

<b>Manuscript Number:</b>	GIGA-D-17-00007R2	
<b>Full Title:</b>	Construction of the third generation Zea mays haplotype map	
<b>Article Type:</b>	Research	
<b>Funding Information:</b>	National Science Foundation (IOS #1238014)	Dr. Edward S Buckler
	Agricultural Research Service	Dr. Edward S Buckler
	National Institute of Food and Agriculture (2009-65300-05668)	Dr. Edward S Buckler
	Bill and Melinda Gates Foundation (US)	Dr. Yunbi Xu
	National Natural Science Foundation of China (CN) (#31271736)	Not applicable
	National Key Basic Research Program of China (2014CB138206)	Not applicable
<b>Abstract:</b>	<p><b>Background</b>            Characterization of genetic variations in maize has been challenging, mainly due to deterioration of collinearity between individual genomes in the species. An international consortium of maize research groups combined resources to develop the maize haplotype version 3 (HapMap 3), built from whole genome sequencing data from 1,218 maize lines, covering pre-domestication and domesticated Zea mays varieties across the world.</p> <p><b>Results</b>            A new computational pipeline was set up to process over 12 trillion bp of sequencing data, and a set of population genetics filters were applied to identify over 83 million variant sites.</p> <p><b>Conclusions</b>            We identified polymorphisms in regions where collinearity is largely preserved in the maize species. However, the fact that the B73 genome used as the reference only represents a fraction of all haplotypes is still an important limiting factor.</p>	
<b>Corresponding Author:</b>	Qi Sun Cornell University Ithaca, NY UNITED STATES	
<b>Corresponding Author Secondary Information:</b>		
<b>Corresponding Author's Institution:</b>	Cornell University	
<b>Corresponding Author's Secondary Institution:</b>		
<b>First Author:</b>	Robert Bukowski	
<b>First Author Secondary Information:</b>		
<b>Order of Authors:</b>	Robert Bukowski	
	Xiaosen Guo	
	Yanli Lu	
	Cheng Zou	
	Bing He	
	Zhengqin Rong	
	Bo Wang	

	Dawen Xu Xu
	Bicheng Yang
	Chuanxiao Xie
	Longjiang Fan
	Shibin Gao
	Xun Xu
	Gengyun Zhang
	Yingrui Li
	Yinping Jiao
	John F Doebley
	Jeffrey Ross-Ibarra
	Anne Lorant
	Vince Buffalo
	M. Cinta Romay
	Edward S Buckler
	Yunbi Xu
	Doreen Ware
	Jinsheng Lai
	Qi Sun
<b>Order of Authors Secondary Information:</b>	
<b>Response to Reviewers:</b>	In this revised version, the reference to the supporting data has been added, and new NCBI accessions have been added for each data set.
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<b>Experimental design and statistics</b>	Yes
Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a> . Information essential to interpreting the data presented should be made available in the figure legends.	
Have you included all the information requested in your manuscript?	
<b>Resources</b>	Yes
A description of all resources used, including antibodies, cell lines, animals and software tools, with enough	

<p>information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.</p> <p>Have you have met the above requirement as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	<p>Yes</p>

# Construction of the third generation *Zea mays* haplotype map

Robert Bukowski<sup>1</sup>, Xiaosen Guo<sup>2,3</sup>, Yanli Lu<sup>4</sup>, Cheng Zou<sup>5</sup>, Bing He<sup>2</sup>, Zhengqin Rong<sup>2</sup>, Bo Wang<sup>2</sup>, Dawen Xu<sup>2</sup>, Bicheng Yang<sup>2</sup>, Chuanxiao Xie<sup>5</sup>, Longjiang Fan<sup>6</sup>, Shibin Gao<sup>4</sup>, Xun Xu<sup>2</sup>, Gengyun Zhang<sup>2</sup>, Yingrui Li<sup>2</sup>, Yinping Jiao<sup>7</sup>, John Doebley<sup>8</sup>, Jeffrey Ross-Ibarra<sup>9</sup>, Anne Lorant<sup>9</sup>, Vince Buffalo<sup>9</sup>, M. Cinta Romay<sup>10</sup>, Edward S. Buckler<sup>10,11</sup>

## *Corresponding authors:*

Yunbi Xu<sup>5,12</sup>, Jinsheng Lai<sup>13</sup>, Doreen Ware<sup>7</sup>, and Qi Sun<sup>1</sup>

## *Authors' affiliations:*

<sup>1</sup>Bioinformatics Facility, Institute of Biotechnology, Cornell University, Ithaca, NY 14853

<sup>2</sup>BGI-Shenzhen, Shenzhen 518083, China

<sup>3</sup>Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

<sup>4</sup>Maize Research Institute, Sichuan Agricultural University, Wenjiang 611130, Sichuan, China

<sup>5</sup>Institute of Crop Science, Chinese Academy of Agricultural Sciences/National Key Facilities for Crop Gene Resource and Genetic Improvement, Beijing 100081, China

<sup>6</sup>Institute of Crop Science and Institute of Bioinformatics, Department of Agronomy, Zhejiang University, Hangzhou 310058, China

<sup>7</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

<sup>8</sup>Department of Genetics, University of Wisconsin, Madison, Wisconsin, USA

<sup>9</sup>Department of Plant Sciences, University of California, Davis, California, USA

<sup>10</sup>Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853

<sup>11</sup>US Department of Agriculture-Agricultural Research Service, Ithaca, NY 14853

<sup>12</sup>International Maize and Wheat Improvement Center (CIMMYT), El Batan 56130, Texcoco, Mexico

<sup>13</sup>National Maize Improvement Center, China Agricultural University (CAU)

1  
2  
3  
4 **ABSTRACT**  
5  
6

7 **Background**  
8  
9

10  
11 Characterization of genetic variations in maize has been challenging, mainly due to deterioration of  
12  
13 collinearity between individual genomes in the species. An international consortium of maize research  
14  
15 groups combined resources to develop the maize haplotype version 3 (HapMap 3), built from whole  
16  
17 genome sequencing data from 1,218 maize lines, covering pre-domestication and domesticated *Zea mays*  
18  
19 varieties across the world.  
20  
21

22  
23  
24 **Results**  
25  
26

27  
28 A new computational pipeline was set up to process over 12 trillion bp of sequencing data, and a set of  
29  
30 population genetics filters were applied to identify over 83 million variant sites.  
31  
32

33 **Conclusions**  
34  
35

36  
37 We identified polymorphisms in regions where collinearity is largely preserved in the maize species.  
38  
39 However, the fact that the B73 genome used as the reference only represents a fraction of all haplotypes  
40  
41 is still an important limiting factor.  
42  
43

44 **KEYWORDS**  
45  
46

47  
48 *Zea mays*, sequencing, haplotype map, genotyping, variant discovery, linkage disequilibrium, identity by  
49  
50 descent, imputation  
51  
52

53  
54 **BACKGROUND**  
55  
56

57  
58 Maize, one of the most important cereals in the world, also happens to be among the crop species with  
59  
60 the most genetic diversity. Advances in the next generation sequencing technologies made it possible to  
61  
62

1  
2  
3  
4 characterize genetic variations in maize at genomic scale. The previously released maize HapMap2 were  
5  
6 constructed with whole genome sequencing data of 104 maize lines across pre-domestication and  
7  
8 domesticated *Zea mays* varieties [1]. Since then, more maize lines have been sequenced by the  
9  
10 international research community, and a consortium was formed to develop the next generation  
11  
12 haplotype map. The maize HapMap 3 consortium includes, among others, BGI-Shenzen, Chinese Academy  
13  
14 of Agricultural Sciences, China Agricultural University (CAU), International Maize and Wheat Improvement  
15  
16 Center (CIMMYT). High-coverage data for 31 European and US Flint and Dent lines is also available in Ref.  
17  
18 [2]. Altogether, in this work we used a total of 1218 maize lines sequenced with depth varying from below  
19  
20  
21  
22  
23 1x to 59x.

24  
25  
26 A common approach in today's genetic diversity projects is to map the shotgun sequencing reads from  
27  
28 each individual onto a common reference genome to identify DNA sequence variations, and the physical  
29  
30 positions of the reference genome is used as a coordinate system for the polymorphic sites. A good  
31  
32 example is the human 1000 genome project [3]. The computational data processing pipeline developed  
33  
34 for the human project, GATK, has been widely adopted for identifying genetic variations in many other  
35  
36 species [4].  
37  
38  
39  
40

41 As the sequencing technology is improved and sequencers' base calling error model gets more accurate,  
42  
43 the computational challenges in genotyping by short-read sequencing have shifted from modeling  
44  
45 sequencer machine artifacts errors to resolving genotyping errors derived from incorrect mapping of short  
46  
47 reads to the reference genome. The problem is associated with the experimental design that uses the  
48  
49 single-reference genome as coordinate system. Taking maize as an example, the reference being used is  
50  
51 a 2.1 Gb assembly from an elite inbred line B73 that represents 91% of the B73 genome [5], and was  
52  
53 estimated to capture only ~70% of the low-copy gene fraction of all inbred lines [6]. The sequence  
54  
55 alignment software, however, can map 95-98% of the whole genome sequencing reads to the reference.  
56  
57  
58 That suggests a high percentage of the reads were mapped incorrectly, either being mapped to the  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 paralogous loci or highly repetitive regions under-represented in the reference assembly. The genetic  
5  
6 variants called from the miss-mapped reads need to be corrected computationally. The maize HapMap2  
7  
8 relied on linkage disequilibrium in the population to purge most of the bad markers caused by alignment  
9  
10 errors. To construct maize HapMap 3, a new computational pipeline was developed from scratch to  
11  
12 handle the sequencing data from 10 times more lines, and also took advantage of the high quality genetic  
13  
14 map constructed from the GBS technology [7, 8] which was not present when HapMap2 was constructed.  
15  
16  
17

18  
19 Genome structure variation in the population, including transposition, deletion, duplication and inversion  
20  
21 of the genomic segments, poses another challenge in the HapMap projects. As the physical genomes of  
22  
23 each of the individuals included in the HapMap projects vary both by size and structure, and there is no  
24  
25 co-linearity of all the sequence variants between the reference and genomes of each of the individuals, it  
26  
27 is not always possible to anchor all genetic variants in a population onto a single reference coordinate  
28  
29 system. As a compromise, markers included in the maize HapMap are defined as sites of which the  
30  
31 physical positions of the B73 alleles matching the markers' consensus genetic mapped positions.  
32  
33  
34

35  
36 Here we present maize haplotype map version 3 (HapMap 3), which is a result of coordinated efforts of  
37  
38 the international maize research community. The build includes 1,218 lines and over 83 million variant  
39  
40 sites anchored to the B73 reference genome version AGP v3.  
41  
42  
43  
44

## 45 DATA DESCRIPTION

46  
47  
48 The sequencing data used in this work is comprised of 12,497 billion base pairs in a total of 113,702 billion  
49  
50 Illumina paired-end reads, originating from 1,218 maize and teosinte lines. The data was collected from  
51  
52 several sources over several years, and varies in quality, read length, and coverage. Basic information  
53  
54 about various datasets and stages of the HapMap 3 project they were used in are summarized in Table 1.  
55  
56 Each of the 1,218 lines were sequenced at depth varying from below 1x to 59x, using reads of lengths  
57  
58 ranging from 44 through 201 bp, averaging 110 bp. All reads were aligned to maize reference genome B73  
59  
60  
61  
62  
63  
64  
65

version AGP v3 using BWA mem aligner [9]. Overall, 95-98% of the reads were mapped to the reference genome, although only about 50-60% with non-zero mapping quality.

Table 1: Sequence datasets used in various stages of HapMap 3

Dataset	# Taxa	Coverage per taxon			3.1.1	3.2.1unimp	3.2.1imp
		Minimum	Maximum	Average			
HapMap2	103	1	18.5	4.1	+	+	+
Hapmap2 extra	44	4.2	42	11.5	+	+	+
CAU	725	0.06	36.8	1.75	+	+	+
CIMMYT/BGI	89	1.1	19	11	+	+	+
282-2x	271	0	9	2.2	-	+	+
282-4x	270	0.6	34.5	4.4	-	+	-
German, Ref. [2]	31	8.3	59	17.4	-	+	+

Taxa from sets “HapMap2”, “HapMap2 extra”, and “CAU” partially overlap. The “282” libraries, sequenced twice represent 271 taxa. A “+” means that the dataset was used in a given stage, “-“ that it was not.

All sequence data used in this work is publically available. Collection and publishing of this data does not violate any local or international legislation or guidelines.

## ANALYSIS

### Initial variant discovery

The HapMap 3 pipeline is summarized in Figure 1. First, polymorphic sites were called for a set of 916 taxa from datasets HapMap2 through CIMMYT/BGI (7,191 billion base pairs, 74,643 million reads). In the first step, a custom built software tool was used to determine genotypes for each taxon at each site of the genome based on allelic depths at that site. Bases counted towards depth had base quality score of at least 10 and originated from reads with mapping quality ( $MAPQ = -\text{int}(10 \log P)$ , where  $P$  – calculated by the BWA mem aligner - is the probability of the reported read location being wrong) at or above 30. This cut-off was chosen at mid-point between the highest MAPQ value reported by the aligner,

1  
2  
3  
4 corresponding to unambiguous alignments (60) and that of the most ambiguous ones (0). Analysis of the  
5  
6 inbreeding coefficient (Section HapMap 3.1.1) and of MAPQ distributions shows that our choice of cut-off  
7  
8 leads to decent quality genotypes while allowing for over 80% of alignments with MAPQ>0 to be included.  
9  
10 Only sites where at least 10 taxa had coverage of 1 or more were considered. Following Ref. [1], at each  
11  
12 site the allelic read depths were subject to segregation test (ST – see METHODS section for details). For a  
13  
14 population of inbred lines at true variant sites, one expects depths corresponding to minor and major  
15  
16 alleles to be concentrated in roughly different subset of taxa rather than being randomly distributed. The  
17  
18 purpose of the ST test is to find and eliminate sites for which allelic depth distribution appears random,  
19  
20 as such randomness, indicating high heterozygosity, is likely caused by alignment and sequencing errors.  
21  
22 A measure of the randomness is the p-value of the ST test (the smaller the p-value, the less random the  
23  
24 distribution). A p-value threshold of 0.01 was used in this study. This choice was somewhat arbitrary,  
25  
26 aimed at reducing the number of tentative variant sites to manageable size before further, more stringent  
27  
28 filters were applied. In this first, ST-based round of filtering, a total of 196 million tentative polymorphic  
29  
30 sites were selected. In the second step, these sites were filtered using the identity by descent (IBD)  
31  
32 information derived from about 0.5 million of high-quality polymorphisms obtained previously [8] using  
33  
34 the Genotyping-By-Sequencing (GBS) approach [7]. These GBS variants (GBS anchor) were used to  
35  
36 determine regions of IBD, where certain pairs of taxa are expected to have identical haplotypes. The  
37  
38 tentative polymorphic sites violating these IBD constraints were then filtered out, leaving 96.8 million  
39  
40 sites. At roughly half of the sites surviving this filter, minor allele was not present in taxa involved in the  
41  
42 tested IBD relationships. At such sites (typically with low minor allele frequency), the satisfied IBD  
43  
44 constraints do not confirm the existence of a variant. They are therefore less reliable and have been  
45  
46 marked with “IBD1” flag in the VCF files (see Table 2 for summary of flags and parameters present in  
47  
48 HapMap 3 VCF files). The ST- and IBD-filtered variant sites were then used in two separate procedures,  
49  
50 leading to two versions of HapMap 3 genotypes, referred to as HapMap 3.1.1 and HapMap 3.2.1.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 2: Flags and parameters used in INFO field of VCF files in various HapMap 3 versions.

Parameter	3.1.1	3.2.1 unimp	3.2.1 imp	Description
DP	+	+	+	Total read depth at the site
NZ	+	+	+	Number of taxa with called genotypes
AD	+	+	+	Allelic depths (reference, alternative in order listed in ALT field)
AC	+	+	+	Numbers of alternative alleles in order listed in ALT field
AQ	+	+	+	Average allele base qualities (reference, alternative in order listed in ALT field) computed in HapMap 3.1.1 from 916 taxa
GN	+	+	+	Numbers of genotypes (AA,AB,BB or AA,AB,AC,BB,BC,CC if 2 alt alleles present)
HT	+	+	+	Number of heterozygotes
EF	+	+	+	EF=heterozygosity/(presence_frequency*minor_allele_frequency); computed in HapMap 3.1.1 from 916 taxa
PV	+	+	+	p-value from segregation test, computed in HapMap 3.1.1. from 916 taxa
MAF	+	+	+	Minor allele frequency (summed up over all alternative alleles)
MAFO	-	-	+	Minor allele frequency in unimputed HapMap 3.2.1.
FH	+	-	-	Fraction of heterozygous taxa among the 506 taxa with more than 50% non-missing genotypes on chr 10
FH2	+			Site with FH greater than 2%
IBD1	+	+	+	only one allele present in IBD contrasts - based on 916 taxa of HapMap 3.1.1
LLD	+	+	+	Site in local LD with GBS map - based on 916 taxa of HapMap 3.1.1
NI5	+	+	+	Indel or site within 5 bp of a putative indel - from 916 taxa of HapMap 3.1.1
INHMP311	-	+	+	Site present in HapMap 3.1.1
ImpHomoAccuracy	-	-	+	Fraction of homozygotes imputed back into homozygotes
ImpMinorAccuracy	-	-	+	Fraction of minor allele homozygotes imputed back into minor allele homozygotes
DUP	-	-	+	Site with heterozygotes frequency > 3% - based on unimputed HapMap 3.2.1 genotypes

“+” and “-” indicate presence or absence, respectively, of a parameter or flag in a given version of HapMap. For example, “-+-” means the parameter is present in VCF file of both unimputed and imputed HapMap 3.2.1, and absent from HapMap 3.1.1. VCF files. Unless indicated otherwise, all parameters are computed from depths and genotypes in the current VCF file.

## HapMap 3.1.1

The HapMap 3.1.1 procedure involved checking for linkage disequilibrium of each site against the GBS anchor map [7, 8], which consists of markers located in hypo-methylated and genetically stable regions. Sites giving only very weak or only nonlocal (i.e., outside of 1 Mb radius) linkage Disequilibrium (LD) hits were eliminated, which resulted in the final set of 61,228,639 polymorphisms. For slightly less than 40% of these sites, LD could not be conclusively calculated due to small minor allele frequencies (MAF), whereas the remaining sites, confirmed to be in local LD with the GBS anchor, have been marked with flag “LLD”. Among the sites surviving all filtering steps, 8.7 million are indels or are located near (within 5 bp) of an indel. These have been marked with the flag “NI5”. Since a procedure to achieve consistent alignment across all reads covering the same indels - local realignment - is not computationally feasible at this scale and has not been performed, genotyping errors could occur, and, consequently, most such sites are tentative and should be treated with caution.

Figure 2 shows overlaps between various classes of variants of HapMap 3.1.1. First, we notice a rather small overlap between sites in confirmed local LD (“LLD” flag) and those marked “IBD1”. This is understandable, as the IBD1 sites represent mostly low MAF cases, where LD assessment could not be done. Indels and vicinity (labeled “NI5”) constitute about 15% of sites in each of the LLD, IBD1, and the union of LLD and IBD1 sets. Only a very small fraction of sites does not carry LLD or IBD1 flag, i.e., they are strongly confirmed by the IBD filter, but could not be classified with LD. The subset of 29.8 million sites in local LD and away from indels should be considered the most reliable.

