Accurate detection of somatic small variants with attention-based deep neural networks

Jing Meng[1,2#], Jiangyuan Wang[3#], Wenkai Song[4], Jingze Liu[1,2], Taijiao Jiang[1,2,3*]

[1] Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China
[2] Suzhou Institute of Systems Medicine, Suzhou 215123, China
[3] Guangzhou Laboratory, Guangzhou 510005, China
[4] College of Computer Science, Sichuan University, Chengdu 610065, China

#These authors contributed equally: Jing Meng, Jiangyuan Wang.
*Correspondence and requests for materials should be addressed to Taijiao Jiang.

Cancer diagnosis and precision medicine greatly benefit from accurate identification of somatic small variants in the tumor samples. A number of somatic mutation callers have been developed. However, there is a lack of a model that can capture the influences of the flanking genomic sites on the candidate somatic sites in the context sequence. Here, to fill the gap, we employ the multi-head attention-based deep neural networks to develop a tool called TransSSVs for detection of somatic small variants. TransSSVs focuses on the mapping features in both the local and global views, and enables a reliable representation of the intersite interactions in the context sequence by the multi-head attention mechanism. The benchmarking experiments show that TransSSVs achieves significantly better performances than the five state-of-the-art competitors on the four real tumors. Also, we investigate the contributions of the flanking genomic sites to the accurate detection of somatic mutations, and reveal the different attention weight patterns for the positive and negative somatic sites. We believe that TransSSVs provides a valuable insight into somatic mutation detection.

Keywords: somatic small variants, multi-head attention mechanism, intersite interactions, attention-based deep neural networks, context sequence

Introduction

Somatic mutations play the critical roles in cancer genesis and mutational procession[1,2]. Identifying such mutations is considered a key step for precision medicine[3]. It is convenient to get the sequencing data due to high-throughput sequencing technologies[4]. However it is hard to detect somatic mutations accurately from the sequencing data, due to biological and technological noises, such as intra-tumor heterogeneity, sample contamination, uncertainties in read alignment and base sequencing artifacts[5,6]. To achieve the accurate detection of somatic mutations, a major challenge is how to model biological and technological noises and set the

effective filters to remove the possible false positive predictions[7].

Somatic mutation callers, including Mutect2[8], EBCall[9], VarScan2[10], Strelka2[11], SomaticSeq[12] and MutationSeq[13], basically belong to two types: statistical-based and machine learning-based methods. Mutect2 applies two Bayesian classifiers to call a variant in the tumor and reduce false positives, respectively. Strelka2 contains germline and somatic analyses based on the mixture model, which holds that the normal sample is a mixture of germline variation and noise while the tumor sample is a mixture of the normal sample and somatic variation. VanScan2 reads data from both samples simultaneously, and uses a heuristic and statistical algorithm to detect variants and classifies them by somatic status. However, there are some limitations for these methods. For statistical-based tools, they depend largely on the prior probability and the likelihood probability that a given genotype could have generated the observed data[14]. For machine learning-based methods, their performances rely on the feature engineering, which requires the domain knowledge[3]. Both of the two types of somatic callers share the disadvantages that they cannot model the flanking genomic sites contributing to the somatic state of the genuine somatic sites.

Deep learning can extract the features automatically and learn feature representations directly from the raw data[15-17]. To the best of our knowledge, there are two somatic mutation detection callers based on the deep learning, including DeepSSVs[18] and NeuSomatic[19]. NeuSomatic is a CNN-based approach to detect somatic mutations, which encodes the mapping information into a 3D matrix as input for the deep residual model. In our earlier work, DeepSSVs encodes the raw mapping information of the context sequence to distinguish true somatic mutations from biological and technological noises. However, it cannot investigate the intersite interactions in the context sequence, and cannot weight the influences of the mapping features of the flanking genomic sites to the candidate somatic sites.

To overcome the shortcomes of the currently available somatic callers, we develop a tool called TransSSVs to detect somatic small variants, which employs the attention-based deep neural networks[20]. The core of TransSSVs is to use the multi-head self-attention to obtain a reliable representation of the interactions of the flanking genomic sites and the candidate somatic sites in the context sequence. TransSSVs can effectively extract the mapping features in the context sequence in global and local views for predictions. The benchmarking experiments show that TransSSVs achieves significantly better performances than the five state-of-the-art competitors on the four real tumors. Also, we weight the contributions of the flanking genomic sites to the accurate detection of somatic mutations, and investigate the different attention weight patterns for the positive and negative somatic sites.

