# Genomic Resources for Evolutionary Studies in the Large, Diverse, Tropical Genus, *Begonia*

**Adrian Christopher Brennan · Stephen Bridgett · Mobina Shaukat Ali · Nicola Harrison · Andrew Matthews · Jaume Pellicer · Alex David Twyford · Catherine Anne Kidner**

**Abstract** *Begonia* is one of the ten largest angiosperm genera with over 1,500 species found throughout the tropics. To use this group as a model for the evolution of diversity in tropical herbaceous plants, we have produced three species transcriptomes, physical genome size measures, and two backcross genetic maps. We chose to focus on two Central American species, *B. conchifolia* and *B. plebeja*, and one SE Asian species, *B. venusta,* allowing us to pose questions at widely different evolutionary scales within the genus. We used next generation sequencing of cDNA libraries to produce annotated transcriptome databases for each of the three species. Though *Begonia* is functionally diploid, transcriptome analysis suggested a genome duplication occurred at or near the base of the *Begonia* clade. The genetic maps were built from first generation backcrosses in both directions between *B. plebeja* and *B.conchifolia* using 105 SNP markers in genes known to regulate development that were identified from the transcriptomes and the map bulked out with 226 AFLP loci. The genetic maps had 14 distinct linkage groups each and mean marker densities of between 3.6 and 5.8 cM providing between 96 and 99 % genomic coverage within 10 cM. We measured genome size 1C value of 0.60 and 0.63 pg for *B. conchifolia* and *B. plebeja* corresponding to recombination rates of between 441 and 451 Kb per cM in the genetic maps. Altogether, these new data represent a powerful new set of molecular genetic tools for evolutionary study in the genus *Begonia*.

**Keywords** *Begonia* · Genetic map · Transcriptome · Tropical diversity · Genome duplication

A. C. Brennan
Estación Biológica de Doñana (EBD-CSIC),
Sevilla, Spain

S. Bridgett
Genepool, University of Edinburgh,
Mayfield Road,
Edinburgh EH9 3JU, UK

M. Shaukat Ali · N. Harrison · A. D. Twyford · C. A. Kidner (✉)
Royal Botanic Garden Edinburgh,
20a Inverleith Row,
Edinburgh EH3 5LR, UK
e-mail: catherine.kidner@ed.ac.uk

N. Harrison
East Malling Research,
New Road, East Malling,
Kent ME19 6BJ, UK

A. Matthews
Queen Mary, University of London,
Mile End Road,
London E1 45NS, UK

J. Pellicer
Royal Botanic Gardens, Kew,
Kew, Richmond TW9 3AB, UK

A. C. Brennan · M. Shaukat Ali · N. Harrison · A. D. Twyford · C. A. Kidner
Institute of Molecular Plant Sciences, University of Edinburgh,
Mayfield Road,
Edinburgh EH9 3JU, UK

## Introduction

Owing to the historical nature of speciation, it is difficult to test hypotheses concerning the underlying evolutionary forces that originally drove divergence. Convergent evolution offers a chance to tease out signals from the noise as the independent evolution of the trait in each lineage can be examined to determine whether similar selective pressures and similar genetic changes were responsible (Elmer and Meyer 2011; Wood et al. 2005). Some of the most exciting insights into adaptation have come from analysis of convergent traits in large lineages (for example: Drummond et al. 2012; Kaeuffer et al. 2012; Melo et al. 2011; Strecker et al. 2012; Whitehead et al. 2012; Woodard et al. 2011).

With over 1,500 species growing in a wide range of environments, *Begonia* L. (Begoniaceae) offers many examples of convergent evolution in lineages separated by a few or many millions of years (Neale et al. 2006). This allows the evolution of traits to be examined on a range of scales. *Begonia* is found throughout the tropics in habitats which range from wet rainforest to seasonally dry scrubland. Species radiations in the New World, Southeast Asia and Africa have generated many examples of parallel evolution, for example in leaf form, plant architecture, inflorescence arrangement and drought tolerance (Neale et al. 2006). *Begonia* is also a major horticultural crop, being important in both bedding plants and indoor plants markets (Hvoslef-Eide and Munster 2007) and consequently large, accessible collections of species and hybrids are available (Tebbit 2005). Additionally, methods of in-vitro propagation and genetic transformation that have been developed for commercial exploitation of *Begonia* can also be applied to research (Kishimoto et al. 2002; Xu et al. 2011).

Most *Begonia* species are local endemics and population genetic studies using a range of marker types have shown strong population structures (Hughes et al. 2003; Hughes and Hollingsworth 2008; Matolweni et al. 2000; Nakamura et al. 2012). However, in the greenhouse, many species are interfertile (Tebbit 2005; Dewitte et al. 2011). Interfertility between species opens up the possibility of using genetic analysis in interspecific crosses to identify loci associated with species-level variation. Until recently the power of genetics to investigate evolutionary events has been largely limited to model species or their close relatives, but the advent of next generation methods of sequencing and high throughput genotyping techniques have lowered the barrier to developing genetic resources for any organism allowing us to investigate interesting evolutionary patterns wherever they occur (Harrison and Kidner 2011; Ekblom and Galindo 2011; Wheat 2010).

Here we report the development of genetic resources for *Begonia* which will allow study of evolution at different levels, from populations to species through to the whole genus. We present transcriptome data for three species of *Begonia* that provide the basis of a new reference sequence database and suggest a whole-genome duplication may have occurred early in the lineage leading to *Begonia*. Genetic variation identified between transcriptomes was used to produce the first genetic recombination maps for *Begonia*. These genetic map resources provide tools for mapping the genes that underlie interspecific and population level traits and for identifying the functional genetic basis of these quantitative differences.

## Results

We chose two closely related but ecologically contrasting species from Central America for transcriptome sequencing and to generate mapping populations. *Begonia plebeja* is widespread in northern Mexico and found in seasonally dry forests (Fig. 1a) whilst *B. conchifolia* has a more restricted distribution in wet rainforests of Southern Mexico and Central America (Burt Utley 1985; Fig. 1c). These species are both members of a phylogenetically well supported Central American section *Gireoudia* within *Begonia* with characteristic chromosome number, 2n = 28 (Fig. 1d, Dewitte et al. 2009; Legro and Doorenbos 1971; Thomas et al. 2011).

To represent more of the range of diversity in *Begonia* a third species, *B. venusta,* that is more distantly related to *B. conchifolia* and *B. plebeja* (Fig. 1b), was included in the transcriptome analysis. This species grows in wet rainforests and field margins along the Malaysian Peninsula (Kiew 2005). *Begonia venusta* belongs to the parallel radiation of *Begonia* in Southeast Asia. The lineage leading to *B. conchifolia* and *B. plebeja* split from that leading to *B. venusta* very early during *Begonia* evolution (Fig. 1d). *Begonia venusta* is placed in section *Platycentrum* and has a chromosome number of 2n = 44, in contrast to the 2n = 28 of section *Gireoudia* (Dewitte et al. 2009; Legro and Doorenbos 1971; Thomas et al. 2011).

### Genome Size Measurements

The genome sizes of *B. conchifolia* and *B. plebeja*, expressed as 1C-values (i.e. the DNA content of an unreplicated gametic nucleous sensu Greilhuber et al. 2005), were fairly similar (1C=0.60±0.01 pg and 1C=0.63± 0.01 pg respectively), but higher for *B. venusta* (1C= 1.11±0.00 pg). These genome size similarities reflect the different chromosome counts for the two *Gireoudia* species and *B. venusta*.

