


METHOD

Open Access



MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data

Yu Fan¹, Liu Xi², Daniel S. T. Hughes², Jianjun Zhang³, Jianhua Zhang³, P. Andrew Futreal³, David A. Wheeler² and Wenyi Wang^{1*} 

Abstract

Subclonal mutations reveal important features of the genetic architecture of tumors. However, accurate detection of mutations in genetically heterogeneous tumor cell populations using next-generation sequencing remains challenging. We develop MuSE (<http://bioinformatics.mdanderson.org/main/MuSE>), Mutation calling using a Markov Substitution model for Evolution, a novel approach for modeling the evolution of the allelic composition of the tumor and normal tissue at each reference base. MuSE adopts a sample-specific error model that reflects the underlying tumor heterogeneity to greatly improve the overall accuracy. We demonstrate the accuracy of MuSE in calling subclonal mutations in the context of large-scale tumor sequencing projects using whole exome and whole genome sequencing.

Keywords: Somatic mutation calling, Sensitivity and specificity, Bayesian inference, Model-based cutoff finding, Next-generation sequencing

Background

The detection of somatic point mutations is a key component of cancer genomic research that has been rapidly developing since next-generation sequencing (NGS) technology revealed its potential for describing genetic alterations in cancer [1–6]. As the cost of NGS has decreased, the need to thoroughly interrogate the cancer genome has spurred the migration from using whole exome sequencing (WES) to whole genome sequencing (WGS). A critical challenge accompanying this migration is the rigorous requirement of specificity, considering that a false positive rate (FPR) of even 1 per megabase pair (Mbp) results in 3000 incorrect variant calls for WGS data. In addition, the sequencing depth decreases from 100 – 200× for WES

data to 30 – 60× for WGS data, resulting in a lower signal-to-noise ratio and making accurate mutation calling more difficult.

Another nontrivial difficulty is accounting for the influence of tumor heterogeneity that is commonly observed in mutation calling. The presence of both normal cells and tumor subclones in the sample causes this phenomenon to vary from sample to sample [7, 8]. It is thus important to identify sample-specific cutoffs dynamically and report tier-based variant call sets instead of using a fixed cutoff for all samples, which is current common practice. On the other hand, tier-based variant call sets that inherently attach uncertainties will be helpful when evaluating the behavior of low variant allele fraction (VAF) mutations and seeking to understand the effect of tumor heterogeneity.

Here, we present a novel and automatic approach to discovering somatic mutations, *Mutation calling using a Markov Substitution model for Evolution* (MuSE), which

*Correspondence: wwang7@mdanderson.org

¹Department of Bioinformatics and Computational Biology - Unit 1410, The University of Texas MD Anderson Cancer Center, P. O. Box 301402, 77230-1402 Houston, TX, USA

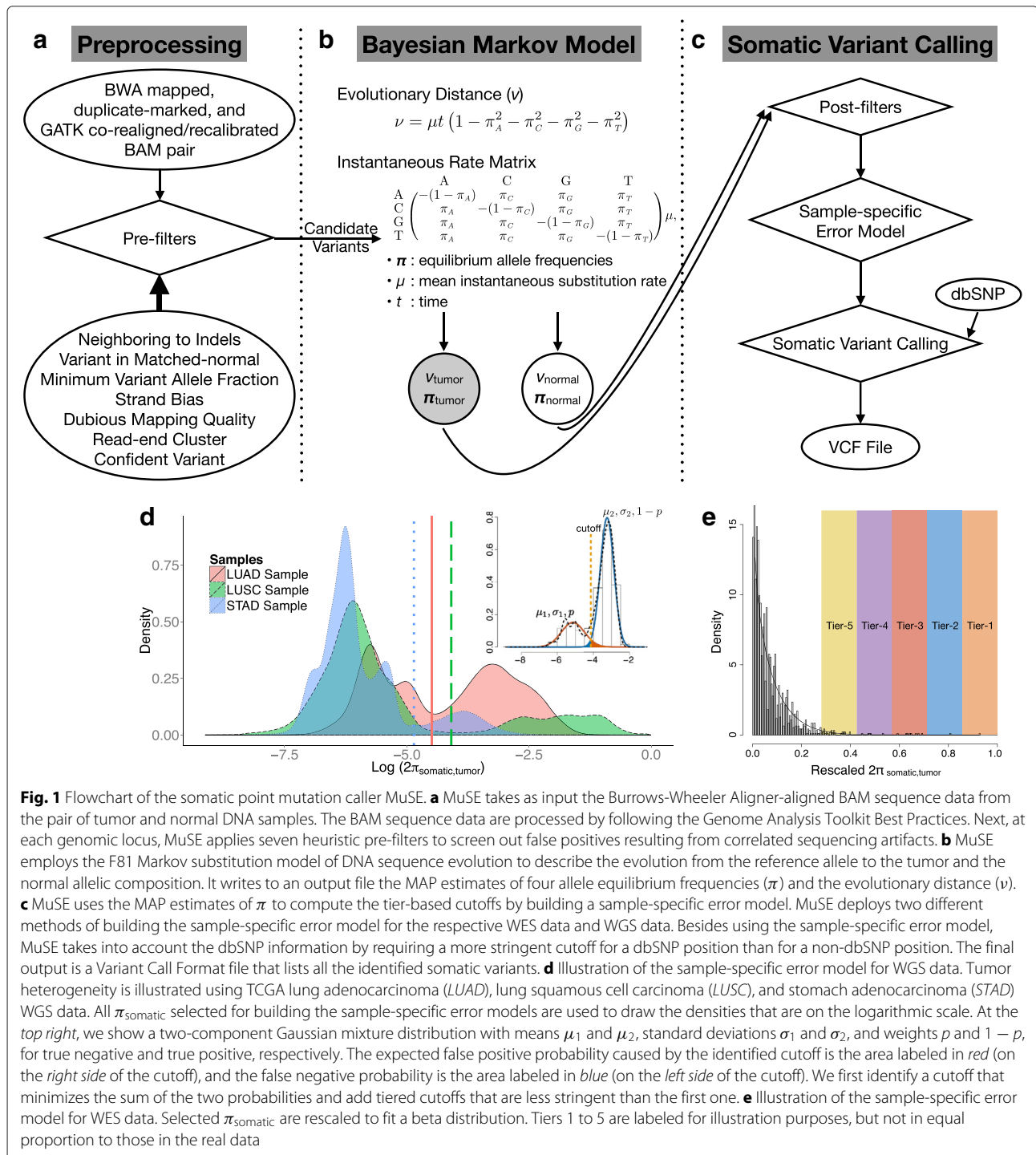
Full list of author information is available at the end of the article

models the evolution of the reference allele to the allelic composition of the tumor and normal tissue at each genomic locus. We further adopt a sample-specific error model to identify cutoffs, reflecting the variation in tumor heterogeneity among samples. We demonstrate the reliable performance of MuSE, a good balance of sensitivity and specificity, with various types of data.

