Data-driven unbiased curation of the *TP53* tumor suppressor gene mutation database and validation by ultradeep sequencing of human tumors

Karolina Edlund[a], Ola Larsson[b], Adam Ameur[c], Ignas Bunikis[c], Ulf Gyllensten[c], Bernard Leroy[d], Magnus Sundström[a], Patrick Micke[a], Johan Botling[a], and Thierry Soussi[b,d,1]

[a]Department of Immunology, Genetics, and Pathology, and [c]SciLifeLab Uppsala, Uppsala University, SE-751 85 Uppsala, Sweden; [b]Department of Oncology-Pathology, Cancer Center Karolinska, Karolinska Institute, SE-171 76 Stockholm, Sweden; and [d]Université Pierre et Marie Curie-Paris6, 75005 Paris, France

Edited by Carol Prives, Columbia University, New York, NY, and approved April 24, 2012 (received for review January 3, 2012)

Cancer mutation databases are expected to play central roles in personalized medicine by providing targets for drug development and biomarkers to tailor treatments to each patient. The accuracy of reported mutations is a critical issue that is commonly overlooked, which leads to mutation databases that include a sizable number of spurious mutations, either sequencing errors or passenger mutations. Here we report an analysis of the latest version of the *TP53* mutation database, including 34,453 mutations. By using several data-driven methods on multiple independent quality criteria, we obtained a quality score for each report contributing to the database. This score can now be used to filter for high-confidence mutations and reports within the database. Sequencing the entire *TP53* gene from various types of cancer using next-generation sequencing with ultradeep coverage validated our approach for curation. In summary, 9.7% of all collected studies, mostly comprising numerous tumors with multiple infrequent *TP53* mutations, should be excluded when analyzing *TP53* mutations. Thus, by combining statistical and experimental analyses, we provide a curated mutation database for *TP53* mutations and a framework for mutation database analysis.

cancer genetics | genomic | locus-specific database

Conventional sequencing using Sanger's methodology has allowed for the discovery of genetic alterations in cancer genes (1). Next-generation sequencing (NGS) techniques have expanded this knowledge by providing a more complete description of each type of alteration, including copy-number variations, translocations, and missense mutations (2, 3). The majority of these mutations are passenger mutations (or hitchhiking mutations) that have no active role in cancer progression and are only coselected with the driver mutations (4).

Since the first publication on *TP53* mutations in 1989, more than 2,700 articles have been published describing more than 35,000 *TP53* mutations in various tumor types and cell lines (5, 6). *TP53* mutation studies have applied a variety of analyses, including molecular epidemiology, clinical surveys, and structural analyses (7, 8). Such studies require highly curated *TP53* mutation data from the Locus Specific Database (LSDB) established and maintained since 1989 (9, 10).

The unique feature of *TP53* compared with other tumor-suppressor genes is its mode of inactivation. Although most tumor-suppressor genes are inactivated by mutations, leading to absence of the protein (or synthesis of a truncated product), more than 80% of *TP53* alterations are missense mutations encoding a stable full-length protein (11). Moreover, each tumor generally harbors a single mutation in the *TP53* gene that reduces the transactivation activity of the *TP53* protein, leading to loss of its antiproliferative and proapoptotic properties.

Previous studies have raised concerns about the accuracy of the various *TP53* databases, because they include all mutations published in peer-reviewed journals (12–14). Statistical analysis showed that the use of nested PCR with DNA obtained from formalin-fixed, paraffin-embedded (FFPE) tissues led to increased detection of spurious *TP53* mutations (13). Furthermore, the dogma that each tumor harbors a single *TP53* mutation has recently been challenged by several studies. In breast cancer, several reports have described a high frequency of *TP53* mutations (more than 60% compared with the general frequency of 20%) with an average of four mutations per tumor (15, 16).

To define an accurate landscape of *TP53* mutations in human cancers, we performed statistical evaluation of 34,453 published *TP53* mutations and analysis of the entire *TP53* gene in human tumors by ultradeep sequencing. The results described below provide important information that can be used in future studies on various molecular aspects of *TP53* and other cancer genes in human patients.

Results

Curating the *TP53* Database Using a Single Quality Criterion. One unique feature of the mutant *TP53* database is the accessibility of quantitative measurements of *TP53* transcriptional activity for most *TP53* mutants found in human cancer (17–19). A clear inverse correlation between the frequency of *TP53* mutants and their transcriptional activity has been observed (Fig. S1). Hot-spot *TP53* mutants sustain a significant loss of transcriptional activity, with a remaining activity ranging from 0 to 20% compared with the wild-type protein. On the other hand, half of the infrequent mutants have an activity greater than 50% compared with wild-type *TP53*, suggesting that the impact on tumor formation—if any—of these mutations is limited (Figs. S1 and S2). Such analysis performed with the latest release of the database (34,453 mutations, 2,756 publications) confirmed our previous analysis done in 2006 based on 21,000 mutations (13, 18). The present analysis was also extended to mutant *TP53* found either in cell lines or in germ line, because the latest release of the database contains sufficient entries for statistical analysis (Fig. 1*A*). Importantly, the database of *TP53* mutations in cell lines has been carefully curated to remove duplicate entries and erroneous cell lines (20). *TP53* mutants reported in germ line or cell lines are less active (the great majority being completely inactive) compared with mutants detected in tumors ($P < 0.0001$, nonparametric Mann–Whitney statistical analysis) (Fig. 1*A*). No statistical difference was observed between cell-line and germ-line *TP53* mutations regarding loss of activity.

