# CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD IRAPUATO

Divergencia fenotípica, transcripcional y regulatoria asociada al proceso de domesticación del chile (*Capsicum annuum*)

Tesis que presenta

**Erik Omar Díaz Valenzuela**

Para Obtener el Grado de

**Maestro en Ciencias**

En la Especialidad de

**Biotecnología de Plantas**

**Director de Tesis:** Dra**.** Angélica Cibrián Jaramillo

Irapuato, Guanajuato

Agosto, 2017

This Project was developed at the Ecological and Evolutionary Genomics Laboratory in Unidad de Genómica Avanzada of the Centro de Investigación y de Estudios Avanzados del Instituto Politécnico nacional, under the guidance of Dra. Angélica Cibrián Jaramillo

## Acknowledgements

First, I want to express my gratitude to my advisor Dra. Angélica Cibrián, for her guidance, patience, trust, and for all the time invested in intense discussions about plant evolution and genomics of domestication. She is the first professor that have truly encouraged me to go beyond my analytic and writing skills.

To Consejo Nacional de Ciencia y Tecnología (CONACyT), for the scholarship granted.

To Melissa Dipp, who is always there to hear and discuss my now-poorly supported hypotheses about plants evolution. But also for all the non-scientific talks, she is one of the most interesting and smart persons I've ever known. Specially, she deserves my respect and gratitude because she has decided to take with me the train of a scientific career, even when the odds of 'success' are small in this country where people is obsessed with slowdown the scientific development.

To Alex Aragon, my PhD candidate friend, who is the only non-biologist guy that challenges me to be a better biologist. He shares with me the adrenaline rush before reading a 50 pages review paper about the role of transcriptional enhancers in morphological evolution of early invertebrates. I respect that, because few people proudly recognize their nerd side.

To my family. They all have supported me not only economically, but emotionally, they trust in my dreams. This years of being away from them have made me understand how much of them I carry on my genes and in my principles.

# INDEX

# Resumen

Las plantas han habitado la tierra desde hace cerca de 470 millones de años y a través de ese tiempo han desarrollado una amplia diversidad de morfologías. La mayor parte de esta diversidad morfológica correlaciona con recableados en redes regulatorias que controlan la expresión génica (GRN), más que con cambios en genes que codifican para proteínas enzimáticas o estructurales. Las GRN controlan la expresión de genes río abajo mediante la interacción entre elementos regulatorios *cis* (promotores) y *trans* (factores de transcripción). Por tanto, una gran fracción de la variación genética asociada a la evolución de la morfología y divergencia de expresión génica debería estar contenida en elementos *cis* y *trans* regulatorios. Casos de evolución morfológica en intervalos cortos de tiempo, como la domesticación de plantas (< 10,000 años), son escenarios ideales para entender cómo se modulan los paisajes transcripcionales durante divergencias fenotípicas, pero también para entender qué cambios regulatorios les acompañan. En este estudio utilizamos el fruto de chile (*Capsicum annuum*) como modelo para investigar cuáles cambios transcripcionales han derivado de la domesticación mediada por humanos, así como diseccionar sus mecanismos genéticos subyacentes.

Empleamos tecnología de secuenciación de RNA para: 1) caracterizar los patrones globales de divergencia transcripcional entre frutos de chiles cultivados (C) y silvestres (W), y 2) diseccionar los mecanismos regulatorios que subyacen la divergencia de expresión génica mediante análisis de expresión alelo-específica (ASE) en los híbridos F1 producidos mediante la cruza de C x W. Encontramos que 51% de los genes compartidos entre C y W muestran divergencia de expresión. Esta divergencia parece estar sesgada hacia la sobreexpresión en los chiles cultivados, y además enriquecida en procesos biológicos como respuesta a estrés biótico y abiótico, y desarrollo de fruto. También, por medio de un pipeline bioinformático nuestro para implementar ASE, descubrimos que 66% de los genes que muestran divergencia de expresión, fueron producto de variación *trans* regulatoria, 44% de variación *cis* regulatoria y 41% de ambas. Además, encontramos evidencia que sugiere que la mayoría de las mutaciones *cis* regulatorias podrían haber surgido como mutaciones dominantes, mientras que las mutaciones *trans* regulatorias parecen ser producto de mutaciones recesivas de pérdida de función (LOF).

Finalmente, también encontramos que los procesos biológicos desarrollo de fruto y reproducción, están exclusivamente enriquecidos en genes que muestran evidencia de variación solo-trans. Lo que nos llevó a hipotetizar que los cambios en la morfología del fruto de *C. annuum*, que derivaron de la domesticación mediada por humanos, podrían estar subyacidos por mutaciones recesivas LOF en factores de transcripción u otro tipo de elementos regulatorios que actúen en trans. Sin embargo, saber si las mutaciones se encuentran en la secuencia codificante o río arriba no es posible sin reconstruir y caracterizar la red regulatoria que controla la morfología del fruto. A pesar de todo el trabajo que falta para entender completamente la evolución de la morfología del fruto, este estudio provee de hallazgos importantes que contribuyen al entendimiento general de los mecanismos genómicos que impulsan la evolución fenotípica de las plantas.

# ABSTRACT

Plants have inhabited the earth since *ca*. 470 Mya and through that time they have evolved an ample diversity of morphologies. Most of this morphological diversity often correlates with changes in gene expression driven by the rewiring of ancient gene regulatory networks (GRN), rather than changes in protein coding genes. GRNs control the expression patterns of genes downstream by means of interactions between *cis* (promoters) and *trans* (transcription factors) regulatory elements. Thus, genetic variation associated to morphology evolution and expression divergence, should be contained in *cis* and *trans* regulatory elements. Short timeframe morphological divergences, as plant domestication, are ideal scenarios to understand how transcriptional landscapes are shaped during rapid phenotypic divergences, but also to understand what regulatory changes accompany them. In this study, we utilized the fruit of *Capsicum annnuum* as a model to investigate what transcriptional changes were derived from human-mediated domestication, and what their underlying genetics are.

We used RNA-seq technology to: 1) characterize the genome-wide transcriptional divergence between wild (W) and cultivated (C) *C. annuum* fruits, and 2) dissect the regulatory mechanisms underlying expression divergence by means of Allele Specific Expression Analyses in F1 hybrids derived from crosses between C and W. We found that 51% of the shared genes between C and W show differential expression. This expression divergence appears to be biased towards overexpression in cultivated chilies, and loaded into biological processes such as response to biotic and abiotic stresses as well as fruit development. Also, by means of a custom-made bioinformatic pipeline to perform ASE, we discovered that, 66% of the genes with expression divergence between W and C, showed evidence of trans-regulatory variation, 44% of cis-acting variation, and 41% of both. Also, we found that most of the cis-regulatory changes could have resulted from dominant mutations, while most of the trans-regulatory changes appear to have arisen from recessive loss of function (LOF) mutations.

In addition to the above-mentioned evidence, we found that, the biological processes reproduction and fruit development, were exclusively loaded into genes that showed trans-only variation. Which led us to hypothesize that changes in fruit morphology during human-mediated domestication could be underlaid by recessive LOF mutations at trans-acting regulatory elements. However, whether the mutations are in the coding sequence of the trans-acting elements or upstream, is not possible to know without reconstructing the whole GRN that controls fruit morphology. Despite of all the ongoing work needed to fully characterize fruit morphology evolution, this study provides important findings that contribute to the overall understanding of genomic mechanisms driving plant phenotypic evolution.

# INTRODUCTION

## The role of regulatory networks in rapid plant evolution

Plants have inhabited the earth since *ca.* 470 Mya and through that time they have evolved an ample diversity of morphologies, from simple and unidimensional tissues as those seen in liverworts, to the three-dimensional tissues and complex structures such as flowers and fruits, seen in angiosperms (Gensel, 2008; Pires & Dolan, 2012; Harrison, 2017). The repertoire in complexity in plant morphologies throughout their evolution often correlates with the expansion and rewiring of ancient gene regulatory networks (GRN) (gene tool kits), rather than changes to the protein coding sequences. For instance, most of the plant developmental transcription factors (TFs) such as KNOX and LEAFY, are conserved at the sequence level between phylogenetically distant groups such as mosses and angiosperms (Pires et al. 2013). However, these TFs have experienced duplications followed by neo-functionalizations or sub-functionalizations (Prince & Pickett, 2002; Riaño-Pachón et al. 2008; Pires et al. 2013), that impact the GRN in which they are in, resulting in increased spatiotemporal expression variation with varying phenotypes (Richardt et al. 2007). In addition to gene duplication and expansions, GRN are rewired through mutations at both *cis* (local acting non-coding DNA sequences) and *trans* (distant acting TFs) regulatory elements that interact with conserved TFs and their targets, affecting the developmental timing and tissue specificity of gene expression (Shubin & Marshalll, 2000; Swinnen et al. 2016). Thus, regulatory mechanisms ensure the genetic integrity of central gene tool kits while fueling plant morphological evolution.

Plant domestication is an instance of often dramatic morphological evolution, albeit at a much shorter temporal scale when compared to longer term evolutionary changes such as the phenotypic radiation of angiosperms that took place in the Early Cretaceous *ca.* 145 Mya (Taylor et al. 2009). We would expect that selection would favor rapid morphological changes driven by changes in regulatory mechanisms rather than changes in coding sequences.

Plant domestication is thus an ideal model to understand how plant morphology evolution begins, how transcriptional landscapes are modulated, and what are the regulatory mechanisms underlying all these changes.

In the present study, the fruits of *Capsicum annuum var. annuum* (cultivated) and its wild relative *Capsicum annuum* var. *glabriusculum* were used as a model to understand: 1) What transcriptional changes distinguish the wild from the domesticated plant; 2) What the main genetic inheritance patterns of those transcriptional changes are; and ultimately, begin to infer 3) How human-mediated domestication may have driven the diverse phenotypic fruit morphology that exists in this cultivated species. In the following paragraphs, I will provide a more detailed overview of the selected topics required to address this model, beginning with basic concepts and ending with our model system.

## Genetics of phenotypic divergence: dissecting *cis* and trans regulatory mutations and its effects of gene expression

Dissecting the genetics of phenotypic variation is a common topic that evolutionary biologists have addressed since the first decades of 20th century. In the most classical approach, phenotypic variation within and among populations is the result of complex interactions between genes and the environment. Until 1975, it was thought that the main source of genotypic variation explaining divergent phenotypes such as different brain size among gorillas, chimpanzees and humans, was product of point mutations at protein coding genes. However, King and Wilson found that blood proteins of those groups shared 99% of sequence identity, evidence that led them to propose that variation in gene expression by means of regulatory mutations, should explain the degree of divergence. With the advent of the high-throughput sequencing and fine cloning technologies, it has been possible to test hypotheses regarding regulatory divergence and its impact on phenotypes. To date, there is robust evidence that support how regulatory mutations at both, coding and non-coding DNA sequences can alter the tempo and mode of gene expression, and consequently bring novel variation at morphological and physiological traits (Wray, 2007; Stern & Orgogozo, 2008; Wagner & Lynch, 2010; Wittkopp & Kalay, 2012; Martin & Orgogozo, 2013).

But, to understand how those types of mutations alter the molecular and morphological phenotypes, we must adopt a view where phenotypes are the outcome of a multi-layered gene network on which highly conserved transcription factors interact with more specific transcription factors and local regulatory DNA sequences such as promoters and enhancers (Carroll, 2008). In the light of this, dissecting the contribution of the mutations at any of the above-mentioned elements of gene networks is required to understand the genetic architecture of simple and complex phenotypes.

## Dissecting *cis* and *trans* regulatory mutations and its effects on gene expression

At the transcriptional level, gene expression is governed by biochemical interactions between local (*cis*) and distant (*trans*) elements. Cis-regulatory elements such as promoters and enhancers are non-coding DNA sequences that contain short motifs called binding sites, where transcription factors or other trans-regulatory molecules, can bind and perform DNA transcription, which is the first level of regulating gene expression (reviewed in Wittkopp & Kalay, 2012).

Taking this into account, one can suggest that mutations affecting either a cis or a trans-regulatory element can result in a rewiring of the regulatory network on which they are involved and thus, affect the expression of the genes that they control downstream (**Figure 1**).

Distinguishing between cis and trans-acting genetic variation that impacts gene expression is important because these mechanisms shape phenotypes. Understanding how cis and trans-acting genetic variation evolved and how each is inherited can help in predicting their effects. For example, gene expression levels, as other quantitative phenotypes such as fruit shape and fruit size can be inherited additively or nonadditively (Gibson et al. 2004; Tanksley, 2004).

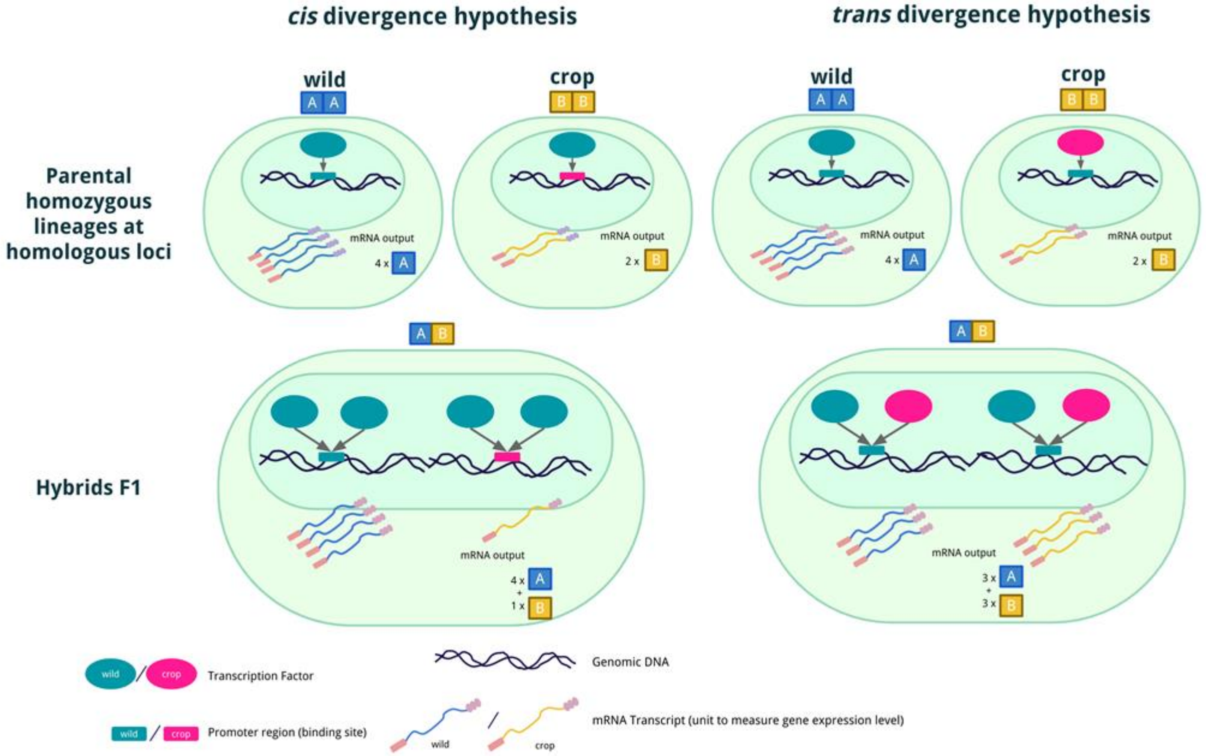


**Figure 1.** Dissection of *cis* and *trans* regulatory divergence in F1 hybrids. Two hypothetical scenarios where mutations at only *cis* or only *trans* regulatory elements led to expression divergence between wild and cultivated plants (crop). The left panel shows how a *cis* regulatory mutation (pink box) causes a reduction in binding affinity for both the conserved wild and cultivated transcription factors (blue ovals). In the right panel, a *trans* mutation in a transcription factor of the cultivated genotype (pink oval) reduced the binding affinity for the conserved *cis* regulatory sequences of wild and cultivated genotypes. The number of mRNA transcripts in the cytoplasm indicates the relative gene expression. In the F1 hybrids, a cis-regulatory mutation only affects the expression level of the cultivated allele as the cis-trans interaction in conserved for the wild allele (allele specific expression). Whereas a trans-regulatory mutation affects the expression of the two alleles as in the hybrid nucleus the two conserved *cis* elements are exposed to both the divergent and the conserved transcription factors. Note that it is needed to have homozygous genetic markers in parental genotypes, and heterozygous in F1 hybrids to avoid ambiguities in allele specific expression assessment.

Indeed, it has been reported that cis-acting regulatory mutations are more likely to be inherited additively than trans mutations because their impacts on phenotypes can be observed in heterozygous F1 hybrids. In contrast, trans-acting variation is harder to detect because it would require knowing all possible targets to be able to predict the resulting phenotypes, but also because  (Yvert et al. 2003; Lemos et al. 2008).

In terms of times of divergence, one can expect that the proportion of *cis* and *trans* variation accounting for expression divergence is highly associated to the timing of populations, eventually leading to speciation or separation of lineages. For instance, two *Drosophila* species that diverged *ca.* 145 Mya showed higher proportion of *cis* variation, while the expression divergence between yeast strains, and between wild and cultivated plants, with relatively short times of divergence (<10,000 years) shows a greater proportion of *trans* divergence (Yver et al. 2003; Gompel et al. 2005).

