Acquiring and integrating data from multiple resources for meta-analysis

Wu Huang[1,2], Alex Twyford[2], Peter Hollingsworth[1]

[1]Royal Botanic Garden Edinburgh; [2]University of Edinburgh

Wu Huang: Corresponding email: hazelhuang1993@gmail.com;

Wu Huang
University of Edinburgh, Royal Botanic Garden Edinburgh

AUG 18, 2023

OPEN ACCESS

DISCLAIMER

DISCLAIMER – FOR INFORMATIONAL PURPOSES ONLY; USE AT YOUR OWN RISK

The protocol content here is for informational purposes only and does not constitute legal, medical, clinical, or safety advice, or otherwise; content added to protocols.io is not peer reviewed and may not have undergone a formal approval of any kind. Information presented in this protocol should not substitute for independent professional judgment, advice, diagnosis, or treatment. Any action you take or refrain from taking using or relying upon the information presented here is strictly at your own risk. You agree that neither the Company nor any of the authors, contributors, administrators, or anyone else associated with protocols.io, can be held responsible for your use of the information contained in or linked to this protocol or any of our Sites/Apps and Services.

DOI: dx.doi.org/10.17504/protocols.io.kxygx3z9og8j/v1

Protocol Citation: Wu Huang, Alex Twyford, Peter Hollingsworth 2023. Acquiring and integrating data from multiple resources for meta-analysis. **protocols.io** https://dx.doi.org/10.17504/protocols.io.kxygx3z9og8j/v1

Protocol status: Working
We use this protocol and it's working

Created: Jul 22, 2023

Last Modified: Aug 18, 2023

PROTOCOL integer ID: 85380

Keywords: Meta-analysis, literature research, acquire public data, plant genomics

ABSTRACT

Acquiring sequence data generated for various purposes and repurposing it to evaluate the nature of genomic differences between plant species requires thorough dataset assessments, efficient data processing, and careful management. In this chapter, I summarise the methodology for searching appropriate datasets, data collection and filtering, and data management. I provide examples of the data formats to guide future downloading and acquiring of the datasets, and the links to curated publicly available bioinformatic tools and self-written scripts to allow reproduction and reuse of this data acquiring and processing pipeline.

There is now a significant number of studies that have sequenced multiple loci from the nuclear genomes of plants and these are available in public databases or in private data repositories (if the study hasn't yet been published). Collectively these datasets have great potential for understanding the nature of genomic differences between plant species. To harness this existing information I have developed a set of workflows, with the logic behind my approach being to

- Develop key criteria for selecting suitable datasets focusing only on datasets which have sampled multiple individuals from multiple congeneric species
- Search the literature and public data repositories for datasets that match these criteria
- Obtain and organise the selected datasets
- Use these datasets to estimate the proportion of species that resolve as monophyletic units as one measure of species discrimination success
- Establish the frequency distribution of taxonomically informative nucleotide substitutions among taxa to better understand the genomic nature of differences between plant species
- Subsample the data to better understand the effectiveness of smaller numbers of loci in recovering maximal species

discrimination, as well as evaluating the attributes of the gene regions that are most effective at telling species apart.

This protocol will focus on the first three bullet points. The protocol for the rest of the bullet points please refer to another protocol "NucBarcoder - a bioinformatic pipeline to characterise the genetic basis of plant species differences".

Developing data request form

1 Sourcing the literature and compiling publications which sequenced multiple nuclear loci for multiple individuals from multiple congeneric species.

1.1 To compile data suitable for meta-analysis, We first searched journal publications from 2013 onwards for studies that sequenced multiple loci from the nuclear genome and which

sampled multiple individuals of multiple congeneric species. The cut-off of 2013 was selected as this reflects the initiation of widespread use of next-generation sequence platforms for the recovery of nuclear sequence data from plants. We used ambiguous matching patterns to search in the Web of Science and University of Edinburgh literature search engines. The matching patterns are listed in Table 1.1

“phylogen*” AND “RAD*" AND “plants*”
“phylogen*” AND “GBS" AND “plants*”
“phylogen*” AND “genome skim*" AND “plants*”
“phylogen*” AND “transcriptome*" AND “plants*”
“phylogen*” AND “target capture" OR "Hyb-seq" AND “plants*”
“phylogen*” AND “WGS" AND “plants*”
“*RAD*" AND “plants*” AND "genus"
“GBS" AND “plants*”AND "genus"
“genome skim*" AND “plants*”AND "genus"
“transcriptome*" AND “plants*”AND "genus"
“target capture" OR "Hyb-seq" AND “plants*”AND "genus"
“WGS" AND “plants*”AND "genus"
Note: all of the advance search patterns could be merged into one query technically (combining by OR and AND), but the search engines give not as comprehensive a result as searching separately.