Curating the *TP53* Database Using Data-Driven Approaches. It is possible that a significant number of *TP53* mutants observed in human tumors, particularly rare mutants with no loss of activity,

Author contributions: T.S. designed research; K.E. and O.L. performed research; U.G., B.L., M.S., P.M., and J.B. contributed new reagents/analytic tools; K.E., O.L., A.A., I.B., U.G., and T.S. analyzed data; and K.E., O.L., and T.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: thierry.soussi@ki.se.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1200019109/-/DCSupplemental.

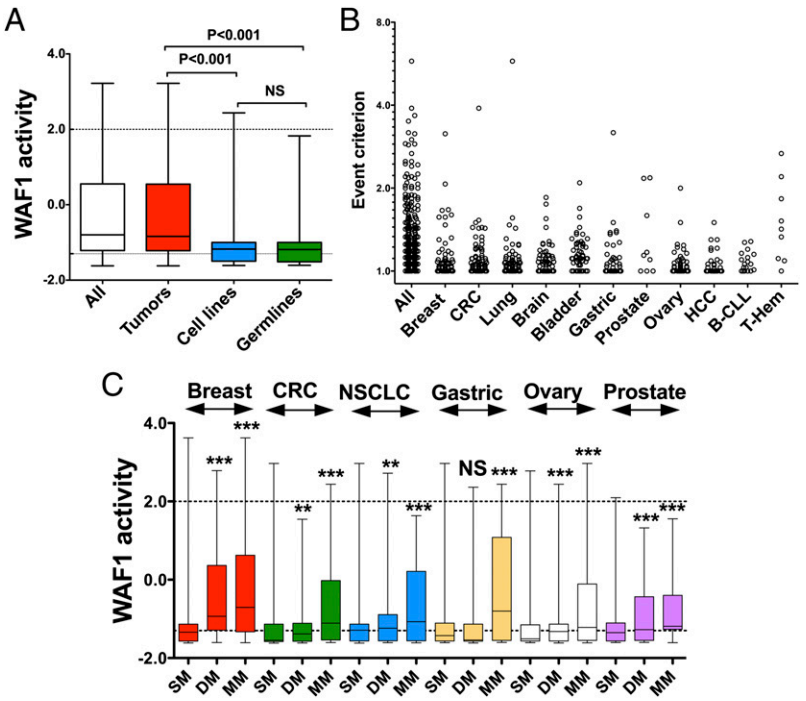

Fig. 1. *TP53* mutation heterogeneity. (*A*) Box-and-whisker analysis of mutant *TP53* activity according to origin. The *y* axis corresponds to the transcriptional activity of *TP53* mutants as reported by Kato et al., and included in the UMD *TP53* database (17, 18). Box-and-whisker plots show the upper and lower quartiles and range (box), median value (horizontal line inside the box), and full-range distribution (whisker line) for *TP53* activity. All: entire database; tumors: tumors only; cell lines: cell lines only; germline: germ line only. For germ-line mutations, the R337H mutation, very frequently found in patients with adrenocortical carcinoma in Brazil, was only added once to the database because it has been shown to be a founder mutation. The Mann–Whitney *U* test was used to evaluate statistical significance. N.S., not significant. (*B*) Distribution of the EVT criterion. The number of *TP53* mutations per tumor is very heterogeneous. The EVT criterion ranges from 1 to 5.7 with 660 publications with a value of 1 and 26 publications with a value greater than 2. This heterogeneity is not cancer-specific and can be observed in all types of neoplasia. (*C*) Activity of mutant *TP53* in tumors with only one mutation (SM), two mutations (DM), or more than two mutations (MM). The Mann–Whitney *U* test was used to evaluate statistical significance. NS, not significant; $^{**}P < 0.001$; $^{***}P < 0.0001$. A log scale was used for the *y* axis.

are passenger mutations. To address this concern, we used other independent criteria [compared with the remaining activity (ACT) that was used previously] to evaluate the quality of the publications in the database (see Table S1 and *SI Materials and Methods* for a full description of the various criteria). FREQ is related to the frequency of the mutations in the database, WT to the frequency of synonymous mutations, and EVT to the number of mutations (events) per tumor. Criteria T2, T3, and T4+ (frequency of tumors with two, three, or more than three mutations, respectively) take into account the dispersion of multiple mutations in single tumors. The last criterion, REC (mutation recurrence), is related to the frequency of each mutant in the publication and allows detection of unusual mutation hot spots (Table S1). The EVT criterion, number of mutation per tumor, is of particular interest because it has increased significantly in the past few years, with recent studies describing large series of tumors with multiple *TP53* mutations. In 50% of all publications, EVT equals one, as only one *TP53* mutation per tumor was detected (Fig. 1*B* and Fig. S3). The remaining publications report multiple *TP53* mutations per tumor, ranging from 2 to 14, leading to an average number of mutations per tumor greater than one (Fig. 1*B*). Tumors with multiple mutations are not restricted to a specific type of cancer or a specific genotype (Fig. 1*B*). Tumors with two or more mutations contain a high frequency of mutants with either partial or no loss of activity, suggesting that they are either spurious or passenger mutations (Fig. 1*C*). Tumors with two mutations could arise from a driver/passenger mutation configuration in which a single "neutral mutation" with WT activity is coselected by an inactive driving mutation. However, analysis of the loss of activity of *TP53* mutants in tumors with two mutations showed that it is a random event.