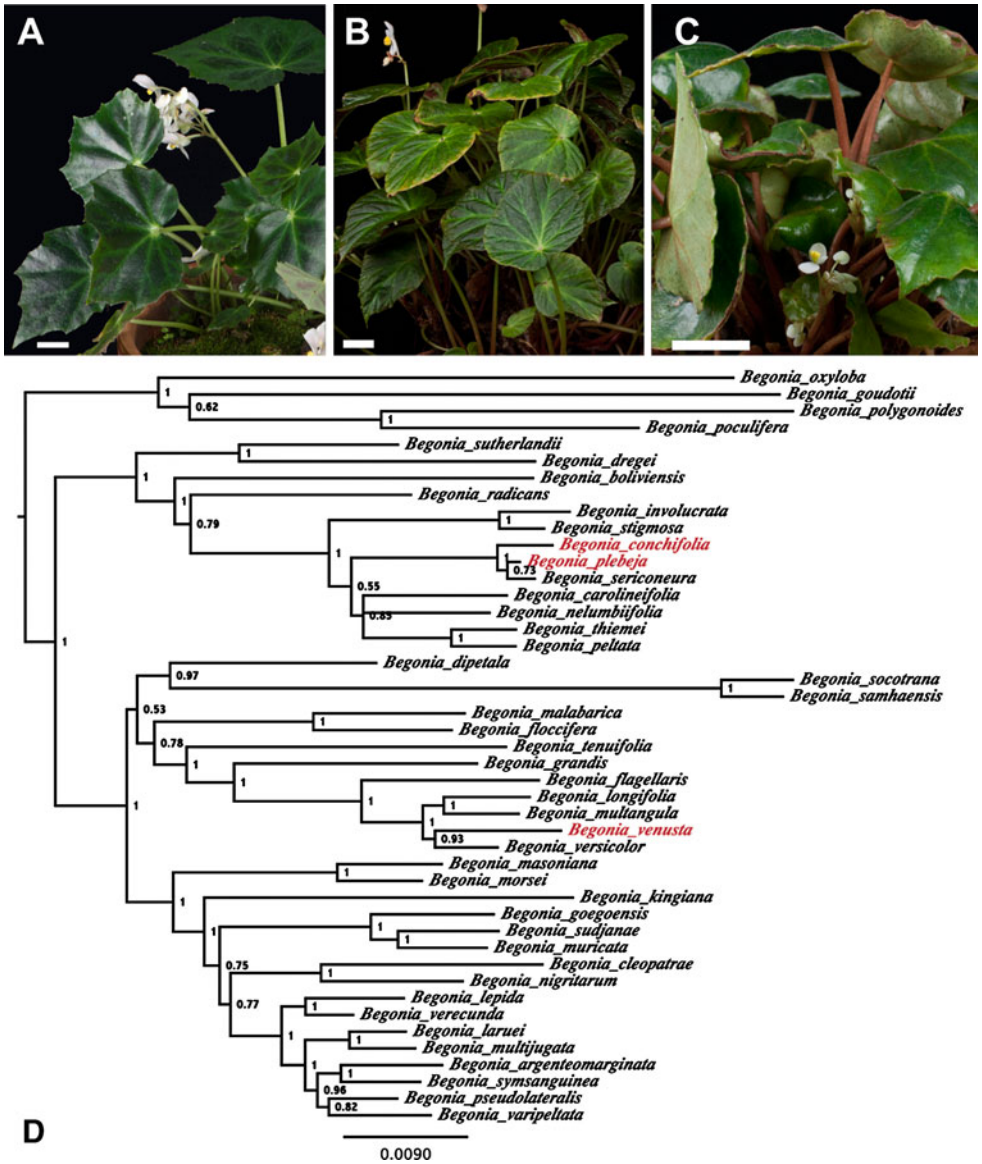



**Fig. 1** *Begonia plebeja* (**a**), *Begonia venusta* (**b**), *Begonia conchifolia* (**c**). Scale bars represent 2 cm. Bayesian phylogeny of South American and Southeast Asian *Begonia* based on the chloroplast sequences *ndhA*, *ndhF-rpl32* and *rpl32-trnL* (**d**)

## Transcriptome Analysis

Summary statistics for the Roche 454 Titanium sequence reads are shown in Table 1. Using total RNA extracted from vegetative apices for cDNA synthesis, between 200,000 and 260,000 sequence reads were obtained for each species using Roche 454 sequencing. Assembly of these reads using Newbler 2.5 resulted in 15,855, 17,385 and 16,206 isotigs (consensus sequence fragments derived from groups of assembled reads) for *B. plebeja, B. conchifolia* and *B. venusta,* respectively, with over half of the isotigs longer than 750 bp (Table 1). We used BLASTX to identify orthologous genes in *Cucumis sativus* (cucumber; the closest related species with a sequenced genome) and found that 82.1 %, 70.7 % and 78.0 % of *B. conchifolia*, *B. plebeja* and *B. venusta* isotigs had a match in the cucumber transcriptome with an expect value of $e^{-10}$ or lower. These matches were to 8611, 8082 and 8563 unique cucumber transcripts respectively, indicating that there were often two *Begonia* isotigs for any one cucumber transcript. Assuming that the assemblies are largely correct, this could indicate a genome duplication in *Begonia*.

The similar numbers of isotigs produced for the different species suggests that, although *B. venusta* has a higher genome size than the other two species, it does not have significantly more genes. An increase in the nuclear DNA content may reflect expansion in non-coding sequence (such as transposable elements) or a very recent genome duplication which

**Table 1** Statistics for sequencing and for Newbler 2.5.2 assemblies

| | PLE | CON | VEN | TRI |
|---|---|---|---|---|
| Number of bases sequenced | 71198213 | 73094984 | 63949869 | |
| Number of reads | 204477 | 260385 | 206884 | |
| N50 read length | 430 | 401 | 386 | |
| GC content of reads | 45.47 % | 46.27 % | 45.39 % | |
| % Bases aligned | 89.30 % | 87.27 % | 87.64 % | 90.81 % |
| % Reads aligned | 88.90 % | 81.54 % | 85.94 % | 88.70 % |
| Number of isogroups | 13929 | 15671 | 14391 | 28526 |
| Number of isotigs | 15855 | 17385 | 16206 | 46774 |
| Isotig N50 | 773 | 759 | 760 | 1041 |
| Number of large contigs (> = 1 kb) | 7482 | 8017 | 7561 | 15597 |
| Numbers of singletons | 14767 | 23341 | 16453 | 27316 |
| % Inferred error | 1.7 | 1.45 | 1.79 | 2.55 |
| % Isogroups with more than 1 isotig | 8.5 | 7.4 | 7.3 | 16.9 |
| Average number of isotigs per isogroup | 1.1 | 1.1 | 1.1 | 1.3 |

(PLE = *B. plebeja*, CON = *B. conchifolia*, VEN = *B. venusta* and TRI = the combined tri-scriptome assembly)

has not had time to diverge sufficiently for reads from different paralogs to be assembled into separate isotigs.

We used two different methods to examine similarities between the three transcriptomes. First, reciprocal BLAST searches between species were used to identify likely orthologs as those sequences that show high similarity in different species datasets (implemented with RBH, Reciprocal Best Hit, Orthologs; Barker et al. 2010) . The overlap between the transcriptomes was low, about a third of the isotigs in any one species had an identified ortholog in another species (Fig. 2a). However, RBH Orthologs is very conservative in its identification of ortholog pairs, requiring an extensive length of match of sequence identity. We examined the transcriptomes again using a much less conservative measure—the presence of a reciprocal BLASTN hit of any length and any e value. By this measure, still fewer than half of *B. conchifolia* or *B. plebeja* isotigs had orthologs in *B. venusta* (40.8 %, 42.0 % respectively) and this percentage was only marginally increased in the comparison between the more closely related *B. conchifolia* and *B. plebeja* (44.3 %). This observation suggests that the poor overlap was possibly due to incomplete coverage of genes with low levels of expression rather than poor conservation of sequences between species.

To investigate the overlap between transcriptomes further, three sets of isotigs (those found in all three species, isotigs found in only a pair of species and isotigs unique to one species, as identified by RBHorthologs) were searched against the NCBI Refseq Green Plant protein database (Pruitt et al. 2012). The percentage of good matches to sequences in the database was higher for isotigs shared between all three species, presumably reflecting the presence of more strongly conserved genes in this set (Fig. 2a). This relatively small conservative subset of 3392 isotigs that have been confirmed against known proteins provides a good starting point for molecular evolutionary studies in this genus.

To study the variation in our transcriptomes, we used GS mapper (Roche) to map the reads from each species onto the assembled isotigs of each species in turn. The results are shown in Table 2. Polymorphism frequency, measured as mean number of polymorphisms per 100 bases, ranged from

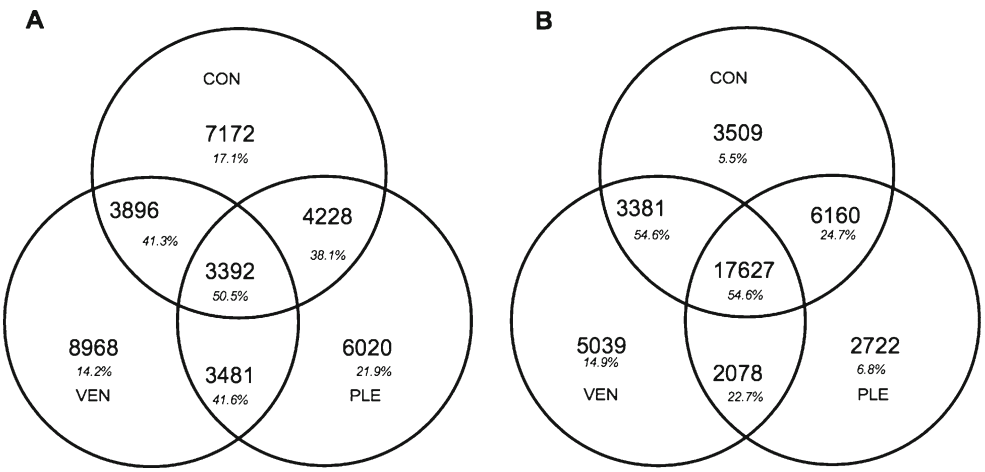



**Fig. 2** Pattern of sharing (**a**) of assembled sequences (isotigs) between the three individual species transcriptomes and (**b**) of reads within assembled sequences from the combined assembly of all three species transcriptomes (tri-scriptome) together. Counts are numbers of isotigs in each sharing category and percentages are those sequences that show good matches with known proteins from the Green Plants RefSeq database

**Table 2** High confidence SNPs identified between and within species

| Comparison | Total SNPs | Species-level SNPs | Mapped bases | % Polymorphisms |
|---|---|---|---|---|
| C onto P | 29988 | 21944 | 41216046 | 0.0532414 |
| P onto C | 31104 | 15522 | 39763005 | 0.0390363 |
| V onto C | 88112 | 70189 | 38231807 | 0.183588 |
| C onto V | 80700 | 66110 | 44100971 | 0.149906 |
| V onto P | 81518 | 65223 | 37893074 | 0.1721238 |
| P onto V | 73073 | 54273 | 39634494 | 0.1369338 |
| P onto P | 6079 | 517 | 60719146 | 0.0100117 |
| C onto C | 1852 | 399 | 60067876 | 0.0030832 |
| V onto V | 4633 | 444 | 52316073 | 0.0088558 |

(P = *B. plebeja*, C = *B. conchifolia*, V = *B. venusta*)

0.05 % for *B. plebeja* mapped onto *B. conchifolia* to 0.18 % for *B. venusta* mapped onto *B. conchifolia*. As expected, given the evolutionary relationships between the species, more high-confidence polymorphisms were detected between either *B. plebeja* or *B. conchifolia* and *B. venusta* than between *B. plebeja* and *B. conchifolia*. Polymorphism frequencies within each species, which are indicative of heterozygosity, ranged from 0.010 % for *B. plebeja* to 0.003 % for *B. conchifolia.*

As an alternative method of comparison, the low variation between transcriptomes allowed us to assemble all three transcriptomes together using Newbler 2.5.2 (Roche) as a joint 'tri-scriptome'. This reduced the number of unassembled reads from 54,561 (14,767+23,341+16,453) for the individual assemblies to 27,316 for the joint assembly and gave us a set of 38,448 isotigs. A total of 72.2 % of the joint tri-scriptome isotigs had sequence from more than one species and 43.5 % contained reads from all three species, confirming that in most cases the program has assembled reads from orthologs together. This gave us a set of 17,627 isotigs representing genes found in all three species, over half of which had a BLASTX match in the Green Plant Refseq database (with e value of $e^{-40}$ or less; Fig. 2b). BLASTX showed also that these isotigs had hits at $e^{-10}$ or below to 12442 unique cucumber genes that represent nearly half of the annotated cucumber proteins (25600). Additional BLASTX matches of tri-scriptome reads against the Uniprot database using Annot8r were used to assign GO-term, EC and KEGG functional annotations to these matching isotigs (Supplementary Table 1).

