nature methods

BRIEF COMMUNICATION

https://doi.org/10.1038/s41592-018-0036-9

DeTiN: overcoming tumor-in-normal contamination

Amaro Taylor-Weiner[1,2,13], Chip Stewart[1,13], Thomas Giordano[3], Mendy Miller[1], Mara Rosenberg[1], Alyssa Macbeth[1], Niall Lennon[1], Esther Rheinbay[1], Dan-Avi Landau[1,4,5,6], Catherine J. Wu[1,7,8,9] and Gad Getz[1,10,11,12]*

Comparison of sequencing data from a tumor sample with data from a matched germline control is a key step for accurate detection of somatic mutations. Detection sensitivity for somatic variants is greatly reduced when the matched normal sample is contaminated with tumor cells. To overcome this limitation, we developed deTiN, a method that estimates the tumor-in-normal (TiN) contamination level and, in cases affected by contamination, improves sensitivity by reclassifying initially discarded variants as somatic.

To accurately detect somatic mutations, it is necessary to distinguish between somatic and germline (inherited) variants. Comparison of DNA-sequencing data from tumor and patient-matched control (normal) tissue allows for the removal of patient-specific inherited variants and locus-specific artifacts that affect both samples. This variant-detection paradigm provides sensitive and specific somatic mutation calls with low false positive rates (<0.5 mutations per megabase)[1], but it relies on sequencing data from matched normal healthy tissue that is free of contaminating tumor cells[1–3]. Procuring pure normal tissue can be challenging[4–7].

Tumor-sample DNA found in the normal sample, known as TiN contamination (Methods), arises from the invasion of healthy compartments by cancer or precancer cells and is reported in leukemias[6,8,9] and in breast, bladder, and gastric cancers[10–12], among others. TiN contamination may cause the rejection of true somatic variants on the basis of tumor-derived reads that support the mutation in the matched normal tissue, thereby decreasing sensitivity for mutation detection and leading to potential misinterpretation of patient sequencing data (Supplementary Fig. 1a). To overcome these challenges, we developed deTiN, a method that estimates TiN levels and salvages many somatic mutations that would otherwise be filtered out as germline or artifactual variants.

deTiN models a normal sample as a mixture of normal cells with an unknown fraction of contaminating tumor cells. We estimate TiN, defined as the relative tumor DNA fraction in normal and tumor samples, by using two independent types of tumor-specific events: (i) somatic single-nucleotide variants (SSNVs), and (ii) genomic regions of allelic imbalance (deletions, amplifications, copy-neutral loss of heterozygosity) extracted from allele-specific somatic copy-number alterations (aSCNAs) (Supplementary Fig. 1b, Methods). DeTiN calculates posterior distributions over TiN values based on each of the two somatic event types separately, and then combines them to identify the maximum a posteriori value and confidence interval. The estimated TiN is used to recover previously rejected SSNVs or insertions/deletions by means of a probabilistic comparison of two scenarios for each candidate variant, in which the alternative allele count in the normal represents either (i) an underlying germline variant or (ii) a somatic variant coming from tumor DNA mixed in with the normal according to the estimated TiN value (Supplementary Fig. 1b, Methods).

We carried out in silico and in vitro simulation experiments to measure deTiN's accuracy in estimating TiN and its ability to recover SSNVs. Somatic mutations in pairs of tumors and artificially contaminated normal samples were first called with MuTect[1] (Methods) and then processed by deTiN. deTiN estimated TiN contamination with mean absolute errors of 0.01 (in silico) and 0.02 (in vitro) over the range of simulated TiN values (Fig. 1a,b, Supplementary Tables 1 and 2).

We quantified the effect of TiN contamination on SSNV-detection sensitivity. MuTect[1], VarScan[3], and Strelka[2] lost sensitivity for SSNV detection at TiN > 0.02 (Fig. 1c,d, Supplementary Tables 1 and 2, Supplementary Fig. 2, Supplementary Results). TiN mostly affects mutations with a high allele fraction in the tumor (AF), as these mutations are more likely to be observed in the contaminated normal and cause the mutation caller to reject the somatic mutation. Indeed, mutations with AF > 0.3 were detected with lower sensitivity than those with AF < 0.3 (Mann–Whitney one-tailed $P = 0.004$; in silico TiN = 0.2) (Supplementary Fig. 3a,b). Application of deTiN's mutation recovery step improved detection sensitivity across all TiN values (Fig. 1c,d). At very high TiN values (>0.75), where germline single-nucleotide polymorphisms (SNPs) were indistinguishable from somatic events, SSNV recovery was less effective. DeTiN-recovered mutations did not substantially increase false positive rates (Fig. 1c,d, Supplementary Fig. 3c–f, Supplementary Results) and, as expected, were enriched with high-AF events (Supplementary Fig. 3a,b). High-AF SSNVs are more likely to be clonal mutations, thus representing many initiating drivers and clinically important oncogenic events. We characterized deTiN's performance by using simulated data over a range of tumor-sample purities, sequencing depths, and mutation rates (Supplementary Fig. 4, Supplementary Results).

[1]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [2]Harvard University, Cambridge, MA, USA. [3]Department of Pathology, University of Michigan, Ann Arbor, MI, USA. [4]Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [5]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [6]New York Genome Center, New York, NY, USA. [7]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [8]Department of Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA. [9]Department of Medicine, Harvard Medical School, Boston, MA, USA. [10]Department of Pathology, Harvard Medical School, Boston, MA, USA. [11]Cancer Center, Massachusetts General Hospital, Boston, MA, USA. [12]Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. [13]These authors contributed equally: Amaro Taylor-Weiner and Chip Stewart. *e-mail: gadgetz@broadinstitute.org

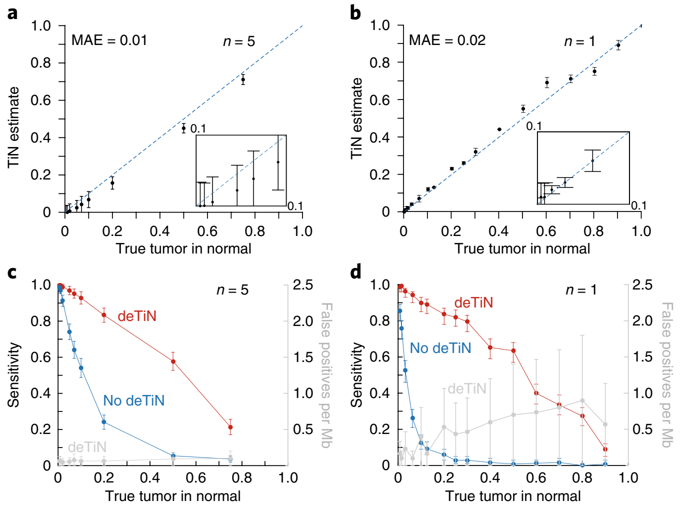

Fig. 1 | Results from in silico and in vitro validation of deTiN. a, TiN estimates at different in silico simulated TiN levels. **b**, deTiN estimates at different in vitro mixed TiN levels. **a**,**b**, Dashed blue lines indicate $y = x$. MAE, mean absolute error. Insets highlight TiN values between 0 and 0.1. **c**,**d**, Sensitivity to detect mutations with (red) and without (blue) deTiN at different (**c**) in silico simulated and (**d**) in vitro mixed TiN levels. False positives per megabase are shown in gray (right-hand *y*-axes). **a**,**c**, deTiN results from $n = 5$ in silico independent simulation experiments. Dots, weighted average; error bars, ±s.e. **b**,**d**, Results from $n = 1$ sequencing experiment; error bars indicate the 95% confidence interval on TiN estimates.

