scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-scale genome assembly and *de novo* annotation of *Alopecurus aequalis*

Jonathan Wright[1], Kendall Baker[1], Tom Barker[1], Leah Catchpole[1], Alex Durrant[1], Fiona Fraser[1], Karim Gharbi[1], Christian Harrison[2,3], Suzanne Henderson[1], Naomi Irish[1], Gemy Kaithakottil[1], Ilia J. Leitch[4], Jun Li[5], Sacha Lucchini[1], Paul Neve[2,6], Robyn Powell[4], Hannah Rees[1,7], David Swarbreck[1], Chris Watkins[1], Jonathan Wood[8], Seanna McTaggart[1], Anthony Hall[1,9] & Dana MacGregor[2] ✉

Alopecurus aequalis **is a winter annual or short-lived perennial bunchgrass which has in recent years emerged as the dominant agricultural weed of barley and wheat in certain regions of China and Japan, causing significant yield losses. Its robust tillering capacity and high fecundity, combined with the development of both target and non-target-site resistance to herbicides means it is a formidable challenge to food security. Here we report on a chromosome-scale assembly of *A. aequalis* with a genome size of 2.83 Gb. The genome contained 33,758 high-confidence protein-coding genes with functional annotation. Comparative genomics revealed that the genome structure of *A. aequalis* is more similar to *Hordeum vulgare* rather than the more closely related *Alopecurus myosuroides*.**

Background & Summary

Alopecurus aequalis, commonly known as shortawn foxtail or orange foxtail, is a winter annual or short-lived perennial bunchgrass of the Poaceae family. It is native to at least 55 different countries across the Northern Hemisphere and northern Southern Hemisphere and has been introduced into Australasian regions[1]. *A. aequalis* has emerged as the dominant agricultural weed of winter canola, barley, and wheat only in certain regions of China and Japan despite its widespread distribution[2]. *A. aequalis* can cause significant yield losses; densities of up to 1,560 plants per m^2 reduce wheat yields by up to 50%[3]. The biology of *A. aequalis*, particularly its robust tillering capacity and high fecundity (with a single plant able to produce over 7,300 highly dispersible seeds), makes it a challenging weed to control[3]. Moreover, the evolution of both target-site resistance (TSR[4,5]) and non-target-site resistance (NTSR[5–7]) means many of the available chemical control methods are ineffective. Therefore, *A. aequalis* is a formidable challenge to food security and novel and innovative control methods are urgently required.

Another Poaceae grass, *Alopecurus myosuroides* (black-grass), has evolved to occupy similar agroecosystem niches. Like *A. aequalis*, *A. myosuroides* is a weed of winter cereals in China and Japan[8,9] and surveys have recorded it as present across the Northern Hemisphere[1]. However, *A. myosuroides* has become the predominant agricultural weed in Western European winter wheat and barley, leading to considerable yield losses and economic consequences[10]. These two species have similar but distinct morphologies and growth habits (Fig. 1). Like *A. aequalis*, black-grass exhibits widespread multiple-herbicide resistance[10–12] and the characterized resistance mechanisms are similar between the two species. Both have TSR mutations that alter equivalent amino acids of homologous herbicide target genes[4] and NTSR correlated with increased xenobiotic-metabolizing enzymes such as cytochrome P450 mono-oxygenases and glutathione s-transferases[6,7]. In black-grass, NTSR is highly heritable with no evidence that it results in a fitness penalty[13], and it is correlated with increased tolerance to drought and

[1]Earlham Institute, Norwich Research Park, Norwich, UK. [2]Rothamsted Research, Protecting Crops and the Environment, Harpenden, UK. [3]University College London, Rayne Building, University Street, London, UK. [4]Royal Botanic Gardens, Kew, Richmond, Surrey, UK. [5]College of Plant Protection, Nanjing Agricultural University, Nanjing, China. [6]University of Copenhagen, Plant and Environmental Sciences, Taastrup, Denmark. [7]Institute of Biological, Environmental & Rural Sciences, Aberystwyth University, Aberystwyth, Wales, UK. [8]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [9]School of Biological Sciences, University of East Anglia, Norwich, UK. ✉e-mail: dana.macgregor@rothamsted.ac.uk

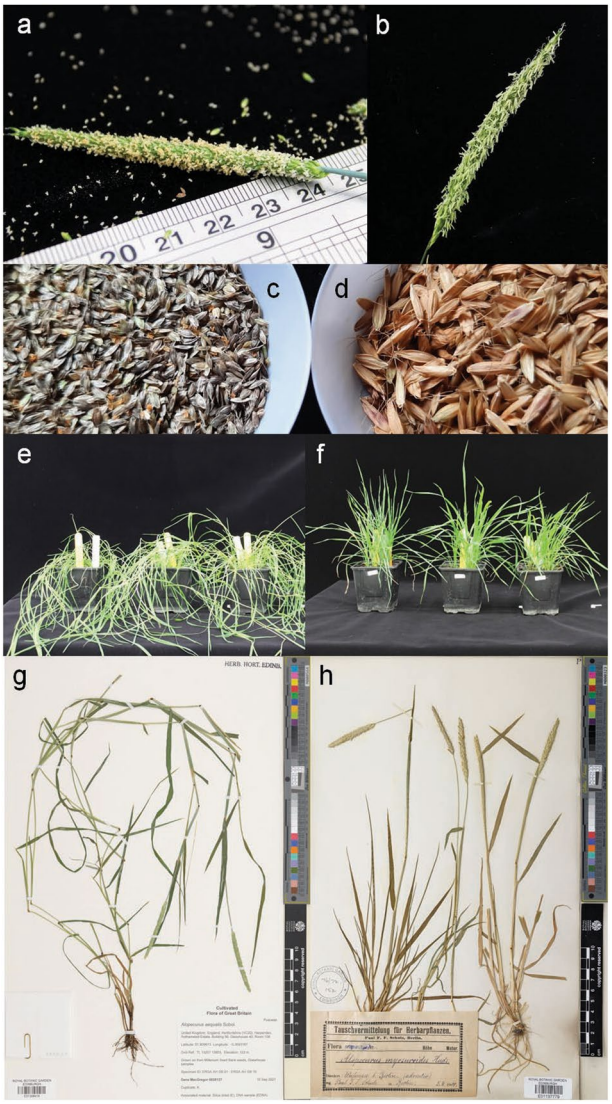

Fig. 1 Images highlighting differences in *Alopecurus aequalis* (shortawn foxtail or orange foxtail; a, c, e, & g) and *Alopecurus myosuroides* (black-grass; b, d, f, & h) morphologies. Images show flowering heads (**a,b**), seeds (**c,d**), at vegetative growth stage (**e,f**) and Kew Herbarium images (**g,h**). The flower spike of *A. aequalis* (**a**) has a blunt end rather than the tapered end in *A. myosuroides* (**b**). While both have single flowered spikelets, the anthers of *A. aequalis* are shorter compared to *A. myosuroides* (**b**). The mature seeds of these two species are easily distinguished (c: *A. aequalis*, d: *A. myosuroides*). *A. aequalis* has a more prostrate growth (**e**) while *A. myosuroides* is more upright (**f**). These differences can be seen in Herbarium images from Royal Botanic Garden Edinburgh (https://data.rbge.org.uk/search/herbarium/) for *A. aequalis* (Barcode: E01358418, g), and *A. myosuroides* (Barcode: E01137779, h).

waterlogging stresses[14,15]. There is evidence that some TSR mutations are associated with fitness costs[16]. These characteristics, combined with an ability to compete with crops for essential resources like nutrients, water, and light, mean that when either foxtail species are present in agricultural fields, they significantly reduce crop yields and overall productivity within agroecosystems[3,10,11,14,15].

Despite geographic isolation and 7.4 million years of divergence[17], these two species have evolved similar herbicide resistance mechanisms and have become problematic in similar winter crops. It is not yet understood whether similarities between these two species are the result of parallel evolution. This lack of direct comparison is in part due to lack of genomic data for either species. Recently, two reference genomes have been produced for biotypes of *A. myosuroides* that are sensitive to all tested herbicides[18,19]. We therefore set out to generate a genome of similar quality for *A. aequalis* as part of the European Reference Genome Atlas[20] (ERGA) pilot programme, which aims to empower research communities to expand the taxonomic coverage of genomic resources to address continent-scale questions at the genomic level.

