Oregon State University OSU College of Agricultural Sciences

Oregon State University

Identifying Plant Resistance Genes in Transcriptomes using Hidden Markov Models

Tyler Schappe, Andy Jones
Department of Botany & Plant Pathology, Oregon State University

Introduction

Negative Density Dependence

Negative density dependence is an ecological theory that posits that there is a stronger negative effect of growing near an individual of the same species compared to growing near an individual of a different species on a plant's growth, survival and reproduction. Past reserach has provided strong evidence for the existence of negative density dependence in forests in tropical Panama [1] as well as in temperate forests of eastern North America [2]. Further evidence has linked negative density dependence in Panamanian forests with soil organisms [3] and a recent study found that species in Panama with greater R-gene diversity experienced lower levels of negative density dependence and were more abundant regionally [4]. Thus, determining R-gene diversity for many co-existing species in a community can provide insights into how genetic diversity of trees and abundances pathogens drive tree communities in natural forest systems.

Plant resistance genes

Plant resistance genes comprise the plant adaptive immune system. Products of R-genes recognize changes within the cell caused by specific pathogen effector proteins and initiate the appropriate resistance response to combat the pathogen. The largest class of R-genes are the NBS-LRRs, the protein products of which contain nucleotide binding site (NBS) and leucine-rich repeat (LRR) motifs. R-genes are often found to contain significantly higher allelic diversity than expected from random genetic drift, suggesting that strong selection acts on them [5].

Here, we test a method for targeting plant R-genes from transcriptome assemblies, which can be used for comparative genomic analyses and to design probes for target enrichment genotyping methods.

