Technische Universität München
Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Professur für Populationsgenetik

Neutral and selective processes underlying genome evolution post-duplication in maize

Saurabh Dilip Pophaly

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

Vorsitzender:	Univ.-Prof. Dr. Dimitrij Frischmann
Prüfer der Dissertation:	1. Univ.-Prof. Dr. Aurélien Tellier
	2. Hon.-Prof. Dr. Klaus F.X.Mayer
	3. Univ.-Prof. Dr. John Parsch

Die Dissertation wurde am 30.11.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 06.02.2017 angenommen.

Neutral and selective processes underlying genome evolution post-duplication in maize

Abstract

Maize is an important model organism with a rich legacy of applied and basic research and is also in the forefront of genomics and modern breeding. The aim of this work was to assay and examine the role of purifying selection in maize, which is a prevalent force maintaining the integrity of the genome. I used available genome data for teosinte (wild progenator) and maize as well as gene expression data to study three aspects involving purifying selection in maize, namely, the recent whole genome duplication (WGD), transposon (TE) proximity to genes, and maize domestication.

The WGD event was followed by gene erosion which generated two subgenomes, maize1 subgenome experiencing fewer deletions than maize2. Differences in purifying selection and gene expression divergence between WGD retained paralog pairs were studied. The relative gene expression of paralogs across tissues demonstrated that 98% of duplicate pairs have either subfunctionalized in a tissuewise manner or have diverged consistently in their expression thereby preventing functional complementation. Dominant gene expression was found to be a strong determinant of the strength of purifying selection, explaining the inferred stronger negative selection on maize1 genes. A novel expression based classification of duplicates was developed which is more robust in explaining observed patterns of polymorphism than the subgenome location. Upstream regions of repressed genes exhibited an enrichment of TEs indicative of a possible mechanism driving expression divergence.

Factors shaping the TE abundance in the gene vicinity were explored in the context of high TE content of maize. Gene regulatory complexity assayed by tissue- specificity and gene functional categories were found to be the dominant factors shaping TE landscape around genes. High upstream TE abundance was found to be linked with weaker purifying selection on genes while downstream TEs were found to weakly influence gene expression.

The role of maize domestication bottleneck in reducing the strength of purifying selection was explored by comparing the polymorphism patterns between maize and teosinte. Both shared and private polymorphisms displayed this reduction. Recombination being a potent force delinking loci and increasing selection efficiency was found to be associated with stronger purifying and positive selection. An increase in linkage disequilibrium post domestication in maize was proposed as a reason for the decrease in the strength of purifying selection.

The genomic and population genetics analysis conducted were indicative of a potent role of purifying selection in shaping the maize genome, a force often neglected when studying the genome evolution of domesticated species.

Zusammenfassung

Mais ist ein wichtiger Modelorganismus und zudem von zentraler Bedeutung für die Züchtung. Im Rahmen dieser Arbeit wurde die Rolle negativer Selektion in Mais untersucht, die eine wichtige Kraft für die Erhaltung der Genomintegrität darstellt. Hierfür nutzte ich verfügbare Genomdaten von Mais und Teosinte, einem nicht domestizierten Verwandten von Mais, sowie Expressionsdaten, um negative Selektion im Zusammenhang mit den folgenden drei Faktoren zu analysieren: die Verdopplung des Maisgenoms (WGD), transposable Elemente (TEs) in unmittelbarer Nähe von Genen und die Domestizierung von Mais. Auf die WGD folgte eine Generosion, die die Etablierung zweier Subgenome zur Folge hatte, wobei das Mais1 Subgenom im Vergleich zum Mais2 Subgenom eine geringere Anzahl an Deletionen aufweist. Hier wurden Unterschiede in der Intensität negativer Selektion und die Divergenz der Genexpression zwischen paralogen Genpaaren, die seit der WGD erhalten blieben, erforscht. Die Analyse der relativen Genexpression von Paralogen über verschiedene Gewebe hinweg zeigte, dass 98% der Genpaare entweder in einer gewebeabhängigen Art subfunktionalisiert vorliegen oder eine konsistente Divergenz in ihrer Expression aufweisen, so dass keine funktionale Komplementierung mehr möglich ist. Dominante Genexpression war hierbei der bestimmende Faktor in Bezug auf die Intensität der negativen Selektion und erklärte somit die als stärkere festgestellte negative Selektion von Mais1-Genen. Überdies wurde eine neue expressionsbasierte Klassifizierung der Genduplikate entwickelt, die in besserem Einklang mit den beobachteten Polymorphismusmustern steht im Vergleich zur bisherigen Erklärung, die auf der Subgenomlokalisierung beruht. Genomische Bereiche, die Genen mit unterdrückter Expression vorgelagert sind, zeigten eine Anreicherung von TEs, die somit mglicherweise an der Divergenz der Genexpression beteiligt sind. Auf Grund des hohen Anteils von TEs im Maisgenom wurden verschiedene Faktoren untersucht, die die Verteilung der TEs in unmittelbarer Nähe von Genen beeinflussen. Die Analysen zeigten, dass die Komplexität der Genregulation, die mittels Gewebespezifität und Genkategorien gemessen wurde, einen maßgeblichen Einfluss auf die Verteilung der TE-Dichte rund um Gene hat. Gene mit einer hohen Dichte an vorgelagerten TEs wiesen eine geringere Selektionsintensität auf, wohingegen nachgelagerte TEs offenkundig Einfluss auf die Expression der Gene hatten. Die Auswirkung der Domestikation von Mais auf eine Änderung in der Intensität negativer Selektion wurde durch den Vergleich der Polymorphismusmuster zwischen Mais und Teosinte abgeschätzt. Sowohl gemeinsame als auch für Mais oder Teosinte spezifische Polymorphismen zeigten wie erwartet eine verringerte Selektionsintensität. Rekombination kann die physikalische Verbindung zwischen genomischen Bereichen aufbrechen und dadurch die Selektionseffizienz deutlich erhöhen. Es konnte gezeigt werden, dass Rekombination mit stärkerer negativer und positiver Selektion einhergeht. Daher wurde die Hypothese aufgestellt, dass

