bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Construction of the third generation *Zea mays* haplotype map

Robert Bukowski[1], Xiaosen Guo[2,3], Yanli Lu[4]

Cheng Zou[5], Bing He[2], Zhengqin Rong[2], Bo Wang[2], Dawen Xu[2], Bicheng Yang[2], Chuanxiao Xie[5], Longjiang Fan[6], Shibin Gao[4], Xun Xu[2], Gengyun Zhang[2], Yingrui Li[2], Yinping Jiao[7], John Doebley[8], Jeffrey Ross-Ibarra[9], Vince Buffalo[9], Edward S. Buckler[10,11]

Corresponding authors:

Yunbi Xu[5,12], Jinsheng Lai[13], Doreen Ware[7], and Qi Sun[1]

Authors' affiliations:

[1]Bioinformatics Facility, Institute of Biotechnology, Cornell University, Ithaca, NY 14850

[2]BGI-Shenzhen, Shenzhen 518083, China

[3]Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

[4]Maize Research Institute, Sichuan Agricultural University, Wenjiang 611130, Sichuan, China

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

[6]Institute of Crop Science and Institute of Bioinformatics, Department of Agronomy, Zhejiang University, Hangzhou 310058, China

[7]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA

[8]Department of Genetics, University of Wisconsin, Madison, Wisconsin, USA

[9]Department of Plant Sciences, University of California, Davis, California, USA

[10]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853

[11]US Department of Agriculture-Agricultural Research Service, Ithaca, NY 14853

[12]International Maize and Wheat Improvement Center (CIMMYT), El Batan 56130, Texcoco, Mexico

[13]National Maize Improvement Center, China Agricultural University

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ABSTRACT

Characterization of genetic variations in maize has been challenging, mainly due to deterioration of collinearity between individual genomes in the species and the fact the B73 genome used as the reference only represents a fraction of all haplotypes. An international consortium of maize research groups combined resources to develop the maize haplotype version 3 (HapMap3), built from whole genome sequencing data from 916 maize lines, covering pre-domestication and domesticated *Zea mays* varieties across the world. A new computational pipeline was set up to process over 7 trillion bp of sequencing data, and a set of population genetics filters were applied to identify over 60 million variant sites in regions where collinearity is largely preserved in the maize species.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

INTRODUCTION

Maize, one of the most important cereals in the world, also happens to be among the crop species with the most genetic diversity. Advances in the next generation sequencing technologies made it possible to characterize genetic variations in maize at genomic scale. The previously released maize HapMap2 were constructed with whole genome sequencing data of 104 maize lines across pre-domestication and domesticated *Zea mays* varieties [1]. Since then, more maize lines have been sequenced by the international research community, and a consortium was formed to develop the next generation haplotype map. The maize HapMap3 consortium includes, among others, BGI-Shenzen, Chinese Academy of Agricultural Sciences, China Agricultural University, International Maize and Wheat Improvement Center (CIMMYT). Together, a total of 916 maize lines were sequenced with depth varying from below 1x to 46x.

A common approach in today's genetic diversity projects is to map the shotgun sequencing reads from each individual onto a common reference genome to identify DNA sequence variations, and the physical positions of the reference genome is used as a coordinate system for the polymorphic sites. A good example is the human 1000 genome project [2]. The computational data processing pipeline developed for the human project, GATK, has been widely adopted for identifying genetic variations in many other species [3].

As the sequencing technology is improved and sequencers' base calling error model gets more accurate, the computational challenges in genotyping by short-read sequencing have shifted from modeling sequencer machine artifacts errors to solving genotyping errors derived from incorrect mapping of short reads to the reference genome. The problem is associated with the experimental design that uses the single reference genome as coordinate system. Taking maize as an example, the reference being used is a 2.1 Gb assembly from an elite inbred line B73 that represents 91% of the B73 genome [4], and was estimated to capture only ~70% of the low-copy gene fraction of all inbred lines [5]. The sequence alignment software, however, can map 95-98% of the whole genome sequencing reads to the reference. That suggests a high percentage of the reads were mapped incorrectly, either being mapped to the paralogous loci or highly repetitive regions under-represented in the reference assembly. The genetic variants called from the miss-mapped reads need to be corrected computationally. The maize HapMap2 relied on linkage disequilibrium in the population to purge most of the bad markers caused by alignment errors. To construct maize HapMap3, a new computational pipeline was developed from scratch to handle the sequencing data from 10 times more lines, and also took advantage of the high quality genetic map constructed from the GBS technology [6,7] which was not present when HapMap2 was constructed.

Genome structure variation in the population, including transposition, deletion, duplication and inversion of the genomic segments, poses another challenge in the HapMap projects. As the physical genomes of each of the individuals included in the HapMap projects vary both by size and structure, and there is no co-linearity between the reference and genomes of each of the individuals, it is not always possible to anchor all genetic variants in a population onto a single reference coordinate system. As a compromise, markers included in the maize HapMap are defined as sites of which the physical positions of the B73 alleles matching the markers' consensus genetic mapped positions.

Here we present maize haplotype map version 3 (HapMap3), which is a result of coordinated efforts of the international maize research community. The build include 916 lines, about 60 million variant sites, and anchored to the B73 reference genome version AGP v3.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

RESULTS

A total of 7,191 billion base pairs sequencing data from a total of 74,643 million Illumina paired-end reads were aligned to maize reference genome B73 version AGP v3 using BWA mem aligner [8]. Each of the 916 maize lines were sequenced at depth varying from 2x to 60x (Figure 1), using reads of lengths ranging from 44 through 200bp, averaging 102 bp. Overall, 95-98% of the reads were aligned to the reference genome, although only about 50-60% with non-zero mapping quality.

Polymorphic sites were called in a three-step process, summarized in Figure 2. 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 at or above 30. Only sites where at least 10 taxa had coverage of 1 or more were considered. The allelic depths were subject to segregation test (ST – see next section), which determines the probability that a given distribution of allelic depths over taxa has been obtained by chance. Sites with high probability, which are likely a result of random sequencing errors, have been eliminated by applying a p-value threshold of 0.01. In this first round, a total of 196 million tentative polymorphic sites were selected. In the second step, these sites were filtered using the identity by descent (IBD) information derived from about 0.5 million of high-quality polymorphisms obtained previously [7] using the GBS approach [6]. These GBS variants (GBS anchor) were used to determine regions of IBD, where certain pairs of taxa are expected to have identical haplotypes. The raw tentative polymorphisms violating these IBD constraints were then filtered out, leaving 95.4 million sites. At roughly half of the sites surviving this filter, minor allele was not present in IBD contrasts. Such sites, typically with low minor allele frequency, are less reliable and have been marked with “IBD1” flag in the VCF files. In the third and final filtering step, each of these sites was checked for linkage disequilibrium with the GBS anchor. Sites giving only very weak or only nonlocal (i.e., outside of 1 Mb radius) LD hits were eliminated, which resulted in the final set of 60.4 million polymorphisms. For slightly less than ½ of these sites, LD could not be conclusively calculated due to small 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, 8.5 million are indels or are located near (within 5bp) of an indel. These have been marked with the flag “NI5”. Since no realignment around indels has been performed, most of these site are tentative and have to be treated with caution.

Figure 3 shows overlaps between various classes of variants. 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 minor allele frequency 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 do 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 29.4 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 threshold equal to 1. Comparison of the variant set obtained this way (q1) with our recommended set (q30) is shown in Figure 4. While the overall number of variant sites is approximately independent of the mapping quality threshold, the two pipelines produce significantly different sets of sites, with only about 71.5% of all “q30” sites reproduced by the “q1” pipeline. Closer inspection shows that this variability is due primarily to the IBD1 sites, for which our filtering strategy was the least stringent. On the other hand, the LLD sites, confirmed to be in

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

local LD with GBS anchor, are much more independent of the mapping quality threshold, which confirms high quality of such sites.

Importance of choosing a sufficiently tight mapping quality threshold is apparent from Figure 5, where the distribution of inbreeding coefficient is shown for variant sets obtained with thresholds 1 and 30. The lower threshold results in a large number of miss-mapped reads being counted towards depth, producing overly heterozygous genotypes, especially for highly covered taxa (the peak below 0.8 is due mostly to CIMMYT lines with 10-15x coverage; these lines have higher heterozygosity than other lines which may also contribute to the peak) and thus shifting the curve to the left. Since most HapMap3 taxa are inbred lines, one should expect the true distribution to be contained within peak around 0.95. In view of this, the "q30" result is definitely an improvement over "q1", although a longer than expected tail extending towards the value 0.8 indicates that the HapMap3 variants may contain too many false heterozygotes.

Seemingly heterozygous sites may result from sequencing errors and from misalignments of reads originating from paralogous regions. To investigate this further, we calculated, for each site, the fraction of heterozygous HapMap3 genotypes within a subset of 506 high-coverage taxa (defined as those with more than 50% non-missing genotypes on chromosome 10). In the VCF files, this fraction has been recorded as parameter "FH". At sites for which this parameter exceeds 2-3%, heterozygotes are likely to originate from misalignments, for example, from tandem and ectopic duplications. Such sites constitute 9% of all HapMap3 sites.

After the genotypes for the 916-taxa set were completed, WGS data became available for 263 taxa from the 282 panel. Among those, 47 taxa were included in HapMap3. We used the remaining 216 taxa to estimate the error rate of HapMap3 variants as follows. The 216 new taxa were genotyped on all ~60 million HapMap3 sites as well as on the GBS anchor sites. These genotypes were then used to construct and IBD filter, analogous to that used in the main pipeline, but based only on the new taxa, not included in HapMap3. This filter was then applied to each of the HapMap3 sites and the fraction of rejected sites was recorded. As shown in Figure 8, this fraction decreases with decreasing genetic distance threshold used to define IBD regions from genotypes on GBS sites. In regions of strong IBD (distance 0.001 and below), an average of 1% of sites are rejected, however, this estimate is based on a relatively small number of IBD regions (about 50) and small average number of IBD contrasts per region (about 4). On the other hand, the rejection rate above 8% obtained with IBD threshold of 0.02 is too pessimistic, since with such threshold, a lot of IBD contrasts (on average, 303 contrasts per IBD region) are probably of low quality, so that not every IBD violation observed in tested genotypes represents an error. A reasonable (although not perfect) compromise for the choice of IBD threshold seems to be 0.005, where the number of IBD regions saturates. At this threshold, the fraction of rejected sites is 2.4%, with 220 IBD regions (representing 95% of the genome) and 40 IBD contrasts per region on average. This number drops to 1.5% when only low-heterozygosity sites are considered (see curve marked "FH<2%" in Fig. 8). The true error or HapMap3 is thus likely to be between 1-3%.