The magnitude and direction of cis and trans-regulatory mutations on global expression divergence can be estimated by comparing the magnitude of the expression difference between two genotypes of interest to the relative allelic expression in F1 hybrids produced by crossing these two genotypes (McManus et al. 2010). This is because allele-specific measures of gene expression in heterozygotes resemble the relative transcription amount of two cis-regulatory alleles in the same trans-regulatory cellular environment (Cowles et al. 2002) (**Figure 1**). The fraction of the total expression difference between the two genotypes of interest that is not explained by cis-regulatory divergence is attributed to trans-regulatory divergence (Wittkopp et al. 2004).

Until recent years, dissecting regulatory mutations underlying divergent phenotypes was limited to measuring expression of a few hundreds of genes using molecular biology and sequencing technologies such as microarrays, pyrosequencing, and qRT-PCR (Wittkopp et al. 2004).

However, with the arrival and fast adoption of RNA-seq technologies (Nagalakshmi, et al. 2008), it has been possible to perform genome-wide measurements of gene expression and understand the dynamics and evolutionary consequences of regulatory divergence. For instance, regulatory divergence has been mostly assessed by means of measurements of allele specific expression (ASE) (Crowley et al. 2015; Brill et al. 2016). Results from these studies show that cis-regulatory variation accounts for morphological evolution while trans-acting variation accounts for adaptive evolution. Also, the proportion of expression divergence explained by *cis* and *trans* variation is very variable.

Studies aiming to dissect regulatory mutations responsible for genome-wide expression divergences are currently limited because of the capabilities of distinguish between two genotype-specific alleles of the same gene in F1 hybrids. Nowadays, most of the studies have achieved ASE analysis by means of using single nucleotide polymorphisms (SNPs) found in mRNA molecules and whose genotype in parental individuals must be homozygous (**see Figure 1 for further explanation**).

In plant domestication and speciation, the consensus is that *trans* variation is linked to adaptive evolution and accounts for most of the expression divergence when compared to the impacts of *cis* variation (Bell et al. 2013, Lemmon et al. 2014; Combes et al. 2015). Which contrast with *cis* variation accounting for most of the regulatory variation and morphological evolution in animal systems (Wittkopp et al. 2004; Wittkopp et al. 2007; McGregor et al. 2007; Carroll, 2008; Wittkopp & Kalay, 2012). This has been partially explained by the fact that *cis* regulatory mutations accumulate nearly-neutral effects over time until they are detected by selection, while mutations at trans-regulatory elements show their impacts almost immediately if they are not deleterious because they have pleiotropic effects.

However, despite this progress, there is still no generalized trends in plants or within plant families, nor in various forms of domestication processes. I will discuss this last topic in next sections, and focus on our model system *Capsicum*, or chili pepper

## Genetics of plant domestication: predominant regulatory mutations

Crops, or domesticated plants, are the result of our continuous selection and use of plants for at least the last 10,000 years (Weiss et al. 2004). Crops have helped us maintain food security while continuing to develop our religious and sociocultural welfare. During the domestication process humans selected plant morphological traits that eventually genetically distinguished the cultivated form from its ancestors, or crop wild relatives and resulted in a 'domestication syndrome' (Meyer & Purugganan, 2013).

Plant traits of the domestication syndrome *sensu stricto* include physiological changes such as losing the ability to disperse, losing seed dormancy, and reduction of toxic compounds, among others (**Figure 2-A**). Some authors (Doebley et al. 2006; Meyer & Purugganan, 2013) have proposed that morphological changes such as increased fruit size or changes in shoot architecture should be recognized as diversification traits because they were acquired in the late stages of domestication (**Figure 2-B**).

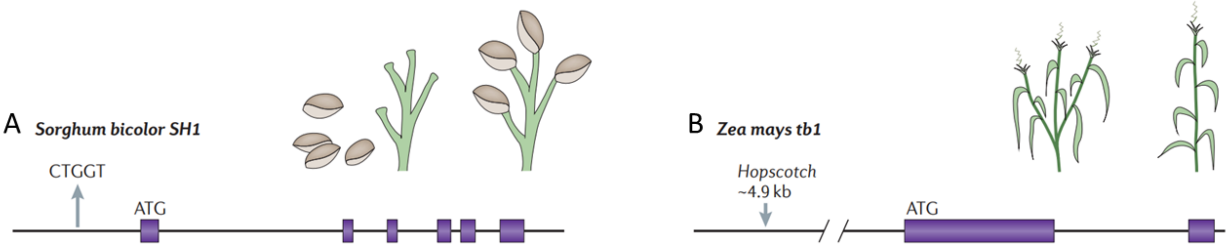


**Figure 2.** Examples of domestication and diversification traits underlaid by mutations at *cis* regulatory regions of orthologous transcription factors. (**A**) The non-shattering phenotype in *Sorghum bicolor* resulted from a deletion of five nucleotides upstream the transcription start site (promoter) of the *SH1* gene. (**B**) The altered shoot architecture of maize resulted from the insertion of the Hopscotch transposable element in the *cis* regulatory region of *Zea mays teosinte branched1* (*tb1*) gene. Modified from Meyer and Purugganan, 2013.

Most of the studies that have aimed to understand the genetic architecture of domestication phenotypes have used fine QTL mapping and molecular complementation analysis. The consensus of the most cited studies suggests that mutations at cis-regulatory elements of transcription factors are the main source of variation associated to evident morphological changes. For instance, the loss of axillary branches and increased apical dominance in maize is product of a mutation the cis-regulatory region of the *teosinte branched1* gene (a transcription regulator) that altered its expression pattern (Rong-Lin et al. 1999; Studer et al. 2011).

There is also evidence of the evolutionary forces shaping the fruit morphology. For example, the increased mass and size that modern tomatoes exhibit when compared to their wild relatives appears to be the product of a single nucleotide polymorphism (SNP) in the cis-regulatory region of the gene *fw2.2,* a transcription factor that regulates the cellular divisions in the fruit (Frary et al. 2000, Lin et al. 2014) (**Figure 3**).

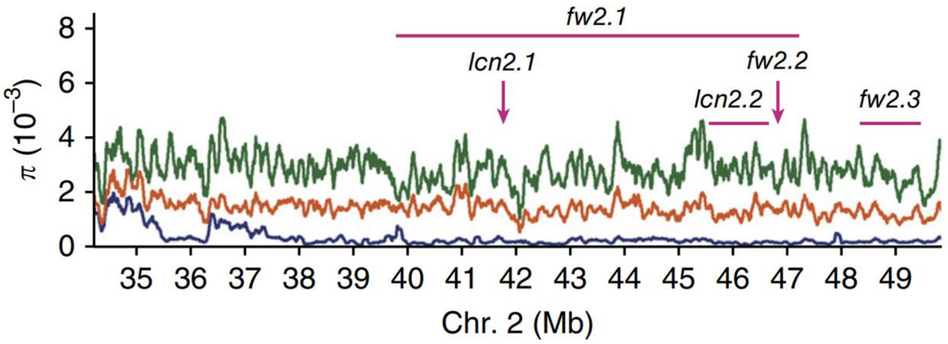


**Figure 3.** Distribution of nucleotide diversity (*π*) of the PIM (green), CER (orange) and BIG (blue) tomato lines within the improvement sweep harbouring the fw2.2 QTL on chromosome 2. Small values of *π* indicates signals of selection. Taken from Lin et al (2014).

## *Capsicum annuum* domestication as a model to understand fruit morphology evolution

Chile (*Capsicum annuum* var. *annuum*) is a diploid member of the Solanaceae family that is thought to be one of the first domesticated crops in America (Kraft et al. 2014). Genetic and anthropological evidence suggest that chiltepin (*Capsicum annuum* var. *glabriusculum* is the wild progenitor of all modern Mexican chilies (Pickersgill, 1971; Loaiza-Figueroa et al. 1989; Aguilar-Meléndez et al. 2009).

Mexican chilies show a wide diversity of morphophysiological traits in shoots, but more markedly in fruit characteristics such as size, shape, colour, and flavour (Kim et al. 2014; Qin et al. 2014). It has been proposed that Mexican chilies show that degree of fruit phenotypic variation due to both, local adaptation to a wide spectrum of environments where it has been cultivated, and differential agricultural practices and uses across diverse ethnic groups (Kraft et al. 2014). Those two factors have resulted in a marked domestication syndrome.

Nowadays there are three published reference genomes of *Capsicum*, Zunla-1, chiltepin and CM334 (Kim et al. 2014; Qin et al. 2014). Data from those genomes agree that cultivated chilies bear larger genomes than wild chilies, for instance the genome size of Zunla-1 was estimated to be 3.26 Gb while the chiltepin genome size was estimated to be 3.07 Gb. Data from those studies also showed that the *Capsicum* genome is highly dynamic because the ~81% of the total genome size is composed of transposable elements, mainly LTR retrotransposons such as *Gypsy* and *Copia.* Sequence divergence between cultivated and wild *Capsicum* has been reported to be between 0.35% and 1.85%. Annotation of the *Capsicum* genomes reported ~35,000 protein coding genes, of which, 17,000 have orthologous in tomato (Kim et al. 2014). *Capsicum* domestication appears to be biased towards fruit traits but also to stress response mechanisms. For instance, 115 genomics regions that show strong selective sweeps contain 511 genes annotated as developmental regulators and stress or defence response (Qin et al. 2014).

In sum, despite various genomic and transcriptomic studies that have aimed to understand the underlying genetics of *Capsicum* domestication, none of them have integrated data from genome-wide expression divergence between wild and cultivated chilies, nor explored causative mutations or mechanisms related to their gene regulation. There is also lacking knowledge regarding how *Capsicum* fruit morphology evolved and how the novel associated traits are inherited. The main goal of this study is to advance the understanding of gene regulatory mechanisms in fruit morphological diversity in the context of domestication, and contribute to the overall understanding of the genomic mechanisms driving plant phenotypic evolution.

# HYPOTHESIS

An excess of trans-acting regulatory variation should be responsible of the genome-wide expression divergence generated during the rapid and intense domestication process of the fruit of *Capsicum annuum*

# OBJECTIVE

## Main objective

To characterize the genome-wide expression divergence between wild and cultivated *Capsicum annuum* fruits as well as its underlying regulatory mechanisms.

## Specific objectives

a) Generate a collection of F1 plants from a cultivated x wild cross.

b) Characterize morphological variation of the parents and the F1 population and carry out a morphometric analysis.

c) Utilize RNA-seq technology to describe genome-wide expression divergence between wild and cultivated *C. annuum* fruits.

d) Perform Allele Specific Expression analysis to dissect the regulatory mechanisms responsible for expression divergence between wild and cultivated *C. annuum.*

# MATERIALS AND METHODS

## Plant accessions

Mexican chilies show a wide spectrum of phenotypic diversity, going from shrub-like and small-fruited morphologies in wild accessions, to monopodial and big-fruited in cultivated ones. Aiming to capture this phenotypic variation and dissect its underlying genetics, a cultivated accession (C) 'Puya' *Capsicum annuum* var*. annuum*, and a wild accession (W) 'Chiltepin' *Capsicum annuum* var. *glabriusculum* (Dunal) Heiser and Pickersgill, were chosen to perform crosses and generate a F1 Hybrid population. Dry fruits of W were collected at 'El Patol', Guanajuato, México, while fruits of C were obtained from a local market at Irapuato, Guanajuato.

## Germination, growing conditions and crosses

Seeds from both, C and W were extracted from dry fruits and treated as follows: a wash with a 10% v/v bleach solution followed by a chemical scarification with 0.05 N HCl at 35° C by 30 min, and a second wash with distilled water. Seeds were then sown into 0.5 L pots containing commercial Peat Moss mix (peat moss, vermiculite, perlite 3:1:1) and let to germinate in a temperature-controlled (18-28° C) greenhouse at LANGEBIO-CINVESTAV, Irapuato. Juvenile and adult plants were fertilized every two weeks with standard NPK fertilizer. Once plants reached their sexual maturity reciprocal crosses were performed as follows: flowers from both, C and W were designated as female (pollen receptor) or male (pollen donor), unopened female flowers were carefully emasculated and pollinated by shaking a male flower overhead the stigma of a female flower and covered with a labeled (♀C x ♂W or ♀W x ♂C) paper bag to avoid cross-pollination from other plants. C and W plants were let to self-fertilize. Fruits were harvested 60 days after anthesis (DAA) and their seeds were collected and stored in paper bags.

## Parental and F1 fruit phenotyping

Seeds from C and W, as well as their F1 population were germinated and grown as described above, however, the ♀W x ♂C cross resulted in abortive fruits and sterile seeds, possibly due to the unilateral incompatibility that has been previously reported in Solanaceae plants, including *Capsicum* genus (Onus and Pickersgill, 2004). 40 DAA has been shown to be the most entropic time point in the transcriptome of chili development, so it contains most of the dynamics and diversity in gene expression (Martínez-López et al. 2014). Taking this into account, the phenotyping and sequencing strategies were designed.

## Fruit phenotyping and morphometric analysis

The phenotyping was performed considering the morphometric variation of 400 individual fruits at 40 DDA: 200 from 5 individuals of ♀C x ♂W (hereafter called CxW), 100 from 5 individuals of C as well as 100 from 5 individuals of W. All fruits where scanned at 600 dpi and the images were analyzed with the ImageJ software (Rasband, 2016) to extract shape descriptors such as area, minor axis, major axis, aspect ratio (major axis/minor axis), roundness (4x[(area)/π(major axis)$^2$]), and circularity (4π[(area)/(perimeter)$^2$]. The morphometric data were subjected to a principal component analysis (PCA) to extract the best descriptors (vectors of variables that describe a large amount of the continuous variation in the whole data set) and use them to dissect the inheritance modes of C and W fruit shape in their CxW population. Best descriptors where analyzed with an Analysis of Variance (ANOVA) to test the follow hypotheses:

- $H_1$: $\mu C = \mu CxW \neq \mu W$ (dominance of C over W in F1)
- $H_2$: $\mu W = \mu CxW \neq \mu C$ (dominance of W over C in F1)
- $H_3$: $\mu W > \mu CxW < \mu C$ (additivity W>C)
- $H_4$: $\mu C > \mu CxW < \mu W$ (additivity C>w)

Both PCA and ANOVA were carried out in R (v. 3.4.1, CRAN).

## Total RNA extraction

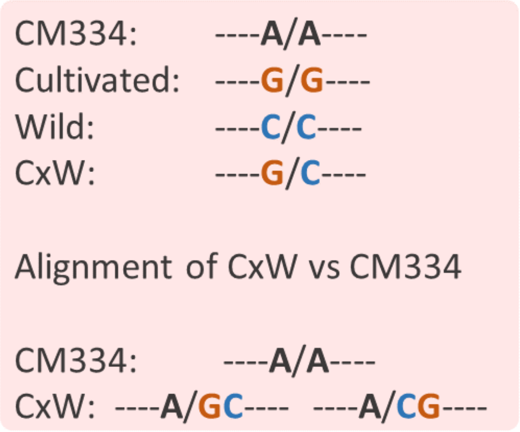
Chilies were harvested and quickly dissected to remove all seeds inside. Then, tissue from placenta and pericarp was collected and immediately homogenized in liquid nitrogen, powdered tissue was transferred to 1.5 uL Eppendorf tubes and deposited in liquid nitrogen. Total RNA was extracted using a standard TRIzol® reagent protocol with the follow modifications: TRIzol® reagent was incubated at 56° C until it was required for the first phase separation, one extra phase separation was performed with chloroform-isoamyl alcohol mixture 24:1, RNA precipitation was enhanced by adding a volume of saline solution (0.8 M sodium citrate and 1.2 M sodium chloride) by one of isopropanol, an extra wash step with 70% ethanol. RNA yield and integrity was evaluated by firstly quantify nucleic acids concentration in NanoDrop 2000™ (Thermo Scientific Nanodrop, USA) and then load a known amount of RNA into a 1.5% agarose gel and let it run at 90 volts for 45 min.

## cDNA library preparation and sequencing

cDNA libraries were prepared at the Genomic Services facility (LANGEBIO-CINVESTAV) according to the Illumina TrueSeq™ RNA Library Preparation Kit v2, which works as follows: 1) mRNA is purified from total RNA by means of the capture of polyA enriched molecules, then chemically fragmented and converted into single-stranded cDNA, which will be used to synthesize the second strand; 2) 'A' bases are added to the double stranded cDNA followed by the ligation of sequencing adapters. The nine cDNA libraries corresponding to the 3 biological replicates of each cDNA libraries where sequenced in one lane of the Illumina Hi-Seq 4000 platform, which yields 1300-1500 Gb of 2x150 bp reads.

## Bioinformatics

To date, there are four high quality published *Capsicum annuum* genomes. However, nor C or W are one of them, which in this study, resulted in two main technical issues: mapping and quantifying transcriptomes can be biased towards either C or W; identifying C or W specific alleles in biallelic-heterozygous positions of CxW can't capture the two 'real' parental alleles (**Figure 4**). Along the following sections I will mention how I solved the above-mentioned technical issues.