To evaluate all these criteria in a combined analysis, we used principal component analysis (PCA) (*Materials and Methods*). The first four components captured 66% of the total variance and were therefore used to calculate the number of SDs by which each sample deviated from the median. We identified 129 studies (9.7%) that deviated from the median by >2 SD (Dataset S1). We then compared the outliers' profiles across quality parameters using hierarchical clustering (Fig. 2), which, as expected, showed that outliers were identified by different patterns across the quality criteria.

To further characterize groups of studies with similar profiles across quality parameters, we used *k*-means clustering. By identifying five clusters, we detected subgroups that were biologically informative and distinct in the PCA. Clusters two and four (1,055 publications) included reports with a low value for ACT, WT, EVT (including T2, T3 and T4), and REC. These clusters differed only by the FREQ parameter that is related to the frequency of mutations. Cluster four had a higher FREQ value compared with cluster two (Fig. S4). Closer examination revealed that group four included studies of colorectal and brain cancers with a high frequency of hot-spot mutations at codons 175, 248, and 273, whereas group two included studies of cancer types, such as head and neck squamous cell carcinoma (SCC), breast carcinoma, and nonsmall-cell lung cancer (NSCLC), in which the distribution of *TP53* mutations was more heterogeneous and did not include hot-spot mutations (Fig. S5). Therefore, publications in both clusters can be considered to have a normal pattern of *TP53* mutations in accordance with the PCA, as they included only 4 of the 129 outliers. On the other hand, cluster one (17 publications) displayed outlier values for all parameters except REC. In the above PCA approach, all these studies deviated from the median by >3 SD. Clusters three and five (247 publications) included publications with infrequent mutants with partial loss of activity and a high value for REC and WT criteria (Fig. S4). These two clusters included 85% of outlier studies.

Data-Driven Quality Scores Are Independent of the ACT Criterion. In previous studies, *TP53* mutation analysis was performed using the ACT criterion only (Fig. S2) (13). We therefore repeated the PCA as described above but omitted the ACT criterion to assess to what extent the ACT criterion contributed to the analysis (Fig. S5*A*). This analysis was very similar to the PCA based on all parameters and identified 127 outlier studies, including 116 (90%) common outlier studies. The remaining 11 novel outliers were all of borderline significance in the PCA, comprising all parameters (SD greater than 1.6 but less than 2).

The overlap between the PCA approach and the previously described ACT-only approach was assessed for colorectal and breast carcinoma (13) (Fig. 3). In colorectal carcinoma, two outlier studies were identified using either PCA or the previous analysis based on the remaining activity (Fig. 3*A* and Fig. S2*A*).

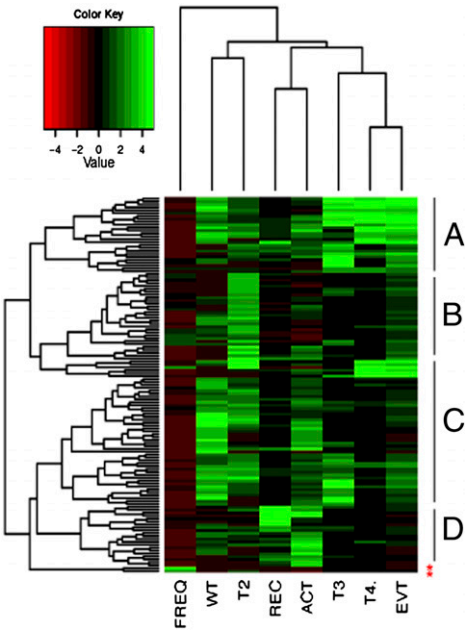

Fig. 2. Quality criteria profiles for outlier studies. Scaled data from all parameters were collected from those studies tagged as outliers (129) and used for hierarchical clustering analysis. Green indicates positive scaled values and red indicates negative scaled values. Four clusters were identified, all including publications with a large number of infrequent *TP53* mutants. Cluster A presented high values for most criteria and low values for FREQ, identifying outliers with the highest SD in the PCA. Cluster B was predominantly composed of tumors with a high frequency of two mutations (T2). Cluster C was driven by publications with an unusually large number of tumors with synonymous mutations (WT), tumors with two mutations (T2), and with high *TP53* activity (ACT). Cluster D included tumors with unusual hot-spot mutations (REC). Interestingly, two publications with a high FREQ criterion were also identified (red asterisks at the bottom of the figure), but low values were observed for the other criteria. Examination of these two publications showed that they included only mutants at hot-spot codons 175, 248, and 273. No methodological bias was observed in these reports.

