PLOS ONE

RESEARCH ARTICLE

A Phylogenomic Approach Based on PCR Target Enrichment and High Throughput Sequencing: Resolving the Diversity within the South American Species of *Bartsia* L. (Orobanchaceae)

Simon Uribe-Convers[1,2,3☯]*, **Matthew L. Settles**[1,2☯], **David C. Tank**[1,2,3]

1 Department of Biological Sciences, University of Idaho, Moscow, Idaho, United States of America, **2** Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, Idaho, United States of America, **3** Stillinger Herbarium, University of Idaho, Moscow, Idaho, United States of America

☯ These authors contributed equally to this work.
* uribe.convers@gmail.com

Citation: Uribe-Convers S, Settles ML, Tank DC (2016) A Phylogenomic Approach Based on PCR Target Enrichment and High Throughput Sequencing: Resolving the Diversity within the South American Species of *Bartsia* L. (Orobanchaceae). PLoS ONE 11(2): e0148203. doi:10.1371/journal.pone.0148203

Editor: Berthold Heinze, Austrian Federal Research Centre for Forests BFW, AUSTRIA

Received: June 19, 2015

Accepted: January 14, 2016

Published: February 1, 2016

Data Availability Statement: All high-throughput sequencing data files are available from the GenBank Sequence Read Archive (SRA) accession numbers: SRR2045582, SRR2045585, SRR2045588, SRR2045589, SRP058302. Additionally, custom R scripts, molecular matrices, sequence data, and analyses and results files are available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592.

Funding: Funding for this work was provided by the National Science Foundation (DEB-1210895, http://www.nsf.gov/awardsearch/showAward?AWD_ID=1210895&HistoricalAwards=false, to DCT for SUC,

Abstract

Advances in high-throughput sequencing (HTS) have allowed researchers to obtain large amounts of biological sequence information at speeds and costs unimaginable only a decade ago. Phylogenetics, and the study of evolution in general, is quickly migrating towards using HTS to generate larger and more complex molecular datasets. In this paper, we present a method that utilizes microfluidic PCR and HTS to generate large amounts of sequence data suitable for phylogenetic analyses. The approach uses the Fluidigm Access Array System (Fluidigm, San Francisco, CA, USA) and two sets of PCR primers to simultaneously amplify 48 target regions across 48 samples, incorporating sample-specific barcodes and HTS adapters (2,304 unique amplicons per Access Array). The final product is a pooled set of amplicons ready to be sequenced, and thus, there is no need to construct separate, costly genomic libraries for each sample. Further, we present a bioinformatics pipeline to process the raw HTS reads to either generate consensus sequences (with or without ambiguities) for every locus in every sample or—more importantly—recover the separate alleles from heterozygous target regions in each sample. This is important because it adds allelic information that is well suited for coalescent-based phylogenetic analyses that are becoming very common in conservation and evolutionary biology. To test our approach and bioinformatics pipeline, we sequenced 576 samples across 96 target regions belonging to the South American clade of the genus *Bartsia* L. in the plant family Orobanchaceae. After sequencing cleanup and alignment, the experiment resulted in ~25,300bp across 486 samples for a set of 48 primer pairs targeting the plastome, and ~13,500bp for 363 samples for a set of primers targeting regions in the nuclear genome. Finally, we constructed a combined concatenated matrix from all 96 primer combinations, resulting in a combined aligned length of ~40,500bp for 349 samples.

and DEB-1253463, http://www.nsf.gov/awardsearch/showAward?AWD_ID=1253463, to DCT) and the University of Idaho Student Grant Program to SUC. Fieldwork was supported in part by the University of Idaho Graduate Student Research Awards to SUC, by the Society of Systematic Biologists (SSB) to SUC, the Botanical Society of America (BSA) to SUC, the American Society of Plant Taxonomists (ASPT) to SUC, and the University of Idaho Stillinger Herbarium Expedition Funds to SUC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Advances in high-throughput sequencing (HTS) have allowed researchers to obtain large amounts of genomic information at speeds and costs unimaginable only a decade ago. The fields of phylogenetics and population genetics have benefitted greatly from these advancements, and large phylogenomic and population genomic datasets are becoming more common [1–3]. Driven by the need to generate homogenous, informative, and affordable multilocus datasets, we present an approach for obtaining affordable, large, multilocus datasets for phylogenetic and population genetic studies, based on microfluidic PCR amplification with the Fluidigm Access Array System (Fluidigm, San Francisco, CA, USA) and HTS. Microfluidic PCR technology has been used extensively in the fields of cancer research (e.g., [4,5]), genotyping of single nucleotide polymorphisms (SNP) (e.g., [6–8]), gene expression (e.g., [9–11]), and targeted resequencing (e.g., [12,13]). Recently, this technology has started to be used for generating phylogenetic datasets in microbial systems [14], haplotyping of commercially important plants [15], and elucidating recent radiations in plants from Madagascar [16]—see also [17] for more on its potential uses in phylogenetics. This approach uses the Fluidigm Access Array System (Fluidigm) and two sets of PCR primers to simultaneously amplify 48 target regions across 48 samples, incorporating sample-specific barcodes and HTS adapters (2,304 amplicons per microfluidic array). This four-primer PCR approach circumvents the need to construct genomic libraries for every sample, avoiding the high costs and time requirements involved in library preparation. Furthermore, by using a dual barcoding strategy, we were able to multiplex 24 Fluidigm Access Arrays on a single Illumina MiSeq run, representing two distinct sets of 48 target regions (plastome and nuclear, in this experiment) across 576 samples (55,296 distinct amplicon sequences) from the South American species of the plant genus *Bartsia* L.—the Neobartsia clade [18] (Fig 1)—and its close relatives in Orobanchaceae, demonstrating the power of this approach for species-level phylogenetics.

In plant phylogenomics, there has been a special focus on the chloroplast genome, also know as the plastome, given its phylogenetic informativeness at all taxonomic scales (e.g., [19–22]), the straightforwardly interpreted results due to its non-recombining nature, conserved gene order and gene content [23], and its historical importance since the beginning of the field (e.g., [24]). Large datasets have been produced with approaches that have involved massively parallel sequencing (e.g., [21]), a compilation of coding regions from both whole plastome sequences (e.g., [22]) and targeted approaches (e.g., [25]), transcriptomics (e.g., [26]), RNA hybridization or capture probes (e.g., [3]), and long-range PCR coupled with HTS (e.g., [27]). Because the chloroplast genome evolves relatively slowly, ~3–5 times slower than the nuclear genome in plants [28–30], the power of these datasets for phylogenetic studies lies in their size; at ~150 kilobases (kb), plastome datasets can provide phylogenetic resolution from the interspecific level (e.g., [21,31]) to the level of major clades (e.g., [22,32]). However, because it is inherited as a single unit, plastome sequences only provide information from a single locus, and although often well-supported, phylogenies based solely on plastome-scale datasets may be misleading because of well known evolutionary processes that can lead to gene tree-species tree discordance [33]. This may be especially problematic at low-taxonomic scales where processes such as coalescent stochasticity and gene flow may be more prevalent. Thus, data from multiple independently evolving loci are necessary to fully understand the evolutionary history of a group of organisms, and to take full advantage of the emerging species-tree paradigm made possible by the integration of population genetic processes into phylogenetic reconstruction via the multispecies coalescent [34–36].

For the nuclear genome, phylogenomic datasets have been obtained in plant systems using genome skimming (e.g., [37]), sequence capture [38–41], and restriction-site associated DNA

Fig 1. Floral diversity of the South American species of *Bartsia*, i.e., the Neobartsia clade. The sections from top left to bottom right are: *Strictae*, *Diffusae*, *Laxae*, and *Orthocarpiflorae*.

doi:10.1371/journal.pone.0148203.g001

sequencing (RADseq) (e.g., [42]). Likewise, the field of phylogenomics in animals has advanced with datasets obtained using targeted amplicon sequencing (TAS) in Pancrustacea [43] and North American tiger salamanders [44], GBS in butterflies [45], fish [46], and beetles [47]. However, genome-scale datasets for animal phylogenetics has been most heavily impacted by sequence capture approaches focused on ultraconserved genomic elements (UCEs) at various taxonomic scales, e.g., vertebrates [1], amniotes [2,48], turtles [49], birds [50], ray-finned fishes [51], and chipmunks [52]. Furthermore, UCEs have been shown to be an important resource for gathering information from museum specimens [53], and to be useful at shallow evolutionary time scales in birds [54].

In plant systems, genome skimming—as implemented in the Alignreads pipeline [55]—has perhaps had the most impact for assembling phylogenetic datasets from HTS data. In contrast

to sequence capture approaches that require preliminary genomic data for capture bait design, genome skimming requires little to no pre-existing genomic information. Genome skimming is a reference-guided approach that takes advantage of high copy regions in the genome, e.g., nuclear rDNA, the plastome, and the mitochondrial genome. By using reference sequences, this method 'skims' out targeted regions from low-coverage genomic data. This approach has been used to recover the mitochondrial and chloroplast genomes in the genus *Asclepias* L. [37], study introgression in *Fragaria* L. species [56], identify horizontal transfer of DNA from the mitochondrion to the chloroplast in *Asclepias syriaca* L. [57], resolve phylogenetic relationships in the family Chrysobalanaceae [58], recover plastomes across multiple genera [27], and was also used to assemble the plastomes used for microfluidic PCR primer design in this study.

Both the UCE sequence capture and genome skimming approaches share similar technical and fundamental constraints that make their utility for phylogenetics at low-taxonomic levels with large sampling strategies limited. First, both of these methods are limited by the need to first construct HTS libraries for each sample in the study, a step that greatly increases the time and costs of the experiment. Second, variable regions flanking the UCEs are often captured at much reduced depths as one moves away from the UCE, or the UCE is lost completely if the target taxon is phylogenetically divergent from the one used in the bait design [3,54]. Smith et al. [54] found that UCEs containing variable flanking regions were usually not recovered across all samples if the variable regions extended more than 300 bp from the UCE probe. This is unfortunate, given that the more variable regions are of potentially greater utility for inter-specific phylogenetic and population genetic studies. Likewise, genome skimming from low-coverage genomic data is most useful for recovering high-copy number regions in the genome; however, regions with lower representation numbers, such as single copy nuclear genes, are likely to be recovered in some samples and missed completely in others [55], depending on the depth of the low-coverage genomic data and the phylogenetic distance of the references used for mapping. Both of these cases result in the introduction of missing data, which could potentially lead to incorrect or misleading phylogenetic inferences [59]. In contrast, the large scale targeting of chloroplast, nuclear rDNA, and multiple independent single-copy nuclear genes using the Fluidigm Access Array and HTS circumvents many of these problems.

Our approach is similar in theory to targeted amplicon sequencing (TAS) methods (e.g., [43,44]), but contains major improvements in efficiency. For example, Bybee et al. [43] implemented a first round of PCRs to amplify each target region, which was then reamplified in a second round of PCRs to incorporate barcodes and HTS adapters. While using this two-reaction approach allows for more flexibility in the annealing temperature of target specific primers, this approach is labor intensive and thus difficult to scale to hundreds of samples and/or a large number of targets to take full advantage of the current yield of most HTS platforms. In their study, Bybee et al. [43] amplified six genes for 44 taxa from Pancrustacea, which translates to performing 12 PCRs for each of the 44 taxa to amplify and tag each amplicon. At this scale, both in terms of the number of samples and the number of loci, this method may be more favorable than the approach proposed here, however, once 48 or 96 different primer pairs are used to amplify hundreds of samples, this method becomes inefficient. We believe that experiments with high numbers of samples and loci are quickly becoming more common (e.g., [14–16]), and that the fields of systematics, phylogenetics, and population biology need more tools to deal with this type of sampling.

In this paper, we test the performance and utility of our targeted approach using the Neotropical clade of the plant genus *Bartsia* L. (Orobanchaceae) (Fig 1). This clade is comprised of approximately 45 closely related species that are part of an ongoing rapid and recent radiation in the páramo ecosystem above tree line (~2900m in elevation) throughout the Andes [18]. Using minimal genomic resources collected via plastome sequencing [27] and low-coverage

Table 1. Sample and sequencing information for preliminary data acquisition.

