Received: 23 November 2021 | Accepted: 11 April 2022

DOI: 10.1002/aps3.11484

SOFTWARE NOTE

Applications in Plant Sciences

Long-read genotyping with SLANG (Simple Long-read loci Assembly of Nanopore data for Genotyping)

Marco Dorfner | **Tankred Ott** | **Philipp Ott** | **Christoph Oberprieler**

Institute of Plant Sciences, University of Regensburg, Universitätsstraße 31, Regensburg, BY, D-93053, Germany

Correspondence

Marco Dorfner, Institute of Plant Sciences, University of Regensburg, Universitätsstraße 31, Regensburg, BY, D-93053, Germany.
Email: marco.dorfner@ur.de

Abstract

Premise: Most phylogenomic library preparation methods and bioinformatic analysis tools in restriction site–associated DNA sequencing (RADseq)/genotyping-by-sequencing (GBS) studies are designed for use with Illumina data. The lack of alternative bioinformatic pipelines hinders the exploration of long-read multi-locus data from other sequencing platforms. The Simple Long-read loci Assembly of Nanopore data for Genotyping (SLANG) pipeline enables locus assembly, orthology estimation, and single-nucleotide polymorphism (SNP) calling using Nanopore-sequenced multi-locus data.

Methods and Results: Two test libraries (*Leucanthemum* spp., *Senecio* spp.; Compositae) were prepared using an amplified fragment length polymorphism (AFLP)-based method to reduce genome complexity, then Nanopore-sequenced, and analyzed with SLANG. We identified 704 and 448 orthologous loci with 12,368 and 10,048 SNPs, respectively. The constructed phylogenetic networks were identical to a GBS network produced using *Leucanthemum* Illumina data and were consistent with *Senecio* species circumscriptions based on morphology.

Conclusions: SLANG identifies orthologous loci and extracts SNPs from long-read multi-locus Nanopore data for phylogenetic inference, population genetics, or phylogeographical studies. Combined with an AFLP-based library preparation, SLANG provides an easily scalable, cost-effective, and affordable alternative to Illumina-based RADseq/GBS procedures.

KEYWORDS

AFLP, genotyping, *Leucanthemum*, Nanopore, *Senecio*

Next-generation sequencing (NGS) techniques such as restriction site–associated DNA sequencing (RADseq; Baird et al., 2008; Davey and Blaxter, 2011; Peterson et al., 2012) and genotyping-by-sequencing (GBS; Elshire et al., 2011) are powerful and cost-effective genome-reduction methods currently used in biosystematic studies to generate anonymous multi-locus data for genotyping, phylogenetics, and species delimitation. For these reasons, RADseq and GBS are considered particularly valuable for use in taxonomically challenging groups (Razkin et al., 2016; Wagner et al., 2020a, 2020b), for which single-marker studies are insufficient for robust taxon delimitation. Many well-developed bioinformatic pipelines are available for the analysis of high-quality, short-read Illumina data, with *ipyrad* (Eaton and Overcast, 2020) and *Stacks* (Rochette and Catchen, 2017) being the most prominent ones for the de novo reconstruction of loci from single reads and the subsequent orthology estimation performed by grouping reads and loci based on similarity.

Although the high quality of short Illumina reads is attractive, long reads harbor enormous potential for phylogenetic studies. In addition to producing more accurate topologies, longer reads have the potential to detect deeper divergences more efficiently for very young and closely related taxa than shorter reads (Rubin et al., 2012; Cariou et al., 2013). Coalescent-based methods also

require highly robust gene trees, which are best inferred from long, non-recombining loci providing a large number of linked single-nucleotide polymorphisms (SNPs; McCormack et al., 2013).

Over the past few years, long-read sequencing has become more attractive because of the commercially available Nanopore sequencer MinION from Oxford Nanopore Technologies (Oxford, United Kingdom). Unlike Illumina sequencers, the MinION system can be established for a low initial cost in any lab. Reusable flow cells allow for the high scalability of sequencing projects, and the small size of the machine permits its use during field excursions for the rapid sequencing of freshly sampled material. For systematic studies, Nanopore sequencers seem to have been overlooked or omitted, most likely because of their relatively high error rates of around 10% (Magi et al., 2017; Fuselli et al., 2018) compared to median error rates of 0.087–0.613% observed across Illumina sequencers (Stoler and Nekrutenko, 2021). The lack of bioinformatic tools specifically tailored for the use of long, error-prone reads in the analysis of multi-locus data has likely also hindered progress toward longer-read genotyping. To kickstart the exploration of long-read genotyping, we developed a pipeline written in Python called Simple Long-read loci Assembly of Nanopore data for Genotyping (SLANG), which is able to analyze error-prone multi-locus data as produced by a Nanopore sequencer, comparable with the *ipyrad* or *Stacks* pipelines developed for Illumina data. Similar to these established pipelines, SLANG's workflow comprises three major segments (Figure 1). (1) During within-sample clustering, loci are assembled for each sample by grouping them according to read similarity. This is followed by (2) the among-samples clustering, in which locus orthology is estimated through the consensus read similarity clustering of all previously assembled loci across all samples. Finally, (3)