das Kopplungsphasenungleichgewicht in Mais nach der Domestikation auf Grund der reduzierten Intensität negativer Selektion zunahm. Die genomischen und populationsgenetischen Analysen wiesen auf eine bedeutende Rolle der negativen Selektion in der Gestaltung des Maisgenoms hin, einer Kraft, die in Studien zur Genomevolution domestizierter Spezies bislang oft nicht berücksichtigt wurde.

Contents

List of Figures

List of Tables

Glossary

DoFE Distribution of fitness effects. 6

LD Linkage Disequilibrium. 5

MRCA Most recent common ancestor. 3

MYA Million Years Ago. 8

Ne Effective population size. 4

NGS Next Generation Sequencing. 3, 7

QTL Quantitative Trait Loci. 13

RFLP Restriction Fragment Length Polymorphism. 3, 8

SNP Single Nucleotide polymorphism. 5

WGD Whole Genome Duplication. 8

Organization

This work involved exploration of three areas in maize. Namely-

- The recent Whole genome duplication (WGD).
- Transposons (TEs) proximal to genes.
- Selection post domestication.

Each of these aspects is discussed in the general introduction, first broadly in the context of grasses and then specifically of maize. Then separate materials and methods, results and discussion are added for each. Finally, all three areas are assimilated together and discussed in the future perspectives section.

The first aspect was published in the following article:

Saurabh D. Pophaly and Aurélien Tellier. Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize. Molecular Biology and Evolution, 32(12):3226-3235, December 2015.

The article is available at the following URL: `http://mbe.oxfordjournals.org/content/32/12/3226`
For the second aspect a manuscript is in preparation.

Candidate's contribution: Conception of work, obtaining raw data, analysis and processing of data, writing of manuscript, revision of the paper.

Introduction

Population genetics and genomics

Genetic-drift, purifying and positive selection are three dynamic evolutionary forces shaping genomes of all living species. Their relative magnitude is still debated and forms a three way tug of war. Mechanistic process manifesting a phenotype at different levels and natural evolutionary processes like mutation, recombination, gene/genome duplications and other factors like demography and domestication influences the balance of these forces and shapes the patterns of polymorphism in a species. Population genetics theory has a rich heritage of providing a comprehensive set of predictions regarding these patterns in relation to various influencing factors. Due to recent breakthroughs in genomics data generation, these predictions can now be tested more extensively, comprehensively and thoroughly.

A substantial part of population genetics is the study of intra species polymorphism patterns and of the forces shaping them. Genetics typically involves studying related individuals and inheritance patterns via crosses, whereas population genetics involves generalizing the outcome of an ensemble of these crosses over time. Variation is crucial to both these fields as at a technical level it 'marks' the inheritance pattern and acts like a tracer. Markers essentially need to be 'polymorphic' whereby they display variants and these should be able to be assayed. The variants of a marker are called alleles. Markers can be phenotypic, for example, the texture of the famed Mendel's peas with two variants wrinkled and smooth. But most markers in use are molecular. Earlier studies primarily used markers like isozymes, RFLPs and microsatellites wherein a change in a DNA sequence is assayed by proxies like altered enzyme activity or by different sized cleavage fragments in a gel. Sequencing technologies provide a more direct view by removing the proxies and deciphering the DNA sequence itself. The advancement of sequencing technologies have made SNPs (Single Nucleotide Polymorphisms) which is a change in a single basepair of DNA, a dominant choice as a marker for population genetics. Analysis typically involves marker variants assayed from a sample of individuals from a population. Although usually markers from a small number of individuals (size n) from the population are analyzed, and statistics reported, it is generally sufficient to capture the overall polymorphism patterns and history (time to most recent common ancestor, MRCA) of the whole population. The statistics used in the current work are explained below.

