



HAL
open science

De novo assembling 19 maize inbred lines of the European germplasm

Carole Iampietro, Camille Ech , Clement Birbes, Andreea Dreau, Erwan Denis, Claire Kuchly, Christophe Klopp, Arnaud Di Franco, Thomas Faraut, Matthias Zytnicki, et al.

► To cite this version:

Carole Iampietro, Camille Ech , Clement Birbes, Andreea Dreau, Erwan Denis, et al.. De novo assembling 19 maize inbred lines of the European germplasm. 63rd Annual Maize Genetics Meeting, Mar 2021, virtuel, United States. hal-03539396

HAL Id: hal-03539396

<https://hal.science/hal-03539396v1>

Submitted on 21 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destin e au d p t et   la diffusion de documents scientifiques de niveau recherche, publi s ou non,  manant des  tablissements d'enseignement et de recherche fran ais ou  trangers, des laboratoires publics ou priv s.

De novo assembling 19 maize inbred lines of the European germplasm

Carole Iampietro (1), Camille Ech  (1), Clement Birbes (2), Andreea Dr au (2), Erwan Denis (1), Claire Kuchly (1), Christophe klopp (2), Arnaud Di-Franco (2), Thomas Faraut (4), Matthias Zytnecki (2), Johann Joets (3), Cl mentine Vitte (3), Alain Charcosset (3), Christine Gaspin (2), Denis milan(1,4), C cile Donnadi u (1)

1- INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France

2- Plateforme Bio-informatique Genotoul, Math matiques et Informatique Appliqu es de Toulouse, INRAE, Castanet-Tolosan, France.

3- Universit  Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, Gif-sur-Yvette, France 4 GenPhySE, Universit  de Toulouse, INRA, INPT, ENVT, Chemin de Borde Rouge, Castanet-Tolosan Cedex, F-31326, France.

Background and objectives

Characterizing the genomic diversity of maize is critical to understand the molecular origin of structural variation, and is a prerequisite to underpin the functional variation underlying phenotypic variation. Whole genome sequence assemblies at the chromosome scale with low amount of missing data are critical resources for answering such questions.

Whole genome sequence assemblies are now available for several maize inbred lines, but this is still missing for a number of key founders used in European maize breeding programs. This concerns the dent genetic pool and the European specific flint genetic group. Therefore, we have established a list of 13 priority lines for *de novo* sequencing as part of the SeqOccln project.

One of the aim of the SeqOccln project (Sequencing Occitanie Innovation, <https://get.genotoul.fr/seqoccln/>) is to acquire expertise on the optimal combination of long fragment sequencing technologies and associated applications to better characterize complex genomes for species of agronomical interest (maize, sheep, cow, pork). In the framework of this project, we are *de novo* assembling whole genome sequences of these 13 maize inbred lines. Here, we present results obtained for one inbred line, highlighting the strategy used to a “high quality assembled genome” with a combination of chosen technologies.

Sequencing data

| | Long reads | Linked reads | Hi-C reads |
|------------------------------|----------------------------------------|------------------------------------------------|-----------------------|
| Preparation kit | SMRTbell Express Template Prep Kit 2.0 | Genome Reagent Kits v2 (10X Genomics-chromium) | Dovetail Hi-C Kit |
| Production plateforme | PacBio Sequel II | Illumina NovaSeq3000 | Illumina NovaSeq3000 |
| Librairie size | 15 kb | >50kb | NA |
| Sequencing Layout | CSS mode | Paired End 2 × 150 bp | Paired End 2 × 150 bp |
| Total number of reads | 4,6 M | 640 M paired-end | 205 M paired-end |
| Total Gb | 67 Gb | 197 Gb | 62 Gb |

Our first goal is to validate the strategy choose to obtained a “high quality assembled genome” with our combination of technologies.

The table above summarizes the sequencing data produce for one of the maize line (MBS847) use to test our strategy.

Genome assembly protocole

1. HiFiasm contig assembly with 28X HiFi reads
2. 3d-dna scaffolding with 26X Hi-C reads
3. Manual modification of Hi-C map
4. Manual verification with 10X chromium map (65X)

PacBio HiFi reads were assembled with Hifiasm v0.9 with default parameters. The assembly was then scaffolded using Hi-C reads which were first aligned to the assembly using juicer with default parameters. A pre-scaffolded assembly was then 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. Genome completeness was qualified using Benchmarking Universal Single-Copy Orthologs (BUSCO) v4.0.2 based on 4,896 BUSCO orthologs derived from the poales lineage.

More about the SeqOccln program

The overall SeqOccln project should enable us to acquire expertise on the optimal combination of long fragment sequencing technologies and associated applications to better characterize complex genomes in agronomical field. Three majors topics are treated in this project: genome variability analysis, epigenetic mark analysis and metagenome analysis. The SeqOccln project is carried out by Get-PlaGe and Genotoul Bioinfo platforms. It is financed by Feder funds (Programme Op rationnel FEDER-FSE_Midi-Pyr n es 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, Ma sAdour, KWS, Caussade Semences, Limagrain, RAGT, Euralis, Syngenta).