Species	Collection Voucher	Platform	Type of read (bp)	No. clean reads (M)	Sequencing Facility	Source
Bartsia inaequalis Benth.	Uribe-Convers 2010–022	Illumina HiSeq 2000	100 single end	0.93	Berkeley	Uribe-Convers et al. 2014
Bartsia stricta (Kunth) Benth.	Uribe-Convers 2010–024	Illumina HiSeq 2000	100 single end	0.9	Berkeley	Uribe-Convers et al. 2014
Bartsia pedicularoides Benth.	Antonelli 574	Illumina HiSeq 2000	100 single end	46.8	Berkeley	This study
Bartsia santolinifolia (Kunth) Benth.	Uribe-Convers 2010–041	Illumina HiSeq 2000	100 paired end	46.4	UO	This study
Bartsia pedicularoides Benth.	Uribe-Convers 2011–064	Illumina HiSeq 2000	100 paired end	52.9	UO	This study
Bartsia serrata Molau	Uribe-Convers 2012–015	Illumina HiSeq 2000	100 paired end	65.1	UO	This study

Sequencing information of the samples used during the preliminary data acquisition step. Type of read refers to the length of the read in base pairs (bp), and if it was single or paired end. Number of reads denotes the number of raw reads in millions (M). Berkeley = Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley; UO = Genomics Core Facility at the University of Oregon. Additional collecting information about these samples can be found in S3 Table.

doi:10.1371/journal.pone.0148203.t001

genome sequencing in representative species of *Bartsia*, we present an approach for designing microfluidic PCR primer combinations for amplifying i) the most variable regions of the plastome (referred to as the chloroplast set henceforth), ii) the commonly sequenced ITS and ETS regions of the nuclear rDNA repeat, and iii) a suite of putatively single-copy nuclear loci (ii and iii are referred to as the nuclear set henceforth). Our targeted approach generated a large multilocus dataset across hundreds of samples, which allowed us to investigate evolutionary relationships at the species level. While shotgun approaches yield more data, the great majority of these data are highly conserved across samples and thus phylogenetically uninformative. By focusing on targeted loci and not whole genomes, we were able to maximize the yield of shared and phylogenetically informative data across a significantly greater number of samples, which is ideal for phylogenetic studies at low taxonomic levels.

Methods

Microfluidics PCR Primer Design and Validation

Preliminary data acquisition. Data used for chloroplast and nuclear microfluidic PCR primer design were compiled from two taxa using long-range PCR to generate plastome DNA templates for HTS [27] and three taxa (4 samples, Table 1) using genome skimming [37]. DNA was extracted from ~0.02 g of silica gel-dried tissue using a modified 2X CTAB method [60], yielding 30 to 70 ng/μL of DNA per sample. Genomic DNAs were sheared by nebulization at 30 psi for 70 sec, yielding an average shear size of 500bp as measured by a Bioanalyzer High-Sensitivity Chip (Agilent Technologies, Inc., Santa Clara, California, USA). Sequencing libraries were constructed using the Illumina TruSeq library preparation kit and protocol (Illumina Inc., San Diego, California, USA) and were standardized at 2nM prior to sequencing. Library concentrations were determined using the KAPA qPCR kit (KK4835) (Kapa Biosystems, Woburn, Massachusetts, USA) on an ABI StepOnePlus Real-Time PCR System (Life Technologies, Grand Island, New York, USA). The long-range PCR libraries and one genome skimming library were sequenced on an Ilumina HiSeq 2000 at the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley, whereas the remaining

genome skimming libraries were sequenced on an Illumina HiSeq 2000 at the Genomics Core Facility at the University of Oregon (Table 1). Raw reads were cleaned using SeqyClean v 1.8.10 (https://bitbucket.org/izhbannikov/seqyclean) using defaults settings to remove sequencing adaptors and low quality reads (<20 Phred Quality Scores).

Chloroplast target selection. For the chloroplast PCR primer set, the cleaned reads were assembled against a reference genome (*Sesamum indicum* L., GenBank accession JN637766) using the Alignreads pipeline v. 2.25 [55], and visually inspected using Geneious R6 v6.1.5 (Biomatters, Auckland, New Zealand). Six complete plastomes—two from the long-range PCR approach and four from genome skimming—were aligned using MAFFT v.7.017b in its default settings [61]. To target the putatively most phylogenetically informative regions of the plastome, we developed custom scripts in R [62] to identify the most variable regions in the alignment. The R script takes a multiple sequence alignment as input and identifies 400–1000bp regions of high sequence divergence that are flanked by conserved regions, and automatically generates an output table where the regions that satisfy the searching criteria are shown. This allowed us to rank and prioritize regions in the plastome for primer design. The script has been deposited in the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.fh592) and was written to be general—although, as commented on the script, different versions of R and the libraries used by the script might affect its functionality.

Nuclear target selection. For the nuclear PCR primer set, cleaned reads from the genome skimming samples were compared to two publicly available genomic databases, a list of the pentatricopeptide repeat genes (PPR) and the conserved orthologous set II (COSII), using the BLAST-Like Alignment Tool (BLAT, tileSize = 7, minIdentity = 80) [63]. We chose these two reference databases because a list of 127 PPR loci was shown to have a single ortholog in both rice (*Oryza sativa* L.) and *Arabidopsis thaliana* (L.) Heynh. [64], and was previously used to successfully infer the phylogenetic relationships of the plant family Verbenaceae and the *Verbena* L. complex [65], and in the plant clade Campanuloideae (Campanulaceae) [66]. Similarly, the COSII genes have been identified to be putatively single-copy and orthologous across the Euasterid plant clade [67], and several loci have been used for phylogenetic reconstructions of closely related species in the plant families Orobanchaceae [68] and Solanaceae [69].

Using a custom R-script (deposited in the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592), reads that matched any reference gene from these two databases were kept, binned with their respective reference locus, and aligned using MAFFT in its default settings. We then used the online tool IntronFinder (http://solgenomics.net/, last accessed in January 2014) from the Sol Genomics Network [70] to predict exon/intron junction positions in the COSII genes. The PPR genes do not contain introns [64] and thus this step was not necessary for these loci. Reference loci that had reads from at least two taxa aligned to them forming conserved 'islands' separated by 400-800bp, including estimated introns with an assumed average length of 100bp, were selected for primer design (Fig 2A). Additionally, an alignment from nuclear ribosomal DNA (rDNA) internal and external transcribed spacers sequences—ITS and ETS, respectively [18]—as well as an alignment of sequences of the *PHOT1* gene and one of the *PHOT2* gene [71] were made in MAFFT with default settings.

Microfluidic PCR primer design and validation. Forward and reverse primers for the selected chloroplast regions and nuclear loci were designed using Primer3 [72–74] following the recommended criteria specified in the Fluidigm Access Array System protocol (Fluidigm, San Francisco, CA, USA), e.g., annealing temperature was set to 60°C (+/- 1°C) for all primers, and no more than three continuous nucleotides of the same base were allowed (Max Poly-X = 3). Furthermore, regions identified as appropriate for primer design that were not present in every taxon in the alignment or that contained ambiguous bases (due to missing data and/or low coverage in our assemblies) were discarded. A complete list of the chosen primers can be

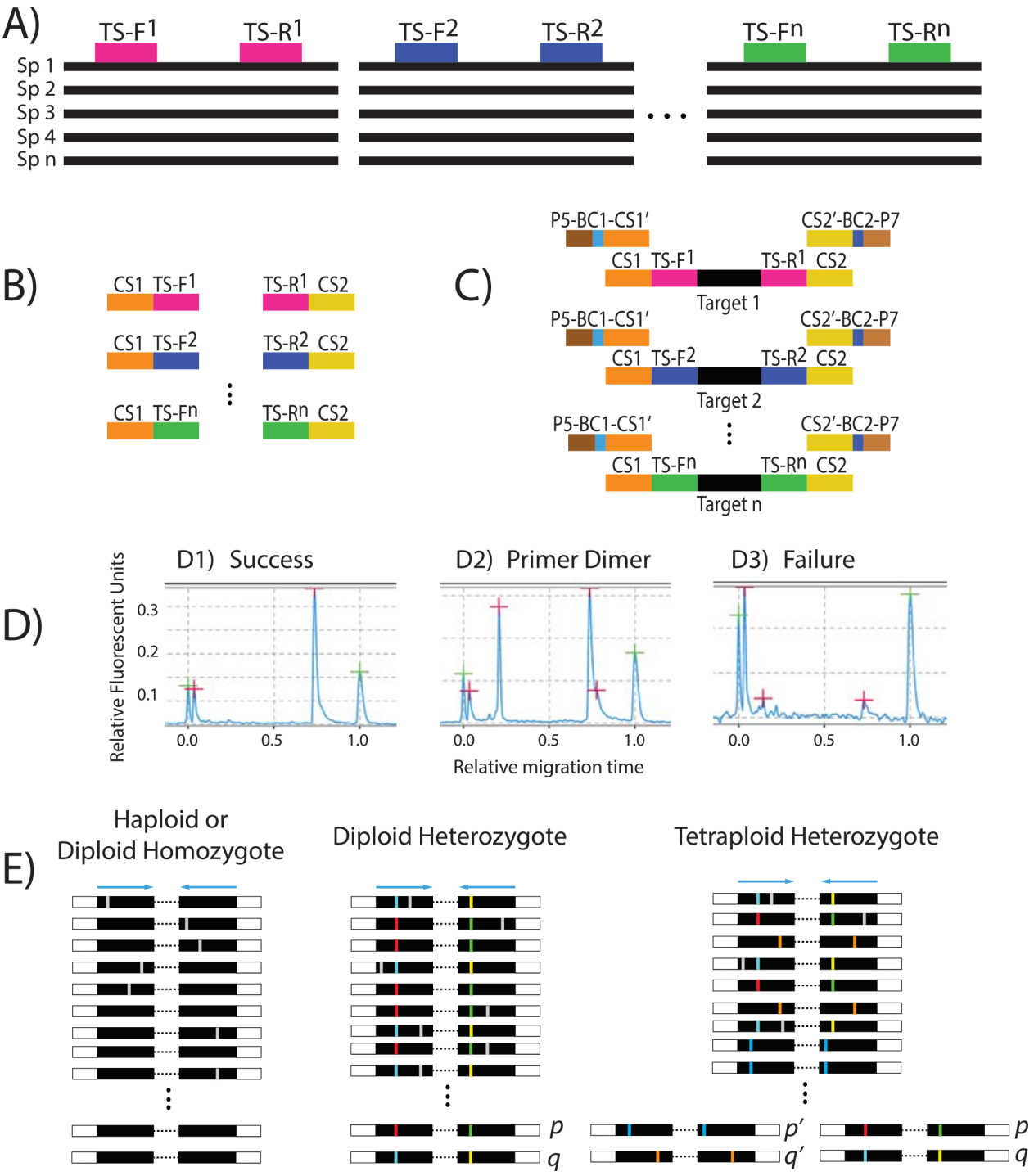

Fig 2. Flowchart describing the method used in this study. (A) Forward and reverse target specific primer combinations (TS-F and TS-R, respectively) designed in Primer3 from a multiple sequence alignment of existing genomic resources obtained in the preliminary data acquisition step. (B) The conserved sequences (CS1 and CS2) are added to the target specific primers at the time of synthesis. (C) Each target specific combination needs to be validated to ensure amplification. This is performed by simulating the microfluidic amplification reaction in a standard thermocycler, with both the first and the second pair of primers added (4-primer reaction). The second pair of primers is comprised of the sequencing adapters (e.g., P5 and P7, for Illumina), a barcode combination (BC) specific for each sample, and the reverse complement of the conserved sequences (CS1' and CS2'). (D) Each reaction is analyzed for successful amplifications (D1), primer dimers (D2), or failed amplifications (D3). Only primer combinations with successful amplification and no primer dimers are chosen. (E) After sequencing, the reads are demultiplexed, sample-specific pools of amplicon sequences are generated, and groups of identical reads are identified in each pool. Pools of identical sequence reads represented by at least 5 reads and representing at least 5% of the total reads for that amplicon/sample are kept as alleles. Three examples are shown to illustrate potential results from one amplicon in one sample: a haploid or diploid homozygote

sample with just one identical sequence, a diploid heterozygote sample with two different sequences (p & q), and a tetraploid heterozygote sample with two sets of homeologs (p & q and p' & q'). Each line represents the demultiplexed reads for that amplicon/sample with PCR and/or sequencing errors denoted in white, and allele-specific substitutions in colors. The final line (after the three vertical dots) is the final result after the groups of identical reads have been identified. The possible resulting alleles are denoted with p & q and p' & q', depending if it is a diploid or a tetraploid sample.

doi:10.1371/journal.pone.0148203.g002

found in S1 Table. Once the initial primer design was completed, a conserved sequence (CS) tail was added to the 5' end of both the forward and reverse primers, CS1 and CS2 respectively (Fluidigm), resulting in the final target specific primers (TS) with universal tails (CS1-TS-F and CS2-TS-R, respectively). The purpose of the added tails (CS1 and CS2) is to provide an annealing site for the second pair of primers, which, starting from the 5' end, are composed of the HTS adapters (e.g., P5 or P7 for Illumina sequencing), a sample specific forward and reverse barcode combination (e.g., BC1 and BC2), and the complementary CS sequence (CS1' or CS2'; Fig 2B and 2C). To avoid confusion, the first pair of primers with universal tails (CS1-TS-F and CS2-TS-R) will be referred to as the 'target specific primers', and the second pair of primers—with complementary universal tails, barcodes, and Illumina adapters; P5-BC-CS1' and P7-BC-CS2'—will be referred to as the 'barcoded primers'. The CS1 and CS2 sequences were obtained from the Fluidigm Access Array System protocol, Illumina Nextera barcode sequences (E5, A5/A7, D5/D7 and N7 barcode sets, total 24 P5 barcodes and 36 P7 barcodes, 8 bp barcodes, each $>$ 2bp distinct) were used to design custom pairs of barcodes (864 total unique pairs) allowing us to dramatically increase the number of samples that can be multiplexed in one sequencing run (S2 Table).

