A critical comparison of technologies for a plant genome sequencing project

Pirita Paajanen1,3,†, George Kettleborough1,†, Elena López-Girona2,4, Michael Giolai1, Darren Heavens1, David Baker1, Ashleigh Lister1, Fiorella Cugliandolo1, Gail Wilde2, Ingo Hein2, Iain Macaulay1, Glenn J. Bryan2 and and Matthew D. Clark1,5,∗

1Earlham Institute, Norwich, UK and 2The James Hutton Institute, Invergowrie, Dundee, UK and 3Department of Cell and Developmental Biology, John Innes Centre, Norwich, UK and 4The New Zealand Institute for Plant & Food Research Limited, Palmerston North, New Zealand and 5Department of Life Sciences, Natural History Museum, London, UK

∗matt.clark@nhm.ac.uk
†Contributed equally.

Abstract

Background A high quality genome sequence of any model organism is an essential starting point for genetic and other studies. Older clone based methods are slow and expensive, whereas faster, cheaper short read only assemblies can be incomplete and highly fragmented, which minimises their usefulness. The last few years have seen the introduction of many new technologies for genome assembly. These new technologies and associated new algorithms are typically benchmarked on microbial genomes or, if they scale appropriately, on larger (e.g. human) genomes. However, plant genomes can be much more repetitive and larger than the human genome, and plant biochemistry often makes obtaining high quality DNA free from contaminants difficult. Reflecting their challenging nature we observe that plant genome assembly statistics are typically poorer than for vertebrates.

Results Here we compare Illumina short read, PacBio long read, 10x Genomics linked reads, Dovetail Hi-C and BioNano Genomics optical maps, singly and combined, in producing high quality long range genome assemblies of the potato species S. verrucosum.

Conclusions The field of genome sequencing and assembly is reaching maturity and the differences we observe between assemblies are surprisingly small. We expect that our results will be helpful to other genome projects, and that these datasets will be used in benchmarking by assembly algorithm developers.

Key words: assembly, long reads, short reads, optical mapping, Pacific Biosciences, PacBio, 10x Genomics.

Developments in high-throughput sequencing have revolutionised genetics and genomics, with lower costs leading to an explosion in genome sequencing project size [1] and number of species [2]. Genomes from many diverse organisms have been sequenced, from marsupials to microbes, plants, phytoplankton, and fungi, amongst many others [3]. For a while it has been feasible for a single lab to sequence and de novo assemble a complex genome (for example, [4]).

The existence of very high quality references [5, 6] has made the human genome popular for demonstrating new sequencing technologies and assembly algorithms. The human genome has now been sequenced and assembled using various technologies including Sanger, 454, IonTorrent, Illumina, Pacific Biosciences (PacBio), 10x Genomics and even nanopore sequencing technologies [7, 8, 9, 10, 11, 12].

Hybrid approaches have also been used which combine complementarity technologies, for example PacBio and BioNano [13].

However, the human genome is not representative of all eukaryotic genomes, plant genomes in particular are typically more repetitive (including multi-kilobase long retrotransposon elements as well as even longer regions comprising of “nested” transposon insertions). Plant biology also poses challenges for the isolation of high quality high molecular weight DNA, due to strong cell walls, copurifying polysaccharides, and secondary metabolites which inhibit enzymes or directly damage DNA [14]. Thus, technologies that work well on vertebrate genomes may not work well for plants [15]. For these reasons slow and expensive clone based minimal tiling path sequencing approaches have persisted in plants [16, 17] long after faster, cheaper short read whole genome assemblies were first demon-
strated for vertebrate genomes [18]. In addition to increased genome repetitiveness and size, polyploidy is common in plants (especially key crops such as cotton, brassicas, wheat, potatoes) as are high levels of heterozygosity, especially where inbreeding is problematic due to generation times [19] or the plants are obligate outcrossers.

Plant biology poses some additional challenges for the isolation of high quality high molecular weight DNA. Plant cells possess strong rigid cell walls not broken by the addition of a detergent and, when physically breaking the cell wall, the DNA can be sheared, rendering the isolation of high molecular weight DNA problematic. A large proportion of the DNA in a plant cell can be from organelles (mitochondrial and chloroplast) [20] which are high copy number and large, for example the mitochondrial genome is 453 kbp in wheat [21] but only 16 kbp in human [22]. Plants are also rich in polyaccharides which can co-purify with DNA, and they produce secondary metabolites to protect themselves from herbivores [14].

Plant genomes also vary hugely in size, from 61 Mbp (Genlisea tuberosa, a member of the bladderwort family [23]) to 150 Gbp (Paris japonica, a relative of lilies [24]), it is still nontrivial to design a de novo assembly project which involves an ensemble of technologies. Each platform comes with its own input requirements, computational requirements, quality of output and, of course, labour and materials costs. Our results can be used as guidance for further sequencing assembly projects and provide a basis for comparative genome studies, as each sequencing strategy and assembly method has its own biases.

