New Phytologist

Research

Next-generation polyploid phylogenetics: rapid resolution of hybrid polyploid complexes using PacBio single-molecule sequencing

Carl J. Rothfels[1], Kathleen M. Pryer[2] and Fay-Wei Li[1,2]

[1]University Herbarium and Department of Integrative Biology, University of California, Berkeley, CA 94720, USA; [2]Department of Biology, Duke University, Durham, NC 27705, USA

Author for correspondence:
Carl J. Rothfels
Tel: +1 919 907 8744
Email: crothfels@berkeley.edu

Received: *1 May 2016*
Accepted: *14 June 2016*

New Phytologist (2016)
doi: 10.1111/nph.14111

Key words: allopolyploidy, Cystopteridaceae, hybridization, parallel-tagged amplicon sequencing, Pipeline for Untangling Reticulate Complexes (PURC), reticulate evolution, species complex, species network.

Summary

- Difficulties in generating nuclear data for polyploids have impeded phylogenetic study of these groups. We describe a high-throughput protocol and an associated bioinformatics pipeline (Pipeline for Untangling Reticulate Complexes (PURC)) that is able to generate these data quickly and conveniently, and demonstrate its efficacy on accessions from the fern family Cystopteridaceae. We conclude with a demonstration of the downstream utility of these data by inferring a multi-labeled species tree for a subset of our accessions.
- We amplified four *c*. 1-kb-long nuclear loci and sequenced them in a parallel-tagged amplicon sequencing approach using the PacBio platform. PURC infers the final sequences from the raw reads via an iterative approach that corrects PCR and sequencing errors and removes PCR-mediated recombinant sequences (chimeras).
- We generated data for all gene copies (homeologs, paralogs, and segregating alleles) present in each of three sets of 50 mostly polyploid accessions, for four loci, in three PacBio runs (one run per set). From the raw sequencing reads, PURC was able to accurately infer the underlying sequences.
- This approach makes it easy and economical to study the phylogenetics of polyploids, and, in conjunction with recent analytical advances, facilitates investigation of broad patterns of polyploid evolution.

Introduction

Nuclear sequence data are critical for plant phylogenetic inference (Sang, 2002; Small *et al.*, 2004; Zimmer & Wen, 2012, 2015). These data are typically faster evolving than plastid or mitochondrial sequences (Wolfe *et al.*, 1987; Rothfels & Schuettpelz, 2014; Minaya *et al.*, 2015) and, because the nuclear genome is so much larger than the organellar genomes, more data from it are available (Zimmer & Wen, 2015); these sequence qualities provide increased resolving power for classic concatenated-data approaches, especially at shallower phylogenetic scales (Duarte *et al.*, 2010; Zhang *et al.*, 2012; Minaya *et al.*, 2015). Moreover, nuclear data provide the opportunity to sample multiple linkage groups, mitigating the risk of misleading results caused by biases or idiosyncrasies of individual linkage groups (e.g. the plastid or mitochondrion; Moore, 1995) and allowing for inferences beyond those possible from mitochondrial or plastid data alone. These novel analyses include the inference of species histories from coalescence patterns across multiple gene genealogies (Maddison & Knowles, 2006; Degnan & Rosenberg, 2009) and the estimation of patterns of introgression across the genome (Yu *et al.*, 2014). Finally, because of their biparental inheritance, nuclear sequence data can be used to infer both homoploid (Zhang *et al.*, 2013) and polyploid hybridization (Triplett *et al.*, 2014).

Nuclear sequence data, however, are difficult and expensive to generate for taxonomic groups with limited available genomic resources, especially for studies of polyploidy or gene family evolution, where multiple gene copies (homeologs or paralogs) are present in individual accessions. In these cases, direct sequencing of PCR products typically results in uninterpretable data as a consequence of polymorphic sites and sequence length variation among the amplified copies. The most frequently adopted solution to this problem is to clone the PCR product into plasmid vectors, use these to transform *Escherichia coli*, and then Sanger-sequence a sufficient number of the resulting bacterial colonies (Sang, 2002; Dufresne *et al.*, 2013). This approach is expensive and labor-intensive and, because of practical limits on the number of colonies that can be sequenced, it is often difficult to identify and correct PCR errors, sequencing errors, and PCR-mediated chimeras (Schuettpelz *et al.*, 2008; Griffin *et al.*, 2011; Dufresne *et al.*, 2013).

There are few existing alternatives to cloning. If sequence data are available for each of the homeologs (or paralogs) then it may be possible to design copy-specific primers (e.g. Marcussen *et al.*, 2012; Meseguer *et al.*, 2014). This approach is labor-intensive,

and requires that all copies be pre-identified and sufficiently distinct from each other, so that copy-specific primers can be designed. Alternatively, if the copies have no length differences, then polymorphisms can be visually identified (e.g. Shepherd *et al.*, 2008). This approach requires that there not be indels among the copies, so it is largely limited to more slowly evolving sequences, and some method is needed to phase the substitutions into haplotypes. Finally, single-molecule PCR (with template so dilute that most successful reactions will start from only a single target sequence) can be used (Marcussen *et al.*, 2012), but this approach is labor-intensive, technically challenging, and prone to high levels of PCR error. These difficulties have long impeded the study of polyploid groups (Ramsey & Schemske, 2002; Schuettpelz *et al.*, 2008; Dufresne *et al.*, 2013; Ramsey & Ramsey, 2014; Soltis *et al.*, 2014), which is especially problematic given the high frequency of polyploidy in plants, and the ecological and evolutionary significance of the process of polyploidization (Otto & Whitton, 2000; Otto, 2007; Martin & Husband, 2009; Wood *et al.*, 2009; Husband *et al.*, 2013; Mable, 2013; De Storme & Mason, 2014; Estep *et al.*, 2014).

A convenient and cost-effective method for generating long-read DNA sequence data for multiple independent loci, from multiple polyploid accessions (and their diploid relatives), is sorely needed to overcome this data deficit. Recent developments in next-generation sequencing technologies have brought us much closer to this goal (Twyford & Ennos, 2011; McCormack *et al.*, 2013; Zimmer & Wen, 2015). In particular, the work of Griffin *et al.* (2011) demonstrated the utility of next-generation sequencing for resolving polyploid complexes. Subsequent workers have elaborated upon this general approach or refined it for other purposes. However, these studies tended to focus on generating single-copy data (so were limited to inferring a maximum of one or two alleles at a given locus; Zieliński *et al.*, 2013; Barrow *et al.*, 2014; Wielstra *et al.*, 2014; Feng *et al.*, 2016), or, as in Griffin *et al.* (2011), were applicable only to short sequences (< 500 bp per locus; Gholami *et al.*, 2012; Wielstra *et al.*, 2014; Uribe-Convers *et al.*, 2016). Approaches that were able to generate longer sequences did so by assembling shorter reads (from 454 Life Sciences, Branford, CT, USA or Illumina MiSeq runs, Illumina Inc., San Diego, CA, USA) and thus required additional bioinformatic steps to phase the sequences (Bybee *et al.*, 2011; Gholami *et al.*, 2012; O'Neill *et al.*, 2013; Zieliński *et al.*, 2013; Barrow *et al.*, 2014; Feng *et al.*, 2016), or required that reference sequences of all potential homeologs be available (Brassac & Blattner, 2015). The former approaches are not able to correctly phase single nucleotide polymorphisms (SNPs) separated by too large a region of invariable sequence, and the latter approach requires cloning and Sanger sequencing and requires all homeologs to be present (and sequenced) in the subset of accessions selected as the references. Finally, these methods typically require relatively involved data preparation steps (DNA shearing, barcode ligations, multiple purifications and quantifications, etc; Bybee *et al.*, 2011; Gholami *et al.*, 2012; O'Neill *et al.*, 2013; Brassac & Blattner, 2015; Feng *et al.*, 2016) and some entail two rounds of PCR amplifications, increasing the prevalence of PCR errors in the amplicon pool (Bybee *et al.*, 2011; Gholami *et al.*, 2012).

Aside from amplicon sequencing approaches, sequence data for polyploids (or for gene families, etc.) could be generated through reduced genome complexity techniques. Eaton (2014), for example, developed a pipeline to extract homologous sequences from RADseq data and methods are available to do the same with transcriptome data (e.g. Yang & Smith, 2014). However, reduced complexity approaches to data generation (RADseq, target enrichment, RNAseq, etc.) have significant requirements before they generate useful phylogenetic data, or are ill-suited to complex polyploid genomes.

Here, we describe an amplicon-sequencing method for generating long (*c.* 1 kb) sequences from multiple loci (nuclear, plastid, or mitochondrial), from multiple accessions, focusing on cases where individual accessions are expected to contain multiple distinct copies of each locus. We first describe a molecular laboratory protocol, utilizing the PacBio sequencing platform (Pacific Biosciences, Menlo Park, CA, USA), that generates data relatively quickly and cheaply and does not require any sequence phasing or assembly. This protocol yields all the copies (alleles, homeologs, or paralogs) amplified by a given primer pair, for each accession. We then describe the Pipeline for Untangling Reticulate Complexes (PURC), the associated bioinformatics package that takes as input the raw PacBio reads and infers the true biological sequences, producing alignments for each locus, with each sequence labeled according to its source accession and its depth of sequencing coverage. The efficacy of this combined wet lab/bioinformatics approach is demonstrated by generating data for four nuclear loci for a sample of mostly polyploid accessions from the fern family Cystopteridaceae. Finally, we use a subset of our accessions to illustrate how this approach can allow for novel evolutionary insights, by inferring a multilocus 'species network' that is analogous to a species tree, but includes the inference of reticulations for the allopolyploid accessions.

Materials and Methods

Molecular laboratory protocol

Our molecular laboratory protocol follows three steps: PCR amplification of the loci of interest, sample pooling, and sequencing on the PacBio RS II platform. The PCR amplifications are performed with barcoded primers, removing any need for ligation steps. For our pilot study we used the 48 16-base barcodes (without padding) provided by Pacific Biosciences (available in the Dryad Digital Repository: doi: 10.5061/dryad.dj82k). To reduce primer costs, we barcoded only the forward primers, and reused barcodes within runs for taxonomic groups that could be distinguished phylogenetically (see the description of the bioinformatics pipeline in the next section). The number of accessions that can be included in a run varies with the length of the target loci, the expected ploidy of the accessions, and the desired coverage; in our pilot study we limited each run to *c.* 700 potential target sequences (where each diploid accession, given that it has two copies of the genome, contributes two target sequences per locus, etc.). This cut-off corresponds to *c.* 50 mostly polyploid accessions for four loci per run (based on the P4-C2 sequencing

chemistry; with the recently released reagents – the P6-C4 chemistry – at least 1000 potential targets could be included). The amplified regions to be pooled in a single sequencing run should all be approximately the same length (Pacific Biosciences recommends that pooled amplicons be within 15% of the mean length of that pool; Pacific Biosciences, 2016), otherwise the short regions may be preferentially sequenced, resulting in poorer coverage of the longer regions. Here, we selected regions that ranged from *c.* 900 to *c.* 1100 bp. No special steps need to be taken for the PCR amplifications, and the resulting amplicons were not cleaned or purified before library preparation (Pacific Biosciences does recommend purifying the pooled amplicons before library preparation – doing so many improve the method's performance; Pacific Biosciences, 2016).

Each PCR product was run on an agarose gel to confirm amplification success, and to allow the samples to be pooled in approximately equal concentrations. We adopted a very coarse quantification scheme by scoring each band, by eye, on a five-point scale ranging from 'very weak' to 'very strong'. Based on our preliminary quantification of DNA concentrations using a nanodrop, we estimated these five band strength categories to roughly correspond to DNA concentrations of 10, 15, 20, 30 and 50 ng μl^{-1}, respectively, and we used these concentration estimates to pool equal masses of DNA from each sample, aiming for a final pooled sample of at least 2 μg of DNA in a 150-μl volume. This method of sample standardization is very quick and inexpensive, and was sufficient to yield good coverage of our target sequences (see the 'Nuclear data from polyploids: a case study' section). However, researchers interested in sequencing more targets per run may wish to standardize their amplicons more rigorously.

The pooled sample was then sequenced on a single PacBio SMRT cell, utilizing PacBio's 'Circular Consensus Sequencing' (CCS) technology (Travers *et al.*, 2010). With average sequencing lengths of > 4 kb on the P4-C2 chemistry (> 10 kb on the latest P6-C4 chemistry; Pacific Biosciences, 2015), a given 1-kb target gets an average of four sequencing laps; the final sequence delivered is the consensus of those laps. Because PacBio sequencing errors occur randomly (Eid *et al.*, 2009), the individual laps are unlikely to share the same errors, and the consensus sequences are highly accurate. An example file for sample tracking and library construction (amplicon pooling) is available in the Dryad Digital Repository (doi: 10.5061/dryad.dj82k).