Primer validation. Due to the complexity of simultaneously using two sets of primers in one PCR, it is necessary to validate each set of primers prior to the actual microfluidic PCR amplification in the Fluidigm Access Array. Primer validation is a crucial step to ensure that no primer dimers are formed and that no interaction and/or competition between the barcoded and target specific primer pairs are negatively affecting the amplification. Primer validation was performed for each primer combination in 10 μL reactions in an Eppendorf Mastercycler ep thermocycler, following the Fluidigm Access Array System protocol. Validation reactions were performed on three species of *Bartsia* (*B. mutica* (Kunth) Benth., *B. crisafullii* N. H. Holmgren, and *B. melampyroides* (Kunth) Benth.), which represent the morphological and geographical diversity in the genus, and a negative control (using water instead of DNA), and included the following: 1μL of 10X FastStart High Fidelity Reaction Buffer without $MgCl_2$ (Roche Diagnostic Corp., Indianapolis, Indiana, USA), 1.8 μL of 25 mM $MgCl_2$ (Roche), 0.5 μL DMSO (Roche), 0.2 μL 10mM PCR Grade Nucleotide Mix (Roche), 0.1 μL of 5 U/μL FastStart High Fidelity Enzyme Blend (Roche), 0.5 μL of 20X Access Array Loading Reagent (Fluidigm), 2 μL of 2 μM barcoded primers, 2μL of 50nM target specific primers, 0.5 μL of 30–70 ng/μL genomic DNA, 1.4 μL of PCR Certified Water (Teknova, Hollister, California, USA). Resulting amplicons from these reactions were analyzed in a QIAxcel Advance System (Qiagen, Valencia, California, USA), and primer pairs that produced a single amplicon and had no (or minimal) primer dimers were selected (Fig 2D, S1 Table).

Sampling, microfluidic PCR and sequencing

We were interested in generating data to investigate the evolutionary history of the Neotropical *Bartsia* clade [75], and thus, we sampled the complete species richness of the group, including multiple individuals per species, and some of its close relatives. A total of 74 species were represented across the 576 samples. These samples encompassed the entire geographic breadth of the Neobartsia clade, with samples ranging from northern Colombia to southern Chile (S3 Table, Fig 3). The majority of samples were collected in the field, dried in silica-gel desiccant,

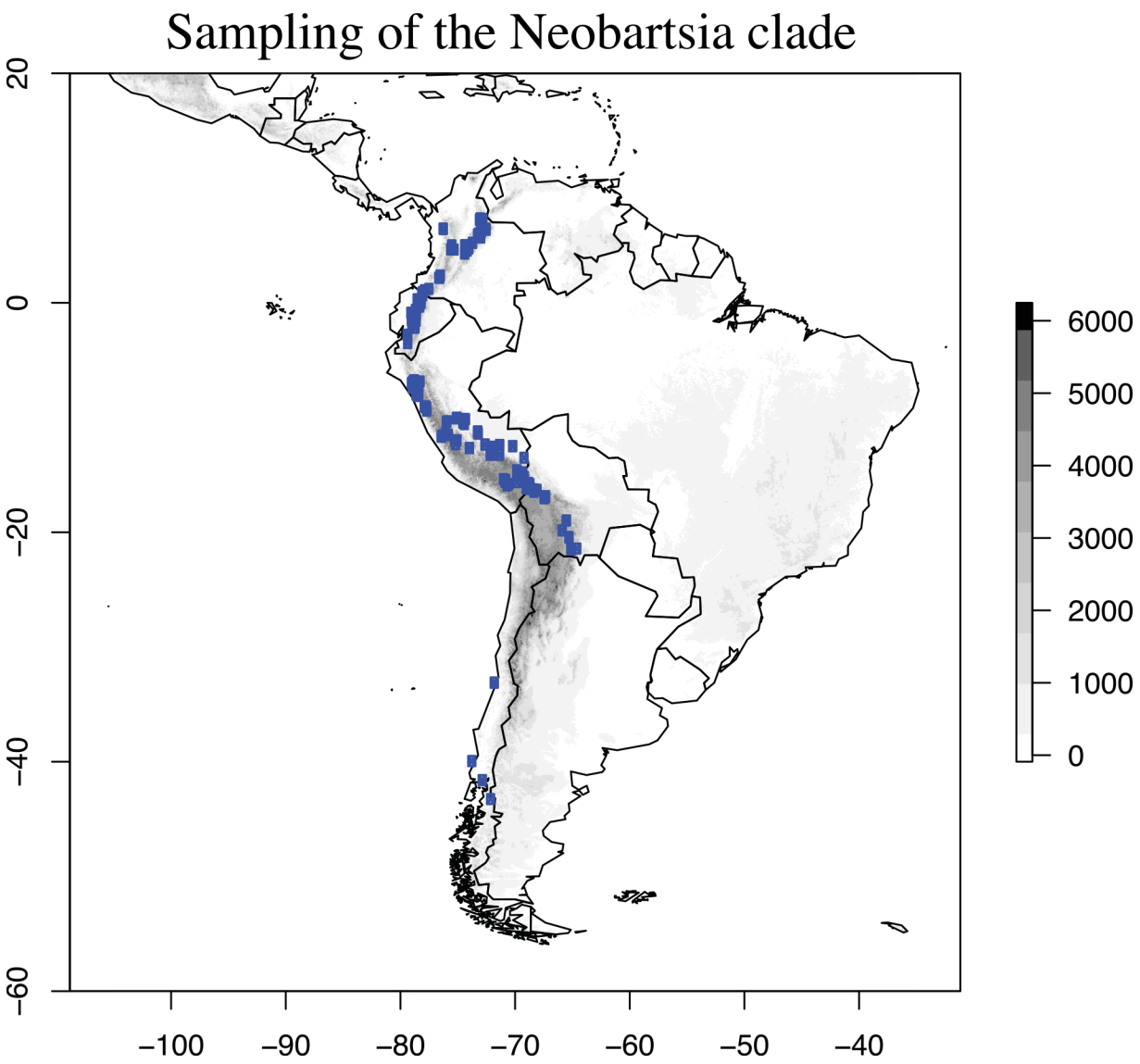

Fig 3. Sampling effort in the Neobartsia clade. A total of 576 samples, localities represented by dots, were collected for this study. Samples of *Bellardia trixago* L., *Bellardia viscosa* (L.) Fisch. & C.A. Mey, and *Parentucellia latifolia* (L.) Caruel not shown. The Y and X axis represent latitude and longitude, respectively, and the gray-scale to the right denotes elevation in meters. This figure was generated in R [62] using the packages maps, maptools, and raster [76–78].

doi:10.1371/journal.pone.0148203.g003

and stored in airtight bags. When field-collected tissue was not available, leaf tissue was sampled from herbarium specimens (S3 Table). Collecting permits were granted by the Ministerio de Medio Ambiente y Agua (Bolivia), Viceministerio de Medio Ambiente, Biodiversidad, Cambios Climáticos y de Gestión y Desarrollo Forestal (Bolivia), Ministerio de Ambiente (Ecuador), and the Dirección General Forestal y de Fauna Silvestre (Peru). Permission from the owner was granted for samples collected on private land. Field studies did not involve endangered or protected species. For all samples, DNA was extracted from ~0.02 g of silica gel-dried tissue using a modified 2X CTAB method [60], yielding ~30 to 70 ng/μl of DNA per sample.

Microfluidic PCR was performed in an Access Array System (Fluidigm) using 24 fully loaded 48.48 Access Array Integrated Fluidic Circuits (Fluidigm) (12 for the chloroplast and 12 for the nuclear set) following the manufacturer's protocols. This particular array allows for 48

samples to be simultaneously amplified across 48 distinct primer pairs, resulting in 2,304 isolated and unique PCR amplicons per array. While we chose here to amplify our 48 chloroplast and 48 nuclear targets in separate microfluidic arrays, we have also had success with multiplexing genomically divergent regions such as these and performing amplification of all 96 primer pairs in a single array (i.e., target specific primers for one chloroplast region and one nuclear locus pooled prior to amplification). The amplicons were harvested from each array as per the Fluidigm Access Array System protocol and pooled per sample in equal volumes. To remove unused reagents and/or undetected primer dimers smaller than 350bp, each pool was purified with 0.6X AMPure XP beads (Agencourt, Beverly, Massachusetts, USA). The purified pools were analyzed in a Bioanalyzer High-Sensitivity Chip (Agilent Technologies, Santa Clara, California, USA) and standardized at 13 pM using the KAPA qPCR kit (KK4835; Kapa Biosystems, Woburn, Massachusetts, USA) on an ABI StepOnePlus Real-Time PCR System (Life Technologies, Grand Island, New York, USA). The resulting pools were multiplexed in an Illumina MiSeq using the Reagent Kit version 3 with a final yield of 21.4 million 300bp paired-end reads. Microfluidic PCR in the Fluidigm Access Array, downstream quality control and assurance, and Illumina sequencing was performed in the University of Idaho Institute for Bioinformatics and Evolutionary Studies Genomics Resources Core facility.

Illumina sequence data processing

Reads from the Illumina MiSeq run were demultiplexed for each sample, using the sample-specific dual barcode combinations, and chloroplast region or nuclear locus, using the target specific primers with the python application dbcAmplicons (https://github.com/msettles/dbcAmplicons), following [79–81]. The dbcAmplicons application was used to preprocess the Illumina double-barcoded amplicons generated in the manner described above. Specifically, the application was used to identify the barcode pairs (edit-distance < = 1bp), identify and remove the target specific primer sequences, associate barcode and primer sequences to a sample and annotate each read with the resulting information for downstream processing. To maximize the number of amplicons recovered for each sample and each DNA region, the dbcAmplicons application allows for target specific primer matching errors less than or equal to four (default, determined by Levenshtein distance), as long as the final bases (4 by default) of the 3' end are exact matches, thus yielding firm ends. Representative sequences for each amplicon were then identified using the reduce_amplicons R script (https://github.com/msettles/dbcAmplicons/blob/master/scripts/R/reduce_amplicons.R). This postprocessing pipeline allows for an optional trimming step, where the number of trimmed bases can be set independently for both the forward and reverse reads (using the "—trim-1 and—trim-2" flags in the reduce_amplicons R script). This step is especially useful in instances of high sequencing error rates towards the end of a read, particularly in read 2 on the MiSeq platform. After experimenting with different trimming values and evaluating the reads visually, we decided to trim 75 bp and 150 bp of our forward and reverse reads, respectively. We found that this amount of trimming produced a good balance between high quality sequence and sufficient data in order to draw supported phylogenetic conclusions. After any read trimming, paired reads that overlap by at least 10 bp (default) are joined into a single continuous sequence using FLASH2 (https://github.com/dstreett/FLASH2).

Finally, after reads are first reduced to the most often occurring amplicon length variant (all other reads are not included in the consensus calculation), for every sample at every locus, the reduce_amplicons R script produces consensus sequences (using the "-p" flag in the reduce_amplicons R script) with IUPAC ambiguity codes (-p ambiguity) for individual sites represented by more than one base when each variant is present in at least 5 reads and 5% of the

total number of reads (these thresholds are adjustable using the "-s" and "–f" flags, respectively), or without ambiguities ("-p consensus"). For allele recovery ("-p occurrence"), reads are reduced to identical pairs or joined paired reads (candidate alleles) and counted. If the candidate allele count represents at least 5% of the total number of reads and contains at least 5 reads (again, adjustable by the user), that amplicon is retained as a filtered candidate allele for the sample and target (Fig 2E). However, if both the sequencing depth and minimum total percentage thresholds are not met for any candidate allele, that sample specific target is discarded. Our allele-recovering method is based on the assumption that, post PCR, the original sample target alleles should dominate the candidate allele pool and reads containing common sequencing errors will be represented at much lower frequencies than actual biological variation. Trimming low quality 3' ends of reads, as described above, can improve the retention of candidate alleles.

