Nucleic Acids Research, 1996, Vol. 24, No. 1 **127–131**

PAH Mutation Analysis Consortium Database: a database for disease-producing and other allelic variation at the human *PAH* locus

Liem Hoang, Susan Byck, Lynne Prevost and Charles R. Scriver* (curators)

The DeBelle Laboratory, McGill University-Montreal Children's Hospital Research Institute, 2300 Tupper Street, Montreal, Quebec, Canada H3H 1P3 and The Department of Human Genetics, McGill University, Montreal, Quebec, Canada

Received August 21, 1995; Revised and Accepted October 4, 1995

ABSTRACT

The PAH Mutation Analysis Consortium (81 investigators, 26 countries) is engaged in mutation detection at the human PAH locus. Ascertainment of probands occurs largely through newborn screening for hyperphenylalaninemia. A relational database records allelic variation (disease-producing and polymorphic) at the locus. Information is distributed by Newsletter, diskette (WINPAHDB software stand-alone executable on IBM compatible hardware), and at a 'real' site on the Worldwide Web (http://www.mcgill.ca/pahdb). The database presently records (Sept. 27, 1995) 248 alleles in 798 different associations (with polymorphic haplotype, geographic region and population) along with additional information. The database, as a record of human genetic diversity, at a particular locus, contributes to the study of human evolution and demic expansion; it also has medical relevance.

Phenylalanine hydroxylase enzyme (PAH) catalyses conversion of the essential amino acid phenylalanine to tyrosine (1). Mutations in the corresponding human gene (symbol *PAH*) may cause deficient enzyme activity and elevate levels of phenylalanine in body fluids unless dietary intake of phenylalanine is reduced. The associated diseases, notably phenylketonuria (PKU), are autosomal recessive and impaired cognitive development and function are their chief manifestations. Screening newborns for hyperphenylalaninemia (HPA) is now one of the most widely used genetic tests in the world, and early treatment of PKU is the prototype for prevention of a 'genetic' disease.

The *PAH* locus

The *PAH* gene is located on chromosome 12q, band region 24.1, and it comprises 13 exons spanning ~90 kb of DNA. Disease-producing allelic variation, ascertained through probands with persistent HPA, has been identified in all exons (1–3). The majority of alleles occur in the 3′ half of the gene which encodes domains in the polypeptide PAH subunit for binding of tetrahydrobiopterin cofactor, the putative active site for catalysis and the contact regions between subunits in the homotetrameric form of the enzyme. Most PAH mutations identified so far result in severe enzyme deficiency with the PKU phenotype, but some cause less severe loss of enzyme activity which is associated with a 'non-PKU HPA' phenotype. A few alleles may be silent polymorphisms without functional significance; 248 variant alleles are presently known (Fig. 1).

A suite of polymorphic diallelic restriction sites (4) and a VNTR (5), in linkage disequilibrium (6), account for over 70 recognized polymorphic haplotypes (7). Of the seven diallelic sites, five are analyzed by the PCR reaction and two (*Eco*RI, *Eco*RV) by Southern blotting. A multiallelic VNTR (5) in the 3′ untranslated region of the PAH gene has at least eight variants in size, and there is additional sequence variation in the 30 bp repeat unit (8). A short tandem tetranucleotide $(TCTA)_n$ repeat (STR) in intron 3 is multiallelic by size variation (9). Locations of these polymorphic markers in the gene are shown in Figure 1. They provide the rich background upon which mutations affecting PAH function are expressed and they serve analysis of the population genetic variation of *PAH* alleles.

THE DATABASE

We have developed a relational database for PAH mutations. Entries originate from members of the Mutation Analysis Consortium (81 investigators in 26 countries, listed in the database), by independent submissions, and from a regular survey of the literature.