Figure 3 shows histograms of the number of reads per isotig for the joint tri-scriptome compared to the number of reads in the individual assemblies, illustrating that the assembly increased the read depth of isotigs. The joint tri-scriptome also increased the isotig lengths (see Fig. 4). The longest isotigs in this assembly were the mostly highly expressed genes such as tubulin, suggesting that assembly was still limited by coverage, but a cytochrome P450 family member and a Armadillo/beta catenin-like repeat were also highly represented.

We used BLASTN to compare the joint tri-scriptome to the individual assemblies. In total, 2,695 isotigs from the

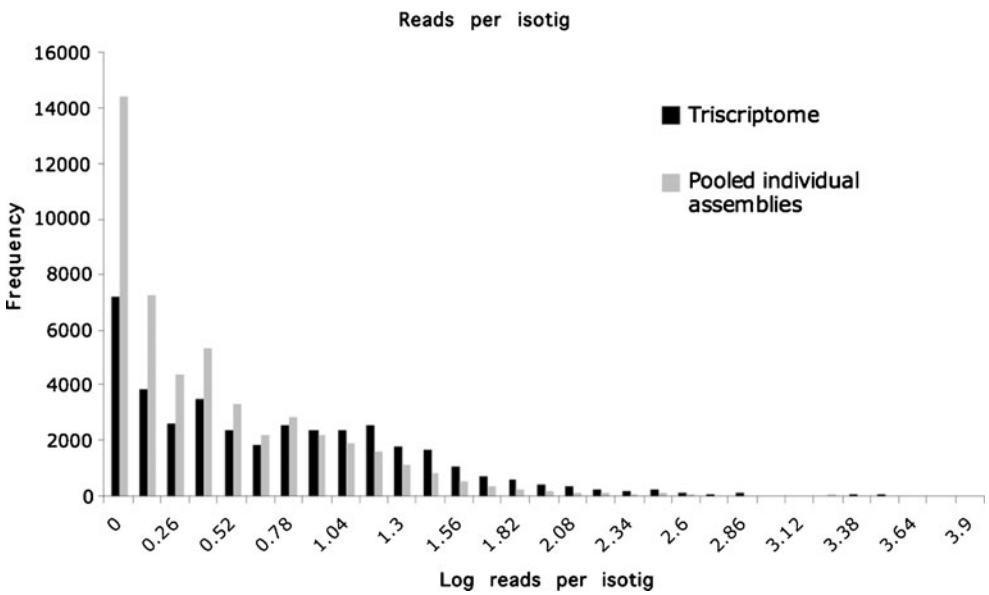



**Fig. 3** The number of reads per assembled sequence (isotig) in the tri-scriptome assembly and in the pooled individual assemblies. More tri-scriptome isotigs have medium to large numbers of reads

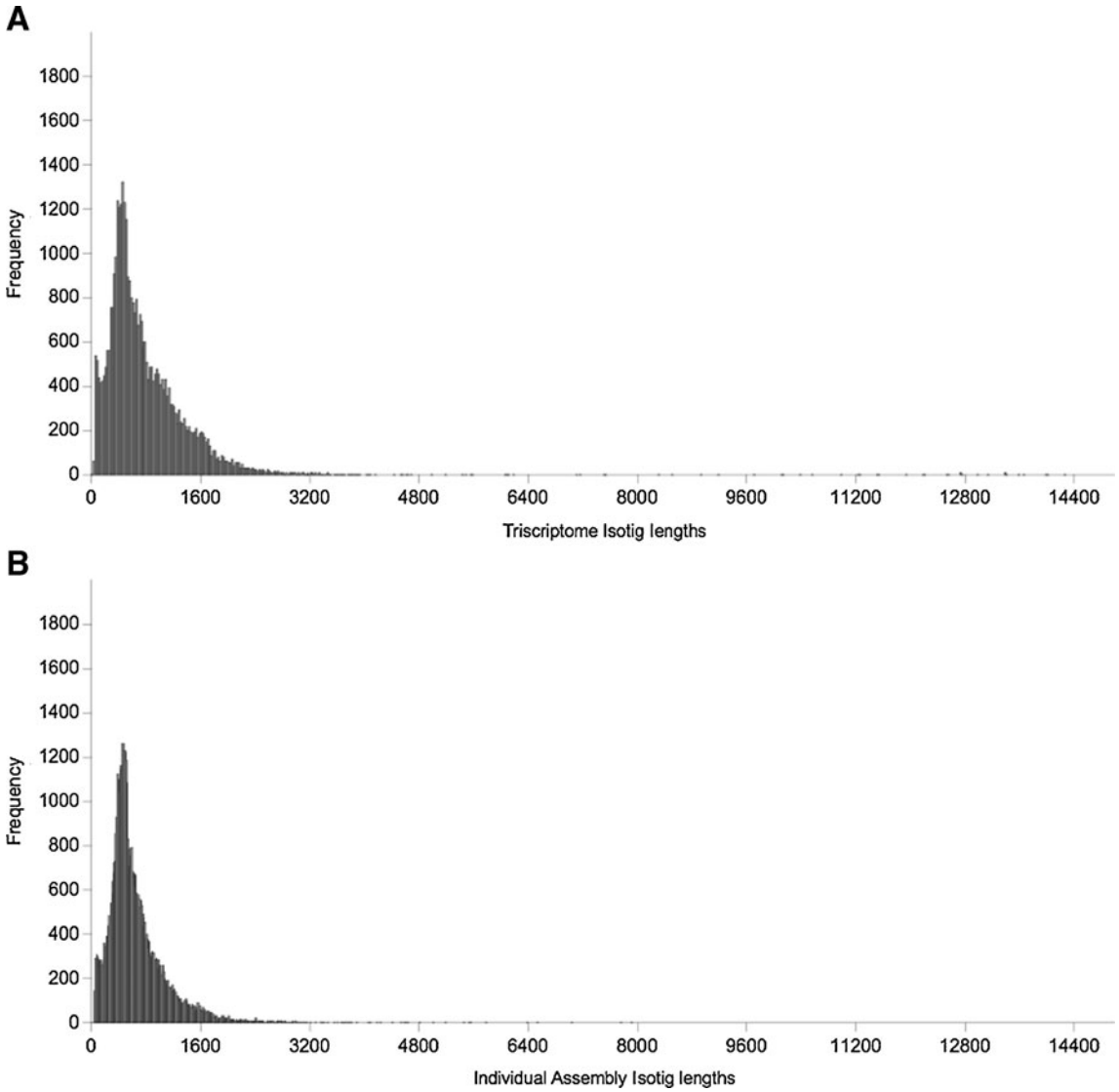



**Fig. 4** Isotig lengths (bp) for the individual assemblies (**a**) and the tri-scriptome assembly (**b**). Note that the tri-scriptome has longer isotigs

joint transcriptome had no corresponding isotig in any individual, suggesting that they represent genes identifiable only through the combination of data from all three transcriptomes. As expected these isotigs had low numbers of reads (2.4±5.37) in comparision to 39.7 (±129.6) for isotigs that were also found in at least one individual assembly. However, 2,358 of the new isotigs had a BLASTX match in the Green Plant database, including 254 matches at e < = e−40, confirming they were likely to represent true genes and not assembly artifacts.

The transcriptomes were produced from non-normalised cDNA pools and therefore the read number per isotig is related to the expression level of the gene. Figure 5 shows plots of the number of reads per isotig from each species. In order to control for the partial nature of the transcriptomes only those isotigs which include at least one read from each species are shown. There was a strong correlation between expression levels in *B. plebeja* and *B. conchifolia* but not between *B. venusta* and either *B. plebeja* or *B. conchifolia* (Fig. 5) suggesting that expression differences have accumulated with evolutionary distance across the genus *Begonia*. While limited coverage of the transcriptomes limits the statistical certainty of observed patterns of gene expression, we were able to make some initial qualitative observations that could be followed up with further confirmatory studies.

