106. Thun, M. J., Henley, S. & Patrono, C. Nonsteroidal anti-inflammatory drugs as anticancer agents: mechanistic, pharmacologic, and clincal issues. *J. Natl Cancer Inst.* **94**, 252–266 (2002).
107. Bernstein, L. *et al.* Tamoxifen therapy for breast cancer and endometrial cancer risk. *J. Natl Cancer Inst.* **91**, 1654–1662 (1999).
108. Bergman-Jungestrom, M., Gentile, M., Lundin, A. C. & Wingren, S. Association between *CYP17* gene polymorphism and risk of breast cancer in young women. *Int. J. Cancer* **84**, 350–353 (1999).
109. United States Department of Health and Human Services. in *Reducing Tobacco Use: A Report of the Surgeon General — Executive Summary*. 7 (United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, Georgia, 2000).

Acknowledgements
We are most thankful for the critical contribution of peer reviewers.

Competing interests statement
The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
ATM | *BRCA1* | *BRCA2* | PSA | *TP53*
National Cancer Institute: http://www.cancer.gov
breast cancer | colon cancer | endometrial cancer | lung cancer | non-Hodgkin lymphoma | prostate cancer

FURTHER INFORMATION

American Cancer Society Cancer Prevention Study II: http://www.cancer.org/docroot/RES/content/RES_6_1_History_and_Accomplishments.asp
IARC Handbooks on Cancer Prevention: http://www.iarc.fr/IARCPress/general/prev.pdf
IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: http://monographs.iarc.fr/
IARC Press: http://www.iarc.fr/IARCPress/index.php
National Center for Health Statistics, National Health and Nutrition Survey: http://www.cdc.gov/nchs/nhanes.htm
NCI Early-Detection Research Network: http://www3.cancer.gov/prevention/cbrg/edrn/
US Surgeon General Reports on Smoking and Health: http://profiles.nlm.nih.gov/NN/ListByDate.html
Access to this interactive links box is free online.

SCIENCE AND SOCIETY

Locus-specific mutation databases: pitfalls and good practice based on the p53 experience

Thierry Soussi, Chikashi Ishioka, Mireille Claustres and Christophe Béroud

Abstract | Between 50,000 and 60,000 mutations have been described in various genes that are associated with a wide variety of diseases. Reporting, storing and analysing these data is an important challenge as such data provide invaluable information for both clinical medicine and basic science. Locus-specific databases have been developed to exploit this huge volume of data. The p53 mutation database is a paradigm, as it constitutes the largest collection of somatic mutations (22,000). However, there are several biases in this database that can lead to serious erroneous interpretations. We describe several rules for mutation database management that could benefit the entire scientific community.

Progress has been made over recent years in the cloning of the genes involved in both monogenic and polygenic disorders, including complex diseases such as cancer[1]. For each of these genes, numerous alterations of various types have also been described, ranging from point mutations to large deletions. The future development of new high throughput methods for the detection of mutations will lead to an enormous increase in the detection of new mutations[2]. It is difficult to evaluate the number of mutations reported in the literature to date (more than 50,000 have been collected in various databases), but a similar number could remain unpublished. It is also impossible to predict how many new mutations will be detected over the next 10 years, and the reporting and analysis of these mutations will therefore constitute a major challenge for the future[3,4]. Nevertheless, a number of points can be predicted. First, knowledge of these mutations will be important for treatment decisions as well as for basic science. And second, changes in our environment will lead to variations in the mutational events that modify our genome. Such changes will alter the distribution and/or pattern of mutations leading to the discovery of new and specific hot-spot mutations, so databases will need to be constantly updated. A good example of this is the specific mutation of *TP53* at codon 249 that is only found in hepatocellular carcinoma (HCC) that occurs in countries with a high exposure to the food contaminant aflatoxin B1. In these regions, the high specificity of the 249 mutation has enabled the development of a very sensitive diagnostic procedure that should not be applied when exposure to aflatoxin B1 is not (or no longer) evident[5].

The practical value of mutation analysis
All studies performed to date show that mutations are, in general, not randomly distributed. Hot-spot regions have been demonstrated, corresponding to a region of DNA that is susceptible to mutations (such as CpG dinucleotides), a codon encoding a key residue in the biological function of the protein, or both (BOX 1). Identification of these hot-spot regions and natural mutants is essential to define crucial regions in an unknown protein. In large genes such as neurofibromin 1 (*NF1*; 59 exons, 2,818 amino acids), retinoblastoma 1 (*RB1*; 27 exons, 928 amino acids), adenomatosis polyposis coli (*APC*; 15 exons, 2,843 amino acids), breast cancer 1 (*BRCA1*; 24 exons, 1,863 amino acids) and the titin gene (*TTN*; 363 exons, approximately 25,000 amino acids), detection of point mutations by direct sequencing analysis is difficult because of the size of the target gene. Identification of a hot-spot region allows analysis to be focused on this region, keeping in mind that a negative result should be viewed with caution.

It has also been clearly demonstrated that alterations in a single gene can cause various types of disorders. For example, mutations in *RET* are associated with multiple endocrine neoplasia types IIA[6] and IIB[7], familial medullary thyroid carcinoma[8] and a non-cancerous disorder known as Hirschsprung disease[9,10]. Mutations seem to be localized in specific domains of the protein for each of these disorders. The site of specific alterations at various positions in a given gene is also associated with different clinical features, as in the case of colon cancer and mutations in *APC*. A mutation in the C-terminus of the protein is specifically associated with a secondary abnormality, congenital hypertrophy of the retinal pigment epithelium[11], whereas mutations in the N-terminus are associated with an attenuated phenotype[12]. Analysis of mutations can also lead to the definition of risk factors. For instance, von Hippel–Lindau (VHL) families with mutations in *VHL* that result in truncated proteins have an increased frequency of renal-cell carcinoma (83%) compared with families with *VHL* missense mutations (54%)[13]. In diseases that are characterized by considerable variations in the clinical phenotype between families and also within the same family, such as Marfan syndrome (a connective tissue disorder), it is very important to confirm or formally exclude the diagnosis in high-risk family members as early as possible because of the potentially fatal cardiovascular complications of the disease.

