DATABASES

UMD (Universal Mutation Database): 2005 Update

Christophe Béroud,[1]* Dalil Hamroun,[1] Gwenaëlle Collod-Béroud,[1] Catherine Boileau,[1,2,3] Thierry Soussi,[4] and Mireille Claustres[1]

[1]*Laboratoire de Génétique Moléculaire, IURC, Montpellier, France;* [2]*INSERM U383, Hôpital Necker Enfants Malades, Clinique Maurice Lamy, Paris, France;* [3]*Laboratoire de Biochimie, d'Hormonologie et de Génétique Moléculaire, Hôpital Ambroise Paré, Boulogne, France;* [4]*Laboratoire de Génotoxicologie des Tumeurs, EA 3493, Service de Pneumologie, Hôpital Tenon, Paris, France*

Communicated by Marc Greenblatt

With the completion of the Human Genome Project, our vision of human genetic diseases has changed. The cloning of new disease-causing genes can now be performed in silico, and thousands of mutations are being identified in diagnostic and research laboratories yearly. Knowledge about these mutations and their association with clinical and biological data is essential for clinicians, geneticists, and researchers. To collect and analyze these data, we developed a generic software called Universal Mutation Databases (UMD®) to create locus-specific databases. Here we report the new release (September 2004) of this freely available tool (www.umd.be), which allows the creation of LSDBs for virtually any gene and includes a large set of new analysis tools. We have implemented new features to integrate noncoding sequences, clinical data, pictures, monoclonal antibodies, and polymorphic markers (SNPs). Today the UMD retains all specifically designed tools to analyze mutations at the molecular level, as well as new sets of routines to search for genotype–phenotype correlations. We also created specific tools for infrequent mutations such as gross deletions and duplications, and deep intronic mutations. A large set of dedicated tools are now available for intronic mutations, including methods to calculate the consensus values (CVs) of potential splice sites and to search for exonic splicing enhancer (ESE) motifs. In addition, we have created specific routines to help researchers design new therapeutic strategies, such as exon skipping, aminoglycoside read-through of stop codons, or monoclonal antibody selection and epitope scanning for gene therapy. Hum Mutat 26(3), 184–191, 2005. **© 2005 Wiley-Liss, Inc.**

KEY WORDS: database; mutation detection; UMD; SNP

INTRODUCTION

Almost 50 years after the DNA double helix was discovered by James Watson and Francis Crick [Watson and Crick, 1953], the International Human Genome Sequencing Consortium announced the successful completion of the Human Genome Project [Abramowicz, 2003]. The tremendous work performed to sequence the 3 billion DNA letters in the human genome has changed our vision of human genetic diseases. New disease-causing genes can now be cloned in silico, and a new field of mutations (intronic mutations) is emerging for already known genes. Concomitantly with this sequencing effort, many biotechnology companies have produced new tools to rapidly scan large sets of samples for mutations. There are new technologies for high-throughput SNP typing, DHPLC, and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF). Every year thousands of variations are thus identified in diagnostic and research laboratories. Knowledge about these variations and their association with clinical and biological data is essential for clinicians, geneticists, and researchers.

The Human Genome Organization-Mutation Database Initiative (HUGO-MDI) was established in the mid-1990s to collect and analyze these data [Cotton, 2000; Cotton et al., 1998]. Since those early days, two main approaches have been developed: 1) "core" databases (also known as central databases), and 2) "locus-specific" databases (LSDBs). Core databases collect published mutations from all genes. The archetype is the Human Gene Mutation Database (HGMD; www.hgmd.org). By December 2004, the database contained an excess of 44,090 different lesions detected in 1,714 different nuclear genes [Stenson et al., 2003]. Each mutation is entered only once in order to avoid confusion between recurrent and identical-by-descent lesions. Furthermore, the phenotypic description associated with each mutation is very limited, which precludes any study on phenotypic variability. These databases are frequently referred as "mile wide and inch deep databases" [Auerbach, 2000] because they include mutations from many genes, but only limited descriptions. Conversely, an LSDB collects all published and unpublished mutations from a specific gene, and the annotation of each mutant includes a full molecular, biological, and phenotypic description. Experts (also known as "curators") validate these data. This validation process is critical for maintaining high-quality data, since up to 10% of the information in various publications is erroneous. The curators are also essential for standardizing the clinical and biological descriptions for each patient. In addition, each mutation can occur more than once in a particular position and hence appear more than once in the database, which allows the identification of

Received 11 October 2004; accepted revised manuscript 16 March 2005.

*Correspondence to: Christophe Béroud, Laboratoire de Génétique Moléculaire, IURC, 641 Avenue du Doyen G. Giraud, 34093 Montpellier, France. E-mail: christophe.beroud@igh.cnrs.fr

Grant sponsor: Association Française Contre les Myopathies (AFM).
DOI 10.1002/humu.20210
Published online in Wiley InterScience (www.interscience.wiley.com).

mutation hot spots. These databases are referred to as "inch wide and mile deep databases" [Auerbach, 2000] because they include a broad range of high-quality data for a specific gene.

