The first draft genome of feather grasses using SMRT sequencing and its implications in molecular studies of Stipa

Evgenii Baiakhmetov, Cervin Guyomar, Ekaterina Shelest, Marcin Nobis, Polina D Gudkova

scientific reports

Check for updates

OPEN

The first draft genome of feather grasses using SMRT sequencing and its implications in molecular studies of *Stipa*

Evgenii Baiakhmetov[1,2]✉, Cervin Guyomar[3,4], Ekaterina Shelest[3,5], Marcin Nobis[1,2]✉ & Polina D. Gudkova[2,6]

The Eurasian plant *Stipa capillata* is the most widespread species within feather grasses. Many taxa of the genus are dominants in steppe plant communities and can be used for their classification and in studies related to climate change. Moreover, some species are of economic importance mainly as fodder plants and can be used for soil remediation processes. Although large-scale molecular data has begun to appear, there is still no complete or draft genome for any *Stipa* species. Thus, here we present a single-molecule long-read sequencing dataset generated using the Pacific Biosciences Sequel System. A draft genome of about 1004 Mb was obtained with a contig N50 length of 351 kb. Importantly, here we report 81,224 annotated protein-coding genes, present 77,614 perfect and 58 unique imperfect SSRs, reveal the putative allopolyploid nature of *S. capillata*, investigate the evolutionary history of the genus, demonstrate structural heteroplasmy of the chloroplast genome and announce for the first time the mitochondrial genome in *Stipa*. The assembled nuclear, mitochondrial and chloroplast genomes provide a significant source of genetic data for further works on phylogeny, hybridisation and population studies within *Stipa* and the grass family Poaceae.

In the year 2000, the *Arabidopsis thaliana* L. genome became the first plant genome to be completely sequenced and assembled[1]. Since then, many genomes from the plant kingdom have been sequenced, e.g. green algae[2,3], bryophytes[4,5], ferns[6], gymnosperms[7,8] and angiosperms[9,10]. In the grass family (Poaceae) the reference assemblies were primarily obtained for crops[11–13] and model plants[14–16]. The advent of second-generation sequencing and the subsequent decreasing of the overall sequencing costs have enabled the determination of whole genome sequences in many non-model plant species[17–20].

Recently, the 1KP project that was aiming to sequence 1,000 green plant transcriptomes[21–23] has been followed by the 10KP project[24]. The later initiative intends to sequence complete genomes from more than 10,000 plants and protists. The project is supposed to be completed in 2023 and it presumes to provide family-level high-quality reference genomes, ideally with chromosome-scale assemblies. Nevertheless, the data at the level of genera may not be processed immediately[24]. In comparison to other kingdoms, plants have very large genomes[13,25,26], high ploidy level[27] and the abundance of repetitive sequences[28–30]. Currently, to face these issues, the third-generation sequencing has been applied. The so-called single-molecule real-time (SMRT) sequencing provided by Pacific Biosciences (PacBio)[31] and nanopore sequencing by Oxford Nanopore Technologies[32] afford a range of benefits, including exceptionally long-read lengths (20 kb or more), resolving extremely repetitive and GC-rich regions and direct variant phasing[32,33].

In the fossil record *Stipa* L., or a close relative genus, is known from about 34 Mya of the upper Eocene[34,35]. For many decades, *Stipa* has been described as a genus with over 300 species common in steppe zones of Eurasia, North Africa, Australia and the Americas[36,37]. According to the recent studies based on both morphological and molecular data, the genus has been reduced and currently includes over 150 species geographically confined to

[1]Institute of Botany, Faculty of Biology, Jagiellonian University, Gronostajowa 3, 30-387, Kraków, Poland. [2]Research Laboratory 'Herbarium', National Research Tomsk State University, Lenin 36 Ave., Tomsk 634050, Russia. [3]German Centre for Integrative Biodiversity Research (iDiv), Puschstrasse 4, 04103 Leipzig, Germany. [4]Institute for Genetics, Environment and Plant Protection (IGEPP), Agrocampus Ouest, INRAE, University of Rennes 1, 35650 Le Rheu, France. [5]Centre for Enzyme Innovation, University of Portsmouth, Portsmouth PO1 2UP, UK. [6]Department of Biology, Altai State University, Lenin 61 Ave., Barnaul, Russia 656049. ✉email: evgenii.baiakhmetov@doctoral.uj.edu.pl; m.nobis@uj.edu.pl

Figure 1. A representative individual of *Stipa capillata*.

Europe, Asia and North Africa[38–42]. Most species of *Stipa* are dominants and/or subdominants in steppe plant communities[43–45] and can be used for their classification[46]. Moreover, some species are of economic importance mainly as pasture and fodder plants, especially in the early phases of vegetation[36,47], they can be used for soil remediation processes[48,49], in studies related to climate change[50–52] and as ornamental plants (e.g. *S. capillata* L., *S. pulcherrima* K. Koch, *S. pennata* L.).

In recent years, large-scale molecular data began to appear for *Stipa*: *de novo* transcriptome assemblies of *S. purpurea* Griseb.[50,53], *S. grandis* P. A. Smirn.[54] and *S. lagascae* Roem. & Schult.[52], whole chloroplast genomes for 19 taxa[57] and raw genomic data available via the NCBI Sequence Read Archive (SRA) for *S. capillata*[58] and *S. breviflora* Griseb.[59]. In addition, nucleolar organising regions (NORs) were sequenced for six *Stipa* taxa[60]. Nevertheless, no complete or draft genome assembly currently exists for any *Stipa* species. In order to fill this gap, here we aim to: (1) present for the first time a single-molecule long-read dataset (nuclear, mitochondrial and chloroplast genomes) generated using the SMRT sequencing on the PacBio Sequel platform; (2) demonstrate and discuss the potential usage of this data in further studies of *Stipa*.

For the goals of the study we chose to sequence the entire genome of *S. capillata* (Fig. 1) as it is the most widespread taxon within the genus, growing on sandy to loamy, nutrient poor soils in the dry grasslands of Eurasia[61]. Currently, this species is increasingly attracting the interest of conservation biologists due to its large distribution range, common occurrence in the Eurasian steppes and pseudosteppes, a limited number of refugia in Europe and both great morphological and genetic variability within its range[62–64].

Results

Assembled nuclear genome. The SMRT sequencing yielded in 23.16-fold genome coverage consisting of 25.84 Gb sequence data with an N50 read length of 17,096 bp (Supplementary Table S1). *De novo* assembling of PacBio reads using Flye v.2.4[65,66] resulted in a genome size of 1,004 Mb[67] with a contig N50 of 351 kb and a GC level of 45.97%. On the other hand, another *de novo* assembly performed with FALCON v.0.2.5[68] demonstrated a smaller genome size of 773 Mb with a GC level of 46.04%. However, the Flye assembly has a better N50 of 350,543 that is almost three times bigger than for FALCON. In case of applying Purge Haplotigs v1.1.1[69], the final genome size was reduced by 177 Mb with an N50 of 381,155 (Table 1) and a GC level of 45.82%.

Metrics	Flye assembly	FALCON assembly	Flye assembly after Purge Haplotigs
Length of assembly, bases	1,003,531,354	773,212,558	826,891,869
Number of sequences	5,931	885	3,683
Largest length of a sequence, bases	2,321,367	590,564	2,321,367
Average length of sequences, bases	169,201	88,015	224,516
N50, bases	350,543	119,836	381,155
Number of sequences with N50	837	2,061	640
N100, bases	1,001	20,078	1,014

Table 1. Statistics of the nuclear genome assemblies.

Metrics	Flye assembly	FALCON assembly	Flye assembly after Purge Haplotigs
Complete BUSCOs	4,557 (93.10%)	2,765 (56.50%)	4,304 (87.90%)
Complete and single-copy BUSCOs	2,383 (48.70%)	2,408 (49.20%)	2,916 (59.60%)
Complete and duplicated BUSCOs	2,174 (44.40%)	357 (7.30%)	1,388 (28.30%)
Fragmented BUSCOs	46 (0.90%)	186 (3.80%)	80 (1.60%)
Missing BUSCOs	293 (6%)	1,945 (39.70%)	512 (10.50%)
Total BUSCO groups searched	4,896 (100%)	4,896 (100%)	4,896 (100%)

Table 2. BUSCO statistics.

Species	Number of chromosomes (n)	Number and the total length of contigs assigned to the reference	Number and the total length of non-assigned contigs
B. distachyon[70]	5	4,061 (950.13 Mb)	1,871 (53.40 Mb)
		94.68%	5.32%
H. vulgare[71]	7	4,036 (945.36 Mb)	1,896 (58.17 Mb)
		94.20%	5.80%
A. tauschii[72]	7	4,161 (954.95 Mb)	1,771 (48.59 Mb)
		95.16%	4.84%
O. sativa[73]	12	3,477 (902.39 Mb)	2,455 (101.15 Mb)
		89.92%	10.08%
T. aestivum[74]	21	2,434 (418.14 Mb)	3,498 (585.40 Mb)
		41.67%	58.33%

Table 3. RaGOO statistics.

The subsequent analysis based on a benchmark of 4,896 conserved genes belonging to the Poales order (dataset poales_odb10) revealed that the Flye assembly has 4,557 (93.10%) completed BUSCO (Benchmarking Universal Single-Copy) genes and only 293 (6%) missing BUSCOs versus 2,765 (56.50%) and 1,945 (39.70%) for the FALCON assembly. The Flye assembly after Purge Haplotigs shows 4,304 (87.90%) completed BUSCOs and 512 (10.50%) missing BUSCOs (Table 2).