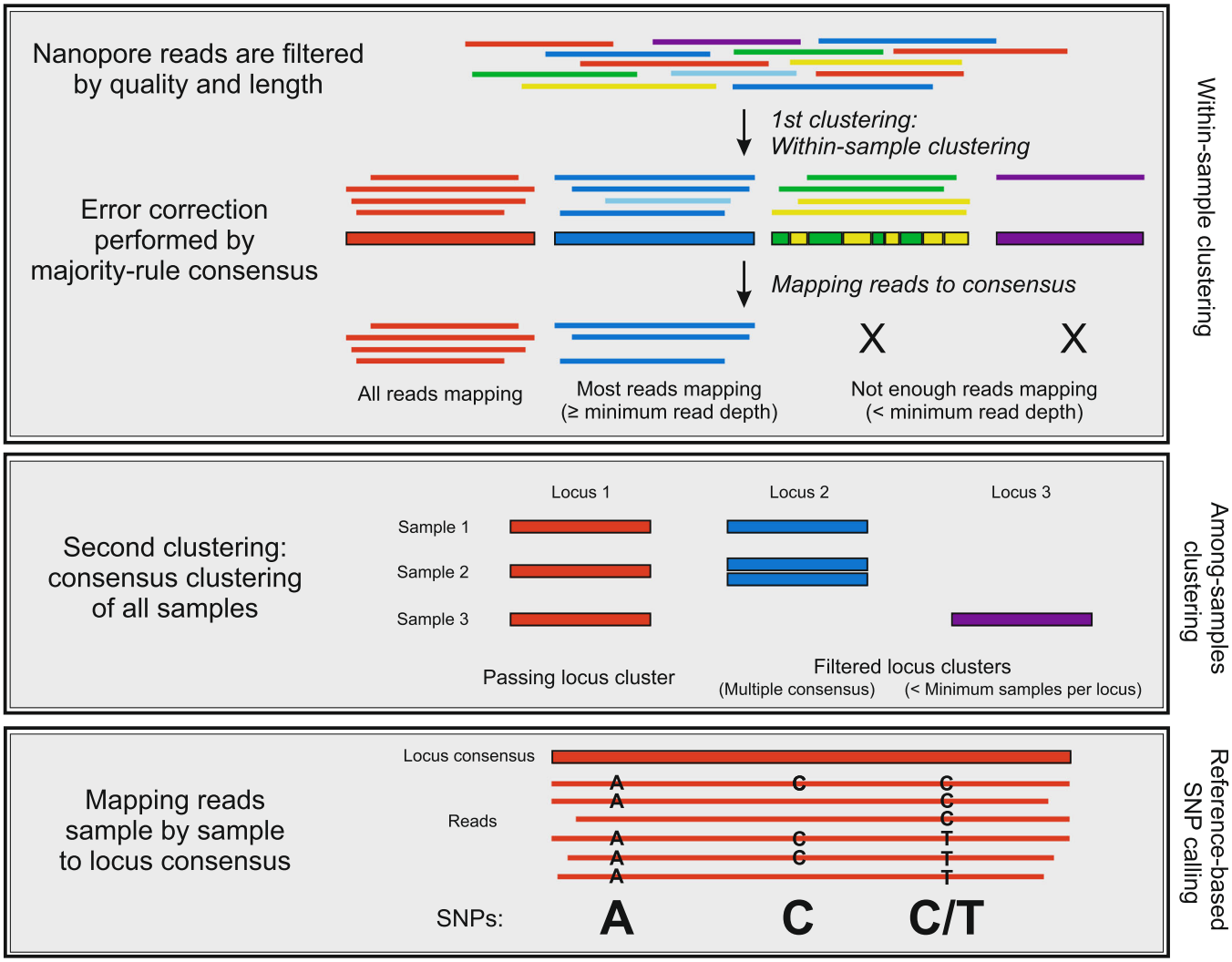


FIGURE 1 The SLANG workflow. During within-sample clustering, quality- and length-filtered reads are clustered according to their locus identity. Reads are then mapped to their cluster-consensus sequence. Unmapped reads are filtered out under the assumption that they do not belong to the locus concerned. Only clusters meeting the mapped-read depth threshold are eligible for the among-samples clustering analysis, where consensus sequences of the passing clusters are clustered to estimate locus orthology across samples. Clusters with only one consensus sequence per sample and enough samples per locus pass the filters. Finally, sequences of the among-samples clusters are mapped to their consensus sequence for reference-based SNP calling.

reference-based SNP calling extracts relevant variants from the orthologs, which can then be used for downstream phylogenetic inference or population genetic analyses.

As a proof-of-concept for SLANG, we prepared two small Nanopore sequencing libraries for analysis. The first library consisted of four samples, one per species, from the genus *Leucanthemum* Mill. (*L. vulgare* Lam., *L. monspeliense* (L.) H. J. Coste, *L. gaudinii* Dalla Torre, and *L. rotundifolium* (Willd.) DC.), and the second library comprised three accessions for each of the three species in the central European *Senecio nemorensis* L. group (*S. hercynicus* Herborg, *S. ovatus* (G. Gaertn., B. Mey. & Scherb.) Willd., and *S. germanicus* Wallr.; Table 1). To generate an anonymous multi-locus data set, we took advantage of the well-known amplified fragment length polymorphism (AFLP)-based genome-reduction approach (Vos et al., 1995) and applied it in the modern setting of NGS. Similar to RADseq/GBS techniques, restriction enzymes were used to produce smaller fragments in the first steps, which are then reduced by selective amplification steps and size selections. While RADseq/GBS accomplishes this through the selective amplification of fragments with both restriction sites, AFLP reduces the locus count by applying primers with additional bases at the 3′ ends in order to only amplify fragments matching the sequence overhang. The individual choice of the length and identity of the selective bases allows an AFLP-based approach to be easily scalable in terms of locus numbers and the sequencing depth necessary for the envisaged project, which makes it a viable alternative to RADseq/GBS. The application of the AFLP technique may also provide an easy entry into NGS, as most plant systematics and ecology labs are familiar with this method. Here, we describe the procedures of preparing AFLP-based sequencing libraries for Nanopore sequencers and how SLANG handles these long-read multi-locus data to extract orthologous loci and SNPs for phylogenetic inference, population genetics, or phylogeographical studies.

