Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 317–325

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Review

The Curation of Genetic Variants: Difficulties and Possible Solutions

Kapil Raj Pandey [1,2], Narendra Maden [1,3,*], Barsha Poudel [4], Sailendra Pradhananga [1,5], Amit Kumar Sharma [1,5]

[1] *Deerwalk Services, Kathmandu 44602, Nepal*
[2] *Department of Microbiology, Bangalore University, Bangalore 560001, India*
[3] *Central Department of Microbiology, Tribhuvan University, Kathmandu 44613, Nepal*
[4] *Department of Bioinformatics, Wageningen University, Wageningen 6700 AA-6799 ZZ, Netherlands*
[5] *Department of Biotechnology, Kathmandu University, Kavrepalanchok 45200, Nepal*

Received 5 February 2012; revised 27 May 2012; accepted 20 June 2012
Available online 29 November 2012

Abstract

The curation of genetic variants from biomedical articles is required for various clinical and research purposes. Nowadays, establishment of variant databases that include overall information about variants is becoming quite popular. These databases have immense utility, serving as a user-friendly information storehouse of variants for information seekers. While manual curation is the gold standard method for curation of variants, it can turn out to be time-consuming on a large scale thus necessitating the need for automation. Curation of variants described in biomedical literature may not be straightforward mainly due to various nomenclature and expression issues. Though current trends in paper writing on variants is inclined to the standard nomenclature such that variants can easily be retrieved, we have a massive store of variants in the literature that are present as non-standard names and the online search engines that are predominantly used may not be capable of finding them. For effective curation of variants, knowledge about the overall process of curation, nature and types of difficulties in curation, and ways to tackle the difficulties during the task are crucial. Only by effective curation, can variants be correctly interpreted. This paper presents the process and difficulties of curation of genetic variants with possible solutions and suggestions from our work experience in the field including literature support. The paper also highlights aspects of interpretation of genetic variants and the importance of writing papers on variants following standard and retrievable methods.

Keywords: Difficulties in curation; Automated curation; Manual curation; Interpretation of variants

Introduction

Information about genetic variants in biomedical literature is already extensive and will increase exponentially in future as we are embarking on whole genome sequencing using high throughput next generation sequencing (NGS) technology [1]. While it is time-consuming to manually curate genetic variants from such an extensive and ever growing literature, automated tools that speed up curation are evolving or already in use [2,3]. Although automated curation is fast, nomenclature and presentation issues of variants as well as false positive and negative results inherent to the process lower the sensitivity and specificity [4]. Manual curation is adopted as the gold standard and used to compare the qualitative measures of automated tools. Most databases of variants rely on manual curation for data extraction and entry [5]. There are nearly 2000 locus-specific databases (LSDBs) or variant databases which serve as web-based platforms for submission, repository and extraction of overall information of variants [6,7]. While a LSDB can be built using specific guidelines, standard guidelines for establishing all LSDBs have been described by Vihinen et al. [7], which include the essential steps for the establishment, maintenance, information structure and ethics of LSDB. Furthermore, steps on the

* Corresponding author.
E-mail: maden.narendra@gmail.com (Maden N).

http://dx.doi.org/10.1016/j.gpb.2012.06.006

development of a centralized database have been made to better manage the deluge of variant data coming from whole genome sequencing via NGS [1,8]. For broad utility in clinical practice and research, variant databases should contain information regarding genotypic and phenotypic information, consequences of variants (predictive, structural and functional), and clinical and demographic data. Moreover, it is of utmost necessity to gather maximum information about variants before drawing any conclusions about their disease-causing potentials. Thus, it is necessary to understand thoroughly what difficulties surround the curation and ways to solve them effectively. In this paper, we discuss the steps and difficulties in curation with their possible solutions, automated curation, aspects of interpretation of the variants and importance of following a standard nomenclature of variants.

Curation of variants

Curation of variants should start with ascertaining reference sequences, standard ontology and nomenclature system of genes and proteins [7,9,10]. Gene selection for curation can be adopted as phenotype- or gene-specific; a common practice is to select a genetic disease and then consider all its linked genes for curation. Which and how many genes are required for curation is dependent upon the findings from linkage analysis, genotype-phenotype correlation and genome wide association study (GWAS). Information from genetic testing and counseling bodies, online Mendelian inheritance in man (OMIM), and phenotype-specific online databases (such as catalogue of somatic mutations in cancer (COSMIC)) are also useful for identifying genetic links to disease. Reference DNA and protein can be procured from NCBI (http://www.ncbi.nlm.nih.gov/genebank/index) and UniProt (www.uniprot.org), respectively. Recently, locus reference genome (LRG) reference sequence is gaining popularity to obtain reference sequences of DNA and protein (www.lrg-sequence.org/) [11]. Likewise, many databases may also provide reference DNA and protein sequences.