To check the sensitivity of the obtained variant set to the mapping quality threshold imposed on the reads counted towards allelic depths, we repeated the pipeline using the mapping quality threshold equal to 1. Comparison of the variant set obtained this way (referred to as q1) with our recommended set (q30) is shown in Figure 3. While the overall number of variant sites is approximately independent of the mapping

1  
2  
3  
4 quality threshold, the two pipelines produce significantly different sets of sites, with only 72% of all q30  
5  
6 sites reproduced by the q1 pipeline. Closer inspection shows that this variability is due primarily to the  
7  
8 IBD1 sites, for which our filtering strategy was the least stringent. On the other hand, the LLD sites,  
9  
10 confirmed to be in local LD with GBS anchor, are much more independent of the mapping quality  
11  
12 threshold, which confirms high quality of such sites.  
13  
14

15  
16  
17 For a population of inbred lines considered here, insight into genotype quality may be obtained from the  
18  
19 inbreeding coefficient, calculated here for each taxon using the VCFtools program [11] from the formula  
20  
21

$$F_{inbr} = (O - E)/(N - E)$$

22  
23  
24  
25 where  $O$  is the observed number of homozygotes for a given taxon,  $N$  is the number of sites at which the  
26  
27 taxon was genotyped, and  $E$  is the expected number of homozygotes given by  $E = \sum_k(1 - 2p_kq_k)$ .  
28  
29 Summation in the latter formula runs over  $N$  genotyped sites,  $p_k$  is the minor allele frequency at site  $k$   
30  
31 (computed from all taxa in the population with non-missing genotypes at this site), and  $q_k = (1 - p_k)$ .  
32  
33 Low values of the inbreeding coefficient, indicating high heterozygosity, are mostly due to genotyping  
34  
35 errors. Importance of choosing a sufficiently tight mapping quality threshold for the quality of genotypes  
36  
37 is apparent from Figure 4, where the distribution of inbreeding coefficient for chromosome 10 is shown  
38  
39 for q1 and q30 variant sets. The lower MAPQ threshold results in a large number of miss-mapped reads  
40  
41 being counted towards depth, producing overly heterozygous genotypes, especially for highly covered  
42  
43 taxa (the peak below 0.8 is due mostly to CIMMYT lines with 10-15x coverage; these lines have higher  
44  
45 heterozygosity than other lines which may also contribute to the peak) and thus shifting the curve to the  
46  
47 left. Since most HapMap 3 taxa are inbred lines, one should expect the true distribution to be contained  
48  
49 within peak around 0.95. In view of this, the q30 result is definitely an improvement over q1, although a  
50  
51 longer than expected tail extending towards the value 0.8 indicates that the HapMap 3 variants may  
52  
53 contain too many false heterozygotes.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 Seemingly heterozygous sites may result from either sequencing errors or misalignments of reads  
5  
6 originating from paralogous regions. To investigate this further, we calculated, for each site, the fraction  
7  
8 of heterozygous HapMap 3.1.1 genotypes within a subset of 506 high-coverage taxa (defined as those  
9  
10 with more than 50% non-missing genotypes on chromosome 10). In HapMap 3.1.1 VCF files, this fraction  
11  
12 has been recorded as parameter “FH”. At sites for which this parameter exceeds 2-3%, heterozygotes are  
13  
14 likely to originate from misalignments, for example, from tandem and ectopic duplications. Such sites  
15  
16 constitute 9% of all HapMap 3.1.1 sites.  
17  
18  
19  
20

## 21 HapMap 3.2.1

22  
23  
24 The 96.8 million ST- and IBD-filtered variant sites were the starting point for the HapMap 3.2.1 procedure  
25  
26 (Figure 1). On these sites, genotypes were called on the 263 taxa from the “282” panel of Ref. [10] using  
27  
28 “282 2x” dataset, and on the 31 high-coverage (on average 17 x) “German” taxa [2], for the total of 1210  
29  
30 taxa. Some of the taxa present in the “282” and “German” sets carry the same names as the ones included  
31  
32 in the 916-taxon HapMap 3.1.1 set. Since despite identical names such taxa often originate from different  
33  
34 germplasm sources, they have been kept separate during genotyping, i.e., reads from different sources  
35  
36 were not merged and separate genotypes were computed for each source. In the resulting VCF files, the  
37  
38 names of the overlapping taxa have been prefixed by “282set\_” and “german\_”. For example, in the case  
39  
40 of B73, there are three columns representing different datasets for this taxon: “B73” (the original 916-  
41  
42 taxa set), “282set\_B73” (sequence from the more recent “282” libraries), and “german\_B73” (from Ref.  
43  
44 [2]).  
45  
46  
47  
48  
49  
50

51 To further eliminate the false positives resulting from sequencing errors, an additional depth-based filter  
52  
53 was applied to the 96.8 million sites. Referred to as “>1,>2” filter it accepts sites for which the read support  
54  
55 of minor allele was greater than 1 in at least one taxon and greater than 2 across all taxa. Genotypes on  
56  
57 the surviving 83,153,144 sites, referred to as “un-imputed HapMap 3.2.1”, were then processed through  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 the LD KNN imputation procedure based on Ref. [12], where the “nearest neighbors” of a given line are  
5  
6 selected based on sites in good local LD with the target site. Whenever possible, the procedure filled up  
7  
8 missing genotypes with imputed ones, but the non-missing genotypes were left unchanged, even if  
9  
10 imputation classified them differently. Non-imputable missing genotypes at the sites with (pre-  
11  
12 imputation) MAF below 1% were assumed to be major allele homozygotes. Imputation reduces the  
13  
14 fraction of missing genotypes from 50% to 7%. Most of the originally missing genotypes (about 85%) are  
15  
16 imputed to major allele homozygotes. Accuracy of the genotype dataset can be assessed by comparing  
17  
18 the original genotypes with imputed ones. As shown in Table 3, 99.8% of major allele homozygotes are  
19  
20 imputed back into the same class. While the accuracies of minor allele homozygotes and genotypes  
21  
22 including indels are both above 90%, only 11% of heterozygotes are imputed back into the same class,  
23  
24 while 47% of them fail imputation altogether. This reflects the inherent difficulty in calling heterozygotes.  
25  
26 In the single-reference approach to maize genotyping employed here, heterozygous sites represent true  
27  
28 residual heterozygosity as well as misalignments of reads from tandem and ectopic duplications. Since  
29  
30 residual heterozygosity in our population of predominantly inbred lines should not exceed 2-3%, all  
31  
32 heterozygotes with frequency  $\geq 3\%$  can be considered a result of misalignments. About 10% of all  
33  
34 heterozygotes present in HapMap 3.2.1 set satisfy this condition. In the VCF files, these sites have been  
35  
36 flagged with flag DUP (“duplicated regions”). Other parameters generated by the imputation procedure  
37  
38 and recorded for each variant site in the INFO field are ImpHomoAccuracy fraction of all homozygotes  
39  
40 imputed back into homozygotes and ImpMinorAccuracy fraction of minor allele homozygotes imputed  
41  
42 back to the same class. The INFO field also contains flags IBD1, LLD, and NI5, computed from the initial  
43  
44 916 taxa in the HapMap 3.1.1 procedure. Genotypes resulting from the imputation procedure are referred  
45  
46 to as “imputed HapMap 3.2.1”.

47  
48 Table 3: Accuracy of various genotype classes based on statistics from imputation in HapMap 3.2.1  
49  
50  
51  
52  
53  
54  
55  
56

Genotype class	Accuracy within class [%]	% unimputed
Major allele homozygote	99.8	1.2
Heterozygote	11.1	47.0
Minor allele homozygote	94.4	14.2
Indel	92.2	17.3

Accuracy computed as percentage of the original number of genotypes in a given class (excluding genotypes that could not be imputed) imputed into the same class. The last column shows the fraction of genotypes within a class which could not be imputed.

Relationship between variant sites included in HapMap 3.1.1 and 3.2.1 is shown in Figure 5. Both pipelines start from the same set of IBD-filtered genotypes and subject them to different kinds of filtering, with that of HapMap 3.1.1 being more stringent. It is therefore not surprising that HapMap 3.2.1 recovers the majority (86%) of HapMap 3.1.1 sites, including over 99% of those flagged LLD (i.e., confirmed in local LD). In addition, 30.3 million extra sites are retained in HapMap 3.2.1, which failed the LD filter in HapMap 3.1.1 pipeline. On the other hand, the depth-based “>1,>2” filter applied in HapMap 3.2.1 eliminated 8.2 million sites present in HapMap 3.1.1, including about 0.2 million LLD ones.

After the HapMap 3.2.1 release was completed, “282-2x” sequencing data became available for additional 8 taxa from the “282” panel. Libraries for all 271 taxa were also re-sequenced at a higher depth (average of about 4.4x), leading to another dataset, “282-4x” (as this re-sequencing failed for one of the taxa, this dataset only contained 270 taxa). Therefore, the un-imputed HapMap 3.2.1 genotypes for all 271-taxa from the “282” panel were re-called using the full available sequencing depth, creating a separate variant dataset for the “282” panel.

## DISCUSSION

The maize genome, 2.3 GB in size [5], is smaller than the human genome. But some of its distinctive features makes it more challenging for variants identification. First, a recent whole genome duplication occurred 12 million years ago resulted in homologous segments that complicate the short read

1  
2  
3  
4 alignments; second, the rampant activities of transposable elements within last 1-5 million year not only  
5  
6 resulted in accumulation of large amount of relatively young repetitive elements in the intergenic regions,  
7  
8 but also extraordinary structural variations within species [5, 6]. In this study, the genome of the B73  
9  
10 maize line was used as the reference for variant calling from short sequencing reads. Structural variations  
11  
12 between B73 and other individuals has been the major challenge for identification of true variants. In  
13  
14 particular, short reads derived from regions missing in the reference genome could be mismatched to  
15  
16 other paralogous regions, which lead to false positive genotypes. In the human 1000 genome project, a  
17  
18 new HaplotypeCaller was used [4], which performs local *de novo* assembly to identify the most likely  
19  
20 haplotypes for each individual and thus improve the genotyping results. However, HaplotypeCaller is  
21  
22 computationally very expensive, and not always applicable in species like maize, where the single  
23  
24 reference genome misses many haplotypes presents in the species and has a lot more mismatched  
25  
26 paralogous reads that would disrupt the local assembly. To filter out these false positive variants called  
27  
28 from the mismatched reads, we relied on the *Zea* GBS map [7, 8], which was obtained from GBS markers  
29  
30 located primarily in hypo-methylated chromosomal regions. GBS maps were used to identify IBD regions  
31  
32 between the individual genomes, and 100 million markers with high percentage of mismatched genotype  
33  
34 calls in the IBD regions were filtered out from the initial set of 196 million markers. The highly repetitive  
35  
36 genomic regions derived from recent transposition activities are in general easier to identify, because the  
37  
38 templates of these repeats are well represented on the reference genome, and sequencing reads mapped  
39  
40 to these regions, flagged with low mapping quality, can be removed at the early stage of the analysis  
41  
42 pipeline. For HapMap 3, reads with mapping quality lower than 30 were not included in the build.  
43  
44  
45  
46  
47  
48  
49  
50  
51

52 One of the goals of HapMap 3.1.1 is to identify genetic markers in regions where collinearity is preserved  
53  
54 in majority of maize lines. The LD filter in the pipeline was applied for this purpose. To do this, we  
55  
56 genetically mapped the presence/absence of the minor alleles using the GBS genetic map, and these  
57  
58 mapped genetic positions were compared to the physical positions on the B73 reference. Among the 96.8  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 million sites surviving the IBD filter, 25% did not have enough non-missing data or sufficient minor allele  
5  
6 frequency for genetic mapping to be meaningful. For 38% of sites, at least one genetically mapped  
7  
8 position matching the physical positions on B73 reference was found, 24% have no significant hits from  
9  
10 genetic mapping, probably due to no consensus positions in the HapMap 3 population, and 13% have  
11  
12 genetic positions not matching the B73 physical positions. Markers from the latter two categories (37% of  
13  
14 all IBD-filtered markers) were removed by the LD filter, leaving slightly over 61 million sites, about 60% of  
15  
16 which were confirmed in local LD and marked with a flag “LLD” in VCF files.  
17  
18  
19  
20

21 The IBD and LD filters applied in the HapMap 3.1.1 project effectively remove majority of the false positive  
22  
23 genetic variants caused by paralogous genomic regions, as well as markers with lost collinearity between  
24  
25 the species. However, not all the genotyping errors have been removed from the release. 23,839,286 of  
26  
27 the sites do not have sufficient minor allele frequency for genetic test (these are missing the “LLD” label  
28  
29 in the INFO field of the VCF files). Another source of errors are paralogous regions evolved from tandem  
30  
31 duplications. Misalignments of reads from such regions result in false heterozygous genotypes with  
32  
33 relatively high frequency and in local LD, and therefore difficult to filter out. Given enough sequencing  
34  
35 depth, the tandem duplications can be identified either as copy number variation or imputation errors.  
36  
37 However, majority of the HapMap 3 lines have very low sequencing depth, and fail to sample all  
38  
39 paralogous loci or all alleles, which makes it difficult to flag all sites complicated by tandem duplications.  
40  
41  
42  
43  
44

45 Local LD filter based on a large, diverse population may be too stringent, as some markers, good within  
46  
47 certain sub-populations, may be thrown out. Therefore, the LD filter was not used in the HapMap 3.2.1  
48  
49 release, which contains a total of 83 million variant sites, subject only to ST and IBD filters and an  
50  
51 additional depth-based filter aimed to improve reliability of rare allele calls. Although those sites are likely  
52  
53 to have higher misalignment rates, they are still likely to capture a true association with phenotypes.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 In the un-imputed HapMap 3.2.1, at about 10% of all variant sites, fraction of heterozygous taxa exceeds  
5  
6 3%. Such sites are marked “DUP”, as most likely originating from duplication misalignments. Figure 6  
7  
8 shows the distribution of fraction of heterozygous sites per taxon for different versions of HapMap 3.2.1  
9  
10 release. While for the un-imputed genotypes the distribution peaks slightly below 1%, imputation  
11  
12 significantly shifts the peak to the left, down to about 0.5%. This is a consequence of most missing  
13  
14 genotypes being imputed to homozygotes. Interestingly, considering only sites in good local LD (marked  
15  
16 with the “LLD” flag) leads to distributions (both imputed and un-imputed) shifted towards higher  
17  
18 heterozygosities. This is understandable, as the LLD sites are typically those with higher minor allele  
19  
20 frequencies, where the chance of encountering a heterozygote is higher.  
21  
22  
23  
24  
25

26 In summary, besides the addition of more maize lines, the HapMap 3.2.1 release differs from the 3.1.1  
27  
28 release in three major aspects: 1) Improved rare allele calls: to increase the accuracy of the variants with  
29  
30 rare allele, the HapMap 3.2.1 pipeline applied more stringent read depth thresholds instead of the  
31  
32 population genetics based LD filter that could not be applied to sites with very low MAF; 2) The sites with  
33  
34 high percentage of heterozygous calls were flagged in the VCF files; 3) Missing data was imputed using  
35  
36 the LD KNN method. As summarized in Table 2, the VCF files of both datasets contain labels that flag the  
37  
38 characteristics of each of the sites. To effectively use this resource, it is recommended to filter the sites  
39  
40 based on the flags that are appropriate to the purpose of each project.  
41  
42  
43  
44  
45

46 When constructing the maize HapMap 3, the most serious problems we were facing can be attributed to  
47  
48 the use of a genome from a single individual (B73) as a reference for other, often very different species.  
49  
50 This is becoming the single limiting factor in the study of maize diversity, as well as breeding practice. The  
51  
52 only remedy is to move away from a single genome-based reference coordinate and adopt a pan-genome  
53  
54 based reference system that incorporates all major haplotypes of the species.  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## METHODS

### Plant material

Plant material used in this study was obtained mostly from maize inbred lines representing wide range of *Zea mays* diversity. 103 of these lines, used previously in the HapMap2 project [1], include 60 improved lines, including the parents of the maize nested association mapping (NAM) population [13], 23 maize landraces and 19 wild relatives (teosinte lines: 17 *Z. mays ssp. parviglumis* and 2 *Z. mays ssp. mexicana*). Sequence datasets originating from these lines are referred to in Table 1 as “HapMap2” and “HapMap2 extra”. Majority of the remaining inbred lines originated from CAU (sequence dataset “CAU”) and include, among others, “Chinese NAM” parent lines. Additional 89 inbred lines were provided by CIMMYT and sequenced at BGI (dataset “CIMMYT/BGI”). The HapMap 3 population also contained one *Tripsacum* line (TDD39103), one “mini-maize” line (MM-1A), and a few newly sequenced landraces. Overall, the number of taxa in the initial, variant-discovery stages of the HapMap 3.1.1 project was 916.

The sequence of 271 taxa from the libraries of the “282” panel [10] were added at a later stage (HapMap 3.2.1). DNA to construct these libraries was obtained from the collection that the North Central Regional Plant Introduction Station (NCRPIS) distributes all over the world. Additionally, the high-coverage data of Ref. [2], originating from 31 European and US inbreds was also included. The total number of taxa genotyped in the HapMap 3.2.1 build is 1218.

In this study, individuals with the same taxa name but contributed by different members of the consortium were kept as separate entries in genotyping pipeline - a decision prompted by comparison of genotypes obtained from different datasets. For example, the newly sequenced CML103 is significantly more heterozygous from CML103 studied previously in the HapMap2 project. Also, the Mo17 sequence originating at CAU has been treated as taxon separate from Mo17 and CAUMo17. In most of those cases,

1  
2  
3  
4 a prefix or suffix indicating the origin of the sequence data has been added to the taxa name (e.g.,  
5  
6 "282set\_" or "german\_", "-chin").  
7  
8

## 9 10 Sequencing

11  
12 Sequencing has been performed over several years using various generations of Solexa/Illumina  
13  
14 instruments and library preparation protocols, giving paired end reads from 44 to 201 bp long. Overall,  
15  
16 113.7 billion reads were obtained on 1,218 lines, containing 12,497 billion base pairs, giving on average  
17  
18 4.4x coverage per line (assuming 2.3 Gb genome size). However, as shown in Table 1, coverage was not  
19  
20 uniform among all lines. For a few lines, sequence generated previously in the context of HapMap2 project  
21  
22 was augmented with reads from recent re-sequencing which brought the median coverage of the  
23  
24 HapMap2 lines to 5x, with average coverage equal to 7.8x and standard deviation of 7.2x. All NAM parent  
25  
26 lines are covered to 10x or higher. Most of the 89 lines provided by CIMMYT and sequenced at BGI have  
27  
28 coverage exceeding 10x. The recent re-sequencing of the "282" panel resulted in coverage between 1.7  
29  
30 and 36x, averaging 6.5x. Coverage of the 31 "German" lines for Ref. [2] ranges from 8.3 to 59x, with  
31  
32 average of 17.4x. Majority of the inbred lines originated from CAU have been sequenced at a lower  
33  
34 coverage (1-2x). The list of all lines used in HapMap 3 with the corresponding coverage is given in  
35  
36  
37  
38  
39  
40  
41 Additional file 1.  
42  
43

## 44 Alignment

45  
46 Due to the use of different versions of Solexa/Illumina sequencing equipment, the base qualities in  
47  
48 different input FASTQ files are given in different encodings. Prior to alignment, all base qualities have been  
49  
50 converted to phred+33 scale. Reads were then aligned to B73 reference (AGP v3) as paired-end using bwa  
51  
52 mem aligner (1) with default options. In 72 read sets (Illumina lanes), for technical reasons a high (6%-  
53  
54 54%) fraction of paired-end fragments was found to be shorter than reads, so that the two ends contained  
55  
56 a part of Illumina adapter and were reverse complements of each other. For such "read-through"  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 fragments, the remnants of Illumina adapter sequences were clipped using TRIMMOMATIC [14] and only  
5  
6 one read was used and aligned as single-end. The bwa mem aligner is capable of clipping the ends of reads  
7  
8 and splitting each read in an attempt to map its different parts to different location on the reference. As  
9  
10 a result, typically over 95% of reads are reported as mapped. However, the fraction of reads with non-  
11  
12 zero mapping quality (negative log of the probability that a read has been placed in a wrong location) is  
13  
14 much lower – typically only 40-50%. Figure 7 shows a typical distribution of the mapping quality obtained  
15  
16 from bwa mem alignment. In practice, we only used alignments with mapping quality of at least 30. A  
17  
18 base was counted towards allele depth if its base quality score was at least 10.  
19  
20  
21  
22

23  
24 It is well known that alignment may be especially ambiguous when reads contain indels with respect to  
25  
26 the reference. In such cases, multiple-sequence realignment approaches have been proposed [4] to find  
27  
28 the correct sequence and location of an indel and avoid spurious flanking SNPs. Since indels are not the  
29  
30 primary focus of this work and since the realignment is computationally very expensive, it has not been  
31  
32 performed by the HapMap 3 pipeline. Thus, although indels and SNPs in their immediate vicinity have  
33  
34 been retained in the HapMap 3 VCF files, they are less reliable and have therefore been marked with “NI5”  
35  
36 label for easy filtering.  
37  
38  
39  
40

## 41 Genotyping pipeline

42  
43  
44 Raw genotypes were obtained using a custom-built multi-threaded java code. First, the code executes  
45  
46 samtools mpileup command (thresholds on the base and mapping quality are imposed here) for each  
47  
48 taxon individually, processing a certain portion of the genome. On a multi-core machine, several such  
49  
50 pileup processes (i.e., for several taxa) can be run concurrently as separate threads. Since we are  
51  
52 predominantly interested in calling SNPs, we use a simplified indel representation where insertions and  
53  
54 deletions with respect to reference are treated as additional alleles “I” and “D”, respectively, regardless  
55  
56 of length and actual sequence of the indel. Read depths and average base qualities of all six alleles (A, C,  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 G, T, I, and D) are extracted from samtools mpileup output for each taxon at each genomic position and  
5  
6 stored in an array shared between all threads. The amount of memory available on the machine along  
7  
8 with the number of taxa determine the upper limit on the size of this array, and therefore – the maximum  
9  
10 size of chromosome chunk which can be processed at one time. As base quality of I and D alleles we took  
11  
12 the value corresponding to the base directly preceding the indel on the reference.  
13  
14

15  
16 Extraction of allelic depths for all genomic positions is time consuming, which presents a major obstacle  
17  
18 if joint genotyping needs to be re-run, for example, upon extending the taxa set. It is therefore  
19  
20 advantageous to run the depth extraction only once for each taxon and save the obtained depths on disk  
21  
22 to be retrieved (rather than re-calculated) during the genotyping step. This way, when the taxa set for  
23  
24 genotyping is extended, mpileup step has to be run only for the newly added taxa. Thus, the program  
25  
26 features an option to save allelic depths and average qualities in specially designed data structures stored  
27  
28 in HDF5 files – one such file per taxon per chromosome. To save space, each allele depth and average  
29  
30 quality is stored as one byte, which allows exact representation of integers from 0 to 182, while higher  
31  
32 integers (up to about 10,000) are represented approximately by negative byte values through a  
33  
34 logarithmic formula with carefully chosen base. Depths and qualities are stored only for sites with non-  
35  
36 zero coverage. The details of the storage format and integer representation in terms of byte variables are  
37  
38 given in Additional file 2.  
39  
40  
41  
42  
43  
44

45  
46 Once the allelic depths for all taxa and a given chunk of the genome are available in shared memory, each  
47  
48 site is evaluated for presence of a tentative SNP. On a multi-core machine, the set of sites within the  
49  
50 genome chunk is divided into subsets processed in parallel on different cores. Sites with less than 10 taxa  
51  
52 with read coverage and those with only reference allele present are ignored. For all other sites, genotypes  
53  
54 are called for all taxa using a simple likelihood model with a uniform error rate [15] assumed at 1%.  
55  
56 Alternative alleles are then sorted according to their allele frequencies and up to two most abundant  
57  
58 alleles are kept, as decided by the segregation test described in the next Section. Sites for which all taxa  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 turn out to be reference homozygotes (which may happen despite non-reference alleles being present in  
5 the mapped reads) are skipped. Raw variant set obtained in this way is then subject to extensive filtering  
6  
7 with the intention of reducing the number of false positives resulting from misalignments.  
8  
9

## 10 11 Filtering

### 12 Segregation test (ST) filter

13  
14  
15 For each pair of alleles obtained in the genotyping step, a 2 by N (where N is the number of taxa)  
16  
17 contingency table is constructed, containing depths of the first allele in row 1 and depths of the second  
18  
19 allele in row 2. The Fisher exact test (FET) is then performed to assess how likely such a table is to occur  
20  
21 by chance. If the expected values of the array elements are sufficiently large, the p-value from FET is  
22  
23 approximated by that from the computationally efficient chi-square test. However, in most cases  
24  
25 encountered here, expensive simulation is needed to obtain sufficiently accurate p-value. To reduce  
26  
27 computational burden, we adopted a hybrid approach based on an empirical observation that for  
28  
29 statistically insignificant cases (p-values larger than 0.2) the chi-square test results in a de facto lower  
30  
31 bound to exact p-values. Thus, the chi-square test is performed first for each site and if the p-value from  
32  
33 this test is below 0.2, more exact p-value is obtained from a simulation procedure. The simulation  
34  
35 procedure used here, implemented in Java, is the same as the one implemented in R package [16]. An  
36  
37 alternative allele is kept if at least one contingency table involving this allele has p-value smaller or equal  
38  
39 to 0.01. If none of the alternative alleles survive the ST filter, the site is skipped (not reported in output).  
40  
41  
42 The ST filter tends to eliminate variant sites resulting from random sequencing errors.  
43  
44  
45  
46  
47  
48  
49  
50

### 51 GBS anchor map and IBD filter

52  
53  
54 Given a set of trustworthy SNPs and a diverse set of 916 taxa it is possible to identify, for an arbitrary  
55  
56 region of the genome, a number of taxa pairs which are identical by descent (IBD) and are therefore  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 expected to have identical genotypes in this region. If known, these IBD pairs can be used as a powerful  
5  
6 filter eliminating variant which violate IBD constraints.  
7  
8

