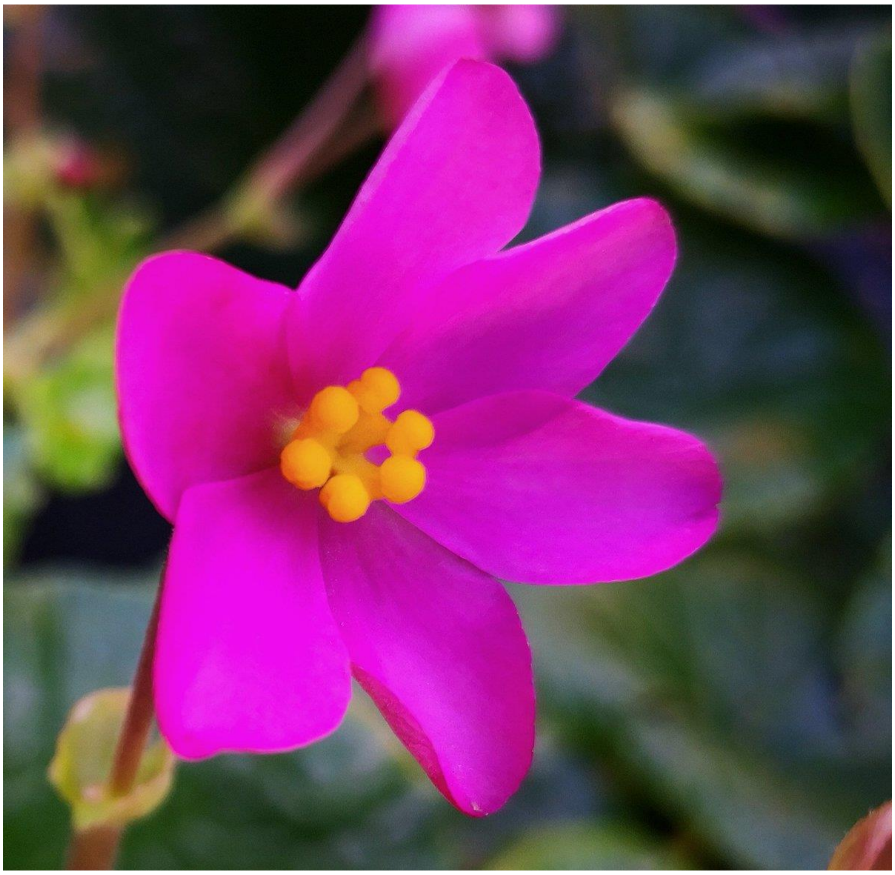
An Investigation into the Introgression of *Begonia socotrana* Alleles into Horticultural *Begonia* Varieties



Philippa Stone

Thesis submitted for partial credit for the MSc Biodiversity and Taxonomy of Plants

Abstract

Begonia socotrana is a species endemic to Socotra, which, unusually, flowers during winter. Following collection in the 1880s, *Begonia socotrana* was crossed with many varieties of *Begonia* to produce winter-flowering hybrids. It is thought that a single accession of *Begonia socotrana* is parent to all modern winter-flowering *Begonia*. This study uses modern genetic techniques to determine the level of *Begonia socotrana* introgression into a range of horticultural *Begonia* hybrids.

By comparing variants in coding sequences shared between *Begonia socotrana* and known species of Begoniaceae, genomic variants specific to *Begonia socotrana* were identified. The proportion of these *Begonia socotrana* specific genomic variants found in horticultural *Begonia* hybrids was used to assess introgression of *Begonia socotrana* alleles. Variation in non-coding regions of the genome was studied through analysis of repeat sequences.

The horticultural *Begonia* hybrids examined contained large, but varying, proportions of the *Begonia socotrana* specific genomic variants. This suggests there is some diversity in the *Begonia socotrana* alleles present in horticultural varieties, even though it is probable they all originated from a single *Begonia socotrana* specimen. There was also evidence for high genetic variation in the non-coding regions of the genomes, likely due to elevated levels of transposon activity following hybridisation events.

Acknowledgements

Firstly, I am, and always will be, extremely grateful to my supervisor, Catherine Kidner, for having introduced me to *Begonia* research last year and patiently guiding me through my second project.

I also thank Thibauld Michel for giving me a huge head start on my project by completing all the lab work needed for this project, Lucia Campos Dominguez for getting me started with various softwares and helping me make sense of sometimes baffling outputs, Mark Hughes for giving me the expert's opinion on *Begonia socotrana*, Sven Oostvogels for giving me an insight into *Begonia* breeding, and Thea Kongsted for her advice on repeat analyses.

Finally, I am grateful to Caroline Stone for always looking at my work with a critical eye and improving my writing, and to Katherine Durrell for her proofreading skills and ability to make me laugh even in the worst parts of this experience.

Cover photo: *Begonia socotrana* at Royal Botanic Garden Edinburgh (Hughes, 2019)

Table of Contents

Appendices

List of Tables

List of Figures

Chapter 1 - Introduction

1.1 Socotra

Socotra is an island that at just 132 kilometres (82 miles) in length and 49.7 kilometres (31 miles) in width is the largest island of its namesake, the Socotran archipelago. Socotra is situated in the Arabian Sea 240 kilometres (150 miles) east of the coast of Somalia and 380 kilometres (240 miles) south of the Arabian Peninsula, making it one of the most isolated continental landforms on Earth (figure 1). While politically Socotra is a part of Yemen, geologically the Socotran archipelago is considered to be a part of the African continent. Socotra has an arid climate, with an annual daily mean temperature of over 28°C and an annual rainfall of 193mm, which is usually restricted to falling in just three months of the year (Deutscher Wetterdienst, 2019). This rainfall, however, is extremely unevenly distributed across the island, with most falling in the mountains of the Hajhir massif. The unusual landscape of Socotra has been described as "the most alien-looking place on Earth" (Huntingford, 1980) and its famously unusual flora is likely to be at least partially responsible for this accolade.

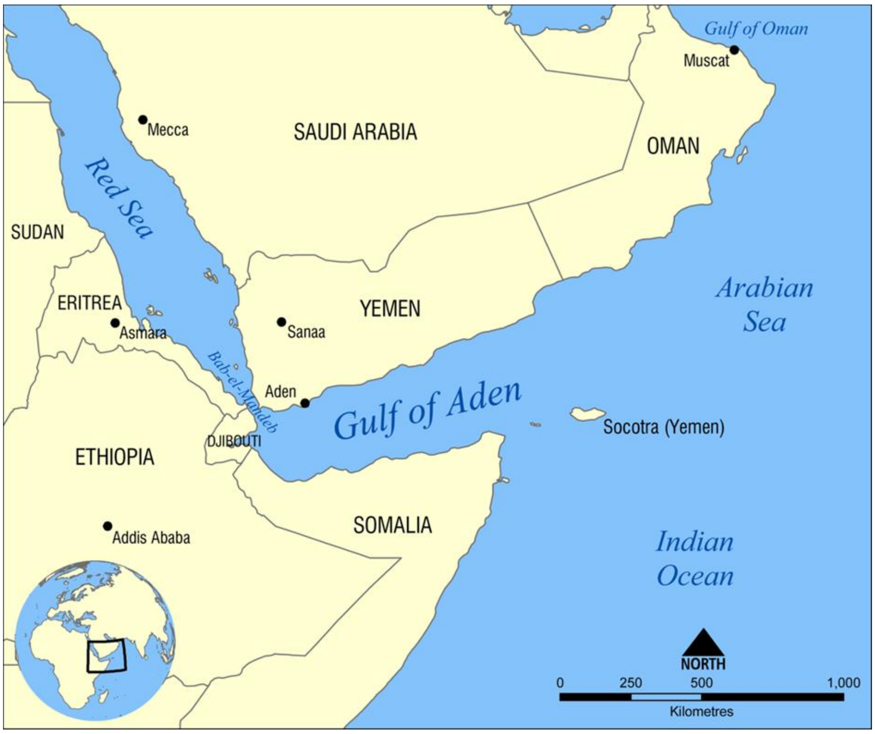


Figure 1 Map of the Gulf of Aden.

Socotra is extremely isolated from mainland Africa and the Middle Eastern Peninsula (Einstein, 2009).

1.2 *Begonia socotrana*

One third of the plant species found on Socotra are thought to be endemic to the island, with *Begonia socotrana* being one of them. *Begonia socotrana* is a small herb with pink flowers that grows in the limestone peaks of the Hajhir massif (figure 2, figure 3). The rest of the island is too dry for *Begonia* species, known to favour moist environments (Hughes and Miller, 2002). It is surprising to find any *Begonia* in Socotra at all - given their usual preference for moist conditions. When *Begonia socotrana* was described by Hooker, he noted that Socotra was "one of the last places in the world in which a *Begonia* could have been expected to occur" (Hooker, 1881).

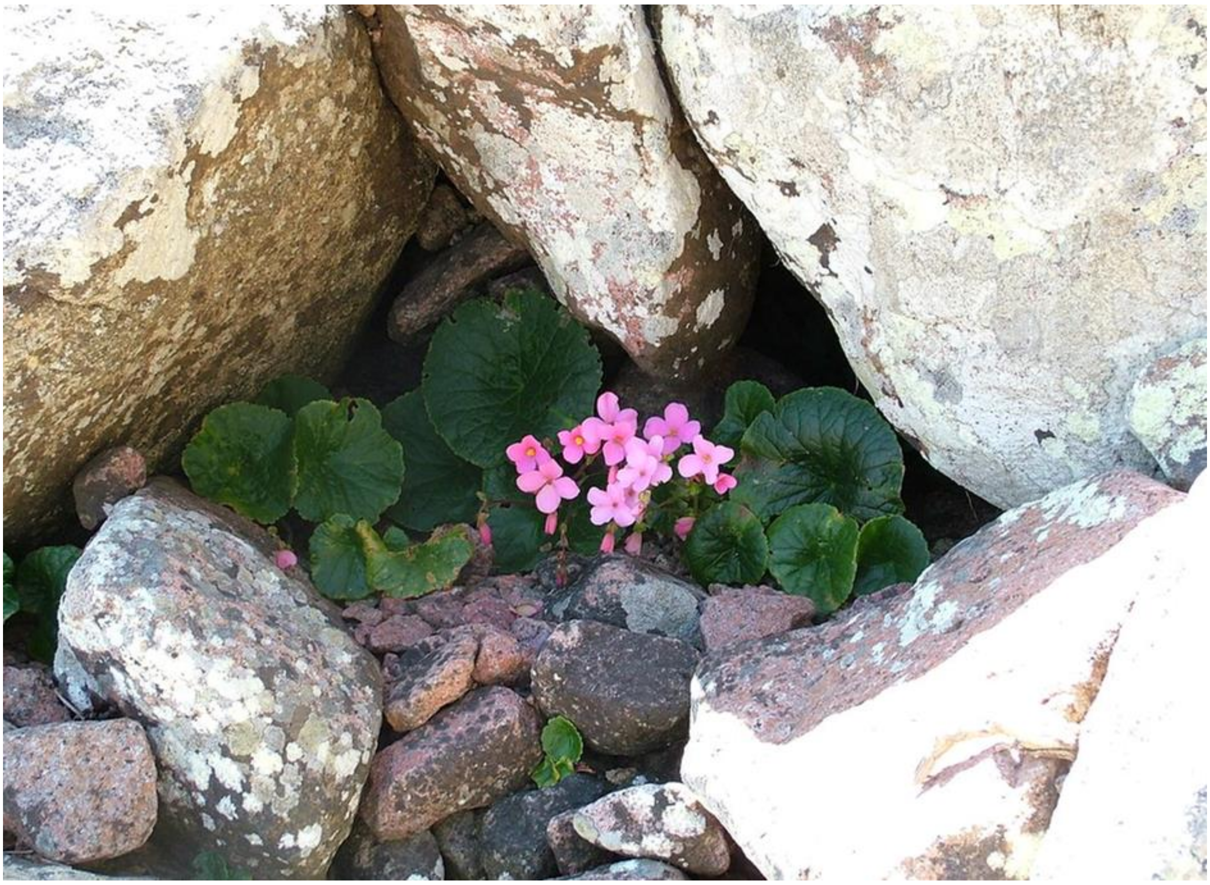


Figure 2 *Begonia socotrana* in the Hajhir massif.

Begonia socotrana grows within crevices in the limestone rock of the Hajhir massif (Scepi, 2005).

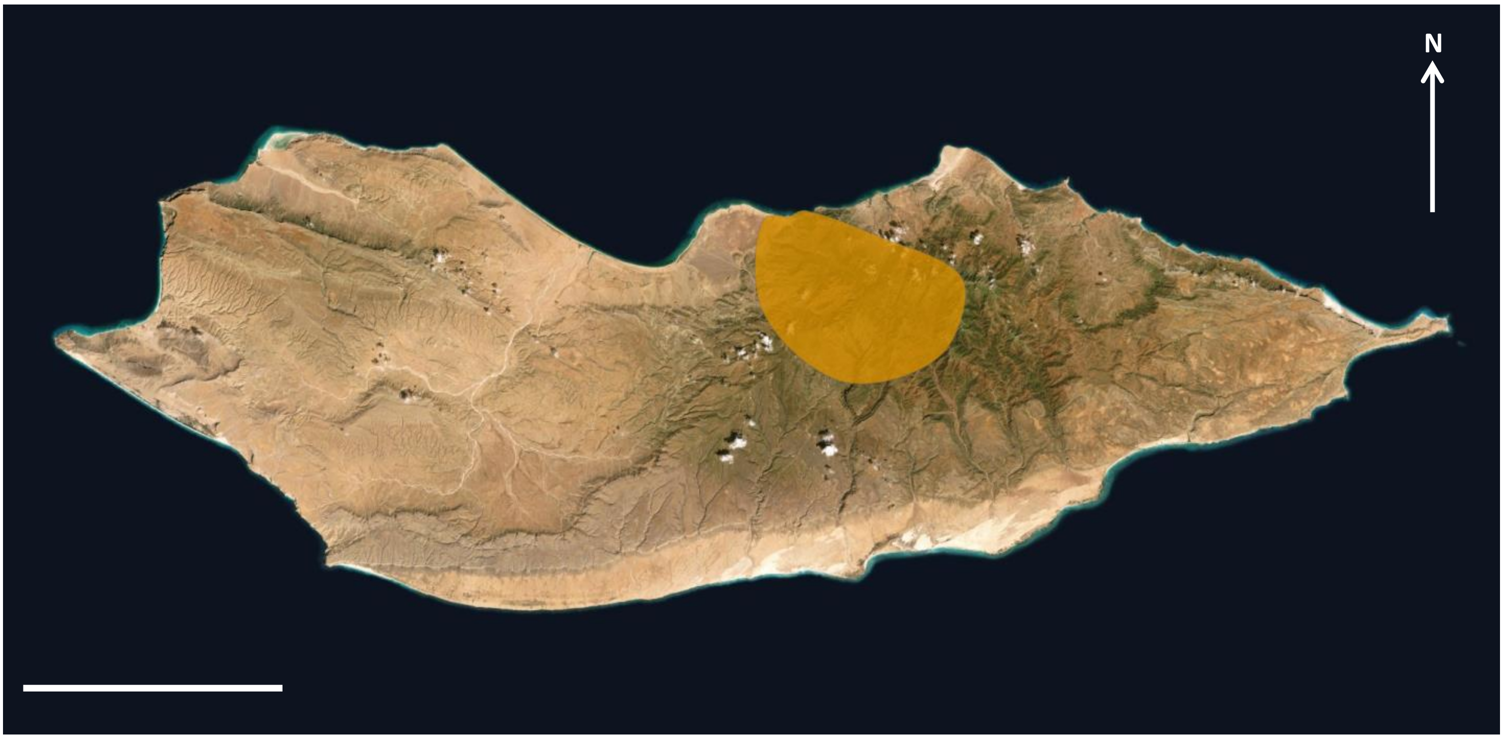


Figure 3 Map of Socotra showing distribution of *Begonia Socotrana*.

The Yellow transparent layer over the satellite image shows the range of Begonia socotrana (Miller, 2004a). Scale bar represents 25km.

Begonia socotrana caused immediate interest in the botanical community after its discovery in 1880 during Balfour's expedition to the Socotran archipelago. Due to the extremely hot and dry conditions of the Socotran summer season, *Begonia socotrana* can only flower during the rainy season, which lasts from November to February (Hughes and Miller, 2002). This makes *Begonia socotrana* a short day flowering plant. The unusual flowering time combined with attractive glabrous, circular, peltate leaves and bright pink, showy flowers meant that *Begonia socotrana* had traits highly desirable for horticultural *Begonia* (figure 3), and was an instant hit with breeders (Gleed, 1961). Notably, *Begonia socotrana* was indeed responsible for the production of the first winter flowering hybrid *Begonia* (Hughes, 2001).

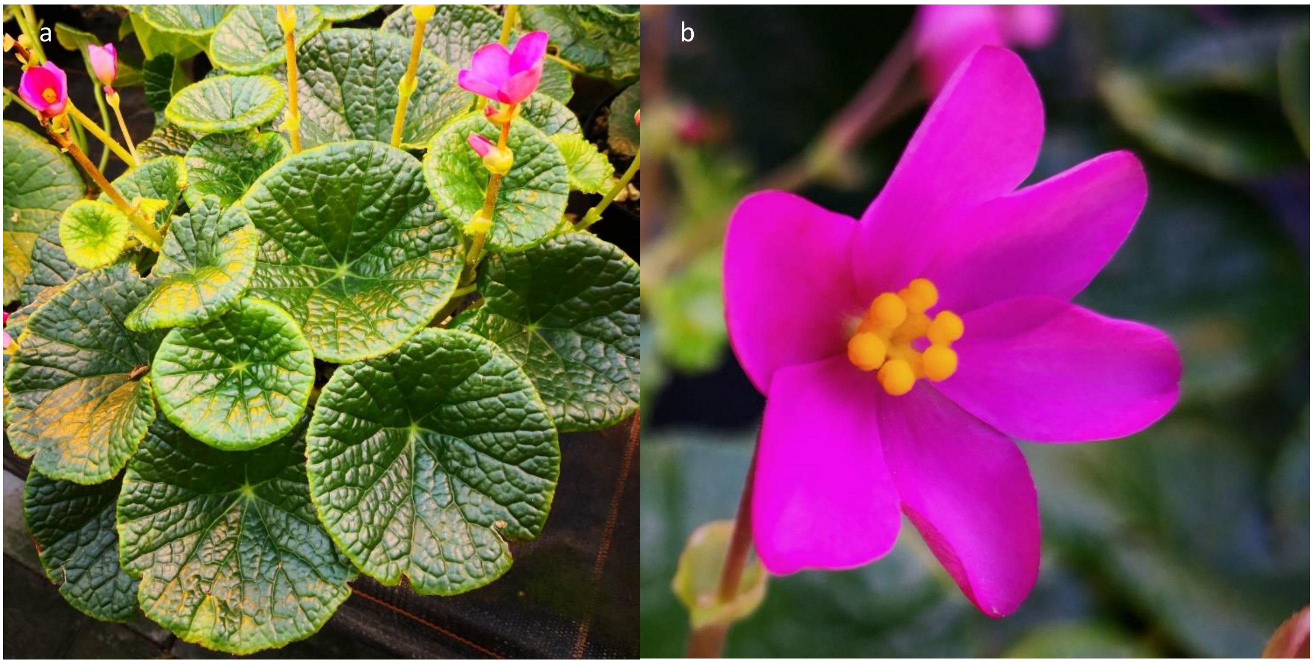


Figure 4a-b *Begonia socotrana* in the research glasshouses at Royal Botanic Garden Edinburgh.

