Tumors from individuals with cancer are frequently genetically profiled for information about the driving forces behind the disease. We present the CancerMine resource, a text-mined and routinely updated database of drivers, oncogenes and tumor suppressors in different types of cancer. All data are available online (http://bionlp.bcgsc.ca/cancermine) and downloadable under a Creative Commons Zero license for ease of use.

To interpret the somatic events present in a patient sample, it is necessary to know which genes have important roles in the development of the corresponding cancer type. This normally requires substantial literature review. We have developed a text-mining approach that identifies mentions of genes as drivers, oncogenes or tumor suppressors. This approach is used to create a resource that will be kept up-to-date with monthly releases and is amenable for data analysis pipelines. We also provide an online tool, based on these data, for identifying cancer genes from a gene list and interactive plots of cancer-type clustering, which are helpful for understanding the somatic landscapes of cancer.

Oncogenes are genes that (in either their normal or aberrated form) promote the development of cancer, whereas tumor suppressors act against carcinogenesis. "Drivers" refers to genes that are important in cancer development and can be either oncogenes or tumor suppressors. Some genes (for example, NOTCH1) have been identified as oncogenes in one type of cancer and as tumor suppressors in another type1. Furthermore, many genes are important in only certain types of cancer and are probably irrelevant in others. The type of cancer provides important context when one is interpreting the relevance of somatic aberrations in a patient sample.

Different methods exist to identify potential cancer-related genes, including statistical analysis of mutation frequency in large genomic cohorts2 and in vitro studies of gene knockdown3. Several resources have been built to catalog the roles of genes in cancer. The Cancer Gene Census (CGC)4 uses data from the Catalogue of Somatic Mutations in Cancer (COSMIC) to list known oncogenes and tumor suppressors. The Network of Cancer Genes5 builds upon the CGC and integrates a wide variety of additional contextual data, such as the frequency of mutations. IntOGen6 uses data from large-scale sequencing projects (for example, the Cancer Genome Atlas (TCGA)) to collate the importance of cancer genes. ONGene7 and TSGene8 list oncogenes and tumor suppressors but do not associate them with specific cancer types. The Clinical Interpretation of Variants in Cancer (CIViC) database curates clinically relevant variants and contains a set of genes that are relevant for different types of cancer.

All manually curated databases face the overwhelming curation burden of expert curator time and costs necessary to stay up-to-date. Text-mined databases offer an automated approach that can provide high-quality results with regular updates. To enable this approach, we built on previous text-mining approaches for extracting gene-disease relations9. We extracted titles, abstracts and available full-text articles from PubMed, PubMed Central Open Access (PMCOA) subsets and PubMed Central Author Manuscript Collection (PMCAMC). We then identified sentences that mentioned a gene name, a cancer type and keywords suggestive of a gene playing a role in cancer (Supplementary Table 1). We manually annotated the drivers, oncogenes and tumor suppressors discussed in 1,500 sentences (Supplementary Table 2, three annotators, with mean inter-annotator agreement F1 score of 0.77). These sentences were from throughout the papers and could relate to a new result or a previous result. An example of a sentence describing two oncogene associations is, "KRAS is a known oncogene in lung cancer and pancreatic cancer." More examples can be found in Supplementary Table 3. An individual sentence could contain zero, one or many drivers, oncogenes and tumor suppressor genes associated with different types of cancer.

We trained a Kindred relation classifier10 that learns the characteristics of these sentences by building a logistic regression classifier on word frequencies and semantic features. The classifier was then able to predict the pairs of genes and cancer types with their associated roles (driver, oncogene, or tumor suppressor) from a given sentence. A knowledge base that contains a high number of incorrect associations would quickly lose users and be unusable for other analyses. Therefore, we controlled the precision-recall trade-off by applying a high threshold on classifier scores to reduce the number of false positives with the accepted increase in false negatives (Fig. 1a and Supplementary Table 4, error analysis in Supplementary Table 5). This gave an average precision of 85.6% and recall of 29.4% across the three gene role types. We are most interested in well-established drivers, oncogenes and tumor suppressors, which will be mentioned in many papers—first in the discovery paper(s), and then in other papers. Even with a low recall of 29.4%, the redundancy in the literature means that the association needs to be mentioned in only seven papers for a >90% chance of being identified. Important cancer gene associations are identified in hundreds of papers by CancerMine. However, users should be aware that this method cannot provide an exhaustive list of all known cancer–gene associations.

