

# RNA-seq based analysis of population structure within the maize inbred B73

Zhikai Liang<sup>1</sup> and James C. Schnable<sup>2\*</sup>

\*Correspondence:  
schnable@unl.edu

<sup>2</sup>Center for Plant Science  
Innovation & Department of  
Agronomy and Horticulture,  
University of Nebraska-Lincoln,  
Lincoln, NE, USA, Lincoln, NE,  
USA  
Full list of author information is  
available at the end of the article

## Abstract

B73 is a variety of maize (*Zea mays* ssp. *mays*) widely used in genetic, genomic, and phenotypic research around the world. B73 was also served as the reference genotype for the original maize genome sequencing project. The advent of large-scale RNA-sequencing as a method of measuring gene expression presents a unique opportunity to assess the level of relatedness among individuals identified as variety B73. The level of haplotype conservation and divergence across the genome were assessed using 27 RNA-seq data sets from 20 independent research groups in three countries. Several clearly distinct clades were identified among putatively B73 samples. A number of these blocks were defined by the presence of clearly defined genomic blocks containing a haplotype which did not match the published B73 reference genome. In a number of cases the relationship among B73 samples generated by different research groups recapitulated mentor/mentee relationships within the maize genetics community. A number of regions with distinct, dissimilar, haplotypes were identified in our study. However, when considering the age of the B73 accession – greater than 40 years – and the challenges of maintaining isogenic lines of a naturally outcrossing species, a strikingly high overall level of conservation was exhibited among B73 samples from around the globe.

**Keywords:** Maize, RNA-seq, Reference Genotypes

## Background

A great deal of biological research depends on reference genotypes that allow researchers around the world to work with material that is genetically identical or nearly identical. For many decades, assessing whether two samples labeled as coming from genetically identical sources truly were identical was a costly, time consuming, and often inconclusive process. Recent advances in genotyping and sequencing technology have revealed a number of cases where sample names and sequence information significantly different stories. One study of human cell cultures found that 18% of cell lines were either contaminated or something entirely different from what they were labeled as [1] with the widely used HeLa cell line being one of the most frequent offenders [2]. Among plants, a recent resequencing study of *Arabidopsis* demonstrated that a line believed to carry a mutation for the *ABP1* gene in an otherwise Col-0 background actually contained a wide range of other nonsense and missense mutations as well as a large region on chromosome 3 which came from a different *Arabidopsis* accession [3]. In soybean (*Glycine max*), segregating variation was observed among various inbred sources of the line Williams82 which was used in the construction of the soybean reference genome [4].

Here we set out to quantify how severely these issues of divergence among samples labeled as belonging to the same genetic background impact maize (*Zea mays*), the preeminent model for plant genetics over the past 100 years. Unlike soybean and *Arabidopsis*, maize is a naturally outcrossing species, so reference genotypes must be maintained by controlled self-pollination in each generation. This study focuses

specifically on the maize reference genotype B73, which was developed in Iowa and first registered in 1972 [5], widely used in commercial hybrid seed production across the United States for much of the 1970s and 1980s [6] and is represented in the parentage of many elite lines even today [7]. B73 has also been widely used by plant biologists conducting basic genetic research in maize, and was employed in the sequencing and assembly of the first maize reference genome [8].

## Methods

### Data sources

A search of NCBI's sequence read archive identified 25 Illumina RNA-seq data sets deposited by 19 independent research group in three countries (Table 1). Two additional RNA-seq data sets were constructed from B73 seed requested from Iowa State and the USDA's Germplasm Resources Information Network (Control 1 and Control 2 respectively). For these two samples RNA was extracted from 12-day old B73 seedlings grown at the University of Nebraska-Lincoln (Table 1). In four cases where the total amount of data per run was limited (USA 6, USA 8, USA 9 and USA 17), data from multiple sequencing runs labeled as coming from the same sample were grouped together for analysis. In one case, SRR514100, the total quantity of data was excessive, so only 1/10th of the total data set was employed.

**Table 1 B73 RNA-seq data sets sources**

Sample Name	Run Accession	Library Layout (bp)	Institute
Control 1	-	Paired (101)	University of Nebraska - Lincoln
Control 2	-	Single (51)	University of Nebraska - Lincoln
USA 1[9]	SRR651051	Paired (51)	University of Minnesota
USA 2[10]	SRR1819621	Paired (52)	University of Minnesota
USA 3[11]	SRR404150	Single (76)	University of Wisconsin - Madison
USA 4[12]	SRR514100	Paired (151)	University of Wisconsin - Madison
USA 5[13]	SRR940300	Single (101)	University of Wisconsin - Madison
USA 6[14]	<i>SRR395191, SRR395192</i> <i>SRR395194, SRR395208</i>	Single (40)	Iowa State University
USA 7	SRR445245	Paired (102)	Iowa State University
USA 8[15]	<i>SRR039505, SRR039506</i>	Single (35)	Danold Danforth Center
USA 9[16]	<i>SRR755252, SRR762349</i> <i>SRR762350, SRR762351</i> <i>SRR764626, SRR764627</i>	Single (35)	Danold Danforth Center
USA 10[17]	SRR1656746	Single (101)	University of Nebraska - Lincoln
USA 11[18]	SRR1567899	Paired (50)	Iowa State University
USA 12*[19]	SRR504480	Single (100)	University of California - Berkeley
USA 13[20]	SRR1587038	Single (101)	University of Wisconsin - Madison
USA 14[21]	SRR1231518	Single (100)	Cornell University
USA 15[22]	SRR1272115	Paired (50)	DuPont Pioneer
USA 16[23]	SRR640263	Single (35)	Yale University
USA 17[24]	<i>SRR520998, SRR520999</i>	Paired (51)	Cold Spring Harbor Laboratory
USA 18[25]	SRR536834	Single (76)	Virginia Tech
USA 19[26]	SRR999052	Paired (50)	Cold Spring Harbor Laboratory
USA 20[27]	SRR248565	Paired (81)	Stanford University
CHN 1[28]	SRR491307	Paired (76)	China Agricultural University
CHN 2[29]	SRR1522119	Paired (102)	China Agricultural University
CHN 3[30]	SRR910231	Paired (91)	China Academy of Agricultural Sciences
DEU 1[31]	SRR924107	Single (96)	MPIPZ
DEU 2[32]	SRR1030995	Single (85)	University of Bonn

\*:USA 12 harbors an long introgression on chromosome 2.