These photographs clearly show the bright pink flowers and glossy leaves that are of interest to *Begonia* breeders. (Hughes, 2019).

1.3 Horticultural *Begonia* varieties

The *Begonia x* Hiemalis group of winter-flowering hybrid *Begonia* originated in 1883 from crosses performed by Veitch and Sops of Chelsea between *Begonia socotrana* and *Begonia* from the *Begonia x* tuberhybrida group of hybrid *Begonia*. The *Begonia x* tuberhybrida group of hybrid *Begonia* originated in 1865 from repeated hybridisations between four species of South American tuberous *Begonias*; *Begonia veitchii*, *Begonia rosiflora*, *Begonia pearcei*, and *Begonia boliviensis* (Hughes, 2002). A second group of winter-flowering *Begonia* were produced using *Begonia socotrana* as a parent species in 1891. The *Begonia x* Cheimantha group originated from crosses between *Begonia socotrana* and *Begonia dregei*, a tuberous species native to southern Africa. The most famous of the *Begonia* x Cheimantha group is the *Begonia* “Gloire de Lorraine”, produced by Lemoine. While the *Begonia* “Gloire de Lorraine” itself was never hugely commercially successful due to its difficult growing conditions, it was heavily utilised by *Begonia* breeders (Hvoslef-Eide et al., 1995). It has been long been suspected that all winter flowering varieties of *Begonia* originate from original collection of *Begonia socotrana* made by Balfour back in the 19th century (Gleed, 1961).

1.4 Importance of horticultural trade and research

Despite the fact that the UK horticulture industry has an annual footprint of £24.2 billion, little scientific research is carried out on horticultural plants when compared to other industries such as agriculture (Oxford Economics, 2018). *Begonia* themselves are economically important plants, largely due to their popularity as bedding plants. In the USA the annual trade for Semperflorens hybrids, derived from *Begonia cucullata* crossed with *Begonia schmidtiana* and *Begonia* x tuberhybrida cultivars, alone was estimated to be worth over £68 million when adjusted for inflation (Tebbit, 2005). Despite their widespread use, the genetics of commercial winter flowering varieties of *Begonia* have not yet been optimised. *Begonia socotrana* has poor disease resistance, leaving *Begonia socotrana* hybrid cultivars vulnerable to plant pathogens, in particular *Botrytis* and *Fusarium* (S Oostvogels pers comm., 2019). The variation in petal colour seen in winter flowering *Begonia* hybrids could also be improved. Gaining a better understanding of the diversity of *Begonia socotrana* alleles in current varieties of horticultural *Begonia* hybrids will inform further breeding efforts, which may enable the production of improved winter flowering *Begonia* cultivars. The number of winter flowering Begonia purchased is decreasing, and the Norwegian production of *Begonia* x cheimantha halved between 1977 and 1992 to just over one million plants annually (Hvoslef-Eide et al., 1995). The production of new and improved varieties of winter *Begonia* may enable winter flowering begonia to compete with more established competitors in the Christmas pot plant market such as *Euphorbia pulcherrima.*

1.5 Experimental design

The postulate that all winter flowering varieties of *Begonia* originate from Balfour's original 1880 collection of *Begonia socotrana* has not been tested using modern genetic techniques. Research into the genome composition of *Begonia socotrana* and *Begonia* x hiemalis has been carried out, however past studies have focussed on using cytological methods (Marasek-Ciolakowska et al., 2009). In recent years, the advent of Next Generation Sequencing (NGS) has revolutionised genomic research, allowing large amounts of data to be sequenced at high speed and incredibly low costs of less than £0.012 per one million base pairs of DNA at the time of writing (Wetterstrand, 2019). Royal Botanic

Garden Edinburgh is a unique place to do research into *Begonia* genomics, as a large amount of data is already available, and there is a huge selection of live material in the research collections. This study will take advantage of the low cost of sequencing and availability of *Begonia* genomic data, which will be utilised in an attempt to elucidate the degree of *Begonia socotrana* introgression into the genomes of horticultural hybrid varieties of *Begonia*.

1.5.1 Next Generation Sequencing techniques

1.5.1.1 Genome skims

Genome skims are a relatively simple technique in the world of NGS. To make a genome skim, genomic DNA is extracted from the organism and sequenced at a low-coverage, typically up to 5% of the genome (0.05x coverage) (Dodsworth, 2015). This enables a large amount of data to be garnered at a fraction of the cost of a whole genome sequence. Genome skims consist of sequences that are high-copy in the genome or in the cell. Genome skims will typically be made up of plastid and mitochondrial DNA, high-copy nuclear genes, and nuclear genomic repeats including retroelements, DNA transposons and satellites (Straub et al., 2012).

1.5.1.2 Bait sequences

Whole genome in solution capture (WISC) is method used to increase the proportion of coding sequence extracted from a DNA sequencing library. Baits are made of oligonucleotides sequences based on the reverse complements of the genes of interest to be captured from the DNA sequencing library. The fragments of DNA containing the sequence of interest will hybridise to the bait with the reverse complement of their sequence. Once hybridised, the oligonucleotide baits are removed, and the DNA fragments can be detached and sequenced. This process is known as hybrid bait sequencing (Carpenter et al., 2013). Hybrid bait sequencing was not used in this study, as it decreases the content the sequence data that is non-coding, and non-coding regions of DNA can be highly informative

(Dodsworth, 2015). The sequences of the baits designed for use in WISC, however, were used. The sequences of a *Begonia* bait set that had been designed using a *Begonia luzhaiensis* transcriptome were able to be utilised. There are over 1200 baits present in this set, and each bait is based on the coding sequence of a *Begonia luzhaiensis* gene.

In this study, the *Begonia* bait set was used in lieu of a genome when mapping reads from genomes and genome skims to a reference. This method has disadvantages when compared to mapping to a genome. Firstly, not all genes are present in the bait set, only 1200. Secondly, no variation in non-coding regions of the genome is detected. However, time and computational constraints meant that mapping to the baits was chosen over mapping to a genome for this project. Variation in non-coding regions of DNA was quantified using other techniques discussed further in the methodology.

1.5.1.3 Analysing genome variation

To assess variation in an organism, genomic reads are aligned, or mapped, to a reference sequence or series of references sequences, using an alignment tool such as Bowtie2 or BWA (Langmead and Salzberg, 2012 and Li and Durbin, 2009). The alignment parameters can be altered to change how exact the match has to be between a genomic read and reference sequence for the genomic read to be committed to mapping by the alignment tool's algorithm. The more closely related the organism being mapped is to the reference organism, the fewer variations will be found between the genomic reads and the reference sequence assembly. Different types of variation in an alignment are coded in different ways. Single Nucleotide Polymorphisms (SNPs) are variants that each represent a difference in a single nucleotide from the reference sequence. Insertions or deletions of bases between one to ten thousand base pairs long in the genomic read being aligned are called indels (Gagliano et al., 2018). Indels are a more powerful indicator of divergence from a reference genome, especially in this study, as reads are being aligned to coding sequences. The redundancy of the genetic code means that a point mutation may not cause a change in amino acid sequence. The presence of an indel in a coding region means that the amino acid sequence of the protein that gene codes for is likely to have been altered.

Table 1. Identity and origin of Horticultural *Begonia* accessions sampled.

Cross	Cultivar	Flower Colour	Location of acquisition	RBGE accession number	Type of sequence data
Begonia x hiemalis	Valentino Pink	Red	Tapei flower market	N/A	Genome skim
Begonia x hiemalis	Unknown	Pink	Tapei flower market	N/A	Genome skim
Begonia x hiemalis	Solenia	Red	Glasgow Botanic Garden	N/A	Genome skim
Begonia x hiemalis	Solenia	Pink	Glasgow Botanic Garden	N/A	Genome skim
Begonia x hiemalis	Solenia	White	Glasgow Botanic Garden	N/A	Genome skim
Begonia x hiemalis	Baladin	Red	Beekenkamp, Maasdijk	N/A	Genome skim
Begonia x hiemalis	Baladin	Red	J&P ten Have, Dunbar	N/A	Genome skim
Begonia x cheimantha	F1 "Gloire de Lorraine"	Pink	Royal Botanic Garden Edinburgh	Unknown	Genome skim
Unknown	Unknown	Pink	OPGC, Texas	20130776	Genome

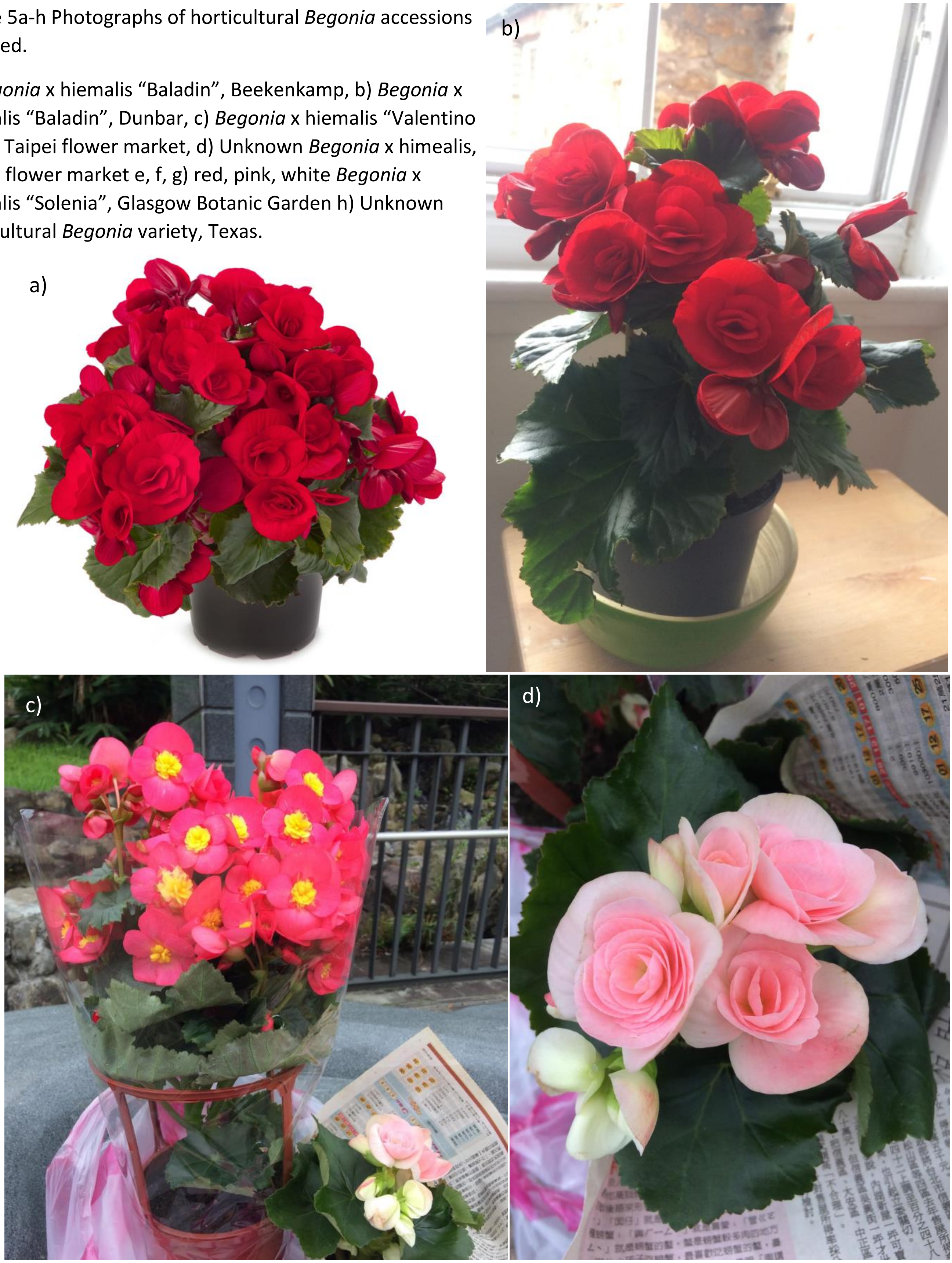
1.5.2 Plant material

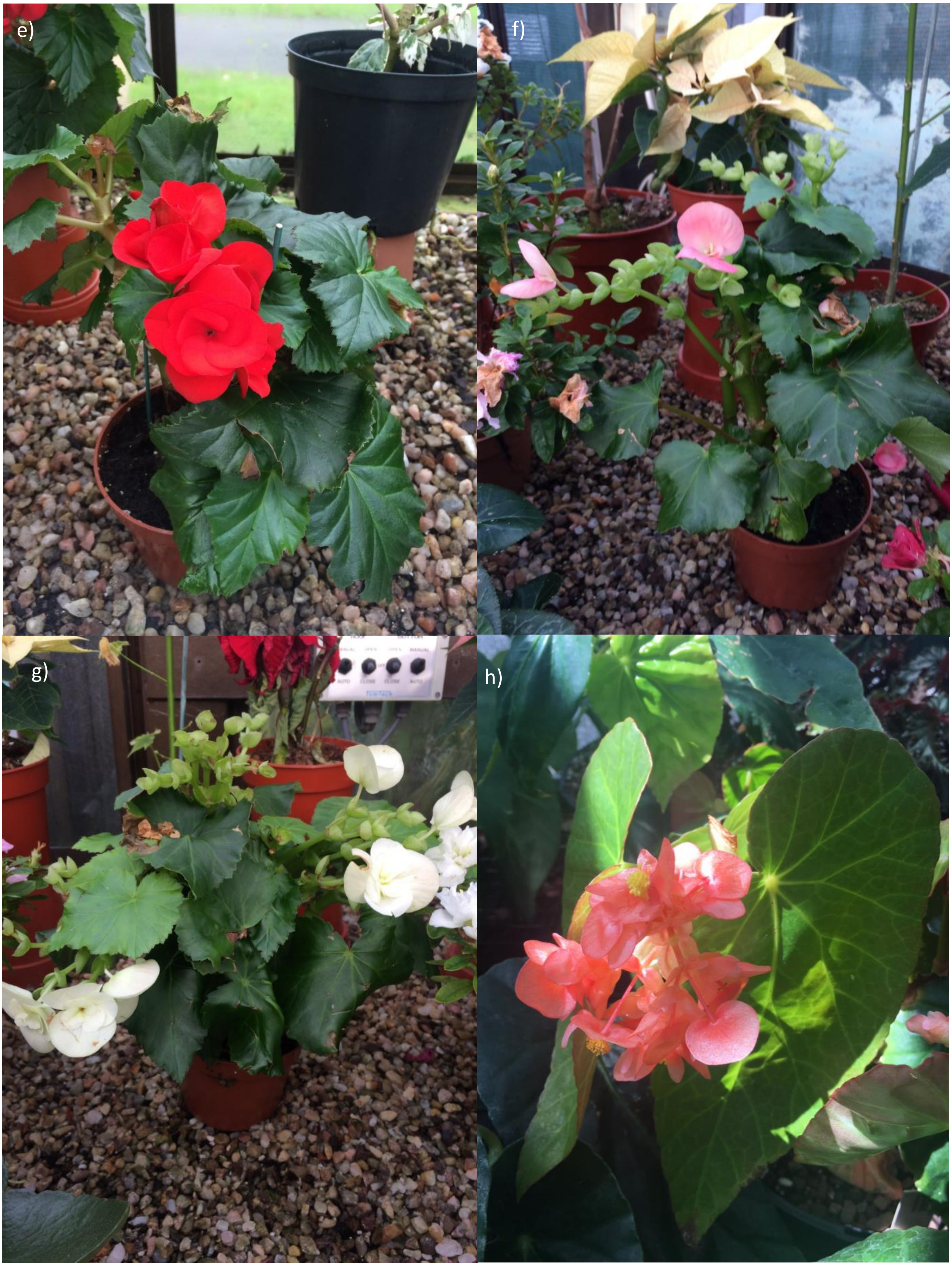
Nine horticultural accessions of *Begonia* were used in this study, sourced from locations as diverse as Dunbar and Taipei (Table 1). Seven of the nine species belong to the *Begonia* x hiemalis group of winter flowering *Begonia*. There was a wide range of floral morphologies present in this group, and lineages have clearly been selected for large and showy flowers (figure 5). The two specimens from the Taipei flower market are particularly can be seen to have particularly derived flowers when compared to the picture of *Begonia socotrana* growing in the wild (figure 2).

An F1 from a cross between *Begonia socotrana* and *Begonia dregei*, the same parentage as that of the *Begonia* "Gloire de Lorraine" was also sampled (figure 6), as was an accession of horticultural *Begonia* from unknown genetic background (Table 1). While no information is available about the parentage of the final horticultural accession, the morphology of leaves and flowers suggests it has largely South American ancestry (Figure 5h).