To federate this new field of bioinformatics (mutations study), the Human Genome Variation Society was created in 2001 [Horaitis and Cotton, 2004]. Following HGVS guidelines, we modified the original Universal Mutation Database generic software (UMD®), which allows the creation of LSDBs for virtually any gene and includes a large set of analysis tools [Beroud et al., 2000]. The availability of the human genome sequence led us to implement new features that allow the integration of noncoding sequences (intronic, and 5′ and 3′ sequences) as well as SNPs from the dbSNP (www.ncbi.nlm.nih.gov/SNP). To analyze these new types of mutations, we developed a set of tools to facilitate the interpretation of intronic mutations. In addition, we created specific routines to help researchers design new therapeutic strategies, such as exon skipping, aminoglycoside read-through of stop codons, and monoclonal antibody selection and epitope scanning to monitor gene therapy. For clinicians, we created routines to look for genotype–phenotype and phenotype–genotype correlations. Finally, we implemented new graphical displays to illustrate many different analyses.

Here we report the new release (September 2004) of this freely available tool, which can be downloaded at www.umd.be.

MATERIALS AND METHODS

Database Structure

The UMD® software was developed using the 4th Dimension® language from 4D (www.4D.com). While this language is not frequently used to build genetic databases, and is often compared unfavorably with other solutions (such as MySQL and PHP), we believe the 4th Dimension® language is one of the few products that provide a complete data set for the development of databases and web servers. Furthermore, it employs a broad language that includes hundreds of commands to work with spreadsheets, graphics, pictures, and web pages.

An analysis of the LSDBs available on the Internet (www.hgvs.org/) reveals that most databases display mutations as simple tables that can be obtained from various search engines [Claustres et al., 2002]. A minority of them have clinical or biochemical features and provide access to graphical displays (www.phdb.mcgill.ca [Scriver et al., 2003]; http://imgt.cines.fr [Lefranc et al., 2003]), such as the distribution of mutations, and only the UMD-LSDBs provide access to dynamic graphical displays. This feature results from the unique structure of the UMD® software.

The previous version of the UMD included only the reference coding sequence of each gene because only a few intronic sequences were available. With the completion of the human genome project, intronic sequences are now available for almost all genes. We therefore decided to include this information. We created an import interface that allows an easy input from the NCBI Genome Annotated Genomic Contig named NT_xxxxxx (for example, the genomic contig for the VHL gene is NT_022517) of intronic sequences, 5′ and 3′ sequences, and SNPs. An automatic process searches for exonic sequences in the contig and extracts noncoding sequences. Furthermore, it collects all annotated SNPs and calculates their position in the corresponding noncoding segment. These data are stored in two new tables (intronic sequences and polymorphisms). A third table includes information about monoclonal antibodies (the monoclonal antibody's name, clone, type, epitope, and the first and last amino acids from the epitope). A fourth table was created to include a virtually unlimited number of pictures (depending on the storage capacities of the computer) to illustrate a specific sample, such as immunostaining, dHPLC, or specific clinical features. A fifth table includes pedigree pictures to illustrate family trees. Finally, we added a sixth table, called mutant activities, to store information about in vitro activity of mutants. This table has proved to be very useful for TP53 analysis [Soussi et al., 2004].

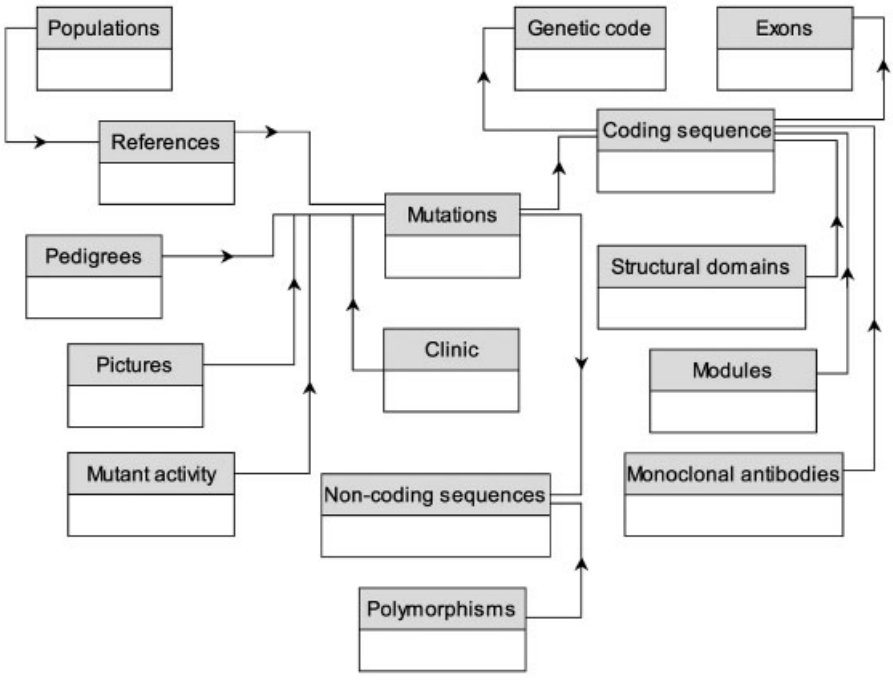

FIGURE 1. **The relational database structure of the UMD. Each rectangle represents a table. Each table contains a variable number of fields (up to 75 for the mutation table). Arrows indicate links between the tables.**

Similar data have been already well documented for other databases, such as the PAHdb [Scriver et al., 2003]. In order to build true LSDB generic software, we modified the exon table to include noncoding exons, which have been described for some genes. The new UMD structure is shown in Figure 1.

