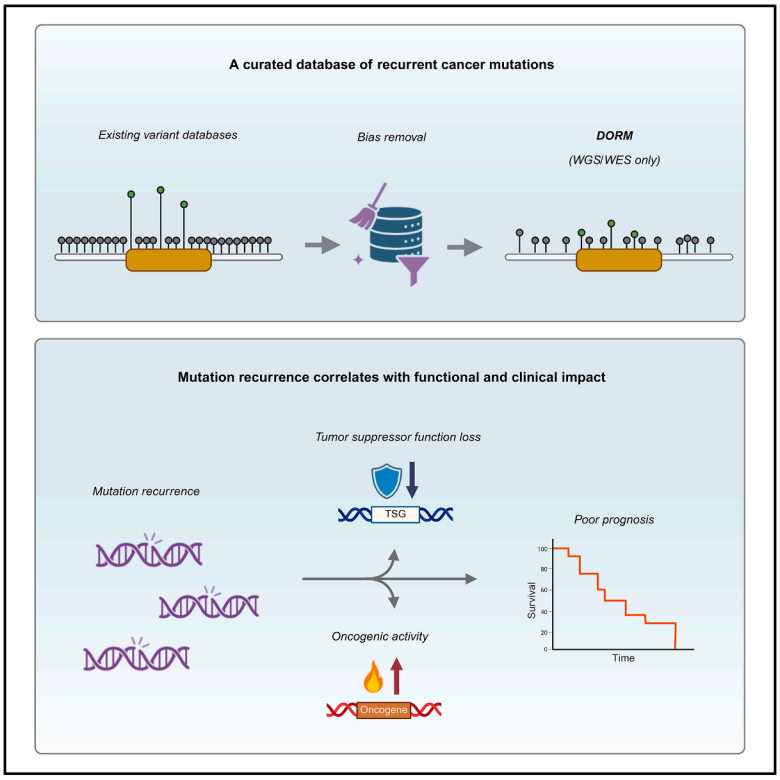
iScience

Article

Database of recurrent mutations, an unbiased web resource to browse recurrent mutations in cancers

Graphical abstract



Highlights

- DORM is a fast, open-source web resource for analyzing recurrent cancer mutations
- DORM mitigates biases introduced by targeted sequencing and duplicate samples
- Mutation recurrence correlates with oncogenicity and poor patient survival

Authors

Deepankar Chakroborty, Katri Vaparanta, Bishwa Ghimire, Ilkka Paatero, Kari J. Kurppa, Klaus Elenius

Correspondence

klaus.elenius@utu.fi

In brief

Biocomputational method; Computational bioinformatics; Cancer

Chakroborty et al., 2026, iScience *29*, 114561
February 20, 2026
https://doi.org/10.1016/j.isci.2025.114561

CellPress

iScience

CellPress
OPEN ACCESS

Article

Database of recurrent mutations, an unbiased web resource to browse recurrent mutations in cancers

Deepankar Chakroborty,[1,2,3,4] Katri Vaparanta,[1,2,5] Bishwa Ghimire,[1,5,6] Ilkka Paatero,[2] Kari J. Kurppa,[1,2] and Klaus Elenius[1,2,4,5,7,8,*]

[1]Institute of Biomedicine and Medicity Research Laboratories, University of Turku, 20520 Turku, Finland
[2]Turku Bioscience Center, University of Turku and Åbo Akademi University, 20520 Turku, Finland
[3]Turku Doctoral Programme of Molecular Medicine, 20520 Turku, Finland
[4]Research Oncology, Genentech, 1 DNA Way, South San Francisco, CA 94080, USA
[5]InFLAMES Research Flagship Center, University of Turku, 20520 Turku, Finland
[6]Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Science (HiLIFE), University of Helsinki, 00014 Helsinki, Finland
[7]Department of Oncology, Turku University Hospital, 20521 Turku, Finland
[8]Lead contact
*Correspondence: klaus.elenius@utu.fi
https://doi.org/10.1016/j.isci.2025.114561

SUMMARY

Existing cancer-associated variant databases contain biases arising from duplicate entries and the inclusion of targeted sequencing panels, which interfere with accurate estimation somatic mutation frequency in cancer cohorts. To address this, we developed the Database of Recurrent Mutations (DORM), a web resource derived exclusively from whole-genome and whole-exome sequencing data. By filtering out targeted screens and non-recurrent variants, our analysis reveals that mutation recurrence significantly correlates with oncogenic activity, loss of tumor suppressor function, and unfavorable patient prognosis. In a pan-cancer analysis of EGFR, DORM identified frequent mutations outside the kinase domain that are underrepresented in other databases. This resource offers a streamlined, unbiased platform for mutation frequency analysis, enhancing biomarker discovery and the assessment of clinical variant significance.

INTRODUCTION

The fast-paced development of next-generation sequencing (NGS) technology and its use to study cancer specimens has led to an accumulation of large quantities of data and the establishment of expansive databases that have propelled the discovery of predictive and therapeutic biomarkers for various cancers.[1] Large-scale sequencing efforts have pinned somatic mutations as the most common cause of human cancers.[2] Mutations in several oncogenes are well-characterized driver events in various cancers, e.g., mutations in KRAS G12 residue in pancreatic and lung cancer,[3] BRAF V600 in melanoma,[4] and the EGFR L858 in lung cancer.[5,6] Despite their frequent observations in the clinic, these hotspot mutations make up only a small proportion of all cancer-associated mutations and there are a large number of recurrent "non-hotspot" mutations.[7] These recurrent mutations are highly insightful for the underlying biological mechanisms of cancer. Cancer cells are under evolutionary selection pressure and the recurrence of a mutation in the population may indicate its potential to increase cancer cell fitness.[8]

Databases presenting cancer-associated mutations like COSMIC (https://cancer.sanger.ac.uk),[9] AACR GENIE (https://genie.cbioportal.org),[10] and cBioPortal (https://www.cbioportal.org/)[11,12] present a well-designed interface that provides access to rich data. However, by design, these databases with comprehensive information use a significant amount of bandwidth as well as require multiple steps to access key pieces of information, like the frequency of mutations and the affected amino acid residues. In addition, the mutation frequency information that can be retrieved from these databases without additional manual data processing is affected by several biases. One, the inclusion of information from targeted sequencing can overestimate the mutational frequency of the mutations in the regions included in the targeted sequencing panel design. Consequently, the frequency of the mutations in excluded regions is underestimated due to the increase in the number of samples. Two, the mutation frequency information in the databases can be biased by duplicate records from the same sample. This bias can arise from the inclusion of the same sample in multiple studies as well as the mapping of the mutation to several transcripts. Three, the requirement for the user to select the datasets for the mutation frequency estimation introduces sample-selection bias. Since mutation frequency estimation is continuously used for clinical decision-making as well as biomarker and oncogenic variant

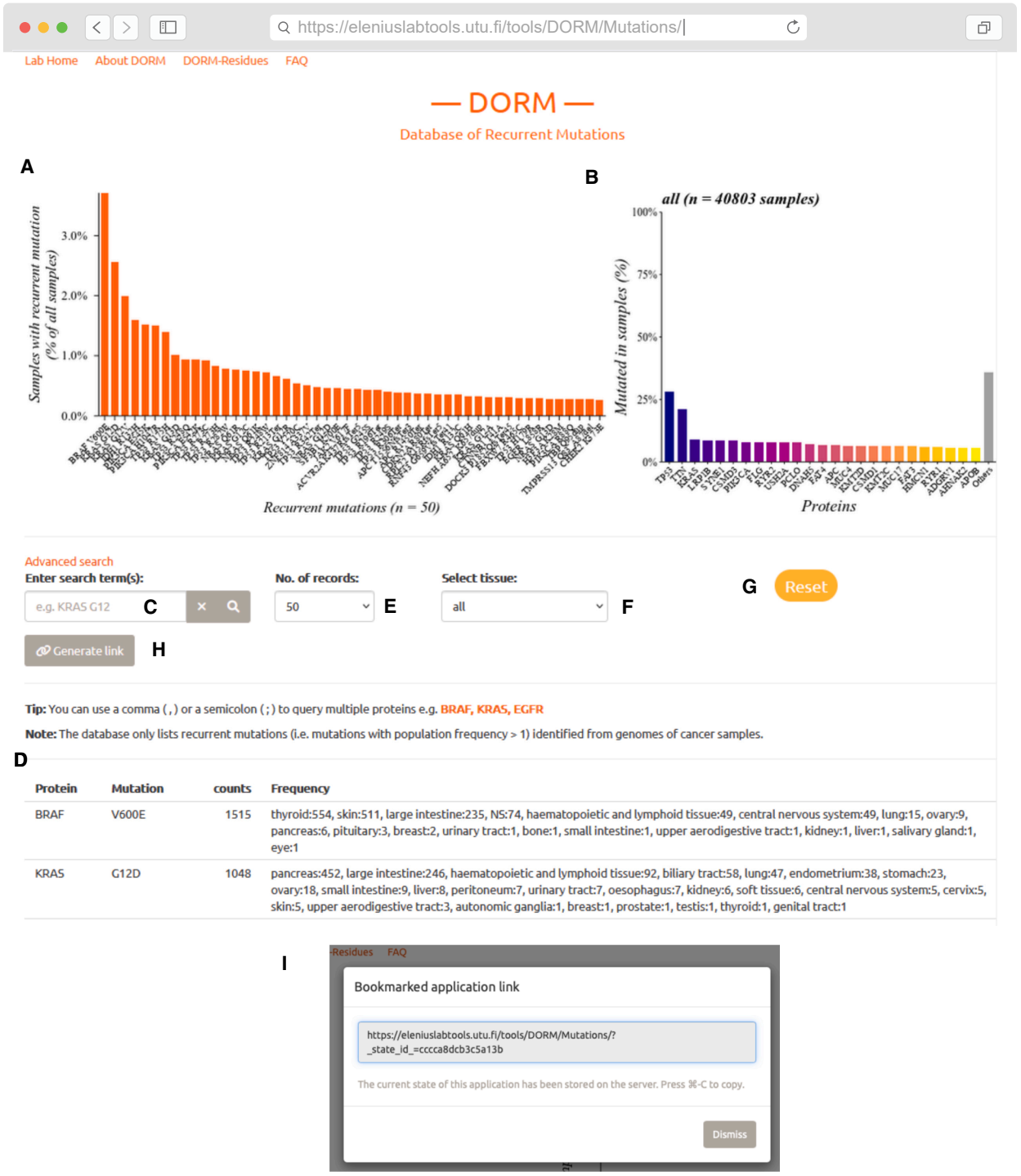


Figure 1. User interface for DORM: Database of recurrent mutations

The default GUI of DORM, hosted at https://eleniuslabtools.utu.fi/tools/DORM/Mutations/, shows information about the top 50 most-recurrent mutations identified from genomes of cancer samples.

