Nucleic Acids Research, 1996, Vol. 24, No. 1 **141–146**

Somatic point mutations in the *p53* gene of human tumors and cell lines: updated compilation

M. Hollstein*, B. Shomer[1], M. Greenblatt[2], T. Soussi[3], E. Hovig[4], R. Montesano[5] and C.C. Harris[6]

German Cancer Research Center, Im Neuenheimer Feld 280,D-69120 Heidelberg, Germany, [1]EMBL Outstation–European Bioinformatics Institute, Hinxton Hall, Hinxton, Cambridge, UK, [2]University of Vermont College of Medicine, Health Sciences Complex, Burlington, VT 05405, USA, [3]Institut de Genetique Moleculaire, Unite 301 INSERM, 27 rue Juliette Dodu, 75010 Paris, France, [4]Institute for Cancer Research, Norwegian Radium Hospital, 0310 Oslo, Norway, [5]International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France and [6]Laboratory of Human Carcinogenesis, National Cancer Institute, Bethesda, MD 20892, USA

Received October 11, 1995; Accepted October 16, 1995

ABSTRACT

In 1994 we described a list of ~2500 point mutations in the *p53* gene of human tumors and cell lines which we had compiled from the published literature and made available electronically through the file server at the EMBL Data Library. This database, updated twice a year, now contains records on 4496 published mutations (July 1995 release) and can be obtained from the EMBL Outstation–the European Bioinformatics Institute (EBI) through the network or on CD-ROM. This report describes the criteria for inclusion of data in this database, a description of the current format and a brief discussion of the current relevance of *p53* mutation analysis to clinical and biological questions.

INTRODUCTION

The *p53* tumor suppressor gene encodes a phosphoprotein with cancer inhibiting functions. Development of cancer in humans typically involves loss of wild-type p53 activity and consequently of the growth control it confers. Point mutation is a common route to this loss of wild-type p53 function in tumors. Since the discovery in 1989 by Vogelstein and colleagues of tumor-specific, missense mutations in the *p53* gene of two patients with colorectal cancer there has been an exponential increase in the number of *p53* mutations identified in human tumors, which will surpass the 5000 mark by the end of 1995 (Fig. 1). Simple inspection of these data, such as analysis of the frequency, kind and distribution of mutations along the *p53* gene, can be used to make inferences about potentially important sources of mutation in the human setting. The mutations are also informative with regard to the biochemistry of interactions between the p53 protein and its macromolecular targets and the biological factors contributing to selection of mutant clones during tumorigenesis in different tissues. Finally, *p53* mutations have a potentially important clinical role in diagnostic and therapeutic procedures.

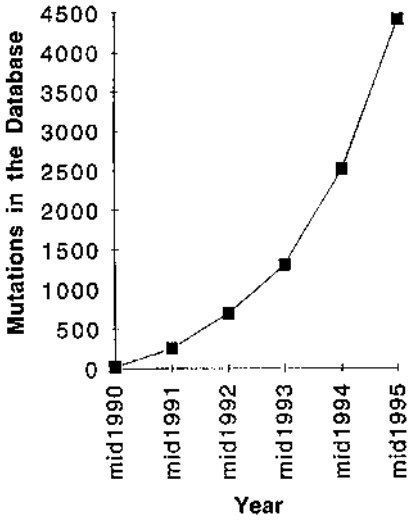

Figure 1. Human tumor mutations in the *p53* gene published since 1989.

SCOPE OF THE MUTATION COMPILATION AND CRITERIA FOR INCLUSION OF DATA

The purpose of our database is to retrieve and arrange data from the accumulating literature on *p53* mutations in a standardized electronic format. This provides a powerful means for manipulation, comparison, search and retrieval of records describing the nature of *p53* mutations in various cancers. Software which enables more sophisticated mutation spectra analysis is also available (1). In the forthcoming year updates of our database may also be available from EBI in Microsoft Access™ format to

* To whom correspondence should be addressed

facilitate a better interchangeability of data with other programs. The database will be made searchable on-line at EBI through the network by using the SRS indexing system (2).