9  
10 To determine the IBD regions, we used the first step of our pipeline to call genotypes for our 916 taxa on  
11  
12 the set of GBS v2.7 sites [7, 8] which tend to concentrate in relatively well-conserved low-copy regions of  
13  
14 the genome and can therefore be considered reliable. This set of 954,384 sites was filtered to include only  
15  
16 SNP (not indel) sites for which the p-value from the segregation test was below 0.05 and which were more  
17  
18 than 5 bp away from any indel. The set of genotypes at 475,272 sites obtained in this way, which will be  
19  
20 referred to as GBS anchor, agree well with those from GBS on 167 taxa present in both sets. Alleles  
21  
22 detected by the HapMap 3 pipeline agreed with those from GBS at 94% of the GBS sites. At 90% of the  
23  
24 sites, fraction of (non-missing data) taxa with genotypes in agreement with those from GBS was at or  
25  
26 above 85%. Genotypes different from GBS ones were observed for 82 taxa. These differences were most  
27  
28 frequent (up to 19% of all sites) for teosinte lines.  
29  
30  
31  
32

33  
34 The GBS anchor was used to compute the genetic distance (identity by state) between any two of the 916  
35  
36 lines in windows containing 2000 GBS sites each (about 8.5 Mbp on average). If the genetic distance within  
37  
38 such a window was  $\leq 0.02$  (about 10 times smaller than the mean distance across all pairs), the two lines  
39  
40 were considered to be in IBD. At least 200 comparable GBS sites (i.e., non-missing data simultaneously on  
41  
42 both lines being compared) were assumed necessary to make the genetic distance calculation feasible.  
43  
44 This allowed for good distance estimate while keeping the number of detected IBD relationships large.  
45  
46  
47

48  
49 The number of taxa involved in IBD relationships in any given window were between 385 (start of  
50  
51 chromosome 10) and 757 (middle of chromosome 7) and averaged 588, leading to large numbers of IBD  
52  
53 contrasts, ranging from 3,710 (beginning of chromosome 4) to 42,890 (middle of chromosome 7), and  
54  
55 averaging 13,500.  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 The tentative (ST-filtered) variant sites were confronted with the IBD information as follows: for each site,  
5  
6 pairs of lines in IBD were determined as described above. Genotypes of IBD-related lines were compared  
7  
8 and the numbers of allele matches and mismatches, summed over all IBD pairs, were counted for each  
9  
10 allele present at the site. If the match/mismatch ratio was at least 2 for at least two alleles, or if only one  
11  
12 allele was present in all IBD contrasts, the site is considered as passing the IBD filter. Such a filter is less  
13  
14 powerful for sites where all bases in IBD lines are major allele homozygotes, i.e., the variant being  
15  
16 evaluated occurs in lines not involved in IBD pairs. Formally, such a site passes our IBD filter, but the actual  
17  
18 variant is not strongly confirmed. These uncertain sites, mostly with low minor allele frequency, are  
19  
20 labeled "IBD1" in the HapMap 3 VCF files and constitute about 50% of all HapMap 3 sites.  
21  
22  
23  
24  
25

#### 26 Linkage Disequilibrium (LD) filter

27  
28 Any true SNP should be in local linkage with other nearby SNPs. This observation is the origin of another  
29  
30 filter used in this work, referred to as the LD filter. For each variable site surviving the ST and IBD filters,  
31  
32 we evaluated LD with each site of the GBS anchor. As the LD measure we chose the p-value from a 2 by 2  
33  
34 contingency table of haplotype counts AB, Ab, aB, ab. For the purpose counting haplotypes, heterozygous  
35  
36 genotypes were treated as homozygous in minor allele, so that each taxon only contributed at most one  
37  
38 haplotype. This tends to somewhat strengthen the LD signal and simplify the calculation. For a pair of sites  
39  
40 to be tested for LD, the following three conditions had to be satisfied to make the calculation meaningful:  
41  
42  
43 i) the two sites were at least 2,500 bp apart, ii) there were at least 40 taxa with non-missing genotypes at  
44  
45 both sites being compared, and iii) at least 2 taxa with minor allele had to be present at each of the two  
46  
47 sites.  
48  
49  
50  
51  
52

53 Filtering procedure executed for each site is summarized in Figure 8. First, LD between the given site and  
54  
55 all sites in GBS anchor was computed and up to 20 best LD hits (the ones with lowest p-values) were  
56  
57 collected. If the p-value of the best hit exceeded  $1E-6$  (which roughly corresponds to the peak of the  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 overall distribution of p-values), the site was rejected. Otherwise, it was determined whether the set of  
5  
6 best hits contained any local hits, i.e., hits to GBS sites on the same chromosome within 1 Mbp of the site  
7  
8 in question and with the p-value smaller than 10 times the p-value of the best hit. If no such local hits  
9  
10 were found, the site was rejected, otherwise it was kept and marked as a site in Local LD using the flag  
11  
12 “LLD”. Note that the procedure defined this way filters out sites with only non-local LD hits as well as  
13  
14 those with only weak LD signal. Sites in local LD as well as those for which LD could not be assessed  
15  
16 (because of low minor allele frequency or missing data) pass the filter.  
17  
18  
19  
20

## 21 Imputation

22  
23  
24 In the HapMap 3.2.1 pipeline, the ST- and IBD-filtered genotypes, after the application of the additional  
25  
26 “>1,>2” depth-based filter, were processed through the LD KNN imputation procedure based on Ref. [12]  
27  
28 to fill in the missing data. The procedure is a version of the “K nearest neighbors” routine where the  
29  
30 “nearest neighbors” of a given taxon are selected based on genetic distance computed using variant sites  
31  
32 in good local LD. Specifically, for a given target site, a list of up to 70 sites in best LD (as given by the  $R^2$   
33  
34 measure) with it is compiled by checking all surrounding sites within 600Kb characterized by  
35  
36 heterozygosity lower than 3% and more than 50% taxa with non-missing genotypes. Capping this list at  
37  
38 70 sites leads to good compromise between distance accuracy and computation speed. Then, at the same  
39  
40 target site, for each target taxon, up to 30 “nearest neighbor” taxa are selected, with lowest genetic  
41  
42 distances from the target taxon. Genetic distances are computed using the set of local LD sites selected  
43  
44 in the previous step. Taxa with more than 50% missing genotypes at LD sites, missing genotype at the  
45  
46 target position, having distance from the current taxon larger than 0.1, or resulting in less than 10 common  
47  
48 LD sites on which the distance can be calculated, are excluded from distance calculation process.  
49  
50 Genotypes of the selected nearest neighbor taxa at the target site are stored in memory along with the  
51  
52 genetic distances from the target taxon. This information is used to compute a weight  $w_i$  of each neighbor  
53  
54 genotype  $g$  as follows:  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

$$w_g = \sum_i \frac{1}{1 + 70d_{gi}},$$

where the summation index  $i$  runs over all neighboring taxa with genotype  $g$  at the target site, and  $d_{gi}$  is the distance of taxon  $i$  from the target taxon. The genotype with the highest weight is considered the imputed genotype (of the target taxon at the target site) provided its weight is at least 10 times larger than that of the second-best candidate genotype. Otherwise the imputation is considered inconclusive and the imputed genotype is set to “unknown” (missing data), as it is in the case when no close neighbors of the current taxon could be found. If a genotype imputed to “unknown” occurs at a site where MAF<1%, it is automatically converted into major allele homozygote.

The imputation procedure is run for each genotype in the input file. However, in the output only the originally missing genotypes are updated to imputed ones, whereas all others are left unchanged, even if classified differently. On the other hand, all imputed genotypes are used during a run to collect imputation statistics. The “transition matrix” showing how many genotypes originally in a given class were imputed into other classes is an indication of the accuracy of the input genotypes. Error rates calculated from the transition data are given in Table 3.

## AVAILABILITY OF DATA

At present, reads from all datasets is available in the form of BAM files (with reads aligned to AGP v3 reference) on CYVERSE data store (formerly iPlant, <http://www.cyverse.org/data-store>), in directories `/iplant/home/shared/panzea/raw_seq_282/bam` (dataset “282-2x” and “282-4x”, also available from NCBI BioProject PRJNA389800), `/iplant/home/shared/panzea/hapmap3/bam_germanlines` (“German” dataset; raw reads available from NCBI BioProject PRJNA260788), and `/iplant/home/shared/panzea/hapmap3/bam` (other datasets).

1  
2  
3  
4 The set of HapMap 3.1.1. polymorphisms determined for 916 taxa (from datasets “HapMap2”,  
5  
6 “HapMap2 extra”, “CAU”, and “CIMMYT/BGI”) is available in VCF format on CYVERSE data store in the  
7  
8 directory /iplant/home/shared/panzea/hapmap3/hmp311, in files  
9  
10 c\*\_hmp311\_q30.vcf.gz (one file per chromosome, where “\*” stands for chromosome 1-10).  
11  
12 Additionally, files c\*\_hmp311\_q1.vcf.gz (in the same location) contain test results obtained with  
13  
14 mapping quality threshold equal to 1.  
15  
16  
17  
18

19 The HapMap 3.2.1 variants for 1210 taxa (916 initial Hapmap 3.1.1 taxa + 263 taxa from “282-2x” set +  
20  
21 31 “German” lines) are available from  
22  
23 /iplant/home/shared/panzea/hapmap3/hmp321/unimputed, in files  
24  
25 mergedflt\_c\*.vcf.gz (un-imputed results) and in  
26  
27 /iplant/home/shared/panzea/hapmap3/hmp321/imputed, in files  
28  
29 mergedflt\_c\*.imputed.vcf.gz (imputed results).  
30  
31  
32  
33

34 Files c\*\_282\_corrected\_onHmp321.vcf.gz in CYVERSE directory

35  
36 /iplant/home/shared/panzea/hapmap3/hmp321/unimputed/282\_libs\_2015 contain  
37  
38 un-imputed genotypes on HapMap 3.2.1 sites from the full depth data available for the “282” panel (271  
39  
40 taxa, datasets “282-2x” + “282-4x”).  
41  
42  
43  
44  
45  
46  
47

## 48 LIST OF ABBREVIATIONS

49  
50 International Maize and Wheat Improvement Center (CIMMYT), Segregation test (ST), Identity-by-descent  
51  
52 (IBD), Genotyping-by-sequencing (GBS), Linkage Disequilibrium (LD), Minor Allele Frequency (MAF),  
53  
54 Nested Association Mapping (NAM), Single Nucleotide Polymorphism (SNP), Insertion-Deletion (Indel),  
55  
56 read mapping quality (MAPQ), base quality (BQ), linkage disequilibrium-base K nearest neighbor  
57  
58 imputation (LDKNN), sites in local LD (LLD), variant call format (VCF)  
59  
60  
61  
62  
63  
64  
65

## FUNDING

This work has been funded by grants from National Key Basic Research Program of China (2014CB138206), National Science Foundation of China (Grant #31271736), Bill & Melinda Gates Foundation (Yunbi Xu), National Science Foundation IOS #1238014, USDA-ARS, and USDA NIFA grant 2009-65300-05668.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## REFERENCES

1. Chia J-M, Song C, Bradbury PJ, Costich D, Leon N de, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet.* 2012;44:803–7. doi:10.1038/ng.2313.
2. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics.* 2014;15:823. doi:10.1186/1471-2164-15-823.
3. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65. doi:10.1038/nature11632.
4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43:491–8. doi:10.1038/ng.806.
5. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science.* 2009;326:1112–5. doi:10.1126/science.1178534.
6. Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of maize. *Science.* 2009;326:1115–7. doi:10.1126/science.1177837.
7. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE.* 2014;9:e90346. doi:10.1371/journal.pone.0090346.
8. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* 2013;14:R55. doi:10.1186/gb-2013-14-6-r55.
9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.
10. Flint-Garcia SA, ThUILlet A-C, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 2005;44:1054–64. doi:10.1111/j.1365-313X.2005.02591.x.
11. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–8. doi:10.1093/bioinformatics/btr330.
12. Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S. LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3 (Bethesda).* 2015;5:2383–90. doi:10.1534/g3.115.021667.

13. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. Genetic properties of the maize nested association mapping population. *Science*. 2009;325:737–40. doi:10.1126/science.1174320.
14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.
15. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda)*. 2011;1:171–82. doi:10.1534/g3.111.000240.
16. Patefield WM. Algorithm AS 159: An Efficient Method of Generating RXC Tables with Given Row and Column Totals. *Applied Statistics*. 1981;30:91–7.

## FIGURES

Figure 1: Overview of HapMap 3 pipeline. Initial set of tentative variant sites was obtained from 916 taxa using reads with mapping quality (MAPQ) at least 30 and bases with base quality (BQ) at least 10. At least 10 taxa had to have non-zero read coverage and the p-value from the segregation test (ST) on allelic depths had to be at most 0.01. This initial set of sites was subject to filtering based on identity by descent (IBD). Application of linkage disequilibrium (LD) filter eliminated sites with only non-local LD hits, leading to HapMap 3.1.1 variant set. An alternative route, leading to HapMap 3.2.1 genotypes, involved K nearest neighbors (KNN) imputation in which distances were computed using sites in good local LD (hence – LD KNN). See text for detailed explanation of methods and acronyms. The exact numbers of variant sites in HapMap 3.1.1 and HapMap 3.2.1 are 61,228,639 and 83,153,144, respectively.

Figure 2: Overlap between various classes of HapMap 3.1.1 polymorphic sites. All sites listed passed the ST and IBD filters. LLD sites are those found in local LD with the GBS anchor. Sites flagged IBD1 passed the IBD filter, however, no alternative allele was present in IBD contrasts. Such sites do not violate IBD, but

1  
2  
3  
4 the existence of a variant is not confirmed. NI5 flag is used to mark indels and sites within 5 bp of an indel.  
5  
6 Since no local re-alignment was done, the NI5 sites are not reliable.  
7  
8  
9

10  
11 Figure 3: Polymorphic sites detected by HapMap 3.1.1 pipeline based on two read mapping quality  
12 thresholds:  $MAPQ \geq 1$  (q1) and  $MAPQ \geq 30$  (q30). Tightening of MAPQ threshold affects mostly the sites  
13  
14 flagged with IBD1 (least reliable), while the LLD sites (in local LD with GBS anchor) are mostly independent.  
15  
16  
17  
18  
19

20  
21 Figure 4: Distribution of inbreeding coefficient for HapMap 3.1.1 variant sets obtained with two read  
22 mapping quality thresholds:  $MAPQ \geq 1$  (q1) and  $MAPQ \geq 30$  (q30). Lower MAPQ threshold leads to lower  
23  
24 values of inbreeding coefficient (i.e., higher heterozygosities) resulting from misaligned reads.  
25  
26  
27  
28  
29

30  
31 Figure 5: Overlap between HapMap 3.1.1 and HapMap 3.2.1 variant sites. 86% of HapMap 3.1.1 sites (99%  
32 of those in local LD) are recovered by the HapMap 3.2.1 pipeline.  
33  
34  
35  
36

37  
38 Figure 6: Distribution of fraction of heterozygous sites per taxon for un-imputed and imputed HapMap  
39 3.2.1. Curves marked LLD have been obtained considering only sites verified in HapMap 3.1.1 to be in  
40  
41 good local LD with GBS anchor.  
42  
43

44  
45 Figure 7: Cumulative distribution of mapping quality from BWA mem alignment of 125.4 million 150 bp  
46 reads from taxon A272.  
47  
48  
49  
50

51  
52 Figure 8: Linkage Disequilibrium-based filtering flowchart. The procedure eliminates sites with weak or  
53 non-local only LD hits. Sites with good local LD hits as well as those for which LD could not be probed  
54  
55 (because of low MAF) are retained.  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## ADDITIONAL FILES

Additional file 1: HapMap3TaxaAndCoverage.xlsx – spreadsheet with a list of all lines used in HapMap 3 with their corresponding coverage

Additional file 2: DepthFormatDetails.pdf - details of byte representation and storage format used for allelic depths

# Construction of the third generation *Zea mays* haplotype map

Robert Bukowski<sup>1</sup>, Xiaosen Guo<sup>2,3</sup>, Yanli Lu<sup>4</sup>, Cheng Zou<sup>5</sup>, Bing He<sup>2</sup>, Zhengqin Rong<sup>2</sup>, Bo Wang<sup>2</sup>, Dawen Xu<sup>2</sup>, Bicheng Yang<sup>2</sup>, Chuanxiao Xie<sup>5</sup>, Longjiang Fan<sup>6</sup>, Shibin Gao<sup>4</sup>, Xun Xu<sup>2</sup>, Gengyun Zhang<sup>2</sup>, Yingrui Li<sup>2</sup>, Yinping Jiao<sup>7</sup>, John F. Doebley<sup>8</sup>, Jeffrey Ross-Ibarra<sup>9</sup>, Anne Lorant<sup>9</sup>, Vince Buffalo<sup>9</sup>, M. Cinta Romay<sup>10</sup>, Edward S. Buckler<sup>10,11</sup>

## *Corresponding authors:*

Yunbi Xu<sup>5,12</sup>, Jinsheng Lai<sup>13</sup>, Doreen Ware<sup>7</sup>, and Qi Sun<sup>1</sup>

## *Authors' affiliations:*

<sup>1</sup>Bioinformatics Facility, Institute of Biotechnology, Cornell University, Ithaca, NY 14853

<sup>2</sup>BGI-Shenzhen, Shenzhen 518083, China

<sup>3</sup>Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

<sup>4</sup>Maize Research Institute, Sichuan Agricultural University, Wenjiang 611130, Sichuan, China

<sup>5</sup>Institute of Crop Science, Chinese Academy of Agricultural Sciences/National Key Facilities for Crop Gene Resource and Genetic Improvement, Beijing 100081, China

<sup>6</sup>Institute of Crop Science and Institute of Bioinformatics, Department of Agronomy, Zhejiang University, Hangzhou 310058, China

<sup>7</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

<sup>8</sup>Department of Genetics, University of Wisconsin, Madison, Wisconsin, USA

<sup>9</sup>Department of Plant Sciences, University of California, Davis, California, USA

<sup>10</sup>Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853

<sup>11</sup>US Department of Agriculture-Agricultural Research Service, Ithaca, NY 14853

<sup>12</sup>International Maize and Wheat Improvement Center (CIMMYT), El Batan 56130, Texcoco, Mexico

<sup>13</sup>National Maize Improvement Center, China Agricultural University (CAU)

1  
2  
3  
4 **ABSTRACT**  
5

6  
7 **Background**  
8  
9

10  
11 Characterization of genetic variations in maize has been challenging, mainly due to deterioration of  
12  
13 collinearity between individual genomes in the species. An international consortium of maize research  
14  
15 groups combined resources to develop the maize haplotype version 3 (HapMap 3), built from whole  
16  
17 genome sequencing data from 1,218 maize lines, covering pre-domestication and domesticated *Zea mays*  
18  
19 varieties across the world.  
20  
21

22  
23  
24 **Results**  
25

26  
27 A new computational pipeline was set up to process over 12 trillion bp of sequencing data, and a set of  
28  
29 population genetics filters were applied to identify over 83 million variant sites.  
30  
31

32  
33 **Conclusions**  
34  
35

36  
37 We identified polymorphisms in regions where collinearity is largely preserved in the maize species.  
38  
39 However, the fact that the B73 genome used as the reference only represents a fraction of all haplotypes  
40  
41 is still an important limiting factor.  
42  
43

44 **KEYWORDS**  
45

46  
47  
48 *Zea mays*, sequencing, haplotype map, genotyping, variant discovery, linkage disequilibrium, identity by  
49  
50 descent, imputation  
51  
52

53  
54 **BACKGROUND**  
55  
56

57  
58 Maize, one of the most important cereals in the world, also happens to be among the crop species with  
59  
60 the most genetic diversity. Advances in the next generation sequencing technologies made it possible to  
61  
62

1  
2  
3  
4 characterize genetic variations in maize at genomic scale. The previously released maize HapMap2 were  
5  
6 constructed with whole genome sequencing data of 104 maize lines across pre-domestication and  
7  
8 domesticated *Zea mays* varieties [1]. Since then, more maize lines have been sequenced by the  
9  
10 international research community, and a consortium was formed to develop the next generation  
11  
12 haplotype map. The maize HapMap 3 consortium includes, among others, BGI-Shenzen, Chinese Academy  
13  
14 of Agricultural Sciences, China Agricultural University (CAU), International Maize and Wheat Improvement  
15  
16 Center (CIMMYT). High-coverage data for 31 European and US Flint and Dent lines is also available in Ref.  
17  
18 [2]. Altogether, in this work we used a total of 1218 maize lines sequenced with depth varying from below  
19  
20  
21  
22  
23 1x to 59x.

24  
25  
26 A common approach in today's genetic diversity projects is to map the shotgun sequencing reads from  
27  
28 each individual onto a common reference genome to identify DNA sequence variations, and the physical  
29  
30 positions of the reference genome is used as a coordinate system for the polymorphic sites. A good  
31  
32 example is the human 1000 genome project [3]. The computational data processing pipeline developed  
33  
34 for the human project, GATK, has been widely adopted for identifying genetic variations in many other  
35  
36 species [4].  
37  
38  
39  
40

41 As the sequencing technology is improved and sequencers' base calling error model gets more accurate,  
42  
43 the computational challenges in genotyping by short-read sequencing have shifted from modeling  
44  
45 sequencer machine artifacts errors to resolving genotyping errors derived from incorrect mapping of short  
46  
47 reads to the reference genome. The problem is associated with the experimental design that uses the  
48  
49 single-reference genome as coordinate system. Taking maize as an example, the reference being used is  
50  
51 a 2.1 Gb assembly from an elite inbred line B73 that represents 91% of the B73 genome [5], and was  
52  
53 estimated to capture only ~70% of the low-copy gene fraction of all inbred lines [6]. The sequence  
54  
55 alignment software, however, can map 95-98% of the whole genome sequencing reads to the reference.  
56  
57  
58 That suggests a high percentage of the reads were mapped incorrectly, either being mapped to the  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 paralogous loci or highly repetitive regions under-represented in the reference assembly. The genetic  
5  
6 variants called from the miss-mapped reads need to be corrected computationally. The maize HapMap2  
7  
8 relied on linkage disequilibrium in the population to purge most of the bad markers caused by alignment  
9  
10 errors. To construct maize HapMap 3, a new computational pipeline was developed from scratch to  
11  
12 handle the sequencing data from 10 times more lines, and also took advantage of the high quality genetic  
13  
14 map constructed from the GBS technology [7, 8] which was not present when HapMap2 was constructed.  
15  
16

17  
18  
19 Genome structure variation in the population, including transposition, deletion, duplication and inversion  
20  
21 of the genomic segments, poses another challenge in the HapMap projects. As the physical genomes of  
22  
23 each of the individuals included in the HapMap projects vary both by size and structure, and there is no  
24  
25 co-linearity of all the sequence variants between the reference and genomes of each of the individuals, it  
26  
27 is not always possible to anchor all genetic variants in a population onto a single reference coordinate  
28  
29 system. As a compromise, markers included in the maize HapMap are defined as sites of which the  
30  
31 physical positions of the B73 alleles matching the markers' consensus genetic mapped positions.  
32  
33

34  
35  
36 Here we present maize haplotype map version 3 (HapMap 3), which is a result of coordinated efforts of  
37  
38 the international maize research community. The build includes 1,218 lines and over 83 million variant  
39  
40 sites anchored to the B73 reference genome version AGP v3.  
41  
42

## 43 44 45 **DATA DESCRIPTION**

46  
47  
48 The sequencing data used in this work is comprised of 12,497 billion base pairs in a total of 113,702 billion  
49  
50 Illumina paired-end reads, originating from 1,218 maize and teosinte lines. The data was collected from  
51  
52 several sources over several years, and varies in quality, read length, and coverage. Basic information  
53  
54 about various datasets and stages of the HapMap 3 project they were used in are summarized in Table 1.  
55  
56 Each of the 1,218 lines were sequenced at depth varying from below 1x to 59x, using reads of lengths  
57  
58 ranging from 44 through 201 bp, averaging 110 bp. All reads were aligned to maize reference genome B73  
59  
60  
61

version AGP v3 using BWA mem aligner [9]. Overall, 95-98% of the reads were mapped to the reference genome, although only about 50-60% with non-zero mapping quality.

