# Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples

Published in final edited form as:
*Nat Biotechnol*. 2013 March ; 31(3): . doi:10.1038/nbt.2514.

# Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples

**Kristian Cibulskis**[1], **Michael S. Lawrence**[1], **Scott L. Carter**[1,2], **Andrey Sivachenko**[1], **David Jaffe**[1], **Carrie Sougnez**[1], **Stacey Gabriel**[1], **Matthew Meyerson**[1,3], **Eric S. Lander**[1,4,5], and **Gad Getz**[1,6]

Gad Getz: gadgetz@broadinstitute.org

[1]The Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA

[2]Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts, USA

[3]Divisions of Medical Oncology and Cancer Biology and Center for Cancer Genome Discovery, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[4]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA

[5]MIT Department of Biology, 31 Ames Street, Cambridge, Massachusetts, USA

[6]Massachusetts General Hospital Cancer Center and Department of Pathology

## Abstract

Detection of somatic point substitutions is a key step in characterizing the cancer genome. Mutations in cancer are rare (0.1–100/Mb) and often occur only in a subset of the sequenced cells, either due to contamination by normal cells or due to tumor heterogeneity. Consequently, mutation calling methods need to be both specific, avoiding false positives, and sensitive to detect clonal and sub-clonal mutations. The decreased sensitivity of existing methods for low allelic fraction mutations highlights the pressing need for improved and systematically evaluated mutation detection methods. Here we present MuTect, a method based on a Bayesian classifier designed to detect somatic mutations with very low allele-fractions, requiring only a few supporting reads, followed by a set of carefully tuned filters that ensure high specificity. We also describe novel benchmarking approaches, which use real sequencing data to evaluate the sensitivity and specificity as a function of sequencing depth, base quality and allelic fraction. Compared with other methods, MuTect has higher sensitivity with similar specificity, especially for mutations with allelic fractions as low as 0.1 and below, making MuTect particularly useful for studying cancer subclones and their evolution in standard exome and genome sequencing data.

## INTRODUCTION

Somatic single-nucleotide substitutions are an important and common mechanism for altering gene function in cancer. Yet, they are difficult to identify. First, they occur at a very low frequency in the genome, ranging from 0.1 to 100 mutations per megabase, depending on tumor type[1–7]. Second, the alterations may be present only in a small fraction of the DNA molecules originating from the specific genomic locus for reasons including: contaminating

**Author Contributions**

D.J. posed the concept of using a statistical method and filters to detect somatic mutations. G.G. and K.C. conceived and designed MuTect and the analysis. K.C implemented the algorithm and performed the analysis. M.S.L. conceived of and initially developed the panel of normals filter. G.G. and S.C. developed the power calculations and investigated sub-clonal events detected with MuTect. C.S. and A.S. assisted in the generation and interpretation of validation data. D.J, C.S. and M.M. provided critical review of the manuscript. K.C., G.G. and E.S.L. wrote the manuscript. G.G, M.M., S.G. and E.S.L provided leadership for the project.

normal cells in the analyzed sample; local copy-number variation within the cancer genome; and presence of a mutation within only a sub-population of the tumor cells[8–11] ('subclonality'). The fraction of DNA molecules harboring an alteration ('allelic fraction') has been reported to be as low as 0.05 for highly impure tumors[8]. The study of the subclonal structure of tumors is not only critical to understanding tumor evolution both in disease progression and response to treatment[12], but also for developing reliable clinical diagnostic tools for personalized cancer therapy[13].

Recent reports on subclonal events in cancer used non-standard experiments; they have either inferred subclonal status by looking for shared clonal events among several metastases from the same patient[14], resorted to ultra-deep sequencing[11] or sequenced very small numbers of single cells[15–17]. In contrast, tens of thousands of tumors are being sequenced at standard depths of 100–150x for exomes and 30–60x for whole genome as part of large scale cancer genome projects, such as The Cancer Genome Atlas (TCGA)[1,2,7] and the International Cancer Genome Consortium (ICGC)[18]. In order to detect clonal and sub-clonal mutations present in these samples there is a need for a highly sensitive and specific mutation calling method. Although specificity can be controlled through subsequent experimental validation, this is an expensive and time-consuming step that is impractical for general application.

The sensitivity and specificity of any somatic-mutation calling method varies along the genome. They depend on several factors, including the following: depth of sequence coverage in the tumor and a patient-matched normal sample; the local sequencing error rate; the allelic fraction of the mutation; and the evidence thresholds used to declare a mutation. Understanding how sensitivity and specificity depend on these factors is necessary for designing experiments with adequate power to detect mutations at a given allelic fraction, as well as for inferring the mutation frequency along the genome, which is a key parameter for understanding mutational processes and significance analysis[19,20].

To meet these critical needs of high sensitivity and specificity, which are not adequately addressed by the available methods in the field[21–23], we have developed a somatic point mutation caller, MuTect. During its development MuTect was used in numerous studies[1–4,7,19,24–35]. Here we describe the final and publicly available version of MuTect including the rationale behind its different components. We also estimate its performance as a function of the aforementioned factors using benchmarking approaches that, to our knowledge, have not been described before; through independent experimental validation in previous studies[3,4,7,19,24–30]; as well as by applying our method to datasets analyzed in other publications[21,36,37]. We demonstrate that our method is several times more sensitive than other methods for events at low allelic fractions while remaining highly specific, allowing for deeper exploration of the mutational landscape of highly impure tumor samples and the subclonal evolution of tumors.

MuTect is freely available for non-commercial use at http://www.broadinstitute.org/cancer/cga/mutect

# RESULTS

## Benchmarks for assessing mutation callers

Many mutation detection methods have been developed, but there are few systematic approaches for benchmarking their performance on real sequencing data. Previous publications described simulation methods ranging from fully synthetic models[21] to ones that better capture real sequencing errors[11]. However, none of these methods model the full diversity of non-random sequencing errors of both the reference and alternate alleles at the genomic site. To better evaluate the performance of mutation detection methods, we have used two benchmarking approaches, down-sampling and virtual tumors.