Results and discussion

MuSE design

MuSE comprises two steps (Fig. 1). The first step, ‘MuSE call’ (Fig. 1a, b), takes as input the binary sequence alignment map (BAM) formatted sequence data that require special preparation from the pair of tumor and normal DNA samples. The results of our investigation



avored the co-local realignment of tumor and matched-normal BAMs rather than the local realignment of tumor and matched-normal BAMs separately (data not shown). MuSE carries out pre-filtering on every genomic locus, which is a common practice (e.g., see [5]) ahead of variant detection in order to accelerate the computing speed and remove potential false positives. Next, MuSE accomplishes variant detection by employing the F81 Markov substitution model [9], which provides the estimates of equilibrium frequencies for all four alleles ($\pi_A, \pi_C, \pi_G, \pi_T$), and the evolutionary distance (ν). In practice, we report the maximum a posteriori (MAP) estimates of π and ν instead of exploring the full posterior distribution.

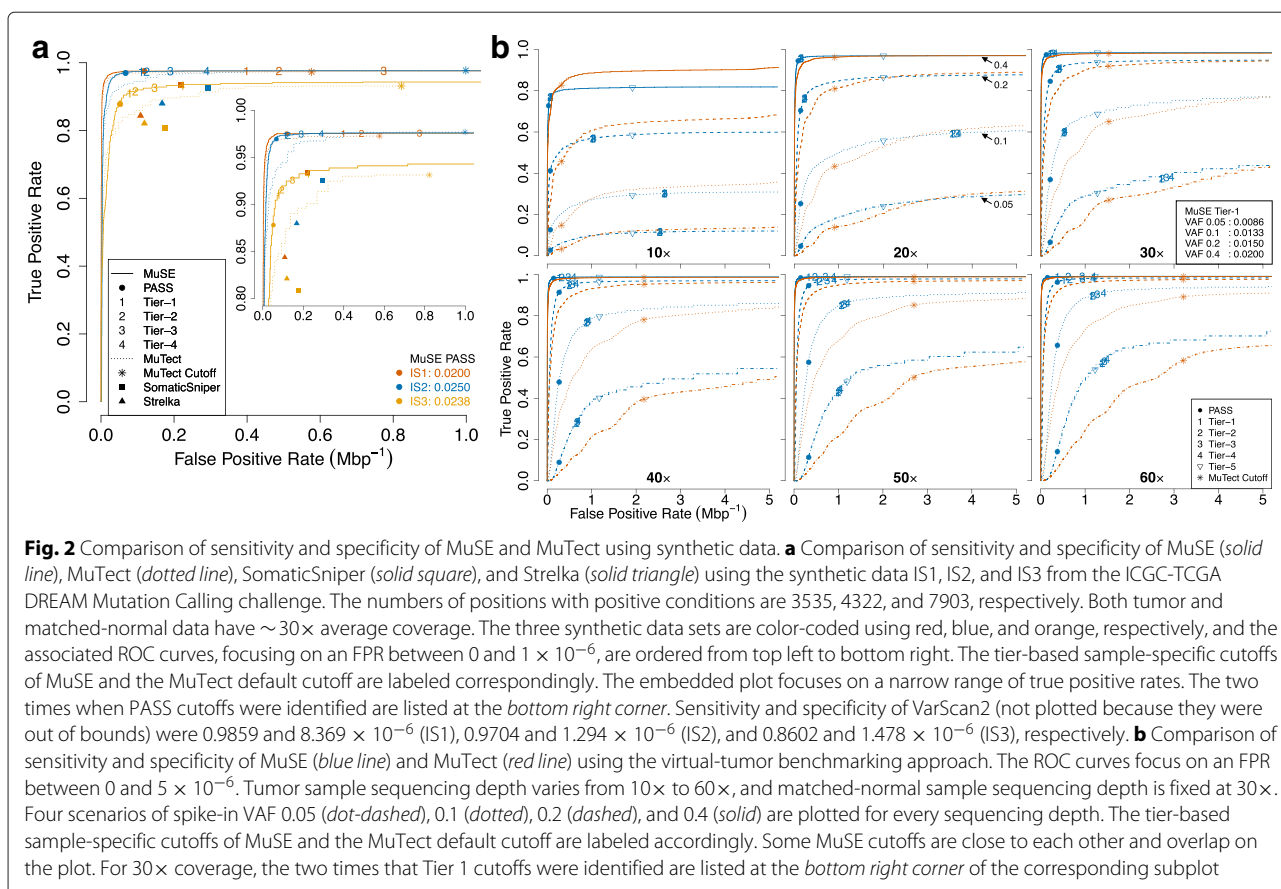
The second step, 'MuSE sump' (Fig. 1c), takes as input the post-filtered $\pi_{\text{somatic,tumor}}$ and computes tier-based cutoffs from a sample-specific error model. As a unique feature of MuSE, the tier-based cutoffs (PASS, Tiers 1 to 5) address the large variations observed in the distributions of $\pi_{\text{somatic,tumor}}$ across tumor samples (Fig. 1d). With WGS data, we fit a two-component Gaussian mixture model to $\log(2\pi_{\text{somatic,tumor}})$ across positions in order to separate the two major modes, one from true mutations that represent tumor growth dynamics that vary largely across samples and one from reference positions with variations in noise arising from sequencing machine errors or mapping errors. However, the degree of difference between the true negative (reference) and true positive (somatic mutation) positions and whether it is detectable depend on the sequencing depth and the VAFs of the mutations. When true mutations are of low VAF, presenting a distribution that largely overlaps with that of the true negative positions, we use a cutoff of 0.005 as our lowest boundary (Tier 5) to control the number of false positives. When the number of true positives is relatively minimal compared to that of true negatives, as in most WES data (mutation rate up to 10/Mbp; [10]), we model $\pi_{\text{somatic,tumor}}$ as a beta distribution (Fig. 1e) and call mutations as the extreme and rare events on the right tail of the fitted distribution. We take into account the dbSNP information by requiring a cutoff that is two times more stringent for a dbSNP position than for a non-dbSNP position. The final output of the second step is a Variant Call Format (VCF) file that lists the identified somatic variants.

