RESEARCH

Sequencing smart: *De novo* sequencing and assembly approaches for a non-model mammal

Graham J. Etherington [iD]*, Darren Heavens [iD], David Baker [iD], Ashleigh Lister [iD], Rose McNelly, Gonzalo Garcia, Bernardo Clavijo [iD], Iain Macaulay [iD], Wilfried Haerty [iD] and Federica Di Palma [iD]*

The Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK

*Correspondence address. Graham J. Etherington, The Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK; E-mail: graham.etherington@earlham.ac.uk [iD] http://orcid.org/0000-0002-5003-1425; Federica Di Palma, The Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK; E-mail: Federica.di-palma@earlham.ac.uk [iD] http://orcid.org/0000-0002-4394-0102

Abstract

Background: Whilst much sequencing effort has focused on key mammalian model organisms such as mouse and human, little is known about the relationship between genome sequencing techniques for non-model mammals and genome assembly quality. This is especially relevant to non-model mammals, where the samples to be sequenced are often degraded and of low quality. A key aspect when planning a genome project is the choice of sequencing data to generate. This decision is driven by several factors, including the biological questions being asked, the quality of DNA available, and the availability of funds. Cutting-edge sequencing technologies now make it possible to achieve highly contiguous, chromosome-level genome assemblies, but rely on high-quality high molecular weight DNA. However, funding is often insufficient for many independent research groups to use these techniques. Here we use a range of different genomic technologies generated from a roadkill European polecat (*Mustela putorius*) to assess various assembly techniques on this low-quality sample. We evaluated different approaches for *de novo* assemblies and discuss their value in relation to biological analyses. **Results:** Generally, assemblies containing more data types achieved better scores in our ranking system. However, when accounting for misassemblies, this was not always the case for Bionano and low-coverage 10x Genomics (for scaffolding only). We also find that the extra cost associated with combining multiple data types is not necessarily associated with better genome assemblies. **Conclusions:** The high degree of variability between each *de novo* assembly method (assessed from the 7 key metrics) highlights the importance of carefully devising the sequencing strategy to be able to carry out the desired analysis. Adding more data to genome assemblies does not always result in better assemblies, so it is important to understand the nuances of genomic data integration explained here, in order to obtain cost-effective value for money when sequencing genomes.

Keywords: polecat; vertebrate; non-model organism; Illumina; chromium; Bionano; assembly; sequencing

Introduction

Starting in 1990, the Human Genome Project used low-throughput, high-cost Sanger sequencing platforms to create the first draft human genome at a cost of US $300 million. Fast-forward 20 years and the cost of sequencing a human genome has decreased to roughly US $1,000. Short-read technologies producing high-throughput, low per-base cost next-generation sequencing (NGS) means that genomics is no longer restricted to large sequencing consortiums and has opened up the field to even the smallest of research groups. The recently formed Vertebrate Genomes Project aims to produce near-gapless, chromosome-scale phased genome assemblies for

~66,000 extant vertebrate species [1]. The assembly pipeline consists of 60× coverage Pacific Biosciences (PacBio) long-read sequencing, followed by 10x Genomics linked reads, Bionano optical mapping, and Arima Genomics' Hi-C profiles. These long-read technologies provide highly contiguous genome assemblies. Similar consortiums and sequencing initiatives have been formed to sequence a range of target organisms such as Bat1K, Bird10K, Oz Mammals Genomics, and Earth BioGenome Project (including Darwin UK Tree of Life, Colombia EBP, and so forth) [2]. Although such efforts make it possible to achieve highly contiguous, chromosome-level genome assemblies, the cost of generating this amount of data and assembling them is considerable and often only within reach of a few of these consortiums. It is important for smaller independent research groups or initiatives to consider value for money against biological questions as a key factor when planning the generation of genomic sequencing data.

Non-model organisms have the potential to provide new knowledge related to phenotypic and genotypic variation. Through comparative genomics, it is possible to identify how different organisms are related to each other, how they adapt to novel environments, or the genetic basis underlying novel phenotypes. These new findings can be applied to further research, such as in the biomedical and food industries through breeding programs with the development of marker-assisted selection and in conservation biology [3–12].

De novo assembly of endangered species, followed by low-coverage population-level sequencing, provides unprecedented information about the amount of genetic diversity within populations, past and ongoing gene flow between different populations, and the level of inbreeding in small populations [13–17].

However, there are a number of difficulties when working with non-model mammals. First, the genome size is not always known, hampering the assessment of the completeness of the "assembled" genome and of the sequencing depth. Additionally, the availability and quality of the samples used for sequencing non-model organisms are often substandard. Tissue and blood samples are often obtained from wild populations and may need to be acquired from remote locations, delaying the time between collection and DNA extraction. Another common issue relates to samples that may have been stored in collections such as museums, zoos, and tissue collections and subjected to a number of different preservation methods such as freezing, storage in ethanol, or formalin-fixed paraffin-embedded. Many current sequencing technologies rely on high molecular weight DNA with varying optimum molecule lengths (e.g., PacBio HiFi reads 15–20 kb, Bionano >150 kb, and 10x Genomics >50 kb). Degraded DNA, as is commonly observed in samples from wild populations, is usually sub-optimal for use in many advanced sequencing methods. It is therefore difficult, or sometimes impossible, to leverage the full application of these technologies.

Many samples from non-model organisms originate from wild populations that are highly heterozygous, leading to numerous challenges during the assembly step. Allelic differences in a diploid genome generate branches and bubbles in the assembly graph [18]. Even though most graph-based assemblers have functions to search for and remove these structures, high density variation can still make assembly of heterozygous organisms challenging. Conversely, high levels of homozygosity, characteristic of endangered (and typically inbred) species, hamper the efforts of creating phased genome assemblies because the ability to phase haplotypes is dependent on linked sequences spanning polymorphisms. Additionally, non-model organisms vary in their ploidy, chromosome number, repeat con-

tent, sequence composition, and GC content, adding further confounding factors to genome assembly.

The European polecat (*Mustela putorius*, NCBI:txid9668) is a medium-sized carnivore found across Europe and the Middle East. It is purported to be the ancestral species of the domestic ferret (*M. putorius furo*) [19]. Across most of mainland Europe the polecat is in widespread decline [20]. In the United Kingdom, the European polecat has a chequered history. Persecuted to the verge of extinction in the early 1900s, when it was confined to unmanaged forests in central Wales, it has since seen a population increase and is now found throughout Wales and across much of central, south-western, and eastern England [21].

Here, a roadkill sample of European polecat from the Vincent Wildlife Trust collection (VWT 693) was used to assess short-read and long-range *de novo* sequencing strategies for non-model mammals. Comparisons between combinations of PCR-free Illumina libraries, Nextera long mate pair (LMP) libraries, 10x Genomics Chromium libraries, and Bionano optical maps are made to assess optimum sequencing and assembly strategies.

Sequencing Technologies

Short-read sequencing

The market leader in short-read high-throughput NGS is Illumina [22]. Machines produce read lengths of 100 bp and above, and a single Illumina Novaseq run is currently capable of generating 3,000 Gb of read data. An advantage to Illumina sequencing is the generation of paired-end (PE) reads, in which the sequence from both ends of each DNA molecule is synthesized. As the input molecules are of an approximate known length, the acquisition of PE data provides a greater amount of information. Additionally, using a PCR-free library preparation removes bias in genomic coverage previously incorporated by a PCR amplification step in older library preparation procedures [23]. Although requiring input DNA of a degree of magnitude greater than PCR-amplified libraries, PCR-free libraries are expected to capture unbiased coverage of the genome, usually reflected by an increased size of the assembly and less duplication in single-copy regions of the genome compared with PCR-amplified libraries [24]. They also provide superior coverage in GC-rich regions of the genome, enabling access to regions that were previously difficult to sequence [25]. PCR-free Illumina sequencing requires a minimum of 2 μg of genomic DNA (gDNA) at a minimum concentration of 35 ng/μL in 60 μL.

Long mate pair sequencing

Long DNA fragments up to ~40 kb can be sequenced to provide PE reads that bridge long repeats, thus producing longer contiguous genome assemblies as well as characterizing structural variants. Under the Nextera LMP protocol [26], a transposase enzyme attaches 19-bp biotinylated adaptors to both ends of each long DNA fragment. The DNA is then circularized, where the biotinylated ends become joined. The circularized DNA is then fragmented and biotin enrichment is used to process the fragments containing the adaptors that mark the junction. During sequencing, reads are produced from both ends of a fragment, resulting in inward-facing reads that read toward and through the adaptors. Twelve libraries covering a wide range of jump sizes can be constructed using this protocol, thus ensuring production of the best LMP libraries from a given DNA sample. For Illumina Nextera LMP sequencing the Nextclip tool can then be

used to trim adaptors and de-duplicate reads [27]. Nextera LMP sequencing requires a minimum of 4 μg of gDNA for the 12 libraries, at a minimum concentration of 30 ng/μL in 300 μL.