Accessibility

The Consortium distributes a Newsletter and hardcopy of the database. The electronic form has a dedicated software package (WINPAHDB) using Microsoft FOXPRO 2.6A, designed to be stand-alone executable on an IBM compatible computer with MS-DOS 3.3 and Microsoft Windows 3.0 or higher; at least 4 Meg of RAM and 5 Meg of disk space are needed. Various fields can be searched and records generated from a Browse Window

* To whom correspondence should be addressed

Figure 1. General structure of the human PAH gene (~90 kb at 12q24.1). Introns are numbered 3′ to corresponding exon (vertical bars). Polymorphic markers are indicated below the gene (open boxes for diallelic markers; shaded boxes for the two multiallelic markers). Mutation types are indicated by symbols (see box for code); those placed below the gene are splice-deficient alleles; those above it are substitutions, deletions and insertions. Nomenclature follows the convention of Beaudet and Tsui (11).

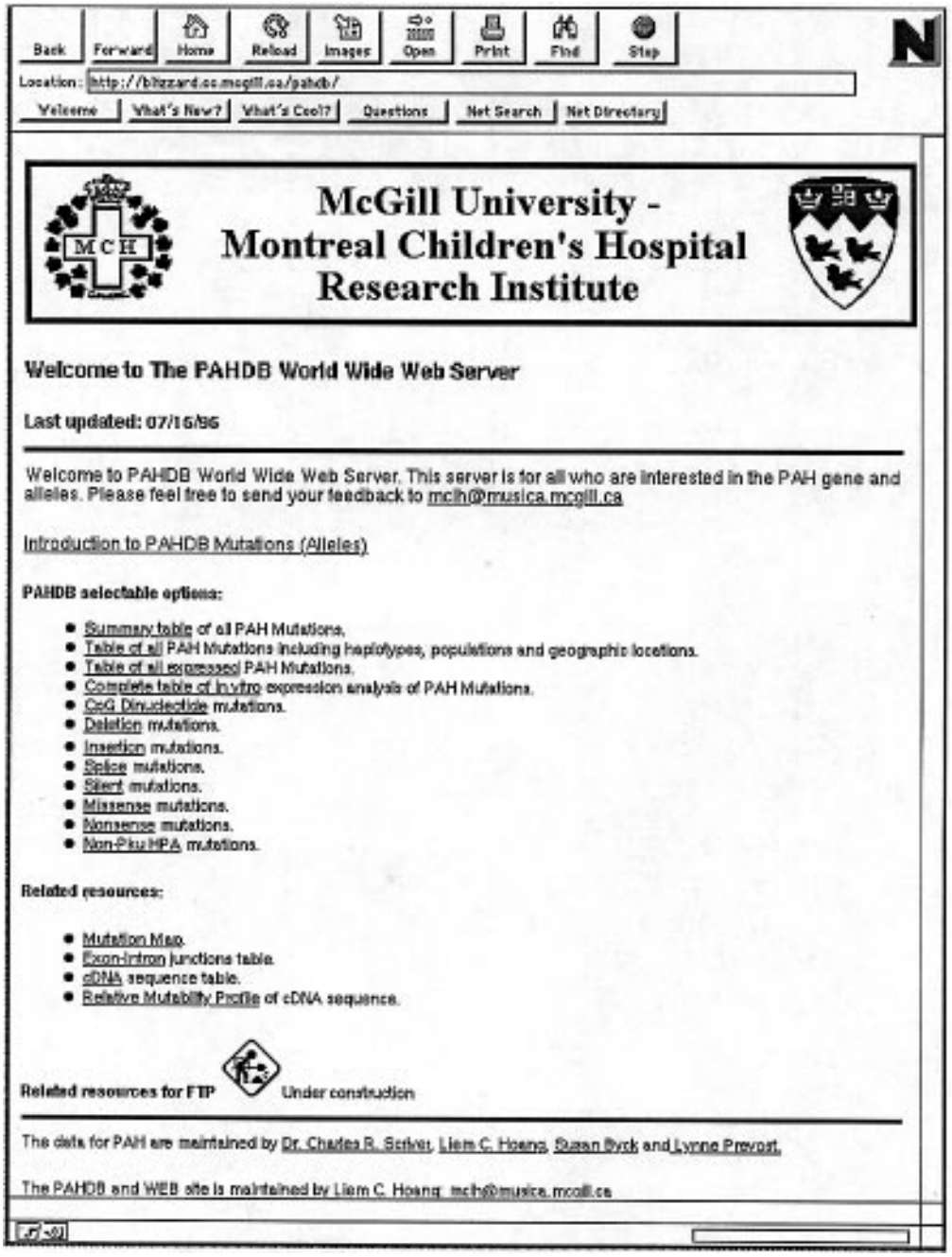

Figure 2. The home page of the World Wide Web server of the PAH Mutation Analysis Consortium Database.

can be printed or saved as a text file. It is also possible to print a single record with all associated information. The Consortium has opened a Web site at http://www.mcgill.ca/pahdb, where the database is 'real' rather than 'virtual' (10) and can be accessed on line. Figure 2 shows the home page of the World Wide Web server of the database. Downloading can be achieved at increased speed by disabling the graphic component.

Content

The database contains extensive information relevant to the PAH gene and allelic variation at this locus (Table 1). Nucleotide change (mutation) is given along with associations between the variant allele, the polymorphic haplotype, the population of origin and corresponding geographic region. Relative frequencies of alleles, an important feature in the study of population genetic variation, can be derived from the database. Deletion or creation of a restriction site [including any amplification created restriction site (ACRS)] is recorded and the source of information (publication or dated personal submission to the Database with accession number) is given. Figure 3 shows a sample of data.

Nomenclature

Mutations are described by a conventional nomenclature (11). Splice mutations are identified by their intronic nucleotide number;