This database resides in a spreadsheet containing published data on human somatic point mutations in cell lines, primary tumors, neoplastic and pre-neoplastic tissues. The list is updated twice a year and the July 1995 release contained 4496 mutation records. Germ line mutations, including those identified in families with Li–Fraumeni syndrome and known polymorphisms of the human *p53* gene, are thus not included. Experimentally induced mutations in tumor cells or cell lines *in vitro*, unselected mutations in histologically normal tissue (see for example 3) and *p53* mutations in animal tumors are also beyond the present scope of our task.

The *p53* point mutations as selected from the literature and entered into the database have been identified by DNA sequencing of either PCR-amplified material or cloned PCR products. Some investigators have included a preliminary mutation screening step, such as the SSCP or DGGE/CDGE techniques. Analysis is usually confined to exons 5–8, where most mutations cluster. A bias against identification of DNA sequence alterations outside this region can thus be expected. Mutations identified by digestion of DNA with restriction enzymes and demonstration of an RFLP are not entered, nor are mutations in tumors that were screened in only part of the mutation cluster mid-region (exons 5–8). If identical samples and mutations were published in more than one article only one of the reports is referenced and the data are entered only once in the database. If the identical mutation was found in two separate samples from the same patient, for example in the primary tumor and in the metastatic tissue or in the cell line derived from the tumor or in two separate biopsies at the same site and from the same patient, the mutation is presumed to be a single event and is entered only once.

Information that does not permit us to identify the nature or location of the mutation has not been entered, such as band shift by SSCP of a PCR product without subsequent DNA sequencing. Data on loss of heterozygosity at the *p53* locus, gene rearrangements and immunohistochemical analysis with anti-p53 antibody are not part of this compilation. (Note that many papers initially retrieved by standard bibliographic search systems using 'p53' and 'mutation' as keywords will thus not figure in the reference list constituting the source of mutations in this database.) Mutations identified in tumors are presumed to be somatic unless: (i) analysis of normal tissue from the same patient demonstrated that the mutation was constitutional in that individual; (ii) the mutation corresponded to one of the known constitutional polymorphisms of the human *p53* gene (at codons 21, 31, 47, 72 and 213), as these are unlikely to be mutations that arose in the tumors sequenced.

OBTAINING THE DATABASE

The data described are provided to the scientific community in several formats. The database is available as an Excel™ spreadsheet, which requires use of the Microsoft Excel™ program on either an MS-DOS™ system or an Apple Macintosh™. The data has also been converted into a flatfile format modeled after the standard used by the EMBL nucleotide sequence database. In this format the data are stored in an ASCII text file with each column of the spreadsheet represented by a special line type. The flatfile format can be used on any computer system and with standard text editors. The data can be obtained from the EBI network server by using one of the following methods: (i) anonymous ftp to ftp.ebi.ac.uk, in the directory/pub/databases/p53; (ii) World Wide Web access using the URL http://www.ebi.ac.uk/, selecting 'documentation and software' and going to the databases selection; (iii) using a gopher to the site gopher.ebi.ac.uk (port 70), selecting the option 'access to various databases' and 'p53'; (iv) sending an email message containing the line 'help p53' to netserv@ebi.ac.uk.

The p53 directory contains the original spreadsheet file as a Macintosh binHex4.0 self-extracting archive. The release notes are included in the file p53.doc and the references are in p53.ref. Also included are the database in flatfile format (p53.dat) and the data in tab-delimited (data.tab) and comma-delimited (data.comma) formats for usage by other data management systems.

DESCRIPTION OF SPREADSHEET FORMAT

Each row (record) represents a single tumor mutation with an arbitrarily assigned unique identity number. The spreadsheet columns contain the following information and abbreviations. [Note: important additions or changes to the original format (4) are written in italics.]

