Title:

Field-based species identification in eukaryotes using single molecule, real-time sequencing.

Authors:

Joe Parker[1]*, Dion Devey[1], Andrew J. Helmstetter[1] & Alexander S.T. Papadopulos[1]*

[1]Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey UK. TW9 3AB
*Correspondence to a.papadopulos@kew.org and joe.parker@kew.org

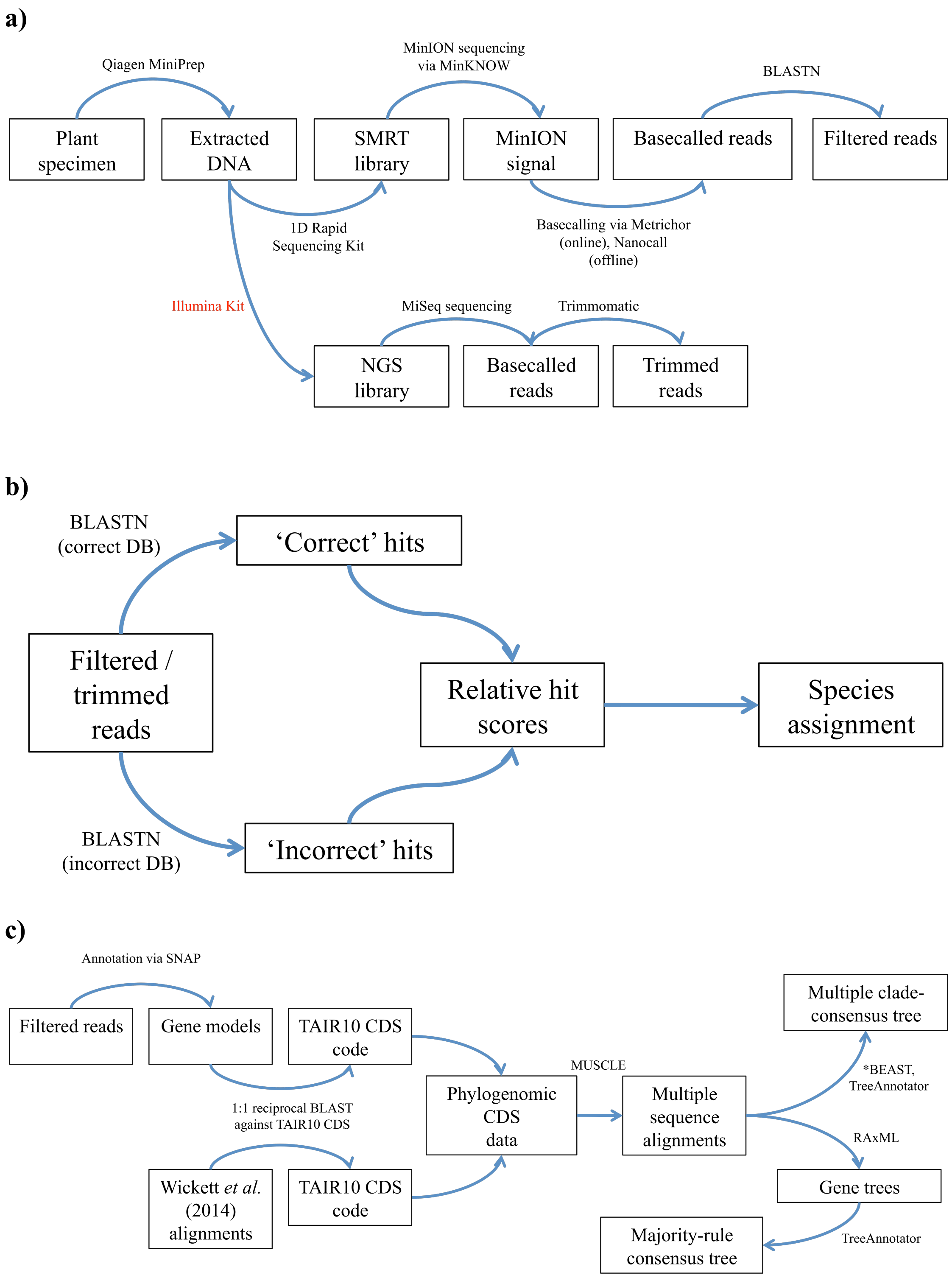
Keywords:

Nanopore, SMRT, MinION, onsite DNA sequencing, phylogenomics

Extended Data

Species	*Arabidopsis lyrata ssp. petraea*	*A, lyrata*	*A. thaliana*
Version	1.0	1.0	TAIR10
Date accessed	25/05/2016	17/05/2016	17/05/2016
Accession / ID	http://www.ncbi.nlm.nih.gov/assembly?LinkName=bioproject_assembly_all&from_uid=235280	http://www.ncbi.nlm.nih.gov/genome/493?genome_assembly_id=29434	http://www.ncbi.nlm.nih.gov/genome/4?genome_assembly_id=22492
Total assembly length	202,972,003	206,667,935	119,667,750
Total gap length	20,456,163	22,960,134	185,644
Number of scaffolds	281,536	695	7
Scaffold N50	7,848	24,464,547	23,459,830
Scaffold L50	6,426	4	3
Number of contigs	369,168	3,645	102
Contig N50	2,321	227,391	11,194,537
Contig L50	16,831	247	5
Total chromosomes & plasmids	0	0	7

Extended Data Table 1: Statistics of reference genomes used.



Extended Data Figure 1: Schematic of experimental workflows. a, Sampling-to-sequencing workflow. **b,** Sample identification workflow via BLASTN. **c,** Outline for direct annotation of raw SMRT reads followed by phylogenomic inference. See Methods for details.

Species	*Arabidopsis thaliana*	*Arabidopsis thaliana*	*Arabidopsis lyrata ssp. petraea*	*Arabidopsis lyrata ssp. petraea*	*(Totals)*
Reaction chemistry	R7.3, 1D	R9, 1D	R7.3, 1D	R9, 1D	
Run IDs	2507 2126 3637	5913 2144 0509	4901 1842 1201	1222 5458 1958 0824 2912 5201	
Start time	18/05/2016 21:36	18/05/2016 18:21	19/05/2016 16:10	19/05/2016 16:39:20	
Latest read	19/05/2016 15:17:00	19/05/2016 15:18:39	24/05/2016 01:14	22/05/2016 07:08:56	
Disc space (raw)	2.9Gb	106.8Gb	3.0Gb	35.1Gb	
Metrichor ID	115360	115459	115375	115432	
# reads	4,152	92,693	2,387	23,452	122,684
Yield: total, bp	7,351,585	233,244,147	16,092,487	46,118,754	302,806,973
Yield: mean, bp	1,771	2,516	6,742	1,967	3,249
Yield: median, bp	586	1,396	120	305	602
Yield: min, bp	7	5	6	5	5
Yield: max, bp	434,377	170,598	1,114,970	177,310	1,114,970
Yield: N25, bp	19,244	9,651	574,112	24,254	156,815
Yield: N50, bp	4,771	4,410	309,034	7,926	81,535
Yield: N75, bp	2,041	2,121	62,360	3,374	17,474

Extended Data Table 2: Performance of MinION sequencing runs. Yield summary statistics refer to untrimmed raw reads, including phage-lambda experimental control in the case of *A. thaliana* R9 data. See Methods for details.

	A. lyrata *ssp. petraea*	*A. lyrata* *ssp. petraea*	*A. thaliana*	*A. thaliana*
ID	AL1a	AL2a	AT1a	AT2a
MinION repeat[1]	Y			Y
# reads[2]	8,143,010	7,048,060	8,924,824	8,033,488
Yield (bp)	2,451,046,010	2,121,466,060	2,686,372,024	2,418,079,888

Extended Data Table 3: Performance of Illumina MiSeq sequencing runs. Field-extracted DNA was returned to the laboratory and libraries prepared according to the manufacturer's instructions for paired-end sequencing with a 300bp insert size (see Methods.) [1]Samples AL1a and AT2a were also sequenced by SMRT; AL2a and AT1a were not. [2]Paired reads only.

