Impatiens glandulifera (Himalayan balsam) chloroplast genome sequence as a promising target for populations studies

Giovanni Cafa[1], Riccardo Baroncelli[2], Carol A. Ellison[1] and Daisuke Kurose[1]

[1] CABI Europe, Egham, Surrey, UK
[2] University of Salamanca, Instituto Hispano-Luso de Investigaciones Agrarias (CIALE), Villamayor (Salamanca), Spain

ABSTRACT

Background: Himalayan balsam *Impatiens glandulifera* Royle (Balsaminaceae) is a highly invasive annual species native of the Himalayas. Biocontrol of the plant using the rust fungus *Puccinia komarovii* var. *glanduliferae* is currently being implemented, but issues have arisen with matching UK weed genotypes with compatible strains of the pathogen. To support successful biocontrol, a better understanding of the host weed population, including potential sources of introductions, of Himalayan balsam is required.

Methods: In this molecular study, two new complete chloroplast (cp) genomes of *I. glandulifera* were obtained with low coverage whole genome sequencing (genome skimming). A 125-year-old herbarium specimen (HB92) collected from the native range was sequenced and assembled and compared with a 2-year-old specimen from UK field plants (HB10).

Results: The complete cp genomes were double-stranded molecules of 152,260 bp (HB92) and 152,203 bp (HB10) in length and showed 97 variable sites: 27 intragenic and 70 intergenic. The two genomes were aligned and mapped with two closely related genomes used as references. Genome skimming generates complete organellar genomes with limited technical and financial efforts and produces large datasets compared to multi-locus sequence typing. This study demonstrates the suitability of genome skimming for generating complete cp genomes of historic herbarium material. It also shows that complete cp genomes are solid genetic markers for population studies that could be linked to plant evolution and aid with targeting native range and natural enemy surveys for biocontrol of invasive species.

Subjects Genomics, Molecular Biology, Plant Science
Keywords Balsaminaceae, Chloroplast genome, *Impatiens glandulifera*, Whole genome sequencing, Genome skimming, Phylogenetic analyses

INTRODUCTION

Impatiens glandulifera Royle (Balsaminaceae), commonly known as Himalayan balsam, is an annual plant species native to the foothill of the Indian and Pakistani Himalayas, where it can be found in mixed plant communities at altitudes from 2,000 to 3,000 m. It was introduced into the UK in 1839 as an ornamental plant (*Beerling & Perrins, 1993*)

Submitted 26 July 2019
Accepted 12 February 2020
Published 24 March 2020

Corresponding author
Giovanni Cafa, G.Cafa@cabi.org

Academic editor
Andrea Case

Additional Information and Declarations can be found on page 10

DOI **10.7717/peerj.8739**

and is now considered to be one of the worst invasive, non-native plant species in the UK particularly in riverine habitats (*Environment Agency, 2010*), although it also flourishes in damp woodlands and waste grounds (*Tanner et al., 2014*). Himalayan balsam is the tallest annual plant in Europe, attaining a height of 2.5 m, significantly taller than seen in its native range. Each plant can produce up to 2,500 seeds that are forcibly ejected up to 7 m away from the parent plant and able to survive for up to 2 years in the soil seed bank. Seeds which land in rivers are transported downstream to new sites, establishing particularly well in disturbed habitats (*Tanner et al., 2014*; *Tanner, 2017*). It forms dense monocultures and outcompetes native plant species (*Perrins, Fitter & Williamson, 1993*; *Hulme & Bremner, 2006*; *Pattison et al., 2016*); this is aided by seed germination early in the growing season, its extremely fast growth rate and potential allelopathic effect on soil mycobiota (*Lobstein et al., 2001*; *Clements et al., 2008*; *Tanner & Gange, 2013*). The showy flowers rich in nectar lure pollinators away from native plants, reducing pollination opportunities (*Beans & Roach, 2015*). As the plant dies back in the autumn in riverine habitats, banks laid bare have an increased risk of erosion and dead plant material entering the water course can cause blockages leading to flooding. Himalayan balsam stands have also been shown to alter invertebrate communities and negatively affect detritivores, which are reliant on native plant species (*Tanner et al., 2013*).

Biological control, using a rust fungus (*Puccinia komarovii* var. *glanduliferae*) collected from the native range, is currently being implemented in the UK, but issues have arisen with matching the genotype of the weed with a compatible strain of the rust fungus (*Tanner et al., 2015*; *Varia, Pollard & Ellison, 2016*). In order to enhance the likelihood of successful biological control, molecular studies are carried out to better understand the host weed population and its relationship with pathogen genotypes in both the native and introduced ranges of a weed (*Gaskin et al., 2011*). In the introduced range, it is important to ascertain how genetically diverse the invasive population is. A single genetic entity of a natural enemy may be unilaterally effective if the invasive weed population is genetically uniform (*Tomley & Evans, 2004*). However, an invasive plant is often introduced on more than one occasion and from different parts of the native range, resulting in increased genetic diversity in the introduced range. This information is particularly important with co-evolved biotrophic pathogens, such as rusts, because they often express intraspecies specificity (*Evans, 2002*; *Ellison, Evans & Ineson, 2004*). Hence, more than one strain of a pathogen sourced from the area where the introduction originated may be required.

