University of Crete
ICS-FORTH

Master Thesis

Evolutionary forces in the genomic neighborhoods of polymorphic transposable elements in plant populations

Author:
Joanna Garefalaki

Committee:
Pavlos Pavlidis
Kiriakos Kotzabasis
Panagiotis Sarris

A thesis submitted in fulfillment of the requirements for the degree of MSc in Molecular Biology-Plant Biotechnology

in the

Biology Department, University of Crete, Greece

September 2019

“*The history of the earth is recorded in the layers of its crust; the history of all organisms is inscribed in the chromosomes.* ”

H. Kihara

University of Crete

Department of Biology

ICS-FORTH

MSc in Molecular Biology-Plant Biotechnology

Evolutionary forces in the genomic neighborhoods of polymorphic transposable elements in plant populations

by Joanna Garefalaki

Abstract

Transposable Elements (TEs) have been shown to evolve under the effects of either positive and/or negative selection. Some of them is believed to be beneficial to organisms and they can even be domesticated by their host genome, contributing to genomic diversity and adaptation of natural populations or crops. The polymorphic landscape of TE at the population level of many species as well as their adaptive capacity is still unknown. Our goal is first, to characterize the evolutionary forces in the genomic neighborhoods of polymorphic TEs in natural populations of economically important plants. Second, to detect and characterize Transposable Element Insertion Polymorphisms (TIPs) in a large gene pool, and third, to gain access to genetic diversity at species level. Here, we traced TIPs from the previously published 3000 rice genome project database, using Mobile Element Locator Tool (MELT), a software developed specifically for the detection of (TIPs) from large datasets. Using SweeD, a maximum-likelihood approach to infer selective sweeps and localize recent and strong positive selection, we detected the action of recent and strong Darwinian selection in TE polymorphic genomic regions following the insertion of the TE or its absence.

Acknowledgements

I started this project on my first day of second year of grad school, back in October of 2018. I have learned so much along the way (including learning how to code from scratch) and it has been amazing experience. I take this opportunity to express my gratitude and thank my Supervisor, Professor, Mentor and Friend Dr. Pavlos Pavlidis for accepting me and giving me the opportunity to be a part of his great team. Without his patience and support I would never be able to accomplish this MSc Thesis. I would also like to thank the other two members of my advisory committee: Dr. Kiriakos Kotzabasis and Dr. Panagiotis Sarris. They where both a great teachers and mentors to me. I feel really grateful for having such amazing laboratory members and friends. Special thanks to Maria Malliarou for being supportive to me on my first steps learning Python and of course Antonis Kioukis for teaching me Bioinformatics from day one until today. I would also like to thank Ioannis Koutsoukos for his advice and help during the development of this project. Nothing would be accomplished without my friends Artemis, Nefeli, Agapi, Christos, Olympia, Thimios and Alexis. I would like to thank them for being here for me all the time and supported me by any manner of means. I am also grateful to Alexandros Marantos for believing in me during this year and for all his care and support. This thesis is dedicated firstly to my mother, my father, my brother and to all my family who made my studies possible all these years and support me in any way possible.

Contents

Background

1.1 Introduction

Transposable elements (TEs) have been characterized for a long time as selfish sequences parasiting eukaryotic genomes which are outnumbered by TEs (Kidwell, 1993). It is known that TEs are potent mutagenic agents able to shape their hosts genomes, as a TE insertion may disrupt functional regions of the genome. They appear to have great diversity in structure and mechanisms of transposition. However, all of them are able to transpose and increase their copy number within eukaryotic chromosomes (Wicker et al., 2007). Last years many studies propose that the properties which lead TEs to be labeled as 'junk DNA' might have drived TEs to develop plasticity to genomes providing diversity and might have played a crucial role in enhancing the evolutionary potential of their hosts (Kidwell and Lisch, 2000). Recent progress in new sequencing technologies together with genomic material from large number of accessions has given the opportunity to quantify genome-wide polymorphisms for annotated and novel TE insertions. Still many aspects of TE dynamics at the population level remain unclear. Fortunately, a broad variety of computational methods and software tools have being developed for comprehensive genomic analyses that discover, annotate new TE families and reveal polymorphic TE to get new insights and explain the impact of fixed and polymorphic TEs in genomes and to disentangle selfish behaviour from coopted functions (Goerner-Potvin and Bourque, 2018).

1.2 Transposable elements drive genetic variation

1.2.1 Transposable element classification

There are two main classes of TEs according to their mechanism of transposition, which may be RNA-mediated or DNA-mediated. Retrotransposons which can be described as either cut and paste (Class I TEs) and DNA transposons, copy and paste (Class II TEs) (Wicker et al., 2007) .

1. Retrotransposons: Class I TEs replicate in two steps. The are transcribed and RNA intermediates are reverse transcribed using reverse transcriptase in the process. The copied DNA is then inserted back into the genome at a new genomic location. Retrotransposons are divided into three main orders: (i) Retrotransposons (LTR retrotransposons), with long terminal repeats (LTRs), which encode reverse transcriptase, similar to retroviruses; (ii) Retroposons (non-LTR retrotransposons), long interspaced nuclear elements (LINEs, LINE-1s, or L1s), which encode reverse transcriptase but

lack LTRs, and (iii) short interspersed nuclear elements(SINEs) (non-LTR retrotransposons) which do not encode reverse transcriptase.Retroviruses can also be considered TEs.

2. DNA transposons: Class II TEs use cut-and-paste transposition mechanism, which involves only transpose enzymes. Some type of transposases may target non-specifically any site of DNA, while others bind to specific target sequences. A staggered cut is made at the target site producing sticky ends. The DNA transposon is cut down and ligates it into the target site. DNA polymerases repair the resulting gaps by filling them in and DNA ligase joins the two ends. This procedure leads to Target Site Duplication (TSD) and the insertion sites of DNA transposons are easy to distinguish by short direct repeats followed by inverted repeats. These sites are important for the TE excision by transposase. It is possible for cut-and-paste TEs to be duplicated if their transposition takes place during the S phase of the cell cycle.

Autonomous or non-autonomous transposition may happen in both Class I and Class II TEs. Autonomous TEs mobile by themselves and non-autonomous TEs require the presence of another TE to move.

1.2.2 Rice Transposable Elements

Over 40% of rice genome is repetitive DNA and the majority of it is related to TEs. The class 1 (LTR) retrotransposons comprise the largest component of rice mobilome (14% of the genomic DNA) but, numerically, the short (<500 bp), non-autonomous class II Miniature Inverted–repeat Transposable Elements (MITEs) form the largest group with over 100,000 elements divided into hundreds of families comprising about 6% of the genome (Jiang et al., 2003).

LTR retrotransposons have contributed in the transposition-driven genome dynamics which shaped the architecture and size of the rice genome and have also been found to play a major role in the process of speciation and diversification of this crop (Zhang and Gao, 2017). Rice genome harbors 300 families of LTR-retrotransposons, belonging to either Gypsy or Copia superfamilies (Chaparro et al., 2007).

MITES have been found to be located very close to plant genes and perhaps providing coding sequences or poly(A) signals, affecting the expressions of the nearby genes, leading to hypotheses that MITEs also play major role to gene regulation and evolution (Lu et al., 2012). In rice genome among the studied TEs, MITEs exist at the highest copy number with hundreds of MITE families discovered in the rice genome, but the full picture of the transpositional landscape in this crop still remains unknown (Oki et al., 2008; Jiang et al., 2004).

To test the hypothesis that TEs represent a source of evolution, it is important to directly observe TE families that are still active and attain high copy numbers. In *O. sativa* genome most MITEs are fixed except from the currently active *mPing* family. The fact that the non-autonomous *mPing* elements avoid inserting into exons, but prefers promoter regions and that has evolved to target neutral regions creating new alleles and novel regulatory networks, makes it a great example to further explore the transpositional landscape of this element in a population scale (Naito et al., 2014; Naito et al., 2009; Jiang et al., 2003; Lu et al., 2017). Also, non-autonomous *Copia-like* LTR retrotransposons *Tos17* have been described as active in cultivated rice and with the *Gypsy-like fam106* are also found to be inserted very close to genes (Sabot, 2014; Carpentier et al., 2019). *Karma*, a LINE non-LTR retrotransposon was also identified as transpositionally active in rice and is affected by *Tos17*'s mobilization (Huang et al., 2009).

1.3 Genetic variation

The term variant can be used to describe an alteration between two genomes. There are three categories of genetic variants:

1. Single Nucleotide Variants SNVs (or Single Nucleotide Polymorphisms-SNPs) define a substitution of a single nucleotide in a specific region, and may be a transition or a transversion leading to synonymous or non-synonymous, missense or nonsense variants.
2. Indels are small insertions or deletions of hundreds of base-pairs.
3. Structural Variants are larger alterations in the DNA sequence like Chromosomal rearrangements (Deletions, Insertions, Inversions, Duplications and Reciprocal translocations) and Copy Number Variations.

As TEs can create many types of rearrangements, the mutagenic activity of mobile DNA is a double-edged sword. On one hand if the alteration happens in important genomic sequences, they will have negative effects on the fitness of the host. On the other hand TE-mediated mutations can be beneficial to the host under certain conditions (Volff, 2006).

1.4 3000 genome project

To locate TEs in a large gene pool requires genomic data for a comprehensive sample of accessions and a good-quality reference genome sequence from which TEs have been well characterized. These resources are available for a few crop species. One of the most suitable model species for this kind of study is Rice (*Oryza sativa*). 3000 genome project (3KGP) is a collection of resequenced 3,024 rice accessions from 89 countries with good average sequencing depth (14×), high average genome coverages and mapping rates of 94.0% and 92.5%, respectively (Li, Wang, and Zeigler, 2014; Alexandrov et al., 2015). Approximately 18.9 million single nucleotide polymorphisms (SNPs) were detected in rice after the alignment to the reference genome of the *Oryza japonica* Nipponbare variety. These data give great opportunity for large-scale bioinformatics analysis of polymorphic TE insertions in order to understand the genomic diversity within *O.sativa* at a higher level of detail (Access, 2014).

1.5 Detecting TE insertions

1.5.1 Definitions

Reference genome refers to the genome on which the sequence mapping was performed. It is the digital DNA sequence database, assembled as the representative example of a species' genome.

Mapping refers to the process of aligning short reads to a reference sequence, whether the reference is a complete genome, transcriptome, or de novo assembly.

Concordant pairs are properly aligned reads.

Discordant pairs are improperly aligned reads, important to identify genome alteration events.

Both concordant and discordant pairs refer to paired-end reads. Their distinction is related to whether they fulfil certain criteria. Typically, the R1 mate should be in the forward direction, whereas the R2 mate in reverse. Also, the distance between them should be within a certain range (in Illumina paired-end reads, this distance is about 500bp, $\pm 1SD$).

Genetic variation in the presence or absence of TEs is an important source of variability between individuals of the same species. In order to accurately map the locations of TE presence/absence variants with respect to a reference genome in Whole Genome Sequencing (WGS) data as the 3KGP, the urge for scaled bioinformatics tools to meet the demands of these data-intensive projects is more important than ever. Various approaches for detection of TE polymorphisms between one individual and the reference genome have been implemented, but few ensure fast and accurate analysis to succesfully unrevealing the transpositional landscape in a population level. Short-read TE detection is currently the most suitable way to detect TE insertions in existing data from population-scale WGS projects (Goerner-Potvin and Bourque, 2018).

1.6 Detecting Natural Selection

1.6.1 Selective Sweeps

When natural selection benefits a new allele, positive selection is operating favoring the individuals that carry it. The favored allele increases in frequency and if it will manage to overcome the effect of random genetic drift, it will eventually fixate in the population. As the beneficial allele increases, *neutral* genetic variants that happen to be present in the proximal genetic background of the beneficial allele, will also become more prevalent. This phenomenon is called genetic hitchhiking (John Maynard Smith, 1974). Because of genetic hitchhiking, the neighboring linked diversity diminishes, creating so-called selective sweeps. Positive selection can then be detected in genomes by searching for distinct footprints introduced by selective sweeps, such as (i) regions of reduced variation, (ii) a specific shift of the site frequency spectrum, and (iii) particular Linkage Disequilibrium (LD) patterns in the region (Pavlidis et al., 2013; Pavlidis and Alachiotis, 2017).

1.6.2 Composite Likelihood Ratio tests (CLR)

An efficient approach to analyze Next Generation Sequencing (NGS) data from whole genomes at different geographic locations and environmental conditions, are Composite Likelihood approaches. Composite Likelihood calculates likelihoods in a subset of the genetic data, and then combines them as if each subset of the data were independent (Nielsen, 2005). This method supports the separation of a large dataset into smaller pieces, for each of which the likelihood function can be calculated. Calculation of a likelihood score for the possible existence of selection in regions of sampled genotypes, under a neutral model, provides likelihood scores under the null hypothesis. Thus, obtaining values of the (likelihood) statistic under the null hypothesis of no selection it can be used to perform hypothesis testing and calculate threshold values.

1.7 Purpose

In 2016, Wildschutte et al. developed a pipeline to discover polymorphic HERV-K retrovirus insertions in human populations using data from the 1000 Human Genome Project (1KGP) (Wildschutte et al., 2016). In 2018, during a study at the CBML (EvoLab group) in FORTH-ICS under the supervision of P.Pavlidis, Wildschutte's results were used to investigate the selection forces in the nearby genomic regions of HERV-K insertion sites of haplotypes from homozygous human individuals either for the presence or the absence of the retrovirus. The application of selective sweep detection algorithms in 26 insertions (that were fulfilling certain quality criteria) suggested positive selection in 8 insertions. Also, five reported sites under positive selection are related to individuals for which their homologous genomic regions do not contain a retrovirus. In addition, subsequent expression analysis of nearby genes revealed differential expression leading to hypothesis that viral insertions affected actually the genomic areas around them. Such results motivated us to examine the evolutionary forces in genomic neighborhoods of polymorphic TE insertions in other species. We applied the Mobile Element Locator Tool (MELT), a population-scale mobile element for discovery of new TE insertions on samples from 3000 Rice Genome Project (3GKP), already used for TE detection in 1KGP (Gardner Eugene J. et al., 2017). We also used data published from the Panaud team in 2019 using TRACKPOSON tool which identified 32 families of retrotransposons and more than 50,000 TE insertion polymorphisms in the 3000 rice genomes (Carpentier et al., 2019). Insertion sites used in application of selective sweep detection algorithms to provide evidence for recent and strong positive selection around these regions (Pavlidis et al., 2013).

Materials and Methods

2.1 Data

The previously published 3000 rice genome raw sequencing data provide Genotype, Phenotype and Variety information data for rice (*Oryza sativa* L.) called against Nipponbare reference Os-Nipponbare-Reference-IRGSP-1.0, which are available from GigaScience Database (`https://doi.org/10.5524/200001`). In order to make our analysis to run fast, we collected from this gene pool a random sample of 100 individuals. From these 100 samples, 86 BAM files were available and used for de novo discovery of TE insertion sites. In addition, all meta-information for the 86 random samples was extracted from available tables from the International Rice Genebank Collection at the International Rice Research Institute and from the China National Crop Genebank and the Chinese Academy of Agricultural Sciences (CAAS) working collections. We also recreated the SNPs and allele information matrix of 20 million SNPs × 3000 rice lines from the International Rice Informatics Consortium `http://snp-seek.irri.org/_download.zul`) using PLINK (BEDtoVCF) to examine selective sweeps near the insertion loci.

On January 2019, a full matrix of presence/absence of TE insertions in the 3000 rice genomes for 32 families of retrotransposons was created by Panaud's team and became available at `http://gamay.univ-perp.fr/~Panaudlab/TRACKPOSON_Results.tar.gz` (Carpentier et al., 2019). We extracted all insertion sites for 3 TE families for the same 86 accessions we used for our initial TE insertion discovery. Furthermore, from the information tables, we tracked all 44 accession names and information sequenced from the region of Nepal and extracted all insertion sites from the presence/absence matrix for the same 3 TE families for the 44 Nepal accessions.

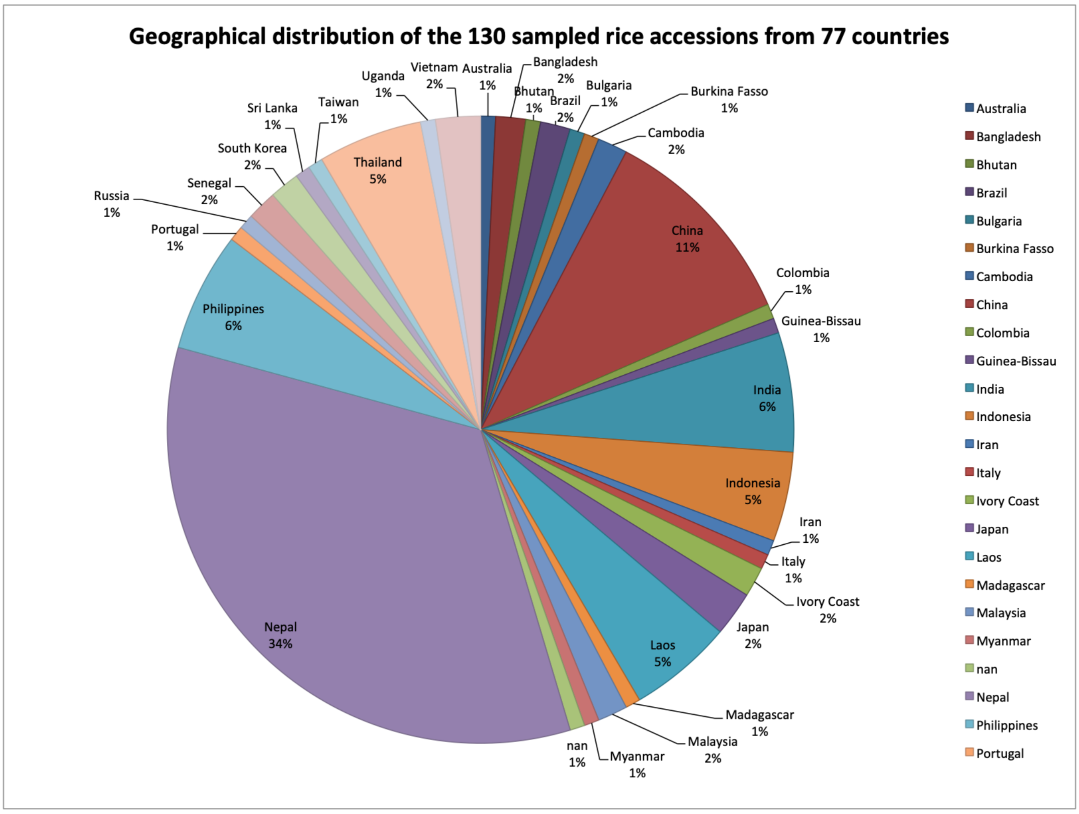
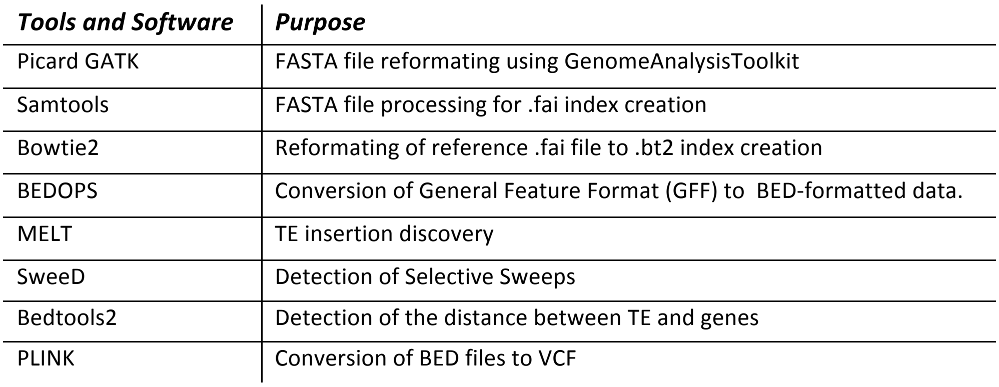


Table 2.1: Geographical distribution of the 130 sampled rice accessions from 77 countries used for selective sweep analysis.

2.2 Pipelines Software and Tools

We developed an open source pipeline available at `https://github.com/Joannagare` to perform TE insertion discovery on a population-level and also detect recent and strong positive selection in the TE insertion sites. All scripts and pipelines used for this analysis are documented in the aforementioned GitHub repository. Briefly, the repository contains the two main software were employed for the MSc thesis (Mobile Element Locator Tool-MELT and SweeD), a collection of tools and in-house written scripts. Several code scripts were properly modified for the specific analysis from the github repository of another thesis of the group `https://github.com/kutsukos/SweeDKutsukosWorkflow`.



Tools and Software	***Purpose***
Picard GATK	FASTA file reformating using GenomeAnalysisToolkit
Samtools	FASTA file processing for .fai index creation
Bowtie2	Reformating of reference .fai file to .bt2 index creation
BEDOPS	Conversion of General Feature Format (GFF) to BED-formatted data.
MELT	TE insertion discovery
SweeD	Detection of Selective Sweeps
Bedtools2	Detection of the distance between TE and genes
PLINK	Conversion of BED files to VCF

Table 2.2: Table of Software used

2.3 Novel TE insertions discovery

To perform the discovery and annotation of non-reference TE insertion sites we developed a TE insertion detection pipeline using (MELT). Mobile Element Locator Tool (MELT), was developed as part of the 1000 Genomes Project and has already been tested for TE insertion discovery on such a large scale. For example, in a 2017 study MELT outperformed existing TE insertion discovery tools in terms of speed, scalability, specificity, and sensitivity, revealing extensive TE insertion diversificatiion across distinct human populations (Gardner Eugene J. et al., 2017). MELT collects all discordant pairs from a WGS alignment, aligning them to provided TE insertion reference sequences. Next it 'walks' across the reference genome classifying putative TE insertions based on total read support at each putative site. It then merges the initial TE insertion calls across the available datasets, and analyzes in detail the breakpoints for each putative TE insertion. All sites are genotyped and filtered based on true positive calls.