Allele frequency is the simplest statistic to calculate and is the frequency of a particular variant of a marker in the sample. A SNP can have more than two alleles but but it is rare and analysis usually is done on a biallelic SNP and frequencies of both alleles add to one. Single nucleotide insertion and deletions are also usually excluded. In this work only biallelic SNPs were used. When multiple SNPs are assayed, the variation is also displayed as a 'allele frequency spectrum' which is a histogram plot of allele frequencies. For a biallelic SNP, since the frequency of both alleles add to one, only the information about frequency of one allele is enough to calculate the frequency of another. The question arises that which allele's frequency should be reported. Studies sometimes make a distinction based on smaller or larger allele frequency and correspondingly report a 'minor allele frequency' or 'major allele frequency'. NGS based technologies sequence short fragments of DNA from samples and these fragments are then 'mapped' to a reference sequence. Reference is a an independently sequenced and

assembled whole genome of usually one individual (but sometimes multiple) of the species involved in the study. In this case a 'reference allele frequency' and an 'alternate allele frequency' is respectively reported based on if a basepair in a sample matches a reference basepair or not. Another way of reporting is deciphering the historical (or ancestral) direction of mutation which caused a SNP, for example, if a SNP has two alleles A and G, the mutation could have happened from A→G or G→A. The older allele is called as the 'ancestral allele' and the newer one the 'derived allele. Obtaining this information is called as 'polarizing' a SNP. It involves assaying the state of the site in a closely related species and this state is assumed as ancestral. Most calculations are done according to an 'infinite sites model' where an important assumption made is that there are infinite number of sites and a mutation can only strike once at a particular site [123].

Humans usually have an intuitive understanding of the concept of 'diversity' which they qualitatively associate with more variants seen for a particular trait. But quantitative measures of diversity are essential for a rigorous analysis and to test predictions. Diversity for a population sample of SNPs depends on two variables, first is the number of SNPs and another is the frequency of each SNP. Watterson's theta (θ_w) is a measure of diversity purely on the number of segregating sites (SNPs) [270]. Another measure denoted as Π or θ_Π gives a convenient way of capturing both [167] and was used in this work. Π is estimated as the average of the number of differences seen in two random individuals from the sample (also called as 'average pairwise difference'). One way of calculating it involves performing all possible pairwise comparisons of individuals and adding the nucleotide differences seen for each comparison and then dividing by the total number of comparisons (which is (n(n-1))/2). Naturally, Π depends positively on the length of the region of the genome sampled, so a normalized value is usually obtained by dividing the locus length (per basepair value of Π). Per basepair Π can be calculated over the genome in sliding windows, for a gene/region, or for a particular type of site for example, for synonymous and non-synonymous sites. In the later case the normalization is done by dividing by total number of synonymous or non-synonymous sites assayed.

Sewall Wright and Ronald Fisher reconciled Mendelian genetics with Darwinian evolution and explained how numerous individual crosses affect the allele frequency and intra and interspecies differences. The Wright-Fisher model provides a simplistic null model wherein alleles in one generation are randomly sampled to form another generation. Allele frequencies thereby stochastically vary and although there is an expectation for the next generation, actual value can only be determined by sampling the next generation akin to the famous urn model of statistics [207]. This random sampling is also known as 'genetic drift' and the size of the urn is the 'effective population size (Ne)' which also is a measure of the fidelity of allele frequency to remain same between generations. Such models are crucial to draw inferences and provide a framework to orient oneself in empirical data analysis. But it is the deviation of the observed data from these models which make the most interesting and sought after cases. These deviations are generally caused by violation of the assumptions of the model. An elegant example of quantification of the deviation is a statistic called as Tajima's D [243] wherein under the assumptions of the Wright-Fisher (a constant population size) θ_w and Π should be equivalent but forces like natural selection and demography upset this balance. Tajima's D quantifies the difference between Π and θ_w. When an allele has a selective advantage, then sampling is not random

as it is 'preferentially sampled' because the individual harboring the allele has a better chance of making to the next generation compared to other individuals without the selected allele. Allele frequency of that variant then rises more rapidly than expected under neutrality and nearby sites also increase in frequency due to linkage (see below) clearing the area of variation when the preferred allele and linked variants reach the frequency of one (fixation). This phenomena is also called as 'selective sweep and hitchhiking', a term coined by Maynard-Smith and Haig [229]. Mutation then introduces new variation and initially, new variants have low frequency since increase of frequency by drift needs generations of sampling. Since Π depends more on allele frequency whereas θ_w does not, their values differ and Tajima's D becomes negative. In this case natural selection in terms of adaptive evolution is the force that disturbs the equilibrium and violates the assumptions of the Wright-Fisher model. Many other statistics/measure/methods exist which detect such deviations, particularly for finding regions under positive selection (adaptive evolution) and are commonly used for genomic data [261].