(A) Dynamically updated bar plot that responds to search queries and settings of dropdown menus in "E" and "F."

(B) A bar chart showing the 25 most-frequently mutated genes (color gradient) across all samples in the selected tissue (which can be changed from menu "F"). The "Others" bar represents the percentage of samples not containing mutations in any of the top 25 genes (bars with color gradient).

(C) The search bar can be used to query the database with several terms, as well as, regular expressions; an example is displayed in gray text.

(legend continued on next page)

discovery, misguided directions in patient care as well as cancer research might be unknowingly selected due to a biased information set.

We sought to address these shortcomings and built a database of recurrent mutations using the large COSMIC cancer registry as a model. Our goal with this project was to develop and deploy a fast and lightweight web-resource to give a user a quick-and-easy way to check the status of a particular mutation of interest in cancer samples in an easy-to-understand format. In addition to direct time-savings, we believe initiatives like ours help further cancer research and its global outreach by improving accessibility to well-summarized and actively debiased information. Moreover, we hope that our open-source framework enables applications to other public cancer registries and diversification to other frontiers of healthcare genomics.

RESULTS

Website to browse the recurrent mutations

The DORM database was created to provide a reliable, fast, protein-centric, and user-friendly resource with reduced bias to analyze substitution mutation frequency in different cancers. The processed database is hosted on a web server at the University of Turku and can be accessed at the URL https://eleniuslabtools.utu.fi/tools/DORM/Mutations (Figure 1). DORM was developed as a tool for cancer researchers with limited bioinformatics expertise. While it can also be used by clinicians, it is intended for research purposes and lacks regulatory approval for diagnostic use. At the top of the page is a plot panel consisting of two dynamic plots that are updated in real-time in response to the user's search queries. The bar plot on the left shows the cumulative frequency of the individual recurrent mutations in the population (Figure 1A). The bar plot on the right shows the 25 most frequently mutated proteins across all the samples for the selected tissue (Figure 1B). The plot is rendered as a high-resolution image in the user's web browser following the browser's dimensions and can be saved as an image directly from the browser. Query term(s) can be entered in the search bar (Figure 1C), which updates the results in the table (Figure 1D) showing the protein, the mutation, the aggregate frequency in the population, and the frequencies categorized by the primary cancer site. The results displayed in the table can be readily copied to a spreadsheet. There is a dropdown menu (Figure 1E) adjacent to the search bar to limit the number of results displayed in the table and the plot. The search or browsing can be restricted to a particular tissue from the menu (Figure 1F). In addition to a button to reset the website and all the parameters to their default value (Figure 1G), there is a button to generate a direct link to a particular search (Figure 1H). Clicking this button opens a dialogue box (shown in Figure 1I), displaying a link that can be used to perform the same search with the exact selected parameters. Clicking this button saves the search term(s) and the set parameter(s) anonymously on our server (i.e., no identifiable information is stored). Such links facilitate the sharing of the results, and make it easy to repeat a search without re-entering the terms and/or setting the search parameters individually. The search bar (shown in Figure 1C) in the DORM database supports advanced search using regular expressions. A brief description and examples of how to query with regular expressions are provided in the supplemental guide in Document S1.

Mitigating sources of bias in the calculation of mutation frequency

To reduce bias in the mutation frequency estimation, several filtering steps were performed on the source data in COSMIC to create DORM (Figure S1). The source information of all the major cancer databases (like COSMIC, cBioPortal, and GENIE) from multiple institutions share the common problem of mutations being reported multiple times due to the same samples being included in different publications and/or studies. In addition, the source data for the COSMIC database includes individual mutations from the same sample mapped to several transcripts. These duplicate entries constituted a major portion of the source data in COSMIC (71% of all entries) and were consequently removed along with silent mutations and mutations of unknown consequence to create DORM. Since the inclusion of targeted screen data can potentially cause an over-representation or under-representation of genes and their mutations, a phenomenon that has been previously noted,[13,14] data from targeted panels and selected sequencing were not included in DORM. Finally, the non-recurrent mutations (61% of non-synonymous coding alterations) were filtered out to minimize the inclusion of mutations derived from sequencing errors.

To confirm that the mutation frequency information in DORM is less biased than in other available resources, the mutation frequency was estimated from the information in DORM and compared to the estimates from other databases. First, the effect of targeted screening data inclusion on the mutation frequency estimation was analyzed. To this end, the mutation frequency was estimated with data from the COSMIC database and DORM. The source data in COSMIC which includes data from targeted screens was processed as the data in DORM. As expected, the frequency of the well-known hotspot mutations such as JAK2 V671F, EGFR L858R, and GNAS R201 C/H was observed to be grossly overestimated when data from targeted screens was included (Figure 2A). In contrast, the mutation frequency of most mutations was underestimated due to the

(D) Table showing the protein name, mutation (displayed as amino acid change), the number of samples with that exact mutation, and the breakdown of the sample count by primary site of the cancer.
(E) Dropdown menu that changes the number of records displayed in the table "D" and plotted in the bar plot "A".
(F) Dropdown menu to limit the search to a specific tissue type.
(G) Button to reset the website and various parameters to their default values.
(H) Button to generate a direct link to repeat a search with the exact search terms and parameters. Clicking this opens the dialogue box "I" which shows the link.
(I) Dialogue box showing the direct link which can be used to conduct the exact search again without having to manually enter search term(s) and set the parameters.

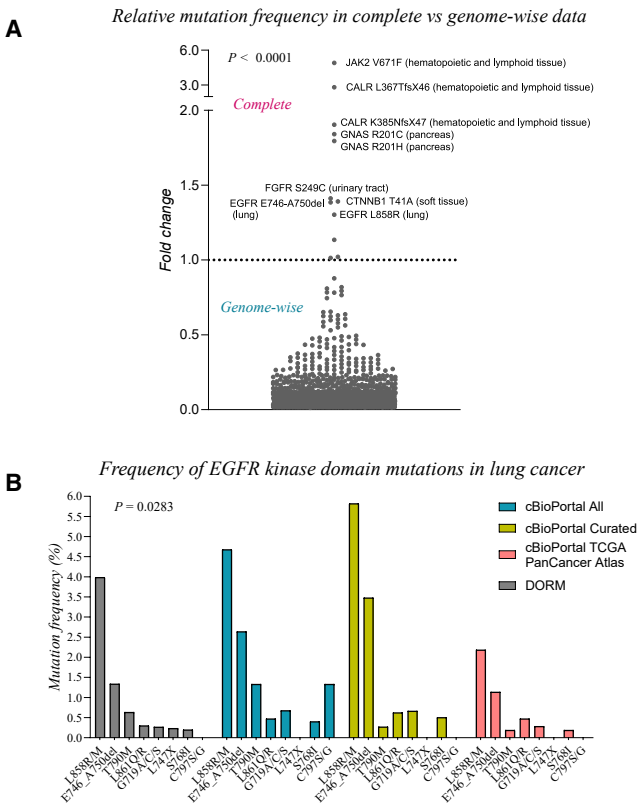


Figure 2. Inclusion of targeted sequencing data and sample selection bias skew mutation frequency estimation

(A) Fold change of the relative mutation frequency estimates of somatic mutations in complete (genome-wide and targeted) vs. genome-wide (genome-wide data only) data visualized as a scatterplot. Fold change above 1 or below 1 indicates higher frequency estimates in complete or genome-wide data, respectively. Each dot corresponds to one cancer mutation in a specific tissue. Only mutations with $n > 10$ in the genome-wide data were included in the analysis.

(B) Frequency estimates of the EGFR kinase domain mutations in lung cancer. The frequencies were estimated from subsets of studies in cBioPortal or from DORM. All: all lung cancer studies in cBioPortal included for frequency estimation. Curated: lung cancer studies included in the curated set of non-redundant studies (default setting in cBioPortal). TCGA PanCancer Atlas: lung cancer studies included in the TCGA PanCancer Atlas (default setting in cBioPortal).

inclusion of the targeted screen data. The mutation frequency of only a few mutations was similar in both cases.

Second, the effect of the bias introduced by sample selection was analyzed by estimating the mutation frequency of the most common *EGFR* kinase domain mutations in lung cancer with cBioPortal and DORM. In the user interface of cBioPortal, the user needs to define the studies to use for the mutation frequency estimation. Two default settings are provided that include either the datasets from TCGA PanCancer Atlas or a curated set of non-redundant studies. The mutation frequency was estimated from the lung cancer datasets from these default selections as well as all available lung cancer study records included in the cBioPortal database (Figure 2B). The different sample selection choices had a significant impact on the mutation frequency estimates ($p = 0.0238$). While the two most frequent mutations were consistently identified across all sample selection choices, the estimates were highly variable even for these recurrent mutations (more than a 2.5-fold difference between the lowest and highest frequency estimates; Figure 2B). These observations indicate that the mutation frequency information is highly influenced by the inclusion of targeted screen data as well as sample selection bias.