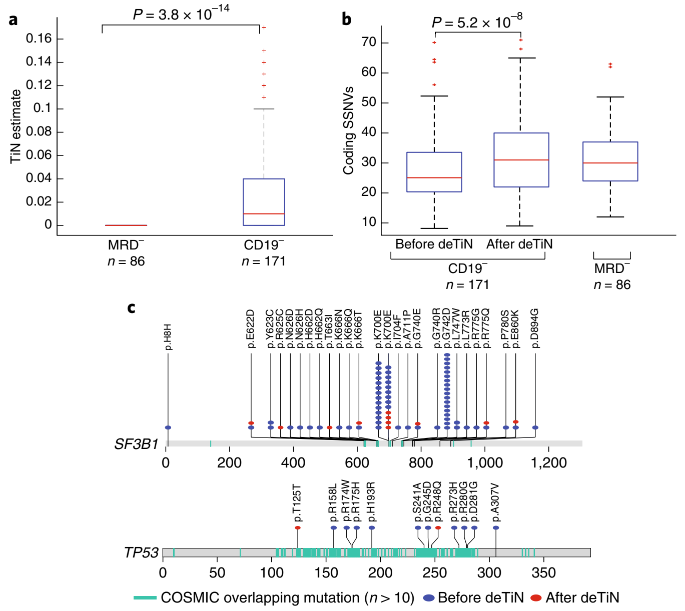

Fig. 2 | Application of deTiN to CLL sequencing data. a, TiN estimates for CD19$^-$ selected (normal) blood compared with whole blood from MRD$^-$ patients. Red line, median; box edges, first and third quartiles; whiskers, full data range excluding outliers; red crosses, outliers defined as >1.5× the interquartile range beyond the first or third quartiles. *P* value determined by two-tailed Mann–Whitney test ($n = 257$ independent patient samples). **b**, Mutation rate in samples before and after application of deTiN, stratified by normal sample type. Box plot elements and *P* value as in **a**. **c**, Stick plots showing mutation data for *SF3B1* and *TP53*. Amino acid positions of recurrent COSMIC mutations are highlighted in teal. Blue circles indicate variants detected before deTiN; red circles indicate variants recovered by deTiN.

Our assumption to this point has been that all tumor cells contaminating the normal sample harbor the same somatic events as the tumor cells in the tumor sample. However, this assumption may be invalid if (i) the tumor cells in a tumor-adjacent normal tissue sample (a common source of 'normal' tissue) contain tumor subclones that differ from the dominant clone in the tumor sample, or (ii) normal-appearing cells are the descendants of a premalignant precursor and share a subset of clonal events with the neighboring tumor cells[5,11,13]. Thus, multiple TiN values may be required to describe the contaminating clones in a single normal sample. We used the tumor and normal cell lines selected for the in vitro experiments as a model to test this phenomenon. At each simulated TiN fraction, deTiN identified two distinct TiN levels: (i) the intended mixing fraction and (ii) a fraction corresponding to a shared precursor subclone (Supplementary Fig. 5). Presence of the parental clone did not interfere with TiN estimation.

We applied deTiN to a whole-exome-sequencing dataset generated from a cohort of 257 tumor–normal paired samples from subjects with chronic lymphocytic leukemia (CLL)[9]. Leukemic DNA was extracted from selected CD19$^+$ cells, and matched germline DNA was derived from either the negative fraction ('sorted CD19$^-$ cells') or matched post-treatment samples without molecularly detectable disease (MRD$^-$; Fig. 2a). DeTiN identified higher TiN contamination in sorted CD19$^-$ cells than in MRD$^-$ samples (Fig. 2a; Mann–Whitney $P < 0.001$). In one case, the CD19$^-$ normal sample was contaminated, but a corresponding saliva-derived sample was not (Supplementary Fig. 6, Supplementary Results). Consistent with the simulation results, mutation calling without deTiN on 171 tumors with CD19$^-$ normals resulted in a markedly lower mutation rate (Mann–Whitney $P < 0.001$). After deTiN application, CD19$^-$ and MRD$^-$ mutation rates became similar ($P = 0.56$; Fig. 2b, Supplementary Table 3). The fraction of candidate mutations at sites from the dbSNP database was not statistically different between tumor samples paired with CD19$^-$ normals and those with MRD$^-$ normals, which suggests that deTiN did not increase the putative false positive SNV rate ($P = 0.27$; Supplementary Table 3). DeTiN recovered mutations in known CLL drivers[9] (Supplementary Fig. 7) at previously reported hot spots, supporting their functional oncogenic role[14] (Fig. 2c).

We also assessed contamination in tumor-adjacent histologically normal tissue[7,15–17], and found significant TiN in 161 of 1,477 tumor and adjacent normal sample pairs (i.e., TiN ≥ 0.02) (Supplementary Table 4). The fraction of samples containing detectable TiN varied by tumor type. Breast invasive carcinoma and testicular germ cell tumors (both non–Cancer Genome Atlas (TCGA) cohorts) showed a significantly higher fraction of cases with TiN > 0.02 (Mann–Whitney $P < 0.01$) and TiN levels per case (Fig. 3a, Supplementary Fig. 8), perhaps owing to the different tissue-collection protocols compared with those used for the TCGA samples. For 304 of 1,477 cases, a matched germline peripheral blood sample was also available and was uncontaminated. Comparison of the mutation calls detected from the tissue-adjacent and blood normal samples demonstrated deTiN's improved sensitivity (Fig. 3b, Supplementary Results).

Histological review by a pathologist blinded to the TiN estimates identified areas of malignant cells in three of eight selected high-TiN cases (prostate adenocarcinoma cells, evidence of dysplastic glands, areas of pancreatic intraepithelial neoplasia-2 (PANIN-2) (Fig. 3c)), but none in eight uncontaminated (TiN = 0) control cases. Notably, deTiN detected *KRAS* G12A mutations in one sample pair, and large copy-number events in all eight contaminated samples (Fig. 3c, Supplementary Fig. 9), suggesting that somatic lesions can be present in histologically nonmalignant tissue and occur before full transformation[18]. Because the sequencing samples originated from

