Journal of
Next Generation Sequencing & Applications

Bellec et al., Next Generat Sequenc & Applic 2016, 3:2
DOI: 10.4172/2469-9853.1000128

Research Article Open Access

Long Read Sequencing Technology to Solve Complex Genomic Regions Assembly in Plants

Arnaud Bellec[1]*, Audrey Courtial[1], Stephane Cauet[1], Nathalie Rodde[1], Sonia Vautrin[1], Genseric Beydon[1], Nadege Arnal[1], Nadine Gautier[1], Joelle Fourment[1], Elisa Prat[1], William Marande[1], Yves Barriere[2] and Helene Berges[1]

[1]*French Plant Genomic Center, Centre National des Ressources Génomiques Végétales, INRA-CNRGV, Castanet-Tolosan, France*
[2]*INRA, UR889, Unité de Génétique et d'Amélioration des Plantes Fourragères, 86600 Lusignan, France*

Abstract

Background: Numerous completed or on-going whole genome sequencing projects have highlighted the fact that obtaining a high quality genome sequence is necessary to address comparative genomics questions such as structural variations among genotypes and gain or loss of specific function. Despite the spectacular progress that has been made in sequencing technologies, obtaining accurate and reliable data is still a challenge, both at the whole genome scale and when targeting specific genomic regions. These problems are even more noticeable for complex plant genomes. Most plant genomes are known to be particularly challenging due to their size, high density of repetitive elements and various levels of ploidy. To overcome these problems, we have developed a strategy to reduce genome complexity by using the large insert BAC libraries combined with next generation sequencing technologies.

Results: We compared two different technologies (Roche-454 and Pacific Biosciences PacBio RS II) to sequence pools of BAC clones in order to obtain the best quality sequence. We targeted nine BAC clones from different species (maize, wheat, strawberry, barley, sugarcane and sunflower) known to be complex in terms of sequence assembly. We sequenced the pools of the nine BAC clones with both technologies. We compared assembly results and highlighted differences due to the sequencing technologies used.

Conclusions: We demonstrated that the long reads obtained with the PacBio RS II technology serve to obtain a better and more reliable assembly, notably by preventing errors due to duplicated or repetitive sequences in the same region.

Keywords: Long read sequencing; Genome assembly; Bacterial artificial chromosomes; Repeated elements; Plant genomes

Introduction

During the last decade, we have observed remarkable advances in sequencing technology and bioinformatics analysis. The turning point came when the pyrosequencing technologies became available for the scientific community. Following Sanger's method, pyrosequencing has provided a massive increase in sequencing throughput combined with a huge decrease in the cost per sequenced base. Thus, it became possible to sequence any genome in a faster and more affordable way. Nevertheless, the read length of pyrosequencing technology (i.e. 200-400 bp) has constituted a step backward compared to standard sizes obtained using Sanger technology, (i.e. 800 bp) and is a weak point for de novo genome assembly [1,2]. The sequence length is critical to overcome assembly problems linked to particular features of genomes such as large genome size, high repetitive DNA ratio and various ploidy levels. These features are frequently found to be combined in complex genomes of plants [3]. Thus, obtaining high-quality sequences of complex genomes is still very challenging. Gaps and collapsed repeat sequences remain frequent in final assemblies of several released genomes [4]. To overcome these problems, most plant genome sequencing projects have used the information from the large insert DNA libraries, mainly Bacterial Artificial Chromosome (BAC) libraries. The BAC clones were physically ordered and sequenced, to obtain a genome backbone. For some genomes like barley, data from the whole genome shotgun sequencing technology were added to increase the genome coverage and decrease the error rate [5]. Nevertheless, sequence assembly from one BAC clone is still difficult, even with Roche-454 technology that produces the longest reads among second-generation sequencing technologies. It is common to obtain several contigs for a single BAC that may be very limiting in genomic research such as structural sequence analysis. This has been reported for BAC clones from complex regions of human and chimpanzee genomes and we have encountered the same problems when performing the assembly of BAC clone sequences of various plant species (unpublished observations) [6]. Nowadays, the emergence of third generation sequencers, and more particularly the Pacific Biosciences RS (PacBio RS II) platform by increasing read length, has radically improved genome assemblies of many bacteria, animals and chloroplast genomes [6-9]. In addition to the read length, the PacBio RS II produces the sequence from a single DNA molecule and no PCR amplification step is needed, thus reducing the errors due to the multiple DNA copies. Regarding the matter of error rates, the massive throughput obtained with the NGS methods allows high sequence coverage and consequently an acceptable degree of confidence. For complex genomes, the very long reads are promising.

In this study, we assessed the performance of PacBio RS II technology by sequencing plant BAC clones with the expectation that the read length would solve the assembly problems reported above. To do that, we compared the efficiency of PacBio RS II and Roche-454 technologies to sequence and assemble pools of BAC clones. Thus, we sequenced with both technologies BAC clones from six different plant species, and we compared their assembly. This work

***Corresponding authors:** Arnaud Bellec, French Plant Genomic Center : Centre National des Ressources Génomiques Végétales, INRA-CNRGV, Castanet-Tolosan, France, Tel: +33561285562; Fax: +33561285564; E-mail: arnaud.bellec@toulouse.inra.fr

Received May 13, 2016; **Accepted** July 07, 2016; **Published** July 09, 2016

Citation: Bellec A, Courtial A, Cauet S, Rodde N, Vautrin S, et al. (2016) Long Read Sequencing Technology to Solve Complex Genomic Regions Assembly in Plants. Next Generat Sequenc & Applic 3: 128. doi:10.4172/2469-9853.1000128

is a useful comparison of two sequencing technologies highlighting the advantages or weaknesses of each approach. We showed that long-read technology provides a more reliable assembly, preventing possible collapses of duplicated regions which can have a significant impact in the study of structural variation at the genome scale.

Methods

BAC clones selection, plasmid DNA preparation and insert size estimation

We selected nine BAC clones from six different plant species that had previously been sequenced and analyzed with Roche-454, and we sequenced these BAC clones with the PacBio RS II technology for the purpose of this comparative analysis. The plant species represented were strawberry (*Fragaria* × *ananassa*-1BAC clone), sunflower (*Helianthus annuus*-1BAC clone), barley (*Hordeum vulgare* -1 BAC clone), sugarcane (*Saccharum hybridum*-2 BAC clones), bread wheat (*Triticum aestivum* - 2BAC clones) and maize (*Zea mays*-2 BAC clones) (Table 1). BAC clones from the same species covered different genomic regions with no indication of a homology relationship. BAC clone sizes ranged from 85 kb to 170 kb and represented a total of 1135 kb (Figure 1). Mean insert size was estimated to be around 126 kb.

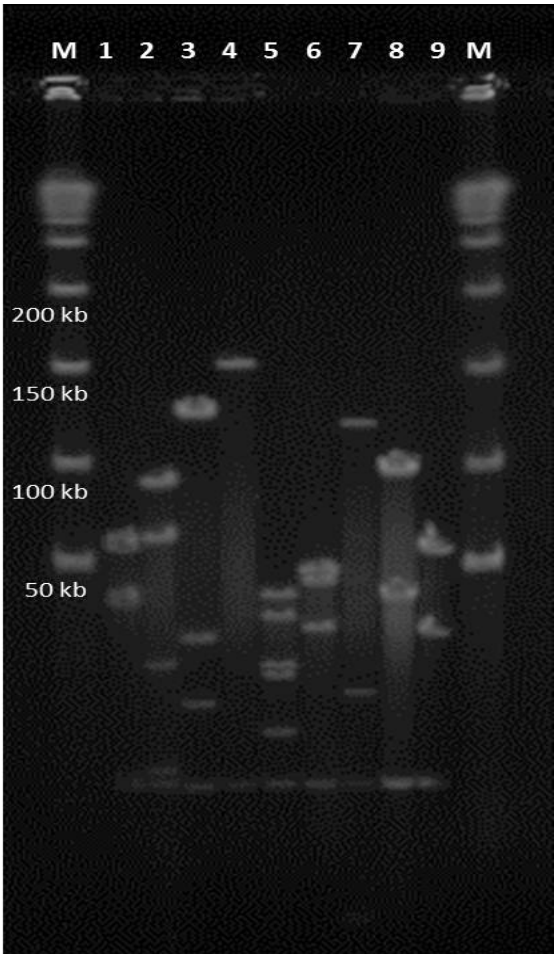
BAC DNA from nine BAC clones (Table 1) was isolated using the Nucleobond Xtra Midi Plus kit (Macherey Nagel) following the manufacturer's instructions, using 100 mL of LB media with a Chloramphenicol selective marker (at a concentration of 12.5 µg/mL). We obtained about 15 µg of DNA for each BAC clones (at a concentration of around 150 ng/µL). To estimate insert size, 150 ng of each BAC was digested with the fast NotI enzyme (Fermentas) for 40 minutes at 37°C and size-fractioned by PFGE (6 V/cm, 5 up to 15 s switch time, 16 h run time, 12.5°C) followed by ethidium bromide staining and visualization. Each insert size was assessed by comparison with PFGE standard size markers (New England Biolabs) using Genetools software (Syngene).

