Banques de données de mutations : enjeux et perspectives pour les maladies génétiques orphelines

M Humbertclaude, Sylvie Tuffery-Giraud, C. Bareil, C. Thèze, P Paulet, D Desmet, D. Hamroun, David Baux, G Girardet, Gwenaëlle Collod-Béroud, et al.

Banques de données de mutations : enjeux et perspectives pour les maladies génétiques orphelines

Genetic mutation databases: Stakes and perspectives for orphan genetic diseases

V. Humbertclaude [a,b,*], S. Tuffery-Giraud [b,c], C. Bareil [a], C. Thèze [a], D. Paulet [c], F.-O. Desmet [b], D. Hamroun [a,b], D. Baux [a], A. Girardet [a], G. Collod-Béroud [b], P. Khau Van Kien [a], A.-F. Roux [a], M. des Georges [a,b], C. Béroud [a,b,c], M. Claustres [a,b,c]

[a] *Laboratoire de génétique moléculaire, CHU de Montpellier, 34000 Montpellier, France*
[b] *Inserm U827, 34000 Montpellier, France*
[c] *UFR de médecine, université Montpellier-1, 34000 Montpellier, France*

INFO ARTICLE

Mots clés :
Banques de données génétiques
LSDB
Mutations
Maladies génétiques

Keywords:
Human variation database
LSDB
Mutations
Genetic disorders

RÉSUMÉ

Des technologies toujours plus puissantes d'analyse des gènes et de leurs mutations ont contribué à l'accélération fulgurante de l'identification des gènes responsables de maladies humaines et de la révélation de la variabilité du génome humain. Le rôle des banques génétiques devient majeur pour transformer ces masses de données en « informations » utilisables par la communauté scientifique et médicale. Les banques gène-spécifiques (LSDBs) sont les plus informatives car maintenues par des curateurs experts et motivés. Cette revue présente les principaux types de banques de données de mutations responsables de maladies génétiques, les perspectives qu'elles ouvrent pour de nouvelles pistes thérapeutiques et les enjeux qu'elles représentent pour la médecine du futur.

ABSTRACT

New technologies, which constantly become available for mutation detection and gene analysis, have contributed to an exponential rate of discovery of disease genes and variation in the human genome. The task of collecting and documenting this enormous amount of data in genetic databases represents a major challenge for the future of biological and medical science. The Locus Specific Databases (LSDBs) are so far the most efficient mutation databases. This review presents the main types of databases available for the analysis of mutations responsible for genetic disorders, as well as open perspectives for new therapeutic research or challenges for future medicine. Accurate and exhaustive collection of variations in human genomes will be crucial for research and personalized delivery of healthcare.

1. Introduction

L'existence de sites portails tels Orphanet (http://www.orpha.net), Genetests ou Genreviews (http://www.ncbi.nlm.nih.gov/sites/GeneTests/) a considérablement facilité l'accès au diagnostic moléculaire de nombreuses maladies génétiques rares en Europe. Ces portails très utilisés ne sont cependant pas conçus pour répondre aux besoins spécialisés des laboratoires impliqués dans le diagnostic moléculaire ou la recherche dans le domaine des maladies rares. L'utilisation de technologies toujours plus performantes pour étudier les gènes génère en effet une énorme quantité de données, qui vont considérablement s'accumuler lorsque les techniques de séquençage à haut débit seront utilisables dans les laboratoires de diagnostic. Cette information sur la variabilité du génome humain normal et pathologique représente une valeur inestimable pour la médecine. Pour qu'elle soit utilisable, il est nécessaire de développer de robustes systèmes bio-informatiques (banques de données) avec lesquels des « curateurs » experts d'un gène ou d'une maladie pourront collecter, classer, archiver des données de qualité recueillies de façon homogène, leur assigner une signification biologique, les relier à des descriptions phénotypiques standardisées et les rendre ensuite accessibles à la communauté scientifique et médicale [1].

* Auteur correspondant.
Adresse e-mail : veronique.humbertclaude@inserm.fr (V. Humbertclaude).

Les banques de mutations et leurs outils d'analyse *in silico* facilitent l'interprétation des variations de séquences, contribuent à une meilleure compréhension des mécanismes mutationnels et apportent des informations irremplaçables sur la structure, la fonction et/ou l'expression du gène, ouvrant ainsi la voie à de nouvelles pistes thérapeutiques. Elles permettent également un partage de données standardisées entre cliniciens, biologistes et chercheurs spécialistes d'une maladie. Ce travail en réseau stimule le développement de stratégies consensuelles de diagnostic et de prise en charge des maladies héréditaires. L'accès libre et gratuit via Internet de données anonymisées permet enfin une diffusion rapide et évolutive des connaissances sur les gènes et les maladies.

L'objectif de cette revue est d'illustrer, à partir de quelques exemples, le contexte, les enjeux et les perspectives des banques de mutations.

2. Les principaux types de banques de données de mutations

On distinguait jusqu'ici les collections généralistes de données sur l'ensemble des gènes (« *mile wide and inch deep databases* ») et les banques spécialisées dans l'analyse d'un gène responsable d'une maladie génétique (« *inch wide and mile deep* ») [2] (Tableau 1) [3]. Un troisième type de banque s'est développé récemment, plus spécialisé dans les données ethniques et nationales [4]. Les banques de données existantes sont répertoriées dans le site web dédié « HGVbase » (Human Genome Variation base), développé par la Human Genome Variation Society (HGVS, http://www.hgvs.org), qui présente les règles internationales concernant la nomenclature des mutations ainsi que les critères de qualité et recommandations concernant les banques de mutations.

2.1. Banques de données centrales ou générales (core or central mutation databases)

Leur objectif est de collecter toutes les mutations décrites dans tous les gènes, chaque mutation étant décrite de façon succincte. Elles donnent une bonne vision globale des mutations ayant des conséquences cliniques. En général, les données proviennent de la littérature.

2.1.1. Online Mendelian Inheritance in Man (OMIM)

Online Mendelian Inheritance in Man est la version électronique de Mendelian Inheritance in Man (MIM), première banque de données génétiques développée par Victor McKusick dans les années 1960 [5]. OMIM recensait en avril 2009 près de 3800 maladies secondaires à des mutations dans 2250 gènes. Chaque entrée est associée à un texte détaillant le phénotype et divers liens (banques de données, séquences ADN ou protéines, références...). Les données sont issues de la littérature biomédicale et mises à jour régulièrement. Cependant, la collection des mutations est loin d'être exhaustive. La description phénotypique en texte libre rend plus difficile l'extraction des données par des outils d'analyse automatisés pour leur intégration dans d'autres banques. OMIM, « *knowledgebase* » de référence des maladies génétiques, est très utilisée pour la génétique clinique.

2.1.2. Human Gene Mutation Database

Human Gene Mutation Database (HGMD) contient les descriptions des mutations germinales de gènes nucléaires responsables de maladies humaines (85 000 mutations dans 3253 gènes en décembre 2008) [6,7] et leur classification selon les mécanismes mutationnels. Les données sont issues de la littérature, à partir d'une combinaison de procédures automatiques et manuelles, via une analyse de plus de 500 revues scientifiques et médicales. Chaque mutation n'est saisie qu'une fois, ce qui ne permet pas d'évaluer sa fréquence dans la population ni d'étudier des corrélations génotype–phénotype. Le répertoire des mutations n'est pas tenu à jour pour tous les gènes et les informations peuvent être fragmentaires. De plus, l'accès public libre et gratuit est retardé d'un an après la saisie dans la banque, HGMD ayant dû conclure un accord avec une entreprise commerciale pour des raisons budgétaires.

