Purpose

This project aims to investigate the genetic differences between the sister species Begonia fluvialis and Begonia sublobata, and how these differences may contribute to speciation. By identifying species-specific mutations and comparing intra- and inter-species variation, we seek to understand the genomic basis of their divergent leaf morphology and ecological adaptations.

Background

Begonia is one of the most diverse flowering plant genera, containing over 2000 species distributed across America, Asia, and Africa (Xiao et al., 2025; Goodall-Copstake et al., 2010; Moonlight et al., 2018). However, most research on Begonia has historically focused on morphological characterization (Girmansyah et al., 2022) and phylogenetic reconstructions (Goodall-Copstake et al., 2010; Tseng et al., 2022; Ardi et al., 2022). Genetic investigations, when conducted, have typically relied on transcriptomic analysis (Emelianova and Kidner, 2022) or chromosome counting (Kono et al., 2023), both of which provide limited insights into the comprehensive genetic landscape of the genus.

In contrast, whole-genome sequencing (WGS) offers a comprehensive view of genetic variation, including structural variants, regulatory elements, and non-coding regions that are not detectable through transcriptomic or chromosome-based approaches (Ng & Kirkness, 2010). To date, only three studies have assembled whole genomes of Begonia species (Chen et al., 2024; Li et al., 2022; Xiao et al., 2025), each with distinct limitations. Xiao focused solely on genome assembly without further genetic analyses (Xiao et al., 2025). Chen examined broad genomic divergence and mutation load within the *B. masoniana* complex but did not investigate specific genes under selection or their functional roles in adaptation (Chen et al., 2024). Li concentrated on genomic conservation and phylogenetic relationships across multiple species, providing limited insight into species-specific adaptive divergence (Li et al., 2022). Notably, none of these studies explored the gene-level basis of divergence between closely related species occupying distinct ecological niches.

To address this gap, this study focuses on *Begonia fluvialis*, a newly identified species closely related to *Begonia sublobata* but exhibiting distinct morphological traits. Both species were discovered in Sumatra, Indonesia, yet they occupy contrasting ecological niches. *B. fluvialis* is adapted to a riverine habitat, featuring smaller and narrower leaves suited to its aquatic environment, typically found growing on moss-covered rocks within flowing water (Hughes et al., 2015). Conversely, *B. sublobata* inhabits relatively drier conditions, typically growing under moist but non-submerged rocks with broader leaves (Hughes and Girmansyah, 2011; Moonlight et al., 2018). These morphological differences are hypothesized to result from genetic adaptations to their respective environments; however, the underlying genetic mechanisms remain unexplored. Additionally, some Begonia species are highly valued as ornamentals and medicine (Xiao et al., 2025). Identifying genes associated with morphological traits adapted to high-moisture habitats may enable horticulturists to develop new cultivars that are more resilient under diverse moisture conditions, expanding both the commercial appeal and conservation of Begonia.

We hypothesize that the riverine adaptation and leaf morphological differences between *B. fluvialis* and its sister species *B. sublobata* are driven by species-specific genetic changes, such as gene family expansions/contractions, structural variants, indels, and SNPs. To test this hypothesis, we will utilize newly generated whole-genome sequencing datasets (e.g. long-read PacBio and short-read Illumina) to test whether these genomic differences are associated with the distinct ecological and morphological traits observed between the two species.

Method

Several sequencing datasets will be available for this study. PacBio long-read sequencing data will be provided for both species, with *B. sublobata* (11.6 GB) and *B. fluvialis* (6.6 GB) assembled by other PhD students in the lab. Additionally, Illumina skim sequencing data will be available for three *B. fluvialis* samples and two *B. sublobata* samples from different individuals. Furthermore, Nanopore MinION sequencing data from additional individuals will also be accessible.

First, PacBio sequences of *B. fluvialis* and *B. sublobata* will be annotated using BRAKER3 (Gabriel et al., 2024). RNA-seq data from *B. bipinnatifida* will serve as a reference during genome annotation, as this species belongs to Begonia sect. *Petermannia*, which is closely related to Begonia sect. *Jackia*, the section *B. fluvialis* and *B. sublobata* belongs to (Ardi et al., 2022). Standard pre-processing steps like converting RNA-seq into aligned BAM format using SAMtools (Danecek et al., 2021) will be applied before annotation.