10x Genomics

The Chromium system from 10x Genomics uses oil emulsion and multiple displacement amplification to ligate short molecular barcodes to reads from each fragment of DNA, followed by PE Illumina sequencing [28]. Each fragment receives its own unique barcode, and hence reads with the same barcodes represent clusters of reads from the same region in the genome. These "linked reads" provide the long-range information missing from standard Illumina sequencing and are then used to assemble phased assemblies *de novo*. 10x Chromium libraries require a minimum of 1.25 ng of high molecular weight gDNA at a concentration of 1 ng/μL. In order to take full advantage of the technology, gDNA should >50 kb in length.

Optical mapping (Bionano)

Bionano technology produces optical maps of nicking/restriction enzyme sites across kilobase-long stretches of DNA molecules, providing a high-throughput tool for ordering and orienting contigs of physical maps and validation of genome assemblies [29]. Bionano optical maps can be compared to *in silico* restriction maps produced from an NGS genome assembly for validation purposes, to improve contiguity by assigning the shorter NGS scaffolds to the longer optical maps, and identifying structural variants. A total of 600 ng of raw gDNA at a concentration of 35–200 ng/μL is typically enough DNA to generate ~120 μL of labelled molecules—enough to provide adequate coverage for analysis of a human-sized genome (3 Gb).

Genome contiguity has an effect on what analyses can be achieved (Table 1), so it is important to appreciate the power and limitations of each sequencing strategy and technology.

Materials and Methods

Sequencing

Using the same sample of a roadkill European polecat stored in 100% ethanol, 2 lanes of PCR-free Illumina HiSeq2500 250-bp PE reads (77× coverage), 2 Illumina LMP libraries of size 5 kb (27× coverage) and 7 kb (9× coverage), and 4 lanes of 150-bp PE 10x Genomics Chromium (totalling 85× coverage) using an Illumina HiSeq2500 were generated (Illumina HiSeq 2500 System, RRID:SCR_016383).

We extracted DNA from 4 European polecat samples (all from the VWT collection) and analysed the molecule distribution using an Agilent TapeStation (Supplementary Fig. S3). Sample VWT 693 had the highest concentration of the longest molecules where the distribution of molecule lengths peaked at just under 60 kb and was used for all further sequencing. For this sample 50% of the molecules were >51 kb. The mean molecule length of the remaining 50% of molecules (i.e., those <51 kb) was 15 kb. This was not of good enough quality to generate Bionano data (recommended >150 kb). Because the domestic ferret and its polecat ancestor diverged only ~2,000 years ago and they fully interbreed, we do not expect substantial divergence and structural differences between the 2 species [19, 30–34]. Therefore, the original sample used for the domestic ferret genome assembly [35] was obtained and 1 chip of Bionano Genomics optical

genome maps was generated (Saphyr, RRID:SCR_017992). This was used to create Bionano hybrid-scaffold assemblies for the European polecat genomes assembled with the previously mentioned short-read data, using the Bionano Solve software [36]. We generated 664 Gb of Bionano data, with an N50 size of 185 kb and a contig coverage of 261×. Of this, 40% of the molecules aligned back to the Bionano *de novo* assembly, leaving an effective coverage of 110×. A more detailed description of the library preparation methods can be found in the Supplementary Methods, and the protocols are also available in protocols.io [37].

Assemblies

Ten different genome assemblies were generated as summarized in Fig. 1 (with additional information in Supplementary Table S1), and detailed as follows:

Assembly A1 (w2rap)

The PCR-free Illumina reads from polecat were assembled using the w2rap-contigger [38]. The w2rap-contigger (w2rap) originated from a fork of the popular DISCOVAR de novo program (Discovar, RRID:SCR_016755) [39], and then a number of improvements were made to reduce memory usage and processing time, enhance parameterization, improve repeat resolution, and increase accuracy and contiguity. It also benefits from requiring less computational resources than other popular assemblers such as ALLPATHS-LG (ALLPATHS-LG, RRID:SCR_010742) [40]. w2rap is predominantly a contig assembler—reads are used to construct an assembly graph, which is then traversed to create a contig assembly. A final step involves using the PE information to scaffold contigs not joined during the initial assembly process. Using w2rap, 4 different assemblies were created using a range of k-mers ($k = 180, 200, 224,$ and 240), and simple assembly stats were run to examine contiguity across the assemblies (for all contigs and filtering for contigs > 1 kb). From these statistics, the assembly constructed with $k = 224$ was selected as the final assembly.

Assembly A2 (w2rap + lmp)

We analysed the distribution and coverage from the 12 Nextera LMP libraries and selected the 5- and 7-kb libraries owing to their tight distribution and higher coverage, when compared to the other 10 libraries. Using SSPACE (SSPACE, RRID:SCR_005056) [41], the 5- and 7-kb Nextera LMPs were used to scaffold the w2rap assembly from assembly A1. For all SSPACE LMP assemblies the reads were used only for scaffolding and not for contig extension.

Assembly A3 (10x)

The 10x Genomics Chromium library was assembled using the 10x Genomics Supernova software [28], using default parameters. Default parameters automatically cap the number of reads to 1,200 M, which, after trimming and filtering, resulted in an effective coverage of 52.18× with a mean molecule length of 38.42 kb. Similar to w2rap, Supernova creates an initial contig assembly but then scaffolds using the molecule-specific barcode information in the reads to join contigs known to be from the same molecule [28]. The output style of the resulting assembly was "pseudohap," which creates 1 haplotype per scaffold at random.

Assembly A4 (10x + lmp)

SSPACE was used with the 5- and 7-kb Nextera LMPs to scaffold the 10x assembly generated in assembly A3. As in assembly A2,

Table 1: Information regarding the possible resolution for various *de novo* genome sequencing technologies

Assembly resolution	Paired-end	Paired-end + long mate pair	Bionano	10x Genomics
Gene content	Yes	Yes	No	Yes
Gene order	Yes	Yes	No	Yes
Repeat spanning	No	Yes	Yes	Yes
Structural variants	No	Yes	Yes	Yes
Haplotype resolution (phased genomes)	No	No	No	Yes

When planning a genome assembly project, it is important to understand the strengths and limits of the various sequencing strategies available.
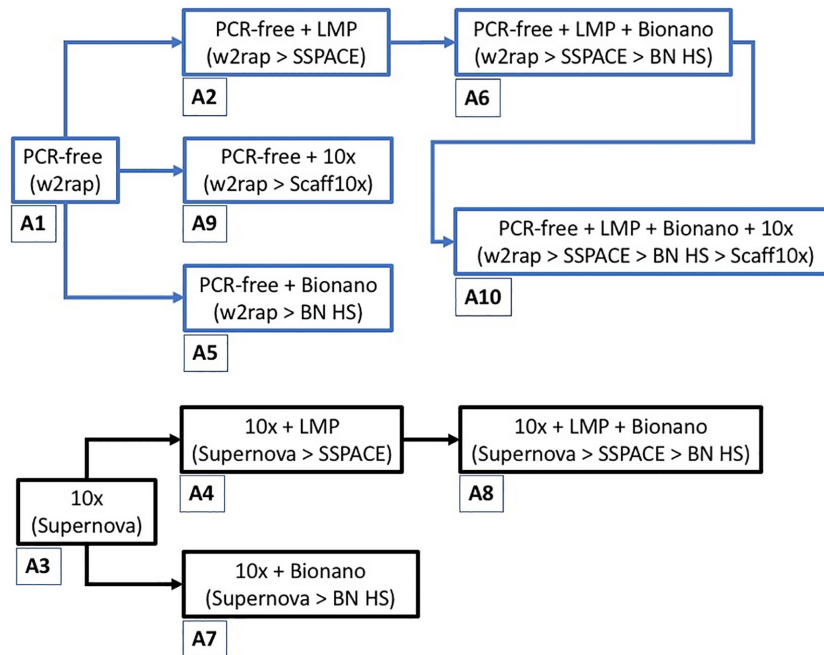
Figure 1: Ten different assembly strategies using a variety of different data types: PCR-free Illumina short-read ("PCR-free"), long mate pair ("LMP"), 10x Genomics Chromium library ("10x"), and Bionano Genomics optical maps ("Bionano"). The blue-boxed assemblies all originate from the same PCR-free w2rap assembly (A1), and the black-boxed assemblies all originate from the same 10x Genomics Supernova assembly (A3). Information in parentheses refers to assembly software pipeline, and assembly numbers are annotated below each assembly.

the LMP reads were used only for scaffolding and not for contig extension.

The Bionano data were assembled *de novo* and then were used to position and orient scaffolds from previous assemblies, creating a Bionano hybrid-scaffold as follows:

Assembly A5 (w2rap + bionano). Bionano hybrid-scaffolding with w2rap assembly (Assembly A1).
Assembly A6 (w2rap + lmp + bionano). Bionano hybrid-scaffolding with the w2rap + lmp assembly (Assembly A2).
Assembly A7 (10x + bionano). Bionano hybrid-scaffolding with the 10x assembly (Assembly A3).
Assembly A8 (10x + lmp + bionano). Bionano hybrid-scaffolding with the 10x + lmp assembly (Assembly A4).

Finally, the 30× coverage of 10x Genomics data (from the same data generated for assembly A3, henceforth referred to "10x-scaffolding") was used to scaffold 2 assemblies using the scaff10x program from Phusion2 [42], as follows:

Assembly A9 (w2rap + 10x). The w2rap-only assembly (Assembly A1) with 10x-scaffolding.

