Virmid: accurate detection of somatic mutations with sample impurity inference

*Genome Biology* 2013, 14:R90  doi:10.1186/gb-2013-14-8-r90

Sangwoo Kim (sak042@cs.ucsd.edu)
Kyowon Jeong (kwj@ucsd.edu)
Kunal Bhutani (kbhutani@eng.ucsd.edu)
Jeong Ho Lee (jhlee4246@kaist.ac.kr)
Anand Patel (adp002@ucsd.edu)
Eric Scott (escott55@gmail.com)
Hojung Nam (hjnam@gist.ac.kr)
Hayan Lee (hlee@cshl.edu)
Joseph G Gleeson (joglesson@ucsd.edu)
Vineet Bafna (vbafna@cs.ucsd.edu)

**ISSN**  1465-6906

**Article type**  Method

**Submission date**  30 April 2013

**Acceptance date**  22 July 2013

**Publication date**  29 August 2013

**Article URL**  http://genomebiology.com/2013/14/8/R90

This peer-reviewed article can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Genome Biology* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Genome Biology* go to

http://genomebiology.com/authors/instructions/

© 2013 Kim et al.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Virmid: accurate detection of somatic mutations with sample impurity inference

Sangwoo Kim\textsuperscript{1,7\ast}, Kyowon Jeong\textsuperscript{2,7}, Kunal Bhutani\textsuperscript{3}, Jeong Ho Lee\textsuperscript{3,6}, Anand Patel\textsuperscript{1}, Eric Scott\textsuperscript{3}, Hojung Nam\textsuperscript{4}, Hayan Lee\textsuperscript{5}, Joseph G Gleeson\textsuperscript{3} and Vineet Bafna\textsuperscript{1\ast}

\textsuperscript{1}Department of Computer Science and Engineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
\textsuperscript{2}Department of Electrical and Computer Engineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
\textsuperscript{3}Institute for Genomic Medicine, Rady Children's Hospital, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
\textsuperscript{4}School of Information and Communications, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju, 500-712, Republic of Korea
\textsuperscript{5}Department of Computer Science, Stony Brook University, 100 Nicolls Road, NY 11794, USA
\textsuperscript{6}Graduate School of Medical Science and Engineering, KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon 305-701, Republic of Korea
\textsuperscript{7}These authors contributed equally to this work

Email: Sangwoo Kim - sak042@cs.ucsd.edu; Kyowon Jeong - kwj@ucsd.edu; Kunal Bhutani - kbhutani@eng.ucsd.edu; Jeong Ho Lee - jhlee4246@kaist.ac.kr; Anand Patel - adp002@cs.ucsd.edu; Eric Scott - escott55@gmail.com; Hojung Nam - hjnam@gist.ac.kr; Hayan Lee - hlee@cshl.edu; Joseph G Gleeson - jogleeson@ucsd.edu; Vineet Bafna - vbafna@cs.ucsd.edu;

\ast Corresponding author

Abstract

Detection of somatic variation using sequence from disease-control matched datasets is a critical first step. In many cases including cancer, however, it is hard to isolate pure disease tissue, and the impurity hinders accurate mutation analysis by disrupting overall allele frequencies. Here, we propose a new method, Virmid, that explicitly determines the level of impurity in the sample, and uses it for improved detection of somatic variation. Extensive tests on simulated and real sequencing data from breast cancer and hemimegalencephaly demonstrate the power of our model. A Software implementation of our method is available at http://sourceforge.net/projects/virmid/

Background

Identifying mutations relevant to a specific phenotype is one of the primary goals in sequence analysis. With the advent of massively parallel sequencing technologies, we can produce an immense amount of genomic information to estimate the landscape of sequence variations. However, the error rate of base-call
and read alignment still remains much higher than the empirical frequencies of single nucleotide variations (SNVs) and de novo mutations [1]. Many statistical methods have been proposed to strengthen mutation discovery in the presence of confounding errors [2–4]. Finding somatic mutations is one particular type of variant calling, which constitutes an essential step of clinical genotyping. Unlike the procedures used for germline mutation discovery, the availability of matched control sample is indispensable. Here, sequence variants that exist in the control sample are used as a basis for measuring individual polymorphisms, while the disease-only mutations are generally regarded as candidate somatic mutations. Traditional approaches call variants from each sample to estimate the sequential differences [5,6]. But most recent studies that calculate joint probabilities of the disease-control genotype pairs showed higher efficiency in separating true somatic mutations from germline mutations by considering correlations between two samples [7–9]. With the aid of probabilistic variant calling models, whole genome/exome sequencing data have been used to identify potential de novo mutations in various studies including schizophrenia [10], autism [11], and cancer [12]. However, there are many cases where mutation discovery might be confounded. One big hurdle is the impurity and heterogeneity of the disease sample. For example, gastric and breast cancer tissues usually contain large amount of stromal cells to make the acquisition of pure cancer sample not feasible [13]. More importantly, there are many cases in which this type of contamination is not only inevitable but dominating the sample constitution. Focal malformation of cortical developments including focal cortical dysplasia and hemimegalencephaly is the most common cause of childhood intractable epilepsy and contains disease cells in affected brain regions with a high proportion of normal cells [14]. Similar problem arises when detecting small amount of target genome mixed in the control samples. In organ transplant, an increased level of circulating cell free DNAs (~10%) of the donor in the recipient’s blood indicates a higher risk of failure [15]. Cell-free DNAs are also found in pregnancy; small amount of fetal DNAs (~13%) are detectable in maternal plasma [16]. In both cases, accurate identification of the target genotypes will provide a basis for a non-invasive and low cost diagnostic method. Conventional methods for somatic mutation profiling are severely compromised in highly contaminated samples because the abundant short reads originated from control genomes obscure true allele frequency (AF) at the site of de novo mutations. This usually results in a failure to call true variants. We have two questions that arise: (a) estimate the contamination level, defined as the proportion of control sample in the mixed disease sample (α, 0 ≤ α ≤ 1) and (b) use α in SNV calling. A natural approach to estimating α has been adapted by many previous studies [15,17–20]. For any heterozygous mutation, the B allele frequency
(BAF) is expected to be close to 50%. A significant and consistent deviation from this value is indicative of the existence and level of control sample inclusion. We found, however, there are two substantial problems in this approach. First, it needs an initial SNV calling procedure either from sequencing or SNP array data, which takes extra time and cost. Second, and more importantly, the initial mutation call is not representative; higher BAF is likely to be observed in the selected sites causing underestimation of $\alpha$. We will show that the bias is significantly large in highly contaminated samples and describe the way to resolve it. Incorporating the estimated $\alpha$ in SNV calling model is another important problem. There are only a few studies that consider $\alpha$ or similar concept in SNV calling [7,9]. More rigorous and explicit use of $\alpha$ by a tight parameterization in a probabilistic model will improve the accuracy of final calls.

Here, we describe a novel probabilistic model Virmid (virtual microdissection for variant calling) which estimates 1) the sample contamination level ($\alpha$), and 2) the disease genotypes including somatic mutations (Figure 1). In the core of Virmid lies a maximum likelihood estimator (MLE) of $\alpha$ and the joint probabilities of control and disease genotypes represented by the joint genotype probability matrix $G$ (see Methods), driven from a local distribution of BAF. It does not require any prior SNV calling nor does it attempt to find variants beforehand; we show that this not only saves computation but also greatly improves the accuracy by reducing sampling bias. Our model also incorporates other sources of noise including sequencing error, mapping error, read mapping bias and mappability [21] of the genomic regions for more accurate modeling, as well as the effect of copy number variations. More importantly, the tight coupling of $\alpha$ and $G$ implemented in a single integrated model enables rigorous recalibration of genotype probabilities with the given $\alpha$. We demonstrate on simulated and real exome sequencing data that Virmid significantly increases overall precision and recall in variant finding. Even in some intractable cases, where the target genome exists in a very small amount ($\alpha \geq 80\%$) in the sample, Virmid shows a near-robust performance. We expect that this improvement will contribute to many related problems from cancer somatic mutation profiling to contaminant genome identification.

**Results and discussions**

**Virmid workflow:** The Virmid workflow is shown in Figure 1. The input to Virmid includes short reads sequenced from a pure control sample and a potentially mixed disease sample. As a pre-processing step, the reads were aligned to the reference genome to generate sequence alignments. Second, the alignments were corrected using post-processing tools such as GATK’s IndelRealigner [22]. Third, BAF was calculated from the corrected alignments for every nucleotide position. Fourth, initial filters were applied for quality
control as well as reduction in sample size. Due to the large size of usual genomic data, we implemented a multi-tier sampling strategy (please see Additional file 1, Figure S1), which reduced overall running time and disk usage about 7 to 10 fold. Finally, two filtered alignment files (pileup format) from control and disease sample were prepared as input.