Figure 5a-h Photographs of horticultural *Begonia* accessions sampled.

a) *Begonia* x hiemalis “Baladin”, Beekenkamp, b) *Begonia* x hiemalis “Baladin”, Dunbar, c) *Begonia* x hiemalis “Valentino Pink”, Taipei flower market, d) Unknown *Begonia* x himealis, Taipei flower market e, f, g) red, pink, white *Begonia* x himealis “Solenia”, Glasgow Botanic Garden h) Unknown horticultural *Begonia* variety, Texas.




e)
f)
g)
h)

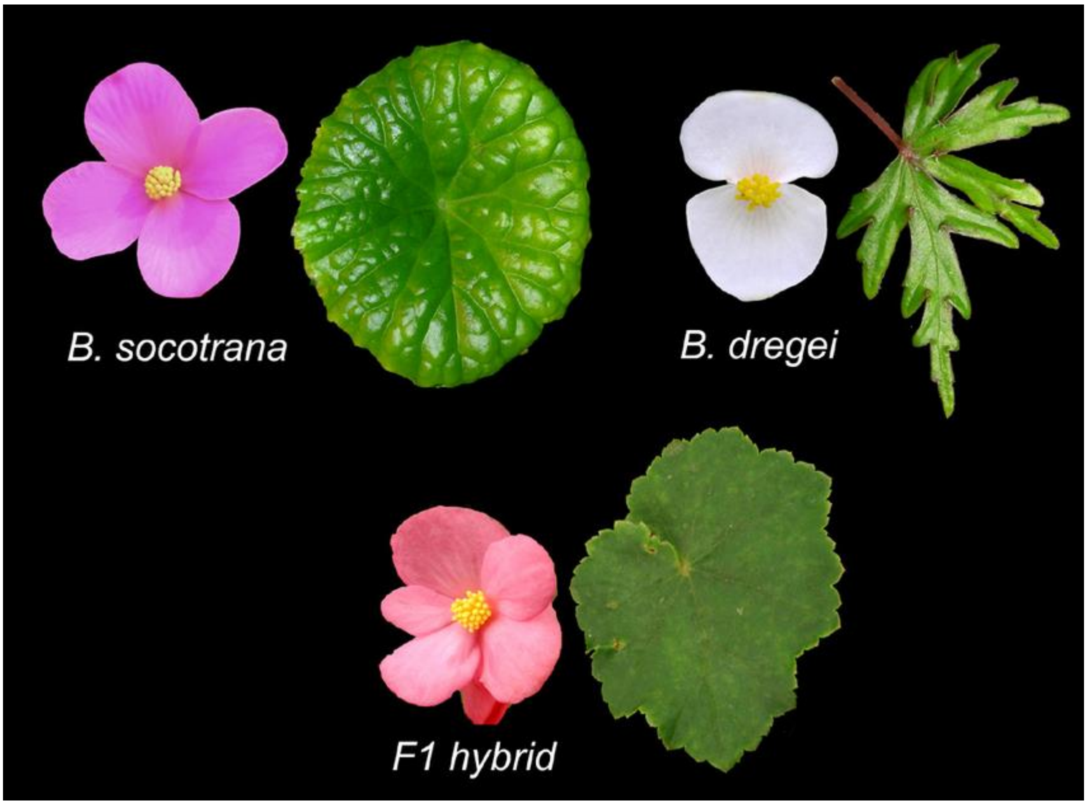


Figure 5 Recreation of *Begonia* “Gloire de Lorraine”.

A modern day *Begonia* “Gloire de Lorraine” produced from its parent species *Beognia socotrana* and *Begonia dregei* in Royal Botanic Garden Edinburgh (Neale et al., 2006).

1.5.3 Study description

Genome skims from a range of horticultural hybrid *Begonia* were carried out. Genomic reads from known species in the family Begoniaceae already present in RBGE data store, including *Begonia socotrana*, were mapped to the sequences of the *Begonia* bait set, which was used as a reference. Variants only present in *Begonia socotrana* were identified. The genomic reads form horticultural hybrid *Begonia* genome skims were then also mapped to the *Begonia* bait set. For each horticultural hybrid accession, the SNPs and indels shared with the *Begonia socotrana* specific SNPs and indels were found and counted to assess the degree of *Begonia socotrana* introgression. The *Begonia socotrana* alleles deemed to have introgressed into the horticultural hybrid accessions were identified, linked to phenotypic changes that may have been selected for in the horticultural hybrids where possible. As the *Begonia* bait set only comprises coding sequences, the diversity of the non-coding regions of the genomes was examined using alternative methods discussed further.

Chapter 2 - Methodology

2.1 Sequence data

The genome sequences of *Begonia socotrana*, *Begonia dregei*, *Begonia conchifolia*, *Begonia luxurians*, and *Begonia bipinnatifida* were available from RBGE datastore (Campos Dominguez, unpublished). The genome of *Hillebrandia sandwichensis* was also available from the RBGE datastore (Martinez Martinez, 2017). *Hillebrandia* is the only other genus in the family Begoniaceae apart from *Begonia*, and is monotypic. Two genome skims from two accessions of *Begonia heracleifolia* were also present in the RBGE data store (table 2).

The genome sequence of a horticultural variety of *Begonia* of unknown genetic background from Ornamental Plant Germplasm Center (OPGC), Texas, Index Seminum number 11ohci01 and RBGE accession number 20130776, was available on the RBGE datastore. While the genetic background of this *Begonia* is unknown, the leaf and floral morphology suggest it may be of South American origin. The rest of the sequence data used in this study came from genome skims of eight accessions of horticultural varieties of *Begonia* (table 1).

Table 2. Identity and origin of Begoniaceae species accessions sampled

The *Begonia heracleifolia* accessions are differentiated by A and “B” since no information of biological distinguishing features between the two accessions could be located.

Genus	Species	RBGE accession number	Type of sequence data
Begonia	*bipinnatifida*	20090801	Genome
Begonia	*conchifolia*	20042082	Genome
Begonia	*dregei*	20030633	Genome
Begonia	*heracleifolia* A	Unknown	Genome skim
Begonia	*heracleifolia* B	Unknown	Genome skim
Begonia	*luxurians*	19689454	Genome
Begonia	*socotrana*	200000294	Genome
Hillebrandia	*sandwichensis*	N/A	Genome

2.2 Genome skims

DNA was extracted by Thibauld Michel using a CTAB protocol for extraction of ultrashort DNA fragments (Gutaker et al., 2017). This atypical protocol was used due to the high oxalic acid content of the *Begonia* leaves. Sequencing was carried out by NOVOgene using an Illumina hiSeq X Ten, and reads were 300bp long. Reads were trimmed of adapters and low quality bases with a phred score less than or equal to 30 were trimmed from the ends of each read prior to adapter removal using Cutadapt v2.4 (Martin, 2011). Any flanking N bases from each read were also removed. Sequence quality metrics were garnered and analysed using FastQC (Andrews et al., 2010) before and after quality control.

Raw reads were uploaded to the NCBI short reads archive on 15th August 2019 (SUB6183516, PRJNA560636: Horticultural varieties of *Begonia*), and will be made available following necessary curator review.

2.3 Analysis of genome repeat content

Repeat content of Begoniaceae species and horticultural variety genomic reads was analysed using RepeatExplorer2 (Novák et al., 2013) and TAREAN (Novák et al., 2017). RepeatExplorer2 characterises the number and type of repetitive elements found in the input short sequences using a graph-based, sequence-clustering algorithm that facilitates *de novo* repeat identification. This means there is no need for a reference database of known elements, making it suitable for non-model species such as those in Begoniaceae (Novák et al., 2013). TAREAN is a pipeline that detects satellite repeats, also directly from unassembled short reads (Novák et al., 2017). For each specimen, three random samples of 300 000 reads taken using seqtk, (Li, 2013) were run through RepeatExplorer2, the maximum number RepeatExplorer2 and TAREAN will analyse in one run. Both RepeatExplorer2 and TAREAN were run using the default settings for paired and unpaired reads as appropriate. For accessions sequenced using unpaired end reads, only the forward unpaired reads were sampled, as unpaired reverse reads had lower quality scores. The mean percentage of sequences in analysed clusters was calculated for each accession. The means of the mean percentages of sequences in analysed clusters were calculated for all horticultural samples and all species samples separately, as was the variances and standard deviation between these groups. A two-tailed Welch's t-test for unequal variances was used to test if the differences between the mean percentage of sequences in analysed clusters in the horticultural varieties and species was significant. The identity and similarity of the sequences within the largest cluster of each horticultural hybrid and *Begonia socotrana* was examined.

2.4 Read mapping

Reads from the Begoniaceae species and horticultural variety genomes and genome skims were aligned to the *Begonia* bait set. The *Begonia* bait set was designed by Catherine Kidner based on *Begonia luzhaiensis* sequences (Kidner, unpublished). Alignment of the genomic reads was carried out using Bowtie2 2.3.5.1 (Langmead and Salzberg, 2012). Bowtie2 parameters were optimised to map reads as many reads to the *Begonia* bait set as accurately as possible. To ascertain what parameters to use with Bowtie2 when aligning reads from the *Begoniaceae* genomes and genome skims, and horticultural genome skims to the *Begonia* bait set, 30 sets of mapping were carried out. Random subsets of three million reads were sampled from the *Begonia conchifolia* and *Begonia socotrana* genomes using seqtk (Li, 2013). Using Bowtie2, these random subsets of three million reads from the *Begonia socotrana* and *Begonia conchifolia* genomes were aligned back to the *Begonia socotrana* genome, while altering the minimum score threshold of Bowtie2. For Bowtie2 to consider an alignment valid, and map a read, the alignment must have an alignment score equal to or greater than the minimum score threshold. In end-to-end alignment mode, the default minimum score threshold is `-0.6 + -0.6 * L`, where `L` is the read length. The random samples of reads from *Begonia socotrana* and *conchifolia* genomes were aligned back to the *Begonia socotrana* genome with 15 decreasing minimum score thresholds until the percentage of reads mapped to the bait set for each species stopped increasing and plateaued (figure 11).

The genome skims and genomes for each accession were mapped back to the *Begonia* bait set using bowtie2 in end-to-end alignment mode with minimum score threshold `-2.2 + -2.2 * L`. SAM files output by bowtie2 were converted to VCF (variant call format) files using SAMtools v1.9, a suite of utilities for interaction with and manipulation of short DNA sequence read alignments (Danecek et al., 2011). The variant call format is a format of text file used to store variations in gene sequence such as SNPs and INDELs. A combination of BCFtools v1.9 (Li, 2011) and VCFtools v0.1.13 (Danecek et al., 2011) was used for manipulation of VCF files. BCFtools is a faster version of VCFtoools, and both provide a suite of commands that allow basic functions to be carried out on VCF files. VCFtools was used to generate statistics about the variation seen in each accession from the *Begonia* bait sequences. A full list of commands used is found in appendix A.

2.5 Identification of specific and shared SNPs and indels

VCF files for each accession contained the SNPs and indels found in that specific accession. The mean number of SNPs and INDELs was calculated for horticultural varieties and species separately, and for all samples as a whole. The difference in means between the horticultural varieties and species groups was tested using a two-tailed Welch's t-test for unequal variances. The SNPs unique to *Begonia socotrana* were found using VCF-isec to create a new VCF file containing only SNPs that were found in *Begonia socotrana* and no other species sampled within Begoniaceae. VCF-isec was then used again, to determine how many of the SNPs unique to *Begonia socotrana* from all Begoniaceae species SNPs could be found in each of the horticultural hybrids. Further VCF files containing the *Begonia socotrana* specific SNPs and INDELs common to all horticultural *Begonia* varieties sampled and found in only *Begonia socotrana* and the *Begonia socotrana* x *Begonia* dregei F1 were created. *Begonia baits* based on genes related to control of flowering time and other important *Begonia socotrana* traits were located in the bait set. These baits were then searched for in the *Begonia socotrana* SNPs and indels and if found, these socotrana specific SNPs are found in the SNPs and indels shared by *Begonia socotrana* and the hybrid horticultural *Begonia* varieties.

Chapter 3 - Results

3.1 Genome skims and Quality control

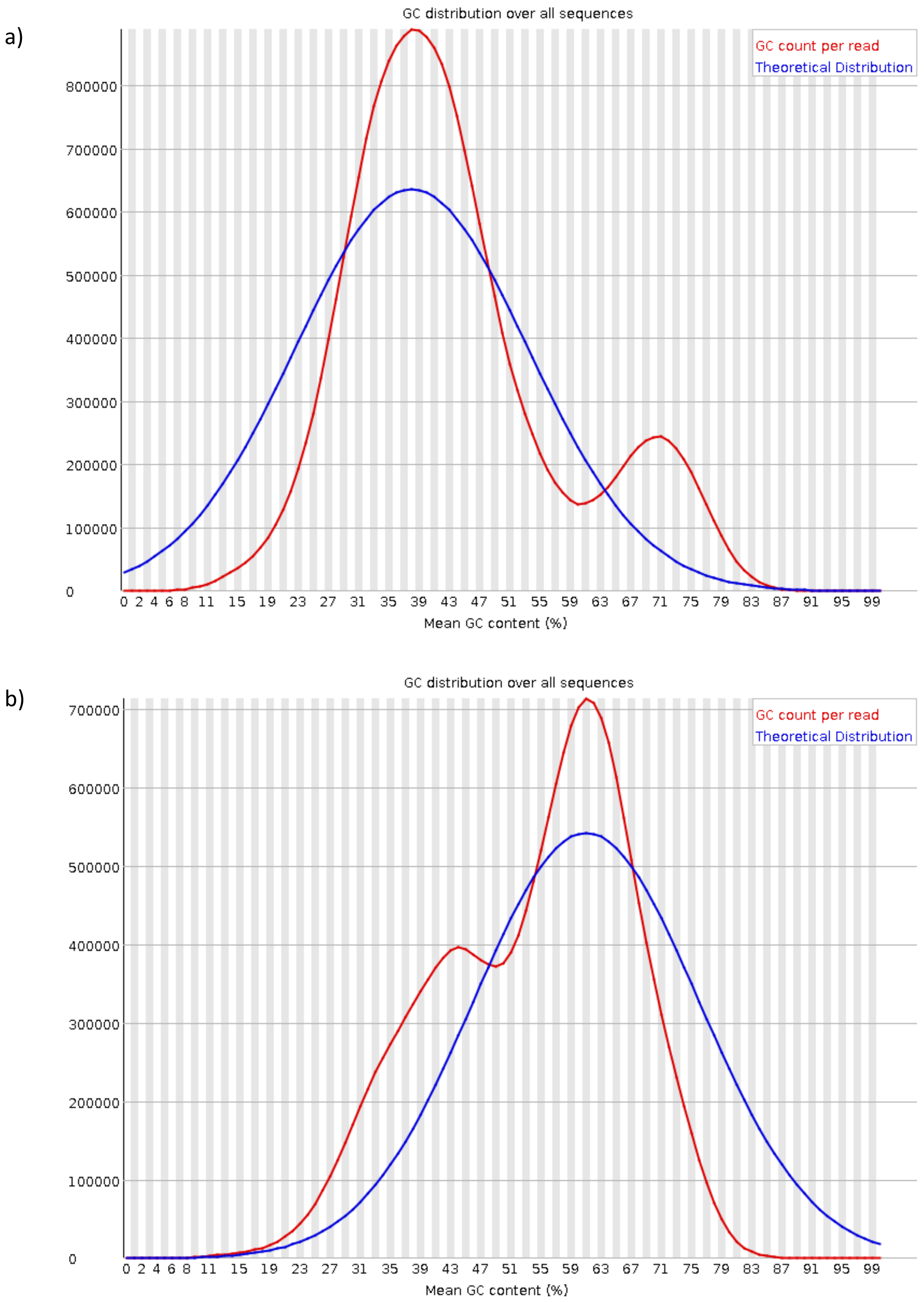
In total, over 50 billion base pairs of DNA were sequenced from the genome skims for the eight horticultural hybrids. FastQC results indicated that the genome skims were of good quality, as the median per quality base score was over 25 for each base and the number of non-unique sequences made up less than 20% of the total number of sequences (Andrews et al., 2010). Less than 2% of reads were discarded from each horticultural genome skim following quality control (table 3). However, the *Begonia* x hiemalis "Baladin" from Beekenkamp and *Begonia* x cheimantha F1 had GC content distribution curves with two peaks (figure 7). FastQC results for Begoniaceae species sequences showed that paired reads were of higher quality than unpaired reads, so paired reads were used where possible for input into RepeatExplorer2 and TAREAN.

Table 3 Quality control of horticultural *Begonia* variety genome skims.