Next, gene variation between the two species will be analyzed using the GFF3 annotation files produced by BRAKER3. Gene counts will be directly compared using the annotation files. Synteny analysis will be performed using MCScanX (Wang et al., 2012) to examine gene arrangement differences between the two genomes. To further assess gene family evolution, *B. fluvialis* will be compared to *B. sublobata* using OrthoFinder (Emms and Kelly, 2019) to identify gene family expansion and contraction. Additionally, DupGene Finder (Qiao et al., 2019) will be used to investigate the duplication modes of genes within orthogroups exhibiting differences in gene numbers.

Beyond gene content and structural organization, mutations between the two species will also be examined to identify evolutionary differences at the nucleotide level. To achieve this, raw PacBio reads from *B. fluvialis* will first be aligned to the *B. sublobata* reference genome using Minimap2 (Li, 2018), followed by conversion and sorting into BAM format using SAMtools (Danecek et al., 2021). This will enable the identification of three main types of mutations: structural variations (SVs) involving large insertions or deletions greater than 50 bp, single nucleotide polymorphisms (SNPs), and frameshift mutations caused by small Indels.

For the detection of structural variants, Sniffles2 (Smolka et al., 2024) will be employed to generate a VCF file containing the identified SVs. Insertions and deletions exceeding 50 bp will then be extracted from this file using BCFtools (Danecek et al., 2021). Meanwhile, SNPs and smaller Indels will be identified separately through variant calling with BCFtools, producing comprehensive VCF files for subsequent analysis. To determine the genomic distribution of these variants, the SNPs and Indels located within gene regions will be identified by intersecting the VCF files with annotated gene regions in the GFF3 file using BEDTools (Quinlan, 2014). From this, the SNP and Indel density per gene will be calculated to highlight highly polymorphic genes. Furthermore, non-synonymous mutations, which may have functional consequences on protein structure and activity, will be extracted and analyzed using BCFtools.

Apart from inter-species variation, within-species genetic polymorphism in *B. fluvialis* will also be examined. Illumina skims from three different individuals will be aligned to the PacBio assembly derived from a separate individual using BWA-MEME (Jung and Han, 2022) and SAMtools (Danecek et al., 2021). BWA-MEME is preferred over Minimap2 here, as it is specifically optimized for short-read alignment (Jung and Han, 2022). SNPs and Indels within gene regions will be identified by intersecting the VCF output with the GFF3 annotation file using BCFtools (Danecek et al., 2021) and BEDTools (Quinlan, 2014). Highly polymorphic genes and SNPs predicted to have functional impacts on protein structure will be further analyzed with BCFtools. Due to the limited sample size, population genomic approaches such as PCA and D-statistics are not feasible in this study.

A comparison between intra-species and inter-species variation will be conducted to identify species-specific mutations. Gene Ontology (GO) enrichment analysis will then be performed using topGO (Alexa and Rahnenführer, 2009) in R, based on GO annotations derived from BRAKER3. The analysis will assess whether these genes are functionally associated with leaf morphology and water adaptation. Statistical significance will be determined using Fisher's exact test.