The classifier was applied to the remaining corpora to identify 38,106 sentences containing 6,843 mentions of drivers, 26,909 mentions of oncogenes and 14,460 mentions of tumor suppressors. By aggregating the results, we found 4,038 genes linked to 425 cancer types as drivers, oncogenes or tumor suppressors, with the highest number of associations in well-established cancer genes (for example, TP53, MYC and KRAS) (Fig. 1b). Of the cancer gene associations, 26.6% are discussed in more than one paper. The resulting knowledge base incorporates information from 28,334 papers.
CancerMine is updated monthly, and every new release adds new cancer gene associations. These may be completely new discoveries or mentions of known gene associations that the system failed to capture successfully from previous papers. As evidence that automatically curated knowledge bases need to be regularly updated, we found that in 2018, approximately 186 new cancer gene associations were extracted from the literature each month. Delving deeper into these new cancer gene associations, we found a strong association between their novelty to CancerMine and their location within the paper (Supplementary Fig. 1). Specifically, we found that novel associations were more likely to appear ($\chi^2 = 270.85$, d.f. = 1, $P < 2.2 \times 10^{-16}$) in the non-introductory sections of a paper (for example, Results and Discussion). Examples are shown in Supplementary Table 6, which shows some of the frequent differences between novel and non-novel mentions. We found that, overall, 58.1% of mentions of cancer gene roles were found in the main text as opposed to the title or abstract (Supplementary Fig. 2). Furthermore, 53.7% of all cancer gene roles were found only in full-text articles. This also underlines the need to text-mine full-text articles, as cancer-gene associations are found in all subsections of papers (Supplementary Fig. 1) and not only their abstracts, and highlights the need for greater access to articles for text mining.

CancerMine can also identify genes that serve dual roles as oncogenes in some types of cancer and as tumor suppressors in other types. We identified genes that were mentioned in at least four papers that have strong support as an oncogene in at least one cancer type and also as a tumor suppressor in at least one other type. NOTCH1, FOXP1 and several other genes were identified as having dual roles (Supplementary Table 7).

We compared CancerMine with the CGC and several other cancer-gene-related resources (Fig. 2a) and found that CancerMine contains substantially more cancer gene associations than other resources but does not show substantial overlap with the CGC or IntOGen. The IntOGen project provides inferences of driver genes from large datasets such as TCGA, and many will not yet have been examined in publications, which may explain the poor overlap with CancerMine. The CGC is a manually curated resource based on COSMIC, and the poor overlap may also be explained by a lack of publications discussing all of the data in COSMIC. When the threshold of the classifier was lowered to reduce the false negative rate, the overlap improved, but not substantially (Supplementary Fig. 3), suggesting that many of the gene associations in the CGC and IntOGen are not mentioned in the literature in a form that CancerMine can capture. CancerMine does not extract associations from tables or figures, or across multiple sentences.

The number of papers that discuss a gene as important for a cancer provides a metric of the importance of that gene in cancer development. Using this information, we generated profiles for each cancer type. As expected, similar cancer types clustered together.
Fig. 2 | Comparison to other resources and cancer-type clustering by gene role. a, A comparison with other resources revealed that CancerMine contains more gene–cancer associations but has poor overlap with the CGC and IntOGen. b, CancerMine overlaps substantially with the genes listed in the TSGene, ONGene and CIViC resources. Using the number of papers that discuss a gene role with a cancer, a profile can be created for each cancer type. Colors of increasing intensity indicate higher importance scores.

(Fig. 2b), providing an excellent summary of how different cancer types differ on the basis of their genomic profiles. We then validated these profiles, specifically for tumor suppressors, using data from TCGA. We reasoned that probable deleterious point mutations would often target important tumor suppressors. We matched each TCGA sample (in seven TCGA cancer-type projects) with the CancerMine profile that most closely overlapped its somatic mutations. We found that five of the seven CancerMine profiles had their highest proportion matches with the corresponding TCGA project (Supplementary Fig. 4).

