AACR Abstract Number 240

Gene fusion calling from RNA sequencing data: An ensemble learning approach

Kenneth Brad Thomas, Yanglong Mou, Christophe Magnan, Tibor Gyuris, Eve Shinbrot, Fernando Lopez Diaz, Steven Lau-Rivera, Segun Jung, Vincent Funari, Lawrence M. Weiss.

Abstract

Introduction: Our goal is to improve gene fusion detection via RNA sequencing by combining multiple fusion callers through machine learning techniques.

Background: Gene Fusion events are important drivers of malignancy. RNA sequencing (RNAseq) methods for detection of fusions have the advantage that multiple markers can be targeted at one time. Unlike DNA methods, in which it is challenging to capture fusion breakpoints, in RNA methods fusions are readily identified through chimeric transcripts. While many fusion calling algorithms exist for use on RNAseq data, sensitive fusion callers, needed for samples of low tumor content, often present high false positive rates - a result of aligning chimeric transcripts. Further, there currently is no single feature in NGS data that can be used to filter out false positive fusion calls. In order to achieve higher accuracy in fusion calls than can be achieved using individual fusion callers, we have weighted and combined the results of multiple fusion callers by systematic and objective means: an ensemble learning approach based on random forest models. Our method selects from data generated by three independent fusion callers supplemented by metrics obtained from in-house methods. It presents a metric that can be immediately interpreted as the probability that a candidate fusion call is a true fusion call.

Methods: Random forest models were generated by use of the randomForest package in R, with tuning by the R caret package. Training data sets consisted of a balanced set of 394 fusion calls from clinical samples of solid tumors. For training, fusion calls with at least 10 supporting reads were deemed true or false based on manual review via IGV, and orthogonal methods including PCR with Sanger sequencing and the commercial Archer™ fusion CTL and Sarcoma panels. We present the results of training on data from the three well-known fusion callers Arriba, STAR-Fusion, and FusionCatcher, together with additional data from an in-house developed junction counting method, and fusion membership in a list of known fusions (a “white list”). Models were validated by 10-fold cross-validation.

Results: In performance evaluations, false positive and false negative calls were presumed false based on orthogonal determinations. On that basis, our current best model has an accuracy of 94.9% (sensitivity 93.4%, specificity 96.7%). Currently, High Confidence fusion calls (calls with probability score greater than 70%) are the most common positive calls. These have been confirmed with 100% success.

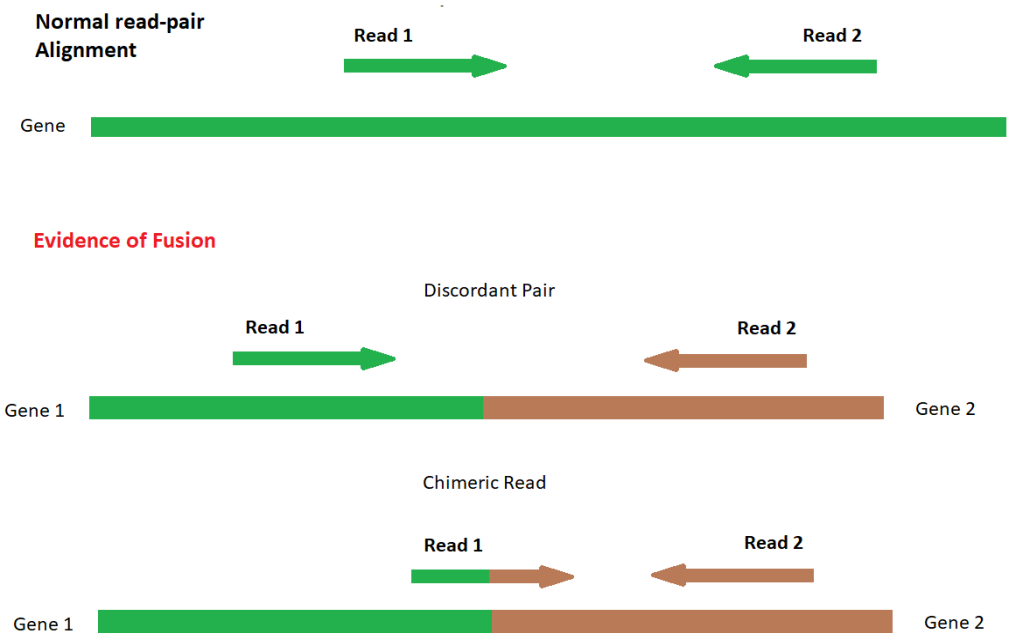
Conclusion: We have successfully integrated multiple fusion callers by means of random forest models. Our current model is validated for use on our solid tumor fusion calling pipeline.