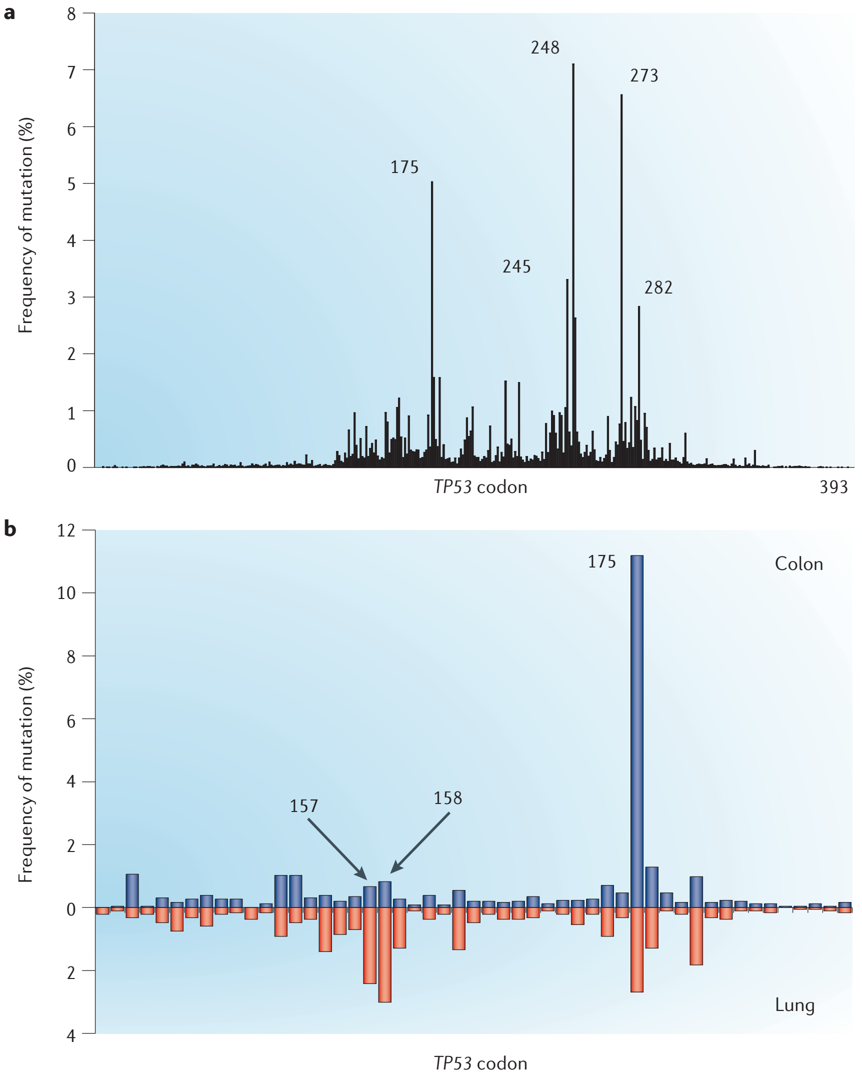
Therefore, it is important to consider mutation databases not only as tools that can provide essential information on protein structure and function, but also as an important framework for the development of new molecular-based diagnostic strategies and patient management.

Collecting mutations and LSDBs

Historically, collections of mutations and variations in human genes have been reported in the published literature. In the mid-1980s, several of these variations were available in the form of various databases, such as the Genome DataBase (GDB)[14], GenBank[15], the European Molecular Biology Laboratory (EMBL)[16] and Swiss-Prot[17]. However, because of the structure of these databases, the extraction of relevant information concerning mutations was almost impossible, so specific software had to be developed to facilitate this. Several teams started to develop specific databases to collect and document mutations in human genes. Today, several hundred locus-specific databases (LSDBs) are available through the Internet and have been recently reviewed[4]. Many of these databases are just a simple list of mutations that cannot be searched. They are also highly heterogeneous in terms of quality and content[4]. One of the main problems of these LSDBs concerns their follow-up. A recent survey of 138 known LSDBs of human gene mutations with available follow-up information found that 40 databases have not been updated since the year 2000 and that another 44 had only been updated between 2001 and 2003. Although the creation of a mutation database can be exciting and gratifying (in terms of publication), follow-up is time-consuming and less stimulating. As recently highlighted in a special report in *Nature*, financial issues are also involved and several databases, including the Asthma and Allergy gene database, were closed due to lack of funding[18]. These LSDBs are sponsored by only a few grants and they are usually developed 'on the side' (the Universal Mutation Database (UMD) for p53, created in 1991, only received one grant from a charity organization in 1995)[19]. Projects to generate central databases that regroup information from multiple LSDBs have been developed, but except for the Human Gene Mutation Database (HGMD), most of them have been abandoned for technical reasons such as incompatibility of the various LSDBs or lack of funding. The HGMD reports more than 40,000 mutations in 1,500 nuclear genes, but only germline

Box 1 | **Origin of p53 mutations**

Distribution of mutations in the p53 protein showing the various mutation hot spots (21,717 mutations in the entire database). Hot spots for somatic mutations can be explained both at the DNA level (the codon is highly susceptible to modification) and the protein level (the residue is essential for the function of the protein). In the case of p53, both explanations prevail. Codon 175 (similar to the other hot-spot codons 245, 248, 273 and 282, shown in part **a**) contains a CpG dinucleotide. In mammalian cells, the cytosine in this dinucleotide is often methylated and it has been shown that the 42 CpG sites of the *TP53* gene are methylated in normal tissue[50]. The higher deamination rate of 5-methylcytosine leading to a T–G mismatch that is not efficiently repaired leads to this high rate of transition in the *TP53* gene. Deamination of cytosine leads to a U–G mismatch that can be removed more efficiently. Various studies have also demonstrated that exogenous carcinogens have a higher affinity for methylated CpG dinucleotides than for their unmethylated counterparts[51,52]. Mutations at every CpG dinucleotide of the *TP53* gene have been reported in the p53 database, albeit at different frequencies[43]. Exogenous carcinogens can also target some specific residues of the *TP53* gene, such as benzo(a)pyrene [B(a)P] diolepoxide (BPDE), one of the carcinogens of tobacco smoke, which binds specifically to codons 157 and 158 *in vitro* (shown in part **b**). This observation explains the predominance of mutations found at these two codons in lung cancers from smokers compared with non-smokers or other cancers.



mutations are included (which excludes most cancers) and each mutation is reported only once[20].

In the early 1990s we decided to develop not just a simple repository of locus-specific mutations but a dynamic database that would include various software tools for the analysis of locus-specific mutations. This project ultimately led to the development of the UMD software[21], which is now recognized by the Human Genome Organization (HUGO) and the Human Genome Variation Society (HGVS) as a reference tool with which to build an LSDB. It was first used to create the UMD p53 database in 1991 and subsequently to develop databases for various genes involved in cancers, such as *APC* in colon cancer[22], *BRCA1* and *BRCA2* in breast cancer, *MEN1* in multiple endocrine neoplasia type 1 (REF. 23), the sulfonylurea receptor, *SUR1*, in hyperinsulinism[21], *RB1* in retinoblastoma, *VHL* in von Hippel–Lindau syndrome[24] and *WT1* in Wilms tumour[25]. UMD software is also used in other databases of genes that are involved in genetic disease, such as *FBN1* in Marfan syndrome[26], *LDLR* in hypercholesterolaemia[27], *VLCAD* in very-long-chain acyl-CoA dehydrogenase deficiency and *DMD* in Duchenne muscular dystrophy. The useful nature of these databases is illustrated by the findings that have been prompted by the information contained within them. The UMD APC LSDB is the only *APC* gene mutation database available to the scientific community. It was the basis for the discovery of the dependence of the second hit (somatic mutations) on the site of the first hit (germline mutations)[28]. It also led to the demonstration that in families with inherited colon cancer that showed high levels of somatic G:C–T:A transversions in *APC*, the disease was caused by Mut-Y homologue (*MYH*) germline mutations[29]. Moreover, analysis of the pattern of *TP53* mutations was the basis for a large number of molecular epidemiology studies demonstrating the relationship between exposure to exogenous carcinogens and p53 mutations in different cancer types[30,31] (BOX 1).

