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

Authors:
Joe Parker[1], Andrew J. Helmstetter[1], Dion Devey[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

Abstract
Advances in DNA sequencing and informatics have revolutionised biology over the past four decades, but technological limitations have left many applications unexplored[1,2]. Recently, portable single molecule, real-time (SMRT) technologies have become available. These offer opportunities to rapidly collect and analyse genomic data anywhere[3–5]. However, the generation of datasets from large, complex genomes has been constrained to laboratories[6,7]. The portability and long DNA sequences of SMRT offer great potential for field-based species identification, but the feasibility and accuracy of these technologies for this purpose have not been assessed. Here, we show that a field-based SMRT analysis of closely-related plant species (*Arabidopsis spp.*)[8] has many advantages over laboratory-based high-throughput sequencing (HTS) methods for species level identification-by-sequencing and *de novo* phylogenomics. Samples were collected and sequenced in a single day by SMRT using a portable, "*al fresco*" laboratory. Our analyses demonstrate that correctly identifying unknown reads from matches to a reference database with SMRT reads enables rapid and confident species identification. Individually annotated SMRT reads can be used to infer the evolutionary relationships of *A. thaliana*. Furthermore, hybrid genome assembly with SMRT and HTS reads substantially improved upon a genome assembled from HTS reads alone. Field-based SMRT makes real-time, rapid specimen identification and genome wide analyses possible. These technological advances are set to revolutionise research in the biological sciences[9] and have broad implications for conservation, taxonomy, border agencies and citizen science.

Introduction
DNA sequencing used to be a slow undertaking, but the past decade has seen an explosion in HTS methods[2,10]. DNA barcoding (i.e., the use of a few, short DNA sequences to identify organisms) has benefited from this sequencing revolution[11,12], but has never become fully portable. Samples must be returned to a laboratory for testing and the discrimination of closely related species using few genes can be problematic due to evolutionary phenomena (e.g. lineage sorting, shared polymorphism and hybridisation)[10]. While typical barcoding approaches have been effective for generic level identification, accuracy is much more limited at the species level[11,13] and concerns remain[14]. Species delimitation using limited sequencing information has also been problematic and is thought to heavily underestimate species diversity[11,15]. Consequently, increasingly elaborate analytical methods have been spawned to mitigate the inherent limitations of short sequences[13,16]. The Oxford Nanopore Technologies® MinION® is one of a new generation of SMRT DNA sequencers that is small enough to be portable for fieldwork and produces data within minutes[17,18]. These properties suggest species identification could be conducted using genome scale data generated at the point of sample collection. Furthermore, the large number of long reads generated[17] may provide more accurate species-level identification than current approaches. This application offers great potential for conservation, environmental biology, evolutionary biology and combating wildlife crime, however, this potentially exciting combination of methods has not yet been tested in the field for eukaryotes.

Our experiment was designed to determine whether DNA reads produced entirely in the field could accurately identify and distinguish samples from closely-related species (*A. thaliana* (L.) Heynh. and *A. lyrata* (L.) O'Kane & Al-Shehbaz). Recent analyses have shown that gene flow has been common and shared polymorphisms are abundant between the morphologically distinct species in *Arabidopsis.* Indeed, the two study species share >20,000 synonymous SNPs[8], making this a good stress test of genome scale SMRT sequencing for species discrimination.

