CANCER

A machine learning approach for somatic mutation discovery

Derrick E. Wood[1], James R. White[1], Andrew Georgiadis[1], Beth Van Emburgh[1], Sonya Parpart-Li[1], Jason Mitchell[1], Valsamo Anagnostou[2], Noushin Niknafs[2], Rachel Karchin[2,3], Eniko Papp[1], Christine McCord[1], Peter LoVerso[1], David Riley[1], Luis A. Diaz Jr.[4], Siân Jones[1], Mark Sausen[1], Victor E. Velculescu[2]*, Samuel V. Angiuoli[1]*

Variability in the accuracy of somatic mutation detection may affect the discovery of alterations and the therapeutic management of cancer patients. To address this issue, we developed a somatic mutation discovery approach based on machine learning that outperformed existing methods in identifying experimentally validated tumor alterations (sensitivity of 97% versus 90 to 99%; positive predictive value of 98% versus 34 to 92%). Analysis of paired tumor-normal exome data from 1368 TCGA (The Cancer Genome Atlas) samples using this method revealed concordance for 74% of mutation calls but also identified likely false-positive and false-negative changes in TCGA data, including in clinically actionable genes. Determination of high-quality somatic mutation calls improved tumor mutation load–based predictions of clinical outcome for melanoma and lung cancer patients previously treated with immune checkpoint inhibitors. Integration of high-quality machine learning mutation detection in clinical next-generation sequencing (NGS) analyses increased the accuracy of test results compared to other clinical sequencing analyses. These analyses provide an approach for improved identification of tumor-specific mutations and have important implications for research and clinical management of cancer patients.

INTRODUCTION

Comprehensive molecular profiling of cancer through next-generation sequencing (NGS) approaches is increasingly used in oncology for diagnostic and therapeutic management decisions (*1*–*8*). Tumor-specific (somatic) mutations, including single-nucleotide alterations and small insertions or deletions, are known to affect key driver genes early during tumorigenesis. Over time, cancers accumulate additional mutations that may influence the underlying biology of the tumor cells, representing new opportunities for therapeutic intervention.

Cancer therapies targeting specific driver genes altered in tumors are now commonly used in the clinic (*9*, *10*). Examples include vemurafenib and trametinib in BRAF-mutated melanoma (*11*), and erlotinib and osimertinib in epidermal growth factor receptor (EGFR)–mutated non–small cell lung cancer (NSCLC) (*12*, *13*). The immune checkpoint inhibitor pembrolizumab recently received accelerated approval by the U.S. Food and Drug Administration (FDA) for treatment of patients with solid tumors from any tissue type where the tumor was determined to have mismatch repair deficiency (dMMR) resulting in microsatellite instability (MSI-H) and higher mutation loads (*14*). High tumor mutation burden, resulting from repair defects or exposure to mutagens, has been associated with durable clinical response to a variety of immune checkpoint inhibitors, including nivolumab, ipilimumab, and atezolizumab (*15*–*19*), and may serve as a predictive biomarker for treatment response (*20*).

Systematic discovery of somatic alterations in new driver genes and pathways began with large-scale sequencing analyses in various human cancers using Sanger sequencing and NGS (*21*–*30*). These efforts were extended by The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium, which profiled the molecular landscape of 33 common cancer types and more than 10,000 patients (*31*, *32*). Because of the scale and complexity of TCGA, initial analysis efforts to characterize somatic mutations within tumors were typically unique to each cancer type (*33*–*42*). More recently, the PanCanAtlas project has generated a unified set of consensus mutation calls (Multi-Center Mutation Calling in Multiple Cancers or "MC3") across the entire TCGA cohort (*43*), although little, if any, validation of mutations has been performed for these data sets.

On the basis of these large-scale genomic efforts, targeted NGS approaches in clinical oncology have begun to be rapidly adopted to identify genetic alterations and make decisions regarding patient therapy and management (*3*, *5*, *44*–*47*). Hundreds of laboratories in the United States provide NGS cancer profiling, issuing tens of thousands of reports each year. Nearly all NGS-based analyses for clinical cancer mutation detection are Clinical Laboratory Improvement Amendments/College of American Pathologists (CLIA/CAP)–regulated laboratory-developed tests and have not been approved by the FDA for use as in vitro diagnostics for analyses across tumor types (*48*–*51*). The importance of analyzing matched tumor and normal sequencing from the same individual has been reported for improved identification of somatic and germline alterations in cancer patients (*5*, *52*, *53*), although few laboratories providing laboratory-developed tests use these approaches. Despite the development of recommendations for validation of NGS testing (*54*), many challenges remain in somatic mutation detection, including sensitive detection of alterations that are subclonal or in low-purity tumor samples, as well as distinguishing these from germline changes or from artifacts related to polymerase chain reaction (PCR) amplification or sequencing. Head-to-head comparisons of laboratory-developed NGS tests have reported a wide range of mutation concordance estimates, leading to concern regarding the accuracy of such tests (*55*–*58*).

Here, we describe the development of a novel method for somatic mutation identification that uses machine learning strategies to

[1]Personal Genome Diagnostics, Baltimore, MD 21224, USA. [2]The Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA. [3]Department of Biomedical Engineering, Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD 21218, USA. [4]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.
*Corresponding author. Email: velculescu@jhmi.edu (V.E.V.); angiuoli@personalgenome.com (S.V.A.)

optimize sensitivity and specificity for detection of true alterations. We compared the overall accuracy of this approach to existing methods for somatic mutation identification using simulated and experimentally validated whole-exome and targeted gene analyses. We evaluated the overall concordance of our method with mutation calls from TCGA exomes, examined the underlying causes of erroneous calls, including those in actionable driver genes, and assessed the effects of discordant mutation calls on tumor mutational burden (TMB) and clinical response to cancer immunotherapy. To assess the importance of high-quality mutation analysis in clinical testing, we performed head-to-head comparisons of clinical sequencing with or without our machine learning approach. Overall, these analyses highlight the importance of improved somatic mutation detection for the interpretation of large-scale genome studies and for the application of these approaches to clinical practice.

RESULTS

Overview of approach

We created a method for analyzing next-generation cancer sequence data, called Cerebro, that uses machine learning to identify high-confidence somatic mutations while minimizing false positives (Fig. 1). Cerebro uses a specialized random forest classification model that evaluates a large set of decision trees to generate a confidence score for each candidate variant (Materials and Methods). The model was trained using a normal peripheral blood DNA sample, where exome regions were captured and sequenced twice using NGS methods. More than 30,000 somatic variants comprising substitutions, insertions, and deletions at mutant allele fractions from 1.5 to 100% were introduced in silico into one set of NGS data to provide the classifier with a training set of "tumor-specific" mutations as well as a real-world representative source of more than 2 million NGS errors and artifacts that

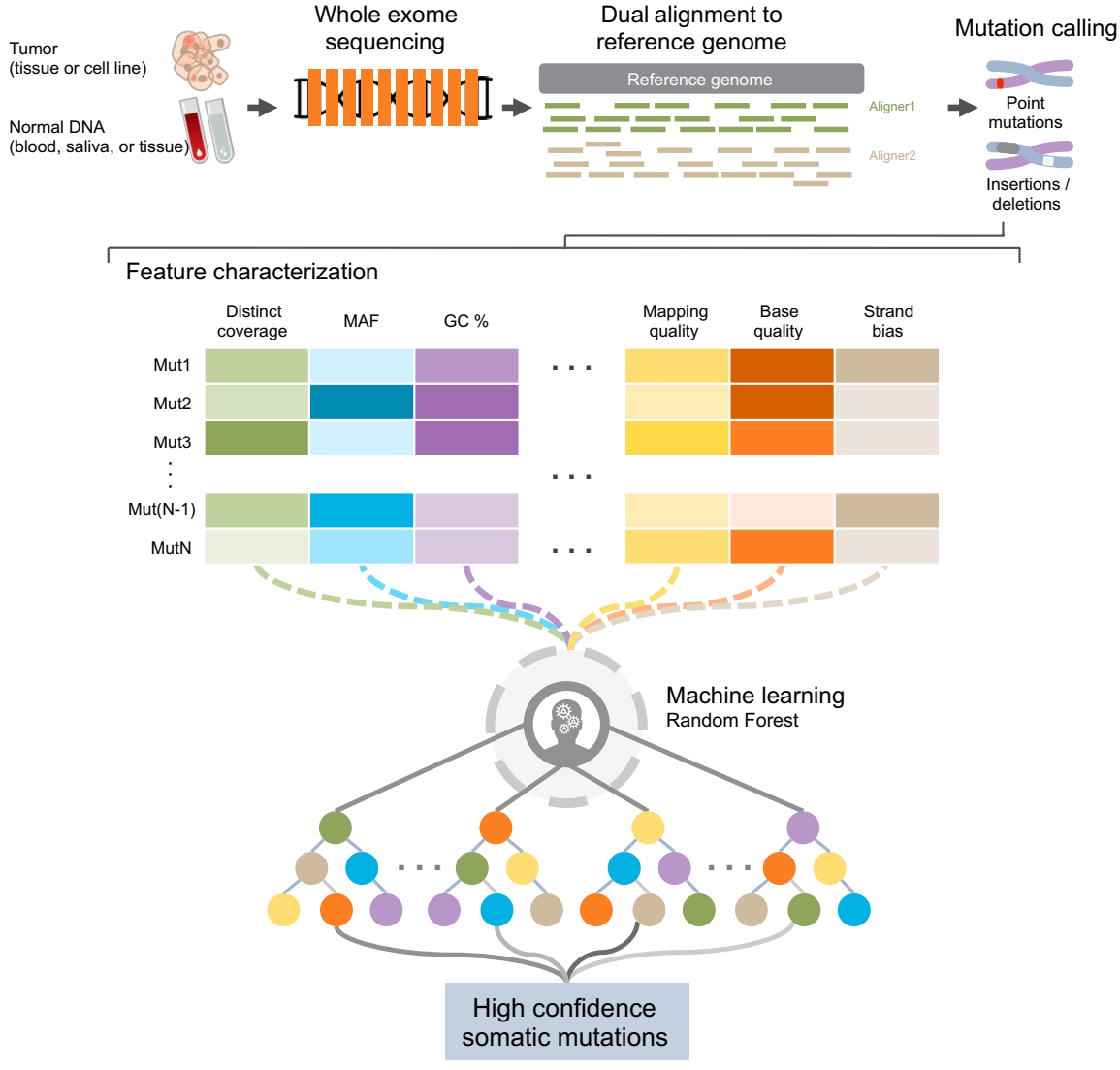

Fig. 1. Overview of Cerebro for somatic mutation detection. Paired tumor-normal whole-exome sequence data were mapped to the human reference genome using a dual alignment protocol for consensus mutation calling. Candidate mutations were assessed for confidence using an extremely randomized trees classification model (Cerebro) that evaluates a large set of decision trees to generate a confidence score for each candidate variant. A variety of characteristics are considered by the Cerebro model including distinct coverage, mutant allele frequency (MAF), GC content, and mapping quality.

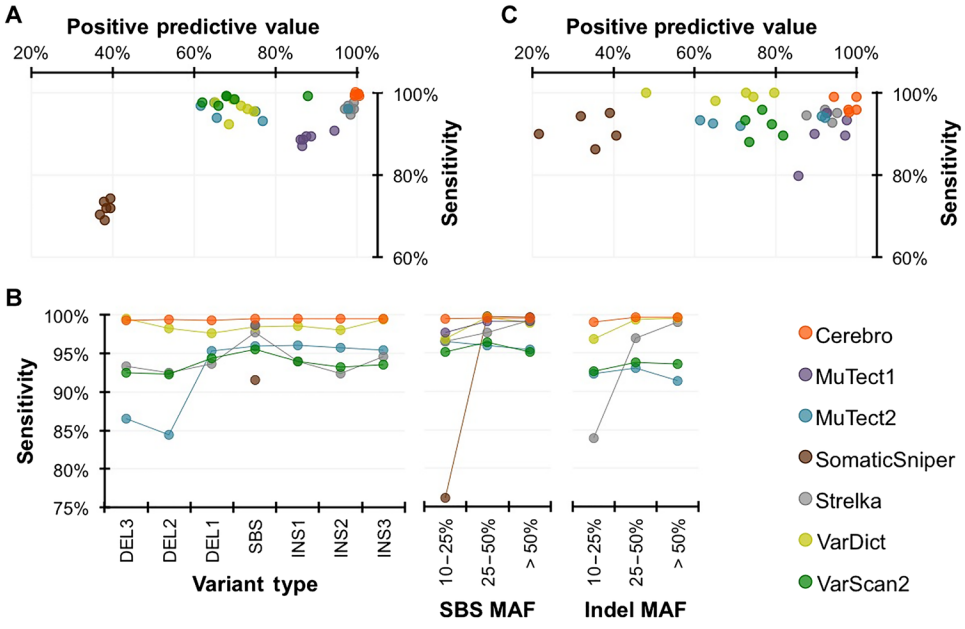

Fig. 2. Comparison of Cerebro to mutation detection methods. (**A**) Positive predictive value versus sensitivity for simulated low-purity tumor data sets created from normal cell line sequence data. (**B**) Sensitivity stratified by mutation type, calculated from simulated mutations across mutation types and allele frequencies. (**C**) Positive predictive value versus sensitivity for cell line data sets with experimentally validated somatic mutations. DEL, deletion; INS, insertion; 1, 2, or 3, length of indel.

might otherwise be mislabeled as variants. The second NGS sequence data set from the same sample was used as the matched normal, and the combined data sets were analyzed for the detection of somatic variants.

The manner in which this training set was created is an integral part of the machine learning approach that we present here. With this training set, we provided our classifier a representative set of experimentally obtained artifactual changes and germline alterations across the exome, along with a large number of true in silico mutations comprising a wide spectrum of allele frequencies and genomic contexts. Adding in silico mutations allowed us to include training data in regions where experimentally obtained alterations would not have provided sufficient sensitivity across the exome.

We considered more than 300 features that might optimize performance for identifying true somatic variants, ultimately selecting 15 feature categories from two separate alignment programs that included alignment characteristics (mapping quality and mismatches), sequence quality information (coverage and base quality), and information related to specific alterations (allele frequency, nearby sequence complexity, and presence of alteration in matching normal specimen) (table S1). Once implemented, Cerebro used 1000 decision trees for analysis of each mutation, with each tree evaluating a unique combination of the selected information supporting a candidate variant. The resultant confidence score from the Cerebro model represented the proportion of decision trees that would classify a candidate variant as somatic.