### Alignment and initial SNP calling

Low quality sequences were removed using Trimmomatic-0.33 with settings LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, MINLEN:36 [33]. Trimmed reads were aligned to the repeat masked version of the maize reference genome (version B73 RefGen v3) [8] downloaded from Ensemble ([ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea\\_mays/dna/](ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/zea_mays/dna/)) using GSNAP in version 2014-12-29 (with parameters -N 1,-n 2,-Q) [34]. Output files were converted from SAM to BAM format, sorted, and indexed using SAMtools [35]. SNPs were called in parallel along ten chromosomes of the maize version 3 using SAMtools mpileup (-I -F 0.01) and bcftools call (-mv -Vindels -Ob).

### SNP list generation

The view function of Bcftools was combined with in-house Python scripts to extract the content of bcf files and classify SNPs based on the number of reference and non-reference alleles on every screened SNP locus. In detail, if the total number of reads covering a particular SNP in a particular sample was below 5, then the site was treated as missing data. When 99% reads on the locus of a sample were from the non-reference allele the sample was coded as homozygous non-reference allele. The same criteria were used for calling a site as homozygous reference allele. When the reads containing reference and non-reference alleles totaled more than 90% of all reads and each allele was represented by more than 20% of aligned reads the site was coded as heterozygous. If two or more alleles were present at >1% of aligned reads but the above criteria were not satisfied, the site was also coded as missing data. To reduce the prevalence of false SNPs resulting from the alignment of reads from multiple paralogous loci to a single position in the reference genome, sites which were scored as heterozygous in more than 20% of all genotyped individuals were discarded. In total, 13,360 SNPs were used in downstream analysis. For each of these SNPs, the impact of the SNP on gene function was estimated using SnpEff v4.1 and SnpEff databases (*AGPv3.26*) [36].

### Population structure analysis

The distribution of the three possible genotypes (homozygous reference allele, homozygous non-reference allele and heterozygous allele) over each of the ten chromosomes of maize was visualized using matplotlib. PhyML 3.0 [37] was used to construct a phylogenetic tree with 100 bootstrap replicates, and 13,360 SNPs in total of 27 data sets.

### Expression bias test

Individual FPKM (Frequency per kilobase of exon per million reads) value for each gene in each data set was calculated using Cufflinks v2.2.1 [38]. Expression values were averaged across all China and USA South samples (excluded USA 12 sample that contained a unique introgressed region) separately. Only genes with average FPKM values  $\geq 10$  in both groups were retained for testing expression bias. The remaining genes were sorted into two groups: genes located in the 7 chromosome intervals where USA South and China showed different haplotypes and genes outside these intervals.

### Origins of haplotype blocks

The origins of haplotype blocks observed in some B73 accessions but not in the published reference genome were investigated using data from diverse maize lines in the HapMap2 project [39]. In order to make comparisons to these data, alignments and SNP calling were performed a second time as above using B73 RefGen v2. All of samples in China or USA North clade were combined to generate a consensus sets of SNP calls with reduced missing data. In examining region c2r2, sample USA 12 was used individually in addition to the combined China and USA North sequences (Additional file 1: Figure S1). In the analysis of region c5r2 (Additional file 1: Figure S1), USA 10, USA 14 and USA 15 were combined to generate a consensus set of SNP calls for the UC-Berkeley clade. The resulting SNP sets were employed for phylogenetic analysis as described above, with the alteration that the an approximate likelihood ratio test (aLRT) method with SH-like was employed. The resulting trees were visualized using FigTree v1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>).

## Results

### Relationship among accessions labeled as B73

After alignment, SNP calling, and filtering (see Methods), a total of 13,360 high confidence segregating SNPs were identified among the 27 RNA-seq samples labeled as B73 employed in this study, substantially lower than the ~64,000 high quality SNPs identified by RNA-seq in a population segregating for a single non-B73 haplotype [40]. Phylogenetic analysis identified three distinct clades of samples separated by long branches with 100% bootstrap support (Figure 1). One clade consisted entirely of Chinese samples, one clade of samples from US research groups from Minnesota and Wisconsin, and the final clade encompassed the majority samples from US research groups as well as all German samples and the published reference genome for B73. We designated these clades "China", "USA North", and "USA South" respectively. Notably, the USA North clade is paraphyletic with respect to the China clade, suggesting B73 samples in China are likely derived from this group while both German samples are clearly part of the USA South Clade.

The USA South clade was somewhat arbitrarily divided into three subclades with at least 60% bootstrap support, as well as a number of singleton lineages (USA 1, USA 13, USA 19). Two of these clades contained control samples generated for this study, one from B73 seed requested through the USDA Germplasm Resource Network, and one from B73 seed requested from Iowa State. The subclade containing the known USDA B73 sample also contained the B73 reference genome sequence, consistent with the reported seed source for the B73 used in the construction of the reference genome. The final subclade did not contain any control samples. However, it was notable that four of the six samples placed in this clade originated in research groups whose PIs had conducted either PhD or Postdoctoral training with Michael Freeling at UC-Berkeley, and none of the samples outside of this clade originated in research groups linked to UC-Berkeley. Based on these, we designated the final USA South subclade "UC-Berkeley". This accessions has also been described as "Freeling B73" [41].