This status did not change when the ACT criterion was omitted from the PCA, indicating that other criteria were abnormal in these two publications (Fig. 3*A*). Similar results were observed for other types of cancer, such as head and neck SCC or NSCLC. In breast cancer, among the eight outlier studies detected by PCA, five were previously tagged as outliers based solely on the ACT analysis (Fig. 3*B*). Within each of the three novel outlier studies, more than half of the tumors harbored multiple *TP53* mutations and numerous synonymous mutations (Fig. 3*B*, Table 1, and Fig. S2). Omitting the ACT criterion from the PCA led to reclassification of only a single outlier study (2108). Twenty percent of all breast cancer *TP53* mutations included in the database were derived from the eight outlier publications identified. These outliers did not show similar profiles across the criteria, and were therefore not defined as outliers because *TP53* mutations in breast cancer differ from those observed in other cancer types. Furthermore, these abnormal features were not observed for the remaining 50 publications of *TP53* mutations using breast cancer tissues or cell lines. Notably, two independent studies of different patients published by the same laboratory displayed identical anomalies, suggesting local technical problems (Table 1). These observations indicate that ACT is not the only important parameter in the PCA approach. Repeating PCA without the EVT criteria, another parameter that varies considerably between publications, did not induce a major shift in the results, as 125 (95%) of all publications remained outliers (SD > 2) compared with the analysis using all criteria (Fig. S5*B*). Thus, the PCA-analysis approach is robust and captures additional information compared with previous rankings based on ACT only.

***TP53* Mutations in Cell Lines, Nonneoplastic Tissue, or Germ Lines.** The genetics of tumor cell lines are very similar to primary tumors, including *TP53* status, which remains identical between a cell line and the original tumor. Molecular analysis of *TP53* alterations in cell lines is not subject to methodological bias because the obtained high DNA quality does not demand nested PCR or complicated prescreening procedures. Furthermore, for numerous cell lines, the *TP53* status has been confirmed by several independent laboratories and the *TP53* database has been curated for duplicate entries (20). The majority of all cell lines that harbor *TP53* mutations (1,492; *TP53* database 2011) have a single *TP53* mutation with low remaining activity (Fig. 1). Most of the published cell-line reports described only a few mutations (i.e., <6) and we therefore grouped these into a single set to assess their performance in the PCA-based approach described above. The distance from the median for the 1,492 cell lines in PCA components one to four was 0.39 SD, indicating that the distribution of *TP53* mutations was very similar to that of nonoutlier studies. This analysis suggests that assessing the genetic status of cell lines is a good training set for analysis of tumor genomes. Using a similar strategy for germ-line *TP53* mutations resulted in a value of 0.27 SD (322 mutations), confirming that most of these mutations are correct.

Occasional *TP53* mutations have been reported in various nonneoplastic diseases, including gastritis, liver cirrhosis, or rheumatoid arthritis. Sufficient rheumatoid arthritis cases have been published to allow analysis (62 cases, 120 mutations). PCA showed that *TP53* mutations in rheumatoid arthritis deviated from the median by >2 SD (2.47). For other nonneoplastic diseases, the number of SD was 1.7, suggesting that caution is required before drawing any definitive conclusions.

Deep-Sequencing of the *TP53* Gene Validates the *TP53* Database Curation Strategy. The *k*-means approach indicated that the number of *TP53* mutations per tumor (EVT) is markedly heterogeneous among the various *TP53* reports (Fig. S4). Although EVT was between 1 and 1.1 in the majority of publications, it was higher in a few publications describing multiple mutations in various tumors. Ultradeep sequencing using NGS could elucidate this issue, as a high coverage allows identification of low-frequency mutations. We focused our analysis on lung, colorectal, and breast tumors, because these three types of cancer are very frequent in the human population, display a high frequency of *TP53* alterations, and have already been extensively analyzed for *TP53* mutations using conventional sequencing.

One-hundred NSCLC tumors were first fully analyzed for *TP53*, KRAS, and EGFR mutations using a conventional sequencing methodology (Dataset S2). The frequency of the various genetic alterations found in this series of patients was congruent with the literature, and the pattern of mutation events for *TP53* showed a high frequency of G→T transversions compatible with smoking behavior (Fig. S6). Twenty representative lung tumors were analyzed using the SOLiD platform for *TP53* mutations in exons 2–10 at a very high coverage (mean depth 18,000) (Fig. S7). Although such a high coverage is not necessary for detection of a mutation in samples containing large amounts of tumor material, it is suitable for detection of minor clones that could contain *TP53* mutations that may not be detected by conventional sequencing. The TP73 gene, a member of the *TP53* family that is not mutated in human cancer, was analyzed as a negative control in a subset of tumors (Fig. S7). Ultradeep sequencing confirmed the *TP53* status detected by conventional DNA sequencing in all but one tumor (Dataset S2). An exonic mutation in a lung cancer specimen previously found to be wild-type was detected, but review of the results of the Sanger chromatogram indicated that this mutation was present but missed during base calling (Dataset S2). Furthermore, NGS identified two previously uncharacterized exonic mutations in exons 3 and 10 that were not analyzed by conventional sequencing (Dataset S2). No TP73 mutations were identified in this analysis,