CM334: ----**A/A**----
Cultivated: ----**G/G**----
Wild: ----**C/C**----
CxW: ----**G/C**----

Alignment of CxW vs CM334

CM334: ----**A/A**----
CxW: ----**A/GC**---- ----**A/CG**----

**Figure 4**. SNP calling is biased when performing alignments of CxW against the CM334 reference genome because neither the Cultivated nor the Wild genotype bear the same allele as the CM334 reference genome.

## Pre-processing of Illumina reads

To keep only high-quality sequencing reads, Trimmomatic software ver. 0.32 (Bolger et al, 2014) was run in pair-ended mode to remove Illumina adapters and filtering out reads under a quality value of 15 and a total length of 70 bp. Trimmed reads were analyzed with the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to verify their quality and length distribution as well as the absence of Illumina adapters.

## Pseudo-alignment and quantification

As the scope of this study was to detect expression variation between the transcriptomes of cultivated and wild chili accessions, as well as dissect its underlaying regulatory and evolutionary mechanisms, my strategy was dependent of an accurate and non-reference-biased transcript quantification. To achieve this, I opted for a new-generation transcriptomics pipeline (Kallisto + Sleuth). Kallisto software (Bray et al. 2015) is a de Bruijn graph-based software that offers a fast and highly accurate transcript quantification by means of a pseudo-alignment against an annotated transcriptome. As Kallisto works on de Bruijn graphs built from *k*-mers of the reference transcripts, only perfect matching *k*-mers from sample reads are assigned to a reference transcript *k*-mer, regardless of the position of individual bases. This property allowed me to pseudo-align and quantify the nine non-reference transcriptomes against the same reference without sequence bias (Bray et al. 2015). The nine transcriptomes were pseudo-aligned against the CM334 v.1.55 reference transcriptome (Kim et al. 2014) and abundance matrices were extracted to perform downstream analysis such as differential gene expression and Gene Ontology analysis.

## Differential expression and gene enrichment analysis of GO terms

Pseudo-aligned reads from the three replicates of C and the three of W were analyzed to identify differentially expressed genes at *q*-value of 0.05 (only 5% of the genes categorized as differentially expressed are likely to be false positives) with the Sleuth R package (Pimentel et al. 2016), which leverages the bootstrap estimates of Kallisto as well as the variance across independent sets of samples to be analyzed to be more accurate.

The output matrix of Sleuth was utilized to perform a gene enrichment and gene ontology against all transcriptomes available in the panther database (pantherdb.org/), and overrepresentation analysis against *Capsicum* transcriptome with the AgriGO Analysis tool kit (http://bioinfo.cau.edu.cn/agriGO/analysis.php).

Top 100 differentially expressed genes as well as genes involved in fruit ripening and capsaicinoids biosynthesis were analyzed to identify co-expressed genes by means of Euclidean distance of expression values. Differential expression analysis and heatmap analyses were performed in R (v. 3.4.1, CRAN).

## Inheritance classifications

It has been proposed that genes whose expression in hybrids deviate at least between |1.25| and |2.0| foldchange from that showed in parents can be considered to have non-conserved inheritance (McManus et al. 2010). Considering this, a bash script with a conservative threshold of 2-fold was written to classify the inheritance modes of C and W expression patterns on CxW. Modes were defined as additive, dominant, or transgressive according to the magnitude and direction of divergence between CxW and the two parental transcriptomes (**Script Box 1**).

```
#!/bin/bash

#values in matrix are the averaged expression of 3 biological replicates
of genes whose expression was > 4 Kallisto ‘est_counts’

#$1=Log2(CxW-C)
#$2=Log2(CxW-W)

# C dominant over W: CxW expression deviate from W but not from C

      awk ‘sqrt($1*$1) < 2 && sqrt($2*$2) >= 2’ > C_dom_W.txt
# W dominant over C: CxW expression deviate from C but not from W

      awk ‘sqrt($1*$1) >= 2 && sqrt($2*$2) < 2’ > W_dom_C.txt

# additive effects |C>W|: CxW expression is the midpoint between C and W

      awk ‘$1 <= -2 && $2 >= 2’ > add_C_W.txt

# additive effects |W>C|: CxW expression is the midpoint between C and W

      awk ‘$1 >= 2 && $2 <= -2’ > add_C_W.txt

# transgressive upregulation: CxW expression is above both C and W

      awk ‘$1 >= 2 && $2 >=2’ > up_transgressive.txt

# transgressive downregulation: CxW expression is under both C and W

      awk ‘$1 <= -2 && $2 <= -2’ > down_transgressive.txt
```

**Script-box 1**. Bash script written to assign inheritance modes of C and W expression on CxW. Actual code and explanation of each of the modes is shown.

## Allele Specific Expression Assignment

In order to identify the C and W specific alleles in CxW transcriptomes, I developed a bioinformatic pipeline considering the pseudoreference approach (Sarver et al. 2017). This pipeline takes advantage of the Genome Analysis Toolkit (GATK) as well as GATK's RNA-seq variant calling pipeline recommended software's (**Figure 5**). This pipeline captures the two real alleles in biallelic-heterozygous SNPs in F1 Hybrids by means of two independent alignments against two custom pseudo-references that contain the parental specific allele of homozygous positions (it means, the real alleles that F1 must bear). With this pipeline, the above-mentioned issue was solved (**Figure 4).**

The output of my pipeline is a matrix that includes the allele-specific expression values of C and W in CxW, which correspond to the summed read depth of all segregating sites along the same gene ID. The intersection of those gene ID's with the parental expression data base (Kallisto expression values) resulted in the final matrix, on which iterative statistical tests were applied.

## *cis* and *trans* regulatory divergence assignment

As I mentioned above, at the transcriptional level, gene expression is governed by interaction between cis and trans-regulatory elements, thus, mutations affecting each of them or both, impact the gene expression of downstream genes. In this context, any gene with evidence of expression divergence between parental lineages can be dissected in its evolutionary and regulatory mechanisms. To perform the assignment of regulatory mechanisms, three independent statistical tests were performed with conservative thresholds ($p<0.005$) in R (v. 3.4.1, CRAN) (**Script Box 2**). Firstly, iterative $\chi^2$ tests were used to identify genes with expression divergence between parental C and W ($Log_2C$ - $Log_2W \neq 0$). Secondly, Iterative $\chi^2$ tests were also used to test for evidence of *cis* regulatory divergence by means of comparing proportion of C and W alleles in CxW (Log2HybC - Log2HybW ≠ 0). And thirdly, Fisher's exact tests were performed to test for evidence of *trans* effects by comparing the ratios between parental and Hybrids ($Log_2C$ - $Log_2W$ ≠ Log2HybC - Log2HybW).

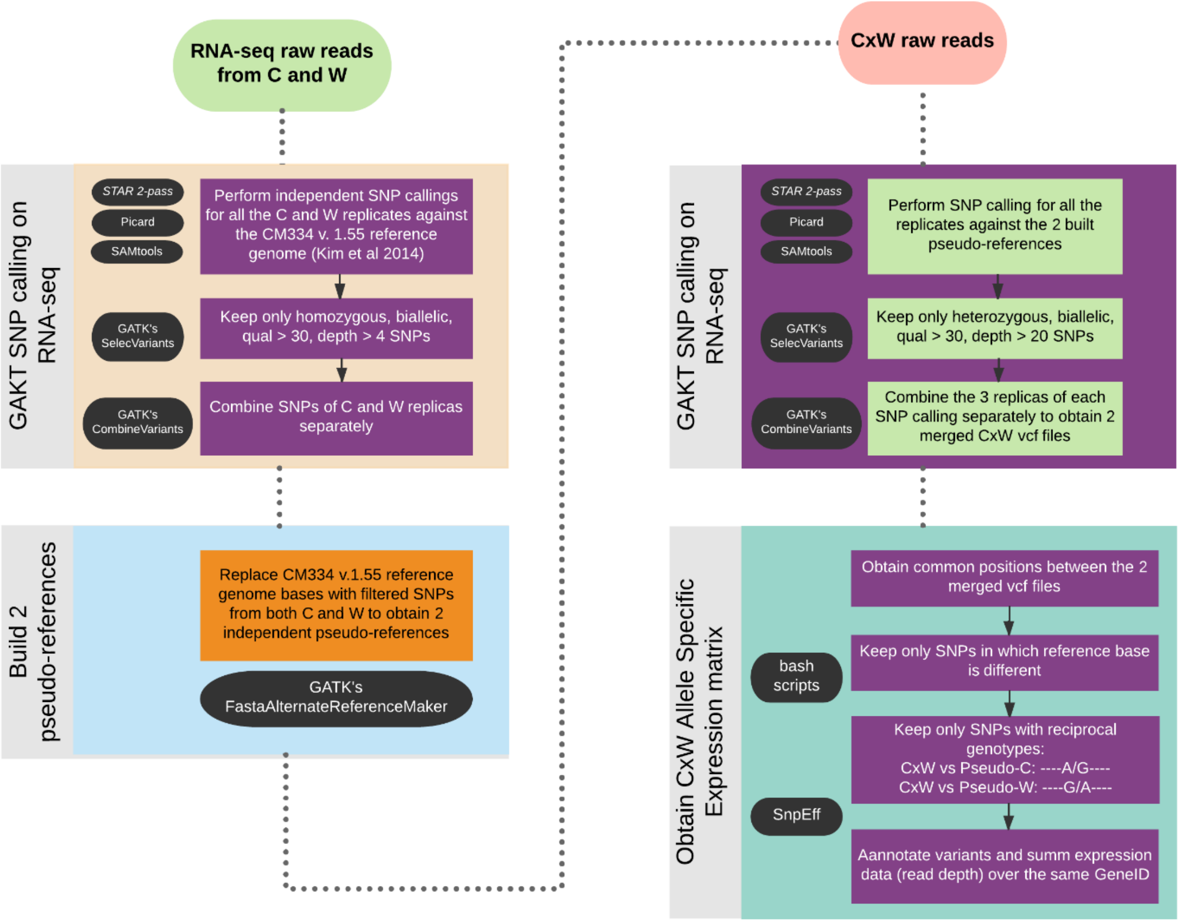

RNA-seq raw reads from C and W
CxW raw reads
GAKT SNP calling on RNA-seq
STAR 2-pass
Picard
SAMtools
Perform independent SNP callings for all the C and W replicates against the CM334 v. 1.55 reference genome (Kim et al 2014)
GATK's SelecVariants
Keep only homozygous, biallelic, qual > 30, depth > 4 SNPs
GATK's CombineVariants
Combine SNPs of C and W replicas separately
Build 2 pseudo-references
Replace CM334 v.1.55 reference genome bases with filtered SNPs from both C and W to obtain 2 independent pseudo-references
GATK's FastaAlternateReferenceMaker
GAKT SNP calling on RNA-seq
STAR 2-pass
Picard
SAMtools
Perform SNP calling for all the replicates against the 2 built pseudo-references
GATK's SelecVariants
Keep only heterozygous, biallelic, qual > 30, depth > 20 SNPs
GATK's CombineVariants
Combine the 3 replicas of each SNP calling separately to obtain 2 merged CxW vcf files
Obtain CxW Allele Specific Expression matrix
Obtain common positions between the 2 merged vcf files
bash scripts
Keep only SNPs in which reference base is different
Keep only SNPs with reciprocal genotypes:
CxW vs Pseudo-C: ----A/G----
CxW vs Pseudo-W: ----G/A----
SnpEff
Aannotate variants and summ expression data (read depth) over the same GeneID

**Figure 5.** A homemade pipeline to identify parental-specific alleles in F1 Hybrids (CxW)**.** In the first step, homozygous SNPs from each parental transcriptome (C and W) are identified, filtered and imputed into the reference genome to build 2 pseudo-references. In the second step, the F1 Hybrids are aligned against the two pseudo-references in two independent jobs. After functional annotation and fine filters, the expression of reciprocal genotypes is extracted and summed over all segregating sites that share that same Gene ID. Currently, GATK's pipeline is the most fast and accurate referring to SNP calling in RNA-seq data because it integrates the 'best' current bioinformatic software's: GATK (https://software.broadinstitute.org/gatk/best-practices/ ;Van der Auwera et al 2013) STAR (https://github.com/alexdobin/STAR ; Dobin et al. 2013), Picard (http://broadinstitute.github.io/picard/ ), SAMtools (http://samtools.sourceforge.net/ ), and SnpEff (http://snpeff.sourceforge.net/ ; Cingolani et al. 2012).

I then wrote a bash script to sort genes into seven divergence mechanisms:

- cis-only: significant differential expression between C and W, evidence of cis divergence. No evidence of *trans* effects.
- trans-only: significant differential expression between C and W, evidence of *trans*. No evidence of *cis* divergence.
- cis + trans: significant differential expression between C and W, evidence of *cis* and *trans*. $Log_2$ transformed allele specific ratios have the same sign in parental and hybrids. Regulation of these genes has diverged such that *cis* and *trans* regulatory changes favor the expression of the same allele.
- cis x trans: significant differential expression between C and W, evidence of *cis* and *trans*. $Log_2$ transformed allele specific ratios have the opposite signs in parental and hybrids. Regulation of these genes has diverged such that *cis* and *trans* regulatory changes favor the expression of opposite alleles.
- Compensatory: expression is conserved between parents but not in hybrids (C=W; HybC ≠ HybW). Evidence *cis* and *trans* effects. Regulation of these genes has diverged such that cis and trans regulatory changes perfectly compensate each other, which results in no expression difference between Cultivated and Wild chilies.
- Conserved: expression is conserved between parents and hybrids. These genes are expressed at similar level in each parent as well as in the hybrid, which indicates conserved regulation.
- Ambiguous: all other patterns of significance tests, which have no evident biological interpretation at transcriptional level.

The bash script (**Script-box 3**) intersects sets of genes according to the significance of the mentioned-above statistical tests, as well as the foldchange of expression values.

```
#chi square test for expression divergence between Parents
DE_par <- read.table("Parental_ratio.txt", header = TRUE, sep="\t")
rownames(DE_par) <- DE_par$Gene_ID
DE_matrix <- as.matrix(DE_par[2:3])
head(DE_matrix)
d <- data.frame(PUYA = DE_matrix[,1], CHILT = DE_matrix[,2])
DIFF_EXP <- cbind(d, t(apply(d, 1, function(x) {
  ch <- chisq.test(x)
  c(unname(ch$statistic), ch$p.value)})))
colnames(DIFF_EXP)[3:4] <- c('x-squared', 'p-value')
write.xlsx(DIFF_EXP, "Parent_DE_SNPs.xlsx")
#chi square test for cis regulatory variation
ASE_HYB <- read.table("Hybrid_ratio.txt", header = TRUE, sep="\t")
rownames(ASE_HYB) <- ASE_HYB$Gene_ID
ASE_matrix <- as.matrix(ASE_HYB[2:3])
head(ASE_matrix)
cis <- data.frame(PUYA_HYB = ASE_matrix[,1], CHILT_HYB = ASE_matrix[,2])
cis_test <- cbind(cis, t(apply(cis, 1, function(x) {
  ch <- chisq.test(x)
  c(unname(ch$statistic), ch$p.value)})))
colnames(cis_test)[3:4] <- c('x-squared', 'p-value')
write.xlsx(cis_test, "Hybrid_DE_SNPs.xlsx")
#Fisher exact test to test for trans effects in F1
trans_test <- read.table("trans_eff.txt", header = TRUE, sep="\t")
trans_evidence <- apply(as.matrix(trans_test[,2:5]), 1, function(x)
  fisher.test(matrix(round(x), ncol=2), workspace=1e9)$p.value)
  colnames(trans_evidence)[6] <- c('p-value')
write.xlsx(trans_evidence, "TRANS_SNPs.xlsx")
```

**Script-box 2**. R script written to dissect regulatory mechanisms underlaying expression divergence between C and W. Chi-square and Fisher's exact tests were iteratively executed ($p<0.005$) across ~8,000 genes. Expression values of the parental genes as well as those of both alleles in CxW were $Log_2$-transformed to avoid bias in downstream analysis.

```
#!/bin/bash

#$1=Log2(C)- Log2(W)
#$2=Log2(C.Hyb) - Log2(W.Hyb)

#cis only
awk ‘sqrt($1*$1) >= 1.25 && sqrt($2*$2) >= 1.25\
 && sqrt($1*$1-$2*$2) < 1.25’ > cis_only.txt

#trans only
awk ‘sqrt($1*$1) >= 1.25 && sqrt($2*$2) < 1.25\
 && sqrt($1*$1-$2*$2) >= 1.25’ trans_only.txt

#cis + trans (-)
awk ‘$1 <= -1.25 && $2 <= -1.25\
 && sqrt($1*$1-$2*$2) >= 1.25’ > cis_plus_trans_neg.txt

#cis + trans (+)
awk ‘$1 >= 1.25 && $2 >= 1.25\
 && sqrt($1*$1-$2*$2) >= 1.25’ > cis_plus_trans_pos.txt

#cis x trans (+)
awk ‘$1 >= 1.25 && $2 <= -1.25\
 && sqrt($1*$1-$2*$2) > 1.25’ > cis_by_trans_pos.txt

#cis x trans (-)
awk ‘$1 <= -1.25 && $2 >= 1.25\
 $$ sqrt($1*$1-$2*$2) > 1.25’ > cis_by_trans_neg.txt

#compensatory
awk ‘sqrt($1*$1) < 1.25 && sqrt($2*$2) >= 1.25\
 && sqrt($1*$1-$2*$2) >= 1.25’ > compensatyory.txt

#conserved
awk ‘sqrt($1*$1) < 1.25 && sqrt($2*$2) < 1.25\
 && sqrt($1*$1-$2*$2)’ < 1.25

#ambiguous: all genes lacking a category
```

**Script-box 3**. Bash script used to classify genes according to their divergence mechanism. A foldchange of |1.25| was set as threshold, this value was used because significance ($p<0.005$) was reached at 0.5-fold, which categorizes this analysis as conservative.