Previous studies (*Kurose, Pollard & Ellison, 2018*) sampled leaves from across the native range in India and Pakistan and the introduced range in the UK and Ireland with specimens dating from 1881 to 2016. Six regions of the cp genome were included in this study with a multi-locus sequence typing (MLST) approach, which represented a limited portion of the cp genome of around 2.5% of the complete sequence of ~150 Kb genome. These were *trnL-F*, *atpB-rbcL*, *rps16* Intron, *trnG* Intron, *psbA-trnH*$^{(GUG)}$ and *rpl32-trnL*$^{(UAG)}$ (*Kurose, Pollard & Ellison, 2018*). The study showed that the investigated

specimens segregated into three groups, suggesting that Himalayan balsam in the UK was introduced at least three times from the native range. However, one of the biotypes found in the UK has not been matched in the native range and therefore additional molecular testing is needed for further investigation. Studies targeting complete cp genomes might provide better resolution as they can examine additional genes beyond the six previously investigated by MLST. Surveys targeting the three regions identified (Pakistan, Indian Kashmir and Nepalese border region with India) could provide rust fungus isolates more pathogenic to Himalayan balsam genotypes invasive in the UK. In addition, sequencing of complete cp genomes from historic herbarium material could allow the comparison with recently sampled specimens and gain knowledge to establish the origin of invasive alien weed populations. Herbarium collections have already shed new light on the mechanisms shaping species adaptation and diversification and biodiversity patterns over space and time (*Besnard et al., 2018*). A specimen stored in an herbarium is an archive document, which proves species distribution in a certain territory and habitat at a certain time (*Jukonienė, Subkaitė & Ričkienė, 2019*). The use of historic herbarium specimens can inform about long-term effects on plants of at least four of the main drivers of global change: pollution, habitat change, climate change and invasive species (*Lang et al., 2019*).

Whole organellar genomes such as the cp genome can be obtained by genome skimming. *Straub et al. (2012)* first showed that "genome skimming" is suitable for the recovery of highly repetitive genome regions such as the ribosomal cluster (rRNA) or organelle genomes and the generation of entire genome data at a relatively low sequence depth. In addition, genome skimming has been used to recover plastid DNA and rRNA from up to 83-year-old herbarium specimens (*Zeng et al., 2018*). *Li et al. (2018)* published the first complete cp genomes from the genus *Impatiens* and the monotypic genus *Hydrocera*, the only two genera within the family Balsaminaceae. The study provides the genomic data that clearly positions the Balsaminaceae as sister to the rest of the families of the order Ericales. Chloroplast (cp) genomes of Balsaminaceae are of ~150 Kb and include 110–130 distinct genes, with approximately 80 Protein-coding genes (PCG), 30 tRNA and 4 rRNA genes, consistent with those of other angiosperms (*Jansen & Ruhlman, 2012*). Most of the genetic regions regularly examined for the phylogeny of *Impatiens* with MLST (*Yu et al., 2015*; *Kurose, Pollard & Ellison, 2018*) are in the cp genome, such as the regions *trnL-F*, *atpB-rbcL*, *rps16* Intron, *trnG* Intron, *psbA-trnH*$^{(GUG)}$ and *rpl32-trnL*$^{(UAG)}$.

The objectives of this study were to test low coverage whole genome sequencing (genome skimming) and assembly protocols for the generation of complete cp genomes to identify invasion sources of *I. glandulifera*. Specifically, they were to: (i) determine the possibility of successfully generating a complete cp genome of a 125-year-old *I. glandulifera* herbarium specimen by genome skimming; (ii) determine the complete sequence of the cp genomes of an introduced and a native specimen of *I. glandulifera*; (iii) document pairwise differences between the complete cp genomes of an introduced and a native specimen of *I. glandulifera*.

MATERIALS AND METHODS

DNA isolation and template preparation

DNA was isolated from fresh and herbarium leaves of *I. glandulifera* (Himalayan Balsam). Himalayan Balsam leaves of specimen HB10 were obtained in 2016 from Silwood park, Berkshire, UK (Latitude 51.40756, Longitude −0.64374), dried in silica gel and kept at 4 °C. The other sample HB92, collected from Liddar Valley, Jammu and Kashmir, India (Latitude 33.9131, Longitude 76.4797) in 1893, was obtained from the herbarium of the Natural History Museum, UK (Catalogue Number BM001254006). DNA was isolated from HB10 with DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) and from HB92 with Power Plant Pro DNA Isolation Kit (MO BIO Laboratories Inc., West Carlsbad, CA, USA). DNA extracts were quantified with Qubit 3.0 Fluorometer (Thermofisher Scientific, Loughborough, UK).

Illumina library preparation

Genomic DNA was fragmented after quantification and quality assessment for library preparation. A total of 100 ng of DNA were diluted in 50 µL of 10 mM TRIS–HCl pH 8.5—Ethanol 20% (v/v) solution (Sigma–Aldrich Merck, Gillingham, UK) and sheared with Covaris M220 Focused-ultrasonicator (Covaris Ltd., Brighton, UK) with the following settings for a 350 bp insert size: duty factor (%) 20.0, peak power 50.0, cycles/Burst 200 and duration of 65 s at 20 °C. Libraries were prepared with a Truseq Nano DNA Library prep kit (Illumina, Cambridge, UK), according to the manufacturers' instruction. Steps included fragmentation, repair ends and library size selection, addition of adenylate 3′ ends, ligation of adapters, enrichment of DNA Fragments, normalization and pool. The paired end libraries were generated with the Illlumina MiSeq at CABI (Egham, UK), with an average insert size of 350 bp and run on an Illumina MiSeq 250-PE V2 Cartridge (500 cycles) (Illumina, Cambridge, UK).