Evaluation of mutation calling accuracy

To systematically assess the accuracy of Cerebro for mutation detection, we designed a series of validation studies using simulated and experimental cancer exomes and evaluated the performance of this approach as compared to existing software tools commonly used for somatic variant identification in research and clinical genomic analyses (Fig. 2A and table S2) (*59*–*63*). We performed three studies using a set of six normal cell lines with NGS exome data. Two of those studies contained simulated mutations, with in silico somatic mutations spiked in to a normal DNA sample. In our first study, we simulated low-purity tumors by incorporating 132 coding somatic mutations [120 single-base substitutions (SBSs), and 12 insertions and deletions (indels)] with mutant allele frequencies ranging from 10 to 25% in the exome data (Fig. 2A). To characterize false-positive rates for the various tools, we analyzed technical replicate exome pairs of the six normal cell lines (fig. S1 and table S3). We additionally created simulated exomes with in silico spike-ins of 7000 somatic SBS and indel changes with variable mutant allele frequencies (ranging from 10 to 100%), comprising a total of 42,000 somatic changes across the six samples that could be detected through these approaches (Fig. 2B). This last study allowed us to examine the sensitivity of the various tools for specific mutation types and allele frequencies. In all cases, the location, type, and level of in silico alterations were different from those used in the training of the Cerebro algorithm. Overall, we observed substantial variation in sensitivity and positive predictive value among the tested variant classification programs (Fig. 2, A and B, and tables S4 to S6). Cerebro maintained the highest level of sensitivity and positive predictive value, whereas other methods resulted in moderate to high false-positive rates (Fig. 2A and tables S4 to S6). False-positive calls by other methods were frequently associated with indicators such as poor tumor/normal coverage, low mutant base quality, and low mutant mapping quality (fig. S1).

To assess the mutation detection performance of Cerebro using independently obtained experimental data, we next analyzed five matched tumor and normal specimens for which somatic mutations had been previously identified and validated through independent whole-exome sequencing (WES) (*28*, *30*). These previous analyses carefully evaluated the entire coding sequences of the samples through PCR amplification of 173,000 coding regions and Sanger sequencing of the amplification products (*28*, *30*). Any observed alteration was resequenced in the tumor and normal sample to confirm its tumor origin. Because the Sanger sequencing analyses were designed to identify only clonal or near-clonal alterations, we supplemented the Sanger-validated alterations previously observed in this set of samples ($n = 314$) with additional bona fide changes that had either been identified by a consensus of multiple NGS mutation callers ($n = 163$) or detected by up to two mutation callers and validated using droplet digital PCR (ddPCR) ($n = 18$; fig. S2 and table S7), a highly sensitive method for detection of alterations in a subset of DNA molecules (*64*). Comparison of results from all mutation callers with this reference set of alterations revealed that Cerebro had the highest overall accuracy compared to other methods (Fig. 2C, fig. S3, and table S8).

Evaluation of tumor exomes from TCGA

We assessed whether the improved capabilities of Cerebro could be used to increase the accuracy of mutation calling in large-scale cancer

Table 1. WES analyses and somatic mutation loads.

Tumor type*	Number of samples	Cerebro					TCGA MC3				
		Median mutation load	% Load > 100	% Load > 250	% Load > 500	% Load > 1000	Median mutation load	% Load > 100	% Load > 250	% Load > 500	% Load > 1000
Lung adenocarcinoma (LUAD)	473	141	60.3	29.4	8.9	2.1	152	62.8	31.3	10.1	2.1
Lung squamous cell (LUSC)	134	178	82.8	27.6	6.7	1.5	188	84.3	34.3	7.5	1.5
Bladder (BLCA)	360	136	64.4	23.1	5.6	1.1	141.5	65.0	23.9	5.8	1.1
Enriched for high–mutation load tumors											
Liver (LIHC-H)	141	75	25.5	3.5	1.4	0.0	72	24.1	3.5	1.4	0.7
Melanoma (SKCM-H)	105	308	70.5	57.1	33.3	9.5	317	71.4	58.1	35.2	11.4
Colon (COAD-H)	44	786.5	88.6	65.9	63.6	34.1	892.5	88.6	61.4	59.1	45.5
Uterine (UCEC-H)	41	78	46.3	46.3	26.8	0.0	78	46.3	46.3	22.0	0.0
Kidney renal clear cell (KIRC-H)	25	64	8.0	0.0	0.0	0.0	62	0.0	0.0	0.0	0.0
Head and neck (HNSC-H)	15	493	100.0	86.7	46.7	13.3	486	100.0	93.3	46.7	13.3
Stomach (STAD-H)	6	998	100.0	100.0	100.0	50.0	872.5	100.0	100.0	100.0	50.0
Sarcoma (SARC-H)	5	540	100.0	100.0	60.0	0.0	522	60.0	60.0	60.0	0.0
Other-H	19	734	84.2	73.7	68.4	36.8	668	94.7	84.2	68.4	36.8

*-H, tumor samples enriched for high mutation load.

genome sequencing efforts including TCGA, because these serve as the basis for various research efforts in human cancer. We used Cerebro to analyze paired tumor-normal exomes from 1368 patients in TCGA, focusing on tumors that would be relevant for both targeted therapies and immunotherapy. This set consisted of all available patients with non–small cell lung adenocarcinoma, non–small cell lung squamous cell carcinoma, or bladder urothelial carcinoma, as well as selected patients with higher mutational loads that had melanoma or colorectal, gastric, head and neck, hepatocellular, renal, or uterine cancer (Table 1). A total of 365,539 somatic alterations were identified, with an average of 267 somatic alterations per tumor (range, 1 to 5871).

The total number of somatic mutations measured across the various tumor types was largely similar to previous analyses of TCGA exomes, indicating that estimated mutational loads derived from Cerebro were consistent and representative of the TMB (Fig. 3A and fig. S4). We found a significant positive correlation between mutation loads called by Cerebro and the TCGA PanCanAtlas MC3 method (*43*) that uses consensus calls among seven different mutation callers (Pearson correlation coefficient = 0.93, $P < 0.0001$; Fig. 3A), with 74.0% of somatic mutations shared between the two approaches. However, we found that 10.3% of calls detected by Cerebro ($n = 44{,}439$) were apparently missed by TCGA, whereas 15.7% of alterations identified by TCGA ($n = 68{,}138$) were not considered somatic alterations by Cerebro (Fig. 3B and table S9). Individual tumors that were reanalyzed by Cerebro had mutation loads that differed by as many as 390 (95%) fewer or 729 (800%) more alterations compared to original calls. Comparison of detected alterations between Cerebro and other TCGA call sets [MC3, MuTect2 (*59*), and FIREHOSE original calls] showed increasing concordance with Cerebro calls from MuTect2 (least concordant; average, 60.2%) to MC3 (most concordant; average, 75.8%) (figs. S4 and S5). These observations are consistent with our analyses of individual mutation callers (Fig. 2) and support the notion that the use of multiple approaches in MC3 is likely to reduce the false-positive observations resulting from individual methods but may still have additional errors compared to Cerebro.

To more carefully evaluate discordant alterations in TCGA, we investigated somatic mutation calls for a subset of 66 well-characterized cancer driver genes (table S10). Of the 1368 evaluated tumors, 1257 (92%) had a mutation in this gene set, with 4037 shared somatic mutations between TCGA and Cerebro, and a total of 777 alterations that were discordant between the analyses. Further examination revealed that most of the 429 mutations called only by TCGA were associated with poor sequence quality and alignment issues that were likely to not represent bona fide alterations (Fig. 4A). Among driver genes associated with FDA-approved therapies or ongoing clinical trials (Fig. 4B), we found low-confidence TCGA mutations in 211 (16.8%) patients analyzed. We found that mutations uniquely called by TCGA were of significantly lower confidence than mutations observed by both platforms ($P < 0.0001$; Fig. 4C) or those that were uniquely detected by Cerebro ($P < 0.0001$; Fig. 4C), as determined by measures of sequence and alignment quality at those positions as well as the presence of the alterations in the matching normal sample. Analysis of COSMIC (Catalogue Of Somatic Mutations In Cancer) hotspot mutations in these genes identified 38 alterations that were missed by TCGA analyses, representing 4.3% of all COSMIC hotspots analyzed (table S10). The alterations that were missed included those in 15 patients with *KRAS* hotspots at codons 12 and 13 that had an

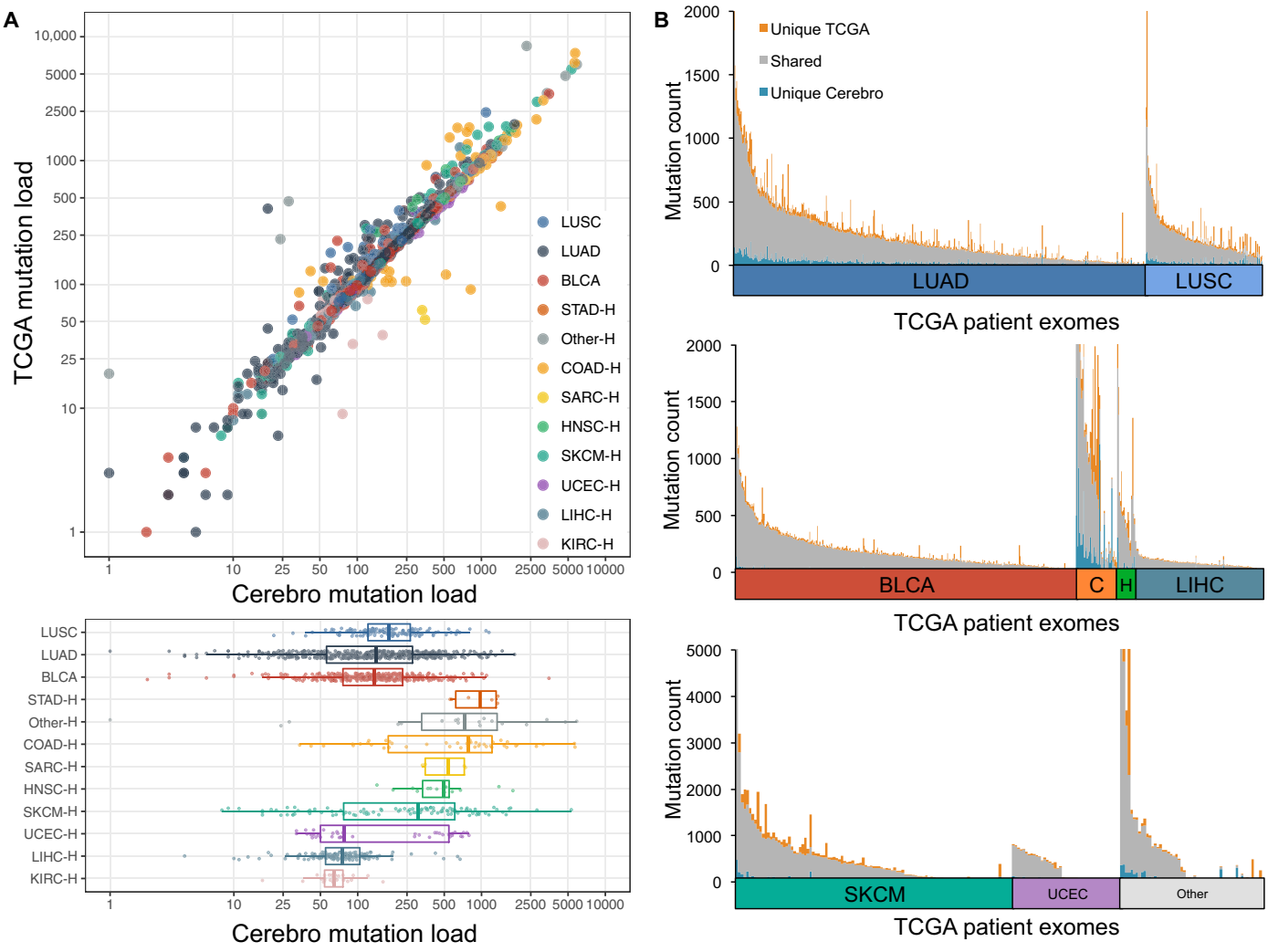

Fig. 3. Comparison of TCGA and Cerebro mutations for 1368 exomes. Somatic mutations from the TCGA MC3 project and Cerebro were compared for concordance. (**A**) Total mutational loads between the two mutation calls shared by cancer type. Mutation loads were defined as the total number of nonsynonymous mutations per sample. LUSC, lung squamous cell carcinoma; LUAD, lung adenocarcinoma; BLCA, bladder; STAD, stomach; COAD, colorectal; SARC, sarcoma; HNSC, head and neck squamous cell; SKCM, melanoma; UCEC, uterine; LIHC, liver; KIRC, kidney; *-H, set enriched for high–mutation load samples. (**B**) Unique/shared status for somatic mutations across all samples. C, colorectal; H, head and heck squamous cell.

average mutant allele fraction of 28%. In contrast, the Cerebro approach may have missed only one *KRAS* hotspot alteration in a tumor because of the fact that this alteration was also present in the matching normal sample of that patient. Overall, these results suggest that current TCGA data sets contain a substantial number of false-negative and false-positive somatic alterations in both passenger and driver genes among many tumor types, and the new mutation calls determined here provide an improved resource for analysis of TCGA sequence data.

Somatic mutation burden and response to cancer immunotherapy

To evaluate recently reported associations between total somatic mutation count (that is, mutational burden) and response to immune checkpoint blockade, we next obtained paired tumor-normal exome data from two recent studies including a study of response to anti–PD-1 therapy for 34 NSCLC patients (*19*) and a study of response to anti–CTLA-4 in 64 melanoma patients (*17*). We compared mutations called in these NGS data by Cerebro to the mutations reported in the original publications, limiting our analyses to nonsynonymous SBS changes because other types of alterations were not included in the published analyses. Across the NSCLC cohort, 9049 and 6385 mutations were identified in the original study and our reanalysis, respectively (table S11). In the melanoma cohort, 25,753 and 32,092 mutations were identified by the original publication and our reanalysis, respectively. Among all mutations in the NSCLC set, 48.2% were concordant between Cerebro and the original report, whereas 61.9% of mutations in the melanoma cohort were concordant (Fig. 5A). We performed an in-depth characterization of mutations that were identified in the original publications but that would be considered false positives using Cerebro and found that the large majority of such calls could be attributed to systematic issues such as limited observations of the mutation in distinct read pairs, poor base quality at the mutation position, and inaccurate alignment (Fig. 5B).

