scientific data

OPEN

DATA DESCRIPTOR

High-quality chromosome-scale genome assemblies of 29 maize inbred lines of European breeding relevance

Camille Marcuzzo[1,8], Clément Birbes[2,8], Camille Eché[1,8], Arnaud Di Franco[3], Thomas Faraut[3], Erwan Denis[1], Claire Kuchly[1], Caroline Vernette[1], Sébastien Praud[4], Alain Charcosset[5], Christine Gaspin[2], Denis Milan[1,3], Stéphane D. Nicolas[5], Cécile Donnadieu[1]✉, Clémentine Vitte[6]✉, Christophe Klopp[7]✉ & Carole Iampietro[1]✉

Although several maize genome assemblies are publicly available, those of lines important to European breeding programs are underrepresented. Using PacBio long-read sequencing, we assembled high-quality chromosome-level genomes of 29 key lines of European breeding relevance, encompassing Northern flint and European flint lines used for adaptation to Northern European climate, lines derived from European landraces of tropical origin, and American temperate dent lines adapted to European regions. Genome assembly sizes range from 2.17 to 2.35 gigabases, with scaffold N50s ranging from 219 to 254 megabases. Completeness assessment revealed BUSCO scores ranging from 97.7 to 98.5 and merqury completeness scores ranging from 96.62 to 98.30. Calling structural variants and SNPs relative to the B73 reference sequence revealed the expected separation of inbred groups. Flint lines contribute the highest number of novel variants, thus emphasizing the importance of sequencing flint material to complete the maize pangenome. These high-quality genome assemblies therefore provide new opportunities to understand the dynamics of maize structural variation, and to identify the functional variations underlying maize phenotypic diversity.

Background & Summary

Maize (*Zea mays* ssp. mays) is known for its large genetic diversity, which allowed the species to adapt to a multitude of environments, including tropical and temperate climates. Maize is now grown throughout the world and is the cereal with the highest production worldwide[1]. Its extensive genetic and phenotypic variation has also been the foundation of modern hybrid breeding. In the U.S., complementary heterotic groups within the dent germplasm - Stiff Stalk Synthetic and non-Stiff Stalk Synthetic, including Lancaster and Iodent lines - have been developed to generate highly productive hybrids, while in Europe, heterotic effects between dent and flint lines have been exploited to develop productive hybrids adapted to cooler climate. In addition to its role as a major food crop, maize is also a model organism in biology, particularly for genome dynamics, due to its large amount of intra-specific structural variation[2] and its massive transposable elements content[3,4]. The discovery that non-coding polymorphisms contribute significantly to a wide range of phenotypic traits[5] also led to the establishment of maize as a model for the study of gene expression regulation[6–8], including the integration of *cis*-regulatory elements into gene regulatory networks[9]. Characterizing

[1]INRAE, GeT-PlaGe, Genotoul, 31326, Castanet-Tolosan, France. [2]Université Fédérale de Toulouse, INRAE, MIAT, BioinfOmics, 31326, Castanet-Tolosan, France. [3]Université de Toulouse, INRAE, GenPhySE, 31326, Castanet-Tolosan, France. [4]Groupe Limagrain, Centre de Recherche, Route d'Ennezat, Chappes, France. [5]Université Paris-Saclay, INRAE, AgroParisTech, GQE, Le Moulon, 91190, Gif-sur-Yvette, France. [6]Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE, Le Moulon, EMR GEvAD, 91190, Gif-sur-Yvette, France. [7]Université Fédérale de Toulouse, INRAE, MIAT, Sigenae, BioInfo Genotoul, BioinfOmics, 31326, Castanet-Tolosan, France. [8]These authors contributed equally: Camille Marcuzzo, Clément Birbes, Camille Eché. ✉e-mail: cecile.donnadieu@inrae.fr; clementine.vitte@inrae.fr; christophe.klopp@inrae.fr; carole.iampietro@inrae.fr

Inbred line	Group	Subgroup	Pedigree	Obtentor or Developer	Accession Code
A632	Dent	Stiff Stalk Synthetic[a,b]	B14	University of Minnesota	A632_usda
B14	Dent	Stiff Stalk Synthetic[a,b]	B14	Iowa State University	B14_usda
B37	Dent	Stiff Stalk Synthetic[a,b]	B37	Iowa State University	B37_usa
CM174	Dent	Stiff Stalk Synthetic[a,b]	B14	Manitoba Agriculture Canada Research	CM174_usa
CO255	Flint	European Flint[a]	Mixed Flint	Ontario Agriculture Canada Research	CO255_usda
DK3IIH6	Dent	Iodent[b]	Mixed Dent	Dekalb	PI 564754
DKFBLL	Dent	Stiff Stalk Synthetic[b]	B73	Dekalb	PI 546481
DKMM501D	Dent	Lancaster[b]	Oh43	Dekalb	PI 564752
DKPB80	Dent	Stiff Stalk Synthetic[b]	B73	Dekalb	PI 60144
EA1197	Tropical	Tropical Spanish[a]	Mollar Almeria (Spain)	CSIC	EM1197_inra
F120	Flint	European Flint[b]	Mixed Flint	INRAE	F120_inra
F2	Flint	European Flint[a]	Lacaune (France)	INRAE	FV2_inra / FRA2711759
F252	Dent	Lancaster/Iodent[a]	CO125	INRAE	FV252_MLN
F283	Flint	European Flint[a]	Mixed Flint	INRAE	FV283_inra
F331	Tropical	Tropical Highland[a]	POB 86 (CIMMYT)	INRAE	FV331_inra
F353	Dent	Iodent[a]	Mixed Dent	INRAE	FV353_inra
F4	Flint	Northern Flint[a]	Etoile de Normandie (France)	INRAE	FV4_inra
F7130	Flint	European Flint[b]	Aranga (Spain)	INRAE	F7130_inra
GF111	Flint	Northern Flint[b]	Gaspe (Canada)	University of Bologna	GF111_unibo
LH123Ht	Dent	Lancaster[b]	Mixed Dent	Holden's	LH123Ht_usda
LH82	Dent	Lancaster[b]	Oh43	Holden's	LH82_usda
Lo3	Flint	Italian Flint[a,b]	Nostrano dell'Isola (Italy)	CREA	Lo3_inra
MBS847	Dent	Iodent[a]	Mixed Dent	Mike Brayton Seeds	MBS847_MLN
PHG35	Dent	Lancaster[b]	Oh43	Pioneer Hi	PI 601008
PHG39	Dent	Stiff Stalk Synthetic[b]	B73	Pioneer Hi	PI 600981
PHN82	Dent	Lancaster[b]	Oh43	Pioneer Hi	PI 601783
PHP02	Dent	Iodent[b]	Mixed Dent	Pioneer Hi	PI 601570
PHR03	Dent	Lancaster[b]	Lancaster/Iodent	Pioneer Hi	PI 548803
PHW52	Dent	Stiff Stalk Synthetic[b]	B73	Pioneer Hi	PI 601575