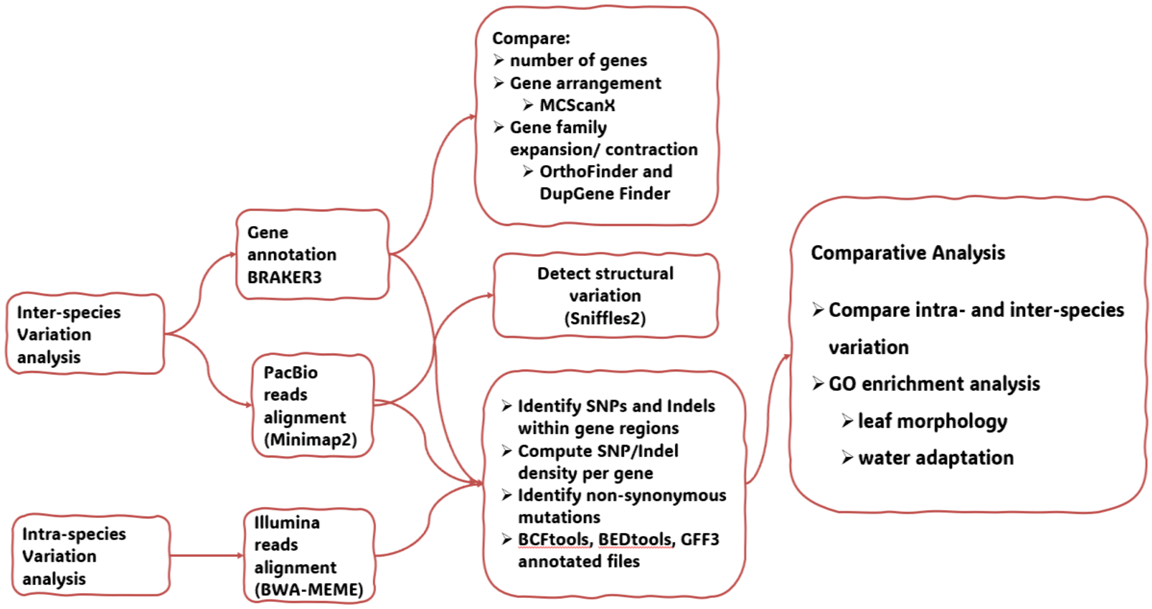

Compare:
➢ number of genes
➢ Gene arrangement
➢ MCScanX
➢ Gene family expansion/ contraction
➢ OrthoFinder and DupGene Finder
Gene annotation BRAKER3
Inter-species Variation analysis
Detect structural variation (Sniffles2)
PacBio reads alignment (Minimap2)
Comparative Analysis
➢ Compare intra- and inter-species variation
➢ GO enrichment analysis
➢ leaf morphology
➢ water adaptation
➢ Identify SNPs and Indels within gene regions
➢ Compute SNP/Indel density per gene
➢ Identify non-synonymous mutations
➢ BCFtools, BEDtools, GFF3 annotated files
Intra-species Variation analysis
Illumina reads alignment (BWA-MEME)

Figure 1. This Workflow diagram illustrates the main steps for genome alignment, annotation, variation detection, and comparative analysis between *Begonia fluvialis* and *B. sublobata*.

Evaluation

First, the quality of the assembled PacBio genomes will be evaluated using QUAST (Gurevich et al., 2013), based on metrics such as N50, GC content, and contig alignment plots. Second, the annotation quality will be assessed using OMArk, a recently developed tool that evaluates completeness, structural consistency, and potential contamination of the annotated sequences (Nevers et al., 2025). If the initial PacBio assemblies prove to be low quality, Nanopore MinION data can be used to refine or scaffold them, thus improving overall genome continuity and quality.

To evaluate the quality of predicted orthogroups used for gene family expansion/contraction analysis, BLASTP can be used. For example, a random sequence from a given orthogroup can be selected and aligned against the other sequences in the same orthogroup. A strong match (e.g. low E-values) would indicate that the group contains functionally related sequences, providing confidence in the orthogroup prediction.

SNP and indel variants will be evaluated and filtered during the variant calling and post-processing steps using BCFtools (Danecek et al., 2021). Variants will be filtered based on thresholds previously applied in Chen et al. (2024): variants with QUAL < 30 and multi-allelic SNPs will be removed; those with more than 30% missing data will be excluded to reduce mapping bias; and SNPs with MAF < 0.05 will be filtered out to reduce noise from rare variants. For structural variants (SVs), quality control will involve filtering on key metrics such as minimum read support, variant length, and overall variant quality, using information from Sniffles2 output (Smolka et al., 2024) and BCFtools.