When considering several loci (sites), a more common, fundamental and nearly ubiquitous violation of the assumptions of the Wright- Fisher model is called as linkage disequilibrium (LD). It violates the assumption of independent sampling of different sites. LD is a measure of non-random association of alleles between sites [90], this association, when caused by the physical proximity of the markers on a chromosome is also called simply as 'linkage'. Specific variants of proximal markers, due do their location on the same chromosome, are sampled as a block (which is also called as a haplotype) thereby linking their sampling, allele-frequency and fate. Recombination due to cross-overs exchanges (swaps) homologous regions in the chromosome from a population thereby decreasing LD. LD typically decreases with the distance between sites in the genome as the likelihood of a recombination event increases with this distance. This make sites independent to some degree. LD measurement needs haplotype data but for SNPs usually only genotypes are available. Deducing haplotype from a genotypes is called as 'phasing' and there are specialized softwares for it. Sometimes the samples are inbred by selfing which drastically reduces heterozygous sites and the genotype data can then be represented as a haplotype and no phasing is needed. This was the case in this work as the maize samples used were of inbred lines. A 'four gamete test' [103] is one of the simplest ways of detecting a recombination event. It uses combination of marker variants seen between two markers in several individuals to obtain an estimate for the number of recombination events in the history of the sample. As a simplistic example, if two markers/variants (A/a and B/b) are in physical proximity and assayed in a population, the possible combinations that can be seen are AB, ab, Ab, aB. If in four individuals AB, ab, Ab and aB are seen then at-least one recombination can be inferred which switched the allele variants. It is important to note that the number of recombination events reported are nearly always an underestimate as the events which are not 'flagged' by the marker variants can not be determined. An extreme example would be a sample with no diversity in a population (only one marker variant is seen for each marker), although the recombination events might have happened, they can not be determined as the method relies on marker variants.

Purifying or negative selection maintains the 'status quo' by weeding out variants which negatively affect the fitness of the individual harboring them in the population. They violate the random sampling assumption by negatively affecting the chance of a variant to be sampled thereby decreasing

its allele frequency. Since this work involved identifying and studying the role of purifying selection in shaping the maize genome, several measures were used for assaying it. Some of the methods involve using genome annotation and biological knowledge. An example is a variant which introduces a frame-shift or a premature stop codon which is most likely to affect the protein function. Non-synonymous variants change the amino acid thereby are more likely to be detrimental than synonymous. Conservation of regions in interspecies comparisons is a strong indicator of function and purifying selection. The Ka/Ks ratio is a commonly used statistic in molecular evolution [286] based on conservation and biological knowledge, where the nucleotide changes in each category between species are normalized by the total number of sites in each category. $\Pi n/\Pi s$ measures the ratio of diversity between non-synonymous and synonymous sites (analogous to Ka/Ks) and indicates the strength of purifying selection acting on the coding regions on a shorter timescale. DoFE (Distribution of Fitness Effects) [58] is the frequency distribution of mutations in different fitness classes. It gives the proportion of mutations in various classes of selective effects. DoFE thus gives what fraction of mutations are neutral, deleterious and very deleterious. This distribution can be obtained by experiments involving fitness assays. But many population genetics based approaches have been developed to obtain DoFE from sequence polymorphism data [58]. Like $\Pi n/\Pi s$, population genetics based methods usually need two classes of sites, one class for which DoFE is obtained (e.g non-synonymous) and another which is assumed to be neutral (neutral standard) (e.g synonymous sites). These methods are based on the premise that mutations in sites in the selected class will be few and kept at a lower frequency by purifying selection and the fraction of mutations in different classes of selection strength is obtained by comparing the number and frequency of SNPs between two classes of sites (selected v.s neutral). The Eyre-Walker and Keightley method as implemented in the software DoFE was used in this work [59]. As a test example this method was run on classical genes in maize v.s 15000 randomly chosen maize protein coding genes with cDNA evidence. Classical genes are genes which are well studied in maize and are more likely to exhibit a mutant phenotype [215] so purifying selection is expected to be stronger for them. The results are depicted in Figure 2.1 where the strength of selection is represented as a product of selection coefficient (S) and recent effective population size (Ne). Higher values of -NeS indicates stronger purifying selection. Higher fraction of mutations in higher -NeS classes shows stronger purifying selection acting on classical genes compared to a set of random genes. Another observation seen here and in general [101] is that a large fraction of non-synonymous mutations are very deleterious and purifying selection is thereby a pervasive force. One common confusion in interpreting DoFE results is that the inferred fraction of mutations in different classes in the population is reported and not the fraction of SNPs in the data. The SNP data only helps in inferring the distribution and is not directly represented in the results. For example, in Figure 2.1 classical genes show a higher inferred fraction of mutations in highly deleterious class (-Nes$>$100) and thereby stronger purifying selection on non-synonymous sites, but this does not represent the fraction of SNPs which belong to that class (-NeS$>$100). In the DoFE distribution, the class -Nes$>$100 also covers mutations which are too deleterious or even lethal to be observed in the polymorphism data. Also DoFE obtained by this method can not predict the purifying selection strength on an individual given mutation.

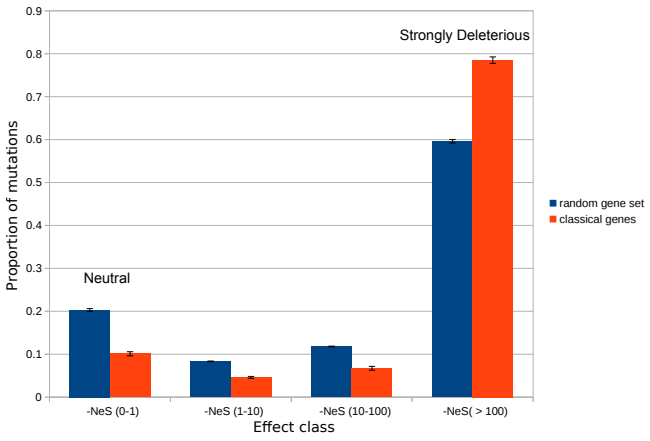


Figure 2.1: Distribution of Fitness Effects (DoFE) for classical maize genes vs 15000 random genes with cDNA evidence. Y-axis gives the fraction of mutations and x-axis gives the deleterious effect of mutations scaled as product of effective population size and selection coefficient (-Nes). Higher values of -Nes indicate stronger deleterious effect.