The set of polymorphisms determined in this work is available in VCF format on iPlant data store in the directory `/iplant/home/shared/panzea/hapmap3/hmp31`, in files `c*_hmp31_q30.vcf.gz` (one file per chromosome, where "*" stands for chromosome 1-10). Additionally, files `c*_hmp31_q1.vcf.gz` (in the same location) contain test results obtained with mapping quality threshold equal to 1. Files `*_fndrs37.vcf.gz` contain HapMap3 genotypes for the 37 "American" and "Chinese" founder lines, extracted from the complete 916-taxa files.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

DISCUSSION

As a species start to diverge, genomic collinearity between individuals deteriorates. The loss of collinearity is by far the biggest challenge to the construction of haplotype maps. The highly repetitive genomic regions are in general easy to identify, because the templates of these repeats are well represented on the reference genome, and sequencing reads mapped to these regions are flagged with low mapping quality, and can be removed at the early stage of the analysis pipeline. For HapMap3, reads with mapping quality lower than 30 were not included in the build. The loss of collinearity of the low copy genomic regions, however, causes vast majority of the genotyping errors, and is not easy to identify computationally, especially for the data sets with limited sequencing depth, which is the case for the maize HapMap3 project.

The biggest issue in loss of collinearity is the deletion of genomic segment in the individual that is used as the reference genome. The sequencing reads derived from these regions, instead of being removed, would be mapped to other, paralogous regions of the reference genome by the alignment software. In this study, 95%-98% of the reads were mapped by BWA to the reference genome, many of these incorrect mappings result in false positive variant sites. In the human 1000 genome project, a new HaplotypeCaller was used [3], which performs local haplotype re-assembly to identify the two most likely haplotypes for each individual and thus improve the genotyping results. However, HaplotypeCaller is computationally very expensive, and not always applicable in the species like maize, where reference genome misses much more haplotypes of the pan-genome and has much more paralogous duplications than human. To construct HapMap3, we relied on the *Zea* GBS map [6,7], which was constructed from GBS SNP markers mostly located in un-methylated chromosomal regions. GBS results was used to identify Identity-by-descent (IBD) regions between the individual genomes, and 101 million markers with high percentage of mismatching genotype callings in the IBD regions were removed from the initial set of 196 million markers.

As the goal of HapMap3 is to identify genetic markers in regions where collinearity is preserved in majority of the maize lines. The LD filter in the pipeline was applied for this purpose. To do this, we genetically mapped the presence/absence of the minor alleles using the GBS genetic map, and these mapped genetic positions were compared to the physical positions on the B73 reference. Among the 95.4 million sites surviving the IBD filter, 25% did not have enough non-missing data or sufficient minor allele frequency for genetic mapping to be meaningful. For the rest of the sites, 51% have at least one genetically mapped position matching the physical positons on B73 reference, 32% have no significant hits from genetic mapping, probably due to no consensus positions in the HapMap3 population, and 17% have genetic positions not matching the B73 physical positions.

The two major filters applied in the HapMap3 project effectively remove majority of the false positive genetic variants caused by paralogous genomic regions, as well as markers with lost collinearity between the species. However, not all the genotyping errors have been removed from the release. 23489911 of the sites do not have sufficient minor allele frequency for genetic test (these are missing the “LLD” label in the INFO field of the VCF files). Another source of errors are paralogous regions evolved from tandem duplications. Given enough sequencing depth, the tandem duplications can be identified either as copy number variation or imputation errors. However, majority of the HapMap3 lines have very low sequencing depth, and fail to sample all paralogous loci or all alleles, which makes it difficult to flag all sites complicated by tandem duplications.

When constructing the maize HapMap3, most serious the problems we were facing can be attributed to the use of a genome from a single individual (B73) as a reference for other, often very different species.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

This is becoming the single limiting factor in the study of maize diversity, as well as breeding practice. The only remedy is to move away from a single genome-based reference coordinate and adopt a pan-genome based reference system that incorporates all major haplotypes of the species.

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 [9], 23 maize landraces and 19 wild relatives (teosinte lines, 17 *Z. mays ssp. parviglumis* and 2 *Z. mays ssp. mexicana*). Majority of the remaining inbred lines originated from CAU and include, among others, "Chinese NAM" parent lines. Additional 89 inbred lines were provided by CIMMYT and sequenced at BGI. During the course of this work it was discovered that new sequence marked as originating from line CML103 actually represents material different (significantly more heterozygous) from the line with the same name studied previously in HapMap2 project. This new sequence has been treated as coming from a separate line. Also, the Mo17 sequence originating at CAU has been treated as taxon separate from Mo17 and CAUMo17. The HapMap3 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 HapMap3 project was 916.

Sequencing

Sequencing has been performed over several years using various generations of Solexa/Illumina instruments and library preparation protocols, giving paired end reads from 44 to 200 bp long. Overall, 74,231 million reads were obtained, containing 7,129 billion base pairs, giving on average 3.4x coverage per line (assuming 2.3 Gb genome size). However, as shown in Figure 1, coverage was not uniform among all lines. For a few lines, sequence generated previously in the context of HapMap2 project was augmented with reads from recent re-sequencing which brought the median coverage of the HapMap2 lines to 5x, with average coverage equal to 7.8x and standard deviation of 7.2x. All NAM parent lines are covered to 10x or better. Most of the additional 89 lines provided by CIMMYT and sequenced at BGI have coverage exceeding 10x. Majority of the remaining inbred lines originated from CAU and have been sequenced at a lower coverage (1-2x). The list of all lines used in HapMap3 with the corresponding coverage is given in supplementary Table S1.

Alignment

Due to the use of different versions of Solexa/Illumina sequencing equipment, the base qualities in different input FASTQ files are given in different encodings. Prior to alignment, all base qualities have been converted to phred+33 scale. Reads were then aligned to B73 reference (AGP v3) as paired-end using bwa mem aligner (1) with default options. In 72 read sets (Illumina lanes), for technical reasons a high (6%-54%) fraction of paired-end fragments were found to be shorter than reads, so that the two ends contained a part of Illumina adapter and were reverse complements of each other. For such "read-through" fragments, the remnants of Illumina adapter sequences were clipped using TRIMMOMATIC [10] and only one read was used and aligned as single-end. The bwa mem aligner is capable of clipping the ends of reads and splitting each read in an attempt to map its different parts to different location on the

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

reference. As a result, typically over 95% of reads are reported as mapped. However, the fraction of reads with non-zero mapping quality (negative log of the probability that a read has been placed in a wrong location) is much lower – typically only 40-50%. Figure 6 shows a typical distribution of the mapping quality obtained from bwa mem alignment. In practice, we only used alignments with mapping quality of at least 30. A base was counted towards allele depth if its base quality score was at least 10.

It is well known that alignment may be especially ambiguous when reads contain indels with respect to the reference. In such cases, multiple-sequence realignment approaches have been proposed [3] to find the correct sequence and location of an indel and avoid spurious flanking SNPs. Since indels are not the primary focus of this work and since the realignment is computationally very expensive, it has not been performed by the HapMap3 pipeline. Thus, although indels and SNPs in their immediate vicinity have been retained in the HapMap3 VCF files, they are less reliable and have therefore been marked with "NI5" label for easy filtering.

Genotyping pipeline

Raw genotypes were obtained using a custom-built multi-threaded java code. First, the code executes samtools mpileup command (thresholds on the base and mapping quality are imposed here) for each taxon individually, processing a certain portion of the genome. On a multi-core machine, several such pileup processes (i.e., for several taxa) can be run concurrently as separate threads. Since we are predominantly interested in calling SNPs, we use a simplified indel representation where insertions and deletions with respect to reference are treated as additional alleles "I" and "D", respectively, regardless of length and actual sequence of the indel. Read depths and average base qualities of all six alleles (A, C, G, T, I, and D) are extracted from samtools mpileup output for each taxon at each genomic position and stored in an array shared between all threads. The amount of memory available on the machine along with the number of taxa determine the upper limit on the size of this array, and therefore – the maximum size of chromosome chunk which can be processed at one time. As base quality of I and D alleles we took the value corresponding to the base directly preceding the indel on the reference.

Extraction of allelic depths for all genomic positions is time consuming, which presents a major obstacle if joint genotyping needs to be re-run, for example, upon extending the taxa set (the so-called "N+1 problem"). It is therefore advantageous to run the depth extraction only once for each taxon and save the obtained depths on disk to be retrieved (rather than re-calculated) during the genotyping step. This way, when the taxa set for genotyping is extended, mpileup step has to be run only for the newly added taxa. Thus, the program features an option to save allelic depths and average qualities in specially designed data structures stored in HDF5 files – one such file per taxon per chromosome. To save space, each allele depth and average quality is stored as one byte, which allows exact representation of integers from 0 to 182, while higher integers (up to about 10,000) are represented approximately by negative byte values through a logarithmic formula with carefully chosen base. Depths and qualities are stored only for sites with non-zero coverage. The details of the storage format and integer representation in terms of byte variables are given in Supplementary Material.

Once the allelic depths for all taxa and a given chunk of the genome are available in shared memory, each site is evaluated for presence of a tentative SNP. On a multi-core machine, the set of sites within the genome chunk is divided into subsets processed in parallel on different cores. Sites with less than 10 taxa with read coverage and those with only reference allele present are ignored. For all other sites, genotypes are called for all taxa using a simple likelihood model with a uniform error rate [11] assumed at 1%. Alternative alleles are then sorted according to their allele frequencies and up to two most abundant

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

alleles are kept, as decided by the segregation test described in the next Section. Sites for which all taxa turn out to be reference homozygotes (which may happen despite non-reference alleles being present in the mapped reads) are skipped. Raw variant set obtained in this way is then subject to extensive filtering with the intention of reducing the number of false positives resulting from misalignments.

Filtering

Segregation test (ST) filter

For each pair of alleles obtained in the genotyping step, a 2 by N (where N is the number of taxa) contingency table is constructed, containing depths of the first allele in row 1 and depths of the second allele in row 2. The Fisher exact test (FET) is then performed to assess how likely such a table is to occur by chance. If the expected values of the array elements are sufficiently large, the p-value from FET is approximated by that from the computationally efficient chi-square test. However, in most cases encountered here, expensive simulation is needed to obtain sufficiently accurate p-value. To reduce computational burden we adopted a hybrid approach based on an empirical observation that for statistically insignificant cases (p-values larger than 0.2) the chi-square test results in a de facto lower bound to exact p-values. Thus, the chi-square test is performed first for each site and if the p-value from this test is below 0.2, more exact p-value is obtained from a simulation procedure. The simulation procedure used here, implemented in Java, is the same as the one implemented in R package [12]. An alternative allele is kept if at least one contingency table involving this allele has p-value smaller or equal to 0.01. If none of the alternative alleles survive the ST filter, the site is skipped (not reported in output). The ST filter tends to eliminate variant sites resulting from random sequencing errors.

