Bioinformatics Advances, 2021, 1–3
doi: 10.1093/bioadv/vbab033
Advance Access Publication Date: 11 November 2021
Application Notes

Sequence analysis

ResCap: plant resistance gene prediction and probe generation pipeline for resistance gene sequence capture

Sandeep K. Kushwaha [1,2,*] **Inger Åhman** [1] **and Therése Bengtsson** [1]

[1]Department of Plant Breeding, Swedish University of Agricultural Sciences, Lomma 234 22, Sweden and [2]Bioinformatics, National Institute of Animal Biotechnology, Hyderabad 500 032, India

*To whom correspondence should be addressed.

Associate Editor: Alex Bateman

Received on August 5, 2021; revised on September 15, 2021; editorial decision on October 27, 2021; accepted on November 7, 2021

Abstract

Summary: The discovery of novel resistance genes (R-genes) is an important component in disease resistance breeding. Nevertheless, R-gene identification from wild species and close relatives of plants is not only a difficult but also a cumbersome process. In this study, ResCap, a support vector machine-based high-throughput R-gene prediction and probe generation pipeline has been developed to generate probes from genomic datasets. ResCap contains two integral modules. The first module identifies the R-genes and R-gene like sequences under four categories containing different domains such as TIR-NBS-LRR (TNL), CC-NBS-LRR (CNL), Receptor-like kinase (RLK) and Receptor-like proteins (RLPs). The second module generates probes from extracted nucleotide sequences of resistance genes to conduct sequence capture (SeqCap) experiments. For the validation of ResCap pipeline, ResCap generated probes were synthesized and a sequence capture experiment was performed to capture expressed resistance genes among six spring barley genotypes. The developed ResCap pipeline in combination with the performed sequence capture experiment has shown to increase precision of R-gene identification while simultaneously allowing rapid gene validation including non-sequenced plants.

Availability and implementation: The ResCap pipeline is available at http://rescap.ltj.slu.se/ResCap/

Contact: sandeep.kushwaha@slu.se or sandeep@niab.org.in

Supplementary information: Supplementary materials are available at *Bioinformatics Advances* online.

1 Introduction

Plant breeding efforts to develop resistant varieties do still mainly rely on the introgression of major dominant disease or pest resistance genes (R-genes) from resistant cultivars or from landraces through repeated backcrossing. R-genes play a key role in the recognition of specific pathogen effector molecules, leading to an induction of plant defence signalling often associated with local hypersensitive response at the infection site (McHale *et al.*, 2006). Based on current knowledge, plant R-genes can be divided into at least five major classes, such as coiled–coiled nucleotide-binding leucine-rich repeat (CNL), Toll/interleukin-1 receptor-nucleotide-binding site leucine-rich repeat (TNL), Receptor-like kinase (RLK) and Receptor-like protein (RLP), and others (Sanseverino *et al.*, 2013). One strategy to improve the efficiency and durability of resistance is to stack R-genes and precede the rapidly evolving effector genes in pathogens. However, finding of R-genes from landraces and close relatives to crops is a difficult and laborious process. In this context, the SeqCap technique can make it possible to target regions of interest, while minimizing the fraction of off-targets at a large scale. The SeqCap technique picks up nucleotide fragments of interest from genomic and transcriptomic pools through a user-designed set of probes. Recently, the sequence capture technique has been used successfully for R-gene enrichment sequencing (RenSeq) in potato (Witek *et al.*, 2016), tomato (Andolfo *et al.*, 2014; de Oliveira *et al.*, 2018) and wheat (Steuernagel *et al.*, 2016; Zhang *et al.*, 2020).

Mostly, sequence and motif similarity, domain matching and domain association-based methods are in use for resistance gene identification such as Disease Resistance Analysis and Gene Orthology (DRAGO) pipeline (Sanseverino *et al.*, 2013), R-gene analogues pipeline (RGAugury) (Li *et al.*, 2016) and NLR-parser (Steuernagel *et al.*, 2015). Prediction of R-proteins on the basis of sequence and domain similarity with a small set of reference R-genes is challenging due to the high level of diversity, as R-genes are under high selection pressure to adapt their immunity to the rapidly evolving effector genes in the pathogens (Marone *et al.*, 2013). R-gene identification from a plant species or landraces through traditional methods would be difficult to perform at large scale. But presently, a large number of plant genomes and transcriptomes have been sequenced and assembled. Despite the availability of draft genome and genome sequences, R-gene identification and validation are still difficult due to poor gene annotation model. However, machine learning techniques-based webservers and tools such as NBSPred (Kushwaha *et al.*, 2016) and DRPPP (Pal *et al.*, 2016) enabled *in silico* exploration of R-genes. However, the prediction results of these tools were never validated experimentally. Here, as an integrated solution, ResCap an automated pipeline has been developed for R-gene identification, nucleotide sequence extraction of R-gene from genome and transcriptome sequences, and probe generation to perform experimental validation.

