

THE UNIVERSITY
of EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

A study on the evolution of genome dynamics across the mega-diverse genus *Begonia* L. (Begoniaceae)

Lucía Campos-Domínguez

Doctor of Philosophy



School of Biological Sciences

Institute of Molecular Plant Sciences

Royal Botanic Garden Edinburgh

The University of Edinburgh

October 2020

A study on the evolution of genome dynamics across the mega-diverse genus *Begonia* L. (Begoniaceae)

Lucía Campos-Domínguez

The University of Edinburgh

School of Biological Sciences

Institute of Molecular Plant Sciences

Royal Botanic Garden Edinburgh

Supervisors

Catherine A. Kidner[1,2], Alexander D. Twyford[2]

[1]Royal Botanic Garden Edinburgh

[2]The University of Edinburgh

”-Por qué anda usted en zigzag, señor Nge?

-Porque así se tarda más en hacer el recorrido

y se piensa mejor adónde va uno, hijo.”

José Luis Cuerda

Amanece, que no es poco (1989)

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Lucía Campos Domínguez

31st October 2020

Abstract

Understanding the drivers of angiosperm species diversity has always been one of the major goals of research in evolutionary biology. In this thesis, the genus *Begonia* L. (Begoniaceae) is used as a model to study some of these drivers. *Begonia* is one of the most species-rich angiosperm genera with 1991 pantropically distributed species currently identified. The current distribution of *Begonia* species abundance overlaps with angiosperm species richness, dominating the tropical areas and making this genus a good model for the study of large angiosperm radiations. Recent evolutionary work on *Begonia* provides a robust phylogenetic background for the analysis of evolutionary patterns across the group and confirms speciation in *Begonia* is not only driven by adaptive processes, but rather by a "random walk". As well as extensive phenotypic variation, previous research has established that *Begonia* is also very variable in genome size chromosome numbers. This suggests that large scale genomic re-arrangements and bursts of transposon activity may be common in *Begonia* evolution. This thesis investigates patterns of genome dynamics across the *Begonia* genus, and its potential role in rapid species diversification.

To have a better overview of the genome structural variation present in *Begonia*, chromosome numbers for hundreds of species have been gathered and taxonomically confirmed according to the latest phylogeny. Results show chromosome number diversity is higher in some but not all of the larger clades, and the most variable karyotypes belong to taxonomically complex or unresolved *Begonia* sections. Our results also indicate that genome size variation does not correlate with changes in chromosome number. This suggests that *Begonia* genome dynamics is not only caused by large-scale duplications, rearrangements or changes in ploidy levels, but rather that changes in the repetitive fraction of the genome also play an important role, likely causing changes in chromosome size.

Whole Genome Sequencing (WGS) has been used to identify TE content and variation among different *Begonia* clades. In addition, the genome of the only *Begonia* sister species *Hillebrandia sandwicensis* has been assembled to compare these two genera at a genomic level and better understand the specific factors involved in the repeated

radiation events described in *Begonia*. Our results suggest that *Begonia* species show more complex, repetitive and dynamic genomes overall than *Hillebrandia*. There is a wide variation in both content and types of DNA repeats across the different *Begonia* clades. Our Long Terminal Repeat (LTR) dating results show that species from older and more depauperate clades show more stable and generally lower recent LTR activity. In contrast, we have identified LTR-*Copia* and *Gypsy* retrotransposon families being recently expanded in two different highly speciose South-East Asian sections, and LTR elements and satellite DNA families being expanded in three Neotropical radiations. One of the largest Neotropical *Begonia* radiations is represented by section *Gireoudia*. To further investigate the genomic factors involved in this radiation, the repeat content, variation and transposon insertion times of 32 *Gireoudia* species was studied. These results show recent LTR and wide satellite DNA expansions across *Gireoudia* species, in contrast with its species poor, closely related sections. Moreover, we found a common LTR-*Gypsy* expansion shared by all *Gireoudia* species studied, that dates to the origin of this section, suggesting TE dynamics may have played a role in the origin of the *Gireoudia* radiation.

To additionally comprehend whether intra-specific genome structure variation drives speciation by introducing crossing barriers between populations, repeatomes from different *Begonia heracleifolia* (*B.* section *Gireoudia*) accessions showing variable genome sizes have also been studied. This revealed a markedly higher satellite repeat and LTR-*Gypsy* content in the accessions with larger genome sizes, associated with the expansion of a few abundant repeat lineages. Furthermore, experimental crosses between the accessions demonstrated that seed viability is impaired in crosses between individuals with different genome sizes. Therefore, we have also identified an association between changes in genomic repeat content and genetic barriers to reproduction within a *Begonia* species.

Lay summary

Flowering plants are great example of unusually rich diversity emerging in a relatively brief period of time. Understanding the process of species diversification is still a major challenge in the field of evolutionary biology. Any process of adaptation and speciation is determined by changes in their DNA. These can be triggered by spontaneous events or induced by specific factors. This thesis investigates the most variable and dynamic part of the genome, which is called the repeatome. The repeatome is composed by a mixture of repetitive DNA elements comprising transposons, simple sequence repeats and satellite DNA. Transposons (also called Transposable Elements or TEs) are DNA fragments that can integrate into new positions in the genome using mechanisms of "cut and paste" and "copy and paste". Simple sequence repeats are very short regions of repeated DNA motifs of a few basepairs, and satellite DNA consists of large regions of these repeated motifs. The potential dynamism of all these elements can rapidly trigger genomic changes, that eventually could have an effect in the evolutionary trajectory of a species.

In this study, I investigate the potential role of the repeatome in the species diversification in one of the largest genera of flowering plants which has almost 2000 species, *Begonia*. Chromosome numbers for hundreds of *Begonia* species were studied to understand how their genome structure has changed during the evolution of the genus. Results show that chromosome number is more variable in *Begonia* clades with large numbers of species, and generally stable within *Begonia* the smaller category of sections. I also investigated variation in the genome size (amount of DNA in the genome), and the results indicate genome size does not correlate with chromosome numbers, which suggests that genome dynamics in *Begonia* involves changes in chromosome size as well as numbers, indicating that repeatome changes also play an important role.

To explore repeatome variation across this genus, I have sequenced and analysed the genome of 12 *Begonia* species from its main clades, which have different levels of species richness. The only species from the *Begonia* sister genus (*Hillebrandia*) was also selected as an outgroup for this study. Results suggest that *Begonia* species overall show more complex, repetitive and dynamic genomes than *Hillebrandia*. I also found a wide

variation in both content and types of DNA repeats across the different *Begonia* clades. Clades with more species diversity show, in general, more repeatome diversity and recent expansions of transposons, when compared to clades with less species. To have a closer look into the repeatome diversity across more closely related species, I studied the genome of other 32 *Begonia* species from a large Neotropical section, and compared them to 2 others from their sister, species-poor sections. Results from this approach further confirm that *Begonia* species repeatomes are rapidly changing, and transposons and satellite DNA repeats are highly variable and recently expanded even among very recently diverged species. Moreover, I found a specific transposon expansion for this very large section that dates to the origin of this group of species, further supporting that changes in the repeatome had a role in shaping *Begonia* species diversification.

To better understand how these changes in the repeatome can introduce crossing barriers between populations and trigger species diversification, I have studied repeatome differences among populations from the same *Begonia* species (*B. heracleifolia*). These populations show variable genome sizes, and our results revealed that these differences were caused by satellite repeat and transposon expansions. We also did experimental crosses between the populations and demonstrated that crosses between individuals that show genome size differences yield fewer viable seeds, suggesting a link between changes in genomic repeat content and reproductive barriers within a *Begonia* species.

Acknowledgements

I am most grateful to the people that have taught me so much during these four years. First of all to Catherine because I have absolutely no doubt I got the best PhD supervisor I could have asked for. Thank you for your kindness, your patience and for all the valuable lessons. Thanks for being always available and finding the time for quick meetings to solve any struggles, helping me deal with my constant insecurities and giving me all the support I needed to achieve this PhD. Thank you for all the quick feedback, your kind and supporting words, and having the ability to transform my "end of the world" feelings to "it is actually ok" ones whenever something doesn't go as expected. I look forward to continuing to learn and collaborate with such brilliant scientist and person. To Alex, huge thanks for always being there, giving me so much support and helping me gain confidence in my work. Talking to you always gives me an extra kick of motivation and new perspectives. I am extremely excited and look forward to your mentoring on new genomic adventures. I am also extremely grateful for my two Kew supervisors Jaume and Ilia, who have provided constant support and much-appreciated feedback. Thanks Jaume for your help during my visit to Kew.

My PhD was funded by the Darwin Trust of Edinburgh, to whom I have to thank for the immense learning opportuning they have provided me with. A special thanks to Professor David Finnegan, who has always been excited to hear about my project.

Although not officially my supervisor, many thanks to one of the people that has represented a massive help in this project and from whom I have learnt the most, Aure Bombarely. Thanks for your constant bioinformatics support, and for the great help and company during my days in Milan. Thanks for the great discussions during those Italian heat wave days that despite the melting brains hugely inspired me and motivated me. It was a nice surprise to learn that a "paisano" was one of our collaborators at the beginning of my PhD, and you have become one of the greatest scientists and people I have met during these last four years. I look forward to more encounters and new collaborations in the future.

Also thanks to the other members of my thesis committee Andrew Hudson and Mark

Hughes who have given me such great support and feedback along the way. I am really grateful for the great discussions and advice you have given me.

One of the best experiences as a PhD student has been being able to help supervising and mentoring other students' projects. I have been lucky to meet such amazing people, from whom I have learnt probably more than I have been able to teach them. Thank you to Andrea and Thea, you were brilliant students and became great friends and I am extremely grateful for your hard work. Thank you also to Antoine and PJ who were excellent students and great office buddies.

The completion of this thesis would not have been possible without the exceptional help and generosity of our collaborators in Taipei. Thank you Dr. Daisy Hu and Prof. Kuo-Fang Chung, for the data sharing and all the ongoing collaborations. I look forward to keeping working with you in the future. Special thanks to Yu-Hsin Tseng for being such great friend and colleague during all these years. Thanks for all the great help and *Begonia* wisdom, and also for showing us around Taipei during our lovely visit.

Many thanks to all the RBGE people that have participated in our *Begonia* lab and NGS helpgroup meetings for the great discussions from which I have learnt so much. Special thanks to Peter Moonlight, Mark Hughes and Hannah Wilson for all the *Begonia* taxonomy and evolution knowledge you have shared with me. Also thanks to all the RBGE horticulture staff that make our research possible by keeping all our plants alive. Special thanks David Tricker, Gunnar Ovstebo, Neil Watherston and Sadie Barber for your massive help growing and propagating our favourite begonias. To the RBGE STS personnel Ruth Hollands, Frieda Christie, Michelle Hart and Laura Forrest, thanks for your kindness and help setting things up in the lab. Extra thanks to the people that made my sample shipment possible in the middle of a pandemic lockdown: Sadie Barber and Louise Galloway for collecting the samples and Ruth Hollands for preparing and sending them.

I have met many great scientists during these last four years that I have to thank for somehow contributing to my research. Thanks to Dr. Lingfei Li for your genome data and your discussions and interest on *Begonia* transposons. Many thanks to Dustin

Wolkis for your kindness and help with sending the Hillebrandia seeds, and to Prof. Katharina Pawlowski for kindly sending us *Begonia foliosa* material. I am also extremely grateful to Dr. Andrew Romanowski for your great help and making my first and only try at a MinION run so successful.

I could not be more grateful for the incredible human beings (and scientists) I have gotten to meet during these years. RBGE has given me the best friends, and each one of them has helped me grow as a scientist and made me a better person. Thanks Surabhi for timidly introducing yourself in the RBGE canteen during one of my first days and becoming the first friend I made in Edinburgh (we have come a long way!). Thanks to Pakkapol and Yun-Yu, for being my first office buddies when everyone else was in the field, for always offering your bioinformatics and botanical knowledge and for your valued friendship. Thanks for starting the "cooking club" (or basically everytime Pakkapol ended up cooking for twelve) that gave us such great (and delicious) times. Thank you Flavinha for always making the office so alive, for your friendship and wisdom, and for even giving me a bed when I was homeless. Thanks for letting me bother you with bioinformatics and LaTeX questions all the time, I really owe you ten thousand pints (If we are ever in the same place again!). Thanks to every office mate for the laughs and coffee (and the pub evenings too!): Julieth, Andrés, Mauro, Andy, Thibauld, Jess, Hazel, Isuru, Gail, Nadia. Thanks to Natalia, Gustavo, Ozan and Édeline for being such good friends and always so fun to be around. To the best begonia gals, Hann, Madh, Cynth: thanks for always offering your support and not falling asleep during my long lab meeting presentations on genome assembly. Cynth, special thanks for always letting me vent about anything and being such a great listener and amazing friend. Hann, Madh, huge thanks for being my Stockbridge family and always being there. Special thanks to Kristine, the lichen master, best flatmate and lockdown companion. I am super grateful for all the time we spend together, all I have learnt from you and all I have done thanks to your pushing motherly moods. I really love you all. These years would have been way too boring without you.

Doing a PhD (with the addition of a world pandemic in the last year!) is an emotional rollercoaster which can be impossible to face without the support of friends and family.

I am so happy I can say I now have a family in Edinburgh of friends that I love and who have become essential in my life. Apart from all the above mentioned RBGE friends, thanks to everyone else I have met (or re-met) these years and have become the best company. Thanks Ana for being the light at the beginning of this journey, for being my bit of the south when I most needed it. To Aisha, for being the best yoga, travel and coffee buddy, and an amazing friend overall. To Maria (and Jon and Bella), because I wasn't expecting to feel so at home with someone I had not seen in 13 years. Thanks to anyone that I have met during this journey and has made it somehow a bit better. Also huge thanks to Sito, because I blame you for the start of this adventure.

Mil gracias también a todo el mundo que me ha regalado su cariño y apoyo desde el sur. Eterna gratitud hacia Antonio Flores, mi primer mentor y una de las mejores personas y científicos con los que me he cruzado en mi camino, gracias a quien me entró el "gusanillo" por la evolución vegetal. Muchísimas gracias por tus años de amistad y apoyo profesional. Gracias también a mis amigos de siempre. A todo el que me ha mandado ánimos y cariño desde la distancia cuando lo necesitaba. Mil gracias a todos los que me habéis dicho "puedes hacerlo" cuando necesitaba oirlo. A los que me habéis visitado al otro lado del muro. Os quiero a rabiar Menchu, Ale, Meri, Marina, Patri, Mar, Ro, Juanki.

A Antón, porque aunque pienses que las plantas no son seres vivos, creo que al final hasta he conseguido que las begonias y sus transposones te interesen un poquito (o quizás es puro peloteo). Gracias por hacerlo todo fácil. Por hacerme reir siempre, por el apoyo infinito y los incontables abrazos. Gracias por acogerme en tu guarida y regalarme un pedacito de mar al que mirar mientras escribo. Ah, y por las diez mil tablas formateadas. Gracias también a las demás personas que han hecho de "la Galisia" un lugar acogedor para esta sureña estos meses, especialmente a Ale y a Joaquín.

Por último, gracias a mi familia. Por ser el eje fundamental de cariño y apoyo en esta y todas las etapas. Por preguntar e interesaros por mi trabajo constantemente, aún sabiendo que probablemente obtendríais una retahíla de quejas como respuesta. Por los ánimos continuados y por haber sabido mantenerme y mantenernos fuertes pese a distancias, pandemias y otros palos. Os quiero.

Table of Contents

Declaration viii

Abstract xi

Lay summary xiii

Acknowledgements xvii

List of figures xxiii

List of tables xxv

1 General introduction **1**

1.1 The questions around angiosperm diversification 3

1.2 Dynamic genomes: the role and effects of repetitive DNA in angiosperm evolution 6

1.3 Evolution and speciation in the mega-diverse genus *Begonia* L. 9

1.4 Objectives 12

2 Evolutionary patterns of genome size and chromosome number variation across the Begoniaceae family **23**

2.1 Introduction 27

2.2 Results and discussion 31

2.2.1 Chromosome variation across the *Begonia* phylogeny 31

2.2.2 Genome size variation across *Begonia* 41

2.3 Material and Methods 45

2.3.1 Data collection, curation and analysis 45

2.3.2 Genome size estimations 46

2.A Appendix 53

2.A.1 Chromosome numbers of *Begonia* 53

2.A.2 Genome sizes of *Begonia* 66

3 Generating genomic resources to study *Begonia* evolution: de novo genome sequencing, assembly and gene annotation of seven Begoniaceae species 69

3.1 Introduction 73

3.2 Results and discussion 78

3.2.1 Genome sequencing of seven Begoniaceae species 78

3.2.2 K-mer analysis and genome characteristics of seven Begoniaceae species 79

3.2.3 Draft genome assemblies of seven Begoniaceae species 81

3.2.4 Protein-coding gene sets of seven Begoniaceae species 84

3.3 Materials and methods 85

3.3.1 Plant material 85

3.3.2 DNA extraction, library preparation and genome sequencing . . 86

3.3.3 Genomic data QC and pre-processing 87

3.3.4 Assembly of the *Begonia conchifolia* genome using long-read data 88

3.3.5 Assembly of six Begoniaceae genomes using short-read data . . . 88

3.3.6 Contaminant removal 89

3.3.7 Gene annotation of seven Begoniaceae species 90

3.A Appendix 95

3.A.1 Assembly and annotation stats of other Begoniaceae genomes . . 95

3.A.2 Contaminant removal of the *B. johnstonii* genomic dataset . . . 96

3.A.3 *B. conchifolia* sequencing: long-read datasets statistics 98

3.A.4 Assembly pipeline of the assembly of *Begonia conchifolia* using long and short reads 98

3.A.5 Assembly pipeline of the assembly of 6 Begoniaceae genomes using short reads 99

3.A.6 Assembly metrics of all the different assemblies performed 100

3.A.7 Contamimant removal from final assemblies 104

3.1.8 Smudgeplot of *B. bipinnatifida* 114

4 An in-depth analysis of the DNA repeat content and dynamics across Begoniaceae 115

4.1 Introduction . . . 117

4.2 Results and Discussion . . . 120

4.2.1 Repeat annotations . . . 120

4.2.2 Temporal dynamics of *Begonia* LTRs . . . 135

4.3 Materials and Methods . . . 143

4.3.1 Repeat and LTR annotation of the 12 Begoniaceae genome assemblies . . . 143

4.3.2 Cluster-based repeat annotation of seven Begoniaceae species . . 143

4.3.3 Insertion date estimation of assembled and annotated LTRs of the 12 Begoniaceae genomes . . . 144

4.3.4 LTR temporal dynamics estimation using cluster-based LTR annotation . . . 144

4.A Appendix . . . 152

4.A.1 Raw numeric data of repeat proportions shown in figures . . . 152

4.A.2 LTR dynamics on long vs. short read assembly of the same species (*B. conchifolia*) . . . 154

5 A closer look: genome dynamics in the highly specious *Begonia* section *Gireoudia* 155

5.1 Introduction . . . 159

5.2 Results and Discussion . . . 162

5.2.1 Genomic repeat landscapes of 34 closely related Neotropical *Begonia* . . . 162

5.2.2 LTR insertion dynamics of 34 Neotropical *Begonia* . . . 171

5.3 Materials and Methods . . . 175

5.3.1 Plant material . . . 175

5.3.2 Data acquisition . . . 176

5.3.3 Data pre-processing and analysis . . . 176

5.A Appendix . . . 181

5.A.1 Raw numeric data of repeat proportions shown in figures . . . 181

6 An even closer look: intra-specific genome dynamics and population divergence in *Begonia heracleifolia* (*B.* sect. *Gireoudia*) 183

6.1 Introduction . . . 187

6.2 Results and discussion . . . 190

6.2.1 Genome size differences of 10 *Begonia heracleifolia* populations . 190

6.2.2 WGS and repeat content analysis of 10 *Begonia heracleifolia* populations . . . 191

6.2.3 Assessment of reproductive isolation in intraspecific crosses . . . 198

6.3 Materials and Methods . . . 201

6.3.1 Plant Material . . . 201

6.3.2 Genome size estimations by flow cytometry . . . 202

6.3.3 DNA extraction, library preparation and sequencing . . . 202

6.3.4 Analysis of genomic repeats . . . 203

6.3.5 Intra and inter-population crosses of *Begonia heracleifolia* . . . 203

6.3.6 Seed viability tests . . . 203

6.A Appendix . . . 208

6.A.1 RepeatExplorer2 summary data table . . . 208

6.A.2 Table of crosses . . . 208

6.A.3 Germination rate distributions . . . 208

6.A.4 Seed count and germination rate data . . . 209

7 General conclusions 211

7.1 General conclusions . . . 213

7.2 Whole Genome Sequencing projects for non-model species and the challenges of analysing repetitive DNA . . . 214

7.3 Chromosome evolution in *Begonia*: karyotype and genome size instability at different taxonomic levels . . . 216

7.4 *Begonia* genomes are more dynamic in large radiations . . . 217

7.5 Potential evolutionary consequences of LTR and satDNA expansions in *Begonia* . . . 222

7.6 Future approaches . . . 223

List of Figures

1.1 Biological factors involved in angiosperm speciation. 5

1.2 Schematic representation of the main genomic elements within a chromosome 7

1.3 Potential roles of TEs in speciation 9

1.4 *Begonia* sectional phylogeny 11

2.1 *Begonia* chromosome number data sampling within sections and clades . 32

2.2 Chromosome number distribution in the different *Begonia* clades 35

2.3 Chromosome number distribution in the different *Begonia* sections . . . 38

2.4 Relationship between chromosome number variation and number of species in the sections of *Begonia* 40

2.5 C-value distribution in the different *Begonia* clades 43

2.6 Genome size variation *versus* chromosome number in *Begonia* 45

3.1 *Begonia* simplified phylogeny with the clade position of all species used in whole genome sequencing. 76

3.2 K-mer spectra (K=31) of the 7 genomes sequenced 80

4.1 Schematic structure of LTR-*Copia* and LTR-*Gypsy* superfamilies of retrotransposons 118

4.2 Repeatome annotation results of the 12 Begoniaceae genome assemblies using RepeatMasker 121

4.3 Comparison of the amount of LTR elements found by two different approaches in 12 Begoniaceae genomes 123

4.4 Comparison of genomic proportions taken by LTR superfamilies in the 12 Begoniaceae genomes 126
4.5 Comparison of the number of full length LTR elements found in each of the 12 Begoniaceae genome assemblies 127
4.6 Repeat landscape from read-clustering analyses in 8 Begoniaceae species 129
4.7 Relative amounts of species data in each repetitive DNA cluster across a Begoniaceae comparative analysis 132
4.8 Repeat annotation and quantification method comparison 134
4.9 Insertion times of all full-length LTR-*Copia* elements found in each genome assembly 137
4.10 Insertion times of all full-length LTR-*Gypsy* elements found in each genome assembly 138
4.11 Insertion times of all full-length Unclassified LTRs found in each genome assembly 139
4.11 Retrotranspositional dynamics of different LTR clusters from 8 Begoniaceae species 142

5.1 *B.* sect. *Gireoudia* phylogeny indicating the species sequenced for repeatome analysis 161
5.2 Repeatome annotation of 34 Neotropical *Begonia* 164
5.3 Graphical representation of the different repeat families expanded between *Gireoudia* species and outgroup sections. 167
5.4 Type of repeats most differentially expanded within the studied groups of species 170
5.5 LTR temporal dynamics in Neotropical *Begonia* 172

6.1 Geographical location of the 7 *B. heracleifolia* populations studied . . . 189
6.2 Haploid genome size differences found between the *B. heracleifolia* populations 190
6.3 GC content distribution from raw datasets of all *B. heracleifolia* populations sequenced 192
6.4 Repeatome annotation results from 10 *B. heracleifolia* accessions 193

6.5 Graphic representation of two SatDNA clusters differentially expanded between two *B. heracleifolia* populations 195
6.6 Graphic representation of two SatDNA clusters differentially expanded within the same *B. heracleifolia* population 196
6.7 Proportion of viable seeds of *B. heracleifolia* inter and intra-population crosses 199

List of Tables

3.1 Genome size and chromosome number information from all Begoniaceae species with sequenced genomes 77

3.2 Sequencing output statistics of the 7 species sequenced 78

3.3 Final genome assembly metrics for the seven Begoniaceae species sequenced 81

3.4 BUSCO results for the seven Begoniaceae genome assemblies 83

3.5 Gene annotation results of the 7 genomes assembled 84

3.6 Plant material used for whole genome sequencing 86

5.1 Species ID, collection and accession ID from all specimens sequenced in this chapter. 175

6.1 List of differentially expanded repeat clusters between two *B. heracleifolia* populations 194

6.2 List of differentially expanded repeat clusters within the same *B. heracleifolia* population 196

6.3 Population number, RBGE accession id (last three digits used as accession number throughout this chapter), and collection id of all the *B. heraclaeifolia* specimens studied in this chapter. 201

Chapter 1

General introduction

1.1 The questions around angiosperm diversification

The vast diversity of living organisms on this planet has always fascinated humans and stimulated the curiosity of scientists. The processes of elucidating and understanding the drivers of this diversity have laid some of the foundations of modern biology, especially in the fields of genetics and evolution. One of the distinctive features of biological diversity is its unevenness (Donoghue and Sanderson, 2015): some groups of organisms are highly diverse and others have only a few representative species. Among the living groups with higher species diversity, angiosperms have always been singled out as a great example of unusually rich taxonomic diversity emerging in a relatively brief period of time (Heywood, 1993). Within flowering plants, this species richness is also unevenly distributed: a few angiosperm families include thousands of described species, others count only with one or two. However, it could be discussed that this unevenness is partially due to taxonomists splitting and lumping genera, in particular to reflect structured biological diversity (Scotland and Sanderson, 2004; Strand and Panova, 2015; Barraclough and Humphreys, 2015; Campos et al., 2016).

Despite the recent highlighting of difficulties in modelling diversification rates (Louca and Pennell, 2020) it is clear that such rates in angiosperm taxa are highly variable (Magallon and Sanderson, 2001) and the uneven branching patterns of most angiosperm phylogenies suggests that the traits involved in species diversification are heterogeneously distributed and therefore have independently evolved and been lost many times in the evolutionary history of flowering plants (Doyle and Donoghue, 1993; Sanderson and Donoghue, 1994). One of the major challenges in the field of evolutionary biology is describing and analysing all potential factors underlying speciation, which would aid in the understanding of the remarkable differences in species diversification levels between angiosperm taxa.

The generation of a large number of species in a brief evolutionary period defines an evolutionary radiation (Simões et al., 2016). The triggering factors of radiations in different angiosperm clades have for many years been proposed to be mainly environmental and ecological (De Queiroz, 2002; Bromham and Cardillo, 2003; Davies et al., 2004b). Davies et al. (2004a) suggested that shifts in species diversification are likely

to be attributed to the effect of a complex process that reflects the simultaneous interaction of many ecological and environmental factors, rather than being the effect of a few key evolutionary innovations. Under this assumption, many species radiations have been studied (Richardson et al., 2001; Ellis et al., 2006; Valente et al., 2010; Marazzi and Sanderson, 2010). Among these, Donoghue and Sanderson (2015) highlighted the generality of these species radiations usually being related to species-poor lineages, or "depauperons", and uses this framework to study the radiation of Bromeliaceae species. Bouchenak-Khelladi et al. (2015) proposed a framework of variables impacting species radiations in different angiosperm clades, which are still only divided between intrinsic (ecological) and extrinsic (environmental) variables. Nonetheless, in recent years, the suggestion that other biological intrinsic factors such as hybridisation and polyploidisation can play an important role in angiosperm diversification patterns has gained strength (Hegarty and Hiscock, 2008; Fawcett and Van de Peer, 2010; Jiao et al., 2011). Leitch and Leitch (2012) suggest that the effect and interaction of a series of intrinsic biological factors that include not only ecological traits, but also non-adaptive genomic traits with an impact on genome diversification and evolution. Such factors, (shown in Figure 1.1), would have a varying influence among different plant groups contributing to their evolutionary rates. According to Leitch and Leitch (2012), it may have been underestimated how the genomic changes triggered by non-homologous recombination, transposition, hybridisation or polyploidy events can rapidly increase the genetic diversity levels among individuals or populations and trigger radiations in flowering plants. On the other hand pollination ecology, population size, growth rates and life cycle are among the ecological factors also playing a major role on the impact of genetic drift and selection among individuals or populations. Combined, all these factors may represent the key to better understand the speciation rate variation among angiosperms.

Variation in these factors and their interactions is likely responsible for the unevenly distributed species diversity of flowering plants. The largest angiosperm groups provide us with an opportunity to examine which factors are key to the dramatic radiations described in some of these groups and better understand the origins of biodiversity bursts. According to Frodin (2004), there are 57 angiosperm genera with more than

500 species, although this number may have changed in the last years due to the growth or split of some taxa. Some of these genera are very familiar to us and their radiations have been well studied (e.g. *Solanum* L., *Piper* L., *Begonia* L.), but others are more obscure (*Astragalus* L., *Miconia* Ruiz & Pav., *Psychotria* L.). However, working with large genera also leads to problems. The largest angiosperm genus has around 3000 species (*Astragalus* L.), and it is followed by genera with approximately 2000 species currently identified (*Bulbophyllum* Thouars, *Begonia* L., *Piper* L. and *Euphorbia* L.). For this reason, evolutionary studies on large genera at the genus wide level are rare. One of the few exceptions is Sanderson and Wojciechowski (1996), where the speciation rates of *Astragalus* L. (Fabaceae) were analysed, and it was found that actually these were not exceptionally high compared with its closest relatives. This suggested that in the case of *Astragalus*, the factors that drove the shift to higher speciation rates were not unique to this genus, but that it was rather a combined effect involving biological and ecological factors. Moreover, previous research on *Solanum* L. (Solanaceae) suggests mainly ecological factors (such as climate changes and dispersal events) as the main drivers of an adaptive radiation in Australian *Solanum*. However, more studies on the evolution of large genera are necessary to test for the involvement of intrinsic genomic properties in their diversification (Twyford, 2012).

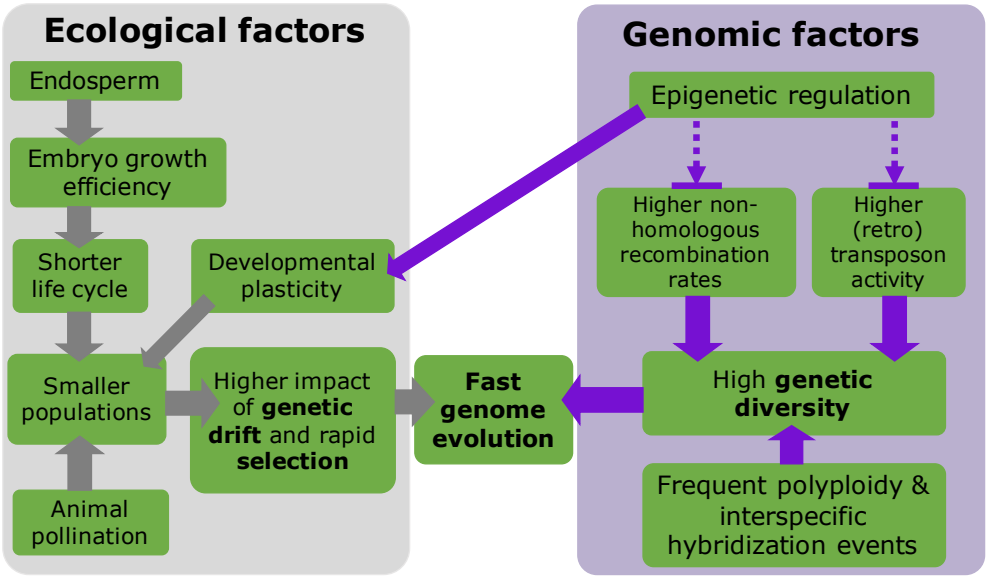


Figure 1.1: Key biological factors shaping angiosperm speciation. They are divided in two main groups: ecological factors, conditioned by the environment, and genomic factors, conditioned by sexual reproduction rates and genome dynamism. Figure modified from Leitch and Leitch (2012).

In the last years, the introduction of genomic studies into speciation research has brought the understanding of this evolutionary process to another level. New approaches that include the generation of large amounts of genomic data of non-model organisms have started to bring new and unexpected insights into the genetics of speciation (Seehausen et al., 2014). New genus-wide studies on the factors driving large angiosperm radiations specially involving research on genomic factors would give us new perspectives to better understand the evolutionary drivers of mega-diversity.

1.2 Dynamic genomes: the role and effects of repetitive DNA in angiosperm evolution

Any process of adaptation, genetic isolation and speciation is determined by genomic variation, caused by genetic changes that may arise spontaneously (e.g. point mutations, recombination errors) or be induced by stress conditions (e.g. some transposable elements activations). This natural genetic variation represents the driving force of evolution (Alonso-Blanco and Koornneef, 2000; Flood et al., 2011). As discussed above, the rates at which a genome can change and evolve are affected by a series of factors (Figure 1.1) that make it more or less dynamic. This genome dynamism is not only determined by point mutation rates, but also by a range of large-scale chromosomal changes that can be associated to the activation or expansion of repetitive DNA elements. The fraction of the genome containing these elements is called the repeatome.

The repeatome is dispersed across the genome. Each chromosome has loosely condensed euchromatic regions (where genes are present) that is interspersed with highly condensed heterochromatic regions (Schmidt and Heslop-Harrison (1998); King (2002), Figure 1.2). Along with other chromosomal elements such as the telomere, centromere and NORs (Nuclear Organising Regions), the heterochromatin constitutes the repeatome of the genome. The repeatome is composed by a mixture of repetitive elements comprising transposons, simple sequence repeats and satellite DNA (Choi et al., 2020). The repeatome is potentially the most variable part of the genome and their activity can cause severe changes in the genomic architecture (McClintock, 1953; Kidwell and Lisch, 1997). In angiosperms, the huge genome size variation found (Pellicer et al., 2018) is

largely due to variation in the size of the repeatome (Maumus and Quesneville, 2014).

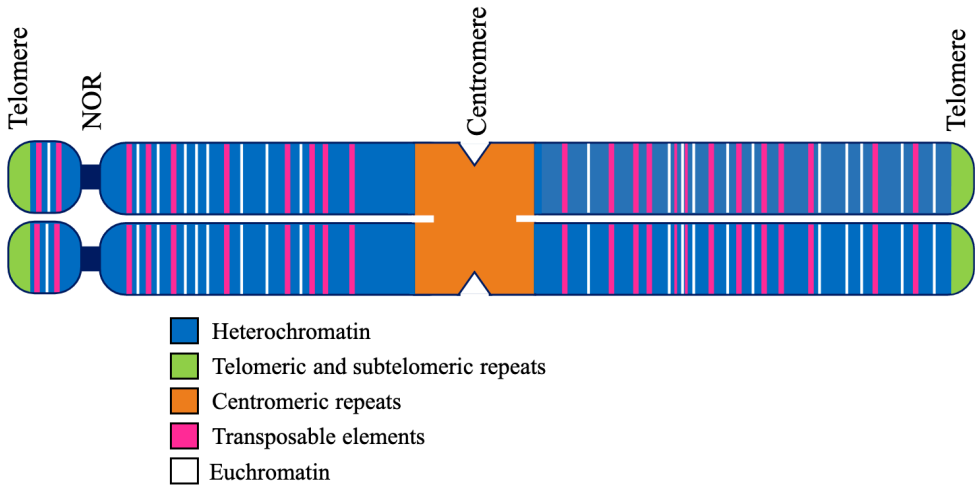


Figure 1.2: Schematic representation of the main genomic elements within a chromosome. Figure from Choi et al. (2020).

The elements in the repeatome have variable structure. Satellite DNA consists of large regions of tandemly repeated short repetitive sequences, whereas simple sequence repeats (or microsatellites) are much shorter regions of tandemly repeated DNA motifs of a few basepairs. Unlike transposons, these elements do not contain any coding domains. Transposons (also called Transposable Elements or TEs) are DNA fragments that can integrate into new positions in the genome and some of which are competent to increase their copy number over time. These elements can either be cut out and transposed into a new position (class I TE or DNA elements) or alternatively be copied and transposed keeping both the old and new positions (class II elements or retrotransposons), and their structure includes coding regions that makes them autonomous elements. All these elements are normally distributed in different chromosomal regions. Highly repeated satellite DNA and specific transposon families (LTR-*Gypsy* retrotransposons) are normally found in centromeric regions, whereas satellite DNA is also found in telomeres, and other transposon families and simple repeats and are dispersed along the chromosome (Figure 1.2, Choi et al. (2020); Biscotti et al. (2015)). Although TEs and satellite DNAs (satDNA) present structural differences, their homology has led to support the hypothesis that satDNA can be derived from TEs (Meštrović et al., 2015).

Transposons were first described in plants (McClintock, 1953), and their ability to alter

gene and/or genome structure explains why their discoverer Barbara McClintock first named them "controlling elements" (McClintock, 1950). Phenotypic variation caused by TEs was first described in maize (McClintock, 1953) and petunia (Gerats et al., 1990). Later, TEs have been described to trigger phenotypic variation in different plant species using different mechanisms (for a review, see Lisch (2013)). According to Lisch (2013), emerging research on comparative plant genomics and phenomics could provide better insights on how much of the variation that has been important for plant speciation and evolution has been driven by TE activity. Research suggests that TEs have played a more important role than polyploidy in plant genome size variation (Kellogg and Bennetzen, 2004; Ibarra-Laclette et al., 2013; Bennetzen and Wang, 2014). In the recent years, the extension of genome sequencing beyond model organisms has enabled research on repeatome structure and variation in an increasing range of angiosperm lineages (e.g. Novák et al. (2014); Beulé et al. (2015); Zhang et al. (2015); Jouffroy et al. (2016); Niu et al. (2019); Belyayev et al. (2020)).

The idea that TEs can drive speciation has been around for many years (Ginzburg et al., 1984). TEs are generally silenced across the genome, but certain factors (such as stress or genomic shock after hybridisation or polyploidisation) can stimulate their activation (Levin and Moran, 2011; Bousios and Gaut, 2016). Active TEs in a genome have the potential to accelerate genetic variation through generations, either through triggering genome rearrangements or by causing important gene disruptions or duplications (Kidwell and Lisch, 2000; Dodsworth et al., 2015). Eventually, this could give rise to crossing barriers between populations and consequently, genetic drift and speciation (for general reviews, see Belyayev (2014); Serrato-Capuchina and Matute (2018); for a plant-specific review on LTRs and speciation, see Galindo-González et al. (2017)). Figure 1.3 shows different case scenarios in which TEs can play a role in genomic diversification and speciation events (from Levy (2013) and Belyayev (2014)), either by gradual speciation, by their activation as a response to genomic shock (Figure 1.3 A), or activated by any other stress factors (Figure 1.3 B).

Recently the roles of TEs in evolution have been studied in different angiosperm genera, including *Nicotiana* (Parisod et al., 2012), *Aegillops* (Senerchia et al., 2013, 2015),

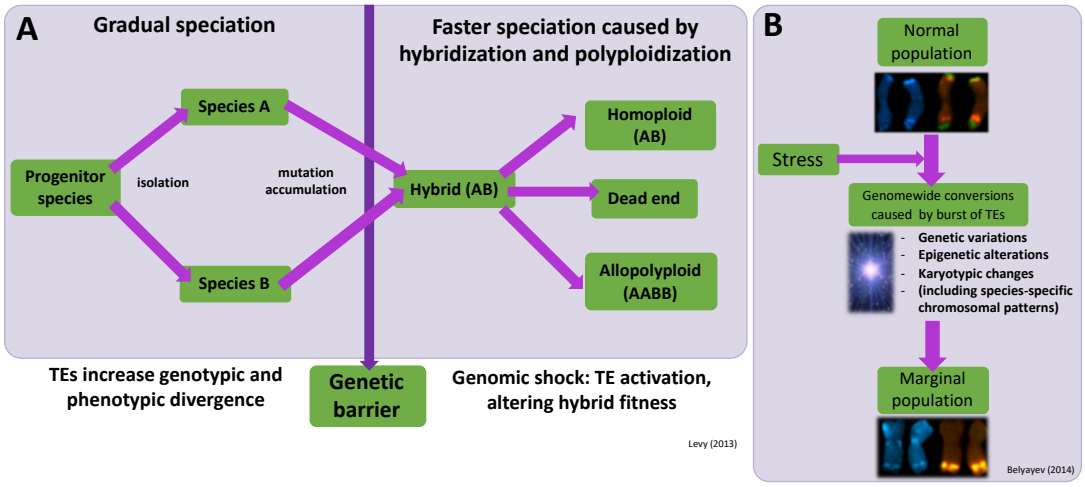


Figure 1.3: Examples of scenarios in which transposons could play important evolutionary roles. This contemplates from increase of natural variation by random activations or variations in TE silencing after genomic shock (A), to transposon bursts activated by stress factors (B). Figure modified from Levy (2013) and Belyayev (2014)

.

Biscutella (Bardil et al., 2015), *Asparagus* (Harkess et al., 2016), *Oryza* (Zhang Gao, 2017), *Helianthus* (Mascagni et al., 2017), *Cucumis* (Morata et al., 2018), *Arabidopsis* (Quadrana et al., 2019) or *Citrus* (Meng et al., 2020). However, most of these studies focus on the evolution of small taxonomic groups or of specific TE families. Although they are widely considered one of the main sources of genetic variation in plants, genome-wide studies on the role of the repeatome in the evolution of large plant radiations are still not very abundant.

1.3 Evolution and speciation in the mega-diverse genus *Begonia* L.

To better understand the drivers of plant mega-diversity and speciation, the genus *Begonia* is a suitable model. This pan-tropically distributed genus has the same species richness distribution patterns than the angiosperms (data from Hughes et al. (2015)). Moreover, this genus is among the five largest genera of flowering plants, with more species being described every year (over 1990 described species, as currently stated in Hughes et al. (2015), expected 2,000th species to be described in 2021 according to Hughes, M. pers. comm.). Previous research established that early diverging African *Begonia* groups dispersed and derived into the recently radiated Neotropical and Asian

Begonia clades (Goodall-Copestake et al., 2010). Recent collaborative work on *Begonia* taxonomy has put together its largest genus-wide phylogeny to date, and suggests a natural sectional division of the genus (Figure 1.4; Moonlight et al. (2018)). *Begonia* species are known for its wide phenotypical diversity, particularly in leaf shape (Neale, 2005; Ali, 2013). Additionally, *Begonia* species grow in a wide range of habitats of varying light and humidity levels (**?**). This group also has the advantage of being suitable for genetic studies because it (i) produces many seeds per cross, has a manageable genome size, detailed cultivation studies have already been undertaken and (iv) there are a growing number of genetics and genomic resources (Neale et al., 2006; Dewitte et al., 2009; Brennan et al., 2012).

The mega-diversity of *Begonia* contrasts with its only sister genus in the family Begoniaceae, *Hillebrandi*a Oliv. (Hughes, 2002), which has only one species (*H. sandwciensis*) and is restricted to the Hawaiian Islands (Clement et al., 2004). This agrees with Donoghue and Sanderson (2015) on the generalisation of "depauperon" sister clades of large radiations. Moreover, within the Cucurbitales order and compared to other Cucurbitaceae genera, *Begonia* is the most species-rich genus (according to the RBGKew database POWO, plantsoftheworldonline.org). Within Begonia, its early diverging clades from mainland Africa are relatively species-poor. In the Neotropics, Asia and Madagascar, the genus is exceptionally rich and exhibits considerably high speciation rates compared to its African sister clades (Moonlight et al. (2015, 2018), Hughes M. and Wilson H. P., pers comm). *Begonia*, therefore, offers several phylogenetic contrasts for the analysis of species radiations.

Begonia seems to gather the four factors that according to Sanderson and Wojciechowski (1996) influence species diversification in large plant genera: strong population structure, herbaceous habit, rapid chromosome evolution and frequent transitions between ecological specialisations (Twyford, 2012). However, although *Begonia* habitat and leaf diversity could lead to the assumption that these radiations are adaptive, recent research suggests otherwise (Moonlight, 2017). Instead, research evidence suggests *Begonia* evolution rather follows a "random walk" determined by non-adaptive, mutation-order speciation events. According to Nosil and Flaxman (2011), in this model of

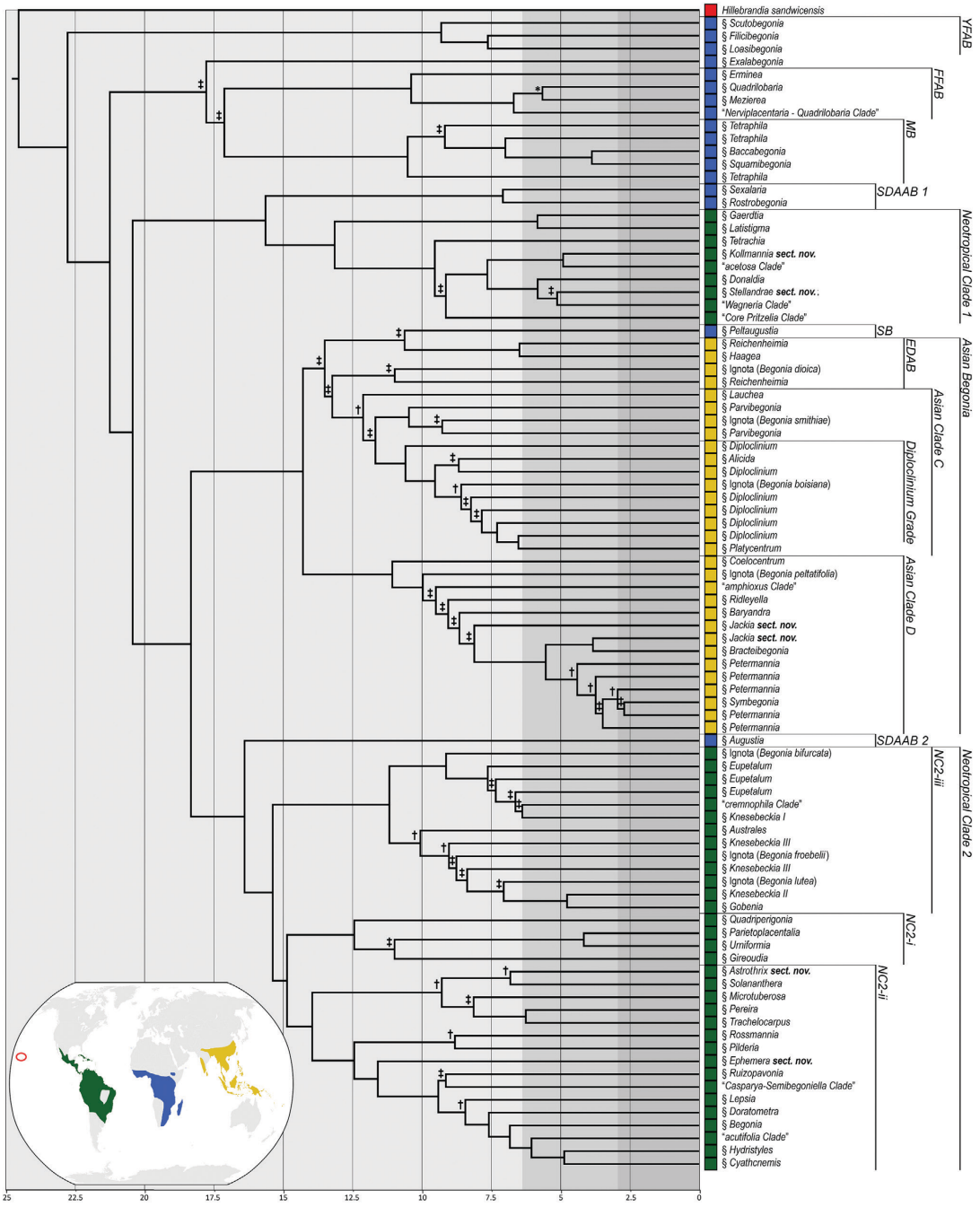


Figure 1.4: Sectional time-calibrated *Begonia* phylogeny from Moonlight et al. (2018). Squares near each branch indicate section distributions (yellow, Africa; blue, Asia; green, America; and red, Hawaii).

speciation incompatible mutations fix in different populations and adapt to the same selective pressure. Many of the conditions needed for this type of speciation process have already been described in *Begonia*, including non-adaptive niche evolution (Moonlight, 2017), low gene flow and high population structure (Hughes and Hollingsworth, 2008; Twyford et al., 2014), high hybrid incompatibility and similar phenotypes formed from different genetic pathways (Kidner, unpublished). Furthermore, (Dewitte et al., 2009) evidenced some genome size and chromosome number variation in *Begonia*, suggesting this genus presents some level of genome dynamics, which could be another of the factors stimulating the hypothesized non-adaptive speciation events for this genus. All the above provide a solid reason to investigate the role of genome dynamics in the recent, rapid radiations of the mega-diverse genus *Begonia*.

1.4 Objectives

This project aims to bring new insights on which genomic factors have influenced speciation processes in *Begonia*. The main objective is to better understand the role of the repeatome in the large radiations of *Begonia* species by studying variation in repeat structure in *Begonia* genomes through several approaches.

The main hypothesis of this project states that the evolution of the *Begonia* genus has been shaped by bursts of transposon activity and repeatome dynamics, which have triggered genomic re-arrangements and contributed to reproductive isolation and to higher speciation rates in some highly specious *Begonia* sections.

To test this hypothesis, each of my thesis chapters will look into *Begonia* genome dynamics at different levels. The specific aims and approaches taken are:

- Chapter 2: Documenting the patterns of chromosome change across the *Begonia* phylogeny by gathering all known chromosome number and genome size data available for *Begonia*. In this chapter, I generate up to date chromosome number and genome size datasets for Begoniaceae, in order to study the distribution of these data across its family-wide phylogeny.

- Chapter 3: Assembling genomes of a phylogenetically balanced set of *Begonia*

species for comparative analyses across four radiations. With this aim, in Chapter 3 I describe the sequencing and assembly of seven Begoniaceae genomes from all main clades, and perform the gene annotation of all available genome assemblies of *Begonia* to date. This chapter addresses the challenges of assembling complex genomes of non-model organisms, and describes the genome assembly data that will be used in Chapter 4.

- Chapter 4: Studying and comparing the repeat landscapes of 12 Begoniaceae species to look for correlations with radiations. With this aim, I use Chapter 3 data and different repeat annotation approaches to describe and compare the repeat landscapes and transposon dynamics of 12 Begoniaceae species from across the different *Begonia* radiations.

- Chapter 5: Examining the repeatome dynamics across a highly speciose *Begonia* section, to associate specific changes in the repeatome with species diversification events. In this chapter I use a genome skimming approach to take a closer look into one of the largest *Begonia* radiations (*B.* sect. *Gireoudia*). Genomic data of 34 species is used to test whether species from this section have more dynamic repeatomes than those from their species-poor relative clades, and whether these repeat changes can be dated to the time of specific radiation or speciation events.

- Chapter 6: Investigating intraspecific repeatome dynamics to link repeat variation with reproductive barriers. In this chapter, a genome skimming approach is again used to test the genomic factors driving population divergence in seven different *Begonia heracleifolia* natural populations.

Finally, in Chapter 7 I synthesise the main findings from all five data chapters, discuss the main significance of this study in a broader context and point at the directions for future studies in *Begonia* speciation genomics.

Bibliography

Ali MS. 2013. Genetic architecture of species level differences in Begonia Section Gireoudia. Ph.D. thesis, University of Edinburgh.

Alonso-Blanco C, Koornneef M. 2000. Natural variation, an underexploited resource of genetic variation for plant genetics. *Trends in Plant Science*. 5:22–29.

Bardil A, Tayalé A, Parisod C. 2015. Evolutionary dynamics of retrotransposons following autopolyploidy in the Buckler Mustard species complex. *The Plant Journal*. 82:621–631.

Barraclough TG, Humphreys AM. 2015. The evolutionary reality of species and higher taxa in plants: a survey of post-modern opinion and evidence. *New Phytologist*. 207:291–296.

Belyayev A. 2014. Bursts of transposable elements as an evolutionary driving force. *Journal of evolutionary biology*. 27:2573–2584.

Belyayev A, Josefiová J, Jandová M, Mahelka V, Krak K, Mandák B. 2020. Transposons and satellite dna: on the origin of the major satellite dna family in the chenopodium genome. *Mobile DNA*. 11:1–10.

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology*. 65:505–530.

Beulé T, Agbessi MD, Dussert S, Jaligot E, Guyot R. 2015. Genome-wide analysis of ltr-retrotransposons in oil palm. *BMC genomics*. 16:1–14.

Biscotti MA, Olmo E, Heslop-Harrison JSP. 2015. Repetitive DNA in eukaryotic genomes. *Chromosome Research*. 23:415–420.

Bouchenak-Khelladi Y, Onstein RE, Xing Y, Schwery O, Linder HP. 2015. On the complexity of triggering evolutionary radiations. *New Phytologist*. 207:313–326.

Bousios A, Gaut BS. 2016. Mechanistic and evolutionary questions about epigenetic conflicts between transposable elements and their plant hosts. *Current opinion in plant biology*. 30:123–133.

Brennan AC, Bridgett S, Ali MS, Harrison N, Matthews A, Pellicer J, Twyford AD, Kidner CA. 2012. Genomic resources for evolutionary studies in the large, diverse, tropical genus, begonia. *Tropical plant biology*. 5:261–276.

Bromham L, Cardillo M. 2003. Testing the link between the latitudinal gradient in species richness and rates of molecular evolution. *Journal of evolutionary biology*. 16:200–207.

Campos L, Salvo AE, Flores-Moya A. 2016. Natural taxonomic categories of angiosperms obey benford's law, but artificial ones do not. *Systematics and Biodiversity*. 14:431–440.

Choi IY, Kwon EC, Kim NS. 2020. The c-and g-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy. *Genes & genomics*. 42:699–714.

Clement WL, Tebbitt MC, Forrest LL, Blair JE, Brouillet L, Eriksson T, Swensen SM. 2004. Phylogenetic position and biogeography of hillebrandia sandwicensis (begoniaceae): a rare hawaiian relict. *American Journal of Botany*. 91:905–917.

Davies TJ, Barraclough TG, Chase MW, Soltis PS, Soltis DE, Savolainen V. 2004a. Darwin's abominable mystery: insights from a supertree of the angiosperms. *Proceedings of the National Academy of Sciences*. 101:1904–1909.

Davies TJ, Savolainen V, Chase MW, Moat J, Barraclough TG. 2004b. Environmental energy and evolutionary rates in flowering plants. *Proceedings of the Royal Society of London. Series B: Biological Sciences*. 271:2195–2200.

De Queiroz A. 2002. Contingent predictability in evolution: key traits and diversification. *Systematic biology*. 51:917–929.

Dewitte A, Leus L, Eeckhaut T, Vanstechelman I, Van Huylenbroeck J, Van Bockstaele E. 2009. Genome size variation in begonia. *Genome*. 52:829–838.

Dodsworth S, Leitch AR, Leitch IJ. 2015. Genome size diversity in angiosperms and its influence on gene space. *Current opinion in genetics & development*. 35:73–78.

Donoghue MJ, Sanderson MJ. 2015. Confluence, synnovation, and depauperons in plant diversification. *New Phytologist*. 207:260–274.

Doyle JA, Donoghue MJ. 1993. Phylogenies and angiosperm diversification. *Paleobiology*. pp. 141–167.

Ellis AG, Weis AE, Gaut BS. 2006. Evolutionary radiation of "stone plants" in the genus argyroderma (aizoaceae): unraveling the effects of landscape, habitat, and flowering time. *Evolution*. 60:39–55.

Fawcett JA, Van de Peer Y. 2010. Angiosperm polyploids and their road to evolutionary success. *Trends in Evolutionary Biology*. 2:e3–e3.

Flood PJ, Harbinson J, Aarts MG. 2011. Natural genetic variation in plant photosynthesis. *Trends in plant science*. 16:327–335.

Frodin DG. 2004. History and concepts of big plant genera. *Taxon*. 53:753–776.

Galindo-González L, Mhiri C, Deyholos MK, Grandbastien MA. 2017. Ltr-retrotransposons in plants: Engines of evolution. *Gene*. 626:14–25.

Gerats A, Huits H, Vrijlandt E, Maraña C, Souer E, Beld M. 1990. Molecular characterization of a nonautonomous transposable element (dtph1) of petunia. *The Plant Cell*. 2:1121–1128.

Ginzburg LR, Bingham PM, Yoo S. 1984. On the theory of speciation induced by transposable elements. *Genetics*. 107:331–341.

Goodall-Copestake WP, Perez-Espona S, Harris DJ, Hollingsworth PM. 2010. The early evolution of the mega-diverse genus begonia (begoniaceae) inferred from organelle dna phylogenies. *Biological journal of the Linnean society*. 101:243–250.

Harkess A, Mercati F, Abbate L, McKain M, Pires JC, Sala T, Sunseri F, Falavigna A, Leebens-Mack J. 2016. Retrotransposon proliferation coincident with the evolution of dioecy in asparagus. *G3: Genes, Genomes, Genetics*. 6:2679–2685.

Hegarty MJ, Hiscock SJ. 2008. Genomic clues to the evolutionary success of polyploid plants. *Current biology*. 18:R435–R444.

Heywood V. 1993. Flowering plants of the world. *Oxford Univ. Press New York*. .

Hughes M. 2002. Population structure and speciation in Begonia L. Ph.D. thesis, University of Glasgow.

Hughes M, Hollingsworth P. 2008. Population genetic divergence corresponds with species-level biodiversity patterns in the large genus begonia. *Molecular ecology*. 17:2643–2651.

Hughes M, Moonlight P, Jara A, Tebbitt M, Wilson H, Pullan M. 2015. Begonia resource centre. online database available from http://padme.rbge.org.uk/begonia/.

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, et al. (11 co-authors). 2013. Architecture and evolution of a minute plant genome. *Nature*. 498:94–98.

Jiao Y, Wickett NJ, Ayyampalayam S, et al. (11 co-authors). 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*. 473:97–100.

Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F. 2016. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC genomics*. 17:624.

Kellogg EA, Bennetzen JL. 2004. The evolution of nuclear genome structure in seed plants. *American Journal of Botany*. 91:1709–1725.

Kidwell MG, Lisch D. 1997. Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences*. 94:7704–7711.

Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. *Trends in ecology & evolution*. 15:95–99.

King GJ. 2002. Through a genome, darkly: comparative analysis of plant chromosomal dna. *Plant molecular biology*. 48:5–20.

Leitch A, Leitch I. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist*. 194:629–646.

Levin HL, Moran JV. 2011. Dynamic interactions between transposable elements and their hosts. *Nature Reviews Genetics*. 12:615–627.

Levy AA. 2013. Transposons in plant speciation. In: Fedoroff NV, editor, Plant Transposons and Genome Dynamics in Evolution, John Wiley & Sons, chapter 9, pp. 165–179.

Lisch D. 2013. How important are transposons for plant evolution? *Nature Reviews Genetics*. 14:49–61.

Louca S, Pennell MW. 2020. Extant timetrees are consistent with a myriad of diversification histories. *Nature*. 580:502–505.

Magallon S, Sanderson MJ. 2001. Absolute diversification rates in angiosperm clades. *Evolution*. 55:1762–1780.

Marazzi B, Sanderson MJ. 2010. Large-scale patterns of diversification in the widespread legume genus senna and the evolutionary role of extrafloral nectaries. *Evolution: International Journal of Organic Evolution*. 64:3570–3592.

Mascagni F, Giordani T, Ceccarelli M, Cavallini A, Natali L. 2017. Genome-wide analysis of ltr-retrotransposon diversity and its impact on the evolution of the genus helianthus (l.). *BMC genomics*. 18:634.

Maumus F, Quesneville H. 2014. Deep investigation of arabidopsis thaliana junk dna reveals a continuum between repetitive elements and genomic dark matter. *PLoS One*. 9:e94101.

McClintock B. 1950. Mutable loci in maize. carnegie inst. wash. *Year Book*. 49:157–167.

McClintock B. 1953. Induction of instability at selected loci in maize. *Genetics*. 38:579.

Meng H, Feng J, Bai T, Jian Z, Chen Y, Wu G. 2020. Genome-wide analysis of short interspersed nuclear elements provides insight into gene and genome evolution in citrus. *DNA Research*. 27:dsaa004.

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M. 2015. Structural and functional liaisons between transposable elements and satellite dnas. *Chromosome Research*. 23:583–596.

Moonlight PW. 2017. Linking speciation and the niche: taxonomy, phylogeny and niche evolution in neotropical Begonia. Ph.D. thesis, University of Glasgow.

Moonlight PW, Ardi WH, Padilla LA, et al. (29 co-authors). 2018. Dividing and conquering the fastest-growing genus: Towards a natural sectional classification of the mega-diverse genus begonia (Begoniaceae). *Taxon*. 67:267–323.

Moonlight PW, Richardson JE, Tebbitt MC, Thomas DC, Hollands R, Peng CI, Hughes M. 2015. Continental-scale diversification patterns in a megadiverse genus: The biogeography of neotropical begonia. *Journal of Biogeography*. 42:1137–1149.

Morata J, Tormo M, Alexiou KG, Vives C, Ramos-Onsins SE, Garcia-Mas J, Casacuberta JM. 2018. The evolutionary consequences of transposon-related pericentromer expansion in melon. *Genome biology and evolution*. 10:1584–1595.

Neale S. 2005. Evolution of LEAFY homologues in Begonia L. Ph.D. thesis, University of Glasgow (United Kingdom).

Neale S, Goodall-Copestake W, Kidner C, et al. (4 co-authors). 2006. The evolution of diversity in begonia.

Niu XM, Xu YC, Li ZW, et al. (11 co-authors). 2019. Transposable elements drive rapid phenotypic variation in capsella rubella. *Proceedings of the National Academy of Sciences*. 116:6908–6913.

Nosil P, Flaxman SM. 2011. Conditions for mutation-order speciation. *Proceedings of the Royal Society B: Biological Sciences*. 278:399–407.

Novák P, Hřibová E, Neumann P, Koblížková A, Doležel J, Macas J. 2014. Genome-wide analysis of repeat diversity across the family musaceae. *PloS one*. 9:e98918.

Parisod C, Mhiri C, Lim KY, Clarkson JJ, Chase MW, Leitch AR, Grandbastien MA. 2012. Differential dynamics of transposable elements during long-term diploidization of nicotiana section repandae (solanaceae) allopolyploid genomes. *PLoS One*. 7:e50352.

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. 2018. Genome size diversity and its impact on the evolution of land plants. *Genes*. 9:88.

Quadrana L, Etcheverry M, Gilly A, et al. (11 co-authors). 2019. Transposition favors

the generation of large effect mutations that may facilitate rapid adaption. *Nature communications*. 10:1–10.

Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM. 2001. Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science*. 293:2242–2245.

Sanderson MJ, Donoghue MJ. 1994. Shifts in diversification rate with the origin of angiosperms. *Science*. 264:1590–1593.

Sanderson MJ, Wojciechowski MF. 1996. Diversification rates in a temperate legume clade: are there "so many species" of astragalus (fabaceae)? *American Journal of Botany*. 83:1488–1502.

Schmidt T, Heslop-Harrison J. 1998. Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends in Plant Science*. 3:195–199.

Scotland RW, Sanderson MJ. 2004. The significance of few versus many in the tree of life. *Science*. 303:643–643.

Seehausen O, Butlin RK, Keller I, et al. (11 co-authors). 2014. Genomics and the origin of species. *Nature Reviews Genetics*. 15:176–192.

Senerchia N, Felber F, Parisod C. 2015. Genome reorganization in F1hybrids uncovers the role of retrotransposons in reproductive isolation. *Proceedings of the Royal Society B: Biological Sciences*. 282:20142874–20142874.

Senerchia N, Wicker T, Felber F, Parisod C. 2013. Evolutionary dynamics of retrotransposons assessed by high-throughput sequencing in wild relatives of wheat. *Genome Biology and Evolution*. 5:1010–1020.

Serrato-Capuchina A, Matute DR. 2018. The role of transposable elements in speciation. *Genes*. 9:254.

Simões M, Breitkreuz L, Alvarado M, Baca S, Cooper J, Heins L, Herzog K, Lieberman B. 2016. The evolving theory of evolutionary radiations. *Trends in ecology & evolution*. 31:27–34.

Strand M, Panova M. 2015. Size of genera–biology or taxonomy. *Zoologica Scripta*. 44:106–116.