All data are free to view and download. We hope that CancerMine will be a valuable resource to the cancer research community and that the methods will be of interest to others working in the field of biomedical knowledge-base curation.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41592-019-0422-y.

Received: 30 July 2018; Accepted: 10 April 2019; Published online: 20 May 2019

References
1. Radtke, F. & Raj, K. Nat. Rev. Cancer 3, 756 (2003).
2. Kristensen, V. N. et al. Nat. Rev. Cancer 14, 299–313 (2014).
3. Zender, L. et al. Cell 135, 852–864 (2008).
4. Patrée, P. A. et al. Nat. Rev. Cancer 4, 177 (2004).
5. Repana, D. et al. Genome Biol. 20, 1 (2019).
6. Gonzalez-Perez, A. et al. Nat. Methods 10, 1081 (2013).
7. Liu, Y., Sun, J. & Zhao, M. J. Genet. Genom. 44, 119–121 (2017).
8. Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. Nucleic Acids Res. 44, D1023–D1031 (2015).
9. Griffith, M. et al. Nat. Genet. 49, 170 (2017).
10. Chun, H.-W. et al. Pac. Symp. Biocomput. 2006, 4–15 (2006).
11. Singhal, A., Simmons, M. & Lu, Z. PLoS Comput. Biol. 12, e1005017 (2016).
12. Lever, J. & Jones, S. BioNLP 2017, 176–183 (2017).

Acknowledgements
J.L. was supported by a Vanier Canada Graduate Scholarship. Funding for S.J.M.J. and M.R.J. was provided through the Personalized Oncogenomics (POG) program, which is generously supported by the BC Cancer Foundation and Genome British Columbia (project B20POG). The authors would like to thank Compute Canada for the use of computational infrastructure for this research.

Author contributions
J.L., M.R.J. and S.J.M.J. conceived the idea. J.L. implemented the software and carried out the analysis. J.L., E.Y.Z. and J.G. annotated the sentence data. All authors contributed to the writing of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0422-y.
Reprints and permissions information is available at www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.J.M.J.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

Processing of corpora. PubMed abstracts and full-text articles from PMCOA subset and PMCAMC were downloaded from the National Center for Biotechnology Information (NCBI) FTP website using the PubRunner framework (https://github.com/jakelever/pubrunner). They were then converted to BioC format using PubRunner’s convert functionality. This strips out formatting tags and other metadata, and retains the Unicode text of the title, abstract and, for PubMed Central, using PubRunner’s convert functionality. This strips out formatting tags and other metadata, and retains the Unicode text of the title, abstract and, for PubMed Central, the full article. The source of the text (title, abstract, article) is also encoded.

Entity recognition. Lists of cancer types and gene names were built using a subset of the Disease Ontology (DO) and NCBI gene lists. These were complemented by matching to the Unified Medical Language System (UMLS). For cancer types, this was achieved using the associated ID in DO or through exact string matching of the DO item title. For gene names, the Entrez ID was used to match with UMLS IDs. The cancer type was then associated with a DO ID, and the gene names were associated with their HUGO gene name. These cancer and gene lists were then pruned with a manual list of stop words with several custom additions for alternative spellings or acronyms of cancers. All cancer terms with fewer than four letters were removed except for a select set of abbreviations (for example, GBM for glioblastoma multiforme). The term lists are managed in the BioWordlists Github repository (https://github.com/jakelever/biowordlists).

The corpus text was loaded in BioC format and processed using the Kindred Python package that, as of v2.0, uses the Spacy IO parser. Using the tokenization, entities were identified through exact string matching against tokens. Longer entity names with more tokens were prioritized and removed from the sentence as entities were identified. Non-fusion terms, which are mentions with multiple gene symbols that actually refer to a single gene (for example, HER2/NEU), were identified when two genes with matching HUGO IDs were attached and combined to be a single non-fusion gene entity. Genes mentioned in the context of pathways were also removed (for example, MTOR pathway) using a list of pathway-related keywords.