All scripts utilized in this study were run in MAZORKA, a high-performance computing cluster at LANGEBIO-CINVESTAV, except by the R scripts, which were run in a desktop computer running ubuntu Linux OS.

# RESULTS

## No reciprocal crosses were obtained between Cultivated and Wild chili accessions

The F1 progeny derived from the reciprocal crosses performed between C and W showed unidirectional fertility as only the seeds labeled as (♀C x ♂W) could germinate, while the seeds labeled as (♀W x ♂C) showed 0% germination rate, nevertheless, many of the seeds were sterile as no embryo was observed in the seeds. This unilateral incompatibility could be result of pollen grain too large in the cultivated parent, or genetic variation that affect self-incompatibility systems (Onus & Pickersgill, 2004).

## Fruit morphometric analysis

As a first approach in the understanding of fruit morphological divergence between cultivated and wild chilies, I characterized the shape of C, W, and CxW fruits by means of quantitative image analysis. Nevertheless, differences are visually evident, CxW fruit size is biased towards W (**Figure 6**). A principal component analysis showed that differences in shape descriptors between C, W and CxW, account for at least 85.8% of continuous morphometric variability (**Figure 7**). Area, major and minor axes, circularity, and aspect ratio variables conform the PCA1 and showed 0.93, 0.96, 0.91, -0.90, and -0.92 of correlation values respectively, or in other words, the loading values of that variables into the eigenvectors.

As shown in **figure 7**, genotypes cluster into two groups, one includes only the C genotype, while the other includes W and CxW. Area, major axis, aspect ratio, and circularity shape descriptors, were further analyzed and showed significance when testing for differences between C and W, and between C and CxW ($p<0.05$), on the other hand, only aspect ratio showed significance when testing for differences between W and CxW ($p<0.05$).

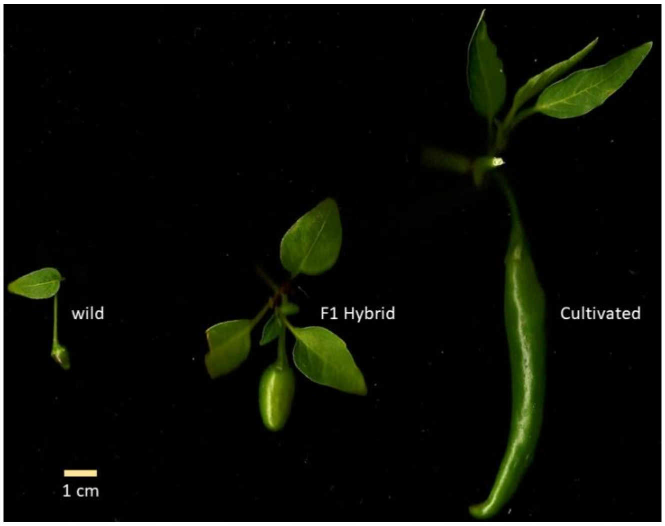


**Figure 6**. Morphological differences between cultivated and wild chilies as well as their F1 hybrid. Images correspond to 40 days after anthesis fruits.

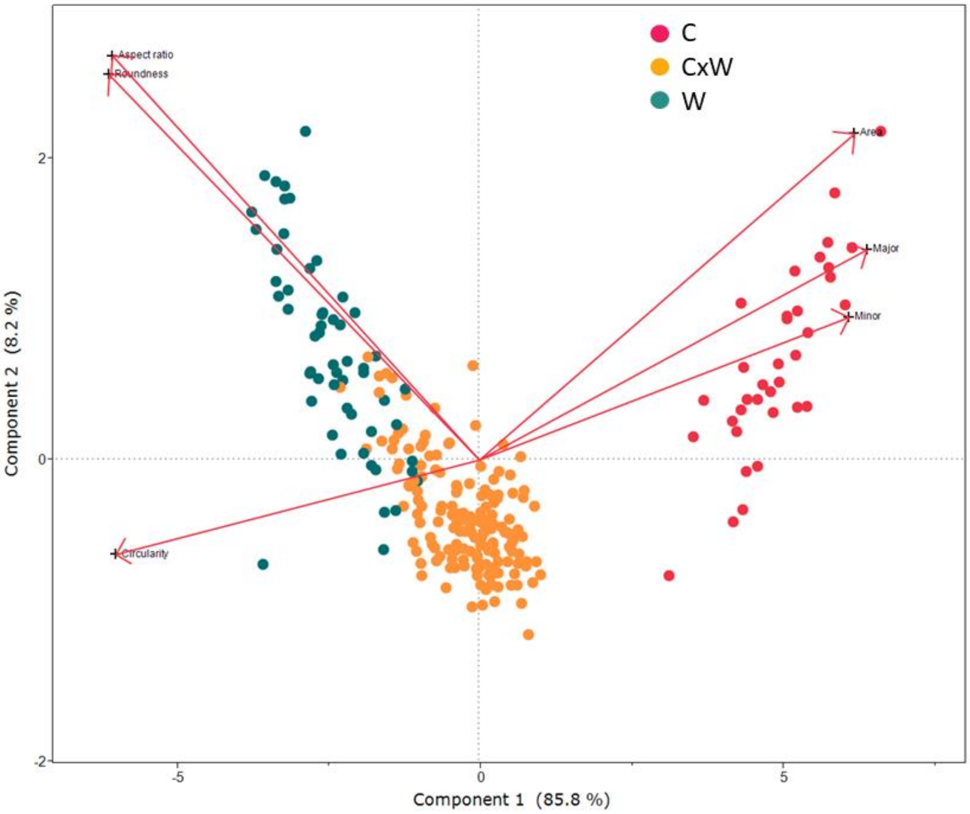


**Figure 7.** PCA from shape descriptors data. Scatterplot of the first against the second principal components of fruit shape descriptors from C (Cultivated), W (Wild) and CxW (F1 Hybrid) genotypes. Each dot represents a single fruit and its color indicates the genotype.

By contrasting this results to the hypotheses that I proposed above (see methods), it was possible to assign dominance or additivity inheritance modes for each of the analyzed shape descriptors as follows: area, major axis and circularity (**Figure 8 A, B, D, respectively**) showed dominant effects of W in F1 hybrids as the mean of those variables does not show significant differences between W and CxW, whereas aspect ratio (**Figure 8-C**) showed additive effects of both C and W on CxW, as the value of CxW is the midpoint between C and W (Fig. n C).

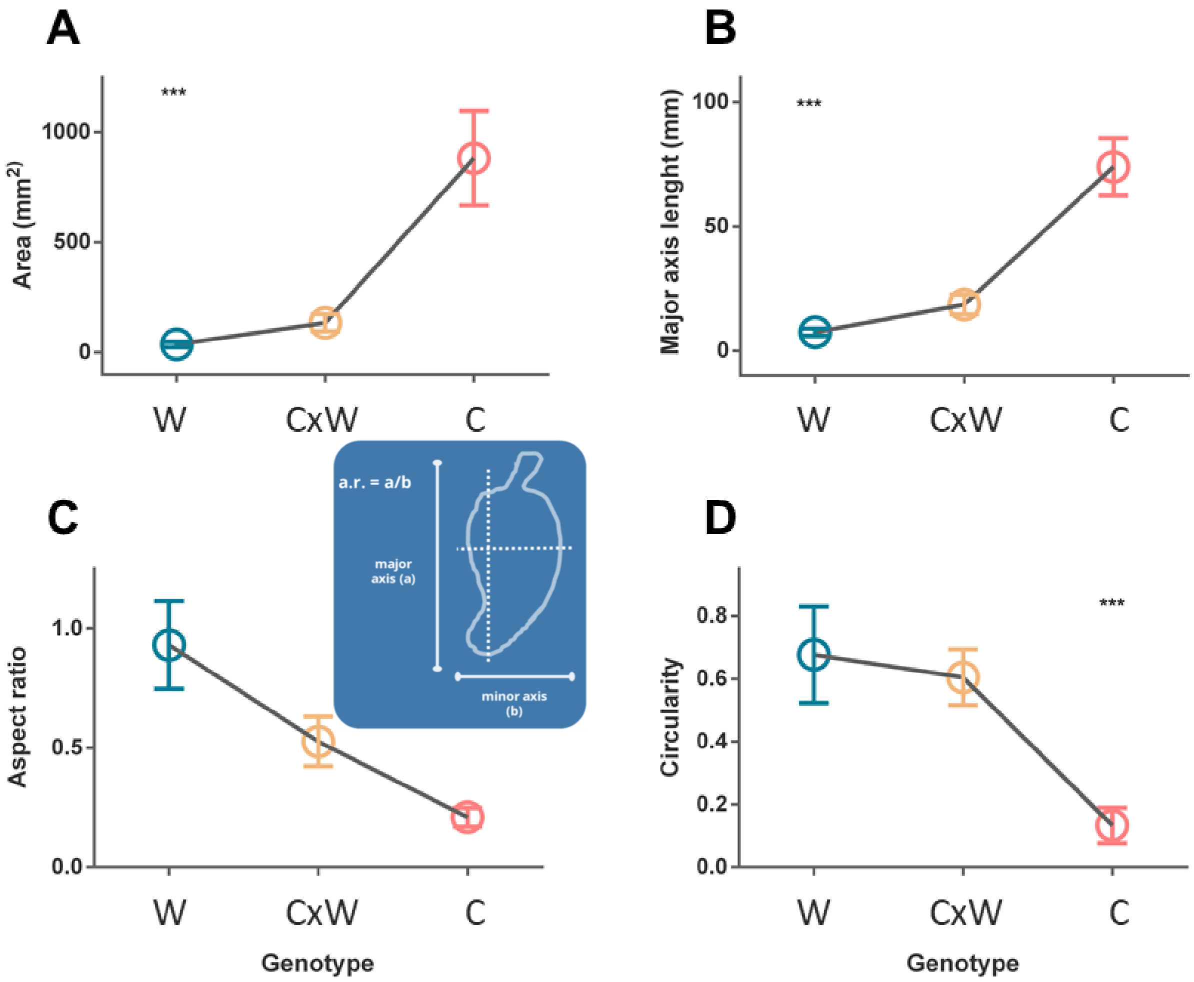


**Figure 8.** Independent analysis for each of the four shape descriptors extracted from PCA1. (**A-D**) Th mean and standard deviation were plotted for each shape descriptor in each of the genotypes. *** indicates significant differences between C and W ($p<0.05$). Blue box specifies the shape descriptor aspect ratio (a.r.).

## RNA-seq yield and quality controls

In this study nine cDNA libraries were sequenced under the Illumina Hi-seq 4000 platform, the average yield obtained for the three biological replicates of each genotype was 3.21 Gbp for the C transcriptomes, 2.78 Gbp for the W transcriptomes, and 2.6 Gbp for the CxW transcriptomes. After quality filtering (LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:15 MINLEN:70), an average of 91% of pair-end reads survived (**Table 1**). FastQC reports showed that after applying filters the mean length was 100 bp, Phred score was above 30, and no adapter content was detected.

## Kallisto's pseudo-alignment recovered more reads than conventional sequence alignment software's

As I mentioned above, one of the bioinformatic challenges of working with non-model organisms is that alignments of two or more different genotypes against the same reference, are usually biased towards the less divergent of the genotypes when compared to the reference genome/transcriptome. By using Kallisto, which is based on perfect *k*-mer matching between indexed reference and samples, I obtained a higher and non-biased averaged-percentage of aligned reads (65%) (see table n) compared to that obtained with a conventional sequence alignment software (Bowtie2) with default settings, which gave an average of 60% of aligned reads for W and 78% for C.

## Expression divergence between Cultivated and Wild chilies

Testing the hypothesis of expression divergence between C and W was achieved by means of using the Sleuth R package, which output is a matrix that includes Gene ID, foldchange, *q*-value, and *p*-value data. However, before exploring differentially expressed genes as well as direction and magnitude of changes, some quality control analyses most be performed to assess for technical bias due to the nature of the reference transcriptome or wrong sample labeling.

**Table 1.** Illumina Hi-seq 4000 reads before and after applying filters and perform transcriptome pseudo-alignments with Kallisto software.

| | Cultivated | | | Wild | | | F1 Hybrids | | |
|---|---|---|---|---|---|---|---|---|---|
| | C_1 | C_2 | C_3 | W_1 | W_2 | W_3 | CxW_1 | CxW_2 | CxW_3 |
| Raw reads (millions)* | 15.1 | 14.5 | 16.8 | 15.4 | 12.3 | 13.7 | 12.1 | 12.2 | 15.6 |
| Filtered reads (millions) | 13.5 | 13.4 | 15.7 | 14.1 | 11.4 | 12.7 | 13.1 | 11.2 | 11.5 |
| Pseudo-aligned reads (millions) | 10.0 | 9.8 | 11.9 | 9.9 | 7.6 | 9.3 | 9.5 | 6.8 | 8.3 |
| Portion of pseudo-aligned reads (%) | 66.0 | 68.0 | 71.0 | 64.0 | 62.0 | 68.0 | 79.0 | 55.0 | 53.0 |
| Average per sample (%) | | 68.3 | | | 64.7 | | | 62.3 | |

* Illumina paired-end reads

For instance, in **Figure 9-A**, the expression values (est_counts) are plotted against the foldchange of W/C, and true significant differentially expressed genes (red dots) showed to be symmetrically distributed in the plot, which suggest no bias in the transcript pseudo-alignment nor in the quantification. Also, in **Figure 9-B** color intensity reflects Jensen-Shannon divergence, which resulted to be smaller within biological replicates of each genotype than that observed between genotypes ($\chi^2$ test, $p<0.05$). PCA (**Figure 9-C**) showed two clusters, one for C samples and one for W samples, which serves as measure of within-genotype expression variability, and as a first approach to measure the degree of divergence in the transcriptional landscape of *C. annuum.*

Sleuth's algorithm is highly accurate because it leverages the bootstrap estimates of Kallisto and the variance across the independent sets of samples to be analyzed, but also because filters out all genes whose expression was < 4 'est_counts'. Therefore, no bias because of $\text{Log}_2$ transformation is expected to arise in downstream analysis.

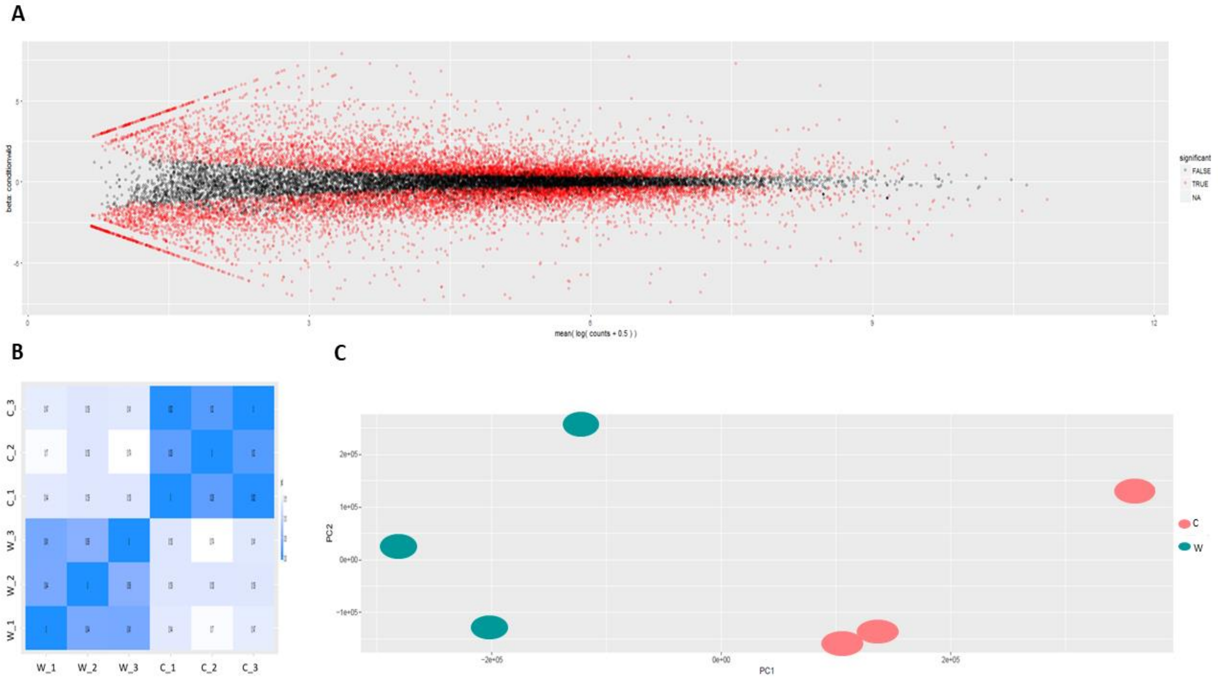


**Figure 9**. Quality control plots. (**A**) MA plot of differential expression analysis, Log of mean gene expression is plotted against foldchanges. Symmetry in MA plots indicates no reference bias when performing quantification. (**B**) Heat map of Jensen-Shannon divergence, darker blue points out higher similarity. (**C**) Principal Component Analysis of expression values of C and W.

