1 GENOME ANALYSIS

Efficient Detection of Highly Mutated Regions with Mutations Overburdening Annotations Tool (MOAT)

Lucas Lochovsky[1,2,†], Jing Zhang[1,2,†], and Mark Gerstein[1,2,3*]

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA

[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

[3]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: Identifying genomic regions with a higher-than-expected mutation burden is useful for cancer driver detection. Here, we introduce the Mutations Overburdening Annotations Tool (MOAT), a new software that can perform mutation burden analysis at high speed. MOAT makes no assumptions about the mutation process, except that the background mutation rate (BMR) changes smoothly with other genomic features. This nonparametric scheme randomly permutes the variants or target regions on a relatively large scale to provide robust burden analysis in cancer driver detection. Furthermore, we provide two versions of the variant permutation method in the MOAT software suite. The first version permutes variants locally, while the second version permutes variants in a global context, using BMR covariates to select similar regions for permuted variants. MOAT also offers the option to evaluate the functional impact within annotations for burden analysis. In conclusion, MOAT is useful for a broad range of analyses that would benefit from variant and target region permutation.

Availability and Implementation: MOAT is available at moat.gersteinlab.org

Contact: pi@gersteinlab.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

2 INTRODUCTION

A common analysis strategy in cancer driver detection is to look for genomic elements with high variant accumulation across patients. However, the background mutation rate (BMR) is highly heterogeneous across the genome due to numerous influences. Inaccurate modeling of BMRs could in turn introduce false positives into cancer driver detection. Our Mutations Overburdening Annotations Tool (MOAT) differs from parametric schemes and does not make any assumption except that the BMR remains constant within a local context.

MOAT offers an annotation-centric algorithm (MOAT-a), a variant-centric algorithm (MOAT-v), and a somatic variant simulator (MOAT-sim) built on MOAT-v's variant placement algorithm. Moreover, we can use MOAT to gauge the functional impact burden of annotations relative to the surrounding genome. MOAT is useful for comparing observed and permuted variant impact scores. Here, we provide an example using Funseq2 scores (Fu, et al., 2014). In the following sections, we describe MOAT's implementation and recall of known noncoding cancer drivers.

3 METHODS

Several covariates jointly affect the BMR in a complicated and dynamic manner, making variant burden analysis very challenging (Lawrence, et al., 2013). The length of the test region usually varies from hundreds to thousands of bases, while external features such as replication timing can work at up to a megabase resolution. To address these challenges, MOAT circumvents the need for parametric models by explicitly permuting the variants or annotations within a region where the levels of all the covariates are essentially constant. One important issue with these permutation algorithms is that their running times do not scale well to whole-genome annotation sets. We addressed this issue by taking advantage of large-scale graphics processing unit (GPU) parallelization.

3.1 MOAT-a: Annotation-Centric Permutation

MOAT requires two input files: an annotation file (*afile*) and a variant file (*vfile*). MOAT-a uses NVIDIA's compute unified device architecture language (Nickolls, et al., 2008) for general-purpose GPU acceleration (Figure 1a). MOAT-a iterates through each annotation, computing the intersecting variant count. It defines a genomic block with user-defined boundaries for permuting the annotation n times. MOAT-a then finds the variant counts of the n random bins, and compares them to the annotation's observed variant count to provide empirical p-values. When MOAT-a is used with a functional impact signal file, it generates observed and permuted annotation impact scores by summing the intersecting variants' impact scores to calculate p-values.

We can adjust the boundaries of the intervals for choosing permuted annotations—*d_min* and *d_max*—to scale the surrounding genome context with respect to the size of the original annotation.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Deleted: can confer several useful applications such as

Deleted: , using any functional whole-genome signal

Deleted: scores (of any functional whole-genome set) to evaluate functional impact

Deleted: file

Deleted: s

Deleted: (

Deleted:)

Deleted: However, t

Deleted: Therefore

Deleted: Furthermore, the MOAT software suite incorporates MOAT-sim, a somatic variant simulator that randomly permutes the input variants with effective covariate control.

