Cite as: J. D. Cohen *et al*., *Science* 10.1126/science.aar3247 (2018).

First release:

amounts of ctDNA, even when extremely sensitive techniques are used to identify them (*11*, *20*). Many proteins potentially useful for early detection and diagnosis of cancer have been described in the literature (*25*–*27*). We searched this literature to find proteins that had previously been shown to detect at least one of the eight cancer types described above with sensitivities > 10% and specificities > 99%. We identified 41 potential protein biomarkers (table S3) and evaluated them in preliminary studies on plasma samples from normal individuals as well as from cancer patients. We found that 39 of these proteins could be reproducibly evaluated through a single immunoassay platform and we then used this platform to assay all plasma samples (table S3). Eight of the 39 proteins proved to be particularly useful for discriminating cancer patients from healthy controls (table S3).

We then used CancerSEEK to study 1,005 patients who had been diagnosed with Stage I to III cancers of the ovary, liver, stomach, pancreas, esophagus, colorectum, lung, or breast. No patient received neo-adjuvant chemotherapy prior to blood sample collection and none had evident distant metastasis at the time of study entry. The median age at diagnosis was 64 (range 22 to 93). The eight cancer types were chosen because they are common in western populations and because no blood-based tests for their earlier detection are in common clinical use. The histopathological and clinical characteristics of the patients are summarized in table S4. The most common stage at presentation was American Joint Commission on Cancer (AJCC) stage II, accounting for 49% of patients, with the remaining patients harboring stage I (20%) or stage III (31%) disease. The number of samples per stage for each of the eight tumor types is summarized in table S11. The healthy control cohort consisted of 812 individuals of median age 55 (range 17 to 88) with no known history of cancer, high-grade dysplasia, autoimmune disease, or chronic kidney disease.

CancerSEEK evaluates levels of 8 proteins and the presence of mutations in 2,001 genomic positions; each genomic position could be mutated in several ways (single base substitutions, insertions, or deletions). The presence of a mutation in an assayed gene or an elevation in the level of any of these proteins would classify a patient as positive. It was therefore imperative to employ rigorous statistical methods to ensure the accuracy of the test. We used log ratios to evaluate mutations and incorporated them into a logistic regression algorithm that took into account both mutation data and protein biomarker levels to score CancerSEEK test results (Supplementary Materials). The mean sensitivities and specificities were determined by ten iterations of 10-fold cross-validations. The receiver operating characteristic (ROC) curves for the entire cohort of cancer patients and controls in one representative iteration is shown in Fig. 2A.

The median sensitivity of CancerSEEK among the eight cancer types evaluated was 70% ($P < 10^{-96}$ one-sided binomial test) and ranged from 98% in ovarian cancers to 33% in breast cancers (Fig. 2C). At this sensitivity, the specificity was > 99%; i.e., only 7 of the 812 individuals without known cancers scored positive. We could not be certain that the few "false positive" individuals identified among the healthy cohort did not actually have an as yet undetected cancer, but classifying them as false positives provided the most conservative approach to classification and interpretation of the data.

The features of the test that were most important to the algorithm were the presence of a ctDNA mutation followed by elevations of cancer antigen 125 (CA-125), carcinoembryonic antigen (CEA), cancer antigen 19-9 (CA19-9), prolactin (PRL), hepatocyte growth factor (HGF), osteopontin (OPN), myeloperoxidase (MPO), and tissue inhibitor of metalloproteinases 1 (TIMP-1) protein levels (table S9). Waterfall plots for each of the ctDNA and protein features used in CancerSEEK illustrate their distribution among individuals with and without cancer (fig. S2). The importance ranking of the ctDNA and protein features used in CancerSEEK are provided in table S9 and a principal component analysis displaying the clustering of individuals with and without cancer is shown in fig. S3. The complete dataset, including the levels of all proteins studied and the mutations identified in the plasma samples, are provided in tables S5 and S6. The probabilistic rather than deterministic nature of the approach used here to call a sample positive is evident from fig S4; each panel represents the sensitivity of CancerSEEK when one specific feature was excluded from the analysis.