Synthetic data

We measured the performance of MuSE using synthetic data and compared the sensitivity and specificity of MuSE with that of other state-of-the-art callers [1, 4, 5, 11]. MuSE is intended to run with little or no human curation. For that reason, all callers were evaluated without human curation to yield a uniform comparison, although in practice, output from mutation callers is often curated. We first made the comparison using the synthetic data IS1, IS2, and IS3 (9.11 gigabase pairs (Gbp)) from the

ICGC-TCGA DREAM Mutation Calling challenge [6]. The complexity of the three data sets increased because of elevating mutation rates, declining VAFs, and incorporating multiple subclones. This increased data complexity affected the performance of all callers, which was evident in the synchronized decreases in sensitivity (Fig. 2a). In all three data sets, MuSE was more sensitive and specific than MuTect, SomaticSniper, Strelka, and VarScan2. Moreover, MuSE identified cutoffs varying by the sample (Fig. 2a, bottom right). These cutoffs at the PASS level are located at the top left corners of the receiver operating characteristic (ROC) curves, which suggests an ideal balance between sensitivity and specificity. Since IS1 was the least complex and furthest away from real data, additional tiers were not able to improve the sensitivity.

Furthermore, using the virtual-tumor benchmarking approach [5], we studied the impact of sequencing depth ($10\times$ to $60\times$) and VAFs (0.05, 0.1, 0.2, and 0.4) on MuSE and MuTect in whole genomes (18.2 Gbp; Fig. 2b, Additional file 1: Table S1). From moderate ($30\times$) to high ($60\times$) coverage, the MuSE curves stayed on top of the MuTect curves. At low ($10\times$ and $20\times$) coverage, the two curves crossed as FPR increased. These two low coverage data sets had low signal-to-noise ratios and were most sensitive to losing true positives from post-filtering. Nevertheless, for the segment of the curve that contained the MuTect default cutoff, the MuSE curve was still on top of its counterpart, except for one scenario, $10\times$ and VAF of 0.4. The incremental changes in calling accuracy from Tier 1 to Tier 4 were more evident in scenarios with high VAFs than in those with low VAFs. Different from the DREAM challenge data, in this data set, Tier 1 cutoffs showed the biggest improvement in sensitivity compared to one level up, PASS, and moved closer to the top left corners of the ROC curves in all simulation scenarios from $30\times$ to $60\times$ coverage, except for $30\times$ and VAF of 0.05. For different VAF spike-in scenarios, again, MuSE identified Tier 1 cutoffs that were distinct from each other (Fig. 2b, subplot on $30\times$). At low ($10\times$ and $20\times$) coverage, PASS performed reasonably well. MuSE could not identify a cutoff comparable to the MuTect default cutoff for $20\times$ coverage, VAF of 0.1 and 0.05. Tier 5 was helpful in improving sensitivity while maintaining a low FPR at low coverage and low VAF. Looking across the data sets with varied coverage but fixed low VAF (VAF = 0.05), we observe that MuSE achieved higher sensitivity than MuTect at the same level of specificity. Therefore, MuSE will be helpful for calling subclonal mutations in studies of the heterogeneity and subclonal evolution of tumors. Although MuSE demonstrated better accuracy than MuTect using the virtual-tumor benchmarking data, the two callers generated intersecting sets (Additional file 1: Figure S1), which provides a conspicuous demonstration of the importance and necessity of using multiple callers in somatic variant detection.