Results

Overview of TransSSVs

We show the overview of TransSSVs in Figure 1. The inputs to TransSSVs are candidate somatic mutations identified from a mixed pileup file, which is converted from sequence alignments for tumor and normal samples. We first select the candidate somatic sites based on our defined criteria (more details are provided in section

Methods). Next, the model encodes the mapping information around a candidate somatic mutation site. For each candidate site, we generate a feature matrix for the context sequence (i×52), including i genomic sites in total. The mapping information，including base count, mapping quality and base quality, corresponds to the reference allele and variant allele in the tumor and normal samples, respectively. Then, the attention-based deep neural network captures the interactions of the genomic sites in the context sequence to achieve a new feature representation of the genomic context. The proposed model consists of four blocks, including preparation block, embedding block, encoder block and optimization block (Figure 2). Mapping information matrix of the candidate somatic site is fed to the hidden layer feature space to generate a new feature matrix through preparation block, which is a combination of fully connected layers. The embedding block adds positional embedding to the context sequence (positional vector in Supplementary Figure S1), and the dropout layer is used to improve the robustness. The embedding for context sequence is the input of the encoder block, which is the core step of the TransSSVs. The encoder block contains the multi-head attention mechanism and feature optimization, including dropout layer, normalization layer, fully connected layer and activation layer. Finally, these features are used to predict the somatic probability for the candidate somatic sites by the optimization block, which is a combination of fully connected layers. Considering the imbalanced training and test sets, focal loss is used to achieve optimal performance[21]. TransSSVs is compared to the state-of-the-art somatic mutation detection approaches. TransSSVs learns local and global interactions in the context sequence by the multi-head attention, and achieved better performances.

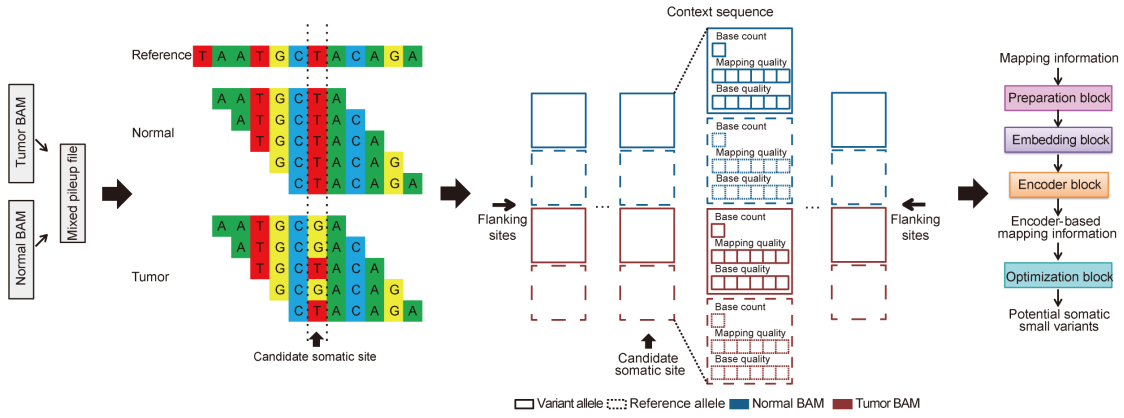

Figure 1. Overview of TransSSVs. The BAM files of tumor and normal samples are first converted into a mixed pileup file where the candidate somatic sites are identified. Then, the mapping information that corresponds to the reference and variant alleles in the tumor and normal samples is extracted and encoded. Next, the intersite interactions in the context sequence are captured by the multi-head attention-based neural network to obtain a new feature representation of the genomic context. Finally, the new feature representation is used to predict the somatic state of the candidate somatic sites.