GBS anchor map and IBD filter

Given a set of trustworthy SNPs and a diverse set of 916 taxa it is possible to identify, for an arbitrary region of the genome, a number of taxa pairs which are identical by descent (IBD) and are therefore expected to have identical genotypes in this region. If known, these IBD pairs can be used as a powerful filter eliminating variant which violate IBD constraints.

To determine the IBD regions, we used the first step of our pipeline to call genotypes for our 916 taxa on the set of GBS v2.7 sites [6,7] which tend to concentrate in relatively well-conserved low-copy regions of the genome and can therefore be considered reliable. This set of 954,384 sites was filtered to include only SNP (not indel) sites for which the p-value from the segregation test was below 0.05 and which were more than 5 bp away from any indel. The set of genotypes at 475,272 sites obtained in this way, which will be referred to as **GBS anchor**, agree well with those from GBS on 167 taxa present in both sets. Alleles detected by the HapMap3 pipeline agreed with those from GBS at 94% of the GBS sites. At 90% of the sites, fraction of (non-missing data) taxa with genotypes in agreement with those from GBS was at or above 85%. Genotypes different from GBS ones were observed for 82 taxa. These differences were most frequent (up to 19% of all sites) for teosinte lines.

The GBS anchor was used to compute the genetic distance (identity by state) between any two of the 916 lines in windows containing 2000 GBS sites each (about 8.5 Mbp on average). If the genetic distance within such a window was <= 0.02 (about 10 times smaller than the mean distance across all pairs), the two lines

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

were considered to be in IBD. At least 200 comparable GBS sites (i.e., non-missing data simultaneously on both lines being compared) were assumed necessary to make the genetic distance calculation feasible.

The number of taxa involved in IBD relationships in any given window were between 385 (start of chromosome 10) and 757 (middle of chromosome 7) and averaged 588, leading to large numbers of IBD contrasts, ranging from 3,710 (beginning of chromosome 4) to 42,890 (middle of chromosome 7), and averaging 13,500.

The raw (ST-filtered) genotypes were checked against the IBD pairs in various regions, using a procedure which counts, for each site, numbers of base matches and mismatches for each allele present at the site. If the match/mismatch ratio is at least 2 for at least two alleles, or if only one allele is present in all IBD contrasts, the site is considered as passing the IBD filter. Such a filter is less powerful for sites where all bases in IBD lines are major allele homozygotes (i.e., the SNP being evaluated occurs in lines not involved in IBD pairs). Formally, such a site passes IBD filter, but this is statistically easier to achieve than agreement involving minor alleles, so the actual SNP is not strongly confirmed. These uncertain sites, mostly with low minor allele frequency, are labeled “IBD1” in the HapMap3 VCF files and constitute about 50% of all HapMap3 sites.

Linkage Disequilibrium (LD) filter

Any true SNP should be in local linkage with other nearby SNPs. This observation gives rise to another filter used in this work, referred to as the LD filter. For each variable site surviving the ST and IBD filters, we evaluate LD with each site of the GBS anchor. As the LD measure we chose the p-value from a 2 by 2 contingency table of taxa counts corresponding to the four haplotypes (AB,Ab,aB,ab). For simplicity, heterozygous genotypes were treated as homozygous in minor allele. For a pair of sites to be tested for LD, the following three conditions had to be satisfied to make the calculation meaningful: i) the two sites were at least 2,500bp apart, ii) there were at least 40 taxa with non-missing genotypes at both sites being compared, and iii) at least 2 taxa with minor allele had to be present at each of the two sites.

Filtering procedure executed for each site is summarized in Figure 7. First, LD between the given site and all sites in GBS anchor was computed and up to 20 best LD hits (the ones with lowest p-values) were collected. If the p-value of the best hit exceeded 1E-6 (which roughly corresponds to the peak of the overall distribution of p-values), the site was rejected. Otherwise, it was determined whether the set of best hits contained any local hits, i.e., hits to GBS sites on the same chromosome within 1 Mbp of the site in question and with the p-value smaller than 10 times the p-value of the best hit. If no such local hits were found, the site was rejected, otherwise it was kept and marked as a site in Local LD using the flag “LLD”. Note that the procedure defined this way filters out sites with only non-local LD hits as well as those with only weak LD signal. Sites in local LD as well as those for which LD could not be assessed (because of low minor allele frequency or missing data) pass the filter.

Acknowledgments

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), and National Science Foundation IOS #1238014 and the USDA-ARS.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

REFERENCES

1. Chia J, 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: 803–807. doi: 10.1038/ng.2313.
2. 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.
3. 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: 491–498. doi: 10.1038/ng.806.
4. 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: 1112–1115. doi: 10.1126/science.1178534.
5. Gore MA, Chia J, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, et al. A first-generation haplotype map of maize. Science. 2009; 326: 1115–1117. doi: 10.1126/science.1177837.
6. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, et al. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE. 2014; 9: e90346. doi: 10.1371/journal.pone.0090346.
7. 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.
8. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. doi: 10.1093/bioinformatics/btp352.
9. 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: 737–740. doi: 10.1126/science.1174320.
10. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120. doi: 10.1093/bioinformatics/btu170.
11. 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: 171–182. doi: 10.1534/g3.111.000240.
12. Patefield WM. Algorithm AS 159: An Efficient Method of Generating RXC Tables with Given Row and Column Totals. Applied Statistics. 1981; 30: 91–97.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Material

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

BYTE REPRESENTATION OF ALLELIC DEPTHS

Most of the per taxon allelic depths encountered in the HapMap3 datasets are lower than 100. It is thus possible, without significant loss of accuracy, to represent these depths as byte variables rather than integers and therefore save disk space. Depths up to 127 can be represented directly as byte variables, whereas for higher depths one can utilize negative byte values as follows:

$$B = \begin{cases} I & \text{for} \quad I \leq 127 \\ 127 - I & \text{for} \quad 127 < I \leq M \\ \mathrm{Int}(\max[-\log_b(I-o), -128]) & \text{for} \quad I > M \end{cases}$$

The inverse transformation is as follows:

$$I = \begin{cases} B & \text{for} \quad B \geq 0 \\ 127 - B & \text{for} \quad 0 > B \geq 127 - M \\ \mathrm{Int}(o + b^{\frac{1}{2}-B}) & \text{for} \quad B < 127 - M \end{cases}$$

The operator Int() truncates the real value to the smallest integer and the parameters b, o, and M were chosen as 1.0746, 126, and 182, respectively. With this choice, all the byte values map to unique integers and encoding is exact not only for depths from 0 to 127, but also somewhat beyond (up to $M = 182$). Determining M is a bit tricky, as one needs to make sure the corresponding negative byte values won't be also needed in the exponential approximation. For base $b = 1.0746$, we found that bytes -1 through -55 can be used directly, allowing exact encoding for depths up to 127+55=182. For depths slightly above 182 the relative error envelope is about 1%, then it grows to 3% for depths around 1,000 to asymptotically reach 3.5% for larger depths (Figure S1). The asymptotic error rate is determined by the base b and equals $(b-1)/2$. The value of b also determines the largest depth that can be effectively approximated. In our case, it is 10,117 approximated as 10,482 (all larger depths are encoded as -128 and then decoded as 10,482). Decreasing the value of b would lead to a better approximation, but smaller maximum representable depth. For example, using 1.05545 allows for depths only up to about 1000, but values up to 210 would be encoded exactly and asymptotic relative error would be 2.5% instead of 3.5%.

STORAGE MODEL FOR ALLELIC DEPTHS AND AVERAGE BASE QUALITIES

The read depths for all six alleles extracted for each taxon and genomic position using samtools mpileup are stored on disk to be retrieved during the joint genotyping step. As shown in Fig. S2, the array of allelic depths is usually sparse, especially at low coverage levels, when only one allele (or not at all) is present at each position. To save disk space, the following scheme was used to represent this array within HDF5 structure. Each HDF5 file (one per taxon per chromosome) consists of three byte arrays. Each byte of the first array represent a position on the chromosome (all positions are represented) and shows – through set bits - which alleles have non-zero depths at this position. The second array stores

the byte-encoded non-zero allelic depths. The genomic position and the allele represented by each entry of this array is determined by non-zero bits of the first array, as shown in Fig. S2. The average base quality scores are stored in a similar way in the third array.

Effectiveness of this compression scheme depends on the sparsity of the original array and is the highest for taxa with low coverage.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures and Tables

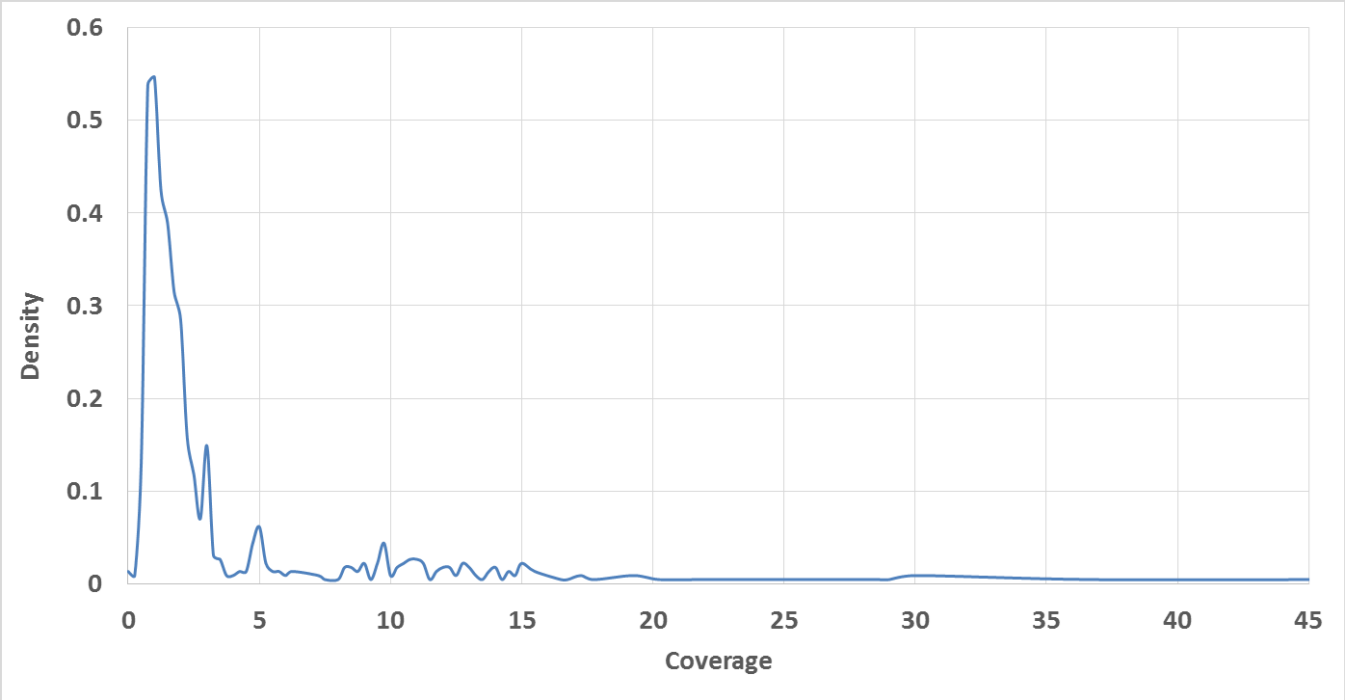

Figure 1: Distribution of nominal coverage (assuming 2.3 GB genome size) among 916 taxa used in HapMap3.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