The sequence with the highest number of reads was a peroxidase expressed very highly in *B. conchifolia* (2,960 reads from the single species assembly) and *B. plebeja* (3,437 reads) compared to *B. venusta* (491 reads). The *Arabidopsis thaliana* ortholog of this gene (*At4g21960*, *PRXR1*) is also highly expressed, at 24.5 times the average transcript level (TAIR www.arabidopsis.org). Peroxidases

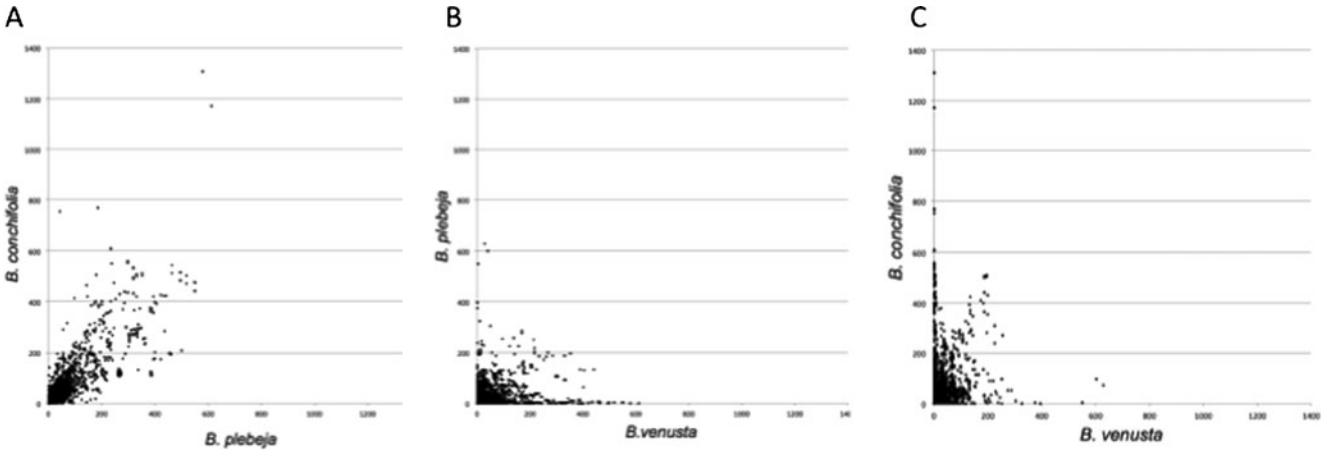



**Fig. 5** Paired-species comparisons of transcriptome read numbers per isotig ortholog. *B. conchifolia* and *B. plebeja* (**a**), *B. plebeja* and *B. venusta* (**b**), *B. conchifolia* and *B. venusta* (**c**)

are involved in stress response, in particular regulation of damage from photorespriation (Zamocky et al. 2008). Higher expression in *B. conchifolia* and *B. plebeja* than *B. venusta* could indicate a greater tolerance to high light levels, or it could simply be due to transient stresses in the individual plants assayed.

A *SHEPHERD* (*SHP*)-like gene was expressed highly in *B. venusta* (635 reads) in comparison to *B. plebeja* (49 reads) and *B. conchifolia* (102 reads). This is a heat-shock protein family member and the *A. thaliana* orthlog (*At4g24190*) may be involved in the regulation of meristem size via regulation of CLAVATA (CLV) protein folding in response to abiotic stress (Ishiguro et al. 2002; Song et al. 2009). It is also highly expressed in *A. thaliana* at 10.5× average transcription levels (TAIR). High levels of heat shock proteins are associated with greater tolerance of abiotic stresses such as temperature (Kadota and Shirasu 2012; Sangster et al. 2008).

Another interesting isogroup, which may represent a type 2 metallothionein metal-binding protein, was very highly expressed in all transcriptomes (*B. conchifolia* 1,365 reads, *B. plebeja* 1,016 reads, *B. venusta* 590 reads) but does not have a clear ortholog in the Green Plant Protein or NCBI databases (all BLASTX matches have relatively weak e values greater than $e^{-20}$). Although the expression pattern and genus-wide sequence variation of this gene could be easily determined, without functional information from a model species it would be difficult to formulate a hypothesis as to its role.

An episode of increased gene duplication, including a whole-genome event, results in an intermediate peak in the negative exponential frequency distribution of sequence divergence between duplicate genes (Vision et al. 2000; Paterson et al. 2004; Schranz and Mitchell-Olds 2006; Barker et al. 2008). A peak of synonymous substitution values (Ks = 0.51−0.58) for paralogous genes was observed for all three *Begonia* transcriptomes (Fig. 6). The presence of this peak in all three species suggests that a whole genome duplication occurred in the lineage leading to the genus *Begonia* or early during *Begonia* evolution. Likely dates for the origin of *Begonia* range from 24 to 45 MYA (Goodall-Copestake et al. 2009) and the origin of the Southeast Asian lineage, to which *B. venusta* belongs is dated to ~22 MYA (Thomas et al. 2012). The *Begonia* lineage split from that of cucumber (which lacks this genome duplication) ~82.1 MYA (Schaefer et al. 2009). *Begonia* isotigs whose paralog pairs fall in the peak of Ks have an average Ks of 1.159 with their cucumber orthologs (as identified by reciprocal TBLASTX hits), suggesting that the *Begonia*-lineage duplication occurred sometime after the split from cucumber, but before 22 MYA. Conservation of paralogs over this period of time suggests diversification of gene functions and that these three transcriptomes will provide valuable data for studying the evolution of gene families.

Analysis of the ratio of non-synonymous to synonymous substitutions (Ka/Ks values) for *Begonia*-cucumber ortholog pairs also identified several with elevated Ka/Ks ratios, and several with reduced ratios. Orthologs with elevated Ka/Ks values included disease resistance and signaling genes such as NB-ARC domain-containing disease resistance protein (Ka/Ks = 1.20) and a receptor like protein of the RPL44 family (Ka/Ks = 1.22) (Wang et al. 2008). The reduced Ka/Ks ortholog set included key functional genes such as Cyclin-dependent protein kinase CYCB1 ortholog (Ka/Ks = 0.07) (Day and Reddy 1998), and an NKS1 ortholog, likely involved in ion homeostatis (Ka/Ks = 0.044) (Choi et al. 2011). Further analysis of the relative rates of change will be undertaken once the paralogs can be more clearly distinguished by comparison to a draft genome sequence.

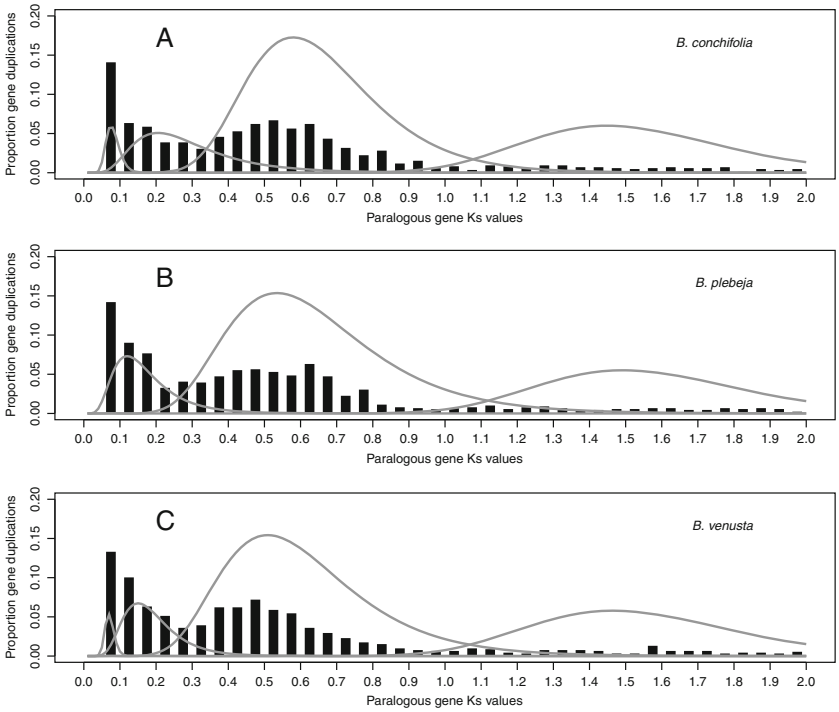



**Fig. 6** Evidence for genome duplication from patterns of Ks in ortholog clusters. *B. conchifolia* (**a**), *B. plebeja* (**b**) and *B. venusta* (**c**) all show a significant peak in the same position (Ks = 0.51−.58), supporting a genome wide duplication in early *Begonia* evolution or before the emergence of the genus. Heights of peaks indicate qualitatively the strength of evidence for a peak at that position

## Genetic Mapping

Genetic maps for the genomes of *B. conchifolia* and *B. plebeja* were constructed for the two backcross mapping families (Fig. 7, Supplementary Table 1, and 2). A total of 117 individuals from the backcross to *B. conchifolia* (CBC) and 225 individuals from the backcross to *B. plebeja* (PBC) were genotyped for mapping. A total of 4,867 potential AFLP loci were obtained using 12 primer combinations and 255 used for mapping in one or both populations following quality control analysis. SNPs in 105 genes with potential roles in development were identified from the transcriptomes and genotyped by competitive allele specific PCR technology (KASP) and these SNP genotypes were then mapped in one or both backcross families. The combined SNP and AFLP datasets consisted of 199 markers genotyped across 117 individuals for the CBC family and 330 markers genotyped across 225 individuals for the PBC family (Supplementary Table 1, and 2).