Because the chloroplast genome is haploid, consensus sequences were generated for each of the 48 regions (using the "-p consensus" flag in the reduce_amplicons R script). In cases where the read 1 and read 2 consensus sequences did not overlap, the reads were concatenated into a single continuous sequence. Each region was aligned with MUSCLE v3.8.31 [82] in its default settings, alignments were cleaned with Phyutility v2.2.4 [83] at a 50 percent similarity threshold to minimize missing data due to ambiguous alignment sites, visually inspected in Geneious R6 v6.1.5 (Biomatters), and any misaligned or ambiguous sequences were discarded. Finally, the 48 chloroplast alignments were concatenated with Phyutility into a single locus.

To accommodate putative heterozygosity at nuclear loci, we generated consensus sequences with IUPAC ambiguity codes (using the "-p ambiguity" flag in the reduce_amplicons R script) for every sample at every locus, as well as individual allelic occurrences for each sample when applicable (using the "-p occurrence" flag in the reduce_amplicons R script). As with the chloroplast set, paired reads that did not overlap were concatenated, and each region was independently aligned using MUSCLE and cleaned with Phyutility at a 50 percent threshold. Because the ITS and ETS regions are physically linked in the rDNA repeat, theses regions were concatenated and treated as a single locus using Phyutility.

This data processing pipeline resulted in alignments for 47 independent nuclear loci–with allelic information, when relevant–and a concatenated chloroplast dataset. To compare alternative strategies for phylogenetic analyses, consensus sequences with IUPAC ambiguity codes for each of the 47 nuclear loci were concatenated into a single dataset (~13,500bp; the concatenated nuclear dataset), and the concatenated nuclear dataset and concatenated chloroplast dataset were combined into a single alignment of more than 40,500bp after cleaning.

Phylogenetic analyses

The concatenated chloroplast dataset was analyzed with PartitionFinder [84,85] to find the best partitioning scheme while also identifying the best-fit model of sequence evolution for each possible partition. Using these partitioning schemes and models of sequence evolution, we conducted maximum likelihood (ML) analyses as implemented in GARLI v2.0.1019 [86] with ten independent runs, each with 50 nonparametric bootstrap replicates. Bootstrap support was assessed with the program SumTrees v3.3.1 of the DendroPy v3.12.0 package [87]. Likewise, we analyzed the dataset in a Bayesian framework as implemented in MrBayes v3.2.1 [88] with the individual parameters unlinked across the data partitions. We ran two independent runs with four Markov chains each using default priors and heating values. Independent runs were started from a randomly generated tree and were sampled every 1000 generations. Convergence of the chains was determined by analyzing the plots of all parameters and the–lnL using Tracer v.1.5 [89]. Stationarity was assumed when all parameters values and the–lnL had

stabilized; the likelihoods of independent runs were considered indistinguishable when the average standard deviation of split frequencies was < 0.001. A consensus trees was obtained using the sumt command in MrBayes.

The nuclear dataset was analyzed in multiple ways. First, we inferred individual gene trees for each locus using RAxML v.8.0.3 [90] to ensure that the each locus was indeed single copy. Second, we analyzed the concatenated nuclear dataset with RAxML with no topological restrictions. Third, a second analysis of the nuclear concatenated dataset with RAxML, but this time constraining the topology to make every species monophyletic (concatenation with monophyly constraints; CMC) [91]. Although not a formal coalescent-based species tree method, comparisons of the CMC approach to coalescent-based species tree approaches have found them comparable and potentially the least sensitive to taxonomic sampling [91]. Furthermore, the CMC approach is a much more computationally tractable approach than currently available coalescent-based species tree approaches on datasets of the size that we are analyzing here—but see [92] for a potentially scalable approach. Finally, the combined dataset (chloroplast and nuclear loci)—with and without monophyly constraints—was analyzed with RAxML. Although we understand the importance of analyzing this type of dataset in a coalescent framework, the scope of the present study is not to infer the species tree or make systematic conclusions for the clade in question, which is the focus of ongoing and future work, but rather to demonstrate the efficacy of this targeted approach for generating large phylogenetic datasets using microfluidic PCR in the Fluidigm Access Array and HTS.

Results

Preliminary Data Acquisition

Low coverage genomes were sequenced for genome skimming from four samples representing three species of *Bartsia*: *B. pedicularoides* Benth. (two samples), *B. santolinifolia* (Kunth) Benth., and *B. serrata* Molau. *Bartsia pedicularoides* 1 yielded ~51.6 million 100 bp single-end reads (sequenced at the Vincent J. Coates Genomics Sequencing Laboratory at the University of California, Berkeley). The other three samples (sequenced at the Genomics Core Facility at the University of Oregon) yielded an average of ~51.4 million 100bp paired-end reads per library (Table 1; GenBank Sequence Read Archive (SRA): SRR2045582, SRR2045585, SRR2045588, SRR2045589). Seqyclean v 1.8.10 (https://bitbucket.org/izhbannikov/seqyclean) processing resulted in ~46.8 million reads for *B. pedicularoides* 1, ~52.9 million for *B. pedicularoides* 2, ~46.4 million for *B. santolinifolia*, and ~65.1 million for *B. serrata*. The plastomes assembled with the Alignreads pipeline from these samples had an average sequencing depth of 995x (Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592.

Chloroplast and Nuclear Primer Design and Validation

The *Bartsia* chloroplast alignment of six plastomes (including only one copy of the inverted repeat) had a length of ~125kb (Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592). From this alignment, we were able to design a total of 74 primer pairs that spanned the entire plastome. Following primer validation, 53 primer pairs (72% success rate) passed the validation criteria. From these, a final set of the most variable 48 primer combinations was chosen, with an average sequence variability of 2.7% (0.8%– 7.5%) (S1 Table).

For the nuclear set, we identified 51 PPR and 762 COSII loci that matched our criteria for further primer design (i.e., enough reads matching from low-coverage genomic data to attempt primer design). The nuclear rDNA, *PHOT1*, and *PHOT2* alignments (aligned lengths of 6,711bp, 578 bp, 1,272bp—Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592) all contained multiple places to design primers based on our criteria. A total of 188 primer

pairs were designed from all datasets (S1 Table). From those, 44 belonged to the PPR gene family, 130 to COSII, 8 to the nrDNA repeat, 3 to *PHOT1*, and 3 to *PHOT2*. After validation, 26 primer pairs were chosen for PPR (59.1% success rate), 25 for COSII (19.2% success rate), 7 for the nuclear rDNA (87.5% success rate), 0 for *PHOT1* (0% success rate), and 3 for *PHOT2* (100% success rate). Finally, the primer pair amplifying the longest target sequence was chosen among the various possibilities for the nuclear rDNA and *PHOT2* loci.

Sampling, Microfluidic PCR and Sequencing

To fully capture the morphological, genetic, and geographical diversity of the Neobartsia clade, and to demonstrate the efficiency of this approach for molecular phylogenetic studies at low-taxonomic levels, we included 576 samples (S3 Table) that represented 46 species of the clade and 28 related taxa as outgroups (included primarily to evaluate how far outside of the target group primers would successfully amplify targeted loci). Microfluidic amplification of the samples using 24 48.48 Access Array Integrated Fluidic Circuits (Fluidigm) resulted in up to 96 amplicons per sample (a total of 55,296 microfluidic reactions). After pooling and normalizing amplicons for each sample, pools were sequenced on an Illumina MiSeq platform with the Reagent Kit version 3, yielding ~20.3 million 300 bp paired-end reads. Raw reads were deposited in the GenBank Sequence Read Archive (SRA SRP058302).

Data Processing

Following processing with dbcAmplicons, ~16.9 million reads (77.7%) were sufficiently matched to both barcodes (sample specific) and primers (target specific). Discarded reads (~4.5 million reads) were a combination of PhiX Control v3 (Illumina; ~3.2 million reads, or 15%) and reads that did not pass our criteria for matching both barcodes and the primers (~1.3 million reads, or 7.3%).

Chloroplast set. Of the 576 samples used in this study, 528 (91.7%) amplified at least one chloroplast DNA amplicon and 486 (84.0%) produced more than 40 amplicons (>21,300bp) (Fig 4A). The majority of the samples that did not amplify efficiently belonged to outgroup taxa that are distantly related to the Neobartsia clade, suggesting that the designed primers were too specific to work efficiently outside this clade. This highlights the importance of careful primer design and validation that is in line with the taxonomic breath of the intended study. Because our primary focus was on the Neobartsia clade, and primers were designed with plastomes from this clade, these results were not surprising. Following processing with reduce_amplicons, multiple sequence alignment, and cleaning, the final chloroplast dataset included 486 samples and had an aligned length of ~25,300bp. The majority of the samples belonged to the Neobartsia clade (472), with the remaining 12 samples representing the three most closely related species (*Bellardia trixago* (L.) All., *Bellardia viscosa* (L.) Fisch. & C.A. Mey, and *Parentucellia latifolia* (L.) Caruel).

Nuclear set. For the nuclear DNA set, 47 out the 48 regions were amplified from the majority of samples (Fig 4B), and only one region did not satisfy our amplicon demultiplexing criteria, i.e., did not meet both barcode and primer matching criteria. From these 47 regions, we were able to recover sequence data from an average of 443 samples (76.0%), ranging from 520 (90.3%) to 318 (55.2%) depending on the region. Allele recovery (from processing with reduce_amplicons) resulted in 85.41% of these samples having one allele, 13.67% having two, 0.82% having three, and 0.10% having four alleles (S4 Table). After summarizing across every sample for each species, 17 taxa presented a combination of one and two alleles across loci (diploid as the minimum ploidy level), while the remaining 33 taxa had a combination of three and four alleles (tetraploid as the minimum ploidy level) (S4 Table). Following alignment clean

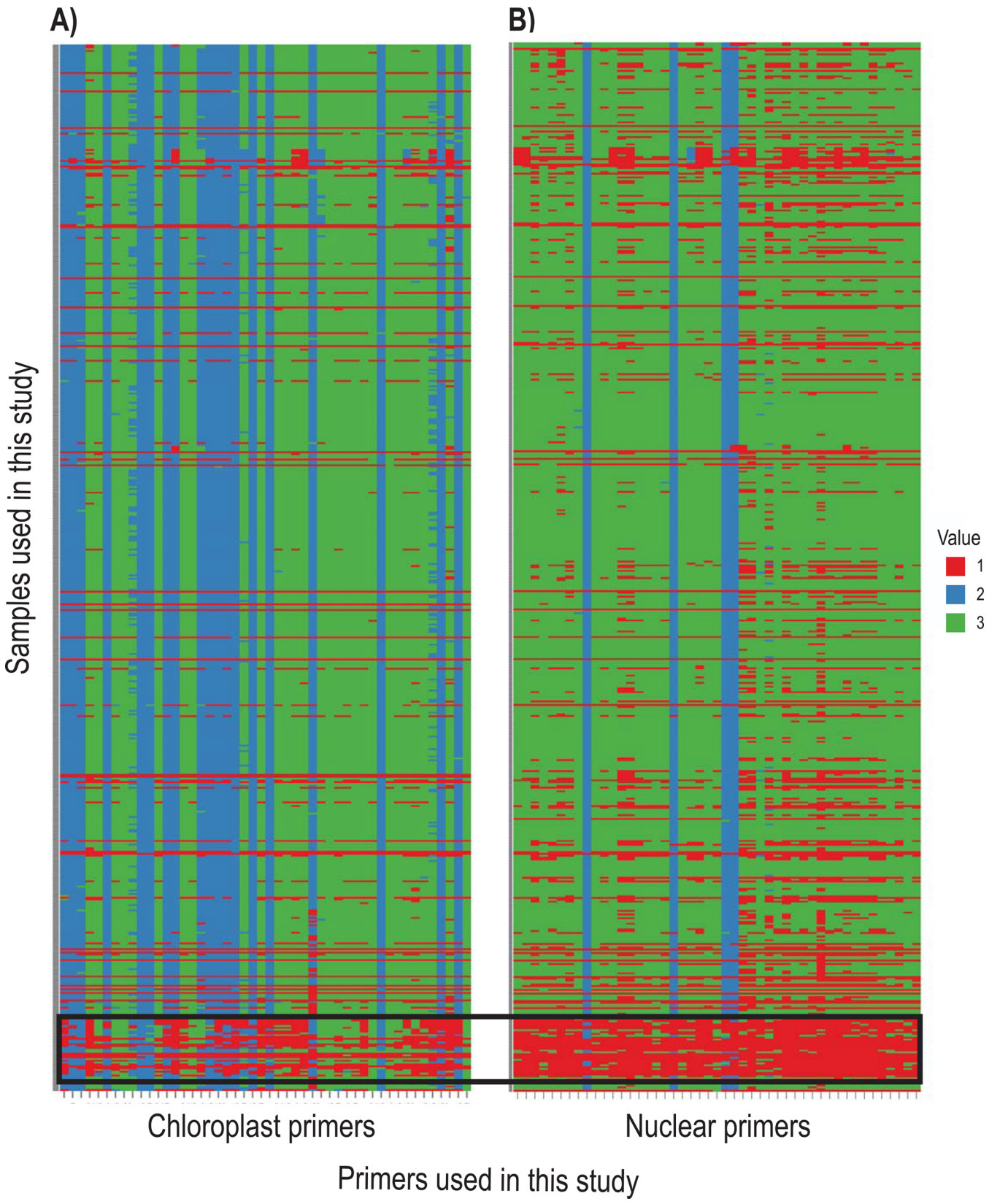
A)
B)
Samples used in this study
Value
1
2
3
Chloroplast primers
Nuclear primers
Primers used in this study