### Genomic distribution of within-B73 polymorphisms

The polymorphic SNPs identified in this study could originate from one of several sources including de novo mutations or the introgression of non-B73 haplotypes in one or more lineages. SNPs originating from de novo mutations would be expected to show a distribution approximating that of gene density across the maize chromosomes. SNPs resulting from introgression of other haplotypes into B73 should be tightly clustered.

When the positions of the SNPs identified in this study were plotted it became clear that 55.3% SNPs fall within a small number of dense genomic blocks on chromosomes 2, 3, 4, 5, and 6 (Figure 2). The distribution of non-reference-genome-like haplotype blocks is consistent with the clade relationships identified above. The USA North clade can be defined by a large block of SNPs on chromosome 2, and smaller blocks on chromosomes 2, 3, and 5, all of which are shared with the China B73 clade. In addition to the blocks shared with the USA North B73 clade, samples from the China B73 clade all share a number of additional non-reference-genome-like blocks on chromosomes 2, 4, and 6. There are no non-reference-genome-like blocks shared by all members of the USA South clade, however a single non-reference-genome-like block on chromosome 5 is shared by the UC-Berkeley subclade of USA South. This block appears to share one breakpoint but not both with a block present in the USA North and China samples. The large block non-reference-genome-like block like SNPs observed only on chromosome 2 on USA 12 can likely be explained by the unique origin of this sample from wild type siblings of knotted1 mutants backcrossed into B73 [42]. The remaining USA South samples, including the USDA GRIN, Iowa State, and German samples do not contain any obvious SNP blocks.

### Functional impact of within-B73 polymorphism

Because the data used here came entirely from RNA-seq studies, our ability to detect SNPs was limited to genes which were consistently expressed at high enough levels to provide coverage of target regions. A total of 25,644 genes were expressed at levels >10 FPKM when at least one of data sets analyzed in this study. Of these genes, 633 (2.5%) fell within regions with non-reference-genome-like SNP blocks in one or more B73 clades. Using SnpEff, we identified 10 cases where SNPs produced “high impact” change such as the gain or loss of a stop code or the alteration of a splice donor or splice acceptor site and 396 cases which produced missense mutations which altered protein sequence. Only three genes with reported mutant phenotypes (whp1, mop1, and gol1) were in these regions, which only constituted at 2.7% of 112 classical identified maize genes with reported mutant phenotypes [43]. However, it must be noted that this is likely an underestimate of the true number of changes, nonsense mediated decay may reduce or eliminate the expression of alleles of genes containing high impact SNPs, reducing the chances these SNPs will be detected from RNA-seq data.

### Impact of within-B73 polymorphism on estimated gene expression

The alignment rate for RNA-seq data from non-B73 genotypes to the B73 reference genome is approximately 13% lower than the alignment rate of RNA-seq data generated from B73 plants [44]. To test whether there is a bias towards lower estimated

expression levels from RNA-seq data for genes in non-reference-genome-like blocks, the expression of highly expressed genes (ie average expression  $\geq 10$  FPKM) was compared between samples in the USA South clade (excluding USA 12) and samples in the China clade. Genes within introgressed regions showed a 5.6% reduction on expression in China samples, relative to a control set of genes outside introgressed regions between B73 USA South and B73 China (see Methods). This reduction approximately half as large as would be predicted if the reduced alignment rate of data from non-B73 samples resulted solely from increased difficulty of aligning reads containing SNPs to the reference genome. Potentially, the other half of the reduced alignment rate for non-B73 samples is the result of the expression of lineage specific genes, as previously suggested [44].

#### Origins of polymorphic regions in B73 accessions

A total of 7 chromosome intervals (referred to here as c2r1, c2r2, c4r1, c5r1, c5r2, c6r1 and c6r2) containing non-reference genome haplotypes were identified in two or more samples (Table 2; Additional file 1: Figure S1). SNP calls were extracted from individual non-reference-genome-like blocks using the previous version of the maize reference genome (B73 RefGen v2) and compared to genotype calls generated from 103 diverse inbreds resequenced by the Maize HapMap2 project [39]. One example, c2r1 is shown in Figure 4. The non-reference genome haplotype present in this block for the Chinese samples clusters very closely with W22, an older inbred developed in Wisconsin which has also been widely used in the maize genetics research community. Analysis of the other six large haplotype blocks produced longer branch lengths relative to the accessions represented in the Maize HapMap2 dataset (Table 2). However, in each case the haplotypes generated from each clade containing a non-reference-genome-like block clustered together, confirming that these regions did not result from parallel introgressions covering the same regions of the genome. These was also true from c5r2 which was represented in both the USA North and China clades as well as the UC-Berkeley subclade (Additional file 2: Figure S2D). A constraining the c2r2 region to only cover that portion of the genome which contained a block of SNPs in the USA North clade, the China clade and sample USA 12 revealed that USA North and China clustered together while USA 12 was placed at a different location on the tree (Additional file 2: Figure S2A).

## Discussion

The maize community has long speculated that significant differences exist among B73 from different sources. Recently that it has become feasible to quantify the specific differences among B73 accessions. Here we employed previously published RNA-seq data sets from a large number of independent research groups to assess the diversity among B73 accessions. No cases of samples which were labeled as originated from B73 but were clearly not B73 based on SNP data were identified in this study. Despite a 40+ generation reproductive history distributed across at least three continents, this analysis shows that 97.7% of the gene space of the maize genome is represented by a single consistent haplotype across all B73 accessions included in this study.