Down-sampling uses subsets of reads from primary sequencing data of validated somatic mutations to measure the sensitivity with which a mutation caller identifies the known mutations. Subsets are generated by randomly excluding reads from the experimentally-derived data set until a desired depth of coverage is reached. Notably, down-sampling preserves the expected allelic fraction of the original mutation because reads are removed regardless whether or not they contain the mutant allele. The down-sampling approach is limited in four respects: (i) the number of validated events is typically small, resulting in larger error bars for the sensitivity estimate; (ii) because allele fractions are preserved, only previously validated allele fractions can be explored; (iii) the analysis excludes any mutations that were not originally detected and hence may overestimate the true sensitivity; and (iv) specificity cannot be measured.

To address the issues with down-sampling, we developed a benchmarking procedure that involves creating ‘virtual tumors’ in which we know all true mutations with certainty (**Online Methods**, Supplementary Fig. 1). To measure specificity, we created virtual tumors and normals, at controlled depths, from sequencing data generated by two different sequencing experiments of the same normal sample (designated A). All mutations identified are necessarily false positives. To measure sensitivity, we simulated somatic mutations at controlled allele fractions by replacing selected reads in the virtual tumor with reads from a second sample (designated B) at loci where sample A is reference and sample B harbors a high confidence germline heterozygous event. We then assess the ability of an algorithm to detect these simulated somatic mutations. In this manner, we can measure sensitivity using real sequencing data at a desired depth of coverage and allelic fraction.

The two benchmarking approaches are complementary. Down-sampling uses real somatic mutations but is limited in the parameter regimes it can explore, and it cannot measure specificity directly. In contrast, the virtual tumor approach does not have these limitations. However, it simulates somatic mutations using germline events, which differ from somatic mutations in their nucleotide substitution frequencies and context. As recalibrated base qualities vary for the different bases (owing to biases in machine errors), there is variable sensitivity to detect different substitutions (Supplementary Fig. 2). Because the difference in sensitivity is minimal, we have chosen to use all the germline events. However, it is possible with the virtual tumor approach to simulate the mutation spectrum of a specific tumor type by reweighting the germline events to match the expected mutation spectrum of the tumor.

## Somatic mutation detection with MuTect

MuTect takes as input sequence data from matched tumor and normal DNA after alignment of the reads to a reference genome and standard preprocessing steps[38–40], which include marking of duplicate reads, recalibration of base quality scores and local realignment. The method operates on each genomic locus independently and consists of four key steps (Fig. 1): (i) Removal of low-quality sequence data (Supplementary Methods); (ii) variant detection in the tumor using a Bayesian classifier; (iii) filtering to remove false positives resulting from correlated sequencing artifacts that are not captured by the error model; and (iv) designation of the variants as somatic or germline by a second Bayesian classifier.

**Variant detection—**Variants in the tumor are identified by analyzing the data at each site under two alternative models: (i) a reference model, $M_0$, which assumes there is no variant at the site and any observed non-reference bases are due to random sequencing errors; and (ii) a variant model, $M_f^m$, which assumes the site contains a true variant allele $m$ at allele fraction *f* in addition to sequencing errors. The allele fraction *f* is unknown and is estimated as the fraction of tumor reads that support *m*. This explicit modeling of *f* instead of assuming a heterozygous, diploid event makes our method more sensitive than other methods[21,22]. We declare *m* to be a candidate variant if the log-likelihood ratio of the data under the variant and reference models (that is, the LOD score (log odds)) exceeds a predefined decision threshold that depends on the expected mutation frequency and the desired false positive rate (**Online Methods**). The choice of decision threshold can be used to control the tradeoff between specificity and sensitivity, as described by a Receiver Operating Characteristic (ROC) curve (Fig. 2a, dashed line). We use a fixed threshold of 6.3 for all results presented unless indicated otherwise. This threshold corresponds to a $10^{6.3}$:1 odds ratio in favor of the reference model, which is reasonable because the frequency of mutations in many tumors is only 1–10 per Mb and thus the *a priori* odds of a site harboring a mutation may be as low as 1:$10^5$ or 1:$10^6$.

The LOD score is useful as a threshold for detection, as observed in the concordance of predicted sensitivity and measured sensitivity from the virtual tumor approach (Fig. 2a, solid grey vs. dashed line; Fig. 2b, solid lines vs. circles). Nonetheless, the LOD score cannot be immediately translated into the probability that a variant is due to true mutation rather than to sequencing error because the LOD score is calculated under an assumption of independent sequencing errors and accurate read placement. As we discuss below, these assumptions are incorrect and as a result, although direct application of the LOD score accurately estimates the sensitivity to detect a mutation, it substantially underestimates the false positive rate.

**Variant filtering—**To eliminate these additional false positives due to inaccurate read placement and non-independent sequencing errors, we developed six filters (Fig. 1, Table 1). In addition, we use a panel of normal samples as controls to further eliminate both germline events and artifacts (**Online Methods**). Subsets of these filters define several versions of the method (Fig. 1): (i) Standard (STD), which applies no filters and thus includes all detected variants; (ii) High Confidence (HC), which applies the six filters and (iii) High Confidence + Panel of Normals (HC+PON), which additionally applies the Panel of Normals filter.

We tested the utility of these filters by applying them to the virtual tumors benchmark and re-comparing the results with the calculations (Fig. 2a). The sensitivity estimated for both with (HC) and without (STD) filters is similar, indicating that the model is accurate with respect to detection and that the filters do not adversely impact sensitivity. However, after applying the filters (HC), the specificity increases and closely follows the calculations, suggesting that the filters have largely eliminated the systematic false positives (Fig. 2a, Supplementary Fig. 3).

**Variant classification—**Finally, each variant detected in the tumor is designated as somatic (not present in the matched normal), germline (present in the matched normal) or variant (present in the tumor, but indeterminate status in the matched normal due to insufficient data). To perform this classification, we use a LOD score that compares the likelihood of the data under models in which the variant is present as a heterozygote or absent in the matched normal **(Online Methods)**. We declare that there are insufficient data for classification if the power to make a germline classification is less than 95%. We also make use of public germline variation databases[41] as a prior probability of an event being germline.

## Sensitivity

We applied several benchmarking methods to evaluate the sensitivity of our method to detect mutations as a function of sequencing depth and allelic fraction (Fig. 2b). First, we calculated the sensitivity under a model of independent sequencing errors and accurate read placement using our statistical test given an allelic fraction, tumor sequencing depth and assuming all bases have a fixed base quality score of Q35 (approximate mean base quality score in simulation data; **Online Methods**; Supplementary Fig. 4).

