Bayesian inference of negative and positive selection in human cancers

Donate Weghorn[1,2] & Shamil Sunyaev[1,2]

Cancer genomics efforts have identified genes and regulatory elements driving cancer development and neoplastic progression. From a microevolution standpoint, these are subject to positive selection. Although elusive in current studies, genes whose wild-type coding sequences are needed for tumor growth are also of key interest. They are expected to experience negative selection and stay intact under pressure of incessant mutation. The detection of significantly mutated (or undermutated) genes is completely confounded by the genomic heterogeneity of cancer mutation[1]. Here we present a hierarchical framework that allows modeling of coding point mutations. Application of the model to sequencing data from 17 cancer types demonstrates an increased power to detect known cancer driver genes and identifies new significantly mutated genes with highly plausible biological functions. The signal of negative selection is very subtle, but is detectable in several cancer types and in a pan-cancer data set. It is enriched in cell-essential genes identified in a CRISPR screen[2], as well as in genes with reported roles in cancer.

From an evolutionary perspective, cancer is a complex system that evolves asexually and is subject to antagonistic selective forces. Oncogenes and tumor suppressors evolve under strong positive selection when mutated and have been the prime target for cancer studies. In contrast, even though highly important for a full understanding of cancer vulnerabilities, negative selection has been an elusive phenomenon. Both positive and negative selection can be inferred by comparing the observed mutations at a given locus to the expectation under the sole action of the mutation process. However, this inference is entirely confounded by the variation in mutation rate along the cancer genome, and the concept of a constant background mutation rate has to be abandoned[3]. The problem is exacerbated by technical factors such as uneven sequencing coverage and mappability, as well as ploidy and purity of the samples. Initial approaches to address the problem of estimating the local background mutation rate relied on local gauging of mutation density by the observed number of neutral mutations. New methods to estimate local mutation rate, taking into account biological covariates such as replication timing, expression level, chromatin state or sequence context, have moved the field forward[1]. However, it is impossible to test whether local mutation rate estimates are exactly accurate at every locus. Even if errors in local mutation rate estimates are limited on average, they amplify in the case of extreme data points that masquerade as cancer driver genes. In contrast to point estimates, the statistical properties of the overall distribution of mutation densities can be fully validated with available data.

Using this concept, we present a probabilistic framework that addresses the problem of mutation rate variation, allowing for models that fit the observed data with high accuracy. Specifically, assuming that synonymous mutations evolve neutrally, we estimate the distribution of per-gene mutation probabilities by fitting synonymous mutation counts across the entire set of genes.

$$\begin{aligned} P(s\,|\,\theta) &= \int \mathrm{d}\lambda_s\, P(s\,|\,\lambda_s)\, P(\lambda_s\,|\,\theta) \\ &= \int \mathrm{d}\lambda_s\, \mathrm{Pois}(\lambda_s)\, P(\lambda_s\,|\,\theta) \end{aligned} \qquad (1)$$

Here $P(s|\lambda_s)$ is the probability of s synonymous mutations in a gene with gene-specific expectation λ_s (encapsulating local mutation density and synonymous target size). This probability is naturally modeled by the Poisson distribution. The expectation λ_s is assumed to be a random variable distributed according to $P(\lambda_s|\boldsymbol{\theta})$ (with parameter vector $\boldsymbol{\theta}$), which fully describes mutation rate heterogeneity and heterogeneity of mutation detection along the genome. For the specific assumption of gamma-distributed expected values, the observed counts s follow a negative binomial distribution, a special case presumed by Nik-Zainal *et al.*[4] (see discussion in the **Supplementary Note**).

Importantly, equation (1) implicitly models all known and unknown covariates of mutation rate, making it independent of their separate knowledge or inference. When applying the equation to different cancer types, we thus obtain for each a parametric model of the distribution of per-gene mutation densities that provides an excellent fit to the observed synonymous counts and, therefore, a baseline to the analysis of selection. Using this baseline, we derive the expected distributions of the numbers of missense and nonsense mutations under neutral evolution. These distributions follow from a rescaling of λ_s with the per-gene ratio of nonsynonymous to synonymous target size, which takes into account the cancer-type-specific context dependence of mutations and does not require additional parameters. With sequence context explicitly accounted for, this approach represents a calibration

[1]Department of Medicine, Division of Genetics, Brigham and Women's Hospital/Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. Correspondence should be addressed to S.S. (ssunyaev@rics.bwh.harvard.edu).