Table 1.1. Advance literature search keywords and matching patterns

1.2 With the full-text publications downloaded, the next step was selecting publications manually. I only kept studies if they satisfied the following criteria.

1) Three or more un-linked nuclear loci were sequenced.

2) More than two species had multiple individuals successfully sequenced and retained.

3) A phylogeny is available either in visual format or in a machine readable text format (newick, phylip, or nexus formats).

4) The species identities on the phylogeny could be interpreted and related to the sampled species.

2 Reading the methods and concluding the two popular NGS analysis pipelines are HybPiper and iPyRAD. Analyse the files used in each step and evaluate the pros and cons of acquiring data from that step.

2.1

	A	B	C	D	E	F
	Data_stages	Features_of_the_stage	Pros	Cons	File_format	Note
	Clean_reads	The original data	1. Intact information, can recover potential heterozygous sites, sequencing error, and coverage	1. Big files (10 Mb * N)2. Parsing reads file needs comprehensive computational analysis3. Need extra step to analyse if there are samples to exclude	fastq	file size denote a relative numberN - number of samples
	Cluster reads by alignment	Clustered and aligned reads data to the reference sequence (bwa). Will filter reads according to standards.	discarded putative contaminations remained potential heterozygous sites, sequencing error, and coverage info	1. Big files (4 Mb * N)2. Computationally expensive (can easily generate vcf file)3. Need extra step to analyse if there are samples to exclude	bam	I think this file as well as the reference_gene.fasta would be the best choice.
	Sequence Assembly	Assembled each cluster in each samples	Small files (0.003 Mb * N)	1. Lost heterozygous information. No sequence error info, no coverage info. Still need alignments2. Need extra step to analyse if there are samples to exclude	fasta	
	Multi-align assembled sequence	Align all the assembles sequence together	1. Small files (0.003 *N Mb)2. Easy to analyse	Lost heterozygous information. No sequence error info, no coverage info.	fasta	I am currently use this for Inga and Geonoma analyses

	A	B	C	D	E	F
	Variant calling file	Variation called	1. Small files (0.001 *N Mb)2. Easy to analyse	Lost heterozygous information. No sequence error info, no coverage info. No non-variant sites info	fasta	

Table 2.1. Pros and cons of acquiring data from the main steps of HybPiper.

3 Collecting information about the file formats frequently used in the data repositories that are popular in this particular field. (Normally read >20 related publications and the picture will be clear.) Choose the most frequently used ones as input formats for the following analytical pipelines. Three important files are:

3.1 Metadata. This includes information that corresponds to the sample IDs in the consensus sequence files which links sequences of individuals to their scientific names (species identities). This information is also useful for tracking changes of names wherever this has been applied, as well as monitoring and understanding the inclusion and exclusion of individuals and loci.

Sample ID	Species name
Antirrhinum_boissieri_L18_2	*Antirrhinum_boissieri*
Antirrhinum_boissieri_L104	*Antirrhinum_boissieri*
Antirrhinum_braun-blanquetii_E20	*Antirrhinum_braun-blanquetii*
Antirrhinum_charidemi_E23_1	*Antirrhinum_charidemi*
Antirrhinum_charidemi_E23_2	*Antirrhinum_charidemi*
Antirrhinum_cirrigherum_L114_1	*Antirrhinum_cirrigherum*
Antirrhinum_cirrigherum_L114_2	*Antirrhinum_cirrigherum*

Figure 3.1. Sample IDs corresponding to species' names. Data from M. Durán-Castillo et al., 2022

3.2 Sequence alignment file. This alignment covers the sequence variation in all individuals per dataset, and depending on how the original raw reads were processed, takes the form of either multiple-aligned sequences in .fasta, .phylip, or .nex formats, or SNP matrix in .vcf or .fasta format.