Next, to apply the down-sampling benchmark, we used 3,753 validated somatic mutations, stratified by allele fraction (median=0.28, range=0.07–0.94), in colorectal cancer[7] with deep-coverage (≥100x) exome-capture sequencing downloaded from dbGAP (phs000178). Finally, to apply the virtual tumor benchmark, we used deep-coverage data from two high coverage whole-genome samples (NA12878 and NA12981) sequenced on Illumina HiSeq instruments by the 1000 Genomes Project[42] and another previous study[43], across 1 Gb of genomic territory. Note that we cannot use the Panel of Normals filter (HC+PON) in the virtual tumor sensitivity benchmark, because it discards common germline sites.

Sensitivity estimates based on these three approaches were highly consistent with each other (median coefficients of variation for each depth of 3.1%). This suggests that the benchmarking approaches accurately estimate the sensitivity of mutation calling methods, and also that the calculated sensitivity is robust across a large range of parameter values enabling us to confidently extrapolate to higher depths and lower allele fractions (Supplementary Table 1).

Based on this analysis, we observe that MuTect is a highly sensitive detection method. It detects mutations at a site with 30x depth in the tumor (typical of whole genome sequencing) and an allele fraction of 0.2 with 95.6% sensitivity. The sensitivity can be increased to 99.9% by sequencing deeper (to 50x), and drops to 58.9% for detecting mutations with allelic fraction of 0.1 (at 30x) (Fig. 2b, Supplementary Table 1). Furthermore, with 150x depth (typical of exome sequencing) we have 66.4% sensitivity for 3% allele fraction events. It is this sensitivity to detect low-allele fraction events that uniquely positions MuTect to analyze samples with low purity or with complex subclonal structure.

This detailed understanding of the factors determining sensitivity is critical for targeting the appropriate depth of sequencing. Because the allelic fraction of a mutation depends on the tumor purity, local copy-number and clonality[8], one can calculate the sequencing depth required for a desired sensitivity on a tumor-specific basis. Also, given a sequencing data set we can calculate the sensitivity to have detected a mutation with a particular allelic fraction for each base along the genome. This allows us to assert the absence of a mutation (with a specified allele fraction), which is particularly important in a clinical setting.

## Specificity

It is trivial to create an extremely sensitive somatic mutation detection method by identifying any site with a single non-reference read as a candidate mutation. Clearly, such an approach would have an enormous false positive rate. Therefore in evaluating the performance of a mutation detection method, it is critical to thoroughly characterize its specificity. There are two sources of false positives: (i) over-calling events in the tumor and (ii) under-calling true germline events in the matched normal. Over-calling in the tumor is typically due to sequencing errors and inaccurate read placements whereas under-calling of true germline events in the matched normal is often due to low sequencing depth in the normal.

To measure the false positive rate due to tumor over-calling, we used the virtual tumor approach across 1 Gb of NA12878 at various depths in the virtual tumor and at 30x in the virtual normal. All detected events are false positives, but to eliminate those due to under-calling germline events from consideration, we excluded all known germline variant sites. Using no filters (STD) the false positive rate increased with depth (from 6.7/mb at 5x to 20.1/mb at 30x) (Fig. 3a). This is due to the increased power to call mutations with lower allele fractions, which are enriched with false positives (Fig. 3b). The HC filters reduce the false positive rate by an order of magnitude (1.00/mb at 30x). The Panel of Normals (HC +PON) then filters out remaining rare, but recurrent, artifacts (0.51/mb at 30x). Certain filters, such as the Poor Mapping filter, have the biggest effect at low depths whereas other filters are more invariant to depth, such as the Proximal Gap filter (Fig. 3c). The Clustered Position filter rejects the most sites exclusively. However, the majority of false positives are rejected by several filters.

We then studied the errors owing to under-calling of true germline events in the matched normal with the same approach but instead using the ~1 million germline variant loci in the same territory (Fig. 3d–f). In classifying an event as germline or somatic, MuTect uses different prior probabilities at sites of common germline variation versus the rest of the genome, and therefore we report the false positive rates separately for these two scenarios (Fig. 3d) along with the power to have classified such events (Fig. 3e–f). We observe that with ≤ 7 reads in the normal at novel germline sites (Fig. 3e) or with ≤ 18 reads at sites of known germline variation (Fig. 3f), there is insufficient data to classify a variant as being somatic or germline, and hence we keep such sites as 'variant' and never make false positive somatic calls in these cases. Once there is sufficient data to make a classification, the error rate drops rapidly from $2.4\times10^{-3}$ at 8x in the normal to below $0.2\times10^{-3}$ at 12x, which corresponds to less than one misclassified germline in the entire exome (~30mb in the exome × 50 novel germline variants/mb × $0.2\times10^{-3}$ error rate).

Finally, we have used MuTect in several recent studies and found a consistent validation rate of ~95% in coding regions based on multiple orthogonal validation technologies[3,4,7,19,24–30] (Table 2) These studies used earlier versions of MuTect which were less sensitive, however a recent publication[13] using this version was able to detect mutations present at 7% allelic fraction (8 reads out of 102) which were subsequently validated by ultra-deep sequencing (~6,000x). In fact, the validation rate is not the best measure for comparing false positive rates across studies because it depends on the ratio of false positive to true mutations, which varies across tumor types. We therefore also report the false positive rate itself (Table 2). We observe a median false positive rate of 0.16/Mb, which is lower than the rate we report using whole genome data (Fig. 3) but consistent with the rate measured when restricting to coding regions (Supplementary Fig. 5), indicating that coding regions are less prone to sequencing and alignment errors.

## Comparison to other methods

We used the down-sampling and virtual tumor benchmarking approaches to compare MuTect against other commonly used methods: SomaticSniper[21], JointSNVMix[22] and Strelka[23]. Each method was tested in two configurations, standard (STD) and high confidence (HC), with thresholds chosen to produce similar false positive rates across the methods. For SomaticSniper (v1.0.0), we used the published configurations and for JointSNVMix (v0.7.5) we used a detection threshold of *P(Somatic)* ≥ 0.95 for STD and *P(Somatic)* ≥ 0.9998 for HC. For Strelka (v0.4.7) we used the recommended configuration with a quality score ≥ 15 for HC and ≥ 1 for STD.

We evaluated the sensitivity of the methods with regard to allele fraction and tumor sequencing depth using the virtual tumor (Fig. 4a) and downsampling (Supplementary Fig.