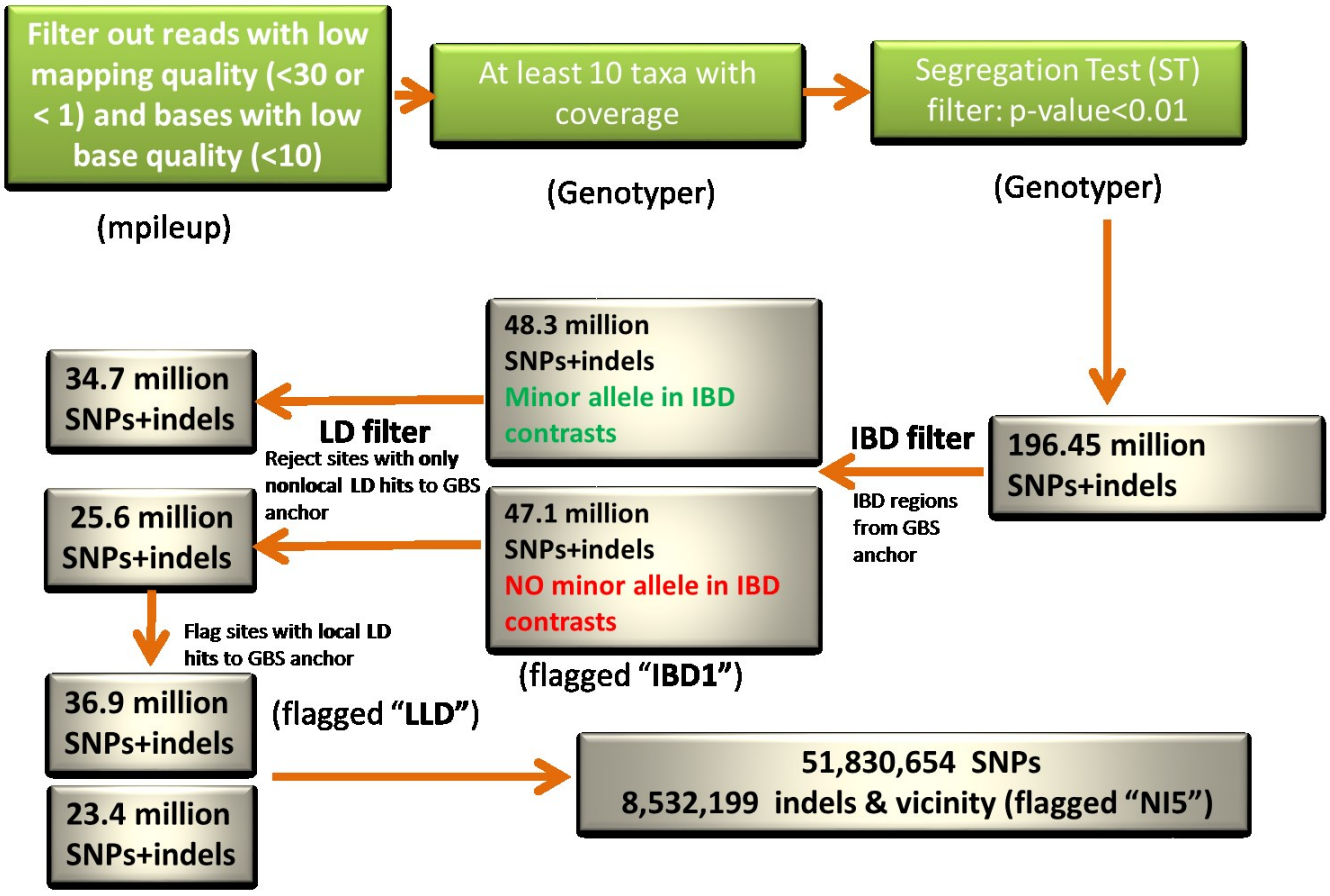

Figure 2: HapMap3 pipeline and the effect of filtering on the number of variants detected.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

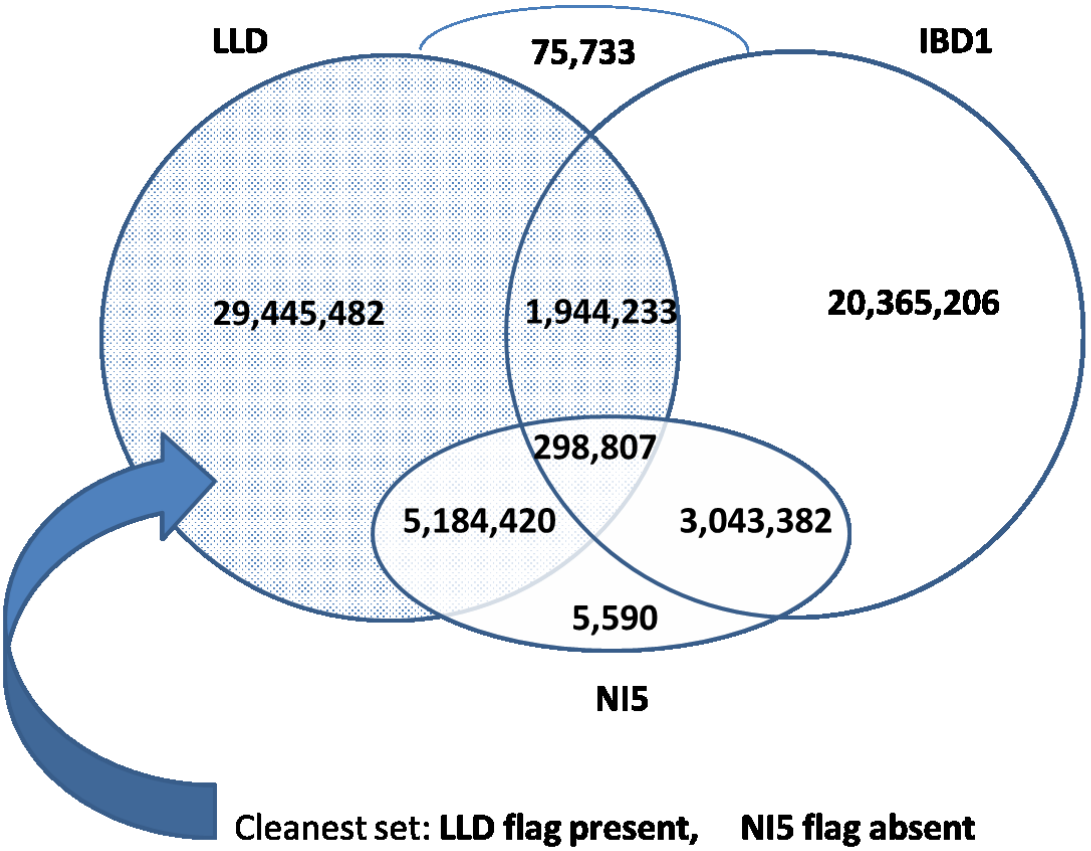

Figure 3: Overlap between various classes of HapMap3 polymorphic sites.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

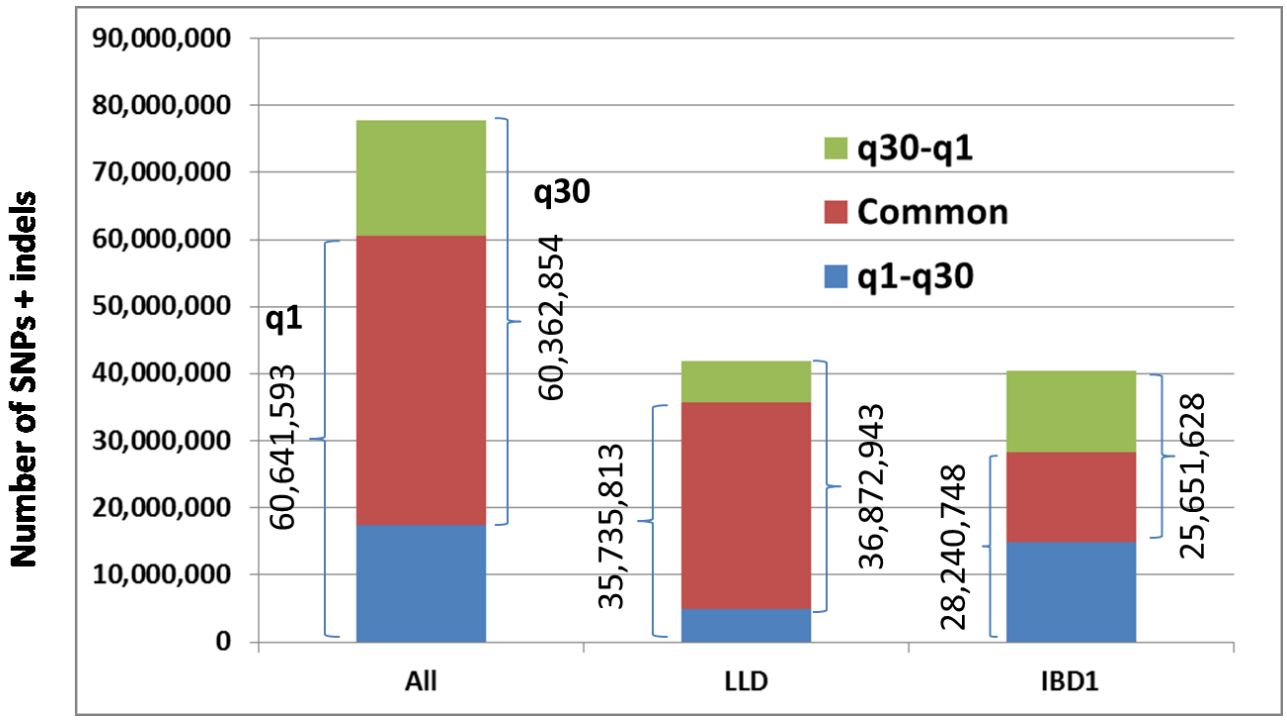

Figure 4: Polymorphic sites detected by HapMap3 pipeline based on two read mapping quality thresholds (q=30 and q=1).