Assembly A10 (w2rap + lmp + bionano + 10x). The w2rap + lmp + bionano assembly (Assembly A6), with 10x-scaffolding.

Analyses
Genome contiguity

For each genome assembly, a number of assembly statistics, such as contig N50, scaffold N50, the number of scaffolds greater than given lengths, and scaffolded genome size were calculated. To calculate contig N50, any scaffolded contigs that were joined by ≥25 Ns were broken. The percentage of the genome contained in scaffolds >25 kb (the average length of a vertebrate gene [43]) and the number of scaffolds >39 Mb (the length of the smallest chromosome in a recent chromosome-scale assembly of a closely related mustelid [44]) were also calculated.

k-mer analysis

The K-mer Analysis Toolkit (KAT) version 2.3.4 [45] was used to examine *k*-mers across reads and assemblies. KAT enables users

to assess levels of errors, bias, and contamination at various stages of the assembly process. Using the KAT "comp" program with a k-mer size of 31, k-mers in the PCR-free Illumina reads were compared with those in the resulting assemblies (omitting the Bionano assemblies because this technology adds negligible sequence content), and for each assembly, the k-mer spectra were plotted.

Gene content

BUSCO (v3.0.2) was used to search for single-copy orthologs in each assembly (BUSCO, RRID:SCR_015008) [46]. BUSCO reports the number of single-copy orthologs discovered in the input assembly and categorizes them as "complete," "single-copy," "multi-copy," or "fragmented." Mammalia_odb9 was used for the "lineage" parameter in BUSCO and "human" for the Augustus species parameter.

Repeat content

To examine repeat content and compare how repeats were resolved in each genome assembly, RepeatMasker v 4.0.7 with library dc20170127-rb20170127 (RepeatMasker, RRID:SCR_012954) [47] was used (with default values) to identify repeat families in each assembly, using all Carnivora-specific repeats. As well as identifying repeat sequences, the mean deletion, insertion, and divergence for each family were also calculated, as well as the mean values overall. Mean divergence is calculated as "mismatches/(matches + mismatches)" between queries and matches for all repeats.

Assembly errors and misassemblies

REAPR (REAPR, RRID:SCR_017625) [48] was used to evaluate the accuracy of each genome assembly by separately mapping PCR-free PE and LMP reads back to each assembly. The fragment coverage distribution (FCD) error for each assembly was calculated. FCD is the fragment depth from only the reads that are mapped to a given base of a fragment. The FCD error is the difference between the theoretical and observed FCD and is used to identify assembly errors in the regions containing a run of high FCD errors. Mapping information such as the FCD and insert size distribution is analysed to locate misassemblies as well as more local per-base accuracies. The "smalt map" option in REAPR was used, which uses SMALT (SMALT, RRID:SCR_005498) [49] to align the PCR-free PE and LMP reads back to each assembly, utilizing the option to map PE reads independently. This ensures that read pairs are not artificially forced to map as proper pairs within a given insert size. REAPR was then used to identify perfectly and uniquely mapped reads in the PE PCR-free alignment, to accurately call error-free bases in the assembly, and further used the LMP reads to identify features consistent with misassemblies. Error-free bases have ≥5× perfect and unique coverage of PE reads. REAPR summary scores were calculated for each assembly by multiplying the number of error-free bases by the square of the REAPR broken scaffold N50 length, and then dividing by the original scaffold N50, i.e., "No. error-free bases ∗ (broken N50²/assembly N50)." This test was first used to evaluate genome assemblies in the Assemblathon series [43] and rewards local accuracy, overall contiguity, and correct scaffolding of an assembly. To independently assess the performance of each data type for scaffolding, the numbers of REAPR breaks were compared between the w2rap-only assembly (A1) and that assembly scaffolded with 1 data type, namely, LMP (A2), Bionano

(A5), and 10x (A9). The same analyses were also performed using the 10x-only assembly (A3).

Value for money

Cost is a huge factor in research and ultimately affects decisions made regarding the technologies used. A metric was created to reflect "value for money" by estimating the cost of each assembly and the N50 achieved. This metric is provided as N50/$1,000 and calculated for contig N50, scaffold N50, and the REAPR broken scaffold N50.

Ranking assemblies

Each assembly was given a rank score according to its position in each of the 7 metrics. The top-placed assembly that performed best in a given metric was given a rank score of 10, the second-placed assembly was given a rank score of 9, and so on, down to the bottom-placed assembly, which was given a rank score of 1.

Assemblies were ranked for the following metrics:

1. Scaffold N50
2. REAPR broken scaffold N50
3. Contig N50
4. Percentage of genome represented by scaffolds >25 kb
5. Single-copy BUSCO orthologs
6. REAPR summary score
7. REAPR broken scaffold N50/$1,000

z-scores

z-scores were used to combine scores from datasets with different means, ranges, and standard deviations and have the benefit of rewarding/penalizing those assemblies with exceptionally high/low scores in any 1 metric. The influence of each of the 7 metrics was tested by removing each metric in turn and recalculating the z-score for each assembly. These recalculations were then used to produce error bars for the final z-score figure, by providing the minimum and maximum z-score that might have occurred if any combination of 6 metrics was used.

Results
Assembly contiguity and connectivity

Assembly statistics
After assembling the 10 genomes as described in Fig. 1, a number of metrics were calculated for each assembly to examine contiguity and connectivity, measured by the lengths and distribution of the scaffolds within each assembly (Table 2). The mean assembly size for all genomes was 2.52 Gb, slightly larger than the 2.41-Gb assembly of the domestic ferret [35]. The 10x-based assemblies erred on having smaller genome assembly sizes (2.46–2.50 Gb), with the larger assemblies (2.47–2.66 Gb) being from the PCR-free Illumina-based assemblies.

Contig N50 for the assemblies varied between 183 and 271 kb. Scaffold N50 for the assemblies varied between 300 kb and 21 Mb. The increase from contig N50 to scaffold N50 varied greatly (Fig. 2). The addition of LMP data to an initial short-read assembly had a varying effect. On the relatively fragmented w2rap assembly (A1), the addition of LMP reads led to an almost 9-fold increase of the scaffold N50, but adding LMPs to the more contiguous 10x assembly (A3) resulted in a 2-fold increase. This is not unexpected because the N90 value for the 10x assembly (800 kb) is 20 times greater than that of the w2rap assembly (40 kb);

Table 2: Genome assembly statistics (for sequences >1 kb) for all assemblies

No.	Assembly	No. scaffolds (%)			% genome ≥25 kb	Longest scaffold (Mb)	Contig N50 (kb)	Scaffold N50 (Mb)	Assembly size (Gb)
		>100 kb	>1 Mb	>39 Mb					
A1	w2rap	6,290 (10.7)	176 (0.3)	0	94.9	2.52	182.93	0.30	2.47
A2	w2rap + lmp	1,680 (4.9)	682 (2.0)	0	94.8	15.65	271.16	2.62	2.60
A3	10x	1,023 (3.9)	501 (1.9)	0	93.3	32.15	207.98	5.26	2.46
A4	10x + lmp	669 (4.2)	346 (2.2)	3	94.7	58.16	210.72	10.33	2.50
A5	w2rap + bionano	4,361 (7.7)	626 (1.1)	0	93.8	6.89	182.93	0.85	2.66
A6	w2rap + lmp + bionano	990 (3.0)	468 (1.4)	0	94.8	34.30	271.16	5.73	2.60
A7	10x + bionano	604 (2.3)	336 (1.3)	3	97.5	46.79	207.98	10.84	2.48
A8	10x + lmp + bionano	409 (2.6)	218 (1.4)	9	97.6	104.38	210.72	21.01	2.50
A9	w2rap + 10x	1,097 (2.4)	467 (1.0)	0	97.6	35.44	182.93	5.58	2.47
A10	w2rap + lmp + bionano + 10x	447 (1.4)	235 (0.7)	6	97.5	65.13	271.16	14.05	2.60

Percentage scores refer to percentage of scaffolds greater than a given threshold. The quantity 39 Mb is the size of the smallest chromosome in a recent chromosome-scale assembly of a closely related mustelid and hence an indication of the number of chromosome-sized scaffolds. A more thorough list of genome statistics can be found in Supplementary Table S2.
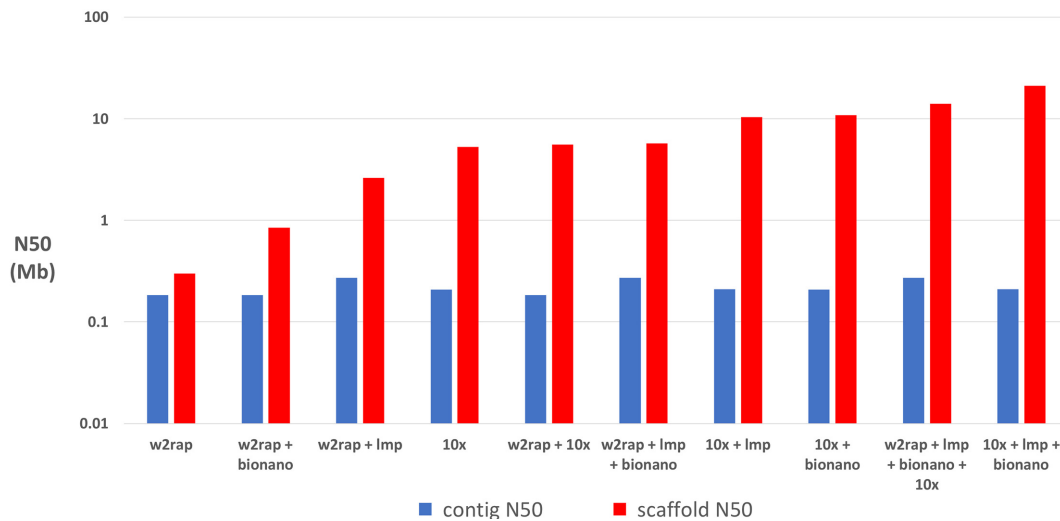
Figure 2: Log-scale lengths of contig N50 (blue) and scaffold N50 (red) of all 10 assemblies, sorted (left to right) by scaffold N50.