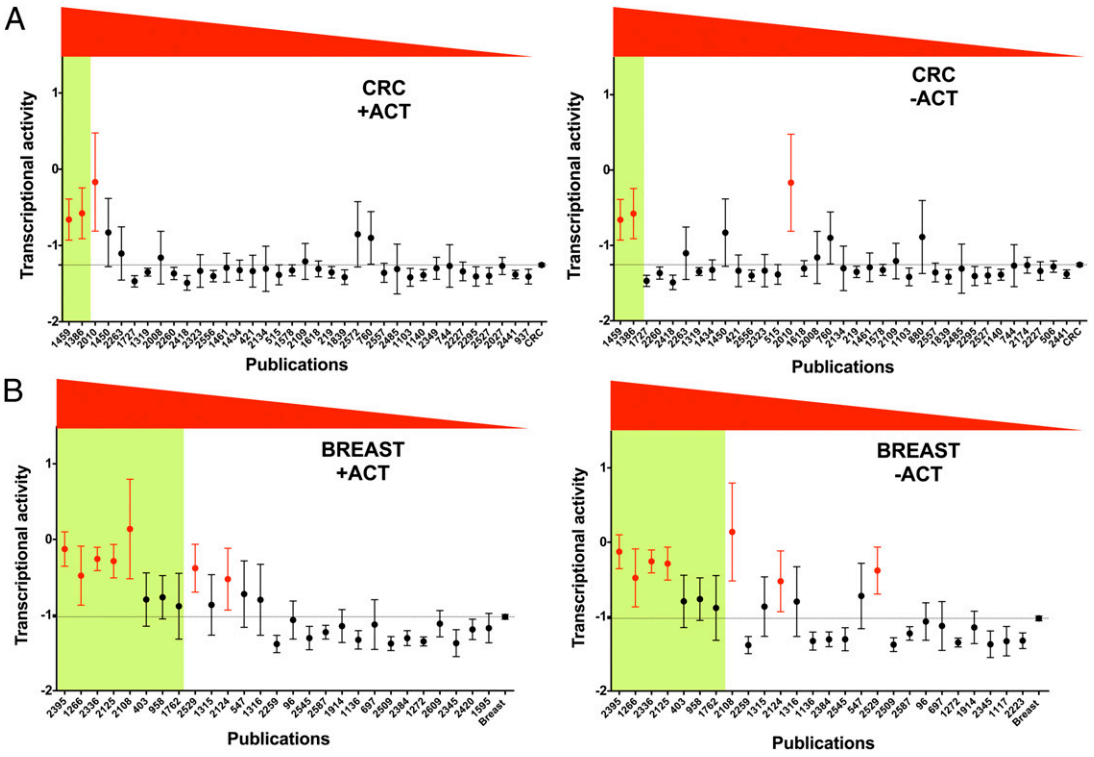

Fig. 3. Ranking *TP53* reports in colorectal and breast cancer. For each publication describing *TP53* mutations in colorectal and breast cancer, the mean (dots) and 99% confidence interval (bars) of *TP53* activity were graphically displayed. Data for all studies on colorectal (*A*) or breast cancer (*B*) are shown on the far right of the graph. The *y* axis corresponds to *TP53* transactivation activity, with a value of −1.23 for the negative control and a value of 2.03 for 100% of wild-type activity (see *SI Materials and Methods* and Fig. S2). A publication code is indicated on the *x* axis. Studies are presented from left to right in decreasing order using data from the PCA. A green box indicates outlier studies obtained by PCA, whereas studies displayed in red are outliers detected exclusively by using the ACT criterion. PCA analysis was performed using either all criteria including ACT (+ACT) or without ACT (−ACT). A change of status was observed for only one study (2018) in breast cancer. The mean activity of *TP53* mutants described in this publication is the highest for breast cancer indicating that the ACT parameter was a strong component in the analysis (distance from the median decreased from 2.2 to 1.8 SD). No changes were observed for colorectal cancer.

confirming that a stringent analysis approach was used. In 20 breast cancers, all but one exonic *TP53* mutations were confirmed and a single previously undescribed exonic mutation was detected in a tumor that was negative by conventional sequencing (Dataset S2). In colorectal cancers that were not analyzed by conventional sequencing, exonic *TP53* mutations were found in 14 of 20 tumors (Dataset S2). In breast and colorectal carcinoma, intronic mutations were found in two and six samples, respectively (Dataset S2).

Table 1. *TP53* mutations in breast cancer from outlier studies

Reference	WT	ACT	T1	T2	T3+	Detection
2395	22% (37/167)	50% (65/130)	28% (15/53)	19% (10/53)	53% (28/53)	PCA and ACT alone
1266	38% (22/58)	39% (14/36)	14% (4/28)	72% (20/28)	14% (4/28)	PCA and ACT alone
2336*	0/158	57% (90/158)	70% (87/125)	30% (38/125)	0/125	PCA and ACT alone
2125	36% (75/206)	47% (61/131)	55% (68/124)	33% (41/124)	12% (15/124)	PCA and ACT alone
2108	11% (3/26)	56% (13/23)	82% (18/22)	18% (4/22)	(0/22)	PCA and ACT alone
403	19% (7/36)	25% (6/24)	52% (12/23)	31% (7/23)	17% (4/23)	PCA only
958	18% (9/50)	39% (15/38)	68% (23/34)	26% (9/34)	6% (2/34)	PCA only
1762	11% (2/48)	40% (9/22)	76% (28/37)	19% (7/37)	5% (2/37)	PCA only
2124*	(0/33)	39% (13/33)	76% (19/25)	20% (5/25)	4% (1/25)	ACT alone[†]
2529	(0/68)	42% (29/68)	100% (68/68)	(0/68)	(0/68)	ACT alone[†]
All breast cancer	7% (229/3,173)	15.5% (483/3,119)	93% (2,544/2,793)	5% (184/2,793)	2% (68/2,793)	
485[‡]	(0/85)	7% (6/85)	100% (85/85)	0	0	
2392[‡]	(0/71)	1% (1/71)	99% (69/70)	1% (1/70)	0	
Breast cancer cell lines	0/68	2.9% (2/68)	98.5% (67/68)	1.5% (1/68)	0	

ACT, Mutant with partial or full activity vs. all mutants; T1, tumor with a single mutation vs. all tumors with p53 mutation; T2, tumor with two mutations vs. all tumors with p53 mutation; T3+, Tumor with three or more than three mutations vs. all tumors with p53 mutation; WT, synonymous mutation vs. total mutation.