In **Figure 10**, the distribution of expression values (est_counts) for C and for W shows them almost overlapping, which suggest mitigated reference bias. However, it can be observed that there is a slightly difference between C and W transcriptomes in density values of genes expressed within $Log_2(2)$ and $Log_2(3)$, which means that Cultivated chili transcriptomes bear more genes expressed at those level that the Wild chili transcriptomes.

A total of 20,456 CM334 reference transcriptome 'genes' survived for both C and W transcriptomes after sleuth normalization and filtering. Of them, 10,468 (51%) showed statistical significance for being differentially expressed ($q$-value < 0.05). Differential expression appears to be symmetric between W and C, as shown in **Figure 11**. However, 5,094 W transcripts show upregulation when compared to C orthologs, while 5374 C transcripts show upregulation when compared to W orthologs. This difference is statistically significant ($Z$ = -2.72, $p$ = 0.003).

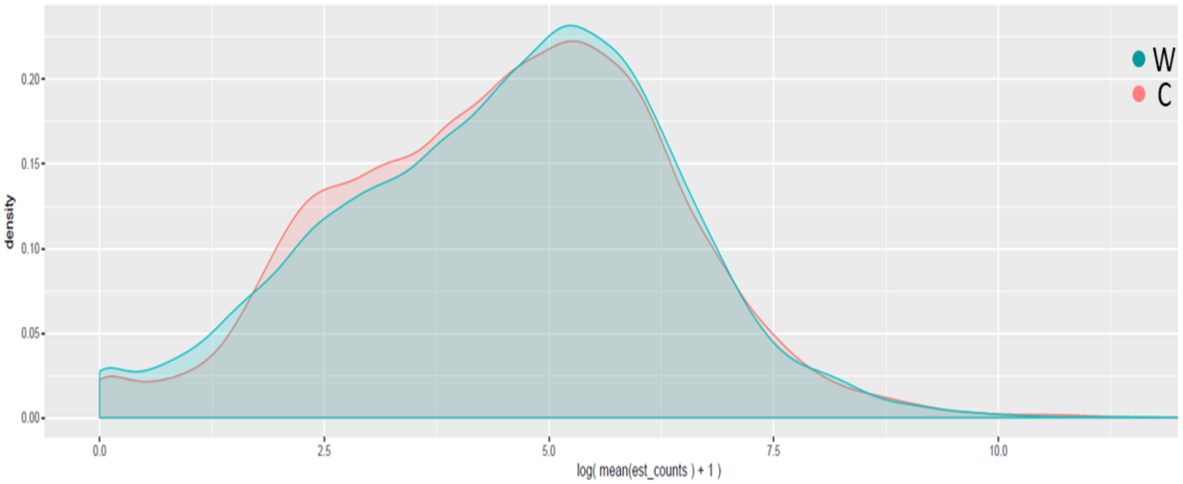


**Figure 10.** Distribution plot of gene expression values for Cultivated (salmon) and Wild (Aqua) transcriptomes.

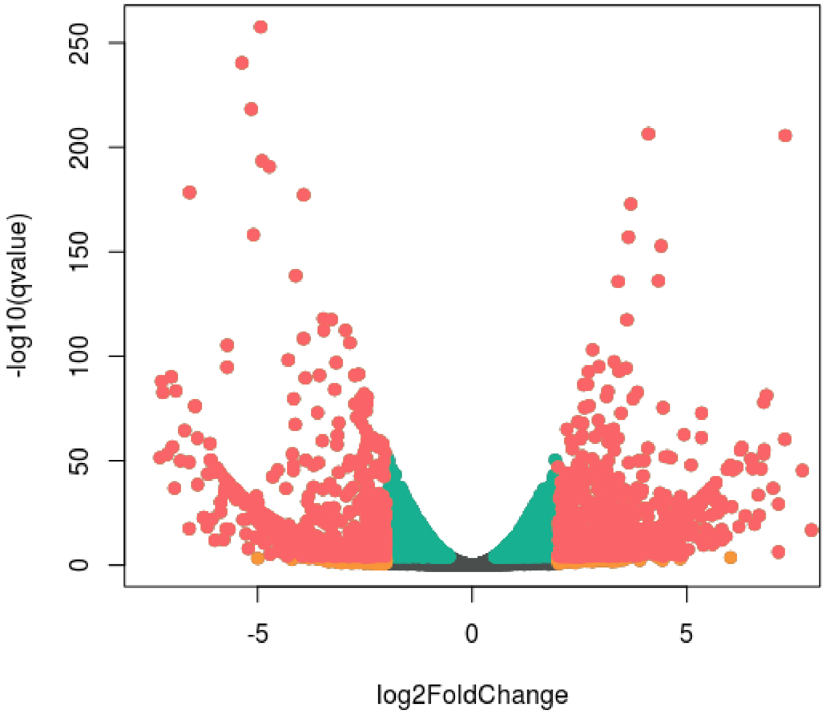


**Figure 11.** Volcano plot of 20,456 genes tested for differential expression between C and W. x-axis shows the foldchange of W/C, while y-axis the -log transformed *q*-values. Gray dots (conserved expression genes) indicate genes whose *q*-values > 0.0001 and absolute value of fold-change '|fold-change|' is < 2. Green dots indicate genes whose *q*-values <= 0.0001 and |fold-change| < 2. Yellow dots show genes whose *q*-values > 0.0001 and |fold-change| >= 2. Finally, salmon dots represent 'confident' differentially expressed genes, their *q*-values <= 0.0001 and |fold-change| >= 2.

Data in **Figure 12**, suggest a slightly bias towards having more upregulated genes in C transcriptomes when compared to W transcriptomes within a range of |0.5-2| fold-change, however, this difference failed to be statistically significant ($\chi^2 = 0.029$, DF = 1, $p = 0.86$).

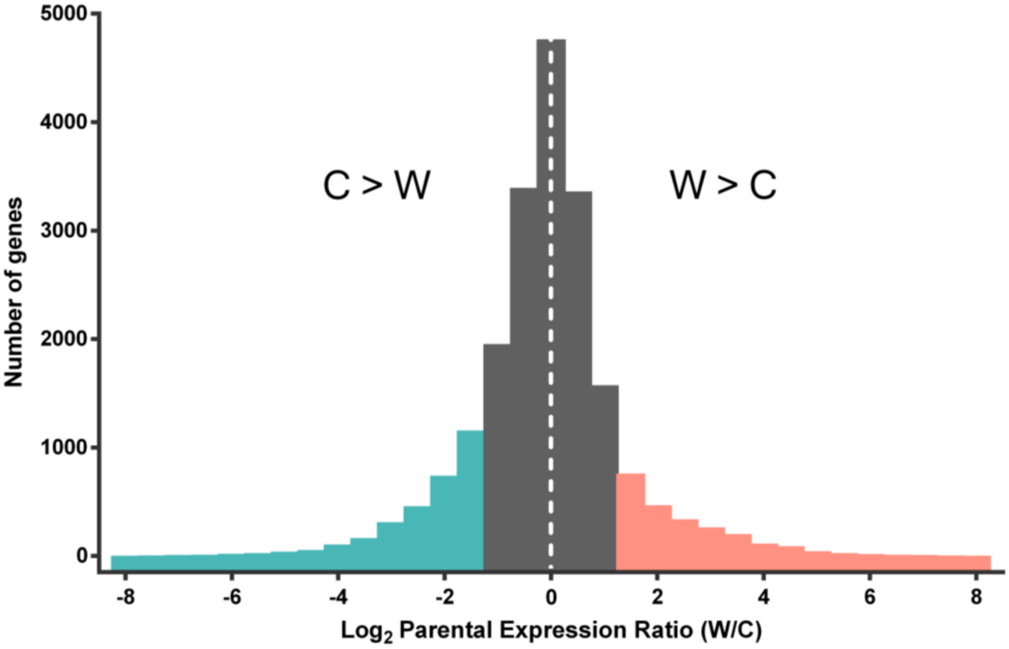


**Figure 11**. Differences in gene expression between Wild and Cultivated chilies. Histogram shows the direction and magnitude of expression divergence between W and C. Blue bars refer to genes whose expression is higher in C when compared to W. Salmon bars refer to genes whose expression is higher in W when compared to C.

## Overrepresentation of Gene Ontology terms and clustering of differentially co-expressed genes

The set of differentially expressed genes at q<0.05 was subjected to a statistical overrepresentation test (p<0.05) of GO terms, which considers the deviation between observed and expected number of genes for each GO term based on a reference transcriptome (see methods). As shown in **Figure 13 A-B**, Biological process is the more enriched category with 87% of the total overrepresented GO terms, followed by Cellular component with 7.3% and Molecular function with 5.7%.

After having scrutinized the most overrepresented GO terms of Biological process category (smaller *p*-values), those related to responses against biotic and abiotic stresses, dynamics of transcriptional landscape, as well as fruit and reproductive system development, were the most prominent. Additionally, antioxidant and oxidoreductase activity were the most highly overrepresented GO terms for Molecular function and Cellular component categories, respectively (**Figure 13-C**).

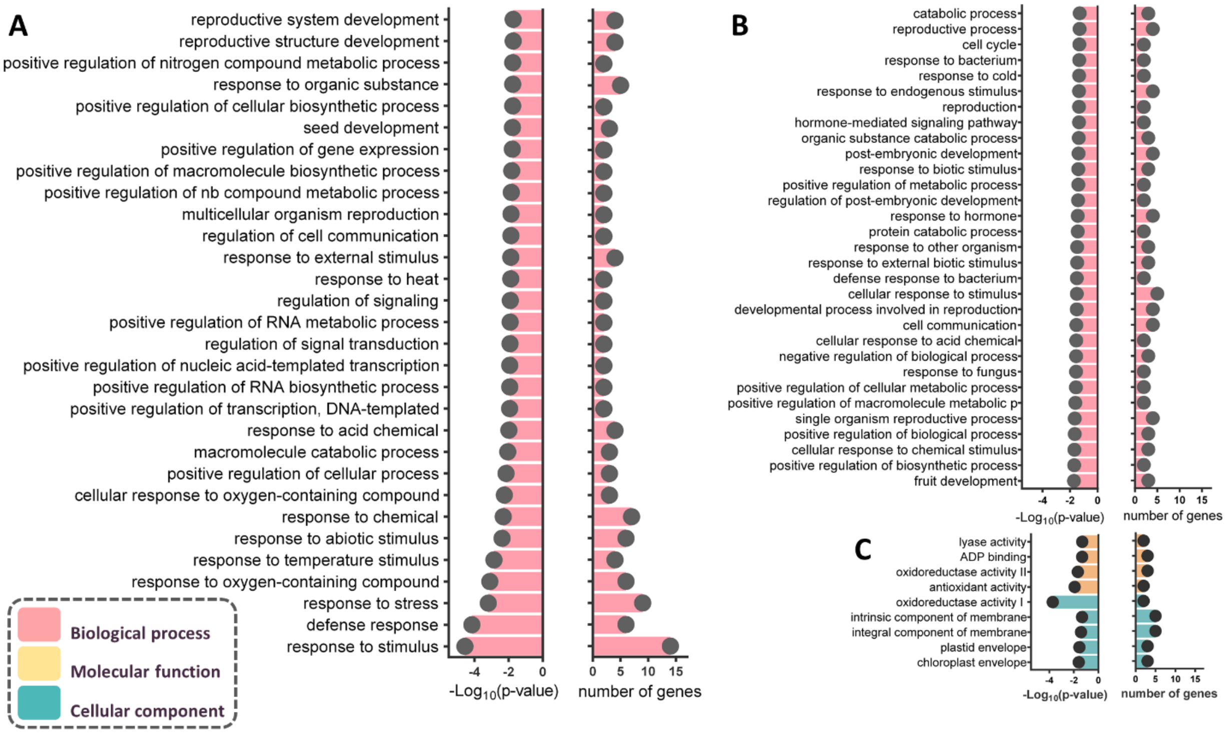


**Figure 13**. GO term overrepresentation test for differentially expressed genes between C and W transcriptomes. Each panel contains information about the significance of the overrepresentation (*p*-value) as well as the number of genes that support those *p*-values. (**A-B**) show GO terms belonging to Biological process category (salmon). (**C**) GO terms overrepresented for Molecular function (yellow) and Cellular component (aqua) categories.

The top 100 differentially expressed genes were examined to identify co-expressed genes by means of Euclidean distance clustering, but also to achieve a deeper examination of the magnitude and direction of the expression divergence between Wild and Cultivated *C. annuum* fruits.

For instance, the two purple clusters that correspond to heat-shock proteins in **Figure 14-A** denote genes that are co-expressed towards upregulation in the C transcriptomes. On the other hand, capsaicinoids biosynthesis genes (**Figure 14-B**) are grouped in two clusters, one of them comprising *Kas I*, *PAL* and *Fat A* genes, whose are co-expressed towards downregulation also in the C transcriptomes.

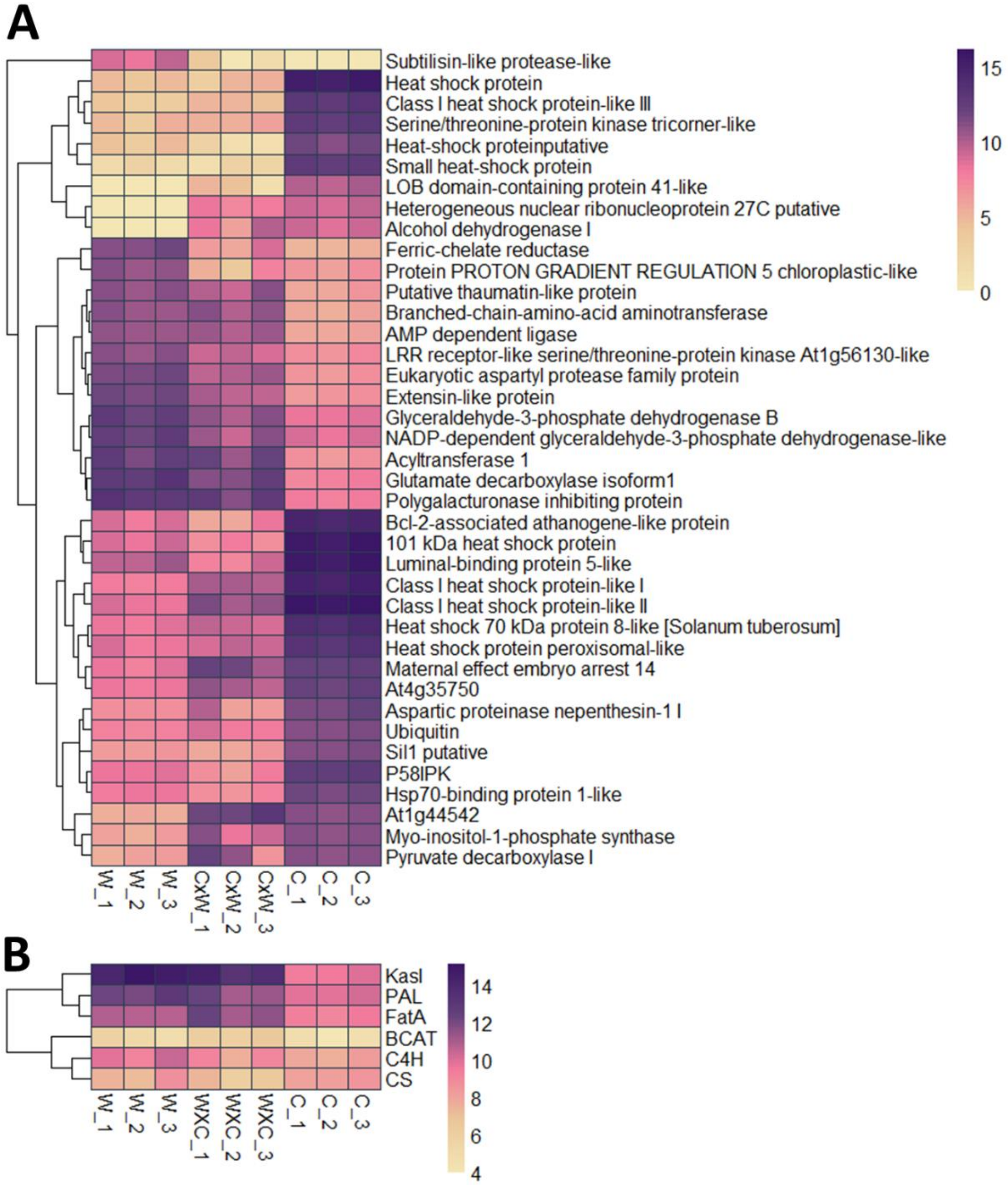


**Figure 14**. Heat map of Log$_2$ normalized 'est_counts' data from the W, CxW, and C transcriptomes. Clusters reflect Euclidean distance between pairwise comparisons. (**A**) Shows data from the first half of top 100 more differentially expressed genes between the C and W transcriptomes. (**B**) Shows Capsaicinoids biosynthesis genes (gene ID's were taken from Kim et al. 2014).