Cp genome sequence assembly and annotation

High quality reads were filtered from Illumina raw reads with the MiSeq Reporter analysis software v3.0 (Illumina, Cambridge, UK). Assembly was performed with SPAdes v3.11.1 (*Bankevich et al., 2012*). Scaffolds belonging to the cp genome were identified by BLAST searches to a custom database built with the cp genome of *Impatiens pinfanensis* (MG162586). Also, an assisted assembly was performed on *Impatiens pinfanensis* (MG162586); raw reads were mapped to the reference genome using Bowtie2 v2.3.5.1 (*Langmead & Salzberg, 2012*), the process was then repeated on the obtained scaffolds in order to fill the gaps. The final fragment was circularized using Geneious v.2019.0.3 (Biomatters Ltd., Auckland, New Zealand). The annotations of the complete cp genomes were performed with GeSeq (*Tillich et al., 2017*). The circular cp genome maps were generated with the online tool OGDraw v1.2 (*Lohse, Drechsel & Bock, 2007*) with default settings. The complete cp genome sequences were submitted to GenBank database, accession numbers *I. glandulifera* HB10 (MK358446) and *I. glandulifera* HB92 (MK358447). The two reference genomes of *I. pinfanensis* (MG162586) and

Hydrocera triflora (MG162585) (*Li et al., 2018*) represent the two genera of the Balsaminaceae family.

Whole genome alignment was performed on annotated merged contigs with progressive MAUVE 2.4.0 (*Darling, Mau & Perna, 2010*) and rearrangements were visualized using Geneious v.2019.0.3 (Biomatters Ltd., Auckland, New Zealand).

Phylogenetics

Nucleotide sequences produced in this work as well as references previously published (*Fujihashi, Akiyama & Ohba, 2002*; *Lim et al., 2014*; *Aust, Ahrendsen & Kellar, 2015*; *Li et al., 2018*) were aligned using MAFFT v.7.407 (*Katoh & Standley, 2013*), manually checked and if necessary, trimmed on both ends to have comparable nucleotides. Sequence alignments (1,262 ± 0 bp) were exported to MEGA7 (*Kumar, Stecher & Tamura, 2016*) and the best-fit substitution model using Maximum Likelihood statistical method was calculated (T92 + G + I) and used in MrBayes v.3.2.6 (*Ronquist & Huelsenbeck, 2003*). The Markov chain Monte Carlo (MCMC) algorithm was performed to generate phylogenetic trees with Bayesian posterior probabilities. Four MCMC chains were run simultaneously for random trees for 5,000,000 generations. Samples were taken every 1,000 generations. The first 25% of trees were discarded as burn-in phase of each analysis and posterior probabilities were determined from the remaining trees. A second phylogenetic tree was built using Fast Tree v.2.1.5 (*Price, Dehal & Arkin, 2010*) as implemented in Geneious v.2019.0.3 (Biomatters Ltd., Auckland, New Zealand). The trees obtained showed the same topology and therefore only the Bayesian inference phylogenetic tree is shown in results.

RESULTS

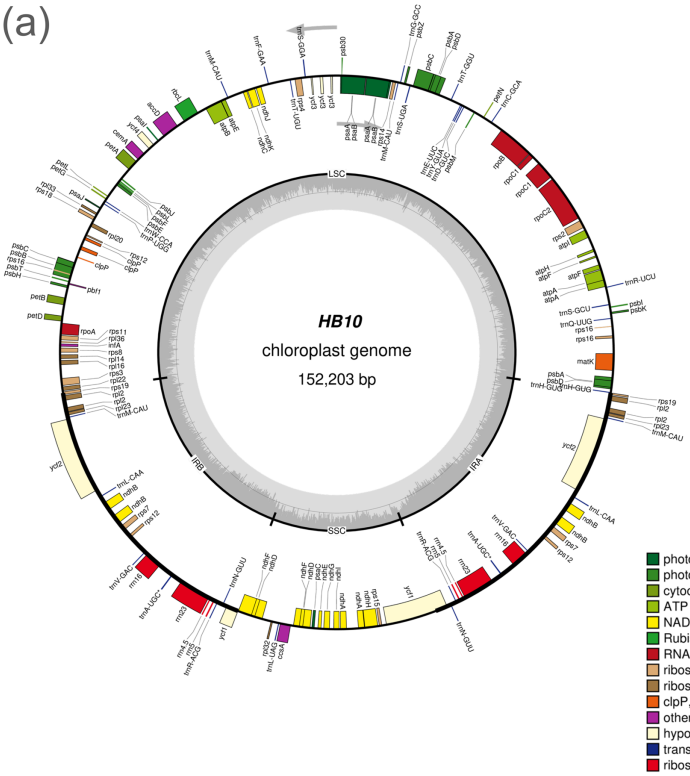
DNA isolation and sequencing of the complete chloroplast (cp) genome of the 125-year-old herbarium specimen of *I. glandulifera*.

DNA yield was of 33.0 ng/µL for *I. glandulifera* HB92 (native herbarium specimen) as compared to the fresh leaves of *I. glandulifera* HB10 (introduced specimen) with 99.2 ng/µL. The number of raw DNA reads after sequencing with an Illumina MiSeq was of 1,104,294 for HB92 and of 1,696,776 for HB10, with the quality control showing no differences between the libraries of HB92 and HB10. On average, lower coverage was obtained for HB92 than HB10, with an average coverage of the nodes including fragments of the cp genome of 15X and of 45X, respectively.