We run the MELT algorithm for the discovery of transposon MITE/*mPing* in 86 individual samples of *O.sativa*. To generate TE insertion call sets, we first downloaded the 86 BAM files from GigaScience Database (`https://doi.org/10.5524/200001`), the MITE/*mPing* reference sequence from `https://www.ncbi.nlm.nih.gov/genbank/`and the *O.sativa* japonica Nipponbare reference sequence from `http://rice.plantbiology.msu.edu`.

We edited MITE/*mPing* sequence NormalizeFasta within the Picard Tools package (`http://broadinstitute.github.io/picard/`) to further use it the preprocess of reference sequences in MELT which performed using MELT-BuildTransposonZIP module.

The error rate for mPing was set to 3, that is the number of allowed mismatches by MELT per 100 bases of the TE insertion reference during alignment. *mPing* is a transposon which does excise precision and has low mutation rates (Kazuhiro et al., 2003; Nakazaki et al., 2003; Lin et al., 2006). *mPing* discovery was performed using MELT-Split runtime (Figure 2.1) with default parameters in all cases, except for the coverage sequencing depth which was set to 14×. Only PASS sites were included in final VCF files used for further analysis. TE insertions that could not be genotyped (. / .) were also filtered out.

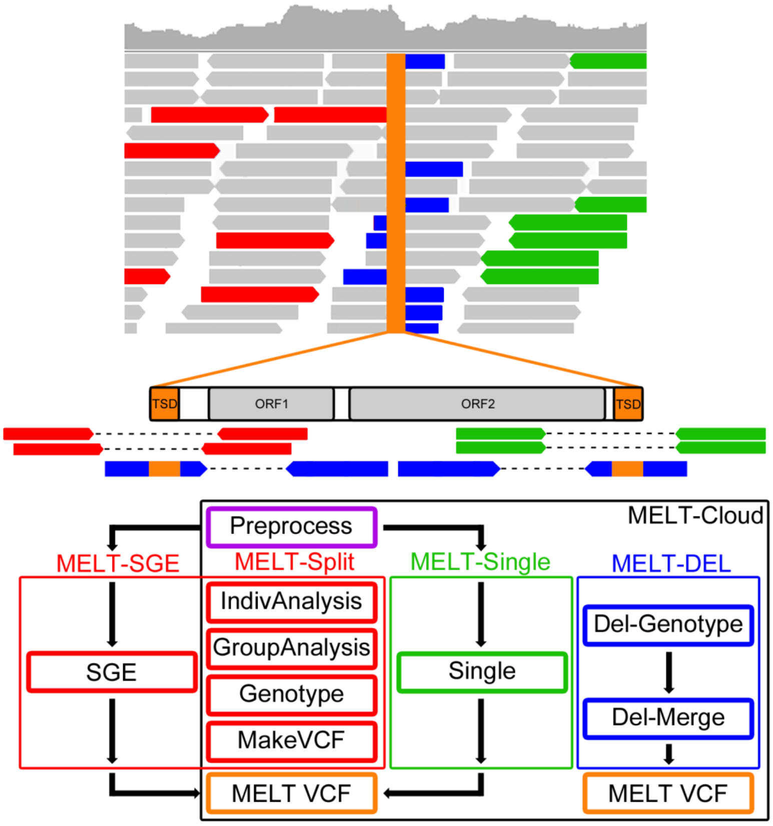


Figure 2.1: MELT performs TE insertion discovery using Illumina WGS paired end reads. (A) MELT uses two types of evidence to ascertain the location of MEIs: discordant read pairs (DRPs) and split reads (SRs). MELT first uses DRPs that map to both the reference genome (top panel) and an ME sequence (bottom) on both the left (red arrows) and right (green arrows) side of the insertion site to determine the approximate location of an TE insertion. MELT then uses SRs (blue arrows) that align to both the reference genome (top) and the TE (bottom panel) to determine the precise location of the insertion site and the target site duplication (TSD; Orange). (B) MELT performs non-reference and reference TE insertion discovery through multiple processing pipelines. Analysis of population scale data (red box) can be performed using either the built-in SGE scheduler (MELT-SGE), or adapted to other parallel computing environments (using MELT-Split). MELT also can rapidly analyze a single genome (green box) using MELT-Single, or genotype reference TE insertions (blue box) using the MELT-DEL pipeline.

2.4 Filtering the insertions loci

In order to filter and keep only the polymorphic TE insertions that are present (in homozygous state) in at least 10 individuals and absent (in homozygous state) in at least 10 individuals of each sample used, we applied an in-house python script that provides a filtered output VCF file appropriate for further use and selective sweep detection. The script is available from `https://github.com/kutsukos/VCFilterbySampleQuan`.

2.5 Distance between TE insertion and gene estimation

Distance estimation between TE insertion and the reference gene locations was performed with bedtools (v. 2.25) between the Nipponbare gtf annotation file (IRGSP-1.0-predicted-transcript-exon-2019-06-26.gtf) and the output of MELT (filtered and not filtered dataset).

2.6 Detection of Selective Sweeps

We investigated the adaptive role of polymorphic TE insertions. Our hypothesis is that some TE insertions may be beneficial, producing the characteristic footprints of positive selection in the proximal genomic locations (for example, the shift of the Site Frequency Spectrum).

2.6.1 SweeD

We used a high performance software, called SweeD, to detect loci on which positive selection has recently operated. The Sweep Detector (SweeD) is a tool based on likelihood calculation and detects sweeps in whole genomes by analyzing Site Frequency Spectra of Single Nucleotide Variant frequencies in a given sample (Pavlidis et al., 2013). SweeD is a High Performance Computation (HPC) software able to analyze thousands of whole genome datasets in relatively small computer clusters or off-the-shelf laptops within a few hours. Our purpose was to detect candidate TE insertions for positive selection by exploring the selective sweeps footprints in the haplotypes that carried the TE insertions. We applied the SweeD algorithm in a window of 500,000 base pairs around the TE insertion site. Within that locus, we chose to test for that kind of events every 5,000 base pairs (gridpoints).

Significance threshold was defined to set above which likelihood score the variation was considered as under positive selection. For each dataset we ran SweeD in multiple null positions of the whole chromosomes that were distant from the TE insertion tested in the same range of window and gridpoints. From each dataset we sampled the maximum value. Those values consisted the points of the null distribution. The threshold was set at the 99.5% max of the distribution, above of which we denoted a loci as positive for selection.

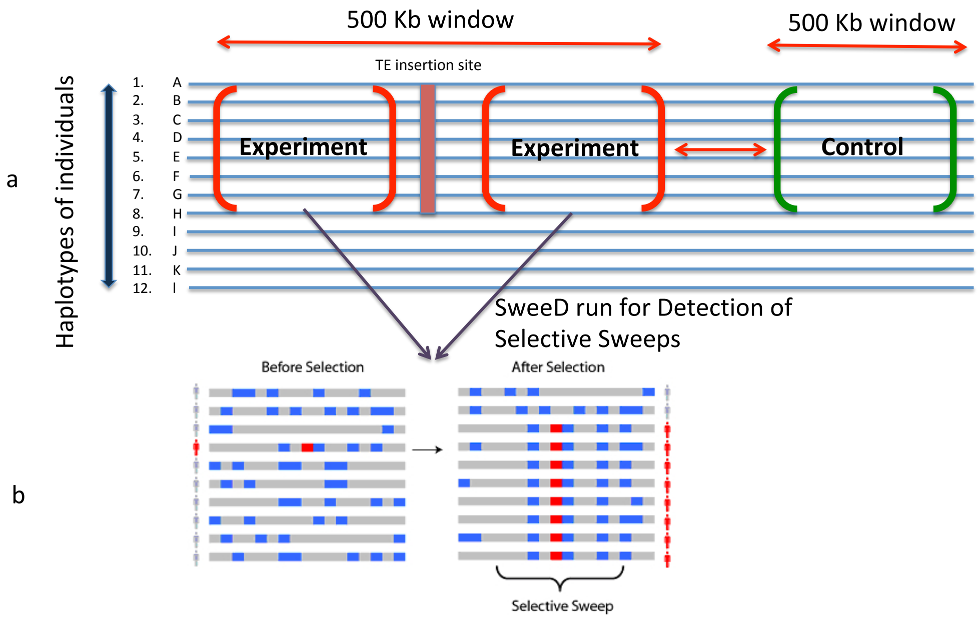
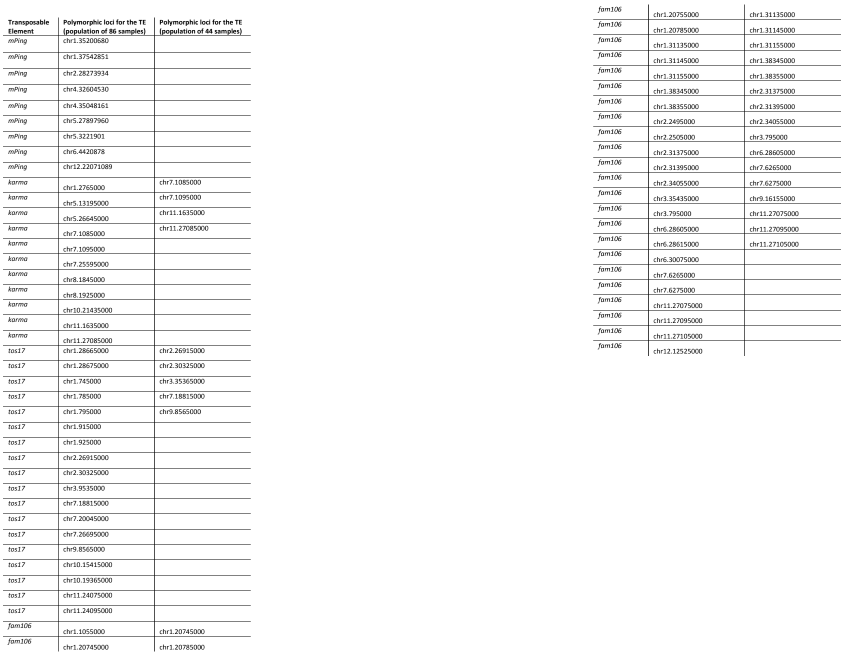


Figure 2.2: a. The pipeline used for the investigation of selective sweeps events only in the haplotypes that carried the polymorphic TE insertion using SweeD tool. b. Schematic representation the polymorphisms along a chromosome of a population, including the selected allele, before and after selection, in which neighboring linked alleles on the chromosome 'hitchhike' along with it to high frequency, creating a 'selective sweep' (`https://www.nature.com/scitable/topicpage/evolutionary-adaptation-in-the-human-lineage-12397`)

(Larribe F., 2011)



Transposable Element	Polymorphic loci for the TE (population of 86 samples)	Polymorphic loci for the TE (population of 44 samples)
mPing	chr1.35200680	
mPing	chr1.37542851	
mPing	chr2.28273934	
mPing	chr4.32604530	
mPing	chr4.35048161	
mPing	chr5.27897960	
mPing	chr5.3221901	
mPing	chr6.4420878	
mPing	chr12.22071089	
karma	chr1.2765000	chr7.1085000
karma	chr5.13195000	chr7.1095000
karma	chr5.26645000	chr11.1635000
karma	chr7.1085000	chr11.27085000
karma	chr7.1095000	
karma	chr7.25595000	
karma	chr8.1845000	
karma	chr8.1925000	
karma	chr10.21435000	
karma	chr11.1635000	
karma	chr11.27085000	
tos17	chr1.28665000	chr2.26915000
tos17	chr1.28675000	chr2.30325000
tos17	chr1.745000	chr3.35365000
tos17	chr1.785000	chr7.18815000
tos17	chr1.795000	chr9.8565000
tos17	chr1.915000	
tos17	chr1.925000	
tos17	chr2.26915000	
tos17	chr2.30325000	
tos17	chr3.9535000	
tos17	chr7.18815000	
tos17	chr7.20045000	
tos17	chr7.26695000	
tos17	chr9.8565000	
tos17	chr10.15415000	
tos17	chr10.19365000	
tos17	chr11.24075000	
tos17	chr11.24095000	
fam106	chr1.1055000	chr1.20745000
fam106	chr1.20745000	chr1.20785000
fam106	chr1.20755000	chr1.31135000
fam106	chr1.20785000	chr1.31145000
fam106	chr1.31135000	chr1.31155000
fam106	chr1.31145000	chr1.38345000
fam106	chr1.31155000	chr1.38355000
fam106	chr1.38345000	chr2.31375000
fam106	chr1.38355000	chr2.31395000
fam106	chr2.2495000	chr2.34055000
fam106	chr2.2505000	chr3.795000
fam106	chr2.31375000	chr6.28605000
fam106	chr2.31395000	chr7.6265000
fam106	chr2.34055000	chr7.6275000
fam106	chr3.35435000	chr9.16155000
fam106	chr3.795000	chr11.27075000
fam106	chr6.28605000	chr11.27095000
fam106	chr6.28615000	chr11.27105000
fam106	chr6.30075000	
fam106	chr7.6265000	
fam106	chr7.6275000	
fam106	chr11.27075000	
fam106	chr11.27095000	
fam106	chr11.27105000	
fam106	chr12.12525000	

Table 2.3: List of polymorphic loci used for the detection of positive selection

Results

3.1 Identification of *mPing* insertions and their locations using MELT

We provide a novel detection of *mPing* transposable element from MITE family in a population level as a representative for its putative adaptive role in the euchromatinic regions of *O.sativa*. MELT detected *mPing* insertions by searching the discordant read pairs (DRPs) and split reads (SRs) in Illumina WGS data that are enriched at sites containing new, non- reference TE insertions. MELT algorithm is designed to analyze BAM files, the most common format output of Illumina WGS data sets.

In the output VCF file of MELT, in INFO column when an insertion is characterized as homozygous for the individual it is shown as 1/1, and when it is shown as 0/0 it means that the sample is homozygous for the empty site. heterozygous (0/1) and no call (./.) sites are also documented in the output dataset. The total TE insertion number detected in our sample of 86 sequenced data from 76 countries was 140. Polymorphic insertion loci number that have more than 10 samples having the insertion in both haploids (1/1) and more than 10 samples with the insertion in none of the haploids (0/0) are found to be only 10 in number. It is important to mention that we worked only with the insertions that were based on the number of samples that have or not the insertion in both haploids, because the VCF file, was not phased, so in the case of (0/1), it was impossible to distinguish which haploid have an insertion or not. The following graph 3.1 demonstrates the non filtered and filtered output of MELT shown near the reference genome of *O.sativa japonica*.

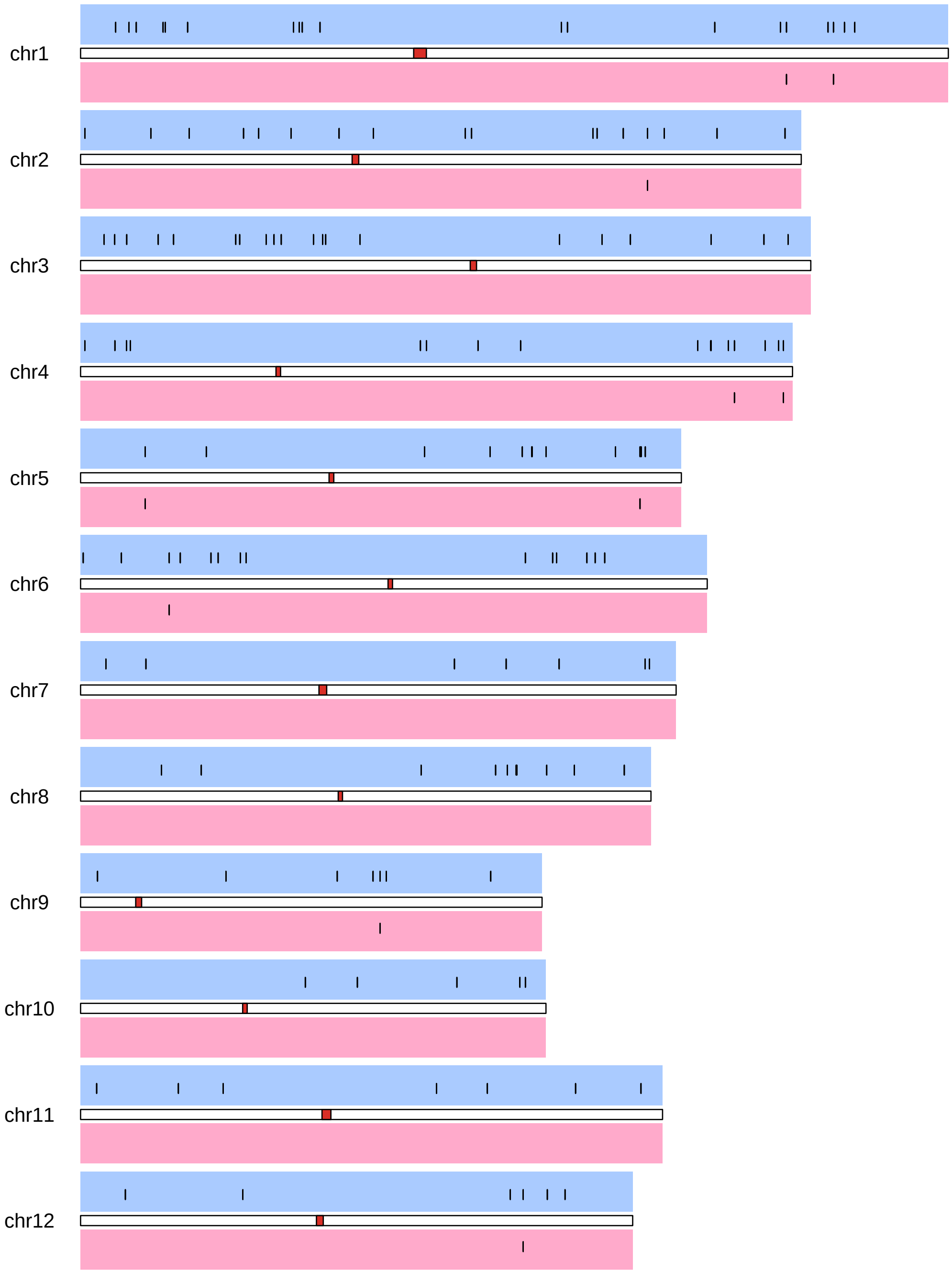
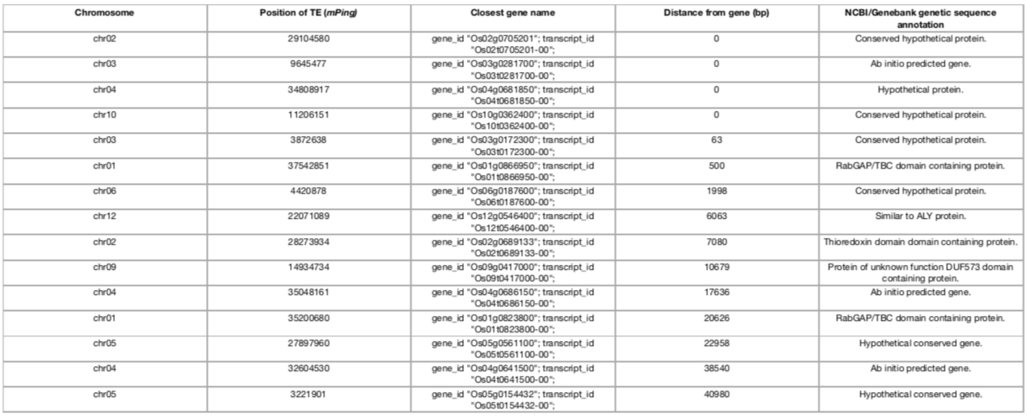


Figure 3.1: Mappability of *mPing* representation. The black tips show the mappability of the *mPing* insertions on the 12 chromosomes of rice genome. For each chromosome, the red tips correspond to the centromere. In blue background the tips show the total number of different *mPing* insertions on each chromosome that MELT software identified. In pink background the black tips show the polymorphic insertions after the filtering.

3.2 Distance between TE insertion and gene estimation

MELT algorithm requires transposon reference sequence and evaluates the exact position of the TE on the genome. On the contrary, the matrices used from Professor's Panaud work provided a window of 10.000 bp in which the TEs are located. For that reason we calculated the distance only between *mP*-

ing element insertions and the closest gene to them. The distances between the closest genes and all *mPing* insertions (polymorphinc and non-polymorphic) can be found at tables 1 and 2. Four insertion sites (chr02.29104580, chr03.9645477, chr04.34808917 and chr10.11206151) were located inside other genes. The NCBI/Genebank genetic sequence annotation is described at the table 3.1. Insertion sites chr03.3872638 and chr01.37542851 were also located very close to genes. As far as polymorphic insertion sites for *mPing* element, chr01.37542851, chr06.4420878, chr02.28273934 and chr12.22071089 were the closest to other genes with distances 500bp, 1998bp, 7080bp and 6063 respectively (table 1). We also checked weather our novel insertion sites are close to genes known to have undergone selection during rice domestication: CLDGR 16(sh4) in chr4.3451967-34765623 and CLDGR 21 in chr7.2778802-3148679, but none of them was found inside or nearby to them (Civán et al., 2015).

