**KEY STEPS FOR DATA PROVIDER**

1. First pass analysis to see if the data set likely contains errors of identification or imperfect taxonomy. Revise taxon as appropriate (Following the convention specified in the section ‘WORKING PRINCIPLE FOR DATA CLEANING’), and keep track of the changes as a text file.

e.g. Taxonomy\_tracking\_changes\_for\_Salix.txt

S. herbaceae\_sample1 -> S. reticulata

? S. herbaceae\_sample2

? S. alba\_sample1

\*\* A -> B : taxon revision

\*\* ? : uncertain which species the sample belongs to

1. Upload the following files onto shared point
(Google drive - <https://drive.google.com/drive/folders/1uvl-kf1yeC7a9zOTtb10wL1SUgPsKU5g?usp=sharing> )

	1. Tree files (**Species tree** which includes all samples on the tip)
	File format: Newick (recommended), Nexus, phylip
	\* also encourage to send visualized tree (File format: pdf, png, jpeg, bmp)
	2. Sequence/SNP datasets (which show variations among all individuals, encourage to include constant loci)
	File format for RAD-seq: vcf (recommended, include GT/DP in the individual identifier columns), phylip, bam
	File format for Target capture: aligned fasta, phylip, vcf (include GT/DP in the individual identifier columns), nex
	3. Taxonomy revision tracking file
	4. Sample list & Sample to Species names corresponding file if the sample names do not speak themselves(which species does the samples belong to)
	5. Reference genome/genes if applicable.

Please refer to the example files [here](https://drive.google.com/drive/folders/1KtNK4JQPrJrL5RUoY_0iopor5ikO-el8?usp=sharing) if needed.

1. Fill in the metadata form (you’ve already filled in)

**WORKING PRINCIPLE FOR DATA CLEANING**

* Correct or delete samples where there is an obvious error of identification (e.g. most samples of most species form a monophyletic cluster apart from a small number of clearly misplaced samples, which when re-examined with morphology turn out to be misidentified)
* Samples from a species do not resolve as monophyletic, but instead fall into two or more distinct clades, consistent with the presence of cryptic taxa. This may warrant splitting of the taxon (and naming accordingly (informally or formally)) to avoid ‘good’ cryptic species being treated as a non-monophyletic single taxon.
* Add a question mark to a taxon name if there is some uncertainty as to whether it is correctly identified (e.g. if there is simply low confidence that the identification is correct because the group is complex or because the material was difficult to identify or verify)
* Do not simply artificially ‘clean up’ the data. E.g. we want to avoid that monophyly is overestimated by reidentification of any samples that are in the wrong cluster, e.g.
	+ Species from well identified material turn out to be non-monophyletic with individuals among related species appearing interdigitated (mixed up). This may just occur because of recent speciation, and may be a genuine signature
	+ Leave in paraphyletic species if it is simply a case that one species is nested within the range of variation of another

**ANALYSES TO CARRY OUT**

1. Calculate Monophyletic Ratio (MR) based on the phylogeny

Monophyletic Ratio (MR) is defined as the number of species that are monophyletic out of the total number of species which have multiple individuals sampled. Figure 1 serves as an example of how to calculate the MR for the plant genus *Pedicularis1.*



Figure 1. Phylogeny of Pedicularis, highlighting species with multiple sampled individuals. In this example, 6 species are sequenced (I do not count subspecific taxa. E.g. P. tham. subsp. cup and P. tham. subsp. tham are both regarded as P. tham); 4 of the species sampled multiple individuals; 3 species in yellow boxes are monophyletic; 1 in green box is non-monophyly. The monophyletic ratio is 0.75.

1. Integrating all datasets: General view of MR for multiples genus.



Figure 2. Proportion of monophyletic clades in multiple genus

1. Extract diagnostic DNA sequence variants for every species in each genus

A variant is treated as species specific when it is present in all individuals of one taxon, and no individuals of any other taxa. If this variant is taxon specific, it is diagnostic to this taxon.

I will explore two types of variants –

1. SNPs

One of the four types of nucleotides (ATCG) appears only in one taxon

1. Single gene based haplotypes

Multiple SNP sites linked together and form a haplotype which is uniq to one taxon

Here is an example which analysed SNP-type of variants. This study sequenced 66 multiple-sampled co-generic species. It used target capture method to sequence ~800 genes on the nuclear genomes.



Figure 3. Distribution of the number of Species Specific Alleles (SNPs) for 66 co-generic species.

1. Integrating all datasets: General view of the distribution of the number of Species Specific Alleles (SNPs) for multiples genus.



Figure 4 Integrated Species Specific alleles (SNPs) for all genus

**Reference**

1. Eaton, D.A. & Ree, R.H. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). *Syst Biol* **62**, 689-706 (2013).