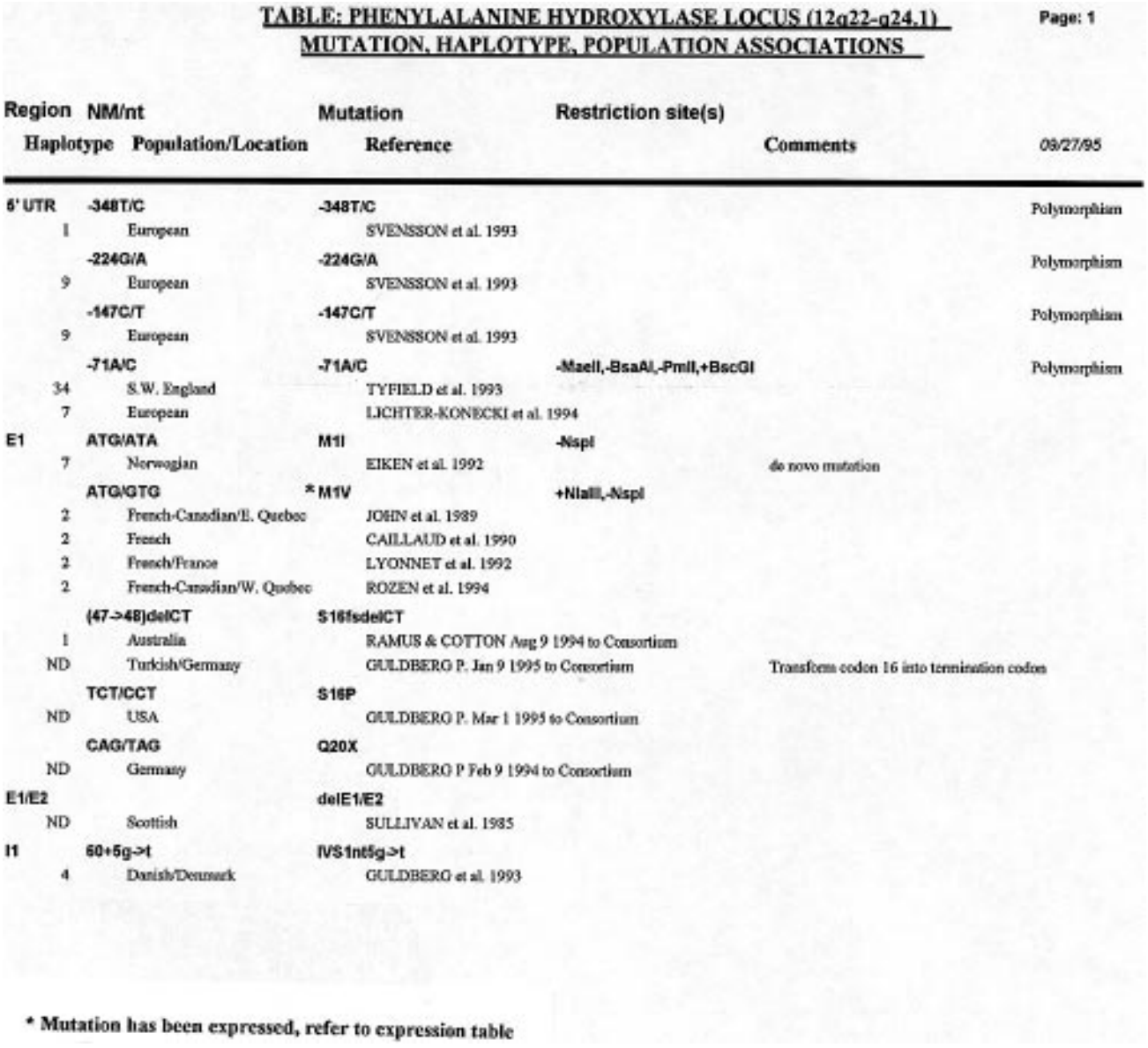

TABLE: PHENYLALANINE HYDROXYLASE LOCUS (12q22-q24.1) Page: 1
MUTATION, HAPLOTYPE, POPULATION ASSOCIATIONS

Region	NM/nt	Mutation	Restriction site(s)		
Haplotype	Population/Location	Reference		Comments	09/27/95
5' UTR	-348T/C	-348T/C			Polymorphism
1	European	SVENSSON et al. 1993			
	-224G/A	-224G/A			Polymorphism
9	European	SVENSSON et al. 1993			
	-147C/T	-147C/T			Polymorphism
9	European	SVENSSON et al. 1993			
	-71A/C	-71A/C	-MaeII,-BsaAI,-PmlI,+BscGI		Polymorphism
34	S.W. England	TYFIELD et al. 1993			
7	European	LICHTER-KONECKI et al. 1994			
E1	ATG/ATA	M1I	-NspI		
7	Norwegian	EIKEN et al. 1992		de novo mutation	
	ATG/GTG	* M1V	+NlaIII,-NspI		
2	French-Canadian/E. Quebec	JOHN et al. 1989			
2	French	CAILLAUD et al. 1990			
2	French/France	LYONNET et al. 1992			
2	French-Canadian/W. Quebec	ROZEN et al. 1994			
	(47->48)delCT	S16fsdelCT			
1	Australia	RAMUS & COTTON Aug 9 1994 to Consortium			
ND	Turkish/Germany	GULDBERG P. Jan 9 1995 to Consortium		Transform codon 16 into termination codon	
	TCT/CCT	S16P			
ND	USA	GULDBERG P. Mar 1 1995 to Consortium			
	CAG/TAG	Q20X			
ND	Germany	GULDBERG P Feb 9 1994 to Consortium			
E1/E2		delE1/E2			
ND	Scottish	SULLIVAN et al. 1985			
I1	60+5g->t	IVS1nt5g>t			
4	Danish/Denmark	GULDBERG et al. 1993			

* Mutation has been expressed, refer to expression table

Figure 3. A sample of data in the PAH Mutation Database. The asterisk indicates that the mutation has been analyzed by *in vitro* expression analysis; results described in a separate field in the database.

positive numbers originate from the 5′ end of the intron while negative numbers indicate an allele correspondingly located from the 3′ end of the intron. The polymorphic haplotypes on which mutations are found are named according to PAH-specific conventions (5,7,9).

