BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3414–3416
doi:10.1093/bioinformatics/btu577

Sequence analysis

Advance Access publication August 27, 2014

RVboost: RNA-seq variants prioritization using a boosting method

Chen Wang[1,†], Jaime I. Davila[1,†], Saurabh Baheti[1], Aditya V. Bhagwate[1], Xue Wang[2], Jean-Pierre A. Kocher[1], Susan L. Slager[1], Andrew L. Feldman[3], Anne J. Novak[4], James R. Cerhan[5], E. Aubrey Thompson[6] and Yan W. Asmann[2,*]

[1]Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First Street SW, Rochester MN 55905, [2]Department of Health Sciences Research, Mayo Clinic, 4500 San Pablo Road South, Jacksonville FL 32224, [3]Department of Laboratory Medicine and Pathology, [4]Division of Hematology, Department of Internal Medicine, [5]Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester MN 55905 and [6]Department of Cancer Biology, Mayo Clinic, 4500 San Pablo Road South, Jacksonville FL 32224, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: RNA-seq has become the method of choice to quantify genes and exons, discover novel transcripts and detect fusion genes. However, reliable variant identification from RNA-seq data remains challenging because of the complexities of the transcriptome, the challenges of accurately mapping exon boundary spanning reads and the bias introduced during the sequencing library preparation.

Method: We developed RVboost, a novel method specific for RNA variant prioritization. RVboost uses several attributes unique in the process of RNA library preparation, sequencing and RNA-seq data analyses. It uses a boosting method to train a model of 'good quality' variants using common variants from HapMap, and prioritizes and calls the RNA variants based on the trained model. We packaged RVboost in a comprehensive workflow, which integrates tools of variant calling, annotation and filtering.

Results: RVboost consistently outperforms the variant quality score recalibration from the Genome Analysis Tool Kit and the RNA-seq variant-calling pipeline SNPiR in 12 RNA-seq samples using ground-truth variants from paired exome sequencing data. Several RNA-seq–specific attributes were identified as critical to differentiate true and false variants, including the distance of the variant positions to exon boundaries, and the percent of the reads supporting the variant in the first six base pairs. The latter identifies false variants introduced by the random hexamer priming during the library construction.

Availability and implementation: The RVboost package is implemented to readily run in Mac or Linux environments. The software and user manual are available at http://bioinformaticstools.mayo.edu/research/rvboost/.

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

Received on May 30, 2014; revised on August 11, 2014; accepted on August 20, 2014

1 INTRODUCTION

RNA-seq has become popular with the decreasing cost and its potential to quantify exon/transcript levels over a large dynamic range, discover novel transcripts, identify various splicing mechanisms and detect fusion genes (Costa *et al.*, 2013) (Asmann *et al.*, 2011). However, while the variant identification from DNA sequencing is becoming a routine practice, the variant detection from RNA-seq remains challenging because of the complexity of the transcriptome, the ambiguities in mapping exon boundary spanning reads and the artifacts introduced in RNA-seq library protocols (Piskol *et al.*, 2013a). Because expressed genetic variants have more immediate impact on the protein function compared with the DNA variants, we were motivated to develop a reliable RNA-seq variant prioritization method.

In general, variant detection from massive parallel sequencing data involves two steps. First is variant calling, which outputs all positions with any evidence of alternative alleles compared with reference. An essential next step is variant prioritization and filtering to obtain reliable variants of high confidence. For DNA sequencing data, the most widely used variant prioritization method is a mixture model-based classifier, variant quality score recalibration (VQSR), within the Genome Analysis Toolkit (GATK) (DePristo *et al.*, 2011). VQSR integrates multiple attributes/annotations of the variants, all of which are based on features of sequencing, including the depth of coverage, strand bias, mapping qualities and variant position bias toward the end of the reads. VQSR uses variants reported in HapMap as the training source to calculate a filtering criterion, and then predicts true 'novel' variants. Another method SNPiR proposes a series of arbitrary hard thresholds to filter and reduce the number of false variants (Piskol *et al.*, 2013b).

After careful examination of the RNA-seq variant detection process, we proposed to include RNA-specific attributes/annotations for the variant prioritization model in addition to the features included in GATK. Furthermore, we observed that the Gaussian mixture model and the parameter selection used in VQSR are not ideal for modeling these features and proposed to use a boosting method that uses generalized linear models as its base learners. This method, called the RNA Variant Boosting (RVboost), is a ranking machine to (a) train a model based on common variants in HapMap and (b) rank the variants accordingly.

*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.