The first step for Virmid is the estimation of $\alpha$. Here we denote $A$ for the reference allele, and $B$ for non-reference. The set of diploid genotypes is, thus, given by $G = \{AA, AB, BB\}$. As $\alpha$ is a global parameter that affects all positions equally, a small subset of positions is sufficient for estimation. To obtain robust and unbiased estimates, we used a number of criteria: 1) We used only the positions with no $B$ allele observed in controls to maximize the chance of getting true somatic mutations; 2) we eliminated positions with very high or low coverage suggestive of CNV; 3) more filters were applied so that the selected positions had mapping and sequencing quality values above a certain threshold; and, 4) the known mappability [21] of the corresponding reference region had to be above a certain threshold (see Methods for detailed setting of Virmid). Finally, the sites were filtered to remove alleles with BAF lower than a parameter $R$ ($0 < R \leq 1$). While this makes the filtered list biased for higher BAF mutations, the explicit parameter value $R$ was incorporated in our model to correct that bias (see Methods).

Virmid estimated $\alpha$ from the sampled sites using a Maximum Likelihood Estimator (MLE) [23] with gradient descent search, and simultaneously estimated a joint genotype probability matrix $G$, based on the estimated $\alpha$. The estimated $\alpha$ value and the matrix $G$ were used to call the most likely genotype at every nucleotide position. Finally, somatic mutation filters were applied to reduce false positives (see Methods).

The overall pipeline including data preprocessing is implemented as a single Java program. We utilized open source libraries such as samtools [2] and picard [24] to increase efficiency and compatibility of the program. We also reduced significant amount of memory and disk usage by minimizing generation of temporary files.

**Test on simulated data**

Simulated control and disease genomes were prepared from human chromosome 1 (hg19) by introducing random mutations. Out of 275,814 germline (mutation rate: $10^{-3}$) and 2,522 somatic mutations (mutation rate: $10^{-5}$), 47,796 and 257 mutations were located in non-detectable regions (e.g. reference genotype is unavailable) leaving only 228,018 and 2,265 mutations as a true answer set. Disease samples were generated by artificially mixing two genomes in 11 different portions ($\alpha = 1\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, \text{and } 90\%$). The Illumina-like short reads (read length=100 bp) in a medium (40x)
coverage were mapped to the reference and used as the input of the Virmid pipeline (see Methods for the complete protocol).

**Contamination level estimation**

The estimated sample contamination levels on the 11 different mixtures are shown in Figure 2 (red line with circles) and Table 1. The overall accuracy was near perfect with only 0.53% mean deviation from the true value. To test robustness, we ran Virmid 20 times for each data varying the sampling parameter $R$ (minimum BAF to control the number of sampling points). All replicated results were bounded within 2% ($0.19 \leq \text{st.dev.} \leq 0.53$) showing that the estimation with MLE is robust (Table 1). We found there is only a minor ($\pm \sim 1$) overestimation in very lowly ($\alpha \leq 5\%$) and underestimation in very highly ($\alpha \geq 80\%$) contaminated samples. However the error size was negligible compared to a conventional call-based calculation (see Methods), which estimates $\alpha$ based on initially identified somatic mutations (Figure 2, green line with circles).

We note two types of biases in the call-based method (see Figure 2, green line), loss of reads (LOR) and loss of variants (LOV) that lead to overestimation and underestimation of $\alpha$ respectively. LOR originates from the difference of mappability among short reads at the site of somatic mutation. Assume a disease genome has a heterozygous somatic mutation ($\text{AB}$) at position $i$. As the reference genome has an $\text{A}$ genotype, reads with $\text{A}$ at position $i$ are more likely to be mapped. This results in underrepresentation of $\text{B}$ allele, followed by an overestimation of $\alpha$ (see Methods for details). LOV is caused from the tendency that the variant calling is more favorable to the regions of higher BAF. Assume that a disease sample of $\alpha$ contamination has $\text{AB}$ heterozygous mutations. In these positions, BAF follows a binomial (or similar) distribution with a probability of choosing $\text{B}$ allele of $(1 - \alpha)/2$. In conventional SNV calling algorithms, the positions with higher BAF are easier to be discovered. Therefore, the distribution of BAF of the called mutations is shifted upward, which results in overrepresentation of $\text{B}$ allele, followed by underestimation of $\alpha$.

The effects of the two estimation biases are dependent on $\alpha$. The difference in the number of mapped reads with $\text{A}$ and $\text{B}$ allele is proportional to the absolute number of reads generated from disease genome. So, the LOR bias is inversely proportional to $\alpha$. On the other hand, LOV effect is proportional to $\alpha$ because the SNV calling performance remains robust in the low contamination samples. The combined effect explains the bimodal error distribution of call based method. Eventually, the estimation result shows the suggested biases exist and are corrected efficiently in Virmid (see Methods).

Because we do not rely on initial mutation calling, the sites used for $\alpha$ estimation may contain
non-mutated positions. As we already filtered out all the positions where control sample has one or more B alleles, only three possible joint genotypes remain: (AA-AA, AA-AB, and AA-BB). Thus, Virmid estimates the frequencies of these genotypes along with $\alpha$. Since the likelihood we used is dependent on $\alpha$ and the frequencies of the genotypes, we attempt to find the combination of $\alpha$ and the genotype frequencies that maximizes the likelihood. We showed empirically that the likelihood space is convex and maximized near the true answers (please see Figure 3A, and Figure S3/S4 in Additional file 1 for all $\alpha$ values). Therefore, we used a fast gradient descent search algorithm to get MLE estimates, instead of slower EM like algorithms (Figure 3B, see Methods).

**Somatic mutation calling**

We ran Virmid to predict the most probable genotype for each nucleotide position in the simulated data set using the estimated $\alpha$. Somatic mutations were called based on the predicted genotype probabilities after filtering. To evaluate the influence of $\alpha$, we compared the result with those from other SNV calling tools including JointSNVMix2, Strelka and VarScan2 (Figure 4). Virmid and VarScan2 can take tumor purity and ran in two different modes (with and without $\alpha$); note that VarScan2 does not estimate $\alpha$ and was provided with our estimation. Strelka generates two outputs, a standard and a filtered mutation lists (see Methods for detailed protocol). Evaluation was done against the 2,265 true somatic mutations cataloged from simulation procedure based on precision-recall curves (Figure 4A) where exact genotype probabilities are available (Virmid and JointSNVMix2), or a single precision, recall, and f-score (Figure 4B) where only final mutation lists are provided (Strelka and VarScan2).

The performance of all algorithms was comparable for relatively low contamination ($\alpha \leq 50$), but varied considerably for higher $\alpha$ values. Generally, tools that incorporated contamination level (Virmid w/$\alpha$, VarScan2 w/$\alpha$, Strelka; Strelka has a non-explicit noise level that may indicate tumor purity) outperformed the ones that did not (Virmid w/o $\alpha$, VarScan2 w/o $\alpha$, JointSNVMix2). The point is clearer when the same tool with different $\alpha$ parameters are compared (Virmid and VarScan2 w/$\alpha$ ↔ w/o $\alpha$).

A detailed analysis on the BAF distribution in different call sets provides a second test of performance (Figure 5). Note that the mean BAF is given by $(1 - \alpha)/2$. As expected, the BAF distribution of the true mutation set (Figure 5, pink bar) decreases as $\alpha$ increases; With low $\alpha$, there is no major problem in detecting somatic mutations because BAF is high enough to be distinguished from non-mutational sequencing/mapping error frequencies. However, in cases with high $\alpha$, algorithms start to fail in calling somatic mutations with relatively low BAF. The most extreme case is that there is no B allele observed in
disease sample due to the low portion of true disease genome and variance. For example, 316 out of 2,265 somatic mutations sites had no reads with B allele in the sample with \( \alpha = 90\% \). As there is no feasible way to detect these sites, the called mutation set must have higher BAF distribution.

Finally, we revisit the coverage issue in SNV calling. Although moderate (40x) coverage is generally considered sufficient for SNV calling (calculated from [25]), high contamination needs higher coverage. For example, with 90% contamination, only 5% of reads (or 2 reads in 40x coverage) will sample B allele. Higher coverage adds more confidence to each position’s genotype probability by providing more reads to observe. To see the effect of higher coverage, we generated 100x simulation data sets from three highly contaminated data (\( \alpha = 70\% \), 80%, and 90%). The data sets were analyzed using Virmid. Table 2 shows the improvement of prediction performance (especially in recall). With 80% contamination, Virmid could identify 94% of the true somatic mutations with almost perfect accuracy. Even with 90% contamination, 68% of true somatic mutations could be discovered, which is improved by more than 250% (611 to 1545) from 40x coverage result with better precision (0.96 to 0.98). From the result, we can conclude that deeper coverage 1) greatly improves the mutation finding in highly contaminated samples and 2) should be considered when sample purity is questionable.

Although testing on simulated data gives a unique benefit to test exact precision and recall by proving true answer set, there is some limitation too. Many difficulties in somatic mutation detection come from ambiguous read mapping. In simulation, the same reference genome assembly is used in artificial read generation. However, in real data, donor genome contains significantly more variations other than SNVs such as copy number variations and structural variations [26]. Therefore, we proceed to test on publicly available disease data for more extensive validation of Virmid’s performance.

**Test on breast cancer data**

To test with real disease data, we ran Virmid on 15 whole exome breast cancer dataset generated from The Cancer Genome Atlas (TCGA) project [27] (please see Additional file 1, Table S2 for the sample list).

Breast cancer is known to contain a large amount of stromal cells in the tumor mass [28], which makes the relevant genetic studies more challenging. In this context, estimating and considering the level of impurity might be necessary for more accurate analysis of somatic mutation finding.