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

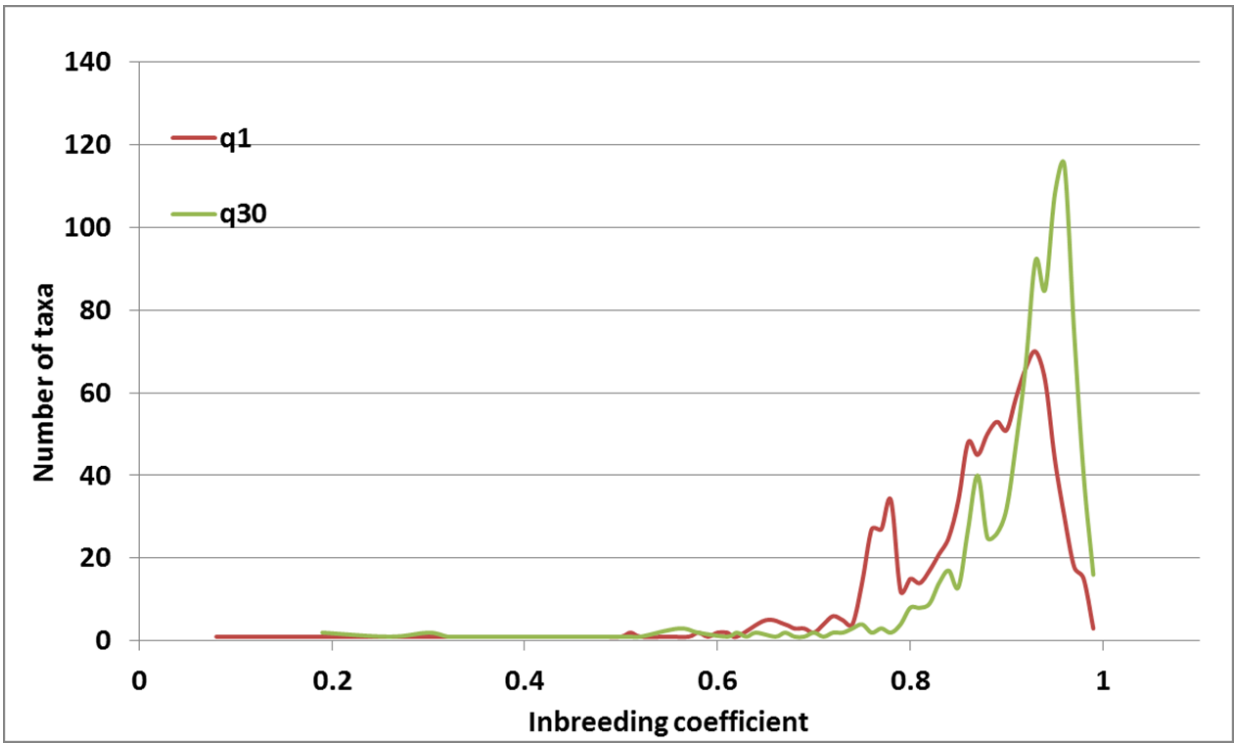

Figure 5: Distribution of inbreeding coefficient for variant sets obtained with two read mapping quality thresholds (q=30 and q=1).

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

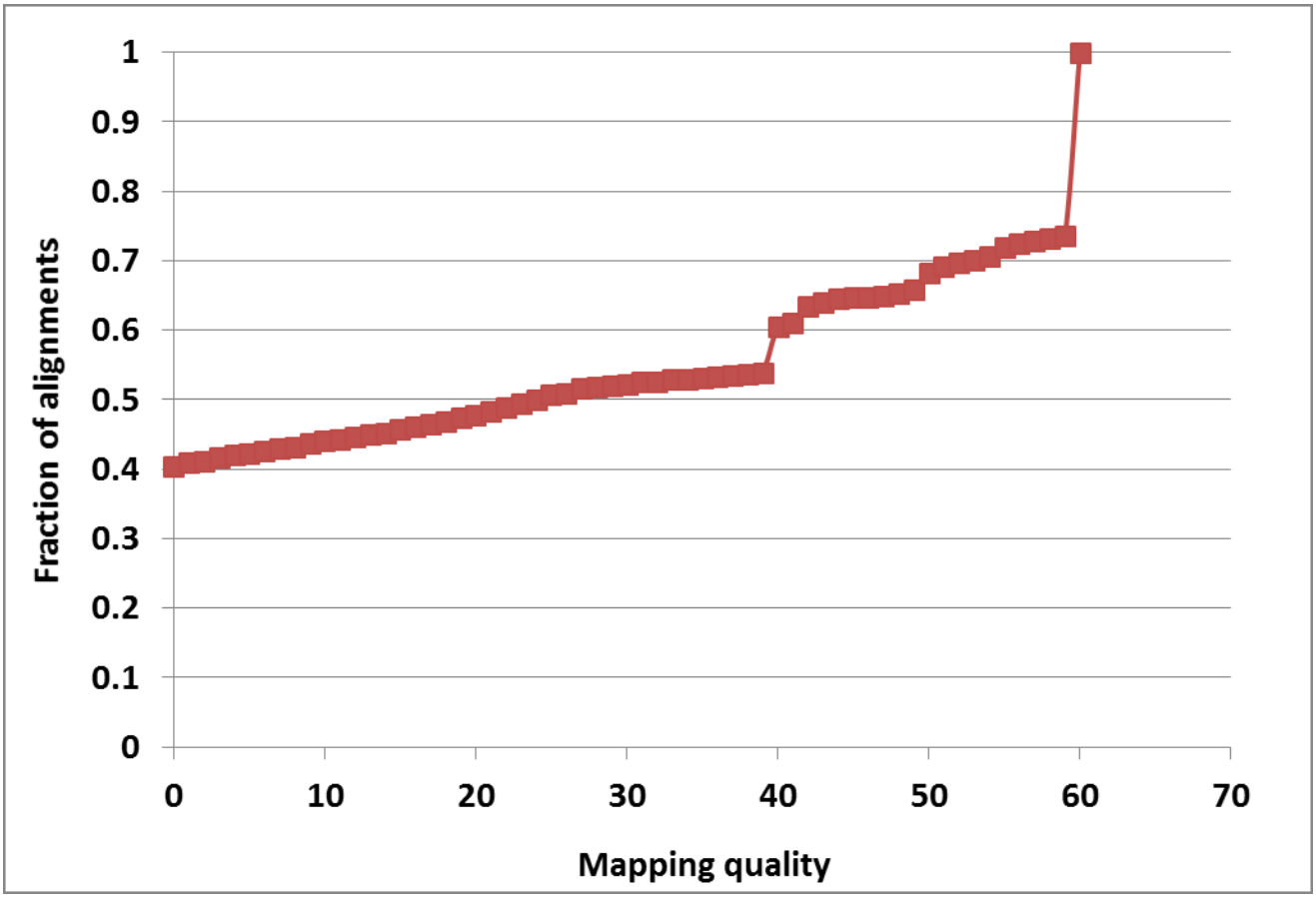

Figure 6: Cumulative distribution of mapping quality from BWA mem alignment of 125,441,950 150bp reads from line A272.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

