256–258 *Nucleic Acids Research, 1998, Vol. 26, No. 1* *© 1998 Oxford University Press*

Software and database for the analysis of mutations in the VHL gene

Christophe Béroud[1,2,*], Dominique Joly[1], Catherine Gallou[1], Frédéric Staroz[1], Marie Thérèse Orfanelli[1] and Claudine Junien[1]

[1]INSERM U383, Hôpital Necker-Enfants Malades, Université René Descartes Paris V, Paris, France and [2]Service de Pharmacie, Hôpital Broussais, Paris, France

Received September 2, 1997; Revised and Accepted October 15, 1997

ABSTRACT

VHL is a tumor suppressor gene localized on chromosome 3p25–26. Mutations of the VHL gene were described at first in the heritable von Hippel-Lindau disease and in the sporadic Renal Cell Carcinoma (RCC). More recently, VHL has also been shown to harbor mutations in mesothelioma and small cell lung carcinoma. To date more than 500 mutations have been identified. These mutations are mainly private with only one hot spot at codon 167 associated with pheochromocytoma. The germline mutations are essentially missense while somatic mutations include deletions, insertions and nonsense. To standardize the collection of these informations, facilitate the mutational analysis of the VHL gene and promote the genotype–phenotype analysis, a software package along with a computerized database have been created. The current database and the analysis software are accessible via the internet and world wide web interface at the URL: http://www.umd.necker.fr

INTRODUCTION

The von Hippel-Lindau disease is a dominantly inherited familial cancer syndrome, with an incidence of 1 in 36 000, predisposing to the development of retinal, cerebellar and spinal hæmangioblastomas, pancreatic cysts and carcinoma and RCC (1,2). RCC, the most frequent malignancy in the adult kidney, is usually sporadic and its phenotype is extremely heterogeneous. Both sporadic and familial RCC bear relation as clear cells subtype accounts for most tumors. It has been postulated that these tumors have a common carcinogenic pathway and that at least one gene should be altered in both forms. This hypothesis has been confirmed with the cloning of the VHL gene and the identification of germline and somatic mutations of this gene in VHL patients and sporadic RCC. The human VHL gene encodes a 213 amino acid (aa) protein expressed in all tissues (3). To understand its functions, different groups have identified and characterized proteins which interacts with pVHL. Three groups have shown that pVHL stably associates with the two regulatory subunits B and C of the transcription elongation factor, elongin (4–6). Duan *et al*. showed that pVHL can inhibit transcriptional elongation by elongation factor sIII (4). Pause and colleagues showed that the trimeric pVHL-elongin B-C specifically associates with the Hs-CUL-2, a member of the Cdc53 family of proteins (7). Tsuchiya *et al*. have shown that a novel protein, VBP1, can form a complex with pVHL *in vivo* (8) but its role remains unknown. Concurrently, investigations were performed on identified mutations of this gene. Today, 314 mutations have been described in VHL families (3,9–18), 143 reported in sporadic RCC (17,19–24) and 49 in cell lines (3,17,22,25). If most germline VHL mutations are missense alterations, most of the sporadic mutations result in truncated protein. The collection of a large number of mutations is then necessary to understand this discrepancy, to identify key residues in the biological function of the protein and to establish correlations between the localization of the mutation and specific phenotypes.

To handle the flow of all known VHL mutations and study genotype/phenotype relations, we adapted the 'universal mutation database' software to this gene (26–28).

DATABASE AND SOFTWARE

The database of VHL mutations was developed using the 'Universal Mutation Database' tool (26–28). It contains all mutations localized in the coding region of the VHL gene. In an attempt to standardize the numbering system used to describe VHL mutations, all mutations which have been previously reported using the numbering scheme of Latif *et al*. (3) have been renumbered. Amino acid positions are derived from the numbering system used in Kuzmin *et al*. (29) who identified the VHL gene promotor and defined the initiation codon at position +71 from the original sequence described by Latif *et al*. (3). The current version of the database contains 507 entries. When the same mutation was reported in more than one article, only the first report was taken into account. For each mutation, information is provided at several levels: at the gene level (exon and codon number, wild type and mutant codon, mutational event, mutation name), at the protein level (wild type and mutant amino acid), at the clinical level (angioma, hemangioblastoma, pheochromocytoma, RCC), at the molecular level (LOH of the second allele) and at the histological level (Robson stage).

