MissMap: A pipeline for visualizing sequence data availability in plant clades.

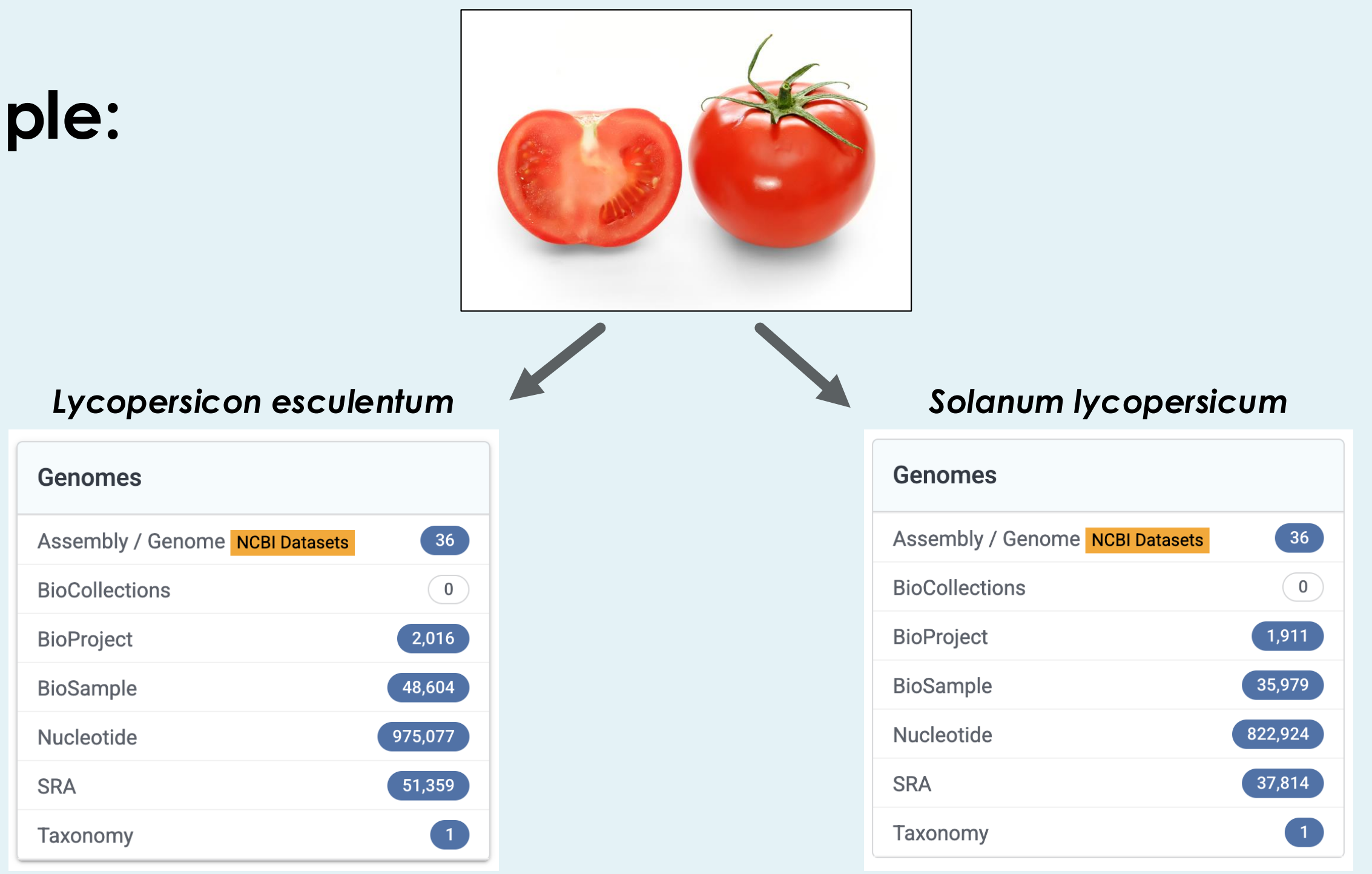
Linda J. Mansour, Joseph F. Walker
Biological Sciences Department, University of Illinois at Chicago, IL USA

UIC UNIVERSITY OF ILLINOIS AT CHICAGO

The Problem

Taxonomy is rapidly changing and presents a challenge for databases that are not always up to date with the most recent changes, or if they are up to date, not all the molecular data gets changed.