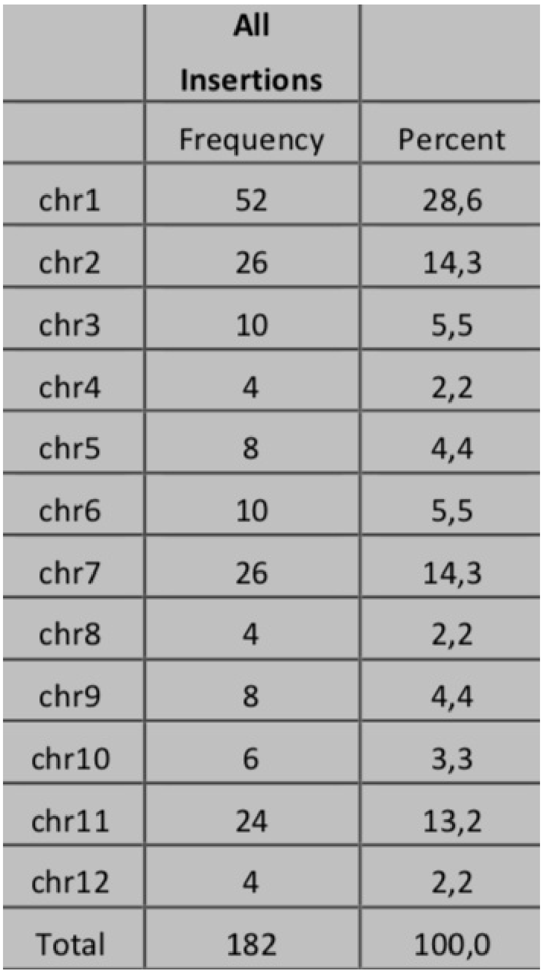


Chromosome	Position of TE (*mPing*)	Closest gene name	Distance from gene (bp)	NCBI/Genebank genetic sequence annotation
chr02	29104580	gene_id "Os02g0705201"; transcript_id "Os02t0705201-00";	0	Conserved hypothetical protein.
chr03	9645477	gene_id "Os03g0281700"; transcript_id "Os03t0281700-00";	0	Ab initio predicted gene.
chr04	34808917	gene_id "Os04g0681850"; transcript_id "Os04t0681850-00";	0	Hypothetical protein.
chr10	11206151	gene_id "Os10g0362400"; transcript_id "Os10t0362400-00";	0	Conserved hypothetical protein.
chr03	3872638	gene_id "Os03g0172300"; transcript_id "Os03t0172300-00";	63	Conserved hypothetical protein.
chr01	37542851	gene_id "Os01g0866950"; transcript_id "Os01t0866950-00";	500	RabGAP/TBC domain containing protein.
chr06	4420878	gene_id "Os06g0187600"; transcript_id "Os06t0187600-00";	1998	Conserved hypothetical protein.
chr12	22071089	gene_id "Os12g0546400"; transcript_id "Os12t0546400-00";	6063	Similar to ALY protein.
chr02	28273934	gene_id "Os02g0689133"; transcript_id "Os02t0689133-00";	7080	Thioredoxin domain domain containing protein.
chr09	14934734	gene_id "Os09g0417000"; transcript_id "Os09t0417000-00";	10679	Protein of unknown function DUF573 domain containing protein.
chr04	35048161	gene_id "Os04g0686150"; transcript_id "Os04t0686150-00";	17636	Ab initio predicted gene.
chr01	35200680	gene_id "Os01g0823800"; transcript_id "Os01t0823800-00";	20626	RabGAP/TBC domain containing protein.
chr05	27897960	gene_id "Os05g0561100"; transcript_id "Os05t0561100-00";	22958	Hypothetical conserved gene.
chr04	32604530	gene_id "Os04g0641500"; transcript_id "Os04t0641500-00";	38540	Ab initio predicted gene.
chr05	3221901	gene_id "Os05g0154432"; transcript_id "Os05t0154432-00";	40980	Hypothetical conserved gene.

Table 3.1: Sorted distance from the closest genes found from *mPing* element.

3.3 Extraction of *Karma*, *Tos17* and *Fam106* polymorphic loci

Insertion sites for 86 individuals were extracted from the full matrix of presence/absence of TE insertions found in the 3000 rice genomes for 32 families of retrotransposons (Carpentier et al., 2019). *Karma*, *Tos17* and *Fam106* TE families were extracted for further analysis using an inhouse written script. Furthermore, insertions sites for the same 3 TE families were extracted for more 44 individuals originated from Nepal. Statistics about distribution of insertions for the 4 mobile elements (*mPing*, *Karma*, *Tos17* and *Fam106*) per chromosome are represented at table 3.2. Chromosomes 1, 2, 7 and 11 show the highest frequencies of TE insertions.

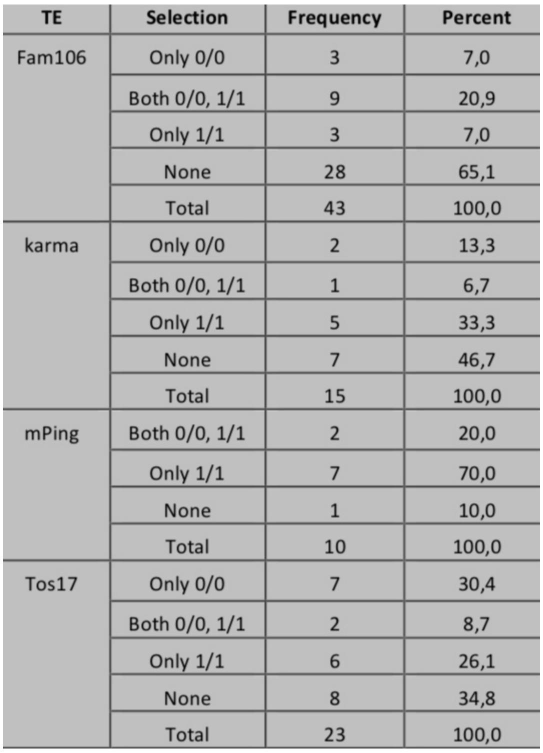


	All Insertions	
	Frequency	Percent
chr1	52	28,6
chr2	26	14,3
chr3	10	5,5
chr4	4	2,2
chr5	8	4,4
chr6	10	5,5
chr7	26	14,3
chr8	4	2,2
chr9	8	4,4
chr10	6	3,3
chr11	24	13,2
chr12	4	2,2
Total	182	100,0

Table 3.2: Frequencies of total polymorphic Transposable Element insertions per chromosome. Chromosomes 1, 2, 7 and 11 show the highest frequencies.

3.4 Likelihood-based detection of selective sweeps using SweeD

We ran SweeD for 182 haplotypes, which we analyzed separately for each TE variant and population (Table 2.3). We searched for selective footprints in 20 haplotypes for *mPing*, 30 haplotypes for *karma*, 46 haplotypes *tos17* and 86 haplotypes for *fam106*. From the 182 haplotypes we scanned for signs of strong positive selection, 94 of them (51.6%) were found to show positive selection around the insertion site either with presence or absence of the TE. 42 haplotypes (23.1%) show positive selection only on when haplotypes carried the insertion. 24 haplotypes (13.2%) showed positive selection only on haploids homozygous for the absence of the TE. In addition, 28 haplotypes (15.4%) showed selective footprints both for absence and presence of the TE. 66 of 94 haplotypes with positive selection in total, concern samples which the TEs are present. We hypothesized that those under selective events would be present in higher frequencies. There is a significant difference among the frequencies in loci which 1/1 haploids have selective footprint and none of the respective 0/0 haploids show positive selection, among the group of loci which 0/0 haploids show positive selection and the respective 1/1 haploids are negative. Similar significant percentage was found inside some TE families specifically. For example from 20 haplotypes, (meaning 10 insertion sites) checked for positive selection for *mPing* element, 7 of 10 sites found to show positive selection only when the TE was present (1/1) 3.3. Nevertheless, there was also a significant number of haploids 88/182 (48,4%), which did not show any positive selection in either category.



TE	Selection	Frequency	Percent
Fam106	Only 0/0	3	7,0
	Both 0/0, 1/1	9	20,9
	Only 1/1	3	7,0
	None	28	65,1
	Total	43	100,0
karma	Only 0/0	2	13,3
	Both 0/0, 1/1	1	6,7
	Only 1/1	5	33,3
	None	7	46,7
	Total	15	100,0
mPing	Both 0/0, 1/1	2	20,0
	Only 1/1	7	70,0
	None	1	10,0
	Total	10	100,0
Tos17	Only 0/0	7	30,4
	Both 0/0, 1/1	2	8,7
	Only 1/1	6	26,1
	None	8	34,8
	Total	23	100,0

Table 3.3: Frequencies of positive selection per Transposable Element. *fam106* show high frequency of no selection on scanned sites, *karma* shows almost double frequencies for positive selection for haplotypes that carry the insertion (1/1) and do not show selection in respective 0/0, than the opposite scenario, There is a significant high frequency (70%) for *mPing* TE when the haplotypes are selected only for the presence (1/1) of the TE.

Selective sweep analysis was done separately in two different groups of samples. The first group included 86 samples originated from 76 countries randomly selected from the 3KGP database and included analysis on 128 haplotypes. The second group included 44 samples originated from the region of Nepal from which 54 haplotypes were analysed. The purpose of this separation was in order to obtain data from an isolated region with extreme conditions, in this case the altitude, because in Nepal the cultivation of *O.sativa* is of the highest elevation of the world. This analysis will give the ability to compare these data with another isolated population with different conditions in the future. Still, we did a comparison between the two populations to test if the mean CLR scores of every haplotype can predict the separation in one of the four categories for positive selection found (only in presence of the TE, only in the absence, both in presence/absence or none). The population from Nepal showed 34.5% predictability and the sample from the variety of countries showed 1.6% predictability. There was also significant correlation ($p<0.05$) between mean CLR scores and the above separation of four categories in the population from Nepal. In the figures 3.2, 3.3,3.4, 3.5,3.6 3.7,3.8 are represented in the y axis the likelihood scores (CLR) of the 4 TE analysed by SweeD for 500kb windows with 5kb step. The samples are separated by haplotypes with absence/presence of the TE variant and polymorphic positions are shown concatenated in x axis separated by chromosome. Black and grey dots correspond to all scores calculated and colored dots correspond to CLR scores found significant for strong positive selection.

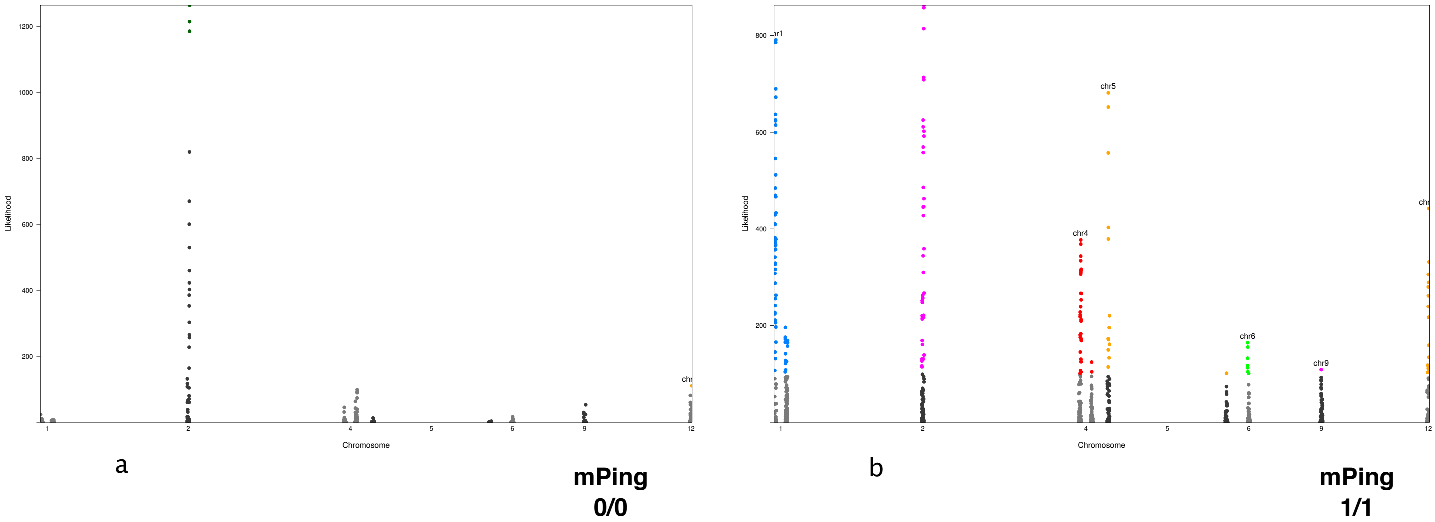


Figure 3.2: Selective sweep analysis for *mPing* family for all positions found polymorphic for the absence (a) and presence (b) of the insertion. Manhattan plot represents in y axis the Likelihood scores for each polymorphic position shown at the x axis of the TE insertion evaluated by SweeD (see Methods). TE insertion-Likelihood association CLR >thresholds in color are significative.

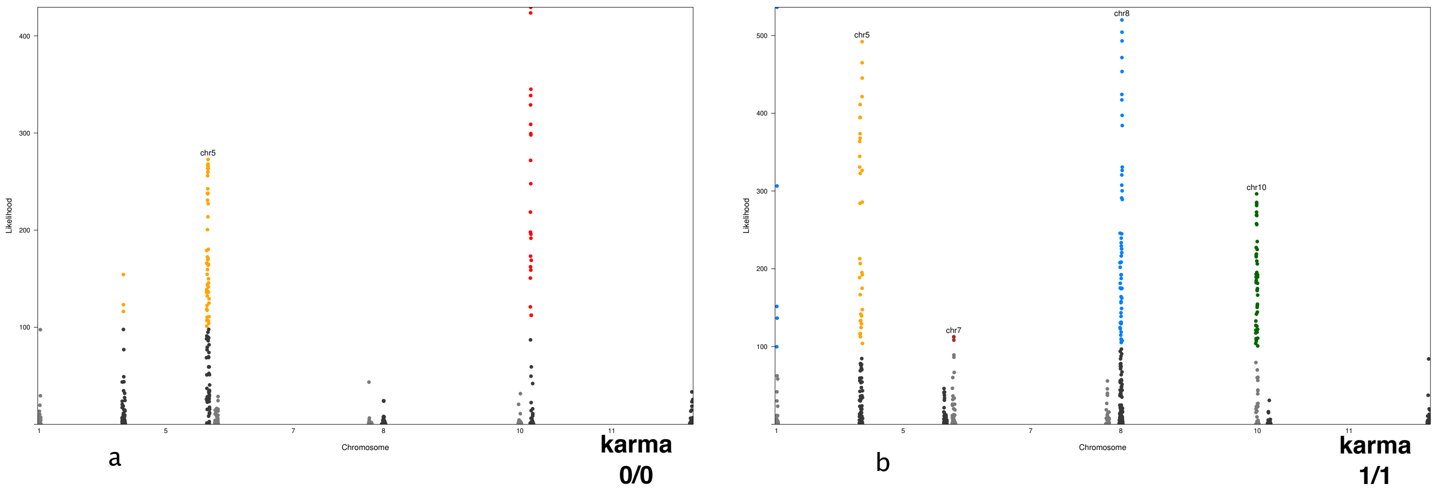


Figure 3.3: Selective sweep analysis for all polymorphic positions for *karma* family insertions for samples homozygous for the absence (a) and presence (b) of the TE for the population of 86 samples. TE insertion-Likelihood association CLR >threshold in color are significative.

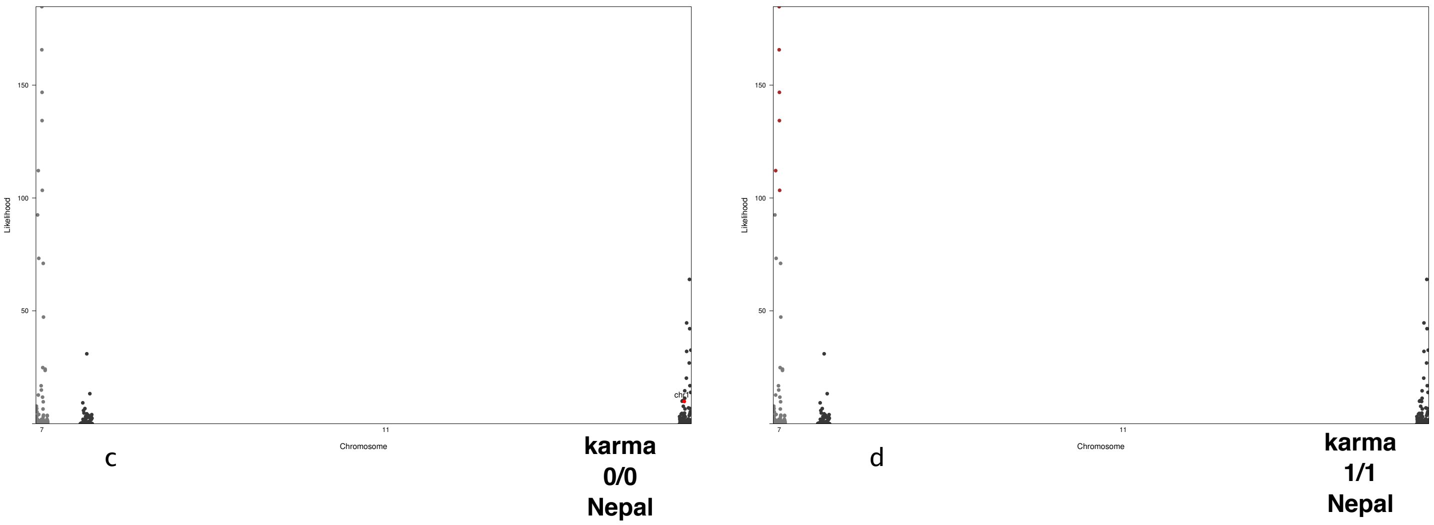


Figure 3.4: Selective sweep analysis for all polymorphic positions for *karma* family insertions for samples homozygous for the absence (c) and presence (d) of the TE for the population of 44 samples from Nepal. TE insertion-Likelihood association CLR >threshold in color are significative.

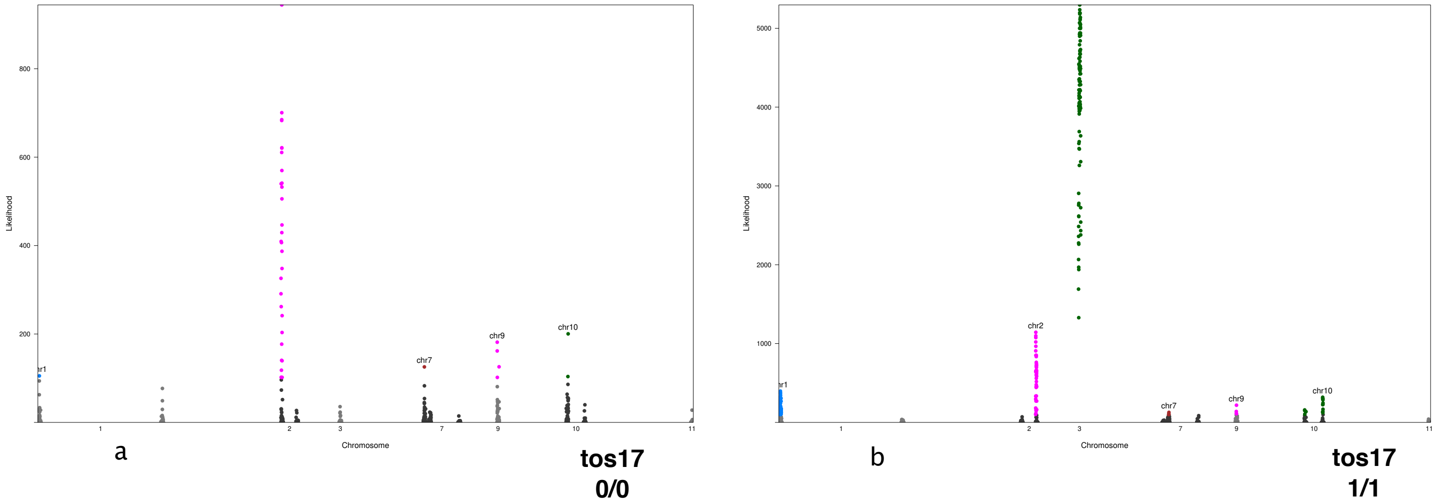


Figure 3.5: Scan of all polymorphic positions for *tos17* family for individuals homozygous for the absence (a) and presence (b) of the insertion for the population of 86 samples (top) and the 44 samples from Nepal (bottom). TE insertion-Likelihood association CLR >thresholds in color are significative.