2.1.3. Autres banques

dbSNP : développé par le National Center for Biotechnology Information (NCBI) est le répertoire le plus complet des *single nucleotide polymorphisms* (SNP). D'autres types d'informations sont disponibles sur le portail HGVS ou sur des sites spécifiques tels HapMap.

PharmGKB est une banque de données d'accès public, fournissant des informations cliniques et génétiques concernant les variations génétiques individuelles impliquées dans les différentes réactions aux médicaments.

WayStation est un site permettant la soumission en ligne de variations génomiques. Après vérification par le comité de révision, l'information est transmise à la banque locus spécifique correspondante ou à une banque centrale. À la date du 26 juin 2009, 761 banques locus spécifiques étaient répertoriées.

2.2. Banques de données nationales et ethniques (National and Ethnic Mutation Databases)

Le spectre des mutations d'un gène et d'une maladie étant le plus souvent variable selon les populations et les groupes ethniques, l'intégration de ces données dans des banques est indispensable pour reconstituer l'histoire génographique des populations humaines. Les banques National and Ethnic Mutation Databases (NEMDB) sont aussi très utiles pour aider à l'interprétation de tests diagnostiques ambigus, optimiser le service de diagnostic moléculaire à l'échelle nationale et diffuser les informations adéquates auprès des professionnels de santé et des patients [4,8,9]. Treize NEMDB étaient actives en 2006 [4]. Certaines banques (National Mutation Frequency Databases) indiquent, par interrogation simple, la fréquence des principales mutations responsables de maladies génétiques « locales » dont le spectre mutationnel est bien caractérisé. D'autres banques (National Genetic Databases) présentent un répertoire détaillé des maladies génétiques d'une population ou d'un groupe ethnique, mais peu ou pas de données sur les mutations. La banque israélienne, qui utilise également le logiciel ETHNOS, contient des données sur 442 maladies. La banque Singapore Human Mutation/Polymorphism Database (SHMPD) analyse les fréquences des mutations et des polymorphismes héréditaires ainsi que des données issues d'études d'associations de gènes candidats [10]. Frequency of Inherited Disorders database (FINDbase) offre un accès à des fréquences de mutations pathogènes dans diverses populations à travers le monde, avec des données sur 32 maladies, 25 gènes et 98 groupes de population disponibles par interrogation simple [11].

Des critères de qualité spécifiques pour ce type de banque ont été récemment définis et un effort international pour harmoniser et coordonner les NEMDB est en cours [4].

2.3. Banques de données gène-spécifiques (Locus Specific Databases [LSDBs])

Ces banques sont les plus exhaustives car elles sont en général développées, entretenues et mises à jour par un ou plusieurs

Tableau 1
Principaux types de banques de données des mutations de maladies génétiques[a].

Banques de données	Sites web
Banques de données centrales ou générales	
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
Ensembl	http://www.ensembl.org/index.html
HapMap	http://www.hapmap.org
Human Gene Mutation Database (HGMD)	http://www.hgmd.cf.ac.uk/ac/index.php
OMIM	http://www.ncbi.nlm.nih.gov/omim/
PharmGKB	http://www.pharmgkb.org
Banques de données nationales ou ethniques (National/Ethnic Mutation Databases [NEMDBs])	
Frequency of INherited Disorders db (FINDbase)	http://www.findbase.org
Finnish Database	http://www.findis.org
National Genetic Databases	http://www.goldenhelix.org
National Mutation Frequency Databases	http://www.goldenhelix.org
SHMPD	http://shmpd.bii.a-star.edu.sg/
Banques de données gène-spécifiques (locus specific databases [LSDB]) : principaux logiciels	
LOVD	http://www.lovd.nl/2.0/
Mutation View (Keio Mutation Databases)	http://mutview.dmb.med.keio.ac.jp
UMD	http://www.umd.be
WayStation	http://www.centralmutations.org
Portail d'accès à l'ensemble des banques de données génétiques	
Human Genome Variation base (HGVbase)	http://www.hgvbaseg2p.org

[a] Le tableau n'inclut pas les banques spécifiques des variations structurales du génome humain (*copy number variations* [CNVs]), qui peuvent être consultées dans un article dédié [3].

experts ou « curateur(s) » ayant un intérêt particulier pour le gène. Une LSDB rassemble dans un même outil informatique et de façon standardisée les mutations et polymorphismes identifiés par les laboratoires dans les familles présentant une maladie mendélienne donnée. Les plus perfectionnées intègrent aussi des informations cliniques et biologiques utiles pour le diagnostic, l'évaluation du pronostic ou l'étude des corrélations génotype–phénotype [12], des informations sur l'origine géographique et/ou ethnique, la fréquence des mutations et variations dans la population, les mutations récurrentes et les points chaud mutationnels...

Les données sont issues de la littérature et/ou soumises directement au curateur, ce qui permet de recueillir un grand nombre d'informations qui seraient, sinon, perdues pour la communauté (on estime que plus de 50 % des données des laboratoires de diagnostic ne sont pas publiées). Les LSDB contiennent donc plus de données par gène que les banques centrales. Grâce aux efforts des curateurs qui vérifient chaque entrée, la qualité des informations est en général élevée [1] et supérieure à celle de la littérature spécialisée [13].

Un des problèmes rencontrés est de s'assurer que toutes les mutations et variations identifiées sont collectées. Certaines LSDBs s'organisent à un niveau national ou international, à travers un réseau rassemblant les laboratoires spécialistes du gène. Par exemple, la banque UMD-DMD France dédiée au gène *DMD* regroupe les données des 14 laboratoires français qui assurent le diagnostic moléculaire des dystrophinopathies [14]. Cette banque est reconnue comme un modèle de « *knowledgebase* » nationale en raison de la qualité et de l'exhaustivité des informations qu'elle contient (2405 mutations et patients en janvier 2009) [15].

Les 730 LSDB répertoriées en 2009 sur le site HGVS varient considérablement dans leurs objectifs, leur contenu, leur exhaustivité, le niveau d'expertise des curateurs et le temps consacré à la curation, ainsi que par le logiciel informatique utilisé (plus de dix logiciels différents) [1]. Faute de financement pérenne, certaines LSDBs ne sont pas mises à jour régulièrement, deviennent rapidement obsolètes ou même disparaissent [1,8]. De nombreuses LSDBs présentent les données sous la forme d'un simple fichier texte, les requêtes possibles sont souvent limitées à un ou deux critères simples, les résultats étant présentés sous forme de listes. Les logiciels spécialisés pour la création de LSDBs (Tableau 1) permettent des analyses plus complexes, multicritères et statistiques, les résultats apparaissant sous forme de graphiques ou de tableaux. Par exemple, le système Universal Mutation Database (UMD) [16] propose de nombreux outils d'analyse des données moléculaires (Fig. 1) [13,17–19].

Suite à l'analyse des LSDBs existantes en 2001, nous avions défini les éléments d'une LSDB idéale et insisté sur l'importance de regrouper les experts dans l'analyse du gène et les cliniciens spécialistes de la pathologie au sein d'un consortium permettant de créer un langage commun pour décrire, collecter, analyser les allèles, les génotypes et les phénotypes à l'échelle nationale et internationale [20]. Ce type d'organisation permettrait de réaliser les études de corrélations génotype–phénotype ainsi que les études d'épidémiologie moléculaire qui font défaut actuellement (distribution des allèles mutés et des allèles sauvages chez les individus atteints, les apparentés porteurs cliniquement asymptomatiques, les apparentés porteurs symptomatiques, les apparentés non porteurs et les individus de la population générale...). Le réseau multidisciplinaire pourrait être complété par des physiologistes cellulaires et des biochimistes pour aider à la classification des variants. Le financement pérenne de curateur(s) spécialiste(s) permettrait d'assurer le contrôle de chaque entrée, les corrections, les mises à jour régulières, le respect des critères de qualité [12], et l'évolution des systèmes de requête et d'exploitation de la banque.