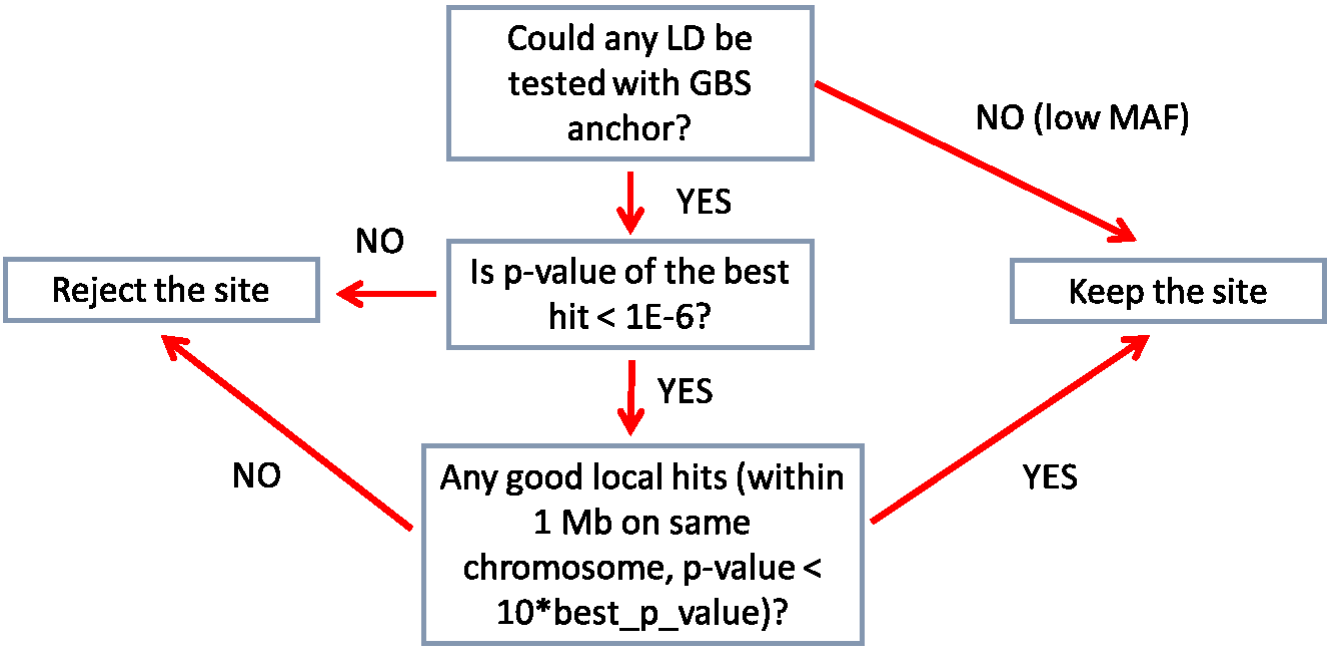

Figure 7: Linkage Disequilibrium-based filtering flowchart

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

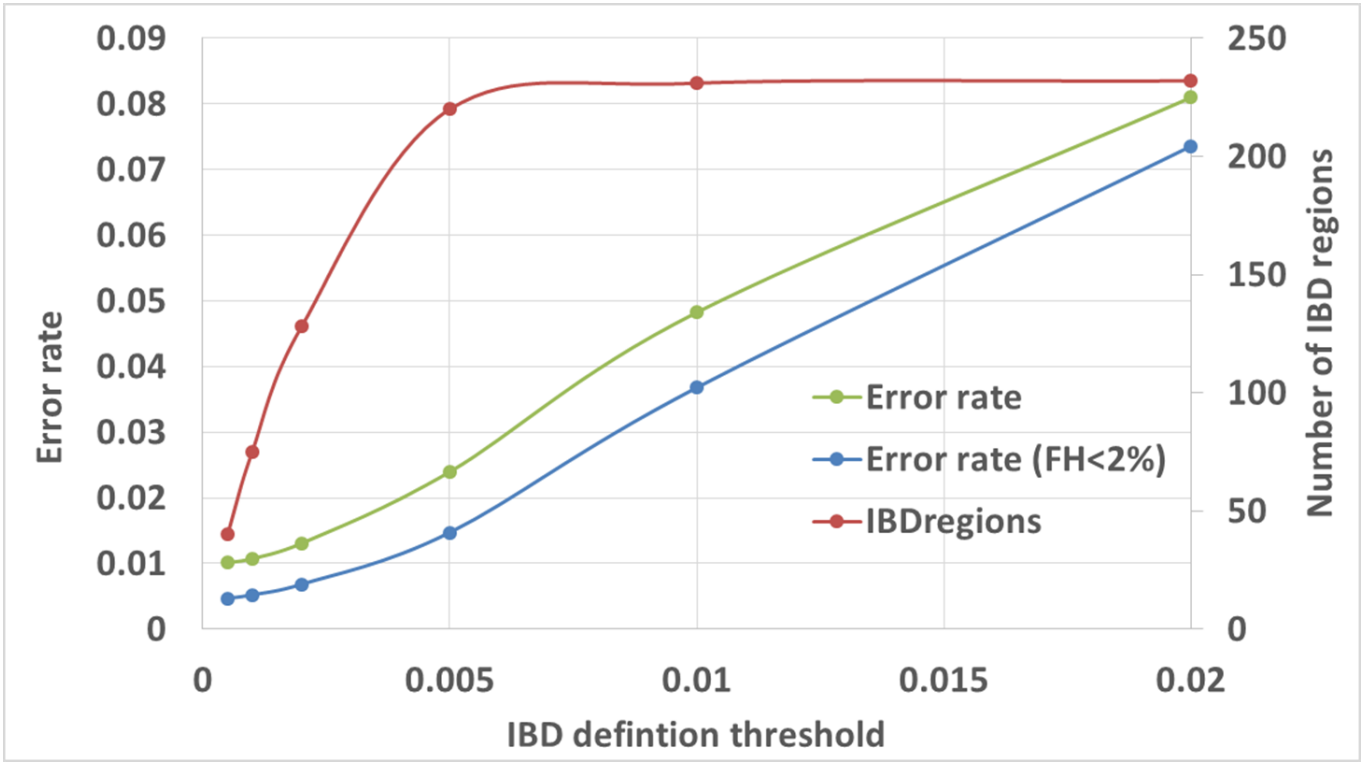

Figure 8: Error rate estimate of HapMap3. The FH<2% curve has been obtained using only sites for which the frequency of heterozygotes computed over 506 high-coverage taxa set was less than 2%.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

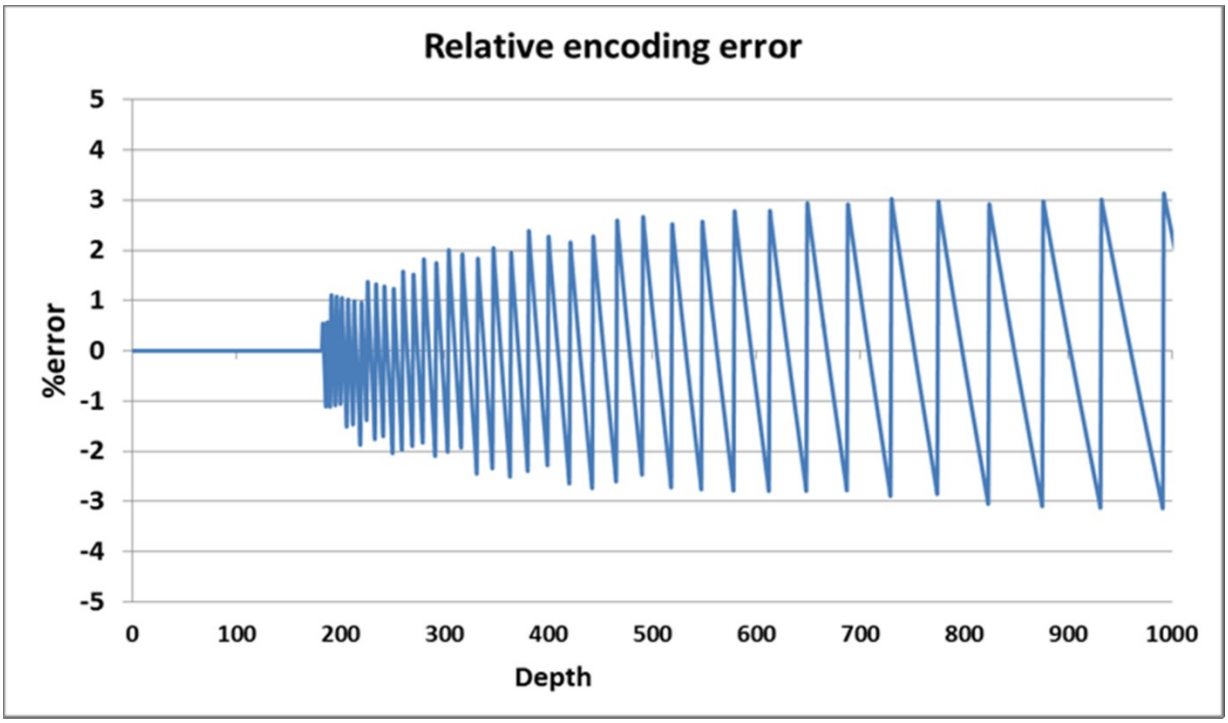

Figure S1: Relative error of depth encoding as byte variables. All values up to 182 are represented exactly. Encoding values 182 through 10,117 and decoding them leads to error of no more than 3.5%. Integers larger than 1 are encoded as -128 and decoded as 10,482.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

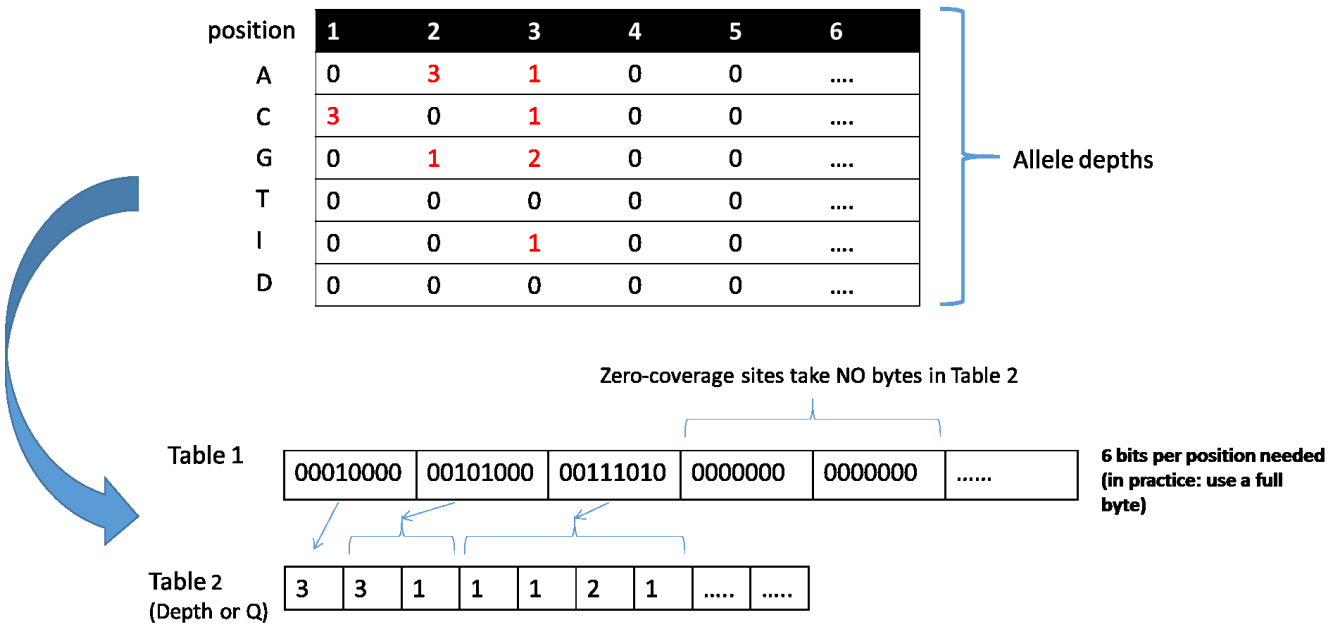

Figure S2: Representing the array of allele depths (or average base qualities) in HDF5 file.

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S1: 916 maize lines included in HapMap3 sorted according to sequencing coverage (assuming 2.3Gbase genome size).