**Table 2 Relationship of Non-Reference-Genome Like SNP Blocks to Haplotypes Surveyed by HapMap2**

Genomic blocks	Chr	Start (kb)	Stop (kb)	Closest haplotypes	Branch length	Present in
c2r1	2	40000	44300	W22	0.00000029	China
c2r2	2	212450	224250	BKN027	0.38965362	China
				BKN027	0.38965366	USA North
				M162W	0.31485540	USA 12
c4r1	4	169650	191550	W22	0.62055176	China
c5r1	5	201200	203000	no single best match	-	China
				no single best match	-	USA North
c5r2	5	209732	211540	no single best match	-	China
				no single best match	-	USA North
				no single best match	-	USA South
c6r1	6	120	8800	CML511	0.59243725	China
c6r2	6	20900	24670	OH7B	0.08886910	China

The interspersed SNPs distributed over ten chromosomes of maize may result from de-novo mutations, segregation of heterozygous loci in the original B73 founder accession [4], or false positive SNP calling errors. However, the polymorphic differences identified among B73 samples in this study primarily fell into a small number of dense non-reference-genome-like blocks which would be consistent with introgression of non-B73 germplasm into a B73 background. It is important to note that the B73 reference genome was sequenced recently relative to the total age of the B73 accession. Therefore, it is not possible to infer whether a given non-reference-genome-like block originated from introgression into the line in which the non-reference-genome SNPs are observed or introgression into the B73 lineage which was ultimately employed in the creation of the B73 reference genome. However, in either case the relatively small size of these non-reference genome like blocks suggests multiple generations of backcrossing to the original B73 line, which would not be consistent with a model based on unrecognized pollen contamination.

Instead we propose a model based on the results from Sample USA 12. Sample 12 consists of homozygous wild-type plants selected from family segregating for the Knotted1 [19]. Therefore the block on chromosome 2 (~1% of the total maize genome) likely represents residual sequence from the knotted1 mutant donor parent line and is consistent with at least 5 generations of backcrossing (expected contribution of the donor parent = ~1.56%). Similar accidental fixations of unlinked regions may have occurred during the intentional introgression of other traits into a B73 background, such as disease resistance genes [45].

The monophyletic placement of Chinese B73 datasets suggests that the B73 seed available in China likely originated from a single transfer from the USA, apparently of seed belonging to the USA North clade and is an indicator of current tight controls on seed import/export which limit the ease with which seed change be exchanged between collaborators in China and the United States. Samples from Germany did not consistently form a monophyletic group. The concordance of academic lineages and genomic relationships in the UC Berkeley subclade acts as a remarkable positive control. More extensive sampling of B73 samples from many labs which employ this genotype in maize genetics research but have not, to date, published RNA-seq datasets may identify further B73 clades and subclades and additional cases where

specific genomic variations have dispersed across the country as graduate students and postdocs leave a given lab for faculty positions of their own.

## Conclusions

The existence of genomic variation among samples labeled as belonging to the same accession creates barriers to reproducibility, one of the core requirements of the scientific method. In this study no examples of sample mislabeling were identified. However, a number of non-reference-genome-like blocks were identified in B73 samples originated from some sources. These blocks were shown to contain missense and nonsense mutations and measurably lower estimated expression values for genes in these regions. The identification of the relationships among different variants of B73 and the genomic locations of non-reference-genome-like regions will allow these differences to be controlled for future studies. With the rapid rise of sequencing-based assays such as RNA-seq, the strategy employed here may be a good one to apply in any case where one or more reference genotypes are widely employed in research across institutions, countries, and continents.

**Figure 1 Phylogenetic tree of 27 data sets (A) Distance-scaled branch lengths; (B) Un-scaled tree.** Only bootstrap values greater than or equal to 60 are displayed.

**Figure 2 SNP distribution pattern along 6 chromosomes of 27 data sets.** Blue dots are non-reference like homozygous alleles and red dots are heterozygous alleles.

**Figure 3 Zoomed haplotype regions of c2r1, c2r2 and c5r2.** (A) Selected haplotype region c2r1 on Chromosome 2; (B) Selected haplotype region c2r2 on Chromosome 2; (C) Selected haplotype region c5r2 on Chromosome 5. Blue line is homozygous non-reference allele and red line is heterozygous allele. Regions within green bars are identified haplotype regions.

**Figure 4 Origin of haplotype region of c2r1.**

### Competing interests

The authors declare that they have no competing interests.

### Author's contributions

JCS and ZL designed the study, ZL collected the data, ZL performed the analysis, and JCS and ZL wrote the manuscript.

### Acknowledgements

This project is supported by start-up funds from University of Nebraska-Lincoln to JCS.

### Author details

<sup>1</sup>Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, USA, Lincoln, NE, USA.

<sup>2</sup>Center for Plant Science Innovation & Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, USA, Lincoln, NE, USA.