Bioinformatics pipeline

To process the raw sequencing reads, we developed an integrated, command-line based pipeline: Pipeline for Untangling Reticulate Complexes (PURC). PURC is written in PYTHON and settings are controlled via a configuration file. It requires PYTHON 2.7 or later (but is not compatible with PYTHON 3), BIOPYTHON v.1.6 or later (Cock *et al.*, 2009), BLAST+ v.2.2.30 or later (Camacho *et al.*, 2009), and comes packaged with three additional dependencies: CUTADAPT (Martin, 2011), MUSCLE (Edgar, 2004), and USEARCH (Edgar, 2010). PURC takes as input a FASTA file of raw amplicon sequence reads, de-multiplexes the amplicon pool, and then clusters the sequences into alleles (henceforth we will use 'alleles' loosely, to refer to all unique biological sequences present, including segregating alleles, homeologs, and paralogs). The full pipeline consists of five main steps: (1) de-multiplexing: primer and barcode sequence removal; annotation of reads with locus and source accession name; separation of reads by locus and accession; (2) read clustering and chimera removal (performed four times, iteratively); (3) consensus sequence computation, for each cluster; (4) final clustering and chimera removal; (5) inferring alignments for each locus.

Preliminary investigations indicated that a small fraction of the PacBio CCS reads were interlocus 'concatemers' (chimeras of two separate loci), presumably introduced during the PacBio library preparation, through SMRTbell-mediated sequence fusion (Fichot & Norman, 2013). Therefore, step 1 begins with the option to split these sequences into their component single-locus sequences and recycle them back into the data for analysis. PURC then uses BLAST (Camacho *et al.*, 2009) to identify the barcode sequences, with the option of using a more sensitive (and slower) Smith–Waterman local alignment approach (Smith & Waterman, 1981) on those sequences where BLAST fails to locate a barcode (Fig. 1). PURC additionally includes the option to restrict the search for barcodes to the ends of the sequence reads, allowing the user to avoid spurious matches to nonbarcode regions in the middle of the sequences. The reads are annotated with the barcode number, and PURC removes the primers and any residual barcode sequences using CUTADAPT (Martin, 2011).

PURC then BLASTs the reads against a list of user-supplied reference sequences (Fig. 1), extracting the locus and group information from the best-matching reference sequence. The 'group' indicates the user-determined division of the accessions into phylogenetically distinct groups (e.g. different genera); this allows the de-multiplexing of sequences from divergent taxa that share the same barcode. PURC then completes the annotation by matching each unique combination of barcode, locus, and group attributes with the source accession, using a user-provided mapping file, and separates the annotated reads according to their locus and accession.

To infer the alleles from these raw amplicon sequences – step 2 – PURC performs four rounds of clustering using USEARCH's cluster_fast algorithm (Edgar, 2010). For each round, the consensus sequences of the previous round's clusters are used as input, and between each clustering step PURC interjects USEARCH's UCHIME function (Edgar *et al.*, 2011) to remove potential PCR-mediated recombinants (chimeras; Fig. 1). Users can adjust the minimum per cent sequence identity necessary for sequences to be clustered together, and the minimum cluster size necessary for the cluster to be retained, for each of the four clustering rounds; the UCHIME settings can also be changed. In step 3, the constituent reads from each remaining cluster are aligned with MUSCLE (Edgar, 2004) and their consensus sequences calculated. This step corrects any errors in the initial consensus sequences that were propagated through the iterative clustering steps, and allows us to use MUSCLE instead of USEARCH for the alignments underlying these final consensus sequences. In step 4, these consensus sequences are in turn clustered and run through UCHIME one more time (Fig. 1), producing a final set of allele sequences for each accession for each locus. In the final step (step 5), PURC gathers the

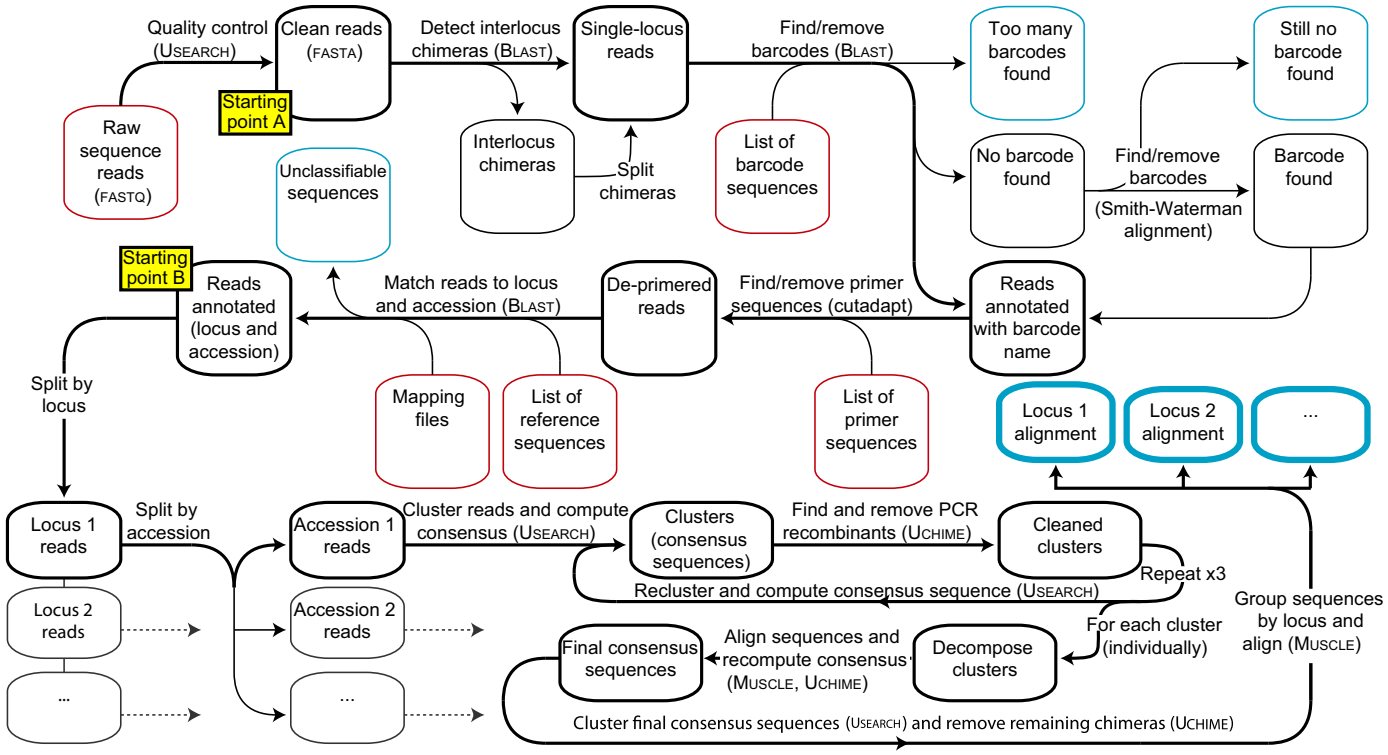

Fig. 1 The bioinformatic Pipeline for Untangling Reticulate Complexes (PURC). Input files are outlined in red and output files in blue. The two versions of PURC (PURC.PY and PURC_RECLUSTERER.PY) operate on the input files at starting points A and B, respectively (highlighted in yellow).

allele sequences for each locus from all accessions, and aligns them using MUSCLE (Edgar, 2004), producing a set of single-locus alignments ready for downstream phylogenetic analyses (Fig. 1).

Two versions of PURC are available: one that performs the entire workflow (PURC.PY), and another (PURC_CLUSTERER.PY) that performs only the clustering and chimera-killing steps – the later script takes as input the FASTA file produced by the full PURC, with the sequences already annotated with their locus and source accession (starting point B in Fig. 1). The settings for PURC_CLUSTERER.PY are set at the command line, rather than through a configuration file, allowing for easy batch scripting. Both PURC scripts are available from Bitbucket (https://bitbucket.org/crothfels/purc), along with installation instructions, tutorials, examples, and additional documentation.

Nuclear data from polyploids: a case study

We tested our molecular laboratory protocol and bioinformatics pipeline on accessions from the fragile fern family (Cystopteridaceae). Using the transcriptome-based 'all-in' alignments from Rothfels *et al.* (2013a), we designed intron-spanning primers to amplify regions *c.* 1-kb long from four single-copy genes (*ApPEFP_C*, *gapCpSh*, *IBR3* and *pgiC*; henceforth *APP*, *GAP*, *IBR* and *PGI*, respectively; Table 1), and added the Pacific Biosciences barcodes to the forward primers. *APP* and *GAP* were amplified in 20-μl reactions consisting of 2 μl of 10× Denville buffer (Denville Scientific, Holliston, MA, USA), 2 μl of dNTPs (2 mM each), 0.2 μl of bovine serum albumin (BSA) (10 mg ml^{-1}), 0.2 μl of Denville Choice Taq (5 U μl^{-1}), 1 μl of each primer (10 μM), and 13.6 μl of water. *IBR* and *PGI* were amplified in 19-μl reactions each with 4 μl of 5× Phusion HF buffer (New England Biolabs, Ipswich, MA, USA), 2 μl of dNTPs (2 mM each), 0.2 μl of Phusion HF polymerase, 2 μl of each primer (10 μM), and 8.8 μl of water. The thermocycling conditions for all loci consisted of an initial period at the melting temperature (3 min for *APP* and *GAP*; 1 min for *IBR* and *PGI*), followed by 35 cycles of 30 s at the melting temperature, 30 s at the annealing temperature, and 1.5 min at the elongation temperature; reactions ended with 10 min at the elongation temperature (the respective temperatures for each locus are listed in Table 1).

After an initial test run ('R1'; data not shown), we amplified these regions for three sets of Cystopteridaceae accessions (the full list with voucher information is available in the Dryad Digital Repository: doi: 10.5061/dryad.dj82k), which we submitted as individual PacBio sequencing runs ('R2', 'R3', and 'R4', respectively). DNA was extracted from silica-dried leaf tissue (or, more rarely, herbarium material) in the Fern Lab Silica Archive (http://fernlab.biology.duke.edu/) using a 96-well modification (Beck *et al.*, 2011a,b) of a standard CTAB protocol (Doyle & Dickson, 1987). Within each sequencing run, we used individual barcodes up to three times per locus: once for a *Gymnocarpium* accession (which we designated as group A), once for a member of the *Cystopteris fragilis* (L.) Bernh. complex (group C), and once for an accession of *Acystopteris* or a 'non-*fragilis* complex' *Cystopteris* species (group B). Two microliters of each PCR product were run on an agarose gel to confirm amplification success and to

estimate DNA concentration (see the 'Molecular laboratory protocol' section, above). Our three sequencing runs each included between 51 and 53 accessions (Dryad Digital Repository: doi: 10.5061/dryad.dj82k), and, because of variation in amplification success and the ploidy of the included samples, between 666 and 710 target sequences (Table 2).

We pooled each amplicon (the PCR products were not cleaned or purified) within a run proportional to its ploidy and inversely proportional to its estimated DNA concentration. Example spreadsheets for sample tracking and pooling are available in the Dryad Digital Repository (doi: 10.5061/dryad.dj82k). Sequencing libraries were prepared using the medium insert PacBio 1-kb protocol (Pacific Biosciences, 2016). Each amplicon pool was sequenced on a single SMRT cell on a PacBio RSII sequencer using the P4-C2 chemistry and a 3-h movie length. Sequencing libraries were diffusion loaded on the SMRT cells and sequence data were analyzed using CCS SMRT ANALYSIS software v.2.2.0 and four filtering passes. Library preparation and sequencing were performed at the Sequencing and Genomic Technologies Core Facility of the Duke University Center for Genomic and Computational Biology.