Fig 4. Bird's-eye view heat map showing the target coverage (sequencing success) for each sample. Target coverage or success (horizontal) for each sample (vertical) in (A) the chloroplast set and in (B) the nuclear set. These regions were processed to generate consensus sequences with ambiguities. Red indicates no amplicon was recovered either due to lack of successful amplification or mismatching of the barcodes and primers (see text for more details). Blue indicates that the forward and reverse paired reads overlapped by at least 10bp and were joined. Green indicates that the forward and reverse reads did not overlap. The group of 'failed' samples inside the box along the bottom of each panel are distantly-related taxa that were not included in primer design or phylogenetic analyses (see text).

doi:10.1371/journal.pone.0148203.g004

up of the consensus sequences with ambiguities (from processing with reduce_amplicons) with the program Phyutility, the nuclear gene regions had an average aligned length of 332bp (from 267 to 459bp). Preliminary ML phylogenetic analyses in RAxML revealed three loci where paralogous copies were amplified, and another three loci that were too variable to be unambiguously aligned—due to off-target amplification and/or spurious amplification. These six loci were removed prior to downstream phylogenetic analyses, resulting in a concatenated nuclear dataset of 41 regions (nrDNA plus 40 single copy nuclear gene regions) with an aligned, cleaned length of ~13,500bp from 363 samples. Finally, we constructed a combined concatenated matrix (nuclear and chloroplast) with an aligned, cleaned length of ~40,500bp, including 349 samples (all sequences deposited in the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592).

Phylogenetic analyses

Individual nuclear gene trees were largely unresolved, and paralogous loci were identified in three of the 48 loci. The concatenated chloroplast dataset was first analyzed with PartitionFinder to identify the best partitioning scheme, while also selecting for the best-fit model of sequence evolution for each possible partition. This analysis resulted in 11 partitions with the following models of sequence evolution: K81uf+I+G, K81uf+I, TrN+I+G, TVM+I+G, F81, K80+I, TVMef+I+G, TVMef+I+G, F81, TVM+I+G, TVM+I+G (data partitions and corresponding models can be found in Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592). Analyses in ML and Bayesian frameworks, in GARLI and MrBayes, respectively, resulted in the same overall phylogenetic relationships among the samples. The same is true for the nuclear concatenated, the combined (nuclear and chloroplast), and the concatenation with monophyly constraints (CMC) analyses, which resulted in the same overall relationships among species. Because every analysis resulted in a very similar tree, the results and discussion will be based on the combined concatenated dataset, with and without constraints (all tree files deposited in the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.fh592).

Several clades that correspond to relationships between outgroup taxa and South American *Bartsia* species were recovered with 100% bootstrap support (BS) and a posterior probability (PP) of 1.0. First, all individuals included for the outgroup taxa *Bellardia trixago*, *Bellardia viscosa*, and *Parentucellia latifolia* were reciprocally monophyletic. Second, all accessions of South American *Bartsia* species (the large majority of the sampling in this study) were monophyletic, and *P. latifolia* is the sister group to this clade. Third, *Bellardia trixago* and *B. viscosa* formed a distinct clade, and this clade is sister to the *P. latifolia* plus the Neobartsia clade (Fig 5).

Support for backbone relationships in the Neobartsia clade is low, and thus, very few systematic conclusions can be made at this point. First, we did not recover four monophyletic groups corresponding to the four morphological sections (sensu [93]). However, we did find several clades comprised of a few species from the same morphological section, albeit with moderate support (Fig 5). Furthermore, individuals of most of the species were recovered in multiple different clades, and in fact only two species (*B. filiformis* Wedd. and *B. adenophylla* Molau) were monophyletic. This is not surprising, given the fact that the Neobartsia clade has been shown to be a recent and rapid radiation [18], and processes like coalescent stochasticity,

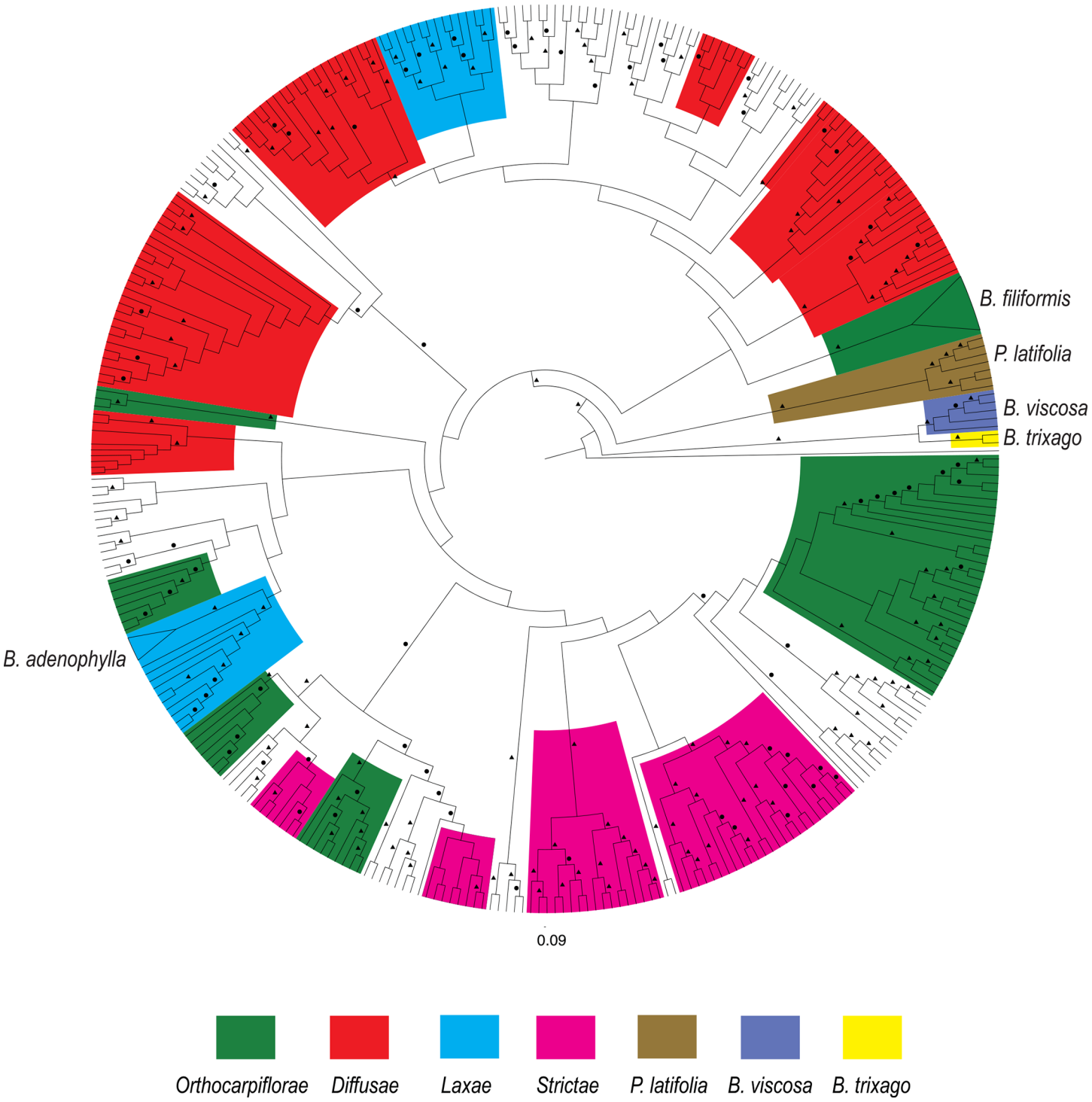

Fig 5. Phylogram of the phylogenetic relationships of the South American *Bartsia* species and closely related taxa. Phylogram based on the maximum likelihood analysis in RAxML on the combined (chloroplast and nuclear) nucleotide unconstrained dataset (~40,500bp and 349 samples). Maximum likelihood bootstrap support (BS) is represented above the branches by a circle (>50 BS) or a triangle (>85 BS). Clades comprised of a few species from the same morphological section of the South American species (sensu [93]) have been colored, as well as the three closely related taxa (*Bellardia trixago* (L.) All., *Bellardia viscosa* Fisch. & C.A. Mey, and *Parentucellia latifolia* (L.) Caruel. The only two taxa that were recovered as monophyletic are indicated on the tree and their clade collapsed in a triangle.

doi:10.1371/journal.pone.0148203.g005

hybridization, and introgression may be playing a large role in the evolution of these taxa. It will be necessary to conduct species tree (e.g., coalescent-based) and network analyses to confidently elucidate relationships among species in this clade.

The CMC 'pseudo-species tree' analysis recovered most of the same clades containing species within the same taxonomic sections. Interestingly, enforcing monophyly of individuals belonging to named species reduced BS support of the backbone relationships even further (Fig 6), indicating that sequences from some of the individuals that were constrained to be monophyletic clearly violate this assumption.

Discussion

Regardless of the HTS method used, it is clear that the field of phylogenetics, and the study of evolution in general, are quickly migrating towards larger and larger molecular datasets. The ability to produce more and longer reads, as well as reduced sequencing costs and increased computing power are making this transition easier and faster. Here, we presented an approach to generate large multilocus, homogeneously distributed, and targeted datasets using microfluidic PCR in the Fluidigm Access Array and HTS.

One of the main advantages of this approach is circumventing the necessity to construct expensive genomic shotgun libraries for each sample in the experiment. This step greatly increases the time and cost of any HTS approach, effectively reducing the sample size possible for any experiment. The approach presented here takes advantage of a four-primer PCR amplification to efficiently tag multiple genomic targets with sample specific barcodes and HTS adapters. By doing so, the resulting amplicons are ready to be sequenced following standard pooling and quality control. Furthermore, the use of a sample-specific dual barcoding strategy allows for a high level of multiplexing with far fewer PCR primers. Commonly used commercial barcoding kits currently offer either 96 (NEXTflex DNA Barcode kit; Bioo Scientific, Austin, Texas, USA) or 386 barcodes (Fluidigm), but we are able to theoretically multiplex up to 1,152 samples with only 72 barcoded primers (48 forward and 24 reverse), although only 576 samples were used in this study. This expands the possibilities during experimental design and takes full advantage of the yield of current HTS platforms, while maintaining low upfront costs.