Here we report a *de novo* annotated, chromosome-level assembly of *A. aequalis*. PacBio HiFi reads (32.9x coverage) were used to assemble the genome resulting in a contig assembly of 2.83 Gb with a contig-N50 of 374.7 Mb. The assembled size was identical to the estimated genome size from *k-mer* based methods. Omni-C

reads (56.7x coverage) were used to anchor and orient the contigs into seven pseudomolecules. Protein-coding genes were annotated using REAT[21] an evidence-guided pipeline, making use of RNA-Seq alignments, transcript assemblies from Iso-Seq reads and alignment of protein sequences. In total, 33,758 high-confidence protein-coding genes were identified. Whole genome alignment between *A. aequalis*, *A. myosuroides* and *Hordeum vulgare* indicated that the genome structure of *A. aequalis* was more similar to *H. vulgare* than to the more closely related *A. myosuroides*. This genomic resource provides a much-needed foundation for investigating the molecular mechanisms underlying weedy traits, such as widespread multiple-herbicide resistance, in *Alopecurus aequalis*, and will be an invaluable tool for the research community in devising more effective weed management strategies.

Methods

Alopecurus aequalis plants and materials. Seeds of *Alopecurus aequalis* (orange foxtail), donated by a private Individual in 2014 to the Royal Botanic Gardens, Kew Millennium Seed Bank (Serial Number 828127), were used for genome sequencing and annotation. While the collection location was not recorded, *A. aequalis* is not an agricultural weed in the UK, making it unlikely that it would have been exposed to herbicide(s) selection. To establish a seed stock, 26 plants were grown and allowed to intrabreed in isolation for one generation at Rothamsted Research. For genome size estimation by flow cytometry, fresh leaf material from four individual plants (ID 828127 A, B, C and D) of this second generation were analysed using the fluorochrome propidium iodide with the 'one-step method'[22]. Nuclei were isolated in the LB01 nuclei isolation buffer[23] and *Petroselinum crispum* 'Champion Moss Curled' was used as the internal calibration standard with an assumed 2C-value of 4.5 pg[24]. The mean relative fluorescence of nuclei from *A. aequalis* and *P. crispum* were used to estimate the genome size of *A. aequalis* using the following equation: 2C-value of *A. aequalis* (pg) = (Mean peak position of *A. aequalis*/ mean peak position of *P. crispum*) × 4.5.

To generate sufficient material for genome sequencing, a single plant (ID 828217 A) was grown and vegetatively cloned twice following protocols described in the supplementary material in Cai *et al.*[18]. DNA for genome sequencing was isolated using the protocols described below from young leaves and meristem material from plants that had been dark-adapted for five days, flash frozen in liquid nitrogen, and stored at −80 °C until shipping on dry ice. Similarly, RNA for Iso-Seq and RNA sequencing for annotation was sourced from flag leaves and flowering heads from clones of plants 828217 A and 828217B. Additional RNA came from bulked shoot or root material derived from five technical replicates of 5–8 plants from the 828217 seed stock. These plants were grown under sterile hydroponic conditions, as per supplementary material of Cai *et al.* [18], for a total of 42 days before separating root or shoot material from the seed. The flash frozen material was stored at −80 °C until it was shipped on dry ice for processing. One clone from 828217 A was allowed to flower and preserved by preparing a herbarium voucher which is stored at the Royal Botanic Garden Edinburgh (RGBE) (Fig. 1g, https://data.rbge.org.uk/herb/E01358418).

DNA extraction. HMW DNA extraction was performed using the Nucleon PhytoPure kit, with a slightly modified version of the manufacturer's protocol. One gram of snap-frozen leaf material was ground under liquid nitrogen for 10 minutes. The tissue powder was thoroughly resuspended in Reagent 1 using a 10 mm bacterial spreader loop until the mixture appeared completely homogeneous, at which point 4 µl of 100 mg/ml RNase A (Qiagen cat no. 19101) was added and mixed again. After incubation on ice, 200 µl of resin was added along with chloroform. The chloroform extraction was followed by a phenol:chloroform extraction. Here, an equal volume of 25:24:1 phenol:chloroform:isoamyl alcohol was added to the previous upper phase, mixed gently at 4 °C for 10 minutes and then centrifuged at 3000 g for 10 minutes. The upper phase from this procedure was then transferred to another 15 ml Falcon tube and precipitation proceeded as recommended by the manufacturer's protocol. The final elution was left open in a chemical safety cabinet for two hours to allow residual phenol and ethanol to evaporate, and the DNA sample was left at room temperature overnight before further processing.

PacBio HiFi library preparation and sequencing. In total, 65 µg of genomic DNA was split into 5 aliquots and manually sheared with the Megaruptor 3 instrument (Diagenode, P/N B06010003), according to the Megaruptor 3 operations manual, loading each aliquot of 150 µl at 21 ng/µl, with a shear speed of 31. Each sheared aliquot underwent AMPure® PB bead (PacBio®, P/N 100-265-900) purification and concentration before undergoing library preparation using the SMRTbell® Express Template Prep Kit 2.0 (PacBio®, P/N 100-983-900). The HiFi libraries were prepared according to the HiFi protocol version 03 (PacBio®, P/N 101-853-100) and the final libraries were size fractionated using the SageELF® system (Sage Science®, P/N ELF0001), 0.75% cassette (Sage Science®, P/N ELD7510), running on a 4.55 hr program with 30 µl of elution buffer per well post elution. The libraries were quantified by fluorescence (Invitrogen Qubit™ 3.0, P/N Q33216) and the size of the library fractions were estimated from a smear analysis performed on the FEMTO Pulse® System (Agilent, P/N M5330AA).

The loading calculations for sequencing were completed using the PacBio® SMRT®Link Binding Calculator 10.2. Sequencing primer v2 was annealed to the adapter sequence of the HiFi libraries. The libraries were bound to the sequencing polymerase with the Sequel® II Binding Kit v2.0. Calculations for primer and polymerase binding ratios were kept at default values for the library type. Sequel® II DNA internal control 1.0 was spiked into each library at the standard concentration prior to sequencing. The sequencing chemistry used was Sequel® II Sequencing Plate 2.0 (PacBio®, P/N 101-820-200) and the Instrument Control Software v 10.1.

The libraries were sequenced on the Sequel IIe across 15 Sequel II SMRT®cells 8 M. The parameters for sequencing per SMRT cell were diffusion loading, 30-hour movie, 2-hour immobilisation time, 4-hour pre-extension time, 60–80 pM on plate loading concentration. We generated 11.7 million PacBio HiFi reads (227 Gb of sequence), corresponding to a haploid genome coverage of 32.9x. The average HiFi read length was 19.4 Kbp.

Dovetail Omni-C library preparation and sequencing. Sample material for the Omni-C library preparation was 300 mg of the youngest leaf and meristem material from plants dark adapted for 5 days then flash frozen at harvest. The Omni-C library was prepared using the Dovetail® Omni-C® Kit (SKU: 21005) according to the manufacturer's protocol[25].

Briefly, the chromatin was fixed with disuccinimidyl glutarate (DSG) and formaldehyde in the nucleus. The cross-linked chromatin sample was then digested *in situ* with 0.05 µl DNAse I. Following digestion, the cells were lysed with SDS to extract the chromatin fragments and the chromatin fragments were bound to Chromatin Capture Beads. Next, the chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, the crosslinks were reversed, the associated proteins were degraded, and the DNA was purified before conversion into a sequencing library (NEBNext® Ultra™ II DNA Library Prep Kit for Illumina® (E7645)) using Illumina-compatible adaptors (NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1) (E7335)). Biotin-containing fragments were isolated using streptavidin beads prior to PCR amplification. The Omni-C library was quantified by qPCR using a Kapa Library Quantification Kit (Roche Diagnostics 7960204001). The pool was diluted to 2 nM and denatured using 2 N NaOH before diluting to 20pM with Illumina HT1 buffer. The denatured pool was loaded on an Illumina MiSeq Sequencer for quality control with a 300 cycle MiSeq Reagent Kit v2 (Illumina MS-102-2002) at 10pM concentration with a 1% phiX control v3 spike (Illumina FC-110-3001). The MiSeq was run using control software version 4.0, RTA v1.18.54.430, and the data was demultiplexed and converted to fastq using bcl2fastq2. Following quality control analysis, the library was sequenced on one lane of a 300 cycle NovaSeq X Series 10B Reagent Kit (Illumina 20085594). For this run, the library was diluted down to 0.75 nM using EB (10 mM Tris pH8.0) in a volume of 40 µl before spiking in 1% Illumina phiX Control v3. This was denatured by adding 10 µl 0.2 N NaOH and incubating at room temperature for 5 mins, after which it was neutralised by adding 150 µl of Illumina's preload buffer, of which 160 µl was loaded onto the NovaSeq X Plus for sequencing. The NovaSeq X Plus was run using control software version 1.0.1.7385 and was set up to sequence 150 bp paired-end reads. The data was demultiplexed and converted to fastq using bcl2fastq2. We generated 1.28 billion reads representing 56.7x haploid genome coverage.