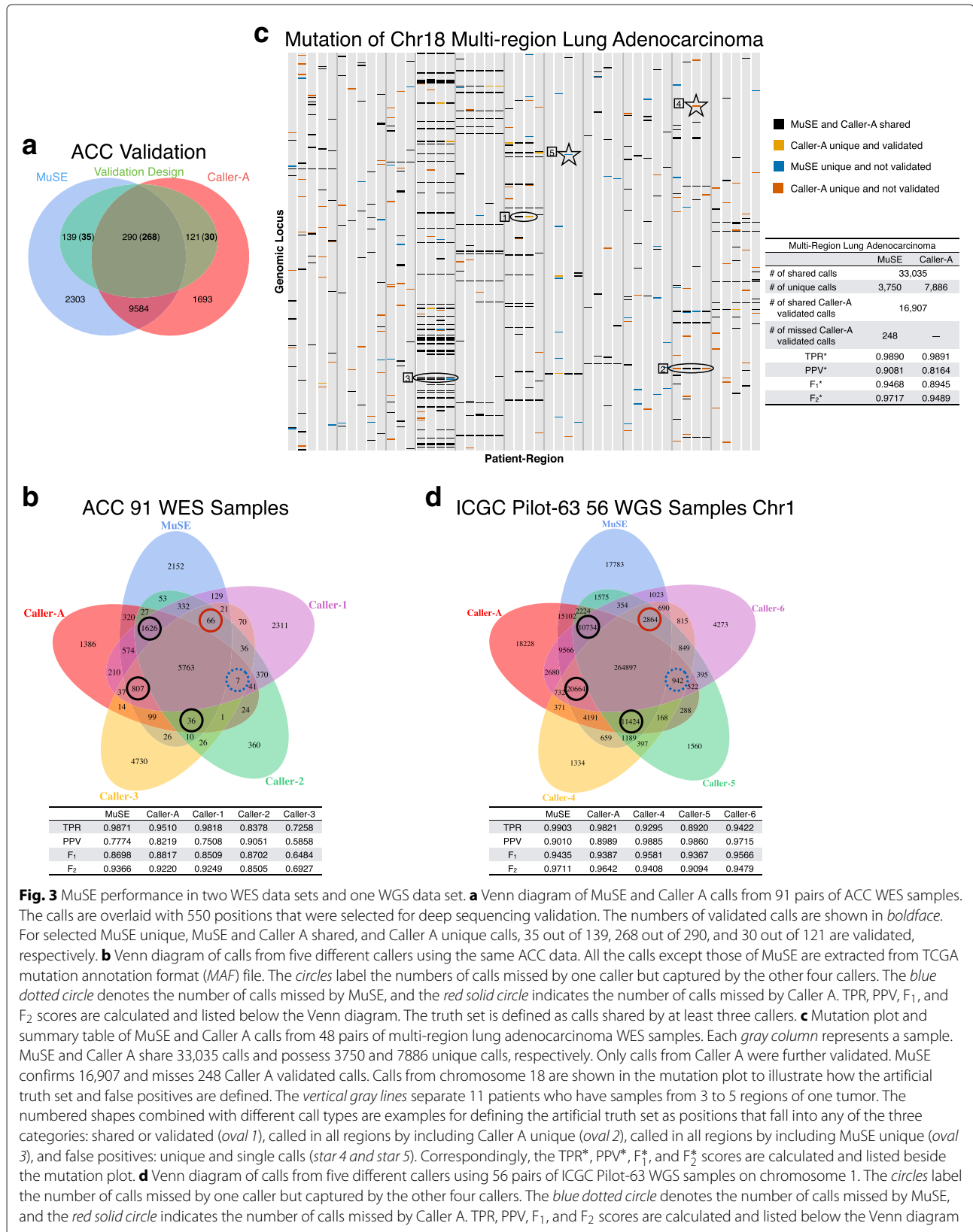


Real data

We evaluated MuSE using multiple real WES and WGS data sets and compared MuSE with other calling pipelines (anonymous). Specifically we focused on comparing with Caller A, which is one of the best-in-breed mutation callers based on the ICGC-TCGA DREAM Mutation Calling challenge. With TCGA and ICGC samples, we used calls that were prepared and provided by the corresponding institutes where individual calling pipelines were run. We first tested the performance of MuSE using data from 91 tumor-normal paired WES samples (3.21 Gbp) from patients with adrenocortical carcinoma (ACC; [12]) (Fig. 3a). Taking into account the tier-based distribution of MuSE calls (Additional file 1: Table S2), we computed the validation rates of MuSE total calls and unique calls, and obtained 84.50 % and 26.34 %, respectively. We repeated a similar calculation for Caller A, which gave the respective validation rates of total calls and unique calls as 87.39 % and 24.79 %. Considering that the validation rate could not measure sensitivity, we extracted the multi-center somatic variant calls from the TCGA mutation annotation format (MAF) file, made an artificial truth set by taking calls that were shared by at least three callers, and computed a sensitivity of 98.71 % for MuSE and a sensitivity of 95.10 % for Caller A (Fig. 3b).

Moreover, MuSE missed only 7 calls that were captured by the other four callers, compared with 66, 36, 807, and 1626 missed calls from Caller A, Caller 1, Caller 2, and Caller 3, respectively. As an alternative to the deep sequencing validation on a small set of positions, we regarded all calls outside of the artificial truth set as false positives to calculate positive predictive values (PPVs). In agreement with previous findings of the validation rates, Caller A benefited from its low number of unique calls and obtained the second best PPV, which in turn helped Caller A acquire a better F_1 score [13]. However, using the F_2 score, which placed a relatively higher weight on sensitivity, we demonstrated the good performance of MuSE ($F_2 = 0.9366$). When we used more stringent tiers, we obtained a smaller number of MuSE unique calls, changing from 2152 to 378, without losing much sensitivity; i.e., the number of missed calls that were shared by the other four callers increased from 7 to 14 (Additional file 1: Figure S2).

We then applied MuSE to WES data from 48 multi-region tumor-normal paired samples (2.46 Gbp) from 11 patients with lung adenocarcinoma, which provided 17,155 deep sequencing validated calls that were originally selected from all calls made by Caller A [14]. MuSE confirmed 16,907 and missed 248 Caller A validated calls,



a sensitivity of 98.55 %, given 3750 unique calls compared with 7886 unique calls from Caller A. In contrast to the ACC data, this validated data set could not provide unbiased evaluation of the two callers. However, the multi-region design of this data set was unique. We therefore built our artificial truth set by taking all validated calls (Fig. 3c; orange in oval 1), all shared calls (Fig. 3c; black in oval 1), and all trunk mutation calls that occurred at the same genomic locus in all tumor regions of one patient (Fig. 3c; ovals 2 and 3). This design allowed us to consider unique and unvalidated calls from each caller as true positives when they appeared as trunk mutations (Fig. 3c; red in oval 2 and blue in oval 3). We regarded all other calls that were subclonal as false positives (Fig. 3c; red in five-pointed star 4 and blue in five-pointed star 5). The F_1^* (0.9468) and F_2^* (0.9717) scores acquired by MuSE were higher than those of Caller A.

We further compared MuSE with other callers using 56 pairs of ICGC Pilot-63 WGS samples on chromosome 1 (14.0 Gbp; [15]). We downloaded the related somatic VCF files that were generated by multiple callers from the ICGC Pilot-63 study. In accordance with the ACC multi-caller result, MuSE missed 942 calls that were captured by the other four callers, which was the least number of missed calls and therefore indicated the highest sensitivity among all five callers (Fig. 3d). Caller 4 and Caller 6 gave the best and the second best F_1 scores due to their high PPVs (Fig. 3d). Caller 5, which had low sensitivity, could not achieve a better F_1 score, although its PPV was higher than that of Caller 6. The F_1 score of MuSE was higher than those of Caller A and Caller 5, but could not compete with those of Caller 4 and Caller 6. However, considering that Caller 4, Caller 5, and Caller 6 respectively missed 10,734, 20,664, and 11,424 calls that were shared by the other four callers, the loss of sensitivity as a tradeoff for greater specificity may raise concerns. Among all five callers, MuSE had the best F_2 score, emphasizing the importance of sensitivity.