The raw PacBio CCS FASTQ files for each run were cleaned using USEARCH's fastq_filter command (Edgar, 2010), set to remove all reads that were < 600 bases long or that had greater than five expected errors (the number of expected errors is calculated as the sum of the error probabilities of each base in the sequence). The cleaned sequences were output in FASTA format and annotated with PURC.PY, incorporating the options to split and recycle back into the analysis concatemers, Smith–Waterman primer searching, and searches for primers across the entire sequence rather than just at the ends. The reference sequences used during the annotation process were derived from earlier studies (Rothfels *et al.*, 2013a, 2014; C. J. Rothfels, unpublished) and by repeated iterations of the annotation step using inferred sequences from one round as reference sequences for the next (verifying each time that the resulting trees matched earlier broad phylogenetic hypotheses; Rothfels *et al.*, 2013b). The annotated sequences from each run were analyzed using six different PURC clustering regimes by repeatedly calling PURC_CLUSTERER.PY via a shell script (an example shell script is available in the Dryad Digital Repository: doi: 10.5061/dryad.dj82k). The first three regimes (a, b and c) were run with the default UCHIME settings, and differed in their sequence similarity cut-offs in the four clustering rounds (Table 3). The second three regimes (aStr, bStr, and cStr) used the same clustering cut-offs as the first three, but with the UCHIME settings changed to be more stringent (Table 3). In all cases, all sequences (clusters) were retained after each of the first three clustering rounds, but only sequences representing clusters of at least four original reads were retained after the fourth clustering (singletons, etc., were discarded). The final clusters were decomposed back into their constituent reads, which were aligned in MUSCLE, and their consensus sequence was recalculated (Fig. 1). These final consensus sequences (for each accession) were merged if they were at least 99.9% identical, and UCHIME was run one final time, before the sequences for each loci were extracted and aligned (again with MUSCLE). PURC configuration, output, and log files are available in the Dryad Digital Repository (doi: 10.5061/dryad.dj82k), as are the R scripts (R Core Team, 2013) used to summarize the results.

Pipeline evaluation

We assessed the ability of our workflow to infer the true biological sequences present in a sample in three ways. First, for each

Table 1 Primers used in the case study

Locus (abbrev.)	Primers	Primer sequence (5′ to 3′)	PCR program	Length
ApPEFP_C (*APP*)	4218CF4*	GGACCTGGSCTYGCTGARGAGTG	95.65.71	*c.* 930 bp
	4218CR12*	GCAACRTGAGCAGCYGGTTCRCGRGG		
gapCpSh (*GAP*)	gapCpShF1Cys	CYACMAACTGCCTTGCRCCTCTTGCC	95.59.71	*c.* 900 bp
	ESGAPCP11R1ShCys**	GTATCCCCACTCRTTATCATACC		
IBR3 (*IBR*)	4321F2*	TCTGCMCATGCMATTGAAAGAGAG	98.65.72	*c.* 840 bp
	4321R2*	CCCARKGTYGAAAGYTCCCAATC		
pgiC (*PGI*)	CRpgicF2U	GAGYGTTTGGAATGTYTCWTTCCTYGG	98.68.72	*c.* 900 bp
	CRpgicR2U	TCGTCGTGGTTGCTCACAACTTCCC		

*From Rothfels *et al.* (2013a). **Modified from ESGAPCP11R1 (Schuettpelz *et al.*, 2008). The three values listed for each PCR program are the melting temperature, annealing temperature, and elongation temperature, respectively.

Table 2 Data set statistics, before clustering

Run	PacBio CCS reads returned	Reads > 600 bp	Reads > 600 bp and ee < 5	Number of interlocus concatemers	Number (%) of barcode fails	Number of unclassifiable reads	Estimated maximum target number
R2	38 610	30 732	18 189	90	9.48%	25	666
R3	51 001	44 721	27 994	149	8.16%	220	710
R4	44 065	38 982	28 550	180	8.21%	89	680

The estimated maximum target number is a count of the potential number of distinct sequences present in the amplicon pool (e.g. a tetraploid accession would contribute four targets per locus, a diploid would contribute two, etc.). CCS, circular consensus sequencing; ee, expected errors.

Table 3 The six Pipeline for Untangling Reticulate Complexes (PURC) analysis regimes used in the case study

Regime	Cluster thresholds				UCHIME settings			
a	0.995	0.995	0.995	0.995	2.0	0.28	8.0	1.4
b	0.995	0.995	0.990	0.990	2.0	0.28	8.0	1.4
c	0.997	0.995	0.995	0.995	2.0	0.28	8.0	1.4
aStr	0.995	0.995	0.995	0.995	1.1	0.20	3.0	0.5
bStr	0.995	0.995	0.990	0.990	1.1	0.20	3.0	0.5
cStr	0.997	0.995	0.995	0.995	1.1	0.20	3.0	0.5

The cluster thresholds show the per cent similarity necessary, at each of the four main clustering rounds, for two sequences to be clustered together. UCHIME settings are listed in the following order: abskew, minh, xn, dn. For explanations of each setting, see the USEARCH manual at http://drive5.com/usearch/manual/UCHIME_score.html.

locus in our R2 data we combined the resulting sequences from the six analysis regimes into a single alignment (using MUSCLE and ALIVIEW; Edgar, 2004; Larsson, 2014) and inferred a maximum parsimony tree in PAUP v.4.0a147 (Swofford, 2002). We then, by manual examination of these four phylogenies and incorporating prior information about the ploidy level of each accession, estimated the number of true biological sequences present for each accession, for each locus. This process was necessarily subjective, but cases of ambiguity were rare: for the great majority of accessions, the same number of alleles was inferred for each locus, and that number was consistent with the known ploidy of that accession (see the Results section). This process resulted in a list of 'true' allele counts per accession for each locus, which we could then use to compare the performance of individual analysis regimes.

As a second evaluation, we included two accessions (*Cystocarpium moupinensis* Franch. #8735 and *Gymnocarpium oyamense* (Baker) Ching #8739) twice in the case study, in different sequencing runs. These accessions allow us to examine the repeatability of inference across different repetitions of the molecular laboratory and sequencing components of the workflow, and, in addition, we can compare the inferences across different PURC analysis regimes within a single sequencing run. Our third evaluation was to compare the results of our PURC allele inferences with the sequences obtained from the classic cloning and Sanger-sequencing approach. To do so, we created a combined alignment (MUSCLE and ALIVIEW; Edgar, 2004; Larsson, 2014) of *GAP* sequences from our PURC-based species network inference (see the next section) and our earlier Sanger-based studies (Rothfels *et al.*, 2014, 2015) for those accessions that were shared across the two data sets. From this combined alignment we then inferred a maximum parsimony tree in PAUP v.4.0a147 (Swofford, 2002), to visualize the similarities and differences between the PURC and Sanger-based inferences.

Untangling polyploid complexes: Cystopteridaceae species network

To provide an example of the potential applications of this approach, we used a subset of data from our first PacBio run (R2) to infer a multilabeled species tree using ALLOPPNET (Jones *et al.*, 2013) as implemented in BEAST 1.8.2 (Drummond & Rambaut, 2007). This tree is a species tree in the sense that it is inferred from multiple unlinked gene trees using the multispecies coalescent; however, the tips are 'species-genomes' rather than species *per se* (an allotetraploid species would occur twice in the species tree, once for each of its homeologous genomes). We analyzed the raw PacBio reads with the six analysis regimes described earlier (Table 3), and, for each locus, used MUSCLE's profile-profile alignment option (Edgar, 2004), followed by manual alignment adjustment, to infer a single alignment containing the output from all six regimes. From each of these alignments, we inferred a maximum parsimony tree in PAUP 4.0a147 (Swofford, 2002), and selected, by visual inspection of this tree, the final set of allele sequences for each accession (these all-regimes trees are available in the Dryad Digital Repository: doi: 10.5061/dryad.dj82k). The application of the current implementation of ALLOPPNET is limited to diploid and tetraploid accessions; we therefore removed the higher ploidy accessions from the data set. Because our initial runs failed to converge after 400 million generations, we reduced the sample to nine diploid species and 19 tetraploids, selected to span the phylogenetic diversity of Cystopteridaceae (Supporting Information Table S1).

ALLOPPNET requires that all input sequences be assigned to a species, an individual, and a genome (Jones, 2012). For the polyploids, we conservatively considered accessions to be conspecific only if they had a high degree of sequence similarity across all four loci. One accession – × *Cystocarpium roskamianum* Fraser-Jenk. – is tetraploid; however, it is itself a hybrid of two allotetraploids and thus has four homeologous (nonrecombining) subgenomes (Rothfels *et al.*, 2015). To accommodate this accession within the ALLOPPNET model, we divided its homeologs into two 'species', corresponding to each of the two parental tetraploids. Our final sample of nine diploid species and 19 tetraploids comprised 13 and 21 individuals, respectively (Table S1).

We excluded any areas of ambiguous alignment in each of the locus alignments and produced the final alignments (with the sequences renamed and blank sequences added for alleles that are missing from particular loci) using the ABIOSCRIPTS 0.9.4 –seqconcat function (Larsson, 2010). We used the R (R Core Team, 2013) scripts provided with ALLOPPNET to generate a BEAST XML file from these final data, and manually edited that XML to incorporate a starting tree for each gene (generated from a short BEAST run on the full data), and to enforce the known monophyly of three clades: *Gymnocarpium*, *Acystopteris*, and *Acystopteris* + *Cystopteris*. All other features – notably an HKY substitution model (Hasegawa *et al.*, 1985) and strict molecular clock applied to each gene—were left at their ALLOPPNET-generated defaults. These data were run four times independently in BEAST v.1.8.2 (Drummond & Rambaut, 2007) on the CIPRES gateway (Miller *et al.*, 2010). Each chain was run for 400 million generations, with the chains sampled every 10 000 generations, and convergence was assessed by examining the parameter traces and the effective sample sizes using TRACER v.1.6 (Rambaut *et al.*, 2007). The burnin was excluded and

the remaining posterior summarized on the maximum clade credibility tree using TreeAnnotator v.1.8.0 (packaged with Beast; Drummond & Rambaut, 2007).

Results

Our three PacBio sequencing runs returned between 38 000 and 51 000 CCS reads, of which between 18 000 and 28 000 were > 600 bp long and had fewer than five expected errors (Table 2; Fig. 2). These raw reads, despite the CCS technology and the post-run quality control, are heterogeneous (far more distinct sequences are recovered than expected given the ploidy of the sequenced accessions; Fig. 3), presumably as a result of chimeras and PCR and/or sequencing errors. However, Purc's iterative clustering, chimera-killing, and consensus-calculation approach succeeds in removing these errors (Fig. 3), as demonstrated by the high repeatability across runs and across analysis regimes within runs (Fig. 4; see also the all-regimes trees in the Dryad Digital Repository: doi: 10.5061/dryad.dj82k). Specifically, for the two accessions (*C. moupinensis* #8735 and *G. oyamense* #8739) that were included in both the R2 and R3 runs, the same alleles were inferred by all six analysis regimes, across both data sets, and the exceptions to this rule were low-coverage sequences easily identifiable as chimeras or were high-coverage sequences resulting from overclustering in the Purc pipeline (Fig. 4). Similarly, the alleles inferred with this workflow are consistent with those from the classic cloning and Sanger-sequencing approach: for the 19 accessions for which we have both Purc and 'classic' Sanger data, the two methods inferred the same number of alleles 15 times, and in the four other accessions Purc inferred an additional allele (Fig. S1). In all four of these cases, the additional allele makes sense phylogenetically and is consistent with the ploidy level of the accession, and relatively few cloned sequences were generated for that accession (Fig. S1); these cases almost certainly represent examples of conventional cloning approaches failing to detect all sequences present. In addition, for the 33 cases where both methods inferred the same allele (or close to it), the Purc/PacBio-generated allele is identical to the most frequently obtained cloning/Sanger-generated allele 22 times, and to the consensus of the cloning/Sanger sequences (in those cases where only unique clone sequences were obtained) five times. In four cases the two methods produced very similar, albeit not identical sequences, with the cloning-derived ones differing by single apomorphies that are very probably the result of PCR error. Finally, in two cases, the Purc allele is identical to one of the minority clone sequences (rather than to the most commonly obtained clone sequence). In these instances, the most commonly obtained clone sequence differs from the Purc and minority clone sequence by a small number of apomorphies. These 'majority' clone sequences are also probably attributable to PCR error that happened early in the PCR cycles, and thus propagated to multiple clones (Fig. S1). There is no indication, for any accession, of the cloning and Sanger-sequencing approach outperforming the Purc-based inferences.

The number of final alleles inferred for each accession varied across loci, mainly because of amplification differences (not all accessions successfully amplified for all loci) and because some loci (i.e. *IBR*) showed a higher propensity for producing difficult-to-eliminate chimeric sequences (Tables 4, 5; Fig. 5). In general, however, the number of alleles inferred was strongly indicative of the ploidy of the accession (Table 5): diploids have one or two alleles (diploids were more often heterozygous than the polyploids); tetraploids have two or sometimes three alleles; and hexaploids have three or occasionally four alleles. In those cases where more alleles were inferred than expected, these excess sequences were typically rare (low coverage; Table 5), recovered in a minority of the analysis regimes, and easily identifiable as chimeras.