Table 1. List of inbred lines with genotype information. [a]Based on structure analysis[37,38], [b]based on pedigree.

the genomic diversity of maize is essential for understanding the contribution of structural variants to this diversity, and is a prerequisite to underpinning the functional variation underlying phenotypic variation. Near complete high quality chromosome-scale genome assemblies are critical resources to address these questions.

Despite this wide genetic diversity, for decades, most knowledge about the genomic structure and function of maize has been obtained from a single genotype, B73, an American temperate dent line, therefore representing only a subset of the genetic variability and biology of the species, with a bias towards genetics of the Stiff Stalk Synthetic germplasm. In the past years, efforts have been made to *de novo* assemble full genome sequences of several other maize lines[10–14], including flint material of interest for Europe[15,16]. While providing first insights into maize structural variation, these studies nevertheless remained limited in characterizing the maize pangenome, as they were generated by different laboratories, using different assembly and annotation strategies. This issue has been overcome by the production of a pangenome analysis of a set of 26 founder inbred lines representing a large fraction of maize diversity, including lines from temperate, subtropical and tropical origin, as well as lines from sweet corn and popcorn germplasm[17]. The production of high-quality assemblies with high contiguity over repetitive regions revealed large amounts of structural variants. Although most of the variants discovered were in high linkage disequilibrium with SNPs, over 6% of the genomic regions found associated with phenotype were solely detected with structural variants and not with SNPs, indicating their biological relevance and their agronomic value. The cumulative number of pan genes found from this set of 26 lines did not reach a plateau, highlighting the need to explore more extensively genome sequences of the maize germplasm to discover the entire set of maize genes. In particular, the absence of flint material in this dataset hampered a global analysis of the maize germplasm and likely caused an under-appreciation of maize genetic variation. This also limits the use of this pangenome for breeding programs using flint material.

In this study, we expand the current collection of maize whole-genome assemblies by generating high-quality PacBio HiFi-based assemblies for 29 key inbred lines of major relevance to European breeding programs. These include Northern and European flint lines used for adaptation to Northern European climates, inbred lines derived from European landraces of tropical origin, and American dent lines that complete the diversity of the 26 American founder lines (see Table 1).

