Structural variants detection and *de novo* genome assembly of a Maize line

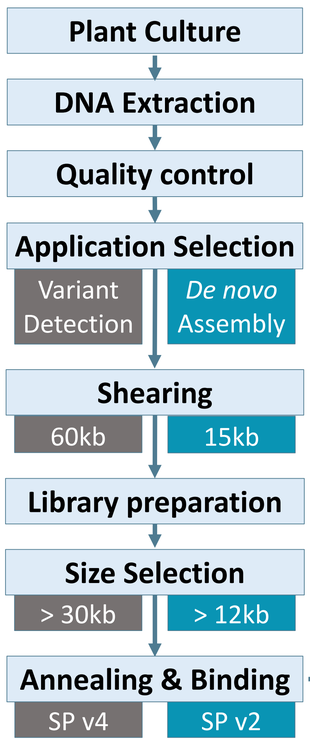
Eché, Camille[1] ; Birbes, Clement[2] ; Iampietro, Carole[1] ; Di Franco, Arnaud[2] ; Dréau, Andreea[2] ; Kuchly, Claire[1] ; Klopp, Christophe[2] ; Faraut, Thomas[5] ; Zytnicki, Matthias[2] ; Denis, Erwan[1] ; Praud, Sébastien[4] ; Joets, Johann[3] ; Vitte, Clémentine[3] ; Charcosset, Alain[3] ; Gaspin, Christine[2] ; Milan, Denis[1 5]; Donnadieu, Cécile[1]

1- INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France ; 2- Plateforme Bio-informatique Genotoul, Mathématiques et Informatique Appliquées de Toulouse, INRAE, Castanet-Tolosan, France. ;3- Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, Gif-sur-Yvette, France. ; 4- Limagrain, Clermont-Ferrand, France ; 5- GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet-Tolosan Cedex, F-31326, France.

Characterizing the genomic diversity of species is critical to understand the molecular origin of phenotypic variations. Whole genome sequence assemblies at the chromosome scale with low amount of missing data are critical resources for answering such questions.

Our aim is to explore combination of technologies to answer those questions, and to find the best one. Using a specific Maize line, we have tested different sequencing applications to build a "high quality genome" and identify a large collection of genome variants. Here we demonstrate that the use of HIFI reads combined with Hi-C and Linked-reads generate a chromosome-scale genome assembly with a better contiguity than the reference genome B73. Moreover we show, that PacBio CLR reads allow a wide detection of structural variants of the genomes.

Workflow for PacBio libraries preparation



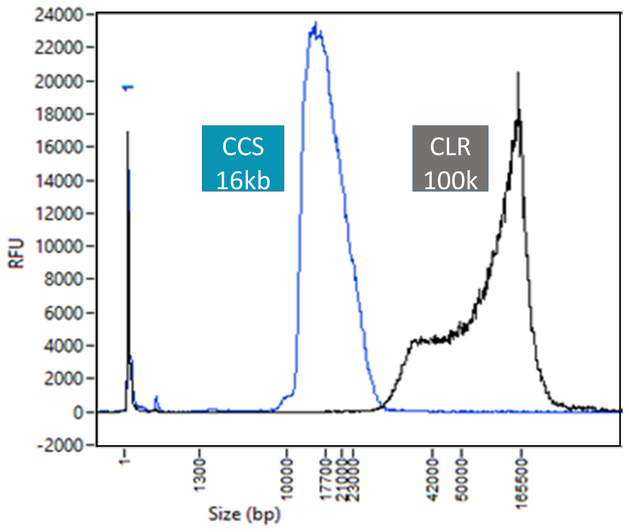
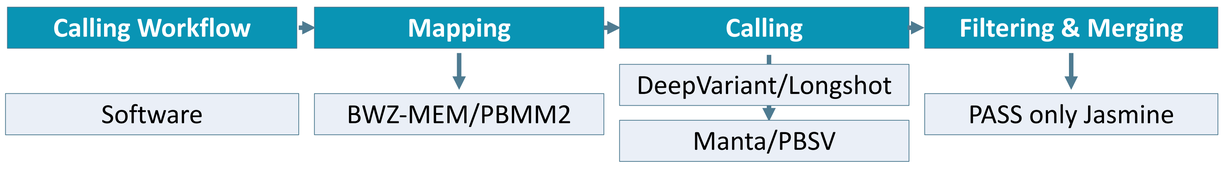


Figure 1 : Fragment length distribution in average of SMRTbell Express libraries of maize sample. (Agilent Femto Pulse)

Lab workflow optimization details

DNA Extraction : E.Z.N.A Plant DNA kit; Quality control : Femto-Qubit-Nanodrop ; Shearing : Megaruptor 3 Library preparation (10µg); Library: SMRTbell Express Template Prep Kit 2.0 + Nuclease Treatment of SMRTbell Library (for HIFI only); Size Selection : BluePippin (5µg).