Use of different internet search engines (such as PubMed, Scholar Google and Google) and databases has become routine practice in obtaining relevant literature [12]. PubMed is one of the richest sources of biomedical papers citing over 21 millions papers. A list of common public databases for genetic variants can be found at human genome variation society (HGVS) (http://www.hgvs.org/dblist/glsdb.html) and UMD database (http://www.umd.be/). By using Boolean queries (single or combined form) and similarity-query, papers are mostly extracted manually but an automated method has also been used to extract a large set of literature [4,13,14]. Essentially, search strings to extract papers on variants should include the following terms: gene name(s) / protein name(s) / phenotype(s) / genetic or gene variant / variant / types of genetic variants / variation / polymorphism / [15,16]. For example, search strings utilized to extract papers for Cystic fibrosis transmembrane regulator (CFTR) gene variants should include: (CFTR) (mutations) / (CFTR) (genetic variants or gene variants) / (CFTR) (Cystic fibrosis or CF) / CFTR) (pancreatitis) (mutations or variants) / (CFTR) (congenital absence of vas deferens or CAVD) (mutations or gene variants or variants) / (CFTR) (variation or polymorphism) / (CFTR) (missense mutation) / (cystic fibrosis) (frequency) (mutations) / (CFTR) (gene functions). For a variant-specific search, a number of search strings can be applied utilizing gene name and descriptive phrases of the variant. The search strings vary according to the type of variants and correction factors. Legacy names of variants, if present, must always be included. The limitation of an internet search is that the variants listed only in tabular forms, image files, and supplementary materials may not be retrieved effectively. A stepwise flowchart for the curation of variants taking an example for the CFTR gene is presented in Figure S1.

Difficulties in curation and their possible solutions

Curation of variants requires meticulous work and the curator has to tackle many difficulties such as nomenclature issues, typos, and errors in papers [17]. Many problems can be gene-specific: for example, in the past, legacy names for HBB gene variants used to be assigned by geographical names such as Hb N-Baltimore, Hb D-LA, Hb O-Panjab, *etc*. In contrast, legacy names of CFTR variants followed the standard numbering of amino acids. Regular expression of variants and correction factors also differ among genes. During curation, nomenclature issues are frequent imposing major difficulties [17]. At times, variants are not decipherable by the standard nomenclature. Nomenclature issues should be prioritized at topmost for disambiguation because, if a variant that accompanies important information is not retrievable due to such an issue, interpretation of the variant in a clinical setting might be misleading.

Nomenclature issues

Variants are presented in a paper following a specific nomenclature. However, disparate conventions as well as nonstandard naming of variants across the literature may be encountered. To maintain uniformity, standard nomenclature recommended by HGVS (http://www.hgvs.org/mutnomen) should be followed [17]. Some of the important aspects of this nomenclature system are reiterated below.

(i) Numbering of variants at the cDNA level should start from the translation initiation site. Naming of substitution variants should be in the form 'c.# wild-type nucleotide>mutated nucleotide'(such as c.372G>A). For deletion and insertion variants, formats are 'c.#del (or ins) nucleotide'(such as c.42delC; c.1_2insC) for single nucleotide involvement, and 'c.#_#del (or ins) nucleotides' for multiple nucleotide involvement (such as c.1_2delAT and c.1_2insCCTAC). Large deletion and insertion variants

format is: 'c.#_#del (or ins) number of nucleotides' (such as c.20_60del42; c.20_21ins21). Though it is not obligatory, for deletion and insertion variants that involve fewer than 20 nucleotides, all deleted or inserted nucleotides should be mentioned. Wherever applicable, deletion and insertion variants especially in coding regions should be expressed at the most 3' position and insertion variants should be duplicated. However, an insertion variant in intronic or untranslated regions should not be 3-primed so as to extend to the coding region. For example, DSP variant c.-1_1insA should not be expressed as c.1dupA. For indel variants, format should be 'c.#delins nucleotide(s)' and 'c.#_#delins nucleotides' (*e.g.*, c.4delinsT; c.4_6delinsA; c.4_5delins40; c.del4_5CCinsA *etc*). An intronic substitution should be reported as the relative distance from exon-intron or intron-exon boundary (such as IVS10+1G>A; c.490+1G>A that lies downstream of the exon and IVS10-G>A; c.550-1G>A that lies upstream of the exon). For intronic deletions, insertion and indel variants, the same rules apply as in coding region; only the formats are different (examples: IVS10+25delA; c.490+25_490+26dupA; c.490+25_490+35delinsAAT; *etc*). For untranslated regions (UTR), the formats can be represented as c.-12A>G and *57C>G or c.*+57C>G for the 5'UTR and 3' UTR, respectively.

(ii) Amino acid numbering should start from the translation initiation codon counted as first amino acid. Use of three-letter codes for amino acids should be emphasized; however, one letter codes are also extensively used for presentational ease. For missense mutations, the format is 'p.wild-type amino acid# mutated amino acid' (such as p.Thr124Ile or p.T124I). Similarly, synonymous variants should be represented as p.Thr124Thr. Substitution at the initiation codon should be expressed as 'p.Met?' and at stop codon as 'p.#mutated amino acid ext*#' (p.117Trpext*25). For nonsense mutations, the format can be represented as Gln243X, p.Gln243* or p.Gln243Ter and for frameshift mutations, it is p.Gln243fsX35. For coding in-frame deletion variants, the formats are 'p.wild-type amino acid#del' and 'p.wild-type amino acid# _wild-type amino acid #del' (p.Gln243del, p.Gln243_Leu244del). In the same ways, formats can be represented as p.Gln243insGly, p.243Gln_244LeuinsGly, p.Gln243dup, p.Gln243_Leu244dup for insertion and duplication variants and as p.243delinsLeu and p.243Gln_244LeudelinsProGly for indel variants. For recessive diseases, two variants in the same allele should be expressed as p.[(variant;variant)] and in different alleles or in the compound heterozygous state as p.[(variant)];[(variant)]. When only one variant is identified the format is p.[variant];[?] and when one allele carries a variant while the other is normal, the format is p.[variant];[=].