Germline versus somatic mutations

Broadly, DNA alterations have two origins: germline or somatic. Germline mutations (also called constitutional mutations) are found in all cells of the organism, including germline cells, and can be transmitted to the offspring. They are the cause of most hereditary genetic diseases, such as cystic fibrosis, and most myopathies or heamoglobinopathies. They are also involved in inherited cancer syndromes such as familial adenomatous polyposis (caused by mutations in *APC*), familial breast cancer (caused by mutations in *BRCA1* or *BRCA2*) or Li–Fraumeni syndrome (LFS, caused by mutations in *TP53*). Somatic mutations are acquired in somatic tissue during the subject's lifetime and predominantly give rise to neoplastic disease with mutations restricted to the tumour.

Germline mutations are easy to detect when using specific methodology and an adequate screening strategy, but this process is time-consuming and costly for large genes. DNA or RNA is usually extracted from blood cells, leading to large amounts of good quality genetic material. Depending on the type of disease — autosomal-dominant or recessive — the cells will carry either one or two mutations. The biological significance of deletions, insertions or nonsense mutations is usually obvious, but the main problem concerns germline missense mutations. It is difficult to determine whether the detected sequence variant is a causal mutation or a neutral (polymorphic) variation without any effect on phenotype[32]. Only extensive studies that include segregation analysis of the mutation inside the family, phylogenetic conservation of the targeted codon, or statistical analysis can unravel the pathogenicity of the mutation, but often the question remains unresolved. It is usually assumed that variants that are found in more than 1% in the normal population could correspond to polymorphisms, but this feature might be heterogeneous between various regions of the genome.

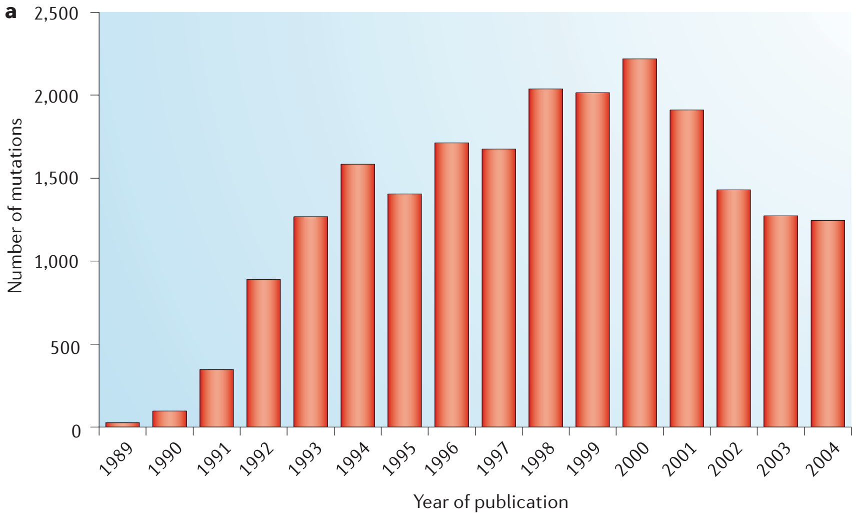
For somatic mutations, this problem usually does not arise, as comparison of normal and neoplastic tissue from the same patient will unambiguously identify the mutation. Although most primary tumours have a few somatic mutations in different genes — in other words, 'driver mutations' selected for a growth advantage — some tumours with a mutator phenotype can have hundreds of mutations. These mutations, termed either 'hitch-hiking mutations' or 'passenger mutations', do not confer any selective advantages and are co-selected with a driver mutation[33,34]. The frequency of these passenger mutations is unknown at the present time and could vary according to the type of cancer. As there is no easy way to distinguish between these two types of mutations, numerous passenger mutations will probably contaminate all mutation databases. Detection of these passenger mutations constitutes an important challenge, as these mutations do not have any clinical relevance and could interfere with database analysis.

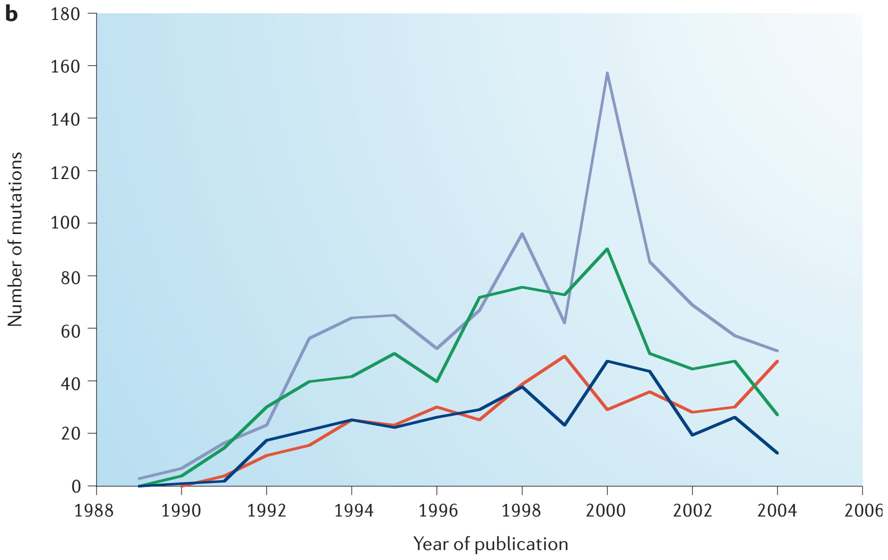
An important problem for somatic mutations concerns the difficulty of their detection. Tumour samples can be highly heterogeneous in terms of content and origin, both of which have a profound impact on the diagnosis. The main problem is contamination of the tumour by normal cells such as stromal cells or infiltrating lymphocytes. This contamination can be low for surgical specimens (less than 20%), but can be more than 90% for biopsies. For biological specimens such as serum, sputum, faeces or urine, tumour cells can account for less than 0.01% of all cells. Pre-screening methodologies such as single-strand conformation polymorphism (SSCP), denaturing high-performance liquid chromatography (DHPLC) or denaturing gradient gel electrophoresis (DGGE) can increase the sensitivity of detection, but not when the target constitutes less than 5% of the original sample material. The second problem concerns the protocol used to conserve the sample. Although material frozen directly after collection generates good quality DNA, DNA from fixed and paraffin-embedded tissue is often degraded and analysis can lead to the artefactual discovery of mutations. Furthermore, the low yield of DNA obtained from these samples can necessitate the use of two rounds of PCR amplification (nested PCR), a procedure known to be error-prone if not carefully controlled.