METHODS AND RESULTS

Leucanthemum and *Senecio nemorensis* group sequencing library preparations

All samples were silica-dried and their DNA was extracted using the cetyltrimethylammonium bromide (CTAB) method (Doyle and Dickson, 1987; Doyle and Doyle, 1987). Beginning with the AFLP-based genome reduction, the fragment length distributions of the genomic DNA digested with different restriction enzymes were screened. While longer fragments are most desirable, one should keep in mind that they demand a higher flow cell capacity than shorter fragments at the same read coverage per locus, so either fewer samples should be sequenced per flow cell or fewer loci can be covered. We selected *Mse*I as a suitable restriction enzyme for the present demonstration, with fragment lengths around 500 bp. For the combined restriction–ligation reaction, 4 μL genomic DNA (12.5 ng/μL) was mixed with 1 μL T4 Ligase Buffer (Thermo Fisher Scientific, Waltham, Massachusetts, USA), 1 μL NaCl (0.5 M), 0.5 μL BSA (1 mg/mL), 0.5 μL Mse adapter (40 μM), 0.5 μL *Mse*I (10 U/μL; Thermo Fisher Scientific), 0.5 μL T4 Ligase (5 U/μL; Thermo Fisher Scientific), and 2 μL H_2O and placed into a thermocycler for 2 h at 37°C with an additional 15 min at 70°C for enzyme denaturation. For the Mse adapter,

TABLE 1 Sample information. *Senecio* reads were filtered for read lengths between 50 and 1000 bp, while *Leucanthemum* was filtered for reads between 200 and 1000 bp. A total of 310,336,638 bp of *Leucanthemum* sequences and 244,902,300 bp of *Senecio* sequences passed the Q7 quality filter.

Sample	Sample ID	Longitude	Latitude	Raw reads (after qcat)	Raw bases (after qcat)	Reads after filtering	Bases after filtering
L. vulgare Lam.	120-02	43.8925	3.2477222	156,281	65,540,162	139,944	58,525,941
L. monspeliense (L.) H. J. Coste	131-01	44.141167	3.7316389	207,747	87,943,849	181,131	76,063,161
L. gaudinii Dalla Torre	276-01	46.860333	13.817233	197,957	80,059,219	169,044	68,019,760
L. rotundifolium DC.	495-02	45.404022	22.885686	180,232	76,793,408	155,960	65,887,521
S. hercynicus Herborg	01-02	47.699850	10.183917	79,547	27,804,961	62,028	21,381,063
	01-03	47.699850	10.183917	69,116	22,901,337	54,964	17,649,906
	01-04	47.699850	10.183917	69,087	24,695,263	53,836	18,869,793
S. ovatus (G. Gaertn., B. Mey. & Scherb.) Willd.	02-02	49.049767	12.257717	85,092	30,998,678	66,178	20,887,198
	02-01	49.049767	12.257717	75,204	26,453,432	60,319	20,887,198
	02-05	49.049767	12.257717	80,383	29,748,871	63,851	23,136,493
S. germanicus Wallr.	03-03	49.052850	11.973900	74,672	27,267,933	59,789	21,595,843
	03-04	49.052850	11.973900	76,213	27,754,247	59,332	21,004,957
	03-05	49.052850	11.973900	72,262	27,277,578	57,889	21,225,204

the two oligos 5′-GACGATGAGTCCTGAG-3′ and 5′-TACTCAGGACTCAT-3′ were annealed by heating the samples to 95°C for 5 min and reducing the temperature to 22°C by 2°C every 5 s. The number of loci is first reduced in the preselective PCR (94°C at 2 min; 30 cycles of 94°C for 20 s, 56°C for 15 s, and 72°C for 2 min; followed by a final elongation at 72°C for 2 min), which uses a primer with an additional cytosine nucleotide added to the 3′ end (5′-GATGAGTCCTGAGTAAC-3′). To every 2 μL of 1 : 10 diluted restriction–ligation product, 5 μL *Taq* DNA Polymerase Master Mix RED (Ampliqon, Odense, Denmark), 0.5 μL preselective primer (10 μM), and 2.5 μL H_2O were added. Further loci reduction was accomplished using a selective PCR, where two additional bases, 'AA', are added to the 3′ end and Nanopore barcode adapter sequences are tailed to the 5′ end (5′-TTTCTGTTGGTGCTGATATTGCGATGAGTCCTGAG TAACAA-3′ and 5′-ACTTGCCTGTCGCTCTATCTTCGA TGAGTCCTGAGTAACAA-3′) of the primers, as suggested in the 'Ligation sequencing amplicons - PCR barcoding (SQK-LSK109 with EXP-PBC001)' protocol by Oxford Nanopore Technologies, substituting a subsequent ligation of the Nanopore barcode adapter. To ensure specific binding with long and tailed primers, a two-step variation of the preselective PCR was conducted (94°C for 2 min; followed by 30 cycles of 94°C for 20 s and 72°C for 2 min; and a final step at 72°C for 2 min). To every 2 μL of 1 : 10 diluted preselective PCR product, 5 μL *Taq* DNA Polymerase Master Mix RED, 0.25 μL of each 10 μM tailed selective primer, and 2.5 μL H_2O were added.