In this paper we compare several practical de novo assembly projects of a Mexican wild potato species Solanum verrucosum. We chose this genome because Solanum verrucosum is a self-compatible, diploid, tuber-bearing, wild potato species which we inbred further to produce the line Ver-54. The estimated genome size based on k-mer content is 722 Mbp. In addition, recent cytogenetic and molecular studies have shown it likely represents a genome donor to Mexican allopolyplod potatoes [25, 26] and as such is taxonomically distant from the genetically characterised cultivars and landraces, although it has been classified into the same larger phylogenetic potato clade (Clade 4) as cultivated potatoes [27]. The Mexican allopolyplodis in Series Longipedicellata and Demissa have very high levels of resistance to Phytophthora infestans (encoded by several R-genes) as does S. verrucosum. Thus, the S. verrucosum genome can be a highly useful genetic resource and a “potato model” for forward/reverse genetic studies relating to its high level of blight resistance, its unusually high level of self-fertility, and because it produces tubers, albeit small inedible ones. The Solanaceae, or nightshades which can co-purify with DNA, and they produce secondary metabolites to protect themselves from herbivores [14].

Contig assembly and scaffolding

The first stage of an assembly is to piece together reads to form long contiguous sequences, or contigs for short. These contigs can be ordered and oriented using longer-range information such as jumping/mate pair libraries. Throughout this paper we will refer to different contig assemblies that have been scaffolded. We use a naming convention which shows all of the steps used to construct the assembly. Each assembly name contains the steps used in order, separated by a hyphen. For example, the discovar-sp-dt-bn assembly is the discovar contig assembly scaffolded first with mate-pairs, then Dovetail and finally BioNano.

Illumina contig assembly

Two libraries were constructed for Illumina assembly. The first is a PCR-free library with insert size 500 bp (±40%) which was sequenced with 250 bp paired-end reads on a single Illumina HiSeq 2500 run. We refer to this below as the DISCOVAR library. The coverage of the library was 120×. The second library is a PCR-free “Tight and Long Library” (TALL) with insert size 650 bp (±20%) sequenced with 100 bp and 150 bp paired-end reads on two Illumina HiSeq 2500 runs. The coverage of this library was 135×.

We analysed the TALL library reads with preqc, part of the SGA assembler [32], and giving a genome size estimate at 702 Mbp, while the same analysis on the DISCOVAR library yielded 722 Mbp. The latter agrees better with the 727 Mbp size of the potato genome assembly [28].

The TALL library was assembled with ABySS (ABySS, RRID:SCR_016755) [34] producing a total contig assembly scaffolded first with mate-pairs, then Dovetail and finally BioNano.

The results for these two Illumina assemblies are similar in contiguity and shown in Table 1. However, while ABySS assembled about 8% longer total length, the number of small contigs was larger leading to very similar contig N50 to DISCOVAR. One additional feature was that ABySS performed more scaffolding using the paired end data but did not fill many of the introduced gaps leading to about 100 times higher percentage of N bases than DISCOVAR. These assemblies are more contiguous than the equivalent contig assemblies of the S. tuberosum genome where the reported contig N50 from paired-end reads is 22.4 kbp [28].

Contig assembly statistics of Illumina and PacBio assemblies, with a minimum contig/scaffold size of 1 kbp. abyss uses the TALL library, discovar uses the DISCOVAR library, and hgap, canu and falcon use the PacBio library. For a more comprehensive summary, see Supplementary Table 7.
Illumina scaffolding

A Nextera long mate-pair (LMP) library was made with insert size 10000 bp (±20%) and sequenced on two lanes of an Illumina MiSeq with fragment size 500 bp and 300 bp reads. The total coverage of the LMP library was 15x after we had filtered out duplicates 23.4% of reads, reads that did not contain a Nextera adapter or were too short to be useful.

We scaffolded both the discovar and abyss assemblies separately using Soapdenovo2 (soapdenovo2, RRID:SCR_014986) [35] producing discovar−mp and abyss−mp, respectively. The contiguity of both was increased significantly as shown in Table 1. Here the discovar−mp scaffolds were slightly better so we used this assembly to take forward for longer range scaffolding with other data types.

PacBio assembly

A gel size selected PacBio library with fragment lengths of at least 20 kbp was made according to the manufacturer’s instructions. The library was sequenced using a PacBio RSII instrument and P6C4 chemistry. We sequenced 65 SMRT cells total, each giving about 500 MB of data and a total coverage of 50x. The N50 of the fragments was 13 499 bp and total number of reads 9 768 980.

We conducted three long read assemblies on the same data using HGAP3 [36], part of SMRT-analysis (version 2.3.0p5) (SMRT-Analysis, RRID:SCR_002942), Canu (Canu, RRID:SCR_015880) [37] (version 1.0), and Falcon (Falcon, RRID:SCR_016089) [38] (version 0.3.0) producing the hgap, canu and falcon assemblies, respectively. The assembly statistics for each is shown in Table 1. Another long read assembler, that we chose not to use, because it does not include any error correction is miniasm [39]. This is a fast lower computational power alternative to the ones that we used in this paper and is useful for many purposes e.g. empirical testing of long read assemblies.