hence, the chance of mate pairs spanning the same contig and not adding to the contiguity of the assembly is much higher in the already contiguous 10x assembly. The addition of Bionano data to assemblies leads to a similar scaffold N50 increase across all assemblies, namely, between a 2.0- and 2.8-fold increase. Finally, 10x-scaffolding data were added to scaffold assembly A1 (w2rap) and assembly A6 (w2rap + lmp + bionano). As might be expected, the effect of 10x scaffolding data on less contiguous genomes was greater than that on more contiguous genomes. There was an 18.6-fold increase in N50 between assembly A1 (w2rap) and assembly A9 (w2rap + 10x), whereas the increase in N50 between assembly A6 (w2rap + lmp + bionano) and assembly A10 (w2rap + lmp + bionano + 10x) was less contrasting at 2.5-fold.

Generally speaking, assemblies created with 1 or 2 data types, where 1 of the data types was Illumina short reads, showed the smallest increase from contig N50 to scaffold N50 (Fig. 2).

Assembly errors and misassemblies

REAPR was used to assess the accuracy of the polecat genome assemblies by looking at low-quality regions, breakpoints (Ta-

ble 3), and summary scores (Fig. 3). The percentage of error-free bases for each assembly varied between 76.05 and 85.9%. All the w2rap-based assemblies were on the low end of the scale (76.05–81.09%), whilst 10x-based assemblies were on the high end (84.65–85.9%). Conversely, there was a trend for w2rap-based assemblies to be less affected by misassemblies (excluding those with 10x scaffolding). Their REAPR broken N50 size reduced between 2 and 64%, whilst 10x-based assemblies reduced in N50 size between 68 and 91%. A similar pattern is seen with the number of FCD errors, where all w2rap-based assemblies (bar A10, with 10x scaffolding) have <8,214 FCD errors and all 10x-based assemblies have ≥9,095 errors.

Finally, the performance of each technology was independently assessed for scaffolding by comparing the number of REAPR breaks between the w2rap assembly (A1) and those scaffolded with only 1 data type (LMP, Bionano, and 10x scaffolding) (Table 4). After accounting for the 2,756 breaks introduced by REAPR in the w2rap-only assembly (A1), it was found that Bionano (assembly A5) clearly performed best, containing only 729 more breaks than the original assembly (A1). Conversely, LMP (6,843 more breaks) and 10x scaffolding (7,353 more breaks) data types had ≥9 times more breaks introduced by REAPR than Bio-

Table 3: REAPR statistics showing the percentage of error-free bases in the assembly, N50s before and after breaking at breakpoints, the percentage decrease in scaffold N50 after breaking, and the fragment coverage distribution (FCD) errors including errors across gaps

No.	Assembly name	% error-free	Original N50 (Mb)	REAPR broken N50 (Mb)	% reduction	FCD errors
A1	w2rap	80.83	0.3	0.29	2	6,065
A2	w2rap + lmp	79.10	2.61	1.13	57	8,213
A3	10x	85.90	5.26	1.69	68	11,379
A4	10x + lmp	85.35	10.33	1.86	82	9,095
A5	w2rap + bionano	76.05	0.85	0.52	38	4,523
A6	w2rap + lmp + bionano	78.38	5.73	2.06	64	7,392
A7	10x + bionano	84.65	10.84	2.00	82	13,068
A8	10x + lmp + bionano	84.75	21.00	1.86	91	11,531
A9	w2rap + 10x	81.09	5.58	0.57	90	7,601
A10	w2rap + lmp + bionano + 10x	77.80	14.05	1.75	88	9,488
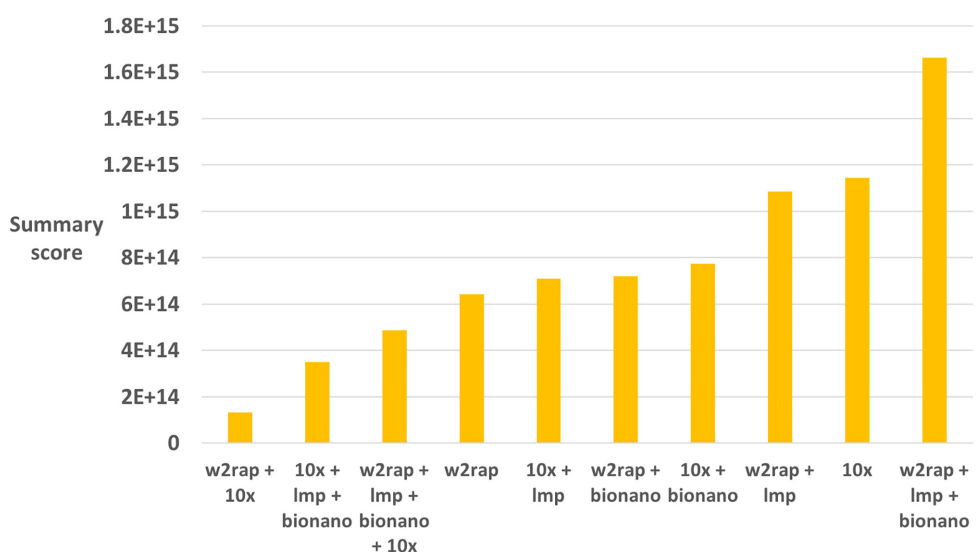
Figure 3: REAPR summary scores for each polecat assembly. REAPR summary scores were calculated for each assembly by multiplying the number of error-free bases by the square of the REAPR broken scaffold N50 length and then dividing by the original scaffold N50.

Table 4: Comparison of the number of breaks introduced by REAPR for each of the technologies used to scaffold the w2rap-only assembly (A1)

No.	Assembly name	Assembled sequences	No. sequences after breaking	REAPR breaks
A1	w2rap	929,245	932,001	2,756
A2	w2rap + lmp	887,887	897,486	9,599 (6,843)
A5	w2rap + bionano	927,316	930,801	3,485 (729)
A9	w2rap + 10x	916,014	926,123	10,109 (7,353)

The number of breaks in parentheses represents the number of breaks after accounting for the 2,756 breaks introduced into the comparison assembly (A1).

Table 5: Comparison of the number of breaks introduced by REAPR for each of the technologies used to scaffold the 10x assembly (A3)

No.	Assembly name	Assembled sequences	No. sequences after breaking	REAPR breaks
A3	10x	26,253	31,505	5,252
A4	10x + lmp	16,018	24,055	8,037 (2,785)
A7	10x + bionano	25,834	31,147	5,313 (61)

The number of breaks in brackets represent the number of breaks after accounting for the 5,252 breaks introduced into the comparison assembly (A3).

Assembly completeness

k-mer content

"KAT comp" [44] was used to compare k-mers in the Illumina PCR-free reads with k-mers in the non-Bionano assemblies (A1–A4 and A9). "KAT plot" was then used to visualize the output (Fig. 4 and Supplementary Fig. S1). The plots all show a similar distribution of k-mers. The black distribution at the start of the x-axis represents sequencing errors in reads, and its increased width represents an increased number of errors in the reads. The

nano. A comparison was made between the number of breaks (5,252) in the 10x assembly (A3) to the 10x + lmp assembly (A4) and the 10x + bionano assembly (A7) (Table 5). A similar pattern as above was found, with the LMP assembly having 2,785 more breaks than the 10x assembly but with the Bionano assembly having only 61 more breaks, again demonstrating the accuracy of Bionano for scaffolding.

k-mers in these reads have not been incorporated into the final assembly. The extension of the black line along the x-axis (up to a *k*-mer multiplicity of 40 on the x-axis) represents collapsed haplotypes, where *k*-mers from 1 side of a bubble in the assembly graph have been removed to construct a linear path through the graph. Any extension of the black line along the x-axis into the main red distribution (>40 *k*-mer multiplicity) represents a small number of high-copy *k*-mers in the reads missing from the assembly. The red area in all graphs represents a normal distribution of *k*-mers found in the reads and occurring once in the assembly. The absence of any additional colours, representing *k*-mers appearing once in the reads but multiple times in the assembly, reflects the presence of only unique content throughout the assembly, with *k*-mers in the reads occurring no more than once in the assembly.