Twyford AD. 2012. Speciation and gene flow in Central American Begonia L.(Begoniaceae). Ph.D. thesis, The University of Edinburgh.

Twyford AD, Kidner CA, Ennos RA. 2014. Genetic differentiation and species cohesion in two widespread central american begonia species. *Heredity*. 112:382–390.

Valente LM, Reeves G, Schnitzler J, Mason IP, Fay MF, Rebelo TG, Chase MW, Barraclough TG. 2010. Diversification of the african genus protea (proteaceae) in the cape biodiversity hotspot and beyond: equal rates in different biomes. *Evolution: International Journal of Organic Evolution*. 64:745–760.

Zhang Y, Cheng C, Li J, Yang S, Wang Y, Li Z, Chen J, Lou Q. 2015. Chromosomal structures and repetitive sequences divergence in cucumis species revealed by comparative cytogenetic mapping. *BMC genomics*. 16:1–13.

Chapter 2

Evolutionary patterns of genome size and chromosome number variation across the Begoniaceae family

This chapter includes the collaborative work and unpublished data of Prof. Kuo-Fang Chung, Dr. Daisy Hu and Dr. Yoshiko Kono (Academia Sinica, Taiwan) on *Begonia* chromosome evolution.

Also included here is our collaborative work with Dr. Jaume Pellicer-Moscardó and Dr. Ilia J. Leitch (Royal Botanic Gardens, Kew) on *Begonia* genome size evolution.

2.1 Introduction

Changes in chromosome structure and number can trigger speciation events by establishing nuclear incompatibilities that lead to crossing barriers (Lowry and Willis, 2010; Winterfeld et al., 2014; Baack et al., 2015). Events such as chromosomal rearrangements, polyploidy, and dysploidy (gain or loss of single chromosomes by chromosome fusion or fission) are frequent events in plant evolution (Wood et al., 2009; Winterfeld et al., 2020). This makes angiosperms one of the groups of organisms with the widest ranges of chromosome numbers: from n=2 (described in approx. 6 sp, Vanzela et al. (1996); Cremonini (2005)) to n > 320 (Uhl, 1978). Cytogenetic and evolutionary studies have detected an ancient polyploidisation event in the early divergence of seed plants, as well as other various ancestral polyploidy events in the crown nodes of specific angiosperm clades (Jiao et al., 2011; Li et al., 2015), suggesting that such events play an important role in many plant lineages. Although changes in ploidy levels generate high levels of evolutionary novelty (Doyle and Coate, 2019) and have a widely stablished evolutionary role in the diversification of major angiosperm families (Stebbins and others, 1971; Husband and Sabara, 2004; Knight et al., 2005; Han et al., 2020), dysploidy events are known to be more frequent than polyploidy in flowering plants (Grant, 1981). Moreover, more recent studies demonstrate these events have an important evolutionary impact in specific plant lineages, and that these chromosomal changes persist longer than polyploid changes (Escudero et al., 2014). Studying karyotype variation across a taxonomic group provides wide information about its evolutionary history, and can provide insights into the role such changes have had in diversification events. In the context of this thesis, studying chromosome number diversity and evolution of the Begoniaceae can bring critical insights on how much genomic structure variation is observed within this group, and how strong the taxonomic signal and evolutionary role of this variation are.

In lineages with stable chromosome counts, genome size changes due to repetitive DNA dynamics are responsible for most karyotype variation and evolution. In angiosperms, genome size (traditionally also referred to as the C-value (1C) and defined by the size of a haploid genome) shows the widest range known of all eukaryotes: from 61

to 148,852 Mbp, a 2440-fold difference (Schubert and Vu, 2016; Pellicer et al., 2018). The evolutionary forces behind this variation have been widely studied (for reviews, see Schubert and Vu (2016); Carta et al. (2020); Blommaert (2020)). Many research studies have previously focused on C-value dynamics of a specific phylogenetic frame and reported different levels of variation within a specific taxonomic group. Among others, this has been studied at the genus level in *Sorghum* and other genera (Price et al., 2005; Boutanaev and Nemchinov, 2020), at the family level in Liliaceae (Leitch et al., 2007), Brassicaceae (Lysak et al., 2009), Orchidaceae (Leitch et al., 2009) or Melanthiaceae (Pellicer et al., 2014), and even at the order level on some groups, such as Nymphaeales (Pellicer et al., 2013). Although absolute C-values are not correlated to speciation rates, it has been found that genome size variation rates are indeed linked to speciation rates in angiosperms. This means that taxonomic groups with higher genome size variation (but not necessarily with larger genomes) show higher speciation rates (Puttick et al., 2015). In the context of this thesis, studying the variation of c-values within the *Begonia* genus as well ats its phylogenetic distribution will provide insights of the genome dynamics not explained by karyotype changes, but likely by the presence and proliferation of repetitive DNA.

The genus *Begonia* is very large and taxonomically complex, and some species relationships are still not well resolved (Moonlight et al., 2018). Studying karyotype variation across its phylogeny will reveal the extent of chromosome dynamics in this genus. Most previous studies in *Begonia* cytology (Legro and Haegeman, 1971; Forrest, 2000; Dewitte, 2010; Hong-Zhe et al., 2005; Hughes et al., 2011; Peng et al., 2014) are focused on karyotype compatibility of commercial hybrids and on species descriptions. Although useful, these do not show a general overview of the *Begonia* karyotype variation or how stable these values are across different *Begonia* groups. Other previous studies look into *Begonia* karyotype diversity, but mostly focus on specific sections or clades (Legro and Doorenbos, 1969; Oginuma and Peng, 2002). Dewitte (2010) and Dewitte et al. (2009) gathered a small list of genus-wide chromosome data and grouped it by continent of origin, and concludes that the karyotypic variation found made impossible to assign a basic chromosome number within the genus. This work not only describes genome size variation, polyploidy, aneuploidy and presence of B chromosomes in this genus but also

indicates that *Begonia* "subgroups" can be characterised by particular chromosome numbers. It concludes nonetheless that a larger chromosome number dataset is needed for a deeper understanding of chromosome evolution in *Begonia*.

In the last few years, several newly published descriptions of *Begonia* species has included karyotype information, and the numbers of species karyotyped has increased (Tseng et al., 2017; Peng et al., 2017; Liu et al., 2020). However, there are difficulties that make karyotype description in this group uncommon. For example, *Begonia* species have small chromosomes (Dewitte 2009), and counts can be confused by the presence of satellite chromosomes and by high levels of dysploidy, resulting in variable published values within species (Rice et al., 2015).

Genome size variation in *Begonia* species has also been previously studied. Dewitte et al. (2009) used a flow cytometry approach to report the genome sizes of 37 species and 23 *Begonia* hybrids. This data showed 7-fold c-value differences (from 0.23 to 1.46 pg), and up to 12-fold difference in chromosome size. These results showed for the first time large differences in genome size in this genus, and suggest they are the result of chromosome evolution after polyploidisation events, likely followed by chromosome number decrease and genome stabilisation, which was also previously suggested by Oginuma and Peng (2002). Although Dewitte et al. (2009) found chromosome size trends when grouping species from different continents, the species sampling was not large enough to provide a good overview of the evolution of this trait across all different *Begonia* radiations. More genome size data to confirm these patterns, and further investigation on which genomic factors are triggering genome dynamics within the different *Begonia* groups would provide a better insight on the role of these dynamics on the evolution of *Begonia* species.

Bearing these premises in mind, the specific aims of this chapter are:

- Collecting all available chromosome number data in *Begoniaceae*
- Classifying and curating these data according to the latest nomenclature changes and taxonomy (Hughes et al., 2015; Moonlight et al., 2018) and elaborating an updated, complete chromosome dataset for *Begonia*.

- Mapping chromosome number variation in the latest *Begonia* phylogeny by Moonlight et al. (2018), arranging it by clade and section to find the evolutionary patterns of chromosome change in this genus.

- Determining whether larger rates of chromosome variation are associated with large radiation events.

- Collecting all available genome size data available for *Begonia*, as well as generating some for the species of study in Chapters 3, 4 and 6.

- Studying genome size variation across the *Begonia* phylogeny by clade to find specific *Begonia* groups with especially interesting variation patterns.

- Studying the correlation levels of C-values with chromosome numbers of *Begonia* species to investigate the level of chromosome size variation that could be linked to gain/loss of repetitive DNA.

This in-depth, updated family-wide study of the *Begonia* genome architecture will help investigate the evolutionary consequences of chromosomal change in *Begonia* speciation events.

2.2 Results and discussion

2.2.1 Chromosome variation across the *Begonia* phylogeny

Data compilation

For this chapter, 434 Begoniaceae species chromosome numbers were gathered from different sources (full table in appendix 2.A.1). Out of these, 41 were from African species, 151 from American and 242 were from Asian *Begonia* species, which represent between 23 and 25% of the number of species described to date in each continent. For all of these, species names and taxonomic classification were updated, assigning them section and clade according to the latest phylogeny (Moonlight et al., 2018). According to the Begonia Resource Centre (Hughes et al., 2015), there are 1991 species of *Begonia* (last checked 18th September 2020). This means only approximately 22% of the genus has known chromosome numbers.

Moonlight et al. (2018) classifies all *Begonia* species into 6 African clades: Yellow-Flowered African Begonia or YFAB (6 sections, 58 species), Fleshy-Fruited African Begonia or FFAB (5 sections, 51 species), Malagasy Begonia or MB (3 sections, 36 species), Socotran Begonia or SB (1 section, 2 species), and Seasonally-Dry Adapted African Begonia clades 1 and 2 or SDAAB1-2 (2 sections, 10 species and 1 section, 12 species respectively), 2 main Neotropical clades: NC1 (8 sections, 201 species) and NC2, which is divided into 3 sub-clades NC2-i (4 sections, 134 species), NC2-ii (15 sections, 194 species) and NC2-iii (5 sections, 102 species), and 3 main Asian clades: Early Diverging Asian Begonia or EDAB (3 sections, 24 species), Asian clade C (8 sections, 328 species) and Asian clade D (8 sections, 677 species). 80 *Begonia* species are classified under unknown section (*Begonia* Sect. Ignota in Moonlight et al. (2018)).

The proportion of chromosome counts found for each of these clades is shown in Figure 2.1 A. Although variation in the extent of chromosome count data available between clades has been found, there is chromosome information on more than 10% of all these clades except for the African clade FFAB. This could be due to difficulty of keeping these species in cultivation. However, the taxonomic groups with higher proportion of data are African clades too, because these also are, in general terms, the smallest

Begonia clades. Both American and Asian clades show approximately the same proportion of karyotyped species, with some clades such as NC2-i and Asia C containing the highest relative number of chromosome data. This is probably due to the presence of sections *Gireoudia* and *Platycentrum* in these clades, which are two of the largest and most widely studied *Begonia* sections (Burt-Utley, 1985; Dewitte et al., 2011; Twyford et al., 2013; Ali, 2013; Tebbitt and Dickson, 2000; Tebbitt et al., 2006; Nguyen, 2004). These sections have 110 and 180 species currently described, respectively, and data from 42 and 91 of them were gathered, which represents approx. the 38% and 50% of the size of each section.

Out of the 70 *Begonia* sections described to date, for 24 of them no chromosome numbers were found. Out of these, 12 were American sections, 5 were African and 8 were Asian sections. For the remaining sections, the numbers of chromosome counts foun were highly variable. The percentage of species in each section with karyotype data is mostly below 60, as seen in Figure 2.1 B. This is probably enough for most sections, but it highly depends on each section's size, taxonomic complexity and resolution. For most sections, their karyotype variation level can be appreciated without having all species data but a good enough fraction. However, some sections may need some special attention as in some cases, each observed species shows a different chromosome number and their basic numbers are not clear. For 6 small sections (with from 1 to 5 species), the chromosome data of all (100%) their species were gathered.

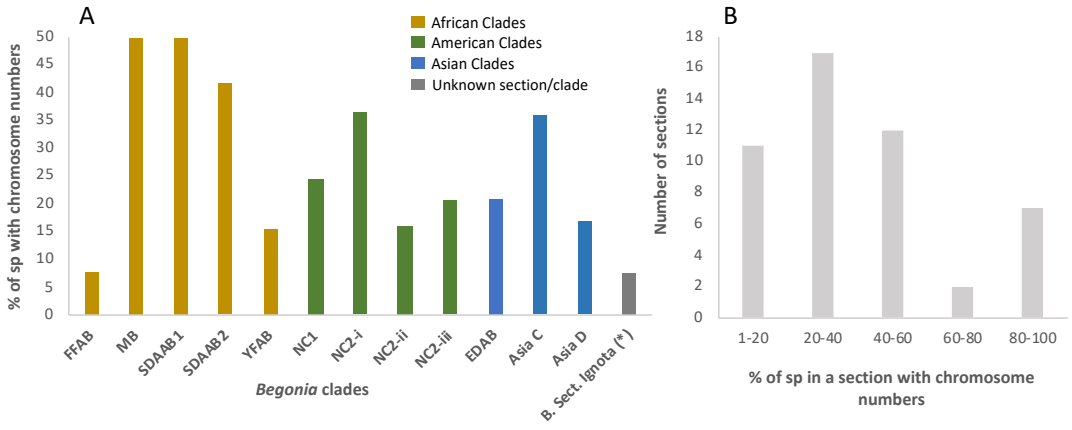


Figure 2.1: A. Relative completeness of chromosome data found across *Begonia* clades. B. Numbers of sections with relative chromosome counts between the specified intervals. For most sections, the collected data represents only less than 60% of their species number. For 6 sections, all species chromosome numbers have been found.

The dataset in this chapter represents a complete and updated summary of all chromosome data that has been observed and published in this family to date. It includes *Begonia* haploid chromosome numbers that range from n=7 to n=52. Although the median and mode of the chromosome numbers found is n=15, chromosome number is highly variable across the phylogeny. N=15 being the most common is likely due to 15 being the most present number in several Asian sections, and our dataset is dominated by species from these groups.

Karyotype evolution in *Begonia* has previously shown to be complex and no consensus in a basic chromosome number has been established (Dewitte, 2010). In most large plant genera, chromosome number information on 22% of the species and most clades and sections would be enough to elucidate their karyotype evolution and basic karyotypes. Other large genera such as *Astragalus* also show karyotype variation and levels of dysploidy. Chromosome numbers from n=6 to n=45 have been described in 836 karyotyped species, 27% of the genus size (according to the Chromosome Count Database CCBD, Rice et al. (2015) and the Plants of the World Online database POWO, http://plantsoftheworldonline.org/). However, a clear basic number is inferred from this data of x=8, found in nearly the 80% of its karyotyped species. Other examples of the largest genera such as *Bulbophyllum* Thouars (n=18 to n=24, but n=19 in 85% of karyotyped species), or *Psychotria* L. (n=11 to n=66 but 11 is the basic number in 85% of karyotyped species) have only published karyotypes of between 2 and 4% of their species (Rice et al. (2015) and POWO, http://plantsoftheworldonline.org/). This means that it is very likely that *Begonia* cytology is not understudied but rather that this genus shows a complex chromosome evolutionary history that is not found in most of the other mega-diverse genera. Interestingly, only the case of the also large and taxonomically complex genus *Euphorbia* L. is comparable to *Begonia* in this context: with 1976 species, 261 (13%) karyotypes are published (Rice et al. (2015) and POWO, http://plantsoftheworldonline.org/). This genus shows a wide range of haploid chromosome numbers (n=6 to n=100), and no basic number or clear consensus on its karyotype evolution. In addition, genome size variation in *Euphorbia* shows a 48-fold difference (342 Mbp to 14.08 Gbp, Pellicer and Leitch (2020)). There are not many studies on the cause of this genomic variation, but previous research has established

that although natural hybridisation is infrequent in *Euphorbia* (Sheidai et al., 2010), aneuploidy and polyploidy have played a role in *Euphorbia* speciation (Perry, 1943).

Which *Begonia* clades show more karyotype stability?

Our data suggests we have a good representation of karyotype information for most *Begonia* groups. Figure 2.2 shows the variation of haploid chromosome numbers across the main *Begonia* clades established in (Moonlight et al., 2018). This figure displays a wide range of chromosome number variation in large Neotropical and Asian radiations, in contrast with African clades, which only show three most common karyotypes: n=13 (in 9 sp present in SDAAB1 and 2 and YFAB), n=18 (in 10 sp from the MB clade), and n=19 (in 16 sp in FFAB, MB, SDAAB1 and YFAB).

Figure 2.2 also displays different levels of chromosome variation in Neotropical clades. Both NC1 and NC2-ii show highly variable chromosome numbers: the most abundant are n=19 (9 sp) and n=28 (29 sp) in NC1, and n=17 (6 sp), n=26 (6 sp) and n=28 (9 sp) the NC2-ii. On the other hand, NC2-i and NC2-iii show more stable values: 48 out of the 49 species with karyotype data in NC2-i have 14 chromosomes, and 16 out of 21 in NC2-iii also have n=14. Although chromosome number variation across Neotropical *Begonia* is highly variable, because two of the most represented numbers in our dataset for this group are n=14 and n=28, we suggest x=14 as the basic number for Neotropical *Begonia* as a whole. However, the clades with highly common n=28 (NC1 and NC2-ii) are of different origin. This means the duplication event from n=14 to n=28 happened at least twice, which suggests that these clades show not only different levels of dysploidy but also at least two independent polyploidisation events. Moreover, since the NC1 and NC2-ii clades also show larger variation, this could indicate that these events were in many cases followed by dysploidy, mainly chromosome loss or genome reduction, which has been studied as a common phenomenon after polyploidisation (Mandáková and Lysak, 2018).

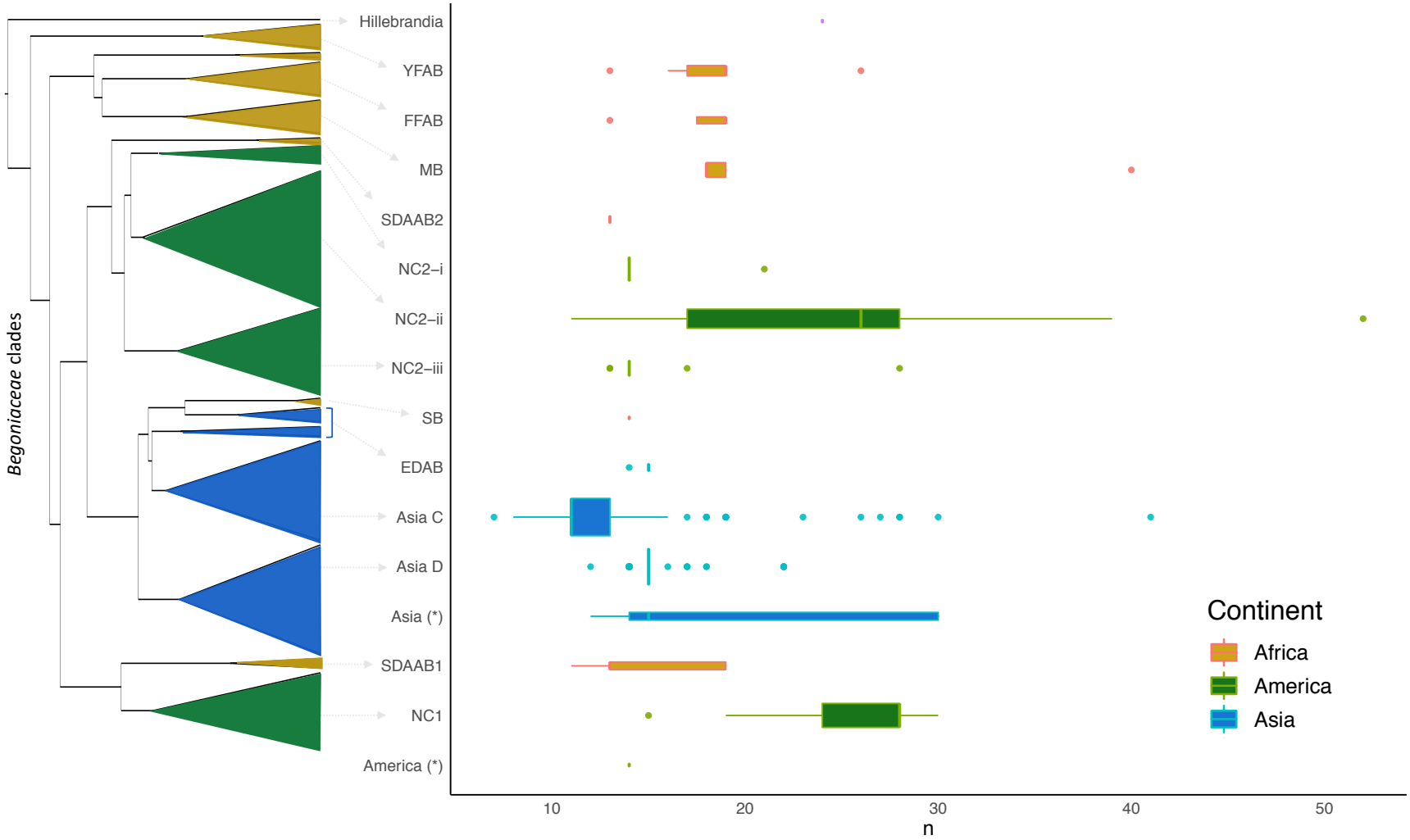


Figure 2.2: Variation of haploid chromosome number across all *Begonia* clades described in Moonlight et al. (2018). The colours of clades in the tree and boxes in the plot indicate the continent where these groups are from. Thickness of the boxes and outlier dots are proportional to sample size.

Regarding Asian clades, the observed variation is higher. These are larger groups, and include some very species-rich, taxonomically complex and non-monophyletic sections (Moonlight et al., 2018). Figure 2.2 shows that although sample sizes are different and there are outliers, these EDAB, Asia C and Asia D have 11 and 15 as their most common haploid chromosome numbers (n=11 in Asia C and n=15 in EDAB and Asia D). There is high variation within Asian clades C and D but 57% and 76% of the species analysed in these respective clades retained these karyotypes. Alternatively, it must be acknowledged that these may be only considered basic number candidates because our dataset is biased by overrepresented sections such as *Coelocentrum* (84 sp) and *Baryandra* (79 sp) in the Asian clade D (which show n=15 in 44 and 23 of their species, respectively), and section *Platycentrum* (196 species) in Asian clade C (n=11 in 54 sp on our dataset). In contrast, other large Asian sections such as *Petermannia* (451 sp) or *Jackia* (60 sp) show much lower sampling with wider variation. This means that although having so much data from large radiations is useful, the large number of cytological data from these sections could be biasing the distributions observed in Figure 2.2. Therefore larger, and more phylogenetically balanced sampling of *Begonia* chromosome numbers will be needed to determine whether 11 and 15 are the basic chromosome numbers of these *Begonia* clades. Moreover, the high number of outliers in these clades (even in sections such as *Platycentrum*) also indicates that dysploidy is frequent in Asian *Begonia*.

The dataset elaborated for this chapter evidences the wide chromosome number distribution across clades. These taxonomic groups are large and complex, and are represented by several *Begonia* radiations defined by specific sections. Therefore, studying the distribution of haploid chromosome numbers by section may help defining better chromosome changes across the evolution of this complex genus.

Are chromosome numbers stable within *Begonia* sections?

Moonlight et al. (2018) established natural sectional classification of *Begonia*. There are to date 70 sections, most of which are represented in the Moonlight et al. (2018) phylogeny and classified into clades. Because the previously discussed *Begonia* clades are significantly large and in most cases it was difficult to find a consensus regarding

basic chromosome numbers, *Begonia* chromosome numbers were also classified into sections to get a more defined view of where the karyotype variation is; whether a few sections have highly unstable karyotypes or many of them just had very different but stable basic numbers. Figure 2.3 shows the variation of chromosome number found across *Begonia* sections linked to their phylogenetic position.

This figure shows that the level of chromosome number variation within *Begonia* sections fluctuates across the genus, but most sections have moderately conserved chromosome numbers with a variable number of outliers, indicating there is a phylogenetic signal at a section level in *Begonia* chromosome variation. It must be considered that to reduce noise and simplify the search for basic chromosome numbers, within-section outlier polyploid species were simplified, and therefore this variation is not represented in Figure 2.3. However, not many within section polyploids of their sectional basic number were detected; only between 1 and 3 karyotypes were simplified in sections *Begonia*, *Doratometra*, *Eupetalum*, *Knesbeckia* and *Platycentrum*, suggesting within-section Whole Genome Duplication (WGD) events are not common. Out of these, only *B.* Sect. *Begonia* shows wide variation in "simplified" chromosome numbers (Figure 2.3), suggesting not only changes in ploidy levels but also high levels of chromosome gain/loss (dysploidy) within this section.

In addition, a few other sections show significantly wide variation in karyotypes, such as B. Sect. *Lepsia* (NC2-ii), B. Sect *Diploclinium* (Asia clade C), B. sect *Jackia* (Asia Clade D) or *Rostrobegonia* (African Clade SSDAB2).

Sections *Lepsia* and *Rostrobegonia* are small sections (9 and 8 species) with cytological data for 2 and 4 species respectively. Species in these sections all have one of two karyotypes: n=30 and n=52 in *Lepsia* and n=13 and n=19 in *Rostrobegonia*, indicating recent chromosomal gain/loss events, but not a simple genome duplication, so basic numbers cannot be determined. There is also the possibility that the unsampled species have very different chromosome numbers and more chromosome numbers from this group would be needed to elucidate the karyotype evolution of these sections.

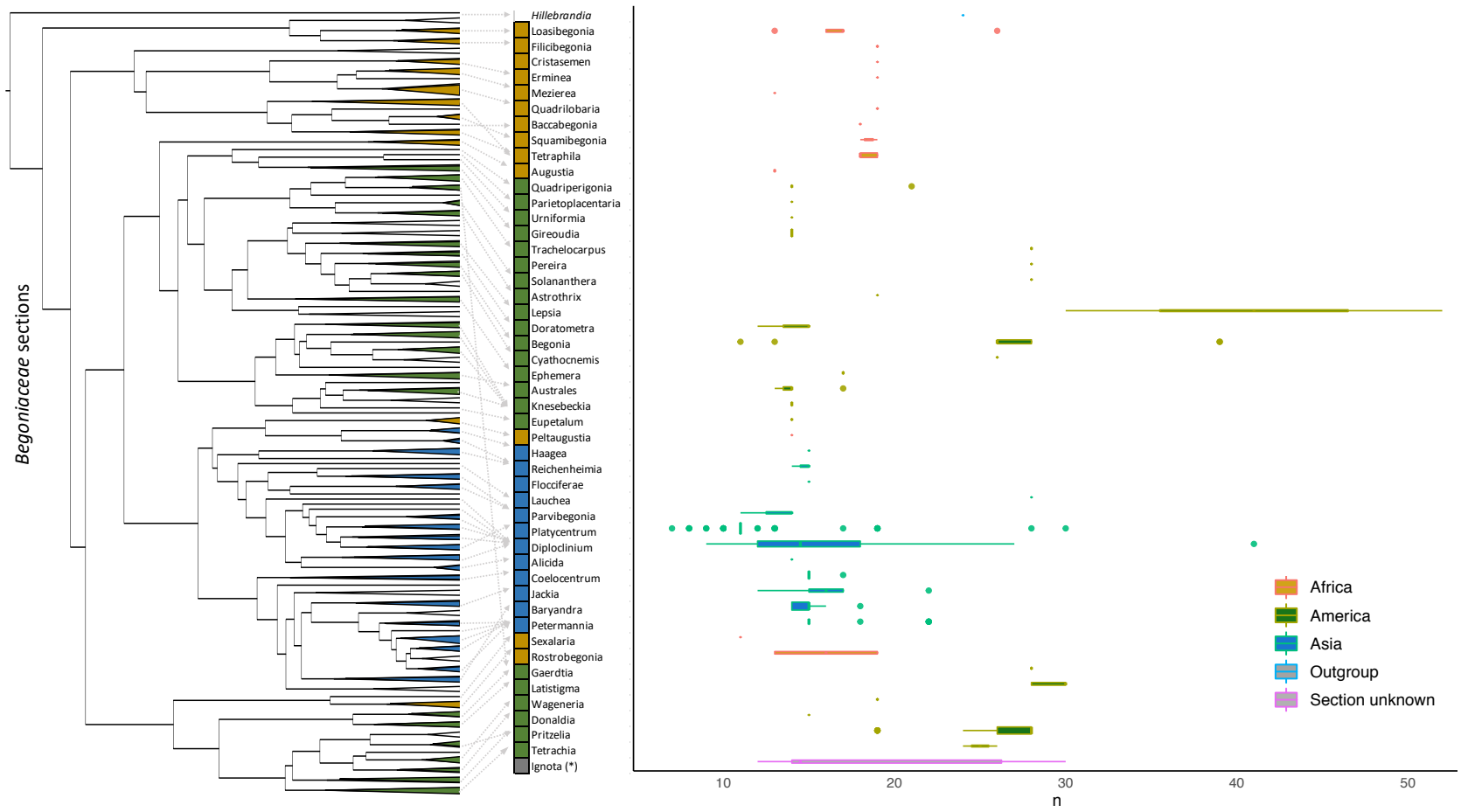


Figure 2.3: Variation of haploid chromosome number across the *Begonia* sections described in Moonlight et al. (2018) with chromosome data. The colours in the tree and boxes in the plot indicate the continent where these groups are from. Species with unknown section are grouped by section *Ignota* and marked with a (*). Thickness of the boxes and outlier dots are proportional to sample size.

In the case of the Asian sections *Jackia* and *Diploclinium*, the variation found is higher. As mentioned earlier, *Jackia* is a section with 60 species of which we only found chromosome data for 8. Karyotype variation oscillates between n=12, 15, 17 and 22. In this case, only 13% of the section has been karyotyped, and in this case additional data on this section would likely help resolve its karyotype evolutionary history. On the other hand, *Diploclinium* is a section with 105 species and we have karyotypes of nearly a quarter (22), varying between n=11, 12, 13, 14, 15, 16, 18, 23, 27 and 41. As it can be observed in Figure 2.3 and according to Moonlight et al. (2018), *Diploclinium* is a polyphyletic section that requires additional morphological and phylogenetic work to recircumscribe the section into natural groups. This implies that *Diploclinium* is a rather "unnatural" section with an undefined single origin, and future taxonomic work could split it into smaller groups or re-assign the species initially assigned to *B.* sect. *Diploclinium* in Moonlight et al. (2018) to other sections. Chromosome data gathered in this chapter supports the view that although the species classified into the *Diploclinium* clade are morphologically similar, this group is not a "natural" taxon. Alternatively this section could have abnormally high rates of chromosomal change. Both a better-resolved phylogeny and more chromosome counts would be needed in this case to help resolve the evolutionary trajectory of chromosome numbers in this group. Something to take into account is that *Diploclinium* is not the only polyphyletic section in Figure 2.3, but it is the one with most striking karyotype variation. In the case of other paraphyletic *Begonia* sections such as *Knesbeckia* (already divided into 3 different groups by Moonlight et al. (2018)), *Petermania* or *Pritzelia*, chromosome numbers may or not be completely conserved but they oscillate between a small range. Although these sections would also require further phylogenetic studies to investigate their karyotype evolution, our chromosome data indicates that out of the most taxonomically complex *Begonia* sections, *Diploclinium* is the one showing the least "natural" or "parsimonious" karyotype variation.

One of the main hypotheses to test in this chapter was whether karyotypic changes, as a form of genome dynamics, are correlated with higher speciation rates. Figure 2.4 represents the relationship between the standard deviation among haploid chromosome numbers and species number within sections. With the data presented in this chapter

(22% of the genus species), this graph shows no simple correlation (R^2=0.04, Pearson p-value 0.82) between species number and chromosome variation (represented by the standard deviation of haploid chromosome numbers) in *Begonia* sections. On the contrary, it indicates that there are sections with low number of species and very high karyotype variation (e.g. *Lepsia* or *Begonia*). Further cytological studies on sections with wider chromosome number variation with larger sampling may also help understanding the nature of such karyotype dynamics in these groups. On the other hand, other sections with over 100 species show more stable chromosome numbers (e.g. *Gireoudia*). This suggests that the correlation between speciation and genome size variation studied by Puttick et al. (2015) does not only depend on the genome size variation caused by chromosome number changes, but also by chromosome size variation.

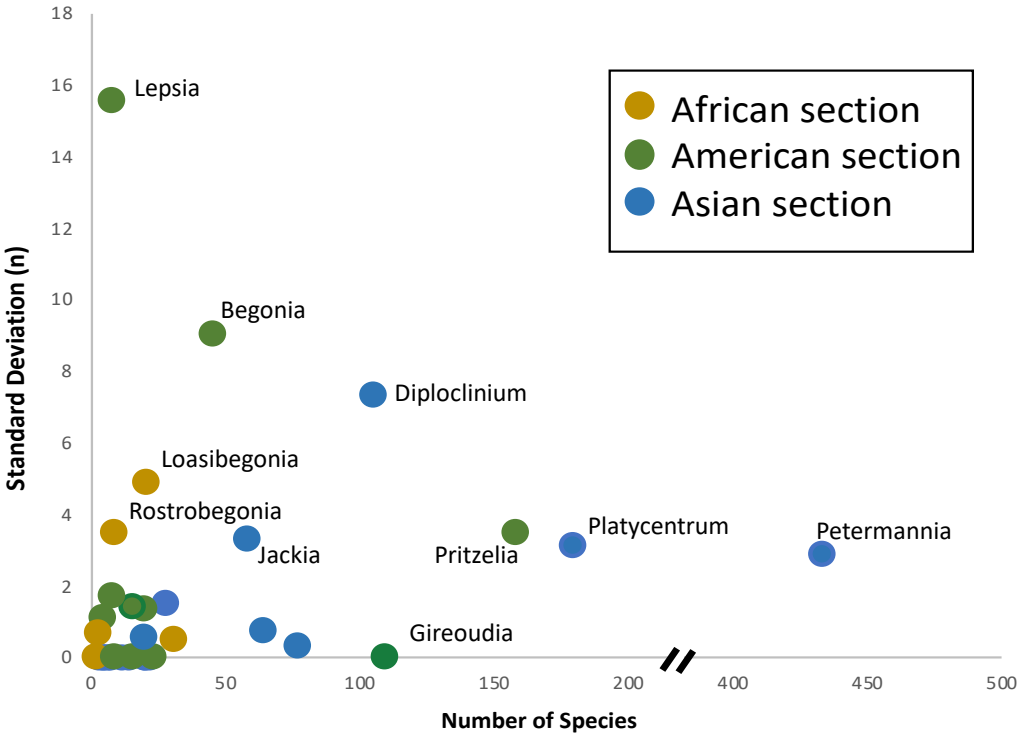


Figure 2.4: Scatterplot representing the standard deviation of the haploid chromosome number between species of the same section versus the number of species of those sections. Dots (sections) with a standard deviation over 2 or species number over 100 are labelled. Colours indicate the continent where each section is from.

Some of the largest sections such as *Pritzelia*, *Platycentrum* and *Petermannia* show moderate karyotype variation (standard deviation >4). Sections *Diploclinium* and *Gireoudia*, also with over 100 species, show very different karyotype change patterns.

While all studied *Gireoudia* species show the very same chromosome numbers (n=14), *Diploclinium* is, as above discussed, the *Begonia* section with the widest range of chromosome numbers. However, the polyphyly of this section means there is no evolutionary connotation of this variation until the species relationships in this group are better defined. Our data incompleteness (as we only have chromosome information of 22% of all *Begonia* species), the different ages and levels of taxonomic resolution/polyphyly of sections, and the specific chromosome evolution rates of each group make the above question difficult to answer. Our data rather suggest that karyotype stability is highly variable among all and within the largest *Begonia* clades, and more conserved within sections, but still diverse across them, especially in Neotropical and Asian groups. We did not find any support for a correlation between highly specious sections and higher chromosome instability, although sampling biases must be considered and acknowledged. A better overview and estimation of the chromosomal changes across the evolutionary history of *Begonia* could have been obtained by mapping the species with data directly in the Moonlight et al. (2018) phylogeny. However, due to the incongruence between the species on this phylogeny and the species we gathered data from, it was impossible to place all our data in a species level *Begonia* tree.

2.2.2 Genome size variation across *Begonia*

Data compilation

For this chapter, genome size information was also gathered from various sources (Plant C-value database, Pellicer and Leitch (2020); Du et al. (2018) and Campos and Pellicer, unpublished), and as done with cytological data, the species were classified into section and clade, and paired with their chromosome count if available (full dataset is in appendix 2.A.2). In total, the genome sizes of 71 species were gathered, including 70 *Begonia* species and *Hillebrandia*. This includes four African species (representing three sections of clades SDAAB2, SDAAB1 and SB), 32 American (11 sections, all clades represented) and 33 Asian species (four sections, Asia C and Asia D clades).

As previously described by Dewitte et al. (2009), high levels of variation were found in *Begonia* genome size data. C-values range between 245 Mbp in *Begonia ulmifolia* (n=15 section *Donaldia*, NC1) and 2,497 Mbp *Begonia fomosana* (n=30, section *Platycentrum*,

Asia clade C). This represents only a 2 fold difference in chromosome number but more than a 10 fold difference in genome size, suggesting this difference in DNA content could be triggered by high transpositional activity and repetitive DNA dynamics in *Begonia*. Here, I will explore the genome size differences across *Begonia* clades as well as whether this variation is essentially driven by chromosome gain and loss (polyploidisation or dysploidisation), or other genetic elements such as DNA repeats are playing a role in genome size variation. Because of the low number of *Begonia* genome sizes found, only data distribution by clade will be graphically represented.

How does genome size vary across *Begonia* clades?

Figure 2.5 shows the variation in haploid genome size (Mbp) across all *Begonia* clades. Wide variation in genome size across clades was expected, since chromosome numbers are also variable in a wide range at this phylogenetic level. As mentioned above, the obtained dataset includes a very small amount of African species, and therefore not much variation can be observed. African genomes have sizes between 360 and 630 Mbp approximately, which is larger than the *Hillebrandia* genome (332.6 Mbp). Additional genome size data across African *Begonia* clades would give a better overview of the genome size evolution and stability in the early *Begonia* clades.

Neotropical and Asian radiations show wider genome sizes. As in Figure 2.2, NC2-ii and Asian clades show the highest level of variation. It must be considered variation observed is, however, biased by a uneven sampling throughout the genus: sections such as *Platycentrum* are overrepresented and therefore the variation observed in Asian clade C is wider. This clade shows a range of C-values from 367 Mbp to 2,497 Mbp. This clade has the highest representation high representation in our dataset (33% of the species included belong to this clade), and therefore the variation observed here could also be present in other clades and more C-value data is needed across the genus. However, because karyotype variation is also high in the *Begonia* Asian clade C, we can confirm this group undergoes high levels of genome structural variation, chromosomal gain and loss, polyploidisation, etc.

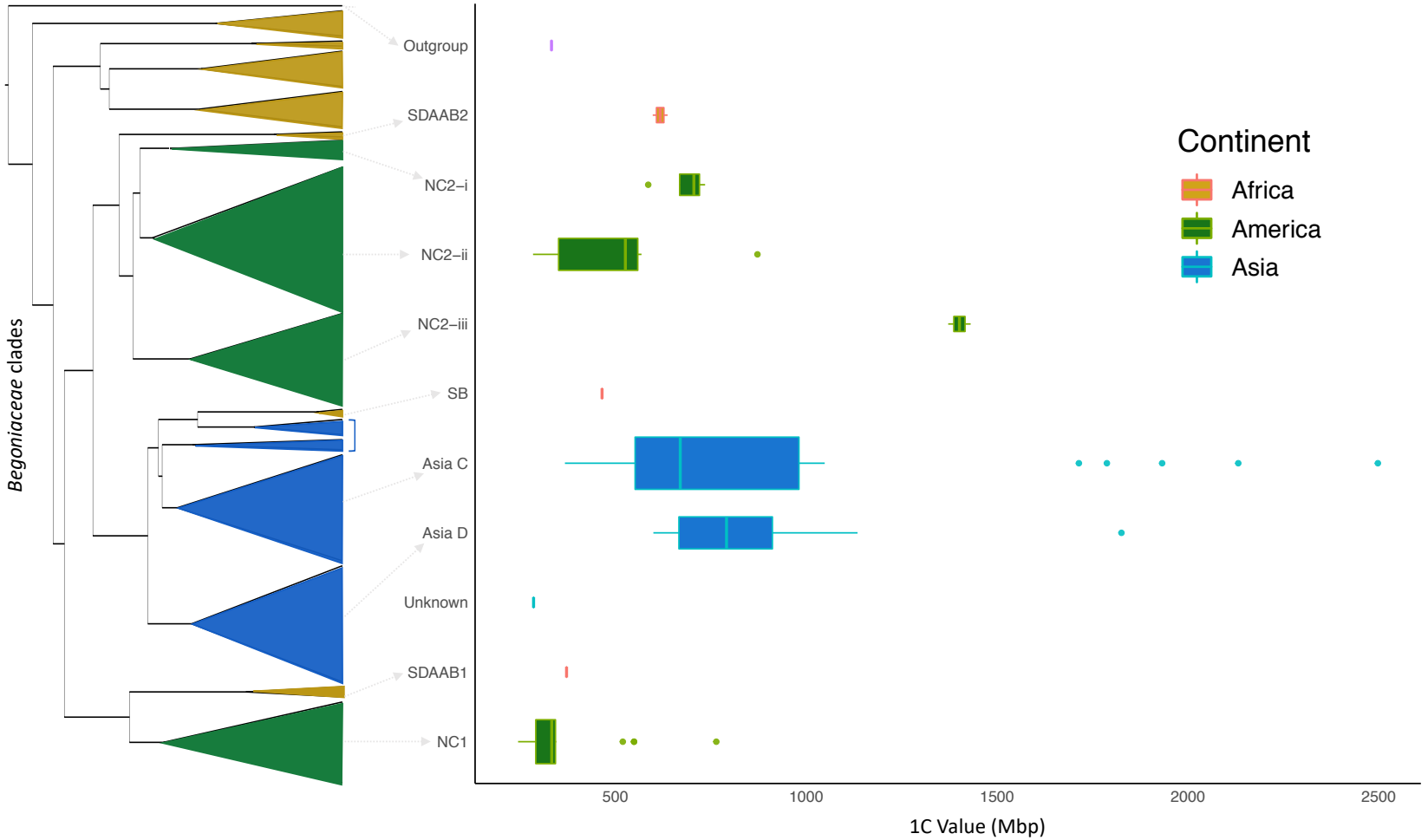


Figure 2.5: Variation of haploid genome size (1C, Mbp) across all *Begonia* clades described in Moonlight et al. (2018) and Hughes et al. (2015). The colours of clades in the tree and boxes in the plot indicate the continent where these groups are from.

Figure 2.5 also shows that in general, Neotropical species seem to smaller genomes than Asian species, with the exception of species in clade NC2-iii. Genome sizes within Neotropical clades also seem more conserved than within Asian clades, but between-clade variation is much higher. This is, there are barely any overlaps between the genome sizes of NC1, NC2i, ii and iii, even in those clades with the same basic chromosome numbers (NC2-1 and NC2-iii). This means there is some phylogenetic signal in genome size in Neotropical radiations, but genome size variation changes across the phylogeny do not overlap with karyotype changes, and other events driving chromosome size variation must play a role in the genome evolution of these species. Asian clades, however, show larger and overlapping ranges of DNA content, as well as chromosome numbers (Figure 2.2). This suggests that large scale genome changes were common early in the Neotropical radiations but have been rarer since, and that across all the Asian radiation there have been more recent changes in their genome structure. However, genome size variations found at smaller scale within groups with stable chromosome numbers also indicates levels of transposition and genome dynamics in Neotropical *Begonia*.

Is there a correlation with chromosome numbers?

To test whether the variation found in our genome size data is caused mostly by karyotypic changes, the relationship between genome sizes and chromosome numbers is plotted in Figure 2.6. Results show that, with the data published to date, there is no correlation (R^2=0.02, p-value 0.87) between chromosome number and genome size in *Begonia*. It is easily appreciated, however, that sections with conserved chromosome numbers such as *Platycentrum* (196 species), *Coelocentrum* (84 sp), *Ephemera* (14 sp) , *Gireoudia* (110 sp), *Petermannia* (451 sp) and *Pritzelia* (158 sp) show variable genome sizes. This means that it is highly likely that these genome size differences are caused by variation in repetitive elements. Most of these setions have the largest number of species and represent recent and rapid *Begonia* species radiations (Moonlight et al., 2018). Therefore, they are a good model group to test the influence of repetitive DNA and Transposable Elements (TEs) on *Begonia* speciation events. In-depth genomic studies are needed to address this question and will be carried out in the following

chapters.

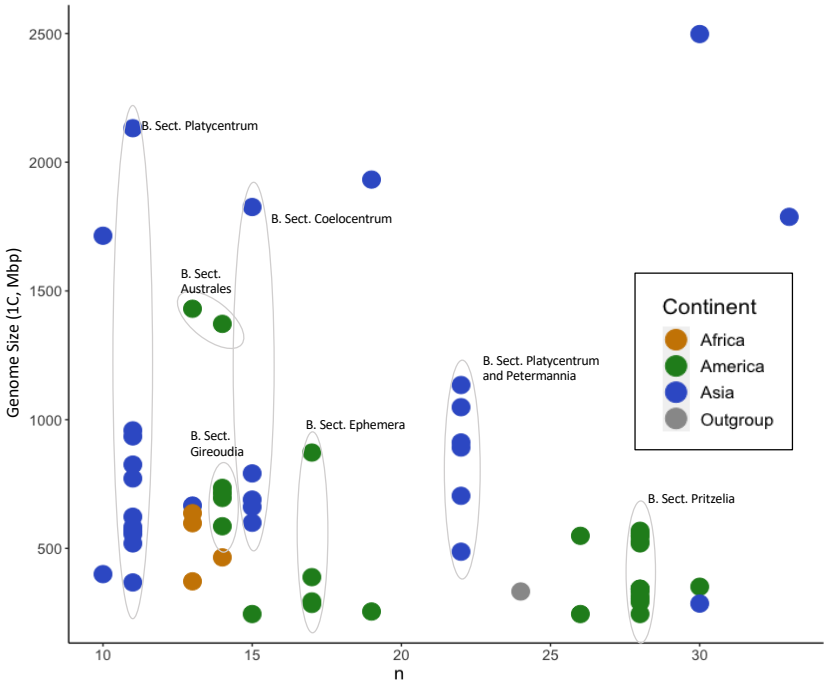


Figure 2.6: Scatterplot representing the genome sizes and chromosome numbers of all the *Begonia*species found in literature for both features (64 *Begonia* and *Hillebrandia*). Species in the same sections are labelled with a circle. Colour code indicates the continent where each species is from, and *Hillebrandia* is labelled as "outgroup".

2.3 Material and Methods

2.3.1 Data collection, curation and analysis

Chromosome number and genome size data were obtained from different sources (see Appendix 2.A.1 and 2.A.2). Species names were curated and updated using the RBGE Begonia Resource Centre (Hughes et al., 2015). Unnamed species (listed as *Begonia sp.*), horticultural hybrids and those with several chromosome numbers associated and with no consensus were removed. Then, data was classified by *Begonia* section and clade, according to the latest phylogeny (Moonlight et al., 2018). Within sections, chromosome numbers were reduced to their greatest common divider to avoid variation caused by polyploids. For species with different C-values associated, an average value was used.

2.3.2 Genome size estimations

Fresh leaf tissue from the Begoniaceae species in the RBGE collection selected for WGS in chapters 3 and 6 (table 1) was collected and taken to the Jodrell Laboratories at the Royal Botanic Gardens, Kew. There, nuclear DNA contents (C-values) were measured following the two-step protocol described in Pellicer and Leitch (2014). Intact leaf tissues of a sample and reference standard (*Solanum lycopersicum* in all cases except with *B. bipinnatifida*, where *Petroselinum crispum* was used) were chopped together in a Petri dish containing 500 μl of ice-cold Nuclei Extraction Buffer (CyStain UV precise P, Partec) using a sharp razor blade. The crude suspension was filtered through a nylon mesh filter (30 m pore size, CellTrics Disposable Filters, Partec). After a 5-min incubation at room temperature, 2ml μl of Staining Buffer (CyStain UV precise P, Partec) were added. Samples were then analysed using a CyFlowSL Partec flow cytometer (Partec GmbH, Goettingen, Germany) fitted with a 100mW green lamp (532 nm solid-state Cobalt Samba laser; Cobolt AB, Solna, Sweden). The resulting flow histograms were analysed using the Partec software for flow cytometry FLOMAX 2.7. At least 5,000 nuclei were analysed per sample. Three tissue samples of each species were analysed and each one was measured three times (9 measurements in total per individual plant sample). The reference standards used in this study have the following 2C-values: *Solanum lycopersicum* = 2.0 pg (Praça-Fontes et al., 2011) and *Petroselinum crispum* 'Champion Moss Curled' cultivar = 4.5 pg DNA (Obermayer et al., 2002). Nuclear DNA content was estimated using the formula:

$$2C_S = \frac{G1_S * 2C_{ST}}{G1_{ST}}$$

Where: $2C_S$ = Sample 2C nuclear DNA content [pg] ; $G1_S$ = Sample G1 peak mean ; $2C_{ST}$ = Standard 2C nuclear DNA content [pg] ; $G1_{ST}$ = Standard G1 peak mean

DNA amounts in picograms can be converted to the number of base pairs using the conversion factor 1 pg DNA= 978 Mbp (Dolezel, 2003).

Bibliography

Ali MS. 2013. Genetic architecture of species level differences in Begonia Section Gireoudia. Ph.D. thesis, University of Edinburgh.

Baack E, Melo MC, Rieseberg LH, Ortiz-Barrientos D. 2015. The origins of reproductive isolation in plants. *New Phytologist*. 207:968–984.

Blommaert J. 2020. Genome size evolution: towards new model systems for old questions. *Proceedings of the Royal Society B*. 287:20201441.

Boutanaev A, Nemchinov L. 2020. Genome size dynamics within multiple genera of diploid seed plants. *Russian Journal of Genetics*. 56:684–692.

Burt-Utley K. 1985. A revision of central american species of begonia section gireoudia (begoniaceae). *Tulane studies in zoology and botany*. 25:1–131.

Carta A, Bedini G, Peruzzi L. 2020. A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytologist*. 228:1097–1106.

Cremonini R. 2005. Low chromosome number angiosperms. *Caryologia*. 58:403–409.

Dewitte A. 2010. Exploitation of 2n pollen to create genetic variation in the genus Begonia. Ph.D. thesis, Ghent University.

Dewitte A, Leus L, Eeckhaut T, Vanstechelman I, Van Huylenbroeck J, Van Bockstaele E. 2009. Genome size variation in begonia. *Genome*. 52:829–838.

Dewitte A, Twyford A, Thomas D, Kidner C, Van Huylenbroeck J. 2011. The origin of diversity in begonia: genome dynamism, population processes and phylogenetic patterns. In: Grillo O, Venora G, editors, The dynamical processes of biodiversity-case studies of evolution and spatial distribution, InTech Open, volume 448, chapter 2, pp. 29–51.

Dolezel J. 2003. Nuclear dna content and genome size of trout and human. *Cytometry Part A*. 51:127–128.

Doyle JJ, Coate JE. 2019. Polyploidy, the nucleotype, and novelty: the impact of genome doubling on the biology of the cell. *International Journal of Plant Sciences*. 180:1–52.

Du WW, Wang XN, Duan Q, Wen-Jie J, Ma LL, Cui G, Wang J. 2018. Genome size comparison in 34 begonia l. species. *Journal of Plant Genetic Resources*. 19:370–378.

Escudero M, Martín-Bravo S, Mayrose I, et al. (11 co-authors). 2014. Karyotypic changes through dysploidy persist longer over evolutionary time than polyploid changes. *PLoS One*. 9:e85266.

Forrest L. 2000. Phylogeny of Begoniaceae. Ph.D. thesis, University of Glasgow.

Grant V. 1981. Plant speciation. Columbia University Press.

Han TS, Zheng QJ, Onstein RE, Rojas-Andrés BM, Hauenschild F, Muellner-Riehl AN, Xing YW. 2020. Polyploidy promotes species diversification of allium through ecological shifts. *New Phytologist*. 225:571–583.

Hong-Zhe L, Hong M, Kai-Yun G, Peng CI. 2005. Begonia rubinea (sect. platycentrum, begoniaceae), a new species from guizhou, china. *Botanical Bulletin of Academia Sinica*. 46.

Hughes M, Moonlight P, Jara A, Tebbitt M, Wilson H, Pullan M. 2015. Begonia resource centre. online database available from http://padme.rbge.org.uk/begonia/.

Hughes M, Rubite RR, Kono Y, Peng CI, et al. (5 co-authors). 2011. Begonia blancii (sect. diploclinium, begoniaceae), a new species endemic to the philippine island of palawan. *Bot Stud*. 52:203–9.

Husband BC, Sabara HA. 2004. Reproductive isolation between autotetraploids and their diploid progenitors in fireweed, chamerion angustifolium (onagraceae). *New Phytologist*. 161:703–713.

Jiao Y, Wickett NJ, Ayyampalayam S, et al. (11 co-authors). 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*. 473:97–100.

Knight CA, Molinari NA, Petrov DA. 2005. The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany*. 95:177–190.

Legro R, Doorenbos J. 1969. Chromosome numbers in begonia. *Netherlands Journal of Agricultural Science*. 17:189–202.

Legro RAH, Haegeman J. 1971. Chromosome numbers of hybrid tuberous begonias. *Euphytica*. 20:1–13.

Leitch I, Kahandawala I, Suda J, Hanson L, Ingrouille MJ, Chase M, Fay M. 2009. Genome size diversity in orchids: consequences and evolution. *Annals of botany*. 104:469–481.

Leitch IJ, Beaulieu J, Cheung K, Hanson L, Lysak M, Fay M. 2007. Punctuated genome size evolution in liliaceae. *Journal of Evolutionary Biology*. 20:2296–2308.

Li Z, Baniaga AE, Sessa EB, Scascitelli M, Graham SW, Rieseberg LH, Barker MS. 2015. Early genome duplications in conifers and other seed plants. *Science advances*. 1:e1501084.

Liu Y, Tseng YH, Yang HA, Hu AQ, Xu WB, Lin CW, Kono Y, Chang CC, Peng CI, Chung KF. 2020. Six new species of begonia from guangxi, china. *Botanical studies*. 61:1–23.

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol*. 8:e1000500.

Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. 2009. The dynamic ups and downs of genome size evolution in brassicaceae. *Molecular Biology and Evolution*. 26:85–98.

Mandáková T, Lysak MA. 2018. Post-polyploid diploidization and diversification through dysploid changes. *Current opinion in plant biology*. 42:55–65.

Moonlight PW, Ardi WH, Padilla LA, et al. (29 co-authors). 2018. Dividing and conquering the fastest-growing genus: Towards a natural sectional classification of the mega-diverse genus begonia (Begoniaceae). *Taxon*. 67:267–323.

Nguyen HQ. 2004. Systematic revision of the indigenous Begonia sect. Platycentrum in Vietnam. Ph.D. thesis, University of Missouri-Saint Louis.

Obermayer R, Leitch IJ, Hanson L, Bennet MD. 2002. Nuclear dna c-values in 30 species double the familial representation in pteridophytes. *Annals of Botany*. 90:209–217.

Oginuma K, Peng CI. 2002. Karyomorphology of taiwanese begonia (begoniaceae): taxonomic implications. *Journal of Plant Research*. 115:0225–0235.

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ. 2018. Genome size diversity and its impact on the evolution of land plants. *Genes*. 9:88.

Pellicer J, Kelly L, Magdalena C, Leitch I. 2013. Insights into the dynamics of genome size and chromosome evolution in the early diverging angiosperm lineage nymphaeales (water lilies). *Genome*. 56:437–449.

Pellicer J, Kelly LJ, Leitch IJ, Zomlefer WB, Fay MF. 2014. A universe of dwarfs and giants: genome size and chromosome evolution in the monocot family m elanthiaceae. *New Phytologist*. 201:1484–1497.

Pellicer J, Leitch IJ. 2014. The Application of Flow Cytometry for Estimating Genome Size and Ploidy Level in Plants. In: Molecular Plant Taxonomy: Methods and Protocols, Methods in Molecular Biology, volume 1115, pp. 279–307.

Pellicer J, Leitch IJ. 2020. The plant dna c-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist*. 226:301–305.

Peng CI, Rubite RR, Lin CW, Hughes M, Kono Y, Chung KF. 2017. Three new species of begonia sect. baryandra from panay island, philippines. *Botanical studies*. 58:28.

Peng CI, Wang H, Kono Y, Yang HA. 2014. Begonia wui-senioris (sect. platycentrum, begoniaceae), a new species from myanmar. *Botanical studies*. 55:1–6.

Perry BA. 1943. Chromosome number and phylogenetic relationships in the euphorbiaceae. *American Journal of Botany*. pp. 527–543.

Praça-Fontes MM, Carvalho CR, Clarindo WR, Cruz CD. 2011. Revisiting the dna c-values of the genome size-standards used in plant flow cytometry to choose the "best primary standards". *Plant Cell Reports*. 30:1183–1191.

Price HJ, Dillon SL, Hodnett G, Rooney WL, Ross L, Johnston JS. 2005. Genome evolution in the genus sorghum (poaceae). *Annals of Botany*. 95:219–227.

Puttick MN, Clark J, Donoghue PC. 2015. Size is not everything: rates of genome size evolution, not c-value, correlate with speciation in angiosperms. *Proceedings of the Royal Society B: Biological Sciences*. 282:20152289.

Rice A, Glick L, Abadi S, Einhorn M, Kopelman NM, Salman-Minkov A, Mayzel J, Chay O, Mayrose I. 2015. The chromosome counts database (ccdb)–a community resource of plant chromosome numbers. *New Phytologist*. 206:19–26.

Schubert I, Vu GT. 2016. Genome stability and evolution: attempting a holistic view. *Trends in plant science*. 21:749–757.

Sheidai M, Ghazei M, Pakravan M. 2010. Contribution to cytology of the genus euphorbia in iran. *Cytologia*. 75:477–482.

Stebbins GL, et al. (2 co-authors). 1971. Chromosomal evolution in higher plants. Edward Arnold Ltd., London.

Tebbitt MC, Dickson JH. 2000. Amended descriptions and revised sectional assignment of some asian begonias (begoniaceae). *Brittonia*. 52:112–117.

Tebbitt MC, Lowe-Forrest L, Santoriello A, Clement WL, Swensen SM. 2006. Phylogenetic relationships of asian begonia, with an emphasis on the evolution of rain-ballist and animal dispersal mechanisms in sections platycentrum, sphenanthera and leprosae. *Systematic botany*. 31:327–336.

Tseng YH, Kim YD, Peng CI, Htwe KM, Cho SH, Kono Y, Chung KF. 2017. Begonia myanmarica (begoniaceae), a new species from myanmar, and molecular phylogenetics of begonia sect. monopteron. *Botanical studies*. 58:1–9.

Twyford AD, Ennos RA, Kidner CA. 2013. Development and characterization of microsatellite markers for central american begonia sect. gireoudia (begoniaceae). *Applications in Plant Sciences*. 1:1200499.

Uhl CH. 1978. Chromosomes of mexican sedum ii. section pachysedum. *Rhodora*. 80:491–512.

Vanzela AL, Guerra M, Luceno M. 1996. Rhynchospora tenuis link (cyperaceae), a species with the lowest number of holocentric chromosomes. *Cytobios*. 88:219–228.

Winterfeld G, Ley A, Hoffmann MH, Paule J, Röser M. 2020. Dysploidy and polyploidy trigger strong variation of chromosome numbers in the prayer-plant family (marantaceae). *Plant Systematics and Evolution*. 306:1–17.

Winterfeld G, Schneider J, Perner K, Röser M. 2014. Polyploidy and hybridization as main factors of speciation: complex reticulate evolution within the grass genus helictochloa. *Cytogenetic and Genome Research*. 142:204–225.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the national Academy of sciences*. 106:13875–13879.

2.A Appendix

2.A.1 Chromosome numbers of *Begonia*

Species	Section	Clade	n	Source
B. bogneri	*Erminea*	Africa (FFAB)	19	CCBD
B. seychellensis	*Mezierea*	Africa (FFAB)	13	CCBD
B. francoisii	*Quadrilobaria*	Africa (FFAB)	19	Forrest, 2000
B. goudotii	*Quadrilobaria*	Africa (FFAB)	19	Kono and Peng, unpub.
B. baccata	*Baccabegonia*	Africa (MB)	18	Forrest, 2000
B. crateris	*Baccabegonia*	Africa (MB)	18	Forrest, 2000
B. poculifera	*Squamibegonia*	Africa (MB)	19	CCBD
B. ampla	*Squamibegonia*	Africa (MB)	40	Kono and Peng, unpub.
B. elaeagnifolia	*Tetraphila*	Africa (MB)	18	CCBD
B. gabonensis	*Tetraphila*	Africa (MB)	18	Kono and Peng, unpub.
B. horticola	*Tetraphila*	Africa (MB)	18	CCBD
B. karperi	*Tetraphila*	Africa (MB)	18	CCBD
B. longipetiolata	*Tetraphila*	Africa (MB)	18	CCBD
B. oxyanthera	*Tetraphila*	Africa (MB)	18	CCBD
B. polygonoides	*Tetraphila*	Africa (MB)	18	CCBD
B. subscutata	*Tetraphila*	Africa (MB)	18	Kono and Peng, unpub.
B. eminii	*Tetraphila*	Africa (MB)	19	CCBD
B. komoensis	*Tetraphila*	Africa (MB)	19	Kono and Peng, unpub.
B. loranthoides	*Tetraphila*	Africa (MB)	19	CCBD
B. mannii	*Tetraphila*	Africa (MB)	19	CCBD
B. molleri	*Tetraphila*	Africa (MB)	19	CCBD
B. squamulosa	*Tetraphila*	Africa (MB)	19	CCBD
B. socotrana	*Peltaugustia*	Africa (SB)	14	CCBD
B. engleri	*Rostrobegonia*	Africa (SDAAB1)	13	CCBD
B. johnstonii	*Rostrobegonia*	Africa (SDAAB1)	13	CCBD
B. rostrata	*Rostrobegonia*	Africa (SDAAB1)	19	CCBD
B. wollastonii	*Rostrobegonia*	Africa (SDAAB1)	19	CCBD
B. annobonensis	*Sexalaria*	Africa (SDAAB1)	11	Forrest, 2000
B. dregei	*Augustia*	Africa (SDAAB2)	13	CCBD
B. homonyma	*Augustia*	Africa (SDAAB2)	13	CCBD
B. princeae	*Augustia*	Africa (SDAAB2)	13	CCBD
B. sutherlandii	*Augustia*	Africa (SDAAB2)	13	CCBD
B. wakefieldii	*Augustia*	Africa (SDAAB2)	13	Kono and Peng, unpub.