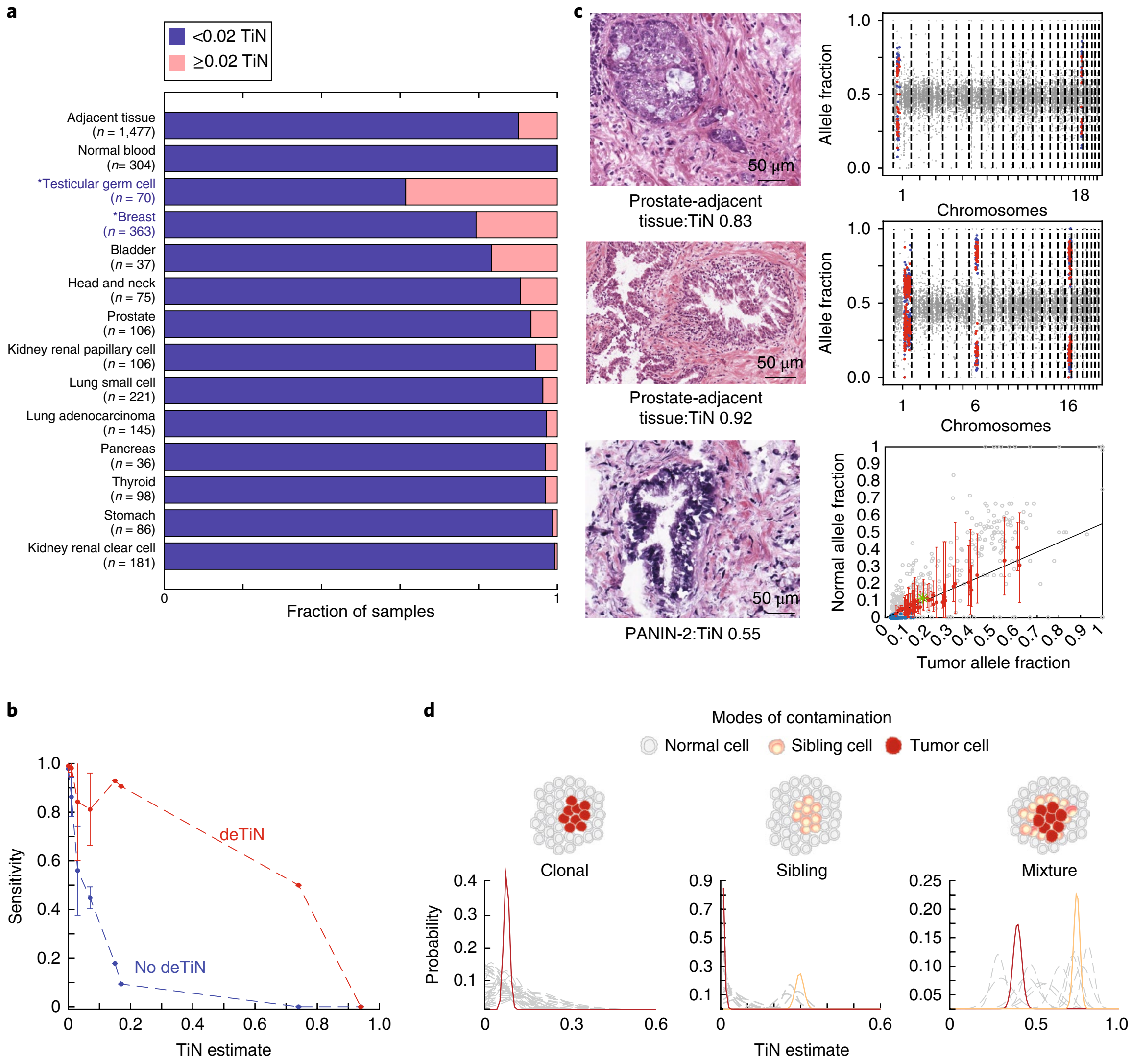

Fig. 3 | Application of deTiN to analysis of solid tumors with adjacent normal controls. **a**, Fraction of contaminated samples (pink; TiN ≥ 0.02) observed when different sources were used for normal tissue (tumor-adjacent normal tissue and peripheral blood) and, in cases with tumor-adjacent normals, stratified by tumor type. Asterisks indicate non-TCGA cohorts. **b**, Mean sensitivity for detecting mutations with (red) and without (blue) deTiN. Means were derived from 256 of the 304 tumors that were matched with both a tumor-adjacent and a blood normal sample and had a sufficient number of somatic events for robust estimation of TiN (TiN = 0, *n* = 230; TiN = 0.01, *n* = 9; TiN = 0.03, *n* = 9; TiN = 0.07, *n* = 4; TiN = 0.15, *n* = 1; TiN = 0.17, *n* = 1; TiN = 0.74, *n* = 1; TiN = 0.94, *n* = 1). Error bars indicate ± s.e. **c**, Histology images of selected adjacent tissue samples with evidence supporting TiN (*n* = 1 patient sample for each image and plot). deTiN aSCNA data supporting the TiN estimate are shown for the top two samples. Points indicate the allele fraction of heterozygous germline SNPs: blue (tumor) and red (normal) points were used for TiN estimation, and gray points were not used by deTiN. The bottom plot shows deTiN somatic variant data supporting the TiN estimate for the bottom sample. Points indicate the allele fraction of variants in the tumor and normal samples; error bars, 95% beta confidence interval; green asterisk, *KRAS* G12V mutation; red points, SSNVs recovered by deTiN; blue points, called before deTiN; gray points, rejected by deTiN and MuTect as germline or artifact. Each plot shows data supporting TiN from a single tumor–normal pair corresponding to the image on the left (*n* = 1). **d**, Illustration of three modes of contamination. Posterior distributions for TiN based on aSCNA data are shown clustered (red and orange curves) and unclustered for individual events (dashed gray curves). In the mixture scenario, TiN has two possible values: the lower represents events unique to the tumor cells (red), and the higher represents events shared between the tumor cells and the sibling precursor cells (orange).

tissue blocks and the histologically evaluated image reflects only the top and bottom slices, we cannot rule out the presence of cancer cells in the sequenced sample due to spatial heterogeneity.

Spatial heterogeneity can result in three TiN contamination types: (i) clonal, sharing all somatic events at a consistent ratio; (ii) one or more sibling clones (e.g., precursor cells), sharing only a subset of events; and (iii) both types i and ii (Fig. 3d). We identified 13 sample pairs from six different tumor types demonstrating sibling or mixture relationships (Supplementary Table 5). In one breast invasive carcinoma–adjacent normal pair, chr1q and chr16q amplifications were present in both samples, but all other aSCNAs were absent, which suggested that the amplifications occurred in a shared precursor clone (sibling model; Fig. 3d, Supplementary Table 5). In a prostate adenocarcinoma–adjacent normal pair, most aSCNAs were consistent with TiN = 0.4, but some focal deletions were present at 0.7 TiN (mixture model; Fig. 3d). Manual review of deTiN's output showed that two adjacent normal samples contained arm-level aSCNAs that were absent in the tumor. In one particularly striking case, deTiN's allele-specific model discerned that a chr1q amplification was present in both breast carcinoma and its adjacent normal, but on opposite alleles, thus demonstrating convergent evolution (Supplementary Fig. 10).

In summary, deTiN is a mixture model that integrates evidence from candidate somatic events and copy-number alterations to provide robust TiN estimates used to infer the somatic status of candidate variants. Our analysis quantified TiN in cases with both adjacent normal tissue and normal blood. In particular, TiN contamination may affect normal samples derived retrospectively from formalin-fixed, paraffin-embedded tumor blocks. Although no TiN was identified in 304 TCGA blood normal samples, TiN may be a factor in metastatic cases. TCGA samples, mostly obtained from untreated resected primary tumors, may have lower circulating tumor cells and DNA levels[19,20]. DeTiN is currently used in large-scale cancer analyses and in the International Cancer Genome Consortium/TCGA Pan-Cancer Analysis of Whole Genomes (https://dcc.icgc.org/pcawg) project (the Supplementary Note, Supplementary Table 6, and Supplementary Fig. 11 present details related to the running of deTiN). Future developments of deTiN (or similar) methods could exploit additional data sources to improve accuracy, including independent sequencing (e.g., RNA-seq), additional patient-matched biopsies, and structural variants.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41592-018-0036-9.

