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

Supplementary Methods

Methods

An overview of study design is presented in Extended Data Figure 1.

Study site and sample collection: On consecutive days, tissue was collected from three specimens each of *A. thaliana* and *A. lyrata subsp. petraea* in Snowdonia National Park and sequenced and analysed in a tent. *A. lyrata* was collected from the summit of Moelwyn Mawr (52.985168° N, 4.003754° W; OL17 65554500; SH6558244971) and *Arabidopsis thaliana* was collected at Plâs Tan-y-bwlch (52.945976° N 4.002730° W; OL18 65604060; SH6552940610). Representative voucher specimens of each species are deposited at RBG, Kew. DNA extractions, library preparation and DNA sequencing with the MinION technology were all conducted using portable laboratory equipment in the Croesor valley on the lower slopes of Moelwyn Mawr immediately following sample collection (52.987463°N 4.028517° W; OL17 63904530; SH6392745273). Laboratory reagents were stored in passively-cooled polystyrene boxes with temperatures monitored using an Arduino Uno. Only basic laboratory equipment was used (two MinION sequencers and three laptops; see supplementary information).

DNA extraction used a modified version of the standard Qiagen DNeasy plant mini prep kit with the exception that the two batches were pooled at the DNeasy mini spin column step to maximise the DNA yield. An R7.3 and R9 1D MinION library preparation were performed for each species according to the manufacturer's instructions using an early version of the Nanopore RAD-001 library kit (Oxford Nanopore Technologies). No PCR machine was used. Lambda phage DNA was added to *A. thaliana* R9 library for quality control. For *A. thaliana,* the MinION experiment generated 96,845 1D reads with a total yield of 204.6Mbp over fewer than 16h of sequencing. Data generation was slower for *A. lyrata*, possibly due to temperature-related reagent degradation or unknown contaminants in the DNA extraction. Over ~90h sequencing, 25,839 1D reads were generated with a total yield of 62.2Mbp; this included three days of sequencing at RBG Kew following a 16h drive, during which reagents and flowcell were stored suboptimally (near room-temperature). BLASTN 2.4.0 (Camacho *et al.*, 2008) was used to remove 5,130 reads with identity to phage lambda. Data are given in **Extended Data Table 2.** Given available reference genome statistics for the *A. thaliana* TAIR10 release and the two *A. lyrata* assemblies (see **Extended Data Table 1**), these sequencing yields equate to approximate gross MinION coverage of 2.0x, 0.3x, and 0.31x for *A. thaliana*, *A. lyrata*, and *A. lyrata ssp. petraea*, respectively; and 20.2x, 11.9x and 12.0x respectively for paired MiSeq reads. The following week in a laboratory, NEBNext Ultra II sequencing libraries were prepared for four field-extracted samples (two individuals from each species) and sequenced on an Illumina MiSeq (300bp, paired end). In total, 11.3Gbp and 37.8M reads were generated (each ~ 8M reads and 2Gbp; see **Extended Data Table 3**).

Field offline basecalling and bioinformatics in real-time. Offline basecalling using nanocall 0.6.13 [https://github.com/mateidavid/nanocall] was applied to the R7.3 data as no offline R9 basecaller was available at the time. Basecalled reads were compared to the reference genomes of *A. thaliana* (TAIR10 release) and *A. lyrata subsp petraea* (1.0 release). 119 reads were processed in real-time with six reads making significant hits by BLASTN that scored correctly:incorrectly for species ID in a 2:1 ratio. After the sequencer had been halted a larger dataset of 1,813 reads gave 281 hits, with correct : incorrect : tied identifications in a 223:30:28 ratio.

Accuracy and mapping rates of short- and long-read data: Both trimmed, lab-sequenced short-reads and untrimmed, field-sequenced long-reads were aligned to the appropriate reference genomes using the BWA v0.7.12-r1039 (Li & Durbin, 2009) and LAST v581 (Kielbasa *et al.* 2011) programs, to estimate depth of coverage and nominal error rate in mapped regions. For all *A. thaliana* datasets (short and long-read), average mapped read depths were approximately equal to the gross coverage (see **Extended Data Table 4)**. MinION reads could be aligned to 53Mbp of the reference genome with LAST (approx. 50% of the total genome length). The nominal average error rate in these alignments was 20.9%, with indels and mismatches present in roughly equal proportions (**Supplementary Table 3**). For both MinION and MiSeq datasets, mapping and alignment to the *A. lyrata* and *A. lyrata ssp. petraea* assemblies was more problematic. For alignable MinION reads, error rates were slightly higher than for *A. thaliana* at 22.5% and 23.5%, estimated against *A. lyrata* and *A. lyrata ssp. petraea* assemblies, respectively. We note that these assemblies are poorer quality than the *A. thaliana* TAIR10 release; total genome lengths differ (206Mbp and 202Mbp,) and contiguity is relatively poor in both (695 and 281,536 scaffolds).