One of the most important attributes of a screening test is the ability to detect cancers at relatively early stages. The median sensitivity of CancerSEEK was 73% for the most common stage evaluated (Stage II), similar (78%) for Stage III cancers, and lower (43%) for Stage I cancers (Fig. 2B). The sensitivity for the earliest stage cancers (Stage I) was highest for liver cancer (100%) and lowest for esophageal cancer (20%).

One limitation of liquid biopsies is their inability to determine the cancer type in patients who test positive, which poses challenges for clinical follow-up. To examine whether the CancerSEEK test can help identify a cancer's tissue of origin, we used supervised machine learning to predict the underlying cancer type in patients with positive CancerSEEK tests. The input algorithm took into account the ctDNA and protein biomarker levels as well as the gender of the patient (Supplementary Materials). One of the main purposes of such predictions is to determine the most appropriate follow-up test for cancer diagnosis or monitoring after a positive CancerSEEK test. We therefore grouped together patients with esophageal and gastric cancers, as endoscopy would be the optimal follow-up in both instances. We then used this algorithm (Supplementary Materials) to study the 626 cancer patients scoring positive in the CancerSEEK Test. Without any clinical information about the patients, we were able to localize the source of the cancer to two anatomic sites in a median of 83% of these patients (Fig. 3, table S8; $P < 10^{-77}$ one-sided binomial test). Furthermore, we were able to localize the source of the positive test to a single organ in a median of 63% of these patients (Fig. 3, table S8; $P < 10^{-47}$ one-sided binomial test). Given that driver gene mutations are usually not tissue-specific, the vast majority of the localization information was derived from protein markers. The accuracy of prediction varied with tumor type; it was highest for colorectal cancers and lowest for lung cancers (Fig. 3 and table S10).

In summary, we have designed a multi-analyte blood test that can detect the presence of eight common solid tumor types. The advantage of combining completely different agents, with distiu3tanin c.6 386n t9 (05)-1 9 (re)-1i1

8. C. G. M , T. R. F , J. S. Mac a , D. G. Ha , J. A. La , C. M. Ta , J. S. U , W. A. E , D. C. T , J. H. G c , M. H. V , J. A. Ma a , F. ac a a c a a , a c a III c ca c a: A a . *Ann. Intern. Med.* **122**, 321 326 (1995). :10.7326/0003-4819-122-5-199503010-00001 M
9. A. C. H a , M. A. P , R. J. O , R. M c , B. B c , S. Ma , W. X , S. Ha , J. R. G , B. W , M. A a , D. K , K. S. Pa a a , C. Ca a, P. W , F. Q a a , B. W bb , K. D'A a, K. E. Pa , R. S. H a , R. P. S a , J. M. Sc , S. McG a , S. K a , S. M. G , R. H. V , R. K. A a a a , G. C. Ka a , L. M. Sc c , X. X , K. L. Na a , J. D. W c , T. C. Ga a a , E. J. W , T-c a b a a c a a -PD-1 . *Nature* **545**, 60 65 (2017). :10.1038/ a 22079 M
10. P. F. P , P. C. P , B. S. K a , P a Ca c Sc - A P c C S a E c . *N. Engl. J. Med.* **376**, 1285 1289 (2017). :10.1056/NEJM b1616281 M
11. C. B a, M. Sa , R. J. L a , I. K , Y. Wa , N. A a a, B. R. Ba , H. Wa , B. L b , R. M. A a , E. S. A a a , N. S. A a , A. Ba , H. B , J. L. Ca , C. C. L , L. A. F c , G. L. Ga a, P. G bb , D. L , R. L. G , M. G , M. D. H a , M. H , S.-M. H , Y. J a , H. H. J , J. J. K , G. S a a, D. A. La , C. La c a, M. L , E. J. L , S. K. N. Ma , G. J. N , K. S. O , A. O , L. O , G. J. R , A. Sa -B a c , K. Sc , M. S , S. M. Oba-S , S. S a, D. T c , J. T , T. T. Ha , S. V , T.-L. Wa , J. D. W a , C. L. W a , L. D. W , D. X , R. H. H ba , J. W , P. J. A , C. M. Sc , M. A. C , V. E. V c c , K. W. K , B. V , N. Pa a , L. A. D a J ., D c c c a DNA a - a a - a a a a c . *Sci. Transl. Med.* **6**, 224 a24 (2014). (K) (.)93.7 ( )-.16701 T -2.3 (c 0 T 3)-5.8 (5