2 Methods

R-gene and non-R-gene sequences were retrieved from public databases such as NCBI, Uniprot and PRGdb. Redundancy removal among extracted sequences was performed through clustering. A domain-based approach was used to generate the final datasets referred to as the positive and negative dataset. R-gene classes were identified among extracted sequences on basis of the occurrence of well-known R-gene domains such as NB-ARC, TIR, CC, kinase, LRR, Serine/threonine-LRR and Kinase-LRR. Sequences containing these domains are referred to as the positive dataset, whereas the negative dataset included all kind of sequences except R-gene and R-gene like sequences. Sequence compositional frequencies (amino acid frequency, dipeptide frequency, tripeptide frequency, multiplet frequency, charge and hydrophobicity composition) were calculated (Supplementary File Section S2), and all the calculated properties were gathered as a numerical feature vector for each sequence of the positive and negative dataset (Chaudhuri *et al.*, 2011; Ramana and Gupta, 2010). The SVM$^{\text{light}}$ package modules (SVM_learn and SVM_classify) (Joachims, 1999) were used to generate SVM classifier for R-gene prediction. Best binary classifiers for each family were identified through 5-fold cross-validation technique (Supplementary File Section S3). Augustus gene prediction software was used in the pipeline for the annotation of plant genome (Stanke and Morgenstern, 2005). TransDecoder (Grabherr *et al.*, 2011) was used to generate protein sequences from transcripts. The flowchart of the pipeline is given in Figure 1. For the validation of ResCap pipeline, coding sequences of plants of poaceae family from the Gramene database (Gupta *et al.*, 2016) were extracted and processed through the ResCap pipeline and generated probes were