The software package contains routines for the analysis of the VHL database that were developed with the 4th dimension″ (4D) package from ACI. The use of the 4D SGDB gives access to

*To whom correspondence should be addressed at: INSERM UR 383, Hôpital Necker-Enfants Malades, 149-161, rue de Sèvres, 75745 Paris Cedex 15, France. Tel: +33 1 44 49 44 84; Fax: +33 1 47 83 32 06; Email: beroud@necker.fr

Table 1. A section of the database in Excel spreadsheet format

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
122	1	262	88	TGG	AGG	T->A	Tv	No	Kind1255	Trp	Arg	VHL	Germ line		3/3	3/3	0/3	2/3		6
123	1	263	88	TGG	TAG	G->A	Ts	No	T193	Trp	Stop	Clear Cell Carcinoma	Tumor						I	2
124	1	263	88	TGG	TCG	G->C	Tv	No	CrgJap8	Trp	Ser	VHL	Germ line		0/5	0/5	0/5	5/5		7
125	1	263	88	TGG	TCG	G->C	Tv	No	Casen°2	Trp	Ser		Tumor			1/1				14
126	1	263	88	TGG	TCG	G->C	Tv	No	Kind8YO	Trp	Ser	VHL	Germ line				0/1			17
127	1	263	88	TGG	TTG	G->T	Tv	No	9	Trp	Leu		Tumor							29
128	1	266	89	CTC	CAC	T->A	Tv	No	H 28	Leu	His	Mesothelioma	Cell line							26
129	1	266	89	CTC	CAC	T->A	Tv	No	UMRC 3	Leu	His		Cell line						IV	13
130	1	266	89	CTC	CCC	T->C	Ts	No	VHL3	Leu	Pro	VHL	Germ line		3/11	7/11		5/11		12
131	1	266	89	CTC	CCC	T->C	Ts	No	kind8ce	Leu	Pro	VHL	Germ line				0/1			31
132	1	266	89	CTC	CCC	T->C	Ts		Kind21CE	Leu	Pro	VHL	Germ line				0/1			17
133	1	266	89	CTC	CCC	T->C	Tv	No	Kind 59	Leu	Pro	VHL	Germ line				0/1			9
134	1	266	89	CTC	CGC	T->G	Tv	No	SKRC 9M	Leu	Arg	Clear and Granular Cells	Cell line						IV	13
135	1	269	90	AAC	ATC	A->T	Tv	No	UOK 135g	Asn	Ile	Clear Cell Carcinoma	Cell line						IV	13
136	1	268	90	AAC	TAC	A->T	Tv	No	C35	Asn	Tyr	Clear Cell Carcinoma	Tumor	Yes						11
137	1	274	92	GAC	del1b	Stop at 158	Fr.		T60	Asp	Fr.	Clear Cell Carcinoma	Tumor						I	2
138	1	277	93	GGC	AGC	G->A	Ts	Yes	Kind4873	Gly	Ser	VHL	Germ line				1/1			31
139	1	277	93	GGC	AGC	G->A	Ts	Yes	Kind62ce	Gly	Ser	VHL	Germ line				2/2			31
140	1	277	93	GGC	del1a	Stop at 158	Fr.		Kind16	Gly	Fr	VHL	Germ line							29
141	1	278	93	GGC	GAC	G->A	Ts	No	Kind2547	Gly	Asp	VHL	Germ line		3/5	4/5	2/5	0/5		6
142	1	280	94	GAG	del11b	Stop at 127	Fr.		T225	Glu	Fr.	Clear Cell Carcinoma	Tumor						I	2