Received 4 March; accepted 11 October; published online 6 November 2017; doi:10.1038/ng.3987

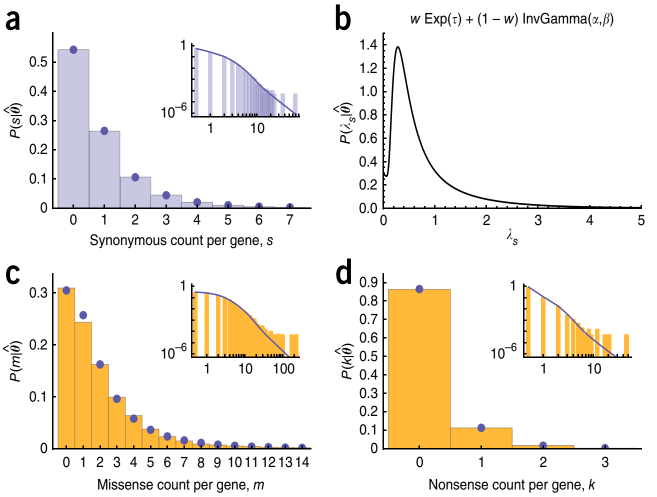

Figure 1 Observed and expected neutral distributions of mutation counts per gene. Histograms show the genome-wide distributions of the number of mutations per gene observed in each of the functional categories: synonymous (s), missense (m), and nonsense (k). Blue dots and blue lines mark the respective model distributions under the assumption of neutral evolution, $P(x|\hat{\theta})$, $x \in \{s, m, k\}$. The displayed distributions are for HNSC. (**a**) The synonymous class is fit according to equation (1) (sum over bins of squared deviations between observed and expected histograms 2×10^{-6}). (**b**) Distribution of per-gene expected values for the synonymous mutation count, λ_s, with estimated parameters $\hat{\theta}$. (**c**,**d**) Missense (**c**) and nonsense (**d**) class distributions. The missense mutation class shows a significant departure from neutrality ($P_{mis} = 5 \times 10^{-5}$, $P_{non} = 3 \times 10^{-1}$, χ^2 test). Negative selection causes deviations in the regimes with small m and k values, while positive selection causes deviations at large m and k values (insets show log–log scale).

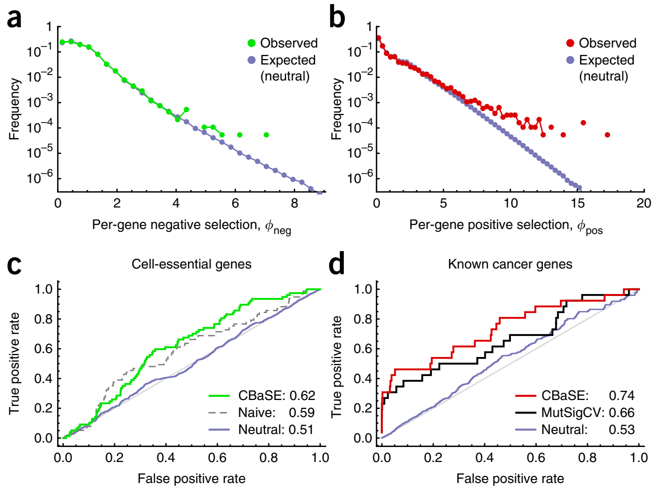

Figure 2 Per-gene inference of selection and validation. (**a**) Histogram of the meta-statistic measuring the per-gene negative selection signal ϕ_{neg} in real data (green) and from simulation under the neutral model (blue). (**b**) As in **a** for positive selection, ϕ_{pos} (real data in red). For illustration purposes, the x axis was terminated at $\phi_{pos} = 20$, and the following genes (with corresponding ϕ_{pos} values in parentheses) are therefore not shown: *EPHA2* (22.0), *PIK3CA* (26.5), *NOTCH1* (34.2), *CASP8* (35.2), *NSD1* (37.5), *FAT1* (39.2), *KMT2D* (44.7), and *TP53* ($p_m^{pos} = 0$); both distributions are shown for bins of width 0.3. (**c**) ROC curves of gene ranking according to the significance level q_{neg}, using as true positives 77 predicted cell-essential genes from Wang *et al.*[2] ($P_{AUC} = 2 \times 10^{-4}$). The dashed gray line corresponds to the ROC curve when using uninformed ranking based only on hypomutation in all mutational categories. (**d**) ROC curves of 26 genes described as causally implicated in HNSC by the CGC (red; $P_{AUC} = 4 \times 10^{-4}$). Predictions from the mutation rate covariate clustering algorithm MutSigCV[1] are shown in black. Blue lines in **c** and **d** correspond to mean ROC curves from neutral simulations (**Supplementary Note**). AUCs are given in the insets. All data shown are for HNSC.

of the conditions affecting mutations at a given gene, only assuming that every influencing local factor other than selection has a similar effect on synonymous and nonsynonymous variation (**Supplementary Fig. 1**; see the **Supplementary Note** for discussion).