Cross	Cultivar	Flower colour	Raw data (Gbp)	Raw reads	Reads left after trimming	Reads trimmed
Begonia x hiemalis	Valentino Pink	Red	7.7364651	25788217	25514862	273355
Begonia x hiemalis	Unknown	Pink	7.1958507	23986169	23634532	351637
Begonia x hiemalis	Solenia	Red	6.5066706	21688902	21328866	355698
Begonia x hiemalis	Solenia	Pink	7.3487595	24495865	24211713	286602
Begonia x hiemalis	Solenia	White	6.3534957	21178319	20947475	230844
Begonia x hiemalis	Baladin	Red	6.6087252	22029084	21641372	387712
Begonia x hiemalis	Baladin	Red	6.0921528	20307176	20079736	227440
Begonia x cheimantha	F1 "Gloire de Lorraine"	Pink	5.8513065	19504355	19239096	265259

GC content of each horticultural variety was compared with GC content of Begoniaceae species (table 4). The mean GC content across all accessions sampled was 38%. All accessions sampled had a GC content between a range of 10% from 34% to 43%, except the *Begonia* x hiemalis "Baladin" from Beekenkamp. The *Begonia* x hiemalis "Baladin" from Beekenkamp had the highest GC content of 53%, 10% higher than the next highest, *Begonia* x cheimantha F1 which had 43%. The mean GC content of the horticultural varieties was 41%, 5% higher than the mean of the Begoniaceae species

GC contents at 36%. This difference in means of GC content between the Begoniaceae species and horticultural varieties was statistically significant (P=0.02).



c)

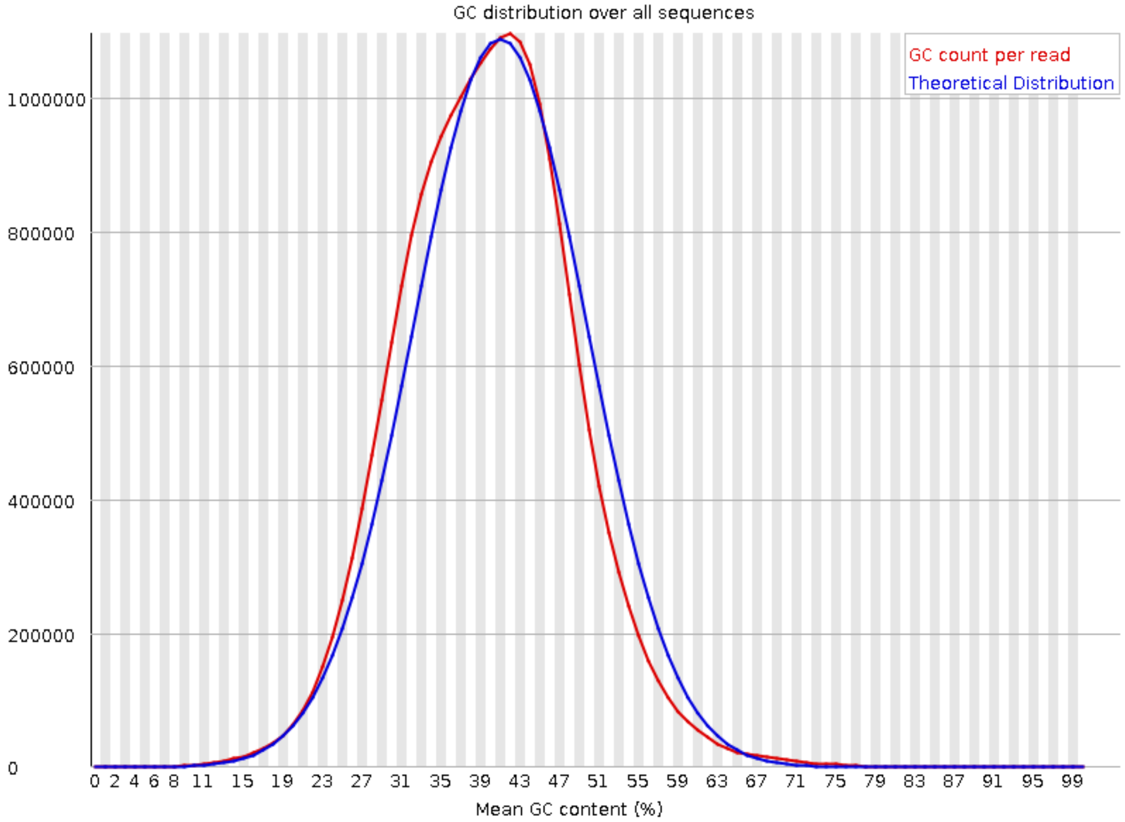


Figure 7a-b GC content curves of selected horticultural hybrid *Begonia* accessions

a) *Begonia* x cheimantha F1 b) Beekenkamp *Begonia* x hiemalis “Baladin” c) *Begonia* x hiemalis “Valentino Pink”. *Begonia* x hiemalis “Valentino Pink” has a typical GC content curve shape, while *Begonia* x cheimantha and Beekenkamp *Begonia* x hiemalis “Baladin” have double-peak curves.

Table 4 GC content of all accessions sampled.

Accession	GC%
Begonia heracleifolia B	35
Begonia heracleifolia A	36
Begonia conchifolia	38
Begonia luxurians	36
Hillebrandia sandwichensis	34
Begonia dregei	38
Begonia bipinnatifida	37
Begonia socotrana	36
Begonia x hiemalis “Valentino Pink”	39
Begonia x hiemalis unknown	39
Red *Begonia* x hiemalis "Solenia"	39
Pink *Begonia* x hiemalis "Solenia"	38
White *Begonia* x hiemalis "Solenia"	39
Beekenkamp *Begonia* x hiemalis "Baladin"	53
Dunbar *Begonia* x hiemalis "Baladin"	38
Begonia socotrana x *Begonia dregei* F1	43
Unknown Texan horticultural hybrid	40

3.2 RepeatExplorer2 and TAREAN

Results from RepeatExplorer 2 show that there is a large variation in repeat content of genomes sampled (figure 8). The horticultural accessions of *Begonia* sampled have a higher mean proportion of sequences in analysed clusters than the *Begonia* species by almost 10%, with a mean of 45.63% compared to 34.25%. A t-test shows that this result is statistically significant at 95% (P=0.042). The plants with the highest and lowest proportion of sequences in analysed clusters were both horticultural specimens; 61% of the sequences from the unknown horticultural *Begonia* specimen from Texas were identified as repeat elements compared to just 10% of the *Begonia* x hiemalis "Baladin" from Beekenkamp. As a result of this, the standard deviation of proportion of clustered sequences in reads from horticultural varieties was almost five times higher than that of the Begoniaceae species reads (16.7% vs 5.5%). The individual accession with the highest variation in repeat content between samples was *Hillebrandia sandwichensis*, all other accessions had standard deviations from their mean repeat content of less than 2.5%. The results from TAREAN were extremely similar to the results from RepeatExplorer2 (figure 9). This suggests almost all of the repeats placed in analysed clusters by RepeatExplorer2 are also tandem repeats, or that repeat content and tandem repeat content are strongly correlated. The most common repeat elements in *Begonia socotrana* were Ty3-*gypsy* retrotransposons. This was also the case for seven of the horticultural *Begonia* hybrids (figure 10). The most common repeated sequences in the unknown *Begonia* variety from Texas and pink *Begonia* x hiemalis "Solenia" from Glasgow were unable to be categorised into a class of repeat elements by RepeatExplorer2. The shape of the cluster in the pink *Begonia* x hiemalis "Solenia" from Glasgow shows that most repeats have very high sequence similarity, suggesting simple repeats or the birth of a new type of repeat element.

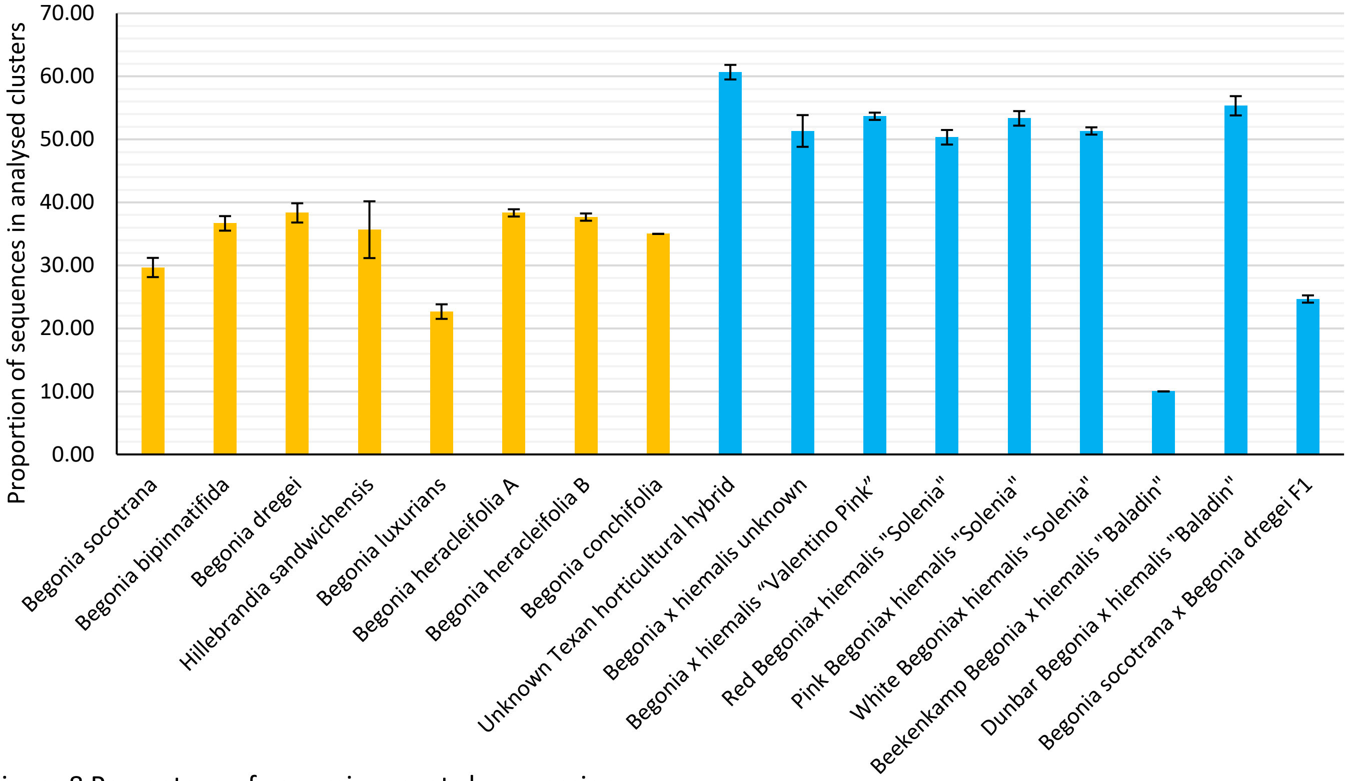


Figure 8 Percentage of genomic repeats by accession

Displays the mean percentage of accession's genome that has been pulled into analysed clusters of repeats by RepeatExplorer2 from three replicates. Error bars show ± Standard deviation.

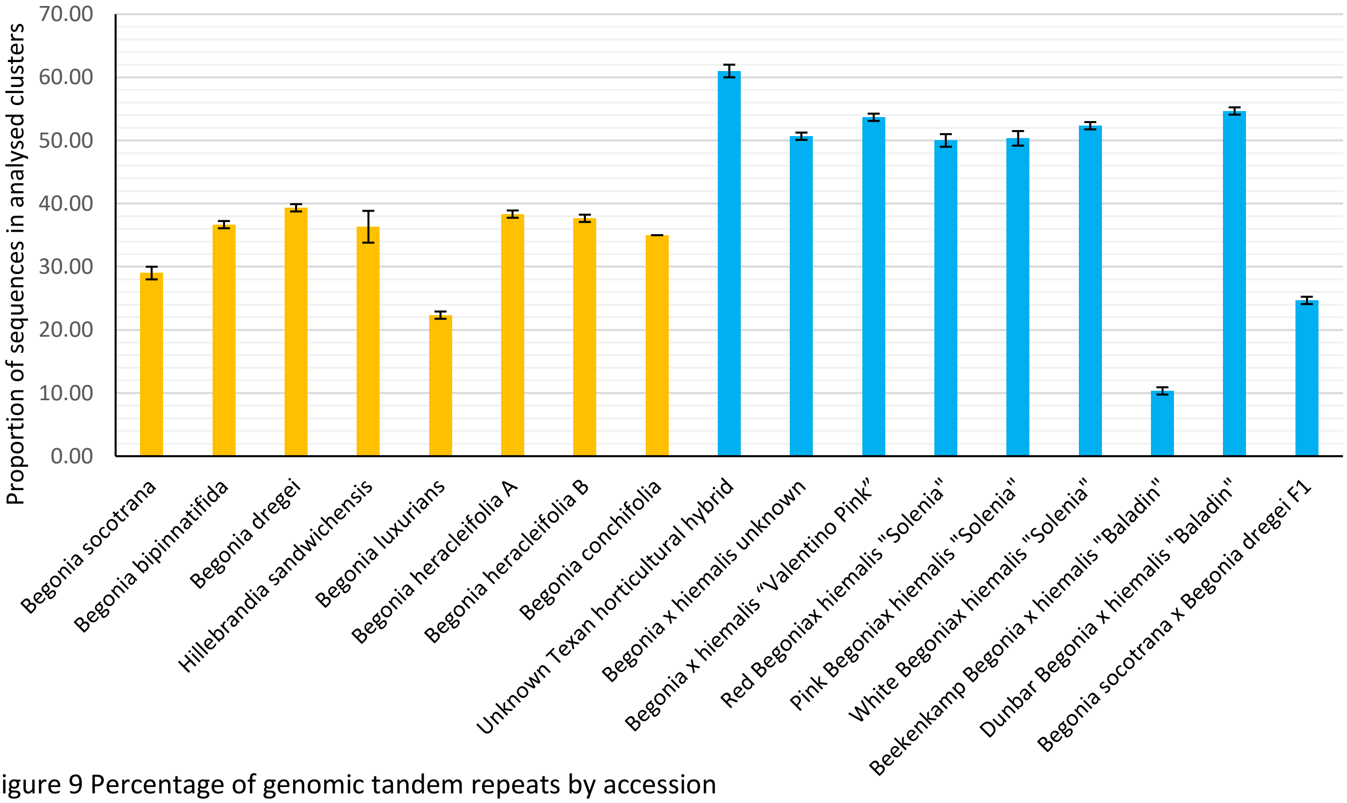


Figure 9 Percentage of genomic tandem repeats by accession

Displays the mean percentage of accession's genome that has been pulled into analysed clusters of tandem repeats by TAREAN from three replicates. Error bars show ± Standard deviation.

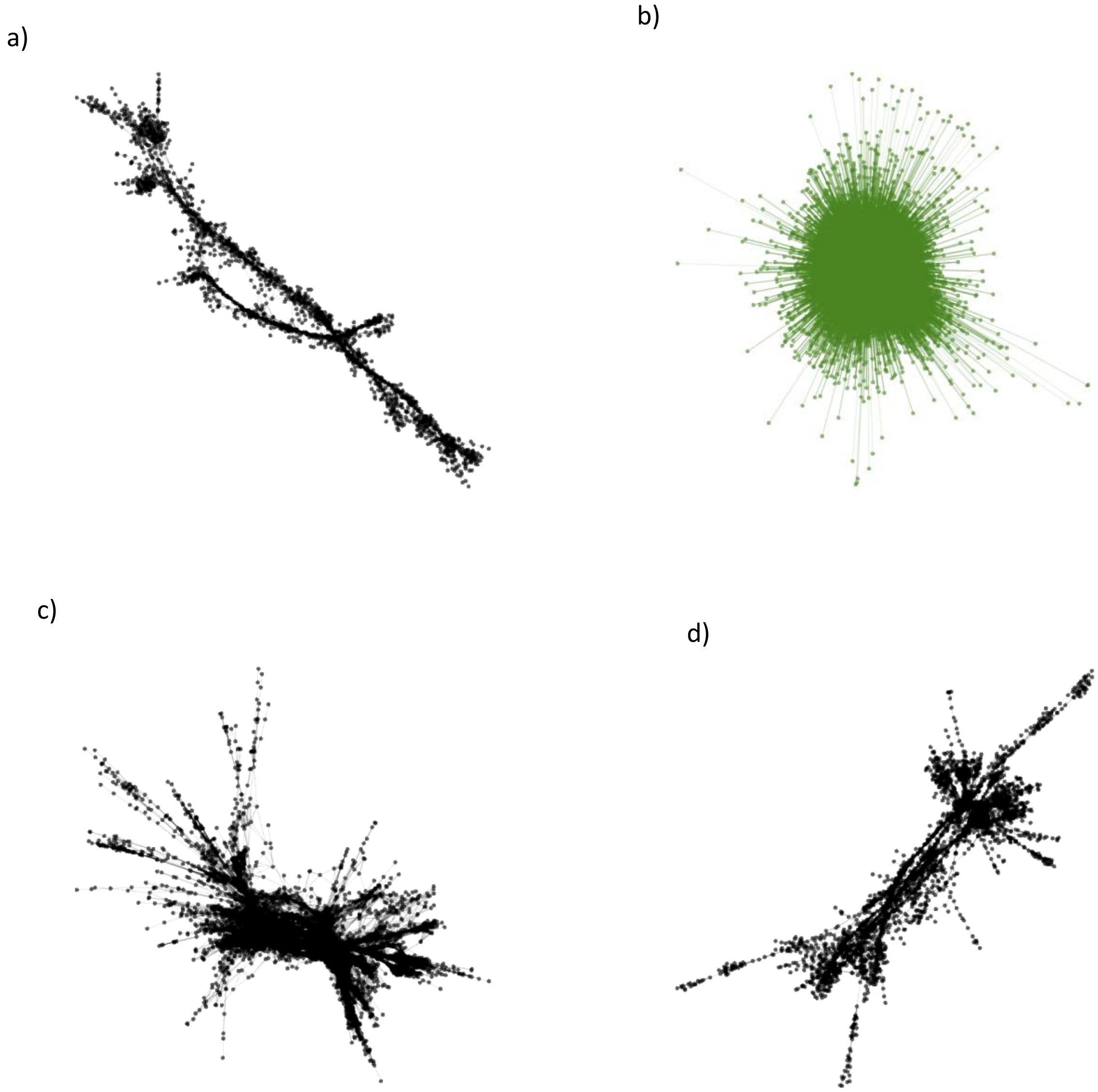
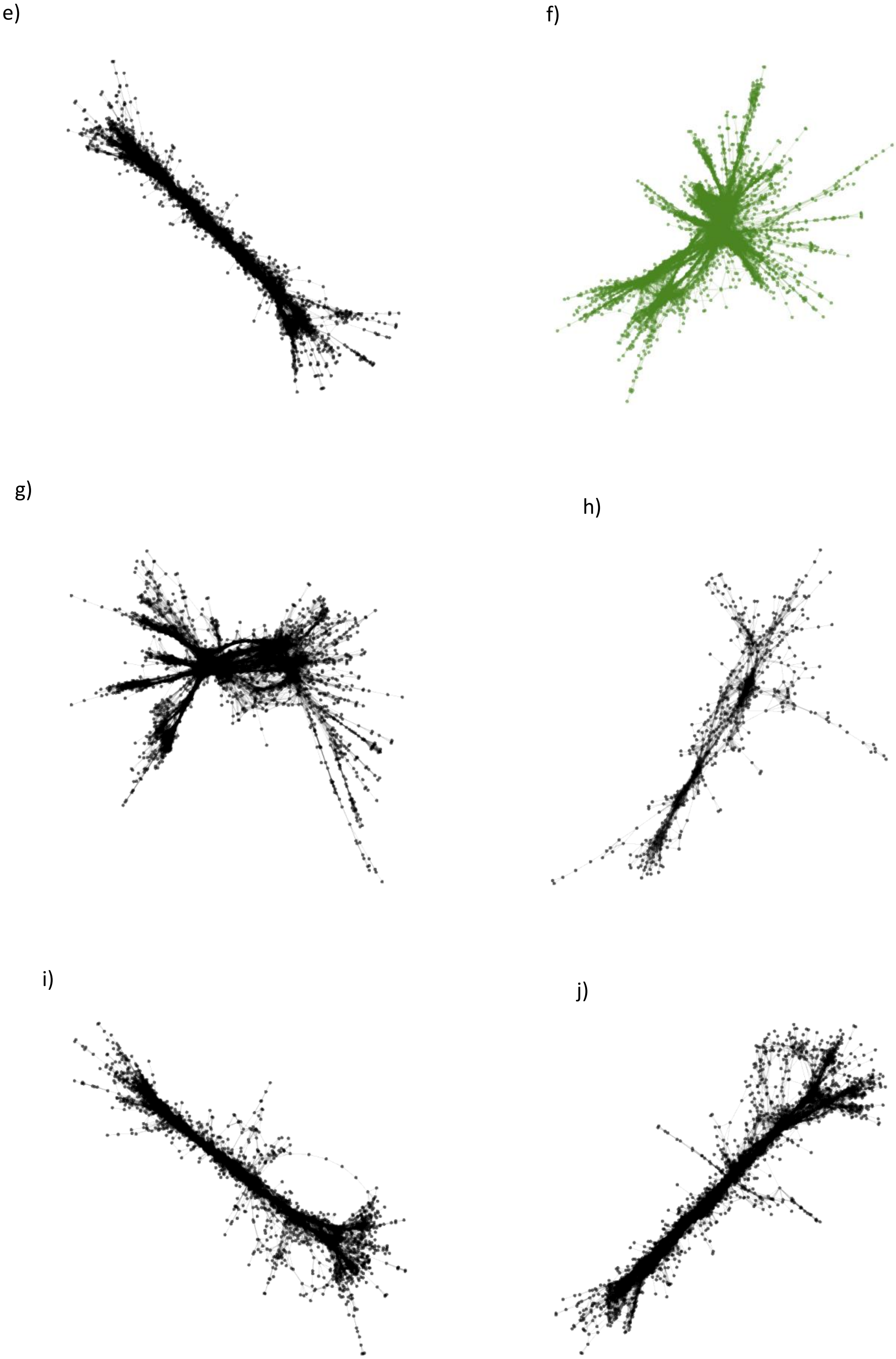


Figure 10a-j Similarity of sequences in largest cluster of repeats by accession

a) *Begonia socotrana*
b) unknown horticultural variety of *Begonia*
c) *Begonia* x hiemalis “Valentino Pink”, Taipei flower market,
d) Unknown *Begonia* x himealis, Taipei flower market,
e, f, g) red, pink, white *Begonia* x himealis “Solenia”, Glasgow Botanic garden,
h) *Begonia* x hiemalis “Baladin”, Beekenkamp,
i) *Begonia* x hiemalis “Baladin”, Dunbar,
j) *Begonia socotrana* x *Begonia dregei* F1, RBGE.