Taxon	Coverage	Taxon	Coverage	Taxon	Coverage	Taxon	Coverage
CML69	46.07	RIMMA0438.1	36.87	CAUCHANG72	34.41	CAUZHENG58	32.99
CAU5003	32.92	CAU478	32.38	MM-1A	27.06	Oh43	26.69
TIL15	25.14	M37W	23.59	KI3	23.01	TIL10	21.82
W22	20.29	ZEAxppRBRDIAAPEI-3	19.27	W64A	17.98	TIL25-TIP489	17.96
TIL05	17.68	MAIdgiRAPDIAAPEI-12	17.35	ZEAxppRBSDIAAPEI-4	17.29	Mo17	17.2
ZEAxppRCODIAAPEI-9	16.75	MAIdgiRAVDIAAPEI-4	16.5	CML103-run248	16.38	MAIdgiRAXDIAAPEI-6	15.44
MAIdgiRAWDIAAPEI-5	15.38	MAIdgiRAYDIAAPEI-7	15.11	TIL03	15.05	MAIdgiRCKDIAAPEI-9	15.05
MAIdgiRABDIAAPEI-2	14.94	MAIdgiRAKDIAAPEI-9	14.9	MAIdgiRARDIAAPEI-1	14.8	MAIdgiRAMDIAAPEI-10	14.64
MAIdgiRCIDIAAPEI-7	14.6	Ms71	14.58	TIL11	14.41	MAIdgiRASDIAAPEI-2	14.26
MAIdgiRAODIAAPEI-11	14.08	ZEAxppRCQDIAAPEI-11	14.03	ZEAxppRBFDIAAPEI-3	13.97	CML52	13.91
MAIdgiRAGDIAAPEI-5	13.87	TIL09	13.73	ZEAxujRANDIBAPE	13.66	TIL14-TIP498	13.41
ZEAxujRAKDIBAPE	13.41	MAIdgiRAEDIAAPEI-4	13.25	MAIdgiRACDIAAPEI-3	13.18	MAIdgiRAIDIAAPEI-6	13.11
ZEAxujRALDIBAPE	13	MAIdgiRAADIAAPEI-1	12.94	ZEAxujRAHDIBAPE	12.89	CML333	12.76
ZEAxujRASDIAAPE	12.75	N6	12.75	HP301	12.74	ZEAxujRAMDIBAPE	12.65
ye8112	12.48	ZEAxujRAODIBAPE	12.46	ZEAxujRAXDIAAPE	12.37	ZEAxujRAYDIAAPE	12.32
ZEAxujRBBDIAAPE	12.22	ZEAxujRBADIAAPE	12.15	ZEAxujRAGDIBAPE	12.03	ZEAxujRARDIAAPE	12.01
ZEAxppRCVDIAAPEI-2	12	ZEAxujRAQDIAAPE	11.91	TIL01-JD	11.85	MO18W	11.78
ZEAxujRAPDIAAPE	11.77	ZEAxujRADDIBAPE	11.63	ZEAxujRAZDIAAPE	11.58	TDD39103	11.3
ZEAxujRATDIAAPE	11.29	ZEAxppRCUDIAAPEI-1	11.23	ZEAxujRAWDIAAPE	11.16	ZEAxujRBCDIBAPE	11.15
ZEAxujRAUDIAAPE	11.04	ZEAxppRCHDIAAPEI-6	11.04	A632	11.04	Tx601	10.95
ZEAxppRCXDIAAPEI-4	10.89	ZEAxppRDHDIAAPEI-12	10.89	ZEAxppRCPDIAAPEI-10	10.86	ZEAxujRAVDIAAPE	10.85
ZEAxujRBDDIBAPE	10.76	CML103	10.76	A272	10.71	ZEAxppRAZDIAAPEI-7	10.71
ZEAxppRCJDIAAPEI-7	10.67	ZEAxppRCFDIAAPEI-4	10.58	ZEAxujRAJDIBAPE	10.53	ZEAxppRBCDIAAPEI-2	10.5
ZEAxppRBUDIAAPEI-5	10.49	ZEAxppRCYDIAAPEI-5	10.39	ZEAxppRAHDIAAPEI-6	10.32	ZEAxppRDLDIAAPEI-2	10.32
MAIdgiRCCDIAAPEI-10	10.31	ZEAxujRBFDIBAPE	10.27	ZEAxppRAQDIAAPEI-11	10.1	ZEAxppRBMDIAAPEI-6	9.96
ZEAxppRANDIAAPEI-10	9.83	ZEAxppRDQDIAAPEI-3	9.8	ZEAxppRDIDIAAPEI-1	9.79	CML277	9.76
ZEAxppRBXDIAAPEI-1	9.7	ZEAxppRDGDIAAPEI-11	9.66	ZEAxppRALDIAAPEI-9	9.65	ZEAxppRDCDIAAPEI-7	9.64
ZEAxppRAFDIAAPEI-5	9.55	ZEAxppRBLDIAAPEI-2	9.55	M162W	9.46	ZEAxppRBHDIAAPEI-1	9.38
ZEAxppRADDIAAPEI-4	9.28	Tx303	9.1	dan340	8.99	Il14H	8.94
Ki21	8.93	Tzi8	8.88	Ki11	8.83	ZEAxppRAJDIAAPEI-7	8.8
Ky21	8.63	B73	8.58	Oh7B	8.58	P39	8.53
CML247	8.45	NC350	8.41	ZEAxppRBYDIAAPEI-2	8.32	huangzaosi	8.31
CML322	8.19	Oh40B	8.05	CML228	7.44	NC358	7.31
CAUMo17	7.22	Mo17-chin	7.17	ZEAxppRATDIAAPEI-12	6.27	zong31	6.17
L578	6.17	D340	5.75	ZEAxppRAUDIAAPEI-1	5.56	ZEAxppRCBDIAAPEI-3	5.54
H127	5.41	CAU178	5.37	ZEAxppRCGDIAAPEI-5	5.27	ZEAhwcRAXDIAAPE	5.19
ZEAxppRCDDIAAPEI-12	5.18	BKN015	5.15	BKN011	5.09	BKN029	5.04
BKN027	5.04	BKN014	5.01	BKN022	5	BKN030	4.99

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

BKN009	4.99	dupl-178	4.98	BKN020	4.97	BKN023	4.94
BKN026	4.94	BKN025	4.92	ZEAxppRCWDIAAPEI-3	4.88	BKN031	4.88
BKN010	4.87	BKN017	4.84	BKN016	4.83	dan9046	4.81
BKN033	4.78	B97	4.72	BKN035	4.71	BKN018	4.71
dan599	4.66	BKN034	4.64	BKN019	4.61	TIL06-TIP260	4.61
BKN032	4.57	CT52C	4.3	F7	4.15	changK	4.13
TIL04-TIP454	4.09	TIL08	4.08	W344	3.67	B95	3.61
78599	3.53	TIL01	3.53	dupl-478	3.52	La2-4	3.47
XF223	3.42	lu65	3.41	TIL07	3.39	D1139	3.36
zhongyin15	3.3	TIL17	3.3	TIL16	3.26	87001	3.2
yu87-1	3.2	shen135	3.15	longkang11	3.09	SC-14	3.09
danhuang02	3.06	T24	3.05	CML133	3.04	CML192	3.04
CML202	3.04	CML206	3.04	CML312SR	3.04	CML330	3.04
CML341	3.04	CML411	3.04	CML418	3.04	CML479	3.04
CML504	3.04	CML505	3.04	CML511	3.04	CML84	3.04
CML85	3.04	CML96	3.04	CML99	3.04	H16	3.04
P1	3.04	VL0512447	3.04	VL05128	3.04	VL054178	3.04
VL05610	3.04	VL056883	3.04	VL062784	3.04	A801	3.02
xiangai3	2.94	ES40	2.9	mei68113	2.89	200-24-13413	2.88
K12	2.87	shuang741	2.86	TIL06-TIP496	2.86	N42	2.85
yue89E4-2	2.83	150-4	2.82	C521	2.82	LP5	2.79
698-3	2.79	dupl-Zheng58	2.77	W668	2.76	SX-6-7	2.75
68139	2.7	C103	2.68	M131-5	2.68	IBB15	2.64
2005	2.63	DM101B	2.62	xi502	2.62	H21	2.62
B98	2.62	Los-6	2.6	WIL900	2.56	Ay420	2.56
K22	2.54	lv28	2.51	huangC	2.5	L473	2.49
LH128	2.47	LH132	2.47	BKN040	2.47	ji846	2.46
qiong51	2.44	N138	2.42	yan156	2.42	LH74	2.42
1145	2.4	LH156	2.4	B12	2.39	68122	2.39
chengzi2142	2.38	1205A	2.38	huangyesi3	2.38	91huang5	2.38
PHR63	2.37	17564	2.36	HHe01	2.36	4F1	2.35
PHN82	2.34	624	2.34	D864	2.34	D15	2.33
LH208	2.33	D857	2.29	zheng32	2.28	85bai64	2.27
luosu3	2.27	D1049	2.27	L005	2.24	GN4095	2.24
ernan24	2.22	XF134	2.21	PHW30	2.2	CML125-2pianma	2.2
FR14	2.2	ye52106	2.2	dupl-5003	2.19	yan172	2.19
tai184	2.18	equn3	2.17	PHN73	2.17	Ay3566	2.17
zong3	2.16	jiu138-5	2.16	897	2.16	fusheA	2.15
ye107	2.15	dan9064	2.15	030-1	2.14	cheng18	2.14
3489a	2.13	D729	2.12	fangyin	2.1	ys06	2.1
PHZ51	2.09	huotanghuang17	2.08	Feb-48	2.07	807	2.07

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

764	2.07	PHP55	2.07	shen137	2.06	79028	2.06
BT1	2.06	SC24-1	2.05	1614	2.05	LH39	2.05
E200	2.05	SC11-1	2.05	yan103	2.05	78551S	2.05
5023	2.04	liao2202	2.04	84-126-15-1	2.04	PHG86	2.04
BCC03	2.04	PHW03	2.03	jingnuo2	2.03	F939	2.03
PHW17	2.02	SG17	2.01	PHW79	2	Lx9801	2
DF20	1.99	fanrong2	1.99	9702	1.99	GY3	1.99
M3	1.97	ji53	1.97	LH143	1.97	cheng698-3	1.97
chengzi108	1.96	9706	1.96	xingK36	1.95	wu109	1.95
M3736	1.94	B76	1.93	DM07	1.93	D23	1.93
KP3130	1.93	785	1.93	XOP2	1.92	ji432	1.92
1538	1.92	zong548-1521	1.92	mu6	1.92	MP	1.92
suwan1611	1.91	PHM10	1.9	MBUB	1.9	3335	1.9
Q1261	1.89	LH82	1.89	TIL12	1.89	68202	1.88
802	1.88	yan812	1.88	chang69	1.87	OQ603	1.87
20564	1.87	PHP85	1.87	CML58	1.86	hai1134	1.86
cheng435	1.86	W23	1.86	MBPM	1.85	RS710	1.84
DH65232	1.84	Seagu1117	1.84	8982	1.84	LH60	1.84
91huang10	1.83	DF27	1.83	C13	1.82	PHW20	1.81
W9706	1.81	SC14	1.81	MBST	1.8	e220	1.79
shuangtaiwu	1.78	ji870	1.78	D883	1.77	longkang1	1.77
DH40	1.77	wenhuang31413	1.76	Q381	1.76	R25	1.76
changD	1.74	BC4B	1.74	Timpunia-1	1.73	y9961	1.73
yue39-4	1.73	Max	1.72	C166	1.72	Yd6	1.71
H66-6	1.71	bai197	1.71	shen5005	1.7	jiao05	1.7
ji833	1.69	XF197	1.69	D856	1.69	tangsipingtou	1.69
LH150	1.68	S001	1.68	PHP76	1.68	411	1.68
ning24	1.68	794	1.68	L061F	1.67	dan3115	1.67
C8605-2	1.67	W117	1.67	78371A	1.66	H114	1.66
LH127	1.66	PHN37	1.66	yun147	1.66	D20	1.65
PHW51	1.65	ZPON7	1.65	zhonghuang64	1.64	K514	1.64
Cwu215B	1.64	XZ19	1.64	zhong128	1.64	huangyesi	1.63
PHK35	1.63	xing230	1.63	Sg1533	1.61	832	1.6
98F1	1.6	wu312	1.59	32	1.59	huobai	1.58
757	1.58	PN2	1.58	chang7daxian1	1.58	zhongyin10	1.58
DF24	1.58	D801	1.58	DH138	1.57	S37	1.56
96201	1.56	huangchanga	1.56	da255	1.56	chi545	1.56
huangchangb	1.56	liao4271	1.55	luyuan133	1.55	jiutai1	1.55
L069	1.55	fu8538	1.55	PHJ33	1.54	CM105	1.54
7146	1.53	ning37	1.53	W2H03	1.52	chang72	1.52
IA5125	1.52	yan38	1.52	SZ3	1.52	xuan6	1.52