Received: 10 November 2017; Accepted: 25 April 2018;
Published online: 25 June 2018

References

1. Cibulskis, K. et al. *Nat. Biotechnol.* **31**, 213–219 (2013).
2. Saunders, C. T. et al. *Bioinformatics* **28**, 1811–1817 (2012).
3. Koboldt, D. C. et al. *Genome Res.* **22**, 568–576 (2012).
4. Stieglitz, E. et al. *Nat. Genet.* **47**, 1326–1333 (2015).
5. Wei, L. et al. *BMC Med. Genomics* **9**, 64 (2016).
6. The Cancer Genome Atlas Research Network. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
7. Taylor-Weiner, A. et al. *Nature* **540**, 114–118 (2016).
8. Welch, J. S. et al. *Cell* **150**, 264–278 (2012).
9. Landau, D. A. et al. *Nature* **526**, 525–530 (2015).
10. Deng, G., Lu, Y., Zlotnikov, G., Thor, A. D. & Smith, H. S. *Science* **274**, 2057–2059 (1996).
11. Försti, A., Louhelainen, J., Söderberg, M., Wijkström, H. & Hemminki, K. *Eur. J. Cancer* **37**, 1372–1380 (2001).
12. Leung, W. K. et al. *Cancer* **91**, 2294–2301 (2001).
13. Braakhuis, B. J. M., Tabor, M. P., Kummer, J. A., Leemans, C. R. & Brakenhoff, R. H. *Cancer Res.* **63**, 1727–1730 (2003).
14. Forbes, S. A. et al. *Nucleic Acids Res.* **43**, D805–D811 (2015).
15. Rheinbay, E. et al. *Nature* **547**, 55–60 (2017).
16. Van Allen, E. M. et al. *Nat. Med.* **20**, 682–688 (2014).
17. Giannakis, M. et al. *Cell Rep.* **15**, 857–865 (2016).
18. Kanda, M. et al. *Gastroenterology* **142**, 730–733 (2012).
19. Bettegowda, C. et al. *Sci. Transl. Med.* **6**, 224ra24 (2014).
20. Schwarzenbach, H., Hoon, D. S. B. & Pantel, K. *Nat. Rev. Cancer* **11**, 426–437 (2011).

Acknowledgements

This work was funded in part by the NIH TCGA Genome Data Analysis Center (U24CA143845 to G.G.), the Paul C. Zamecnick, MD, Chair in Oncology at MGH Cancer Center (G.G.), the National Human Genome Research Institute, NIH (T32 HG002295 to A.T.-W.), and the NIH (NCI P01CA206978-01, R01CA182461-01, U10CA180861-01, and R01CA184922-02 to A.T.-W., C.S., and C.J.W.). C.J.W. is a Scholar of the Leukemia and Lymphoma Society.

Author contributions

A.T.-W., C.S., and G.G. outlined and planned development. A.T.-W. and C.S. developed the method. A.T.-W. and M.R. performed genomic analysis of large data cohorts. A.T.-W., A.M., and N.L. performed and analyzed in vitro simulations. A.T.-W. and C.S. performed and analyzed in silico simulations. E.R., D.-A.L., and C.J.W. enabled sample acquisition for data analysis. T.G. provided histopathology review of TCGA healthy tumor-adjacent tissue samples. A.T.-W., M.M., C.S., C.J.W., and G.G. prepared the manuscript and figures.

Competing interests

C.J.W. is a cofounder of Neon Therapeutics and a member of its scientific advisory board.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0036-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to G.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Overview of deTiN. DeTiN measures TiN (θ) contamination by comparing sequencing data from matched tumor and normal samples. DeTiN uses two statistical (generative mixture) models to estimate TiN. The first uses aSCNAs, and the second uses SSNVs. Each model generates a posterior probability distribution for TiN. If both models are used, deTiN computes the joint posterior distribution. DeTiN reports the maximum a posteriori point estimate for TiN and a 95% confidence interval based on each model and their combination. Next, deTiN uses the TiN estimate to reclassify candidate variants detected in the tumor sample as either somatic or germline, on the basis of the allele counts observed in the normal at the corresponding sites. Below, we describe the inference steps in which we estimate TiN via an expectation-maximization procedure using SSNVs, and maximum a posteriori estimation using aSCNAs, as well as the application of these estimates for somatic variant reclassification (i.e., rescue of previously rejected somatic variants).

Defining TiN. DeTiN estimates the relative abundance of tumor DNA in the normal sample compared with that in the tumor sample.

$$\theta \;=\; \mathrm{TiN} \;=\; \left(\frac{\text{DNA from tumor cells in the normal sample}}{\text{Total DNA in the normal sample}}\right) \times \left(\frac{\text{Total DNA in the tumor sample}}{\text{DNA from tumor cells in the tumor sample}}\right)$$

Note that, for simplicity, we define TiN as the relative abundance of tumor DNA to circumvent the need to estimate the purity (percentage of tumor cells) and ploidy (average DNA content of the tumor cells) of the tumor sample. Thus, in the uncommon scenario in which the normal sample has a higher fraction of tumor-derived DNA than the tumor sample, TiN may theoretically exceed 1. In our analysis, we assume that TiN ≤ 1, and in reality it is typically $\ll 1$. If the purity (α) and ploidy (τ) of the tumor cells are known (or estimated, for example, by ABSOLUTE[21]), then the TiN estimate (θ) can be used to calculate the actual fraction of tumor cells in the normal sample (β) via the equation (Supplementary Fig. 11)

$$\theta = \left(\frac{\beta}{\beta\tau + 2(1-\beta)}\right)\left(\frac{\alpha\tau + 2(1-\alpha)}{\alpha}\right)$$

Input data. The raw inputs to deTiN are (i) prefiltered variants (including SNVs and indels, both somatic and germline (see "Filtering of SSNVs")) that are observed in the tumor sample, annotated with the corresponding read counts from both tumor and normal samples; and (ii) segmented tumor aSCNAs.

(i) For each variant v, the random variables f_v^{n} and f_v^{t} denote the underlying alternative allele fractions in the tumor (t) and normal (n) samples, respectively. The variables follow beta distributions conditional on the observed read counts for the reference and alternative alleles in the tumor and normal, $(r_v^{\mathrm{t}}, r_v^{\mathrm{n}})$ and $(a_v^{\mathrm{t}}, a_v^{\mathrm{n}})$, respectively. The total coverage in each sample $(h_v^{\mathrm{n}}, h_v^{\mathrm{t}})$ is taken as the sum of the alternative and reference counts (ignoring the other alleles).

$$f_v^{\mathrm{n}} \,|\, a_v^{\mathrm{n}}, r_v^{\mathrm{n}} \sim \mathrm{Beta}(a_v^{\mathrm{n}} + 1, r_v^{\mathrm{n}} + 1)$$
$$f_v^{\mathrm{t}} \,|\, a_v^{\mathrm{t}}, r_v^{\mathrm{t}} \sim \mathrm{Beta}(a_v^{\mathrm{t}} + 1, r_v^{\mathrm{t}} + 1)$$

(ii) The aSCNA input data for the tumor are a set of segments, **S**, representing aSCNAs (see "Filtering of segments and SNPs"), each with a corresponding tumor total copy ratio R_s^{t} and a set of associated heterozygous germline SNPs within the segment, $(v_1, \ldots, v_{N_S})$. Using the normal data, we first calculate the mean allele fraction (of the non-reference allele) across all heterozygous SNPs (N) to represent the balanced allele fraction (which can slightly deviate from 0.5 owing to hybrid capture bias toward the reference):

$$\mu^{\mathrm{n}} = \frac{1}{N}\sum_{v=1}^{N} \frac{a_v^{\mathrm{n}}}{a_v^{\mathrm{n}} + r_v^{\mathrm{n}}}$$

Model. DeTiN compares two models: (i) H_0, no tumor-in-normal, where $\theta = 0$; and (ii) H_1, some tumor-in-normal, where $0 < \theta \leq 1$. The prior probability of H_1, π, is set according to the estimated risk of contamination from malignant cells in the normal, which can depend on the tumor type and the source of normal sample. For example, when using a tissue-adjacent normal, we set $\pi = 0.5$, and when using a blood normal we use $\pi = 0.05$. Under model H_1 we assume a uniform prior distribution for θ.