The Canu assembly was made with reads that were first error-corrected by the HGAP3 pipeline because the first attempt using raw reads resulted in an excessive amounts of small scaffolds and a genome size more than 50% longer than expected.

The canu and hgap assemblies contain slightly more sequence content (as measured by the total length of the assembly), and a lower percentage of unknown bases (as measured by the percentage of bases denoted by N) than the short read assemblies. This may be due to their capturing of additional difficult sequences, especially repeat elements which short read assemblies are known to have problems traversing. The falcon assembly has the highest N50, and while canu is closest to the estimated genome length. Falcon also produced 9.9 Mb of alternate contigs, likely from residual heterozygosity, which will be useful for interpreting downstream genetic results e.g. forward and reverse genetic screens. We also found this assembly was easier and faster to run than HGAP3. We also found the basepair accuracy of canu read correction to be lower than HGAP3 read correction. For these reasons we chose the falcon assembly to take forward to hybrid scaffolding. We first polished it using Quiver as part of SMRT-analysis (version 2.3.0p5).
Longer-range scaffolding

To achieve higher contiguity, newer technologies have been developed to complement the previous methods and, in some cases, each other. In this section we investigate using longer range scaffolding methods to increase the contiguity of the Illumina discovar-mp assembly and the falcon PacBio assembly. We also investigate the 10x Genomics platform constructs a new paired-end library which is roughly equivalent to 350× coverage.

Dovetail

Dovetail Genomics provides a specialised library preparation method called Chicago and an assembly service using a custom scaffold called HiRise. The Chicago library preparation technique is based on the Hi-C method, producing deliberately “chimeric” inserts linking DNA fragments from distant parts of the original molecule [40]. This is followed by standard Illumina paired-end sequencing of the inserts. Since the separation of the original fragments follows a well-modelled insert size distribution, the scaffold is able to join contigs to form scaffolds spanning large distances, even up to 500 kbp [40].

Dovetail Genomics, LLC (Santa Cruz, CA, USA) received fresh leaf material from us from which they constructed a Chicago library. This was sequenced at Earlham Institute using Illumina 250 bp paired-end reads. The total read coverage of the Chicago library was 105×. Dovetail used their HiRise software to further scaffold the discovar-mp assembly, increasing the N50 from 858 kbp to 871 kbp, and the falcon assembly, increasing the N50 from 712 kbp to 2553 kbp. These assemblies are called discovar-mp-dt and falcon-dt, respectively.

BioNano

The BioNano Genomics Irys platform constructs a physical map using very large DNA fragments digested at known sequence motifs with a specific nicking enzyme, to which a polymerase adds a fluorescent nucleotide. The molecules are scanned, and the distance between nicks generates a fingerprint of each molecule which is then used to build a whole genome physical map. Sequence-based scaffolds or contigs can be integrated by performing the same digestion in silico then ordering and orienting the contigs according to the physical map [41].

We collected BioNano data from 16 runs by repeatedly running the same chip. After filtering fragments less than 100 kbp, the yield varied from 0.8 Gb to 25.8 Gb, with the earlier runs yielding more whereas the molecule N50 was higher in later runs (ranging from 135 kbp to 240 kbp). The total yield of BioNano data was 252 Gbp which is roughly equivalent to 350× coverage.

We performed hybrid scaffolding on the discovar-mp and falcon assemblies. The in silico digest suggested a label density of 8.1/100 kbp for discovar-mp and 8.4/100 kbp for falcon whilst the actual observed density was only 6.8/100 kbp. We used the BioNano pipeline (v2.0) (BioNano Irys, RRID:SCR_016754) to scaffold discovar-mp, increasing the N50 from 858 kbp to 1260 kbp, and falcon, increasing the N50 from 710 kbp to 1500 kbp. These assemblies are called discovar-mp-bn and falcon-bn, respectively.

10x Genomics

10x Genomics provides an integrated microfluidics based platform for generating linked reads (a cloud of non-contiguous reads with the same barcode from the same original DNA molecule) and customised software for their analysis [11]. Large fragments of genomic DNA are combined with individually barcoded gel beads into micelles in which library fragments are constructed and then sequenced as a standard Illumina library. Using the barcodes the reads from the same gel bead can be grouped together.

Unlike the previous two longer-range scaffolding approaches, the 10x Genomics platform constructs a new paired-end library which can be sequenced and then assembled into large scaffolds by one assembly program: Supernova.

A 10x Genomics Chromium library was made according to manufacturer’s instructions and a lane of Illumina HiSeq 250 bp paired-end reads were generated with a coverage of about 92×. Supernova (version 1.1.1) (Supernova, RRID:SCR_016756) produced the supernova assembly with length 641 Mbp and a scaffold N50 of 2.33 Mbp. Trimming reads back to 150 b for reducing sequencing depth to 56× which are the read length and depth recommended by 10x Genomics, generated very similar results (see Supplemental Section ??) compared to the ones reported above.