Before reporting accuracy, exact meaning of sensitivity and specificity of the test should be defined. Note that in the absence of a complete list of true positives, predicted but not confirmed calls cannot be treated as false positives. To test specificity, we generate false tumor/normal pairs from the same sample, where
every call is a false positive. We applied a “virtual tumor” approach, suggested by MuTect study [29] for the specificity test.

We first measured the sample impurity by estimating $\alpha$ (Figure 6-A, blue area). The values were ranged from 0.41 to 0.77. Unlike other monoclonal disease, we note that there is a chance of overestimation due to the genetic heterogeneity in cancer (independently addressed in other study [19]). However, we found that the impurity range is generally consistent with the previous measurement from 21 breast cancers [28] and with TCGA’s sample quality control step (see Methods).

There were 1654 experimentally validated mutations in the 15 exome data (see Methods). Figure 6-A shows the measured sensitivity from four different callers (Virmid, Strelka, JointSNVMix and MuTect); we excluded VarScan2 because it has been used to generate the initial TCGA callset. The recent JointSNVMix contains its own filtering module. We found the filtered version of JointSNVMix is always better than unfiltered by not losing a single true answer. So, we will only use the filtered version for JointSNVMix for comparison. For Strelka, we found that the default Strelka filter eliminates most of the true answers (see Methods). So, we manually disabled its read depth filter (DP) that is too strict to exome sequencing data to get the best filtered result of Strelka. Out of 1654, we only missed 23 validated mutations (36 in unfiltered Strelka, 47 in MuTect, 95 in filtered Strelka, 255 in JointSNVMix) to mark the best sensitivity 98.61% (97.82% in unfiltered Strelka, 97.16% in MuTect, 94.26% in filtered Strelka, 84.58% in JointSNVMix). As pre-calculated $\alpha$ increases, a clear decrease of sensitivity was observed in JointSNVMix, which does not estimate sample impurity (Figure 6-A, black trend line). The relatively smaller sensitivity increase compared to Strelka and MuTect, which are shown to be consistent in the given $\alpha$ range, can be explained by other features such as unbiased estimation of sample purity and more rigorous filtering.

We then analyzed the types of negative calls. Out of 47, 95 and 255 false negative calls from filtered Strelka, MuTect and JointSNVMix, Virmid recovered 32, 81, and 252 mutations, which correspond to 68.1%, 85.3%, and 98.8% of each false negatives. The read depths (of normal and tumor sample) and allele frequencies of the recovered mutations are shown in Figure 6-B. We found the main reason of missing calls in JointSNVMix is low allele frequency (yellow dots in Figure 6-B upper, $\mu = 13.3\%$), while many from Strelka and MuTect resulted from low read depth (blue and purple dots, median read depth=18). Because the subtle changes in genotype probability more critically affect in lower read depth, this results also proves the reliability of Virmid’s genotype calculation model. We also analyzed the 23 false negatives of Virmid (Figure 6-B lower). In most case (15/23), the read depth in normal sample was extremely low (red dots in Figure 6-B lower, median read depth=7). We found Virmid called these mutations as germline
mutations (AB-AB). Although missed in consequence, we are convinced our call is not wrong in theory, because in low read depth, the probability of having only reference sequences from AB genotype (solely by chance, calculated from binomial distribution) is much higher than the prior probability of having somatic mutation (AA-AB). The only solution for this, is increasing (e.g. by balancing out) read depth in normal sample, because calling these regions will greatly increase false positive rate. Overall, the false negatives from Virmid were partially recovered by other tools (8, 9, and 20 by Strelka, MuTect and JointSNVMix).

In this test, sensitivity increases monotonically according to the total number of calls. It is informative to compare the number of calls to achieve a similar level of sensitivity. Figure 6-C shows the total number of predicted mutations. We found that unfiltered Strelka (∼ 5646 per sample) and JointSNVMix (∼ 7362 per sample) predicted far more mutations than others (941, 336 and 738 per sample in Virmid, filtered Strelka, and MuTect respectively). Although we cannot assert non-validated calls are all false, we can suspect more false positive calls in those two tools.

To ensure the specificity, we designed “virtual tumor” by dividing high depth (> 80) normal samples into two artificial samples including one faked tumor and one faked control sample. Because all the reads are originally generated from the same genome, any positive calls on these samples can be considered as false positives. We ran all tools on five “virtual tumor” sets with same parameters used in breast cancer data to get estimated false positive rates (Figure 6-D). Filtered Strelka showed a surprising specificity (< 0.01 false calls per million base-pair (Mbp); note that the sensitivity was limited. Virmid and MuTect also showed a satisfactory performance (∼ 1 per Mbp). Unfiltered Strelka, which showed almost comparable sensitivity in breast cancer data with Virmid, however, contained more false positives (∼ 3 per Mbp). JointSNVMix, even after applying its own filtering method, is shown to contain more false positives than other tools.

We note that on simulated data, Virmid’s performance compared to other tools was similar for samples with up to 50% contamination, but becomes progressively better for higher contamination levels such as those exceeding 70%. The experimental data presented here is at mid levels of contamination (41% to 77%), which is not ideal to showcase Virmid’s strengths. In the next section, we also discuss a new data set with higher levels of contamination, but without independently validated mutations. As validated data sets grow in number, the advantages of calling mutations after estimating α will be more apparent.

Application to HME exome sequencing data

We applied Virmid to the recently sequenced disease/normal paired data of five hemimegalencephaly (HME) patients [14]. HME is a rare disease characterized by the enlargement and malformation in one
cerebral hemisphere and is known to be an important cause of epilepsy and developmental delay. One distinctive histopathological feature of HME is that the dysmorphic and immature neurons are dispersed in the disease lesion. In this condition, the brain samples from surgical resection are expected to contain significant amount of non-disease cells. Also, the mutation burden measured by whole exome sequencing and mass spectrometry from three previously validated mutations (AKT3 c.49C>T in HME-1565, MTOR c.4448C>T in HME-1563, and PIK3CA c.1633G>A in HME-1573) assured the compromised sample purity (please see Additional file, Table S1). The dropped allele frequencies (9.7%~30.38%) are far less than the expected (50% for heterozygous, 100% for homozygous mutation sites) indicating the existence of reference alleles (AA) from normal cells. As the mutations are believed to occur during early cerebral development and the surgical treatment is done in infants, the low mutation burden is less explained by disease subpopulation.

To estimate the sample contamination level, we ran Virmid 20 times for each whole exome sequencing data with variating sampling parameters (see Methods). The estimated $\alpha$ for the five samples (HME-1563, HME-1565, HME-1573, HME-1574 and HME-1620) are surprisingly high ranging from 64.0% to 84.8% (Figure 6A), which indicates that only 15.2% to 36% of the sample is disease cell. The low variance (<0.075) within the same sample gives high confidence to the estimated values. While our manuscript was under review process, an independent study [30] measured mutation burden of the same disease using 100 single cell sequencing, which reported a consistent result (39% in NeuN + and 27% in NeuN - populations).

We also checked the distribution of BAF at the site of candidate somatic mutations (Figure 6B). Note the overall drop of BAF towards zero as shown in the high $\alpha$ simulated examples (Figure 5 right). The average BAF was perfectly negatively correlated with the expected BAF calculated from $(1 - \alpha)/2$. For example, the sample HME-1573, which has been predicted to have lowest contamination ($\alpha$=64.0%), had the highest BAF distribution. This negative correlation is consistent with our assumption that higher $\alpha$ leads to lower BAF.

Although there is no efficient way to measure the true contamination levels in the sample, the allele frequencies of validated mutations (AKT3, MTOR and PIK3CA) provide a good reference. In HME-1563 and HME-1573, the validated BAF values (Figure 6B, red triangles) were very close to the expected heterozygous BAF. In HME-1565, the validated BAF (28%) is twice as the expected BAF. We checked the genotype probability of the corresponding mutation and found that its probability of being homozygous is significantly high (7.1%, ranked 2nd out of 496 mutations). In all cases, mutation burdens measured from peak intensity of mass spectrometry (Figure 6B, blue squares) are also bounded in a low BAF range.
Although the allele frequencies at a single position is still highly variable to be a sample level estimator, we are convinced that the HME samples contain large amount of normal cells and the estimated $\alpha$ are reasonably ranged by aggregating various inspections. Identifying and quantifying more somatic mutations will lead to better validation of the sample contamination level.

