

CanPests V1.0: A Reference Dataset for Arthropod Pests of Canada Integrating DNA Barcodes

Sameer Padhye *, Chris Ho, Dirk Steinke  and Paul D. N. Hebert 

Centre for Biodiversity Genomics, University of Guelph, Guelph, ON N1G 2W1, Canada; hoc@uoguelph.ca (C.H.); dsteinke@uoguelph.ca (D.S.); phebert@uoguelph.ca (P.D.N.H.)

* Correspondence: spadhye@uoguelph.ca

Abstract

Arthropod pest species represent a serious threat to agriculture and forestry. Canada is no exception, with over 1200 recorded pest species. Although a consolidated dataset would benefit research, management, and policy, information on these species has never been compiled into a unitary database. This publication merges available information to create a dataset for the agroforest pest insects of Canada and updates the taxonomy for these records based on the Catalog of Life and GBIF. Each species record includes (1) their global distribution, (2) feeding guild for adults and larval stages (when available), (3) host plant data (when available) and (4) identifiers to obtain DNA barcodes from the Barcode of Life Data Systems (when available).

Dataset: <https://doi.org/10.5281/zenodo.15684684>.

Dataset License: CC BY 4.0

Keywords: insects; mites; DNA sequences; COI; distribution

1. Summary

Approximately 9000 species of arthropods are known to damage crops worldwide; they cause substantial losses in crop and lumber production through direct damage and by transmitting diseases [1–6]. In aggregate, they reduce global crop production by 18–26%, creating an annual loss of \$470 billion [7]. Severe outbreaks can also hinder tree growth and survival [3,8]. Although many pests are native species, invasive arthropods pose additional challenges [9]. Human activities have facilitated the dispersal of such non-native pest species [4,6,10–12] or forced native species to become pests [13].

Over 1200 insect pest species are known in Canada [10,13], but estimates suggest the actual total may exceed 2000 [10]. Prominent examples include the emerald ash borer (*Agrilus planipennis* (Fairmaire, 1888), *Lygus lineolaris* (Palisot de Beauvois, 1818), and *Lygus keltoni* (Schwartz & Footitt, 1998)), all of which are causing significant damage to forests and crops [5,6,11,14]. Several checklists are available for pests and introduced species as well as species with the potential to establish a persistent population in Canada [10,12,15] in addition to entries in the Global Register of Introduced and Invasive Species (GRIIS) [16]. However, these inventories lack trait information such as feeding guild [17,18] and host plant/s information, which could aid evaluation of their interactions and the design of potential pest management strategies [17–21].



Academic Editor: Juraj Gregaň

Received: 11 February 2026

Revised: 11 March 2026

Accepted: 12 March 2026

Published: 17 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Although many pest species have been registered, morphological identifications are often difficult, as taxonomic specialists are scarce and keys are lacking for some life stages. DNA barcoding [22] has proven an effective solution to taxonomic uncertainty, offering a faster and more cost-effective way to assess terrestrial biodiversity [23–26]. However, its use requires comprehensive DNA barcode reference libraries. Despite their usefulness for efficient and economic pest monitoring, they are often limited in scope [27].

This study generates an inclusive dataset for Canadian arthropod pest species (native and non-native) and pest species that are regulated by Canadian authorities and related to agroforestry. It consolidates various data sources and adds new aspects to the resulting reference library. Specifically, we provide new information on (a) feeding guilds and host plant(s), (b) DNA barcodes, and (c) detailed country and continental scale occurrences for each species.

2. Data Description

The CanPests dataset includes three TSV files containing all text-based fields, a .doc file, and an R script to extract and transform data from important references (for field definitions refer to Supplementary Data). Table S1 (14 columns) provides the species list with taxonomic classification, feeding guild information (for adult and larval stage), host plant information and its ‘Global Biodiversity Information Facility’ (GBIF) and ‘Catalog Of Life’ (COL) ids. Table S2 (5 columns) includes country and continent-based geographic coverage of the dataset for all species where relevant information is available. Tables S3 and S4 provide important references that were used to obtain the data. The TSV files can be easily opened with any spreadsheet software, Business Intelligence tool or Integrated Development Environment (IDE).