Illumina RNA-Seq library preparation and sequencing. Five root samples and five leaf samples were used to generate RNA-Seq libraries. This was performed on the Perkin Elmer (formerly Caliper LS) Sciclone G3 (PerkinElmer PN: CLS145321) using the NEBNext Ultra II RNA Library prep for Illumina kit (NEB#E7760L) NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB#E7490L) and NEBNext Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs) (E6440S/L) at a concentration of 10 µM. One µg of RNA was purified to extract mRNA with a Poly(A) mRNA Magnetic Isolation Module. Isolated mRNA was then fragmented for 12 minutes at 94 °C, and converted to cDNA. NEBNext Adaptors were ligated to end-repaired, dA-tailed DNA. The ligated products were subjected to a bead-based purification using Beckman Coulter AMPure XP beads (A63882) to remove unligated adaptors. Adaptor Ligated DNA was then enriched by 10 cycles of PCR (30 secs at 98 °C, 10 cycles of: 10 secs at 98 °C _75 secs at 65 °C _5 mins at 65 °C, final hold at 4 °C). The size of the resulting libraries was determined using a Perkin Elmer DNA High Sensitivity Reagent Kit (CLS760672) with DNA 1 K/12 K/HiSensitivity Assay LabChip (760517) and the concentration measured with a Quant-iT™ dsDNA Assay Kit, high sensitivity (Plate Reader) assay from ThermoFisher (Q-33120). The final libraries were pooled equimolarly and quantified by qPCR using a Kapa Library Quantification Kit (Roche Diagnostics 7960204001).

The pool was diluted down to 0.5 nM using EB (10 mM Tris pH8.0) in a volume of 18 µl before spiking in 1% Illumina phiX Control v3. This was denatured by adding 4 µl 0.2 N NaOH and incubating at room temperature for 8 mins, after which it was neutralised by adding 5 µl 400 mM tris pH 8.0. A master mix of DPX1, DPX2, and DPX3 from Illumina's Xp 2-lane kit was made and 63 µl added to the denatured pool leaving 90 µl at a concentration of 100pM. This was loaded onto a single lane of the NovaSeq SP flow cell using the NovaSeq Xp Flow Cell Dock before loading onto the NovaSeq. 6000. The NovaSeq was run using NVCS v1.7.5 and RTA v3.4.4 and was set up to sequence 150 bp PE reads. The data was demultiplexed and converted to fastq using bcl2fastq2. A total of 253 million and 234 million reads were generated for root and shoot samples respectively.

PacBio Iso-Seq library preparation and sequencing. Iso-Seq libraries were generated for root and shoot samples. The five shoot extractions were pooled for the shoot library and a single root sample was used for root. The libraries were constructed starting from 356–482 ng of total RNA per sample. Reverse transcription cDNA synthesis was performed using NEBNext® Single Cell/Low Input cDNA Synthesis & Amplification Module (NEB, E6421). Each cDNA sample was amplified with barcoded primers for a total of 12 cycles. Each library was prepared according to the guidelines laid out in the Iso-Seq protocol version 02 (PacBio, 101-763-800) using SMRTbell express template prep kit 2.0 (PacBio, 102-088-900). The library pool was quantified using a Qubit Fluorometer 3.0 (Invitrogen) and sized using the Bioanalyzer HS DNA chip (Agilent Technologies, Inc.).

The loading calculations for each Iso-Seq library were calculated using the PacBio SMRTlink Binding Calculator v.10.2.0.133424 and prepared for sequencing according to the library type. Sequencing primer v4 was annealed to the Iso-Seq library pool and complexed to the sequencing polymerase with the Sequel II binding kit v2.1 (PacBio, 101-843-000). Calculations for primer to template and polymerase to template binding ratios were kept at default values for the library type. Sequencing internal control complex 1.0 (PacBio, 101-717-600) was spiked into the final complex preparation at a standard concentration before sequencing for all preparations. The sequencing chemistry used was Sequel® II Sequencing Plate 2.0 (PacBio®, 101-820-200) and the Instrument Control Software v10.1.0.125432.

Each Iso-Seq library was sequenced on the Sequel IIe instrument with one Sequel II SMRT®cell 8 M cell per library. The parameters for sequencing per SMRTcell were diffusion loading, 30-hour movie, 2-hour immobilisation time, 2-hour pre-extension time, 60pM on plate loading concentration.

Sample	Mean peak position of *Alopecurus aequalis*	Mean peak position of *Petroselinum crispum* (internal standard)	Genome size (2C-value; pg)	Genome size (1C-value; Gb)*	CV of sample peak (%)	CV of calibration standard peak (%)	SD(R) sample
828217 A	245.16	156.12	7.07	3.45	3.58	3.36	0.030
828217B	221.8	140.67	7.10	3.47	4.53	4.61	0.050
828217 C	238.42	151.45	7.08	3.46	4.35	3.95	0.005
828217D	250.47	161.25	6.99	3.42	3.49	3.91	0.003

Table 1. Genome size estimation by flow cytometry for four individual plants of *Alopecurus aequalis* line 828127. (*Flow cytometry estimates in pg were converted to Gb using the conversion factor 1 pg = 0.978 Gb[69]). "CV" is the coefficient of variation for the peaks in our flow cytometry data which are an indication of the reliability of the genome size estimate and "SD(R)" is the standard deviation of the genome size estimate.

We generated 3.9 million CSS reads for the root sample and 3.8 million CCS reads for the shoot sample. These were filtered to identify Full-Length Non-Concatamer (FLNC) reads resulting in 3.6 and 3.4 million FLNC reads for root and shoot, respectively. Clustering these individually generated 258,543 transcripts for root and 212,332 transcripts for shoot. In addition, root and shoot FLNC reads were combined and clustered to generate a set of 410,286 transcripts.

Genome survey. A previous study reported that *A. aequalis* has seven chromosomes[26]. Prior to genome sequencing, we first estimated genome size using flow cytometry for four individual plants of *A. aequalis* line 828217 which gave a mean estimated genome size of 3.45 Gb/1 C (Table 1).

We used FastK[27] (v1.1) to count k-mers (k = 31) in the HiFi reads, then GenomeScope.FK (based on GenomeScope 2.0[28]) to explore genome characteristics. In order to estimate genome size, the haploid k-mer coverage was set to 33.

```
FastK -k31 -T16 -M200 -P. -Ntmp_fastk reads.fastq.gz
Histex -G reads.hist>hist_all.out
GeneScopeFK.R -i hist_all.out -o all_out -k 31 --kmercov 33
```

The estimated genome size from GenomeScope was 2.9 Gb (Fig. 2a) using k = 31, which is 0.55 Gb smaller than the flow cytometry estimate of 3.45 Gb. However, genome size estimates using k-mers are often smaller than those obtained by flow cytometry due to the challenges of assembling the repetitive fraction of the genome, and can also be significantly affected by polyploidy and heterozygosity. Computational methods try to avoid overestimating genome size by ignoring high frequency k-mers which come from a mix of real repeats and artifacts such as organelles or small contaminants. The choice of k-mer size can also affect k-mer based genome size estimation but in this case, increasing the k-mer size to 61 made little difference to the estimated genome size (2.911 Gb compared to 2.908 Gb). GenomeScope estimated the heterozygous percentage of the genome as 0.058% and the k-mer profile shows a large single peak centered around coverage 73 representing the homozygous part of the genome. There is a slight deviation from the model on the left side of the peak (marked with a red arrow) representing the heterozygous part of the genome. This indicates the genome is less heterozygous than expected for an outcrossing species.

Genome assembly. Hifiasm[29] v0.16.1 was used to generate a contig assembly and the homozygous coverage (--hom_cov parameter) was set to 73, corresponding to the position of the homozygous peak in the GenomeScope plot generated from the reads (Fig. 2a). Hifiasm generates assembly graphs for primary contigs and haplotypes 1 and 2.

```
hifiasm -o fx_run1.asm -t 64 --hom-cov 73 0.5_hifi_reads.fastq
```

Reads assembled into 1,034 primary contigs with a total size of 2.9 Gb and a contig-N50 of 374.7 Mb (Table 2). The assembly size was similar to the k-mer based estimate and smaller than the flow-cytometry estimate. The total size of the individual haplotypes was similar to the primary contigs but the contiguity was lower (348 Mb and 244 Mb for haplotypes 1 and 2 respectively).