Each line represents a single VHL mutation. The columns contain the following information and abbreviations:
Column A: file number.
Column B: exon number at which the mutation is located. Exons are numbered with respect to the translational initiation site given by Kuzmin *et al*. (29).
Column C: nucleotide position at which the mutation is located, numbered as above.
Column D: codon number at which the mutation is located, numbered as above. If the mutation spans more than one codon, e.g., there is a deletion of several bases, only the first (5′) codon is entered.
Column E: normal base sequence of the codon in which the mutation occurred.
Column F: mutated base sequence of the codon in which the mutation occurred. If the mutation is a base pair deletion or insertion this is indicated by 'del' or 'ins' followed by the number of bases deleted or inserted and the position of this deletion or insertion in the codon (a, b or c). The nucleotide position is the first that is deleted or the one following the insertions. For example, 'del66b' is a deletion of 66 bases including the second base of the codon; 'ins4b' is an insertion of 4 bases occurring between the first and the second base of the codon.
Column G: concerns base substitutions. It gives the base change, by convention, read from the coding strand. If the mutation predicts a premature protein-termination, the novel stop codon position is given, e.g., 'Stop at 115'.
Column H: mutational event (transition/transversion or frameshift).
Column I: indicates if the mutation is a transition occurring at a CpG dinucleotide.
Column J: name of the tumor/patient/cell line as given by the authors.
Column K: wild type amino acid.
Column L: mutant amino acid. Deletion and insertion mutations which result in frameshift are designated by 'Fr'.
Column M: cancer.
Column N: origin of the mutation (tumor, cell line, xenograft or germline).
Column O: LOH, if available. 2, two alleles remaining; 1, only one allele remaining; ?, no information available or non-informative.
Columns P–S: clinical information if available. P, angioma; Q; hemangioblastoma; R, pheochromocytoma; S, RCC. For each clinical location, the number of affected patients/number of patients carrying the mutation is indicated.
Column T: tumor stage according to Robson staging system.
Column U: reference number indicating the publication in which the mutation is described. Full citations (authors, year, title, volume, pages) are provided with the database. If the same mutation has been reported for the same patient in different papers only one entry is made. If the same mutation has been reported for unrelated patients, a separate entry is made for each patient.

optimized multicriteria research and sorting tools to select records from any field. Moreover, all routines already developed for other databases were added to the VHL package: (i) 'Position' studies the distribution of mutations at the nucleotide level to identify preferential mutation sites; (ii) 'Mutational events' is comparable to (i) but also indicates the type of mutational event. The result can either be displayed as a table or in a graphic representation; (iii) 'Frequency of mutations' allows one to study the relative distribution of mutations at all sites and to sort them according to their frequency. A graphic representation is also available and displays a cumulative chart of mutation distributions; (iv) 'Frequency of events' is similar to (iii) but also indicates if mutations are localized in a CpG dinucleotide; (v) 'Distribution of mutations' and 'Binary comparison' are two graphical routines which allow one to compare the distribution of mutations in up to eight groups of records and display the result either in one or two graphics; (vi) 'Stat exons' studies the distribution of mutations in the different exons. It enables detection of a statistically significant difference between observed and expected mutations; (vii) 'Insertions and deletions analysis' searches for repeated sequences surrounding the mutation and possibly involved in the mutational mechanism.

Subsequently, the software will be expanded as the database grows and according to the requirements of its users. New functions could be implemented. Table 1 describes a section of the database.

AVAILABILITY

The current database and subsequent updated versions are (will be) available on request to C.B. on floppy disc either Apple or PC formatted. Notification of omissions and errors in the current

version as well as specific phenotypic data would be gratefully received by the corresponding author. If you use this database, please cite the present article as reference. The software package is available on a collaborative basis.

The current database and the analysis software will be accessible via the internet and world wide web interface in January 1998 at the URL: http://www.umd.necker.fr

ACKNOWLEDGEMENTS

This work was supported by the Ligue Nationale contre le Cancer and the Association pour la Recherche sur le Cancer (ARC).