The selective bases were chosen based on a screening procedure with multiple combinations of selective bases. For both libraries, the selective bases 'CAA' resulted in a consistent fragment length distribution of around 500 bp, which is crucial for a more even sequencing depth of the selected loci. The selective PCR products were cleaned using AmpliClean magnetic beads (NimaGen, Nijmegen, the Netherlands). Nanopore barcodes were added by PCR (95°C for 2 min; followed by 25 cycles of 98°C for 20 s and 72°C for 2 min; with a final extension at 72°C for 1 min), in which 12.5 μL KAPA HiFi HotStart ReadyMix (Roche, Basel, Switzerland), 0.5 μL Nanopore PCR Barcode of the PBC001 PCR Barcoding Expansion 1-12 (Oxford Nanopore Technologies), and 10 μL H_2O were added to 2 μL of cleaned selective PCR product (10 ng/μL).

The samples were then equimolarly multiplexed and loaded onto a 1.5% agarose gel for size selection at ~500 ± 100 bp by gel excision (QIAquick PCR and Gel Cleanup Kit; Qiagen, Hilden, Germany). Finally, the Nanopore sequencing libraries were prepared following the 'Ligation sequencing amplicons (SQK-LSK109)' protocol (Oxford Nanopore Technologies) and sequenced with the MinION Mk1B using a R9.4.1 flow cell.

Read data processing

The FAST5 files were basecalled using the Guppy basecalling software (version 3.2.4; Oxford Nanopore Technologies) and demultiplexed with qcat version 1.1.0 (Oxford Nanopore Technologies), with the '*--detect-middle*', '*--trim*', '*--min-read-length 1*', and '*--guppy*' parameters. NanoFilt version 2.7.1 (De Coster et al., 2018) was then used to filter for reads with Phred quality scores of 7 or higher ('*--quality 7*'). Primer sequences were removed with BBDuk from the BBTools software suite version 38.87 (Bushnell, 2014). First, the 5′ ends were trimmed by only allowing matches of the primer sequence within the first 100 bases ('*ktrim = l*', '*restrictleft = 100*', '*literal = GATGAGTCCTGAGTAACAA*', '*rcomp = f*'). Considering the Nanopore read quality, a *k*-mer of 10 ('*k = 10*') with up to one error ('*edist = 1*') was considered sufficient for it to be accepted as a match. In order for BBDuk to properly work with the provided Nanopore reads, the '*ignorebadquality*' and '*qin = 33*' parameters were necessary. To trim the 3′ ends, the respective parameters were adjusted ('*ktrim = r*', *restrictright = 100*', '*literal = TTGTTACTCAG-GACTCATC*'). Finally, BBDuk was used to filter for read lengths between 200 and 1000 bp in the *Leucanthemum* data set and between 50 and 1000 bp in the *Senecio* data set.

De novo locus assembly with SLANG (within-sample clustering)

The quality- and length-filtered reads were then subjected to the SLANG pipeline, starting with the within-sample clustering, in which all sequenced loci with adequate read depth are assembled for each sample. Similar established software adapted to Illumina data, such as *ipyrad* (Eaton and Overcast, 2020), address de novo locus assembly with clustering algorithms, as provided by VSEARCH version 2.15.0 (Rognes et al., 2016). SLANG also adapts VSEARCH for clustering reads on the basis of a read similarity–threshold parameter, assuming that reads group according to their locus identity. Even at an optimized similarity threshold, where most reads are correctly assigned to their clusters, some clusters do contain reads unassociated with the locus that the cluster represents, which could negatively influence the SNP-calling quality and must therefore be removed. Overcoming this problem is an even more apparent challenge to address for error-prone Nanopore reads. For this reason, a majority-rule consensus sequence based on all reads of a cluster is computed and used as a mapping reference in the Nanopore read–specialized Minimap2 version 2.17 (Li, 2018), to which all reads of the cluster concerned are mapped. Individual reads that strongly deviate from the majority of the clustered reads due to an incorrect assignment or because they are too flawed by sequencing errors are prevented from mapping to the reference and will thus be filtered. Finally, a cluster read depth filter removes all clusters that are not covered by sufficient mapped reads for high-quality SNP calling. To preserve as much read data as possible while still being able to make sophisticated SNP calls (see "Reference-based SNP calling" below), we chose a minimum read depth of 10 for the *Leucanthemum* and *Senecio* data sets.

The correct inference of phylogenies using methods based on read-similarity clustering relies heavily on the choice of the similarity threshold values, as they determine

the orthology of reads during both within-sample clustering and among assembled loci (among-samples clustering, see below). Setting the similarity threshold too low results in reads of different loci clustering together (locus undersplitting), while setting it too high will split reads of the same locus into different clusters (locus oversplitting). Many valid and sophisticated methods have been proposed for the choice of optimized similarity thresholds (Ilut et al., 2014; Harvey et al., 2015; Mastretta-Yanes et al., 2015; McCartney-Melstad et al., 2019), each with a different focus on weighting the importance of certain metrics. We conceptualized and applied another methodology that chooses an optimized within-sample clustering threshold by multiplying two metrics: the total number of clusters and the number of clusters containing unmapped reads, the former being high in cases with prevalent oversplitting and the latter indicating locus undersplitting by larger values. The product of the two metrics was calculated for similarity thresholds in incremental steps of 0.05. By doing so, a similarity threshold value of 0.75 (75%) was determined for both the *Leucanthemum* and the *Senecio nemorensis* group data sets (Appendix 1), which can be interpreted as quite stringent when considering the 10% Nanopore error rate (Magi et al., 2017; Fuselli et al., 2018), basecalling errors (Wick et al., 2018), and PCR artifacts. Additional free space for allowing allelic variation within clusters must be considered, otherwise locus oversplitting will be an issue.