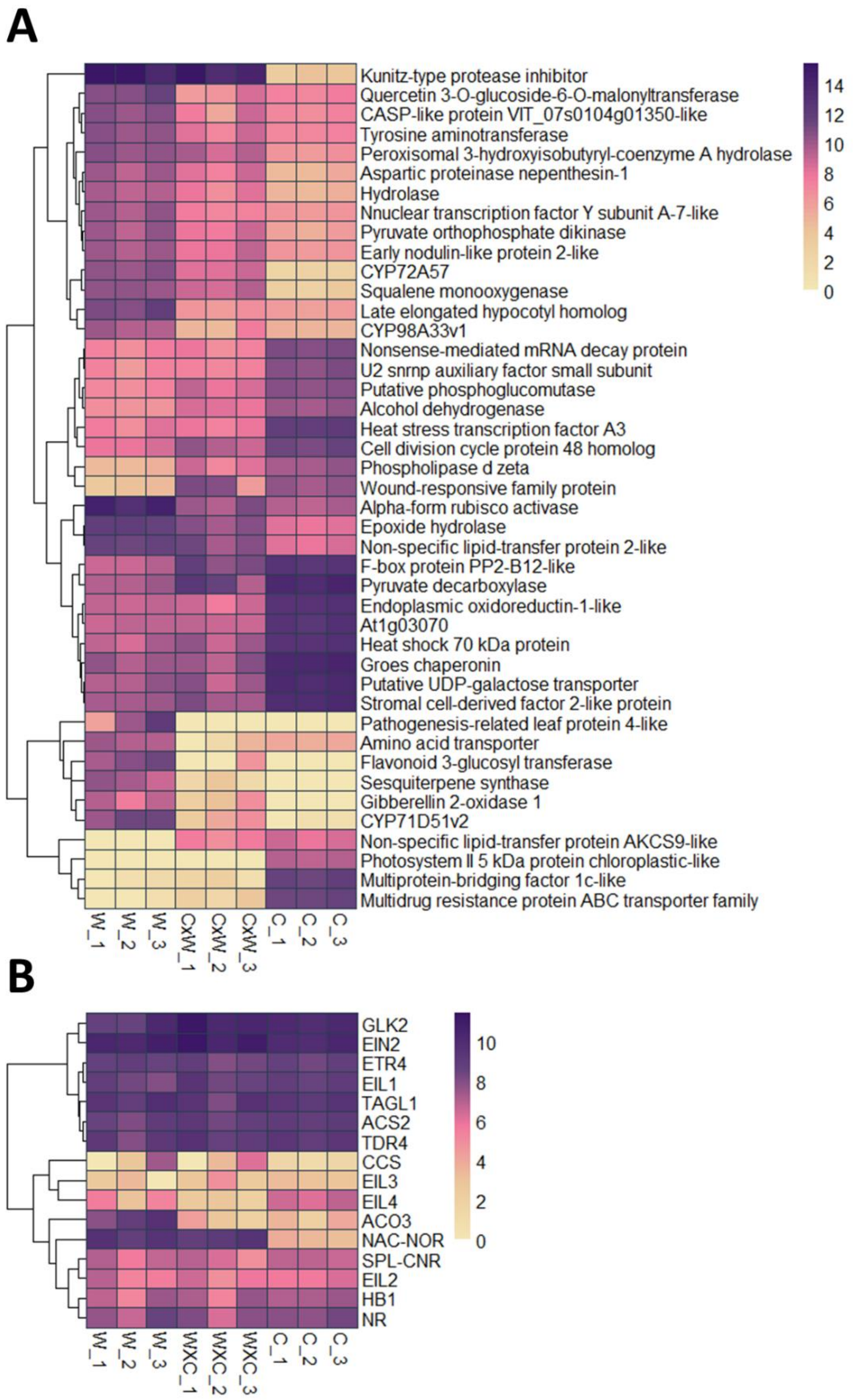


**Figure 15**. Heat map of Log$_2$ normalized 'est_counts' data from the W, CxW, and C transcriptomes. Clusters reflect Euclidean distance between pairwise comparisons. (**A**) Shows data from the second half of top 100 more differentially expressed genes between the C and W transcriptomes. (**B**) Shows fruit ripening genes (gene ID's were taken from Kim et al. 2014).

In the second half of the top 100 differentially expressed genes (**Figure 15-A**), a 'block' of co-expressed C genes towards downregulation when compared to their expression in W genes, can be observed; pathogenesis related protein, flavonoid transferase, and gibberellin-2 transferase are the most dramatic instances of this block, supported by a fold-change > 10. In **Figure 15-B** two clusters were constructed, one includes genes that are highly co-expressed in C, CxW and W transcriptomes (purple), and other that includes fruit ripening genes with no clear expression patterns between C, CxW and W transcriptomes. The one exception is the NAC-NOR gene, whose expression is towards downregulation in C.

Data in the heatmaps also function as a proxy of the inheritance mode of C and W genes on the CxW hybrid. For example, a Kunitz-type protease inhibitor in **Figure 14-A** shows 14-fold of $Log_2$ transformed expression in W and CxW when compared to C, which suggest dominant inheritance. These are only a few of many examples of expression inheritance patterns that are worth examining, but a deeper discussion of these results is outside the scope of this thesis.

## Modes of expression inheritance of Cultivated and Wild chilies into their F1 hybrid

To assess for the mode and direction of parental inheritance to the F1 hybrids, a series of hierarchical classifications according to the expression deviation between hybrids and parental transcriptomes was carried out (**Script-box 1**). This classification considers a set of 20,468 genes whose expression in both, parental and F1 hybrids have at least four reads that support their expression. This criterion, plus the fact that significant *q*-values were obtained in genes whose $Log_2$ transformed-expression foldchange was smaller than two (see **Figure 11**), resulted in a reduction of genes classified as divergent, from 10,468 to 6,633.

Of the 6,633 classified as differentially expressed between C and W, 82% showed dominant effects in CxW. In 55% of these cases, W showed dominance over C, whereas in 45% of the cases C showed dominance over W (**Figure 16**). This difference indicates a clear excess of dominance of W genes in CxW ($Z = 8.4$, $p < 000001$).

309 genes showed additive effects, of those, 33% are cases where C had higher expression than W, and 67% for the opposite case.

The definition of transgressiveness in gene expression is a mode of inheritance where the hybrid expression is higher or smaller than that observed in both parents. 861 genes showed transgressive inheritance mode. Of them, 487 cases correspond to transgressivity upward, while 374 showed transgressivity downward, suggesting a pattern of transgressive preferentially upregulated ($Z = 3.81$, $p = 0.00006$).

.

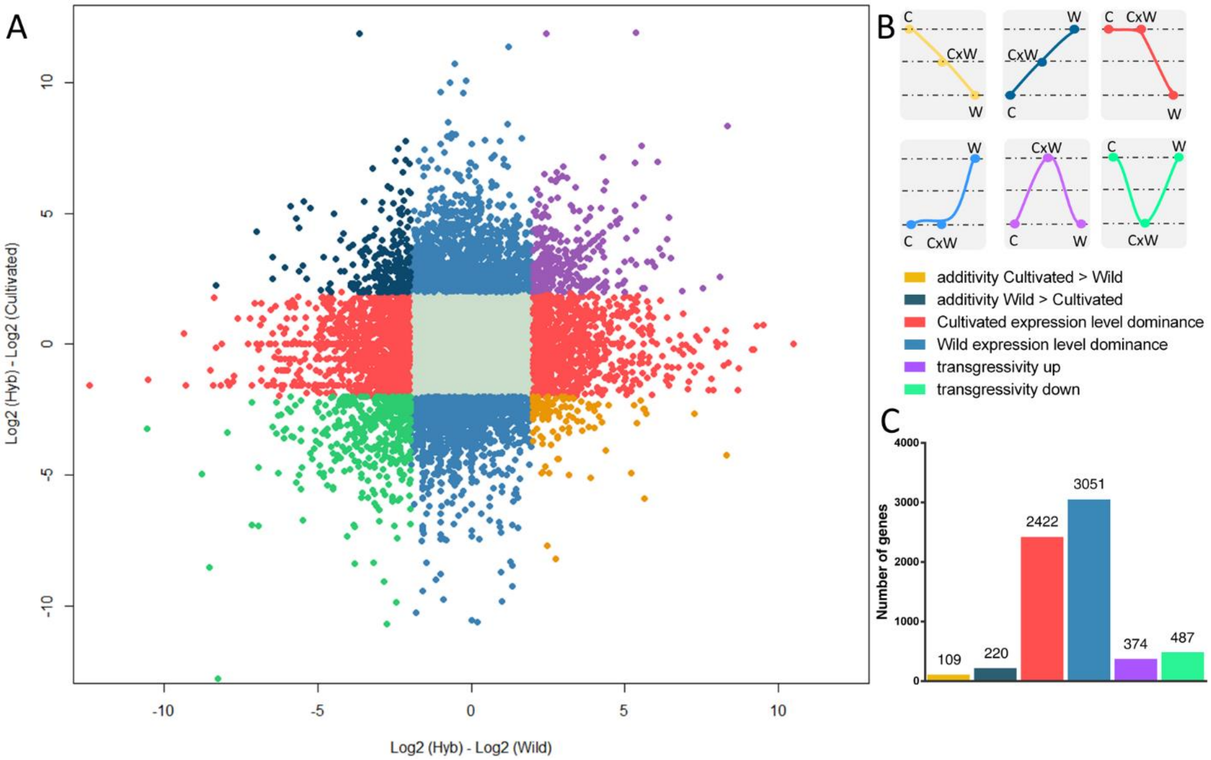


**Figure 16**. Inheritance mode of gene expression in F1 hybrids (CxW). (**A**) Shows the gene expression deviation between hybrid and parental genotypes for each of 20,468 genes whose expression was quantified in the 'C-CxW-W' trio. x-axis shows the expression deviation of CxW against W while y-axis axis the deviation between CxW and C. Given the data transformation, changes of $Log_2|10|$ implies a 1024-foldchange. (**B**) Shows hypothetical patterns of expression in parental (C and W) and their hybrid (CXW), each color corresponds to an inheritance mode such as additivity, dominance or transgressivity, gray color pinpoints to genes with conserved expression (see methods to further explanation). (**C**) Bar plots show the number of genes classified in each of the inheritance modes.

Genes exhibiting dominant mode of expression for both, Wild and Cultivated chilies into their Hybrid, were loaded in the Biological process GO terms: response to stimulus, developmental process, cellular process, metabolic process, biological regulation, cellular component organization or biogenesis, and localization. While reproduction, multicellular organismal process, and immune system process, were found as unique GO terms for genes where C is dominant over W.

### Identification of parental-specific alleles in F1 hybrid

The main challenge of performing Allele Specific Expression Analysis (ASE) from transcriptomic data, is to recognize each of the two parental alleles in the F1 hybrid. However, by means of building two parental-specific pseudo-references (see methods), the C and W specific alleles were identified in CxW.

After aligning the CxW transcriptomes against C pseudo-reference, 164,634, 153,435, and 183,290 SNPs were called for each of the biological replicates. Of this total, 290,990 positions were shared. The alignment of CxW against W pseudo-reference yielded 171,722, 158,608 and 191,011 SNPs for each of the biological replicates, whose merge resulted in 47,713 SNPs.

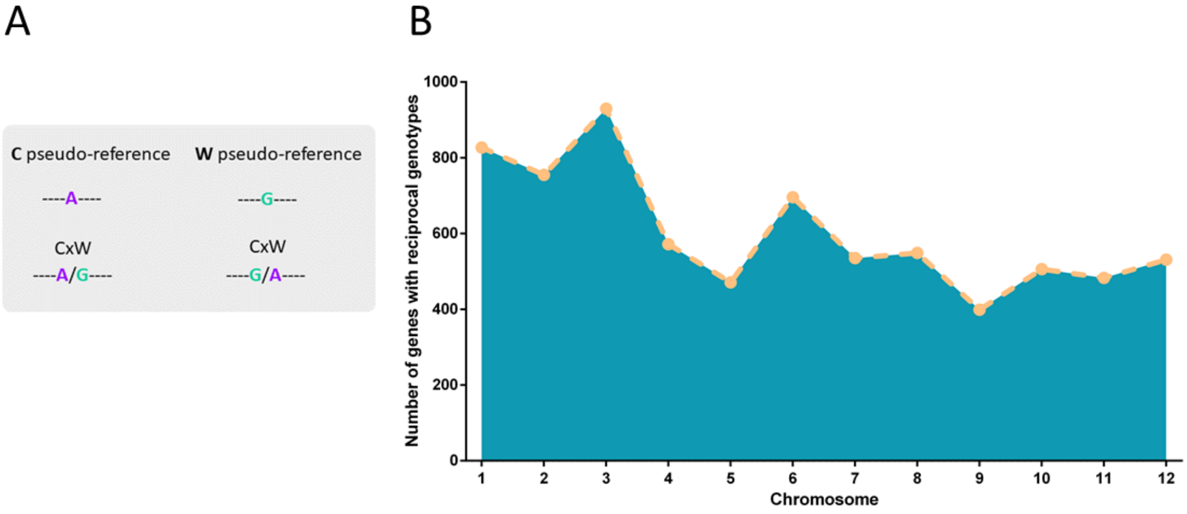


**Figure 17.** (**A**) shows an example of the reciprocal genotypes obtained from the two independent alignments against each of the two pseudo-references. (**B**) Shows the distribution of number of genes with reciprocal SNPs obtained by chromosome.

These two merged SNP files were compared to obtain common positions whose genotypes were reciprocal. This means that, the reference allele of one position is the alternate allele of the other (**Figure 17-A**). Finally, all SNPs related to the same gene ID were summed over all segregating sites and only SNPs supported by the three biological replicates were kept, and then concatenated by chromosome (**Figure 17-B**). The final allele-specific matrix contained 7254 gene IDs as well as the expression values for both C and W alleles in CxW.

### Allele Specific Expression Analysis and regulatory divergence assignment

Dissecting the underlying genetics of differential expression between cultivated plants and their wild relatives can be achieved by means of testing for *cis* and *trans* regulatory divergence (see methods).

In this study, the regulatory mechanisms that led to gene expression variation between cultivated (C) and wild chilies (W) were dissected for a set of 7254 stringently filtered-genes (**Figure 18-A**). Of them, 1734 were classified as having conserved expression between C and W. From the non-conserved expression gene set of 4023, 2513 (44%) were categorized as having cis-regulatory variation, while 3878 (66%) showed *trans* variation. Also, 2368 genes (41%) showed evidence of both, *cis* and *trans* divergence.

Given the direction and the magnitude of gene expression divergence between parents, and between their alleles in a F1 hybrid, one can dissect genes that show either cis-only or trans-only divergence; or genes whose expression differences resulted from effects of both, *cis* and *trans* effects, either to favor the expression of one ortholog or the other (see methods to further description).

Deeper characterization of the regulatory mechanisms (**Figure 18-B**), showed that 145 genes (2.5%) experienced cis-only divergence; 1510 (26%) trans-only divergence; whereas 113 (1.9%) cis + trans; and 130 (2.25%) cis x trans divergence.