Sanger sequencing of BAC-ends and internal sequences

Sanger sequencing reactions were done using Big Dye Terminator chemistry v3.1 (Applied {Bibliography}Biosystems). Reactions were carried out on 2 µL of plasmid DNA (around 300 ng) following the protocol described using T7 and M13r universal primers for BAC-end sequencing and two designed primers for internal sequencing [10]. The last primers were designed on PacBio RS II sequence using Primer 3 software and the forward primer was designed at position 22591 bp (5'-TGGCATTTTCGTTCCCACAC-3') and reverse primer was designed at 24112 bp (5'-ATCCATTCCATCCATCGGCA-3') [11]. Reaction products were analyzed on an ABI 3730 DNA Analyzer (Applied Biosystems) at GeT-PlaGe, Toulouse [12]. A consensus sequence for the two internal sequences was obtained using Codon Code Aligner software (v. 5.1.5, Codon Code Corporation). Alignment between Sanger, PacBio RS II and Roche-454 sequences was then performed using MultAlin with Dayhoff symbol comparison table [13].

Clone Name	Common Names	Species	Cloning Vector	Estimated Insert Size (kb)
Frag-55O19	Strawberry	*Fragaria ananassa*	pIndigoBAC-5	90
Heli-337E08	Sunflower	*Helianthus annuus*	pIndigoBAC-5	170
Hord-155N13	Barley	*Hordeum vulgare*	pIndigoBAC-536	155
Sacc-241H10	Sugarcane	*Saccharum hybridum*	pIndigoBAC-5	150
Sacc-276O20	Sugarcane	*Saccharum hybridum*	pIndigoBAC-5	100
Trit-136P19	Bread wheat	*Triticum aestivum*	pIndigoBAC-5	120
Trit-131J6	Bread wheat	*Triticum aestivum*	pIndigoBAC-5	130
Zeam-34K24	Maize	*Zea mays*	pIndigoBAC-5	135
Zeam-100L1	Maize	*Zea mays*	pIndigoBAC-5	85
			Total Insert size	1 135

Table 1: Clone summary.



PFGE pattern of the sequenced BAC clones after NotI restriction enzyme digestion. M: Lambda Ladder PFG Marker (NEB), 1 to 9: BAC clone DNA digested with NotI enzyme. Lane 1: Frag-55019; lane 2: Heli-337E08; lane 3: Hord-155N13; lane 4: Sacc-241H10; lane 5: Sacc-276O20; lane 6: Trit-136P19; lane 7: Trit-131J6; lane 8: Zeam-34K24; lane 9: Zeam-100L1.

Figure 1: Insert size estimation.

Roche-454 sequencing

For the Roche-454 strategy, each BAC clone was individually tagged to allow DNA reads clustering prior to assembly. 1 µg of each individual BAC DNA was used to prepare individual tagged libraries using the GS FLX Titanium Rapid Library Preparation Kit (Roche). The BAC DNA was mechanically sheared to a size of approximately 1.5 kb using the Covaris DNA shearing system (M220 type, Kbioscience). The final size and the quality of the library were analyzed before sequencing using the Bioanalyzer

System (Agilent). We obtained DNA fragments bigger than 800 bp with an average of 1.5 kb. All the libraries were pooled to proceed with the emulsion-based clonal amplification (emPCR). After amplification, the beads binding the DNA were loaded on a GS FLX+ sequencer flowcell (2 regions) according to the manufacturer's instructions.

Roche-454 assembly

Raw datas were clustered according to tagging information to generate a standard flowgram format (sff) file per BAC clone. In order to eliminate low quality and highly repeated reads, the raw reads were cleaned with Pyrocleaner [14,15]. Then the remaining data were filtered against "*Escherichia coli* str. K12DH10B, complete genome" to eliminate contaminated reads. Finally, a standard assembly was performed on the cleaned reads (Newbler 2.9) with the *vs* option for screening the vector (Figure 2). Assembly was performed using clusters and tools of the GenoToul Bioinformatics platform Toulouse Midi-Pyrenees [16,17].

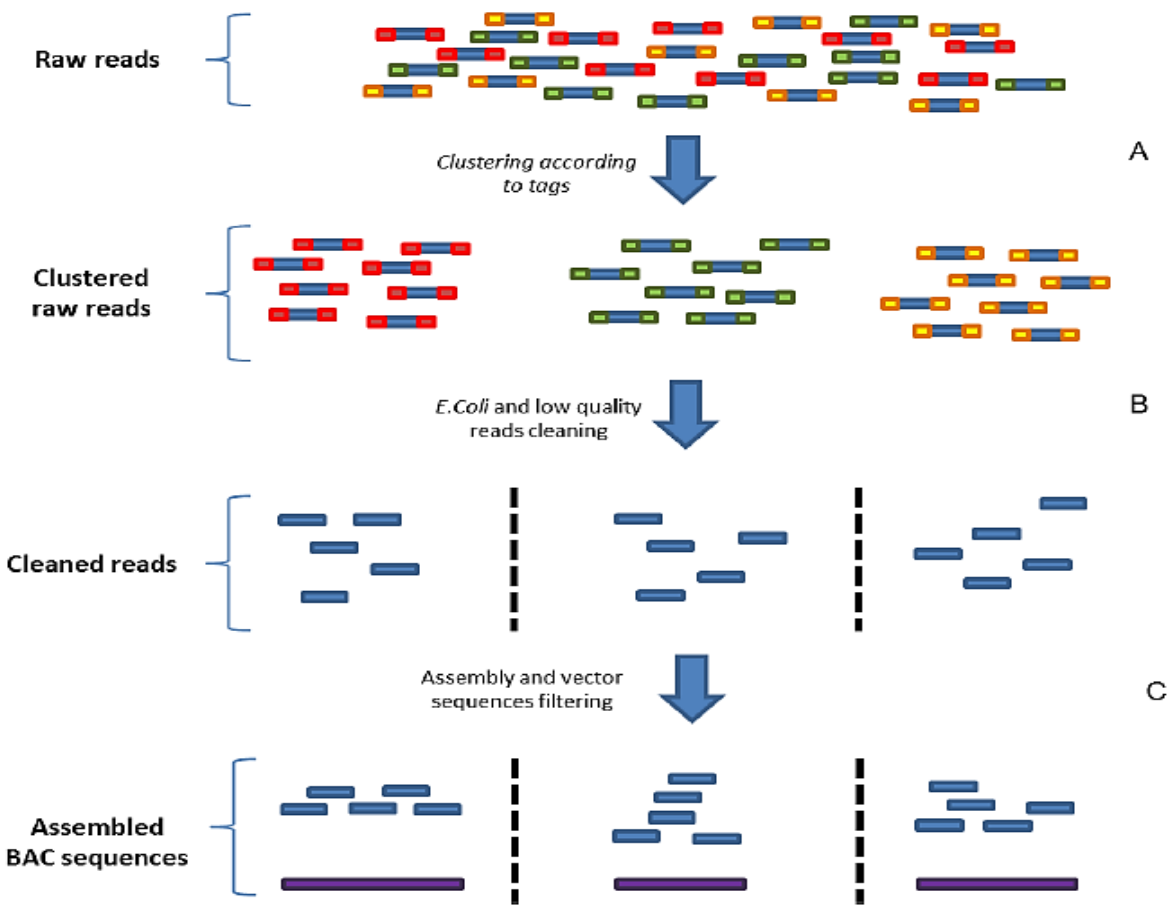
PacBio RS II sequencing

For PacBio RS II sequencing, BAC clone DNAs were pooled and sequenced all together without individual tags, in contrast to Roche-454 sequencing. 2 µg of each individual BAC DNA were pooled to obtain a total amount of 18 µg. One library was generated using the standard Pacific Biosciences library preparation protocol for 10 kb libraries. It was sequenced in one SMRT Cell using the P4 polymerase in combination with the C2 chemistry. The work was conducted following the standard operating procedures of the manufacturer (sequencing service provider using Pacific Biosciences PacBio RS II platform was GATC Biotech) [18].

PacBio RS II assembly

Assembly of the PacBio RS II reads was performed following the HGAP workflow [19]. The SMRT® Analysis (v2.2.0) software suite was used for HGAP implementation [20].

Reads were first aligned using BLASR against "*E. coli* str. K12 substr. DH10B, complete genome". Identified *E. coli* reads and low quality reads (read quality <0.80 and read length <500 bp) were removed from data used for the BAC clone sequences assembly [21,22]. Filtered reads were then preassembled to generate long and highly accurate



(A) Raw data are clustered according to tagging information. After clustering reads from each BAC clone are processed individually.
(B) Reads are filtered against E.coli sequence (command: blastall -p blastn -i reads.fasta -d ecoli -e 1e-5 -F F -b 150 -v 150 >ecoli.blast.seq). Among remaining reads, low quality and highly repeated reads are cleaned with Pyrocleaner (command: pyrocleaner.py -Indc -r 1000 -b 100 -e 10 -j 2 -k 300 -m 600 -s 4 -t 70 -v 100 -x 5 -y 40 -i reads_clean.sff -o output_folder). Remaining
(C) A standard assembly is performed on the cleaned reads by using Newbler 2.9 with the -vs option for screening the vector.