Largest clusters are determined by RepeatExplorer2. Each dot in the cluster represents one read, and distance of the dots is determined by genetic similarity with more similar reads being shown as closer together dots. Clusters of Ty3-*gypsy* retrotransposons repeats are black, and clusters of repeats of unknown identity are green.


e)
f)
g)
h)
i)
j)

3.3 Optimisation of Bowtie2 parameters

The proportion of reads from the *Begonia socotrana* genome that mapped back to the *Begonia socotrana* genome was consistently higher than the proportion of reads from the *Begonia conchifolia* genome that mapped back to the *Begonia socotrana* genome using the same parameters (figure 11). The percentage of reads mapped back the *Begonia socotrana* genome increases as the minimum score value decreases until a plateau is reached, with percentage of reads mapped increasing more quickly at high minimum score thresholds than low minimum score thresholds. The *Begonia socotrana* reads reach a plateau at a higher minimum score threshold than the *Begonia conchifolia* reads. For *Begonia socotrana*, a plateau was reached at a minimum score of approximately -2.25, and the percentage of reads mapped back to the genome did not increase beyond 97.7%, even when the minimum mapping threshold was decreased to -3. When mapping *Begonia conchifolia* reads back to the *Begonia socotrana* genome with decreasing minimum scores, the results did not follow a simple curve like the *Begonia socotrana* reads did. Similarly to the *Begonia socotran*a reads, the percentage of *Begonia conchifolia* reads mapped back to the *Begonia socotrana* genome increases as minimum score threshold decreases. However, unlike the *Begonia socotrana* reads, the percentage of reads mapped back does not increase quickly at high minimum score thresholds. The sharper increase in percentage of *Begonia conchifolia* reads mapped back to the *Begonia socotrana* genome occurs between minimum score thresholds of approximately -2.2 and -2.5, where the percentage of reads mapped back to the *Begonia socotrana* genome increases from 14.14% to 22.52%. A plateau in the number of *Begonia conchifolia* reads mapped back to the *Begonia socotrana* genome is reached at a minimum score threshold of approximately -2.7, and the percentage of *Begonia conchifolia* reads aligned does not increase past 22.76% at minimum score thresholds lower than -2.7.

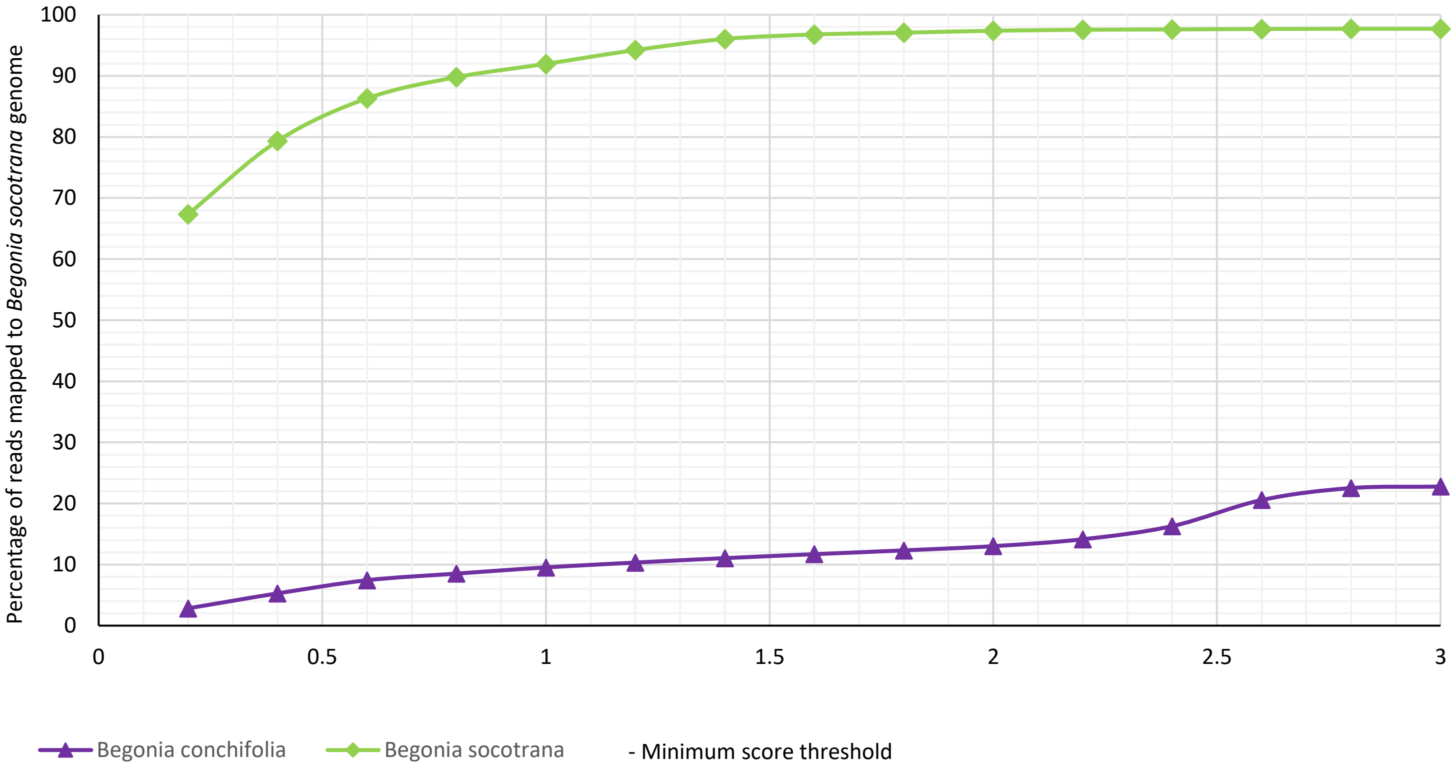


Figure 11 Percentage of genomic reads mapped to *Begonia socotrana* genome using different minimum score thresholds of Bowtie2

Results from using genomic reads from *Begonia socotrana* are shown in green, and results using genomic reads from *Begonia conchifolia* are shown in purple.

3.4 Mapping of genomic reads from horticultural varieties and Begoniaceae species to the *Begonia* bait set

The percentage of each accession's genome or genome skim that mapped back to the *Begonia* bait set using Bowtie2 was low (table 5). The mean percentage of each accession mapped back to the *Begonia* bait set was 0.90%, with a standard deviation of 0.24%. *Hillebrandia sandwichensis* had the lowest percentage of its genome mapped to the *Begonia* bait set of all the accessions sequenced, at 0.57%. Due to the large number of reads present in each genome and genome skim, the numbers of SNPs and indels found in each species using Bowtie2 was high (table 5). As to be expected, the number of indels found was consistently lower than the number of SNPs. The mean numbers of SNPs and indels found in a single accession were 89 803 and 1017 respectively. *Hillebrandia sandwichensis* had only 77 indels, over five times lower than the number found in *Begonia luxurians*, the accession with the second least number of indels. Interestingly, however, *Begonia bipinnitafida* had the lowest number of SNPs, not *Hillebrandia sandwichensis*. Using a t-test, it can be seen that horticultural varieties have a significantly higher number of SNPs and indels than the Begoniaceae species, to the P=0.05 and P=0.01 thresholds respectively. The SNPs and indels identified follow the same pattern.

Table 5 Genomic reads mapped to the *Begonia* bait set

Accession	bowtie2 alignment %	Number of SNPs	Number of indels
Begonia heracleifolia A	0.98	100172	972
Begonia heracleifolia B	0.96	99741	1005
Begonia conchifolia	1.08	6398	592
Begonia luxurians	1.06	94585	444
Hillebrandia sandwichensis	0.57	73769	77
Begonia dregei	0.69	42947	513
Begonia bipinnatifida	0.67	31894	490
Begonia socotrana	1.29	48024	523
Begonia x hiemalis “Valentino Pink”	1.44	119576	1379
Begonia x hiemalis unknown	1.16	107547	1231
Red *Begonia* x hiemalis "Solenia"	1.1	110512	1253
Pink *Begonia* x hiemalis "Solenia"	0.37	108830	1224
White *Begonia* x hiemalis "Solenia"	1.09	105589	1197
Beekenkamp *Begonia* x hiemalis "Baladin"	0.95	44037	507
Dunbar *Begonia* x hiemalis "Baladin"	1.08	96690	1077
Begonia socotrana x *Begonia dregei* F1	1.1	133471	1790
Unknown Texan horticultural hybrid	0.89	102871	1017

3.5 Identification of *Begonia socotrana* specific SNPs and indels

Begonia socotrana had 48024 SNPs and 523 indels. This is in the fourth quartile of number of SNPs and indels. Of the 48 024 SNPs and 523 indels, 2730 SNPs and 56 indels were found in none of the Begoniaceae species except *Begonia socotrana*, and hence these SNPs and indels are specific to *Begonia socotrana*.

All horticultural varieties of *Begonia* sequenced in this study contained *Begonia socotrana* specific SNPs and indels (table 6), and again, the variation of SNPs and indels followed the same patterns. Six of the eight known horticultural varieties contained close to 2000 socotrana specific SNPs. The F1 from *Begonia socotrana* x *Begonia dregei* had 2519 socotrana specific SNPs, the highest number of socotrana specific SNPs from any of the horticultural varieties. The *Begonia* x hiemalis “Baladin” from Dunbar had the highest number of *Begonia socotrana* specific indels at 54. The *Begonia* x hiemalis “Baladin” from Beekenkamp had roughly half the number of SNPs and indels as the other horticultural varieties sampled, containing 1092 SNPs and 507 indels. Of the *Begonia socotrana* specific SNPs and indels, 954 SNPs and 17 indels were found to be common to all known horticultural varieties of *Begonia* sampled. The unknown *Begonia* from Texas contained the least number of *Begonia socotrana* specific SNPs and indels by far, containing just 286 SNPs and 3 indels out of the set specific to *Begonia socotrana*.

Table 6 Number of *Begonia socotrana* specific variants found in horticultural hybrid *Begonias*

Accession	Number of Begonia *socotrana* specific SNPs	Percentage of total Begonia *socotrana* specific SNPs	Number of Begonia *socotrana* specific indels	Percentage of total Begonia *socotrana* specific indels
Begonia x hiemalis “Valentino Pink”	2279	83.48	48	85.71
Begonia x hiemalis unknown	2078	76.12	47	83.93
Red *Begonia* x hiemalis "Solenia"	2127	77.91	45	80.36
Pink *Begonia* x hiemalis "Solenia"	2033	74.47	40	71.43
White *Begonia* x hiemalis "Solenia"	2047	74.98	42	75.00
Beekenkamp *Begonia* x hiemalis "Baladin"	1092	40.00	20	35.71
Dunbar *Begonia* x hiemalis "Baladin"	1989	72.86	41	73.21
Begonia socotrana x *Begonia dregei* F1	2519	92.27	54	96.43
Unknown Texan horticultural hybrid	286	10.48	3	5.36

Chapter 4 - Discussion

4.1 GC content

The horticultural varieties of *Begonia* had a significantly higher GC mean content than the mean GC content of the Begoniaceae species (table 3). GC content tends to correlate positively with plant genome size (Vesely et al., 2011). Polyploid hybrids will naturally have larger genomes than their parent plants, and most likely also have larger genomes than species closely related to their parent plants. All *Begonia x* hiemalis are triploid, as while *Begonia socotrana* is diploid, the *Begonia* x tuberhybrida are tetraploid. This is likely to account for the increase in GC content seen in the *Begonia* x hiemalis accessions sampled. Results suggest that the unknown *Begonia* cultivar may also be polyploid as it also has a high GC content of 40%. Flow cytometry could be used to ascertain the ploidy level of the horticultural hybrid varieties to further the understanding of the GC content of each genome. Hybridisation can cause activation of transposable elements in plant genomes, and high levels transposable elements can increase genome size considerably (Kidwell, 2002). The possibility of GC rich transposons would further explain the higher GC content of the horticultural varieties of *Begonia* when compared to the GC content of the *Begonia*ceae species.

4.1.1 Unusually high GC content

While the GC content of all accessions sampled was relatively high on the scale of plant GC content, (Šmarda et al., 2014) the *Begonia* x hiemalis "Baladin" from Beekenkamp in particular had an especially high GC content of 53%. This was surprising seeing as the GC content of the *Begonia* x hiemalis from Dunbar, also assumed to be a "Baladin" is much lower at 38%. This value of 38% is far more in line with what is seen in the rest of the accessions sampled as part of this study. A GC content of 53% is unlikely to be biologically real, meaning error or contamination has most likely occurred during lab work or sequencing. This is supported by the fact that the highest known GC content of any plant species is 49% (Šmarda et al., 2014).

The Beekenkamp *Begonia* x hiemalis “Baladin” had a GC content curve with two peaks (figure 7). For a double-peaked GC content curve, two smooth peaks as opposed to one sharp and one smooth, suggests there is contamination from another organism with a different GC content rather than from a proliferation of repeats, which would group together more closely to form a sharper peak. If the sample was contaminated, the high GC content of the second GC peak suggests the cause of the contaminant was not a plant. BlobTools partitions reads by GC content to tease out reads from potential contaminants and uses BLAST to identify the taxonomy of contaminants present (Laetsch and Blaxter, 2017). Running the genomic reads from the Beekenkamp *Begonia* x hiemalis “Baladin” specimen through BlobTools may indicate presence of a contaminant. The presence of a bacterial endophyte present in the leaf used for DNA extraction sampled could explain the second peak of the GC content curves, as some species of bacteria can have high genome GC contents of up to 75% (Brocchieri, 2014).

4.2 RepeatExplorer2 and TAREAN

The results of the RepeatExplorer2 and TAREAN analysis were extremely similar (figures 8 and 9). This suggests that all repeats detected using RepeatExplorer2 were tandem repeats, or at least had a high percentage of tandem repeats within the repeated sequence. This is surprising, as a large proportion of the horticultural hybrid *Begonia* genomes were expected to be comprised of transposable elements, for reasons that are further discussed. As the results from both outputs are so similar, only the results from RepeatExplorer2 will be discussed as it can be assumed the same reasoning would apply to analysis of the TAREAN results.

4.2.1 Repeat content of accessions

It was expected that the horticultural hybrid varieties of *Begonia* would have higher repeat content than the *Begonia* species. When two different plant genomes come together in a hybrid, a burst of retrotransposon activity can be triggered (Zhang et al., 2018). This phenomenon is known as genome shock, and can occur under a variety of conditions including extreme heat and cold. It is theorised that in response to a genome shock, the genome must rearrange itself to overcome a threat to its survival (McClintock, 1984). As all known horticultural varieties sampled came from hybrid backgrounds, the combination of the parent genomes is likely to have caused genome shock, giving rise to the proliferation of repeat elements in all the horticultural varieties. This is especially true given that, of the horticultural hybrid varieties sampled with known parents, the parents are distantly related. The parents of the *Begonia* x hiemalis group are *Begonia socotrana* and *Begonia* x *tuberhybrida. Begonia* x *tuberhybrida* has been formed from several South American species of *Begonia*, including *Begonia boliviensis* and *Begnoia pearcei* (Hughes, 2002). The South American clade of *Begonia* diverged from the African clade, containing *Begonia socotrana*, over 20 million years ago (Moonlight et al., 2018). During these years, each plant will have adapted to considerably different conditions, and their genomes will have diverged accordingly. Hence it is likely that genomes of the parents are very different and caused a major genome shock event when coming together, activating many repeat elements. The higher mean repeat content of the horticultural varieties supports this. The unknown horticultural *Begonia* had the highest repeat content of all accessions sampled, further supporting the assumption that the specimen is indeed a hybrid of some sort, most likely polyploid, and not a true species.

4.2.2 Repeat content in the genus *Begonia*

The genus *Begonia* underwent a whole genome duplication at base of the genus (Brennan et al., 2012). Whole genome duplication events can also trigger genome shock, causing proliferation of repeatable elements in the genome. However, Hillebrandia sandwichensis, outside of the genus

Begonia had a mean repeat content of 36%, much in line with the *Begonia* species sampled. It is possible that the whole genome duplication occurred at the base of the family Begoniaceae rather than the genus *Begonia*, or that Hillebrandia has also undergone some form of genome shock to elevate the level of repeat elements in the genome. Overall, as only seven species were sampled from the genus *Begonia* which, is the sixth largest angiosperm genus containing almost 2000 species (Moonlight et al., 2018), insight into the evolution and diversity of the whole genus is unlikely to be seen from this small sample.