Results and Discussion

The first goal was to extract and sequence shotgun genomic data from higher plant species in the field using SMRT technology in sufficient quantity for downstream analyses within hours of the collection of plant tissue. On consecutive days, tissue was collected from three specimens each of *A. thaliana* and *A. lyrata subsp. petraea* (Figs. 1b,c) in Snowdonia National Park, and prepared, sequenced and analysed outdoors in the Croesor Valley (Fig. 1a). Only basic laboratory equipment was used for DNA extraction and MinION sequencing-library preparation; we did not use a PCR machine (Fig. 1d; see Supplementary Methods for details). One specimen of each species was sequenced with both R7.3 and R9 MinION chemistries. For *A. thaliana,* the SMRT experiment generated 97k reads with a total yield of 204.6Mbp over fewer than 16h of sequencing (see Extended Data Table 2). Data generation was slower for *A. lyrata*, over ~90h sequencing (including three days of sequencing at RBG Kew following a 16h drive), 26k reads were generated with a total yield of 62.2Mbp. At the time, a limited implementation of local basecalling was available for the R7.3 data only. Of 1,813 locally basecalled reads, 281 had successful BLAST matches to the reference databases with a correct to incorrect species ID ratio of 223:30. The same samples were subsequently sequenced using HTS short read technology (Illumina MiSeq™, paired-end, 300bp; Extended Data Table 3). Mapping reads to available reference genomes for the *A. thaliana* (TAIR10 release[19]) and two *A. lyrata* assemblies[20,21] indicates approximate SMRT coverage of 2.0x, 0.3x, and 0.3x for *A. thaliana*, *A. lyrata*, and *A. lyrata ssp. petraea*, respectively; and 19.5x, 11.9x and 12.0x respectively for HTS reads (Table 1, Tables S1 and S4). These results demonstrate that the entire process (from sample collection thorough to genome scale sequencing) is now feasible for eukaryotic species within a few hours in field conditions.

As expected given the developmental stages of the technologies, the quality and yield of field sequenced SMRT data was lower than the HTS data (Extended Data Table S4). *Arabidopsis thaliana* SMRT reads could be aligned to approx. 50% of the reference genome (53Mbp) with an average error rate of 20.9%. Indels and mismatches were present in similar proportions. The *A. lyrata* SMRT data were more problematic with significantly poorer mapping to the two *A. lyrata* assemblies, whereas, the HTS data performed relatively well. For the limited number of alignable SMRT reads, error rates were slightly higher than for *A. thaliana* (22.5% and 23.5%). The poorer SMRT results for *A. lyrata* may be a consequence of temperature-related reagent degradation in the field or due to unknown contaminants in the DNA extraction that inhibited library preparation and/or SMRT sequencing. Despite the smaller yield and lower accuracy of the SMRT compared to HTS data, the SMRT reads were up to four orders of magnitude longer than the HTS reads and we predicted they would be useful for species identification, hybrid genome assembly and phylogenomics.

To explore the utility of these data for species identification, the statistical performance of field-sequenced (SMRT) and lab-sequenced (HTS) read data was assessed. Datasets for each species were compared to two databases via BLASTN, retaining single best-hits: one database contained the *A. thaliana* reference genome and the second was composed of the two draft *A. lyrata* genomes combined. Reads which matched a single database were counted as positive matches for that species. The majority of matching reads hit both databases, which is expected given the close evolutionary relationships of the species. In these cases, positive identifications were determined based on four metrics; a) the longest alignment length, b) the highest % sequence identities and c) the largest number of sequence identities d) the lowest *E*-value. Test statistics for each of these metrics were calculated as the difference of scores (length, % identities, or *E*-value) between 'correct' and 'incorrect' database matches. The performance of these difference statistics for binary classification was assessed by investigating the true and false positive rates (by reference to the known sample species) across a range of threshold difference values (Figs. S2, S3, S4 & Table S5). For both short- and

long-read data at thresholds greater than 100bp, the differences in total alignment lengths (ΔL_T) or number of identities (ΔL_I) are superior to e-value or % identity biases (Figs. 2a-d). Furthermore, at larger thresholds (i.e., more conservative tests), SMRT reads retained more accuracy in true- and false-positive discrimination than HTS data. This proves that whole genome shotgun SMRT is a powerful method for species identification. We posit that the extremely long length of the observed 'true positive' alignments compared with an inherent length ceiling on false-positive alignments in a typical BLASTn search is largely responsible for this property.