(iii) Genomic numbering starts from the first nucleotide of the gene and should be represented by the suffix 'g.'. The format for genomic numbering involving substitution is 'g.# wild-type nucleotide>mutated nucleotide' (for example g.345555A>G).

Although the majority of papers use the translation initiation site as the start of numbering, depending upon genes, it is common to find papers that start numbering from other sites such as the transcription initiation site, specific domain of a protein, and signal or leader peptide in some proteins. Numbering system includes both plus and minus numerals excluding zero. Positional discordance of nucleotide or amino acid numbering between paper and reference sequence has to be resolved by calculating correction factor. For example, in the nomenclature of CFTR gene variants, although amino acid numbering starts from the translation start in both older and standard nomenclature systems, nucleotide numbering in the older nomenclature starts from the transcription initiation site which is 132 nucleotides upstream of the translation start [17]. Similarly, in LDLR gene, many older papers used numbering starting from the signal peptide which is 21 amino acids upstream of the initiation codon following the LDLR sequence described by Yamamoto et al. [18]. Mostly, the correction factor calculated for a gene applies to all its variants described in the literature.

Nucleotide numbering issues. In many papers, the amino acid numbering of variants is usually the same for papers as well as reference sequence but numbering of the nucleotide sequence may vary. Suppose in a paper, a variant in a gene is provided as: 608G>A (G202D). If the nucleotide at position 608 in that gene is 'C' in the reference sequence, numbering of the nucleotide is different between the paper and reference sequence. In the latter, if the amino acid at codon 202 is Gly, it means that numbering of the amino acid sequence is identical in both. If Gly 202 is encoded by GGC (c.604-606) in the reference sequence, the nucleotide change given in the paper therefore is c.605G>A in reference numbering. Thus the correction factor is paper -3 nucleotides= reference nucleotide. In the CFTR gene, nucleotide numbering in older nomenclature starts from the transcription start site which is 132 nucleotides upstream of the 'A' of initiation codon. Hence, +1 nucleotide in standard numbering represents +133 in older numbering. Thus +133, +134, +135, +136, +137,... in older numbering converts to +1, +2, +3, +4, +5,..., respectively, in standard numbering which gives correction factor: paper -132 nucleotides= reference nucleotide. Similarly, +132, +131, +130, +129,... in older numbering converts to −1, −2, −3, −4,..., respectively, in standard numbering which gives the correction factor: paper -133 nucleotides=reference nucleotide [19]. In some papers only legacy names of variants are given resulting in ambiguity at first sight, for example, T/TG counts at intron 9 of CFTR and legacy names for HBB variants. Standard names for such variants can be determined by observing their descriptions in multiple papers published over a range of years and by using information from public databases.

Amino acid numbering issues. To describe variants at the amino acid level, authors may not always use standard numbering for some genes. Suppose for a gene, variants are mentioned in a paper as G204A, R216W, K236X,

and M250V. In the reference protein sequence, if these amino acids are found at codons 201, 213, 233 and 247, respectively, the correction factor is paper-3 amino acids = reference amino acid. Use of the correction factor in papers may be different depending on the positions because there is no zero in numbering schemes (*i.e.*, 1 is followed by −1). For example, suppose numbering schemes of amino acids of a protein in nomenclature systems A and B are −3, −2, −1, 1, 2, 3,... and −1, 1, 2, 3, 4, 5,..., respectively, such that amino acid '3' in nomenclature system A represents amino acid '5' in nomenclature system B and so on. In this situation, the correction factor to get from A to B is A + 3 = B if A is negative and B is positive (*e.g.*, at A = −2); but it is A + 2 = B if both A and B are either negative (*e.g.*, at A = −3) or positive (*e.g.*, at A = 1). For example, in growth hormone 1 (GH1) gene, numbering of the amino acid in the new nomenclature system starts from initiation codon of leader peptide which is 26 amino acids upstream of the numbering start in old nomenclature. In old numbering system, a leader sequence used to be indicated by minus numbering. Hence +1 in old numbering is +27 (1 + 26) in reference numbering and −1, −2, −3, −4, −5, −6,... in older numbering system converts to +26, +25, +24, +23, +22, +21,..., respectively, in reference numbering system, giving the correction factor paper+27 amino acids = reference amino acid. Similarly, +1, +2, +3, +4, +5... in older numbering converts to +27, +28, +29, +30, +31,..., respectively, in reference numbering system, giving the correction factor paper+26 amino acid = reference amino acid [20].