4.2.3 High GC content and low repeat count

Two samples of horticultural varieties of *Begonia* had surprisingly low repeat content. The *Begonia* x cheimantha F1 had a mean repeat content of 25%, while the *Begonia* x hiemalis "Baladin" from Beekenkamp had a mean repeat content of just 10% (figure 8). While 25% is not especially low repeat content in comparison to the other results, it is odd given the fact that it is the direct offspring from two distantly related species of *Begonia*. It is more surprising given the fact that both parents had higher repeat contents than the F1 they produced. It is worth remembering that only one *Begonia dregei* and one *Begonia* was sampled, and that the F1 was not produced from the *Begonia dregei* and *Begonia socotrana* sequenced. Hence it is possible that there is a high level of variation between individual plants within the parent species that is unseen due to the sequencing plan employed for this project. However two accessions of *Begonia heracelifolia* were sequenced, and both accessions had a mean repeat content of 38%, with a low standard deviation of just 0.57%. The pedigree of these two specimens is not known, and the two accessions could be very closely related, but it does suggest that within species variation of repeat content is low.

The repeat content of the *Begonia* x hiemalis "Baladin" from Beekenkamp is extremely low compared to the other horticultural varieties and Begoniaceae species sequenced as part of this study (figure 8). The combination of low repeat content and high GC content found in the *Begonia* x hiemalis "Baladin" from Beekenkamp marks it out as an unusual specimen. The low repeat content found means that the high GC content of the *Begonia* x hiemalis “Baladin” from Beekenkamp cannot be attributed to high levels GC rich repeatable elements or GC tandem repeats. Therefore, it seems likely that either the *Begonia* x hiemalis “Baladin” from Beekenkamp was tainted with a GC rich contaminant during labwork, or there were errors made in sequencing causing a high GC content. *Begonia* x cheimantha F1 was the other specimen of *Begonia* horticultural variety that had a double peaked GC content curve, and it also had a lower repeat content than expected. It is unlikely that the second peak on the GC content curves of these specimens was caused by a high level of repeated sequences which were too divergent to be recognised as a cluster by RepeatExplorer2 or TAREAN. This is because the high GC content sequences must have been similar in order to have caused a second peak on the GC content curve.

4.2.4 Analysis of repeat clusters

The most common repeat elements in seven of the horticultural hybrid accessions of *Begonia* were Ty3-*gypsy* retrotransposons (figure 10). This was also the most common class of repeat in *Begonia socotrana*. Ty3-*gypsy* retrotransposons are a type of Long Terminal Repeat retrotransposon (LTR), the most common class of repetitive sequence in plant genomes (Vitte and Panaud, 2005). The presence of long terminal repeat regions that flank the internal coding region of the retrotransposon may explain why the results of TAREAN and RepeatExplorer2 were so similar. The shape of the clusters of Ty3-*gypsy* retrotransposon sequences differs considerably between accessions, and each cluster is fairly spread out. This indicates that there is a wide genetic diversity within the Ty3-*gypsy* retrotransposons found in the horticultural hybrid accessions.

RepeatExplorer2 was unable to categorise the most common repeated sequences in the unknown *Begonia* variety from Texas and pink *Begonia* x hiemalis “Solenia” from Glasgow. The shape of the cluster in the pink *Begonia* x hiemalis “Solenia” from Glasgow shows that most repeats have very high sequence similarity, suggesting simple short repeats e.g. tandem repeats. Alternatively, this high-density cluster may be indicative of a recent burst of transposon activity. Under this scenario, the proliferated regions of DNA would not yet have accumulated random mutations over time, so still have very similar sequences. The fact that this sequence is unknown may mean that the pink *Begonia* x hiemalis "Solenia" from Glasgow contains a new type of transposon.

4.3 *Begonia socotrana* specific SNPs

Relatively high levels of the *Begonia socotrana* specific SNPs and indels were shared with seven of the eight horticultural varieties of *Begonia* sequenced as part of this study. All "Solenia" specimens had similar numbers of indels and SNPs shared with *Begonia socotrana*, as was expected given their similar genetic background (table 5). The high level of *Begonia socotrana* specific SNPs in the horticultural varieties is surprising, as it implies that a large proportion of the *Begonia socotrana* genome is present in the horticultural variety genomes, and that specific alleles have not been selected for by breeders.

4.3.1 *Begonia socotrana* has higher levels of genetic diversity than expected

Of the 2730 SNPs and 56 indels identified as being specific to *Begonia socotrana*, 2519 of the SNPs and 41 of the indels were also found to be present in the F1 from the *Begonia dregei x socotrana* cross (table 6). This cross can be seen as a control for the identification of the *Begonia socotrana* SNPs; SNPs within this accession can only be attributed to *Begonia socotrana* or *Begonia dregei*. It was thought that because *Begonia socotrana* is highly homozygous, the number of *Begonia socotrana* SNPs within the F1 from the *Begonia dregei x socotrana* cross would be the same as the total number of *Begonia socotrana* specific SNPs, as all would have been imparted to progeny. However, the *Begonia dregei x socotrana* cross contained 211 fewer SNPs and 15 fewer indels. Even when accounting for the low heterozygosity of the genome, which has been estimated at 0.37% based on k-mer analyses (L Campos Dominguez pers comm., 2019), assuming random distribution of SNPs throughout the genome one would expect 2725 SNPs to be found (2730 – (0.0037 x 2730)/2).

The fact that not all of the *Begonia socotrana* specific SNPs were found in the F1 *Begonia socotrana x dregei* genome skim suggests that the parent of the cross was not the specific accession of *Begonia socotrana* that had its genome sequenced. The eight percent difference in SNP number implies that there is relatively wide amount of genetic diversity in the *Begonia socotrana* collections kept at RBGE. This is surprising given the highly homozygous nature of the genome, and implies that the accession sequenced and the accession that parented the cross may have been genetically isolated and have been collected from separate populations of *Begonia socotrana* on Socotra.

4.3.2 Identity of the unknown *Begonia*

The unknown Texas *Begonia* shared the fewest SNPs with *Begonia socotrana*, as was expected based on morphology alone. However, this unknown cultivar still shared 10% of the *Begonia socotrana* specific SNPs, a higher proportion than expected (table 6). It is worth noting that the 283 SNPs found in the unknown Texas *Begonia* is similar number to the 211 SNPs that were not found in the *Begonia* x cheimantha F1. This pattern is also seen in the indels; three indels were shared between *Begonia socotrana* and the unknown Texan horticultural *Begonia*, while two were "missing" from the *Begonia* x cheimantha F1.

4.3.3 *Begonia* x hiemalis "Baladin" from Beekenkamp

The number of SNPs shared between the *Begonia* x hiemalis "Baladin" from Beekenkamp was approximately half of what one would expect it to have been based on the results of the other horticultural hybrid *Begonia*. This is especially true given the results of the other *Begonia* x hiemalis "Baladin" from Dunbar. The fact that the number of SNPs was so close to exactly half raises the theory that the Beekenkamp *Begonia* x hiemalis "Baladin" may actually have been an F2, and crossed with another *Begonia* following an initial cross that involved *B. socotrana.* This would follow simple Mendelian genetics. However, the *Begonia* breeder at Beekenkamp thought that while not impossible, it was unlikely that this mistake had occurred (S. Oostvogels pers comm., 2019).

This raises the question; with such a high GC content, low repeat content and low shared percentage of SNPs, does the sequence from the Beekenkamp *Begonia* x hiemalis “Baladin” even represent Begonia? Further mapping experiments was carried out using the *Begonia socotrana* genome to examine if this was the case, using the F1 *Begonia dregei x socotrana* as a control. Reads from the F1 *Begonia dregei x socotrana* cross were mapped back first, and 55.3% of the F1’s reads mapped back/ This is roughly what would be expected from an F1 hybrid of *Begonia socotrana*. In contrast, just 9.1% of reads from the Beekenkamp *Begonia* x hiemalis “Baladin” mapped back to the *Begonia socotrana* genome. 9.1% is an extremely low given twice as many reads from *Begonia conchifolia* mapped back to the *Begonia socotrana* genome under the same parameters (figure 11), and Begonia conchifolia diverged from Begonia socotrana over 20 million years ago. Hence it seems the sequence data attributed to the Beekenkamp *Begonia* x hiemalis “Baladin” has come from another source, so the Beekenkamp *Begonia* x hiemalis “Baladin” results should not be used in any further analyses.

4.4 Baits of interest pertaining to *Begonia socotrana*

It was hypothesised that *Begonia socotrana* specific SNPs and indels would be located on baits for genes that encode traits of interest to breeders. Breeders would presumably have performed many crosses between *Begonia socotrana* and *Begonia* x tuberhybrida or *Begonia dregei,* and only propagated the individuals with the most desirable phenotypes. Wintering flowering time was the trait of most interest to *Begonia* breeders in the 18th century, and remains the most valuable to this day because winter-flowering plants are popular due to their scarcity (Hvoslef-Eide et al., 1995). Other *Begonia socotrana* traits that are of interest to *Begonia* breeders are hairless leaves and longevity of flowers (S. Oostvogels pers comm., 2019).

4.4.1 *FLOWERING LOCUS T (FT)*

FT is critical for floral induction in all angiosperms it has been studied in so far (Adeyemo et al., 2017). *FT* codes for a phosphatidylethanolamine-binding proteins that promotes the developmental transition from vegetative growth to reproductive growth in plants. *FT* expression is induced by the transcription factor CONSTANS (CO), which is regulated by the circadian clock. *FT* expression is tightly controlled by further transcription factors and photoreceptors to ensure flowering occurs at the correct day length (Wickland and Hanzawa, 2015). For *Begonia socotrana,* this is short day length, so the transition to flowering will only occur once the length of the night exceeds the critical photoperiod for the species. Unfortunately there was no bait based on the coding sequence for FT or CO in the *Begonia* bait set used.

Using the *Begonia* BLAST service hosted by Royal Botanic Garden Edinburgh, the position of FT in the *Begonia socotrana* genome was located by using the coding sequence for FT from the *Cucumis sativus*, hosted on Phytozome (Goodstein et al., 2012). 1000bp upstream of the start of the coding sequence of FT in the *Begonia socotrana* genome was selected and run through LTR_Finder and MEGANTE to look for changes in the promoter region of FT in *Begnoia socotrana* that may affect the initiation of flowering time (Zhao and Wang, 2007). The promoter is a region of DNA that enables the initiation of transcription of a specific gene, and changes in promoter sequence can have large impacts on gene expression, as they can alter binding affinities and abilities of transcription factors (Li and Zhang, 2014). There were no obvious signs of a retrotransposon or other large repeat elements, which if present would be likely to mark a clear change in the ability of transcription factors to bind to the *FT* promoter region.

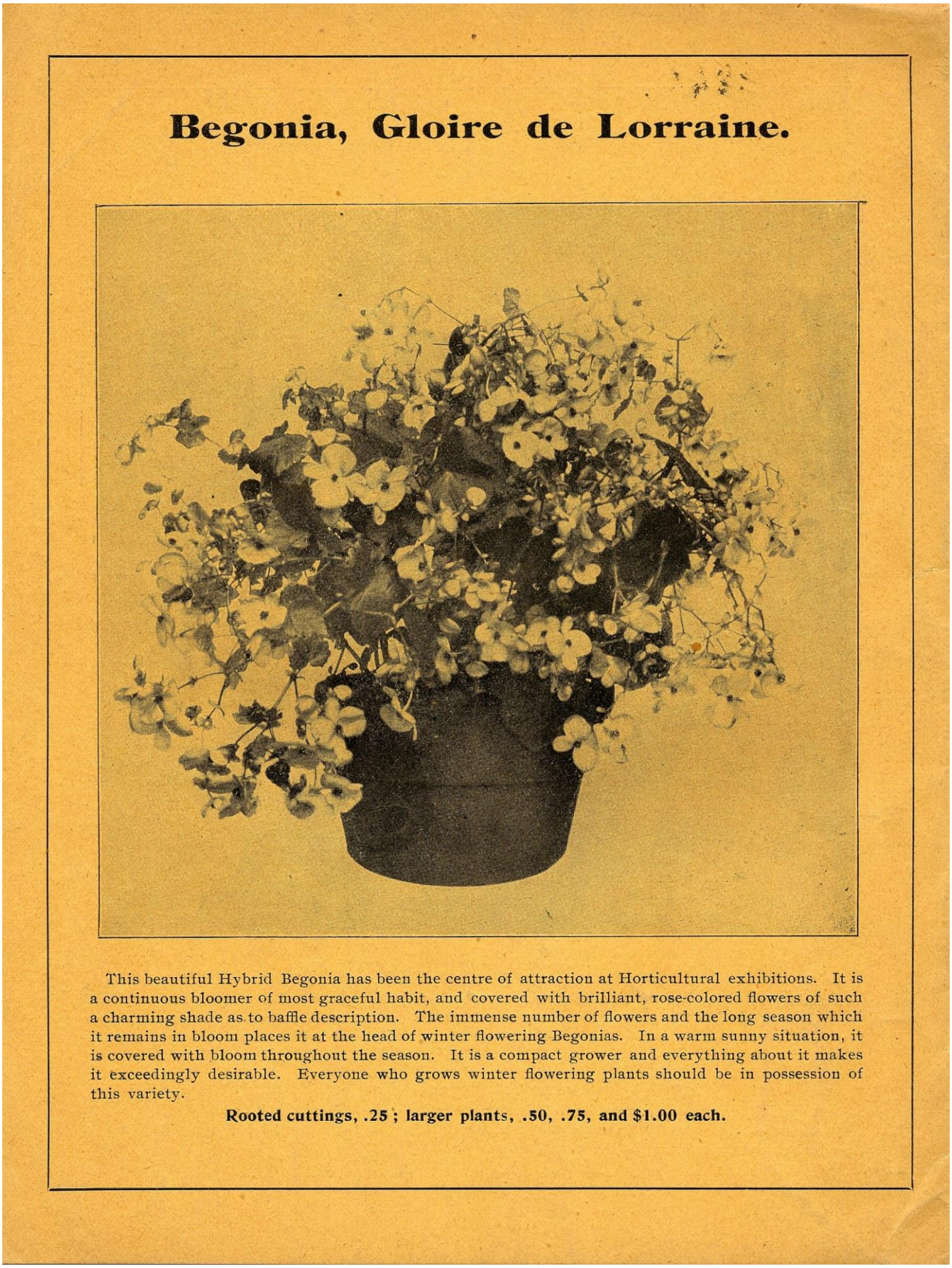
Changes in the *FT* gene sequence may cause the unusual short-day flowering time of *Begonia socotrana*, or the change could have occurred upstream of *FT* in the developmental pathway that regulates the transition from vegetative to reproductive growth. Phytochrome B (PHYB) is the member of the phytochrome family of photoreceptors responsible for detecting day length and determining flowering time (Endo et al., 2013). While a bait based on the coding sequence for PHYB was present in the *Begonia* bait set, no SNPs or INDELs specific to *Begonia socotrana* mapped to the bait based upon *PHYB*. This shows that there is no version of phytochrome B specific to *Begonia socotrana*, and that again, the genetic change leading to the shifted flowering time must be located elsewhere in the regulatory pathway for floral initiation.

4.4.2 *PHYTOCHROME AND FLOWERING TIME 1 (PFT1)*

PFT1 encodes a subunit of Mediator, a protein complex comprised of 30 subunits that acts as a molecular bridge between transcription factors and RNA polymerase II to initiate transcription. *PFT1* has been shown to be involved in the regulation of a diverse range of functions, including flowering time and trichome development. PFT1 negatively regulates the PHYB pathway to promote flowering, and the *PFT1* nucleotide sequence contains a series of short tandem repeats (STR). STR have very high mutation rates, yet small changes in regions of STR can cause vast phenotypic differences. It has been shown in *Arabidopsis thaliana* that flowering time can be delayed by increasing the number of STR in *PFT1* (Rival et al., 2014). Due to the high mutation rate of STR regions, the regions are more likely to differ between species, supporting this identity of *Begonia socotrana* specific SNP in this study. Flowering time is a trait of key interest to *Begonia* breeders working with *Begonia socotrana*, so SNPs were expected to be found on baits involved in the regulation of this process. Interestingly PFT1 is also involved in the regulation of trichome development, as it promotes papillae growth (Fornero et al., 2018). Hence a *Begonia socotrana* allele of *PFT1* may be partly responsible for the lack of leaf hairs seen in *Begonia* x hiemalis plants.

4.4.3 *TEOSINTE BRANCHED 1 (TB1)*

TB1 is a transcription factor that regulates branching architecture in plant growth (Doebly et al., 1995). As seen in figure 12, *Begonia* “Gloire de Lorraine”, the F1 from *Begonia dregei* and *Begonia socotrana* is a “compact grower” with a very dense and bushy habit, suggesting a highly specialised branching architecture (Schlegel and Fottler company, 1899). The fact that this growth habit is listed in an advertisement suggests the trait is desirable, and it is likely that this characteristic growth habit was selected for by breeders to maximise the number of flowers to come from one plant. As well as regulating vegetative branching architecture, TB1 interacts with FT to control branching in the inflorescence. Wheat plants overexpressing TB1 produce denser inflorescences with a greater number of flowers compared to wild type plants (Dixon et al., 2018). It is possible that the copy of *TB1* from *Begonia socotrana* is responsible for the bushy compact habit of *Begonia* “Gloire de Lorraine”. *Begonia dregei*, while phenotypically variable, tends to have a looser branching pattern, and has fewer flowers than *Begonia “*Gloire de Lorraine”.



Begonia, Gloire de Lorraine.

This beautiful Hybrid Begonia has been the centre of attraction at Horticultural exhibitions. It is a continuous bloomer of most graceful habit, and covered with brilliant, rose-colored flowers of such a charming shade as to baffle description. The immense number of flowers and the long season which it remains in bloom places it at the head of winter flowering Begonias. In a warm sunny situation, it is covered with bloom throughout the season. It is a compact grower and everything about it makes it exceedingly desirable. Everyone who grows winter flowering plants should be in possession of this variety.

Rooted cuttings, .25 ; larger plants, .50, .75, and $1.00 each.