Scaffolding of contigs. Nearly all contigs of *S. capillata* genome can be assigned to the reference chromosomes of *Brachypodium distachyon* L., *Hordeum vulgare* L. and *Aegilops tauschii* Coss., whereas genomes of *Oryza sativa* L. and especially *Triticum aestivum* L., have much less homology to the feathergrass assembly. In particular, 95.16% contigs of *S. capillata* genome were assigned to seven chromosomes of *A. tauschii* genome, 94.68% to five chromosomes of *B. distachyon*, 94.20% to seven chromosomes of *H. vulgare*, 89.92% to 12 chromosomes of *O. sativa* and only 41.67% to 21 chromosomes of *T. aestivum*. The total length of non-assigned contigs was reasonably low for *A. tauschii* (48.59 Mb), *B. distachyon* (53.40 Mb) and *H. vulgare* (58.17 Mb), whereas for *O. sativa* and *T. aestivum* it was about 101.15 Mb and 585.40 Mb, respectively (Table 3). In addition, the RaGOO grouping confidence and orientation confidence scores per chromosome ranged from 57.81 to 76.11% and from 80.03 to 95.11%, respectively, indicating that the contigs could be placed on a chromosome with an acceptable level of confidence (Supplementary Table S2). The only exception is *T. aestivum* for which scores ranged from 30.49 to 47.76% for the grouping confidence score and from 57.81 to 70.19% for the orientation confidence score. Nevertheless, based on the location confidence score, the exact position of the contigs on a chromosome could not be accurately estimated, reflecting a low level of synteny to the reference genomes. In

Type of repeats	Number of elements	Total (bp)	% of genome
Class I: Retrotransposon:	123,524	161,756,598	16.12
SINEs	6,211	2,422,254	0.24
LINEs	26,453	19,189,619	1.91
LTR elements	90,860	140,144,725	13.97
Class II: DNA-transposon:	99,245	72,448,468	7.22
Hobo-Activator	6,824	3,826,368	0.38
Tc1-IS630-Pogo	619	500,988	0.05
PiggyBac	1	75	0.00
Tourist/Harbinger	11,326	3,980,231	0.40
Other	2	113	0.00
Unclassified	758,908	344,622,074	34.34
Total repeats	981,677	578,827,140	57.68
Rolling-circles	3,306	2,797,158	0.28
Low complexity	18,762	1,145,428	0.11
Simple repeats	114,826	5,716,291	0.57

Table 4. Statistics of repetitive elements.

particular, the score was in a range of 31.30–43.66% for *O. sativa*, 26.06–39.13% for *B. distachyon*, 19.56–31.41% for *H. vulgare*, 17.47–24.15% for *A. tauschii* and 10.30–38.23% for *T. aestivum.*

Transposable elements and nuclear genome annotation. Identification of transposable elements (TEs) revealed that more than half of the *S. capillata* genome (57.68%) is occupied by repetitive sequences. Particularly, retrotransposons represent at least 16.12% and transposons are reaching no less than 7.22% of the genome. Nonetheless, 34.34% of TEs are currently unclassified. Among classified repeats, long terminal repeats (LTRs) were the most abundant elements within retrotransposons, whereas Tourist/Harbinger elements were more common amid DNA-transposons. In total, 114,826 sequences were identified as simple repeats and occupy 0.57% of the genome. In addition, rolling-circles (0.28% of the genome) and low complexity sequences (0.11% of the genome) were found (Table 4).

The subsequent structural annotation of the masked genome revealed 53,535 nuclear genes (Supplementary File 1). On the other hand, the unmasked genome has 154,755 structurally annotated genes and 94,237 of them have BLAST hits in the NCBI non-redundant database. Nonetheless, among the 94,237 genes of the unmasked genome, 12,094 sequences are related to transposable elements. In particular, 2,925 genes associated with transposons, and 9,859 assigned to retrotransposons. In addition, 229 genes encode transposase-related proteins. Thus, except transposable elements the unmasked genome has 81,224 genes that can be associated with already known proteins (Supplementary File 2).

SSR markers. In total, 77,614 perfect repeat motifs were identified for the nuclear genome assembly using Krait[75] (Supplementary File 3). Within those, di- and tri-nucleotides were the most common types, accounting 28,365 (36.55%) and 25,794 (33.23%) repeats, respectively. Tetra-nucleotide motifs were the third most abundant repeats with 9,777 SSRs (12.60%), followed by mono-nucleotides with 6,572 SSRs (8.47%) and penta-nucleotides with 4,629 SSRs (5.96%). Hexa-nucleotides were the rarest motifs with 2,477 SSRs (3.19%). Only four mono-nucleotide, four di-nucleotide and three tetra-nucleotide motifs were found in the mitochondrial and chloroplast genomes. However, a total length of those SSRs was in a range of 12–16 bp. In addition, in total 58 unique repeats present only in a single copy in a range 101–325 bp were retrieved from the analysis of TEs. Within those were four hexa-, 35 hepta-, nine octa-, five nona- and five deca- nucleotide motifs (Supplementary Table S3).

Divergence time of *Stipa*. The Bayesian phylogenetic reconstruction based on the five loci within NORs revealed the divergence time of *Stipa* from *Brachypodium* around 30.00–35.52 Mya and the putative origin of feather grasses about 2.90–6.02 Mya (Fig. 2). Although not all branches were well supported within the genus, the current analysis confirmed the monophyly of *Stipa* and the general grouping of the analysed species regarding their taxonomic positions. In particular, *S. capillata* and *S. grandis* represent the section *Leiostipa* Dumort; *S. magnifica* Junge, *S. narynica* Nobis, *S. lipskyi* Roshev. and *S. caucasica* Schmalh. belong to the section *Smirnovia* Tzvelev. The remaining three groups include (1) *S. orientalis* Trin. and *S. pennata* L., (2) *S. richteriana* Kar. & Kir., *S. lessingiana* Trin. & Rupr., *S. heptapotamica* Golosk. and *S. korshinskyi* Roshev, (3) *S. lagascae* and *S. breviflora* currently have a discrepancy between morphological and molecular data. In addition, the divergence time estimation indicates that the potential origin of the clade comprising *S. capillata* and *S. grandis* is in a range of 0.67–2.93 Mya while the sister clade has the 95% credibility intervals for that parameter in a range of 2.38–4.78 Mya. Furthermore, the lowest genetic divergence time was registered for *S. lessingiana* and *S. richteriana* (0.00–0.48 Mya) as well as for the split between *S. heptapotamica* and the two above-mentioned species (0.01–0.78 Mya). The divergence times for the rest of taxa are present in Table 5.

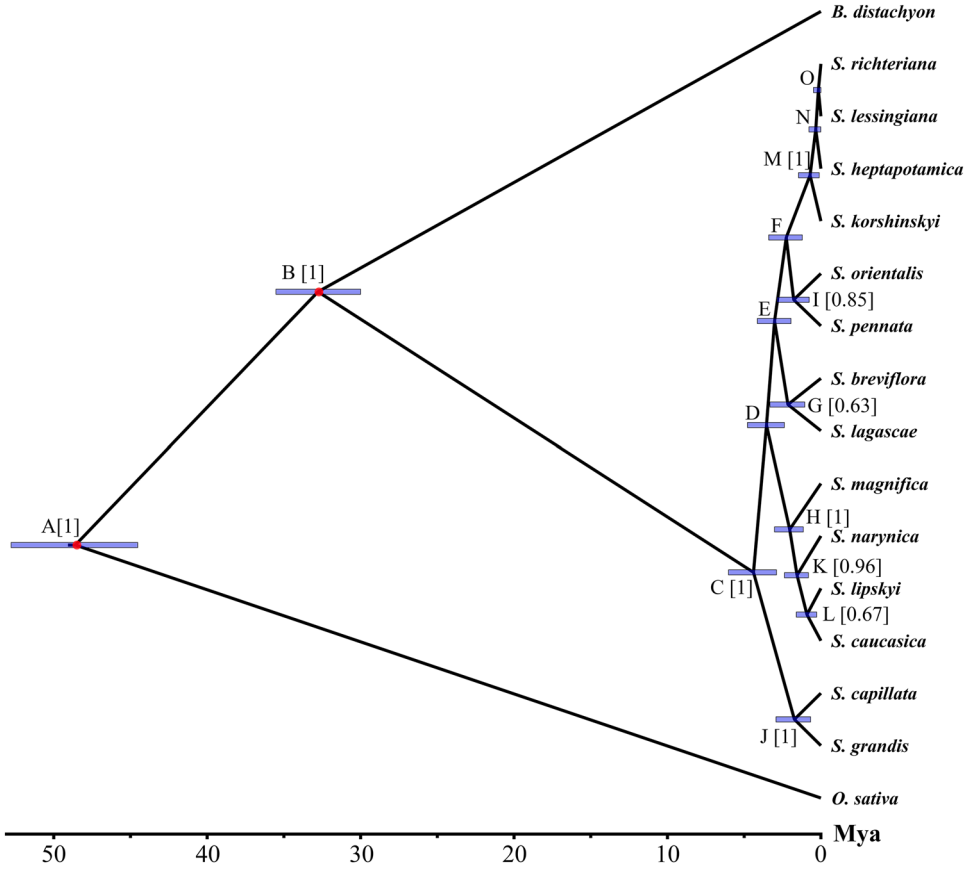

Figure 2. Phylogeny and divergence time estimation by molecular clock analysis. Letters at each node refer to Table 5. Numbers in brackets represent the Bayesian posterior probabilities (BPP > 0.50 only). The blue rectangles on the nodes indicate the 95% credibility intervals (CI) of the estimated posterior distributions of the divergence times. The red circles indicate the presumed divergence time splits set as a reference. The scale on the bottom shows divergence time in Mya. The figure was created using Figtree v1.4.4, https://tree.bio.ed.ac.uk/software/figtree/.