Background

- Our work is intended to improve the utility of RNA sequencing as a means of detecting gene fusions.
- Available fusion calling algorithms were found to be sensitive, but prone to high false positive rates.
- There is a need for an alternative to time-intensive manual review as a means of improving specificity.
- To serve that need, we chose to combine metrics from multiple fusion calling algorithms that are currently available. For that purpose we chose an ensemble learning approach – random forest models.

Evidence of fusions in paired-end RNA sequencing data

Fusion calling from high throughput RNA sequencing data relies, for the most part, on two types of evidence available from paired end sequencing: Discordant read pairs and chimeric reads.



Methods (1)

- RNA sequencing data was generated by paired-end sequencing of samples from a wide variety of solid tumors from many sites. RNA templates were pulled down using capture probes of a new panel, the Solid Tumor Gene Fusion RNA-Seq panel developed by NeoGenomics. The panel includes baits for 250 genes clinically relevant for fusions in solid tumors. It was designed to target 2,232 know fusions and is capable of detecting novel fusions which may involve only one of the panel genes.
- We analyzed the RNA seq data using four distinct fusion-calling pipelines
    1. Arriba
    2. STAR-Fusion
    3. FusionCatcher
    4. An in-house-developed method for counting specific chimeric reads
- From fusion callers 1-3, we gathered metrics on the chimeric read counts, counts of discordant read pairs that were associated with candidate fusion junctions, and depth of coverage on either side of the fusion junction. Our chimeric junction read counter counted chimeric reads matching the predicted junction sequence for the candidate fusion.
- The final data set consisted of 431 candidate fusions. Of these, 147 were from defined samples (known true positive fusions). These were standard samples at several dilutions. The remainder of the samples were assigned true positive or true negative status on the basis of orthogonal confirmation by PCR/Sanger sequencing or by the commercial Archer™ fusion CTL and Sarcoma panels. The breakdown of the training set for the final model was 230 true positive / 200 true negative.

Methods (2)

- Software for modeling consisted of the R application *randomForest,* which was run within the *caret* optimization environment.
- Training employed 1500 trees (ntrees) with 10-fold cross-validation.
- Training Data comprised - candidate fusion calls – each made by at least two of the fusion callers Arrriba, STAR-Fusion and FusionCatcher. Also, using an internally-developed tool, counts for junction-associated chimeric reads were obtained for each of the candidate fusions.

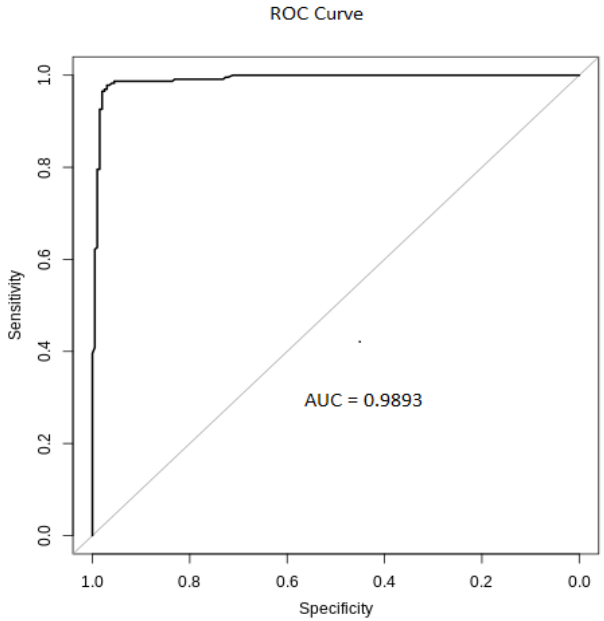
Contribution of importance to model development