In general, the formula presented favors similarity thresholds where high numbers of clusters are formed, but prefers to have as few clusters with unmapped reads as possible. The similarity threshold resulting in the highest value represents a tradeoff between high cluster count and potential paralogous read groupings within these clusters. As a result, the resulting optimized similarity threshold keeps over- and undersplitting at low levels. Nonetheless, despite intensive efforts to determine the optimal similarity threshold, high topological accuracy can be expected across a wide range of similarity threshold values, as long as non-extreme values are chosen (Rubin et al., 2012). In order to give users the opportunity to implement their own approaches on choosing similarity thresholds, no automatic choice is made by the SLANG pipeline, and optimization as described above is suggested.

Identification of orthologous loci (among-samples clustering)

The described procedure for within-sample clustering identifies potential loci for each individual sample in the form of majority-rule consensus sequences. Subsequently, the among-samples clustering aggregates all consensus sequences of every sample with the goal of forming groups of orthologous sequences, which are referred to as locus clusters. Again, VSEARCH is used to group similar sequences based on a similarity threshold, a practice of orthology inference that has been shown to be effective in *ipyrad* for the analysis of Illumina data. In order to only filter for clusters containing orthologous sequences, a specific adaption of SLANG is to filter all locus clusters with more than one consensus sequence of a single sample. Multiple consensus sequences of a single sample may indicate potential paralogous groupings or oversplit loci (in the within-sample clustering step) coming together; because both will introduce erroneous SNP calls, these locus clusters are therefore omitted. Singleton clusters consisting of unique loci only found in a single sample are also filtered, as well as clusters not passing the minimum samples-per-locus parameter, which by default must be set to at least two samples. Similar to the within-sample clustering, multiple similarity threshold values must be explored to allow an optimized choice to be made. Overly stringent similarity thresholds will only result in singleton locus clusters, which would be excluded; on the other hand, if they are too lax, more same-sample clusters are generated, which would also be filtered out. With regard to the filtering procedure, the optimal similarity threshold value results in the highest number of passing locus clusters, where same-sample consensus sequence groupings and singleton clusters are minimized. For both the *Leucanthemum* and *Senecio* data sets, we left the minimum samples-per-locus parameter at the default (at least two samples), and after testing the similarity thresholds in incremental steps of 0.05, values of 0.91 and 0.90 were found to provide the highest numbers of passing locus clusters (704 and 448, respectively).

Reference-based SNP calling

The orthologs are established through locus clusters, which means the SNPs are then called using a reference-based approach. For the locus clusters, majority-rule consensus sequences are used as mapping references, referred to as "locus consensus sequences" below. Subsequently, the filtered reads from the within-sample clustering are mapped, sample by sample, to their respective locus consensus sequence using Minimap2. With BCFtools mpileup (Danecek et al., 2021), all possible variants are collected, except for indels ('*--skip-indels*'), which were excluded due to the high probability of being frequently occurring Nanopore sequencing artifacts. We disabled the reconsideration of the per-base alignment quality (BAQ; '*--no-BAQ*') due to many variants being excluded when the parameter was active. Moreover, variants are gathered without reconsideration of their base quality score ('*--min-BQ 0*'). Multiallelic sites are split using BCFtools norm '*-m-*', facilitating a downstream assessment of each individual allelic variation. SNPs are then filtered with BCFtools view. Assuming diploid individuals and a minimum read depth of 10, we retained SNPs with a frequency over 0.25 that appeared at least twice at the position concerned, with a minimum total read depth of 5. It was previously shown that, despite reported Nanopore read error rates around 10% (Magi et al., 2017; Fuselli

et al., 2018), read depths between 2 and 5 were sufficient for a 75% SNP-calling accuracy at heterozygous sites, and even up to 98% at homozygous sites (Malmberg et al., 2019). Finally, multiallelic sites are restored with BCFtools norm '*-m+*', and all variant data are combined into a single Variant Call Format (VCF) file, which can be used as an input for most downstream phylogenetic applications or transformed into other common input formats.