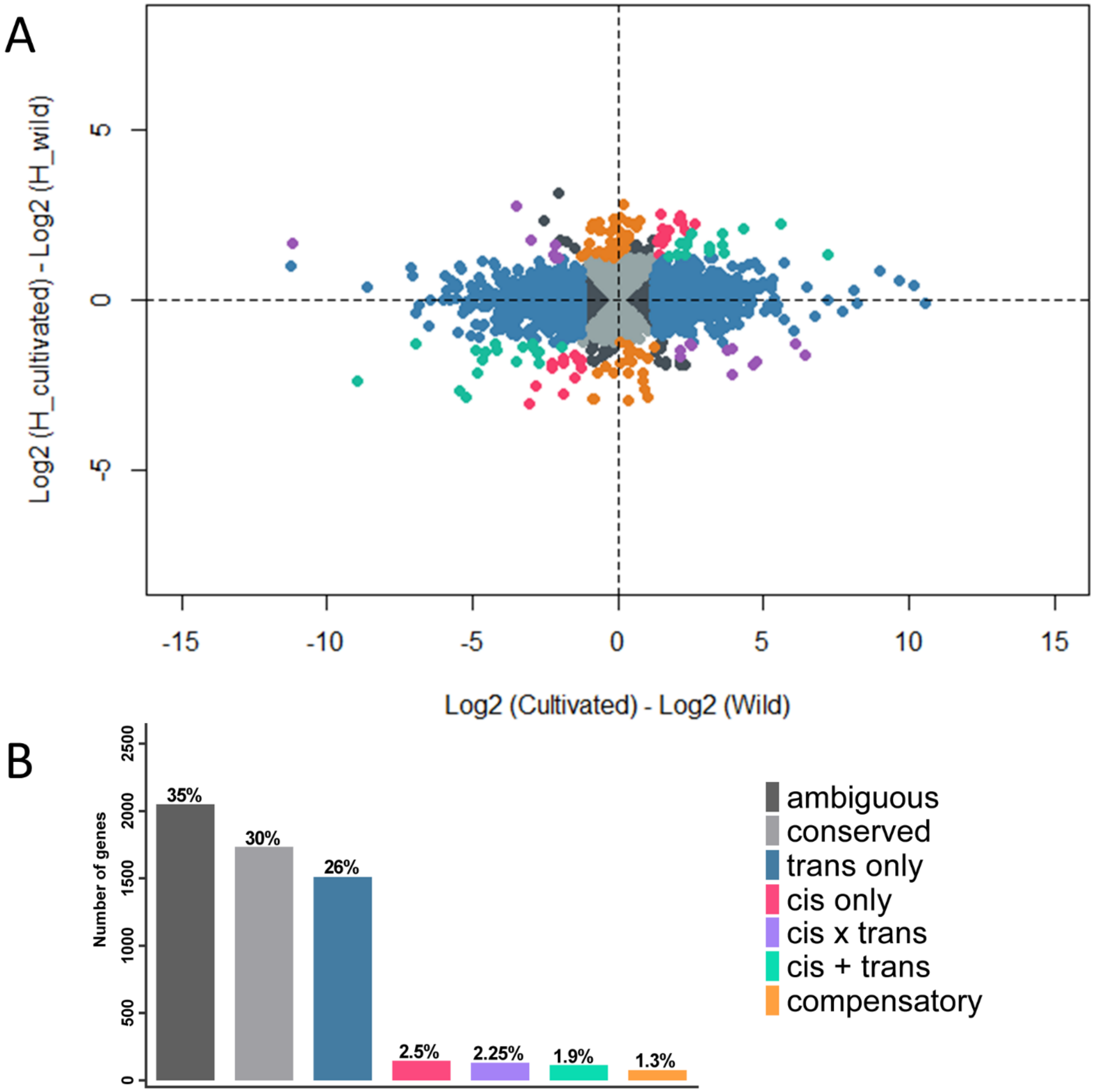


**Figure 18.** Dissection of the regulatory mechanisms underlying transcriptional divergence between wild and cultivated chilies. (**A**) The parental (x-axis) versus F1 hybrid (y-axis) allele-specific expression ratios of 7254 genes are plotted against each other. Color indicates each of the seven regulatory divergence categories that were assigned according to different patterns of significance for the hierarchical statistical tests (see methods). (**B**) Barplot that shows the proportion of genes falling in each of the regulatory categories as well as the actual value.

Genes that showed conserved expression in parents but differential expression in hybrids were classified as compensatory, while genes that showed statistical significance for the tests, but no clear biological explanation at transcriptional level were classified as ambiguous. In this study 75 (1.3%) genes revealed compensatory divergence and a substantial proportion, 2050 (35%), showed ambiguous divergence.

## Implications of the regulatory divergence mode in gene expression inheritance modes

Understanding how the regulatory changes that alter gene expression are inherited, can help to predict their impacts on molecular phenotypes, but also in generating a deeper knowledge about the nature of these regulatory changes (i.e. whether they are dominant-recessive or codominant).

In this study, data sets from inheritance mode of expression and regulatory divergence classification were intersected to investigate: 1) what proportion of the cis-only and trans-only regulatory changes are inherited additively or non-additively (dominant-recessive), and 2) whether are the *cis* or the *trans* regulatory changes more likely to be inherited additively or non-additively.

It was found that, of 145 genes categorized as cis-only, 12% showed dominant effects in gene expression in CxW. Of this subset, 62% of the cases correspond to C dominant over W, while 38% of the cases showed the opposite. Of the cis-only genes, the 2% showed additivity for cases where W expression > C expression, though, none of them showed additivity when C expression > W expression.

Also, in terms of the 1509 genes classified as trans-only divergence, 19% showed dominant effects. Of those, in 71% of the cases W was dominant over C, while in 29% C showed dominance of over W. 3% of the genes classified as trans-only showed also additive effects in cases where W expression was higher than C's, while 0.1%. showed the opposite.

All this data together indicates: 1) that most of the cis-only regulatory changes that altered gene expression between wild and cultivated chilies arose from non-additive mutations, 2) that, a considerable fraction of the trans-only genes that showed non-additive effects (1/4), could have arisen from recessive alleles, as the wild alleles were dominant over the cultivated.

We observed that none of the genes categorized as cis-only showed transgressive inheritance mode. The 3% of the genes classified as trans-only, were also classified as transgressive upwards, while 0.2% transgressive downwards. These data suggest that trans-acting variation have stronger impacts than cis-acting variation in gene networks, as they offer pleiotropic effects to cause transgressive expression variation.

## Enriched GO terms in each of the regulatory divergence mechanisms

Genes belonging to all the regulatory divergence classes (except by conserved and ambiguous), were examined to identify biological processes related to their functional annotation (**Figure 19**).

The GO terms localization, cellular process, and metabolism were identified in all the divergence classes, each in relatively high proportion (<20%). It was notable that the GO terms developmental process, and reproduction were exclusively identified in the trans-only genes. On the other hand, the GO term response to stimulus was identified in all the divergence classes but not in the cis-only class.

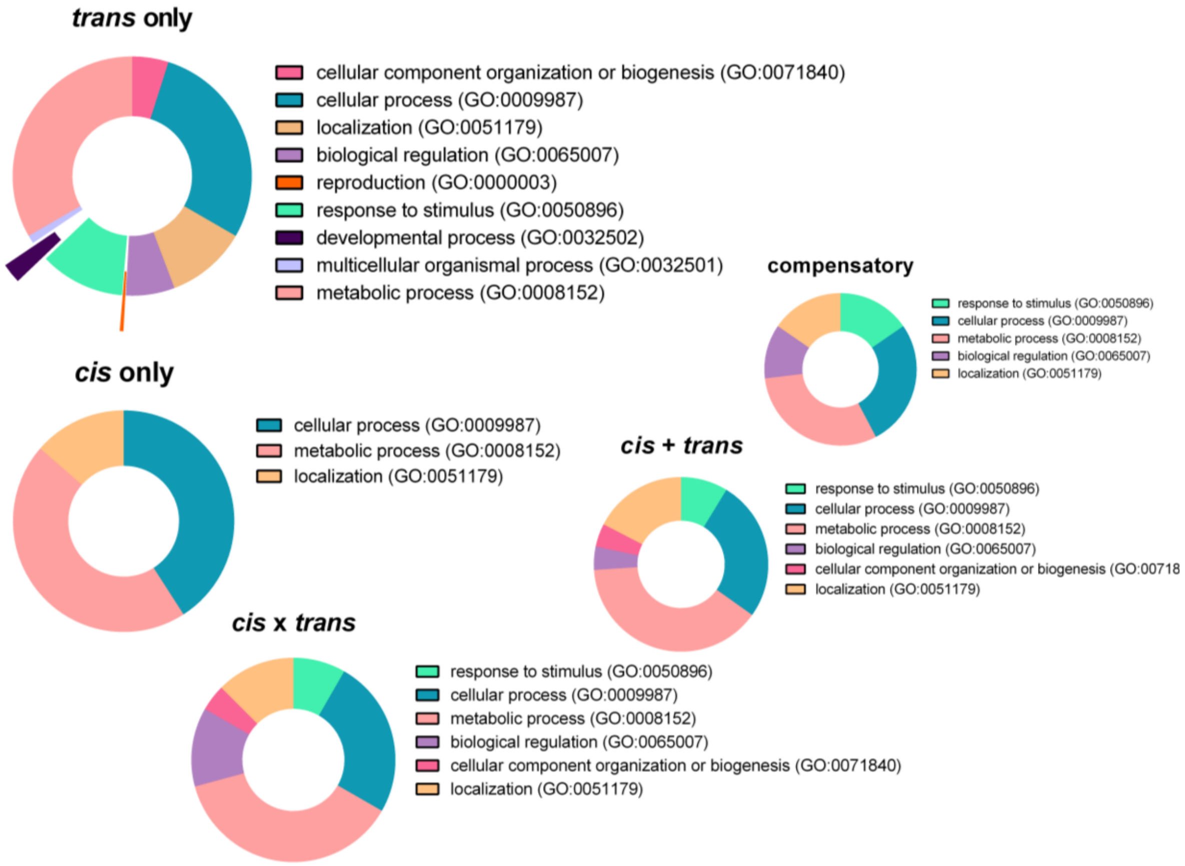


**Figure 19.** GO terms loaded into the regulatory mechanisms underlying expression divergence between wild and cultivated chilies. Each plot shows the GO terms identified in each divergence mode. Color key is GO term-specific.

# DISCUSSION

Given quick evolutionary timeframe in which genetic changes occur during plant domestication, this is a great model to understand how plant morpho-physiology is shaped by changes in gene expression via mutations in regulatory regions.

Data from this study showed that 51% of the shared genes between wild and cultivated *Capsicum annuum* fruits showed expression divergence. This expression divergence appears to be biased towards overexpression in cultivated chilies, and in genes that can be classified in biological processes such as response to biotic and abiotic stresses, and fruit development.

By means of a custom-made bioinformatic pipeline it was possible to dissect the regulatory mechanisms underlying this expression divergence, and it was found that trans-acting variation accounts for 66% of the expression divergence, while cis-acting variation accounts for 44%. Furthermore, it was shown that most of the cis-only variation resulted from non-additive mutations, while the fraction of the trans-only divergent genes that showed non-additive effects (1/4), could have arisen from recessive mutations. In addition, genes whose divergence during *C. annuum* domestication were product of trans-only mutations, are also uniquely enriched in developmental and reproduction biological processes GO terms, which include genes such as ANNEXIN, ABR1, and GLK1.

## Divergent expression between wild and cultivated chilies as a consequence of domestication

The proportion of differentially expressed genes between wild and cultivated *C. annuum* was 51%. This result is in the range of expression divergence found in similar comparative transcriptomic studies that measured expression divergence between closely related wild and cultivated tomato (38%) (Koenig et al. 2013), between invasive and native populations of the thistle *Cirsium arvense* (70%) (Bell et al. 2013), between two African *Coffea* subspecies (33%) (Combes et al. 2015), and between wild and cultivated maize (70%) (Lemmon et al. 2014).

However, despite of the proportion of differentially expressed genes (DEG) and the type of expression divergence (domestication or speciation) of the above-mentioned studies, our results converge with them in the enrichment of the following GO terms: cell division, response to biotic and abiotic stresses and pathogen resistance.

Thus, it can be argued that the domestication process of distinct species, involves human-mediated selection on ortholog traits that are underlaid by ortholog regulatory networks. This had been previously hypothesized by Darwin and Vavilov, and formalized time after, as parallel evolution hypothesis (reviewed in Wood et al. 2005). An example of extreme parallel evolution is that reported by Miller and colleagues (2007), they found that the same regulatory mutation that affected pigmentation levels in stikleback fishes and resulted in a color radiation, also caused pigmentation divergence in ancient human populations. Our results suggest that indeed, expression divergence of *Capsicum annuum* could have been shaped by a similar set of selective pressures as species that have undergone rapid and dramatic changes in their distribution and local environmental conditions as a result of domestication (i.e. tomato, coffee, maize) or rapid invasion (i.e. *Cirsium arvense*), and that these processes impacted the modulation of homologous regulatory networks.

To support this hypothesis, a set of co-expressed heat shock proteins were found in the top 100 of DEG. These proteins exhibit up-regulation in cultivated *C. annuum*, and relatively high expression in wild *C. annuum* (**Figures 13-14**), which could be result of a hyper-sensitivity to heat stress driven by domestication. The same phenomenon was observed between wild heat-tolerant tomato and heat-sensitive cultivated tomato (Bita et al. 2011). Biologically, wild plants exhibit constitutively elevated levels of heat shock proteins because of their need to adapt to various changing environmental conditions, while cultivated plants that tend to be cultivated more homogenously, could have less robust heat stress response machinery, and thus when they face heat stress, exhibit a more dramatic transcriptional response in heat shock proteins.

In terms of traits associated to the domestication syndrome in *Capsicum*, pungency and ripening showed patterns congruent with artificial selection rather than environmental adaptation. Although pungency is often a highly selected trait in *Capsicum*, both the wild and the cultivated accessions used in this study have a pungent phenotype. Thus, no dramatic differences were found in the capsaicinoids biosynthesis machinery genes. A notable exception was *Kas1* a member of a gene family in chili (Kim et al. 2014), which was recently shown not to be related to pungency, but to general biotic stress response instead (Arce-Rodríguez & Ochoa-Alejo, 2017). Our study confirms this observation and provides information about the dual roles of duplicated enzymes in the *Capsicum* genome

The *NAC-NOR* transcription factor, showed dominant high expression levels in W and CxW transcriptomes, but under-expression in C transcriptomes (**Figure 14-B**). NAC-NOR gene has been previously reported to control ripening timing in late stages of fruit development of tomato fruits (Martel et al. 2011), additionally, it was reported that the mutant *nac-nor* shows long shelf life (Casals et al. 2012). This evidence, compared to our data, suggests that the delayed ripening observed in cultivated chilies, which was lost in the F1 hybrid (personal observation in this study) could be result of a recessive mutation in a regulatory element upstream the *NAC-NOR* gene.

We also found a Kunitz-type protease inhibitor in the top 100 DEG was, which was highly expressed in wild chilies and downregulated in cultivated chilies, and could be an unintended side-product of intense artificial selection and the domestication syndrome. This gene was characterized in flower development of *Arabidopsis thaliana* and it was found that its downregulation results in defective fruits and sterile seeds (Boex-Fontvielle et al. 2015), which could explain the high abortion rates reported for cultivated *C. annuum* (Wubs et al. 2009).

A thoroughly analysis of each of the 100 genes is required, but outside of the scope of this thesis. Overall, the information obtained from genome-wide transcriptomics analyses provide a fantastic opportunity to study genes as isolated cases and infer how physiology is altered during domestication.

The time and experiments needed to dissect all cases seems prohibitive for this study. Nevertheless, by means of analyses of the more DEG (lowest q-values), one can begin to visualize what functions were dramatically altered in a rapid human-mediated morphophysiological *C. annuum* evolution.

## Main trends in expression divergence explained by *cis* or *trans* regulatory variation in *Capsicum*

Beyond of analyzing the gene-specific transcriptional divergence, the main goal of this study was to dissect the regulatory mechanisms that have shaped the transcriptional landscape through *Capsicum* domestication.

The proportion of expression divergence between wild and cultivated chilies explained by *cis* (44%), and by *trans* (66%) regulatory divergence, is consistent with results of studies that measured expression divergence in relatively short timeframes (< 10000 years), which have found that most of the expression divergence is explained by trans-regulatory variation (Yvert et al. 2003; Tirosh et al. 2009; Lemmon et al. 2014). In contrast to cases where divergence time extend million years, as in fruit fly evolution (Wittkopp et al. 2004; McGregor et al. 2007; McManus et al. 2010).

This is explained because cis-regulatory variation takes more time than *trans* variation to be perceived by selection, as trans-regulatory changes often show pleiotropic effects, that are not deleterious, can be quickly fixed (Carrol, 2008; Wittkopp & Kalay, 2012). Additionally, cis-regulatory variation has been more strongly associated to finely modulate spatiotemporal patterns of gene expression, which can fuel morphological evolution via gene-specific expression divergence without compromising fitness components that could arise from pleiotropic effects.

In the light of this, it can be suggested that, as domestication is an accelerated morphophysiological evolution where both human-mediated selection, and highly homogenous environment quickly fix traits of interest and its underlying genetic variation (Meyer & Purugganan, 2013). An excess of genes whose expression was affected via mutations at important nodes of regulatory networks with pleiotropic effects would be required, and that is what *trans* variation offers.

## Most of the expression divergence between wild and domesticated Capsicum annuum is result of recessive mutations

Plant domestication, in contrast to other processes of evolutionary divergences such as speciation, results in unique genomic features. One of the most remarkable characteristic is that human-mediated selection during domestication, quickly fixes traits of interest despite underlying recessive loss of function (LOF) mutations. This would be unusual outside of a domestication scenario because: 1) recessive mutations usually show low allele frequency in natural populations, and 2) most of recessive mutations result in deleterious phenotypes in natural populations (Lester, 1989).