The dataset comprises 679 species from 450 genera, belonging to 10 different arthropod orders. Hemiptera, Lepidoptera, and Coleoptera are the three largest orders based on species richness (Figure 1). The coleopteran family Curculionidae ($n = 89$) and the hemipteran family Aphididae contribute to most species ($n = 114$). Adult feeding guild information is available for 485 species (71%) (Figure 2). Host plant information is available for 145 species (21%), while DNA barcodes are available for 594 species (794 BINs) (~88%) (as of 8 October 2025). BOLD records for 429 species (501 BINs) (~63%) are available from Canada (Figure 3) while records for 160 species are available from other countries. In total, 85 species are reported from six continents, while five are reported from seven continents (*Pieris rapae* (Linnaeus, 1758), *Plutella xylostella* (Linnaeus, 1758), *Agrotis ipsilon* (Hufnagel, 1766), *Tetranychus urticae* (C. L. Koch, 1836) and *Acyrtosiphon pisum* (Harris, 1776)). Thirteen species are reported from over 100 countries, while three species (*Vanessa cardui* (Linnaeus, 1758), *Aphis gossypii* (Glover, 1877), *Aphis craccivora* (C.L.Koch, 1854)) are reported from over 150 countries.

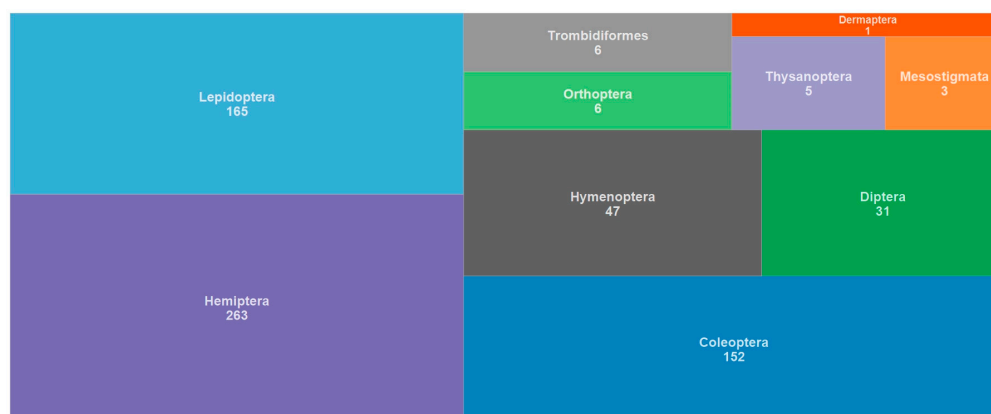


Figure 1. Number of pest species in the 10 arthropod orders of the CanPests dataset.

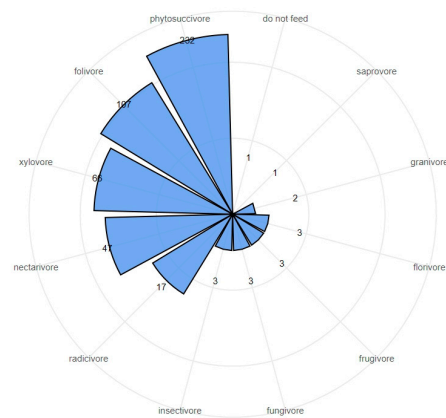


Figure 2. Species richness for different feeding guilds (adults).

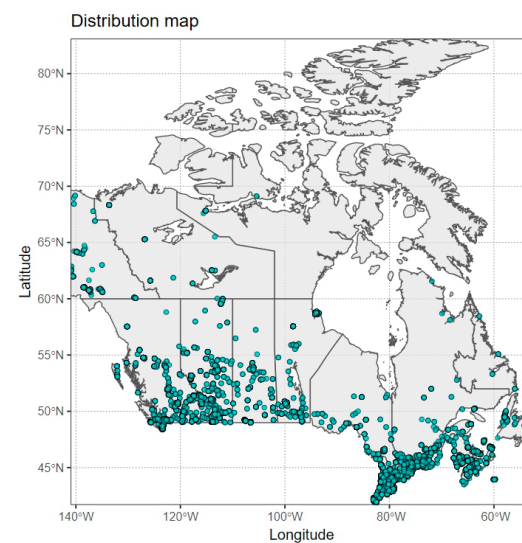


Figure 3. Map showing collection locations for the records of pest species that have been barcoded from Canada (data points as of 1 October 2025).