Figure 2: Roche-454 assembly workflow.

sequences. To perform this step, we separated the smallest and longest reads (e.g. >11 kbp) in order to correct read errors by mapping the first ones onto the second ones. Obtained sequences were filtered against vector sequences, and the Celera assembler is used to assemble the data and obtain a draft assembly. The last step of HGAP workflow was the "polishing" that significantly reduced the remaining InDel and base substitution errors in the draft assembly. The Quiver algorithm was used, which is a quality-aware consensus algorithm that uses the rich quality scores embedded in Pacific Biosciences bas.h5 files [23]. Once the "polished assembly" was obtained, each BAC was individualized by matching their BES on the ends of assembled sequences with BLAST. In order to calculate the error rate of PacBio RS II technology, the raw reads were aligned on the assembled sequences with BLASR and alignment results were analyzed using Qualimap software [24].

Sequence alignment

Comparison between Roche-454 and PacBio RS II assemblies for each clone was done using NCBI's MEGABLAST, using default settings. Alignment identity was calculated by dividing the total number of base matches by the total length of alignment (including substitutions, insertions, and deletions).

In order to compare the Roche-454 assembly and PacBio RS II assembly in greater detail, further analyses were performed on Zeam-34K24, Sacc-241H10 and Sacc-276O20 BAC clones. A dot plot between Roche-454 assembly and PacBio RS II assembly was created with YASS software with an E-value of 1.0E-10 [25].

In order to estimate read coverage for both technologies, alignments were performed using appropriate tools. Roche-454 reads were aligned on the Roche-454 assembly with the BWA-SW aligner [26,27]. PacBio RS II reads were aligned on the PacBio RS II assembly with BLASR. Coverage of each alignment was extracted from .bam files with Samtools in a .csv file. Graph coverage was generated from this file with Microsoft Excel [28,29].

The representation of homology between the Roche-454 assembly and the PacBio RS II assembly was generated using Synthviewer script based on MEGABLAST results [30].

Sequence annotation

BAC annotation was performed with TriAnnot pipeline using the maize default analysis template [31]. In addition, repeat elements were annotated using plant repbase databank and MIPS databank. LTR retroelements were located more precisely thanks to the identification of the LTRs terminate in the highly conserved TG…CA motifs and their target site duplication (TSD). Visualizations of annotations were done using Artemis genome browser [32].

Results and Discussion

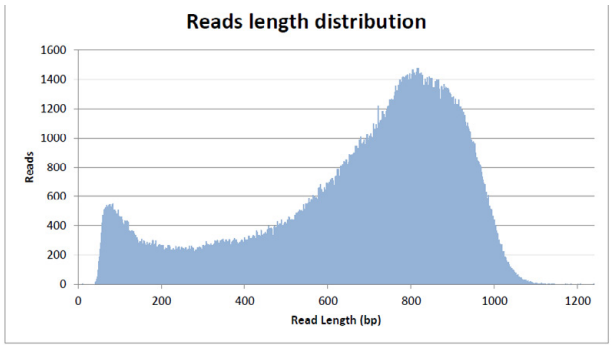
In the context of various BAC library screening projects, we sequenced the BAC clones that possess the region of interest using Roche-454 technology. The read lengths were >600 bp in average, which were the longest sequences produced by the different NGS platform. We found varying results that ranged from 1 contig per BAC to more than 20, mainly due to the mis-assembly of repeated sequences. Recently, PacBio RS II technology has emerged and seemed very promising for the assembly of complex genome or genomic regions due to the production of very long reads (>10 kb). To test this hypothesis, we compared assembled sequences of BAC clones obtained with either PacBio RS II technology or Roche-454 technology.

Several contigs per BAC clone with Roche-454 assembled sequences data

The GS Flex+ sequencing platform provided 416 Mbp of raw reads showing a mean insert size of 656 bp (Figure 3). After the cleaning process (as described in Materials and Methods), 312 Mbp were used for BAC clone assembly. The coverage per BAC clone ranged from 54 fold to 647 fold (with an average of 275 fold) (Table 2). For the comparative analysis of the two sequencing technologies, we decided to use the best assembly among various tries obtained by varying the coverage of raw reads used to assemble BAC sequences (data not shown). For seven BACs, we obtained the lowest numbers of contigs when using the whole set of reads. For the two remaining BAC clones, Frag-55O19 and Zeam-100L1, that were highly covered (602 fold and 940 fold of raw reads respectively), decreasing the coverage has led to a better assembly with a noticeable reduction in the contig number (Table 3). For these two BACs, we selected the assemblies obtained with a subset of raw reads of 200 fold for Frag-55O19 clone and 400 fold for Zeam-100L1 clone. Large contig numbers varied from 1-17 with an average of 7.3 large contigs per BAC clone (Table 4). The total length of the 66 large contigs represented 1138253 bp.

One contig per BAC clone with PacBio RS II assembled sequences data

One SMRT Cell of the PacBio RS II platform (P4C2 chemistry) provided a total amount of 404 Mbp of raw reads, showing a mean insert size of 2.69 kbp (Figure 4 and Table 5). The data were processed by our assembly pipeline (Figure 5). The first step of the pipeline, which consists of *E. coli* reads removal and low quality reads filtration, resulted in 71690 subreads, showing a mean insert size of 3.74 kbp for a total amount of sequences usable for assembly of 268 Mbp (Table 5). Maximum subread length was 25.31 kbp. Average sequence coverage was 233 fold per BAC clone. We used these subreads to perform reading error correction. We applied an 11 kbp threshold and used shorter subreads to correct longer subreads by alignment. At this stage, cloning vector sequences were trimmed. We obtained an average of 15 fold coverage with a 7.5 kbp mean size of long and accurate sequences that we used for the assembly. After a stage of sequence polishing, we obtained a final number of nine different linear contigs, which was consistent with the number of BAC clones that were pooled. The "finished" sequence of the nine contigs represented a total of 1159327 bp (Table 6), which was also consistent with the total BAC size of 1135 kbp experimentally estimated on pulsed field gel electrophoresis.



The GS Flex sequencing platform provided 634 927 raw reads for a total of 416 532 893 bp showing a mean insert size of 656 bp. The distribution shows a peak of read number around 800 bp. After remove low quality, contaminated and very short reads, we increased mean size to 690 bp.

Figure 3: Roche-454 sequencing metrics and raw reads distribution.

Raw Reads Mean Size	Raw Reads Total Size	Raw Read Coverage	% of E. *coli* Reads	Clean Reads Number	Clean Reads Total Size	Clean Reads Coverage
649	58366748	602	3	63988	43559941	449
669	48986008	258	6	56081	39484720	207
622	56714261	354	10	63141	41120486	257
659	10009146	69	13	11438	7855226	54
679	38384069	349	6	42124	30152470	274
657	44556769	446	8	49275	33869154	338
655	46362127	346	7	51913	35556692	265
666	47347835	364	9	50667	35374048	272
664	65805930	940	6	65473	45307125	647
Before performing assembly, Roche-454 raw reads were clustered according to tags in order to individualize each BAC and cleaned to eliminate low quality, highly repeated and contaminated reads. metrics of raw reads have been calculated after clustering for each BAC, and metrics of clean reads after removal of low quality and E. *coli* contaminated reads.						

Table 2: Per BAC sequencing metrics for Roche-454.

Clone name	Read coverage used for assembly	940X	600X	400X	300X	200X	100X	40X
Frag-55o19			4	3	3	2*	3	2
Zeam-100L01		3	1*	1	1	1	1	1
* Assembly retained for clone sequence analysis. For 7 of the 9 BAC clones analyzed in our study the whole sets of sequences have been used for assembly. Nonetheless for 2 BAC clones (Frag-55o19 and Zeam-100L1) that were highly covered (602-fold and 940-fold of raw reads respectively) we performed assemblies with different coverage.The decreasing of the coverage has led to a better assembly with a reduction of contig number.								

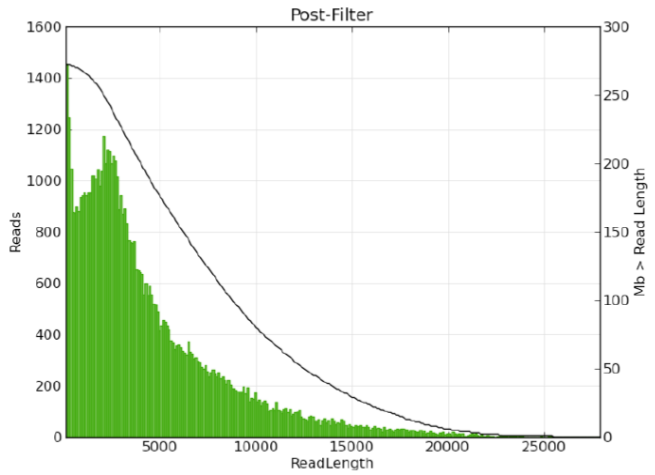
Table 3: Impact of read coverage used for Roche-454 assemblies on large contig number.

Clone Name	Contig Number	Large Contig Number (>500 bp)	N50 Length	Large Contig Max Size	Large Contig Mean Size	Large Contig Total Size
Frag-55o19	3	2	60315	59340	45124	90248
Heli-337E08	6	6	111775	111775	28912	173471
Hord-155N13	21	10	39023	62822	15555	155553
Sacc-241H10	20	17	18359	44090	7921	142570
Sacc-276o20	1	1	105854	105854	105854	105854
Trit-136p19	5	2	105476	105476	62395	124789
Trit-131J06	19	17	15872	34546	7379	125436
Zeam-34K24	12	10	20463	33950	11641	128036
Zeam-100L01	1	1	86647	86647	86647	86647
Assembly of Roche-454 reads performed with Newbler using the -vs option for screening the vector.For each BAC assembly we indicated the number of obtained contigs, the longest, the total and the mean size of contigs. N50 length corresponds to the length of the contig for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs.						

