



POLITÉCNICA



E.T.S. DE INGENIERÍA AGRONÓMICA,
ALIMENTARIA Y DE BIOSISTEMAS

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA AGRONÓMICA,
ALIMENTARIA Y DE BIOSISTEMAS
GRADO EN BIOTECNOLOGÍA

DEPARTAMENTO DE MEJORA ASISTIDA POR GENÓMICA (CBGP)

***Detección de variaciones en el número de copias en
germoplasma de Zea mays adaptado a la altitud.***

***"Copy number variations detection on contrasting altitude
adapted Zea mays germplasm"***

TRABAJO FIN DE GRADO

Autor: Pablo Atienza López

Tutor: Dr. Julio Isidro y Sánchez

Cotutor: Dr. Tiago Loureira de Carvalho

Junio de 2021



POLITÉCNICA



E.T.S. DE INGENIERÍA AGRONÓMICA,
ALIMENTARIA Y DE BIOSISTEMAS

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA
AGRONÓMICA, ALIMENTARIA Y DE BIOSISTEMAS

GRADO DE BIOTECNOLOGÍA

**DETECCIÓN DE VARIACIONES EN EL NÚMERO DE COPIAS EN GERMOPLASMA DE ZEA MAYS
ADAPTADO A LA ALTITUD.**

TRABAJO FIN DE GRADO

PABLO ATIENZA LÓPEZ

MADRID, 2021

Directores: Julio Isidro y Sánchez (Departamento de Biotecnología-Biología Vegetal)
Tiago Loureira de Carvalho (Centro de Biotecnología y Genómica de Plantas)



POLITÉCNICA



E.T.S. DE INGENIERÍA AGRONÓMICA,
ALIMENTARIA Y DE BIOSISTEMAS

**TITULO DEL TFG- DETECCIÓN DE VARIACIONES EN EL NÚMERO DE COPIAS EN
GERMOPLASMA DE ZEA MAYS ADAPTADO A LA ALTITUD.**

**Memoria presentada por Pablo Atienza López para la obtención del título de Graduado
en Biotecnología por la Universidad Politécnica de Madrid**

Fdo: Pablo Atienza López

VºBº Tutor y Director del TFG

ISIDRO
SANCHEZ JULIO -
79202367G
Digitally signed
by ISIDRO
SANCHEZ JULIO
- 79202367G
Date:
2021.06.23
14:56:17 +02'00'

Tiago
Loureiro
de
Carvalho
Firmado
digitalmente por
Tiago Loureiro
de Carvalho
Fecha:
2021.06.24
02:41:38 +02'00'

**D Julio Isidro y Sánchez
Jefe de grupo del departamento de mejora asistida por genómica
Dpto de Biotecnología-Biología vegetal
ETSIAAB - Universidad Politécnica de Madrid**

Madrid, 24 de junio de 2021

Abstract

...

Maize is one of the most important cultivars around the world, and it has its domestication origin in the lowland areas of Mexico. As it was spread during domestication, it adapted to the highland regions of the country, where it encountered several setbacks, such as a cooler and dryer environment and the imposing presence of *Puccinia sorghi*, a fungus that causes the maize common rust. In several studies, a specific type of structural variations known as Copy Number Variations (CNVs) have been linked to adaptation traits. However, the genetic mechanisms underlying this adaptation of Mexican maize to the highland areas has not yet been related with the presence of CNVs. Throughout this study, we have detected the presence and assessed the effect of CNVs in a population of Mexican maize adapted to high altitude. CNVs were called in two populations of Mexican maize, one from lowland areas and one of highland areas. By applying a CNV calling method based on read depth that only looks for CNVs in the coding regions, CNVs in both populations have been detected, as well as duplications and deletions unique for each population. When analyzing more in detail the genes called as a duplication only in the highlands population, eleven genes have been found associated with phenotypic traits. We found three CNVs in genes regions that had a significant effect on resistance to abiotic stresses or to *Puccinia sorghi* fungi. These results demonstrate the role of CNVs in the functional adaptability to specific environments. We believe that this work serves as basis for future studies where the presence of these genes is experimentally validated. Moreover, this first approach hints possible unknown effects of different genes on the Mexican highland maize phenotype.

Resumen

...

El maíz es uno de los cultivos más importantes del mundo, y tiene su origen de domesticación en las zonas bajas de México. A medida que se fue extendiendo durante su domesticación, se adaptó a las regiones del altiplano del país, donde encontró varios contratiempos, como un ambiente más frío y seco y la imponente presencia de *Puccinia sorghi*, un hongo que causa la roya común del maíz. En varios estudios, un tipo específico de variaciones estructurales conocidas como variaciones en el número de copias (CNV) se han relacionado con los rasgos de adaptación. Sin embargo, los mecanismos genéticos que subyacen a esta adaptación del maíz mexicano a las zonas del altiplano no se han relacionado aún con la presencia de CNVs. En este estudio, hemos detectado la presencia y evaluado el efecto de CNVs en una población de maíz mexicano adaptada a la altura. Los CNVs fueron llamados en dos poblaciones de maíz mexicano, una de zonas bajas y otra de zonas altas. Aplicando un método de llamada de CNVs basado en la profundidad de lectura que solo busca CNVs en las regiones codificantes, se han detectado CNVs en ambas poblaciones, así como duplicaciones y deleciones únicas para cada población. Al analizar con más detalle los genes llamados como duplicación solo en la población de las tierras altas, se ha encontrado asociación entre ellos y algunos rasgos fenotípicos, como la resistencia a la sequía y a *Puccinia sorghi*. Estos rasgos, entre otros encontrados en este estudio, están relacionados con la adaptación del maíz del altiplano mexicano a su entorno. Creemos que este trabajo sirve de base para futuros estudios donde se valide experimentalmente la presencia de estos genes. Además, esta primera aproximación deja entrever posibles efectos desconocidos de diferentes genes en el fenotipo del maíz del altiplano mexicano.

Acknowledgements

...

In the first place, I would like to thank the Department of Genomic Assisted Breeding, specially to Dr. Julio Isidro y Sanchez and Dr. Tiago Loureiro de Carvalho, for all their help they have lent me. Tiago has helped me with every issue I had, giving always fast and useful answers despite the simplicity or complexity of the issue. He has been aware of every need I had and much of what I have learnt is due to him.

I want to deeply thank Julio for allowing me to work in his team since the beginning, not being treated as an undergraduate, but as part of the department in our weekly meeting and presentations. He has helped me to improve my scientific and writing skill to a great extent.

As well, I would like to give a special mention to Eric Gonzalez, the person who introduced me in the world Structural Variations and maize. During the last year in Munich he taught me how to work in this field, and he gave me the idea for this thesis. Not only that, but he has been paying close attention to my work and even helped me in his spare time.

Finally, I would like to thank my whole family for their patience and comprehension during this period. Last but not least, I wish to thank M^a Ángeles for bearing my thousand explanations of what I was doing every single day. If there is somebody who knows my thesis better than me, it is her.

Contents

Abstract	iv
Resumen	v
Acknowledgements	vi
List of abbreviations	x
Chapter 1. Introduction	1
Chapter 2. Material and Methods	3
2.1 Material	3
2.1.1 Plant material	3
2.1.2 Reference genome and gene models	3
2.1.3 B73 functional annotation	4
2.2 Methods	4
2.2.1 Preprocessing	4
2.2.2 Determination of low mappability regions (LMRs)	5
2.2.3 Copy Number Variation calling	5
2.2.4 Copy Number Variations clustering	6
2.2.5 Variant Calling	7
2.2.6 Variant summary	8
2.2.7 Functional enrichment analysis	8
2.2.8 CNV Visualization	8
Chapter 3. Results	9
3.1 Plant material analysis	9
3.2 Mappability filter	9
3.3 Copy number and variation calling preprocessing	10
3.4 Variant Calling	11
3.4.1 Base Quality Score Recalibration	11
3.4.2 Single nucleotide polymorphisms and Indels quantification	11
3.4.3 Population structure	12
3.5 CNV calling	12
3.5.1 Gene size modification	12
3.5.2 CNV genes count	12
3.6 Functional enrichment analysis	14

Chapter 4. Discussion	17
Chapter 5. Conclusion	20
Chapter 6. Bibliography	21
A Annex	30

List of Tables

Table 1	Genes function summary	16
Table S1	Plant material	30
Table S2	Genes function	30

List of Figures

Figure 1	Mexican climates	3
Figure 2	CNV calling workflow	6
Figure 3	Summary of maize samples	9
Figure 4	Mappability results	10
Figure 5	BAM files coverage	11
Figure 6	PCA and heatmap	13
Figure 7	CNVs among populations	15
Figure S1	Quality of the reads	31
Figure S2	Base Quality Score Recalibration	32
Figure S3	Gene size variation	33
Figure S4	Coverage vs CNV number	34
Figure S5	CNV number for 85% threshold	35
Figure S6	CNV number with no mappability filter	36

List of abbreviations

- **BAM:** Binary sequence alignment map
- **BED:** Browser extensible data
- **bp:** base pair
- **BQSR:** Base quality score recalibration
- **CNV:** Copy number variant
- **DNA:** Deoxyribonucleic acid
- **DRP:** Differentially regulated protein
- **FDR:** False discovery rate
- **G6P-DH:** Glucose-6-phosphate dehydrogenase
- **HL:** Highlands
- **IGV:** Integrative genome viewer
- **kb:** kilobase
- **LL:** Lowlands
- **LMR:** Low mappability region
- **mRNA:** Messenger ribonucleic acid
- **NGS:** Next generation sequencing
- **PAV:** Presence absence variation
- **PC:** Principal component
- **PCA:** Principal component analysis
- **PME:** Pectinmethyl esterase
- **PPP:** Pentoses-phosphate pathway
- **PR:** Paired read
- **RD:** Read depth
- **RNA:** Ribonucleic acid
- **SAM:** Sequence alignment map
- **SNP:** Single nucleotide polymorphism
- **SV:** Structural variation
- **UCI:** Unilateral-Cross incompatibility
- **WGS:** Whole genome sequencing
- **VCF:** Variant call format

Chapter 1. Introduction

Differences between plant genomes range from single nucleotide polymorphisms (SNPs) to large-scale polymorphism, the latter are termed structural variants. Structural variations include rearrangements, such as inversions and translocations and also copy number variations (CNVs). The presence-absence variation (PAV) is the most extreme form of CNV, in which a specific sequence can be present in some individuals and missing in others (Swanson-Wagner et al., 2010). There are evidences that demonstrate that structural variations or CNVs or PAVs can lead to variation in the genome content. (Haberer et al., 2020). Copy number variation studies have been well reported in many mammalian species, such as in humans (Wong et al., 2007; Conrad et al., 2006; Craddock et al., 2010), mice (Graubert et al., 2007), rats (Guryev et al., 2008), monkeys (Lee et al., 2008), and canines (Chen et al., 2009). In plants, the extent and role of CNVs in resistance against pathogens and pests exploration has just begun (Dolatabadian A., 2017). For example, Bai et al. (2016) validated 28 functional CNV genes implying that CNVs might be involved in control of flowering time, insect resistance, RNA interference, response to salt and dehydration stress. Other CNVs have no apparent influence on phenotype, but evidence also indicates that additional genetic or environmental factors may influence whether CNVs have a detectable phenotypic effect (Clancy, 2008).

Recent studies in plant genetics have shown that CNVs may control a considerable amount of variation at the trait level (Schiessl et al., 2017; Dolatabadian et al., 2017; Saxena et al., 2014), with most CNVs being detected by specific algorithms from next-generation sequencing data (NGS) (Abyzov et al., 2011; Knaus, Grünwald, 2018; Mayrhofer et al., 2016). Structural genomic variations in maize *Zea mays* ssp. *mays*, such as copy gain (duplication) and copy loss (deletion), may have a variety of effects on gene dosage and phenotype and thus directly influence gene expression (Brohammer et al., 2018; Chia et al., 2012; Hirsch et al., 2014; Lu et al., 2015; Maron et al., 2013).

In maize the recent completion of a reference genome B73, affords the opportunity to assess structural variation and complexity within this species (Schnable et al., 2009). Nearly 85% of the maize genome is composed of hundreds of families of transposable elements (Schnable, 2012) that can produce an increase of structural variation among maize genotypes. In maize, cytogenetics studies (McClintock et al., 1981; Kato et al., 2004) and sequencing of multiple haplotypes (Wang, Dooner, 2006) have identified structural variations.

Here, we focused on two populations of Mexican maizes that are adapted to the lowland and highland regions of Mexico. Mexican highland maize presents better cold

and drought resistance than lowland maize (LL), as well as resistance to *Puccinia sorghi*, a fungus that causes the Maize's Common Rust (Eagles, Lothrop, 1994).

In this study, we used NGS technologies from pair-end reads to assess the frequency and identity of genes affected by CNVs within a contrasting maize population. In this sense, the aims of my research were i) to assess the presence of CNVs in two populations of maize landraces with different environmental conditions, ii) to detect genes that have CNVs among landraces, and iii) to perform a functional characterization of the CNVs found, to assess whether CNVs have an impact in the phenotypic adaptation of Mexican highland maize.