## References

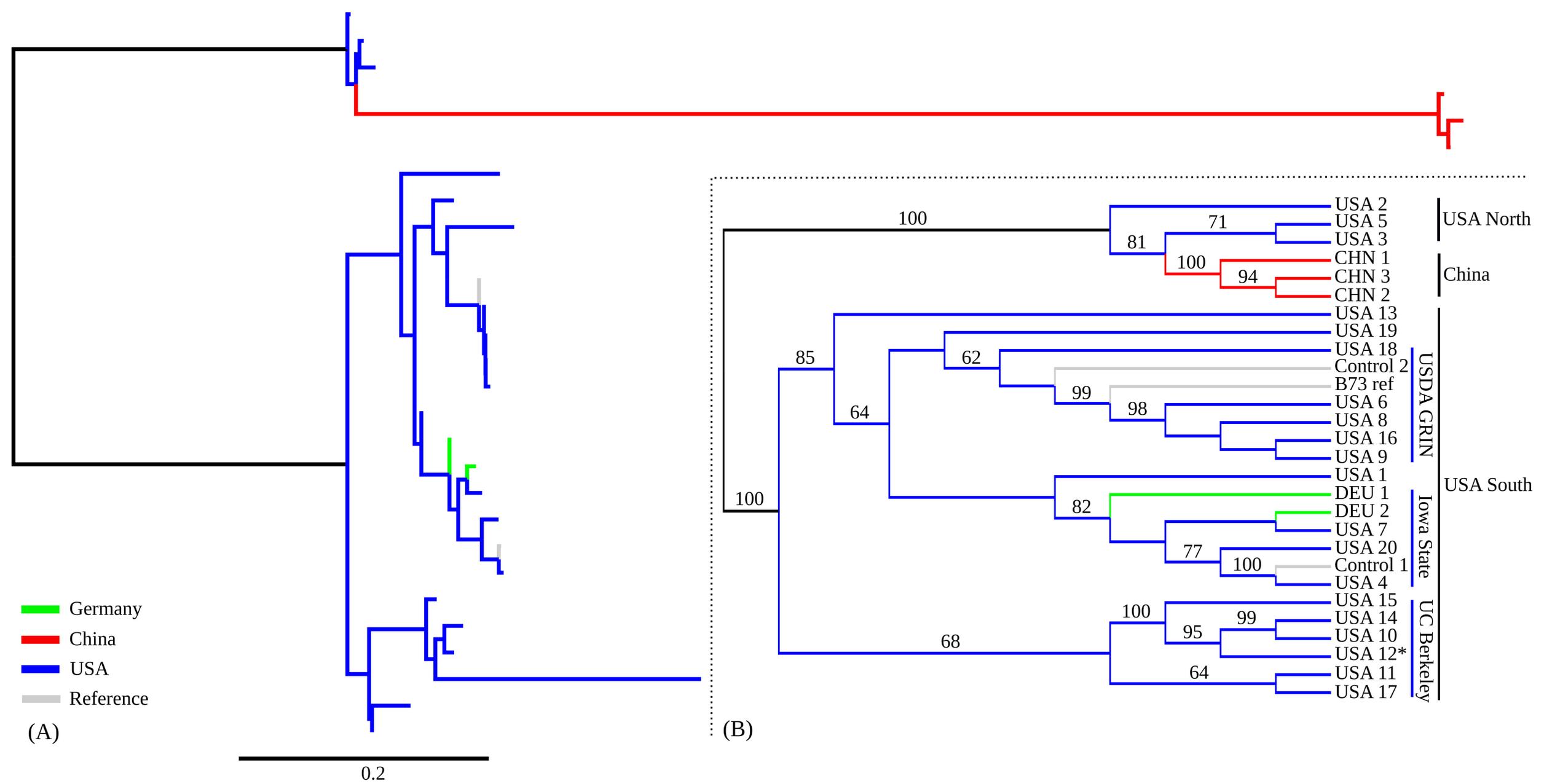
1. MacLeod, R.A., Dirks, W.G., Matsuo, Y., Kaufmann, M., Milch, H., Drexler, H.G.: Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *International Journal of Cancer* **83**(4), 555–563 (1999)
2. Lucey, B.P., Nelson-Rees, W.A., Hutchins, G.M.: Henrietta lacks, hela cells, and cell culture contamination. *Archives of pathology & laboratory medicine* **133**(9), 1463–1467 (2009)
3. Enders, T.A., Oh, S., Yang, Z., Montgomery, B.L., Strader, L.C.: Genome Sequencing of Arabidopsis abp1-5 Reveals Second-Site Mutations That May Affect Phenotypes. *The Plant Cell* **27**(7), 1820–1826 (2015)
4. Haun, W.J., Hyten, D.L., Xu, W.W., Gerhardt, D.J., Albert, T.J., Richmond, T., Jeddloh, J.A., Jia, G., Springer, N.M., Vance, C.P., et al.: The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant physiology* **155**(2), 645–655 (2011)
5. Russell, W.: Registration of B70 and B73 Parental Lines of Maize1 (Reg. Nos. PL16 and PL17). *Crop Science* **12**(5), 721–721 (1972)
6. Darrah, L., Zuber, M.: 1985 united states farm maize germplasm base and commercial breeding strategies. *Crop science* **26**(6), 1109–1113 (1986)
7. Mikel, M.A.: Genetic diversity and improvement of contemporary proprietary North American dent corn. *Crop science* **48**(5), 1686–1695 (2008)
8. Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al.: The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**(5956), 1112–1115 (2009)
9. Eichten, S.R., Briskine, R., Song, J., Li, Q., Swanson-Wagner, R., Hermanson, P.J., Waters, A.J., Starr, E., West, P.T., Tiffin, P., et al.: Epigenetic and genetic influences on dna methylation variation in maize populations. *The Plant Cell Online* **25**(8), 2783–2797 (2013)
10. Makarevitch, I., Waters, A.J., West, P.T., Stitzer, M., Hirsch, C.N., Ross-Ibarra, J., Springer, N.M.: Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS genetics* **11**(1), 1004915 (2015)
11. Sekhon, R.S., Briskine, R., Hirsch, C.N., Myers, C.L., Springer, N.M., Buell, C.R., de Leon, N., Kaeppeler, S.M.: Maize gene atlas developed by rna sequencing and comparative evaluation of transcriptomes based on rna sequencing and microarrays. *PLoS One* **8**(4), 61005 (2013)
12. Martin, J.A., Johnson, N.V., Gross, S.M., Schnable, J., Meng, X., Wang, M., Coleman-Derr, D., Lindquist, E., Wei, C.-L., Kaeppeler, S., et al.: A near complete snapshot of the zea mays seedling transcriptome revealed from ultra-deep sequencing. *Scientific reports* **4** (2014)
13. Stelpflug, S.C., Sekhon, R.S., Vaillancourt, B., Hirsch, C.N., Buell, C.R., de Leon, N., Kaeppeler, S.M.: An expanded maize gene expression atlas based on rna sequencing and its use to explore root development. *The Plant Genome* (2015)
14. Paschold, A., Jia, Y., Marcon, C., Lund, S., Larson, N.B., Yeh, C.-T., Ossowski, S., Lanz, C., Nettleton, D., Schnable, P.S., et al.: Complementation contributes to transcriptome complexity in maize (zea mays l.) hybrids relative to their inbred parents. *Genome research* **22**(12), 2445–2454 (2012)
15. Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S.L., Kebrom, T.H., Provart, N., Patel, R., Myers, C.R., et al.: The developmental dynamics of the maize leaf transcriptome. *Nature genetics* **42**(12), 1060–1067 (2010)
16. Wang, L., Czedik-Eysenberg, A., Mertz, R.A., Si, Y., Tohge, T., Nunes-Nesi, A., Arrivault, S., Dedow, L.K., Bryant, D.W., Zhou, W., et al.: Comparative analyses of c4 and c3 photosynthesis in developing leaves of maize and rice. *Nature biotechnology* (2014)
17. Campbell, M.T., Proctor, C.A., Dou, Y., Schmitz, A.J., Phansak, P., Kruger, G.R., Zhang, C., Walia, H.: Genetic and molecular characterization of submergence response identifies *sub1a* as a major submergence tolerance locus in maize. *PLoS one* **10**(3), 0120385 (2015)
18. Yi, G., Neelakandan, A.K., Gontarek, B.C., Vollbrecht, E., Becraft, P.W.: The naked endosperm genes encode duplicate indeterminate domain transcription factors required for maize endosperm cell patterning and differentiation. *Plant physiology* **167**(2), 443–456 (2015)
19. Bolduc, N., Yilmaz, A., Mejia-Guerra, M.K., Morohashi, K., O'Connor, D., Grotewold, E., Hake, S.: Unraveling the knotted1 regulatory network in maize meristems. *Genes & development* **26**(15), 1685–1690 (2012)
20. Lemmon, Z.H., Bukowski, R., Sun, Q., Doebley, J.F.: The role of cis regulatory evolution in maize domestication. *PLoS Genet* **10**(11), 1004745 (2014)
21. Frank, M.H., Edwards, M.B., Schultz, E.R., McKain, M.R., Fei, Z., Sørensen, I., Rose, J.K., Scanlon, M.J.: Dissecting the molecular signatures of apical cell-type shoot meristems from two ancient land plant lineages. *New Phytologist* (2015)
22. Thatcher, S.R., Zhou, W., Leonard, A., Wang, B.-B., Beatty, M., Zastrow-Hayes, G., Zhao, X., Baumgarten, A., Li, B.: Genome-wide analysis of alternative splicing in zea mays: landscape and genetic regulation. *The Plant Cell* **26**(9), 3472–3487 (2014)
23. He, G., Chen, B., Wang, X., Li, X., Li, J., He, H., Yang, M., Lu, L., Qi, Y., Wang, X., et al.: Conservation and divergence of transcriptomic and epigenomic variation in maize hybrids. *Genome Biol* **14**(6), 57 (2013)
24. Regulski, M., Lu, Z., Kendall, J., Donoghue, M.T., Reinders, J., Llaca, V., Deschamps, S., Smith, A., Levy, D., McCombie, W.R., et al.: The maize methylome influences mrna splice sites and reveals widespread paramutation-like switches guided by small rna. *Genome research* **23**(10), 1651–1662 (2013)
25. Kakumanu, A., Ambavaram, M.M., Klumas, C., Krishnan, A., Batlang, U., Myers, E., Grene, R., Pereira, A.: Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by rna-seq. *Plant physiology* **160**(2), 846–867 (2012)
26. Eveland, A.L., Goldshmidt, A., Pautler, M., Morohashi, K., Liseron-Monfils, C., Lewis, M.W., Kumari, S., Hiraga, S., Yang, F., Unger-Wallace, E., et al.: Regulatory modules controlling maize inflorescence architecture. *Genome research* **24**(3), 431–443 (2014)