2.4. Le projet HVP : intégration de toutes les banques de données

Sous l'impulsion de HGVS, le consortium international Human Variome Project (HVP : http://www.humanvariomeproject.org/) propose de créer une cyber-infrastructure qui serait capable d'établir la connexion entre toutes les banques centrales, nationales et gène-spécifiques, qui partageraient ainsi une architecture, des ontologies et des classifications semblables [21]. À plus long terme, il est envisagé de capturer les données directement à leur source. Ces projets se heurtent à de nombreux problèmes techniques (comment extraire les données de centaines de banques non standardisées), éthiques (législations très différentes entre les pays concernant des données informatiques) ou concernant la propriété intellectuelle des LSDBs, la reconnaissance du travail des curateurs et leur financement. Pour inciter le dépôt des données dans les banques centrales, un système de « *microattribution* » a même été imaginé [22].

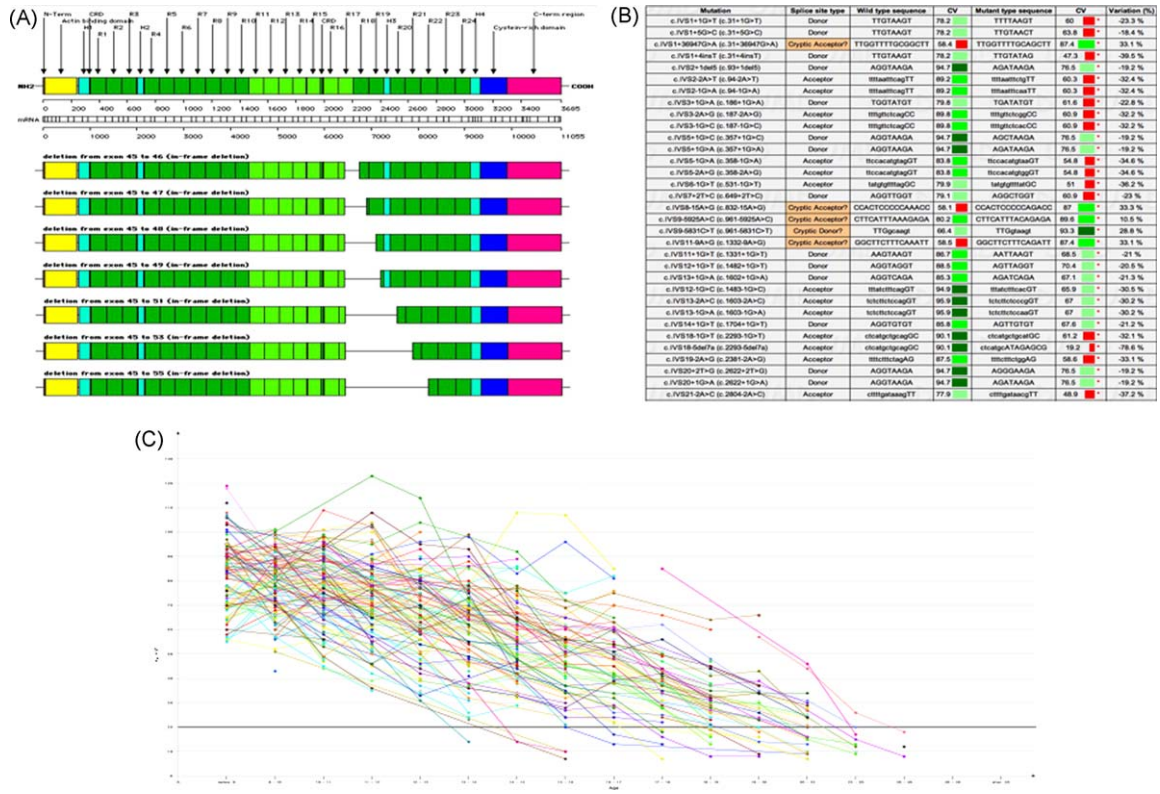

Fig. 1. Exemples d'outils d'analyse des génotypes (A, B) et phénotypes (C) disponibles dans des LSDBs (données extraites de la banque UMD-DMD France http://www.umd.be/DMD/). A. Protéines tronquées résultant de grandes délétions du gène *DMD*. Les différents domaines structuraux de la dystrophine sont présentés avec des codes couleur. Les deux protéines résultant des délétions des exons 45 à 53 et 45 à 55 sont considérées comme les meilleures candidates pour la thérapie par saut d'exons multiples [17,18]. B. Conséquences des mutations introniques sur les signaux d'épissage du gène *DMD*. Les valeurs des sites donneurs et accepteurs sauvages et mutants sont indiquées ainsi que la variation de leur valeur (CV). NB. Des mutations activant des sites cryptiques sont également correctement prédites par ce type d'outils [13,19]. C. Représentation de l'évolution de la capacité vitale des patients présentant une myopathie de Duchenne.

3. Enjeux

3.1. Nomenclature des mutations

Plusieurs systèmes d'annotation des diverses séquences génomiques se sont développés de façon disparate, aboutissant à une impasse sémantique dans la nomenclature des mutations. Sachant que ce problème peut être lourd de conséquences (la dénomination erronée d'une mutation peut conduire à un diagnostic prénatal erroné), il fut décidé, après des années de débats passionnés, d'adopter une nomenclature « universelle » [23] permettant de décrire de façon unique et non équivoque toute variation de séquence pathogène ou non. Les règles [24] sont régulièrement mises à jour sur le site HGVS (http://www.HGVS.org/mutnomen/). Les séquences de référence génomiques et codantes sont répertoriées dans la banque de données : http://www.ncbi.nlm.nih.gov/RefSeq/. Les variations de séquence étant le plus souvent décrites par rapport à la séquence codante du gène, les nucléotides sont numérotés à partir du codon d'initiation de la traduction ATG (le A est numéroté +1).

Ces règles n'étant pas encore systématiquement appliquées, il existe dans les LSDBs et dans la littérature des variations de séquence dont le nom est erroné. De plus, pour les gènes décrits avant cette nomenclature, on continue le plus souvent à utiliser une terminologie « historique », comme c'est le cas par exemple avec le gène CFTR (http://www.genet.sickkids.on.ca/cftr/app). Ainsi, « 3659delC » correspond à deux mutations différentes dans le même exon selon la nomenclature utilisée (Tableau 2) : la coexistence de deux nomenclatures peut conduire à des erreurs de diagnostic. Il est donc impératif afin d'éviter toute confusion de nommer correctement les mutations en suivant les recommandations internationales et en mentionnant la séquence de référence utilisée et sa version. La nomenclature évoluant au fur et à mesure de la découverte de nouveaux types de mutations ou de nouvelles séquences dans les gènes, le rôle des curateurs de banques de données gène-spécifique est essentiel pour inciter tous les laboratoires à utiliser la même nomenclature.

L'outil Mutalyzer (http://www.humgen.nl/mutalyzer/1.0.1/) permet de générer et/ou de vérifier la nomenclature des variations de séquence [25]. Certaines LSDBs ont incorporé un système informatique de vérification automatique de la nomenclature au moment de la saisie de chaque mutation ou variation : c'est le cas des LSDBs développées avec les logiciels UMD [16,26] ou LOVD [27].