Hybrid scaffolding

It is possible to iteratively combine these longer-range scaffolding approaches. We tested several hybrid approaches using the discovar, falcon and supernova assemblies. For example the discovar-mp assembly was scaffolded using Dovetail and then BioNano producing discovar-mp-dt-bn with an N50 of 7.0 Mbp, the highest contiguity
of any assembly reported here. The falcon assembly when scaffo-
folded with both produced scaffolds with an N50 of only 3.09 Mbp, 
lower than with BioNano alone. Finally we scaffolded the supernova 
assembly with BioNano producing supernova-bn which increased 
the N50 from 2.33 Mbp to 2.85 Mbp.

Most scaffolding steps add gaps of unknown sequence, so we 
also used long reads from PacBio to scaffold and to perform “gap-
filling” on the assemblies, replacing regions of unknown sequence (N 
stretches) with a PacBio consensus sequence. This also presents an 
opportunity to use lower coverage PacBio data to improve an Illu-
mina assembly, which may be more cost effective than a de novo 
assembly using PacBio. PBJelly (version 15.2.20) [42] was used to 
perform gapfilling using only 10 SMRTcells of PacBio data (8x 
depth). The Supernova assembly increased in size from 641 Mbp 
to 671 Mbp, and N50 from 2.33 Mbp to 2.64 Mbp, and the amount 
of Ns present reduced from 7.58% to 5.14%. The discovar-mp-dt 
assembly increased in size from 656 Mbp to 680 Mbp and N50 from 
4.69 Mbp to 4.87 Mbp, with Ns reduced from 3.03% to 1.28%. How-
ever, how gaps and percentage Ns are generated differs between as-
sembly methods (see Discussion).

Assembly evaluation

Achieving a genome assembly with high levels of contiguity is poten-
tially useless if it does not faithfully represent the original genome 
sequence. We assessed errors in assemblies by comparison to the 
raw data used to make the assemblies, as well as measuring gene 
content, local accuracy (BAC assemblies), and long-range synteny 
with the close relative Solanum tuberosum.

K-mer content

Analysis of the k-mer content of an assembly gives a broad overview 
of how well the assembly represents the underlying genome. We 
used the PCR-free Illumina Discovar library as our reference for 
the k-mer content of the genome. Due to the high accuracy of the 
reads we expect the k-mer spectra for a library to form a number of 
distributions which correspond to read errors, non-repetitive, and 
repetitive content in the genome. These distributions can be seen 
by observing only the shapes and ignoring the colours in Figure 2. 
The reader is referred to the KAT (KAT, RRID:SCR_016741) doc-
umentation for further details [43].

In Figure 2 we compare the k-mer contents of the three contig 
assemblies—discovar, falcon, and supernova—to the Discovar 
library. To minimise the effects of the differences between Illumina and 
PacBio sequencing error profiles the falcon assembly has been pol-
ished with the Illumina reads using Pilon [44] (see Supplemental 
Figure 7 for the unpolished plot).

The small red bar on the origin in some plots shows content 
which appears in the assembly but not in the Illumina reads. The 
discovar assembly is very faithful to the content in the library. The 
black area denotes sequences in the reads but not in the assembly;
those clustering at the origin are predicted sequence errors in the
reads, the small amount between 50 to 100 on the z-axis is sequence 
missing from the assembly. The dominant red peak (1x, around 
multiplicity 77), which is the vast majority of all assemblies here, 
contains content in the Illumina reads which appears once in the as-
sembly (homozygous sample). Green areas on top of the main peak 
in Falcon and Supernova represents possible duplications in the as-
sembly, whereas the green (2x) small peak to the right of the main 
peak is probably true duplicates—as these sequences are present 
twice in the assembly and at twice the expected read counts. At 
the main peak (k-mer multiplicity 77), the amount of potentially 
duplicated content in the assemblies (that is, number of k-mers ap-
pearing more than once in the assembly) is 0.66% in falcon, 1.3% 
in supernova, and 0.15% in discovar.

Gene content

We assessed the gene content of the three most contigu-
assemblies—discovar, falcon, and supernova—using two datasets. The first is with BUSCO and its 
embryophytaodb9 (plants) dataset (BUSCO, RRID:SCR_015008) 
[45] and the second is all the predicted transcript sequences from the S. tuberosum genome [28].

We found that each of the three assemblies shows at least 95% of 
BUSCOs as complete, with just a small difference of only 2% to 3% missing. The results are shown in Figure 4.

We aligned the S. tuberosum representative transcript sequences 
to each genome assembly using blast [46] and then measured how much of each transcript sequence was represented in the assembly according to various minimum percentage identity cutoffs. As expected when comparing between species, as the threshold approaches 100% nucleotide identity the transcript completeness drops closer 
to zero. Using a threshold between 96% to 98% we find the median transcript completeness is highest in discovar, followed 
by falcon, and then supernova. However, the difference between the assemblies is small. Figure 5 shows a box and whisker 
plot of completeness of the representative transcript sequences.