Across all three data sets, and all four loci, average coverage for the final allele sequences was roughly comparable, even with our decidedly coarse amplicon standardization scheme (Table 4). There was, however, considerable variance about those averages (Table 4; Fig. 6): while the vast majority of accessions had sufficient coverage for Purc to correctly identify the homeologs present (Table 6), more rigorous standardization schemes would probably increase the evenness of coverage.

Our data for the AlloPPNet allopolyploid network inference comprised four loci from 28 species (Table S1). The individual locus alignments ranged in length from 862 to 1132 base pairs, and included between 4.2 and 15.5% missing data. The missing data in the locus alignments were attributable to indels; the combined data set had a greater percentage of missing data because of

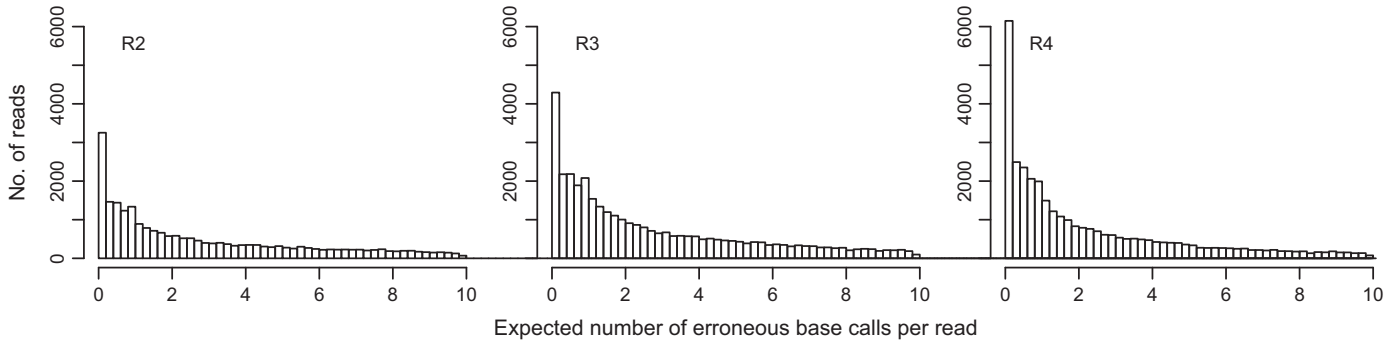

Fig. 2 Expected number of sequencing errors in the raw PacBio circular consensus sequencing (CCS) data, for each of the three sequencing runs. The expected errors in a read are calculated by summing the probabilities of error (from the fastq file) for each base in that read.

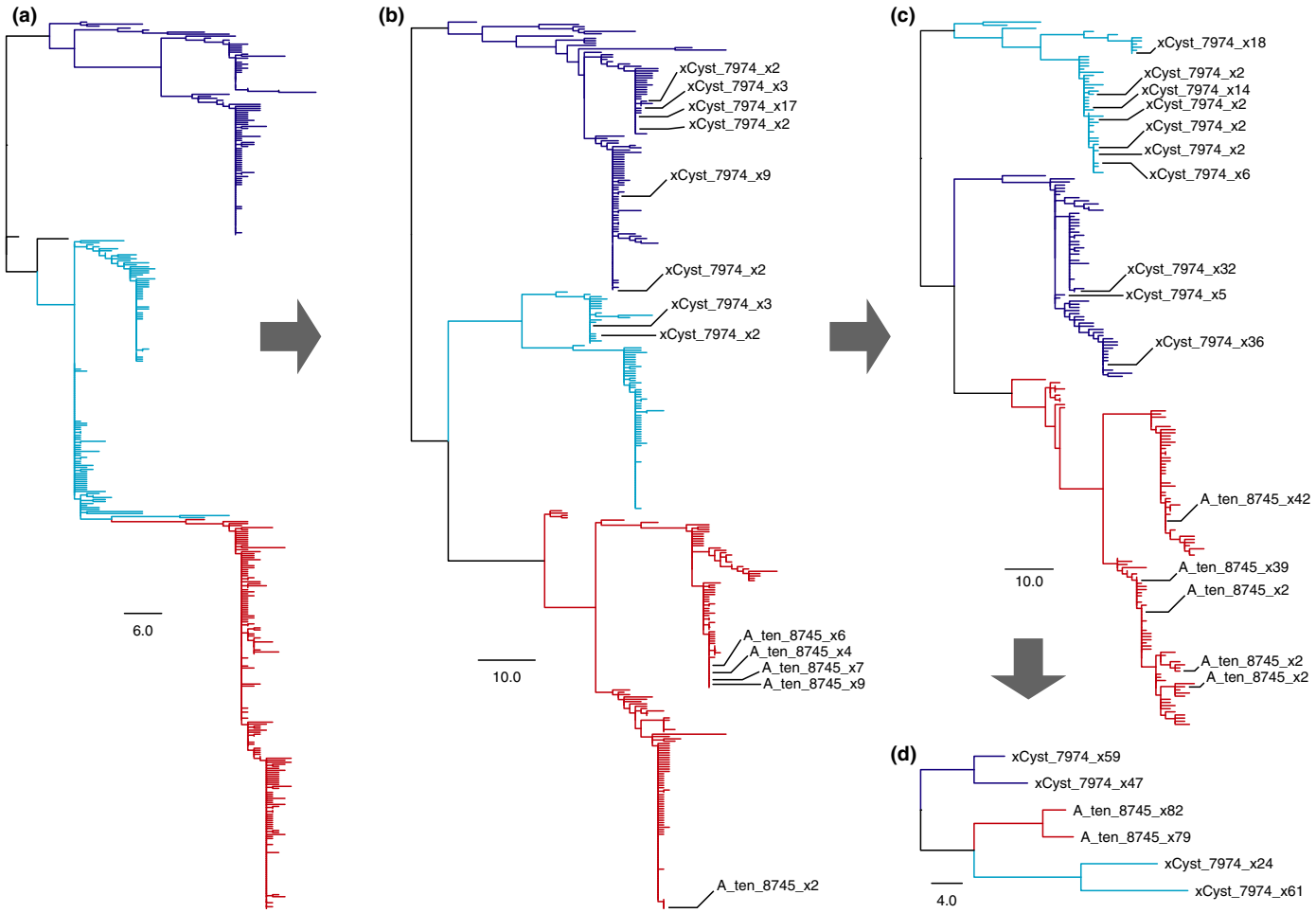

Fig. 3 Example of Pipeline for Untangling Reticulate Complexes (PURC)'s iterative clustering and chimera removal approach. The displayed trees are most parsimonious trees of the *APP* locus for barcode 12 at each round of the analysis (in these data, two accessions had barcode 12: an accession of the *Cystopteris–Gymnocarpium* tetraploid hybrid × *Cystocarpium roskamianum* (#7974) and a tetraploid accession of *Acystopteris tenuisecta* (#8745)). Cluster consensus sequences are labeled with their taxon and size (singleton sequences are unlabeled), and branches are colored according to their phylogenetic identity: dark blue for the *Gymnocarpium* homeologs of × *Cystocarpium*; light blue for the *Cystopteris* homeologs of × *Cystocarpium*; and red for *Acystopteris*. (a) The raw sequencing reads are highly heterogeneous, with far more variants than expected given the ploidy of the included accessions. (b) After one round of clustering and chimera removal, dominant sequences start to emerge, and the tree structure is clearer. This pattern continues with additional clustering and chimera removal (c). (d) After a final round of clustering/chimera removal and the elimination of rare sequence types (those with coverage < 4), four alleles are recovered for × *Cystocarpium* and two for *Acystopteris*, consistent with earlier results for these species.

accessions that were heterozygous at a subset of the loci, necessitating the addition of corresponding blank sequences in the other loci (Table 7). Each of the four runs converged to the same region of parameter space by the 100 millionth generation; we took the last 200 million generations from one of the runs as our posterior sample. This sample comprised 10 000 trees and post-burnin effective sample sizes for all parameters were > 500.

Discussion

Single-molecule amplicon sequencing of polyploids

The combined wet laboratory and bioinformatics approach we describe here is highly effective in generating low-copy nuclear data, including sequences from each homeolog present in the polyploid accessions. Requiring only PCR with barcoded primers, and capable of generating high-coverage data for four loci for *c.* 50 mostly polyploid accessions per run (or equivalent; Table S1; all-runs voucher table, available in the Dryad Digital Repository: doi: 10.5061/dryad.dj82k), this approach is both cost-effective and time-efficient. With the upcoming PacBio Sequel platform, the sequencing costs should go down significantly and even more samples can be pooled into one run. However, even with high sequencing coverage, analysis of the amplicon sequencing results was challenging – sequence errors introduced during PCR, PCR-mediated recombinants (chimeras), and errors in the PacBio sequencing all appeared to be common in our data (Figs 3, 5). Fortunately, the PURC pipeline, with its iterative clustering and chimera-killing steps, was able to detect and correct the majority of these errors. For our data, regime c (default UCHIME settings, and clustering cut-offs of 99.7, 99.5, 99.5, and 99.5% similarity, followed by the

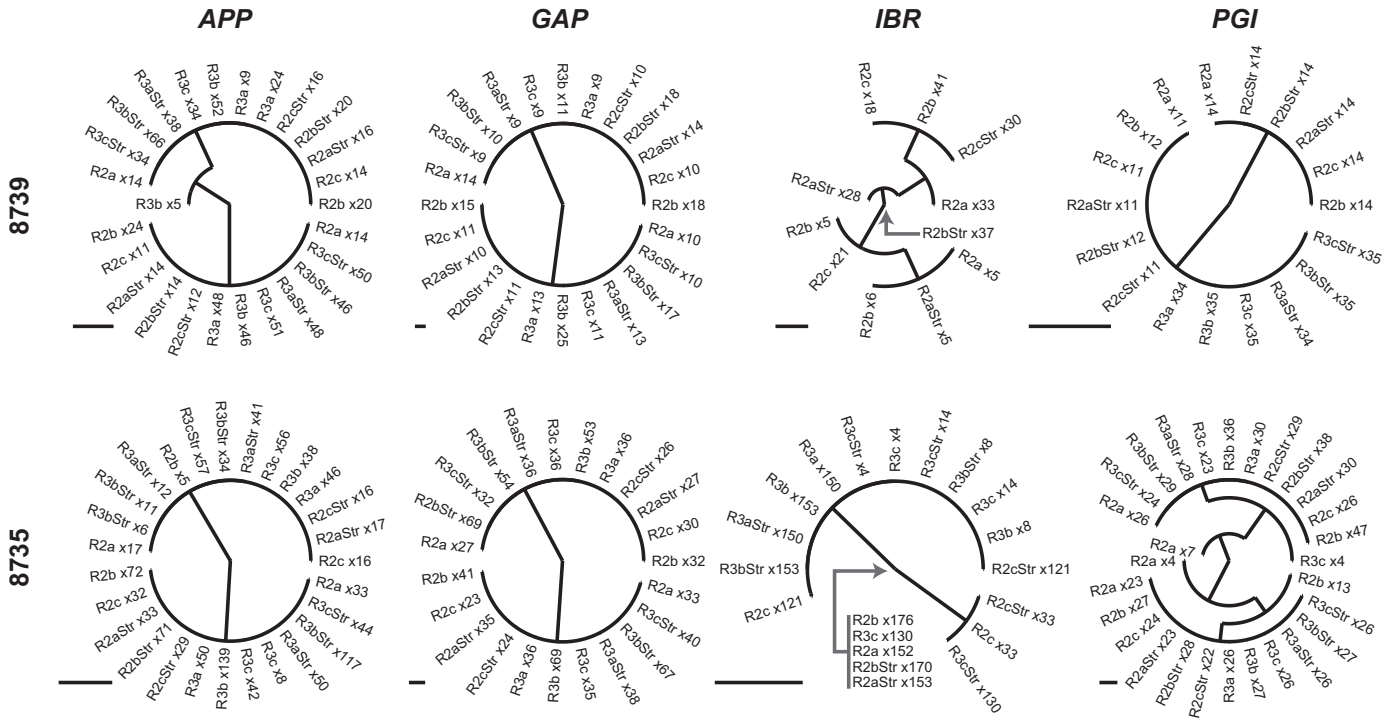

Fig. 4 Repeatability of allele inference. Two allotetraploid accessions (*Cystopteris moupinensis* #8735 and *Gymnocarpium oyamense* #8739) were included in both the R2 and R3 sequencing rounds, allowing for the examination of repeatability across different repetitions of the molecular laboratory workflow. Unrooted most-parsimonious trees for each accession, for each locus, show high levels of repeatability across sequencing rounds (R2 vs R3) and across PURC analysis regimes (a, c, e, aStr, cStr and eStr; see Table 3); in the majority of cases the different analysis regimes infer the same two sequences for each accession, in both rounds. In cases where the regimes vary in the sequences they infer, as in the *IBR* and *PGI* loci for accession 8735, there are still two clear 'good' sequence types inferred, with the other sequences occupying a more central position in the tree, indicating that they are the result of chimerism or overclustering in the Pipeline for Untangling Reticulate Complexes (PURC). The R3 *IBR PCR* amplification for accession 8739 failed, so that locus–accession combination was not included in the R3 sequencing run. Bars indicate a single substitution.