To evaluate the speed with which species identification can be carried out, we performed *post hoc* analyses by subsampling the SMRT *A. thaliana* dataset. This simulated the rate of improvement in species assignment confidence over a short SMRT run. We classified hits among the subsampled reads based on (i) ΔL_I over a range of threshold values (ii) mean ΔL_I and (iii) aggregate ΔL_I (Fig. 3). This demonstrates that a high degree of confidence can be assigned to species identifications over the timescales needed to generate this much data (i.e., < one hour) and that variation in the accuracy of identifications quickly stabilises above 1000 reads. Aggregate ΔL_I values rapidly exclude zero (no signal) or negative (incorrect assignment) values, making this simple and rapidly-calculated statistic particularly useful for species identification. In a multispecies context, the slopes of several such log-accumulation curves could be readily compared, for example.

Field-sequencing large and complicated eukaryotic genomes with SMRT data alone would require a greater volume of data than available here[7,22,23]. As expected, *de novo* assembly of SMRT data performed poorly, likely due to insufficient coverage. However, these data do have potential for hybrid genome assembly approaches. We assembled the HTS data *de novo* using ABYSS[24] and produced a hybrid assembly with both SMRT and HTS datasets using HybridSPAdes[25]. The hybrid assembly was an improvement over the HTS-only assembly (see Table S6) with fewer contigs, a total assembly length closer to the reference (119.0Mbp), N50 and longest contig statistics both increasing substantially and estimated completeness (CEGMA[26]) of coding loci increased to ~99%. These results suggest that relatively small quantities of long and short reads can produce useful genome assemblies when analysed together, an important secondary benefit of field–sequenced data.

The length of typical SMRT reads is similar to that of genomic coding sequences (1-10kb)[17]. This raises the possibility of extracting useful phylogenetic signal from such data, despite the relatively high error rates of individual reads. We annotated individual raw *A. thaliana* reads directly, without genome assembly, which recovered over 2,000 coding loci from the data sequenced in the first three hours (Fig. 2e). These predicted gene sequences were combined with a published dataset spanning 852 orthologous, single-copy genes[27], downsampled to 6 representative taxa. Of our gene models, 207 were present in the Wickett *et al.*[27] dataset and the best 56 matches were used for phylogenomic analysis (see Supplementary Methods for details). The resulting phylogenetic trees (Fig. 2f) are consistent with the established intergeneric relationships[27]. Although the taxonomic scale used here for phylogenomics is coarse it highlights an additional benefit to rapid, in-the-field sequencing for evolutionary research.

This experiment is the first to demonstrate field-based sequencing of higher plant species. When directly compared to lab-based HTS, our experiment highlights key discriminatory metrics for highly accurate species identifications using portable SMRT sequencing. Few approaches can boast this level of discriminatory power and none of these have the same degree of portability[10,11]. The data produced for identification is also useful for genome

assembly. Entire coding sequences can be recovered from single reads and incorporated into evolutionary analyses. Clearly, data generated with the goal of accurate species identification has much broader usefulness for genomic and evolutionary research. Few technical barriers remain to prevent the adoption of portable SMRT by non-specialists, or even keen amateurs and schoolchildren. As these tools mature, and the number of users expands, portable SMRT sequencing can revolutionise the way in which researchers and practitioners can approach ecological, evolutionary and conservation questions.

Methods summary:
Genomic DNA was extracted from two plant specimens and sequenced on Oxford Nanopore MinION devices according to manufacturers' recommendations in a portable outdoor laboratory. Offline basecalling software and local BLAST (v2.2.31) were used to identify individual reads on-site. Short reads were sequenced in the laboratory from the same extracted DNA using an Illumina MiSeq. Local BLAST was used to identify reads from all four datasets (2 field x 2 species) by comparison to available published reference genomes. Gene models were predicted directly from individual DNA reads using SNAP (v2006-07-28), matched to existing phylogenomic datasets and used to infer plant phylogenies using MUSCLE (v3.8.31) and RAxML (v7.2.8). *de novo* genome assemblies were performed using Abyss (v1.5.2) and Hybrid-SPAdes (v3.5.1) with completeness assessed with QUAST (v4.0) and CEGMA. R (v3.1.3) was used to perform statistical analyses. Additional details are given in the Supplementary Methods.