Chapter 2. Material and Methods

2.1 Material

2.1.1 Plant material

A total set of 11 individuals, 6 from Mexican Highlands population (HL) and 5 from Mexican Lowlands population (LL), was selected from a total of 31 maize landrace accessions (Wang et al., 2017) (GenBank BioProject PRJNA72137). DNA was extracted from leaves using a standard cetyltrimethyl ammonium bromide (Doyle, Doyle, 1987). The samples were sequenced using Illumina HiSeq 2000 sequencing (100-bp paired-end) following the established protocols by the Beijing Genomics Institute (<https://www.bgi.com/global/>). From the 11 individuals, 10 had a sequencing depth higher than 26x, and one HL individual had a sequencing depth of 7,5 (Supplementary Fig. S1). We mapped the geographical coordinates of each sample using the libraries developed by Schweinberger (2021)

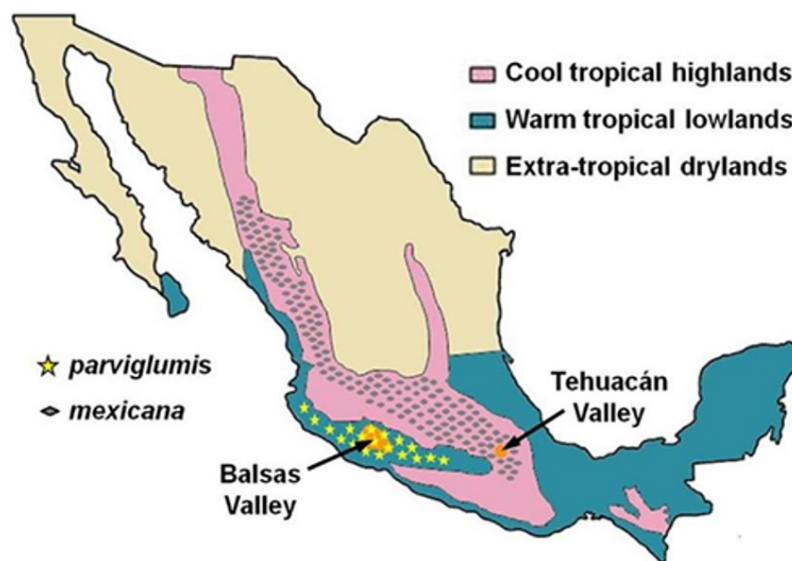


Figure 1: Representation of the climates in relation with the altitude. The cool highland areas are located in the south center region of Mexico. (Sawers, Sanchez-Leon, 2011)

2.1.2 Reference genome and gene models

To compare the sequence with the reference genome and perform the alignments, we used B73v5 reference genome, downloading it from MaizeGDB.org (Portwood et al., 2019) (<https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/Zm-B73-REFERENCE-NAM-5.0.fa.gz>). The gene set used for the study was downloaded

from MaizeGDB.org (https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/Zm-B73-REFERENCE-NAM-5.0_Zm00001eb.1.gff3.gz). From the whole set, only the positions of the genes were extracted, without taking into account their transcripts. The Zm00001eb.1 is the recommended gene model set for the B73v5 genome, containing 39.034 genes. Some smaller genes (under 1kb) may be in regions that were called as a deletion by the read depth method that result to be false positives deletions when visualizing. Paired-read and split-read methods perform better for deletions under 1kb in size (Trost et al., 2018a). Therefore, an alternative gene model was created by expanding up and downstream those genes from the original set that were less than 1000 basepairs in size. The difference between the gene size and 1000 was calculated, and half of it was added to each side of the gene.

2.1.3 B73 functional annotation

The functional annotation for the B73v5 gene models was created by mixing two different functional annotations, since there is no complete functional annotation for B73v5. The main functional annotation was obtained from the GAMER annotation for B73v4, which contains a total of 39.324 genes (Wimalanathan et al., 2018). To use it for the B73v5 release, the gene IDs were cross-referenced with a liftoff file (https://datacommons.cyverse.org/browse/iplant/home/maizegdb/maizegdb/B73v5_JBROWSE_AND_ANALYSES/B73v3-B73v5_and_B73v4-B73v5_gene_model_associations/B73v4_B73v5_liftoff_genemodel_CDS_xref_shortened.txt.gz). The final gene number of the B73v5-adapted GAMER functional annotation was of 32.290 genes.

The second functional annotation was obtained from the InterProScan (Blum et al., 2021) results of the B73v5 proteins (https://download.maizegdb.org/Zm-B73-REFERENCE-NAM-5.0/Zm-B73-REFERENCE-NAM-5.0_Zm00001eb.1_interproscan.tsv.gz). It contained 16.948 annotated genes. From this set, 1.329 were not included in the original B73v5-GAMER annotation, and were added to create the B73v5 GAMER-InterProScan Annotation used for the study, comprising a total of 33.169 genes.

2.2 Methods

2.2.1 Preprocessing

The quality of the reads was visualized using FastQC v0.11.8 (Andrews et al., 2012). Afterwards, we used Trimmomatic v0.38 (Bolger et al., 2014) to remove low quality sequences. The trim parameters were set as follows: Leading: 3 Trailing: 3

SlidingWindow: 4:15 MinLen: 36. Once the trimming of the sequence is done, FastQC was used again to confirm the correct working of Trimmomatic-0.38. The alignment was performed against the B73v5 reference genome using Bwa-mem-0.7.17 (Li, Durbin, 2009) using default parameters with the flag -M for Picard tools compatibility. The resulting Sequence Alignment Map file (SAM) is converted to binary (BAM) for storage reasons output was sorted using Picard tools v.2.4.1 (<http://broadinstitute.github.io/picard>). Both single-end and paired-end BAM files for each individual were merged using Samtools v.1.9 (Li et al., 2009), and duplicated reads removed with Picard Tools using default parameters and the flag Remove_Duplicates=true. The last two steps were the indexation of the BAM files using Picard Tools, and the indel-realignment with the Genome Analysis Toolkit (GATK) v.3.5.0 (Auwera Van der, D O'Connor, 2020).

2.2.2 Determination of low mappability regions (LMRs)

The workflow for the determination of the mappability of the reference genome is described in <http://lh3lh3.users.sourceforge.net/snpable.shtml>. The first step was to split the genome in one step overlapping artificial reads of 35bps of length (35-mers). Afterwards, using *bwa aln*, all the 35-mers were mapped against the genome. Then the 35-mers were used to assign a mappability score to each position in the genome; 0 for those positions where all thirty-five 35-mers cannot be mapped unambiguously, 1 if the majority of the 35-mers cannot be mapped unambiguously, 2 if the majority can be mapped unambiguously but with one or more mismatches, and 3 if the majority can be mapped unambiguously without mismatches. Only sites of the genome with a mappability score of 3 were considered as mappable (Ramos-Madrigal et al., 2016).

The rest of the genome was soft masked by switching these regions to lower case nucleotides, and the file containing the positions of the low mappability regions was created (BED file). The filtering of the detected CNVs was carried out using the script “compare_with_RLCR_definition.py” (<https://github.com/bjtrost/TCAG-WGS-CNV-workflow>) created by the group of Trost et al. (2018b) using the BED file of the low mappability regions (score <3).

2.2.3 Copy Number Variation calling

Figure 2 shows the steps followed during the CNV calling workflow. We filtered the BAM files for a threshold mapping quality of 30, using samtools v1.9. Per gene mean read depth was calculated with bedtools v2.28.0 (Quinlan, Hall, 2010) with the module coverage, option '-mean', using a BED file with the positions of the genes.

The set containing 39.172 genes was used to determining the regions were bedtools would extract the coverage. To determine CNV, the mean genes read depth was considered as average read depth. Duplications were called when the genes presented a 3-fold increase of the average read depth. Genes with a 0.1-fold the average read depth were considered as deletions. These RD comparisons were calculated in R programming version 6 (R Core Team, 2020). The set of CNVs obtained was filtered against a BED file containing the low mappability regions of the maize genome, using a script developed by Trost et al. (2018b). Those CNVs which presented an overlap with the low mappability regions greater than 70% were discarded from the final set. In further analyses, the threshold has been raised to 85% and 100%.

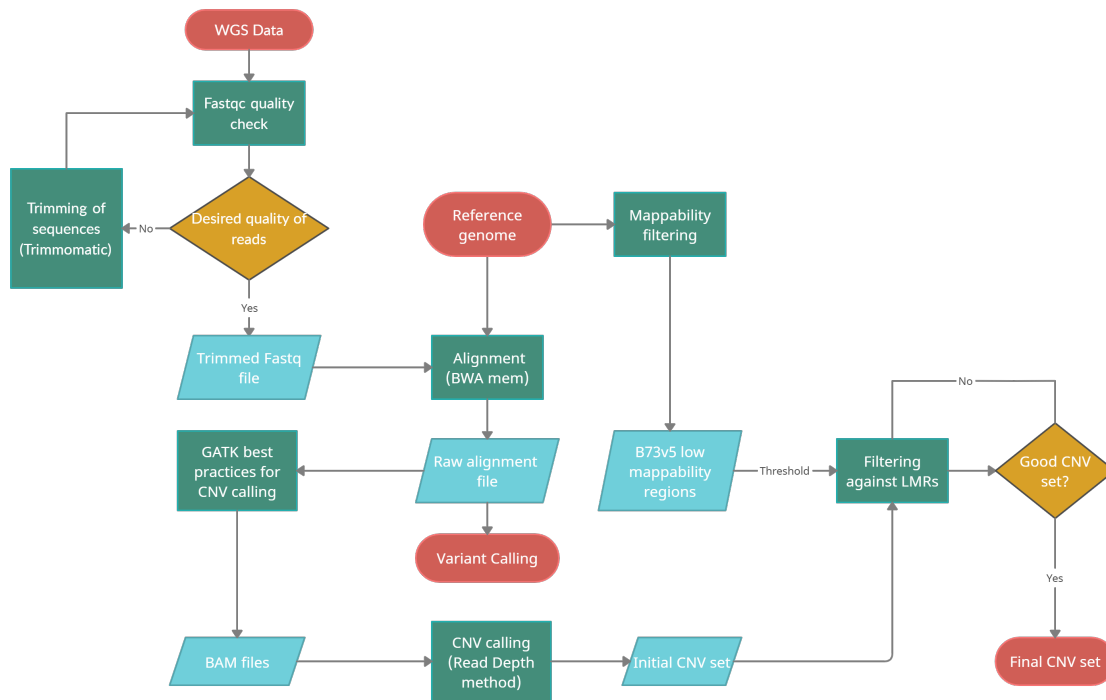


Figure 2: CNV calling workflow used in this study.

2.2.4 Copy Number Variations clustering

The CNVs called for all the individuals were stored in a table with 39.172 columns, one per gene, and 10 rows, one per individual. Values 1, -1 and 0 were given for duplicated, deleted and normal genes respectively. In this thesis, duplicated genes are named duplications and deleted genes, deletions, even though not being proper CNVs regions (a CNV is a region of DNA, not necessarily a gene). Total CNVs and per population were counted, as well as the number of deletions and duplications. Finally, we compared the CNVs among populations, and we created six groups: HL

only duplications, HL only duplications, common duplications, common deletions, LL only deletions and LL only deletions. These groups are used in the functional enrichment analysis to assess differences between the two populations.

2.2.5 Variant Calling

Through the variant calling, the process by which we identify variants from sequence data, the SNPs and Indels were called. The best practices for variant calling workflow proposed by GATK was used to call for SNPs and insertions-deletions (Indels) across the 10 individuals. The pipeline started from the raw mapped reads with removed duplicates. The Base Quality Score Recalibration (BQSR) was performed using two different training sets for the base recalibration: all the genome and only chromosome one. After recalibrating the BAM files, the variant calling of each individual was carried out. In these steps, the calling were split by chromosome, to reduce the computing time. A database was generated to store all the variants called per sample and afterwards, the variants from all of the individuals were processed jointly to generate a single VCF file with the unfiltered variants per chromosome. The variants were separated by SNPs and Indels, and both were filtered with different parameters. The SNPs filtering thresholds were: lower than 2.0 for QualByDepth (QD), the normalization of the variant quality; higher than 60.0 for FisherStrand (FS), which is the Phred-scaled probability that there is a strand bias at the site; higher than 4.0 for StrandOddsRatio (SOR), being another estimation of strand bias of variants that cover both alleles; lower than 40.0 for the RMSMappingQuality (MQ), the root-mean-square mapping quality over all the reads at the site; lower than -12.5 for MappingQualityRankSumTest (MQRankSum), an approximation of the Rank Sum Test for mapping qualities; less than -8.0 for the ReadPosRankSumTest, a comparison of whether the positions of the reference and alternate alleles differ within reads. Different threshold values were applied for the Indels: lower than 2.0 for QD, higher than 200.0 for FD and higher than 10.0 for SOR. The variants that reached one of the previous thresholds were deleted from the final set.

The filtered VCF files of SNPs and Indels per chromosome were joined, separately, and annotated using SnpEff version 5.0 (Cingolani et al., 2012). An HTML report and a summary table were generated. Tassel (Bradbury et al., 2007) was used to generate the Hapmap files from the VCF files, as well as the numerical files (coded 0; 0.5; 1).

2.2.6 Variant summary

The summary of the variant calling was done using R version 3.6.1 (R Core Team, 2020) using all the information from the numerical files. SNPs and Indels were used to represent the proportion of missing alleles, the minor allele frequency, the distribution of variants align chromosomes and principal coordinate analyses. A heatmap with the individuals grouped by population clusters was created as well.