Node	Node age (Mya)	BPP	95% CI
A	48.59	1.00	44.53–52.78
B	32.77	1.00	30.00–35.52
C	4.39	1.00	2.90–6.02
D	3.55	0.40	2.38–4.78
E	3.02	0.28	1.95–4.14
F	2.26	0.40	1.21–3.40
G	2.15	0.63	1.05–3.32
H	2.04	1.00	1.15–3.02
I	1.77	0.85	0.76–2.87
J	1.73	1.00	0.67–2.93
K	1.56	0.96	0.81–2.38
L	0.91	0.67	0.28–1.60
M	0.71	1.00	0.11–1.46
N	0.33	0.39	0.01–0.78
O	0.16	0.28	0.00–0.48

Table 5. Node ages, BPP and CI related to Fig. 2.

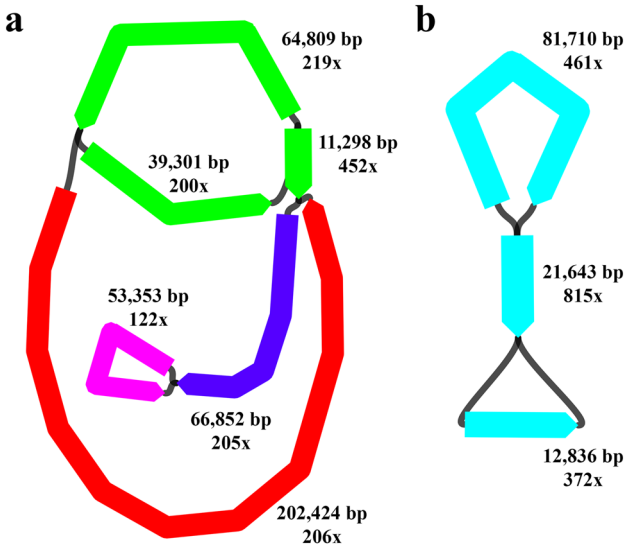

Figure 3. Visualisation of the *de novo* mitochondrial and chloroplast genome assemblies using Bandage v.0.8.1[85]. (**a**) Contigs representing mitochondrion. (**b**) Contig representing chloroplast. Different colours represent different contigs; length (in bp) and coverage (x) of edges within contigs are shown. The figure was created using Bandage v.0.8.1, https://rrwick.github.io/Bandage/.

Assembled mitochondrial and chloroplast genomes. The resulting Flye assembly contained four mitochondrial contigs with a total length of 438,037 bp[76–79] represented by six edges and an entire 137,832 bp-long circular chloroplast genome combining a long single copy region (LSC) of 81,710 bp, a short single copy region (SSC) of 12,836 bp and two inverted repeats (IR) of 21,643 bp each (Fig. 3). However, after a manual checking in IGV v.2.8.6[80] the final size of the chloroplast genome was slightly reduced to 137,823 bp. In addition, an analysis using Cp-hap[81] detected two structural haplotypes of the chloroplast genome: haplotype A[82] (LSC—IR, reverse-complement (rc)—SSCrc—IR) and haplotype B[83] (LSC—IRrc—SSC—IR). We also obtained one assembly using Unicycler v.0.4.8[84] resulted in 76 linear contigs from which 29 can be assigned to mitochondrial sequences with a total length of 1,668,569 bp. Due to the Unicycler assembly being more complex and none of the obtained contigs were likely to be circular in nature, for the downstream genome annotation we used the Flye assembly.

In total, 112 and 133 genes were functionally annotated for mitochondrial and chloroplast genomes, respectively. The mitochondrial annotation resulted in 78 protein-coding genes, 4 ribosomal RNA genes and 30 tRNA genes. The chloroplast annotation contained 85 protein-coding genes, 8 ribosomal RNA genes and 40 tRNA genes. The chloroplast genome size of 137,823 bp generated with Flye and the number of annotated genes in the current study were similar to the known assemblies for *S. capillata* obtained by Illumina sequencing[57]. However, the previous genome assemblies were slightly longer, specifically 137,830 bp[86] and 137,835 bp[87].

DArTseq markers. The DArT pipeline analysis resulted in 61,328 Silico markers and in 52,970 sequences with SNPs. The BLAST process revealed 58,701 Silico markers and 52,252 sequences with SNPs that were successfully mapped to 4,361 and 3,935 genome contigs, respectively. Thus, the current genome assembly has 95.72% of Silico markers and 98.64% of sequences with SNPs that are represented in 73.52% (the total length of 969.30 Mb) and 66.34% (940.37 Mb) of the contigs, respectively. In addition, we established that 50,953 Silico markers and 47,181 sequences with SNPs were present only in a single copy in the genome. Finally, we identified 30 Silico markers and 10 sequences with SNPs aligned to the mitochondrial genome and only 2 Silico markers and 4 sequences with SNPs that were found in the chloroplast genome.

Discussion

The number of sequenced plant genomes is rapidly increasing year by year serving as a fundamental resource for various genomic studies. In the current work, we present a 1004 Mb genome with the 23 × coverage of the most widespread feather grass species, *S. capillata*, using SMRT PacBio sequencing. The current assembly comprises 5,931 sequences with a contig N50 length of 351 kb (Table 1). The BUSCO completeness score of 93.10% (Table 2), the observation of a large portion of TEs (57.68%, Table 4) and the presence of Silico (95.72%) and SNPs (98.64%) markers derived from the DArT platform indicate that the assembly is of high quality. Moreover, the proportion of TEs has been reported for the first time in the genus due to the previous *de novo* assemblies which were performed exclusively based on transcriptomic data[50,52,54]. In addition, here we also attempted to perform a reference-guided scaffolding of the assembled contigs. Nevertheless, although nearly all contigs of the *S. capillata* genome were assigned to the chromosomes of *B. distachyon*, *H. vulgare* and *A. tauschii*, it was not possible to estimate their proper position on the reference with an acceptable level of confidence (Table 3

and Supplementary Table S2). In general, in the absence of a high-density genetic linkage map the task of reconstructing pseudomolecules of chromosomes seems to be challenging. On the other hand, we believe that in order to improve the contiguity of the long-read assembly the high-throughput chromosome conformation capture (Hi-C)[88] technique should be applied. Currently, many studies on non-model species successfully utilised a combination of long-read techniques and Hi-C data to perform assemblies at chromosome scale[89–91]. Moreover, an additional key for improving this genome assembly in the future is merely to get more sequencing reads. Recently, it was shown that contig length metrics are positively correlated with both read length and sequence coverage. Specifically, long-read assemblies in maize demonstrated that the highest contig N50 of 24.54 Mb was reached with a subread N50 of 21,166 bp and a 75-fold depth of coverage while the longest contig of 79.68 Mb was observed with the same subread N50 but with a 60-fold depth[92].

The newly generated genome has a GC content of 45.97% that is similar to the known estimates for species in *Stipa* varying in a range of 46.61–49.05%[93], and more broadly to grasses ranging from 43.57% in *O. sativa* to 46.90% in *Z. mays*[94]. Recently, it was shown that a higher GC content in monocots is associated with adaptation to extremely cold and/or dry climates[95]. The genus *Stipa* highly supports this hypothesis due to the fact that all feather grasses are adapted to temperate, dry climates[36]. In addition, a positive correlation between the GC content and genome size was established[96] suggesting insertion of LTR retrotransposons as a potential driving force of genome enlargement[97]. Similarly, here we showed that the expansion of the *S. capillata* genome also resulted from insertions of repetitive sequences that occupy 57.68% of the genome including LTR retrotransposons (13.97%). However, among all repetitive sequences around 34.34% are currently unclassified (Table 4). Nonetheless, the total proportion of TEs in *S. capillata* in comparison to other species within the Poaceae family is close to *Oryza minuta* J. Presl (58.35%) and *O. alta* Swallen[98] (57.54%), bigger than in *B. distachyon*[99] (28.10%) and *O. sativa*[100] (45.52%) and smaller than in *O. granulata* Nees & Arn.[101] (67.96%), *Avena sativa* L.[102] (69.47%) and *T. aestivum*[103] (84.67%).

Importantly, the presented genome size is roughly twice smaller than the expected size of 2,355 Mb and twice bigger than the expected monoploid size of 589 Mb estimated using flow cytometry[93]. Considering that we were unable to remove redundant sequences due to possible heterozygosity and the number of duplicated BUSCOs (Tables 1 and 2), it may be presumed that the current genome assembly combines two very distinct genomes. To the current knowledge, the vast majority of *Stipa* species have 44 (2n = 4x) chromosomes and are supposed to be tetraploids[41,104]. In addition, recently it was shown that a single-copy region *ACC1* and a low-copy nuclear gene *At103* have two different copies in *Stipa*[104,105]. Thus, it may suggest that *S. capillata*, and the genus *Stipa* in general, has arisen through hybridisation between genetically distant diploid species (2n = 22) and the subsequent allopolyploidisation via whole genome duplication (WGD) rather than via one WGD event of an ancestral species. Well-documented examples of natural allopolyploid taxa in the Pooideae subfamily are *Triticum turgidum* L. (2n = 4x = 28, genome constitution AABB) and *T. aestivum* (2n = 6x = 42, AABBDD) formed through hybridisation and successive chromosome doubling of ancestral diploid species *T. urartu* (2n = 2x = 14, AA), *Aegilops speltoides* Tausch. (2n = 2x = 14, BB) and *A. tauschii* (2n = 2x = 14, DD)[106]. Moreover, in the tribe Stipeae based on the *At103* gene allopolyploidy was reported for the genus *Patis* Ohwi (2n = 46, 48)[105]. Heretofore, at least three hypotheses were considered regarding the base chromosome number in Stipeae: x = 7[107], x = 11[108,109] and x = 12[110]. Recently, it was suggested that the latter two are more plausible[41,104]. Thus, in order to better assemble the *S. capillata* genome and verify if *Stipa* is an allopolyploid genus we suggest sequencing at chromosome level the close relative diploid species (2n = 22) from genera representing, e.g. *Ptilagrostis* Griseb., *Achnatherum* P. Beauv., e.g. *A. calamagrostis* L. (2n = 22 + 0−2B), or *Piptatheropsis* Romasch., P. M. Peterson & Soreng (2n = 20, 22, 24)[41,104].