Most tumour-suppressor genes can be mutated constitutionally and somatically in family syndromes and sporadic cancers. Nevertheless, for each gene, databases for germline and somatic mutations must be clearly dissociated, as they have different interpretations and applications. For somatic alterations, hot-spot regions have been demonstrated, corresponding to a DNA region that is susceptible to mutations, or a codon that encodes a key residue in the biological function of the protein, or both. Whatever the explanation, it is synonymous with a high rate of *de novo* mutations at this position. This information can be useful for pinpointing important regions of the protein. For germline mutations, mutation hot-spots can also be associated with a founder effect in which most carriers of the mutation descend from a single ancestor. Selection of some founder mutations over large periods of time can be explained by their association with certain specific selective advantages. The best example is the ΔF508 mutation found in the cystic fibrosis transmembrane conductance regulator gene (*CFTR*) in 70% of Caucasian patients with cystic fibrosis. Genetic analysis has shown that all of these patients originate from a single ancestor and that this mutation is a

Box 2 | **Growth of p53 mutation publications**

Since the first publication in 1989, there was a constant increase in the number of publications describing *TP53* mutations, culminating in 2000 (see part **a**). The decrease first observed in 2001 is continuing. This is not because of a lack of interest in p53, but because of the difficulty in publishing *TP53* mutations in peer-reviewed journals, owing to a lack of novelty. Furthermore, in recent publications, *TP53* mutations are not fully described owing to journal space considerations. Many laboratories have unpublished *TP53* mutations that are not included in the database. This problem is not specific to *TP53* — it also concerns other genes and raises an important issue related to the publication of mutations. It is estimated that 50% of all mutations in various databases are unpublished and have been collected by the curators. This unpublished information has not been peer-reviewed, which raises the problem of its accuracy and how it should be curated before being entered in a database. The trend for publishing p53 mutation data has also decreased in most instances. The publication trend for two hot-spot mutations (R175H and R248Q) and one moderate mutation (G245S) follows the publication trend observed above with a decrease since 2000 (see part **b**). The publication rate for unique missense mutations has not decreased and has even showed a slight increase in 2004. Several non-exclusive explanations can account for this progression: more thorough analysis of the whole *TP53* gene instead of exons 5 to 8 (50% of unique mutations lie outside exons 5 to 8), analysis of early neoplastic lesions that do not have full loss of p53 activity, and less stringent experimental procedures. The lines shown in part **b** represent the publication trends for R175H (lilac), R248Q (green) 245S (blue) and unique missense mutations (red).





unique molecular event[35]. A founder effect is therefore associated with social or geographical isolation, and the number of patients associated with the same mutation does not correspond to the mutation rate. The increasing rate of discovery of founder mutations in human populations and their heterogeneity in disease penetrance will lead to the development of mutation databases that will require extensive social and clinical information that might preclude their usefulness.

Structure and accessibility

LSDBs often originate as loosely organized compilations of data. Curators choose from the available database management systems or create their own system, depending on their abilities. As no standard has yet been adopted, the way data is presented in LSDBs varies enormously. Most curators use flat file, plain text databases or spreadsheet programs (such as Microsoft Excel) as a simple means to collate and store data on mutations, but neither the search nor the retrieval of specific data are possible. More sophisticated databases use MySQL, an open source database management software that runs on most platforms. A minority of curators use specialized or generic software such as the UMD[36], the Mutation Storage and Retrieval Program (MuStaR)[37], or the Leiden Open Source Variation Database (LOVD)[38]. The use of these complex relational databases allows specific analysis of either the entire database or any customized subset.

One of the greatest needs for the future is the capacity to link, merge and interrogate multiple gene mutation databases. It is now well-accepted that neoplastic transformation of normal cells requires mutations in multiple genes to achieve inactivation of the various pathways that control cell growth, apoptosis or genome integrity. Extensive analysis of well-characterized tumours indicates that there is a non-random pattern in gene inactivation. Alteration of *BRAF* and *KRAS* are mutually exclusive in various types of cancers. A similar situation can be observed in colon cancer with mutation occurring in either *CTNNB1* (the gene that encodes β-catenin) or *APC*. This situation is easily explained by the fact that such gene pairs belong to the same biological pathways and there is usually no need for redundant inactivation.

Although some observations are obvious and can be easily detected, more subtle correlations can be revealed by more extensive database searches. It would therefore be highly desirable to link LSDBs for various genes to evaluate the 'mutation profile' of tumours. Theoretically, this should be easy

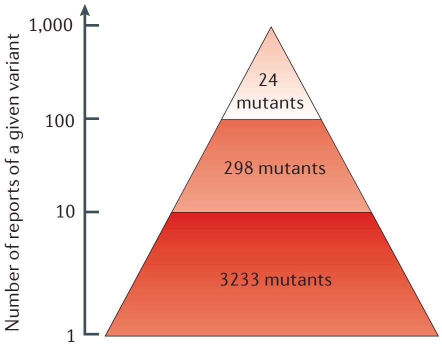


Figure 1 | **Frequency of p53 mutants in the Universal Mutation Database.** There are 3,555 different p53 mutants in the database, but their frequency is heterogeneous. 3,233 mutants are found at frequencies ranging from 1 to 10 times, with 1,874 mutants described only once. 298 mutants have been described at intermediate frequencies (between 11 and 99 times). Only 24 mutants are found more than 100 times, with the highest frequency of 979 times for the R175H mutant. Overall, the latest version of the p53 database includes 21,717 mutations.

because most clinical laboratories work on limited subsets of informative tumours that are used for multiple genetic analyses. Unfortunately, all of this information cannot be linked at the present time. Mutation analyses for various genes are published over different periods in various journals using different sample names for similar tumours, preventing cross-referencing. Only a few papers have performed multigene analysis, but most of the papers on mutations are restricted to single genes. It would be highly advisable to improve this situation to allow meta-analysis with linked LSDBs. A first step would be to ensure that authors always use the same nomenclature for sample labelling. Another step would be to define an international nomenclature for samples that could include various types of information such as country, clinical centre and unique ID. As this information is already indicated in the material section of publications, this type of nomenclature would not raise any ethical issues.

The other advantage of using a homogeneous label is that it would facilitate merging of information from databases developed on different platforms or with different generic software. Most of the fields of the various mutation databases are similar (codon position, exon, wild-type and mutant codon, mutation ID, and so on) and the nomenclature for each mutation is already well-established and many journals require the use of this name in their publications[39]. The feasiblilty of merging databases, once homogeneous standardization has been achieved, was illustrated by the work on the p53 databases performed at the EMBL European Bioinformatics Institute (EBI). Although six p53 databases are available, only two have been reproducibly updated. Despite being developed on different platforms, these two p53 databases can be easily merged by using certain fields such as tumour ID or references to identify common entries.

Access to most mutation databases has not been a crucial issue as they were developed in academic laboratories and most of them are freely available, whereas only a few require registration. However, the issue of intellectual property rights concerning biological databases has been a major concern and these issues are far from being resolved. A legal framework has been developed for the protection of databases both in Europe and in the USA, but some remaining differences and gaps have prevented perfect legislation[40,41].