Local accuracy

As BACs are easier to assemble due to smaller size and a much 
more limited amount of repetitive DNA content than a whole gen-
ome, we assessed the performance of our three assemblies at a local 
scale using BAC assemblies. We randomly selected, sequenced, and 
assembled 96 BAC clones from S. verrucosum BAC library. We 
chose 20 high-quality BAC assemblies (single scaffolds/contigs with 
Illumina or PacBio) to measure the accuracy of the whole genome 
assemblies.
We used dnaDiff [47] to compare the BAC sequences to the supernova-bn, discovar-mp-dt-bn, and falcon-dt-bn assemblies finding sequence identities of 99.40%, 99.97%, and 99.87%, respectively. As in the previous section, the discovar-mp-dt-bn assembly shows the highest accuracy, with supernova-bn the lowest, though the differences are small.

To illustrate the performance of the different technologies sequencing different genomic features we mapped whole genome reads and assemblies to single BACs as shown in Figure 3. None of our three whole genome assemblies are able to reconstruct BAC 22; each breaking at a large (more than 12 kbp) repeat. The Discovar library (paired-end), mate-pair library and Dovetail library were each mapped and only reads mapping to a high quality and exhibiting up to one mismatch are shown in the figure. The mapping reveals several areas of high repetition, for example the arms and middle of a retrotransposon, and there are areas lacking coverage completely which suggests a sequence which is difficult for our Illumina sequence data to resolve. We also see drops in coverage at some sites with high concentrations of homopolymers, as marked by coloured lines in the GC content, for example an A rich region of ~7 kbp. Interestingly the repeat arms are also rich in homopolymers.

We note that the discovar-mp-dt-bn assembly leaves the largest gap around the repeat. The falcon assembly was able to completely cover an area with no mapping paired-end Illumina reads which explains some of extra k-mer content in Figure 2 noted earlier in this assembly. The supernova-bn assembly was able to reconstruct more of the difficult region, but it also contains duplications in the homopolymer rich flanking regions that is not seen in the other assemblies.

The mate-pair library was not able to scaffold the discovar contigs due to the size of this repeat being larger than its 10 kbp insert size. The mate-pair fragments also map to a great depth in the repeat. Dovetail data, however, shows a much smoother fragment distribution and was able to scaffold the two discovar contigs in the correct order and orientation as it could scaffold up to 50 kbp (the cutoff used by the HiRise scaffold). However, the gap length was not estimated with Dovetail and was arbitrarily set to 100 Ns when in reality the gap is over 12000 bp long. While BioNano software estimates gap sizes, we note that BioNano data was not able to close this particular gap in any of the assemblies.

Long-range accuracy using synteny to S. tuberosum

As all our assemblies are de novo, in the sense that we used no prior information from other Solanaceae genomes, we reasoned that more accurate long range scaffolding would be apparent as longer syntenic blocks to a closely related species. We used nucmer [47] to analyse the synteny of our assemblies to the pseudomolecules of the S. tuberosum genome [48]. Figure 6 shows the mummer plot for chromosome 11 of S. tuberosum against our three assemblies. We saw the falcon-dt-bn assembly showed the best synteny with the discovar-mp-dt-bn being the worst. The plots for the remaining chromosomes are shown in Supplemental Figures 7?, 7? and 7?

Using synteny we identified two cases of chimerism, i.e. scaffolds that align well to two different pseudomolecules of S. tuberosum genome. Both cases are in discovar-mp-dt-bn but not falcon-dt-bn. The first 1.5 Mbp of scaffold ScEqE3Q_528 maps to pseudomolecule 7 while the last 2.9 Mbp map to pseudomolecule 2 in the S. tuberosum genome. There is no conflict reported with the BioNano Genomics optical map in this area, but we can exclude the possibility that these are real chromosome structural arrangements in S. verrucosum because we have GbS markers on each end of this scaffold which also map in an S. verrucosum cross to these different linkage groups (López-Girona unpublished). The other case is a scaffold ScEqE3Q_633 in which the first 1.4 Mbp map to pseudomolecule 8 and the remainder to pseudomolecule 3, here BioNano Genomics does report a conflict which would highlight this error, and S. verrucosum genetic markers also support the chimera classification.

Discussion

The quality and quantity of DNA available, whether it is from fresh or frozen tissue, and ease of its extraction will often dictate which sequencing technologies and assembly algorithms. Instead they differ mostly in the length of the ungapped sequence and scaffolds, with much smaller differences in missing sequence and gene content, duplicated regions, and per base accuracy.

A Discovar assembly is the cheapest and easiest to construct, and the resulting assembly is very accurate, albeit highly fragmented. Adding a long mate-pair library is a proven method of increasing the contiguity of a short read assembly by scaffolding. The 10x
Genomics based assembly using Supernova was as easy to obtain as the Discovar assembly. The two most remarkable features of this assembly are the low cost and input DNA requirement: for only slightly higher cost than a Discovar assembly, and considerably less than with only one long mate-pair library, we obtained an assembly comparable to what one would expect from multiple long mate-pair libraries.