Specific Features for Intronic Mutations and Splicing Consequences

The introduction of noncoding sequences paved the way for intronic mutation analysis. First, to avoid typing errors, an automatic check is performed for each input. Second, the consensus values (CVs) of potential donor and acceptor splice sites in the vicinity of the mutation are calculated according to the CVs for each nucleotide at each splice site's position. The global splice-site CVs are then calculated using an algorithm derived from Senapathy et al. [1990] and Cartegni et al. [2002]. Results for the wild-type and mutant sequences are displayed as either a graph (Fig. 2) or a table.

In the past few years, major progress has been achieved in identifying regulatory elements involved in the splicing machinery that is localized in the introns or exons. Various exonic splicing enhancers (ESEs) have been identified, and it has been shown that they correspond to binding sites for specific serine/arginine-rich (SR) proteins, a family of structurally related and highly conserved splicing factors that are characterized by one or two RNA-recognition motifs (RRM) and a distinctive C-terminal domain highly enriched in RS dipeptides (the RS domain). The RRMs mediate sequence-specific binding to the RNA and thus determine substrate specificity, whereas the RS domain appears to be involved mainly in protein–protein interactions. SR proteins bound to ESEs can promote exon definition by directly recruiting the splicing machinery through their RS domain and/or by antagonizing the action of nearby silencer elements. To analyze the potential loss or gain of ESEs, we used previously established sequences to score probable ESE motifs of four human SR proteins: SF2/ASF, SC35, SRp40, and SRp55 [reviewed by Cartegni et al., 2002]. This feature displays ESE site differences between wild-type and mutant sequences as a graphic display. Color codes allow for the rapid identification of variations between the two sequences (data not shown). If this information cannot be used to interpret any variants in the absence of functional data, the prediction tool will be progressively optimized.

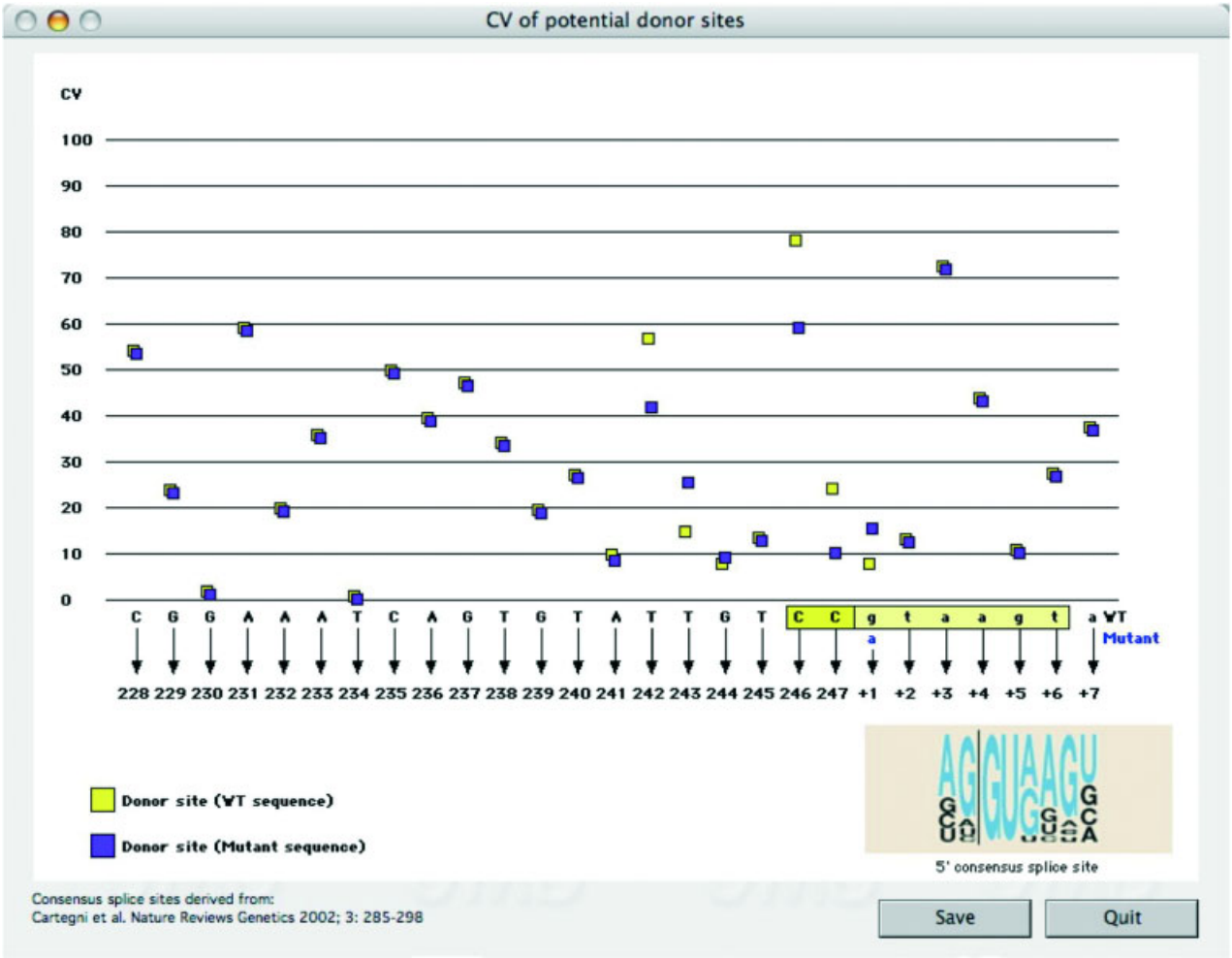

FIGURE 2. **Consequences of the IVS2+1G>A mutation from the *FBN1* gene. Exonic sequence = capital letters with corresponding nucleotide numbers (228 to 247); Intronic sequence = small letters with corresponding nucleotide numbers (+1 to +7). The normal donor splice site is shown in yellow (dark yellow for exonic nucleotides, light yellow for intronic nucleotides). The CV for each potential donor splice site (eight-nucleotide sequence) in the vicinity of the mutation is displayed according to Cartegni's rule (100 = strong splice site; 0 = not a splice site). Wild-type sequence CV = yellow square; mutant sequence CV = purple square. Each square represents the CV of the potential splice site beginning with the corresponding nucleotide. For example, the fourth nucleotide (A at position 231) gives the CV value of the potential splice site: AAATCAGT.**