Example:



Lycopersicon esculentum

Genomes

Assembly / Genome NCBI Datasets	36
BioCollections	0
BioProject	2,016
BioSample	48,604
Nucleotide	975,077
SRA	51,359
Taxonomy	1

Solanum lycopersicum

Genomes

Assembly / Genome NCBI Datasets	36
BioCollections	0
BioProject	1,911
BioSample	35,979
Nucleotide	822,924
SRA	37,814
Taxonomy	1

In theory, using synonyms should yield identical results, but that's not always the case, and in many instances, you end up with a large set of the same results, along with some different ones. So why are they different, and are you missing information by only searching for one?

Addressing the Problem

The introduction of large language models (LLMs) has changed the way biological data can be summarized and queried. While LLMs have limitations and may be error prone, they excel at processing large amounts of text data and summarizing relevant information (e.g., doomharvest). Using MissMap, users can query taxa, which will utilize the NCBI Taxonomy Database and then LLMs to identify historical synonyms. The tool will summarize the results, determine the similarities and differences, and provide insights into the historical context behind taxonomic synonyms.

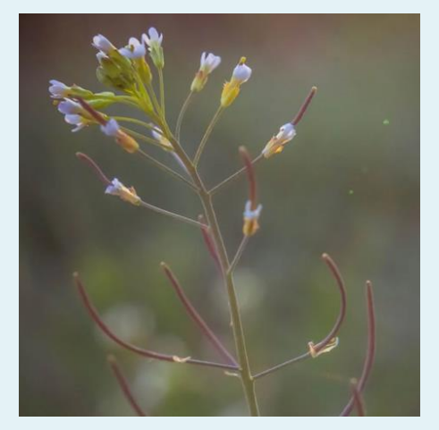
Objective & Research Question

Develop a tool that can provide all available molecular data for a given list of taxa or a specific taxonomic rank

To what extent do taxonomic synonyms affect the molecular data obtained from public databases such as NCBI?

Species Tested

Solanum lycopersicum


Arabidopsis thaliana

Zea mays

Oryza sativa

Lemna gibba

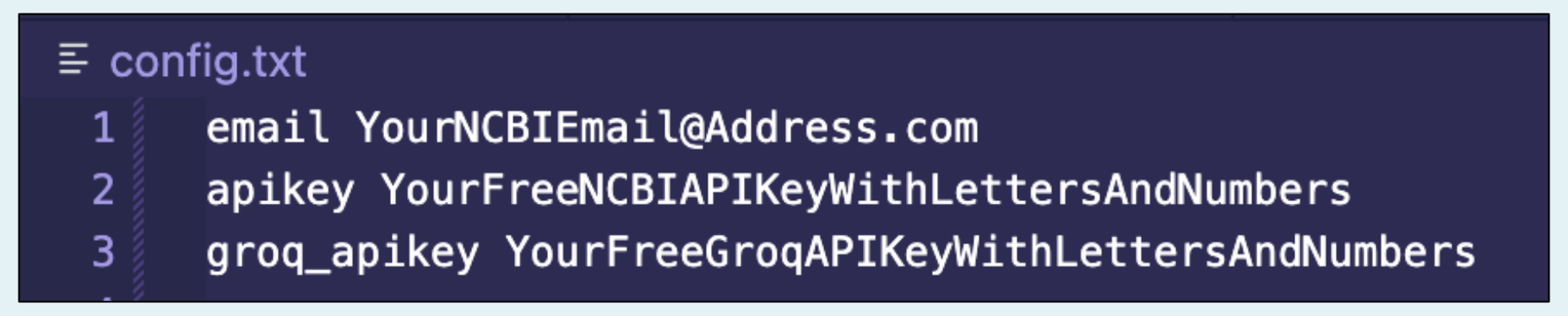
Linum usitatissimum

For this poster, we tested these model taxa as they are easy to check manually, but in the future, the program will be run across a broader sample of the plant tree of life