Table 1: Sequence datasets used in various stages of HapMap 3

Dataset	# Taxa	Coverage per taxon			3.1.1	3.2.1unimp	3.2.1imp
		Minimum	Maximum	Average			
HapMap2	103	1	18.5	4.1	+	+	+
Hapmap2 extra	44	4.2	42	11.5	+	+	+
CAU	725	0.06	36.8	1.75	+	+	+
CIMMYT/BGI	89	1.1	19	11	+	+	+
282-2x	271	0	9	2.2	-	+	+
282-4x	270	0.6	34.5	4.4	-	+	-
German, Ref. [2]	31	8.3	59	17.4	-	+	+

Taxa from sets “HapMap2”, “HapMap2 extra”, and “CAU” partially overlap. The “282” libraries, sequenced twice represent 271 taxa. A “+” means that the dataset was used in a given stage, “-“ that it was not.

All sequence data used in this work is publically available. Collection and publishing of this data does not violate any local or international legislation or guidelines.

## ANALYSIS

### Initial variant discovery

The HapMap 3 pipeline is summarized in Figure 1. First, polymorphic sites were called for a set of 916 taxa from datasets HapMap2 through CIMMYT/BGI (7,191 billion base pairs, 74,643 million reads). In the first step, a custom built software tool was used to determine genotypes for each taxon at each site of the genome based on allelic depths at that site. Bases counted towards depth had base quality score of at least 10 and originated from reads with mapping quality ( $MAPQ = -\text{int}(10\log P)$ , where  $P$  – calculated by the BWA mem aligner - is the probability of the reported read location being wrong) at or above 30. This cut-off was chosen at mid-point between the highest MAPQ value reported by the aligner,

1  
2  
3  
4 corresponding to unambiguous alignments (60) and that of the most ambiguous ones (0). Analysis of the  
5  
6 inbreeding coefficient (Section HapMap 3.1.1) and of MAPQ distributions shows that our choice of cut-off  
7  
8 leads to decent quality genotypes while allowing for over 80% of alignments with  $MAPQ > 0$  to be included.  
9  
10 Only sites where at least 10 taxa had coverage of 1 or more were considered. Following Ref. [1], at each  
11  
12 site the allelic read depths were subject to segregation test (ST – see METHODS section for details). For a  
13  
14 population of inbred lines at true variant sites, one expects depths corresponding to minor and major  
15  
16 alleles to be concentrated in roughly different subset of taxa rather than being randomly distributed. The  
17  
18 purpose of the ST test is to find and eliminate sites for which allelic depth distribution appears random,  
19  
20 as such randomness, indicating high heterozygosity, is likely caused by alignment and sequencing errors.  
21  
22 A measure of the randomness is the p-value of the ST test (the smaller the p-value, the less random the  
23  
24 distribution). A p-value threshold of 0.01 was used in this study. This choice was somewhat arbitrary,  
25  
26 aimed at reducing the number of tentative variant sites to manageable size before further, more stringent  
27  
28 filters were applied. In this first, ST-based round of filtering, a total of 196 million tentative polymorphic  
29  
30 sites were selected. In the second step, these sites were filtered using the identity by descent (IBD)  
31  
32 information derived from about 0.5 million of high-quality polymorphisms obtained previously [8] using  
33  
34 the Genotyping-By-Sequencing (GBS) approach [7]. These GBS variants (GBS anchor) were used to  
35  
36 determine regions of IBD, where certain pairs of taxa are expected to have identical haplotypes. The  
37  
38 tentative polymorphic sites violating these IBD constraints were then filtered out, leaving 96.8 million  
39  
40 sites. At roughly half of the sites surviving this filter, minor allele was not present in taxa involved in the  
41  
42 tested IBD relationships. At such sites (typically with low minor allele frequency), the satisfied IBD  
43  
44 constraints do not confirm the existence of a variant. They are therefore less reliable and have been  
45  
46 marked with “IBD1” flag in the VCF files (see Table 2 for summary of flags and parameters present in  
47  
48 HapMap 3 VCF files). The ST- and IBD-filtered variant sites were then used in two separate procedures,  
49  
50 leading to two versions of HapMap 3 genotypes, referred to as HapMap 3.1.1 and HapMap 3.2.1.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 2: Flags and parameters used in INFO field of VCF files in various HapMap 3 versions.

Parameter	3.1.1	3.2.1 unimp	3.2.1 imp	Description
DP	+	+	+	Total read depth at the site
NZ	+	+	+	Number of taxa with called genotypes
AD	+	+	+	Allelic depths (reference, alternative in order listed in ALT field)
AC	+	+	+	Numbers of alternative alleles in order listed in ALT field
AQ	+	+	+	Average allele base qualities (reference, alternative in order listed in ALT field) computed in HapMap 3.1.1 from 916 taxa
GN	+	+	+	Numbers of genotypes (AA,AB,BB or AA,AB,AC,BB,BC,CC if 2 alt alleles present)
HT	+	+	+	Number of heterozygotes
EF	+	+	+	EF=heterozygosity/(presence_frequency*minor_allele_frequency); computed in HapMap 3.1.1 from 916 taxa
PV	+	+	+	p-value from segregation test, computed in HapMap 3.1.1. from 916 taxa
MAF	+	+	+	Minor allele frequency (summed up over all alternative alleles)
MAF0	-	-	+	Minor allele frequency in unimputed HapMap 3.2.1.
FH	+	-	-	Fraction of heterozygous taxa among the 506 taxa with more than 50% non-missing genotypes on chr 10
FH2	+			Site with FH greater than 2%
IBD1	+	+	+	only one allele present in IBD contrasts - based on 916 taxa of HapMap 3.1.1
LLD	+	+	+	Site in local LD with GBS map - based on 916 taxa of HapMap 3.1.1
NI5	+	+	+	Indel or site within 5 bp of a putative indel - from 916 taxa of HapMap 3.1.1
INHMP311	-	+	+	Site present in HapMap 3.1.1
ImpHomoAccuracy	-	-	+	Fraction of homozygotes imputed back into homozygotes
ImpMinorAccuracy	-	-	+	Fraction of minor allele homozygotes imputed back into minor allele homozygotes
DUP	-	-	+	Site with heterozygotes frequency > 3% - based on unimputed HapMap 3.2.1 genotypes

“+” and “-” indicate presence or absence, respectively, of a parameter or flag in a given version of HapMap. For example, “-+-” means the parameter is present in VCF file of both unimputed and imputed HapMap 3.2.1, and absent from HapMap 3.1.1. VCF files. Unless indicated otherwise, all parameters are computed from depths and genotypes in the current VCF file.

## HapMap 3.1.1

The HapMap 3.1.1 procedure involved checking for linkage disequilibrium of each site against the GBS anchor map [7, 8], which consists of markers located in hypo-methylated and genetically stable regions. Sites giving only very weak or only nonlocal (i.e., outside of 1 Mb radius) linkage Disequilibrium (LD) hits were eliminated, which resulted in the final set of 61,228,639 polymorphisms. For slightly less than 40% of these sites, LD could not be conclusively calculated due to small minor allele frequencies (MAF), whereas the remaining sites, confirmed to be in local LD with the GBS anchor, have been marked with flag “LLD”. Among the sites surviving all filtering steps, 8.7 million are indels or are located near (within 5 bp) of an indel. These have been marked with the flag “NI5”. Since a procedure to achieve consistent alignment across all reads covering the same indels - local realignment - is not computationally feasible at this scale and has not been performed, genotyping errors could occur, and, consequently, most such sites are tentative and should be treated with caution.

Figure 2 shows overlaps between various classes of variants of HapMap 3.1.1. First, we notice a rather small overlap between sites in confirmed local LD (“LLD” flag) and those marked “IBD1”. This is understandable, as the IBD1 sites represent mostly low MAF cases, where LD assessment could not be done. Indels and vicinity (labeled “NI5”) constitute about 15% of sites in each of the LLD, IBD1, and the union of LLD and IBD1 sets. Only a very small fraction of sites does not carry LLD or IBD1 flag, i.e., they are strongly confirmed by the IBD filter, but could not be classified with LD. The subset of 29.8 million sites in local LD and away from indels should be considered the most reliable.

To check the sensitivity of the obtained variant set to the mapping quality threshold imposed on the reads counted towards allelic depths, we repeated the pipeline using the mapping quality threshold equal to 1. Comparison of the variant set obtained this way (referred to as q1) with our recommended set (q30) is shown in Figure 3. While the overall number of variant sites is approximately independent of the mapping

1  
2  
3  
4 quality threshold, the two pipelines produce significantly different sets of sites, with only 72% of all q30  
5  
6 sites reproduced by the q1 pipeline. Closer inspection shows that this variability is due primarily to the  
7  
8 IBD1 sites, for which our filtering strategy was the least stringent. On the other hand, the LLD sites,  
9  
10 confirmed to be in local LD with GBS anchor, are much more independent of the mapping quality  
11  
12 threshold, which confirms high quality of such sites.  
13  
14

15  
16  
17 For a population of inbred lines considered here, insight into genotype quality may be obtained from the  
18  
19 inbreeding coefficient, calculated here for each taxon using the VCFtools program [11] from the formula  
20  
21

$$F_{inbr} = (O - E)/(N - E)$$

22  
23  
24  
25 where  $O$  is the observed number of homozygotes for a given taxon,  $N$  is the number of sites at which the  
26  
27 taxon was genotyped, and  $E$  is the expected number of homozygotes given by  $E = \sum_k(1 - 2p_kq_k)$ .  
28  
29 Summation in the latter formula runs over  $N$  genotyped sites,  $p_k$  is the minor allele frequency at site  $k$   
30  
31 (computed from all taxa in the population with non-missing genotypes at this site), and  $q_k = (1 - p_k)$ .  
32  
33 Low values of the inbreeding coefficient, indicating high heterozygosity, are mostly due to genotyping  
34  
35 errors. Importance of choosing a sufficiently tight mapping quality threshold for the quality of genotypes  
36  
37 is apparent from Figure 4, where the distribution of inbreeding coefficient for chromosome 10 is shown  
38  
39 for q1 and q30 variant sets. The lower MAPQ threshold results in a large number of miss-mapped reads  
40  
41 being counted towards depth, producing overly heterozygous genotypes, especially for highly covered  
42  
43 taxa (the peak below 0.8 is due mostly to CIMMYT lines with 10-15x coverage; these lines have higher  
44  
45 heterozygosity than other lines which may also contribute to the peak) and thus shifting the curve to the  
46  
47 left. Since most HapMap 3 taxa are inbred lines, one should expect the true distribution to be contained  
48  
49 within peak around 0.95. In view of this, the q30 result is definitely an improvement over q1, although a  
50  
51 longer than expected tail extending towards the value 0.8 indicates that the HapMap 3 variants may  
52  
53 contain too many false heterozygotes.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 Seemingly heterozygous sites may result from either sequencing errors or misalignments of reads  
5  
6 originating from paralogous regions. To investigate this further, we calculated, for each site, the fraction  
7  
8 of heterozygous HapMap 3.1.1 genotypes within a subset of 506 high-coverage taxa (defined as those  
9  
10 with more than 50% non-missing genotypes on chromosome 10). In HapMap 3.1.1 VCF files, this fraction  
11  
12 has been recorded as parameter “FH”. At sites for which this parameter exceeds 2-3%, heterozygotes are  
13  
14 likely to originate from misalignments, for example, from tandem and ectopic duplications. Such sites  
15  
16 constitute 9% of all HapMap 3.1.1 sites.  
17  
18  
19  
20

## 21 HapMap 3.2.1

22  
23  
24 The 96.8 million ST- and IBD-filtered variant sites were the starting point for the HapMap 3.2.1 procedure  
25  
26 (Figure 1). On these sites, genotypes were called on the 263 taxa from the “282” panel of Ref. [10] using  
27  
28 “282 2x” dataset, and on the 31 high-coverage (on average 17 x) “German” taxa [2], for the total of 1210  
29  
30 taxa. Some of the taxa present in the “282” and “German” sets carry the same names as the ones included  
31  
32 in the 916-taxon HapMap 3.1.1 set. Since despite identical names such taxa often originate from different  
33  
34 germplasm sources, they have been kept separate during genotyping, i.e., reads from different sources  
35  
36 were not merged and separate genotypes were computed for each source. In the resulting VCF files, the  
37  
38 names of the overlapping taxa have been prefixed by “282set\_” and “german\_”. For example, in the case  
39  
40 of B73, there are three columns representing different datasets for this taxon: “B73” (the original 916-  
41  
42 taxa set), “282set\_B73” (sequence from the more recent “282” libraries), and “german\_B73” (from Ref.  
43  
44 [2]).  
45  
46  
47  
48  
49  
50

51 To further eliminate the false positives resulting from sequencing errors, an additional depth-based filter  
52  
53 was applied to the 96.8 million sites. Referred to as “>1,>2” filter it accepts sites for which the read support  
54  
55 of minor allele was greater than 1 in at least one taxon and greater than 2 across all taxa. Genotypes on  
56  
57 the surviving 83,153,144 sites, referred to as “un-imputed HapMap 3.2.1”, were then processed through  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 the LD KNN imputation procedure based on Ref. [12], where the “nearest neighbors” of a given line are  
5  
6 selected based on sites in good local LD with the target site. Whenever possible, the procedure filled up  
7  
8 missing genotypes with imputed ones, but the non-missing genotypes were left unchanged, even if  
9  
10 imputation classified them differently. Non-imputable missing genotypes at the sites with (pre-  
11  
12 imputation) MAF below 1% were assumed to be major allele homozygotes. Imputation reduces the  
13  
14 fraction of missing genotypes from 50% to 7%. Most of the originally missing genotypes (about 85%) are  
15  
16 imputed to major allele homozygotes. Accuracy of the genotype dataset can be assessed by comparing  
17  
18 the original genotypes with imputed ones. As shown in Table 3, 99.8% of major allele homozygotes are  
19  
20 imputed back into the same class. While the accuracies of minor allele homozygotes and genotypes  
21  
22 including indels are both above 90%, only 11% of heterozygotes are imputed back into the same class,  
23  
24 while 47% of them fail imputation altogether. This reflects the inherent difficulty in calling heterozygotes.  
25  
26 In the single-reference approach to maize genotyping employed here, heterozygous sites represent true  
27  
28 residual heterozygosity as well as misalignments of reads from tandem and ectopic duplications. Since  
29  
30 residual heterozygosity in our population of predominantly inbred lines should not exceed 2-3%, all  
31  
32 heterozygotes with frequency  $\geq 3\%$  can be considered a result of misalignments. About 10% of all  
33  
34 heterozygotes present in HapMap 3.2.1 set satisfy this condition. In the VCF files, these sites have been  
35  
36 flagged with flag DUP (“duplicated regions”). Other parameters generated by the imputation procedure  
37  
38 and recorded for each variant site in the INFO field are ImpHomoAccuracy fraction of all homozygotes  
39  
40 imputed back into homozygotes and ImpMinorAccuracy fraction of minor allele homozygotes imputed  
41  
42 back to the same class. The INFO field also contains flags IBD1, LLD, and NI5, computed from the initial  
43  
44 916 taxa in the HapMap 3.1.1 procedure. Genotypes resulting from the imputation procedure are referred  
45  
46 to as “imputed HapMap 3.2.1”.

47  
48 Table 3: Accuracy of various genotype classes based on statistics from imputation in HapMap 3.2.1  
49  
50  
51  
52  
53  
54  
55  
56

Genotype class	Accuracy within class [%]	% unimputed
Major allele homozygote	99.8	1.2
Heterozygote	11.1	47.0
Minor allele homozygote	94.4	14.2
Indel	92.2	17.3

Accuracy computed as percentage of the original number of genotypes in a given class (excluding genotypes that could not be imputed) imputed into the same class. The last column shows the fraction of genotypes within a class which could not be imputed.

Relationship between variant sites included in HapMap 3.1.1 and 3.2.1 is shown in Figure 5. Both pipelines start from the same set of IBD-filtered genotypes and subject them to different kinds of filtering, with that of HapMap 3.1.1 being more stringent. It is therefore not surprising that HapMap 3.2.1 recovers the majority (86%) of HapMap 3.1.1 sites, including over 99% of those flagged LLD (i.e., confirmed in local LD). In addition, 30.3 million extra sites are retained in HapMap 3.2.1, which failed the LD filter in HapMap 3.1.1 pipeline. On the other hand, the depth-based “>1,>2” filter applied in HapMap 3.2.1 eliminated 8.2 million sites present in HapMap 3.1.1, including about 0.2 million LLD ones.

After the HapMap 3.2.1 release was completed, “282-2x” sequencing data became available for additional 8 taxa from the “282” panel. Libraries for all 271 taxa were also re-sequenced at a higher depth (average of about 4.4x), leading to another dataset, “282-4x” (as this re-sequencing failed for one of the taxa, this dataset only contained 270 taxa). Therefore, the un-imputed HapMap 3.2.1 genotypes for all 271-taxa from the “282” panel were re-called using the full available sequencing depth, creating a separate variant dataset for the “282” panel.

## DISCUSSION

The maize genome, 2.3 GB in size [5], is smaller than the human genome. But some of its distinctive features makes it more challenging for variants identification. First, a recent whole genome duplication occurred 12 million years ago resulted in homologous segments that complicate the short read

1  
2  
3  
4 alignments; second, the rampant activities of transposable elements within last 1-5 million year not only  
5  
6 resulted in accumulation of large amount of relatively young repetitive elements in the intergenic regions,  
7  
8 but also extraordinary structural variations within species [5, 6]. In this study, the genome of the B73  
9  
10 maize line was used as the reference for variant calling from short sequencing reads. Structural variations  
11  
12 between B73 and other individuals has been the major challenge for identification of true variants. In  
13  
14 particular, short reads derived from regions missing in the reference genome could be mismatched to  
15  
16 other paralogous regions, which lead to false positive genotypes. In the human 1000 genome project, a  
17  
18 new HaplotypeCaller was used [4], which performs local *de novo* assembly to identify the most likely  
19  
20 haplotypes for each individual and thus improve the genotyping results. However, HaplotypeCaller is  
21  
22 computationally very expensive, and not always applicable in species like maize, where the single  
23  
24 reference genome misses many haplotypes presents in the species and has a lot more mismatched  
25  
26 paralogous reads that would disrupt the local assembly. To filter out these false positive variants called  
27  
28 from the mismatched reads, we relied on the *Zea* GBS map [7, 8], which was obtained from GBS markers  
29  
30 located primarily in hypo-methylated chromosomal regions. GBS maps were used to identify IBD regions  
31  
32 between the individual genomes, and 100 million markers with high percentage of mismatched genotype  
33  
34 calls in the IBD regions were filtered out from the initial set of 196 million markers. The highly repetitive  
35  
36 genomic regions derived from recent transposition activities are in general easier to identify, because the  
37  
38 templates of these repeats are well represented on the reference genome, and sequencing reads mapped  
39  
40 to these regions, flagged with low mapping quality, can be removed at the early stage of the analysis  
41  
42 pipeline. For HapMap 3, reads with mapping quality lower than 30 were not included in the build.  
43  
44  
45  
46  
47  
48  
49  
50  
51

52 One of the goals of HapMap 3.1.1 is to identify genetic markers in regions where collinearity is preserved  
53  
54 in majority of maize lines. The LD filter in the pipeline was applied for this purpose. To do this, we  
55  
56 genetically mapped the presence/absence of the minor alleles using the GBS genetic map, and these  
57  
58 mapped genetic positions were compared to the physical positions on the B73 reference. Among the 96.8  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 million sites surviving the IBD filter, 25% did not have enough non-missing data or sufficient minor allele  
5  
6 frequency for genetic mapping to be meaningful. For 38% of sites, at least one genetically mapped  
7  
8 position matching the physical positions on B73 reference was found, 24% have no significant hits from  
9  
10 genetic mapping, probably due to no consensus positions in the HapMap 3 population, and 13% have  
11  
12 genetic positions not matching the B73 physical positions. Markers from the latter two categories (37% of  
13  
14 all IBD-filtered markers) were removed by the LD filter, leaving slightly over 61 million sites, about 60% of  
15  
16 which were confirmed in local LD and marked with a flag “LLD” in VCF files.  
17  
18  
19  
20

21 The IBD and LD filters applied in the HapMap 3.1.1 project effectively remove majority of the false positive  
22  
23 genetic variants caused by paralogous genomic regions, as well as markers with lost collinearity between  
24  
25 the species. However, not all the genotyping errors have been removed from the release. 23,839,286 of  
26  
27 the sites do not have sufficient minor allele frequency for genetic test (these are missing the “LLD” label  
28  
29 in the INFO field of the VCF files). Another source of errors are paralogous regions evolved from tandem  
30  
31 duplications. Misalignments of reads from such regions result in false heterozygous genotypes with  
32  
33 relatively high frequency and in local LD, and therefore difficult to filter out. Given enough sequencing  
34  
35 depth, the tandem duplications can be identified either as copy number variation or imputation errors.  
36  
37 However, majority of the HapMap 3 lines have very low sequencing depth, and fail to sample all  
38  
39 paralogous loci or all alleles, which makes it difficult to flag all sites complicated by tandem duplications.  
40  
41  
42  
43  
44