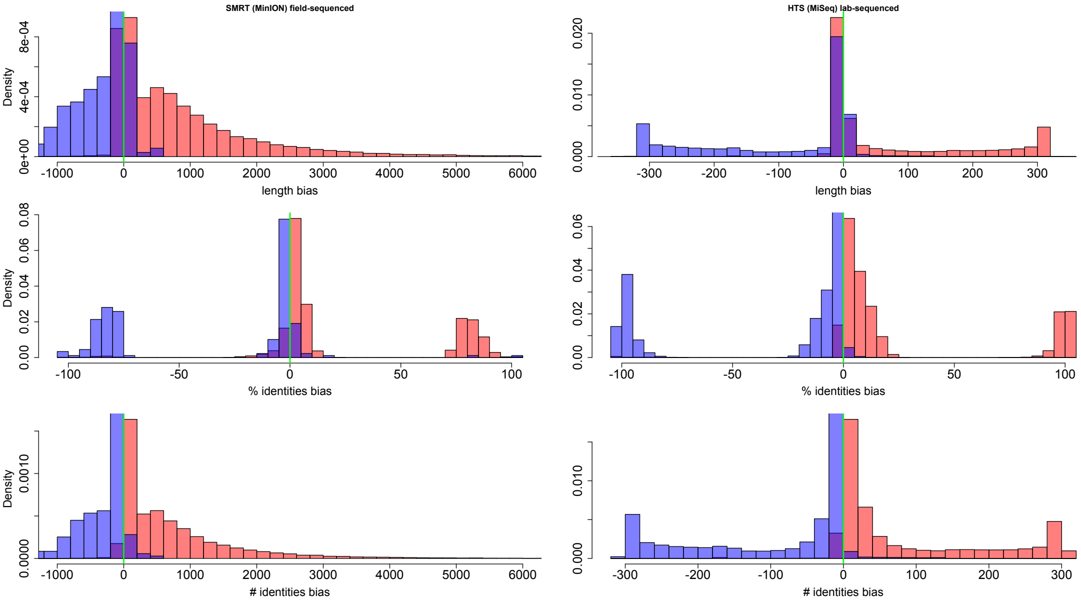
Species sample	Reference assembly	Type[1]	Total yield (bp)	Approx. coverage[2]	Read depth[3]	Aligned length[4]	Nominal error[4]
Arabidopsis thaliana	*A. thaliana*	ONT	240,597,532	2.01	1.82	54,722,05	0.216
A. lyrata ssp. petraea	*A. lyrata*[5]	ONT	62,211,241	0.30	4.07	980,358	0.225
A. lyrata ssp. petraea	*A. lyrata ssp. petraea*[6]	ONT	"	0.31	4.36	811,232	0.235
A. thaliana	*A. thaliana*	ILL	2,418,079,888	20.21	19.49	▨	▨
A. lyrata ssp. petraea	*A. lyrata*[5]	ILL	2,451,046,010	11.86	13.76		
A. lyrata ssp. petraea	*A. lyrata ssp. petraea*[6]	ILL	"	12.08	14.93		

Extended Data Table 4: Quality of sequenced reads mapped against available reference assemblies. [1]ONT: Oxford Nanopore MinION 1D (rapid) Sequencing kit, developer version; ILL: Illumina MiSeq paired-end 300bp. [2]Gross average coverage calculated as reference length divided by yield. [3]Inferred from BWA read mapping. [4]Inferred from reads aligned by LAST. [5]*Arabidopsis lyrata ssp. petraea* sample against *A. lyrata* assembly. [6]*A. lyrata ssp. petraea* sample against *petraea ssp.* assembly. See Methods for details.