a) fasta format

```
>FG_186
CATTGTTCCTCCATAACACACAGATTTTGGCCGGA
>KD_13
CATTGTTCCTCCATAACANACAAATTTTGGCCGGA
>LA_2023
CATTGTTCCTCCATAACACACAAATTTTGGCCGGA
>M46A
CATTGTTCCTCCATAACACACANATTTTGGCCGGA
```

b) phylip format

```
Sample_ID_103     CATACATCTTCAGCACTACAGNTATCT
Sample_ID_1300    CATACATCTTCAGCACTACAGTTATCT
Sample_ID_32      CATACATCTTCAGCACTACAGTTATCT
Sample_ID_8577    CATACATCTTCAGCACTACAGTTATCT
Sample_ID_8578    CATACATCTTCAGCACTACAGTTATCT
Sample_ID_8579    CATACATCTTCAGCACTACAGTTATCT
```

c) nexus format

```
begin characters;
        dimensions nchar=31;
        format datatype=dna missing=? gap=-;
        matrix
        Sample_ID_1 CCATGACTTGATTAGCATCTGTCAAAATCCC
        Sample_ID_2 CCATGACTTGATTAGCACCTGTCAAAATCCCC
        Sample_ID_3 CCATGACTTGATTAGCACCTGTCAAAATCCCC
        Sample_ID_4 CCATGACTTGATTAGCACCTGTCAAAATCCC
    ;
end;
```

d) vcf format

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample_1.sorted.bam
Chr1	31150	.	G	A	999	.	.	GT:PL:DP:SP:ADF:ADR:AD	./.:0,0,0:0:0:0,0:0,0:0,0
Chr1	31174	.	A	T	70	.	.	GT:PL:DP:SP:ADF:ADR:AD	./.:0,0,0:0:0:0,0:0,0:0,0
Chr1	252833	.	G	T,A	243	.	.	GT:PL:DP:SP:ADF:ADR:AD	0/0:0,9,98,9,98,98:3:0:3,0,0:0,0,0:3,0,0

Figure 3.2. Examples of aligned sequences in a) fasta, b) phylip, and c) nex formats, and SNP matrix in d) vcf format.

3.3 Phylogenetic trees. Phylogenetic trees were recovered from each study to enable easy estimation of the proportion of species that resolve as monophyletic. The preferred format is for this to be a machine-readable text format such as the Newick format. However, where only a graphical representation of the tree was available, this was also retained and used, to maximise the number of studies analysed.

a) the phylogenies in newick format

```
((FG_186:0.0004627773,M46A:0.0003625627)100:0.0002685118,(KD_13:0.0003441657,LA_2023:0.0003470779)100
:0.0003440200)100:0.0013278046)100:0.0003416399,
```

b) visualized phylogeny of a)

Figure 3.3. An exemplar phylogeny in a) newick format and b) visualisation. Four samples and the lengths of the branches are included.

Acquiring data

4 Through the literature review, identify published data that could be reused.

4.1 Email the corresponding author of published data asking for their consent of using their data (though it's always free to use published data, it helps to build rapport and potentially will gain additional insights into the data used). Attached is an example of the email I used for this request. Cold email requesting data reuse.docx

4.2 Extract metadata from the publication manually.

4.3 Download data from the data repositories, such as Zenodo and Dryad. Email the correspondent author for required data or metadata if the files are not self-explanatory. The chance of having a helpful reply is low though.

5 In addition to mining the published literature, I contacted potential collaborators to request access to unpublished datasets. This involved designing a data request form and developing a list of metadata entries. This is for standardising metadata and agree with a data-sharing scheme. Meta-data_info.xlsx

5.1 Identify potential working groups that have unpublished data. Cover letter.docx

5.2 Ask people to fill in the metadata entries.

5.3 Send them the guidance for sharing files, including what files to share, where to share, and how the data will be used and authored. Requesting data form.docx

Data storage and backup

6 The shared data were uploaded to Google Drive (https://www.google.co.uk/intl/en-GB/drive/), and then downloaded and managed on the UK crop diversity bioinformatics HPC platform (https://www.cropdiversity.ac.uk).
Two copies of the backup are continually updated. A full version of data is backed up on the local hard drive and an abridged version ison the laptop.