It has been known for many years that intronic mutations can disrupt the splicing machinery. These mutations are usually localized in the vicinity of splice sites; however, recent findings have identified deep intronic mutations leading to the incorporation of cryptic exons in the mRNA. The first such observation was made in the β-globin gene [Treisman et al., 1983]. Since then, a few examples have been reported in other genes, such as ornithine-∂-aminotransferase [Mitchell et al., 1991], *CFTR* [Chillon et al., 1995; Highsmith et al., 1994], estrogen receptor [Wang et al., 1997], β-glucuronidase [Vervoort et al., 1998], *NF1* [Ars et al., 2000], *ATM* [Pagani et al., 2002], α-galactosidase [Ishii et al., 2002], and *DMD* [Beroud et al., 2004; Tuffery-Giraud et al., 2003]. To facilitate the interpretation of such intronic mutations, previously described functions (splice and ESE analysis) can be performed on any intronic sequence or cryptic exon.

Specific Routines for Designing Therapeutic Strategies

One of the major purposes of an LSDB is to provide valuable information for researchers who are attempting to develop new therapeutic strategies to cure patients. In collaboration with experts from the neuromuscular field, we developed specific routines for exon-skipping strategies, aminoglycoside read-through of stop codons, and monoclonal antibody selection and epitopes scanning for gene therapy. These tools are available in the generic software and can potentially be used for all genes.

Exon Skipping

The archetype for exon skipping is the DMD gene associated with Duchenne and Becker muscular dystrophy. Most of the *DMD* mutations consist of large genomic deletions. The "out-of-frame" deletions elicit the formation of premature stop codons and consequent abortion of the translation process, which result in dystrophin deficiencies and severe phenotypes (e.g., Duchenne muscular dystrophy). In contrast, deletions that produce "in-frame" mRNAs leading to shorter proteins are responsible for a milder myopathy (e.g., Becker muscular dystrophy). In many DMD patients, as well as in animal models (e.g., the *mdx* mouse and the GRMD dog), rare dystrophin-positive fibers have been reported [Crawford et al., 2001; Wilton et al., 1997]. It has been suggested that restoration of the reading-frame by exon skipping is the most likely cause of this natural phenomenon. This has prompted many groups to investigate the possibility of designing strategies for gene repair/modulation based on the use of compounds that interfere with splicing and thus induce exon skipping [Aartsma-Rus et al., 2004; Goyenvalle et al., 2004; Kapsa et al., 2003]. Because the *DMD* gene contains 79 exons, more than 3,000 potential transcripts can be produced by exon skipping (one or more exons can be additionally deleted), and should be investigated to search for frame restoration. Therefore, we developed an automatic tool that displays the largest in-frame protein resulting from exon skipping (i.e., the lowest number of additional deleted exons). Other exon-skipping combinations that potentially can restore the frame are also directly available. These data can be used to select specific exon skipping, which results in frame restoration for most patients.

Stop Codons

Over the last decade, studies have demonstrated that the genetic code may be more flexible than was previously supposed. For example, the ribosome can change the decoding frame during elongation, either forward or backward, or even skip a part of the message. These alternative readings of the genetic code were termed "recoding" by Gesteland et al. [1992]. Usually these mechanisms result in the bypass of a stop codon and the synthesis of a longer polypeptide than that produced by conventional decoding. These alternative ways of reading the genetic code are programmed by signals present in specific mRNAs at defined locations in the messenger. Such recoding events generally occur in competition with standard decoding, and allow the synthesis of two or more polypeptides, at a defined ratio, from a single mRNA molecule (for review see Namy et al. [2004]). Today the RECODE database of translational recoding events (programmed ribosomal frameshifting, codon redefinition, and translational bypass) is available at http://recode.genetics.utah.edu/ [Baranov et al., 2003]. This recoding has gained attention with the demonstration that aminoglycosides can suppress disease-causing nonsense mutations [Howard et al., 2000; Kerem, 2004; Sleat et al., 2001; Wilschanski et al., 2003]. Thus, knowledge about nonsense mutations is crucial for developing certain therapeutic approaches.

We designed simple tools to identify all codons from a specific gene that can produce a stop codon by a simple mutational event. Only a few of these potential stop codons are found in patients and are therefore eligible for read-through strategies. The various tools included in the UMD software (such as "potential stop codons" or "PSC surrounding sequences") should provide valuable help in addressing this issue.