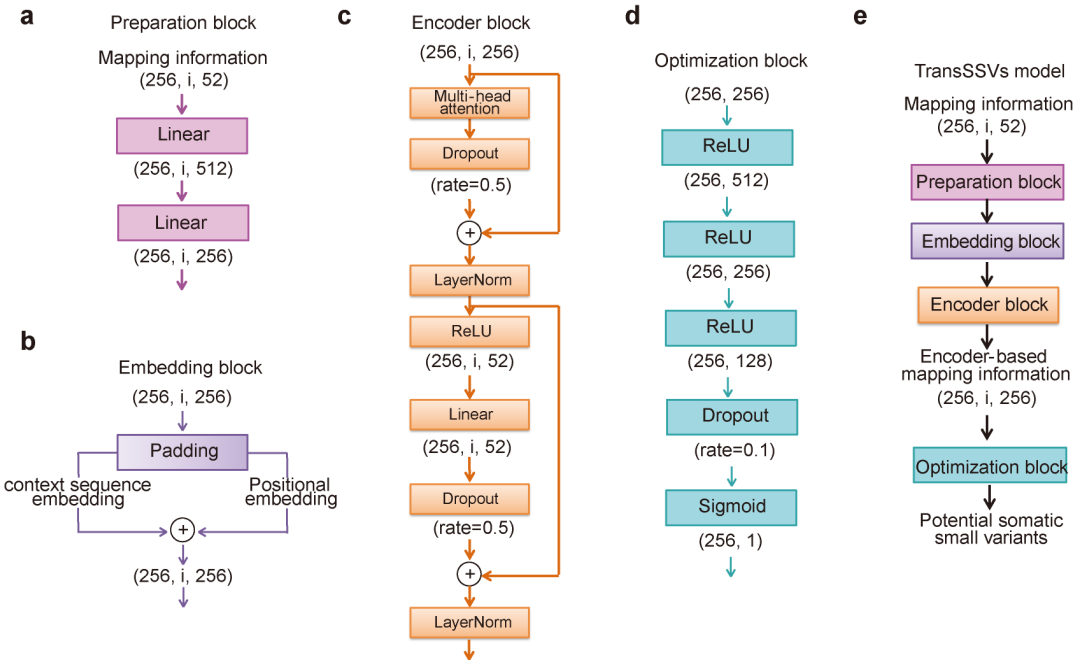

Figure 2. Sub-modules of the TransSSVs and their parameters. a. The preparation block. b. The embedding block. c. The encoder block. d. The optimization block. e. The proposed TransSSVs model. 256: batch_size; i : the length of the flanking genomic sites; 1, 52, 128, 256, 512: the number of neurons of linear layers in preparation block, encoder block and optimization block; 256: embedding size for context sequence in embedding block.

Comparison on the COLO829 dataset

We compared the performance of TransSSVs against the five state-of-the-art somatic callers, including Strelka2, NeuSomatic, Mutect2, VarScan2 and DeepSSVs on the four real tumor datasets (COLO829[22], MB[7], CLL[23] and AML[24]genomes). First we applied a 5-fold cross-validation analysis to do the comparison on the COLO829 dataset. The five test subsets (COLO829-1, COLO829-2, COLO829-3, COLO829-4 and COLO829-5) contain 6811, 6946, 6974, 6935 and 7025 ground truth somatic sites, of which 6738, 6870, 6893, 6851 and 6946 are somatic SNVs, respectively. Considering that the real tumors are lowly mutated, we chosen the negative somatic mutations to be three times the number of the positive somatic mutations in the training set. To explore how the flanking genomic sites around the candidate site affect the performance of the TransSSVs, we set a series of the number of flanking genomic sites and trained the corresponding models of TransSSVs. We test the different numbers of attention headers for TransSSVs, and found that four attention headers achieved better performances (Supplementary Figure S2). We ran these five somatic callers on the five test subsets of the COLO829 genome with default settings.

Table 1 shows the average performance of these callers on the five test subsets, and the performances on each of the five test subsets are showed in the Supplementary Tables S1-S6. The results show that TransSSVs with the different flanking genomic sites have higher F1-scores and MCC than its five competitors. Five models of TransSSVs achieve the top overall F1-scores and MCC values with 0.9268, 0.9265, 0.9257, 0.9257 and 0.9254, and 0.8973, 0.8967, 0.8957, 0.8956 and 0.8963,

respectively, which are higher than DeepSSVs with 0.9238 and 0.8937 and Strelka2 with 0.9068 and 0.8713, respectively. For all of the somatic detection methods, the performances of somatic SNVs are significantly different from that of INDELs. The F1-scores and MCC of somatic SNVs for all of the somatic callers range from 0.8741 to 0.9355 and from 0.8092 to 0.9017, respectively, while the F1-scores and MCC of INDELs are between 0.2641 and 0.5223, and between 0.2925 and 0.5638, respectively (Strelka2 with 0.4987 and 0.5638, TransSSVs from 0.4559 to 0.4894 and from 0.4999 to 0.5293, Mutect2 with 0.3370 and 0.4102, DeepSSVs with 0.3240 and 0.3409, and VarScan2 with 0.2641 and 0.2925, respectively). These differences may result from the hardly distinguishing noise in the somatic INDELs.

Table 1. The performances of somatic callers on the COLO829 dataset

Callers	Somatic INDELs				Somatic SNVs				Overall			
	Recall	Precision	F1-score	MCC	Recall	Precision	F1-score	MCC	Recall	Precision	F1-score	MCC
Strelka2	0.9924	0.3330	0.4987	0.5638	0.9938	0.8484	0.9154	0.8718	0.9938	0.8338	0.9068	0.8713
NeuSomatic	0.7608	0.3976	0.5223	0.5390	0.8920	0.8570	0.8741	0.8092	0.8905	0.8475	0.8685	0.8160
Mutect2	0.9008	0.2073	0.3370	0.4102	0.9510	0.8631	0.9050	0.8489	0.9505	0.8348	0.8889	0.8387
VarScan2	0.8906	0.1550	0.2641	0.2925	0.9867	0.8783	0.9294	0.8151	0.9856	0.8383	0.9060	0.7913
DeepSSVs	0.2366	0.5138	0.3240	0.3409	0.9314	0.9262	0.9288	0.8921	0.9235	0.9240	0.9238	0.8937
TransSSVs-06	0.8422	0.3126	0.4559	0.4999	0.9480	0.9234	0.9355	0.9016	0.9468	0.9055	**0.9257**	**0.8956**
TransSSVs-07	0.8473	0.3214	0.4661	0.5090	0.9450	0.9250	0.9349	0.9008	0.9438	0.9077	**0.9254**	**0.8953**
TransSSVs-40	0.8575	0.3327	0.4794	0.5218	0.9500	0.9215	0.9355	0.9016	0.9489	0.9051	**0.9265**	**0.8967**
TransSSVs-50	0.8524	0.3418	0.4880	0.5277	0.9457	0.9255	0.9355	0.9017	0.9447	0.9097	**0.9268**	**0.8973**
TransSSVs-60	0.8550	0.3429	0.4894	0.5293	0.9432	0.9256	0.9344	0.9000	0.9422	0.9097	**0.9257**	**0.8957**