The contiguity of this contig assembly was exceptionally high with the 13 longest contigs comprising 98% of the assembly. To determine whether these contigs represented whole chromosome arms we aligned contigs to the assembly of *A. myosuroides*[18] and *H. vulgare*[30] 'Morex' V3. Ten contigs were identified that represented chromosome or chromosome-arm level sequences. We also found two contigs with telomere sequences (TTTAGGG) on both ends and seven with telomere sequences on one end indicating that we had assembled chromosome or chromosome-arm level sequences. The low heterozygosity in *A. aequalis* estimated by GenomeScope was confirmed by identifying single-copy k-mers from a comparison of both haplotypes to the reads using KAT[31] (Fig. 2b).

Contig scaffolding was performed using the Illumina paired-end Omni-C reads. The Arima Genomics Hi-C mapping pipeline[32] was used to map the reads to the contigs before scaffolding with YaHS[33]. The mapping pipeline uses BWA[34] (v.0.7.12) to map reads 1 and 2 separately before combining them into a single BAM file which is then deduplicated using Picardtools[35] MarkDuplicates (v.2.25.7). The deduplicated BAM file was used as input

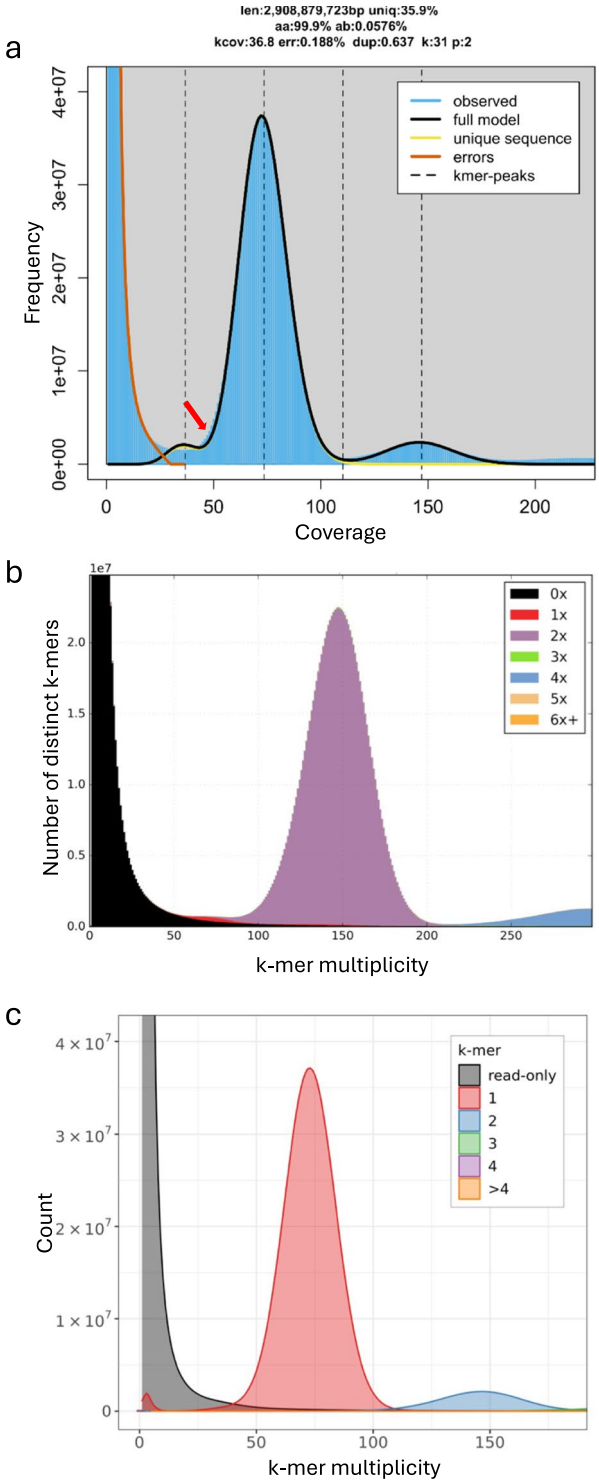

Fig. 2 K-mer spectra used to estimate genome size and to assess assembly quality of *Alopecurus aequalis*. (**a**) GenomeScope.FK profile of *Alopecurus aequalis* HiFi reads. K-mer count distribution (in blue) shows a single homozygous peak at a coverage of 73, and a peak representing repeats at approximate coverage 150. The summary above the plot shows the estimated genome size in bp (len), unique fraction (uniq), homozygous fraction (aa), heterozygous fraction (ab), k-mer coverage of the heterozygous peak (kcov), PCR error estimation (err), PCR duplication estimation (dup), k-mer length (k), and ploidy (p). The red arrow indicates where the observed deviates from the model representing the heterozygous part of the genome. (**b**) K-mer copy-number spectra comparing the combined haplotypes of *Alopecurus aequalis* to the reads. The majority of the kmers are present twice (once in each haplotype: purple peak), the red single-copy region represents kmers that exist in a single haplotype (**c**) K-mer copy-number spectra comparing k-mers from reads to k-mers in the primary contig assembly of *Alopecurus aequalis*.

	Primary contigs	Haplotype 1	Haplotype 2
Total bases (bp)	2,895,609,631	2,856,592,317	2,873,613,452
GC content (%)	45.57	45.55	45.58
Contig number	1,034	1,067	943
Longest contig (bp)	476,254,116	400,977,190	373,165,844
Contig N50 (bp)	374,747,030	347,700,133	244,747,959
Contig L50	4	4	5
Contig N90 (bp)	57,656,729	27,353,843	41,286,451
Contig L90	9	16	15
k-mer completeness (%)	98.0	97.7	97.9
BUSCO	C:93.9% [S:89.9%, D:4.0%], F:1.6%, M:4.5%, n:4896	C:94.0% [S:89.9%, D:4.1%], F:1.5%, M:4.5%, n:4896	C:94.1% [S:89.8%, D:4.3%], F:1.5%, M:4.4%, n:4896

Table 2. Contiguity and completeness statistics from the contig assembly of *Alopecurus aequalis* showing primary contigs and the two assembled haplotypes. BUSCO abbreviations are C: Complete, S: Single copy, D: Duplicated, F: Fragmented, M: Missing, n: total BUSCO genes.

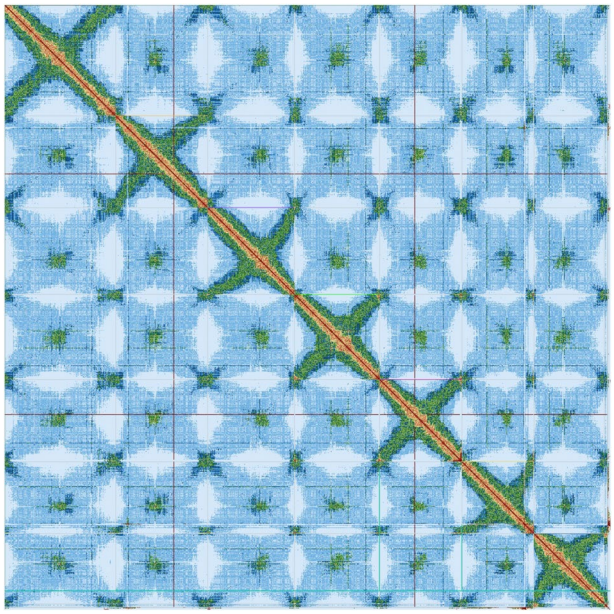

Fig. 3 The PretextView contact map for *Alopecurus aequalis* after scaffolding and curation. Seven chromosomes in descending size order are represented by the diagonal red line with centromeric repeats showing as green regions. This map was generated after removal of one scaffold identified as a haplotypic duplication and the merging of scaffold_7 and scaffold_8.

to YaHS. This generated seven chromosome-level scaffolds, in agreement with previous cytogenetic studies[26], and 421 sequences not incorporated into scaffolds.

To validate the scaffolding, Omni-C reads were mapped back to the scaffolds using BWA and the resulting contact map visualised using PretextView[36]. This identified one scaffold as retained haplotypic duplication and indicated a potential join between two larger scaffolds (scaffold_7 and scaffold_8). It also highlighted that many of the smaller scaffolds represented contamination. We removed the scaffold representing haplotypic duplication and joined scaffolds 7 and 8 into a larger scaffold. The final contact map is shown in Fig. 3.

Contaminated short sequences were identified and removed using NCBI BLAST+[37] megablast (v2.12.0). The final assembly contained seven chromosome-level sequences and 145 shorter sequences not included in scaffolds. The chromosome-level scaffolds represented 2.8 Gb of sequence (99.5% of the assembly) and the total size of the 145 unassigned scaffolds was 15 Mb. Chromosome-level scaffolds were labeled in descending size order (Table 3). The contiguity of the contigs was such that only 11 joins were made between contigs and no contigs were broken. Two contigs generated chromosomes without any scaffolding (chr_4 and chr_6).