Pipeline Overview

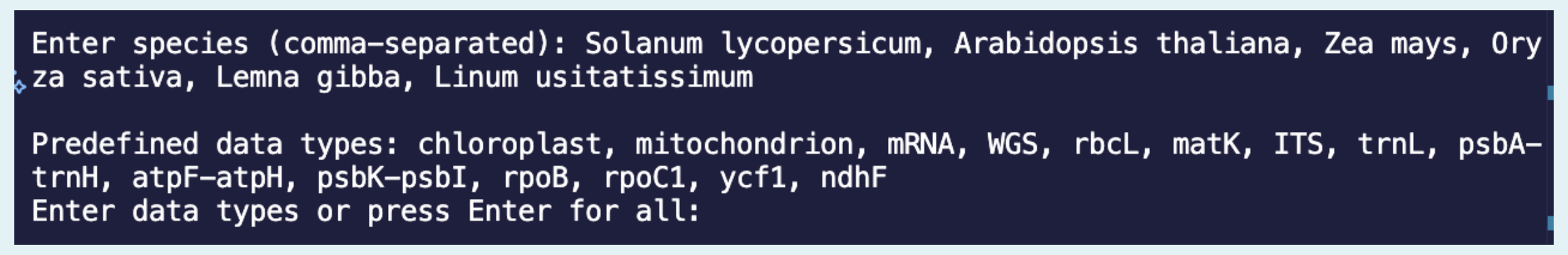
*Disclaimer: Groq AI is separate of Grok AI (Elon Musk). Additionally, since this is a large language model, outputs may not be 100% accurate and discretion is advised.

The user prepares a config file, and MissMap loads the NCBI Entrez API and Groq AI credentials



```
config.txt
email YourNCBIEmail@Address.com
apikey YourFreeNCBIAPIKeyWithLettersAndNumbers
groq_apikey YourFreeGroqAPIKeyWithLettersAndNumbers
```

MissMap parses through the user's species and data type input arguments and initializes the NCBI Entrez API and Groq AI

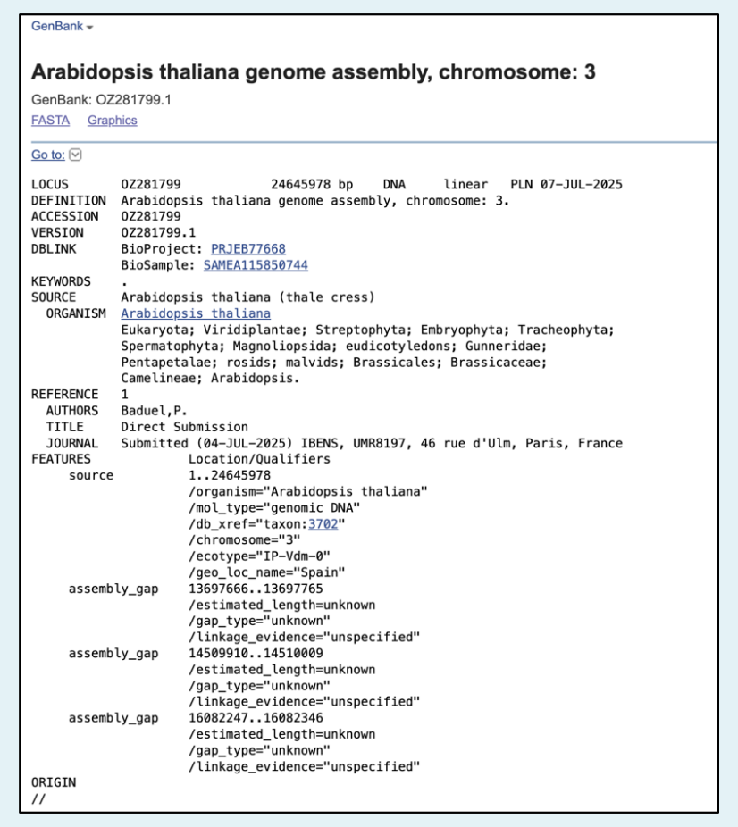


```
Enter species (comma-separated): Solanum lycopersicum, Arabidopsis thaliana, Zea mays, Ory
za sativa, Lemna gibba, Linum usitatissimum

Predefined data types: chloroplast, mitochondrion, mRNA, WGS, rbcL, matK, ITS, trnL, psbA-
trnH, atpF-atpH, psbK-psbI, rpoB, rpoC1, ycf1, ndhF
Enter data types or press Enter for all:
```