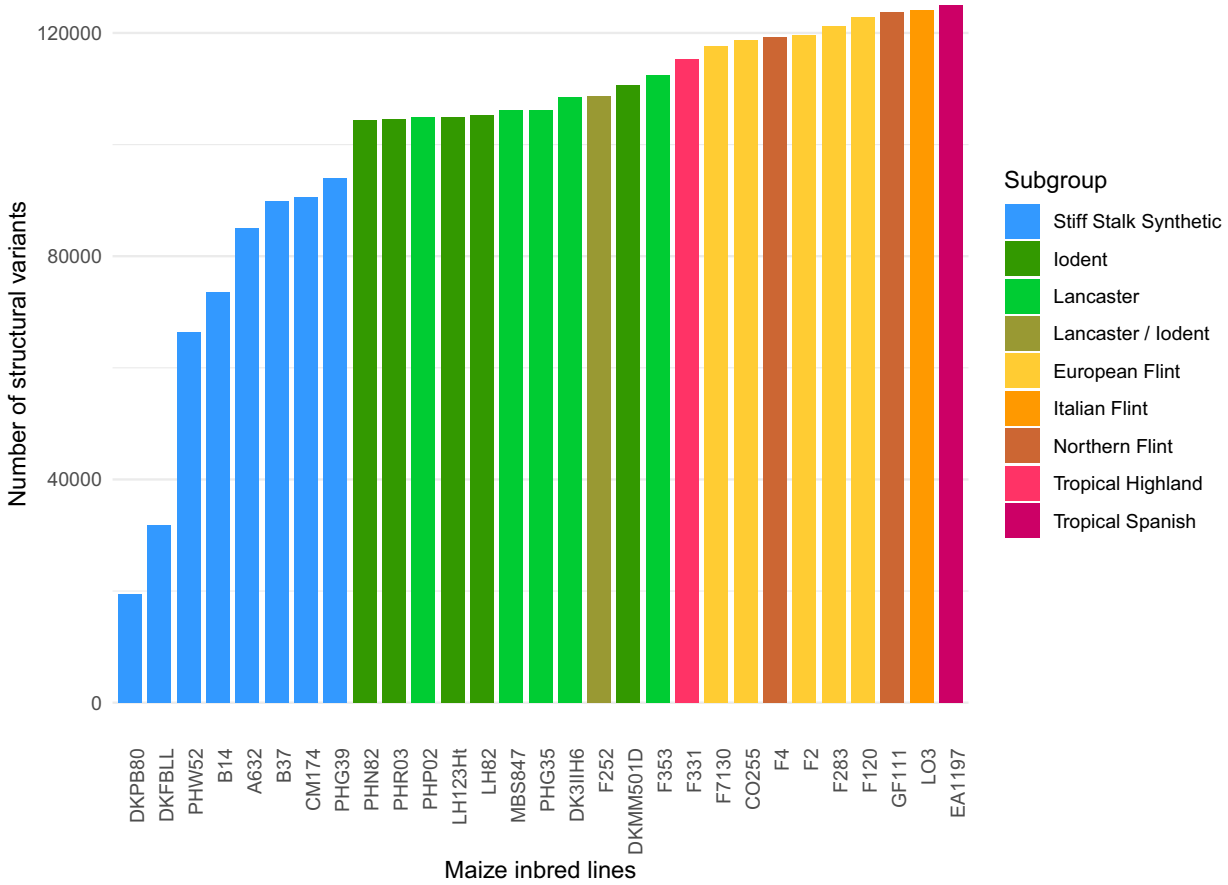

Fig. 1 Quantities of structural variants detected for each inbred line as compared to B73.

Methods

Sample collection and genomic DNA extraction. Plants were grown in standard conditions (growth chamber) up to emergence, then moved to obscurity for 2 to 5 days. Young etiolated leaf samples were flash frozen in liquid nitrogen upon collection. Leaf DNA extractions were carried out using three different protocols: EZNA SQ plant kit (Omega, D3095), Mayjonade *et al.*[18] and Nucleobond HMW DNA Kit (Macherey-Nagel, Ref: 740160.20). The protocol used was tracked for each sample and can be found in the DNA samples metadata. DNA was quantified using the Qubit fluorimetry system, with the High Sensitivity kit (Thermo Fisher, Q32854). Fragment size distributions were assessed using the Agilent Fragment Analyzer. Purity measurements were performed using a Thermo Fisher Nanodrop system, thus ensuring absence of contaminants.

Genome sequencing. *Generation of HIFI reads using PacBio Sequel II - CCS.* Library preparation was performed according to the manufacturer's instructions "Procedure & Checklist Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit 2.0 or 3. 0". 5 to 10 μg of DNA was purified and sheared to reach 20kb size using the Megaruptor3 system (Diagenode). Size selection with a 10–15 kb cutoff was performed on the BluePippin Size Selection system or the Pippin HT system (Sage Science). Libraries were sequenced on 2 to 4 SMRTcells on a Sequel II instrument with a 2 hours pre-extension and a 30 hours movie, aiming to reach a 25X HIFI reads genome coverage.

Hi-C library preparation and sequencing. Hi-C libraries were prepared from the F2, F4, F252 and MBS847 samples, using isolated nuclei as starting material. The nuclei were obtained from 1g of young leaves, following the method described in Workman *et al.*[19]. All nuclei obtained where then fixed in 1.5% formaldehyde and used to perform Hi-C using the Dovetail Hi-C Kit according to the manufacturer's protocol (Ref: DG-HiC). Briefly, fixed *in situ* chromatin was digested with *Dpn*II, DNA ends were labeled with Biotin and proximity ligation was performed. After reverse-crosslinking, 1 µg of purified DNA was then sheared to reach a mean fragment size of ~550 bp (Covaris) and used to build a sequencing library using Illumina adapters. Biotin-containing fragments were isolated using M280 streptavidin Dynabeads (Invitrogen) before PCR enrichment of the library (10 PCR cycles). The libraries were sequenced on an Illumina NovaSeq6000 platform to generate 2 × 150 bp pair-end reads, producing a minimum of 48 Gb of Hi-C read data per library.

Inbred line	HiFi reads accession	Number of reads	Number of nucleotides	Average read length
A632	ERR14085326, ERR14085330	3,424,702	57,543,104,624	16802.4
B14	ERR14085312, ERR14085317, ERR14085318	4,192,820	63,372,322,878	15114.5
B37	ERR14085313, ERR14085314, ERR14085315, ERR14085321	4,004,846	50,688,610,963	12656.8
CM174	ERR14085367, ERR14085369	4,670,902	76,685,308,638	16417.7
CO255	ERR14085316, ERR14085319, ERR14085322,ERR14085324	4,929,741	74,399,642,716	15092.0
DK3IIH6	ERR14085309, ERR14085310, ERR14085311	5,020,853	77,950,301,448	15525.3
DKFBLL	ERR14085334	2,376,358	28,630,681,311	12048.1
DKMM501D	ERR14085338, ERR14085339	4,757,310	75,620,011,379	15895.5
DKPB80	ERR14085337, ERR14085340	3,666,901	61,466,645,359	16762.6
EA1197	ERR14085361, ERR14085364	3,617,240	76,561,159,902	21165.6
F120	ERR14085327, ERR14085331	3,550,992	59,310,705,466	16702.6
F2	ERR14085295, ERR14085303, ERR14085304	5,976,900	75,601,454,246	12648.9
F252	ERR14085296, ERR14085299, ERR14085300	5,836,950	80,114,716,897	13725.4
F283	ERR14085328, ERR14085332, ERR14085333	3,231,162	55,225,844,986	17091.6
F331	ERR14085358, ERR14085359	3,159,062	65,331,493,229	20680.7
F353	ERR14085360, ERR14085363	3,630,976	80,056,704,698	22048.3
F4	ERR14085298, ERR14085301, ERR14085302	6,886,772	85,753,750,380	12452.0
F7130	ERR14085325, ERR14085329	3,645,386	58,321,890,027	15998.8
GF111	ERR14085366, ERR14085368	4,488,140	82,146,256,235	18303.0
LH123Ht	ERR14085336, ERR14085341	4,907,664	72,219,007,386	14715.6
LH82	ERR14085320, ERR14085323	3,583,487	58,631,023,192	16361.4
Lo3	ERR14085362, ERR14085365	3,784,213	79,426,343,442	20988.9
MBS847	ERR14085293, ERR14085294, ERR14085297	4,477,179	67,020,456,487	14969.3
PHG35	ERR14085342, ERR14085343	4,095,547	64,042,826,836	15637.2
PHG39	ERR14085347, ERR14085348, ERR14085349	3,026,793	46,322,128,190	15304.0
PHN82	ERR14085350, ERR14085351	3,484,281	57,295,832,824	16444.1
PHP02	ERR14085352, ERR14085354, ERR14085356	4,401,998	65,723,820,887	14930.5
PHR03	ERR14085344, ERR14085345, ERR14085346	2,131,704	30,800,884,018	14448.9
PHW52	ERR14085353, ERR14085355, ERR14085357	4,648,997	68,432,090,419	14719.8

Table 2. Read sets accessions and statistics.