Finally, we ran Virmid on the same dataset to find novel somatic mutations that might have been missed due to the sample impurity. Virmid predicted totally 2,787 somatic mutations from five individual data sets, only 653 (23.4%) of which were predicted from the previous finding. (see Table 3). Note that the number of newly predicted mutations is correlated with the estimated $\alpha$; the higher the $\alpha$ is, the more somatic mutations could be missed from conventional approaches. Out of 2,134 newly predicted sites, 1,377 were located in exon region containing 923 missense events. As we expected, Virmid successfully discovered all the previously reported de novo mutations of PIK3CA, AKT3, and MTOR with very high confidence ($p \sim 1.0$). We focused on the two (HME-1574 and HME-1620) samples where no meaningful somatic missense mutations were not detected in the previous study. In HME-1574, Virmid discovered a novel somatic mutation of MTOR (MTOR p.Ala1517Thr) located nearby one of the validated mutations (MTOR p.Cys1483Tyr). At this site, only 3 out of 54 mapped reads represented B allele (BAF=5.56%) while no B allele was found in the control sample; we could not find any sign of mapping ambiguity nor compromised base call quality. We are convinced that the extremely reduced BAF is the major reason of the unsuccessful finding in conventional approaches. A functional analysis showed that the region is highly conserved (phastCons [31] score=1.0) and no other mutation has been known at the same site so far. We expect a further experimental validation can verify the mutation in other patients as well as the more accurate effects on the protein activity (e.g. gain of function). Virmid also detected more candidates of de novo somatic mutations in HME-1620, but we could not find meaningful missense mutations directly linked to the PI3K-ATK-mTOR pathway or HME pathogenesis. It is possibly due to either that disease-causing mutations exist other than in coding regions or that the combinatorial effect of the low coverage of current exome sequencing ($\sim$30x) and the high $\alpha$ (84.8%). As we showed in the simulation (Table 2), a much higher coverage ($\sim$100x) might be necessary to secure a prediction power to retrieve sufficient candidate mutations in such a highly impure sample.

Conclusions

In this study, we presented a novel probabilistic method for SNV calling, with two significant contributions. First, Virmid can estimate accurate sample composition, or level of contamination of disease sample
without genotyping. This not only reduces extra time and cost, but removes severe estimation biases that come from initial SNV calling. Second, Virmid increases genotyping accuracy, especially somatic mutation profiling, by rigorously integrating the sample composition parameter into SNV calling model. We showed Virmid outperformed all recent tools in finding somatic mutations particularly in highly contaminated samples. In applying Virmid to HME disease/normal paired data sets, we discovered previously unknown sample impurity and somatic mutations. Our results suggest that it is important to estimate sample composition for all tumor/normal paired data when the sample purity is questionable and explicitly consider the composition in SNV calling if the sample is highly (>50%) contaminated. The robustness of Virmid to high contamination also makes it applicable to mutation identification in other challenging cases, such as low amounts of fetal DNA in maternal plasma.

Materials and methods

Virmid model

Virmid consists of three parts: $\alpha$ estimator, $G$ estimator, and caller (Figure 1). In the $\alpha$ estimator and $G$ estimator, we use the Maximum Likelihood Estimation (MLE) method. The caller calls somatic variants using the Bayesian inference with the estimated joint genotype probability matrix as the prior distribution. To describe the estimation methods in detail, we first define the likelihood function and then describe how the likelihood function is used for each part of Virmid.

Likelihood function

Denote the set of the reads from control sample by $C$. Given $C$ and a nucleotide position $i$, the values of the reads mapped to the position $i$ are represented by a vector (read vector) denoted by $C^i$ where the $j$-th element of $C^i (C^i_j)$ is given by the value (i.e., A or B) of the $j$-th mapped read. For the set of the reads from disease-control mixture sample $D$, $D^i$ is defined similarly.

For the likelihood function, the parameters are $\alpha$ and $G$. $\alpha$ is the proportion of control in the disease sample. The joint genotype probability matrix $G$ is a $3 \times 3$ matrix that describes the average rate of the control-disease joint genotypes $g$ (control genotype) and $g'$ (disease genotype). Given a joint genotype $g$ and $g'$, the element of $G$ that corresponds to the joint genotype is specified by $G_{g,g'}$. For example, $G_{AA,AB}$ is the rate at which the joint genotype of a random position is AA and AB. Note that $G$ is not position-specific; it describes the average rate of each joint genotype over the whole positions. The position specific rates are calculated in the caller using $G$ as the prior distribution of joint genotypes (see below).
We assume i) all reads at different positions are independent and ii) all reads at the same position are independent given the joint genotype of that position. Given \( \theta := \{ \alpha, \mathcal{G} \} \), the likelihood function is written by

\[
L(\theta | C, D) = \prod_{i} \left( \sum_{(g, g') \in G} \left\{ \frac{1}{\alpha} P_\theta(C_i | g) \cdot \prod_{j=1}^{|C^i|} P_\theta(C_j | g) \cdot \prod_{j=1}^{|D^i|} P_\theta(D_j | g, g') \right\} \right)
\]

where \( G \) is the set of all possible joint genotypes, and \(|C^i| \) and \(|D^i| \) denote the read depths of \( C^i \) and \( D^i \), respectively (see the supplementary note for the derivation of the likelihood function).

The probability in (a) is defined by \( P_\theta(g, g') := \mathcal{G}_g \cdot g' \). The probabilities in (b) and (c) are defined so that their definitions incorporate the read error rate, mapping error rate, and the loss of reads (LOR) bias. The LOR bias is that reads with more mismatches are less mappable to the reference sequence, which is due to the maximum number of allowed edit distance. First we derive the probabilities in (b) and (c) without considering LOR bias. Denote the read and mapping error probability of \( C_{ij} \) by \( r \) and \( m \), respectively. The probability in \( P_\theta(C_{ij} | g) \) is defined by

\[
P_\theta(C_{ij} | g) := \begin{cases} 
\frac{1}{\alpha} \cdot r + \left(1 - r\right) \beta & \text{if } C_{ij} = A \\
1 - m \cdot \left(\frac{1}{\alpha} \cdot r + \left(1 - r\right) \beta\right) - \left(1 - m\right) \cdot \left(\frac{1}{\alpha} \cdot r + \left(1 - r\right) \gamma\right) & \text{if } C_{ij} = B
\end{cases}
\]

where \( \gamma \) is the probability that an error-free read (i.e., without mapping or read error) has \( A \) allele given \( g \) and \( \beta \) is the probability that an incorrectly mapped read has \( A \) allele. \( \gamma \) can be calculated as

\[
\gamma = \begin{cases} 
1 & \text{if } g = AA \\
\frac{1}{2} & \text{if } g = AB \\
0 & \text{if } g = BB
\end{cases}
\]

\( \beta \) may depend on experimental settings but is simply set to be \( \beta = 0.99 \). For simplicity, we denote the right hand side of the equation (2) as a function \( \mu_{g}(C_{ij}) \).

The probability \( P_\theta(D_{ij} | g, g') \) in (c) is given by

\[
P_\theta(D_{ij} | g, g') := \alpha \mu_{g}(D_{ij}) + (1 - \alpha) \mu_{g'}(D_{ij}).
\]

The right hand side of the equation (4) is denoted by \( \nu_{g'}(D_{ij}, \alpha) \).

Second we derive the probabilities in (b) and (c) considering the LOR bias. Denote the probability that a read with at least one \( A \) (or \( B \)) allele is mappable (i.e., the edit distance of the read is less than the maximum allowed edit distance) by \( x(A) \) (or \( x(B) \)). Then we obtain

\[
P_\theta(C_{ij} | g) = \frac{\mu_{g}(C_{ij}) \cdot x(C_{ij})}{\mu_{g}(A) \cdot x(A) + \mu_{g}(B) \cdot x(B)}
\]
The right hand side of the equation (5) is denoted by $f_g(C_{ij})$. We also obtain

$$P_\theta(D_i^j | g, g') = \frac{\nu_g'(D_i^j, \alpha) \cdot x(D_i^j)}{\nu_g(A, \alpha) \cdot x(A) + \nu_g'(B, \alpha) \cdot x(B)}.$$ (6)

The right hand side of the equation (6) is denoted by $h_g'(D_i^j, \alpha)$. The detailed derivation of the above probabilities $x(C_{ij}), \mu_g(C_{ij}), \nu_g'(D_i^j, \alpha), f_g(C_{ij})$, and $h_g'(D_i^j, \alpha)$ is given in the supplementary note.

Overall, we have

$$L(\theta | C, D) = \prod_i \left( \sum_{(g, g')} \left\{ \nu_{g'}(D_i^j, \alpha) \cdot \prod_{j=1}^{\mid C_i \mid} f_g(C_{ij}) \cdot \prod_{j=1}^{\mid D_i^j \mid} h_g'(D_i^j, \alpha) \right\} \right).$$ (7)

**Basic model**

Using the likelihood function defined above, the MLE of $\theta = (\alpha, G)$ can be obtained by

$$\hat{\theta} = \arg \max_{\theta} L(\theta | C, D)$$ (8)

with proper constraints (e.g., $0 \leq \alpha \leq 1$ and $G_{AA,AA} > 100 \cdot G_{AA,AB}$ - see the supplementary note for the constraints we used). Since the exact global maximum point cannot be derived analytically, one needs to use a numerical approach to find it. To make a numerical approach work, one should carefully avoid the local maximum points. However, even if we impose strong constraints, many local maximum points may be present in the likelihood function. Moreover, in terms of the estimation of $\alpha$, not all read vectors are useful; some read vectors deteriorate the estimation (see below). Therefore, we try to estimate $\alpha$ and then estimate all elements in $G$ (with the estimated $\alpha$).