Results

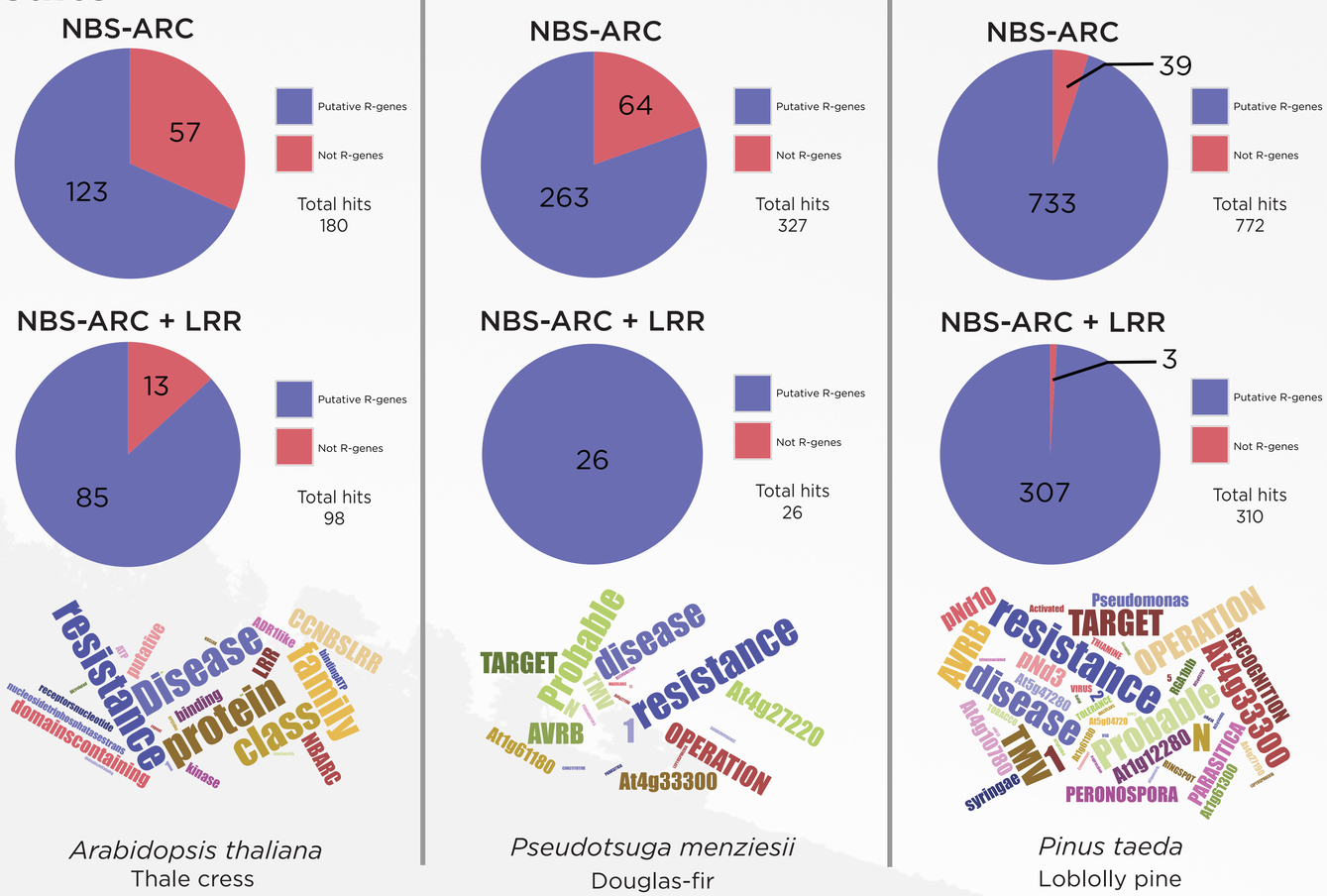

Figure 1. Pie charts showing the number of genes that matched putative plant resistance genes and that did not matcch putative plant resistance proteins in a BLASTP analysis. Results from using just the hidden Markov model for the NBS-ARC protein family are shown in the upper row of pie charts and results from using hidden Markov models for both the NBS-ARC and the LRR protein families are shown in the lower row of pie charts. The size of words in the word clouds indicate their relative proportion in the gene ontology terms resulting from a blast2GO analysis of all putative R-genes recovered from the transcriptomes. From left to right, the columns indicate results from transcriptome assemblies of *Arabidopsis thaliana*, *Pseudotsuga menziesii* and *Pinus taeda*.

Methods

Transcriptome assemblies were downloaded for 3 plant species from publicly available online resources [6, 7] and translated using TransDecoder 2.0.1 [8]. Raw hidden Markov models were downloaded from Pfam for the NBS-ARC (PF00931) and LRR_1 (PF00560) protein families [9] and either the NBS-ARC alone or both models were used to extract putative R-genes using HMMER [10]. Blast2GO was used to collect gene ontology information for each putative gene [11]. Grep was used to find the proportion of gene ontology (GO) terms that contain the word 'resistance' and a wordcloud of the GO terms was constructed for qualitative visualization purposes [12]. Pie charts were made in R [13] with the ggplot2 package [14].

Conclusions

Our hidden Markov model method was successful at finding a large number of putative plant resistance genes quickly in large transcriptomes and at minimizing the number of false positive hits. Using both the NBS-ARC and LRR_1 models resulted in fewer false positives but in the case of Douglas-fir was too stringent. A large percentage of genes that resulted from both models received strong support for orthology to known or putative plant resistance genes in protein databases, but hand-curation will still be neccesary because a small number of non-resistance genes were retained.

References

1. Comita, L. S.; Muller-Landau, H. C.; Aguilar, S.; Hubbell, S. P. Asymmetric Density Dependence Shapes Species Abundances in a Tropical Tree Community. Science 2010, 329, 330–332.
2. Johnson, D. J.; Beaulieu, W. T.; Bever, J. D.; Clay, K. Conspecific Negative Density Dependence and Forest Diversity. Science 2012, 336, 904–907.
3. Mangan, S. A.; Schnitzer, S. A.; Herre, E. A.; Mack, K. M. L.; Valencia, M. C.; Sanchez, E. I.; Bever, J. D. Negative plant-soil feedback predicts tree-species relative abundance in a tropical forest. Nature 2010, 466, 752–755.
4. Comita, L.S.; Mangan, S.A. 2016. Personal communication.
5. Marone, D. et al. Plant Nucleotide Binding Site–Leucine-Rich Repeat (NBS-LRR) Genes: Active Guardians in Host Defense Responses. Int. J. Mol. Sci. 2013, 14, 7302–7326.
6. Neale, D. B. et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 2014, 15, 1–13.
7. Berardini, T. Z.; Reiser, L.; Li, D.; Mezheritsky, Y.; Muller, R.; Strait, E.; Huala, E. The arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. genesis 2015, 53, 474–485.
8. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 2013, 8.
9. Finn, R. D. et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016, 44, D279–D285.
10. Durbin, R.; Eddy, S. R.; Krogh, A.; Mitchison, G. Biological sequence analysis: probabilistic models of proteins and nucleic acids; Cambridge university press, 1998.
11. Conesa, A.; Götz, S.; García-Gómez, J. M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005, 21, 3674–3676.
12. Davies, J. https://www.jasondavies.com/wordcloud/
13. R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/
14. Wickham, H. ggplot2; Springer New York: New York, NY, 2009.