Bold entries indicate the top five overall F1-scores and MCC

Generalization abilities of somatic callers on the three real datasets

Then, we ran benchmarking experiments on the three real tumor genomes, including MB, CLL and AML to compare the performances of TransSSVs and five state-of-the-art somatic mutation detection approaches. Compared with the COLO829 genome, more false-positive somatic mutations were predicted by the somatic callers in these three tumor genomes. The low mutational loads of the three tumors may lead to the more false-positive predictions. The COLO829 genome is highly mutated, whereas the MB, CLL and AML genomes have low mutational burdens, including only around one mutation per megabase, which may be challenging to distinguish true signals from noise.

TransSSVs significantly outperformed all the five methods, as showed in Figure 3 and Figure 4. The results show that TransSSVs leads in the overall somatic mutations on the MB, CLL and AML, with F1-scores and MCC of 0.7707, 0.6346 and 0.7974, and 0.7686, 0.6302 and 0.7986 respectively, which outperformed Strelka2 by 0.2011, 0.2253 and 0.09196, and 0.1571, 0.1561 and 0.07098, respectively, and Mutect2 by 0.1685, 0.2024 and 0.1061, and 0.1409, 0.1401 and 0.08587, respectively. And the performances of TransSSVs are better than our previous approach DeepSSVs, which has F1-scores and MCC of 0.6210, 0.2982 and 0.7124 and 0.6279, 0.3274 and 0.7145 for MB, CLL and AML, respectively. The extremely lower precision values lead to

the lower performances for the five competitors. For example, TransSSVs yielded recall values of 0.4795 and 0.5911, and precision values of 0.3271 and 0.7607 on the CLL genome for INDELs and SNVs, respectively, while Strelka2 yielded recall values of 0.9041 and 0.9150, and precision values of 0.05160 and 0.4153 on the CLL genome for INDELs and SNVs, respectively (Supplementary Tables S7-S9). The similar phenomena were also observed for the other four somatic callers. Overall TransSSVs can identify somatic mutations with higher confidence and reduce false positive rates compared with the other methods.

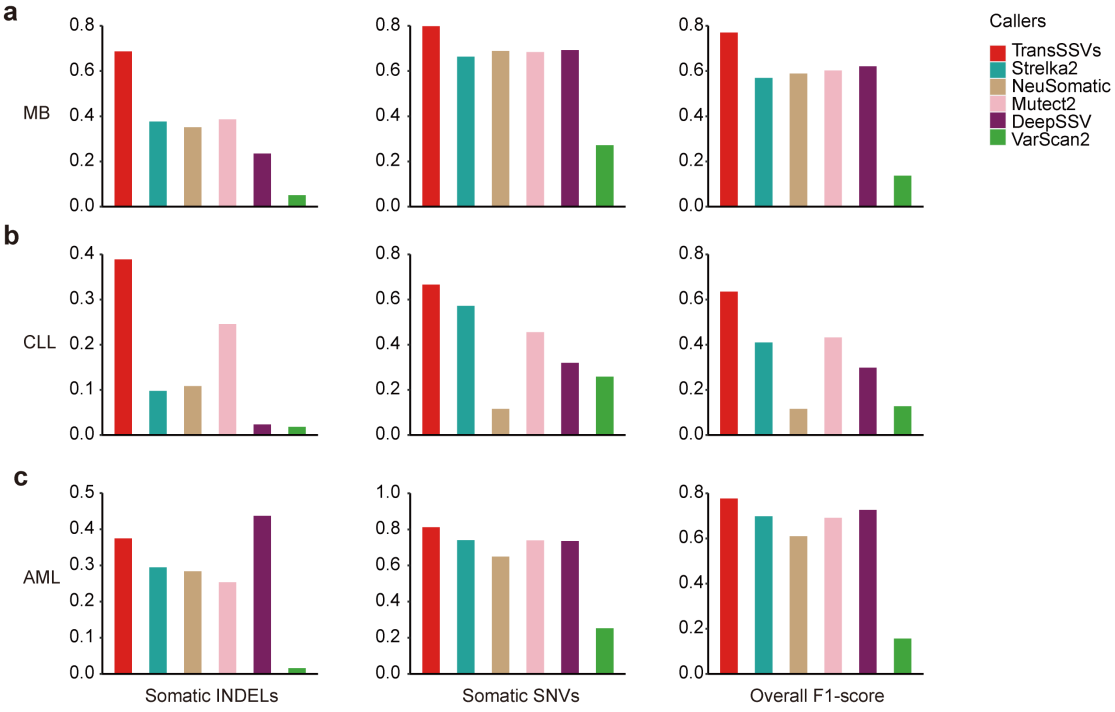

Figure 3. Comparisons of TransSSVs with five state-of-the-art methods on F1-score. (a) The performances on the MB genome. (b) The performances on the CLL genome. (c) The performances on the AML genome.