A p53 mutation database

The first *TP53* mutations were published in 1989. At present, more than 2,000 publications have reported the description of p53 alterations in various neoplasms and also in other diseases, such as rheumatoid arthritis. The latest version of the UMD p53 database contains 21,717 mutations, which is approximately 30% of all mutations found in human diseases reported to date (April 2005 release). The decreasing number of reports describing *TP53* mutations since 2001 is mainly due to the difficulty of publication (BOX 2). Several thousand mutations in the *TP53* gene, and other genes, are currently unpublished and unavailable in mutation databases. To address this issue, the HGVS has initiated a project to collect all this information by means of newly developed web software called the Waystation Project. However, this software is still at the stage of testing and more volunteers are required for validation of this important project.

Bias in the p53 mutation database

In 2001 and then in 2003, several reservations were expressed concerning the biological significance of some *TP53* mutations[42,43]. First, there is a marked difference in the frequency between the various mutations, with occurrences ranging from once (401 mutants) to 979 times (mutant R175H) (FIG. 1). Structural and biological studies have

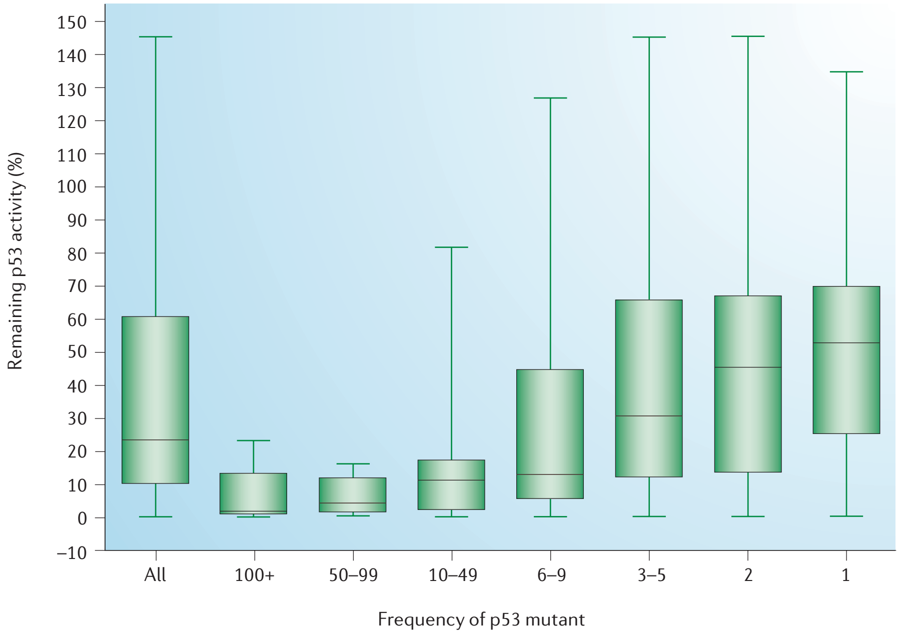


Figure 2 | **Activity of p53 mutants according to their frequency in the Universal Mutation Database.** The 3, 555 p53 mutants have been classified into 7 categories according to their frequencies in the database, ranging from more than 100 times (hot-spot mutants) to rare mutants (frequency shown on the x-axis). The y-axis corresponds to the transcriptional activity of p53 mutants, in which 100% corresponds to the activity of the wild-type protein. Box and whisker plots show the upper and lower quartiles and range (box), median value (horizontal line inside the box), and full range distribution (whisker line). Loss of p53 activity is clearly observed for hot-spot mutants, whereas rare mutants behave more heterogeneously, with a high proportion of mutants that do not show any loss of activity. This figure is an updated version of the analysis described in REF. 46.

