

1 **Extensive intra-specific gene order and gene structural variations**
2 **between Mo17 and other maize genomes**

3

4 Silong Sun^{1*}, Yingsi Zhou^{1*}, Jian Chen^{1*}, Junpeng Shi^{1*}, Haiming Zhao^{1*}, Hainan
5 Zhao¹, Weibin Song¹, Mei Zhang^{1, 12}, Yang Cui¹, Xiaomei Dong¹, Han Liu¹, Xuxu
6 Ma¹, Yinping Jiao³, Bo Wang³, Xuehong Wei³, Joshua C. Stein³, Jeff C. Glaubitz⁴, Fei
7 Lu^{5,13,14}, Guoliang Yu⁶, Chengzhi Liang⁷, Kevin Fessler⁸, Bailin Li⁹, Antoni Rafalski⁹,
8 Patrick S. Schnable¹⁰, Doreen H. Ware^{3, 11}, Edward S. Buckler^{5, 11}, Jinsheng Lai^{1,2#}

9

10 ¹State Key Laboratory of Agrobiotechnology and National Maize Improvement
11 Center, Department of Plant Genetics and Breeding, China Agricultural University,
12 Beijing, 100193, China

13 ²Center for Crop Functional Genomics and Molecular Breeding, China Agricultural
14 University, Beijing, 100193, China

15 ³Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

16 ⁴Genomics Facility, Institute of Biotechnology, Cornell University, Ithaca, New York
17 14850, USA

18 ⁵Institute for Genomic Diversity, 175 Biotechnology Building, Cornell University,
19 Ithaca, New York 14850, USA

20 ⁶Nextomics Biosciences Co., Ltd., Wuhan, Hubei, 430205, China

21 ⁷State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental
22 Biology, Chinese Academy of Sciences, Beijing 100101, China

23 ⁸DuPont Pioneer, Johnston, IA 50131, USA

24 ⁹DuPont Pioneer, Wilmington DE 19880, USA

25 ¹⁰Department of Agronomy, Iowa State University, Ames, IA 50011, USA

26 ¹¹USDA-ARS, NEA Robert W. Holley Center, Cornell University, Ithaca, New York
27 14853, USA

28 ¹²Department of biology, Stanford University, Stanford, CA 94305, USA

29 ¹³State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of

30 Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101,
31 China

32 ¹⁴CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS),
33 Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai
34 200032, China

35

36 *These authors contributed equally to this work.

37 #Corresponding author: Jinsheng Lai

38 China Agricultural University, Beijing 100193, P.R. China

39 Tel: 86-10-62731405

40 Email: jlai@cau.edu.cn

41 **Supplementary Table**

42 **Supplementary Table 1.** Summary of sequencing data of Mo17 genome

	Pacbio raw data	Bionano raw data	Illumina PCR-free raw data
Total length (Gb)	200.8	267.7	251.75
Genome depth*	~90x	~122x	~114x
N50 size (Kb)	13.3	278.3	-
Longest length (Mb)	0.054	2.24	-
Number of sequences	24,110,142	987,732	968,268,842

43 * Estimated with the assembled genome size of 2.18 Gb.

44 **Supplementary Table 2.** Details of the Mo17 genome assembly

	Pacbio assembled Contigs		Pacbio+Bionano hybrid Scaffolds	
	Size (Mb)	Number	Size (Mb)	Number
N10	3.36	54	25.07	8
N20	2.63	127	17.10	19
N30	2.16	218	14.46	32
N40	1.83	326	11.85	49
N50	1.48	457	10.20	69
N60	1.19	620	7.99	94
N70	0.93	822	5.98	125
N80	0.68	1,092	4.64	166
N90	0.40	1,491	2.31	231
Longest	7.26	-	32.18	-
Total number	-	4,257	-	2,560
Total size	2,147.5	-	2,182.6	-

45 **Supplementary Table 3.** Assessment of the Mo17 genome assembly using 9 published BAC sequences