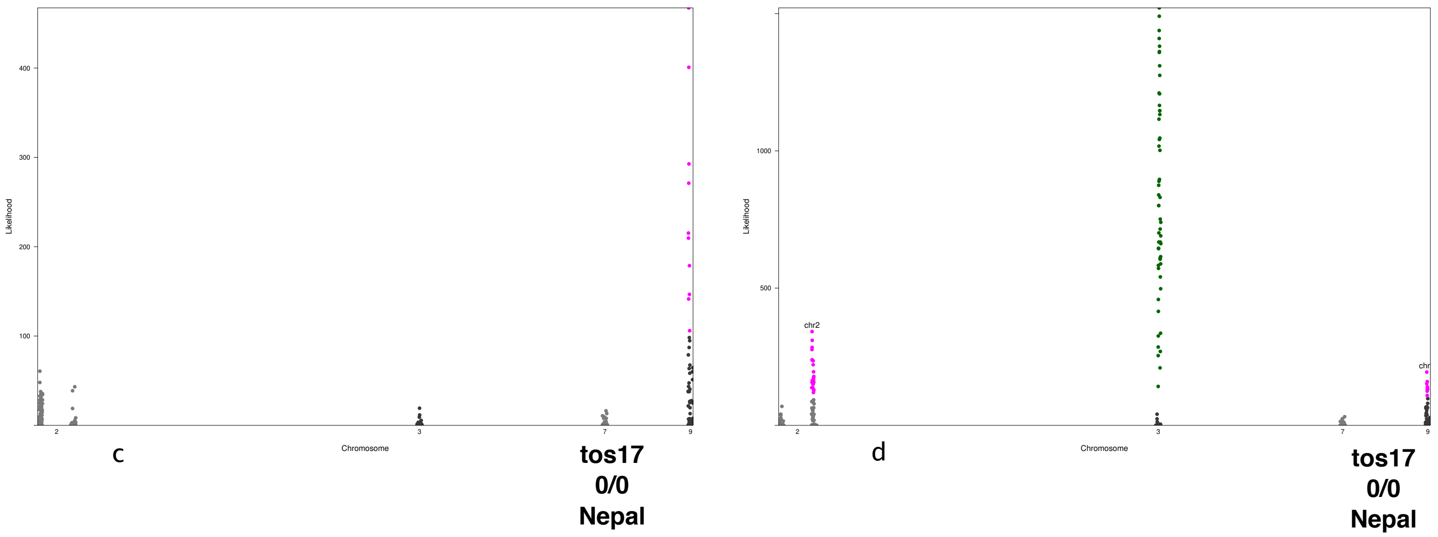


Figure 3.6: Scan of all polymorphic positions for *tos17* family for individuals homozygous for the absence (c) and presence (d) of the insertion for the population of 44 samples from Nepal. TE insertion-Likelihood association CLR >thresholds in color are significative.

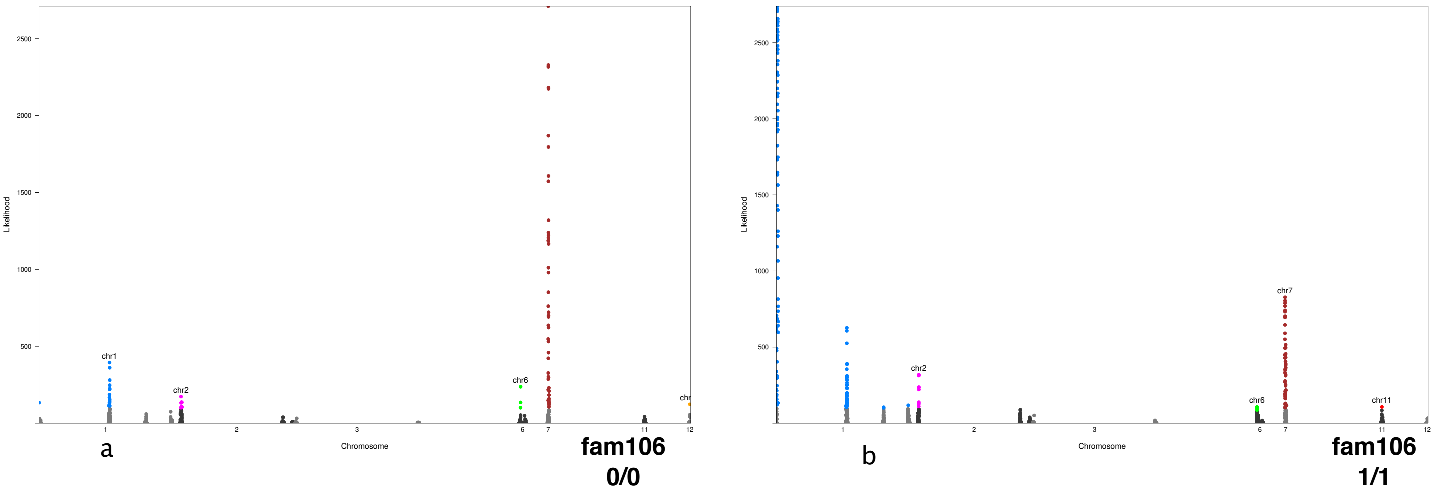


Figure 3.7: Selective sweep analysis for all polymorphic positions of *fam106* family for individuals homozygous for the absence (a) and presence (b) of the TE for the population of 86 samples. TE insertion-Likelihood association CLR >threshold in color are significative.

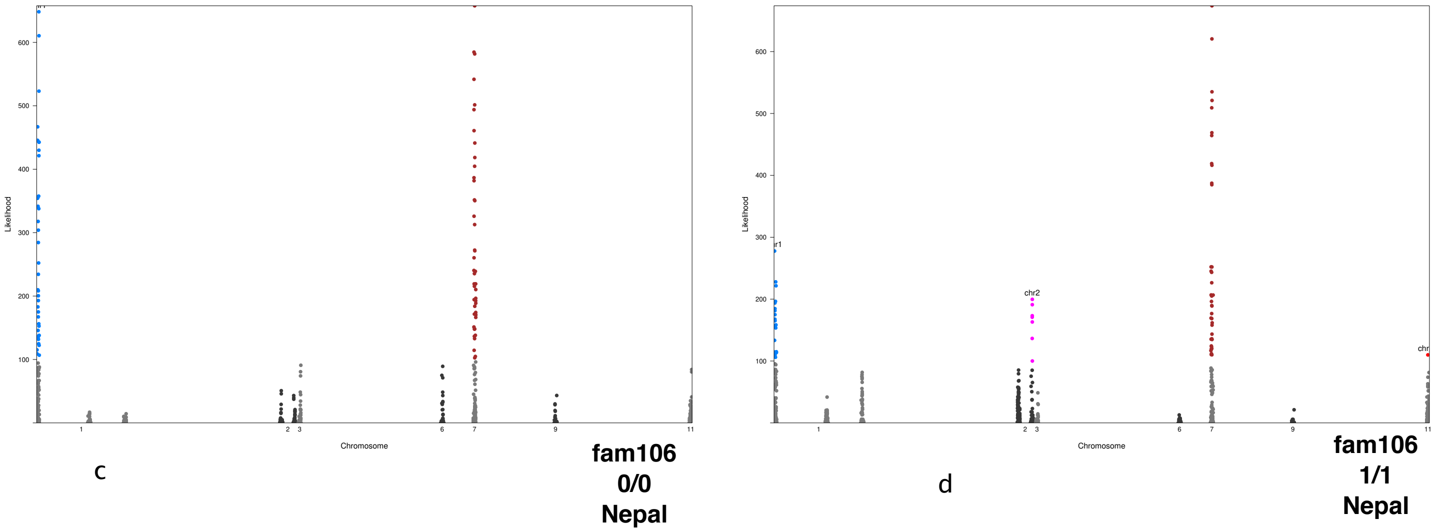


Figure 3.8: Selective sweep analysis for all polymorphic positions of *fam106* family for individuals homozygous for the absence (c) and presence (d) of the TE for the population of 44 samples from Nepal (bottom). TE insertion-Likelihood association CLR >threshold in color are significative.

Figures below 3.9, 3.10,3.11,3.12, 3.13, 3.14, 3.15, 3.16,3.17,3.18,3.19,3.20,3.21 represent the Likelihood scores in y axis for each TE at each polymorphic insertion loci separately shown in x axis for haplotypes with absence/presence respectively. Figures are grouped in 4 categories: for positive selection found in only 0/0 samples, only 1/1 samples, both samples or none of the haplotypes of each insertion loci.

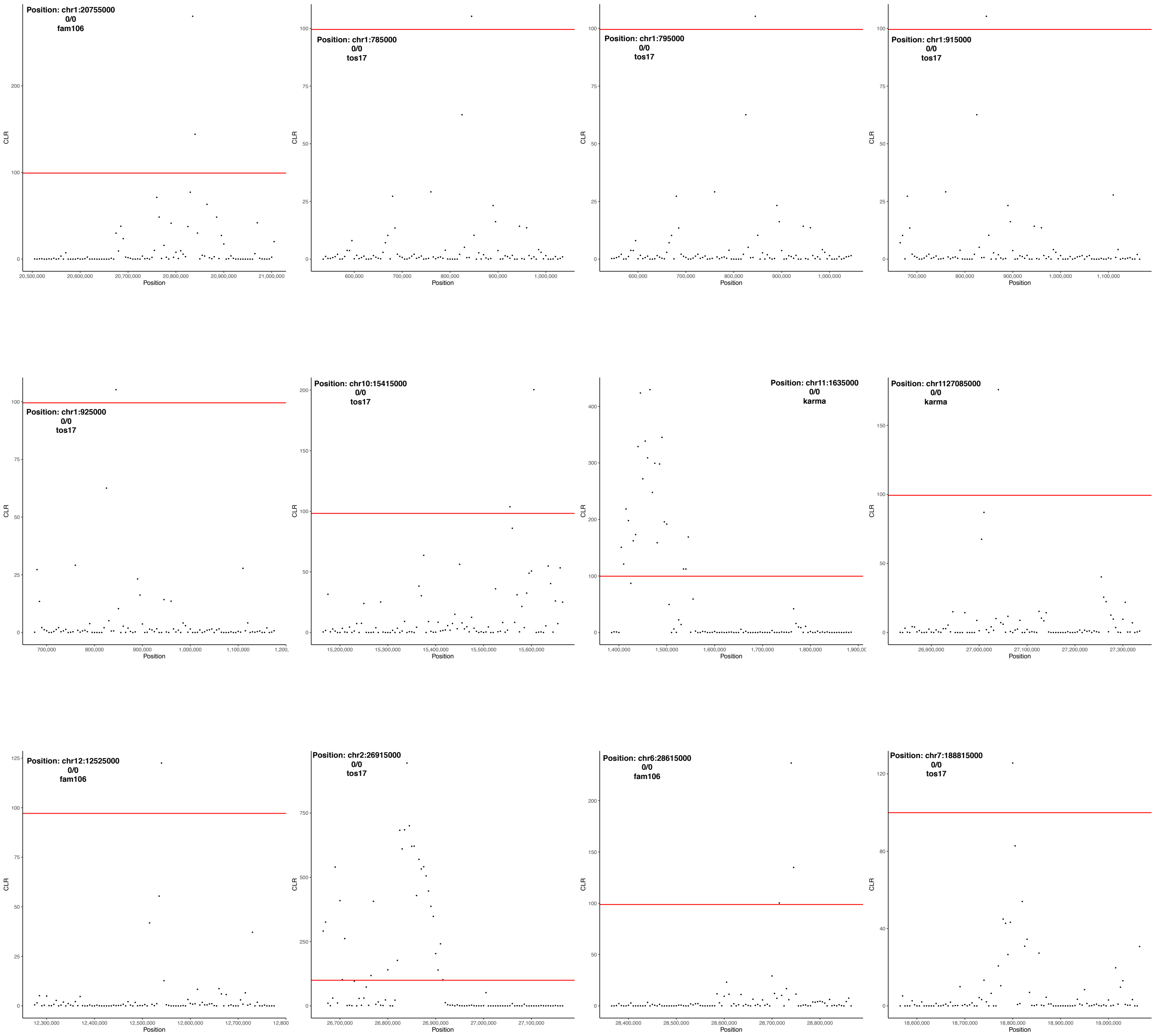


Figure 3.9: Scan of 12 polymorphic positions found significant for individuals homozygous for the absence (0/0) of each TE variant. The respective scores for 12 samples (not shown in figure) that were homozygous for the presence (1/1) of the same TE variant did not found to be significant.

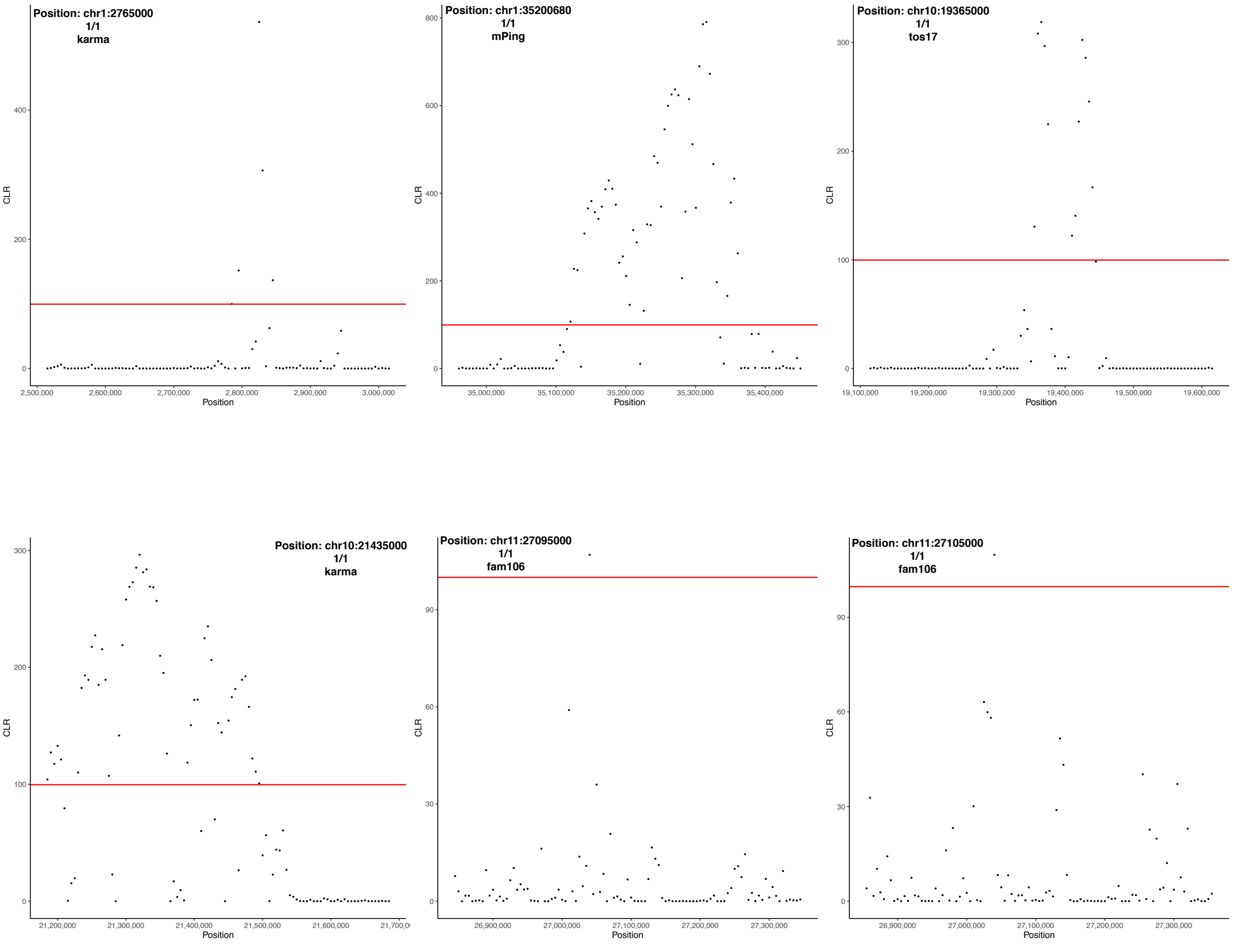


Figure 3.10: Scan of 21 polymorphic positions found significant for individuals homozygous for the presence (1/1) of each TE variant. The respective scores for 21 samples (not shown in figure) that were homozygous for the absence (0/0) of the same TE variant did not found to be significant.

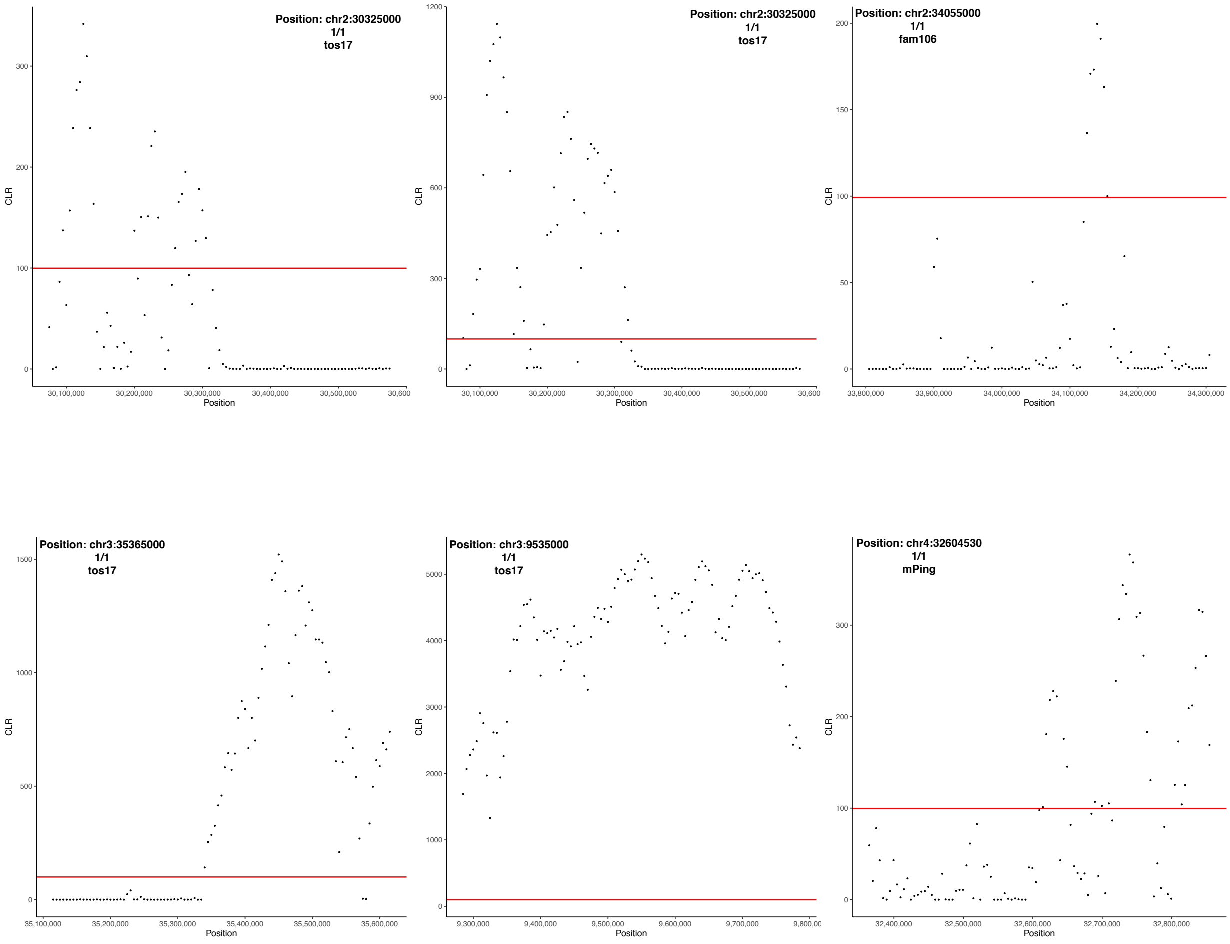


Figure 3.11: Scan of 21 polymorphic positions found significant for individuals homozygous for the presence (1/1) of each TE variant. The respective scores for 21 samples (not shown in figure) that were homozygous for the absence (0/0) of the same TE variant did not found to be significant.