Figure 1 shows the observed and expected neutral distributions in the three mutational categories—synonymous, missense (m), and nonsense (k) mutations—for head and neck squamous cell carcinoma (HNSC) (see **Supplementary Figs. 2–25** for other cancer (sub)types). The synonymous count distribution was precisely fit (**Fig. 1a**). The best model for the distribution $P(\lambda_s|\boldsymbol{\theta})$ in this cancer was a mixture of an exponential component and an inverse gamma distribution component (**Fig. 1b**; model 4 in the **Supplementary Note**). The nonsynonymous count distributions, although relatively closely matching the neutral expectation, showed deviations suggesting the action of selection (**Fig. 1c**,**d** and **Supplementary Table 1**). When we extended the summation over tumor types to a pan-cancer analysis, enabling inferences about selective pressures acting across the entire range of cancer etiologies, we found a substantial signature of positive selection (**Supplementary Fig. 26**).

In combination with the observed count of synonymous mutations at each gene, the distribution $P(\lambda_s|\boldsymbol{\theta})$ opens up a Bayesian route to the estimation of per-gene probabilities for missense and nonsense mutations under the neutral evolution hypothesis. From this, we assess the deficit or excess in observed missense and nonsense mutations jointly in statistics ϕ_{neg} and ϕ_{pos}, which quantify the strength of negative and positive selection, respectively, at the gene level. Briefly, ϕ is a meta-statistic of P values (from m and k) similar to that obtained with Fisher's method, whose value increases with increasing selection (Online Methods and **Supplementary Fig. 1**). The observed distributions of ϕ_{neg} and ϕ_{pos} as well as the expectation under neutrality from simulation are shown in **Figure 2a**,**b**. From the distributions of ϕ_{neg} and ϕ_{pos}, we derive q values q_{neg} and q_{pos}, respectively, for each gene in each cancer type, which control the false discovery rate at q and are given in **Supplementary Table 2**. We call this method cancer Bayesian selection estimation (CBaSE).

Although signals of positive selection have been the primary instrument in identifying cancer drivers, detection of negative selection in tumors is arguably of equal importance. Negative selection aimed at conservation of genetic material is the dominant evolutionary mode across biological systems. Surprisingly, however, it has proven almost impossible to detect in cancer, although recent findings point in a new direction[5–8]. In part, the elusive nature of negative selection can be explained by sparsity of mutation data, which results in lower statistical power to show a significant deficit of mutations in comparison to a mutational excess. However, if negative selection is an important factor in cancer evolution, it may determine the dynamics of tumor progression and identify cancer vulnerabilities in the form of genes the tumor cannot afford to lose[9].

Indeed, the genome-wide signal of negative selection appeared to be exceedingly weak in our analysis. In line with previous reports, we found it to be far less pronounced than the signal of positive selection on driver genes. However, we detected a significant negative selection signal in several cancer types and subtypes urothelial bladder carcinoma (BLCA), $P = 8 \times 10^{-3}$; colorectal cancer (CRC), $P = 1 \times 10^{-2}$; *POLE*-aberrant CRC (CRC_POLE), $P = 3 \times 10^{-2}$; melanoma (MEL), $P = 3 \times 10^{-6}$; uterine corpus endometrial carcinoma (UCEC), $P = 5 \times 10^{-8}$;