Table 4: Assembly metrics of Roche-454 reads.

Metric name	Metric value
Raw reads number	150292
Raw reads total size (bp)	404105727
Raw reads mean size (bp)	2689
Raw reads coverage (X)	349
E.*coli* contamination (number of E.*coli* reads out of raw reads expressed in percentage)	1.12%
Filtered subreads number	71690
Filtered subreads total size (bp)	268149966
Filtered subreads mean size (bp)	3740
Filtered subreads max size (bp)	25312
Filtered subreads coverage (X)	233
Filtered subreads greater than 25 kbp	2
Filtered subreads greater than 20 kbp	57
Filtered subreads greater than 15 kbp	702
Filtered subreads greater than 10 kbp	3698
Filtered subreads greater than 5 kbp	16069
Filtered subreads greater than 3 kbp	32529
Assembly of PacBio RS II reads was performed following HGAP workflow. The first step of the pipeline was to filter E. *coli* and low quality reads from raw data. The table shows metrics on raw data and after filtering step.	

Table 5: PacBio RS II sequencing metrics.



Distridution of subreads obtained after filtration (Assembly workflow on Figure 5) shows a pic shorter reads around 2500 bp. Shorter reads are used to correct error of the longer subreads (>11 kbp).

Figure 4: PacBio Rs II subreads length distribution after filtration.

Because the BACs were not tagged, we used Sanger sequences of BAC-ends to link contigs to BAC clones (Table 7). The nine contigs were assigned to the nine BAC clones unambiguously and the sequence size was equivalent to the BAC clone size estimated on agarose gel. We mapped the raw reads on the nine contigs in order to estimate the coverage per BAC. The coverage ranged from 77-372 fold per BAC clone (Table 6).

Gaps and collapsing issues solved with long reads technology

One contig per BAC clone was obtained with PacBio RS II sequencing while an average of 7.3 large contigs per BAC clones were obtained with Roche-454 (Table 8). For all 9 BAC clones, the total length of the sequence was relatively similar with both technologies (1138253 bp Roche-454 *vs* 1159327 bp PacBio RS II), but systematically smaller for each individual BAC clone with the Roche-454 technology. The differences in length increased significantly when the BACs were assembled in more than 6 contigs with Roche-454, except for the Hord-155N13 BAC clone. For each clone, this difference in length found was probably due to missing data in Roche-454 assembled sequences. The two main reasons that could explain these differences in length are: (i) the limitation on read size of the Roche-454 technology, resulting in the collapse of repeated regions, (ii) the sequence contexts in terms of GC- or AT- rich regions or palindromic sequences, which are known to frequently not be covered by the 2nd generation sequencing technologies including Roche-454 and, conversely, to be resolved by PacBio RS II technology [33-36]. To determine the reasons for the differences in length between PacBio RS II and Roche-454 assemblies, we aligned each PacBio RS II sequence on the corresponding Roche-454 assembled sequence contigs using NCBI's MEGABLAST with default settings [37]. These alignments showed that some regions of Pacbio RS II sequences do not align to Roche-454 sequences while all Roche-454 sequences align to Pacbio RS II sequences. It implied missing data in Roche-454 sequences. Moreover, for the four BAC clones that showed significant differences in length between PacBio RS II and Roche-454 sequences, some Roche-454 contigs aligned in two PacBio RS II sequence positions. These regions correspond to repeat elements that are collapsed in Roche-454 assembly and correctly assembled with the PacBio RS II technology. These regions are longer than 700 bp and present a percentage of identity greater than 99%. Thanks to long read sequencing, PacBio RS II technology served to overcome this problem.

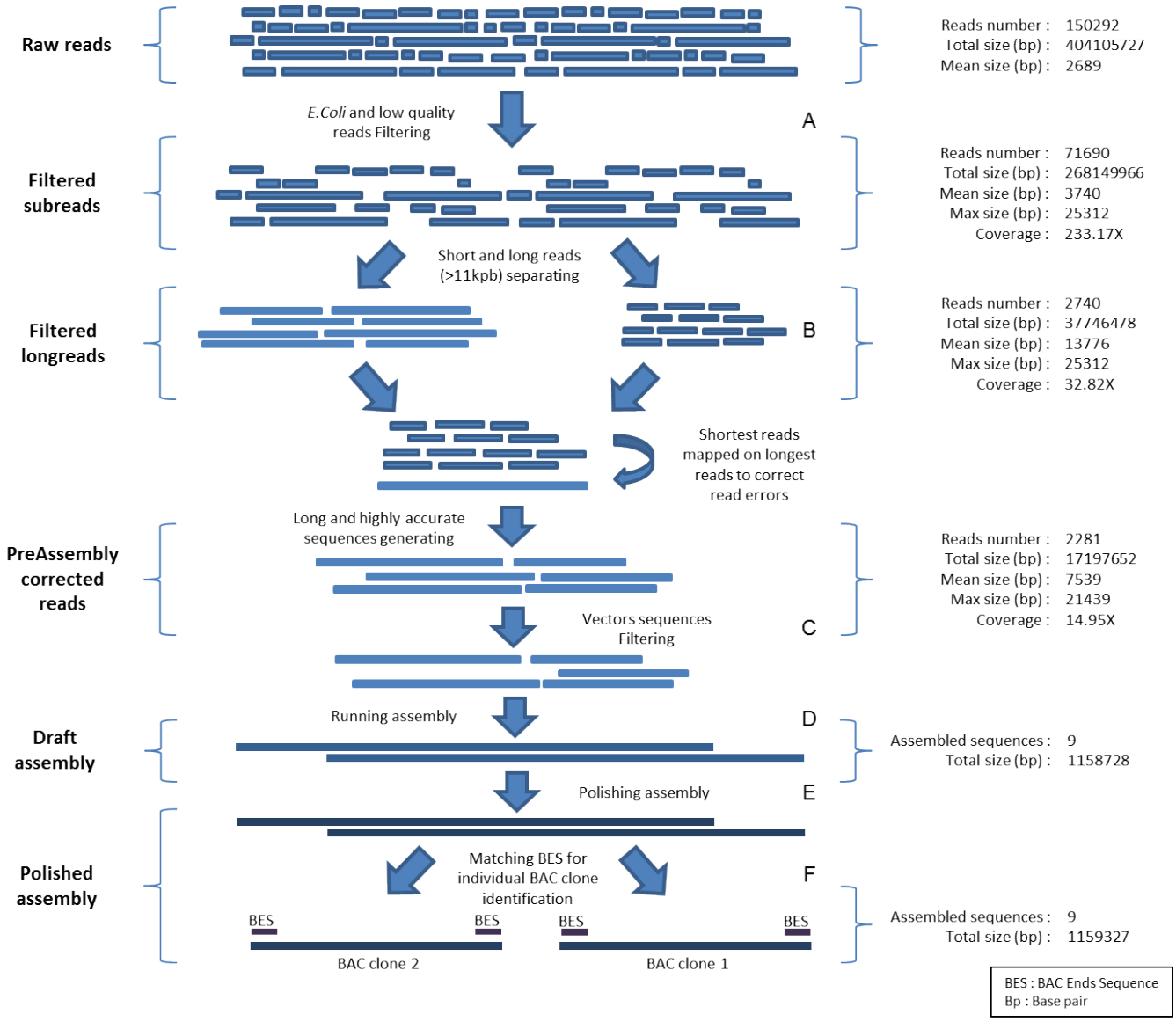
Error rate corrected

We have noted that PacBio RS II sequencing allowed us to obtain a high quality assembly that minimizes the number of contigs and resolves gaps and repeated region problems. In order to complete these results, it was important to evaluate the reliability of the sequences obtained in terms of nucleotide calling accuracy. Read errors in PacBio RS II raw reads is known to be high and randomly distributed in single-molecule sequencing [38]. By mapping the raw reads on the nine contigs we estimated average error rates in raw reads ranging from 14.6%-14.96% depending on the BAC clones (Table 6). Our assumption was that randomly distributed errors would be corrected by using a sufficient coverage of raw reads. Thus the consensus sequence obtained would represent the biological reality. We identified differences in terms of substitution and insertions/deletions between PacBio RS II and Roche-454 sequences based on the previous alignment between each PacBio RS II and Roche-454 assembled sequence. We found a total of 167 mismatches (unique events) on a total sequence alignment of 1132 kbp, resulting in a 99.98% identity between the two assemblies (Table 9). The estimation was done by taking into account the number of substitutions and insertions/deletions as unique events or by base. Generally, we found very few substitutions. Among the nine BAC clones analyzed, five exhibited no substitution at all, and three had 1-4 substitutions for an average sequence size of 126 kb. Overall, substitutions represented 14% of the mismatches detected. The major part of the mismatches corresponds in fact to insertions (86% of the mismatches). It is worth noting that there is a difference in the two assemblies as most of the insertions were found in the PacBio RS II assembly compared to the Roche-454 assembly. To determine if there are true insertions in the PacBio RS II sequencing or bases that are missing in the Roche-454 sequencing we have sequenced by Sanger technology 4 loci from the Heli-337E08 BAC clone that exhibited the highest number of insertions in the PacBio RS II assembly among the 9 BAC clones studied. For the 6 cases Sanger sequencing confirms PacBio assembly and it appears that bases are deleted by the 454 technology (Figure 6). On the whole, the error rate measured in PacBio RS II raw reads was corrected by using sufficient coverage. The sequences obtained after the polishing assembly stage (Figure 5) showed a strong

Clone Name	Contig Number	Mean Coverage	QV*	Assembly Total Size	Reads Error Rate (%)
Frag-55O19	1	90	48.56	90557	14.72
Heli-337E08	1	141	48.55	175563	14.96
Hord-155N13	1	235	48.54	155889	14.68
Sacc-241H10	1	77	48.54	152547	14.62
Sacc-276O20	1	372	48.54	105851	14.8
Trit-136P19	1	206	48.54	125714	14.77
Trit-131J6	1	81	48.54	133334	14.7
Zeam-34K24	1	119	48.55	133221	14.6
Zeam-100L1	1	229	48.55	86651	14.74

*QV (Qualite Value): A prediction of the error probability of a basecall (QV40=1 in 10,000, QV50=1 in 100,000); Assembly of PacBio RS II reads was performed following HGAP workflow (see Assembly workflow on Figure 5). For eacg BAC assembled we indicated the number of generated contigs, the mean coverage, the mean QV and the total size. The coverage was estimated by mapping the raw reads on the assembly contigs with assembly contigs with the BLASR software. The reads erroe rate was estimated with Qualimap software.