Given the association of TMB with clinical outcome in patients treated with immune checkpoint blockade (*8*, *17*, *19*), we wondered whether our analyses could be used to improve the classification of patients

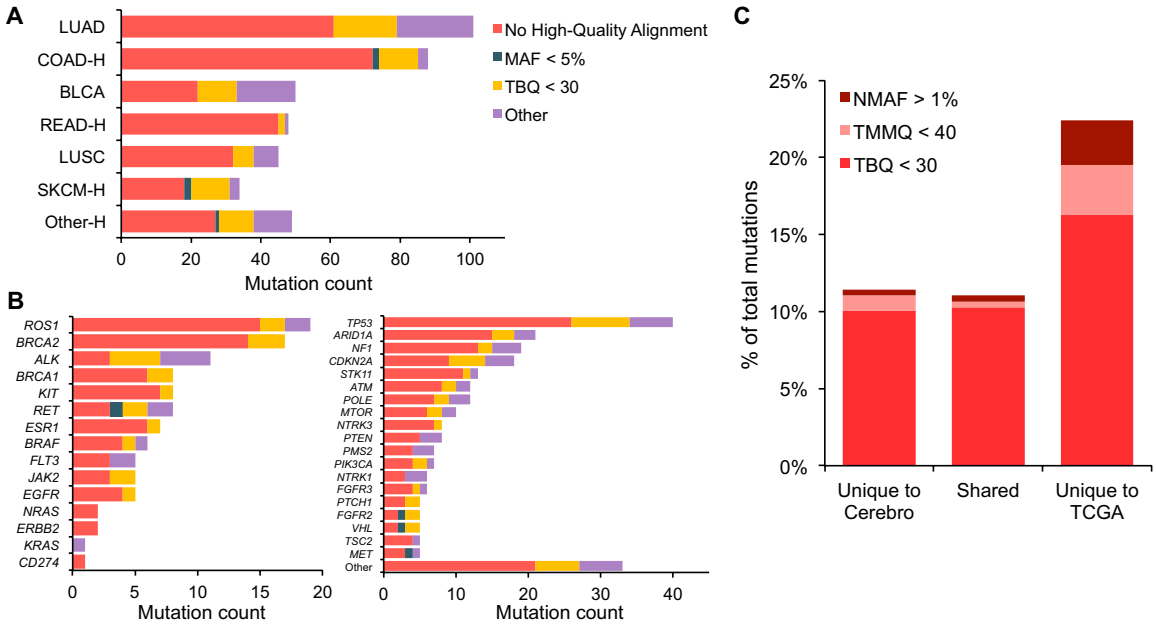

Fig. 4. Analysis of cancer driver gene mutations. Evaluation of mutations in 66 oncogenes and tumor suppressor genes indicated (**A**) a large number of low-confidence mutations unique to TCGA associated with various problematic features. No High-Quality Alignment, no consistent alignment found with at least one mutant base with quality higher than 30; MAF < 5%, mutant allele frequency below 5%; TBQ < 30, tumor base quality below 30. (**B**) Distribution of problematic TCGA driver gene calls by genes with approved FDA therapies (left panel) or ongoing clinical trials (right panel). (**C**) Quality characteristics of mutations uniquely called by TCGA were more problematic than other identified mutations. NMAF, normal mutant allele frequency; TMMQ, tumor mutant mapping quality; TBQ, tumor mutant base quality).

into mutator groups with different clinical outcomes. Cerebro analyses revealed that the average SBS mutational burdens for NSCLC and melanoma groups were 187 and 501, respectively, both substantially different than the original publications (266 and 402, respectively). Previously, melanoma analyses were performed using SomaticSniper (*60*) and led to a lower number of detected mutations, consistent with the lower sensitivity that we observed with this method (Fig. 2, A and C). Using the individual TMB from our analyses, we observed an improved prediction compared to previous mutation calls for progression-free survival (NSCLC) and overall survival (melanoma) for each entire patient cohort (Fig. 5, C and D, and fig. S6), and improved prediction compared to the previous validation set for the melanoma cohort (fig. S7). We examined whether these analyses would lead to changes in the classification of patients from high TMB to low TMB or vice versa. We found that of the 34 NSCLC patients treated with anti–PD-1 therapy, Cerebro mutation calls resulted in 4 patients (11.8%) switching from high TMB in the original publication to low TMB in our analysis (Fig. 5, C and D). These four patients had an average progression-free survival of 3.25 months. In contrast, of the 64 melanoma patients treated with anti–CTLA-4, Cerebro mutation calls resulted in 9 patients (14.0%) switching from low TMB classification in the original publication to high TMB in our analysis. These nine patients had an average overall survival of 40 months. Overall, these analyses suggest that improved mutation determination may have even greater impact on clinical outcome for immune checkpoint blockade than previously anticipated.

Evaluation of somatic mutations in clinical NGS analyses

To evaluate the effect of somatic mutation detection methods on NGS clinical cancer sequencing tests, we compared the Cerebro method trained for PGDx CancerSELECT 125 to three approaches for mutational profiling: the Thermo Fisher Oncomine Comprehensive Assay, the Illumina TruSeq Amplicon—Cancer Panel, and the MSK-IMPACT (Memorial Sloan Kettering–Integrated Mutation Profiling of Actionable Cancer Targets) panel. The analysis was performed using two sample cohorts, including replicates of formalin-fixed paraffin-embedded (FFPE) or frozen tumor samples from 22 lung cancer patients analyzed by CancerSELECT 125, Oncomine, and TruSeq or replicates of 15 breast, lung, and colorectal cancers with matched normal samples analyzed by both CancerSELECT 125 and MSK-IMPACT. For each sample, adjacent interspersed FFPE sections were evaluated using these approaches. Samples from three patients could not be analyzed using the TruSeq Amplicon—Cancer Panel due to insufficient DNA and were excluded from comparative analyses. Putative somatic mutations for the remaining patients were used for concordance evaluation and were limited to genomic regions comprising those common to the approaches in each analysis. Mutations were considered true positives if they were detected in at least two assays or if identified in only one of the assays and independently confirmed using ddPCR.

The first analysis resulted in a set of true somatic mutations consisting of 30 SBSs and six indels in commonly analyzed regions among the 19 patients (Table 2). The CancerSELECT 125 panel using Cerebro achieved 100% sensitivity and positive predictive values for all alterations, including SBSs and indels, and outperformed Oncomine and TruSeq. The Oncomine assay resulted in a sensitivity of 97% for SBSs and 83% for indels. The two alterations detected by CancerSELECT 125 and missed by Oncomine were confirmed using ddPCR (fig. S8). The TruSeq panel was unable to detect 15 SBSs and three indels, resulting in a lower sensitivity of 50% for both types of alterations (Table 2 and fig. S9). The CancerSELECT 125, Oncomine, and TruSeq approaches resulted in 0, 17, and 175 false positives, respectively. Seventeen false-positive SBSs reported by the Oncomine or TruSeq

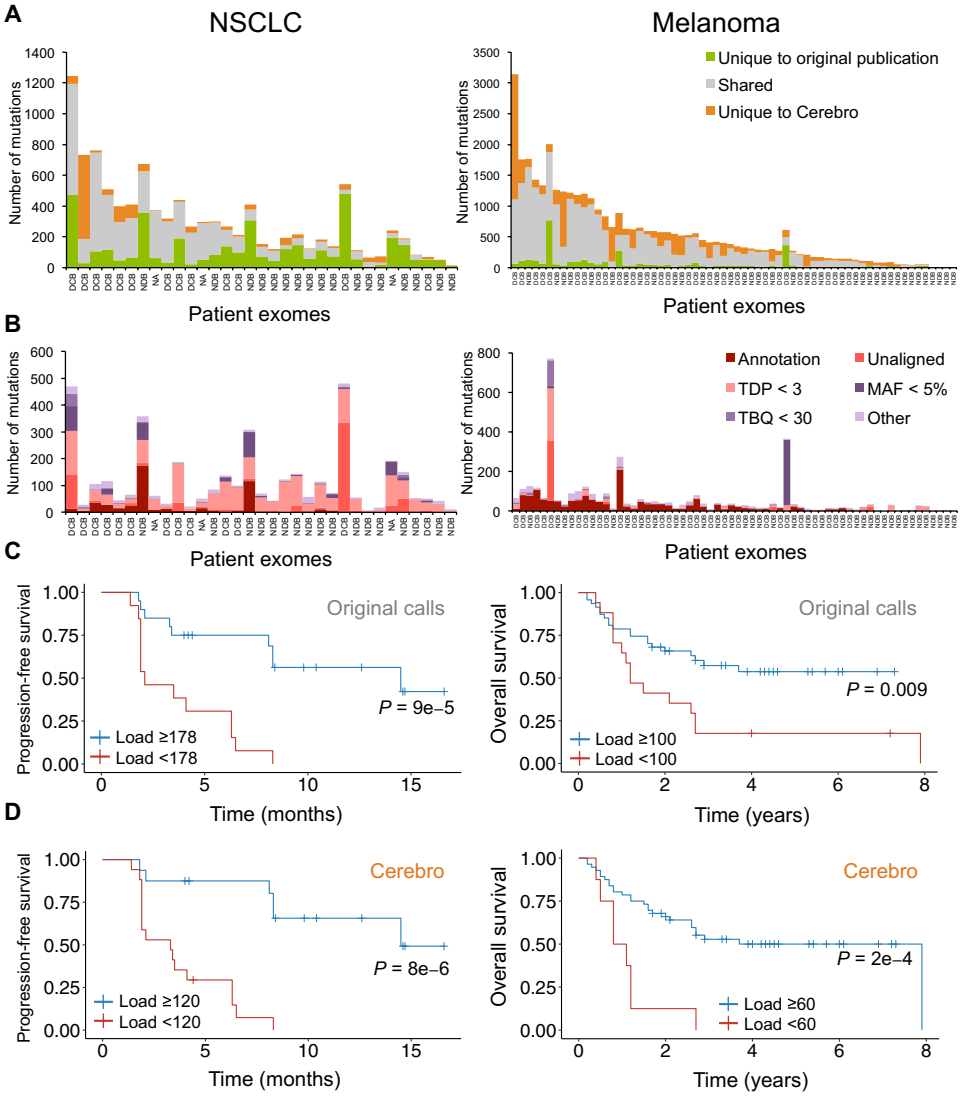

Fig. 5. Analysis of TMB in patients treated with immune checkpoint blockade. Comparison of Cerebro mutation calls with published calls associated with NSCLC (left panels) or melanoma (right panels). (**A**) Unique/shared mutation status for all patients. (**B**) Problematic mutations unique to original publications annotated by characteristic issue. Annotation, conflicting consequence; TDP < 3, tumor distinct pairs less than 3; TBQ < 30, tumor base quality less than 30; Unaligned, no alignment of mutated reads to the mutation position; MAF < 5%, mutant allele frequency less than 5%. (**C**) Kaplan-Meier analysis of progression-free survival (left) or overall survival (right) using tumor mutation loads from original publications. (**D**) Kaplan-Meier analysis of the same samples using Cerebro mutational loads. Log-rank *P* value shown for each survival plot. DCB, durable clinical benefit; NDB, no durable benefit.

panels were single-nucleotide polymorphisms (SNPs) that could be identified in databases of known germline variants and had been removed by the Cerebro approach. However, the large number of false-positive results identified by the TruSeq panel suggests additional underlying technical causes, including PCR artifacts and sequencing error, resulting in a positive predictive value of 8% (Table 2 and fig. S9). Because of the high number of TruSeq nongermline false positives, including in putative driver hotspots, we performed a detailed evaluation of these candidate mutations in the NGS data obtained using the CancerSELECT 125 and Oncomine platforms, and found no evidence of these mutations predicted by the TruSeq approach (table S12).

In our second comparison, we identified 46 alterations among the 15 patients (Table 3). The CancerSELECT 125 panel achieved 100% sensitivity and positive predictive values for all alterations identified. The MSK-IMPACT assay, which included a manual review of all candidate alterations, identified all but one missense mutation in *PMS2* that had been detected by CancerSELECT 125. This alteration was confirmed to be a bona fide change using ddPCR (fig. S8). To further demonstrate the importance of the Cerebro base caller within CancerSELECT 125, we evaluated the NGS data from our analyses using two mutation callers and demonstrated superior sensitivity and positive predictive values for Cerebro (table S13). Overall, these analyses establish that the Cerebro-trained CancerSELECT 125 had superior performance compared to other commonly used clinical NGS platforms, and highlight the importance of high-accuracy somatic mutation detection for identification of bona fide alterations in clinical NGS assays.

DISCUSSION

This study describes the development of a machine learning approach for optimizing somatic mutation detection in human cancer. Our analyses demonstrated that high-accuracy mutation detection can improve identification of bona fide alterations to determine total mutational burden to predict outcomes to immunotherapy, as well as to detect alterations in potentially actionable driver genes. These data highlight the challenges of detecting somatic sequence alterations in human cancer and provide a broadly applicable means for detecting such changes that is more accurate than existing approaches.

Our assessment of mutation calling approaches revealed that existing methods for somatic mutation detection may be influenced by factors that can lead to considerable false positives and false negatives. The machine learning approach that we described here, together with the large amount of experimental and in silico training data, identified key features in NGS sequence data to minimize false-positive calls and to improve sensitivity for bona fide alterations. An important aspect of our approach is the use of dual sequence alignments to improve identification of bona fide alterations (Supplementary Materials and Methods), effectively removing half of the erroneous mutation calls. Our overall approach provides the highest sensitivity and specificity of all methods that we analyzed. Although it is conceivable that other mutation callers could achieve similar performance

Table 2. Comparison of NGS cancer sequencing panels. TP, true positive; FN, false negative; FP, false positive; 95% CI, 95% confidence interval; Indel, insertion or deletion; PPV, positive predictive value.