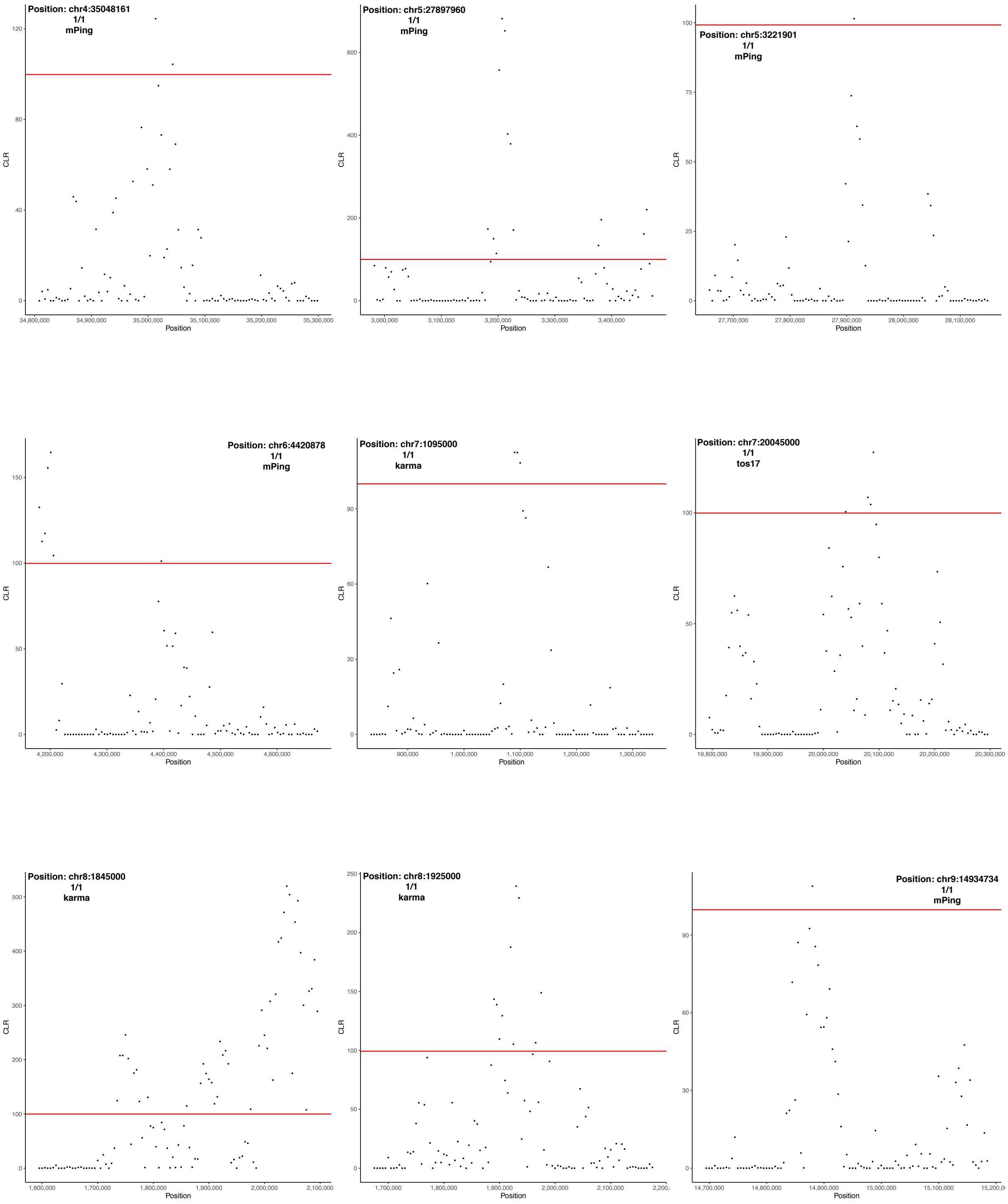


Figure 3.12: Scan of 21 polymorphic positions found significant for individuals homozygous for the presence (1/1) of each TE variant. The respective scores for 21 samples (not shown in figure) that were homozygous for the absence (0/0) of the same TE variant did not found to be significant.

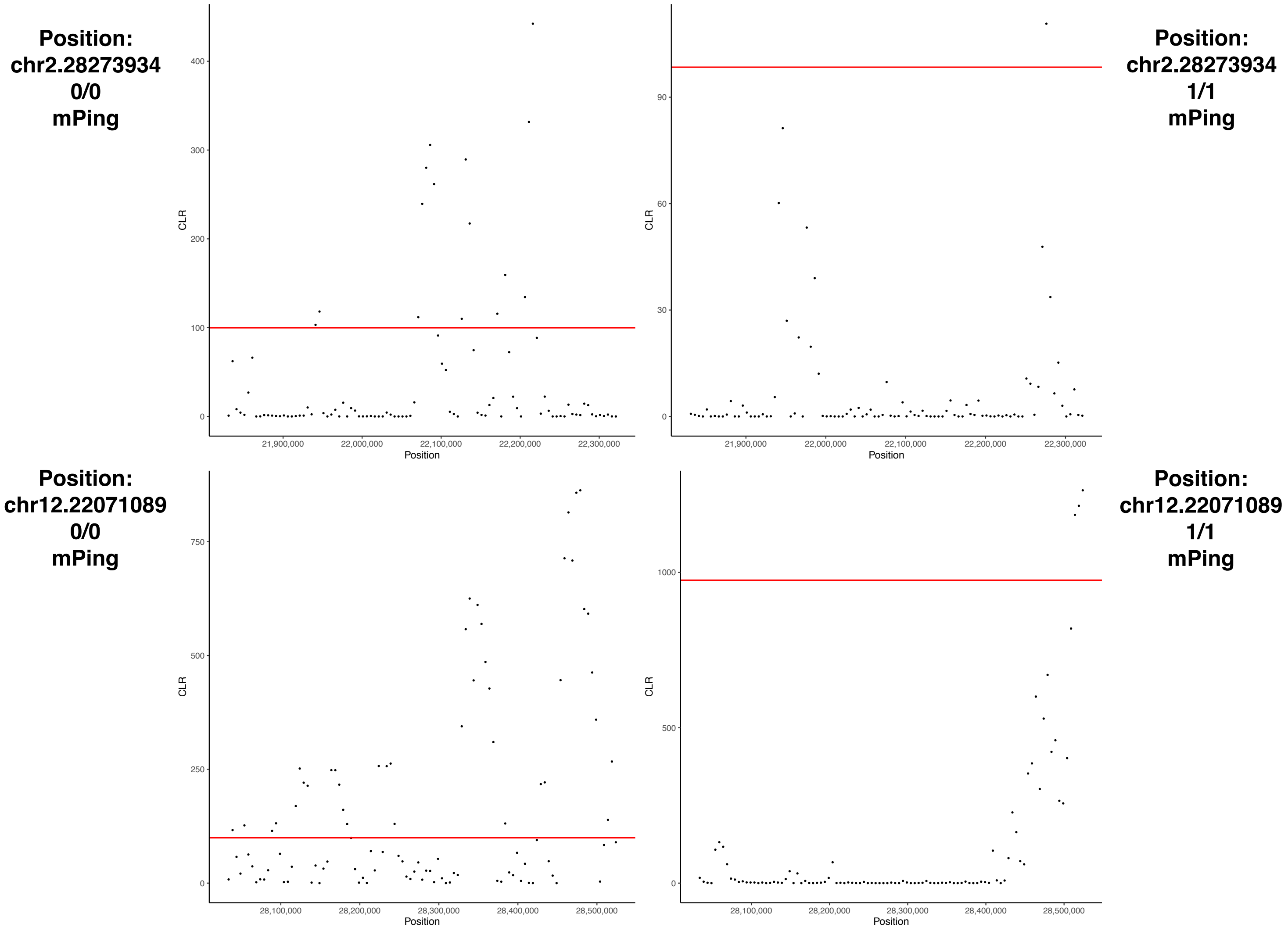


Figure 3.13: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *mPing* TE.

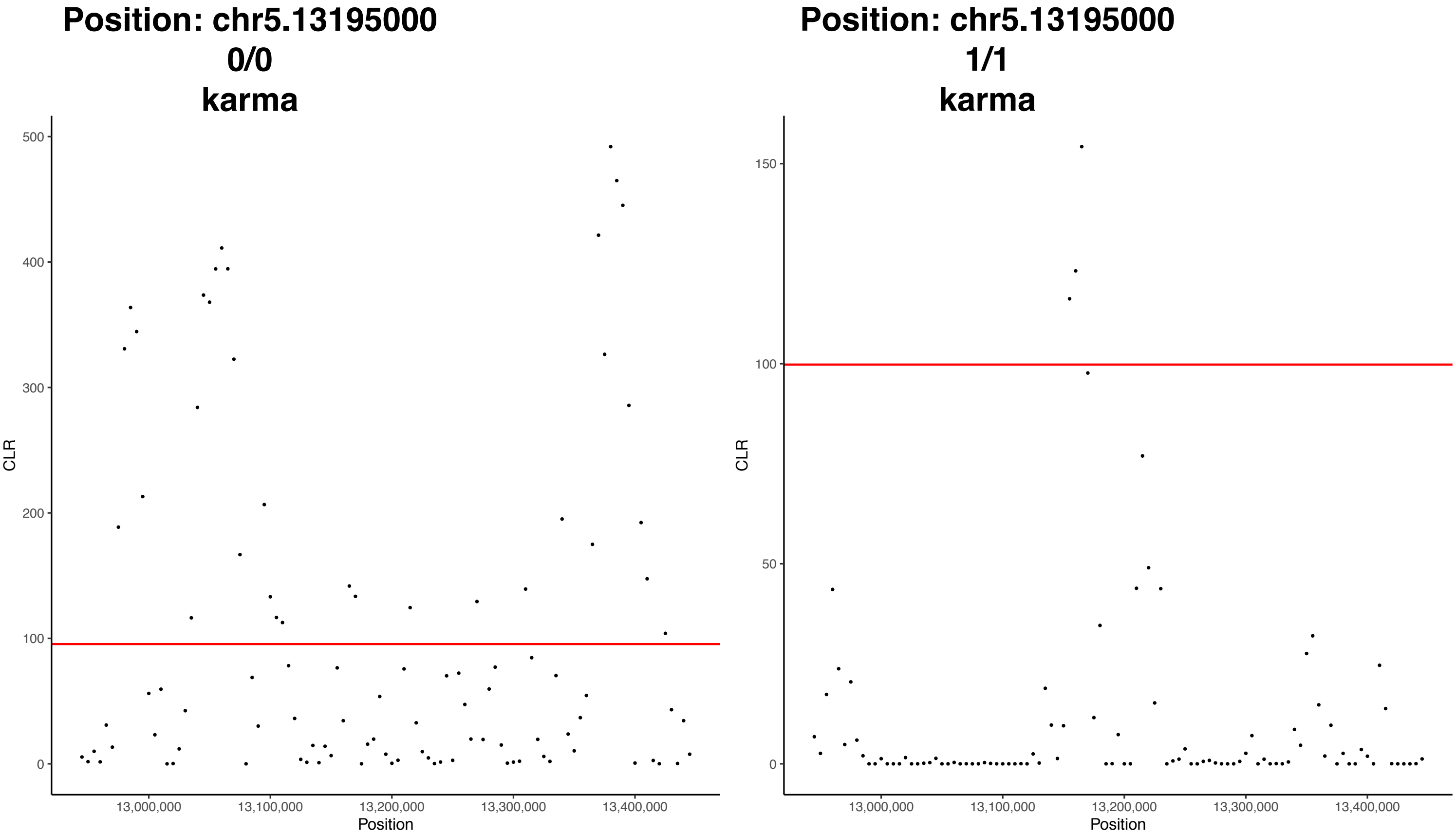


Figure 3.14: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Karma* TE.

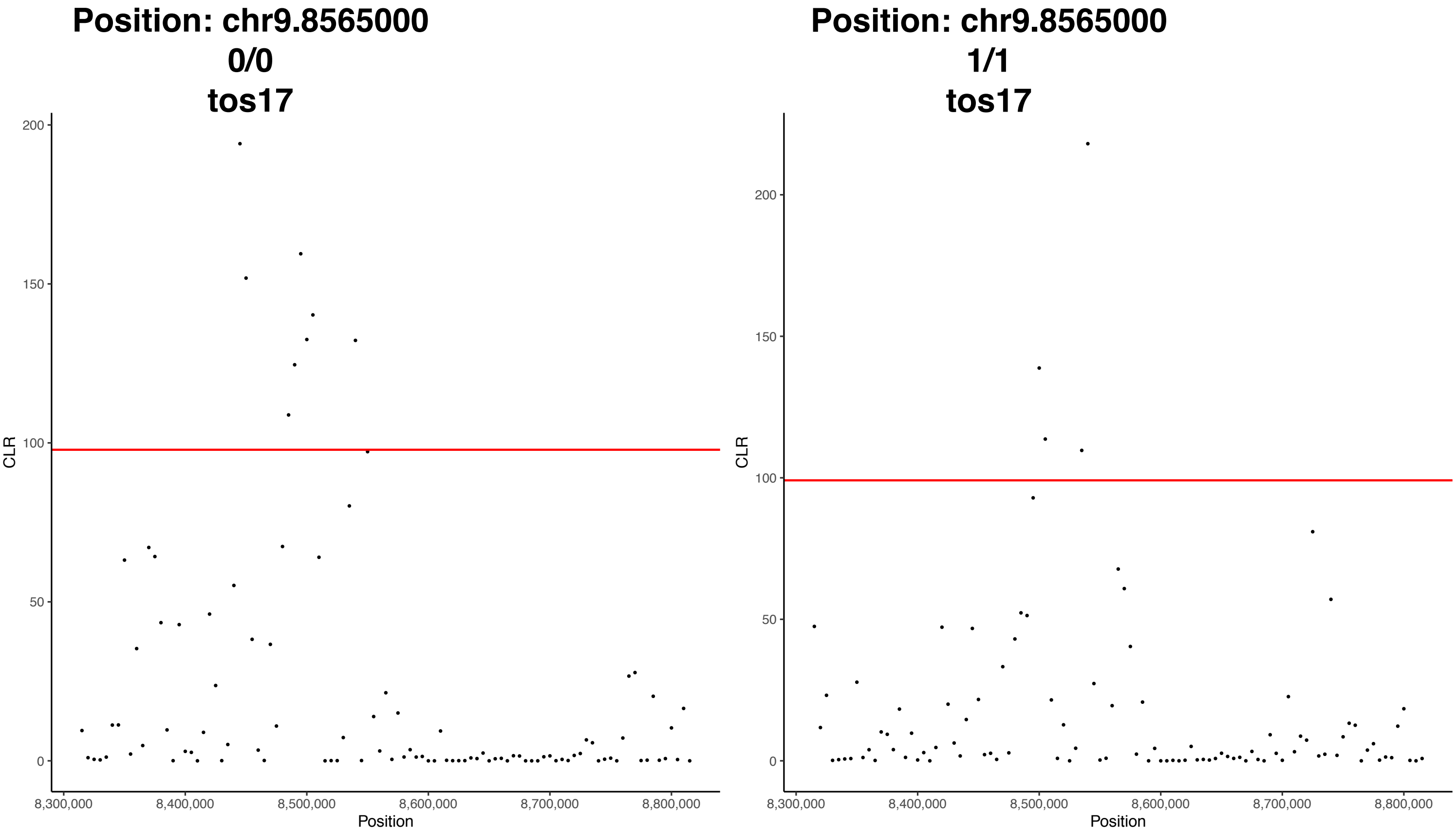


Figure 3.15: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Tos17* TE.

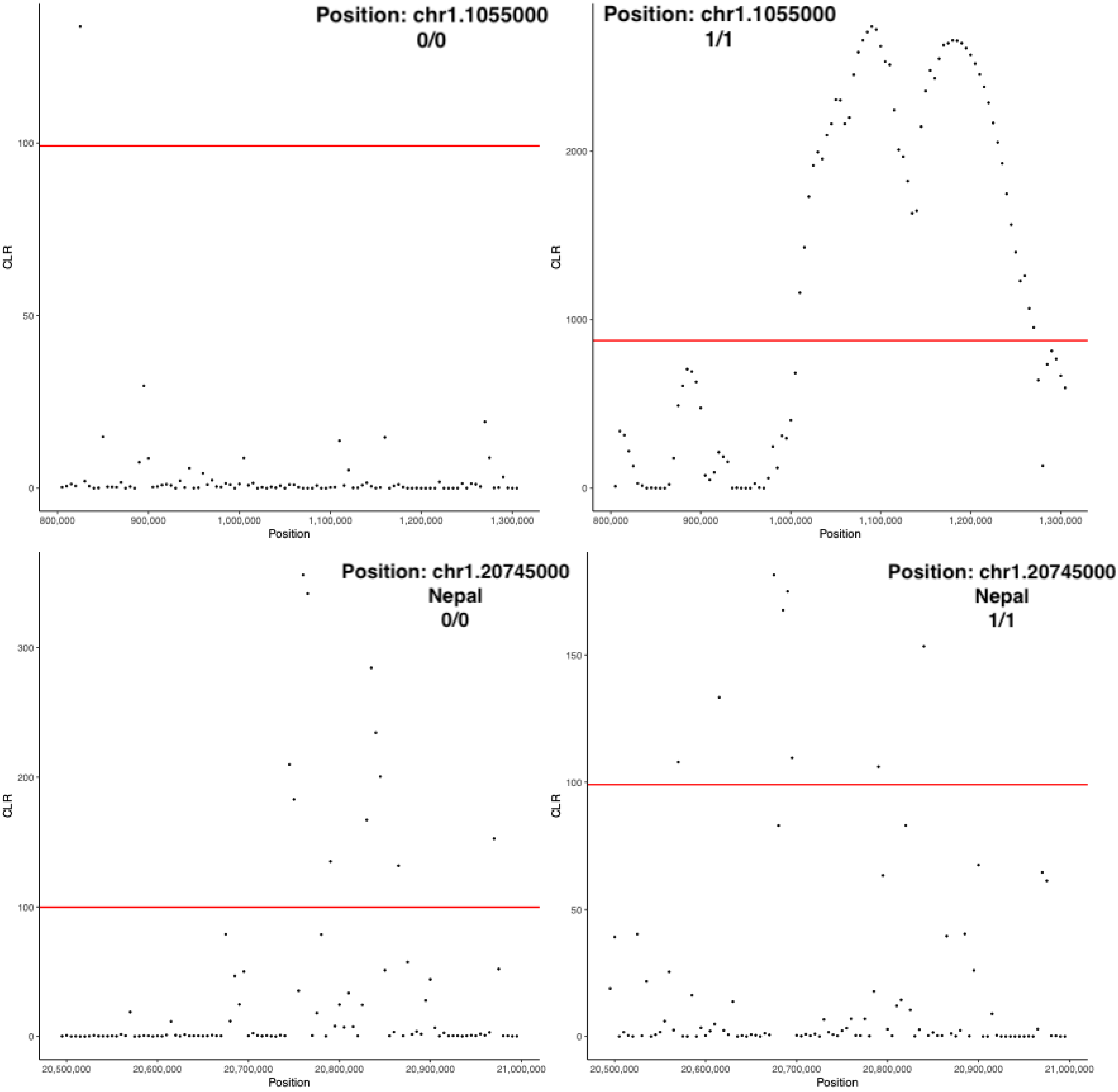


Figure 3.16: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Fam106* TE.

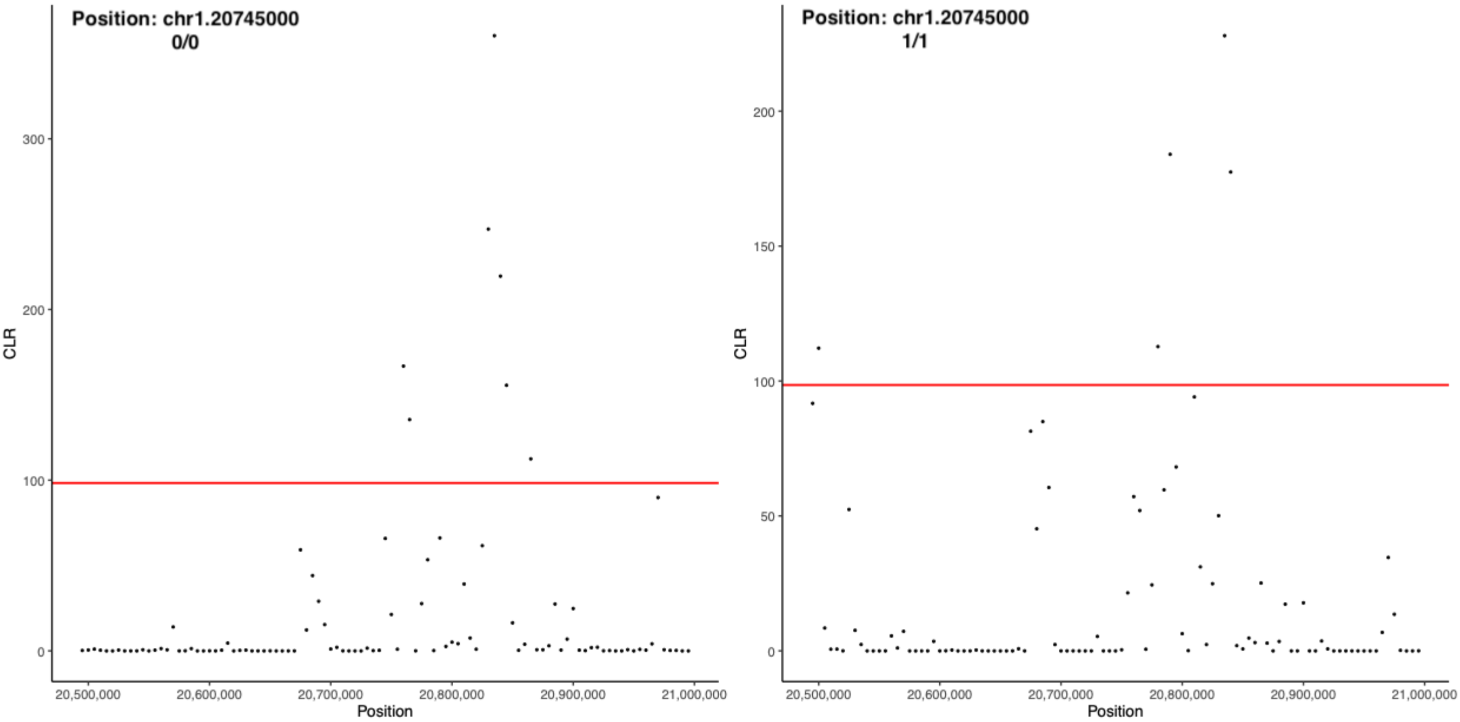


Figure 3.17: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Fam106* TE.

