Supplementary Results

Assessment of deTiN performance (ROC and AUC)

To assess the performance of deTiN for somatic mutations at various TiN levels, we generated receiver operating characteristic (ROC) curves using the *in-vitro* simulated data. We used MuTect[1] somatic mutation calls with an uncontaminated normal (i.e. TiN=0) as a ground truth set. We then calculated the true positive and false positive rates for a range of deTiN parameters as well as the number of false positives when run as default, represented as fractions among the candidate variants that deTiN attempts to rescue (Supplementary Figure 3e; Supplemental Note). For TiN > 0.5, DeTiN had an AUC > 0.88 (Supplementary Figure 3f).

deTiN with low mutation rate tumors and gene panel data

The ability of deTiN to estimate the level of tumor-in-normal contamination depends on the number of candidate somatic variants, the number and size of aSCNAs, the tumor purity, and the depth of sequencing. First, to assess how deTiN mutation-based model performs with fewer true somatic mutations, we performed a down-sampling experiment using the in vitro simulations. We ran deTiN a 100 times, using a contaminated normal with TiN=0.1, while keeping a subset of the 241 somatic mutations detected when run using the uncontaminated normal. We tested different sizes of subsets, ranging from 4 to 150 true somatic mutations. In each of the 100 iterations and number of sites, we recorded the TiN estimate and the 95% confidence interval. We found that, on average, deTiN provides estimates of TiN consistent with the true values across the range of number of mutations. The confidence intervals, however, were larger when using fewer somatic sites (S. Figure 4a,b). This analysis can be used to estimate the performance of deTiN when sequencing only a panel of genes. For example, using a targeted panel of 400 of the 20,000 genes, one would expect to find 5 of the 241 mutations (in our simulations) that will result in a TiN estimate of 0.1 ± 0.03 and a confidence interval size of 0.07 ± 0.02. We note that these results serve as a lower bound since they do not take into account the improved accuracy of deTiN due to the typically deeper coverage of targeted assays.

deTiN with low-pass whole genome sequencing and gene panel data

To study the effect of sequencing coverage on the performance of the aSCNA model of deTiN, we down-sampled the sequencing reads in the contaminated normal, decreasing the coverage from 123x down to 1.24x. Here we used TiN=0.015, to demonstrate that even at such low contamination levels we are able to produce consistent TiN estimates based on copy-number events (S. Figure 4c). We then ran deTiN's aSCNA based model and recorded the TiN estimates and their confidence intervals. The TiN estimates (maximum likelihood) were between 0.01 and 0.03 (deTiN's estimates are given at 0.01 increments) for all down-sampled coverage values (the 95% CIs did include 0 in all cases). As expected, when we reduced the coverage, the size of the 95% CIs increased, due to the increased uncertainty in the allele fractions of the SNPs.

By downsampling coverage, we demonstrated that deTiN's aSCNA estimate is consistent with the true value (TiN = 0.015) even at very low coverages (**Supplementary Figure 4c**). We further note that DeTiN's estimate was not biased by reduced coverage; 4/10 of the point estimates were below 0.015 TiN and 6/10 were above (mean TiN estimate = 0.017). With low-pass whole genome sequencing data, we expect deTiN to perform better since each copy-number segment will have approximately 50-100 times more germline SNPs, due to the increase in covered genome territory, which could mitigate the effects of lower coverage and thus improve deTiN's accuracy and confidence.