In general, the number of genes in Poaceae varies from 28,835 in the smallest known genome, *Oropetium thomaeum* Trin. (2n = 20; genome size of 245 Mb)[111], to 107,891 in *T. aestivum* (2n = 42; 14,547 Mb)[112]. Here, we reported 53,535 nuclear genes that were structurally annotated for the masked genome assembly. Such a number of genes was roughly 1.8 and 1.6 times smaller than previously determined for *S. grandis* (94,674 genes)[54] and *S. purpurea* (84,298 genes)[50], respectively. On the other hand, the annotation analysis of the unmasked genome resulted in 81,224 genes associated with already known proteins. In comparison, only 65,047 functionally annotated genes were reported for *S. grandis* while *S. purpurea* had 58,966. Nonetheless, as RNA-seq data is currently unavailable for *S. capillata*, we believe that the current version of the genome annotation demands a further investigation to properly characterise the genes sets when the appropriate information will be available.

SSR markers are widely distributed across the genome and they are commonly applied in establishing genetic structure in *Stipa*. Previously, polymorphic microsatellite primers were reported in populations of *S. purpurea* (11[113], 15[114] and 29[115] loci), *S. pennata* (7 loci[116]), *S. breviflora* (21 loci[117]) and *S. glareosa* (9 loci[118]). In the present study, we identified 77,614 perfect SSR markers (Supplementary File 3) and 58 imperfect repeat motifs presented only in a single copy (Supplementary Table S3). Although we did not test them on the population level we are confident that such a number of new loci will be a valuable source for the farther development of SSR markers in *S. capillata*, and more generally in the genus *Stipa*. Additionally, the revealed loci could be used for the designing dominant inter simple sequence repeat (ISSR) markers[119]. Recently, the usefulness of applying ISSRs were shown for studies in *S. bungeana*[120], *S. ucrainica* and *S. zalesskii*[121], *S. tenacissima*[122] and the hybrid complex *S. heptapotamica*[123].

According to the previous studies, based on three chloroplast loci[124] and four chloroplast loci and one nuclear region[105], it was shown that the origin of Stipeae can be estimated in a range of 30.60–47.30 Mya and 21.20–39 Mya, respectively. Here, based on the five loci within NORs we demonstrated that the potential split between *Stipa* representing the tribe Stipeae and *Brachypodium* (the tribe Brachypodieae) took place approximately 30–35.52 Mya that supports the previous findings[105,124,125]. The present results also suggest that the genus *Stipa* likely originated ca. 4.39 (2.90–6.02) Mya. On the other hand, one previous study indicated the origin of feather grasses at about 12.90 Mya[124] while another one showed different estimates based on chloroplast loci (21.20 Mya, 13–22) and the *At103* region[105]. Specifically, two copies of *At103* had the following suggested ages: 15.78

(6.30–26.60) Mya for the Eurasian Stipeae lineage and 5.62 (0–6.50) Mya for the American Stipeae lineage[105]. Thus, the latter estimate is close enough to the origin-age calculated in the current study. In addition, our data on the divergence time among *S. richteriana*, *S. lessingiana* and *S. heptapotamica* (Fig. 2 and Table 5) conforms to the previous findings on the ongoing hybridisation among these taxa[123] suggesting NORs as a useful tool for revealing species of putative hybrid origin. Nonetheless, we believe that the current and previous estimates regarding the origin of *Stipa* should be treated with caution. Firstly, to our knowledge, there is still no available fossil data for any *Stipa* species from the Old World that can properly calibrate the historical diversification in the genus. Currently, the earliest definite *Stipa* caryopses were found in central Poland and are dated ca. 4,000 BC[126]. Secondly, available data demonstrate incongruence between chloroplast and nuclear loci analyses. In further studies we suggest utilising single-copy nuclear genes derived from whole genome sequencing projects. Thirdly, different sets of species and parameters used for inferring diversification dates may result in different estimates[127].

Finally, we report a 137,823 bp chloroplast genome that is similar to the known assemblies in *Stipa* and specifically in *S. capillata*[57]. Here we highlight the applicability of a long-read sequencing technology like PacBio for the straightforward assembling of plastomes using Flye[67,68]. In addition, due to the long-reads we were able to identify two haplotypes presented in *S. capillata*. This result supports the previous findings in Poaceae[81] suggesting that plastome structural heteroplasmy can be a common state in feather grasses. Moreover, for the first time in the genus *Stipa*, here we present a 438,037 bp mitochondrial genome. The current size of this genome is close to *Alloteropsis semialata* (R.Br.) Hitchc. (442,063 bp)[128], *T. aestivum* (452,526 bp)[129], *Sorghum bicolor* L. (468,628 bp)[130] and *A. speltoides* (476,091 bp)[131]. Nevertheless, the present version of the genome is constituted by four contigs rather than one circular sequence. Although the general acceptance among mitochondrial biologists is that plant mitochondrial genomes have a variety of configurations[132–134], in order to verify if a more accurate assembly could be performed, we suggest reusing our data for a more comprehensive analysis of the mitochondrial structures within *Stipa*.

Materials and methods

Plant material and DNA extraction. Our research complies with relevant institutional, national, and international guidelines and legislation. A *S. capillata* sample from Kochkor River Valley, central Kyrgyzstan (Supplementary Table S4), was selected for genome sequencing. The sample was stored in silica gel at ambient temperature until DNA extraction was performed. Total genomic DNA was isolated from dried leaves after a six-month storage period using a CTAB large-scale DNA extraction protocol (Supplementary information S1, described in Supplementary File 6). DNA extraction was performed by SNPsaurus (USA). In addition, we isolated DNA from dried leaves using a Genomic Mini AX Plant Kit (A&A Biotechnology, Poland). Subsequently, quality check, quantification and concentration adjustment were accomplished using a NanoDrop One (Thermo Scientific, USA) and agarose gel electrophoresis visualisation. The concentration of the sample was adjusted to 50 ng/μL. The purified DNA sample (1 μg) was sent to Diversity Arrays Technology Pty Ltd (Canberra, Australia) for sequencing and DArT marker identification. Moreover, to test the phylogenetic power of NORs in *Stipa*, we supplemented the study with five specimens of *S. richteriana* Kar. & Kir, three of *S. lessingiana* Trin. & Rupr., four of *S. heptapotamica* Golosk. and four of *S. korshinskyi* Roshev. (Supplementary Table S4). The isolation of genomic DNA was performed from dried leaf tissues using a modified CTAB method[135].

Library construction and sequencing. In total, 5 ug of *S. capillata* genomic DNA were used to construct a PacBio library according to the 20 kb PacBio template preparation protocol omitting a shearing step. The size selection cut-off was set at 15 kb. The library preparation followed by sequencing on three PacBio Sequel SMRT cells (Pacific Biosciences, Menlo Park, CA, USA) was carried out by SNPsaurus, LLC. Prior to the assembly, reads from each SMRT cell were inspected and quality metrics were calculated using SequelQC v.1.1.0[136]. A high-density assay using the DArT complexity reduction method for *S. capillata* was performed according to a previously reported procedure[137].

For the rest of the specimens used in the current study, the quality control using a fluorometer (PerkinElmer Victor3, USA) and gel electrophoresis, library construction using a TruSeq Nano DNA Library kit (350 bp insert size; Illumina, USA) and sequencing using 100 bp paired-end reads on an Illumina HiSeq 2500 platform (Illumina, USA) were performed by Macrogen Inc. (South Korea).

Nuclear genome assembly and validation. The execution of this work involved using many software tools, whose versions, settings and parameters are described in Supplementary information S2 (available in Supplementary File 6). The *de novo* assembly of the PacBio data was performed using Flye v.2.4[65,66]. The draft assembly was cleaned by running BLASTn v.2.10.0[138] against the NCBI nucleotide database v.5, and subsequently sending each BLAST hit to the JGI taxonomy server (https://taxonomy.jgi-psf.org/) with a downstream step of keeping only plant contigs. Thereafter, Qualimap v.2.2.2[139] was used to identify mean coverage for each contig. In the final assembly we kept only contigs with an average coverage of more than 10x. In addition, overrepresented contigs (>60x) were BLASTed against the NCBI nucleotide database v.5 and sequences assigned to chloroplasts and mitochondria were removed.

Due to the final assembly performed with Flye v.2.4 being roughly twice bigger than an expected monoploid genome size of 589 Mb[93], we accomplished an additional assembly with FALCON v.0.2.5[68] and applied Purge Haplotigs v1.1.1[69] to filter redundant sequences due to possible heterozygosity. The assemblies' statistics were analysed using assembly-stats v.1.0.1[140]. In addition, in order to assess the completeness of the genome assemblies, we investigated the presence of highly conserved orthologous genes using BUSCO v.4.0.6[141].