BACs Accession No.	BAC size (bp)	Matched scaffold	Scaffold size (bp)	BAC region matched by scaffolds	Scaffold region Matched by BACs	Coverage of BAC region (%)	Identity of aligned regions (%)
AY664416	203,581	scaffold149	5,129,234	1 - 203,581	4,500,886 - 4,702,664	100%	99.99
AY664417	366,120	scaffold43	12,596,929	1 - 366,120	1,921,911 - 2,317,476	100%	99.83
AY664418	282,600	scaffold19	17,096,410	1 - 282,600	8,711,156 - 9,005,593	100%	99.92
AY664419	405,672	scaffold55	11,263,077	1 - 405,672	5,463,406 - 5,871,148	100%	99.99
DQ002406	92,777	scaffold11	22,050,599	1 - 92,777	17,970,669 - 18,062,443	100%	99.99
DQ002408	57,607	scaffold39	13,268,609	1 - 57,607	1,996,841 - 2,054,100	100%	99.96
AY455286	94,829	scaffold48	11,935,030	1 - 94,829	9,118,069 - 9,208,027	100%	99.98
AY146798	10,854	scaffold10	23,028,026	1 - 10,854	7,046,579 - 7,057,432	100%	99.99
GQ923937	13,749	scaffold103	7,477,135	1 - 13,749	2,828,763 - 2,842,524	100%	99.97

46 **Supplementary Table 4.** Details of the 10 Mo17 pseudo-chromosomes

Chromosome	Length of pseudomolecule (bp)	# Anchored scaffolds	# Anchored annotated genes
1	306,176,853	55	5,727
2	243,312,244	43	4,535
3	240,957,546	34	4,089
4	249,748,977	43	3,956
5	220,382,597	35	4,388
6	168,720,447	31	3,124
7	178,574,342	31	3,098
8	182,612,305	30	3,429
9	164,939,053	34	2,820
10	149,041,351	26	2,664
Total anchored	2,104,430,515 (96.42%)	362	37,830 (97.95%)
Unassigned	78,184,926(3.58%)	2,198	790 (2.05%)
Total	2,182,615,441	2,560	38,620

47 **Supplementary Table 5.** Comparison of repetitive elements between B73, Mo17 and PH207

Classification	Mo17		B73		PH207	
	Length (bp)	Percentage of genome (%)	Length (bp)	Percentage of genome (%)	Length (bp)	Percentage of genome (%)
Class I: Retrotransposon	1,615,808,455	75.24	1,590,805,799	75.62	1,237,393,323	72.20
SINE	1,186,211	0.06	1,157,458	0.06	1,156,854	0.07
LINE	18,525,059	0.86	18,416,979	0.87	17,759,793	1.04
LTR-Retrotransposon	1,596,097,185	74.32	1,571,231,362	74.69	1,218,476,676	71.09
<i>Copia</i>	549,129,783	25.57	543,580,110	25.84	374,649,221	21.86
<i>Gypsy</i>	1,044,412,734	48.63	1,025,189,781	48.73	841,483,696	49.10
Class II: DNA Transposon	131,369,364	6.12	129,769,009	6.17	111,255,792	6.49
hAT	20,797,357	0.98	20,555,145	0.98	19,578,748	1.14
MULE	16,091,283	0.75	16,042,769	0.76	13,794,997	0.80
CMC-EnSpm	57,358,930	2.67	56,299,448	2.68	41,293,002	2.41
Harbinger	17,713,498	0.82	17,594,763	0.84	17,428,901	1.02
TcMar-Stowaway	2,544,246	0.12	2,477,258	0.12	2,546,850	0.15
Helitron	15,017,854	0.70	14,948,170	0.71	14,757,676	0.86
Unclassified	36,922,258	1.72	36,313,300	1.73	31,890,727	1.86
Satellites	16,162,294	0.75	5,060,924	0.24	7,742,641	0.45
Total Content	1,800,262,371	83.83	1,761,949,032	83.76	1,388,282,483	81.00

48 **Supplementary Table 6.** Summary of the RNA-seq data

Tissue Description	Read type	Read Length (bp)	Raw reads	Mapped reads
B73 materials				
12Day_endosperm rep1	PE	100	19,049,102	17,916,798
12Day_endosperm rep2	PE	100	16,612,376	15,722,273
14Day_seedling rep1	PE	100	13,424,192	12,852,797
14Day_seedling rep2	PE	100	14,969,460	14,332,393
Bract rep1	PE	100	20,349,526	13,512,828
Bract rep2	PE	100	23,677,876	18,527,727
Root rep1	PE	100	24,039,714	18,003,245
Root rep2	PE	100	21,336,116	12,796,563
Stem rep1	PE	100	23,431,080	12,558,028
Stem rep2	PE	100	19,739,992	7,592,632
Mo17 materials				
12Day_endosperm rep1	PE	100	18,836,188	18,421,811
12Day_endosperm rep2	PE	100	14,262,930	13,976,102
14Day_seedling rep1	PE	100	13,975,918	13,756,473
14Day_seedling rep2	PE	100	13,437,504	13,214,650
Bract rep1	PE	100	22,501,512	10,106,459
Bract rep2	PE	100	17,526,886	11,553,124
Root rep1	PE	100	21,587,728	19,329,334
Root rep2	PE	100	21,666,142	19,843,595
Stem rep1	PE	100	19,144,244	9,713,435
Stem rep2	PE	100	23,043,870	14,653,307