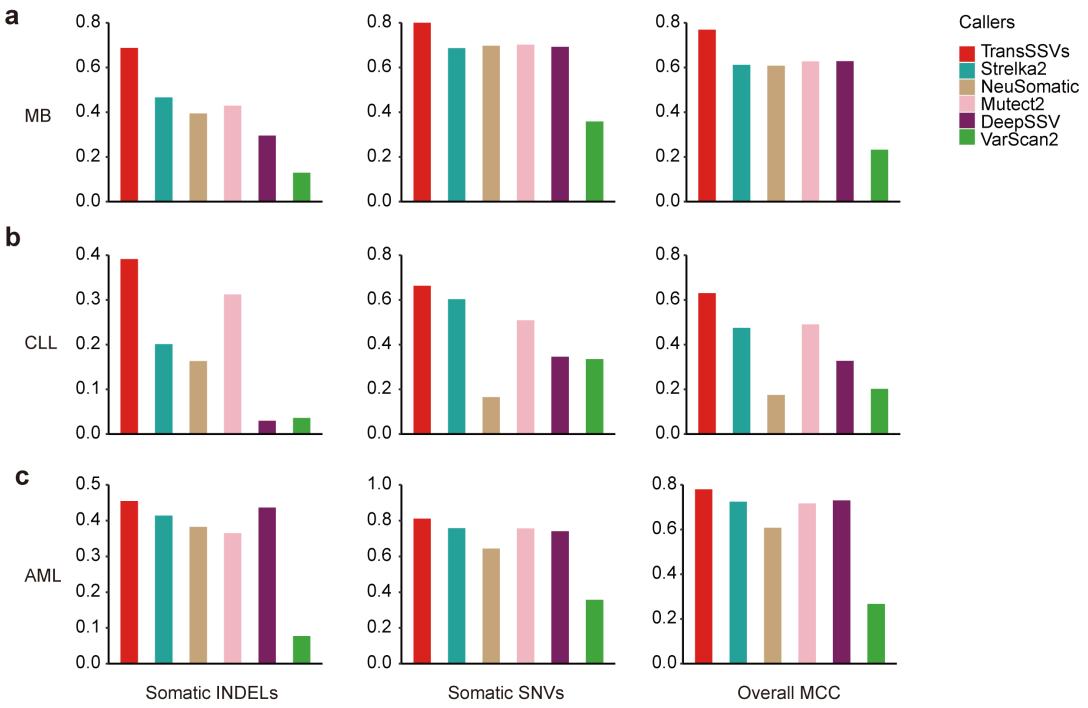

Figure 4. Comparisons of TransSSVs with five state-of-the-art methods on MCC. (a) The performances on the MB genome. (b) The performances on the CLL genome. (c) The performances on the AML genome.

TransSSVs reveals the underlying patterns of the intersite interactions in the context sequence

The self-attention mechanism plays the key role in TransSSVs. It can focus on the different positions of the contextual genome sequence and extract the essential messages from a large number of features. In this section, we attempt to explain the attention mechanism by the attention score. We extract the attention weight matrix from the multi-head attention encoder of TransSSVs using four attention headers (supplementary Table S10-S12). We find that the headers learn features from different views. As shown in figure 3a, it is clear that, for the positive samples with 7 flanking genomic sites, the genomic sites in the context sequence have similar attention weights in the first header, which means that the first header extract features in the global view. In the last three attention headers, the attention weights of the candidate somatic site are significant higher than its flanking genomic sites, which means that the features are extracted in the local view. For negative samples with 7 flanking genomic sites, the fourth attention header pays attention to the global features, while the other three headers focus on the local features.