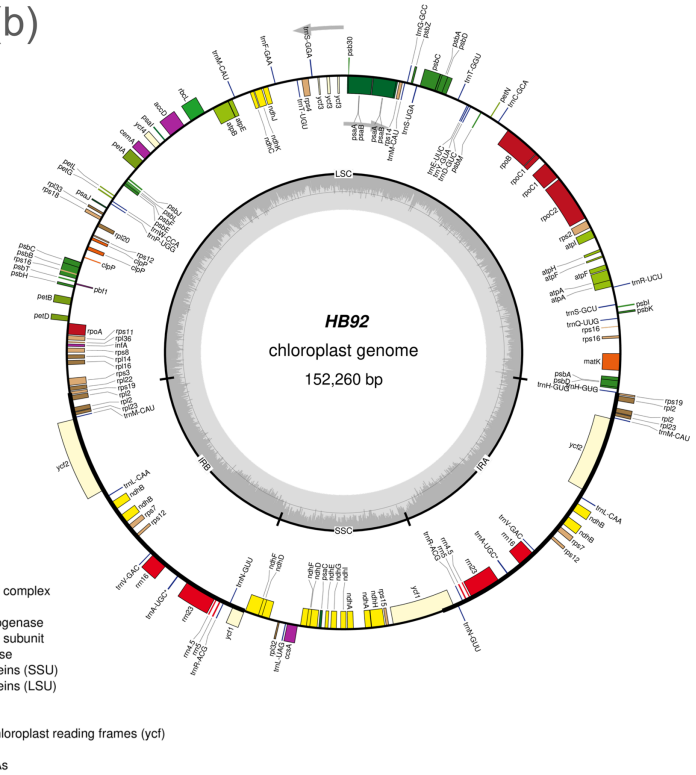
The *I. glandulifera* complete cp genome structure and gene content

Unbalanced distribution of coverage was generated by the assembler software in the scaffolds of the cp genome. A reference assisted approach was necessary to obtain the complete cp genome. The cp genomes of the two specimens were fragmented in three scaffolds, with one of the three having double value of average coverage. For example, HB10 had the three scaffolds of 83.2 Kb, 18.7 Kb and 17.7 Kb with an average coverage of 44X, 94X and 46X, respectively. The final cp genomes could be represented as circular molecules and were of 152,203 bp for the specimen HB10 of *I. glandulifera* (the introduced

(a)



(b)



(c)

Isolate	Fragments	Length (bp)	%GC	Base Content				Accession No			
				A	T	C	G	Cp Genomes	BioProject	BioSample	SRA
HB10	1	152,203	36.80	47,446	28,514	27,561	48,682	MK358446	PRJNA512444	SAMN10668721	SRR8393251
HB92	1	152,260	36.80	47,465	28,518	27,560	48,717	MK358447	PRJNA512445	SAMN10668723	SRR8393261

Figure 1 Gene maps of *Impatiens glandulifera* chloroplast (cp) genome. Genes drawn inside of the circle are transcribed counterclockwise, while genes outside the circle are transcribed clockwise. Gene functional groups are color-coded. The dark gray area in the inner circle corresponds to GC content while the light gray corresponds to the adenine-thymine (AT) content of the genome: (A) isolate HB10; (B) isolate HB92; (C) summary of sequencing and GenBank accession data. Full-size DOI: 10.7717/peerj.8739/fig-1

specimen) and of 152,260 bp for the specimen HB92 (the native herbarium specimen) (Fig. 1). The genomes contained 86 protein-coding genes (PCG), 31 transfer RNA and four ribosomal RNA genes.

The *I. glandulifera* cp genomes of the native and introduced specimens

A preliminary phylogenetic investigation was carried out on the assembled genomes by extracting the *rbcL* genes identified by BLASTn and comparing those with publicly available data. The *rbcL* gene is a key marker universally employed in MLST. BLAST analyses of the gene allowed the identification of two publicly available complete cp genomes closely related to *I. glandulifera*. These two genomes were of *I. pinfanensis* (MG162586) and of *H. triflora* (MG162585) (*Li et al., 2018*). The analysis of the key MLST