Final Assembly statistics

| | MBS847 | B73 |
|-------------------------------------------------------------------|---------------|---------------|
| 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 MBS847 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. As presented in Figure 2, the first 10 scaffolds represent an almost complete assembly of the maize chromosomes and 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.

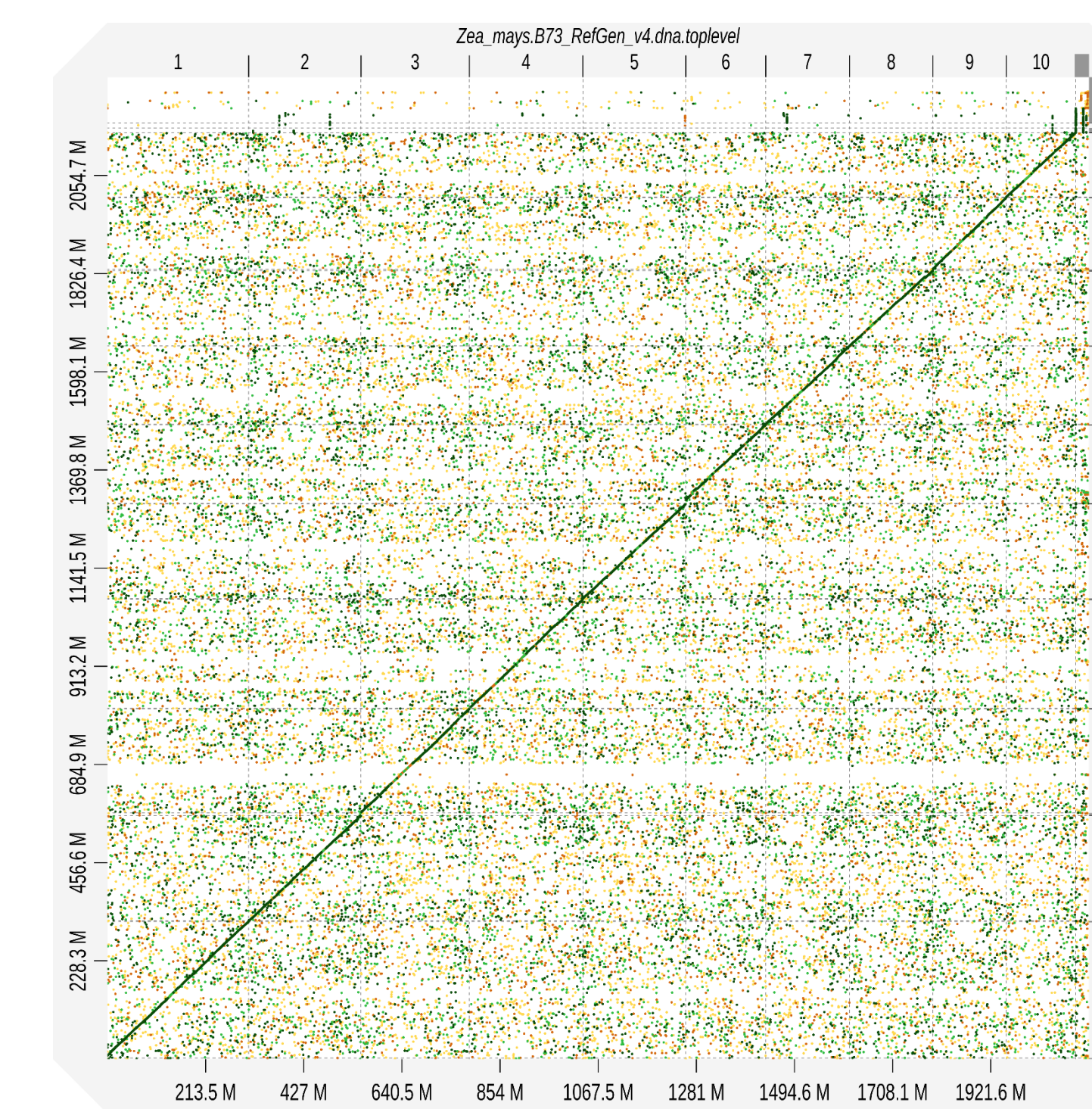


Figure 2: Alignment results of MBS847 assembly to the B73 reference sequence: The correctness of our assembly is indicated by the high quality alignment to the maize reference genome. The first 10 scaffolds represent an almost complete assembly of the maize chromosomes and we obtained also a significant number of short scaffolds originated from highly repetitive regions that have not been assembled in the reference.