Variants detection



Summary of Variant Calling workflow.

Variants were called using Illumina, CLR and HiFi reads (60, 53 and 28-fold respectively). Reads were first mapped against the reference genome B73v4 and two softwares were used for calling small variants (SNPs) and large variants (SV). Finally, all VCF (Variant Calling Format) files were filtered for proper variant size (0-50bp for small variants and 50bp+ for SV), only entries with FILTER field equivalent to PASS and useful genotyping information (no 0/0 or ./.). Illumina reads were mapped with bwa-mem (default parameters) and variants called using DeepVariant 1.0.0 (WGS model) and Manta 1.6.0. PacBio CLR reads were mapped with pbmm2 1.3.0 and variants called using Longshot 0.4.1 (default parameters) and pbsv 2.4.0 (default parameters). PacBio HiFi reads were mapped with pbmm2 1.3.0 (--css option) and variants called using DeepVariant 1.0.0 (PACBIO model) and pbsv 2.4.0 (default parameters). SV from the three sequencing technologies were merged using Jasmine.

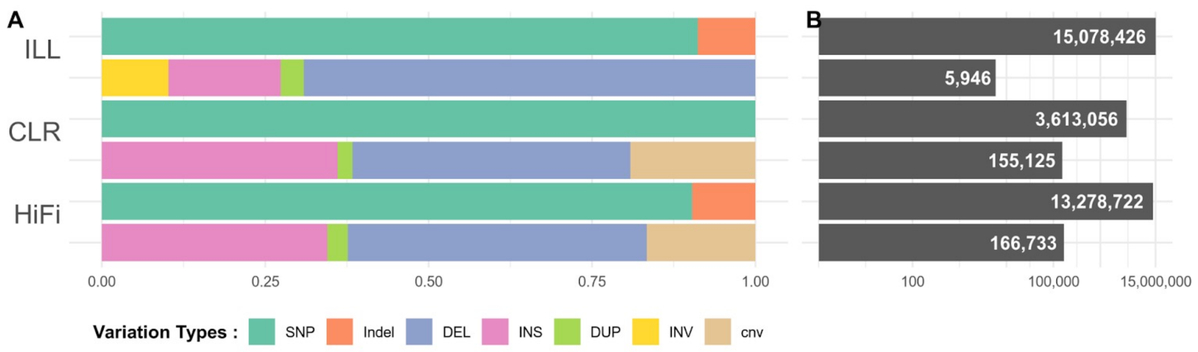
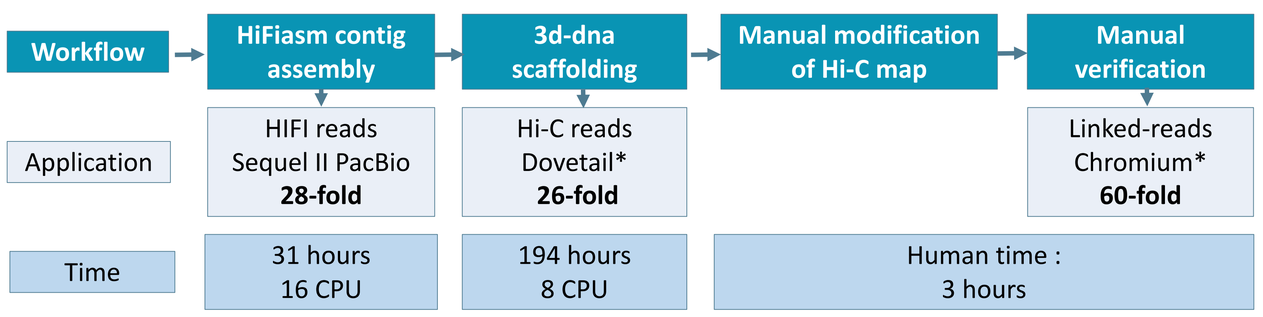


Figure 2 : Variant calling comparison between sequencing technologies on Maize1.