27. Chettoor, A.M., Givan, S.A., Cole, R.A., Coker, C.T., Unger-Wallace, E., Vejlupekova, Z., Vollbrecht, E., Fowler, J.E., Evans, M.: Discovery of novel transcripts and gametophytic functions via rna-seq analysis of maize gametophytic transcriptomes. *Genome Biol* **15**(414), 10–1186 (2014)
28. Zhang, M., Xie, S., Dong, X., Zhao, X., Zeng, B., Chen, J., Li, H., Yang, W., Zhao, H., Wang, G., *et al.*: Genome-wide high resolution parental-specific dna and histone methylation maps uncover patterns of imprinting regulation in maize. *Genome research* **24**(1), 167–176 (2014)
29. Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A., Lai, J.: Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant physiology* **166**(1), 252–264 (2014)
30. Fu, J., Cheng, Y., Linghu, J., Yang, X., Kang, L., Zhang, Z., Zhang, J., He, C., Du, X., Peng, Z., *et al.*: Rna sequencing reveals the complex regulatory network in the maize kernel. *Nature communications* **4** (2013)
31. Urbany, C., Benke, A., Marsian, J., Huettel, B., Reinhardt, R., Stich, B.: Ups and downs of a transcriptional landscape shape iron deficiency associated chlorosis of the maize inbreds b73 and mo17. *BMC plant biology* **13**(1), 213 (2013)
32. Opitz, N., Paschold, A., Marcon, C., Malik, W.A., Lanz, C., Piepho, H.-P., Hochholdinger, F.: Transcriptomic complexity in young maize primary roots in response to low water potentials. *BMC genomics* **15**(1), 741 (2014)
33. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 170 (2014)
34. Wu, T.D., Nacu, S.: Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**(7), 873–881 (2010)
35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., *et al.*: The sequence alignment/map format and samtools. *Bioinformatics* **25**(16), 2078–2079 (2009)
36. Cingolani, P., Platts, A., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., Ruden, D.M.: A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of *drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**(2), 80–92 (2012)
37. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O.: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phylml 3.0. *Systematic biology* **59**(3), 307–321 (2010)
38. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L.: Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nature protocols* **7**(3), 562–578 (2012)
39. Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., *et al.*: Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics* **44**(7), 803–807 (2012)
40. Liu, S., Yeh, C.-T., Tang, H.M., Nettleton, D., Schnable, P.S.: Gene mapping via bulked segregant rna-seq (bsr-seq). *PLoS One* **7**(5), 36406 (2012)
41. Johnston, R., Wang, M., Sun, Q., Sylvester, A.W., Hake, S., Scanlon, M.J.: Transcriptomic analyses indicate that maize ligule development recapitulates gene expression patterns that occur during lateral organ initiation. *The Plant Cell* **26**(12), 4718–4732 (2014)
42. Vollbrecht, E., Reiser, L., Hake, S.: Shoot meristem size is dependent on inbred background and presence of the maize homeobox gene, *knotted1*. *Development* **127**(14), 3161–3172 (2000)
43. Schnable, J.C., Freeling, M.: Genes identified by visible mutant phenotypes show increased bias toward one of two subgenomes of maize. *PLoS one* **6**(3), 17855 (2011)
44. Hirsch, C.N., Foerster, J.M., Johnson, J.M., Sekhon, R.S., Muttoni, G., Vaillancourt, B., Peñagaricano, F., Lindquist, E., Pedraza, M.A., Barry, K., *et al.*: Insights into the maize pan-genome and pan-transcriptome. *The Plant Cell Online* **26**(1), 121–135 (2014)
45. Lipps, P., Pratt, R., Hakiza, J.: Interaction of ht and partial resistance to *exserohilum turcicum* in maize. *Plant disease* **81**(3), 277–282 (1997)