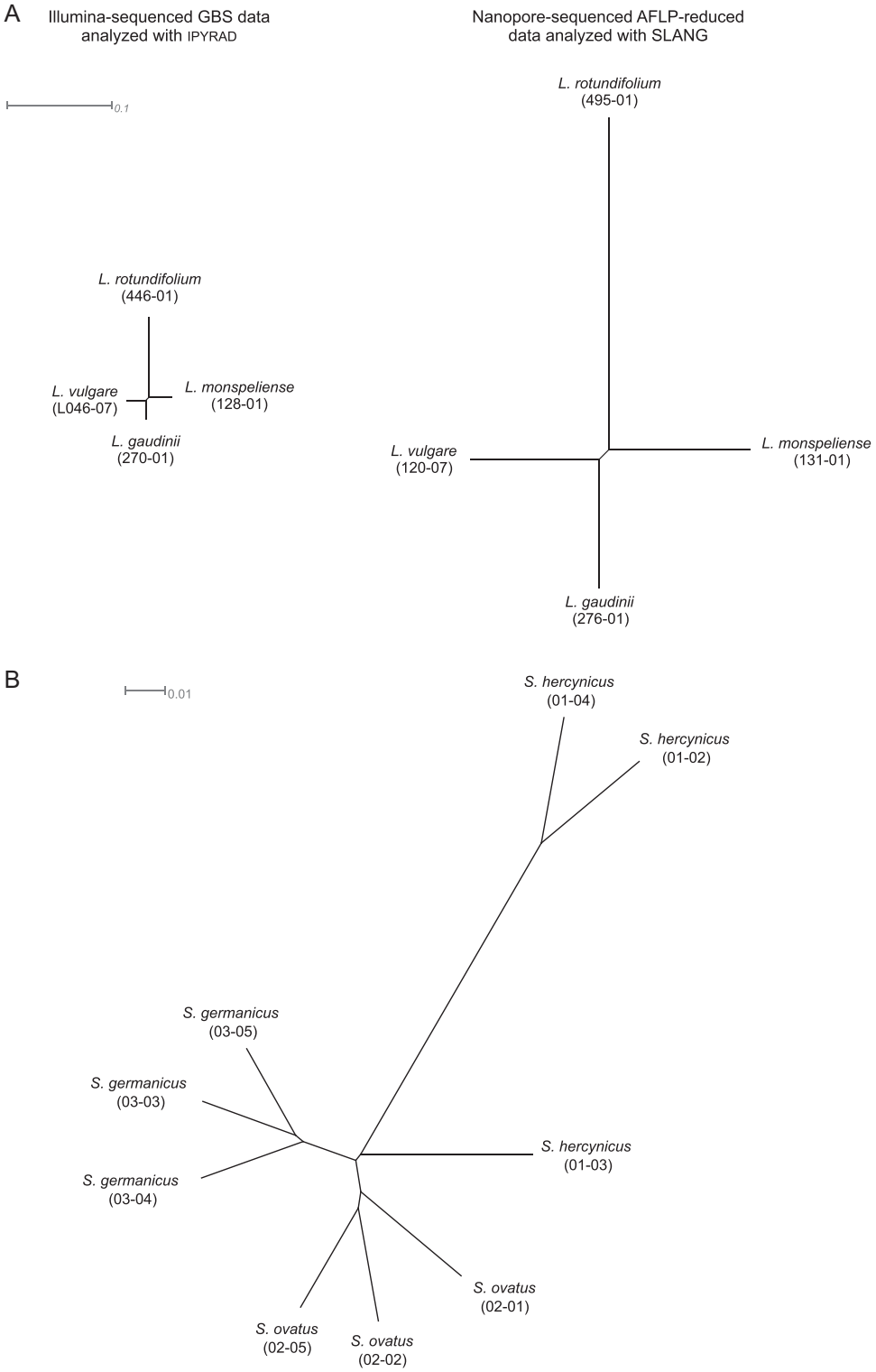


FIGURE 2 Phylogenetic network reconstructions of the *Leucanthemum* and *Senecio* data sets. (A) Phylogenetic network reconstruction of the *Leucanthemum* data set based on GBS high-quality short Illumina reads assembled using *ipyrad* (left) and AFLP-based Nanopore reads assembled using SLANG (right). (B) Phylogenetic network reconstruction of the *Senecio nemorensis* group data set produced using SLANG. Nei–Li distances were calculated based on the base frequencies in the VCF file and used as inputs in SplitsTree version 4.16.1 (Huson and Bryant, 2006).

SLANG in action—part I: *Leucanthemum* data set

In total, 742,217 reads and ~310.34 Mbp were sequenced for the four *Leucanthemum* representatives (Table 1). After read preprocessing, 505,715 reads (268.5 Mbp) passed the Q7 quality filter and had a length between 200 and 1000 bp. The read length distribution after preprocessing was similar among all *Leucanthemum* samples, indicating the representation of comparable loci across samples (Appendix 2), although in the range of around 430 bp, *L. rotundifolium* and *L. monspeliense* reads were overrepresented compared to the numbers observed in the other accessions. The preprocessed sequence data were then subjected to the SLANG pipeline after the clustering threshold optimization, performed as described above. This resulted in the assembly of 704 orthologous loci and the extraction of 12,368 SNPs. To evaluate the quality of the methodology's locus assembly, orthology estimation, and SNP-calling process, we calculated the base frequency–sensitive Nei–Li distances (https://github.com/TankredO/nei_vcf) and constructed a phylogenetic network using the unweighted pair group method with arithmetic mean (UPGMA) method in SplitsTree version 4.16.1 (Huson and Bryant, 2006). The resulting network grouped *L. rotundifolium* with *L. monspeliense*, and *L. vulgare* with *L. gaudinii* (Figure 2A). Exactly the same relationships with comparable relative genetic distances and topology were revealed based on a preceding Illumina short-read GBS data set (T. Ott [University of Regensburg], M. Schall [University of Regensburg], R. Vogt [Free University of Berlin], and C. Oberprieler [University of Regensburg], unpublished manuscript) analyzed with *ipyrad* version 0.9.54 (Eaton and Overcast, 2020). Different accessions were included in the Illumina GBS data set and in the Nanopore SLANG analysis; however, the sampled *Leucanthemum* species are so distinctly separated from each other that it is highly unlikely for the accessions to break their species boundaries. The only notable difference between the two networks are the longer branch lengths of the Nanopore SLANG analysis, which indicates a higher background noise caused by erroneous SNP calls. This is somewhat expected from error-prone Nanopore reads even after extensive filtering. An additional source of erroneous SNPs could be introduced by PCR artifacts caused by the use of a standard *Taq* polymerase instead of a high-fidelity polymerase during the numerous cycles of preselective and selective PCRs. Nonetheless, this background noise is evenly distributed across all samples and does not impair the correct phylogenetic inference. SLANG therefore successfully identified loci, estimated orthology, and extracted predominantly meaningful SNP information for a correct reconstruction of phylogenetic relationships among the four *Leucanthemum* samples.