Table 4 Summary of Pipeline for Untangling Reticulate Complexes (PURC) results for the six analysis regimes on each of the three sequencing data sets

		Locus			
Regime	Data set	*APP*	*GAP*	*IBR*	*PGI*
a	R2	99 (26 ± 15.2)	88 (27.4 ± 17.2)	118 (29.9 ± 28.8)	98 (25.7 ± 25.4)
	R3	110 (48.9 ± 61.1)	90 (37.3 ± 35.2)	70 (54.6 ± 58.2)	69 (51.9 ± 70.8)
	R4	106 (34.9 ± 29.9)	97 (41 ± 35.4)	124 (48.5 ± 61.3)	74 (34 ± 28.1)
b	R2	102 (36.8 ± 24.2)	91 (34.4 ± 23.8)	101 (43.4 ± 43.4)	86 (34.9 ± 31.8)
	R3	123 (62.8 ± 98.7)	95 (51.5 ± 49.4)	79 (59.6 ± 70)	66 (67.4 ± 87.4)
	R4	109 (50.2 ± 50.3)	98 (60.2 ± 56)	128 (58.2 ± 74)	69 (44.1 ± 38.1)
c	R2	99 (25 ± 16.1)	92 (24.8 ± 17.1)	114 (29.5 ± 26.8)	97 (25.2 ± 24.8)
	R3	113 (48.3 ± 72.4)	96 (33.1 ± 32.2)	79 (46.7 ± 52.2)	69 (50.3 ± 68.2)
	R4	106 (34.9 ± 29.9)	97 (41 ± 35.4)	124 (48.5 ± 61.3)	74 (34 ± 28.1)
aStr	R2	95 (26.4 ± 14.9)	84 (27.7 ± 17.3)	93 (33.7 ± 28.1)	89 (27.6 ± 25.6)
	R3	108 (48.5 ± 60.2)	84 (39 ± 35.5)	66 (54.4 ± 55.1)	61 (55.8 ± 72.4)
	R4	92 (37.9 ± 28.5)	91 (44.8 ± 34.1)	94 (61.5 ± 66.6)	65 (37.8 ± 28.9)
bStr	R2	90 (38.8 ± 23.7)	88 (33.3 ± 22.7)	82 (45.3 ± 37.5)	83 (33.7 ± 30.7)
	R3	106 (66.7 ± 94.2)	91 (50.6 ± 47.5)	61 (67.1 ± 73)	60 (66.3 ± 85.5)
	R4	100 (48.1 ± 39)	93 (57.8 ± 55.3)	98 (66.1 ± 75.4)	62 (44 ± 35.9)
cStr	R2	98 (24 ± 15.6)	91 (24.8 ± 17)	94 (31.4 ± 26.1)	91 (26.1 ± 24.7)
	R3	106 (48.6 ± 55.6)	87 (35.9 ± 32.7)	70 (47.1 ± 48)	61 (53.6 ± 69.7)
	R4	98 (36 ± 29.3)	90 (42.7 ± 31.6)	101 (55.4 ± 65.2)	64 (35.9 ± 28.5)

Regime details are listed in Table 3. The counts are the number of final consensus sequences ('alleles') inferred for each regime/data set/locus combination; they are followed by the mean coverage of each allele and the ± SD of the coverage. Only successful PCR amplifications are included.

Table 5 Allele counts and coverage, by accession

	Ploidy	*APP*	*GAP*	*IBR*	*PGI*
A_jap_7978	2x	1 (22)	1 (23)	1 (103)	1 (92)
A_tai_4870	4x	3 (38, 32, 4)	2 (26, 24)	4 (62, 4, 68, 14)	2 (67, 59)
A_tai_6137	4x	2 (12, 14)	2 (13, 12)	2 (33, 29)	2 (22, 40)
A_ten_4831	2x	3 (51, 27, 8)	1 (71)	1 (163)	1 (67)
A_ten_8704	4x	2 (13, 60)	3 (23, 6, 29)	3 (44, 4, 27)	2 (111, 4)
A_ten_8745	4x	3 (8, 22, 52)	2 (39, 35)	4 (35, 25, 12, 19)	2 (11, 55)
C_alp_7920	6x	4 (41, 64, 26, 21)	3 (24, 28, 27)	4 (61, 66, 9, 36)	3 (9, 10, 12)
C_bul_7650	2x	2 (45, 5)	na	4 (41, 23, 7, 4)	2 (21, 8)
C_dia_5316	4x	3 (11, 11, 18)	2 (11, 11)	2 (61, 45)	3 (18, 4, 28)
C_dia_6380	4x	2 (26, 36)	2 (21, 20)	na	2 (29, 18)
C_dou_6378	6x	3 (54, 34, 12)	4 (16, 13, 13, 12)	5 (76, 6, 4, 8, 20)	6 (5, 12, 4, 14, 7, 5)
C_fra_7009	4x	2 (11, 18)	1 (100)	2 (52, 41)	2 (20, 36)
C_fra_7248	4x	2 (66, 29)	2 (41, 45)	5 (27, 5, 7, 48, 13)	2 (26, 16)
C_fra_7625	6x	3 (27, 16, 22)	4 (25, 26, 4, 24)	4 (6, 14, 4, 22)	2 (6, 4)
C_hau_7034	6x	4 (41, 12, 5, 38)	3 (40, 27, 33)	7 (45, 5, 4, 56, 16, 26, 4)	4 (18, 11, 4, 14)
C_lau_8484	6x	3 (27, 27, 52)	3 (10, 17, 17)	3 (5, 10, 20)	2 (14, 64)
C_mem_6732	2x	1 (78)	1 (45)	1 (49)	1 (72)
C_mon_6969	4x	2 (22, 39)	2 (25, 14)	4 (55, 6, 6, 64)	2 (64, 4)
C_mon_7943	4x	2 (30, 25)	2 (23, 29)	2 (111, 11)	2 (11, 47)
C_mou_4861	2x	1 (30)	1 (63)	2 (56, 4)	1 (59)
C_mou_8735	4x	2 (14, 11)	2 (11, 10)	2 (18, 21)	2 (11, 14)
C_pel_6055	4x	2 (14, 12)	1 (9)	2 (26, 10)	2 (29, 25)
C_pel_6060	4x	na	1 (6)	2 (29, 14)	2 (16, 22)
C_pro_6359	2x	2 (25, 15)	2 (19, 27)	1 (47)	1 (25)
C_pro_6362	2x	3 (9, 4, 15)	2 (14, 18)	2 (43, 8)	2 (11, 13)
C_ree_6342	4x	2 (32, 49)	2 (15, 27)	na	2 (15, 10)
C_sud_7980	4x	2 (16, 11)	3 (9, 17, 9)	na	2 (7, 20)
C_sud_8674	4x	2 (9, 58)	2 (22, 24)	2 (66, 37)	2 (12, 31)
C_tas_6379	4x	2 (8, 6)	2 (19, 12)	1 (34)	2 (11, 19)
C_tenu_6387	4x	3 (41, 23, 6)	2 (67, 35)	3 (29, 12, 46)	2 (27, 22)
C_uta_6848	4x	2 (24, 20)	2 (7, 8)	na	1 (8)
G_aok_7984	4x	2 (19, 15)	2 (48, 41)	2 (37, 35)	2 (21, 26)
G_app_7639	2x	2 (6, 16)	2 (14, 7)	1 (33)	3 (4, 6, 5)
G_app_7800	2x	1 (30)	1 (62)	1 (59)	1 (34)
G_con_6979	4x	1 (10)	2 (5, 13)	2 (47, 37)	2 (36, 44)
G_dis_4710	2x	1 (23)	1 (8)	3 (39, 37, 8)	2 (22, 20)
G_dis_7751	2x	1 (40)	2 (12, 34)	3 (20, 5, 14)	1 (32)
G_dry_7981	4x	2 (29, 17)	3 (18, 17, 24)	3 (4, 4, 17)	3 (11, 4, 5)
G_dry_8031	4x	3 (5, 26, 25)	2 (51, 38)	3 (8, 40, 17)	1 (77)
G_jes_6059	4x?	na	na	5 (13, 15, 6, 10, 4)	na
G_oya_6399	2x	1 (43)	1 (75)	2 (53, 23)	2 (48, 12)
G_oya_8702	4x	2 (8, 9)	2 (42, 36)	2 (71, 5)	4 (27, 7, 12, 7)
G_oya_8739	4x	2 (16, 32)	2 (23, 30)	2 (121, 33)	2 (26, 24)
G_rem_4862	4x	2 (25, 16)	2 (11, 11)	2 (18, 22)	2 (35, 37)
G_rob_7945	4x	3 (15, 9, 51)	2 (25, 26)	2 (50, 30)	2 (153, 12)
G_sp_7979	4x	2 (30, 21)	2 (32, 16)	2 (20, 21)	2 (56, 17)
xCyst_7974	4x	4 (26, 51, 44, 15)	4 (15, 9, 25, 22)	4 (6, 4, 21, 20)	4 (27, 9, 12, 11)

The data summarized are from the R2 sequencing run, analyzed with regime c (see Table 3). Full voucher data are available in the all-runs table, available in the Dryad Digital Repository (doi: 10.5061/dryad.dj82k). na, failed PCR amplification.

final 99.9% clustering; Table 3) generally performed best. It tended to slightly overestimate the number of biological sequences present (Table 6), which was preferable, at least for our purposes, to overlumping the raw reads. However, rather than relying on a single 'best' analysis regime, we found that a particularly helpful approach was to run PURC multiple times with different analysis regimes (different clustering-similarity settings and different chimera-detecting stringencies) and look for commonalities among the results from those analyses (see Figs 4, S1). This was especially true because the results for individual accessions and loci were often idiosyncratic. For example, for some locus–accession combinations the more inclusive clustering settings resulted in the merger of two distinct alleles apparent under less inclusive settings, whereas for another accession the inclusive settings revealed an allele that was below the coverage threshold in the less inclusive settings. These are the minority of cases, however; in nearly all instances the different analysis regimes inferred the same alleles, and the

rarely inferred sequences were identifiable as chimeras, clustering artifacts, or the result of PCR error (Fig. 4; all-regimes trees, available in the Dryad Digital Repository: doi: 10.5061/dryad.dj82k).

This method for rapidly and cheaply generating low-copy nuclear data has the potential to facilitate a wide variety of evolutionary investigations. Most obviously, it allows researchers to easily generate information-rich long-read data for phylogenetic inference of polyploid complexes, an enterprise that previously relied heavily on expensive and time-consuming cloning approaches (e.g. Brysting *et al.*, 2007; Kim *et al.*, 2008; Mason-Gamer, 2008; Grusz *et al.*, 2009; Ishikawa *et al.*, 2009; Nitta *et al.*, 2011; Dyer *et al.*, 2012; Sessa *et al.*, 2012; Metzgar *et al.*, 2013; Sigel *et al.*, 2014). In addition, such low-copy nuclear data are useful for investigating questions unrelated to polyploidy, such as 'classical' nonpolyploid phylogenetic inference (Zhang *et al.*, 2012), inference of hybridization (Govindarajulu *et al.*, 2011; Tripp *et al.*, 2013; Rothfels *et al.*, 2015), horizontal gene transfer (Li *et al.*, 2014), and studies of gene family and genome evolution (Popp & Oxelman, 2004; Rauscher *et al.*, 2004; Flagel & Wendel, 2009; Weiss-Schneeweiss *et al.*, 2011; Larsen *et al.*, 2014; Li *et al.*, 2015). In particular, this approach allows researchers to easily sequence both alleles for heterozygous accessions (of any ploidy level), providing highly informative dominant markers for coalescent-based analyses or inferences of

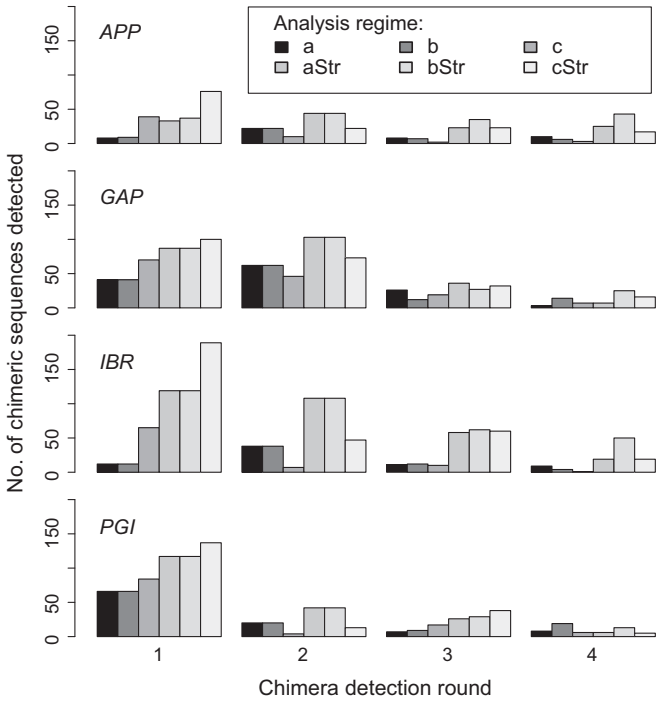

Fig. 5 Number of PCR-mediated recombinant sequences (chimeras) found in the R2 data set. The six analysis regimes (Table 3) are indicated by different shades of gray. The four rounds of cluster detection are sequential, and each is preceded by a round of clustering. The chimeric sequences detected may be raw sequencing reads, or may be consensus sequences from clusters of multiple reads.

