

p53 REVIEW ARTICLE

The UMD-p53 Database: New Mutations and Analysis Tools

Christophe Bérout^{1*} and Thierry Soussi²¹Laboratoire de Génétique Moléculaire, CHU de Montpellier, Institut Universitaire de Recherche Clinique, Montpellier Cedex, France;²Institut Curie and Université P. & M. Curie, Laboratoire de Génomotoxicologie des Tumeurs, Paris Cedex, France

For the p53 Special Issue

The tumor suppressor gene TP53 (p53) is the most extensively studied gene involved in human cancers. More than 1,400 publications have reported mutations of this gene in 150 cancer types for a total of 14,971 mutations. To exploit this huge bulk of data, specific analytic tools were highly warranted. We therefore developed a locus-specific database software called UMD-p53. This database compiles all somatic and germline mutations as well as polymorphisms of the TP53 gene which have been reported in the published literature since 1989, or unpublished data submitted to the database curators. The database is available at www.umd.necker.fr or at <http://p53.curie.fr/>. In this paper, we describe recent developments of the UMD-p53 database. These developments include new fields and routines. For example, the analysis of putative acceptor or donor splice sites is now automated and gives new insight for the causal role of “silent mutations.” Other routines have also been created such as the prescreening module, the UV module, and the cancer distribution module. These new improvements will help users not only for molecular epidemiology and pharmacogenetic studies but also for patient-based studies. To achieve these purposes we have designed a procedure to check and validate data in order to reach the highest quality data. *Hum Mutat* 21:176–181, 2003. © 2003 Wiley-Liss, Inc.

KEY WORDS: TP53; p53; cancer; tumor; database; mutation analysis; Locus Specific Database; LSDB; UMD

DATABASES:

TP53 – OMIM: 191170; GenBank: NM_000546 (mRNA)

<http://p53.curie.fr/> (p53 Web Site at Institut Curie)

INTRODUCTION

Historically, collection of mutations and variations in human genes has taken place in the published literature. In the mid-eighties, a few of these variations were available in various databases such as the Genome Database (GDB) [Pearson, 1991; Cuticchia, 2000], GenBank [Bilofsky et al., 1986], EMBL [Hamm and Cameron, 1986], and SWISS-PROT [Bairoch and Boeckmann, 1991]. Nevertheless, because of the structure of these databases, the extraction of relevant information about mutations was almost impossible. In this context, it became clear that specific software had to be developed.

A few teams then started to build local databases to collect and document mutations in human genes. In the early nineties, we decided to develop not a simple repository of locus-specific mutations but a dynamic database including various computerized tools for their analysis. This project ultimately led to the Universal Mutation Database software (UMD) [Bérout et al., 2000], which is today recognized by the Human Genome Organization (HUGO) and the

Human Genome Variation Society (HGVS) as a reference tool to build the Locus Specific Database (LSDB). It was first used to create the UMD-p53 database in 1994 [Cariello et al., 1994] and subsequently for various genes involved in cancers (APC [Bérout and Soussi, 1996]; BRCA1, BRCA2, MEN1 [Wautot et al., 2002]; SUR1, RB, VHL [Bérout et al., 1998]; and WT1 [Jeanpierre et al., 1998]) and genetic diseases (FBN1 [Collod-Bérout et al., 1997], LDLR [Varret et al., 1998], VLCAD, MCAD, ATP7B, DMD, LMNA, STA).

The UMD-p53 database includes somatic and germline mutations of the TP53 gene (MIM# 191170) collected from the literature or directly

*Correspondence to: Christophe Bérout, Laboratoire de Génétique Moléculaire, CHU de Montpellier, Institut Universitaire de Recherche Clinique, 641 avenue du Doyen Gaston Giraud, 34093 Montpellier Cedex 5, France.

E-mail: Christophe.Beroud@igh.cnrs.fr

DOI 10.1002/humu.10187

Published online in Wiley InterScience (www.interscience.wiley.com).

submitted to the curator of the database. This tumor suppressor gene codes for a phosphoprotein expressed at a very low level in the nucleus of normal cells that functions as a tetrameric transcription factor with multiple, anti-proliferative functions activated in response to several forms of cellular stress. Somatic mutations of this gene are the most common genetic alteration found in human cancers. Germline mutations have been reported in Li-Fraumeni syndrome (MIM# 151623), which is a hereditary cancer predisposition syndrome. Other germline mutations have been found in patients who developed various tumors: adrenocortical carcinoma, astrocytoma, B-acute lymphoblastic leukemia, glioblastomas, colorectal carcinomas, osteosarcomas, and rhabdomyosarcomas.