The population genetics based methods for assaying strength of purifying selection rely on allele frequency calculated from the sample. Another method which was used in this work utilizes site conservation only and provides an independent estimate is SIFT [170]. SIFT measures the inter species conservation at the site of the SNP to generate a score estimating the deleterious nature of the SNP [170]. A good agreement was in general seen between the SIFT and other methods in this work which is further discussed in chapter 2 and 4.

The decade-old and still ongoing advances in Next Generation Sequencing (NGS) and allied technologies have influenced all areas of biology. The steep decline in costs and the sharp incline in the data density have even surpassed the famed 'moores law' meaning that the 'Sequencing revolution' is unparalleled to even the 'Semiconductor revolution' [87]. Unprecedented number of genomes are now available both within and across species making way for detailed intra and interspecific comparisons and generation of deep intra-species variation catalogs. NGS based deep expression quantification via RNA-Seq, genome occupancy assays via Chip-Seq and NGS variants to detect epigenetic DNA modifications are closing the knowledge gap between genotype and phenotype. Fundamental changes are also happening in the way biological research is done. The scope of hypothesis testing has expanded to a system level, and power to test theories, predictions and effects has received a boost. In addition to testing a hypothesis, the generated data is also aiding human curiosity and intuition in creating new predictions and linking different levels of biological organization. Needless to say that plant biology with all its flavors including crop breeding and domestication has seen a massive percolation by NGS. Population genetics traditionally had a strong theoretical bent and empirical data analysis and testing of predictions have been limited by the data availability. NGS technologies have lifted this and genome scale population surveys involving multiple individuals is now common giving rise to the field of 'population genomics' which involves fusion of genomics with traditional population genetics [148, 30].

The grass family and their genomic circle

The family of grasses is usually in the limelight in the field of plant biology given their single largest contribution to the global food supply [119]. The very word 'food grain' implies endosperm of a grass species. In addition to direct human consumption as food, grasses also are indispensable for their indirect application as animal feed and biofuel source [176]. The food value has been achieved by domestication of many individual grass species by humans as far back in time as epipalaeolithic period [188]. It would not be far-fetched to say that the success of human civilization has been and still hinges on the continuing molding of grass species to suit our needs. The growing human population only underlines the need and urgency to better understand grasses and suitably apply the acquired knowledge for higher yielding and robust grass species [54]. Grasses also form a well suited study system for a range of applied and basic biological and evolutionary questions. From a pure evolutionary-genomics perspective, grasses form an excellent test case to study fundamental processes shaping genome evolution post domestication as many grass species have been independently domesticated. They have also colonized a variety of different habitats and encompass a spectrum of phenotypic variation, and added to this are the phenotypic changes introduced by domestication. Despite the variability amongst different species in appearance, initial studies indicated that grass genomes display a marked collinearity. These studies were based on morphological, isozyme and RFLP markers and formed the basis for a view that grasses are made up of similar linkage blocks. The information in one linkage block in one species can be transposed to the same linkage block in another species by a circular cross-species linkage map [161]. There was even the hypothesis of a single grass pan genome and transposing loci responsible for phenotypes amongst grass species [12]. The transfer of markers from one species to another is still common in grasses [69], but the coarseness of these markers meant that small deviations from synteny were not detected. Alignments of whole genome sequences can nowadays give a complete picture. A fine scale and unbiased analysis of synteny is readily possible today by whole genome alignments of different species and web based tools like SynMap [149]. Contrasting to the collinearity is the genome size variation in grasses which is extensive, for example wheat has a haploid genome size of $\sim$17Gb whereas for rice the genome size is around $\sim$400Mb [53]. Also, the genome size changes can be attained in relatively short evolutionary timescale, for example, maize and sorghum are closely related ($\sim$12MYA) [241] but the genome size is double in the former. The major factors which explain this are Whole Genome Duplication (WGD) and heterogeneity in the abundance of Transposable Elements (TEs).