Figure 12 *Begonia* “Gloire de Lorraine” advertisement

This 19th century advertisement for *Begonia* “Gloire de Lorraine” lists which traits are attractive to buyers, and hence also breeders (Schlegel and Fottler company, 1899).

4.5 Implications of findings

4.5.1 Genetic diversity of *Begonia socotrana*

Findings from this study imply that there is a higher than expected level of genetic diversity within the living collections of *Begonia socotrana* kept at Royal Botanic Garden Edinburgh. *Begonia socotrana* is an island endemic species, and many island endemic species have low levels of genetic diversity (Hamabata et al., 2019). The mountainous terrain of the Hajhir massif in which *Begonia socotrana* is found means that individual populations have a high degree of geographical isolation from one another. This geographical isolation can often translates to a high degree of reproductive isolation, as pollinators are unable to move between different populations easily and quickly. Despite this, the populations of *Begonia socotrana* have a mean estimated outcrossing rate of 0.86, with no single population having an inbreeding coefficient significantly greater than zero (Hughes et al., 2003). This study indicated that *Begonia socotrana* was in sufficiently good health as a species to have its IUCN conservation status changed from “endangered”, as it was classified in 1978, to “least concern” (Miller, 2004b). Species categorised by IUCN as “least concern” are thought to widespread and abundant in their habitat, and with no need for any conservation work to ensure the longevity of the species.

The decision to change the conservation status of *Begonia socotrana* was made based on work done on the population structure of the species done in the early 2000s (Hughes et al., 2003). It is now thought that the conclusions reached are outdated and may no longer valid (M. Hughes pers comm. 2019). When the decision was made to lower the IUCN Red List status of *Begonia socotrana* to least concern, the scale of the current climate crisis had not been fully understood. While *Begonia socotrana* is not threatened by land-use change as the Hajhir massif has been consistently grazed by goats, climate change is causing globally increased temperatures and decreased rainfall in the region (Deutscher Wetterdienst, 2019). In an already arid environment, *Begonia socotrana* may soon struggle to find enough water to survive. Furthermore, as *Begonia socotrana* populations are presently found at high altitudes, populations may not be able to migrate to land further north or to even higher altitudes in order to reach suitable habitats experiencing lower temperatures. Hence, it is likely that the populations of *Begonia socotrana* are decreasing in size and genetic diversity. The sister species of *Begonia socotrana* is *Begonia samhaienses* (figure 13), a species native to the small island of Samha in the Socotran archipelago (Hughes and Miller, 2002). *Begonia samhaensis* was considered endangered in 2004 due to threats from increased agriculture and severe drought (Miller and Morris 2004a), and it is thought that this may now be the status of *Begonia socotrana* fifteen years later (M. Hughes pers comm., 2019).

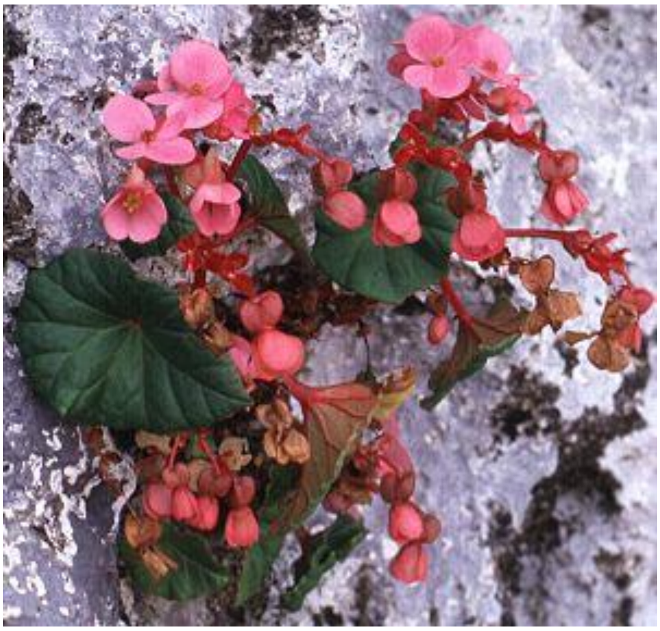


Figure 12 *Begonia samhaensis*

Begonia samhaensis grows on the hotter, drier island of Samha in the Socotran archipelago (Hughes, 2002).

A higher than expected level of genetic variation seen in samples of *Begonia socotrana* taken from the glasshouses at Royal Botanic Garden Edinburgh. These accessions were grown from seed collected from Socotra in February 2000. This study shows SNPs identified as specific to *Begonia socotrana* appear to vary by approximately 8% between accessions of *Begonia socotrana* grown in the research collections at Royal Botanic Garden Edinburgh. It is likely that the alleles found in horticultural varieties originate from the original 1880 collections. This is supported by the considerable degree of variation between *Begonia socotrana* specific SNPs found in accessions containing alleles from wild *Begonia socotrana* grown from seed in Edinburgh, and in the horticultural hybrids. Hence it can be said that these collections act as snapshots of the genetic diversity of *Begonia socotrana*, each taken at a time when the population of *Begonia socotrana* would have been greater in size, and likely more genetically diverse. It is probable that samples sequenced for this study contain alleles no longer found in the wild populations of *Begonia socotrana*. Collections of *Begonia socotrana* at Royal Botanic garden Edinburgh should be considered ex situ conservation. If the genetic diversity of *Begonia socotrana* decreases beyond a certain critical point, the species may undergo mutational meltdown (Lynch et al., 1995). If this were to happen, alleles from accessions of *Begonia socotrana* in Edinburgh, and potentially even from horticultural hybrids could be introgressed into the natural population of *Begonia socotrana* to artificially increase levels of genetic diversity.

4.5.2 *Begonia* breeding

The high level of genetic variation seen in the Edinburgh research collections may also be of interest to commercial *Begonia* breeders. Approximately 8% of the SNPs identified as specific to *Begonia socotrana* in this study may vary between individual accessions of horticultural varieties of *Begonia* which contain *Begonia socotrana* specific SNPs. This suggests that, while it is completely possible and perhaps likely that all Christmas *Begonia* did originate from the same accession collected by Balfour in the 1880s, there is a relatively large amount of genetic variation in the *Begonia socotrana* DNA within the horticultural hybrids. This could be attributed to the smuggling of other accessions of *Begonia socotrana*, which may have been used to found separate lineages of horticultural *Begonia* hybrids. However, there is no explicit evidence for this. It is more likely due to the prolonged period of increased transposon activity following the initial genome shock on hybridisation (Parisod et al,. 2010). The proliferation of transposable elements would cause different horticultural *Begonia* accessions to gain increasing numbers of different, random somatic mutations, hence becoming more genetically distant from one another. This is supported by the diversity of variation seen in the clusters of Ty3-*gypsy* retrotransposons in figure 10. Either way the surprisingly large amount of genetic variation present in the *Begonia socotrana* specific SNPs found in horticultural varieties sampled, suggests there is a higher level of diversity present in horticultural varieties than previously thought (Neale et al., 2006). The results of this study suggest there is ample genetic scope for breeders to utilise in their quest to create new and improved varieties of *Begonia*. Current targets for modern *Begonia* breeders include increased disease resistance, increasing the range of colours available, and production (S. Oostvogels pers comm., 2019).

Chapter 5 - Caveats and Future Research

5.1 Caveats

5.1.1 Use of bait set for alignment

Using the *Begonia* bait set as a reference sequence for this study came with large disadvantages. The main disadvantage is that any variations present in non-coding regions of the genome were unable to be utilised for variant calling. Non-coding regions of DNA contain many types of regulatory element including promoters, enhancers, silencers, and insulators. Changes in the nucleotide sequence of any of these regulatory elements have the potential to alter gene expression substantially, but changes in regulatory elements of genes of interest could not be looked into as part of this study as the baits are based on coding sequence only. Examination of non-coding regions may also have enabled mechanisms of posttranscriptional regulation to be examined. Nucleotide sequence changes in microRNAs (miRNAs) responsible for the posttranscriptional regulation of gene activity in diverse processes, including floral induction, could have vast impacts on phenotype. In relation to the winter flowering time of *Begonia socotrana*, miRNAs such as miR156 and miR172 that have been shown to play a key role in the vegetative to reproductive transition may be of interest (Nag and Jack, 2010).

Furthermore, the bait set used to align reads from the Begoniaceae species and horticultural hybrid *Begonia* accessions only contained one copy of FT gene. It was thought that FT was likely to be a *Socotrana* specific allele found in the horticultural hybrid *Begonia*, as the winter flowering time of *Begonia socotrana* was a trait that interested breeders greatly. Using a bait set that contains baits designed on coding sequences of multiple member of the FT gene family, as well as other key members of the developmental pathway that regulates the initiation of the change from vegetative to reproductive growth in angiosperms. Other important genes to look for in this pathway include the members of the *PHYTOCHROME-DEPENDENT LATE-FLOWERING (PHL)* family, *PHYTOCHROME INTERACTING FACTOR (PIF)* family, and the *LEAFY (LFY)* family. *Begonia socotrana* is not only valued for its flowering time by *Begonia* breeders; glabrous leaves and the long lasting flowers are also valuable traits (S. Oostvogels pers comm., 2019). Hence baits based on the sequences of coding sequences of genes involved in the trichome development pathway such as *GLASSY HAIR* and those in the *GLABRA (GL)* family should also be added to a new list of baits. Flower senescence and petal abscission in *Beognia* is ethylene sensitive (Scariot et al., 2014), so genes involved in the ethylene perception pathway, such as the *ETHYLENE RECEPTOR (ETR)* family, should also be included in a further bait set.

5.1.2 Identification of SNPs

The SNPs identified in this study may not truly be specific to *Begonia socotrana*. In this study, the only qualification for SNPs and indels to be identified as *Begonia socotrana* specific was that the SNPs were not present in any of the other seven accessions from one of the six species sampled. Hence the *Begonia socotrana* specific SNPs and indels could be found in one of the many other species of *Begonia*. This is particularly important to consider when thinking of the other species of *Begonia* that contributed to the horticultural varieties of *Begonia* sampled. To produce *Begonia* x hiemalis, *Begonia socotrana* was crossed with plants containing alleles from a range of South American tuberous species including *Begonia boliviensis* and *Begonia pearcei* (Hughes, 2002). SNPs from these South American *Begonia* species that make up parentage of *Begonia* horticultural varieties may be the same as SNPs found in *Begonia socotrana*. However, *Begonia* bait set was designed using coding sequences from *Begonia luzhaiensis,* and Asian species. Therefore, SNPs and indels found within horticultural varieties of *Begonia* may be attributed as coming from *Begonia socotrana* when their genetic origin lies in South American species. Three species of *Begonia* from *South America* were included in this study: *Begonia luxurians, Begonia conchifolia,* and the two accessions of *Begonia heracleifolia*. Hence, any SNPs specific to the broad clade of South American *Begonia* species should have been removed from the list of *Begonia socotrana* specific SNPs and indels. This is because as they would have been removed from the file due to being present in one of *Begonia luxurians, Begonia conchifolia*, and the two accessions of *Begonia heracleifolia.* However, the unknown *Begonia* from Texas possessed approximately 10% of the *Begonia socotrana* specific SNPs and indels. It is possible that the unknown *Begonia* from Texas does share alleles with *Begonia socotrana*, but it is perhaps more likely that these "*Begonia socotrana specific*" SNPs were present by chance and not through a shared genetic background.

5.1.3 Variant calling

The methodology for variant calling used in this study has considerable potential to be improved upon. Genome Analysis Toolkit (GATK) is a pipeline developed by the Broad Institute used to identify SNPs, indels, and other forms of variants (DePristo et al., 2011). GATK is the "industry standard" for identification of SNPs and indels, and has been used extensively across the literature, making it well trusted in the scientific community. The use of GATK for variant calling was beyond the scope of this study due to time and computational constraints. Furthermore, GATK is designed for use on human genetic data, making work on any plant species challenging, even more so for non-model plant species such as *Begonia*. A caveat that is inevitable from using genome skimming to generate the genomic reads is that SNPs, indels and other variants found from the reference sequence will have low depth. This is because of the very low coverage rate of genome skims, up to 0.05%, and means that variants called from genome skim reads are called with less confidence than genomic reads generated by higher coverage sequencing approaches.

5.1.4 Computational constraints

The methodology of this study was constrained by the availability of computational resources. Decisions for methodology were made in order to reduce the computational power required, in order that a wider variety of species could be analysed within the time available for the research.

The extent to which bowtie2 parameters were optimised was a compromise that allowed progression of the project within the allotted time whilst still yielding relatively good resolution. Had more time been available to optimise parameters further, more data points could have been gathered for parameter optimisation, especially around the -2.0 to -2.5 range of the minimum score threshold to increase the resolution seen here. It took between two to six hours to align each set of three million reads to the *Begonia socotrana* genome. As the previous results from bowtie2 for one set of parameters were needed in order to inform the parameters of the next run, parameter optimisation was a lengthy process. Having more data points for parameter optimisation this would enable a better understanding of how the mapping percentage increases with minimum score threshold decrease. This would have been especially informative for the *Begonia conchifolia* reads, as they did not follow a typical curve (figure 11).

It took up to six hours for *Begonia socotrana* reads to be mapped back to the genome under the minimum score threshold of -2.2 decided on after parameter optimisation. When mapping genomic reads from accessions to the *Begonia* baits instead of the *Beognia socotrana* genome the process took around two hours. This led to the decision to map genomic reads from all accessions to the bait set, in order to save time and streamline processing. To increase the scope of this research, access to a server with more disk access, a higher amount of processing power and a better network connection would be needed.

5.2 Future research

5.2.1 Identification of historical *Begonia socotrana* specific alleles

Recent advancements in biotechnology have permitted development of methods allowing relatively high amounts of good quality DNA to be extracted from dried plant material on herbarium specimens (Gutaker et al., 2017). By sequencing accessions *Begonia socotrana* collected in the 19th century from herbarium specimens, much more insight could be gained into the nature of introgression of alleles from *Begonia socotrana* into horticultural varieties used today. Using the same methodology as this study, SNPs and indels specific to the original herbarium specimens could be identified. Comparison of the variants shared between horticultural varieties, the 19th century *Begonia socotrana* specimens, and the modern *Begonia socotrana* specimens would not only enable elucidation of whether *Begonia socotrana* alleles found in the modern day horticultural varieties of *Begonia* are more similar to modern day *Begonia socotrana,* or those from the 19th century. Of particular interest would be the original specimen collected by Balfour in 1880, housed in Kew herbarium. Combined with higher resolution techniques, there is potential to elucidate whether all socotrana hybrid *Begonia* did indeed originate from this first specimen as has been postulated. Sequencing of early *Begonia socotrana* horticultural hybrids would also provide insight into genome evolution within horticultural varieties. The herbarium of the Royal Botanic Garden Edinburgh possesses a specimen dating to 1928. While this is 37 years after the origin of this horticultural hybrid, comparison of the genetics between this 1928 specimen and the F1 hybrid produced in RBGE 90 years later would be interesting.

5.2.2 Population level sampling

This study would have benefitted from having more than one accession of *Begonia socotrana* sequenced. This would have improved the identification of SNPs during the variant calling process. By having sequences from only one specimen of *Begonia socotrana*, it is impossible to tell whether this plant is typical of its species. In order to gather more robust SNPs and indels specific to *Begonia socotrana*, a population level analysis of the species would need to be carried out. As the Republic of Yemen has been embroiled in civil war for over four years at the time of writing, it seems unlikely that fieldwork will be possible for quite some time. Due to these difficulties, a large sample size of herbarium specimens of *Begonia* socotrana, as well accessions from living collections may prove useful in lieu of further fieldwork and collections for the foreseeable future.

Chapter 6 - Conclusions

The results from this study support the hypothesis that a single accession of *Begonia socotrana* was an ancestor to all horticultural *Begonia x Hiemalis* and *Begonia x chiemantha* due to the similar levels of *Begonia socotrana* specific SNPs seen in the horticultural hybrid varieties of *Begonia* sampled. The combination of results from RepeatExplorer2 and GC content analysis suggest that despite the likely single origin of the horticultural varieties of *Begonia*, there is a high level of genetic diversity of *Begonia socotrana* alleles to be found in the horticultural hybrids. However, the regions of high genetic diversity are located in the non-coding regions of the plants' genomes. Prolonged elevation of transposon activity following genome shock on hybridisation is thought to be the reason for this.

Comparisons of the SNPs and repeat content of newly created *Begonia socotrana* F1 hybrid *Begonia* and the historical hybrid *Begonia* accessions analysed would provide further insight into the nature and rate of genome change in *Begonia* hybrids over time. This would assist in elucidation of the causes behind the high level of diversity that this study suggests is present in the non-coding regions of the horticultural hybrid *Begonia* sampled. Aligning reads from specimens to a full genome as opposed to a set of baits designed around coding sequences would provide further insight into the nature of diversity found in the non-coding regions of the genome. Aligning genomic reads to a reference genome should enable the identification of more robust *Begonia socotrana* specific SNPs and indels, which could be called with higher confidence.

This study demonstrates that noncoding regions of the genome can be as interesting as coding regions, and demonstrates the importance of horticultural research. Horticulture not only contributes hugely to the economy, but also has other applications such as *ex situ* conservation, as explored in this work. Furthermore, horticulture can also be seen as the lynchpin to huge amounts of botanical research by enabling the maintenance of living collections.

References

Adeyemo, O., Chavarriaga, P., Tohme, J., Fregene, M., Davis, S. and Setter, T. (2017). Overexpression of Arabidopsis FLOWERING LOCUS T (FT) gene improves floral development in cassava (Manihot esculenta, Crantz). *PLOS ONE*, 12(7), p.e0181460.

Wetterstrand, K. (2019). DNA Sequencing Costs: Data. *National Human Genome Research Institute*. Available at: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data [Accessed July 17, 2019].

Brennan, A.C., Bridgett, S., Ali, M.S., Harrison, N., et al. (2012) Genomic Resources for Evolutionary Studies in the Large, Diverse, Tropical Genus, Begonia. *Tropical Plant Biology*. 5 (4), pp.261–276.

Brocchieri, L. (2014). The GC Content of Bacterial Genomes. *Journal of Phylogenetics & Evolutionary Biology*, 02(01).

Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., et al. (2013) Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. *The American Journal of Human Genetics*. 93 (5), 852–864.

Danecek, P. et al. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), pp.2156–2158.