2.2.7 Functional enrichment analysis

We used Blast2GO 5.2 (Götz et al., 2008) to perform the enrichment analysis. The functional annotation was loaded into the software and the enrichment analysis was carried out for the different gene sets (HL only duplications, HL only duplications, etc.). Blast2GO performs enrichment analysis with a Fisher's Exact Test, applying a False Discovery Rate (FDR) correction. GO-terms with an FDR value under 0.05 were considered as enriched functionally in respect to the genome. Firstly, the set of duplications that are only found in the HL population was analysed. The different gene sets and all the HL and LL CNVs were tested as well. Different thresholds of mappability were tested using Blast2GO, such as the same groups but for the 85% of threshold and for the HL populations as well.

The CNVs that are duplications in only the HL population were blasted using BlastX and BlastN (Boratyn et al., 2013) against genes databases of *Zea mays* and *Arabidopsis thaliana*. The function of the genes and all related information was searched in Maize GDB (Portwood et al., 2019).

2.2.8 CNV Visualization

A set of at least 15 CNVs from each set was visualized with the tool Integrative Genomic Viewer (IGV) v2.8.2 (Thorvaldsdottir et al., 2013), importing the BAM files and reference genome. Therefore, a validation of the CNVs called can be achieved by concentrating on the specific regions where the CNV is expected to take place (Trost et al., 2018b). According to Trost et al. (2018b), a CNV is considered as valid, if true deletions are determined with a sudden drop in coverage in the region, while duplications with a sudden increase in coverage.

Chapter 3. Results

3.1 Plant material analysis

We mapped the global positioning system geographical location of both maize populations in Mexico, and also the average altitude of plant growth (Figure 3). It can be seen that HL maize samples were located further north than LL maize (Figure 3A). The LL population showed more map distribution, even a sample was found outside of Mexico, although they shared the same environmental conditions (Fig 1). The average altitude for HL and LL was 2.300 meters and for LL 277 meters over sea level (Fig 3B). From this figure, we can see that geographically both populations presented a clear separation, except in the center of Mexico where the environments were more homogeneous.

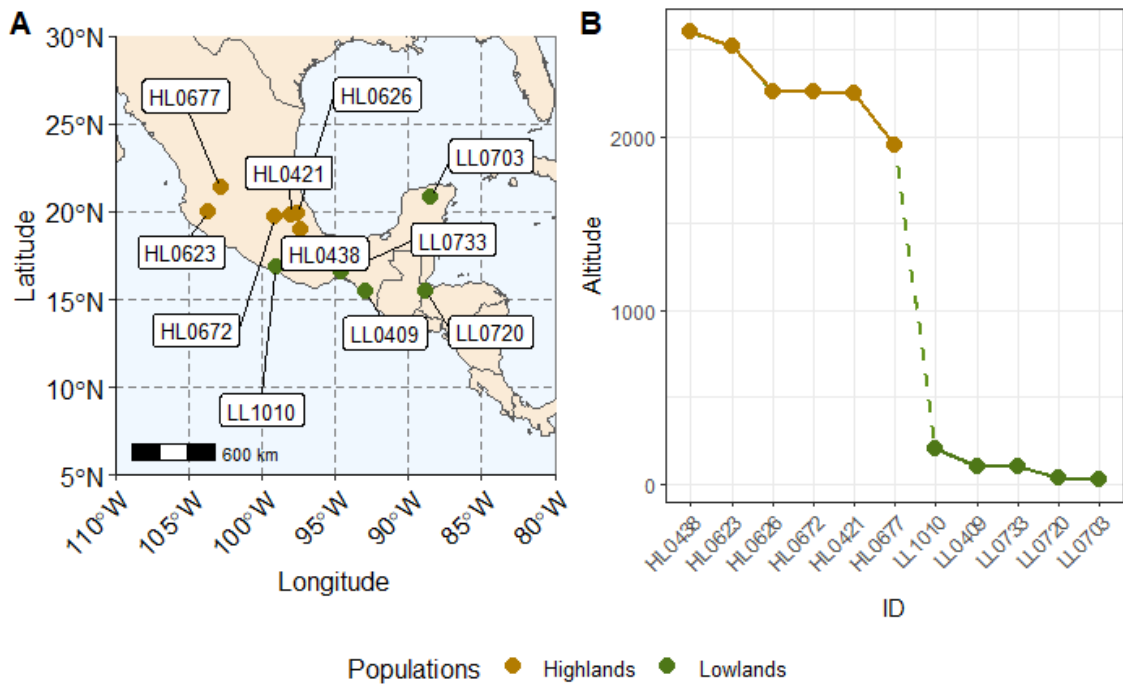


Figure 3: Summary of the maize samples. **A.** Geographical location of the maize samples. **B.** Representation of the altitude where the plants had grown.

3.2 Mappability filter

The results of the mappable percentage of the reference genome (B73v5) and of the genes that are located in mappable regions of the genome is shown in Figure 4. Results indicate that 20% of the genome was found to be mappable and the remaining 80% was determined as low mappability reads (LMRs) (Fig. 4A). In regard

to the genes, we assessed that 78% of the gene set used for this study had less than 70% of overlap with LMRs, thus being in unique regions of the maize genome (Fig 4).

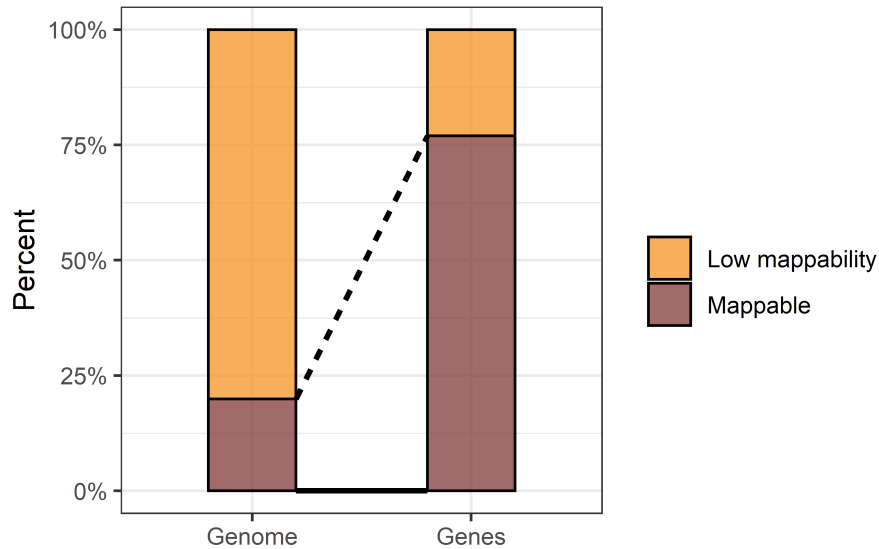


Figure 4: Percentages of mappable vs low mappability regions of the reference genome and number of genes. As is represented by the colored bars, 20% of the genome was found to be mappable, meanwhile 78% of the gene set was found to be mappable. The lines in the plot represent the relation between the percentage of mappable genome and the percentage of genes in unique regions. Even though only 20% of the genome is mappable, around 78% of the genes are located in those regions.

3.3 Copy number and variation calling preprocessing

To be able to have clean samples for the CNVs analysis, we did a CNV pre-processing following the pipeline developed by Trost et al. (2018b). Figure 5 showed that the obtained BAM files had an average coverage of 26.5, ranging from 15 to 40. One of the BAM files, obtained from HL0623, presented a coverage under 5, which is significantly low to perform the CNV calling. Thus, it was discarded from the analysis set, leaving 5 HL and 5 LL samples. We did not find association between the BAM files coverage over 15 (Supplementary Fig. S4) and the number of CNVs called.

These results are not spurious or due to sequencing errors because we checked the quality of the fasta files to assess whether adapter content was present and needed to be removed, and also to set the base quality filtering threshold with Trimmomatic (Supplementary Figure S1). After setting the base quality at a minimum of 20, the overall quality of the reads improved (Supplementary Figure S1). The reads highlighted in yellow in Supplementary Figure S1B are from HL0627, the individual that had lower sequencing depth than the rest.

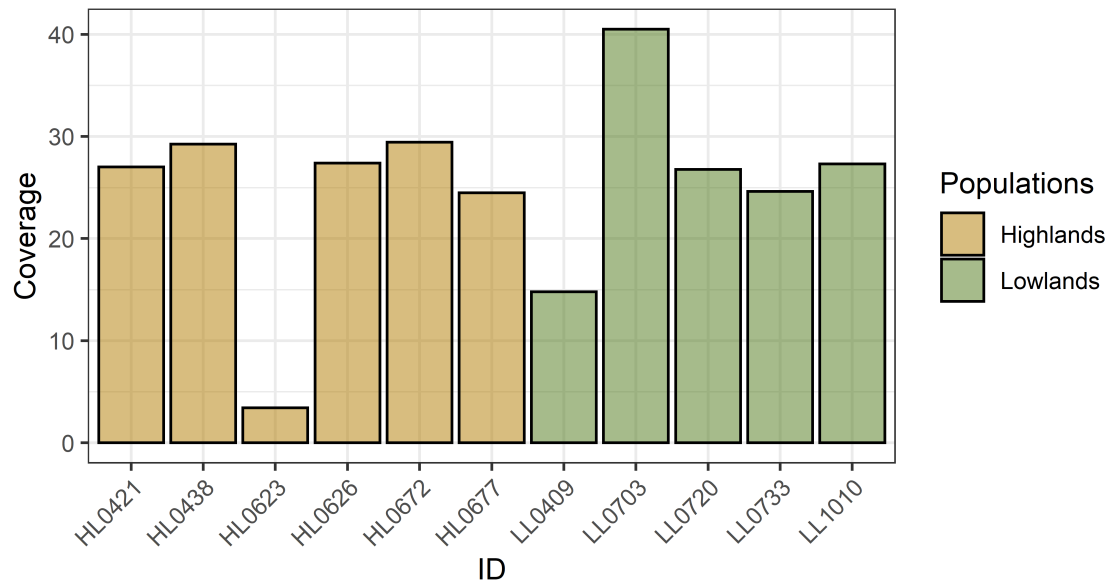


Figure 5: Coverage per BAM file is represented. HL0623 presents the lower coverage, therefore is discarded from the final CNV calling set. The remaining individuals show an average coverage of 26.5.

3.4 Variant Calling

3.4.1 Base Quality Score Recalibration

Supplementary Figure S2 shows the difference in the quality score covariate of the recalibrated BAM files using the whole genome or chromosome one as training set. As can be seen, the quality difference between both sets is considerably smaller for one chromosome, while the computational time was reduced tenfold for all the individuals. Since the rest of parameters, such as the empirical quality score, showed no significant differences, using one chromosome to obtain the variants for the training set was selected as the best option to perform the BQSR.

3.4.2 Single nucleotide polymorphisms and Indels quantification

The quantification of the SNPs and Indels yielded a total of 42.4 million SNPs and 8.5 million Indels. Around 2 millions out of the total SNPs were multi-allelic. The total number of annotated effects was of 85.5 million, being around 45% located in intergenic regions, followed by a 30% of effects located in downstream and upstream regions (https://github.com/TheRocinante-lab/PabloAtienza/blob/main/Results/Snp_summary.html).

The Indels were less numerous, as expected. In total, 8.5 million Indels were processed, with a variant rate of 1 indel every 250. From this set, 18.4 million effects were annotated, having 4% a high impact, and 41.5% located in intergenic regions

(https://github.com/TheRocinante-lab/PabloAtienza/blob/main/Results/Indels_summary.html).

SNPs were used to perform a principal component analysis (PCA) and a heatmap of the relations between the individuals.

3.4.3 Population structure

We performed PCA to summarize the genetic variation in the maize dataset. The results revealed structure in the population (Fig. 6). The maize clusters analysis revealed that all of the clusters can be separated in the first two PC axes that accounted for 25% (PC1) and 13% (PC2) of the genetic variance (Fig. 6A). The number of lines per cluster were five in both HL and LL populations. Two genotypes (HL0677 and LL1010) were more genetically distant to each group.

3.5 CNV calling

3.5.1 Gene size modification

The analysis of gene size modification indicated that there was a decrease in the number of CNVs called, although it wasn't significant (Supplementary Fig. S3). The number of duplications remained unchanged, although sometimes one or more duplications were called CNVs by increasing the gene size. Regarding the number of deletions, the number was reduced in all samples, but not enough to be considered significant (Supplementary Fig S3B). Nevertheless, the quality of the deletions in the 1kb size gene set improved.