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. thomeana	*Cristasemen*	Africa (YFAB)	19	Forrest, 2000
B. elatostemmoides	*Filicibegonia*	Africa (YFAB)	19	CCBD
B. sciaphila	*Filicibegonia*	Africa (YFAB)	19	Forrest, 2000
B. sessilifolia	*Filicibegonia*	Africa (YFAB)	19	CCBD
B. potamophila	*Loasibegonia*	Africa (YFAB)	13	Forrest, 2000
B. prismatocarpa	*Loasibegonia*	Africa (YFAB)	16	CCBD
B. microsperma	*Loasibegonia*	Africa (YFAB)	17	CCBD
B. staudtii	*Loasibegonia*	Africa (YFAB)	17	CCBD
B. quadrialata	*Loasibegonia*	Africa (YFAB)	26	Forrest, 2000
B. froebelii	*Ignota (*)*	America (*)	14	CCBD
B. ulmifolia	*Donaldia*	America (NC1)	15	CCBD
B. coccinea	*Gaerdtia*	America (NC1)	28	CCBD
B. dichroa	*Gaerdtia*	America (NC1)	28	Forrest, 2000
B. maculata	*Gaerdtia*	America (NC1)	28	CCBD
B. undulata	*Gaerdtia*	America (NC1)	28	CCBD
B. olbia	*Latistigma*	America (NC1)	28	CCBD
B. teuscheri	*Latistigma*	America (NC1)	28	CCBD
B. aconitifolia	*Latistigma*	America (NC1)	30	CCBD
B. leathermanniae	*Latistigma*	America (NC1)	30	Forrest, 2000
B. platanifolia	*Latistigma*	America (NC1)	30	CCBD
B. acetosa	*Pritzelia*	America (NC1)	19	CCBD
B. acida	*Pritzelia*	America (NC1)	19	CCBD
B. itaguassuensis	*Pritzelia*	America (NC1)	19	CCBD
B. paleata	*Pritzelia*	America (NC1)	19	CCBD
B. reniformis	*Pritzelia*	America (NC1)	19	CCBD
B. valida	*Pritzelia*	America (NC1)	19	CCBD
B. princeps	*Pritzelia*	America (NC1)	24	CCBD
B. scabrida	*Pritzelia*	America (NC1)	24	CCBD
B. paulensis	*Pritzelia*	America (NC1)	26	CCBD
B. venosa	*Pritzelia*	America (NC1)	26	CCBD
B. angularis	*Pritzelia*	America (NC1)	27	CCBD
B. arborescens	*Pritzelia*	America (NC1)	28	CCBD
B. capanemae	*Pritzelia*	America (NC1)	28	CCBD
B. caraguatatubensis	*Pritzelia*	America (NC1)	28	CCBD
B. dietrichiana	*Pritzelia*	America (NC1)	28	CCBD
B. echinosepala	*Pritzelia*	America (NC1)	28	CCBD

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. epipsila	*Pritzelia*	America (NC1)	28	CCBD
B. hispida	*Pritzelia*	America (NC1)	28	CCBD
B. hookeriana	*Pritzelia*	America (NC1)	28	CCBD
B. hugelii	*Pritzelia*	America (NC1)	28	CCBD
B. incisoserrata	*Pritzelia*	America (NC1)	28	CCBD
B. listada	*Pritzelia*	America (NC1)	28	Forrest, 2000
B. luxurians	*Pritzelia*	America (NC1)	28	CCBD
B. metallica	*Pritzelia*	America (NC1)	28	Forrest, 2000
B. odeteiantha	*Pritzelia*	America (NC1)	28	CCBD
B. olsoniae	*Pritzelia*	America (NC1)	28	CCBD
B. oxyphylla	*Pritzelia*	America (NC1)	28	Forrest, 2000
B. paranaensis	*Pritzelia*	America (NC1)	28	CCBD
B. pulchella	*Pritzelia*	America (NC1)	28	CCBD
B. rufoserica	*Pritzelia*	America (NC1)	28	CCBD
B. sanguinea	*Pritzelia*	America (NC1)	28	CCBD
B. scharffii	*Pritzelia*	America (NC1)	28	CCBD
B. stenophylla	*Pritzelia*	America (NC1)	28	CCBD
B. tomentosa	*Pritzelia*	America (NC1)	28	Forrest, 2000
B. kuhlmannii	*Tetrachia*	America (NC1)	24	CCBD
B. egregia	*Tetrachia*	America (NC1)	26	CCBD
B. convolvulacea	*Wageneria*	America (NC1)	19	CCBD
B. fagifolia	*Wageneria*	America (NC1)	19	CCBD
B. glabra	*Wageneria*	America (NC1)	19	CCBD
B. alice-clarkiae	*Gireoudia*	America (NC2-i)	14	Forrest, 2000
B. aridicaulis	*Gireoudia*	America (NC2-i)	14	CCBD
B. barkeri	*Gireoudia*	America (NC2-i)	14	CCBD
B. bettinae	*Gireoudia*	America (NC2-i)	14	Forrest, 2000
B. bowerae	*Gireoudia*	America (NC2-i)	14	CCBD
B. carolineifolia	*Gireoudia*	America (NC2-i)	14	Forrest, 2000
B. chivatoa	*Gireoudia*	America (NC2-i)	14	CCBD
B. conchifolia	*Gireoudia*	America (NC2-i)	14	CCBD
B. crassicaulis	*Gireoudia*	America (NC2-i)	14	CCBD
B. fusca	*Gireoudia*	America (NC2-i)	14	CCBD
B. heracleifolia	*Gireoudia*	America (NC2-i)	14	CCBD
B. hispidi-villosa	*Gireoudia*	America (NC2-i)	14	CCBD
B. hydrocotylifolia	*Gireoudia*	America (NC2-i)	14	Forrest, 2000

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. imperialis	*Gireoudia*	America (NC2-i)	14	CCBD
B. involucrata	*Gireoudia*	America (NC2-i)	14	CCBD
B. kenworthyae	*Gireoudia*	America (NC2-i)	14	CCBD
B. kortsiae	*Gireoudia*	America (NC2-i)	14	CCBD
B. lindleyana	*Gireoudia*	America (NC2-i)	14	CCBD
B. ludicra	*Gireoudia*	America (NC2-i)	14	CCBD
B. manicata	*Gireoudia*	America (NC2-i)	14	CCBD
B. mazae	*Gireoudia*	America (NC2-i)	14	CCBD
B. mexicana	*Gireoudia*	America (NC2-i)	14	CCBD
B. nelumbiifolia	*Gireoudia*	America (NC2-i)	14	CCBD
B. peltata	*Gireoudia*	America (NC2-i)	14	CCBD
B. philodendroides	*Gireoudia*	America (NC2-i)	14	CCBD
B. pinetorum	*Gireoudia*	America (NC2-i)	14	CCBD
B. plebeja	*Gireoudia*	America (NC2-i)	14	CCBD
B. polygonata	*Gireoudia*	America (NC2-i)	14	CCBD
B. popenoei	*Gireoudia*	America (NC2-i)	14	CCBD
B. pringlei	*Gireoudia*	America (NC2-i)	14	Forrest, 2000
B. pruinata	*Gireoudia*	America (NC2-i)	14	CCBD
B. purpusii	*Gireoudia*	America (NC2-i)	14	CCBD
B. pustulata	*Gireoudia*	America (NC2-i)	14	CCBD
B. sartorii	*Gireoudia*	America (NC2-i)	14	CCBD
B. sericoneura	*Gireoudia*	America (NC2-i)	14	CCBD
B. squarrosa	*Gireoudia*	America (NC2-i)	14	CCBD
B. stigmosa	*Gireoudia*	America (NC2-i)	14	Forrest, 2000
B. strigillosa	*Gireoudia*	America (NC2-i)	14	CCBD
B. theimei	*Gireoudia*	America (NC2-i)	14	CCBD
B. urophylla	*Gireoudia*	America (NC2-i)	14	CCBD
B. vestita	*Gireoudia*	America (NC2-i)	14	Forrest, 2000
B. violifolia	*Gireoudia*	America (NC2-i)	14	Forrest, 2000
B. udisilvestris	*Parietoplacentaria*	America (NC2-i)	14	Forrest, 2000
B. biserrata	*Quadriperigonia*	America (NC2-i)	14	CCBD
B. bulbillifera	*Quadriperigonia*	America (NC2-i)	14	CCBD
B. gracilis	*Quadriperigonia*	America (NC2-i)	14	CCBD
B. uruapensis	*Quadriperigonia*	America (NC2-i)	14	CCBD
B. balmisiana	*Quadriperigonia*	America (NC2-i)	21	Forrest, 2000
B. heydei	*Urniformia*	America (NC2-i)	14	Forrest, 2000

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. crispula	*Astrothrix*	America (NC2-ii)	19	CCBD
B. wrightiana	*Begonia*	America (NC2-ii)	11	Forrest, 2000
B. bissei	*Begonia*	America (NC2-ii)	13	Forrest, 2000
B. cubensis	*Begonia*	America (NC2-ii)	26	CCBD
B. odorata	*Begonia*	America (NC2-ii)	26	Forrest, 2000
B. rotundifolia	*Begonia*	America (NC2-ii)	26	CCBD
B. dominicalis	*Begonia*	America (NC2-ii)	28	Legro and Doorenbos, 1969
B. obliqua	*Begonia*	America (NC2-ii)	28	CCBD
B. minor	*Begonia*	America (NC2-ii)	39	CCBD
B. serratifolia	*Begonia*	America (NC2-ii)	52	Forrest, 2000
B. acutifolia	*Begonia*	America (NC2-ii)	78	CCBD
B. bracteosa	*Cyathocnemis*	America (NC2-ii)	26	CCBD
B. roezlii	*Cyathocnemis*	America (NC2-ii)	26	CCBD
B. hirsuta	*Doratometra*	America (NC2-ii)	12	CCBD
B. wallichiana	*Doratometra*	America (NC2-ii)	15	CCBD
B. humilis	*Doratometra*	America (NC2-ii)	30	CCBD
B. cucullata	*Ephemera*	America (NC2-ii)	17	Forrest, 2000
B. fischeri	*Ephemera*	America (NC2-ii)	17	CCBD
B. hirtella	*Ephemera*	America (NC2-ii)	17	Forrest, 2000
B. mollicaulis	*Ephemera*	America (NC2-ii)	17	CCBD
B. subvillosa	*Ephemera*	America (NC2-ii)	17	CCBD
B. foliosa	*Lepsia*	America (NC2-ii)	30	CCBD
B. guaduensis	*Lepsia*	America (NC2-ii)	52	CCBD
B. edmundoi	*Pereira*	America (NC2-ii)	28	Forrest, 2000
B. lubbersii	*Pereira*	America (NC2-ii)	28	CCBD
B. radicans	*Solananthera*	America (NC2-ii)	28	CCBD
B. solananthera	*Solananthera*	America (NC2-ii)	28	CCBD
B. fulvosetulosa	*Trachelocarpus*	America (NC2-ii)	28	Forrest, 2000
B. herbacea	*Trachelocarpus*	America (NC2-ii)	28	CCBD
B. lanceolata	*Trachelocarpus*	America (NC2-ii)	28	Forrest, 2000
B. velloziana	*Trachelocarpus*	America (NC2-ii)	28	Forrest, 2000
B. cinnabarina	*Australes*	America (NC2-iii)	13	CCBD
B. pearcei	*Australes*	America (NC2-iii)	13	CCBD
B. boliviensis	*Australes*	America (NC2-iii)	14	Forrest, 2000
B. micranthera	*Australes*	America (NC2-iii)	14	CCBD
B. veitchii	*Australes*	America (NC2-iii)	14	CCBD

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. weddelliana	*Australes*	America (NC2-iii)	14	CCBD
B. crinita	*Australes*	America (NC2-iii)	17	CCBD
B. geraniifolia	*Eupetalum*	America (NC2-iii)	14	CCBD
B. octopetala	*Eupetalum*	America (NC2-iii)	14	CCBD
B. pleiopetala	*Eupetalum*	America (NC2-iii)	42	CCBD
B. cavum	*Knesebeckia (core)*	America (NC2-iii)	14	CCBD
B. falciloba	*Knesebeckia (core)*	America (NC2-iii)	14	CCBD
B. incarnata	*Knesebeckia (core)*	America (NC2-iii)	14	CCBD
B. monophylla	*Knesebeckia (core)*	America (NC2-iii)	14	CCBD
B. uniflora	*Knesebeckia (core)*	America (NC2-iii)	14	Forrest, 2000
B. viscida	*Knesebeckia (core)*	America (NC2-iii)	14	CCBD
B. ignea	*Knesebeckia (core)*	America (NC2-iii)	28	CCBD
B. acerifolia	*Knesebeckia (I)*	America (NC2-iii)	14	CCBD
B. ludwigii	*Knesebeckia (III)*	America (NC2-iii)	14	Forrest, 2000
B. wollnyi	*Knesebeckia (III)*	America (NC2-iii)	14	CCBD
B. wollnyi	*Knesebeckia (III)*	America (NC2-iii)	14	CCBD
B. capensis	*Ignota (*)*	Asia (*)	12	CCBD
B. physandra	*Ignota (*)*	Asia (*)	14	CCBD
B. boisiana	*Ignota (*)*	Asia (*)	15	CCBD
B. malabarica	*Ignota (*)*	Asia (*)	30	CCBD
B. peltatifolia	*Ignota (*)*	Asia (*)	30	Li et al unpub.
B. floccifera	*Flocciferae*	Asia (EDAB)	15	CCBD
B. dipetala	*Haagea*	Asia (EDAB)	15	CCBD
B. parvula	*Reichenheimia*	Asia (EDAB)	14	CCBD
B. henryi	*Reichenheimia*	Asia (EDAB)	15	CCBD
B. morelii	*Reichenheimia*	Asia (EDAB)	15	CCBD
B. alicida	*Alicida*	Asia C	14	Kono and Peng, unpub.
B. wenshanensis	*Diploclinium*	Asia C	9	Kono and Peng, unpub.
B. fimbristipula	*Diploclinium*	Asia C	11	CCBD
B. howii	*Diploclinium*	Asia C	11	Kono and Peng, unpub.
B. picta	*Diploclinium*	Asia C	11	CCBD
B. scintillans	*Diploclinium*	Asia C	11	Kono and Peng, unpub.
B. guaniana	*Diploclinium*	Asia C	12	Forrest, 2000
B. labordei	*Diploclinium*	Asia C	12	CCBD
B. lithophila	*Diploclinium*	Asia C	12	CCBD
B. muliensis	*Diploclinium*	Asia C	12	CCBD

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. grandis	*Diploclinium*	Asia C	13	CCBD
B. acaulis	*Diploclinium*	Asia C	14	CCBD
B. cavaleriei	*Diploclinium*	Asia C	15	CCBD
B. gigabracteata	*Diploclinium*	Asia C	15	Kono and Peng, unpub.
B. coptidimontana	*Diploclinium*	Asia C	16	Kono and Peng, unpub.
B. parviflora	*Diploclinium*	Asia C	16	CCBD
B. garrettii	*Diploclinium*	Asia C	18	Thamdee (1999)
B. modestiflora	*Diploclinium*	Asia C	18	Thamdee (1999)
B. ravenii	*Diploclinium*	Asia C	18	CCBD
B. concanensis	*Diploclinium*	Asia C	23	CCBD
B. chuyunshanensis	*Diploclinium*	Asia C	26	CCBD
B. wilsonii	*Diploclinium*	Asia C	27	CCBD
B. tengchiana	*Diploclinium*	Asia C	41	CCBD
B. crenata	*Lauchea*	Asia C	28	CCBD
B. tenuifolia	*Parvibegonia*	Asia C	11	CCBD
B. integrifolia	*Parvibegonia*	Asia C	14	Thamdee (1999)
B. variabilis	*Parvibegonia*	Asia C	14	Kono and Peng, unpub.
B. lacei	*Platycentrum*	Asia C	7	Kono and Peng, unpub.
B. griffithiana	*Platycentrum*	Asia C	8	Tseng et al., 2017
B. nepalensis	*Platycentrum*	Asia C	8	Tseng et al., 2017
B. rockii	*Platycentrum*	Asia C	8	Kono and Peng, unpub.
B. purpureofolia	*Platycentrum*	Asia C	9	Nakata et al., 2007
B. tamdaoensis	*Platycentrum*	Asia C	9	Peng et al., 2015
B. cathayana	*Platycentrum*	Asia C	10	CCBD
B. ceratocarpa	*Platycentrum*	Asia C	10	CCBD
B. hekouensis	*Platycentrum*	Asia C	10	Kono and Peng, unpub.
B. hongkongensis	*Platycentrum*	Asia C	10	Kono and Peng, unpub.
B. longicarpa	*Platycentrum*	Asia C	10	CCBD
B. polytricha	*Platycentrum*	Asia C	10	Kono and Peng, unpub.
B. abdullahpieei	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. acetosella	*Platycentrum*	Asia C	11	CCBD
B. aequilateris	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. aptera	*Platycentrum*	Asia C	11	CCBD
B. balansana	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. baviensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. burkillii	*Platycentrum*	Asia C	11	Kono and Peng, unpub.

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. cathcartii	*Platycentrum*	Asia C	11	CCBD
B. circumlobata	*Platycentrum*	Asia C	11	CCBD
B. coptidifolia	*Platycentrum*	Asia C	11	CCBD
B. daweishanensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. decora	*Platycentrum*	Asia C	11	CCBD
B. deliciosa	*Platycentrum*	Asia C	11	CCBD
B. diadema	*Platycentrum*	Asia C	11	CCBD
B. dielseiana	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. dryadis	*Platycentrum*	Asia C	11	CCBD
B. edulis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. gagnepainiana	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. hahiepiana	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. handelii	*Platycentrum*	Asia C	11	CCBD
B. hatacoa	*Platycentrum*	Asia C	11	CCBD
B. hemsleyana	*Platycentrum*	Asia C	11	CCBD
B. herveyana	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. jinyunensis	*Platycentrum*	Asia C	11	Ding et al., 2014
B. klossii	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. laminariae	*Platycentrum*	Asia C	11	CCBD
B. lancangensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. limprichtii	*Platycentrum*	Asia C	11	CCBD
B. lipingensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. longanensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. longialata	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. longifolia	*Platycentrum*	Asia C	11	CCBD
B. menglianensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. oreodoxa	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. palmata	*Platycentrum*	Asia C	11	CCBD
B. paupercula	*Platycentrum*	Asia C	11	CCBD
B. pavonina	*Platycentrum*	Asia C	11	CCBD
B. perakensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. psilophylla	*Platycentrum*	Asia C	11	CCBD
B. pulchrifolia	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. rex	*Platycentrum*	Asia C	11	CCBD
B. roxburghii	*Platycentrum*	Asia C	11	CCBD
B. rubinea	*Platycentrum*	Asia C	11	Li et al., 2005

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. rubropunctata	*Platycentrum*	Asia C	11	Nakata et al., 2007
B. siamensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. sikkimensis	*Platycentrum*	Asia C	11	CCBD
B. silletensis	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. sizemoreae	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. smithiana	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. sphenantheroides	*Platycentrum*	Asia C	11	Peng et al., 2015
B. subhowii	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. truncatiloba	*Platycentrum*	Asia C	11	CCBD
B. versicolor	*Platycentrum*	Asia C	11	CCBD
B. wyepingiana	*Platycentrum*	Asia C	11	Kono and Peng, unpub.
B. xanthina	*Platycentrum*	Asia C	11	CCBD
B. discrepans	*Platycentrum*	Asia C	12	Kono and Peng, unpub.
B. iridescens	*Platycentrum*	Asia C	12	Kono and Peng, unpub.
B. annulata	*Platycentrum*	Asia C	13	CCBD
B. beddomei	*Platycentrum*	Asia C	13	CCBD
B. manhaoensis	*Platycentrum*	Asia C	13	Kono and Peng, unpub.
B. emeiensis	*Platycentrum*	Asia C	17	Kono and Peng, unpub.
B. austrotaiwanensis	*Platycentrum*	Asia C	19	CCBD
B. bouffordii	*Platycentrum*	Asia C	19	CCBD
B. chitoensis	*Platycentrum*	Asia C	19	CCBD
B. myanmarica	*Platycentrum*	Asia C	19	Tseng et al., 2017
B. pinglinensis	*Platycentrum*	Asia C	19	CCBD
B. taiwaniana	*Platycentrum*	Asia C	19	CCBD
B. cucurbitifolia	*Platycentrum*	Asia C	22	CCBD
B. pedatifida	*Platycentrum*	Asia C	22	Nakata et al., 2007
B. tetralobata	*Platycentrum*	Asia C	22	Kono and Peng, unpub.
B. thomsonii	*Platycentrum*	Asia C	22	Kono and Peng, unpub.
B. venusta	*Platycentrum*	Asia C	22	CCBD
B. villifolia	*Platycentrum*	Asia C	22	Nakata et al., 2007
B. lukuana	*Platycentrum*	Asia C	26	CCBD
B. wutaiana	*Platycentrum*	Asia C	26	CCBD
B. baliensis	*Platycentrum*	Asia C	28	Kono and Peng, unpub.
B. formosana	*Platycentrum*	Asia C	30	CCBD
B. robusta	*Platycentrum*	Asia C	33	Kono and Peng, unpub.
B. multangula	*Platycentrum*	Asia C	44	Kono and Peng, unpub.

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. chingipengii	*Baryandra*	Asia D	14	Peng et al., 2017
B. chloroneura	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. copelandii	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. elmeri	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. fenicis	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. gueritziana	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. hernandioides	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. luzonensis	*Baryandra*	Asia D	14	Forrest, 2000
B. merrilliana	*Baryandra*	Asia D	14	Peng et al., 2017
B. mindorensis	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. neopurpurea	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. sharpeana	*Baryandra*	Asia D	14	Forrest, 2000
B. sykakiengii	*Baryandra*	Asia D	14	Peng et al., 2017
B. tandangii	*Baryandra*	Asia D	14	Peng et al., 2017
B. taraw	*Baryandra*	Asia D	14	Peng et al., 2017
B. tayabensis	*Baryandra*	Asia D	14	Kono and Peng, unpub.
B. acuminatissima	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. anisoptera	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. biliranensis	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. blancii	*Baryandra*	Asia D	15	Forrest, 2000
B. calcicola	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. castilloi	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. cleopatrae	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. culasiensis	*Baryandra*	Asia D	15	Peng et al., 2017
B. elnidoensis	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. gabaldonensis	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. gitingensis	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. hughesii	*Baryandra*	Asia D	15	Peng et al., 2017
B. klemmei	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. longiscapa	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. manillensis	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. nigritarum	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. obtusifolia	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. oxysperma	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. rubiteae	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. rufipila	*Baryandra*	Asia D	15	Kono and Peng, unpub.

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. suborbiculata	*Baryandra*	Asia D	15	Peng et al., 2017
B. tagbanua	*Baryandra*	Asia D	15	Peng et al., 2017
B. trichochila	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. wadei	*Baryandra*	Asia D	15	Kono and Peng, unpub.
B. subnummularifolia	*Baryandra*	Asia D	16	Legro and Doorenbos 1971
B. parva	*Baryandra*	Asia D	18	Legro and Doorenbos 1969
B. arachnoidea	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. aurantiflora	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. auritistipula	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. austroguangxiensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. bamaensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. biflora	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. calciphila	*Coelocentrum*	Asia D	15	Peng et al., 2015
B. chongzuoensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. curvicarpa	*Coelocentrum*	Asia D	15	CCBD
B. danxianensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. debaoensis	*Coelocentrum*	Asia D	15	CCBD
B. fangii	*Coelocentrum*	Asia D	15	CCBD
B. filiformis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. guixiensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. jingxiensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. kui	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. langsonensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. lanternaria	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. leipingensisis	*Coelocentrum*	Asia D	15	Li et al (2016)
B. leprosa	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. liuyanii	*Coelocentrum*	Asia D	15	CCBD
B. locii	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. longa	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. longgangensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. longistyla	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. luochengensis	*Coelocentrum*	Asia D	15	CCBD
B. luzhaiensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. masoniana	*Coelocentrum*	Asia D	15	Li et al unpub.
B. melanobullata	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. montaniformis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.

Continued on next page

Table 2.1 – continued from previous page

Species	Section	Clade	n	Source
B. ningmingensis	*Coelocentrum*	Asia D	15	CCBD
B. ornithophylla	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. pengii	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. phuthoensis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. picturata	*Coelocentrum*	Asia D	15	CCBD
B. porteri	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. pseudodaxinensis	*Coelocentrum*	Asia D	15	CCBD
B. pseudodryadis	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. retinervia	*Coelocentrum*	Asia D	15	CCBD
B. semiparietalis	*Coelocentrum*	Asia D	15	CCBD
B. setulosa-peltata	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. sinofloribunda	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. subcoriacea	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. umbraculifolia	*Coelocentrum*	Asia D	15	Kono and Peng, unpub.
B. guangxiensis	*Coelocentrum*	Asia D	17	Kono and Peng, unpub.
B. muricata	*Jackia*	Asia D	12	Forrest, 2000
B. bonii	*Jackia*	Asia D	15	Kono and Peng, unpub.
B. rajah	*Jackia*	Asia D	15	CCBD
B. goegoensis	*Jackia*	Asia D	17	CCBD
B. sublobata	*Jackia*	Asia D	17	Kono and Peng, unpub.
B. nurii	*Jackia*	Asia D	22	CCBD
B. abbreviata	*Petermannia*	Asia D	15	Peng et al., 2015
B. aequata	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. beryllae	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. borneensis	*Petermannia*	Asia D	15	CCBD
B. cauliflora	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. darthvaderiana	*Petermannia*	Asia D	15	Li et al unpub.
B. hainanensis	*Petermannia*	Asia D	15	Hu, Cheng et al., unpub.
B. isoptera	*Petermannia*	Asia D	15	CCBD
B. jiewhoei	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. labiensis	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. merrittii	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. naumoniensis	*Petermannia*	Asia D	15	CCBD
B. palawanensis	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. paracauliflora	*Petermannia*	Asia D	15	Kono and Peng, unpub.
B. pseudolateralis	*Petermannia*	Asia D	15	Kono and Peng, unpub.

Continued on next page

Table 2.1 – continued from previous page

Species	**Section**	**Clade**	**n**	**Source**
B. pubescens	*Petermannia*	Asia D	18	CCBD
B. augustae	*Petermannia*	Asia D	22	CCBD
B. bipinnatifida	*Petermannia*	Asia D	22	Peng et al., 2015
B. brevirimosa	*Petermannia*	Asia D	22	CCBD
B. serratipetala	*Petermannia*	Asia D	22	CCBD
B. holttumii	*Petermannia*	Asia D	30	Peng et al., 2015
H. sandwichensis		Outgroup	24	Forrest, 2000

2.A.2 Genome sizes of *Begonia*

Species	Section	Clade	1C-value (Mbp)	n	Original reference
B. socotrana	*Peltaugustia*	Africa (SB)	465	14	Campos et al., unpub.
B. johnstonii	*Rostrobegonia*	Africa (SDAAB1)	372	13	Campos et al., unpub.
B. sutherlandii	*Augustia*	Africa (SDAAB2)	597.8	13	Dewitte et al.,2009
B. dregei	*Augustia*	Africa (SDAAB2)	636.8	13	Campos et al., unpub.
B. ulmifolia	*Donaldia*	America (NC1)	245	15	Dewitte et al.,2009
B. maculata	*Gaerdtia*	America (NC1)	519.4	28	Dewitte et al.,2009
B. undulata	*Gaerdtia*	America (NC1)	548.8	28	Dewitte et al.,2009
B. coccinea	*Gaerdtia*	America (NC1)	548.8	28	Dewitte et al.,2009
B. corallina	*Gaerdtia*	America (NC1)	764.4	-	Dewitte et al.,2009
B. venosa	*Pritzelia*	America (NC1)	245	26	Dewitte et al.,2009
B. dietrichiana	*Pritzelia*	America (NC1)	245	28	Dewitte et al.,2009
B. hispida var. cucullifera	*Pritzelia*	America (NC1)	291.4	28	Wu et al., 2018
B. listada	*Pritzelia*	America (NC1)	303.8	28	Dewitte et al.,2009
B. soli-mutata	*Pritzelia*	America (NC1)	303.8	-	Dewitte et al.,2009
B. echinosepala	*Pritzelia*	America (NC1)	313.6	28	Dewitte et al.,2009
B. tomentosa	*Pritzelia*	America (NC1)	333.2	28	Dewitte et al.,2009
B. luxurians	*Pritzelia*	America (NC1)	342.3	28	Campos et al., unpub.
B. hispida	*Pritzelia*	America (NC1)	343	28	Dewitte et al.,2009
B. metallica	*Pritzelia*	America (NC1)	343	28	Dewitte et al.,2009
B. glabra	*Wageneria*	America (NC1)	254.8	19	Dewitte et al.,2009
B. conchifolia	*Gireoudia*	America (NC2-i)	586	14	Campos et al., unpub.
B. manicata	*Gireoudia*	America (NC2-i)	695.8	14	Dewitte et al.,2009
B. bowerae	*Gireoudia*	America (NC2-i)	715.4	14	Dewitte et al.,2009
B. heracleifolia	*Gireoudia*	America (NC2-i)	735	14	Dewitte et al.,2009
B. odorata	*Begonia*	America (NC2-ii)	548.8	26	Dewitte et al.,2009
B. subvillosa	*Ephemera*	America (NC2-ii)	284.2	17	Dewitte et al.,2009
B. cucullata	*Ephemera*	America (NC2-ii)	294	17	Dewitte et al.,2009
B. subvillosa	*Ephemera*	America (NC2-ii)	388.2	17	Wu et al., 2018
B. fischeri	*Ephemera*	America (NC2-ii)	872.2	17	Dewitte et al.,2009
B. foliosa	*Lepsia*	America (NC2-ii)	351.1	30	Wu et al., 2018
B. radicans	*Solananthera*	America (NC2-ii)	526.2	28	Wu et al., 2018
B. solananthera	*Solananthera*	America (NC2-ii)	558.6	28	Dewitte et al.,2009
B. radicans	*Solananthera*	America (NC2-ii)	568.4	28	Dewitte et al.,2009
B. boliviensis	*Australes*	America (NC2-iii)	1372	14	Dewitte et al.,2009
B. pearcei	*Australes*	America (NC2-iii)	1,430.8	13	Dewitte et al.,2009
B. peltatifolia	*Ignota*	Asia (Unknown sect.)	285.6	30	Wu et al., 2018
B. laminariae	*Platycentrum*	Asia C	367.7	11	Wu et al., 2018
B. augustinei	*Platycentrum*	Asia C	375.5	-	Wu et al., 2018
B. cathayana	*Platycentrum*	Asia C	400	10	Wu et al., 2018

Continued on next page

Table 2.2 – continued from previous page

Species	Section	Clade	1C-value (Mbp)	n	Original reference
B. pedatifida	*Platycentrum*	Asia C	487	22	Wu et al., 2018
B. baviensis	*Platycentrum*	Asia C	519.3	11	Wu et al., 2018
B. crocea	*Platycentrum*	Asia C	539.9	-	Wu et al., 2018
B. hemsleyana	*Platycentrum*	Asia C	555.5	11	Wu et al., 2018
B. acetosella	*Platycentrum*	Asia C	567.2	11	Wu et al., 2018
B. diadema	*Platycentrum*	Asia C	568.4	11	Dewitte et al.,2009
B. palmata	*Platycentrum*	Asia C	582.9	11	Wu et al., 2018
B. psilophylla	*Platycentrum*	Asia C	623	11	Wu et al., 2018
B. lacerata	*Platycentrum*	Asia C	635.7	-	Wu et al., 2018
B. cucurbitifolia	*Platycentrum*	Asia C	704.2	22	Wu et al., 2018
B. gulinqingensis	*Platycentrum*	Asia C	765.8	-	Wu et al., 2018
B. dryadis	*Platycentrum*	Asia C	771.6	11	Wu et al., 2018
B. rex	*Platycentrum*	Asia C	825.4	11	Wu et al., 2018
B. longialata	*Platycentrum*	Asia C	935	11	Wu et al., 2018
B. handelii	*Platycentrum*	Asia C	958.4	11	Wu et al., 2018
B. tetralobata	*Platycentrum*	Asia C	1,048.4	22	Wu et al., 2018
B. ceratocarpa	*Platycentrum*	Asia C	1,714.4	10	Wu et al., 2018
B. handelii	*Platycentrum*	Asia C	1,787.8	33	Wu et al., 2018
B. pinglinensis	*Platycentrum*	Asia C	1,932.5	19	Wu et al., 2018
B. deliciosa	*Platycentrum*	Asia C	2,132.0	11	Wu et al., 2018
B. fomosana	*Platycentrum*	Asia C	2,497.8	30	Wu et al., 2018
B. masoniana	*Coelocentrum*	Asia D	661.1	15	Wu et al., 2018
B. fangii	*Coelocentrum*	Asia D	689.5	15	Wu et al., 2018
B. wangii	*Coelocentrum*	Asia D	791.2	15	Wu et al., 2018
B. pseudodryadis	*Coelocentrum*	Asia D	1,824	15	Wu et al., 2018
B. cavaleriei	*Diploclinium*	Asia D	599.51	15	Wu et al., 2018
B. grandis	*Diploclinium*	Asia D	666.4	13	Dewitte et al.,2009
B. serratipetala	*Petermannia*	Asia D	891.8	22	Dewitte et al.,2009
B. brevirimosa	*Petermannia*	Asia D	911.4	22	Dewitte et al.,2009
B. bipinnatifida	*Petermannia*	Asia D	1,134.5	22	Campos et al., unpub.
H. sandwicensis	*Outgroup*	Hillebrandia	332.6	24	Campos et al., unpub.

Chapter 3

Generating genomic resources to study *Begonia* evolution: de novo genome sequencing, assembly and gene annotation of seven Begoniaceae species

This chapter includes the work and results of the Bioinformatics MSc Thesis of Andrea Martínez Martínez (University of Edinburgh) on the genome assembly of *Hillebrandia sandwicensis*, as well as the internship project work and results of Antoine Cossard (École Nationale Supérieure Agronomique de Toulouse) on the genome annotation of *Hillebrandia sandwicensis*.
Also included here is the collaborative work on the long-read sequencing and assembly of *B. conchifolia* with Dr. Aureliano Bombarely Gomez (Università degli Studi di Milano Statale).

3.1 Introduction

Studying genetic data to understand the basis of species diversity has for many years been a major and standard procedure in evolutionary biology studies. Traditional phylogenetic methods are based on the analysis of single loci or small number (<10) of genomic regions. This approach may be sufficient for studies of distantly related, slowly divergent taxa, but would likely be biased and uninformative when studying fast-evolving taxa or closely-related species (Maddison, 1997; Nichols, 2001; Edwards, 2009). Genomic changes can be highly heterogeneous, and evolutionary processes do not only involve slow and stochastic genetic mutations. Dramatic and immediate large-scale changes such as hybridisation, polyploidisation or other types of genomic rearrangements are common in plants (Otto and Whitton, 2000). Such processes would be difficult to prove or detect when only studying a small number of genomic loci, and Next-Generation Sequencing (NGS) approaches such as whole-genome sequencing or target-capture allow resolving evolutionary relationships with a higher resolution (Nater et al., 2015).

Results from Chapter 2 described evidence of *Begonia* genome dynamics at the genome architecture level. It was shown that this genus has highly variable chromosome number and genome size, and that many sections show stable chromosome numbers but variable genome sizes. Genome size variation correlated positively with levels of species diversity in some sections (e.g., the *Petermannia*, *Coelocentrum* or *Gireoudia* sections). To better understand the possible role of these genome dynamics in *Begonia* radiations, whole-genome analysis of transposon content, genomic landscape and gene content diversity are needed. Unless economically important, most plant families and genera have one or two species fully sequenced (see Sequenced Plant Genomes Database: plabipd.de), and these are used as reference in evolutionary studies. However, although some NGS data is currently available for *Begonia* (Brennan et al., 2012; Emelianova, 2017; Tseng et al., 2017; Griesmann et al., 2018), given the level of genome variation and the wide species diversity found in this genus (addressed in Chapter 2), we believe that sequencing a carefully selected small number of genomes across the genus would contribute to obtaining a better overview of the evolutionary history of this phylogenetically complex,

mega-diverse genus.

Although NGS is increasingly cheaper and new techniques and protocols are making possible to sequence almost any living organism (Goodwin et al., 2016; Lewin et al., 2018), there are certain caveats to consider when starting a genome sequencing project on non-model organisms (Ellegren, 2014). Different species present various genomic traits that affect genome assembly pipelines and have an effect on the resultant genome assemblies. For example, high heterozygosity levels can be a major issue when assembling non-model organism genomes. Genomes from polymorphic species with considerable genetic diversity are a mosaic of homozygous and heterozygous regions, and in particular paralogous genes and heterozygous regions can be confused in fragmented sequences, especially when only short read sequencing data is used (Vinson et al., 2005; Barrière et al., 2009). Sequencing divergent heterozygous haplotypes impedes standard assembly approaches, and requires dedicated assembly of haplotypes prior to building the scaffold graph (Kajitani et al., 2019). Model organisms are usually inbred and therefore highly homozygous. However, non-model organisms are sometimes hard to inbreed due to long generation times or self-incompatibility. Moreover, other genome characteristics such as large genome size and high ploidy level or/and repeat content will create difficulties for genome assembly. Long read and other sequencing technologies such as Hi-C can overcome or help dealing with these issues, but then project costs increase to levels where sequencing multiple species is unaffordable. Nonetheless, whole genome sequencing using short-read Illumina data is still highly useful and scalable, and although it generally generates fragmented assemblies, these often contain a good representation of gene space, therefore allowing them to be used for some comparative genomics studies. This approach does not produce high-contiguity genomes, but in many cases these aren't required for many of the questions addressed in evolutionary and ecological studies. The need for long-read sequencing approaches for evolutionary studies on repeatome comparative genomics will be further discussed in Chapters 4 and 7.

This chapter presents the Whole Genome Sequencing (WGS) project of seven Begoniaceae species. To avoid high costs and be able to generate data for as many species as

possible, we selected one species for long-read sequencing (PacBio and Nanopore), and six more for short-read Illumina sequencing. While the data in this chapter was being generated and processed, the first *Begonia* genome assembly was published (Griesmann et al., 2018), and more *Begonia* genome sequencing projects started (Li et al., in prep). This involved the sequencing and assembly of other five *Begonia* genomes from across the *Begonia* phylogeny that will also be included in some of the analyses in Chapter 4. The assembly and annotation metrics of these genomes are shown in appendix 3.A.1. All Begoniaceae species currently sequenced and assembled and their position in a simplified phylogeny are shown in Figure 3.1.

For this chapter, the only species of the *Begonia* monotypic sister genus *Hillebrandia sandwicensis*, was selected as an outgroup. Within the *Begonia* genus, we sequenced three African species (*B. dregei*, section *Augustia*; *B. socotrana*, section *Peltaugustia* and *B. johsntonii*, section *Rostrobegonia*, sections with 12, 2 and 9 species, respectively) from small clades that represent the sister groups to the three main *Begonia* radiations: the two Neotropical clades (Neotropical clade 1 and 2), from which *B. luxurians* (sect. *Pritzelia*, 158 species) and *B. conchifolia* (sect. *Gireoudia*, 110 species) were selected for sequencing, respectively, and the main Asian clade, from which *B. bipinnatifida* (sect. *Petermannia*, 451 species) was sequenced. Griesmann et al. (2018) made publicly available the genome of the Neotropical species *B. foliosa* var. *miniata* (referred to by its synonym *B. fuchsioides* in Griesmann et al. (2018)), from section *Lepsia* (with 8 species) , in the Neoropical Clade 2-ii. Moreover, our collaborator's project (Li et al., unpublished) selected one African species (*B. loranthoides*, section *Tetraphilla*, from the Malagasy Begonia clade), and three Asian species from the Asian Clade D (*B. masoniana, B. peltatifolia* and *B. darthvaderiana*, from sections *Coelocentrum*, *Ignota*, and *Petermannia*, with 84, unknown and 451 species respectively). Althought *B. peltatifolia* has no assigned section (sect. *Ignota* refers to "unknown section"), Moonlight et al. (2018) places this species between *Coelocentrum* and *Petermannia*.

These species represent not only a wide phylogenetic range but also a broad variation in genome size and chromosome number even within clades, as Table 3.1 shows.

With this WGS project, we aim to generate a series of genomic resources that together

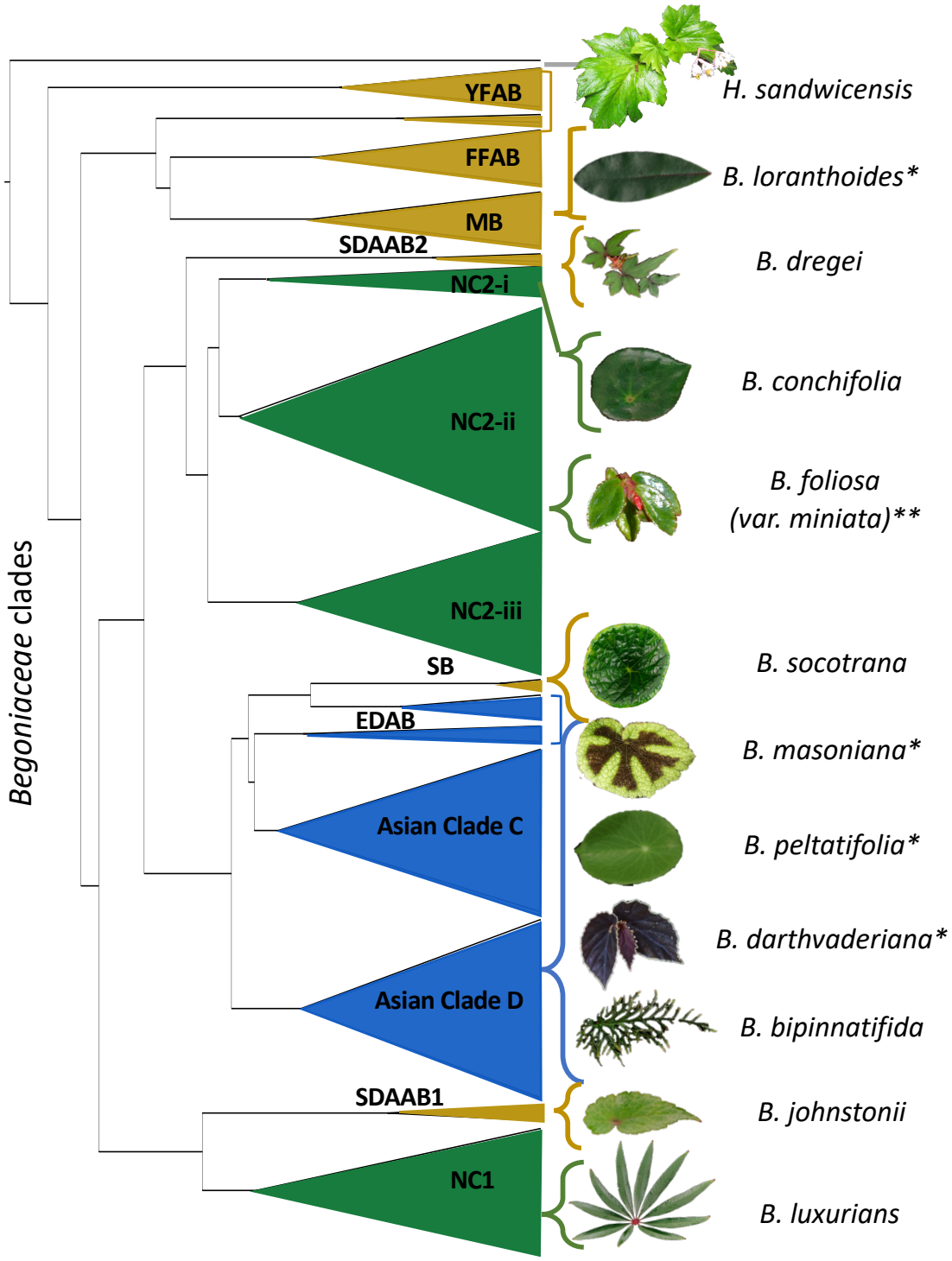


Figure 3.1: Leaf images and phylogenetic possition of the 12 Begoniaceae species currently used for whole genome sequencing approaches in a simplified *Begonia* clade phylogeny (Moonlight et al., 2018). The species sequenced by Li et al (in prep.) are indicated with an asterisk, and the species sequenced in Griesmann et al. (2018) is indicated with two asterisks.

with those currently available, can help address many questions still to be acknowledged about *Begonia* evolution. In the context of this thesis, I will use these genome assemblies to test the hypothesis that genome dynamics are drivers of *Begonia* radiation events by comparing genome landscapes across the genus, looking for changes at the root of radiations in Asia and South America.

Table 3.1: Genome sizes, chromosome numbers (from data available in Chapter 2) and average chromosome sizes of the 12 Begoniaceae species currently sequenced and assembled. For *B. loranthoides* and *B. darthvaderiana* there is no flow cytometry data available, and the genome size estimations indicated here were inferred from their k-mer distributions (Li et al., in prep).

Species name	**Haploid genome size (Mbp)**	**Haploid chromosome number**	**Average chromosome size (Mbp)**
H. sandwicensis	332.6	24	13.9
*B.loranthoides**	724.0	19	38.1
B. dregei	635.7	13	48.9
B. conchifolia	586.8	14	41.9
B. foliosa (var. *miniata*)	351.1	30	11.7
B. socotrana	464.6	14	33.2
B. masoniana	661.1	15	44.1
B. peltatifolia	285.6	15	19.0
*B. darthvaderiana**	797.0	15	53.1
B. bipinnatifida	1,134.5	22	51.6
B. johnstonii	371.6	13	28.6
B. luxurians	342.3	28	12.2

This chapter presents the whole genome assemblies and annotation of seven species, and aims to explore and compare different approaches needed to assemble genomes of diverse complexity and characteristics. Moreover, I intend to analyse and assess the limitations of generating whole-genome assemblies using only short-read data for complex genomes, in contrast with the use of more cost intensive long-read technologies.

3.2 Results and discussion

3.2.1 Genome sequencing of seven Begoniaceae species

We sequenced the genomes of seven Begoniaceae species to medium-high coverage (50-300x) using short-read Illumina platforms. The amount of short-read data for each species (before and after trimming and non-target organism detection) is described in Table 3.2.

Table 3.2: Whole Genome Sequencing (WGS) output statistics before and after filtering low quality reads and contaminants.

Species name	**Total no. of reads**	**Initial coverage (x fold 1C)**	**No. of surviving reads after trimming and filtering***	**% of initial reads**	**Final coverage (x fold 1C)**
H. sandwicensis	240,036,376	108	232,862,700	97.0	105
B. dregei	283,503,242	67	275,316,402	97.1	65
B. conchifolia	477,059,644	122	475,449,966	99.7	122
B. socotrana	278,380,786	90	271,296,132	97.5	88
B. bipinnatifida	906,638,326	120	897,663,880	99.0	119
B. johnstonii	646,118,134	261	498,180,280	77.1	201
B. luxurians	317,545,952	139	308,306,180	97.1	135

(*) Contaminant filtering was only done at this step for the *B. johnstonii* dataset, as the k-mer distribution profile suggested it was highly contaminated. As much contaminant species DNA as possible was removed and raw reads were recovered using Blobtoolkit (initial k-mer profile and blobplots in appendix 3.A.2). None of the other species data needed to be cleaned for contaminants at this step.

This table supports the high quality of the short-read data obtained as, after filtering the low-quality reads, over 95% of the data survived in most species and was used for k-mer analysis. Due to the striking genome size differences between species, the sequence coverage varied widely between species, from 65x in *B. dregei* to 201x in *B. johnstonii*.

We sequenced the *Begonia conchifolia* genome at low coverage using PacBio and Nanopore long-read technologies (∼19x and ∼10x, respectively). A total of 1,673,858 PacBio reads and 568,783 Oxford Nanopore reads were obtained for the assembly of the *Begonia conchifolia* genome. A summary of the statistics from the obtained datasets is shown in appendix 3.A.3.

3.2.2 K-mer analysis and genome characteristics of seven Begoniaceae species

Using the trimmed and filtered reads showed in Table 3.2, I generated k-mer distribution profiles for every species and k-mer sizes from 21 to 121 in intervals of 10 nucleotides. K-mer profiles provide useful information on the sequenced genomes such as heterozygosity and ploidy levels, repeat content, genome coverage (if the genome size was unknown), as well as potential contaminants. Figure 3.2 shows the 31-mer graphs obtained for each of the species data.

Our data gives very different k-mer profiles for each species. *Hillebrandia sandwicensis* shows a simple single-peak curve, whereas all *Begonia* show, to some extent, two-peak curves. This implies potentially all the sequenced *Begonia* species have heterozygous genomes (Li and Harkess, 2018). These heterozygosity levels are also quite variable, as their haploid (right) and diploid (left) peaks also show high variability in height. The higher the haploid peak, the more heterozygous the genome: this means *B. dregei* and *B. bipinnatifida* have the highest heterozygosity rates, with approximate values of 2.07 and 2.43%, respectively. This could be due to a possible hybrid or polyploid origin of these species; while the *B. dregei* specimen sequenced is from cultivated origin and can potentially be a hybrid, the *B. bipinnatifida* genome was sequenced from a wild collected specimen, and its chromosome number and genome sizes indicate it could be a species of polyploid origin. The heterozygosity levels of the rest of the sequenced *Begonia* species are variable, ranging from 0.1% (in *B. johnstonii*) to 0.98% (in *B. luxurians*), with *B. socotrana* showing an approximate heterozygosity rate of 0.64% and *B. conchifolia* of 0.47%. All these values are also considerably high compared to the low heterozygosity of *Hillebrandia* (0.03%; heterozygosity values obtained by GenomeScope, from Ranallo-Benavidez et al. (2020)). In the case of *Begonia johnstonii*, there is no haploid peak but the 31-mer profile shows a main diploid peak with a slight hump at 1/3 of its coverage. Although most of the contaminant sequences were filtered, this hump represents a low number of contaminant reads that could not be filtered out. However, it is expected that due to the low coverage, these sequences will not be present in the final assembly. This also means *B. johnstonii* is the least heterozygous of all the *Begonia* species sequenced.

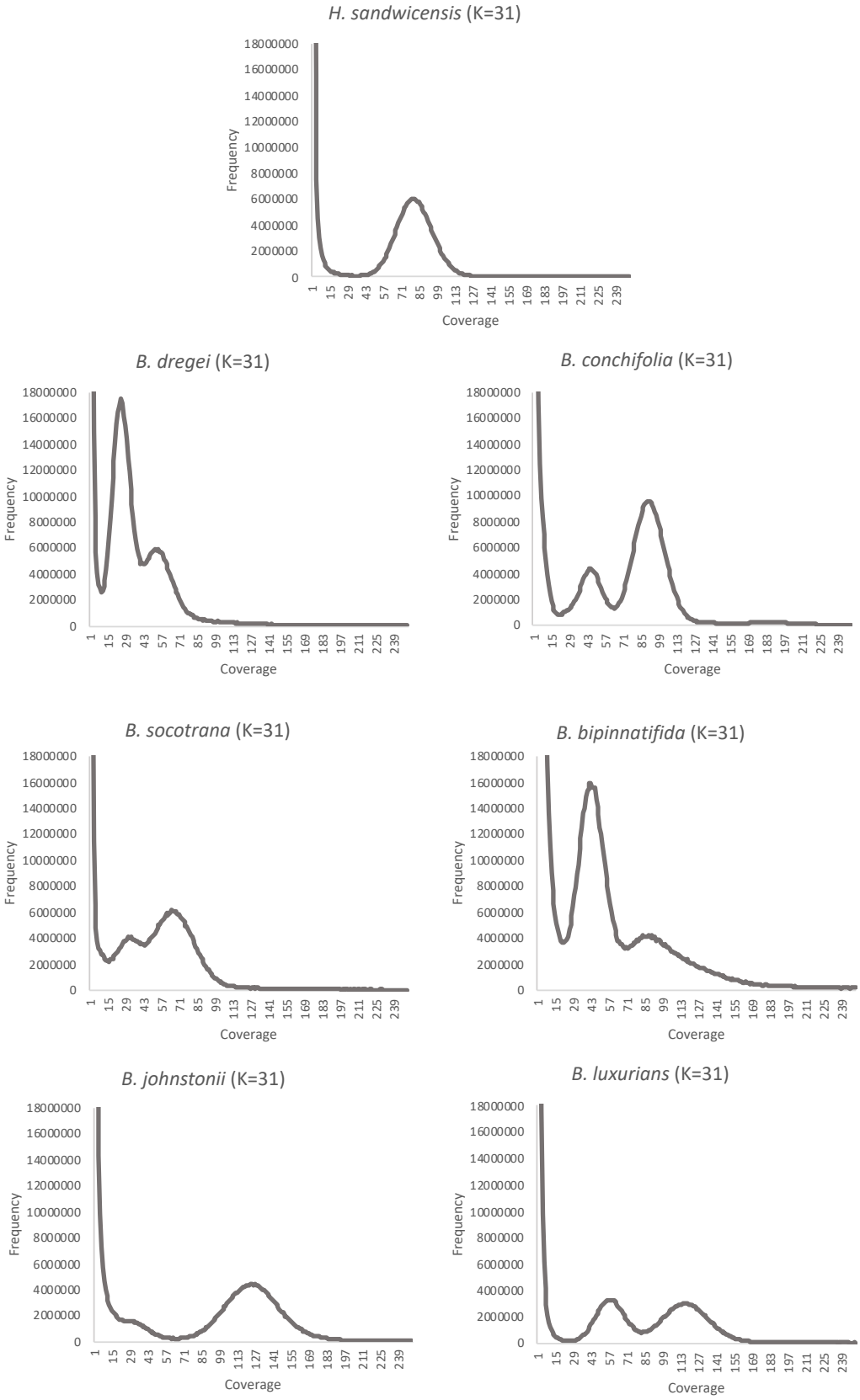


Figure 3.2: 31-mer distribution profiles for all the Begoniaceae species sequenced in this chapter.

Levels of repetitive DNA in the genome can also be inferred from the k-mer profile shape: the longer the haploid peak tail, the more repetitive the genome (Li and Harkess, 2018). *Begonia* k-mer profiles generally show a slight tail (especially *B. bipinnatifida*), which indicates abundant repetitive DNA. In contrast, the *Hillebrandia* curve is more abrupt, indicating this is a less repetitive genome. In Chapter 4 the differences in repeat content in all these genomes will be further studied using different approaches.

3.2.3 Draft genome assemblies of seven Begoniaceae species

Several assembly approaches were tested and compared at this step. Different assembly strategies were used for the genome with long-read data in contrast with the other six. For the *B. conchifolia* genome data, several long-read assemblers were tested using the PacBio data. The pipeline used and all assembly metrics from the different assemblers tested are specified in appendices 3.A.4 and 3.A.6. For the remaining six genomes, 13 different k-mer sizes and between 10 and 12 assembly pipelines were tested. Pipelines used and assembly stats of all these assemblies are shown in appendices 3.A.5 and 3.A.6 Assembly metrics of the best pipelines are shown in Table 3.3. Once selected, all final genome assemblies went through a contaminant filtering step to remove any contaminant sequences (blobplots of assemblies before and after filtering in appendix 3.A.7).

Table 3.3: Final genome assembly metrics for the seven Begoniaceae species sequenced. Best genome metrics are coloured in light green, while light orange and red highlight the lower quality ones.

Species name	**Read type available**	**Best assembler used**	**Final assembly span (Mb)**	**Final assembly span (% gsize)**	**No contigs**	**N50 (kb)**	**Longest contig (kb)**
H. sandwicensis	short	SOAPdenovo2	248.53	74.73	4,579	145.89	885.50
B. dregei	short	Platanus-allee	561.48	88.33	69,315	21.85	389.10
B. conchifolia	long+short	Flye	564.91	96.27	3,860	314.60	1,820
B. socotrana	short	Platanus-allee	304.65	65.58	48,818	19.89	496.01
B. bipinnatifida	short	Platanus-allee	1099.57	96.92	175,180	11.71	2,793.18
B. johnstonii	short	Platanus-allee	291.15	78.34	13,383	58.63	645.15
B. luxurians	short	Platanus-allee	254.89	74.46	20,233	36.86	373.71

As indicated by Table 3.3 and expected, results show that the assembly of *B. conchifolia* has a much higher N50 and contiguity than the rest, and therefore is of better quality, due to the use of long-read data. Flye (Kolmogorov et al., 2019) was the long-read assembler that proved to produce the best assembly of our data, whereas for most

of the short-read data assemblies, Platanus-allee was found to be the best assembler. This is expected as this software is written specifically for highly heterozygous plant genomes (Kajitani et al., 2019). Table 3.3 also shows that lowest assembly qualities are from those species with larger genomes and higher heterozygosity values, and the best illumina assembly being the *H. sandwicensis* due to a combination of low heterozygosity and high sequence coverage. These results also show that although coverage is an important and limiting factor, others such as heterozygosity and repeat content limit the assembly quality even in species where assembly is higher than 100x. For instance, in the case of *B. bipinnatifida*, even though we sequenced this species in high coverage, the high content of repeats and heterozygosity make this the least contiguous assembly. In fact, one of the main factors having a negative impact on genome assembly quality is the genomic repeat content, which has been shown to be highly variable in Begoniaceae, and will be further discussed in the next chapter.

As the assembly data in appendix 3.A.1 shows, the metrics of our assembled genomes are only comparable with the genome of *B. foliosa* by Griesmann et al. (2018). The genome assemblies by Li et al (in prep) show much higher levels of contiguity, represented by N50 values of almost 1000 higher than our genomes. This is mostly due to the high coverage data obtained, as well as the use of 10x and Hi-C technologies. Moreover, Li et al., (in prep) obtained long read data for two of the genomes *B. masoniana* and *B. darthvaderiana*, which clearly show higher contiguity than *B. loranthoides* and *B. peltatifolia*.

Determining whether a genome assembly is good enough for annotation is a key step that is difficult to take because of the numerous pipelines available for genome assembly, polishing, gap-filling, etc. For these genomes, several approaches were attempted and, although the resulting assemblies present variable characteristics and levels of contiguity, at some point it was decided they were good enough to proceed with annotation. Since N50 values give the most information on assembly contiguity these must be compared to select the best assembly (Del Angel et al., 2018). Other studies suggest that N50s over 6 kbp are good enough for gene prediction analysis (Stevens, 2020). Once the best version of our genomes were selected, genome quality and completeness

were assessed using BUSCO (version 4.0) with the viridiplantae odb10 dataset (Seppey et al., 2019). Table 3.4 shows BUSCO results for all the final assemblies.

Table 3.4: BUSCO results for gene set completeness obtained for the seven Begoniaceae genome assemblies.

Species name	**No. complete single BUSCOs (%)**	**No. duplicated BUSCOs (%)**	**No. fragmented BUSCOs (%)**	**No. missing BUSCOs (%)**
H. sandwicensis	91.80	4.50	1.20	2.10
B. dregei	88.20	4.00	6.10	1.70
B. conchifolia	83.80	11.30	2.60	2.30
B. socotrana	93.40	3.50	2.10	1.00
B. bipinnatifida	36.50	35.80	22.60	5.20
B. johnstonii	84.20	4.20	8.90	2.70
B. luxurians	83.10	9.90	4.90	2.10

As Table 3.4 shows, the total of full BUSCOs found in all genomes (single+duplicated) is over 85% in five out of the six assembled genomes, and over 90% in five out of seven. This suggests that the assembly completeness is high in most genomes, with only a few predicted genes found to be fragmented or missing. Best results are from *H. sandwicensis* and *B. socotrana*, even though the assembly of the last one did not show great levels of contiguity. A clear outlier of these gene prediction results is the *B. bipinnatifida* assembly, which shows only a 36.5% of single complete BUSCOs and almost the same amount of duplicated ones. This indicates there is a large number of duplicated sequences in this assembly. As mentioned above, *Begonia bipinnatifida* has the largest genome of the seven sequenced species, together with a high chromosome number (2n=48), which could indicate a possible tetraploid origin of this species. This could explain the high heterozygosity levels found in its k-mer profile, as well as why a high number of BUSCO sequences are duplicated in this assembly. However, the 31-mer profile of *B. bipinnatifida* does not show four different peaks as observed in recent polyploids datasets (Ranallo-Benavidez et al., 2020). I tried confirming the polyploid origin of *B. bipinnatifida* analysing *B. bipinnatifida* k-mer data on Smudgeplot (Ranallo-Benavidez et al., 2020) but the model also suggested this is a diploid species (plot in appendix 3.A.8). This, however, only indicates that this species has has been diploidised for a while (and lost most of its sub-genome alleles) but this does not disprove the polyploid origin of the species. Additional genomic analysis would be needed in order to prove a Whole Genome Duplication (WGD) in this species compared to

its relatives to confirm the polyploid origin of *B. bipinnatifida*. Moreover, the genome assembly of this species also shows the highest numbers of fragmented and missing BUSCOs, which correlates with the low contiguity of the assembly shown in Table 3.3. This confirms is a highly fragmented, possibly highly repetitive assembly potentially with a high portion of duplicated sequences and unresolved heterozygosity.

Most of our BUSCO results are as good as the ones shown in appendix 3.A.1 for other *Begonia* species. *B. bipinnatifida*, however, shows a much lower number of complete BUSCOs than any other species and the highest number of duplicated ones. Considering *B. darthvaderiana* is also a species from the *Petermannia* section, this indicates the polyploidy event suggested for *B. bipinnatifida* happened after the split of these two species.

3.2.4 Protein-coding gene sets of seven Begoniaceae species

All seven genomes were annotated for gene content using the MAKER-P pipeline (Campbell et al., 2014), and these gene sequences were classified into "bins" or functional groups using MapMan4 (Schwacke et al., 2019). Table 3.5 shows the summary of the gene prediction and functional annotation results.

Table 3.5: Annotation results summary table of the genomes assembled in this chapter. Relative number of scaffolds with genes found, number of genes found as well as the proportion of these with a functional annotation, together with the number of functional groups (bins) found are shown for each species.

Species name	**Annotated scaffolds**	**Annotated scaffolds (%)**	**Genes found**	**Genes annotated (%)**	**Functional groups (bins)**
H. sandwicensis	2,937	64.1	25,165	27.78	2517
B. dregei	22,437	32.5	62,761	28.69	3261
B. conchifolia	3,546	91.9	48,654	45.18	4134
B. socotrana	5,951	28.2	22,588	22.98	1543
B. bipninatifida	61,693	35.2	134,981	29.97	3759
B. johnstoni	6,092	45.5	24,671	42.56	2931
B. luxurians	7,991	39.5	28,656	57.73	3391

This table shows that the percentage of scaffolds where genes were found is highly variable, and much lower in short-read assemblies (28-64%) compared to the *B. conchifolia* genome (92%). This indicates a high number of short-read assembled scaffolds

are either too fragmented to have full gene sequences found, or contain non-resolved repetitive sequences. However, these percentages of annotated scaffolds do not correlate with the number of genes found, which likely means unannotated scaffolds contain mostly unresolved and fragmented sequences of repetitive DNA. The number of genes found oscillates between 22,588 in *B. socotrana* and 134,981 in *B. bipinnatifida*. Most species, however, show a gene content between 22,000 and 28,000 genes, which is supported by the data from Li et al., (in prep) in appendix 3.A.1. Three of our species represent an exception to these values: *B. dregei*, *B. conchifolia* and *B. bipinnatifida* show much higher gene content. The potential hybrid origin of the *B. dregei* specimen sequenced and polyploid origin of *B. bipinnatifida* previously discussed would explain the higher gene content in these two genomes. However, our *B. conchifolia* assembly shows around twice the number of genes than most of the other *Begonia* assemblies. Although this could be associated with assembly artefacts or unresolved heterozygosity, the other sequenced species from the Neotropical Clade II *B. foliosa* shows a similar gene content (appendix 1). Hybridisation is highly common in *Begonia* and it has been studied in Neotropical species, and recent research indicates the possible hybrid origin of Neotropical *Begonia* (Hughes et al., in prep). This would explain the higher gene content of species in this clade compared to African and Asian species.

3.3 Materials and methods

3.3.1 Plant material

Flower and leaf buds of five *Begonia* species (*B. dregei, B. socotrana, B. bipinnatifida, B. johnstonii and B. luxurians*) were collected from the *Begonia* living collections at the Royal Botanic Garden, Edinburgh (RBGE) and kept in silica gel for approximately three months. For additional DNA extractions, fresh flowers and flower buds were collected and frozen in liquid nitrogen, and DNA was extracted on the same day. The *Hillebrandia sandwicensis* silica-dried flower tissue samples were sent by collaborators from the National Tropical Botanical Gardens in Hawaii (USA). Table 3.6 shows the reference accession IDs (if provided), location of the collection where the material was obtained from, type of material used for the DNA extraction selected for sequencing

and DNA sample RBGE reference (EDNA numbers).

Table 3.6: Plant species, accession number and location, type of material and DNA reference number of the plant samples used for whole-genome sequencing.

Taxon name	**Accession ID**	**Collection location**	**Plant material used**	**EDNA ref.**
H. sandwicensis	n/a	NTBG (USA)	Silica-dried	EDNA17-0046719
B. dregei	20030633	RBGE (UK)	Silica-dried	EDNA17-0046714
B. conchifolia	20042082	RBGE (UK)	Fresh	n/a
B. socotrana	20000294	RBGE (UK)	Silica-dried	EDNA17-0046716
B. bipinnatifida	20090801b	RBGE (UK)	Fresh	EDNA17-0046718
B. johnstonii	20130776	RBGE (UK)	Fresh	EDNA17-0046717
B. luxurians	19685494	RBGE (UK)	Silica-dried	EDNA17-0046715

3.3.2 DNA extraction, library preparation and genome sequencing

For the DNA extractions of the seven species for illumina and PacBio platforms, extractions were performed using the QIAGEN DNeasy plant mini-kit using silica gel tissue samples. For cases where a lower quality/quantity of DNA was obtained, DNA extraction from fresh material was performed. DNA samples were quantified using the qubit 2.0 fluorimeter (Thermo Fisher Scientific) and absence of DNA degradation was confirmed by running the DNA on 1% agarose gels.