Tableau 2
Co-existence de deux nomenclatures : exemple du gène *CFTR*.

	Numérotation usuelle	Numérotation internationale HGVS
Nucléotides	+132 (initiation de la transcription)	+1 (initiation de la traduction)
Mutation 1	3659delC[a]	c.3528delC (RefSeq NM_000492.3)
Mutation 2	3791delC	c.3659delC[a] (RefSeq NM_000492.3)

[a] La nomenclature usuelle est basée sur une numérotation des exons de 1 à 24 alors que le gène en comporte 27 et sur une numérotation des nucléotides commençant au site d'initiation de la transcription et non pas au site d'initiation de la traduction. Le nom de la variation de séquence est identique dans les deux nomenclatures alors qu'il s'agit de deux mutations différentes.

3.2. Classification des variants d'effet inconnu

L'analyse des gènes révèle diverses catégories de variations de la séquence nucléotidique (par rapport à la séquence de référence) dans les exons et dans les introns selon leur degré d'exploration

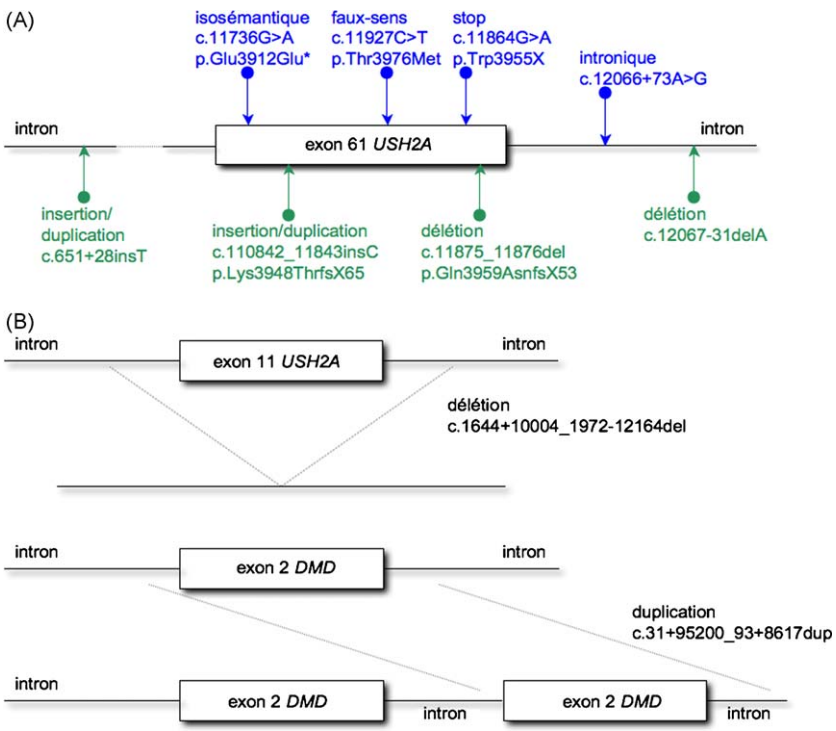

Fig. 2. Différents types de mutations décrites au niveau nucléotidique et conséquences prédites au niveau protéique. A. Substitutions et délétions/insertions d'un petit nombre de nucléotides (au-dessus : substitutions, en dessous : microdélétions, insertions/duplications, pouvant entraîner, lorsqu'elles sont exoniques, un décalage de la phase de lecture de l'ARN messager).
*La nomenclature protéique du variant c.11736G > A devrait être p. =, cependant, par souci de compréhension, l'ancienne notation a été conservée ici). B. Grands réarrangements. Délétions/insertions de régions génomiques incluant un ou plusieurs exons ou gènes. RefSeq *USH2A* NM_206933.3 – *DMD* NM_004006.2.

(Fig. 2). Pour de nombreux variants (faux-sens, iso-sémantiques ou introniques), il est impossible d'affirmer d'emblée un lien causal avec le phénotype. Ces *unclassified variants* (UV) posent des problèmes récurrents d'interprétation, d'autant que les tests fonctionnels qui permettraient de tester leur effet in vitro ne sont pas disponibles ou ne sont pas réalisables.

Les SNP non synonymes ou variants faux-sens, qui induisent une substitution d'acide aminé dans la protéine, sont particulièrement problématiques en raison de leur fréquence (estimée entre 100 000 et 200 000). De nombreux algorithmes utilisant des méthodes génétiques, biochimiques et informatiques ont été développés pour tenter de prédire quels variants sont pathogènes et lesquels ne le sont pas [28]. Il a été montré récemment que la valeur prédictive peut atteindre 88 % si on combine les diverses ressources étudiant la conservation et la structure (SIFT, PolyPhen, Align-GVGD, BLOSUM62) [29] et plus de 95 % si on rajoute d'autres propriétés biochimiques et l'impact sur l'épissage [30].

Les informations disponibles dans les banques de données centrales (UniProtKB/Swissprot [31,32] ou dbSNP [33]) doivent être examinées avec la plus grande attention. L'appellation SNP ne signifie pas « sans effet clinique ». Par exemple, la référence rs34691738 correspond à un variant du gène *GPR98* impliqué dans le syndrome de Usher type IIC (MIM #605472) ; sa nomenclature exacte est NM_032119.3:c.1278_1279insG, prédisant l'apparition d'un codon stop prématuré (p.Asn428GlnfsX14) avec une probabilité élevée de protéine tronquée non fonctionnelle. Certaines banques telles SM2PHDB, MutDB [34] ou SAAPDB [35] s'efforcent de centraliser les annotations issues de diverses sources, tout en permettant des analyses prédictives à partir d'alignements ou de structures tridimensionnelles concernant des variants connus ou non. Les outils de prédiction doivent être utilisés avec prudence, la fiabilité des informations dépendant fortement non seulement des algorithmes utilisés pour la création des alignements et des structures, mais aussi de la qualité des annotations de départ.

Les LSDBs peuvent être utilisées pour répertorier la fréquence allélique des UV chez les malades, chez les apparentés asymptomatiques, dans la population générale, ou pour vérifier la co-occurrence éventuelle d'une mutation plus délétère sur le même allèle. . ., critères plus fiables que les prédictions.

3.3. Corrélations génotype–phénotype, gène–fonction

Les banques de données peuvent jouer un rôle majeur dans l'étude de ces corrélations, pour la compréhension des mécanismes physiopathologiques, l'identification de marqueurs pronostiques, voire pour la sélection de patients éligibles pour certaines thérapies.

3.3.1. Hétérogénéité génétique et phénotype

De nombreuses maladies peuvent être le résultat de mutations dans des gènes différents, tels les gènes *APOB*, *LDLR* ou *PCSK9* dans l'hypercholestérolémie, les gènes *FBN1, TGFBR1* ou *TGFBR2* dans le syndrome de Marfan, ou encore la cinquantaine de gènes identifiés dans les surdités non syndromiques [36]. L'anémie de Fanconi, avec 12 groupes de complémentation, représente un bon modèle d'hétérogénéité génétique dans lequel les corrélations génotype–phénotype viennent de suggérer une hypothèse d'intérêt majeur concernant l'effet des variants FA sur le vieillissement accéléré [37].

3.3.2. Position des mutations dans le gène

Grâce à la banque *FBN1* (étude d'un millier de probands), une relation a pu être établie entre les mutations situées dans la région comprenant les exons 24 à 32 et les formes sévères du syndrome de Marfan (formes néonatales ou formes associant plusieurs critères de sévérité) [38].