Column A

Unique mutation identity number. *The prefix 'ID' next to this number in the original format has now been removed, in order to facilitate sorting records by identity number.* Tandem mutations (two adjacent base substitutions) are considered as one mutation event and are entered together, therefore tandem mutations have only one identity number and are a single record.

Column B

Codon number at which the mutation is located (1–393). If a tandem dinucleotide mutation spans two codons both codons are entered. If other mutations span more than one codon, e.g. there is a deletion of several bases, only the first (5′) codon is entered (see note below regarding deletion, insertion and complex mutations). If the mutation is located in intron sequences this is indicated by 'intron' and the intron number.

Column C

Normal and 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'.

Column D

Nucleotide position at which the mutation is located (1–1179), numbered from the ATG codon to the termination codon. This information is not entered in the present versions for deletion, insertion, intron and complex mutations (see note below).

Column E

Base change, read from the coding strand by convention, for base substitutions. For deletions (indicated by '–') and insertions (indicated by '+') the number of bases deleted or inserted is given in parentheses.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	1	143	GTG to GCG	428	T to C	Cx3	Colon	1				Val->Ala		
2	2	175	CGC to CAC	524	G to A	Cx1	Colon	1			yes	Arg->His		
3	3	132	AAG to CAG	394	A to C	BT 20	Breast	2		L		Lys->Gln		
4	4	249	AGG to AGC	747	G to C	BT 549	Breast	2		L		Arg->Ser		
5	5	280	AGA to AAA	839	G to A	MDA231	Breast	2		L		Arg->Lys		
6	6	285	GAG to AAG	853	G to A	BT 474	Breast	2		L		Glu->Lys		
7	7	157	GTC to TTC	469	G to T	OZ1	HCC	3				Val->Phe		
8	8	249	AGG to AGT	747	G to T	OZ2	HCC	3				Arg->Ser		
9	9	249	AGG to AGT	747	G to T	OZ3	HCC	3				Arg->Ser		
10	10	249	AGG to AGT	747	G to T	OZ4	HCC	3				Arg->Ser		
11	11	286	del		(-8)	OZ5	HCC	3				frameshift		
12	12	152	CCG to CTG	455	C to T	Ca 4	Esophagus	4	Barretts		yes	Leu->Pro		revised 12/94
13	13	155	ACC to GCC	463	A to G	Ca19	Esophagus	4	Barretts			Thr->Ala		
14	14	175	CGC to CAC	524	G to A	Ca 5	Esophagus	4	Barretts		yes	Arg->His		

Figure 2. Section of the database in Excel spreadsheet format showing record 1 as an example.

Column F

The name or number given by the authors to the tumor sample or cell line is entered here. If the name is not distinctive, e.g. if the publication refers to samples as tumors 1, 2, 3, etc., then we have arbitrarily assigned a name, usually the first letters of the first author's name followed by the number in the series. Thus entirely different tumors may have, by chance, a similar or identical trivial tumor sample name (see recommendations to authors below). If more than one mutation has been found in the same sample the tumor name in column F is suffixed with an apostrophe.

Column G

Anatomical site or type of tumor as described in the publication cited. Abbreviations used in this column are: HCC, hepatocellular carcinoma; Leuk/Lym, leukemias and lymphomas; cholangioca, cholangiocarcinoma; choriocarc, choriocarcinoma; colon, cancers of the colon or rectum.

Column H

Reference number indicating the publication in which the mutation is described. The full citation (authors, title, journal, pages, year) is given as a separate text file in the electronic version of the database. If the same mutation in the same tumor sample or cell line has been published in more than one article only one report is referenced and the data are entered only once in the database (see Notes to authors, below).

Column I

This is a column with heterogeneous notes, usually containing comments regarding the tumor or the patient, such as histological type of tumor or exposure history or other information emphasized by authors reporting the mutations. The terminology used by the authors has been retained and no attempt has been made to complete the data with unpublished information or to standardize the entries.