3.5.2 CNV genes count

For the sake of simplicity in the following sections, I refer as a CNVs to the genes that had been called as a CNV, despite the fact that they might not be in a CNV region *per se*. The summary of the CNVs called per individual is shown in the files found at our group's GitHub (<https://github.com/TheRocinante-lab/PabloAtienza/tree/main/Results>). These files have 39 thousand columns, representing the genes and each row corresponded to one individual. In case a gene had been called as a duplication, it was stored as 1 and if it had been called a deletion, it was stored as -1. Figure 7 represents the number of CNVs shared by each number of individuals. To assess the amount of CNVs present in at least one individual we calculated the accumulative CNV frequency. These calculations were made in both

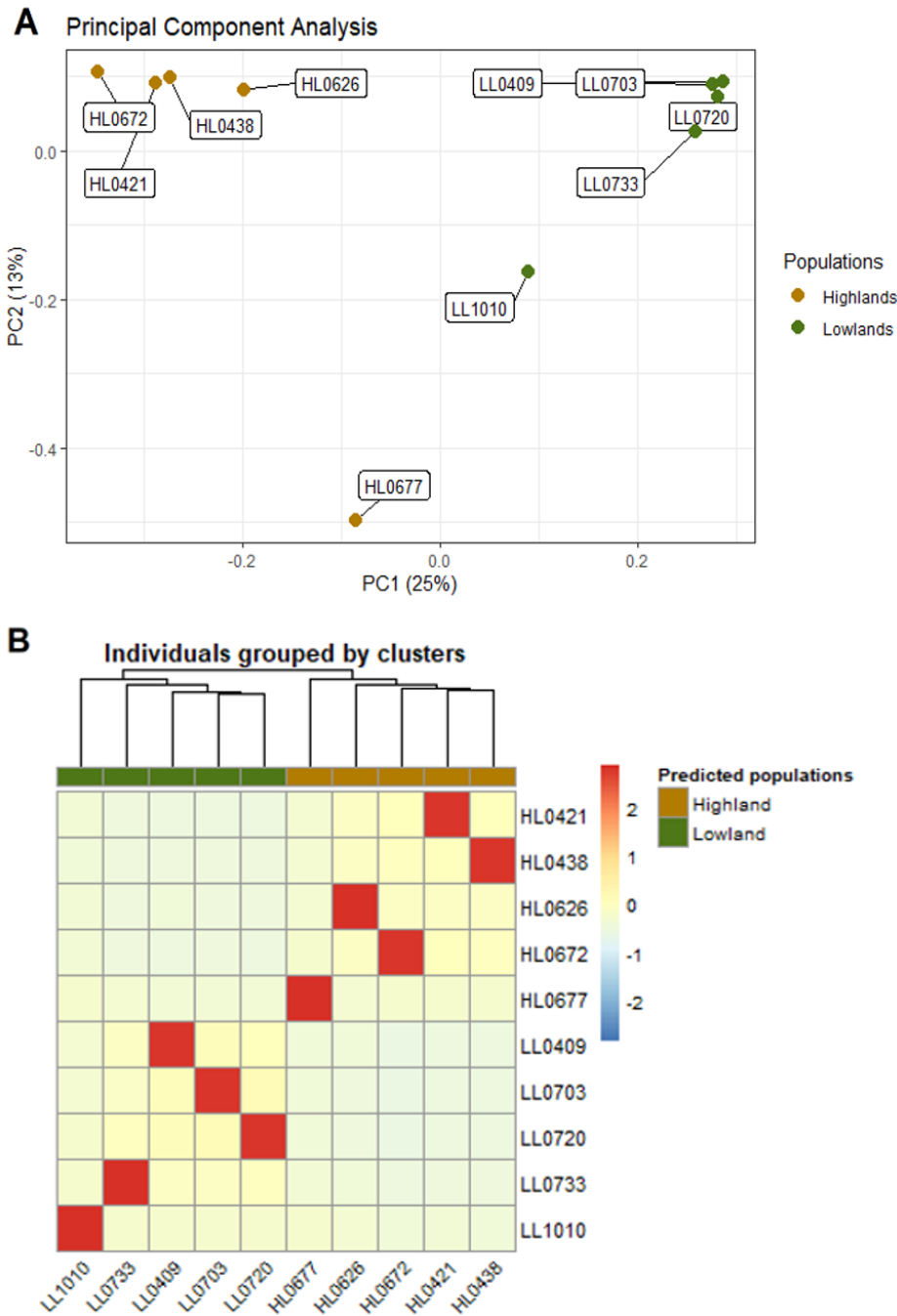


Figure 6: Summary of the analyses carried out with the SNPs data **A.** Comparison of the total number of CNVs in each population. **B.** Comparison of the number of duplications in each population. As can be seen, there are slightly more duplications in the HL, but not a significant difference. **C.** Comparison of the number of deletions in each population. As can be seen, there are slightly more deletions in one individual in the HL, but not a significant difference. For the rest of numbers of individuals, there are more deletions in the LL population than in the HL

populations and compared against each other, to evaluate if there is a significant difference in the number of CNVs called. (Fig. 7)

In general, the number of CNVs decreased when increasing the number of individuals that shared them (Fig 7A). As well, more deletions were called than duplications (Fig 7BC). These trends were constant for the set of CNVs that had less than 70% of overlap with low mappability regions, as well as when the threshold of overlap was raised to 85% (Supplementary Fig S5). However, when no mappability filter is applied, the number of CNVs that are present in all 5 individuals of a population increased significantly (Supplementary Fig. S6A). When the CNVs were separated in duplications and deletions, the duplication number regained the expected trend, while the deletions showed an increased number in all individuals of a population (Supplementary Fig. S6BC).

Finally, the set of CNVs that are present in 5 individuals for both populations was chosen. The genes present in each set were separated in deletions and duplications and compared against each other. The HL had a total of 71 CNVs, 17 duplications and 54 deletions. The LL population had a total of 67 CNVs, 10 duplications and 47 deletions. After the comparison between both gene sets, 35 deletions and 4 duplications were only found in the HL population; 22 deletions and 6 duplications were common for both populations and 32 deletions and 11 duplications were only found in the LL populations.

3.6 Functional enrichment analysis

Blast2GO detected no enriched GO terms for none of the different CNVs sets used for the analysis. The two blast analysis showed two different results. Firstly, BlastN did not yield significant results. Nevertheless, we found ten associations with proteins when using BlastX (one of the genes had no hits). We evaluated these genes in MaizeGDB to find the gene function, as well as, the synteny with other maize models or finding gene orthologs function in other plants genus (Supplementary Fig S2). Only one of the genes (Zm00001eb357130) resulted in having no correspondence with previous maize gene sets or orthologs. The remaining genes, had association with either version 4 of B73 reference genome or version 3. A total of 5 genes had orthologs at least in rice, and some of them in other plant organisms like Arabidopsis or Sorghum (Supplementary Table S2). Finally, in either the Blast X results or by the Maize GDB search, 8 out of 11 genes had an associated protein or domain, which allowed identifying possible functions or processes. Regarding the gene function, two main clusters can be observed: genes related to plant growth and genes related to stress response (Table 1). In the plant growth cluster, there are 4 genes, mainly

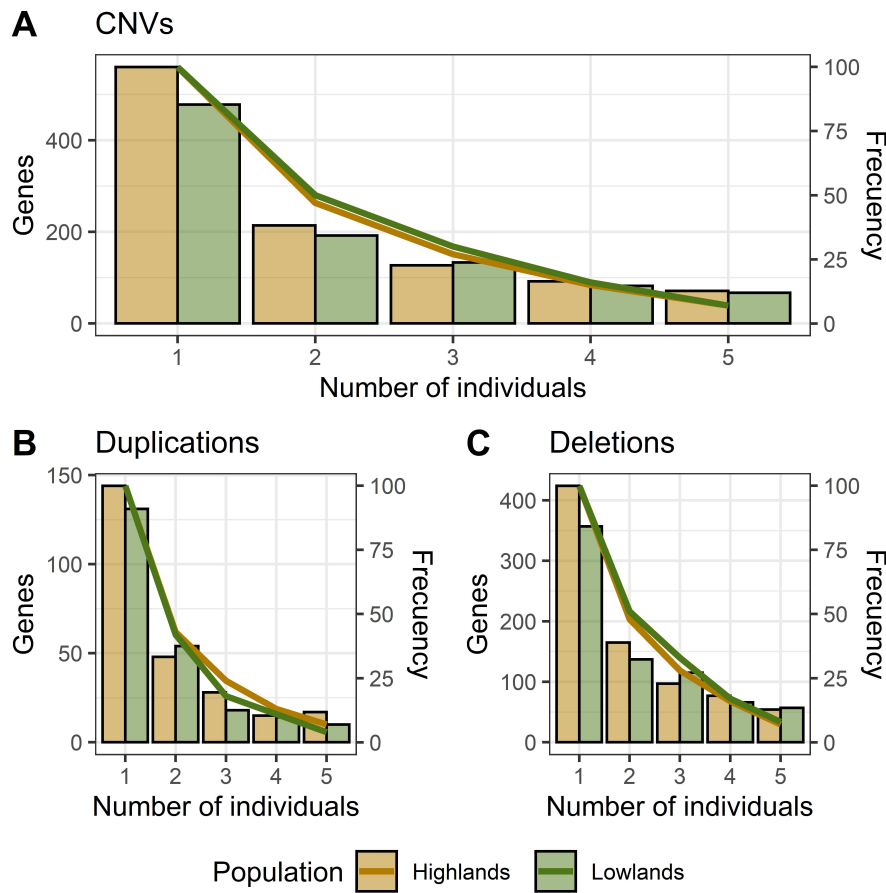


Figure 7: Number of copy number variants among populations. The bars represent the number of genes that are called as variant in a number of individuals. The lines represent the accumulated percentage of variants, going from the most restrictive (five individuals) to the variants in at least one individual. **A.** Comparison of the total number of CNVs in each population. **B.** Comparison of the number of duplications in each population. As can be seen, there are slightly more duplications in the HL, but not a significant difference. **C.** Comparison of the number of deletions in each population. As can be seen, there are slightly more deletions in one individual in the HL, but not a significant difference. For the rest of numbers of individuals, there are more deletions in the LL population than in the HL. *Quito explicación de en qué grupo hay más dups o dels de la caption?*

associated with plant growth and development. In regard to the second cluster, 5 out of 11 genes present response to stress related functions, mainly biotic and drought stress. In this study, the focus will be set upon three genes: Zm00001eb046710, which may have relation with the resistance against *Puccinia sorghi*; Zm00001eb167600, which has high similarity with the PME previously mentioned; and Zm00001eb421480, which has relations with the penthose-phosphate pathway (PPP).

The first gene, Zm00001eb046710 is located in chromosome 1 presented no clear area of the plant to be expressed. When blasted against the maize Database, the protein with the highest identity is the Splicing Factor 3B Subunit 1, which takes part in the process of pre-mRNA splicing as a component of the splicing factor SF3B complex (Papasaikas, Valcárcel, 2016). Nonetheless, it presents similarity

with biological stress response proteins in Maize, Rice and Arabidopsis. Some of the proteins it has similarity with are the RGA4, the putative rp3 protein and, a *Puccinia sorghi* resistance protein. The e-value of this protein and the gene was of 1e-28, being the fifth lowest of the associations.

The gene Zm00001eb167600, located in chromosome 4, had correspondence with Zm00001d048936, from the B73 reference genome 4.0 (B73v4). Due to the association, the characteristics from Zm00001d048936 will be applied to Zm00001eb167600. This gene has two synonyms: putative pectinesterase 63 and ZmGA1P, which is a pollen/male determinant of the *Ga1* locus. Thus, it presents relation with two loci: *Gametophyte factor (Ga1)* and *Pectin methylesterase (Pme3)*.

Finally, Zm00001eb421480 is located in chromosome 10 and mainly expressed in internodes 6 days after pollination, according to B73v4 annotation. This gene has a protein associated in version 4, a glucose-6-phosphate dehydrogenase (G6P-DH), which is involved in the pentoses-phosphate pathway (PPP). At the same time, when blasting the translation of the transcript generated by Zm00001eb421480, the main hits were of an Enolase 1-like protein. The enolase is an enzyme with catalytic activity, breaking the 2-phosphoglycerate into a water molecule and phosphoenolpyruvate in a reversible reaction. This is part of the glycolysis, taking part during the Krebs cycle (Gao et al., 2013).

B73v5 name	Expression on the plant	Associated function	Associated protein
Zm00001eb015360	Pericarp/Aleurone	Plant growth: Maturation and radical emergency	MADS box interactor-like (Integral membrane protein)
Zm00001eb046710	NO DATA	Stress response: biotic. Possible resistance to <i>Puccinia sorghi</i>	Splicing factor 3B subunit / RGA3 disease resistance protein
Zm00001eb049370	Pericarp/ Endosperm,	Uncharacterized protein	Possible pyruvate dehydrogenase kinase 2
Zm00001eb072540	Primary root and coleoptile	Plant growth: Seedling emergency. Seed imbibition stage (B73v3)	Transposable element
Zm00001eb144990	Germinating seed	Plant growth: Primary shoot system and leaf expansion	NO DATA
Zm00001eb167600	Low expression	Plant development: Modifications on cell wall. Unilateral Cross-Incompatibility	Pectinesterase
Zm00001eb168270	Internodes, pollen and pericarp	Stress response: saline and biotic	Ankyrin repeat protein SKIP35
Zm00001eb260240	Primary root and ear	Stress response: heat and biotic	GDP-Mannose transporter GONST1
Zm00001eb260250	Anthers, mature leaves and pericarp	Stress response: biotic. Pollen and leaf development. Fruit ripening	F-box domain containing protein
Zm00001eb421480	Internodes post pollination	Stress response: drought. Basal metabolism. Penthose Phosphate Pathway.	Enolase 1
Zm00001eb357130	NO DATA	-	NO DATA

Table 1: Summary table with the location in the plant, associated function and protein of the 11 genes called as duplication only in the HL population.

Chapter 4. Discussion

In this study, one of the main goals was to test whether or not, CNVs found in two populations geographically distinct, were associated with the phenotypic adaptation to their climate conditions. When we assessed if the populations were genetically clustered according to their environments, we observed that LL1010 and HL0677 were slightly separated from each of their own cluster (Fig. 3 and Fig. 6). This could be explained by the fact that those individuals were found to be geographically located in the edge of both environments (Fig. 1). This also can be observed in Figure 6B where the individual LL0733, the closest variety to LL1010 (Fig. 3), indicating that geographic location (environment) plays an important role on the genetic makeup of these populations.