Deleted: We conclude that

Deleted: mark.gerstein@yale.edu

Deleted: numerous

Deleted: other

Deleted: s

Deleted: If we run MOAT with a functional impact score option, we can compute and use the variants' scores to calculate annotation scores by summing the scores of the intersecting variants.

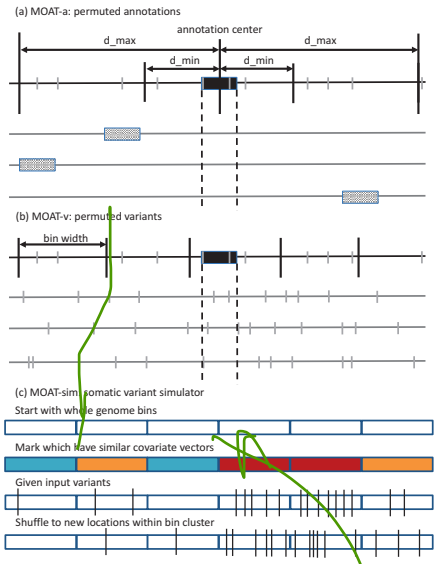

Figure 1 (a) MOAT-a shuffles each annotation to a new location within the local genome context bounded by user-defined parameters *d_min* and *d_max*, producing *n* permutations. (b) In MOAT-v, the whole genome is divided into bins of user-defined size *bin width*, within which variants are moved to new coordinates, thereby preserving the local mutation context. As with MOAT-a, MOAT-v produces *n* permutations. (c) MOAT-sim bins the entire genome, whereupon it calculates the covariate values for each bin. The program then clusters bins with similar covariate values, represented here as bins with the same color (referred as equivalent class). The input variants that fall within each cluster are then permuted to new locations chosen from the bins within the same cluster, honoring trinucleotide context preservation if requested.

Ideally, the permutation intervals will provide enough range to enable non-overlapping sampling. As a rule of thumb, the choice of *d_min* should be large enough to avoid potential mutation burden signal from "bleeding" into the permutation intervals. Simultaneously, the selected *d_max* must be small enough that the BMR covariates remain approximately constant within the permutation intervals. For example, in our analysis of transcription start site(s) (TSS) mutation burdens, where TSS are roughly 100 bp in length, we used a *d_min* of 2kb and a *d_max* of 50kb.

3.2 MOAT-v: Variant-Centric Permutation

MOAT-v creates permuted datasets by assigning new coordinates to each variant within a local genomic region to account for the covariate effects from known genomic features (Figure 1). MOAT-v (and similarly in MOAT-sim) offers the option to preserve the tri-nucleotide context of the original variant when choosing a new variant location (see supplement). This constraint reflects the differential mutation probabilities of different tri-nucleotides while preserving the mutational signatures. MOAT-v generates a permuted dataset by subdividing the genome into blocks of a user-defined size within which variants are permuted, thus generating *n* permutations (Figure 1b). We can determine the empirical *p*-value for each annotation based on the fraction of permutations with variants equal to or greater than the observed variant count. Unlike MOAT-a, we designed MOAT-v to parallelize its workflow across multiple central processing unit (CPU) cores using OpenMPI framework (Gabriel, et al., 2004), due to the more memory intensive nature of the tri-nucleotide context preservation.

The ability to adjust the width of the whole-genome bins in MOAT-v enables users to select a width that represents regions in which the BMR covariates are expected to be approximately constant. Hence, the permutations that MOAT-v creates will honor the expected density of regional mutations due to these covariates. Our analyses of a few of the most significant covariates, such as DNA-replication timing, histone marks, and guanine-cytosine content, indicate that a suitable bin size range is 50 – 100 kb (see supplement).