Genome sequence assembly and validation. Genome sequence assemblies were performed in two consecutive steps, first building contigs from HiFi reads, then organizing these contigs into chromosomes. For a first set of 4 lines, contigs were scaffolded using Hi-C data. These lines were chosen to represent material with various degree of relatedness to B73: two non stiff stalk lines belonging to two different subgroups (F252 and MBS847), and two flint lines representing European flints (F2) and Northern flints (F4). We observed no major rearrangements as compared to B73 for any of the assembled genome sequences (see Supplementary Fig. 1 for a genome comparison illustration using D-GENIES[20]), and all these were included within contigs. This indicates that our contig length was large enough to ensure good scaffolding using B73 as a reference. We therefore generated reference-guided assemblies for all other inbred lines using B73v5 sequence as reference.

Contig assembly. HiFi reads were assembled in contigs with hifiasm[21] version 0.16.1 using default parameters. Contig assembly metrics were generated using the assemblathon_stats.pl script found at https://github.com/KorfLab/Assemblathon.

Contig scaffolding. For F2, F4, F252 and MBS847 lines, Illumina Hi-C reads were aligned onto the contigs with Juicer[22], and contigs were scaffolded with 3D-DNA[23]. Resulting contact maps were manually corrected with Juicebox[24]. For all three software packages, default parameters were used. Read quantity, read coverage and Hi-C link metrics are presented in Table 6. For all other maize lines, contig sets were scaffolded with ragtag[25] version 2.0.1 using default parameters, using the Zm-B73-REFERENCE-NAN-5.0.fa sequence as reference, downloaded from the NCBI website https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_902167145.1/. For each maize line, contigs were organized into 10 pseudo-chromosomes, with unplaced contigs corresponding to only 0.9 to 7.2% of the assembly total length.

Inbred line	Assembly accession	Number of contigs	Total size of contigs	N50 contig length	L50 contig count
A632	GCA_964658895	645	2,258,924,671	63,658,720	10
B14	GCA_964657075	706	2,275,428,870	85,618,270	10
B37	GCA_964657055	636	2,231,219,587	62,251,460	10
CM174	GCA_964657175	1631	2,228,253,082	73,401,490	11
CO255	GCA_964656985	2233	2,282,614,402	136,188,670	7
DK3IIH6	GCA_964657035	1586	2,243,029,690	50,387,573	17
DKFBLL	GCA_964657165	725	2,238,763,432	103,816,895	9
DKMM501D	GCA_964657045	766	2,289,768,348	114,918,082	8
DKPB80	GCA_964657015	602	2,247,613,153	166,046,607	6
EA1197	GCA_964657185	1495	2,338,259,689	63,357,351	10
F120	GCA_964657095	546	2,218,004,528	88,675,532	8
F2	GCA_964656995	1877	2,179,294,636	51,283,957	14
F252	GCA_964656955	3084	2,252,981,446	51,100,982	15
F283	GCA_964657145	434	2,168,383,650	121,137,930	7
F331	GCA_964657155	1806	2,348,188,253	66,810,458	13
F353	GCA_965119405	1268	2,274,377,264	120,452,905	7
F4	GCA_964657005	2651	2,223,982,854	56,851,000	13
F7130	GCA_964657025	769	2,196,217,465	97,749,953	8
GF111	GCA_964657105	1208	2,222,670,629	74,743,008	10
LH123Ht	GCA_964656965	1181	2,306,050,812	135,232,566	7
LH82	GCA_964657065	260	2,219,224,844	13,5124,035	7
Lo3	GCA_964657125	1170	2,292,843,525	67,403,620	11
MBS847	GCA_964656975	2981	2,282,760,051	53,262,804	15
PHG35	GCA_964657135	640	2,274,023,514	101,361,818	8
PHG39	GCA_964657195	940	2,295,771,951	48,408,447	14
PHN82	GCA_964657115	593	2,260,355,512	121,963,513	7
PHP02	GCA_964658885	562	2,224,964,976	116,332,906	8
PHR03	GCA_964657085	1198	2,270,845,736	11,758,156	61
PHW52	GCA_964658625	799	2,308,517,212	93,206,125	9

Table 3. Genome assembly: contig metrics.

Scaffold validation. Scaffold metrics were produced using the assemblathon_stats.pl script[26] and the BUSCO (Benchmarking Universal Single-Copy Orthologs)[27] metrics with version 5.1.2 using the `poales_odb10` lineage. Kmer completeness and sequence quality value of the scaffolds were assessed using Merqury[28] version 1.3 with default parameters.

SNPs and structural variants detection. SNPs and structural variants were detected from the raw HiFi reads, aligning the fastq reads from each maize line to the maize reference assembly B73_RefGen_v4 using pbmm2 (https://github.com/PacificBiosciences/pbmm2) with the CSS preset flag. SNPs were detected using DeepVariant (1.3.0) using default parameters (see snp_detection rules in https://github.com/SeqOccin-SV/SeqOccinVariants). Structural variants were detected using the Sniffles[29] (https://github.com/fritzsedlazeck/Sniffles) in a two round process. Sniffles was first used to detect variant on an individual basis with the following parameters (–minsupport 12 –minsvlen 100 –max-splits-base 2 –max-splits-kb 0 –min-alignment-length 5000 –minsvlen 20) with default values for the other parameters. The resulting vcf files were filtered to keep only variant with PASS filter and merged using the jasmine software[30]. BND (breakend) and TRA (translocation) variants were filtered out and the merged SVs were provided as input (–genotype-vcf) to Sniffles along with the BAM files on each individual line, leading to a set of SV genotyped on all the individuals (see Fig. 1).