An additional technical advantage of this approach is the high throughput achieved with smaller amounts of DNA, reagents, and labor. A commercially available platform–the Fluidigm Access Array System–facilitates simultaneous amplification of 48 samples with 48 distinct primer pairs (2,304 reactions) using only 15 U of *Taq* polymerase and 1μL of 30-60ng/μL genomic DNA per sample. By conducting a simultaneous four-primer reaction, one avoids the necessity of performing multiple rounds of manual PCR to incorporate barcodes and adapters–a limitation of the Targeted Amplicon Sequencing (TAS) strategy of Bybee et al. [43]. For example, following the TAS approach, to produce tagged amplicons for the 576 samples and 96 gene regions targeted in this study, it would have required >1,100 96-well plates (or >275 384-well plates) of PCR to produce the same number of barcoded amplicons. While the TAS approach does allow for more flexibility in terms of primer design, i.e., primer annealing temperatures do not need to all be the same, and it may be possible to incorporate ambiguities into primer design, to take advantage of performing PCR in plates, significant PCR optimization would also need to be performed. Nevertheless, with high levels of taxon-by-gene region samplings, TAS becomes unpractical. Using the microfluidic PCR approach to amplicon generation and tagging, only 24 Fluidigm Access Arrays were necessary to amplify and tag the 55,296 amplicons. While studies with smaller sampling strategies (e.g., [43]) would likely benefit from the two-reaction TAS approach, with the ever increasing sequencing read length and

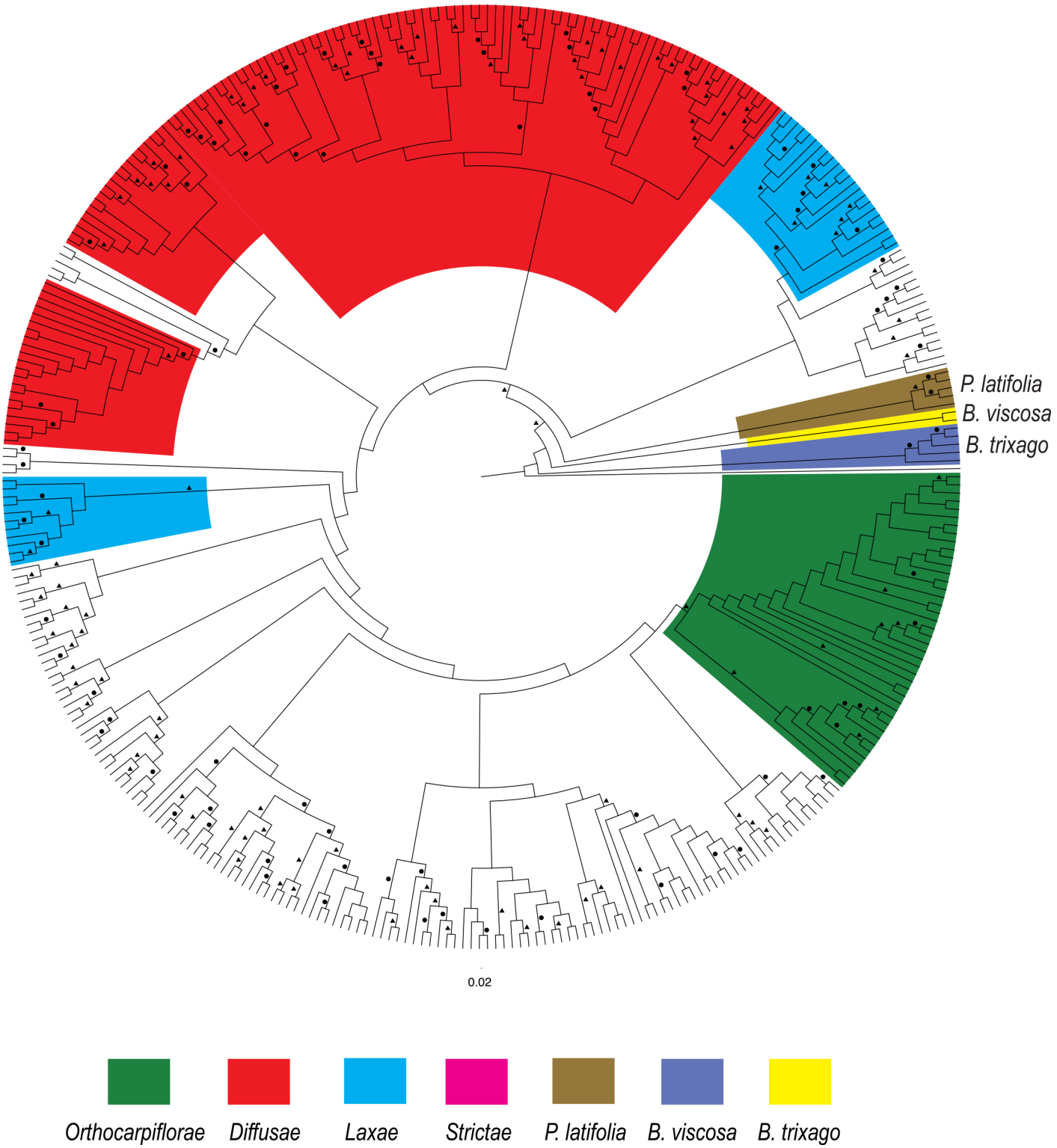

Fig 6. Phylogram of the phylogenetic relationships of the South American *Bartsia* species and closely related taxa. Phylogram based on the maximum likelihood analysis in RAxML on the combined (chloroplast and nuclear) dataset using the concatenation with monophyly constraints (CMC) approach on the combined dataset (~40,500bp and 349 samples). Maximum likelihood bootstrap support (BS) is represented above the branches by a circle (>50 BS) or a triangle (>85 BS). Clades comprised of a few species from the same morphological section of the South American species (sensu [93]) have been colored, as well as the three closely related taxa (*Bellardia trixago* (L.) All., *Bellardia viscosa* Fisch. & C.A. Mey, and *Parentucellia latifolia* (L.) Caruel.

doi:10.1371/journal.pone.0148203.g006

throughput of HTS platforms, the microfluidic PCR approach presented here allows researchers performing large phylogenetic or population genetic studies to maximize data collection using HTS techniques.

Perhaps the most significant advantage of generating phylogenetic datasets using a targeted amplicon approach is that using our 'read frequency' approach (Fig 2E), it is possible to distinguish individual alleles at heterozygous nuclear loci without requiring additional assembly or mapping steps. Because targeted loci were specifically amplified one locus per reaction per sample, when paired-end sequences from each specific barcode and primer combination are identified, heterozygous loci in diploid species appear as high frequency amplicons, allowing us to straightforwardly determine both alleles. This contrasts with methods in which genomic DNA is sheared and selected for a specific length, e.g., sequence capture and GBS/RADseq, where an assembly and/or reference-based mapping strategy is necessary to compile consensus sequences. These additional steps introduce known problems associated with the large number of de novo assemblers and mappers, e.g., varying numbers of resulting contigs, the performance of the assembler or mapper based on the error model of the sequencing platform used, and computational power and time (see [94] for further details). More importantly, most phylogenomic studies that have included nuclear data generated using HTS techniques like these, have ignored the challenge that heterozygosity presents by using ambiguity coding [1] or by selecting only one allele and discarding others [2,54]. For phylogenetic and population genetic studies using modern coalescent-based approaches, allelic information is important when reconstructing the evolutionary histories of the genes sampled, and in cases where there is a large amount of coalescent stochasticity and/or gene flow, discarding or masking allelic information may be misleading. In a population genetic study of the North American tiger salamander (*Ambystoma tigrinum* Green) species complex, O'Neill et al. [44] used a haplotype phasing strategy to computationally determine individual alleles. For statistical phasing approaches, the number of individuals per population present in a sample is a critical factor in determining how well haplotype phase can be estimated [95], and therefore may only be appropriate for the deep population-level sampling in population genetic studies such as this, but will likely not be useful for most phylogenetic studies.

Furthermore, polyploidy is common in many plant groups, as well as in select groups of insects [96], fish [97], amphibians [98], and reptiles [99], and therefore, this is an important consideration that complicates the issue of heterozygosity even more. For example, a tetraploid species may be heterozygous at both homeologous loci, and in this scenario, one would expect to identify four sets of reads with high frequencies dominating the amplicon pool. Likewise, a tetraploid may be homozygous at one homeolog and heterozygous at the other; in this case, we would expect to identify three-sets of high frequency reads dominating the amplicon pool. Finally, for many species of plants, ploidy levels are often unknown, or variable within a species (e.g., [100]), and material appropriate for determining ploidy via chromosome counts and/or flow cytometry is not available. While at any one nuclear locus, a polyploid species may or may not be heterozygous at one or more of the homeologs, by having multiple nuclear loci in one experiment, it may be possible to calculate the frequencies of alleles across all loci and not only recover individual alleles, but potentially estimate ploidy level–or at least a minimum ploidy level depending on levels of heterozygosity. In plants, this could be especially useful for evaluating hypothesized allopolyploid events, as well as the evolutionary and ecological consequences of polyploidy when these data are analyzed in a comparative phylogenetic context.

Within the Neobartsia clade, only 23 of 45 species have published chromosome numbers, and seven of these have been characterized as tetraploids based on these counts (S4 Table) [93]. Likewise, chromosome counts have been published for the European *Bartsia alpina* (diploid), and the Mediterranean species *Bellardia viscosa* (tetraploid) and *Parentucellia latifolia*

(tetraploid) [93]. The reduce_amplicons pipeline employed here recovered at least three alleles in one locus for six of these species, suggesting that their minimum ploidy level might be tetraploid. Although we only recovered one or two alleles in the remaining four species (suggesting that the taxa may be diploid), it is important to keep in mind that these species comprise a very recently diverged clade, and most species occur in small, isolated populations [18], where low sequence divergence, autopolyploidy, and homozygosity may mask true ploidy levels. Something to keep in mind, however, is the number of samples (individuals) per species that were recovered as having more than two alleles. For some of these species, e.g., *Bartsia camporum* Diels, *Bartsia serrata* Molau, and *Bellardia viscosa*, the majority of individuals (more than 65%) presented at least three alleles—supporting that these taxa are indeed tetraploid. On the other hand, species such as *Bartsia pyricarpa* Molau, *Bartsia pedicularoides* Benth., and *Bartsia patens* Benth.—all of which are suggested to be tetraploid based on chromosome counts (S4 Table) [93]—were recovered with most of their individuals presenting one or two alleles and only 2% to 15% of the individuals having at least three alleles. This highlights the necessity of including more than one individual from more than one population when assessing levels of ploidy. To fully investigate the utility of this approach for bioinformatically estimating ploidy levels, chromosome counts and/or flow cytometry data from these same samples would be necessary, but this was beyond the scope of this study. Nevertheless, these results are encouraging as they open the door for future comparisons between 'bioinformatically karyotyped' samples and traditional ploidy estimation experiments.

A notable limitation of the microfluidic approach that we present here is the necessity to design a relatively large number of target-specific primers to fill a Fluidigm Access Array. To do this in an efficient manner, it is necessary to first have at least some genomic resources available for your clade of interest. In our case, we had both whole plastome sequences, as well as low-coverage genomic data for a small, but representative, set of species. With these preliminary data in mind, we developed an effective approach for primer design that allowed us to target 1) the most variable regions of the plastome in the Neobartsia clade, 2) the ITS [101] and ETS [102] regions of the nrDNA repeat that have been used extensively at the interspecific level in plants, 3) multiple, independent nuclear genes from the intronless PPR gene set developed by Yuan et al. [64] and shown to be phylogenetically informative at the family level in Verbenaceae [103] and at the subfamily level in Campanuloideae [66], and 4) intron-spanning regions from within the COSII gene set developed by Wu et al. [67] and used within Orobanchaceae [68], and the Phototropin 2 gene used at the interspecific level in *Glandularia*, *Junellia* and *Verbena* in the Verbenaceae [71]. By specifically targeting the variable regions of the plastome, commonly sequenced regions of the nrDNA repeat (e.g., ITS, ETS), and multiple independent nuclear loci that range from more conserved (e.g., intronless PPR genes) to rapidly evolving nuclear introns (e.g., COSII), we were able to assemble a large, multilocus, homogeneously distributed dataset with high levels of intraspecific sampling for a complete clade of recently diverged Andean plants. Although we took, and advocate, the genome skimming approach [37,55] to develop the necessary genomic resources used here for primer development, there are a growing number of publically available databases that could also be used–e.g., for plants, Phytozome (http://www.phytozome.net), One Thousand Plants Project (1KP; http://onekp.com), IntrEST [104], and Genome 10k for animals [105]–as well as a recently published bioinformatics pipeline for identifying single-copy nuclear genes and designing target specific primers for phylogenomic analyses using existing transcriptome data [106].

The Neobartsia Clade

This is the first time that the interspecific relationships of the species in the Neobartsia clade have been studied with such deep taxonomic sampling and with so much molecular data. From our results, it is clear that in order to fully understand the evolutionary history of the clade, phylogenetic species tree methods that explicitly incorporate mechanisms that lead to gene tree-species tree discordance (e.g., coalescent stochasticity, divergence with gene flow) are needed. However, these detailed analyses are beyond the scope of this study (which is focused on data collection approaches), and therefore, the phylogenetic results presented here are preliminary. Nevertheless, the monophyly of the group is highly supported by all analyses, and is in agreement with recent biogeographic study of the clade [18]. Interspecific relationships, however, have very little support–a pattern commonly seen in rapid radiations like this–and only two species were found to be monophyletic. These two species are taxa with small and restricted geographic distributions with likely small effective population sizes. Given that the time to coalescence is directly linked to effective population size [34], it is not unexpected that individuals from these species were monophyletic in our unconstrained analyses. Conversely, when we look at species with a large geographic distributions, and larger effective population sizes (e.g., *B. pedicularoides* Benth.), we see that the individuals are recovered in multiple different groups across the tree (Fig 5).

Enforcing the monophyly of species has been used as a 'pseudo-species tree' method with good results [91], and some relationships recovered here make evolutionary sense–*Bartsia sericea* Molau and *B. crisafullii* N. Holmgren were recovered as sister species (Fig 5) with high support. Both species are extremely similar morphologically, only differing in their life history and ploidy level (perennial vs. annual, and diploid vs. tetraploid, respectively). However, in some instances, enforcing monophyly of the species reduced the BS support of deeper branches. There are several possibilities for this result, including violations of our *a priori* species designations (i.e., incorrect species delimitations, cryptic species, etc.), severe coalescent stochasticity, ancient and/or contemporary introgression, and/or hybrid speciation. Given the recent and rapid nature of this Andean diversification, each of these (or any combination of) are potential mechanisms that increase the phylogenetic complexity of this group, and incorporating these processes into species tree estimation in this clade is the focus of ongoing systematic studies using these data.