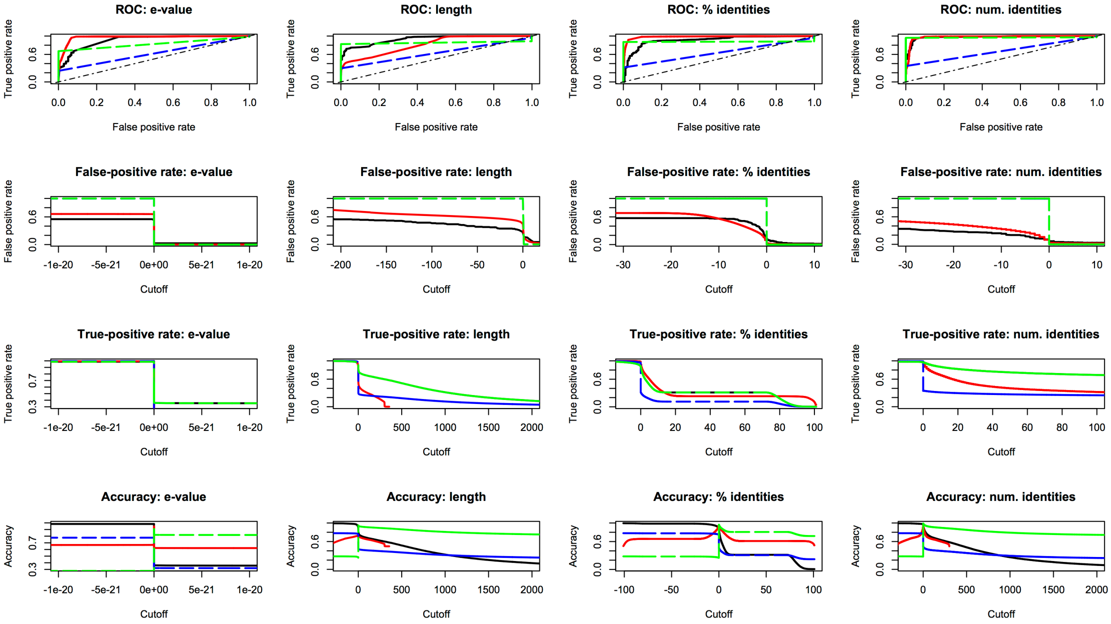
Source data / sample 1	**A. thaliana**	**A.lyrata combined[1]**	**A. thaliana**	**A.lyrata combined**
Pairwise comparison	**A.lyrata combined**	**A. thaliana**	**A.lyrata combined**	**A. thaliana**
Source data	**ONT 1D**	**ONT 1D**	**MiSeq**	**MiSeq**
# Reads, total	91,715	25,839	9,476,598	9,659,489
# Reads, 1-way TRUE[2]	10,322	76	2,140,403	2,907,921
# Reads, 1-way FALSE[3]	378	2	53,056	24,329
# Reads, 2-way BOTH [4]	22,386	101	7,098,032	6,256,969
# Reads, ZERO hits[5]	58,629	25,660	185,107	470,270
Proportion false both	*0.639*	*0.993*	*0.020*	*0.049*
Proportion true 1-way	*0.113*	*0.003*	*0.226*	*0.301*
Proportion false 1-way	*0.004*	*0.000*	*0.006*	*0.003*
Proportion 2-way (both)	*0.244*	*0.004*	*0.749*	*0.648*
Biases[6]:				
Cumulative[7] *length*	29,636,139	70,523	355,442,635	417,433,705
Cumulative identities	24,958,479	58,090	409,528,919	448,987,073
Cumulative % identities	850,070	6,203	128,227,752	162,303,707
Cumulative E-*values*	0.01	0.01	0.05	0.01
Mean[8] *length*	1,323.87	698.25	83.61	108.99
Mean identities	1,115	575	96	117
Mean % identities	37.97	61.41	30.16	42.38
Mean E-*values*	4.80E-07	1.05E-04	1.07E-08	3.52E-09

Extended Data Table 5: Sample identification via BLASTN.