49 **Supplementary Table 7.** Statistics of B73, Mo17 and PH207 gene models

	Mo17 genes	B73 genes	PH207 genes
Number of gene model	38,620	39,324	37,613
Number of transcripts	46,856	131,318	55,039
Mean exons per gene [#]	4.68	4.93	4.63
Mean CDS length [#]	1,135	1,178	1,055
Numbers of genes with gaps in intragenic region	12	14	13,265
Expressed genes with CDS coverage > 90% ^{#*}	21,136	21,625	19,038
Expressed genes with CDS coverage > 50% ^{#*}	28,228	29,759	27,933
Homology of swiss-prot	26,085	28,090	26,106
Supported by Pfam domain	26,757	28,551	26,146

50

51 [#] Analyzed based on the longest transcript.

52 ^{*}B73 and Mo17 genes were evaluated using the RNA-seq data of bract, root, stem,
53 seedling and endosperm generated in this study. PH207 genes were evaluated by
54 RNA-seq data generated in the PH207 genome paper.

55 **Supplementary Table 8.** Summary of aligned sequences, SNPs and Indels in B73, Mo17 and
56 PH207 genomes

	B73 to Mo17	Mo17 to PH207	B73 to PH207
	B73	Mo17	B73
One-to-One syntenic	1287.5Mb (61.13%)	1071.7Mb (50.93%)	1101.7Mb (52.30%)
blocks	Mo17	PH207	PH207
	1287.4Mb (61.18%)	1071.6Mb (52.01%)	1101.4Mb (53.46%)
SNPs	9,867,466	8,598,810	8,290,900
Insertions (<100bp)	630,711	610,716	522,433
Deletions (<100bp)	791,735	580,110	556,289

57 **Supplementary Table 9.** Summary of PAV clusters longer than 500 kb between B73 and PH207
58 genomes

PAVs cluster in B73			PAVs cluster in Mo17		
Region	Size (bp)	Gene number*	Region	Size (bp)	Gene number*
chr6:22,483,501-25,398,000	2,914,500	66	chr6:63,965,001-66,467,900	2,502,900	10
chr6:100,489,301-101,117,900	628,600	13	chr2:235,809,501-236,562,100	752,600	20
chr2:228,268,301-228,815,200	546,900	8	chr2:237,529,501-238,164,700	635,200	23
chr5:5,422,601-5,965,600	543,000	21	chr9:153,099,801-153,638,800	539,000	16
chr4:114,977,101-115,505,300	528,200	9			

59 *Gene number in PAV clusters represented all annotated genes, with part of them identified as
60 PAV genes.

61 **Supplementary Table 10.** Summary of PAV genes in B73, Mo17 and PH207 genomes

	B73 to Mo17	Mo17 to B73	PH207 to B73	PH207 to Mo17
PAV genes had orthologs in sorghum	19	11	5	23
PAV genes present in the wild relatives	70	49	21	74
Total	72	50	22	75

62 **Supplementary Table 11.** Genome distribution of SNPs and indels between B73 and Mo17
63 genomes

Region	SNPs		B73 Deletion /Mo17 Insertion		B73 Insertion/Mo17 Deletion	
	Number [#]	Percent	Number	Percent	Number	Percent
Intergenic	5,844,655	59.23	195,956	31.07	318,425	40.22
Upstream 5kb	1,538,530	15.59	162,031	25.69	176,223	22.26
3'UTR	75,202	0.76	13,647	2.16	14,473	1.83
CDS	207,856	2.11	11,091	1.76	12,162	1.54
Intron	603,083	6.11	75,225	11.93	81,591	10.31
5'UTR	75,662	0.77	13,626	2.16	14,397	1.82
Downstream 5kb	1,522,478	15.43	159,135	25.23	174,464	22.04
Total	9,867,466	100	630,711	100	791,735	100

64 [#] The statistics was based on the annotation of B73 genome.