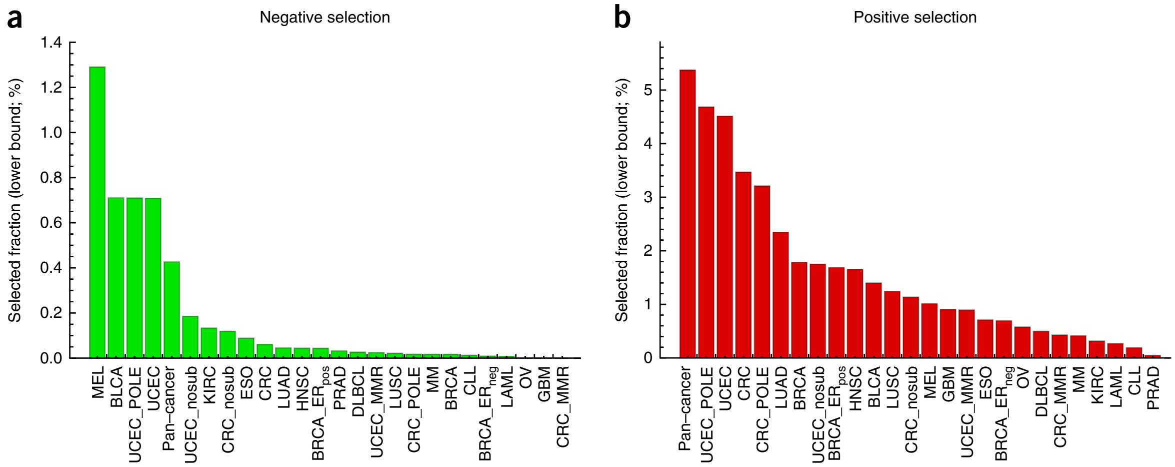

Figure 3 Genome-wide intensity of selection. (**a**,**b**) Lower bounds on the fractions of genes under negative (**a**) and positive (**b**) selection in each cancer (sub)type, inferred from the excess in the respective regime of large values of the meta-statistic ϕ. Standard error estimates derived from varying the binning scheme for ϕ are 0.017% and 0.020% for negative and positive selection, respectively (**Supplementary Note**). Abbreviations for all cancer types are defined in the Online Methods..

UCEC no subtypes (UCEC_nosub), $P = 2 \times 10^{-2}$; *POLE*-aberrant (UCEC_POLE), UCEC $P = 1 \times 10^{-13}$; χ^2 test; **Supplementary Table 3**), as well as in the pan-cancer data set ($P = 4 \times 10^{-4}$). Notably, these cancer types have high mutation densities, increasing the power of the analysis. Given that one of the potential causes of negative selection is the maintenance of genes that are responsible for basal cellular functions, we first compared CBaSE gene-specific predictions to the set of likely cell-essential genes from a cancer cell line knockout assay[2]. For 18 of the 25 cancer (sub)types and the pan-cancer data set, there was significant enrichment of our negative selection measure for cell-essential genes ($5 \times 10^{-6} \leq P_{\mathrm{AUC}} \leq 0.05$; **Fig. 2c** (HNSC), **Supplementary Figs. 2–25** and **27**, and **Supplementary Table 4**). Notably, rank-based measures like the area under the receiver operating characteristic (ROC) curve (AUC) can be affected by variability in statistical power across genes (P_{AUC} is corrected for this, and the neutral expectation is shown in the figures as a blue line; **Supplementary Note**).

Negative selection is also expected to act on neoantigens and oncogenes activated by noncoding or epigenetic driver events, expanding the scope of our predictions beyond cell essentiality[10,11]. Analysis of the cancer type with the strongest overall signal of negative selection, MEL, showed that, of the five most conserved genes, four have a reported oncogenic role in cancer (*MKL1*, *NPY5R*, *RMDN2*, and *DIAPH1*)[12–15]. Even in cancer types where we lacked power to detect a significant genome-wide signal of negative selection, strong biological candidates were found among the ten most highly protected genes across cancer types. For example, *BCL2* ($q_{\mathrm{neg}} = 0.03$), *BCL11B* ($q_{\mathrm{neg}} = 0.13$), and *PREX2* ($q_{\mathrm{neg}} = 0.14$), identified in diffuse large B cell lymphoma (DLBCL), chronic lymphocytic leukemia (CLL), and esophageal cancer (ESO), respectively, are either established apoptosis suppressors in the corresponding cancer types or tumor-suppressor inhibitors[16–18]. To concentrate on functionally important genes under negative selection, we selected the subset of genes with q_{neg} <0.6 ($n = 421$) that also showed significantly increased expression in tumor versus normal tissue in at least 50% of available cancer tissues or are likely cell essential (CRISPR score < 0 or GTS (gene-trap score) < 0.2 from Wang *et al.*[2]; **Supplementary Table 5**). Clustering this subset of 88 (21%) genes putatively indispensable in cancer by functional annotation using DAVID, we found enrichment for proteins involved in RNA processing (group enrichment score (ES; geometric mean of negative log-transformed *P* values in an annotation cluster) = 2.1; **Supplementary Table 6**). **Supplementary Table 7** shows the per-gene estimates of negative selection inferred from the pan-cancer analysis. Here, among 13 genes with q_{neg} <0.4, we found 6 (46%) cancer-promoting genes (*ATAT1*, *BCL2*, *CLIP1*, *GALNT6*, *CKAP5*, and *REV1*)[17,19–23].