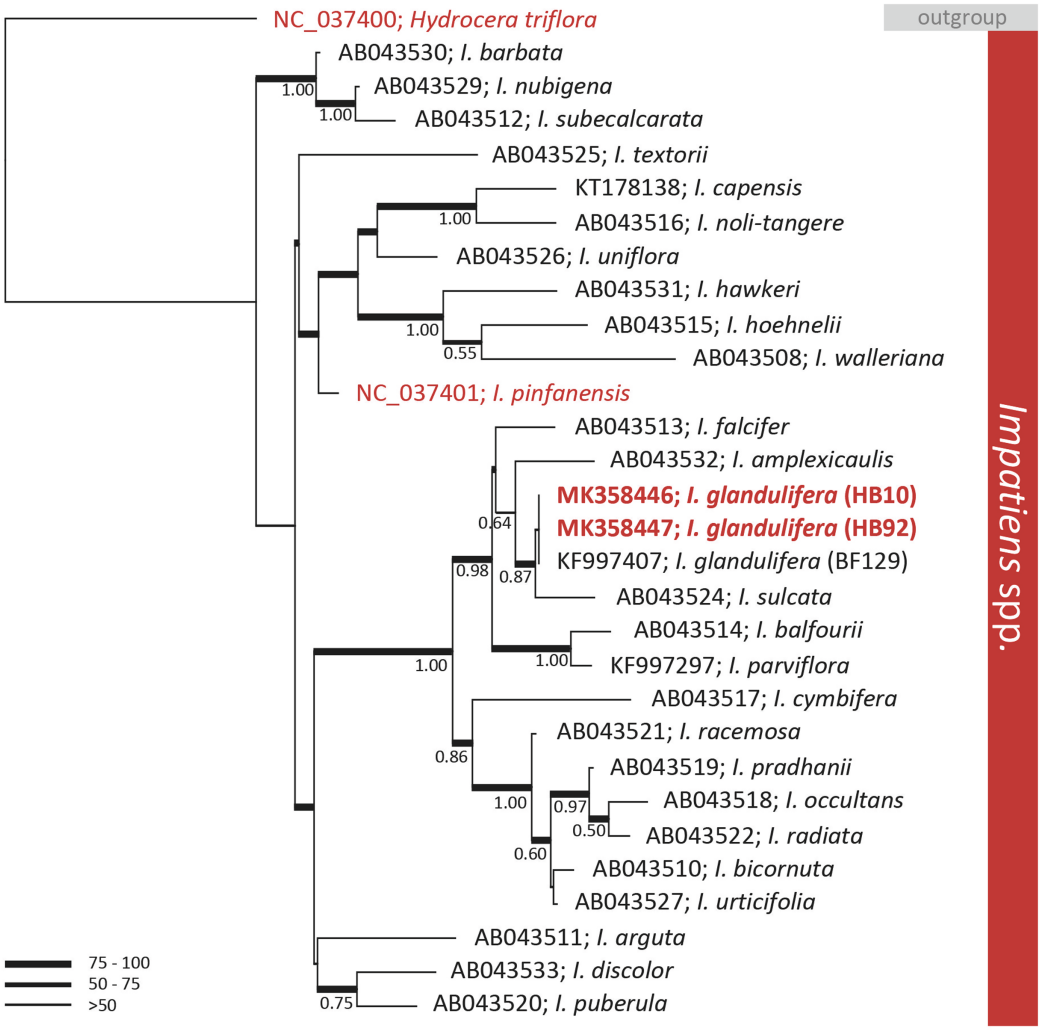


Figure 2 Phylogenetic tree of the genus *Impatiens*. Bayesian inference phylogenetic tree reconstructed from the alignment of 29 ribulose-1, 5-bisphosphate carboxylase large subunit (*rbcL*) sequences belonging to the *Impatiens* genus (*Fujihashi, Akiyama & Ohba, 2002*; *Lim et al., 2014*; *Aust, Ahrendsen & Kellar, 2015*; *Li et al., 2018*). The tree was rooted with the closely related species *Hydrocera triflora*. ML bootstrap values are represented with branch thickness while Bayesian posterior probabilities (BPP) values (above 0.50) are reported next to the node. Reference sequences from GenBank and taxonomic designation are reported. In red are highlighted sequences retrieved by full chloroplast genomes while in bold those produced in this work. Full-size DOI: 10.7717/peerj.8739/fig-2

gene *rbcL* did not discriminate among geographic isolates of *I. glandulifera*. The *rbcL* sequence of the native and introduced to the UK specimens were identical (Fig. 2). However, the gene *rbcL* did differentiate *I. glandulifera* and *I. pinfanensis*, where 1.1% of the bases were different (14/1,262 bp). The overall diversity of *rbcL* among *I. glandulifera*, *I. pinfanensis* and *H. triflora* was of 2.9% of the bases (37/1,262 bp).

Similarly to the analysis of *rbcL*, complete cp genomes were most similar between the two *I. glandulifera* genotypes and least similar between genera. However, unlike *rbcL*, the cp genomes of HB10 and HB92 were not identical. A small percentage of the bases (206/152,301 bp-0.14%) were variable between the isolates HB10 and HB92, while 4.85%

Table 1 Pairwise similarity of *Impatiens glandulifera* cp genomes HB10 (Silwood park, Berkshire, UK), HB92 (LiddarValley, Jammu and Kashmir, India), and references *I. pinfanensis* and *Hydrocera triflora*.

Pairwise similarity (%)	*I. glandulifera* HB10	*I. glandulifera* HB92	*I. pinfanensis*	*H. triflora*
I. glandulifera HB10	100.00	99.87	95.21	92.64
I. glandulifera HB92		100.00	95.24	92.68
I. pinfanensis			100.00	92.58
H. triflora				100.00

of the bases (7,480/154,134 bp) were variable between the two *I. glandulifera* (HB10 and HB92) as compared to *I. pinfanensis*. Finally, the diversity among the two of *I. glanfulifera* HB10 and HB92, *I. pinfanensis* and *H. triflora* was of 9.54% of the bases (14,958/156,779 bp). Pairwise similarity (Table 1) of the complete cp genomes HB10 and HB92 showed 206 differences, which included multiple substitutions, insertions and deletions for a total of 97 variable sites 27 sites differed within 14 genes (Table S1), while 70 sites differed in intergenic regions (Table S2). Intragenic variability of key genes of MLST genes was further investigated. For example, no differences were found for *rbcL*, while the gene *matK* was different for the two specimens with a substitution of C in HB10 to T in HB92 in position 2,600 of the aligned genomes (Table S1).

Additional synteny analysis with the Mauve multiple-genome alignment was undertaken and showed that the four genomes had the same structure and organization (Fig. S1).

DISCUSSION

This study demonstrates the success of low coverage whole genome sequencing (genome skimming) for generating complete chloroplast (cp) genomes of historic herbarium material and the suitability of the method to document differences between the complete cp genomes of the 125-year-old herbarium specimen with an introduced specimen of *I. glandulifera*. Historic herbarium specimens contain a large repository of historical and geographical information (*Crawford & Hoagland, 2009*) and contribute to understanding invasion dynamics and developing management strategies (*Lang et al., 2019*). The fight against invasive species can benefit from molecular studies targeting host weed populations (*Gaskin et al., 2011*) and support successful biocontrol with a better understanding of the sources of introductions. The typical maternal inheritance of cpDNA in angiosperms help to retain any genetic structure that may have originated with introductions, when compared to nuclear markers that are subject to gene flow via both pollen and seeds (*Gaskin, Zhang & Bon, 2005*). In this study, the complete cp genome of a 125-year-old herbarium specimen from the native range of *I. glandulifera* was successfully sequenced and used to document pairwise differences with the complete cp genome of an introduced specimen of *I. glandulifera*. The two complete cp genomes were double-stranded molecules of 152,260 bp (HB92, native herbarium specimen) and 152,203 bp (HB10, introduced specimen) in length and when aligned and mapped with two closely related genomes used as references, showed a total variability of 97 sites.