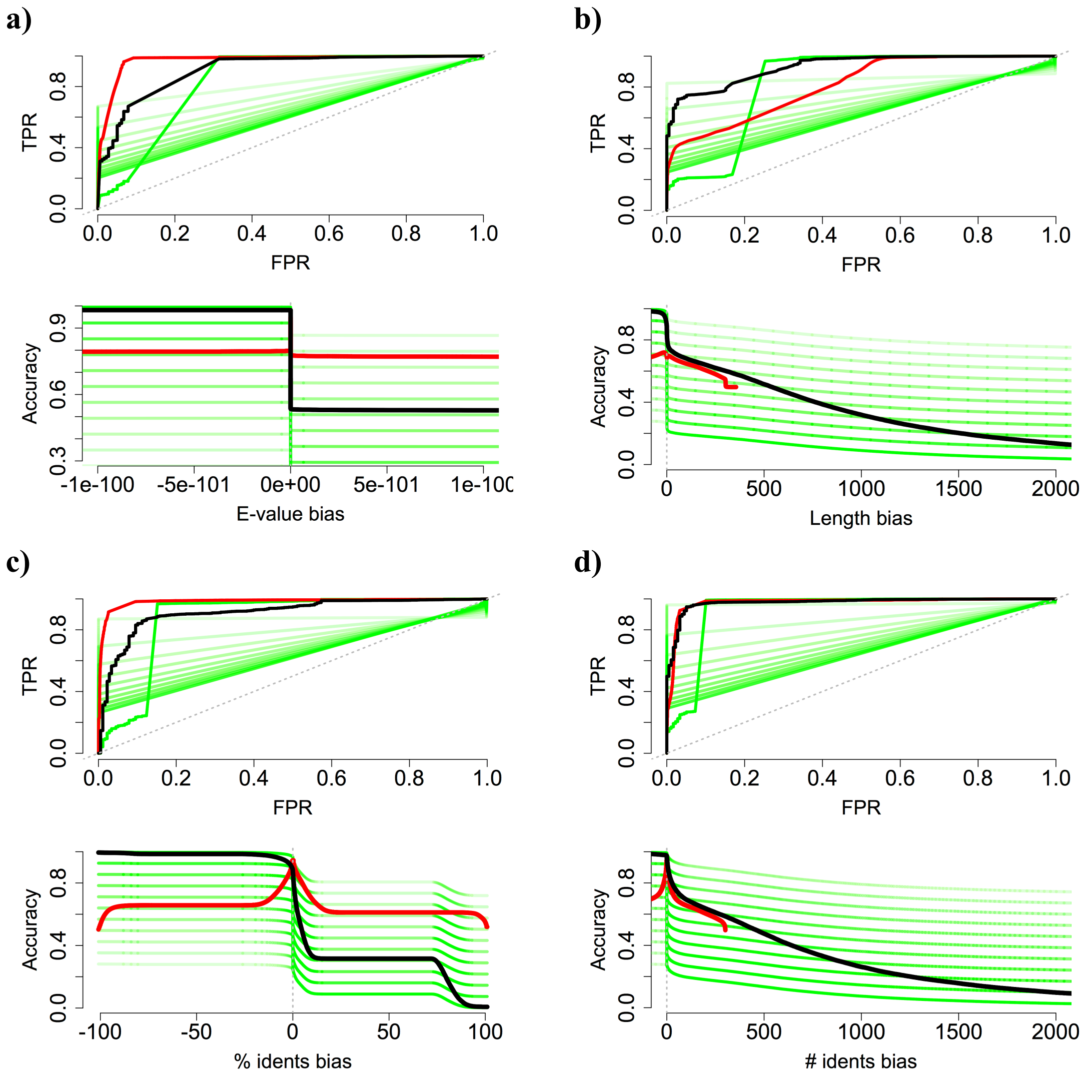
Notes: [1]*A. lyrata* and *A. lyrata ssp. petraea* databases combined, see Supplementary Information; [2]Total number of reads matching only conspecific database ('true-positives'); [3]Total number of reads matching only pairwise-compared database ('false-positives' in the case of a mixed /multiplexed sample, or 'false-negatives' in the case of a single sample); [4]Total number of reads matching both databases – further analysis would ordinarily be required; [5]Total number of reads with no hits in either comparison, e.g. 'false-negatives'; [6]Difference statistics for each query read calculated as (score conspecific comparison – score pairwise comparison), for BLASTN alignment length, alignment identities, alignment % identities and *E*-value; [8]Cumulative bias across all reads; [7]Mean bias across all reads.



Extended Data Figure 2: Distribution of difference statistics in BLASTN comparisons for congeneric species ID. Performance of test statistics for each-way congeneric sample ID (binary classification) using BLASTN evaluated for MinION (*left column*) and MiSeq (*right column*) platforms. Difference (test) statistics were calculated for each alignment as (true positive (TP) score – false positive (FP) score) for each of length, % identities and number of identities (product of length * % identities). For reads matching a single database only, a T or F assignment was made and difference statistic calculated by masking unobserved alignment scores as 'extreme' (-1 each for length and % identities, 999 for *e*-value). Reads sequenced from *A. thaliana* samples (comprising nominal true positives and false-positives (contaminants) are shown in red; reads from *A. lyrata* samples (nominal true negatives) shown in blue.