In some cases, to achieve the desired DNA concentration (an optimal concentration of 37 ng/μl and minimum volume of 60μl was required for PCR-free libraries) DNA from different extractions were pooled together and precipitated using alcohol and then re-suspended. This was done by adding 0.1 volume of 3M sodium acetate (pH5.2), 1 volume of isopropanol and incubating on ice between 15 mins to 1 hour. The sample was then centrifuged at 12,000 xg for 15 minutes and the pellet was washed with 70% ethanol and air-dried. Samples were then re-suspended in the desired amount of TE buffer to achieve the optimal DNA concentration and re-run on a 1% agarose gel to check there had been no DNA degradation during the procedure.

Pair-End (PE) illumina libraries (350bp insert size) for *H. sandwicensis* and *B. dregei, B. socotrana, B. bipinnatifida, B. johnstonii* and *B. luxurians* were obtained using the Tru-Seq DNA PCR-Free (kit and protocol as specified by manufacturer) by Edinburgh Genomics (UK). A PE illumina library (470bp insert size) for *B. conchifolia* was prepared using the Illumina TruSeq DNA kit. Illumina whole-genome sequencing of the

six species sequenced in Edinburgh Genomics was done in 3 lanes of Illumina Hi-SeqX. For *B. conchifolia*, the library was sequenced in a full flowcell of a HiSeq 2500 Illumina system (rapid-run mode).

For the long-read sequencing of *Begonia conchifolia* using PacBio technologies, 4 μg of genomic DNA were sent to Genome Quebec and MacGill University sequencing facility according to their specifications, where a large insert library (SMRTbell) was prepared following the Pacific Bioscience specifications. Our sample was sequenced in 15 SMRT Cells of a PacBio RSII sequencer.

The DNA extraction of *B. conchfiolia* for nanopore sequencing was done using a 3-day protocol optimised by RBGE scientists for High-Molecular-Weight (HMW) plant DNA (Nishii and Hart, available at https://stories.rbge.org.uk/archives/30792).

The DNA library for Nanopore sequencing was obtained using the Ligation Sequencing Kit (SQ-LSK109, Nanopore, UK), and this library was loaded onto a MinION flow cell for sequencing.

3.3.3 Genomic data QC and pre-processing

The quality of the Illumina raw reads was checked using FastQC (Andrews, 2010). After this, reads were trimmed using trimmomatic (Bolger et al., 2014) to remove adapter sequences and low-quality reads (quality threshold: 10). The adapter sequences used for trimming were retrieved from the trimmomatic default Illumina adapter file and the overrepresented sequences found by FastQC. An extra FastQC analysis was run to check the quality of the trimmed reads. Before and after trimming the data, coverage was calculated using the following equation (Lander and Waterman, 1988):

$$C = \frac{L * N}{G}$$

Where C= coverage, L= read length (bp), N= no. reads and G= genome size (bp).

The trimmomatic output files containing both forward and reverse paired reads were then used as input for KmerGenie (Chikhi and Medvedev, 2014). Such software per-

forms k-mer counting using different k-mer sizes, taking from 21 to 121 bp in intervals of 10 by default. Output files were used as input for GenomeScope2 (Ranallo-Benavidez et al., 2020), which uses k-mer count data to adjust the analysed genomes to different models and provides together with the k-mer profiles, estimates of the genome size, % heterozygosity and % of genomic repeats estimation.

PacBio reads were corrected using the Illumina data with the program LoRDEC (Salmela and Rivals, 2014), and nanopore read quality and statistics were visualised using nanoplot (De Coster et al., 2018). Nanopore reads smaller than 500bp were filtered out using Printseq-lite (Schmieder and Edwards, 2011).

3.3.4 Assembly of the *Begonia conchifolia* genome using long-read data

The full pipeline for the *B. conchifolia* genome assembly using PacBio, Nanopore and Illumina is shown on appendix 3.A.4. For this assembly, we tried different long-read assemblers with the PacBio data, such as Sprai (documentation available at sprai-doc.readthedocs.io), Miniasm and five rounds of Racon for consensus calling (Li, 2016; Vaser et al., 2002), Canu (Koren et al., 2017) and Flye (Kolmogorov et al., 2019). All these assemblies were polished using the Illumina data and Pilon (Walker et al., 2014), and the assembly with the best metrics was selected for re-scaffolding with the Nanopore data using SSPACE-longread (Boetzer and Pirovano, 2014) and LRScaf (Qin et al., 2019). LRScaf gave the best assembly, and these scaffolds were again polished with Pilon (Walker et al., 2014) and gaps were closed using GapFiller (Nadalin et al., 2012). Smallest contigs were filtered out (>1000bp) with Printseq-lite (Schmieder and Edwards, 2011).

3.3.5 Assembly of six Begoniaceae genomes using short-read data

The full pipeline for the Illumina genome assemblies of *H. sandwicensis* and *B. dregei, B. socotrana, B. bipinnatifida, B. johnstonii* and *B. luxurians* is shown in appendix 3.A.5. For each of the 7 datasets, I ran contig assemblies using Minia (Chikhi and Rizk, 2013) for 14 different k-mer sizes: from K17 to K121 in intervals of 8 nt (metrics of all these assemblies in appendix 6). I then selected the best contig assemblies for each

species and performed scaffold assembly using 4 different assemblers: SOAPdenovo2 (Luo et al., 2012), Sspace (Boetzer et al., 2011), Redundans (Pryszcz and Gabaldón, 2016) and Pseudohaploid (https://github.com/schatzlab/pseudohaploid). In addition, I ran Platanus-allee (Kajitani et al., 2019) on all the *Begonia* species datasets. Best assemblies were selected and smallest contigs were filtered out (>1000bp) Printseq-lite (Schmieder and Edwards, 2011).

3.3.6 Contaminant removal

The *B. johnstonii* dataset showed a contamination peak (a number of k-mers with a lower coverage than the main peak) on the k-mer profile, and therefore only a preliminary assembly was used (minia contigs, K=121) on blobtoolkit (Challis et al., 2020) for contamination removal (appendix 2). Contigs were filtered by blast hit, GC content and coverage using the blobtools filter command. Sequences that did not have a blast hit with a database of *Begonia* genomes, with a GC content smaller than 0.3 or larger than 0.6, or with a coverage lower than 80x were removed. Filtered reads were used for the final genome assembly.

For all of the seven assembled species, a contaminant removal step was carried out on the final assemblies. Blobplots of all species are shown in appendix 3.A.7.

For the *B. conchifolia* genome assembly, no significant contaminants were found and no sequences were filtered out. For the *H. sandwicensis* blobplot, a previous blobtools version (Laetsch and Blaxter, 2017) was used, and the detected contaminants were removed by aligning the reads to the reference sequences retrieved for a selection of the contaminant taxa. For the rest of the assemblies, contaminant sequences were removed using the blobtools 2.0 filter command; sequences with blast hits other than "Streptophyta" and "no-hit" and with GC content below 0.2 and over 0.6 were filtered out. Remaining scaffolds were then considered to be the final assembly used for downstream analyses.

3.3.7 Gene annotation of seven Begoniaceae species

All seven genomes were annotated for gene content using the MAKER-P pipeline (Campbell et al., 2014). RNA-seq data available from female flowers, male flowers, vegetative buds, mature leaves, roots and petioles of *B. conchifolia* (Emelianova, 2017) and from buds and male flowers of *B. luzhaiensis* (Tseng et al., 2017), as well as the complete protein sets from 4 annotated Begonia genomes (Li et al., in prep) were used by MAKER-P for gene prediction. Repeats annotated in a previous *B. conchifolia* assembly version (Bombarely, unpublished) were used to mask the genome. BUSCO-trained Augustus (Stanke et al., 2008) and three rounds of SNAP training (Korf, 2004) were used for de-novo gene prediction. A basic functional annotation of identified proteins was performed using Mercator4, based on the MapMan4 functional annotation framework (Schwacke et al., 2019).

Bibliography

Andrews S. 2010. Fastqc: a quality control tool for high throughput sequence data.

Barrière A, Yang SP, Pekarek E, Thomas CG, Haag ES, Ruvinsky I. 2009. Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome research*. 19:470–480.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding preassembled contigs using sspace. *Bioinformatics*. 27:578–579.

Boetzer M, Pirovano W. 2014. Sspace-longread: scaffolding bacterial draft genomes using long read sequence information. *BMC bioinformatics*. 15:211.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 30:2114–2120.

Brennan AC, Bridgett S, Ali MS, Harrison N, Matthews A, Pellicer J, Twyford AD, Kidner CA. 2012. Genomic resources for evolutionary studies in the large, diverse, tropical genus, begonia. *Tropical plant biology*. 5:261–276.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using maker and maker-p. *Current protocols in bioinformatics*. 48:4–11.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. Blobtoolkit–interactive quality assessment of genome assemblies. *G3: Genes, Genomes, Genetics*. 10:1361–1374.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*. 30:31–37.

Chikhi R, Rizk G. 2013. Space-efficient and exact de bruijn graph representation based on a bloom filter. *Algorithms for Molecular Biology*. 8:22.

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. Nanopack: visualizing and processing long-read sequencing data. *Bioinformatics*. 34:2666–2669.

Del Angel VD, Hjerde E, Sterck L, et al. (11 co-authors). 2018. Ten steps to get started in genome assembly and annotation. *F1000Research*. 7.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution: International Journal of Organic Evolution*. 63:1–19.

Ellegren H. 2014. Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution*. 29:51–63.

Emelianova K. 2017. Using Next Generation Sequencing to investigate the generation of diversity in the genus Begonia. Ph.D. thesis, The University of Edinburgh.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*. 17:333.

Griesmann M, Chang Y, Liu X, et al. (42 co-authors). 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*. 361.

Kajitani R, Yoshimura D, Okuno M, Minakuchi Y, Kagoshima H, Fujiyama A, Kubokawa K, Kohara Y, Toyoda A, Itoh T. 2019. Platanus-allee is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature communications*. 10:1–15.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*. 37:540–546.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*. 27:722–736.

Korf I. 2004. Gene finding in novel genomes. *BMC bioinformatics*. 5:59.

Laetsch DR, Blaxter ML. 2017. Blobtools: Interrogation of genome assemblies. *F1000Research*. 6:1287.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*. 2:231–239.

Lewin HA, Robinson GE, Kress WJ, et al. (11 co-authors). 2018. Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*. 115:4325–4333.

Li FW, Harkess A. 2018. A guide to sequence your favorite plant genomes. *Applications in Plant Sciences*. 6:e1030.

Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*. 32:2103–2110.

Luo R, Liu B, Xie Y, et al. (11 co-authors). 2012. Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 1:2047–217X.

Maddison WP. 1997. Gene trees in species trees. *Systematic biology*. 46:523–536.

Moonlight PW, Ardi WH, Padilla LA, et al. (29 co-authors). 2018. Dividing and conquering the fastest-growing genus: Towards a natural sectional classification of the mega-diverse genus begonia (Begoniaceae). *Taxon*. 67:267–323.

Nadalin F, Vezzi F, Policriti A. 2012. Gapfiller: a de novo assembly approach to fill the gap within paired reads. *BMC bioinformatics*. 13:S8.

Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. 2015. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. *Systematic Biology*. 64:1000–1017.

Nichols R. 2001. Gene trees and species trees are not the same. *Trends in Ecology & Evolution*. 16:358–364.

Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annual review of genetics*. 34:401–437.

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic acids research*. 44:e113–e113.

Qin M, Wu S, Li A, Zhao F, Feng H, Ding L, Ruan J. 2019. Lrscaf: improving draft genomes using long noisy reads. *BMC genomics*. 20:955.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. Genomescope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nature communications*. 11:1–10.

Salmela L, Rivals E. 2014. Lordec: accurate and efficient long read error correction. *Bioinformatics*. 30:3506–3514.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 27:863–864.

Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, Hallab A, Gruden K, Stitt M, Bolger ME, Usadel B. 2019. Mapman4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Molecular plant*. 12:879–892.

Seppey M, Manni M, Zdobnov EM. 2019. Busco: assessing genome assembly and annotation completeness. In: Gene Prediction, Springer, pp. 227–245.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics*. 24:637–644.

Stevens L. 2020. Genome evolution in the genus Caenorhabditis. Ph.D. thesis, The University of Edinburgh.

Tseng YH, Huang HY, Xu WB, Yang HA, Liu Y, Peng CI, Chung KF. 2017. Development and characterization of est-ssr markers for begonia luzhaiensis (begoniaceae). *Applications in Plant Sciences*. 5:1700024.

Vaser R, Sović I, Nagarajan N, Šikić M. 2002. Racon-rapid consensus module for raw de novo genome assembly of long uncorrected reads. *Bioinformatics*. 18:452–464.

Vinson JP, Jaffe DB, O'Neill K, et al. (11 co-authors). 2005. Assembly of polymorphic genomes: algorithms and application to ciona savignyi. *Genome research*. 15:1127–1135.

Walker BJ, Abeel T, Shea T, et al. (11 co-authors). 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*. 9:e112963.

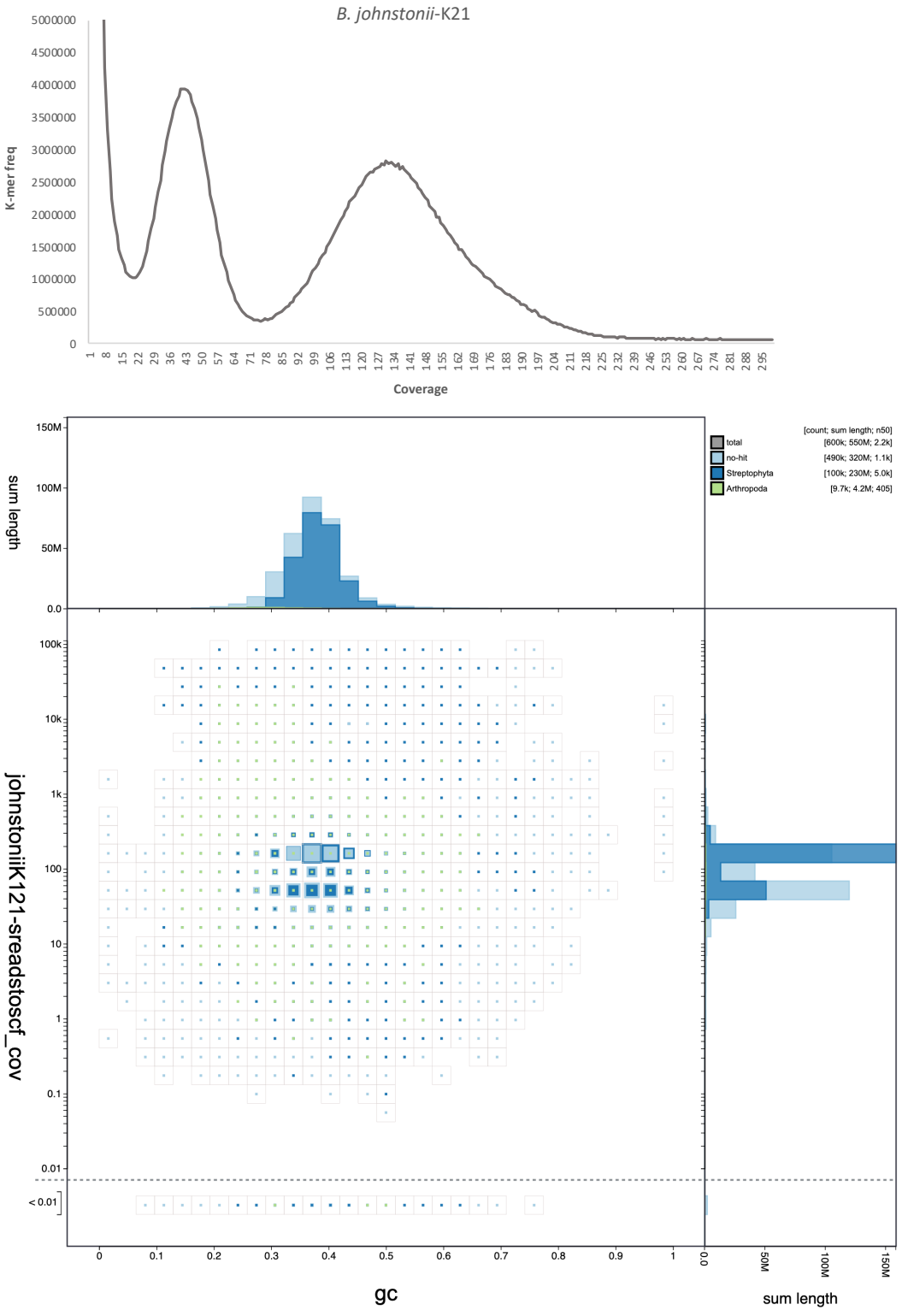
3.A Appendix

3.A.1 Assembly and annotation stats of other Begoniaceae genomes

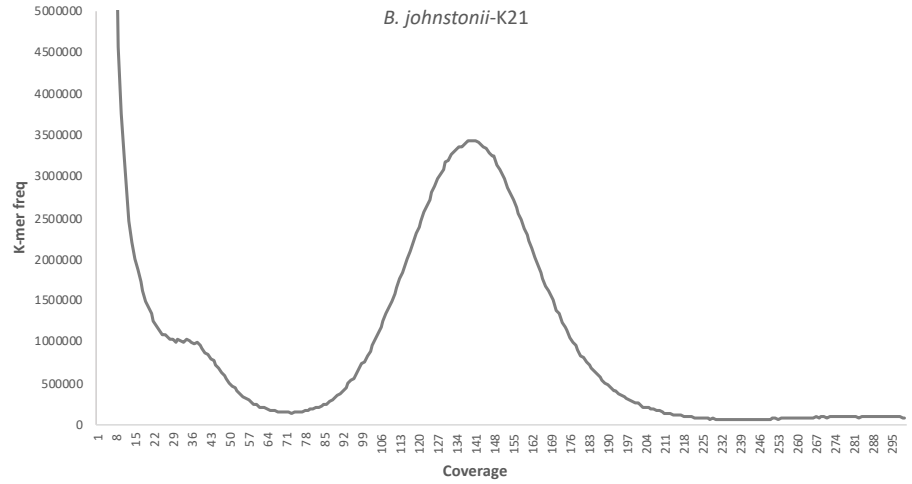
	B. foliosa (var. miniata)	*B. loranthoides*	*B. masoniana*	*B. peltatifolia*	*B. darthvaderiana*
Read type used	short	short	long+short	short	long+short
Final assembly span (Mb)	322.40	707.55	799.83	331.75	786.43
Assembly span (% 1C genome)*	91.82	97.73	120.98	116.17	98.67
No contigs	47,355	16,502	885	7,168	1,787
N50 (kb)	152.57	30,629.29	52,515.07	18,565.98	54,626.12
Longest contig (Mb)	1.90	57.64	84.93	42.45	89.91
% Complete BUSCOs	88.20	97.90	90.80	97.70	92.90
Genes found	44,123	22,059	22,861	23,010	23,444
Annotated genes (%)	63	75	69	73	70
N of bins (functional groups)	2,453	4,092	3,897	4,088	3,931

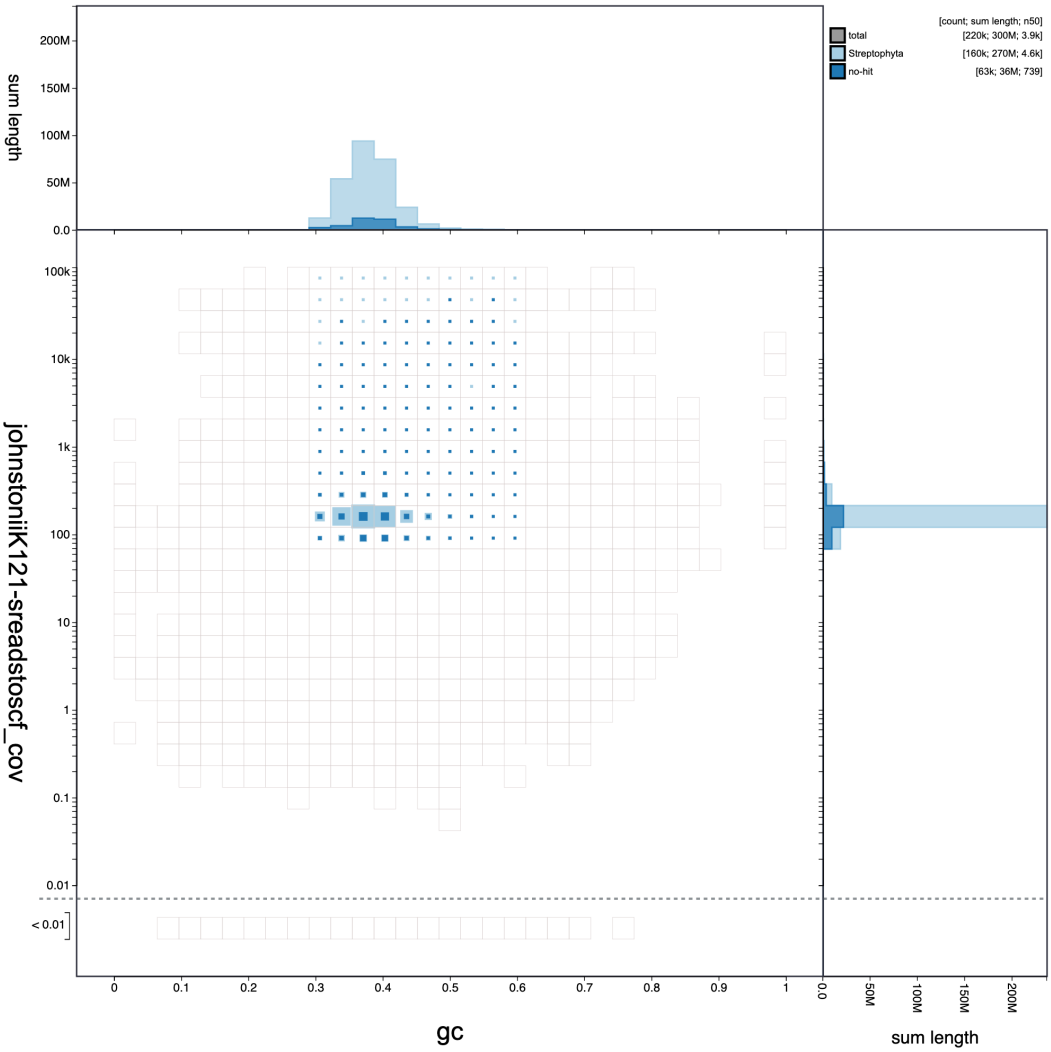
3.A.2 Contaminant removal of the *B. johnstonii* genomic dataset

21-mer graph and blobplot of *B. johnstonii* raw dataset



21-mer graph and blobplot of *B. johnstonii* after filtering

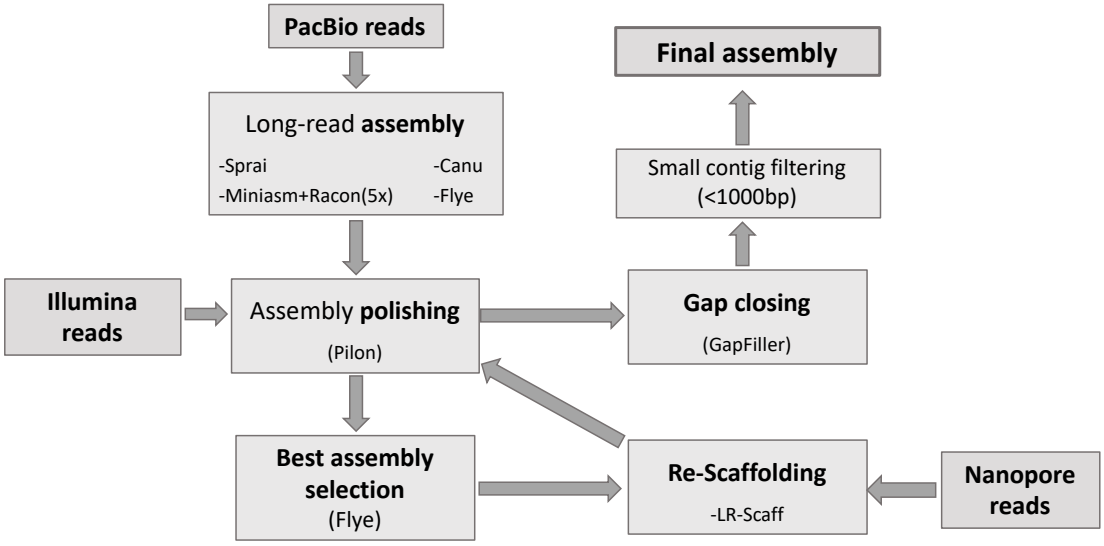




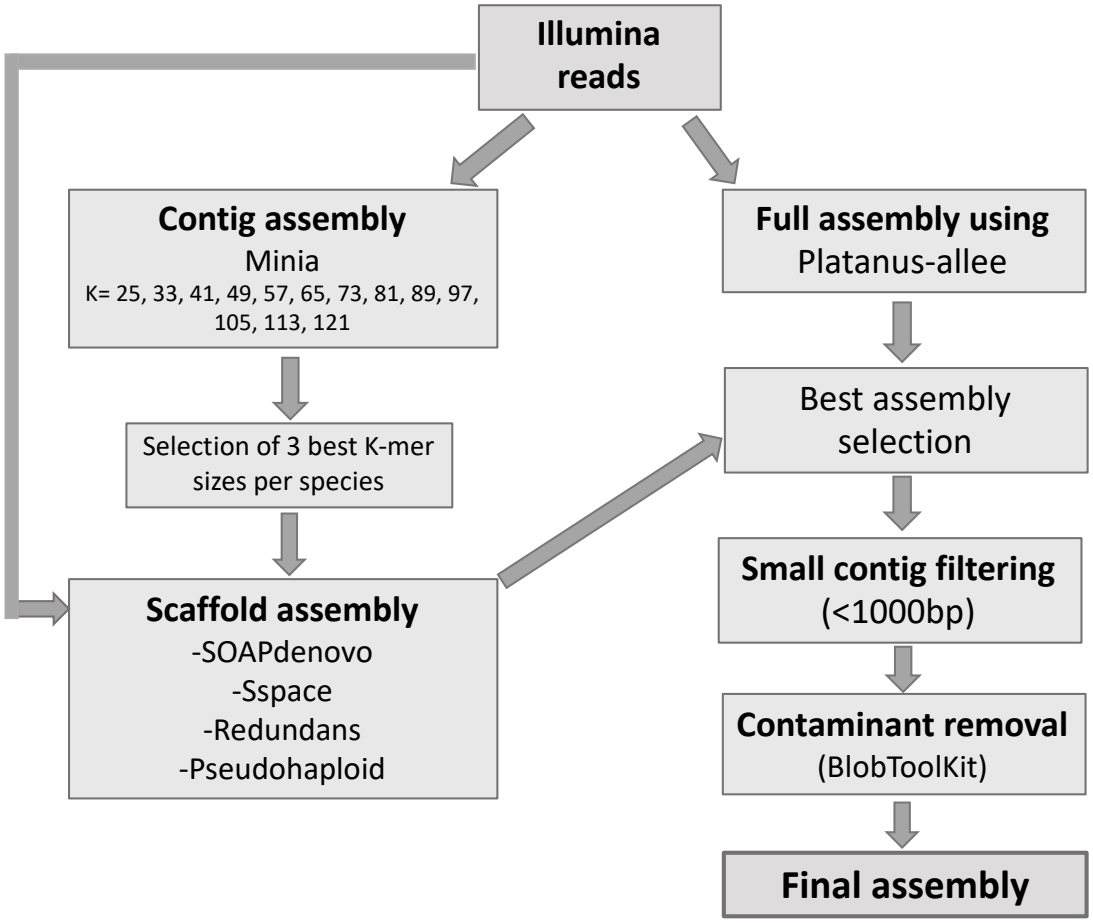
3.A.3 *B. conchifolia* sequencing: long-read datasets statistics

	PacBio	Nanopore
Total reads	1,673,858	568,783
Total (Gb)	13.14	5.41
Avg. length (bp)	7,852	9,513
Median length (bp)	6,035	3,536
N50 (bp)	12,754	25,920

3.A.4 Assembly pipeline of the assembly of *Begonia conchifolia* using long and short reads



3.A.5 Assembly pipeline of the assembly of 6 Begoniaceae genomes using short reads



3.A.6 Assembly metrics of all the different assemblies performed

Assembly metrics of all the Pacbio assemblies performed

	Sprai (Becon104)	miniasm (no polishing)	miniasm + racon04	minisasm + racon04 + pilon	canu	canu + pilon	flye	flye + pilon	canu - run2	canu2 + pilon
size (Mbp)	736.0	491.6	485.0	484.5	486.0	487.2	498.2	498.9	513.9	515.1
% genome	125.4	83.8	82.7	82.6	82.8	83.0	84.9	85.0	87.6	87.8
no scaffolds	25304	5074	5014	5014	9110	9110	6066	6066	8117	8117
Longest sequence (Mb)	0.6	1.0	1.0	1.0	0.7	0.7	1.1	1.1	0.9	0.9
N50 (kb)	61.3	144.6	143.2	143.0	93.6	93.9	192.3	192.8	124.1	124.4
BUSCOS (%complete)	68.0	0.0	36.3	67.0	49.5	72.3	59.0	78.7	47.5	78.6

Assembly metrics of all the re-scaffolding, gap closing and polishing with Nanopore datasets

	SSPACE-longread	Lrscaf	Lrscaf+ GapFiller	Lrscaf+ GF+ Pilon	Lrscaf+GF+ Pilon+ filtered1000
size (Mbp)	499.1	565.8	565.8	565.0	564.9
% genome	85	96	96	96	96
no scaffolds	6031.0	4068.0	4068.0	4068.0	3860.0
Longest sequence (Mb)	1.1	1.8	1.8	1.8	1.8
N50 (kb)	194.1	314.6	314.6	314.6	314.6
BUSCOS (%complete)	-	-	-	-	95.1

Assembly metrics of all the non-selected contig assemblies performed using short-read data

H. sandwicensis	**K17**	**K25**	**K33**	**K41**	**K49**	**K57**	**K65**	**K73**	**K81**	**K89**	**K97**	**K105**	**K113**	**K121**
size (Mbp)	280.7	332.3	292.4	278.9	272.5	268.5	265.6	263.4	259.5	256.9	255.2	254.8	254.9	254.9
% genome	84.4	99.9	87.9	83.9	81.9	80.7	79.8	79.2	78.0	77.2	76.7	76.6	76.7	76.6
no contigs	158884401	5354703	2353602	1314345	821932	555217	399943	302931	222857	168327	130651	104901	88866	79120
Longest sequence (bp)	186	19285	286210	327542	329249	352882	328534	155795	225779	323456	323585	323626	235973	166315
N50 (bp)	16	122	350	1053	2533	4795	7877	11336	15428	19116	22184	23770	23238	19589
B. dregei	**K17**	**K25**	**K33**	**K41**	**K49**	**K57**	**K65**	**K73**	**K81**	**K89**	**K97**	**K105**	**K113**	**K121**
size (Mbp)	3686.6	501.4	515.0	558.0	613.1	670.2	722.7	767.8	804.2	830.4	840.3	823.0	761.9	636.3
% genome	579.9	78.9	81.0	87.8	96.4	105.4	113.7	120.8	126.5	130.6	132.2	129.5	119.9	100.1
no contigs	208760673	11032995	7276203	5611630	4702743	4110905	3671319	3308274	2988989	2699537	2413217	2099674	1751130	1482529
Longest sequence (bp)	158	9655	13704	14797	17690	17841	17656	22201	22217	23231	21185	19733	15737	22776
N50 (bp)	17	43	82	133	172	211	252	296	343	393	446	499	524	444
B. socotrana	**K17**	**K25**	**K33**	**K41**	**K49**	**K57**	**K65**	**K73**	**K81**	**K89**	**K97**	**K105**	**K113**	**K121**
size (Mbp)	3284.66	585.76	520.85	451.61	430.40	426.37	426.12	428.74	432.91	436.42	438.44	436.80	427.28	399.06
% genome	707.06	126.09	112.12	97.22	92.65	91.78	91.73	92.29	93.19	93.94	94.38	94.03	91.98	85.90
no contigs	185834792	14347761	7577911	3650057	2161995	1515079	1164610	972216	861931	787464	727232	663631	585136	487415
Longest sequence (bp)	197	16005	25472	42475	57361	59569	71546	77764	96307	101843	97152	173722	115691	138312
N50 (bp)	17	31	86	247	504	749	967	1120	1227	1320	1396	1492	1616	1674
B. bipinnatifida	**K17**	**K25**	**K33**	**K41**	**K49**	**K57**	**K65**	**K73**	**K81**	**K89**	**K97**	**K105**	**K113**	**K121**
size (Mbp)	-	1240.1	1075.0	1100.4	1141.0	1178.3	1201.2	1218.5	1225.3	1232.2	1238.0	1241.7	1227.3	1149.2
% genome	-	109.3	94.8	97.0	100.6	103.9	105.9	107.4	108.0	108.6	109.1	109.5	108.2	101.3
no contigs	-	28454587	14895148	11252712	9100258	7556371	6309968	5381122	4667042	4125118	3681568	3305129	2943254	2501441
Longest sequence (bp)	-	63698	79870	174109	174897	166198	224252	393315	398634	567228	360949	359207	159654	38631
N50 (bp)	-	39	98	146	176	207	242	283	331	387	451	524	598	643
B. johnstonii	**K17**	**K25**	**K33**	**K41**	**K49**	**K57**	**K65**	**K73**	**K81**	**K89**	**K97**	**K105**	**K113**	**K121**
size (Mbp)	6828.3	465.2	385.9	361.2	357.3	360.5	368.3	375.6	375.3	372.9	368.8	363.8	355.4	324.7
% genome	1837.3	125.2	103.8	97.2	96.1	97.0	99.1	101.1	101.0	100.3	99.2	97.9	95.6	87.4
no contigs	394097468	9754283	4253192	2295152	1471353	1072005	894965	791389	676462	586025	507080	438084	371409	235699
Longest sequence (bp)	135	14082	30927	26543	29638	29648	45271	52063	69088	106679	106695	115738	174081	174097
N50 (bp)	17	48	181	432	757	1099	1374	1612	1935	2307	2751	3311	4357	10745
B. luxurians	**K17**	**K25**	**K33**	**K41**	**K49**	**K57**	**K65**	**K73**	**K81**	**K89**	**K97**	**K105**	**K113**	**K121**
size (Mbp)	3696.6	311.9	286.9	286.1	306.2	305.5	315.6	324.3	331.3	336.3	350.7	346.9	352.1	350.4
% genome	1079.9	91.1	83.8	83.6	89.4	89.3	92.2	94.7	96.8	98.2	102.4	101.3	102.9	102.4
no contigs	209825675	4612039	2304010	1562277	1507398	1144052	1040699	952185	879663	803432	829293	693371	657027	601018
Longest sequence (bp)	201	13299	29953	43036	65300	65316	67667	96978	103614	107029	107040	95624	82814	84645
N50 (bp)	17	147	287	387	387	454	486	529	581	643	669	764	814	875

Assembly metrics of all the non-selected scaffold assemblies performed using short-read data

H. sandwicensis	K25-sspace	K25-SDN	K81-sspace	K81-SDN	K105-sspace	K105-SDN	K113-SDN
size (Mbp)	331.0	181.8	259.0	250.9	254.6	251.9	252.2
% genome	99.5	54.7	77.9	75.4	76.5	75.8	75.8
no scaffolds	5251972	282149	180222	95766	72544	65996	54141
.ongest sequence (kbp)	35.8	235.6	346.6	471.6	542.8	494.8	658.3
N50 (bp)	121	5175	37011	50327	54267	70716	76530

B. dregei	K57-sspace	K57-SDN	K57-redundans	K89-sspace	K89-SDN	K89-redundans	K121-sspace	K121-SDN	K121-redundans
size (Mbp)	665.2	524.0	179.0	827.4	754.1	239.1	633.8	579.1	221.9
% genome	104.6	82.4	28.2	130.2	118.6	37.6	99.7	91.1	34.9
no scaffolds	3980936	1327783	165632	2217510	1554676	260427	924559	762635	266331
Longest sequence (bp)	57405	45265	25802	60047	68325	25690	66996	40719	58602
N50 (bp)	211	810	1968	780	1113	1371	1050	1214	1100

B. socotrana	K41-sspace	K41-SDN	K41-redundans	K89-sspace	K89-SDN	K89-redundans	K105-sspace	K105-SDN	K105-redundans
size (Mbp)	449.3	312.4	171.7	435.3	415.0	227.1	435.9	419.8	233.1
% genome	96.7	67.3	37.0	93.7	89.3	48.9	93.8	90.4	50.2
no scaffolds	3546505	437333	60204	580430	453395	55193	440415	435330	66786
Longest sequence (bp)	82540	167338	60997	301625	496007	165189	441577	474904	241775
N50 (bp)	267	7731	6709	3758	7499	12646	4317	6720	12187

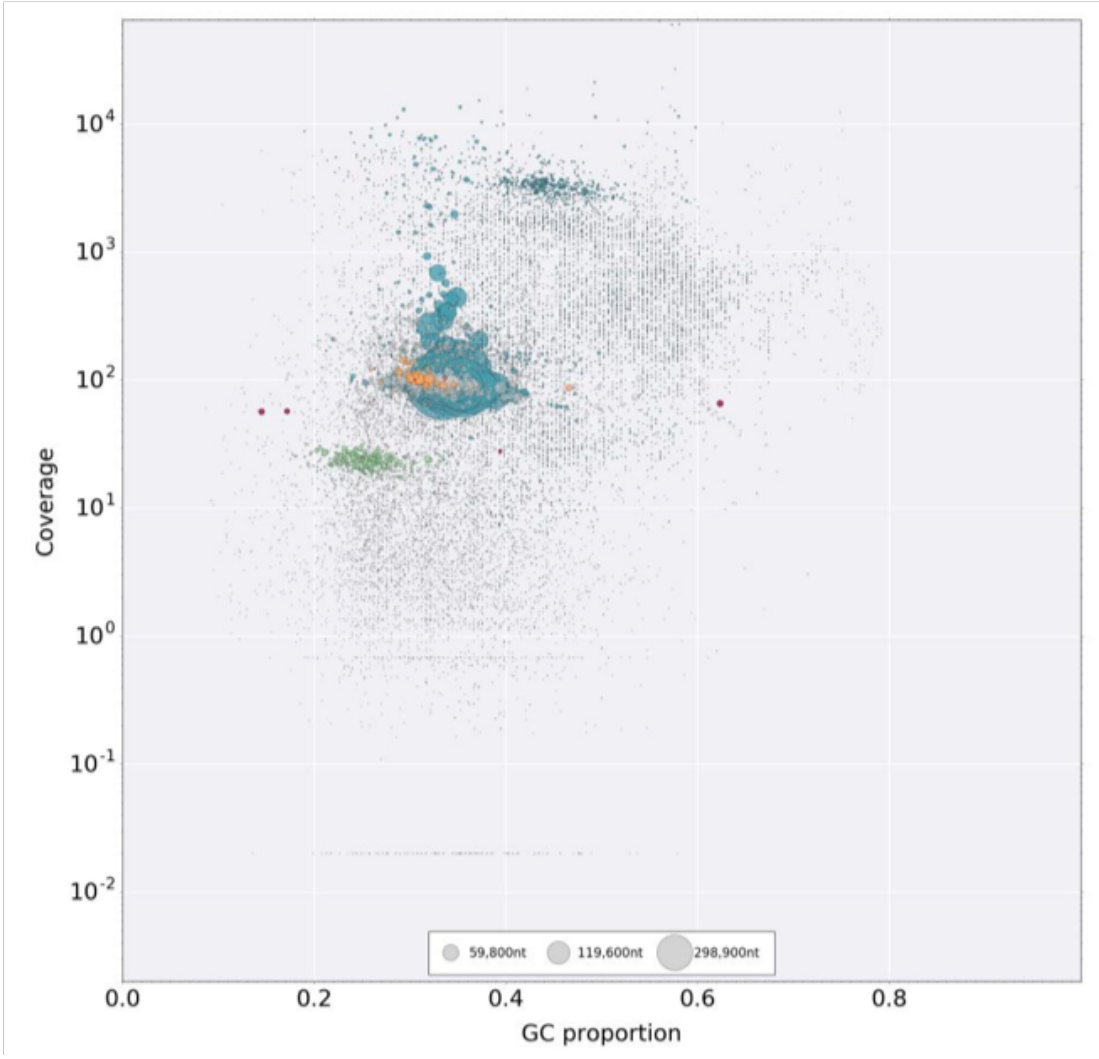
B. bipinnatifida	K49-sspace	K49-SDN	K49-redundans	K89-sspace	K89-SDN	K89-redundans	K105-sspace	K105-SDN
size (Mbp)	1134.9	788.7	319.4	1225.9	1122.4	394.7	1236.6	1176.4
% genome	100.0	69.5	28.1	108.1	98.9	34.8	109.0	103.7
no scaffolds	8889000	2208795	462472	3333482	2519071	485026	2458344	2404170
Longest sequence (bp)	174897	738379	174897	567228	2792252	567228	359207	1213800
N50 (bp)	176	582	1218	962	1166	1457	1650	1627

B. johnstonii	K105-SDN	K113-SDN	K121-SDN
size (Mbp)	686.5	629.4	563.8
% genome	184.7	169.4	151.7
no scaffolds	2170649	1691478	1269547
Longest sequence (bp)	339941	432407	432654
N50 (bp)	565	1096	1596

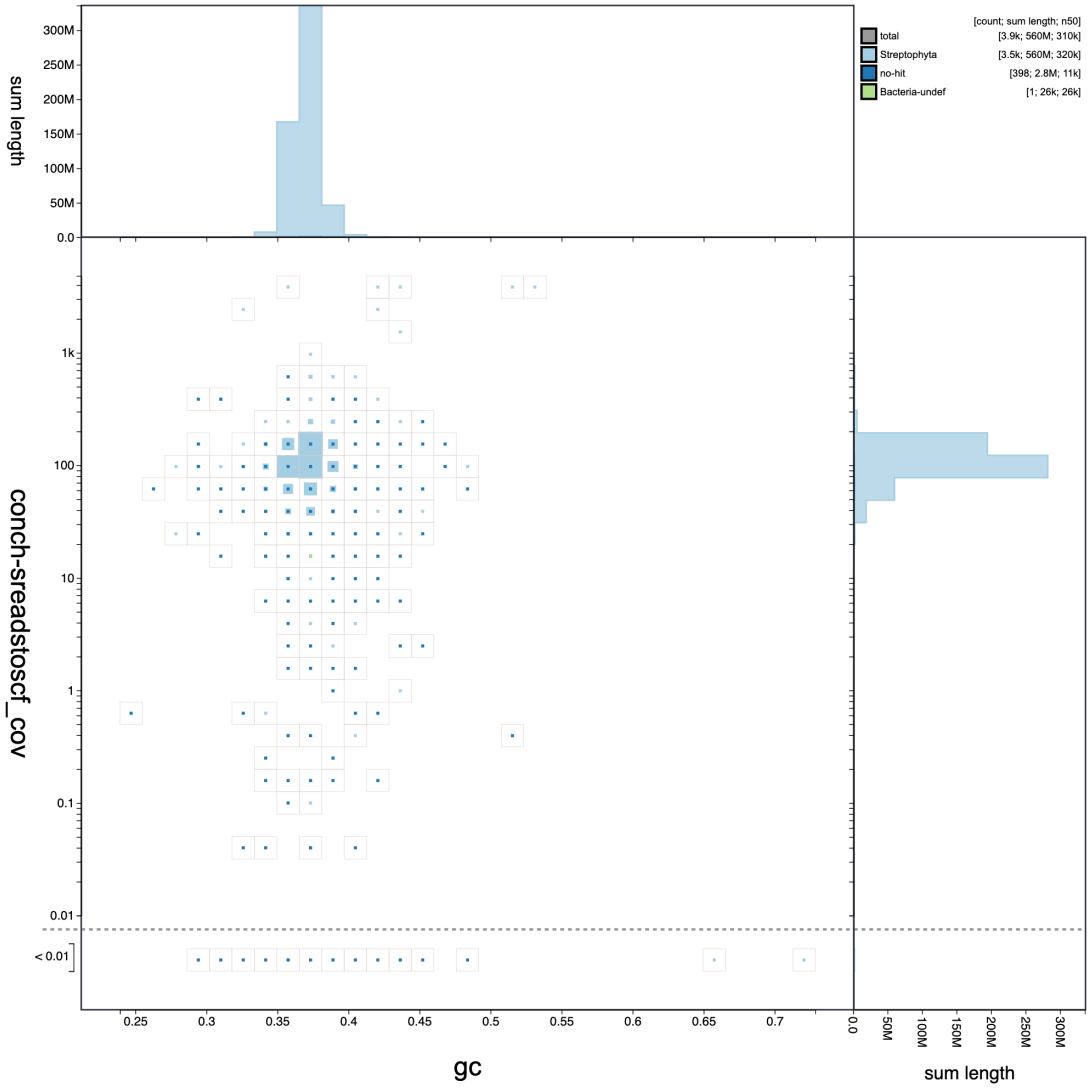
B.luxurians	K73-sspace	K73-SDN	K73-pseudohapl	K73-redundans	K81-sspace	K81-SDN	K81-pseudohapl	K81-redundans	K89-sspace	K89-SDN	K89-pseudohapl	K89-redundans
size (Mbp)	322.2	296.9	296.7	171.8	329.5	306.2	306.0	174.3	334.6	314.8	314.6	176.2
% genome	94.1	86.7	86.7	50.2	96.2	89.4	89.4	50.9	97.8	92.0	91.9	51.5
no scaffolds	717976	493225	493186	70128	619909	490319	490281	67016	529984	487166	487124	63225
Longest sequence (bp)	185777	249389	249389	120054	185456	307220	307220	113617	185370	248258	248258	127171
N50 (bp)	1813	1948	1946	7396	2201	1954	1952	7280	2555	2008	2005	7183

3.A.7 Contamimant removal from final assemblies

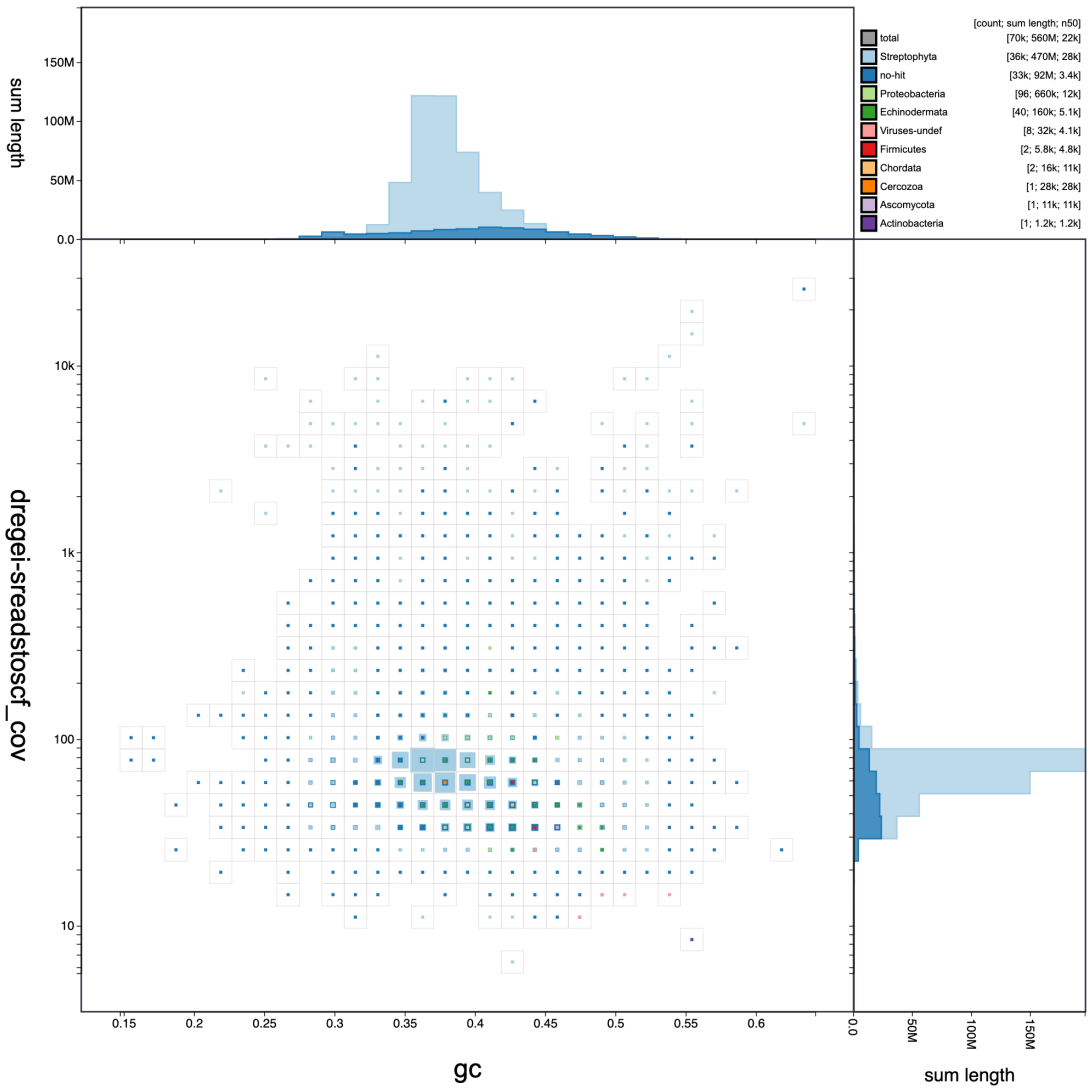
H. sandwicensis **blobplot**



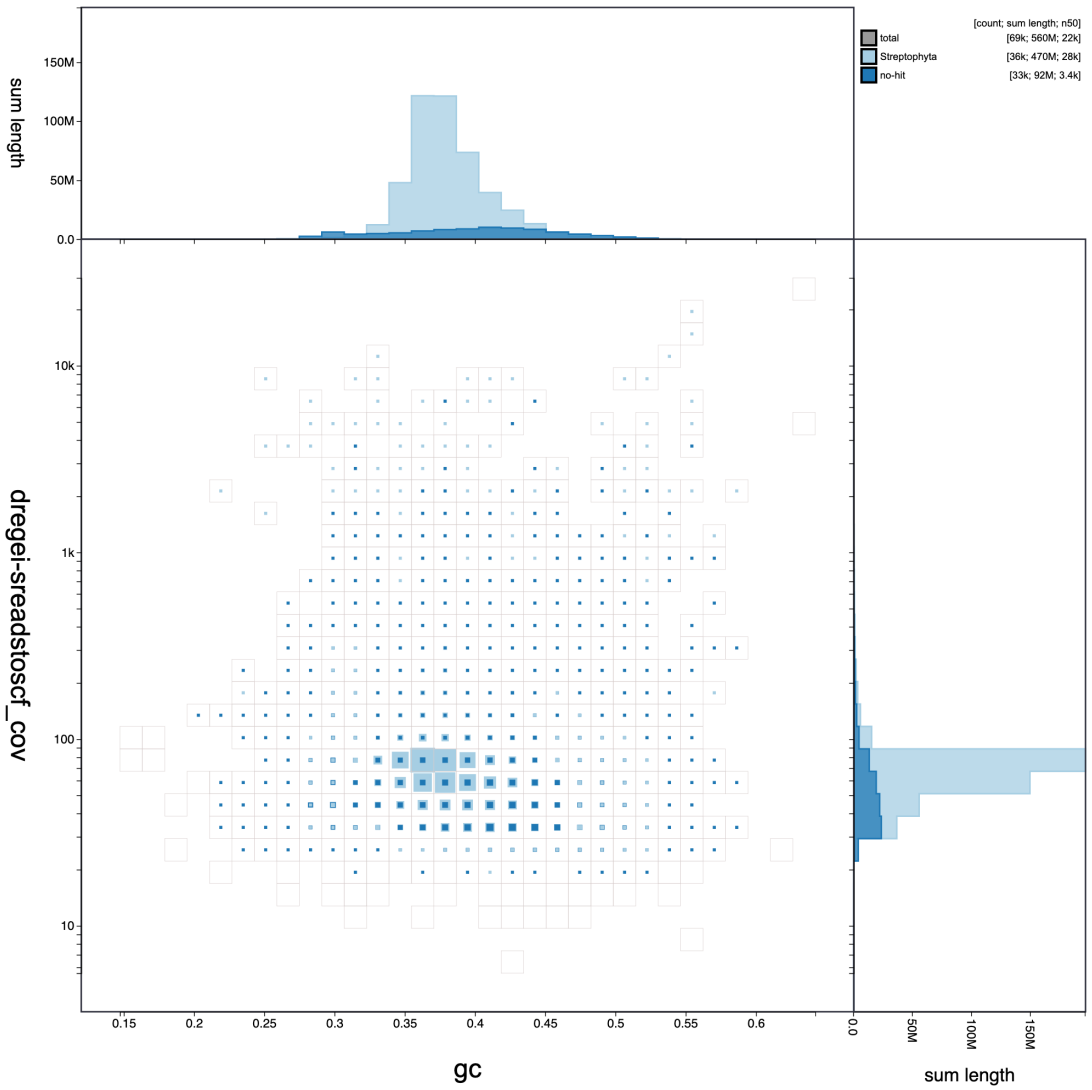
B. conchifolia blobplot



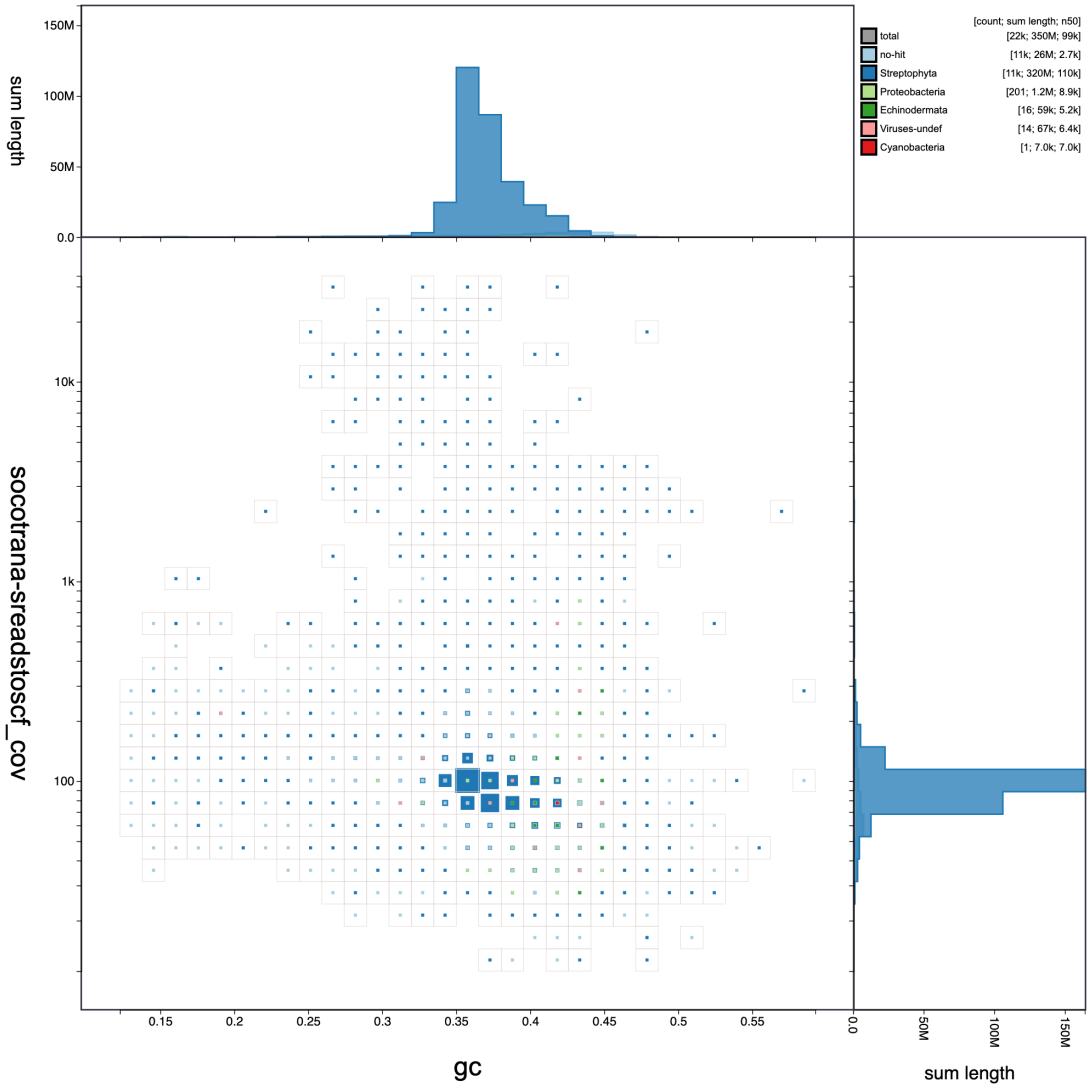
B. dregei blobplot before filtering



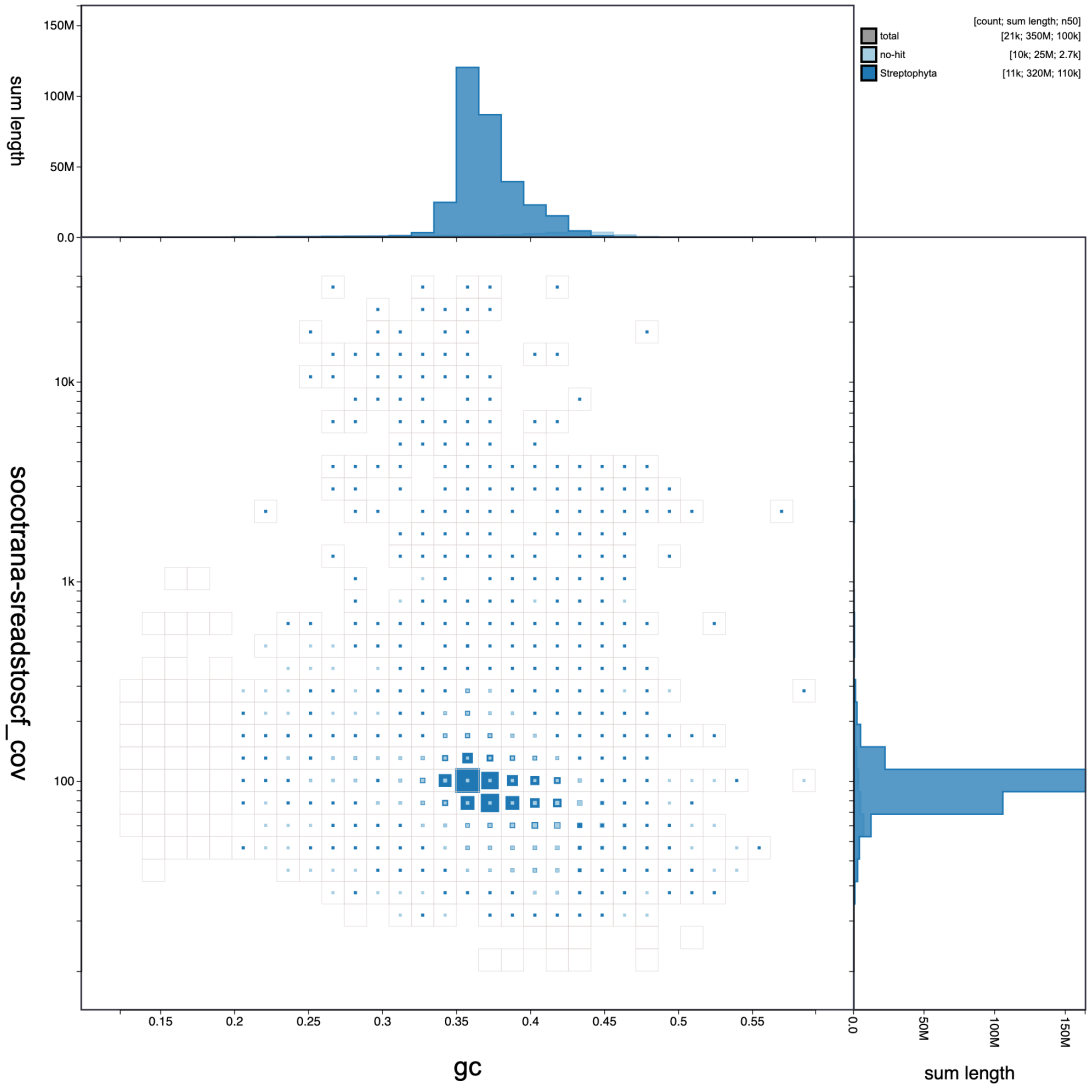
B. dregei blobplot after filtering



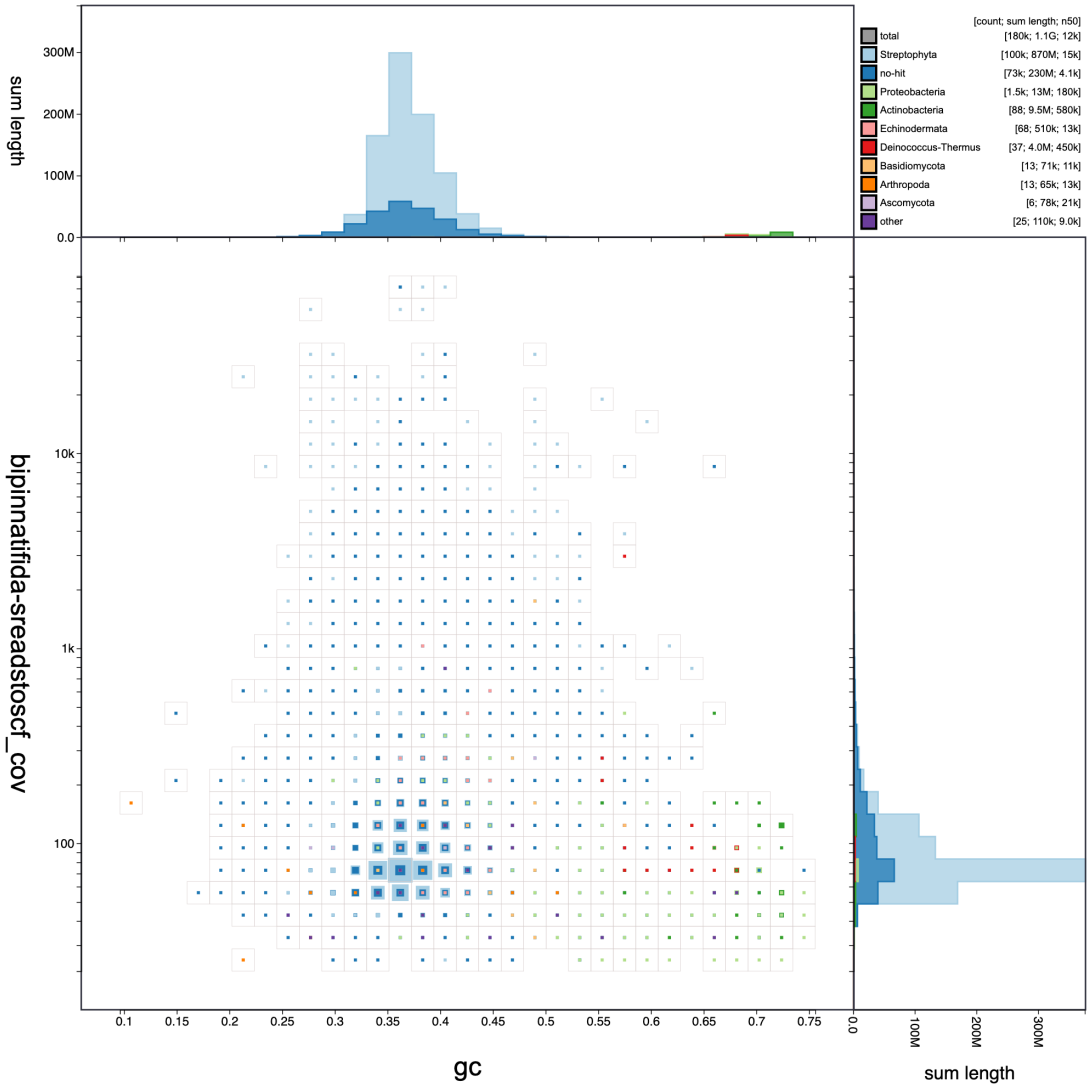
B. socotrana blobplot before filtering



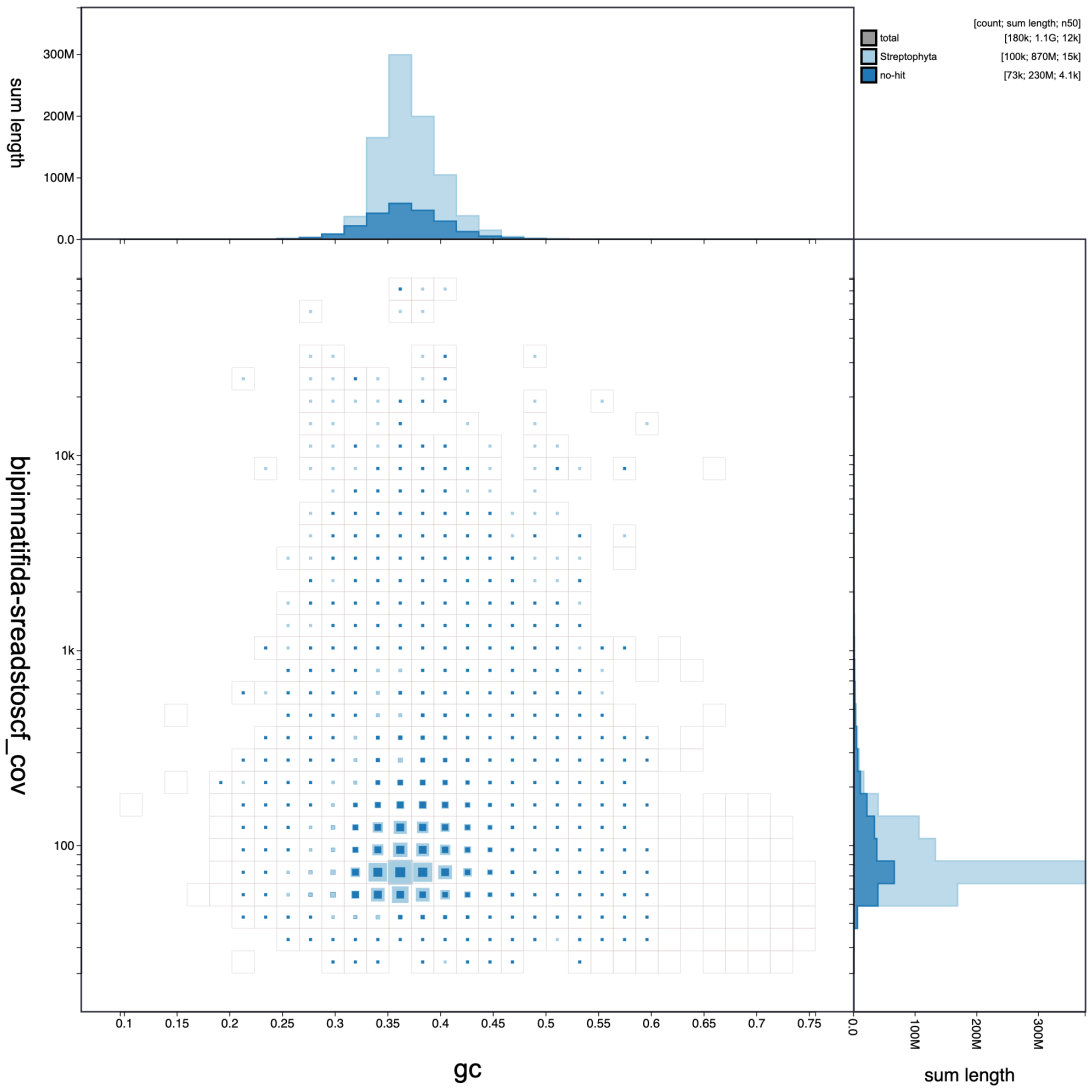
B. socotrana blobplot after filtering



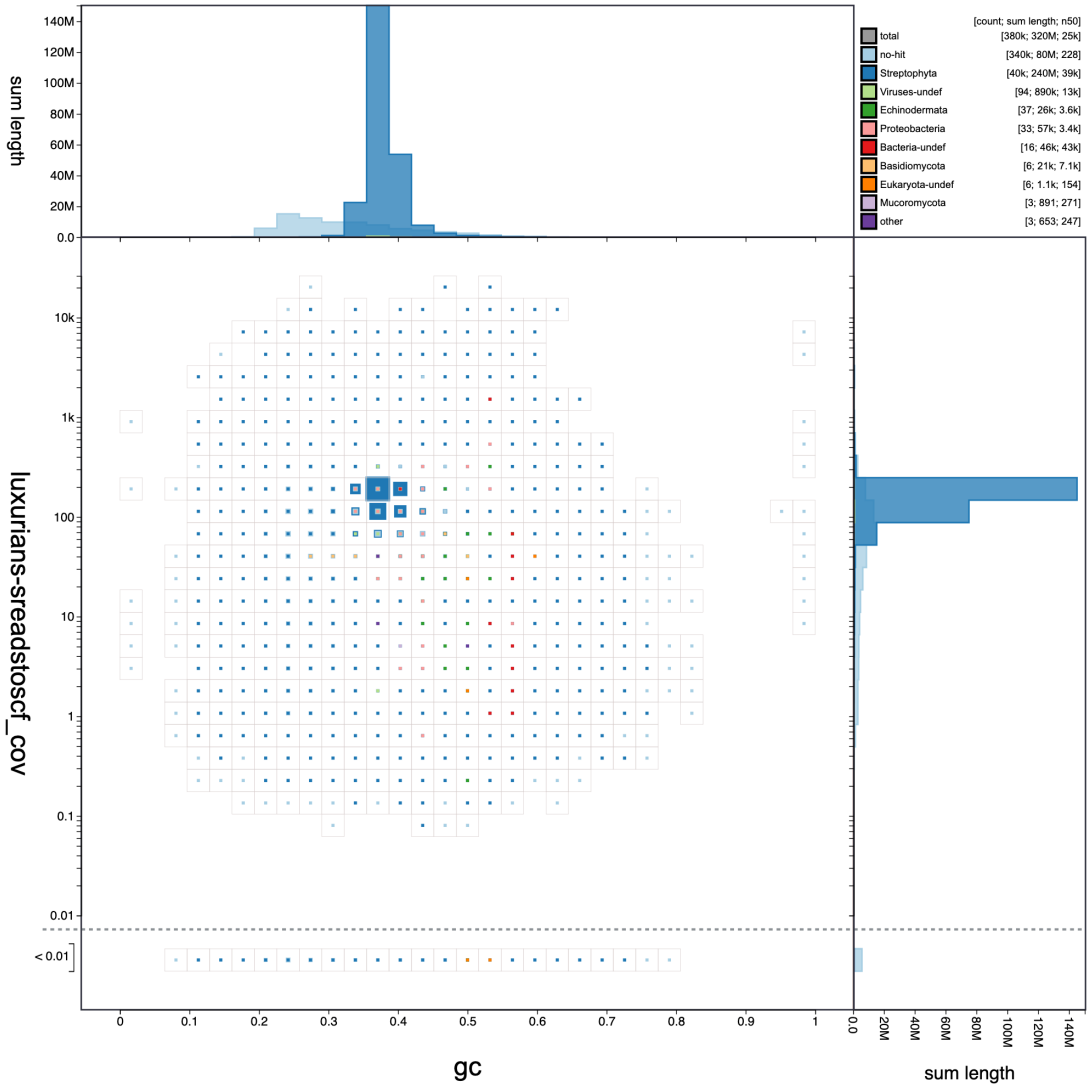
B. bipinnatifida blobplot before filtering