Model based on aSCNAs. The model based on aSCNAs compares the tumor allelic imbalance with the allelic imbalance observed in the normal sample at the same genomic segment. Because aSCNAs may arise independently, we treat each segment as an independent measure of TiN. This enables us to detect multiple TiN values in one normal sample, representing different modes of contamination. Assuming we know the segment's TiN value (θ_s), we can calculate, for each heterozygous SNP in the segment, the expected underlying allele fraction of non-reference reads in the normal sample ($\widehat{f}^{n}$) (see "Derivation of f_v^{n} as a function of R_v^{t} and θ"):

$$C_s^{\mathrm{n}} = \frac{R_s^{\mathrm{t}}}{R_s^{\mathrm{t}}\theta_s + 2(1-\theta_s)}$$
$$\psi(f_v^{\mathrm{t}}) = |\mu^{\mathrm{n}} - f_v^{\mathrm{t}}|$$
$$\widehat{f}_v^{\mathrm{n}}(f_v^{\mathrm{t}}, \theta_s, C_s^{\mathrm{n}}(\theta_s, R_s^{\mathrm{t}})) = \mu^{\mathrm{n}} + \theta_s C_s^{\mathrm{n}} \psi(f_v^{\mathrm{t}})$$

The expected normal allele fraction is equal to the tumor allele imbalance ($\psi(f_v^{\mathrm{t}})$) relative to the midpoint (μ^{n}) multiplied by the TiN and the ratio of total copy ratios, R_s^{t} and R_s^{n} ($R_s^{\mathrm{n}} = R_s^{\mathrm{t}}\theta_s + 2(1-\theta_s)$); see below). The phase of the SNP with respect to its neighbors, d_v^{t}, is based on the tumor data and equals 1 if it is above the midpoint and −1 otherwise. Because the true somatic allele fraction of each SNP is unknown, we integrate over the distribution of possible allele fractions (f) given the observed tumor reads. To calculate the likelihood function for each segment, we calculate the joint likelihood considering all SNPs in each segment.

$$p(\widehat{f}_v^{\mathrm{n}} \,|\, a_v^{\mathrm{t}}, r_v^{\mathrm{t}}, a_v^{\mathrm{n}}, r_v^{\mathrm{n}}, \theta_s, C_s^{\mathrm{n}}) = \int_0^1 p(\widehat{f}_v^{\mathrm{n}}(\theta_s, f, C_s^{\mathrm{n}}) \,|\, a_v^{\mathrm{n}}, r_v^{\mathrm{n}})\, p(f \,|\, a_v^{\mathrm{t}}, r_v^{\mathrm{t}})\, df$$
$$L_s(\theta_s \,|\, \hat{\mathbf{f}}^{\mathbf{n}}, \mathbf{v_s}) = \prod_{v=1}^{N_S} p(\widehat{f}_v^{\mathrm{n}} \,|\, a_v^{\mathrm{t}}, r_v^{\mathrm{t}}, a_v^{\mathrm{n}}, r_v^{\mathrm{n}}, \theta_s, C_s^{\mathrm{n}})$$

We perform k-means clustering on the segment TiN estimates (see "Clustering of aSCNA data") and calculate the posterior distribution of TiN over all clustered segments assigned to cluster K:

$$L(\theta \,|\, \mathbf{S}, \hat{\mathbf{f}}^{\mathbf{n}}, \mathbf{v}) = \prod_{s \in K} L_s(\theta_s \,|\, \hat{\mathbf{f}}^{\mathbf{n}}, \mathbf{v_s})$$

Inference using aSCNAs. We calculate the posterior probability for each value of θ (over a grid $(0, 0.01, 0.02, \ldots, 1)$) and determine $\theta^{\star}_{\mathrm{aSCNA}}$, the maximum a posteriori estimate of θ.

$$\theta^{\star}_{\mathrm{aSCNA}} = \underset{\theta \in [0, 0.01, \ldots, 1]}{\mathbf{argmax}}\; l(\theta \,|\, \mathbf{S}, \mathbf{F^n}, \mathbf{v})$$

Model based on SSNVs. The model based on SSNVs compares the tumor allele fractions of candidate variants with the allele fractions in the normal sample (f_v^{n}). For each candidate SSNV i, we assign a latent Bernoulli indicator variable z_i that indicates whether the SSNV is classified as a somatic mutation. The prior probability of a candidate SSNV being somatic, ϕ, is set on the basis of the expected ratio of somatic to rare inherited germline variants, which varies by tumor type (for example, the somatic mutation frequency in CLL is ~1 mutation per megabase, and the rate of rare germline SNPs is ~10 mutations per megabase; therefore, ϕ is set as 1/11). For most sites with sufficient coverage (depth > 20), the prior has effectively no impact on classification as a somatic mutation.

To calculate the probability of each variant being somatic, we consider the probability of the observed data under three scenarios: (i) the variant is a somatic mutation and thus the normal allele fraction is due to TiN ($z_v = 1, a_v^{\mathrm{tin}} \approx f_v^{\mathrm{t}} C_s^{\mathrm{n}} \theta h_v^{\mathrm{n}}, r_v^{\mathrm{tin}} \approx h_v^{\mathrm{n}} - a_v^{\mathrm{tin}}$); (ii) the variant is a germline polymorphism and the allele fraction is determined as described above (SNP) ($z_v = 0, a_v^{\mathrm{het}} \approx \widehat{f}_v^{n}(\theta, f, C_s^{n}) h_v, r_v^{\mathrm{het}} \approx h_v^{n} - a_v^{\mathrm{het}}$); and (iii) the variant is an artifact and the underlying allele fractions are equal in both samples ($z_v = 0, a_v^{\mathrm{t}}, r_v^{\mathrm{t}}$). A priori we consider candidate variants to be equally likely to be germline variants or artifacts. Conceptually the normal allele fraction is generated in three ways:

$$\widehat{f}_v^{\mathrm{n}} \,|\, \mathrm{Somatic}, z_v = 1 \sim \mathrm{Beta}(a_v^{\mathrm{tin}} + 1, r_v^{\mathrm{tin}} + 1)$$
$$\widehat{f}_v^{\mathrm{n}} \,|\, \mathrm{SNP} \sim \mathrm{Beta}(a_v^{\mathrm{het}} + 1, r_v^{\mathrm{het}} + 1)$$
$$\widehat{f}_v^{\mathrm{n}} \,|\, \mathrm{Artifact} \sim \mathrm{Beta}(a_v^{\mathrm{t}} + 1, r_v^{\mathrm{t}} + 1)$$