Genome annotation. Annotation was performed on contigs due to the high contiguity of the assembly at this stage. Rather than reannotate the scaffolds, the annotation was lifted over from contigs as scaffolding made very few contig joins and no contig breaks.

Scaffold name	Length (bp)
chr_1	521,103,429
chr_2	432,403,959
chr_3	408,022,704
chr_4	400,977,190
chr_5	386,775,718
chr_6	345,309,186
chr_7	339,394,970
Total	2,833,987,156

Table 3. Final scaffold lengths and total assembly size for *Alopecurus aequalis*.

Organism Scientific Name	Assembly Accession
Sorghum bicolor	GCF_000003195.3
Brachypodium distachyon	GCF_000005505.3
Setaria italica	GCF_000263155.2
Oryza sativa	GCF_001433935.1
Lolium rigidum	GCF_022539505.1
Panicum hallii	GCF_002211085.1
Lolium perenne	GCF_019359855.1
Panicum virgatum	GCF_016808335.1
Zea mays	GCF_902167145.1
Hordeum vulgare	GCF_904849725.1
Triticum aestivum	iwgsc_refseqv2.1 (HC genes)

Table 4. List of species used for cross species protein alignment.

Repeat identification. Repeat annotation was conducted using the EI-Repeat pipeline[38] (v1.3.4), which incorporates third-party tools for repeat detection. De novo identification of repetitive elements in the assembled *A. aequalis* genome was carried out using RepeatModeler[39] (v1.0.11) with a customized repeat library. Unclassified repeats were compared against a custom BLAST database containing organellar genomes (mitochondrial and plastid sequences from Pooideae in the NCBI nucleotide division), and repeat families matching organellar DNA were hard-masked. RepeatMasker[40] (v4.0.7) was employed to identify repeats, utilizing both a RepBase[41] embryophyte library and the customized RepeatModeler library. Overall, 79% of the assembly was identified as repetitive and masked.

Reference guided transcriptome reconstruction. Gene models were constructed using RNA-Seq reads, Iso-Seq transcripts (HQ + LQ), and FLNC reads, leveraging the REAT transcriptome workflow as outlined in Grewal *et al.*[42]. This process utilised several tools, including minimap2[43] (v2.18-r1015) for alignment, Portcullis[44] (v1.2.4) for splice junction refinement, StringTie2[45] (v2.1.5) and Scallop[46] (v0.10.5) for transcript assembly, and Mikado[47] for gene model integration and refinement.

Cross-species protein alignment. Protein sequences from 11 Poaceae species (Table 4) were aligned to the *A. aequalis* assembly using the REAT Homology workflow[21] which aligns proteins with spaln[48] (v2.4.7) and miniprot[49] (v0.3). The aligned proteins from both methods were clustered into loci and a consolidated set of gene models were derived via Mikado[47].

Evidence guided gene prediction. Protein-coding genes were annotated using the evidence-guided REAT prediction workflow, which incorporated repeat annotations, RNA-Seq alignments, transcript assemblies, protein sequence alignments and gene prediction with AUGUSTUS[50] and EVidenceModeler[51]. For details refer to Grewal *et al.*[42], with the modification that REAT transcriptome and homology Mikado models were categorised based on alignments to uniprot magnoliopsida proteins.

Projection of gene models from *A. myosuroides*. *A. myosuroides* gene models[18] were projected onto the *A. aequalis* assembly with Liftoff-1.5.1[52], only those models that transferred fully with no loss of bases and identical exon/intron structure were retained (ei-liftover pipeline[53]).

Gene model consolidation. The final set of gene models was selected using Minos[54], a pipeline that integrates metrics from protein, transcript, and expression datasets to produce a unified and optimized set of gene models. The final consolidated gene models were derived by integrating the following sources of evidence-guided annotations and predictions:

1. Augustus Gene Builds: Three alternative evidence-guided builds generated from the REAT prediction workflow.

Biotype	Confidence	Genes	Transcripts
protein_coding_gene	High	35,149	53,355
protein_coding_gene	Low	20,304	21,383
transposable_element_gene	Low	3,617	3,691
transposable_element_gene	High	3,467	3,600
predicted_gene	Low	2,337	2,366
ncrna_gene	Low	575	1,313
Total		65,449	85,708

Table 5. Minos classified gene models on the contig assembly of *Alopecurus aequalis*.

	HC	LC	HC + LC
Number of genes	33,758	19,331	53,089
Number of transcripts	51,946	20,409	72,355
Transcripts per gene	1.54	1.06	1.36
Number of monoexonic genes	7,933	6,035	13,968
Number of monoexonic transcripts	8,423	6,106	14,529
Transcript mean size cDNA (bp)	2,151.48	1,262.47	1,900.72
Transcript median size cDNA (bp)	1,906.00	1,103.00	1,653.00
Min cDNA length (bp)	156	159	156
Max cDNA length (bp)	17,668	8,838	17,668
Total exons	328,060	61,029	389,089
Mean number of exons per transcript	6.32	2.99	5.38
Exon mean size (bp)	340.67	422.19	353.46
CDS mean size (bp)	249.71	326.41	261.35
Transcript mean size CDS (bp)	1473.53	876.93	1,305.25
Transcript median size CDS (bp)	1,248	690.00	1,095.00
Min CDS length (bp)	156	138	138
Max CDS length (bp)	16,185	8,337	16,185
Intron mean size (bp)	448.30	968.72	515.04
5′ UTR mean size (bp)	262.00	141.15	227.91
3′ UTR mean size (bp)	415.94	224.39	367.55

Table 6. Genome annotation statistics for *Alopecurus aequalis*. High-confidence protein coding genes (HC), low-confidence protein coding genes (LC) and both sets combined (HC + LC).

2. EVM-Mikado Gene Models: Outputs from the REAT prediction workflow integrated through EVM and Mikado.
3. Transcriptome-Derived Gene Models: Models constructed based on the REAT transcriptome analysis.
4. Homology-Based Gene Models: Predictions produced from the REAT homology-guided workflow.
5. LiftOff-Derived Models: Gene annotations transferred and refined using LiftOff.

Gene models were classified into the biotypes protein_coding_gene, predicted_gene, ncRNA_gene, and transposable_element_gene. Each model was further designated as either high-confidence or low-confidence following the criteria outlined in Grewal *et al.*[42], with a modification: ncRNA genes were defined as gene models lacking CDS features, having a protein-coding potential score < 0.25 (calculated using CPC2[55] v0.1), and an expression score > 0.6 (results in Table 5).

Annotation liftover from contigs to scaffolds. From the contig annotation (GFF3) we removed genes on contigs that were identified as contamination. Mikado[47] was used to generate statistics from this GFF3 file. An AGP file was created to reflect the changes made during validation of the scaffolds and this was used as input to the *fromagp* function in the JCVI utility libraries[56] (v0.8.12) to generate a chain file. Then the *gff* function of CrossMap[57] (v0.3.4) was used to generate the liftover annotation (GFF3) from the modified contig GFF3. Mikado and GFFRead[58] (v0.12.2) were used to confirm that all gene models had transferred correctly. A total of 33,758 high-confidence protein coding genes were annotated with a mean transcript length of 2,151 bp (Table 6). An additional 19,331 genes were classified as low-confidence.

Functional annotation. All the proteins were annotated using the eifunannot pipeline[59] (v1.4.0) incorporating AHRD[60] (v.3.3.3). Databases and configuration as described in Grewal *et al.*[42].

