]	1	17
Mutation 4			
BRCA2		3906	
BRCA2	D2723H	1	9
BRCA2	Asp2723His[a]	1	18
Mutation 5			
BRCA1		6317	
BRCA1	V772A	0	
BRCA1	Val772Ala[a]	0	

Quote marks combined search terms.
[a]Upper case letters gave the same result as lower case.

and the clinicians/patients they serve. Consequently, there is now a plethora of divergent databases with varying datasets, formats, links and search capabilities, making it difficult to find information about a mutation of interest in a reasonable time-frame. In an ideal world, a complete dataset for each particular mutation would be rapidly obtainable using simple, uncomplicated search terminology. In this study, we demonstrate that the current situation for locating five model mutations of varying clinical importance, with associated information pertaining to their phenotypic effect, using a set of simple search terms relating to amino acid and base changes, to be less than ideal.

Amino acid-related search terms showed that only three of 12 general databases supplied information (including phenotype) for all five mutations. All were found in their relevant LSDBs, with phenotype data, with the exception of PAHdb, however. Search times for these were generally less than five minutes, although the initial searches often took longer. All of these resources have pros and cons. Most are fairly user-friendly, but HGMD has not been updated for public access since 2006 and access to current data requires a subscription. MutDB lacks recorded times for data updates (the database system was last updated in 2007). OMIM only contains the first reported or most significant variations and hence is not complete. Ensembl and PAHdb do not have any phenotypic data. The InSiGHT database appeared unwieldy to use initially, but this improved for subsequent searches. BIC requires a password (provided to members who abide by a set of rules) and has not been updated for BRCA1 and -2 since 2007. Although phenotypic data could be found for some mutations, the detail provided varied greatly. The inability to find phenotype data for mutation 1 in the PAH database was not due to the restrictive search terms, as searching for a mutation (known to be present) using this format was successful, indicating an incomplete dataset in this database. Generally, the ability to find mutations was related to their potential clinical importance.

In the group of databases that did not display any of the mutations using the specified search terms, dbSNP[10] was the most difficult to navigate (available search queries relate to ID numbers—such as RefSNP—or nomenclature—such as HGVS name). Mutation data in the format used in this study is present in the database but cannot be located directly using the search terms. A method of searching the 'Clinical/LSDB submissions' entries directly for a particular mutation would be ideal. This could potentially be done by using the nucleotide sequence (eg ACGTACGT(N/N)ACGTACGT) as a search term, which would

identify the now-existing graphic summary sequence format. Alternatively, an ability to search within the entries for other expressions of nucleotide or amino acid change would be useful. For a recently initiated database (MedRefSNP), it is specifically noted that there is a bias against the inclusion of some data due to the sole reliance on RefSNP identifiers.[19] Hence, it is not able to capture data (and presumably to be searched) using amino acid or nucleotide change terminology.