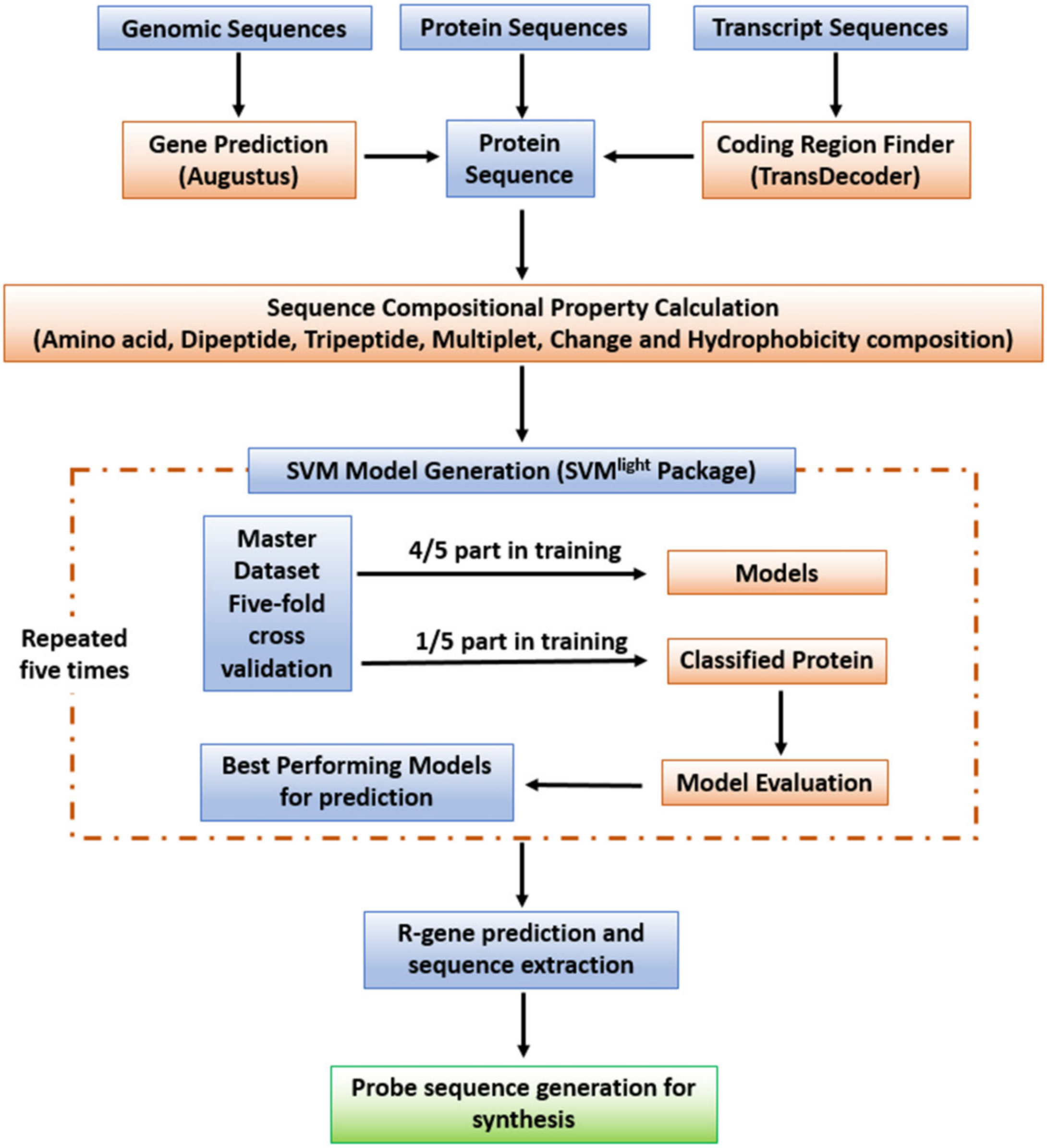
Genomic Sequences
Protein Sequences
Transcript Sequences
Gene Prediction (Augustus)
Protein Sequence
Coding Region Finder (TransDecoder)
Sequence Compositional Property Calculation
(Amino acid, Dipeptide, Tripeptide, Multiplet, Change and Hydrophobicity composition)
SVM Model Generation (SVM^{light} Package)
Repeated five times
Master Dataset Five-fold cross validation
4/5 part in training
Models
1/5 part in training
Classified Protein
Model Evaluation
Best Performing Models for prediction
R-gene prediction and sequence extraction
Probe sequence generation for synthesis

Fig. 1. ResCap data processing workflow for R-gene identification and probe generation

synthesized using SeqCap EZ HyperCap, Nimblegen, Roche, USA. Six spring barley genotypes (142-31, 142-93, 252-33, 252-61, Barke and Lina) were selected for the experimental validation (Åhman and Bengtsson, 2019).

All the genotypes were grown under highly controlled experimental conditions (Supplementary Section S5) and homogenized leaf samples were used for RNA extraction. Library preparation, sequence capture experiment and sequencing were performed at Centre for Genomic Research, University of Liverpool, UK, and bioinformatics analysis was performed at Swedish University of Agricultural Sciences, Sweden. Generated sequence data are available at NCBI SRA public repository (PRJNA740109).

3 Implementation

Dell PowerEdge T440 Server E5-2430 with 16 core processors of 2.1 GHz, running on Ubuntu 20.04 LTS was used to host ResCap pipeline, and freely accessible as a web interface which was developed in PHP version 8.0. ResCap pipeline provides email confirmation for each submission and email notification upon job completion.