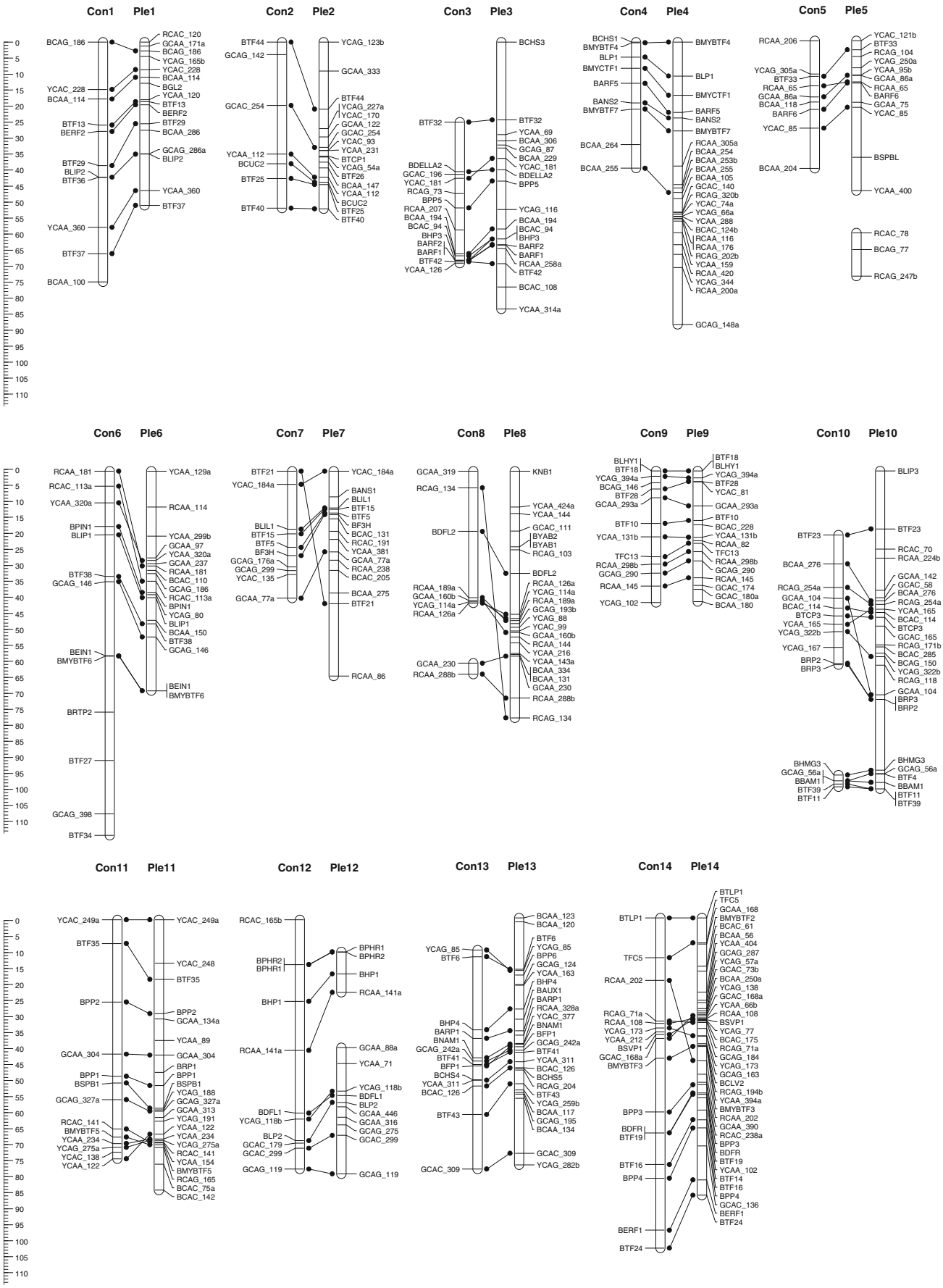
In total, 165 and 294 markers were placed on the final CBC and PBC genetic maps respectively, with 136 of these markers shared between maps. Other genotyped markers could not be mapped with high confidence, either because they appeared unlinked to all other markers (linkage scores < logarithm of odds [LOD] 4) or showed anomalous strong linkage to markers in different well supported linkage groups. A more detailed breakdown of marker composition is provided in Supplementary Table 1. In general, the *B. plebeja* parent was heterozygous at more markers than *B. conchifolia*, generating more segregating polymorphisms that could be mapped.

Both genetic maps consisted of 16 linkage groups. However a pair of linkage groups in each map was represented by weakly linked ends of a single chromosome in the other map indicating that these extra linkage groups are likely to be artifacts of the mapping method, rather than biological reality and that the number of distinct linkage groups in each map is actually 14 (Fig. 7). Map summary statistics are shown in Supplementary Table 2. The total map distance was estimated to be between 1,034 and 1,043 cM for the CBC map and between 1,099 and 1,110 cM for the PBC map for the add2s and method4 genome length estimators, respectively. Similarly, mean observed linkage group length was slightly smaller for CBC at 53.6 cM (st. dev. 30.5 cM), than for PBC at 61.6 cM (st. dev. 26.3 cM).

Consistent with the different numbers of markers mapped onto each map, mean marker distance was greater for CBC at 5.8 cM (st. dev. 5.5 cM), than for PBC at 3.6 cM (st. dev. 4.1 cM). The percentage of the genome within 5 cM of a

**Fig. 7** Genetic linkage maps based on reciprocal backcross populations between *B. plebej*a and *B. conchifolia*. Map distances in Kosambi centiMorgans are shown in the left scale bar. Linkage group labels from the backcross to *B. conchifolia* start with "Con", while those from the backcross to *B. plebeja* start with "Ple". Linkage groups are ordered in pairs, one from each genetic map that share markers (i.e. orthologous chromosomes) as shown by linking lines ▶

T


Con1 Ple1
Con2 Ple2
Con3 Ple3
Con4 Ple4
Con5 Ple5
Con6 Ple6
Con7 Ple7
Con8 Ple8
Con9 Ple9
Con10 Ple10
Con11 Ple11
Con12 Ple12
Con13 Ple13
Con14 Ple14

mapped marker was estimated to be 79 % in the CBC map and 93 % in the PBC map while 96 % or 99 % of the genomes were estimated to be within 10 cM of a marker according to the estimator of Fishman et al. (2001). However, markers were not evenly distributed in either map according to dispersion chi square tests. Uneven distribution of markers across genetic maps is typical of highly structured eukaryotic genomes. We speculate that clusters of SNP markers could indicate gene rich regions, while clusters of AFLPs could indicate repetitive sequence regions with a relatively high frequency of restriction sites for the enzymes used to generate AFLPs. Alternatively, since mapping families represent interspecies crosses, local chromosomal rearrangements between species that limit homologous recombination in hybrids could appear as clusters of markers on genetic maps.

Extensive synteny in order of shared homologous markers was observed between homologous linkage groups in the two backcross maps. Across 136 shared markers, only four markers on separate linkage groups (BTF21, RCAG_134, GCAA_104, and RCAA_202 on linkage groups 6, 7, 10, and 14, respectively, Fig. 7) showed linkage order differences that involved more than four adjacent markers (Fig. 7, Supplementary Table 1). Other minor differences in linkage order are more likely to reflect genotyping errors (mean 1.96 % among all AFLP loci) or missing data (mean 7.63 and 3.31 % AFLPs and SNPs, respectively) than differences in gene order.

## Discussion

*Begonia* is a highly speciose genus of tropical plants but their success appears to rest on high diversity and endemism between regions rather than on plasticity and wide species distributions (Hughes and Hollingsworth 2008). We describe here a set of tools that can be used to study the genetic basis of this diversity in the face of ongoing threats to tropical biodiversity in general (Myers et al. 2000).

We used two closely related species from Mexico and Central America (*B. plebeja* and *B. conchifolia*) and a third species from a parallel radiation of *Begonia* species in Southeast Asia (*B. venusta*). The genome size 1C values for these three species were determined to be 0.63 pg, 0.60 pg and 1.11 pg respectively. Previous research in the genus showed that *Begonia* is characterized by having small genomes (Dewitte et al. 2009), but differences up to around 6-fold have been reported between species with similar chromosome numbers, suggesting extensive gene duplication, transposon activity or repeat expansion can occur in *Begonia*. The distribution of Ks values for paralog pairs in *B. venusta* provides no evidence for a recent genome duplication suggesting that the larger genome of *B. venusta* reflects an increase in the number transposons and repetitive elements. However, a very recent duplication may result in paralogs so similar that they would be assembled together into a single isotig leaving no evidence in the transcriptome sequences of the duplication. A proportion of the polymorphisms detected in *B. venusta* could represent sequence variation between recent paralogs rather than alleles.

Partial transcriptomes were generated from vegetative buds from each species and orthologs and paralogs identified. Overlap between the transcriptomes was about a third, and 3,392 genes with orthologs in all three species were identified. Polymorphism rates vary from 0.003 % for *B. conchifolia* to 0.010 % for *B. plebeja*. The higher polymorphism frequency in *B. plebeja* may reflect more frequent outcrossing in this species, as it has larger, showier flowers with more stamens than *B. conchifolia* and occurs in more open habitats conducive to insect mediated cross-pollination (Burt Utley 1985).

As the polymorphism frequencies between the species were low (maximum of 0.17 % between *B. conchifolia* and *B. venusta*), we assembled all three transcriptomes together into a joint tri-scriptome in order to increase the number of isotigs assembled and their read depth and length. This generated a set of 17,627 isotigs including reads from all three species. By using all three species, we have come closer to the typical number of genes expressed in angiosperms (25,000–35,000 Bennetzen et al. 2004; Sterck et al. 2007), bearing in mind that some of these isotigs could be splice variants and chimeric sequences that have escaped detection in our analyses and that these transcriptomes were made from vegetative buds rather than a full range of tissue types and ages. We concluded that the joint tri-scriptome assembly was a useful means of identifying additional *Begonia* genes, although further sequencing would be required to eliminate the possibility that some isotigs had been assembled from paralogous gene families.

The read numbers per species from this set of 17,627 isotigs can be used as an indication of comparative expression levels. The Central American species, *B. conchifolia* and *B. plebeja* showed patterns of expression more similar to each other than to the Southeast Asian *B. venusta*. This presumably reflects changes in gene expression patterns over evolutionary time. The outliers, with highly different expression between species, are candidates for differential regulation. Alternatively, apparent differences in representation might reflect sequence divergence, as divergent transcripts might be assembled into species-specific contigs which would not be detected in the combined species transcriptome analyses. High levels of sequence divergence would also mark a gene out as worthy of further study as a candidate for divergent selection between species.

An intermediate peak in the distributions of substitution rates (Ks values) in paralogous isotigs identified in each species support the hypothesis of a whole genome duplication early in *Begonia* evolution to at least before 22 MYA.

Whole genome duplications are common in angiosperm lineages and contribute to potential for diversification (Otto and Whitton 2000; Petit and Thompson 1999; Vamosi and Dickinson 2006). The variation in chromosome number and genome size in *Begonia* as a whole suggests that genome duplication could be an important driver of diversity within the group (Dewitte et al. 2009; Legro and Doorenbos 1971). Ongoing research to identify and analyse orthologous and paralagous gene sets will undoubtedly provide further insights into the interplay between genome duplication and species diversification in *Begonia*.