Degenerate codon. Papers that mention variants only at the amino acid level may result in ambiguity due to the presence of degenerate codons when deriving such variants at the cDNA level. For example, if Leu (suppose encoded by TTA) is mutated to Phe, the change will either be A > C or A > T. Here, either both nucleotide changes have to be considered or the corresponding author of the paper has to be contacted to ascertain the cDNA change. Public databases can also be checked to clarify such ambiguous variants. Since it is DNA which is sequenced and not protein, authors should understand that giving only amino acid changes is providing the consequences only, not the actual changes. If possible, it is good practice to mention nucleotide as well as amino acid changes of the coding variants. Some synonymous variants can form cryptic splice-sites for which giving nucleotide changes is more relevant. For example, hemoglobin beta (HBB) gene variant c.75T>A leads a silent change (p.Gly25Gly) but the variant is demonstrated to result in skipping of exon 2 by activating a cryptic splice-site and hence is disease-causing [21].

Inter-conversion of nucleotide numbering. Exonic as well as intronic variants should preferably be mentioned at the cDNA level to reflect their effects on translation. Knowing relative positions of intronic variants from exon–intron boundaries is informative because variants at splice-sites are deleterious. For example, it is better to express a CFTR variant as c.489+1G>T than its genomic numbering (AJ574942.1) g.240G>T [13]. To get cDNA numbering of a genomic variant, manual as well as automated sequence alignment can be done. For manual conversion, alignment of genomic variant over full length cDNA sequence should be performed using a repetitive sequence a few nucleotides upstream or downstream of the variant. Genomic numbering can also be changed to cDNA numbering by software tools that perform the task by a sequence alignment process. By utilizing Mutalyser (https://www.mutalyzer.nl/), numbering of a genomic variant can easily be converted to its cDNA numbering.

Referencing issues of single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) may be presented in NCBI dbSNP as referenced complementarily to that of reference sequence. To get the correct nucleotide change and the frequency data of alleles, alignment of sequence flanking the SNP with the reference sequence is required. For example, FASTA sequence of rs35062203 (ALMS1 gene) in dbSNP is given as GGGGAGAATG[S]TTTT CTCTCA (8822-8842) where 'S' represents C or G at c.8832. The reverse complementary sequence is TGAGA GAAAA[S]CATTCTCCCC. The reference sequence for ALMS1 NM_015120 in this region is TGAGAGAAA A(C)CATTCTCCCC where C is present at 8832. Since negative strand has been referenced in this case, variant allele (*i.e.,* minor allele) in FASTA sequence is C at 8832. Hence, while using the variant allele's frequency information from population diversity data forc.8832C>G, data of C must be used, not of G. Thus, dbSNP data are strand-dependent which is indicated on 'Primary Assembly Mapping' row at SNP to Chr column. The UCSC genome browser provides reference sequences as well as dbSNP dataset aligned to the positive strand (http://www.genome.ucsc.edu/cgi-bin/hgTables?command=start), which can be used directly without need for sequence alignment.

Others

A gene can have more than one alias but papers may not be consistent in using a particular name for it. To standardize the use of gene symbols, the human genome organization gene nomenclature committee (HGNC) approved that gene symbols (http://www.genenames.org/) should be used by all authors, curators and databases. Furthermore, for some genes, the curator has to discriminate pseudogene-specific variants that have no pathological roles by examining the experimental set-up in the paper. Similarly, a protein can have more than one isoform; hence, the standard mRNA isoform of the protein should be predetermined. Papers usually provide information about reference proteins (either mRNA accession or UniProt or Ensemble ID); if such information is not given, a few other papers related to that protein need to be examined to determine the standard isoform of the protein. Primarily, the longest and

most predominantly expressed isoform of a protein should be utilized for curation.

Because primary research papers provide original reports of variant-related information, they should be prioritized for the curation. It is often found that, if an original paper has mentioned variants only at the amino acid level, writers citing the variants from that paper do not mention the variants at the cDNA level. Similarly, if variants in a primary paper are ambiguous due to nomenclature or other issues, authors of secondary papers may not resolve the ambiguities. It should be the responsibility of authors who cite the primary papers to clarify any ambiguous variants. Even if colloquial and/or legacy names of variants need to be retained, the authors should provide their standard names as well.

In papers reporting a large number of variants, positional and alphabetical typos for the amino acid and nucleotide changes may be encountered. For papers that provide both amino acid and DNA changes of the variants, the numbering that is consistent with that of the reference sequence should be used to derive their standard names. In rare cases, lack of uniformity in using single letter code for amino acids might be encountered in old papers. For example, papers may use 'L' to denote leucine or lysine, 'B' for aspartic acid or asparagine, 'P' for proline or phenylalanine and 'Z' for glutamate or glutamine. Many papers are also written in native languages that contain genetic variants which need to be translated to the language curators require.