Monoclonal Antibody Selection and Epitope Scanning for Monitoring Gene Therapy

The idea behind gene therapy is to induce cells to produce gene products that will replace proteins that are nonfunctional due to gene mutation. In such experiments, the introduction of the wild-type gene (or cDNA) can be monitored using monoclonal antibodies that recognize epitopes absent from the endogenous mutated protein. With the increasing number of described mutants for a single gene and monoclonal antibodies, it became apparent that the availability of an automatic tool to search for antibodies, such as those with an epitope localized in the lost part of a mutant protein, could be very useful for experiment design. We thus created a specific table to store data for monoclonal antibodies, and specific routines to search for specific epitopes localized in deletions. These data are automatically processed during mutation input and can be displayed in a table or graphic format. Additionally, the user can search for mutant proteins that harbor or lack a specific epitope. This last option can be found in the "specific tools" category, and can be activated for a specific gene upon request.

Genotype–Phenotype Correlations

Genotype–phenotype correlations and phenotype–genotype correlations are probably the most challenging aspects of LSDB. With the rapidly growing collection of mutant descriptions (including clinical data), it is clear that the development of predictive medicine is under way. In the near future, it is reasonable to believe that specific mutation profiles associated with specific clinical features will be identified, as has been reported for various genes, such as the *VHL* gene [Gallou et al., 1995]. This knowledge will benefit both patients (e.g., by enabling better screening and reducing hospitalizations) and the medical community (e.g., by focusing on specific routine analyses, and reducing biological and medical investigations).

The UMD structure includes a clinical table linked to the mutations table (Fig. 1). This simple relation allowed us to develop genotype–phenotype and phenotype–genotype tools. The first one

allows the selection of a specific phenotype and displays, for each clinical symptom, the distribution of the various associated severities. This analysis may be performed on a subset of the database. The mirroring tool displays the various clinical symptoms with the distribution of the various associated severities. The user can select one or more clinical symptoms, and the software will display the various combinations of symptom/severity with the associated genotypes. As before, this analysis can be performed on a subset of the database.

These analyses can only be performed on high-quality data. Obtaining a clinical description has always been a challenge for databases. To solve this problem, several UMD curators developed a network of international experts in specific clinical fields and asked them to fill out a questionnaire. The UMD software allows the easy input of such text tab delimited files.

Other New Analysis Tools

Geographic distribution of mutations. Analyses of human diseases have shown that relatively common disease-causing mutations can result from a common ancestor. One of the best examples is the Phe508del mutation that causes cystic fibrosis, which occurs at the highest frequency in Denmark (87.2%) and the lowest in Algeria (26.3%) [Estivill et al., 1997]. This geographic distribution indicates that the disease arose from a common ancestor, distinct from any present European group, and spread throughout Europe. To facilitate such studies, we included a new tool that can display the distribution of mutations on a world map. Each country is defined as an object, which can be specifically addressed by the software to modify its color. The user can choose which mutations should be selected for the analyses, and whether the resulting values (incidence) should be displayed either as crude data or relative to a mutation type (all, missense, deletions, insertions, nonsense, and out-of-frame mutations).

Haplotypes. With the development of rapid scanning and sequencing technologies, it is now much easier to study a large genomic fragment. This has led to the identification of numerous polymorphisms that can be used to define a haplotype. We have included haplotype-dedicated routines in the new release of the UMD software. A new field, called "allele," was added to the mutation table. The user can thus specify for each genetic variant (either mutation or polymorphism) the chromosome with which it is associated (maternal or paternal). The software thus reorders the variants according to their position on the gene and defines the haplotypes. Specific tools allow the analysis of haplotype–haplotype associations, allele–allele associations, or disequilibrium between specific variants and haplotypes (allele–haplotype associations). These data can also be used to build phylogenic trees (data not shown).

Large rearrangements. Small rearrangements account for most mutations (92.4%; 25,266 missense or nonsense, 4,182 splice, 483 regulatory, 7,387 small deletions, 2,912 small insertions, and 428 small indels); however, large rearrangements have also been reported (7.6%; 406 gross insertions and deletions, 514 complex rearrangements including inversions, and 2,421 gross deletions). These data were extracted from the HGMD database (12/13/2004 version; www.hgmd.org) [Stenson et al., 2003]. While this last group of mutations generally occurs infrequently, for some genes it represents the majority of mutations. The archetype is the DMD gene, for which 60% of mutations are large deletions and 5% are large duplications. To facilitate the input of such mutations, we included an easy to use interface. The user selects the extent of the large rearrangement at the exonic level (for example, a deletion from exon 4 to exon 8) and the software computes the nomenclature at the cDNA level, as well as transcriptional and translational impact of the mutation and the theoretical exon-skipping pattern that restores a reading frame. In addition, polymorphic markers located in the deleted portion of the gene are directly accessible. Furthermore, monoclonal antibodies with epitopes fully or partially disrupted by the mutation are also available. While this is not useful for most genes, in which large rearrangements can affect the folding of the total protein, it is valuable for proteins with repeated motifs, such as dystrophin. Finally, we developed two new analysis tools called "deletion and duplication distribution" and "deletion map." The first one gives access to a graphical presentation of the extent of the large rearrangements at the exonic level, with the number of reported mutations. The user can select a specific portion of the gene and/or a specific type of mutation. The second one allows a graphic display of large deletions at the protein level. The user can now specify a given color code for already defined structural domains. The software then builds a scheme with various objects corresponding to structural domains and exons. Two scales (amino acids and nucleotides) are also displayed. For each deletion, its consequence at the translational level is shown as well as the extent of the deletion. Moreover, if monoclonal antibody epitopes have been defined, zooming in on the deleted region shows the epitope extent of the various antibodies (Fig. 3).