Supplementary Table 12. Variations within genes between B73 and PH207, Mo17 and PH207 genomes

Variation type	PH207 genes to B73 genome		PH207 genes to Mo17 genome	
	Syntenic genes	Nonyntenic genes	Syntenic genes	Nonyntenic genes
1. Structurally conserved genes	27,004 (81.50%)	1,112(33.86%)	25,092 (78.46%)	1,547 (34.59%)
Without amino acid substitutions	12,979	251	11,445	395
No DNA variation in CDS	10,606	191	9,176	286
No DNA variation in CDS and intron	6,800	114	5,739	174
No DNA variation in genic region*	1,627	8	1,226	26
With amino acid changes	14,025	861	13,647	1,152
With missense mutation in CDS	13,786	851	13,391	1,138
With 3n indel in CDS	4,951	171	4,930	293
2. Genes with large effect mutations	3,649 (11.01%)	602 (18.33%)	4,209 (13.16%)	906 (20.26%)
Start codon mutation	195	54	227	99
Stop codon mutation	247	91	265	150
Splice donor mutation	62	19	101	45
Splice acceptor mutation	119	55	172	100
With 3n±1 indel in CDS	2,379	256	2,630	389
Pre-mature of stop codon	2,151	418	2,578	586
3. Genes with large structural variations	2,479 (7.49%)	1,548 (47.14%)	2,679 (8.38%)	1,944 (43.47%)
At least one exon missing	901	1,351	1,007	1,635
4. PAV genes	-	22	-	75
Total	33,132[#]	3,284[#]	31,980[#]	4,472[#]

[#] Only genes and their best hits in the counterpart genome anchored on 10 pseudomolecules were included for the analysis.

* Genic regions include the 2 kb upstream and downstream of the gene body.

71 **Supplementary Table 13. Summary of variations genes in maize and rice genomes**

	B73 genes to Mo17 genome	Mo17 genes to B73 genomes	PH207 genes to B73 genome	PH207 genes to Mo17 genome	Nipponbare genes to R498 genome	R498 genes to Nipponbare genome
Large effect and Large structure variation genes	9,058 (23.35%)	8,153 (21.68%)	8,278 (22.73%)	9,738 (26.71%)	5,455 (15.29%)	7,534 (20.06%)
Non-syntenic genes	5,105 (13.16%)	4,008 (10.66%)	3,284 (9.02%)	4,472 (12.27%)	1,283 (3.60%)	1,962 (5.23%)
Total analyzed genes	38,786	37,605	36,416	36,416	35,679	37,549

72

73

74 **Supplementary Table 14.** Summary of duplicated genes and corresponding gene families in B73, Mo17 and PH207 genomes

	B73 annotation			Mo17 annotation			PH207 annotation		
	Genes	TFs (% *)	Families (Genes)	Genes	TFs (% *)	Families (Genes)	Genes	TFs (% *)	Families (Genes)
Singleton	3,589	33	2,129 (2,162)	5,044	42	2,186 (2,251)	4,987	46	2,436 (2,532)
	(9.13%)	(0.92%)		(13.06%)	(0.83%)		(13.26%)	(0.92%)	
Dispersed	20,874	863	6,439 (19,531)	19,468	782	5,956 (18,382)	20,082	875	6,330 (18,697)
	(53.08%)	(4.13%)		(50.41%)	(4.01%)		(53.39%)	(4.36%)	
Proximal	1,331	43	546 (1,287)	1,324	42	551 (1,300)	942	37	384 (909)
	(3.38%)	(3.23%)		(3.43%)	(3.17%)		(2.5%)	(3.93%)	
Tandem	2,281	95	845 (2,203)	2,109	74	749 (2,033)	1,694	63	639 (1,641)
	(5.80%)	(4.16%)		(5.46%)	(3.67%)		(4.5%)	(3.72%)	
WGD/segmental	11,249	1,202	3,719 (11,036)	10,675	1,138	3,573 (10,539)	9,908	987	3,309 (9,741)
	(28.61%)	(10.68%)		(27.64%)	(10.66%)		(26.34%)	(9.96%)	
Total	39,324	2,236	10,771 (36,219)	38,620	2,078	10,160 (34,505)	37,613	2,008	10,394 (33,520)

75 *: The percentage of TFs relative to the genes number of each category.