Data Records

Reads and assembled genome sequences were deposited in European National Archive under bioproject PRJEB67812[31], (see Tables 2–6 for details). SNPs and structural variants data were deposited in the European Variant Archive (Study ID: PRJEB106599)[32] and in the 'Recherche Data Gouv' repository: https://doi.org/10.57745/7AUTOL[33].

Inbred line	Assembly accession	Number of scaffolds	Total size of scaffolds	N50 scaffold length	L50 scaffold count
A632	GCA_964658895	588	2,258,930,371	229,849,476	5
B14	GCA_964657075	664	2,275,433,070	228,416,585	5
B37	GCA_964657055	578	2,231,225,387	247,052,023	5
CM174	GCA_964657175	1556	2,228,260,582	228,036,651	5
CO255	GCA_964656985	2198	2,282,617,902	225,910,932	5
DK3IIH6	GCA_964657035	1369	2,243,051,390	219,774,651	5
DKFBLL	GCA_964657165	670	2,238,768,932	238,920,698	5
DKMM501D	GCA_964657045	673	2,289,777,648	241,113,593	5
DKPB80	GCA_964657015	491	2,247,624,253	229,108,405	5
EA1197	GCA_964657185	1353	2,338,273,889	228,304,669	5
F120	GCA_964657095	493	2,218,009,828	237,609,080	5
F2	GCA_964656995	1778	2,179,344,136	222,818,903	5
F252	GCA_964656955	2739	2,253,153,946	222,268,000	5
F283	GCA_964657145	384	2,168,388,650	233,703,039	5
F331	GCA_964657155	1668	2,348,202,053	231,213,771	5
F353	GCA_965119405	998	2,274,404,264	253,768,414	4
F4	GCA_964657005	2535	2,224,040,854	221,853,500	5
F7130	GCA_964657025	692	2,196,225,165	234,581,251	5
GF111	GCA_964657105	1079	2,222,683,529	224,781,681	5
LH123Ht	GCA_964656965	866	2,306,082,312	228,166,395	5
LH52	GCA_964657065	195	2,219,231,344	228,046,157	5
Lo3	GCA_964657125	1118	2,292,848,725	229,037,764	5
MBS847	GCA_964656975	2564	2,282,968,551	219,798,000	5
PHG35	GCA_964657135	528	2,274,034,714	246,415,018	5
PHG39	GCA_964657195	848	2,295,781,151	228,837,152	5
PHN82	GCA_964657115	537	2,260,361,112	241,736,781	5
PHP02	GCA_964658885	513	2,224,969,876	219,473,517	5
PHR03	GCA_964657085	735	2,270,892,036	219,736,915	5
PHW52	GCA_964658625	738	2,308,523,312	229,847,906	5

Table 4. Genome assembly: scaffold metrics.

Technical Validation

We produced about 2.1 to 6.9 million reads per maize line, with an average read length ranging from 12 kb to 22 kb (Table 2). These high quality HiFi reads were first used to assemble the genomes into contigs, with contig number per maize line ranging from 260 to 3084 (average 1221.1, see Table 3) and N50 contig lengths ranging from 11.8 Mb to 166.0 Mb (average 87.1 Mb, see Table 3). For each maize line, chromosome-scale scaffolds were obtained, with cumulative size of assembled chromosomes ranging from 2.18 Gb to 2.35 Gb (Table 4), in line with the genome sizes expected for maize. As anticipated, tropical lines had larger genome sizes (2.32 Gb) than temperate lines (2.25 Gb). Scaffold N50s range from 219.5 Mb to 253.8 Mb, with L50 from 4 to 5. (Table 4). To ensure the quality, integrity and accuracy of the assembled chromosome sequences generated, we carried our several validation approaches.

Completeness of genome assemblies was evaluated using BUSCO version 5.1.2 with the poales_odb10 containing 4,896 proteins, as well as with Merqury version 1.3. Metrics per genome assembly are presented in Table 5. For all assemblies, >97% of the BUSCO proteins were complete. Merqury results showed genome assemblies quality values >60 and completeness >96.62%.

To further validate the quality of the genome assemblies generated and the genotypes of the DNA sequenced, we investigated the polymorphisms (SNPs, indels and structural variants >50bp) of each line relative to reference line B73. As expected, the number of variants reflected the genetic distances of maize lines from B73 (Fig. 1). Stiff Stalk Synthetic lines showed the lowest amount of variants (7,290,142 SNPs, 829,336 indels and 68,850 SVs, Supplementary Table 2), with the lowest amounts found for lines of the B73 subgroup (Fig. 1 and Table 7). In contrast, flint lines showed the highest number of variants (14,901,375 SNPs, 1,490,896 indels and 119,558 SVs) (Supplementary Table 1). Lancaster and Iodent lines had intermediate values, with Lancaster having slightly more variants (12,365,784 SNPs, 1,282,139 indels and 107,607 SVs) than Iodents lines (11,995,606 SNPs, 1,257,735 indels and 105,935 SVs) (Supplementary Table 2). Lines of tropical origin showed slightly less variants than flint lines. Finally, a PCA based on the SNPs recapitulated the genetic groups and relationships among the lines (Fig. 2). Altogether, these results indicate the high quality of the sequences generated and the