Optimizing the performance of DORM

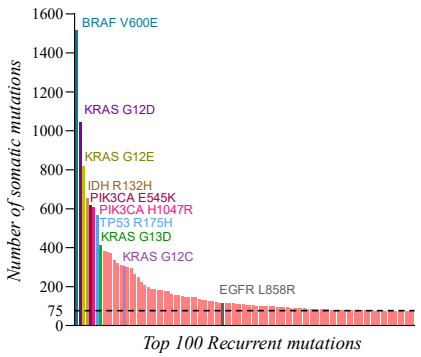
One of the primary goals of DORM was to display the desired focused statistics and results faster than contemporary databases like COSMIC and cBioPortal. Therefore, we designed DORM as a focused tool and prioritized speed over the breadth of the results. To maximize the efficiency throughout the DORM pipeline, common workflows that can be used to resolve computational bottlenecks were benchmarked (Figure S2). R functions related to reading in data, generating frequency tables, searching, search-replace actions, parallelization, and saving data were benchmarked and selected to optimize the speed of DORM (Figure S2). Google Lighthouse was additionally utilized to benchmark the performance of DORM to understand the end-user's experience (Figure S3). The Google Lighthouse performance score is the weighted mean of six individual parameters, namely, First Contentful Paint, Speed Index, Largest Contentful Paint, Time to interactive, Total Blocking Time, and Cumulative Layout Shift (details described in the STAR Methods section titled "benchmarking and testing performance"). DORM had a high Lighthouse performance score and behaved well in other Google Lighthouse metrics indicating fast and efficient performance (Figure S3).

Top recurrent mutations

The top 100 most frequent mutations across cancer types were assessed with DORM. Among the 100 most-recurrent mutations, the highest number of mutations were reported in TP53 (number of variants [n] = 21, frequency in cohort [ν] = 4,346), followed by KRAS ($n = 9$, $\nu = 3{,}287$), PIK3CA ($n = 6$, $\nu = 1{,}907$), BRAF ($n = 1$, $\nu = 1{,}515$), and NRAS ($n = 6$, $\nu = 1{,}075$) (Figure 2). Among the top 20 recurrent mutations, 12 mutations ($\nu = 7{,}220$) were in oncogenes, and 8 mutations ($\nu = 2{,}546$) in tumor suppressor genes. The top three recurrent mutations were the amino acid substitutions BRAF V600E ($\nu = 1{,}515$), KRAS G12D ($\nu = 1{,}048$), and KRAS G12V ($\nu = 817$) (Figure 3).

Mutation recurrence correlates with functional consequence

We cross-referenced the data from DORM and unique mutations from COSMIC with data from *in vitro* screens of activating mutations (iSCREAM).[15–17] Specifically, we compared whether the proportion of recurrent and unique mutations and mutated residues in *EGFR*, *ERBB3*, and *ERBB4* identified in the iSCREAM were different from the proportion of recurrent and unique mutations in all other *EGFR*, *ERBB3*, and *ERBB4* mutations and mutated residues. This comparison indicated that recurrent mutations and mutated residues were enriched in the functional screen, indicating that mutation recurrence is associated with oncogenic properties (Figures 4A and 4B). Several but not all



Rank	Gene	Mutation	Counts
1	BRAF	V600E	1515
2	KRAS	G12D	1048
3	KRAS	G12V	817
4	IDH1	R132H	653
5	PIK3CA	E545K	617
6	PIK3CA	H1047R	611
7	TP53	R175H	570
8	KRAS	G13D	413
9	TP53	R248Q	384
10	PIK3CA	E542K	380
11	TP53	R273C	375
12	TP53	R273H	336
13	TP53	R282W	319
14	NRAS	Q61R	311
15	KRAS	G12C	307
16	NRAS	Q61K	299
17	TP53	R248W	295
18	TP53	R213X	267
19	KRAS	G12R	249
20	TP53	Y220C	222

Figure 3. Distribution of the top 100 recurrent mutations

Bar plots showing the top 100 most-frequently mutated proteins in the genome-wide somatic mutation data from COSMIC release v100. The top 20 mutations are listed in the table on the right, and the mutations in oncogenes are colored in red and the mutations in tumor suppressors are colored in blue.

recurrent mutations were observed in these *in vitro* screens. This was not unexpected as it is likely that not all mutations will provide growth advantage in a simplified *in vitro* experiment. Validation of DORM-identified mutations, on the other hand, indicates that at least some recurrent mutations provide a functional growth advantage also *in vitro*, and are more likely to provide it than unique mutations.

As another validation, we compared the recurrent *TP53* mutations in DORM and unique mutations in COSMIC to information in a well-curated *TP53* database.[23,24] Indeed, predicted (Figure 4C) and experimentally verified (Figure 4D) loss of function of *TP53* was significantly more associated with recurrent mutations compared to unique mutations. These results indicate that recurrence predicts the functional relevance of a mutation.

Use case: Pan-cancer analysis of the frequency of EGFR mutations

As a use case, pan-cancer analysis of the frequency of EGFR mutations was conducted with the graphical user interfaces of DORM, COSMIC, and cBioPortal databases. To access the mutation frequency information in COSMIC the name of the gene (EGFR) was entered in the search bar, the correct transcript of the gene was selected, and the information in the "Variants" table of the "Gene" page was manually processed (Figure S4). To access the mutation frequency information in cBioPortal, the datasets used for the frequency estimation were selected, the name of the gene (EGFR) was supplied to the "Enter Genes" box after selecting the "Query by gene" option, and the "Mutations" tab was selected to view the mutation frequency information as a lollipop plot (Figure S5). To access the mutation frequency information in DORM, the name of the gene (EGFR) was supplied to the search bar, and the mutation frequency information was displayed below (Figure S6).

The extracted frequency information from the databases was supplied to the MutationMapper tool in cBioPortal to visualize the frequencies as lollipop plots (Figure 5). The L858 alterations were discovered to be the most frequent EGFR mutation in all databases (Figure 5). In both cBioPortal and COSMIC, the frequency of alterations in E746, T790, and G719 in the kinase domain of EGFR (yellow box in Figure 5) were estimated to be more frequent than the mutations in other structural regions. In DORM, however, mutations in other structural regions, such as alterations in residues L62, A289, R521, G598, and D1009 were estimated to be more frequent or similarly frequent as the alterations in E746, T790, and G719. This highlights the markedly different estimates of the frequency of EGFR mutations that can be extracted from DORM and other contemporary databases.

Mutations in DORM associate with clinical relevance

To assess the clinical utility of the mutations in DORM, we assessed whether recurrent mutations or unique mutations (e.g., the mutations not listed in DORM) would correlate with prognostic value. Existence of a recurrent mutation in at least one oncogene or tumor suppressor gene associated with significantly poorer overall survival compared to the patients who only had unique mutations in oncogenes or tumor suppressor genes in two separate pan-cancer cohorts (Figures 6A and 6B). Recurrent mutations in a randomized set of genes, however, did not have similar effects. This suggests that recurrent mutations in cancer-relevant genes only affect patient survival. To additionally address the potential clinical utility of DORM, we analyzed the relevance of recurrence in the relation of patient prognosis using the information from the curated *TP53* database (Figure 6C). The recurrent *TP53* mutations were significantly more associated with prognostic value than unique mutations, indicating prognostic relevance for mutation recurrence. Taken together, these results indicate that analysis of mutation recurrence provides useful information at the patient level.

DISCUSSION

NGS of cancer sample series has enabled accurate understanding of cancer biology and helped identify new predictive and therapeutic biomarkers. Here, we present DORM, a fast, less biased and focused (Table 1) web tool, that allows browsing its database derived from an analysis of somatic substitution mutations identified by whole-genome or whole-exome NGS. This strategy avoids the biases introduced due to the use of targeted sequencing panels, inclusion of duplicate entries, and sample selection. Indeed, the mutation frequency estimates of DORM were consistently discovered to differ from estimates derived from other databases that include less filtered data.