To assess how deTiN aSCNA-based model performs with fewer SNPs, we performed a down-sampling experiment using the *in vitro* simulation data. We selected a single copy number segment and ran deTiN 100 times, using a contaminated normal with TiN=0.1. In each run, we selected a subset of the SNPs, ranging from 5 to 200 SNPs, and recorded the resulting TiN estimate and the 95% confidence interval. We found that deTiN provides an unbiased estimate of TiN across the range of number of SNPs. As expected, the confidence intervals were larger when using fewer SNPs consistent with the results from the SSNV-based model (**Supplementary Figure 4d**). The confidence interval size decreased with the number of SNPs following a power law (exponent = $-0.59; \mathrm{CI}_{99}\%[-0.56, -0.63]$) (**Supplementary Figure 4a**, **b**).

deTiN with low purity samples

To investigate possible biases introduced by tumor purity, we used the *in-vitro* simulation data to generate 0.5 TiN mixtures at different tumor purities. For example, in one of the pairs, we considered the sample with 0.8 TiN as the "tumor", and the sample with 0.4 TiN as the contaminated "normal", resulting in a 0.5 ratio of tumor in the normal (0.4/0.8 = 0.5). We performed mutation calling and tumor copy number segmentation with each pair of TiN mixtures. We then used deTiN to estimate TiN using each pair of samples. Since the true TiN value in the *in-vitro* experiments can have small deviations due to inaccuracies in the laboratory mixtures, we also calculated the expected TiN fraction using deTiN's estimates for the tumor and contaminated normal. The resulting deTiN estimates were consistent with these predicted fractions and not biased by purity (**Supplementary Figure 4e**). The confidence intervals on the TiN estimate increased as the tumor purity decreased consistent with less data being available to fit the model.

Comparison with no-normal mutation calling

The normal sample provides important information for detecting somatic events, and, in particular, rare germline events (that are not in standard germline databases which are used by tumor-only methods). Therefore, methods which do not use the normal sequencing data suffer from lower specificity compared to deTiN. To demonstrate this limitation, we ran SomVarIUS[26] on the tumor validation cell line data. SomVarIUS identified 164 (of the 241) true positives mutations (i.e. had 68% sensitivity) and 9,816 false positives (corresponding to 327 false positive mutations per megabase) (Data not shown). DeTiN, on the same data, is more specific for all values of TiN and is also more sensitive at TiN <= 50% (which covers the vast majority of cases) (**Figure 1**). Additionally, no normal approaches which focus on "cancer-hotspots" are unable to do novel variant discovery and are not suitable for general mutation calling.

Performance of deTiN with Strelka and VarScan

Existing mutation detection tools classify variants as somatic by comparing the evidence in the tumor to the normal in various ways. For example, VarScan[3] compares allele count distributions in the tumor and normal samples and thus is slightly more robust to tumor-in-normal contamination (at the expense of lower sensitivity to low AF mutations and reduced specificity) but still loses sensitivity due to contamination at higher TiN levels (S. Figure 2). MuTect or Strelka[1,2] are more specific, but are also more affected by tumor contamination in the normal (S. Figure 2, Methods). By integrating evidence from candidate somatic events and copy-number alterations, deTiN provides a robust estimate of TiN, which can then be used to accurately infer the probability that a variant observed in the contaminated normal was in fact somatic. DeTiN, like any somatic variant caller, relies on the presence of copy number variants and/or candidate somatic nucleotide variants. If the tumor sample has a very low purity, or has neither candidate somatic variants nor copy number events, deTiN will return estimates with much larger confidence intervals due to the lack of data, as seen in the down-sampling analysis in Supplementary Figure 4.

Comparison of CD19- selected normal with saliva normal

DeTiN estimated 31% TiN in the CD19- selected normal, whereas the saliva-derived sample was not found to be contaminated with tumor cells (S. Figure 6). Consistent with these results, deTiN recovered, when using the contaminated case, 58% (42 of 72) of SSNVs detected in the uncontaminated case.