We use observed counts for each candidate SSNV to compute the log-likelihood for θ:

$$p(\widehat{f}_v^{\mathrm{n}}|\mathrm{SNP},\theta)=\int_0^1 p(\widehat{f}_v^{\mathrm{n}}(\theta,f,C_s^{\mathrm{n}})|a_v^{\mathrm{n}},r_v^{\mathrm{n}})p(f|a_v^{\mathrm{t}},r_v^{\mathrm{t}})df$$

$$p(\widehat{f}_v^{\mathrm{n}}|\mathrm{Artifact})=\int_0^1 p(f|a_v^{\mathrm{n}},r_v^{\mathrm{n}})p(f|a_v^{\mathrm{t}},r_v^{\mathrm{t}})df$$

$$\begin{aligned}p(\widehat{f}_v^{\mathrm{n}}|z_v=0,\theta)=&(\Pr(\widehat{f}_v^{\mathrm{n}}|\mathrm{SNP},\theta)(1-\Pr(\widehat{f}_v^{\mathrm{n}}|\mathrm{Artifact}))\\&+(\Pr(\widehat{f}_v^{\mathrm{n}}|\mathrm{Artifact})(1-\Pr(\widehat{f}_v^{\mathrm{n}}|\mathrm{SNP},\theta)))\end{aligned}$$

$$p(\widehat{f}_v^{\mathrm{n}}|z_v=1,\theta)=\int_0^1 p(\widehat{f}_v^{\mathrm{n}}(\theta,f,C_s^{\mathrm{n}})|,a_v^{\mathrm{n}},r_v^{\mathrm{n}})p(f|a_v^{\mathrm{t}},r_v^{\mathrm{t}})df$$

$$L(\theta|\hat{\mathbf{f}}^{\mathbf{n}},\mathbf{v})=\prod_{v=1}^{N}p(\widehat{f}_v^{\mathrm{n}}|z_v=1,\theta)^{z_v}p(\widehat{f}_v^{\mathrm{n}}|z_v=0,\theta)^{1-z_v}$$

$$\begin{aligned}l(\theta|\hat{\mathbf{f}}^{\mathbf{n}},\mathbf{v})=&\sum_{v=1}^{N}((z_v)\log(p(\widehat{f}_v^{\mathrm{n}}|z_v=1,\theta))\\&+(1-z_v)\log(p(\widehat{f}_v^{\mathrm{n}}|z_v=0,\theta)))\end{aligned}$$

Inference using SSNVs. To estimate TiN using SSNVs, we use the expectation-maximization algorithm. Briefly, θ is initialized to 0, and expectations (E) of the variant assignments (z_v) are calculated given θ. Then we find $\theta^{\star}_{\mathrm{SSNVs}}$, which maximizes (M) the likelihood function (over a grid ($\theta=0,0.01,0.02,\ldots,1$)). We repeat this procedure until the estimate on θ converges (typically in a few iterations).

$$\mathrm{E-step}:\mathrm{E}_{\theta}[z_v]=\frac{\phi p(\widehat{f}_v^{\mathrm{n}}|\theta,z_v=1)}{(1-\phi)p(\widehat{f}_v^{\mathrm{n}}|\theta,z_v=0)+\phi p(\widehat{f}_v^{\mathrm{n}}|\theta,z_v=1)}$$

$$\mathrm{M-step}:\theta^{\star}_{\mathrm{SSNVs}}=\underset{\theta\in[0,0.01,\ldots,1]}{\mathbf{argmax}}\ [l(\theta|\mathbf{v},\hat{\mathbf{f}}^{\mathbf{n}},\mathrm{E}_{\theta}[\mathbf{z}])]$$

Inference using the joint likelihood function. The likelihood functions for SSNVs and aSCNAs are nearly independent because they are generated by distinct underlying processes and use different measurements. Therefore, when both data types are available, deTiN calculates the joint TiN estimate ($\theta^{\star}$) and posterior distribution by summing and normalizing the log-likelihood functions for SSNVs and aSCNAs. Next we compare the model $\theta=0$ to $\theta=\theta^{\star}$:

$$\theta^{\star}=\underset{\theta\in[0,0.01,\ldots,1]}{\mathbf{argmax}}\ [l(\theta|\mathbf{S},\hat{\mathbf{f}}^{\mathbf{n}},\mathbf{v})+l(\theta|\mathbf{v},\hat{\mathbf{f}}^{\mathbf{n}},\mathrm{E}[\mathbf{z}])]$$

$$p(\theta=\theta^{\star})=\frac{\pi p(\theta=\theta^{\star})}{\pi p(\theta=\theta^{\star})+(1-\pi)p(\theta=0)}$$

As a final step, if the model $\theta=\theta^{\star}$ is chosen, we recalculate $\mathrm{E}[z_v]$ given $\theta^{\star}$ and classify as somatic candidate variants for which $\mathrm{E}[z_v]>\kappa$ (we use $\kappa=0.5$). Finally, to remove variants that do not fit any of our models, we reject candidate somatic variants in cases where the predicted normal allele fraction is unlikely given the observed normal allele counts.

$$\int_0^{\widehat{f}_v^{\mathrm{n}}}p(f|a_v^{\mathrm{n}},r_v^{\mathrm{n}})df\le 0.01$$

Derivation of f_v^{n} as a function of R_v^{t} and θ. To estimate TiN, we calculate the expected normal allele fraction of each variant given a TiN value, observed tumor allele fractions, and total copy ratio. We define the allele fractions $(f_v^{\mathrm{n}},f_v^{\mathrm{t}})$ and total copy ratios $(R_v^{\mathrm{n}},R_v^{\mathrm{t}})$ as follows, where m is the multiplicity of some variant v, α is the fraction of tumor cells in the tumor sample, β is the fraction of tumor cells in the normal sample, q_v is the local total copy number in the tumor sample, τ is the ploidy of the tumor cells, and 2 is the ploidy of normal cells:

$$f_v^{\mathrm{n}}=\frac{\beta m}{\beta q_v+2(1-\beta)}$$

$$f_v^{\mathrm{t}}=\frac{\alpha m}{\alpha q_v+2(1-\alpha)}$$

$$R_v^{\mathrm{t}}=2\frac{\alpha q_v+2(1-\alpha)}{\alpha\tau+2(1-\alpha)}$$

$$R_v^{\mathrm{n}}=2\frac{\beta q_v+2(1-\beta)}{\beta\tau+2(1-\beta)}$$

We then want to derive a factor Z, which allows us to translate tumor allele fractions f_v^{t} to allele fractions in the normal f_v^{n} given θ:

$$f_v^{\mathrm{n}}=f_v^{\mathrm{t}}Z$$

$$Z=\frac{f_v^{\mathrm{n}}}{f_v^{\mathrm{t}}}=\frac{\frac{\beta m}{\beta q_v+2(1-\beta)}}{\frac{\alpha m}{\alpha q_v+2(1-\alpha)}}=\frac{\beta[\alpha q_v+2(1-\alpha)]}{\alpha[\beta q_v+2(1-\beta)]}$$

$$Z=\frac{\beta}{\alpha}\frac{\alpha q_v+2(1-\alpha)}{\beta q_v+2(1-\beta)}\frac{\alpha\tau+2(1-\alpha)\beta\tau+2(1-\beta)}{\alpha\tau+2(1-\alpha)\beta\tau+2(1-\beta)}$$

$$Z=\frac{\beta[\alpha\tau+2(1-\alpha)]}{\alpha[\beta\tau+2(1-\beta)]}\frac{\beta\tau+2(1-\beta)}{\beta q_v+2(1-\beta)}\frac{\alpha q_v+2(1-\alpha)}{\alpha\tau+2(1-\alpha)}$$

$$Z=\theta\frac{R_v^{\mathrm{t}}}{R_v^{\mathrm{n}}}$$

We can then show that $R_v^{\mathrm{n}}=\theta R_v^{\mathrm{t}}+2(1-\theta)$ and thus derive C_s^{n}:

$$\begin{aligned}\theta R_v^{\mathrm{t}}+2(1-\theta)&=2\frac{\beta}{\alpha}\frac{\alpha\tau+2(1-\alpha)}{\beta\tau+2(1-\beta)}\frac{\alpha q_v+2(1-\alpha)}{\alpha\tau+2(1-\alpha)}+2\\&\quad-2\frac{\beta}{\alpha}\frac{\alpha\tau+2(1-\alpha)}{\beta\tau+2(1-\beta)}\\&=2\frac{\beta}{\alpha}\frac{\alpha q_v+2(1-\alpha)}{\beta\tau+2(1-\beta)}+2\frac{\alpha[\beta\tau+2(1-\beta)]}{\alpha[\beta\tau+2(1-\beta)]}-2\frac{\beta}{\alpha}\frac{\alpha\tau+2(1-\alpha)}{\beta\tau+2(1-\beta)}\\&=2\frac{\beta\alpha q_v+2\beta-2\beta\alpha+\alpha\beta\tau+2\alpha-2\beta\alpha-\beta\alpha\tau-2\beta+2\beta\alpha}{\alpha[\beta\tau+2(1-\beta)]}\\&=2\frac{\beta\alpha q_v-2\beta\alpha+2\alpha}{\alpha[\beta\tau+2(1-\beta)]}=2\frac{\beta q_v+2(1-\beta)}{\beta\tau+2(1-\beta)}=R_v^{\mathrm{n}}\end{aligned}$$

$$C_s^{\mathrm{n}}=\frac{R_v^{\mathrm{t}}}{R_v^{\mathrm{n}}}=\frac{R_v^{\mathrm{t}}}{\theta R_v^{\mathrm{t}}+2(1-\theta)}$$