**Estimation of $\alpha$**

For the estimation of $\alpha$, the disease read vectors generated from the same control genotype $(g)$ and disease genotype $(g')$ are simply noisy sample points conveying no information. Thus, we want to sample the read vectors generated from different $g$ and $g'$, but without the initial calling. Also we want to fix $g = AA$ so that the number of parameters to be estimated can be minimized. Denote the number of B's in a control read vector $C_i$ (or in a disease read vector $D_i$) by $\langle C_i \rangle$ (or $\langle D_i \rangle$). We sample the positions $i$ such that $\langle C_i \rangle = 0$ and $\langle D_i \rangle > R$ (i.e., the BAF of $D_i$ is larger than $R$) for a real value $0 < R \leq 1$. Imposing $\langle C_i \rangle = 0$ minimizes the chance that $g = AA$, and imposing $\langle D_i \rangle > R$ increases the chance that $g' \neq AA$.

If $R$ is too large, however, we may not have sufficient number of samples for the estimation. On the other hand, if $R$ is too small, the samples may contain too many read vectors from $g = AA$ and $g' = AA$ that serve...
as noise. Thus, we estimate $\alpha$ using different values of $R$ and take the median of the estimates. Table 1 shows that our $\alpha$ estimator is quite robust for different values of $R$. We also outputs the asymptotic variance of the estimated $\alpha$ using the outer products of the first derivatives of the log likelihoods (called BHHH estimator [32]).

With the selected samples as above, we only estimate 4 parameters (instead of 10 - $\alpha$ and 9 elements in $G$): $\alpha$, $G_{AA,AA}$, $G_{AA,AB}$, and $G_{AA,AB}$; other elements in $G$ are set to a very small number close to 0. In this step, the parameters except $\alpha$ are estimated simply for better estimation of $\alpha$.

Unfortunately, the sampling described above introduces estimation bias if we use the likelihood function as is because the sampling procedure inflates the number of B’s in the disease read vector (please see Additional file 1, Figure S2 for an example). To take this sampling bias into account, we modify the likelihood function as

$$L_R(\theta|C, D) = P_0(C, D) \frac{|D^i_j|}{|D^i|} > R \text{ for all } i = \prod_i \left( \sum_{(g,g') \in G} \left\{ G_{g,g'} \cdot \prod_j f_j(C_j^i) \cdot \left\{ \prod_{j=1}^{\frac{|D^i|}{|D^i_j|}} h_{g'}^j(D^i_j, \alpha) \right\} P_0(\frac{|D^i_j|}{|D^i|} > R|g,g') \right\}. \right) \tag{9}$$

The denominator in (d) can be efficiently calculated using a dynamic programming algorithm with the time complexity of $O(R \cdot |D^i|^2)$ (see the supplementary note for the algorithm). As above, the parameter $R$ can be readily incorporated in our model in our method (correcting the possible bias); however, it is very hard to make a rigorous model that takes the LOV (loss of variants) bias found in the calling-based methods into account.

The estimates of the 4 parameters ($\hat{\alpha}$, $\hat{G}_{AA,AA}$, $\hat{G}_{AA,AB}$, and $\hat{G}_{AA,AB}$) that maximize the likelihood are found by the feasible direction method [33], a gradient descent search method with constraints. Note that only the estimate of $\alpha$ is retained for the next step.

Figure 3A shows the values of log likelihood over different $\hat{\alpha}$ and $\hat{G}_{AA,AB}$ (for each point, other parameters are optimized). For low $\alpha$, the optimum $\hat{G}_{AA,AB}$ is almost 1. However, when $\alpha$ is larger, the likelihood is maximized for low $\hat{G}_{AA,AB}$. For example, when $\alpha = 0.9$, the maximum likelihood is found when $\hat{G}_{AA,AB} \approx 0.1$. Such estimation results are predicted because for high $\alpha$, even disease read vectors generated with $g' = AB$ would not have a sufficient number of B’s to distinguish between $g' = AB$ and $g' = AA$. Even if we sample disease read vectors with many B’s, there are often many vectors from $g' = AA$ which leads to high value of $\hat{G}_{AA,AB}$. 

15
**Estimation of the joint genotype probability matrix**

In this step, we estimate $G$ with the estimated $\alpha$. We sample 1,000,000 positions except ones at which the number of $B$ in the disease read vector is zero (i.e., $\langle D^i \rangle = 0$). Such positions are not sampled because the Virmid does not analyze such points for SNP calling. We estimate $\hat{G}$ that maximizes this likelihood function in (7) using the feasible direction method.

**Calling genotypes**

Given the estimated $\hat{\theta} = (\hat{\alpha}, \hat{G})$ and a position $i$, we first calculate $G^i$, the posterior distribution of genotypes at the position $i$, (with $\hat{G}$ as the prior distribution) by

$$G^i_{g,g'} = \frac{P^i_{\hat{\theta}}(g,g'|C^i,D^i)}{\sum_{(k,k') \in G} P^i_{\hat{\theta}}(k,k'|C^i,D^i)}$$

$$= \frac{P^i_{\hat{\theta}}(C^i,D^i|g,g') \cdot P^i_{\hat{\theta}}(g,g')}{\sum_{(k,k') \in G} P^i_{\hat{\theta}}(C^i,D^i|k,k') \cdot P^i_{\hat{\theta}}(k,k')}$$

$$= \frac{\hat{G}^i_{g,g'} \cdot \prod_{j=1}^{\left|C^i\right|} f^i_g(C^i_j) \cdot \prod_{j=1}^{\left|D^i\right|} h^{i'}_g(D^i_j, \hat{\alpha})}{\sum_{(k,k') \in G} \hat{G}^i_{k,k'} \cdot \prod_{j=1}^{\left|C^i\right|} f^i_k(C^i_j) \cdot \prod_{j=1}^{\left|D^i\right|} h^{i'}_k(D^i_j, \hat{\alpha})}.$$  

(12)

Then Virmid calls the position $i$ a somatic variant if $1 - (G^i_{AA,AA} + G^i_{AB,AB} + G^i_{BB,BB})$ exceeds 0.5.

**Filtration of read data**

Reads or positions that may contain unreliable information were filtered out from observation. Two filtering criteria have been established depending on applying Virmid step. The first filtering scheme is for selecting observation points for $\alpha$ estimation. The purpose of filtering in this step is to eliminate positions possibly contain following noises: 1) $B$ alleles originated from sequencing error, 2) $B$ alleles originated from mapping error, 3) $B$ alleles originated from non-reference control genotype. The second filtering scheme is for calling somatic mutations. The purpose in this step is to remove false-positive somatic mutations, which are usually one of the following cases: 1) both samples have reference genotype ($AA$-$AA$), but $B$ alleles are observed by sequencing or mapping errors, 2) both samples have non-reference genotype ($AB$-$AB$, germline mutation), but significant BAF differences are observed. To satisfy above criteria, we divided all frequent miscalling events into seven classes. 1) mapping quality (MQ): mutations are filtered out if the mapping quality of their corresponding $B$ allele read is significantly worse ($> 30$ MAPQ score) than $A$ allele reads. Or overall ratio of ambiguously mapped reads ($< 17$ MAPQ score) is more than a threshold ($> 0.4$).
2) read offset filter: if the position of $B$ allele in the read is significantly biased at the both ends. (z-score > 3). 3) indel proximity (PRX): if more than 50% of $B$ alleles are located within 10 bp of nearby indels. 4) tri-allele (TRI): if the major allele frequency is less than 0.9. 5) base quality (BQ): if the mean base call quality of $B$ allele read is less than 20. 6) number of mismathc (NM): the mean number of mismatches per read is bigger than 3. Or more than 60% of the reads are soft/hard clipped. 7) allele frequency (AF): if the absolute number of $B$ allele is less than a certain threshold (3), or BAF in control is larger than one tenth (1/10) of that in disease sample. The filters are differently applied in $\alpha$ estimation and mutation calling. For $\alpha$ estimation, our goal is to eliminate germline mutations ($AB$-$AB$) only allowing reference and somatic mutation alleles. To do so, we strictly apply MQ, PRX and NM filters to prevent potential mapping errors. For mutation calling, we apply all the seven filters with empirically known parameters. These parameters can be also defined by users.

Data preparation

Simulated data

First, two diploid genomes were simulated: a normal genome and a disease genome. The normal genome was created by using the hg19 genome as a template and infusing germline SNPs found in dbSNP 135 [34] at a rate of one SNP per thousand nucleotides. Somatic mutations were introduced by perturbing a nucleotide to any of the other three nucleotides with equal probability at rates of $10^{-5}$ mutations per nucleotide to simulate a disease genome. Both of these simulations were carried out using in-built python functions. The python scripts are available upon request.

GemSim v1.5 [35] was setup to generate paired-end 100 bp reads using the Illumina paired-end error model. The number of reads necessary was calculated using the average coverage of the sample (40x and 100x). The metagenomic mode was configured with four genomes: normal haplotype 1, normal haplotype 2, disease haplotype 1, and disease haplotype 2. The relative abundance of each genome was calculated based on the contamination level ($\alpha=1\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 60\%, 70\%, 80\%, \text{and } 90\%$). For the normal sample, the metagenomic mode was configured with the normal haplotype 1 and normal haplotype 2 in equal abundances. All the reads were aligned using bwa [36] and passed through the GATK data processing pipeline for variant calling including indel realignment and base quality score recalibration. The resulting BAM files were fed into the variant calling tools.
**Breast cancer data**

There were 545 tumor to normal matched samples with verified somatic mutations. The putative somatic mutations were validated using Illumina Capture gDNA technologies. For our purposes of assessing normal cell contamination in tumor specimens, we downloaded the matched tumor/normal samples listed as whole exome sequencing (WXS) on CGHub (https://cghub.ucsc.edu/) under controlled access. As essential post-processing on the mapping such as indel realignment and quality recalibration is time-consuming, we limit our analysis to randomly selected 15 patients for use as our gold standard. (please see Additional file 1, Table S2 for full list). The 1654 validated mutations were extracted from accompanying mutation annotation format (MAF) file by the following criteria: 1) field “Validation_Status” (column 25) is “Valid”, 2) field “Variant_Type” (column 10) is “SNP”.