Autre exemple : le gène *DMD* (79 exons répartis parmi 2,5 millions de nucléotides) génère une série d'isoformes de dystrophine produits par des promoteurs alternatifs et des promoteurs internes. Le produit le plus abondant dans le cerveau est la Dp71 (71 kDa), dont le promoteur et l'exon 1 spécifiques sont situés entre les exons 62 et 63. Il vient d'être démontré que les patients avec une myopathie de Becker et une déficience mentale ont des mutations altérant l'expression de la Dp71. Chez les patients avec une forme Duchenne, les mutations du gène *DMD* situées avant l'exon 62 (qui induisent la perte complète de toutes les isoformes sauf la Dp71) sont le plus souvent associées à des compétences cognitives normales ou limites. Ces données confirment la relation entre absence de Dp71 et sévérité de la déficience mentale dans les dystrophinopathies [39].

3.3.3. Type de mutations

CDH23 est un gène impliqué dans le syndrome de Usher de type I ou dans les surdités non syndromiques d'origine génétique. Les mutations faux-sens peuvent être responsables de l'une ou l'autre des deux pathologies, contrairement aux mutations aboutissant à la formation d'un codon stop. De manière intéressante, les mutations faux-sens, lorsqu'elles entraînent une modification de la séquence protéique de l'un des deux motifs de liaison au calcium (LDRE ou DXNDN) d'un domaine cadhérine, sont associées à une surdité non syndromique (DFNB12). Une mutation faux-sens impactant un autre domaine protéique sera associée à un USHI ou une surdité DFNB12 [40].

3.4. Les thérapies allèle-spécifiques

Une compréhension de plus en plus fine des anomalies moléculaires en cause dans les maladies génétiques a permis de développer des approches visant à réparer les parties défectueuses du gène lui-même (« chirurgie du gène ») plutôt que de transférer un gène normal dans la cellule (« thérapie génique » classique). L'objectif est d'intervenir directement au niveau d'une des étapes de la synthèse d'une protéine afin de corriger in situ le défaut génétique spécifique de chaque malade de façon à produire un peu de protéine fonctionnelle même partiellement. Il est par exemple possible d'intervenir sur les transcrits pour les rendre fonctionnels.

3.4.1. Le saut d'exon thérapeutique (épissothérapie « soustractive »)

Il est possible d'empêcher sélectivement l'intégration d'un ou plusieurs exons dans l'ARNm (saut d'exon) en masquant spécifiquement des séquences essentielles au déroulement normal de l'épissage, grâce à l'utilisation de petites molécules d'ARN antisens (« oligonucléotides antisens »). Cette approche a été particulièrement développée pour la myopathie de Duchenne (DMD) : une manipulation de l'épissage visant à forcer le saut d'un ou plusieurs exons aux bornes de la région délétée pourrait permettre la restauration d'une dystrophine tronquée partiellement fonctionnelle [41]. Un essai clinique de phase I a déjà eu lieu en 2007 [17] et un essai de phase I/II est en cours dont les résultats sont attendus en 2009. Cette approche pourrait aussi s'appliquer dans le cas d'une mutation non-sens : l'élimination de l'exon qui la porte dans moins de 20 % des ARNm peut suffire pour produire une quantité de dystrophine compatible avec un phénotype plus modéré [18].

Les banques spécialisées de type LSDB, en recensant à la fois les mutations identifiées chez les patients et les caractéristiques cliniques, peuvent permettre d'orienter le choix du meilleur saut d'exon(s) à réaliser afin de traiter le plus grand nombre de patients. En effet, les délétions du gène *DMD* étant très hétérogènes, plusieurs modèles de sauts d'exon seront nécessaires. Il est par ailleurs primordial de vérifier que la délétion nouvellement créée est effectivement associée à un phénotype modéré de type myopathie de Becker (BMD) lorsqu'elle survient naturellement chez un patient. Les outils d'analyse présents dans certaines LSDB comme la banque française UMD-DMD France peuvent contribuer à la définition des meilleurs modèles de sauts d'exon grâce à l'exhaustivité et la qualité des données collectées [42].

Il est également envisageable d'utiliser l'épissothérapie pour restaurer la production d'une protéine normale, dans le cas par exemple où une mutation intronique active un site cryptique d'épissage et génère l'intégration d'un nouvel exon dans l'ARNm. Le masquage de ce signal anormal par des oligonucléotides antisens pourrait restaurer la production d'un transcrit normal. Différentes études en ont montré la faisabilité dans la mucoviscidose [19], l'albinisme oculaire de type I [43], l'afibrinigénémie [44] ou encore la forme congénitale d'une maladie métabolique associée à des défauts de glycosylation de type Ia [45]. Cette approche pourrait aussi s'appliquer aux dystrophinopathies dans les cas où ce type de mutations a été décrit [46,47].

3.4.2. L'inclusion d'exon (épissothérapie « additive »)

Les petits ARN antisens peuvent être utilisés pour forcer le maintien dans le transcrit mature d'un exon anormalement éliminé en raison de la présence d'une mutation ou d'un polymorphisme qui altère l'épissage normal, comme c'est le cas pour les amyotrophies spinales infantiles. La faisabilité de cette approche a été démontrée dans divers modèles in vitro et in vivo [48,49]. Les banques de données sont utiles pour recenser les mutations du gène *SMN1* et le nombre de copies du gène *SMN2* chez les patients potentiellement éligibles pour des essais cliniques.

3.4.3. La translecture des codons stop

Des cribles pharmacologiques ont permis d'identifier des molécules chimiques, notamment la molécule PTC124, ayant la propriété de permettre au ribosome de franchir le signal d'arrêt prématuré et de continuer la traduction de l'ARNm jusqu'au codon stop naturel, produisant ainsi une protéine complète [50,51]. L'efficacité de cette translecture forcée « corrigeant » l'anomalie responsable de la maladie peut être variable selon la nature du codon stop (TAA, TAG ou TGA) et son environnement nucléotidique. Les LSDBs sont essentielles pour évaluer l'intérêt d'une telle approche en établissant avec précision la fréquence des mutations non-sens, la répartition des trois types de codon stop concernés et les régions du gène dans lesquelles le remplacement d'un codon stop prématuré par un acide aminé (introduction d'une mutation faux-sens) pourrait a priori être le mieux toléré par la protéine [14]. Parmi les maladies potentiellement concernées, figurent la myopathie de Duchenne pour laquelle un essai international de phase II/III est en cours [52,53], la mucoviscidose pour laquelle des essais cliniques ont déjà été réalisés chez des patients [54,55].

4. Perspectives

4.1. Annotation du génome humain et mutations

Les avancées technologiques qui ont permis de révéler la séquence complète des trois milliards de bases du génome humain [56] ont suscité de nombreux projets de séquençage d'organismes eucaryotes dont une liste exhaustive est accessible sur le site du National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/). L'annotation des génomes peut se faire à trois niveaux. L'annotation « syntaxique » a pour objet d'identifier les régions de séquence présentant une pertinence biologique

(gènes, signaux, répétitions...). L'annotation « fonctionnelle » prédit les fonctions et produits potentiels des gènes préalablement identifiés (similitudes de séquences, motifs, structures...) et utilise les informations expérimentales disponibles (littérature, jeux de données à grande échelle...). L'annotation « relationnelle » détermine les interactions que des cibles biologiques préalablement identifiées sont susceptibles d'entretenir (familles de gènes, réseaux de régulation, réseaux métaboliques...) [57].