The UMD-p53 database is a relational database developed with the 4th Dimension language (4D[®]). Since its first release in 1994, which included 3,000 mutations, a growing rate of about 1,500 mutations per year has been observed (totals: 4,200 in 1996 [Beroud et al., 1996], 7,500 in 1998 [Beroud and Soussi, 1998]). Today the database includes 14,770 somatic mutations and 201 germline mutations from 1,433 publications and unpublished series. All data have been extensively curated to remove all duplicates of the same mutation from the same sample published in various papers and to homogenize data. The current paper describes the new structure and routines of the UMD-p53 database.

DATABASE STRUCTURE

The database includes information from the literature and unpublished data directly submitted to the database curator. Only mutations obtained by sequencing are inserted in the database. According to the HGVS recommendations (www.genomic.unimelb.edu.au/mdi/rec.html), we collected information on the prescreening method used (e.g., SSCP, DGGE, YEAST ASSAY, IHC). Furthermore, to identify screening bias recently identified as the major source of results heterogeneity from clinical studies [Soussi and Beroud, 2001], we also added a description of which exons have been tested. In addition, to be able to characterize the incidence of each mutation in a specific cancer type, we created a new table called "Population Information" linked to the "Reference" table as shown in Figure 1. Overall, for each publication data are collected as follows: 1) the number of tested samples classified according to their cancer type, 2) the number of p53 mutated cases, and 3) the prescreening method used.

To secure the data input, the UMD-p53 database includes the human genetic code and the reference sequence of the TP53 gene (UniGene Cluster Hs.1846). In addition, for each amino acid, data

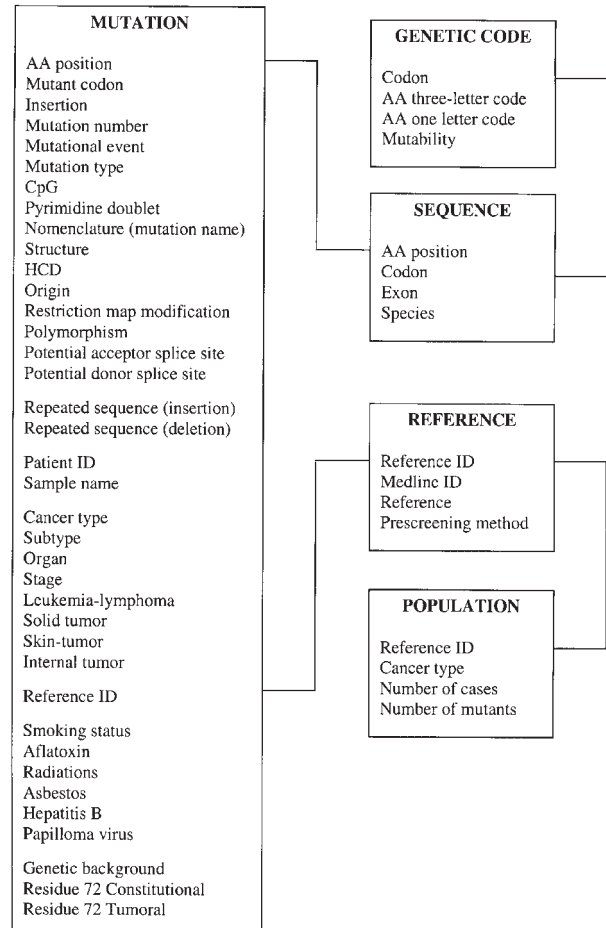


FIGURE 1. UMD-p53 database structure.

about its variation or conservation among species is described.

Finally, the main table reports data of the p53 mutations. Only variations of the coding sequence or flanking intronic regions corresponding to splice sites are included in the database. These variations are mostly mutations but polymorphisms can also be included. Because of the recent inclusion of the "Medline ID" and "ethnic origin" fields in the database structure, the corresponding data are scheduled to be included in the next release of the UMD-p53 database.