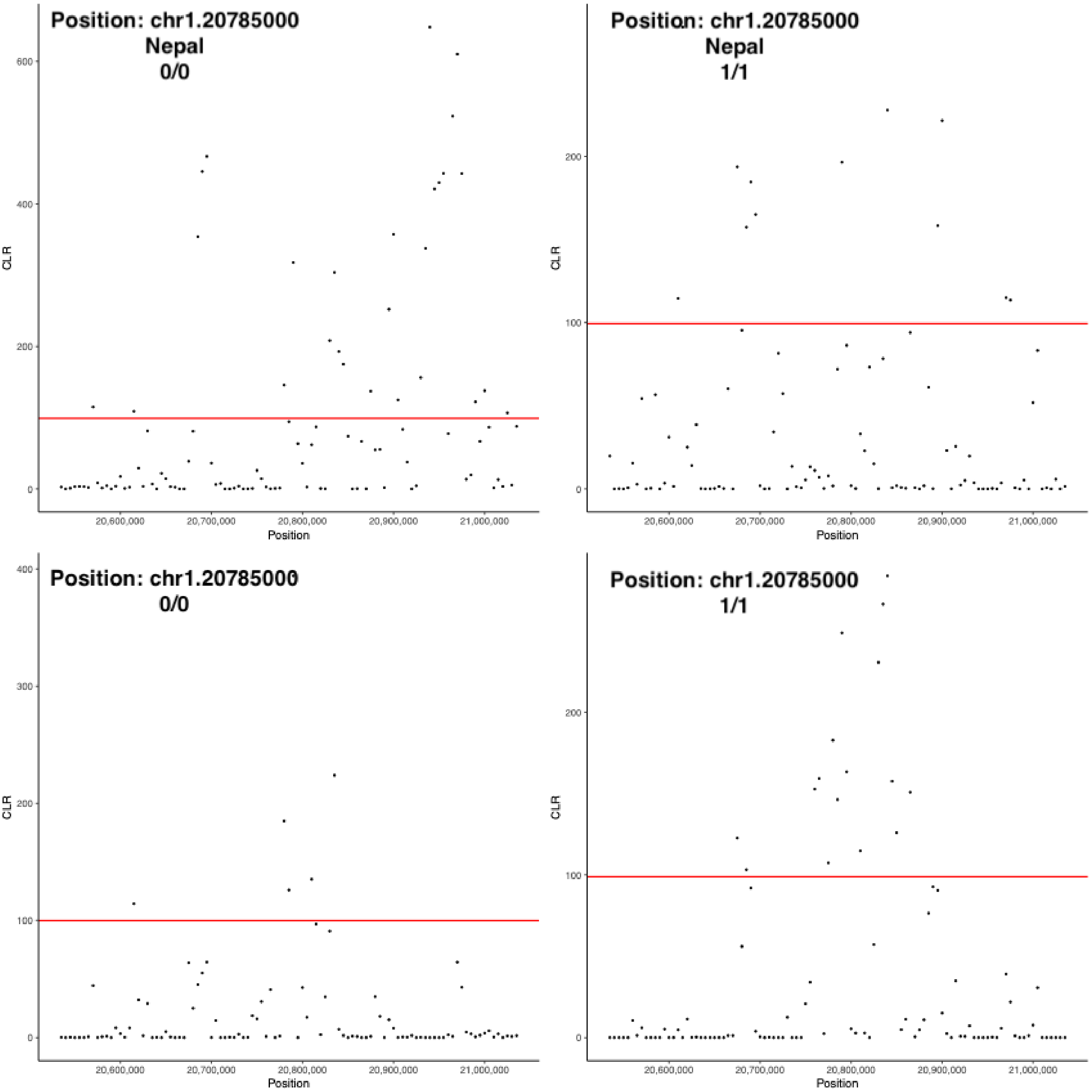


Figure 3.18: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Fam106* TE.

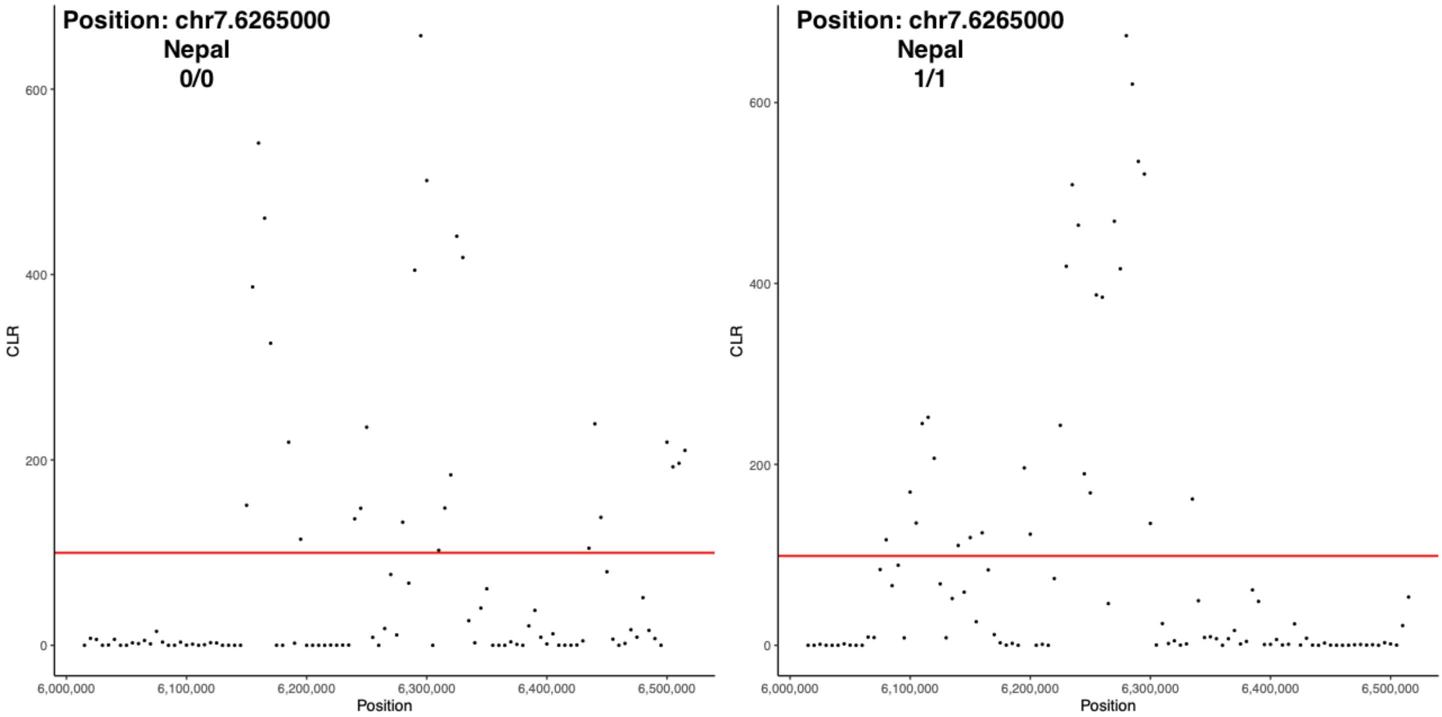


Figure 3.19: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Fam106* TE.

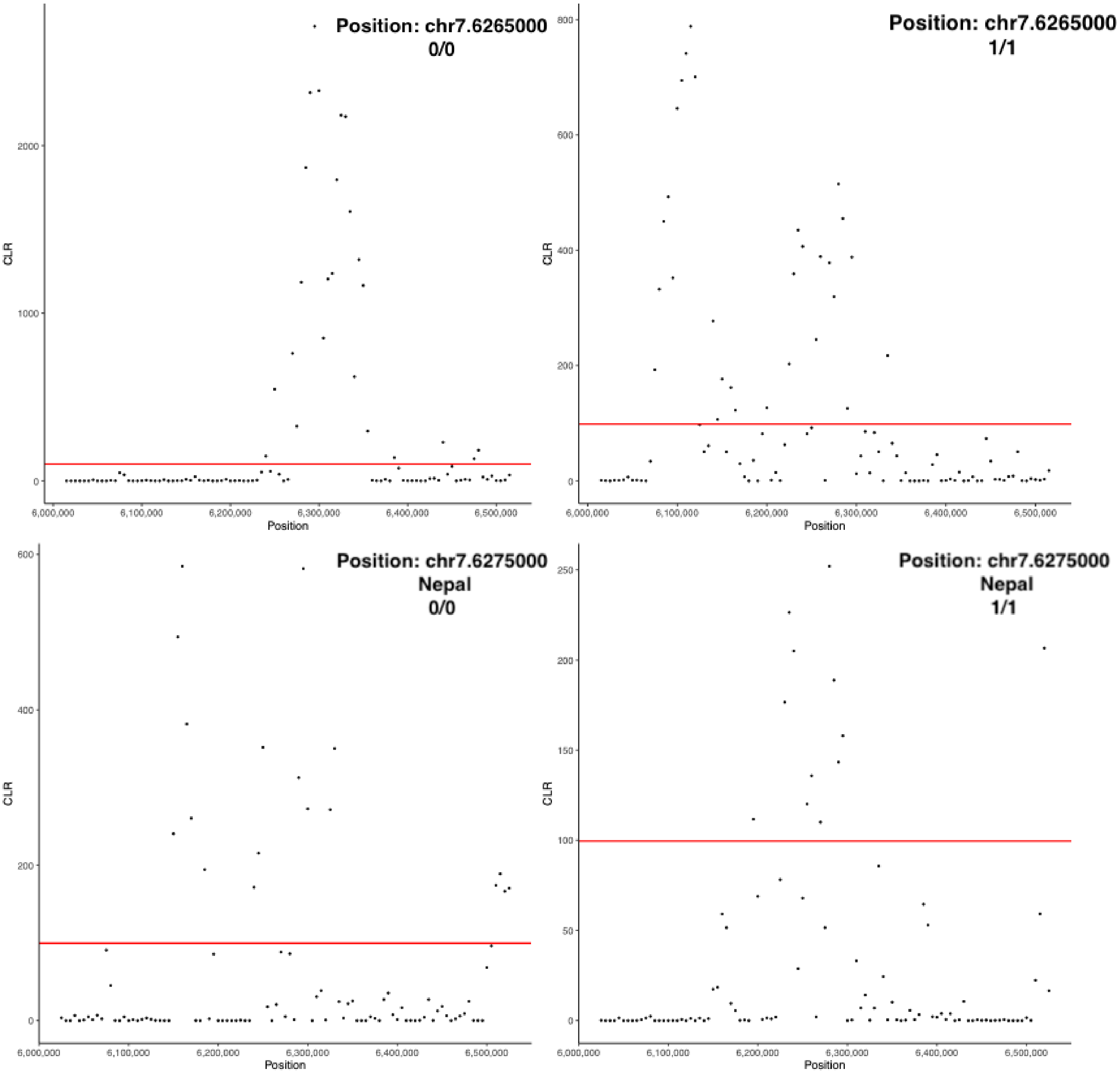


Figure 3.20: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Fam106* TE.

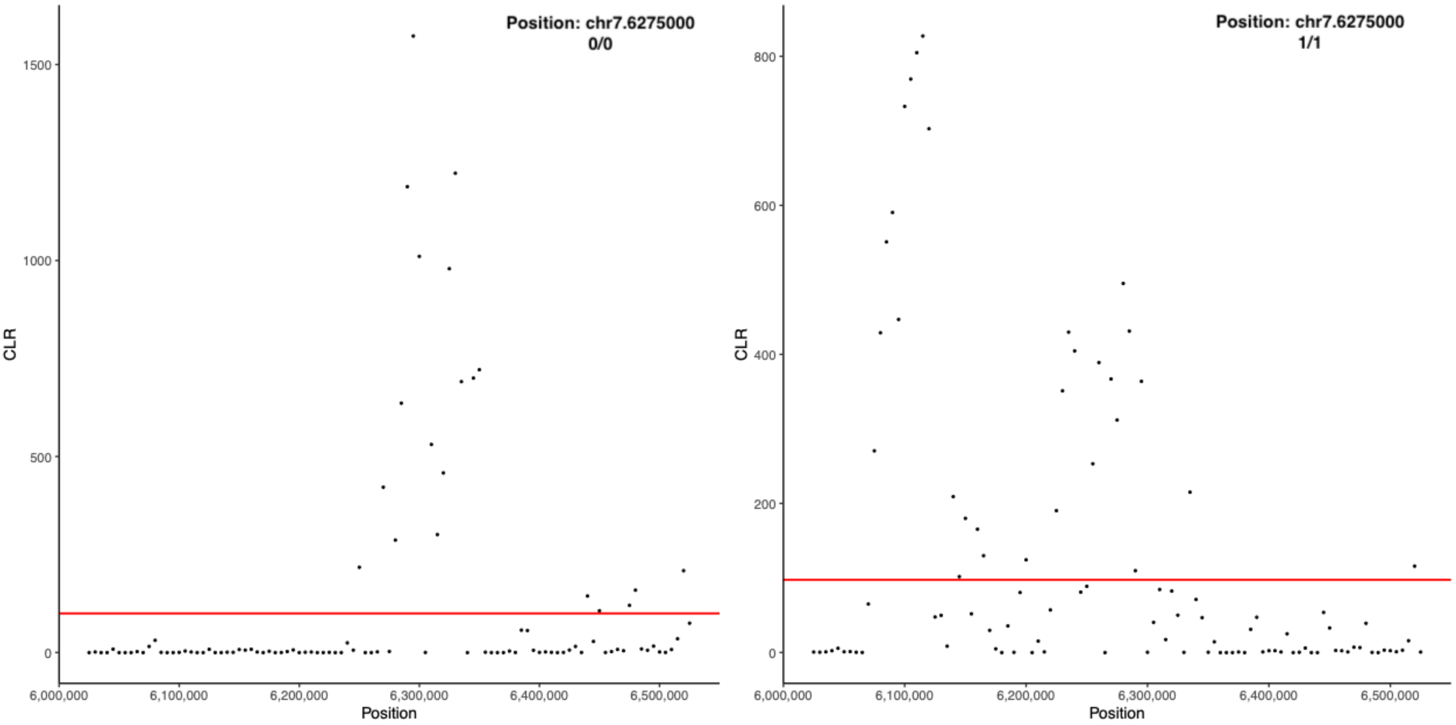
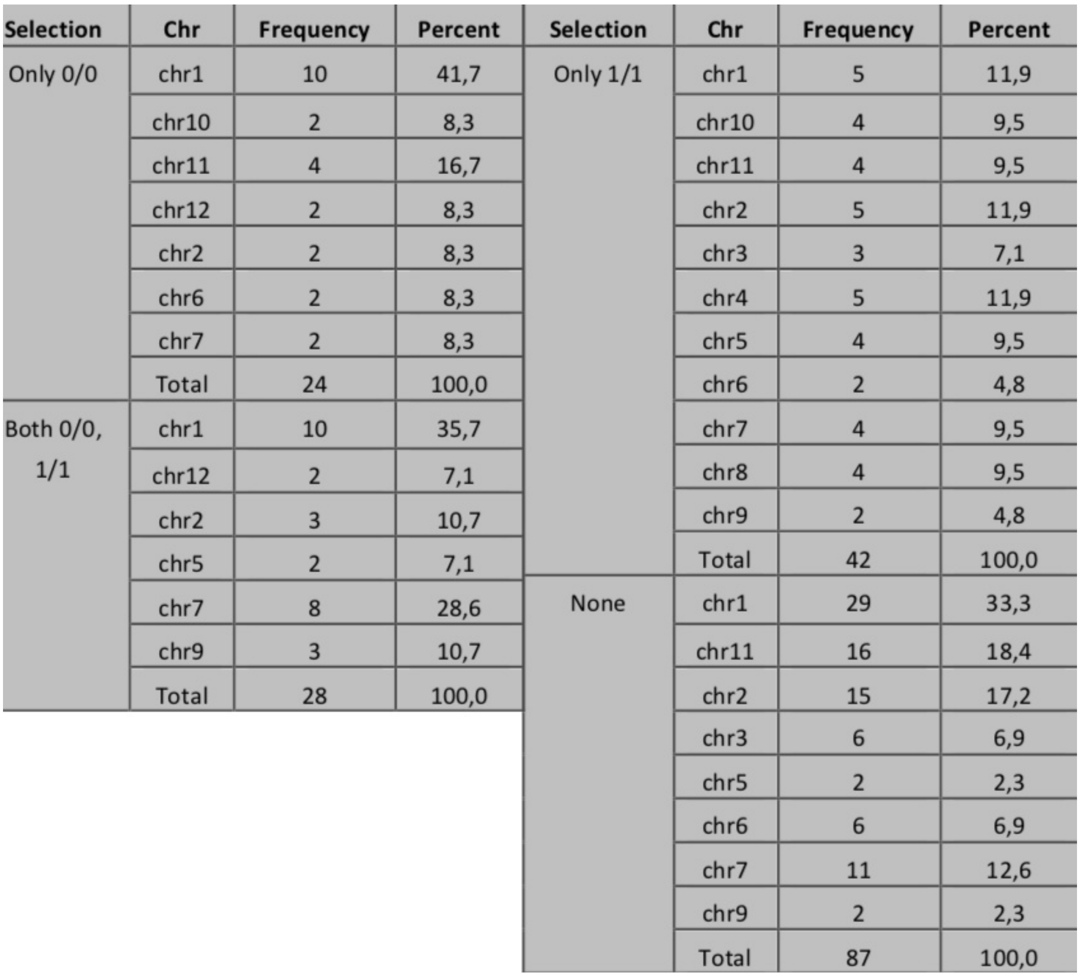


Figure 3.21: Selective sweep analysis of haplotypes significant for both absence (0/0)/presence (1/1) of *Fam106* TE.



Selection	Chr	Frequency	Percent
Only 0/0	chr1	10	41,7
	chr10	2	8,3
	chr11	4	16,7
	chr12	2	8,3
	chr2	2	8,3
	chr6	2	8,3
	chr7	2	8,3
	Total	24	100,0
Both 0/0, 1/1	chr1	10	35,7
	chr12	2	7,1
	chr2	3	10,7
	chr5	2	7,1
	chr7	8	28,6
	chr9	3	10,7
	Total	28	100,0

Selection	Chr	Frequency	Percent
Only 1/1	chr1	5	11,9
	chr10	4	9,5
	chr11	4	9,5
	chr2	5	11,9
	chr3	3	7,1
	chr4	5	11,9
	chr5	4	9,5
	chr6	2	4,8
	chr7	4	9,5
	chr8	4	9,5
	chr9	2	4,8
	Total	42	100,0
None	chr1	29	33,3
	chr11	16	18,4
	chr2	15	17,2
	chr3	6	6,9
	chr5	2	2,3
	chr6	6	6,9
	chr7	11	12,6
	chr9	2	2,3
	Total	87	100,0

Table 3.4: Frequencies of significant scores of haplotypes for absence (0/0), presence (1/1), both absence (0/0)/presence (1/1) and neither of TE variant per chromosome.

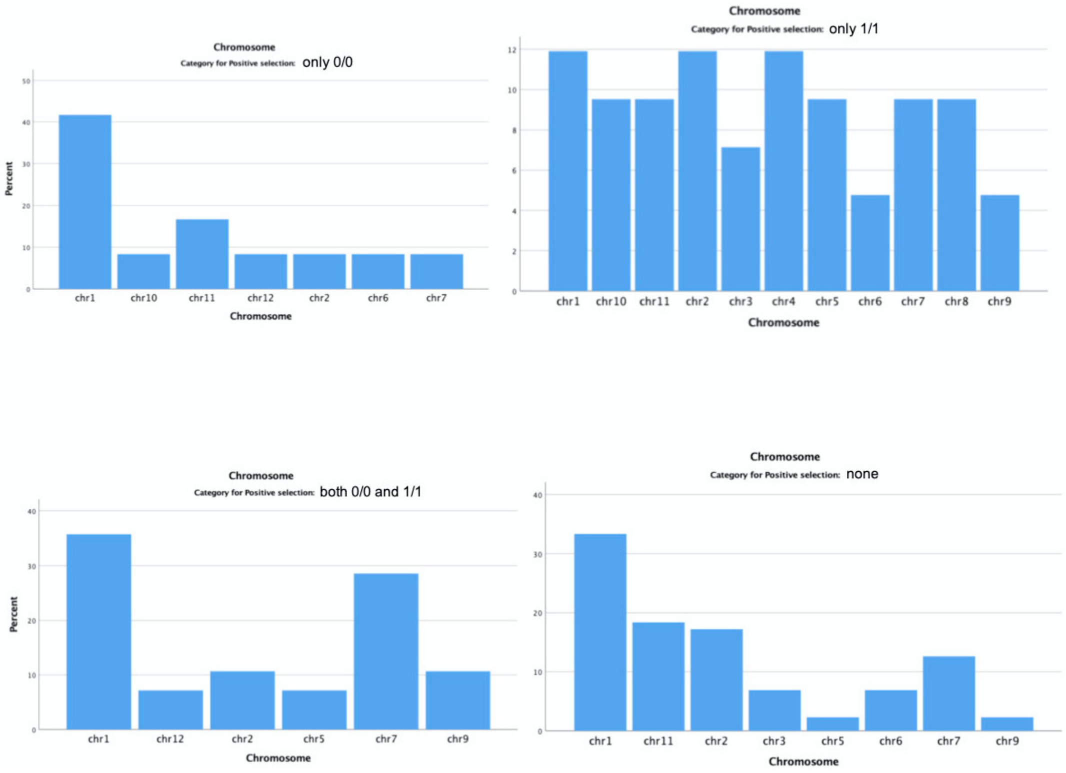


Figure 3.22: Distribution of significant scores calculated for haplotypes for absence (0/0), presence (1/1), both absence (0/0) and presence (1/1) and neither of TE variant per chromosome. Positive selection is unequally distributed on different chromosomes except for haplotypes with positive selection only for the presence (1/1) of TE variant (top-right).