**HME data**

Five paired normal data sets (10 BAM files, 76 bp, 30x coverage) processed in the previous study [14] were downloaded with authors’ permission. The alignments are already post-processed using GATK’s pipeline including IndelRealigner, MarkDuplicate and TableRecalibration. Pileup files are generated using samtools mpileup and indexed with tabix. Possible noise reads that do not pass the quality check or possibly included as PCR duplicates were filtered out using samtools view -F option.

**Call-based estimation of α**

Initial SNV calling was done using Virmid w/o α mode. All the filtration steps are applied after the initial calling. The detailed calculation steps are introduced in the supporting information of Snyder et al [15]. Briefly, the number of reads at the called somatic mutation sites are classified by genotypes and allele types. Donor fraction is estimated from $\frac{2N_{AB}(B)+N_{BB}(B)}{N_{AB}(A)+N_{AB}(B)+N_{BB}(A)+N_{BB}(B)}$ where $N_G(A)$ denotes the number of reads at the site of genotype $G$ with $A$ allele.

**Somatic mutation call sets**

Strelka version 0.4.5 was used for the comparative studies. The program was configured using the provided settings for bwa. The results presented show the calls after the first filtration step and after the final filtration step. For JointSNVMix version 0.7 [8], the results were generated using the JointSnvMix2 mode, which included the base and mapping errors. First the program was trained using the jsm.py train, and then joint genotype calls were made using the jsm.py classify option. All of the configurable settings were
left to their default values. For the AUC curves, we varied the probability cutoff necessary to make a joint genotype call to adjust the specificity and sensitivity of the program. For the filtered version of JointSNVMix2, we used JointSNVMix version 0.8 with “--post-process option”. Other parameters were the same as JointSNVMix version 0.7. For VarScan2 [7], we evaluated its performance using version 2.2.11. The pileup files were created using samtools version 0.1.18 [2]. The somatic option was adjusted with the Virmid derived values of contamination for tumor purity calculations. Note that since VarScan2 requires purity estimation, the values fed into the program were (1-predicted contamination level). We also carried out additional filtering using the default options for VarScan2’s methods somaticFilter and processFilter as well as the perl script fpfilter.pl that is available on VarScan2’s website. The most recent version of bam-readcount per its Github repository was used to create the input files for the perl script. Lastly, we ran MuTect as described in the MuTect website, except for “--cosmic” option since the validated mutations (true answer) are included in the corresponding database.

**Program implementation and optimization**

We implemented the Virmid model and its surrounding workflow using Java (JDK version 1.6), samtools and picard library. Post-processed BAM files are converted to pileup format using samtools `mpileup` program. Mapping quality scores are included in the pileup files using `-s` option. To optimize the overall pipeline, we divided the pileup data into three different layers (Figure S1A). Pileup level 1 is the most fundamental data in which we observe $B$ allele at least one time in the disease genome. Level 2 data contains all the positions that the observed BAF is higher than or equal to 5% as well as there is no $B$ allele observed in the control genome. Lastly, level 3 data was generated by increasing the minimum BAF until the number of satisfying positions is less than a threshold (generally 1k to 10k). $\alpha$ estimation is done using the level 3 disease pileup data. After getting $\alpha$, we call genotypes of the positions included in the level 1 pileup files. This hierarchical model significantly reduced overall search space (Figure S1B). Starting from all the nucleotide regions of chromosome 1 (~240 Mbp), the size of level 1 data is reduced to 9%. Final numbers of data points are reduced to 0.28% and 0.00041% of the original number in level 2 and level 3 data respectively. Due to the successful reduction, we decreased the running time for $\alpha$ estimation to less than a few minutes.

**Competing interests**

The authors declare that they have no competing interests.
Abbreviations
BAF - b-allele frequency, CNV - copy number variation, GATK - Genome Analysis Toolkit, HME - hemimegalencephaly, MLE - maximum likelihood estimator, SNP - single nucleotide polymorphism, SNV - single nucleotide variation, TCGA - The Cancer Genome Atlas

Authors’ contributions
SK, KJ, KB, VB designed the base model. SK, KJ implemented the model. SK, KB generated simulated data and tested the tool. SK, KJ, VB prepared the manuscript. JL, JGG motivated the problem and reviewed biological discoveries. AP prepared and processed the breast cancer dataset. ES worked on functional analysis of discovered mutations. HN worked on data analysis and presentation. HL wrote a module for genome mappability score. All authors read and approved the final manuscript.

Acknowledgments
This work is supported by the National Science Foundation (NSF-CCF-1115206), National Institute of Child Health and Human Development (1P01HD070494-01) and National Institute of Health (5R01-HG004962, U54 HL108460).
References
1. Saksena G, Mermel C, Getz G: Developing Algorithms to Discover Novel Cancer Genes: A look at the challenges and approaches. Signal Processing Magazine, IEEE 2012, 29:89–97.
2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP: The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009, 25:2078–2079.
3. Goya R, Sun MG, Morin RD, Leung G, Ha G, Wieand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP: SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 2010, 26:730–736.
4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011, 43:491–498.
5. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L: VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 2009, 25:2283–2285.
6. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L: SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics 2012, 28:311–317.
7. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK: VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research 2012, 22:568–576.
8. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A, Marra MA, Aparicio S, Shah SP: JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. Bioinformatics 2012, 28:907–913.
9. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK: Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 2012, 28:1811–1817.
10. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, Gogos JA, Karayiorgou M: Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat. Genet. 2011, 43:864–868.
11. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornsson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Gunel M, Roeder K, Geschwind DH, Devlin B, State MW: De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 2012, 485:237–241.
12. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N, Nickerson E, Chae SS, Boysen G, Auclair D, Onofrio RC, Park K, Kitabayashi N, MacDonald TY, Sheikh K, Vuong T, Guiducci C, Cibulskis K, Sivachenko A, Carter SL, Saksena G, Voet D, Hussain WM, Ramos AH, Winckler W, Redman MC, Ardlie K, Tewari AK, Mosquera JM, Rupp N, Wild PJ, Moch H, Morrissey C, Nelson PS, Kantoff PW, Gabriel SB, Golub TR, Meyerson M, Lander ES, Getz G, Rubin MA, Garraway LA: Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 2012, 44:685–689.
13. Makino H, Uetake H, Danenberg K, Danenberg P, Sugihara K: Efficacy of laser capture microdissection plus RT-PCR technique in analyzing gene expression levels in human gastric cancer and colon cancer. BMC Cancer 2008, 8:210.
14. Lee JH, Hyunh M, Silhavy JL, Kim S, Dixon-Salazar T, Heiberg A, Scott E, Bafna V, Hill KJ, Collazo A, Funari V, Russ C, Gabriel SB, Mathern GW, Gleeson JG: De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat Genet 2012.
15. Snyder TM, Khush KK, Valentine HA, Quake SR: Universal noninvasive detection of solid organ transplant rejection. Proceedings of the National Academy of Sciences 2011, 108:6229–6234.
16. Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, Simmons LE, Gammill HS, Rubens CE, Santillan DA, Murray JC, Tabor HK, Banshad MJ, Eichler EE, Shendure J: Noninvasive Whole-Genome Sequencing of a Human Fetus. Science Translational Medicine 2012, 4:137ra76–137ra76.

17. Yau C, Mouradov D, Jorissen R, Colella S, Mirza G, Steers G, Harris A, Ragoussis J, Sieber O, Holmes C: A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biology 2010, 11:R92.

18. Cibulskis K, McKenna A, Fennell T, Bans E, DePristo M, Getz G: ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics 2011, 27:2601–2602.

19. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, Pellman D, Levine ES, Meyerson M, Getz G: Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. 2012, 30:413–421.

20. Su X, Zhang L, Zhang J, Meric-Bernstam F, Weinstein JN: PurityEst: estimating purity of human tumor samples using next-generation sequencing data. Bioinformatics 2012, 28:2265–2266.