Conclusion

We presented an approach to generate large multilocus phylogenomic datasets for a large number of samples and species using microfluidic PCR in the Fluidigm Access Array and HTS. This approach allows for more control in targeting informative regions of the genome to be sequenced, resulting in datasets that are tailored to address the specific questions being asked, and that are orthologous across samples. Additionally, this method is both cost effective and time efficient, as it does not require genomic shotgun libraries to be constructed for every sample, and takes full advantage of the large multiplexing capabilities of HTS platforms. As a case study, we focused on 576 samples of the Neobartsia clade, amplifying and sequencing the 48 most variable regions of the chloroplast genome, as well as 48 nuclear gene regions representing a range of both coding and non-coding data. This targeted strategy for the collection of multilocus data for phylogenetic studies provided us with a large, but modest, set of loci that will be appropriate for sophisticated species tree inference methods (e.g., coalescent-based, networks, concordance analyses), and provided us with the first species level phylogeny for the Neobartsia clade. Furthermore, the bioinformatic approaches employed here allowed for the recovery of individual alleles in heterozygote individuals (without the need for statistical

phasing), and opened the door for the exploration of bioinformatic approaches to estimating ploidy levels—an important and often overlooked consideration at low taxonomic levels.

Supporting Information

S1 Table. List of the primers designed and used in this study. Sequences are written in 5'-3' direction. Variability of the chloroplast regions is based on nucleotides from six plastomes and is given in substitution per site. Primer names contain, whenever possible, the gene in which they are anchored and an approximate location in our six plastome alignment.
(XLSX)

S2 Table. Sequences for conserved sequences 1 and 2, barcodes, and sequencing adapters. The CS1 and CS2 sequences were obtained from the Fluidigm Access Array System protocol, whereas the barcode pairs were custom designed to allow for dual barcoding, in order to dramatically increases the number of samples that can be multiplexed in one sequencing run (see text for a detailed explanation). The sequencing adapters are the Illumina standard adapters, in our case the P5 and P7 adapters.
(XLSX)

S3 Table. Samples used in this study. Herbarium codes follow the Index Herbariorum [107].
(XLSX)

S4 Table. Allele occurrences found for each species in this study. Table A. Summary all Samples per Species. Summary of the number of alleles in every sample for each species, and a percentage of how many of these were recovered with more than two alleles. An estimated ploidy level is also given. A reference table of *Bartsia* species and closely related taxa with published chromosome counts from Molau (1990) is also included. **Table B**. Summary for each Sample. Summary of the number of alleles found in each locus for every sample, and a percentage of how many loci had this specific number of alleles. **Table C**. Allele Counts for every Sample. Raw data of the number of alleles found in each locus for every sample.
(XLSX)

Acknowledgments

We would like to thank Dan New, Tamara Max, and the Institute for Bioinformatics and Evolutionary Studies (IBEST—NIH/NCRR P20RR16448 and P20RR016454) at the University of Idaho for technical assistance and bioinformatic resources. Jonathan Eastman was instrumental in R scripting and data processing. Jack Sullivan, Luke J. Harmon, Eric H. Roalson, Diego F. Morales-Briones, Matthew W. Pennell, Tracy C. Peterson, and two anonymous reviewers offered helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: SUC DCT MLS. Performed the experiments: SUC. Analyzed the data: SUC MLS. Contributed reagents/materials/analysis tools: SUC DCT. Wrote the paper: SUC DCT MLS. Designed the software used in analysis: MLS.

References

1. Lemmon AR, Emme SA, Lemmon EM. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology. 2012; 61: 727–744. doi: 10.1093/sysbio/sys049 PMID: 22605266

2. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology. 2012; 61: 717–726. doi: 10.1093/sysbio/sys004 PMID: 22232343
3. Stull GW, Moore MJ, Mandala VS, Douglas NA, Kates H-R, Qi X, et al. A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes. Applications in Plant Sciences. BioOne; 2013; 1: 1–7. doi: 10.3732/apps.1200497
4. Walter K, Holcomb T, Januario T, Du P, Evangelista M, Kartha N, et al. DNA Methylation Profiling Defines Clinically Relevant Biological Subsets of Non-Small Cell Lung Cancer. Clinical Cancer Research. 2012; 18: 2360–2373. doi: 10.1158/1078-0432.CCR-11-2635-T PMID: 22261801
5. Gaedcke J, Grade M, Camps J, Sokilde R, Kaczkowski B, Schetter AJ, et al. The Rectal Cancer microRNAome—microRNA Expression in Rectal Cancer and Matched Normal Mucosa. Clinical Cancer Research. 2012; 18: 4919–4930. doi: 10.1158/1078-0432.CCR-12-0016 PMID: 22850566
6. Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA. Development and mapping of SNP assays in allotetraploid cotton. Theoretical and Applied Genetics. 2012; 124: 1201–1214. doi: 10.1007/s00122-011-1780-8 PMID: 22252442
7. Bhat S, Polanowski AM, Double MC, Jarman SN, Emslie KR. The Effect of Input DNA Copy Number on Genotype Call and Characterising SNP Markers in the Humpback Whale Genome Using a Nanofluidic Array. Liu Z, editor. PLoS ONE. 2012; 7: e39181. doi: 10.1371/journal.pone.0039181.t004 PMID: 22745712
8. Lu X, Wang L, Chen S, He L, Yang X, Shi Y, et al. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. Nature Publishing Group. Nature Publishing Group; 2012; 44: 890–894. doi: 10.1038/ng.2337
9. Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nature Cell Biology. Nature Publishing Group; 2013; 15: 363–372. doi: 10.1038/ncb2709 PMID: 23524953
10. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality inexpression and splicing in immune cells. Nature. Nature Publishing Group; 2013;: 1–5. doi: 10.1038/nature12172
11. Dominguez MH, Chattopadhyay PK, Ma S, Lamoreaux L, McDavid A, Finak G, et al. Journal of Immunological Methods. Journal of Immunological Methods. Elsevier B.V; 2013; 391: 133–145. doi: 10.1016/j.jim.2013.03.002 PMID: 23500781
12. Lohr JG, Stojanov P, Lawrence MS, Auclair D, Chapuy B, Sougnez C, et al. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. Proceedings of the National Academy of Sciences. National Acad Sciences; 2012; 109: 3879–3884. doi: 10.1073/pnas.1121343109/-/DCSupplemental
13. Moonsamy PV, Williams T, Bonella P, Holcomb CL, Höglund BN, Hillman G, et al. High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Array™ system for simplified amplicon library preparation. Tissue Antigens. 2013; 81: 141–149. doi: 10.1111/tan.12071 PMID: 23398507
14. Hermann-Bank ML, Skovgaard K, Stockmarr A, Larsen N, Mølbak L. The Gut Microbiotassay: a high-throughput qPCR approach combinable with next generation sequencing to study gut microbial diversity. BMC Genomics. 2013; 14: 788. doi: 10.1186/1471-2164-14-788 PMID: 24225361
15. Curk F, Ancillo G, Garcia-Lor A, Luro F, Perrier X, Jacquemoud-Collet J-P, et al. Next generation haplotyping to decipher nuclear genomic interspecific admixture in Citrus species: analysis of chromosome 2. BMC Genetics. 2014; 15: 152. doi: 10.1186/s12863-014-0152-1 PMID: 25544367
16. Gostel MR, Coy KA, Weeks A. Microfluidic PCR-based target enrichment: A case study in two rapid radiations of Commiphora (Burseraceae)from Madagascar. Wen J, Liu J, Ge S, Xiang Q-YJ, Zimmer EA, editors. Journal of Systematics and Evolution. 2015; 53: 411–431. doi: 10.1111/jse.12173
17. Godden GT, Jordon-Thaden IE, Chamala S, Crowl AA, García N, Germain-Aubrey CC, et al. Making next-generation sequencing work for you: approaches and practical considerations for marker development and phylogenetics. Plant Ecology & Diversity. 2012; 5: 427–450. doi: 10.1080/17550874.2012.745909
18. Uribe-Convers S, Tank DC. Shifts in diversification rates linked to biogeographic movement into new areas: An example of a recent radiation in the Andes. American Journal of Botany. 2015; 102: 1854–1869. doi: 10.3732/ajb.1500229 PMID: 26542843
19. Graham SW, Olmstead RG. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. American Journal of Botany. 2000; 87: 1712–1730. PMID: 11080123

20. Moore MJ, Bell CD, Soltis PS, Soltis DE. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proceedings of the National Academy of Sciences. 2007; 104: 19363–19368. doi: 10.1073/pnas.0708072104

21. Parks M, Cronn R, Liston A. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Bioloogy. 2009; 7: 84. doi: 10.1186/1741-7007-7-84

22. Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proceedings of the National Academy of Sciences. 2010; 107: 4623–4628. doi: 10.1073/pnas.0907801107

23. Downie SR, Palmer JD. Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis PS, Soltis DE, Doyle JJ, editors. Molecular Systematics of Plants. Chapman and Hall, New York, NY; 1992. pp. 14–35.

24. Chase MW, Soltis DE, Olmstead RG, Morgan D, Les DH, Mishler BD, et al. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcL. Annals of the Missouri Botanical Garden. JSTOR; 1993;: 528–580.

25. Moore MJ, Hassan N, Gitzendanner MA, Bruenn RA, Croley M, Vandeventer A, et al. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. International Journal of Plant Sciences. JSTOR; 2011; 172: 541–558.

26. Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, et al. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proceedings of the National Academy of Sciences. 2012; 109: 17519–17524. doi: 10.1073/pnas.1205818109

27. Uribe-Convers S, Duke JR, Moore MJ, Tank DC. A Long PCR–Based Approach for DNA Enrichment Prior to Next-Generation Sequencing for Systematic Studies. Applications in Plant Sciences. 2014; 2: 1300063. doi: 10.3732/apps.1300063

28. Wolfe KH, Li WH, Sharp PM. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proceedings of the National Academy of Sciences. 1987; 84: 9054–9058.

29. Wolfe KH, Sharp PM, Li W-H. Rates of synonymous substitution in plant nuclear genes. Journal of Molecular Evolution. Springer; 1989; 29: 208–211.

30. Drouin G, Daoud H, Xia J. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Molecular Phylogenetics and Evolution. 2008; 49: 827–831. doi: 10.1016/j.ympev.2008.09.009 PMID: 18838124

31. Njuguna W, Liston A, Cronn R, Ashman T-L, Bassil N. Insights into phylogeny, sex function and age of Fragaria based on whole chloroplast genome sequencing. Molecular Phylogenetics and Evolution. 2013; 66: 17–29. doi: 10.1016/j.ympev.2012.08.026 PMID: 22982444

32. Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evolutionary Biology. 2014; 14: 23. doi: 10.1186/1471-2148-14-23 PMID: 24533922

33. Maddison W. Gene trees in species trees. Systematic Biology. 1997; 46: 523–536.

34. Rannala B, Yang Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 2003; 164: 1645–1656. PMID: 12930768

35. Edwards SV. Is a new and general theory of molecular systematics emerging? Evolution. 2009; 63: 1–19. doi: 10.1111/j.1558-5646.2008.00549.x PMID: 19146594

36. Knowles LL, Kubatko LS. Estimating species trees: practical and theoretical aspects. Wiley-Blackwell; 2010.

37. Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. American Journal of Botany. 2012; 99: 349–364. doi: 10.3732/ajb.1100335 PMID: 22174336

38. Tennessen JA, Govindarajulu R, Liston A, Ashman T-L. Targeted sequence capture provides insight into genome structure and genetics of male sterility in a gynodioecious diploid strawberry, Fragaria vesca ssp. bracteata (Rosaceae). G3; Genes|Genomes|Genetics. 2013; 3: 1341–1351. doi: 10.1534/g3.113.006288 PMID: 23749450

39. Mandel JR, Dikow RB, Funk VA, Masalia RR, Staton SE, Kozik A, et al. A Target Enrichment Method for Gathering Phylogenetic Information from Hundreds of Loci: An Example from the Compositae. Applications in Plant Sciences. 2014; 2: 1300085. doi: 10.3732/apps.1300085

40. Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, et al. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. Applications in Plant Sciences. 2014; 2: 1400042. doi: 10.3732/apps.1400042.s1

41. Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, et al. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African Oxalis (Oxalidaceae). Molecular Ecology Resources. 2015. In Press doi: 10.1111/1755-0998.12487