A comparison of genome skimming against MLST was carried out. MLST entails large amount of PCR and Sanger sequencing efforts to characterize limited portions of cp genomes. Specifically, *Kurose, Pollard & Ellison (2018)* investigated a portion of the cp genome of around 2.5% of the complete sequence of ~150 Kb and showed that 1.29% of the bases (50/3,865 bp) were variable among 52 samples of Himalayan balsam collected from both the native and introduced range. A total of 32 variable sites were found among them. In the same specimens investigated in this study, HB10 and HB92, six characters differed over a total of 3,852 bp, specifically, one in *rps16* Intron, two in *trnG* Intron and three in *rpl32-trnL*$^{(UAG)}$ (*Kurose, Pollard & Ellison, 2018*). Genome skimming generates complete cp genomes and reveals significantly more variability than MLST with limited technical and financial investment. We estimate that, in this study, the number of informative sites per unit of information was more than 16 times higher for genome skimming than MLST and that the cost per unit of information was more than 100 times cheaper for genome skimming than MLST. The former generates entire genomes, while the latter captures genes or fractions of genes.

Better information on the variability in invasive plant genotypes could improve the use of natural enemies such as the rust fungus *Puccinia komarovii* var. *glanduliferae*, which has great potential as a biocontrol agent of Himalayan balsam (*Gaskin et al., 2011*; *Tanner et al., 2015*). Molecular studies targeting the host weed population in the UK could benefit from additional investigation of variability, as variation in the susceptibility of UK populations to the strains of the rust that have been released is reducing the potential efficacy of the biological control strategy. Fast and cheap approaches such as MALDI-TOF MS can be used for the identification of *I. glandulifera* and closely related *Impatiens* spp. to match the optimal rust based biological control agent (*Reeve, Pollard & Kurose, 2018*). However, the latter approach is described for glasshouse-grown fresh plants and may not be suited for dry specimens.

In this work, we successfully sequenced, assembled and annotated, the complete cp genomes of a 125-year-old herbarium specimen as well as a relatively recently dried leaf specimen from a UK population. The method showed the ability to successfully sequence complete cp genomes from an historic herbarium specimen and facilitated the direct comparison with a recently sampled specimen. Comparison of complete cp genomes reveal more variable sites than a classical MLST approach and provide a platform for further molecular studies targeting the host weed populations of *I. glandulifera*. Phylogenetics can complement the investigation by enabling comparison of multiple cp genomes and therefore we encourage the use of this approach so that large numbers of complete cp genomes can become publicly available and contribute to increased knowledge of the genetic variation of invasive species at the population level.

A better resolution of the origins of invasive plant populations has implications for the selection of co-evolved and virulent pathogens (*Goolsby et al., 2006*; *Prentis et al., 2009*). As cp genomes are effective genetic markers (*Gaskin, Zhang & Bon, 2005*), the investigation of their variability impacts biocontrol strategies that can be better targeted and tailored to host weed populations (*Tanner et al., 2015*; *Varia, Pollard & Ellison, 2016*). This approach provides data that support complete organellar genomes as a cheap and

effective genetic markers for larger case studies that aim to study population genetics. The successful sequencing of complete cp genomes from historic herbarium specimens, which proved to be comparable to recently sampled specimens, might be valuable for establishing the origin of invasive alien weed populations and aiding natural enemy surveys in the native range.

CONCLUSIONS

The complete cp genome sequences of two isolates of *I. glandulifera*, the first two sequenced members of the species, were assembled, annotated and analyzed in this study. The genome structure, gene content and gene order were similar in the two isolates. Complete cp genomes were used to investigate the variability of a 125-year-old herbarium specimen collected from the native range and a 2-year-old specimen from invasive UK field plants. Low coverage whole genome sequencing (genome skimming) was suitable for sequencing complete cp genomes of herbarium material and proved to be more informative than classical MLST. The complete cp genomes generated in this study are of critical importance to increase the number of sequence data available for the Balsaminaceae family and provide a basis for further research to tackle the rapid spread of *I. glandulifera*.

ACKNOWLEDGEMENTS

We thank Dr. Alan G. Buddie, Dr. Matthew J. Ryan and Ms. Kate M. Pollard for helpful comments on the manuscript. We thank Ms. Benedetta Caggiano for the Illumina library preparation, and Dr. Jovita Yesilyurt from the Natural History Museum for providing the herbarium leaf material used in this study.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

CABI is an international intergovernmental organization and receives core financial support from our member countries (and lead agencies) including the United Kingdom (Department for International Development), China (Chinese Ministry of Agriculture), Australia (Australian Centre for International Agricultural Research), Canada (Agriculture and Agri-Food Canada), Netherlands (Directorate-General for International Cooperation) and Switzerland (Swiss Agency for Development and Cooperation). Riccardo Baroncelli's research is supported by the project Escalera de Excelencia CLU-2018-04 co-funded by the P.O. FEDER of Castilla y León, Spain. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Department for International Development, United Kingdom.
Chinese Ministry of Agriculture, China.
Australian Centre for International Agricultural Research, Australia.