21. Lee H, Schatz MC: Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 2012, 28:2097–2105.

22. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER, Wilson RK, Ally A, Balasundaram M, Butterfield YS, Carlsen R, Carter C, Chu A, Chuah E, Chun HJ, Coope RJ, Dhalla N, Guin R, Hirst M, Holt RA, Lee D, Li HI, Mayo M, Moore RA, Mungall AJ, Pleasance E, Robertson J, Schein JE, Shafiee A, Sipahimalani P, Slobodan JR, Stoll D, Tam A, Thiessen N, Varhol RJ, Wye N, Zeng T, Zhao Y, Birol I, Jones SJ, Marra MA, Cherniack AD, Saksena G, Onofrio RC, Pho NH, Carter SL, Schumacher SE, Tabak B, Hernandez B, Gentry J, Nguyen H, Crenshaw A, Ardlie K, Beroukhim R, Winckler W, Getz G, Gabriel SB, Meyerson M, Chin L, Park PJ, Kucherlapati R, Hoadley KA, Auman J, Fan C, Turman YJ, Shi Y, Li L, Topal MD, He X, Chao HH, Prat A, Silva GO, Iglesia MD, Zhao W, Usary J, Berg JS, Adams M, Booker J, Wu J, Gulabani A, Bodenheimer T, Hoyle AP, Simons JV, Soloway MG, Mose LE, Jaffe RS, Balu S, Parker JS, Hayes D, Perou CM, Malik S, Maharukar S, Shen H, Weisenberger DJ, Triche T, Lai PH, Bootwalla MS, Maglinte DT, Berman BP, Van Den Berg DJ, Baylin SB, Laird PW, Creighton CJ, Donehower LA, Getz G, Noble M, Voet D, Saksena G, Gehlenborg N, DiCara D, Zhang J, Zhang H, Wu CJ, Lawrence MS, Zou L, Sivachenko A, Lin P, Stojanov P, Jing R, Cho J, Sinha R, Park RW, Nazaire MD, Robinson J, Thorvaldsdottir H, Mesirov J, Park PJ, Chin L, Reynolds S, Kreisberg RB, Bernard B, Bressler R, Erkkila T, Lin J, Thorsson V, Zhang W, Shmulevich I, Ciriello G, Weinhold N, Schultz N, Gao J, Cerami E, Gross B, Jacobsen A, Sinha R, Aksyev B, Antipin Y, Reva B, Shen R, Taylor BS, Ladanyi M, Sander C, Anur P, Spellman PT, Lu Y, Liu W, Verhaak RR, Mills GB, Afban R, Zhang N, Broom BM, Casasent TD, Wakefield C, Unruh AK, Bagley R, Koombe K, Weinstein JN, Haussler D, Benz CC, Stuart JM, Benz SC, Zhu J, Szeto CC, Scott GK, Yau C, Paull EO, Carlin D, Wong C, Sokolov A, Thusberg J, Mooney S, Ng S, Goldstein TC, Elliott K, Grifrod M, Wilks C, Ma S, Craft B, Yan C, Hu Y, Meierzaman D, Gastier-Foster JM, Bowen J, Ramirez NC, Black AD, Pyatt RE, White P, Zmuda EJ, Frick J, Lichtenberg TM, Brookens R, George MM, Gerken MA, Harper HA, Leraas KM, Wise LJ, Tabler TR, Mccallister C, Barr T, Hart-Kothari M, Tarvin K, Saller C, Sandusky G, Mitchell C, Iacocca MV, Brown J,
Rabeno B, Czerwinski C, Petrelli N, Dolzhansky O, Abramov M, Voronina O, Potapova O, Marks JR, Suchorska WM, Murawa D, Kyed W, Ibbes M, Korski K, Spycha?a A, Murawa P, Braze?ski JJ, Perz H, ?a?niak R, Teresiak M, Tatk? H, Leporowska E, Bogusz-Czerniewicz M, Malicki J, Mackiewicz A, Wisnerowicz M, Le XV, Kohl B, Nguyen VT, Thorp R, Nguyen VB, Sussman H, Bui DP, Hajek R, Nguyen PH, Tran VT, Huynh QT, Khan KZ, Penny R, Mallery D, Curley E, Shelton C, Yena P, Ingle JN, Couch FJ, Lingle WL, King TA, Gonzalez-Angulo AM, Mills GB, Dyer MD, Liu S, Meng X, Patangan M, Waldman F, Stoppler H, Rathmell W, Thorne L, Huang M, Boice L, Hill A, Morrison C, Gaudioso C, Bshara W, Daily K, Egea SC, Pegram M, Gomez-Fernandez C, Dhir R, Bhargava R, Brufsky A, Shriver CD, Hooke JA, Campbell JL, Mural RJ, Hu H, Somiai S, Larson C, Deyarmin B, Kvecher L, Kovatch AJ, Ellis MJ, King TA, Hu H, Couch FJ, Mural RJ, Stricker T, White K, Olopade O, Ingle JN, Lu C, Shen Y, Marks JR, Waldman F, Wisnerowicz M, Bose R, Chang LW, Beck AH, Gonzalez-Angulo AM, Pihl T, Jensen M, Sfeir R, Kahn A, Chu A, Kothiyal P, Wang Z, Snyder E, Pontius J, Ayala B, Backus M, Walton J, Baboud J, Berton D, Nichols M, Srinivasan D, Raman R, Girshik S, Kigonya P, Alonso S, Sanbhadti R, Barletta S, Pot D, Sheth M, Demchok JA, Shaw KR, Yang L, Eley G, Ferguson ML, Tarnuzzer RW, Zhang J, Dillon LA, Buetow K, Fielding P, Ozenberger BA, Guyer MS, Sofia HJ, Palchik JD: Comprehensive molecular portraits of human breast tumours. Nature 2012, 490:61–70.

28. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jonsson G, Mariani O, Boyault S, Miron P, Fatima A, Langer?d A, Aparicio SA, Tutt A, Sierwerts AM, Gorg A, Thomas G, Salomon AV, Richardson AL, B?resen-Dale AL, Futreal PA, Stratton MR, Campbell PJ: The life history of 21 breast cancers. Cell 2012, 149:994–1007.

29. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G: Sensitive detection of somatic point mutations in imput and heterogeneous cancer samples. Nat. Biotechnol. 2013, 31:213–219.

30. Eronen GD, Cai X, Lee E, Hills LB, Ellosary PC, Lehman HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A, Park PJ, Walsh CA: Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 2012, 151:483–496.

31. Siepel A, Bejerano G, Pedersen HS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spleth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research 2005, 15:1034–1050.

32. Berndt EK, Hall BH, Hall RE: Estimation and Inference in Nonlinear Structural Models. Tech. rep., National Bureau of Economic Research, Inc 1974.

33. Bazarra MS, Sherali HD, Shetty CM: Nonlinear Programming: Theory and Algorithms. John Wiley & Sons, Inc. 2005.

34. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 2001, 29:308–311.

35. McElroy KE, Luciani F, Thomas T: GemSIM: general, error-model based simulator of next-generation sequencing data. BMC genomics 2012, 13:74.

36. Li H, Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010, 26:589–595.

**Figure legends**

**Figure 1 - Overall Virmid workflow**

The complete Virmid workflow (a) Disease/control paired data are given (top) to generate an alignment (BAM) file. The mixed disease sample produce short reads of mixed types (blue and orange rectangles).

Somatic mutations, where the control has the reference genotype (AA) and the disease has non-reference
(AB or BB, red dots in alignment), are hard to detect in high contamination due to the significant drop in B-allele frequency (BAF). Virmid takes the disease/control paired data and analyzes 1) the sample level proportion of control cells in the disease sample ($\alpha$) and 2) the most probable disease genotype for each position that can be used to call somatic mutations. (b) An example BAF drop. Without contamination, the expected BAF is 0.5 and 1.0 for heterozygous and homozygous mutations sites respectively. When there is a control sample contamination of $\alpha$, mutation alleles are observed only in $(1 - \alpha)$ of the whole reads. So the expected BAF is dropped to $(1 - \alpha)/2$ and $(1 - \alpha)$. With estimating the accurate $\alpha$, Virmid can detect more true somatic mutations that can be missed in the conventional tools due to the insufficient observation of B alleles.

**Figure 2 - Estimation of contamination**

Estimation of contamination level in mixed disease sample. The proportion of control sample ($\alpha$) is estimated from the simulated mixed data. A total of 11 data sets with different $\alpha$ (1%, 5%, 10% 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%) were generated and tested. Virmid estimated all the $\alpha$ values (red line with circles) with high concordance compared to the true values (black line with squares). Note that there is a significant bias in highly contaminated samples ($\alpha \geq 60$) in the call based method (green line with circles) due to undetectable low BAF mutations; somatic mutations with higher BAF are likely to be called initially causing overestimation of BAF and underestimation of $\alpha$.

**Figure 3 - Maximum likelihood estimation and search**

(a) Log likelihoods over different $\hat{\alpha}$ (the estimate of $\alpha$) and $\hat{G}_{AA,AB}$ (the estimate of $G_{AA,BB}$, the probability that the control genotype is AA and the disease genotype is AB). The surface graph shows that likelihoods are maximized around the true $\alpha$ values (10%(left), 50%(middle), 90%(right)). (b) Searching paths using the feasible direction method is denoted on contour maps. The method efficiently finds the optimum points (green circles) only in a few searching steps. Searching processes from different starting points (blue circles) are finally converged.