Inbred line	Assembly accession	Complete	Single	Double or more	Fragmented	Missing	Quality value	Completeness (%)
A632	GCA_964658895	98.2	83.1	15.1	0.2	1.6	67.69	97.33
B14	GCA_964657075	98.1	83.1	15.0	0.2	1.7	65.66	97.23
B37	GCA_964657055	98.3	83.1	15.2	0.2	1.5	66.12	98.30
CM174	GCA_964657175	98.1	82.7	15.4	0.2	1.7	65.27	97.61
CO255	GCA_964656985	98.3	82.9	15.4	0.2	1.5	58.85	97.70
DK3IIH6	GCA_964657035	98.2	82.7	15.5	0.2	1.6	62.30	96.90
DKFBLL	GCA_964657165	98.0	82.7	15.3	0.3	1.7	67.54	98.12
DKMM501D	GCA_964657045	98.3	82.8	15.5	0.2	1.5	66.75	97.69
DKPB80	GCA_964657015	98.1	82.3	15.8	0.2	1.7	67.53	97.81
EA1197	GCA_964657185	98.3	83.2	15.1	0.2	1.5	64.27	97.08
F120	GCA_964657095	98.2	83.2	15.0	0.2	1.6	67.73	97.52
F2	GCA_964656995	98.3	83.1	15.2	0.2	1.5	63.04	97.79
F252	GCA_964656955	98.2	82.6	15.6	0.3	1.5	61.61	97.63
F283	GCA_964657145	98.1	83.0	15.1	0.2	1.7	68.29	97.47
F331	GCA_964657155	98.1	82.5	15.6	0.3	1.6	63.84	97.30
F353	GCA_965119405	98.2	82.7	15.5	0.2	1.6	64.53	96.78
F4	GCA_964657005	98.0	82.8	15.2	0.2	1.8	61.32	97.97
F7130	GCA_964657025	98.1	82.7	15.4	0.3	1.6	67.53	97.41
GF111	GCA_964657105	98.0	82.6	15.4	0.3	1.7	65.22	97.01
LH123Ht	GCA_964656965	98.2	82.6	15.6	0.2	1.6	65.48	97.91
LH82	GCA_964657065	97.8	82.8	15.0	0.2	2.0	68.20	97.75
Lo3	GCA_964657125	98.0	82.6	15.4	0.3	1.7	65.20	97.20
MBS847	GCA_964656975	98.3	83.0	15.3	0.2	1.5	61.74	97.39
PHG35	GCA_964657135	98.2	83.0	15.2	0.2	1.6	67.93	97.58
PHG39	GCA_964657195	97.7	82.5	15.2	0.3	2.0	65.81	96.62
PHN82	GCA_964657115	98.3	83.0	15.3	0.2	1.5	66.80	97.15
PHP02	GCA_964658885	98.5	83.1	15.4	0.2	1.3	68.15	97.87
PHR03	GCA_964657085	98.1	82.4	15.7	0.3	1.6	63.49	97.74
PHW52	GCA_964658625	98.2	82.8	15.4	0.3	1.5	67.50	98.18

Table 5. BUSCO and merqury scores.

Inbred line	Hi-C reads accessions	Number of read pairs	Cov. (x)	Percent aligned read pairs	Number of V.i.	Percent V.i. inter-contig	Percent V.i. intra-contig <20 kb	Percent V.i. intra-contig >20 kb
F2	ERR14035548	189,262,356.0	26	80.40%	69,830,909.0	33%	39%	28%
F4	ERR14035549	260,294,856.0	35	80.27%	88,151,672.0	29%	40%	30%
F252	ERR14035543, ERR14035550	140,006,067.0	19	75.41%	43,536,237.0	35%	36%	29%
MBS847	ERR14035536, ERR14035537, ERR14035542	206,127,921.0	27	80.19%	66,163,943.0	22%	51%	27%

Table 6. Hi-C metrics. Cov.: coverage, V.i.: Valid interaction.

reliability of the seedlots sequenced. They also highlight the relevance of our dataset to improve knowledge on maize structural diversity, and the importance of including flint lines in sequencing programs to leverage the maize pangenome.

Data availability

All raw sequencing data, assembled genomes, and variant data (VCF files) have been deposited in publicly accessible repositories. The PacBio HiFi and Hi-C sequencing reads, as well as the genomes assembled from these data, have been uploaded to the European Nucleotide Archive (ENA) at www.ebi.ac.uk/ena as part of the SeqOccIn project, PRJEB6007516[34], and are accessible under project PRJEB67812[31]. Structural Variants and SNPs are available to European Variation Archive (EVA) and accessible under the accession PRJEB106599[32]. Variant data are linked to the nucleotide data through the sharing of a single BioSample ID. Variant data are also available at data.gouv.fr repository (https://doi.org/10.57745/7AUTOL)[33].

Inbred line	Number of SNPs	Number of indels	Number of SVs
A632	8,831,074	993,629	85,074
B14	7,370,699	889,197	73,610
B37	11,061,066	1,151,188	89,875
CM174	9,976,071	1,088,490	90,590
CO255	15,072,571	1,505,472	118,730
DK3IIH6	12,979,891	1,322,942	108,468
DKFBLL	3,100,121	434,673	31,717
DKMM501D	13,317,829	1,374,340	110,652
DKPB80	1,601,782	311,382	19,480
EA1197	14,376,326	1,435,168	125,021
F120	15,352,014	1,525,832	122,866
F2	15,823,786	1,570,288	119,594
F252	13,405,714	1,391,609	108,710
F283	14,712,732	1,457,867	121,289
F331	13,264,044	1,333,745	115,224
F353	12,055,207	1,261,391	112,471
F4	16,050,744	1,598,782	119,325
F7130	14,307,725	1,448,508	117,640
GF111	15,148,796	1,492,753	123,741
LH123Ht	11,161,491	1,208,926	104,891
LH82	12,255,189	1,285,313	105,323
Lo3	14,238,292	1,426,952	124,129
MBS847	12,026,058	1,260,694	106,079
PHG35	12,564,429	1,305,163	106,186
PHG39	9,345,231	965,746	94,057
PHN82	11,649,929	1,224,619	104,339
PHP02	12,203,337	1,260,503	104,832
PHR03	11,593,591	1,195,478	104,471
PHW52	7,035,096	800,381	66,397

Table 7. Number of variants detected for each maize line as compared to inbred line B73. SNPs: Single Nucleotide Polymorphisms, indels: insertion/deletions shorter than 50 bp, SVs: structural variants longer than 50 pb. Counts were obtained using the –snps and –indels flag of bcftools, for SNPs and indels respectively with the condition of a least one alternative allele (COUNT(GT ="alt") > 0), and simply using this condition for the number of SVs on the associated structural variants vcf files.