Agriculture and Agri-Food Canada, Canada.
Directorate-General for International Cooperation, Netherlands.
Swiss Agency for Development and Cooperation, Switzerland.
Escalera de Excelencia: CLU-2018-04.
P.O. FEDER of Castilla y León, Spain.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Giovanni Cafa conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Riccardo Baroncelli analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Carol A. Ellison performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Daisuke Kurose performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The sequences are available at GenBank: MK358446 and MK358447.

Data Availability

The following information was supplied regarding data availability:

The data is available at SRA: SRR8393251 and SRR8393261.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.8739#supplemental-information.

REFERENCES

Aust SK, Ahrendsen DL, Kellar PR. 2015. Biodiversity assessment among two Nebraska prairies: a comparison between traditional and phylogenetic diversity indices. *Biodiversity Data Journal* **3**:e5403 DOI 10.3897/BDJ.3.e5403.

Bankevich A, Nurk S, Antipov D, Gurevich A, Dvorkin M, Kulikov AS, Lesin V, Nikolenko S, Pham S, Prjibelski A, Pyshkin A, Sirotkin A, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19(5)**:455–477 DOI 10.1089/cmb.2012.0021.

Beans CM, Roach DA. 2015. An invasive plant alters pollinator-mediated phenotypic selection on a native congener. *American Journal of Botany* **102(1)**:50–57 DOI 10.3732/ajb.1400385.

Beerling DJ, Perrins JM. 1993. *Impatiens glandulifera* Royle (*Impatiens roylei* Walp.). *Journal of Ecology* **81(2)**:367–382 DOI 10.2307/2261507.

Besnard G, Gaudeul M, Lavergne S, Muller S, Rouhan G, Sukhorukov AP, Vanderpoorten A, Jabbour F. 2018. Herbarium-based science in the twenty-first century. *Botany Letters* **165(3–4)**:323–327 DOI 10.1080/23818107.2018.1482783.

Clements DR, Feenstra KR, Jones K, Staniforth R. 2008. The biology of invasive alien plants in Canada. 9. *Impatiens glandulifera* Royle. *Canadian Journal of Plant Science* **88(2)**:403–417 DOI 10.4141/CJPS06040.

Crawford PHC, Hoagland BW. 2009. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? *Journal of Biogeography* **36(4)**:651–661 DOI 10.1111/j.1365-2699.2008.02043.x.

Darling AE, Mau B, Perna NT. 2010. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE* **5(6)**:e11147 DOI 10.1371/journal.pone.0011147.

Ellison CA, Evans HC, Ineson J. 2004. The significance of intraspecies pathogenicity in the selection of a rust pathotype for the classical biological control of Mikania micrantha (mile-a-minute weed) in Southeast Asia. Canberra: CSIRO Entomology, 102–107.

Environment Agency. 2010. Our river habitats: the state of river habitats in England, Wales and the Isle of Man: a snap shot. London: Environment Agency. *Available at http://www.ecrr.org/Portals/27/Publications/Our%20river%20habitats.pdf* (accessed 2 January 2019).

Evans HC. 2002. Plant pathogens for biological control of weeds. In: Waller JM, Lenné JM, Waller SJ, eds. *Plant Pathologist's Pocketbook*. Wallingford: CAB International, 366–378.

Fujihashi H, Akiyama S, Ohba H. 2002. Origin and relationships of the Sino-Himalayan *Impatiens* (Balsaminaceae) based on molecular phylogenetic analysis, chromosome numbers and gross morphology. *Journal of Japanese Botany* **77**:284–295.

Gaskin JF, Bon M-C, Cock MJ, Cristofaro M, De Biase A, De Clerck-Floate R, Ellison CA, Hinz HL, Hufbauer RA, Julien MH, Sforza R. 2011. Applying molecular-based approaches to classical biological control of weeds. *Biological Control* **58(1)**:1–21 DOI 10.1016/j.biocontrol.2011.03.015.

Gaskin JF, Zhang D-Y, Bon M-C. 2005. Invasion of Lepidium draba (Brassicaceae) in the western United States: distributions and origins of chloroplast DNA haplotypes. *Molecular Ecology* **14(8)**:2331–2341 DOI 10.1111/j.1365-294X.2005.02589.x.

Goolsby JA, DeBarro PJ, Makinson JR, Pemberton RW, Hartley DM, Frohlich DR. 2006. Matching the origin of an invasive weed for selection of a herbivore haplotype for a biological control programme. *Molecular Ecology* **15(1)**:287–297 DOI 10.1111/j.1365-294X.2005.02788.x.

Hulme PE, Bremner ET. 2006. Assessing the impact of *Impatiens* glandulifera on riparian habitats: partitioning diversity components following species removal. *Journal of Applied Ecology* **43(1)**:43–50 DOI 10.1111/j.1365-2664.2005.01102.x.

Jansen RK, Ruhlman TA. 2012. Plastid genomes of seed plants. In: Bock R, Knoop V, eds. *Genomics of Chloroplasts and Mitochondria*. Dordrecht: Springer, 103–126.

Jukonienė I, Subkaitė M, Ričkienė A. 2019. Herbarium data on bryophytes from the eastern part of Lithuania (1934–1940) in the context of science history and landscape changes. *Botanica* **25(1)**:41–53 DOI 10.2478/botlit-2019-0005.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4)**:772–780 DOI 10.1093/molbev/mst010.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33(7)**:1870–1874 DOI 10.1093/molbev/msw054.

Kurose D, Pollard KM, Ellison CA. 2018. Searching for host-pathogen compatibility: how cpDNA analysis can aid classical biological control of *Impatiens glandulifera*. In: *Proceedings of the XV International Symposium on Biological Control of Weeds, 26–31 August 2018, Engelberg, Switzerland.* 189.