Table 6: Assembly metrics of PacBio RS II reads.



Assembly of PacBio RS II reads was performed following HGAP workflow. (A) Low quality reads were filtered from raw data. (B) Filtered subreads were then preassembled to generate long and highly accurate sequences. To perform this step smallest and longest reads were partitioned according to an 11 kb cutoff to correct read errors by mapping the first ones on the second ones. (C) Cloning vectors sequences were removed before performing assembly (D). (E) Last step of the assembly was the polishing that significantly reduced the remaining InDel and base substitution errors in the draft assembly. (F) Finally assembled sequences are identified (i.e. linked to BAC clones names) by matching BES.

Figure 5: Assembly workflow of PacBio RS II reads and related metrics.

similarity with Roche-454 sequences.

More accuracy of the assembled data with long reads sequence technology

In order to investigate the type of sequence responsible for Roche-454 and PacBio RS II differences in term of contig numbers after assembly, we focused on Zeam-34K24 BAC clone. Results of assemblies for this clone were 1 contig with PacBio RS II reads and 10 contigs with Roche-454 reads. We aligned Roche-454 and PacBio RS II contigs of Zeam-34K24 BAC clone (Figure 7). Seven of the Roche-454 contigs showed a relevant homology with one single region of the PacBio RS II sequence, while two of them (ctg08 and ctg10) displayed strong homologies with two distinct regions. Overall, contig ctg08 (3305 bp) aligned completely to PacBio RS II sequence at positions 59838 to 63143 and 64452 to 67757 with, respectively, 99.91% and 99.85% homology. Contig ctg10 (1224 bp) aligned completely to PacBio RS II sequence at positions 97538 to 98761 and 105692 to 106914 with, respectively, 100% and 99.26% homology. We analyzed the reads coverage on the contigs generated with the two specific technologies. We observed that Roche-454 reads coverage exhibited two spikes corresponding to these two particular contigs of the Roche-454 assembly, while PacBio RS II reads coverage remained stable. Alignment of the PacBio RS II contig of Zeam-34K24 BAC clone against itself showed repeated sequences that co-localized with these same two mis-assembled

Query (BAC End sequence)	Subject (Clone name)	% Identity	Alignment Length	Mis Matches	Gap Opens	Query Start	Query End	Subject Start	Subject End	e-value
Frag-55O19 T7	Frag-55O19	100.00	825	0	0	1	825	1	825	0.0
Frag-55O19 M13r	Frag-55O19	100.00	722	0	0	1	722	90510	89789	0.0
Heli-337E08 T7	Heli-337E08	100.00	781	0	0	1	781	175561	174781	0.0
Heli-337E08 M13r	Heli-337E08	100.00	780	0	0	7	786	1	780	0.0
Hord-155N13 T7	Hord-155N13	99.87	773	1	0	1	773	155841	155069	0.0
Hord-155N13 M13r	Hord-155N13	100.00	850	0	0	1	850	852	3	0.0
Sacc-241H10 T7	Sacc-241H10	99.87	772	1	0	1	772	152547	151776	0.0
Sacc-241H10 M13r	Sacc-241H10	100.00	729	0	0	53	781	1	729	0.0
Sacc-276O20 T7	Sacc-276O20	100.00	797	0	0	1	797	105851	105055	0.0
Sacc-276O20 M13r	Sacc-276O20	100.00	790	0	0	12	801	1	790	0.0
Trit-136P19 T7	Trit-136P19	100.00	762	0	0	1	762	1	762	0.0
Trit-136P19 M13r	Trit-136P19	99.87	757	1	0	1	757	125712	124956	0.0
Trit-131J6 T7	Trit-131J6	100.00	774	0	0	1	774	133334	132561	0.0
Trit-131J6 M13r	Trit-131J6	100.00	751	0	0	1	751	43	793	0.0
Zeam-34K24 T7	Zeam-34K24	100.00	794	0	0	1	794	1	794	0.0
Zeam-34K24 M13r	Zeam-34K24	99.87	781	1	0	1	781	133219	132439	0.0
Zeam-100L1 T7	Zeam-100L1	100.00	822	0	0	1	822	86651	85830	0.0
Zeam-100L1 M13r	Zeam-100L1	100.00	784	0	0	10	793	1	784	0.0

After assembly, the obtained contigs were individualized by matching the BES (BAC-End Sequences) of the sequenced clones on assembled sequences. The mapping was performed using NCBI's MEGABLAST with default settings. The table shows the output of BLAST results. BES match correctly on the extremities on the corresponding assembled contig. The 4 mismatch that we found are "N" in the BES.

Table 7: Blast results of BES on PacBio RS II assembly.

Clone Name	Estimated Insert Size (kb)	Roche-454 Contigs[a]	PacBio RS II Contigs	Roche-454 Size (bp)	PacBio RS II Sze (bp)	Roche-454/PacBio RS II Size Ratio
Frag-5O19	90	2	1	90248	90557	0.99659
Heli-37E08	170	6	1	173471	175563	0.98808
Hord-155N13	155	10	1	155553	155889	0.99784
Sacc-241H10	150	17	1	142570	152547	0.93460
Sacc-276O20	100	1	1	105837	105851	0.99987
Trit-136P19	120	2	1	124789	125714	0.99264
Trit-131J6	130	17	1	125436	133334	0.94077
Zeam-34K24	135	10	1	128036	133221	0.96108
Zeam-100L1	85	1	1	86647	86651	0.99995

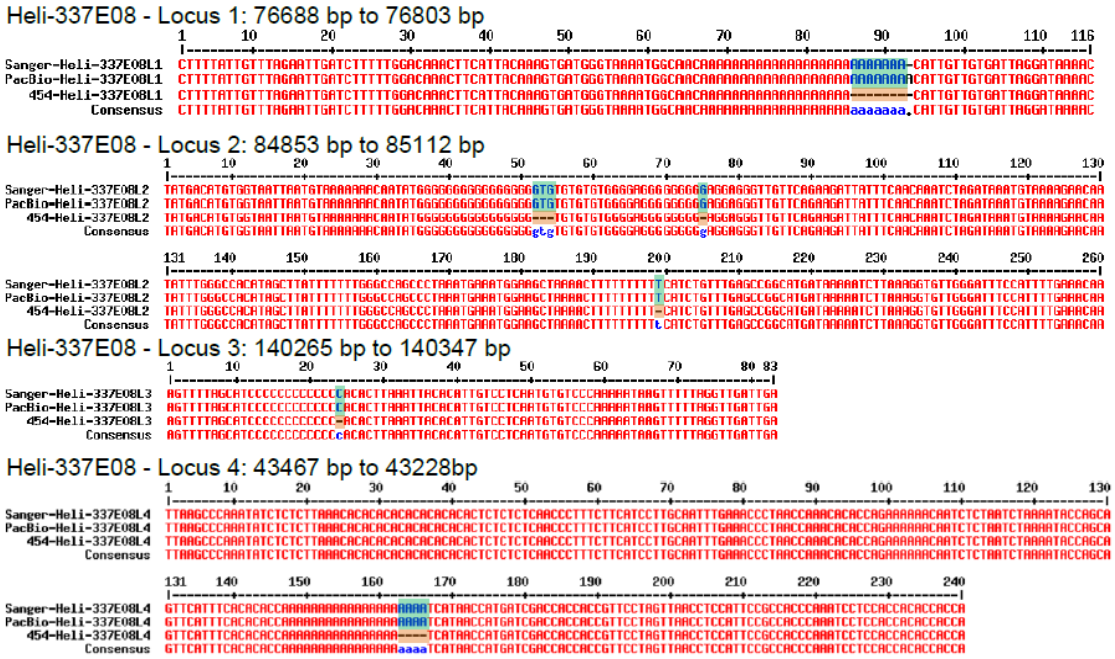
Assembly of PacBio RS II reads was performed following HGAP workflow. A standard assembly of Roche-454 reads was performed with Newbler after filtering low quality, *E. coli*, and vectors reads. Sizes of assemblies were compared by dividing each Roche-454 assembly size by corresponding PacBio RS II assembly size and were confirmed with the insert size estimated by a NotI restriction enzyme digestion. PacBio RS II assembly lowered the number of generated contigs compared to Roche-454 assembly. [a]Large contigs > 500 bp.