*Studies 2336 and 2124 described different sets of patients, but all derived from the same laboratory.

[†]These two studies were not defined as outliers by PCA but they fall just below the 2SD mark (Fig. 3*B* and Dataset S1).

[‡]Reference studies in breast cancer. The *TP53* status from the same tumor set was analyzed using a combination of DNA sequencing (exon 2–11), cDNA sequencing, and functional assay in yeast in different laboratories.

None of these mutations were localized in splice site signals and their consequences are therefore unknown.

Defects in DNA repair can lead to an increased rate of somatic mutations and could be an explanation for the high frequency of multiple *TP53* mutations observed in tumors. Colorectal cancers used for NGS included five tumors with high microsatellite instability, indicating deficient mismatch repair. Five breast cancers had BRCA1 germ-line mutation, a gene associated with reduced capacity to repair DNA double-strand breaks. These 10 tumors with DNA repair-gene defects were not associated with an increased incidence of multiple *TP53* mutations (Dataset S2). Overall, in the 60 tumors analyzed for *TP53* mutations using deep sequencing, the number of events per tumor was 1.12, a value very close to that of 1.11 found for the whole database. These results confirm the validity of the EVT parameter in database curation and indicate that tumors with multiple *TP53* mutations are indeed very infrequent.

Discussion

Over the last 20 years, more than 33,500 *TP53* mutations have been reported in various cancer types. These mutations have been used for many analyses, ranging from molecular epidemiology, clinical stratification of patients, or structure-function studies of the *TP53* protein (6, 7, 21). All these studies depend on the accuracy of the *TP53* LSDB.

Previous studies of the *TP53* database have shown that 6% of reported *TP53* mutations conserved partial or full transcriptional activity (18, 22). Furthermore, a substantial number of tumors harboring multiple *TP53* mutations have been described. Recent studies on cancer genome sequencing have shown that neutral passenger mutations coselected by driver mutations are very frequent in human cancer. However, several observations suggest that the *TP53* mutation database does not include a substantial number of passenger mutations.

First, it has been estimated that the frequency of mutations in a tumor genome is 1.2 mutations per megabase (23). Based on the estimate that 30 Mb of the *TP53* gene has been sequenced (1 kb of DNA for each 35,000 tumors), only 35 passenger mutations would be included in the database. Second, Kato et al. showed that the *TP53* gene is likely a cold spot for passenger mutations (17). Using a library of *TP53* mutants representing all possible amino acid substitutions caused by a point mutation, they showed that the majority of *TP53* mutants for the central DNA binding domain were inactive, whether or not they were found in human cancer. This work (as well as data from other groups) demonstrated the sensitivity of the *TP53* DNA binding domain to modifications. Finally, statistical analysis showed that the majority of these active *TP53* mutants were aggregated in a small number of publications associated with the use of DNA extracted from FFPE tissue and nested PCR (13). It has now been clearly established that this type of DNA material can lead to amplification and sequencing errors if control experiments are not carefully performed (24, 25).

Analysis of a new release of the *TP53* mutation database with an additional 10,000 mutations confirmed these studies by showing a highly heterogeneous distribution of the ACT criterion (Fig. 1 and Figs. S1 and S2). Furthermore, we found that this heterogeneity was specific for tumors and was not observed in cell lines or germ-line mutations. This finding can be related to several technical issues. The genetic material used for analysis of germ-line or cell-line mutations is more homogenous than tumor specimens that can be contaminated by normal cells. Furthermore, genetic material extracted from peripheral blood lymphocytes or cell lines is of higher quality for genetic studies than tumor DNA obtained from tissue. Finally, germ-line mutation analyses are carefully controlled and performed under strict quality procedures, as they are usually linked with a familial analysis that will direct clinical decisions. As cell lines are very similar to their original tumors, they would represent a good control for the various ranking analyses.

Previously, *TP53* mutant ranking was performed with a functional assay based on the transcriptional activity of *TP53* (18). Although this assay is highly correlated with *TP53* function, it is possible that it may not capture the multiple antitumor effects of the *TP53* gene. We therefore used other criteria to evaluate the quality of the reports included in the *TP53* database to provide a curated database. These novel criteria are more objective, unrelated from any *TP53* function, and largely independent of each other. Our PCA approach led to accurate quality ranking of the various publications included in the *TP53* database. Outlier studies included publications with a high rate of tumors with multiple anomalies, such as a high frequency of tumors with multiple mutations, variants that do not change the amino acid, or infrequent mutations. Although detection of a few of these anomalies in a genomic analysis would not be a problem, accumulation of these anomalies in a single study is suspect. Ranking performed by PCA without the ACT or EVT criteria led to a very similar ranking, indicating that this methodology is robust and integrates information from multiple criteria.