76

77 **Supplementary Table 15.** Number of duplicated non-syntenic genes in B73, Mo17 and PH207
78 genomes

	B73 to Mo17	Mo17 to B73	PH207 to B73	PH207-to-Mo17
Singleton	369 (7.23%)	559 (13.95%)	337 (10.26%)	467 (10.44%)
Dispersed	4,173 (81.74%)	2,949 (73.71%)	2,537 (77.25%)	3,349 (74.89%)
Proximal	76 (1.49%)	81 (2.02%)	55 (1.67%)	69 (1.54%)
Tandem	170 (3.33%)	161 (4.02%)	99 (3.01%)	133 (2.97%)
WGD/segmental	317 (6.21%)	258 (6.44%)	256 (7.80%)	454 (10.15%)
Total	5,105	4,008	3,284	4,472

79 **Supplementary Table 16.** Summary of non-syntenic genes in B73, Mo17 and PH207 genomes

	B73 to Mo17	Mo17 to B73	PH207 to B73	PH207 to Mo17
Number of annotated genes	39,498	38,620	37,613	37,613
Number of genes that had orthologs in sorghum	23,032 (58.31%)	21,703 (56.19%)	21,914 (58.23%)	21,914 (58.23%)
Number of non-syntenic genes	5,105	4,008	3,284	4,472
Number of non-syntenic genes that had orthologs in sorghum	578 (11.32%)	518 (12.92%)	597 (18.18%)	1,038 (23.21%)

80

81 **Supplementary Table 17.** Number of genes specifically expressed in B73 or Mo17 tissues

Tissue	Expressed genes		Specifically expressed genes	
	B73	Mo17	B73	Mo17
Bract	20,997	19,731	401	177
Root	20,176	20,297	344	205
Stem	18,722	18,497	317	266
Seedling	20,146	20,081	308	339
Endosperm	15,476	15,229	254	277
Total	24,209	23,947	859	770
Common*	13,414	13,276	55	5
Always specific if expressed	-	-	417	281

82 * The genes showed specifically expressed in B73 or Mo17 for all five tissues.

83 **Supplementary Table 18. Number of expressed genes in B73 and Mo17 tissues with different**
84 **degree of structure variations**

Types	B73			Mo17		
	Total	Expressed genes	Specifically expressed genes	Total genes	Expressed genes	Specifically expressed genes
Structurally conserved genes	29,656	19,605 (66.11%)	439 (1.48%)	29,402	19,403 (65.99%)	369 (1.26%)
Genes with large effect mutations	5,334	2,608 (48.89%)	170 (3.19%)	4,997	2,424 (48.51%)	180 (3.6%)
Genes with large structure variations	3,724	1,660 (44.58%)	215 (5.77%)	3,156	1,542 (48.86%)	180 (5.7%)
PAV genes	72	26 (36.11%)	25 (34.72%)	50	23 (46%)	16 (32%)
Total	38,786	23,889	849	37,605	28,614	745

85 **Supplementary Table 19. Variations within genes between Rice *Japonica* Nipponbare and**
86 **Indica R498 genomes**

Variation type	Japonica genes to Indica genome		Indica genes to Japonica genome	
	Syntenic genes	Nonsyntenic genes	Syntenic genes	Nonsyntenic genes
1. Structurally conserved genes	30,059 (87.39%)	165 (12.86%)	29,613 (82.61%)	402 (20.49%)
Without amino acid substitutions	12,556	25	11,476	306
No DNA variation in CDS	8,859	13	8,065	246
No DNA variation in CDS and intron	6,326	11	5,598	169
No DNA variation in genic region*	1,380	1	1,345	10
With amino acid changes	17,503	140	18,137	910
With missense mutation in CDS	16,862	140	18,137	899
With 3n indel in CDS	4,139	44	4,475	221
2. Genes with large effect mutations	3,636 (11.72%)	224 (17.46%)	4,977(13.88%)	347 (17.69%)
Start codon mutation	306	27	381	45
Stop codon mutation	442	60	665	68
Splice donor mutation	57	7	79	4
Splice acceptor mutation	108	5	105	15
With 3n±1 indel in CDS	1,797	109	2,335	169
Pre-mature of stop codon	2,246	132	3,363	241
3. Genes with large structural variations	701 (4.79%)	674 (52.53%)	997 (2.78%)	931 (47.45%)
At least one exon missing	205	499	239	685
4. PAV genes	-	220	-	282
Total	34,396[#]	1,283[#]	35,847[#]	1,962[#]

87 [#] Only genes and their best hits in the counterpart genome anchored on chromosomes were
88 included for the analysis.

89 * Genic regions include the 2 kb upstream and downstream of the gene body