Scaffolding of contigs. Due to there being no reference genome for any *Stipa* species, here we applied RaGOO v.1.1[142] to verify if a reference-guided scaffolding can be performed for the draft genome contigs based on four genomes from the Pooideae subfamily (*B. distachyon*[70], *H. vulgare*[71], *A. tauschii*[72], *T. aestivum*[74]) and one genome from the Oryzoideae subfamily (*O. sativa*[73]). The subsequent assessment of the scaffolding accuracy was based on three parameters: (1) location confidence score, (2) orientation confidence score and (3) grouping confidence score[142].

Repeat prediction and nuclear genome annotation. The repeat prediction for *S. capillata* was performed using a *de novo* transposable element (TE) family identification and modeling package RepeatModeler v.2.0.1[143] which includes three repeat finding programs; RECON[144], RepeatScout[145], and TRF[146]. The resulting TE library was supplemented by the transposable elements database (Release 19, http://botserv2.uzh.ch/kelldata/trep-db/).[147] Subsequently, the genome assembly was masked for TEs regions by RepeatMasker v.4.1.0[148] (http://repeatmasker.org) with the search engine RMBlast v.2.9.0 + [149] and the custom library created in the previous step. Next, gene and protein sequences were predicted using Augustus v.3.2.3 with the unmasked and v.3.3.3[150] with the masked genome assemblies. The predicted protein sequences of the unmasked assembly were then BLASTed against the NCBI protein database v.5 and the subsequent BLAST hit descriptions were added to GFF (General Feature Format) files.

Genome-wide identification of microsatellite markers. The unmasked nuclear genome, chloroplast and mitochondrial genome assemblies were screened for perfect mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeat motifs using Krait v.1.3.3[75]. We applied the following criteria: mono-nucleotide repeat motifs contain at least 12 repeats, di-nucleotide repeat motifs contain at least seven repeats, tri-nucleotide repeat motifs contain at least five repeats, tetra-, penta- and hexa-nucleotide repeat motifs contain at least four repeats.

Divergence time of *Stipa*. In order to estimate the divergence between *S. capillata* and other *Stipa* species we used the nucleolar organising regions. Firstly, we prepared a set of reference sequences including *S. lipskyi* Roshev.[151], *S. magnifica* Junge[152], *S. narynica* Nobis[153], *S. caucasica* Schmalh.[154], *S. orientalis* Trin.[155] and *S. pennata* L.[156]. Secondly, we mapped raw reads of *S. capillata*, *S. richteriana*, *S. lessingiana*, *S. heptapotamica* and *S. korshinskyi* (Supplementary Table S2) as well as *S. grandis*[55], *S. breviflora*[59], *S. lagascae*[157] to the reference set using Minimap2 v.2.17-r941[158] with keeping only uniquely mapped reads by Samtools v.1.9[159]. Thirdly, the *de novo* assembly of the NORs was performed using Canu v.2.0[160] for *S. capillata* and SPAdes v.3.14.1[161] for the rest of *Stipa* species. Additionally, we added to the analysis *B. distachyon*[162] as an ingroup member of the Pooideae subfamily and *O. sativa*[163] as an outgroup representing the Oryzoideae subfamily within the Poaceae family. Next, all sequences were aligned using MAFFT v.7.471[164]. Subsequently, the aligned sequences were visualised in AliView v.1.26[165] and divided in five loci: (1) 18S ribosomal RNA, (2) Internal Transcribed Spacer 1 (ITS1), (3) 5.8S ribosomal RNA, (4) Internal Transcribed Spacer 2 (ITS2) and (5) 26S ribosomal RNA (Supplementary File 4). Estimation of divergence times was performed in BEAST2 v.2.6.3[166] using the 121,321 substitution model determined by bModelTest[167]. We used the following constraints for time calibrations: 38–48 million years ago (Mya) for the *Brachypodium-Oryza* split[101] and 33–39 Mya for the potential origin and divergence of *Stipa*[34,35]. Then, the divergence time was estimated using the strict clock model and the Yule prior. In total, we ran the analysis three times independently, 50 million Markov chain Monte Carlo (MCMC) generations for each run. The log and tree files were combined using LogCombiner v.2.6.3 (a part of the BEAST package) with the first five million generations discarded as burn-in from each run. Next, Tracer v.1.7.1[168] was used to check the log files regarding Effective Sample Size (ESS) values. As all ESSs exceeded 200, we summarised the final maximum clade credibility tree (Supplementary File 5) in TreeAnnotator v.2.6.3 (a part of the BEAST package). The final tree was visualised and edited using FigTree v.1.4.4[169].

Mitochondrial and chloroplast genomes assembly, annotation and validation. Prior to assembly, we mapped raw reads to 11 reference mitochondrial genomes of species belonging to the Poaceae family (Supplementary Table S5) using Minimap2 v.2.17-r941[158]. Only uniquely mapped reads were kept by Samtools v.1.9[159] for the next step. *De novo* mitochondrial assembly of the 4.08 Mb data was performed using Flye v.2.7.1-b1590.

In the next step, we BLASTed the resulting contigs against the NCBI nucleotide database v.5, and sequences assigned to mitochondria were kept. Then, the PacBio subreads were mapped onto the kept contigs using Minimap2, and only uniquely mapped reads were retained by Samtools. A new *de novo* assembly of the 15.51 Mb data was performed using Flye. In order to check if the mitochondrial contigs obtained by Flye could be merged into larger scaffolds we applied Circlator v.1.5.5[170]. However, the resulting sequences were identical to the Flye contigs. In addition, we used Unicycler v.0.4.8[84] with reads that were mapped onto the Flye contigs as a reference.

Further, to detect all possible structural haplotypes of the chloroplast genome we applied Cp-hap[81]. Next, we mapped raw reads onto the resulting mitochondrial contigs and the chloroplast genomes to manually check in IGV v.2.8.6[80] if any potential SNPs or indels are present. Eventually, annotations of the final mitochondrial contigs of 438,037 bp and the chloroplast genomes of 137,823 bp were performed using Geneious Prime v.2021.1.1 (https://www.geneious.com) based on 85% and 95% similarities to the reference genomes of mitochondria and chloroplasts, respectively (Supplementary Table S5).

In Silico mapping of DArT marker sequences. Since the DArT markers are designed to target active regions of the genome[171], here we use them to validate the completeness of the nuclear genome assembly and

improve the accuracy of data filtering in further genomic studies on *Stipa*. Two data types, Silico and SNPs markers, were mapped to the nuclear genome using BLASTn v.2.10.0. As a query we used trimmed DArT sequences in a range of 29–69 bp with the percent identity values to the reference genome of 95% or greater and removing alignments below 95% of a query.

Data availability

The raw PacBio reads are available at NCBI Sequence Read Archive[172]. The final genome assemblies are deposited into NCBI Assembly database under the following Accession Numbers: nuclear assembly (JAGXJF000000000)[67]; mitochondrion assembly, contig 1 (MZ161090)[76], contig 2 (MZ161091)[77], contig 3 (MZ161093)[78] and contig 4 (MZ161092)[79]; chloroplast assemblies, haplotype A (MZ146999)[82] and haplotype B (MZ145043)[83]. The masked and the unmasked versions of the nuclear genome annotation are presented in the Supplementary File 1 and the Supplementary File 2, respectively.

Received: 17 November 2020; Accepted: 24 June 2021
Published online: 28 July 2021