#### Additional Files

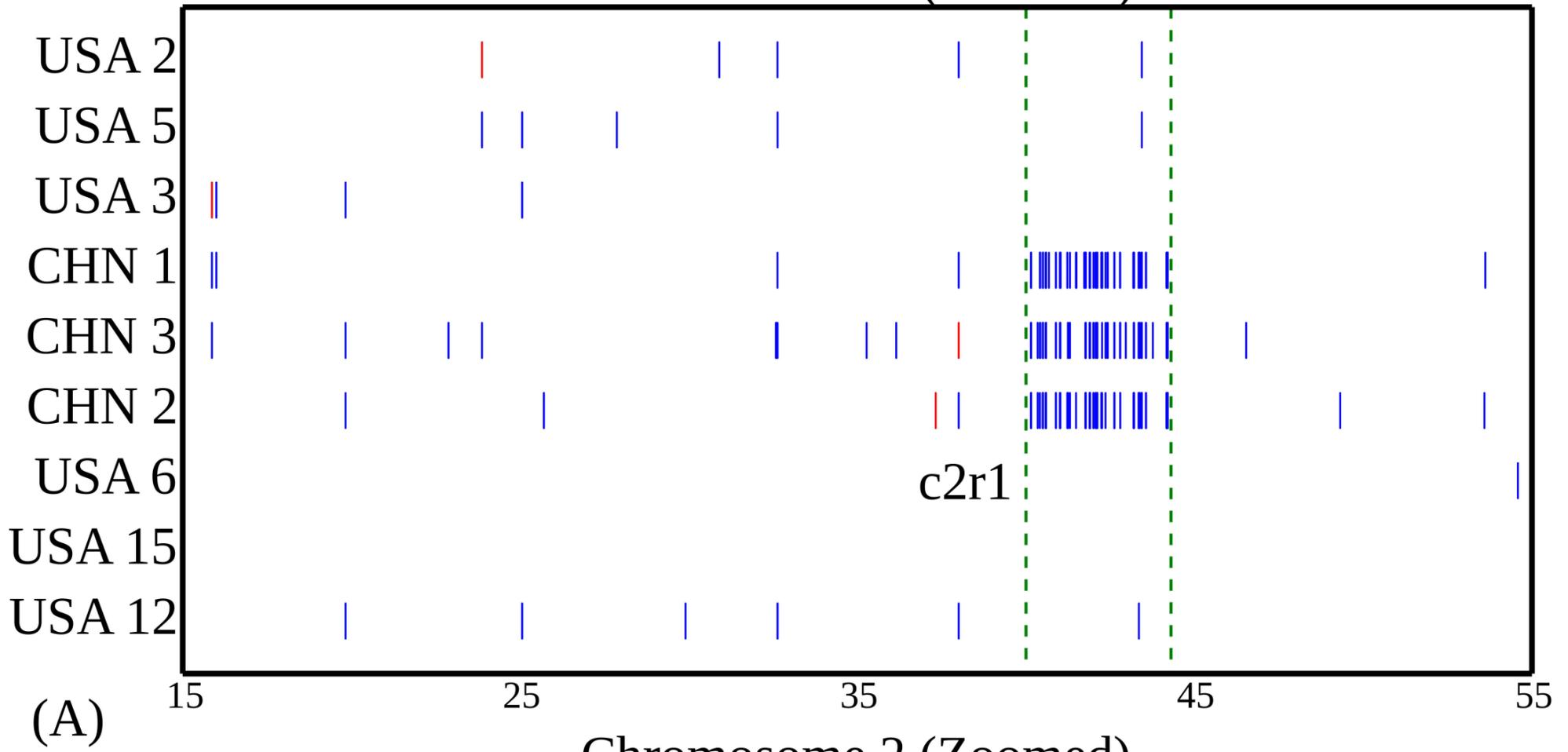
Additional file 1: Figure S1. Identified 7 haplotype blocks and SNP distribution pattern along 10 chromosomes of 27 data sets.