SNPs. We developed an automatic system to collect all polymorphisms from the NCBI annotated contig. A link is provided by the SNP database and is directly accessible on the web version of the software. In addition, we added a graphic display module to show the distribution of polymorphisms along the contig that harbors the gene of interest. Various zoom-in/zoom-out functions allow one to select a specific portion of the gene and obtain information about individual polymorphisms.

Modules

The *FBN1* gene is composed of repeated modules. We previously designed specific tools for the UMD-FBN1 database to align these modules and display the distribution of mutations on the conserved or specific amino acids [Collod-Beroud et al., 1998]. To provide access to this specific tool for all UMD-LSDBs, we used the 4D-view productivity plug-in, which adds spreadsheet functionality. Thus the user can define conserved residues for a repeated module by selecting defined modules and labeling each conserved residue. This spreadsheet can be stored for analysis.

DISCUSSION

The creation of the UMD tool was initiated 10 years ago, and the first generic UMD software was released in 2000 [Beroud et al., 2000]. During this period many developments have been made, and the human genome project has released an almost complete sequence of the human genetic code. Changes in the database structure allowed us to evolve and add this new information. The first release of the UMD was limited to the coding sequence and splice sites of a gene, and thus did not include large deletions or intronic mutations. This problem has been overcome with the new release of UMD, which includes intronic and regulatory sequences. Other major developments include the introduction of clinical data, pictures, monoclonal antibodies, and polymorphic markers. In addition, automatic imports from the NCBI annotated genomic contig sequences allow the SNPs to be easily updated.

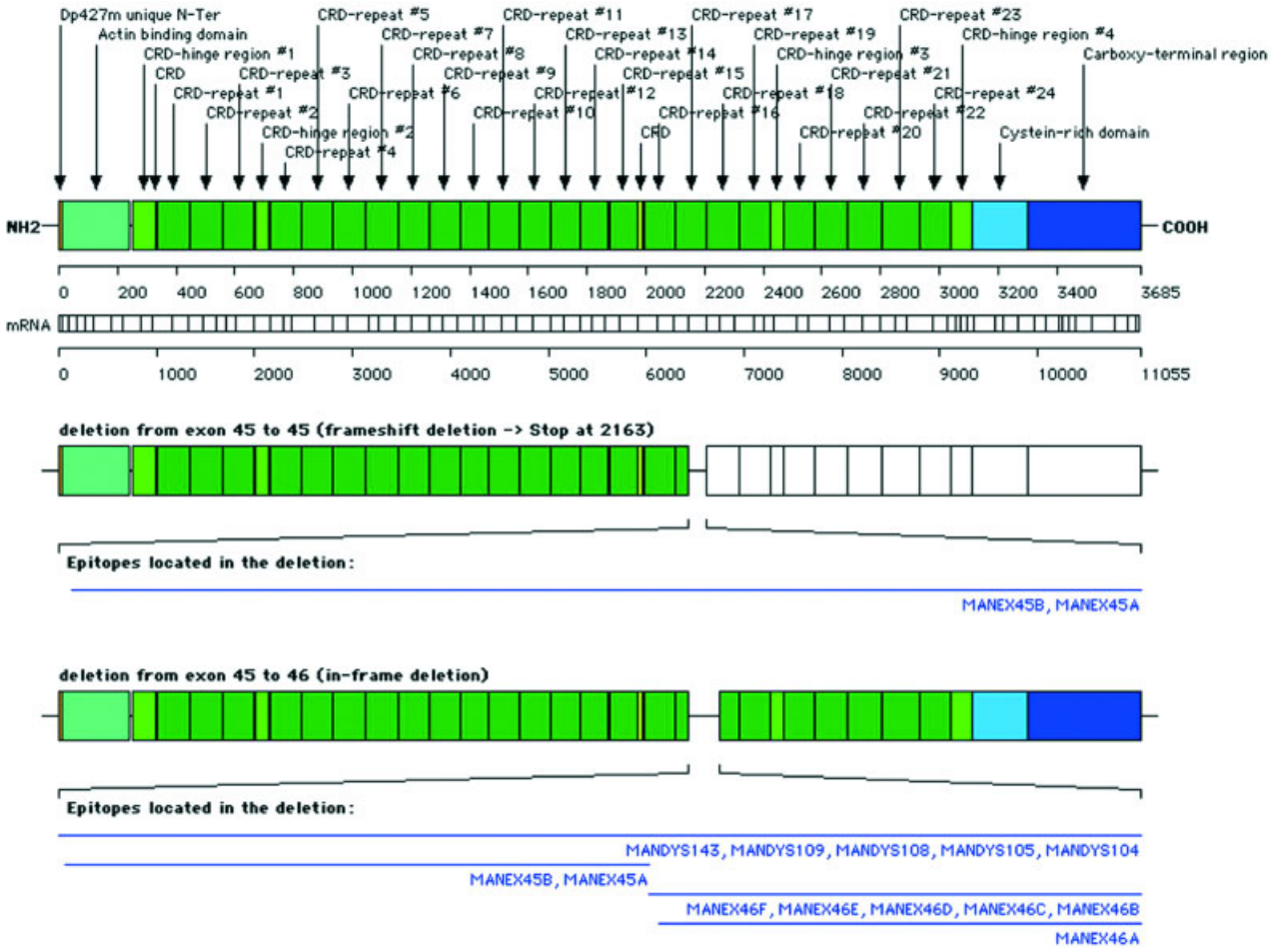

FIGURE 3. **Deletion map routine for the DMD gene. Color squares = structural domains; empty squares = exons; 0–3685 = amino acid scale; 0–11055 = nucleotide scale; blue text = monoclonal antibodies with epitopes disrupted by the mutation. For each mutation the remaining domains of the mutant protein are presented in colors, while absent domains (resulting from a premature stop codon) are presented as empty boxes. For an in-frame deletion (deletion of exons 45–46), all remaining domains appear in color. The empty space between the colored squares represents the extent of the deletion at the mRNA level.**