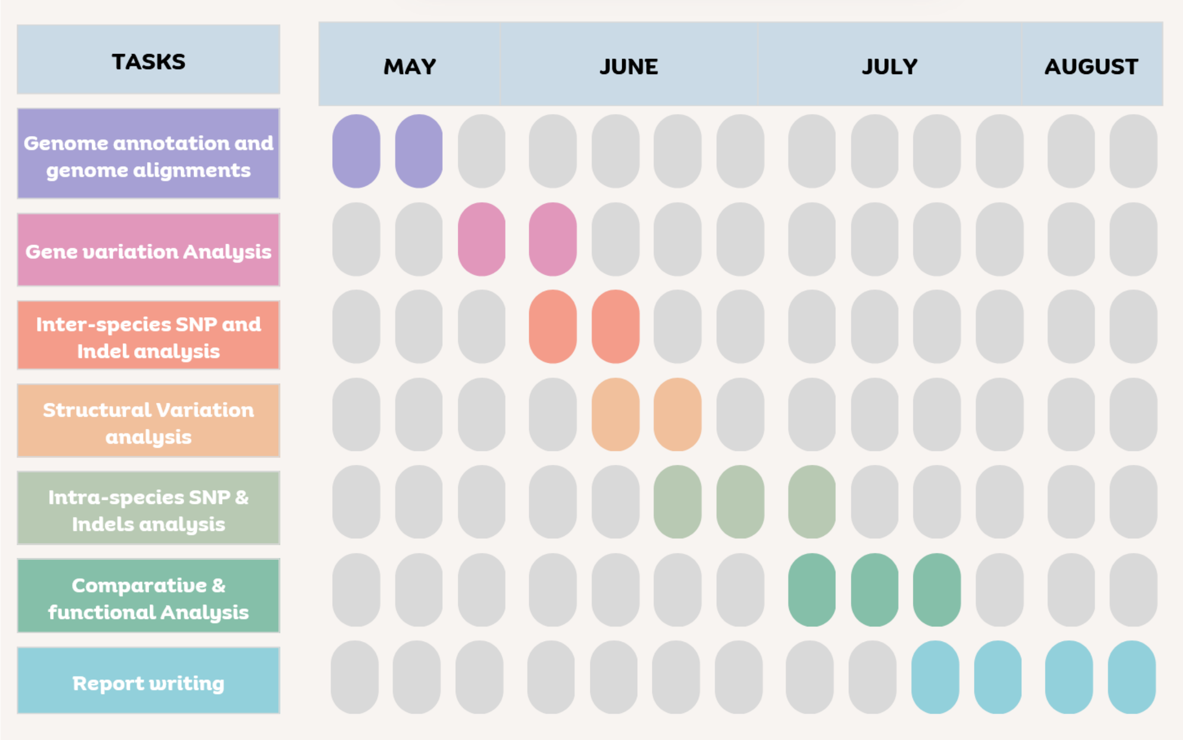
Because multiple testing will be performed during GO enrichment analysis, p-values will be adjusted using false discovery rate (FDR) correction, such as the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), to reduce false positives. Additionally, the biological relevance of significantly enriched GO terms will be evaluated by assessing their association with known pathways related to leaf development and water adaptation, using functional gene databases such as UniProt (UniProt Consortium, 2019) or literature-curated resources.

Expected research output

The annotated data will provide fundamental insights into the genetic differences between *B. fluvialis* and *B. sublobata*. First, the number of genes and their arrangement in both species are expected to be largely similar, with only minor differences. This expectation is based on the previous distance matrix analysis using Illumina skim data, which indicated small genetic differences between the two species. However, as the genetic distance between them is not zero, we anticipate the presence of species-specific genetic variations that contribute to their phenotypic differences. Gene family expansion and contraction analysis will help identify orthogroups that have undergone evolutionary changes like gene duplication. SNP, structural variation, and indel analyses are expected to reveal mutations within coding and regulatory regions. Furthermore, the identification of non-synonymous mutations will help pinpoint SNPs that are likely to affect protein function and contribute to phenotypic changes in these plants. A comparison between intra-species and inter-species SNP variation will allow us to distinguish species-specific mutations from within-species polymorphisms. Finally, GO enrichment analysis will help determine whether genes with species-specific mutations are enriched in pathways related to leaf structural development and environmental adaptation, particularly water-associated traits in *B. fluvialis.*

Workplan.

I will begin genome annotation and sequence preprocessing on May 10th, which is expected to take approximately two weeks. Following this, I will conduct the inter-species variation analysis from late May to around June 20th (details in Figure 2). The intra-species variation analysis, comparative analysis, and GO enrichment analysis will take approximately one month, with all analyses expected to be completed by July 20th. I will then dedicate the following month to writing my final dissertation, aiming to complete it by mid-August.



TASKS	MAY			JUNE				JULY				AUGUST	
Genome annotation and genome alignments	●	●											
Gene variation Analysis			●	●									
Inter-species SNP and Indel analysis				●	●								
Structural Variation analysis					●	●							
Intra-species SNP & Indels analysis						●	●	●					
Comparative & functional Analysis								●	●	●			
Report writing										●	●	●	●

Figure 2. Gantt chart of the project timeline. Created by the author using a template from Canva.

Reference

Alexa, A., & Rahnenführer, J. (2009). Gene set enrichment analysis with topGO. *Bioconductor Improv, 27*(1-26), 776.