Ces différents niveaux d'annotation sont accessibles grâce à des sites de référence tels NCBI, Ensembl (http://www.ensembl.org/index.html) ou UCSC (http://genome.ucsc.edu/). Bien que ces portails soient d'une importance considérable pour la recherche, ils demeurent très pauvres dans l'annotation des variations de séquence. Par exemple, si plus de 25 millions de « polymorphismes » étaient accessibles via dbSNP au 30 juin 2009, il est difficile de distinguer ceux qui sont pathogènes de ceux qui ne le sont pas. Différentes organisations (Human Genome Organization ; HGVS) ou consortiums internationaux (HVP [21] ; GEN2PHEN [http://www.gen2phen.org] [58]) ont décidé de promouvoir une collecte intégrée de l'ensemble des variations de séquence décrites chez l'homme et leur annotation précise.

Les banques de données et leur lien avec le génome de référence et les systèmes d'annotations permettront l'interprétation des énormes quantités de données qui seront générées par les nouvelles technologies de séquençage en phase solide (*next generation sequencing*) [59–61]. Il faudra pouvoir identifier, parmi un flot d'informations et de variations, les mutations délétères pour la santé.

Des registres de patients à l'échelle internationale permettant de relier les informations génétiques et cliniques pourront faciliter les études de faisabilité d'essais cliniques et éventuellement le recrutement des patients, à l'exemple du projet européen TREAT-NMD (http://www.treat-nmd.eu/home.php) [62].

4.2. Développement des descriptions phénotypiques : vers une ontologie clinique bio-informatique intégrant les données génomiques ?

Dans les banques de mutations, les données phénotypiques sont souvent présentées de façon très rudimentaire (voire limitées au seul nom de la maladie), en texte libre et donc difficilement analysable par des outils automatisés. L'exploitation informatique des données phénotypiques nécessite le développement d'ontologies phénotypiques approfondies et d'outils d'analyse spécifiques.

Les principaux systèmes de classification des maladies ont été développés par l'Organisation mondiale de la santé (OMS). La « classification internationale des maladies » (CIM) permet le codage des maladies. La « classification internationale du fonctionnement, du handicap et de la santé » propose une échelle d'évaluation de la sévérité du handicap sur le plan individuel. La « classification internationale des interventions de santé » est un outil permettant de rapporter et d'analyser sur le plan statistique la distribution et l'évolution des interventions de santé.

La CIM-10 (dixième version) est essentiellement utilisée pour l'analyse des dépenses de santé, mais son degré de précision est incompatible avec une description adaptée des données cliniques composant un dossier médical. D'autres « thesauri » de termes médicaux ont été développés. Medical Subjects Headings (MeSH) est un outil d'indexation, de catalogage et d'interrogation des banques de données notamment utilisé pour Medline/Pubmed. Unified Medical Language System (UMLS) a pour objectif de faciliter la recherche et l'intégration des informations biomédicales. Il comprend trois composants : le Métathesaurus, collection de concepts et de termes issus de différents vocabulaires ; le Réseau sémantique, ensemble de catégories et de relations utilisées pour classer et relier les entrées du Métathesaurus ; le Specialist Lexicon, qui comprend des informations lexicographiques pour le traitement du langage biomédical. La nomenclature systématique de médecine ou SNOMED CT (*systematized nomenclature of medicine – clinical terms*) est une nomenclature pluriaxiale qui attribue à chaque concept un code combinant le site anatomique, la cause, les effets physiopathologiques, les circonstances d'apparition et les actions diagnostiques ou thérapeutiques. Cette nomenclature, créée afin de classer les activités cliniques pour la gestion des systèmes d'information hospitaliers, est complexe et lourde à utiliser, le nombre de caractères nécessaires au codage pouvant être très important et les manuels de référence volumineux !

Aucun de ces systèmes n'a été conçu pour décrire avec précision les différents symptômes cliniques présentés par un patient, leur mode évolutif ou leur modification sous traitement. La plupart des maladies génétiques ne bénéficient pas encore d'un consensus international concernant les critères cliniques et génétiques les plus pertinents pour le diagnostic, à l'exemple de la mucoviscidose [63,64] ou du syndrome de Marfan avec les critères de Gand [65].

Si l'on veut que la connaissance biologique associée à la génomique puisse à l'avenir être bénéfique pour la santé, il est nécessaire de développer de nouvelles ressources bio-informatiques cliniques et biologiques structurées comme des ontologies, capables d'intégrer des informations standardisées complexes et évolutives ; elles devront être suffisamment flexibles pour être utilisables par diverses catégories de professionnels [66]. Le dossier médical électronique du futur [67] devra comporter les informations génétiques nécessaires pour répondre aux exigences d'une pratique médicale qui sera à la fois prédictive, préventive, participative et personnalisée (« P4 medicine »).

4.3. Principes éthiques et règles législatives

Les principes éthiques de la protection des personnes faisant l'objet de recherche énoncés par l'Association médicale mondiale (déclaration d'Helsinki) et par l'Unesco (déclaration universelle sur la bioéthique et les droits de l'homme) sont repris par les différentes lois françaises et européennes (Loi informatique et libertés ; Directive 95/46/EC du Parlement européen et du Conseil) (Tableau 3). Les droits d'auteur et la protection de la propriété intellectuelle doivent également être respectés [8]. La protection juridique des banques de données est énoncée par la loi française n° 98-536 et la Directive européenne n° 96/9/EC.

L'« anonymisation » des données est essentielle avant leur intégration dans la banque. Les informations contenues dans les banques de données génétiques sont considérées comme des « données à caractère personnel » et des « données sensibles ». Une donnée à caractère personnel est une information relative à une personne physique qui peut être identifiée directement ou indirectement par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres, tels qu'un prélèvement d'ADN. Toute donnée de santé, considérée comme une donnée sensible, ne peut être recueillie et exploitée qu'avec le consentement explicite du patient. Avant le recueil de ce consentement, le patient doit être informé de l'identité de la personne responsable de la banque de données, des objectifs de cette banque, de la nature des données recueillies et des utilisateurs des données. Un délai de réflexion suffisant doit être respecté entre l'information orale et le recueil du « consentement écrit », afin de permettre au patient de poser toutes les questions qu'il juge nécessaire.

En France, la déclaration d'une banque de données moléculaires et/ou cliniques ayant un objectif de recherche comporte deux étapes. Le Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé (CCTIRS) émet un

Tableau 3
Sites web des principales terminologies cliniques – Accès aux principaux textes éthiques et législatifs.

	Sites web
Ontologies cliniques	
Medical Subjects Headings (MeSH)	http://www.nlm.nih.gov/mesh/
Organisation mondiale de la santé (OMS)	http://www.who.int/classifications/en/
Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)	http://www.ihtsdo.org/snomed-ct/
Unified Medical Language System (UMLS)	http://www.nlm.nih.gov/research/umls/
Éthique et législation	
Code de déontologie médicale	http://www.conseil-national.medecin.fr/?url=rubrique.php&menu=DEOINTEGRAL
Commission nationale de l'informatique et des libertés (CNIL)	http://www.cnil.fr/
Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la Santé (CCTIRS)	http://www.enseignementsup-recherche.gouv.fr/cid20537/cctirs.html
Déclaration d'Helsinki 2008	http://www.wma.net/f/policy/b3.htm
Déclaration universelle sur la bioéthique et les droits de l'homme – Unesco 2005	http://portal.unesco.org/fr/ev.php-URL_ID=31058&URL_DO=DO_TOPIC&URL_SECTION=201.html
Directive du Parlement Européen et du Conseil 95/46/EC	http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:fr:HTML
Loi informatique, fichiers et libertés	http://legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20090709

avis sur la méthodologie de la recherche au regard de la loi, la nécessité du recours à des données à caractère personnel et la pertinence de celles-ci par rapport à l'objectif de la recherche. Après avis favorable, une demande d'autorisation doit être déposée auprès de la Commission nationale de l'informatique et des libertés (CNIL), qui vérifie les garanties présentées par le demandeur, la conformité de sa demande à ses missions ou à son objet social, détermine une durée de conservation des données, et apprécie les dispositions prises pour assurer la sécurité des données et la garantie des secrets protégés par la loi.