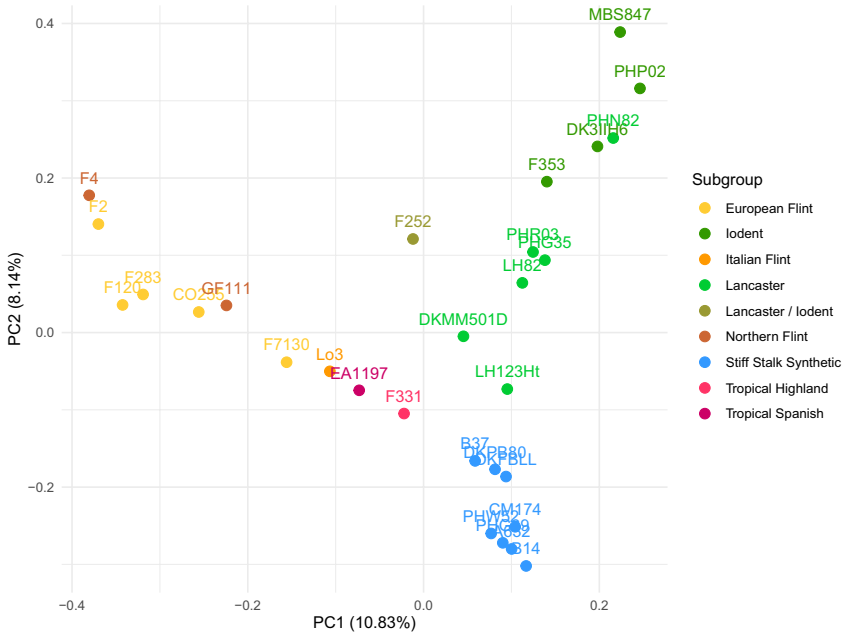

Fig. 2 PCA constructed from the (standardized) genotypes of 1 millions randomly selected SNPs.

Code availability

All the codes used for the analysis can be found on the SeqOccIn project's GitHub page, following the path Data paper/Zea mays data paper: https://github.com/GeTPlaGe/SeqOccIn/tree/main/Data%20paper/Zeamays. The pipeline used for aligning reads and calling variants is available here: https://github.com/SeqOccin-SV/SeqOccinVariants.

Received: 7 July 2025; Accepted: 9 March 2026;
Published online: 19 March 2026

References

1. Wrigley, C. W. & Nirmal, R. C. The major cereal grains: Corn, rice, and wheat, https://doi.org/10.1002/0471238961.23080501.a01.pub3 (2017).
2. Wang, Q. & Dooner, H. K. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proceedings of the National Academy of Sciences* **103**, 17644–17649, https://doi.org/10.1073/pnas.0603080103 (2006).
3. Stitzer, M. C., Anderson, S. N., Springer, N. M. & Ross-Ibarra, J. The genomic ecosystem of transposable elements in maize. *PLOS Genetics* **17**, e1009768, https://doi.org/10.1371/journal.pgen.1009768 (2021).
4. Ou, S. *et al*. Differences in activity and stability drive transposable element variation in tropical and temperate maize. *Genome Research* **34**, 1140–1153, https://doi.org/10.1101/gr.278131.123 (2024).
5. Wallace, J. G. *et al*. Association mapping across numerous traits reveals patterns of functional variation in maize. *PLoS Genetics* **10**, e1004845, https://doi.org/10.1371/journal.pgen.1004845 (2014).
6. Zhou, P., Hirsch, C. N., Briggs, S. P. & Springer, N. M. Dynamic patterns of gene expression additivity and regulatory variation throughout maize development. *Molecular Plant* **12**, 410–425, https://doi.org/10.1016/j.molp.2018.12.015 (2019).
7. Ricci, W. A. *et al*. Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants* **5**, 1237–1249, https://doi.org/10.1038/s41477-019-0547-0 (2019).
8. Marand, A. P. *et al*. The genetic architecture of cell type-specific cis regulation in maize. *Science* **388**, https://doi.org/10.1126/science.ads6601 (2025).
9. Fagny, M. *et al*. Identification of key tissue-specific, biological processes by integrating enhancer information in maize gene regulatory networks. *Frontiers in Genetics* **11**, https://doi.org/10.3389/fgene.2020.606285 (2021).
10. Springer, N. M. *et al*. The maize w22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics* **50**, 1282–1288, https://doi.org/10.1038/s41588-018-0158-0 (2018).
11. Sun, S. *et al*. Extensive intraspecific gene order and gene structural variations between mo17 and other maize genomes. *Nature Genetics* **50**, 1289–1295, https://doi.org/10.1038/s41588-018-0182-0 (2018).
12. Yang, N. *et al*. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nature Genetics* **51**, 1052–1059, https://doi.org/10.1038/s41588-019-0427-6 (2019).
13. Lin, T., Song, Y., Lawrence, P., Kheshgi, H. S. & Jain, A. K. Worldwide maize and soybean yield response to environmental and management factors over the 20th and 21st centuries. *Journal of Geophysical Research: Biogeosciences* **126**, https://doi.org/10.1029/2021jg006304 (2021).
14. Chen, J. *et al*. A complete telomere-to-telomere assembly of the maize genome. *Nature Genetics* **55**, 1221–1231, https://doi.org/10.1038/s41588-023-01419-6 (2023).
15. Darracq, A. *et al*. Sequence analysis of european maize inbred line f2 provides new insights into molecular and chromosomal characteristics of presence/absence variants. *BMC Genomics* **19**, https://doi.org/10.1186/s12864-018-4490-7 (2018).
16. Haberer, G. *et al*. European maize genomes highlight intraspecies variation in repeat and gene content. *Nature Genetics* **52**, 950–957, https://doi.org/10.1038/s41588-020-0671-9 (2020).
17. Hufford, M. B. *et al*. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, 655–662, https://doi.org/10.1126/science.abg5289 (2021).
18. Mayjonade, B. *et al*. Extraction of high-molecular-weight genomic dna for long-read sequencing of single molecules. *BioTechniques* **61**, 203–205, https://doi.org/10.2144/000114460 (2016).
19. Workman, R. *et al*. High molecular weight dna extraction from recalcitrant plant species for third generation sequencing v1. https://doi.org/10.1038/protex.2018.059 (2018).
20. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *journal = PeerJ* **6**, https://doi.org/10.7717/peerj.4958 (2018).
21. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
22. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell systems* **3**, 95–98 (2016).
23. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
24. Durand, N. C. *et al*. Juicebox provides a visualization system for hi-c contact maps with unlimited zoom. *Cell Systems* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
25. Alonge, M. *et al*. Automated assembly scaffolding using ragtag elevates a new tomato system for high-throughput genome editing. *Genome Biology* **23**, https://doi.org/10.1186/s13059-022-02823-7 (2022).
26. Bradnam, K. R. *et al*. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, https://doi.org/10.1186/2047-217x-2-10 (2013).
27. Seppey, M., Manni, M. & Zdobnov, E. M. *BUSCO: Assessing Genome Assembly and Annotation Completeness*, 227–245 (Springer New York, 2019).
28. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, https://doi.org/10.1186/s13059-020-02134-9 (2020).
29. Smolka, M. *et al*. Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology* **42**, https://doi.org/10.1038/s41587-023-02024-y (2024).
30. Kirsche, M. *et al*. Jasmine and Iris: population-scale structural variant comparison and analysis. *Nature Methods* **20**, https://doi.org/10.1038/s41592-022-01753-3 (2023).
31. *European nucleotide archive*. https://www.ebi.ac.uk/ena/browser/view/PRJEB67812 (2025).
32. *European variant archive*. https://www.ebi.ac.uk/eva/?eva-study=PRJEB106599 (2026).
33. The 29 maize lines SNP and SV variant set. https://doi.org/10.57745/7AUTOL (2025).
34. Germplasm Resources Information Network (GRIN) — doi.org. https://doi.org/10.15482/USDA.ADC/1212393.
35. European nucleotide archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB60075 (2023).
36. Byrne, P. F. *et al*. Sustaining the future of plant breeding: The critical role of the usda-ars national plant germplasm system. *Crop Science* **58**, 451–468, https://doi.org/10.2135/cropsci2017.05.0303 (2018).