6) approaches, and observed a sharp distinction in sensitivity, particularly at lower allele fractions. We analyzed data for 30x sequence coverage. In the standard configurations, all methods show ≥ 99.3% sensitivity for mutations at an allele fraction of 0.4. However, in the HC configurations, MuTect, JointSNVMix and Strelka remain sensitive, 98.8%, 96.6% and 98.5% respectively, whereas SomaticSniper drops to 91.5%. At an allele fraction of 0.1, MuTect HC can detect more than half of the mutations (53.2%), whereas Strelka HC detects only 29.7%, JointSNVMix HC drops to 16.8% and SomaticSniper HC falls to 7.4%. At an even lower allele fraction of 0.05, MuTect HC has 16.0% sensitivity but can be increased to 51.9% with 60x coverage. By contrast, JointSNVMix HC and SomaticSniper HC have a sensitivity of ≤ 2.0%, and the sensitivity does not increase appreciably with tumor sequencing depth. Strelka HC detects just 4.6% of the events at 30x and only increases to 20.8% at 60x. Sensitivity for such low allelic fraction events is critical for characterizing impure tumors or subclonal mutations in heterogeneous tumors, and it appears that MuTect is much more sensitive in this regime.

As a more sensitive method may also be less specific, we also compared the performance of the methods with regard to the two kinds of false positives. We observed a very low false positive rate due to miscalled germline sites for all methods given sufficient depth (≥15x) in the matched normal (Fig. 4b). The false positive rates per megabase owing to miscalled reference sites (Fig. 4c) are comparable above 20x in both the STD configuration (median=10.2, range=0.7–20.1) and the HC configuration (median=1.0, range=0.2–3.1).

We can summarize the tradeoff between sensitivity and specificity for each of the methods using a ROC curve, which depends on the sequencing depths in the tumor and normal and the mutation allele fraction. Figure 4d gives an example using tumor depth of 30x, normal depth of 30x and allele fraction of 0.1, showing that MuTect is a generally more sensitive for a given specificity and also has a much smaller decrease in sensitivity for a similar increase in specificity gained by the HC configuration (dashed lines).

Finally, we have also compared the sensitivity of the methods using previously reported sequencing data and validated mutations in the COLO-829 melanoma cell line[37] (Supplementary Table 2). Although MuTect is slightly more sensitive than the other methods, this dataset represents a pure cell line with easily detectable high allelic fractions events (median=0.55) and thus does not expose differences between methods. By running MuTect and the other mutation callers we were able to find additional mutations not originally reported (Supplementary Tables 3,4), underscoring that comparisons to mutations reported in the literature typically underestimate the sensitivity as the complete ground truth set of somatic mutations is often unknown.

## Discussion

As new somatic mutation callers are developed, the cancer genomics community will greatly benefit from a systematic performance measurement using the approaches described here across the entire parameter space of tumor and normal depths and mutation allele fraction. Our method as well as the tools we developed to benchmark mutation detection methods are available, and we encourage methods developers to report the characteristics of their method using these metrics. The approaches described here can also be extended to other alterations such as indels or rearrangements.

Our data suggest that MuTect has an advantage over other methods in terms of its tradeoff between specificity and sensitivity (Fig. 4). The advantage in sensitivity of MuTect is derived from the variant detection statistical test, which includes an estimation of the allele fraction of the event, and the working point chosen along the ROC curve. SomaticSniper and JointSNVMix use a model based on a clonal mutation in a pure, diploid tumor (and thus assume a fixed 50% allele fraction). This assumption reduces sensitivity for lower allele fraction events. In contrast, Strelka specifically considers allele fraction, and thus in the STD configuration has similar sensitivity to MuTect. However, when running in the recommended HC configurations to control false positives, MuTect has only a minor drop in sensitivity compared with the other methods. This is likely because the filters in MuTect were carefully tuned to reject true false positive calls without sacrificing sensitivity.

We have shown that MuTect is much more sensitive at a given specificity than competing methods, allowing us to more comprehensively characterize the landscape of somatic mutations, particularly those present in a small fraction of cancer cells. Moreover, this can be done with standard sequencing depths enabling analysis of the large datasets that are being generated worldwide. Analysis of subclonal mutations and changes in the fractions of cancer cells which harbor them is a powerful way to study the evolution of subclones as they progress during treatment, metastasis and relapse[11,12,44,45]. In particular, we demonstrated that the presence of subclonal mutations in genes involved in driving chronic lymphocytic leukemia (CLL) is an independent prognostic factor beyond the currently used clinical parameters[13]. In fact, using standard exome sequencing data, we were able to detect mutations present in as low as 10% of cancer cells, representing an expected allele fraction of 0.05 (assuming a heterozygous mutations in a diploid region) even before accounting for stromal contamination, which appear to have an effect on time to therapy[13].

Because other methods are not as sensitive to low allelic fraction events, they may miss important subclonal drivers of progression or resistance. Therefore, the sensitivity of MuTect to detect subclonal mutations with low allele-fractions represents a substantial advance, essential to future discoveries regarding the subclonal architecture of cancer and the translation of those discoveries into clinical diagnostics affecting cancer patient treatment and outcomes.

## Online Methods

### Virtual tumor benchmarking approach

The virtual tumor approach begins with deep-coverage data from a high coverage whole-genome sample (NA12878) sequenced on Illumina HiSeq instruments by the 1000 Genomes Project[42] (2 libraries, “Solexa-18483” and “Solexa-18484”, at 30x each) and Gnerre et al.[43] (1 library, “Solexa-23661”, at 30x). These data are publicly available – details are in Supplementary Table 5.

First, we randomly divide the sequencing data into several partitions. We chose to create 6 partitions from each of the 3 libraries (18 partitions total), therefore creating data partitions with ~5x each. We accomplished this by sorting the BAM by name using SortSam from the Picard (http://picard.sourceforge.net) tools to effectively give the reads random ordering. We then randomly allocate each read to one of the partitions and write it to a partition-specific BAM file.

In order to measure specificity, we can designate certain partitions as the tumor and others as the normal and process them through MuTect (or any other method). Somatic mutations identified in this process are false positives as they are either germline events that are under-called in the normal, or erroneous variants due to sequencing noise over-called in the partitions designated as tumor. We chose to draw reads from libraries Solexa-18483 and Solexa-23661 for the tumor and from library Solexa-18484 for the normal.