Abbreviations of tumor subtype or cell type are as follows: SCLC, small cell lung cancer; adenoca, adenocarcinoma; adenosq, mixed adenosquamous carcinoma; medullobl, medulloblastoma; SCC, squamous cell carcinoma; TCC, transitional cell carcinoma; NPC, nasopharyngeal carcinoma. For abbreviations of sarcoma subtypes or leukemia and lymphoma subclassifications, e.g. ATL (adult T cell leukemia), refer to the cited reference. Uniformity of these abbreviations in the different reports has not been verified.

Other abbreviations: UC, ulcerative colitis; FAP, familial adenoma polyposis; XP, xeroderma pigmentosum; HPV+ or HPV–, tumor harboring or lacking human papilloma virus DNA; diff, differentiated tumor; undiff, undifferentiated tumor; CIS, carcinoma *in situ*; premal, premalignant.

Other information. 1. 'metastasis' specifies that the DNA analyzed for the mutation was obtained from metastatic tissue (the primary tumor location is in column G). 2. Examples of exposure history: tobacco smoke; radon gas, etc.

Column J

An entry 'L' indicates that the material examined was from a tumor cell line. If there is no entry the material is from tumor tissue or biopsy (most instances), xenograft or unspecified.

Column K

Mutations that are single base transitions at CpG dinucleotides, i.e. CpG→TpG or CpG→CpA are designated by 'yes'. If there is no entry the mutation does not fall into this category.

Column L

Chain terminating mutations due to single base substitutions are designated by '(three letter amino acid abbreviation)→stop'. Frame-shift mutations are designated by 'frameshift', whereas in-frame deletions and insertions are designated 'deletion' or 'insertion'. Mutations that do not result in an amino acid change are designated 'silent', while mutations that occurred in intron sequences are sometimes indicated by the term 'splicing', even though in most instances it was not determined whether splicing errors did result from the mutation (some of these base changes are likely to be phenotypically silent).

Column M

If the information on the nature or location of the mutation in the reference is ambiguous or contradictory the letter 'e' appears in this column. Where possible we have made a presumptive correction of the published discrepancy in the database entry.