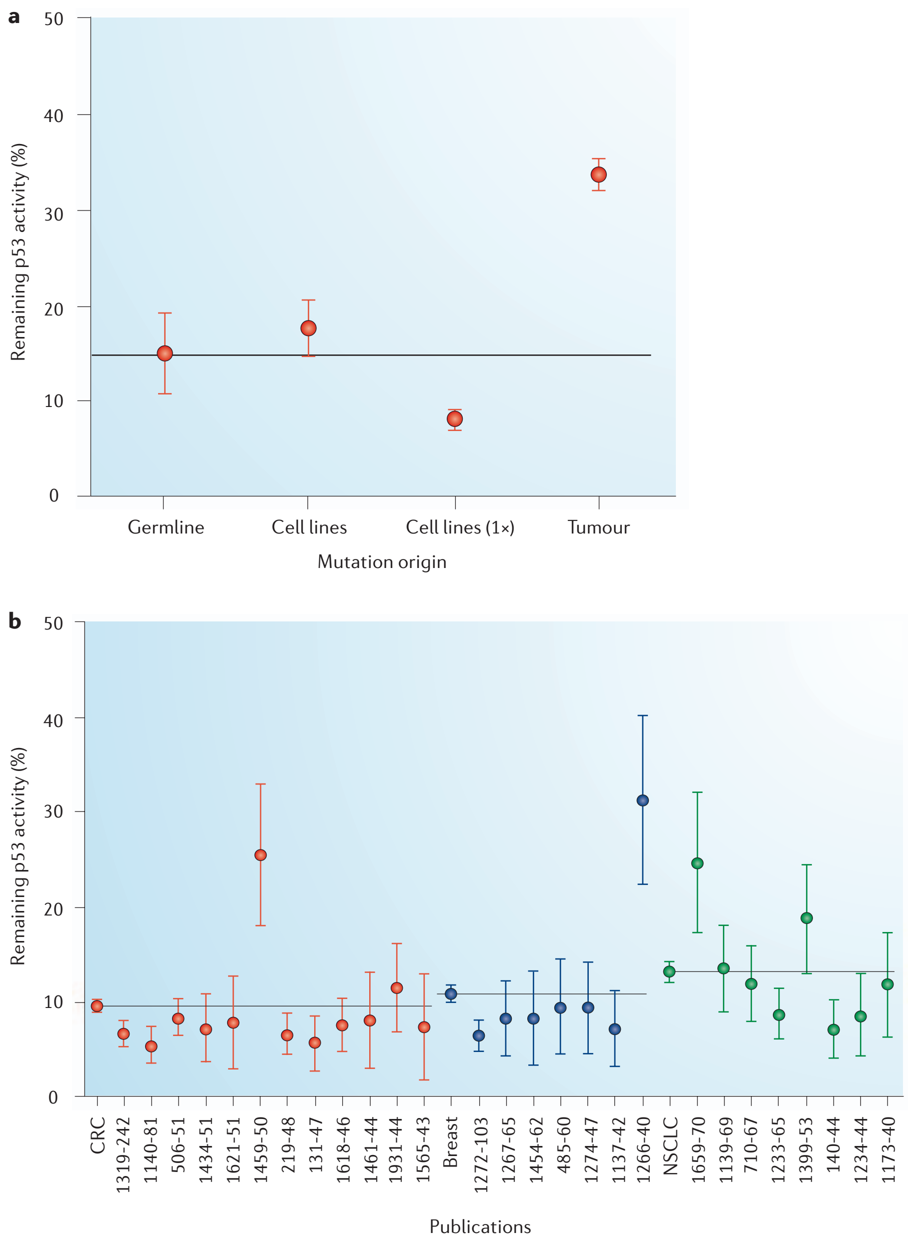


Figure 3 | **p53 loss of function. a** | Distribution of p53 loss of function. The dot and bars show the mean and 95% confidence interval (CI) of p53 activity as measured by transactivation of the *CDKN1A* (which encodes p21) promoter. The y-axis corresponds to p53 transactivation activity as described in FIG. 2, and the x-axis lists the origin of the p53 mutant. The horizontal line represents the mean of the loss of p53 function in germline mutations. Many cell lines have more than one mutation that can be either the same allele or on two different alleles. It has been previously shown that only one of the mutants shows loss of activity, whereas the second mutant is either more leaky or shows no loss of activity, indicating that only one of the mutants is important for the development of the tumour. Cell lines (1x), cell lines with only one mutation; tumour, all tumours in the database. **b** | Meta-analysis of p53 loss of function. For each cancer, the remaining activity of the p53 mutant in each publication is graphically displayed. The dot and bars show the mean and 95% CI of p53 activity, as measured by transactivation of the *CDKN1A* promoter. The horizontal line shows the mean of the combined studies. The publication code is indicated on the x-axis: the first number is an anonymous ID for the publication, and the second number indicates the number of p53 mutants included in this study. The reference value corresponding to the mean and 95% CI of all studies for the specific cancer is shown on the far left of for each cancer type. Studies are presented from left to right in decreasing order of the number of p53 mutants. The y-axis corresponds to p53 transactivation activity as described in FIG. 2. Only studies with 40 or more p53 mutations are shown on the graph. Red, colorectal cancer (CRC); blue, breast cancer; green, non-small-cell lung cancer (NSCLC). Adapted from REF. 53.

shown that mutations at hot-spot positions inactivate p53 growth-suppressive and apoptotic activities, which explains why they are found at a high frequency in human cancers. On the other hand, the significance of the rare mutants remained elusive as they were not the subject of intense scrutiny. The trend for publication of rare mutants has not decreased compared with that of other mutants (BOX 2). Second, the inclusion of reports with unusual and non-reproducible patterns of p53 mutations can alter the quality of the database. Although an unbiased database should contain all publications in the literature, some reports are notoriously dubious. Inclusion of artefactual results has a number of harmful effects, both intellectually when they are quoted indiscriminately by non-specialists, but also for the integrity of the databases. This is clearly illustrated by the debate on the origin of p53 mutations in lung cancer, challenged by the tobacco industry, as recently discussed by Bitton and colleagues[44].

The recent construction and biological analysis of 2,300 p53 mutants shed some new light on p53 mutant activity and allowed some unique analyses[45,46]. The remaining activity of mutants that are often found in the *TP53* database is usually low, ranging from 0% to 20% compared with the normal protein (FIG. 2). For rare mutants, the scatter is very heterogeneous, ranging from 0% to 160%. There is a clear inverse correlation between the frequency of p53 mutants and their activity. Of the mutants that have been found only once, approximately half have an activity greater than 50% when compared with wild-type p53, indicating that the importance, if any, of these mutations is very low. Analysis of the origin of p53 mutations indicates that several methodological biases are responsible for this observation.

TP53 germline mutations that are found in patients with LFS sustain the most drastic loss of activity (FIG. 3a). LFS is a rare autosomal-dominant syndrome in which patients are predisposed to a wide variety of cancers. Diagnosis of *TP53* germline mutations is performed by trained staff in specific laboratories using robust protocols that require analysis both at the DNA and RNA level to identify anomalies that are important for diagnosis and follow-up of these families. The situation is similar for cell lines because good quality genetic material can be analysed and mutations are usually present in 100% of cells. For tumours, the mean loss of activity in p53 mutants is lower than that observed in samples from patients with LFS or from cell lines, and there is a wide range of distribution that reflects a considerable heterogeneity in p53 mutant activity. This is because of the large number of rare mutations that show no loss of activity.

Meta-analysis of the 2,000 publications describing *TP53* mutations led to the discovery of reports that are characterized by having a high frequency of rare variants with no loss of activity[53] (FIG. 3b). Close examination of these publications showed several anomalies that have cast serious doubts on these results, such as multiple mutations in each tumour, a high frequency of 'neutral' mutations or an unusual pattern of mutations in other genes. A methodological bias has also been demonstrated, as 55% of these studies use nested PCR versus 8% of all other studies that have identified a common p53 mutant with loss of activity[53]. Although this methodology is powerful in amplifying minute amounts of DNA, it is error-prone if not carefully controlled. It is usually used for DNA extracted from paraffin-embedded tissue, which is often fragmented. Furthermore, fixation procedures can lead to chemical modification of the DNA that can generate misincorporation during PCR amplification. The use of archival tissue enables the rapid and convenient assay of a large number of tissues from specific diseases that would take years to accumulate prospectively, but it has some drawbacks. In the p53 database, data from these ambiguous reports account for 4% to 8% of all reported mutations depending on the type of cancer. p53 alterations in lung cancers from one publication (1659-70 in FIG. 3) has been a constant problem in the heated debate on the origin of mutations in smokers. Its inclusion in statistical analyses lead to serious bias and a change in the outcome of any study[47].

All these points raise the important question of how to include this information in the various databases. Two different approaches have been adopted. In the p53 database of the International Agency for Research on Cancer (IARC), all p53 mutations are included irrespective of the quality of the study. In the UMD p53 database, a manual curation eliminates ambiguous reports. In both databases, however, careful curation ensures removal of duplicate data, the other major plague of p53 mutation databases. Both approaches have advantages and disadvantages. Combining all publications ensures an unbiased database, but includes all dubious data; whereas curation, although it removes dubious data, could be dangerous because a meaningful report of an unusual p53 mutation could be discarded. The knowledge of p53 mutant activity and the meta-analysis of the p53 mutant database now provide a more objective approach to distinguish dubious studies. In the latest version of the UMD p53 database, all publications have been included, but those that differ statistically from the range of other studies have been tagged with a warning. This procedure will allow accurate analysis to be performed with greater confidence.

Guidelines for LSDB curation

We think that the problems described above are not limited to the p53 databases and extend to other databases. LSDBs are tools developed by the scientific community for the scientific community. As already indicated above, the value of these databases has now been clearly established, but it is essential to guarantee their quality. The pollution by artefactual results has a number of harmful effects, both intellectually when they are quoted indiscriminately by non-specialists, but also for analysis of the databases. Their inclusion in LSDBs can also mask other original studies describing real differences in mutation profiles. Quality control must therefore be applied at all levels. This problem concerns not only laboratories but also journals and anonymous reviewers involved in the publication of this information.