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4936	1.52	R136	1.51	B102	1.51	707	1.51
L105	1.51	tie7922	1.5	7884	1.5	ZEAxppRADDIAAPEI-7	1.5
PHN66	1.5	1121	1.5	wu126	1.5	W499	1.5
444	1.49	huotanghuang	1.49	LH191	1.49	gan41	1.49
wu125	1.48	ji434	1.48	R017	1.48	ye488	1.48
PHV78	1.47	PHJ31	1.47	xun92-6	1.47	chang3	1.47
LH57	1.46	H99	1.46	LH205	1.46	Mo24W	1.45
dan598	1.45	nan21-3	1.45	6	1.45	Aug-64	1.45
9710	1.45	18	1.44	72-125	1.44	PHR58	1.43
LH59	1.43	953	1.43	ji853	1.43	928	1.42
WN11H	1.42	LH38	1.42	LIBC4	1.42	duzi	1.41
P25	1.41	ji444	1.41	LH194	1.41	yu374	1.4
W238	1.4	207	1.39	liao7794	1.39	daqing133	1.39
11430	1.39	433-7	1.39	29MIBZ2	1.38	shangyin110-1	1.37
DH65232-DH9	1.36	200B	1.36	ZEAxppRBHDIAAPEI-5	1.36	dai6	1.36
MBSJ	1.36	9058	1.36	yi12	1.36	6103	1.36
ziyu3	1.35	7026B	1.35	qi318	1.35	SX-4-21	1.35
99122	1.35	PHG72	1.35	hu803	1.34	yuanwu05	1.34
PHT77	1.33	ML606	1.33	3H2	1.32	IBC2	1.32
Lo1067	1.31	R08	1.31	qi35	1.31	PHP02	1.3
jiu22	1.3	ning55	1.3	L05-6	1.3	M101	1.3
92huang40	1.3	zheng22	1.29	LH54	1.29	NC268	1.29
PHW43	1.29	hai014	1.28	zun90110	1.28	PHG50	1.28
PHW52	1.27	chihuang14	1.27	huo17	1.27	d140	1.27
1610	1.26	LH1	1.25	P167	1.25	han21	1.24
D33A	1.23	ye515	1.23	bao3040	1.23	R150	1.23
LH149	1.23	chaoxianbai	1.23	W64a	1.23	DH149	1.23
1313	1.23	PHM81	1.23	Va35	1.22	ziduosui	1.22
L061M	1.22	PHN47	1.21	TIL02	1.21	guangyou5	1.21
S22	1.21	PHN11	1.21	L127	1.2	LH52	1.2
D88	1.2	83IBI3	1.18	PHG35	1.18	PHG71	1.18
PHT10	1.18	E8501	1.18	qi319	1.18	R98	1.17
zaC546	1.17	M14	1.17	jian1495a	1.17	ben7884	1.17
ji4112	1.16	926	1.16	R1656	1.16	LH61	1.16
W153R	1.16	equn4	1.16	7327	1.16	W967	1.16
SS99	1.15	4003	1.15	NN14B	1.14	NQ508	1.14
W182bn	1.14	HD568	1.14	ZEAxppRAFDIAAPEI-1	1.14	PN2-8	1.14
E601	1.13	W222	1.13	4N506	1.13	S7913	1.13
fu8527	1.13	yi49	1.12	W172	1.12	ZEAxppRDCDIAAPEI-2	1.12
ND246	1.11	PHK24	1.1	xun92-8	1.1	PHT55	1.1
zhonger02	1.09	SW153	1.09	tian77	1.09	SC9	1.08

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

78010	1.08	PHG83	1.08	B37	1.08	ye5237	1.07
196	1.07	82huangzao4	1.07	Beck	1.07	P007	1.07
D869	1.07	CT109	1.07	fu8529	1.06	LH192	1.06
hai9-21	1.06	LH51	1.06	yu82	1.06	PHR32	1.05
PHJ75	1.05	795	1.05	B394	1.05	ning45	1.04
liao2204	1.04	787	1.04	N192	1.04	K14	1.04
2005-4	1.04	Maxa	1.03	8701	1.03	FAP1360A	1.03
L-1	1.03	R31	1.03	6M502	1.03	5032	1.02
fu8521	1.02	PHN29	1.02	PHK93	1.02	jinsui54	1.01
han49	1.01	zhong451	1.01	TIL04-TIP285	1.01	shan89	1.01
FC-13	1	W8304	1	W8555	1	Lo	1
77	0.99	PHBA6	0.99	XF77	0.99	su75	0.99
chen322	0.99	LH162	0.98	A554	0.98	468-3	0.98
7903E	0.97	LH193	0.97	S8324	0.97	SC30-1	0.97
75-364	0.97	CR1HT	0.97	4676A	0.96	Hda-5	0.96
H105W	0.96	4722	0.96	D892	0.96	S311	0.96
zhangxi28	0.96	BM	0.96	1101	0.96	2FACC	0.95
shen142	0.95	PB80	0.95	he344	0.95	M22	0.95
P136	0.95	F7584	0.94	PHR30	0.94	2369	0.94
MDF-13D	0.93	jiao51	0.93	PHV53	0.93	oh07B	0.93
140	0.93	LH119	0.93	PHPR5	0.93	S53	0.92
yuanwu02	0.92	79131	0.92	D886	0.92	PHR47	0.92
PHG39	0.92	3IBZ2	0.92	P39-chin	0.92	M9	0.92
9101-7	0.91	75-24	0.91	PHR55	0.91	441950	0.91
fu8701	0.91	CN165	0.91	IB014	0.91	7236	0.9
ji63	0.9	Lp215D	0.9	1127	0.9	jian1495b	0.89
P138	0.89	PHJ40	0.89	78004	0.89	PHT22	0.89
L139	0.89	WIL903	0.89	IB02	0.89	A679	0.88
zi330	0.88	J8606	0.88	E588	0.88	CR14	0.87
NS501	0.87	dong327	0.87	ye8001	0.87	mu4	0.87
shen977	0.87	LH295	0.86	680	0.86	LH220Ht	0.86
NS701	0.86	W968	0.86	HB8229	0.86	X314	0.86
LD61	0.86	779	0.86	MM402A	0.85	H8431	0.85
M1016	0.85	SC-9	0.85	87916W	0.84	yue89A12-1	0.84
LH190	0.84	PHV37	0.84	W969	0.84	K10	0.84
lv45	0.84	siyi	0.84	PHK42	0.83	3189	0.83
PHG84	0.83	yi36	0.83	LH160	0.83	Va26	0.83
LH65	0.82	D619	0.82	S8326	0.82	zhongxi042	0.82
PHK05	0.82	18-599	0.82	lian87	0.82	LH196	0.81
PHG47	0.81	B8	0.81	PH207	0.81	zhengbai11	0.8
Y223	0.8	91huang15	0.8	Pa91	0.8	PHT69	0.8

bioRxiv preprint doi: https://doi.org/10.1101/026963; this version posted September 16, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

baihe43	0.8	W966	0.8	yu537A	0.8	MBNA	0.8
3514	0.8	PHP60	0.79	PHM57	0.79	WIL901	0.78
OH7	0.78	2MA22	0.78	D978	0.78	LH93	0.78
LH145	0.77	D881	0.76	chihuang32	0.76	PHM44	0.76
697	0.75	ZS01250	0.75	IBB14	0.75	VG85-5	0.75
H84	0.75	5707	0.75	fengke1	0.74	PHNV9	0.73
D1051	0.73	MO113	0.73	619	0.73	chong72	0.72
zhen367	0.72	B47	0.72	ji533	0.72	XF117	0.72
B4	0.72	PHH93	0.71	A619	0.71	CN104	0.71
807D	0.71	L135	0.71	fu96	0.7	LH123HT	0.7
790	0.69	Z31B	0.69	LM-2	0.69	811A	0.69
WIL500	0.69	806A	0.68	FBHJ	0.68	De811	0.68
xun971	0.68	E28	0.68	PHK76	0.68	LH85	0.68
792	0.68	BS110	0.68	yi67	0.67	F42	0.67
PHWG5	0.67	PHJ70	0.67	75-14gao	0.67	DF32	0.67
jiao3	0.67	Co109	0.67	PHR36	0.66	PHR25	0.66
D854	0.65	B100	0.65	LP1	0.65	B09	0.64
cai11-8	0.64	PHT60	0.64	DJ7	0.63	zhangjin6	0.63
hua160	0.63	qi205	0.63	shuangM9B-1	0.62	daMO	0.62
20762	0.62	B7	0.61	H2	0.61	PHG24	0.61
FAPW	0.6	baiU8112	0.6	LH215	0.6	B68	0.59
E600	0.59	niu2-1	0.59	lan766-4-2	0.59	Lo1125	0.58
B14	0.58	guan17-1	0.58	N68a	0.57	zheng653	0.57
HBAI	0.57	xinlun5-9	0.57	LH202	0.55	20837	0.55
78002A	0.54	song1145	0.53	740	0.53	wei3322	0.53
ye832	0.5	XF27	0.5	liao5110	0.5	743	0.46
luyuan92	0.25	LH146HT	0.23	PHR62	0.11	qi410	0.06