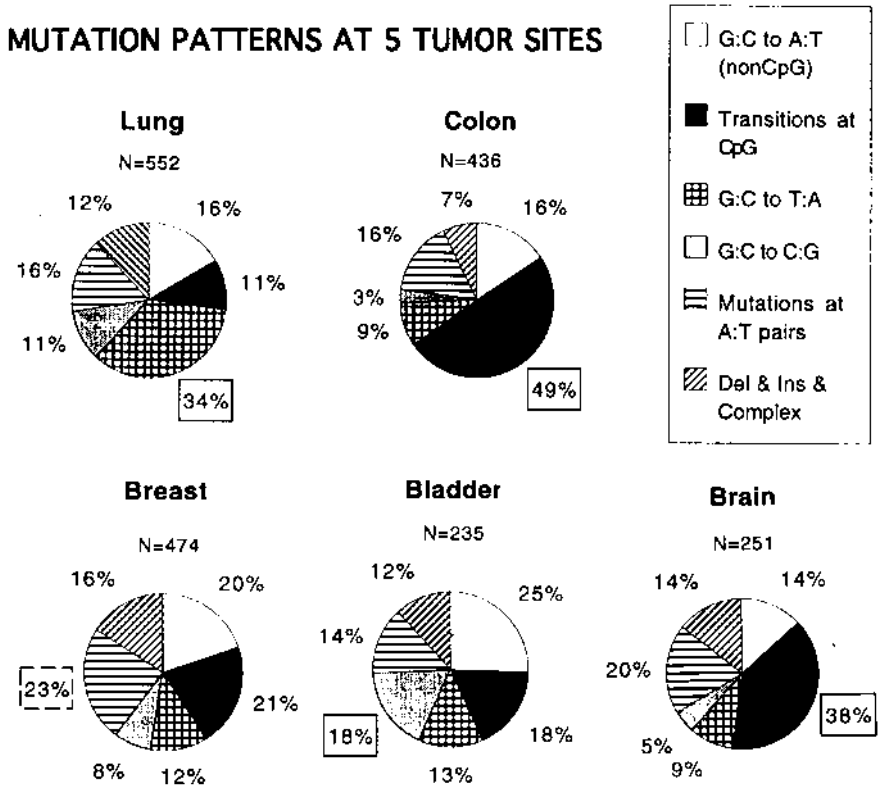

Figure 3. Examples of *p53* mutation patterns by tumor site (July 1995 release).

Column N

When a record requires correction or information is added the date of the revision will be noted in this column. Record revisions are due to inherent inconsistencies or ambiguities in the published report or omissions or errors during data entry.

Example

The information on the mutation identified as 1 on the spreadsheet (Fig. 2) is as follows.

Column A Unique mutation identity number is 1.
Column B Mutation is located in codon 143.
Column C The wild-type codon sequence is GTG and the mutated allele sequence is GCG.
Column D The nucleotide position of the base change is 428.
Column E The base change is T→C.
Column F The tumor sample name is Cx3.
Column G The DNA harboring the mutation was obtained from a tumor of the colon or rectum.
Column H The mutation is reported in reference 1 of the database (5).
Column I (Blank: no comments regarding histology or patient data were entered for this record.)
Column J (Blank: the DNA analyzed was not obtained from a cell line.)
Column K (Blank: the mutation is not a base transition substitution at a CpG dinucleotide.)
Column L The amino acid change is from valine to alanine.
Column M (Blank: no errors to the published information have been found.)
Column N (Blank: no corrections or additions to the record have been made since it was entered.)

Comments regarding insertion, deletion and complex mutations

Since precise information describing these kinds of mutations requires a more expanded format than the current spreadsheet and requires conventions for presenting sequence changes that are inherently ambiguous, e.g. base deletions in repetitive sequences, our compilation presently gives minimal information on these mutations. A HUGO initiative is underway to standardize and establish conventions for specific locus mutation databases. We will attempt to follow the recommendations in revising our format and in providing additional sequence information on this class of mutation.

Notes to authors

When tumor sample mutations are described for the second time in a new publication describing the *p53* mutations in another context we recommend this be stated in a footnote to the table where the mutations are re-listed, also indicating which tumor mutations were reported previously. Providing tumor samples with unique case numbers would also help to avoid redundancies in the database.

The inherent inconsistencies we have detected (~2% of all records) in reported mutations were usually traceable to typo-

graphical errors in the publication (e.g. codon 223 instead of 232), to reading the genetic code from the wrong DNA strand or to misnumbering the codons in sequencing film illustrations.

The electronic form of the database may be cited by referencing this *Nucleic Acids Research* article.

p53 MUTATIONS AND CANCER

Mutation patterns by tumor site and patient risk group (reviewed in 6,8,9)

Simple inspection of the mutation data in human tumors when only 5% of the current data were available had already shown that mutation patterns could be quite different from one tumor site to another (7). This is now indisputably clear (Fig. 3), with certain mutation pattern differences surpassing a statistical significance of $P < 0.001$ (χ^2 statistic). Mutation profile differences between histological subtypes of cancer are also emerging as more detailed clinical information on histopathological diagnosis of the tumor samples sequenced is included in published reports. There are still few instances where exposure to a specific carcinogenic risk factor can be linked to a distinctive mutation profile that corresponds to the known mutagenic specificity of the agent. The clearest examples remain: (i) aflatoxin-associated hepatocellular carcinoma (usually in conjunction with HBV chronic infection) and a G→T hot spot mutation at *p53* codon 249; (ii) UV exposure and tandem base substitutions at dipyrimidine sequences in skin cancers; (iii) tobacco smoking and G→T transversions in lung tumors. It is not unexpected that examples of associations between mutation pattern and risk factors are still few, since exposures are typically complex or difficult to assess and because more subtle differences in mutation patterns are not seen until large numbers of samples are examined. *p53* mutation studies with exposure cohorts matched for various parameters that could influence mutation patterns (e.g. genetic background, age, other exposures, etc.) are, by nature, long-term epidemiological studies.