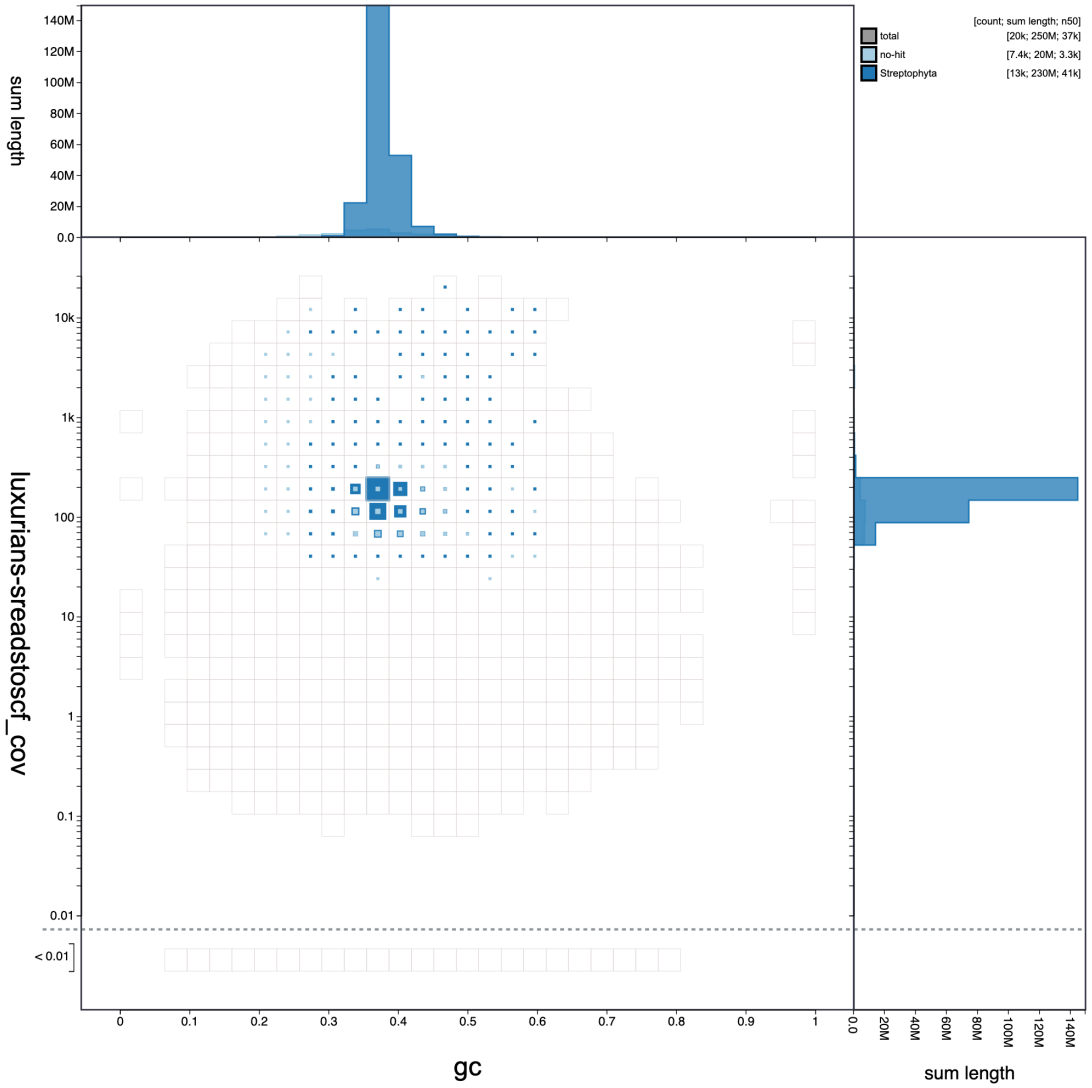
B. bipinnatifida blobplot after filtering



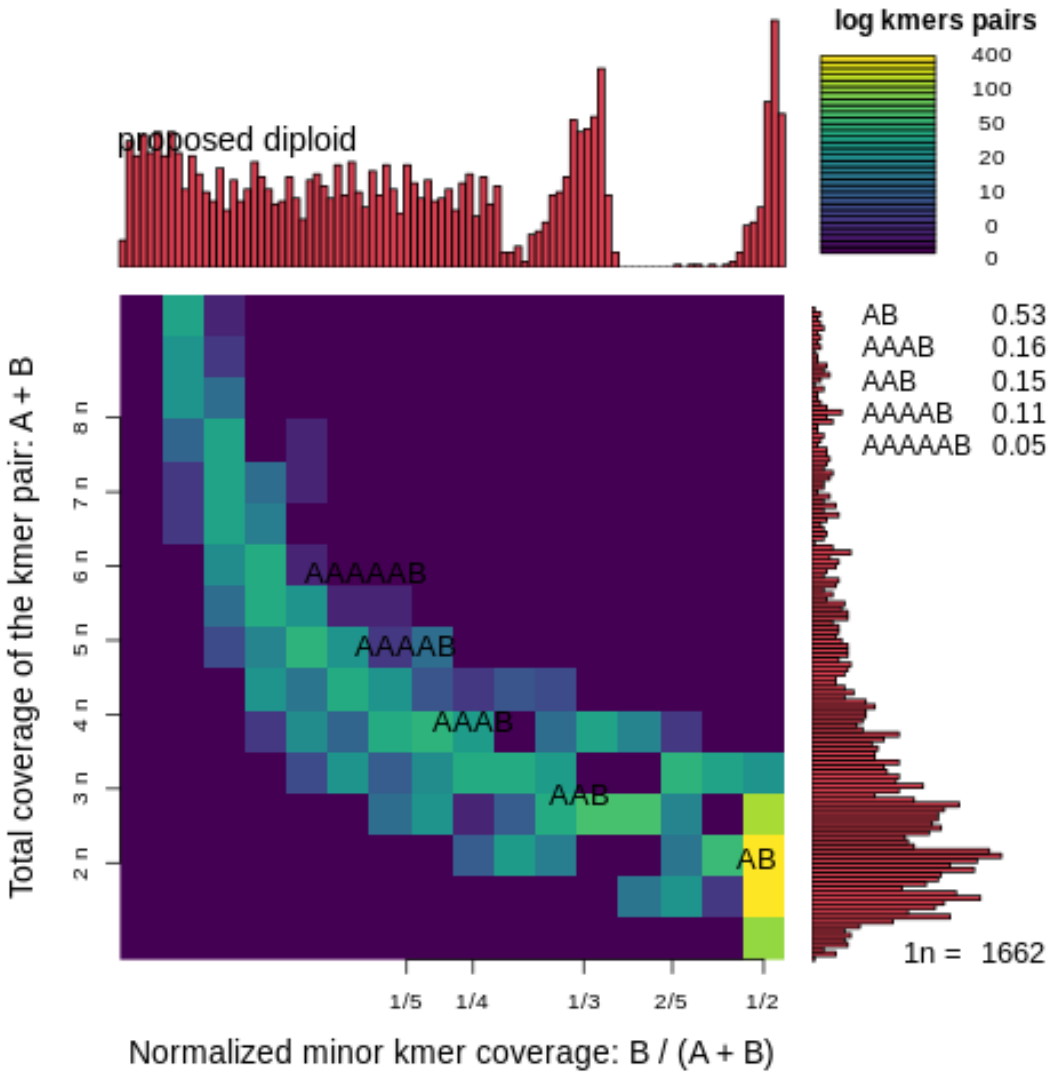
B. luxurians blobplot before filtering



B. luxurians blobplot after filtering



3.1.8 Smudgeplot of *B. bipinnatifida*



Chapter 4

An in-depth analysis of the DNA repeat content and dynamics across Begoniaceae

4.1 Introduction

As described in Chapters 1 and 2, genome size in plants is highly variable. This is largely due to a combination of genome multiplication events and variation in the number of repetitive DNA elements (Neumann et al., 2006; Piegu et al., 2006). In plants, repetitive DNA can comprise anywhere from around 14% of the genome in *Arabidopsis thaliana* (Kaul et al., 2000), to over 80% in maize (Schnable et al., 2009).

As previously introduced in Chapter 1, the repeatome can be broadly classified into satellite DNA (short tandem repeats), normally localised at specific chromosomal sites, and Transposable Elements (TEs or transposons), more interspersed and with their own transposition mechanism that moves or copies them using specific enzymes (transposases). Some plant species can accumulate for as much as 10–20% of satellite repeats in their genome (Ingham et al., 1993; Kelly et al., 2015), but the main bulk of genomic repeats is usually made up of TEs (Tenaillon et al., 2011; Novák et al., 2014; Piednoël et al., 2012; Macas et al., 2011, 2007). There are several types and families of TEs depending on their origin and transposition mechanism, each being present or more abundant in different major groups of organisms (Wicker et al., 2007). In plants, the most abundant groups of TEs are LTRs (Long-Terminal-Repeats), a Class I (RNA-mediated) type of TE (Neumann et al., 2006; Piegu et al., 2006). Their size ranges go from a few hundred base pairs to over 10 kb (Bennetzen, 2000). In maize, LTRs alone make up to between 49 and 78% of the genome length (Levin, 2002).

Most plant LTR retrotransposons belong to two superfamilies: Ty1-*Copia* (also LTR-*Copia* or *Copia*) elements and Ty3-*Gypsy* (also LTR-*Gypsy* or *Gypsy*) elements (Wicker et al., 2007; Bennetzen and Wang, 2014). These are distinguished by the organization of their internal domains, the coding regions for retrotransposition proteins (Figure 4.1, Wicker et al. (2007)). They also differ in their genomic niches: LTR-*Gypsy* elements insert more frequently in heterochromatic regions (Presting et al., 1998; Staton et al., 2009; Cavallini et al., 2010; Cossu et al., 2012), while LTR-*Copia* elements have a less specific insertion pattern and therefore they are more often found closely associated with genes (White et al., 1994; Kumar et al., 1997; Pereira, 2004; Lockton and Gaut, 2009; González and Deyholos, 2012).

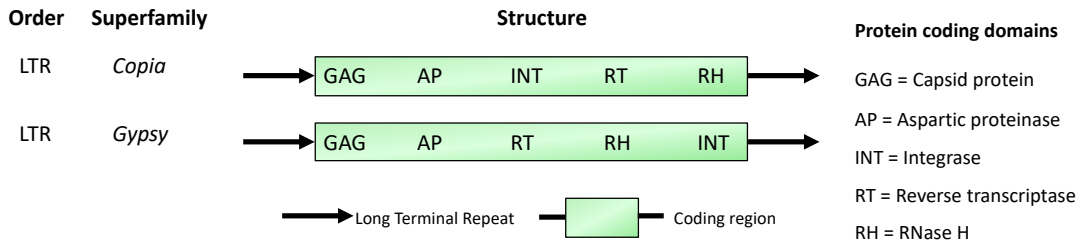


Figure 4.1: Schematic structure of the two LTR superfamilies present in plant genomes. The difference between the two resides in the order of the coding domains between the LTR regions. The coding regions are known to be better conserved than the LTR sequences, which are responsible for the known LTR variation in length. Figure modified from (Wicker et al., 2007)

The role of Transposable Elements in genome evolution at different levels was also introduced in Chapter 1. Their activation can facilitate chromosomal rearrangements such as deletions, translocations and inversions. At a smaller scale, transposons can cause point mutations by inserting in genic regions. They can also determine heterochromatin patterns and therefore their activity can modify gene expression through epigenetic regulation (Lippman et al., 2004). Studying and comparing genome structure enables understanding the distribution and diversity of repetitive elements and their impact on genome evolution. Finding out how they vary between species is essential for the understanding of the evolutionary consequences of genome dynamics (da Costa et al., 2019).

Comparison of repeats between species is complicated by the difficulties in assembling genomic repeats. Such sequences are often unresolved in short-read assemblies, and can be the cause of assembly gaps (Peona et al., 2018). A comparison of repetitive DNA patterns between genome assemblies of varying levels of completeness and contiguity can be affected by different biases, created by differences in sequencing method (length and accuracy of reads), sample quality, library preparation protocols and assembly strategies. Identification and classification of repeats and TEs must be as accurate as possible to allow comparison between species and the analysis of their evolutionary role (Wierzbicki et al., 2020).

Although the number of publicly available chromosome level assemblies is increasing (Kelly, 2019), these are often single genomes used as model for a specific taxonomic

group. As discussed in Chapter 3, we have invested effort and resources in sequencing a number different *Begonia* species. However, our assemblies are of very different quality and contiguity. This may cause difficulties in comparison of genome structure between the species. To control for assembly issues a second, assembly free approach to TE analysis was also taken in this chapter.

A potential alternative to full genome sequencing and assembly for the study of TEs is the use of the Repeatexplorer2 pipeline (Novak et al., 2013). This tool performs all-to-all comparisons and clustering analyses of low coverage short-read datasets, and identifies the most abundant genomic repeats (Renny-Byfield et al., 2013; Dodsworth et al., 2015; Mascagni et al., 2017, 2020). By only using a small subset of reads (<1x coverage), this pipeline represents a time and cost-effective method for the identification of genomic repeats and the study of repetitive DNA landscapes across a range of organisms (Macas et al., 2015). However, by itself it cannot provide information on where these repeats are inserted or the specific number of copies of a LTR lineage in the genome of interest.

In this chapter I use both approaches, compare them and discuss their advantages and disadvantages in addressing the question of whether TE dynamics have contributed to *Begonia* evolution. Specifically, the objectives of this chapters are:

- To annotate and compare repeats in the 12 Begoniaceae genome assemblies described in Chapter 3. Understanding and comparing the genomic repeat landscapes of the species across Begoniaceae will help us understand which TE lineages are most active and how this varies across this family. With this aim, two different approaches are taken: repeat annotation of the genome assemblies from Chapter 3, and repeat clustering from short-read non assembled data using RepeatExplorer2 of all those species we have raw data from. The comparison of these two approaches will also allow to evaluate completeness and contiguity of our genome assemblies.

- To compare patterns of LTR activity across the sequenced species. The aim is to study the distribution of insertion dates for LTR lineages of interest using two

different methods based on results from the two different approaches taken for LTR analysis. Using the full-length annotated LTRs, I aim to date their insertion times by comparing their two LTR flanking sequences (SanMiguel et al., 1998). In addition, using the repeat clustering analysis results, I aim to calculate the RT-coding sequence divergence between different copies of specific LTR lineages.

4.2 Results and Discussion

4.2.1 Repeat annotations

Repeat annotation of the 12 Begoniaceae genome assemblies

A standard repeat annotation procedure was implemented on the 12 Begoniaceae genome assemblies by obtaining a repeat library from RepeatModeler (Flynn et al., 2020) using these 12 assemblies, which then was used to mask such repeats in the assemblies (Smit et al., 2019). Results of all the masked repeats and the genome proportion they represent are shown in Figure 4.2 (data table in appendix 4.A.1). Figure 4.2 shows variation in repeat types and content across the annotated genomes. The total amount of masked repeats go from 30.61% in *B. luxurians* to 73.6% in *B. darthvaderiana*. These values change according to the size of the genome assemblies, and therefore with their genome sizes (from Chapter 2, also indicated in Figure 4.2). This means the larger the genome, the larger the repetitive DNA proportion found in their assemblies. However, and although the species with the largest genome is *B. bipinnatifida*, this one is not the one with the largest proportion of repeats found in the assembly. Species such as *B. masoniana* and *B. darthvaderiana* show a larger masked repeat content in proportion to their assembled genome size. This is an additional support to our WGD hypothesis in *B. bipinnatifida*, together with the large number of duplicated BUSCOs found in Chapter 3.

Although the repeat libraries built with RepeatMasker were *Begonia*-specific, a high percentage of the masked repeats were not classified. The genomic proportion of unclassified repeats goes from 16,7% in the *B. luxurians* genome to 48.17% in the *B. darthvaderiana* genome, which is proportional to the total amount of masked repeats in each assembly.

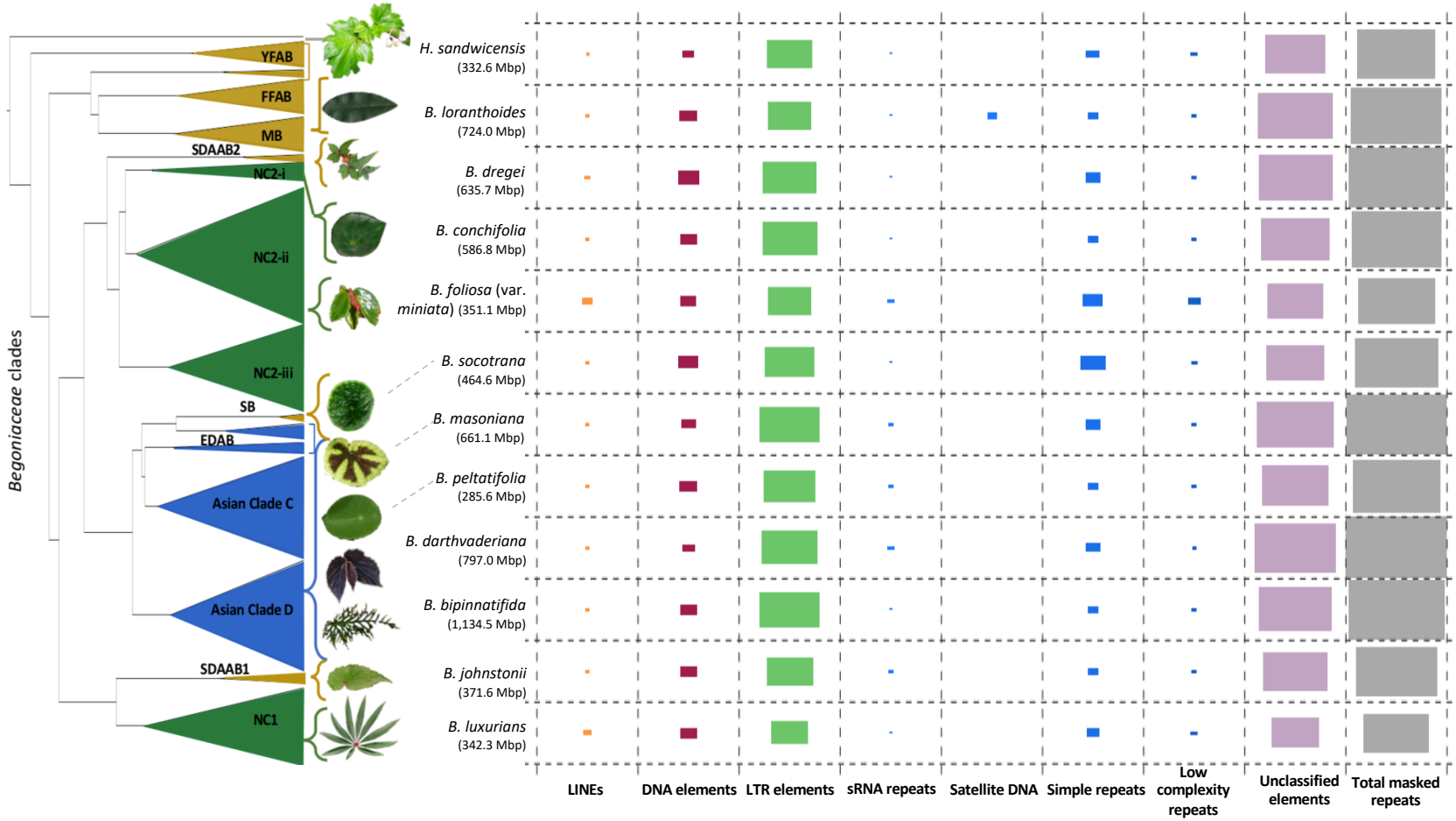


Figure 4.2: Repeatome content and variation of 12 Begoniaceae genome assemblies and their phylogenetic position in a *Begonia* clade phylogeny (simplified from Moonlight et al. (2018)). Each species names is accompanied by their genome sizes (1C (Mbp), Chapter 3). Box sizes represent the % of each type of repetitive element indicated below found by RepeatMasker in each genome assembly. Colours of clades in the tree indicate the continent of origin, colours of boxes indicate the TE order or repeat type. Only TE types represented with more than 0.5% of the reads in at least 1 species are shown.

There are many TE annotation tools and little general consensus on which are the best in finding the most transposon and the least false positives (Ou et al., 2019). Recent work by Yan et al. (2020) uses an original approach to test a wide list of TE annotation tools and suggests that for a *de novo* repeat library construction, RepeatModeler (and MITE_hunter for small tandem repeats) is the best candidate. This is supported by numerous articles on newly sequenced and assembled genomes where similar approaches involving RepeatModeler and RepeatMasker are taken (Lu et al., 2019; Wang et al., 2019; Hong et al., 2020). However, this approach may be useful for general genomic descriptions but not accurate enough in detail for TE-specific studies. Perhaps a RepeatModeler approach combining our *Begonia* repeat libraries with existing repeat libraries from other plant groups (Repbase; Jurka et al. (2005)) as, for example, done in Zhou et al. (2020) or Xia et al. (2020) would improve the annotations by increasing accuracy including homologous sequences. Including repeat libraries of other Cucurbitales species in our *de novo* repeat identification approach may increase the number of identified repeats in our Begoniaceae genomes and must be considered or future analyses.

The classification system in RepeatMasker includes several types of short simple repeats (Yates et al., 2020), such as small RNA repeats (non-functional copies of RNA genes which have been reintegrated into the genome), Satellite DNA repeats (repeated patterns of variable length of the same short nucleotide sequence), simple repeats (duplications of simple sets of 1-5 DNA bases) and low complexity repetitive regions (poly-purine or poly-pyrimidine stretches, or regions of extremely high AT or GC content). These small, simple types of repetitive DNA are normally not well assembled and are therefore underrepresented in genome assemblies. This problem is most severe in short-read assemblies (Peona et al., 2018), and can clearly be seen in Figure 4.2. Type II TEs (DNA elements) and non-LTR type I (LINEs) also make up to small, stable genomic proportions. As expected, the most represented type of repeat in all species according to our RepeatMasker results are LTRs.

To further identify, annotate and analyse LTR regions, a LTR-specific annptation pipeline was run, which included different LTR annotation softwares such as LTR_finder,

LTRharvest and LTR_retriever (Xu and Wang, 2007; Ellinghaus et al., 2008; Ou and Jiang, 2018). As shown in Figure 4.3, the genomic proportions identified as LTRs by this pipeline are, in most species, very similar to the proportions of LTRs and RepeatMasker's unclassified repeats together. This implies that this LTR-specific pipeline was able to identify most unannotated repeats as LTRs. The use of LTR-specific annotation tools improved our results accuracy, and a LTR-specific annotation pipeline is recommended for a more accurate and complete repeat annotation.

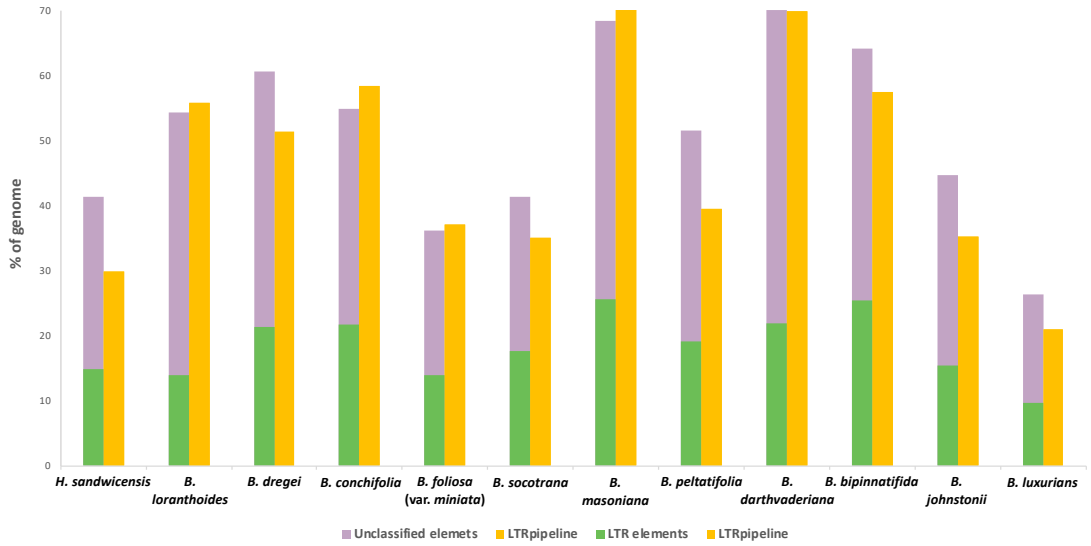


Figure 4.3: Comparison between the assembly proportions annotated as LTR (green) and unknown (purple) repeats found by RepeatMasker and all the LTRs found by the LTR pipeline (orange).

The LTR pipeline implemented is not only more accurate at identifying LTRs but also it classifies them into the LTR-*Gypsy* and LTR-*Copia* superfamilies, which was not part of the RepeatMasker output in Figure 4.2. Figure 4.4 shows the percentage of each genome assembly comprised of these LTRs, together with the total proportions of LTRs and masked repeats (data table in appendix 4.A.1). As previously reported in other plant species, there is a larger proportion of LTR-*Gypsy* elements in all genomes compared to LTR-*Copia* or unknown LTRs, although the ratio *Gypsy*/*Copia* is highly variable among species (Mascagni et al., 2017). This is expected as it was stated before LTR-*Gypsy* elements have preference for insertion in heterochromatic regions and are less likely to be negatively selected than those inserted in gene-rich regions.

The total LTR content found in these species is again proportional to the total amount

of masked repeats and to their genome sizes, and found to be higher in most species from Asian clades with a total LTR content of 71.9, 70.0 and 57.4% in *B. masoniana, B. darthvaderiana* and *B. bipinnatifida* respectively, with the exception of *B. peltatifolia*, which has a much smaller genome and repeat content overall compared to related species. Amongst the species with higher LTR content there is also a Neotropical species, *B. conchifolia*, which shows a LTR genomic proportion of 58.42%.

Overall, Figure 4.4 indicates a wide variation in LTR content across Begoniaceae species, with the lowest proportions found in the smaller genomes such as *B. luxurians* (21.0%), *H. sandwicensis* (29.9%), some African species such as *B. socotrana* and *B. johnstonii* (35.1 and 35.2 %) and Neotropical and Asian species belonging to smaller or unidentified sections such as textit *B. foliosa* and *B. peltatifolia*. An exception to the tendency of species in smaller and older clades having lower LTR contents are the African species *B. loranthoides* and *B. dregei*. These show similar repeat proportions and genome sizes, and only slightly lower LTR content than their closely related Neotropical species *B. conchifolia*.

Figure 4.4 also shows that the LTR-*Copia* superfamily seems less abundant than LTR-*Gypsy*, but these numbers are more closely associated with clade species richness. The largest genomic proportion of *Copia* elements (15.7%) is found in the Neotropical species *B. conchifolia* (section *Gireoudia*, 110 spp.), and it is also larger in Asian species belonging to large sections, such as *B. masoniana* (*B.* sect. *Coelocentrum*, 84 spp.) and *B. darthvaderiana* and *B. bipinnatifida* (*B.* sect *Petermannia* $\sim$ 451 spp.). These LTR-*Copia* expansions seem clade or section-specific since their closest relatives from more depauperate sections and clades *B. dregei*, *B. foliosa B. socotrana* or *B. peltatifolia* show lower genomic proportions of this retrotransposon superfamily. An exception to this pattern is the *B. johnstonii- B. luxurians* species pair, in which both genome size and LTR content differences seem to follow opposite patterns. Statistical analysis showed no significant differences in the number of LTR-*Copia* elements between species from large radiations (from sections with >50sp) and from species-poor clades. This could be influenced by this exception, since larger proportion of repetitive sequences is found in the genome assembly of the African species *B. johnstonii*,

compared to the Neotropical species *B. luxurians*, which also has a smaller genome size than its depauperon.

It is important however, to keep in consideration that the differences in sequencing strategies taken to sequence these species may have played a role not only in the number of LTR elements assembled, but also in the assembly quality of repetitive elements in general. To highlight the importance of long-read sequencing for LTR assembly, Figure 4.5 shows the number of each type of full-length LTR found by this pipeline in all genomes. There is a remarkable difference found between those assemblies that include long-read PacBio data and those which do not. The most striking difference is found between *B. bipinnatifida* and both *B. masoniana* and *B. darthvaderiana*. Even though the genomic proportions of LTR sequences found in these species are similar (between 60 and 70%), and having *B. bipinnatifida* a much larger genome and therefore a much larger number of identified LTRs, Figure 4.5 shows that the number of full-length LTR elements is significantly lower in this assembly compared to those assembled with long-read data (indicated with an asterisk). This highlights the importance of the investment in long-read sequencing for the accurate assembly and downstream analysis of TEs. To be able to evaluate the bias created by genome assembly contiguity and quality on TE assembly and analysis, a non assembly-dependant transposon annotation method would be of help.

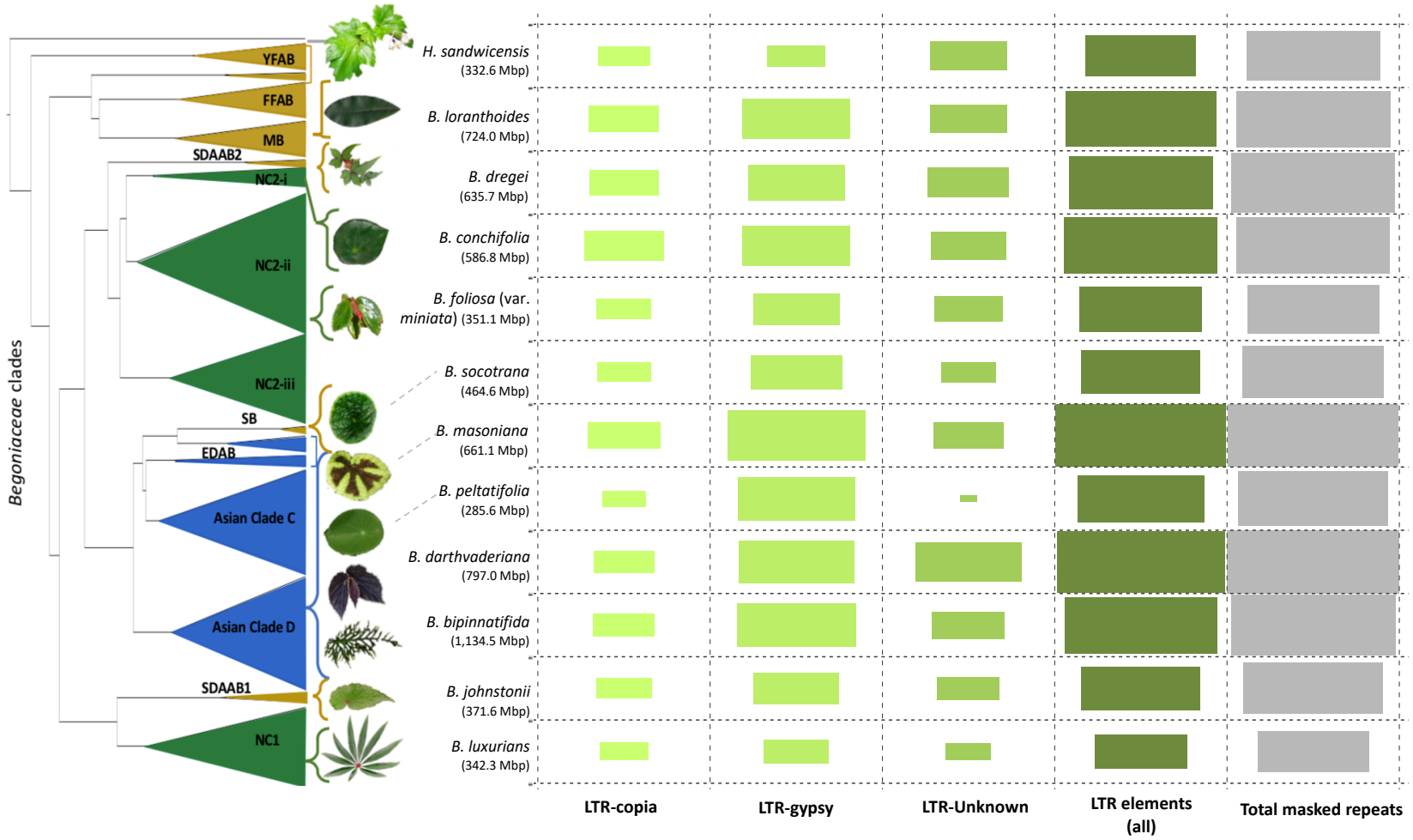


Figure 4.4: LTR content variation of 12 Begoniaceae genome assemblies and their phylogenetic position. Each species names is accompanied by their genome sizes (1C (Mbp), Chapter 3). Box size represents the % of each type of repetitive element indicated below found by LTR_retriever in each genome assembly. Colours of clades in the tree indicate the continent of origin, colours of boxes indicate the TE order, varying shades indicate the LTR superfamilies.

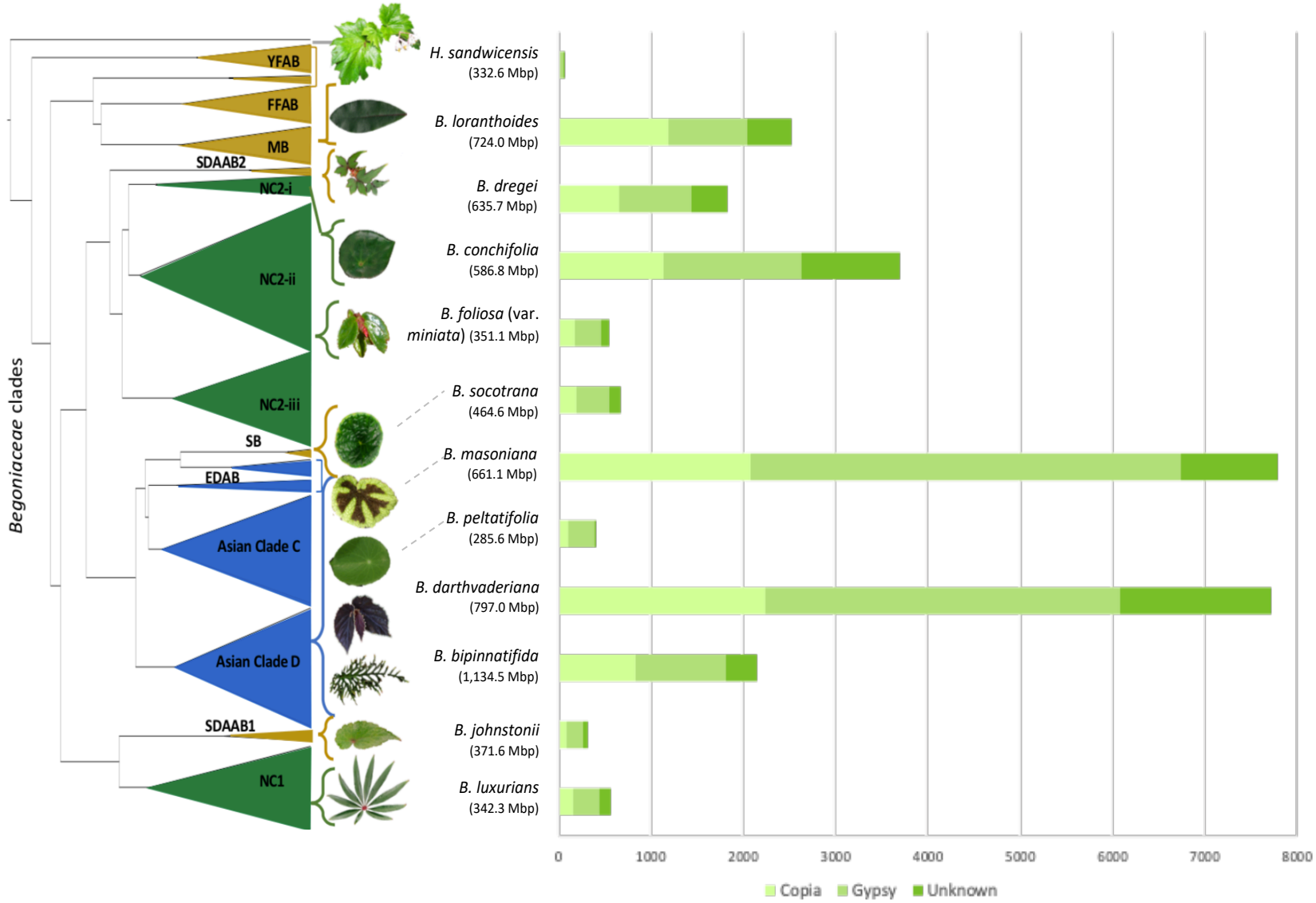


Figure 4.5: Number of full-length LTRs identified in of the 12 Begoniaceae genome assemblies and their phylogenetic position in a *Begonia* clade phylogeny (simplified from (Moonlight et al., 2018)). Each species names is accompanied by their genome sizes (1C (Mbp), Chapter 3). Colour shades of the bars indicate the LTR superfamily. Asterisks indicate the species sequenced with long read technologies in addition to Illumina.

Read clustering-based repeat analysis of eight Begoniaceae genomes

RepeatExplorer2 (Novak et al., 2013) is a repeat identification pipeline that uses sets of clean, random Illumina reads and clusters them according to similarity. It then classifies such clusters using a plant-specific repeat database (Repbase, Jurka et al. (2005)). One drawback of this method is that only species with available raw reads can be tested and analysed. Since the *B. loranthoides*, *B. masoniana*, *B. peltatifolia* and *B. darthvaderiana* genomes were not sequenced for this project and raw datasets have not been made available (Li et al., pers. comm.), these species are not included in the following RepeatExplorer2 analysis. Figure 4.6 shows the proportions of analysed reads from clusters matching specific TE types for *H. sandwicensis*, *B. dregei*, *B. conchifolia*, *B. foliosa* var. *miniata* (Griesmann et al., 2018), *B. socotrana*, *B. bipinnatifida*, *B. johnstonii* and *B. luxurians* (data table in appendix 4.A.1). The most noticeable difference between the repeat annotations by RepeatMasker and RepeatExplorer2 is the greater level of detail in repeat classification of by RepeatExplorer, reaching LTR family or lineage level. This makes the among-species comparison more interesting. This also makes possible to focus our interest in specific LTR families that show the most significant differences across the phylogeny.

In the case of *Begonia*, Figure 4.6 shows that the most abundant LTR types are the LTR-*Copia*-SIRE (*Copia*-SIRE), the LTR-*Gypsy*-Chromovirus-CRM (*Gypsy*-CRM) and the LTR-*Gypsy*-Chromovirus-Tekay (*Gypsy*-Tekay) families. More specifically, the *Gypsy*-Tekay family seems to be shared across all studied species, being the only LTR family present in a high proportion of the *H. sandwicensis* genome. This family is present in high proportion in some species, representing a 27.1% of the reads in *B. conchifolia* and a 22.7% in *H.sandwicensis*, but also its presence is reduced in African species such as *B. socotrana* and *B. johnstonii*, and the Neotropical species *B. luxurians* where it only takes up to 1.05, 3.7 and 2.1%, respectively. Another LTR family found in high proportion in all *Begonia* species (but barely present in *Hillebrandia*) is the LTR-*Gypsy*-CRM group. The presence of this family of LTRs across the analysed *Begonia* species is variable, ranging from 2.8% in *B. conchifolia* to 12.9% in *B. socotrana*. In this case, differences among species are not as striking as in the other two LTR families, but African species seem to show the largest proportion of them.

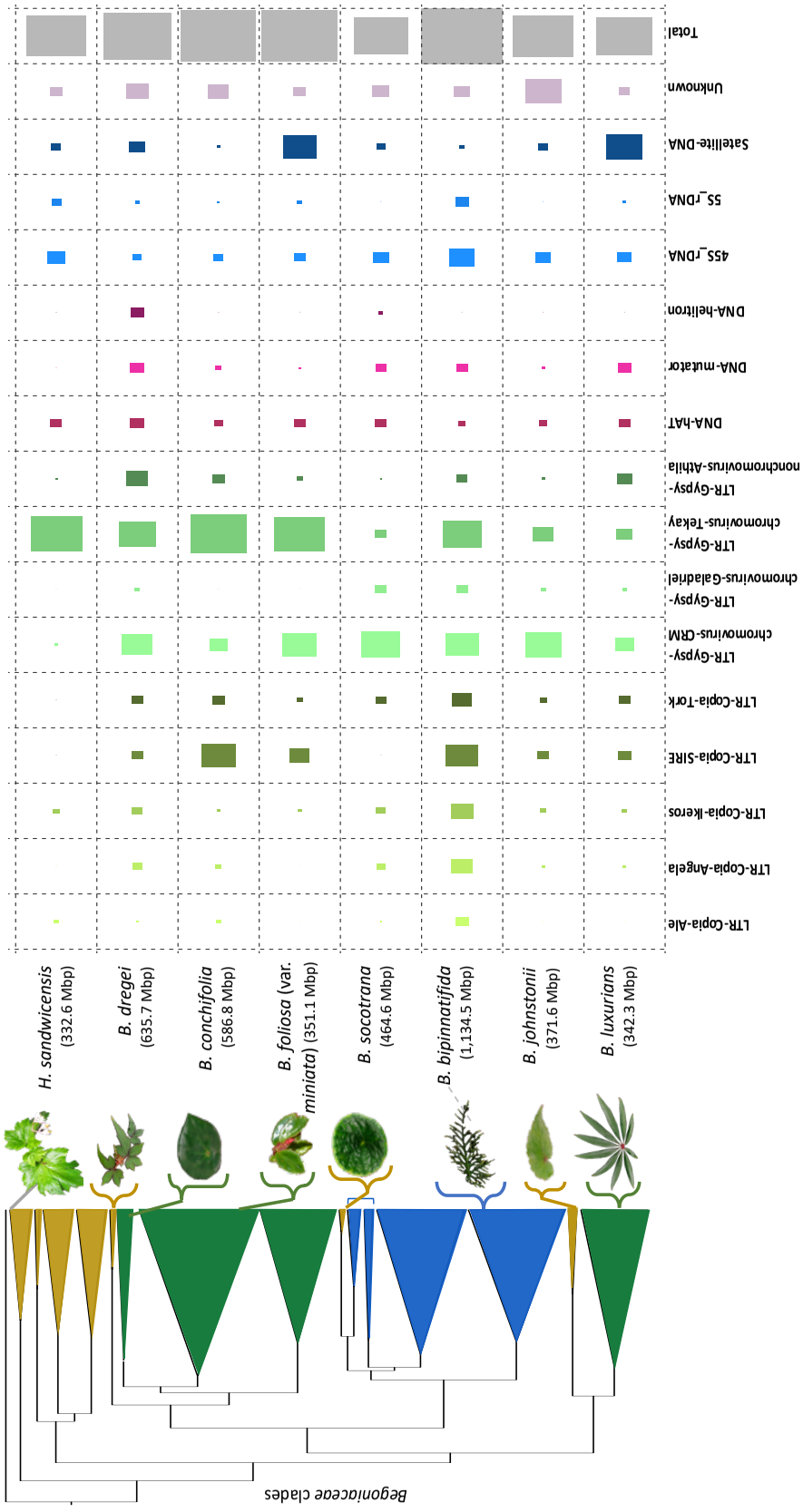


Figure 4.6: Variation of repeat and transposon content found in the genomic read datasets of eight Begoniaceae species, and their phylogenetic position (simplified phylogeny from Moonlight et al. (2018)). The clade colours in the tree indicate the continent of origin. Each species names is accompanied by their genome sizes (1C, Mbp). Different colours indicate different repeat or TE type, shades of the same colours indicate different families or superfamilies. Only TE types represented with more than 0.5% of the reads in at least one species are shown.

The last LTR family showing interesting differences across species is the *Copia*-SIRE group. Although in smaller proportion compared to *Gypsy* families, *Copia*-SIRE content varies from a 10.0% in *B. conchifolia* and a 8.8% in *B. bipinnatifida* to a 0.0% in *B. socotrana*. This LTR group is clearly the most variable across species and specially abundant in species belonging to highly speciose sections and clades, with the exception of *B. luxurians*. To have a better insight of the TE activity and insertion time variation across *Begonia*, these three LTR families were selected for further LTR insertion dynamics analysis, and will be further discussed in Section 4.2.3.

Another advantage of RepeatExplorer2 compared to annotating full assemblies is the detection and quantification of small simple repeats that are normally misrepresented and misassembled (even more in short-read only assemblies). While RepeatMasker identified 0% satellite DNA in 11 out of 12 genome assemblies, RepeatExplorer2 provides a more interesting picture of highly variable amounts of Satellite DNA across Begoniaceae species. Although present in all species, this type of repeat is highly abundant in *B. foliosa* var. *miniata* and *B. luxurians*. Both of these are Neotropical species with small genomes, in which LTRs are not highly abundant. Our results show that species from large radiations with large genomes have significant enrichment of *Copia*-SIRE elements, while those from large radiations and smaller genomes (¡500Mbp), show a significant enrichment of satDNA sequences (p-values 8.3E-05 and 3.4E-06, respectively). This reveals that almost all Neotropical and Asian species analysed have either a LTR or satellite-enriched genome. The exception is *B. peltatifolia*, which could not be analysed using this pipeline and has an undefined phylogenetic position (unassigned section).

Although it is interesting to see what the most abundant types of LTRs are and the large differences in genomic content some of them represent, it can be difficult to interpret RepeatExplorer2 results as the repeats are not presented in a genomic context, making it impossible to assess whether these TEs are shared across species or are species-specific. To have a better overview on what LTR lineages are the result of old LTR activity and which are recently expanded in specific taxa, reads from all species were used for a RepeatExplorer2 comparative analysis. This analysis of the RepeatExplorer2

pipeline clusters these reads by similarity, annotates them and distinguishes which reads belong to each species within each clusters. This reveals which TE types are most shared among Begoniaceae species and which TE categories are clade, section or species specific. Figure 4.7 summarises the results of the RepeatExplorer2 comparative analysis, showing the cluster sizes and the proportion of the different species reads within each cluster. Results show that the clusters common to all species are mostly annotated as organellar and rDNA. Among those shared by all datasets in different proportions, the largest ones include two Tekay and one CRM clusters. The two *Gypsy*-Tekay clusters are more abundant in *B. conchifolia* and *B. bipinnatifida*, while the larger CRM cluster shows to be more represented by the three African species. These two are the types of LTR family that are most abundant in most species, specially *H. sandwicensis* (Tekay only) and the African species *B. dregei* (both Tekay and CRM), *B. socotrana* (CRM only) and *B. johnstonii* (CRM and satellite DNA). On the other hand, the *Gypsy*-Tekay and *Copia*-SIRE are the most represented LTR families in the largest species-specific clusters for *B. bipinnatifida* and *B. conchifolia*. This correlates with the data shown in Figure 4.6.

The other two Neotropical species specific clades show a large proportion of Satellite DNA (as also seen in Figure 4.6). Interestingly, the *B. bipinnatifida Copia*-SIRE cluster seems to be largely clade-specific, as well as the *B. conchifolia* one, as it is partially shared with *B. foliosa* var. *miniata*, which belongs to the same clade (but different subclades, as seen in Chapter 2).

The Begoniaceae repeatome annotations analysed in this chapter show that *Begonia* repeat landscapes are dominated by LTR elements, and among these the *Gypsy* superfamily is the most abundant. These landscapes are similar to the observed in other species (Mascagni et al., 2017), in which LTR-*Gypsy* elements are also the most abundant and correlate with genome size. Results shown in both previous figures also show *Copia*-SIRE elements to be enriched in genomes from species-rich clades such as *B. conchifolia*, *B. foliosa* and *B. bipinnatifida*.

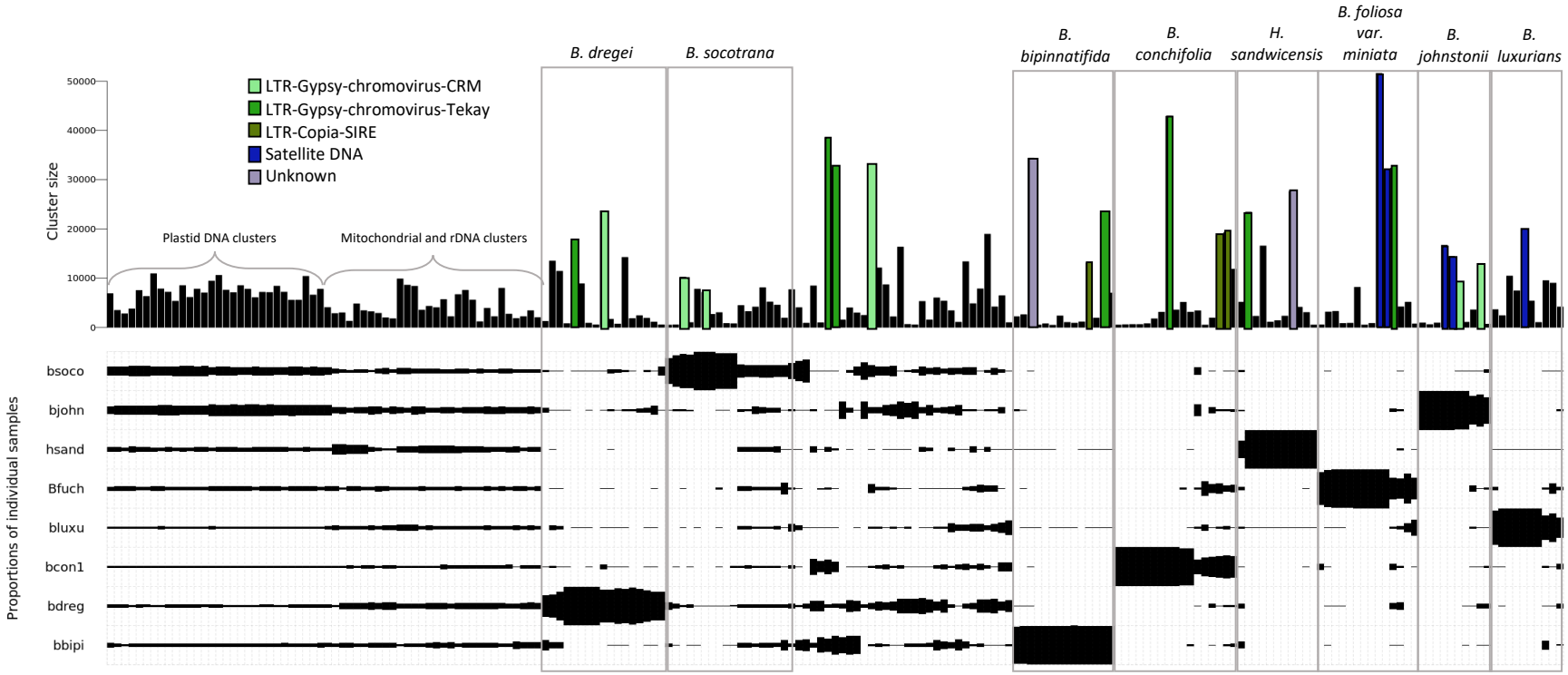


Figure 4.7: RepeatExplorer2 comparative analysis results on repeat and transposon variation found in the genomic datasets of seven Begoniaceae species. The barchart in the upper half indicates cluster size (n of reads). Only the 3-4 larger clusters within a species-specific group were coloured to show their annotation. The barchart in the lower half indicates the proportion of reads from each species belonging to a specific cluster. Species codes in the y axis are: bsoco foe *B. socotrana*, bjohn fore *B. johnstonii*, hsand for *H. sandwicensis*, Bfuch for *B. foliosa* var. *miniata*, bluxu for *B. luxurians*, bcon1 for *B. conchifolia*, bdreg for *B. dregei* and bbipi for *B. bipinnatifida*. Groups of species-specific and single-species rich clusters are included in large grey boxes.

To be able to verify the link between TE activation and speciation bursts events, TEs do not only need to be more abundant in larger taxa, but also more active during radiation and speciation events, and shared among all species of a specific radiation. Section 4.2.3 will explore the temporal dynamics of these LTR groups to compare LTR dynamics across radiations, and see whether these TEs were likely active at the radiation of these *Begonia* species.

How does genome assembly quality affect repeat annotation?

The two different TE annotation pipelines used give similar pictures of the repeat and transposon landscape across the 12 Begoniaceae genomes studied. Both these approaches are needed when analysing and trying to understand repeat landscapes from genomes of different characteristics that have been sequenced and assembled using different approaches. RepeatExplorer2 represents an assembly-independent method that produces accurate and specific results. However, RepeatMasker provides a more precise quantification of the repeats found. Quantifications given by RepeatExplorer2 are based on a random subset of genomic reads, being subject to misrepresentation. In addition, organellar DNA reads take up a good portion of the input datasets and are difficult to identify and remove. The % estimations given by RepeatExplorer must be handled with care, since transforming the number of reads of our sample into genomic percentage cannot be taken as a 100% unbiased method. A series of assay runs with different genomic proportions were also performed, and a coverage of 0.2x was considered the minimum optimal coverage for all Begoniaceae species. In addition, repeated RepeatExplorer2 runs of the same datasets with different random read samples will approximately keep the TE type proportions, but annotations can vary slightly, and so do TE quantifications (data not shown, Campos-Dominguez, unpublished). In this chapter, this was avoided by a manual curation of cluster annotation, which helps with the annotation accuracy but cannot compensate for the lack of full TE sequence and the genomic proportion values inaccuracy. However, manual curation, optimisation analyses and assays take time and make the RepeatExplorer2 pipeline results analysis more time and resources-consuming.

Figure 4.8 shows a comparison between the different types of repeat content identified

by both approaches. In all cases except from *B. foliosa*, the proportion of repeats found in the assembly is higher than the ones estimated by RepeatExplorer. In the case of the *B. foliosa* genome, however, this pattern is not seen and the RepeatExplorer2 results show a larger repeat proportion than the annotated genome. It is highly likely that the available *B. foliosa* genome assembly had shorter contigs removed and the smaller, misassembled repetitive contigs are missing in the assembly but found by RepeatExplorer2. As mentioned earlier, Satellite DNA is often not assembled in genomes sequenced using only short-read data. On the other hand, the amount of unknown elements as well as the annotation depth and the identification of small repeats such as satellite DNA is much improved by RepeatExplorer2.

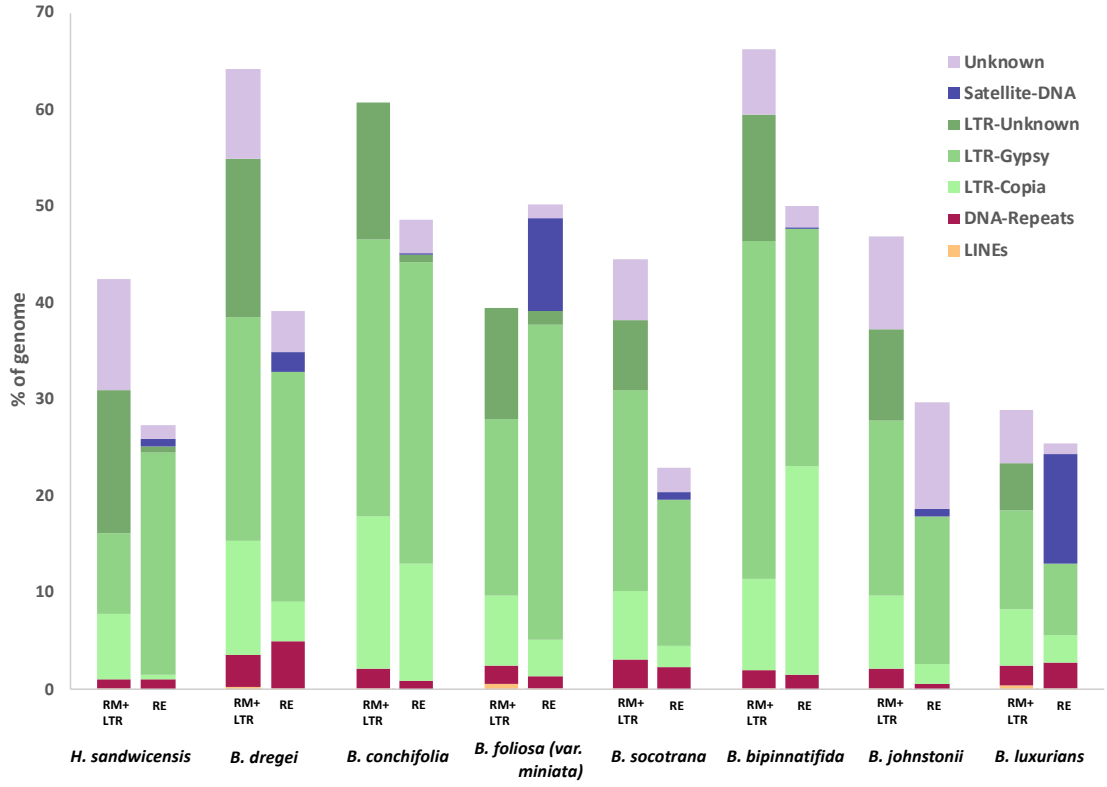


Figure 4.8: Comparison of repeat and transposon content of 8 Begoniaceae genomes according to their genome annotations (left) and their RepeatExplorer2 results (right). Different colours indicate different repeat or TE type, shades of the same colours indicate different families or superfamilies.