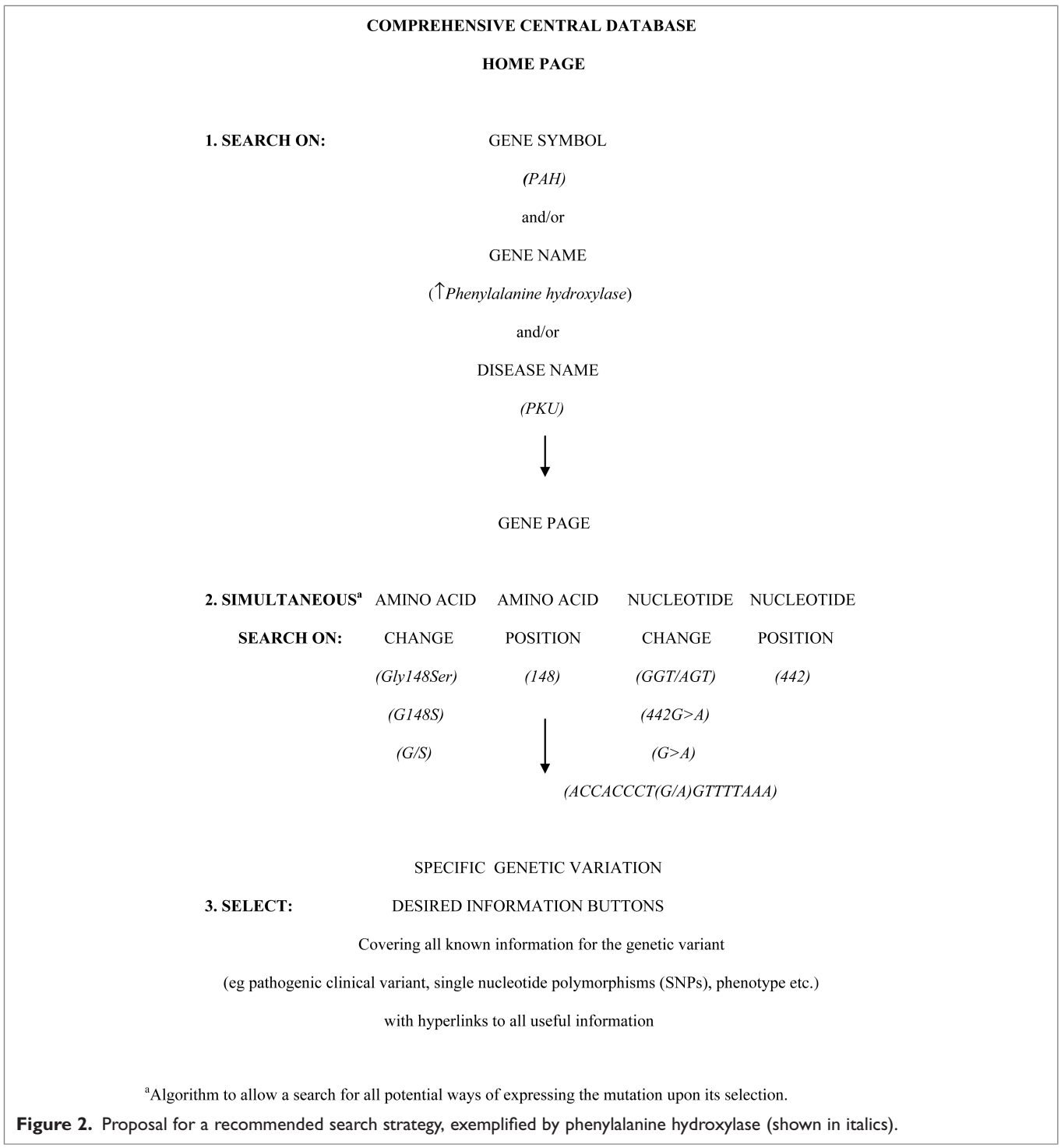

Figure 2. Proposal for a recommended search strategy, exemplified by phenylalanine hydroxylase (shown in italics).

As perhaps expected, Google searches provided direct access to some databases and provided results for some, but not all, of the mutations. Interestingly, the search terms MLH1 K618A and Lys618Ala generated hundreds of results, which showed little overlap when the unique entries were compared. This indicates a vital need to use multiple versions of associated nomenclature in order to capture maximum information. PubMed (and also Google) searches displayed various journal article information on some mutations, but sifting through this information can be time consuming. Cataloguing data from publications is clearly critical, as, without a database, dozens of publications would need to be read to obtain data. In addition, the information is not curated and hence its reliability would need to be assessed by the user. Thus, the need for curated databases is underlined. The use of nucleotide-related search terms on a subset of databases for mutation 1 proved less useful than using amino acid terminology. In fact, this type of search would be preferable to using amino acid change, as the latter is not always the outcome of a mutation and the actual sequence data generated could be used. In particular, as discussed above, the graphic summary nucleotide sequence displayed in dbSNP[10] in particular would seem to lend itself to a nucleotide-based search.

The results obtained here point to the need for new, improved search capabilities and strategies to allow easier access to state-of-the-moment mutation information. These could well be developed in association with the formation of a central data repository. The argument for a single repository is compelling, as this would allow for standardisation and provide a 'one-stop shop' for all mutation information. It has been suggested that this role be taken by one of the centralised databases[7]—the European Bioinformatics Institute or the National Center for Biotechnology Information through dbSNP[10]—which would receive already-curated data from LSDBs. One of the tasks of the central repository could be to provide systems that enable searchers easily to extract all known information relating to each mutation in short time-frames with minimal CMCs (a universal search strategy). We propose a process along the lines outlined in Figure 2, where two CMCs would be required to obtain the specific genetic variation sought, with a selection of search options on this page providing access to all current knowledge concerning that variant. The first search would locate the gene page and the second would allow for simultaneous interrogation using all of the suggested search terms in Figure 2, triggered by the keying-in of any one of these, using a yet-to-be-developed tool.

In conclusion, these results serve to highlight the disparate nature of current cataloguing of human gene variations and the difficulties encountered in interrogating the various web-based resources using straightforward terms. While all existing databases have attributes concomitant to their aims, we have demonstrated an urgent need for a universal, broad-reaching approach where all information pertaining to a specific mutation can be readily accessed using a set of simple search terminologies.

Acknowledgments

The authors wish to thank Val Hyland for useful discussions on the need for improved searching systems.