A genetic map was constructed by genotyping backcross populations between *B. conchifolia* and *B. plebeja* for 105 SNPs in developmental genes and 255 AFLP markers. Individual maps for each backcross show that synteny is highly preserved between the two species suggesting that their divergence has not been accompanied by large scale genomic restructuring. The most closely related species to *Begonia*, which has a genetic map and genome sequence is cucumber, *Cucumis sativus* (Huang et al. 2009; Zhang et al. 2012). We looked for synteny between the sequenced *Begonia* markers and their orthologs in cucumber by BLASTX searching the isotigs containing the markers against the cucumber genome to find cucumber best hits and comparing relative positions. This revealed only one case of co-linearity: in markers that are 1.8 cM apart in *B. plebeja* and 350 Kb apart in *C. sativus*. Absence of large-scale synteny is perhaps not surprising for species whose lineages most probably split in the late Cretaceous (81 MYA) (Schaefer et al. 2009), but future studies with higher density genetic maps or genome assemblies are required to better assess and understand the pattern and scale at which synteny breaks down between these lineages.

Mean marker distance is 5.8 cM for the *B. conchifolia* map and 3.6 cM for the *B. plebeja* map covering 96 % and 99 % of the genome within 10 cM, respectively. As mentioned above, the non-replicated holoploid nuclear DNA contents (1C-value) of *B. conchifolia* and *B. plebeja* were estimated to be 591.5 and 616 Mb respectively (1 pg equates to 978 Mb, Doležel et al. 2003). Although slightly smaller, these genome size values are comparable to those previously published for *Begonia*, where 1C-values of 694.3 to 733.5 Mb were reported for three species in Section *Gireoudia* (Dewitte et al. 2009), to which these two species belong. According to this physical genome size estimate, 1 cM corresponds to average recombination rates of 441 and 451 Kb in the CBC and PBC backcross maps respectively. These observed recombination frequencies are somewhat lower than those seen in other plants (50 Kb/cM for *Carica papaya* (Yu et al. 2009) and 260 Kb/cM for *A. thaliana* (Singer et al. 2006)). Although recombination frequency can vary considerably across the genome and is dependent on the specific demographic history of the individual or species considered, the low recombination frequency observed in this *Begonia* mapping family holds promise for association mapping of genes underlying *Begonia* diversity in natural populations with relatively low marker density. Conversely, fine mapping of the genes will be a greater challenge requiring large mapping families, but will be offset by the rapidly decreasing costs of genotyping that makes construction of high density maps feasible.

## Conclusions

We have generated three transcriptomes and a genetic map that increase the potential of the mega diverse genus *Begonia* for evolutionary studies. We plan to use the genetic map to understand the genetic basis of speciation between this closely related species pair and to study traits which may have been key in their adaptations to very different seasonally dry tropical forest and wet rainforest habitats. The transcriptomes will prove useful in analysing future *Begonia* whole genome sequences that we are in the process of developing. With these improved sequence data, we hope to better unravel gene orthology and paralogy in order to understand the consequences of the genome duplication early in *Begonia's* evolutionary history and how it has contributed to the diversity of the genus. The sequence data we have produced that show low sequence divergence between Central American and Southeast Asian sections of the genus will also allow the generation of large numbers of genome-wide molecular genetic markers to examine the evolutionary forces generating and maintaining this diversity at the population-level for species in this genus. For example, this transcriptome sequence data has already been used to develop microsatellite markers for population genetic studies of other *Begonia* species in Mexico (Twyford et al., in prep).

The relative ease with which genetic resources can be developed from scratch has largely overcome the limitations on the choice of species which can be studied genetically. With the resources we have developed, we have added over 1,500 *Begonia* species to the lists of plants with characterised transcriptomes and molecular recombination maps in close relatives.

## Methods

### Plants

An individual each of *B. conchifolia* and *B. plebeja* in long-term cultivation at Royal Botanic Garden Edinburgh (RBGE accession numbers: 20042082 and 20051406, respectively) were cross pollinated in the glasshouse during the winter of 2007 with *B. plebeja* acting as the maternal parent and *B. conchifolia* as paternal parent. The F1 was

only partially self fertile so backcrosses to each parent were used to generate mapping families of sufficient size. F1 seeds were germinated on sterilised compost (16 Bark: 3 Peat: 1 Perlite plus finely sieved osmocote) in a warm 28°C humid environment. F1 seedlings (family number CKB137) were moved to individual 15 cm diameter pots of compost (as above) and grown in a glasshouse with supplemental lights and an average temp of 28°C. Seedlings from a backcross between CKB137.6 (maternal) and *B. conchifolia* ($n$=117) and from a backcross between CKB137.8 (maternal) and *B. plebeja* paternal ($n$=277), were sown in the spring of 2009 and plants were subsequently maintained under the same conditions as the parents and F1s.

## Genome Size Measurements

Nuclear DNA contents (C-values) were assessed by flow cytometry. Fully expanded leaf tissue from each accesion (about 1 $cm^2$) was chopped along with an internal standard (*Solanum lycopersicum* 'Stupiké polní rané' 2C=1.96 pg or *Petroselinum crispum* 'Champion Moss Curled' 2C=4.45 pg) in 2 mL of 'General purpose isolation buffer' (GPB; Loureiro et al. 2007) with 3 % PVP-40 following the one-step procedure described by Doležel et al. (2007). The nuclear suspension was then filtered through a nylon mesh (30 μm) to remove debris, stained with Propidium Iodide (Sigma-Aldrich) at a final concentration 60 $\mu g \cdot mL^{-1}$ and supplemented with 100 $mg \cdot mL^{-1}$ ribonuclease A (RNase A; Sigma-Aldrich). Samples were kept on ice for 15 min and 5,000 particles were recorded using a Partec Cyflow SL3 (Partec GmbH, Canterbury, UK) flow cytometer fitted with a 100 mW green solid state laser (Cobolt Samba, Solna, Sweden). Nine replicates per accesion were processed and the resulting flouresscence histograms were analysed with FlowMax software (Partec GmbH).

## DNA Extraction

Approximately 20 μg of silica dried inforesence or leaf was pulverized using a TissueRupter milling machine (Qiagen, Crawley, UK) for three rounds of 20 Hz for 90 s with added sand. Alternatively, fresh young leaf tissue (100 μg approx.) was ground to a powder after freezing with liquid nitrogen. Individual DNeasy Plant Mini kits (Qiagen) were used with the following changes to the supplier's instructions. The initial digestion at 65°C was extended to 30 min and two elutions (subsequently combined) were performed with 50–100 ul AE buffer each for 1–15 h to improve DNA yield. DNA was extracted twice from the parents and from ~10 % of backcross progeny to assess and control for genotyping errors.

## Phylogenetics

A phylogeny of 45 representative *Begonia* species was constructed to place *B. conchifolia*, *B. plebeja* and *B. venusta* into phylogenetic context. Sequences of the *Begonia* chloroplast *ndh*A intron, the *ndh*F-*rpl32* and *rpl32-trn*L spacers were obtained either from the National Centre for Biotechnology Information (http://www.ncbi.nlm. nih.gov/) or from whole *Begonia* chloroplast genome sequences (Harrison and Kidner in prep). Sequences were aligned using MAFFT v6.717 (Multiple Alignment using Fast Fourier Transform) (Katoh and Toh 2008) applying the iterative refinement method (FFT-NS-i) and using default parameter settings (gap opening penalty: 1.53, offset-value: 0.0) and then manually adjusted.

The three nucleotide regions were concatenated to produce an alignment of 4086 characters and analysed as an unpartitioned matrix in a Bayesian analysis, using parameters: GTR + I- optimised proportion of invariable sites and across site rate variation. The model used was the same as that used in Thomas et al. (2012). The analysis was performed in *MrBayes* v3.1.2 with 4 chains and 1,000,000 generations (Huelsenbeck and Ronquist 2001).

## RNA Extraction

Three vegetative buds were harvested from mature plants of *B. plebeja* and *B. conchifolia* (the same individuals used for mapping) and from two individuals of *B. venusta* (RBGE accession numbers 20021604 and 20021596). Buds were dissected to leave apices with three developing leaves, the longest of which was less than 5 mm long. Dissected buds were frozen in liquid nitrogen and ground for RNA extraction.

Total RNA was isolated using Invitrogen Plant RNA purification reagent. Complementary DNA (cDNA) was synthesized using the SMART cDNA protocol (Zhu et al. 2001) by the Evrogen synthesis service (Evrogen, Moscow, Russia). cDNA fragment libraries were prepared for Roche 454 Titanium sequencing and sequenced on the 454-FLX titanium platform (Roche, West Sussex, UK), followed by signal-processing and base-calling using the Roche Shotgun signal-processing software, gsRunProcessor version 2.0.01.12 (Roche) for *B. conchifolia* and *B. plebeja*, and gsRunProcessor version 2.3 (Roche) for *B. venusta*. The reads have been deposited in the Genbank Sequence Read Archive, accession number ERP001195.

## Transcriptome Analysis

Repeat elements were identified by BLASTX (Altschul et al. 1997) searching reads against the Plant Repeat Element Database (Ouyang and Buell 2004) and plastid, mitochondrial and viral sequences from the RefSeq database and

these sequences were removed before subsequent analysis. Reads were assembled using Newbler 2.5 (Roche) with options enabled for transcriptome assembly ("-cdna"), and trimming of SMART adapters ("-vt SMARTAdapters"). The "-urt" option ("use read tips") was enabled, to yield longer assemblies ("isotigs"), although this may result in less accurate calling of low coverage bases. In addition to assembled isotigs of reads, Newbler also outputs isogroup data for groups of overlapping but distinct isotigs that probably correspond to alternative transcripts for individual genes. Summary statistics for the 454 reads including numbers of bases and reads, median read length (N50), percentage GC content, percentage alignments, numbers of isogroups and isotigs, median isotigs lengths, numbers of unaligned singleton reads and percentage inferred error were calculated using a custom Perl script (available upon request).