Our PacBio assembly using Falcon achieved contiguity similar to that of discovar-mp (Discovar plus long mate-pair scaffolding). PacBio sequencing has a considerably higher cost and material requirement than Illumina sequencing, but the Falcon assembly contains truly contiguous sequence as opposed to discovar-mp which contains gaps patched with Ns. The PacBio read lengths (N50=13.5kbp) were similar to the insert size of mp library (mean 10kb), and the read coverage was higher for PacBio (50×) than with Dovetail scaffolds (N50=858kbp).

The addition of Dovetail showed the most striking increase in contiguity by scaffolding. We note that our Dovetail scaffolds provided the order and orientation of the constituent contigs but no estimate for the length of the gaps between them. This should be taken into consideration if true physical length of sequences is important, and for specific downstream uses. Both Illumina (Discovar-mp) and PacBio (Falcon) assemblies are amenable to the addition of Dovetail, but the scaffolds produced from the Falcon contigs (4× increase) were not as long as those from the Illumina assembly (5.5× increase). This could be due to the Falcon assembly has been polished with PacBio reads, it retains some PacBio errors and so some Dovetail (Illumina) reads do not pass stringgent mapping filters. If true, Pilon polishing with Illumina reads could help, as it improved the k-mer spectra (Figure 2).

With BioNano Genomics restriction enzyme digest based optical maps we obtained less (~2× increase) scaffolding improvement than with Dovetail (4× to 5.5× increase). This could be due to three issues: first that assembly gaps are not correctly sized which prevents real, and in silico, restriction maps matching (as information is purely encoded in the distances between sites). We see that the ungapped PacBio assemblies improve more than scaffolds Illumina, and Dovetail scaffolds (with arbitrary 100 bp gaps) hardly increase at all. Secondly, because the method produces low information density (one enzyme site per ~12kbp) long fragments with many sites are needed to create significant matches, and our DNA was not sufficiently long (best run N50 was 240kbp). Longer DNA (over 300kbp), and perhaps multiple enzyme maps with iterative scaffolding could have improved the results. Thirdly we observe that the in silico restriction rates for Illumina and PacBio assemblies are similar (8.1× to 8.4 sites/100 kbp) whereas the actual observed rates from the physical map is much lower at 6.8 sites/100 kbp, suggesting that there could be a fraction of the genome missing from our assemblies which is very low in sites such as centromeric or telomeric regions where the BioNano Genomics map can not scaffold through.

Gapfilling using PBJelly offers an attractive method of using the long read data from PacBio to improve an existing Illumina based assembly. This closed many of the gaps in the scaffolds thereby decreasing the fraction of unknown sequence (Ns) and also increasing the contiguity. The increase in contiguity of the 10x Genomics assembly was the highest. It was intriguing to see if an assembly approach combining Chromium data with long reads (directly on the assembly graph) can combine the best attributes of both data types to resolve complex regions. Analysis of the k-mer content of the supernova, discovar, and falcon assemblies showed that the k-mer spectra of each assembly is very clean. We see slightly higher level of sequence duplication in the supernova assembly, and to a lesser extent in the falcon assembly. All three assembly algorithms are diploid aware, meaning they are able to preserve both haplotypes. The gene content of each assembly was very similar with all three of our long assemblies showing a high percentage of the expected genes. The 10x Genomics based assembly showed a slightly lower count in both of our assessments but the difference is very small.

We used multiple BAC assemblies of ~100 kb insert size to illustrate the technical limitations of each method. Short read methods cannot resolve many areas of repetition within a WGS assembly. This is especially noticeable in a plant genome with higher repeat content, and is one of the major reasons for breaks in contiguity in these assemblies. In our example in Figure 3, the long mate-pair library alone is not sufficient. It takes the larger fragment lengths within the Dovetail Chicago library to finally make the join in the whole genome assembly.

Long read technologies do not suffer as much with repeats and, in the case of PacBio, tend to have more random rather than systematic errors [49]. We can see in our exemplar that the falcon assembly covers some of the repetitive region. The underlying BAC assembly was also obtained with PacBio and gave us a single true contig for the entire BAC. On close inspection we noticed that difficult region was spanned by reads of length 22kbp to 26kbp. This shows that
long reads are certainly able to span such regions of difficulty, and to assemble them.

Recently ultra-long reads with an N50 of 99.7 kbp (max. 882 kbp) with ~92% accuracy have been produced with the new MinION R9.4 chemistry using high molecular weight DNA from a human sample [12]. If this is also achievable on plant material the remaining (mostly repetitive) fraction of genomes should become visible. The recent S. pennellii Nanopore assembly [50] reported average read length 12.7 kbp and error rate of 18% to 20%.