Whole genome duplication (WGD)

WGDs are also aptly called as paleopolyploides meaning an ancient event of polyploidy, only whose remnants can be seen in the genome today. WGD are a common occurrence in plant phylogeny and seem to be well tolerated compared to animals (Figure 2.2) [14, 136]. All grasses share two rounds of WGD which happened in the pregrass ancestor $\sim$70 Mya [180], an additional WGD event happened in maize lineage as it diverged from Sorghum [73] and seems to be the only well characterized and

studied WGD in grasses post divergence. Although WGD events are uncommon in grasses post divergence from the MRCA, when compared to other flowering plant lineages, discovering new events can not be ruled out as more data on different species becomes available. The simplest depiction of a WGD is a doubling of the chromosome number back in time and its returning to the preduplication number (2.1). Susumu Ohno not only proposed a WGD in vertebrate ancestor which was supported by later work but also highlighted the importance or WGDs in evolution [173]. WGD generated duplicate pairs are also called as homeologs or ohnologs in the honor of Ohno.

$$2N \longrightarrow 4N \longrightarrow 2N \tag{2.1}$$

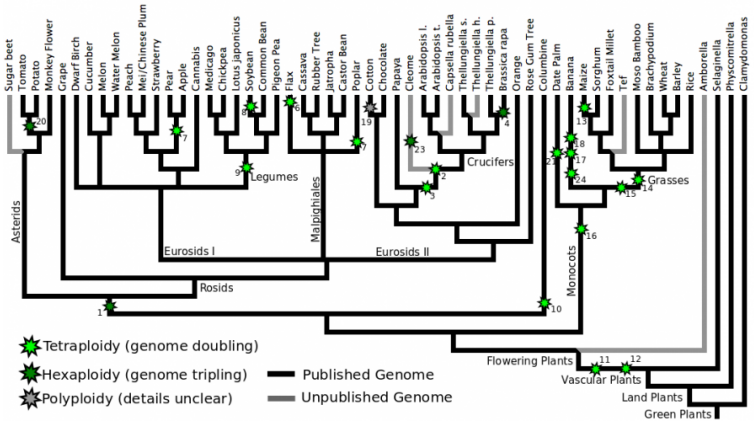


Figure 2.2: WGD events in Flowering Plant phylogeny. Obtained from `www.genomeevolution.org`

A WGD not only duplicates every gene but also preserves the intergenic and regulatory interactions. Paralogous pairs (ohnologs) resulting from WGD are thus 'born equal' but then they diverge in sequence, expression, and sometimes function. After WGD a process of diploidization ensues, which is typically accompanied by rapid and massive gene loss, a process called 'fractionation' [131]. For example, in maize only approximately 20% of genes are retained as duplicate copies since the maize specific WGD [217]. Fractionation thus explains why gene numbers in plants are still moderate even after multiple rounds of WGD [218]. Fractionation is shown to be biased in many studies done in different plant species which means that some genomic regions of the diploidized polyploid shed more genes (sensitive subgenome) compared with other (dominant subgenome) [249, 164, 35, 65, 28] although WGD events without such a bias also exist [72]. It has been proposed that biased fractionation is associated with allotetraploidies whereas autotetraploidies are associated with no fractionation bias [72]. Not only gene deletions but gene expression has also shown to be biased in the favor of the dominant subgenome [35, 65, 83]. In this context an often caused confusion was highlighted by Grover et.al [83] between 'genome dominance' and 'homeolog expression bias'. Genome dominance implies the combined expression is similar to one of the polyploid progenitors whereas homeolog expression bias references to difference in relative expression of ohnologs [83]. This work involved studying homeolog expression bias in maize which has been seen in many other plant species [193, 1]. While analyzing divergent relative expression of duplicate pairs across many tissues two possible patterns exist. First a unidirectional expression divergence (UED) defined as one member of the pair being

consistently highly expressed (UED-dominant) than its counterpart (UED-repressed). Such a consistent decrease in expression of one member was called as 'regulatory hypofunctionalization' [52]. The second scenario is bidirectional expression divergence (BED) for which both genes are alternatively dominant and repressed in different tissues. The expression difference between ohnolog pairs has been shown to be quickly established after the formation of the synthetic allopolyploids in the cases of Cotton [65] and Arabidopsis [263] and for natural allotetraploids of *Tragopogon miscellus* [19], *Brassica rapa* [35] and maize [217]. The mechanisms operating behind the initial expression differences and divergence between ohnologs is not clearly understood. Epigenetic effects like DNA methylation have been tested but have not been proven. Parkin et al. [179] found subgenome dominance for expression in *Brassica oleracea* but methylation profiles did not correlate with dominance for individual genes. No differences were found in gene body methylation between maize subgenomes [55]. Initial differences in upstream transposable elements (TEs) caused by allotetrapolidy has been proposed [217]. Repression of an upstream TE might cause an inadvertent decrease in expression of the nearby gene but it was shown not to be working in cotton [199].

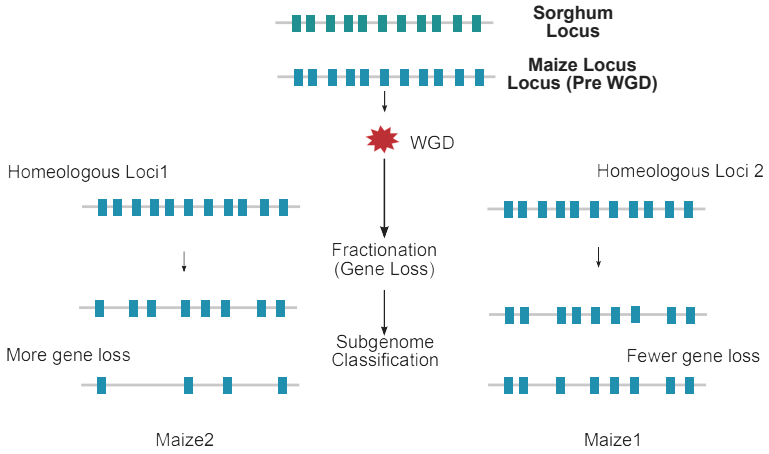


Figure 2.3: Post WGD Fractionation and creation of subgenomes after maize specific WGD