The species and data types lists are split by comma, and sequence counts from GenBank are fetched manually using Entrez and stored in a Pandas data frame for each species

NIH NLM National Center for Biotechnology Information



GenBank

Arabidopsis thaliana genome assembly, chromosome: 3

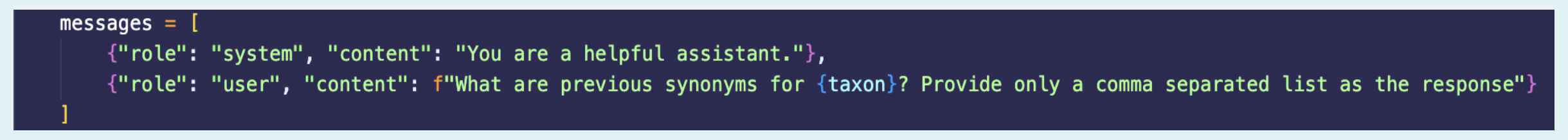
GenBank: OZ281799.1

FASTA Graphics

Go to:

```
LOCUS       OZ281799          24645978 bp    DNA     linear   PLN 07-JUL-2025
DEFINITION  Arabidopsis thaliana genome assembly, chromosome: 3.
ACCESSION   OZ281799
VERSION     OZ281799.1
DBLINK      BioProject: PRJEB77668
            BioSample: SAMEA115850744
KEYWORDS    .
SOURCE      Arabidopsis thaliana (thale cress)
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliopsida; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; malvids; Brassicales; Brassicaceae;
            Camelineae; Arabidopsis.
REFERENCE   1
  AUTHORS   Baduel,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (04-JUL-2025) IBENS, UMR8197, 46 rue d'Ulm, Paris, France
FEATURES             Location/Qualifiers
     source          1..24645978
                     /organism="Arabidopsis thaliana"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:3702"
                     /chromosome="3"
                     /ecotype="IP-Vdm-0"
                     /geo_loc_name="Spain"
     assembly_gap    13697666..13697765
                     /estimated_length=unknown
                     /gap_type="unknown"
                     /linkage_evidence="unspecified"
     assembly_gap    14509910..14510009
                     /estimated_length=unknown
                     /gap_type="unknown"
                     /linkage_evidence="unspecified"
     assembly_gap    16082247..16082346
                     /estimated_length=unknown
                     /gap_type="unknown"
                     /linkage_evidence="unspecified"
ORIGIN
//
```

Groq AI utilizes the default prompt template to obtain more species synonyms using its LLM



```
messages = [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": f"What are previous synonyms for {taxon}? Provide only a comma separated list as the response"}
]
```

The data frame with data counts and synonyms is saved to a CSV file, and a table is printed to the terminal

Species	Total sequences	chloroplast	mitochondrion	mRNA	WGS	rbcL	matK	ITS	trnL	ycf1	ndhF	Total Count of Data Found
Solanum lycopersicum	812407	239	46	363339	316	100	55	133	14	39	40	363940
Arabidopsis thaliana	2702695	2402	193	1658637	6369	48	42	27	37	15	23	1667601
Zea mays	4885606	399	442	2248738	30118	39	45	80	15	0	31	2279697
Oryza sativa	2297855	3845	315	1364003	22511	467	225	124	31	1	32	1390674
Lemna gibba	203	104	2	32	0	15	13	1	0	0	0	138
Linum usitatissimum	380624	161	0	289559	45	19	22	19	3	4	8	289765