Table 6 Accuracy of allele number inference

Regime	*APP*	*GAP*	*IBR*	*PGI*
a	16 (0.178)	7 (0.022)	31 (0.674)	14 (0.217)
b	22 (0.267)	16 (0.089)	32 (0.326)	15 (−0.065)
c	21 (0.333)	10 (0.178)	28 (0.605)	12 (0.261)
aStr	16 (0.089)	8 (−0.089)	12 (0.093)	4 (0)
bStr	12 (0)	12 (−0.089)	25 (−0.209)	12 (−0.13)
cStr	22 (0.222)	9 (0.156)	16 (0.14)	12 (0.13)

Values indicate the absolute value of the difference between the inferred number of alleles and the true number, summed across all accessions (if Pipeline for Untangling Reticulate Complexes (PURC) perfectly inferred the biological sequences, these numbers would all be zero), followed, in brackets, by the per-accession average deviation from the truth (positive numbers indicate that the method infers too many sequences on average and negative numbers indicate the opposite). The 'true' allele number for each accession for each locus was determined by visual inspection of the phylogenies of the results from all analysis regimes, from the read coverage, and from prior information on the ploidy of each accession; see the Materials and Methods section and Table 5. The data in this table are from the R2 sequencing run. Analysis regime characteristics are listed in Table 3.

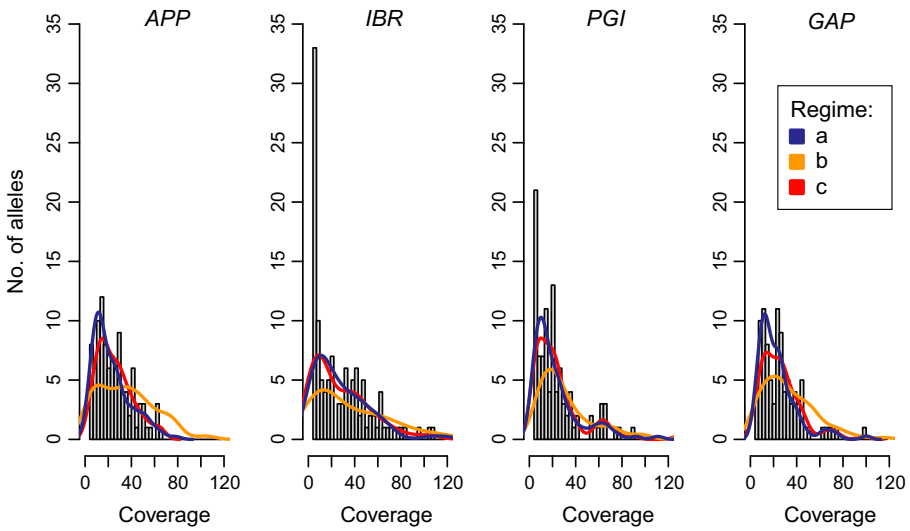

Fig. 6 Depth of coverage for the inferred allele sequences. The histogram shows the number of inferred alleles at each depth of coverage, for Pipeline for Untangling Reticulate Complexes (PURC) analysis regime a (see Table 3); color curves show the smoothed fits for each of the analysis regimes a, b and c.

Table 7 Data set characteristics for the AlloPPNet analyses

Data set	Tips	Alignment length	Missing data (%)	Pars. informative sites
APP	44	1006	6.6	127
GAP	48	1068	7.7	212
IBR	45	862	4.2	103
PGI	48	1132	15.5	172
Combined	53	4068	19.5	614

Missing data for the individual-locus data sets are attributable to indels; missing data for the combined data set include indels and blank sequences inserted for loci that were, for example, homozygous at an accession where other loci were heterozygous.

population structure. Finally, our laboratory protocol and bioinformatics pipeline are not restricted to PacBio-generated low-copy nuclear data. Purc works with amplicon sequences generated by any sequencing platform and can accommodate dual-barcoded data, which allows more samples to be pooled into one run. Our approach is also effective for sequencing plastid and mitochondrial data, and such loci can be pooled with nuclear data in a single PacBio sequencing run (data not shown; see also Fior *et al.*, 2013; Uribe-Convers *et al.*, 2016).

The main drawbacks of our method include its reliance on PCR – it thus requires primer design, which can be a significant difficulty for poorly studied groups (e.g. Schuettpelz *et al.*, 2008; Rothfels *et al.*, 2013a), and can result in difficulties in removing PCR-mediated chimeras (Figs 3–5). However, even modest genomic resources can greatly reduce the challenges of primer design (e.g. Uribe-Convers *et al.*, 2016), and future refinements of the molecular laboratory approaches and bioinformatics pipelines could help address both these issues. Specifically, there is great potential in combining our approach with automated primer design (e.g. MarkerMiner; Chamala *et al.*, 2015) and microfluidic PCR (e.g. Uribe-Convers *et al.*, 2016). Additional potential improvements could be gained from modifications to the PCR protocol (the use of high-fidelity polymerases, and pooling multiple reactions for each accession), the library preparation (incorporating a more rigorous method of standardizing DNA concentrations across reactions, and adding library purifications; Pacific Biosciences, 2016), and the chimera detection steps (incorporating reference-based chimera detection).

Cystopteridaceae case study: allopolyploid species network

Our inferred multilabeled Cystopteridaceae 'species tree' is generally consistent with earlier gene trees of the family, including those inferred from plastid (Rothfels *et al.*, 2013b, 2014; Wei & Zhang, 2014) and single-locus nuclear data (Rothfels *et al.*, 2014): *Acystopteris* and *Cystopteris* are sister genera; *Gymnocarpium* is sister to the rest of the family; *Cystopteris* comprises the *Cystopteris montana* clade, *Cystopteris sudetica* clade, *Cystopteris bulbifera* clade, and the *C. fragilis* complex; *C. montana* is sister to the rest of the genus; and *Cystopteris protrusa* is sister to the rest of the *C. fragilis* complex (Fig. 7a). Within *Gymnocarpium*, the concordance with earlier studies is less clear. The resolution of deeply isolated *Gymnocarpium robertianum* and *Gymnocarpium disjunctum* clades mirrors the results of Rothfels *et al.* (2013b); however, we do not find their 'core *Gymnocarpium*' clade. Instead, that group is rendered paraphyletic in our phylogeny by the placement of the *G. disjunctum* clade (Fig. 7a). This result, however, is perhaps not surprising, given the historical difficulty in finding support for the deep relationships within *Gymnocarpium* (Rothfels *et al.*, 2013b).

Our analyses also infer extensive reticulation (allopolyploidy) within the family (Fig. 7b–e). Rampant allopolyploidy in Cystopteridaceae has been inferred before with single-locus data (Rothfels *et al.*, 2014, 2015), but our study is the first to do so with multiple nuclear loci while accounting for incomplete lineage sorting. Novel inferences from our analyses include: tetraploid *G. oyamense* is relatively distantly related to our diploid *G. oyamense* accession, suggesting at least one undiscovered (or extinct) diploid species within *G. oyamense*; there is extensive reticulation within *Acystopteris*, a clade thought to include only three species (our data suggest that there are at least three distinct extant tetraploids and at least two extant diploids, plus at least two unsampled diploid lineages; Fig. 7); there are probably at least two unsampled diploids related to *C. montana*; and allopolyploidy is common in the *sudetica* clade (our data indicate at least six distinct lineages – two tetraploid and four diploid – within this clade which contains only three recognized species). Within the *C. fragilis* complex, our results are even more extreme. Our sample includes two diploid accessions (providing the first evidence that *C. membranifolia* is diploid; Mickel, 1972) and five distinct tetraploids, two of which would be referred to *Cystopteris fragilis* even in the strictest application of that name. Furthermore, analysis of even this small sample implies the existence of approximately nine unsampled diploid lineages, providing an indication of extensive unsampled diversity within the complex.

Our data also reveal further evidence for wide hybridization within the family (Fig. 7e). First, within *Cystopteris*, they provide the first sequence-based evidence that the parents of *Cystopteris utahensis* span two major clades, with one parent being *C. bulbifera* and the other a member of the *C. fragilis* complex; this parentage was previously hypothesized based on isozyme profiles and the intermediate morphology of *C. utahensis* (Haufler & Windham, 1991). Second, they corroborate the single-locus inference that × *Cystocarpium roskamianum* is an allotetraploid hybrid between *G. dryopteris* and a member of the *C. fragilis* complex (Rothfels *et al.*, 2015) and that it formed very recently – it shows very little divergence from either of its parents. *Cystopteris* and *Gymnocarpium* last shared a common ancestor *c.* 60 million yr ago (Rothfels *et al.*, 2015), making the hybridization event that formed × *Cystocarpium* one of the deepest yet documented.

Polyploid evolution

In addition to specific details of Cystopteridaceae phylogeny, our polyploid species tree (Fig. 7) provides evidence for some general patterns of polyploid evolution. For example, most of the polyploidization events inferred in our Cystopteridaceae sample are attributable to allopolyploidy rather than autopolyploidy. Early