The fate of genes after WGD is generally seen in the light of two hypotheses. First, the gene dosage balance hypothesis, which predicts that selection acts on maintaining the stoichiometric ratios of protein amount between interacting gene partners [13]. Thus regulatory genes and genes involved in multi-protein complexes which typically have many interactions are more likely to be retained after WGD [8, 200]. Few studies have also indicated absolute gene dosage to be an important determinant of retention after WGD [2, 27, 199, 156]. This is in stark contrast with the gene retention after tandem duplication where genes with lower dosage constraints are preferentially retained [200]. This causes the genome post WGD to get enriched in certain functional categories which include regulation. Many implications and consequences of this enrichment have been proposed and reported [256] which include increase in regulatory and organismal complexity [118, 106], diversification [97], speciation [203], and evolutionary innovation [23].

The second hypothesis implicated in the fate of WGD duplicates is the subfunctionalization and neofunctionalization hypothesis which states that the fate of the duplicated gene pair broadly follows two known outcomes, subfunctionalization, where the ancestral function is partitioned between du-

plicate copies [68], and neofunctionalization where one of the duplicate copy evolves a new function. Both outcomes can be achieved at the level of gene expression or protein function [64]. Expression-based subfunctionalization can be readily assayed by analyzing relative expression of duplicates. Duplicates can be expressed differentially across tissues, developmental stages, or environmental conditions or one copy can attain a novel expression profile [52, 31, 144, 253]. Generally a substantial number of WGD duplicates display divergence in expression [52]. 50% of duplicates were reported to have diverged in expression in a study in soyabeen [204]. A study in Arabidopsis found that 85% of duplicate genes show evidence of regulatory subfunctionalization and/or neofunctionalization [52, 198]. A study in cotton reported a near complete expression divergence between WGD duplicates [198].

The WGD event shared by grasses is very old ($\sim$70MYA) [180, 269] and its remnants can be seen in whole genome sequences of individual grass species. This also gives a chance to study the fate of each ohnolog pair separately in each individual species. Occurrence of paralog pairs in studies involving analysis of gene expression, annotation, function, gene family and phylogeny is thus common [268, 265, 281, 160, 185, 114]. WGD paralogs have been implicated in various phenotypes in grasses including C4 photosynthesis [268] and grain hardness [289]. A study reported preferential retention of starch synthesis genes post WGD in grasses when compared to arabidopsis [280], this is particularly important as it indicates that the seeds for domestication of grasses and their food value were sown $\sim$70 million years ago by the WGD. The WGD has been implicated in many instances in domestication induced phenotype change in grasses [181, 254]. The differing changes specific to each species in WGD duplicate pair which include gene gain by tandem duplication, gene loss and location or amount of expression can have multiple consequences. These can include convergent evolution where a same phenotype is achieved by different changes in duplicates, interspecific phenotypic diversity where changes in paralogs result in species specific phenotypic differences and adaptive evolution. Overall the ancient WGD in the grasses presents an excellent system to study replicated instances of rewiring of gene interactions and the resulting similar, different or novel outcomes.

Taming of the grasses

The relationship of humans with grasses can be described as been active and reciprocal. Humans molded different grass species via domestication and grasses provided a staple and stable food source thereby changing the hunter gatherer lifestyle to more stable permanent settlements [212]. Domestication traditionally involved artificial selection based on traits and was largely empirical [89]. The timescales involved in domestication are minuscule compared to the phylogeny of grasses [76], yet the resulting phenotypic changes are nothing short of the word 'impressive'. For example, in case of maize, the appearance of the plant was so different (Figure 2.4) that no species was clearly identified as the 'wild progenator', purely based on morphology, [46] and finally molecular data resolved this issue [47]. Independent domestication events have happened for many grass species like rice, wheat, maize, sorghum etc making grasses an excellent overall system for studying domestication itself.

The generalized collection of changes induced by domestication has been dubbed as 'domestication syndrome' [88]. This convergence of phenotypes can be seen in different aspects of biological organization including morphology (e.g branching patters and seed shape and size) life history (e.g seed dormancy) and biochemical composition (e.g altered starch composition and toxicity) [175]. Several examples of convergent phenotypic evolution have emerged in grasses due to their independent domestication [89, 183], although the molecular mechanisms and genes involved can be different for each species [78]. Some examples of such changes include non shattering seeds (which is crucial for harvesting), changes in branching patterns (single branching in maize as shown in Figure 2.4) and waxy phenotype which increases consistency after cooking and selection for flowering time [175]. Studies have identified different mutations in different genes achieving these phenotypes in individual species [107, 78].