Despite all of the assemblies being compared to the PCR-free Illumina short reads, virtually the same distribution of *k*-mers between the reads and assemblies was observed, showing an almost identical distribution of *k*-mers from all the different read sequences and their resulting assemblies. The KAT-plots involving 10x assemblies (Supplementary Fig. S1C and D) are also characterized by some high-copy read *k*-mers missing from the assemblies. This suggests that the minimum size of contigs included in the final assembly (1 kb) may be too high. This may also explain the slightly smaller assembly sizes obtained from the 10x-based assemblies when compared to the w2rap-based assemblies (Table 2).

Gene content
BUSCO was used to look at single-copy orthologs in the assemblies (Fig. 5) and examine the number of single-copy, duplicated, fragmented, and missing orthologs. The number of complete and single-copy orthologs reconstructed varied from 3,748 (91%) in Assembly A1 and 3,885 (95%) in Assembly A10. Of the 4,104 mammalian orthologs examined 3,603 (88%) were found in single copies across every assembly, 65 were missing, 30 were fragmented, and 21 duplicated across all assemblies. The w2rap assembly (A1) had the highest number of missing orthologs (117) and fragmented orthologs (198), probably owing to the fragmented nature of the assembly. Adding scaffold datasets improved ortholog reconstruction in all but 1 case, where we found that adding only Bionano (A7) or only LMP (A4) data to the 10x assembly (A3) fragmented a few orthologs (6 and 7, respectively) although adding Bionano data to the 10x + lmp assembly (A4, resulting in assembly A8) increased the number of single-copy orthologs by connecting 13 fragmented ones. Generally, the addition of LMP data had the least beneficial effect on ortholog reconstruction, followed by Bionano, and then 10x scaffolding. Indeed, the lowest-ranking assembly (A1) jumped to the second-highest-ranking assembly merely by the addition of 10x scaffolding data, which reduced the number of fragmented orthologs from 198 to 94.

Repeats
RepeatMasker was used to look at Carnivora-specific repeat content in the assemblies. Long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) were by far the most common classes of repeats, and these are concentrated on here. A very similar picture was found among all datasets. The percentage of the genome assemblies that was masked for repeats varied between 35.82 and 39.49%, with SINEs varying between 8.40 and 9.81%. The w2rap-based assemblies were on the lower end of both of these scales, with the 10x-based assemblies on the higher end.

A slightly different pattern was found when examining LINEs, the composition of which varied between 19.20 and 20.73%. In these repeats the w2rap-based assemblies clustered at the lower end of the scale, with the exception of assembly A1 (w2rap) and assembly A9 (w2rap + 10x), which grouped with the 10x assemblies at the higher end of the scale.

Mean divergence between each assembly and all repeat families was also calculated. It was found that the divergence between assemblies was small (24.52–24.60), with no defined grouping of the assemblies by data type. This suggests an overall similar ability of each data type to accurately reconstruct repeat sequences (Table 6).

Value for money
The N50/$1,000 metric (see Methods) was calculated in order to provide a metric for value for money when considering the choice of technology and the return on money spent (Fig. 6). For contig N50/$1,000, the w2rap-based assemblies provide by far the best value for money, with the exception of those with 10x scaffolding. Value for money decreases as more data are added to the w2rap assemblies. So, for contig assemblies a basic PCR-free Illumina short-read assembly provides the best value for money.

However, when looking at scaffold N50/$1,000, the trend changes. Five of the 6 lowest-scoring assemblies constitute w2rap-based assemblies, generated with between 1 and 3 data types. The 10x-based assemblies show better performance when looking at scaffold N50/$1,000, with 3 of the 4 highest-scoring assemblies being 10x-based. The difference in scaffold N50 between the w2rap-based and 10x-based assemblies might be expected because the short-read Illumina data do not contain the additional molecule-specific linked-read information present in 10x data. Another trend is that adding more scaffolding data to a "base" assembly (A1 and A3) increased the scaffold N50/$1,000. Hence, adding more data to increase scaffold contiguity provides value for money, although one must judge whether the amount of increase justifies the extra cost.

Ranking assemblies
Assemblies were ranked on a number of key metrics (see Methods), allocated a final rank score (Supplementary Fig. S2), and z-scores were calculated for each assembly (Fig. 7). The order of ranking in the rank-scoring method is similar to the z-score ranking, although when the z-scores are calculated the 10x assembly (A3) and w2rap + lmp + bionano + 10x assembly (A10) both rank 2 places higher, with the 10x + lmp + bionano assembly (A8) ranking a place lower. The z-scores provide a better assessment of the performance of each assembly across all the metrics and not just their position in the final ranking. The general trend was that the more data included, the higher the assembly ranked, although this was not always the case. For example, the second-highest-ranked assembly was A10, the only assembly with 4 different data types (w2rap + lmp + bionano + 10x), but the highest-placed assembly was assembly A6 (assembly A10, but without the final 10x scaffolding data). Also, the 10x-only assembly (A3) ranked 1 place higher than the 10x + lmp assembly (A4).

Discussion

Although chromosome-scale assemblies are now achievable, it is often not possible or necessary to assemble the genomes of non-model organisms to such precision. A number of difficulties are faced when sequencing and assembling non-model organ-
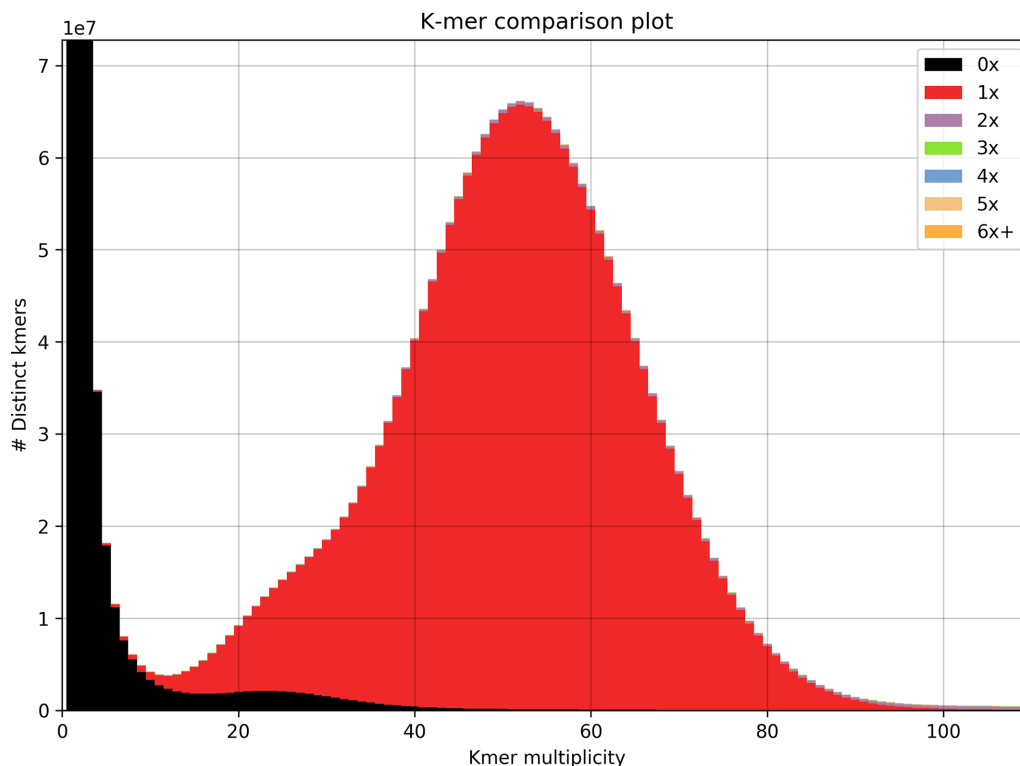
Figure 4: KAT k-mer plots comparing k-mer content of Illumina PCR-free reads with w2rap assembly (A1). The black area of the graph represents the distribution of k-mers present in the reads but not in the assembly, and the red area represents the distribution of k-mers present in the reads and once in the assembly.
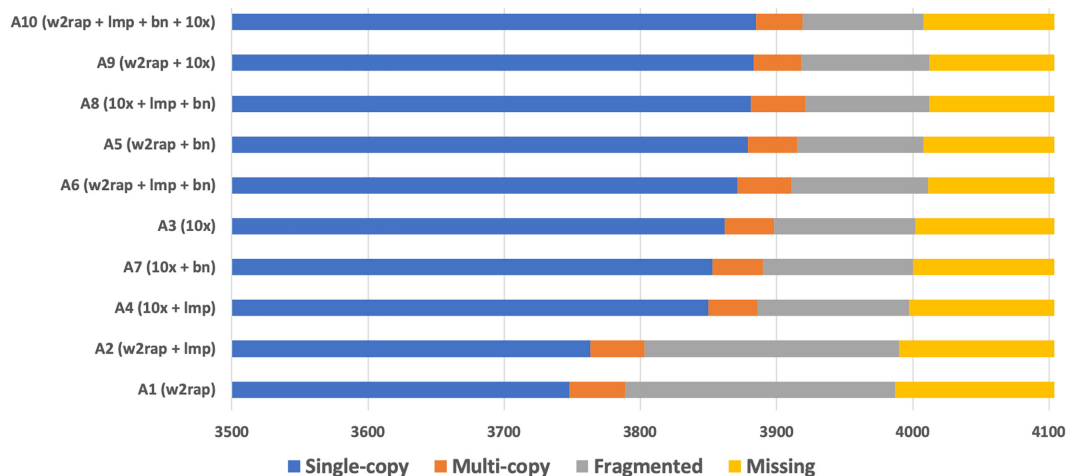

Figure 5: Number of single-copy (blue), duplicated (orange), fragmented (grey), and missing (yellow) orthologs from BUSCO. To visualize the number of duplicated, fragmented, and missing orthologs, the first 3,500 single-copy orthologs present in each assembly are truncated.

isms. Genome size (and as a consequence, sequencing depth), chromosome number, sequence composition, and GC content are often unknown or inaccurate, the species can be highly heterozygous, and samples are often degraded. We address some of these factors and identify which sequencing and assembly strategies are required to answer various biological questions. PCR-free Illumina short-read, 10x Genomics linked-read, long mate paired read, and Bionano optical maps were generated from a roadkill European polecat to create 10 different genome assemblies, using different combinations of the data. The assemblies were assessed using a range of tools and ranked using 7 key metrics. We find that although some genomes assemble to high contiguity, this is often at the expense of accuracy and it is often not necessary to spend additional funds on increasing contiguity to answer biological questions.