Sentence selection. After Kindred parsing, the sentences with tagged entities were searched for those containing at least one cancer type and at least one gene name. These sentences were then filtered using the terms ‘tumor suppressor’, ‘oncogenen’ and ‘driver’ to enrich for sentences that were probably discussing these gene roles.

Annotation. From the complete set, 1,600 of the sentences were then selected randomly and output into the BioNLP Shared Task format for ingestion into an online annotation platform. This platform was then used by three expert annotators who were all PhD students actively engaged in precision cancer projects. Within each sentence, the platform presents each possible pair of a gene and cancer, and the user must select driver, oncogene and tumor suppressor annotations. The first 100 sentences were used to help the users to understand the system, evaluate initial inter-annotator agreement and adjust the annotation guidelines (available at the GitHub repository). The results were then discarded and the complete 1,500 sentences were annotated by the first two annotators. Of the 1,500 sentences, 379 were found to contain disagreements. The third annotator then annotated the sentences that the first two disagreed on and approximately 400 sentences that did not contain disagreements for more accurate inter-annotator agreement calculations. The inter-annotator agreement was calculated using the F1 score. A ‘gold’ corpus was created using the majority vote of the annotations of the three annotators.

Relation extraction. To create a training and test split, we used 75% of the 1,500 sentences as a training set and trained a Kindred relation classifier with an underlying logistic regression model for all three gene roles (driver, oncogene and tumor suppressor). The threshold was varied to generate the precision-recall curves with evaluation on the remaining 25% of sentences. With the selection of the optimal thresholds (Supplementary Table 4), a complete model was trained using all 1,500 sentences. This model was then applied to all sentences found in PubMed, PMCOA and PMCAMC that fit the sentence requirements. The associated gene and cancer type IDs were extracted, entity names were normalized and the specific sentence was extracted.

We note examples that our system would be unable to extract. If the gene name or cancer name are missing from the almost exhaustive nomenclature sets, the system cannot identify the entities, and therefore no relation can be extracted. The DO has a large list of cancer gene names, but slight variations of them may be missed. For example, a sentence discussing ‘lung carcinogenesis’ instead of ‘lung cancer development’ would not be tagged for an appropriate cancer entity. Some missing nomenclature can be fixed with future versions of the resource. However, substantially different wording used to describe the actual relations would require re-annotation of data and retraining of the classifier.

Web portal. The resulting cancer gene role data were aggregated by the triples (gene, cancer, role) to count the number of citations supporting each cancer gene role. This information was then presented in tabular and chart form using a Shiny web application. The data can be searched by gene or cancer or using a gene list that highlights the important cancer genes.

Resource comparisons. The data from the CGC, IntOGen, TSGene and ONGene resources were downloaded for comparison. HUGO gene IDs in CancerMine were mapped to Entrez Gene IDs. CGC data were mapped to DO cancer types using a combination of the cancer synonym list created for CancerMine and manual curation. Oncogenes and tumor suppressors were extracted using the presence of ‘oncogene’ or ‘TSG’ in the ‘Role in Cancer’ column. The mapped CGC data were then compared against the set of oncogenes and tumor suppressors in CancerMine, taking the role into account. IntOGen cancer types were mapped manually to corresponding DO cancer types and compared against all of CancerMine without the role, as IntOGen does not provide role information. The TSGene, ONGene and CIVIC gene sets were compared against the CancerMine gene sets without an associated cancer type.

CancerMine profiles and TCGA analysis. The number of papers that relate a gene role (for example, RUNX3 as a tumor suppressor) to a cancer type (for example, stomach cancer) is used as a metric of importance for that gene role in the cancer type. The paper number is log transformed and divided by the largest value for that specific cancer type. This normalizes for the cancer type, as many more papers discuss breast cancer than basal cell carcinoma. For each cancer type, we then have a set of importance values (a CancerMine profile) for all gene roles, with the most important having a value of 1 and all others having lower values. The top 25 cancer types were then selected, and the top 25 gene roles were identified for them. These profiles were then hierarchically clustered to generate a heat map using heatmaply.