Polymorphisms between species were identified using GSmapper (Roche) by mapping reads from one species onto assembled isotigs of the second species. High confidence SNPs between species were identified as those for which all reads from one species differed from the isotig assembly sequence, and involved at least 3 reads in each direction or at least 7 reads in the same direction.

Reciprocal BLAST searches of the assembled transcriptomes were performed using RBH Orthologs at Evopipes.net (http://evopipes.net, Barker et al. 2010). This program searches for reciprocal best hits with MEGABLAST (Zhang et al. 2000; Ma et al. 2002). Although MEGABLAST is not recommended for cross species comparisons (Korf et al. 2003), the three species are so closely related (Fig. 1d) that it should function acceptably well. This analysis identified isotigs that were either shared between two or three species or unique to one species (Fig. 2a). BLASTX was used to determine the number of contigs in each of these categories that matched sequence from the Green Plant Ref set (National Centre for Biotechnology Information 2002; Pruitt et al. 2012) with expect (e) values of less than −40. The transcriptomes for individual species were annotated with BLASTX hits at less than $1 \times e^{-40}$ against Uniref and the Green Plant reference databases. We also used Annot8r (Schmid & Blaxter, 2008) to examine the tri-scriptome alignment of all three species to identify genes and annotate gene functions. This program uses BLASTX searches against the uniprot database (The Uniprot Consortium 2011; http://www.uniprot.org/) and subsequent GO-term, EC and KEGG annotations to determine the likely function of genes. Annotation data, along with reciprocal BLASTX, was uploaded into a partigene database (Parkinson et al. 2004) available at http://genepool-blast.bio.ed.ac.uk/partigene/Begonia.

Patterns of gene duplication that can be informative about past polyploidization events were analysed using DupPipes at Evolpipes.net (www.evopipes.net, Barker et al. 2008, 2010). Briefly, the program used MEGABLAST to identify gene family members within the transcriptome with at least 40 % nucleotide similarity over 300 bp and BLASTX to identify gene matches in the Green Plant Ref set database (NCBI) with at least 30 % similarity over 150 bp. The amino acid sequences of remaining isotig families were deduced, trimmed for non-coding sequence, and transcribed back to DNA sequences using Genewise v2.2.2 (Birney et al. 2004), MUSCLE v3.6 (Edgar 2004) and vRevTrans 1.4 (Wernersson and Pedersen 2003). Synonymous substitution Ks values, a measure of neutral divergence, for each duplicate pair were estimated under the F3-4 maximum likelihood model (Goldman and Yang 1994) using PAML (Yang 1997). Finally, only node Ks values for each gene family cluster and Ks values >0.05 and >2.0 were retained, removing biases related to gene family size and potentially identical genes missed by read alignments. The natural log Ks distribution was analysed to identify peaks of duplication relative to a null hypothesis of an exponentially declining distribution due to background gene birth-death processes using mixture analysis as implemented by the mclust package in R v13 software (Fraley and Raftery 2006; R Development Core Team 2012).

## SNP Genotyping

The assembled transcriptomes for *B. plebeja* and *B. conchifolia* were interrogated for SNPs within potential developmental genes for marker development. The first set of selected genes are known to influence morphology and micro-morphology in other plant species. A second larger set of genes consisted of potential transcription factors based on the BLAST annotation of the *Begonia* transcriptomes. Transcription factors were targeted because changes in their expression frequently lead to large phenotypic effects. The final set consisted of genes with potential influences on morphology or responses to common developmental cues, such as light, were also isolated using searches of the BLAST annotation in the partigene database. To identify sequence polymorphisms for genetic mapping, isotig sequences from one parent species were BLAST searched against the transcriptome for the other parent to identify potential orthologs. Probable orthologs were identified as having BLASTN e-values of less than $e^{-100}$, deemed to be a suitably stringent similarity threshold between these closely related species. When likely orthologs were found, a reciprocal BLAST search was done to confirm orthology and to identify any other closely related sequences. If potential paralogs were detected, a locus was not considered further for SNP marker development. Aligned reads were then inspected to identify highly supported SNPs between species surrounded on each side by approximately 50–60 bp of high quality conserved sequence against which to design primers.

As a final check, the presence of introns near to the SNP was investigated by comparing the target sequence (~120 bp) to the annotated genome sequence for the closest fully sequenced relative of *Begonia*; *Cucumis sativa* (Cucurbitaceae) using BLASTX at the online plant genome resource website Phytozome v5.0 (http://www.phytozome.net/).

In total, 130 SNPs in target genes were identified using this approach. Summary data for these loci are provided in the supplementary Table 1. Primer design and genotyping was performed by KBiosciences (Hoddesdon, UK) using their proprietary competitive allele-specific PCR genotyping system (KASP). This led to 112 SNP markers being genotyped across 288 samples representing two repeats of each parent species, two F1s, 162 PBC mapping individuals, and 117 CBC1 mapping individuals. A final set of 105 SNPs were retained for further analysis following inspection of SNP genotype results for interpretable heritability in mapping families.

## AFLP Genotyping

A PCR amplification protocol for amplified fragment length polymorphisms (AFLPs) was modified from the online protocol (http://bioweb.usu.edu/wolf/aflp_protocol.htm). The complete AFLP methods and the sequences of primers are provided in supplementary information (Supplementary Note 1). In total, all 12 combinations of four fluorescent dye-labeled *Eco*RI primers and three non-labelled *Mse*I primers were genotyped across all individuals. Genotyping was performed on an ABI 3730 capillary sequencer system at the Genepool Sequencing and Bioinformatics Facility at the University of Edinburgh. Trace data were viewed and analyzed using GeneMapper v4.0 (Applied Biosystems Inc., Foster City, CA, USA) and the presence or absence of fragments between 55 bp and 500 bp scored. Samples that exhibited a summed AFLP peak height for a particular primer combination that was in the lower ~10 % of the distribution were removed. Loci that could not differentiate *B. conchifolia* and *B. plebeja* and fragments that occurred at significantly higher or lower frequencies than expected according to Mendelian inheritance (>0.90 or <0.25) were also excluded. Frequency distributions of peak heights were then examined at the position of each remaining locus (typically giving a bimodal distribution corresponding to presence or absence of the fragment), allowing a threshold intensity to be set for filtering the presence of a fragment. Filtered data was then analysed in AFLPscore v1.4b (Whitlock et al. 2008), that compared normalized peak heights between duplicated AFLP samples to estimate error rates for different filter threshold combinations of locus median peak height and sample relative peak heights. Optimal filtering procedures were chosen that typically led to estimated AFLP genotyping error rates from 2 % to 5 % per primer combination.

## Genetic Mapping

Genetic maps were constructed using the demonstration version of Joinmap v4 (Plant Research International, Wageningen, Netherlands). The populations were treated as an outcrossed mapping family (CP type) because heterozygous parental species genotypes resulted in many informative AFLP and SNP loci that with alleles that were present in approximately 3:1 proportions that could therefore not have been mapped with parameters restricted for backcross populations.

Linkage groups of loci were identified at >4 logarithm of odds (LOD) score and a maximum of 25 Kosambi centi-Morgan (cM) between neighbouring loci. Genetic maps were estimated using Joinmap's default parameters for regression mapping. Map quality was assessed by examining goodness of fit $G^2$ likelihood ratio statistics and by comparing locus sharing and order between homologous linkage groups in the two mapping families. Large initial linkage groups with poor statistical support that were not recovered from both populations were re-analysed and frequently split into two well-supported linkage groups after removal of a few markers that reduced their overall support. Genetic map data was graphed using MapChart v2.2 (Voorrips 2002). Total map length was estimated by either adding twice mean marker distance to the length of each linkage group (add2s method; Fishman et al. 2001) or by multiplying by the correction factor (marker number +1)/(marker number −1) (method 4; Chakravarti et al. 1991). The percentage of the genome present within 5 and 10 cM distance of a mapped marker, assuming markers are randomly distributed on genetic maps, was estimated according to 1-exponent ((−2*distance*marker number)/map length) according to Fishman et al. (2001). The extent of marker clustering on the genetic maps was tested using a $chi^2$ dispersion test against a null poisson distribution of evenly distributed markers separated by mean marker distance.

**Acknowledgments** Mark Hughes, Neil Watherston, Keith Gardner, Michelle Hollingsworth, Laura Forrest, at RBG Edinburgh, Andrew Hudson at University of Edinburgh, Ilia Leitch at RBG Kew, David Menzies and Ewan Donaldson at Glasgow Botanic Gardens.

**Funding was Provided by** Biological and Biotechnology Research Council (UK), M. L. MacIntyre Begonia Trust, Rural and Environmental Research and Analysis Directorate (RERAD) grant to Royal Botanic Gardens, Edinburgh. ACB was supported during the writing of this study by funding from FP7-REGPOT 2010-1, Grant No. 264125 EcoGenes.

