

NLRome dataset from 124 genomes of plants in the Solanaceae family

Yu Sugihara, AmirAli Toghani, Sophien Kamoun, Jiorgos Kourelis

The Sainsbury Laboratory, University of East Anglia, Norwich Research Park, UK

We report a dataset of 66,665 NLR immune receptor sequences extracted using NLRtracker from the proteomes of 124 genomes of plants in the Solanaceae family. These include samples from the following genera: *Solanum* (108 genomes, 30 species), *Physalis* (1, 1), *Capsicum* (7, 3), and *Nicotiana* (8, 5). This dataset includes NLR protein sequences and their domain annotations. The dataset is made available to the community as an open access resource as part of the OpenPlantNLR initiative.

Introduction

Nucleotide-binding domain and leucine-rich repeat (NLR) immune receptors represent the primary class of intracellular resistance genes in plants, playing a pivotal role in pathogen defence (Kourelis & van der Hoorn, 2018). Plant NLRs are characterised by a central NB-ARC (nucleotide-binding adaptor shared by Apaf-1, certain *R* genes, and CED-4) domain and typically contain a C-terminal LRR (leucine-rich repeat) domain (Kourelis et al., 2021; Takken & Goverse, 2012). The NLRtracker pipeline, leverages these and other canonical NLR features to extract and functionally annotates NLR sequences from plant proteomes, thereby enabling comprehensive phylogenetic and evolutionary studies across plant species (Kourelis et al., 2021).

In the Solanaceae family, the NLR-REQUIRED FOR CELL DEATH (NRC) clade represents a massively expanded group of NLRs (Goh et al., 2023; Sakai et al., 2023; Wu et al., 2017). Our team along with our collaborators have been studying the diversity and evolution of the NRC family. In particular, the evolutionary transitions and functional diversification of NLRs in the Solanaceae family have remained largely unexplored. To bridge this knowledge gap, we extracted and annotated 66,665 NLRs from 124 curated proteomes from four genera in the Solanaceae family. The sequences and functional annotations deposited here are publicly accessible, providing a comprehensive resource for analysing both conserved and variable regions, thereby offering insights into their evolutionary dynamics and functional roles in plant immunity.

Materials and methods

Dataset curation and NLR extraction

We curated a dataset comprising 124 publicly available proteomes from the Solanaceae family, including 108 *Solanum*, 1 *Physalis*, 7 *Capsicum*, and 8 *Nicotiana* proteomes (detailed in **Table 1**). Proteome sequences for these genomes were directly downloaded for analysis when available. In instances where transcriptome or proteome files were absent, we employed gffread (Pertea & Pertea, 2020) to extract sequences directly from genome annotations. For transcriptome files, nucleotide sequences were translated into proteins. During this process, we applied a filtering step to exclude sequences not in multiples of three, ensuring the preservation of reading frame integrity. Additionally, all stop codons were removed from the dataset to enhance the identification of potential NLR sequences.

To extract and functionally annotate NLRs from these 124 Solanaceae proteomes, we used the NLRtracker pipeline (Kourelis et al., 2021). NLRtracker is designed to specifically extract and annotate NLRs, using predefined criteria based on NLR domain architecture and sequence features. An updated description of the NLRtracker pipeline's capabilities and methodologies is available at <https://github.com/slt666666/NLRtracker> (**Dataset 1**).

Results and Conclusions

Using NLRtracker (Kourelis et al., 2021), we obtained 66,665 NLRs from 124 curated proteomes representing 39 species from four genera of Solanaceae (**Table 1, Dataset 1**). This dataset enables analyses of conserved and variable regions, as well as the identification of evolutionary transitions in NLR proteins.

Table 1: Summary of 124 Solanaceae genomes and NLRtracker results.

Dataset 1: NLRtracker outputs for 124 proteomes of 39 species of Solanaceae.

References

- Goh, F.-J., Huang, C.-Y., Derevnina, L., & Wu, C.-H. (2023). NRC immune receptor networks show diversified hierarchical genetic architecture across plant lineages. *bioRxiv*. <https://doi.org/10.1101/2023.10.25.563953>
- Kourelis, J., Sakai, T., Adachi, H., & Kamoun, S. (2021). RefPlantNLR is a comprehensive collection of experimentally validated plant disease resistance proteins from the NLR family. *PLOS Biology*, 19(10), e3001124. <https://doi.org/10.1371/journal.pbio.3001124>
- Kourelis, J., & van der Hoorn, R. A. L. (2018). Defended to the Nines: 25 Years of Resistance Gene Cloning Identifies Nine Mechanisms for R Protein Function. *The Plant Cell*, 30(2), 285–299. <https://doi.org/10.1105/tpc.17.00579>
- Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, 9, 304. <https://doi.org/10.12688/f1000research.23297.1>
- Sakai, T., Martinez-Anaya, C., Contreras, M. P., Kamoun, S., Wu, C.-H., & Adachi, H. (2023). The NRC0 gene cluster of sensor and helper NLR immune receptors is functionally conserved across asterid plants. *bioRxiv*. <https://doi.org/10.1101/2023.10.23.563533>
- Takken, F. L., & Goverse, A. (2012). How to build a pathogen detector: Structural basis of NB-LRR function. *Current Opinion in Plant Biology*, 15(4), 375–384. <https://doi.org/10.1016/j.pbi.2012.05.001>
- Wu, C.-H., Abd-El-Haliem, A., Bozkurt, T. O., Belhaj, K., Terauchi, R., Vossen, J. H., & Kamoun, S. (2017). NLR network mediates immunity to diverse plant pathogens. *Proceedings of the National Academy of Sciences*, 114(30), 8113–8118. <https://doi.org/10.1073/pnas.1702041114>