DESCRIPTION OF MUTATIONS

The UMD-p53 database is a relational database with a main table corresponding to the mutation. For each mutation, either somatic or germline, an automatic process computes information and displays the mutation name according to the international nomenclature system [Antonarakis, 1998; den Dunnen and Antonarakis, 2000, 2001] as well as the mutational event and the structure and/or highly

conserved domain of the protein where the mutation is located. For specific purposes, such as the study of mutations resulting from an UV exposure, the software also searches for pyrimidine doublets. Finally, usual information about the mutational event (e.g., transition, transversion) and possible CpG involvement are also displayed. A detailed analysis of TP53 mutations at CpG dinucleotides utilizing the UMD-p53 database is reported in Soussi and Beroud [2003].

To facilitate the mutation confirmation by a PCR digest method, modifications induced by the mutation in the restriction map are also available. Because polymorphisms are included in the database, and because some pathogenic mutations may appear as polymorphisms (silent mutations), we recently added a new routine that checks automatically for the creation of a putative acceptor or donor splice site according to the consensus sequences described by Senapathy et al. [1990]. Among the 652 silent mutations collected in the actual release of the UMD-p53 database, we have detected seven variations resulting in the creation of a potential acceptor or donor splice sites (Table 1). The strength of these sites has been evaluated using the calculation method described by Senapathy et al. [1990]. The average score increased from 62.3 to 76.8 (+23%), supporting the hypothesis of a possible pathogenic involvement of these new splice sites. Among the other 14,118 mutations, 753 also result in a putative creation of a potential acceptor or donor splice sites (data not shown).

Besides molecular information, data from the sample are also available. They include the sample type, cancer type and sub-type, affected organ, and tumor stage. Data from the genetic background are collected in a "background" field associated with a "constitutional polymorphism 72" and a "tumoral polymorphism 72" field that stores information about the Pro/Arg p53 polymorphism described at codon

position 72. Homozygosity for arginine at residue 72 was associated with an increased risk for bladder cancer [Soulitzis et al., 2002] while the p53 Pro homozygotes have a higher risk of development of nasopharyngeal carcinoma [Tsai et al., 2002].

To facilitate the molecular epidemiology and pharmacogenetic studies, we have collected information about various exposure risks: smoking status, exposure to aflatoxin B1, exposure to radiation, exposure to asbestos, and development of various pathologies such as hepatitis B and papillomavirus.

OBJECTIVE: HIGH QUALITY DATA

The UMD-p53 database is a typical LSDB that differs from the various core databases such as the Human Gene Mutation Database (HGMD; www.hgmd.org), which comprises various types of mutations within the coding regions of human nuclear genes causing inherited disease but no somatic mutations. Furthermore, in a core database each mutation is entered only once to avoid confusion between recurrent and identical-by-descent lesions. In the UMD-p53 database we have collected all available mutations from the p53 gene.

In order to achieve the highest quality data, we have developed an efficient curation process performed by experts in the fields of molecular biology and oncogenetic. For each newly described mutation, we search if it has previously been reported using the sample name as criteria. If this step reports it as unpublished, we then query the database using two or three author names as search criteria. This step is necessary because some teams have reported the same set of mutations in up to four different papers using a different nomenclature to describe tumor samples. We complete this analysis by a comparison of the reported cancer types and mutations frequency among the various publications from the same authors.

TABLE 1. Putative Acceptor or Donor Splice Sites in TP53 Created by Silent Mutations

Mutation	AA position	Wt AA	Mu AA	Cancer	Splice site	Sequence	Score ^a	Reference
642T>C	214	His	His	Breast carcinoma	Acceptor	Wt: acttttcgacatagt Mu: acttttcgaca C agt	77.6 87.3	Kucera et al. [1999]
642T>C	214	His	His	Breast carcinoma	Acceptor	Wt: acttttcgacatagt Mu: acttttcgaca C agt	77.6 87.3	Lou et al. [1997]
834T>A	278	Pro	Pro	Colorectal carcinoma	Acceptor	Wt: ttgtgcctgtcctgg Mu: ttgtgcctgtcc A gg	76.9 92.8	Ramnani et al. [1999]
681T>A	227	Ser	Ser	Ovarian carcinoma	Acceptor	Wt: tgaggttggctc A ga Mu: tgaggttggctc A ga	54.7 70.7	Wen et al. [1999]
744G>C	248	Arg	Arg	Larynx SCC	Acceptor	Wt: ggcatgaaccggagg Mu: ggcatgaaccg C agg	61.9 74.3	Suzuki et al. [1994]
426T>A	142	Pro	Pro	Head and neck SCC	Acceptor	Wt: caagacctgcctgt Mu: caagacctgccc A gt	53.4 69.6	Mineta et al. [1998]
822T>G	274	Val	Val	Sarcoma	Donor	Wt: gcgtgttt Mu: gcgtgt G t	48.9 63.6	Nakanishi et al. [1998]

^aThe scoring system from Senapathy [1990] ranges from 0 to 100 (strongest splice site).