Although 84% of the tumors described in the database carry only one mutation, the status of the remaining tumors were more heterogeneous, leading to a high heterogeneity of the EVT criteria. Several biological explanations can be proposed to account for this observation, including passenger mutations, a hypermutator phenotype, or heterogeneous tumors with multiple minor clones harboring different *TP53* mutations.

We therefore confirmed in silico results that deemed several studies with high EVT as outliers by sequencing the *TP53* gene of 60 tumors from different cancer types using NGS. This approach allows for detection of very rare variants with a much higher sensitivity than that of the Sanger method (mean depth 18,000). This method was applied to cancer types representing different groups with distinct mutation etiology, including tobacco exposure (lung cancer) or DNA-repair deficiency (BRCA1 mutation in breast cancer or microsatellite instability in colorectal carcinoma). The observed frequency of *TP53* mutations per tumor was identical to the entire database and no major differences were found between Sanger sequencing and NGS. To further examine the frequency of *TP53* mutations, we also analyzed the complete genome sequences published for various types of cancer using NGS methodology. Data mining for *TP53* mutations confirmed our results and showed that the majority of human tumors harbor a single *TP53* mutation.

Taken together, our multicriteria analysis showed that the *TP53* mutation database contains a nonnegligible number of publications with artifactual results (129 of 1,315). Although only publications with six or more mutations were analyzed, they corresponded to 86% of the mutations included in the database.

The present study has multiple implications, some of them far beyond the scope of the *TP53* mutation database. Previous versions of the *TP53* database included warning information solely based on the loss of *TP53* transcriptional activity. The latest release of this database will therefore include a quality score (the number of SDs from the median in the PCA) that will allow each user to select filtered content (http://p53.free.fr). Although some publications are notoriously erroneous, we believe that it is important to keep them in the database because they can be used for future ranking analysis. One of the major changes concerns breast cancer, because the eight outlier studies correspond to 20% of *TP53* mutations for breast cancer and three have a high frequency of tumors with multiple infrequent *TP53* mutation (Dataset S1). In one study, multiple discordant *TP53* mutations were detected in both tumors and stromal cells. Data presented in this publication have been highly debated and have not been reproduced by independent laboratories (26–28). This substantial refinement of the database calls for a new meta-analysis to reassess the clinical importance of *TP53* mutations in breast cancer.

Here we propose a unique approach involving multiple independent discriminating criteria and unbiased statistical analysis to detect error-prone publications. The criteria we applied are not restricted to the *TP53* database and can therefore be applied to other genes. Indeed, contamination of other mutation

databases is a well-known problem and for example, studies on EGFR or KRAS mutations found in FFPE tissue have also been highly debated (29–31).

Our results shed further light on the dark side of molecular genetics, as quoted by Kern and Winter (32). Using a scoring system based on methodology and mutation frequency, they showed that 50% of mutation reports could be erroneous, with a higher score for studies that use FFPE tissue (33).

LSDBs for cancer gene mutations have been developed using data from the published literature. With progress in cancer genomics and the development of integrated analysis of tumors, data from these LSDB will be integrated into a central database. Expert curation procedures will therefore be essential to ensure that these central databases are not contaminated by spurious information.

Materials and Methods

***TP53* Mutation Database Analysis.** The *TP53* database used in the present study contains 34,453 mutations derived from 2,756 publications (2012 R1 release at http://p53/free.fr). Selection of the various subsets used for this analysis is described in detail in *SI Materials and Methods*. PCA-analysis was performed on 1,319 publications (27,048 mutations) that described six or more mutations. Germ line and cell lines were not included in the PCA analysis initially but added separately to assess their performance (Fig. S8).

PCA Analysis. In PCA, the dimensionality of the data are reduced and components that capture nonoverlapping variance of the data are obtained. The components are ordered so that the first component captures most of the variance. Each study can then be described in terms of the difference compared with the median in a single component or a Euclidian distance in a combination of several components. All quality parameters were collected for all studies with more than five tumors. Data were scaled (within the quality parameters) and used directly as an input for PCA [using the prcomp function in R (www.r-project.org)]. The first four components were used to derive medians per component. The values for the first four components for each study were then compared with the medians across all studies using Euclidian distances. The SD for the distances obtained across all studies was calculated and the distance to the median for each study was related to the SD across all studies to obtain a measure of the outlier behavior for each study. Studies that showed a distance/SD > 2 were tagged as outliers and were compared using hierarchical clustering using the scaled data that was used as input for the PCA (values >5 were set to 5 to improve visualization in the heatmap).

ACKNOWLEDGMENTS. T.S. is supported by Cancerföreningen i Stockholm, the Swedish Cancer Society, and the Swedish Research Council (VR); O.L. is supported by the Swedish Research Council, the Swedish Cancer Foundation, the Jeansson Foundation, and the Cancer Society in Stockholm; P.M. is in part supported by the Swedish Cancer Society; and J.B. is supported by the Swedish Cancer Society and the Lions Cancer Research Fund, Uppsala. Next-generation DNA sequencing was performed at the Uppsala node of the Swedish National Infrastructure for Large-Scale DNA Sequencing, financed by the Swedish Research Council.