Performance metric	CancerSelect 125				Oncomine comprehensive assay				TruSeq Amplicon—Cancer Panel			
	TP	FN	Point estimate (%)	95% CI	TP	FN	Point estimate (%)	95% CI	TP	FN	Point estimate (%)	95% CI
SBS sensitivity	30	0	100	85.9–100	29	1	96.7	80.9–99.8	15	15	50.0	33.2–66.8
Indel sensitivity	6	0	100	51.7–100	5	1	83.3	36.4–99.1	3	3	50.0	18.8–81.2
	TP	FP	Point estimate (%)	95% CI	TP	FP	Point estimate (%)	95% CI	TP	FP	Point estimate (%)	95% CI
SBS PPV	30	0	100	85.9–100	29	17	63.0	47.5–76.4	15	169	8.2	4.8–13.3
Indel PPV	6	0	100	51.7–100	5	0	100	46.3–100	3	6	33.3	9.0–69.1

Table 3. Comparison of clinical cancer sequencing panels.

Performance metric	CancerSelect 125				MSK-IMPACT			
	TP	FN	Point estimate (%)	95% CI	TP	FN	Point estimate (%)	95% CI
SBS sensitivity	36	0	100	88.0–100	35	1	97.2	83.8–99.9
Indel sensitivity	10	0	100	65.5–100	10	0	100	65.5–100
	TP	FP	Point estimate	95% CI	TP	FP	Point estimate (%)	95% CI
SBS PPV	36	0	100	88.0–100	35	0	100	87.7–100
Indel PPV	10	0	100	65.5–100	10	0	100	65.5–100

characteristics, it would be challenging to identify appropriate parameters for these callers without taking a large-scale training and validation approaches similar to ours. Machine learning methods have previously been proposed for somatic mutation discovery including MutationSeq (*65*), SomaticSeq (*66*), and SNooPer (*67*). However, these tools were not trained with extensive data, nor were they validated using independent (non-NGS) sequencing approaches.

Despite the advantages of our approach, this study has several limitations. For example, our training data sets were specific to certain genomic analyses. Because we have focused on coding regions within exome or targeted sequencing with >150× coverage using Illumina platforms, analyses of NGS data with different characteristics, such as other sequencing technologies or lower sequence coverage levels, would not be expected to be as successful using the current set of features, parameters, and training data. Additionally, analyses of whole-genome sequencing data sets containing noncoding and repeat elements would also likely require expanded training data to reflect the unique characteristics of these regions. Our approach is not intended for identification of germline variants, although other methods have been designed for this purpose (*68*). Furthermore, our analyses were largely focused on tumors with high tumor cellularity, and we have not comprehensively evaluated detection of low-frequency alterations, nor did we evaluate structural changes in these tumors. Although our approach has been optimized for somatic mutation identification using a common NGS technology, we expect that with sufficient training data, our method could be successfully used with new sequencing platforms and applications.

Given the fundamental importance of somatic genomic alterations in human cancer, the improvements that we have developed are likely to have meaningful implications for research and clinical analyses. Our reanalysis of TCGA and exome data from patients treated with cancer immunotherapy identified that a substantial fraction of existing mutation calls are likely to be false-positive changes associated with low-quality evidence and that many true alterations may have been missed in current databases. We estimated that 16% of alterations are likely inaccurate in current TCGA mutation databases and an additional 10% of true alterations may have been missed in these data sets. If these ratios are accurate across TCGA (with ~2 million somatic mutations across 10,000 exomes), then the overall number of false-positive and false-negative changes in TCGA is likely to be >500,000. Such discrepancies are likely to be important across a variety of additional efforts, as much of TCGA has been incorporated into mission-critical databases such as COSMIC (*69*–*71*), gnomAD/ExAC (*72*), Genomic Data Commons (*73*), and the International Cancer Genome Consortium (*74*).

Our analyses may have important implications for therapies using genomic information, including targeted therapies and immune therapy approaches targeting mutation-associated neoantigens. Although MSI testing is currently used to identify the small number of patients

with dMMR that benefit from immune checkpoint blockade, there are large-scale efforts in developing TMB as a biomarker for immunotherapy response because this approach could identify a larger number of patients likely to benefit from this therapy. Although the number of samples that we analyzed was small, our use of Cerebro to determine TMB in NSCLC and melanoma patients improved the accuracy of stratifying patients into likely responders and nonresponders; in particular, 13% of patients from these two studies were reclassified using the Cerebro mutation analyses. Improved discrimination of bona fide alterations may facilitate development of mutation load–based predictive biomarkers for immune checkpoint blockade as well as for understanding changes in cancer genomes during immune therapy (*75*). Development of mutation-specific vaccines and immune cell–based therapies will require high-confidence identification of alterations that may be unique to individual patients.

This study has implications beyond analyses of tumor tissues, including, for example, in deep sequencing of cell-free DNA (cfDNA) to identify somatic mutations in the circulation (*76*). Current approaches for cfDNA analyses use sequence data with over ~30,000× coverage to identify alterations with concentrations as low as 0.05% (*64*, *76*–*85*). Highly accurate analyses of these sequences will be needed to provide robust differentiation of true-positive and true-negative mutation calls, especially in cases without previous knowledge of sequence alterations from tumor tissue.

Overall, as cancer genomic analyses continue to expand and gain acceptance in the clinical community, the ability to effectively design and validate methods for mutation identification remains challenging. Our approach, which is entirely automated, would eliminate the need for expert review of sequence data, a practice commonly used in the clinical setting and likely unsustainable for widespread NGS analyses. Large-scale validation of these methods will provide new opportunities for the treatment and management of patients with cancer.

MATERIALS AND METHODS

Study design

This study provides a somatic mutation detection algorithm using a machine learning approach. We trained our method using in silico mutations spiked into replicate exome sequencing runs of a well-characterized peripheral blood sample. We estimated our method's sensitivity and specificity using five matched tumor/normal cancer cell line pairs with somatic variants previously identified and validated through Sanger sequencing and other means. Additionally, we compared our method's accuracy against six other mutation detection methods using the same validated mutation data. We determined concordance between our method and a consensus somatic mutation call set using exome data from 1368 patients available through the TCGA. We also evaluated the association between clinical response to immune checkpoint blockade and TMB using two cohorts of paired tumor-tissue and normal exome samples obtained before immune therapy. These cohorts were composed of stage IV NSCLC patients (n = 34) treated with anti–PD-1 and metastatic melanoma patients (n = 64) treated with anti–CTLA-4. We compared the performance of the method to three additional NGS clinical cancer sequencing tests using two sample cohorts, including replicates of FFPE or frozen tumor samples from breast, lung, and colorectal cancers (n = 37). All analyzed samples were obtained under Institutional Review Board–approved protocols with informed consent for research use at participating institutions.

Development of Cerebro

In choosing a machine learning model for Cerebro's development, we focused on models that could quickly process large amounts of training data during model fitting. This was an important consideration given the large amount of training data used (more than 2 million candidate mutations) and the need to build many models during initial testing. We used the scikit-learn software package (*86*) to implement our models, and this also informed our model decision.

We considered machine learning techniques such as support vector machines (SVMs) available through scikit-learn, but as the SVM training algorithm scales more than quadratically with the number of training examples, we found that SVMs were not applicable to our training data. Additionally, the implementations of models that use either adaptive or gradient boosting did not support parallelization of model fitting, which also made them unsuitable for our training data.

Random Forest classifiers (*87*) provide support for parallelized model fitting and are well suited to large training data sets. The closely related approach of extremely randomized trees (or "Extra-Trees") (*88*) is able to fit models more rapidly because of the Extra-Trees classifier's selection of random thresholds versus the Random Forest classifier's computationally expensive determination of optimal thresholds for each examined feature. We used default settings in the Extra-Trees model, with the exception of (i) parallelization options to use all available processing cores and (ii) increasing the number of decision trees in the model to 1000. All other parameters were left as default values as provided by scikit-learn (criterion="gini"; max_features="auto"; max_depth=None; min_samples_split=2; min_samples_leaf=1; min_weight_fraction_leaf=0; max_leaf_nodes=None; bootstrap=False; oob_score=default; random_state=None; warm_start=False; class_weight=None).

On the basis of review of visual analyses of bona fide somatic mutations, we chose a set of features that described mutation, sequence quality, genomic context, and alignment characteristics. These are described in table S1. We used two common aligners [BWA-MEM (Burrows-Wheeler Aligner–Maximal Exact Match) (*89*) and Bowtie2 (*90*)] to provide multiple estimates of alignment-related features. We did not make any attempt to remove redundant features from our model because these are appropriate for Extra-Trees models.

Processing of exome data with Cerebro

Reads from tumor cell lines and matched normal samples sequenced at PGDx were adapter-masked and demultiplexed using bcl2fastq (http://support.illumina.com). All read data, including those from PGDx, TCGA, and immunotherapy whole-exome studies, were aligned with BWA-MEM (*89*) and Bowtie2 (*90*) to the hg19 reference assembly of the human genome (*91*), with unplaced and unlocalized contigs and haplotype chromosomes removed. Then, Cerebro identified candidate somatic mutations by examining alignments in the matched tumor and normal samples. Alignment data were filtered to remove nonprimary alignment records, reads mapped into improper pairs, and reads with more than six edits. Individual bases were excluded from mutant coverage calculation if their Phred base quality was <30 in tumor samples and <20 in normal samples. Only candidate somatic variants found in both pairs of alignments (BWA-MEM and Bowtie2) were scored using our confidence scoring model (see table S1). Candidate variants with somatic confidence scores <0.75, <3 distinct mutant fragments in the tumor, <10% mutant allele fraction (MAF) in the tumor, or <10 distinct coverage in the normal

sample were removed. For our analysis of cancer immunotherapy response–associated data (*17*) or NSCLC (*19*), we included mutations ≥5% MAF to compensate for the low tumor purity that appeared to be present in some samples.

For mutations found in at least 50 samples according to the COSMIC v72 database ("hotspots"), we applied relaxed cutoffs. For such hotspot mutations, we only excluded bases in the tumor sample if their Phred base quality was <20. We also only removed candidate hotspot mutations if they had somatic confidence scores <0.25, <2 distinct mutant fragments in the tumor, or <5% MAF in the tumor. Because sequence data obtained from TCGA were often less than 100 bp in length, which we found to reduce Bowtie2 alignment sensitivity for long indels, we also created a set of relaxed cutoffs for hotspots that were indels >8 bp in length. These relaxed indel hotspot filtering criteria focused only on the BWA-MEM alignments and removed mutations with <5 distinct fragments in the tumor, a left-tailed Fisher's exact test $P > 0.01$, <5% MAF in the tumor, or any mutant fragments in the normal sample.

Variants were further filtered for coding consequence using VEP (Variant Effect Predictor) (*92*) and CCDS (Consensus Coding Sequence)/RefSeq (*93*) to remove intragenic and synonymous mutations. Finally, variants that were listed as Common in dbSNP (Single Nucleotide Polymorphism database) (*94*) version 138 were removed.

Processing of exome data with external variant callers

Read data were aligned with BWA-MEM (*89*) to an hg19 reference assembly of the human genome (*91*), with unplaced and unlocalized contigs and haplotype chromosomes removed. The Picard (*95*) MarkDuplicatesWithMateCigar program was used on the resulting BAM files to find optical and PCR duplicates. Each external variant caller was run with default parameters and filters as described in Supplementary Materials and Methods. In the case of Strelka, we used the reported "tier 2" set of variants. Similarly to the processing with Cerebro, for all variant callers, we removed variants with MAF < 10%, as well as intragenic and synonymous mutations, and variants listed as Common in dbSNP138. Variants failing a caller's default set of filters were also removed. For VarDict, we also removed variants that hit either of two filters suggested by one of VarDict's authors (see the Supplementary Materials).

Confidence scoring model and training

A sample derived from normal peripheral blood (CRL-2339, American Type Culture Collection) was used to generate a genomic library for exome capture as previously described (*5*) and analyzed twice using NGS. One of these sequencing runs was designated the "training tumor" and had novel variants spiked into it using BAMSurgeon (*96*). Novel coding variants were randomly generated across the exome, at MAFs ranging from 1.5625 to 100%, according to a distribution defined by 2^{-R}, where R is a uniform random variable between 0 and 6. After accounting for the variants that could not be inserted because of low coverage or presence of polymorphisms, these novel variants were a mixture of 16,958 substitutions, 6675 insertions, and 6720 deletions. The range of MAFs used for training was intended to begin well below the expected limit of detection (5 or 10% MAF) to ensure that calls near the limit would be accurate. Novel indels ranged from 1 to 18 bp in length. Additional novel indels were spiked in by locating short-tandem repeat tracts (either mono-, di-, or trinucleotide repeats) within the exome and inserting one or two repeat unit contractions or extensions of the tracts. After spiking the training tumor with BAMSurgeon, the read data from the training tumor were realigned with both BWA-MEM and Bowtie2, and then all candidate somatic variants supported by at least one tumor read were reported using Cerebro. Candidate somatic variants found in both pairs of alignments formed the training set for the scoring model, which included the 30,353 spiked "true somatic" mutations (class 1) and 2,016,867 artifactual or germline mutations (class 0).

For each candidate somatic variant, Cerebro considers the mutation, sequence quality, genomic context, and alignment characteristics (listed in table S1). These values were calculated for the two sets of alignments and were concatenated together to form a feature vector for a candidate somatic variant. The training set for the scoring model consisted of the feature vectors for each candidate variant and a class labeling indicating whether or not the variant was spiked into the training tumor (that is, if the candidate is a somatic variant or not). The scoring model itself is an extremely randomized trees model (*88*) with 1000 decision trees, implemented using the scikit-learn library (*86*), with the reported confidence score being the percentage of the model's trees that would classify the variant as somatic.

Validation of somatic variants

We analyzed five matched tumor/normal breast cancer cell line pairs in which several hundred somatic variants had previously been identified and validated through Sanger sequencing analyses (*28*, *30*). To add to this set, we performed exome sequencing of the cell lines using an NGS approach as previously described (*5*). We then analyzed the NGS data through three variant calling programs (VarDict, MuTect 1, and our Cerebro pipeline). Somatic variants called by all three programs were considered to be validated, provided that they passed a visual inspection to remove possible artifacts; mutations in this set that did not clearly pass visual inspection were tested using ddPCR. Somatic variants called by one or two of those programs with a reported MAF of at least 20% (by at least one program) were visually inspected for alignment artifacts. Those variants passing visual inspection and having at least 10 reads covering the locus in the normal sample were then tested using ddPCR, and variants validated by ddPCR formed the remainder of our validated variant set.