Automated curation

To curate variants, automated curation tools use expression patterns such as annotation of variants, contextual features, distance-metrics, graph-metrics, and rule-based systems such as pattern matching. For example, two common tools MEMA and MuteXt have been developed that can use a dictionary search for protein and gene names and differentiate protein names based on the measurement of word proximity distance for extraction of variants: a variant in a paper is closer to the names of its related genes/proteins rather than names of other proteins/genes that are also present in the paper [16,22]. There are many automated tools that mainly differ on extraction strategies and efficiency. A list of common automated tools and their extraction strategies are briefly presented in **Table 1**. Though variants curated using automated tools need to be validated for false positive (FP) and false negative (FN) results, less time and resources needed make them attractive for the construction of large LSDBs. Many of them just curate the variants mainly from Medline abstracts but a few have been reported that can curate the variants from full text papers [16]. Moreover, automated tools that can check the nomenclature of the variants have also been developed. Mutalyzer (http://www.lovd.nl/mutalyzer/), a web-based software application, can be used for assessment of nomenclature of the variants extracted from publications [2]. Similarly, a web-based software application called COMUS can detect not only variants from sequencing files (AB1 files

Table 1 Common automated mutation curation tools and their extraction strategies and quality measures

Tool	Extraction approach	Extraction pair	Literature set used	Quality measures (P; R; F)	Refs
MuteXt	Regular expression, word proximity, Swiss-Prot entry	Variant-protein (at amino acid level)	GPCR and NR protein related full texts and abstracts	0.87; 0.87; U[#]	[26]
MEMA	Regular expression, word proximity	Variant – gene (at amino acid and DNA levels)	Medline abstracts	0.93;0.35;U[*]	[4,16]
Mutation GraB	Regular expression, graph metric, sequence check	Variant–protein–organism (at amino acid level)	Full text articles	0.84;0.90;0.87	[16]
Mutation miner	Regular expression, sentence co-mention	Variant-organism (at amino acid level)	Abstracts	0.91;0.46;0.61	[10,16]
Mutation finder	Regular expression	Gene-variant (at amino acid level)	Full text articles	0.98;0.81;0.81	[31]
Yip et al., 2007	Regular expression, rule-based system	Gene-variant (at amino acid level)	Full text articles	0.89;U;U	[32]
coagMDB	Regular expression, graph metric, sequence check	Gene-variant (at amino acid level)	Full text articles; serine protease	87-93;96-99;U	[33]
MuGeX	Regular expression	Gene-variant (at protein and DNA levels)	Medline abstracts; Alzheimer's disease associated genes	88.9;91.3;U	[34]
Krallinger et al., 2009	Regular expression, residue disambiguation and classification	Gene-variant (at protein level); natural vs artificial variants	Abstract and full text articles; kinase protein	72;U;U and 93.88;U;U for natural vs artificial variants	[35]
PolySearch	Sentence co-mention, word association	SNP detection; gene-variant	Abstracts, full text articles	U;U;U	[36]

Note: U indicates undetermined; [#], G-protein-coupled receptor (GPCR) mutations; NR, nuclear hormone receptor. [*] For example, when 100 abstracts were tested by MEMA for cited mutations in one letter code for variant-gene extraction pair, the quality measures, P and R values, were 0.93 and 0.35, respectively. P, precision; R, recall; F, F-score. See more details in the text.

Table 2 Descriptive patterns or expression of variants in literature

Level		Descriptive phrases
Amino acid	Missense	p.Arg30Gln; Arg30Gln; R30Q; Arg30 to Gln; Arg30 > Gln; Arg^{30} → Gln; Arg_{30}Gln; Arg30toGln; Arg(30)-Gln; Arg-30 → Gln; Arg30 → Gln; Gln30; 30Gln; Arg/Gln at codon 30; Arginine to Glutamine (substitution) at (codon) 30; Arg > Gln change at amino acid 30; Glutamine for Arginine at residue 30; RQ30; Q30; Arg30 → Gln
	Nonsense	R30X; p.Arg30X; R30Ter; R30*. R30Stop
	Frameshift	R14fsX4; DeltaR30; ΔR30; 30delArg; Ins30Arg; deletion (or insertion) at codon 30
	Silent	R30R; Arg30Arg; p.Arg30=; p.Arg30Arg.
DNA	Substitution	c.90G>A; G90A; 90G/A; G-90>A; 90→A; 90G-A; G(90)→A; UTR: c.-90G>A; -90 G→A; G to A at -90; c.*90G>A; c.*+90G>A
	Frameshift	c.90delG; c.90del; 90delG; c.90insG; 90insG; 90del2; 1-bp del, 90G; c.89_90insG; c.89_90delinsA; 90delinsA; c.90dupG; insertion of G at position 90; Arg30fsX2; R30fs; insertion (or duplication) of G at position 90; deletion of 2 bp at codon 30.
	Intronic	A to G at splice acceptor of intron 2; IVS31AS, A-T, -2; 3061(-1)G –> A; IVS32DS, G-A, +1; IVS2-2A>G; IVS2+1G>A; IVS2+1(G>A); Intron 2 nt-51A>G; 401(-1)G –> A; IVS1, G-A, -1; c.400+30A/G; c.400+30A>G; 400+30A>G; c.-8C>G; Intron 2 (-8G->A); IVS1+15del3; 400+30delG; 400+30insG; *etc.*
	Large deletions/ duplications	### bp deletion; del exon 1, c.del exons 2_4; c.dup exons 2_4; *etc.*
SNP		rs# or ss#; for example: rs5495
Haplotype		Haplotype description is gene/locus specific; for example, (TG)m(T)n, *i.e.,* TG and T repeats at intron 9 of CFTR gene: 7T, 5T, 5T/TG10, *etc.*

from Sanger sequencing) but also check the HGVS compliant nomenclature of variants [23].