References

1. Hebert, P. D. N., Hollingsworth, P. M., Hajibabaei, M. & Hebert, P. D. N. From writing to reading the encyclopedia of life. *Philos. Trans. R. Soc. London B Biol. Sci.* **371,** 1–9 (2016).
2. Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R. & Golding, G. B. A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philos. Trans. R. Soc. London B Biol. Sci.* **371,** 20150330 (2016).
3. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–32 (2016).
4. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8,** 97 (2016).
5. Edwards, A., Debbonaire, A. R., Sattler, B., Mur, L. A. & Hodson, A. J. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. *bioRxiv* 73965 (2016). doi:10.1101/073965
6. Schmidt, K. *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.* dkw397 (2016). doi:10.1093/jac/dkw397
7. Datema, E. *et al.* The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. *bioRxiv* (2016). doi:10.1101/084772
8. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48,** 1077–1082 (2016).
9. Erlich, Y. A vision for ubiquitous sequencing. *Genome Res.* **25,** 1411–1416 (2015).
10. Mallo, D. & Posada, D. Multilocus inference of species trees and DNA barcoding. *Philos. Trans. R. Soc. London B* **371,** 20150335 (2016).
11. CBOL Plant Working Group *et al.* A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12794–7 (2009).
12. Hollingsworth, P. M., Li, D.-Z., van der Bank, M. & Twyford, A. D. Telling plant species apart with DNA: from barcodes to genomes. *Philos. Trans. R. Soc. B Biol. Sci.* **371,**

20150338 (2016).
13. Little, D. P. DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS One* **6,** (2011).
14. Collins, R. A. & Cruickshank, R. H. The seven deadly sins of DNA barcoding. *Mol. Ecol. Resour.* **13,** 969–975 (2013).
15. Tang, C. Q. *et al.* The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proc. Natl. Acad. Sci.* **109,** 16208–16212 (2012).
16. Zhang, A. B. *et al.* A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Mol. Ecol.* **21,** 1848–1863 (2012).
17. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **3,** 1–8 (2015).
18. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 13770–3 (1996).
19. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40,** D1202–D1210 (2012).
20. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43,** 476–81 (2011).
21. Akama, S., Shimizu-Inatsugi, R., Shimizu, K. K. & Sese, J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucleic Acids Res.* **42,** (2014).
22. Goodwin, S., Gurtowski, J., Ethe-sayers, S., Deshpande, P. & Michael, C. Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome. (2015).
23. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14,** 1097–1102 (2014).
24. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. & Jones, S. J. M. ABySS : A parallel assembler for short read sequence data ABySS : A parallel assembler for short read sequence data. 1117–1123 (2009). doi:10.1101/gr.089532.108
25. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. HybridSPAdes: An algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32,** 1009–1015 (2016).
26. Parra, G., Bradnam, K. & Korf, I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).
27. Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.* **111,** E4859–E4868 (2014).

Acknowledgements
This work was funded by a Pilot Study Grant to JDP and a Howard Lloyd Davies legacy grant to ASTP. JDP was also supported by funding from the Calleva Foundation Phylogenomic Research Programme and the Sackler Trust. The authors also thank The Botanical Society of Britain & Ireland. Natural Resources Wales, Tim Wilkinson, Robyn Cowan and Patricia and David Brandwood for assistance.

Conflicts of interest and author information
Oxford Nanopore Technologies provided free reagents and consumables to this study, as well as technical advice. JDP and ASTP received travel remuneration and free tickets to present an early version of this work at a conference (London Calling 2016). Basecalled read data for Illumina and Oxford Nanopore sequencing runs are available via the EBI ENA at XXX. Scripts for downstream analyses and a list of required software are available at http://github.com/lonelyjoeparker/XXXX.