Assembly contiguity and connectivity

As a general rule, adding more data to an assembly increases the contiguity (scaffold N50). This was observed in the assemblies here, with each assembly having a higher scaffold N50 than any "parent" assembly before it. The linked reads from 10x

Table 6: Repeat content of assemblies

No.	Assembly name	% Masked	% SINEs	% LINEs	Mean divergence
A1	w2rap	38.31	8.99	20.53	24.52
A2	w2rap-lmp	37.24	8.75	19.95	24.56
A3	10x	39.49	9.81	20.73	24.53
A4	10x-lmp	39.10	9.71	20.52	24.54
A5	w2rap-bionano	35.82	8.40	19.20	24.52
A6	w2rap-lmp-bionano	36.79	8.64	19.71	24.52
A7	10x-bionano	39.11	9.72	20.53	24.60
A8	10x-lmp-bionano	38.89	9.66	20.41	24.58
A9	w2rap 10x	38.32	8.99	20.54	24.55
A10	w2rap_lmp_bionano_10x	36.79	8.64	19.70	24.53

% Masked refers to the amount of the genome masked for all repeats, % SINEs and % LINEs reflect the percentage of the genome found to contain each of these classes, and mean divergence is calculated as "mismatches/(matches + mismatches)" between queries and matches for all repeats.
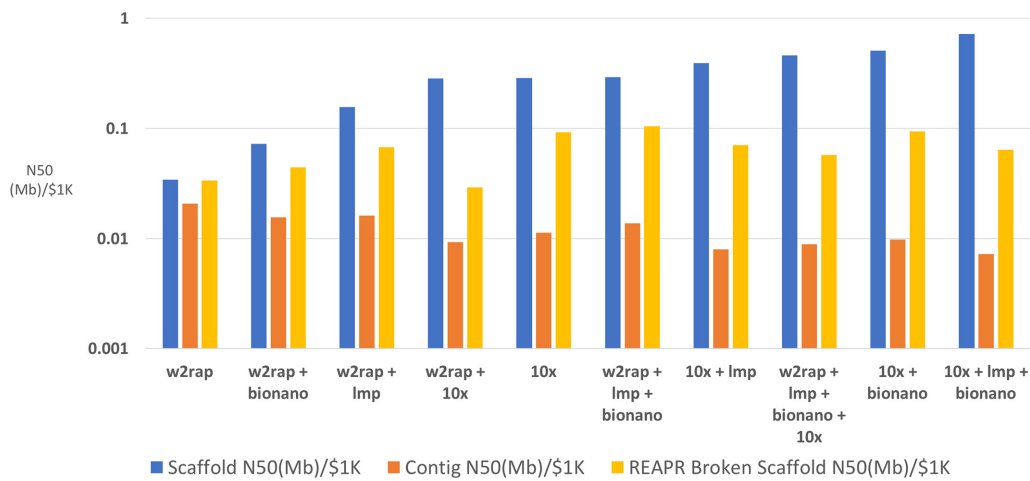
Figure 6: N50/$1,000, providing an estimate to the cost of scaffold (blue), contig (orange), and REAPR broken scaffold (yellow) contiguity for each genome assembly. Assemblies are ranked in order of scaffold N50/$1,000.
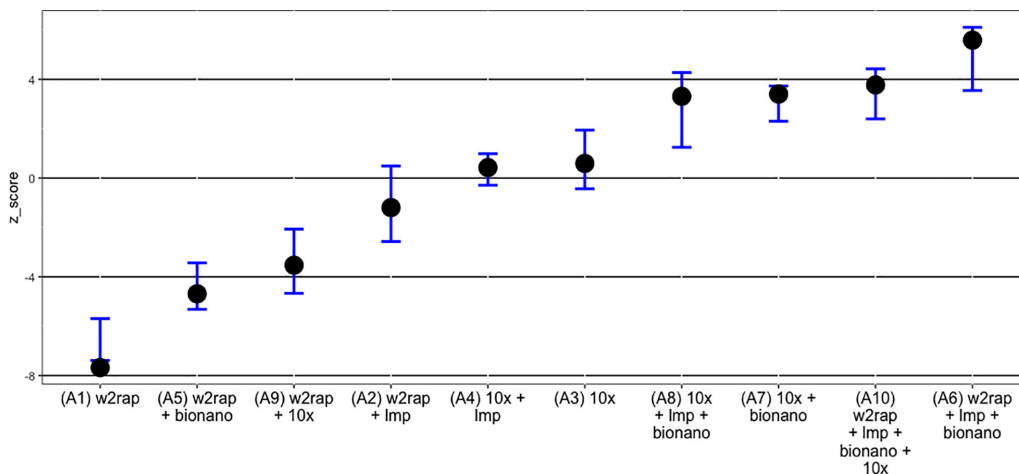
Figure 7: Cumulative z-scores of assemblies (solid black circles). Error bars represent the minimum and maximum cumulative z-score after removing each metric in turn and recalculating the z-score for each assembly. Wide error bars show assemblies that are strongly affected by a given metric. For example, the 10x + lmp + bionano assembly (A8) has a long lower-boundary error bar because it has an exceptionally high scaffold N50 z-score (double that of the next nearest ranking assembly) and hence omitting this metric results in the assembly scoring much lower.

Genomics data constantly outperform the equivalent PCR-free short-read–based assemblies, with the barcoded linked reads acting as an additional scaffolding dataset. The assemblies with the best contig N50 were those based on w2rap + lmp (namely, A2, A6, and A10). For scaffold N50 and percentage of the genome represented by scaffolds >25 kb, 10x + lmp + bionano (A8) provided by far the best contiguity. When REAPR breaks are taken into consideration, the w2rap + lmp + bionano assembly (A6) provides the best scaffold N50, although assembly A8, with the best initial contig N50, is still ranked third. It should be noted that Bionano and 10x (for scaffolding) added no sequence data to the assemblies and hence do not extend the contig lengths or have not connected contigs together without the need to add N's. Those assemblies scaffolded with LMPs however do increase in contig N50, reflecting previously unconnected contigs being joined without N's.

An increase in REAPR summary scores was seen when LMP and Bionano data are added to PCR-free short-read assemblies, but a decrease in summary scores when 10x scaffolding reads are included. For 10x-based assemblies, the addition of extra data leads to a reduction in summary scores. Additionally, 10x-based assemblies tended to have more FCD errors, and the breaking of assemblies at these errors affected 10x-based assemblies to a greater degree than w2rap-based assemblies. Finally, the number of breaks created by REAPR for each scaffolding technology showed that Bionano-scaffolded genomes had significantly fewer breaks than both LMPs and 10x scaffolding. The addition of 10x-scaffolding data led to an overall reduction in summary scores, suggesting that although 10x-scaffolded assemblies provide a good increase in scaffold N50, much of this increase is through misassemblies.

The increase in misassemblies with the addition of extra data is understandable. An initial, 1-technology *de novo* assembly will have all the "easy joins" put together, and most of those will be correct. When a new data type is added, it will have the "difficult joins" to put together, making it very likely that a significant number of these will be incorrect. Bionano performs best at connecting these difficult joins.

Assembly completeness

Gene content usually increased after adding scaffolding data, with the exception of the 10x assembly. Here, adding only LMP data or only Bionano data fragmented a few orthologs, but incorporating both technologies led to an increase in ortholog reconstruction. Other than this it is clear that 10x data perform exceptionally well in the gene space, both for *de novo* sequencing and for scaffolding. Bionano also performs well, with LMP data having the smallest impact.

There was a small amount of difference in repeat content between assemblies. The tendency of 10x assemblies to have a slightly higher percentage of the genome assembled as repeats probably reflects the ability of this technology to better resolve repeats than the standard short-read assemblies, which collapse a large proportion of the repeats. Those repeats that were resolved in assemblies all showed a very similar divergence, regardless of the data types used.