Table 8: Summary of assembly results.

Clone Name	Matches[a]	Substitutions[b]	PacBio RS II Insertions[b]	Roche-454 Insertions[b]	Mismatches[b]	Per-Base identity[c]	Per Event Identity[d]
Frag-55O19	90248	0 (0)	23 (15)	0 (0)	23 (15)	0.999745	0.999834
Heli-337E08	173471	0 (0)	85 (42)	0 (0)	85 (42)	0.999510	0.999758
Hord-155N13	155546	4 (4)	15 (11)	1 (1)	20 (16)	0.999871	0.999897
Sacc-241H10	142520	1 (1)	18 (15)	10 (10)	29 (26)	0.999797	0.999818
Sacc-276O20	105837	0 (0)	14 (12)	0 (0)	14 (12)	0.999868	0.999887
Trit-136P19	124788	0 (0)	5 (5)	0 (0)	5 (5)	0.999960	0.999960
Trit -131J6	125194	16 (16)	28 (20)	4 (4)	48 (40)	0.999617	0.999681
Zeam-34K24	128029	3 (3)	5 (5)	0 (0)	8 (8)	0.999938	0.999938
Zeam-100L1	86644	0 (0)	7 (3)	0 (0)	7 (3)	0.999919	0.999965

Comparison between Roche-454 and PacBio RS II assemblies for each clone was performed with BLAST using NCBI's default MEGABLAST setting. Only best hit was conserved for each Roche-454 contigs. The table brings out the high similarity and low mismatches between Roche-454 and PacBio RS II.
[a] Matching bases determined by BLAST alignments of PacBio RS II and Roche-454 assemblies.
[b] Total differences between assemblies by base and by unique event in parentheses.
[c] Percentage of identity between assemblies based on total matches divided by matches plus mismatch bases.
[d] Percentage of identity between assemblies based on total matches divided by matches plus mismatch events.

Table 9: Summary of alignments between PacBio RS II and Roche-454 assemblies.



For the 6 cases Sanger sequencing confirms PacBio assembly and it appears that inserted bases are the result of deletions in the Roche-454 assembly.

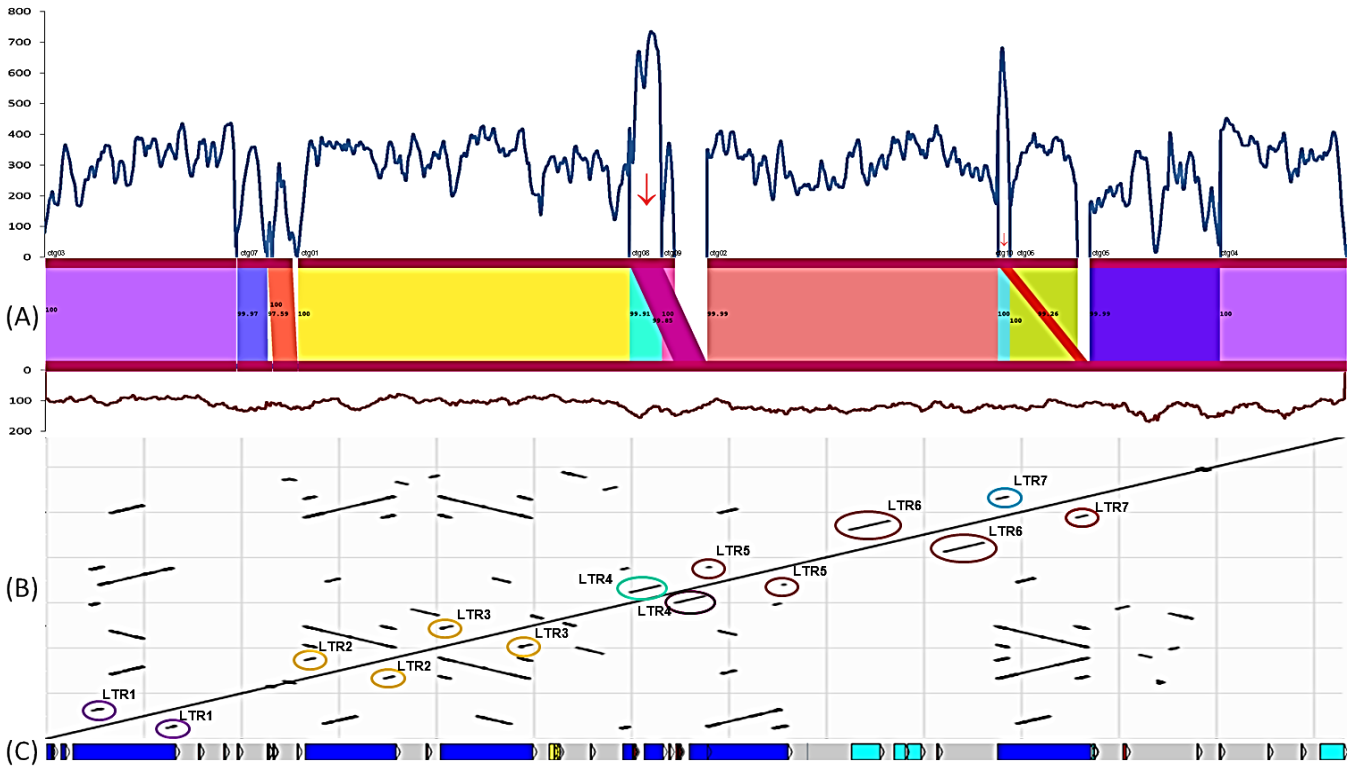
Figure 6: Sanger sequencing of 4 loci that exhibits insertions in PacBio RS II assembly when compared to Roche-454 assembly.

Roche-454 contigs. These two pairs of repeated sequences presented a high sequence similarity (>99%) and were longer than 1000 bp. These repeated sequences correspond to the long terminal repeat (LTR) of retro-elements belonging to the copia superfamily based on the BAC sequence annotation. Other LTR retro-elements were predicted among this sequence, but either the LTRs were longer than 1000 bp and presented less than 97% of homology, or LTRs presented strong homology (>99%) and were shorter than 600 bp. In these last two cases, the Roche-454 sequence was well assembled onto two different regions instead of one collapsed region. All these results on repeated sequences confirmed the previous observation done on the four BAC clones which presented a great length difference. Moreover, we analyzed the content of repeated sequences in two sugarcane BAC clones exhibiting high variation in the Roche-454 assembly results (1 contig for Sacc-276O20 BAC clone versus 17 contigs for Sacc-241H10 BAC clone). We showed that Sacc-241H10 BAC clone possesses more complex and larger repeated sequences than Sacc-276O20 BAC clone (Figure 8). A high level of similarity in large sequences is likely to explain the failure of assembly of Roche-454 reads whereas the length of PacBio RS II reads allows one to rely on sequences that flank repeated sequences to obtain a true assembly.

Contig 7 of Roche-454 Zeam-34K24 BAC clone assembly showed a difference from the PacBio RS II sequence. The first 3115 bp of this contig aligned with the PacBio RS II sequence at positions 19652-22766 with 99.97% homology and the following 2533 bp of this contig aligned with the PacBio RS II sequence at positions 23312-25844 with 100.0% homology, thus generating a region of 545 bp in the PacBio RS II sequence (22767-23311 bp) that was not covered by the Roche-454 sequence. This PacBio RS II region was partially covered by a 415 bp region present in contig ctg07 according to blast results with only 97.59% homology. This 415 bp region also aligned with the PacBio RS II sequence at positions 22747-22766 and 23312-23706 regions included in the previous alignments with more than 99.9% homology. This 415 bp region is thus present twice in the PacBio RS II sequence and only once in the Roche-454 sequence. The remaining 131 bp present in the PacBio RS II sequence and absent in the Roche-454 sequence corresponds to AC-, AT- and AG- tandem repeat regions. In order to confirm whether the PacBio RS II sequence or the Roche-454 sequence is the accurate sequence, we designed two primers that were flanking this region and sequenced them using the Sanger method. We aligned the consensus sequence obtained in Sanger with PacBio RS II and Roche-454 sequences and showed that the Sanger sequence confirmed the PacBio RS II assembly (Figure 9). There was only one gap between these two sequences corresponding to the absence of two nucleotides (TA) in $(AC)_{19}(AT)_{26}(GA)_{13}$ motif in the Sanger sequence (position 654-655 bp of Sanger). Moreover, Sanger sequencing confirmed the absence of a 545 bp region in Roche-454 sequencing. This missing region in Roche-454 sequencing corresponds to a 497 bp repeated sequence in tandem based on blast results and a lack of (TA) repetition. Besides the correct assembly of a repeated region, PacBio RS II sequencing served to assemble correctly an AC-, AT- and AG- rich region where Roche-454 sequencing failed.