Conclusions

In summary, we present a somatic point mutation caller, MuSE. We design MuSE as an automatic approach with two steps. The first step, ‘MuSE call’, implements the heuristic pre-filters and uses the Markov substitution model to describe the evolution of the reference allele to the allelic composition of the matched tumor and normal tissue at each genomic locus, which provides the summary statistics π_{somatic} . The $\pi_{\text{somatic,tumor}}$ associated ROC curve is shown to stand above that from Caller A, suggesting a good ability to discriminate mutations from references of the MuSE pipeline. The second step, ‘MuSE sump’, identifies tier-based cutoffs on $\pi_{\text{somatic,tumor}}$. We build a sample-specific error model to account for tumor heterogeneity

and to identify cutoffs that are unique to each sample, achieving high accuracy in mutation calling. With the two steps, we aim at mitigating users’ curation of output. We provide five tiers. From experience, we suggest using calls up to Tier 4 for WES data, and calls up to Tier 5 for WGS data. These suggested cutoffs are derived based on our observation of real data and serve the goal of maximizing sensitivity and maintaining a good specificity. Typically, the ‘MuSE call’ step takes ~ 4 hours to process a tumor-normal paired WGS sample with $30 - 60\times$ coverage when the WGS data is divided into ~ 50 equal-sized blocks and each block is assigned with 1 CPU core and 2 GB memory, and the ‘MuSE sump’ step requires ~ 1 hour for WGS data given 1 CPU core and 4 GB memory.

We demonstrate the reliable performance of MuSE using both synthetic and real data, such as the ICGC-TCGA DREAM Mutation Calling challenge WGS data, the virtual-tumor benchmarking approach, TCGA ACC WES data, the multi-region lung adenocarcinoma WES data, and the ICGC PanCancer Pilot-63 WGS data. We demonstrate the superior sensitivity of MuSE, especially to low VAF mutations, and its capacity to identify an appropriate balance of sensitivity and specificity in each sample with varying levels of heterogeneity. This feature is essential for downstream analyses, such as finding tumor subclonal structures and understanding the evolution of tumors, a broad interest in the cancer community and beyond. So far, we have found substantially more subclones using MuSE calls (up to Tier 5) than using calls from other callers in ICGC PanCancer Analysis of Whole Genomes (data not shown; [16]).

Copy number aberration (CNA), tumor purity, and tumor subclonality commonly exist in our data, both synthetic and real. All influences of CNA, tumor purity, and tumor subclonality on the mutant chromosome content of a tumor reduce to the same question of VAF, and the mechanism of creating or changing the VAF is not as important as the VAF itself in terms of somatic mutation calling. Therefore, we use the F81 Markov substitution model to capture the VAF dynamics at each locus. Our $\pi_{\text{somatic,tumor}}$ is directly related with the configuration of local copy number variation, purity, and subclonality of the position. Our two-component Gaussian mixture model was motivated when we tested the performance of MuSE using the virtual-tumor benchmarking approach (Additional file 1: Figure S3). Therefore, we aim to deconvolute two $\log(2\pi_{\text{somatic,tumor}})$ distributions, one from true mutations that represent tumor growth dynamics and one from reference positions that arise from sequencing machine errors or mapping errors. When there are multiple peaks in each distribution, as often observed in real data (Fig. 1d), our assumption that true mutations from multiple subclones and reference positions from machine

or mapping errors can be separated by finding two major modes is supported by the high sensitivity and specificity of MuSE calls in the validation data. We chose the Gaussian mixture model because of its robustness to model assumption and easy implementation with a closed-form likelihood function. However, alternative distributions, for example the gamma mixture distribution, may also be appropriate due to the fact that $\log(2\pi_{\text{somatic,tumor}})$ is bounded by 0.

We considered two aspects when using the F81 model: (1) the number of free parameters in the model should remain small to allow for higher accuracy in estimation for each position; and (2) the F81 model can be extended to take into account mutational contexts, which will be our future work. One potential benefit of considering the mutational contexts is to further reduce false positives. We accessed MuSE calls in annotated CpG islands (UCSC Genome Browser CpG Island Annotation Track) using the TCGA WES data from ACC. The validation rate of MuSE total calls decreased from 0.8450 to 0.7245, and the validation rate of MuSE calls shared with Caller A decreased from 0.9889 to 0.8829.

We will further validate MuSE through participation in the ICGC-TCGA DREAM Mutation Calling challenge and the ICGC Pilot-63 study, both of which have promised independent experimental validations. We have also applied MuSE to analyze the WES data of chromophobe renal cell carcinoma (KICH; [17]) and liver hepatocellular carcinoma (LIHC), which are part of the TCGA project. The corresponding calls have been made available to the TCGA community. MuSE is being used by two new ongoing consortium projects: TCGA PanCanAtlas and ICGC PanCancer Analysis of Whole Genomes, which includes WGS data from more than 2800 pairs of tumor and matched-normal samples.

Despite the satisfactory performance of MuSE, we contend that there is no comprehensive caller that can replace all the others; each caller has strengths and unique attributes. We support the trend to incorporate call sets from multiple callers in future NGS analyses, for example, using SomaticSeq [18]. Due to its ensemble nature, SomaticSeq relies on the performance of its callers, and is bounded by the best sensitivity among individual callers [18]. Therefore, when MuSE is included as one of the callers to be integrated, we expect SomaticSeq to generate results that are more accurate than it can produce currently. We welcome the usage of other post-filtering methods on MuSE calls, for instance, panel of normal samples, when data from the appropriate control samples are available. Our method can be extended for calling binucleotide, triplet, or small insertion-deletion variants by modifying the F81 Markov substitution model.

Methods

BAM preparation

All the sequence reads were aligned against the hg19 reference genome using the Burrows-Wheeler Aligner (BWA) with either the backtrack or the maximal exact matches (MEM) algorithm [19]. In addition, data sets (3), (4), and (5) were processed by following the Genome Analysis Toolkit (GATK) Best Practices [20–22] that include marking duplicates, realigning the paired tumor-normal BAMs jointly, and recalibrating base quality scores.

Variant heuristic pre-filters

In order to detect context-based sequencing artifacts, remove potential false positives, and accelerate the computing speed, we apply heuristic pre-filters to every genomic locus in advance of variant detection.