42. Eaton DAR, Ree RH. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (Pedicularis: Orobanchaceae). Systematic Biology. 2013; 62: 689–706. doi: 10.1093/sysbio/syt032 PMID: 23652346

43. Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, et al. Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. Genome Biology and Evolution. 2011; 3: 1312–1323. doi: 10.1093/gbe/evr106 PMID: 22002916

44. O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, et al. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (Ambystoma tigrinum) species complex. Molecular Ecology. 2013; 22: 111–129. doi: 10.1111/mec.12049 PMID: 23062080

45. Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, et al. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Molecular Ecology. 2012; 22: 814–826. doi: 10.1111/j.1365-294X.2012.05730.x PMID: 22924870

46. Jones JC, Fan S, Franchini P, Schartl M, Meyer A. The evolutionary history of Xiphophorusfish and their sexually selected sword: a genome-wide approach using restriction site-associated DNA sequencing. Molecular Ecology. 2013; 22: 2986–3001. doi: 10.1111/mec.12269 PMID: 23551333

47. Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G, et al. Empirical Assessment of RAD Sequencing for Interspecific Phylogeny. Molecular Biology and Evolution. 2014; 31: 1272–1274. doi: 10.1093/molbev/msu063 PMID: 24497030

48. McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Research. 2012; 22: 746–754. doi: 10.1101/gr.125864.111 PMID: 22207614

49. Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biology Letters. 2012; 8: 783–786. doi: 10.1098/rsbl.2012.0331 PMID: 22593086

50. McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. Alvarez N, editor. PLoS ONE. 2013; 8: e54848. doi: 10.1371/journal.pone.0054848.s004 PMID: 23382987

51. Faircloth BC, Sorenson L, Santini F, Alfaro ME. A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). Moreau CS, editor. PLoS ONE. 2013; 8: e65923. doi: 10.1371/journal.pone.0065923.s001 PMID: 23824177

52. Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. BMC Genomics; 2012; 13: 1–1. doi: 10.1186/1471-2164-13-403 PMID: 22900609

53. Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. Unlocking the vault: next-generation museum population genomics. Molecular Ecology. 2013; 22: 6018–6032. doi: 10.1111/mec.12516 PMID: 24118668

54. Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Systematic Biology. Oxford University Press; 2014; 63: 83–95. doi: 10.1093/sysbio/syt061 PMID: 24021724

55. Straub SCK, Fishbein M, Livshultz T, al E. Building a model: developing genomic resources for common milkweed (Asclepias syriaca) with low coverage genome sequencing. BMC Genomics. 2011; 12. doi: 10.1186/1471-2164-12-211

56. Salamone I, Govindarajulu R, Falk S, Parks M, Liston A, Ashman T-L. Bioclimatic, ecological, and phenotypic intermediacy and high genetic admixture in a natural hybrid of octoploid strawberries. American Journal of Botany. 2013; 100: 939–950. doi: 10.3732/ajb.1200624 PMID: 23579477

57. Straub SCK, Cronn RC, Edwards C, Fishbein M, Liston A. Horizontal Transfer of DNA from the Mitochondrial to the Plastid Genome and Its Subsequent Evolution in Milkweeds (Apocynaceae). Genome Biology and Evolution. 2013; 5: 1872–1885. doi: 10.1093/gbe/evt140 PMID: 24029811

58. Malé P-JG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, et al. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Molecular Ecology Resources. 2014. doi: 10.1111/1755-0998.12246

59. Lemmon AR, Brown JM, Stanger-Hall K, Lemmon EM. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology. 2009; 58: 130–145. doi: 10.1093/sysbio/syp017 PMID: 20525573

60. Doyle JJ, Doyle JL. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin. 1987; 19: 11–15.

61. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013; 30: 772–780. doi: 10.1093/molbev/mst010 PMID: 23329690

62. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0, Available: http://wwwR-projectorg. 2013.

63. Kent WJ. BLAT––The BLAST-Like Alignment Tool. Genome Research. 2002; 12: 656–664. doi: 10.1101/gr.229202 PMID: 11932250

64. Yuan Y, Liu C, Marx H, Olmstead R. The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. New Phytologist. 2009; 182: 272–283. doi: 10.1111/j.1469-8137.2008.02739.x PMID: 19192190

65. Yuan Y-W, Liu C, Marx HE, Olmstead RG. An empirical demonstration of using pentatricopeptide repeat (PPR) genes as plant phylogenetic tools: phylogeny of Verbenaceae and the Verbena complex. Molecular Phylogenetics and Evolution. 2010; 54: 23–35. doi: 10.1016/j.ympev.2009.08.029 PMID: 19733248

66. Crowl AA, Mavrodiev E, Mansion G, Haberle R, Pistarino A, Kamari G, et al. Phylogeny of Campanuloideae (Campanulaceae) with Emphasis on the Utility of Nuclear Pentatricopeptide Repeat (PPR) Genes. Louis EJ, editor. PLoS ONE. 2014; 9: e94199. doi: 10.1371/journal.pone.0094199.s022 PMID: 24718519

67. Wu F, Mueller LA, Crouzillat D, Pétiard V, Tanksley SD. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics. 2006; 174: 1407–1420. doi: 10.1534/genetics.106.062455 PMID: 16951058

68. Li M, Wunder J, Bissoli G, Scarponi E, Gazzani S, Barbaro E, et al. Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics. 2008; 24: 727–745. doi: 10.1111/j.1096-0031.2008.00207.x

69. Tepe EJ, Farruggia FT, Bohs L. A 10-gene phylogeny of Solanum section Herpystichum (Solanaceae) and a comparison of phylogenetic methods. American Journal of Botany. 2011; 98: 1356–1365. doi: 10.3732/ajb.1000516 PMID: 21795733

70. Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, Fischer-York T, et al. The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Research. 2011; 39: D1149–55. doi: 10.1093/nar/gkq866 PMID: 20935049

71. Yuan Y, Olmstead R. Evolution and phylogenetic utility of the PHOT gene duplicates in the Verbena complex (Verbenaceae): dramatic intron size variation and footprint of ancestral recombination. American Journal of Botany. 2008; 95: 1166–1176. doi: 10.3732/ajb.0800133 PMID: 21632434

72. Rozen S, Skaletsky H. Primer3 on the WWW for General Users and for Biologist Programmers. In: Misener S, Krawetz SA, editors. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ; 2000. pp. 365–386.

73. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007; 23: 1289–1291. doi: 10.1093/bioinformatics/btm091 PMID: 17379693

74. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Research. 2012; 40: e115–e115. doi: 10.1093/nar/gks596 PMID: 22730293

75. Uribe-Convers S. Phylogenomic Insights into the Radiation of an Andean Group of Plants. Tank DC, editor. Ph.D. Dissertation, University of Idaho. Available: http://digital.lib.uidaho.edu/cdm/ref/collection/etd/id/585. 2014.

76. Deckmyn M, Deckmyn A. maps: Draw Geographical Maps. CRAN R-Project. Available: http://CRAN.R-project.org/package=maps

77. Bivand R, Lewin-Koh N. maptools: Tools for Reading and Handling Spatial Objects. CRAN R-Project. 2015;Available: http://CRAN.R-project.org/package=maptools.

78. Hijmans RJ. raster: Geographic Data Analysis and Modeling. Available: http://CRAN.R-project.org/package=raster.

79. Carrothers JM, York MA, Brooker SL, Lackey KA, Williams JE, Shafii B, et al. Fecal Microbial Community Structure Is Stable over Time and Related to Variation in Macronutrient and Micronutrient Intakes

in Lactating Women. The Journal of Nutrition. 2015; 145: 2379–2388. doi: 10.3945/jn.115.211110 PMID: 26311809

80. Dai J, Gliniewicz K, Settles ML, Coats ER, McDonald AG. Bioresource Technology. Bioresource Technology. Elsevier Ltd; 2015; 175: 23–33. doi: 10.1016/j.biortech.2014.10.049

81. Liang S, Gliniewicz K, Mendes-Soares H, Settles ML, Forney LJ, Coats ER, et al. Bioresource Technology. Bioresource Technology. Elsevier Ltd; 2015; 179: 268–274. doi: 10.1016/j.biortech.2014.12.032 PMID: 25545096

82. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32: 1792–1797. doi: 10.1093/nar/gkh340 PMID: 15034147

83. Smith S, Dunn C. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics. 2008; 24: 715. doi: 10.1093/bioinformatics/btm619 PMID: 18227120

84. Lanfear R, Calcott B, Ho SYW, Guindon S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution. 2012; 29: 1695–1701. doi: 10.1093/molbev/mss020 PMID: 22319168

85. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evolutionary Biology. 2014; 14: 82. doi: 10.1186/1471-2148-14-82 PMID: 24742000

86. Zwickl DJ. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. PhD Dissertation, The University of Texas at Austin. The University of Texas at Austin; 2006.

87. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. Bioinformatics. 2010; 26: 1569–1571. doi: 10.1093/bioinformatics/btq228 PMID: 20421198

88. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology. 2012; 61: 539–542. doi: 10.1093/sysbio/sys029 PMID: 22357727

89. Rambaut A, Drummond AJ. Tracer. University of Edinburgh, Edinburgh, UK. Available http://treebioedacuk/software/tracer/. 2004.

90. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. doi: 10.1093/bioinformatics/btu033 PMID: 24451623

91. Linnen C, Farrell B. Comparison of Methods for Species-Tree Inference in the Sawfly Genus Neodiprion (Hymenoptera: Diprionidae). Systematic Biology. 2008; 57: 876–890. doi: 10.1080/10635150802580949 PMID: 19085330

92. Chifman J, Kubatko L. Quartet Inference from SNP Data Under the Coalescent Model. Bioinformatics. 2014;in press. doi: 10.1093/bioinformatics/btu530

93. Molau U. The genus Bartsia (Scrophulariaceae—Rhinanthoideae). Opera Botanica. 1990; 102: 1–100.

94. Bao S, Jiang R, Kwan W, Wang B, Ma X, Song Y-Q. Evaluation of next-generation sequencing software in mapping and assembly. Journal of Human Genetics. 2011; 56: 406–414. doi: 10.1038/jhg.2011.43 PMID: 21525877

95. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics. 2011; 12: 703–714. doi: 10.1038/nrg3054 PMID: 21921926

96. Lokki J, Saura A. Polyploidy in insect evolution. Basic Life Sciences. 1980; 13: 277–312.

97. Leggatt RA, Iwama GK. Occurrence of polyploidy in the fishes. Reviews in Fish Biology and Fisheries. 2003; 13: 237–246.

98. Cannatella DC, De Sá RO. Xenopus laevis as a model organism. Systematic Biology. Oxford University Press; 1993; 42: 476–507.

99. Bogart JP. Evolutionary implications of polyploidy in amphibians and reptiles. Basic Life Sciences. 1980; 13: 341–378. doi: 10.1007/978-1-4613-3069-1_18

100. Judd WS, Soltis DE, Soltis PS, Ionta G. Tolmiea diplomenziesii: A new species from the Pacific Northwest and the diploid sister taxon of the autotetraploid T. menziesii (Saxifragaceae). Brittonia. Springer; 2007; 59: 217–225.

101. Baldwin BG. Phylogenetic Utility of the Internal Transcribed Spacers of Nuclear Ribosomal DNA in Plants: An Example from the Compositae. Molecular Phylogenetics and Evolution. 1992; 1: 3–16. PMID: 1342921

102. Baldwin BG, Markos S. Phylogenetic Utility of the External Transcribed Spacer (ETS) of 18S–26S rDNA: Congruence of ETS and ITS Trees of Calycadenia (Compositae). Molecular Phylogenetics and Evolution. Elsevier; 1998; 10: 449–463. PMID: 10051397

103. Marx HE, O'Leary N, Yuan Y-W, Lu-Irving P, Tank DC, Múlgura ME, et al. A molecular phylogeny and classification of Verbenaceae. American Journal of Botany. 2010; 97: 1647–1663. doi: 10.3732/ajb.1000144 PMID: 21616800

104. Ilut DC, Doyle JJ. Selecting Nuclear Sequences for Fine Detail Molecular Phylogenetic Studies in Plants: A Computational Approach and Sequence Repository. Systematic Botany. 2012; 37: 7–14. doi: 10.1600/036364412X616576

105. Genome 10K Community of Scientists. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. Journal of Heredity. 2009; 100: 659–674. doi: 10.1093/jhered/esp086 PMID: 19892720

106. Chamala S, García N, Godden GT, Krishnakumar V, Jordon-Thaden IE, De Smet R, et al. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. Applications in Plant Sciences. 2015; 3. doi: 10.3732/apps.1400115

107. Thiers B. Index Herbariorum: A global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium Available: http://sweetgum.nybg.org/ih/.