3.3 MOAT-sim: Simulated Somatic Variant Datasets

In addition to the main MOAT programs, we developed a variant simulator, MOAT-sim, that reflects the levels of whole-genome covariates that directly influence the background mutation rate. MOAT-sim evaluates covariate signals over a set of whole-genome bins. The simulator then clusters these bins based on their covariate signal profiles, and allows variants to be permuted not just within their local genome context, but across all bins that share the same covariate signal profile (referred as equivalent class). Specifically, MOAT-sim clusters the whole-genome bins using *k* means, which use the distances between the bins' covariate signal profiles to group them into a predefined number of clusters (see supplement).

4 RESULTS

4.1 MOAT-a

We demonstrated the parallel speedup by running MOAT-a on datasets of various sizes. Using a dataset of ~8 million cancer variants from (Alexandrov, et al., 2013; Wang, et al., 2014), we used three different annotation sets to demonstrate the scalability of MOAT-a (Harrow, et al., 2012; Thurman, et al., 2012; Yip, et al., 2012). We demonstrate that the GPU version of MOAT-a scales very well with respect to the number of annotations (e.g. ~9-fold speedup on ~3 million annotations), and with respect to the number of permutations (e.g. ~256-fold speedup on 100,000 permutations), resulting in dramatically improved running times (Supp Table 1).

Due to the lack of a gold standard, assessing MOAT's predictions is challenging. Nevertheless, we used the aforementioned cancer variant dataset to demonstrate how MOAT-a can find elevated mutation burdens in genomic elements by identifying highly mutated GENCODE elements. TERT, which has well-documented cancer-associated promoter mutations, carried a significant mutation burden. Other well-known cancer-associated TSS sites, such as TP53, LMO3, and AGAP5, also had significant mutation burdens.

4.2 MOAT-v & MOAT-sim

Using the same set of cancer variants as in the MOAT-a tests, we evaluated MOAT-v's running time. The running time scales close to linear with the number of CPUs, indicating an even division of labor between each CPU core. MOAT-sim's running time exhibited similar characteristics (data not shown).

Deleted: s from a
Deleted: resolution
Deleted: and (Wang, et al., 2014)
Deleted: <sp>
Deleted: en
Deleted: -
Deleted: were

We then applied MOAT-v on the same variant and annotation sets to find elevated cancer mutation burdens. MOAT-v produced comparable results as MOAT-a, flagging the same known cancer-associated TSS sites as significant.

5 DISCUSSION

Here, we introduce MOAT, a new software tool to facilitate identification of high mutation burden. We demonstrate the usefulness of this tool for flagging putative noncoding cancer drivers, and provide parallelized versions that dramatically improve running time. Given the demand for efficient and meaningful analysis of genome sequence data, which scientists are producing at very high rates, we believe that MOAT's provision of such analysis for genetic disease drivers is timely.

Deleted: Identifying genomic elements with a high mutation burden can help narrow down the exact site of functional disruption.

Deleted: such analyses

Deleted: . We also

Deleted: increase the speed of mutation burden analysis

Deleted: very

Funding: This work was supported by the National Institutes of Health [grant number 5U41HG007000-04].

REFERENCES

Alexandrov, L.B., *et al*. Signatures of mutational processes in human cancer. *Nature* 2013;500(7463):415-421.

Fu, Y., *et al*. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome biology* 2014;15(10):480.

Gabriel, E., *et al*. Open MPI: Goals, concept, and design of a next generation MPI implementation. *Springer* 2004:97-104.

Harrow, J., *et al*. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* 2012;22(9):1760-1774.

Lawrence, M.S., *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499(7457):214-218.

Nickolls, J., *et al*. Scalable parallel programming w/CUDA. *Queue* 2008;6(2):40-53.

Thurman, R.E., *et al*. The accessible chromatin landscape of the human genome. *Nature* 2012;489(7414):75-82.

Wang, K., *et al*. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics* 2014;46(6):573-582.

Yip, K.Y., *et al*. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome biology* 2012;13(9):R48.