Although there are not many studies that combine our two approaches, some prove RepeatExplorer2 analyses can be extremely useful even when having a good quality genome assembly (Wan et al., 2018), in particular the use of comparative analysis. The results of this chapter suggest that both approaches provide useful and comparable

insights of the transposon landscape. While the unassembled data pipeline can generate a highly useful overview and detailed classification of the repeat landscape, a good genome assembly can add the possibility of syntheny studies and LTR insertion site analysis. An even better way of implementing both analyses together would be the analysis of the repeat clusters found by RepeatExplorer2 in the assembled genome with tools such as PROFREP (also available from RepeatExplorer2, Novak et al. (2013)). In the case of the genome assemblies generated in this study, a full repeat annotation with this tool would be an inconvenient task due to the high levels of fragmentation of most of our assemblies. Nevertheless, using this tool to find the scaffold locations of a cluster of interest would be a useful approach. Further analysis in which specific LTR insertions and their genomic environment are studied would provide a more complete overview of the effects of LTR variation and activity in the *Begonia* genome evolution.

4.2.2 Temporal dynamics of *Begonia* LTRs

LTR insertion dating across Begoniaceae using full-length assembled LTRs

In addition to finding the type and amount of repeats present in the Begoniaceae genomes analysed in the present study, the objective of this chapter is also trying to understand and compare the temporal dynamics of such transposons across *Begonia*. One of the tools used for the annotation of LTRs, LTR_retriever (Ou and Jiang, 2018), provides the insertion times of each full-length LTR found in the genome assembly, based on the genetic divergence of the two Long-Terminal-Repetitive regions at both ends of the retrotransposons, that ought to be identical upon insertion (SanMiguel et al., 1998).

Figure 4.9, Figure 4.10 and Figure 4.11, show the distribution of the LTR insertion times of the *Copia*, *Gypsy* and Unknown full-length LTRs from Figure 4.5. All three figures show the same pattern of *H. sandwicensis* and the four African species showing thinner, shorter distributions with widenings indicating older LTR bursts. On the other hand, Neotropical and most Asian species would show thicker (larger) distributions, with a higher number of TE bursts in the recent side of the timescale. This is yet another evidence of *Begonia* species having overall more and more recently active retrotransposons than *Hillebrandia*, as well as a pattern of species belonging to larger

clades showing higher and more recent retrotranspositional activity compared to their African relatives.

This method is, again, conditioned by the sequencing technology and assembly approaches taken for each genome. As seen in Figure 4.5, the *B. masoniana*, *B. darthvaderiana* and *B. conchifolia* genomes have a larger number of full length identified LTRs. The similarity of the terminal sequences in the assembly could also have been altered depending on the approach, as highly similar and long repeats are more likely to be misassembled when using short read data. This could explain the differences seen in these species compared to the rest. To test this, the LTR sequences were also annotated on a *Begonia conchifolia* assembly from illumina only data. The violin plots comparing the LTR dynamics of full-length LTR sequences from these short-read only and long-read assemblies are shown in Appendix 4.A.2. To test whether these distributions are statistically significant between each other t-tests were performed (as previously done by Jiménez-Ruiz et al. (2020)), and distributions were found to be significantly different between *Copia*, *Gypsy*, and Unknown elements (p-values are 0.0 in the three cases). The general pattern, however, is not as influenced by this since not only the Pacbio sequenced species but all Neotropical and most Asian species show the same LTR insertion patterns compared to their closest African relatives. Nonetheless, to test for significant differences between LTR dynamics across species, only short-read assemblies were compared between each other. Results show that all species from large radiations (*B. conchifolia*, *B. foliosa*, *B. bipinnatifida*, and *B. luxurians*) show significantly different LTR activity distributions than their closest depauperons *B. dregei*, *B. socotrana* and *B. johnstonii*, respectively (all p-values <0.01).

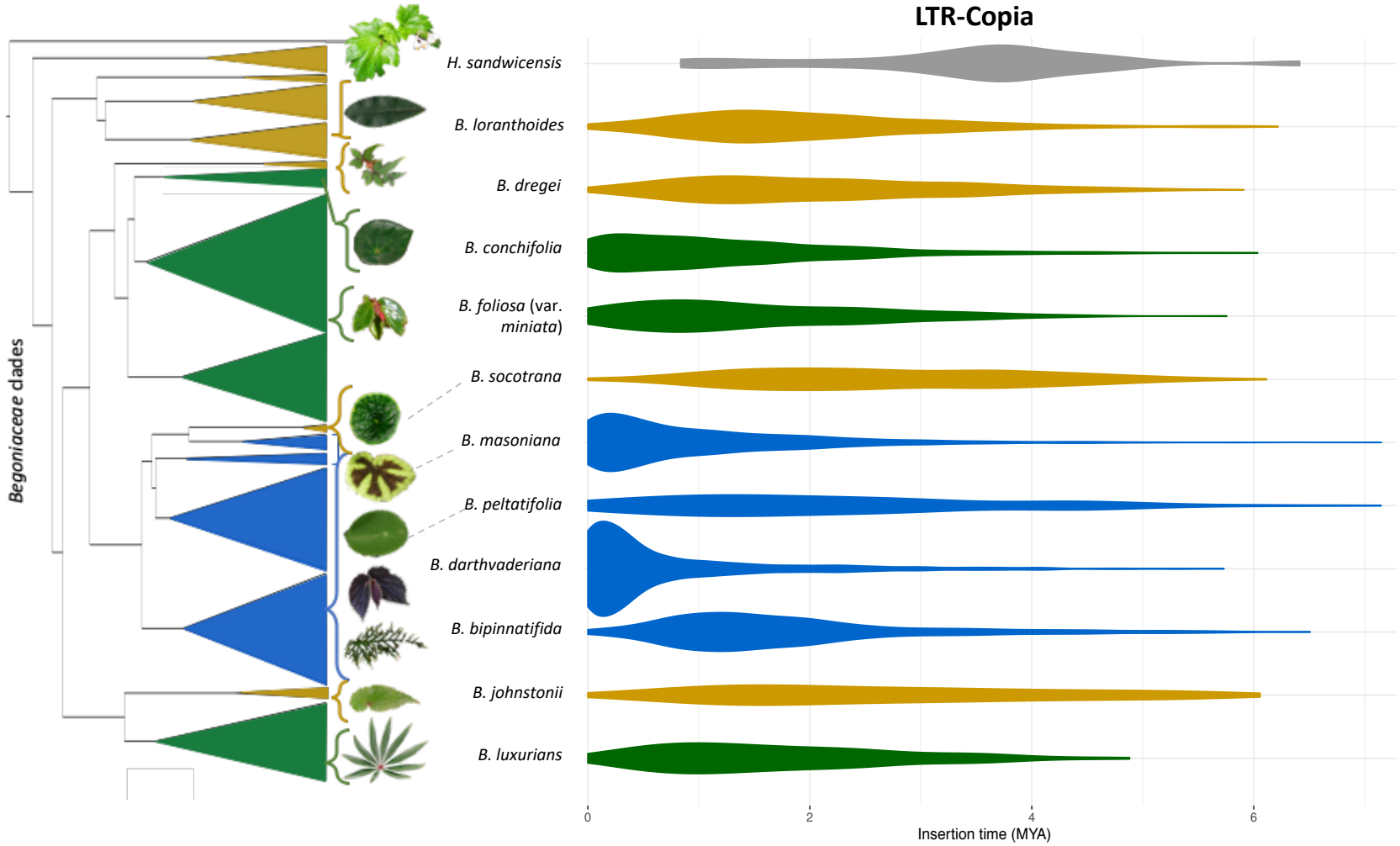


Figure 4.9: LTR-*Copia* insertion times of all full-length elements found in each genome assembly by LTR_retreiever (Ou and Jiang, 2018), and the phylogenetic position of each species. Colours of the violins and tree clades indicate the continent of origin of each clade and species.

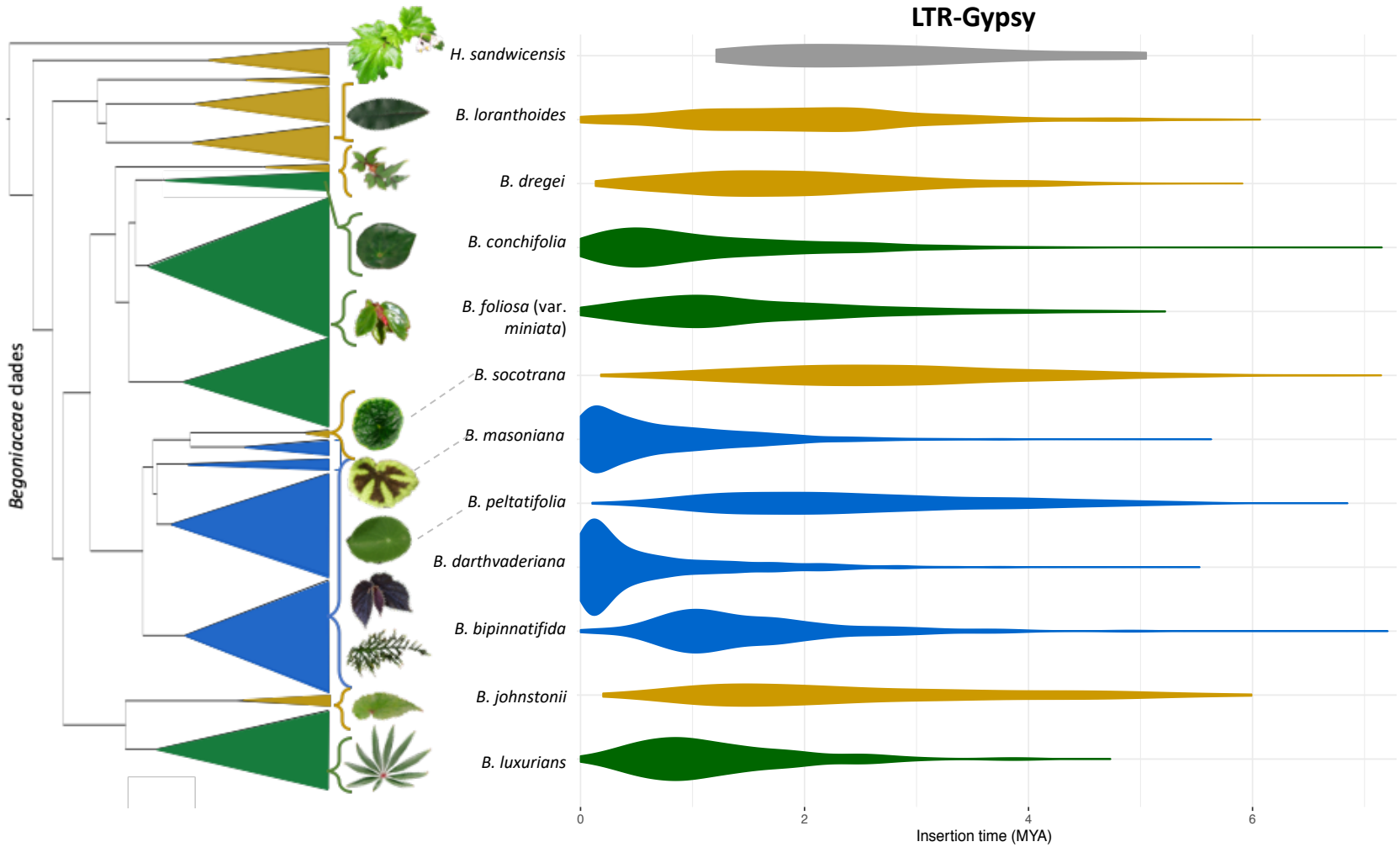


Figure 4.10: Insertion times of all full-length LTR-*Gypsy* elements found in each genome assembly by LTR_retreiever (Ou and Jiang, 2018), and the phylogenetic position of each species. Colours of the violins and tree clades indicate the continent of origin of each clade and species.

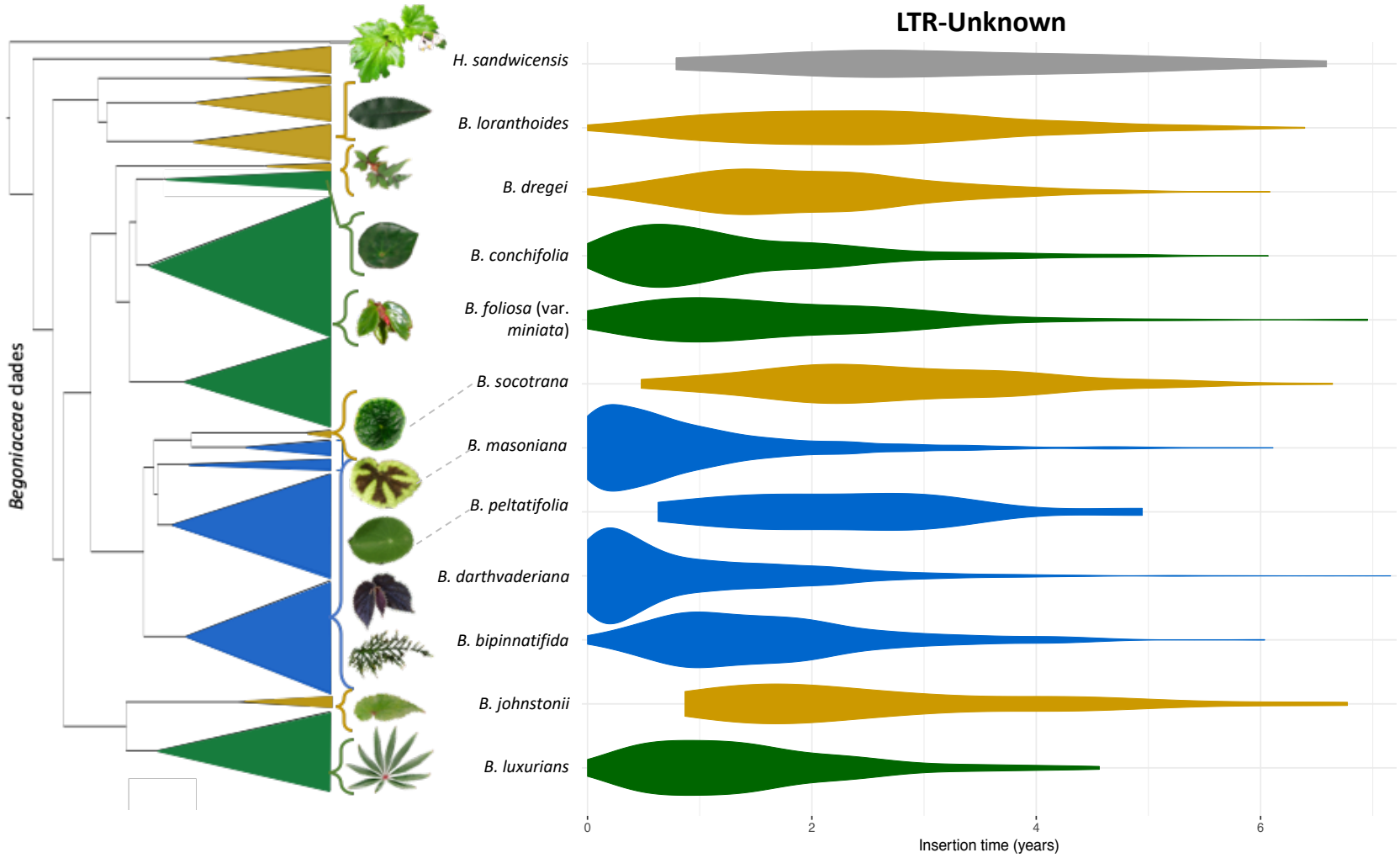


Figure 4.11: Insertion times of all full-length LTR elements of unidentified superfamily found in each genome assembly by LTR_retreiever (Ou and Jiang, 2018), and the phylogenetic position of each species. Colours of the violins and tree clades indicate the continent of origin of each clade and species.

LTR insertion dynamics across Begoniaceae using cluster-based repeat annotations

A different method for studying LTR temporal dynamics is analysing the sequence divergence of RT-domains found by RepeatExplorer2, as previously done in Usai et al. (2017) and Mascagni et al. (2020). This consists on determining the genetic distance of the reads mapping to RT-domains of different clusters. As previously found in Section 4.2.1, the LTR families of *Gypsy*-CRM, *Gypsy*-Tekay and *Copia*-SIRE are the ones of most interest for comparison among the Begoniaceae. For all species studied, all clusters identified as any of these three types of LTRs, and with more than 1% of their reads with hits on RT domains were selected for analysis.

The number of clusters of each type found per species are highly variable. As observed in Figure 4.6, *Gypsy*-CRM are more abundant in all *Begonia* species, and therefore the number of analysed clusters of this type is larger (18). On the other hand, *Copia*-SIRE elements are less abundant and only present in large enough amounts in a few species, and only nine clusters of this type were dated. Figure 4.11 shows the retrotransposition activity based on the RT-domain sequence divergence of these clusters.

Overall, most clusters seem to follow similar trends and their patterns overlap. However, there are a few ones showing cluster-specific peaks too. In all analysed LTR lineages, there are clusters showing specific peaks (translated as bursts of retrotransposition activity) compared to the rest (indicated with an arrow in Figure 4.11). Both *H. sandwicensis* clades in the figure (HsandCL6 and HsandCL240) show older peaks compared to the rest of the species. This, as well as Figure 4.9, Figure 4.10 and Figure 4.11, indicate *H. sandwicensis* transposable elements bursts of activity are older than the ones observed in all *Begonia* species studied. In addition, all the cluster-specific peaks and recent retrotransposition peaks observed in Figure 4.11, are from clusters found in either Neotropical or Asian species. This, also supported by Figure 4.9, Figure 4.10 and Figure 4.11, supports the hypothesis that species from more specious clades have more recently active transposons and therefore more dynamic genomes. More specifically, two *B. conchifolia* clusters and one *B. bipinnatifida* show recent increases in retrotransposition activity in *Gypsy*-CRM and *Gypsy*-Tekay clusters, and *Copia*-SIRE,


A
LTR-Gypsy-CRM
% Pairwise comparisons
Genetic distance (MYA)
B. conchifolia CL94
H. sandwicensis CL240
B. conchifolia CL174
HsandCL240
BdreCL12
BconCL94
BconCL116
BconCL174
BfucCL5
BfucCL13
BfucCL85
BsocCL4
BsocCL7
BscoCL9
BbipCL51
BbipCL96
BjohnsCL17
BjohnsCL40
BjohnsCL55
BjohnsCL97
BluxCL27
B
LTR-Gypsy-Tekay
% Pairwise comparisons
Genetic distance (MYA)
B. conchifolia CL49
B. foliosa var. miniata CLs 24 and 43
H. sandwicensis CL6
HsandCL6
BdregCL15
BconCL9
BconCL23
BconCL48
BconCL49
BfucCL24
BfucCL43
BfucCL82
BbipiCL6
BbipiCL7
BbipiCL42
BbipiCL122
BjohnsCL10
C
LTR-Copia-SIRE
%Pairwise comparisons
Genetic Distance (MYA)
B. bipinnatifida CL111
B. luxurians CL46
B. conchifolia CL24
BdregCL27
BconCL24
BconCL45
BfucCL11
BfucCL20
BbipCL23
BbipCL111
BjohnsCL13
BluxCL46

Figure 4.11: (Previous page.) Timing of different LTR clusters retrotranspositional activity in the analysed Begoniaceae species based on pairwise comparisons of Illumina reads to the cluster contig regions that match RT-encoding domains. Line colours indicate the continent of origin (green for New Tropics species, blue for Asian species and yellow/orange for African species), colour shade indicates different clusters. species codes in the legend are Hsand for *H. sandwicensis*, Bdreg for *B. dregei*, Bcon for *B. conchifolia*, Bfuc for *B. foliosa*, Bbipi for *B. bipinnatifida*, Bjohns for *B. johnstonii* and Blux for *B. luxurians*.

respectively. Whether these TEs are currently active could be further confirmed with RNA-seq or transposon display experiments.

Are LTR-insertion bursts comparable using different calculation methods?

Both methods used and presented above show certain limitations that must be kept in mind. As discussed earlier, when comparing LTR regions of annotated genomes that show high repeat content and heterozygosity, it must be considered that the nucleotide substitutions found in the repetitive regions of the assemblies will include some artefacts. This, together with the use of different sequencing technologies and assembly methods, introduces an error factor to the calculation of genetic distance between LTRs that must be accounted for. High quality, long-read sequencing and chromosome level assemblies of all species included in this study would be the best way to study and compare LTR elements. This was, however, unrealistic in this project.

The alternative method based on read clustering, although it avoids assembly artefacts, it is based on only a subset of the genome and relies on clusters having reads mapping to RT domains. It is also more time consuming (as clusters need to be chosen and sequences extracted and compared manually). A number of large clusters annotated within the lineages of interest had to be removed because no RT domains were found. Maybe replicates of the RepeatExplorer2 analyses would help finding these RT sequences among the read datasets, but this would again be highly time consuming and inefficient.

Moreover, the time-scales showed in Figure 4.9, Figure 4.10, Figure 4.11 shows that either LTR_retriever was only able to identify young transposons (<7MYA), or the dating of these sequences is inaccurate. Although specific nucleotide substitution rates

for Begoniaceae were calculated using neutral intronic sequences, it is possible that LTR sequences do not mutate at the same rate as introns. Another option to consider is that LTR_retriever was only able to identify relatively young full-length LTRs as older ones would not have been fully annotated.

In Figure 4.11, the timescale is similar to in Figure 4.9, Figure 4.10, Figure 4.11, which means that even though the dating methods and nucleotide substitution rates were different, similar TE activity dates were obtained. Although very roughly, these graphs can also be broadly compared since in both the case of *Gypsy* and *Copia* elements, the main LTR activity is found between one and five million years ago, and in both methods, Neotropical and Asian species show more recent LTR activity than African ones and *Hillebrandia*. It is important, however, to keep in mind that translation between genetic distances and insertion dates is prone to error. Nevertheless, despite the possible timing inaccuracies and technical artefacts caused by long-read vs short-read assemblies, the comparison of retrotransposition waves between species still provides clear evidence of differential LTR activation across this group of Begoniaceae.

4.3 Materials and Methods

4.3.1 Repeat and LTR annotation of the 12 Begoniaceae genome assemblies

A Begoniaceae repeat library was obtained using RepeatModeler2 (Flynn et al., 2020) and used to annotate all genomic repeats using RepeatMasker (Smit et al., 2019). For a more accurate LTR annotation, LTR_finder (Xu and Wang, 2007) and LTRharvest (Ellinghaus et al., 2008) were run separately and both outputs were used as input of LTR_retriever (Ou and Jiang, 2018), which generated the final LTR annotations.

4.3.2 Cluster-based repeat annotation of seven Begoniaceae species

The *Begonia foliosa* var. *miniata* reads were downloaded from the Short Nucleotide Archive (SRA study code SRP100595). For all species, the number of trimmed and clean reads equivalent to 20% of the genome length (0.2x coverage) was randomly sampled for each dataset using seqtk (Li, 2012). Reads were then uploaded, pre-processed

and clustered and annotated by similarity using the RepeatExplorer2 pipeline, (Novak et al., 2013) implemented in a Galaxy server environment (https://repeatexplorer-elixir.cerit-sc.cz). For the species that showed high satellite DNA content (*B. foliosa* and *B. luxurians*), a second RepeatExplorer2 run with satellite filtering parameters was performed. For each of the species, cluster annotations were then reviewed and curated by hand, and all those clusters identified as plastid or mitochondrial DNA were removed prior to any downstream analysis.

For the comparative analysis, 440,000 random sequences were sampled for each species, pre-processed and code-named to be pooled into the same dataset. Reads were then clustered and annotated using the satellite filtering and comparative analysis RepeatExplorer2 settings.

4.3.3 Insertion date estimation of assembled and annotated LTRs of the 12 Begoniaceae genomes

Module 4 of the LTR_retriever pipeline includes the insertion dates estimation of all LTRs found by the pipeline. This step is based on aligning and calculating the genetic divergence of the two terminal repeat sequences from the LTRs (which ought to be identical upon insertion). However, insertion dates obtained by LTR_retriever are based on rice substitution rates (Ma and Bennetzen, 2004). A Begoniaceae specific neutral substitution rate was calculated by aligning the introns of 10 ortholog genes (based on 10 BUSCO single copy sequences) and estimating the number of nucleotide variants between *Cucumis sativus* and Begoniaceae sequences. The nucleotide substitution rate was obtained by dividing this proportion of variants by the divergence time between Cucurbitaceae and Begoniaceae (80 Mya). Substitution rates were obtained for the introns of 10 randomly selected genes and their average was used to re-calculate LTR insertion dates.

4.3.4 LTR temporal dynamics estimation using cluster-based LTR annotation

LTR lineages found of *Gypsy*-CRM, *Gypsy*-Tekay and *Copia*-SIRE subfamilies were chosen for temporal dynamics analyses. From the individual clustering results of each

species, a list of clusters belonging to such subfamilies and with over 100 reads annotated as RT-domain sequences were selected. RT-domain sequences were identified within the contigs of those clusters using the RepeatExplorer2 tool DANTE (Hostakova, 2020). Such sequences were extracted and the cluster reads were mapped back to them using the mapping software bwa (default parameters, Li and Durbin (2009)). The bam files were run through the BBmap reformat script (Bushnell, 2014) to obtain the % identity between each read and the reference contig. The % identities were converted into p-distances (% of nucleotide changes), and these into MYA using a Begoniaceae specific substitution rate. This was calculated by aligning the coding sequences of 10 ortholog genes (based on 10 BUSCO single copy sequences) and estimating the number of nucleotide variants between *Cucumis sativus* and Begoniaceae sequences. The nucleotide substitution rate was obtained by dividing this proportion of variants by the divergence time between Cucurbitaceae and Begoniaceae (80 Mya). Substitution rates were obtained for 10 randomly selected genes and their average was used. The transformation from p-distances to insertion dates were done by doubling the obtained substitution rates, with the consideration that transposon sequences mutate at higher rates (SanMiguel et al., 1996; Cossu et al., 2012).

Bibliography

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant molecular biology*. 42:251–269.

Bennetzen JL, Wang H. 2014. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology*. 65:505–530.

Bushnell B. 2014. Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Cavallini A, Natali L, Zuccolo A, et al. (11 co-authors). 2010. Analysis of transposons and repeat composition of the sunflower (helianthus annuus l.) genome. *Theoretical and applied genetics*. 120:491–508.

Cossu RM, Buti M, Giordani T, Natali L, Cavallini A. 2012. A computational study of the dynamics of ltr retrotransposons in the populus trichocarpa genome. *Tree Genetics & Genomes*. 8:61–75.

da Costa ZP, Cauz-Santos LA, Ragagnin GT, Van Sluys MA, Dornelas MC, Berges H, de Mello Varani A, Vieira MLC. 2019. Transposable element discovery and characterization of ltr-retrotransposon evolutionary lineages in the tropical fruit species passiflora edulis. *Molecular Biology Reports*. 46:6117–6133.

Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, Piednoël M, Weiss-Schneeweiss H, Leitch AR. 2015. Genomic repeat abundances contain phylogenetic signal. *Systematic biology*. 64:112–126.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. *BMC bioinformatics*. 9:18.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. Repeatmodeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*. 117:9451–9457.

González LG, Deyholos MK. 2012. Identification, characterization and distribution of

transposable elements in the flax (linum usitatissimum l.) genome. *BMC genomics*. 13:644.

Griesmann M, Chang Y, Liu X, et al. (11 co-authors). 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*. 361.

Hong Z, Li J, Liu X, Lian J, Zhang N, Yang Z, Niu Y, Cui Z, Xu D. 2020. The chromosome-level draft genome of dalbergia odorifera. *GigaScience*. 9:giaa084.

Hostakova N. 2020. Domain based annotation of transposable elements - dante. *Nina Hostakova, GitHub repository. https://github. com/kavonrtep/dante*. .

Ingham L, Hanna WW, Baier J, Hannah L. 1993. Origin of the main class of repetitive dna within selected pennisetum species. *Molecular and General Genetics MGG*. 238:350–356.

Jiménez-Ruiz J, Ramírez-Tejero JA, Fernández-Pozo N, et al. (11 co-authors). 2020. Transposon activation is a major driver in the genome evolution of cultivated olive trees (olea europaea l.). *The Plant Genome*. 13:e20010.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*. 110:462–467.

Kaul S, Koo HL, Jenkins J, et al. (11 co-authors). 2000. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*. 408:796–815.

Kelly LJ, Renny-Byfield S, Pellicer J, et al. (11 co-authors). 2015. Analysis of the giant genomes of f ritillaria (l iliaceae) indicates that a lack of dna removal characterizes extreme expansions in genome size. *New Phytologist*. 208:596–607.

Kelly S. 2019. Editorial overview: Harvesting the fruits of plant genomics. *Current opinion in plant biology*. 48:A1.

Kumar A, Pearce SR, McLean K, Harrison G, Heslop-Harrison J, Waugh R, Flavell AJ. 1997. The ty1-copia group of retrotransposons in plants: genomic organisation, evolution, and use as molecular markers. In: Evolution and Impact of Transposable Elements, Springer, pp. 205–217.

Levin DA. 2002. The role of chromosomal change in plant evolution. Oxford University Press.

Li H. 2012. seqtk. *Heng Li, GitHub repository. https://github. com/lh3/seqtk*. .

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*. 25:1754–1760.

Lippman Z, Gendrel AV, Black M, et al. (11 co-authors). 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature*. 430:471–476.

Lockton S, Gaut BS. 2009. The contribution of transposable elements to expressed coding sequence in arabidopsis thaliana. *Journal of molecular evolution*. 68:80–89.

Lu S, Yang J, Dai X, et al. (11 co-authors). 2019. Chromosomal-level reference genome of chinese peacock butterfly (papilio bianor) based on third-generation dna sequencing and hi-c analysis. *GigaScience*. 8:giz128.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences*. 101:12404–12410.

Macas J, Kejnovskỳ E, Neumann P, Novák P, Koblížková A, Vyskot B. 2011. Next generation sequencing-based analysis of repetitive dna in the model dioceous plant silene latifolia. *PLoS One*. 6:e27335.

Macas J, Neumann P, Navrátilová A. 2007. Repetitive dna in the pea (pisum sativum l.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and medicago truncatula. *BMC genomics*. 8:427.

Macas J, Novak P, Pellicer J, Čížková J, Koblížková A, Neumann P, Fukova I, Doležel J, Kelly LJ, Leitch IJ. 2015. In depth characterization of repetitive dna in 23 plant genomes reveals sources of genome size variation in the legume tribe fabeae. *PloS one*. 10:e0143424.

Mascagni F, Giordani T, Ceccarelli M, Cavallini A, Natali L. 2017. Genome-wide analysis of ltr-retrotransposon diversity and its impact on the evolution of the genus helianthus (l.). *BMC genomics*. 18:634.

Mascagni F, Vangelisti A, Giordani T, Cavallini A, Natali L. 2020. A computational comparative study of the repetitive dna in the genus quercus l. *Tree Genetics & Genomes*. 16:11.

Moonlight PW, Ardi WH, Padilla LA, et al. (11 co-authors). 2018. Dividing and conquering the fastest–growing genus: Towards a natural sectional classification of the mega–diverse genus begonia (begoniaceae). *Taxon*. 67:267–323.

Neumann P, Koblížková A, Navrátilová A, Macas J. 2006. Significant expansion of vicia pannonica genome size mediated by amplification of a single type of giant retroelement. *Genetics*. 173:1047–1056.

Novák P, Hřibová E, Neumann P, Koblížková A, Doležel J, Macas J. 2014. Genome-wide analysis of repeat diversity across the family musaceae. *PloS one*. 9:e98918.

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. Repeatexplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*. 29:792–793.

Ou S, Jiang N. 2018. Ltr_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology*. 176:1410–1422.

Ou S, Su W, Liao Y, et al. (11 co-authors). 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology*. 20:1–18.

Peona V, Weissensteiner MH, Suh A. 2018. How complete are "complete" genome assemblies?—an avian perspective. *Molecular ecology resources*. 18:1188–1195.

Pereira V. 2004. Insertion bias and purifying selection of retrotransposons in the arabidopsis thaliana genome. *Genome biology*. 5:R79.

Piednoël M, Aberer AJ, Schneeweiss GM, Macas J, Novak P, Gundlach H, Temsch EM, Renner SS. 2012. Next-generation sequencing reveals the impact of repetitive dna across phylogenetically closely related genomes of orobanchaceae. *Molecular biology and evolution*. 29:3601–3611.

Piegu B, Guyot R, Picault N, et al. (11 co-authors). 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in oryza australiensis, a wild relative of rice. *Genome research*. 16:1262–1269.

Presting GG, Malysheva L, Fuchs J, Schubert I. 1998. A ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *The Plant Journal*. 16:721–728.

Renny-Byfield S, Kovarik A, Kelly LJ, Macas J, Novak P, Chase MW, Nichols RA, Pancholi MR, Grandbastien MA, Leitch AR. 2013. Diploidization and genome size change in allopolyploids is associated with differential dynamics of low-and high-copy sequences. *The Plant Journal*. 74:829–839.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. *Nature genetics*. 20:43–45.

SanMiguel P, Tikhonov A, Jin YK, et al. (11 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science*. 274:765–768.

Schnable PS, Ware D, Fulton RS, et al. (11 co-authors). 2009. The b73 maize genome: complexity, diversity, and dynamics. *Science*. 326:1112–1115.

Smit A, Hubley R, Green P. 2019. 2013–2015. repeatmasker open-4.0.

Staton SE, Ungerer MC, Moore RC. 2009. The genomic organization of ty3/gypsy-like retrotransposons in helianthus (asteraceae) homoploid hybrid species. *American Journal of Botany*. 96:1646–1655.

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. 2011. Genome size and transposable element content as determined by high-throughput sequencing in maize and zea luxurians. *Genome biology and evolution*. 3:219–229.

Usai G, Mascagni F, Natali L, Giordani T, Cavallini A. 2017. Comparative genome-wide analysis of repetitive dna in the genus populus l. *Tree Genetics & Genomes*. 13:96.

Wan T, Liu ZM, Li LF, et al. (11 co-authors). 2018. A genome for gnetophytes and early evolution of seed plants. *Nature Plants*. 4:82–89.

Wang L, Wu J, Liu X, Di D, Liang Y, Feng Y, Zhang S, Li B, Qi XG. 2019. A high-quality genome assembly for the endangered golden snub-nosed monkey (rhinopithecus roxellana). *GigaScience*. 8:giz098.

White SE, Habera LF, Wessler SR. 1994. Retrotransposons in the flanking regions of normal plant genes: a role for copia-like elements in the evolution of gene structure and expression. *Proceedings of the National Academy of Sciences*. 91:11792–11796.

Wicker T, Sabot F, Hua-Van A, et al. (13 co-authors). 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*. 8:973–982.

Wierzbicki F, Schwarz F, Cannalonga O, Kofler R. 2020. Generating high quality assemblies for genomic analysis of transposable elements. *bioRxiv*. .

Xia E, Tong W, Hou Y, et al. (11 co-authors). 2020. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Molecular plant*. 13:1013–1026.

Xu Z, Wang H. 2007. Ltr_finder: an efficient tool for the prediction of full-length ltr retrotransposons. *Nucleic acids research*. 35:W265–W268.

Yan H, Torchiana F, Bombarely A. 2020. Guideline for genome transposon annotation derived from evaluation of popular te identification tools. *Preprints*. .

Yates AD, Achuthan P, Akanni W, et al. (11 co-authors). 2020. Ensembl 2020. *Nucleic acids research*. 48:D682–D688.

Zhou QS, Luo A, Zhang F, Niu ZQ, Wu QT, Xiong M, Orr MC, Zhu CD. 2020. The first draft genome of the plasterer bee colletes gigas (hymenoptera: Colletidae: Colletes). *Genome Biology and Evolution*. 12:8603–866.

4.A Appendix

4.A.1 Raw numeric data of repeat proportions shown in figures

Repeat annotation results (indicated by genome %) using the genome assemblies

	H. sand	*B. lor*	*B. dre*	*B. con*	*B. fol*	*B. soco*	*B. mas*	*B. pel*	*B. dar*	*B. bipi*	*B. john*	*B. lux*
LINEs	0.07	0.1	0.32	0.15	0.69	0.18	0.1	0.18	0.12	0.16	0.15	0.47
DNA elements	1.01	2.36	3.31	2.1	1.75	2.92	1.44	2.29	1.16	1.9	1.96	2
LTR elements	14.81	13.94	21.36	21.77	13.88	17.74	25.61	19.24	22.01	25.53	15.39	9.7
Small RNA repeats	0.04	0.05	0.05	0.04	0.35	0.06	0.19	0.27	0.42	0.08	0.15	0.05
Satellite DNA	0	0.7	0	0	0	0	0	0	0	0	0	0
Simple repeats	1.36	0.94	1.71	0.84	2.9	4.52	1.74	0.82	1.57	0.78	0.75	1.34
Low complexity repeats	0.38	0.23	0.2	0.2	1.17	0.31	0.17	0.21	0.15	0.19	0.18	0.35
Unclassified elements	26.61	40.5	39.25	33.23	22.26	23.65	42.77	32.27	48.17	38.58	29.37	16.7
Total masked repeats	44.28	58.82	66.2	58.33	43	49.38	72.02	55.28	73.6	67.22	47.95	30.61

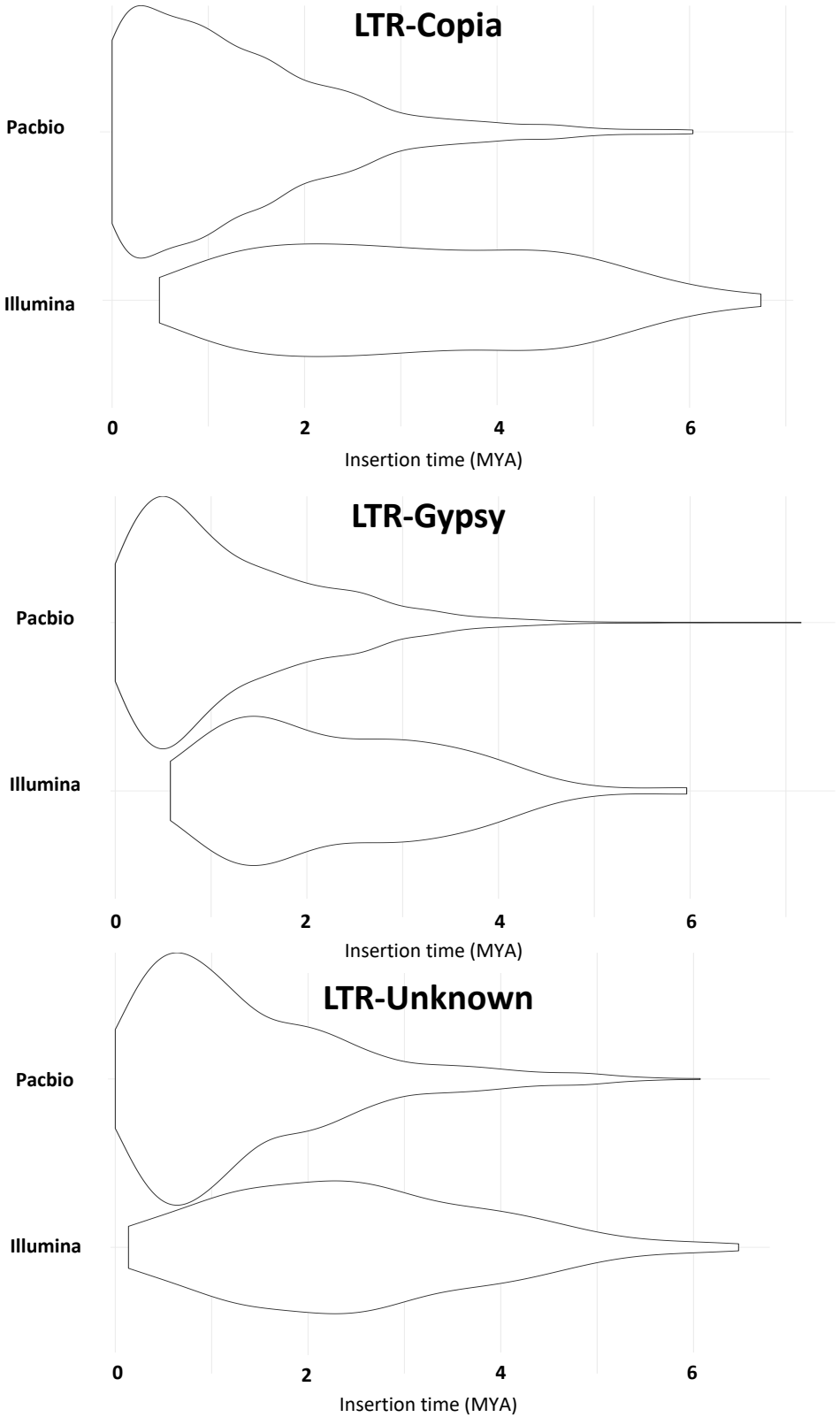
LTR annotation results (indicated by genome %) using the genome assemblies

	H. sand	*B. lor*	*B. dre*	*B. con*	*B. fol*	*B. soco*	*B. mas*	*B. pel*	*B. dar*	*B. bipi*	*B. john*	*B. lux*
Copia	6.77	12.22	11.77	15.67	7.27	7.09	12.96	4.63	9.08	9.35	7.64	5.8
Gypsy	8.4	28.95	23.16	28.71	18.27	20.78	46.77	34.22	33.22	35.06	18.02	10.3
Unknown	14.76	14.75	16.41	14.04	11.51	7.28	12.22	0.68	27.73	13.03	9.54	4.94
LTR-total	29.92	55.92	51.34	58.42	37.04	35.14	71.95	39.52	70.03	57.44	35.2	21.03
Total masked repeats	44.28	58.82	66.2	58.33	43	49.38	72.02	55.28	73.6	67.22	47.95	30.61

Repeat annotation results (indicated by genome %) using the raw reads

	H. sand	*B. dre*	*B. con*	*B. fol*	*B. soco*	*B. bipi*	*B. john*	*B. lux*
LTR-*Copia*-Ale	0.21	0.06	0.23	0	0.02	1.44	0	0
LTR-*Copia*-Angela	0	0.77	0.31	0	0.64	3.87	0.09	0.09
LTR-*Copia*-Ikeros	0.36	0.93	0.11	0.12	0.72	4.24	0.26	0.21
LTR-*Copia*-SIRE	0	1.09	9.97	3.46	0	8.81	1.16	1.51
LTR-*Copia*-Tork	0	1.14	1.39	0.31	0.9	3.19	0.45	1.03
LTR-*Gypsy*-chromovirus-CRM	0.1	8.07	2.81	10.27	12.87	9.51	11.28	3.04
LTR-*Gypsy*-chromovirus-Galadriel	0	0.21	0	0	1.19	1.04	0.23	0.16
LTR-*Gypsy*-chromovirus-Tekay	22.74	11.49	27.15	21.93	1.05	13.08	3.68	2.11
LTR-*Gypsy*-nonchromovirus-Athila	0.05	4.03	1.29	0.35	0.02	1.02	0.07	2
DNA-hAT	1.06	1.74	0.6	1.16	1.09	0.39	0.57	1.18
DNA-mutator	0	1.61	0.35	0.03	1.02	1.12	0.07	1.57
DNA-helitron	0	1.6	0	0	0.15	0	0	0
45S_rDNA	2.84	0.71	0.83	1.06	2.11	5.59	2.02	1.71
5S_rDNA	0.78	0.17	0.03	0.18	0	1.57	0	0.07
Satellite-DNA	0.86	2.12	0.07	9.71	0.69	0.18	0.81	11.39
Unknown	1.4	4.28	3.49	1.33	2.53	2.17	11.09	1
Total	30.4	40.01	48.64	49.93	25	57.24	31.77	27.09

4.A.2 LTR dynamics on long vs. short read assembly of the same species (*B. conchifolia*)



Chapter 5

A closer look: genome dynamics in the highly specious *Begonia* section *Gireoudia*

This chapter is a collaboration with Dr. Daisy Hu, Dr. Yu-Hsin Tseng and Prof. Kuo-Fang Chung from the Biodiversity Research Center at Academia Sinica (Taipei, Taiwan), who kindly provided their genome-skimming data for repeat analysis.

5.1 Introduction

In the previous chapters, I have explored genome dynamics across the Begoniaceae family. Chapter 2 describes the high karyotype diversity across *Begonia*, and shows that karyotypes are generally conserved within sections, also in the highly speciose ones, and that genome size is not strongly correlated to karyotype (Figure 2.6). Chapter 2 also describes genome size variation across large sections with stable karyotypes (Figure 2.6) such as the Southeast Asian sections *Patycentrum*, *Coelocentrum* or *Petermannia*, and the Neotropical sections *Gireoudia*, *Pritzelia* or *Ephemera*. Under the hypothesis that these genome size differences could be triggered by the expansion of repetitive regions in the genome, in the present chapter I will use one of these sections to study the causes of genome size variation between closely related species. Chapter 4 offers a description of the repetitive landscapes of the genomes of four species from four of these highly speciose sections (*B. conchifolia* from sect. *Gireoudia*, *B. masoniana* from sect. *Coelocentrum*, *B. bipinnatifida* from sect. *Petermannia* and *B. luxurians* from sect. *Pritzelia*). Results from these chapters confirmed that retrotransposons made up a high proportion of the genomes of *B. conchifolia, B. masoniana* and *B. bipinnatifida*, and that lower TE proportions but large contents of satellite DNA (satDNA) are present in the genomes of *B. luxurians* and *B. foliosa*. The results in Chapter 4 also suggest that the TEs found in these species show higher recent activity levels than their close relatives from smaller clades. However, it must be acknowledged that our findings and conclusions at this point are drawn from comparing a single species per section. To gain a clearer view of the dynamics of repetitive elements during the radiation of large *Begonia* sections, repetitive landscape comparisons within these sections are needed.

Of all the sections studied in Chapters 2 and 4, *Gireoudia* is the one with the most stable karyotype. Forty two of its 110 species have been karyotyped, all with 2n=28. However, Legro and Doorenbos (1969) describes supernumerary B chromosomes in at least three *Gireoudia* species. Chapter 2 describes genome sizes in *Gireoudia* ranging from 586 to 735 Mbp, although this data represents only 4 species. This indicates that although *Gireoudia* species have stable karyotypes, genomic-level changes are still occurring between these species. *Gireoudia*'s phylogenetic position makes it convenient

for genomic comparisons between specious and depauperate sections, as according to (Moonlight et al., 2018), its two sister sections (*Parietoplacentalia* and *Urniformia*) have 3 and 1 species respectively.

Gireoudia is a large, monophyletic section that has been the subject of many previous *Begonia* studies. *Gireoudia* species have been found to show low rates of gene flow (Twyford, 2012) and these species (as most *Begonia*) are also found to be generally narrow endemics (Dewitte et al., 2011). Additionally, *Gireoudia* species were studied by Moonlight (2017) as a case of non-adaptive radiation, meaning rapid genome dynamics followed by genetic drift may have played a principal role in this section's rapid radiation. Under the hypothesis that large rates of repetitive DNA change have played a role in the *Gireoudia* radiation, different evolutionary scenarios can be considered. On the one hand, a specific transposon/repeat expansion at the origin of the radiation could have triggered high levels of genomic variability leading to faster genetic isolation of populations. On the other hand, many different and independent bursts of repetitive DNA could have led to specific reproductive isolation between populations and speciation. These two scenarios, however, do not exclude one another, and it could be indeed the case that both an initial transposon burst and subsequent repetitive DNA expansions across the evolution of this section contributed to such large and explosive radiation.

In this chapter, genomic data from 34 species (32 *Gireoudia* and 2 outgroup species from *Parietoplacentalia* and *Urniformia*) will be analysed and repeat content compared. Species sampled for genomic analysis and their phylogenetic position are presented in Figure 5.1. Specifically, this chapter aims to:

- Use a genome skimming approach to identify differences in the repeat landscape between the species in the speciose section *Gireoudia* and the depauperate sections *Parietoplacentalia* and *Urmifornia*.
- Determine whether the expansion and loss of repetitive elements is the cause of genome size variation within section *Gireoudia*.
- Date *Gireoudia*-specific repeat expansions relative to the origin of this section, to

determine whether genome dynamics could have a role the rapid radiation of this section.

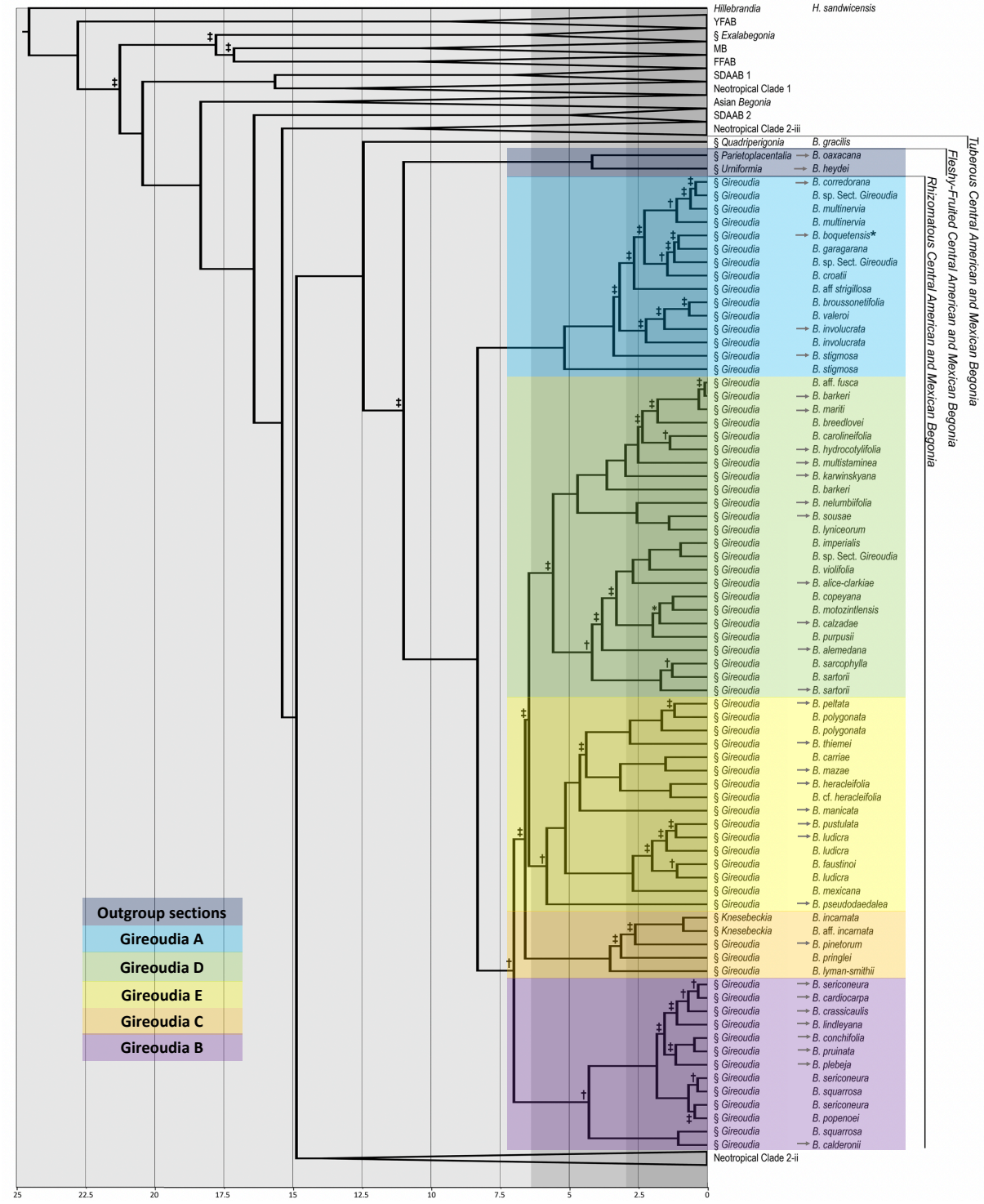


Figure 5.1: Neotropical clade 2-i phylogeny as shown in Moonlight et al. (2018), modified with colour codes for each *B.* sect. *Gireoudia* group within the section. Arrows next to species name indicate the species sequenced for this chapter.

5.2 Results and Discussion

5.2.1 Genomic repeat landscapes of 34 closely related Neotropical *Begonia*

The genomic data from 34 species was analysed in the exact same way as in the clustering analysis of read subsets presented in Chapter 4. The RepeatExplorer2 pipeline (Novak et al., 2013) was the main tool used for both individual and comparative analyses, and its output data was analysed in several ways to retrieve the most information possible about the repetitive DNA landscape in these species, and to test the hypothesis of repetitive DNA as a species radiation driver in *Gireoudia*.

Figure 5.2 shows the summarised results of the individual species repeat clustering analysis (data table in appendix 5.A.1). It represents the percentage of input data clustered as each repeat family, but only those families that are represented by over 1% of the data in at least one species are shown (data table in appendix 1). As expected and previously seen in Chapter 4, most of the repetitive elements are LTRs. Among these, the different *Gypsy* and *Copia* families show variable abundance, with most of the data clustered as LTR-*Gypsy*-Tekay (6 to 40%), and LTR-*Copia*-SIRE (4 to 15%), as also seen across the genus in Chapter 4. Results show that both these families are significantly more abundant among *Gireoudia* species than in the *Urmifornia* and *Parietoplacentalia* sections (p-values 0.02 and 0.00). Less abundant but also taking between approximately 0.5 and 5% of the data in most species are the LTR-*Copia*-Tork, and LTR-*Gypsy* CRM and Athila families. Two of these families seem to be in constant proportions in all analysed species (Tork and Athila), however the *Gypsy*-CRM family is found to be significantly more abundant in outgroup sections than across *Gireoudia* (p-value 0.01). All species show fairly constant content (0-1.5%) of other LTRs. DNA elements are also more abundant in the outgroup sections compared to *Gireoudia* species (p-values 0.0 for both DNA transposon families shown), although these proportion differences could be due to the lower LTR content found in the outgroups.

Figure 5.2 also shows that Satellite DNA (satDNA) is extremely variable across the 32 *Gireoudia* species analysed. Percentages vary between 0% in *B. alice-clarkiae*, *B.*

pruinata and *B. plebeja*, and 25.7% in the closely related *B. calderonii*. These values are highly variable even within *Gireoudia* sub-sections or groups, showing species-specific expansions of satellite DNA in most cases, which suggests very recent and likely ongoing expansions of satDNA families.

The species analysed from *Gireoudia* clade A (Gir A, blue) do not show large satellite expansions. In *Gireoudia* clade B (Gir B, purple), most species show low (0-5%) satellite DNA (satDNA) except for *B. calderonii*, which shows a satDNA content of 25%. This represents the largest difference within clades, and considering this species diverged from its closest relatives about $\sim$ 2.5-5 MYA (Figure 5.1) and assuming an average *Gireoudia* genome size of $\sim$ 600 Mbp, it suggests either a satDNA expansion of $\sim$ 150Mbp in under 5MY, or strong selective pressure against these satellite sequences in all other species in the lineage leading to multiple losses. This is also the case for *Gireoudia* clades D (Gir D, green) and E (Gir E, yellow) with satellite DNA contents that oscillate between 0 (*B. alice-clarkiae*) and 23.5% (*B. sartorii*) in Gir D, and between 0.5 (*B. pseudoadalea*) and 24% in *B. pustulata* in Gir E. This implies very rapid gain/loss of satDNA within *B.* sect. *Gireoudia* that could have led to genomic incompatibilities linked with rapid speciation. The rapid evolution of centromere sequences (which are generally made up by satDNA) has previously been studied and proposed to lead to genome instability (Balzano et al., 2020), reduced compatibility of homologous chromosomes and ultimately postzygotic isolation, consequently triggering speciation (Dawe and Henikoff, 2006; Ma et al., 2007; Plohl et al., 2014).

The variation in satDNA content found, and its correlation with the genome size values shown in Chapter 2 suggest that differences in satDNA content are the main cause of genome size variation in *Gireoudia*. However, satDNA estimations using RepeatExplorer2 can be biased by, for example, different library preparation methods (Macas and Novák, personal communication). Library replicates in some species as well as more *Gireoudia* genome size estimations would be needed to further test this correlation.

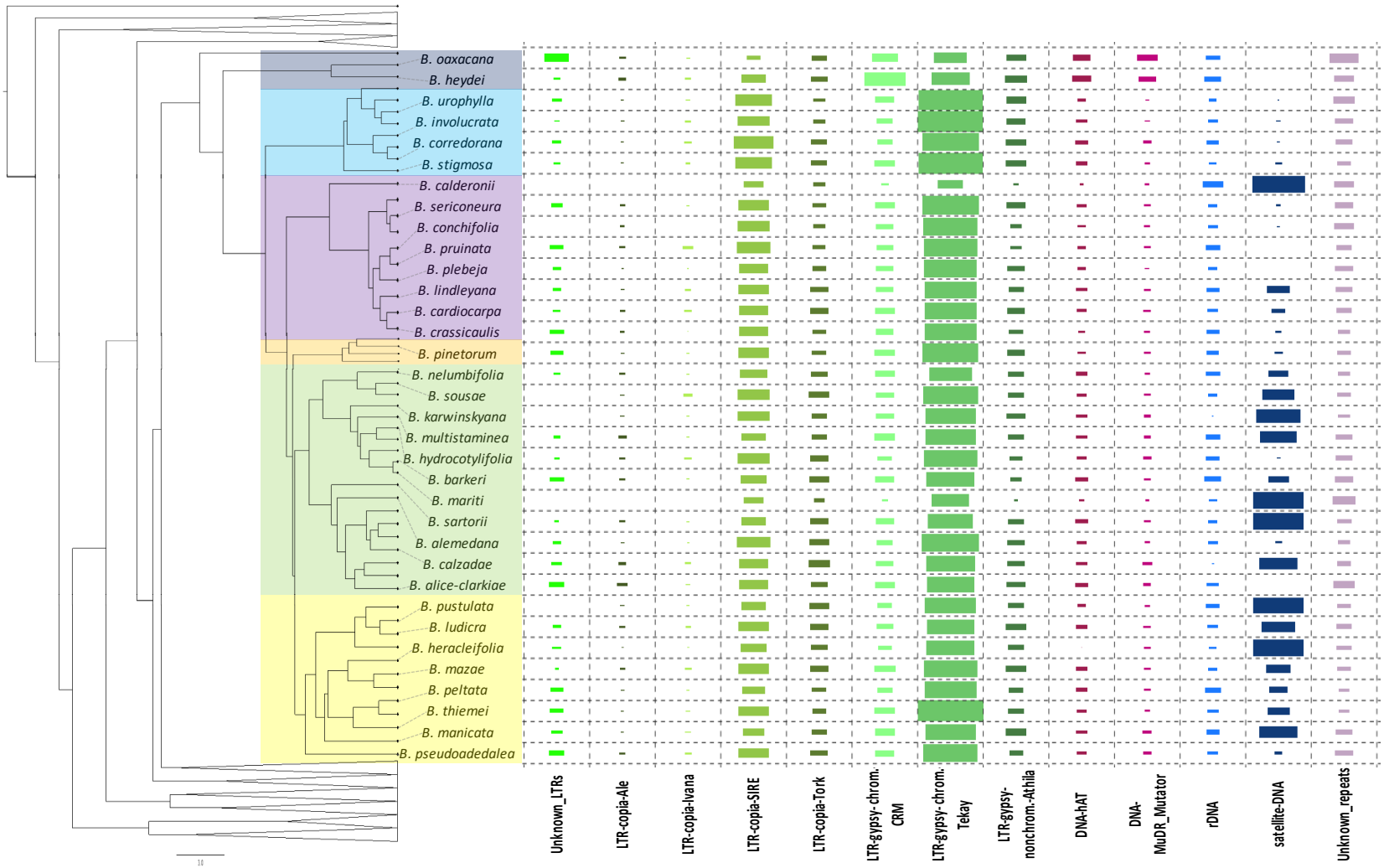


Figure 5.2: Distribution of percentages of reads clustered and annotated as different types of repeats by RepeatExplorer2 in each of the 34 species analysed in this chapter, and their phylogenetic position. Phylogeny modified from Moonlight et al. (2018) by E. Gagnon, unpublished. LTR lineages are indicated in different shades of green, DNA TEs in maroon and small repeats in blue. Only repeat families represented by over 1% of the reads in at least one species are shown.

Together with these rapid changes in satDNA, variation in LTR content across *Gireoudia* can be observed in Figure 5.2. A clear pattern is that species with low LTR content (*B. calderonii, B. maritii*) have higher satDNA content. This is likely due to the variable number of reads processed by RepeatExplorer2 for each accession. Although the same genome coverage was used as input for all species, the computational power this tool uses is limited and the analysis of satellite sequences by TAREAN (Novák et al. (2017), implemented in RepeatExplorer2) is computationally intensive. Samples with large proportions of satellite DNA are not analysed in full, which implies that transposons and other non-satellite repeats will be under-represented in the results. To further confirm whether these LTR expansions/losses are an artefact, I performed RepeatExplorer2 comparative analysis between all species. However, because 34 species are too many to compare in the same run, I performed 6 different comparative runs: a first one selecting the outgroup species and one or two representatives of all five *Gireoudia* clades, and then five runs including all species from the outgroup and clades A, B, D and E.

Are any specific types of repeat more expanded in *Gireoudia* species than in sister sections?

The first comparative analysis, shows an overview of what types of transposons and repeats are shared across *Gireoudia*, but not shared or less expanded in the two sister sections *Urmifornia* and *Parietoplacentalia*. *B. oaxacana* and *B. heydei* were included in a comparative analysis with *B. corredorana* and *B. stigmosa* from Gir A, *B. calderonii* and *B. cardiocarpa* from Gir B, *B. pinetorum* from Gir C, *B. hydrocotylifolia* and *B. calzadae* from Gir D, and *B. pseudoadalea* and *B. ludicra* from Gir E. Figure 5.3 shows an overview of the results of this comparative analysis. Graphical representations of the most abundant repeat families superclusters (SCs, groups of clusters that share paired-end data) are shown in this figure, indicating which clusters are equally shared across all samples, and which are expanded/unique in a group of species.

Figure 5.3 shows that, as previously seen in Figure 5.2, the most abundant repeat families are LTR-*Gypsy*-Tekay, LTR-*Copia*-SIRE, satDNA and LTR-*Gypsy*-CRM. However, it is interesting to observe the proportion of each SC from each species. The

LTR-*Gypsy*-Tekay and LTR-SIRE-*Copia* superclusters show a low relative abundance of *B. oaxacana* and *B. heydei* reads, suggesting these two families are expanded in the *Gireoudia* section. Across the nine *Gireoudia* species included in this analysis, most Tekay and SIRE clusters seem to be equally shared, but a few of them (CL1, CL3, CL22, CL31, SC 38) seem expanded in *B. corredorana* and *B. stigmosa*, the two species from *Gireoudia* clade A. Other minor ones seem approximately equally shared by all species, representing LTR lineages with unequal expansion or loss across these species (SC17). This could either imply a concerted activity after these lineages separated, or a shared expansion before the split.

Among the main repeat families shown in Figure 5.3, as expected the most unequally expanded and specific for certain species are satellite DNA clusters. The dramatic expansion/loss of satDNA between closely related species in *Gireoudia* was already observed in Figure 5.2, and results of this comparative analysis show that these satDNA dynamics are not caused only by one satDNA family expanding across species, but rather by several different ones. SC3 seems present in most but not all species, and this satDNA seems to be most expanded in *B. calderonii* (Gir B) and *B. ludicra* (Gir E). On the other hand, SC15 and SC23 seem mostly expanded in single species (*B. calderonii* and *B. calzadae*), but also present in other species (from different Gireoudia subclades) in small number. As well as SC37, which is equally expanded in *B. cardiocarpa* (Gir B) and *B. pseudoadalea* (Gir E). This indicates there are several satDNA families expanded across *Gireoudia*, and these are not conserved within subclades. Further analysis on the origin of these satDNA sequences and their position within the chromosome would help understanding the processes underlying this rapid satDNA dynamism and its evolutionary consequences.