Supplemental Note

Generating input data for deTiN

DeTiN requires a candidate somatic site metrics file, two files listing heterozygous site allele counts in the tumor and normal samples, and a ".seg" file with tumor segmented allelic copy number. We used MuTect version 1.1.6 (HG19) to generate "call_stats" files as the candidate somatic site metrics input using the following parameters:

–cosmic gs://firecloud-tcga-open-access/tutorial/reference/hg19_cosmic_v54_120711.vcf

–dbsnp gs://firecloud-tcga-open-access/tutorial/reference/dbsnp_134_b37.leftAligned.vcf

–fraction_contamination = 0.0001

–downsample_to_coverage = 10000

DeTiN is effective for other callers (**Supplemental Figure 2**), however, MuTect presents several advantages over other callers for use with deTiN. First, MuTect emits every candidate somatic site which meets the tumor log odds threshold. This is a key feature since deTiN is only able to rescue and measure TiN from sites which are available in the input files. Second, MuTect explicitly reports which filters were used to reject a candidate variant. This allows deTiN to consider as candidate somatic mutations sites which were filtered due to evidence in the normal rather than mapping quality or other artifact modes (see **Filtering of SSNVs**).

To generate the heterozygous allele counts and segmented allelic copy data we used GATK4. For a complete guide to running GATK4CNV see these documents: (How to) Call somatic copy number variants using GATK4 CNV, Overview of GetBayesianHetCoverage for heterozygous SNP calling and, Description and examples of the steps in the ACNV case workflow. These modules are available for download and use in Firecloud (upon request). We used the following parameter settings:

GetBayesianHetCoverage
–readDepthThreshold = 15, –minimumMappingQuality = 30, –minimumBaseQuality = 20, –hetCallingStringency = 30, –minimumAbnormalFraction = 0.8, –maximumAbnormalFraction = 0.9, –maximumCopyNumber = 2, –quadratureOrder = 200, –errorAdjustmentFactor = 1

PadTargets
–padding = 250

PerformSegmentation
–alpha = 0.01, –eta = 0.05 , –kmax = 25 , –minWidth = 2 , –nmin = 200 , –nperm = 10000 , –trim = 0.025 , –undoPrune = 0.05 , –undoSD = 3 , –undoSplits = NONE , –pmethod = HYBRID

AllelicCNV
–smallSegmentThreshold = 3, –numSamplesCopyRatio = 100, –numBurnInCopyRatio = 50, –numSamplesAlleleFraction = 200, –numBurnInAlleleFraction = 100, –intervalThresholdCopyRatio = 2, –intervalThresholdAlleleFraction = 2, –maxNumIterationsSimSeg = 25, –maxNumIterationsSNPSeg = 25, –useAllCopyRatioSegments = false

Overview of deTiN command line arguments

We provide an example dataset and commands on Github https://github.com/broadinstitute/deTiN. DeTiN install time is approximately 1 min. For general use the following command line arguments should be set in DeTiN (**Supplementary Figure 10**). For more advanced customization see (**Supplementary Table 6**) or the deTiN github wiki page:

Input files
–mutation_data_path MuTect call stats file (or similar variants file)
–cn_data_path GATK4 AllelicCNV seg file
–tumor_het_data GATK4 tumor het cov file
–normal_het_data GATK4 normal het cov file
–exac_data_path pickle file of MAF >0.01 ExAC sites
–indel_data_path Strelka or MuTect2 unfiltered indel vcf
–indel_data_type Either ”Strelka” or ”MuTect2”
Parameters
–output_name sample name
–output_dir output directory
–mutation_prior fraction of rare germline sites to somatic sites. Default = 0.08
–TiN_prior fraction of samples which user expects to be contaminated. This is used for model selection, set to 0.5 if unknown. Default = 0.5

Additionally we provide an example jupyter/ipython notebook for running deTiN with example data. Example data includes ExAC[22] pickle file for filtering, and MuTect/GATK4 inputs generated using the 0.1 TiN *in-vitro* simulated normal and MuTect 2 indels generated on the same data. The notebook, example data, and example outputs are hosted on github:
https://github.com/broadinstitute/deTiN/blob/master/deTiN_example_0.1_TiN_sim.ipynb. Demo run time is approximately 1 min.