SLANG in action—part II: *Senecio nemorensis* group analysis

For the nine *Senecio* samples, 681,576 reads and ~244.9 Mbp were sequenced, from which 538,186 reads and ~186.64 Mbp remained after Q7 quality and 50–1000 bp length filtering. With the SLANG pipeline, 448 orthologous loci were assembled, containing 10,048 SNPs. As with the *Leucanthemum* data, the UPGMA network based on Nei–Li distances was reconstructed (Figure 2B). All accessions of the three species grouped together according to their species affiliation, with *S. ovatus* showing greater genetic similarity with *S. germanicus* than with *S. hercynicus*. A comparison of read length distributions revealed major deviations caused by one of the *S. hercynicus* samples (accession 01-03), originating from divergent locus sampling during library preparation. This subsequently led to a conspicuously low number of orthologs found as observed in the VCF output, which explains its outlying position in the network. In fact, several attempts were needed to successfully amplify this sample and even then, in hindsight, the fragment length distribution on the agarose gel slightly differed from those of the other samples. Difficulties during PCRs could indicate lower-quality DNA, stressing the importance of high-quality DNA for NGS library preparation.

CONCLUSIONS

The SLANG pipeline proved to be a simple and effective tool for fast genotyping with error-prone long-read Nanopore data. It successfully de novo assembled loci and estimated orthology through clustering, as well as extracting meaningful SNP data for phylogenetic reconstructions. By preparing, sequencing, and analyzing two independent data sets, two phylogenetic networks were generated from Nei–Li distances calculated from the extracted SNPs, which in the case of the *Leucanthemum* data set could be compared to a topologically identical network based on high-quality short-read Illumina data. The phylogenetic network for the *Senecio nemorensis* group data set comprised distinct groups corresponding to species affiliation. These results were based on the well-known AFLP method, which proved to still be a viable option to reduce genome complexity in the age of NGS for its ability to produce a reproducible multi-locus data set. SLANG could therefore have the potential to expand the growing interest in long-read genotyping methods alongside the more common short-read procedures. SLANG is freely available at Github: https://github.com/DorfnerM/SLANG.

AUTHOR CONTRIBUTIONS

The present study was conceptualized by M.D. and C.O. C.O. provided the plant samples and acquired the funding for this project. M.D. planned and conducted all laboratory experiments, as well as the SLANG analysis for the *Leucanthemum* data set. P.O. conducted laboratory work and the SLANG analysis for the *Senecio* data set under the supervision of M.D. M.D. designed and developed the SLANG Python script with additional methodological input by C.O. T.O. developed the Nei–Li distance script and provided advice on all programming. M.D. wrote the initial draft of the manuscript; T.O. and C.O. contributed to the final version of the manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

This work was funded by a grant from the Deutsche Forschungsgemeinschaft (DFG) in the SPP 1991 project "Taxon-omics—New Approaches for Discovering and Naming Biodiversity" to C.O. (OB 155/13-1). Two anonymous reviewers and the editors contributed considerably to an improved version of our present contribution. Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

SLANG is open source and freely available as a Python script at https://github.com/DorfnerM/SLANG. The Nei–Li distance script is open source and freely available at https://github.com/TankredO/nei_vcf.

ORCID

Marco Dorfner https://orcid.org/0000-0002-5667-2063
Tankred Ott https://orcid.org/0000-0001-6748-0510
Christoph Oberprieler https://orcid.org/0000-0002-7134-501X

REFERENCES

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3: e3376.

Bushnell, B. 2014. BBTools software package. Website: http://sourceforge.net/projects/bbmap [accessed 8 November 2021].

Cariou, M., L. Duret, and S. Charlat. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. *Ecology and Evolution* 3: 846–852.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10: giab008.

Davey, J., and M. L. Blaxter. 2011. RADSeq: Next-generation population genetics. *Briefings in Functional Genomics* 10: 108.

De Coster, W., S. D'Hert, D. T. Schultz, M. Cruts, and C. Van Broeckhoven. 2018. NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics* 34: 2666–2669.

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

Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.

Eaton, D. A., and I. Overcast. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics* 36: 2592–2594.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6: e19379.

Fuselli, S., R. P. Baptista, A. Panziera, A. Magi, S. Guglielmi, R. Tonin, A. Benazzo, et al. 2018. A new hybrid approach for MHC genotyping: High-throughput NGS and long read MinION nanopore sequencing, with application to the non-model vertebrate Alpine chamois (*Rupicapra rupicapra*). *Heredity* 121: 293–303.

Harvey, M. G., C. D. Judy, G. F. Seeholzer, J. M. Maley, G. R. Graves, and R. T. Brumfield. 2015. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ* 3: e895.

Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.

Ilut, D. C., M. L. Nydam, and M. P. Hare. 2014. Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering. *BioMed Research International* 2014: 675158.

Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094–3100.

Magi, A., B. Giusti, and L. Tattini. 2017. Characterization of MinION nanopore data for resequencing analyses. *Briefings in Bioinformatics* 18: 940–953.

Malmberg, M. M., G. C. Spangenberg, H. D. Daetwyler, and N. O. I. Cogan. 2019. Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (*Brassica napus* L.). *Scientific Reports* 9: 8688.

Mastretta-Yanes, A., N. Arrigo, N. Alvarez, T. H. Jorgensen, D. Piñero, and B. C. Emerson. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources* 15: 28–41.

McCartney-Melstad, E., M. Gidiş, and H. B. Shaffer. 2019. An empirical pipeline for choosing the optimal clustering threshold in RADseq studies. *Molecular Ecology Resources* 19: 1195–1204.

McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* 66: 526–538.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* 7: e37135.

Razkin, O., G. Sonet, K. Breugelmans, M. J. Madeira, B. J. Gómez-Moliner, and T. Backeljau. 2016. Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex *Pyramidula*: The power of RADseq data. *Molecular Phylogenetics and Evolution* 101: 267–278.

Rochette, N. C., and J. M. Catchen. 2017. Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols* 12: 2640–2659.

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* 4: e2584.

Rubin, B. E., R. H., Ree, and C. S. Moreau. 2012. Inferring phylogenies from RAD sequence data. *PLoS ONE* 7: e33394.