Conclusion

Eukaryotic genomes are more complex than prokaryotic genomes, and among eukaryotes, plant genomes are much more complex than any other [39]. Plant genomes are complex in many aspects, including large sizes that often reach gigabases, high ploidy levels and a high percentage of repetitive elements that may represent the majority of the genome size [40]. The next-generation sequencing (NGS) technologies have revolutionized genomic research in several domains by providing the capacity to obtain large amounts of DNA sequences in a short time. However, this approach is not sufficient to decipher the high complexity of plant genomes, mainly due to the small size of the reads



(A) Alignment between Roche-454 (top) and PacBio RS II (bottom) assembled sequences of Zeam-34K24 BAC clone (respectively 10 contigs and 1 contig) and reads coverage. Alignment shows that 2 Roche-454 contigs align each at two distinct positions on the PacBio RS II contig (Red arrows). For these 2 contigs, Roche-454 reads coverage exhibits 2 spikes (at more than 600-fold coverage) while PacBio RS II reads coverage is uniform (around 100 fold coverage). This is consistent with a real duplication of the region. (B) Dot plot of PacBio RS II Zeam-34K24 BAC clone sequences against itself shows repeated sequences corresponding to LTR elements (LTR1: 1193 bp with 96.76% identity; LTR2: 1169 bp with 95.82% identity; LTR3: 1247 bp with 96.49% identity; LTR4: 3305 bp with 99.82% identity; LTR5: 516 bp with 99.81% identity; LTR6: 4041 bp with 96.72% identity; LTR7: 1328 bp with 99.17% identity; colors of the circles were given according to the Figure A). (C) Annotation of Zeam-34K24 PacBio RS II contig. In yellow: gene, in dark blue: LTR retroelement belonging to the copia superfamily, in light blue: LTR retroelement belonging to the Gypsy superfamily, in red: repeated motif. The two contigs of the Roche-454 assembly that align each at 2 positions on PacBio RS II contig correspond to the 2 pairs of LTR longer than 1000 bp with more than 99% of identity. Roche-454 contig 7 that is split at 2 positions on PacBio RS II contig colocalizes with repeated motives $(AT)_{52}$, $(TACA)_{20}$, $(AT)_{34}$.

Figure 7: Comparison of Zeam-34K24 PacBio RS II and Roche-454 assemblies.

produced. Indeed, high quality genome sequence references are still lacking despite the progress made with NGS technology.

By dividing genomes into relatively small units and thus reducing their complexity, BAC libraries are still a relevant tool to answer various scientific questions in order to isolate genes of interest, to characterize trait locus, to contribute to positional cloning, to study polymorphism, genome evolution and genetic variability, or to explore biodiversity. We have focused our expertise on a strategy aimed at directly targeting genomic regions of interest in specific genotypes and rapidly isolating clones that span a genomic region. This strategy, combined with NGS technology, has proven to be an efficient way to directly target genomic regions of interest and explore the variability among specific genotypes (unpublished observations) [41]. However, even when reducing the complexity of a whole genome analysis using BAC clones, it is not always easy to obtain a high-quality and reliable sequence assembly. The complexity of the assembly usually depends on the presence of repeat sequences, which complicates the assembly. With the advent of PacBio RS II single-molecule, real-time sequencing technology, the read length has been significantly increased. The aim of this study was to compare two technologies (Roche-454 and PacBio RS II technologies) in order to determine if longer reads can affect the quality and reliability of the sequence assembly. Our work, although done on a short number of BAC clones, clearly showed the value of long reads technology in the assembly of genome sequences and consequently in the accuracy of the data generated (avoiding collapsing of duplicated regions due to sequence homology for example). Defining which technology is the most reliable in terms of assembly, and consequently in terms of annotation, is very critical.

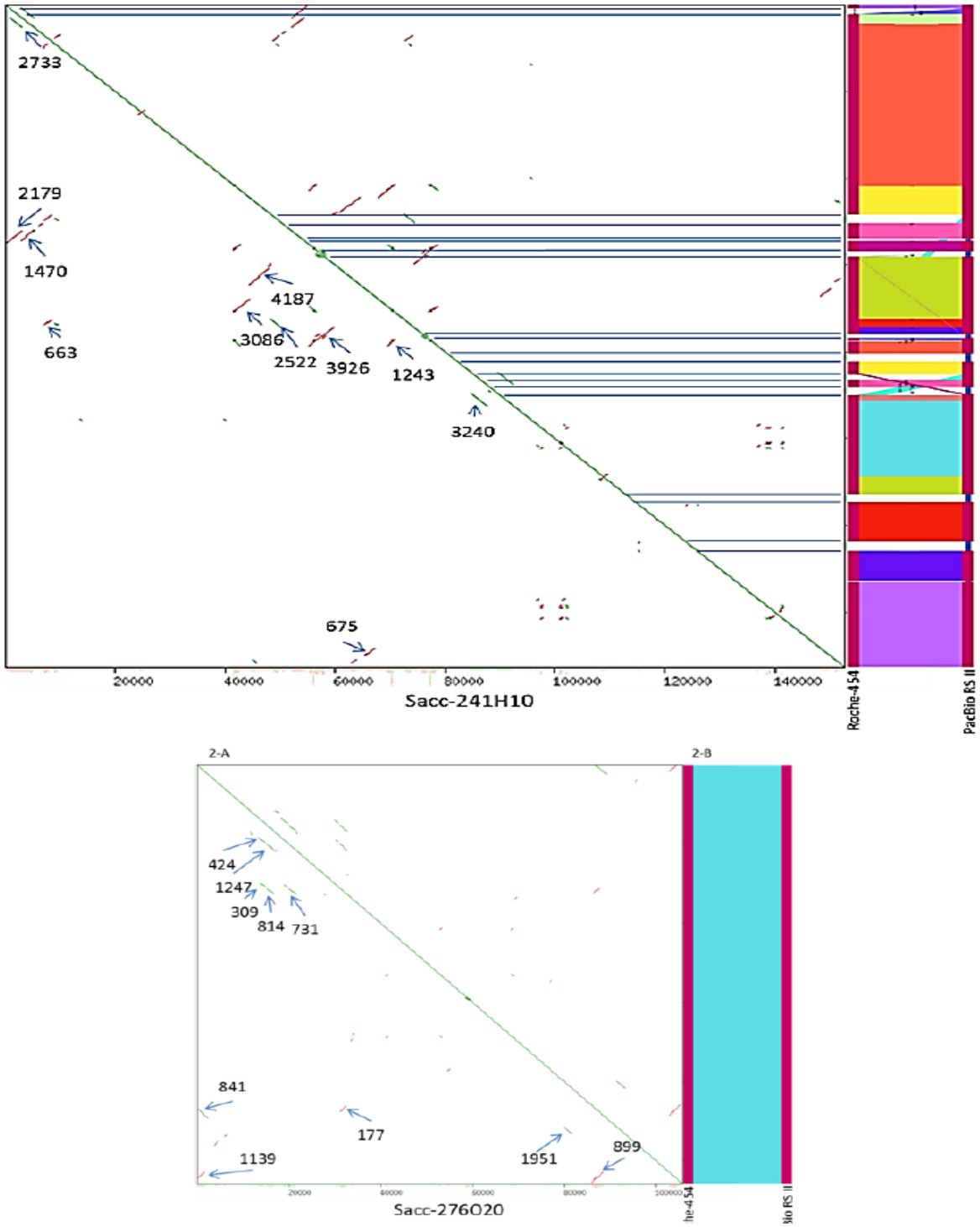
Overall, we conclude that the existence of very long reads is clearly an important advantage when working with complex genomes such as plant genomes with duplicated genes, with high level of repetitive elements or genome duplication.

Availability of Supporting Data

The sequences assemblies of the *Zea mays* Zeam-34K24 BAC clone have been submitted to the NCBI GenBank under accession number KR233325.