Lang PL, Willems FM, Scheepens JF, Burbano HA, Bossdorf O. 2019. Using herbaria to study global environmental change. *New Phytologist* **221(1)**:110–122 DOI 10.1111/nph.15401.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI 10.1038/nmeth.1923.

Li Z-Z, Saina JK, Gichira AW, Kyalo CM, Wang Q, Chen J. 2018. Comparative genomics of the balsaminaceae sister genera *Hydrocera triflora* and *Impatiens pinfanensis*. *International Journal of Molecular Sciences* **19(1)**:319 DOI 10.3390/ijms19010319.

Lim J, Crawley MJ, De Vere N, Rich T, Savolainen V. 2014. A phylogenetic analysis of the British flora sheds light on the evolutionary and ecological factors driving plant invasions. *Ecology and Evolution* **4**:4258–4269 DOI 10.1002/ece3.1274.

Lobstein A, Brenne X, Feist E, Metz N, Weniger B, Anton R. 2001. Quantitative determination of naphthoquinones of *Impatiens* species. *Phytochemical Analysis* **12(3)**:202–205 DOI 10.1002/pca.574.

Lohse M, Drechsel O, Bock R. 2007. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* **52(5–6)**:267–274 DOI 10.1007/s00294-007-0161-y.

Pattison Z, Rumble H, Tanner RA, Jin L, Gange AC. 2016. Positive plant-soil feedbacks of the invasive *Impatiens glandulifera* and their effects on above-ground microbial communities. *Weed Research* **56(3)**:198–207 DOI 10.1111/wre.12200.

Perrins J, Fitter A, Williamson M. 1993. Population biology and rates of invasion of three introduced *Impatiens* species in the British Isles. *Journal of Biogeography* **20(1)**:33–44 DOI 10.2307/2845737.

Prentis PJ, Sigg DP, Raghu S, Dhileepan K, Pavasovic A, Lowe AJ. 2009. Understanding invasion history: genetic structure and diversity of two globally invasive plants and implications for their management. *Diversity and Distributions* **15(5)**:822–830 DOI 10.1111/j.1472-4642.2009.00592.x.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLOS ONE* **10(3)**:e9490 DOI 10.1371/journal.pone.0009490.

Reeve MA, Pollard KM, Kurose D. 2018. Differentiation between closely-related *Impatiens* spp. and regional biotypes of *Impatiens glandulifera* using a highly-simplified and inexpensive method for MALDI-TOF MS. *Plant Methods* **14(1)**:60 DOI 10.1186/s13007-018-0323-6.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **12(12)**:1572–1574 DOI 10.1093/bioinformatics/btg180.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. *American Journal of Botany* **99(2)**:349–364 DOI 10.3732/ajb.1100335.

Tanner RA. 2017. Information on measures and related costs in relation to species included on the union list: *Impatiens glandulifera*. *Available at https://circabc.europa.eu/sd/a/90d96077-f628-4975-a760-3e6249d0e297/TSSR-2016-003%20Impatiens%20glandulifera(0).pdf* (accessed 2 January 2019).

Tanner RA, Ellison CA, Seier MK, Kovacs GM, Kassai-Jager E, Berecky Z, Varia S, Djeddour D, Singh MC, Csiszar A, Csontos P, Kiss L, Evans HC. 2015. *Puccinia komarovii* var. *glanduliferae* var. nov.: a fungal agent for the biological control of Himalayan balsam

(*Impatiens glandulifera*). *European Journal of Plant Pathology* **141(2)**:247–266 DOI 10.1007/s10658-014-0539-x.

Tanner RA, Gange AC. 2013. The impact of two non-native plant species on native flora performance: potential implications for habitat restoration. *Plant Ecology* **214(3)**:423–432 DOI 10.1007/s11258-013-0179-9.

Tanner RA, Jin L, Shaw HR, Murphy ST, Gange AC. 2014. An ecological comparison of *Impatiens glandulifera* Royle in the native and introduced range. *Plant Ecology* **215(8)**:833–843 DOI 10.1007/s11258-014-0335-x.

Tanner RA, Varia S, Eschen R, Wood S, Murphy ST, Gange AC. 2013. Impacts of an invasive non-native annual weed, *Impatiens glandulifera*, on above- and below-ground invertebrate communities in the United Kingdom. *PLOS ONE* **8(6)**:e67271 DOI 10.1371/journal.pone.0067271.

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq–versatile and accurate annotation of organelle genomes. *Nucleic Acids Research* **45(W1)**:W6–W11 DOI 10.1093/nar/gkx391.

Tomley AJ, Evans HC. 2004. Establishment of, and preliminary impact studies on, the rust, Maravalia cryptostegiae, of the invasive alien weed, Cryptostegia grandiflora in Queensland, Australia. *Plant Pathology* **53(4)**:475–484 DOI 10.1111/j.1365-3059.2004.01054.x.

Varia S, Pollard K, Ellison CA. 2016. Implementing a novel weed management approach for Himalayan balsam: progress on biological control in the UK. *Outlooks on Pest Management* **27(5)**:198–203 DOI 10.1564/v27_oct_02.

Yu S-X, Janssens S-B, Zhu X-Y, Liden M, Gao T-G, Wang W. 2015. Phylogeny of *Impatiens* (Balsaminaceae): integrating molecular and morphological evidence into a new classification. *Cladistics* **32(2)**:179–197 DOI 10.1111/cla.12119.

Zeng C-X, Hollingsworth P-M, Yang J, He Z, Zhang Z, Li D, Yang J-B. 2018. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Methods* **14(1)**:43 DOI 10.1186/s13007-018-0300-0.