3. Methods

Published datasets and research articles related to “pest” species of Canada were downloaded from Google Scholar or governmental websites. We also queried GBIF for datasets specifically pertaining to these search terms. We used the R package ‘tabulapdf’ [28] to extract tables from PDF files. Species names from some resources were manually added to the list as well. Datasets were harmonized using custom R scripts (available via Zenodo) that utilized the ‘dplyr’ and ‘tidyr’ packages [29,30]. Common features included adding/removing columns, removing trailing spaces, adjusting text with respect to cases (lower vs. upper), removing instances where a species name was either unavailable or inconsistent (sp. vs. sp1/2/3, etc.), and adding/updating the relevant terms used in categorical columns (feeding guilds, country codes) and merging them. This was followed by aligning the species taxonomy with the Catalog of Life (COL) by using the ‘get_sp_ids’ and ‘get_sp_names’ functions from the ‘taxadb’ package [31] and the Global Biodiversity Information Facility (GBIF) (especially where Catalog of Life could not provide the name) using the ‘name_backbone_checklist’ function of the ‘rgbif’ package [32]. We also added the ‘GBIF usage key’ and the ‘COL id’ in the data to distinguish the information obtained from the two resources.

Updated names were replaced in the ‘species’ column of the ‘species_information’ dataset while the names given in the original references were retained in the ‘origi-

nal_ref_name' column in the same dataset. Please note that COL id is not an exhaustive list since it uses the database created internally by the package. This will be updated as and when the package database is updated [31]. Feeding guild information was initially adopted from [12] and expanded by incorporating information on adult as well as larval feeding guilds from various sources (references provided in the dataset). Host plant information was obtained from [33] and other published literature (references provided in the dataset). This information is currently not exhaustive, but data will be updated regularly. The names are either scientific or common names based on the resources used. Publicly available information on the pest species was compiled as a dataset (DSCANPESTS) from the Barcode of Life Data System (BOLD) data repository. It contains DNA sequence-based information along with metadata such as taxonomy and the Barcode Identification Number (BIN), geography, attributions, etc., for every record [34–36]. We first searched and selected all the Canadian records available for these species and expanded the search to North America and other countries for the remaining species not yet reported from Canada. In addition, only records that had (a) base pair count of 500 or more and (b) no flags (stop codons, indels etc.) were retained in the final dataset. The dataset is available for download either by the BOLD portal [37] or via the 'BOLDconnectR' package [38]. To obtain the global distribution data for each species, we first used the R script by Waller and Grosjean [39], specifically the 'occ_download' function, to schedule downloads for very large datasets from the Global Biodiversity Information Facility (GBIF) (full references of GBIF datasets available in the 'additional_references' doc file on Zenodo as references 13 and 14 respectively). The data was imported into R using the 'vroom' package [40] that processes large data in an efficient manner. It was then converted into an in-memory database (DuckDB) using the 'duckdb' package [41] for easier downstream data wrangling. Continent data was added by using the 'country_codes' function from the 'geodata' package [42] and then joined with the main dataset using the 'country_codes' column. The resulting dataset was then summarized for each species to get country and continental distribution, respectively.

Some of the reference datasets ([43,44]) include species that are not currently reported to occur in Canada but are regulated to prevent infestations. We retained these species in our list to ensure that DNA barcode information can be used in regulatory contexts.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/data11030060/s1>.

Author Contributions: Conceptualization: S.P. and D.S. Data curation: S.P., D.S. and C.H. Formal analysis: S.P. and D.S. Funding acquisition: P.D.N.H. Investigation: S.P. and D.S. Validation: S.P. and C.H. Visualization: S.P. Writing—original draft: S.P. and D.S. Writing—review and editing: P.D.N.H. and C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This study was enabled by support to P.D.N.H. from the Government of Canada through Genome Canada and Ontario Genomics (OGI-208 and OGI-233), and by the New Frontiers in Research Fund (NFRFT-2020-00073). Infrastructure required to carry out the work was acquired with grants from the Canada Foundation for Innovation (CFI) and the Gordon and Betty Moore Foundation. A Major Science Infrastructure award (MSI 42450) from CFI sustains the CBG's informatics and sequencing platforms.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data is available on Zenodo.