Next, we applied CBaSE to detect cancer driver genes, the longstanding area of method development in cancer genomics. **Figure 2d** (HNSC) and **Supplementary Figures 2–25** and **27** show that the method had high sensitivity in recovering known cancer-type-specific genes listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC)[24] for the 17 original cancer types and 8 subtypes, as well as for the pan-cancer data set ($0 \leq P_{\mathrm{AUC}} \leq 0.04$; **Supplementary Table 4**; **Supplementary Figs. 28** and **29** show corresponding precision–recall curves). The AUC exceeded that of the covariate-clustering-based algorithm MutSigCV[1] (run with default mutation rate covariates) for HNSC and in aggregate (for 21 of 25 cancer (sub)types). The performance of CBaSE depends on the availability of synonymous mutations, and the four cancer types for which it currently does not outperform MutSigCV indeed had the fewest synonymous mutations (prostate adenocarcinoma (PRAD), CLL, acute myeloid leukemia (LAML), and multiple myeloma (MM)). To prioritize genes under positive selection further, we derived properties expected from cancer driver genes, such as clustering of missense alterations in 3D protein structures (according to the algorithm mutation3D) and aberrant expression in cancers (**Supplementary Table 8**).

Overall, of the 58 genes that had q_{pos} <0.002 in any of the cancer types, 74% ($n = 43$) are already listed in the CGC. Among the 15 remaining strongest novel driver candidates, we found 6 genes that have previously been associated with tumorigenesis (*ARHGAP35*, *TRAF3*, *EPHA2*, *AJUBA*, *RBL2*, and *MED23*)[4,25–29]. In the larger group of 452 non-CGC candidate driver genes with q_{pos} <0.1, 32% ($n = 144$) had either largely reduced expression in cancer tissues relative to primary cells or oncogenic features (largely elevated expression or missense mutation clusters with $P < 0.1$). These genes showed a strong enrichment for cell adhesion and cell morphogenesis (ES = 5.9 and 4.3, respectively; **Supplementary Table 9**), consistent with typical targets for positive selection in cancer. Genes with highly significant signals of positive selection from the pan-cancer analysis replicated many of the findings for individual cancer (sub)types (**Supplementary Table 10**). Of the 36 pan-cancer genes with $q_{\mathrm{pos}} = 0$, 35 (97%) are listed in the CGC. The remaining gene, *ARHGAP35*, has been implicated as a cancer driver[25].

From the distributions of ϕ shown in **Figure 2a**,**b** and **Supplementary Figures 2**–**25** and **27**, we inferred lower bounds on the genome-wide fractions of genes under negative and positive selection. The minimum fraction of negatively selected genes varied from 0 to 1.3% across individual cancer (sub)types, while positive selection acted on 0.05 to 4.7% of genes (**Fig. 3** and **Supplementary Table 3**). In the pan-cancer data set, at least 5.4% of genes appeared to be under positive selection and 0.4% of genes appeared to be under negative selection, which is broadly compatible with findings from a recent study[8].

Unlike in most other biological systems, detection of negative selection in cancer is a formidable task. Besides the fact that negative selection coefficients may on average be small because of the dispensability of most of the gene repertoire for cancer cell survival, their impact of negative fitness effects is also often weakened by strongly selected driver mutations drawing deleterious passenger mutations to fixation[9,30,31]. Also, most mutations in our analysis are heterozygous, and many cell-essential genes seem to be haplosufficient[2,7]. Conversely, positive selection on synonymous variation in oncogenes would contribute to the signal[32]. Lastly, detectability of significant hypomutation is constrained by mutation data availability, necessitating fine-tuned approaches to separate signal from noise. The inference of selection is generally also complicated by the possible presence of unfiltered germline variants in cancer sequencing data, statistical uncertainty in the inferred context-dependent mutation models, and heterogeneity in the context dependence of mutations along the genome (**Supplementary Figs. 30** and **31**, and **Supplementary Note**). With the rapidly growing number and size of cancer sequencing data sets, hierarchical models can be adopted to incorporate the uncertainty for many underlying biological and technical variables. These methods may boost power for cancer driver detection and identify negative selection to help find genes indispensable for cancer.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Schiffels and L. Mirny for helpful discussion during the conception of the project and feedback on the manuscript, as well as P. Polak and Z. Li for critical reading of the manuscript. We also thank V. Seplyarskiy for advice on mutational processes and I. Adzhubey for support with the setup of the web interface for CBaSE. This work was supported by US National Institutes of Health (NIH) grants U54 CA143874, R01 MH101244, and R01 GM078598 (S.S.).

