Supplementary Information for

MutSpot: detection of non-coding mutation hotspots in cancer genomes

Yu Amanda Guo[1*], Mei Mei Chang[1*] & Anders Jacobsen Skanderup[1]

[1] Computational and Systems Biology, Agency for Science Technology and Research, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore

*These authors contributed equally to this work

Correspondence should be addressed to Anders Jacobsen Skanderup (skanderupamj@gis.a-star.edu.sg) or Yu Amanda Guo (yg246@gis.a-star.edu.sg)

Supplementary Methods

Input data for MutSpot

MutSpot requires a mutation file in the MAF format, and genomic/epigenomic features in the bigwig or bed formats. Clinical features and other sample specific features can also be supplied in plain text format (optional). MutSpot provides a default set of 135 epigenetic features including replication timing, transcription factor binding profiles and APOBEC editing sites. The user has the option to include additional features such as cancer type specific epigenetic profiles. Discrete epigenetic features such as peak calls of histone modification should be provided in the bed format, and continuous features such as replication timing profile should be provided in the bigwig format. Continuous genomic features will be discretized into 10 equally sized bins, and the mean value of each bin is used for regression.

MutSpot automatically computes sequence features from the input mutation file. Sequence features for SNVs include identity of mutated base (A/T or CG), the trinucleotide and pentanucleotide context of the mutated site, and the 1bp and 2bp left and right flanks of the mutated site. Sequence features for indels include the presence of mononucleotide repeats of 5bp or longer at the mutated sites. Finally, MutSpot also computes the local mutation rate in 100kb non-overlapping bins to account for additional unknown covariates of regional mutation rates.

Feature selection using LASSO regression

MutSpot uses LASSO logistic regression to identify the most predictive features of somatic mutation variation in the provided cancer genomes. It is computationally intensive to fit a regression model over all positions in the non-coding genome with a large number of predictor variables. To keep the computational cost of feature selection tractable, MutSpot randomly samples 2 million mutated sites from the input MAF file (or all mutated sites if the total number of mutations is less than 2 million) and an equal number of non-mutated sites as input for the LASSO logistic regression model. Then, MutSpot regresses the binary mutation status of each site against the candidate sequence or epigenetic features. The regularization parameter is chosen by 10-fold cross-validation such that the error of the selected model is within 1 standard deviation from the minimum error. MutSpot uses the 'glmnet' package for LASSO regression and cross validation.

$$glmnet(y \sim \beta X, family = logistic)$$

To select for the most robust predictive features, MutSpot bootstraps 100 samples with 50% of the data in each bootstrap, and performs LASSO regression using the bootstrap samples. By default, MutSpot uses epigenomic features selected in more than 75% of the bootstrap samples and sequence features selected in more than 90% of the bootstrap sample for the final regression model. The user can adjust these thresholds to control the number of features included in the final background mutation model.

Sample- and position-specific background mutation model

MutSpot estimates the patient specific background mutation probabilities by fitting a logistic regression model on all genomic sites after masking CDS regions, immunoglobin loci and poorly mappable regions. MutSpots fits the logistic regression model on the frequency table of the counts of mutated and non-mutated bases for each combination of the covariates.

Since SNVs and indels arise from different mutational processes, MutSpot builds separate models to estimate the background mutation probabilities of SNVs and indels.

$$glm(y \sim \beta X, family = logit)$$

Here, X includes sequence and epigenetic features selected by LASSO regression as well as sample specific features such as tumor mutation count and clinical features.

Poisson binomial model of mutation recurrence

For a specific region of the genome, the probability of mutation in tumor *i,* p_i, is a function of the length of the region and the expected mutation rates of individual nucleotides in that region given the background mutation model. Assuming $q_{i,j}$ is the mutation probability of nucleotide *j* in tumor *i*, and *l* is the length of the region of interest:

$$p_i = 1 - \prod_{j=1}^{l}(1 - q_{i,j})$$

The *P* value of mutation recurrence of the given region is modelled using the Poisson binomial distribution, which accounts for variations in mutation rate across tumours. For a specific region of the genome, the probability of having mutations in *k* or more tumors is given by:

$$\Pr(K \geq k) = \sum_{m=k}^{n} \sum_{A \in F_m} \prod_{i \in A} p_i \prod_{j \in A^c} (1 - p_j)$$

Here *n* is the total number of tumors sequenced, *k* is the number of tumors with mutations in the region of interest, F_m is the set of all subsets of *k* integers selected from {1,2,…,n}, *A* is a subset of F_m, A^c is the complement of set A, p_i is the probability of mutation in tumor *i*, and p_j is the probability of mutation in tumor *j*. The Poisson binomial probability is calculated with the 'poibin' R package.