Acknowledgments: The authors thank the anonymous reviewers for their comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pimentel, D. Pesticides and Pest Control. In *Integrated Pest Management: Innovation-Development Process*; Peshin, R., Dhawan, A.K., Eds.; Springer: Dordrecht, The Netherlands, 2009; pp. 83–87.
2. Lehmann, P.; Ammunét, T.; Barton, M.; Battisti, A.; Eigenbrode, S.D.; Jepsen, J.U.; Kalinkat, G.; Neuvonen, S.; Niemelä, P.; Terblanche, J.S.; et al. Complex Responses of Global Insect Pests to Climate Warming. *Front. Ecol. Environ.* **2020**, *18*, 141–150. [\[CrossRef\]](#)
3. Canelles, Q.; Aquilué, N.; James, P.M.A.; Lawler, J.; Brotons, L. Global Review on Interactions between Insect Pests and Other Forest Disturbances. *Landsc. Ecol.* **2021**, *36*, 945–972. [\[CrossRef\]](#)
4. Poland, T.M.; Patel-Weynand, T.; Finch, D.M.; Miniati, C.F.; Hayes, D.C.; Lopez, V.M. Correction to: Invasive Species in Forests and Rangelands of the United States. In *Invasive Species in Forests and Rangelands of the United States*; Poland, T.M., Patel-Weynand, T., Finch, D.M., Miniati, C.F., Hayes, D.C., Lopez, V.M., Eds.; Springer International Publishing: Cham, The Netherlands, 2021; p. C1.
5. Thurston, G.S.; Slater, A.; Nei, I.; Roberts, J.; McLachlan Hamilton, K.; Sweeney, J.D.; Kimoto, T. New Canadian and Provincial Records of Coleoptera Resulting from Annual Canadian Food Inspection Agency Surveillance for Detection of Non-Native, Potentially Invasive Forest Insects. *Insects* **2022**, *13*, 708. [\[CrossRef\]](#)
6. Cárcamo, H.; Tansey, J.; Beres, B.; Catton, H.A.; Tidemann, B.; Reid, P.; Vankosky, M.A. Trap Cropping for Insect Pests in the Canadian Prairies: A Review and a Case Study. *Arthropod-Plant Interact.* **2024**, *18*, 1201–1218. [\[CrossRef\]](#)
7. Culliney, T.W. Crop Losses to Arthropods. In *Integrated Pest Management*; Pimentel, D., Peshin, R., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 201–225.
8. Van Lierop, P.; Lindquist, E.; Sathyapala, S.; Franceschini, G. Global Forest Area Disturbance from Fire, Insect Pests, Diseases and Severe Weather Events. *For. Ecol. Manag.* **2015**, *352*, 78–88. [\[CrossRef\]](#)
9. Tamburello, N.; Litt, M.A. Multiple Impacts of Invasive Species on Species at Risk: A Case Study in British Columbia, Canada. *FACETS* **2023**, *8*, 1–13. [\[CrossRef\]](#)
10. Langor, D.W.; DeHaas, L.J.; Footitt, R.G. Diversity of Non-Native Terrestrial Arthropods on Woody Plants in Canada. In *Ecological Impacts of Non-Native Invertebrates and Fungi on Terrestrial Ecosystems*; Langor, D.W., Sweeney, J., Eds.; Springer: Dordrecht, The Netherlands, 2009; pp. 5–19.
11. Dukes, J.S.; Pontius, J.; Orwig, D.; Garnas, J.R.; Rodgers, V.L.; Brazee, N.; Cooke, B.; Theoharides, K.A.; Stange, E.E.; Harrington, R.; et al. Responses of Insect Pests, Pathogens, and Invasive Plant Species to Climate Change in the Forests of Northeastern North America: What Can We Predict? This Article Is One of a Selection of Papers from NE Forests 2100: A Synthesis of Climate Change Impacts on Forests of the Northeastern US and Eastern Canada. *Can. J. For. Res.* **2009**, *39*, 231–248. [\[CrossRef\]](#)