Evaluation of variant caller accuracy

For simulated data sets, true positives (TPs) were those spiked-in somatic variants found by a program, false positives (FPs) were variants called by a program that were not spiked in, and false negatives (FNs) were spiked-in variants not called by a program. For these simulated data sets, sensitivity is defined as TP/(TP + FN), positive predictive value (PPV) is defined as TP/(TP + FP), and false-positive rate (FPR) is the number of false positives reported per megabase of the exome (51.5 Mbp). For the cell line data sets, we created a validated variant set as described above and selected those variants called by at least one caller as having a MAF of at least 20%; these selected variants created the validated comparison set. When evaluating the variant callers on cell line data, TP were comparison variants found by a program; FP were variants not in the comparison set that were called by a program with a MAF of at least 20%; and FN were comparison variants not found by a program. Because we only validated variants reported to have a MAF of at least 20%, we defined PPV for the cell line data as $X/(X + \mathrm{FP})$, where X is the number of TP variants that a program claimed had a MAF of at least 20%. This approach allowed us to compensate for the variation in reported allele frequency between variant calling programs by restricting PPV calculation to only

those variants that a program reported as being over the validation MAF threshold of 20%. The following scores were used for thresholding for Precision-Recall/ROC curve generation: Cerebro: Cerebro score; MuTect1: t_lod_fstar ("Log of (likelihood tumor event is real / likelihood event is sequencing error)"); MuTect2: TLOD (same as MuTect1); SomaticSniper: tumor mutant coverage; Strelka: "Quality score" in VCF ("QSS"/"QSI" from SNV/indel); VarDict: Variant quality in VCF (log(AD) * mean BQ); VarScan: –log(somatic *P* value).

Simulation experiments

We performed three simulation experiments designed to evaluate the accuracy of the various variant calling programs. In each experiment, we used a set of simulated tumor-normal pairs created by sequencing six exome-captured normal cell lines twice to create six sample pairs; one of the samples in each pair was designated the "tumor," and both samples were aligned to the human genome using Bowtie2. Depending on the experiment, a set of artificial coding somatic variants were inserted into the tumors using BAMSurgeon. After BAMSurgeon was run, the read sequence data were extracted from the modified BAM file, and the resultant FASTQ data were aligned again using the methods described above.

Our first simulation experiment was designed to simulate low-purity tumors with 120 SBSs and 12 indels inserted into each tumor at MAFs ranging from 10 to 25%. In this final simulation, we evaluated several accuracy metrics, including sensitivity, PPV, FPR, and F-score. The second experiment was solely a test of specificity, where we did not insert any somatic variants into our simulated tumor; this examination of somatic variant calling specificity with technical replicates is similar to that discussed by Saunders *et al.* (*61*) in their presentation of Strelka. In our final simulation experiment, in which we focused on sensitivity, we inserted 7000 coding variants, consisting of 1000 SBSs and 6000 indels (1000 each of 1-, 2-, and 3-bp insertions and 1-, 2-, and 3-bp deletions). The second experiment's variants were inserted at MAFs ranging from 10 to 100%.

ddPCR methods

ddPCR forward and reverse primers as well as wild-type and mutant probes were created using the Bio-Rad ddPCR Custom Design Portal (www.bio-rad.com/digital-assays). Genomic DNA corresponding to 10,000 genome equivalents was added to 10 μl of 2× ddPCR Supermix (Bio-Rad), 1 μl of 20× target primers and probe (FAM), and 1 μl of 20× reference primers and probe (HEX) and brought to a 22-μl volume with nuclease-free water to create a reaction mix. A DG8 cartridge in a DG8 cartridge holder (Bio-Rad) was loaded with 20 μl of reaction mix and 70 μl of Droplet Generation Oil (Bio-Rad). The cartridge was placed in a QX200 Droplet Generator to generate about 20,000 nanoliter-sized droplets. Droplets were loaded into a twin.tec 96-well, semi-skirted plate (Eppendorf) and sealed with a foil heat seal using a PX1 PCR Plate Sealer (Bio-Rad). Subsequent PCR cycling was performed on a C1000 Touch Thermal Cycler (Bio-Rad) with the following conditions: 95°C for 10 min, followed by 40 cycles of 94°C for 30 s and 55°C for 1 min, and ending with 98°C for 10 min. The plate was then loaded on a QX200 Droplet Reader (Bio-Rad), and PCR-positive and PCR-negative droplets were quantified. Raw droplet data were analyzed with QuantaSoft software (Bio-Rad). Thresholds were manually assigned using two-dimensional amplitude clustering plots and the crosshair tool for each tumor and normal pair. Tumor samples were run in duplicate, and the average mutant allele fraction was taken. For the comparative evaluation of clinical targeted sequencing panels, the ddPCR protocol was modified as follows: 5500 genome equivalents were used, and 1 μl of a 20× mixture of target primers and probes and reference primers and probes (FAM and HEX, respectively) were used (Bio-Rad). Because there were no matched normal samples, wild-type and mutant oligomers were designed as controls for each target investigated (Operon Biotechnologies).

WES data extraction and annotation

The results shown here are in part based on data generated by the TCGA Research Network (http://cancergenome.nih.gov), as outlined in the TCGA publication guidelines http://cancergenome.nih.gov/publications/publicationguidelines. TCGA WES data sets (bam alignment files) represented untreated primary tumors and paired normal tissue samples obtained from the Cancer Genomics Hub (http://cghub.ucsc.edu). WES-derived somatic mutation calls were obtained from the MC3 project. TCGA somatic mutation calls (v0.2.8) were also obtained from the Synapse repository (http://synapse.org; Synapse ID, syn7214402). Variant-supporting coverage and total coverage were extracted and manually reviewed for consistency. To normalize across cancer types, mutations with fewer than three variant-associated reads or less than 10% mutant allele frequency were filtered before downstream comparative analysis. For mutations found in at least 50 samples according to the COSMIC v72 database (hotspots), we allowed for 5% minimum mutant allele frequency. Additional somatic mutation call sets generated by MuTect2 were downloaded from the Genomic Data Commons (http://gdc.cancer.gov) and Broad GDAC Firehose (http://gdac.broadinstitute.org) using the *firehose_get* download client prioritizing the *BIFIREHOSE_Oncotated_Calls* somatic mutation call sets compiled from various TCGA Genome Sequencing Centers' bioinformatics pipelines. For comparisons of Cerebro to other call sets, mutations were required to fall within a common region of interest (ROI) set. The source of the primary mutation calling tools used for each cancer type may be found in the corresponding TCGA marker publications (http://cancergenome.nih.gov/publications). Individual shared and unique mutations between Cerebro and TCGA MC3 are displayed in table S14. Concordance analysis of somatic mutations from 66 oncogenes and tumor suppressor genes included manual review to determine shared status of mutations within the same or adjacent codons. Whole-exome melanoma (*17*) or NSCLC (*19*) immunotherapy data sets were obtained via National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGaP) (accession numbers phs000980 and phs001041). We excluded one sample from (*19*) in our comparative analysis (patient ID: SB010944) because of conflicting information regarding TMB in the original publication.

Clinical study design and analysis

The first comparison of NGS assays included 22 total samples, 17 FFPE samples, and 5 frozen tumor tissue specimens obtained from lung cancer patients from ILSBio/Bioreclamation that were analyzed for the presence of sequence mutations using three targeted cancer gene panels from independent vendors. The second comparison of clinical NGS assays used 15 matched tumor-normal samples from breast ($n = 9$), lung ($n = 5$), and colorectal ($n = 1$) cancers from ILSBio/Bioreclamation or Indivumed. All samples were obtained under Institutional Review Board–approved protocols with informed consent for research use at participating institutions. One set of patient samples was processed and analyzed for sequence mutations by Personal Genome Diagnostics using the CancerSELECT 125 panel. In brief, samples were reviewed

by a pathologist to determine the percent tumor purity followed by macrodissection and DNA extraction. DNA was fragmented and used for CancerSELECT 125 library preparation and capture. Libraries were sequenced using HiSeq 2500 instruments (Illumina). Sequencing output was analyzed and mutations were identified using Cerebro, which was retrained to reflect the composition and coverage of the CancerSELECT 125 panel. In our first comparison of NGS assays, an identical set of FFPE and frozen tumor tissue specimens were sent to MolecularMD along with a hematoxylin and eosin–stained image for processing and NGS analysis using two cancer-specific panels supplied by two independent vendors: Oncomine Comprehensive Assay (ThermoFisher) and TruSeq Amplicon—Cancer Panel (Illumina). In the second comparison of NGS assays, an identical set of FFPE tumor tissue specimens and matched normal blood specimens were sent to Memorial Sloan Kettering Cancer Center along with a hematoxylin and eosin–stained image for processing and NGS analysis using the MSK-IMPACT panel. To limit the effects of tumor heterogeneity on analysis, slides were distributed in nonsequential order for testing. For orthogonal analysis, comparisons were limited to ROIs that were included in a comparison (table S15). Samples that failed quality check in one or more panels were excluded. A sequence mutation was considered a true positive (TP) if there was positivity in at least two panels and a false positive (FP) if only detected in one panel. Sequence mutations were considered true negatives (TNs) if there was negativity in at least two panels. A position with no mutation detected was considered a false negative (FN) in a panel if that position was concordantly positive in the other two panels. Genomic positions that were masked on the basis of known SNPs in the CancerSELECT 125 panel were considered FP in Oncomine and TruSeq analyses. Because there were >150 FPs detected with the TruSeq panel, discordant resolution was limited to FPs or FNs obtained by CancerSELECT 125 and Oncomine (not considered SNPs) and were resolved using ddPCR. Searches for TruSeq false positives in CancerSELECT 125 and Oncomine raw bam files used Samtools *mpileup* (*97*) with minimum Phred quality thresholds of 0 and 25, respectively.

Statistical methods

The Mann-Whitney *U* test was used to compare quantitative measures (for example, nonsynonymous mutational load) between groups of interest. Comparisons of relative frequencies used Fisher's exact test. The log-rank test was used to evaluate differences between Kaplan-Meier curves for overall survival or progression-free survival in the melanoma (*17*) or NSCLC (*19*) immunotherapy data sets, respectively. Ninety-five percent CIs for proportions in clinical NGS comparisons were calculated using the method described by Newcombe (*98*) and Wilson (*99*) with no continuity correction.

SUPPLEMENTARY MATERIALS

www.sciencetranslationalmedicine.org/cgi/content/full/10/457/eaar7939/DC1
Materials and Methods
Fig. S1. False-positive evaluation for somatic mutation callers.
Fig. S2. ddPCR mutation validation analyses.
Fig. S3. Precision-recall and ROC curve analyses of Cerebro and other mutation callers using experimentally validated alterations.
Fig. S4. Mutation loads of TCGA exomes using different mutation calling methods.
Fig. S5. Concordance rates (% of total mutations) of Cerebro compared to other mutation call sets for TCGA exomes.
Fig. S6. Response to checkpoint inhibitors associated with mutational load.
Fig. S7. Survival analysis by mutation load stratified by discovery and validation cohorts.
Fig. S8. Alterations confirmed by ddPCR in clinical NGS panel comparison.
Fig. S9. Comparative results of three clinical sequencing panels.
Table S1. Confidence scoring model features for the Cerebro machine learning algorithm.
Table S2. Performance results for simulated low-purity tumors.
Table S3. False-positive rates of mutation calling methods.
Table S4. Sensitivity results by variant type of mutation calling methods.
Table S5. Sensitivity results for substitutions by MAF.
Table S6. Sensitivity results for insertion-deletions by MAF.
Table S7. Results of ddPCR validation of somatic mutations.
Table S8. Performance results for cell lines with validated somatic mutations.
Table S9. Unique and shared mutation load results for Cerebro and TCGA (MC3).
Table S10. Concordance results for Cerebro and TCGA for driver oncogenes and tumor suppressor genes.
Table S11. Comparison of mutation calls in immunotherapy publications and Cerebro reanalysis.
Table S12. Evaluation of TruSeq false-positive calls using raw CS125 and Oncomine sequence data.
Table S13. Comparison of mutation callers for clinical samples.
Table S14. Shared and unique somatic mutation calls between Cerebro and TCGA.
Table S15. Genomic ROIs used in clinical panel comparisons.