Identification of mutation hotspots

Finally, Mutspot considers all mutated focal regions by taking *m* bp flanks on each side of each mutation (default *m*=10, corresponding hotspots of 21bps). For each potential hotspot region with n or more mutated samples (default n=2), MutSpot calculates the *P* value of mutation recurrence using the Poisson binomial model described in the previous section. MutSpot adjusts for multiple testing using the Benjamini–Hochberg approach, where the total number of hypothesis tested is equal to the number of bases in the masked non-coding genome.

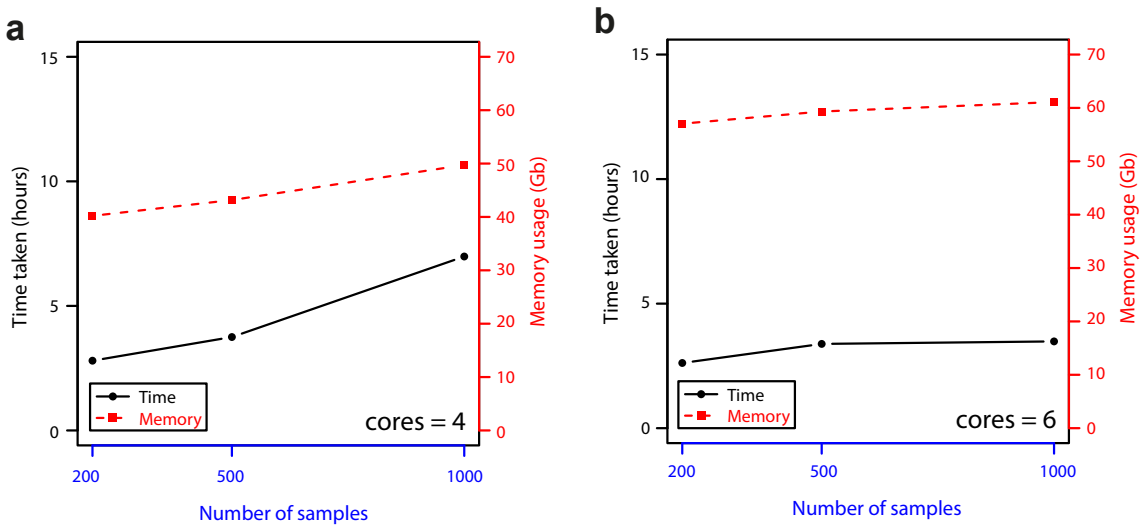
Benchmarking MutSpot on simulated cancer genomes

To benchmark the memory and time usage of MutSpot for large tumor cohorts, we simulated SNV profiles of 1000 cancer genomes from whole genome sequencing data of 168 non-hypermutated gastric cancer tumors using the following steps:

1. Identify the 10th (A) and 90th (B) quantiles of tumor mutation burden from 168 gastric cancer whole genomes.

2. Simulate the mutation burden of 1000 genomes by choosing 1000 random numbers ($x_1, x_2, x_3, x_4, \ldots x_{1000}$) between A and B.
3. For the i^{th} sample, we sample x_i mutations from the 168 gastric cancer genomes. Then for each mutation, we add a random number between -1000 to 1000 to the position to shift the location of the mutation.

We benchmarked MutSpot on 200, 500 and 1000 simulated cancer whole genomes (**Supplementary Fig. 1**). We ran MutSpot on the simulated tumours using 4 cores and 6 cores on a R4 machine (8vCPU, 61GiB) on Amazon Web Services. We used the default cutoff for epigenetic feature selection (0.75) and chose the top 5 nucleotide context features for each background model.



Supplementary Figure 1 Time and memory usage of MutSpot ran on 200, 500 and 1000 cancer whole genomes using **(a)** 4 cores and **(b)** 6 cores.