Results

Species	Total sequences	NCBI Taxonomy DB Synonyms	Groq Enhanced Synonyms	Synonym-only count	% Synonym-only
Solanum lycopersicum	812407	Lycopersicon esculentum,Lycopersicon esculentum var. esculentum,Solanum esculentum,Solanum lycopersicum var. humboldtii,tomato	Lycopersicon esculentum,Solanum lycopersicum var. lycopersicum,Lycopersicon lycopersicum,Solanum pomiferum,Lycopersicon pomiferum,Solanum melongenum,Lycopersicon melongenum	812407	100.0
Arabidopsis thaliana	2702695	Arabis thaliana,thale cress	Arabidopsis thaliana,Arabidopsis thaliana (L.) Heynh.,Arabidopsis thaliana (L.) Knyl. & Grunth.,Sisymbrium thalianum L.,Arabidopsis thaliana var. typica,Arabidopsis thaliana var. thaliana	2702695	100.0
Zea mays	4885606	Zea mays var. japonica	Zea mays var. indurata,Zea indurata,Zea mays var. rugosa,Zea rugosa,Zea mays var. tunicata,Zea tunicata,Zea mays var. saccharata,Zea saccharata	4885606	100.0
Oryza sativa	2297855	Asian cultivated rice	Oryza sativa var. japonica,Oryza sativa var. indica,Oryza sativa subsp. japonica,Oryza sativa subsp. indica	2297855	100.0
Lemna gibba	203	swollen duckweed	Lenticula gibba,Lemna gibba var. gibba,Lemna gibba var. lecontei,Lemna lecontei,Lemna mexicana,Lemna minor var. gibba,Lemna trisulca var. gibba,Lenticularia gibba	203	100.0
Linum usitatissimum	380624	flax	Linum creticum,Linum edule,Linum humile,Linum macrosepalum,Linum pallidum,Linum sativum,Macrothymus linum	380624	100.0

Figure 1: MissMap table output table representing synonyms and synonym-only count for each species