Discussion

Here we present the methodology we followed to explain the putative adaptive role of TEs in rice genome by investigating the genomic patterns (SNP patterns) of their neighborhoods in the genome in rice. For this, we applied specialized computational tools recently developed for discovery and annotation of novel TE insertions in a rice population, in combination with data published from another similar, current analysis published this year (Carpentier et al., 2019) to obtain as much information as possible of the polymorphic picture of active TEs in *O.sativa*. In order to gain a more comprehensive understanding of the role of TEs in adaptation we localized the action of positive selection by detecting selective sweeps. Thus, we test the following hypothesis: if a TE insertion might be beneficial, even if it has not reached fixation yet, it could show hitchhiking effect due to natural selection at least in the subset of sequences carrying it.

Our first observation is that rice genome is flourished of *mPing* mobile element, but the polymorphic insertions represent a small fraction of this (only 7.14%). *mPing* element was also detected inside/or very close to, genes, a finding that may be explained by the very small size of this TE (only 433bp). Structurally, MITEs have very small size (<600 base pairs) and might cause lower genomic disruption compared to other mobile elements (Jiang et al., 2003). The frequency of *mPing* found inside genic regions is small and did not pass our filter for further selective sweep analysis. This findings agree with earlier studies claiming that *mPing* elements avoid inserting into exons (Naito et al., 2014; Naito et al., 2009; Jiang et al., 2003; Lu et al., 2017). Regarding the insertion location biases, we document a chromosomal insertion preference in chromosomes 1, 2, 7 and 11. We hypothesise that these chromosomal insertion biases may have more to do with the chromosomal environment surrounding that area, than whole chromosomes. It is known that transposons often have specific targeting mechanisms that exploit ‘safe havens’ in the genome, such as noncoding or transcriptionally repressed regions (Martin and Garfinkel, 2003).

Our population survey of the frequency of selective sweeps in 182 haplotypes from 130 individuals showed sweep signatures in the flanking regions of all four of the putatively adaptive TEs analyzed. Almost half of the total haplotypes scanned showed significant likelihood scores for positive selection and 23% of them where homozygous for the TE presence, a number almost double from the haplotypes found significant for the absence of them. Our results indicate that individuals with TE insertions in their genome show significant stronger positive selection around that area than the respective ones not carrying the TE, but those loci were not inside other genes. However, the number of identified adaptive TEs is still too small to draw any general conclusions about the TE-induced adaptive process and is definitely not conclusive of adaptive evolution role of TEs. Although identification of a selective sweep provides considerable evidence for positive selection, they are not entirely conclusive of adaptive evolution for several reasons. First, there is still some uncertainty about the exact demographic model for *O.sativa*. There are still some contradictions on the domesticated scenarios of Asian rice which leaded to the picture we have today (Izawa, 2008; Civáň and Brown, 2018; Stein et al., 2018). Analyzing patterns of polymorphism without taking into account the demographic history of the populations can lead to spurious inference of positive selection. Here, we used control regions, i.e., regions where no TE insertion has taken place to generate a distribution of CLR scores representing the null hypothesis. Second, it is possible that a mutation located near the scanned region is associated with the sweep and not the TE itself. Third, the value of detecting genetic variation is small, if it is not accompanied by the study of its possible effects on associated biological pathways. Investigation of the epigenetic landscape that reacts with the specific TEs would give important information on the evolutionary arms race that happens between mobile elements and their hosts. Selection forces could also shape the piRNA, tasiRNA or natsiRNA clusters on the genome that get activated when they try to silence these elements. In conclusion, the selective pattern found around the insertions that we analyzed is consistent with our hypothesis. However, further selective sweep analysis needs to be done to get more clear picture of the evolutionary forces that act nearby TEs. For that, we aim to run similar analyses in larger sample for the same or other putative adaptive TEs and also repeat the analysis on another isolated population to compare it with our results obtained from Nepal.

Appendix

Origin of data used

DNA unique ID	Country of origin	Variety
B002	China	Temperate japonica
B024	Thailand	Indica
B040	Uganda	Indica
B049	Nepal	Aus
B084	China	Intermediate type
B126	China	Indica
B142	China	Intermediate type
B146	China	Indica
B162	China	Temperate japonica
B202	China	Indica
B268	China	Indica
CX100	Nepal	Indica
CX113	nan	Tropical japonica
CX128	Nepal	Indica
CX207	China	Indica
CX225	Philippines	Indica
CX230	China	Indica
CX303	China	Indica
CX313	China	Indica
CX76	Sri Lanka	Indica
CX90	Nepal	Indica
IRIS 313-8268	Nepal	Intermediate type
IRIS 313-9053	India	Intermediate type
IRIS 313-8744	Indonesia	Indica
IRIS 313-8956	Indonesia	Indica
IRIS 313-8957	India	Indica
IRIS 313-8985	Thailand	Indica
IRIS 313-8996	Vietnam	Indica
IRIS 313-9112	Thailand	Indica
IRIS 313-9281	Thailand	Indica
IRIS 313-9924	South Korea	Indica
IRIS 313-10007	Nepal	Indica
IRIS 313-10016	Iran	Basmati/sadri
IRIS 313-10374	Philippines	Indica
IRIS 313-9342	Vietnam	Indica
IRIS 313-8293	Senegal	Indica
IRIS 313-9188	Indonesia	Indica
IRIS 313-9406	Thailand	Indica
IRIS 313-9995	South Korea	Temperate japonica
IRIS 313-10073	Japan	Japonica
IRIS 313-10075	Japan	Indica
IRIS 313-10077	Japan	Japonica
IRIS 313-8694	Brazil	Tropical japonica
IRIS 313-8205	Italy	Temperate japonica

IRIS 313-8076	Australia	Temperate japonica
IRIS 313-8123	Portugal	Temperate japonica
IRIS 313-8143	Bulgaria	Indica
IRIS 313-8155	Russia	Tropical japonica
IRIS 313-7664	Colombia	Tropical japonica
IRIS 313-8011	Vietnam	Tropical japonica
IRIS 313-7933	Nepal	Tropical japonica
IRIS 313-11433	India	Tropical japonica
IRIS 313-11478	India	Temperate japonica
IRIS 313-8147	India	Indica
IRIS 313-11428	Brazil	Temperate japonica
IRIS 313-11524	Ivory Coast	Indica
IRIS 313-11525	Guinea-Bissau	Tropical japonica
IRIS 313-11516	Philippines	Indica
IRIS 313-11561	Nepal	Indica
IRIS 313-11563	Nepal	Basmati/sadri
IRIS 313-11564	Nepal	Indica
IRIS 313-11565	Nepal	Indica
IRIS 313-11566	Nepal	Basmati/sadri
IRIS 313-11567	Nepal	Indica
IRIS 313-11568	Nepal	Indica
IRIS 313-11585	China	Temperate japonica
IRIS 313-11624	Nepal	Indica
IRIS 313-11625	Nepal	Basmati/sadri
IRIS 313-11626	Nepal	Basmati/sadri
IRIS 313-11627	Nepal	Intermediate type
IRIS 313-11628	Nepal	Aus/boro
IRIS 313-11629	Nepal	Basmati/sadri
IRIS 313-11630	Nepal	Basmati/sadri
IRIS 313-11632	Nepal	Indica
IRIS 313-11671	Nepal	Temperate japonica
IRIS 313-11672	Nepal	Tropical japonica
IRIS 313-11691	Bhutan	Indica
IRIS 313-11704	Thailand	Indica
IRIS 313-11706	Thailand	Temperate japonica
IRIS 313-11868	China	Indica
IRIS 313-11939	Burkina Fasso	Indica
IRIS 313-11943	Nepal	Indica
IRIS 313-11944	Nepal	Indica
IRIS 313-11956	Nepal	Intermediate type
IRIS 313-11959	Philippines	Indica
IRIS 313-11999	Cambodia	Indica
IRIS 313-12083	Madagascar	Indica
IRIS 313-12093	Nepal	Indica
IRIS 313-12076	Laos	Tropical japonica

IRIS 313-12139	Nepal	Aus/boro
IRIS 313-12180	Nepal	Indica
IRIS 313-12182	Nepal	Indica
IRIS 313-12183	Nepal	Aus/boro
IRIS 313-12190	Laos	Indica
IRIS 313-12207	Laos	Indica
IRIS 313-12261	Laos	Indica
IRIS 313-12352	Laos	Tropical japonica
IRIS 313-10429	Taiwan	Temperate japonica
IRIS 313-10518	Myanmar	Indica
IRIS 313-10623	Nepal	Aus/boro
IRIS 313-10702	Malaysia	Indica
IRIS 313-10706	Malaysia	Indica
IRIS 313-10727	Senegal	Indica
IRIS 313-10731	Nepal	Indica
IRIS 313-10732	Nepal	Basmati/sadri
IRIS 313-10733	Nepal	Indica
IRIS 313-10734	Nepal	Aus/boro
IRIS 313-10735	Nepal	Aus/boro
IRIS 313-10736	Nepal	Aus/boro
IRIS 313-10737	Nepal	Aus/boro
IRIS 313-10768	Indonesia	Indica
IRIS 313-10859	India	Indica
IRIS 313-10874	India	Tropical japonica
IRIS 313-10918	Philippines	Japonica
IRIS 313-10924	Nepal	Indica
IRIS 313-10925	Nepal	Aus/boro
IRIS 313-10926	Nepal	Basmati/sadri
IRIS 313-10927	Nepal	Aus/boro
IRIS 313-10946	Indonesia	Tropical japonica
IRIS 313-10985	Bangladesh	Indica
IRIS 313-11004	Indonesia	Tropical japonica
IRIS 313-10990	Philippines	Indica
IRIS 313-11077	Laos	Tropical japonica
IRIS 313-11122	Philippines	Indica
IRIS 313-11086	Laos	Indica
IRIS 313-11088	Cambodia	Indica
IRIS 313-11210	Bangladesh	Aus/boro
IRIS 313-11232	India	Aus/boro
IRIS 313-11330	Philippines	Indica
IRIS 313-11376	Ivory Coast	Tropical japonica

The table above includes the accession names of the 130 samples used in this analysis, the country that was originated and the variety

Commands and Pipelines

Picard GATK

For FASTA file reformating we used the JAVA-written tool below to in order all lines of sequence to be of the same length.

All reference FASTA sequences for MELT tool applied with the command below :

```
$ java -jar /path/to/picard.jar NormalizeFasta
I=input__reference_TE_Sequence.fa
O=normalized_TE_sequence.fasta
```

Samtools

Indexing for reference FASTA sequences for MELT tool applied using Samtools with the command below :

```
$ samtools faidx reference_sequence.fa
```

Data download

A python script was written and used for the creation of the full links directing to the BAM files to be downloaded from an initial MANIFEST file. The script is documented in the following GitHub repository (`https://github.com/Joannagare`). The BAM files were downliaded using the output of the above script in the command below:

```
wget --no-check-certificate -t 100 -i file
```

TransposonZip file

MELT requirement for transposon.zip file to direct transposable element discovery created using MELT-BuildTransposonZIP runtime with the following command:

```
$ java -Xmx1G -jar MELT.jar BuildTransposonZIP
TransposonSequence.fa| Transposon.bed
NAME[mPingElement] ERROR[3]
```

Preprocessing BAM Files for MELT

In order to to speed up MELT's runtime BAM files where preprocessed using the command below:

```
$ java -Xmx2G -jar MELT.jar Preprocess -bamfile sorted.bam -h IRGSP-1.0_g
```

Running MELT-SPLIT

The pipeline for the Transposable Element discovery follows 4 general steps (IndivAnalysis – TE insertion discovery in individual samples, GroupAnalysis – Merge discovery information across all genomes in project, Genotype – Genotyping all samples using merged TE insertion discovery information and MakeVCF – Performing final filtering and merging of individual samples into final VCF). The usage we followed in this study is described below:

```
$ java –Xmx6G −jar MELT.jar IndivAnalysis
–w mPingElement −c 14
–h IRGSP−1.0_genome.fasta
−t mPingElement_MELT.zip

$ java –Xmx6G −jar MELT.jar GroupAnalysis
–w /path/meltsplit  −discoverydir /path/mPingElement
–h IRGSP−1.0_genome.fasta
−t mPingElement_MELT.zip –n bedfile

$ java –Xmx2G −jar MELT.jar Genotype
–w /path/meltsplit –p /path/meltsplit
–h IRGSP−1.0_genome.fasta
−t mPingElement_MELT.zip

$ java –Xmx2G −jar MELT.jar MakeVCF
−genotypingdir /path/meltsplit –w /path/meltsplit
–h IRGSP−1.0_genome.fasta
−t mPingElement_MELT.zip
–p /path/meltsplit –o ./
```

Running SweeD

For the SweeD analysis, on control chromosomal positions the following commands were used:

Creation of osf files, for whole chromosomes:

```
$ sweed/SweeD–P
–name chr.position.at.ctrl.chr.position.run
–input ctrl.chr.postition.sf
−gridFile ctrl.chr.position −grid 1312 −threads 2
```

SweeD runs for control chromosomal positions:

```
$ sweed/SweeD–P –name ctrl.chr.position.sfrun
−threads 2 −osf ctrl.chr.position.sf
–input 3kSNP_chr.vcf −sampleList chr.position.00/11.out
```

For specific genetic regions: GridFiles, which contain information about chromosome and position, were created by GridFileCreator.py script.

```
$ sweed/SweeD–P −threads 6 –name chr.position.run
–input 3kSNP_chr.vcf −gridFile points.chr.start.end.out
–grid 1312420 −sampleList name.chr.position.00/11.out
```

bedtools closest -a teins.bed -b idsorted -d ¿ mpinfclosestalltogenes.txt

Distance between TE insertion and closest gene estimation

```
chr01    37542851 37543284 chr01    37542029 37542352 gene_id "Os01g0866950"; transcript_id "Os01t0866950-00";    500
chr06    4420878  4421311  chr06    4423308  4423523  gene_id "Os06g0187600"; transcript_id "Os06t0187600-00";    1998
chr12    22071089 22071522 chr12    22064278 22065027 gene_id "Os12g0546400"; transcript_id "Os12t0546400-00";    6063
chr02    28273934 28274367 chr02    28266547 28266855 gene_id "Os02g0689133"; transcript_id "Os02t0689133-00";    7080
chr09    14934734 14935167 chr09    14945845 14946597 gene_id "Os09g0417000"; transcript_id "Os09t0417000-00";    10679
chr04    35048161 35048594 chr04    35066229 35066636 gene_id "Os04g0686150"; transcript_id "Os04t0686150-00";    17636
chr01    35200680 35201113 chr01    35179642 35180055 gene_id "Os01g0823800"; transcript_id "Os01t0823800-00";    20626
chr05    27897960 27898393 chr05    27921350 27922411 gene_id "Os05g0561100"; transcript_id "Os05t0561100-00";    22958
chr04    32604530 32604963 chr04    32643502 32643554 gene_id "Os04g0641500"; transcript_id "Os04t0641500-00";    38540
chr05    3221901  3222334  chr05    3180284  3180922  gene_id "Os05g0154432"; transcript_id "Os05t0154432-00";    40980
```

Table 1: Sorted distance from the closest genes found from the filtered polymorphic *mPing* element insertions.