45 Local LD filter based on a large, diverse population may be too stringent, as some markers, good within  
46  
47 certain sub-populations, may be thrown out. Therefore, the LD filter was not used in the HapMap 3.2.1  
48  
49 release, which contains a total of 83 million variant sites, subject only to ST and IBD filters and an  
50  
51 additional depth-based filter aimed to improve reliability of rare allele calls. Although those sites are likely  
52  
53 to have higher misalignment rates, they are still likely to capture a true association with phenotypes.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 In the un-imputed HapMap 3.2.1, at about 10% of all variant sites, fraction of heterozygous taxa exceeds  
5  
6 3%. Such sites are marked “DUP”, as most likely originating from duplication misalignments. Figure 6  
7  
8 shows the distribution of fraction of heterozygous sites per taxon for different versions of HapMap 3.2.1  
9  
10 release. While for the un-imputed genotypes the distribution peaks slightly below 1%, imputation  
11  
12 significantly shifts the peak to the left, down to about 0.5%. This is a consequence of most missing  
13  
14 genotypes being imputed to homozygotes. Interestingly, considering only sites in good local LD (marked  
15  
16 with the “LLD” flag) leads to distributions (both imputed and un-imputed) shifted towards higher  
17  
18 heterozygosities. This is understandable, as the LLD sites are typically those with higher minor allele  
19  
20 frequencies, where the chance of encountering a heterozygote is higher.  
21  
22  
23  
24  
25

26 In summary, besides the addition of more maize lines, the HapMap 3.2.1 release differs from the 3.1.1  
27  
28 release in three major aspects: 1) Improved rare allele calls: to increase the accuracy of the variants with  
29  
30 rare allele, the HapMap 3.2.1 pipeline applied more stringent read depth thresholds instead of the  
31  
32 population genetics based LD filter that could not be applied to sites with very low MAF; 2) The sites with  
33  
34 high percentage of heterozygous calls were flagged in the VCF files; 3) Missing data was imputed using  
35  
36 the LD KNN method. As summarized in Table 2, the VCF files of both datasets contain labels that flag the  
37  
38 characteristics of each of the sites. To effectively use this resource, it is recommended to filter the sites  
39  
40 based on the flags that are appropriate to the purpose of each project.  
41  
42  
43  
44  
45

46 When constructing the maize HapMap 3, the most serious problems we were facing can be attributed to  
47  
48 the use of a genome from a single individual (B73) as a reference for other, often very different species.  
49  
50 This is becoming the single limiting factor in the study of maize diversity, as well as breeding practice. The  
51  
52 only remedy is to move away from a single genome-based reference coordinate and adopt a pan-genome  
53  
54 based reference system that incorporates all major haplotypes of the species.  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## METHODS

### Plant material

Plant material used in this study was obtained mostly from maize inbred lines representing wide range of *Zea mays* diversity. 103 of these lines, used previously in the HapMap2 project [1], include 60 improved lines, including the parents of the maize nested association mapping (NAM) population [13], 23 maize landraces and 19 wild relatives (teosinte lines: 17 *Z. mays ssp. parviglumis* and 2 *Z. mays ssp. mexicana*). Sequence datasets originating from these lines are referred to in Table 1 as “HapMap2” and “HapMap2 extra”. Majority of the remaining inbred lines originated from CAU (sequence dataset “CAU”) and include, among others, “Chinese NAM” parent lines. Additional 89 inbred lines were provided by CIMMYT and sequenced at BGI (dataset “CIMMYT/BGI”). The HapMap 3 population also contained one *Tripsacum* line (TDD39103), one “mini-maize” line (MM-1A), and a few newly sequenced landraces. Overall, the number of taxa in the initial, variant-discovery stages of the HapMap 3.1.1 project was 916.

The sequence of 271 taxa from the libraries of the “282” panel [10] were added at a later stage (HapMap 3.2.1). DNA to construct these libraries was obtained from the collection that the North Central Regional Plant Introduction Station (NCRPIS) distributes all over the world. Additionally, the high-coverage data of Ref. [2], originating from 31 European and US inbreds was also included. The total number of taxa genotyped in the HapMap 3.2.1 build is 1218.

In this study, individuals with the same taxa name but contributed by different members of the consortium were kept as separate entries in genotyping pipeline - a decision prompted by comparison of genotypes obtained from different datasets. For example, the newly sequenced CML103 is significantly more heterozygous from CML103 studied previously in the HapMap2 project. Also, the Mo17 sequence originating at CAU has been treated as taxon separate from Mo17 and CAUMo17. In most of those cases,

1  
2  
3  
4 a prefix or suffix indicating the origin of the sequence data has been added to the taxa name (e.g.,  
5  
6 "282set\_" or "german\_", "-chin").  
7  
8

## 9 10 Sequencing

11  
12 Sequencing has been performed over several years using various generations of Solexa/Illumina  
13  
14 instruments and library preparation protocols, giving paired end reads from 44 to 201 bp long. Overall,  
15  
16 113.7 billion reads were obtained on 1,218 lines, containing 12,497 billion base pairs, giving on average  
17  
18 4.4x coverage per line (assuming 2.3 Gb genome size). However, as shown in Table 1, coverage was not  
19  
20 uniform among all lines. For a few lines, sequence generated previously in the context of HapMap2 project  
21  
22 was augmented with reads from recent re-sequencing which brought the median coverage of the  
23  
24 HapMap2 lines to 5x, with average coverage equal to 7.8x and standard deviation of 7.2x. All NAM parent  
25  
26 lines are covered to 10x or higher. Most of the 89 lines provided by CIMMYT and sequenced at BGI have  
27  
28 coverage exceeding 10x. The recent re-sequencing of the "282" panel resulted in coverage between 1.7  
29  
30 and 36x, averaging 6.5x. Coverage of the 31 "German" lines for Ref. [2] ranges from 8.3 to 59x, with  
31  
32 average of 17.4x. Majority of the inbred lines originated from CAU have been sequenced at a lower  
33  
34 coverage (1-2x). The list of all lines used in HapMap 3 with the corresponding coverage is given in  
35  
36  
37  
38  
39  
40  
41 Additional file 1.  
42  
43

## 44 Alignment

45  
46 Due to the use of different versions of Solexa/Illumina sequencing equipment, the base qualities in  
47  
48 different input FASTQ files are given in different encodings. Prior to alignment, all base qualities have been  
49  
50 converted to phred+33 scale. Reads were then aligned to B73 reference (AGP v3) as paired-end using bwa  
51  
52 mem aligner (1) with default options. In 72 read sets (Illumina lanes), for technical reasons a high (6%-  
53  
54 54%) fraction of paired-end fragments was found to be shorter than reads, so that the two ends contained  
55  
56 a part of Illumina adapter and were reverse complements of each other. For such "read-through"  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 fragments, the remnants of Illumina adapter sequences were clipped using TRIMMOMATIC [14] and only  
5  
6 one read was used and aligned as single-end. The bwa mem aligner is capable of clipping the ends of reads  
7  
8 and splitting each read in an attempt to map its different parts to different location on the reference. As  
9  
10 a result, typically over 95% of reads are reported as mapped. However, the fraction of reads with non-  
11  
12 zero mapping quality (negative log of the probability that a read has been placed in a wrong location) is  
13  
14 much lower – typically only 40-50%. Figure 7 shows a typical distribution of the mapping quality obtained  
15  
16 from bwa mem alignment. In practice, we only used alignments with mapping quality of at least 30. A  
17  
18 base was counted towards allele depth if its base quality score was at least 10.  
19  
20  
21  
22

23  
24 It is well known that alignment may be especially ambiguous when reads contain indels with respect to  
25  
26 the reference. In such cases, multiple-sequence realignment approaches have been proposed [4] to find  
27  
28 the correct sequence and location of an indel and avoid spurious flanking SNPs. Since indels are not the  
29  
30 primary focus of this work and since the realignment is computationally very expensive, it has not been  
31  
32 performed by the HapMap 3 pipeline. Thus, although indels and SNPs in their immediate vicinity have  
33  
34 been retained in the HapMap 3 VCF files, they are less reliable and have therefore been marked with “NI5”  
35  
36 label for easy filtering.  
37  
38  
39  
40

## 41 Genotyping pipeline

42  
43  
44 Raw genotypes were obtained using a custom-built multi-threaded java code. First, the code executes  
45  
46 samtools mpileup command (thresholds on the base and mapping quality are imposed here) for each  
47  
48 taxon individually, processing a certain portion of the genome. On a multi-core machine, several such  
49  
50 pileup processes (i.e., for several taxa) can be run concurrently as separate threads. Since we are  
51  
52 predominantly interested in calling SNPs, we use a simplified indel representation where insertions and  
53  
54 deletions with respect to reference are treated as additional alleles “I” and “D”, respectively, regardless  
55  
56 of length and actual sequence of the indel. Read depths and average base qualities of all six alleles (A, C,  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 G, T, I, and D) are extracted from samtools mpileup output for each taxon at each genomic position and  
5  
6 stored in an array shared between all threads. The amount of memory available on the machine along  
7  
8 with the number of taxa determine the upper limit on the size of this array, and therefore – the maximum  
9  
10 size of chromosome chunk which can be processed at one time. As base quality of I and D alleles we took  
11  
12 the value corresponding to the base directly preceding the indel on the reference.  
13  
14

15  
16 Extraction of allelic depths for all genomic positions is time consuming, which presents a major obstacle  
17  
18 if joint genotyping needs to be re-run, for example, upon extending the taxa set. It is therefore  
19  
20 advantageous to run the depth extraction only once for each taxon and save the obtained depths on disk  
21  
22 to be retrieved (rather than re-calculated) during the genotyping step. This way, when the taxa set for  
23  
24 genotyping is extended, mpileup step has to be run only for the newly added taxa. Thus, the program  
25  
26 features an option to save allelic depths and average qualities in specially designed data structures stored  
27  
28 in HDF5 files – one such file per taxon per chromosome. To save space, each allele depth and average  
29  
30 quality is stored as one byte, which allows exact representation of integers from 0 to 182, while higher  
31  
32 integers (up to about 10,000) are represented approximately by negative byte values through a  
33  
34 logarithmic formula with carefully chosen base. Depths and qualities are stored only for sites with non-  
35  
36 zero coverage. The details of the storage format and integer representation in terms of byte variables are  
37  
38 given in Additional file 2.  
39  
40  
41  
42  
43  
44

45  
46 Once the allelic depths for all taxa and a given chunk of the genome are available in shared memory, each  
47  
48 site is evaluated for presence of a tentative SNP. On a multi-core machine, the set of sites within the  
49  
50 genome chunk is divided into subsets processed in parallel on different cores. Sites with less than 10 taxa  
51  
52 with read coverage and those with only reference allele present are ignored. For all other sites, genotypes  
53  
54 are called for all taxa using a simple likelihood model with a uniform error rate [15] assumed at 1%.  
55  
56 Alternative alleles are then sorted according to their allele frequencies and up to two most abundant  
57  
58 alleles are kept, as decided by the segregation test described in the next Section. Sites for which all taxa  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 turn out to be reference homozygotes (which may happen despite non-reference alleles being present in  
5 the mapped reads) are skipped. Raw variant set obtained in this way is then subject to extensive filtering  
6  
7 with the intention of reducing the number of false positives resulting from misalignments.  
8  
9

## 10 11 Filtering

### 12 Segregation test (ST) filter

13  
14  
15 For each pair of alleles obtained in the genotyping step, a 2 by N (where N is the number of taxa)  
16  
17 contingency table is constructed, containing depths of the first allele in row 1 and depths of the second  
18  
19 allele in row 2. The Fisher exact test (FET) is then performed to assess how likely such a table is to occur  
20  
21 by chance. If the expected values of the array elements are sufficiently large, the p-value from FET is  
22  
23 approximated by that from the computationally efficient chi-square test. However, in most cases  
24  
25 encountered here, expensive simulation is needed to obtain sufficiently accurate p-value. To reduce  
26  
27 computational burden, we adopted a hybrid approach based on an empirical observation that for  
28  
29 statistically insignificant cases (p-values larger than 0.2) the chi-square test results in a de facto lower  
30  
31 bound to exact p-values. Thus, the chi-square test is performed first for each site and if the p-value from  
32  
33 this test is below 0.2, more exact p-value is obtained from a simulation procedure. The simulation  
34  
35 procedure used here, implemented in Java, is the same as the one implemented in R package [16]. An  
36  
37 alternative allele is kept if at least one contingency table involving this allele has p-value smaller or equal  
38  
39 to 0.01. If none of the alternative alleles survive the ST filter, the site is skipped (not reported in output).  
40  
41  
42 The ST filter tends to eliminate variant sites resulting from random sequencing errors.  
43  
44  
45  
46  
47  
48  
49  
50

### 51 GBS anchor map and IBD filter

52  
53  
54 Given a set of trustworthy SNPs and a diverse set of 916 taxa it is possible to identify, for an arbitrary  
55  
56 region of the genome, a number of taxa pairs which are identical by descent (IBD) and are therefore  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 expected to have identical genotypes in this region. If known, these IBD pairs can be used as a powerful  
5  
6 filter eliminating variant which violate IBD constraints.  
7  
8

9  
10 To determine the IBD regions, we used the first step of our pipeline to call genotypes for our 916 taxa on  
11  
12 the set of GBS v2.7 sites [7, 8] which tend to concentrate in relatively well-conserved low-copy regions of  
13  
14 the genome and can therefore be considered reliable. This set of 954,384 sites was filtered to include only  
15  
16 SNP (not indel) sites for which the p-value from the segregation test was below 0.05 and which were more  
17  
18 than 5 bp away from any indel. The set of genotypes at 475,272 sites obtained in this way, which will be  
19  
20 referred to as GBS anchor, agree well with those from GBS on 167 taxa present in both sets. Alleles  
21  
22 detected by the HapMap 3 pipeline agreed with those from GBS at 94% of the GBS sites. At 90% of the  
23  
24 sites, fraction of (non-missing data) taxa with genotypes in agreement with those from GBS was at or  
25  
26 above 85%. Genotypes different from GBS ones were observed for 82 taxa. These differences were most  
27  
28 frequent (up to 19% of all sites) for teosinte lines.  
29  
30  
31  
32

33  
34 The GBS anchor was used to compute the genetic distance (identity by state) between any two of the 916  
35  
36 lines in windows containing 2000 GBS sites each (about 8.5 Mbp on average). If the genetic distance within  
37  
38 such a window was  $\leq 0.02$  (about 10 times smaller than the mean distance across all pairs), the two lines  
39  
40 were considered to be in IBD. At least 200 comparable GBS sites (i.e., non-missing data simultaneously on  
41  
42 both lines being compared) were assumed necessary to make the genetic distance calculation feasible.  
43  
44 This allowed for good distance estimate while keeping the number of detected IBD relationships large.  
45  
46  
47

48  
49 The number of taxa involved in IBD relationships in any given window were between 385 (start of  
50  
51 chromosome 10) and 757 (middle of chromosome 7) and averaged 588, leading to large numbers of IBD  
52  
53 contrasts, ranging from 3,710 (beginning of chromosome 4) to 42,890 (middle of chromosome 7), and  
54  
55 averaging 13,500.  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 The tentative (ST-filtered) variant sites were confronted with the IBD information as follows: for each site,  
5  
6 pairs of lines in IBD were determined as described above. Genotypes of IBD-related lines were compared  
7  
8 and the numbers of allele matches and mismatches, summed over all IBD pairs, were counted for each  
9  
10 allele present at the site. If the match/mismatch ratio was at least 2 for at least two alleles, or if only one  
11  
12 allele was present in all IBD contrasts, the site is considered as passing the IBD filter. Such a filter is less  
13  
14 powerful for sites where all bases in IBD lines are major allele homozygotes, i.e., the variant being  
15  
16 evaluated occurs in lines not involved in IBD pairs. Formally, such a site passes our IBD filter, but the actual  
17  
18 variant is not strongly confirmed. These uncertain sites, mostly with low minor allele frequency, are  
19  
20 labeled "IBD1" in the HapMap 3 VCF files and constitute about 50% of all HapMap 3 sites.  
21  
22  
23  
24  
25

#### 26 Linkage Disequilibrium (LD) filter

27  
28 Any true SNP should be in local linkage with other nearby SNPs. This observation is the origin of another  
29  
30 filter used in this work, referred to as the LD filter. For each variable site surviving the ST and IBD filters,  
31  
32 we evaluated LD with each site of the GBS anchor. As the LD measure we chose the p-value from a 2 by 2  
33  
34 contingency table of haplotype counts AB, Ab, aB, ab. For the purpose counting haplotypes, heterozygous  
35  
36 genotypes were treated as homozygous in minor allele, so that each taxon only contributed at most one  
37  
38 haplotype. This tends to somewhat strengthen the LD signal and simplify the calculation. For a pair of sites  
39  
40 to be tested for LD, the following three conditions had to be satisfied to make the calculation meaningful:  
41  
42  
43 i) the two sites were at least 2,500 bp apart, ii) there were at least 40 taxa with non-missing genotypes at  
44  
45 both sites being compared, and iii) at least 2 taxa with minor allele had to be present at each of the two  
46  
47 sites.  
48  
49  
50  
51  
52

53 Filtering procedure executed for each site is summarized in Figure 8. First, LD between the given site and  
54  
55 all sites in GBS anchor was computed and up to 20 best LD hits (the ones with lowest p-values) were  
56  
57 collected. If the p-value of the best hit exceeded  $1E-6$  (which roughly corresponds to the peak of the  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 overall distribution of p-values), the site was rejected. Otherwise, it was determined whether the set of  
5  
6 best hits contained any local hits, i.e., hits to GBS sites on the same chromosome within 1 Mbp of the site  
7  
8 in question and with the p-value smaller than 10 times the p-value of the best hit. If no such local hits  
9  
10 were found, the site was rejected, otherwise it was kept and marked as a site in Local LD using the flag  
11  
12 “LLD”. Note that the procedure defined this way filters out sites with only non-local LD hits as well as  
13  
14 those with only weak LD signal. Sites in local LD as well as those for which LD could not be assessed  
15  
16 (because of low minor allele frequency or missing data) pass the filter.  
17  
18  
19  
20

## 21 Imputation

22  
23  
24 In the HapMap 3.2.1 pipeline, the ST- and IBD-filtered genotypes, after the application of the additional  
25  
26 “>1,>2” depth-based filter, were processed through the LD KNN imputation procedure based on Ref. [12]  
27  
28 to fill in the missing data. The procedure is a version of the “K nearest neighbors” routine where the  
29  
30 “nearest neighbors” of a given taxon are selected based on genetic distance computed using variant sites  
31  
32 in good local LD. Specifically, for a given target site, a list of up to 70 sites in best LD (as given by the  $R^2$   
33  
34 measure) with it is compiled by checking all surrounding sites within 600Kb characterized by  
35  
36 heterozygosity lower than 3% and more than 50% taxa with non-missing genotypes. Capping this list at  
37  
38 70 sites leads to good compromise between distance accuracy and computation speed. Then, at the same  
39  
40 target site, for each target taxon, up to 30 “nearest neighbor” taxa are selected, with lowest genetic  
41  
42 distances from the target taxon. Genetic distances are computed using the set of local LD sites selected  
43  
44 in the previous step. Taxa with more than 50% missing genotypes at LD sites, missing genotype at the  
45  
46 target position, having distance from the current taxon larger than 0.1, or resulting in less than 10 common  
47  
48 LD sites on which the distance can be calculated, are excluded from distance calculation process.  
49  
50 Genotypes of the selected nearest neighbor taxa at the target site are stored in memory along with the  
51  
52 genetic distances from the target taxon. This information is used to compute a weight  $w_i$  of each neighbor  
53  
54 genotype  $g$  as follows:  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

$$w_g = \sum_i \frac{1}{1 + 70d_{gi}},$$

where the summation index  $i$  runs over all neighboring taxa with genotype  $g$  at the target site, and  $d_{gi}$  is the distance of taxon  $i$  from the target taxon. The genotype with the highest weight is considered the imputed genotype (of the target taxon at the target site) provided its weight is at least 10 times larger than that of the second-best candidate genotype. Otherwise the imputation is considered inconclusive and the imputed genotype is set to “unknown” (missing data), as it is in the case when no close neighbors of the current taxon could be found. If a genotype imputed to “unknown” occurs at a site where  $MAF < 1\%$ , it is automatically converted into major allele homozygote.

The imputation procedure is run for each genotype in the input file. However, in the output only the originally missing genotypes are updated to imputed ones, whereas all others are left unchanged, even if classified differently. On the other hand, all imputed genotypes are used during a run to collect imputation statistics. The “transition matrix” showing how many genotypes originally in a given class were imputed into other classes is an indication of the accuracy of the input genotypes. Error rates calculated from the transition data are given in Table 3.

## AVAILABILITY OF DATA

At present, reads from all datasets is available in the form of BAM files (with reads aligned to AGP v3 reference) on CYVERSE data store (formerly iPlant, <http://www.cyverse.org/data-store>), in directories `/iplant/home/shared/panzea/raw_seq_282/bam` (dataset “282-2x” and “282-4x”, also available from NCBI BioProject PRJNA389800), `/iplant/home/shared/panzea/hapmap3/bam_germanlines` (“German” dataset; raw reads available from NCBI BioProject PRJNA260788), `/iplant/home/shared/panzea/hapmap2/bam` (“hapmap2” dataset; raw reads available from NCBI BioProject SRP011907), and

1  
2  
3  
4 /iplant/home/shared/panzea/hapmap3/bam (dataset “hmp3.1.1”, also available from NCBI  
5  
6 BioProject PRJNA399729).

7  
8  
9  
10 The set of HapMap 3.1.1. polymorphisms determined for 916 taxa (from datasets “HapMap2”,  
11  
12 “HapMap2 extra”, “CAU”, and “CIMMYT/BGI”) is available in VCF format on CYVERSE data store in the  
13  
14 directory /iplant/home/shared/panzea/hapmap3/hmp311, in files

15  
16  
17 c\*\_hmp311\_q30.vcf.gz (one file per chromosome, where “\*” stands for chromosome 1-10).

18  
19 Additionally, files c\*\_hmp311\_q1.vcf.gz (in the same location) contain test results obtained with  
20  
21 mapping quality threshold equal to 1.

22  
23  
24  
25 The HapMap 3.2.1 variants for 1210 taxa (916 initial Hapmap 3.1.1 taxa + 263 taxa from “282-2x” set +  
26  
27 31 “German” lines) are available from

28  
29 /iplant/home/shared/panzea/hapmap3/hmp321/unimputed, in files

30  
31  
32 mergedflt\_c\*.vcf.gz (un-imputed results) and in

33  
34 /iplant/home/shared/panzea/hapmap3/hmp321/imputed, in files

35  
36  
37 mergedflt\_c\*.imputed.vcf.gz (imputed results).