Overall, PacBio technologies discover far more structural variations than Illumina while the latter performs the best on small variants. **A: Proportion of variant types** (small and large variants) for Illumina, PacBio CLR and PacBio HiFi. Around 1,300,000 Indels are detected by DeepVariant with both Illumina and HiFi reads. However, no Indels can be detected at the moment with our method for CLR reads. Regarding large variants, insertions are enriched with PacBio technologies compared to Illumina. NB: Actually, majorities of entry type were breakends, not represented in this picture. They were 7332, 482,888 and 317,856 breakend reported for Illumina, CLR and HiFi analyzes, respectively. This depicts the complexity of rearrangement in Maize lines. **B**: **Amount of small and large variants per technology**. HiFi shows ambivalent results on both small and large variants where Illumina excels only on small ones and CLR only on large ones.

De novo assembly



Optimized genome assembly workflow

PacBio HiFi reads were assembled with Hifiasm v0.9 (default parameters). The assembly was scaffolded using Hi-C reads which were first aligned to the assembly using juicer (default parameters). A pre-scaffolded assembly was generated with 3D-DNA (version 180114) with -r 0 flag (no iterative rounds for mis-join correction). In addition, 10X Chromium reads were aligned on the scaffolds and using tag continuity as 3D-DNA input file was generated to help scaffolding validation. Both 10X chromium and Hi-C links were uploaded in Juicebox to manually review the assembly.

*Hi-C and Chromium (Genome reagent v2) datas were sequenced on an Illumina NovaSeq6000 using a paired-end read length of 2x150 pb.

Final Assembly statistics

	Maize1	B73 reference
Number of scaffolds	2 564	685
Total size of scaffolds	*2 282 968 551*	2 182 075 994
Total scaffold length as percentage of assumed genome size	*95,1 %*	90,9%
% of estimated genome that is useful	94,7%	90,9%
N50 scaffold length	219 798 000	226 353 449
Scaffold %N	0,01	0,17
Number of contigs	2 981	2787
N50 contig length	53 262 804	1 279 966

Contiguity statistics of our maize assembly versus B73 reference sequence:

The use of HiFi reads allowed us to obtain significantly longer contigs compared to the maize reference sequence. With a longer total assembly size (the expected size is ~2.4Gb) and lower gap percentage, our results offer a more complete image of the maize genome. The difference in scaffold number between our assembly and the reference sequence is due to a high number of short scaffolds originated from highly repetitive regions that have not been assembled in the reference. These assembly statistics were done with an old version of the Hifiasm assembler (0.9). The new version (0.15.1) significantly will improve the quality of the assembly: larger contigs and a higher BUSCO score.

Figure 3 : Alignment results of Maize1 assembly to the B73 reference sequence.

The correctness of our assembly is indicated by the high quality alignment to the maize reference genome. The first 10 scaffolds represent an almost complete assembly of the maize chromosomes and we obtained also a significant number of short scaffolds originated from highly repetitive regions that have not been assembled in the reference.

Figure 4 : Busco (v4.0.2) statistics of Maize 1 assembly versus B73 reference sequence.

The number of genes found in our assembly is similar to the results of the maize reference sequence. A potential cause for the missing genes in our assembly is the variability between different maize lines.

Perspectives

All the results, presented here are, have been obtained within the framework of the "Sequencing Occitanie Innovation" project, which aims to acquire expertise on the optimal combination of long fragment sequencing technologies and associated applications to better characterize complex genomes for species with agronomical interest (maize, sheep, cow).

Thanks to the expertise acquired with the assembly of this first Maize genome, we will now apply a similar strategy to gain knowledge on specific Maize lines of interest (Flint and Dent lines) adapted to European regions.

More about the SeqOccIn program : https://get.genotoul.fr/seqoccin/

Three majors topics are treated in this project: genome variability analysis, epigenetic mark analysis and metagenome analysis. It's carried out by Get-PlaGe and Genotoul Bioinfo platforms and financed by Feder funds (Programme Opérationnel FEDER-FSE_Midi-Pyrénées et Garonne 2014-2020). The project benefits from the contributions of INRAe research units GenPhySE, MIAT, GABI, GQE. 25 private partners are involved in the project, among them, the AMAIZING project partners (Biogemma, MaïsAdour, KWS, Caussade Semences, Limagrain, RAGT, Euralis, Syngenta).