Wt, wild type sequence; Mu, mutant sequence. Nucleotide variations and the corresponding scores are given in boldface.

When a mutation has been through these various processes, we use the automatic routines described previously to name the mutation and to compute molecular information. If the report is correctly described at the molecular level, we check the clinical description to standardize it using description criteria such as: organ, tumor type, sub-type, and stage. Finally, we decide if this mutation has been correctly described and can be added to the database. If not, we contact the authors to get additional information. This curation process allowed us to achieve high quality data and to discard more duplicates than other p53 databases.

DATABASE SEARCH AND ANALYSIS TOOLS

The chosen structure for the UMD-p53 database authorizes the collection of multiple mutations from the same tumor sample or from various tumor samples from the same patient. “sample name” links mutations from the same sample while “patient ID” links mutations from the same patient. This allows various levels of analysis including not only molecular epidemiology studies but also patient-based analysis such as clinical or pharmacogenetic studies. Because most of these analyses require computation, specific routines have been developed. They can be performed on the overall database or just on a subset, which can be extracted using a multi-criteria searching tool. Results can usually be exported to a text file. The available routines of the UMD-p53 database are:

1. “Insertions and deletions analysis” are two functions used to search if flanking repeated sequences could be involved in the mutational process as described in Bérout et al. [2000].
2. The “detailed mutational events” routine is used to get detailed information about p53 mutations. It displays the number of p53 variants (1,579 in the present release) and the distribution of these variants at the codon level. For example, codon 248 (CGG) is the most frequent site of mutation of the p53 gene as 1,109 mutations have been collected in the database (8.34%). They correspond to 13 different mutational events. In addition, a table displays the number of variants harboring a specific number of mutations. For example, one variant was reported 614 times while 115 are detected only four times and 581 only once. This raises the question of the significance of the very rare variants.
3. The “distribution of events” routine is presented as a table and in the form of a histogram. It also indicates the distribution of G:C>A:T transitions localized in a CpG dinucleotide.
4. The “distribution of mutations” is a graphical routine, which shows the number of p53 mutations along the protein.

5. “Distribution by exons” displays the frequency of each type of mutation (point mutation, frame shift mutation, and complex mutation) in each of the 11-p53 exons.
6. “Structure” is used to define the site of p53 mutations according to two parameters: the structure of the region in which these mutations occur and the phylogenetic preservation of this region according to Soussi and May [1996].
7. The “global analysis” function gives an overall view of the various mutation types found. It is usually performed on a subset of the database.

The following functions have been specifically designed for the UMD-p53 database:

1. The “cancer distribution” function displays the various cancer types with the corresponding number of associated p53 mutations.
2. “UV induced mutations,” “tandem mutations,” and “coding strand mutations” are three routines specifically dedicated to the study of UV exposure. The first one searches for G:C>A:T transitions on pyrimidine doublets, the second searches for tandem mutations characteristic of UV damage, and the third searches for preferential mutations on the non-coding strand of the p53 gene. For example, when we compared mutations found in internal tumors to mutations from skin cancers, we observed that only the second group is associated to mutations involving a pyrimidine doublet. The same difference is observed for tandem mutations, which are characteristic to the action of UV rays at the DNA level resulting in the formation of dipyrimidines. It is therefore easy to determine the strand that carries the putative mutagenic lesion. The result shows that the distribution of mutations found in skin cancers preferentially occurs in the non-coding strand as expected, as it is known that the repair process of the coding strand is more efficient. If this type of analysis is easy to perform in the case of UV-related cancers, it is much more difficult for other cancers where the mutagenic targets of the various associated carcinogens have not been formally identified.
3. The “prescreening” function is used to analyze the strategy adopted by authors to search for p53 mutations. It displays the p53 region scanned for mutations, the number of references using this strategy, and the number of found mutations. In the current release of the UMD-p53 database, 39% of reports have only scanned the exons 5 to 8 region, while only 13% have studied the entire coding region of the p53 gene. The prescreening function also allows compiling information about the prescreening technique used. Today 60% of analyses have used a prescreening step—being in 50% of cases the SSCP.