Since the beginning we have focused on creating specific analysis tools. We have come a long way since the first analysis of the TP53 gene [Beroud et al., 1996] and the development of molecular epidemiology. Today the UMD retains all of the specifically designed tools to analyze mutations at the molecular level, as well as new sets of routines to search for genotype–phenotype correlations. With the creation of more than 20 LSDBs with the UMD software, a large set of mutations is now included in the UMD-LSDBs (more than 28,000). We also created specific tools for infrequent mutations, such as gross deletions and duplications and deep intronic mutations. A large set of dedicated tools are now available for intronic mutations, such as methods to calculate the CVs of potential splice sites and to search for ESE motifs.

Other major innovations include the development of graphical displays for various analyses and, more recently, geographic distribution of mutations. Concomitantly, we created specific routines to help researchers design new therapeutic strategies, such as exon skipping, aminoglycosides read-through of stop codons, and monoclonal antibody selection and epitope scanning for gene therapy. These tools pave the way for clinical trials to assess associations between genotypes and drug response.

The UMD generic software is a reference tool to build LSDBs. It includes the largest set of analysis tools available and is adapted to genes involved in genetic diseases and cancers, the "gene-type" option allowing the activation of cancer-related fields. Since the beginning we have made this tool freely accessible. It can be downloaded from our website at www.umd.be. We also offer the various curators our hosting capacities for their UMD-LSDBs.

Finally, we want to emphasize that the potential of these LSDBs can only be realized with the use of high-quality data. The role of curators today is even more critical for the integration of clinical data. Consensus submission forms reduce the complexity of this task, but we regret the increasing number of poorly or even erroneous descriptions of mutations. We highly recommend that researchers use the international nomenclature for mutations (www.genomic.unimelb.edu.au/mdi/mutnomen/) [den Dunnen and Antonarakis, 2001], which will benefit the entire community.

ACKNOWLEDGMENTS

We thank Professor Jean-Claude Kaplan, Dr. Luis Garcia, Dr. Jean-Pierre Rousset, Dr. Sylvie Tuffery-Giraud, Dr. Irène Ceballos,

and Dr. Etienne Rouleau for helpful discussions. We also thank all of the curators of the various UMD-LSDBs.

REFERENCES

Aartsma-Rus A, Janson AA, Kaman WE, Bremmer-Bout M, van Ommen GJ, den Dunnen JT, van Deutekom JC. 2004. Antisense-induced multiexon skipping for Duchenne muscular dystrophy makes more sense. Am J Hum Genet 74:83–92.

Abramowicz M. 2003. The Human Genome Project in retrospect. Adv Genet 50:231–261; discussion 507–510.

Ars E, Serra E, Garcia J, Kruyer H, Gaona A, Lazaro C, Estivill X. 2000. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 9:237–247.

Auerbach AD. 2000. 8th International HUGO-Mutation Database Initiative Meeting, April 9, 2000, Vancouver, Canada. Hum Mutat 16:265–268.

Baranov PV, Gurvich OL, Hammer AW, Gesteland RF, Atkins JF. 2003. Recode 2003. Nucleic Acids Res 31:87–89.

Beroud C, Verdier F, Soussi T. 1996. p53 Gene mutation: software and database. Nucleic Acids Res 24:147–150.

Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. 2000. UMD (Universal Mutation Database): a generic software to build and analyze locus-specific databases. Hum Mutat 15: 86–94.

Beroud C, Carrie A, Beldjord C, Deburgrave N, Llense S, Carelle N, Peccate C, Cuisset JM, Pandit F, Carre-Pigeon F, Mayer M, Bellance R, Recan D, Chelly J, Kaplan JC, Leturcq F. 2004. Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene. Neuromuscul Disord 14:10–18.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298.

Chillon M, Dork T, Casals T, Gimenez J, Fonknechten N, Will K, Ramos D, Nunes V, Estivill X. 1995. A novel donor splice site in intron 11 of the CFTR gene, created by mutation 1811+ 1.6kbA–>G, produces a new exon: high frequency in Spanish cystic fibrosis chromosomes and association with severe phenotype. Am J Hum Genet 56:623–629.

Claustres M, Horaitis O, Vanevski M, Cotton RG. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 12:680–688.

Collod-Beroud G, Beroud C, Ades L, Black C, Boxer M, Brock DJ, Holman KJ, de Paepe A, Francke U, Grau U, Hayward C, Klein HG, Liu W, Nuytinck L, Peltonen L, Alvarez Perez AB, Rantamaki T, Junien C, Boileau C. 1998. Marfan database (third edition): new mutations and new routines for the software. Nucleic Acids Res 26:229–233.

Cotton RG. 2000. Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. Hum Mutat 15:4–6.

Cotton RG, McKusick V, Scriver CR. 1998. The HUGO mutation database initiative. Science 279:10–11.

Crawford GE, Lu QL, Partridge TA, Chamberlain JS. 2001. Suppression of revertant fibers in mdx mice by expression of a functional dystrophin. Hum Mol Genet 10:2745–2750.

den Dunnen JT, Antonarakis SE. 2001. Nomenclature for the description of human sequence variations. Hum Genet 109:121–124.