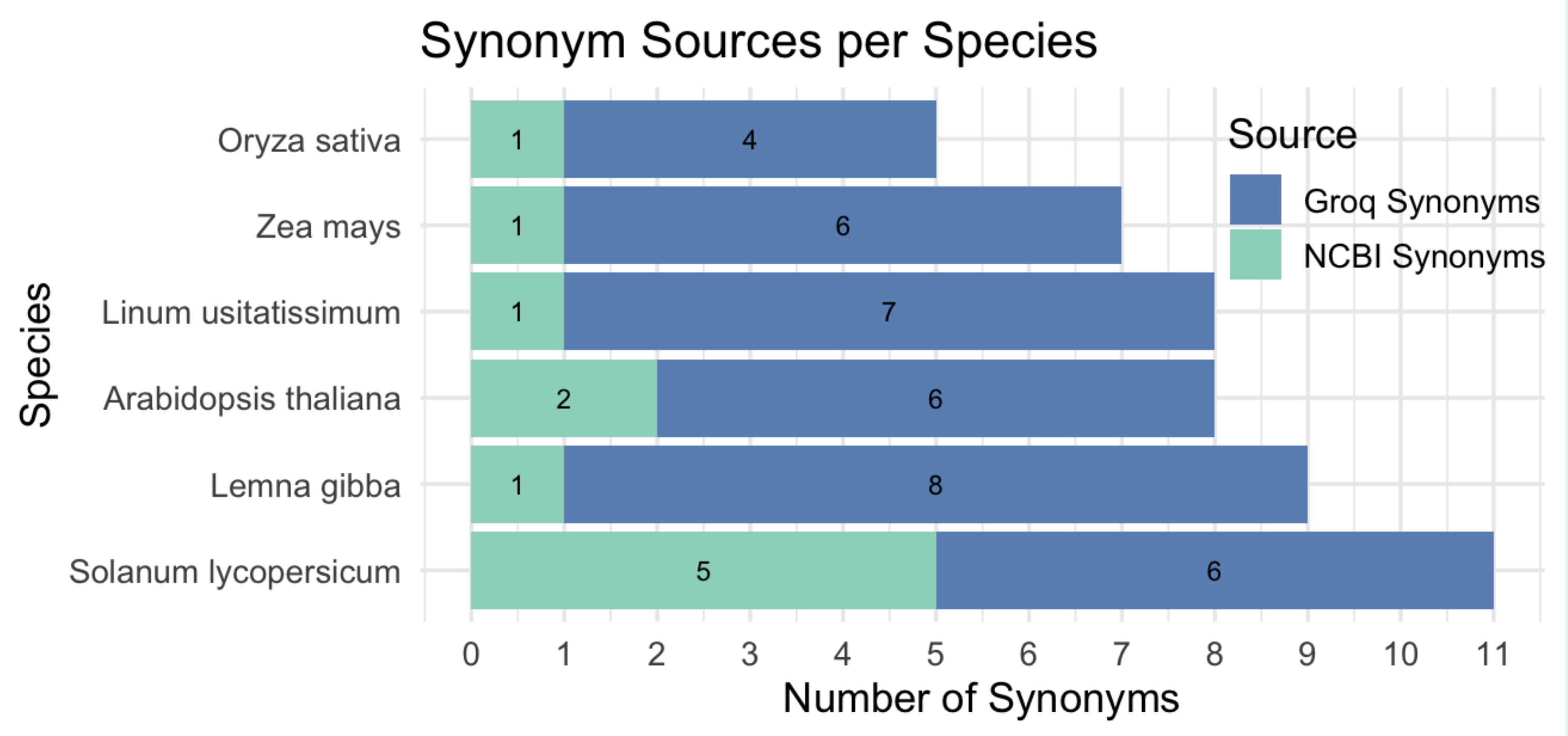


Figure 2: Stacked bar chart representing the number of synonyms extracted from NCBI Taxonomy database compared to synonym retrieval using Groq AI.

Conclusion

The resulting output from the test species suggests that for some species, like *Arabidopsis thaliana* or *Zea mays*, most sequence data on NCBI can be captured through the NCBI taxonomy database synonyms alone. However, this is only based on the small number of model taxa, which are typically highly curated. The difference in the number of results appears to be due to studies where nucleotide data from secondary species (e.g., bacterial data from an infection) are uploaded using a historical name of the primary species.

This explains the inconsistencies between the synonyms *Lycopersicon esculentum* and *Solanum lycopersicum*. Although these two synonyms represent the same tomato species and share the same NCBI Taxon ID, NCBI records return significantly different counts depending on which synonym is used. This discrepancy occurs because users have uploaded bacterial data under either *Lycopersicon esculentum* or *Solanum lycopersicum*, which, when they are not the focal species, have not been synonymized.

From this preliminary look, if you are interested in the focal species (e.g., the plant), then NCBI taxonomy works, but if you are interested in which species have been associated with the focal species (e.g., a bacteria that infects the plant), you will want to incorporate synonyms for the focal species.

Future Directions

MissMap is still in the early stages of development, and to fulfill its purpose of summarizing information available on NCBI, its functionality will be expanded to encompass more aspects of NCBI, making it easier to navigate. Since one of MissMap's strengths is to obtain data that would otherwise be difficult to parse manually from NCBI, one future goal is to test this on all species in the plant tree of life. The functionality will also be expanded to help navigate NCBI for literature searches and other areas where synonyms may be meaningful. Furthermore, the goal will be to eventually use the LLM to parse the metadata deposited with sequences, as it is not always deposited in standard formats.

Acknowledgements

THE UNIVERSITY OF ILLINOIS AT CHICAGO

We are grateful for the support, revisions, and feedback made by the members of the Walker and Igic labs at UIC.

Feedback Form

References and Citations

Federhen, Steven. "The NCBI Taxonomy database." Nucleic Acids Research 40 (2012).
Hollingsworth, Michelle L., Alex A. Clark, et al. "Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants." Molecular Ecology Resources 9, no. 2 (2009): 439–457.
National Center for Biotechnology Information. "Entrez Programming Utilities Help." NCBI Bookshelf NBK25501 (2023).
Smith, Stephen A., and Joseph F. Walker. "PyPHLAWD: a python tool for phylogenetic dataset construction." Methods in Ecology and Evolution 10, no. 1 (2019): 104–108.