REFERENCES AND NOTES

1. R. Kamps, R. D. Brandão, B. J. Bosch, A. D. Paulussen, S. Xanthoulea, M. J. Blok, A. Romano, Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. *Int. J. Mol. Sci.* **18**, E308 (2017).
2. H. L. Rehm, Disease-targeted sequencing: A cornerstone in the clinic. *Nat. Rev. Genet.* **14**, 295–300 (2013).
3. G. M. Frampton, A. Fichtenholtz, G. A. Otto, K. Wang, S. R. Downing, J. He, M. Schnall-Levin, J. White, E. M. Sanford, P. An, J. Sun, F. Juhn, K. Brennan, K. Iwanik, A. Maillet, J. Buell, E. White, M. Zhao, S. Balasubramanian, S. Terzic, T. Richards, V. Banning, L. Garcia, K. Mahoney, Z. Zwirko, A. Donahue, H. Beltran, J. M. Mosquera, M. A. Rubin, S. Dogan, C. V. Hedvat, M. F. Berger, L. Pusztai, M. Lechner, C. Boshoff, M. Jarosz, C. Vietz, A. Parker, V. A. Miller, J. S. Ross, J. Curran, M. T. Cronin, P. J. Stephens, D. Lipson, R. Yelensky, Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
4. L. A. Garraway, P. A. Jänne, Circumventing cancer drug resistance in the era of personalized medicine. *Cancer Discov.* **2**, 214–226 (2012).
5. S. Jones, V. Anagnostou, K. Lytle, S. Parpart-Li, M. Nesselbush, D. R. Riley, M. Shukla, B. Chesnick, M. Kadan, E. Papp, K. G. Galens, D. Murphy, T. Zhang, L. Kann, M. Sausen, S. V. Angiuoli, L. A. Diaz Jr., V. E. Velculescu, Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl. Med.* **7**, 283ra253 (2015).
6. E. M. Van Allen, N. Wagle, P. Stojanov, D. L. Perrin, K. Cibulskis, S. Marlow, J. Jane-Valbuena, D. C. Friedrich, G. Kryukov, S. L. Carter, A. McKenna, A. Sivachenko, M. Rosenberg, A. Kiezun, D. Voet, M. Lawrence, L. T. Lichtenstein, J. G. Gentry, F. W. Huang, J. Fostel, D. Farlow, D. Barbie, L. Gandhi, E. S. Lander, S. W. Gray, S. Joffe, P. Janne, J. Garber, L. MacConaill, N. Lindeman, B. Rollins, P. Kantoff, S. A. Fisher, S. Gabriel, G. Getz, L. A. Garraway, Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **20**, 682–688 (2014).
7. L. A. Diaz Jr., D. T. Le, PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **373**, 1979 (2015).
8. D. T. Le, J. N. Durham, K. N. Smith, H. Wang, B. R. Bartlett, L. K. Aulakh, S. Lu, H. Kemberling, C. Wilt, B. S. Luber, F. Wong, N. S. Azad, A. A. Rucki, D. Laheru, R. Donehower, A. Zaheer, G. A. Fisher, T. S. Crocenzi, J. J. Lee, T. F. Greten, A. G. Duffy, K. K. Ciombor, A. D. Eyring, B. H. Lam, A. Joe, S. P. Kang, M. Holdhoff, L. Danilova, L. Cope, C. Meyer, S. Zhou, R. M. Goldberg, D. K. Armstrong, K. M. Bever, A. N. Fader, J. Taube, F. Housseau, D. Spetzler, N. Xiao, D. M. Pardoll, N. Papadopoulos, K. W. Kinzler, J. R. Eshleman, B. Vogelstein, R. A. Anders, L. A. Diaz Jr., Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
9. K. L. Redmond, A. Papafili, M. Lawler, S. Van Schaeybroeck, Overcoming resistance to targeted therapies in cancer. *Semin. Oncol.* **42**, 896–908 (2015).
10. F. Stegmeier, M. Warmuth, W. R. Sellers, M. Dorsch, Targeted cancer therapies in the twenty-first century: Lessons from imatinib. *Clin. Pharmacol. Ther.* **87**, 543–552 (2010).
11. P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, D. Hogg, P. Lorigan, C. Lebbe, T. Jouary, D. Schadendorf, A. Ribas, S. J. O'Day, J. A. Sosman, J. M. Kirkwood, A. M. Eggermont, B. Dreno, K. Nolop, J. Li, B. Nelson, J. Hou, R. J. Lee, K. T. Flaherty, G. A. McArthurBRIM-3 Study Group, Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **364**, 2507–2516 (2011).
12. J. D. Moyer, E. G. Barbacci, K. K. Iwata, L. Arnold, B. Boman, A. Cunningham, C. DiOrio, J. Doty, M. J. Morin, M. P. Moyer, M. Neveu, V. A. Pollack, L. R. Pustilnik, M. M. Reynolds, D. Sloan, A. Theleman, P. Miller, Induction of apoptosis and cell cycle arrest by CP-358,774, an inhibitor of epidermal growth factor receptor tyrosine kinase. *Cancer Res.* **57**, 4838–4848 (1997).

13. R. Rosell, E. Carcereny, R. Gervais, A. Vergnenegre, B. Massuti, E. Felip, R. Palmero, R. Garcia-Gomez, C. Pallares, J. M. Sanchez, R. Porta, M. Cobo, P. Garrido, F. Longo, T. Moran, A. Insa, F. De Marinis, R. Corre, I. Bover, A. Illiano, E. Dansin, J. de Castro, M. Milella, N. Reguart, G. Altavilla, U. Jimenez, M. Provencio, M. A. Moreno, J. Terrasa, J. Muñoz-Langa, J. Valdivia, D. Isla, M. Domine, O. Molinier, J. Mazieres, N. Baize, R. Garcia-Campelo, G. Robinet, D. Rodriguez-Abreu, G. Lopez-Vivanco, V. Gebbia, L. Ferrera-Delgado, P. Bombaron, R. Bernabe, A. Bearz, A. Artal, E. Cortesi, C. Rolfo, M. Sanchez-Ronco, A. Drozdowskyj, C. Queralt, I. de Aguirre, J. L. Ramirez, J. J. Sanchez, M. A. Molina, M. Taron, L. Paz-Ares; Spanish Lung Cancer Group in collaboration with Groupe Français de Pneumo-Cancérologie and Associazonr Italiana Oncologia Toracica, Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): A multicentre, open-label, randomised phase 3 trial. *Lancet Oncol.* **13**, 239–246 (2012).
14. L. Chang, M. Chang, H. M. Chang, F. Chang, Microsatellite instability: A predictive biomarker for cancer immunotherapy. *Appl. Immunohistochem. Mol. Morphol.* **26**, e15–e21 (2017).
15. D. P. Carbone, M. Reck, L. Paz-Ares, B. Creelan, L. Horn, M. Steins, E. Felip, M. M. van den Heuvel, T. E. Ciuleanu, F. Badin, N. Ready, T. J. N. Hiltermann, S. Nair, R. Juergens, S. Peters, E. Minenza, J. M. Wrangle, D. Rodriguez-Abreu, H. Borghaei, G. R. Blumenschein Jr., L. C. Villaruz, L. Havel, J. Krejci, J. Corral Jaime, H. Chang, W. J. Geese, P. Bhagavatheeswaran, A. C. Chen, M. A. Socinski; CheckMate 026 Investigators, First-line nivolumab in stage IV or recurrent non–small-cell lung cancer. *N. Engl. J. Med.* **376**, 2415–2426 (2017).
16. M. J. Overman, R. McDermott, J. L. Leach, S. Lonardi, H. J. Lenz, M. A. Morse, J. Desai, A. Hill, M. Axelson, R. A. Moss, M. V. Goldberg, Z. A. Cao, J. M. Ledeine, G. A. Maglinte, S. Kopetz, T. André, Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): An open-label, multicentre, phase 2 study. *Lancet Oncol.* **18**, 1182–1191 (2017).
17. A. Snyder, V. Makarov, T. Merghoub, J. Yuan, J. M. Zaretsky, A. Desrichard, L. A. Walsh, M. A. Postow, P. Wong, T. S. Ho, T. J. Hollmann, C. Bruggeman, K. Kannan, Y. Li, C. Elipenahli, C. Liu, C. T. Harbison, L. Wang, A. Ribas, J. D. Wolchok, T. A. Chan, Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* **371**, 2189–2199 (2014).
18. J. E. Rosenberg, D. P. Petrylak, M. S. Van Der Heijden, A. Necchi, P. H. O'Donnell, Y. Loriot, M. Retz, J. L. Perez-Gracia, J. Bellmunt, P. Grivas, R. W. Joseph, L. Fong, E. E. Kadel, Z. Boyd, D. Nickles, G. M. Frampton, R. Bourgon, P. S. Hegde, S. Mariathasan, T. Powles, PD-L1 expression, Cancer Genome Atlas (TCGA) subtype, and mutational load as independent predictors of response to atezolizumab (atezo) in metastatic urothelial carcinoma (mUC; IMvigor210). *J. Clin. Oncol.* **34**, 104 (2016).
19. N. A. Rizvi, M. D. Hellmann, A. Snyder, P. Kvistborg, V. Makarov, J. J. Havel, W. Lee, J. Yuan, P. Wong, T. S. Ho, M. L. Miller, N. Rekhtman, A. L. Moreira, F. Ibrahim, C. Bruggeman, B. Gasmi, R. Zappasodi, Y. Maeda, C. Sander, E. B. Garon, T. Merghoub, J. D. Wolchok, T. N. Schumacher, T. A. Chan, Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**, 124–128 (2015).
20. A. M. Goodman, S. Kato, L. Bazhenova, S. P. Patel, G. M. Frampton, V. Miller, P. J. Stephens, G. A. Daniels, R. Kurzrock, Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).
21. N. Agrawal, M. J. Frederick, C. R. Pickering, C. Bettegowda, K. Chang, R. J. Li, C. Fakhry, T.-X. Xie, J. Zhang, J. Wang, N. Zhang, A. K. El-Naggar, S. A. Jasser, J. N. Weinstein, L. Treviño, J. A. Drummond, D. M. Muzny, Y. Wu, L. D. Wood, R. H. Hruban, W. H. Westra, W. M. Koch, J. A. Califano, R. A. Gibbs, D. Sidransky, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, D. A. Wheeler, K. W. Kinzler, J. N. Myers, Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in *NOTCH1*. *Science* **333**, 1154–1157 (2011).
22. A. Bardelli, D. W. Parsons, N. Silliman, J. Ptak, S. Szabo, S. Saha, S. Markowitz, J. K. V. Willson, G. Parmigiani, K. W. Kinzler, B. Vogelstein, V. E. Velculescu, Mutational analysis of the tyrosine kinome in colorectal cancers. *Science* **300**, 949 (2003).
23. H. Davies, G. R. Bignell, C. Cox, P. Stephens, S. Edkins, S. Clegg, J. Teague, H. Woffendin, M. J. Garnett, W. Bottomley, N. Davis, E. Dicks, R. Ewing, Y. Floyd, K. Gray, S. Hall, R. Hawes, J. Hughes, V. Kosmidou, A. Menzies, C. Mould, A. Parker, C. Stevens, S. Watt, S. Hooper, R. Wilson, H. Jayatilake, B. A. Gusterson, C. Cooper, J. Shipley, D. Hargrave, K. Pritchard-Jones, N. Maitland, G. Chenevix-Trench, G. J. Riggins, D. D. Bigner, G. Palmieri, A. Cossu, A. Flanagan, A. Nicholson, J. W. Ho, S. Y. Leung, S. T. Yuen, B. L. Weber, H. F. Seigler, T. L. Darrow, H. Paterson, R. Marais, C. J. Marshall, R. Wooster, M. R. Stratton, P. A. Futreal, Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002).
24. S. Jones, X. Zhang, D. W. Parsons, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, H. Kamiyama, A. Jimeno, S.-M. Hong, B. Fu, M.-T. Lin, E. S. Calhoun, M. Kamiyama, K. Walter, T. Nikolskaya, Y. Nikolsky, J. Hartigan, D. R. Smith, M. Hidalgo, S. D. Leach, A. P. Klein, E. M. Jaffee, M. Goggins, A. Maitra, C. Iacobuzio-Donahue, J. R. Eshleman, S. E. Kern, R. H. Hruban, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, K. W. Kinzler, Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
25. D. W. Parsons, S. Jones, X. Zhang, J. C.-H. Lin, R. J. Leary, P. Angenendt, P. Mankoo, H. Carter, I.-M. Siu, G. L. Gallia, A. Olivi, R. McLendon, B. A. Rasheed, S. Keir, T. Nikolskaya, Y. Nikolsky, D. A. Busam, H. Tekleab, L. A. Diaz Jr., J. Hartigan, D. R. Smith, R. L. Strausberg, S. K. Marie, S. M. O. Shinjo, H. Yan, G. J. Riggins, D. D. Bigner, R. Karchin, N. Papadopoulos, G. Parmigiani, B. Vogelstein, V. E. Velculescu, K. W. Kinzler, An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).
26. Y. Samuels, Z. Wang, A. Bardelli, N. Silliman, J. Ptak, S. Szabo, H. Yan, A. Gazdar, S. M. Powell, G. J. Riggins, J. K. Willson, S. Markowitz, K. W. Kinzler, B. Vogelstein, V. E. Velculescu, High frequency of mutations of the PIK3CA gene in human cancers. *Science* **304**, 554 (2004).
27. M. Sausen, R. J. Leary, S. Jones, J. Wu, C. P. Reynolds, X. Liu, A. Blackford, G. Parmigiani, L. A. Diaz Jr., N. Papadopoulos, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, M. D. Hogarty, Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat. Genet.* **45**, 12–17 (2013).
28. T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, S. Szabo, P. Buckhaults, C. Farrell, P. Meeh, S. D. Markowitz, J. Willis, D. Dawson, J. K. Willson, A. F. Gazdar, J. Hartigan, L. Wu, C. Liu, G. Parmigiani, B. H. Park, K. E. Bachman, N. Papadopoulos, B. Vogelstein, K. W. Kinzler, V. E. Velculescu, The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
29. Z. Wang, J. M. Cummins, D. Shen, D. P. Cahill, P. V. Jallepalli, T. L. Wang, D. W. Parsons, G. Traverso, M. Awad, N. Silliman, J. Ptak, S. Szabo, J. K. Willson, S. D. Markowitz, M. L. Goldberg, R. Karess, K. W. Kinzler, B. Vogelstein, V. E. Velculescu, C. Lengauer, Three classes of genes mutated in colorectal cancers with chromosomal instability. *Cancer Res.* **64**, 2998–3001 (2004).
30. L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, J. Ptak, N. Silliman, S. Szabo, Z. Dezso, V. Ustyanksky, T. Nikolskaya, Y. Nikolsky, R. Karchin, P. A. Wilson, J. S. Kaminker, Z. Zhang, R. Croshaw, J. Willis, D. Dawson, M. Shipitsin, J. K. Willson, S. Sukumar, K. Polyak, B. H. Park, C. L. Pethiyagoda, P. V. Pant, D. G. Ballinger, A. B. Sparks, J. Hartigan, D. R. Smith, E. Suh, N. Papadopoulos, P. Buckhaults, S. D. Markowitz, G. Parmigiani, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
31. K. Tomczak, P. Czerwinska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68–A77 (2015).
32. International Cancer Genome Consortium, T. J. Hudson, W. Anderson, A. Artez, A. D. Barker, C. Bell, R. R. Bernabé, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard, A. Guttmacher, M. Guyer, F. M. Hemsley, J. L. Jennings, D. Kerr, P. Klatt, P. Kolar, J. Kusada, D. P. Lane, F. Laplace, L. Youyong, G. Nettekoven, B. Ozenberger, J. Peterson, T. S. Rao, J. Remacle, A. J. Schafer, T. Shibata, M. R. Stratton, J. G. Vockley, K. Watanabe, H. Yang, M. M. Yuen, B. M. Knoppers, M. Bobrow, A. Cambon-Thomsen, L. G. Dressler, S. O. Dyke, Y. Joly, K. Kato, K. L. Kennedy, P. Nicolás, M. J. Parker, E. Rial-Sebbag, C. M. Romeo-Casabona, K. M. Shaw, S. Wallace, G. L. Wiesner, N. Zeps, P. Lichter, A. V. Biankin, C. Chabannon, L. Chin, B. Clément, E. de Alava, F. Degos, M. L. Ferguson, P. Geary, D. N. Hayes, T. J. Hudson, A. L. Johns, A. Kasprzyk, H. Nakagawa, R. Penny, M. A. Piris, R. Sarin, A. Scarpa, T. Shibata, M. van de Vijver, P. A. Futreal, H. Aburatani, M. Bayés, D. D. Botwell, P. J. Campbell, X. Estivill, D. S. Gerhard, S. M. Grimmond, I. Gut, M. Hirst, C. López-Otín, P. Majumder, M. Marra, J. D. McPherson, H. Nakagawa, Z. Ning, X. S. Puente, Y. Ruan, T. Shibata, M. R. Stratton, H. G. Stunnenberg, H. Swerdlow, V. E. Velculescu, R. K. Wilson, H. H. Xue, L. Yang, P. T. Spellman, G. D. Bader, P. C. Boutros, P. J. Campbell, P. Flicek, G. Getz, R. Guigó, G. Guo, D. Haussler, S. Heath, T. J. Hubbard, T. Jiang, S. M. Jones, Q. Li, N. López-Bigas, R. Luo, L. Muthuswamy, B. F. Ouellette, J. V. Pearson, X. S. Puente, V. Quesada, B. J. Raphael, C. Sander, T. Shibata, T. P. Speed, L. D. Stein, J. M. Stuart, J. W. Teague, Y. Totoki, T. Tsunoda, A. Valencia, D. A. Wheeler, H. Wu, S. Zhao, G. Zhou, L. D. Stein, R. Guigó, T. J. Hubbard, Y. Joly, S. M. Jones, A. Kasprzyk, M. Lathrop, N. López-Bigas, B. F. Ouellette, P. T. Spellman, J. W. Teague, G. Thomas, A. Valencia, T. Yoshida, K. L. Kennedy, M. Axton, S. O. Dyke, P. A. Futreal, D. S. Gerhard, C. Gunter, M. Guyer, T. J. Hudson, J. D. McPherson, L. J. Miller, B. Ozenberger, K. M. Shaw, A. Kasprzyk, L. D. Stein, J. Zhang, S. A. Haider, J. Wang, C. K. Yung, A. Cros, Y. Liang, S. Gnaneshan, J. Guberman, J. Hsu, M. Bobrow, D. R. Chalmers, K. W. Hasel, Y. Joly, T. S. Kaan, K. L. Kennedy, B. M. Knoppers, W. W. Lowrance, T. Masui, P. Nicolás, E. Rial-Sebbag, L. L. Rodriguez, C. Vergely, T. Yoshida, S. M. Grimmond, A. V. Biankin, D. D. Bowtell, N. Cloonan, A. deFazio, J. R. Eshleman, D. Etemadmoghadam, B. B. Gardiner, J. G. Kench, A. Scarpa, R. L. Sutherland, M. A. Tempero, N. J. Waddell, P. J. Wilson, J. D. McPherson, S. Gallinger, M. S. Tsao, P. A. Shaw, G. M. Petersen, D. Mukhopadhyay, L. Chin, R. A. DePinho, S. Thayer, L. Muthuswamy, K. Shazand, T. Beck, M. Sam, L. Timms, V. Ballin, Y. Lu, J. Ji, X. Zhang, F. Chen, X. Hu, G. Zhou, Q. Yang, G. Tian, L. Zhang, X. Xing, X. Li, Z. Zhu, Y. Yu, J. Yu, H. Yang, M. Lathrop, J. Tost, P. Brennan, I. Holcatova, D. Zaridze, A. Brazma, L. Egevard, E. Prokhortchouk, R. E. Banks, M. Uhlén, A. Cambon-Thomsen, J. Viksna, F. Ponten, K. Skryabin, M. R. Stratton, P. A. Futreal, E. Birney, A. Borg, A. L. Børresen-Dale, C. Caldas, J. A. Foekens, S. Martin, J. S. Reis-Filho, A. L. Richardson, C. Sotiriou, H. G. Stunnenberg, G. Thoms, M. van de Vijver, L. van't Veer,