AUTHOR CONTRIBUTIONS

D.W. designed the statistical framework, wrote code, analyzed and interpreted data, created the web interface, and wrote the manuscript. S.S. supervised the project, gave technical and conceptual advice, and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
2. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
3. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
4. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
5. Rooney, M.S., Shukla, S.A., Wu, C.J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
6. Lindeboom, R.G.H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
7. Van den Eynden, J., Basu, S. & Larsson, E. Somatic mutation patterns in hemizygous genomic regions unveil purifying selection during tumor evolution. *PLoS Genet.* **12**, e1006506 (2016).
8. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. Preprint at *bioRxiv* http://dx.doi.org/10.1101/132324 (2017).
9. McFarland, C.D., Korolev, K.S., Kryukov, G.V., Sunyaev, S.R. & Mirny, L.A. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. USA* **110**, 2910–2915 (2013).
10. Melton, C., Reuter, J.A., Spacek, D.V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (2015).
11. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
12. Scharenberg, M.A., Chiquet-Ehrismann, R. & Asparuhova, M.B. Megakaryoblastic leukemia protein-1 (MKL1): increasing evidence for an involvement in cancer progression and metastasis. *Int. J. Biochem. Cell Biol.* **42**, 1911–1914 (2010).
13. Sheriff, S. *et al.* Neuropeptide Y Y5 receptor promotes cell growth through extracellular signal–regulated kinase signaling and cyclic AMP inhibition in a human breast cancer cell line. *Mol. Cancer Res.* **8**, 604–614 (2010).
14. Law, M.H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat. Genet.* **47**, 987–995 (2015).
15. Nürnberg, A., Kitzing, T. & Grosse, R. Nucleating actin for invasion. *Nat. Rev. Cancer* **11**, 177–187 (2011).
16. Grabarczyk, P. *et al.* Inhibition of BCL11B expression leads to apoptosis of malignant but not normal mature T cells. *Oncogene* **26**, 3797–3810 (2007).
17. Adams, J.M. & Cory, S. The Bcl-2 apoptotic switch in cancer development and therapy. *Oncogene* **26**, 1324–1337 (2007).
18. Fine, B. *et al.* Activation of the PI3K pathway in cancer through inhibition of PTEN by exchange factor P-REX2a. *Science* **325**, 1261–1265 (2009).
19. Boggs, A.E. *et al.* α-Tubulin acetylation elevated in metastatic and basal-like breast cancer cells promotes microtentacle formation, adhesion, and invasive migration. *Cancer Res.* **75**, 203–215 (2015).
20. Bilbe, G. *et al.* Restin: a novel intermediate filament–associated protein highly expressed in the Reed–Sternberg cells of Hodgkin's disease. *EMBO J.* **11**, 2103–2113 (1992).
21. Park, J.H. *et al.* Critical roles of mucin 1 glycosylation by transactivated polypeptide *N*-acetylgalactosaminyltransferase 6 in mammary carcinogenesis. *Cancer Res.* **70**, 2759–2769 (2010).
22. Fielding, A.B., Lim, S., Montgomery, K., Dobreva, I. & Dedhar, S. A critical role of integrin-linked kinase, ch-TOG and TACC3 in centrosome clustering in cancer cells. *Oncogene* **30**, 521–534 (2011).
23. Xie, K., Doles, J., Hemann, M.T. & Walker, G.C. Error-prone translesion synthesis mediates acquired chemoresistance. *Proc. Natl. Acad. Sci. USA* **107**, 20792–20797 (2010).
24. Futreal, P.A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
25. Biname, F. *et al.* Cancer-associated mutations in the protrusion-targeting region of p190RhoGAP impact tumor cell migration. *J. Cell Biol.* **214**, 859–873 (2016).
26. Baud, V. & Karin, M. Is NF-κB a good target for cancer therapy? Hopes and pitfalls. *Nat. Rev. Drug Discov.* **8**, 33–40 (2009).
27. Pasquale, E.B. Eph receptors and ephrins in cancer: bidirectional signalling and beyond. *Nat. Rev. Cancer* **10**, 165–180 (2010).
28. Tanaka, I. *et al.* LIM-domain protein AJUBA suppresses malignant mesothelioma cell proliferation via Hippo signaling cascade. *Oncogene* **34**, 73–83 (2015).
29. Ullah, F. *et al.* Promoter methylation status modulate the expression of tumor suppressor (RbL2/p130) gene in breast cancer. *PLoS One* **10**, e0134687 (2015).
30. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756 (1974).
31. Good, B.H. & Desai, M.M. Deleterious passengers in adapting populations. *Genetics* **198**, 1183–1208 (2014).
32. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).

ONLINE METHODS

Data sets. We analyzed 17 cancer types with a total of 4,476 patients from the published Tumor Portal data set[33] (http://www.tumorportal.org/): urothelial bladder carcinoma (BLCA), breast invasive carcinoma (BRCA), chronic lymphocytic leukemia (CLL), colorectal cancer (CRC), diffuse large B cell lymphoma (DLBCL), esophageal cancer (ESO), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), acute myeloid leukemia (AML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), melanoma (MEL), multiple myeloma (MM), ovarian cancer (OV), prostate adenocarcinoma (PRAD), and uterine corpus endometrial carcinoma (UCEC). Four additional cancer types (carcinoid, medulloblastoma, neuroblastoma, and rhabdoid tumors) were omitted because of their low overall mutation counts. Breast (BRCA), colorectal (CRC), and endometrial (UCEC) cancers were further subdivided into common subtypes: ER-positive ($\mathrm{ER}_{\mathrm{pos}}$) and ER-negative ($\mathrm{ER}_{\mathrm{neg}}$) cases for breast cancer, and for colorectal and endometrial cancers polymerase ε–aberrant (POLE) and mismatch-repair-deficient (MMR) subtypes, and samples belonging to neither of these categories (nosub) (**Supplementary Table 11**).