chr02	29104580	29105013	chr02	29102407	29108066	gene_id "Os02g0705201";	transcript_id "Os02t0705201-00";	0
chr03	9645477	9645910	chr03	9645540	9646541	gene_id "Os03g0281700";	transcript_id "Os03t0281700-00";	0
chr04	34808917	34809350	chr04	34807961	34809028	gene_id "Os04g0681850";	transcript_id "Os04t0681850-00";	0
chr04	34808917	34809350	chr04	34808987	34809028	gene_id "Os04g0681850";	transcript_id "Os04t0681850-00";	0
chr10	11206151	11206584	chr10	11203386	11206622	gene_id "Os10g0362400";	transcript_id "Os10t0362400-00";	0
chr10	11206151	11206584	chr10	11206581	11206622	gene_id "Os10g0362400";	transcript_id "Os10t0362400-00";	0
chr03	3872638	3873071	chr03	3872169	3872576	gene_id "Os03g0172300";	transcript_id "Os03t0172300-00";	63
chr04	21942344	21942777	chr04	21942984	21943259	gene_id "Os04g0440700";	transcript_id "Os04t0440700-00";	208
chr04	21942344	21942777	chr04	21942984	21946604	gene_id "Os04g0440700";	transcript_id "Os04t0440700-00";	208
chr06	131071	131504	chr06	130355	130738	gene_id "Os06g0100850";	transcript_id "Os06t0100850-00";	334
chr08	21754342	21754775	chr08	21755178	21755432	gene_id "Os08g0446051";	transcript_id "Os08t0446051-00";	404
chr01	37542851	37543284	chr01	37542029	37542352	gene_id "Os01g0866950";	transcript_id "Os01t0866950-00";	500
chr05	27956036	27956469	chr05	27957109	27957354	gene_id "Os05g0561950";	transcript_id "Os05t0561950-00";	641
chr11	27945567	27946000	chr11	27946834	27946861	gene_id "Os11g0688666";	transcript_id "Os11t0688666-00";	835
chr11	27945567	27946000	chr11	27946834	27947159	gene_id "Os11g0688666";	transcript_id "Os11t0688666-00";	835
chr01	38092041	38092474	chr01	38093559	38093794	gene_id "Os01g0877700";	transcript_id "Os01t0877700-00";	1086
chr01	38092041	38092474	chr01	38093559	38094202	gene_id "Os01g0877700";	transcript_id "Os01t0877700-00";	1086
chr01	11055159	11055592	chr01	11053657	11053902	gene_id "Os01g0300850";	transcript_id "Os01t0300850-00";	1258
chr03	34077518	34077951	chr03	34074396	34075694	gene_id "Os03g0813300";	transcript_id "Os03t0813300-00";	1825
chr03	34077518	34077951	chr03	34075532	34075694	gene_id "Os03g0813300";	transcript_id "Os03t0813300-00";	1825
chr06	4420878	4421311	chr06	4423308	4423523	gene_id "Os06g0187600";	transcript_id "Os06t0187600-00";	1998
chr04	17247989	17248422	chr04	17250604	17250686	gene_id "Os04g0360401";	transcript_id "Os04t0360401-00";	2183
chr04	17247989	17248422	chr04	17250604	17252023	gene_id "Os04g0360401";	transcript_id "Os04t0360401-00";	2183
chr11	7110256	7110689	chr11	7113206	7114000	gene_id "Os11g0233900";	transcript_id "Os11t0233900-00";	2518
chr04	2289084	2289517	chr04	2292073	2292933	gene_id "Os04g0135100";	transcript_id "Os04t0135100-00";	2557
chr02	14599323	14599756	chr02	14602725	14602774	gene_id "Os02g0449101";	transcript_id "Os02t0449101-00";	2970
chr02	14599323	14599756	chr02	14602725	14603195	gene_id "Os02g0449101";	transcript_id "Os02t0449101-00";	2970
chr12	21428926	21429359	chr12	21432423	21432641	gene_id "Os12g0538200";	transcript_id "Os12t0538200-00";	3065
chr02	19186320	19186753	chr02	19182475	19182945	gene_id "Os02g0524950";	transcript_id "Os02t0524950-00";	3376
chr02	19186320	19186753	chr02	19182789	19182945	gene_id "Os02g0524950";	transcript_id "Os02t0524950-00";	3376
chr05	17150495	17150928	chr05	17154443	17154871	gene_id "Os05g0360100";	transcript_id "Os05t0360100-00";	3516
chr05	22027352	22027785	chr05	22023035	22023739	gene_id "Os05g0448650";	transcript_id "Os05t0448650-00";	3614
chr09	20454264	20454697	chr09	20449962	20450243	gene_id "Os09g0523450";	transcript_id "Os09t0523450-00";	4022
chr06	23538621	23539054	chr06	23534104	23534598	gene_id "Os06g0597301";	transcript_id "Os06t0597301-00";	4024
chr06	7963848	7964281	chr06	7968584	7968829	gene_id "Os06g0253675";	transcript_id "Os06t0253675-00";	4304
chr01	34897492	34897925	chr01	34902303	34902431	gene_id "Os01g0819150";	transcript_id "Os01t0819150-00";	4379
chr01	34897492	34897925	chr01	34902303	34903598	gene_id "Os01g0819150";	transcript_id "Os01t0819150-00";	4379
chr03	9252789	9253222	chr03	9257792	9258217	gene_id "Os03g0274350";	transcript_id "Os03t0274350-00";	4571
chr02	12891464	12891897	chr02	12885926	12886309	gene_id "Os02g0322102";	transcript_id "Os02t0322102-00";	5156
chr04	31426322	31426755	chr04	31432434	31432952	gene_id "Os04g0618900";	transcript_id "Os04t0618900-00";	5680
chr09	12798353	12798786	chr09	12791129	12792529	gene_id "Os09g0379750";	transcript_id "Os09t0379750-00";	5825
chr09	12798353	12798786	chr09	12792212	12792529	gene_id "Os09g0379750";	transcript_id "Os09t0379750-00";	5825
chr12	22071089	22071522	chr12	22064278	22065027	gene_id "Os12g0546400";	transcript_id "Os12t0546400-00";	6063
chr12	22071089	22071522	chr12	22064950	22065027	gene_id "Os12g0546400";	transcript_id "Os12t0546400-00";	6063
chr02	19498258	19498691	chr02	19491686	19491976	gene_id "Os02g0530401";	transcript_id "Os02t0530401-00";	6283
chr03	4637247	4637680	chr03	4643969	4644286	gene_id "Os03g0188701";	transcript_id "Os03t0188701-00";	6290
chr03	1174087	1174520	chr03	1181014	1181046	gene_id "Os03g0121450";	transcript_id "Os03t0121450-00";	6495
chr03	1174087	1174520	chr03	1181014	1182699	gene_id "Os03g0121450";	transcript_id "Os03t0121450-00";	6495
chr04	19818151	19818584	chr04	19810813	19811199	gene_id "Os04g0400300";	transcript_id "Os04t0400300-00";	6953
chr07	21221382	21221815	chr07	21214013	21214351	gene_id "Os07g0538966";	transcript_id "Os07t0538966-00";	7032
chr02	28273934	28274367	chr02	28266547	28266855	gene_id "Os02g0689133";	transcript_id "Os02t0689133-00";	7080
chr11	20284863	20285296	chr11	20276950	20277402	gene_id "Os11g0549400";	transcript_id "Os11t0549400-00";	7462
chr11	801353	801786	chr11	793190	793744	gene_id "Os11g0117550";	transcript_id "Os11t0117550-00";	7610
chr05	26675277	26675710	chr05	26667198	26667491	gene_id "Os05g0537001";	transcript_id "Os05t0537001-00";	7787
chr01	10619873	10620306	chr01	10628391	10629401	gene_id "Os01g0292300";	transcript_id "Os01t0292300-00";	8086
chr04	31441253	31441686	chr04	31432434	31432952	gene_id "Os04g0618900";	transcript_id "Os04t0618900-00";	8302
chr06	8253870	8254303	chr06	8245149	8245442	gene_id "Os06g0257850";	transcript_id "Os06t0257850-00";	8429
chr06	4972278	4972711	chr06	4963254	4963643	gene_id "Os06g0197700";	transcript_id "Os06t0197700-00";	8636
chr01	2414152	2414585	chr01	2423261	2423512	gene_id "Os01g0145101";	transcript_id "Os01t0145101-00";	8677
chr09	14585292	14585725	chr09	14594643	14595030	gene_id "Os09g0411050";	transcript_id "Os09t0411050-00";	8919
chr09	14585292	14585725	chr09	14594643	14595768	gene_id "Os09g0411050";	transcript_id "Os09t0411050-00";	8919
chr05	28160441	28160874	chr05	28171399	28171644	gene_id "Os05g0565966";	transcript_id "Os05t0565966-00";	10526
chr09	14934734	14935167	chr09	14945845	14946597	gene_id "Os09g0417000";	transcript_id "Os09t0417000-00";	10679
chr03	13937187	13937620	chr03	13925931	13926191	gene_id "Os03g0358950";	transcript_id "Os03t0358950-00";	10997
chr11	4876435	4876868	chr11	4864538	4864765	gene_id "Os11g0197301";	transcript_id "Os11t0197301-00";	11671
chr06	26135962	26136395	chr06	26123167	26124150	gene_id "Os06g0642000";	transcript_id "Os06t0642000-00";	11813
chr01	38606686	38607119	chr01	38594450	38594743	gene_id "Os01g0888132";	transcript_id "Os01t0888132-00";	11944
chr09	15242689	15243122	chr09	15229458	15230650	gene_id "Os09g0421400";	transcript_id "Os09t0421400-00";	12040
chr09	15242689	15243122	chr09	15230396	15230650	gene_id "Os09g0421400";	transcript_id "Os09t0421400-00";	12040
chr10	13801413	13801846	chr10	13788424	13788992	gene_id "Os10g0404700";	transcript_id "Os10t0404700-00";	12422
chr10	13801413	13801846	chr10	13788862	13788992	gene_id "Os10g0404700";	transcript_id "Os10t0404700-00";	12422
chr08	16981365	16981798	chr08	16994594	16995049	gene_id "Os08g0366300";	transcript_id "Os08t0366300-00";	12797
chr03	23885554	23885987	chr03	23898856	23898879	gene_id "Os03g0627000";	transcript_id "Os03t0627000-00";	12870
chr03	23885554	23885987	chr03	23898856	23899879	gene_id "Os03g0627000";	transcript_id "Os03t0627000-00";	12870
chr06	2035338	2035771	chr06	2021816	2022094	gene_id "Os06g0137800";	transcript_id "Os06t0137800-00";	13245
chr06	23734515	23734948	chr06	23720181	23720483	gene_id "Os06g0600550";	transcript_id "Os06t0600550-00";	14033
chr11	24682794	24683227	chr11	24697981	24697991	gene_id "Os11g0630200";	transcript_id "Os11t0630200-00";	14755
chr11	24682794	24683227	chr11	24697981	24699421	gene_id "Os11g0630200";	transcript_id "Os11t0630200-00";	14755
chr03	26005811	26006244	chr03	25990597	25990938	gene_id "Os03g0661850";	transcript_id "Os03t0661850-00";	14874
chr01	37273038	37273471	chr01	37288423	37288443	gene_id "Os01g0861200";	transcript_id "Os01t0861200-00";	14953
chr01	37273038	37273471	chr01	37288423	37296952	gene_id "Os01g0861200";	transcript_id "Os01t0861200-00";	14953
chr09	844898	845331	chr09	829334	829915	gene_id "Os09g0108800";	transcript_id "Os09t0108800-00";	14984
chr02	3502596	3503029	chr02	3518105	3518186	gene_id "Os02g0165300";	transcript_id "Os02t0165300-00";	15077
chr02	3502596	3503029	chr02	3518105	3519196	gene_id "Os02g0165300";	transcript_id "Os02t0165300-00";	15077
chr02	25553669	25554102	chr02	25537385	25537882	gene_id "Os02g0636851";	transcript_id "Os02t0636851-00";	15788
chr02	8125072	8125505	chr02	8106666	8108741	gene_id "Os02g0243700";	transcript_id "Os02t0243700-00";	16332
chr01	11939630	11940063	chr01	11956868	11956978	gene_id "Os01g0316550";	transcript_id "Os01t0316550-00";	16806
chr01	11939630	11940063	chr01	11956868	11958170	gene_id "Os01g0316550";	transcript_id "Os01t0316550-00";	16806
chr01	1748896	1749329	chr01	1731553	1731945	gene_id "Os01g0131200";	transcript_id "Os01t0131200-00";	16952
chr08	24622997	24623430	chr08	24640403	24640645	gene_id "Os08g0499225";	transcript_id "Os08t0499225-00";	16974
chr04	35048161	35048594	chr04	35066229	35066636	gene_id "Os04g0686150";	transcript_id "Os04t0686150-00";	17636
chr07	3258233	3258666	chr07	3240261	3240557	gene_id "Os07g0160232";	transcript_id "Os07t0160232-00";	17677
chr08	4034192	4034625	chr08	4053038	4053376	gene_id "Os08g0169201";	transcript_id "Os08t0169201-00";	18414
chr02	25755512	25755945	chr02	25736687	25737079	gene_id "Os02g0640900";	transcript_id "Os02t0640900-00";	18434
chr01	2776621	2777054	chr01	2757759	2758049	gene_id "Os01g0151001";	transcript_id "Os01t0151001-00";	18573
chr04	16947508	16947941	chr04	16925446	16928752	gene_id "Os04g0354300";	transcript_id "Os04t0354300-00";	18757
chr04	16947508	16947941	chr04	16928624	16928752	gene_id "Os04g0354300";	transcript_id "Os04t0354300-00";	18757
chr02	10492618	10493051	chr02	10511994	10512209	gene_id "Os02g0282150";	transcript_id "Os02t0282150-00";	18944
chr10	18765107	18765540	chr10	18745564	18746055	gene_id "Os10g0493700";	transcript_id "Os10t0493700-00";	19053
chr05	6268708	6269141	chr05	6248958	6249293	gene_id "Os05g0200220";	transcript_id "Os05t0200220-00";	19416
chr09	7249080	7249513	chr09	7229093	7229311	gene_id "Os09g0297700";	transcript_id "Os09t0297700-00";	19770
chr02	35128275	35128708	chr02	35107927	35108322	gene_id "Os02g0818201";	transcript_id "Os02t0818201-00";	19954
chr07	18642854	18643287	chr07	18663640	18664521	gene_id "Os07g0497800";	transcript_id "Os07t0497800-00";	20354
chr11	17748308	17748741	chr11	17769275	17769697	gene_id "Os11g0499101";	transcript_id "Os11t0499101-00";	20535
chr01	35200680	35201113	chr01	35179642	35180055	gene_id "Os01g0823800";	transcript_id "Os01t0823800-00";	20626
chr05	27897960	27898393	chr05	27921350	27922411	gene_id "Os05g0561100";	transcript_id "Os05t0561100-00";	22958
chr05	27897960	27898393	chr05	27921350	27923744	gene_id "Os05g0561100";	transcript_id "Os05t0561100-00";	22958
chr06	6501384	6501817	chr06	6524834	6525202	gene_id "Os06g0226066";	transcript_id "Os06t0226066-00";	23018
chr08	6019708	6020141	chr08	5995644	5996216	gene_id "Os08g0203250";	transcript_id "Os08t0203250-00";	23493
chr02	5416385	5416818	chr02	5440341	5440685	gene_id "Os02g0197050";	transcript_id "Os02t0197050-00";	23524
chr01	4222417	4222850	chr01	4198348	4198833	gene_id "Os01g0180100";	transcript_id "Os01t0180100-00";	23585
chr03	27418804	27419237	chr03	27394860	27395102	gene_id "Os03g0686600";	transcript_id "Os03t0686600-00";	23703
chr12	2232165	2232598	chr12	2256678	2256778	gene_id "Os12g0145933";	transcript_id "Os12t0145933-00";	24081
chr12	2232165	2232598	chr12	2256678	2257245	gene_id "Os12g0145933";	transcript_id "Os12t0145933-00";	24081
chr03	12223360	12223793	chr03	12190565	12199136	gene_id "Os03g0331200";	transcript_id "Os03t0331200-00";	24225
chr03	12223360	12223793	chr03	12199128	12199136	gene_id "Os03g0331200";	transcript_id "Os03t0331200-00";	24225
chr03	2301775	2302208	chr03	2276860	2277298	gene_id "Os03g0141300";	transcript_id "Os03t0141300-00";	24478
chr03	2301775	2302208	chr03	2277074	2277298	gene_id "Os03g0141300";	transcript_id "Os03t0141300-00";	24478
chr02	8881839	8882272	chr02	8855510	8856100	gene_id "Os02g0256900";	transcript_id "Os02t0256900-00";	25740
chr05	20419720	20420153	chr05	20393084	20393476	gene_id "Os05g0416801";	transcript_id "Os05t0416801-00";	26245
chr05	20419720	20420153	chr05	20393367	20393476	gene_id "Os05g0416801";	transcript_id "Os05t0416801-00";	26245
chr01	10904743	10905176	chr01	10931624	10931971	gene_id "Os01g0298200";	transcript_id "Os01t0298200-00";	26449
chr01	10904743	10905176	chr01	10931624	10932347	gene_id "Os01g0298200";	transcript_id "Os01t0298200-00";	26449
chr04	34135026	34135459	chr04	34108013	34108521	gene_id "Os04g0668501";	transcript_id "Os04t0668501-00";	26506
chr04	34135026	34135459	chr04	34108417	34108521	gene_id "Os04g0668501";	transcript_id "Os04t0668501-00";	26506
chr03	11613054	11613487	chr03	11585658	11585969	gene_id "Os03g0320825";	transcript_id "Os03t0320825-00";	27086
chr01	4113126	4113559	chr01	4141380	4141700	gene_id "Os01g0179450";	transcript_id "Os01t0179450-00";	27822
chr08	20695916	20696349	chr08	20727624	20728048	gene_id "Os08g0429600";	transcript_id "Os08t0429600-00";	31276
chr08	20695916	20696349	chr08	20727624	20740891	gene_id "Os08g0429600";	transcript_id "Os08t0429600-00";	31276
chr08	21719826	21720259	chr08	21753310	21753909	gene_id "Os08g0446001";	transcript_id "Os08t0446001-00";	33052
chr03	7739011	7739444	chr03	7704831	7705184	gene_id "Os03g0246000";	transcript_id "Os03t0246000-00";	33828
chr03	12068386	12068819	chr03	12103108	12103413	gene_id "Os03g0329400";	transcript_id "Os03t0329400-00";	34290
chr04	30776393	30776826	chr04	30811261	30811758	gene_id "Os04g0608700";	transcript_id "Os04t0608700-00";	34436

Table 2: Sorted distance from the closest genes found from all *mPing* element insertions.

Bibliography

Access, Open (2014). “The 3 , 000 rice genomes project”. In: pp. 1–6.

Alexandrov, Nickolai et al. (2015). “SNP-Seek database of SNPs derived from 3000 rice genomes”. In: *Nucleic Acids Research* 43.D1, pp. D1023–D1027. ISSN: 13624962. DOI: `10.1093/nar/gku1039`.

Carpentier, Marie Christine et al. (2019). “Retrotranspositional landscape of Asian rice revealed by 3000 genomes”. In: *Nature Communications* 10.1. ISSN: 20411723. DOI: `10.1038/s41467-018-07974-5`. URL: `http://dx.doi.org/10.1038/s41467-018-07974-5`.

Chaparro, Cristian et al. (2007). “RetrOryza: A database of the rice LTR-retrotransposons”. In: *Nucleic Acids Research* 35.SUPPL. 1, pp. 66–70. ISSN: 03051048. DOI: `10.1093/nar/gkl780`.

Civán, Peter et al. (2015). “Three geographically separate domestications of Asian rice”. In: *Nature Plants* 1.November, pp. 1–5. ISSN: 2055026X. DOI: `10.1038/nplants.2015.164`.

Civáň, Peter and Terence A. Brown (2018). “Misconceptions regarding the role of introgression in the origin of oryza sativa subsp. Indica”. In: *Frontiers in Plant Science* 871.November, pp. 1–4. ISSN: 1664462X. DOI: `10.3389/fpls.2018.01750`.

Gardner Eugene J. et al. (2017). “The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology”. In: *Genome Research* 27.11, pp. 1916–1929. ISSN: 1088-9051. DOI: `10.1101/gr.218032.116`.

Goerner-Potvin, Patricia and Guillaume Bourque (2018). “Computational tools to unmask transposable elements”. In: *Nature Reviews Genetics* 19.11, pp. 688–704. ISSN: 14710064. DOI: `10.1038/s41576-018-0050-x`. URL: `http://dx.doi.org/10.1038/s41576-018-0050-x`.

Huang, Jian et al. (2009). “Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the hAT family in rice”. In: *Genomics* 93.3, pp. 274–281. ISSN: 08887543. DOI: `10.1016/j.ygeno.2008.11.007`. URL: `http://dx.doi.org/10.1016/j.ygeno.2008.11.007`.

Izawa, Takeshi (2008). “The Process of rice domestication: A new model based on recent data”. In: *Rice* 1.2, pp. 127–134. ISSN: 19398425. DOI: `10.1007/s12284-008-9014-7`.

Jiang, Ning et al. (2003). “An active DNA transposon family in rice”. In: *Nature* 421.6919, pp. 163–167. ISSN: 00280836. DOI: `10.1038/nature01214`.

Jiang, Ning et al. (2004). “Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs)”. In: *Current Opinion in Plant Biology* 7.2, pp. 115–119. ISSN: 13695266. DOI: `10.1016/j.pbi.2004.01.004`.

John Maynard Smith, John Haigh (1974). “Hitch-Hiking Effect of a Favourable Gene”. In: *Population Genetics* 1974, pp. 109–114. DOI: `10.1007/978-1-4613-3924-3_7`.

Kazuhiro, Kikuchi et al. (2003). “The plant MITE mPing is mobilized in anther culture”. In: *Nature* 421.6919, p. 167.

Kidwell, M. (1993). “Lateral Transfer in Natural Populations of Eukaryotes”. In: *Annual Review of Genetics* 27.1, pp. 235–256. ISSN: 00664197. DOI: 10.1146/annurev.genet.27.1.235.

Kidwell, Margaret G and Damon R Lisch (2000). “MG Kidwell et al 2000 Review.pdf”. In: 15.3, pp. 95–99.

Larribe F., Fearnhead P. (2011). “On composite likelihoods in statistical genetics”. In: *Statistica Sinica* 21.1, pp. 43–69. ISSN: 1017-0405.

Li, Jia Yang, Jun Wang, and Robert S. Zeigler (2014). “The 3,000 rice genomes project: New opportunities and challenges for future rice research”. In: *GigaScience* 3.1, pp. 1–3. ISSN: 2047217X. DOI: 10.1186/2047-217X-3-8.

Lin, Xiuyun et al. (2006). “In planta mobilization of mPing and its putative autonomous element Pong in rice by hydrostatic pressurization”. In: *Journal of Experimental Botany* 57.10, pp. 2313–2323. ISSN: 00220957. DOI: 10.1093/jxb/erj203.

Lu, Chen et al. (2012). “Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in oryza sativa”. In: *Molecular Biology and Evolution* 29.3, pp. 1005–1017. ISSN: 07374038. DOI: 10.1093/molbev/msr282.

Lu, Lu et al. (2017). “Tracking the genome-wide outcomes of a transposable element burst over decades of amplification”. In: *Proceedings of the National Academy of Sciences* 114.49, E10550–E10559. ISSN: 0027-8424. DOI: 10.1073/pnas.1716459114.

Martin, Sandra L. and David J. Garfinkel (2003). “Survival strategies for transposons and genomes”. In: *Genome Biology* 4.4. ISSN: 14656906. DOI: 10.1186/gb-2003-4-4-313.

Naito, Ken et al. (2009). “Unexpected consequences of a sudden and massive transposon amplification on rice gene expression”. In: *Nature* 461.7267, pp. 1130–1134. ISSN: 00280836. DOI: 10.1038/nature08479. URL: http://dx.doi.org/10.1038/nature08479.

Naito, Ken et al. (2014). “mPing: The bursting transposon”. In: *Breeding Science* 64.2, pp. 109–114. ISSN: 1344-7610. DOI: 10.1270/jsbbs.64.109.

Nakazaki, Tetsuya et al. (2003). “Mobilization of a transposon in the rice genome”. In: *Nature* 421.6919, pp. 170–172. ISSN: 00280836. DOI: 10.1038/nature01219.

Nielsen, Rasmus (2005). “Molecular Signatures of Natural Selection”. In: *Annual Review of Genetics* 39.1, pp. 197–218. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.39.073003.112420.

Oki, Nobuhiko et al. (2008). “A genome-wide view of miniature inverted-repeat transposable elements (MITEs) in rice, Oryza sativa ssp. japonica”. In: *Genes & Genetic Systems* 83.4, pp. 321–329. ISSN: 1341-7568. DOI: 10.1266/ggs.83.321.

Pavlidis, Pavlos and Nikolaos Alachiotis (2017). “A survey of methods and tools to detect recent and strong positive selection”. In: *Journal of Biological Research-Thessaloniki* 24.1, pp. 1–17. ISSN: 2241-5793. DOI: 10.1186/s40709-017-0064-0.

Pavlidis, Pavlos et al. (2013). “SweeD: Likelihood-based detection of selective sweeps in thousands of genomes”. In: *Molecular Biology and Evolution* 30.9, pp. 2224–2234. ISSN: 07374038. DOI: 10.1093/molbev/mst112.

Sabot, Francois (2014). “Tos17 rice element: Incomplete but effective”. In: *Mobile DNA* 5.1, pp. 2–5. ISSN: 17598753. DOI: 10.1186/1759-8753-5-10.

Stein, Joshua C. et al. (2018). “Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza”. In: *Nature Genetics* 50.2, pp. 285–296. ISSN: 15461718. DOI: 10.1038/s41588-018-0040-0. URL: http://dx.doi.org/10.1038/s41588-018-0040-0.

Volff, Jean Nicolas (2006). “Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes”. In: *BioEssays* 28.9, pp. 913–922. ISSN: 02659247. DOI: 10.1002/bies.20452.

Wicker, Thomas et al. (2007). “A unified classification system for eukaryotic transposable elements”. In: *Nature Reviews* 8, pp. 973–982. ISSN: 14710056. DOI: doi:10.1038/nrg2165. URL: https://www.nature.com/articles/nrg2165.pdf.

Wildschutte, Julia Halo et al. (2016). “Discovery of unfixed endogenous retrovirus insertions in diverse human populations”. In: *Proceedings of the National Academy of Sciences* 113.16, E2326–E2334. ISSN: 0027-8424. DOI: 10.1073/pnas.1602336113.

Zhang, Qun-Jie and Li-Zhi Gao (2017). “ Rapid and Recent Evolution of LTR Retrotransposons Drives Rice Genome Evolution During the Speciation of AA-Genome Oryza Species ”. In: *G3&#58; Genes—Genomes—Genetics* 7.6, pp. 1875–1885. DOI: 10.1534/g3.116.037572.