The fourth most abundant repeat family found is the LTR-*Gypsy*-CRM family, which as can also be observed in Figure 5.2. This is more abundant in the outgroup sections species *B. oaxacana* and *B. heydei*, and less present in *Gireoudia* in most superclusters but 39. Figure 5.3 also shows CRM superclusters 7 and 31 are more expanded in *B. heydei* than in *B. oaxacana*, suggesting this LTR family was active after the split of these two sections.

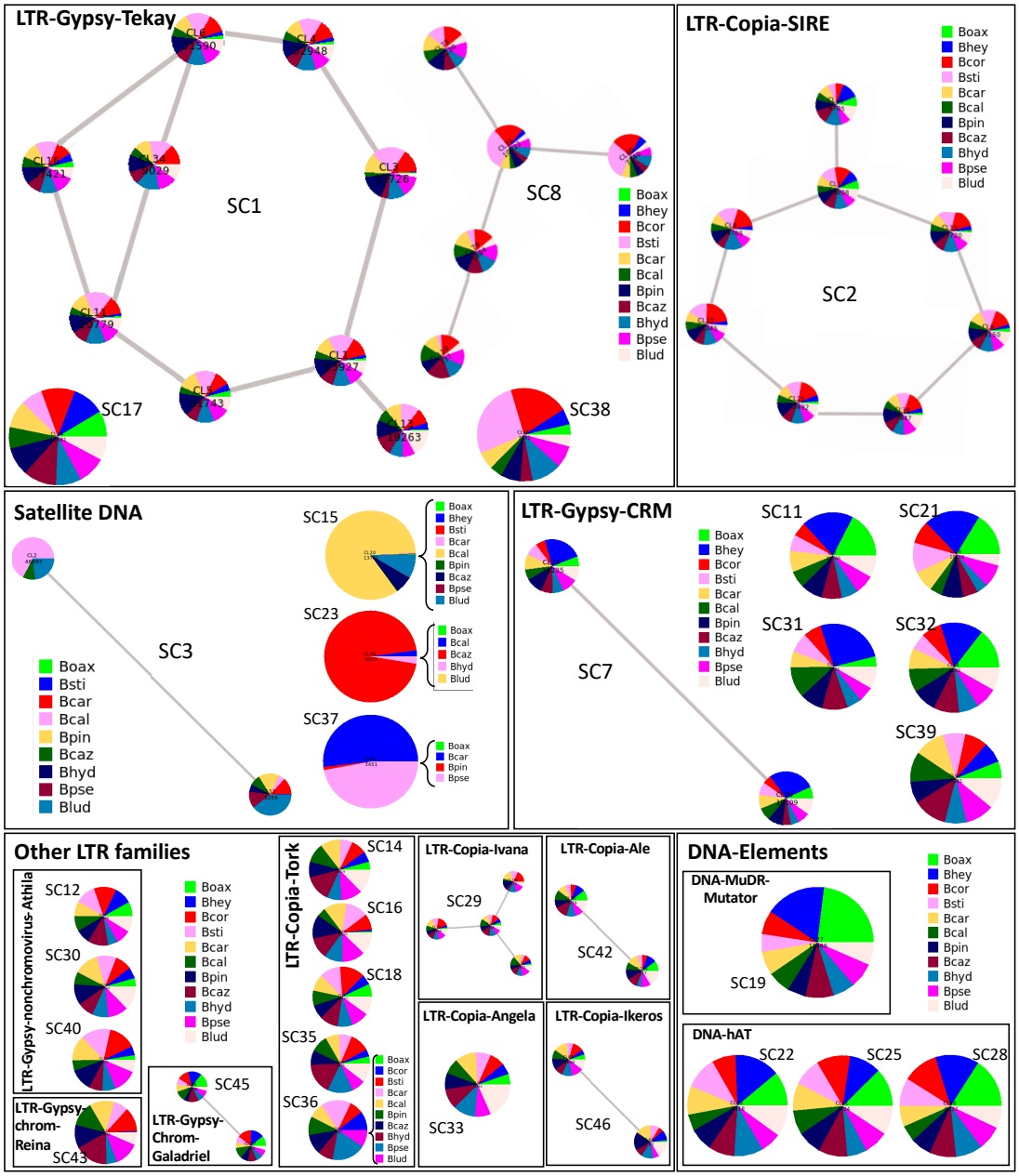


Figure 5.3: Main superclusters (SCs) and the relative abundance of the analysed species across *B.* sect. *Urmifornia, Parietoplacentalia* and *Gireoudia* in each of them. SC are represented by the all the repeat clusters (here represented by each piechart) they share paired data with. SCs shown in this figure are only the annotated as the main repeat families found among the first (most abundant) fifty. Because some SC do not have representation of all species, some pie-charts show their own legends by their side. The legends show 4 characters codes for all the species present in each supercluster, being Boax for *B. oaxacana*, Bhey for *B. heydei*, Bcor for *B. corredorana*, Bsti for *B. stigmosa*, Bcar for *B. cardiocarpa*, Bcal for *B. calderonii*, Bpin for *B. pinetorum*, Bcaz for *B. calzadae*, Bhyd for *B. hydrocotylifolia*, Bpse for *B. pseudoadalea*, and Blud for *B. ludicra*. Grey lines connect clusters within the same supercluster.

The rest of the LTR families shown in Figure 5.3 seem to be more expanded in *Gireoudia* species than in the outgroup sections, except from the LTR-*Gypsy*-Galadriel family, which shows more unequal distribution and more expanded in *B. oaxacana* and B. *heydei*, as well in some specific *Gireoudia* species. Other LTR families such as LTR-*Copia* Angela, Ikeros and Ale although generally expanded in *Gireoudia*, show unequal expansions in specific *Gireoudia* species, suggesting gain and loss of these LTR families after the split of closely related species. In the case of DNA elements, however, all shown superclusters seem to be expanded in the outgroup species compared to *Gireoudia*, which also confirms what the species individual RepeatExplorer2 analyses indicated in Figure 5.2 and indicate DNA element expansions after the *Gireoudia* split.

The results of this comparative analysis provide evidence of a *Gireoudia*-specific expansion of three out of four of the most abundant repeat families (as well as a few minor ones) found in these 34 *Begonia* species. SatDNA seems the most dynamic repeat family within *Gireoudia* in subclades B, D and E. The only *Gireoudia* group that did not show any variation in satDNA (Gir A) shows a larger amount of Tekay and SIRE elements in some clusters of SC1, 2 and 8, as well as in SC38 (Figure 5.3). The fact that we found no satDNA variation in Gir A could however also be due to an artefact in the cluster sampling, being sequenced only the species without satDNA expansions. However, this seems unlikely since we sampled and sequenced half the species of this clade, which is a similar proportion to that sequenced for the other clades. Therefore, the genomic event that triggered the satDNA dynamics in *Gireoudia* either happened after the split of Gir A, or satDNA expansions have been lost or negatively selected in Gir A species. The only other *Gireoudia* clade where satDNA dynamics cannot be confirmed is in Gir C, which is underrepresented in this chapter and only data of one out of four species is shown.

The only repeat family that seems more enriched in the outgroup species is the LTR-*Gypsy*-CRM family. This pattern of CRMs being more expanded and active in small, depauperate clades compared to sister speciose ones was also seen in the comparisons made in Chapter 4 between African and Neotropical and Asian species. *B. dregei*, *B. socotrana* and *B. johnstonii* contain larger species-specific CRM clusters than their

Neotropical and Asian relatives (Figure 4.6 and Figure 4.7). This could be an indicative that CRM LTRs accumulate in older species with more stable genomes, and they are less present in proportion to other LTRs in species with more dynamic genomes. This family is, however, not abundant in *Hillebrandia*. Figure 5.3 confirms this is not due to the relative difference of other LTRs in *Gireoudia*, but rather an absolute difference in the number of CRM elements. This LTR family was initially assigned to centromeric regions, but research on the sequence and distribution of these elements across a number of plant species described that this group of elements is not genuinely centromeric (Neumann et al., 2011). This study also concludes that CRM activity plays a role in normal centromere function, suggesting perhaps the lower CRM activity/content described in some *Begonia* triggers abnormal centromere function. A low amount and activity of CRMs was also linked with the evolution of other LTR elements and the speciation of *Oryza* species (Zhang and Gao, 2017).

Having taken an overview of repeat dynamics in *Gireoudia* clades compared to depauperate outgroups, my next aim is observing whether these genomic changes are conserved within *Gireoudia* clades, and what repeat families are the most recently active and recently expanded. Figure 5.3 shows that the repeats showing most species and group specific expansions are satellite DNA clusters, so it is expected that species within clades will show variation in satDNA. However, from the previous results it is unclear how active/differentially expanded other LTRs are within clades. To answer these questions, five RepeatExplorer2 comparative analyses were performed between species of the different groups in Figure 5.1 (Outgroup species and Gir A, B, D and E). From the output of each of these analyses, I gathered all the clusters that were unequally expanded by selecting those showing a species distribution ratio $\pm$ 50% of the expected in at least one species. Pie charts showing the proportion of these from each main repeat family for each species group are shown in Figure 5.4.

These results confirm that the most dynamic elements in these species are LTR-*Gypsy* elements and satDNA. More specifically, *Gypsy*-CRM elements are differentially expanded outside *Gireoudia*, whereas other *Gypsy* element expansions are seen in *Gireoudia* clades. *Gypsy*-Tekay is the main family showing differences between closely re-

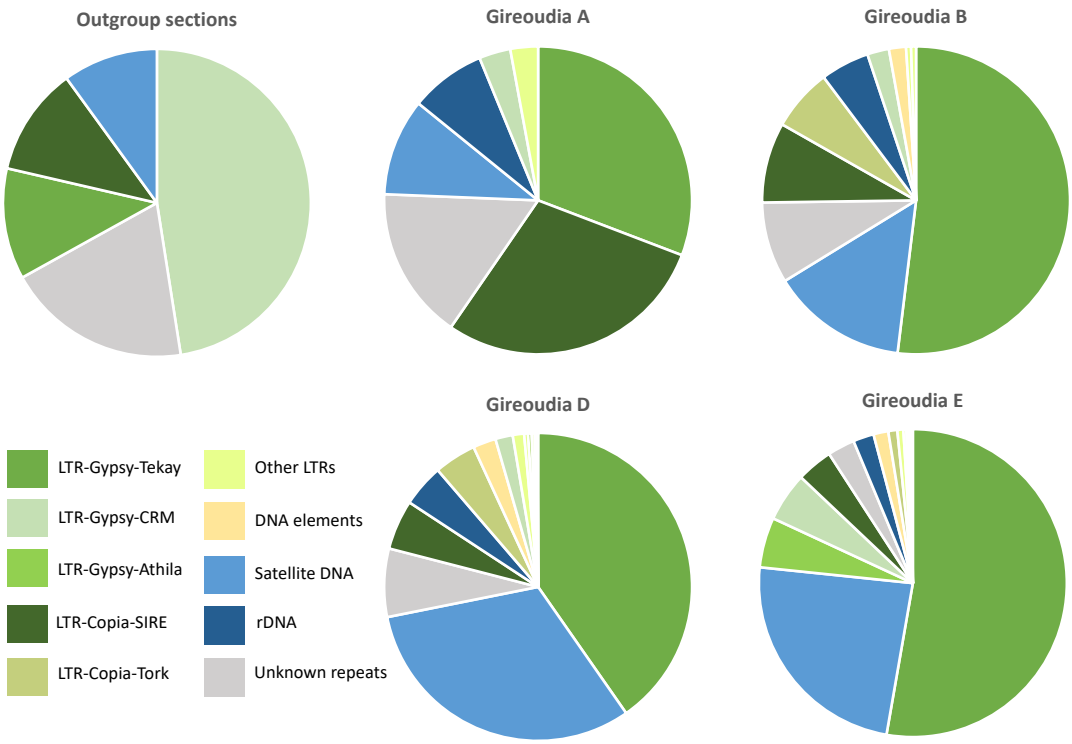


Figure 5.4: Total size (reads) of all the clusters unevenly expanded (± 50% of the expected ratios) in each species sub-section, grouped by repeat family.

lated species, but in clade Gir A the SIRE family is almost as differentially expanded as Tekay. In the rest of *Gireoudia* sub-sections satellite DNA is the second most common differentially expanded group compared to Tekay elements.

All *Gireoudia* groups show high levels of differentially expanded ribosomal DNA (rDNA). Ribosomal genes compose the Nucleus Organiser Region (NOR), and are generally present in several copies, organised in tandem arrays. These genomic regions tend to be conserved in phylogenetically close species, and are therefore generally used for cytogenetic studies to characterise a given species, genus or group (Garcia et al., 2009, 2012). Recent work (Picart-Picolo et al., 2020) describes the effects of rDNA loss in *Arabidopsis thaliana* mutants. This research provides evidence of rapid and large tandem duplications, changes in chromatin organisation and gene expression changes in a few generations of these rDNA mutants. If rDNA changes have such consequences in such short evolutionary time, these differences in rDNA levels across closely related *Begonia* species would need to be further investigated to link them to satDNA expansions and other genomic changes.

These last comparative analysis results confirm the high levels of genome dynamics within *Gireoudia*. SatDNA and TE variation was found even between species recently diverged. This provides further evidence and confirms the pattern found in Chapter 4 of higher TE and satDNA dynamics in highly speciose *Begonia* groups. Nonetheless, these results show a static view of the genomes and make it difficult to link higher TE activity and general genome dynamics to large radiation or speciation events. There is also the possibility that what we are actually seeing is consequence of repeat loss instead of expansions. To obtain better proof and further test our main hypothesis, linking a specific repeat expansion event or TE burst with a specific species radiation is needed. With this aim, I obtained as much information as possible from our RepeatExplorer2 data, extracted the RT-LTR sequences found from all the 34 species studied in this chapter and analysed their temporal dynamics by plotting their sequence divergence levels as done in Chapter 4.

5.2.2 LTR insertion dynamics of 34 Neotropical *Begonia*

All clusters showing more than 100 reads matching RT domains within the first (largest) hundred of each dataset were selected for RT divergence calculations. Considering normally around 200-300 clusters are annotated in RepeatExplorer2 runs, this represents a variable portion of the data output. However, after a hundred clusters, it was difficult to find LTR clusters with 100 RT sequences, and that is why the threshold of the first hundred clusters was established. Between 10 and 30 LTR clusters of each of the 34 analysed species were dated. To provide an overview of the different LTR dynamics in each subclade, Figure 5.5 shows the RT divergence dynamics of one or two clusters per LTR family and species. Any overlapping RT activity profiles were removed to leave only one, with the aim of displaying LTR activity variation as clearly as possible.

Figure 5.5 shows the LTR temporal dynamics varies across species and between LTR families. The activity of some LTR families is roughly correlated as seen in the overlapping plots, and some LTR families show species or group-specific peaks of activity. TE activity of SIRE elements roughly correlated across all *Gireoudia* groups with the exception of Gir A which shows more recent activity and clade-specific peaks.

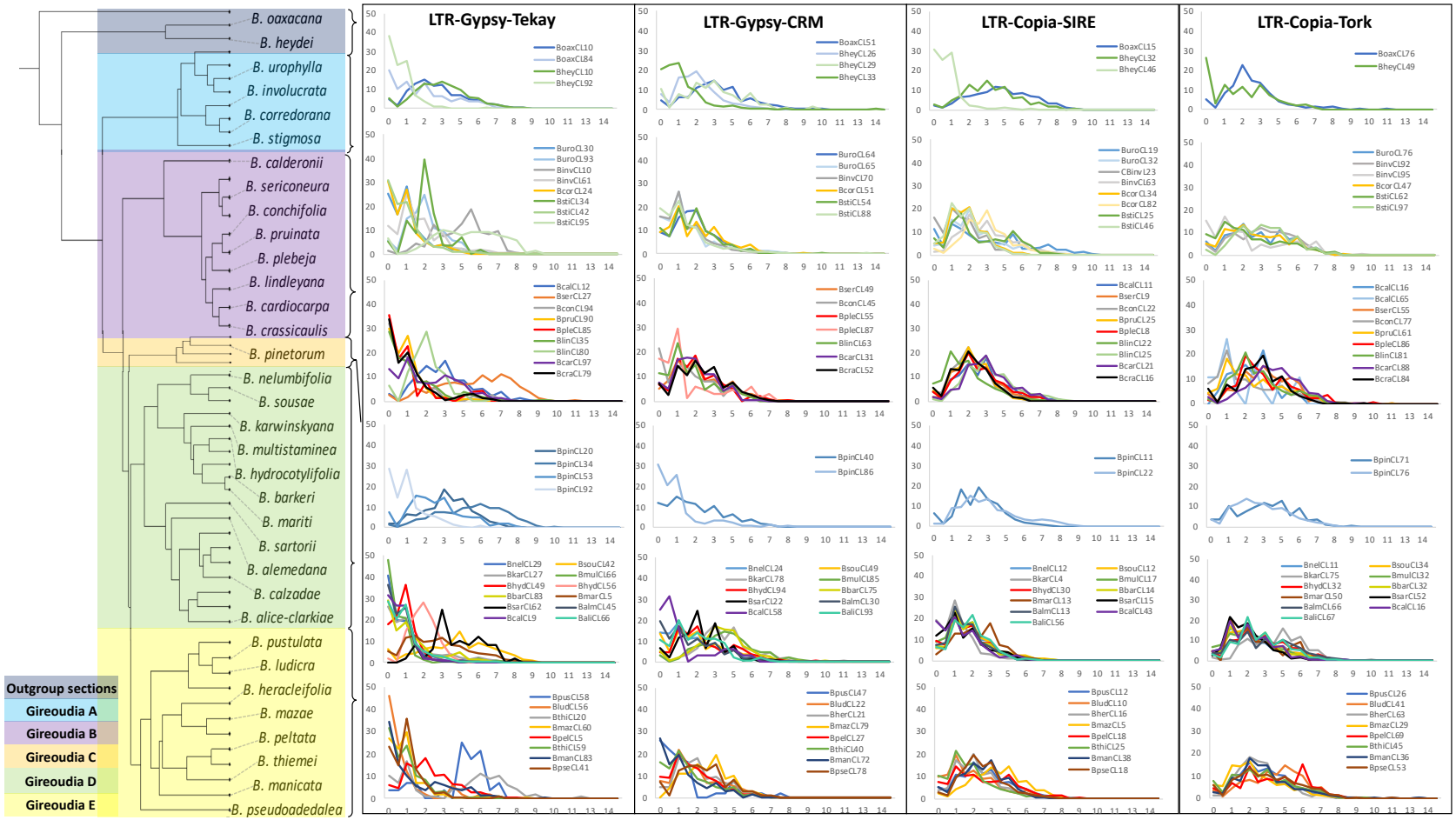


Figure 5.5: Timing of different LTR clusters retrotranspositional activity in the analysed *Begonia* species based on pairwise comparisons of Illumina reads to the cluster contig regions that match RT-encoding domains, and the phylogenetic position of these species (Moonlight et al., 2018). Each plot shows the RT activity of a specific species group and LTR family. Their X axis indicate the divergence time in MYA, and their Y axis indicate the percentage of reads showing that specific divergence to the reference. Line colours indicate different species, colour shade indicates different clusters. Species codes in the legend indicate the first three letters of each species name.

This agrees with the specific GirA SIRE expansion seen in Figure 5.3 and Figure 5.4. Tork elements also seem more recently active in Gir B, and a differential expansion can also be seen in Figure 5.4. However as expected, the LTR family with the highest and most extensive peaks of activity is the *Gypsy*-Tekay family. These peaks are unique to *Gireoudia*, because outgroup species show smoother activity curves and no main peaks. As would be expected given the results shown in Figure 5.4, the two outgroup species do show more unequal and species-specific peaks of *Gypsy*-CRM activity.

So far our RT-LTR temporal dynamics analysis results confirm what was already suggested by our comparative analysis results, in which Tekay (and SIRE in Gir A) elements are found to be differentially expanded between all these closely related species. However, another main aim of this analysis is finding any *Gireoudia*-specific LTR activity that could be linked to this section's origin. Figure 5.5 shows that the oldest LTR activity peaks we have been able to date are Tekay elements, which show their first (datable by this method) expansions $\sim$ 9 MYA. Figure 5.1 indicates that according to (Moonlight et al., 2018), the origin of the *Gireoudia* section has been dated between 10 and 7.5 MYA. However, and as discussed in Chapter 4, molecular dating methods are prone to error and our estimations are rough. This ancient LTR-*Gypsy*-Tekay activation shared in *Gireoudia* species but not in their closest sections could have played a role in the *Gireoudia* radiation.

In Chapter 4, we were able to in-depth examine the repeat landscape of 12 Begoniaceae species across the genus. We found several specific transposon expansions, and in general that most genomes from species in highly speciose clades were more dynamic than their African relatives. However, these clades were too old and large to be able to link any specific genome expansions or repetitive element dynamics to a specific radiation event. In this chapter, I have used the monophyletic, highly speciose *Begonia* section *Gireoudia* to analyse genome dynamics at a smaller phylogenetic scale, study repeat landscape changes between closely related species, and try to associate specific retrotransposition activity to the origin of this clade. Analysis on the repeatome of 34 species does not only prove that some repeat families are highly dynamic within *Gireoudia* and more stable in the sister and small sections *Parietoplacentalia* and *Urniformia*, but also

provides evidence of a *Gireoudia* specific expansion of LTR-*Gypsy*-Tekay elements that is dated roughly at the same time as the origin of the section *Gireoudia*. Further analysis on the genomic location of these Tekay elements would help understand whether these expansions had an impact on the evolutionary trajectory of this major *Begonia* section.

This chapter highlights satDNA as another highly dynamic repeat family in *Begonia*. In Chapter 4 we also found high amounts of satDNA in the genomes of two Neotropical species studied, *B. foliosa* and *B. luxurians*. Whether the expansion of these satDNA families are linked to LTR activity or, as mentioned above, rDNA changes, cannot be inferred from this data. Recent work (Vondrak et al., 2020) found that satDNA families are frequently originated by LTR elements that ocasionally expand in length. This suggests that the highly active Tekay elements could be responsible for the satDNA expansion found in *Gireoudia*. On the other hand, as previously mentioned, the origin of large satDNA duplications has also been described to be triggered by rDNA changes (Picart-Picolo et al., 2020). Both investigations used nanopore reads to get a better view of the location and impact of these elements. In the case of *Begonia*, the nanopore data of *B. conchifolia* originally used in Chapter 3 to improve its assembly could be used to further investigate the origin of these satDNA families. However, in this chapter *B. conchifolia* was found to be one of the few *Gireoudia* species without an expansion of satDNA and therefore data from other species may be needed for comparison. Whatever the origin of these tandem repeats is, Picart-Picolo et al. (2020) describe satDNA expansions can modify the genome structure and stability, as well as trigger gene duplications and change expression patterns. This could potentially play a role in speciation events by generating rapid genetic incompatibilities between populations. Comparative analysis of satDNA families across different plant groups also indicate the rapid diversification of some satDNA families (Helm and Hemleben, 1997; Cuadrado and Jouve, 2002; Zhang et al., 2014), but the evolutionary consequences are still unclear. The role of satDNA in *Begonia* speciation will be further studied and discussed in Chapter 6.

5.3 Materials and Methods

5.3.1 Plant material

The specimens sequenced were part of both the living collection of the Biodiversity Centre in Academia Sinica (Taipei) and the Royal Botanic Garden Edinburgh. *B. oxacana* material was provided by the RBG Kew (A. Monro, personal communication). A full list of all species accession number and collection of origin is shown in Table 5.1.

Table 5.1: Species ID, collection and accession ID from all specimens sequenced in this chapter.

Species	Section	Collection	Accession No.
B. urophilla	*Gireoudia*	Academia Sinica	Chang221
B. barkeri	*Gireoudia*	Academia Sinica	Chang233
B. mazae	*Gireoudia*	Academia Sinica	Chang239
B. plebeja	*Gireoudia*	Academia Sinica	Chang244
B. thiemei	*Gireoudia*	Academia Sinica	Chang249
B. peltata	*Gireoudia*	Academia Sinica	Chang254
B. ludicra	*Gireoudia*	Academia Sinica	Chang265
B. karwinskyana	*Gireoudia*	Academia Sinica	Chang275
B. hydrocotylifolia	*Gireoudia*	Academia Sinica	Chang291
B. sartorii	*Gireoudia*	Academia Sinica	Hu334
B. sericoneura	*Gireoudia*	Academia Sinica	Hu337
B. heracleifolia	*Gireoudia*	Academia Sinica	Hu344
B. mariti	*Gireoudia*	Academia Sinica	Hu345
B. pustulata	*Gireoudia*	Academia Sinica	Hu347
B. involucrata	*Gireoudia*	Academia Sinica	Hu354
B. alice-clarkiae	*Gireoudia*	Academia Sinica	Hu367
B. nelumbiifolia	*Gireoudia*	Academia Sinica	Hu377
B. lindleyana	*Gireoudia*	Academia Sinica	Hu395
B. stigmosa	*Gireoudia*	Academia Sinica	Hu404
B. sousae	*Gireoudia*	RBGE	20100334
B. multistaminea	*Gireoudia*	RBGE	20071056
B. almedana	*Gireoudia*	RBGE	20100266
B. urophylla	*Gireoudia*	RBGE	20082084
B. pseuduodaedalea	*Gireoudia*	RBGE	20100379
B. pruinata	*Gireoudia*	RBGE	20070430
B. manicata	*Gireoudia*	RBGE	20100428
B. corredorana	*Gireoudia*	RBGE	20071055
B. cardiocarpa	*Gireoudia*	RBGE	20051398
B. pinetorum	*Gireoudia*	RBGE	20100268
B. calzadae	*Gireoudia*	RBGE	20100348
B. crassicaulis	*Gireoudia*	RBGE	20131495
B. calderonii	*Gireoudia*	RBGE	20100288
B. heydei	*Urniformia*	RBGE	N/A
B. oaxacana	*Parietoplacentalia*	RBGKew	N/A

5.3.2 Data acquisition

The data used in this chapter comes from two different sources. The 32 *Gireoudia* species were sequenced by our colleagues in Academia Sinica (Hu et al., personal communication). They used NEBNext Ultra II DNA Library Prep Kits for Illumina (New England BioLabs, US)for library preparation and performed Illumina HiSeq 2500 sequencing, obtaining around 600 MB paired end data (150 bp reads) per sample.

DNA isolation and library preparation of *B. heydei* and *B. oaxacana* were carried out by AllGenetics & Biology SL (www.allgenetics.eu) using the Isolate II Plant kit (Bioline, Meridian Bioscience, US) and the NEBNext Ultra II DNA Library Prep kit (New England Biolabs, US). Samples were then shipped to Novogene (UK) to be sequenced (Illumina NovaSeq), where a total of 4 gigabases of data were obtained.

5.3.3 Data pre-processing and analysis

Cluster-based repeat annotation of 34 *Begoniaceae* species

An average genome size using all the known C-values for *Gireoudia* was obtained, and the number of trimmed and clean reads equivalent to 20% of this value (0.2x coverage) was randomly sampled for each dataset using seqtk (Li, 2012). Reads were then uploaded, pre-processed and clustered and annotated by similarity using the RepeatExplorer2 pipeline, (Novak et al., 2013) implemented in a Galaxy server environment (https://repeatexplorer-elixir.cerit-sc.cz). For the species that showed high satellite DNA content, second RepeatExplorer2 runs with satellite filtering parameters were performed. For each of the species, cluster annotations were then reviewed and curated by hand to resolve automated annotation conflicts, and all those clusters identified as plastid or mitochondrial DNA were removed prior to any downstream analysis. Statistical differences between species groups were found with t-tests.

For the comparative analysis, 300,000 random sequences were sampled for each species, pre-processed and code-named to be poolled into the same dataset. Due to the large number of datasets, two species were sampled from the larger species groups (A, B, D

and E) and a general comparative analysis was done. In addition, different comparative analyses with all species from each group (Outgroup, Gireoudia A, B, D and E) were performed. Reads were then clustered and annotated using the satellite filtering and comparative analysis RepeatExplorer2 settings.

RT-divergence estimation using cluster-based LTR annotations

For all species individual repeat analyses, all LTR lineages found between the 100 first (largest) clusters were chosen for temporal dynamics analyses. First, all LTR clusters with over 100 reads annotated as RT-domain sequences were selected. RT-domain sequences were then identified within the contigs of those clusters using the RepeatExplorer2 tool DANTE (Hostakova, 2020). Such sequences were extracted and all the reads from such cluster were mapped back to them using the mapping software bwa (default parameters, Li and Durbin (2009)). The bam files were ran through the BBmap reformat script (Bushnell, 2014) to obtain the % identity between each read and the reference contig. The % identities were converted into p-distances (% of nucleotide changes), and these into MYA using a Begoniaceae specific substitution rate. This was calculated by aligning the coding sequences of 10 ortholog genes (based on 10 BUSCO single copy sequences) and estimating the number of nucleotide variants between *Cucumis sativus* and Begoniaceae sequences. The nucleotide substitution rate was obtained by dividing this proportion of variants by the divergence time between Cucurbitaceae and Begoniaceae (80 Mya). Substitution rates were obtained for 10 randomly selected genes and their average was used to re-calculate LTR insertion dates. The transformation from p-distances to insertion dates were done by doubling the obtained substitution rates, with the consideration that transposon sequences mutate at higher rates (SanMiguel et al., 1996; Cossu et al., 2012).

Bibliography

Balzano E, Pelliccia F, Giunta S. 2020. Genome (in) stability at tandem repeats. In: Seminars in Cell & Developmental Biology. Elsevier.

Bushnell B. 2014. Bbmap: a fast, accurate, splice-aware aligner. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Cossu RM, Buti M, Giordani T, Natali L, Cavallini A. 2012. A computational study of the dynamics of ltr retrotransposons in the populus trichocarpa genome. *Tree Genetics & Genomes*. 8:61–75.

Cuadrado A, Jouve N. 2002. Evolutionary trends of different repetitive dna sequences during speciation in the genus secale. *Journal of Heredity*. 93:339–345.

Dawe RK, Henikoff S. 2006. Centromeres put epigenetics in the driver's seat. *Trends in biochemical sciences*. 31:662–669.

Dewitte A, Twyford A, Thomas D, Kidner C, Van Huylenbroeck 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*. 448:2029.

Garcia S, Garnatje T, Kovařík A. 2012. Plant rdna database: ribosomal dna loci information goes online. *Chromosoma*. 121:389–394.

Garcia S, Garnatje T, Pellicer J, McArthur ED, Siljak-Yakovlev S, Valles J. 2009. Ribosomal dna, heterochromatin, and correlation with genome size in diploid and polyploid north american endemic sagebrushes (artemisia, asteraceae). *Genome*. 52:1012–1024.

Helm MA, Hemleben V. 1997. Characterization of a new prominent satellite dna of cucumis metuliferus and differential distribution of satellite dna in cultivated and wild species of cucumis and in related genera of cucurbitaceae. *Euphytica*. 94:219–226.

Hostakova N. 2020. Domain based annotation of transposable elements - dante. *Nina Hostakova, GitHub repository. https://github. com/kavonrtep/dante*. .

Legro R, Doorenbos J. 1969. Chromosome numbers in begonia. *Netherlands Journal of Agricultural Science*. 17:189–202.

Li H. 2012. seqtk. *Heng Li, GitHub repository. https://github. com/lh3/seqtk*. .

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*. 25:1754–1760.

Ma J, Wing RA, Bennetzen JL, Jackson SA. 2007. Plant centromere organization: a dynamic structure with conserved functions. *TRENDS in Genetics*. 23:134–139.

Moonlight PW. 2017. Linking speciation and the niche: taxonomy, phylogeny and niche evolution in neotropical Begonia. Ph.D. thesis, University of Glasgow.

Moonlight PW, Ardi WH, Padilla LA, et al. (11 co-authors). 2018. Dividing and conquering the fastest–growing genus: Towards a natural sectional classification of the mega–diverse genus begonia (begoniaceae). *Taxon*. 67:267–323.

Neumann P, Navrátilová A, Koblížková A, Kejnovskỳ E, Hřibová E, Hobza R, Widmer A, Doležel J, Macas J. 2011. Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA*. 2:1–16.

Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. 2017. Tarean: a computational tool for identification and characterization of satellite dna from unassembled short reads. *Nucleic acids research*. 45:e111–e111.

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. Repeatexplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*. 29:792–793.

Picart-Picolo A, Grob S, Picault N, et al. (11 co-authors). 2020. Large tandem duplications affect gene expression, 3d organization, and plant–pathogen response. *Genome Research*. 30:1–10.

Plohl M, Meštrović N, Mravinac B. 2014. Centromere identity from the dna point of view. *Chromosoma*. 123:313–325.

SanMiguel P, Tikhonov A, Jin YK, et al. (11 co-authors). 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science*. 274:765–768.

Twyford AD. 2012. Speciation and gene flow in Central American Begonia L. (Begoniaceae). Ph.D. thesis, The University of Edinburgh.

Vondrak T, Robledillo LÁ, Novák P, Koblížková A, Neumann P, Macas J. 2020. Characterization of repeat arrays in ultra-long nanopore reads reveals frequent origin of satellite dna from retrotransposon-derived tandem repeats. *The Plant Journal*. 101:484.

Zhang H, Koblížková A, Wang K, et al. (11 co-authors). 2014. Boom-bust turnovers of megabase-sized centromeric dna in solanum species: rapid evolution of dna sequences associated with centromeres. *The Plant Cell*. 26:1436–1447.

Zhang QJ, Gao LZ. 2017. Rapid and recent evolution of ltr retrotransposons drives rice genome evolution during the speciation of aa-genome oryza species. *G3: Genes, Genomes, Genetics*. 7:1875–1885.

5.A Appendix

5.A.1 Raw numeric data of repeat proportions shown in figures

	Unknown LTRs	LTR *Copia* Ale	LTR *Copia* Ivana	LTR *Copia* SIRE	LTR *Copia* Tork	LTR *Gypsy* chrom CRM	LTR *Gypsy* chrom Tekay	LTR *Gypsy* nonchrom Athila	DNA hAT	DNA MuDR Mutator	rDNA	satDNA	Unkwn repeats	Total in clusters
B.oaxacana	5.46	0.41	0.13	1.67	2.26	6.31	9.91	3.64	3.03	4.14	2.06	0.00	7.64	48.29
B.heydei	0.44	0.57	0.30	5.89	2.62	15.69	13.55	4.60	3.16	2.99	2.62	0.00	3.78	57.82
B.urophylla	0.90	0.09	0.20	12.78	1.46	3.45	38.52	3.71	0.72	0.15	0.59	0.02	4.50	67.35
B.involucrata	0.30	0.12	0.41	9.17	1.51	2.45	39.52	3.40	1.24	0.25	0.91	0.19	2.91	63.06
B.corredorana	0.77	0.08	0.55	14.38	2.38	2.09	29.79	3.90	1.43	0.60	1.44	0.09	2.65	60.90
B.stigmosa	0.39	0.09	0.21	12.13	2.41	3.88	38.04	3.83	1.09	0.29	0.57	0.43	1.87	65.98
B.calderonii	0.00	0.00	0.00	3.43	1.43	0.57	5.97	0.24	0.14	0.19	3.85	25.74	3.56	45.20
B.sericoneura	1.12	0.24	0.11	8.45	1.85	3.57	29.52	3.26	0.94	0.46	0.76	0.16	3.35	54.79
B.conchifolia	0.00	0.23	0.00	9.97	1.39	2.81	27.15	1.29	0.60	0.35	0.86	0.07	3.49	48.64
B.pruinata	1.72	0.33	1.06	10.70	1.60	2.77	26.54	1.12	0.65	0.40	1.97	0.00	2.03	51.53
B.plebeja	0.60	0.04	0.02	8.03	1.74	2.87	25.40	2.69	0.66	0.21	0.77	0.00	3.02	46.72
B.lindelyana	0.59	0.14	0.41	8.67	3.10	2.80	24.99	2.15	1.31	0.42	1.52	4.89	2.43	54.67
B.cardiocarpa	0.57	0.28	0.48	7.86	3.23	3.77	25.15	2.94	0.92	0.53	1.07	1.87	2.27	51.89
B.crassicaulis	1.85	0.22	0.02	7.86	1.75	2.81	24.99	1.87	0.50	0.36	1.54	0.39	1.30	46.13
B.pinetorum	1.59	0.11	0.08	9.03	1.94	3.97	29.08	3.01	0.67	0.40	1.41	0.69	1.69	54.46
B.nelumbifolia	0.48	0.32	0.07	7.06	2.74	3.53	16.77	2.39	1.15	0.27	1.94	3.56	1.47	43.02
B.sousae	0.00	0.23	0.73	9.53	3.92	2.98	27.93	2.56	0.79	0.41	0.84	9.71	1.70	62.42
B.kawinskyana	0.00	0.19	0.12	9.44	2.13	3.24	23.42	3.25	1.16	0.48	0.03	18.04	1.39	64.13
B.multistaminea	0.49	0.64	0.10	5.27	2.04	3.95	23.70	2.31	1.10	0.48	1.93	12.71	2.53	58.21
B.hydrocotylifolia	0.26	0.33	0.50	9.77	3.03	1.89	26.34	1.53	0.95	0.66	1.81	0.14	2.60	50.52
B.barkeri	2.00	0.33	0.12	6.40	3.72	3.33	21.29	1.21	1.46	0.46	2.55	3.87	3.06	50.87
B.mariti	0.00	0.00	0.00	3.89	0.98	0.30	12.71	0.11	0.28	0.15	0.63	23.40	4.95	47.52
B.sartorii	0.19	0.39	0.12	5.64	3.09	3.23	18.96	2.50	1.46	0.45	0.81	23.55	1.96	63.41
B.almedana	0.60	0.16	0.06	10.10	3.53	2.45	30.05	2.91	0.85	0.25	0.90	0.44	1.29	54.77
B.calzadae	1.10	0.60	0.36	7.61	4.04	2.77	22.05	2.81	1.36	0.96	0.04	13.31	1.86	61.49
B.alice-clarkiae	2.22	1.10	0.16	7.83	2.66	3.53	21.11	3.25	1.56	0.54	1.32	0.00	4.32	51.06
B.pustulata	0.00	0.19	0.06	5.79	3.49	1.95	23.98	2.54	0.72	0.31	1.76	23.94	1.77	67.53
B.ludicra	0.73	0.34	0.28	8.33	3.26	2.52	20.82	3.74	1.34	0.54	1.06	10.37	2.35	56.96
B.heracleifolia	0.82	0.04	0.05	6.24	2.55	1.81	21.68	2.48	0.01	0.39	0.55	23.22	2.28	62.53
B.mazae	0.16	0.29	0.42	9.05	3.04	4.29	26.82	3.72	1.08	0.38	0.79	5.44	1.61	58.69
B.peltata	1.45	0.11	0.09	4.82	1.96	2.27	24.93	2.09	1.34	0.41	2.41	3.05	1.14	47.06
B.thiemei	1.74	0.07	0.19	9.00	1.91	3.85	40.00	2.27	0.87	0.32	1.22	4.61	0.96	67.48
B.manicata	1.18	0.07	0.33	4.51	2.35	3.77	23.90	4.11	1.38	0.75	1.61	13.51	2.51	60.86
B.pseudoadalea	2.13	0.35	0.44	8.28	2.91	3.36	27.09	1.79	1.01	0.64	1.12	0.52	3.04	53.83

Chapter 6

An even closer look: intra-specific genome dynamics and population divergence in *Begonia heracleifolia* (*B.* sect. *Gireoudia*)

This chapter includes the work and results of the Plant Sciences Honours Project of Thea E. Kongsted (University of Edinburgh). Most of the work described here was equally shared by both of us and therefore plural pronouns will be used throughout this chapter.

Also included here is part of the collaborative work on *Begonia* genome size variation with Dr. Jaume Pellicer-Moscardó (Royal Botanic Gardens, Kew), who estimated the *Begonia heracleifolia* genome size values studied in this chapter.

6.1 Introduction

In previous chapters, I have attempted to link radiation events to repatome dynamics by analysing differences at the genome architecture level of species within the Begoniaceae family (Chapters 2 and 4) and within one of the largest sections (Chapter 5). In these chapters, I have been able to describe genome dynamics at different levels, and find changes in TE activity that are specific to individual radiations. Although the LTR retrotransposition dynamics analyses performed can give us some evidence of recent TE activity, it is difficult to link a specific transposition, repeat expansion or chromosomal rearrangement event to a specific speciation event to confirm their evolutionary role. Comparing genomic data from different species can only provide an estimation of sequence divergences, and comparing these with species diversification dates can only roughly place bursts of TE expansion in the evolutionary timeline. This makes the link between TE bursts and species radiation an indirect one. Moreover, to understand the mechanisms facilitating or constraining speciation it is important to study the effects of the many factors involved not only at the end of the process, but at different points along the speciation continuum (Via, 1999). Strong evidence that TE activity is linked to species radiations will require comparative genomics at an intraspecific level. By analysing genomic changes and their link to population divergence levels, we would be catching "speciation in the act", and being much closer to find which specific factors drive the genomic incompatibilities that can lead to speciation.

Chapter 5 results indicate that species from the large section *B* sect. *Gireoudia* show higher levels of satDNA variation and LTR activity than their depauperate sister sections, suggesting an association of these repeat expansions with the repeatome dynamics found in the *Gireoudia* radiation. Twyford et al. (2014) studied genetic differentiation and species cohesion in two *Gireoudia* species from central America: *B. heracleifolia* and *B. nelumbifolia*. Their research found significant population structure and species differentiation in *B. heracleifolia*, and a 20% reduction in pollen viability in interpopulation crosses involving a divergent *B. heracleifolia* population that was found to have a genome size 10% larger than the species mean C-value (Dewitte et al., 2009). Their data suggests a weak gene flow connection between populations, low enough to allow

allopatric divergence over the likely timescales, suggesting that the population with a larger genome size could be in a potential route towards speciation.

Few other studies have investigated reproductive isolation in diverse tropical taxa, and those that have (Pinheiro et al., 2013) suggest that this variation is likely the outcome of small effect changes at several loci. Alternatively, the genome size differences found in Twyford et al. (2014) suggest that this reproductive isolation could also have been driven by large-scale genomic changes in *B. heracleifolia*. To better understand the relationship between C-value differences and reproductive isolation a genome sequencing approach to identify the causes of genome size differences, as well as more genome size estimations of other populations and more genetic data testing crossing barriers are needed.

In this chapter, we aim to further investigate the C-value variation across these natural *B. heracleifolia* populations, identify the factors driving this intraspecific genome divergence and measure its effects on reproductive isolation. With the main hypothesis being that repetitive DNA elements are driving these genomic differences and the emergence of genetic barriers between populations, the main objectives of this chapter are the following:

- Fully describing genome size variation across more *B. heracleifolia* populations, as well as across individuals from the same population, as suggested in Twyford et al. (2014). In this initial study, five populations were measured for genome size differences. For our analyses, we include three more plus an artificial F1 population from an interpopulation cross.

- Investigating whether the population-specific C-value increase described by Twyford et al. (2014) is due to the expansion of repetitive elements, and if that would be the case, which specific elements are being expanded. For this purpose, low-coverage whole-genome sequencing data was generated for 10 *B. heracleifolia* accessions.

- Testing whether the described genome size variation is associated with intraspecific crossing barriers by analysing seed viability of inter and intra-specific crosses.

  A subset of seven *B. heralceifolia* populations from the described in Twyford

et al. (2014) comprising the most geographic areas possible were selected for this study. The geographic location, population and accession numbers of the accessions studied in this chapter are specified in Figure 6.1.

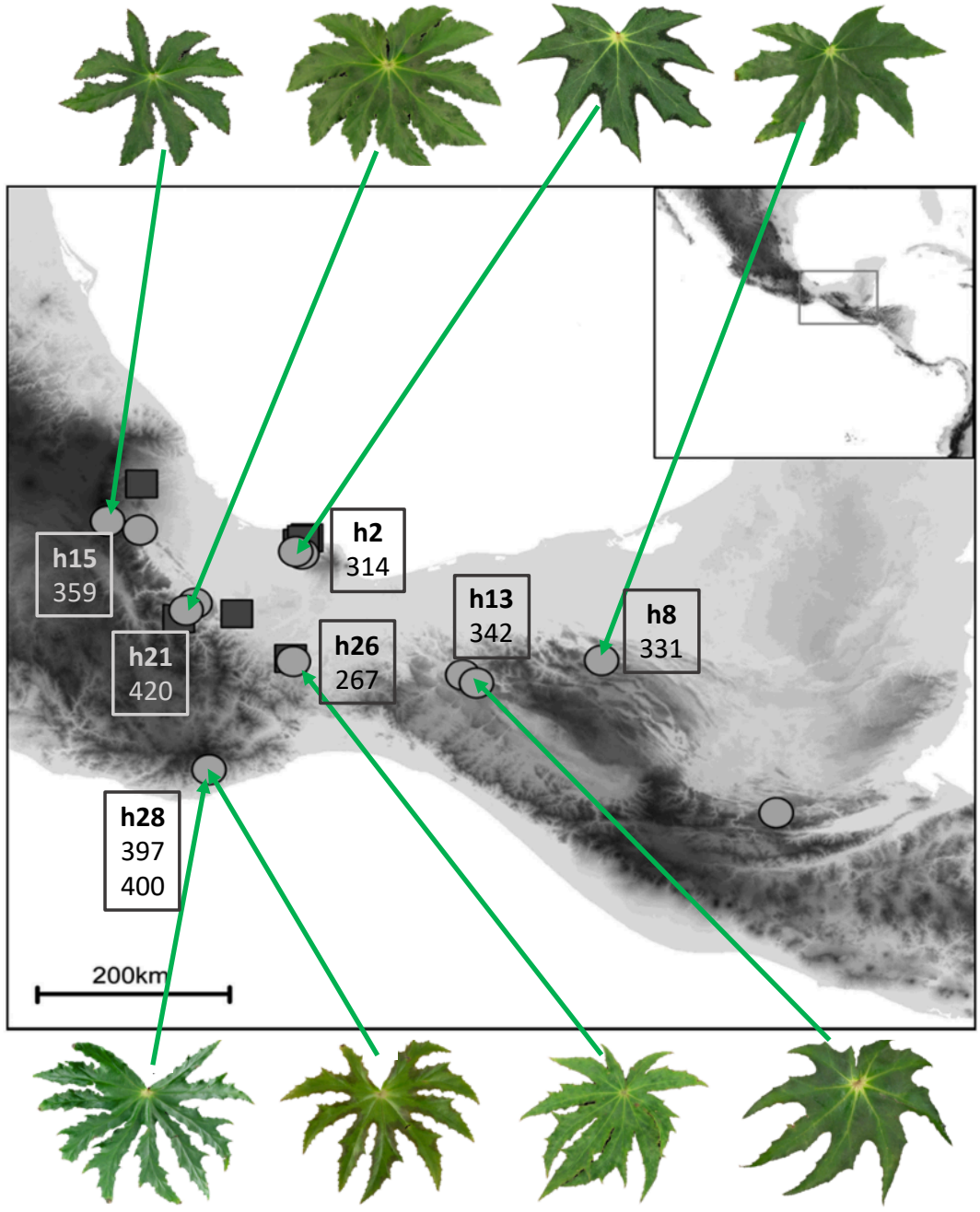


Figure 6.1: Geographical location of the 7 *B. heracleifolia* populations studied in this chapter. Different circles and squares in the map represent the location of all the *B. heracleifolia* and *B. nelumbifolia* populations analysed in Twyford et al. (2014). The *B. heracleifolia* populations (circles) studied in this chapter are indicated with green arrows and boxes next to them, containing the population reference and the accession number of the individuals analysed. The arrows indicate leaf pictures of each accession. Modified from Twyford et al. (2014)

.

6.2 Results and discussion

6.2.1 Genome size differences of 10 *Begonia heracleifolia* populations

Although genome size differences between some of these *B. heracleifolia* populations are already described in Twyford et al. (2014), leaves of all the individuals included in our WGS experiment were measured to confirm their C-values. Results are shown in Figure 6.2.

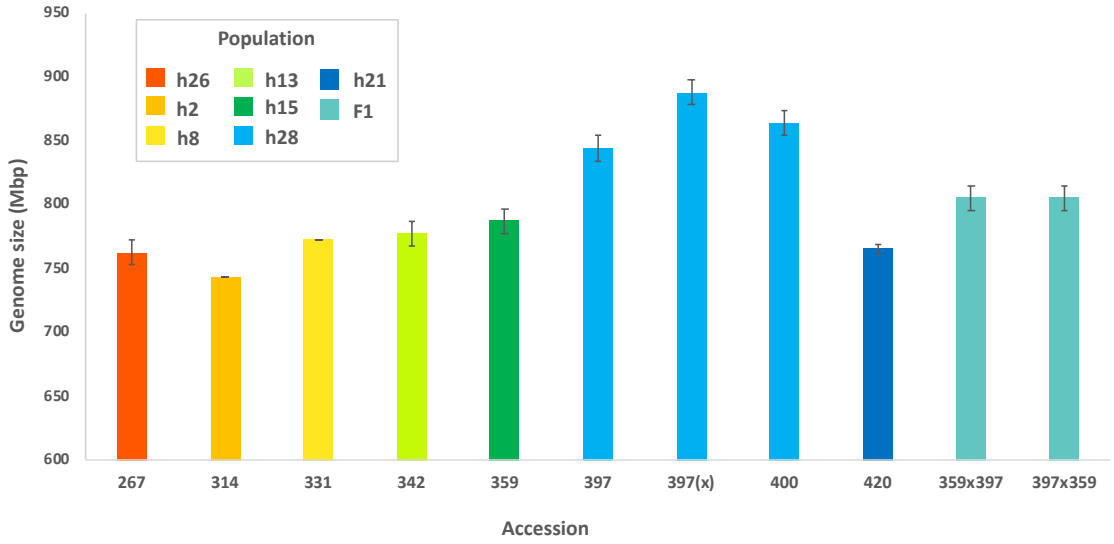


Figure 6.2: Haploid genome size differences found between the *B. heracleifolia* populations from Twyford et al. (2014) one selfed individual (397(x)) and two inter-population hybrids between h15 and h28. Bars indicate average values and error bars indicate the standard deviations of 3 measurements per sample.

Our data shows slight genome size variation in populations h2 to h26 and confirms that, as indicated by Twyford et al. (2014), population h28 individuals have a larger genome than all the other populations. Among the h28 analysed individuals, C-value variation can be seen too, representing a difference of 44 Mbp between accessions 397 (844 Mbp) and 397(x) (888 Mbp). The population with the smaller genome size is h2 (743 Mbp), and the difference with h28 accessions is up to 145Mbp, which means C-value differences between populations represent between 10 and 20% of their genomes. Genome sizes of inter-population hybrids show the same intermediate genome sizes between their two parent accessions (805 Mbp), but these are slightly closer values to their smaller genome parent. This, together with the larger genome of 397(x) compared to 397, indicates this "extra DNA" may not follow mendelian segregation. However,

the fact that both F1 accessions show the same genome sizes suggests the opposite. Additional F1 genome size data would be needed to further test how this inheritance works.

6.2.2 WGS and repeat content analysis of 10 *Begonia heracleifolia* populations

We obtained low-coverage Illumina datasets (around 1 Gb per sample of 150bp reads from 350 bp insert libraries) from 10 accessions from seven different populations, which are the same ones from Figure 6.2. We performed standard quality control analysis for each of the sequenced accessions (Andrews and others, 2010), and GC-distribution curves showed a variable presence of AT-richer regions in accessions 397, 400 and the two F1. These GC distribution curves of all datasets are shown in Figure 6.3. Secondary peaks with lower GC content can be seen in all accessions with larger genomes, being lower in the F1 accessions. These peaks could be indicative of the presence of higher amounts of satellite DNA in the larger genome accessions, since satellite repeats are often found to be AT-rich regions (Garrido-Ramos, 2015). The intermediate height of the F1 secondary peaks correlate with their intermediate genome size, and suggests these AT-rich regions are likely to be the cause of genome size variation.

Random read subsets representing 20% of the genome (0.2x genomic coverage) were sampled for each accession, and used as input for the repetitive DNA analysis with RepeatExplorer2 (Novak et al., 2013). The repeat proportion distributions of each accessions are represented in Figure 6.4 (data table in appendix 6.A.1). As expected from a *B.* sect. *Gireoudia* species (Chapter 5), large amounts of LTR-*Gypsy*-Tekay (11 to 25%) LTR-*Copia*-SIRE (3.5 to 7.8%) elements are found in all accessions. These, together with satellite DNA are the most abundant repetitive DNA lineages found by RepeatExplorer2.

Figure 6.4 shows that the amount of Satellite DNA found in each population is variable, and higher in h28 and F1 individuals. The proportions of analysed reads clustered as satellite repeats varies from 10.6% in accession 314 (also known to have the smallest genome, as seen in Figure 6.2) to 25% in accession 397(x). Satellite DNA content

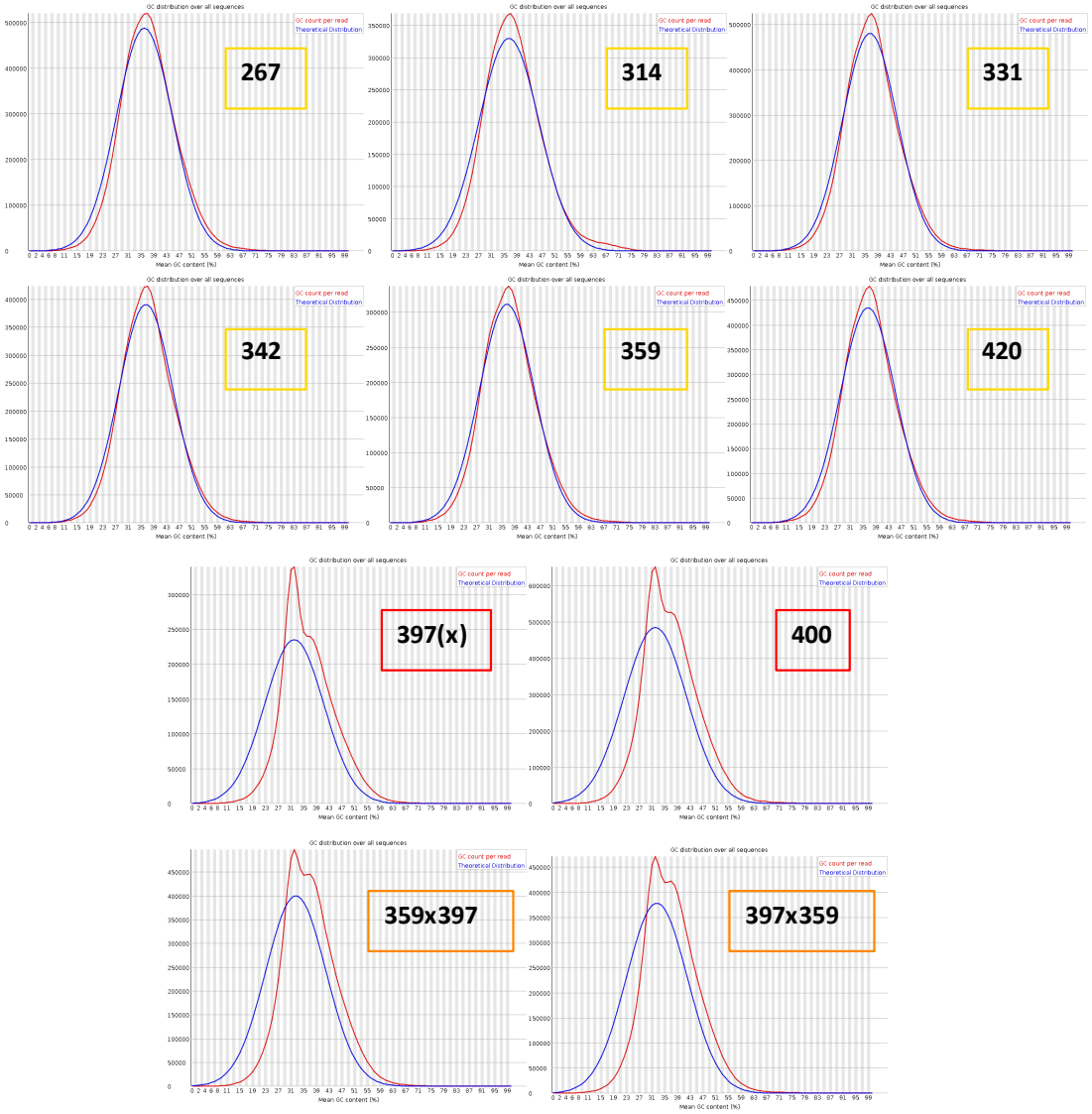


Figure 6.3: GC content distribution from raw datasets of all populations sequenced. All smaller genome size accessions (yellow squares) show normal GC content distributions. Individuals from the h28 population (red squares) show a peak of AT-rich sequences that disrupt the normal distribution of the data. F1 individuals (orange squares) show peaks too but of smaller height.

differences are proportional to genome size differences between h28 and the rest of populations, suggesting satellite DNA expansions are mainly responsible for the increase in genome size of h28 individuals. As also mentioned in Chapter 5, differences found on satDNA between samples can be a product of sequencing bias due to, for example, different library preparation methods. This would include random differences on satDNA amounts identified by RepeatExplorer2 and must be acknowledged and, if possible, library replicates should be included in the experimental approach. However, in this case the AT-rich regions seen in Figure 6.3 found only in larger genome accessions and segregating with the F1 individuals, supports our hypothesis that the satellite DNA enrichment observed in the analysis is not only due to artefact biases.

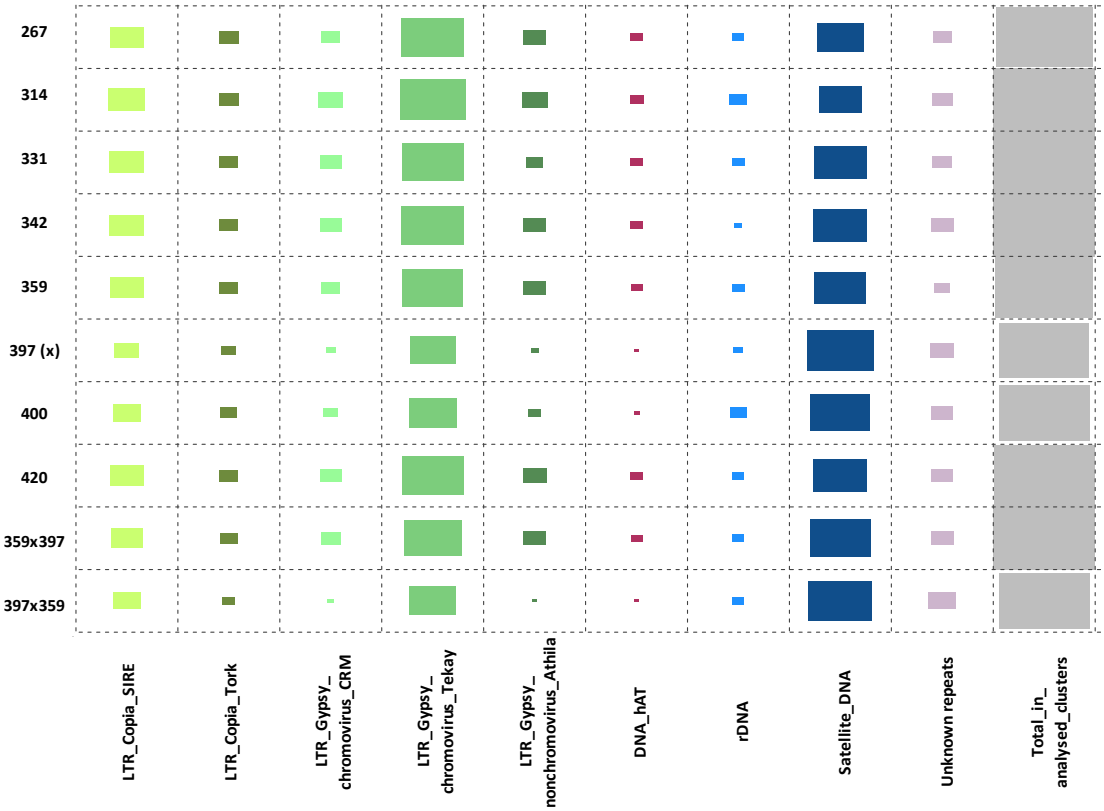


Figure 6.4: Distribution percentages of reads clustered and annotated as different types of repeats by RepeatExplorer2 in each accession. LTR lineages are indicated in different shades of green, DNA TEs in maroon and small repeats in blue. Only repeat groups and lineages represented by over 1% of the reads in at least one accession are shown.

The proportion of reads clustered as LTR elements is also variable across populations, but in this case the correlation with C-values is opposite: the more satellite DNA is found in an accession, the less LTR content found. This is likely due to the variable number of reads processed by RepeatExplorer2 for each accession caused by the variable

amounts of satDNA present in each sample, as also discussed in Chapter 5. This also means that it is likely transposons and other types of repeats will be misrepresented in the analysis of accessions with larger satDNA content, and explains the lower amount of LTR sequences found in h28 and F1 individuals. To further confirm that average genome size populations do not show LTR expansions, and elucidate whether LTR clusters expansions have also caused genome size differences between populations, we performed RepeatExplorer2 comparative analysis between h28 (397(x)) and h15 (359) individuals. In addition, to explore repeat variation within the h28 population, we performed another comparative analysis between the two h28 accessions 397(x) and 400.

Table 6.1: Clusters from RepeatExplorer2 comparative analysis showing expansions in 397 and 359. Cluster and supercluster number, size (number of reads) and annotations are specified. Only clusters showing ratios over 55:45% and under 45:55% are shown.

Cluster	**Super-cluster**	**Size_adjusted**	**Supercluster_best_hit**	**397:359**
11	3	42854	Satellite_DNA	90:10
34	3	12460	Satellite_DNA	68:32
68	7	750	LTR_*Gypsy*_chromovirus_Tekay	64:36
73	37	616	Satellite_DNA	62:38
31	7	2229	LTR_*Gypsy*_chromovirus_Tekay	60:40
5	5	6583	Satellite_DNA	58:42
39	7	1721	LTR_*Gypsy*_chromovirus_Tekay	57:43
35	3	1889	Satellite_DNA	56:44
87	49	455	LTR_*Gypsy*_non-chromovirus_Athila	40:60
92	53	357	Satellite_DNA	36:64
81	44	501	Unknown_repeats	31:69
93	54	336	Satellite_DNA	02:98

Data from the comparative analysis results was curated and sorted to find the clusters showing most differences between accessions samples in content. In most of the clusters a 50:50 ratio between populations was observed. Table 6.1 indicates all clusters (except those annotated as organelle) showing differences above 55:45 in their read content. Since RepeatExplorer2 comparative analyses require short codes of only three characters to distinguish between datasets, 397(x) will be addressed as 397 from now on. Our results show the largest clusters with major differences between populations are labelled as satellite DNA and LTR-Gypsy-Tekay. This indicates that not only satellite DNA expansions are responsible for the genome size differences found, but likely these

LTRs are also expanded in h28. Our results indicate that superclusters 3, 5, 7 and 37 represent the major expansions that explain C-value differences between these two populations, but none of them are unique for h28. Other small clusters (87, 92, 81, 93) are found to be more expanded in h15 (accession 359).

The two largest and most expanded clusters in h28 compared to h15 are satellite DNA repeats and they could represent a good proportion of the 100 Mbp difference these populations show in genome size. Graphic representations of these clusters, together with their tandem DNA consensus sequences are shown in Figure 6.5.