Value for money

For contig assembly, a basic PCR-free short-read assembly provides the best value for money (A1). Adding more data does not increase the contig N50 enough to warrant the extra expense. For scaffold assembly, the story is very different. The 10x + lmp

+ bionano (A8) offers the best value for money. The more data added to an initial assembly, the higher the scaffold N50/$1,000. When REAPR broken assemblies are taken into consideration, the 10x-based and LMP-scaffolded assemblies provide the best value, with the w2rap + lmp + bionano assembly (A6) being ranked top (Fig. 6). Another feature when considering REAPR broken assemblies is the poor performance of 10x scaffolding (A9 and A10). Compared to an Illumina PCR-free library, 10x Chromium libraries are expensive to produce owing to higher cost of the library preparation and the additional hands-on time required associated with the protocol. This increased cost for the 10x Genomics scaffolding data and the high misassembly rate when used as a scaffolding technology means that it scores low in this metric. Nextera LMP libraries are even more expensive to produce than 10x libraries and take a similar amount of preparation time but are less susceptible to misassembly and score higher in this metric.

In summary, when looking at genome contiguity, if contigs are all that are required from a genome assembly, then PCR-free short-read assemblies with no additional data types provide the best value for money. If accurate scaffolds are more important, 10x data, often augmented with LMPs or Bionano, provide good value for money, with Bionano misassembling significantly fewer scaffolds than LMPs.

Ranking assemblies

As expected, the general trend was that the more data included, the higher the assembly ranked, with the addition of Bionano or 10x data providing the most powerful scaffolding technique.

Application to non-model mammals

Although the sample quality used for non-model organisms is often sub-standard, sequencing technologies and software are still successful in assembling these samples into highly contiguous genomes. As a general rule, adding more data to an assembly increases the contiguity (scaffold N50), but the additional expense of incorporating additional data to increase contiguity is not always necessary.

For population genetics approaches, SNP calling, and large multi-species comparisons, basic short-read assemblies such as w2rap (A1) or 10x (A3) provide enough accuracy and contiguity to achieve interpretable results. The 10x assemblies also have the added advantage of haplotype resolution (phased genomes). Where structural variation, long repeat content, gene order, or gene clusters are of importance, an additional scaffolding dataset is often necessary to obtain the required precision for these analyses (A2, A4, A5, A7, and A9), with 10x scaffolding or Bionano being the better data to incorporate if working in the gene space. Examples where this might be important include when dealing with gene clusters of similar genes, such as immune-related gene clusters (e.g., MHC, interleukin, toll-like receptors). When looking at more long-range features, such as genome synteny, Bionano provides additional contiguity. Bionano, however, is dependent on high-quality high molecular weight DNA, which might not be available for many organisms and appears to be the first data type to be hindered by sample degradation. This was apparent from our polecat sample, where the distribution of molecule lengths peaked at <60 kb, and of which the smallest 50% of material had a mean molecule length of 15 kb (Supplementary Fig. S3).

Experimental design

Assemblies with both short contig lengths and a high number of misassemblies can sometimes be found in very heterozygous species. Knowing the distribution of molecule lengths from a sample will provide information about the limitations of which sequencing technologies can be successfully supplied. Researchers can then design their assembly and analysis pipeline to accommodate the limitations of the sample. For example, if the molecule lengths are only in the region of 1 kb, then PCR-free Illumina PE sequencing is the only viable option. Longer molecules, between 10 and 40 kb, allow the preparation of LMP libraries and between 20 and 100 kb permit the inclusion of 10x Genomics data. Beyond that ($\geq$100 kb), Bionano optical maps may also be included. Low-coverage 10x Genomics sequencing has recently been shown to produce a high-quality, cost-effective *de novo* assembly in a non-model mammal [50]. Using 25$\times$ coverage, a *de novo* assembly of the African wild dog produced a reference genome with contig and scaffold N50s of 50 kb and 15.3 Mb, respectively, providing another avenue of assembly approach.

Recently, Hi-C sequencing has also been used to good effect to scaffold genomes of a number of different organisms [51–54]. The technique is based on cross-linking DNA, then digesting the DNA with restriction enzymes. The DNA fragment ends are then re-ligated, which will contain 2 fragments of DNA that were far apart in the genome but still maintaining some degree of physical proximity (e.g., on the same molecule). By sequencing the ends of these fragments, PE reads can be mapped to *de novo* genome assemblies and used to scaffold and order contigs, creating chromosomes-scale assemblies [55, 56]. Although the Hi-C protocol does not specify a minimum molecule length (the protocol is carried out in cells or tissue), it relies on fresh DNA long enough to form distant cross-linked fragments.

Adding long-read data, such as low-coverage PacBio or Nanopore data, will often be the only solution to overcoming complexities such as high heterozygosity or long repeats. Unfortunately, long-read data rely on high molecular weight DNA with long molecules, but as described previously, DNA samples from non-model organisms are often of low quality and the application of these technologies may not be suitable. The quality of sample should reflect the experimental design and assembly pipeline. Development in new DNA extraction (e.g., Nanobind Magnetic Disks [57]) and sequencing technologies may provide access to low quantity and quality of DNA, which may be a potential solution to overcome the sample extraction issues.

As mentioned, with longer molecules, using long-read technologies such as PacBio and Nanopore becomes a possibility, but these require significantly more DNA (>20 ng) to work successfully, as well as being associated with a much higher cost. This overcomes some of the limitations of short-read assemblies, such as characterizing structural variation, sequencing through extended repetitive regions, discriminating paralogous genes, and detecting disease-associated mutations, although with the drawback of requiring high coverage due to the lower base accuracy of long-read sequencing.

Limitations of this study

In this study a combination of 4 different technologies have been used to create 10 different genome assemblies. An exhaustive assessment would produce many more different assemblies, so a choice of what was considered a good representation of all practical combinations was used. Additionally, different assembly software (including versions thereof) may produce slightly different results depending on the algorithms used within them. Finally, test metrics can bias results. For example, the inclusion of more cost-related metrics would bias rankings to favour cheaper assemblies, whereas more contiguity-related tests would bias results for assemblies with higher N50s. The choice of metrics was made to encapsulate genome contiguity, accuracy, error, biologically meaningful content, and cost whilst not unduly biasing the results towards any 1 feature of the assemblies.

Summary

We address how different sequencing and assembly strategies are required to answer various biological questions in non-model mammals. We find that although some genomes assemble to high contiguity, this is often at the expense of accuracy and it is often not necessary to spend additional funds on increasing contiguity to answer biological questions.

Sequencing technologies and assembly software are always progressing, with new sequencing chemistry releases providing longer and more accurate read sequences. Also, novel assembly algorithms promise more contiguous and accurate assemblies. Often each algorithm is dependent on the input of specific data types, with some new assembly software providing more contiguous assemblies at the expense of accuracy. It is important to fully assess the performance of an assembly by using a number of different quality assessment approaches as shown in our study, rather than relying on simple statistics such as scaffold N50, which itself can be biased by the exclusion of shorter sequences from the calculations.

Finally, given the accuracy of PCR-free assemblies and the contiguity of the 10x linked-read technology, if a PCR-free linked-read sequencing technology existed, it would provide accurate, contiguous, and cheap assemblies.

Availability of Supporting Data and Materials

The submission of sequencing data was brokered by the COPO platform [58], funded by the BBSRC (BB/L024055/1), and supported by CyVerse UK, part of the Earlham Institute National Capability in e-Infrastructure. All datasets supporting the results of this article are available in the ENA repository under umbrella project accession No. PRJEB34131. Optical maps, annotations, and other results are available from the *GigaScience* GigaDB repository [59], and protocols are available from protocols.io [37].

Additional Files

Figure S1. KAT *k*-mer plots comparing *k*-mer content of Illumina PCR-free reads with (A) w2rap assembly (A1), (B) w2rap + lmp assembly (A2), (C) 10x assembly (A3), (D) 10x + lmp assembly (A4), and (E) w2rap + 10x assembly (A9). The black area of the graphs represents the distribution of *k*-mers present in the reads but not in the assembly, and the red area represents the distribution of *k*-mers present once in the reads and once in the assembly.
Figure S2. Rank scoring for assemblies. Each assembly was ranked over 7 key metrics. A rank score of 10 was given to the highest-ranking assembly, down to a rank score of 1 for the assembly that performed worst for the given metric.
Figure S3. Agilent TapeStation traces for DNA extractions of VWT samples 448, 557, 639, and 693. The traces show the rel-

ative intensity of the number of molecules against the length of the molecules in base pairs.

Table S1. Ten different assembly strategies using a variety of different data types: PCR-free Illumina short-read (ISR), long mate-pair (LMP), 10x Genomics Chromium library, and Bionano Genomics optical maps.

Table S2. Full assembly statistics for all assemblies.

Supplementary Methods

Abbreviations

BBSRC: Biotechnology and Biological Sciences Research Council; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; FCD: fragment coverage distribution; Gb: gigabase pairs; GC: guanine-cytosine; gDNA: genomic DNA; kb: kilobase pairs; LINE: long interspersed nuclear element; LMP: long mate pair; Mb: megabase pairs; MHC: major histocompatibility complex; NCBI: National Center for Biotechnology Information; NGS: next-generation sequencing; PacBio: Pacific Biosciences; PE: paired end; SINE: short interspersed nuclear element; VWT: Vincent Wildlife Trust.