22. F. D. , M. L , D. D a , Y. H , D. S. , S. S ab , L. A. D a J ., S. N. G a , K. A. Da , H. J , K. W. K , B. V , D c a a ca a a a a c c a . *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16368 16373 (2005). :10.1073/ a .0507904102 M
23. S. A. F b , D. B a , H. B a , S. Ba , N. B a , J. Ta , C. G. C , S. Wa , E. Da , L. P , R. S a c , B. Ha a, C. Y. K , M. J a, H. J bb, Z. S a, S. T. , T. D , P. J. Ca b , COSMIC: S a c ca c c a - . *Nucleic Acids Res.* **45** (D1), D777 D783 (2017). :10.1093/ a / 1121 M
24. I. K , J. W , N. Pa a , K. W. K , B. V , D c a a ca a a a a , a a c . *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9530 9535 (2011). :10.1073/ a .1105422108 M
25. L. A. L a, E. F. P c 3 , T. c . *Clin. Adv. Hematol. Oncol.* **1**, 460 462 (2003). M
26. H. Wa , T. S. , W.-J. Q a , T. L , J. Ka a , S. S a a, R. D. S , K. D. R a , D. G. Ca 2 , T. c ca ac c a c LC-MS ca c b a c , a ca . *Expert Rev. Proteomics* **13**, 99 114 (2016). :10.1586/14789450.2016.1122529 M
27. E. F. Pa J ., M. J. Ca 5 ( )11.1V8 ( )18.04 36 50 2 0.5 ( )1..6 (cBC)3.9 (0.5 (,.1 9 (.)9 8.04 Ba M.( )16.8 G.7 (.8 0 121)-0.7 6T ( ))1 ( )6 0.0 19 Tc 0 T.( )16.8 Ka M