In order to measure sensitivity, we turn to additional sequencing data on a second individual (Supplementary Table 5). In this case we chose NA12891 that was also sequenced to 60x as part of the 1000 Genomes Project. Using the published high confidence SNP genotypes for those samples from the 1000 Genomes Project, we identify a set of sites that are heterozygous in NA12891 and homozygous for the reference in NA12878. We then used a second utility, SomaticSpike, which is part of the MuTect software package, to perform a mixing experiment in-silico. At each of the selected sites, this utility attempts to replace a number of reads determined by a binomial distribution using a specified allelic fraction in the NA12878 data with reads from the NA12891 data, therefore simulating a somatic mutation of known location, type and expected allele fraction. If there are not enough reads in NA12891 to replace the desired reads in NA12878 the site is skipped. The output of this process is a virtual tumor BAM with the in-silico variants and a set of locations of those variants. Sensitivity is then estimated by attempting to detect mutations at these sites.

## Variant detection

For each site we denote the reference allele as $r \in \{A, C, G, T\}$ and denote by $b_i$ and $e_i$ the called base of the $i$-th read ($i$=1…$d$) that covers the site and the probability of error of that base call (each base has an associated Phred-like quality score $q_i$ where $e_i=10^{-\frac{q_i}{10}}$). To call a variant in the tumor we try to explain the data using two models: (i) a model, $M_0$, in which there is no variant at the site and all non-reference bases are explained by sequencing noise; and (ii) a model, $M_f^m$, in which a variant allele $m$ truly exists at the site with an allele fraction $f$ and, as in $M_0$, reads are also subject to sequencing noise. Note that $M_0$ is equivalent to $M_f^m$ with $f$=0.

The likelihood of the model $M_f^m$ is given by

$$L\left[M_f^m\right]=P(\{b_i\}|\{e_i\},r,m,f)=\prod_{i=1}^{d}P(b_i|e_i,r,m,f)$$

assuming the sequencing errors are independent across reads. If all substitution errors are equally likely, i.e. occur with probability $e_i$/3, we obtain

$$P(b_i|e_i,r,m,f)=\left\{\begin{array}{cc} f\,e_i/3+(1-f)\,(1-e_i) & if\,b_i=r \\ f(1-e_i)+(1-f)\,e_i/3 & if\,b_i=m \\ e_i/3 & otherwise \end{array}\right\}$$

Variant detection is performed by comparing the likelihood of both models and if their ratio, i.e. the LOD score (log odds) exceeds a decision threshold ($\log_{10}\delta_T$) we declare $m$ as a candidate variant at the site. We calculate

$$LOD_T(m,f)=\log_{10}\left(\frac{L\left[M_f^m\right]\,P(m,f)}{L\left[M_0\right]\,(1-P(m,f))}\right)\geq\log_{10}\delta_T$$

and set $\delta_T$ to 2 to ensure that we are at least twice as confident that the site is variant as compared to noise. We can then also rewrite $LOD_T$ as:

$$LOD_T(m,f)=\log_{10}\left(\frac{L\left[M_f^m\right]}{L\left[M_0\right]}\right)\geq\log_{10}\delta-\log_{10}\left(\frac{P(m,f)}{(1-P(m,f))}\right)=\theta_T$$

To determine $P(m,f)$, we first assume that $P(m)$ and $P(f)$ are statistically independent, and that $P(f)$ is uniformly distributed (i.e. $P(f)$=1) and $P(m)$ is a 1/3 of the expected mutation frequency for the studied tumor type (representing equal prior for all substitutions). In practice, we use a typical mutation frequency of $3\times10^{-6}$ which yields $\theta_T$=6.3.

We find the maximum $LOD_T$ across all three values of $m$ and in order to set the unknown allelic fraction parameter $f$, we could use maximum likelihood estimation, i.e. find $f$ that maximizes $LOD_T$. However, for computational efficiency, we instead estimate

$\text{estimate }\hat{f}_{ML}\text{ as }\hat{f}=\frac{\#\text{of mutant reads}}{\#\text{of total reads}}.$

A common source of false positive mutation calls is contamination of the tumor DNA with DNA from other individuals. Germline SNPs in the contaminating DNA appear as somatic mutations. We have previously demonstrated that such contamination can yield a large number of false positives and developed a tool, ContEst[46] to estimate the contamination level, $f_{\text{cont}}$, in sequencing data. Low-level contamination of DNA is a common phenomenon and even 2% contamination can give rise to 166 false positive calls per megabase and 10/Mb when excluding known SNP sites[46]. To protect against this type of false positives and enable analysis of contaminated samples, we replace the reference model with a variant model $M_{f_{\text{cont}}}^m$. This guarantees that variants are called only when they are highly unlikely to be explained by contamination.

## Variant Filters

**Panel of Normals—**To further reduce false positives and miscalled germline events, we employ a panel of normal samples as a filter. To create this filter we run MuTect on a set of normals as if they were tumors without a matched normal in STD mode. From this data, a VCF file is created for the sites that were identified as variant by MuTect in more than one normal.

This VCF is then supplied to the caller, which rejects these sites. However, if the site was present in the supplied VCF of known mutations (--cosmic) it is retained because these sites could represent known recurrent somatic mutations which have been detected in the panel of normal when the normal are from adjacent tissue or have some contamination tumor DNA.

The more normal samples used to construct this panel, the higher the power will be to detect and remove rare artifacts. Therefore, we typically we use all the normal samples readily available. The results presented here were obtained by using a panel of whole genome sequencing data from blood normals of 125 solid tumor cancer patients. The samples used as part of the virtual tumor approach were not included in this panel.

## Variant Classification

To perform this classification, we use a similar classifier to the one described above. In this case $f$, in $M_f^m$, is conservatively set to 0.5 for a germline heterozygous variant. Thus we have

$$LOD_N=\log_{10}\left(\frac{L\left[M_0\right]\ P(m,f)}{L\left[M_{0.5}^m\right]\ P(germline)}\right)\geq\log_{10}\delta_N$$

which can be rewritten as

$$LOD_N=\log_{10}\left(\frac{L\left[M_0^m\right]\ P(m,f)}{L\left[M_{0.5}^m\right]\ P(germline)}\right)\geq\log_{10}\delta_N-\log_{10}\left(\frac{P(m,f)}{P(germline)}\right)=\theta_N$$

Note that here the terms are inverted since we want to be confident that alteration was not present. For $\delta_N$ we set a threshold of 10, which is higher than the threshold for $\delta_T$ because we want to be more confident in our variant classification as misclassified germline events will quickly appear to be significant in downstream somatic analysis due to their elevated population frequency at recurrent sites as compared to real somatic events.