DORM only includes recurrent mutations. The biological relevance of recurrent mutations remains to be elucidated and will represent an interesting avenue for scientific research for many years to come. The clearest evidence for relevance is provided by so called "hot-spot" mutations (= highly recurrent mutations), for which there is extensive and long-term evidence for biological significance. For example, KRAS G12[28] recurrent mutations were identified already in 1980s and have been the subject of extensive research, confirming their functional relevance. Similar evidence for recurrent mutations is available for many other classical oncogenes such as BRAF[29] and tumor suppressors such as TP53.[30] In our analyses, mutation recurrence associated with oncogenic properties, loss of function of tumor suppressor genes and prognostic value. Theoretically, these mutations recur because they confer a selective advantage, enhancing the

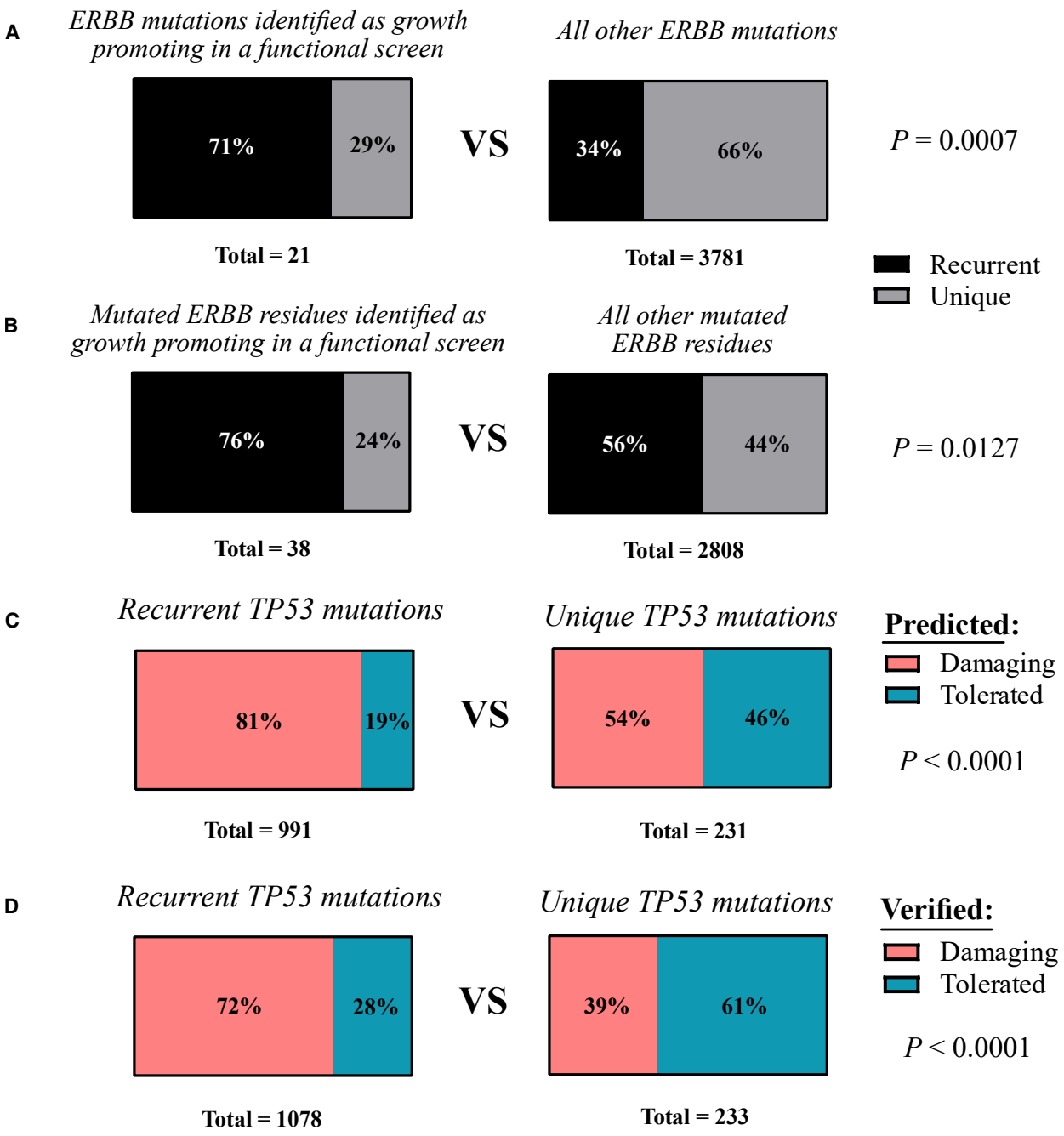


Figure 4. Recurrence is associated with oncogenicity and the loss of function of tumor suppressor genes

Recurrent mutations in DORM and unique mutations in COSMIC release v100 were cross-referenced with results from functional screens[18–20] and annotations from curated TP53 database.[21,22] (A and B) The proportions of recurrent and unique ERBB mutations (A) or mutated residues (B) were compared between those identified in functional screens and those not identified. (C and D) The proportions of predicted (C) or experimentally verified (D) damaging and tolerated TP53 mutations were compared between recurrent and unique TP53 mutations. Fisher's exact test was used for statistical testing.

evolutionary fitness of the cancer cells within their specific microenvironment niche.[31] In addition, some changes may be biochemically more prone to mutation[18] but without a selection advantage these are less likely to be enriched. In summary, the recurrence of a mutation is an interesting signal and may indicate that the mutation may improve cancer cells fitness. While it is possible that similar mutations occur by chance, the likelihood of that decreases as the frequency of re-occurrence for the specific mutation increases. Frequency estimates may be also confounded by the existence of germ-line mutations. While

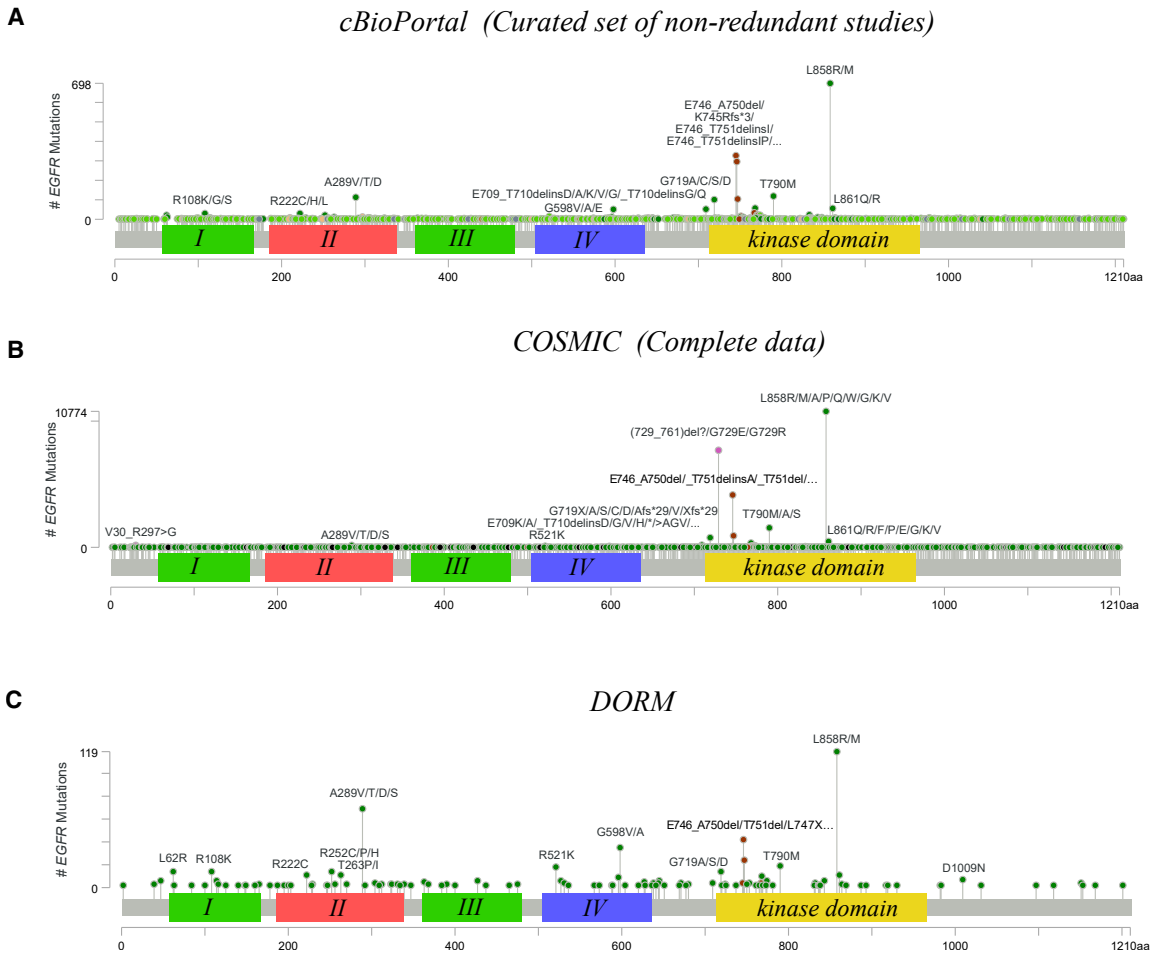


Figure 5. Lollipop plots of the pan-cancer analysis of the frequency of EGFR mutations

The frequency of EGFR mutations across cancer types was estimated with the information extracted from cBioPortal (A), COSMIC (B) and DORM (C) databases and visualized as lollipop plots.

recurrence of *de novo* germ-line mutations would be of importance, inherited mutations may skew the frequency estimates of somatic mutations. Although the COSMIC database used here accounts for mutation's somatic status, utilizing additional filters in DORM for germ-line mutations could be useful.