Parameters for automated curation

Variants can be described mainly at five levels: (i) DNA level, (ii) amino acid level, (iii) SNP (in the form of rs or ss number), (iv) haplotype and (v) RNA level. Though variants are predominantly described at DNA and amino acid levels, authors also present some variants as haplotype, rs record and RNA change. Authors provide the DNA and/or amino acid change of variants mentioned at RNA level. Apart from a specific nomenclature, variants are also mentioned in papers using annotations or phrases which form various expression patterns. The various descriptive ways by which variants are presented in biomedical literature are given in **Table 2**. For automated curation tools, these form signaling texts which have to be read and extracted precisely [17,24,25]. Various automated tools use different strategies to extract variants but organism-gene-variant extraction pair is the most specific. To exclude FP results, point mutation-like terms and overlapping expressions (**Table 3**) should be created along with their contextual meanings and used for the validation of variants curated by automated tools [13,26].

Three parameters—precision, recall and F-score can be used to assess the quality of curation by automated tools [16]. Manual curation results serve as the gold standard or reference for evaluation of these quality measures. Suppose an automated tool has an extraction strategy of variant-protein-gene: variants that the automated tool as well as manual curation assign to a same protein/gene are classified as true positive (TP); variants that the tool assigns to a protein/gene but are manually classified as discordant are classified as false positive (FP); and variants that are manually classified as TP, but assigned to the wrong protein/gene by the tool are classified as (FN). Precision is calculated as $P = TP/(TP + FP)$; recall $R = TP/(TP + FN)$; and $F\text{-score} = 2 \times P \times R/(P + R)$. F-score is a number between 0 and 1, and a value of 1 is attained only when an automated tool produces neither FP nor FN. P, R and F can also be applied for extraction of variants alone.

Table 3 Mutation-like terms and overlapping names that need to be validated against the variants curated by automated tools

Item	Description
Cell lines	T47D (breast cancer cells); L5178Y (lymphoblasts); C33A (human cervical cancer cells); V600E (BRAF thyroid cancer cells); H293K, T98G (Human glioblastoma cell lines); M14T (T-cell line); H294R (adrenocortical line); A375M, F30K, F5K, T14D, T24C, T20C (cancer cell lines); *etc.*
Gene names	L23A, E2F, H4M, ER, *etc.*
Protein names	A2V, S100D, S100C, S100E, P34S(sperm surface protein), C184L, A10L(viral), A11L(viral), A52R(viral), *etc.*
Taxonomic entities	*Escherichia coli* K12S; *A. Viscous* T14V, *Pneumocystis pneumoniae* R36A, *A. Naeslundii* T14V, *Mycoplasma* spp. G145T, *Aeromonas* spp. F713E, *Bacillus* spp. G100I, Candida spp. N12C, *Synecoccus* spp. D120S, *Symbiodinium* spp. H10K, Yeast strain S288C, clone identifiers (eg W12I and W12E), plasmids (eg *E. coli* plasmids P15A), transgenic mouse model G93A), *etc.*
Overlapping names	A13G, C13T.
Others	M24R (filter); A83586C (antibiotic), A27L (immunogen), A9145C (antifungal), *etc.*

Limitations of automated curation

Some inherent limitations of automated tools that reduce effectiveness are as follows. (i) They are not devised to collect information on experimental procedures and effects of the variants [26]. (ii) They may not effectively extract variants listed in the figures in the papers. (iii) They have primarily been optimized to extract point mutations only. (iv) They may not grab the artificial variants cited in papers properly. For example, (a) a variant can be mentioned as CTGTA(G)TGTGT to CTGTA(A)TGTGT where G is changed to A and such a variant may be missed by the tool; (b) similarly, a paper may mention 'functional importance of alanine at position 234 by changing it into His, Arg, Cys and Trp'. Such artificial variants are worthy of curation due to their functional implication(s) but are prone to be missed by the automated curation tools). (v) Precision and recall rate may vary depending upon the type of literature sets and proteins.

Interpretation of variants

Interpretation of a genetic variant primarily depends upon its type and all available information on its pathogenicity determinants. Variant-related information is predominantly extracted manually [5]. A list of various types of information that needs extract during interpretation of a variant is presented in Figure S1. All information about pathogenicity determinants of a variant should be integrated in order to come to a correct yet safe conclusion on pathogenicity of the variant.