The models we generate are classifiers on two classes:
F1 - the candidate fusion is a true positive (fusion)
F0 - the candidate is a false positive (not a fusion)
The models assign a probability the candidate fusion is a true positive (F1).
In training, metrics (importance) are generated that gauge the contribution of features to the final model. This is done by repeated perturbation of the model's features as the accuracy and discriminatory power (gini) of the model are tracked. If changing the value of a feature has greater impact, the score is higher. We used importance to guide the selection of variables for our models. Large scores in our final model indicate significant contribution by all of the variables:

	F0	F1	MeanDecreaseAccuracy	MeanDecreaseGini
Arriba				
split_reads1	18.45683	15.49705	21.80706	6.35341
split_reads2	23.89929	13.53787	26.02563	8.895947
coverage1	11.79315	12.28107	15.96037	3.386475
coverage2	13.28937	17.71542	21.17263	3.419395
STAR-fusion				
JunctionReadCount	35.65961	21.24654	39.50196	33.243242
SpanningFragCount	18.9742	15.90403	23.1746	10.653908
CoverageInLeftBreakpoint	10.94178	20.17112	22.69532	6.326348
CoverageInRightBreakpoint	20.96596	19.73391	26.27631	9.622533
FusionCatcher				
Spanning_pairs	32.47737	18.0114	33.66282	14.023083
Spanning_unique_reads	29.70224	11.93428	31.63528	18.495653
Junction Counter				
donor side junction count	25.21965	19.20357	29.60951	20.212577
acceptor side junction count	34.54436	25.53614	38.52344	34.614434
total junction count	39.12586	23.99656	42.68013	44.001445

Performance of our final model

Confusion Matrix			
		Reference	
		F0	F1
Prediction	F0	191	4
	F1	10	226



Accuracy : 0.9675
95% CI : (0.9461, 0.9821)
No Information Rate : 0.5336
P-Value [Acc > NIR] : <2e-16

Kappa : 0.9346

Mcnemar's Test P-Value : 0.1814

Sensitivity : 0.9826
Specificity : 0.9502
Pos Pred Value : 0.9576
Neg Pred Value : 0.9795
Prevalence : 0.5336
Detection Rate : 0.5244
Detection Prevalence : 0.5476
Balanced Accuracy : 0.9664

Conclusion

We did achieve our goal of a caller with over accuracy greater than 0.95. Our model showed sensitivity : 0.9826 and specificity : 0.9502.

More generally, we found that random forest models provide an effective, objective and defensible means of combining results from several fusion callers.

Software List

Arriba

Sebastian Uhrig, Julia Ellermann, Tatjana Walther, Pauline Burkhardt, Martina Fröhlich, Barbara Hutter, Umut H. Toprak, Olaf Neumann, Albrecht Stenzinger, Claudia Scholl, Stefan Fröhling and Benedikt Brors: Accurate and efficient detection of gene fusions from RNA sequencing data. Genome Research. Published in Advance January 13, 2021.

https://github.com/suhrig/arriba

STAR-Fusion

Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. Haas, Brian J.; Dobin, Alexander; Li, Bo; Stransky, Nicolas; Pochet, Nathalie; Regev, Aviv; Genome Biology; 2019. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1842-9

https://github.com/STAR-Fusion/STAR-Fusion

FusionCatcher

D. Nicorici, M. Satalan, H. Edgren, S. Kangaspeska, A. Murumagi, O. Kallioniemi, S. Virtanen, O. Kilkku, FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data, bioRxiv, Nov. 2014

https://github.com/ndaniel/fusioncatcher

CRAN (R) package caret

https://cran.r-project.org/web/packages/caret

CRAN (R) package randomForest

https://cran.r-project.org/web/packages/randomForest


NEO
GENOMICS