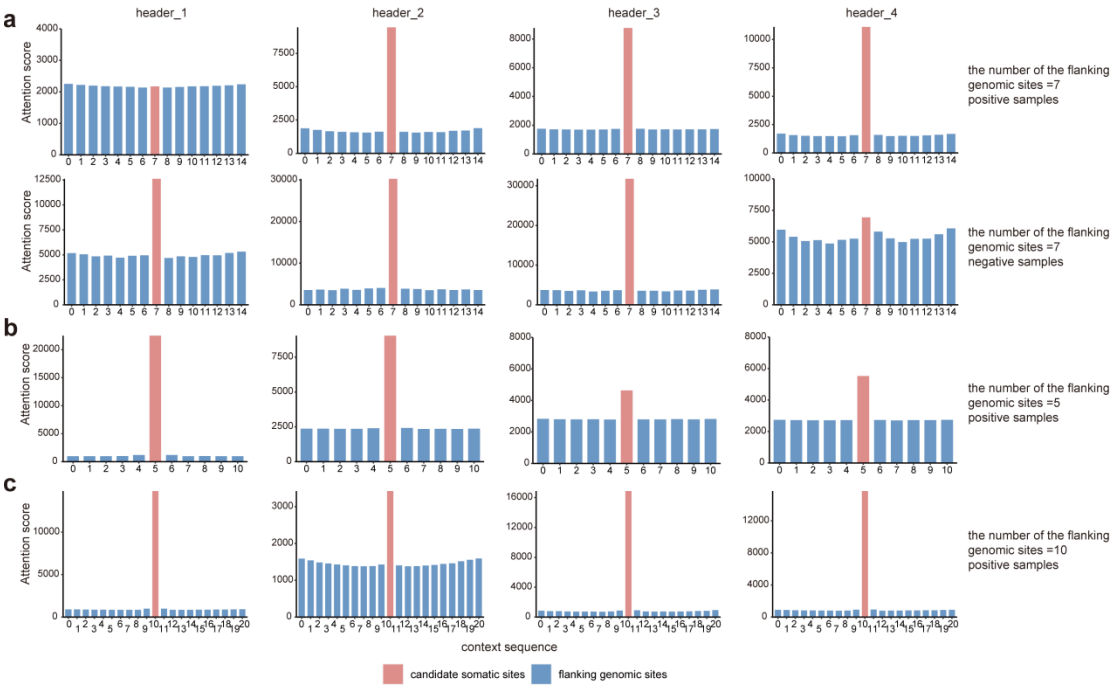

Figure 5. Attention weights. (a) The attention weights associated with correctly predicted positive and negative samples for 7 flanking genomic sites. (b) The attention weights associated with correctly predicted positive samples for 5 flanking genomic sites. (c) The attention weights associated with correctly predicted positive samples for 10 flanking genomic sites.

We also investigate the contributions of the flanking sites with the different numbers. The attention weights of the samples with 5 flanking genomic sites are different from that with 10 flanking genomic sites. For positive samples with 5 flanking genomic sites, the first two attention headers of TransSSVs focus on local features, whereas the last two headers focus on the global features. Compared with the positive samples with 7 flanking genomic sites, the flanking sites in the context sequence in the global view represent a smaller proportion of the attention weights for the positive samples with 5 and 10 flanking genomic sites. It may explain the better performances for the models of TransSSVs with 7 flanking genomic sites. In summary, TransSSVs can focus on the different genomic sites in the context sequence in local and global views to provide a reliable feature representation for somatic mutation detection.

Discussion

Cancer diagnosis and precision medicine greatly benefit from accurate identification of somatic mutations in the tumor samples. Based on statistical and machine learning methods, there are a number of somatic callers that have been developed to identify somatic mutations. In this study, we employ the advantages of the multi-head attention mechanism to develop a tool called TransSSVs to detect somatic small variants. It uses different headers to capture the intersite interactions in the context sequence in global and local views to generate a reliable representation of the mapping features for accurate detection of somatic mutations.

The ground truths of the COLO829 genome were used to train TransSSVs by a 5-fold cross-validation analysis, and we compared the performances of TransSSVs with five state-of-the-art somatic methods on the test subsets of the COLO829. Furthermore, we did the benchmarking experiments on the other three real tumor datasets, which have low mutational loads. The benchmarking results showed that TransSSVs outperforms its five competitors with significantly higher overall F1-scores and MCC in detecting somatic mutations.

We also analyzed the contributions of the different numbers of the flanking genomic sites in the context sequence. TransSSVs can pay attention to both the local and global features by the multi-head attention mechanism to capture a reliable feature representation for accurately identifying somatic mutations. Through investigating the attention weights, we found that the better performances of the TransSSVs models with 7 flanking genomic sites may result from the interactions of the flanking genomic sites and the candidate somatic sites in the context sequence. Also, the positive samples and negative samples display different feature representations, which may explain their different attention weight patterns. For the negative samples, flanking genomic sites attract more attentions, while TransSSVs focuses more on the candidate somatic sites in the context sequence for the positive samples. We believe that TransSSVs provides a valuable insight into somatic mutation detection.

Methods

Datasets

In this study, four real tumor and matched normal whole genome sequencing datasets were selected to assess TransSSVs and the other five somatic callers. The first dataset is a paired COLO829 and COLO829BL cell lines with sequencing depths of 80x, which were provided by Translational Genomics Research Institute (TGen)[25]. The second and third dataset are medulloblastoma (MB) and chronic lymphocytic leukemia (CLL) from the International Cancer Genome Consortium (ICGC) with sequencing depths of 40x for tumor and 30x for normal samples[7]. The remaining dataset is a primary acute myeloid leukemia (AML) with sequencing depth of 312x and its matched normal sample with sequencing depth of 121x[24].