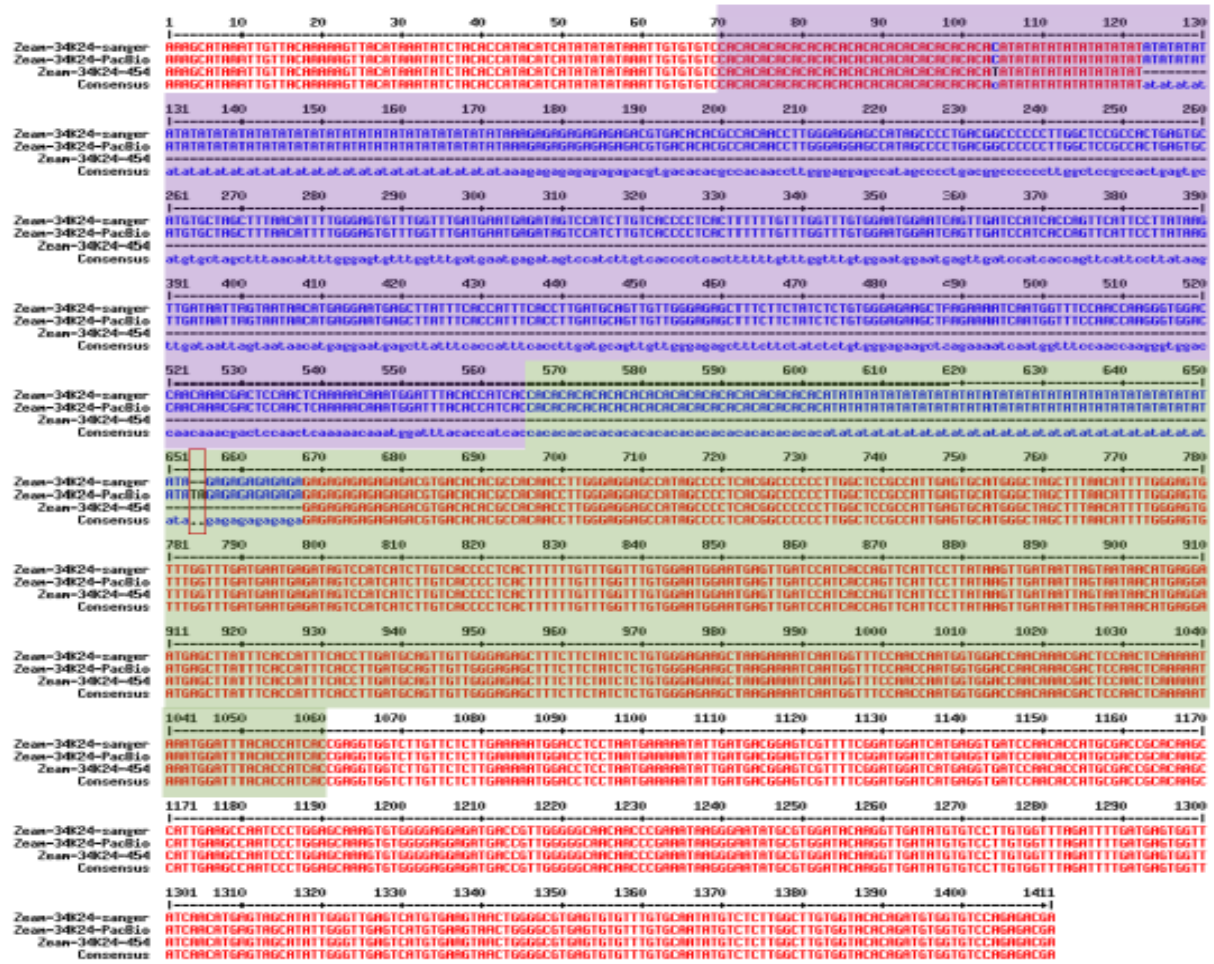
Acknowledgements

We would like to thank Vitte, Denoyes-Rothan, Munos, Augusto, the seed companies involved in the PROMAÏS - INRA ZeaWall network (Caussade Semences, Limagrain Genetics, MaïsAdour, Monsanto SAS, Pioneer Génétique, Pau Euralis, R2n RAGT Semences, Syngenta seeds), for their constructive collaboration and their agreement in using the data obtained from the selected BAC clones described in this article. Audrey Courtial was funded by the PROMAÏS - INRA ZeaWall network.



The repeated sequences content of 2 sugarcane BAC clones (Sacc-241H10 and Sacc-276O20 assembled in 17 and 1 contigs respectively with Roche-454 technology) was analyzed. The PacBio RS II assembly of each BAC was alligned against itself by a dot plot (1-A and 2-A). Secondary diagonals on the graph correspond to reapeted sequences. For each BAC clone the tenth highest alignment Bit scores are indicated on the dot plot (blue arrows). A high value of the Bit score corresponding to long regions with high homology. Sacc-241H10 BAC clone shows higher levels of Bit scores compared to Sacc-276O20 BAC clone. For each BAC clone, contigs assembled with Roche-454 technology were aligned onto the corresponding cintig(s) assembled with PacBio RS II sequences (1-B and 2-B. The alignment breakthroughs for Sacc-241H10BAC clone clearly occured in repeated regions with high value of Bit scores (highlighted with the blue lines linking 1-A and 1-B parts of the graph).

Figure 8: Repeated sequences content of two BAC clones showing high variation in contig numbers after Roche-454 assembly (1 *versus* 17 contigs).



Sanger sequencing was performed using two promers designed on PacBio RS II sequence (forward primer designed in position 22591 bp and reverse primer designed in 24112 bp). The consensus sequence obtained from the two Sanger sequences was aligned with PacBio RS II and Roche-454 sequences using Multalin.

Sanger sequencing confirmed PacBio RS II sequencing with only a gap between these two sequences corresponding to a missing of two nucleotides (TA) in (AC)19(AT)26(GA)13 motif in Sanger sequence (position 654-655 pb of Sanger, nucleotides circled in red). Moreover, Sanger sequencing confirmed the missing of a 545 bp region in Roche-454 sequencing. This missing region in Roche-454 sequencing correspond to a 497bp repeated sequence in tandem (one repeat underlined in purple and one repeat underlined in green) according to blast results and a lack of (TA) repetition.

Figure 9: Alignment of Sanger, PacBio RS II and Roche-454 sequences of Zeam-34K24 BAC clone.

References

1. Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, et al. (2014) Plant genome sequencing - applications for crop improvement. Curr Opin Biotechnol 26: 31-37.
2. Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011) Crop genome sequencing : lessons and rationales. Trends Plant Sci 16: 77-88.
3. Schatz MC, Witkowski J, Mccombie WR (2012) Current challenges in de novo plant genome sequencing and assembly. Genome biol 13: 243-249.
4. Doležel J, Vrána J, Cápal P, Kubaláková M, Burešová V, et al. (2014) Advances in plant chromosome genomics. Biotechnology Adv 32: 122-136.
5. Ariyadasa R, Stein N (2012) Advances in BAC-based physical mapping and map integration strategies in plants. J Biomed Biotechnol .
6. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, et al. (2014) Reconstructing complex regions of genomes using long-read sequencing technology. Genome research 24: 688-696.
7. Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitakeet K, et al. (2014) Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. BMC Genomics 15: 688-675.
8. English AC, Richards S, Han Y, Wang M, Vee V, et al. (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PloS one 7: e47768.
9. Ferrarini M, Moretto M, Ward J a, Šurbanovski N, Stevanović V, et al. (2013) An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics 14: 670.
10. Santos A, Penha H, Bellec A, Munhoz C De, Pedrosa-Harand A, et al. (2014) Begin at the beginning: A BAC-end view of the passion fruit (Passiflora) genome. BMC Genomics 15: 816.
11. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, et al. (2012) Primer3: new capabilities and interfaces. Nucleic acids research 40: e115.
12. GeT (Genome et Transcriptome) GenoToul website. http://get.genotoul.fr/.
13. Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Research 16: 10881-10890.
14. PyroCleanerversion.https://mulcyber.toulouse.inra.fr/plugins/mediawiki/wiki/pyrocleaner/index.php/Pyrocleaner.
15. Mariette J, Noirot C, Klopp C (2011) Assessment of replicate bias in 454 pyrosequencing and a multi-purpose read-filtering tool. BMC Research Notes 4: 149.
16. 454 Life Sciences website. http://www.454.com/products/analysis-software/. Version 2.9.

17. BioinfoGenoToul website. http://bioinfo.genotoul.fr/.

18. GATC Biotech website. https://www.gatc-biotech.com/en/index.html.

19. HGAP on Pacific Biosciences repository website. https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/HGAP. Accessed 28 Febr 2014.

20. The SMRT® Analysis. http://www.pacb.com/support/software-downloads/. Version 2.2.0.

21. Chaisson MJ, Tesler G (2012) Taxon ordering in phylogenetic trees by means of evolutionary algorithms Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC bioinformatics 13: 238.

22. BLASRon Pacific Biosciences repository website. https://github.com/PacificBiosciences/blasr. Accessed 17 Sept 2015.

23. Quiver embedded in The SMRT® Analysis Versions 2.2.0.http://www.pacb.com/support/software-downloads/.

24. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, et al. (2012) Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics (Oxford, England) 28: 2678–2679.

25. Noe L, Kucherov G (2005) YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Research 33: W540-W543.

26. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

27. Burrows-Wheeler Aligner website. http://bio-bwa.sourceforge.net/. Version 0.7.

28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

29. SAMtools website. http://samtools.sourceforge.net/. Version 1.1.

30. Synthviewer website. http://polebio.lrsv.ups-tlse.fr/synthviewer/. Accessed 3 March 2015.

31. Leroy P, Guilhot N, Sakai H, Bernard A, Choulet F, et al. (2012) TriAnnot: A Versatile and High Performance Pipeline for the Automated Annotation of Plant Genomes. Frontiers in plant science 3: 5.

32. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis : sequence visualization and annotation. Bioinformatics 16: 944-945.

33. Phillippy AM, Schatz MC, Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. Genome biology 9: R55.

34. Chin CS, Alexander DH, Marks P, Klammer A, Drake J, et al. (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature methods 10: 563-569.

35. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, et al. (2014) Resolving the complexity of the human genome using single-molecule sequencing. Nature 517: 608-611.

36. Shin SC, Ahn DH, Kim SJ, Lee H, Oh TJ, et al. (2013) Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PloS one 8: e68824.

37. Madden T (2002 updates 2003) The BLAST Sequence Analysis Tool. In: Bethesda MD (2002) The NCBI Handbook Chapter 16 (Internet) (2002). The National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/books/NBK21097/.

38. Koren S, Schatz MC, Walenz BP, Martin J, Howard J, et al. (2013) Hybrid error correction and de novo assembly of single- molecule sequencing reads. Nature biotechnology 30: 693-700.

39. Murat F, Van de Peer Y, Salse J (2012) Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. Genome biology and evolution 4: 917-928.

40. Salse J (2012) *In silico* archeogenomics unveils modern plant genome organisation, regulation and evolution. Current opinion in plant biology 15: 122–130.

41. Mago R, Tabe L, Vautrin S, Šimková H, KubalákováM, et al. (2014) Major haplotype divergence including multiple germin-like protein genes, at the wheat Sr2 adult plant stem rust resistance locus. BMC plant biology 14: 379.