Finally we have the following expression translating a tumor allele fraction to a normal allele fraction given TiN:

$$f_v^{\mathrm{n}}=f_v^{\mathrm{t}}\frac{\theta R_v^{\mathrm{t}}}{\theta R_v^{\mathrm{t}}+2(1-\theta)}$$

Filtering of segments and SNPs. DeTiN uses only large segments (≥200 capture probes) that have at least 20 balanced heterozygous SNPs (ensuring the same number of SNPs with allele fractions below and above 0.5, in the normal sample, by downsampling the more abundant allele). DeTiN ensures an equal number of SNPs above and below 0.5 in the normal sample to remove mapping artifacts. Mapping artifacts are often associated with false positive calls at low allele fractions. Therefore, segments that cover low-mappability regions accumulate reads with errors. These errors tend to be at low allele fractions, and some are miscalled as germline SNPs. Accumulation of these spurious germline SNPs can cause methods that estimate allelic copy numbers to incorrectly infer allelic imbalance at these loci. It is important to account for this accumulation of low-allele-fraction errors because they occur equally in the tumor and normal sample and thus will negatively affect the accuracy of deTiN.

After segment and variant filtering, for each segment s in the tumor data, we calculate the average absolute shift of the allele fractions from balance, ψ_s^{t}, and its population variance, σ_s^2.

$$\psi_s^{\mathrm{t}}=\frac{1}{N_s}\sum_{v=1}^{N_s}\left|\frac{a_v^{\mathrm{t}}}{a_v^{\mathrm{t}}+r_v^{\mathrm{t}}}-\mu^{\mathrm{n}}\right|$$

$$\sigma_s^2=\frac{1}{N_s}\sum_{v=1}^{N_s}\left(\psi_s^{\mathrm{t}}-\left|\mu^{\mathrm{n}}-\frac{a_v^{\mathrm{t}}}{a_v^{\mathrm{t}}+r_v^{\mathrm{t}}}\right|\right)^2$$

DeTiN uses segments with (ψ_s^{t}) greater than T_{aSCNA} (we use 0.1) and absolute allele shift variance less than 0.025 ($\sigma_s^2<0.025$).

Filtering of SSNVs. DeTiN uses candidate SSNVs that are labeled as somatic or are rejected solely on the basis of evidence observed in the normal. With MuTect, SSNVs are considered candidates if and only if the judgment column is "KEEP" or the column indicating reasons for failure contains only "normal_lod" or "alt_allele_in_normal" (or both). Next, we annotate each variant as representing a likely germline SNP or a potential SSNV on the basis of its allele frequency in the ExAC database[22]. Variants with an ExAC population frequency ≥0.01 are considered germline SNPs, and variants with <0.01 allele frequency are considered candidate SSNVs. Variants with <15 reads in either sample or an allele fraction below 15% in the tumor are not used for TiN estimation but are considered for SSNV recovery.

Clustering of aSCNA data. To identify multiple modes of TiN contamination, deTiN performs k-means clustering on the posterior TiN distributions of the aSCNAs. DeTiN considers $K \in \{1, 2, 3\}$ clusters and then performs model selection using the Bayesian information criterion (BIC). When N is the total number of segments, N_k is the number of segments assigned to cluster k, N_s is the number of variants (v) in segment s, θ_v refers to the maximum a posteriori TiN estimate for a SNP, μ_k is the cluster center, and RSS_k is the residual sum of squares for k clusters. We determine the BIC score for each number of clusters:

$$\text{RSS}_k = \sum_{k=1}^{K} \sum_{s \in k}^{N_k} \sum_{v=1}^{N_s} (\mu_k - \theta_v)^2$$

$$\text{BIC}_k = N \log\left(\frac{\text{RSS}_k}{N}\right) + k \log(N)$$

We disregard values of k for which the minimum distance between clusters is less than $2\sigma_k$, where σ_k represents the within-cluster s.d. for solution k. We then select the number of clusters (K^*) with the minimal BIC, and ensure that $\text{BIC}_{k^*-1} - \text{BIC}_{k^*} > 10$.

Role of tumor-derived phasing in deTiN. Phasing information derived from the tumor sample is important because it reduces the uncertainty of the estimated allele shift. Given a segment with an allele shift in the tumor data, one would require two steps to estimate the allele imbalance in the normal: (i) comparison of the evidence for allele shift with the evidence for balance (the null hypothesis); and (ii) estimation of allele shift from the count data. Using the phasing data, we can directly compute the best estimate of the allele shift. Without the phasing data, there is an additional step of accounting for the uncertainty of the phase of each SNP. In this scenario, each SNP has a probability, which depends on its allele counts, of representing the higher allele (allele fraction > 50%) or lower allele (allele fraction < 50%). For example, a SNP with 20 alternative reads and 20 reference reads has equal probability of belonging to each allele, but a SNP with 30 alternative reads and 10 reference reads is more likely to represent the higher allele. In the case of a small allele shift in the normal (most SNPs are close to balance) or in cases of low coverage, there is more uncertainty in the phase of the SNP. The uncertainty in the phasing yields greater uncertainty in the estimate of the allele shift in the normal because for each SNP we need to account for the probability of it being generated by each allele. Users will obtain less accurate results if they ignore the phase information coming from the tumor sample.

Data generation. *In silico simulations.* We selected tumor–normal pairs for in silico simulations from TCGA. We applied the following criteria for sample selection: high coverage (200× in the tumor and 80× in the normal), high purity (ABSOLUTE[21] purity estimate > 95%), somatic mutation frequency > 1 mutation per Mb, and at least one arm-level aSCNA. With these criteria we obtained five tumor–normal peripheral blood sample pairs from three tumor types (bladder cancer, glioblastoma multiforme (×3), and a malignant melanoma; Supplementary Table 1).

To create the simulations, we first downsampled each BAM file with SAMtools[23] to establish uniform coverage (120× in tumors and 60× in normals). Then, we downsampled the normals and tumors in ratios corresponding to the TiN mixtures and mixed each of the resulting BAM files and fixed read groups with Picard tools. For example, to generate a 0.5-TiN simulation, we downsampled a normal to 0.5 (30×) and downsampled the matched tumor to 0.25 (30×), and then mixed them together to generate a 50% TiN mixture (at 60×).

In vitro simulations. To evaluate the performance of deTiN on experimentally derived sequencing data, we mixed tumor and normal cell lines in various ratios. For the tumor sample we selected the cell line CRL-2321D, and for the normal we used CRL-2362D. DNA from these samples was mixed in equal amounts to generate a 0.5-TiN pool with total mass of 500 ng. We then mixed pure tumor and pure normal with this pool to generate the other mixtures. We volume-checked samples on a NanoDrop to ensure that we had achieved the desired mixtures.