We implemented a comprehensive workflow using the framework of GATK that incorporates RVboost to facilitate reliable RNA-seq variant prioritization (Fig. 1). This workflow outputs called variants with detailed annotations in both the standard variant calling format (VCF) and the tab-delimited text format. We have shown that RVboost outperforms VQSR in 12 RNA-seq samples with paired exome sequencing data. We also introduced a key concept of 'train-set quantile score', or the train-Q score, to help users determine their preferred precision/recall trade-off.

2 FEATURES

2.1 Attributes selection

After testing, six attributes were included in RVboost. We kept three attributes from GATK's Unified Genotyper, which are routinely used in VQSR: (i) Quality score over depth, (quality by depth, QD); (ii) Positional bias (ReadPosRankSum); and (iii) Fisher's exact test-based Strand bias (FS). In addition, we added three novel attributes that are specific and unique for RNA-seq: (i) the percent of variant-supporting reads with variant positions in the first six bases of the reads (PctExtPos). This is to model the false variants introduced during the random hexamer priming of the cDNA synthesis step during the RNA-seq library protocol. The mismatches allowed between the hexamer primers and the RNA templates resulted in substantial amount of false variants (Fig. 1 of the Supplementary Material); (ii) distance to the exon–intron boundary (DJ); and (iii) the uniqueness of the read mapping in the genome and transcriptome (ED). More details are available in Supplementary Section 1.1.

2.2 Input, output and major modules

RVboost takes an aligned RNA-seq BAM file [e.g. the BAM file generated by TopHat (Trapnell *et al.*, 2009)] and processes it through three major components (Fig. 1): (i) Unified Genotyper from GATK for raw variant calling in the target region and generation of the annotations including all GATK classic annotations and the three novel attributes described above in Section 2.1; (ii) annotation of each variant with additional attributes, including all functional annotations from SnpEff (Cingolani *et al.*, 2012), and whether the variant position is a known RNA-editing site according to a RNA-editing database (Ramaswami and Li, 2014); (iii) Variant prioritization and ranking using RVboost. This module includes two components: a novel boosting method to train the variant classifying model using high-confidence variants (e.g. common variants in HapMap as GATK recommended); and ranking of the likely true variants using a confidence score (details in the method section below). The output is a VCF file of all variants, with full annotations. Users can also generate a text file with selected attributes.

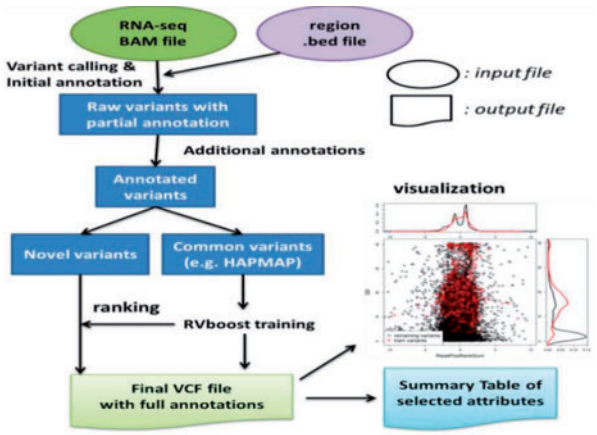

Fig. 1. The overall workflow of RVboost

2.3 Description of the variant prioritization methods

With the selection of six attributes, we formulated the variant prioritization as a ranking problem where only likely true variants (e.g. common variants from HapMap) are used for model training. We describe it as a mathematical process to find a good $F(.)$, which outputs ranking score $\mathbf{y}$ from data X with minimum error defined by a loss function $L(.)$:

$$F_{opt}(\mathbf{X}) = \arg\min_{F(\cdot)} L(\mathbf{y}, F(\mathbf{X})), \quad (1)$$

where elements of $\mathbf{y}$ is 1 or 0 to indicate likely true or false variants, respectively, $\mathbf{X}$ is a variant by attribute matrix used to rank true variants and 'arg min' stands for the argument of the minimum error of the loss function. Different from mixture model-based VQSR, in which the construction of $F(.)$ requires explicitly the number of Gaussian kernels and the percent of worst variants that are used as negative sets, we proposed to use a more flexible boosting method to rank variants. Boosting methods construct such a function $F(\mathbf{X})$ by additive combinations of M 'base' learners $h(\mathbf{X})$ (e.g. linear regression models) $F(\mathbf{X}) = \sum_{m=1}^{M} \beta_m h(\mathbf{X}; \theta_m)$, with corresponding combination coefficients β_m and parameters θ_m of m-th learner (Friedman, 2001; Bühlmann and Hothorn, 2007). By leveraging implemented boosting methods in R package 'gbm', we chose three boosting options for which response variable values range from 0 to 1: 'adaboost' with AdaBoost exponential loss function, 'bernoulli' with logistic regression loss function, and 'huberized' with modified Huber loss function (Ridgeway, 2005). It is often found that these three distribution options lead to similar performance and converge well before 20 000 iterations. Hence, we chose AdaBoost model and 20 000 iterations as default settings for RVboost.