Ardi, W., Campos-Dominguez, L., Chung, K. F., Dong, W. K., Drinkwater, E., Fuller, D., ... & Wilson, H. (2022). Resolving phylogenetic and taxonomic conflict in Begonia. *Edinburgh Journal of Botany,* 79.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Chen, Y., Dong, L., Yi, H., Kidner, C., & Kang, M. (2024). Genomic divergence and mutation load in the Begonia masoniana complex from limestone karsts. *Plant Diversity, 46*(5), 575-584.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., ... & Li, H. (2021). Twelve years of SAMtools and BCFtools. *Gigascience, 10*(2), giab008.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology, 20*, 1-14.

Emelianova, K., & Kidner, C. (2022). Comparative transcriptome analysis of two closely related Begonia species reveals divergent patterns in key light-regulated pathways. *Edinburgh Journal of Botany, 79*, 1-18.

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, *AUGUSTUS, and TSEBRA. Genome Research, 34*(5), 769-777.

Girmansyah, D., Hughes, M., Ardi, W. H., & Chikmawati, T. (2022). Six new species of Begonia (Sect. Jackia, Begoniaceae) from Sumatra, Indonesia. *Taiwania, 67*(1).

Goodall-Copestake, W. P., Pérez-Espona, S., Harris, D. J., & Hollingsworth, P. M. (2010). The early evolution of the mega-diverse genus Begonia (Begoniaceae) inferred from organelle DNA phylogenies. *Biological journal of the Linnean society, 101*(2), 243-250.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072-1075.

Jung, Y., & Han, D. (2022). BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics, 38*(9), 2404-2413.

Hughes, M., & Girmansyah, D. (2011). Searching for Sumatran Begonia described by William Jack: following in the footsteps of a 19th century Scottish botanist. *Gardens' Bulletin Singapore, 63*(1&2), 83-96.

Hughes, M., Girmansyah, D., & Ardi, W. H. (2015). Further discoveries in the ever-expanding genus Begonia (Begoniaceae): fifteen new species from Sumatra. *European Journal of*

Taxonomy, (167).

Kono, Y., Peng, C. I., Oginuma, K., Yang, H. A., & Chung, K. F. (2023). Cytological study of Begonia sections Jackia, Platycentrum and Reichenheimia (Begoniaceae) with chromosome numbers 2n= 34 and 38. *Cytologia, 88*(2), 161-166.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*(18), 3094-3100.

Li, L., Chen, X., Fang, D., Dong, S., Guo, X., Li, N., ... & Liu, H. (2022). Genomes shed light on the evolution of Begonia, a mega-diverse genus. *New Phytologist, 234*(1), 295-310.

Moonlight, P. W., Ardi, W. H., Padilla, L. A., Chung, K. F., Fuller, D., Girmansyah, D., ... & Hughes, M. (2018). Dividing and conquering the fastest–growing genus: towards a natural sectional classification of the mega–diverse genus Begonia (Begoniaceae). *Taxon, 67*(2), 267-323.

Nevers, Y., Warwick Vesztrocy, A., Rossier, V., Train, C. M., Altenhoff, A., Dessimoz, C., & Glover, N. M. (2025). Quality assessment of gene repertoire annotations with OMArk. *Nature biotechnology, 43*(1), 124-133.

Ng, P. C., & Kirkness, E. F. (2010). Whole genome sequencing. *Genetic variation: Methods and protocols*, 215-226.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., ... & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome biology, 20*, 1-23.

Quinlan, A. R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics, 47*(1), 11-12.

Smolka, M., Paulin, L. F., Grochowski, C. M., Horner, D. W., Mahmoud, M., Behera, S., ... & Sedlazeck, F. J. (2024). Detection of mosaic and population-level structural variants with Sniffles2. *Nature biotechnology, 42*(10), 1571-1580.

Tseng, Y. H., Hsieh, C. L., Campos-Domínguez, L., Hu, A. Q., Chang, C. C., Hsu, Y. T., ... & Chung, K. F. (2022). Insights into the evolution of the chloroplast genome and the phylogeny of Begonia. *Edinburgh Journal of Botany*, 79, 1-32.

UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research, 47*(D1), D506-D515.

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., ... & Paterson, A. H. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids research, 40*(7), e49-e49.

Xiao, T. W., Wang, Z. F., & Yan, H. F. (2025). A chromosomal-level genome assembly of Begonia fimbristipula (Begoniaceae). *Scientific Data, 12*(1), 429.