Applications of *p53* mutation analysis to biological research

Tumor mutations in cancer genes are potentially sequence changes that have provided a growth advantage and allowed clonal outgrowth of mutant cells. Analysis of *p53* tumor mutations suggests that they are, indeed, for the most part biologically selected sequence changes (6–11). The biological properties of *p53* mutants vary inherently and in different cellular contexts (9,12).

Although >250 codons in the *p53* gene are potential human tumor mutation sites, mutations at only five of these comprise 25% of all mutations found thus far in human tumors (Fig. 4). Several superimposing factors are responsible for such *p53* mutation hot spots: (i) inherent vulnerability of a given DNA sequence to base sequence changes (13), such as repetitive DNA sequences leading to base slippage and deletion or insertion mutations (14,15), or CpG dinucleotides in codons of conserved residues, prone to transition mutations following deamination of 5-methylcytosine (16); (ii) vulnerabilty to attack by mutagenic agents, e.g. UV-induced base damage at dipyrimidine sequences (17); (iii) relative susceptibility of the non-transcribed DNA strand to persistence of DNA damage and mutation due to

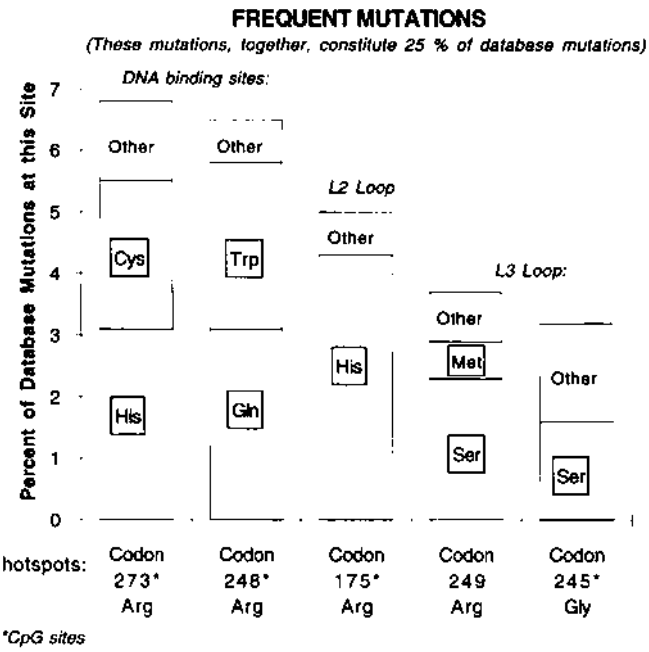

Figure 4. Frequent sites of human tumor mutations in the *p53* coding sequence.

preferential repair of the transcribed strand by transcription-coupled repair processes (7,18); (iv) a slow rate of transcription-coupled repair at some DNA sites (19,20); (v) codons corresponding to biologically critical amino acids that have a crucial role in maintaining the three-dimensional structure of the p53 protein or that make direct contact with the DNA target sequence. Co-crystallization of the p53 core domain with target DNA revealed that human tumor hot spot mutations correspond to amino acid residues at the DNA–protein interface or that maintain the structural integrity of p53 necessary for its tumor suppressor activity (21–23).

Clinical applications (24–26 and references therein)

Since tumor mutations are so heterogeneous, they can be used to investigate the clonality of tumor cell foci in a patient. Finding identical mutations would provide molecular evidence in favor of clonality. Screening exfoliated cells for *p53* mutation can be used to test for relapse in former cancer patients with a previously identified mutation in the primary tumor. Mutation analysis of lymph nodes and surgical margins can be used to assess tumor spread. Since p53 antibodies are found in the serum of some cancer patients with tumor mutations, this is being explored as a diagnostic tool and to measure treatment response. The exact nature of the mutation is an important determinant of humoral immune response and *p53* mutation studies in this context promise to bring insight in the field of cancer immunology.