1. Bamford S, et al. (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* 91:355–358.
2. Forbes SA, et al. (2011) COSMIC: Mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 39(Database issue):D945–D950.
3. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 11:685–696.
4. Chanock SJ, Thomas G (2007) The devil is in the DNA. *Nat Genet* 39:283–284.
5. Soussi T (2011) Advances in carcinogenesis: A historical perspective from observational studies to tumor genome sequencing and TP53 mutation spectrum analysis. *Biochim Biophys Acta* 1816:199–208.
6. Vousden KH, Lane DP (2007) p53 in health and disease. *Nat Rev Mol Cell Biol* 8: 275–283.
7. Robles AI, Harris CC (2010) Clinical outcomes and correlates of TP53 mutations and cancer. *Cold Spring Harb Perspect Biol* 2:a001016.
8. Joerger AC, Fersht AR (2010) The tumor suppressor p53: From structures to drug discovery. *Cold Spring Harb Perspect Biol* 2:a000919.
9. Caron de Fromentel C, Soussi T (1992) TP53 tumor suppressor gene: A model for investigating human mutagenesis. *Genes Chromosomes Cancer* 4:1–15.
10. Hollstein M, Sidransky D, Vogelstein B, Harris CC (1991) p53 mutations in human cancers. *Science* 253:49–53.
11. Soussi T (2011) TP53 mutations in human cancer: Database reassessment and prospects for the next decade. *Adv Cancer Res* 110:107–139.
12. Soussi T, Ishioka C, Claustres M, Béroud C (2006) Locus-specific mutation databases: Pitfalls and good practice based on the p53 experience. *Nat Rev Cancer* 6:83–90.
13. Soussi T, et al. (2006) Meta-analysis of the p53 mutation database for mutant p53 biological activity reveals a methodologic bias in mutation detection. *Clin Cancer Res* 12:62–69.
14. Olivier M, Hollstein M, Hainaut P (2010) TP53 mutations in human cancers: Origins, consequences, and clinical use. *Cold Spring Harb Perspect Biol* 2:a001008.
15. Patocs A, et al. (2007) Breast-cancer stromal cells with TP53 mutations and nodal metastases. *N Engl J Med* 357:2543–2551.
16. Holstege H, et al. (2009) High incidence of protein-truncating TP53 mutations in BRCA1-related breast cancer. *Cancer Res* 69:3625–3633.
17. Kato S, et al. (2003) 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.
18. Soussi T, Kato S, Levy PP, Ishioka C (2005) Reassessment of the TP53 mutation database in human disease by data mining with a library of TP53 missense mutations. *Hum Mutat* 25:6–17.
19. Resnick MA, Inga A (2003) Functional mutants of the sequence-specific transcription factor p53 and implications for master genes of diversity. *Proc Natl Acad Sci USA* 100: 9934–9939.
20. Berglind H, Pawitan Y, Kato S, Ishioka C, Soussi T (2008) Analysis of p53 mutation status in human cancer cell lines: A paradigm for cell line cross-contamination. *Cancer Biol Ther* 7:699–708.
21. Joerger AC, Fersht AR (2007) Structure-function-rescue: The diverse nature of common p53 cancer mutants. *Oncogene* 26:2226–2242.
22. Hamroun D, et al. (2006) The UMD TP53 database and website: Update and revisions. *Hum Mutat* 27:14–20.
23. Greenman C, et al. (2007) Patterns of somatic mutation in human cancer genomes. *Nature* 446:153–158.
24. Williams C, et al. (1999) A high frequency of sequence alterations is due to formalin fixation of archival specimens. *Am J Pathol* 155:1467–1471.
25. Akbari M, Hansen MD, Halgunset J, Skorpen F, Krokan HE (2005) Low copy number DNA template can render polymerase chain reaction error prone in a sequence-dependent manner. *J Mol Diagn* 7:36–39.
26. Zander CS, Soussi T (2008) Breast-cancer stromal cells with TP53 mutations. *N Engl J Med* 358:1635–author reply 1636.
27. Campbell IG, Qiu W, Polyak K, Haviv I (2008) Breast-cancer stromal cells with TP53 mutations. *N Engl J Med* 358:1634–1635, author reply 1636.
28. Qiu W, et al. (2008) No evidence of clonal somatic genetic alterations in cancer-associated fibroblasts from human breast and ovarian carcinomas. *Nat Genet* 40: 650–655.
29. Gallegos Ruiz MI, et al. (2007) EGFR and K-ras mutation analysis in non-small cell lung cancer: Comparison of paraffin embedded versus frozen specimens. *Cell Oncol* 29: 257–264.
30. Marchetti A, Felicioni L, Buttitta F (2006) Assessing EGFR mutations. *N Engl J Med* 354: 526–528, author reply 526–528.
31. Lamy A, et al. (2011) Metastatic colorectal cancer KRAS genotyping in routine practice: Results and pitfalls. *Mod Pathol* 24:1090–1100.
32. Kern SE, Winter JM (2006) Elegance, silence and nonsense in the mutations literature for solid tumors. *Cancer Biol Ther* 5:349–359.
33. Winter JM, Brody JR, Kern SE (2006) Multiple-criterion evaluation of reported mutations: A proposed scoring system for the intragenic somatic mutation literature. *Cancer Biol Ther* 5:360–370.