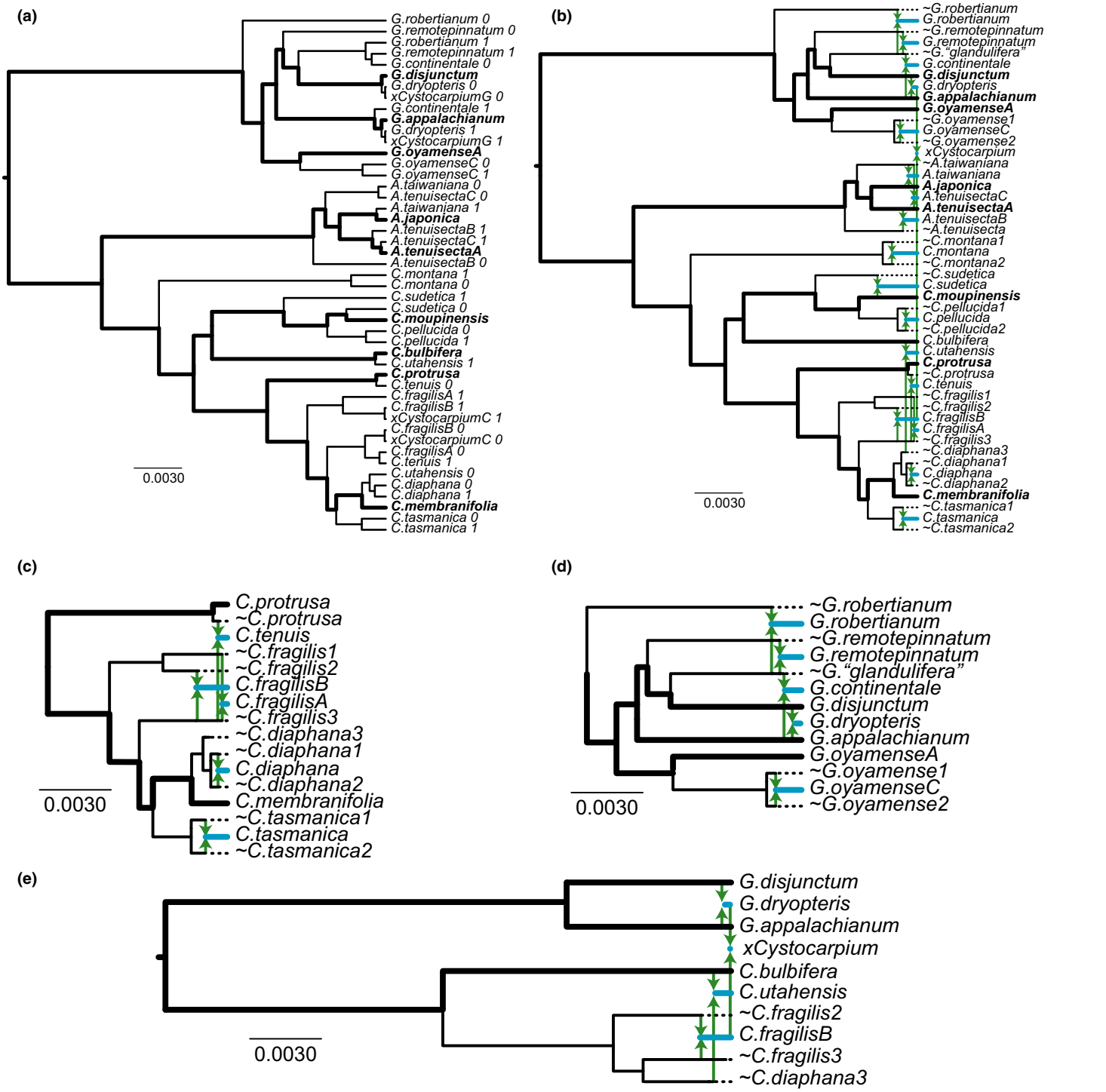

Fig. 7 Polyploid evolution in Cystopteridaceae. (a) The maximum clade credibility multi-labeled tree inferred by AlloPPNet (Jones *et al.*, 2013) from four single-copy nuclear loci. The subtree connecting extant diploid species has thickened branches and extant diploid species names are in bold face. Zeros and ones following species' names indicate homeolog pairs (each tetraploid accession will have two homeologs present). (b) An explicit allopolyploid network consistent with the multi-labeled tree. Green arrows indication genome donors in polyploidy events. Dashed lines indicate unsampled diploid lineages (these may be extinct, or extant but unsampled). Blue branches indicate tetraploid lineages. (c, d) Close-ups of the (c) *Cystopteris fragilis* and (d) *Gymnocarpium* portions of the network. (e) A pruned version of the network showing the two 'deep hybrids' (× *Cystocarpium* and *Cystopteris utahensis*) and their relatives. (c–e) Colors and branch thickening follow (b).

polyploidy researchers tended to believe that allopolyploids were more common than autopolyploids (Stebbins, 1947; Grant, 1981); however, this consensus has started to move in the opposite direction (Ramsey & Schemske, 2002; Soltis *et al.*, 2007; Parisod *et al.*, 2010; Barker *et al.*, 2016). Our results support other recent studies (e.g. Popp *et al.*, 2005; Brysting *et al.*,

2007; Kim *et al.*, 2008; Shepherd *et al.*, 2008; Marcussen *et al.*, 2012, 2014; Arrighi *et al.*, 2014; Triplett *et al.*, 2014) in providing some limited empirical evidence that allopolyploids are indeed more common. Of particular note is the fact that our methodology, with its ability to detect all gene copies present in an individual accession, is a powerful means of detecting previously unrecognized allopolyploids; these appear to be common in our data. For example, the current taxonomic consensus for the genus *Acystopteris* is that it contains three diploid species (*Acystopteris japonica*, *Acystopteris tenuisecta*, and *Acystopteris taiwaniana*), each with an additional tetraploid cytotype (Rothfels, 2012; Wang *et al.*, 2013; efloras, 2016). If these polyploid cytotypes were assumed to be autopolyploids, as is the typical practice, then it would appear as if autopolyploidy predominates in *Acystopteris*. However, our analyses reveal the opposite pattern: all three of our sampled *Acystopteris* polyploids are allopolyploids, including at least two distinct allopolyploids within '*A. tenuisecta*'. In addition to examples such as *A. tenuisecta*, where polyploid cytotypes are shown to be allo- rather than autopolyploids, our data reveal cases of deeply isolated polyploids – such as *Cystopteris montana* – that are also allopolyploids, despite having no known close diploid relatives. Without the insights provided by low-copy nuclear data, these, too, would be assumed to be autopolyploid in origin.

Another broad pattern apparent in our data is the relatively recent formation of polyploids in Cystopteridaceae – the majority of inferred polyploidization events occur towards the tips of the tree (Fig. 7b). While the timing of polyploid formation is imprecise in some cases (the multi-labeled tree has limited information available about the time of formation if extant representatives of the progenitor diploids are unsampled), this 'twigginess' suggests that polyploids have high rates of extinction and thus the only polyploids sampled are those that have not yet had time to go extinct (Nee *et al.*, 1994). In addition, most of our sampled polyploids formed via hybridization events between diploids rather than by primary speciation of an ancestral polyploid lineage, suggesting that polyploid lineages in the Cystopteridaceae have lower speciation rates than do their diploid relatives. In fact, the closest any of the polyploids in our sample come to participating in a speciation event is the formation of the sterile tetraploid × *Cystocarpium* by the hybridization of two other tetraploids (Fig. 7b,e). Taken together, these two inferences support the 'dead-end' model of polyploid evolution, whereby new polyploids form regularly from diploid species, but are themselves prone to short evolutionary lives because of their high extinction and low speciation rates (this conclusion is increasingly supported by empirical study; reviewed in Mayrose *et al.*, 2014; Rothfels & Otto, 2016).

Acknowledgements

Robert Edgar and Graham Jones provided invaluable assistance with USEARCH and ALLOPPNET, respectively; Olivier Fedrigo, Graham Alexander, Nico Devos, and the other staff at the Duke Center for Computational Biology were instrumental in ensuring the success of our PacBio sequencing runs; and conversations with Olivier Fedrigo, Peter Larson, Dave Weisrock, Michael Windham, and Norm Wickett were critical in informing our experimental design. We also want to thank Li-Yaung Kuo for sharing his Illumina dual-barcoded data set for PURC testing. Finally, comments from Ben Dauphin, Forrest Freund, Ingrid Jordon-Thaden, the UBC Biodiversity Research Centre community (especially Sally Otto, Chris Muir, and Jeremy Draghi), three anonymous reviewers, and members of the Pryer lab greatly strengthened the manuscript. This research was supported by funding from a Natural Sciences and Engineering Research Council (Canada; NSERC) Postgraduate Scholarship–Doctoral and an NSERC Postdoctoral Fellowship to C.J.R., a National Science Foundation Graduate Research Fellowship (to F-W.L.) and National Science Foundation Doctoral Dissertation Improvement Grants to K.M.P. and C.J.R. (DEB-1110767) and to K.M.P. and F-W.L. (DEB-1407158).

Author contributions

C.J.R, K.M.P, and F-W.L. designed and implemented the study. C.J.R and F-W.L. performed the analyses, summarized the results, and drafted the manuscript.

References

Arrighi J-F, Chaintreuil C, Cartieaux F, Cardi C, Rodier-Goud M, Brown SC, Boursot M, D'hont A, Dreyfus B, Giraud E. 2014. Radiation of the Nod-independent *Aeschynomene* relies on multiple allopolyploid speciation events. *New Phytologist* **201**: 1457–1468.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. *New Phytologist* **210**: 391–398.

Barrow LN, Ralicki HF, Emme SA, Lemmon EM. 2014. Species tree estimation of North American chorus frogs (Hylidae: *Pseudacris*) with parallel tagged amplicon sequencing. *Molecular Phylogenetics and Evolution* **75**: 78–90.

Beck JB, Alexander PJ, Allphin L, Al-Shehbaz IA, Rushworth C, Bailey CD, Windham MD. 2011a. Data from: Does hybridization drive the transition to asexuality in diploid *Boechera*? Dryad Data Repository. doi: 10.5061/dryad.11p757 m0.

Beck JB, Alexander PJ, Allphin L, Al-Shehbaz IA, Rushworth CA, Bailey CD, Windham MD. 2011b. Does hybridization drive the transition to asexuality in diploid *Boechera*? *Evolution* **66**: 985–995.

Brassac J, Blattner FR. 2015. Species-level phylogeny and polyploid relationships in *Hordeum* (Poaceae) inferred by next-generation sequencing and *in silico* cloning of multiple nuclear loci. *Systematic Biology* **64**: 792–808.

Brysting AK, Oxelman B, Huber KT, Moulton V, Brochmann C. 2007. Untangling complex histories of genome mergings in high polyploids. *Systematic Biology* **56**: 467–476.

Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, Udall JA, Wilcox ER, Crandall KA. 2011. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution* **3**: 1312–1323.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Chamala S, García N, Godden GT, Krishnakumar V, Jordon-Thaden IE, Smet RD, Barbazuk WB, Soltis DE, Soltis PS. 2015. MarkerMiner 1.0: a new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* **3**: 1400115.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B *et al.* 2009. Biopython: freely available

Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

De Storme N, Mason A. 2014. Plant speciation through chromosome instability and ploidy change: cellular mechanisms, molecular factors and evolutionary relevance. *Current Plant Biology* **1**: 10–33.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24**: 332–340.

Doyle JD, Dickson EE. 1987. Preservation of plant samples for DNA restriction endonuclease analysis. *Taxon* **36**: 715–722.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**: 214.

Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack JH, dePamphilis CW. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* **10**: 1.

Dufresne F, Stift M, Vergilino R, Mable BK. 2013. Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology* **23**: 40–69.

Dyer RJ, Savolainen V, Schneider H. 2012. Apomixis and reticulate evolution in the *Asplenium monanthes* fern complex. *Annals of Botany* **110**: 1515–1529.

Eaton DAR. 2014. PyRAD: assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics* **30**: 1844–1849.

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

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–2200.

efloras 2016. *efloras*. Cambridge, MA, USA: Missouri Botanical Garden, St Louis, MO & Harvard University Herbaria. [WWW document] URL http://www.efloras.org [accessed 1 April 2016].

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al.* 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.

Estep MC, McKain MR, Vela Diaz D, Zhong J, Hodge JG, Hodkinson TR, Layton DJ, Malcomber ST, Pasquet R, Kellogg EA. 2014. Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences, USA* **111**: 15149–15154.

Feng YJ, Liu Q-F, Chen MY, Liang D, Zhang P. 2016. Parallel tagged amplicon sequencing of relatively long PCR products using the Illumina HiSeq platform and transcriptome assembly. *Molecular Ecology Resources* **16**: 91–102.

Fichot EB, Norman RS. 2013. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome* **1**: 1–5.

Fior S, Li M, Oxelman B, Viola R, Hodges SA, Ometto L, Varotto C. 2013. Spatiotemporal reconstruction of the *Aquilegia* rapid radiation through next-generation sequencing of rapidly evolving cpDNA regions. *New Phytologist* **198**: 579–592.

Flagel L, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. *New Phytologist* **183**: 557–564.

Gholami M, Bekele WA, Schondelmaier J, Snowdon RJ. 2012. A tailed PCR procedure for cost-effective, two-order multiplex sequencing of candidate genes in polyploid plants. *Plant Biotechnology Journal* **10**: 635–645.

Govindarajulu R, Hughes CE, Bailey CD. 2011. Phylogenetic and population genetic analyses of diploid *Leucaena* (Leguminosae; Mimosoideae) reveal cryptic species diversity and patterns of divergent allopatric speciation. *American Journal of Botany* **98**: 2049–2063.

Grant V. 1981. *Plant speciation*. New York, NY, USA: Columbia University Press.

Griffin PC, Robin C, Hoffmann AA. 2011. A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to *Poa* grasses. *BMC Biology* **9**: 19.

Grusz AL, Windham MD, Pryer KM. 2009. Deciphering the origins of apomictic polyploids in the *Cheilanthes yavapensis* complex (Pteridaceae). *American Journal of Botany* **96**: 1636–1645.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**: 160–174.

Haufler CH, Windham MD. 1991. New species of North American *Cystopteris* and *Polypodium*, with comments on their reticulate relationships. *American Fern Journal* **81**: 7–23.

Husband BC, Baldwin SJ, Suda J. 2013. The incidence of polyploidy in natural plant populations: major patterns and evolutionary processes. In: Greilhuber J, Dolezel J, Wendel FJ, eds. *Plant genome diversity*. Vienna, Austria: Springer, 255–276.

Ishikawa N, Yokoyama J, Tsukaya H. 2009. Molecular evidence of reticulate evolution in the subgenus *Plantago* (Plantaginaceae). *American Journal of Botany* **96**: 1627–1635.