- (1) *Neighboring to indels*: No less than 3 insertions or 3 deletions are observed in an 11-base window centered on the locus.
- (2) *Variant in matched-normal*: The candidate variant allele is observed no less than twice or its variant allele fraction is no less than 3 % in the matched-normal data; moreover, the sum of the variant allele's base quality scores is more than 20. However, this genomic locus is kept if the candidate variant allele turns out to be the germ-line variant in the matched-normal data and the second variant allele is rejected by the above test.
- (3) *Minimum variant allele fraction*: The candidate variant allele fraction in the tumor data is smaller than 0.005.
- (4) *Strand bias*: The p value that is computed from Fisher's exact test using tumor allele count data comparing sense and antisense strands is less than or equal to $1e-5$.
- (5) *Dubious mapping quality*: The average mapping quality score of reads that carry a candidate variant allele is less than or equal to 10.
- (6) *Read-end cluster*: For each read that has the candidate variant allele, we record the smallest distance there can be from the current genomic locus to either the left end or the right end of the read alignment. We disregard the current genomic locus if the median of all these distances is less than or equal to 10 and the median absolute deviation is less than or equal to 3.
- (7) *Confident variant*: We require there to be at least one variant read that meets the following criteria: (a) the read and its mate are mapped in a proper pair; (b) its mapping quality score is no less than 30; and (c) the base quality score of its candidate variant allele is greater than or equal to 25.

Variant detection

For each genomic locus, we denote the base of read r ($r = 1 \dots N$) that covers the locus as b_r , where $r \in \{1 \dots N\}$ and N is the depth of the locus. By knowing the associated Phred quality score q_r of b_r , we denote the probability of b_r being the four different alleles (A, C, G, T) as

$$L(x) = \begin{cases} 1 - 10^{-\frac{q_r}{10}} & \text{if } b_r = x, \\ \frac{1}{3} 10^{-\frac{q_r}{10}} & \text{if } b_r \neq x, \end{cases}$$

where $x \in \{A, C, G, T\}$. We use a continuous-time Markov chain to describe the DNA evolution from the reference allele R to the allelic composition $\mathbf{b} = (b_r : r \in \{1 \dots N\})$ at each locus, namely, the F81 Markov substitution model [9]. The F81 model can be expressed using a 4-state \times 4-state instantaneous rate matrix, \mathbf{Q} :

$$\mathbf{Q} = \begin{matrix} & \begin{matrix} A & C & G & T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} -(1 - \pi_A) & \pi_C & \pi_G & \pi_T \\ \pi_A & -(1 - \pi_C) & \pi_G & \pi_T \\ \pi_A & \pi_C & -(1 - \pi_G) & \pi_T \\ \pi_A & \pi_C & \pi_G & -(1 - \pi_T) \end{pmatrix} \end{matrix} \mu,$$

where each entry represents the changing rate from allele i to allele j in an infinitesimal time dt , μ stands for the mean instantaneous substitution rate, and $\pi_A, \pi_C, \pi_G, \pi_T$ are the equilibrium allele frequencies. The transition matrix that consists of the probabilities of change between any two states in time t can be calculated from the exponential of the instantaneous rate matrix, $\mathbf{P}(t) = e^{\mathbf{Q}t}$. Specifically,

$$P_{ij}(t) = \begin{cases} \pi_j + (1 - \pi_j) e^{-\mu t} & \text{if } i = j, \\ \pi_j (1 - e^{-\mu t}) & \text{if } i \neq j. \end{cases}$$

Because of the confounding nature of the μt product, it is customary to rescale the instantaneous rate matrix so that the mean substitution rate at equilibrium is 1, and replace t with the evolutionary distance ν that represents the expected number of substitutions per base. Consequently, the transition matrix of the F81 model is altered as

$$P_{ij}(\nu) = \begin{cases} \pi_j + (1 - \pi_j) e^{\{-\nu/(1-\pi_A^2-\pi_C^2-\pi_G^2-\pi_T^2)\}} & \text{if } i = j, \\ \pi_j \left(1 - e^{\{-\nu/(1-\pi_A^2-\pi_C^2-\pi_G^2-\pi_T^2)\}}\right) & \text{if } i \neq j, \end{cases}$$

and the likelihood function $f(\mathbf{b}, R | \boldsymbol{\pi}, \nu)$ can be expressed as

$$f(\mathbf{b}, R | \boldsymbol{\pi}, \nu) = \prod_{r=1}^N \left\{ \sum_{x_m} \pi_{x_m} \left[\sum_{x_h} P_{x_m, x_h} \left(\frac{\nu}{2} \right) L_h^{(r)}(x_h) \right] \times \left[\sum_{x_k} P_{x_m, x_k} \left(\frac{\nu}{2} \right) L_k(x_k) \right] \right\},$$

where (1) $x_m, x_h, x_k \in \{A, C, G, T\}$; (2) ν connects the reference allele R and the allelic composition \mathbf{b} ; (3) h and k denote the \mathbf{b} and R tips of ν , respectively; (4) m denotes

the middle point of h and k so that the evolutionary distance ν from m to h is equal to the distance from m to k , i.e., $\nu/2$. Because of the time-reversible characteristic of the F81 model, m can be any point along the evolutionary distance ν that connects the h and k tips without affecting the final result. We set m as the midpoint for the purpose of calculation convenience; (5) $L_k(x_k) = 1$ if $x_k = R$, and $L_k(x_k) = 0$ otherwise; and (6) all the other notation is the same as that used above.

We obtain the joint posterior probabilities of $\boldsymbol{\pi}$ and ν , $f(\boldsymbol{\pi}, \nu | \mathbf{b}, R)$, by setting the priors of $\boldsymbol{\pi}$ and ν to be the Dirichlet distribution $\text{Dir}(1,1,1,1)$ and the exponential distribution $\text{Exp}(1000)$, respectively. In practice, we employ the Broyden-Fletcher-Goldfarb-Shanno algorithm and Brent's algorithm to search for the maximum a posteriori (MAP) estimates of $\boldsymbol{\pi}$ and ν instead of exploring the full posterior distribution.

We apply the above method to both loci of the tumor-normal paired sequencing data and obtain the $\pi_{\text{somatic, tumor}}$ and the $\pi_{\text{somatic, normal}}$ estimates accordingly. We designate the non-reference and non-germline allele that has the largest π as the somatic variant allele. The somatic variant allele should pass all the pre-filtering examinations.