Our mappability results were in concordance with the percentages obtained by Ramos-Madriral et al. (2016), which indicates that the pipeline used in this study, to find unique sequences within the reference, yielded as expected. After the mappability, the quantification of duplications and deletions from this study were in concordance with other studies (Genova et al., 2018; Seol et al., 2019). In our study, and in Genova et al., 2018 and Seol et al., 2019 deletions were found more present than duplications (Fig. 7BC). This is due to the fact that deletions are easier to detect than duplications using the read depth method applied in this study (Abyzov et al., 2011). In addition, the fact that we removed duplicated reads with low mapping quality makes it easier to find deletions than duplications. Nevertheless, this approach could yield in more false positive callings in the deletions due to the reads eliminated that generate small read count areas (Zhao et al., 2013). We minimized the false positive deletions callings by applying our specific mappability filtering.

We expect to have less number of common CNVs as the number of individuals under study increase (Figure 7), since the higher number of individuals the lower the probability to find the same CNV. The main reason behind this phenomenon, is the fast increase in number of deletions in 5 individuals (Supplementary Fig. S6C). This could be due to the above explained filtering of the BAM files. When applying the mappability filters, which is common for all the individuals, the anomaly. Since it was only found in the deletions, and not in the duplications, this fact supports the aforementioned explanation.

To assess if the differences in phenotype between both populations were caused by the presence of CNVs, we selected the CNVs that were present in all individuals for the comparison study between HL and LL populations (Fig. 7). The six groups mentioned in the results were analyzed for enrichment. Our results showed that no

functions were found to be enriched, which could have been due to the low size of the gene sets. HL duplications were selected to include in further analysis because it has been demonstrated previously its role in adaptation to environments and disease resistance (Hu et al., 2018; Qian, Zhang, 2014; Conrad, Antonarakis, 2007; Panchy et al., 2016). We found associations between the function of the genes duplicated only in HL maize and the different adaptation of traits that were in HL maize. As it was mentioned in the results (Table 1), genes could be classified based on resistance to stress (50%), plant growth and development (40%) and a 10% that had a protein result, but there is no function yet associated in maize to them. Here, we focus on three genes that have been shown to have an important function in plant organisms and maize (Collins et al., 1999; Zhang et al., 2018; Hu et al., 2011; Garzon et al., 2020). Nevertheless, there are other groups of genes, such as exclusive deletions in HL and in LL (Table 1) that may have hidden candidate genes, promoter or regulatory regions that also play a role in the phenotypic differences among populations.

The gene Zm00001eb046710 has been found to be associated with the *Rp1* locus in chromosome 10 of maize, which has been shown to be associated with resistance to *P. sorghi* disease (Chavan et al., 2015). In addition, *Rp1* locus is located in a high recombinogenic region due to the presence of transposons (Collins et al., 1999; Ramakrishna et al., 2002a). This nature of the locus could have led to a duplication event outside its regions associated with Leucine Rich Repeats (Ramakrishna et al., 2002b). These repeats have been demonstrated to have a ubiquitous role in the resistance against pathogens (McHale et al., 2006; Collins et al., 1998).

The second gene, Zm00001eb167600, was already found by Zhang et al. (2018) to be directly related with the unilateral cross-incompatibility (UCI) system in maize. The Pectinesterase associated with the gene is an enzyme found in plants that catalyzes the de-esterification of the cell wall pectines (Mohnen, 2008; Jiang et al., 2001) through the modification of the methyl esterification patterns (Cosgrove, 2001; Micheli, 2001; Catoire et al., 1998). This protein is known to confer the male function to maize, having different compatible alleles (Wakeley et al., 1998). It has not been assessed how the duplication of this gene would affect the UCI system, but several Mexican HL landraces are known to carry the alleles that confer the sporophytic and gametophytic incompatibility (Parra et al., 2012).

Finally, the gene Zm00001eb421480 codes the protein Glucose-6-Phosphate Dehydrogenase (G6P-DH) which has been observed to develop higher activity in tolerant to long-term chilling in maize (Garzon et al., 2020) and other plants (Sarkar et al., 2009). In addition to cold tolerance, G6P-DH has been also characterized as being a differentially regulated protein (DRP) and a key rate-limiting enzyme for maize plants with better growth under saline stress (Luo et al., 2018). In regard to functions

of the enolase related to the Mexican HL maize, it has been assessed that it plays a role in drought stress (Hu et al., 2011; Riccardi et al., 1998), starch filling of the seeds (Cao et al., 2019) and anaerobiosis situations (Lal et al., 1998).

Chapter 5. Conclusion

We found 11 genes, half of them were related with response to stress, mainly drought and biotic stress, and the other half to plant growth and development. Three of them have shown to play a central role in the regulation and adaptation to environmental stresses, as well as in resistance to fungus, demonstrating the possible role of CNVs in the functional adaptability to specific environments.

We found millions of SNPs variants between the two populations. In a future study, we will focus on detecting SNPs variants within the CNVs genes found in this study. The aim is to find markers that could be used in plant breeding programs for marker assisted selection.

This study is the first step towards finding specific genes associated with HL and LL maizes. A validation analysis of these findings, by performing Comparative Genomic Hybridization-array or by quantitative PCR analysis needs to be taking in place to verify the presence of these genes, and how the duplications affect the level of these genes transcription.

Copy number variations are still a developing field in genomics, that faces challenges in organisms with highly repetitive genomes, such as maize. By combining different approaches, like joining a read depth method and the mappability, we have been able to reduce most of the false positive CNV callings. The approach taken for CNV calling in this study has proven to be effective even for small working samples and offers a new perspective in the field.

Chapter 6. Bibliography

- Abyzov Alexej, Urban Alexander E., Snyder Michael, Gerstein Mark.* CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing // Genome Research. JUN 2011. 21, 6. 974–984.
- Abyzov Alexej, Urban Alexander E., Snyder Michael, Gerstein Mark.* CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing // Genome research. 2011. 21, 6. 974–984.
- Andrews Simon, Krueger Felix, Segonds-Pichon Anne, Biggins Laura, Krueger Christel, Wingett Steven.* FastQC. Babraham, UK: Babraham Institute, I 2012.
- Auwer GA Van der, D O'Connor Brian.* Genomics in the Cloud: Using Docker, Gatk, and Wdl in Terra. CA 95472 Sebastopol, Canada. 2020.
- Bai Zetao, Chen Jinfeng, Liao Yi, Wang Meijiao, Liu Rong, Ge Song, Wing Rod A., Chen Mingsheng.* The impact and origin of copy number variations in the Oryza species // BMC Genomics. MAR 29 2016. 17.
- Blum Matthias, Chang Hsin-Yu, Chuguransky Sara, Grego et al.* The InterPro protein families and domains database: 20 years on // Nucleic Acids Research. JAN 8 2021. 49, D1. D344–D354.
- Bolger Anthony M., Lohse Marc, Usadel Bjoern.* Trimmomatic: a flexible trimmer for Illumina sequence data // Bioinformatics. AUG 1 2014. 30, 15. 2114–2120.
- Boratyn Grzegorz M., Camacho Christiam, Cooper Peter S., Coulouris George, Fong Amelia, Ma Ning, Madden Thomas L., Matten Wayne T., McGinnis Scott D., Merezhuk Yuri, Raytselis Yan, Sayers Eric W., Tao Tao, Ye Jian, Zaretskaya Irena.* BLAST: a more efficient report with usability improvements // Nucleic Acids Research. JUL 2013. 41, W1. W29–W33.
- Bradbury Peter J., Zhang Zhiwu, Kroon Dallas E., Casstevens Terry M., Ramdoss Yogesh, Buckler Edward S.* TASSEL: software for association mapping of complex traits in diverse samples // Bioinformatics. 06 2007. 23, 19. 2633–2635.
- Brohammer Alex B, Kono Thomas JY, Springer Nathan M, McGaugh Suzanne E, Hirsch Candice N.* The limited role of differential fractionation in genome content

- variation and function in maize (*Zea mays* L.) inbred lines // The Plant Journal. 2018. 93, 1. 131–141.
- Cao Hanwei, Zhou Yuwei, Chang Ying, Zhang Xiuyan, Li Cui, Ren Dongtao.* Comparative phosphoproteomic analysis of developing maize seeds suggests a pivotal role for enolase in promoting starch synthesis // Plant Science. DEC 2019. 289.
- Catoire L, Pierron M, Morvan C, Penhoat CH du, Goldberg R.* Investigation of the action patterns of pectinmethylesterase isoforms through kinetic analyses and NMR spectroscopy - Implications in cell wall expansion // Journal of Biological Chemistry. DEC 11 1998. 273, 50. 33150–33156.
- Chavan Suchitra, Gray Judy, Smith Shavannor M.* Diversity and evolution of Rp1 rust resistance genes in four maize lines // Theoretical and Applied Genetics. MAY 2015. 128, 5. 985–998.
- Chen Wei-Kang, Swartz Joshua D, Rush Laura J, Alvarez Carlos E.* Mapping DNA structural variation in dogs // Genome research. 2009. 19, 3. 500–509.
- Chia Jer-Ming, Song Chi, Bradbury Peter J, Costich Denise, De Leon Natalia, Doebley John, Elshire Robert J, Gaut Brandon, Geller Laura, Glaubitz Jeffrey C, others .* Maize HapMap2 identifies extant variation from a genome in flux // Nature genetics. 2012. 44, 7. 803–807.
- 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. 2012. 6, 2. 80–92.
- Clancy Suzanne.* DNA damage & repair: mechanisms for maintaining DNA integrity // Nature Education. 2008. 1, 1. 103.
- Collins N, Drake J, Ayliffe M, Sun Q, Ellis J, Hulbert S, Pryor T.* Molecular characterization of the maize Rp1-D rust resistance haplotype and its mutants // Plant Cell. JUL 1999. 11, 7. 1365–1376.
- Collins NC, Webb CA, Seah S, Ellis JG, Hulbert SH, Pryor A.* The isolation and mapping of disease resistance gene analogs in maize // Molecular plant-microbe interactions. 1998. 11, 10. 968–978.
- Conrad Bernard, Antonarakis Stylianos E.* Gene duplication: A drive for phenotypic diversity and cause of human disease // Annual Review of Genomics and Human Genetics. Annual Review of Genomics and Human Genetics. 2007. 8. 17–35.

- Conrad Donald F, Andrews T Daniel, Carter Nigel P, Hurles Matthew E, Pritchard Jonathan K.* A high-resolution survey of deletion polymorphism in the human genome // *Nature genetics*. 2006. 38, 1. 75–81.
- Cosgrove DJ.* Wall structure and wall loosening. A look backwards and forwards // *Plant Physiology*. JAN 2001. 125, 1. 131–134.
- Craddock Nick, Hurles Matthew E, Cardin Niall, Pearson Richard D, Plagnol Vincent, Robson Samuel, Vukcevic Damjan, Barnes Chris, Conrad Donald F, Giannoulatou Eleni, others .* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls // *Nature*. 2010. 464, 7289. 713.
- Dolatabadian A. Edwards D. Batley J. Patel DA.* Copy number variation and disease resistance in plants // *Theor Appl Genet*. 2017. 130. 2479–2490.
- Dolatabadian Aria, Patel Dhvani Apurva, Edwards David, Batley Jacqueline.* Copy number variation and disease resistance in plants // *Theoretical and Applied Genetics*. 2017. 130, 12. 2479–2490.
- Doyle Jeff J, Doyle Jane L.* A rapid DNA isolation procedure for small quantities of fresh leaf tissue. 1987.
- Eagles HA, Lothrop JE.* Highland Maize from Central Mexico - its origin, characteristics, and use in breeding programs // *CRrop Science*. JAN-FEB 1994. 34, 1. 11–19.
- Gao Jie, Zhao Rongrong, Xue Yan, Niu Zuoxing, Cui Kai, Yu Fachang, Zhang Bo, Li Sheng.* Role of enolase-1 in response to hypoxia in breast cancer: Exploring the mechanisms of action // *Oncology Reports*. APR 2013. 29, 4. 1322–1332.
- Garzon Catalina Duran, Lequart Michelle, Rautengarten Carsten, Bassard Solene, Sellier-Richard Helene, Baldet Pierre, Heazlewood Joshua L., Gibon Yves, Domon et al.* Regulation of carbon metabolism in two maize sister lines contrasted for chilling tolerance // *Journal of Experimental Bontany*. JAN 1 2020. 71, 1. 356–369.
- Genova F., Longeri M., Lyons L. A., Bagnato A., Strillacci M. G., Consortium 99Lives.* First genome-wide CNV mapping in FELIS CATUS using next generation sequencing data // *Bmc Genomics*. DEC 10 2018. 19.
- Graubert Timothy A, Cahan Patrick, Edwin Deepa, Selzer Rebecca R, Richmond Todd A, Eis Peggy S, Shannon William D, Li Xia, McLeod Howard L, Cheverud James M, others .* A high-resolution map of segmental DNA copy number variation in the mouse genome // *PLoS Genet*. 2007. 3, 1. e3.