Author contributions

ASTP and JDP conceived the study and obtained funding. ASTP, DD and JDP designed and conducted fieldwork. ASTP designed and conducted field-based labwork with input from JDP, AH and DD. AH conducted lab-based sequencing. JDP conducted bioinformatics and phylogenomic analyses with contributions from AH. ASTP and JDP prepared the manuscript with contributions from DD and AH.

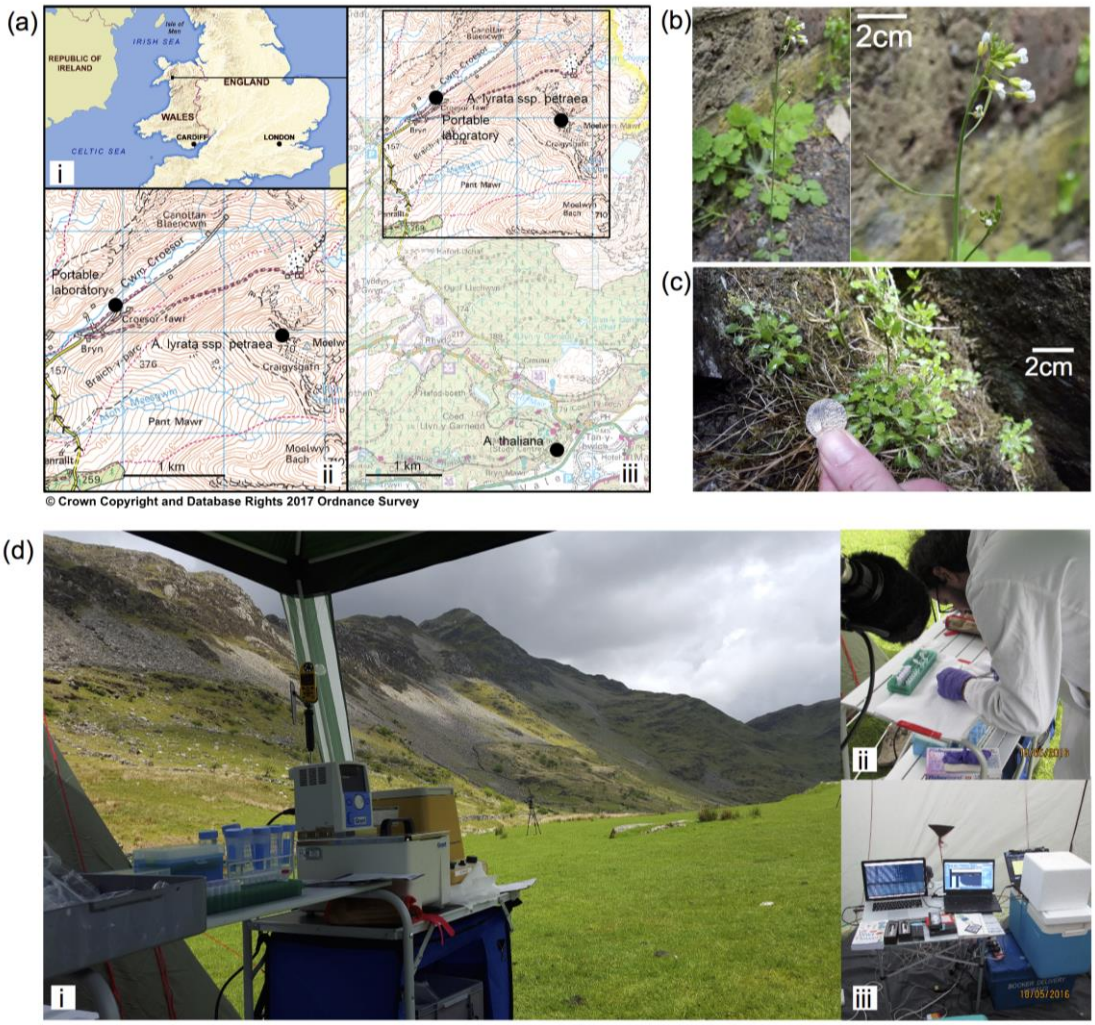

Figure 1: Logistics and scope of field-based sequencing. a, Location of sample collection and extraction, sequencing and analyses in the Snowdonia National Park, Wales. **b,** *Arabidopsis thaliana*. **c,** *A. lyrata ssp. petraea*. **d,** The portable field laboratory used for the research. Ambient temperatures varied between 7-16ºC with peak humidity >80%. A portable generator was used to supply electrical power.

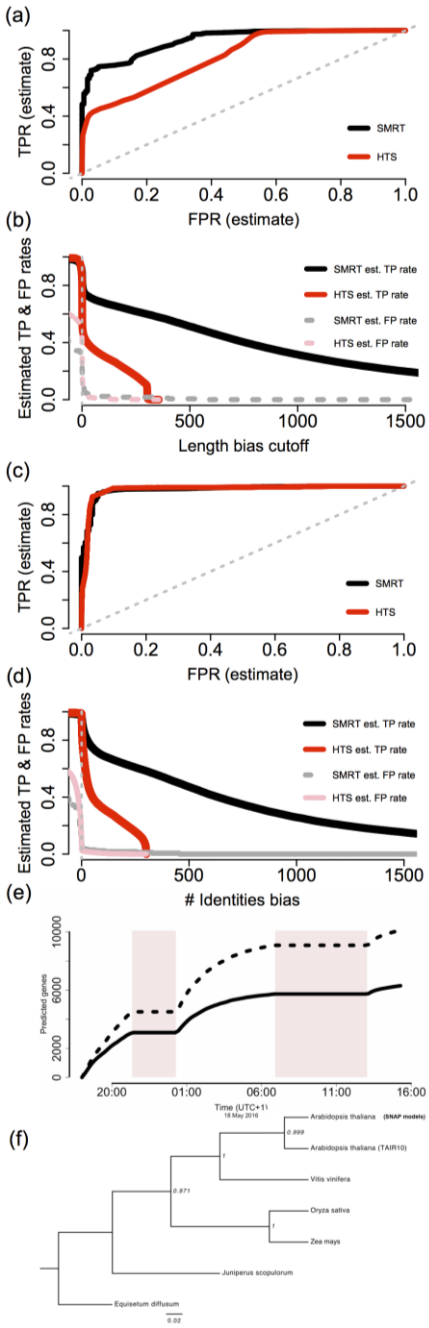

Figure 2: Sample identification and phylogenomics using field-sequenced SMRT data. a-d Orthogonal species identification using BLASTN difference statistics: HTS data (red) and SMRT (black) matched to reference databases via BLASTN. **a, c** Receiver operating characteristic (ROC; estimated false-positive rate vs. estimated true positive rate) and **b, d** estimated true- (solid lines) and false-positive (dashed lines) rates. **a, b** ΔL_T statistic; **c, d,** ΔL_I statistic. **e,** Accumulation curves for *ab initio* gene models predicted directly from individual *A. thaliana* reads over time. Count of unique TAIR10 genes (solid line) and total number of gene models (dashed line). Shaded boxes represent periods where the MinION devices were halted while the laboratory was dismantled and moved. **f,** phylogenetic tree inferred under the multispecies coalescent from SMRT reads.