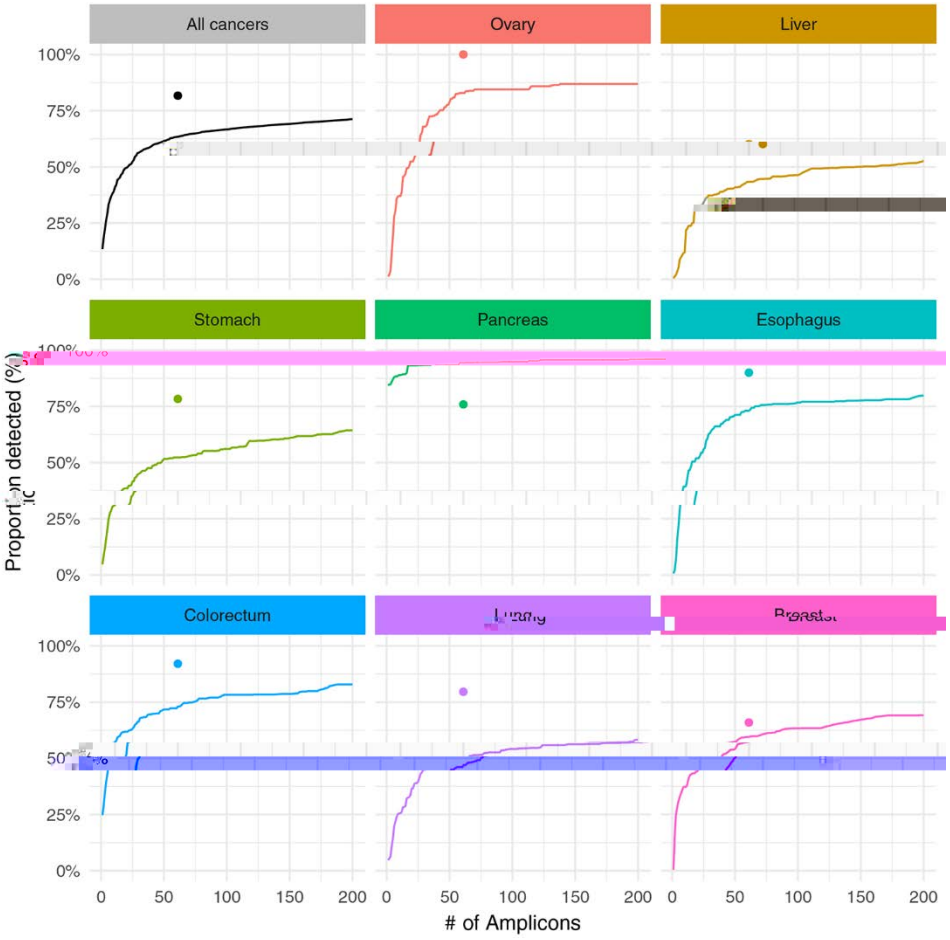

Fig. 1. De el e f a PCR-ba ed a ide if - ecific ai i ha a a le . C l ed c e i dica e he i f ca ce f he eigh e e al a ed i hi d ha ca be de ec ed i h a i c ea i g be f h (< 40 b) a lic . The e i i f de ec i i c ea e i h he be f a lic b ha ea a 60 a lic . C l ed d i dica e he f ac i f ca ce de ec ed i g he 61-a lic a el ed i 805 ca ce e al a ed i d , hich a e aged 82% (ee ai e). P blicl a ailable e e ci g da a e e b ai ed f he Ca al g f S a ic M a i i Ca ce (COSMIC) e i .