38  
39  
40 Files c\*\_282\_corrected\_onHmp321.vcf.gz in CYVERSE directory

41  
42 /iplant/home/shared/panzea/hapmap3/hmp321/unimputed/282\_libs\_2015 contain  
43  
44 un-imputed genotypes on HapMap 3.2.1 sites from the full depth data available for the “282” panel (271  
45  
46 taxa, datasets “282-2x” + “282-4x”).

## 47 48 49 50 51 52 53 LIST OF ABBREVIATIONS

54  
55  
56 International Maize and Wheat Improvement Center (CIMMYT), Segregation test (ST), Identity-by-descent  
57  
58 (IBD), Genotyping-by-sequencing (GBS), Linkage Disequilibrium (LD), Minor Allele Frequency (MAF),  
59

1  
2  
3  
4 Nested Association Mapping (NAM), Single Nucleotide Polymorphism (SNP), Insertion-Deletion (Indel),  
5  
6 read mapping quality (MAPQ), base quality (BQ), linkage disequilibrium-base K nearest neighbor  
7  
8 imputation (LDKNN), sites in local LD (LLD), variant call format (VCF)  
9

## 10 11 FUNDING

12  
13  
14  
15 This work has been funded by grants from National Key Basic Research Program of China (2014CB138206),  
16  
17 National Science Foundation of China (Grant #31271736), Bill & Melinda Gates Foundation (Yunbi Xu),  
18  
19 National Science Foundation IOS #1238014, USDA-ARS, and USDA NIFA grant 2009-65300-05668.  
20  
21

## 22 23 COMPETING INTERESTS

24  
25 The authors have no competing interests to declare.  
26  
27

## 28 29 REFERENCES

- 30  
31 1. Chia J-M, Song C, Bradbury PJ, Costich D, Leon N de, Doebley J, et al. Maize HapMap2 identifies extant  
32 variation from a genome in flux. *Nat Genet.* 2012;44:803–7. doi:10.1038/ng.2313.  
33  
34 2. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool for genome  
35 analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC*  
36 *Genomics.* 2014;15:823. doi:10.1186/1471-2164-15-823.  
37  
38 3. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map  
39 of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65. doi:10.1038/nature11632.  
40  
41 4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation  
42 discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43:491–8.  
43 doi:10.1038/ng.806.  
44  
45 5. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity,  
46 diversity, and dynamics. *Science.* 2009;326:1112–5. doi:10.1126/science.1178534.  
47  
48 6. Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of  
49 maize. *Science.* 2009;326:1115–7. doi:10.1126/science.1177837.  
50  
51 7. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. TASSEL-GBS: a high capacity  
52 genotyping by sequencing analysis pipeline. *PLoS ONE.* 2014;9:e90346.  
53 doi:10.1371/journal.pone.0090346.  
54  
55 8. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive  
56 genotyping of the USA national maize inbred seed bank. *Genome Biol.* 2013;14:R55. doi:10.1186/gb-  
57 2013-14-6-r55.  
58  
59 9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format  
60 and SAMtools. *Bioinformatics.* 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.  
61  
62  
63  
64  
65

10. Flint-Garcia SA, Thuillet A-C, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 2005;44:1054–64. doi:10.1111/j.1365-313X.2005.02591.x.
11. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–8. doi:10.1093/bioinformatics/btr330.
12. Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S. LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3 (Bethesda).* 2015;5:2383–90. doi:10.1534/g3.115.021667.
13. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. Genetic properties of the maize nested association mapping population. *Science.* 2009;325:737–40. doi:10.1126/science.1174320.
14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.
15. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda).* 2011;1:171–82. doi:10.1534/g3.111.000240.
16. Patefield WM. Algorithm AS 159: An Efficient Method of Generating RXC Tables with Given Row and Column Totals. *Applied Statistics.* 1981;30:91–7.
17. Bukowski R, Guo X, Lu Yanli, Zou C, He B, Rong Z, Wang B, Xu D, Yang B, Xie C, Fan L, Gao S, Xu X, Zhang G, Li Y, Jiao Y, Doebley JF, Ross-Ibarra J, Lorant A, Buffalo V, Romay MC, Buckler ES, Xu Y, Lai J, Ware D, Sun Q: Supporting data for "Construction of the third generation *Zea mays* haplotype map". *GigaScience Database.* 2017. <http://dx.doi.org/10.5524/100339>

## FIGURES

Figure 1: Overview of HapMap 3 pipeline. Initial set of tentative variant sites was obtained from 916 taxa using reads with mapping quality (MAPQ) at least 30 and bases with base quality (BQ) at least 10. At least 10 taxa had to have non-zero read coverage and the p-value from the segregation test (ST) on allelic depths had to be at most 0.01. This initial set of sites was subject to filtering based on identity by descent (IBD). Application of linkage disequilibrium (LD) filter eliminated sites with only non-local LD hits, leading to HapMap 3.1.1 variant set. An alternative route, leading to HapMap 3.2.1 genotypes, involved K nearest neighbors (KNN) imputation in which distances were computed using sites in good local LD (hence – LD

1  
2  
3  
4 KNN). See text for detailed explanation of methods and acronyms. The exact numbers of variant sites in  
5  
6 HapMap 3.1.1 and HapMap 3.2.1 are 61,228,639 and 83,153,144, respectively.  
7  
8  
9

10  
11 Figure 2: Overlap between various classes of HapMap 3.1.1 polymorphic sites. All sites listed passed the  
12  
13 ST and IBD filters. LLD sites are those found in local LD with the GBS anchor. Sites flagged IBD1 passed the  
14  
15 IBD filter, however, no alternative allele was present in IBD contrasts. Such sites do not violate IBD, but  
16  
17 the existence of a variant is not confirmed. NI5 flag is used to mark indels and sites within 5 bp of an indel.  
18  
19 Since no local re-alignment was done, the NI5 sites are not reliable.  
20  
21  
22  
23  
24

25  
26 Figure 3: Polymorphic sites detected by HapMap 3.1.1 pipeline based on two read mapping quality  
27  
28 thresholds:  $MAPQ \geq 1$  (q1) and  $MAPQ \geq 30$  (q30). Tightening of MAPQ threshold affects mostly the sites  
29  
30 flagged with IBD1 (least reliable), while the LLD sites (in local LD with GBS anchor) are mostly independent.  
31  
32  
33  
34

35  
36 Figure 4: Distribution of inbreeding coefficient for HapMap 3.1.1 variant sets obtained with two read  
37  
38 mapping quality thresholds:  $MAPQ \geq 1$  (q1) and  $MAPQ \geq 30$  (q30). Lower MAPQ threshold leads to lower  
39  
40 values of inbreeding coefficient (i.e., higher heterozygosities) resulting from misaligned reads.  
41  
42  
43  
44

45  
46 Figure 5: Overlap between HapMap 3.1.1 and HapMap 3.2.1 variant sites. 86% of HapMap 3.1.1 sites (99%  
47  
48 of those in local LD) are recovered by the HapMap 3.2.1 pipeline.  
49  
50  
51

52  
53 Figure 6: Distribution of fraction of heterozygous sites per taxon for un-imputed and imputed HapMap  
54  
55 3.2.1. Curves marked LLD have been obtained considering only sites verified in HapMap 3.1.1 to be in  
56  
57 good local LD with GBS anchor.  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Figure 7: Cumulative distribution of mapping quality from BWA mem alignment of 125.4 million 150 bp reads from taxon A272.

Figure 8: Linkage Disequilibrium-based filtering flowchart. The procedure eliminates sites with weak or non-local only LD hits. Sites with good local LD hits as well as those for which LD could not be probed (because of low MAF) are retained.

## ADDITIONAL FILES

Additional file 1: HapMap3TaxaAndCoverage.xlsx – spreadsheet with a list of all lines used in HapMap 3 with their corresponding coverage

Additional file 2: DepthFormatDetails.pdf - details of byte representation and storage format used for allelic depths

# Construction of the third generation *Zea mays* haplotype map

Robert Bukowski<sup>1</sup>, Xiaosen Guo<sup>2,3</sup>, Yanli Lu<sup>4</sup>, Cheng Zou<sup>5</sup>, Bing He<sup>2</sup>, Zhengqin Rong<sup>2</sup>, Bo Wang<sup>2</sup>, Dawen Xu<sup>2</sup>, Bicheng Yang<sup>2</sup>, Chuanxiao Xie<sup>5</sup>, Longjiang Fan<sup>6</sup>, Shibin Gao<sup>4</sup>, Xun Xu<sup>2</sup>, Gengyun Zhang<sup>2</sup>, Yingrui Li<sup>2</sup>, Yinping Jiao<sup>7</sup>, John F. Doebley<sup>8</sup>, Jeffrey Ross-Ibarra<sup>9</sup>, Anne Lorant<sup>9</sup>, Vince Buffalo<sup>9</sup>, M. Cinta Romay<sup>10</sup>, Edward S. Buckler<sup>10,11</sup>

## *Corresponding authors:*

Yunbi Xu<sup>5,12</sup>, Jinsheng Lai<sup>13</sup>, Doreen Ware<sup>7</sup>, and Qi Sun<sup>1</sup>

## *Authors' affiliations:*

<sup>1</sup>Bioinformatics Facility, Institute of Biotechnology, Cornell University, Ithaca, NY 14853

<sup>2</sup>BGI-Shenzhen, Shenzhen 518083, China

<sup>3</sup>Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

<sup>4</sup>Maize Research Institute, Sichuan Agricultural University, Wenjiang 611130, Sichuan, China

<sup>5</sup>Institute of Crop Science, Chinese Academy of Agricultural Sciences/National Key Facilities for Crop Gene Resource and Genetic Improvement, Beijing 100081, China

<sup>6</sup>Institute of Crop Science and Institute of Bioinformatics, Department of Agronomy, Zhejiang University, Hangzhou 310058, China

<sup>7</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

<sup>8</sup>Department of Genetics, University of Wisconsin, Madison, Wisconsin, USA

<sup>9</sup>Department of Plant Sciences, University of California, Davis, California, USA

<sup>10</sup>Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853

<sup>11</sup>US Department of Agriculture-Agricultural Research Service, Ithaca, NY 14853

<sup>12</sup>International Maize and Wheat Improvement Center (CIMMYT), El Batan 56130, Texcoco, Mexico

<sup>13</sup>National Maize Improvement Center, China Agricultural University (CAU)

1  
2  
3  
4 **ABSTRACT**  
5  
6

7 **Background**  
8  
9

10  
11 Characterization of genetic variations in maize has been challenging, mainly due to deterioration of  
12  
13 collinearity between individual genomes in the species. An international consortium of maize research  
14  
15 groups combined resources to develop the maize haplotype version 3 (HapMap 3), built from whole  
16  
17 genome sequencing data from 1,218 maize lines, covering pre-domestication and domesticated *Zea mays*  
18  
19 varieties across the world.  
20  
21

22  
23  
24 **Results**  
25  
26

27  
28 A new computational pipeline was set up to process over 12 trillion bp of sequencing data, and a set of  
29  
30 population genetics filters were applied to identify over 83 million variant sites.  
31  
32

33 **Conclusions**  
34  
35

36  
37 We identified polymorphisms in regions where collinearity is largely preserved in the maize species.  
38  
39 However, the fact that the B73 genome used as the reference only represents a fraction of all haplotypes  
40  
41 is still an important limiting factor.  
42  
43

44 **KEYWORDS**  
45  
46

47  
48 *Zea mays*, sequencing, haplotype map, genotyping, variant discovery, linkage disequilibrium, identity by  
49  
50 descent, imputation  
51  
52

53  
54 **BACKGROUND**  
55  
56

57  
58 Maize, one of the most important cereals in the world, also happens to be among the crop species with  
59  
60 the most genetic diversity. Advances in the next generation sequencing technologies made it possible to  
61  
62

1  
2  
3  
4 characterize genetic variations in maize at genomic scale. The previously released maize HapMap2 were  
5  
6 constructed with whole genome sequencing data of 104 maize lines across pre-domestication and  
7  
8 domesticated *Zea mays* varieties [1]. Since then, more maize lines have been sequenced by the  
9  
10 international research community, and a consortium was formed to develop the next generation  
11  
12 haplotype map. The maize HapMap 3 consortium includes, among others, BGI-Shenzen, Chinese Academy  
13  
14 of Agricultural Sciences, China Agricultural University (CAU), International Maize and Wheat Improvement  
15  
16 Center (CIMMYT). High-coverage data for 31 European and US Flint and Dent lines is also available in Ref.  
17  
18 [2]. Altogether, in this work we used a total of 1218 maize lines sequenced with depth varying from below  
19  
20  
21  
22  
23 1x to 59x.

24  
25  
26 A common approach in today's genetic diversity projects is to map the shotgun sequencing reads from  
27  
28 each individual onto a common reference genome to identify DNA sequence variations, and the physical  
29  
30 positions of the reference genome is used as a coordinate system for the polymorphic sites. A good  
31  
32 example is the human 1000 genome project [3]. The computational data processing pipeline developed  
33  
34 for the human project, GATK, has been widely adopted for identifying genetic variations in many other  
35  
36 species [4].  
37  
38  
39  
40

41 As the sequencing technology is improved and sequencers' base calling error model gets more accurate,  
42  
43 the computational challenges in genotyping by short-read sequencing have shifted from modeling  
44  
45 sequencer machine artifacts errors to resolving genotyping errors derived from incorrect mapping of short  
46  
47 reads to the reference genome. The problem is associated with the experimental design that uses the  
48  
49 single-reference genome as coordinate system. Taking maize as an example, the reference being used is  
50  
51 a 2.1 Gb assembly from an elite inbred line B73 that represents 91% of the B73 genome [5], and was  
52  
53 estimated to capture only ~70% of the low-copy gene fraction of all inbred lines [6]. The sequence  
54  
55 alignment software, however, can map 95-98% of the whole genome sequencing reads to the reference.  
56  
57  
58 That suggests a high percentage of the reads were mapped incorrectly, either being mapped to the  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 paralogous loci or highly repetitive regions under-represented in the reference assembly. The genetic  
5  
6 variants called from the miss-mapped reads need to be corrected computationally. The maize HapMap2  
7  
8 relied on linkage disequilibrium in the population to purge most of the bad markers caused by alignment  
9  
10 errors. To construct maize HapMap 3, a new computational pipeline was developed from scratch to  
11  
12 handle the sequencing data from 10 times more lines, and also took advantage of the high quality genetic  
13  
14 map constructed from the GBS technology [7, 8] which was not present when HapMap2 was constructed.  
15  
16  
17

18  
19 Genome structure variation in the population, including transposition, deletion, duplication and inversion  
20  
21 of the genomic segments, poses another challenge in the HapMap projects. As the physical genomes of  
22  
23 each of the individuals included in the HapMap projects vary both by size and structure, and there is no  
24  
25 co-linearity of all the sequence variants between the reference and genomes of each of the individuals, it  
26  
27 is not always possible to anchor all genetic variants in a population onto a single reference coordinate  
28  
29 system. As a compromise, markers included in the maize HapMap are defined as sites of which the  
30  
31 physical positions of the B73 alleles matching the markers' consensus genetic mapped positions.  
32  
33  
34

35  
36 Here we present maize haplotype map version 3 (HapMap 3), which is a result of coordinated efforts of  
37  
38 the international maize research community. The build includes 1,218 lines and over 83 million variant  
39  
40 sites anchored to the B73 reference genome version AGP v3.  
41  
42  
43

## 44 45 DATA DESCRIPTION

46  
47  
48 The sequencing data used in this work is comprised of 12,497 billion base pairs in a total of 113,702 billion  
49  
50 Illumina paired-end reads, originating from 1,218 maize and teosinte lines. The data was collected from  
51  
52 several sources over several years, and varies in quality, read length, and coverage. Basic information  
53  
54 about various datasets and stages of the HapMap 3 project they were used in are summarized in Table 1.  
55  
56 Each of the 1,218 lines were sequenced at depth varying from below 1x to 59x, using reads of lengths  
57  
58 ranging from 44 through 201 bp, averaging 110 bp. All reads were aligned to maize reference genome B73  
59  
60  
61

version AGP v3 using BWA mem aligner [9]. Overall, 95-98% of the reads were mapped to the reference genome, although only about 50-60% with non-zero mapping quality.

Table 1: Sequence datasets used in various stages of HapMap 3

Dataset	# Taxa	Coverage per taxon			3.1.1	3.2.1unimp	3.2.1imp
		Minimum	Maximum	Average			
HapMap2	103	1	18.5	4.1	+	+	+
Hapmap2 extra	44	4.2	42	11.5	+	+	+
CAU	725	0.06	36.8	1.75	+	+	+
CIMMYT/BGI	89	1.1	19	11	+	+	+
282-2x	271	0	9	2.2	-	+	+
282-4x	270	0.6	34.5	4.4	-	+	-
German, Ref. [2]	31	8.3	59	17.4	-	+	+

Taxa from sets “HapMap2”, “HapMap2 extra”, and “CAU” partially overlap. The “282” libraries, sequenced twice represent 271 taxa. A “+” means that the dataset was used in a given stage, “-“ that it was not.

All sequence data used in this work is publically available. Collection and publishing of this data does not violate any local or international legislation or guidelines.

## ANALYSIS

### Initial variant discovery

The HapMap 3 pipeline is summarized in Figure 1. First, polymorphic sites were called for a set of 916 taxa from datasets HapMap2 through CIMMYT/BGI (7,191 billion base pairs, 74,643 million reads). In the first step, a custom built software tool was used to determine genotypes for each taxon at each site of the genome based on allelic depths at that site. Bases counted towards depth had base quality score of at least 10 and originated from reads with mapping quality ( $MAPQ = -\text{int}(10\log P)$ , where  $P$  – calculated by the BWA mem aligner - is the probability of the reported read location being wrong) at or above 30. This cut-off was chosen at mid-point between the highest MAPQ value reported by the aligner,

1  
2  
3  
4 corresponding to unambiguous alignments (60) and that of the most ambiguous ones (0). Analysis of the  
5  
6 inbreeding coefficient (Section HapMap 3.1.1) and of MAPQ distributions shows that our choice of cut-off  
7  
8 leads to decent quality genotypes while allowing for over 80% of alignments with MAPQ>0 to be included.  
9  
10 Only sites where at least 10 taxa had coverage of 1 or more were considered. Following Ref. [1], at each  
11  
12 site the allelic read depths were subject to segregation test (ST – see METHODS section for details). For a  
13  
14 population of inbred lines at true variant sites, one expects depths corresponding to minor and major  
15  
16 alleles to be concentrated in roughly different subset of taxa rather than being randomly distributed. The  
17  
18 purpose of the ST test is to find and eliminate sites for which allelic depth distribution appears random,  
19  
20 as such randomness, indicating high heterozygosity, is likely caused by alignment and sequencing errors.  
21  
22 A measure of the randomness is the p-value of the ST test (the smaller the p-value, the less random the  
23  
24 distribution). A p-value threshold of 0.01 was used in this study. This choice was somewhat arbitrary,  
25  
26 aimed at reducing the number of tentative variant sites to manageable size before further, more stringent  
27  
28 filters were applied. In this first, ST-based round of filtering, a total of 196 million tentative polymorphic  
29  
30 sites were selected. In the second step, these sites were filtered using the identity by descent (IBD)  
31  
32 information derived from about 0.5 million of high-quality polymorphisms obtained previously [8] using  
33  
34 the Genotyping-By-Sequencing (GBS) approach [7]. These GBS variants (GBS anchor) were used to  
35  
36 determine regions of IBD, where certain pairs of taxa are expected to have identical haplotypes. The  
37  
38 tentative polymorphic sites violating these IBD constraints were then filtered out, leaving 96.8 million  
39  
40 sites. At roughly half of the sites surviving this filter, minor allele was not present in taxa involved in the  
41  
42 tested IBD relationships. At such sites (typically with low minor allele frequency), the satisfied IBD  
43  
44 constraints do not confirm the existence of a variant. They are therefore less reliable and have been  
45  
46 marked with “IBD1” flag in the VCF files (see Table 2 for summary of flags and parameters present in  
47  
48 HapMap 3 VCF files). The ST- and IBD-filtered variant sites were then used in two separate procedures,  
49  
50 leading to two versions of HapMap 3 genotypes, referred to as HapMap 3.1.1 and HapMap 3.2.1.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

Table 2: Flags and parameters used in INFO field of VCF files in various HapMap 3 versions.

Parameter	3.1.1	3.2.1 unimp	3.2.1 imp	Description
DP	+	+	+	Total read depth at the site
NZ	+	+	+	Number of taxa with called genotypes
AD	+	+	+	Allelic depths (reference, alternative in order listed in ALT field)
AC	+	+	+	Numbers of alternative alleles in order listed in ALT field
AQ	+	+	+	Average allele base qualities (reference, alternative in order listed in ALT field) computed in HapMap 3.1.1 from 916 taxa
GN	+	+	+	Numbers of genotypes (AA,AB,BB or AA,AB,AC,BB,BC,CC if 2 alt alleles present)
HT	+	+	+	Number of heterozygotes
EF	+	+	+	EF=heterozygosity/(presence_frequency*minor_allele_frequency); computed in HapMap 3.1.1 from 916 taxa
PV	+	+	+	p-value from segregation test, computed in HapMap 3.1.1. from 916 taxa
MAF	+	+	+	Minor allele frequency (summed up over all alternative alleles)
MAFO	-	-	+	Minor allele frequency in unimputed HapMap 3.2.1.
FH	+	-	-	Fraction of heterozygous taxa among the 506 taxa with more than 50% non-missing genotypes on chr 10
FH2	+			Site with FH greater than 2%
IBD1	+	+	+	only one allele present in IBD contrasts - based on 916 taxa of HapMap 3.1.1
LLD	+	+	+	Site in local LD with GBS map - based on 916 taxa of HapMap 3.1.1
NI5	+	+	+	Indel or site within 5 bp of a putative indel - from 916 taxa of HapMap 3.1.1
INHMP311	-	+	+	Site present in HapMap 3.1.1
ImpHomoAccuracy	-	-	+	Fraction of homozygotes imputed back into homozygotes
ImpMinorAccuracy	-	-	+	Fraction of minor allele homozygotes imputed back into minor allele homozygotes
DUP	-	-	+	Site with heterozygotes frequency > 3% - based on unimputed HapMap 3.2.1 genotypes

“+” and “-” indicate presence or absence, respectively, of a parameter or flag in a given version of HapMap. For example, “-+-” means the parameter is present in VCF file of both unimputed and imputed HapMap 3.2.1, and absent from HapMap 3.1.1. VCF files. Unless indicated otherwise, all parameters are computed from depths and genotypes in the current VCF file.