**Figure 4 - Performance of somatic mutation detection**

Performance comparison of different methods for somatic mutation detection. (a) Precision-recall curves of Virmid with $\alpha$ (red), Virmid without $\alpha$ (light red), and JointSNVMix2 (blue) at six different $\alpha$ values (1%, 50%, 60%, 70%, 80% and 90%) are shown. Note that the performance is significantly improved when $\alpha$ is
incorporated into the calling model. There is little difference in performance at low contamination levels ($\alpha \leq 50$). (b) Precision and recall scores of final call are generated for each $\alpha$, where mutation probabilities are not available; note that a single point instead of a curve is plotted for each $\alpha$. As $\alpha$ grows, there is a consistent drop in precision, recall, and f-score*. Four tools including Virmid, Strelka, VarScan2 and JointSNVMix2 are tested with the same data. Virmid and VarScan2 are tested in two different modes (with and without $\alpha$). Strelka was also tested in two modes with or without applying quality control. Overall, Virmid with $\alpha$ scored best in f-score, followed by Strelka and VarScan2 with $\alpha$ and JointSNVMix2. Note that the tools with $\alpha$ (Virmid with $\alpha$, Strelka, VarScan2 with $\alpha$) outperformed those without $\alpha$ (Virmid without $\alpha$ and VarScan2 without $\alpha$), denoting that importance of incorporating $\alpha$ in SNV calling.

*F-score=$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

Figure 5 - BAF distribution of call sets

BAF distribution of different call-sets. Box plots are drawn for BAF in true (pink boxes) and called somatic mutations. From low to high contamination, the mean BAF decreases from 50% to 5%. Due to the difficulty in finding low BAF mutations, the call-sets show a slight to significant increase in BAF. Virmid with $\alpha$ represents the closest BAF to that of the true set. Due to the undetectable true somatic mutations that contain no B alleles, there can be a large gap between true and call set BAF distribution ($\alpha=80\%$ and 90%)

Figure 6 - Test on 15 breast cancer exome sequencing data.

(a) Sensitivity results of 4 different tools (5 different modes). The samples have been ordered by our estimated $\alpha$. (b-upper) Negative calls from Strelka (purple), MuTect (light blue) and JointSNVMix (orange). Each false negative resulted from low allele frequency or low read depth. (b-lower) Many of false negatives from Virmid (15/23) resulted from extremely low coverage in normal sample, which causes insufficient likelihood to call somatic mutation (LK, red dots). Remaining 8 false negatives are explained in filtering error. NM=number of mismatches, TRI=tri-allele, OFF=read offset, and PROX=proximity to indel (see Methods for more details of filtering). (c) Total number of calls. (d) Specificity result from “virtual tumor” analysis. The number of false positive calls were normalized by the total size of exome region to calculate number of false positives per million base-pair. Note that the y-axis is log-scaled.
Figure 7 - Estimated $\alpha$ in HME samples.

Analysis of five hemimegalencephaly (HME) samples. (a) Estimated $\alpha$ values are denoted in boxplot. (b) BAF distributions of HME call sets. Each point represents one predicted somatic mutation in the corresponding sample. BAF is calculated from predicted heterozygous (orange), homozygous (green) somatic mutations and loss of heterozygosity (LOH) sites. The mean BAF is consistent with the estimated $\alpha$ for every sample. The BAF calculated from read counts (red triangles) and mass spectrometry peak intensity (blue squares) of three previously validated mutations (MTOR, AKT3 and PI3KCA) are bounded in the predicted BAF ranges.

Tables

Table 1 - Accuracy and robustness of estimated $\alpha$

Virmid was ran 20 times for each sample to estimate $\alpha$ with variance. Generally, all the estimation was clustered very close to the true value. * Estimates without the loss of read (LOR) bias correction. The probability that a read with at least one A is mappable was set to equal the probability that a read with at least one B is mappable, that is, $x(A) = x(B)$ (see Methods). Note that loss of variants (LOV) bias was already addressed here.

| True $\alpha$ (%) | Estimated, $\alpha$ (%) | Stdev ($10^{-2}$) | $\alpha$ range | BAF range |
|-------------------|-------------------------|-------------------|----------------|-----------|
| Call-based | Virmid | Virmid | | |
| 1 | 2.64 | 2.56 | 1.61 | 0.19 | 1.23~1.90 | 11.11~47.50 |
| 5 | 6.31 | 5.56 | 4.74 | 0.28 | 4.07~5.17 | 11.11~46.34 |
| 10 | 10.3 | 10.50 | 9.86 | 0.51 | 9.34~11.03 | 11.43~44.12 |
| 20 | 20.4 | 19.92 | 19.59 | 0.44 | 18.44~20.13 | 11.43~38.46 |
| 30 | 30.4 | 30.33 | 30.28 | 0.48 | 28.48~30.79 | 11.11~33.33 |
| 40 | 39.6 | 40.46 | 39.94 | 0.22 | 39.49~40.51 | 11.43~28.89 |
| 50 | 49.2 | 51.38 | 50.72 | 0.23 | 50.46~51.31 | 11.11~23.53 |
| 60 | 56.2 | 61.16 | 60.62 | 0.54 | 59.78~61.48 | 10.71~19.15 |
| 70 | 62.4 | 70.28 | 70.05 | 0.16 | 69.82~70.44 | 10.00~14.29 |
| 80 | 67.2 | 80.33 | 80.04 | 0.38 | 79.52~80.76 | 9.38~10.81 |
| 90 | 67.4 | 88.91 | 88.88 | 0.53 | 88.06~90.00 | 8.82~9.68 |

Table 2 - Improved mutation calling in higher coverage

Precision and recall values are compared between two different coverages (40x and 100x) at highly contaminated data ($\alpha$=70%, 80%, and 90%). Much deeper coverage is required when severe contamination is expected. $\alpha$=proportion of control sample in disease sample (contamination level), n.Answer=number of true somatic mutations, n.Predict=number of predicted mutations by Virmid, n.Correct=number of correct mutations from prediction, Prc=Precision, Rec=Recall.* rounded up from 0.9954, b rounded up from 0.9958

| $\alpha$ | n.Answer | Coverage=40x | Coverage=100x | F-score |
|----------|----------|--------------|----------------|---------|
|          | n.Predict | Precision | Recall | F-score | n.Predict | Precision | Recall | F-score |
| 70%      | 1999     | 1976        | 0.98  | 0.87  | 0.93    | 2208  | 2198   | 1.00*  | 0.97  | 0.98  |
| 80%      | 2265     | 1551        | 1516  | 0.98  | 0.67  | 0.79   | 2142  | 2133   | 1.00b  | 0.94  | 0.97  |
| 90%      | 638      | 611         | 0.96  | 0.27  | 0.42   | 1572  | 1545   | 0.98  | 0.68  | 0.81  |

Table 3 - Newly predicted somatic mutations in HME data

List of predicted somatic mutations in five HME individual samples are shown. Predicted mutations sets are compared to previous study [14] using JointSNVMix2. Virmid called more novel (and less overlapping) mutations in samples with higher $\alpha$ (HME-1574 and HME-1620). Est.$\alpha$=Estimated contamination level ($\alpha$), std.=standard deviation, n.Total.Mut=number of all called mutations, n.Mut=number of (overlapping or novel) mutations, n.Exon=number of mutations in exon, n.Mis=number
of missense (or nonsense) mutations. * one mutation loci may be counted multiple times when more than one genetic products exist at the position.

| Subject     | Est. α (std.) | n.Total.Mut | n.Mut | n.Exon | n.Mis | Gene  | n.Mut | n.Exon | n.Mis | Gene |
|-------------|---------------|-------------|-------|--------|-------|-------|-------|--------|-------|------|
| HME-1563    | 77.9% (±0.006) | 478         | 108   | 54     | 33    | MTOR  | 370   | 209    | 147   |
| HME-1565    | 71.3% (±0.003) | 494         | 124   | 78     | 50    | AKT3  | 370   | 72     | 43    |
| HME-1573    | 64.0% (±0.017) | 542         | 235   | 167    | 112   | PIK3CA | 307   | 301    | 185   |
| HME-1574    | 83.4% (±0.005) | 579         | 100   | 63     | 34    | PIK3CA | 479   | 412    | 281   | MTOR |
| HME-1620    | 84.8% (±0.004) | 694         | 86    | 56     | 27    |       | 608   | 383    | 267   |

Additional Files
Additional file 1 — Supplementary information
A supplementary PDF file that includes supplementary figures (Figure S1 to S4), table (Table S1 and S2) and methods (Supplementary note 1 to 5 including mathematical proofs, derivations and model descriptions).
Figure 1

A) Control sample
   Mixed disease sample
   Sequencing / Mapping / Post-alignment process:
   (IndelRealigner, MarkDuplicates, TableRecalibration)
   BAF
   α proportion of control sample
   reference alignment
   reads from control sample
   reads from disease sample
   Sampling
   α estimator
   Genotype probability estimator
   Genotype caller

1. Proportion of control sample (α)
2. Most probable disease genotype & somatic mutation probability

|                | Expected value |
|----------------|----------------|
|                | heterozygous   | homozygous    |
| raw frequency: | (1 - α)/2      | 1 - α         |
| corrected frequency: | 0.5 | 1.0 |
Figure 2
Figure 3
Figure 4

A

\( \alpha = 1\% \)

\( \alpha = 50\% \)

\( \alpha = 60\% \)

\( \alpha = 70\% \)

\( \alpha = 80\% \)

\( \alpha = 90\% \)

B

Precision vs. Recall

True \( \alpha \) (%)

Precision vs. Recall

True \( \alpha \) (%)

Precision vs. Score

True \( \alpha \) (%)
Figure 6
Figure 7
Additional files provided with this submission:

Additional file 1: virmidsupplantary.pdf, 1078K
http://genomebiology.com/imedia/1559554893105465/supp1.pdf