Competing Interests

The authors declare that they have no competing interests.

Funding

References

1. Pennisi E. New technologies boost genome quality. Science 2017;**357**(6346):10–1.
2. Lewin HA, Robinson GE, Kress WJ, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A 2018;**115**(17):4325–33.
3. Keller I, Wagner CE, Greuter L, et al. Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. Mol Ecol 2013;**22**(11):2848–63.
4. Prufer K, Munch K, Hellmann I, et al. The bonobo genome compared with the chimpanzee and human genomes. Nature 2012;**486**(7404):527–31.
5. Jones FC, Grabherr MG, Chan YF, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 2012;**484**(7392):55–61.
6. Shao C, Niu Y, Rastas P, et al. Genome-wide SNP identification for the construction of a high-resolution genetic map of Japanese flounder (*Paralichthys olivaceus*): applications to QTL mapping of *Vibrio anguillarum* disease resistance and comparative genomic analysis. DNA Res 2015;**22**(2): 161–70.
7. Dodds PN, Rathjen JP. Plant immunity: towards an integrated view of plant-pathogen interactions. Nat Rev Genet 2010;11(8):539–48.
8. Shaffer HB, Gidiş M, McCartney-Melstad E, et al. Conservation genetics and genomics of amphibians and reptiles. Annu Rev Anim Biosci 2015;**3**:113–38.
9. Kohn MH, Murphy WJ, Ostrander EA, et al. Genomics and conservation genetics. Trends Ecol Evol 2006;**21**(11): 629–37.
10. Attard CRM, Beheregaray LB, Sandoval-Castillo J, et al. From conservation genetics to conservation genomics: a genome-wide assessment of blue whales (*Balaenoptera musculus*) in Australian feeding aggregations. R Soc Open Sci 2018;**5**(1):170925.
11. Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. Nat Rev Genet 2010;**11**(10):697–709.
12. Murgarella M, Puiu D, Novoa B, et al. A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. PLoS One 2016;**11**(3):e0151561.
13. Ekblom R, Brechlin B, Persson J, et al. Genome sequencing and conservation genomics in the Scandinavian wolverine population. Conserv Biol 2018;**32**(6):1301–12.
14. Zhao S, Zheng P, Dong S, et al. Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. Nat Genet 2013;**45**(1):67–71.
15. Locke DP, Hillier LW, Warren WC, et al. Comparative and demographic analysis of orang-utan genomes. Nature 2011;**469**(7331):529–33.
16. Miller W, Schuster SC, Welch AJ, et al. Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci U S A 2012;**109**(36):E2382–90.
17. Der Sarkissian C, Ermini L, Schubert M, et al. Evolutionary genomics and conservation of the endangered Przewalski's horse. Curr Biol 2015;**25**(19):2577–83.
18. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;**18**(5):821–9.
19. Blandford PRS. Biology of the polecat *Mustela-putorius* - a literature-review. Mammal Rev 1987;**17**(4):155–98.
20. Croose E, Duckworth JW, Ruette S, et al. A review of the status of the Western polecat *Mustela putorius*: a neglected and declining species? Mammalia 2018, doi:10.1515/mammalia-2017-0092.
21. Croose E. The distribution and status of the polecat (*Mustela putorius*) in Britain 2014–2015. The Vincent Wildlife Trust; 2016.
22. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;**456**(7218):53–9.
23. Kozarewa I, Ning Z, Quail MA, et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 2009;**6**(4):291–5.
24. Clavijo BJ, Venturini L, Schudoma C, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 2017;**27**(5):885–96.
25. Aird D, Ross MG, Chen WS, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 2011;**12**(2):R18.
26. Heavens D, Accinelli GG, Clavijo B, et al. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. BioTechniques 2015;**59**(1):42–5.

27. Leggett RM, Clavijo BJ, Clissold L, et al. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics 2014;**30**(4):566–8.

28. Weisenfeld NI, Kumar V, Shah P, et al. Direct determination of diploid genome sequences. Genome Res 2017;**27**(5):757–67.

29. Hastie AR, Dong L, Smith A, et al. Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. PLoS One 2013;**8**(2):e55864.

30. Costa M, Fernandes C, Birks JD, et al. The genetic legacy of the 19th-century decline of the British polecat: evidence for extensive introgression from feral ferrets. Mol Ecol 2013;**22**(20):5130–47.

31. Davison A, Birks JDS, Griffiths HI, et al. Hybridization and the phylogenetic relationship between polecats and domestic ferrets in Britain. Biol Conserv 1999;**87**(2):155–61.

32. Birks J, Kitchener A. The Distribution and Status of the Polecat *Mustela putorius* in Britain in the 1990s, London: The Vincent Wildlife Trust; 1999;**152**.

33. Volobuev V. Taxonomic status of ferret based on karyological data. Zool J 1974;**53**:1738–9.

34. Sato JJ, Hosoda T, Wolsan M, et al. Phylogenetic relationships and divergence times among mustelids (Mammalia: Carnivora) based on nucleotide sequences of the nuclear interphotoreceptor retinoid binding protein and mitochondrial cytochrome b genes. Zoolog Sci 2003;**20**(2):243–64.

35. Peng X, Alföldi J, Gori K, et al. The draft genome sequence of the ferret (*Mustela putorius furo*) facilitates study of human respiratory disease. Nat Biotechnol 2014;**32**(12):1250–5.

36. Bionano Genomics. Bionano Solve. 2019. https://bionanogenomics.com/support-page/bionano-solve/.Accessed 21 Apr 2016.

37. Etherington GJ, Heavens D, Baker D, et al. Protocols for "Sequencing smart: De novo sequencing and assembly approaches for a non-model mammal." protocols.io 2020. http://dx.doi.org/10.17504/protocols.io.bd3ri8m6.

38. Clavijo BJ, Venturini L, Schudoma C, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 2017;**27**(5):885–96.

39. Love RR, Weisenfeld NI, Jaffe DB, et al. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. BMC Genomics 2016;**17**:187.

40. Gnerre S, Maccallum I, Przybylski D. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 2011;**108**(4):1513–8.

41. Boetzer M, Henkel CV, Jansen HJ, et al. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 2011;**27**(4):578–9.

42. Mullikin JC, Ning ZM. The phusion assembler. Genome Res 2003;**13**(1):81–90.

43. Bradnam KR, Fass JN, Alexandrov A, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2013;**2**(1), doi:10.1186/2047-217X-2-10.

44. Kliver S, Tamazian G, Krasheninnikova K, et al. Chromosome-level assembly of the endangered black-footed ferret (*Mustela nigripes*) provides insights into male infertility. Poster presented at: Plant and Animal Genome XXVII Conference, San Diego, CA. 2019.

45. Mapleson D, Accinelli GG, Kettleborough G, et al. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 2017;**33**(4):574–6.

46. Simao FA, Waterhouse RM, Ioannidis P. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 2015;**31**(19):3210–2.

47. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015.Accessed 5 Apr 2019

48. Hunt M, Kikuchi T, Sanders M, et al. REAPR: a universal tool for genome assembly evaluation. Genome Biol 2013;**14**(5):R47.

49. Ponstingl H. SMALT. 2010. Accessed 25 Oct 2017

50. Armstrong EE, Taylor RW, Prost S, et al. Cost-effective assembly of the African wild dog (*Lycaon pictus*) genome using linked reads. Gigascience 2019;**8**(2), doi:10.1093/gigascience/giy124.

51. Guo Y, Zhang Y, Liu Q, et al. A chromosomal-level genome assembly for the giant African snail *Achatina fulica*. Gigascience 2019;**8**(10), doi:10.1093/gigascience/giz124.

52. Kang S, Kim J-H, Jo E, et al. Chromosomal-level assembly of *Takifugu obscurus* (Abe, 1949) genome using third-generation DNA sequencing and Hi-C analysis. Mol Ecol Resour 2020;**20**(2):520–30.

53. Xu Z, Dixon JR. Genome reconstruction and haplotype phasing using chromosome conformation capture methodologies. Brief Funct Genomics 2020;**19**(2):139–50.

54. Zhou Y, Xiao S, Lin G, et al. Chromosome genome assembly and annotation of the yellowbelly pufferfish with PacBio and Hi-C sequencing data. Sci Data 2019;**6**(1):267.

55. Belton JM, McCord RP, Gibcus JH, et al. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 2012;**58**(3):268–76.

56. Korbel JO, Lee C. Genome assembly and haplotyping with Hi-C. Nat Biotechnol 2013;**31**(12):1099–101.

57. Liu K, Kim M, Fedak R, et al. Rapid high MW DNA extraction from plant, insect, cell and tissue samples for long-read sequencing using Nanoind Magnetic Disks. Poster presented at: Plant and Animal Genome XXVII Conference, San Diego, CA. 2019.

58. Collaborative Open Plant . https://copo-project.org.Accessed 18 July 2019

59. Etherington GJ, Heavens D, Baker D, et al. Supporting data for "Sequencing smart: De novo sequencing and assembly approaches for a non-model mammal.", GigaScience Database 2020. http://dx.doi.org/10.5524/100731.