Sur le plan international, les grands principes éthiques sont le plus souvent repris dans les différents systèmes législatifs. Les utilisateurs des banques de données doivent connaître ces principes et s'assurer que les fournisseurs de données les respectent. Des recommandations destinées aux curateurs des banques locus spécifiques sont disponibles [68].

5. Conclusion

Les « tests génétiques » ne sont plus seulement la spécialité des laboratoires impliqués dans le diagnostic de maladies rares, mais vont prochainement avoir de nombreuses applications en médecine, depuis la pharmacogénétique jusqu'aux tests prédictifs de susceptibilité aux maladies fréquentes. La complexité que nous connaissons déjà pour les tests « monogéniques » sera sans commune mesure avec celle des tests « polygéniques » (dont certains sont déjà commercialisés via Internet) qui analyseront des dizaines, voire des centaines de gènes simultanément. On estime à plus de 65 % la proportion d'êtres humains qui sont affectés par une ou plusieurs mutations géniques au cours de leur vie. Les banques de mutations capables d'analyser la variabilité génétique sont donc un véritable enjeu de santé publique. Elles sont aussi indispensables pour la recherche, comme en témoignent les tentatives internationales pour les intégrer dans les banques centrales. Il est temps de considérer que les banques de mutations font partie des systèmes de santé et doivent être, à ce titre, financées de façon pérenne comme les centres cliniques et laboratoires de référence des maladies rares.

Remerciements

Le développement et la curation des banques de données LSDBs réalisés à Montpellier sont en grande partie dus au soutien inestimable des associations de patients, notamment l'Association française contre les myopathies (AFM), Vaincre la Mucoviscidose (VLM) et SOS-RP (rétinite pigmentaire). Les auteurs souhaitent leur témoigner leur reconnaissance.

Références

[1] Cotton RGH, Horaitis O. The HUGO-Mutation database initiative. Pharmacogenomics J 2002;2(1):16–9.
[2] Auerbach AD. 8th International HUGO-Mutation database initiative meeting, April 9, 2000, Vancouver, Canada. Hum Mutat 2000;16(3):265–8.
[3] Nemos C, Bursztejn AC, Jonveaux P. Gestion des variations du nombre de séquences génomiques (CNV) en génétique humaine constitutionnelle utilisant l'hybridation génomique comparative en microréseau d'ADN (HGCM). Pathol Biol 2008;56:354–61.
[4] Patrinos GP. National and ethnic mutation databases: recording populations' genography. Hum Mutat 2006;27(9):879–87.
[5] Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's online Mendelian inheritance in man (OMIM (R)). Nucleic Acids Res 2009;37:D793–6.
[6] Krawczak M, Cooper DN. The human gene mutation database. Trends Genet 1997;13(3):121–2.
[7] Stenson PD, Mort M, Ball EV, Howels K, Phillips AD, Thomas NS, et al. The Human gene mutation database: 2008 update. Genome Med 2009;1(1):13 [Epub ahead of print].
[8] Patrinos GP, Brookes AJ. DNA, diseases and databases: disastrously deficient. Trends Genet 2005;21(6):333–8.
[9] Zlotogora J, van Baal S, Patrinos GP. Documentation of inherited disorders and mutation frequencies in the different religious communities in Israel in the Israeli National Genetic Database. Hum Mutat 2007;28:944–9.
[10] Tan EC, Loh M, Chuon D, Lim YP. Singapore human mutation/polymorphism database: a country-specific database for mutations and polymorphismsin inherited disorders and candidate gene association studies. Hum Mutat 2006;27(3):232–5.
[11] van Baal S, Kaimakis P, Phommarinh M, Koumbi D, Cuppens H, Riccardino F, et al. FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. Nucleic Acids Res 2007; 35:D690–5.
[12] Cotton RGH, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AE, et al. Recommendations for locus-specific databases and their curation. Hum Mutat 2008;29(1):2–5.
[13] Gout AM, Ravine D, Consortium AGV. Analysis of published PKD1 gene sequence variants. Nat Genet 2007;39(4):427–8.
[14] Tuffery-Giraud S, Beroud C, Leturcq F, Yaou R, Hamroun D, Michel-Calemard L, et al. Genotype-phenotype analysis in 2,405 patients with a dystrophinopathy using the UMD-DMD database: a model of nationwide knowledgebase. Hum Mutat 2009;30(6):934–45.
[15] Flanigan KM. Mutation-specific database and bioinformatics resource for DMD. Hum Mutat 2009;30(6) [comment on Hum Mutat 2009;30(6):934–45].
[16] Beroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. UMD (Universal Mutation Database): 2005 update. Hum Mutat 2005;26(3):184–91.
[17] van Deutekom JC, Janson AA, Ginjaar IB, Frankhuizen WS, Aartsma-Rus A, Bremmer-Bout M, et al. Local dystrophin restoration with antisense oligonucleotide PRO051. New Engl J Med 2007;357(26):2677–86.
[18] Disset A, Bourgeois CF, Benmalek N, Claustres M, Stevenin J, Tuffery-Giraud S. An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. Hum Mol Genet 2006;15(6):999–1013.
[19] Friedman KJ, Kole J, Cohn JA, Knowles MR, Silverman LM, Kole R. Correction of aberrant splicing of the cystic fibrosis transmembrane conductance regulator (CFTR) gene by antisense oligonucleotides. J Biol Chem 1999;274:36193–9.
[20] Claustres M, Horaitis O, Vanevski M, Cotton RGH. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 2002;12(5):680–8.
[21] Cotton GH, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, et al. The Human Variome Project. Science 2008;322:861–2.
[22] Axton M. Human variome microattribution reviews. Nat genet 2008;40(1):1.