Determination of true- and false-positive detection rates, sensitivity, and specificity of field-sequenced (long) and lab-sequenced (short) read data: Each of the four datasets (short- and long-reads for each species) was matched against two custom databases (the *A. thaliana* reference genome and the two draft *A. lyrata* genomes combined) separately with BLASTN, retaining only the best hit for each query. Queries matching only a single database were counted as positive matches for that species. Queries matching both databases were defined as positives based on: a) longest alignment length (L_T); b) highest % sequence identities, c) longest alignment length counting only identities (L_I), or c) lowest *E*-value. Test statistics for each of these metrics were simply calculated as the difference of scores (length (ΔL_T), % identities, identities (ΔL_I), or *E*-value) between 'true' and 'false' hits. The statistical performance of these statistics (true- and false-positive rates, and accuracy) in putative analyses under varying threshold (cutoff) values were calculated and visualized using the ROCR package in R (Sing *et al.*, 2005). Comparison data are shown in **Extended Data Table 5, Extended Data Figure 3** and **Supplementary Table 4**. The high proportion of reads with significant hits to both species is expected given the close evolutionary relationships of the species. Analyses to determine the best statistics to discriminate between species using reads which aligned to both databases strongly indicated that difference in alignment lengths between the best discriminator, shown in **Figure 2** and **Extended Data Figures 2, 3 & 4.** Overall these show that the difference in alignment length is a powerful indicator for both short- and long-read data at any cutoff ≥ ~100bp. Furthermore, and surprisingly, at this and more conservative (greater difference) cutoffs, long-read field-sequenced reads had substantially more accuracy in true- and false-positive discrimination than short-read data. This suggests that this method provides a powerful means of species identification and we posit that the extremely long length of 'true positive' alignments compared with the natural length ceiling on false-positive alignments is largely responsible for this property.

Accumulation curves for simulated identification: 33,806 pairwise BLASTN hits obtained above in identification against *A. thaliana* and *A.lyrata* genomic reference databases were subsampled without replacement to simulate incremental accumulation of BLASTN hit data during progress of a hypothetical sequencing experiment producing 10,000 reads produced in total. 1,000 replicates were used to calculate means and variances for data accumulation in 0.1 log-increments from r=1 read to 10^4 reads total. For each read, ΔI, 'number of identities bias', was calculated as the difference (number of identities in *A. thaliana* alignment – number of identities in *A. lyrata* alignment). Each read was assigned to *A. thaliana* or not if it ΔL_I exceeded a given threshold, repeated at four possible values, L_{cutoff} ={0, 1, 10, 100}. Mean and aggregate (total) ΔL_I values were also calculated for each replicate over the progress of the simulated data collection. Results are shown in **Figure 3.**

***De novo* genome assemblies:** Short-read data was assembled *de novo* using ABYSS v1.9.0 (**Supplementary Table C**; Simpson *et al.*, 2009). A hybrid assembly with both short- and long-read datasets was performed with HybridSPAdes v3.5.0 (Antipov *et al.* 2016). Assemblies were completed for *A. thaliana* (sample AT2a) and *A. lyrata* (sample AL1a). Assembly statistics were calculated in Quast v4.3 (Gurevich *et al.*, 2013). Completeness of the final hybrid assemblies was assessed using CEGMA v2.5 (Parra *et al.*, 2007). *de novo* assembly using only long-read data for *A. thaliana* was attempted with Canu v1.3 (Koren *at al.*, 2016), but performed poorly (**Supplementary Table D**) due to the low long-read sequencing coverage. Results of *de novo* genome assemblies are given in overview in **Extended Data Table 6** and **Supplementary Table 5** with further details of MiSeq data assemblies via Abyss in **Supplementary Table C**, of Nanopore data via Canu in **Supplementary Table D**, and hybrid data assemblies via hybrid-SPAdes in **Supplementary Table E.** Analyses of genome contiguity and correctness and conserved coding loci completeness indicated that assembly of MiSeq data performed as expected (20x coverage produced ~25,000 contigs covering approx 82% of the reference genome at an N50 of 7,853bp). By contrast, the hybrid assembly of *A. thaliana* illumina MiSeq and Oxford Nanopore MinION data significantly improved on the MiSeq-only assembly: 24,999 contigs reduced to 10,644; total assembly length increased to close to the length of the reference genome (119.0Mbp) with nearly 89% mappable; N50 and longest contig statistics both improved (N50 7,853 → 48,730bp) indicating better contiguity from the addition of long reads. Completeness of coding loci as estimated by CEGMA (**Supplementary Table F**) greatly increased to ~99%. Long reads did not compromise the accuracy of high-coverage short-read data; basewise error rates were not significantly worse.