References

1. Initiative, T. A. G. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815. https://doi.org/10.1038/35048692 (2000).
2. Moreau, H. *et al.* Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* **13**, R74. https://doi.org/10.1186/gb-2012-13-8-r74 (2012).
3. Hamaji, T. *et al.* Anisogamy evolved with a reduced sex-determining region in volvocine green algae. *Commun. Biol.* **1**, 17. https://doi.org/10.1038/s42003-018-0019-5 (2018).
4. Rensing, S. A. *et al.* The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**, 64–69. https://doi.org/10.1126/science.1150646 (2008).
5. Bowman, J. L. *et al.* Insights into Land Plant Evolution Garnered from the *Marchantia polymorpha* Genome. *Cell* **171**, 287–304. https://doi.org/10.1016/j.cell.2017.09.030 (2017).
6. Li, F. W. *et al.* Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat. Plants* **4**, 460–472. https://doi.org/10.1038/s41477-018-0188-8 (2018).
7. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579–584. https://doi.org/10.1038/nature12211 (2013).
8. Mosca, E. *et al.* A reference genome sequence for the European silver fir (*Abies alba* Mill): A community-generated genomic resource. *G3: Genes Genomes, Genetics* **9**, 2039–2049. https://doi.org/10.1534/g3.119.400083 (2019).
9. Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089. https://doi.org/10.1126/science.1241089 (2013).
10. Strijk, J. S., Hinsinger, D. D., Zhang, F. & Cao, K. *Trochodendron aralioides*, the first chromosome-level draft genome in Trochodendrales and a valuable resource for basal eudicot research. *GigaScience* **8**, 11. https://doi.org/10.1093/gigascience/giz136 (2019).
11. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92. https://doi.org/10.1126/science.1068037 (2002).
12. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551–556. https://doi.org/10.1038/nature07723 (2009).
13. International Wheat Genome Sequencing Consortium. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, 705. https://doi.org/10.1126/science.aar7191 (2018).
14. Bennetzen, J. L. *et al.* Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* **30**, 555–561. https://doi.org/10.1038/nbt.2196 (2012).
15. Studer, A. J. *et al.* The draft genome of the C_3 panicoid grass species *Dichanthelium oligosanthes*. *Genome Biol.* **17**, 223. https://doi.org/10.1186/s13059-016-1080-3 (2016).
16. Gordon, S. P. *et al.* Gradual polyploid genome evolution revealed by pan-genomic analysis of *Brachypodium hybridum* and its diploid progenitors. *Nat. Commun.* **11**, 3670. https://doi.org/10.1038/s41467-020-17302-5 (2020).
17. Yagi, M. *et al.* Sequence analysis of the genome of carnation (*Dianthus caryophyllus* L.). *DNA Res.* **21**, 231–241. https://doi.org/10.1093/dnares/dst053 (2014).
18. Cai, J. *et al.* The genome sequence of the orchid *Phalaenopsis equestris*. *Nat. Genet.* **47**, 65–72. https://doi.org/10.1038/ng.3149 (2015).
19. Kim, Y. M. *et al.* Genome analysis of *Hibiscus syriacus* provides insights of polyploidization and indeterminate flowering in woody plants. *DNA Res.* **24**, 71–80. https://doi.org/10.1093/dnares/dsw049 (2017).
20. Li, L. *et al.* Genome sequencing and population genomics modeling provide insights into the local adaptation of weeping forsythia. *Horticulture Res.* **7**, 130. https://doi.org/10.1038/s41438-020-00352-7 (2020).
21. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17. https://doi.org/10.1186/2047-217X-3-17 (2014).
22. Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *PNAS* **111**, 4859–4868. https://doi.org/10.1073/pnas.1323926111 (2014).
23. Leebens-Mack, J. H. *et al.* One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**, 679–685. https://doi.org/10.1038/s41586-019-1693-2 (2019).
24. Cheng, S. *et al.* 10KP: A phylodiverse genome sequencing plan. *GigaScience* **7**, giy013. https://doi.org/10.1093/gigascience/giy013 (2018).
25. Pellicer, J., Fay, M. F. & Leitch, I. J. The largest eukaryotic genome of them all?. *Bot. J. Linn. Soc.* **164**, 10–15. https://doi.org/10.1111/j.1095-8339.2010.01072.x (2010).
26. Stevens, K. A. *et al.* Sequence of the sugar pine megagenome. *Genetics* **204**, 1613–1626. https://doi.org/10.1534/genetics.116.193227 (2016).
27. Meyers, L. A. & Levin, D. A. On the abundance of polyploids in flowering plants. *Evolution* **60**, 1198–1206. https://doi.org/10.1111/j.0014-3820.2006.tb01198.x (2006).
28. Flavell, R. B., Bennett, M. D., Smith, J. B. & Smith, D. B. Genome size and proportion of repeated nucleotide-sequence DNA in plants. *Biochem. Genet.* **12**, 257–269 (1974).
29. Schnable, P. S. *et al.* The B73 maize genome: Complexity diversity and dynamics. *Science* **326**, 1112–1115. https://doi.org/10.1126/science.1178534 (2009).

30. Daron, J. *et al.* Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol.* **15**, 546. https://doi.org/10.1186/s13059-014-0546-4 (2014).
31. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138. https://doi.org/10.1126/science.1162986 (2009).
32. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* **4**, 265–270. https://doi.org/10.1038/nnano.2009.12 (2009).
33. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239. https://doi.org/10.1186/s13059-016-1103-0 (2016).
34. MacGinitie, H. D. *Fossil Plants Of The Florissant Beds, Colorado* (Carnegie Institute of Washington Publication, 1953).
35. Manchester, S. R. Update on the megafossil flora of Florissant Colorado. *Denver Museum Nat. Sci.* **4**, 137–161 (2001).
36. Freitag, H. The genus *Stipa* (Gramineae) in southwest and south Asia. *Notes From Royal Botanic Garden* **42**, 355–489 (1985).
37. Barkworth, M. E. & Everett, J. Evolution in the Stipeae: identification and relationships of its monophyletic taxa. *Grass systematics and evolution* (eds. Soderstrom, T. R., Hilu, K. W., Campbell, C. S. & Barkworth, M. E.) 251–264 (Smithsonian Institution Press, 1987).
38. Hamasha, H. R., von Hagen, K. B. & Röser, M. *Stipa* (Poaceae) and allies in the Old World: molecular phylogenetics realigns genus circumscription and gives evidence on the origin of American and Australian lineages. *Plant Syst. Evol.* **298**, 351–367. https://doi.org/10.1007/s00606-011-0549-5 (2012).
39. Kellogg, E. A. Subfamily Pooideae *in The families and genera of vascular plants* (ed. Kubitzki, K.) 199–229 (Springer International Publishing, 2015).
40. Nobis, M. Taxonomic revision of the Central Asiatic *Stipa tianschanica* complex (Poaceae) with particular reference to the epidermal micromorphology of the lemma. *Folia Geobot.* **49**, 283–308. https://doi.org/10.1007/s12224-013-9164-2 (2014).
41. Romaschenko, K. *et al.* Systematics and evolution of the needle grasses (Poaceae: Pooideae: Stipeae) based on analysis of multiple chloroplast loci, ITS, and lemma micromorphology. *Taxon* **61**, 18–44. https://doi.org/10.1002/tax.611002 (2012).
42. Nobis, M., Gudkova, P. D., Nowak, A., Sawicki, J. & Nobis, A. A synopsis of the genus *Stipa* (Poaceae) in Middle Asia, including a key to species identification, an annotated checklist, and phytogeographic analyses. *Ann. Mo. Bot. Gard.* **105**, 1–63. https://doi.org/10.3417/2019378 (2020).
43. Yunatov, A. A. Main patterns of the vegetation cover of the Mongolian people's republic. *Proc. Mongolian Commission* **39**, 233 (1950).
44. Lavrenko, E. M., Karamasheva, Z. V. & Nikulina, R. I. *Eurasian steppe*. 143 (Nauka, 1991).
45. Nowak, A., Nowak, S., Nobis, A. & Nobis, M. Vegetation of feather grass steppes in the western Pamir Alai Mountains (Tajikistan, Middle Asia). *Phytocoenologia* **46**, 295–315. https://doi.org/10.1127/phyto/2016/0145 (2016).
46. Danzhalova, E. V. *et al.* Indicators of pasture digression in steppe ecosystems of Mongolia. *Exploration Biol. Resour. Mongolia* **12**, 297–306 (2012).
47. Maevsky, V. V. & Amerkhanov, H. H. The note of Poaceae species from former USSR flora, recommended as fodder for agricultural production. *Bull. Botanical Garden Saratov State Univ.* **6**, 80–83 (2007).
48. Brunetti, G., Soler-Rovira, P., Farrag, K. & Senesi, N. Tolerance and accumulation of heavy metals by wild plant species grown in contaminated soils in Apulia region Southern Italy. *Plant Soil* **318**, 285–298. https://doi.org/10.1007/s11104-008-9838-3 (2009).
49. Moameri, M. *et al.* Investigating lead and zinc uptake and accumulation by Stipa hohenackeriana Trin and Rupr in field and pot experiments. *Biosci. J.* **34**, 138–150. https://doi.org/10.14393/BJ-v34n1a2018-37238 (2018).
50. Yang, Y. Q. *et al.* Transcriptome analysis reveals diversified adaptation of *Stipa purpurea* along a drought gradient on the Tibetan Plateau. *Funct. Integr. Genomics* **15**, 295–307. https://doi.org/10.1007/s10142-014-0419-7 (2015).
51. Lv, X., He, Q. & Zhou, G. Contrasting responses of steppe Stipa ssp to warming and precipitation variability. *Ecol. Evolut.* **9**, 9061–9075. https://doi.org/10.1002/ece3.5452 (2019).
52. Schubert, M., Grønvold, L., Sandve, S. R., Hvidsten, T. R. & Fjellheim, S. Evolution of cold acclimation and its role in niche transition in the temperate grass subfamily Pooideae. *Plant Physiol.* **180**, 404–419. https://doi.org/10.1104/pp.18.01448 (2019).
53. *NCBI BioSample*, https://www.ncbi.nlm.nih.gov/biosample/?term=SAMN03178190 (2014).
54. Wan, D. *et al.* De novo assembly and transcriptomic profiling of the grazing response in *Stipa grandis*. *PLoS ONE* **10**, e0122641. https://doi.org/10.1371/journal.pone.0122641 (2015).
55. *NCBI Sequence Read Archive*, https://www.ncbi.nlm.nih.gov/sra/?term=SRP051667 (2020).
56. *ArrayExpress*, https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5300 (2020).
57. Krawczyk, K., Nobis, M., Myszczyński, K., Klichowska, E. & Sawicki, J. Plastid superbarcodes as a tool for species discrimination in feather grasses (Poaceae: *Stipa*). *Sci. Rep.* **8**, 1924. https://doi.org/10.1038/s41598-018-20399-w (2018).
58. *NCBI Sequence Read Archive*, https://www.ncbi.nlm.nih.gov/sra/SRR8208353 (2020).
59. *NCBI Sequence Read Archive*, https://www.ncbi.nlm.nih.gov/sra/SRS3290204 (2020).
60. Krawczyk, K., Nobis, M., Nowak, A., Szczecińska, M. & Sawicki, J. Phylogenetic implications of nuclear rRNA IGS variation in Stipa L (Poaceae). *Sci. Rep.* **7**, 11506. https://doi.org/10.1038/s41598-017-11804-x (2017).
61. Wagner, V. *et al.* Similar performance in central and range-edge populations of a Eurasian steppe grass under different climate and soil pH regimes. *Ecography* **34**, 498–506. https://doi.org/10.1111/j.1600-0587.2010.06658.x (2011).
62. Wagner, V., Durka, W. & Hensen, I. Increased genetic differentiation but no reduced genetic diversity in peripheral vs. central populations of a steppe grass. *Am. J. Botany* **98**, 1173–1179. https://doi.org/10.3732/ajb.1000385 (2011).
63. Durka, W. *et al.* Extreme genetic depauperation and differentiation of both populations and species in Eurasian feather grasses (*Stipa*). *Plant Syst. Evol.* **299**, 259–269. https://doi.org/10.1007/s00606-012-0719-0 (2013).
64. Kirschner, P. *et al.* Long-term isolation of European steppe outposts boosts the biome's conservation value. *Nat. Commun.* **11**, 1968. https://doi.org/10.1038/s41467-020-15620-2 (2020).
65. Lin, Y. *et al.* Assembly of long error-prone reads using de Bruijn graphs. *Proc. Natl. Acad. Sci.* **113**, 8396–8405. https://doi.org/10.1073/pnas.1604560113 (2016).
66. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546. https://doi.org/10.1038/s41587-019-0072-8 (2019).
67. *NCBI Assembly*, https://www.ncbi.nlm.nih.gov/assembly/JAGXJF000000000 (2021).
68. Chin, C.-H. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054. https://doi.org/10.1038/nmeth.4035 (2016).
69. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf.* **19**, 460. https://doi.org/10.1186/s12859-018-2485-7 (2018).
70. *NCBI Assembly*, https://www.ncbi.nlm.nih.gov/assembly/GCF_000005505.3 (2021).
71. *NCBI Assembly*, https://www.ncbi.nlm.nih.gov/assembly/GCA_903813605.1 (2021).
72. *NCBI Assembly*, https://www.ncbi.nlm.nih.gov/assembly/GCA_002575655.1 (2021).
73. *NCBI Assembly*, https://www.ncbi.nlm.nih.gov/assembly/GCF_001433935.1 (2021).
74. *NCBI Assembly*, https://www.ncbi.nlm.nih.gov/assembly/GCA_002220415.3 (2021).
75. Du, L., Zhang, C., Liu, Q., Zhang, X. & Yue, B. Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics* **34**, 681–683. https://doi.org/10.1093/bioinformatics/btx665 (2018).
76. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MZ161090 (2021).

77. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MZ161091 (2021).
78. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MZ161093 (2021).
79. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MZ161092 (2021).
80. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Can. Res.* **77**, 31–34. https://doi.org/10.1158/0008-5472.CAN-17-0337 (2017).
81. Wang, W. & Lanfear, R. Long-reads reveal that the chloroplast genome exists in two distinct versions in most plants. *Genome Biol. Evol.* **11**, 3372–3381. https://doi.org/10.1093/gbe/evz256 (2019).
82. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MZ146999 (2021).
83. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MZ145043 (2021).
84. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, 1–22. https://doi.org/10.1371/journal.pcbi.1005595 (2017).
85. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualisation of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352. https://doi.org/10.1093/bioinformatics/btv383 (2015).
86. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/NC_037026.1 (2020).
87. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MG052599.1 (2020).
88. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chen-Shan, C. Scaffolding of long read assemblies using long range contact information. *BMC Genom.* **18**, 527. https://doi.org/10.1186/s12864-017-3879-z (2017).
89. Carballo, J. *et al.* A high-quality genome of *Eragrostis curvula* grass provides insights into Poaceae evolution and supports new strategies to enhance forage quality. *Sci. Rep.* **9**, 10250. https://doi.org/10.1038/s41598-019-46610-0 (2019).
90. Chen, B. *et al.* The sequencing and de novo assembly of the *Larimichthys crocea* genome using PacBio and Hi-C technologies. *Scientific Data* **6**, 188. https://doi.org/10.1038/s41597-019-0194-3 (2019).
91. Shan, T. *et al.* First genome of the brown alga *Undaria pinnatifida*: Chromosome-level assembly using PacBio and Hi-C technologies. *Front. Genet.* **11**, 140. https://doi.org/10.3389/fgene.2020.00140 (2020).
92. Ou, S. *et al.* Effect of sequence depth and length in long-read assembly of the maize inbred NC358. *Nat. Commun.* **11**, 2288. https://doi.org/10.1038/s41467-020-16037-7 (2020).
93. Šmarda, P. *et al.* Genome sizes and genomic guanine + cytosine (GC) contents of the Czech vascular flora with new estimates for 1700 species. *Preslia* **91**, 117–142. https://doi.org/10.23855/preslia.2019.117 (2019).
94. Singh, R., Ming, R. & Yu, Q. Comparative analysis of GC content variations in plant genomes. *Tropical Plant Biol.* **9**, 136–149. https://doi.org/10.1007/s12042-016-9165-4 (2016).
95. Šmarda, P. *et al.* Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc. Natl. Acad. Sci.* **111**, 4096–4102. https://doi.org/10.1073/pnas.1321152111 (2014).
96. Bureš, P. *et al.* Correlation between GC content and genome size in plants. *Cytometry A* **71**, 764 (2007).
97. Grover, C. E. & Wendel, J. F. Recent insights into mechanisms of genome size change in plants. *J. Bot.* **2010**, 382732. https://doi.org/10.1155/2010/382732 (2010).
98. Zuccolo, A. *et al.* Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. *BMC Evol. Biol.* **7**, 152. https://doi.org/10.1186/1471-2148-7-152 (2007).
99. Vogel, J. *et al.* Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**, 763–768. https://doi.org/10.1038/nature08747 (2010).
100. Sasaki, T. The map-based sequence of the rice genome. *Nature* **436**, 793–800. https://doi.org/10.1038/nature03895 (2005).
101. Wu, Z. *et al.* De novo genome assembly of *Oryza granulata* reveals rapid genome expansion and adaptive evolution. *Commun. Biol.* **1**, 84. https://doi.org/10.1038/s42003-018-0089-4 (2018).
102. Liu, Q. *et al.* The repetitive DNA landscape in *Avena* (Poaceae): Chromosome and genome evolution defined by major repeat classes in whole-genome sequence reads. *BMC Plant Biol.* **19**, 226. https://doi.org/10.1186/s12870-019-1769-z (2019).
103. Wicker, T. *et al.* Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biol.* **19**, 103. https://doi.org/10.1186/s13059-018-1479-0 (2018).
104. Tkach, N. *et al.* Molecular phylogenetics and micromorphology of Australasian Stipeae (Poaceae), and the interrelation of whole-genome duplication and evolutionary radiations in this grass tribe. *Front. Plant Sci.* **11**, 630788, https://doi.org/10.3389/fpls.2020.630788 (2021).
105. Romaschenko, K. *et al.* Miocene-Pliocene speciation, introgression, and migration of *Patis* and *Ptilagrostis* (Poaceae: Stipeae). *Mol. Phylogenet. Evol.* **70**, 244–259. https://doi.org/10.1016/j.ympev.2013.09.018 (2014).
106. Matsuoka, Y., Takumi, S. & Nasuda, S. Genetic mechanisms of allopolyploid speciation through hybrid genome doubling: novel insights from wheat (*Triticum* and *Aegilops*) studies. *Int. Rev. Cell Mol. Biol.* **309**, 199–258. https://doi.org/10.1016/b978-0-12-800255-1.00004-1 (2014).
107. Tzvelev, N. N. On the origin and evolution of the feathergrasses (*Stipa* L.). Problems of ecology, geobotany, botanical geography and floristics (eds. Lebedev, D. V. & Karamysheva, Z. V.) 139–150 (Academiya Nauk SSSR, 1977).
108. Clayton, W. D. & Renvoize, S. A. Genera Graminum. *Kew Bull. Additional Ser.* **13**, 1–389 (1986).
109. Hilu, K. W. Phylogenetics and chromosomal evolution in the Poaceae (grasses). *Aust. J. Bot.* **52**, 13–22. https://doi.org/10.1071/BT03103 (2004).
110. Avdulov, N. P. Karyo-systematische Untersuchung der Familie Gramineen. *Bull. Appl. Bot. Genet. Plant Breed.* **43**, 1–352 (1931).
111. VanBuren, R., Wai, C. M., Keilwagen, J. & Pardo, J. A chromosome-scale assembly of the model desiccation tolerant grass *Oropetium thomaeum*. *Plant Direct* **2**, e00096. https://doi.org/10.1002/pld3.96 (2018).
112. Appels, R. *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, eaar7191. https://doi.org/10.1126/science.aar7191 (2018).
113. Liu, W. *et al.* Morphological and genetic variation along a North-to-South transect in *Stipa purpurea*, a dominant grass on the qinghai-tibetan plateau: implications for response to climate change. *PLoS ONE* **11**, e0161972. https://doi.org/10.1371/journal.pone.0161972 (2016).
114. Liu, W., Liao, H., Zhou, Y., Zhao, Y. & Song, Z. Microsatellite primers in *Stipa purpurea* (Poaceae), a dominant species of the steppe on the Qinghai-Tibetan Plateau. *Am. J. Bot.* **98**, e150–e151. https://doi.org/10.3732/ajb.1000444 (2011).
115. Yin, X., Yang, Y. & Yang, Y. Development and characterization of 29 polymorphic EST-SSR markers for *Stipa purpurea* (Poaceae). *Appl. Plant Sci.* **4**, 1600027. https://doi.org/10.3732/apps.1600027 (2016).
116. Klichowska, E., Ślipiko, M., Nobis, M. & Szczecińska, M. Development and characterization of microsatellite markers for endangered species *Stipa pennata* (Poaceae) and their usefulness in intraspecific delimitation. *Mol. Biol. Rep.* **45**, 639–643. https://doi.org/10.1007/s11033-018-4192-x (2018).
117. Ren, J. *et al.* Development and characterization of EST-SSR markers in *Stipa breviflora* (Poaceae). *Applications in Plant Sciences* **5**, 1600157. https://doi.org/10.3732/apps.1600157 (2017).
118. Oyundelger, K. *et al.* Climate and land use affect genetic structure of *Stipa glareosa* P. *A. Smirn. in Mongolia. Flora* **266**, 151572. https://doi.org/10.1016/j.flora.2020.151572 (2020).
119. Zietkiewicz, E., Rafalski, A. & Labuda, D. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics* **20**, 176–183. https://doi.org/10.1006/geno.1994.1151 (1994).