A pan-cancer analysis of *EGFR* mutation frequencies was performed using DORM and other public databases. While frequency estimates from other databases identified the kinase domain of *EGFR* as a hotspot for recurrent mutations, analysis using DORM revealed that additional structural domains also harbor mutations with comparable recurrence rates. If these findings were applied to the design of a targeted sequencing panel, these data would support screening of the entire *EGFR* coding region to capture all frequently occurring oncogenic mutations. In contrast, reliance on other databases would limit screening to the kinase domain, potentially missing patients who could benefit from EGFR-targeted therapies. Indeed, DORM identifies several non-kinase domain *EGFR* mutations in lung cancer that have clinical implications identified in patients. These include three mutations that correlate with response to EGFR-targeted therapy (two in NSCLC and one in glioblastoma),[19–21] one associated with anti-EGFR therapy resistance (in colorectal cancer)[22] and two identified as biomarkers by the FDA.[25] A classic targeted sequencing screening panel for lung cancer only sequences the *EGFR* exons 18–21 that constitute the kinase domain. This indicates that certain clinically relevant *EGFR* variants are not captured by these targeted sequencing panels. Targeted sequencing panels, however, have been found especially useful in the clinic for assessing tumor mutational burden in low-purity samples.[26]

In addition to enhanced performance and speed DORM offers several advantages (Tables 1 and 2), most notably the ability to directly search for sets of proteins and use regular expressions. Additionally, DORM is the only database that can summarize mutations at the level of amino acid residues (accessible via: https://eleniuslabtools.utu.fi/tools/DORM/Residues/). Data from all other databases require manual processing to retrieve this information. DORM is also the only database that allows the user to view a large amount of data without having to click through numerous pages of results (on DORM, users can choose from a range of 10–10,000

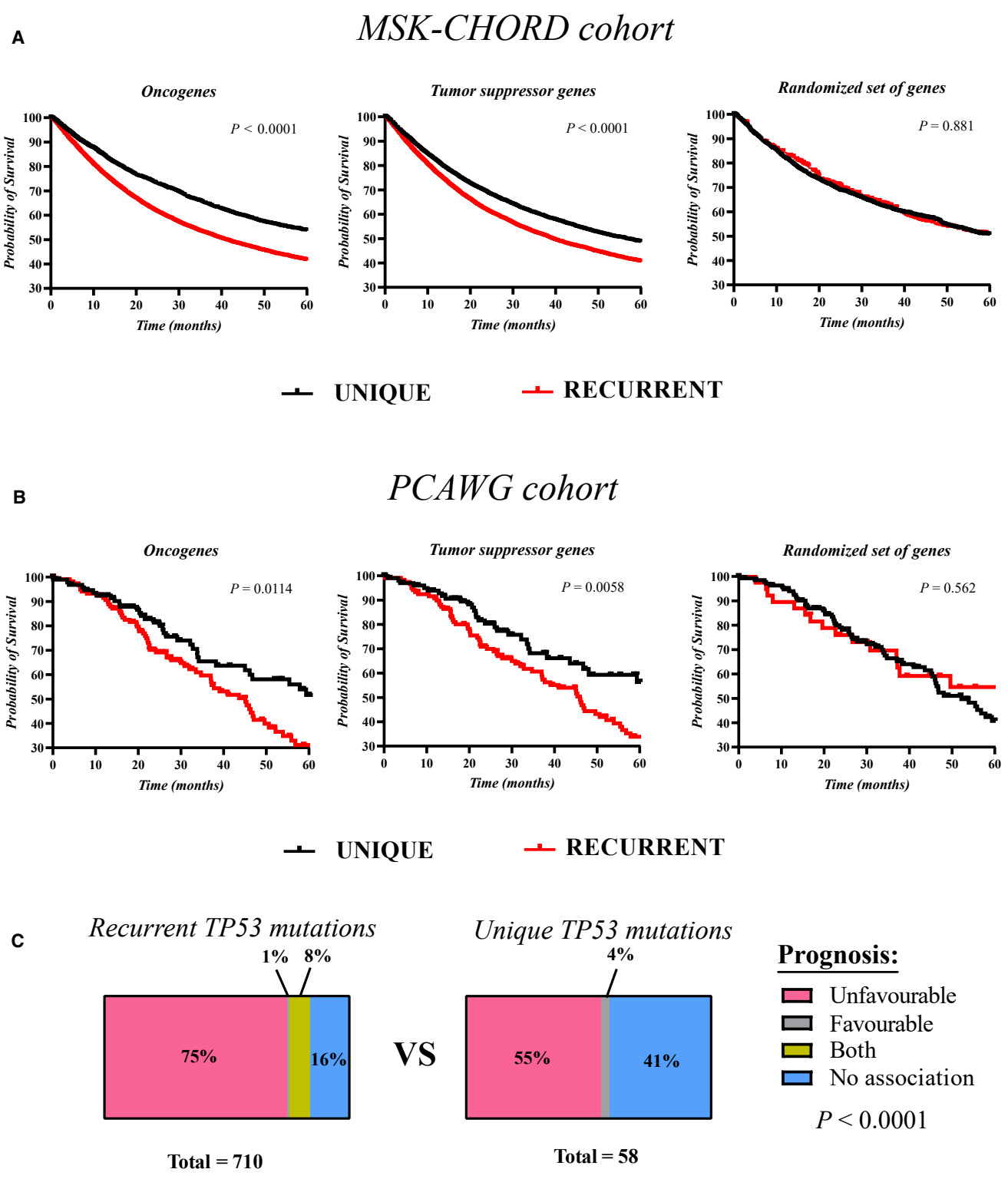


Figure 6. Recurrence of cancer gene mutations is associated with prognostic value

(A and B) The overall survival of cancer patients harboring either only unique mutations or at least one recurrent mutation in oncogenes, tumor suppressor genes, or in a randomized set of genes was estimated from the MSK-CHORD cohort[25] (A) and the PCAWG cohort[26] (B). The Cancer Gene Census resource[27] was used

(legend continued on next page)

Table 1. Comparison of search and querying features between DORM and other public databases

Searching and querying	DORM	COSMIC	cBioportal	AACR Genie
Protein	yes	yes	yes	yes
Individual mutations (e.g., KRAS G12C)	yes	yes	yes[a]	yes[a]
Protein sets	yes	no	yes	yes
Tissues	yes	yes	yes	no
Regular expression	yes	no	no	no
Substring search (searching for RAS shows HRAS, KRAS, NRAS, etc.)	yes	yes	no	no

[a]cBioPortal and Genie require searching for the gene and then the mutation.

Table 2. Comparison of additional features between DORM and other public databases

Additional features	DORM	COSMIC	cBioPortal	AACR Genie
Direct link to save and share search results	yes	yes	yes	yes
Summarize by residue	yes	no	no[a]	no
Duplicate samples	no	yes	yes	yes
Display most recurrent mutations for a tissue or protein set	yes	no	no	no
Show frequency of a protein being mutated in various tissues	yes	yes[b]	yes	no
Number of rows displayed in table	10–10,000	10–100	25	25
Free from cookies, trackers and/or analytics	yes	no	no	no

[a]cBioPortal lollipop occasionally groups hotspot mutations at a residue as a single lollipop.
[b]Possible on COSMIC Cancer Browser.

results to display). DORM is also the only database that is free of cookies, trackers, and any embedded analytics.

While COSMIC, cBioPortal and AACR Genie feature duplicate entries, DORM does not. Individual mutations, such as *KRAS* G12C, can be directly searched on DORM as well as COSMIC. On cBioPortal, the implementation of tissue-specific search filters is similar to DORM (i.e., requires selection from a menu), whereas COSMIC requires users to select the tissue from a table in the "tissue distribution" section.

DORM is lightweight, and can be run on standard consumer hardware using our open-source codebase (see STAR Methods for links to the repositories). DORM is publicly available on a virtual private server allowing resources to scale with an increase in demand. We believe that DORM improves the accessibility of important information regarding recurrent mutations by being faster and more resource-efficient than the competition.

While originally designed for cancer researchers, our discussions with oncologists at Turku University Hospital confirmed its potential value in clinical decision-making. DORM's strengths as a clear, fast, and reliable tool were highlighted particularly in tumor board discussions to identify tumor origin, confirm variant-diagnosis consistency, and assess the functional consequences of rare mutations. Clinicians also noted that reliable updates, error handling, and the inclusion of functional, cancer subtype-specific, and survival data would be essential for broader clinical integration. With these improvements, DORM could complement existing tumor board resources, while in its current form it provides significant value for research, education, and variant annotation in clinical discussions.

Limitations of the study

In the pursuit of speed and performance, certain trade-offs were made that constitute the limitations of DORM (Table 3). For instance, DORM does not incorporate or display the information about copy number variations or structural variations and excludes all the detailed sample- and study-level information. Instead, DORM is protein-centric and focused specifically on substitutions. Consistent with several other databases in our comparison, DORM does not display gene fusions or non-coding mutations, nor does it allow selecting multiple tissues, or display "lollipop" diagrams which is a visualization tool that places the mutations within the context of the protein's primary sequence.

Tumor heterogeneity, variability in tumor purity,[27] and the presence of subclonal mutations can lead to underestimation of mutation frequencies, typically in samples with low relative cancer cell content. This inherent bias may affect the frequency estimates in DORM. However, since tumors are always composed of diverse cell types in addition to malignant cells, this issue is not merely an artifact but an informative biological feature of tumors.[32] Both tumor composition/purity and subclonal mutations are important topics and could be best addressed using single-cell genomics.[33] This represents a promising avenue for future research as single-cell data are accumulating. The frequency estimates in DORM may be refined as single-cell genomic data achieve sufficient sequencing depth and scale.

As the absolute "ground-truth" of the mutations observed in these large-scale sequencing datasets is unknown, the exact rate of bioinformatic artifacts cannot be estimated. Such artifacts can arise from data processing of including raw data, base calling, quality filtering, and database transfer.[34–36] Since all modern sequencing data are subject to numerous

as a reference for a list of oncogenes and tumor suppressor genes. The randomized set of genes was generated by random sampling. Mantel-Cox test was used for statistical testing.

(C) Recurrent mutations in DORM and unique mutations in COSMIC release v100 were cross-referenced with annotations from the *TP53* database. The proportions of *TP53* mutations associated with either unfavorable (poorer survival, resistance to treatment, or recurrence), favorable or both prognoses as well as non-prognostic mutations were compared between the recurrent and unique mutations. Fisher's exact test was used for statistical testing.