We suggest the following guidelines for database curation. All mutations should be reported, including 'neutral' mutations that do not change the amino acid as they can affect splicing or RNA stability. Polymorphisms, such as those at codons 72 or 213 for the *TP53* gene, should not be included as mutations. When mutations are unusual in terms of frequency, multiplicity or profile, confirmation by independent analyses should be performed. This concept of independent repetition has a very broad definition from one laboratory to another. The most rigorous approach consists of repeating the experiment by starting as early as possible in the process — for example, with a second DNA extraction from the sample when available. When using archival samples (paraffin-embedded tissue), a negative control should be performed with DNA extracted from the same sample using the same extraction procedure as for tumour DNA. The properties of most p53 mutants are now available on the UMD website and should be checked for validation. An excessive number of mutations with wild-type activity or rarely reported in the literature should be considered carefully. Typographical errors in mutation tables (approximately 10% to 20% of all reports of *TP53* mutations) should be avoided by carefully checking the data. New tools to generate accurate *TP53* mutation tables are now available on the p53 and UMD website. Using the official nomenclature for description of mutations is a good way to prevent ambiguities and errors[39]. Reviewers and editors must carefully check that all of these procedures have been followed. To facilitate this process, journal editors who are actively involved in the publication of *TP53* mutations should refer authors to the UMD p53 website to check their data. Although these guidelines specifically concern p53, they can obviously be applied to other genes involved in cancer.

Recently, several studies have addressed the biological significance of *BRCA1* germline mutations to identify variants that are causally linked to breast and ovarian cancer[48,49]. Inclusion of this information in the BRCA1 mutation database will be essential to improve the value of genetic counselling.

Application of these simple rules will be beneficial for the entire scientific community. Apart from ensuring the author's compliance with a rigorous scientific and technological approach, reviewers and editors must also act as gatekeepers to ensure that the quality of the information published is maintained at a level of excellence.

Thiery Soussi is at the Université P.M. Curie, 4 place Jussieu, 75005 Paris, France, and the Karolinska Institute, Department of Oncology-Pathology, Cancer Center Karolinska (CCK), SE-171 76 Stockholm, Sweden.

Chikashi Ishioka is at the Department of Clinical Oncology, Institute of Development, Aging, and Cancer, Tohoku University, Sendai 980-8575, Japan.

Mireille Claustres and Christophe Béroud are at the Laboratoire de Génétique Moléculaire et Chromosomique, Institut Universitaire de Recherche Clinique et CHU, CNRS UPR 1142, 641, avenue du Doyen Gaston Giraud, 34093 Montpellier Cedex 5, France.

Correspondence to T.S.
e-mail: thierry.soussi@free.fr

doi:10.1038/nrc1783

1. Collins, F. S. Positional cloning moves from perditional to traditional. *Nature Genet.* **9**, 347–350 (1995).
2. Cotton, R. G. Mutation detection 2001: novel technologies, developments and applications for analysis of the human genome. *Hum. Mutat.* **19**, 313–314 (2002).
3. Paalman, M. H., Cotton, R. G. & Kazazian, H. H. Jr. Variation, databases, and disease: new directions for human mutation. *Hum. Mutat.* **16**, 97–98 (2000).
4. Claustres, M., Horaitis, O., Vanevski, M. & Cotton, R. G. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res.* **12**, 680–688 (2002).
5. Montesano, R., Hainaut, P. & Wild, C. P. Hepatocellular carcinoma: from gene to public health. *J. Natl Cancer Inst.* **89**, 1844–1851 (1997).
6. Mulligan, L. M. *et al.* Germ-line mutations of the *RET* proto-oncogene in multiple endocrine neoplasia type 2A. *Nature* **363**, 458–460 (1993).
7. Hofstra, R. M. *et al.* A mutation in the *RET* proto-oncogene associated with multiple endocrine neoplasia type 2B and sporadic medullary thyroid carcinoma. *Nature* **367**, 375–376 (1994).
8. Xue, F. *et al.* Germline *RET* mutations in MEN 2A and FMTC and their detection by simple DNA diagnostic tests. *Hum. Mol. Genet.* **3**, 635–638 (1994).
9. Edery, P. *et al.* Mutations of the *RET* proto-oncogene in Hirschsprung's disease. *Nature* **367**, 378–380 (1994).
10. Romeo, G. *et al.* Point mutations affecting the tyrosine kinase domain of the *RET* proto-oncogene in Hirschsprung's disease. *Nature* **367**, 377–378 (1994).