To evaluate the longer range accuracy of our genome assemblies we compared them to the closely related S. tuberosum pseudomolecule assembly, which revealed good synteny with all three of our longest assemblies (discovar-mp-dt-bn, falcon-dt-bn and supernova). There are some disagreements especially in the centromeric areas, but as these appeared in all assemblies these could illustrate real structural variation. We detected two chimeric scaffolds in the discovar-mp-dt-bn assembly but neither is present in the falcon-dt-bn. The two Dovetail scaffolding processes shared the same Hi-C sequence data but were conducted many months apart (discovar-mp first and later falcon), and used different versions of Dovetail’s proprietary HiRise software, versions 0.9.6 and 1.3.0, respectively, which might have affected the results. On detailed examination we see that the ScfEqESQ_528 scaffold chimeric join is made by Dovetail hopping through a fragmented area of short (1 kbp to 2 kbp) contigs. Such small contigs do not exist in the Falcon assembly, which maybe why we do not find chimeras. BioNano Genomics finds it hard to map to areas with many Dovetail gaps (as these are set to an arbitrary 100 bp size), and this region also has a high enzyme nicking rate (nearly twice the genome average), including two areas where nicks are less than 200 bp apart and so would be optically merged. In scaffold ScfEqESQ_633 we detect that discovar-mp scaffold123 was correctly split by Dovetail data as chimeric (also highlighted by BioNano Genomics and genetic markers) but the scaffold was not broken at the exact chimeric join, and the remaining sequence from the wrong chromosome was sufficient for Dovetail to propagate the error. Whilst we did not detect a high level of systematic errors in any of our assembly methods, the importance of using BioNano Genomics and genetic markers to identify chimeras that then can be broken is apparent.

Even though we found some surprisingly small differences between assemblies of S. verrucosum, this is an inbred diploid potato species, with a medium size genome and is in no way exceptional. As there are about 300,000 angiosperms alone [51] we remind the reader, that many factors e.g. genome size, the ease of high quality HMW DNA extraction, the types of repeat content, polyploidy or heterozygosity may pose additional hurdles affecting the choice of technology and how well they will perform. Heterozygosity, in particular, complicates the assembly process and if individual haplotypes are desired this places limitations on which strategies can be used. The careful choice of sample where possible, such as a highly inbred plant or doubled haploid, can remove or minimise these problems. This approach was also adopted for the potato DM reference, whereby a completely homozygous “doubled monoploid” was used as the heterozygous diploid RH genotype originally selected for sequencing proved difficult to assemble due to the extremely high level of heterozygosity. Newer methods have recently been developed to assemble diploid genomes into chromosome scale phase blocks [52] or even to exploit the haplotype diversity using a “trio binning” approach developed in [53], so we expect to see more true diploid assemblies in the near future.

Materials and Methods

Project requirements

Each of the assembly methods we have used comes with its own requirements. We have broken this down into material requirements, that is plant and DNA material, monetary requirements, that is the cost of preparation and sequencing, and computational requirements. Table 3 lists the material requirements for each library. Amounts in grams are for fresh/frozen material and amounts in micrograms for DNA. In each case where frozen or flash frozen is stated, fresh material is also acceptable.

<table>
<thead>
<tr>
<th>Library</th>
<th>Tissue type</th>
<th>Material/DNA amount</th>
<th>HMW Fragment length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TALL</td>
<td>Frozen</td>
<td>3 µg</td>
<td>No</td>
</tr>
<tr>
<td>Discovar</td>
<td>Frozen</td>
<td>0.6 µg</td>
<td>No</td>
</tr>
<tr>
<td>Mate-pair</td>
<td>Frozen</td>
<td>4 µg</td>
<td>No</td>
</tr>
<tr>
<td>PacBio</td>
<td>Young frozen</td>
<td>5 g</td>
<td>No</td>
</tr>
<tr>
<td>BioNano</td>
<td>Young fresh</td>
<td>2.5 µg</td>
<td>Yes</td>
</tr>
<tr>
<td>Dovetail</td>
<td>Fresh</td>
<td>20 g</td>
<td>Yes</td>
</tr>
<tr>
<td>Chromium</td>
<td>Flash frozen</td>
<td>0.5 µg</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 3. Material requirements for each library. Amounts in grams are for fresh/frozen material and amounts in micrograms for DNA. In each case where frozen or flash frozen is stated, fresh material is also acceptable.

<table>
<thead>
<tr>
<th>Name of assembly</th>
<th>Approximate runtime</th>
<th>Peak memory</th>
<th>Average memory</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supernova</td>
<td>3 d</td>
<td>1300 GB</td>
<td>Large memory</td>
<td></td>
</tr>
<tr>
<td>Canu (Uncorr)</td>
<td>12 d</td>
<td>47 GB</td>
<td>20 GB</td>
<td>HPC cluster</td>
</tr>
<tr>
<td>Canu (Corr)</td>
<td>4 d</td>
<td>34 GB</td>
<td>14 GB</td>
<td>HPC cluster</td>
</tr>
<tr>
<td>Falcon</td>
<td>5 d</td>
<td>120 GB</td>
<td>60 GB</td>
<td>HPC cluster</td>
</tr>
<tr>
<td>HGAP</td>
<td>2 m</td>
<td>280 GB</td>
<td>Large memory</td>
<td></td>
</tr>
<tr>
<td>Discovar</td>
<td>22 h</td>
<td>260 GB</td>
<td>134 GB</td>
<td>HPC cluster</td>
</tr>
<tr>
<td>AlbySS</td>
<td>1 w</td>
<td>64 GB</td>
<td>Large memory</td>
<td></td>
</tr>
<tr>
<td>BioNano (Asm)</td>
<td>8 h</td>
<td>64 GB</td>
<td>64 GB</td>
<td>HPC cluster</td>
</tr>
<tr>
<td>BioNano (Scaf)</td>
<td>1 d</td>
<td>64 GB</td>
<td>64 GB</td>
<td>HPC cluster</td>
</tr>
</tbody>
</table>

Table 4. Computational requirements.