Table 3. Limitations of DORM in comparison to other public databases presenting somatic mutations identified from cancer samples

Limitations of DORM	DORM	COSMIC	cBioportal	AACR Genie
Copy number variations and structural variations	no	yes	yes	no
Show Lollipop diagram for locating mutations on peptide	no[a]	no	yes	no
Show detailed information (sample and study level)	no	yes	yes	yes
Non-coding mutations	no	yes	yo	no
Fusions	no	yes	yes	yes
Select multiple tissues	no	yes[b]	yes	no
Data download	no	yes	yes	no

[a]DORM shows the distribution of mutations in a single protein in different tissues with a pie chart.
[b]Possible on COSMIC Cancer Browser.

bioinformatic procedures and may therefore contain bioinformatics artifacts, reaching absolute certainty regarding the absence of artifacts is theoretically and epistemologically impossible. Consequently, it is not possible to fully ascertain that all mutation counts in DORM (reflecting the number of times a variant is observed) represent true positives rather than bioinformatic artifacts. However, because DORM lists only recurrent mutations, the probability that a specific mutation entry within the database is a false positive due to recurring data artifacts in the exact same genetic region, remains infinitesimally small.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and data should be directed to and will be fulfilled by the lead contact, Professor Klaus Elenius (klaus.elenius@utu.fi).

Materials availability

This study did not generate any new materials.

Data and code availability

The code for creating and validating DORM has been deposited to GitHub and is publicly available through the following GitHub Repository URLs: https://github.com/KE-group/DORM-2022; https://github.com/dchakro/DORM_Mutations; https://github.com/dchakro/DORM_Residues; https://gist.githubusercontent.com/dchakro/8b1e97ba68563dd0bb5b7be2317692/raw/parallelRDS.R. The source data utilized to create DORM is available on the COSMIC[9] downloads page (https://cancer.sanger.ac.uk/cosmic/download/cosmic) and NCBI's RefSeq database[37] (https://www.ncbi.nlm.nih.gov/refseq/MANE/).

ACKNOWLEDGMENTS

The Cancer Foundation Finland, Novo Nordisk Foundation, Research Council of Finland, Sigrid Juselius Foundation, and Turku University Central Hospital are acknowledged for financial support. The authors wish to thank oncologists Dr. Erika Alanne and Dr. Maria Sundvall (both from Turku University Central Hospital, Finland) for valuable comments on the clinical applicability of DORM.

AUTHOR CONTRIBUTIONS

Conceptualization, D.C., K.E., I.P., and K.J.K.; methodology, D.C. and K.V.; formal analysis, D.C. and K.V.; funding acquisition, K.E.; investigation, D.C. and K.V.; software, D.C., K.V., and B.G.; data curation, D.C. and K.V.; supervision, K.J.K. and K.E.; visualization, D.C. and K.V.; writing – original draft, D.C., K.V., and K.E.; writing – review and editing, D.C., K.V., K.E., and I.P.

DECLARATION OF INTERESTS

K.E. declares research agreements with Boehringer Ingelheim and Puma Biotechnology and ownership in Abomics, Novo Nordisk, Orion, Roche, and Vertex Pharmaceuticals outside of the submitted work. D.C. declares current employment with Roche and Genentech.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

Portions of the article text were edited with the assistance of OpenAI's ChatGPT-4 and ChatGPT-5 to improve clarity. The authors reviewed and revised all generated content as needed and take full responsibility for the final text.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Website and web server
  - Hardware
  - Data and processing of data
  - Benchmarking and testing performance
  - Mutation frequency estimation
  - Comparative evaluation of recurrent and unique mutation effects
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.114561.

Received: March 9, 2025
Revised: October 5, 2025
Accepted: December 23, 2025
Published: December 29, 2025

REFERENCES

1. Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93. https://doi.org/10.1038/s41586-020-1969-6.
2. Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science *349*, 1483–1489. https://doi.org/10.1126/science.aab4082.
3. Hong, D.S., Fakih, M.G., Strickler, J.H., Desai, J., Durm, G.A., Shapiro, G.I., Falchook, G.S., Price, T.J., Sacher, A., Denlinger, C.S., et al. (2020). KRAS G12C Inhibition with Sotorasib in Advanced Solid Tumors.

N. Engl. J. Med. *383*, 1207–1217. https://doi.org/10.1056/nejmoa1917239.

4. Hauschild, A., Grob, J.J., Demidov, L.V., Jouary, T., Gutzmer, R., Millward, M., Rutkowski, P., Blank, C.U., Miller, W.H., Kaempgen, E., et al. (2012). Dabrafenib in BRAF-mutated metastatic melanoma: A multicentre, open-label, phase 3 randomised controlled trial. Lancet *380*, 358–365. https://doi.org/10.1016/S0140-6736(12)60868-X.
5. Lynch, T.J., Bell, D.W., Sordella, R., Gurubhagavatula, S., Okimoto, R.A., Brannigan, B.W., Harris, P.L., Haserlat, S.M., Supko, J.G., Haluska, F.G., et al. (2004). Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. N. Engl. J. Med. *350*, 2129–2139. https://doi.org/10.1056/nejmoa040938.
6. Paez, J.G., Jänne, P.A., Lee, J.C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F.J., Lindeman, N., Boggon, T.J., et al. (2004). EGFR mutations in lung, cancer: Correlation with clinical response to gefitinib therapy. Science *304*, 1497–1500. https://doi.org/10.1126/science.1099314.
7. Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. *34*, 155–163. https://doi.org/10.1038/nbt.3391.
8. Stobbe, M.D., Thun, G.A., Diéguez-Docampo, A., Oliva, M., Whalley, J.P., Raineri, E., and Gut, I.G. (2019). Recurrent somatic mutations reveal new insights into consequences of mutagenic processes in cancer. PLoS Comput. Biol. *15*, e1007496. https://doi.org/10.1371/journal.pcbi.1007496.
9. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. *47*, D941–D947. https://doi.org/10.1093/nar/gky1015.
10. Sweeney, S.M., Cerami, E., Baras, A., Pugh, T.J., Schultz, N., Stricker, T., Lindsay, J., Del Vecchio Fitz, C., Kumari, P., Micheel, C., et al. (2017). AACR project genie: Powering precision medicine through an international consortium. Cancer Discov. *7*, 818–831. https://doi.org/10.1158/2159-8290.CD-17-0151.
11. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095.
12. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1. https://doi.org/10.1126/scisignal.2004088.
13. Fang, H., Bertl, J., Zhu, X., Lam, T.C., Wu, S., Shih, D.J.H., and Wong, J.W.H. (2023). Tumour mutational burden is overestimated by target cancer gene panels. J. Natl. Cancer Cent. *3*, 56–64. https://doi.org/10.1016/J.JNCC.2022.10.004.
14. Chen, X., Listman, J.B., Slack, F.J., Gelernter, J., and Zhao, H. (2012). Biases and Errors on Allele Frequency Estimation and Disease Association Tests of Next Generation Sequencing of Pooled Samples. Genet. Epidemiol. *36*, 549–560. https://doi.org/10.1002/GEPI.21648.
15. Chakroborty, D., Kurppa, K.J., Paatero, I., Ojala, V.K., Koivu, M., Tamirat, M.Z., Koivunen, J.P., Jänne, P.A., Johnson, M.S., Elo, L.L., and Elenius, K. (2019). An unbiased in vitro screen for activating epidermal growth factor receptor mutations. J. Biol. Chem. *294*, 9377–9389. https://doi.org/10.1074/jbc.RA118.006336.
16. Chakroborty, D., Ojala, V.K., Knittle, A.M., Drexler, J., Tamirat, M.Z., Ruzicka, R., Bosch, K., Woertl, J., Schmittner, S., Elo, L.L., et al. (2022). An Unbiased Functional Genetics Screen Identifies Rare Activating ERBB4 Mutations. Cancer Res. Commun. *2*, 10–27. https://doi.org/10.1158/2767-9764.CRC-21-0021.
17. Koivu, M.K.A., Chakroborty, D., Airenne, T.T., Johnson, M.S., Kurppa, K.J., and Elenius, K. (2024). Trans-activating mutations of the pseudokinase ERBB3. Oncogene *43*, 22–2265. https://doi.org/10.1038/S41388-024-03070-9.
18. Brash, D.E. (2015). UV signature mutations. Photochem. Photobiol. *91*, 15–26. https://doi.org/10.1111/PHP.12377.
19. Shen, C.-I., Chang, J.-C., Jain, S., Olsen, S., and Wu, C.-E. (2024). Afatinib Plus Bevacizumab Treatment for a Patient With EGFR S645C–Mutant Non–Small Cell Lung Cancer: A Case Report. JCO Precis. Oncol. *8*, e2400007. https://doi.org/10.1200/PO.24.00007.
20. Wang, W.x., Xu, C., Chen, Y., Cai, X., Fang, Y., Zhang, Q., Zhu, Y.c., Yu, Z., Chen, G., Wang, H., et al. (2019). An EGFR extracellular domain mutation data in the East Asian non-small cell lung cancer populations and response to icotinib: A multicenter study. J. Clin. Oncol. *37*, e13000. https://doi.org/10.1200/JCO.2019.37.15_SUPPL.E13000.
21. Hayes, T.K., Aquilanti, E., Persky, N.S., Yang, X., Kim, E.E., Brenan, L., Goodale, A.B., Alan, D., Sharpe, T., Shue, R.E., et al. (2024). Comprehensive mutational scanning of EGFR reveals TKI sensitivities of extracellular domain mutants. Nat. Commun. *15*, 2742. https://doi.org/10.1038/S41467-024-45594-4.
22. Braig, F., März, M., Schieferdecker, A., Schulte, A., Voigt, M., Stein, A., Grob, T., Alawi, M., Indenbirken, D., Kriegs, M., et al. (2015). Epidermal growth factor receptor mutation mediates cross-resistance to panitumumab and cetuximab in gastrointestinal cancer. Oncotarget *6*, 12035–12047. https://doi.org/10.18632/ONCOTARGET.3574.
23. de Andrade, K.C., Lee, E.E., Tookmanian, E.M., Kesserwan, C.A., Manfredi, J.J., Hatton, J.N., Loukissas, J.K., Zavadil, J., Zhou, L., Olivier, M., et al. (2022). The TP53 Database: transition from the International Agency for Research on Cancer to the US National Cancer Institute. Cell Death Differ. *29*, 1071–1073. https://doi.org/10.1038/S41418-022-00976-3.
24. National Cancer Institute (NCI) (2025). The TP53 Database compiles various types of data and information from the literature and generalist databases on human TP53 gene variations related to cancer. https://tp.cancer.gov.
25. Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis. Oncol. *2017*, 1–16. https://doi.org/10.1200/PO.17.00011.
26. Hong, T.H., Cha, H., Shim, J.H., Lee, B., Chung, J., Lee, C., Kim, N.K.D., Choi, Y.L., Hwang, S., Lee, Y., et al. (2020). Clinical advantage of targeted sequencing for unbiased tumor mutational burden estimation in samples with low tumor purity. J. Immunother. Cancer *8*, e001199. https://doi.org/10.1136/JITC-2020-001199.
27. Xiao, W., Ren, L., Chen, Z., Fang, L.T., Zhao, Y., Lack, J., Guan, M., Zhu, B., Jaeger, E., Kerrigan, L., et al. (2021). Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. Nat. Biotechnol. *39*, 1141–1150. https://doi.org/10.1038/S41587-021-00994-5.
28. Rodenhuis, S., van de Wetering, M.L., Mooi, W.J., Evers, S.G., van Zandwijk, N., and Bos, J.L. (1987). Mutational Activation of the K-ras Oncogene. N. Engl. J. Med. *317*, 929–935. https://doi.org/10.1056/NEJM198710083171504.
29. Thomas, N.E. (2006). BRAF somatic mutations in malignant melanoma and melanocytic naevi. Melanoma Res. *16*, 97–103. https://doi.org/10.1097/01.CMR.0000215035.38436.87.
30. Baugh, E.H., Ke, H., Levine, A.J., Bonneau, R.A., and Chan, C.S. (2018). Why are there hotspot mutations in the TP53 gene in human cancers? Cell Death Differ. *25*, 154–160. https://doi.org/10.1038/CDD.2017.180.
31. Cannataro, V.L., Glasmacher, K.A., and Hampson, C.E. (2024). Mutations, substitutions, and selection: Linking mutagenic processes to cancer using evolutionary theory. Biochim. Biophys. Acta. Mol. Basis Dis. *1870*, 167268. https://doi.org/10.1016/j.bbadis.2024.167268.

32. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. Cell *144*, 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

33. González-Silva, L., Quevedo, L., and Varela, I. (2020). Tumor Functional Heterogeneity Unraveled by scRNA-seq Technologies. Trends Cancer *6*, 13–19. https://doi.org/10.1016/j.trecan.2019.11.010.

34. Sandve, G.K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Comput. Biol. *9*, e1003285. https://doi.org/10.1371/JOURNAL.PCBI.1003285.

35. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W.E., et al. (2013). Low concordance of multiple variant-calling pipelines: Practical implications for exome and genome sequencing. Genome Med. *5*, 28. https://doi.org/10.1186/GM432/FIGURES/5.

36. Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C., and Jaffe, D.B. (2013). Characterizing and measuring bias in sequence data. Genome Biol. *14*, R51. https://doi.org/10.1186/GB-2013-14-5-R51/FIGURES/6.

37. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.

38. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/NATURE193.

39. R Core Team (2023). R: A Language and Environment for Statistical Computing_ (Vienna, Austria: R Foundation for Statistical Computing).

40. Aho, A.V., Kernighan, B.W., and Weinberger, P.J. (1979). Awk— a pattern scanning and processing language. Softw. Pract. Exp. *9*, 267–279. https://doi.org/10.1002/spe.4380090403.

41. Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2022). shiny: Web Application Framework for R. https://shiny.posit.co/tutorial/.

42. Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Hocking, T., Schwendinger, B., Krylov, I., and Srinivasan, A. (2019). data.table: Extension of "data.frame". https://r-datatable.com.Manual.

43. Gagolewski, M. (2022). stringi: Fast and Portable Character String Processing in R. J. Stat. Softw. *103*. https://doi.org/10.18637/jss.v103.i02.

44. Ooms, J. (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. Preprint at bioRxiv. https://doi.org/10.48550/arXiv.1403.2805.

45. Mersmann, O., Beleites, C., Hurling, R., Friedman, A., and Ulrich, J.M. (2021). microbenchmark: Accurate Timing Functions. https://github.com/joshuaulrich/microbenchmark/.

46. Rescorla, E. (2018). The Transport Layer Security (TLS) Protocol Version 1.3. https://doi.org/10.17487/RFC8446.

47. National Institute of Standards and Technology (2001). Advanced encryption standard (AES). https://doi.org/10.6028/NIST.FIPS.197.

48. Morales, J., Pujar, S., Loveland, J.E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C.M., et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature *604*, 310–315. https://doi.org/10.1038/s41586-022-04558-8.

49. Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., et al. (2011). International cancer genome consortium data portal-a one-stop shop for cancer genomics data. Database *2011*, bar026. https://doi.org/10.1093/database/bar026.

50. Jee, J., Fong, C., Pichotta, K., Tran, T.N., Luthra, A., Waters, M., Fu, C., Altoe, M., Liu, S.Y., Maron, S.B., et al. (2024). Automated real-world data integration improves cancer outcome prediction. Nature *636*, 728–736. https://doi.org/10.1038/S41586-024-08167-5.

51. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. https://doi.org/10.1038/S41568-018-0060-1.

52. Benjamini, Y., Krieger, A.M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika *93*, 491–507. https://doi.org/10.1093/biomet/93.3.491.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
COSMIC v100 Genome Screen Mutants data	COSMIC[9]	https://cancer.sanger.ac.uk/cosmic/download/cosmic/v100/genomescreensmutanttsv
COSMIC v100 Samples data	COSMIC[9]	https://cancer.sanger.ac.uk/cosmic/download/cosmic/v100/sample
MANE Select data	RefSeq[38]	https://www.ncbi.nlm.nih.gov/refseq/MANE/
Tumor variants in human tumor samples data file	TP53 database[23]	https://tp.cancer.gov/
Prognostic value of tumor variants file	TP53 database[23]	https://tp.cancer.gov/
Cancer Gene Census data	COSMIC[9]	https://cancer.sanger.ac.uk/census
MSK-CHORD cohort data	cBioPortal[12]	https://cbioportal-datahub.s3.amazonaws.com/msk_chord_2024.tar.gz
PCAWG cohort data	cBioPortal[12]	https://cbioportal-datahub.s3.amazonaws.com/pancan_pcawg_2020.tar.gz
Software and algorithms		
Prism v9 and v10	GraphPad	https://www.graphpad.com/
R	The R project[39]	https://www.r-project.org/
HTML5	WHATWG	https://html.spec.whatwg.org/multipage/
CSS	WORLD WIDE WEB CONSORTIUM, Cascading Style Sheets (CSS) Working Group	https://www.w3.org/Style/CSS/Overview.en.html
JavaScript	MDN	https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference
AWK	GNU[40]	https://www.gnu.org/software/gawk/
R shiny	Posit[41]	https://shiny.posit.co/
GNU Gzip	The Free Software Foundation	https://www.gnu.org/software/gzip/
R data.table package	Barrett et al.[42]	https://r-datatable.com/
R stringi package	Gagolewski, M.[43]	https://stringi.gagolewski.com/index.html
R jsonlite package	Ooms et al.[44]	https://cran.r-project.org/web/packages/jsonlite/index.html
R microbenchmark package	Mersmann et al.[45]	https://cran.r-project.org/web/packages/microbenchmark/index.html
Pigz	Adler, Mark.	https://zlib.net/pigz/
NGINX	NGINX	https://nginx.org/
Transport Layer Security (TLS) 1.3	Rescorla, E.[46]	https://www.rfc-editor.org/rfc/rfc8446.html
Advanced Encryption Standard (AES-256)	National Institute of Standards and Technology	https://doi.org/10.6028/NIST.FIPS.197
Google Lighthouse	Google	https://developer.chrome.com/docs/lighthouse/overview
MutationMapper	cBioPortal[12]	https://www.cbioportal.org/mutation_mapper
Other		
Resource website for the DORM database	This paper	https://eleniuslabtools.utu.fi/main/docs/DORM.html

METHOD DETAILS

Website and web server

The DORM database is accessible at https://eleniuslabtools.utu.fi/tools/DORM/Mutations/, and all requests to the server are handled by an NGINX reverse-proxy (https://nginx.org/) that encrypts the traffic between our server and the end-user's web-browser.

The connection is encrypted using the latest Transport Layer Security (TLS) cryptographic protocol 1.3[38] and an industry standard 256-bit Advanced Encryption Standard (AES-256).[39] As a fallback, the server of DORM also supports connections over TLS 1.2 to support legacy hardware and browsers. The landing page website and the documentation is built using HTML5, CSS and JavaScript. The web tools are built using Shiny[40] and R.[41] These services are hosted on a virtual private server at the premises of University of Turku, Turku, Finland. The source code for deploying DORM as an R Shiny app is available at https://github.com/dchakro/DORM_Mutations and https://github.com/dchakro/DORM_Residues repositories.

Hardware

Database processing & analysis: Apple iMac (early 2013) equipped with Intel Core i5 CPU (4 cores – 3.2 GHz), 24 GB DDR3 RAM, 500 GB SSD running macOS Catalina 10.15.