Post-filtering criteria

After we obtain the $\pi_{\text{somatic, tumor}}$ and $\pi_{\text{somatic, normal}}$, we require that: (1) the minimum coverage of tumor and matched-normal data is 8 at given genomic loci; and (2) the ratio $\frac{\pi_{\text{somatic, normal}}}{\pi_{\text{somatic, tumor}}}$ is less than or equal to 0.05, which tolerates the contamination of matched-normal data with tumor data in a reasonable amount and dynamically changes the constraint on matched-normal data.

Sample-specific error model

We provide two options for building the sample-specific error model. One is applicable to WES data, and the other to WGS data. By plotting the densities of $\log(2\pi_{\text{somatic, tumor}})$ from MuSE on all positions (see Additional file 1: Figure S3), we observed that (1) the density of log-transformed $\pi_{\text{somatic, tumor}}$ showed a bimodal behavior that could be approximated using a Gaussian mixture distribution; (2) the true positives (red) and reference positions (blue) correspond to each of the modes so that a cutoff can be identified to separate the two types of calls; (3) as expected, the separation of two modes becomes easier at higher coverage and higher variant allele fraction (VAF). For most WES data, there are not enough true mutations that can form a detectable second mode as compared to the reference positions. As $\pi_{\text{somatic, tumor}}$ provides a good ranking of true versus false mutations, we fit a beta distribution on the $\pi_{\text{somatic, tumor}}$ in this case and call mutations as the extreme and rare events on the right tail of the fitted distribution. For the WGS data, we transform all post-filtered $\pi_{\text{somatic, tumor}}$ to a logarithmic scale and then fit a two-

component Gaussian mixture distribution on it. Given the means μ_1 and μ_2 , standard deviations σ_1 and σ_2 , and weights p and $1 - p$ of the two Gaussian distributions that are estimated using the expectation-maximization algorithm, we first calculate the cutoff that minimizes the misclassification, the sum of the false positive probability and the false negative probability:

$$\Pr_{FP} = p \int_{\text{cutoff}}^{\infty} N(\mu_1, \sigma_1; x) dx$$

$$\Pr_{FN} = (1 - p) \int_{-\infty}^{\text{cutoff}} N(\mu_2, \sigma_2; x) dx.$$

If the cutoff is larger than 0.01, we consider it as PASS and 0.01 as Tier 1, or vice versa. We take the top 0.1, 0.5, and 1 percentiles of the Tier 1 truncated Gaussian distribution as Tier 2, Tier 3, and Tier 4, respectively. For the WES data, we build the sample-specific error model upon post-filtered $\pi_{\text{somatic,tumor}}$ that are within the interval (0.0025, 0.01). We first rescale all selected $\pi_{\text{somatic,tumor}}$ to the range (0, 1), and then fit a beta distribution on them. We report 0.01 as PASS, and cutoffs that are transformed from the top 0.1, 0.5, 1, and 2 percentiles of the beta distribution as Tier 1, Tier 2, Tier 3, and Tier 4, respectively.

Sensitivity and specificity

For the virtual-tumor benchmarking data, we measured sensitivity and specificity by applying MuSE and MuTect [5] to the combination of 24 spike-in BAMs (4 different variant allele fractions \times 6 distinct depths) with the same depth non-spike-in WGS BAMs. The matched-normal WGS BAM was fixed at 30 \times depth. We considered any missed calls from our in silico spike-in ground truth as false negatives, and any calls from the non-spike-in WGS BAMs as false positives. The denominator for the FPR calculation is the total length of the hg19 reference genome from chromosome 1 to chromosome X.

For the DREAM challenge IS1, IS2, and IS3 data, we took the organizer provided script and the truth VCF files to compute sensitivity and specificity [23]. We extracted the sensitivity and specificity of SomaticSniper, Strelka, and VarScan2 from the DREAM challenge leaderboards. The denominator for the FPR calculation is the total length of the hg19 reference genome from chromosome 1 to chromosome X.

For the multi-region lung adenocarcinoma data, we calculated sensitivity and the positive predictive value (PPV) based on an artificial truth set for the reason that the known validation set was extracted and compiled from the paper’s supplementary document and was biased toward Caller A. The artificial truth set included shared calls (Fig. 3c; black in ovals 1, 2, and 3), validated calls (Fig. 3c; orange in oval 1), and unique-not-validated calls that helped the recognition of trunk mutations (Fig. 3c; red in

oval 2 and blue in oval 3). Here, a trunk mutation was a somatic variant call that all tumor regions of one patient had at the same genomic locus. All the other calls were considered as false positives (Fig. 3c; red in five-pointed star 4 and blue in five-pointed star 5). We evaluated accuracy using the F_1 and F_2 scores, which were defined as

$$F_{\beta} = (1 + \beta^2) \frac{PPV \times TPR}{(\beta^2 \times PPV) + TPR} \quad \beta = 1 \text{ or } 2.$$

To compare the performance of multiple callers in the ACC WES data and the ICGC Pilot-63 WGS data, we also made the artificial truth sets by taking calls that were shared by at least three callers, and computed sensitivity. We regarded other calls as false positives to calculate PPVs. We calculated the F_1 and F_2 scores by following the same equation above.

Validation

To validate variants identified by MuSE and Caller A in the ACC data, we selected 550 patient-specific positions and designed NimbleGen probes correspondingly for the purpose of targeted capture enrichment and deep sequencing. Paired-end Illumina resequencing was carried out to an average sequencing depth at 1500 \times . After mapping the reads against the hg19 reference genome using BWA, we considered a somatic variant as validated if its p value calculated from Fisher’s exact test comparing the tumor and matched-normal samples was not larger than 0.05. The validation rates of MuSE and Caller A were calculated as

validation rate of MuSE unique calls

$$= \frac{1}{139 + 2303} \left[\frac{8}{11} \cdot (11 + 141) + \frac{7}{39} \cdot (39 + 221) \right. \\ \left. + \frac{5}{34} \cdot (34 + 345) + \frac{8}{25} \cdot (25 + 494) + \frac{7}{30} \cdot (30 + 1102) \right] \\ \approx 0.2634,$$

validation rate of MuSE shared calls

$$= \frac{1}{290 + 9584} \left[\frac{125}{125} \cdot (125 + 8900) + \frac{99}{111} \cdot (111 + 472) \right. \\ \left. + \frac{25}{29} \cdot (29 + 109) + \frac{12}{17} \cdot (17 + 52) + \frac{7}{8} \cdot (8 + 51) \right] \\ \approx 0.9889,$$

validation rate of MuSE total calls

$$= \frac{1}{139 + 290 + 2303 + 9584} \\ \times [0.2634 \cdot (139 + 2303) + 0.9889 \cdot (290 + 9584)] \\ \approx 0.8450,$$

validation rate of Caller A unique calls

$$= \frac{30}{121}$$

$$\approx 0.2479,$$

validation rate of Caller A total calls

$$= \frac{1}{121 + 290 + 1693 + 9584}$$

$$\times [0.2479 \cdot (121 + 1693) + 0.9889 \cdot (290 + 9584)]$$

$$\approx 0.8739.$$

Additional file

Additional file 1: Table/figure. (PDF 2826 kb)