We then performed library preparation. Briefly, double-stranded DNA was quantified by PicoGreen fluorescence assay using the provided DNA standards; 100 ng of DNA was fragmented to obtain 150-bp pieces by sonication with a Covaris E210 instrument. Solid-phase reversible immobilization purification and library construction were done with AMPure XP beads, KAPA library preparation and KAPA library amplification kits. Library preparation was performed in 96-well plates on an Agilent Bravo liquid handler.

Finally we carried out hybrid selection, capture, and sequencing. DNA was processed through two hybridization events with the Illumina Content Exome Rapid Capture kit. Samples were normalized to 2 ng/µL and pooled. Quantitative PCR was then performed on the pool to normalize it to 2 nM before 0.1 M NaOH was added to denature the samples. Samples were sequenced on Illumina HiSeq 2500 machines in Rapid Run mode using 76-bp paired-end reads. The BAM files generated by these experiments are publicly available from Google Cloud (bucket ID fc-070aec01-a599-4fe3-9ed0-2f39288f912e), FireCloud (https://portal.firecloud.org/#workspaces/broad-firecloud-testing/deTiN_release_data), and the Sequencing Read Archive (PRJNA422575).

Alignment/assembly and quality control. Exome sequence processing was carried out via established analytical pipelines at the Broad Institute. A BAM file was produced with the Picard pipeline (http://picard.sourceforge.net/), which aligns the tumor and normal sequences to the hg19 human genome build using Illumina sequencing reads. The BAM file was uploaded into the Firehose pipeline (http://www.broadinstitute.org/cancer/cga/Firehose), which manages input and output files to be executed.

Quality control modules for assessment of genotype concordance and cross-contamination using ContEst[24] were applied within Firehose.

Mutation calling and copy-number analysis. MuTect[1], Strelka[2], and Varscan2[3] were applied to identify SSNVs. Strelka[2] was applied to identify small indels. Variants were filtered by a panel of normal samples to remove sequencing variants as previously described[9]. Annotation of identified variants was done with Oncotator[25].

We inferred copy ratios and germline SNPs by using GATK's CNV analysis suite (https://github.com/broadinstitute/gatk). Briefly, read depth at capture probes in tumor samples was normalized via tangent normalization against a panel of normal samples. The resulting normalized coverage ratios were then segmented with the circular binary segmentation algorithm. These data were then transformed into allelic copy-number data via integration of data from informative inherited SNPs. MuTect's "call-stats" raw variant file, allelic copy-number data, and inherited SNPs are the required inputs to deTiN (Supplementary Note).

Statistics and data analysis. For in silico simulation, data points in Fig. 1a,c and Supplementary Fig. 3a show the weighted mean TiN estimate from five independent experiments ($n = 5$ for each TiN level). Error bars in these figures show the s.e. on the weighted mean. For in vitro simulation, panels in Fig. 1b,d and Supplementary Figs. 2a–c and 3b show results from a single experiment ($n = 1$ for each TiN level). Error bars show the 95% confidence interval on the TiN estimate in Fig. 1b and show the 95% confidence interval on the sensitivity calculated using the beta distribution (MATLAB function "betapdf") in Fig. 1d and Supplementary Figs. 2a–c and 3b. TiN estimates and sensitivities are reported in Supplementary Tables 1 and 2. ROC curves and AUCs in Supplementary Fig. 3e,f were calculated using the in vitro sequencing experiment and the Python package scikit-learn function "roc_auc_score." Error bars in Supplementary Fig. 3f show the 95% confidence interval generated via bootstrapping ($n = 100$ iterations). Error bars in Supplementary Fig. 4a,b,d are based on 100 iterations of downsampling. Error bars in Supplementary Fig. 4c,e indicate the 95% confidence interval on the TiN estimate calculated using the in vitro sequencing mixture.

Comparisons of TiN estimates and mutation rates shown in Fig. 2a,b were done by two-tailed Mann–Whitney test (MATLAB function "ranksum"). For each panel, $n = 257$. Error bars shown in Supplementary Fig. 6b (red) and Fig. 3c show 1 s.d. on the allele fraction calculated using the beta distribution. Estimates and mutations are reported in Supplementary Table 3. Error bars in Fig. 3b show the s.e. on mean sensitivities (for TiN = 0, $n = 230$; TiN = 0.01, $n = 9$; TiN = 0.03, $n = 9$; TiN = 0.07, $n = 4$; otherwise no error bar is shown). Normal blood samples were used to generate 'truth set' variants. Calls with coverage lower than 10× in tumor or normal samples and allele fractions lower than 10% in the tumor were excluded from this analysis.

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

Code availability. DeTiN is available for use at https://www.broadinstitute.org/cancer/cga/deTiN, and source code is maintained at https://github.com/broadinstitute/deTiN (the current version is also available as part of the Supplementary Software). Furthermore, deTiN is accessible via the Broad Institute's genomics analysis platform, firecloud (module: broadinstitute_cga/detin_v1.0). Data in this paper were generated with a MATLAB implementation of deTiN (https://hub.docker.com/r/broadinstitute/detin_matlab), which is available from the corresponding author upon request but is no longer being supported.

Data availability. The in vitro validation sequencing data are available in the Sequencing Read Archive (PRJNA422575).

References

21. Carter, S. L. et al. *Nat. Biotechnol.* **30**, 413–421 (2012).
22. Lek, M. et al. *Nature* **536**, 285–291 (2016).
23. Li, H. et al. *Bioinformatics* **25**, 2078–2079 (2009).
24. Cibulskis, K. et al. *Bioinformatics* **27**, 2601–2602 (2011).
25. Ramos, A. H. et al. *Hum. Mutat.* **36**, E2423–E2429 (2015).

nature research

Corresponding author(s): Gad Getz

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ □ A description of all covariates tested

☒ □ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

□ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ □ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

□ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Samtools was used to generate in-silico tumor in normal simulations.
Data analysis	Data analysis was performed in Python-2.7 and MATLAB-2012b. DeTiN was originally developed in MATLAB and then redeveloped in Python 2.7. DeTiN is now supported in Python 2.7 and is available inGithub: https://github.com/broadinstitute/deTiN/.

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

Data

Policy information about availability of data

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

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

DeTiN is available for use https://www.broadinstitute.org/cancer/cga/deTiN and source code is available at https://github.com/broadinstitute/deTiN. Furthermore deTiN is accessible using the Broad Institute’s genomics analysis platform firecloud. Module: broadinstitute_cga/detin v1.0. Data in this paper was generated using a

MATLAB implementation of deTiN which is available upon request but no longer being supported. Additionally, the in-vitro validation sequencing data is available on the Sequencing Read Archive (PRJNA422575)

Field-specific reporting

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

☒ Life sciences   □ Behavioural & social sciences   □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size

No sample size selection was performed. We used all samples from Human patient cohorts which were available.

Data exclusions

No data were excluded from analysis.

Replication

We performed 5 independent in-silico simulation experiments results from each of these experiments is reported in the manuscript.

Randomization

This is not relevant to our study.

Blinding

The expert pathologist was blinded to deTiN estimate during review of deTiN predicted contaminated normal tissues. An equal number of predicted contaminated and uncontaminated samples were provided.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

Cell lines were purchased from ATCC: CRL-2321D and CRL-2321D

Authentication

Cell lines were not authenticated

Mycoplasma contamination

Cell lines were not tested for mycoplasma contamination

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.