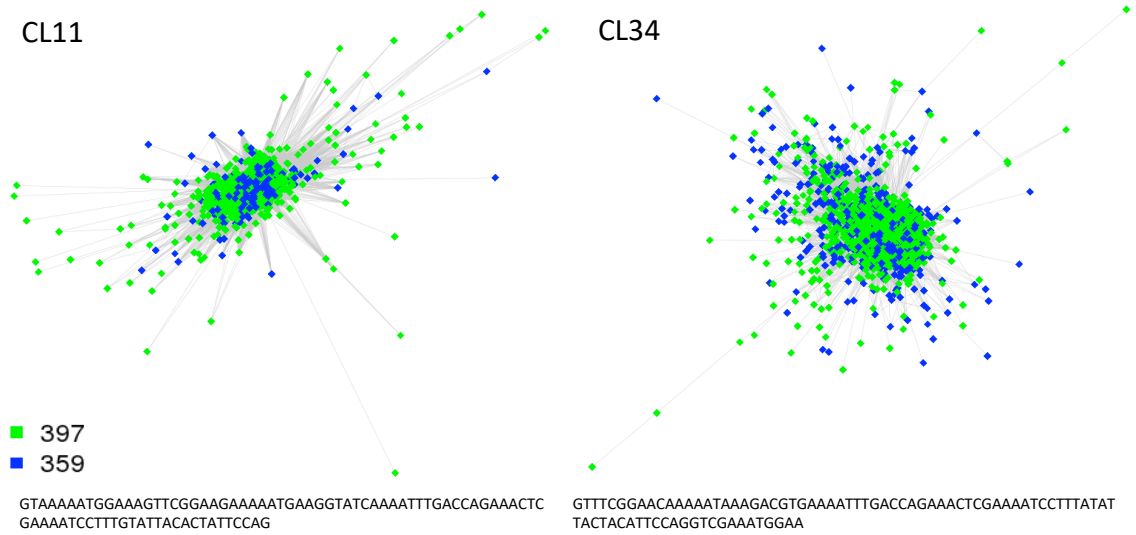


Figure 6.5: Graphic representation of the top two largest clusters showing major expansions in h28 (397) compared to h15 (359) population. Dots indicate each read and colours indicate which accession dataset they come from. Satellite consensus sequences from TAREAN (Novák et al., 2017) are shown below each cluster.

A summary of the results from the comparative analysis between the two h28 accessions 397 and 400 is shown in Table 6.2. This table shows that, again, the largest clusters with variable representation of the two samples are satellite DNA clusters. Although some differences are seen, the sample read ratios are smaller, representing differences of only 40:60 to 60:40. Nevertheless, this indicates that the "extra" satellite DNA in h28 is not fixed and varies even within the population.

Graphical representations of the three largest clusters shown to be slightly expanded in 397 compared to 400, together with their consensus sequences are shown in Figure 6.6. Consensus sequences shown in Figure 6.5 and Figure 6.6 were searched in the PlantSat

Table 6.2: Clusters from RepeatExplorer2 comparative analysis showing expansions in 397 and 400. Cluster and supercluster number, size (number of reads) and annotations are specified. Only clusters showing ratios over 55:45% and under 45:55% are shown.

Cluster	**Super-cluster**	**Size_adjusted**	**Supercluster_best_hit**	**397:400**
11	3	54803	Satellite_DNA	59:41
5	4	50024	Satellite_DNA	55:45
26	4	16279	Satellite_DNA	55:45
52	27	749	Satellite_DNA	55:45
75	47	314	LTR_*Gypsy*_chromovirus_CRM	55:45
95	65	154	Unknown_repeat	45:55
74	45	332	Unknown_repeat	43:57
96	66	140	LTR_*Gypsy*_chromovirus_Reina	43:57
45	5	1227	LTR_*Gypsy*_non-chromovirus_Athila	42:58
77	31	286	DNA_TIR_MuDR_Mutator	42:58
98	68	132	Unknown_repeat	42:58
88	59	169	Unknown_repeat	41:59
80	50	233	Unknown_repeat	40:60

database (Macas et al., 2002), but no sequences with significant similarities were found. Moreover, the consensus sequences of these clusters are not the same as the ones differentially expanded between populations h28 and h15. Further tandem repeat analysis is needed to find out whether these are different families of satDNA.

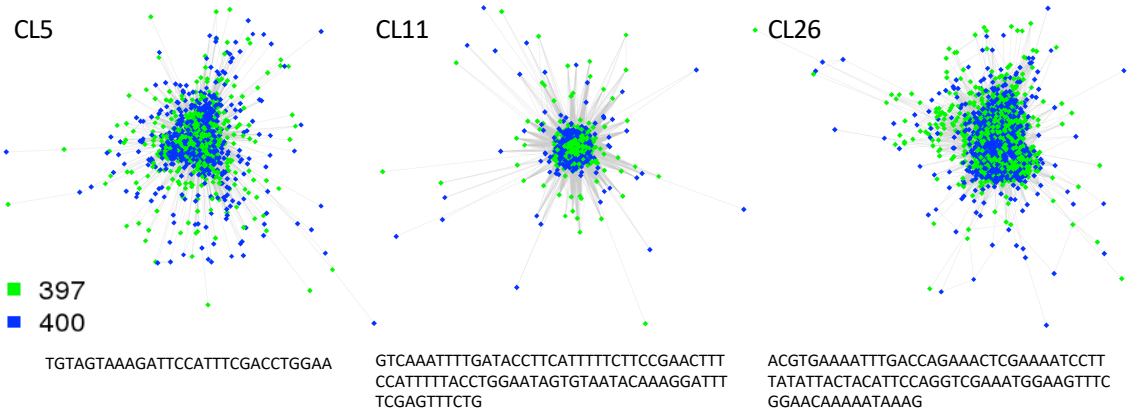


Figure 6.6: Graphic representation of the top three largest clusters showing major expansions the within h28 population, more specifically in accession 397 compared to 400. Dots indicate each read and colours indicate which accession dataset they come from. Satellite consensus sequences from TAREAN (Novák et al., 2017) are shown below each cluster.

Figure 6.2 shows a significant genome size increase between 397 and 397(x), and an intermediate C-value for accession 400. The 397/400 differences found in Tables 6.1 and 6.2 are due to satellite expansion, but this could be likely caused by the selfing

of 397. A better overview of inter-population variation could have been obtained by comparing the original 397 accession with 400, but all the plants from this accession died by the time our WGS approach was performed. Consequently, whether intra-population repeat expansions can be found in the wild cannot be inferred from this experiment.

Our results prove that the genome size differences between *B. heracleifolia* populations are mostly caused by the expansion of satellite DNA and LTR-Gypsy-Tekay elements. These expansions are retained in inter-population hybrids, which show intermediate values of satellite content between their two parental populations. Because the segregation of this repeat expansion does not look random in the hybrids but rather mendelian, it would seem that these repeats are intrachromosomal (Jones, 2018). However, the non-randomness of the segregation of these repeats cannot be inferred from two F1 individuals, and further analysis of F1 accessions would be needed to confirm this genome size variation is not caused by the presence of supernumerary, satellite or B chromosomes. Supernumerary B chromosomes have been previously described in 17 different *Begonia* species (D'Ambrosio et al., 2017), including *B. heracleifolia* (Legro and Doorenbos, 1969). These carry various types of repetitive DNA, not only satellite but also TEs (Tsujimoto and Niwa, 1992; Klemme et al., 2013; Houben et al., 2014), which suggests that random segregation of supernumerary B chromosomes in *B. heracleifolia* h28 populations could be the cause of the LTRs expansions we have observed in the comparative analysis of h28 compared to h15 populations. Additional genetic crosses and cytological work is needed to check for the presence of B chromosomes, but although *Begonia* chromosome count protocols have been previously published (Dewitte et al., 2009; Twyford, 2012), we attempted these with no results as we were not able to correctly perform metaphase chromosome fixation in *Begonia*. If a successful *Begonia* cytology protocol was established in the future, the location of these expanded satellites and LTRs could be observed using chromosome hybridisation techniques and designing probes using the consensus sequences from Figure 6.5 and Figure 6.6.

To determine whether the LTR expansions found in h28 are shared by other populations or whether these elements are currently active, RT-LTR temporal dynamics could be

obtained and compared for the LTR-*Gypsy*-Tekay elements of all populations, as done in Chapter 4 and Chapter 5. The chromosome location of these elements would also give us a better idea of whether these expansions are segregated independently to satellite DNA, or whether they go together and their expansions are correlated.

Satellite DNA expansions across populations have been previously found in other species. Shah et al. (2020) recently found remarkable satellite DNA expansions among populations of grasshoppers, and discussed that such expansions are not the cause of genome size expansions but rather the consequences of selection as means of stabilising the genome previously expanded by other elements. This theory relies on other genomic factors triggering the genome expansion, and in the case of *B. heracleifolia* these factors would be the LTR-*Gypsy*-Tekay lineages also expanded in h28. Another potential scenario would be that the satellite DNA expansion was triggered by transposition mechanisms. Until recently, there was no evidence of the functional and structural links between transposons and satellite DNA (Garrido-Ramos, 2017). However, in recent years these have been found (Meštrović et al., 2015), and research suggests that TEs can act as substrates for satellite DNA origin and mobility (Dias et al., 2014; McGurk and Barbash, 2018). These findings propose that the spread of short tandem repeats can be linked to transposition processes more often than previously thought (Satović et al., 2016), providing genome evolution with a whole new perspective that includes satellite DNA variation as another source of genome variation caused by transposition events.

6.2.3 Assessment of reproductive isolation in intraspecific crosses

To test the evolutionary impact of these expansions and their potential role on population divergence, 174 inter and intra-population crosses were done (table in appendix 6.A.2), and 65 capsules were sampled for seed viability counts. Figure 6.7 shows the proportion distribution of viable seeds in all analysed crosses, classified as Outcrosses of Distinct Genome Sizes (ODGS), Outcrosses of Similar Genome Sizes (OSGS), self-crosses (SELF) and crosses using the interpopulation F1 as parents (F2).

Results show a large variation in seed viability within each cross type. As expected

and previously observed (Twyford, 2012), self-crosses were found to be highly successful (mean = 0.69, SEM = 0.089, n = 7), as well as crosses between accessions with similar genome sizes (OSGS, mean = 0.66, SEM = 0.059, n = 15). On the other hand, crosses from different genome sizes (ODGS, mean = 0.25, SEM = 0.052, n = 15) and F2s (mean = 0.26, SEM = 0.050, n = 28) show much lower seed viability values.

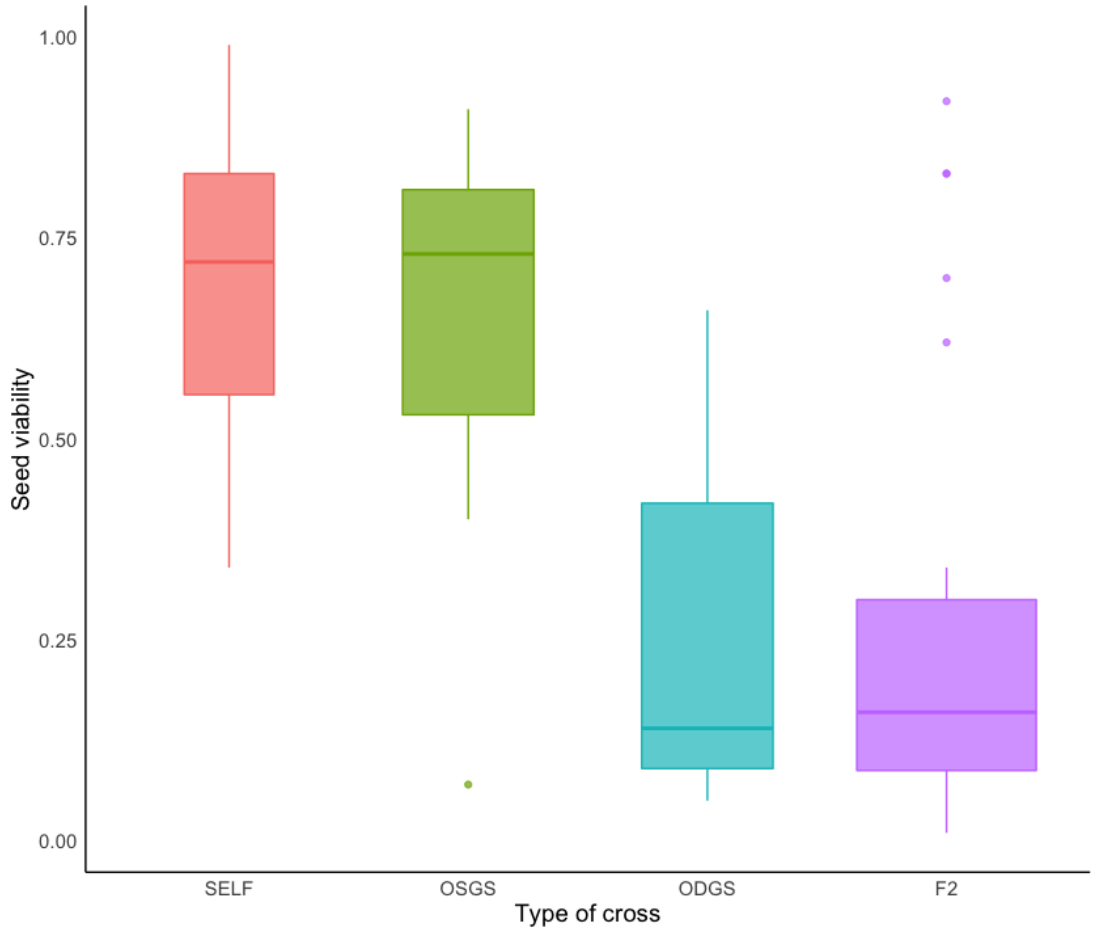


Figure 6.7: Boxplots representing the distribution of seed viability values (number of viable seeds in random samples of 100) in the 65 crosses analysed. Crosses are classified as Outcrosses between accessions of Distinct Genome Sizes (ODGS), Outcrosses between accessions of Similar Genome Sizes (OSGS), self-crosses (SELF) and crosses using the interpopulation F1 as parents (F2).

Non-parametric analysis of variance (Kruskal-Wallis test) revealed the data to be strongly divided into two significantly different groups (SELF and OSGS vs. ODGS and F2, $p < 0.0021$). However, it must be acknowledged that the statistical differences between these types of crosses involve non-independent comparisons, since the same individuals are being used multiple times for different crosses. A statistical test based on testing the differences between a matrix of crosses would help accounting this non-

independence of factors and should be considered for further testing. A table with all the different seed and seedlings counts, and cross details of each of the capsules is in appendix 6.A.4.

Full seeds reported as viable can still abort late in development (Twyford, 2012). To further investigate the crossing barriers between our populations, seed germination rates should also be assessed. We tested this only for 10 seeds per cross, and saw no parental effect on germination rate (boxplot in appendix 6.A.3). However a sample size of 10 seeds per cross, having each capsule hundreds of seeds, possibly has low statistical power. Moreover, only 65 out of the 174 crosses were accounted for seed viability, and variable sample sizes (capsules counted) are shown per type of cross. Larger numbers of seeds and crosses accounted are needed to provide our results with better statistical power.

It is important to keep in mind that in our seed viability and germination rate analyses we have assumed that genome size is confounded with genome-wide divergence, and therefore that the reduction in seed viability is due to the variation in genome size and not other factors (such as for example genetic variation due to geographical distance). Further seed viability tests of crosses between populations differentiated by other factors would help discarding these possibilities. The correlation between genetic isolation and geographical distance is commonly measured by a Mantel test (Diniz-Filho et al., 2013), and should be considered to further distinguish between all the factors that could be involved in the genetic distance found between these *B. heracleifolia* populations. Moreover, any additional ecological information on these populations would also help finding the critical factors that triggered this genetic isolation.

Twyford et al. (2014) described a reduced pollen fertility in h28 inter-population crosses. Our results support the reproductive barriers and population divergence described by (Twyford et al., 2014), and bring light to what underlies the genome size variation across populations. Rapidly changing satellite DNA can be responsible for centromeric changes that may lead to the reproductive isolation of emerging species (Henikoff et al., 2001). A concerted evolution of satellite DNA at the population level has previously been described in pupfish (Elder and Turner, 1994). By inducing heterochromatic and

centromere changes, rapid satellite DNA evolution has been previously described to trigger population differentiation in *Drosophila melanogaster* (Wei et al., 2014) and the beetle species *Tribolium castaneum* (Feliciello et al., 2015). Dynamic variation in satellite DNA content has previously been found at a genus level in other angiosperm genera. Emadzade et al. (2014) describes variation of the satellite repeat PaB6 among different *Prospero* species, and investigates its contribution to genome size and karyotype variation. However, we found no published evidence of interpopulation plant satellite DNA variation driving population divergence.

This chapter broadens our current understanding of the genomic and evolutionary processes behind population divergence in *Begonia*. (Twyford et al., 2014) found low species cohesion and limited gene flow between the populations studied, and we have added to these findings that this allopatric divergence is likely driven by genomic repeat dynamics and genomic incompatibility. This would be an example of non-adaptive radiation (Czekanski-Moir and Rundell, 2019), which has already been hypothesized for *Begonia* due to the low niche specificity found in this genus (Moonlight, 2017).

6.3 Materials and Methods

6.3.1 Plant Material

All the plant individuals analysed in this chapter were collected by Dr. Alex Twyford in his PhD fieldwork Mexico in 2010 and are now part of the living collection at RBGE. The population, collection and accession numbers of these are indicated in Table 6.3.

Table 6.3: Population number, RBGE accession id (last three digits used as accession number throughout this chapter), and collection id of all the *B. heraclaeifolia* specimens studied in this chapter.

Taxon name	Population	Accession id	Collection number
Begonia heracleifolia	h2	20100314	AT46
	h8	20100331	AT294
	h13	20100284	AT505
	h15	20100359	AT545
	h21	20100420	AT711
	h26	20100267	AT1019
	h28	20100397(x)	AT1079
	h28	20100400	AT1079

Accession 20100397 died prior to the WGS experiment and a self-cross from this plant (also shown to have a larger genome) was used instead, and therefore it is indicated with a (x) in Table 6.3. Moreover, we included in this study two more adult plants obtained from bidirectional interpopulation crosses between small and large genome populations, more specifically between the 20100359 and 20100397 accessions (addressed as 359x397 and 397x359). These crosses were performed and germinated by Dr. Catherine Kidner in 2013.

6.3.2 Genome size estimations by flow cytometry

Fresh leaf tissue of all the accessions sequenced in this chapter were shipped to the Jodrell Laboratories at the Royal Botanic Gardens Kew, where the nuclei isolation and flow cytometry protocol described in Chapter 2 was performed by Dr. Jaume Pellicer-Moscardó.

6.3.3 DNA extraction, library preparation and sequencing

Fresh young leaf tissue was harvested and ground in liquid nitrogen. The Qiagen DNeasy Plant Mini Kit (Qiagen, Germantown, MD, USA) was used for DNA isolation. For each sample, four extractions were done and combined at the elution step.

Samples were quantified using a Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and with a Nanovue (GE Healthcare, Chicago, IL, USA), and their integrity was assessed by agarose gel electrophoresis. After quantification and quality control of the elutions, some samples needed to be re-extracted to achieve the minimum DNA amount and concentration required by the sequencing facility. To achieve this, several DNA elutions were combined, and a DNA precipitation step was performed using sodium acetate and isopropanol (protocol from Birren et al. (1997)). Then DNA was diluted in the desired amount of TE buffer.

The sequencing facility (Novogene UK) used a NEBNext DNA Library Prep Kit (New England BioLabs, Ipswich, MA, USA) for 350 bp insert DNA library preparation, and their quality was assessed with a 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) and by quantitative PCR. A minimum of 1 Gb of 150 bp raw sequences was obtained

per sample with Illumina NovaSeq technologies. Quality control and adapter removal were also carried out by the sequencing facility (Novogene, UK).

6.3.4 Analysis of genomic repeats

For all samples, the number of trimmed and clean reads equivalent to 20% of the genome length (0.2x coverage) was randomly sampled for each dataset using seqtk (Li, 2012). Reads were then uploaded, pre-processed and clustered and annotated by similarity using the RepeatExplorer2 pipeline with the satellite filtering settings to be able to input a larger number of reads (Novak et al. (2013), https://repeatexplorer-elixir.cerit-sc.cz). For each of the accessions, cluster annotations were then reviewed and curated by hand, and all those clusters identified as plastid or mitochondrial DNA were removed prior to any downstream analysis. Comparative analysis runs were done between a small genome population (359) and the two individuals from the large genome one (397 and 400) by randomly sampling and pre-processing 300,000 reads of each sample. Comparative analysis results were also reviewed and curated by hand, and all those clusters identified as plastid or mitochondrial DNA were removed prior to any downstream analysis.

6.3.5 Intra and inter-population crosses of *Begonia heracleifolia*

Crosses between *B. heracleifolia* accessions were done as described in Twyford et al. (2014). Daily checks on *B. heracleifolia* plants were done from February to mid-April 2018 and 2019 to score recently opened female flowers. *Begonia* crosses were done by hand-pollinating young female flowers with hand-picked male flowers. Pollinated flowers were labelled and covered to avoid any posterior pollination. Mature seed capsules were harvested when dry.

6.3.6 Seed viability tests

After harvesting the seed capsules, their viability rates were determined by scoring the ratio of viable seeds in random samples of 100 seeds per capsule. For the analysis of seed viability results, crosses were divided into self-crosses (SELF), outcrosses between accessions of similar genome sizes (OSGS), outcrosses between accessions of distinct

genome sizes (ODGS), and crosses done using F1 populations as parents (F2). Seed germination rates were also assessed for a subset of 10 full seeds from each capsule. For this, seeds were germinated in substrate sterilised with boiling water, and sealed in ziploc bags for 14 days. Seedlings were counted on day 14.

Bibliography

Andrews S, et al. (2 co-authors). 2010. Fastqc: a quality control tool for high throughput sequence data.

Birren B, Green ED, Klapholz S. 1997. Genome analysis: a laboratory manual, volume 4. CSHL Press.

Czekanski-Moir JE, Rundell RJ. 2019. The ecology of nonecological speciation and nonadaptive radiations. *Trends in ecology & evolution*. 34:400–415.

D'Ambrosio U, Alonso-Lifante MP, Barros Ferradás K, Kovarik A, Mas de Xaxars G, Garcia S. 2017. B-chrom: A database on b-chromosomes of plants, animals and fungi. *New Phytologist*. 216:635–642.

Dewitte A, Leus L, Eeckhaut T, Vanstechelman I, Van Huylenbroeck J, Van Bockstaele E. 2009. Genome size variation in begonia. *Genome*. 52:829–838.

Dias GB, Svartman M, Delprat A, Ruiz A, Kuhn GC. 2014. Tetris is a foldback transposon that provided the building blocks for an emerging satellite dna of drosophila virilis. *Genome biology and evolution*. 6:1302–1313.

Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, Telles MPdC, Rangel TF, Bini LM. 2013. Mantel test in population genetics. *Genetics and molecular biology*. 36:475–485.

Elder JF, Turner BJ. 1994. Concerted evolution at the population level: pupfish hindiii satellite dna sequences. *Proceedings of the National Academy of Sciences*. 91:994–998.

Emadzade K, Jang TS, Macas J, Kovařík A, Novák P, Parker J, Weiss-Schneeweiss H. 2014. Differential amplification of satellite pab6 in chromosomally hypervariable prospero autumnale complex (hyacinthaceae). *Annals of botany*. 114:1597–1608.

Feliciello I, Akrap I, Brajković J, Zlatar I, Ugarković D. 2015. Satellite dna as a driver of population divergence in the red flour beetle tribolium castaneum. *Genome biology and evolution*. 7:228–239.

Garrido-Ramos MA. 2015. Satellite dna in plants: more than just rubbish. *Cytogenetic and Genome Research*. 146:153–170.

Garrido-Ramos MA. 2017. Satellite dna: an evolving topic. *Genes*. 8:230.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving dna. *Science*. 293:1098–1102.

Houben A, Banaei-Moghaddam AM, Klemme S, Timmis JN. 2014. Evolution and biology of supernumerary b chromosomes. *Cellular and Molecular Life Sciences*. 71:467–478.

Jones R. 2018. Transmission and drive involving parasitic b chromosomes. *Genes*. 9:388.

Klemme S, Banaei-Moghaddam AM, Macas J, Wicker T, Novák P, Houben A. 2013. High-copy sequences reveal distinct evolution of the rye b chromosome. *New Phytologist*. 199:550–558.

Legro R, Doorenbos J. 1969. Chromosome numbers in begonia. *Netherlands Journal of Agricultural Science*. 17:189–202.

Li H. 2012. seqtk. *Heng Li, GitHub repository. https://github. com/lh3/seqtk*. .

Macas J, Meszaros T, Nouzova M. 2002. Plantsat: a specialized database for plant satellite repeats. *Bioinformatics*. 18:28–35.

McGurk MP, Barbash DA. 2018. Double insertion of transposable elements provides a substrate for the evolution of satellite dna. *Genome research*. 28:714–725.

Meštrović N, Mravinac B, Pavlek M, Vojvoda-Zeljko T, Šatović E, Plohl M. 2015. Structural and functional liaisons between transposable elements and satellite dnas. *Chromosome Research*. 23:583–596.

Moonlight PW. 2017. Linking speciation and the niche: taxonomy, phylogeny and niche evolution in neotropical Begonia. Ph.D. thesis, University of Glasgow.

Novák P, Ávila Robledillo L, Koblížková A, Vrbová I, Neumann P, Macas J. 2017. Tarean: a computational tool for identification and characterization of satellite dna from unassembled short reads. *Nucleic acids research*. 45:e111–e111.

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. Repeatexplorer: a galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*. 29:792–793.

Pinheiro F, Cozzolino S, de Barros F, Gouveia TM, Suzuki RM, Fay MF, Palma-Silva C. 2013. Phylogeographic structure and outbreeding depression reveal early stages of reproductive isolation in the neotropical orchid epidendrum denticulatum. *Evolution*. 67:2024–2039.

Satović E, Zeljko TV, Luchetti A, Mantovani B, Plohl M. 2016. Adjacent sequences disclose potential for intra-genomic dispersal of satellite dna repeats and suggest a complex network with transposable elements. *BMC genomics*. 17:997.

Shah A, Hoffman JI, Schielzeth H. 2020. Comparative analysis of genomic repeat content in gomphocerine grasshoppers reveals expansion of satellite dna and helitrons in species with unusually large genomes. *Genome biology and evolution*. 12:1180–1193.

Tsujimoto H, Niwa K. 1992. Dna structure of the b chromosome of rye revealed by in situ hybridization using repetitive sequences. *The Japanese Journal of Genetics*. 67:233–241.

Twyford AD. 2012. Speciation and gene flow in Central American Begonia L.(Begoniaceae). Ph.D. thesis, The University of Edinburgh.

Twyford AD, Kidner CA, Ennos RA. 2014. Genetic differentiation and species cohesion in two widespread central american begonia species. *Heredity*. 112:382–390.

Via S. 1999. Reproductive isolation between sympatric races of pea aphids. i. gene flow restriction and habitat choice. *Evolution*. 53:1446–1457.

Wei KHC, Grenier JK, Barbash DA, Clark AG. 2014. Correlated variation and population differentiation in satellite dna abundance among lines of drosophila melanogaster. *Proceedings of the National Academy of Sciences*. 111:18793–18798.

6.A Appendix

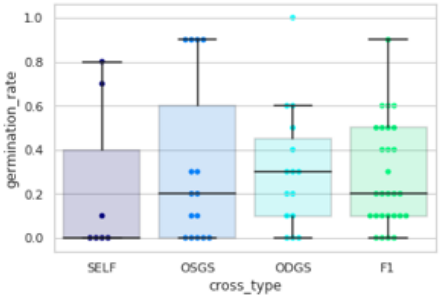
6.A.1 RepeatExplorer2 summary data table

	267	314	331	342	359	397	400	420	359x 397	397x 359
LTR_*Copia*_SIRE	6.72	7.83	7.01	6.80	6.61	3.49	4.59	6.52	5.660	4.334
LTR_*Copia*_Tork	2.30	2.29	2.06	2.01	2.06	1.15	1.69	2.03	1.769	0.951
LTR_*Gypsy*_chrom_CRM	2.09	3.65	2.70	2.66	2.10	0.49	1.16	2.60	2.282	0.213
LTR_*Gypsy*_chrom_Tekay	22.56	25.23	22.28	23.07	21.66	11.94	13.50	22.44	19.178	12.657
LTR_*Gypsy*_non-chrom_Athila	3.07	3.81	1.63	2.85	2.95	0.30	0.85	3.22	2.862	0.139
DNAelement_TIR_hAT	0.85	1.14	0.93	0.89	0.74	0.14	0.21	0.85	0.753	0.112
rDNA	0.82	1.76	0.87	0.30	0.94	0.52	1.56	0.80	0.77	0.78
Satellite_repeat	12.55	10.61	16.33	16.51	15.76	25.74	21.04	16.66	21.692	23.678
Unknown_repeat	2.06	2.38	2.16	2.84	1.48	3.22	2.69	2.61	2.845	4.476
Total_in_clusters	54.66	60.93	59.87	60.16	56.05	46.63	47.79	59.53	60.063	47.748

6.A.2 Table of crosses

	267	314	331	342	359	397(x)	400	420	359x397	397x359
267	2	2	2	2	2	2	1	2	2	2
314	2	2	2	1	2	2	-	-	2	2
331	2	2	2	2	2	2	-	2	2	2
342	2	2	2	2	2	2	2	2	2	2
359	2	-	-	1	2	2	-	-	2	2
397(x)	2	2	2	2	2	3	1	2	3	3
400	-	-	-	1	-	2	1	-	-	-
420	2	2	2	2	2	3	-	3	2	2
359x397	2	2	2	2	2	3	1	2	3	3
397x359	2	2	2	2	2	3	2	2	3	3

6.A.3 Germination rate distributions



6.A.4 Seed count and germination rate data

capsule no	cross ID	female parent	male parent	cross type	full seeds	shrivelled seeds	total seeds counted	seed viability	germination rate
1	TK29	267	267	SELF	90	10	100	0.90	0
2	TK43	267	267	SELF	34	66	100	0.34	0
3	TK105	267	342	OSGS	64	36	100	0.64	0
4	TK106	267	342	OSGS	90	10	100	0.90	0.1
5	TK101	267	359	OSGS	79	21	100	0.79	0
6	TK102	267	359	OSGS	57	43	100	0.57	0.1
7	TK44	267	397x	ODGS	46	54	100	0.46	0.1
8	TK30	267	397x	ODGS	14	86	100	0.14	0
9	TK31	267	400	ODGS	54	46	100	0.54	0.3
10	TK68	267	359x397	F1	5	95	100	0.05	0.1
11	TK69	267	359x397	F1	10	90	100	0.10	0
12	TK66	267	397x359	F1	10	90	100	0.10	0
13	TK67	267	397x359	F1	62	38	100	0.62	0.2
14	TK33	342	267	OSGS	49	51	100	0.49	0.9
15	TK34	342	267	OSGS	49	51	100	0.49	0.3
16	TK48	342	331	OSGS	80	20	100	0.80	0.2
17	TK49	342	331	OSGS	73	27	100	0.73	0
18	TK32	342	342	SELF	76	24	100	0.76	0.1
19	TK1	342	342	SELF	42	58	100	0.42	0
20	TK50	342	359	OSGS	82	18	100	0.82	0.3
21	TK51	342	359	OSGS	91	9	100	0.91	0.9
22	TK4	342	397x	ODGS	38	62	100	0.38	0.3
23	TK2	342	400	ODGS	46	54	100	0.46	0.6
24	TK7	342	397x359	F1	83	17	100	0.83	0.6
25	TK120	359	267	OSGS	68	32	100	0.68	0.2
26	TK169	359	342	OSGS	75	25	100	0.75	0
27	TK119	359	359	SELF	72	28	100	0.72	0
28	TK108	359	397x	ODGS	14	86	100	0.14	0.2
29	TK109	359	397x	ODGS	66	34	100	0.66	0
30	TK117	359	359x397	F1	83	17	100	0.83	0.4
31	TK116	359	359x397	F1	92	8	100	0.92	0.5
32	TK115	359	397x359	F1	70	30	100	0.70	0
33	TK18	397x	267	ODGS	6	94	100	0.06	0
34	TK19	397x	267	ODGS	13	87	100	0.13	0.1
35	TK22	397x	331	ODGS	5	95	100	0.05	0.4
36	TK23	397x	331	ODGS	8	92	100	0.08	0.3
37	TK24	397x	342	ODGS	22	78	100	0.22	0.6
38	TK25	397x	342	ODGS	5	95	100	0.05	0.2
39	TK26	397x	359	ODGS	10	90	100	0.10	0.5
40	TK11	397x	397x	SELF	69	31	100	0.69	0.7
41	TK12	397x	400	OSGS	7	93	100	0.07	0
42	TK16	397x	359x397	F1	34	66	100	0.34	0.5
43	LCD22	397x	359x397	F1	9	91	100	0.09	0.2
44	LCD21	397x	397x359	F1	2	98	100	0.02	0
45	TK15	397x	397x359	F1	17	83	100	0.17	0.4
46	TK61	400	342	ODGS	32	68	100	0.32	1
47	TK28	400	397x	OSGS	89	11	100	0.89	0.9
48	TK59	400	397x	OSGS	40	60	100	0.40	0.9
49	TK60	400	400	SELF	99	1	100	0.99	0.8
50	TK77	359x397	267	F1	21	79	100	0.21	0.1
51	TK83	359x397	342	F1	30	70	100	0.30	0.5
52	TK82	359x397	342	F1	14	86	100	0.14	0.4
53	TK85	359x397	359	F1	8	92	100	0.08	0.2
54	TK84	359x397	359	F1	10	90	100	0.10	0.1
55	TK55	359x397	397x	F1	13	87	100	0.13	0.1
56	LCD24	359x397	397x	F1	23	77	100	0.23	0.2
57	TK58	359x397	400	F1	15	85	100	0.15	0.1
58	LCD23	359x397	359x397	F1	30	70	100	0.30	0.3
59	TK56	359x397	397x359	F1	7	93	100	0.07	0.1
60	TK71	397x359	267	F1	24	76	100	0.24	0.5
61	TK70	397x359	267	F1	1	99	100	0.01	0.1
62	TK74	397x359	342	F1	27	73	100	0.27	0.9
63	LCD27	397x359	397x	F1	17	83	100	0.17	0.6
64	TK35	397x359	397x	F1	8	92	100	0.08	0.2
65	TK38	397x359	397x359	F1	6	94	100	0.06	0.6

Chapter 7

General conclusions

7.1 General conclusions

The importance of genomic factors as drivers of angiosperm speciation have become clearer as genomic data has accumulated in the recent decades (Leitch and Leitch, 2012; Kulmuni et al., 2020). This thesis uses the mega-diverse genus *Begonia* to search for evidence of association between changes in size, diversity and structure of the genomic repeat landscape and bursts of species radiations through the evolution of the group.

In Chapter 2, I generated *Begonia* chromosome number and genome size datasets in order to study the distribution of this data across the *Begonia* phylogeny. Results indicate that on a large scale chromosome number variation correlates with clade size but there is no simple correlation between species number and karyotypic changes in *Begonia*. I also found remarkable genome size differences between closely-related species with the same chromosome numbers, which suggests an effect of repeat dynamics in *Begonia* chromosome size.

In Chapter 3, I described the sequencing and assembly of seven Begoniaceae genomes from all main clades. K-mer analysis revealed high heterozygosity rates, so to obtain the best assemblies I took several assembly approaches to test as many pipelines as possible. Our assembly results reflect a wide variation in *Begonia* genome complexity. I also performed gene annotation of all available genome assemblies of *Begonia* to date. This chapter addresses the challenges of assembling complex genomes of non-model organisms, and describes the genome assembly data that was used in Chapter 4.

In Chapter 4, I used Chapter 3 genomic data and different approaches to describe and compare the repeat landscapes and LTR dynamics of twelve Begoniaceae genomes, with the aim of learning whether genomes from larger and more recent radiations are more dynamic. Results from this chapter suggest larger LTR or satDNA content in most species from large *Begonia* radiations, compared to their depauperons. I also identified peaks of transposon activity in recent radiations in both SE Asian and in Neotropical species.

To take a closer look at repeat dynamics in recent radiations, in Chapter 5 I took a genome skimming approach in one of the most studied and species-rich Neotropical

Begonia sections (*B.* sect. *Gireoudia*).The analysis of genomic data from 34 species revealed that species from species-rich sections have more dynamic repeatomes than their species-poor relatives. In this chapter I also identified section-specific LTR repeat changes that date back to the radiation of section *Gireoudia* (Chapter 5).

In Chapter 6, we also used a genome skimming approach to examine the question at the population level. Results from this chapter show that rapidly changing satellite DNA can be responsible for changes in genome structure that may lead to the reproductive isolation populations in *Begonia heracleifolia*.

7.2 Whole Genome Sequencing projects for non-model species and the challenges of analysing repetitive DNA

The keystone of the work underlying this thesis was obtaining *Begonia* genome resources to test our main hypothesis and to understand the genomic changes occurring during the *Begonia* radiations. In Chapter 3, I presented the whole genome sequencing and assembly of seven Begoniaceae species, and discuss the problems encountered when sequencing genomes of variable complexity including different levels of repeat content and heterozygosity, as well as variable genome sizes and therefore variable sequence coverage.

High-quality genome sequencing and chromosome-level assemblies of non-model plant species have become more common recently (e.g. Xing et al. (2019); Barchi et al. (2019); Zhang et al. (2020b,a); Yi et al. (2020); Chen et al. (2020); Wu et al. (2020)). The comparison of all these genomes is playing an important role in enhancing our understanding of the biology and evolution of plants (Kelly, 2019). However, the bioinformatic challenge of assembling and annotating this data is immense. Plants present highly complex genomes with a high and variable repeat content, making assembly from short read data exceptionally challenging. Assembly of repeat regions using short read data is difficult and error-prone. For this reason, in Chapter 4 I included a repeatome annotation approach using unassembled data.

Naciri and Linder (2020) suggest the need for a large number of high-quality genomes

available for a large number of non-model organisms to better understand the role of TE activity in evolutionary radiations and speciation processes. Luckily, during the time period of this PhD, new and more affordable long-read technologies have become more widely available enabling better quality assemblies and consequently better analysis of repetitive regions (Shahid and Slotkin, 2020; Michael and VanBuren, 2020; Dumschott et al., 2020; Nguyen et al., 2020). In this study, we assembled and annotated seven genomes, and used five more assemblies from collaborators and published data (Griesmann et al. (2018) and Li et al, unpublished). Out of the total of twelve genomes analysed, only three were sequenced using long read technologies (*B. conchifolia* from this project, and *B. masoniana* and *B. darthvaderiana* from Li et al., unpublished).

This approach of comparing long and short read genome assemblies has been used for some evolutionary studies (Shinzato et al., 2021), but these do not focus on the analysis and comparison of the repeatome. According to Wierzbicki et al. (2020), genomic analyses of TEs is only possible with high quality genomes from highly homozygous individuals, sequenced using long reads, assembled using specific approaches, polished with short read data and scaffolded with long reads or Hi-C. For *Begonia*, the creation of homozygous lines would require at least 6 years (minimum 9 month generation time, at least 8 generations of selfing required). However, all the other guidelines were followed in the assembly of *B. conchifolia*, and most of them for the genomes from Li et al (unpublished).

Despite this, it is important to keep in mind that our genome data represents the result of a four-year project of limited budget. The genomes presented in this thesis were of good enough quality to obtain transposon information, the limitations of their quality were taken into account and have been discussed throughout this thesis, and in general, comparisons were mostly made across assemblies from data of the same sequencing technology. It is likely that in the next few years long read technologies will be affordable enough to be able to add several related species to this analysis (Gatter et al., 2020), and the improvement of already available short-read assembled genomes is also possible, as is already being done for some species (Osipowski et al., 2020). With improved genomes the work presented here could be revised and some sources of error

removed.

7.3 Chromosome evolution in *Begonia*: karyotype and genome size instability at different taxonomic levels

In Chapter 2, a large chromosome number dataset for *Begonia* representing karyotypic information of more than 400 species was generated and data was taxonomically classified using the latest phylogeny. This, however, still represents only 1/5 of the species in the genus. Such genus-wide cytogenetic studies are rare, but they tend to include a larger portion of the genus species, as in Leong-Škorničková et al. (2007) for *Curcuma* or Moeglein et al. (2020) for *Vibrunum*, where around 50% of the genus species are karyotyped. However in larger genera, karyological studies are normally done in specific clades within the genus. For example, Molero et al. (2002) studies chromosome number dynamics across one single *Euphorbia* subsection. In other cases, specific geographical ranges of a genus are studied, as done by Choi et al. (2020) in Korean *Iris*. At a larger phylogenetic scale, a recent study looked into the chromosome evolution of one of the largest Leguminoseae clades that includes 45 genera (Moraes et al., 2020). Nonetheless, Chapter 2 represents one of the few genus-wide study on karyotype evolution in one of the largest angiosperm genera.

Before this study, previous research had already established that *Begonia* had no clear basic chromosome number (Dewitte, 2010). Chapter 2 dataset allowed an overview of the karyotype instability at the clade and section levels, and suggests wide chromosome variation within the main clades (specially in the large radiations), but generally higher stability within sections. This supports the rather natural sectional classification suggested by Moonlight et al. (2018), and contrasts with Molero et al. (2002) results, who found at least three basic numbers and two ploidy levels in a single *Euphorbia* subsection. One of the problems of dealing with a large genus has been not being able to map our data on a species-level tree, as for example done in Moeglein et al. (2020) for ancestral state reconstruction. The density of data and the incongruence between the species present in the Moonlight et al. (2018) tree and in our dataset impeded this. Nonetheless, I suggest our datasets could be used to map chromosome numbers

at species level in section-wide or clade-wide *Begonia* phylogenies.

Although many sections had relatively stable chromosome numbers, our genome size data reveals up to almost 10-fold differences in genome size between species of the same section. Genome size differences among species with the same chromosome number generally suggest recent expansions in LTR or other repeats, as has been reported in other taxa (Ma and Bennetzen, 2004; Wong et al., 2019; Baniaga and Barker, 2019; Becking et al., 2020; Petersen, 2020). This was confirmed for *Begonia* sect. *Petermannia* when comparing the *B. bipinnatifida* and *B. darthvaderiana* genomes in Chapter 4 and for section *Gireoudia* in Chapter 5. In other sections where genome size variation was found, we cannot confidently link this to repeatome dynamics and other factors may underlie the change. For example in the case of the genus *Anacyclus* (Asteraceae), recent research reports genome size differences between species with equal chromosome numbers independent of repetitive DNA dynamics, and rather caused by losses or gain of chromosome fragments associated with homoploid hybridisation (Vitales et al., 2020). Our genome size dataset is too small to report genome size variation within all *Begonia* sections, as we have data only for some of the larger sections. Therefore, although we suggest repeat dynamics underlie genome size variation in at least two of the largest *Begonia* sections we cannot say whether this is a general pattern across the genus, and more genome size data would be needed to confirm this.

Finally, the correlation analysis between chromosome number and genome sizes presented in Chapter 2 needs to be tested for phylogenetic signal, and Phylogenetic Generalized Least Squares (PGLS) regression analyses are needed, as were done before for cycad cytological data in Gorelick et al. (2014). This would add robustness to our conclusions and confirm whether genome size dynamics varies across the *Begonia* phylogeny.

7.4 *Begonia* genomes are more dynamic in large radiations

In Chapter 4, I performed a complete repeatome analysis in a number of *Begonia* species from independent radiations using two different approaches, to obtain the best

(not assembly biased) overview possible of the repeat landscape of a phylogenetically balanced selection of species. Results indicate LTR content correlated with genome size, whereas number of full-length LTRs identified correlated with genome assembly contiguity and were strongly influenced by the use of long-read sequencing technologies.

Our genomic analysis on *Hillebrandia sandwicensis* indicate this species has a smaller, less repetitive and more stable genome (or with less recent LTR activity) than most *Begonia* species analysed. Given that the plant sequenced was wild collected, the high homozygosity suggests small and inbred populations. *Hillebrandia* was categorized as in peril of extinction in 1999 (according to NatureServe), supporting the conclusions that it is suffering from severe inbreeding depression, which can imply an evolutionary dead end (Lande Schemske, 1985). According to Forrest (2001), *Hillebrandia* is likely to have been on the Hawaiian archipelago for at least 55 million years. However, Moonlight et al. (2018) estimated a much younger date for the split between *Hillebrandia* and *Begonia* (∼25 MYA). This species is currently only found on the older islands of the archipelago, but there is not much known on its population structure. If, as suggested by Cronk (1989), taxa go through 'bloom and bust' cycles, *Hillebrandia* may be the last dying member of what once was a very species-diverse plant group (Hughes, 2002). Further genetic analyses and additional population structure studies are needed to confirm the levels of genetic isolation and inbreeding that the *Begonia* sister genus undergoes in the wild, and would help update its conservation status.

Within *Begonia*, I found different genome dynamic patterns in each of the radiations studied, but with one similar characteristic: species from large radiations generally show an expansion of at least one repeat family compared with their close relatives. From the first Neotropical radiation (NC1), I sequenced and analysed the genome of *Begonia luxurians*, from the large section *Pritzelia*. Species from this section, including *B. luxurians*, show a large number of chromosomes (2n=56), and the smallest genome sizes found so far in *Begonia* (245 to 345 Mbp), some even smaller than *Hillebrandia*. The high chromosome numbers in all sections from this clade suggest a polyploidisation event early in the diversification of this group. However, the small genome sizes indicate extensive genome downsizing after polyploidisation. Leitch and Bennett (2004)

describe that this phenomenon has been observed in many polyploid species, and suggests this could be caused by (i) homoeologous recombination leading to DNA loss or (ii) elimination of specific DNA sequences (selective gene loss or extensive elimination of low-copy, non-coding DNA). However, although polyploidisation events have been suggested for other *Begonia* clades (NC2-ii), other sections do not show the levels of genome downsizing seen in *Pritzelia*. This possibly suggests either very early genome downsizing event(s) right after polyploidisation in the main common ancestor of this section, or some selective pressure for small genomes for these species. Obtaining more C-value data and further analysis on *Pritzelia* species would provide useful information for understanding genome size evolution in this section. The genome size variation observed among *Pritzelia* species, our repeatome annotation and LTR-dating results suggest that although their LTR content is low, these species show some genome dynamism compared to their closest African relative (*B. johnstonii*). Moreover, the satDNA expansion observed in *B. luxurians* suggests these differences could also be due to satDNA expansions, as also observed in another Neotropical section in Chapters 5 and 6. More genomic data on species from this radiation is needed to answer these questions.

Regarding the second Neotropical radiation (NC2), I analysed the complete genomes of two species: *B. conchifolia* (NC2-i) and *B. foliosa* (var. *miniata* (NC-ii)). Our LTR-dating data in Chapter 4 also confirms more recent LTR activations and specific peaks of LTR-RT activity in these two species compared to their closest African relative (*B. dregei*). The case of *B. foliosa* is very similar to that of *B. luxurians*. Chromosome data from this clade suggest a polyploidisation event, and although genome sizes are not as small in NC2-ii as in NC1, they are smaller than those in sister clades (and have larger chromosome numbers). This, together with the fact that *B. foliosa* also shows a higher amount of satDNA compared to other *Begonia* species, makes us think that species from the NC1 and NC2-ii radiations have undergone similar genome dynamics. *B. conchifolia* from the NC2-i radiation, on the other hand, does show higher LTR content and a slightly larger genome (the largest of the three Neotropical genomes analysed), and its genome dynamism could be triggered by high LTR activity. However, later results from Chapters 5 and 6 will actually suggest that satDNA dynamics are also highly common in this radiation.

Finally, in Chapter 4 I analysed the genomes of four species from the Asian radiation, *B. masoniana* (*B.* sect. *Coelocentrum*), *B. bipinnatifida* and *B. darthvaderiana* (*B.* sect. *Petermannia*), and *B. peltatifolia*, unassigned to a section. Both sections *Coelocentrum* and *Petermannia* are large Asian sections representing two of the main radiations within this clade, and Chapter 2 results shows species in these sections have relatively stable karyotypes and variable genome sizes, suggesting they have dynamic genomes. Chapter 4 suggests high content and recent activity of LTR elements in *B. masoniana*, *B. darthvaderiana* and *B. bipinnatifida*, in contrast with the repeat landscape of *B. peltatifolia*, which also shows a considerably smaller genome. Although there is not much population data published on this species, *B. peltatifolia* is mostly found in small, isolated populations in small ranges of distribution (Xiaobai (1993) and M. Hughes, pers. observ.). Moreover, this species is categorised as endangered according to the IUCN (International Union for the Conservation of Nature, 2017; Tian et al. (2018)). It could be possible that even within large *Begonia* sections, we have encountered some similar cases to the *H. sandwicensis* of highly isolated species with a stable genome in an "evolutionary dead end". This indicates that whatever the genomic factor that triggers genome dynamics in large *Begonia* radiations is, it is not present in all species from these radiations.

To summarise, the in-depth repeatome analysis done on all *Begonia* radiations by comparing them to their African closest relatives suggests that species from Neotropical radiations have smaller genomes, with variable expansions of both LTRs and satDNA, whereas species from large Asian radiations have larger genomes with genome dynamics driven mostly by the high activity of LTR elements.

These findings suggest a link between large radiations and higher rates of repetitive DNA gain/loss, but it is not easy to find the specific factors directly triggering to specific speciation or diversification events when only looking into a few genomes from radiations of hundreds of species. Previous research has linked diversification and TE activity in other taxa, such as for example African cichlid fishes (Carleton et al., 2020), *Hydra* species (Wong et al., 2019) and recently cultivated olive tree varieties (Jiménez-Ruiz et al., 2020). However, the approaches taken were different in each case. Carleton

et al. (2020) found that the phenotypical variation that led to diversification in cichlid fishes was triggered by indels caused by TE insertions using QTL mapping approaches. This would be possible in *Begonia*, since a genetic map has been developed and morphological and physiological leaf traits are being mapped in the *Begonia conchifolia* genome (Ali (2013); Fan and Kidner, unpublished). However, since speciation processes in *Gireoudia* were found not to be adaptive (Moonlight, 2017), any novel traits caused by TE insertions are unlikely to explain speciation patterns. In the *Hydra* study, they found one transposon expansion that is specific for all species in the main brown hydra clade (Wong et al., 2019) by comparing the genomes of all species in the genus. To do this in *Begonia*, too many species would need to be sequenced. Finally, in the olive tree study, the diversification ages are very recent, making it easier to date all transposon expansion and link them to the development of specific varieties. However what is common to all these studies is that they analysed and compared a number of repeatomes from closely related species and populations, which leads to our work in Chapter 5 and Chapter 6.

Moreover, even though non-ecological speciation has been suggested for some *Begonia* groups (Moonlight, 2017), this genus has undergone several evolutionary radiations, and more in-depth evolutionary studies are needed in each of them to fully understand what role do all the different ecological and genomic factors play in these radiations. Naciri and Linder (2020) recently reviewed the different scenarios and factors that lead to different adaptive radiations, including the arrival to a new adaptive zone, geographical and ecological expansion, epigenetic changes, genetic drift, gene flow limitation, intraspecific processes, admixture or hybridisation, aside from the activation and dynamics of repetitive DNA. Evolutionary radiations are complex processes and although TEs represent an important source of genomic variation, it is essential to think about the evolution of *Begonia* in a broad context in which potentially, numerous evolutionary scenarios could have triggered the many species radiations represented by its largest sections.

7.5 Potential evolutionary consequences of LTR and satDNA expansions in *Begonia*

In Chapter 5, I provide a better overview of genome dynamics at a closer phylogenetic level. Results from this chapter indicate that the genome size dynamics found in *Gireoudia* species is caused by different LTR-*Gypsy*, LTR-*Copia* and satDNA expansions. These results contrast with the repeat landscape of the smaller closely related sections, which have less repetitive genomes with smaller satDNA content. Additionally, I found *Gireoudia*-specific LTR-*Gypsy*-Tekay expansion dated at a similar time when the *Gireoudia* section diverged. This provides stronger evidence of genome dynamism driving species radiation in larger *Begonia* sections. Moreover, results from Chapter 6 indicate that inter-population genome size differences are driven by satDNA and LTR-*Gypsy*-Tekay dynamics, and that interpopulation crosses between larger and smaller genome size individuals yield less viable seeds, indicating a potential role in reproductive isolation and speciation.

A study on the centromeric satDNA of flatfish species attributes these elements a possible role in speciation because of the species-specificity of the satDNA families (Rodríguez et al., 2017). Although the main satDNA families found among *Gireoudia* species were not species-specific, a more in-depth satellitome analysis of these species and the *B. heracleifolia* populations from Chapter 6 (as done for *Vaccinium* in Sultana et al. (2020) or for the ladybird beetle in Mora et al. (2020)) would help to better understand the origin and evolution of the differentially expanded satDNA in *Gireoudia*. How population divergence and genetic incompatibility would be driven by these genome dynamics, however, is still unknown.

As discussed in Chapter 1, speciation is a complex process, affected by numerous factors. Studying and trying to understand a continuous process brings some challenges and leads to many open questions. The main question after considering our results from Chapter 6 would be whether the described genomic changes are sufficient to drive the partial reproductive isolation we found towards a stronger isolation. Kulmuni et al. (2020) discuss that this transition is driven by many factors that could be either ecological or of any other nature, but leaves a detectable signal in the genome. Serrato-

Capuchina and Matute (2018) discusses the role of TEs in reproductive isolation and provides examples that mostly rely on phenotypic (adaptive) changes, and concludes that there is still little direct evidence that TEs can facilitate reproductive isolation and ultimately speciation. To better understand what drives reproductive isolation in *Begonia*, high quality genomes of the closest species to *B. heracleifolia*, and of individuals from different genome size populations may help determine which genome changes are involved in the transition from partial to strong reproductive isolation.

7.6 Future approaches

This thesis has provided full-length genome assemblies for a series of species that can be used for many more repeat analyses to study the role of repetitive sequences in genome evolution. For example, the analysis of insertion site of TEs and satDNA to chart their genomic environment and study potential effects on neighbouring genes, or determine whether they have an effect on gene deletion, duplication or interruption. The detection of selection on TE insertions would also be of interest. These genomes bring countless possibilities for bioinformatics and genomic projects on the repeatome evolution of *Begonia*. Moreover, the use of available *Begonia* transcriptomic data (Emelianova, 2017) would give great support to the analysis of LTR expression patterns, for example, in different tissues or species.

Additionally, because of the many still unresolved sections and clades in this genus, many genome skimming datasets are being generated for *Begonia* for phylogenetic purposes. These would be useful to test the same approach taken in Chapter 5 in other *Begonia* radiations. Moreover, another interesting perspective would be brought by the repeatome analysis of genome skims generated using herbarium specimens (Dodsworth et al., 2018), with the aim of studying the evolution of the repeatome of the same populations through time.

Finally, setting up protocols for molecular and cytological studies on *Begonia* would bring us much closer to the understanding of the factors triggering TE activation, the effect of this activation or the location of these elements. Specifically, this could be done for the study of LTR-RT expression levels or insertion site analysis of repeat families of

interest using RT-qPCR or transposon display approaches. Moreover, developing cytological and hybridisation techniques for the chromosomal visualisation of specific repeat families across *Begonia* would provide important information regarding its repeatome evolution.

Bibliography

Ali MS. 2013. Genetic architecture of species level differences in Begonia Section Gireoudia. Ph.D. thesis, University of Edinburgh.

Baniaga AE, Barker MS. 2019. Nuclear genome size is positively correlated with median ltr-rt insertion time in fern and lycophyte genomes. *American Fern Journal*. 109:248–266.

Barchi L, Pietrella M, Venturini L, et al. (11 co-authors). 2019. A chromosome-anchored eggplant genome sequence reveals key events in solanaceae evolution. *Scientific reports*. 9:1–13.

Becking T, Gilbert C, Cordaux R. 2020. Impact of transposable elements on genome size variation between two closely related crustacean species. *Analytical Biochemistry*. p. 113770.

Carleton KL, Conte M, Malinsky M, Nandamuri SP, Sandkam BA, Meier JI, Mwaiko S, Seehausen O, Kocher TD. 2020. Movement of transposable elements contributes to cichlid diversity. *bioRxiv*. .

Chen JD, Zheng C, Ma JQ, Jiang CK, Ercisli S, Yao MZ, Chen L. 2020. The chromosome-scale genome reveals the evolution and diversification after the recent tetraploidization event in tea plant. *Horticulture Research*. 7:1–11.

Choi B, Weiss-Schneeweiss H, Temsch EM, So S, Myeong HH, Jang TS. 2020. Genome size and chromosome number evolution in korean iris l. species (iridaceae juss.). *Plants*. 9:1284.

Cronk Q. 1989. Measurement of biological and historical influences on plant classifications. *Taxon*. 38:357–370.

Dewitte A. 2010. Exploitation of 2n pollen to create genetic variation in the genus Begonia. Ph.D. thesis, Ghent University.

Dodsworth S, Guignard MS, Christenhusz MJ, Cowan RS, Knapp S, Maurin O, Struebig M, Leitch AR, Chase MW, Forest F. 2018. Potential of herbariomics for studying repetitive dna in angiosperms. *Frontiers in Ecology and Evolution*. 6:174.

Dumschott K, Schmidt MHW, Chawla HS, Snowdon R, Usadel B. 2020. Oxford nanopore sequencing: New opportunities for plant genomics? *Journal of Experimental Botany*. .

Emelianova K. 2017. Using Next Generation Sequencing to investigate the generation of diversity in the genus Begonia. Ph.D. thesis, The University of Edinburgh.

Forrest LL. 2001. Phylogeny of Begoniaceae. Ph.D. thesis, University of Glasgow.

Gatter T, Von Loehneysen S, Drozdova P, Hartmann T, Stadler PF. 2020. Economic genome assembly from low coverage illumina and nanopore data. *BioRxiv*. .

Gorelick R, Fraser D, Zonneveld BJ, Little DP. 2014. Cycad (cycadales) chromosome numbers are not correlated with genome size. *International Journal of Plant Sciences*. 175:986–997.

Griesmann M, Chang Y, Liu X, et al. (11 co-authors). 2018. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*. 361.

Hughes M. 2002. Population structure and speciation in Begonia L. Ph.D. thesis, University of Glasgow.

Jiménez-Ruiz J, Ramírez-Tejero JA, Fernández-Pozo N, et al. (11 co-authors). 2020. Transposon activation is a major driver in the genome evolution of cultivated olive trees (olea europaea l.). *The Plant Genome*. 13:e20010.

Kelly S. 2019. Editorial overview: Harvesting the fruits of plant genomics. *Current opinion in plant biology*. 48:A1.

Kulmuni J, Butlin RK, Lucek K, Savolainen V, Westram AM. 2020. Towards the completion of speciation: the evolution of reproductive isolation beyond the first barriers. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 375.

Leitch A, Leitch I. 2012. Ecological and genetic factors linked to contrasting genome dynamics in seed plants. *New Phytologist*. 194:629–646.

Leitch I, Bennett M. 2004. Genome downsizing in polyploid plants. *Biological journal of the Linnean Society*. 82:651–663.

Leong-Škorničková J, Šída O, Jarolímová V, Sabu M, Fér T, Trávníček P, Suda J. 2007. Chromosome numbers and genome size variation in indian species of curcuma (zingiberaceae). *Annals of Botany*. 100:505–526.

Ma J, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences*. 101:12404–12410.

Michael TP, VanBuren R. 2020. Building near-complete plant genomes. *Current opinion in plant biology*. 54:26–33.

Moeglein MK, Chatelet DS, Donoghue MJ, Edwards EJ. 2020. Evolutionary dynamics of genome size in a radiation of woody plants. *American Journal of Botany*. 107:1527–1541.

Molero J, Garnatje T, Rovira A, Garcia-Jacas N, Susanna A. 2002. Karyological evolution and molecular phylogeny in macaronesian dendroid spurges (euphorbia subsect. pachycladae). *Plant Systematics and Evolution*. 231:109–132.

Moonlight PW. 2017. Linking speciation and the niche: taxonomy, phylogeny and niche evolution in neotropical Begonia. Ph.D. thesis, University of Glasgow.

Moonlight PW, Ardi WH, Padilla LA, et al. (11 co-authors). 2018. Dividing and conquering the fastest–growing genus: Towards a natural sectional classification of the mega–diverse genus begonia (begoniaceae). *Taxon*. 67:267–323.

Mora P, Vela J, Ruiz-Ruano FJ, Ruiz-Mena A, Montiel EE, Palomeque T, Lorite P. 2020. Satellitome analysis in the ladybird beetle hippodamia variegata (coleoptera, coccinellidae). *Genes*. 11:783.

Moraes AP, Vatanparast M, Polido C, Marques A, Souza G, Fortuna-Perez AP, Forni-Martins ER. 2020. Chromosome number evolution in dalbergioid legumes (papilionoideae, leguminosae). *Brazilian Journal of Botany*. 43:575–587.

Naciri Y, Linder HP. 2020. The genetics of evolutionary radiations. *Biological Reviews*. 95:1055–1072.

Nguyen D, Peona V, Unneberg P, Suh A, Jern P, Johannesson H. 2020. Transposon- and genome dynamics in the fungal genus neurospora: insights from nearly gapless genome assemblies. *bioRxiv*. .

Osipowski P, Pawełkowicz M, Wojcieszek M, Skarzyńska A, Przybecki Z, Plader W. 2020. A high-quality cucumber genome assembly enhances computational comparative genomics. *Molecular Genetics and Genomics*. 295:177–193.

Petersen M. 2020. Comparative analysis of the insect mobile genetic element repertoire and its influence on genome size dynamics. Ph.D. thesis, Universitäts-und Landesbibliothek Bonn.

Rodríguez FR, de la Herrán R, Navajas-Pérez R, Cano-Roldán B, Sola-Campoy PJ, García-Zea JA, Rejón CR. 2017. Centromeric satellite dna in flatfish (order pleuronectiformes) and its relation to speciation processes. *Journal of Heredity*. 108:217–222.

Serrato-Capuchina A, Matute DR. 2018. The role of transposable elements in speciation. *Genes*. 9:254.

Shahid S, Slotkin RK. 2020. The current revolution in transposable element biology enabled by long reads. *Current Opinion in Plant Biology*. 54:49–56.

Shinzato C, Khalturin K, Inoue J, Zayasu Y, Kanda M, Kawamitsu M, Yoshioka Y, Yamashita H, Suzuki G, Satoh N. 2021. Eighteen coral genomes reveal the evolutionary origin of acropora strategies to accommodate environmental changes. *Molecular Biology and Evolution*. 38:16–30.

Sultana N, Menzel G, Heitkam T, Kojima KK, Bao W, Serçe S. 2020. Bioinformatic and molecular analysis of satellite repeat diversity in vaccinium genomes. *Genes*. 11:527.

Tian D, Xiao Y, Tong Y, Fu N, Liu Q, Li C. 2018. Diversity and conservation of chinese wild begonias. *Plant diversity*. 40:75–90.

Vitales D, Álvarez I, Garcia S, Hidalgo O, Nieto Feliner G, Pellicer J, Vallès J, Garnatje T. 2020. Genome size variation at constant chromosome number is not correlated

with repetitive dna dynamism in anacyclus (asteraceae). *Annals of Botany*. 125:611–623.

Wierzbicki F, Schwarz F, Cannalonga O, Kofler R. 2020. Generating high quality assemblies for genomic analysis of transposable elements. *bioRxiv*. .

Wong WY, Simakov O, Bridge DM, Cartwright P, Bellantuono AJ, Kuhn A, Holstein TW, David CN, Steele RE, Martínez DE. 2019. Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus hydra. *Proceedings of the National Academy of Sciences*. 116:22915–22917.

Wu S, Sun W, Xu Z, et al. (11 co-authors). 2020. The genome sequence of star fruit (averrhoa carambola). *Horticulture Research*. 7:1–10.

Xiaobai J. 1993. The distribution and propagation of six begonia species native to china and assesment of their potential as ornamental plants. In: International Symposium on Cultivar Improvement of Horticultural Crops. Part 3: Flowers 404. pp. 95–99.

Xing Y, Liu Y, Zhang Q, et al. (11 co-authors). 2019. Hybrid de novo genome assembly of chinese chestnut (castanea mollissima). *GigaScience*. 8:giz112.

Yi XG, Yu XQ, Chen J, et al. (11 co-authors). 2020. The genome of chinese flowering cherry (cerasus serrulata) provides new insights into cerasus species. *Horticulture Research*. 7:1–14.

Zhang J, Fu XX, Li RQ, et al. (11 co-authors). 2020a. The hornwort genome and early land plant evolution. *Nature plants*. 6:107–118.

Zhang L, Chen F, Zhang X, et al. (11 co-authors). 2020b. The water lily genome and the early evolution of flowering plants. *Nature*. 577:79–84.