120. Yu, J., Jing, Z. B. & Cheng, J. M. Genetic diversity and population structure of *Stipa bungeana*, an endemic species in Loess Plateau of China, revealed using combined ISSR and SRAP markers. *Genet. Mol. Res.* **13**, 1097–1108. https://doi.org/10.4238/2014.February.20.11 (2014).
121. Kopylov-Guskov, Y. O. & Kramina, T. E. Investigating of *Stipa ucrainica* и *Stipa zalesskii* (Poaceae) from Rostov Oblast using morphological and ISSR analyses. *Bull. Moscow Soc. Nat. Biol. Ser.* **119**, 46–53 (2014).
122. Boussaid, M., Benito, C., Harche, M., Naranjo, T. & Zedek, M. Genetic variation in natural populations of *Stipa tenacissima* from Algeria. *Biochem. Genet.* **48**, 857–872. https://doi.org/10.1007/s10528-010-9367-7 (2010).
123. Nobis, M. *et al.* Hybridisation, introgression events and cryptic speciation in *Stipa* (Poaceae): a case study of the *Stipa heptapotamica* hybrid-complex. *Perspect. Plant Ecol. Evolut. Syst.* **39**, 125457. https://doi.org/10.1016/j.ppees.2019.05.001 (2019).
124. Schubert, M., Marcussen, T., Meseguer, A. S. & Fjellheim, S. The grass subfamily Pooideae: Cretaceous-Palaeocene origin and climate-driven Cenozoic diversification. *Glob. Ecol. Biogeogr.* **28**, 1168–1182. https://doi.org/10.1111/geb.12923 (2019).
125. Hodkinson, T. R. Evolution and taxonomy of the grasses (Poaceae): a model family for the study of species-rich groups. *Annual Plant Rev. Online* **1**, 39. https://doi.org/10.1002/9781119312994.apr0622 (2018).
126. Mueller-Bieniek, A., Kittel, P., Muzolf, B., Cywa, K. & Muzolf, P. Plant macroremains from an early Neolithic site in eastern Kuyavia, central Poland. *Acta Palaeobotanica* **56**, 79–89. https://doi.org/10.1515/acpa-2016-0006 (2016).
127. Brown, R. P. & Yang, Z. Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evol. Biol.* **11**, 271. https://doi.org/10.1186/1471-2148-11-271 (2011).
128. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MH644808.1 (2020).
129. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/MH051716.1 (2020).
130. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/NC_008360.1 (2020).
131. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/NC_022666.1 (2020).
132. Bendich, A. J. Structural analysis of mitochondrial DNA molecules from fungi and plants using moving pictures and pulsed-field gel electrophoresis. *J. Mol. Biol.* **255**, 564–588. https://doi.org/10.1006/jmbi.1996.0048 (1996).
133. Cheng, N. *et al.* Correlation between mtDNA complexity and mtDNA replication mode in developing cotyledon mitochondria during mung bean seed germination. *New Phytol.* **213**, 751–763. https://doi.org/10.1111/nph.14158 (2017).
134. Kozik, A. *et al.* The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genet.* **15**, e1008373. https://doi.org/10.1371/journal.pgen.1008373 (2019).
135. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
136. Hufnagel, D. E., Hufford, M. B. & Seetharam, A. S. SequelTools: a suite of tools for working with PacBio Sequel raw sequence data. *BMC Bioinf.* **21**, 429. https://doi.org/10.1186/s12859-020-03751-8 (2020).
137. Baiakhmetov, E., Nowak, A., Gudkova, P. D. & Nobis, M. Morphological and genome-wide evidence for natural hybridisation within the genus *Stipa* (Poaceae). *Sci. Rep.* **10**, 13803. https://doi.org/10.1038/s41598-020-70582-1 (2020).
138. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinf.* **10**, 421. https://doi.org/10.1186/1471-2105-10-421 (2009).
139. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294. https://doi.org/10.1093/bioinformatics/btv566 (2016).
140. Hunt, M.: Assembly statistics from FASTA and FASTQ files (Version 1.0.1). Github https://github.com/sanger-pathogens/assembly-stats/ (2014).
141. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol. Biol.* **227–245**, 2019. https://doi.org/10.1007/978-1-4939-9173-0_14 (1962).
142. Alonge, M. *et al.* RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224. https://doi.org/10.1186/s13059-019-1829-6 (2019).
143. Flynn, J. M. *et al.* RepeatModeler2: Automated genomic discovery of transposable element families. *PNAS* **117**, 9451–9457. https://doi.org/10.1073/pnas.1921046117 (2020).
144. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276. https://doi.org/10.1101/gr.88502 (2002).
145. Price, A. L., Jones, N. C. & De Pevzner, P. A. novo identification of repeat families in large genomes. *Bioinformatics* **21**, 351–358. https://doi.org/10.1093/bioinformatics/bti1018 (2005).
146. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580. https://doi.org/10.1093/nar/27.2.573 (1999).
147. Wicker, T., Matthews, D. E. & Keller, B. TREP: a database for Triticeae repetitive elements. *Trends Plant Sci.* **7**, 561–562. https://doi.org/10.1016/S1360-1385(02)02372-5 (2002).
148. Smit, A. F. A, Hubley, R. & Green, P. RepeatMasker Open-4.0, http://www.repeatmasker.org, (2020).
149. Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7**, 12. https://doi.org/10.1186/1745-6150-7-12 (2012).
150. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644. https://doi.org/10.1093/bioinformatics/btn013 (2008).
151. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/KY826233 (2020).
152. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/KY826234 (2020).
153. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/KY826235 (2020).
154. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/KY826229 (2020).
155. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/KY826231 (2020).
156. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/KY826232 (2020).
157. *NCBI Sequence Read Archive*, https://www.ncbi.nlm.nih.gov/sra/ERR1744610 (2020).
158. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191 (2018).
159. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
160. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736. https://doi.org/10.1101/gr.215087.116 (2017).
161. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477. https://doi.org/10.1089/cmb.2012.0021 (2012).
162. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/NC_016135.3?report=fasta&from=164020&to=167409 (2020).
163. *NCBI Nucleotide*, https://www.ncbi.nlm.nih.gov/nuccore/KM036284.1 (2020).
164. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. https://doi.org/10.1093/molbev/mst010 (2013).
165. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531 (2014).
166. Bouckaert, R. *et al.* BEAST 25: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, 1006650. https://doi.org/10.1371/journal.pcbi.1006650 (2019).

167. Bouckaert, R. & Drummond, A. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* **17**, 42. https://doi.org/10.1186/s12862-017-0890-6 (2017).
168. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using tracer 17. *System. Biol.* **67**, 901–904. https://doi.org/10.1093/sysbio/syy032 (2018).
169. Rambaut, A. Figtree v1.4.4 https://tree.bio.ed.ac.uk/software/figtree (2018).
170. Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294. https://doi.org/10.1186/s13059-015-0849-0 (2015).
171. Kilian, A. *et al.* Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol. Biol.* **888**, 67–89. https://doi.org/10.1007/978-1-61779-870-2_5 (2012).
172. *NCBI Sequence Read Archive*, https://www.ncbi.nlm.nih.gov/sra/PRJNA726584 (2021).

Acknowledgements

We would like to express our gratitude to Eric Johnson from SNPsaurus, Artem Kasianov from Institute for Information Transmission Problems of the Russian Academy of Sciences (Moscow, Russia) and Igor A. Shmakov from Altai State University (Barnaul, Russia) for their valuable assistance in the genome assembling. We also thank the iDiv High-Performance Computing cluster for providing computing resources for this paper. Finally, we thank two anonymous reviewers for providing valuable comments on the manuscript. The study was supported by the Russian Science Foundation (grant no.19-74-10067). E.B. was supported via the RSF (grant no.19-74-10067) and a DS grant of the Jagiellonian University (DS/D/WB/IB/2/2019). M.N. was supported by the National Science Centre, Poland (grant no. 2018/29/B/NZ9/00313). P.D.G. was supported by the RSF (grant no.19-74-10067). The open-access publication of this article was funded by the BioS Priority Research Area under the program "Excellence Initiative – Research University" at the Jagiellonian University in Krakow.

Author contributions

E.B., P.D.G., M.N. planned the study. E.B. supervised the research. M.N. and P.D.G. identified and collected biological samples. E.B., C.G., E.S. performed the nuclear genome assembly. E.B. performed the remaining bioinformatic analyses and wrote the manuscript. All authors revised the draft, provided comments and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-94068-w.

Correspondence and requests for materials should be addressed to E.B. or M.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.