12. Nealis, V.G.; DeMerchant, I.; Langor, D.; Noseworthy, M.K.; Pohl, G.; Porter, K.; Shanks, E.; Turnquist, R.; Waring, V. Historical Occurrence of Alien Arthropods and Pathogens on Trees in Canada. *Can. J. For. Res.* **2016**, *46*, 172–180. [\[CrossRef\]](#)
13. Langor, D.W.; Cameron, E.K.; MacQuarrie, C.J.K.; McBeath, A.; McClay, A.; Peter, B.; Pybus, M.; Ramsfield, T.; Ryall, K.; Scarr, T.; et al. Non-Native Species in Canada's Boreal Zone: Diversity, Impacts, and Risk. *Environ. Rev.* **2014**, *22*, 372–420. [\[CrossRef\]](#)
14. Howard, R.J.; Garland, J.A.; Seaman, W.L.; Grafius, E.J. Diseases and pests of vegetable crops in Canada. *J. Econ. Entomol.* **1996**, *89*, 1045.
15. Parker, T.J.; Clancy, K.M.; Mathiasen, R.L. Interactions among Fire, Insects and Pathogens in Coniferous Forests of the Interior Western United States and Canada. *Agric. For. Entomol.* **2006**, *8*, 167–189. [\[CrossRef\]](#)
16. Pagad, S.; Bisset, S.; Genovesi, P.; Groom, Q.; Hirsch, T.; Jetz, W.; Ranipeta, A.; Schigel, D.; Sica, Y.V.; McGeoch, M.A. Country Compendium of the Global Register of Introduced and Invasive Species. *Sci. Data* **2022**, *9*, 391. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Wardhaugh, C.W.; Stork, N.E.; Edwards, W. Feeding Guild Structure of Beetles on Australian Tropical Rainforest Trees Reflects Microhabitat Resource Availability. *J. Anim. Ecol.* **2012**, *81*, 1086–1094. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Naveena, N.L.; Subramanya, S.; Setty, S. Guild Structure of Stored Grain Insects Reflects Food Resource Availability in Tropical Forest Ecosystems. *J. Stored Prod. Res.* **2022**, *96*, 101953. [\[CrossRef\]](#)
19. Peeters, P.J.; Read, J.; Sanson, G.D. Variation in the Guild Composition of Herbivorous Insect Assemblages among Co-occurring Plant Species. *Austral Ecol.* **2001**, *26*, 385–399. [\[CrossRef\]](#)
20. Vanderwel, M.C.; Malcolm, J.R.; Smith, S.M.; Islam, N. Insect Community Composition and Trophic Guild Structure in Decaying Logs from Eastern Canadian Pine-Dominated Forests. *For. Ecol. Manag.* **2006**, *225*, 190–199. [\[CrossRef\]](#)
21. Novotny, V.; Miller, S.E.; Hreck, J.; Baje, L.; Basset, Y.; Lewis, O.T.; Stewart, A.J.A.; Weiblen, G.D. Insects on Plants: Explaining the Paradox of Low Diversity within Specialist Herbivore Guilds. *Am. Nat.* **2012**, *179*, 351–362. [\[CrossRef\]](#)
22. Hebert, P.D.N.; Cywinska, A.; Ball, S.L.; deWaard, J.R. Biological Identifications through DNA Barcodes. *Proc. R. Soc. Lond. B* **2003**, *270*, 313–321. [\[CrossRef\]](#)
23. Floyd, R.; Lima, J.; deWaard, J.; Humble, L.; Hanner, R. Common Goals: Policy Implications of DNA Barcoding as a Protocol for Identification of Arthropod Pests. *Biol. Invasions* **2010**, *12*, 2947–2954. [\[CrossRef\]](#)
24. Ball, S.L.; Armstrong, K.F. Rapid, One-Step DNA Extraction for Insect Pest Identification by Using DNA Barcodes. *J. Econ. Entomol.* **2008**, *101*, 523–532. [\[CrossRef\]](#) [\[PubMed\]](#)