For AML, MB and CLL samples, BWA was first used to map the reads in these original FASTQ format files to the human reference genome hg38/GRCh38 with default parameters[26]. Then, the resulting BAM files of these samples were marked PCR and optical duplicates by Picard and adjusted base quality scores and realigned by GATK[27]. Considering that the cancer genomes are generally sequenced to depths between 30x and 50x, the BAM files of COLO829 and AML were downsampled from 80x to 50x for both tumor and normal samples, and from 312x and 121x to 50x for tumor and normal samples, respectively. We chose the highly mutated COLO829 genome as the training set, which enabled us to create a relatively balanced training set. Three lowly mutated real tumors (MB, CLL and AML) were selected to test the generalization abilities of somatic callers.

Preparation of candidate somatic sites

We scan the mixed pileup file created by Samtools from the tumor and matched normal BAM files to identify candidate somatic sites as in DeepSSVs[28]. To make sure the candidate sites are reliable, we select the genomic sites as candidate somatic sites based on the following criteria: (1) the base in the reference genome is standard; (2) the strand-specific frequency is between 10% and 90%; (3) the length of the INDELs in the tumor is no more than 50; (4) the depths of the sites in the tumor and normal samples are larger than 10; and (5) the Phred scaled base quality scores and mapping quality scores of the covering reads are larger than 10.

Input matrix of candidate somatic sites

For each candidate somatic site, the feature matrix of the context sequence (15×52) was generated, where 15 is the length of the context sequence and 52 corresponds to the feature information around the candidate somatic site. The feature information includes the count, the base and mapping qualities of the variant allele and reference allele in tumor and matched normal alignments. Since the length of the context sequence is variable, if the matrix has L sites, the input matrix of a candidate somatic site could be showed as $M = \{M_1, M_2, ..., M_L\}$, where M_i denotes the feature information of the i th site in the input matrix.

Capture of intersite interactions by attention-based neural networks

Each site in the context sequence was encoded into a vector representation as above. To capture the intersite interactions, the feature vector of a site was first encoded by the fully connected layer with mapping information M_i as input,

$$\beta_i^h = W_\beta Linear(M_i),$$

$$\gamma_i^h = W_\gamma Linear(M_i),$$

$$\delta_i^h = W_\delta Linear(M_i),$$

where the number of the attention heads was represented as h. The attention heads can extract more information derived from different sites in different representation spaces. Each head focuses on the different areas that are corresponding to the different positions of the context sequence, and provides information for the model to achieve better site-based features with contextual semantics. W_β, W_γ and W_δ were the parameter matrices. Then the attention weight was formulated by considering the feature vectors in the context sequence,

$$\alpha_i^h = softmax_i\left(\frac{\beta_i^h \gamma_i^{hT}}{\sqrt{d}}\right),$$

where d is the dimension of β_i^h and γ_i^h. The attention weight, which is the weighted

representation of each genomic site, indicates the correlation of the candidate somatic site with the other genomic sites in the context sequence. The signal passed from site to site in the context sequence was formulated as

$$O_i^h = \sum_i \alpha_i^h \delta_i^h .$$

Finally, signal vectors O_i^h from different heads were concatenated to update mapping features by the fully connected layer and the normalization layer[29,30]. By integrating signals in the context sequence, the final feature vector of the candidate somatic site F_i^h represents its interaction with the surrounding genomic sites and is used to predict the somatic state.

Model training

In general, the candidate somatic list includes significantly more non-somatic sites than the true somatic sites. Since the numbers of positive and negative sites in the training set are imbalanced, the focal loss is used for TransSSVs model training[21].

$$L_{fl} = \begin{cases} -\alpha\left(1-y'\right)^{\gamma} \log y' & y=1 \\ -\left(1-\alpha\right) y'^{\gamma} \log\left(1-y'\right) & y=0 \end{cases}$$

There are two factors including balance factor and focusing parameter. Balance factor α can balance the importance of positive and negative somatic sites in the training set, through giving high weights to the true positive sites and small weights to the true negative sites[31]. The parameter γ smoothly adjusts the rate at which easily detected sites are down-weighted. We set the balance factor α=0.25 and focusing parameter γ=2 to address class imbalance during training TransSSVs. We used tensorflow2 to train TransSSVs on the TITAN RTX GPU with 24 GB memory, and used the Adam optimizer with an initial learning rate of 0.001[32].

Parameters of the other five somatic callers

We compare the performance of TransSSVs with five state-of-the-art somatic callers, Strelka2 (v2.9.10), NeuSomatic (v0.2.1), Mutect2 (v2.2) and DeepSSVs with their default settings, and VarScan2 (v2.3.9) with 0.05 as the min_vaf. The records with 'PASS' in the FILTER field were chosen as predictions from the somatic callers for benchmarking.

Performance evaluation metrics

We calculated the precision and recall for benchmarking somatic callers. Precision and recall are the fraction of the true somatic sites in the predictions, and the fraction of true somatic sites that are predicted, respectively[33]. The F1-score is the harmonic average of the precision and recall.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Where TP is true-positive, TN is true-negative, FP is false-positive and FN is false-negative. We also calculated Matthews Correlation Coefficient (MCC) for benchmarking somatic callers[34]. MCC value is between -1 and 1, and this value indicates the correlation of true and predicted values. The higher the correlation of true and predicted values, the better the model.