Dewitte, A., Twyford, A., Thomas, D., Kidner, C. and Van, J. (2011). The Origin of Diversity in Begonia: Genome Dynamism, Population Processes and Phylogenetic Patterns. *The Dynamical Processes of Biodiversity - Case Studies of Evolution and Spatial Distribution.*

Dixon, L., Greenwood, J., Bencivenga, S., Zhang, P., Cockram, J., Mellers, G., Ramm, K., Cavanagh, C., Swain, S. and Boden, S. (2018). TEOSINTE BRANCHED1 Regulates Inflorescence Architecture and Development in Bread Wheat (Triticum aestivum). *The Plant Cell*, 30(3), pp.563-581.

Dodsworth, S. (2015). Genome skimming for phylogenomics. PhD. Queen Mary University of London.

Doebley, J., Stec, A. and Gustus, C. (1995). Teosinte Branched1 and the Origin of Maize: Evidence for Epistasis and the Evolution of Dominance. *Genetics*, 141(1), pp.333-346.

Deutscher Wetterdienst (2019). *Klimatafel von Sokotra*. [online] Dwd.de. Available at: https://www.dwd.de/DWD/klima/beratung/ak/ak_414940_kt.pdf [Accessed 19 Aug. 2019].

Einstein, N. (2009). *Gulf of Aden map*. [online] Commons.wikimedia.org. Available at: https://commons.wikimedia.org/wiki/File:Gulf_of_Aden_map.png [Accessed 19 Aug. 2019].

Endo, M., Tanigawa, Y., Murakami, T., Araki, T. and Nagatani, A. (2013). PHYTOCHROME-DEPENDENT LATE-FLOWERING accelerates flowering through physical interactions with phytochrome B and CONSTANS. *Proceedings of the National Academy of Sciences*, 110(44), pp.18017-18022.

Fornero, C., Suo, B., Zahde, M., Juveland, K., et al. (2017) Papillae formation on trichome cell walls requires the function of the mediator complex subunit Med25. *Plant Molecular Biology*. 95 (4-5), pp.389–398.

Gagliano, S.A., Sengupta, S., Sidore, C., Maschio, A., et al. (2018) Relative impact of indels versus SNPs on complex disease. *Genetic Epidemiology*. 43 (1), pp.112–117.

Gleed, C. J. (1961). Hybrid winter flowering Begonias. J. Roy. Hort. Soc. 80: 319–322.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Rokhsar, D. S. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(1), pp.1178-1186.

Gutaker, R., Reiter, E., Furtwängler, A., Schuenemann, V. and Burbano, H. (2017). Extraction of ultrashort DNA molecules from herbarium specimens. *BioTechniques*, 62(2).

Hamabata, T. et al. (2019). Endangered island endemic plants have vulnerable genomes. *Communications Biology*, 2(1).

Hooker, J. D. (1881). New Garden Plants. *The Gardeners' Chronicle*, 15, pp.8

Hughes, M. (2001). The Begonia of the Socotra Archipelago. *American Begonia Society*, 68, pp.109–213.

Hughes, M. (2019). Twitter. [online] Twitter.com. Available at: https://twitter.com/markhughesis/status/1088763536141139970 [Accessed 19 Aug. 2019].

Hughes, M., Hollingsworth, P. & Miller, A. (2003). Population genetic structure in the endemic Begonia of the Socotra archipelago. *Biological Conservation*, 113(2), pp.277–284.

Hughes, M. & Miller, A.G. (2002). A New Endemic Species of Begonia (Begoniaceae) From The Socotra Archipelago. *Edinburgh Journal of Botany*, 59(2), pp.273–281.

Huntingford, G.W.B. (1980). *Periplus of the Erythraean Sea*, Hakluyt Society.

Hvoslef-Eide, T. & Munster, C. (2006). Begonia: History and breeding. In *Flower breeding and genetics: issues, challenges and opportunities for the 21st century*. Dordrecht: Springer, pp.241–275.

Kidwell, M. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1), pp.49-63.

Kroon, G.H. (1993). Breeding Research in Begonia. *Acta Horticulturae*, (337), pp.53–58.

Ladizinsky, G. (2012). *Plant evolution under domestication*. Springer.

Laetsch, D.R. & Blaxter, M.L. (2017) BlobTools: Interrogation of genome assemblies. *F1000Research*, pp.612-687.

Langmead, B. & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357–359.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), pp.2987–2993.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25 (14), pp.1754–1760.

Li, J. and Zhang, Y. (2014). Relationship between promoter sequence and its strength in gene expression. *The European Physical Journal E*, 37(9), pp.8-12.

Lynch, M., Conery, J. & Burger, R. (1995). Mutational Meltdowns in Sexual Populations. *Evolution*, 49(6), pp.1067.

Marasek-Ciolakowska, A. et al. (2009). Genome composition of 'Elatior'-begonias hybrids analyzed by genomic in situ hybridisation. *Euphytica*, 171(2), pp.273–282.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), pp.10.

Martinez Martinez, A. (2017). A draft genome assembly for Hillebrandia Sandwichensis. MSc. University of Edinburgh.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science,* 226, pp.792–801.

Miller, A. (2004a). Begonia samhaensis. *IUCN Red List of Threatened Species*.

Miller, A. (2004b). Begonia socotrana. *IUCN Red List of Threatened Species*.

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), pp.267–323.

Nag, A., & Jack, T. (2010). Sculpting the Flower; the Role of microRNAs in Flower Development. *Current Topics in Developmental Biology Plant Development*, pp.349–378.

Novak, P. et al. (2013). RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*, 29(6), pp.792–793.

Oxford Economics (2018). *The Economic Impact of Ornamental Horticulture and Landscaping in the UK*. Oxford.

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., Ainouche, M., Chalhoub, B. and Grandbastien, M. (2009). Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytologist*, 186(1), pp.37-45.

Rival, P., Press, M., Bale, J., Grancharova, T., Undurraga, S. and Queitsch, C. (2014). The ConservedPFT1Tandem Repeat Is Crucial for Proper Flowering in Arabidopsis thaliana. *Genetics*, 198(2), pp.747-754.

Scariot, V., Paradiso, R., Rogers, H., & Pascale, S. D. (2014). Ethylene control in cut flowers: Classical and innovative approaches. *Postharvest Biology and Technology*, *97,* pp.83–92.

Scepi, E. (2005). Begonia socotrana. [online] Flickr. Available at: https://www.flickr.com/photos/edoardoscepi/3258157476/in/photostream/ [Accessed 19 Aug. 2019].

Schlegel and Fottler co. 1899. Annual catalogue, bulbs, plants, etc., Boston MA.

Šmarda, P., Bureš, P., Horová, L., Leitch, I., Mucina, L., Pacini, E., Tichý, L., Grulich, V. and Rotreklová, O. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences*, 111(39), pp.E4096-E4102.

Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., et al. (2012) Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*.(2), pp.349–364.

Tebbitt, M.C. (2005) *Begonias: cultivation, identification, and natural history*. Timber Press.

Veselý, P., Bureš, P., Šmarda, P. and Pavlíček, T. (2011). Genome size and DNA base composition of geophytes: the mirror of phenology and ecology?. *Annals of Botany*, 109(1), pp.65-75.

Wickland, D. and Hanzawa, Y. (2015). The FLOWERING LOCUS T/TERMINAL FLOWER 1 Gene Family: Functional Evolution and Molecular Mechanisms. *Molecular Plant*, 8(7), pp.983-997.

Vitte, C., & Panaud, O. (2005). LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic and Genome Research*, *110*(1-4), pp.91–107.

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*(Web Server).

Zhang, M., Liu, X., Fan, W., Yan, D., Zhong, N., Gao, J. and Zhang, W. (2018). Transcriptome analysis reveals hybridization-induced genome shock in an interspecific F1 hybrid from Camellia. *Genome*, 61(7), pp.477-485.

Appendix A - List of commands used for this study

For a similar study on comparable data this set of commands may be used as a pipeline.

Read trimming and quality control

Check read quality

```
FastQC --nogroup example.fq
```

Trim reads

```
Cutadapt –q30 –a YOUR_ADAPTER_SEQUENCE –A YOUR_ADAPTER_SEQUENCE –trim-n –o
example_trimmed.fq example.fq
```

Aligning reads to a reference sequence to produce a variant call file

Create index from reference sequence

```
bowtie2–build index_sequence.fna index_sequence
```

Align reads to reference sequence

```
bowtie2 --no-unal --score-min L,-2.2,-2.2 -x Baits -1 forward_reads.fq -2
reverse_reads.fq -S example.sam 2> example_bowtie_output
```

```
Sort and index sam file and reference
```

```
samtools view -bS example.sam -t index_sequence.fna > example.bam
```

```
samtools sort example.bam –o example_sorted.bam
```

```
samtools index example_sorted.bam
```

```
samtools faidx index.fna
```

Create VCF file

```
samtools mpileup –u -f index.fna example_sorted.bam > example.pileup
```

```
bcftools call -mA example.pileup > example.vcf
```

Create summary of variation within VCF file

```
bcftools stats example.vcf > example_summary.txt
```

VCF-ISEC

Generate vcf.gz file and the index file vcf.gz.tbi

```
bgzip -c file.vcf > file.vcf.gz
```

```
tabix -p vcf file.vcf.gz
```

Modify path to locate perl if necessary

```
export PERL5LIB=/path/to/your/installation/perl
```

Determine variants shared by two VCF files

```
vcf-isec –f -n +2 A.vcf.gz B.vcf.gz | bgzip -c > AB_shared.vcf.gz
```

Determine variants shared by three VCF files

```
vcf-isec –f -n +3 A.vcf.gz B.vcf.gz C.vcf.gz | bgzip -c > AB_shared.vcf.gz
```

Determine variants found in only one of multiple VCF files

```
vcf-isec –f -c A.vcf.gz B.vcf.gz C.vcf.gz D.vcf.gz E.vcf.gz | bgzip -c >
A_unique.vcf.gz
```

Seqtk

Subsample 300 000 read pairs from two paired FASTQ files

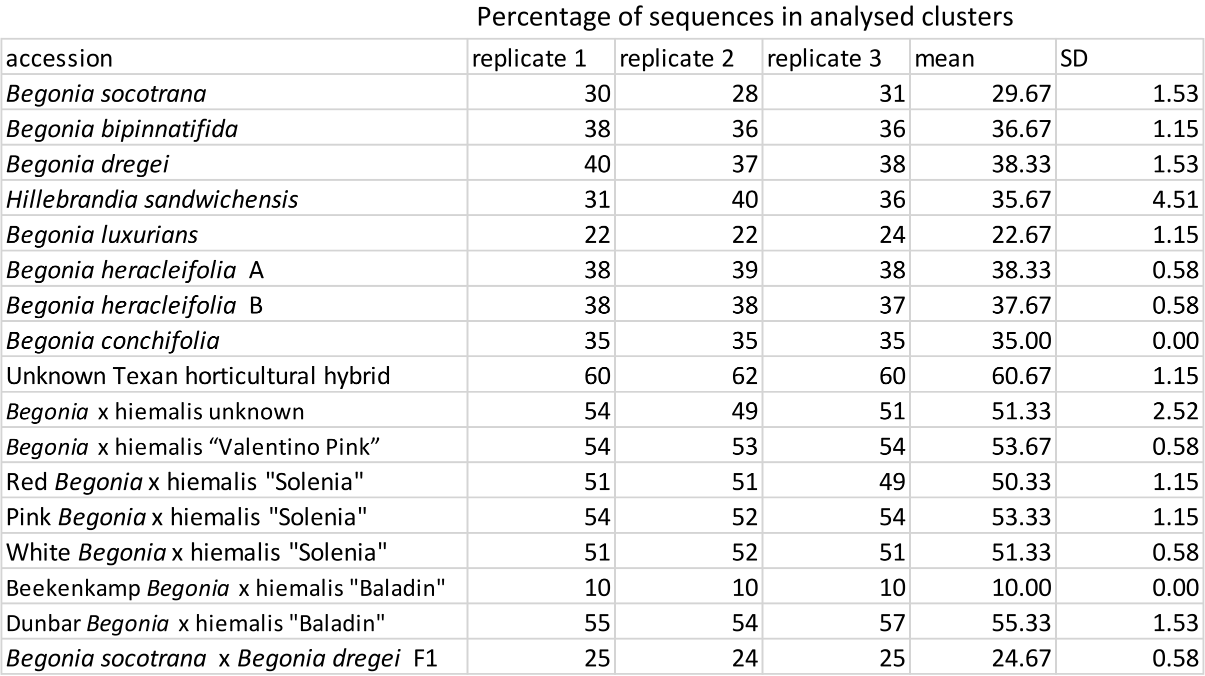
```
seqtk sample -s100 example_1.fq 300000 > subsample_example_1.fq
```

```
seqtk sample -s100 example_2.fq 300000 > subample_example_2.fq
```

Convert FASTQ to FASTA format

```
seqtk seq -a example_sequence.fq.gz > example_sequence.fa
```

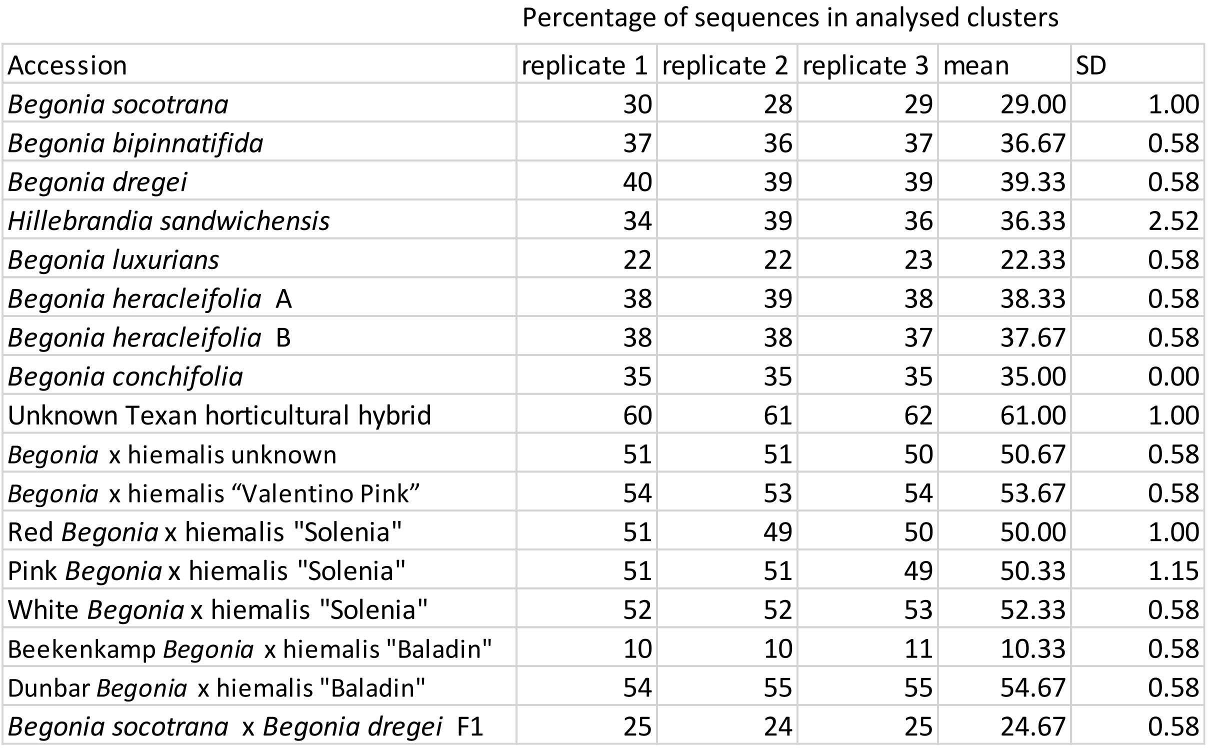
Appendix B – Raw data from RepeatExplorer2



Percentage of sequences in analysed clusters

accession	replicate 1	replicate 2	replicate 3	mean	SD
Begonia socotrana	30	28	31	29.67	1.53
Begonia bipinnatifida	38	36	36	36.67	1.15
Begonia dregei	40	37	38	38.33	1.53
Hillebrandia sandwichensis	31	40	36	35.67	4.51
Begonia luxurians	22	22	24	22.67	1.15
Begonia heracleifolia A	38	39	38	38.33	0.58
Begonia heracleifolia B	38	38	37	37.67	0.58
Begonia conchifolia	35	35	35	35.00	0.00
Unknown Texan horticultural hybrid	60	62	60	60.67	1.15
Begonia x hiemalis unknown	54	49	51	51.33	2.52
Begonia x hiemalis “Valentino Pink”	54	53	54	53.67	0.58
Red *Begonia* x hiemalis "Solenia"	51	51	49	50.33	1.15
Pink *Begonia* x hiemalis "Solenia"	54	52	54	53.33	1.15
White *Begonia* x hiemalis "Solenia"	51	52	51	51.33	0.58
Beekenkamp *Begonia* x hiemalis "Baladin"	10	10	10	10.00	0.00
Dunbar *Begonia* x hiemalis "Baladin"	55	54	57	55.33	1.53
Begonia socotrana x *Begonia dregei* F1	25	24	25	24.67	0.58

Appendix C – Raw data from TAREAN



	Percentage of sequences in analysed clusters				
Accession	replicate 1	replicate 2	replicate 3	mean	SD
Begonia socotrana	30	28	29	29.00	1.00
Begonia bipinnatifida	37	36	37	36.67	0.58
Begonia dregei	40	39	39	39.33	0.58
Hillebrandia sandwichensis	34	39	36	36.33	2.52
Begonia luxurians	22	22	23	22.33	0.58
Begonia heracleifolia A	38	39	38	38.33	0.58
Begonia heracleifolia B	38	38	37	37.67	0.58
Begonia conchifolia	35	35	35	35.00	0.00
Unknown Texan horticultural hybrid	60	61	62	61.00	1.00
Begonia x hiemalis unknown	51	51	50	50.67	0.58
Begonia x hiemalis “Valentino Pink”	54	53	54	53.67	0.58
Red *Begonia* x hiemalis "Solenia"	51	49	50	50.00	1.00
Pink *Begonia* x hiemalis "Solenia"	51	51	49	50.33	1.15
White *Begonia* x hiemalis "Solenia"	52	52	53	52.33	0.58
Beekenkamp *Begonia* x hiemalis "Baladin"	10	10	11	10.33	0.58
Dunbar *Begonia* x hiemalis "Baladin"	54	55	55	54.67	0.58
Begonia socotrana x *Begonia dregei* F1	25	24	25	24.67	0.58