- Guryev Victor, Saar Kathrin, Adamovic Tatjana, Verheul Mark, Van Heesch Sebastiaan AAC, Cook Stuart, Pravenec Michal, Aitman Timothy, Jacob Howard, Shull James D, others* . Distribution and functional impact of DNA copy number variation in the rat // *Nature genetics*. 2008. 40, 5. 538–545.
- Götz Stefan, García-Gómez Juan Miguel, Terol Javier, Williams Tim D., Nagaraj Shivashankar H., Nueda María José, Robles Montserrat, Talón Manuel, Dopazo Joaquín, Conesa Ana*. High-throughput functional annotation and data mining with the Blast2GO suite // *Nucleic Acids Research*. 04 2008. 36, 10. 3420–3435.
- Haberer Georg, Kamal Nadia, Bauer Eva, Gundlach Heidrun, Fischer Iris, Seidel Michael A, Spannagl Manuel, Marcon Caroline, Ruban Alevtina, Urbany Claude, others* . European maize genomes highlight intraspecies variation in repeat and gene content // *Nature genetics*. 2020. 52, 9. 950–957.
- Hirsch Candice N, Foerster Jillian M, Johnson James M, Sekhon Rajandeep S, Muttoni German, Vaillancourt Brienne, Peñagaricano Francisco, Lindquist Erika, Pedraza Mary Ann, Barry Kerrie, others* . Insights into the maize pan-genome and pan-transcriptome // *The Plant Cell*. 2014. 26, 1. 121–135.
- Hu Xiuli, Lu Minghui, Li Chaohao, Liu Tianxue, Wang Wei, Wu Jianyu, Tai Fujun, Li Xiao, Zhang Jie*. Differential expression of proteins in maize roots in response to abscisic acid and drought // *Acta Physiologiae Plantarum*. NOV 2011. 33, 6. 2437–2446.
- Hu Ying, Ren Jie, Peng Zhao, Umana Arnoldo A., Le Ha, Danilova Tatiana, Fu Junjie, Wang Haiyan, Robertson Alison, Hulbert Scot H., White Frank F., Liu Sanzhen*. Analysis of Extreme Phenotype Bulk Copy Number Variation (XP-CNV) Identified the Association of *rp1* with Resistance to Goss's Wilt of Maize // *Frontiers in Plant Science*. FEB 9 2018. 9.
- Jiang CM, Wu MC, Chang WH, Chang HM*. Change in particle size of pectin reacted with pectinesterase isozymes from pea (*Pisum sativum* L.) sprout // *Journal of Agricultural and Food Chemistry*. SEP 2001. 49, 9. 4383–4387.
- Kato Akio, Lamb Jonathan C, Birchler James A*. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize // *Proceedings of the National Academy of Sciences*. 2004. 101, 37. 13554–13559.
- Knaus Brian J, Grünwald Niklaus J*. Inferring variation in copy number using high throughput sequencing data in R // *Frontiers in genetics*. 2018. 9. 123.

- Lal SK, Lee CF, Sachs MM.* Differential regulation of enolase during anaerobiosis in maize // *Plant Physiology*. DEC 1998. 118, 4. 1285–1293.
- Lee Arthur S, Gutiérrez-Arcelus María, Perry George H, Vallender Eric J, Johnson Welkin E, Miller Gregory M, Korbel Jan O, Lee Charles.* Analysis of copy number variation in the rhesus macaque genome identifies candidate loci for evolutionary and human disease studies // *Human molecular genetics*. 2008. 17, 8. 1127–1136.
- Li Heng, Durbin Richard.* Fast and accurate short read alignment with Burrows-Wheeler transform // *Bioinformatics*. JUL 15 2009. 25, 14. 1754–1760.
- Li Heng, Handsaker Bob, Wysoker Alec, Fennell Tim, Ruan Jue, Homer Nils, Marth Gabor, Abecasis Goncalo, Durbin Richard, Proc 1000 Genome Project Data.* The Sequence Alignment/Map format and SAMtools // *Bioinformatics*. AUG 15 2009. 25, 16. 2078–2079.
- Lu Fei, Romay Maria C, Glaubitz Jeffrey C, Bradbury Peter J, Elshire Robert J, Wang Tianyu, Li Yu, Li Yongxiang, Semagn Kassa, Zhang Xuecai, others .* High-resolution genetic mapping of maize pan-genome sequence anchors // *Nature Communications*. 2015. 6, 1. 1–8.
- Luo Meijie, Zhao Yanxin, Wang Yuandong, Shi Zi, Zhang Panpan, Zhang Yunxia, Song Wei, Zhao Jiuran.* Comparative Proteomics of Contrasting Maize, Genotypes Provides Insights into Salt-Stress Tolerance Mechanisms // *Journal of Proteome Research*. JAN 2018. 17, 1. 141–153.
- Maron Lyza G, Guimarães Claudia T, Kirst Matias, Albert Patrice S, Birchler James A, Bradbury Peter J, Buckler Edward S, Coluccio Alison E, Danilova Tatiana V, Kudrna David, others .* Aluminum tolerance in maize is associated with higher MATE1 gene copy number // *Proceedings of the National Academy of Sciences*. 2013. 110, 13. 5241–5246.
- Mayrhofer Markus, Viklund Björn, Isaksson Anders.* Rawcopy: Improved copy number analysis with Affymetrix arrays // *Scientific reports*. 2016. 6, 1. 1–11.
- McClintock Barbara, Kato Takeo Angel, Blumenschein Almiro.* Chromosome constitution of races of maize. its significance in the interpretation of relationships between races and varieties in the americas. 31. 1981.
- McHale Leah, Tan Xiaoping, Koehl Patrice, Michelmore Richard W.* Plant NBS-LRR proteins: adaptable guards // *Genome Biology*. 2006. 7, 4.

- Micheli F.* Pectin methylesterases: cell wall enzymes with important roles in plant physiology // Trends in Plant Science. SEP 2001. 6, 9. 414–419.
- Mohnen Debra.* Pectin structure and biosynthesis // Current Opinion in Plant Biology. JUN 2008. 11, 3. 266–277.
- Panchy Nicholas, Lehti-Shiu Melissa, Shiu Shin-Han.* Evolution of Gene Duplication in Plants // Plant physiology. AUG 2016. 171, 4. 2294–2316.
- Papasaikas Panagiotis, Valcárcel Juan.* The spliceosome: the ultimate RNA chaperone and sculptor // Trends in biochemical sciences. 2016. 41, 1. 33–45.
- Parra José, Ruiz Corral José, Cruz Lino De la, Padilla-Garcia Jose-Miguel, Rivera Moises, González José.* Incompatibilidad gametofítica en las razas mexicanas de maíz // Revista Mexicana de Ciencias Agrícolas. 06 2012. 3. 525–537.
- Portwood John L. II, Woodhouse Margaret R., Cannon Ethalinda K., Gardiner Jack M., Harper Lisa C., Schaeffer et al.* MaizeGDB 2018: the maize multi-genome genetics and genomics database // Nucleic Acids Research. JAN 8 2019. 47, D1. D1146–D1154.
- Qian Wenfeng, Zhang Jianzhi.* Genomic evidence for adaptation by gene duplication // Genome Research. AUG 2014. 24, 8. 1356–1362.
- Quinlan Aaron R., Hall Ira M.* BEDTools: a flexible suite of utilities for comparing genomic features // Bioinformatics. MAR 15 2010. 26, 6. 841–842.
- R Core Team .* R: A Language and Environment for Statistical Computing. Vienna, Austria, 2020.
- Ramakrishna W, Emberton J, Ogden M, SanMiguel P, Bennetzen JL.* Structural analysis of the maize Rp1 complex reveals numerous sites and unexpected mechanisms of local rearrangement // Plant Cell. DEC 2002a. 14, 12. 3213–3223.
- Ramakrishna W, Emberton J, Ogden M, SanMiguel P, Bennetzen JL.* Structural analysis of the maize Rp1 complex reveals numerous sites and unexpected mechanisms of local rearrangement // Plant Cell. DEC 2002b. 14, 12. 3213–3223.
- Ramos-Madrigal Jazmin, Smith Bruce D., Moreno-Mayar J. Victor, Gopalakrishnan Shyam, Ross-Ibarra Jeffrey, Gilbert M. Thomas P., Wales Nathan.* Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication // Current Biology. DEC 5 2016. 26, 23. 3195–3201.

- Riccardi F, Gazeau P, Vienne D de, Zivy M.* Protein changes in response to progressive water deficit in maize - Quantitative variation and polypeptide identification // *Plant Physiology*. AUG 1998. 117, 4. 1253–1263.
- Sarkar Dipayan, Bhowmik Prasanta C., Young-In-Kwon , Shetty Kalidas.* Cold Acclimation Responses of Three Cool-season Turfgrasses and the Role of Proline-associated Pentose Phosphate Pathway // *Journal of the American Society for Horticultural Science*. MAR 2009. 134, 2. 210–220.
- Sawers Ruairidh, Sanchez-Leon Nidia.* Origins of maize: a further paradox resolved // *Frontiers in Genetics*. 2011. 2. 53.
- Saxena Rachit K, Edwards David, Varshney Rajeev K.* Structural variations in plant genomes // *Briefings in functional genomics*. 2014. 13, 4. 296–307.
- Schiessl Sarah, Huettel Bruno, Kuehn Diana, Reinhardt Richard, Snowden Rod.* Post-polyploidisation morphotype diversification associates with gene copy number variation // *Scientific reports*. 2017. 7, 1. 1–18.
- Schnable P. S.* The B73 maize genome: Complexity, diversity, and dynamics (November, pg 1112, 2009) // *Science*. AUG 31 2012. 337, 6098. 1040.
- Schnable Patrick S, Ware Doreen, Fulton Robert S, Stein Joshua C, Wei Fusheng, Pasternak Shiran, Liang Chengzhi, Zhang Jianwei, Fulton Lucinda, Graves Tina A, others .* The B73 maize genome: complexity, diversity, and dynamics // *science*. 2009. 326, 5956. 1112–1115.
- Schweinberger Martin.* Displaying Geo-Spatial Data with R. Brisbane, 2021. 2021.03.21. <https://slcladal.github.io/maps.html>.
- Seol Donghyeok, Ko Byung June, Kim Bongsang, Chai Han-Ha, Lim Dajeong, Kim Heebal.* Identification of Copy Number Variation in Domestic Chicken Using Whole-Genome Sequencing Reveals Evidence of Selection in the Genome // *Animals*. OCT 2019. 9, 10.
- Swanson-Wagner Ruth A, Eichten Steven R, Kumari Sunita, Tiffin Peter, Stein Joshua C, Ware Doreen, Springer Nathan M.* Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor // *Genome research*. 2010. 20, 12. 1689–1699.
- Thorvaldsdottir Helga, Robinson James T., Mesirov Jill P.* Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration // *Briefings in Bioinformatics*. MAR 2013. 14, 2, SI. 178–192.

- Trost Brett, Walker Susan, Wang Zhuozhi, Thiruvahindrapuram Bhooma, MacDonald Jeffrey R., Sung Wilson W. L., Pereira Sergio L., Whitney Joe, Chan Ada J. S., Pellecchia Giovanna, Reuter Miriam S., Lok Si, Yuen Ryan K. C., Marshall Christian R., Merico Daniele, Scherer Stephen W.* A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data // American Journal of Human Genetics. JAN 4 2018a. 102, 1. 142–155.
- Trost Brett, Walker Susan, Wang Zhuozhi et al.* A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data // American Journal of Human Genetics. JAN 4 2018b. 102, 1. 142–155.
- Wakeley PR, Rogers HJ, Rozycka M, Greenland AJ, Hussey PJ.* A maize pectin methylesterase-like gene, ZmC5, specifically expressed in pollen // Plant Molecular Biology . MAY 1998. 37, 1. 187–192.
- Wang Li, Beissinger Timothy M., Lorant Anne, Ross-Ibarra Claudia, Ross-Ibarra Jeffrey, Hufford Matthew B.* The interplay of demography and selection during maize domestication and expansion // Genome Biology. NOV 13 2017. 18.
- Wang Qinghua, Dooner Hugo K.* Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus // Proceedings of the National Academy of Sciences. 2006. 103, 47. 17644–17649.
- Wimalanathan Kokulapalan, Friedberg Iddo, Andorf Carson M., Lawrence-Dill Carolyn J.* Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER) // Plant Direct. 2018. 2, 4.
- Wong Kendy K, deLeeuw Ronald J, Dosanjh Nirpjit S, Kimm Lindsey R, Cheng Ze, Horsman Douglas E, MacAulay Calum, Ng Raymond T, Brown Carolyn J, Eichler Evan E, others .* A comprehensive analysis of common copy-number variations in the human genome // The American Journal of Human Genetics. 2007. 80, 1. 91–104.
- Zhang Zhaogui, Zhang Baocai, Chen Zhibin, Zhang Dongmei, Zhang Huairan, Wang Hang, Zhang Yu’e, Cai Darun, Liu Juan, Xiao Senlin, others .* A PECTIN METHYLESTERASE gene at the maize Gal locus confers male function in unilateral cross-incompatibility // Nature communications. 2018. 9, 1. 1–9.

Zhao Min, Wang Qingguo, Wang Quan, Jia Peilin, Zhao Zhongming. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives // BMC Bioinformatics. SEP 13 2013. 14, 11.