[23] Antonarakis SE. Nomenclature Working Group. Recommendations for a nomenclature system for human gene mutations. Hum Mutat 1996;11:1–3.
[24] den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 2000;15:7–12.
[25] Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat 2008;29(1):6–13.
[26] Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 2000;15(1):86–94.
[27] Fokkema IF, den Dunnen JT, Taschner PE. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 2005;26(2):63–8.
[28] Thusberg J, Vihinen M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 2009;30(5):703–14.
[29] Chan PA, Duraisamy S, Miller PJ, Newel JA, McBride C, Bond JP, et al. Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and TYR. Hum Mutat 2007;28(7):683–93.
[30] Frederic MY, Lalande M, Boileau C, Hamroun D, Claustres M, Beroud C, et al. UMD-predictor, a new prediction tool for nucleotide substitution pathogenicity; application to four genes: FBN1, FBN2, TGFBR1 and TGFBR2. Hum Mutat 2009;30(6):952–9.
[31] The UniProt Consortium. The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 2009;37(Database issue):D169–74.
[32] Bairoch A, Boeckmann B, Ferro S, Gasteiger E. Swiss-Prot: juggling between evolution and stability. Brief Bioinform 2004;5(1):39–55.
[33] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29(1):308–11.
[34] Dantzer J, Moad C, Heiland R, Mooney S. MutDB services: interactive structural analysis of mutation data. Nucleic Acids Res 2005;33(Web Server issue): W311–4.
[35] Hurst JM, McMillan LE, Porter CT, Allen J, Fakorede A, Martin AC. The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat 2009;30(4):616–24.
[36] Hilgert N, Smith RJ, Van Camp G. Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics? Mutat Res 2009;681(2–3):189–96.
[37] Neveling K, Endt D, Hoehn H, Schindler D. Genotype-phenotype correlations in Fanconi anemia. Mutat Res 2009;668(1–2):73–91.
[38] Faivre L, Collod-Beroud G, Loeys BL, Child A, Binquet C, Gautier E, et al. Effect of mutation type and location on clinical outcome in 1,013 probands with Marfan syndrome or related phenotypes and FBN1 mutations: an international study. Am J Hum Genet 2007;81(3):454–66.
[39] Daoud F, Angeard N, Demerre B, Martie I, Benyaou R, Leturcq F, et al. Analysis of Dp71 contribution in the severity of mental retardation through comparison of Duchenne and Becker patients differing by mutation consequences on Dp71 expression. Hum Mol Genet 2009;18(20):3779–94.
[40] Astuto LM, Bork JM, Weston MD, Askew JW, Fields RR, Orten DJ, et al. CDH23 mutation and phenotype heterogeneity: a profile of 107 diverse families with Usher syndrome and nonsyndromic deafness. Am J Hum Genet 2002;71(2): 262–75.
[41] Aartsma-Rus A, Kaman WE, Weij R, den Dunnen JT, van Ommen GJB, van Deutekom JCT. Exploring the frontiers of therapeutic exon skipping for Duchenne muscular dystrophy by double targeting within one or multiple dxons. Mol Ther 2006;14(3):401–7.
[42] Beroud C, Tuffery-Giraud S, Matsuo M, Hamroun D, Humbertclaude V, Monnier N, et al. Multiexon skipping leading to an artificial DMD protein lacking amino acids from exons 45 through 55 could rescue up to 63% of patients with Duchenne muscular dystrophy. Hum Mutat 2007;28(2):196–202.
[43] Vetrini F, Tammaro R, Bondanza S, Surace EM, Auricchio A, de Luca M, et al. Aberrant splicing in the ocular albinism type1 gene (OA1/GPR143) is corrected in vitro by morpholino antisense oligonucleotides. Hum Mutat 2006;27:420–6.
[44] Davis RL, Homer VM, George PM, Brennan SO. A deep intronic mutation in FGB creates a consensus exonic splicing enhancer motif that results in afibrinogenemia caused by aberrant mRNA splicing, which can be corrected in vitro with antisense oligonucleotide treatment. Hum Mutat 2009;30:221–7.
[45] Vega AI, Perez-Cerda C, Desviat LR, Matthijs G, Ugarte M, Perez B. Functional analysis of three splicing mutations identified in the PMM2 gene: toward a new therapy for congenital disorder of glycosylation type Ia. Hum Mutat 2009;30:795–803.
[46] Tuffery-Giraud S, Chambert S, Demaille J, Claustres M. Point mutations in the dystrophin gene: evidence for frequent use of cryptic splice sites as a result of splicing defects. Hum Mutat 1999;14:359–68.
[47] Beroud C, Carrie A, Beldjord C, Deburgrave N, Llense S, Carelle N, et al. Dystrophinopathy caused by mid-intronic substitutions activating cryptic exons in the DMD gene. Neuromuscul Disord 2004;14:10–8.
[48] Meyer K, Marquis J, Trub J, Nlend RN, Verp S, Ruepp MD, et al. Rescue of a severe mouse model for spinal muscular atrophy by U7 snRNA-mediated splicing modulation. Hum Mol Genet 2009;18(3):546–55.
[49] Khoo B, Krainer AR. Splicing therapeutics in SMN2 and APOB. Curr Opin Mol Ther 2009;11(2):108–15.
[50] Linde L, Kerem B. Introducing sense into nonsense in treatments of human genetic diseases. Trends Genet 2008;24(11):552–63.
[51] Welch EM, Barton ER, Zhuo J, Tomizawa Y, Friesen WJ, Trifillis P, et al. PTC124 targets genetic disorders caused by nonsense mutations. Nature 2007; 447(7140):87–96.
[52] Wilton S. PTC124, nonsense mutations and Duchenne muscular dystrophy. Neuromuscul Disord 2007;17(9–10):719–20.
[53] Davies S, Serradell N, Rosa E, Castaner R, Ataluren. Nonsense mutation suppressor, treatment of cystic fibrosis. Treatment of muscular dystrophy. Drugs Fut 2008;33(9):733–6.
[54] Hyde SC, Gill DR. Ignoring the nonsense: a phase II trial in cystic fibrosis. Lancet 2008;372(9640):691–2.
[55] Kerem E, Hirawat S, Armoni S, Yaakov Y, Shoseyov D, Cohen M, et al. Effectiveness of PTC124 treatment of cystic fibrosis caused by nonsense mutations: a prospective phase II trial. Lancet 2008;372(9640):719–27.
[56] Abramowicz M. The Human Genome Project in retrospect. Adv Genet 2003;50:231–61 [discussion 507–10].
[57] Stein L. Genome annotation: from sequence to biology. Nat Rev Genet 2001;2:493–503.
[58] Thorisson GA, Muilu J, Brookes AJ. Genotype-phenotype databases: challenges and solutions for the post-genomic era. Nat Rev Genet 2009;10(1):9–18.
[59] Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The diploid genome sequence of an individual human. PLoS Biol 2007;5:e254.
[60] Wang J, Wang W, Li R, Li Y, Tian G, Goodman. et al. The diploid genome sequence of an Asian individual. Nature 2008;456:60–5.
[61] Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 2008;452:872–6.
[62] Mercuri E, Mayhew A, Muntoni F, Messina S, Straub V, Van Ommen GJ, et al. TREAT-NMD Neuromuscular Network. Towards harmonisation of outcome measures for DMD and SMA within TREAT-NMD; report of three expert workshops: TREAT-NMD/ENMC workshop on outcome measures, 12th–13th May 2007, Naarden, The Netherlands; TREAT-NMD workshop on outcome measures in experimental trials for DMD, 30th June–1st July 2007, Naarden, The Netherlands; conjoint Institute of Myology TREAT-NMD meeting on physical activity monitoring in neuromuscular disorders, 11th July 2007, Paris, France. Neuromuscul Disord 2008;18(11):894–903.
[63] Rosenstein BJ, Cutting GR. The diagnosis of cystic fibrosis: a consensus statement. Cystic Fibrosis Foundation Consensus Panel. J Pediatr 1998;132(4): 589–95.
[64] Castellani C, Cuppens H, Macek Jr M, Cassiman JJ, Kerem E, Durie P, et al. Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice. J Cyst Fibros 2008;7(3):179–96.
[65] De Paepe A, Devereux RB, Dietz HC, Hennekam RC, Pyeritz RE. Revised diagnostic criteria for the Marfan syndrome. Am J Med Genet 1996;62(4): 417–26.
[66] Sam LT, Mendonça EA, Li J, Blake J, Friedman C, Lussier YA, et al. An integrated resource for the multiscale mining of clinical and biological data BMC. Bioinformatics 2009;10(Suppl. 2):S8.
[67] Deshmukh VG, Hoffman MA, Arnoldi C, Bray BE, Mitchelle JA. Efficiency of CYP2C9 genetic tests representation for automated pharmacogenetic decision support. Methods Inf Med 2009;48(3):282–90.
[68] Cotton RGH, Sallee C, Knoppers BM. Locus-specific databases: from ethical principles to practice. Hum Mutat 2005;26(5):489–93.