References

1. Cotton, R.G.H. and Scriver, C.R. (1998), 'Proof of a "disease causing" mutation', *Hum. Mut.* Vol. 12, pp. 1–3.
2. Ely, J.W., Osheroff, J.A., Ebell, M.H., Bergus, G.R. *et al.* (2000), 'Analysis of questions asked by family physicians regarding patient care', *West. J. Med.* Vol. 172, pp. 315–319.
3. Mitchell, G., Ardern-Jones, A., Kissin Mchir, M., Taylor, R. *et al.* (2001), 'A paradox: Urgent BRCA genetic testing', *Fam. Cancer.* Vol. 1, pp. 25–29.
4. George, R.A., Smith, T.D., Callaghan, S., Hardman, L. *et al.* (2008), 'General mutation databases: Analysis and review', *J. Med. Genet.* Vol. 45, pp. 65–70.
5. Claustres, M., Horaitis, O., Vanevski, M., Cotton, R.G. *et al.* (2002), 'Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases', *Genome Res.* Vol. 12, pp. 680–688.
6. Levy, H.P., LoPresti, L. and Seibert, D.C. (2008), 'Twenty questions in genetic medicine — An assessment of world wide web databases for genetics information at point of care', *Genet. Med.* Vol. 10, pp. 659–667.
7. Kaput, J., Cotton, R.G.H., Hardman, L., Watson, M. *et al.* (2009), 'Planning the human variome project: The Spain report', *Hum. Mutat.* Vol. 30, pp. 496–510.
8. Hennermann, J.B., Vetter, B., Wolf, C., Windt, E. *et al.* (2000), 'Phenylketonuria and hyperphenylalaninemia in eastern Germany: A characteristic molecular profile and 15 novel mutations', *Hum. Mutat.* Vol. 15, pp. 254–260.
9. Goldgar, D.E., Easton, D.F., Deffenbaugh, A.M., Monteïro, A.N. *et al.* (2004), 'Integrated evaluation of DNA sequence variants of unknown

clinical significance: Application to BRCA1 and BRCA2', *Am. J. Hum. Genet*. Vol. 75, pp. 535–544.
10. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J. *et al*. (2001), 'dbSNP: The NCBI database of genetic variation', *Nucleic Acids Res.* Vol. 29, pp. 308–311.
11. Mailman, M.D., Feolo, M., Jin, Y., Kimura, M. *et al.* (2007), 'The NCBI dbGaP database of genotypes and phenotypes', *Nat. Genet*. Vol. 39, pp. 1181–1186.
12. Rubio-Del-Campo, A., Salinas-Sánchez, A.S., Sánchez-Sánchez, F., Giménez-Bachs, J.M. *et al.* (2008), 'Implications of mismatch repair genes hMLH1 and hMSH2 in patients with sporadic renal cell carcinoma', *BJU Int*. Vol. 102, pp. 504–509.
13. Blasi, M.F., Ventura, I., Aquilina, G., Degan, P. *et al.* (2006), 'A human cell-based assay to evaluate the effects of alterations in the MLH1 mismatch repair gene', *Cancer Res.* Vol. 66, pp. 9036–9044.
14. Belvederesi, L., Bianchi, F., Loretelli, C., Gagliardini, D. *et al.* (2006), 'Assessing the pathogenicity of MLH1 missense mutations in patients with suspected hereditary nonpolyposis colorectal cancer: Correlation with clinical, genetic and functional features', *Eur. J. Hum. Genet*. Vol. 14, pp. 853–859.
15. Hudler, P., Vouk, K., Liovic, M., Repse, S. *et al.* (2004), 'Mutations in the hMLH1 gene in Slovenian patients with gastric carcinoma', *Clin. Genet*. Vol. 65, pp. 405–411.
16. Guerrette, S., Acharya, S. and Fishel, R. (1999), 'The interaction of the human MutL homologues in hereditary nonpolyposis colon cancer', *J. Biol. Chem*. Vol. 274, pp. 6336–6341.
17. Perera, S. and Bapat, B.. (2008), 'The MLH1 variants p.Arg265Cys and p.Lys618Ala affect protein stability while p.Leu749Gln affects heterodimer formation', *Hum. Mutat*. Vol. 29, p. 332.
18. Pages, S., Caux, V., Stoppa-Lyonnet, D., Tosi, M. *et al.* (2001), 'Screening of male breast cancer and of breast-ovarian cancer families for BRCA2 mutations using large bifluoroescent amplicons', *Br. J. Cancer* Vol. 84, pp. 482–488.
19. Rhee, H. and Lee, J.-S. (2009), 'MedRefSNP : A database of medically investigated SNPs', *Hum. Mutat*. Vol. 30, pp. E460–E466.