Direct gene annotation of single unprocessed field-sequenced reads: The length of typical individual nanopore reads is of similar magnitude to genomic coding sequences. Consequently, useful phylogenomic information could potentially be obtained by annotating reads directly, without a computationally expensive genome assembly step. Raw, unprocessed *A. thaliana* reads were individually annotated directly without assembly via SNAP (Korf, 2004). To verify which gene predictions were genuine, the DNA sequences (and 1kb flanking regions, where available) were matched to available *A. thaliana* (TAIR10) genes with default parameters. BLAST hits were further pruned based on quality (based on 1st-quartile quality scores: alignments length bias $\Delta L_T \geq +570$bp / % identities bias $\geq +78.68$ / E-value bias ≥ 0), reducing the number of hits from 18,098 to 10,615. Sample read alignments and details of SNAP output BLAST score summary statistics are given in **Supplementary Table 6** and encounter curves-through-time are shown in **Supplementary Figure 1.**

Phylogenomics of raw-read-annotated *A. thaliana* genes: Predicted *A. thaliana* gene sequences were combined with a published phylogenomic dataset spanning 852 orthologous, single-copy genes in plants and algae (Wickett *et al.* (2014), downsampled to 6 representative taxa for speed: *Equisetum diffusum, Juniperus scopulorum, Oryza sativa, Zea mays, Vitis vinifera* and *Arabidopsis thaliana*. Our putative gene models were assigned identity based on reciprocal best-hit BLASTN matching with the *A. thaliana* sequences in these alignments, yielding 207 matches, of which the top 56 were used for phylogenomic analysis (**Supplementary Table 7a**), only 18 having no missing taxa in the Wickett *et al*. (2014) dataset (**Supplementary Table 7b**). Alignments were refined and trimmed with a 50% missing-data filter then used to infer species trees in two ways: (i) single gene phylogenies inferred separately and combined into a summary tree; (ii) a species tree inferred directly from the data under the multispecies coalescent (Heled & Drummond, 2010), implemented in *BEAST v2.4.4 (Bouckaert *et al.*, 2014; with adequate MCMC performance confirmed using Tracer v1.5). A maximum clade credibility (MCC) tree was produced using TreeAnnotator v.1.7.4 (Drummond *et al.*, 2012). Results are given in **Extended Data Table 6.** Phylogenies inferred by orthodox (RAxML) and multispecies coalescent (*BEAST) methods are shown in **Supplementary Figure 2** and **Supplementary Figure 3** and agreed with each other and the established phylogeny presented in Wickett *et al.* (2014). Additional protein-based reconstructions were also attempted with more limited data and lower node support values as a result (**Supplementary Table 7c**; **Supplementary Figure 4**).

References

Antipov D, Korobeynikov A, McLean JS, Pevzner PA. (2016) hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**(7):1009-15. doi: 10.1093/bioinformatics/btv688.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30**(15):2114-2120.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C-H., Xie, D., Suchard, MA., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, **10**(4):e1003537.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. (2008) BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.

Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7 *Molecular Biology And Evolution* **29**:1969-1973.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5):1792-97.

Gureyvich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8):1072-1075.

Heled, J. and Drummond, A.J. (2010) Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27** (3):570-580.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**(3):487-93.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R. & Phillippy, A.M. (2016) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. http://dx.doi.org/10.1101/071282.

Korf I. (2004) Gene finding in novel Genomes. *BMC Bioinformatics* **5**:59.

Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25(**14):1754–1760.

Parra, G., Bradnam, K. & Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**(9):1061-1067.

Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven JM Jones, and Inanc Birol. (2009) ABySS: a parallel assembler for short read sequence data. *Genome research* **19**(6):1117-1123.

Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. (2005) ROCR: Visualizing classifier performance in R. *Bioinformatics* **21**(20):3940-3941.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9):1312-1313.

Wickett, N.J., *et al.* (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. *PNAS* **111**(45):E4859-4868.