The Tumor Portal data set is filtered for germline variants[33]. To quantify potential contamination with germline polymorphism in more detail, we compared the expected and observed fractions of somatic mutations overlapping with known SNPs (**Supplementary Note**). We found a high correspondence, suggesting that germline contamination (or overfiltering) in the data set is very low (**Supplementary Fig. 30**). Exonic mutations that were not located in one of 575 known cancer-related genes in the CGC[24] (v80) or in one of 77 likely cell-essential genes from Wang *et al.*[2] were used in the derivation of the nucleotide-context-dependent mutation signature matrix, which determines the expected number of mutations under the null model of neutral evolution. We used a set of 18,666 genes from the UCSC knownGene track, omitting olfactory receptor genes, which cluster anomalously around a gene length of 945 bp (**Supplementary Table 12**).

Predicted negatively selected genes were compared to likely cell-essential genes[2] (**Supplementary Table 13**). For a naive ranking of cell-essential genes according to mutation paucity across all mutation classes (m,k,s), we find PAUC < 0.05 (≤0.002) for 6 (2) of the 25 cancer (sub)types, showing that enrichment of the negative selection signal with cell-essential genes is not driven by general hypomutation (**Supplementary Table 4**). To assess the sensitivity of our positive selection estimates, we tested prediction of cancer-type-specific causally implicated genes from the CGC (**Supplementary Table 14**). We compared the sensitivity to that of the algorithm MutSigCV[1] (v1.2), which is based on clustering genes with respect to mutation rate covariates (**Supplementary Table 15**). In addition to the ROC curves shown in **Figure 2** and **Supplementary Figures 2–25** and **27**, we derived precision–recall curves, which are presented in **Supplementary Figures 28** and **29**. Prediction of clustering for missense alterations was performed using the algorithm mutation3d[34] (http://www.mutation3d.org/) with preset cluster criteria. Tumor–normal gene expression levels were obtained from the TCGA Expression Browser (https://tools.altiusinstitute.org/tcga/) for 22 cancer types. Functional annotation clustering of Gene Ontology biological process terms (GOTERM_BP_FAT) was performed using DAVID[35] (v6.7; http://david.ncifcrf.gov/) with default clustering criteria.

Expected distribution of synonymous mutation counts. Assuming that the synonymous class of mutations evolves neutrally, we model the observed count of synonymous mutations per gene s as a Poisson random variable with expectation λ_s. This expectation is dependent on the synonymous mutation rate at the gene locus, gene length, and the tumor lifetime in generations (**Supplementary Note**). The aim of our probabilistic approach is to infer $P(\lambda_s|\boldsymbol{\theta})$, that is, the distribution of λ_s, across the entire ensemble of genes. Importantly, $P(\lambda_s|\boldsymbol{\theta})$ by construction captures genome-wide mutation rate heterogeneity that is due to, for example, replication timing, expression level, chromatin state, tumor subclonality, and purity or ploidy of the sample, and hence does not depend on explicit knowledge thereof (**Supplementary Table 16**). Summing all synonymous mutations per gene locus over all tumors of a given cancer type, we can write the expected distribution of s as in equation (1). We then fit this neutral model to the observed distribution of the synonymous count per gene s to estimate the parameters $\boldsymbol{\theta}$ of the genome-wide distribution of the expected number of synonymous mutations, $P(\lambda_s|\boldsymbol{\theta})$ (**Fig. 1a**).

Parametric form of $P(\lambda_s|\boldsymbol{\theta})$. The functional form of $P(\lambda_s|\boldsymbol{\theta})$ is not known a priori; however, the main parameter to affect the expected number of synonymous mutations for a given gene is the synonymous target size, which in turn is proportional to the coding-sequence length. Furthermore, we cannot assume that all genes were necessarily covered in the DNA sequencing of a given cancer type, resulting in an inflated weight of λ_s around zero. This motivates the space of parametric functions we explore to describe the genome-wide distribution of λ_s. Briefly, we test six linear combinations of gamma and inverse gamma distributions (**Supplementary Note**). The optimal functional form for each cancer type is then selected from the AIC-penalized log likelihoods. We found that all 17 cancer types could be fit by a linear combination with an inverse gamma distribution component (**Fig. 1b** and **Supplementary Figs. 2–26**).