To calculate P(*germline*) we distinguish two cases: (i) sites which are known to be variant in the population and (ii) all other sites. We use the public dbSNP database[41] to make this distinction.

There are ~30×$10^6$ sites known to be variant in the human population according to dbSNP release 134, which is ~1000 variants/megabase. A given individual typically has ~3×$10^6$ variants in their genome, 95% of which fall on dbSNP sites[41,42]. Therefore we expect ~50 variants/mb not at dbSNP sites, i.e. P(*germline*| non-dbSNP *site*) = 5×$10^{-5}$ and therefore we use $\theta_{N|\text{non-dbSNP site}}$ = 2.2. At dbSNP sites, however, we expect 95% of the ~3×$10^6$ variants to occur in the 30×$10^6$ sites in the dbSNP database, yielding P(*germline*| dbSNP *site*) = 0.095 hence $\theta_{N|\text{dbSNP site}}$ = 5.5.

## Sensitivity Calculation

To calculate the sensitivity to detect a mutation with allelic fraction *f* using *n* reads having a Phred-like quality score *q* (and hence a base error, *e*, of $10^{-\frac{q}{10}}$), we first calculate *k*, the minimum number of reads with the alternate allele that will trigger a variant call using

$$k=\arg\min_x LOD_T(x|n,e)\geq\theta_T$$

The sensitivity is then the probability of observing *k* or more reads given the allelic fraction and depth. The marginal distribution of the number of reads with the alternate allele, either originating from the alternate base or a misread reference base, follows a binomial distribution with a frequency that reflects the true underlying allelic fraction *f* and the probability of error *e* (note that here we take the worst case in which all misread bases convert to the same alternate allele). Therefore we can calculate the probability of having observed *k* or more reads as:

$$\sum_{i=k}^{n} binom(i|n,f(1-e)+(1-f)e)$$

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

This work was supported by grants from the National Human Genome Research Institute (U54HG003067) and the National Cancer Institute (U24CA143845). We would like to thank the Genome Analysis ToolKit (GATK) group, and our beta test users for their valuable feedback.

## References

1. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. [PubMed: 21720365]
2. Cancer Genome Atlas Research Network Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
3. Banerji S, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012; 486:405–409. [PubMed: 22722202]
4. Stransky N, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333:1157–1160. [PubMed: 21798893]
5. Ding L, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008; 455:1069–1075. [PubMed: 18948947]
6. Berger MF, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012; 485:502–506. [PubMed: 22622578]
7. Network TCGA. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337.
8. Carter SL, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012; 30:413–421. [PubMed: 22544022]
9. Walter MJ, et al. Clonal architecture of secondary acute myeloid leukemia. N Engl J Med. 2012; 366:1090–1098. [PubMed: 22417201]
10. So Yeon Park MGHJKFMKP. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. The Journal of Clinical Investigation. 2010; 120:636. [PubMed: 20101094]
11. Nik-Zainal S, et al. The life history of 21 breast cancers. CELL. 2012; 149:994–1007. [PubMed: 22608083]
12. Ding L, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012; 481:506–510. [PubMed: 22237025]
13. Landau DA, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. CELL. 2012 Accepted.
14. Campbell PJ, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010; 467:1109–1113. [PubMed: 20981101]
15. Navin N, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011; 472:90–94. [PubMed: 21399628]
16. Hou Y, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. CELL. 2012; 148:873–885. [PubMed: 22385957]
17. Xu X, et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. CELL. 2012; 148:886–895. [PubMed: 22385958]
18. International Cancer Genome Consortium et al. International network of cancer genome projects. Nature. 2010; 464:993–998. [PubMed: 20393554]
19. Chapman MA, et al. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011; 471:467–472. [PubMed: 21430775]
20. Getz G, et al. Comment on “The consensus coding sequences of human breast and colorectal cancers”. Science. 2007; 317:1500. [PubMed: 17872428]

21. Larson DE, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012; 28:311–317. [PubMed: 22155872]

22. Roth A, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics. 2012; 28:907–913. [PubMed: 22285562]

23. Saunders CT, et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012; 28:1811–1817. [PubMed: 22581179]

24. Barbieri CE, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nature genetics. 2012; 44:685–689. [PubMed: 22610119]

25. Bass AJ, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nature genetics. 2011; 43:964–968. [PubMed: 21892161]

26. Wang L, et al. SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. N Engl J Med. 2011; 365:2497–2506. [PubMed: 22150006]

27. Pugh TJ, et al. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature. 2012; 488:106–110. [PubMed: 22820256]

28. Berger MF, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220. [PubMed: 21307934]

29. Lohr JG, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proc Natl Acad Sci USA. 2012; 109:3879–3884. [PubMed: 22343534]

30. Imielinski M, et al. Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing. CELL. 2012; 150:1107–1120. [PubMed: 22980975]

31. Wang P. Mutations in isocitrate dehydrogenase 1 and 2 occur frequently in intrahepatic cholangiocarcinomas and share hypermethylation targets with glioblastomas. Oncogene. 201210.1038/onc.2012.315

32. Durinck S, et al. Temporal dissection of tumorigenesis in primary cancers. Cancer Discov. 2011; 1:137–143. [PubMed: 21984974]

33. Lee RS, et al. A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. The Journal of Clinical Investigation. 2012; 122:2983–2988. [PubMed: 22797305]

34. Cancer Genome Atlas Research Network et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–525. [PubMed: 22960745]

35. Hodis E, et al. A landscape of driver mutations in melanoma. CELL. 2012; 150:251–263. [PubMed: 22817889]

36. Forbes SA, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011; 39:D945–50. [PubMed: 20952405]

37. Pleasance ED, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature. 2009; 463:191–196. [PubMed: 20016485]

38. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

39. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

40. DePristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics. 2011; 43:491–498. [PubMed: 21478889]

41. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]

42. 1000 Genomes Project Consortium A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

43. Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA. 2011; 108:1513–1518. [PubMed: 21187386]