LSDBs (such as Leiden's open variation database (LOVD), human gene mutation database (HGMD)), dbSNP as well as genetic testing centers interpret variants in a spectrum of interpretation such as: 'pathogenic'; 'probably pathogenic'; 'possibly pathogenic'; 'variant of unknown significance' (VUS); 'possibly nonpathogenic'; 'probably nonpathogenic'; and 'nonpathogenic' [27]. Though the phrases used can differ across the fields concerned, their essence remains the same: HGMD uses 'disease-causing' for 'pathogenic'. While some variants are easy to interpret, interpretation of others might not be straightforward [7,28]. There are also published guidelines for the interpretation and reporting of the variants by ACMG (American College of Medical Genetics) [29]. Nonsense mutations, frameshift mutations, splice-site mutations, large insertion, deletion, indel, duplication mutations, and point mutations at initiation and stop codons are easy to interpret because their effects are deleterious. Missense mutations are the hardest ones to interpret; it is desirable to interpret them as VUS if supporting information is lacking [29,30]. Similarly, if supporting information is lacking for intronic and small in-frame insertion and deletion variants in the coding region, the safest interpretation of them that we can suggest is VUS. Similarly, it is safer to interpret novel synonymous variants with no supporting data as VUS because they can be suspected of forming cryptic splice-sites [21]. The same applies for deep intronic variants—this is supported by the fact that there are many disease causing variants reported deep in the introns of genes like HBB and CFTR (http://www.globin.bx.psu.edu/hbvar/menu.html and www.genet.sickkids.on.ca/, respectively).

Updates of information about a variant over a time may bring about conflict during interpretation of the variant. However, we can safeguard the interpretation so that updates of information do not alter the interpretation so drastically (pathogenic spectrum to nonpathogenic spectrum and vice versa) as to mislead the clinical decisions. Novel missense mutations pose difficulty in making clinical decisions. *In silico* predictions and conservativeness of such variants may help to determine the pathogenicity. If possible, segregation analysis in the family and observation in the ethnicity-matched healthy population should be done to determine the pathogenicity.

The interpretation of a genetic variant by genetic testing, database and research body may not be the same. To standardize interpretation of variants, our suggestions are: (i) there should be universal standard guidelines about the process of determining the pathogenicity of the variants; (ii) a variant should be interpreted in such as way that update in future will not change the interpretation drastically (like a variant previously interpreted as 'pathogenic' should not be 'nonpathogenic' in future) and (iii) a variant which has not been classified as 'not disease causing' or 'disease causing' (*e.g.*, a wide spectrum of interpretation between these two) should be re-interpreted from time to time owing to update of information.

Conclusion

The curation of genetic variants from literature is an essential part of genetic testing as well as various other researches including building of mutational databases. Ideally, it is imperative that curation should produce 100% sensitivity and specificity. Studying the genotype-phenotype correlation or the diagnostic significance of variants is challenging and requires specific as well as reproducible results. It is crucial that papers describing the genetic variants get clear exposure to the information seekers. To avoid or at least reduce difficulties in curation, authors should write papers with variants focusing primarily on standard nomenclature and become more specific by mentioning variants at cDNA level. In recent years, the numbers of papers following the standard nomenclature of variants is increasing which is making curation more effective and easier. Moreover, the authors should avoid mentioning the variants only at amino acid level. Apart from this, to help for effective retrieval and curation, journals and publishers should urge authors to follow the standard nomenclature such as HGNC and HGVS for the gene and variant names, respectively, as conditions for the publication of papers [11]. These practices will increase accuracy, speed, sensitivity, and specificity in manual as well as in

automated curation. For large scale curation, prospective research in bioinformatics should focus on integration of automated and manual curation methods that will obviate the limitations of either approach. To bring uniformity in interpretation of variants, standard guidelines for the interpretation of variants should be developed.

Competing interests

The authors declare that no competing interests exist.

Acknowledgements

We thank Dr. Ute Geigenmuller (Director, Molecular Genetics and Genomics Laboratory at Channing Lab, Boston, MA, USA) and Dr. Heidi Rehm (Assistant Professor of Pathology, Brigham and Women's Hospital and Harvard Medical School, Cambridge, MA, USA) for their constructive suggestions.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2012.06.006.