F. Calvo, D. Birnbaum, H. Blanche, P. Boucher, S. Boyault, C. Chabannon, I. Gut, J. D. Masson-Jacquemier, M. Lathrop, I. Pauporté, X. Pivot, A. Vincent-Salomon, E. Tabone, C. Theillet, G. Thomas, J. Tost, I. Treilleux, F. Calvo, P. Bioulac-Sage, B. Clément, T. Decaens, F. Degos, D. Franco, I. Gut, M. Gut, S. Heath, M. Lathrop, D. Samuel, G. Thomas, J. Zucman-Rossi, P. Lichter, R. Eils, B. Brors, J. O. Korbel, A. Korshunov, P. Landgraf, H. Lehrach, S. Pfister, B. Radlwimmer, G. Reifenberger, M. D. Taylor, C. von Kalle, P. P. Majumder, R. Sarin, T. S. Rao, M. K. Bhan, A. Scarpa, P. Pederzoli, R. A. Lawlor, M. Delledonne, A. Bardelli, A. V. Biankin, S. M. Grimmond, T. Gress, D. Klimstra, G. Zamboni, T. Shibata, Y. Nakamura, H. Nakagawa, J. Kusada, T. Tsunoda, S. Miyano, H. Aburatani, K. Kato, A. Fujimoto, T. Yoshida, E. Campo, C. López-Otín, X. Estivill, R. Guigó, S. de Sanjosé, M. A. Piris, E. Montserrat, M. González-Díaz, X. S. Puente, P. Jares, A. Valencia, H. Himmelbauer, V. Quesada, S. Bea, M. R. Stratton, P. A. Futreal, P. J. Campbell, A. Vincent-Salomon, A. L. Richardson, J. S. Reis-Filho, M. van de Vijver, G. Thomas, J. D. Masson-Jacquemier, S. Aparicio, A. Borg, A. L. Børresen-Dale, C. Caldas, J. A. Foekens, H. G. Stunnenberg, L. van't Veer, D. F. Easton, P. T. Spellman, S. Martin, A. D. Barker, L. Chin, F. S. Collins, C. C. Compton, M. L. Ferguson, D. S. Gerhard, G. Getz, C. Gunter, A. Guttmacher, M. Guyer, D. N. Hayes, E. S. Lander, B. Ozenberger, R. Penny, J. Peterson, C. Sander, K. M. Shaw, T. P. Speed, P. T. Spellman, J. G. Vockley, D. A. Wheeler, R. K. Wilson, T. J. Hudson, L. Chin, B. M. Knoppers, E. S. Lander, P. Lichter, L. D. Stein, M. R. Stratton, W. Anderson, A. D. Barker, C. Bell, M. Bobrow, W. Burke, F. S. Collins, C. C. Compton, R. A. DePinho, D. F. Easton, P. A. Futreal, D. S. Gerhard, A. R. Green, M. Guyer, S. R. Hamilton, T. J. Hubbard, O. P. Kallioniemi, K. L. Kennedy, T. J. Ley, E. T. Liu, Y. Lu, P. Majumder, M. Marra, B. Ozenberger, J. Peterson, A. J. Schafer, P. T. Spellman, H. G. Stunnenberg, B. J. Wainwright, R. K. Wilson, H. Yang, International network of cancer genome projects. *Nature* **464**, 993–998 (2010).

33. Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
34. Cancer Genome Atlas Network, Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
35. Cancer Genome Atlas Network, Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
36. Cancer Genome Atlas Research Network, Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
37. Cancer Genome Atlas Research Network, Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
38. Cancer Genome Atlas Research Network, Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
39. Cancer Genome Atlas Research Network, Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
40. Cancer Genome Atlas Research Network, Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
41. Cancer Genome Atlas Research Network, C. Kandoth, N. Schultz, A. D. Cherniack, R. Akbani, Y. Liu, H. Shen, A. G. Robertson, I. Pashtan, R. Shen, C. C. Benz, C. Yau, P. W. Laird, L. Ding, W. Zhang, G. B. Mills, R. Kucherlapati, E. R. Mardis, D. A. Levine, Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
42. Cancer Genome Atlas Research Network, Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341.e23 (2017).
43. Cancer Genome Atlas Research Network, TCGA PanCanAtlas https://doi.org/10.7303/syn7214402.
44. E. Izumchenko, X. Chang, M. Brait, E. Fertig, L. T. Kagohara, A. Bedi, L. Marchionni, N. Agrawal, R. Ravi, S. Jones, M. O. Hoque, W. H. Westra, D. Sidransky, Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. *Nat. Commun.* **6**, 8258 (2015).
45. M. N. Nikiforova, A. I. Wald, S. Roy, M. B. Durso, Y. E. Nikiforov, Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. *J. Clin. Endocrinol. Metab.* **98**, E1852–E1860 (2013).
46. D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, A. R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi, M. F. Berger, Memorial Sloan Kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
47. R. R. Singh, K. P. Patel, M. J. Routbort, N. G. Reddy, B. A. Barkoh, B. Handal, R. Kanagal-Shamanna, W. O. Greaves, L. J. Medeiros, K. D. Aldape, R. Luthra, Clinical validation of a next-generation sequencing screen for mutational hotspots in 46 cancer-related genes. *J. Mol. Diagn.* **15**, 607–622 (2013).
48. S. Kadri, B. C. Long, I. Mujacic, C. J. Zhen, M. N. Wurst, S. Sharma, N. McDonald, N. Niu, S. Benhamed, J. H. Tuteja, T. Y. Seiwert, K. P. White, M. E. McNerney, C. Fitzpatrick, Y. L. Wang, L. V. Furtado, J. P. Segal, Clinical validation of a next-generation sequencing genomic oncology panel via cross-platform benchmarking against established amplicon sequencing assays. *J. Mol. Diagn.* **19**, 43–56 (2017).
49. K. J. Hampel, F. B. de Abreu, N. Sidiropoulos, J. D. Peterson, G. J. Tsongalis, Variant call concordance between two laboratory-developed, solid tumor targeted genomic profiling assays using distinct workflows and sequencing instruments. *Exp. Mol. Pathol.* **102**, 215–218 (2017).
50. S. Pant, R. Weiner, M. J. Marton, Navigating the rapids: The development of regulated next-generation sequencing-based clinical trial assays and companion diagnostics. *Front. Oncol.* **4**, 78 (2014).
51. S. P. Strom, Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biol. Med.* **13**, 3–11 (2016).
52. D. Mandelker, L. Zhang, Y. Kemel, Z. K. Stadler, V. Joseph, A. Zehir, N. Pradhan, A. Arnold, M. F. Walsh, Y. Li, A. R. Balakrishnan, A. Syed, M. Prasad, K. Nafa, M. I. Carlo, K. A. Cadoo, M. Sheehan, M. H. Fleischut, E. Salo-Mullen, M. Trottier, S. M. Lipkin, A. Lincoln, S. Mukherjee, V. Ravichandran, R. Cambria, J. Galle, W. Abida, M. E. Arcila, R. Benayed, R. Shah, K. Yu, D. F. Bajorin, J. A. Coleman, S. D. Leach, M. A. Lowery, J. Garcia-Aguilar, P. W. Kantoff, C. L. Sawyers, M. N. Dickler, L. Saltz, R. J. Motzer, E. M. O'Reilly, H. I. Scher, J. Baselga, D. S. Klimstra, D. B. Solit, D. M. Hyman, M. F. Berger, M. Ladanyi, M. E. Robson, K. Offit, Mutation detection in patients with advanced cancer by universal sequencing of cancer-related genes in tumor and normal DNA vs guideline-based germline testing. *JAMA* **318**, 825–835 (2017).
53. K. A. Schrader, D. T. Cheng, V. Joseph, M. Prasad, M. Walsh, A. Zehir, A. Ni, T. Thomas, R. Benayed, A. Ashraf, A. Lincoln, M. Arcila, Z. Stadler, D. Solit, D. M. Hyman, L. Zhang, D. Klimstra, M. Ladanyi, K. Offit, M. Berger, M. Robson, Germline variants in targeted tumor sequencing using matched normal DNA. *JAMA Oncol.* **2**, 104–111 (2016).
54. L. J. Jennings, M. E. Arcila, C. Corless, S. Kamel-Reid, I. M. Lubin, J. Pfeifer, R. L. Temple-Smolkin, K. V. Voelkerding, M. N. Nikiforova, Guidelines for validation of next-generation sequencing-based oncology panels: A joint consensus recommendation of the Association for Molecular Pathology and College of American Pathologists. *J. Mol. Diagn.* **19**, 341–365 (2017).
55. G. J. Weiss, B. R. Hoff, R. P. Whitehead, A. Sangal, S. A. Gingrich, R. J. Penny, D. W. Mallery, S. M. Morris, E. J. Thompson, D. M. Loesch, V. Khemka, Evaluation and comparison of two commercially available targeted next-generation sequencing platforms to assist oncology decision making. *Onco Targets Ther.* **8**, 959–967 (2015).
56. R. M. Squillace, G. M. Frampton, P. J. Stephens, J. S. Ross, V. A. Miller, Comparing two assays for clinical genomic profiling: The devil is in the data. *Onco Targets Ther.* **8**, 2237–2242 (2015).
57. M. Misyura, T. Zhang, M. A. Sukhai, M. Thomas, S. Garg, S. Kamel-Reid, T. L. Stockley, Comparison of next-generation sequencing panels and platforms for detection and verification of somatic tumor variants for clinical diagnostics. *J. Mol. Diagn.* **18**, 842–850 (2016).
58. S. Zhang, "One patient, two cancer DNA tests, two different results" (The Atlantic, 2016); www.theatlantic.com/health/archive/2016/12/cancer-biopsy-genetic-test/510656/.
59. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, G. Getz, Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
60. D. E. Larson, C. C. Harris, K. Chen, D. C. Koboldt, T. E. Abbott, D. J. Dooling, T. J. Ley, E. R. Mardis, R. K. Wilson, L. Ding, SomaticSniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28**, 311–317 (2012).
61. C. T. Saunders, W. S. Wong, S. Swamy, J. Becq, L. J. Murray, R. K. Cheetham, Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
62. Z. Lai, A. Markovets, M. Ahdesmaki, B. Chapman, O. Hofmann, R. McEwen, J. Johnson, B. Dougherty, J. C. Barrett, J. R. Dry, VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
63. E. Reble, C. A. Castellani, M. G. Melka, R. O'Reilly, S. M. Singh, VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia. *Psychiatr. Genet.* **27**, 62–70 (2017).
64. M. Sausen, J. Phallen, V. Adleff, S. Jones, R. J. Leary, M. T. Barrett, V. Anagnostou, S. Parpart-Li, D. Murphy, Q. Kay Li, C. A. Hruban, R. Scharpf, J. R. White, P. J. O'Dwyer, P. J. Allen, J. R. Eshleman, C. B. Thompson, D. S. Klimstra, D. C. Linehan, A. Maitra, R. H. Hruban, L. A. Diaz Jr., D. D. Von Hoff, J. S. Johansen, J. A. Drebin, V. E. Velculescu, Clinical implications of genomic alterations in the tumour and circulation of pancreatic cancer patients. *Nat. Commun.* **6**, 7686 (2015).
65. J. Ding, A. Bashashati, A. Roth, A. Oloumi, K. Tse, T. Zeng, G. Haffari, M. Hirst, M. A. Marra, A. Condon, S. Aparicio, S. P. Shah, Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).
66. L. T. Fang, P. T. Afshar, A. Chhibber, M. Mohiyuddin, Y. Fan, J. C. Mu, G. Gibeling, S. Barr, N. B. Asadi, M. B. Gerstein, D. C. Koboldt, W. Wang, W. H. Wong, H. Y. Lam, An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.* **16**, 197 (2015).
67. J. F. Spinella, P. Mehanna, R. Vidal, V. Saillour, P. Cassart, C. Richer, M. Ouimet, J. Healy, D. Sinnett, SNooPer: A machine learning-based method for somatic variant identification from low-pass next-generation sequencing. *BMC Genomics* **17**, 912 (2016).