44. Shah SP, et al. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature. 2009; 461:809–813. [PubMed: 19812674]

45. Yachida S, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature. 2010; 467:1114–1117. [PubMed: 20981102]

46. Cibulskis K, et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics. 2011; 27:2601–2602. [PubMed: 21803805]

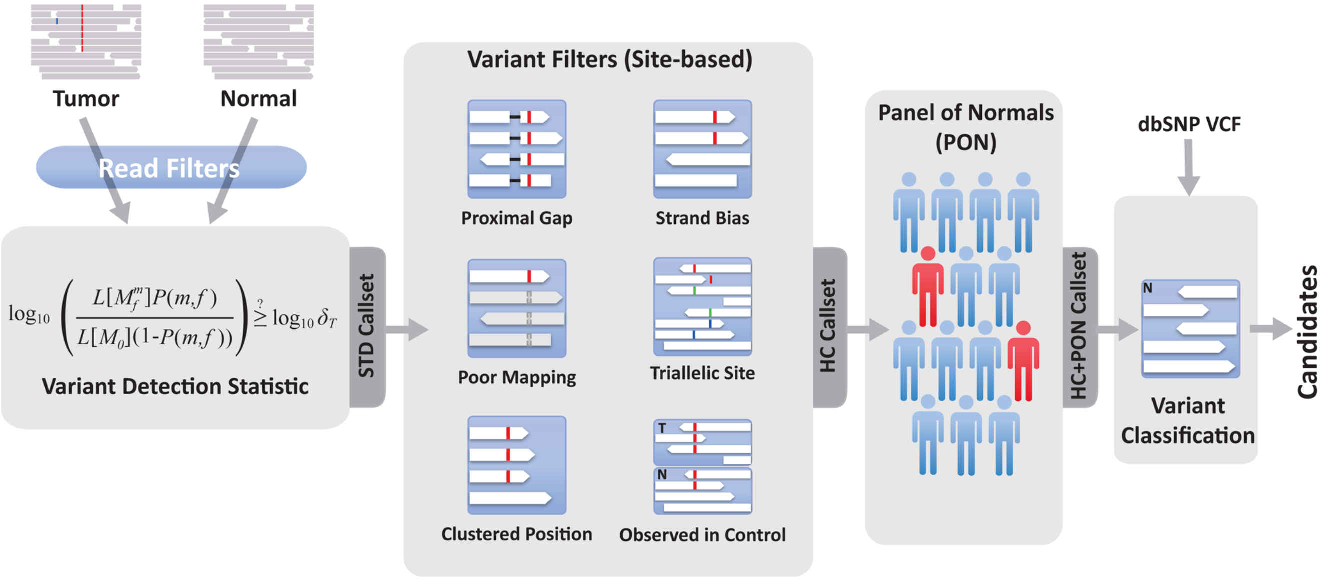

**Figure 1.**

Overview of somatic point mutation detection using MuTect.

MuTect takes as input tumor (T) and normal (N) next generation sequencing data and, after removing low quality reads (Supplementary Methods), determines if there is evidence for a variant beyond the expected random sequencing errors. Candidate variant sites are then passed through six filters to remove artifacts (Table 1). Next, a Panel of Normals is used to screen out remaining false positives caused by rare error modes only detectable in additional samples. Finally, the somatic or germline status of passing variants is determined using the matched normal.

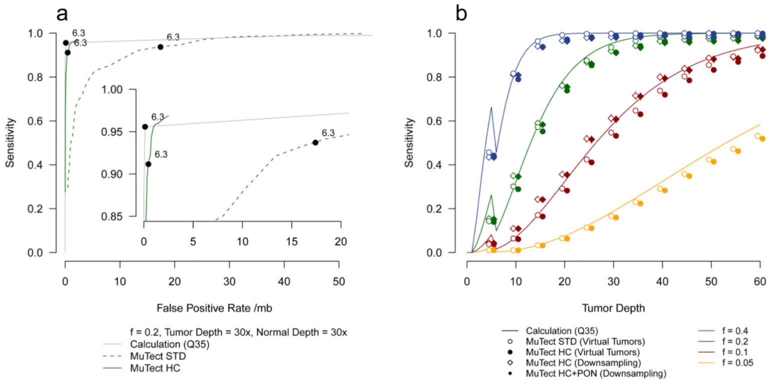

**Figure 2.**

Sensitivity as a function of sequencing depth and allelic fraction.

(**a**) Sensitivity and specificity of MuTect for mutations with an allele fraction of 0.2, tumor depth of 30x and normal depth of 30x using various values of the LOD threshold ($\theta_T$) ($0.1 \leq \theta_T \leq 100$). Results using a model of independent sequencing errors with uniform Q35 base quality scores and accurate read placement (solid grey) are shown as well as results from the virtual tumor approach for the standard (STD, dashed green) and high-confidence (HC, solid green) configurations. A typical setting of $\theta_T = 6.3$ is marked with black circles. (**b**) Sensitivity as a function of tumor sequencing depth and allele fraction (indicated by color) using $\theta_T = 6.3$. The calculated sensitivity using a model of independent sequencing errors and accurate read placement with uniform Q35 base quality scores (solid lines) are shown as well as results from the virtual tumor approach (circles) and the downsampling of validated colorectal mutations[7] (diamonds). Error bars represent 95% CIs.

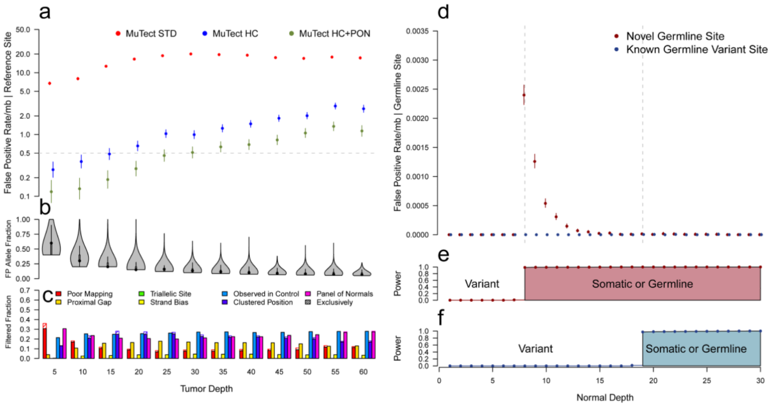

**Figure 3.**

Specificity of variant detection and variant classification using virtual tumor approach. (**a**) Somatic miscall error rate for true reference sites as a function of tumor sequencing depth for the STD (red), HC (blue) and HC+PON (green) configurations of MuTect. Error bars represent 95% CIs. (**b**) Distribution of allele fraction for all miscalls as a function of tumor sequencing depth. (**c**) Fraction of events rejected by each filter; hashed regions indicate events rejected exclusively by each filter. (**d**) Somatic miscall error rate for true germline SNP sites by sequencing depth in the normal when the site is known to be variant in the population (blue) and novel (red). Error bars represent 95% CIs. (**e,f**) Mean power as a function of sequencing depth in the normal to have classified these events as germline or somatic at novel germline sites (e) and known germline variant sites (f).