Additional file 2: Figure S2. Phylogenetic tree of other haplotype regions and corresponding lines in HapmapV2.(Additional file 2:Figure S2A, Origin of c2r2; Additional file 2:Figure S2B, Origin of c4r1; Additional file 2:Figure S2C, Origin of c5r1; Additional file 2:Figure S2D, Origin of c5r2; Additional file 2:Figure S2E, Origin of c6r1; Additional file 2:Figure S2F, Origin of c6r2.)

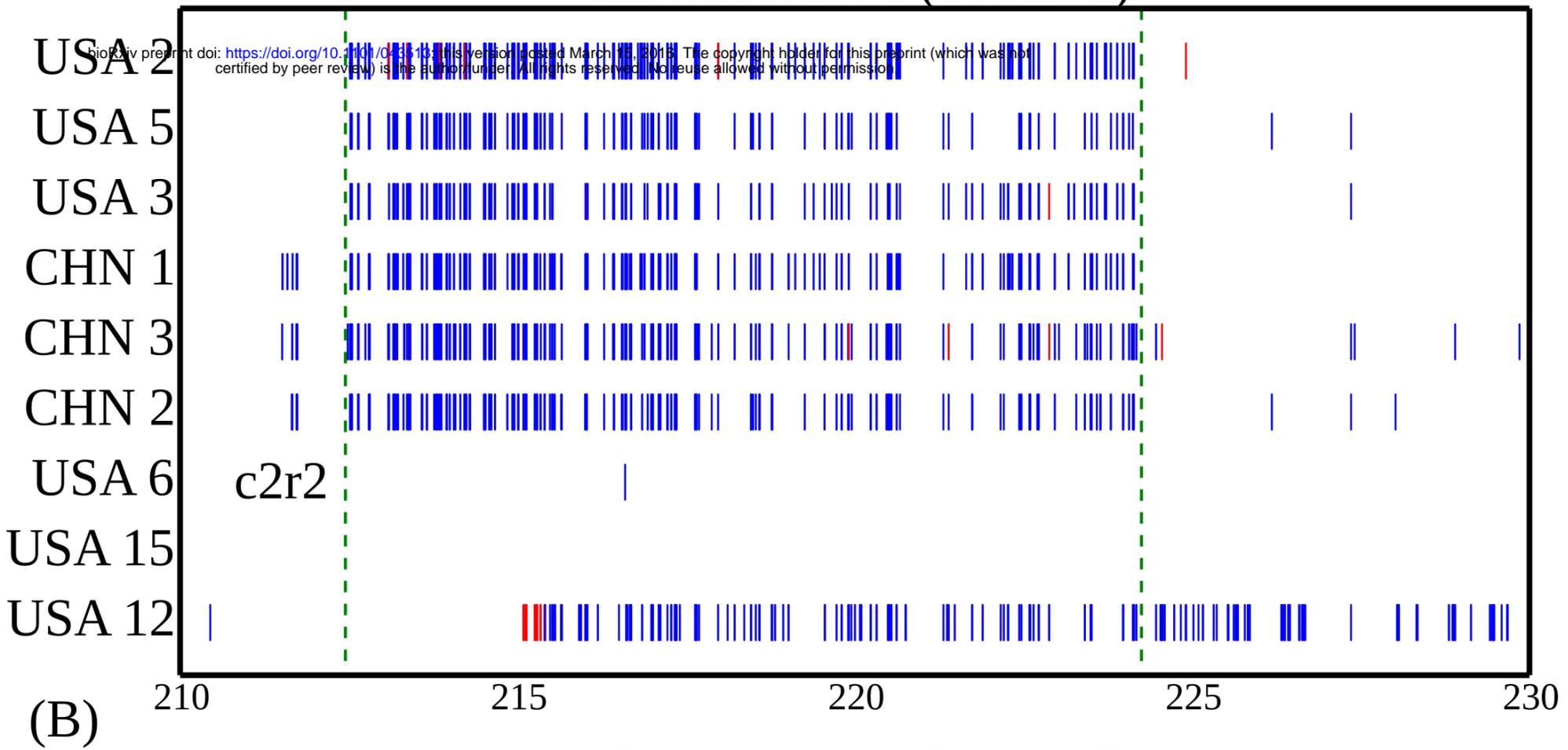




### Chromosome 2 (Zoomed)



### Chromosome 2 (Zoomed)



### Chromosome 5 (Zoomed)