## HapMap 3.1.1

The HapMap 3.1.1 procedure involved checking for linkage disequilibrium of each site against the GBS anchor map [7, 8], which consists of markers located in hypo-methylated and genetically stable regions. Sites giving only very weak or only nonlocal (i.e., outside of 1 Mb radius) linkage Disequilibrium (LD) hits were eliminated, which resulted in the final set of 61,228,639 polymorphisms. For slightly less than 40% of these sites, LD could not be conclusively calculated due to small minor allele frequencies (MAF), whereas the remaining sites, confirmed to be in local LD with the GBS anchor, have been marked with flag “LLD”. Among the sites surviving all filtering steps, 8.7 million are indels or are located near (within 5 bp) of an indel. These have been marked with the flag “NI5”. Since a procedure to achieve consistent alignment across all reads covering the same indels - local realignment - is not computationally feasible at this scale and has not been performed, genotyping errors could occur, and, consequently, most such sites are tentative and should be treated with caution.

Figure 2 shows overlaps between various classes of variants of HapMap 3.1.1. First, we notice a rather small overlap between sites in confirmed local LD (“LLD” flag) and those marked “IBD1”. This is understandable, as the IBD1 sites represent mostly low MAF cases, where LD assessment could not be done. Indels and vicinity (labeled “NI5”) constitute about 15% of sites in each of the LLD, IBD1, and the union of LLD and IBD1 sets. Only a very small fraction of sites does not carry LLD or IBD1 flag, i.e., they are strongly confirmed by the IBD filter, but could not be classified with LD. The subset of 29.8 million sites in local LD and away from indels should be considered the most reliable.

To check the sensitivity of the obtained variant set to the mapping quality threshold imposed on the reads counted towards allelic depths, we repeated the pipeline using the mapping quality threshold equal to 1. Comparison of the variant set obtained this way (referred to as q1) with our recommended set (q30) is shown in Figure 3. While the overall number of variant sites is approximately independent of the mapping

1  
2  
3  
4 quality threshold, the two pipelines produce significantly different sets of sites, with only 72% of all q30  
5  
6 sites reproduced by the q1 pipeline. Closer inspection shows that this variability is due primarily to the  
7  
8 IBD1 sites, for which our filtering strategy was the least stringent. On the other hand, the LLD sites,  
9  
10 confirmed to be in local LD with GBS anchor, are much more independent of the mapping quality  
11  
12 threshold, which confirms high quality of such sites.  
13  
14

15  
16  
17 For a population of inbred lines considered here, insight into genotype quality may be obtained from the  
18  
19 inbreeding coefficient, calculated here for each taxon using the VCFtools program [11] from the formula  
20  
21

$$F_{inbr} = (O - E)/(N - E)$$

22  
23  
24  
25 where  $O$  is the observed number of homozygotes for a given taxon,  $N$  is the number of sites at which the  
26  
27 taxon was genotyped, and  $E$  is the expected number of homozygotes given by  $E = \sum_k(1 - 2p_kq_k)$ .  
28  
29 Summation in the latter formula runs over  $N$  genotyped sites,  $p_k$  is the minor allele frequency at site  $k$   
30  
31 (computed from all taxa in the population with non-missing genotypes at this site), and  $q_k = (1 - p_k)$ .  
32  
33 Low values of the inbreeding coefficient, indicating high heterozygosity, are mostly due to genotyping  
34  
35 errors. Importance of choosing a sufficiently tight mapping quality threshold for the quality of genotypes  
36  
37 is apparent from Figure 4, where the distribution of inbreeding coefficient for chromosome 10 is shown  
38  
39 for q1 and q30 variant sets. The lower MAPQ threshold results in a large number of miss-mapped reads  
40  
41 being counted towards depth, producing overly heterozygous genotypes, especially for highly covered  
42  
43 taxa (the peak below 0.8 is due mostly to CIMMYT lines with 10-15x coverage; these lines have higher  
44  
45 heterozygosity than other lines which may also contribute to the peak) and thus shifting the curve to the  
46  
47 left. Since most HapMap 3 taxa are inbred lines, one should expect the true distribution to be contained  
48  
49 within peak around 0.95. In view of this, the q30 result is definitely an improvement over q1, although a  
50  
51 longer than expected tail extending towards the value 0.8 indicates that the HapMap 3 variants may  
52  
53 contain too many false heterozygotes.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 Seemingly heterozygous sites may result from either sequencing errors or misalignments of reads  
5  
6 originating from paralogous regions. To investigate this further, we calculated, for each site, the fraction  
7  
8 of heterozygous HapMap 3.1.1 genotypes within a subset of 506 high-coverage taxa (defined as those  
9  
10 with more than 50% non-missing genotypes on chromosome 10). In HapMap 3.1.1 VCF files, this fraction  
11  
12 has been recorded as parameter “FH”. At sites for which this parameter exceeds 2-3%, heterozygotes are  
13  
14 likely to originate from misalignments, for example, from tandem and ectopic duplications. Such sites  
15  
16 constitute 9% of all HapMap 3.1.1 sites.  
17  
18  
19  
20

## 21 HapMap 3.2.1

22  
23  
24 The 96.8 million ST- and IBD-filtered variant sites were the starting point for the HapMap 3.2.1 procedure  
25  
26 (Figure 1). On these sites, genotypes were called on the 263 taxa from the “282” panel of Ref. [10] using  
27  
28 “282 2x” dataset, and on the 31 high-coverage (on average 17 x) “German” taxa [2], for the total of 1210  
29  
30 taxa. Some of the taxa present in the “282” and “German” sets carry the same names as the ones included  
31  
32 in the 916-taxon HapMap 3.1.1 set. Since despite identical names such taxa often originate from different  
33  
34 germplasm sources, they have been kept separate during genotyping, i.e., reads from different sources  
35  
36 were not merged and separate genotypes were computed for each source. In the resulting VCF files, the  
37  
38 names of the overlapping taxa have been prefixed by “282set\_” and “german\_”. For example, in the case  
39  
40 of B73, there are three columns representing different datasets for this taxon: “B73” (the original 916-  
41  
42 taxa set), “282set\_B73” (sequence from the more recent “282” libraries), and “german\_B73” (from Ref.  
43  
44 [2]).  
45  
46  
47  
48  
49  
50

51 To further eliminate the false positives resulting from sequencing errors, an additional depth-based filter  
52  
53 was applied to the 96.8 million sites. Referred to as “>1,>2” filter it accepts sites for which the read support  
54  
55 of minor allele was greater than 1 in at least one taxon and greater than 2 across all taxa. Genotypes on  
56  
57 the surviving 83,153,144 sites, referred to as “un-imputed HapMap 3.2.1”, were then processed through  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 the LD KNN imputation procedure based on Ref. [12], where the “nearest neighbors” of a given line are  
5  
6 selected based on sites in good local LD with the target site. Whenever possible, the procedure filled up  
7  
8 missing genotypes with imputed ones, but the non-missing genotypes were left unchanged, even if  
9  
10 imputation classified them differently. Non-imputable missing genotypes at the sites with (pre-  
11  
12 imputation) MAF below 1% were assumed to be major allele homozygotes. Imputation reduces the  
13  
14 fraction of missing genotypes from 50% to 7%. Most of the originally missing genotypes (about 85%) are  
15  
16 imputed to major allele homozygotes. Accuracy of the genotype dataset can be assessed by comparing  
17  
18 the original genotypes with imputed ones. As shown in Table 3, 99.8% of major allele homozygotes are  
19  
20 imputed back into the same class. While the accuracies of minor allele homozygotes and genotypes  
21  
22 including indels are both above 90%, only 11% of heterozygotes are imputed back into the same class,  
23  
24 while 47% of them fail imputation altogether. This reflects the inherent difficulty in calling heterozygotes.  
25  
26 In the single-reference approach to maize genotyping employed here, heterozygous sites represent true  
27  
28 residual heterozygosity as well as misalignments of reads from tandem and ectopic duplications. Since  
29  
30 residual heterozygosity in our population of predominantly inbred lines should not exceed 2-3%, all  
31  
32 heterozygotes with frequency  $\geq 3\%$  can be considered a result of misalignments. About 10% of all  
33  
34 heterozygotes present in HapMap 3.2.1 set satisfy this condition. In the VCF files, these sites have been  
35  
36 flagged with flag DUP (“duplicated regions”). Other parameters generated by the imputation procedure  
37  
38 and recorded for each variant site in the INFO field are ImpHomoAccuracy fraction of all homozygotes  
39  
40 imputed back into homozygotes and ImpMinorAccuracy fraction of minor allele homozygotes imputed  
41  
42 back to the same class. The INFO field also contains flags IBD1, LLD, and NI5, computed from the initial  
43  
44 916 taxa in the HapMap 3.1.1 procedure. Genotypes resulting from the imputation procedure are referred  
45  
46 to as “imputed HapMap 3.2.1”.

47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57 Table 3: Accuracy of various genotype classes based on statistics from imputation in HapMap 3.2.1  
58  
59  
60  
61  
62  
63  
64  
65

Genotype class	Accuracy within class [%]	% unimputed
Major allele homozygote	99.8	1.2
Heterozygote	11.1	47.0
Minor allele homozygote	94.4	14.2
Indel	92.2	17.3

Accuracy computed as percentage of the original number of genotypes in a given class (excluding genotypes that could not be imputed) imputed into the same class. The last column shows the fraction of genotypes within a class which could not be imputed.

Relationship between variant sites included in HapMap 3.1.1 and 3.2.1 is shown in Figure 5. Both pipelines start from the same set of IBD-filtered genotypes and subject them to different kinds of filtering, with that of HapMap 3.1.1 being more stringent. It is therefore not surprising that HapMap 3.2.1 recovers the majority (86%) of HapMap 3.1.1 sites, including over 99% of those flagged LLD (i.e., confirmed in local LD). In addition, 30.3 million extra sites are retained in HapMap 3.2.1, which failed the LD filter in HapMap 3.1.1 pipeline. On the other hand, the depth-based “>1,>2” filter applied in HapMap 3.2.1 eliminated 8.2 million sites present in HapMap 3.1.1, including about 0.2 million LLD ones.

After the HapMap 3.2.1 release was completed, “282-2x” sequencing data became available for additional 8 taxa from the “282” panel. Libraries for all 271 taxa were also re-sequenced at a higher depth (average of about 4.4x), leading to another dataset, “282-4x” (as this re-sequencing failed for one of the taxa, this dataset only contained 270 taxa). Therefore, the un-imputed HapMap 3.2.1 genotypes for all 271-taxa from the “282” panel were re-called using the full available sequencing depth, creating a separate variant dataset for the “282” panel.

## DISCUSSION

The maize genome, 2.3 GB in size [5], is smaller than the human genome. But some of its distinctive features makes it more challenging for variants identification. First, a recent whole genome duplication occurred 12 million years ago resulted in homologous segments that complicate the short read

1  
2  
3  
4 alignments; second, the rampant activities of transposable elements within last 1-5 million year not only  
5  
6 resulted in accumulation of large amount of relatively young repetitive elements in the intergenic regions,  
7  
8 but also extraordinary structural variations within species [5, 6]. In this study, the genome of the B73  
9  
10 maize line was used as the reference for variant calling from short sequencing reads. Structural variations  
11  
12 between B73 and other individuals has been the major challenge for identification of true variants. In  
13  
14 particular, short reads derived from regions missing in the reference genome could be mapped to  
15  
16 other paralogous regions, which lead to false positive genotypes. In the human 1000 genome project, a  
17  
18 new HaplotypeCaller was used [4], which performs local *de novo* assembly to identify the most likely  
19  
20 haplotypes for each individual and thus improve the genotyping results. However, HaplotypeCaller is  
21  
22 computationally very expensive, and not always applicable in species like maize, where the single  
23  
24 reference genome misses many haplotypes presents in the species and has a lot more mismatched  
25  
26 paralogous reads that would disrupt the local assembly. To filter out these false positive variants called  
27  
28 from the mismatched reads, we relied on the *Zea* GBS map [7, 8], which was obtained from GBS markers  
29  
30 located primarily in hypo-methylated chromosomal regions. GBS maps were used to identify IBD regions  
31  
32 between the individual genomes, and 100 million markers with high percentage of mismatched genotype  
33  
34 calls in the IBD regions were filtered out from the initial set of 196 million markers. The highly repetitive  
35  
36 genomic regions derived from recent transposition activities are in general easier to identify, because the  
37  
38 templates of these repeats are well represented on the reference genome, and sequencing reads mapped  
39  
40 to these regions, flagged with low mapping quality, can be removed at the early stage of the analysis  
41  
42 pipeline. For HapMap 3, reads with mapping quality lower than 30 were not included in the build.  
43  
44  
45  
46  
47  
48  
49  
50

51  
52 One of the goals of HapMap 3.1.1 is to identify genetic markers in regions where collinearity is preserved  
53  
54 in majority of maize lines. The LD filter in the pipeline was applied for this purpose. To do this, we  
55  
56 genetically mapped the presence/absence of the minor alleles using the GBS genetic map, and these  
57  
58 mapped genetic positions were compared to the physical positions on the B73 reference. Among the 96.8  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 million sites surviving the IBD filter, 25% did not have enough non-missing data or sufficient minor allele  
5  
6 frequency for genetic mapping to be meaningful. For 38% of sites, at least one genetically mapped  
7  
8 position matching the physical positions on B73 reference was found, 24% have no significant hits from  
9  
10 genetic mapping, probably due to no consensus positions in the HapMap 3 population, and 13% have  
11  
12 genetic positions not matching the B73 physical positions. Markers from the latter two categories (37% of  
13  
14 all IBD-filtered markers) were removed by the LD filter, leaving slightly over 61 million sites, about 60% of  
15  
16 which were confirmed in local LD and marked with a flag “LLD” in VCF files.  
17  
18  
19  
20

21 The IBD and LD filters applied in the HapMap 3.1.1 project effectively remove majority of the false positive  
22  
23 genetic variants caused by paralogous genomic regions, as well as markers with lost collinearity between  
24  
25 the species. However, not all the genotyping errors have been removed from the release. 23,839,286 of  
26  
27 the sites do not have sufficient minor allele frequency for genetic test (these are missing the “LLD” label  
28  
29 in the INFO field of the VCF files). Another source of errors are paralogous regions evolved from tandem  
30  
31 duplications. Misalignments of reads from such regions result in false heterozygous genotypes with  
32  
33 relatively high frequency and in local LD, and therefore difficult to filter out. Given enough sequencing  
34  
35 depth, the tandem duplications can be identified either as copy number variation or imputation errors.  
36  
37 However, majority of the HapMap 3 lines have very low sequencing depth, and fail to sample all  
38  
39 paralogous loci or all alleles, which makes it difficult to flag all sites complicated by tandem duplications.  
40  
41  
42  
43  
44

45 Local LD filter based on a large, diverse population may be too stringent, as some markers, good within  
46  
47 certain sub-populations, may be thrown out. Therefore, the LD filter was not used in the HapMap 3.2.1  
48  
49 release, which contains a total of 83 million variant sites, subject only to ST and IBD filters and an  
50  
51 additional depth-based filter aimed to improve reliability of rare allele calls. Although those sites are likely  
52  
53 to have higher misalignment rates, they are still likely to capture a true association with phenotypes.  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 In the un-imputed HapMap 3.2.1, at about 10% of all variant sites, fraction of heterozygous taxa exceeds  
5  
6 3%. Such sites are marked “DUP”, as most likely originating from duplication misalignments. Figure 6  
7  
8 shows the distribution of fraction of heterozygous sites per taxon for different versions of HapMap 3.2.1  
9  
10 release. While for the un-imputed genotypes the distribution peaks slightly below 1%, imputation  
11  
12 significantly shifts the peak to the left, down to about 0.5%. This is a consequence of most missing  
13  
14 genotypes being imputed to homozygotes. Interestingly, considering only sites in good local LD (marked  
15  
16 with the “LLD” flag) leads to distributions (both imputed and un-imputed) shifted towards higher  
17  
18 heterozygosities. This is understandable, as the LLD sites are typically those with higher minor allele  
19  
20 frequencies, where the chance of encountering a heterozygote is higher.  
21  
22  
23  
24  
25

26 In summary, besides the addition of more maize lines, the HapMap 3.2.1 release differs from the 3.1.1  
27  
28 release in three major aspects: 1) Improved rare allele calls: to increase the accuracy of the variants with  
29  
30 rare allele, the HapMap 3.2.1 pipeline applied more stringent read depth thresholds instead of the  
31  
32 population genetics based LD filter that could not be applied to sites with very low MAF; 2) The sites with  
33  
34 high percentage of heterozygous calls were flagged in the VCF files; 3) Missing data was imputed using  
35  
36 the LD KNN method. As summarized in Table 2, the VCF files of both datasets contain labels that flag the  
37  
38 characteristics of each of the sites. To effectively use this resource, it is recommended to filter the sites  
39  
40 based on the flags that are appropriate to the purpose of each project.  
41  
42  
43  
44  
45

46 When constructing the maize HapMap 3, the most serious problems we were facing can be attributed to  
47  
48 the use of a genome from a single individual (B73) as a reference for other, often very different species.  
49  
50 This is becoming the single limiting factor in the study of maize diversity, as well as breeding practice. The  
51  
52 only remedy is to move away from a single genome-based reference coordinate and adopt a pan-genome  
53  
54 based reference system that incorporates all major haplotypes of the species.  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## METHODS

### Plant material

Plant material used in this study was obtained mostly from maize inbred lines representing wide range of *Zea mays* diversity. 103 of these lines, used previously in the HapMap2 project [1], include 60 improved lines, including the parents of the maize nested association mapping (NAM) population [13], 23 maize landraces and 19 wild relatives (teosinte lines: 17 *Z. mays ssp. parviglumis* and 2 *Z. mays ssp. mexicana*). Sequence datasets originating from these lines are referred to in Table 1 as “HapMap2” and “HapMap2 extra”. Majority of the remaining inbred lines originated from CAU (sequence dataset “CAU”) and include, among others, “Chinese NAM” parent lines. Additional 89 inbred lines were provided by CIMMYT and sequenced at BGI (dataset “CIMMYT/BGI”). The HapMap 3 population also contained one *Tripsacum* line (TDD39103), one “mini-maize” line (MM-1A), and a few newly sequenced landraces. Overall, the number of taxa in the initial, variant-discovery stages of the HapMap 3.1.1 project was 916.

The sequence of 271 taxa from the libraries of the “282” panel [10] were added at a later stage (HapMap 3.2.1). DNA to construct these libraries was obtained from the collection that the North Central Regional Plant Introduction Station (NCRPIS) distributes all over the world. Additionally, the high-coverage data of Ref. [2], originating from 31 European and US inbreds was also included. The total number of taxa genotyped in the HapMap 3.2.1 build is 1218.

In this study, individuals with the same taxa name but contributed by different members of the consortium were kept as separate entries in genotyping pipeline - a decision prompted by comparison of genotypes obtained from different datasets. For example, the newly sequenced CML103 is significantly more heterozygous from CML103 studied previously in the HapMap2 project. Also, the Mo17 sequence originating at CAU has been treated as taxon separate from Mo17 and CAUMo17. In most of those cases,

1  
2  
3  
4 a prefix or suffix indicating the origin of the sequence data has been added to the taxa name (e.g.,  
5  
6 "282set\_" or "german\_", "-chin").  
7  
8

## 9 10 Sequencing

11  
12 Sequencing has been performed over several years using various generations of Solexa/Illumina  
13  
14 instruments and library preparation protocols, giving paired end reads from 44 to 201 bp long. Overall,  
15  
16 113.7 billion reads were obtained on 1,218 lines, containing 12,497 billion base pairs, giving on average  
17  
18 4.4x coverage per line (assuming 2.3 Gb genome size). However, as shown in Table 1, coverage was not  
19  
20 uniform among all lines. For a few lines, sequence generated previously in the context of HapMap2 project  
21  
22 was augmented with reads from recent re-sequencing which brought the median coverage of the  
23  
24 HapMap2 lines to 5x, with average coverage equal to 7.8x and standard deviation of 7.2x. All NAM parent  
25  
26 lines are covered to 10x or higher. Most of the 89 lines provided by CIMMYT and sequenced at BGI have  
27  
28 coverage exceeding 10x. The recent re-sequencing of the "282" panel resulted in coverage between 1.7  
29  
30 and 36x, averaging 6.5x. Coverage of the 31 "German" lines for Ref. [2] ranges from 8.3 to 59x, with  
31  
32 average of 17.4x. Majority of the inbred lines originated from CAU have been sequenced at a lower  
33  
34 coverage (1-2x). The list of all lines used in HapMap 3 with the corresponding coverage is given in  
35  
36  
37  
38  
39  
40  
41 Additional file 1.  
42  
43

## 44 Alignment

45  
46  
47 Due to the use of different versions of Solexa/Illumina sequencing equipment, the base qualities in  
48  
49 different input FASTQ files are given in different encodings. Prior to alignment, all base qualities have been  
50  
51 converted to phred+33 scale. Reads were then aligned to B73 reference (AGP v3) as paired-end using bwa  
52  
53 mem aligner (1) with default options. In 72 read sets (Illumina lanes), for technical reasons a high (6%-  
54  
55 54%) fraction of paired-end fragments was found to be shorter than reads, so that the two ends contained  
56  
57 a part of Illumina adapter and were reverse complements of each other. For such "read-through"  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 fragments, the remnants of Illumina adapter sequences were clipped using TRIMMOMATIC [14] and only  
5  
6 one read was used and aligned as single-end. The bwa mem aligner is capable of clipping the ends of reads  
7  
8 and splitting each read in an attempt to map its different parts to different location on the reference. As  
9  
10 a result, typically over 95% of reads are reported as mapped. However, the fraction of reads with non-  
11  
12 zero mapping quality (negative log of the probability that a read has been placed in a wrong location) is  
13  
14 much lower – typically only 40-50%. Figure 7 shows a typical distribution of the mapping quality obtained  
15  
16 from bwa mem alignment. In practice, we only used alignments with mapping quality of at least 30. A  
17  
18 base was counted towards allele depth if its base quality score was at least 10.  
19  
20  
21  
22

23  
24 It is well known that alignment may be especially ambiguous when reads contain indels with respect to  
25  
26 the reference. In such cases, multiple-sequence realignment approaches have been proposed [4] to find  
27  
28 the correct sequence and location of an indel and avoid spurious flanking SNPs. Since indels are not the  
29  
30 primary focus of this work and since the realignment is computationally very expensive, it has not been  
31  
32 performed by the HapMap 3 pipeline. Thus, although indels and SNPs in their immediate vicinity have  
33  
34 been retained in the HapMap 3 VCF files, they are less reliable and have therefore been marked with “NI5”  
35  
36 label for easy filtering.  
37  
38  
39  
40