37. Camus-Kulandaivelu, L. *et al.* Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the dwarf8 gene. *Genetics* **172**, 2449–2463, https://doi.org/10.1534/genetics.105.048603 (2006).
38. Bouchet, S. *et al.* Adaptation of maize to temperate climates: Mid-density genome-wide association genetics and diversity patterns reveal key genomic regions, with a major contribution of the vgt2 (zcn8) locus. *PLoS ONE* **8**, e71377, https://doi.org/10.1371/journal.pone.0071377 (2013).

Acknowledgements

We thank "La Région Occitanie" and European Union for funding the project as part of the Occitanie Region's "Regional Research and Innovation Platforms" call for projects under the FEDER-FSE MIDI-PYRENEES ET GARONNE 2014-2020 Operational Program. We thank KWS, Maisadour, Euralis, Caussade semences, Syngenta, RAGT and Limagrain for their financial support and their inputs for choosing the genetic material analyzed. We thank Valérie Combes for sample preparation, Delphine Madur and Nathalie Rivière for genotype validation, Gaëtan Givry for EVA data submission and Jorge Duarte and Johann Joets for insightful discussions on maize genome scaffolding. We are grateful to Cyril Bauland for expertise in maize germplasm accession nomenclature. We thank Carine Palaffre and French maize inbred lines seed bank (CRB, INRAE Saint Martin de Hinx), the U.S. National Plant Germplasm System (NPGS)[35] and the USDA Agricultural Research Service Germplasm Resources Information Network (GRIN)[36] for providing seeds with traced seedlots, as well as Adrienne Ressayre and Christine Dillmann (GQE-Le Moulon) for providing seeds of F252 and MBS847, and Silvio Salvi (University of Bologna) for early access to seeds from the GF111 inbred line, Carlotta Balconi (CREA-Research Centre for Cereal and Industrial Crops) for providing access to Lo3, and CSIC (Consejo Superior de Investigaciones Científicas) for authorizing the use of EM1197. GeT core facility https://doi.org/10.15454/1.5572370921303193E12 is supported by France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by the French Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We are grateful to the genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, https://doi.org/10.15454/1.5572369328961167E12) for providing computing and storage resources.

Author contributions

C.D., D.M., and Ch.G. conceived and supervised the whole "SeqOccIn" project. Cl.V. and A.C. conceived the maize-related sub-project of the "SeqOccIn" project. C.D., D.M., Ch.G., Cl.V. and A.C. secured funding. C.I. coordinated data generation and quality control. C.I., C.M., C.E., E.D. produced sequence data. Ch.K., T.F. and Cl.K. supervised bioinformatic analyses. C.B., A.D.F., T.F., J.D., S.N., Cl.V. and Ch.K. analysed the results. S.P. and A.C. coordinated the selection of the inbred lines with private partners. Cl.K. and Ca.V. secured data and submitted them to public databases. C.I., Cl.V., Ch.K., S.N. and T.F. wrote the original draft of the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07055-z.

Correspondence and requests for materials should be addressed to C.D., C.V., C.K. or C.I.