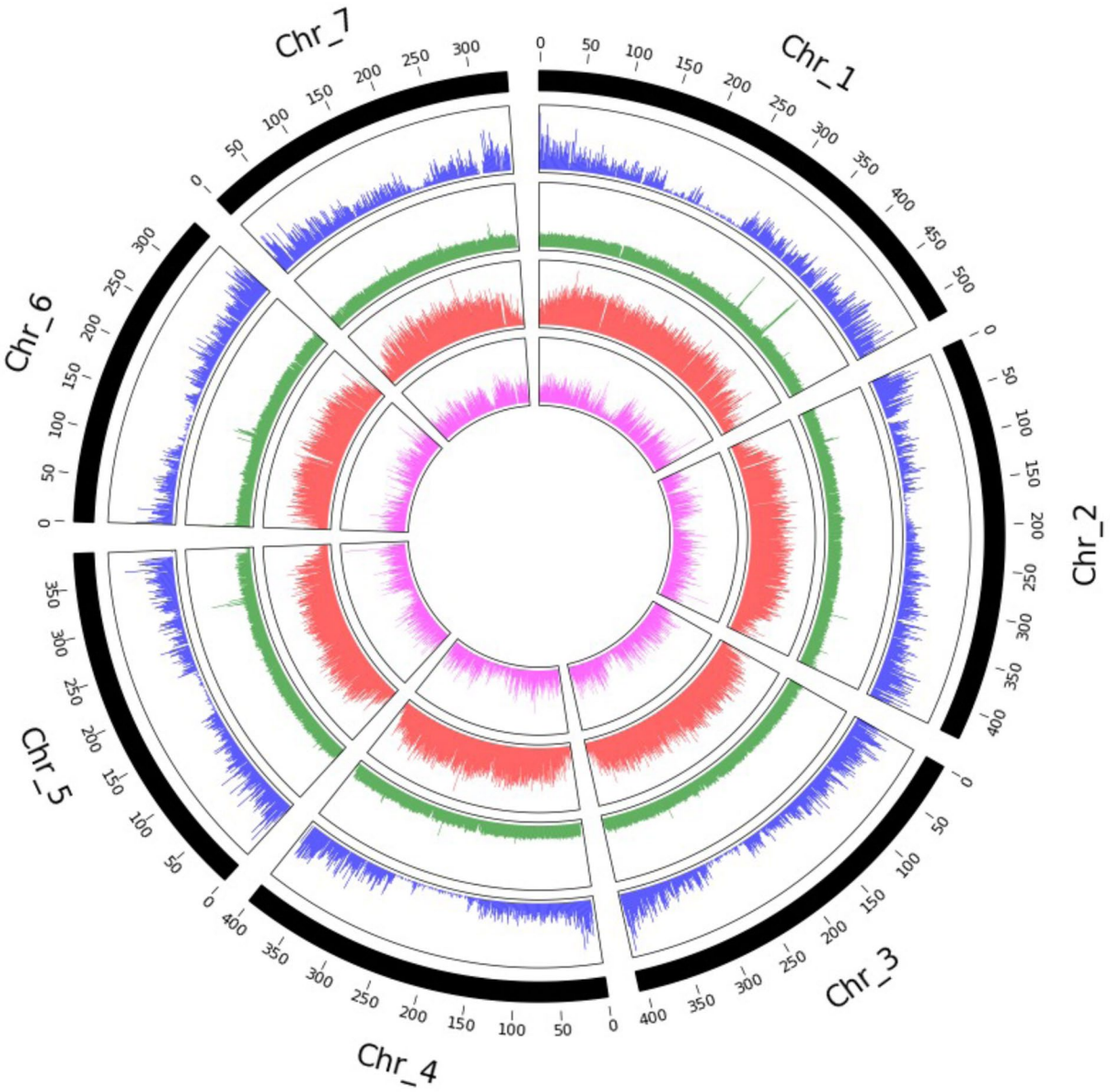

Fig. 4 Overview of the *Alopecurus aequalis* genome. Distribution of high-confidence protein-coding genes (blue), distribution of transposable elements (green), distribution of Ty3/*Gypsy* long-terminal repeats retrotransposons (LTR-RTs) (red) and distribution of Ty1/*Copia* LTR RTs (pink).

Genome overview. The *A. aequalis* genome contains seven chromosomes ranging in size from 521 to 339 Mb (Fig. 4). Gene density is highest at the distal ends of the chromosomes with very few genes in the centromeric regions. The distribution of transposable elements is quite even over the length of the chromosomes. The distribution of Ty3/*Gypsy* LTR retrotransposons (LTR RTs) is highest in gene-poor regions with fewer found in the gene-rich distal ends of the chromosomes. Ty1/*Copia* LTR RTs are distributed throughout the chromosomes with fewer in centromeric regions.

Comparative genomics analysis. GENESPACE[61] was used to assess synteny between *A. aequalis, A. myosuroides*[18] and *Hordeum vulgare* cultivar 'Morex'[30] to identify large structural rearrangements. Although a second *A. myosuroides* genome is available[19], our analyses showed no differences between the two versions. Therefore, we used the genome from Cai *et al.*[18] in all of our analyses.

The seven chromosomes of *A. aequalis* are more similar in structure to *H. vulgare* (Fig. 5a). Six of the seven chromosomes show a high level of synteny to *H. vulgare* with *A. aequalis* chromosome 1 comprising two syntenic blocks from *H. vulgare* chromosomes 4H and 5H. There is also evidence of a small translocation between *A. aequalis* chromosome 6 to *H. vulgare* chromosomes 4H.

Five chromosomes of *A. aequalis* show a high level of synteny to *A. myosuroides*. The chromosome arms of *A. aequalis* chromosome 1 are syntenic to two regions of *A. myosuroides* chromosomes 1 and 6. The break in

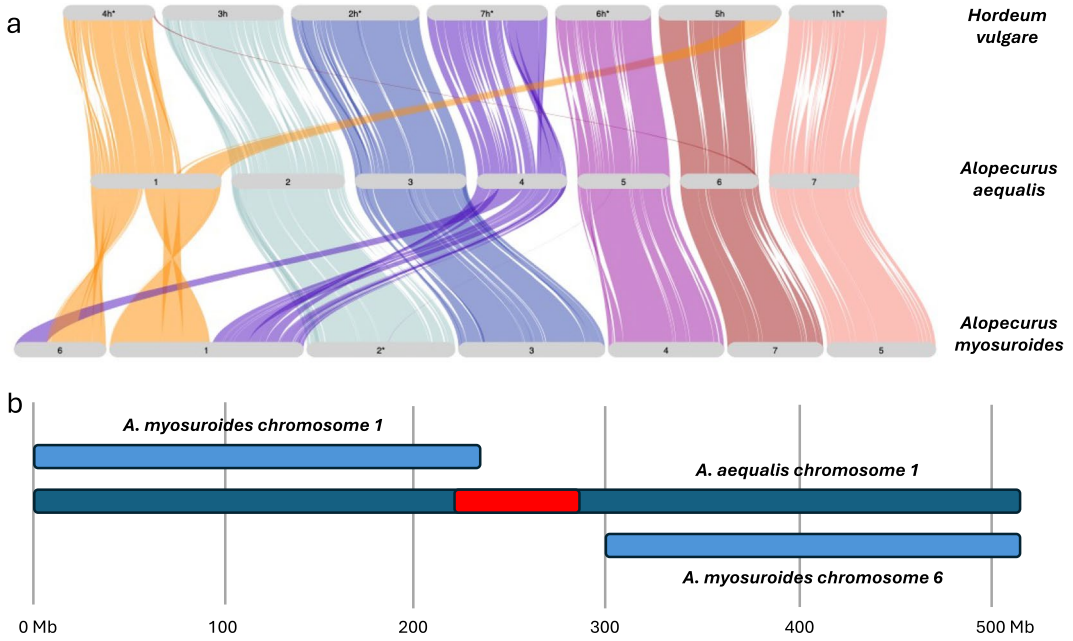

Fig. 5 Whole genome comparison between *Hordeum vulgare*, *Alopecurus aequalis* and *Alopecurus myosuroides*. (**a**) GENESPACE comparison showing synteny between the 3 species (* indicates that the chromosome has been reversed). (**b**) Detailed alignment showing the position of the *A. aequalis* chromosome 1 centromere (in red) in relation to the breakpoint.

synteny appears to occur in the centromeric region of *A. aequalis* chromosome 1, estimated to be between 220 and 290 Mb according to the Hi-C contact map (Fig. 3) which also corresponds to the drop in gene density in this region of the chromosome (Fig. 4). More detailed alignments show *A. aequalis* chromosome 1 aligns to *A. myosuroides* chromosome 1 from 0–240 Mb and to *A. myosuroides* chromosome 6 from 300–521 Mb (Fig. 5b).

Chromosome 4 of *A. aequalis* is syntenic to two large regions on *A. myosuroides* chromosomes 1 and 6. This relationship is more complex, showing several internal rearrangements within the larger syntenic region. It should be noted that the differences between *A. aequalis* chromosomes 1 and 4 compared to *A. myosuroides* chromosomes 1 and 6 were evident in the contig stage of assembly.

Data Records

All read datasets used in the assembly and annotation of *Alopecurus aequalis* are available at the European Nucleotide Archive (ENA) under accession ERP160206[62]. The assembly is available at the ENA under accession GCA_964340505[63]. The genome assembly and annotation files have been deposited online[64].

Technical Validation

Evaluating the quality of the genome assembly. BUSCO[65] (v5.3.2) and Merqury[66] (v1.3) were used to assess assembly completeness. BUSCO analysis showed 4,595 (93.9%) complete BUSCO genes from the poales_odb10 lineage dataset (total: 4,896) were found in the primary contigs. Of these, 4,401 were found as single-copy genes and 194 as duplicated genes. 77 BUSCO genes were fragmented and 224 were missing from the assembly. Merqury computed a k-mer completeness metric of 98.0% meaning that 98% of kmers from the reads are found in the assembly. As the genome is not completely homozygous we would not expect to achieve higher than this without comparing the reads to the combined haplotypes. The QV quality score was 61.6 corresponding to a base level accuracy of 99.9999%.