OTHER TP53 DATABASES

Various TP53 mutation databases are available on the web. However, as suggested by Olivier et al. [2002], not all the sites referring to themselves as p53 databases provide updated information. Today only two databases are regularly updated: the UMD-p53 database (<http://p53.curie.fr/> and www.umd.necker.fr:2001/) and the IARC TP53 database (www.iarc.fr/P53). These two databases provide access to germline and somatic mutations with the possibility to download the entire database. They differ by format, annotations, and characteristics of their respective web-search engines, and the UMD-p53 database includes an optimized on-line searching tool that can be used to select records prior to all analyses. Both websites also provide additional information on TP53 mutations and their biological significance, as well as many links with other websites or database entries on TP53. For additional information about other TP53 databases see Olivier et al. [2002].

Only one piece of software, running a p53 database on a personal computer (either Macintosh[®] or PC[®]), is available today. This UMD-p53 software is freely available. It includes all data and functions described in this paper and can be downloaded at: ftp://umd.necker.fr/UMD-ftp/p53_demo/.

CONCLUSIONS

The mutation in the TP53 gene is the most common alteration in human cancer and the availability of a high-quality database was necessary to make the most of these data. The UMD-p53 database was created to achieve this purpose. From the first compilation, which contained only 4,200 mutations up to the present one with 14,971 entries, several new types of analysis are now possible. A large spectrum of dedicated routines has been created and is essential for molecular epidemiology, pharmacogenetic, and patient-based studies. This database is freely available and can be accessed via the internet at URL: www.umd.necker.fr:2001/. Since 1998, more than 240,000 queries have been addressed to the UMD-p53 online database. Because the UMD-p53 policy is to collect high quality data, we have elaborated an efficient curation process performed by experts in the fields of molecular biology and oncogenetic. This time-consuming curation will be facilitated if authors could follow simple recommendations: 1) clearly identify each re-cited mutation, 2) provide unique identifiers for samples, and 3) provide a detailed description of the mutations and samples. It would be most useful that publishers and editors collaborate with the database and request that authors of accepted publications give the above information and deposit their data to the UMD-p53 mutation database.

ACKNOWLEDGMENTS

We would like to thank the authors who directly submitted information to the database.

REFERENCES

- Antonarakis SE. 1998. Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum Mutat* 11:1–3.
- Bairoch A, Boeckmann B. 1991. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* 19(Suppl):2247–2249.
- Beroud C, Verdier F, Soussi T. 1996. p53 gene mutation: software and database. *Nucleic Acids Res* 24:147–150.
- Beroud C, Soussi T. 1996. APC gene: database of germline and somatic mutations in human tumors and cell lines. *Nucleic Acids Res* 24:121–124.
- Beroud C, Joly D, Gallou C, Staroz F, Orfanelli MT, Junien C. 1998. Software and database for the analysis of mutations in the VHL gene. *Nucleic Acids Res* 26:256–258.
- Beroud C, Soussi T. 1998. p53 gene mutation: software and database. *Nucleic Acids Res* 26:200–204.
- 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.
- Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell CD, Tung CS. 1986. The GenBank genetic sequence databank. *Nucleic Acids Res* 14:1–4.
- Cariello NF, Beroud C, Soussi T. 1994. Database and software for the analysis of mutations at the human p53 gene. *Nucleic Acids Res* 22:3549–3550.
- Collod-Beroud G, Beroud C, Ades L, Black C, Boxer M, Brock DJ, Godfrey M, Hayward C, Karttunen L, Milewicz D, Peltonen L, Richards RI, Wang M, Junien C, Boileau C. 1997. Marfan database (second edition): software and database for the analysis of mutations in the human FBN1 gene. *Nucleic Acids Res* 25:147–150.
- Cuticchia AJ. 2000. Future vision of the GDB human genome database. *Hum Mutat* 15:62–67.
- den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 15:7–12.
- den Dunnen JT, Antonarakis SE. 2001. Nomenclature for the description of human sequence variations. *Hum Genet* 109:121–124.
- Hamm GH, Cameron GN. 1986. The EMBL data library. *Nucleic Acids Res* 14:5–9.
- Jeanpierre C, Beroud C, Niaudet P, Junien C. 1998. Software and database for the analysis of mutations in the human WT1 gene. *Nucleic Acids Res* 26:271–274.
- Kucera E, Speiser P, Gnant M, Szabo L, Samonigg H, Hausmaninger H, Mittlbock M, Fridrik M, Seifert M, Kubista E, Reiner A, Zeillinger R, Jakesz R. 1999. Prognostic significance of mutations in the p53 gene, particularly in the zinc-binding domains, in lymph node- and steroid receptor positive breast cancer patients. Austrian breast cancer study group. *Eur J Cancer* 35:398–405.
- Lou MA, Tseng SL, Chang SF, Yue CT, Chang BL, Chou CH, Yang SL, Teh BH, Wu CW, Shen CY. 1997.