Availability of data and materials

- (1) We downloaded the WGS BAM files of NA12878 and NA12981 [24], and followed the procedures described in Cibulskis et al. [5] to simulate the virtual-tumor benchmarking data set.
- (2) We downloaded the ICGC-TCGA DREAM Mutation Calling challenge IS1, IS2, and IS3 WGS BAM files and VCF files of simulated truths [23].
- (3) According to TCGA cancer types, we downloaded the tumor-normal paired WGS BAM files of 3 patients in LUAD, LUSC, and STAD, and the tumor-normal paired WES BAM files of 91 patients in ACC from the UCSC Cancer Genomics Hub (CGHub).
- (4) We obtained the tumor-normal paired WES BAM files of 11 patients with lung adenocarcinoma, including 48 multi-regions that were collected and generated at MD Anderson Cancer Center [14].
- (5) We obtained somatic SNV VCF files of 56 samples that were generated by multiple callers from the ICGC Pilot-63 study [25].
- (6) The MuSE source code (v1.0 rc) is available on GitHub (<https://github.com/danielfan/MuSE>) under the GNU General Public License, version 2.0 (GPL-2.0). It has also been deposited at Zenodo (<https://zenodo.org/>) with a DOI (<http://dx.doi.org/10.5281/zenodo.57283>).

Acknowledgments

We thank the ICGC PCAWG for use of the ICGC Pilot-63 data and Kyle Ellrott from UCSC for generating the MuSE calls presented in Fig. 3d. We thank Dr. John N. Weinstein from MDACC and Dr. Richard A. Gibbs from BCM HGSC. The results published here are in part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>.

Funding

YF is partially supported by a training fellowship from the Keck Center of the Gulf Coast Consortia for the Computational Cancer Biology Training Program; the Cancer Prevention and Research Institute of Texas (CPRIT) RP140113, PI - Rathindra Bose; and the National Institutes of Health/National Cancer Institute through grant U24 CA143883 02S2 (to Dr. John N. Weinstein) and the Integrative Pipeline for Analysis & Translational Application of TCGA Data, grant 5U24CA143883-04 (to Dr. John N. Weinstein). WW is supported in part by the Cancer Prevention Research Institute of Texas through grant number RP130090 and by NCI through grant numbers 1R01CA174206-01 and P30 CA016672. YF and WW are supported in part by the US National Cancer Institute (NCI); MD Anderson TCGA Genome Data Analysis Center) through grant numbers CA143883, CA083639 and CA183793. JZ and PAF are supported by the Cancer Prevention and Research Institute of Texas through grant number R1205 01, the UT Systems Stars Award (PS100149), the Welch Foundation Robert A. Welch Distinguished University Chair Award (G-0040), the MD Anderson Physician Scientist Award, and the C.G. Johnson Advanced Scholar Award.

Authors' contributions

YF and WW designed the study. YF designed and developed MuSE and performed the analysis. LX, DSTH, JZ, and JZ assisted in the analysis. YF, DAW, and WW wrote the manuscript. PAF critically reviewed the manuscript. WW led the project. All authors discussed the results. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval

Ethics approval was not needed for this study.

Author details

¹Department of Bioinformatics and Computational Biology - Unit 1410, The University of Texas MD Anderson Cancer Center, P. O. Box 301402, 77230-1402 Houston, TX, USA. ²Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Alkek N1419, 77030-3411 Houston, TX, USA. ³Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Boulevard, 77030 Houston, TX, USA.

Received: 18 February 2016 Accepted: 18 July 2016

Published online: 24 August 2016

References

1. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22(3):568–76.
2. Reumers J, De Rijk P, Zhao H, Liekens A, Smeets D, Cleary J, et al. Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nat Biotechnol.* 2012;30(1):61–8.
3. Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics (Oxford, England).* 2012;28(7):907–13.
4. 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 (Oxford, England).* 2012;28(14):1811–7.
5. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol.* 2013;31(3):213–9.
6. Ewing AD, Houlihan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods.* 2015;12(7):623–30.
7. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet.* 2010;11(10):685–96.
8. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.* 2012;366(10):883–92.
9. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 1981;17(6):368–76.
10. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013;499(7457):214–8.
11. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics (Oxford, England).* 2012;28(3):311–7.
12. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al. Comprehensive pan-genomic characterization of adrenocortical carcinoma. *Cancer Cell.* 2016;29(5):723–36.
13. Rijsbergen CJV. Information retrieval, 2nd ed. Butterworth-Heinemann: Newton; 1979.
14. Zhang J, Fujimoto J, Zhang J, Wedge DC, Song X, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science.* 2014;346(6206):256–9.
15. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabe RR, et al. International network of cancer genome projects. *Nature.* 2010;464(7291):993–8.
16. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 2015;16(1):35.

17. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*. 2014;26(3):319–30.
18. Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol*. 2015;16(1):1–13.
19. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*. 2009;25(14):1754–60.
20. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
21. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491–8.
22. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013;11(1110):11.10.1–11.10.33.
23. ICGC-TCGA DREAM Mutation Calling challenge. <https://www.synapse.org/#!Synapse:syn312572/>. Accessed 5 Aug 2014.
24. NCBI FTP site. ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/. Accessed 10 Apr 2013.
25. NCI SFTP site. sftp://dcssftp.nci.nih.gov/pancan/variant_calling_pilot_64. Accessed 5 June 2015.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