The Merqury spectra copy-number plot shows the majority of k-mers from the reads are found in the primary contigs only once at the expected coverage (the red region), and the majority of the low-coverage k-mers (originating from errors in the reads) are not present in the primary contigs (Fig. 2c).

Evaluating the quality of the genome annotation. BUSCO[65] with the poales_odb10 lineage dataset was used to measure the completeness of the high-confidence protein-coding gene set. In total, 99.1% of BUSCO groups were marked as complete (4,855 out of 4,896), 92.1% were complete and single-copy. There were two fragmented and 39 missing BUSCO groups indicating that the high-confidence protein coding gene set represents the *A. aequalis* gene complement accurately. OMArk[67] was used to assess the completeness and consistency of the *A. aequalis* annotation against gene families in the Pooideae clade (Fig. 6). The high-confidence gene set was compared to Hierarchical Orthologous Groups (HOGs) to give an estimate of completeness which showed 93.7% of HOGs were found (83.0% single and 10.7% duplicated) with 6.4% missing. When the combined high and low confidence genes were compared to HOGs we found 95.6% of HOGs with 4.5% missing indicating that some low

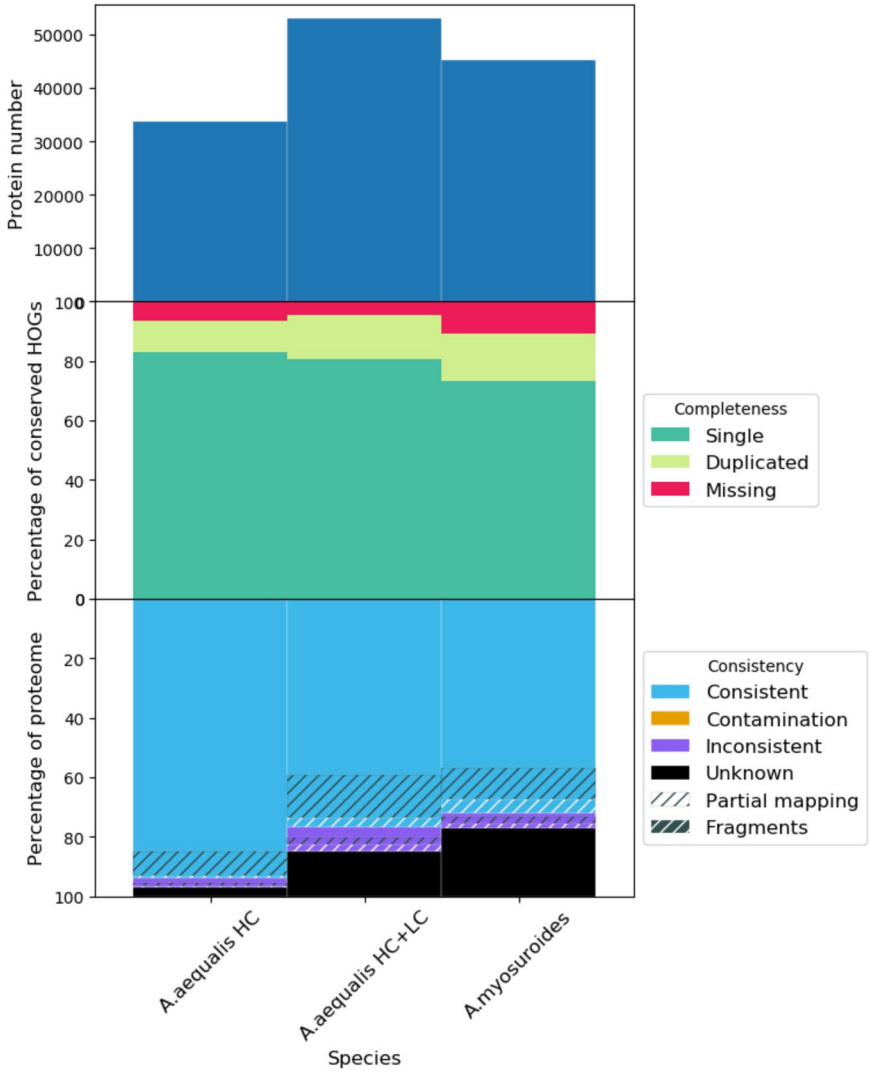

Fig. 6 OMArk plot comparing protein number, completeness and consistency of the *Alopecurus aequalis* genome annotation (HC: high confidence protein-coding genes, LC: low confidence protein-coding genes). OMArk output from the *A. myosuroides* genome annotation is included for comparison.

confidence genes represent valid HOGs. For the high confidence gene set, 93.8% of genes were consistent with gene families in the Pooideae clade, with 3.4% found in different lineages and 2.8% of unknown origin.

We identified 11,505 fewer genes in *A. aequalis* compared to that reported for *A. myosuroides*[18] (45,263), a difference likely caused by the different annotation methods used for each genome and the classification used in this pipeline to separate genes into high and low-confidence. Running OMArk on the *A. myosuroides* annotation (Fig. 6) shows more missing genes (10.8%) indicating a less complete annotation as well as more genes classified as "Unknown" (23.0%) indicating that many of the gene models included in the *A. myosuroides* annotation probably represent low confidence genes. The recent annotation of the closely related barley cultivar 'Morex' identified 35,821 high-confidence gene models[30] which is more similar to the annotation presented here.

We also used OrthoFinder[68] (v2.0.9) to cluster *A. aequalis* high-confidence genes and *A. myosuroides* genes into orthogroups. A total of 50,904 orthogroups were generated, 3,664 containing multiple genes, and 17,375 single-copy orthogroups. 9,660 orthogroups contained a single *A. aequalis* gene and 20,804 orthogroups contained a single *A. myosuroides* gene, also indicating that the *A. myosuroides* gene set contains many genes that are not present in the *A. aequalis* high-confidence gene set.

Code availability

All software used in this study was run according to instructions. The version and parameters are described in the methods. Anything not described in Methods was run with default parameters.

Received: 18 September 2024; Accepted: 2 December 2024;
Published online: 18 December 2024