4 Results and discussion

A total of 1694 (CNL: 447; TNL: 515; RLK: 355; RLP: 377) sequences were involved in the training of four classes of R-gene family. Composition-based amino acid frequencies were used for numerical encoding of training sequences (Supplementary Section S2). In order to find best classifier for each R-gene class, 1176 binary models were created through sequential input of different kernel function and kernel associated parameters for model generation. Polynomial kernel associated d and C parameters were increased stepwise through a combination of 1, 2, 3, 4 … to … 9 for the d, and 10^{-7}, 10^{-6} … to … 10^{13} for C whereas radial basis function kernel parameter gamma (g) was incremented stepwise 10^{-15} … to … 10^{3}, and parameter C from 10^{-5} … to … 10^{15} (Kushwaha *et al.*, 2016). The mean Matthews correlation coefficient and prediction accuracy of the best-performed model, kernel type and kernel associated values are provided in Supplementary File (Table S3). ResCap prediction accuracy was compared with NLR-parser (Supplementary Tables S3–S6) and ResCap has detected higher number of sequences with R-protein domains than NLR-parser. Sequence capture experiment was performed to validate ResCap generated probes. Sequence capture data of six genotypes (142-31, 142-93, 252-33, 252-61, Barke and Lina) were evaluated, and bioinformatics analysis of sequence captured data is given in Supplementary File (Tables S8 and S9). On average, approximately 5 million high-quality paired-end reads were captured for each genotype by using designed probes. Both the pairs were merged and used for BLASTn similarity search against nucleotide sequences used for probe design. Among all captured reads, 27%, 71%, 4% and 0% reads were belonging to the CNL, RLK, RLP and TNL class, respectively. R-gene classes were analysed against the barley genome to identify common and uniquely expressed R-genes among barley genotypes (Supplementary File Figure S2). ResCap pipeline will be highly useful to develop a holistic understanding of disease susceptibility and resistance in crop varieties against pests and pathogens.

Acknowledgements

The authors would like to thank the Department of Plant Breeding, Swedish University of Agricultural Sciences, and Lantmännen ek. för., Sweden for providing necessary resources to conduct this research and Dr Vehbo Hot for maintenance of the plants and aphid rearing.

Funding

The authors would like to thank the Swedish research council FORMAS (2017-00464) and fysiografen (2018) for providing funding to carry out this research.

Conflict of Interest: none declared.

References

Åhman,I. and Bengtsson,T. (2019) Introgression of resistance to *Rhopalosiphum padi* L. from wild barley into cultivated barley facilitated by doubled haploid and molecular marker techniques. *Theor. Appl. Genet.*, **132**, 1397–1408.

Andolfo,G. *et al.* (2014) Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq. *BMC Plant Biol.*, **14**, 120.

Chaudhuri,R. *et al.* (2011) FungalRV: adhesin prediction and immunoinformatics portal for human fungal pathogens. *BMC Genomics*, **12**, 192.

de Oliveira,A.S. *et al.* (2018) The Sw-5 gene cluster: tomato breeding and research toward orthotospovirus disease control. *Front. Plant Sci.*, **9**, 1055.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotech.*, **29**, 644–652.

Gupta,P. *et al.* (2016) Gramene database: navigating plant comparative genomics resources. *Curr. Plant Biol.*, **7–8**, 10–15.

Joachims,T. (1999) Making large-scale support vector machine learning practical. In: Burges, C.J.C., Schölkopf B., and Smola, A.J. (ed.) *Advances in Kernel Methods*. MIT Press, pp. 169–184.

Kushwaha,S.K. *et al.* (2016) NBSPred: a support vector machine-based high--throughput pipeline for plant resistance protein NBSLRR prediction. *Bioinformatics*, **32**, 1223–1225.

Li,P. *et al.* (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genomics*, **17**, 852.

Marone,D. *et al.* (2013) Plant nucleotide binding site–leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. *Int. J. Mol. Sci.*, **14**, 7302–7326.

McHale,L. *et al.* (2006) Plant NBS-LRR proteins: adaptable guards. *Genome Biol.*, **7**, 212.

Pal,T. *et al.* (2016) DRPPP: a machine learning based tool for prediction of disease resistance proteins in plants. *Comput. Biol. Med.*, **78**, 42–48.

Ramana,J. and Gupta,D. (2010) FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. *PLoS One*, **5**, e9695.

Sanseverino,W. *et al.* (2013) PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. *Nucleic Acids Res.*, **41**, D1167–D1171.

Stanke,M. and Morgenstern,B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, **33**, W465–W467.

Steuernagel,B. *et al.* (2015) NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics*, **31**, 1665–1667.

Steuernagel,B. *et al.* (2016) Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nat. Biotechnol.*, **34**, 652–655.

Witek,K. *et al.* (2016) Accelerated cloning of a potato late blight–resistance gene using RenSeq and SMRT sequencing. *Nat. Biotechnol.*, **34**, 656–660.

Zhang,J. *et al.* (2020) How target-sequence enrichment and sequencing (TEnSeq) pipelines have catalyzed resistance gene cloning in the wheat-rust pathosystem. *Front. Plant Sci.*, **11**, 678.