Extended Data Figure 3: Performance of difference statistics in BLASTN comparisons for congeneric species ID. Performance of test statistics for each-way congeneric sample ID (binary classification) using BLASTN evaluated for MinION (*black*) and MiSeq (*red*) platforms. Difference (test) statistics were calculated for each alignment as (true positive (TP) score – false positive (FP) score) for each of e-value, length, % identities and number of identities. For reads matching a single database only, a T or F assignment was made and difference statistic calculated by masking unobserved alignment scores as 'extreme' (-1 each for length and % identities, 999 for *e*-value). Reads that produced no hits to either database might represent false negatives (sequencing error, or genomic regions not represented in the reference genome BLAST databases) or true negatives (sequencing contaminants and sequencer noise). To evaluate the effect of including these nonmatching reads, all were coded with either ‘false’ labels and difference statistic values of zero (dashed green lines), or labelled ‘false’ in the case of reads sequenced from *A. lyrata* and ‘true’ for reads sequenced from *A. thaliana* (blue dashed lines). *Top row:* true-positive (TP) vs false-positive (FP) rate; classical receiver operating curve. *Second row:* FP rate with varying test statistic cutoff. *Third row:* TP rate with varying test statistic cutoff. *Bottom row:* Accuracy with varying test statistic cutoff. Accuracy estimated as (TP+TN / (P + N)). *Columns (L-R):* Difference statistics for *e*-value, total alignment length, % identities, and number of identities, respectively. See Methods for details.



Extended Data Figure 4: Modelling potential affect of incorrectly estimated TN/FN proportions. Red and black lines show empirically estimated statistical performance for species ID via BLASTN comparison of HTS and SMRT reads respectively (see Methods for details). Reads that produced no hits to either database might represent false negatives (sequencing error, or genomic regions not represented in the reference genome BLAST databases) or true negatives (sequencing contaminants and sequencer noise). To model the effect of including these nonmatching reads, dummy rows (one for each nonmatching read SMRT read: 58,629 from the *A. thaliana* experiment, and 25,660 from the *A. lyrata* experiment) were coded with 'false' labels and difference statistic values of zero. Proportions of these dummy reads were recoded with 'true' labels from 0-100% in 10% increments and classifier statistical performance was recalculated and plotted (shown in green; TN:FN mixtures by 10% increments shown from light to dark green shading. Plots (**a-d**) show results for bias statistics in *E*-value; total alignment length; % identities; and total alignment identities, respectively.

Species	**Arabidopsis thaliana**		**A. lyrata ssp. petraea**	
Data	**MiSeq, 300bp**	**MiSeq + MinION**	**MiSeq**	**MiSeq + MinION**
Assembler[1]	**Abyss**	**hybridSPAdes**	**Abyss**	**hybridSPAdes**
Illumina MiSeq NGS reads (300bp paired-end)	8,033,488	*8,033,488*	8,143,010	*8,143,010*
NGS total yield	2,418,079,888	*2,418,079,888*	2,451,046,010	*2,451,046,010*
Oxford Nanopore MinION SMRT reads, R7.3 + R9, N50 ~ 4,410bp	n/a	96,845	n/a	25,839
SMRT reads total yield	n/a	240,597,532	n/a	62,211,241
# contigs	24,999	10,644	37,568	85,599
Largest contig	89,717	413,462	101,114	38,313
Total length	106,455,313	119,031,857	151,562,895	117,256,694
Reference length	*119,667,750[2]*	*119,667,750*	*183,707,801[3]*	*183,707,801*
N50[4]	7,853	48,730	9,605	1,686
Unaligned length	7,121,882	6,737,059	36,669,847	35,287,390
Genome fraction (%)	82.0	88.7	53.4	43.7
Duplication ratio	1.01	1.058	1.17	1.02
# N's per 100 kbp	1.72	5.41	0.22	7.09
# mismatches / 100 kbp	518	588	1,297	1,097
# indels / 100 kbp	120	130	334	271
Largest alignment	76,935	264,039	44,515	17,201
Total aligned length	98,382,255	108,086,256	100,502,092	80,814,492
Coding loci completeness[5]:				
# genes, 'complete'	**219**	**245**	n/a	n/a
% genes, 'complete'	88.31%	98.79%		
# genes 'partial'	**238**	**246**	n/a	n/a
% genes, 'partial'	95.97%	99.19%		

Extended Data Table 6: Performance of *de novo* genome assembly on NGS and SMRT data. [1]*de novo* genome assemblies used either lab-sequenced short-read NGS data only (Abyss) or both NGS and field-sequenced SMRT datasets (Hybrid-SPAdes). [2]TAIR10 release. [3]INSDC: *A. lyrata*: ADBK00000000.1 (Hu *et al.,* 2011); *A. lyrata ssp. petraea*: BASP00000000.1 (Akama *et al.*, 2014). [4]Assembly statistics calculated using QUAST 4.0. [5]Approximate completeness of coding loci assessed via CEGMA. See Methods for details.