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}}$$

Extraction of attention weights

We extracted TransSSVs attention matrices across all heads and layers. We also extracted the row corresponding to the candidate somatic site, which provided the attentions of the flanking genomic sites to the candidate somatic site.

Author contributions

JM and JW conceived the idea, designed and implemented the algorithm, and analyzed the results and drafted the manuscript. JL and WS helped in algorithm implementation. TJ and JM revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32070678,31671371); the Emergency Key Program of Guangzhou Laboratory, Grant No. EKPG21-12; the National Key Research and Development Program of China (2020YFC0840800).

Code availability

The code is freely available at https://github.com/jingmeng-bioinformatics/TransSSVs with GNU General Public License Version 3.

Competing interests

The authors declare that they have no competing interests.

References

1 Aparicio, S. & Caldas, C. The implications of clonal genome evolution for cancer medicine. *N Engl J Med* **368**, 842-851, doi:10.1056/NEJMra1204892 (2013).

2 Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306-313, doi:10.1038/nature10762 (2012).

3 Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput Struct Biotechnol J* **16**, 15-24, doi:10.1016/j.csbj.2018.01.003

(2018).

4 Teer, J. K. An improved understanding of cancer genomics through massively parallel sequencing. *Transl Cancer Res* **3**, 243-259, doi:10.3978/j.issn.2218-676X.2014.05.05 (2014).

5 Bergfeld, S. A. & DeClerck, Y. A. Bone marrow-derived mesenchymal stem cells and the tumor microenvironment. *Cancer Metastasis Rev* **29**, 249-261, doi:10.1007/s10555-010-9222-7 (2010).

6 Mwenifumbo, J. C. & Marra, M. A. Cancer genome-sequencing study design. *Nat Rev Genet* **14**, 321-332, doi:10.1038/nrg3445 (2013).

7 Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat Commun* **6**, 10001, doi:10.1038/ncomms10001 (2015).

8 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

9 Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res* **41**, e89, doi:10.1093/nar/gkt126 (2013).

10 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

11 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods* **15**, 591-594, doi:10.1038/s41592-018-0051-x (2018).

12 Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol* **16**, 197, doi:10.1186/s13059-015-0758-2 (2015).

13 Ding, J. *et al.* Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28**, 167-175, doi:10.1093/bioinformatics/btr629 (2012).

14 Wilkinson, D. J. Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform* **8**, 109-116, doi:10.1093/bib/bbm007 (2007).

15 Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* **35**, 1798-1828, doi:10.1109/TPAMI.2013.50 (2013).

16 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

17 Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw* **61**, 85-117, doi:10.1016/j.neunet.2014.09.003 (2015).

18 Meng, J., Victor, B., He, Z., Liu, H. & Jiang, T. DeepSSV: detecting somatic small variants in paired tumor and normal sequencing data with convolutional neural network. *Brief Bioinform* **22**, doi:10.1093/bib/bbaa272 (2021).

19 Sahraeian, S. M. E. *et al.* Deep convolutional neural networks for accurate somatic mutation detection. *Nat Commun* **10**, 1041, doi:10.1038/s41467-019-09027-x (2019).

20 Ashish Vaswani, N. S., Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)* (2017).

21 Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. in *2017 IEEE International Conference on Computer Vision (ICCV)* 2999-3007 (2017).

22 Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196, doi:10.1038/nature08658 (2010).

23 Puente, X. S. *et al.* Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* **475**, 101-105, doi:10.1038/nature10113 (2011).

24 Griffith, M. *et al.* Optimizing cancer genome sequencing and analysis. *Cell Syst* **1**, 210-223, doi:10.1016/j.cels.2015.08.015 (2015).

25 Craig, D. W. *et al.* A somatic reference standard for cancer genome sequencing. *Sci Rep* **6**, 24607, doi:10.1038/srep24607 (2016).

26 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v2 [q-bio.GN]* (2013).

27 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

28 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, doi:10.1093/gigascience/giab008 (2021).

29 Nitish Srivastava, G. H., Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from overfitting. *Journal of Machine Learning Research* **15**, 1929-1958 (2014).

30 Jimmy Lei Ba, J. R. K., Geoffrey E. Hinton. Layer Normalization. *arXiv:1607.06450* (2016).

31 Nie, L. *et al.* TransPPMP: predicting pathogenicity of frameshift and non-sense mutations by a Transformer based on protein features. *Bioinformatics* **38**, 2705-2711, doi:10.1093/bioinformatics/btac188 (2022).

32 Diederik P. Kingma, J. L. B. Adam: A method for stochastic optimization. *arXiv:1412.6980* (2015).

33 Vijayan, V., Yiu, S. M. & Zhang, L. Improving somatic variant identification through integration of genome and exome data. *BMC Genomics* **18**, 748, doi:10.1186/s12864-017-4134-3 (2017).

34 Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6, doi:10.1186/s12864-019-6413-7 (2020).