## 41 Genotyping pipeline

42  
43  
44 Raw genotypes were obtained using a custom-built multi-threaded java code. First, the code executes  
45  
46 samtools mpileup command (thresholds on the base and mapping quality are imposed here) for each  
47  
48 taxon individually, processing a certain portion of the genome. On a multi-core machine, several such  
49  
50 pileup processes (i.e., for several taxa) can be run concurrently as separate threads. Since we are  
51  
52 predominantly interested in calling SNPs, we use a simplified indel representation where insertions and  
53  
54 deletions with respect to reference are treated as additional alleles “I” and “D”, respectively, regardless  
55  
56 of length and actual sequence of the indel. Read depths and average base qualities of all six alleles (A, C,  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 G, T, I, and D) are extracted from samtools mpileup output for each taxon at each genomic position and  
5  
6 stored in an array shared between all threads. The amount of memory available on the machine along  
7  
8 with the number of taxa determine the upper limit on the size of this array, and therefore – the maximum  
9  
10 size of chromosome chunk which can be processed at one time. As base quality of I and D alleles we took  
11  
12 the value corresponding to the base directly preceding the indel on the reference.  
13  
14

15  
16 Extraction of allelic depths for all genomic positions is time consuming, which presents a major obstacle  
17  
18 if joint genotyping needs to be re-run, for example, upon extending the taxa set. It is therefore  
19  
20 advantageous to run the depth extraction only once for each taxon and save the obtained depths on disk  
21  
22 to be retrieved (rather than re-calculated) during the genotyping step. This way, when the taxa set for  
23  
24 genotyping is extended, mpileup step has to be run only for the newly added taxa. Thus, the program  
25  
26 features an option to save allelic depths and average qualities in specially designed data structures stored  
27  
28 in HDF5 files – one such file per taxon per chromosome. To save space, each allele depth and average  
29  
30 quality is stored as one byte, which allows exact representation of integers from 0 to 182, while higher  
31  
32 integers (up to about 10,000) are represented approximately by negative byte values through a  
33  
34 logarithmic formula with carefully chosen base. Depths and qualities are stored only for sites with non-  
35  
36 zero coverage. The details of the storage format and integer representation in terms of byte variables are  
37  
38 given in Additional file 2.  
39  
40  
41  
42  
43  
44

45  
46 Once the allelic depths for all taxa and a given chunk of the genome are available in shared memory, each  
47  
48 site is evaluated for presence of a tentative SNP. On a multi-core machine, the set of sites within the  
49  
50 genome chunk is divided into subsets processed in parallel on different cores. Sites with less than 10 taxa  
51  
52 with read coverage and those with only reference allele present are ignored. For all other sites, genotypes  
53  
54 are called for all taxa using a simple likelihood model with a uniform error rate [15] assumed at 1%.  
55  
56 Alternative alleles are then sorted according to their allele frequencies and up to two most abundant  
57  
58 alleles are kept, as decided by the segregation test described in the next Section. Sites for which all taxa  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 turn out to be reference homozygotes (which may happen despite non-reference alleles being present in  
5  
6 the mapped reads) are skipped. Raw variant set obtained in this way is then subject to extensive filtering  
7  
8 with the intention of reducing the number of false positives resulting from misalignments.  
9

## 10 11 Filtering

### 12 Segregation test (ST) filter

13  
14  
15 For each pair of alleles obtained in the genotyping step, a 2 by N (where N is the number of taxa)  
16  
17 contingency table is constructed, containing depths of the first allele in row 1 and depths of the second  
18  
19 allele in row 2. The Fisher exact test (FET) is then performed to assess how likely such a table is to occur  
20  
21 by chance. If the expected values of the array elements are sufficiently large, the p-value from FET is  
22  
23 approximated by that from the computationally efficient chi-square test. However, in most cases  
24  
25 encountered here, expensive simulation is needed to obtain sufficiently accurate p-value. To reduce  
26  
27 computational burden, we adopted a hybrid approach based on an empirical observation that for  
28  
29 statistically insignificant cases (p-values larger than 0.2) the chi-square test results in a de facto lower  
30  
31 bound to exact p-values. Thus, the chi-square test is performed first for each site and if the p-value from  
32  
33 this test is below 0.2, more exact p-value is obtained from a simulation procedure. The simulation  
34  
35 procedure used here, implemented in Java, is the same as the one implemented in R package [16]. An  
36  
37 alternative allele is kept if at least one contingency table involving this allele has p-value smaller or equal  
38  
39 to 0.01. If none of the alternative alleles survive the ST filter, the site is skipped (not reported in output).  
40  
41 The ST filter tends to eliminate variant sites resulting from random sequencing errors.  
42  
43  
44  
45  
46  
47  
48  
49  
50

### 51 GBS anchor map and IBD filter

52  
53  
54 Given a set of trustworthy SNPs and a diverse set of 916 taxa it is possible to identify, for an arbitrary  
55  
56 region of the genome, a number of taxa pairs which are identical by descent (IBD) and are therefore  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 expected to have identical genotypes in this region. If known, these IBD pairs can be used as a powerful  
5  
6 filter eliminating variant which violate IBD constraints.  
7  
8

9  
10 To determine the IBD regions, we used the first step of our pipeline to call genotypes for our 916 taxa on  
11  
12 the set of GBS v2.7 sites [7, 8] which tend to concentrate in relatively well-conserved low-copy regions of  
13  
14 the genome and can therefore be considered reliable. This set of 954,384 sites was filtered to include only  
15  
16 SNP (not indel) sites for which the p-value from the segregation test was below 0.05 and which were more  
17  
18 than 5 bp away from any indel. The set of genotypes at 475,272 sites obtained in this way, which will be  
19  
20 referred to as GBS anchor, agree well with those from GBS on 167 taxa present in both sets. Alleles  
21  
22 detected by the HapMap 3 pipeline agreed with those from GBS at 94% of the GBS sites. At 90% of the  
23  
24 sites, fraction of (non-missing data) taxa with genotypes in agreement with those from GBS was at or  
25  
26 above 85%. Genotypes different from GBS ones were observed for 82 taxa. These differences were most  
27  
28 frequent (up to 19% of all sites) for teosinte lines.  
29  
30  
31  
32

33  
34 The GBS anchor was used to compute the genetic distance (identity by state) between any two of the 916  
35  
36 lines in windows containing 2000 GBS sites each (about 8.5 Mbp on average). If the genetic distance within  
37  
38 such a window was  $\leq 0.02$  (about 10 times smaller than the mean distance across all pairs), the two lines  
39  
40 were considered to be in IBD. At least 200 comparable GBS sites (i.e., non-missing data simultaneously on  
41  
42 both lines being compared) were assumed necessary to make the genetic distance calculation feasible.  
43  
44 This allowed for good distance estimate while keeping the number of detected IBD relationships large.  
45  
46  
47

48  
49 The number of taxa involved in IBD relationships in any given window were between 385 (start of  
50  
51 chromosome 10) and 757 (middle of chromosome 7) and averaged 588, leading to large numbers of IBD  
52  
53 contrasts, ranging from 3,710 (beginning of chromosome 4) to 42,890 (middle of chromosome 7), and  
54  
55 averaging 13,500.  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 The tentative (ST-filtered) variant sites were confronted with the IBD information as follows: for each site,  
5  
6 pairs of lines in IBD were determined as described above. Genotypes of IBD-related lines were compared  
7  
8 and the numbers of allele matches and mismatches, summed over all IBD pairs, were counted for each  
9  
10 allele present at the site. If the match/mismatch ratio was at least 2 for at least two alleles, or if only one  
11  
12 allele was present in all IBD contrasts, the site is considered as passing the IBD filter. Such a filter is less  
13  
14 powerful for sites where all bases in IBD lines are major allele homozygotes, i.e., the variant being  
15  
16 evaluated occurs in lines not involved in IBD pairs. Formally, such a site passes our IBD filter, but the actual  
17  
18 variant is not strongly confirmed. These uncertain sites, mostly with low minor allele frequency, are  
19  
20 labeled "IBD1" in the HapMap 3 VCF files and constitute about 50% of all HapMap 3 sites.  
21  
22  
23  
24  
25

#### 26 Linkage Disequilibrium (LD) filter

27  
28 Any true SNP should be in local linkage with other nearby SNPs. This observation is the origin of another  
29  
30 filter used in this work, referred to as the LD filter. For each variable site surviving the ST and IBD filters,  
31  
32 we evaluated LD with each site of the GBS anchor. As the LD measure we chose the p-value from a 2 by 2  
33  
34 contingency table of haplotype counts AB, Ab, aB, ab. For the purpose counting haplotypes, heterozygous  
35  
36 genotypes were treated as homozygous in minor allele, so that each taxon only contributed at most one  
37  
38 haplotype. This tends to somewhat strengthen the LD signal and simplify the calculation. For a pair of sites  
39  
40 to be tested for LD, the following three conditions had to be satisfied to make the calculation meaningful:  
41  
42  
43 i) the two sites were at least 2,500 bp apart, ii) there were at least 40 taxa with non-missing genotypes at  
44  
45 both sites being compared, and iii) at least 2 taxa with minor allele had to be present at each of the two  
46  
47 sites.  
48  
49  
50  
51  
52

53 Filtering procedure executed for each site is summarized in Figure 8. First, LD between the given site and  
54  
55 all sites in GBS anchor was computed and up to 20 best LD hits (the ones with lowest p-values) were  
56  
57 collected. If the p-value of the best hit exceeded  $1E-6$  (which roughly corresponds to the peak of the  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 overall distribution of p-values), the site was rejected. Otherwise, it was determined whether the set of  
5  
6 best hits contained any local hits, i.e., hits to GBS sites on the same chromosome within 1 Mbp of the site  
7  
8 in question and with the p-value smaller than 10 times the p-value of the best hit. If no such local hits  
9  
10 were found, the site was rejected, otherwise it was kept and marked as a site in Local LD using the flag  
11  
12 “LLD”. Note that the procedure defined this way filters out sites with only non-local LD hits as well as  
13  
14 those with only weak LD signal. Sites in local LD as well as those for which LD could not be assessed  
15  
16 (because of low minor allele frequency or missing data) pass the filter.  
17  
18  
19  
20

## 21 Imputation

22  
23  
24 In the HapMap 3.2.1 pipeline, the ST- and IBD-filtered genotypes, after the application of the additional  
25  
26 “>1,>2” depth-based filter, were processed through the LD KNN imputation procedure based on Ref. [12]  
27  
28 to fill in the missing data. The procedure is a version of the “K nearest neighbors” routine where the  
29  
30 “nearest neighbors” of a given taxon are selected based on genetic distance computed using variant sites  
31  
32 in good local LD. Specifically, for a given target site, a list of up to 70 sites in best LD (as given by the  $R^2$   
33  
34 measure) with it is compiled by checking all surrounding sites within 600Kb characterized by  
35  
36 heterozygosity lower than 3% and more than 50% taxa with non-missing genotypes. Capping this list at  
37  
38 70 sites leads to good compromise between distance accuracy and computation speed. Then, at the same  
39  
40 target site, for each target taxon, up to 30 “nearest neighbor” taxa are selected, with lowest genetic  
41  
42 distances from the target taxon. Genetic distances are computed using the set of local LD sites selected  
43  
44 in the previous step. Taxa with more than 50% missing genotypes at LD sites, missing genotype at the  
45  
46 target position, having distance from the current taxon larger than 0.1, or resulting in less than 10 common  
47  
48 LD sites on which the distance can be calculated, are excluded from distance calculation process.  
49  
50 Genotypes of the selected nearest neighbor taxa at the target site are stored in memory along with the  
51  
52 genetic distances from the target taxon. This information is used to compute a weight  $w_i$  of each neighbor  
53  
54 genotype  $g$  as follows:  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

$$w_g = \sum_i \frac{1}{1 + 70d_{gi}},$$

where the summation index  $i$  runs over all neighboring taxa with genotype  $g$  at the target site, and  $d_{gi}$  is the distance of taxon  $i$  from the target taxon. The genotype with the highest weight is considered the imputed genotype (of the target taxon at the target site) provided its weight is at least 10 times larger than that of the second-best candidate genotype. Otherwise the imputation is considered inconclusive and the imputed genotype is set to “unknown” (missing data), as it is in the case when no close neighbors of the current taxon could be found. If a genotype imputed to “unknown” occurs at a site where  $MAF < 1\%$ , it is automatically converted into major allele homozygote.

The imputation procedure is run for each genotype in the input file. However, in the output only the originally missing genotypes are updated to imputed ones, whereas all others are left unchanged, even if classified differently. On the other hand, all imputed genotypes are used during a run to collect imputation statistics. The “transition matrix” showing how many genotypes originally in a given class were imputed into other classes is an indication of the accuracy of the input genotypes. Error rates calculated from the transition data are given in Table 3.

## AVAILABILITY OF DATA

At present, reads from all datasets is available in the form of BAM files (with reads aligned to AGP v3 reference) on CYVERSE data store (formerly iPlant, <http://www.cyverse.org/data-store>), in directories `/iplant/home/shared/panzea/raw_seq_282/bam` (dataset “282-2x” and “282-4x”, also available from NCBI BioProject PRJNA389800), `/iplant/home/shared/panzea/hapmap3/bam_germanlines` (“German” dataset; raw reads available from NCBI BioProject PRJNA260788), `/iplant/home/shared/panzea/hapmap2/bam` (“hapmap2” dataset; raw reads available from NCBI BioProject SRP011907), and

1  
2  
3  
4 /iplant/home/shared/panzea/hapmap3/bam (dataset “hmp3.1.1”, also available from NCBI  
5  
6 BioProject PRJNA399729).

7  
8  
9  
10 The set of HapMap 3.1.1. polymorphisms determined for 916 taxa (from datasets “HapMap2”,  
11  
12 “HapMap2 extra”, “CAU”, and “CIMMYT/BGI”) is available in VCF format on CYVERSE data store in the  
13  
14 directory /iplant/home/shared/panzea/hapmap3/hmp311, in files

15  
16  
17 c\*\_hmp311\_q30.vcf.gz (one file per chromosome, where “\*” stands for chromosome 1-10).

18  
19 Additionally, files c\*\_hmp311\_q1.vcf.gz (in the same location) contain test results obtained with  
20  
21 mapping quality threshold equal to 1.

22  
23  
24  
25 The HapMap 3.2.1 variants for 1210 taxa (916 initial Hapmap 3.1.1 taxa + 263 taxa from “282-2x” set +  
26  
27 31 “German” lines) are available from

28  
29 /iplant/home/shared/panzea/hapmap3/hmp321/unimputed, in files

30  
31  
32 mergedflt\_c\*.vcf.gz (un-imputed results) and in

33  
34 /iplant/home/shared/panzea/hapmap3/hmp321/imputed, in files

35  
36  
37 mergedflt\_c\*.imputed.vcf.gz (imputed results).

38  
39  
40 Files c\*\_282\_corrected\_onHmp321.vcf.gz in CYVERSE directory

41  
42 /iplant/home/shared/panzea/hapmap3/hmp321/unimputed/282\_libs\_2015 contain  
43  
44 un-imputed genotypes on HapMap 3.2.1 sites from the full depth data available for the “282” panel (271  
45  
46 taxa, datasets “282-2x” + “282-4x”).

## 47 48 49 50 51 52 53 LIST OF ABBREVIATIONS

54  
55  
56 International Maize and Wheat Improvement Center (CIMMYT), Segregation test (ST), Identity-by-descent  
57  
58 (IBD), Genotyping-by-sequencing (GBS), Linkage Disequilibrium (LD), Minor Allele Frequency (MAF),  
59  
60

1  
2  
3  
4 Nested Association Mapping (NAM), Single Nucleotide Polymorphism (SNP), Insertion-Deletion (Indel),  
5  
6 read mapping quality (MAPQ), base quality (BQ), linkage disequilibrium-base K nearest neighbor  
7  
8 imputation (LDKNN), sites in local LD (LLD), variant call format (VCF)  
9

## 10 11 FUNDING

12  
13  
14  
15 This work has been funded by grants from National Key Basic Research Program of China (2014CB138206),  
16  
17 National Science Foundation of China (Grant #31271736), Bill & Melinda Gates Foundation (Yunbi Xu),  
18  
19 National Science Foundation IOS #1238014, USDA-ARS, and USDA NIFA grant 2009-65300-05668.  
20  
21

## 22 23 COMPETING INTERESTS

24  
25 The authors have no competing interests to declare.  
26  
27

## 28 29 REFERENCES

- 30  
31 1. Chia J-M, Song C, Bradbury PJ, Costich D, Leon N de, Doebley J, et al. Maize HapMap2 identifies extant  
32 variation from a genome in flux. *Nat Genet.* 2012;44:803–7. doi:10.1038/ng.2313.  
33  
34 2. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, et al. A powerful tool for genome  
35 analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC*  
36 *Genomics.* 2014;15:823. doi:10.1186/1471-2164-15-823.  
37  
38 3. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map  
39 of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65. doi:10.1038/nature11632.  
40  
41 4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation  
42 discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43:491–8.  
43 doi:10.1038/ng.806.  
44  
45 5. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity,  
46 diversity, and dynamics. *Science.* 2009;326:1112–5. doi:10.1126/science.1178534.  
47  
48 6. Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of  
49 maize. *Science.* 2009;326:1115–7. doi:10.1126/science.1177837.  
50  
51 7. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. TASSEL-GBS: a high capacity  
52 genotyping by sequencing analysis pipeline. *PLoS ONE.* 2014;9:e90346.  
53 doi:10.1371/journal.pone.0090346.  
54  
55 8. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive  
56 genotyping of the USA national maize inbred seed bank. *Genome Biol.* 2013;14:R55. doi:10.1186/gb-  
57 2013-14-6-r55.  
58  
59 9. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format  
60 and SAMtools. *Bioinformatics.* 2009;25:2078–9. doi:10.1093/bioinformatics/btp352.  
61  
62  
63  
64  
65

10. Flint-Garcia SA, Thuillet A-C, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.* 2005;44:1054–64. doi:10.1111/j.1365-313X.2005.02591.x.
11. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–8. doi:10.1093/bioinformatics/btr330.
12. Money D, Gardner K, Migicovsky Z, Schwaninger H, Zhong G-Y, Myles S. LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3 (Bethesda).* 2015;5:2383–90. doi:10.1534/g3.115.021667.
13. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. Genetic properties of the maize nested association mapping population. *Science.* 2009;325:737–40. doi:10.1126/science.1174320.
14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30:2114–20. doi:10.1093/bioinformatics/btu170.
15. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: building and genotyping Loci de novo from short-read sequences. *G3 (Bethesda).* 2011;1:171–82. doi:10.1534/g3.111.000240.
16. Patefield WM. Algorithm AS 159: An Efficient Method of Generating RXC Tables with Given Row and Column Totals. *Applied Statistics.* 1981;30:91–7.
17. Bukowski R, Guo X, Lu Yanli, Zou C, He B, Rong Z, Wang B, Xu D, Yang B, Xie C, Fan L, Gao S, Xu X, Zhang G, Li Y, Jiao Y, Doebley JF, Ross-Ibarra J, Lorant A, Buffalo V, Romay MC, Buckler ES, Xu Y, Lai J, Ware D, Sun Q: Supporting data for "Construction of the third generation Zea mays haplotype map". *GigaScience Database.* 2017. <http://dx.doi.org/10.5524/100339>

## FIGURES

Figure 1: Overview of HapMap 3 pipeline. Initial set of tentative variant sites was obtained from 916 taxa using reads with mapping quality (MAPQ) at least 30 and bases with base quality (BQ) at least 10. At least 10 taxa had to have non-zero read coverage and the p-value from the segregation test (ST) on allelic depths had to be at most 0.01. This initial set of sites was subject to filtering based on identity by descent (IBD). Application of linkage disequilibrium (LD) filter eliminated sites with only non-local LD hits, leading to HapMap 3.1.1 variant set. An alternative route, leading to HapMap 3.2.1 genotypes, involved K nearest neighbors (KNN) imputation in which distances were computed using sites in good local LD (hence – LD

1  
2  
3  
4 KNN). See text for detailed explanation of methods and acronyms. The exact numbers of variant sites in  
5  
6 HapMap 3.1.1 and HapMap 3.2.1 are 61,228,639 and 83,153,144, respectively.  
7  
8  
9

10  
11 Figure 2: Overlap between various classes of HapMap 3.1.1 polymorphic sites. All sites listed passed the  
12  
13 ST and IBD filters. LLD sites are those found in local LD with the GBS anchor. Sites flagged IBD1 passed the  
14  
15 IBD filter, however, no alternative allele was present in IBD contrasts. Such sites do not violate IBD, but  
16  
17 the existence of a variant is not confirmed. NI5 flag is used to mark indels and sites within 5 bp of an indel.  
18  
19 Since no local re-alignment was done, the NI5 sites are not reliable.  
20  
21  
22  
23  
24

25  
26 Figure 3: Polymorphic sites detected by HapMap 3.1.1 pipeline based on two read mapping quality  
27  
28 thresholds:  $MAPQ \geq 1$  (q1) and  $MAPQ \geq 30$  (q30). Tightening of MAPQ threshold affects mostly the sites  
29  
30 flagged with IBD1 (least reliable), while the LLD sites (in local LD with GBS anchor) are mostly independent.  
31  
32  
33  
34

35  
36 Figure 4: Distribution of inbreeding coefficient for HapMap 3.1.1 variant sets obtained with two read  
37  
38 mapping quality thresholds:  $MAPQ \geq 1$  (q1) and  $MAPQ \geq 30$  (q30). Lower MAPQ threshold leads to lower  
39  
40 values of inbreeding coefficient (i.e., higher heterozygosities) resulting from misaligned reads.  
41  
42  
43  
44

45  
46 Figure 5: Overlap between HapMap 3.1.1 and HapMap 3.2.1 variant sites. 86% of HapMap 3.1.1 sites (99%  
47  
48 of those in local LD) are recovered by the HapMap 3.2.1 pipeline.  
49  
50  
51

52  
53 Figure 6: Distribution of fraction of heterozygous sites per taxon for un-imputed and imputed HapMap  
54  
55 3.2.1. Curves marked LLD have been obtained considering only sites verified in HapMap 3.1.1 to be in  
56  
57 good local LD with GBS anchor.  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 Figure 7: Cumulative distribution of mapping quality from BWA mem alignment of 125.4 million 150 bp  
5  
6 reads from taxon A272.  
7  
8  
9

10  
11 Figure 8: Linkage Disequilibrium-based filtering flowchart. The procedure eliminates sites with weak or  
12  
13 non-local only LD hits. Sites with good local LD hits as well as those for which LD could not be probed  
14  
15 (because of low MAF) are retained.  
16  
17  
18  
19

## 20 21 ADDITIONAL FILES 22

23  
24 Additional file 1: HapMap3TaxaAndCoverage.xlsx – spreadsheet with a list of all lines used in HapMap 3  
25  
26 with their corresponding coverage  
27

28  
29 Additional file 2: DepthFormatDetails.pdf - details of byte representation and storage format used for  
30  
31 allelic depths  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65