References

[1] Bale S, Devisscher M, Van Criekinge W, Rehm HL, Decouttere F, Nussbaum R, et al. MutaDATABASE: a centralized and standardized DNA variation database. Nat Biotech 2011;29:117–8.
[2] Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in variant databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 2008;29:6–13.
[3] Gieger C, Deneke H, Fluck J. The future of text mining in genome-based clinical research. Biosilico 2003;1:97–102.
[4] Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Comput Biol 2003;10:821–55.
[5] Van Auken K, Jaffery J, Chan J, Muller HM, Sternberg PW. Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC Bioinformatics 2009;10:228.
[6] Mitropoulou C, Webb AJ, Mitropoulous K, Brookes AJ, Patrinos JP. Locus-specific database domain and data content analysis: evolution and content maturation toward clinical use. Hum Mutat 2010;31:1109–16.
[7] Vihinen M, den Dunnen JT, Dalgleish R, Cotton RGH. Guidelines for establishing locus specific databases. Hum Mutat 2012;33:298–305.
[8] Fokkema IFAC, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v. 2.0: the next generation in gene variant databases. Hum Mutat 2011;32:557–63.
[9] Mathiak B, Eckstein S, editors. Five steps to text mining in biomedical literature. Proceedings of the second European workshop on data mining and text mining in bioinformatics. Italy: Pisa; 2004.
[10] Baker CJO, Witte R. Mutation mining—a prospector's tale. Inf Syst Front 2006;8:47–57.
[11] Nature Genetics Editorial. Conventional wisdom. Nat Genet 2010;42:363.
[12] Hunter L, Cohen KB. Biomedical language processing: perspective what's beyond PubMed? Mol Cell 2006;21:589–94.
[13] Ogino S, Gulley ML, den Dunnen JT, Wilson RB. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. J Mol Diag 2007;9:1–6.
[14] McDonald R, Winters RS, Ankuda CK, Murphy JA, Rogers AE, Pereira F, et al. An automated procedure to identify biochemical papers that contain cancer-associated gene variants. Hum Mutat 2006;27:957–64.
[15] Celli J, Dalgleish R, Vihinen M, Taschner PE, den Dunnen JT. Curating gene variant databases (LSDBs): toward a universal standard. Hum Mutat 2012;33:291–7.
[16] Lee LC, Horn F, Cohen FE. Automatic extraction of protein point mutations using a graph bigram association. PLoS Comput Biol 2007;3:84–95.
[17] den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex variants: a discussion. Hum Mutat 2000;15:7–12.
[18] Yamamoto T, Davis CG, Brown MS, Schneider WJ, Casey ML, Goldstein JL, et al. Thehuman LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA. Cell 1984;39:27–38.
[19] Sugarman EA, Rohlfs EM, Silverman LM, Allitto BA. CFTR mutation distribution among US Hispanic and African American individuals: evaluation in cystic fibrosis patient and carrier screening populations. Genet Med 2004;6:392–9.
[20] Millar DS, Lewis MD, Horan M, Newsway V, Easter TE, Gregory JW, et al. Novel mutations of the growth hormone 1 (GH1) gene disclosed by modulation of the clinical selection criteria for individuals with short stature. Hum Mutat 2003;21:424–40.
[21] Goldsmith ME, Humphries RK, Ley T, Cline A, Kantor JA, Nienhuis AW. "Silent" nucleotide substitution in a beta+ thalassemia globin gene activates splice site in coding sequence RNA. Proc Natl Acad Sci U S A 1983;80:2318–22.
[22] Nagel K, Jimeno-Yepes A, Rebolz-Schuhman D. Annotation of protein residues based on a literature analysis: cross validation against UniProtKb. BMC Bioinformatics 2009;10:S4.
[23] Jho S, Kim BC, Ghang H, Kim JH, Park D, Kim HM, et al. COMUS: clinician-oriented locus-specific mutation detection and deposition system. BMC Genomics 2009;10:S35.
[24] Caporaso JG, Deshpande N, Fink JL, Bourne PE, Cohen KB, Hunter L. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. In: Proceedings of PSB online, Department of Bioengineering. Stanford, CA: Stanford University; 2008. p. 640–51.
[25] Rebholz-Schuhmann D, Marcel S, Albert S, Tolle R, Casari G, Kirsch H. Automatic extraction of mutations from Medline and cross-validation with OMIM. Nucleic Acids Res 2004;32:135–42.
[26] Horn F, Lau AL, Cohen FE. Automated extraction of variant data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 2004;20:557–68.
[27] Guerreiro RJ, Baquero M, Blesa R, Boada M, Brás JM, Bullido MJ, et al. Genetic screening of Alzheimer's disease genes in Iberian and African samples yields novel mutations in presenilins and APP. Neurobiol Aging 2010;31:725–31.
[28] Tavtigian SV, Greenblatt MS, Goldgar DE, Boffetta P. Assessing pathogenicity: overview of results from the IARC unclassified genetic variants working groups. Hum Mutat 2008;29:1261–4.
[29] Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hedge MR, et al. ACMG recommendations for interpretation and reporting of sequence variations: revisions 2007. Genet Med 2008;10:294–300.
[30] Kohonen-Corish MRJ, Al-Aama JY, Auerbach AD, Axton M, Barash CI, Bernstein I, et al. How to catch all those mutations – the report of the third Human Variome Project Meeting, UNESCO Paris, May 2010. Hum Mutat 2010;31:1374–81.
[31] Caporaso JG, Baumgartner Jr WA, Randolph DA, Cohen KB, Hunter L. Mutation finder: a high-performance system for extracting point variant mentions from text. Bioinformatics 2007;23:1862–5.
[32] Yip YL, Lachenal N, Pillet V, Veuthey AL. Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot knowledgebase. J Bioinform Comput Biol 2007;5:1215–31.

[33] Saunders RE, Perkins SJ. CoagMDB: a database analysis of missense mutations within four conserved domains in five vitamin K-dependent coagulation serine proteases using a text-mining tool. Hum Mutat 2008;29:333–44.

[34] Erdogmus M, Sezerman OU. Application of automatic mutation-gene pair extraction to diseases. J Bioinform Comput Biol 2007;5:1261–75.

[35] Krallinger M, Izarzugaza JMG, Rodriguez-Penagos C, Valencia A. Extraction of human kinase mutations from literature, databases and genotyping studies. BMC Bioinformatics 2009;10:S1.

[36] Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, variants, drugs and metabolites. Nucleic Acids Res 2008;36:W399–405.