Graphic Content

At present, four items are in the database: (i) a diagram of the PAH gene and location of all alleles (see Fig. 1). (ii) The cDNA sequence, renumbered from the original sequence of Kwok *et al.* (12) to accommodate new findings at the 5′ end reported by Konecki *et al.* (13). The cDNA sequence is numbered positively when moving 3′ from the first translated codon; the 5′ untranslated region is numbered negatively as one moves upstream from the first codon. (iii) The nucleotide sequences at the exon–intron boundaries in the gene (14). (iv) The predicted mutability profile of the gene, derived by using the MUTPRED computer program with the PAH cDNA (15).

Other features

A separate linked database (prepared by Paula Waters Ph.D.) summarizes current information about *in vitro* expression analysis of PAH mutations; plasmid and host system; levels of mRNA, immunoreactive protein and enzyme activity in the mutant phenotype; source of information.

As of Sept. 27, 1995, 4089 mutant PAH chromosomes had been characterized from probands with hyperphenylalaninemia, in a wide array of human populations and geographic regions. They carried 248 different phenotype-modifying alleles in 798 different associations between mutation, haplotype, population and geographic region. The curators welcome all suggestions to improve the database and its accessibility.

COMMENT

Informatics is an essential part of genomic research and mutation detection plays a role in identifying polymorphic markers for gene discovery and in recognition of candidate genes. Our database will be part of the Mutation Newsgroup in the Bio Sci Electronic Newsgroup Network (RGH Cotton, pers. comm. Aug. 1995). It is also linked to OMIM (Catalogue entry 261600) and other genome databases.

Table 1. Contents of PAH Mutation Analysis Consortium Database

A. Allelic variation at locus (Table, fields and figure in database).

- Mutations
  - nucleotide change
  - assigned name (e.g. M1V, IVS12nt1)
  - affected region (e.g. untranslated, exon/codon, intron)
  - type [e.g. substitution (nonsense or missense), deletion, insertion, frameshift, splice, etc.].
  - Restriction site: (deleted, created)
- Polymorphic haplotype defined by
  - diallelic system (7 sites)
  - VNTR (1 site)
  - STR (1 site)

  Together the two sites provide a 'mini-haplotype' in the absence of the full (diallelic) haplotype
- Population (source of allele)
- Geographic region (source of allele)
- Hyperphenylalaninemia Phenotype (associated)
  - Phenylketonuria (PKU): 'severe', 'mild'
  - non-PKU hyperphenylalaninemia
- Citation:
  Submission date/person
  Published article

B. Structural information for *PAH* gene

- cDNA sequence [see refs. (12,13) (Figure in database)]
- exon–intron junction sequences [see ref. (14) (Figure in database)]

C. Predicted mutability profile of *PAH* cDNA (Figure in database) (from MUTPRED analysis profile [see ref. (15)])

D. Consortium membership

The PAH Mutation Analysis Consortium database is a record of mutations and biological memory encoded in the human genome at a particular locus. As a prototype, it is one of an increasing number of databases describing disease-producing allelic variation contributing to the Human Genetic Diversity Project (16). The *PAH* locus is one of at least 500 human loci currently known to harbour more than one disease-producing allele (V.A. McKusick pers. commun. Sept. 16, 1995). Allelic variation at the PAH locus reveals how histories of mutations and histories of human populations are contingent and contribute to the study of human evolution and demic expansion (17). The PAH database also has medical interest because particular mutations alter the corresponding enzymic and metabolic phenotypes with potential consequences for health (18); at the same time this knowledge is likely to have predictive value relevant to the treatment of HPA phenotypes (19).