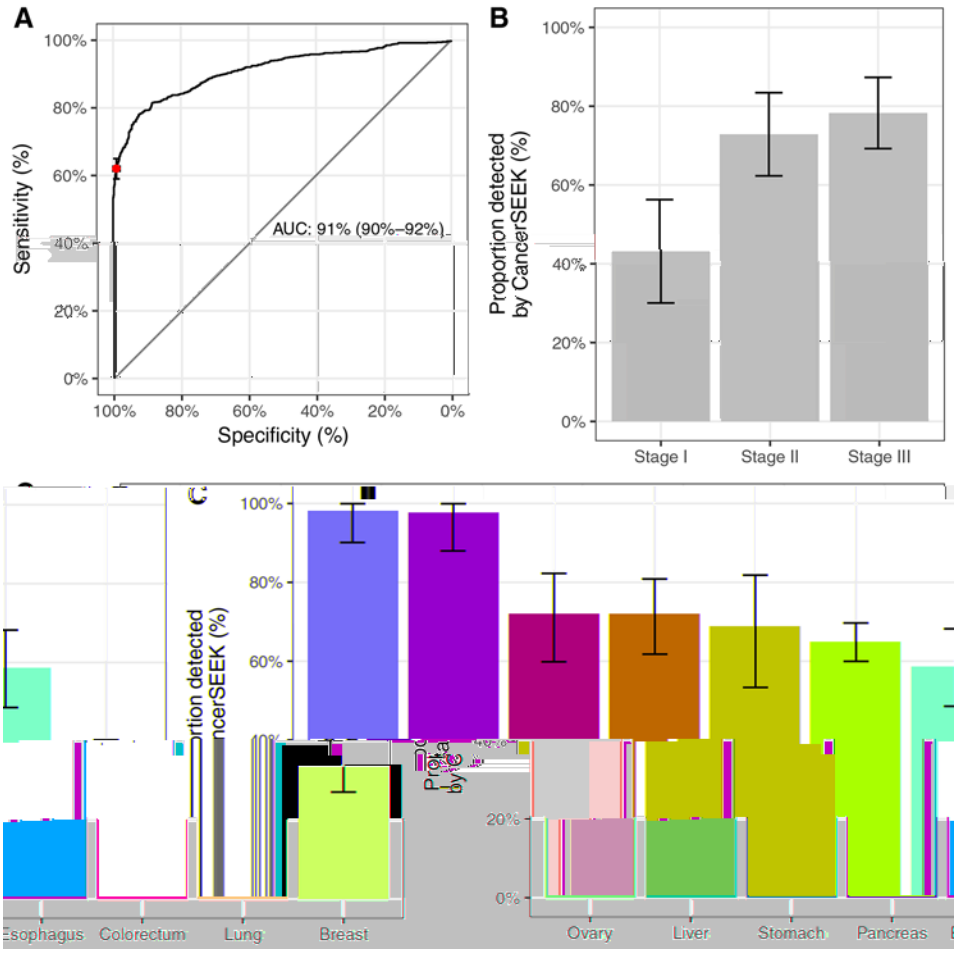

Fig. 2. Pe f a ce f Ca ce SEEK. (A) Recei e a cha ac e i ic (ROC) c e f Ca ce SEEK. The ed i he c e i dica e he e a e age e f a ce (62%)

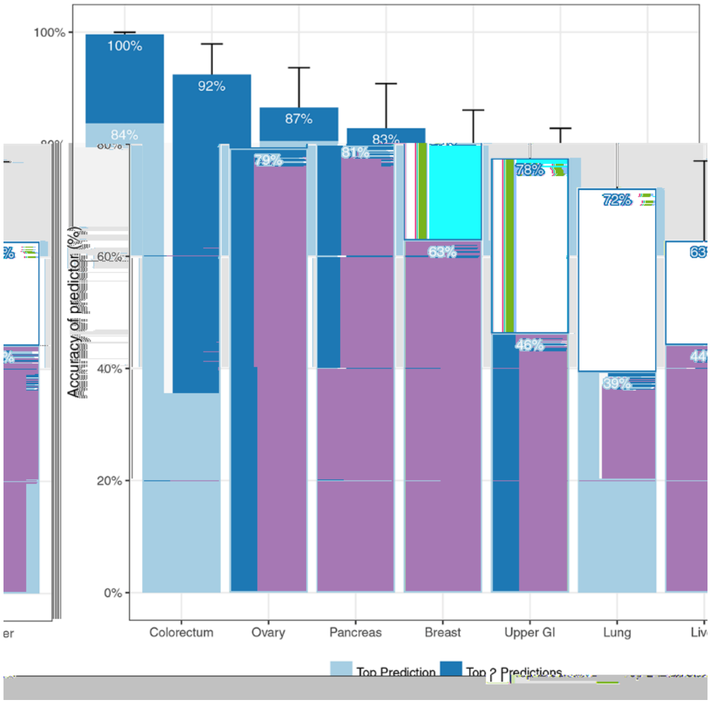

Fig. 3. Identification of cancer type by supervised machine learning for patients classified as positive by CancerSEEK. Percentages correspond to the proportion of patients correctly classified by one of the two most likely types (sum of light and dark blue bars) or the most likely type (light blue bar). Predictions for all patients for all cancer types are provided in table S8. Error bars represent 95% confidence intervals.

Detection and localization of surgically resectable cancers with a multi-analyte blood test

Joshua D. Cohen, Lu Li, Yuxuan Wang, Christopher Thoburn, Bahman Afsari, Ludmila Danilova, Christopher Douville, Ammar A. Javed, Fay Wong, Austin Mattox, Ralph. H. Hruban, Christopher L. Wolfgang, Michael G. Goggins, Marco Dal Molin, Tian-Li Wang, Richard Roden, Alison P. Klein, Janine Ptak, Lisa Dobbyn, Joy Schaefer, Natalie Silliman, Maria Popoli, Joshua T. Vogelstein, James D. Browne, Robert E. Schoen, Randall E. Brand, Jeanne Tie, Peter Gibbs, Hui-Li Wong, Aaron S. Mansfield, Jin Jen, Samir M. Hanash, Massimo Falconi, Peter J. Allen, Shibin Zhou, Chetan Bettegowda, Luis Diaz, Cristian Tomasetti, Kenneth W. Kinzler, Bert Vogelstein, Anne Marie Lennon and Nickolas Papadopoulos

published online January 18, 2018