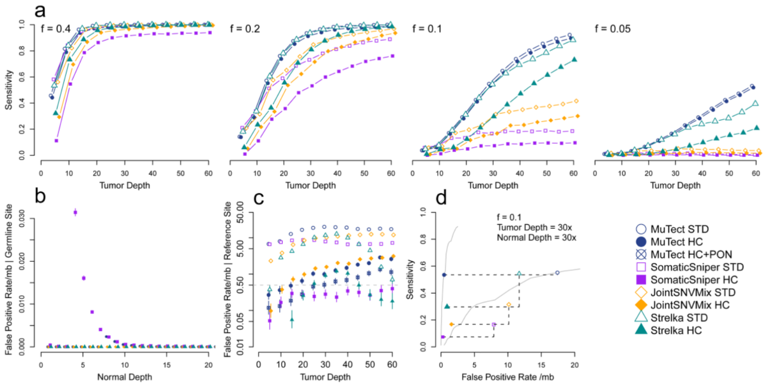

**Figure 4.**

Benchmarking mutation detection methods. (**a**) Comparison of sensitivity as a function of tumor sequencing depth and mutation allele fraction for different mutation detection methods and configurations. (**b**) Comparison of somatic miscall error rate for true germline sites as a function of sequencing depth in the normal. (**c**) Comparison of somatic miscall error rate for true reference sites as a function of tumor sequencing depth. (**d**) Sensitivity as a function of specificity for mutations with an allele fraction of 0.1, tumor depth of 30x and normal depth of 30x between different methods and configurations. Black dotted lines indicate change in sensitivity and specificity between STD and HC configurations for a method. Grey solid lines are the MuTect results of virtual tumor approach from Supplementary Figure 3. (**a–c**) Error bars represent 95% CIs.

## Table 1

Description of variant filters and default thresholds

| Filter Name | Class | Description and Default Thresholds |
|---|---|---|
| Proximal Gap | HC | Remove false positives caused by nearby misaligned small insertion and deletion events. Reject candidate site if there are ≥ 3 reads with insertions within an 11-bp window centered on the candidate mutation, or if there are ≥ 3 reads with deletions within the same 11-bp window |
| Poor Mapping | HC | Remove false positives caused by sequence similarity in the genome, leading to misplacement of reads. Two tests are used to identify such sites: (i) Candidates are rejected if ≥ 50% of the reads in the tumor and normal have a mapping quality of zero (although mapping quality zero reads are discarded in the short-read preprocessing (Supplementary Methods) this filter reconsiders those discarded reads); (ii) Candidates are rejected if they do not have at least a single observation of the mutant allele with a confident mapping (i.e. mapping quality score ≥ 20). |
| Triallelic Site | HC | Reject false positives caused by calling tri-allelic sites where the normal is heterozygous with alleles A/B and MuTect is considering an alternate allele C. Although this is biologically possible, and remains an area for future improvement in mutation detection, calling at these sites generates many false positives and therefore they are currently filtered out by default. However, it may be desirable to review mutations failing only this filter for biological relevance and orthogonal validation and further study the underlying reasons for these false positives. |
| Strand Bias | HC | Reject false positives caused by context specific sequencing errors where the vast majority of the alternate alleles are observed in a single direction of reads. We perform this test by stratifying the reads by direction and then applying the core detection statistic on the two datasets. We also calculate the sensitivity to have passed the threshold given the data (Online Methods). Candidates are rejected when the strand specific LOD is < 2.0 in directions where the sensitivity to have passed that threshold is ≥ 90%. |
| Clustered Position | HC | Reject false positives caused by misalignments hallmarked by the alternate alleles being clustered at a consistent distance from the start or end of the read alignment. We calculate the median and median absolute deviation of the distance from both the start and end of the read and reject sites that have a median ≤ 10 (near the start/end of the alignment) and a median absolute deviation ≤ 3 (clustered) |
| Observed in Control | HC | Eliminate false positives in the tumor by looking at the control data (typically from the matched normal) for evidence of the alternate allele beyond what is expected from random sequencing error. A candidate is rejected if, in the control data, there are (i) ≥2 observations of the alternate allele, or they represent ≥ 3% of the reads; and (ii) their sum of quality scores is > 20. |
| Panel of Normals | HC+PON | Reject artifacts and germline variants by inspecting a panel of normal samples and rejecting candidates that are present in two or more normal samples (**Online Methods**) |

## Table 2

Published validation rates in coding regions

| Tumor type | Mutation rate / Mb (** non-silent) | Validation technology | Validated | Invalidated | Validation rate | False positive rate / Mb |
|---|---|---|---|---|---|---|
| Multiple Myeloma[19] | 2.9 | Sequenom | 87 | 5 | 94.6% | 0.16 |
| Head and Neck [4] | 3.3 | Sequenom | 181 | 8 | 95.8% | 0.14 |
| Breast[3] | 2.9 | Sequenom/PCR/45 | 464 | 27 | 94.5% | 0.16 |
| Prostate[24] | 1.4 | Sequenom | 219 | 10 | 95.6% | 0.06 |
| Colorectal[25] | 5.9 | Sequenom | 292 | 16 | 94.8% | 0.31 |
| CLL[26] | 0.9 | Sequenom | 66 | 5 | 93.0% | 0.06 |
| Medulloblastoma[27] | 0.4** | Fluidigm/PacBio | 19 | 0 | 100.0% | n/a (5 genes) |
| Prostate[28] | 0.9 | Sequenom | 253 | 26 | 90.7% | 0.08 |
| DLBCL[29] | 3.2** | Fluidigm/Illumina | 46 | 1 | 97.9% | n/a (6 genes) |
| TCGA Colorectal[7] | 14 | PCR/454 | 5,713 | 420 | 93.1% | 0.96 |
| Lung Adeno[30] | 12 | Capture/Illumina | 9,458 | 374 | 96.2% | 0.46 |