## References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new

generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol 25:2445–2455

Barker MS, Dlugosch KM, Dinh L, Challa RS, Kane NC, Rieseberg LH (2010) EvoPipes.net: bioinformatic tools for ecological and evolutionary genomics. Evol Bioinformatics Online 6:143–149

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol 7:732–736

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995

Burt Utley K (1985) A revision of Central American species of *Begonia* section *Gireoudia* (Begoniaceae). Tulane Stud Zool Bot 25:1–131

Chakravarti A, Lasher LK, Reefer JE (1991) A maximum likelihood method for estimating genome length using genetic linkage data. Genetics 128:175–182

Choi W, Baek D, Oh DH, Park J, Hong H, Kim WY, Bohnert HJ, Bressan RA, Park HC, Yun DJ (2011) NKS1, Na(+)- and K(+)-sensitive 1, regulates ion homeostasis in an SOS-independent pathway in *Arabidopsis*. Phytochemistry 72:330–336

Day IS, Reddy AS (1998) Isolation and characterization of two cyclin-like cDNAs from *Arabidopsis*. Plant Mol Biol 36:451–461

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 Biodiversity Book 2. InTech p. 27–52

Doležel J, Bartoš J, Voglmayr H, Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. Cytometry 51:127–128

Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nat Protoc 2:2233–2244

Drummond CS, Eastwood RJ, Miotto STS, Hughes CE (2012) Multiple continental radiations and correlates of diversification in *Lupinus* (Leguminosae): testing for key innovation with incomplete taxon sampling. Syst Biol 63:443–460

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107:1–15

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends Ecol Evol 26:298–306

Fishman L, Kelly AJ, Morgan E, Willis JH (2001) A genetic map in the *Mimulus guttatus* species complex reveals transmission ratio distortion due to heterospecific interactions. Genetics 159:1701–1716

Fraley C, Raftery AE (2006) MCLUST version 3 for R: normal mixture modeling and model-based clustering, technical report no. 504, Department of Statistics, University of Washington

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein- coding DNA sequences. Mol Biol Evol 11:725–736

Goodall-Copestake WP, Harris DJ, Hollingsworth PM (2009) The origin of a mega- diverse genus: dating *Begonia* (Begoniaceae) using alternative datasets, calibrations and relaxed clock methods. Bot J Linn Soc 159:363–380

Greilhuber J, Dolezel J, Lysák MA, Bennett MD (2005) The origin, evolution and proposed stabilization of the terms 'genome size' and 'C-value' to describe nuclear DNA contents. Ann Bot 95:255–260

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y et al (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755

Hughes M, Hollingsworth PM (2008) Population genetic divergence corresponds with species-level biodiversity patterns in the large genus *Begonia*. Mol Ecol 17:2643–2651

Hughes M, Hollingsworth P, Miller A (2003) Population genetic structure in the endemic *Begonia* of the Socotra archipelago. Biol Conserv 113:277–284

Hvoslef-Eide AK, Munster C (2007) *Begonia*. History and breeding. In: Anderson NO (ed) Flower breeding and genetics. Springer, pp. 241–275

Ishiguro S, Watanabe Y, Ito N, Nonaka H, Takeda N, Sakai T, Kanaya H, Okada K (2002) SHEPHERD is the *Arabidopsis* GRP94 responsible for the formation of functional CLAVATA proteins. EMBO J 21:898–908

Kadota Y, Shirasu K (2012) The HSP90 complex of plants. Biochim Biophys Acta 1823:689–697

Kaeuffer R, Peichel CL, Bolnick DI, Hendry AP (2012) Parallel and nonparallel aspects of ecological, phenotypic, and genetic divergence across replicate population pairs of lake and stream stickleback. Evolution 66:402–418

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–298

Kiew R (2005) Begonias of peninsular Malaysia. Natural History Publications (Borneo) in association with Singapore Botanic Gardens National Parks Board

Kishimoto S, Aida R, Shibata M (2002) *Agrobacterium tumefaciens*-mediated transformation of Elatior *Begonia* (*Begonia x* hiemalis Fotsch). Plant Sci 162:697–703

Korf I, Yandell M, Bedell J (2003) BLAST. O'Reilly Media

Legro RAH, Doorenbos J (1971) Chromosome numbers in *Begonia*: 2. Neth J Agric Sci 19:176–183

Loureiro J, Rodriguez E, Doležel J, Santos C (2007) Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. Ann Bot 100:875–888

Ma B, Tromp J, Li M (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics 18:440–445

Matolweni L, Balkwill K, McLellan T (2000) Genetic diversity and gene flow in the morphologically variable, rare endemics *Begonia dregei* and *Begonia homonyma* (Begoniaceae). Am J Bot 87:431–439

Melo M, Warren BH, Jones PJ (2011) Rapid parallel evolution of aberrant traits in the diversification of the Gulf of Guinea white-eyes (Aves, Zosteropidae). Mol Ecol 20:4953–4967

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hotspots for conservation priorities. Nature 403:853–858

Nakamura N, Huang CJ, Rubite RR, Leong WC, Kono Y, Yang HA, Peng CI (2012) Isolation of compound microsatellite markers in *Begonia fenicis* (Begoniaceae) endemic to East and Southeast Asian islands. Am J Bot 99:e20–e23

National Centre for Biotechnology Information (2002) Chapter 18; the reference sequence (RefSeq) project In: The NCBI handbook. National Library of Medicine, Bethesda, MD

Neale S, Goodall-Copespeak W, Kidner C (2006) The evolution of diversity in *Begonia*. In: Teixeira da Silva JA (ed) Floriculture, ornamental and plant biotechnology: advances and topical issues. Global Science Books

Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet 34:401–437

Ouyang S, Buell CR (2004) The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:D360–D363

Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M (2004) PartiGene- constructing partial genomes. Bioinformatics 20:1398–1404

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. PNAS 101:9903–9908

Petit C, Thompson JD (1999) Species diversity and ecological range in relation to ploidy level in the flora of the Pyrenees. Evol Ecol 13:45–66

Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130–D135

R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Sangster TA, Salathia N, Undurraga S, Milo R, Schellenberg K, Lindquist S, Queitsch C (2008) HSP90 affects the expression of genetic variation and developmental stability in quantitative traits. PNAS 105:2963–2968

Schaefer H, Heibl C, Renner SS (2009) Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. Proc Biol Sci 276:843–851

Schranz ME, Mitchell-Olds T (2006) Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell 18:1152–1165

Singer T, Fan YP, Chang HS, Zhu T, Hazen S, Briggs S (2006) A high-resolution map of *Arabidopsis* recombinant inbred lines by whole-genome exon array hybridization. PLoS Genet 2:1352–1361

Schmid R, Blaxter ML (2008) Annotator: GO, EC and KEGG annotation of EST datasets. BMC Bioformatics 9:180

Song H, Zhao R, Fan P, Wang X, Chen X, Li Y (2009) Overexpression of *AtHsp90.2*, *AtHsp90.5* and *AtHsp90.7* in *Arabidopsis thaliana* enhances plant sensitivity to salt and drought stresses. Planta 229:955–964

Sterck L, Rombauts S, Vandepoele K, Rouzé P, Van de Peer Y (2007) How many genes are there in plants (… and why are they there)? Curr Opin Plant Biol 10:199–203

Strecker U, Hausdorf B, Wilkens H (2012) Parallel speciation in Astyanax cave fish (Teleostei) in Northern Mexico. Mol Phylogenet Evol 62:62–70

Tebbit M (2005) Begonias: cultivation, identification and natural history. Timber Press

The UniProt Consortium (2011) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39:D214–D219

Thomas DC, Hughes M, Phutthai T, Rajbhandary S, Rubite R, Ardi WH, Richardson JE (2011) A non-coding plastid DNA phylogeny of Asian *Begonia* (Begoniaceae): evidence for morphological homoplasy and sectional polyphyly. Mol Phylogenet Evol 60:428–444

Thomas DC, Hughes M, Phutthai T, Ardi WH, Rajbhandary S, Rubite R, Twyford AD, Richardson JE (2012) West to east dispersal and subsequent rapid diversification of the mega-diverse genus *Begonia* (Begoniaceae) in the Malesian archipelago. J Biogeogr 39:98–113

Vamosi JC, Dickinson TA (2006) Polyploidy and diversification—a phylogenetic investigation in Rosaceae. Int J Plant Sci 167:349–358

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in *Arabidopsis*. Science 290:2114–2117

Voorrips RE (2002) MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93:77–78

Wang G, Ellendorff U, Kemp B, Mansfield JW, Forsyth A, Mitchell K, Bastas K, Liu CM, Woods-Tör A, Zipfel C, de Wit PJ, Jones JD, Tör M, Thomma BP (2008) A genome- wide functional investigation into the roles of receptor-like proteins in *Arabidopsis*. Plant Physiol 147:503–517

Wernersson R, Pedersen AG (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res 31:3537–3539

Wheat CW (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138:1–19

Whitehead A, Pilcher W, Champlin D, Nacci D (2012) Common mechanism underlies repeated evolution of extreme pollution tolerance. Proc Biol Sci 279:427–433

Whitlock R, Hipperson H, Mannarelli M, Butlin RK, Burke T (2008) An objective, rapid and reproducible method for scoring AFLP peak-height data that minimizes genotyping error. Mol Ecol Resour 8:725–735

Wood TE, Burke JM, Rieseberg LH (2005) Parallel genotypic adaptation: when evolution repeats itself. Genetica 123:157–170

Woodard SH, Fischman BJ, Venkat A, Hudson M, Varala K, Cameron SA, Clark AG, Robinson GE (2011) Genes involved in convergent evolution of eusociality in bees. PNAS USA 108:7472–7477

Xu Q-l, Dong J-l, Gao N, Ruan M-Y, Jia H-Y, Zhang L, Want C-Y (2011) Transgenic lines of *Begonia maculata* generated by ectopic expression of *PttKN1*. Biologia 66:251–257

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yu QY, Tong E, Skelton RL, Bowers J (2009) A physical map of the papaya genome with integrated genetic map and genome sequence. BMC Genomics 10:371

Zamocky M, Furtmüller PG, Obinger C (2008) Evolution of catalases from bacteria to humans. Antioxid Redox Signal 10:1527–1548

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214

Zhang W-W, Pan J-S, He H-L, Zhang C, Li Z, Zhao J-L, Yuan X-J, Zhu L-H, Huang S-W, Cai R (2012) Construction of a high density integrated genetic map for cucumber (*Cucumis sativus* L.). Theor Appl Genet 124:249–259

Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: a SMART (TM) approach for full-length cDNA library construction. Biotechniques 30:892–897