Server: Virtual private server (KVM virtualization) with Intel(R) Xeon(R) Gold 5120 CPU (1 core – 2.20 GHz), 6 GB ECC RAM, 100 GB HDD running Ubuntu 22.04 LTS.

Web performance testing: Apple MacBook Pro (early 2015) equipped with an Intel Core i5 CPU (2 cores – 2.7 GHz), 8 GB DDR3 RAM, 500 GB SSD running macOS Catalina 10.15. The device was connected via a 5 GHz Wi-Fi router to the public ISP (i.e., outside the network where the DORM database is hosted) over a 100 Mbps fiber optic broadband connection.

Data and processing of data

Data were acquired from COSMIC release v103 (released November 18, 2025 https://cancer.sanger.ac.uk) as a GNU zip (GZIP) archive of the tab-delimited text file with all mutations identified from genome-wide screens (includes data from whole genome sequencing, and whole exome sequencing). The samples from targeted screens were excluded to ensure our analysis is free from selection bias and to facilitate the direct comparison of the frequency of mutations between different proteins in a particular tissue. No additional germline filtering was performed. As a result, DORM includes: (i) recurrent mutations reported as somatic in other cancer samples, (ii) recurrent mutations confirmed as somatic by comparison of tumor and normal tissue, and (iii) recurrent mutations with unconfirmed somatic status due to lack of normal tissue sampling. As COSMIC does not report metrics for tumor purity or subclonality, these characteristics are unknown for the recurrent mutations in DORM. No additional orthogonal validation was performed.

Pre-processing: The decompressed data is processed using the ‘‘awk’’ programming language[42] to select relevant columns (named, Gene name, Sample name, Primary site, Primary histology, Genome-wide screen, Mutation CDS, Mutation AA). The selected columns were read in R by using the data.table::fread() function.[43] The complete database was stored as standard R object in the .RDS file format, with a notable difference: instead of saveRDS from R base, which uses serialized compression, parallelized GZIP (pigz: https://zlib.net/pigz/) was used for compression – decompression. The functions for reading-writing R objects in .RDS files using parallelized compression-decompression are described in this R script.

Filtering: The duplicate entries for mutations mapped to other than the MANE select transcripts[44] were removed (Figure S1). Mutations with unknown consequences on the protein level were removed. From these, silent mutations were removed (Figure S1). To prevent redundant counting of the same mutation, a unique identifier was generated for each mutation using the sample name, protein name, and amino acid change. Entries with identical mutation IDs were considered duplicates and removed from the dataset (Figure S1). Mutations with single occurrences (i.e., frequency = 1) were removed from the list of unique coding mutations, as they are not part of the pool of recurrent mutations. The filtered database with unique coding mutations was stored as a parallelized GZIP .RDS file, enabling faster load times. Searching and parsing of the text was performed with the ‘stringi’ R package.[45]

Processing: For each mutation, its cumulative frequency of occurrence, as well as its frequency in cancers of various tissues, was calculated and compiled into a table. The table was sorted by mutation frequency (total number of samples across all cancers) and then stored as a parallelized-GZIP .RDS file.

Updates: Since 2004, marking the release of COSMIC v1, the dataset has been updated on average four times per year (range: 11 releases in 2006 and two releases in 2020). The COSMIC data releases need to be acquired from (https://cancer.sanger.ac.uk), then our optimized pipeline can be run with a shell script that automates the processing and generation of the underlying database for DORM.

Benchmarking and testing performance

To evaluate the performance of different code blocks, the ‘microbenchmark’ R package[46] was used to gather data. The data were graphically represented using Graphpad Prism 9 and 10.The code blocks used for testing and benchmarking their performance is available at https://github.com/KE-group/DORM-2022 repository.

The performance of the websites hosting the databases was measured on Google Chrome (v. 97.0.4692.99) with Google Lighthouse (v. 8.5.0) (available in Chrome DevTools). Lighthouse (https://github.com/GoogleChrome/lighthouse) is an open-source tool for automated auditing and assessing performance metrics. A search for EGFR mutations was performed on the five databases (DORM, COSMIC, ICGC,[47] cBioPortal and AACR GENIE), and links (Table S1) to those individual searches were used to test the performance of the databases. This was performed to discount the varying duration required to do the same search on the four databases. Lighthouse 8 produces a performance score which is a weighted average of First Contentful Paint (10%, marks the time at which the first text or image is painted), Speed Index (10%, shows how quickly the contents of a page are visibly populated), Largest

Contentful Paint (25%, marks the time at which the largest text or image is painted), Time to interactive (10%, the amount of time it takes for the page to become fully interactive), Total Blocking Time (30%, measures the total amount of time that a page is blocked from responding to user input), and Cumulative Layout Shift (15%, measures the unexpected movement of page content). The JSON data in the lighthouse reports was parsed using the 'jsonlite' R package[48] and tabulated in R. The data were graphically represented using GraphPad Prism 9.

Mutation frequency estimation

The frequency of a mutation was estimated by dividing the number of patient samples with the mutation by the total number of patient samples from the same cancer type. To analyze the effect of targeted data inclusion on the mutation frequency estimation, the "Genome Screens Mutants" and "Targeted Screens Mutants" data in the download page of the COSMIC database v100[9] were processed as the data in DORM.

To analyze the effect of sample selection bias on mutation frequency estimation, the frequency estimates for common EGFR kinase domain mutations were acquired from cBioPortal v 6.0.12[12] and DORM. Either all lung cancer studies, lung cancer studies related to the TCGA PanCancer Atlas or lung cancer studies included in the curated set of non-redundant studies were selected for the frequency estimation.

To perform the pan-cancer analysis on the frequency of EGFR mutations, frequency estimates from DORM, COSMIC, and cBioPortal databases were acquired. The mutation frequency information in COSMIC was acquired from the "Variants" table of the EGFR Gene page in COSMIC. The studies included in the "curated set of non-redundant studies" default setting were used for the frequency estimation in the cBioPortal database. The lollipop plots were drawn with the MutationMapper tool in cBioPortal.[12]

Comparative evaluation of recurrent and unique mutation effects

Recurrent mutations from DORM and unique mutations from COSMIC v100, which included all mutations in COSMIC that are not listed in DORM, were used for the analyses. The ERBB mutations identified as growth promoting were sourced from the results of *in vitro* screen of activating mutations.[15–17] The predicted and experimentally verified functional classifications of *TP53* mutations were sourced from the *TP53* database.[23,24] Variants classified as damaging or partially damaging in at least two out of five prediction methods in the database (BayesDel, REVEL, AGVGDClass, SIFTClass, Polyphen2) were considered as predicted damaging. Variants classified as damaging or partially damaging in the results of at least one out of three functional screens in the database (TransactivationClass, DNE_LOFclass, DNEclass) were considered as verified damaging.

The prognostic value of *TP53* mutations were sourced from the *TP53* database. Mutations included in studies where *TP53* mutations were associated either with treatment resistance, poorer survival or tumor recurrence were considered unfavourable prognoses. In contrast, *TP53* mutations included in studies where *TP53* mutations were associated with survival, treatment response or no relapse were considered favourable prognoses. *TP53* mutations that were associated with unfavourable and favourable prognoses in separate studies were categorized as both. The mutations that had no prognostic association in the studies were categorized as no association.

Survival analysis of patients in the pan-cancer MSK-CHORD[49] and PCAWG[50] cohorts were conducted using the data sourced from the cBioPortal database.[12] The overall survival was analyzed. All patients with at least one recurrent mutation in any oncogene, tumor suppressor gene or a randomized set of genes were assigned to the recurrent group. Patients with only unique mutations in oncogenes, tumor suppressor genes or a randomized set of genes were assigned to the unique group. The list of oncogenes and tumor suppressor genes were sourced from the Cancer Gene Census resource.[1] The randomized set of genes was generated by random sampling of all mutated genes. The size of the randomized sets was set to equal the number of oncogenes and tumor suppressor genes in the Cancer Gene Census resource (n=325). The process was repeated 10 times and a representative case was visualized.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical testing comparing multiple groups in testing code block performance was performed using Brown Forsythe and Welch ANOVA test and correction for multiple testing was done by controlling the false discovery rate using the two-stage step-up method of Benjamini, Krieger and Yekutieli[51,52] in Grahpad Prism 9. Statistical testing comparing two groups of observations was done using Welch's t-test in Grahpad Prism 9. The statistical significance of the relative fold change of the relative mutation frequency in complete (genome-wide and targeted) vs genome-wide data was calculated with one sample Wilcoxon test against the hypothetical median value of 1. The statistical significance of the sample selection to the mutation frequency estimates was calculated with a two-way ANOVA. The normality and homoscedasticity assumptions were tested with D'Agostino-Pearson omnibus, Anderson-Darling, Shapiro-Wilk, Kolmogorov-Smirnov, and Spearman's tests. The statistical significance of the proportions of unique and recurring mutations and mutated residues in the groups of ERBB mutations identified in the screens and all other ERBB mutations, and damaging and tolerated mutations in the groups of recurrent and unique *TP53* mutations was estimated with the Fisher's exact test. The statistical significance of the proportions of mutations associated with unfavourable, favourable and both prognoses as well as mutations with no prognostic value in the groups of recurrent and unique mutations was estimated with the Fisher's exact

test. The statistical significance of the difference in survival of the recurrent and unique mutation groups was estimated with the Mantel-Cox test. Statistical parameters such as what individual points represent are reported in the Figure Legends.

ADDITIONAL RESOURCES

Access to the DORM database and its documentation and FAQs is available through the weblinks: https://eleniuslabtools.utu.fi/main/docs/DORM.html; https://eleniuslabtools.utu.fi/main/docs/DORM-FAQ.html.