- Novel patterns of p53 abnormality in breast cancer from Taiwan: experience from a low-incidence area. *Br J Cancer* 75:746–751.
- Mineta H, Borg A, Dictor M, Wahlberg P, Akervall J, Wennerberg J. 1998. p53 mutation, but not p53 overexpression, correlates with survival in head and neck squamous cell carcinoma. *Br J Cancer* 78:1084–1090.
- Nakanishi H, Tomita Y, Myoui A, Yoshikawa H, Sakai K, Kato Y, Ochi T, Aozasa K. 1998. Mutation of the p53 gene in postradiation sarcoma. *Lab Invest* 78:727–733.
- Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. 2002. The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat* 19:607–614.
- Pearson PL. 1991. The genome data base (GDB): a human gene mapping repository. *Nucleic Acids Res* 19(Suppl):2237–2239.
- Ramnani DM, Wistuba II, Behrens C, Gazdar AF, Sobin LH, Albores-Saavedra J. 1999. K-ras and p53 mutations in the pathogenesis of classical and goblet cell carcinoids of the appendix. *Cancer* 86:14–21.
- 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.
- Soulitzis N, Sourvinos G, Dokianakis DN, Spandidos DA. 2002. p53 codon 72 polymorphism and its association with bladder cancer. *Cancer Lett* 179:175–183.
- Soussi T, May P. 1996. Structural aspects of the p53 protein in relation to gene evolution: a second look. *J Mol Biol* 260:623–637.
- Soussi T, Beroud C. 2001. Assessing TP53 status in human tumours to evaluate clinical outcome. *Nat Rev Cancer* 1:233–240.
- Soussi T, Beroud C. 2003. Significance of TP53 mutations in human cancer: a critical analysis of mutations at CpG dinucleotides. *Hum Mutat* 21:192–200.
- Suzuki T, Shidara K, Hara F, Nakajima T. 1994. High frequency of p53 abnormality in laryngeal cancers of heavy smokers and its relation to human papillomavirus infection. *Jpn J Cancer Res* 85:1087–1093.
- Tsai MH, Lin CD, Hsieh YY, Chang FC, Tsai FJ, Chen WC, Tsai CH. 2002. Prognostic significance of the proline form of p53 codon 72 polymorphism in nasopharyngeal carcinoma. *Laryngoscope* 112:116–119.
- Varret M, Rabes JP, Thiart R, Kotze MJ, Baron H, Cenarro A, Descamps O, Ebhardt M, Hondelijn JC, Kostner GM, Miyake Y, Pocovi M, Schmidt H, Schuster H, Stuhmann M, Yamamura T, Junien C, Beroud C, Boileau C. 1998. LDLR Database (second edition): new additions to the database and the software, and results of the first molecular analysis. *Nucleic Acids Res* 26:248–252.
- Wautot V, Vercherat C, Lespinasse J, Chambe B, Lenoir GM, Zhang CX, Porchet N, Cordier M, Beroud C, Calender A. 2002. Germline mutation profile of MEN1 in multiple endocrine neoplasia type 1: search for correlation between phenotype and the functional domains of the MEN1 protein. *Hum Mutat* 20:35–47.
- Wen WH, Reles A, Runnebaum IB, Sullivan-Halley J, Bernstein L, Jones LA, Felix JC, Kreienberg R, el-Naggar A, Press MF. 1999. p53 mutations and expression in ovarian cancers: correlation with overall survival. *Int J Gynecol Pathol* 18:29–41.