68. I. Kalatskaya, Q. M. Trinh, M. Spears, J. D. McPherson, J. M. S. Bartlett, L. Stein, ISOWN: Accurate somatic mutation identification in the absence of normal tissue controls. *Genome Med.* **9**, 59 (2017).
69. S. A. Forbes, D. Beare, N. Bindal, S. Bamford, S. Ward, C. G. Cole, M. Jia, C. Kok, H. Boutselakis, T. De, Z. Sondka, L. Ponting, R. Stefancsik, B. Harsha, J. Tate, E. Dawson, S. Thompson, H. Jubb, P. J. Campbell, COSMIC: High-resolution cancer genetics using the catalogue of somatic mutations in cancer. *Curr. Protoc. Hum. Genet.* **91**, 10.11.11–10.11.37 (2016).
70. S. A. Forbes, D. Beare, H. Boutselakis, S. Bamford, N. Bindal, J. Tate, C. G. Cole, S. Ward, E. Dawson, L. Ponting, R. Stefancsik, B. Harsha, C. Y. Kok, M. Jia, H. Jubb, Z. Sondka, S. Thompson, T. De, P. J. Campbell, COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
71. S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague, M. R. Stratton, U. McDermott, P. J. Campbell, COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
72. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur; Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
73. M. A. Jensen, V. Ferretti, R. L. Grossman, L. M. Staudt, The NCI Genomic Data Commons as an engine for precision medicine. *Blood* **130**, 453–459 (2017).
74. J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, A. Kasprzyk, International Cancer Genome Consortium Data Portal—A one-stop shop for cancer genomics data. *Database* **2011**, bar026 (2011).
75. V. Anagnostou, K. N. Smith, P. M. Forde, N. Niknafs, R. Bhattacharya, J. White, T. Zhang, V. Adleff, J. Phallen, N. Wali, C. Hruban, V. B. Guthrie, K. Rodgers, J. Naidoo, H. Kang, W. Sharfman, C. Georgiades, F. Verde, P. Illei, Q. K. Li, E. Gabrielson, M. V. Brock, C. A. Zahnow, S. B. Baylin, R. B. Scharpf, J. R. Brahmer, R. Karchin, D. M. Pardoll, V. E. Velculescu, Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discov.* **7**, 264–276 (2017).
76. H. Husain, V. E. Velculescu, Cancer DNA in the circulation: The liquid biopsy. *JAMA* **318**, 1272–1274 (2017).
77. J. Phallen, M. Sausen, V. Adleff, A. Leal, C. Hruban, J. White, V. Anagnostou, J. Fiksel, S. Cristiano, E. Papp, S. Speir, T. Reinert, M. W. Orntoft, B. D. Woodward, D. Murphy, S. Parpart-Li, D. Riley, M. Nesselbush, N. Sengamalay, A. Georgiadis, Q. K. Li, M. R. Madsen, F. V. Mortensen, J. Huiskens, C. Punt, N. van Grieken, R. Fijneman, G. Meijer, H. Husain, R. B. Scharpf, L. A. Diaz Jr., S. Jones, S. Angiuoli, T. Orntoft, H. J. Nielsen, C. L. Andersen, V. E. Velculescu, Direct detection of early-stage cancers using circulating tumor DNA. *Sci. Transl. Med.* **9**, eaan2415 (2017).
78. C. Bettegowda, M. Sausen, R. J. Leary, I. Kinde, Y. Wang, N. Agrawal, B. R. Bartlett, H. Wang, B. Luber, R. M. Alani, E. S. Antonarakis, N. S. Azad, A. Bardelli, H. Brem, J. L. Cameron, C. C. Lee, L. A. Fecher, G. L. Gallia, P. Gibbs, D. Le, R. L. Giuntoli, M. Goggins, M. D. Hogarty, M. Holdhoff, S. M. Hong, Y. Jiao, H. H. Juhl, J. J. Kim, G. Siravegna, D. A. Laheru, C. Lauricella, M. Lim, E. J. Lipson, S. K. Marie, G. J. Netto, K. S. Oliner, A. Olivi, L. Olsson, G. J. Riggins, A. Sartore-Bianchi, K. Schmidt, M. Shih, S. M. Oba-Shinjo, S. Siena, D. Theodorescu, J. Tie, T. T. Harkins, S. Veronese, T. L. Wang, J. D. Weingart, C. L. Wolfgang, L. D. Wood, D. Xing, R. H. Hruban, J. Wu, P. J. Allen, C. M. Schmidt, M. A. Choti, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, N. Papadopoulos, L. A. Diaz Jr., Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci. Transl. Med.* **6**, 224ra224 (2014).
79. S. J. Dawson, D. W. Tsui, M. Murtaza, H. Biggs, O. M. Rueda, S. F. Chin, M. J. Dunning, D. Gale, T. Forshew, B. Mahler-Araujo, S. Rajan, S. Humphray, J. Becq, D. Halsall, M. Wallis, D. Bentley, C. Caldas, N. Rosenfeld, Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* **368**, 1199–1209 (2013).
80. T. Forshew, M. Murtaza, C. Parkinson, D. Gale, D. W. Tsui, F. Kaper, S. J. Dawson, A. M. Piskorz, M. Jimenez-Linan, D. Bentley, J. Hadfield, A. P. May, C. Caldas, J. D. Brenton, N. Rosenfeld, Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4**, 136ra168 (2012).
81. S. T. Kim, W. S. Lee, R. B. Lanman, S. Mortimer, O. A. Zill, K. M. Kim, K. T. Jang, S. H. Kim, S. H. Park, J. O. Park, Y. S. Park, H. Y. Lim, H. Eltoukhy, W. K. Kang, W. Y. Lee, H. C. Kim, K. Park, J. Lee, A. Talasaz, Prospective blinded study of somatic mutation detection in cell-free DNA utilizing a targeted 54-gene next generation sequencing panel in metastatic solid tumor patients. *Oncotarget* **6**, 40360–40369 (2015).
82. R. J. Leary, M. Sausen, I. Kinde, N. Papadopoulos, J. D. Carpten, D. Craig, J. O'Shaughnessy, K. W. Kinzler, G. Parmigiani, B. Vogelstein, L. A. Diaz Jr., V. E. Velculescu, Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci. Transl. Med.* **4**, 162ra154 (2012).
83. M. Murtaza, S. J. Dawson, D. W. Tsui, D. Gale, T. Forshew, A. M. Piskorz, C. Parkinson, S. F. Chin, Z. Kingsbury, A. S. Wong, F. Marass, S. Humphray, J. Hadfield, D. Bentley, T. M. Chin, J. D. Brenton, C. Caldas, N. Rosenfeld, Non-invasive analysis of acquired resistance to cancer therapy by sequencing of plasma DNA. *Nature* **497**, 108–112 (2013).
84. A. M. Newman, S. V. Bratman, J. To, J. F. Wynne, N. C. Eclov, L. A. Modlin, C. L. Liu, J. W. Neal, H. A. Wakelee, R. E. Merritt, J. B. Shrager, B. W. Loo Jr., A. A. Alizadeh, M. Diehn, An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).
85. A. M. Newman, A. F. Lovejoy, D. M. Klass, D. M. Kurtz, J. J. Chabon, F. Scherer, H. Stehr, C. L. Liu, S. V. Bratman, C. Say, L. Zhou, J. N. Carter, R. B. West, G. W. Sledge, J. B. Shrager, B. W. Loo Jr., J. W. Neal, H. A. Wakelee, M. Diehn, A. A. Alizadeh, Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.* **34**, 547–555 (2016).
86. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
87. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
88. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
89. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
90. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
91. C. Tyner, G. P. Barber, J. Casper, H. Clawson, M. Diekhans, C. Eisenhart, C. M. Fischer, D. Gibson, J. N. Gonzalez, L. Guruvadoo, M. Haeussler, S. Heitner, A. S. Hinrichs, D. Karolchik, B. T. Lee, C. M. Lee, P. Nejad, B. J. Raney, K. R. Rosenbloom, M. L. Speir, C. Villarreal, J. Vivian, A. S. Zweig, D. Haussler, R. M. Kuhn, W. J. Kent, The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* **45**, D626–D634 (2017).
92. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
93. K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, D. Lipman, The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323 (2009).
94. S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
95. A. Wysoker, K. Tibbetts, T. Fennell, Picard tools version 1.90 (2013).
96. A. D. Ewing, K. E. Houlahan, Y. Hu, K. Ellrott, C. Caloian, T. N. Yamaguchi, J. C. Bare, C. P'ng, D. Waggott, V. Y. Sabelnykova; ICGC-TCGA DREAM. Somatic Mutation Calling Challenge participants, M. R. Kellen, T. C. Norman, D. Haussler, S. H. Friend, G. Stolovitzky, A. A. Margolin, J. M. Stuart, P. C. Boutros, Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
97. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin; 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
98. R. G. Newcombe, Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med.* **17**, 857–872 (1998).
99. E. Wilson, Calculating a confidence interval of a proportion. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).

Acknowledgments: We thank members of our laboratories for critical review of the manuscript. **Funding:** This work was supported, in part, by U.S. NIH grants CA121113 and CA180950 (both to V.E.V.), the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation, the Stand Up to Cancer–Dutch Cancer Society International Translational Cancer Research Dream Team Grant (SU2C-AACR-DT1415, to V.E.V.), the Commonwealth Foundation, the Eastern Cooperative Oncology Group–American College of Radiology Imaging Network, the MacMillan Foundation, and the LUNGevity Foundation. Stand Up To Cancer is a program of the Entertainment Industry Foundation administered by the American Association for Cancer Research. **Author contributions:** D.E.W., J.R.W., L.A.D., V.E.V., and S.V.A. designed the

study, performed experiments and analyses, and wrote the paper. A.G., B.V.E., S.P.-L., V.A., N.N., R.K., S.J., and M.S. performed experiments and analyses and wrote the paper. J.M., E.P., C.M., P.L., and D.R. performed experiments and analyses. **Competing interests:** D.E.W., V.E.V., and S.V.A. are inventors on a patent application (62607007) submitted by Personal Genome Diagnostics related to detection of mutations. V.E.V. is a founder of Personal Genome Diagnostics, is a member of its Scientific Advisory Board and Board of Directors, and owns Personal Genome Diagnostics stock, which is subject to certain restrictions under university policy. V.E.V. is also on the Scientific Advisory Board for Ignyta. The terms of these arrangements are managed by Johns Hopkins University in accordance with its conflict of interest policies. L.A.D. is a founder of Personal Genome Diagnostics and PapGene and a stock owner for both entities, a member of the Personal Genome Diagnostics Board of Directors, and a consultant for Personal Genome Diagnostics, Merck, and Cell Design Labs. **Data and materials availability:** Sequence data used in training from this study are available through the European Genome-phenome Archive (EGA) under accession EGAS00001003152. The Cerebro software for training and testing is available at http://github.com/PGDX/cerebro-paper. All other data associated with this study can be found in the main paper or the Supplementary Materials.

Submitted 18 December 2017
Resubmitted 26 May 2018
Accepted 16 August 2018
Published 5 September 2018
10.1126/scitranslmed.aar7939

Citation: D. E. Wood, J. R. White, A. Georgiadis, B. Van Emburgh, S. Parpart-Li, J. Mitchell, V. Anagnostou, N. Niknafs, R. Karchin, E. Papp, C. McCord, P. LoVerso, D. Riley, L. A. Diaz Jr., S. Jones, M. Sausen, V. E. Velculescu, S. V. Angiuoli, A machine learning approach for somatic mutation discovery. *Sci. Transl. Med.* **10**, eaar7939 (2018).

Science Translational Medicine

A machine learning approach for somatic mutation discovery

Derrick E. Wood, James R. White, Andrew Georgiadis, Beth Van Emburgh, Sonya Parpart-Li, Jason Mitchell, Valsamo Anagnostou, Noushin Niknafs, Rachel Karchin, Eniko Papp, Christine McCord, Peter LoVerso, David Riley, Luis A. Diaz, Jr., Siân Jones, Mark Sausen, Victor E. Velculescu, and Samuel V. Angiuoli

Sci. Transl. Med. **10** (457), eaar7939. DOI: 10.1126/scitranslmed.aar7939

Calling it like the algorithm sees it

Somatic mutation calling is essential for the proper diagnosis and treatment of most cancer patients. Wood *et al*. developed a machine learning approach called Cerebro that increased the accuracy of calling validated somatic mutations in tumor samples from cancer patients. Cerebro outperformed six other mutation detection methods by better distinguishing technical sequencing artifacts. An analysis of non–small cell lung cancer and melanoma patient samples revealed that Cerebro more accurately classified patients according to their immunotherapy response, suggesting that the authors' mutation calling approach could favorably affect patient care.