Jones G. 2012. *Manual for using AlloppNET, AlloppMUL on real data. MANUAL: 1–10.* [WWW document] URL www.indriid.com/goteborg/2012-02-14-manual-real-data.pdf [accessed 1 January 2014].

Jones G, Sagitov S, Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Systematic Biology* **62**: 467–478.

Kim S-T, Sultan SE, Donoghue MJ. 2008. Allopolyploid speciation in *Persicaria* (Polygonaceae): insights from a low-copy nuclear region. *Proceedings of the National Academy of Sciences, USA* **105**: 12370–12375.

Larsen PA, Heilman AM, Yoder AD. 2014. The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms. *BMC Genomics* **15**: 720.

Larsson A. 2010. *abioscripts*. [WWW document] URL http://ormbunkar.se/phylogeny/abioscripts/ [accessed 1 January 2015].

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large data sets. *Bioinformatics* **30**: 3276–3278.

Li F-W, Rothfels CJ, Melkonian M, Villarreal JC, Stevenson DW, Graham SW, Wong GKS, Mathews S, Pryer KM. 2015. The origin and evolution of phototropins. *Frontiers in Plant Science* **6**: 637.

Li F-W, Villarreal JC, Kelly S, Rothfels CJ, Melkonian M, Frangedakis E, Ruhsam M, Sigel EM, Der JP, Pittermann J *et al.* 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National Academy of Sciences, USA* **111**: 6672–6677.

Mable BK. 2013. Polyploids and hybrids in changing environments: winners or losers in the struggle for adaptation? *Heredity* **110**: 95–96.

Maddison WP, Knowles L. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* **55**: 21–30.

Marcussen T, Heier L, Brysting AK, Oxelman B, Jakobsen KS. 2014. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). *Systematic Biology* **64**: 84–101.

Marcussen T, Jakobsen KS, Danihelka J, Ballard HE, Blaxland K, Brysting AK, Oxelman B. 2012. Inferring species networks from gene trees in high-polyploid North American and Hawaiian violets (*Viola*, Violaceae). *Systematic Biology* **61**: 107–126.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal* **17**: 10–12.

Martin SL, Husband BC. 2009. Influence of phylogeny and ploidy on species ranges of North American angiosperms. *Journal of Ecology* **97**: 913–922.

Mason-Gamer RJ. 2008. Allohexaploidy, introgression, and the complex phylogenetic history of *Elymus repens* (Poaceae). *Molecular Phylogenetics and Evolution* **47**: 598–611.

Mayrose I, Zhan SH, Rothfels CJ, Arrigo N, Barker MS, Rieseberg LH, Otto SP. 2014. Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis *et al.* (2014). *New Phytologist* **206**: 27–35.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**: 526–538.

Meseguer AS, Sanmartín I, Marcussen T, Pfeil BE. 2014. Utility of low-copy nuclear markers in phylogenetic reconstruction of *Hypericum* L. (Hypericaceae). *Plant Systematics and Evolution* **300**: 1503–1514.

Metzgar JS, Alverson ER, Chen S, Vaganov AV, Ickert-Bond SM. 2013. Diversification and reticulation in the circumboreal fern genus *Cryptogramma*. *Molecular Phylogenetics and Evolution* **67**: 589–599.

Mickel JT. 1972. A "filmy fern" in the genus *Cystopteris*. *American Fern Journal* **62**: 93–95.

Miller MA, Pfeiffer W, Schwartz T 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE)*. New Orleans, LA, USA: IEEE, 1–8.

Minaya M, Díaz-Pérez A, Mason-Gamer R, Pimentel M, Catalán P. 2015. Evolution of the beta-amylase gene in the temperate grasses: non-purifying selection, recombination, semiparalogy, homeology and phylogenetic signal. *Molecular Phylogenetics and Evolution* **91**: 68–85.

Moore WS. 1995. Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees. *Evolution* **49**: 718–726.

Nee S, Holmes EC, May RM, Harvey PH. 1994. Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London B* **344**: 77–82.

Nitta JH, Ebihara A, Ito M. 2011. Reticulate evolution in the *Crepidomanes minutum* species complex (Hymenophyllaceae). *American Journal of Botany* **98**: 1782–1800.

O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea G, Weisrock DW. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology* **22**: 111–129.

Otto SP. 2007. The evolutionary consequences of polyploidy. *Cell* **131**: 452–462.

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

Pacific Biosciences 2015. *Revolutionize genomics with SMRT Sequencing*. [WWW document] URL http://pacificbiosciences.com/brochure 1–9 [accessed 24 September 2015].

Pacific Biosciences 2016. *Procedure checklist – amplicon template preparation and sequencing*. [WWW document] URL http://www.pacb.com/wp-content/uploads/Procedure-Checklist-Amplicon-Template-Preparation-and-Sequencing.pdf [accessed 1 February 2016].

Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. *New Phytologist* **186**: 5–17.

Popp M, Erixon P, Eggens F, Oxelman B. 2005. Origin and evolution of a circumpolar polyploid species complex in *Silene* (Caryophyllaceae) inferred from low copy nuclear RNA polymerase introns, rDNA, and chloroplast DNA. *Systematic Botany* **30**: 302–313.

Popp M, Oxelman B. 2004. Evolution of a RNA polymerase gene family in *Silene* (Caryophyllaceae) – incomplete concerted evolution and topological congruence among paralogues. *Systematic Biology* **53**: 914–932.

R Core Team 2013. *R: a language and environment for statistical computing*, v.3.0.1. Vienna, Austria: R Foundation for Statistical Computing.

Rambaut A, Suchard M, Drummond AJ. 2007. *Tracer v1.6*. [WWW document] URL http://tree.bio.ed.ac.uk/software/tracer/ [accessed 1 April 2016].

Ramsey J, Ramsey TS. 2014. Ecological studies of polyploidy in the 100 years following its discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**: 20130352.

Ramsey J, Schemske DW. 2002. Neopolyploidy in flowering plants. *Annual Review of Ecology and Systematics* **33**: 589–639.

Rauscher JT, Doyle JJ, Brown AHD. 2004. Multiple origins and nrDNA Internal Transcribed Spacer homeologue evolution in the *Glycine tomentella* (Leguminosae) allopolyploid complex. *Genetics* **166**: 987–998.

Rothfels CJ. 2012. *Phylogenetics of Cystopteridaceae: reticulation and divergence in a cosmopolitan fern family*. PhD thesis, Duke University, Durham, NC, USA.

Rothfels CJ, Johnson AK, Hovenkamp PH, Swofford DL, Roskam HC, Fraser-Jenkins CR, Windham MD, Pryer KM. 2015. Natural hybridization between parental lineages that diverged approximately 60 million years ago. *American Naturalist* **185**: 433–442.

Rothfels CJ, Johnson AK, Windham MD, Pryer KM. 2014. Low-copy nuclear data confirm rampant allopolyploidy in the Cystopteridaceae (Polypodiales). *Taxon* **63**: 1026–1036.

Rothfels CJ, Larsson A, Li F-W, Sigel EM, Huiet L, Burge DO, Ruhsam M, Graham SW, Stevenson DW, Wong GK-S *et al.* 2013a. Transcriptome-mining for single-copy nuclear markers in ferns. *PLoS ONE* **8**: e76957.

Rothfels CJ, Otto SP. 2016. Polyploid speciation. In: Kliman RM, ed. *Encyclopedia of evolutionary biology*. Oxford, UK: Academic Press, 317–326.

Rothfels CJ, Schuettpelz E. 2014. Accelerated rate of molecular evolution for vittarioid ferns is strong and not driven by selection. *Systematic Biology* **63**: 31–54.

Rothfels CJ, Windham MD, Pryer KM. 2013b. A plastid phylogeny of the cosmopolitan fern family Cystopteridaceae (Polypodiopsida). *Systematic Botany* **38**: 295–306.

Sang T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology* **37**: 121–147.

Schuettpelz E, Grusz AL, Windham MD, Pryer KM. 2008. The utility of nuclear *gapCp* in resolving polyploid fern origins. *Systematic Botany* **33**: 621–629.

Sessa EB, Zimmer EA, Givnish TJ. 2012. Unraveling reticulate evolution in North American *Dryopteris* (Dryopteridaceae). *BMC Evolutionary Biology* **12**: 104.

Shepherd LD, Perrie LR, Brownsey PJ. 2008. Low-copy nuclear DNA sequences reveal a predominance of allopolyploids in a New Zealand *Asplenium* fern complex. *Molecular Phylogenetics and Evolution* **49**: 240–248.

Sigel EM, Windham MD, Pryer KM. 2014. Evidence for reciprocal origins in *Polypodium hesperium* (Polypodiaceae): a fern model system for investigating how multiple origins shape allopolyploid genomes. *American Journal of Botany* **101**: 1476–1485.

Small RL, Cronn RC, Wendel JF. 2004. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* **17**: 145–170.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.

Soltis DE, Segovia-Salcedo MC, Jordon-Thaden I, Majure L, Miles NM, Mavrodiev EV, Mei W, Cortez MB, Soltis PS, Gitzendanner MA. 2014. Are polyploids really evolutionary dead-ends (again)? A critical reappraisal of Mayrose et al. (2011) *New Phytologist* **202**: 1105–1117.

Soltis DE, Soltis PD, Schemske DW, Hancock JF, Thompson JN, Husband BC, Judd WS. 2007. Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon* **56**: 13–30.

Stebbins GL Jr. 1947. Types of polyploids; their classification and significance. *Advances in Genetics* **1**: 403–429.

Swofford DL. 2002. *PAUP*: Phylogenetic analysis using parsimony (*and other methods)*, v.4.0a147. Sunderland, MA, USA: Sinauer.

Travers KJ, Chin CS, Rank DR, Eid JS, Turner SW. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* **38**: e159.

Triplett JK, Clark LG, Fisher AE, Wen J. 2014. Independent allopolyploidization events preceded speciation in the temperate and tropical woody bamboos. *New Phytologist* **204**: 66–73.

Tripp EA, Fatimah S, Darbyshire I, McDade LA. 2013. Origin of African *Physacanthus* (Acanthaceae) via wide hybridization. *PLoS ONE* **8**: e55677.

Twyford AD, Ennos RA. 2011. Next-generation hybridization and introgression. *Heredity* **108**: 179–189.

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**: e0148203.

Wang Z-R, Haufler CH, Pryer KM, Kato M. 2013. Cystopteridaceae. In: Wu ZY, Raven PH, Hong DY, eds. *Flora of China*. St Louis, MO, USA: Missouri Botanical Garden Press, 257–266.

Wei R, Zhang X-C. 2014. Rediscovery of *Cystoathyrium chinense* Ching (Cystopteridaceae): phylogenetic placement of the critically endangered fern species endemic to China. *Journal of Systematics and Evolution* **52**: 450–457.

Weiss-Schneeweiss H, Blöch C, Turner B, Villaseñor JL, Stuessy TF, Schneeweiss GM. 2011. The promiscuous and the chaste: frequent allopolyploid speciation and its genomic consequences in American daisies (*Melampodium* sect. *Melampodium*; Asteraceae). *Evolution* **66**: 211–228.

Wielstra B, Duijm E, Lagler P, Lammers Y, Meilink WRM, Ziermann JM, Arntzen JW. 2014. Parallel tagged amplicon sequencing of transcriptome-based genetic markers for *Triturus* newts with the Ion Torrent next-generation sequencing platform. *Molecular Ecology Resources* **14**: 1080–1089.

Wolfe KH, Li W-H, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* **84**: 9054–9058.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon P, Rieseberg LH. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences, USA* **106**: 13875.

Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* **31**: 3081–3092.

Yu Y, Dong J, Liu KJ, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences, USA* **111**: 16448–16453.

Zhang R, Liu T, Wu W, Li Y, Chao L, Huang L, Huang Y, Shi S, Zhou R. 2013. Molecular evidence for natural hybridization in the mangrove fern genus *Acrostichum*. *BMC Plant Biology* **13**: 74.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytologist* **195**: 923–937.

Zieliński P, Stuglik MT, Dudek K, Konczal M, Babik W. 2013. Development, validation and high-throughput analysis of sequence markers in nonmodel species. *Molecular Ecology Resources* **14**: 352–360.

Zimmer EA, Wen J. 2012. Using nuclear gene data for plant phylogenetics: progress and prospects. *Molecular Phylogenetics and Evolution* **65**: 774–785.

Zimmer EA, Wen J. 2015. Using nuclear gene data for plant phylogenetics: progress and prospects II. Next-gen approaches. *Journal of Systematics and Evolution* **53**: 371–379.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 A maximum-parsimony tree comparing the PURC PacBio-based allele inferences with those from Sanger sequencing.

Table S1 Voucher table for the ALLOPPNET analyses.

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.