Stoler, N., and A. Nekrutenko. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* 3: Iqab019.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. V. D. Lee, M. Hornes, A. Frijters, et al. 1995. AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research* 23: 4407–4414.

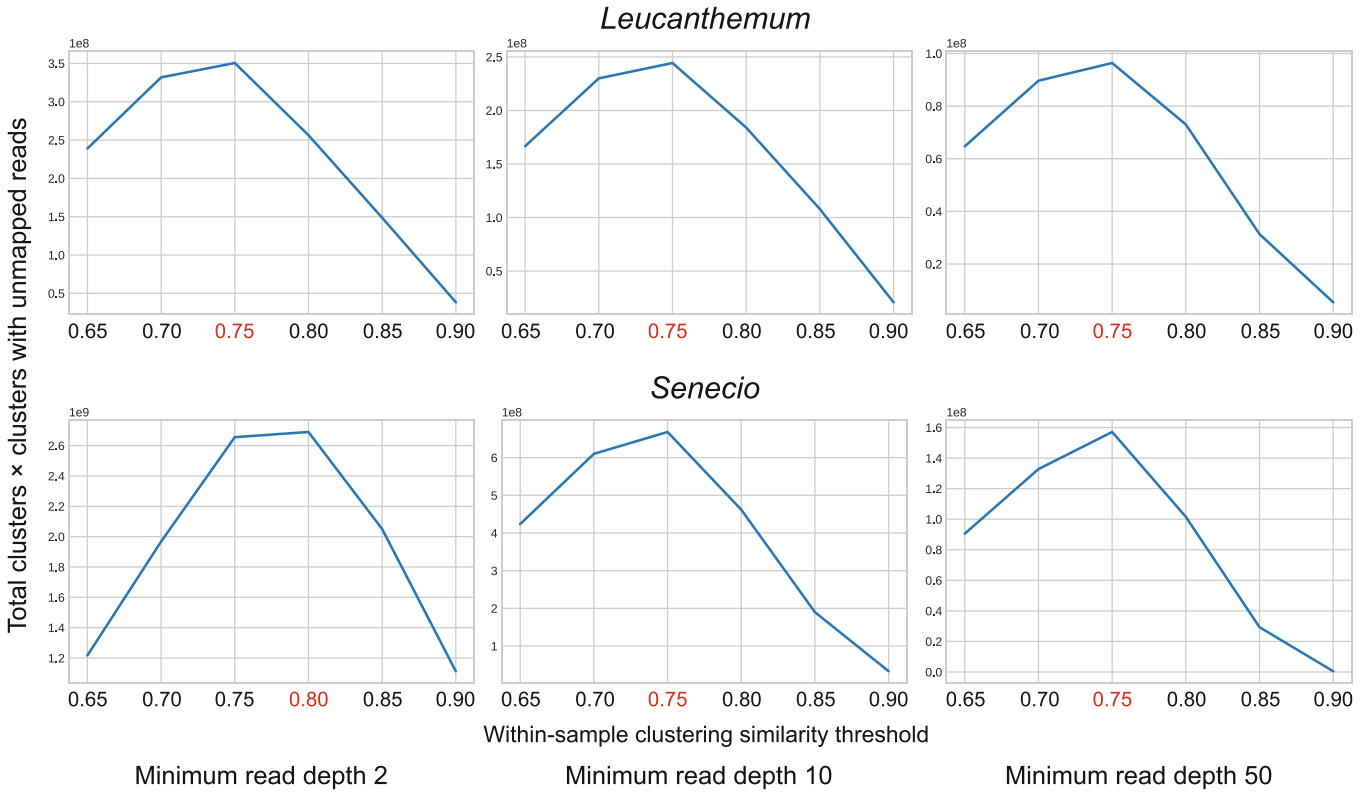
Wagner, F., T. Ott, M. Schall, U. Lautenschlager, R. Vogt, and C. Oberprieler. 2020a. Taming the red bastards: Hybridisation and species delimitation in the *Rhodanthemum arundanum*-group (Compositae, Anthemideae). *Molecular Phylogenetics and Evolution* 144: 106702.

Wagner, N. D., L. He, and E. Hörandl. 2020b. Phylogenomic relationships and evolution of polyploid *Salix* species revealed by RAD sequencing data. *Frontiers in Plant Science* 11: 1077.

Wick, R. R., L. M. Judd, and K. E. Holt. 2018. Deepbinner: Demultiplexing barcoded Oxford Nanopore reads with deep convolutional neural networks. *PLoS Computational Biology* 14: e1006583.

How to cite this article: Dorfner, M., T. Ott, P. Ott, and C. Oberprieler. 2022. Long-read genotyping with SLANG (Simple Long-read loci Assembly of Nanopore data for Genotyping). *Applications in Plant Sciences* 10(3): e11484. https://doi.org/10.1002/aps3.11484

Appendix 1. Similarity threshold optimization for the within-sample clustering. To select an optimized similarity threshold, the sum of the total number of clusters is multiplied by the sum of all clusters containing unmapped reads (potentially paralogous loci). The product of the two metrics was calculated for similarity thresholds in incremental steps of 0.05. This was reproduced at minimum read depths of 2, 10, and 50. For both data sets, 0.75 is the optimal similarity threshold, except when a minimum read depth of 2 is requested for *Senecio*.



Appendix 2. Read length distribution of the *Leucanthemum* and *Senecio nemorensis* group read data sets after read preprocessing. Read length distributions are similar among the samples; only *L. rotundifolium* and *L. monspeliense* were overrepresented around the 430 bp length. *Senecio hercynicus* (01-03) deviated from other *Senecio* samples.