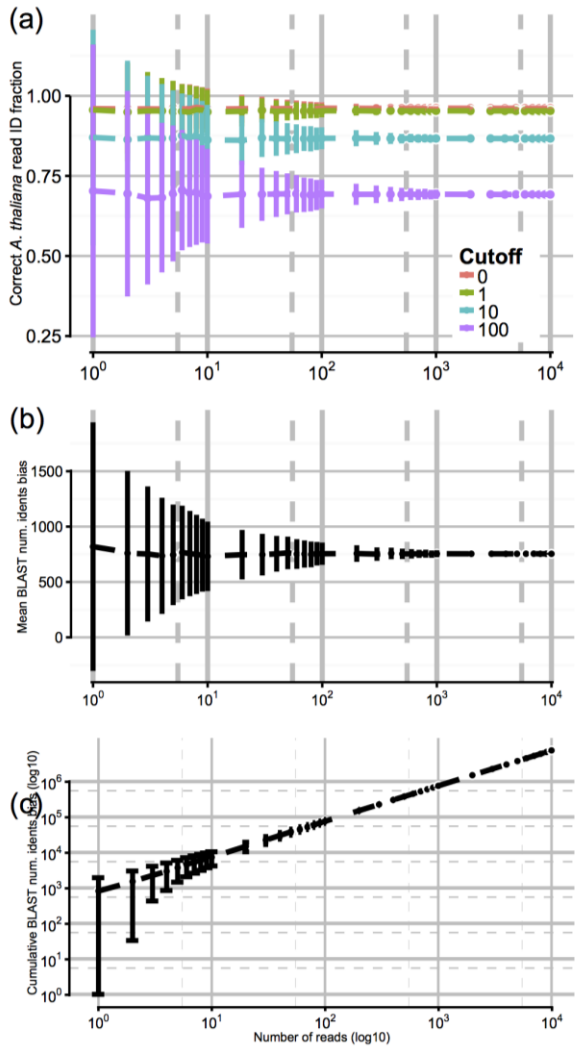

Figure 3: Simulated accumulation curves for rapid species identification by DNA sequencing. 34k pairwise BLASTN hits of *A. thaliana* SMRT reads were subsampled without replacement to simulate an incremental accumulation of data (10^4 reads; 10^3 replicates). For each read the total identities bias (ΔL_I) is the number of identities with the *A. thaliana* reference minus the number of identities with the *A. lyrata* reference. **a)** the proportion of *A. thaliana* reads correctly identified on a per-read basis, classified as *A. thaliana* where ΔL_I > threshold cutoff (0, 1, 10 or 100). **b)** Mean ΔL_I in the simulated dataset rapidly stabilises on the population mean (+754bp, e.g. an average matching read alignment to *A. thaliana* is 754bp longer than to *A. lyrata*). **c)** Cumulative aggregate Δ_{LI}; negative or zero ΔL_I can rapidly be excluded. Typical data throughput rates exceed 10^4 reads per hour of sequencing.

Sequencing platform	MiSeq, 300bp, lab-sequenced	MinION, 1D, field-sequenced
Total reads	9,476,598	91,715
Yield (bp)	2,418,079,888	240,597,532
BLASTN sample identification:		
Zero hits	185,107	58,629
One-way true-positives	2,140,403	10,322
One-way false-positives	53,056	378
Two-way hits[1]	7,098,032	22,386
***de novo* genome assembly:**		
Assembler	**Abyss**	**hybrid-SPAdes**
Coverage[2]	19.49	19.49 + 2.01[3]
Contigs	24,999	10,644
N50	7,853	48,730
Genome fraction (%)	82.0	88.7
Mismatches / indels[4]	518 / 120	588 / 130
Largest alignment	76,935	264,039
CEGMA coding genes	219	245
CEGMA coding fraction	88.31%	98.79%

Table 1: Comparison of BLASTN performance (for sample identification), accuracy, and *de novo* genome assembly of *A. thaliana* by NGS and SMRT sequencing.
Notes: [1]'Two-way' hits matching both reference databases. [2]Approximate coverage inferred using BWA. [3]Hybrid (short- and long-read) assembly; coverage for short and long reads respectively. [4]Per 100kbp.