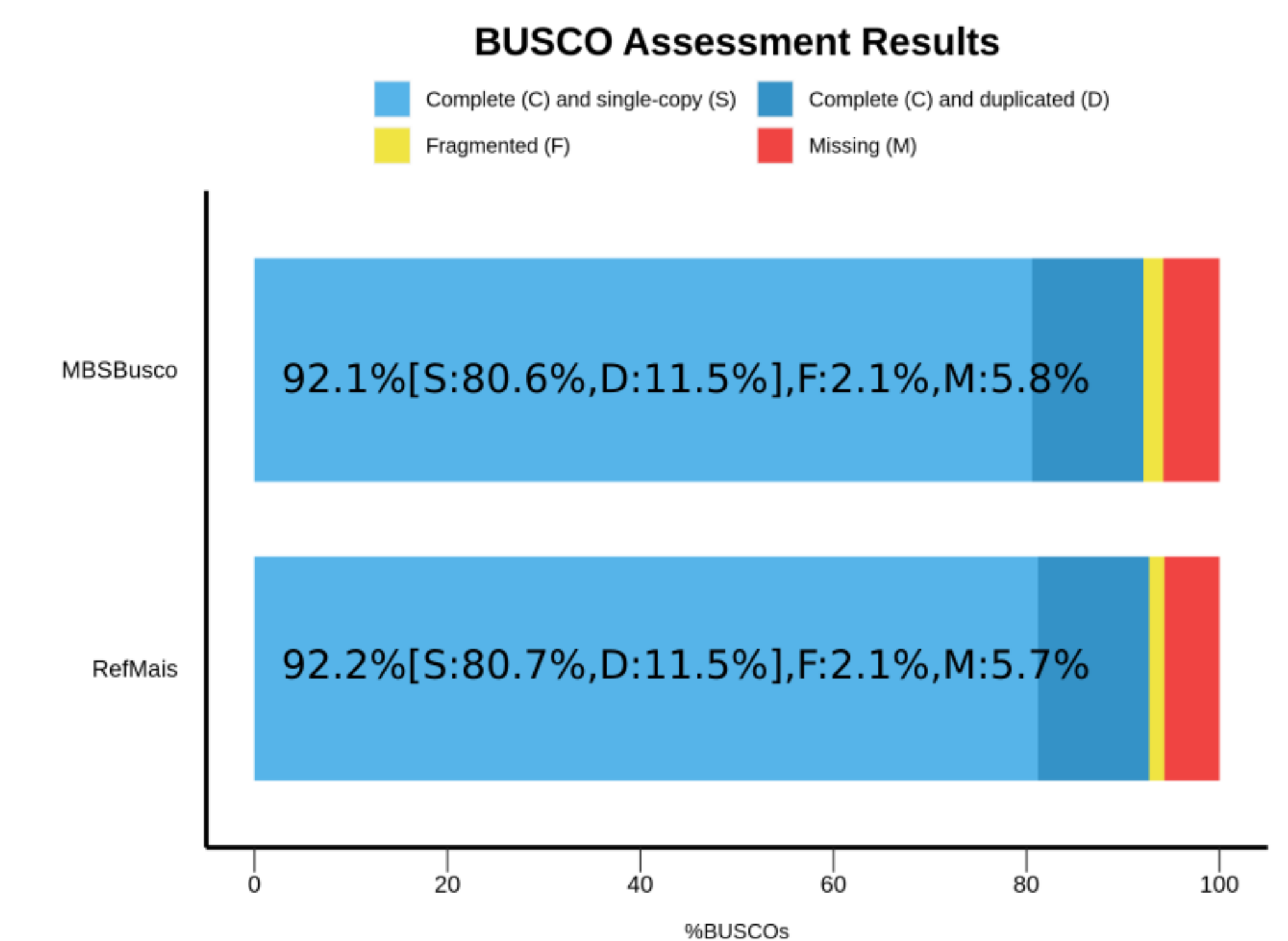


Figure 1: Busco statistics of MBS847 assembly versus B73 reference sequence: The number of genes founded in our assembly is similar to the results of the maize reference sequence. A potential cause for the missing genes in our assembly is the variability between different maize lines.

| Inbred line | CO255 | 3IIH6 | LH82 | B14 B37 | F7130 | F283 | F120 | PHN82 | PHP02 | LH123 | PHR03 | DKMM50 1D | PHG35 | A632 | DKFBLL | DKPB80 | PHG39 | PHW52 |
|---------------|-------|--------|-----------|---------|--------|-------|-------|--------|--------|-----------|-----------|-----------|-----------|------|--------|--------|-------|-------|
| Group | Flint | Iodent | Lancaster | SSS SSS | Flint | Flint | Flint | Iodent | Iodent | Lancaster | Lancaster | Lancaster | Lancaster | SSS | SSS | SSS | SSS | SSS |
| Family | Flint | Iodent | Lanc-Oh43 | B14 B37 | Aranga | F7 | Flint | Iodent | Iodent | Lancaster | Lanc-Iod | Oh43 | Oh43 | B14 | B73 | B73 | B73 | B73 |

Inbred lines and type of sequencing to be applied: We identified by an enquiry process important founder lines of ongoing breeding programs for which genomic data are currently missing to our knowledge. Genetic groups encompass Flint lines specific to North European breeding, and several Dent heterotic groups adapted to different European regions.