The open-access VarScan somatic mutation calls for the seven TCGA projects (BRCA, COAD, LIHC, PRAD, LGG, LUAD and STAD) were downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov). They were filtered for mutations that contained a stop gain or that were classified as probably damaging or deleterious by PolyPhen. Tumor-suppressor-specific CancerMine profiles were generated that used all tumor suppressors for each cancer type. The citation counts were again log transformed and rescaled to produce the CancerMine tumor suppressor profile. Each TCGA sample was represented as a binary vector matching the filtered mutations. The dot-product of a sample vector and a CancerMine profile vector produced the sum of citation weightings and gave the score. For each sample, the score was calculated for all seven cancer types and the highest score was used to label the sample. A sample that did not contain tumor suppressor mutations associated with any of the seven profiles or that could not be labeled unambiguously was labeled as ‘none’.

Online access and updates. A bioinformatics resource is valuable only if it is kept up-to-date. This is very apparent in this area, as several previous knowledge bases that relate to this area have fallen into disrepair and are no longer accessible. To ensure that this project has a long life, we have taken three key steps. First, all code is publicly available and hosted on GitHub, as well as recorded in the Zenodo repository. Second, all data are stored in the Zenodo repository under a Creative Commons Zero license. This repository will guarantee the future of the data for this project. Third, and most important, we have developed a software framework (PubRunner) to make it easier to keep text-mined resources up-to-date. It manages the challenging tasks of downloading the latest publication data, converting formats, executing tools on a cluster and uploading data as required.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data can be viewed and downloaded through the online viewer (https://github.com/jakelever/cancermine). The February 2019 CancerMine release was used for this analysis (https://doi.org/10.5281/zenodo.2557358). All releases can be found at https://doi.org/10.5281/zenodo.1156241.

Code availability

All code for text mining and the analysis in this paper are available in the Github repository (https://github.com/jakelever/cancermine). The specific code release is archived in Zenodo (https://doi.org/10.5281/zenodo.2586207).

References

13. Comeau, D. C. et al. Database 2013, bat064 (2013).
14. Kibbe, W. A. et al. Nucleic Acids Res. 43, D1071–D1078 (2014).
15. Maglott, D. et al. Nucleic Acids Res. 39 (Suppl. 1), D52–D57 (2010).
16. Bodenreider, O. Nucleic Acids Res. 32 (Suppl. 1), D267–D270 (2004).
17. Galili, T., O’Callaghan, A., Sidi, J. & Sievert, C. Bioinformatics 34, 1600–1602 (2017).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

☑ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☑ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☑ The statistical test(s) used AND whether they are one- or two-sided
☑ A description of all covariates tested
☑ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☑ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐ Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection
The text mining portion was undertaken using Python (v3.6.5), specifically the Kindred package (v2.2.5) and PubRunner (v0.5.2). All custom code is open-source and available at the GitHub repository (https://github.com/jakelever/cancermine) and archived at https://doi.org/10.5281/zenodo.2586207.

Data analysis
Data analysis was undertaken using the R statistical programming language (v3.4.1). All custom code is open-source and available at the GitHub repository (https://github.com/jakelever/cancermine) and archived at https://doi.org/10.5281/zenodo.2586207.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data can also be viewed and downloaded through the online viewer (https://github.com/jakelever/cancermine). The February 2019 CancerMine release was used for this analysis (https://doi.org/10.5281/zenodo.2557358). All releases can be found using the DOI: https://doi.org/10.5281/zenodo.1156241.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences  
- Behavioural & social sciences  
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | This is a machine learning analysis that processes all abstracts and all open-access full-text articles available in PubMed and PubMed Central. The 1,125 sentences are used for training the system and 375 are used for evaluation. |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | A custom set of gene names and cancer names that appear as synonyms in the UMLS metathesaurus were removed due to the high level of ambiguity |
| Replication | All code is public and is continually tested using the Travis CI system |
| Randomization | The training and test datasets were split randomly using a 75%/25% split. |
| Blinding | Blinding is not relevant to a machine learning study as the training test split was decided before the analysis began |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | Antibodies |
| ☒   | Eukaryotic cell lines |
| ☒   | Palaeontology |
| ☒   | Animals and other organisms |
| ☒   | Human research participants |
| ☒   | Clinical data |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒   | ChIP-seq |
| ☒   | Flow cytometry |
| ☒   | MRI-based neuroimaging |