

Abstract

Current Mutation Database Quality

Steve Callaghan, Victorian Bioinformatics Consortium, Monash University, VIC 3800, Australia

The primary aim of the Mutation Database Initiative is to provide a single point of entry for and enquiry on all known variant (mutation) data by integrating the existing general mutation and locus specific databases (LSDBs). A past survey of existing general mutation databases [1] describes at a high level what data each contains and how they may be searched. The later survey of LSDBs [2] reports against each of a list of characteristics the percentage of LSDBs that possess that characteristic. Neither survey goes to the depth of assessing data quality or quality of individual query types so we initiated a more detailed survey. This is ongoing but results to date indicate that the existing databases are in general not structured in a manner that facilitates quick queries of the form 'What phenotypes are related to BRCA1 c.5382insC?' and 'What variants are related to prostate cancer?', and there is little if any provision for synonymous phenotype terms. The survey shows further that the reference sequences used as the basis of nucleotide and residue numbering in variant names (descriptions formed according to the standard nomenclature [3]) on some of the databases are not easily (if at all) identifiable. So enquirers may be uncertain as to whether the same reference sequence has been used for position numbering of all the known variants for any particular gene. Similar uncertainty arises where a clinician or researcher has identified a variant (and named it using a particular reference sequence) and then searches the existing databases to see if the variant has been reported before. A variant name based on one reference sequence is likely to be different from a name for the same variant based on another reference sequence, especially if the variant is toward the 3' end of the gene. The approach we are taking to assess the extent of inconsistent numbering is to compare variant names between databases (general and locus specific) and to compare variant names for consistency with reference sequences according to RefSeq (found via LocusLink). For example a variant named c.1357delTAAAG is consistent with a RefSeq only if the 5 nucleotides of the latter starting at position 1357 are indeed TAAAG. Not all variant names can be checked for consistency in this way: DNA variant names for insertions for example. The survey to date shows an unexpectedly high degree of consistency of nucleotide and residue numbering - both between names for the same variant on different databases and between the databases and RefSeqs surveyed. But whenever a reference sequence is superseded by a new sequence then inconsistency between new and existing variant names is likely to arise. And to determine whether a variant recorded in one database is the same as a variant recorded in another is sometimes difficult and always time-consuming. The upshot is that for an enquirer to find all known data concerning a particular variant they must not only identify all the databases that may contain the variant but may also have to ensure that any *apparent* matches between the query variant name and variant names on the databases are *real* matches (rather than accidental matches of position numbers based on different reference sequences). Finding data on phenotypes related to the variant then requires following links to publications and reading and digestion of large amounts of text to extract the required knowledge. Finding variants relating to a particular phenotype requires that the enquirer perform queries for all synonyms for the phenotype, for example 'gastric cancer' and 'stomach cancer'. These processes that enquirers must repeatedly endure could be done just once as part of a data integration exercise - the Mutation Database Initiative. Achieving this integration would mean that enquirers would have a single point of enquiry on all known variant data and would be to quickly retrieve all known variant - phenotype relationships. The positive finding of the survey is that the integration may not involve so much re-numbering of variant positions as we might have expected.

- [1] Porter CJ, Talbot Jr CC, Cuticchia AJ (2000) Central mutation databases - A review. Hum Mut 15: 7-1236-44
- [2] Claustres M., Horaitis O., Vanevski M., and Cotton R.G.H. (2002) Time For A Unified System Of Mutation Description and Reporting: A Review of Locus Specific Mutation Databases. Genome Res 12:680-688
- [3] <http://www.genomic.unimelb.edu.au/mdi/mutnomen/>