25. Mitchell, A.; Gopurenko, D. DNA Barcoding the Heliothinae (Lepidoptera: Noctuidae) of Australia and Utility of DNA Barcodes for Pest Identification in *Helicoverpa* and Relatives. *PLoS ONE* **2016**, *11*, e0160895. [CrossRef]
26. Ashfaq, M.; Hebert, P.D.N. DNA Barcodes for Bio-Surveillance: Regulated and Economically Important Arthropod Plant Pests. *Genome* **2016**, *59*, 933–945. [CrossRef]
27. Steinke, D.; Ashfaq, M.; Ho, C.Y.; Perez, K.H.J.; Sones, J.E.; DeWaard, S.L.; DeWaard, J.R.; Ratnasingham, S.; Zakharov, E.V.; Hebert, P.D.N. Farmland Biodiversity Monitoring Using DNA Metabarcoding. *Diversity* **2025**, *17*, 585. [CrossRef]
28. Leeper, T.J.; Vargas Sepulveda, M.; Paskhalis, T. *tabulapdf: Extract Tables from PDF Documents*, R package version 1.0.5-5; 2024. Available online: <https://cran.r-project.org/web/packages/tabulapdf/index.html> (accessed on 11 March 2026). [CrossRef]
29. Wickham, H.; Vaughan, D.; Girlich, M. *tidyr: Tidy Messy Data*, R package version 1.3.2; 2014. Available online: <https://cran.r-project.org/web/packages/tidyr/index.html> (accessed on 11 March 2026). [CrossRef]
30. Wickham, H.; François, R.; Henry, L.; Müller, K.; Vaughan, D. *dplyr: A Grammar of Data Manipulation*, R package version 1.2.0; 2014. Available online: <https://cran.r-project.org/web/packages/dplyr/index.html> (accessed on 11 March 2026). [CrossRef]
31. Norman, K.E.A.; Chamberlain, S.; Boettiger, C. Taxadb: A High-performance Local Taxonomic Database Interface. *Methods Ecol. Evol.* **2020**, *11*, 1153–1159. [CrossRef]
32. Chamberlain, S.; Oldoni, D.; Waller, J. *rgbif: Interface to the Global Biodiversity Information Facility API*, R package version 3.8.4; 2012. Available online: <https://cran.r-project.org/web/packages/rgbif/rgbif.pdf> (accessed on 1 September 2025). [CrossRef]
33. Beirne, B.P. Pest Insects of Annual Crop Plants in Canada: Part I, Lepidoptera.; II, Diptera.; III, Coleoptera. *Mem. Entomol. Soc. Can.* **1971**, *103*, 1–124. [CrossRef]
34. Ratnasingham, S.; Hebert, P.D.N. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE* **2013**, *8*, e66213. [CrossRef] [PubMed]
35. Ratnasingham, S.; Hebert, P.D.N. Bold: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol. Ecol. Notes* **2007**, *7*, 355–364. [CrossRef]
36. Ratnasingham, S.; Wei, C.; Chan, D.; Agda, J.; Agda, J.; Ballesteros-Mejia, L.; Boutou, H.A.; El Bastami, Z.M.; Ma, E.; Manjunath, R.; et al. BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data. In *DNA Barcoding*; DeSalle, R., Ed.; Methods in Molecular Biology; Springer: New York, NY, USA, 2024; Volume 2744, pp. 403–441.
37. Dataset: DS-CANPESTS. Available online: <https://doi.org/10.5883/DS-CANPESTS> (accessed on 1 December 2025). [CrossRef]
38. Padhye, S.; Ballesteros-Mejia, L.; Ratnasingham, S. *BOLDconnectR: Retrieve, Transform and Analyze the Barcode of Life Data Systems Data*, R package version 1.0.0; 2025. Available online: <https://cran.r-project.org/web/packages/BOLDconnectR/index.html> (accessed on 11 March 2026). [CrossRef]
39. Downloading Occurrences from a Long List of Species in R and Python. Available online: <https://data-blog.gbif.org/post/downloading-long-species-lists-on-gbif/> (accessed on 1 September 2025).
40. Hester, J.; Wickham, H.; Bryan, J. *vroom: Read and Write Rectangular Text Data Quickly*, R package version 1.7.0; 2019. Available online: <https://cran.r-project.org/web/packages/vroom/vroom.pdf> (accessed on 1 September 2025). [CrossRef]
41. Mühleisen, H.; Raasveldt, M. *duckdb: DBI Package for the DuckDB Database Management System*, R package version 1.4.4; 2020. Available online: <https://cran.r-project.org/web/packages/duckdb/index.html> (accessed on 1 September 2025). [CrossRef]
42. Hijmans, R.J. *geodata: Access Geographic Data*, version 0.6-6; 2021. Available online: <https://cran.r-project.org/web/packages/geodata/index.html> (accessed on 1 September 2025). [CrossRef]
43. Philip, H.; Floate, K.; Mori, B. *Field Crop and Forage Pests and Their Natural Enemies in Western Canada: Identification and Management*; Agriculture and Agri-Food Canada: Ottawa, ON, Canada, 2018; pp. 1–148.
44. List of Pests Regulated by Canada. Available online: <https://inspection.canada.ca/en/plant-health/invasive-pests-and-plants/regulated-pests> (accessed on 1 August 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.