2.4 Expected recall rate and train-set quantile score

After training, the user needs to choose a prioritization score as the threshold to call 'true variants'. In practice, it is often difficult to interpret a prioritization score derived from a complex computational model and its implications for precision/recall trade-off. To address this problem, we explicitly defined a monotonic transformation independent of ranking methods, which depends on training set of likely true variants

$$\text{train-Q}[j] = eCDF_{\text{train-set}}(score[j]) \quad (2)$$

where $score[j]$ is score generated by method from high to low indicating the likelihood of the j-th variant to be a true variant.

$eCDF_{\text{train-set}}(\cdot)$ is the empirical cumulative density function learnt from the training dataset.

The train-Q score is intuitive to users, as it directly uses the expected recall rate from the provided training dataset. For example, a cutoff of the train-Q scores of 20% means that using this cutoff, 80% of the variants within the training set will be retained, i.e. we have a 80% expected recall. Through our comparison studies, we suggest a moderate expected recall rate, e.g. 90 or 95%, instead of an aggressive 99%, which is the recommended default by VQSR.

2.5 Comparison studies

We compared specificity and sensitivity of RVboost to VQSR (GATK version 1.6.9) and SNPiR (Piskol *et al.*, 2013b), using the concordance between RNA and DNA variants from eight follicular lymphoma tumor samples and four replicates of MCF-7 cell lines (details described in Supplementary Material 1.4). To make unbiased comparisons, we evaluated the recall/precision on a subset of novel variants that (i) are not in the positive training set and (ii) have at least 10-fold coverage in both RNA-seq and exome-seq. We regarded the genotype calls from exome-seq as the ground truth and computed precision/recall accordingly, under the assumption that RNA-editing sites are a small percentage of RNA-seq variants (Piskol *et al.*, 2013a).

Overall, RVboost consistently outperforms both VQSR and SNPiR in all the tested samples in terms of AUC (Area Under the Curve) of precision/recall curves, and demonstrates superior precision in low train-Q score cutoffs, or equivalently, with high expected recall rates (details in Supplementary Material 2.2). We also investigated the contribution of individual attribute to distinguish true versus false variants, suggesting that the percent of reads supporting the variants in the first six base pairs and QD are the most informative features (details in Supplementary Material 2.3).

3 CONCLUSIONS

We developed RVboost, a software package designed to reliably prioritize and call variants from RNA-seq data. The output of our workflow provides comprehensive annotations to facilitate biological understanding. Variant prioritization is based on a proposed boosting method, which not only outperforms two other methods (SNPiR and VQSR) in overall performance, but also provides great flexibility to users for adjusting of the precision/recall trade-off, and it is superior to *ad hoc* hard-threshold approaches, such as SNPiR. The major modules are wrapped as a comprehensive package.

Funding: Support for this work was provided by gift from Everett and Jane Hauck to the Center for Individualized Medicine at Mayo Clinic Jacksonville Florida, funds from the 26.2 with Donna Foundation and the proceeds of the National Marathon to Fight Breast Cancer and NIH P50 CA097274.

Conflict of interest: none declared.

REFERENCES

Asmann,Y.W. *et al.* (2011) A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. *Nucleic Acids Res.*, **39**, e100.

Bühlmann,P. and Hothorn,T. (2007) Boosting algorithms: regularization, prediction and model fitting. *Stat. Sci.*, **22**, 477–505.

Cingolani,P. *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.

Costa,V. *et al.* (2013) RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.*, **21**, 134–142.

DePristo,M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.

Friedman,J. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.

Piskol,R. *et al.* (2013a) Lack of evidence for existence of noncanonical RNA editing. *Nat. Biotechnol.*, **31**, 19–20.

Piskol,R. *et al.* (2013b) Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.*, **93**, 641–651.

Ramaswami,G. and Li,J.B. (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.*, **42**, D109–D113.

Ridgeway,G. (2005) Generalized Boosted Models: A guide to the gbm package. In: *R CRAN package*.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.