11. Olschwang, S. *et al.* Restriction of ocular fundus lesions to a specific subgroup of APC mutations in adenomatous polyposis coli patients. *Cell* **75**, 959–968 (1993).
12. Spirio, L. *et al.* Alleles of the *APC* gene: an attenuated form of familial polyposis. *Cell* **75**, 951–957 (1993).
13. Gallou, C. *et al.* Mutations of the *VHL* gene in sporadic renal cell carcinoma: definition of a risk factor for VHL patients to develop an RCC. *Hum. Mutat.* **13**, 464–475 (1999).
14. Pearson, P. L. The genome data base (GDB) — human gene mapping repository. *Nucleic Acids Res.* **19** (Suppl.), 2237–2239 (1991).
15. Bilofsky, H. S. *et al.* The GenBank genetic sequence databank. *Nucleic Acids Res.* **14**, 1–4 (1986).
16. Hamm, G. H. & Cameron, G. N. The EMBL data library. *Nucleic Acids Res.* **14**, 5–9 (1986).
17. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19** (Suppl.), 2247–2249 (1991).
18. Merali, Z. & Giles, J. Databases in peril. *Nature* **435**, 1010–1011 (2005).
19. Horaitis, O. & Cotton, R. G. The challenge of documenting mutation across the genome: the human genome variation society approach. *Hum. Mutat.* **23**, 447–452 (2004).
20. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
21. Béroud, C., Collod-Béroud, G., Boileau, C., Soussi, T. & Junien, C. UMD (Universal Mutation Database): a generic software to build and analyze locus-specific databases. *Hum. Mutat.* **15**, 86–94 (2000).
22. Béroud, C. & Soussi, T. *p53* and *APC* gene mutations: software and databases. *Nucleic Acids Res.* **25**, 138–138 (1997).
23. Wautot, V. *et al.* 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 (2002).
24. Béroud, C. *et al.* Software and database for the analysis of mutations in the *VHL* gene. *Nucleic Acids Res.* **26**, 256–258 (1998).
25. Jeanpierre, C., Béroud, C., Niaudet, P. & Junien, C. Software and database for the analysis of mutations in the human *WT1* gene. *Nucleic Acids Res.* **26**, 271–274 (1998).
26. Collod, G., Béroud, C., Soussi, T., Junien, C. & Boileau, C. Software and database for the analysis of mutations in the human *FBN1* gene. *Nucleic Acids Res.* **24**, 137–140 (1996).
27. Varret, M. *et al.* Software and database for the analysis of mutations in the human LDL receptor gene. *Nucleic Acids Res.* **25**, 172–180 (1997).
28. Lamlum, H. *et al.* The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. *Nature Med.* **5**, 1071–1075 (1999).
29. Al-Tassan, N. *et al.* Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nature Genet.* **30**, 227–232 (2002).
30. Soussi, T. The *p53* tumour suppressor gene: a model for molecular epidemiology of human cancer. *Mol. Med. Today* **2**, 32–37 (1996).
31. Harris, C. C. *p53* tumor suppressor gene: at the crossroads of molecular carcinogenesis, molecular epidemiology, and cancer risk assessment. *Environ. Health Perspect.* **104**, 435–439 (1996).
32. Nelson, D. R. 'A variant of uncertain significance' and the proliferation of human disease gene databases. *Hum. Genomics* **2**, 70–74 (2005).
33. Rodin, S. N., Holmquist, G. P. & Rodin, A. S. CPG transition strand asymmetry and hitch-hiking mutations as measures of tumorigenic selection in shaping the p53 mutation spectrum. *Int. J. Mol. Med.* **1**, 191–199 (1998).
34. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4**, 177–183 (2004).
35. Morral, N. *et al.* The origin of the major cystic fibrosis mutation (ΔF508) in European populations. *Nature Genet.* **7**, 169–175 (1994).
36. Beroud, C. *et al.* UMD (Universal Mutation Database): 2005 update. *Hum. Mutat.* **26**, 184–191 (2005).
37. Brown, A. F. & McKie, M. A. MuStaR and other software for locus-specific mutation databases. *Hum. Mutat.* **15**, 76–85 (2000).
38. Fokkema, I. F., den Dunnen, J. T. & Taschner, P. E. LOVD: easy creation of a locus-specific sequence variation database using an 'LSDB-in-a-box' approach. *Hum. Mutat.* **26**, 63–68 (2005).
39. den Dunnen, J. T. & Antonarakis, S. E. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum. Mutat.* **15**, 7–12 (2000).
40. Maurer, S. M., Hugenholtz, P. B. & Onsrud, H. J. Intellectual property. Europe's database experiment. *Science* **294**, 789–790 (2001).
41. Greenbaum, D. & Gerstein, M. A universal legal framework as a prerequisite for database interoperability. *Nature Biotechnol.* **21**, 979–982 (2003).
42. Soussi, T. & Béroud, C. Assessing TP53 status in human tumours to evaluate clinical outcome. *Nature Rev. Cancer* **1**, 233–240 (2001).
43. Soussi, T. & Béroud, C. Significance of TP53 mutations in human cancer: a critical analysis of mutations at CpG dinucleotides. *Hum. Mutat.* **21**, 192–200 (2003).
44. Bitton, A., Neuman, M. D. & Glantz, S. A. The *p53* tumour suppressor gene and the tobacco industry: research, debate and conflict of interest. *Lancet* **365**, 1–10 (2005).
45. Kato, S. *et al.* Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc. Natl Acad. Sci. USA* **100**, 8424–8429 (2003).
46. Soussi, T., Kato, S., Levy, P. P. & Ishioka, C. Reassessment of the *TP53* mutation database in human disease by data mining with a library of *TP53* missense mutations. *Hum. Mutat.* **25**, 6–17 (2005).
47. Krawczak, M. & Cooper, D. N. p53 mutations, benzo[a]pyrene and lung cancer. *Mutagenesis* **13**, 319–320 (1998).
48. Phelan, C. M. *et al.* Classification of BRCA1 missense variants of unknown clinical significance. *J. Med. Genet.* **42**, 138–146 (2005).
49. Fleming, M. A., Potter, J. D., Ramirez, C. J., Ostrander, G. K. & Ostrander, E. A. Understanding missense mutations in the *BRCA1* gene: an evolutionary approach. *Proc. Natl Acad. Sci. USA* **100**, 1151–1156 (2003).
50. Tornaletti, S. & Pfeifer, G. P. Complete and tissue-independent methylation of CpG sites in the *p53* gene: implications for mutations in human cancers. *Oncogene* **10**, 1493–1499 (1995).
51. You, Y. H., Li, C. & Pfeifer, G. P. Involvement of 5-methylcytosine in sunlight-induced mutagenesis. *J. Mol. Biol.* **293**, 493–503 (1999).
52. Denissenko, M. F., Chen, J. X., Tang, M. S. & Pfeifer, G. P. Cytosine methylation determines hot spots of DNA damage in the human *p53* gene. *Proc. Natl Acad. Sci. USA* **94**, 3893–3898 (1997).
53. Soussi, T. *et al.* Meta-analysis of the p53 mutation database for mutant p53 biological activity reveals a methodological bias in mutation detection. *Clin. Cancer Res.* (in the press).

Acknowledgements

Development of the UMD software is supported by grants from AFM to C.B. and M.C. Work of T.S. is supported by the ARC and ligue contre le cancer (comité de Paris). We are grateful to Richard Cotton for critical reading of this manuscript.

Competing interests statement

The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:

Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

APC | *BRCA1* | *BRCA2* | *FBN1* | *NF1* | *RB1* | *RET* | *TP53* | *TTN* | *VHL*

National Cancer Institute: http://www.cancer.gov

breast cancer | colon cancer | hepatocellular carcinoma | lung cancer

OMIM: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

Marfan syndrome | von Hippel–Lindau syndrome

FURTHER INFORMATION

EMBL: http://www.ebi.ac.uk/embl/

GDB Human Genome Database: http://gdbwww.gdb.org/

GenBank: http://www.psc.edu/general/software/packages/genbank/genbank.html

Human Gene Mutation Database: http://www.hgmd.cf.ac.uk/hgmd0.html

Mutation Storage and Retrieval Program: http://www.hgu.mrc.ac.uk/Softdata/Mustar/

Leiden Open Source Variation Database: http://www.dmd.nl/LOVD/1.1.0/

MySQL: http://www.mysql.com/

p53 database of the International Agency for Research on Cancer: http://www.iarc.fr/p53/

Swiss-Prot: http://www.expasy.org/sprot/

The p53 Database: http://www.p53.free.fr

Universal Mutation Database for p53: http://www.umd.be:2072/index.shtml

Waystation Project: http://www.centralmutations.org

Access to this interactive links box is free online.

ONLINE CORRESPONDENCE ✉

Nature Reviews Cancer publishes items of correspondence online. Such contributions are published at the discretion of the Editors and can be subject to peer review. Correspondence should be no longer than 500 words with up to 15 references and should represent a scholarly attempt to comment on a specific Review or Perspective article that has been published in the journal. To view correspondence, please go to our homepage and select the link to New Correspondence, or use the URL indicated below.

The following correspondence has recently been published:

Time to harness the pro-apoptotic property of NFκB?

Radhakrishnan, S. K. and Kamalakaran, S.

http://www.nature.com/nrc/archive/correspondence.html

This correspondence relates to the article:

Nuclear factor-κB inhibitors as sensitizers to anticancer drugs

Nakanishi, C. and Toi, M.

Nature Rev. Cancer **5**, 297–309 (2005)