Expected distribution of nonsynonymous mutation counts under neutrality. The inferred $P(\lambda_s\,|\,\hat{\boldsymbol{\theta}})$ enables derivation of the corresponding distributions in the nonsynonymous mutation categories expected under neutral evolution, which in turn permit detection of signatures of selection. We derive the distributions for missense (m) and nonsense (k) mutation counts from a rescaling of the synonymous Poisson parameter λ_s by the ratio of nonsynonymous to synonymous target size, r_x (**Fig. 1c**,**d**).

$$P(x\,|\,\hat{\boldsymbol{\theta}};r_x)=\int \mathrm{d}\lambda_s\,\mathrm{Pois}(\lambda_s r_x)\,P(\lambda_s\,|\,\hat{\boldsymbol{\theta}}),x\in\{m,k\} \qquad (2)$$

The target size for a given gene depends on the coding-sequence composition in conjunction with the cancer-type-specific mutational signature, which we infer from observed coding-sequence mutations outside of the set of likely selected genes (**Supplementary Note**).

Per-gene selection inference. To quantify selection at the level of individual genes, using $P(\lambda_s\,|\,\hat{\boldsymbol{\theta}})$, we derive a Bayesian expression for the posterior probability of the observed nonsynonymous mutation count x given the observed s.

$$P(x\,|\,s;r_x)=\int \mathrm{d}\lambda_s\,P(x\,|\,\lambda_s;r_x)\,P(\lambda_s|\,s) \qquad (3)$$

This allows corresponding P values to be computed for each gene, where negative selection is quantified by hypomutation in either nonsynonymous category (p_m^{neg} and p_k^{neg}) and positive selection in driver genes is marked by hypermutation (p_m^{pos} and p_k^{pos}) (**Supplementary Note**). We combine the P values from the observed missense and nonsense counts for each selection scenario into a meta-statistic ϕ, analogous to Fisher's method.

$$\begin{aligned}\phi_{\mathrm{neg}} &= -\log p_k^{\mathrm{neg}} - \log p_m^{\mathrm{neg}}\\ \phi_{\mathrm{pos}} &= -\log p_k^{\mathrm{pos}} - \log p_m^{\mathrm{pos}}\end{aligned} \qquad (4)$$

Large values of ϕ indicate small P values and hence strong selection. Using simulated count data generated under the null model of neutral evolution, we then compute the significance of the observed ϕ statistic, resulting in q values for each gene in each selection category (negative or positive), which control the false discovery rate at q.

Inference of the selected fraction and significance of the genome-wide selection signal. The observed excess in the distributions of the meta-statistics ϕ_{pos} and ϕ_{neg} is used to derive a lower bound on the genome-wide fractions of genes under positive and negative selection, respectively. To this end, we compute the difference between the cumulative probabilities contained in the large-ϕ tail of the observed and simulated distributions, making this a conservative estimate. To assess the significance of the genome-wide signals of selection, we perform a χ^2 goodness-of-fit test on the region defined by the highest percentile of the simulated ϕ values in each cancer type (**Supplementary Table 3**).

Code availability. A user interface that allows the cBaSE method to be run with user-provided somatic mutation data can be found at http://genetics.bwh.harvard.edu/cbase. In addition, the code can be downloaded from the same site.

doi:10.1038/ng.3987

Data availability. Data from this study are available from the authors upon reasonable request. A **Life Sciences Reporting Summary** is available.

33. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
34. Meyer, M.J. *et al.* mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome. *Hum. Mutat.* **37**, 447–456 (2016).
35. Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

nature research

Corresponding author(s): Shamil Sunyaev

□ Initial submission □ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

The sample size was given by the published cancer sequencing data set that was used.

e

2. Data exclusions

Describe any data exclusions.

Four out of 21 sequenced cancer types were not analyzed due to insufficient mutation data, which were needed to derive an estimate of the mutational signature. This criterion was pre-established before any downstream analyses.

3. Replication

Describe whether the experimental findings were reliably reproduced.

There are no experimental findings presented in the paper.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

No experiments were carried out.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

No allocation of groups took place.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	□	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☒	□	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	A statement indicating how many times each experiment was replicated
□	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
□	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
□	☒	The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
☒	□	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☒	□	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this stu

The software developed as part of this study will be available as a web interface tool and for download.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

n/a

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

n/a

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

n/a

b. Describe the method of cell line authentication used.

n/a

c. Report whether the cell lines were tested for mycoplasma contamination.

n/a

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

n/a

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

n/a

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

n/a