REFERENCES

1 Maher,E. and Yates,J. (1990) *Q. J. Med.*, **77**, 1151–1163.
2 Richard,S., Olschwang,S., Chauveau,D. and Resche,F. (1995) *Méd. Sci.*, **11**, 43–51.
3 Latif,F., Tory,K., Gnarra,J., Yao,M., Duh,F.-M. *et al.* (1993) *Science*, **260**, 1317–1320.
4 Duan,D.R., Pause,A., Burgess,W.H., Aso,T., Chen,D.Y.T. *et al.* (1995) *Science*, **269**, 1402–1406.
5 Kibel,A., Iliopoulos,O., DeCaprio,J.A. and Kaelin,W.G.J. (1995) *Science*, **269**, 1444–1446.
6 Kishida,T., Stackhouse,T.M., Chen,F., Lerman,M.I. and Zbar,B. (1995) *Cancer Res.*, **55**, 4544–4548.
7 Pause,A., Lee,S., Worrell,R., Chen,D.Y.T., Burgess,W.H. *et al.* (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 2156–2161.
8 Tsuchiya,H., Iseda,T. and Hino,O. (1996) *Cancer Res.*, **56**, 2881–2885.
9 Brauch,H., Kishida,T. and Glavac,D. (1995) *Hum. Genet.*, **95**, 551–556.
10 Chen,F., Kishida,T., Yao,M., Hustad,T., Hustad,T. *et al.* (1995) *Hum Mut.*, **5**, 66–75.
11 Clinical Research Group for VHL in Japan (1995) *Hum. Mol. Genet.*, **4**, 2233–2237.
12 Crossey,P.A., Richards,F.M., Foster,K., Green,J.S., Prowse,A. *et al.* (1994) *Hum. Mol. Genet.*, **3**, 1303–1308.
13 Glavac,D., Neumann,H., Wittke,C., Jaenig,H., Masek,O. *et al.* (1996) *Hum Genet.*, **98**, 271–280.
14 Kanno,H., Shuin,T., Kondo,K., Ito,S., Hosaka,M. *et al.* (1996) *Jpn J. Cancer Res.*, **87**, 423–428.
15 Maher,E., Webster,A., Richards,F., Green,J., Crossey,P. *et al.* (1996) *J. Med. Genet.*, **33**, 328–332.
16 Richards,F., Payne,S., Zbar,B., Affara,N., Ferguson-Smith,M. *et al.* (1995) *Hum. Mol. Genet.*, **4**, 2139–2145.
17 Whaley,J., Naglich,J., Gelbert,L., Hsia,Y., Lamiell,J. *et al.* (1994) *Am. J. Hum. Genet.*, **55**, 1092–1102.
18 Zbar,B., Kishida,T., Chen,F., Schmidt,L., Maher,E. *et al.* (1996) *Hum. Mut.*, **8**, 348–357.
19 Bailly,M., Bain,C., Favrot,M.-C. and Ozturk,M. (1995) *Int. J. Cancer*, **63**, 660–664.
20 Béroud,C., Joly,D., Staroz,F., Gallou,C., Martin,N. *et al.* (1998) Submited for publication.
21 Foster,J., Prowse,A., van den Berg,A., Fleming,S., Hulsbeek,M.M.F. *et al.* (1994) *Hum. Mol. Genet.*, **3**, 2169–2173.
22 Gnarra,J.R., Tory,K., Weng,Y., Schmidt,L., Wei,M.H. *et al.* (1994) *Nature Genet.*, **7**, 85–89.
23 Kanno,H., Kondo,K., Ito,S., Yamamoto,I., Fujii,S. *et al.* (1994) *Cancer Res.*, **54**, 4845–4847.
24 Shuin,T., Kondo,K., Torigoe,S., Kishida,T., Kubota,Y. *et al.* (1994) *Cancer Res.*, **54**, 2852–2855.
25 Sekido,Y., Bader,S., Latif,F., Gnarra,J.R., Gazdar,A.F. *et al.* (1994) *Oncogene*, **9**, 1599–1604.
26 Béroud,C. and Soussi,T. (1996) *Nucleic Acids Res.*, **24**, 121–124. [See also this issue *Nucleic Acids Res.* (1998) **26**, 269–270.]
27 Béroud,C., Verdier,F. and Soussi,T. (1996) *Nucleic Acids Res.*, **24**, 147–150. [See also this issue *Nucleic Acids Res.* (1998) **26**, 200–204.]
28 Collod,G., Béroud,C., Soussi,T., Junien,C. and Boileau,C. (1996) *Nucleic Acids Res.*, **24**, 137–140.
29 Kuzmin,I., Duh,F.-M., Latif,F., Geil,L., Zbar,B. *et al.* (1995) *Oncogene*, **10**, 2185–2194.