A Annex

Table S1: Information summary of the plant material used in the research

ID	Population	SRA_Accession	Latitude	Longitude	Altitude	Location	Depth	Size [Gb]
HL0421	Highlands	SRS1144295	19,850	-97,983	2250	Puebla, Mexico	28,5	50
HL0438	Highlands	SRS1144411	19,000	-97,383	2600	Puebla, Mexico	39,0	56
HL0623	Highlands	SRS1144729	20,033	-103,683	2520	Jalisco, Mexico	7,5	9,7
HL0626	Highlands	SRS1144722	19,883	-97,583	2260	Puebla, Mexico	27,0	43
HL0672	Highlands	SRS1143272	19,683	-99,133	2256	Mexico, Mexico	30,5	47,2
HL0677	Highlands	SRS1144723	21,367	-102,850	1951	Zacatecas, Mexico	26,5	42
LL0409	Lowlands	SRS1144294	15,433	-92,900	107	Chiapas, Mexico	28,0	49,8
LL0703	Lowlands	SRS1144297	20,833	-88,517	30	Yucatan, Mexico	63,5	99,8
LL0720	Lowlands	SRS1143275	15,467	-88,850	39	Guatemala	27,5	45
LL0733	Lowlands	SRS1144724	16,567	-94,617	107	Oaxaca, Mexico	28,5	45,4
LL1010	Lowlands	SRS1144725	16,850	-99,067	201	La Concordia, Guerrero	28,5	45,5

Table S2: Information of the functionally annotated genes. The correspondences in both version 4 and version 3 of B73 genes models are present, as well as the orthologs of each gene, the location of the protein in the cell and in the plant. The functions and name of the proteins related to the genes are shown as well.

B73v5 name	B73v4 correspondence	B73v3 correspondence	Othologs	Expression on the plant
Zm00001eb015360	Zm00001d028965	GRMZM2G145065	Sorghum, Foxtail millet, Rice, Arabidopsis	Pericarp/Aleurone
Zm00001eb046710	-	GRMZM2G090649	Rice	<i>NO DATA</i>
Zm00001eb049370	Zm00001d033115	-	-	Pericarp/ Endosperm, coleoptile and primary root
Zm00001eb072540	Zm00001d002506	GRMZM2G326489	-	Primary root and coleoptile
Zm00001eb144990	-	GRMZM2G064666	Rice	Germinating seed
Zm00001eb167600	Zm00001d048936	-	-	Polen
Zm00001eb168270	Zm00001d049019	GRMZM2G017624	Rice, Sorghum, Foxtail, Arabidopsis	Internodes, pollen and pericarp
Zm00001eb260240	Zm00001d035015	-	Rice, Sorghum, Foxtail, Arabidopsis	Primary root and ear primordium
Zm00001eb260250	Zm00001d035016	GRMZM5G861268	-	Anthers, mature leaves and pericarp
Zm00001eb421480	Zm00001d025288	GRMZM2G133721	-	Internodes post pollination
Zm00001eb357130	-	-	-	-
Location on the cell	Associated function		Associated protein	
	Plant growth: Maturation and radical emergency		MADS box interactor-like (Integral membrane protein)	
-	Stress response: biotic. Possible resistance to Puccinia sorghi		Splicing factor 3B subunit / RGA3 disease resistance protein	
-	Uncharacterized protein		Possible pyruvate dehydrogenase kinase 2	
-	Plant growth: Seedling emergency. Seed imbibition stage (B73v3)		Transposable element	
-	Plant growth: Primary shoot system and leaf expansion		<i>NO DATA</i>	
Cell Wall	Modifications on cell wall. Unilateral Cross-Incompatibility		Pectinesterase	
-	Stress response: saline and biotic		Ankyrin repeat protein SKIP35	
Golgi apparatus	Stress response: heat and biotic		GDP-Mannose transporter GONST1	
-	Stress response: biotic. Pollen and leaf development. Fruit ripening		F-box domain containing protein	
-	Stress response: drought. Basal metabolism. Penthose Phosphate Pathway.		Enolase 1	
-	-		-	

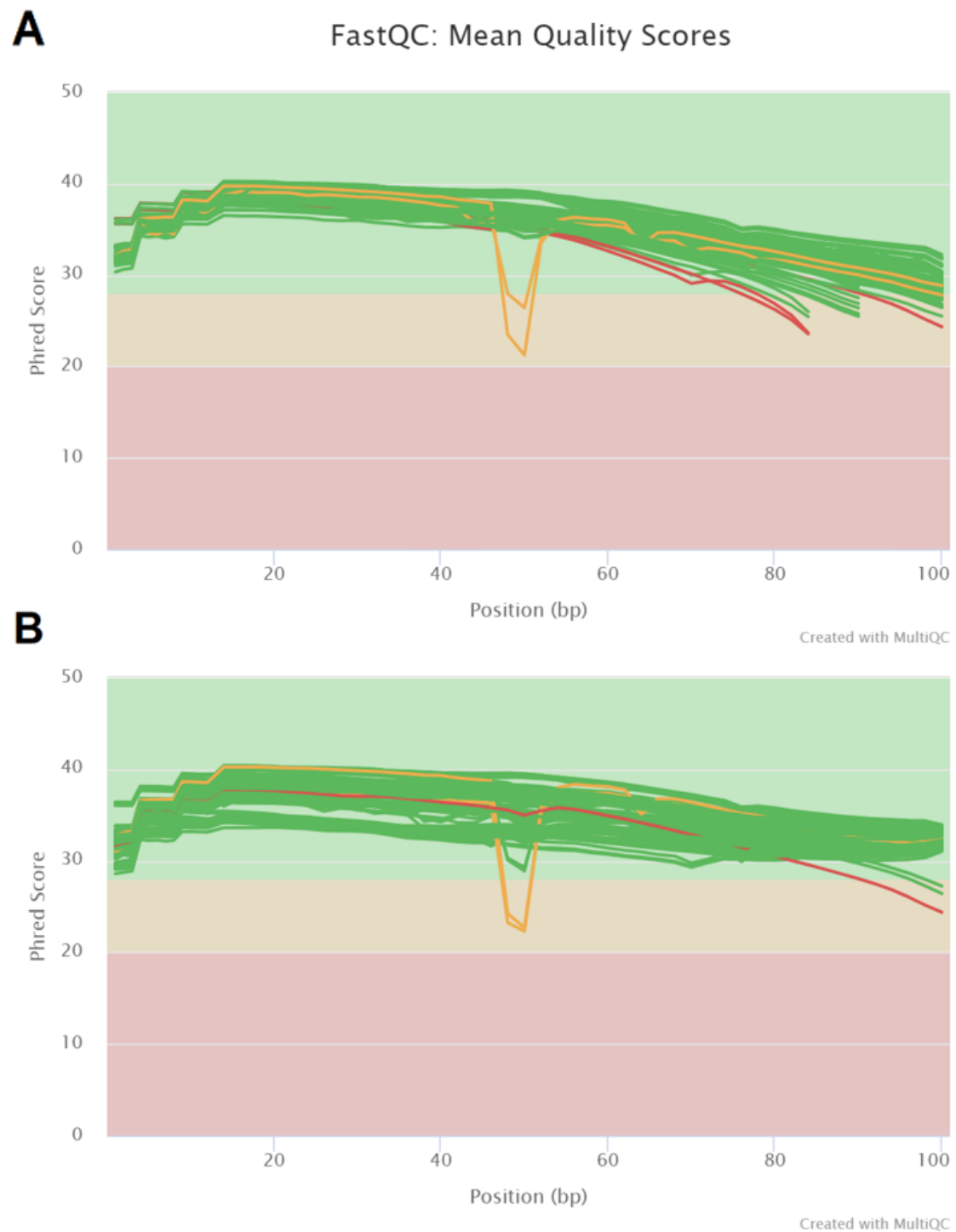


Figure S1: Quality of the reads before (**A**) and after (**B**) filtering with Trimmomatic. The yellow and red lines, which indicate that the reads do not have an optimal quality are due to the low coverage of the individual HL0623.

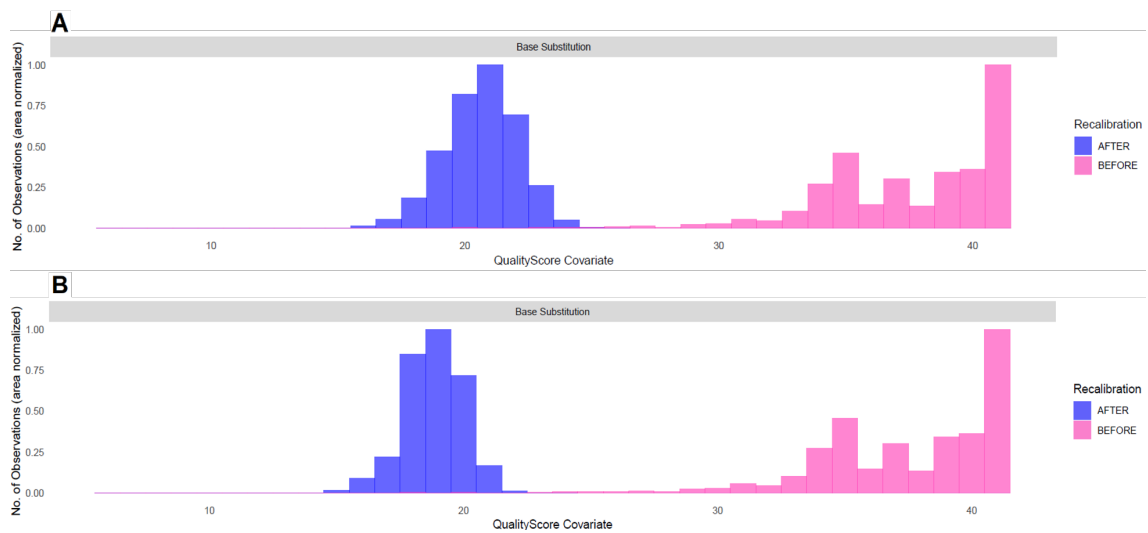


Figure S2: Comparison of the different parameters used for the BQSR. **A.** Quality of the recalibrated BAM files using the variants in the whole genome as training set. **B.** Quality of the recalibrated BAM using the variants in one chromosome as training set

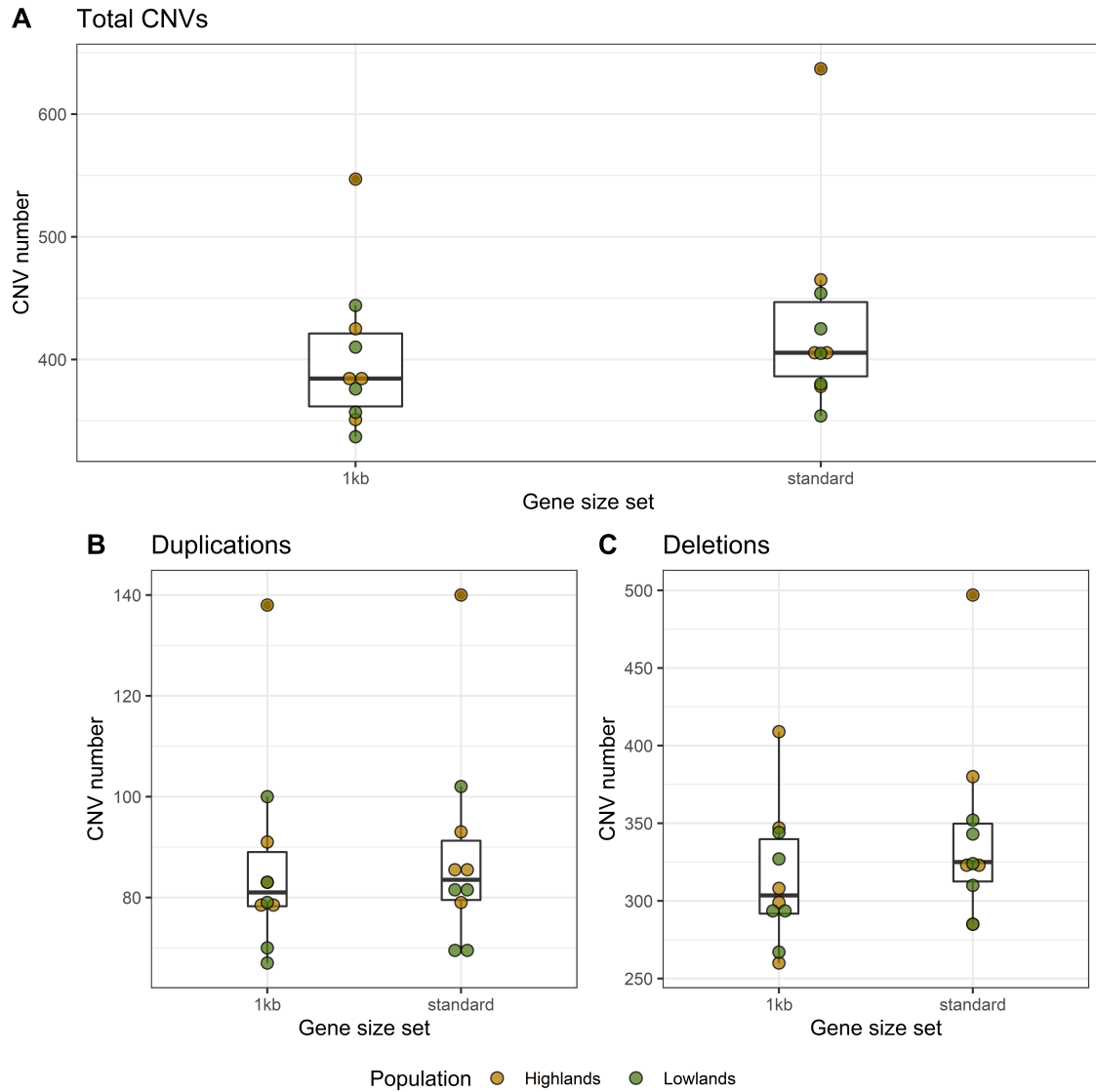


Figure S3: Plots representing the number of CNVs in each individual for the different minimum gene sizes. The dots represent the individuals and the boxes the mean with the standard deviation. No significant difference can be seen between both groups **A**. Number of total CNVs. A small decrease in the set of genes with at least 1kb in size can be observed. **B**. Number of duplications. **C**. Number of deletions. A small decrease in the set of genes with at least 1kb in size can be observed.