In this section we briefly describe methods for library preparation and sequencing. For a comprehensive description, please see the supplementary material.

S. verrucosum accession Ver-54 was grown in the glass house in James Hutton Institute in Scotland. Both fresh and frozen leaves from this accession and its clones were used for DNA extraction.

The TALL library was prepared using 3 µg of DNA and fragments of 650 bp were sequenced with a HiSeq2500 with a 2 x 150 bp read metric. The Discovar library was prepared using 600 ng of DNA and fragments of 500 bp were sequenced with a HiSeq2500 with a 2 x 250 bp read metric. The mate-pair library was prepared using 4 µg of DNA and fragments of 10 kbp were circularised, fragmented and sequenced on a MiSeq with a 2 x 300 bp read metric [54].

A PacBio library was prepared using 5 g of frozen leaf material. A 20 kbp fragment length library was prepared according to manufacturer’s instructions and sequenced on 65 SMRT cells with the P6C4 chemistry on a PacBio RSII.
The 10x Chromium library was prepared according to the manufacturer’s instructions and sequenced on a HiSeq2500 with a 2×250 bp read metric.

For BioNano, DNA was extracted using the IrysPrep protocol. 300ng was used in the Nick, Label, Repair and Stain reaction and loaded onto a single flow cell on a BioNano chip. The chip was run eight times to generate 252Gb of raw data.

Assembly and evaluation

All tools and scripts that were used to perform the evaluation and produce the figures are available on GitHub in the georgek/potato-figures repository.

We used Rampart (Rampart, RRID:SCR_016742) [55] to run ABySS [33] multiple times with different k values. Discovar de novo was run with normal parameters.

Long mate-pair reads were first processed with NextClip (NextClip, RRID:SCR_005465) [56] to remove the Nextera adapter. Soapdenovo2 was then used to perform scaffolding with both the paired-end and mate-pair libraries.

k-mer content was analysed with the kat comp tool [43] (KAT, RRID:SCR_016741). We used default parameters with manually adjusted plot axes to show the relevant information.

We used the BUSCO core plant dataset to evaluate the gene content. The S. tuberosum representative transcripts (PGSC_DM_V403_representative_genes from http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml) were aligned to the assemblies using BLAST and the coverage of transcripts at various thresholds using a tool we developed.

The BACs were sequenced with the Earlham Institute BAC pipeline [57] and were assembled with Discovar de novo using normal parameters after filtering for E. coli and the BAC vector. The PacBio BAC was assembled using HGAP3 [36]. We used GNU parallel [58] for concurrent assembly and analysis.

20 BACs which assembled into a single contig were selected to use as a reference. These BACs are non-redundant to the extent that they do not share any lengths of sequence of more than 95 % identity and over 5000bp long. Short reads were aligned to the BACs using Bowtie2 [59] with default parameters. The assemblies were mapped to the BACs using bwa mem [60]. The mapped sequences were sorted and filtered for quality using sambamba [61]. Fragment coverage was calculated using samtools [62] and bedtools [63].

Synteny was analysed with mummer [64]. We used nucmer to align the assemblies to the S. tuberosum reference v4.04 [65]. Alignments less than 10kb and 90 % identity were filtered out.

Availability of Supporting Data

All read data generated in this study have been submitted to the EMBL-EBI European Nucleotide Archive under the project PRJEB20860. Archival copy of the code, assemblies and other data are available in the GigaScience GigaDB repository [66].

Acknowledgements

We thank Lawrence Percival-Alwyn and Walter Verweij for their assistance in library preparation and analysis, and Michael Bevan for critical reading of this manuscript. This work was funded with BBSRC project grants (BB/K019325/1) and (BB/K019090/1). This work was strategically funded by the BBSRC, Core Strategic Programme Grant (BB/CSP17270/1) at the Earlham Institute. High-throughput sequencing and library construction was delivered via the BBSRC National Capability in Genomics (BB/CCG1720/1) at the Earlham Institute (EI, formerly The Genome Analysis Centre, Norwich), by members of the Platforms and Pipelines Group. This research was supported in part by the NBI Computing infrastructure for Science (CIS) group through the HPC cluster and UV systems. We thank Duke University for providing sequencing costs via Dugsim (https://dugsim.net/).

Authors’ contributions: GB, ELG, IH, and GW prepared the sample. MDC, GK, and PP designed the analysis. DB, GB, FC, ELG, MG, DH, IH, AL, and IM constructed libraries and performed sequencing. GK and PP made the assemblies and GK, ELG, and PP performed the evaluation. MDC, GK, GB, ELG and PP wrote and prepared the manuscript. All authors read and approved the final manuscript.
References