Estivill X, Bancells C, Ramos C. 1997. Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. The Biomed CF Mutation Analysis Consortium. Hum Mutat 10:135–154.

Gallou C, Joly D, Mejean A, Staroz F, Martin N, Tarlet G, Orfanelli MT, Bouvier R, Droz D, Chretien Y, Marechal JM, Richard S, Junien C, Beroud C. 1999. Mutations of the VHL gene in sporadic renal cell carcinoma: definition of a risk factor for VHL patients to develop an RCC. Hum Mutat 13:464–475.

Gesteland RF, Weiss RB, Atkins JF. 1992. Recoding: reprogrammed genetic decoding. Science 257:1640–1641.

Goyenvalle A, Vulin A, Fougerousse F, Leturcq F, Kaplan JC, Garcia L, Danos O. 2004. Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping. Science 306:1796–1799.

Highsmith WE, Burch LH, Zhou Z, Olsen JC, Boat TE, Spock A, Gorvoy JD, Quittel L, Friedman KJ, Silverman LM, Boucher RC, Knowles MR. 1994. A novel mutation in the cystic fibrosis gene in patients with pulmonary disease but normal sweat chloride concentrations. N Engl J Med 331:974–980.

Horaitis O, Cotton RG. 2004. The challenge of documenting mutation across the genome: the human genome variation society approach. Hum Mutat 23:447–452.

Howard MT, Shirts BH, Petros LM, Flanigan KM, Gesteland RF, Atkins JF. 2000. Sequence specificity of aminoglycoside-induced stop condon readthrough: potential implications for treatment of Duchenne muscular dystrophy. Ann Neurol 48:164–169.

Ishii S, Nakao S, Minamikawa-Tachino R, Desnick RJ, Fan JQ. 2002. Alternative splicing in the alpha-galactosidase A gene: increased exon inclusion results in the Fabry cardiac phenotype. Am J Hum Genet 70:994–1002.

Kapsa R, Kornberg AJ, Byrne E. 2003. Novel therapies for Duchenne muscular dystrophy. Lancet Neurol 2:299–310.

Kerem E. 2004. Pharmacologic therapy for stop mutations: how much CFTR activity is enough? Curr Opin Pulm Med 10: 547–552.

Lefranc MP, Giudicelli V, Ginestoux C, Chaume D. 2003. IMGT, the international ImMunoGeneTics information system, http://imgt.cines.fr: the reference in immunoinformatics. Stud Health Technol Inform 95:74–79.

Mitchell GA, Labuda D, Fontaine G, Saudubray JM, Bonnefont JP, Lyonnet S, Brody LC, Steel G, Obie C, Valle D. 1991. Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation. Proc Natl Acad Sci USA 88:815–819.

Namy O, Rousset JP, Napthine S, Brierley I. 2004. Reprogrammed genetic decoding in cellular gene expression. Mol Cell 13: 157–168.

Pagani F, Buratti E, Stuani C, Bendix R, Dork T, Baralle FE. 2002. A new type of mutation causes a splicing defect in ATM. Nat Genet 30:426–429.

Scriver CR, Hurtubise M, Konecki D, Phommarinh M, Prevost L, Erlandsen H, Stevens R, Waters PJ, Ryan S, McDonald D, Sarkissian C. 2003. PAHdb 2003: what a locus-specific knowledge base can do. Hum Mutat 21:333–344.

Senapathy P, Shapiro MB, Harris NL. 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 183: 252–278.

Sleat DE, Sohar I, Gin RM, Lobel P. 2001. Aminoglycoside-mediated suppression of nonsense mutations in late infantile neuronal ceroid lipofuscinosis. Eur J Paediatr Neurol 5 Suppl A:57–62.

Soussi T, Kato S, Levy PP, Ishioka C. 2004. Reassessment of the TP53 mutation database in human disease by data mining

with a library of TP53 missense mutations. Hum Mutat 25: 6–17.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21: 577–581.

Treisman R, Orkin SH, Maniatis T. 1983. Specific transcription and RNA splicing defects in five cloned beta-thalassaemia genes. Nature 302:591–596.

Tuffery-Giraud S, Saquet C, Chambert S, Claustres M. 2003. Pseudoexon activation in the DMD gene as a novel mechanism for Becker muscular dystrophy. Hum Mutat 21: 608–614.

Vervoort R, Gitzelmann R, Lissens W, Liebaers I. 1998. A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene. Hum Genet 103: 686–693.

Wang M, Dotzlaw H, Fuqua SA, Murphy LC. 1997. A point mutation in the human estrogen receptor gene is associated with the expression of an abnormal estrogen receptor mRNA containing a 69 novel nucleotide insertion. Breast Cancer Res Treat 44:145–151.

Watson JD, Crick FHC. 1953. Molecular structure of nucleic acids. Nature 171:737–738.

Wilschanski M, Yahav Y, Yaacov Y, Blau H, Bentur L, Rivlin J, Aviram M, Bdolah-Abram T, Bebok Z, Shushi L, Kerem B, Kerem E. 2003. Gentamicin-induced correction of CFTR function in patients with cystic fibrosis and CFTR stop mutations. N Engl J Med 349:1433–1441.

Wilton SD, Dye DE, Laing NG. 1997. Dystrophin gene transcripts skipping the mdx mutation. Muscle Nerve 20:728–734.