References

1. BSBI. Alopecurus aequalis Sobol. *Online Plant Atlas* https://plantatlas2020.org/atlas/2cd4p9h.x99 (2020).
2. Liu, X. *et al*. Herbicide resistance in China: a quantitative review. *Weed Sci.* **67**, 605–612 (2019).
3. Zhao, N. *et al*. Effect of Environmental Factors on Germination and Emergence of Shortawn Foxtail (Alopecurus aequalis). *Weed Sci.* **66**, 47–56 (2018).
4. Murphy, B. P. & Tranel, P. J. Target-Site Mutations Conferring Herbicide Resistance. *Plants* **8**, 382 (2019).
5. Liu, X. *et al*. Managing herbicide resistance in China. *Weed Sci.* **69**, 4–17 (2021).
6. Gaines, T. A. *et al*. Mechanisms of evolved herbicide resistance. *J. Biol. Chem.* **295**, 10307–10330 (2020).
7. Casey, A. & Dolan, L. Genes encoding cytochrome P450 monooxygenases and glutathione S-transferases associated with herbicide resistance evolved before the origin of land plants. *PLOS ONE* **18**, e0273594 (2023).
8. Lan, Y. *et al*. Mechanism of Resistance to Pyroxsulam in Multiple-Resistant Alopecurus myosuroides from China. *Plants* **11**, 1645 (2022).
9. Li, J. *et al*. A novel naturally Phe206Tyr mutation confers tolerance to ALS-inhibiting herbicides in *Alopecurus myosuroides*. *Pestic. Biochem. Physiol.* **186**, 105156 (2022).
10. Varah, A. *et al*. The costs of human-induced evolution in an agricultural system. *Nat. Sustain.* **3**, 63–71 (2020).
11. Hicks, H. L. *et al*. The factors driving evolved herbicide resistance at a national scale. *Nat. Ecol. Evol.* **2**, 529–536 (2018).
12. Comont, D. *et al*. Evolution of generalist resistance to herbicide mixtures reveals a trade-off in resistance management. *Nat. Commun.* **11**, 3086 (2020).
13. Comont, D. *et al*. Dissecting weed adaptation: Fitness and trait correlations in herbicide-resistant Alopecurus myosuroides. *Pest Manag. Sci.* **78**, 3039–3050 (2022).
14. Zhu, W & Tu, S. Study on damage from Alopecurus aequalis Sobol and its economical threshold in wheat fields of Hubei province. *J.-HUAZHONG Agric. Univ.* **16**, 268–271.
15. Zeller, A. K., Zeller, Y. I. & Gerhards, R. A long-term study of crop rotations, herbicide strategies and tillage practices: Effects on *Alopecurus myosuroides* Huds. Abundance and contribution margins of the cropping systems. *Crop Prot.* **145**, 105613 (2021).
16. Délye, C., Jasieniuk, M. & Le Corre, V. Deciphering the evolution of herbicide resistance in weeds. *Trends Genet.* **29**, 649–658 (2013).
17. Kumar, S. *et al*. TimeTree 5: An Expanded Resource for Species Divergence Times. *Mol. Biol. Evol.* **39**, msac174 (2022).
18. Cai, L. *et al*. The blackgrass genome reveals patterns of non-parallel evolution of polygenic herbicide resistance. *New Phytol.* **237**, 1891–1907 (2023).
19. Kersten, S. *et al*. Standing genetic variation fuels rapid evolution of herbicide resistance in blackgrass. *Proc. Natl. Acad. Sci.* **120**, e2206808120 (2023).
20. Mc Cartney, A. M. *et al*. The European Reference Genome Atlas: piloting a decentralised approach to equitable biodiversity genomics. *Npj Biodivers.* **3**, 1–17 (2024).
21. https://github.com/EI-CoreBioinformatics/reat. EI-CoreBioinformatics (2024).
22. Pellicer, J., Powell, R. F. & Leitch, I. J. The Application of Flow Cytometry for Estimating Genome Size, Ploidy Level Endopolyploidy, and Reproductive Modes in Plants. in *Molecular Plant Taxonomy: Methods and Protocols* (ed. Besse, P.) 325–361, https://doi.org/10.1007/978-1-0716-0997-2_17 (Springer US, New York, NY, 2021).
23. Doležel, J., Binarová, P. & Lcretti, S. Analysis of Nuclear DNA content in plant cells by Flow cytometry. *Biol. Plant.* **31**, 113–120 (1989).
24. Obermayer, R., Leitch, I. J., Hanson, L. & Bennett, M. D. Nuclear DNA C-values in 30 Species Double the Familial Representation in Pteridophytes. *Ann. Bot.* **90**, 209–217 (2002).
25. Cantata Bio. Dovetail Omni-C Kit Non-mammalian Samples Protocol version1.2B, https://dovetailgenomics.com/wp-content/uploads/securepdfs/2022/08/Omni-C-Protocol_Non-mammal_v1.2B.pdf.
26. Sieber, V. K. & Murray, B. G. The cytology of the genus Alopecurus (Gramineae). *Bot. J. Linn. Soc.* **79**, 343–355 (1979).
27. Myers, E. W. Jr. https://github.com/thegenemyers/FASTK. (2024).
28. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
29. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
30. Mascher, M. *et al*. Long-read sequence assembly: a technical evaluation in barley. *Plant Cell* **33**, 1888–1906 (2021).
31. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
32. https://github.com/ArimaGenomics/mapping_pipeline. Arima Genomics, Inc. (2024).
33. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
34. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
35. https://github.com/broadinstitute/picard. Broad Institute (2024).
36. https://github.com/sanger-tol/PretextView. Tree of Life programme (2024).
37. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
38. https://github.com/EI-CoreBioinformatics/eirepeat. EI-CoreBioinformatics (2024).
39. https://github.com/Dfam-consortium/RepeatModeler.
40. Hubley, R. https://github.com/rmhubley/RepeatMasker. (2024).
41. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
42. Grewal, S. *et al*. Chromosome-scale genome assembly of bread wheat's wild relative Triticum timopheevii. *Sci. Data* **11**, 420 (2024).
43. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
44. Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. *GigaScience* **7**, giy131 (2018).
45. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
46. Shao, M. & Kingsford, C. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* **35**, 1167–1169 (2017).
47. Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, giy093 (2018).
48. Gotoh, O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res.* **36**, 2630–2638 (2008).
49. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
50. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**(Web Server), W465-W467, https://doi.org/10.1093/nar/gki458 (2005).
51. Brian, J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Abstract Genome Biology* **9**(1), https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
52. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
53. Venturini, L. https://github.com/lucventurini/ei-liftover (2022).

54. https://github.com/EI-CoreBioinformatics/minos. EI-CoreBioinformatics (2024).
55. Kong, L. *et al*. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345–W349 (2007).
56. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211 (2024).
57. Zhao, H. *et al*. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
58. Pertea, G. & Pertea, M. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research* **9**, 304 https://doi.org/10.12688/f1000research.23297.2 (2020).
59. EI-CoreBioinformatics. https://github.com/EI-CoreBioinformatics/eifunannot https://github.com/EI-CoreBioinformatics/eifunannot (2024).
60. Boecker, F. AHRD: Automatically Annotate Proteins with Human Readable Descriptions and Gene Ontology Terms. (University of Bonn, 2021).
61. Lovell, J. T. *et al*. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. *eLife* **11**, e78526 (2022).
62. Read datasets used in genome assembly and annotation of Alopecurus aequalis. Short Read Archive http://identifiers.org/insdc.sra:ERP160206.
63. Alopecurus aequalis genome assembly v0.2. https://identifiers.org/insdc.gca:GCA_964340505.1 (2024).
64. Chromosome-scale genome assembly and de novo annotation of Alopecurus aequalis. https://doi.org/10.5281/zenodo.14100513.
65. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14 (Springer, New York, NY, 2019).
66. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
67. Nevers, Y. *et al*. Quality assessment of gene repertoire annotations with OMArk. *Nat. Biotechnol.* 1–10 https://doi.org/10.1038/s41587-024-02147-w (2024).
68. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
69. Doležel, J., Bartoš, J., Voglmayr, H. & Greilhuber, J. Letter to the editor. *Cytometry A* **51A**, 127–128 (2003).

Acknowledgements

The authors would like to thank the ERGA executive team, Ann Mc Cartney, Alice Mouton, and Guilio Formenti for initiating the ERGA project and for ongoing support. They also would like to acknowledge Faye McDiarmid Oddy, Richard Hull, Laura Crook, and members of the Rothamsted Research Horticulture and Controlled Environment teams for their support with bulking seed and maintaining plants. The authors would like to thank David Bell of the Royal Botanic Garden Edinburgh for vouchering the *Alopecurus aequalis* sample https://data.rbge.org.uk/herb/E01358418 used in Figure 1g. The authors acknowledge the support of the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation; Earlham Institute (EI) Strategic Programme Grant Decoding Biodiversity BBX011089/1 and its constituent work package BBS/E/ER/230002B (Decode WP2 Genome Enabled Analysis of Diversity to Identify Gene Function, Biosynthetic Pathways, and Variation in Agri/Aquacultural Traits), as well as the Core Capability Grant BB/CCG2220/1 and the work delivered via the Research Computing Groups who manage and deliver High Performance Computing at EI. Part of this work was delivered via the BBSRC-funded National Capability in Genomics and Single Cell Analysis (BBS/E/T/000PR9816) and National Bioscience Research Infrastructure in Transformative Genomics (BBS/E/ER/23NB0006) at Earlham Institute by members of the Technical Genomics and the Core Bioinformatics Groups. Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council of the United Kingdom (BBSRC) and the authors acknowledge support from the Smart Crop Protection Industrial Strategy Challenge Fund (grant no. BBS/OS/CP/000001) and the Growing Health Institute Strategic Programme (BB/X010953/1; BBS/E/RH/230003 A). Part of this work was supported by Wellcome through the Darwin Tree of Life Discretionary Award (218328).

Author contributions

D.M. and A.H. conceived the project and designed the study. C.H. and D.M. grew the plants and sampled them. R.P. performed the flow cytometry analysis and I.J.L. coordinated this analysis. H.R performed initial sample coordination and DNA extraction. F.F. developed methods for and prepared Dovetail Omni-C and RNA-seq libraries. N.I. developed methods for and prepared HiFi and Iso-Seq libraries. A.D. developed the method for and performed the DNA extraction for HiFi library preparation. S.H. prepared RNA-seq libraries. T.B. sequenced Illumina RNA-seq libraries, Dovetail Omni-C library, PacBio Iso-Seq libraries, and PacBio HiFi libraries. J.W. performed genome assembly, genome annotation liftover and comparative analyses. G.K. and D.S. performed genome annotation. J.Wo. assisted with the interpretation of the Hi-C contact map. C.W., K.G., and L.C. planned and supervised data production. S.L. and K.B. coordinated the project from sample submission to data delivery. S.M. managed the project with respect to ERGA. J.W., D.M., C.W. and D.S. wrote the initial manuscript draft and A.H., H.R., J.L., P.N. and I.J.L. contributed editing and improvements. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.M.

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.