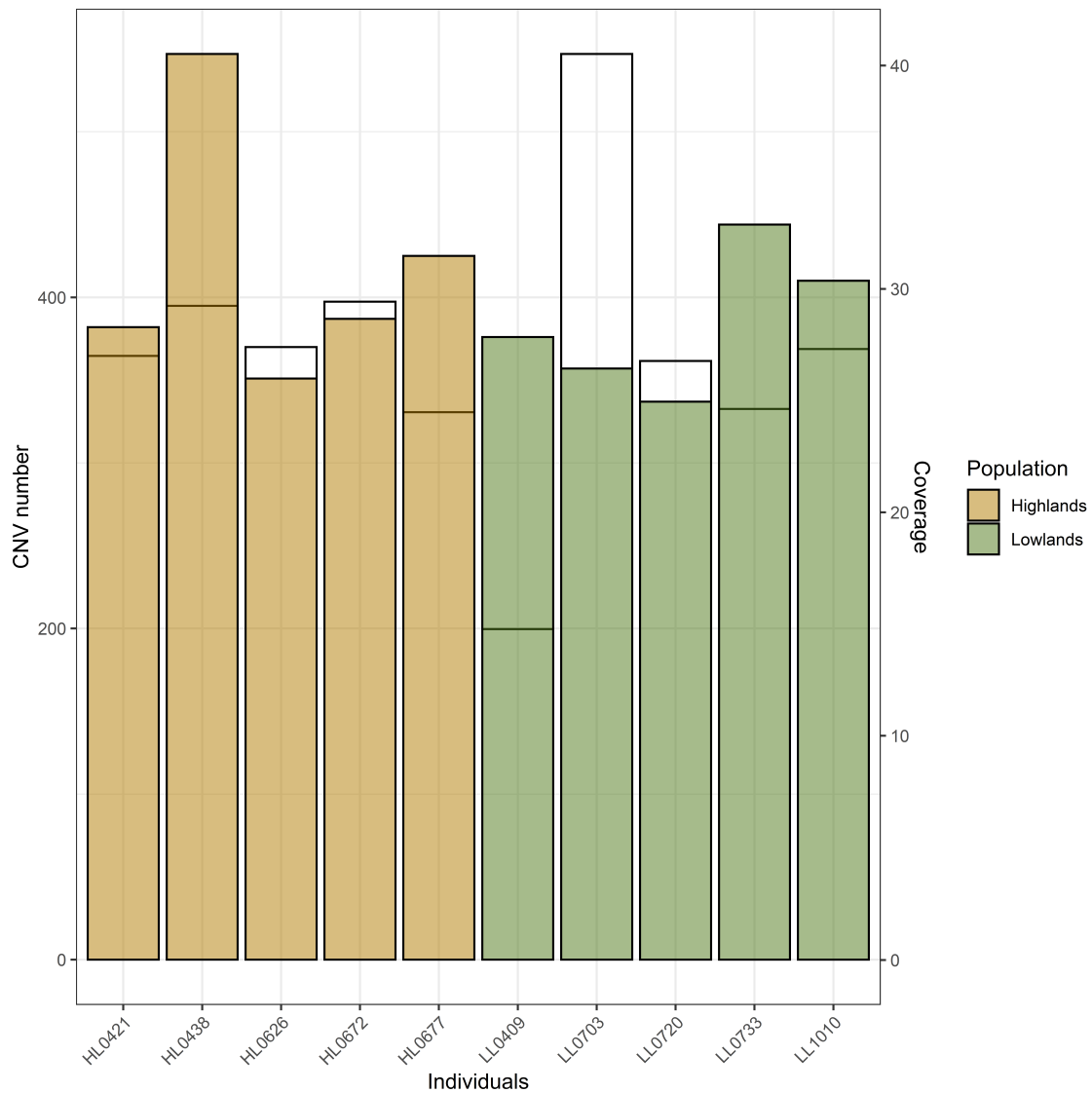


Figure S4: The graphic represents the number of CNVs in relation to the coverage of the BAM files of each individual. The colored bars represent the number of CNVs, which are on the left y-axis. The black lines represent the coverage of the BAM files, which is measured in the right y-axis. As can be seen, there is no association between the number of CNVs and the coverage of the BAM files, since the highest and lowest bars of the coverage do not correspond with the highest and the lowest bars of the number of CNVs

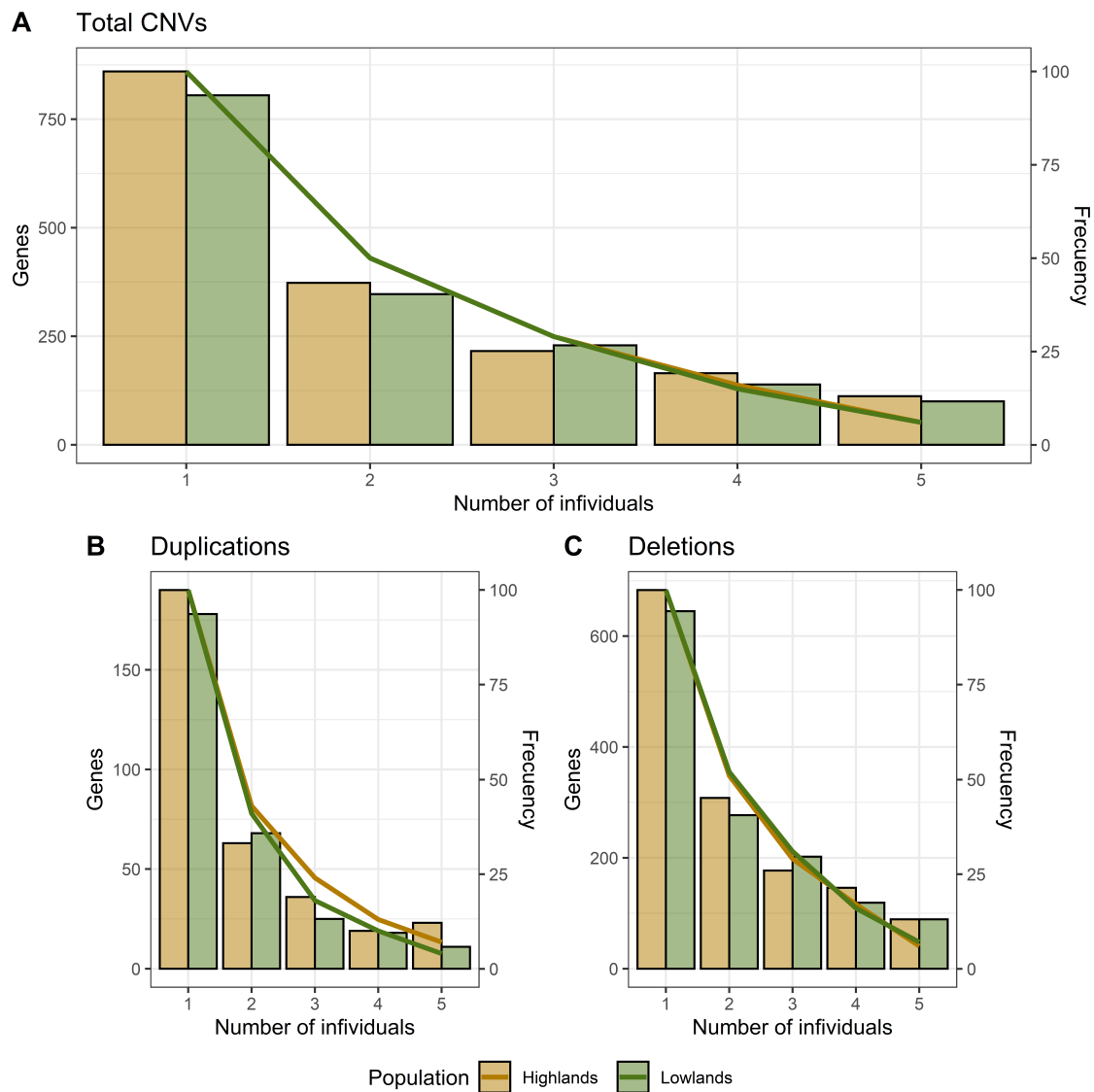


Figure S5: Number of copy number variants among populations when applying an 85% of overlap with low mappability regions. The bars represent the number of genes that are called as variant in a number of individuals. The lines represent the accumulated percentage of variants, going from the most restrictive (five individuals) to the variants in at least one individual. **A.** Comparison of the total number of CNVs in each population. **B.** Comparison of the number of duplications in each population. **C.** Comparison of the number of deletions in each population.

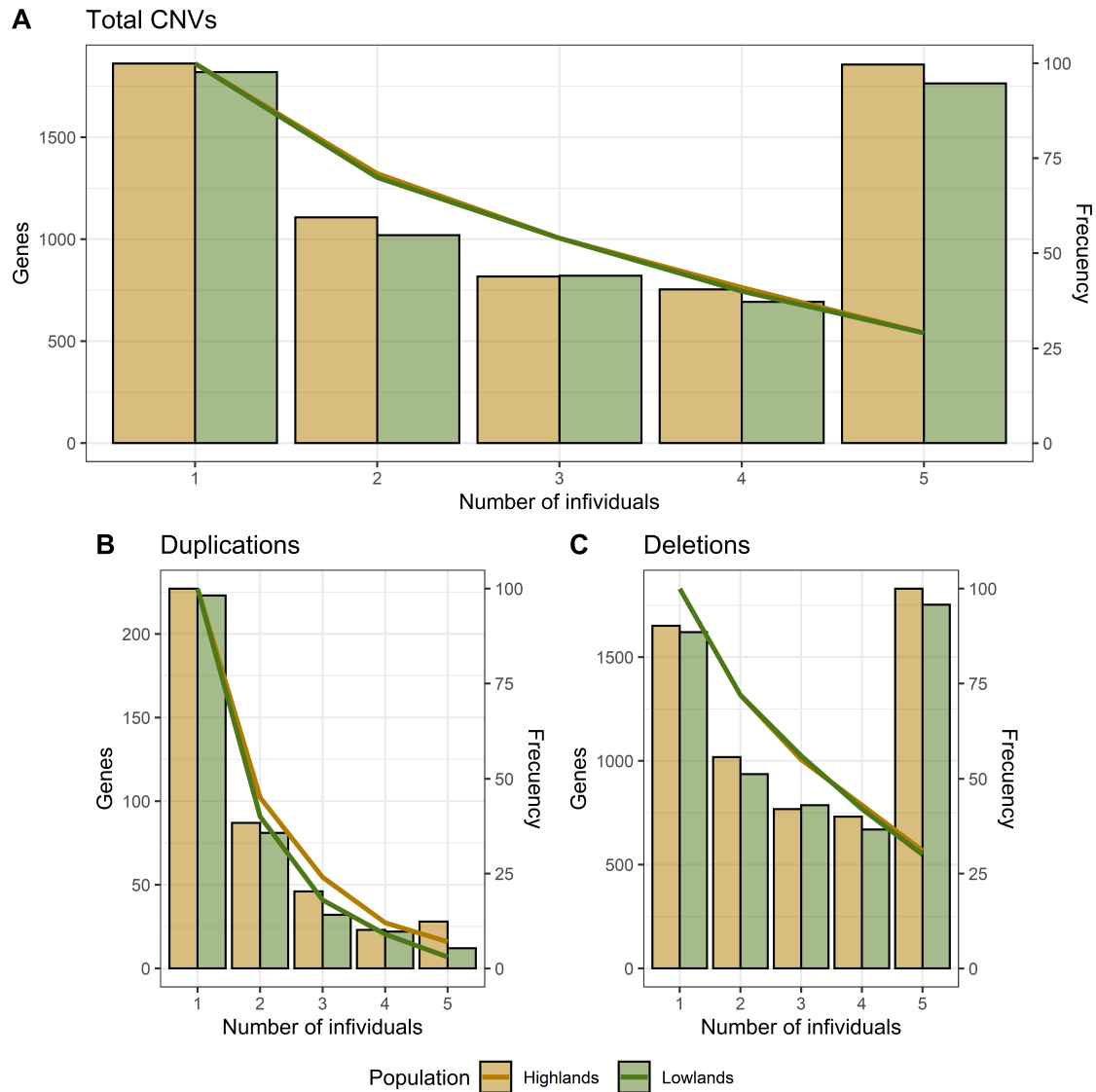


Figure S6: Number of copy number variants among populations without applying the mappability filter. The bars represent the number of genes that are called as variant in a number of individuals. The lines represent the accumulated percentage of variants, going from the most restrictive (five individuals) to the variants in at least one individual. In this case, an unexpectedly high number of CNVs are present in 5 individuals. **A.** Comparison of the total number of CNVs in each population. **B.** Comparison of the number of duplications in each population. The general trend of CNV number occurs as expected and seen in the other graphics. **C.** Comparison of the number of deletions in each population. The graphic shows that the abnormality in the CNV number in 5 individuals is due to the increase in deletions