ACKNOWLEDGEMENTS

We are grateful to Dr R.-P. Kraft at the German Cancer Research Center Library for bibliographic assistance and to Dr F. Hutchinson at Yale University for help and advice.

REFERENCES

1 Cariello,N., Piegorsch,W.W., Adams,W.T., Skopek,T.R. (1994) *Carcinogenesis*, **15**, 2281–2285.
2 Etzold,T. and Argos,P. (1993) *Comput. Appl. Biosci.*, **9**, 49–57.
3 Aguilar,F., Harris,C.C., Sun,T., Hollstein,M. and Cerutti,P. (1994) *Science*, **264**, 1317–1319.
4 Hollstein,M., Rice,K., Greenblatt,M.S., Soussi,T., Fuchs,R., Sorlie,T., Hovig,E., Smith-Sorensen,B., Montesano,R. and Harris,C.C. (1994) *Nucleic Acids Res.*, **22**, 3551–3555.
5 Baker. *et al*. (1989) *Science*, **244**, 217–221.
6 Greenblatt,M.S., Bennett,W.P., Hollstein,M.C. and Harris,C.C. (1994) *Cancer Res.*, **54**, 4855–4878.
7 Hollstein,M., Sidransky,D., Vogelstein,B. and Harris,C.C. (1991) *Science*, **253**, 49–53.
8 Soussi,T., Legros,Y., Lubin,R., Ory,K. and Schlichtholz,B. (1994) *Int. J. Cancer*, **57**, 1–9.
9 Levine,A.J. (1993) *Annu. Rev. Biochem.*, **62**, 623–651.
10 Lane,D.P. (1994) *Int. J. Cancer*, **57**, 623–627.
11 Krawczak,M., Smith-Sorensen,?., Schmidtke,J., Kakkar,V.V., Cooper,D.N. and Hovig,E. (1995) *Hum. Mutat.*, **5**, 48–57.
12 Forrester,K., Lupold,S.E., Ott,V.L., Chay,C.H., Band,V., Wang,X.W. and Harris,C.C. (1995) *Oncogene*, in press.
13 Lindahl,T. (1993) *Nature*, **362**, 709–715.
14 Kunkel,T.A. (1990) *Biochemistry*, **29**, 8003–8011.
15 Jego,N., Thomas,G. and Hamelin,R. (1993) *Oncogene*, **8**, 209–213.
16 Jones,P.A., Buckley,J.D., Henderson,B.E., Ross,R.K. and Pike,M.C. (1991) *Cancer Res.*, **51**, 3617–3620.
17 Brash,D.E., Rudolph,J.A., Simon,J.A., Lin,A., McKenna,G.J., Baden,H.P., Halperin,A.J. and Ponten,J. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 10124–10128.
18 Mellon,P. and Hanawalt,P. (1989) *Nature*, **342**, 995–998
19 Kunala,S. and Brash,D.E. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 11031–11035.
20 Tornaletti,S. and Pfeifer,G. (1991) *Science*, **263**, 1436–1438.
21 Prives,C. (1994) *Cell*, **78**, 543–546.
22 Cho,Y., Gorina,S., Jeffrey,P.D. and Pavletich,N.P. (1994) *Science*, **265**, 346–355.
23 Borresen,A-L. *et al*. (1995) *Genes Chromosomes Cancer*, **14**, 71–75.
24 Harris,C.C. and Hollstein,M. (1993) *New Engl. J. Med.*, **329**, 1318–1327.
25 Sidransky,D. (1995) *J. Natl. Cancer Inst.*, **17**, 17–29.
26 Sidransky,D. and Hollstein,M. (1995) *Annu. Rev. Med.* in press.