The fact that: 1) 76% of the genes showing expression divergence where driven by trans-only variation, while the cis-only, cis + trans, cis x trans, and compensatory, mechanisms together explained 24% of the expression divergence; 2) most of the genes (82%) showed dominant effects in gene expression of the F1 hybrids; 3) in most of the cases showing dominant effects, the wild ortholog was dominant over the cultivated ortholog (55% vs 45%); and, 4) most of the fraction of trans-only genes that showed non-additive effects (71%) showed recessive effects of the cultivated ortholog; suggests that *Capsicum annuum* domestication could be the ‘classic’ domestication scenario (sorghum, rice wheat, tomato) where most of the transcriptional changes are driven by recessive LOF mutations at regulatory elements.

Interestingly, these results contrast with results found in maize domestication, which has had dramatic phenotypic and functional changes when compared to teosinte. Lemmon and colleagues (2014) found that most of the trans-regulatory mutations responsible for expression divergence between maize and teosinte arose as dominant mutations that increased gene expression in maize, not supporting the hypothesis of that most of domestication mutations are recessive LOF. Our results suggest that *Capsicum* domestication resulted in an altered transcriptional landscape biased towards over-expression in cultivated plants by means of recessive mutations at trans-acting regulatory elements with pleiotropic effects.

A partial explanation for this observation is that if trans-only recessive mutations (with pleiotropic effects) that cultivated chilies bear, led to overexpression of the genes controlled downstream; the transcriptome of the F1 hybrid should display expression profiles more alike to the wild transcriptome as consequence of the two divergent transcription factors interacting with the same conserved cis-acting elements, and the impact of the ‘mutant’ transcription factor is therefore masked by the conserved transcription factor.

Our data also suggest that, even when cis-acting variation accounted for a small proportion of the transcriptional divergence between wild and cultivated chilies, most of these changes resulted in dominant mutations. Thus, it could be hypothesized that, as cis-acting elements usually control the expression of one or few genes downstream and are under purifying selection (Tirosh et al. 2009), the mutations affecting this type of regulatory elements should offer an exceptional advantageous phenotype in the context of domestication. For example, stress resistance, increased fruit size, or a trait that facilitate harvesting (Purugganan & Fuller, 2009).

The regulatory mechanisms underlying transcriptional divergence were enriched in common GO terms associated to response to biotic and abiotic stress response, however, the genes classified as having diverged in expression due to trans-only mutations showed two particular GO terms, developmental process and reproduction. This is important because it is the first insight that would suggest that fruit morphology differences in *Capsicum* were modified by means of mutations in transcription factors. Our data also showed that mutations shaping fruit morphology, could be recessive LOF mutations, as the shape descriptors in the F1 hybrid population displayed dominant effects of wild form over the cultivated form (**Figure 6 & 7).**

This evidence, when compared to the evolution of tomato fruit morphology during domestication (Frary et al. 2000), offers the possibility of hypothesizing that *Capsicum* fruit domestication could be convergent with tomato fruit domestication.

Although Paran and Knaap (2007) conclude that chili and tomato fruit evolution overlaps only in a small portion of QTLs, our results still make sense if we assume that fruit morphology is governed by homologous regulatory networks that shared central nodes, but differ in topologies, and in the genes of basal nodes. Indeed, the nature of *Capsicum* domestication has been less dramatic than maize domestication and more similar to crops like tomato, which is also the sister genus of *Capsicum* and thus evolutionary mechanisms may be conserved in both taxa. In *Capsicum*, plant architecture of the wild and the cultivated forms differ in some important traits, but are still found sympatrically throughout the landscape and are known to introgress; pungency has been retained in both wild relative and domesticated forms, and the ability to self-propagate is retained in both wild and domesticated plants. It remains to be seen to what extent these variations in the domestication syndrome and domestication process in crops are explained by common mechanisms in gene regulation.

Fine QTL mapping and reconstruction of regulatory networks of divergent molecular and organismic phenotypes, among other further studies could confirm the hypotheses mentioned throughout this work, yet our results offer novel and exiting evidence in the understanding of Capsicum domestication.

# CONCLUSIONS

Our genome-wide transcriptomic analysis enabled us to characterize both the global patterns of gene expression divergence between wild and cultivated C. annuum fruits, as their underlying regulatory mechanisms.

We found that transcriptional divergence is biased towards over-expression in cultivated chilies, and appears to have been modulated by a combination of cis (44%) and trans regulatory (66%) mutations. Our data suggest that this regulatory variation could have arisen, mainly from recessive mutations in trans-acting transcriptional regulators such as transcription factors. Which fits with the evidence of other plant domestication studies, that have concluded that a rapid evolution as domestication should have been driven by genetic changes with pleiotropic effects.

Genes that showed altered expression as product of trans-only variation were exclusively loaded into the biological processes reproduction and fruit development, which led us to hypothesize that changes in fruit morphology during human-mediated domestication could be underlaid by recessive LOF mutations at trans-acting regulatory elements. Despite of all the ongoing work needed to fully characterize fruit morphology evolution, this study provides important findings that contribute to the overall understanding of genomic mechanisms driving plant phenotypic evolution.

# REFERENCES

Aguilar-Meléndez, A., Morrell, P.L., Roose, M.L. and Kim, S.C., 2009. Genetic diversity and structure in semiwild and domesticated chiles (Capsicum annuum; Solanaceae) from Mexico. *American Journal of Botany*, *96*(6), pp.1190-1202.

Arce-Rodríguez, M.L. and Ochoa-Alejo, N., 2017. An R2R3-MYB Transcription Factor in Capsaicinoid Biosynthesis. *Plant Physiology*, pp.pp-00506.

Bell, G.D., Kane, N.C., Rieseberg, L.H. and Adams, K.L., 2013. RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome biology and evolution*, *5*(7), pp.1309-1323.

Bita, C.E., Zenoni, S., Vriezen, W.H., Mariani, C., Pezzotti, M. and Gerats, T., 2011. Temperature stress differentially modulates transcription in meiotic anthers of heat-tolerant and heat-sensitive tomato plants. *BMC genomics*, *12*(1), p.384.

Boex-Fontvieille, E., Rustgi, S., Reinbothe, S. and Reinbothe, C., 2015. A Kunitz-type protease inhibitor regulates programmed cell death during flower development in Arabidopsis thaliana. *Journal of experimental botany*, *66*(20), pp.6119-6135.

Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), pp.2114-2120.

Bray, N., Pimentel, H., Melsted, P. and Pachter, L., 2015. Near-optimal RNA-Seq quantification. *arXiv preprint arXiv:1505.02710*.

Brill, E., Kang, L., Michalak, K., Michalak, P. and Price, D.K., 2016. Hybrid sterility and evolution in Hawaiian Drosophila: differential gene and allele-specific expression analysis of backcross males. *Heredity*, *117*(2), pp.100-108.

Carroll, S.B., 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, *134*(1), pp.25-36.

Casals, J., Pascual, L., Cañizares, J., Cebolla-Cornejo, J., Casañas, F. and Nuez, F., 2012. Genetic basis of long shelf life and variability into Penjar tomato. *Genetic resources and crop evolution*, *59*(2), pp.219-229.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), pp.80-92.

Combes, M.C., Hueber, Y., Dereeper, A., Rialle, S., Herrera, J.C. and Lashermes, P., 2015. Regulatory divergence between parental alleles determines gene expression patterns in hybrids. *Genome biology and evolution*, *7*(4), pp.1110-1121.

Cowles, C.R., Hirschhorn, J.N., Altshuler, D. and Lander, E.S., 2002. Detection of regulatory variation in mouse genes. *Nature genetics*, *32*(3), p.432.

Crowley, J.J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I.K., Kim, Y., Wang, J.R., Morgan, A.P., Calaway, J.D., Aylor, D.L. and Yun, Z., 2015. Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature genetics*, *47*(4), pp.353-360.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), pp.15-21.

Doebley, J.F., Gaut, B.S. and Smith, B.D., 2006. The molecular genetics of crop domestication. *Cell*, *127*(7), pp.1309-1321.

Du, Z., Zhou, X., Ling, Y., Zhang, Z. and Su, Z., 2010. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic acids research*, *38*(suppl_2), pp.W64-W70.

FastQC A Quality Control tool for High Throughput Sequence Data: *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*by S. Andrews

Frary, A., Nesbitt, T.C., Frary, A., Grandillo, S., Van Der Knaap, E., Cong, B., Liu, J., Meller, J., Elber, R., Alpert, K.B. and Tanksley, S.D., 2000. fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size. *Science*, *289*(5476), pp.85-88.

Gensel, P.G., 2008. The earliest land plants. *Annual Review of Ecology, Evolution, and Systematics*, *39*, pp.459-477.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S. and Wayne, M., 2004. Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. *Genetics*, *167*(4), pp.1791-1799.

Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A. and Carroll, S.B., 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*, *433*(7025), p.481.

Harrison, C.J., 2017. Development and genetics in the evolution of land plant body plans. *Phil. Trans. R. Soc. B*, *372*(1713), p.20150490.

Kim, S., Park, M., Yeom, S.I., Kim, Y.M., Lee, J.M., Lee, H.A., Seo, E., Choi, J., Cheong, K., Kim, K.T. and Jung, K., 2014. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nature genetics*, *46*(3), pp.270-278.

King, M.C. and Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees.

Koenig, S., Lopez-Diaz, D., Antes, J., Boes, F., Henneberger, R., Leuther, A., Tessmann, A., Schmogrow, R., Hillerkuss, D., Palmer, R. and Zwick, T., 2013. Wireless sub-THz communication system with high data rate. *Nature Photonics*, *7*(12), pp.977-981.

Kraft, K.H., Brown, C.H., Nabhan, G.P., Luedeling, E., Ruiz, J.D.J.L., d'Eeckenbrugge, G.C., Hijmans, R.J. and Gepts, P., 2014. Multiple lines of evidence for the origin of domesticated chili pepper, Capsicum annuum, in Mexico. *Proceedings of the National Academy of Sciences*, *111*(17), pp.6165-6170.

Lemmon, Z.H., Bukowski, R., Sun, Q. and Doebley, J.F., 2014. The role of cis regulatory evolution in maize domestication. *PLoS genetics*, *10*(11), p.e1004745.

Lemos, B., Araripe, L.O., Fontanillas, P. and Hartl, D.L., 2008. Dominance and the evolutionary accumulation of cis-and trans-effects on gene expression. *Proceedings of the National Academy of Sciences*, *105*(38), pp.14471-14476.

Lester, R.N., 1989. Evolution under domestication involving disturbance of genic balance. *Euphytica*, *44*(1), pp.125-132.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), pp.2078-2079.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X. and Huang, Z., 2014. Genomic analyses provide insights into the history of tomato breeding. *Nature genetics*, *46*(11), pp.1220-1226.

Loaiza-Figueroa, F., Ritland, K., Cancino, J.A.L. and Tanksley, S.D., 1989. Patterns of genetic variation of the genusCapsicum (Solanaceae) in Mexico. *Plant Systematics and Evolution*, *165*(3), pp.159-188.

Martel, C., Vrebalov, J., Tafelmeyer, P. and Giovannoni, J.J., 2011. The tomato MADS-box transcription factor RIPENING INHIBITOR interacts with promoters involved in numerous ripening processes in a COLORLESS NONRIPENING-dependent manner. *Plant physiology*, *157*(3), pp.1568-1579.

Martin, A. and Orgogozo, V., 2013. The loci of repeated evolution: a catalog of genetic hotspots of phenotypic variation. *Evolution*, *67*(5), pp.1235-1250.

Martínez-López, L.A., Ochoa-Alejo, N. and Martínez, O., 2014. Dynamics of the chili pepper transcriptome during fruit development. *BMC genomics*, *15*(1), p.143.

McGregor, A.P., Orgogozo, V., Delon, I., Zanet, J., Srinivasan, D.G., Payre, F. and Stern, D.L., 2007. Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature*, *448*(7153), p.587.

McManus, C.J., Coolon, J.D., Duff, M.O., Eipper-Mains, J., Graveley, B.R. and Wittkopp, P.J., 2010. Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome research*, *20*(6), pp.816-825.

Meyer, R.S. and Purugganan, M.D., 2013. Evolution of crop species: genetics of domestication and diversification. *Nature Reviews. Genetics*, *14*(12), p.840.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D. and Thomas, P.D., 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, *45*(D1), pp.D183-D189.

Miller, C.T., Beleza, S., Pollen, A.A., Schluter, D., Kittles, R.A., Shriver, M.D. and Kingsley, D.M., 2007. cis-Regulatory changes in Kit ligand expression and parallel evolution of pigmentation in sticklebacks and humans. *Cell*, *131*(6), pp.1179-1189.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, *320*(5881), pp.1344-1349.

Onus, A.N. and Pickersgill, B., 2004. Unilateral incompatibility in Capsicum (Solanaceae): occurrence and taxonomic distribution. *Annals of Botany*, *94*(2), pp.289-295.

Paran, I. and van der Knaap, E., 2007. Genetic and molecular regulation of fruit and plant domestication traits in tomato and pepper. *Journal of Experimental Botany*, *58*(14), pp.3841-3852.

Pickersgill, B., 1971. Relationships between weedy and cultivated forms in some species of chili peppers (genus Capsicum). *Evolution*, *25*(4), pp.683-691.

Pimentel, H.J., Bray, N., Puente, S., Melsted, P. and Pachter, L., 2016. Differential analysis of RNA-Seq incorporating quantification uncertainty. *BioRxiv*, p.058164.

Pires, N.D. and Dolan, L., 2012. Morphological evolution in land plants: new designs with old genes. *Phil. Trans. R. Soc. B*, *367*(1588), pp.508-518.

Pires, N.D., Yi, K., Breuninger, H., Catarino, B., Menand, B. and Dolan, L., 2013. Recruitment and remodeling of an ancient gene regulatory network during land plant evolution. *Proceedings of the National Academy of Sciences*, *110*(23), pp.9571-9576.

Prince, V.E. and Pickett, F.B., 2002. Splitting pairs: the diverging fates of duplicated genes. *Nature reviews. Genetics*, *3*(11), p.827.

Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., Luo, Y. and Yang, Y., 2014. Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proceedings of the National Academy of Sciences*, *111*(14), pp.5135-5140.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rasband, W.S., ImageJ, U. S. National Institutes of Health, Bethesda, Maryland, USA, https://imagej.nih.gov/ij/, 1997-2016.

Riaño-Pachón, D.M., Corrêa, L.G.G., Trejos-Espinosa, R. and Mueller-Roeber, B., 2008. Green transcription factors: a Chlamydomonas overview. *Genetics*, *179*(1), pp.31-39.

Richardt, S., Lang, D., Reski, R., Frank, W. and Rensing, S.A., 2007. PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins. *Plant Physiology*, *143*(4), pp.1452-1466.

Rong-Lin, W., Stec, A., Hey, J., Lukens, L. and Doebley, J., 1999. The limits of selection during maize domestication. *Nature*, *398*(6724), p.236.

Sarver, B.A., Keeble, S., Cosart, T., Tucker, P.K., Dean, M.D. and Good, J.M., 2017. Phylogenomic insights into mouse evolution using a pseudoreference approach. *Genome biology and evolution*, *9*(3), pp.726-739.

Shubin, N.H. and Marshall, C.R., 2000. Fossils, genes, and the origin of novelty. *Paleobiology*, *26*(sp4), pp.324-340.

Stern, D.L. and Orgogozo, V., 2008. The loci of evolution: how predictable is genetic evolution?. *Evolution*, *62*(9), pp.2155-2177.

Studer, A., Zhao, Q., Ross-Ibarra, J. and Doebley, J., 2011. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature genetics*, *43*(11), pp.1160-1163.

Swinnen, G., Goossens, A. and Pauwels, L., 2016. Lessons from domestication: targeting cis-regulatory elements for crop improvement. *Trends in plant science*, *21*(6), pp.506-515.

Tanksley, S.D., 2004. The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *The plant cell*, *16*(suppl 1), pp.S181-S189.

Taylor, E.L., Taylor, T.N. and Krings, M., 2009. *Paleobotany: the biology and evolution of fossil plants*. Academic Press.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J. and Banks, E., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, pp.11-10.

Wagner, G.P. and Lynch, V.J., 2010. Evolutionary novelties. *Current Biology*, *20*(2), pp.R48-R52.

Weiss, E., Wetterstrom, W., Nadel, D. and Bar-Yosef, O., 2004. The broad spectrum revisited: evidence from plant remains. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(26), pp.9551-9555.

Wittkopp, P.J. and Kalay, G., 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature reviews. Genetics*, *13*(1), p.59.

Wittkopp, P.J., 2007. Variable gene expression in eukaryotes: a network perspective. *Journal of Experimental Biology*, *210*(9), pp.1567-1575.

Wittkopp, P.J., Haerum, B.K. and Clark, A.G., 2004. Evolutionary changes in cis and trans gene regulation. *Nature*, *430*(6995), p.85.

Wood, T., Burke, J. and Rieseberg, L., 2005. Parallel genotypic adaptation: when evolution repeats itself. *Genetics of Adaptation*, pp.157-170.

Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics*, *8*(3), p.206.

Wubs, A.M., Heuvelink, E. and Marcelis, L.F.M., 2009. Abortion of reproductive organs in sweet pepper (Capsicum annuum L.): a review. *The Journal of Horticultural Science and Biotechnology*, *84*(5), pp.467-475.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L., 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature genetics*, *35*(1), p.57.