ACKNOWLEDGEMENTS

We thank: Jean Rey and Savio Woo who asked us to be curators of the PAH mutation database and members of the Consortium without whom there would be less to report and with much more difficulty. Richard Cotton, Elizabeth Evans and Victor McKusick, Ken Morgan, Leah Simpkin, Saeed Teebi and Paula Waters gave advice, encouragement and help. This venture into in silico research is supported by the Medical Research Council of Canada, the Canadian Genetic Diseases Network (Networks of Centers of Excellence Program), the Fonds de la Recherche en Santé du Quebec and the Réseau de Génétique Humaine Appliqué (Quebec).

REFERENCES

1 Scriver,C.R., Kaufman,S., Eisensmith,R.C. and Woo,S.L.C. (1995) In Scriver,C.R., Beaudet,A.L., Sly,W.S. and Valle,D. (eds), *The Metabolic and Molecular Bases of Inherited Disease.* McGraw-Hill Inc., New York, Vol. 7, pp. 1015–1075.
2 Scriver,C.R., Eisensmith,R.C., Woo,S.L.C. and Kaufman,S. (1994) *Annu. Rev. Genet.*, **28**, 141–165.
3 Scriver,C.R. (1994) *Acta Paediatr.*, **407**, 11–18.
4 Lidsky,A.S., Ledley,F.D., DiLella,A.G., Kwok,S.C.M., Daiger,S.P., Robson,K.J.H. and Woo,S.L.C. (1985) *Am. J. Hum. Genet.*, **37**, 619–634.
5 Goltsov,A.A., Eisensmith,R.C., Konecki,D.S., Lichter-Konecki,U. and Woo,S.L.C. (1992) *Am. J. Hum. Genet.*, **51**, 627–636.
6 Feingold,J., Guilloud-Bataille,M., Feingold,N., Rey,F., Berthelon,M. and Lyonnet,S. (1993) *Dev. Brain Dysfunct.*, **6**, 26–31.
7 Eisensmith,R.C. and Woo,S.L.C. (1992) *Am. J. Hum. Genet.*, **51**, 1445–1448.
8 Byck,S., Morgan,K., Tyfield,L., Dworniczak,B. and Scriver,C.R. (1994) *Hum. Mol. Genet.*, **3**, 1675–1677.
9 Goltsov,A.A., Eisensmith,R.C., Naughton,E.R., Jin,L., Chakraborty,R. and Woo,S.L.C. (1993) *Hum. Mol. Genet.*, 2, 577–581.
10 Harper,R. (1995) *Trends Genet.*, **11**, 223–228.
11 Beaudet,A.L. and Tsui,L. (1993) *Hum. Mut.*, **2**, 245–248.
12 Kwok,S.C.M., Ledley,F.D., DiLella,A.G., Robson,K.J.H. and Woo,S.L.C. (1985) *Biochemistry*, **24**, 556–561.
13 Konecki,D.S., Wang,Y., Trefz,F.K., Lichter-Konecki,U. and Woo,S.L.C. (1992) *Biochemistry*, **31**, 8363–8368.
14 DiLella,A.G., Kwok,S.C.M., Ledley,F.D., Marvit,J. and Woo,S.L.C. (1986) *Biochemistry*, **25**, 743–749.
15 Cooper,D.N. and Krawczak,M. (1993) *Human Gene Mutation.* Bios Scientific Publishers, Oxford, UK.
16 HUGO (1995) *Genome Digest*, 2, 12–15.
17 Scriver,C.R., Byck,S., Prevost,L., Hoang,L. The PAH Mutation Analysis Consortium. (1995) The phenylalanine hydroxylase locus: a marker for the history of phenylketonuria and human genetic diversity. In Weiss K and Ward R (eds), *Variation in the Human Genome.* CIBA Foundation Symposium No. 197, Wiley and Sons, London, pp. 000.
18 Guldberg,P., Mikkelsen,I., Henriksen,K.F., Lou,H.C. and Güttler,F. (1995) *Eur. J. Pediatr.*, **154**, 551–556.
19 Guldberg,P., Henriksen,F.F., Thöny,B., Blau,N. and Güttler,F. (1994) *Genomics*, **21**, 453–455.