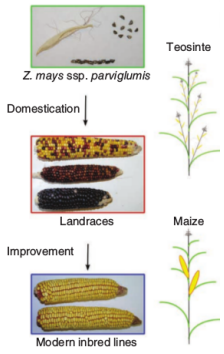


Figure 2.4: Visual changes introduced by domestication in maize from teosinte [104]

Domestication is a fascinating process to study at a genetic and genomic level as it involves discovering the molecular changes selected by humans mostly unaware of genetics. Application of genetics is a powerful tool to understand domestication and its implications in the genome and to discover the variants selected during domestication. In modern days, the phenomena of domestication itself has been domesticated meaning that we understand it much better and can apply this knowledge to design and implement changes to a target species at an accelerated pace. This has happened partly due to the availability of a torrent of molecular information about the phenotypes and the ability to manipulate genetic information to the point that a new phenomenon called as 'super domestication' is emerging [257, 92]. For example, a 5 to 10 fold yield increase is common in a short time span of a few decades due to modern breeding [18] and the so called 'modern' commercial lines for maize produced as double haploids by using genomic selection and knowledge based introgression far surpasses the traditional lines in desirable traits [134]. Traditionally, QTL (quantitative trait loci) mapping has been central to locate the genetic basis of traits in grasses [78]. The development and advancement of methods like GWAS (Genome wide association analysis) and NAM (nested association mapping) with increased availability of dense marker data have accelerated the discovery of implicated loci [175].

Domestication syndrome also happens at the level of genetic polymorphism wherein it generally results in a genomewide decrease in diversity of the domesticated population compared to the wind progenitor. This is due to the fact that few individuals from the wild population become the found-

ing members of the domesticated population (a term called as bottleneck) [45]. The magnitude of reduction in diversity differs for different species and depends on several factors including severity of bottleneck, number of domestication events (single v.s multiple), prevalence of gene flow after domestication and the mating system (selfing v.s outcrossing) [78, 57]. The diversity loss is typically uneven across the genome with the loci favored during domestication incurring stronger reduction. This is because of favored sampling of the desired variant and other linked variants (see introduction) which then increases in frequency clearing the loci of diversity (selective sweep). This scenario is of the case when the phenotype altering variant is new or at a low starting frequency (hard sweep). This may not always be the case if the variant is already at an appreciable frequency in the population (selection from standing genetic variation), a scenario called as 'soft sweep' which does not result in a drastic diversity reduction. Population genetics has been used as a tool to detect loci targeted by domestication. This bottom-up way is in contrast to the topdown approach of going from a phenotype to the causal genetic variant by QTL mapping [202]. The bottom-up approach is typically seen in genome scans for detecting adaptive evolution which are based on the population genetic theory that states that the targeted loci which are undergoing adaptive evolution leave signatures in polymorphism patterns like decreased diversity and longer haplotypes [155, 121]. Purely demographic effects like bottlenecks can also cause such signatures but the effect of demography is usually far less localized and can be seen over the entire genome. Statistics quantifying these signatures are obtained for windows over the genome and in absence of reliable demographic models which are often cumbersome to obtain, an outlier approach is used with extreme values assigned as regions undergoing adaptive evolution. But experimental validation of an adaptive phenotype effect is the only proverbial "proof of the pudding". An augmentation to this approach is to obtain genome scale polymorphism data for the domesticated and the related wild population, such that regions displaying decreased diversity only in the domesticated samples are prioritized [104]. In some rare cases causal variants for phenotypic changes can be found by only comparing the sequences of a domesticated species with a wild ancestor, particularly in case of low number of differences. For example, a naked kernel in maize which makes consumption much easier is caused by only one amino acid substitution between modern maize and its wild cousin teosinte [262]. The success of population genetic methods heavily relies on availability of population scale genome data and understanding of the demographic and stochastic processes involved in domestication. The availability of such datasets is on the rising for different grass species [26, 20, 36, 102]. Post whole genome sequencing, generation of intraspecific variation catalog usually follows for domesticated grasses, maize hapmap consortium, for example, provides whole genome reqsequencing data for about thousand maize lines [20].

Low diversity can adversely affect the species ability to respond to adverse conditions like pathogen pressure, genetic drift is stronger in such cases and the strength of selection is expected to be reduced [78, 122]. The strength of selection is also dependent on the recombination which delinks loci and increases the efficiency of both positive and purifying selection [94, 79]. This was shown elegantly in drosophila by Campos et.al [22] by making dividing genes in bins based on recombination events and then assaying positive and purifying selection differences between bins. Strength of both positive and purifying selection increased with increasing recombination [22]. The decrease in selection

strength is particularly important for domesticated species due to three factors. First concerns the nature of selection itself, traits critical for survival in wild may no longer be important in the domesticated variant due to assistance by humans. For example seed shattering in grasses is important for dispersal in wild but not in domesticated strains [49], conversely fixations of variants for traits under domestication related positive selection result in fixation of linked deleterious variants. Second is the decreased diversity and increased drift due to domestication bottleneck. Third is an increase in LD post domestication which impedes the efficiency of selection [188]. A study in rice reported more deleterious non-synonymous changes when compared with wild relatives [147]. More studies have reported this phenomena in grasses [125, 159]. There is often more focus on detection of adaptation in domesticated species when compared to studying the effect of purifying selection. The availability of polymorphism catalogs of both wild and the corresponding domesticated species would encourage more studies in this regard.

In a nutshell domestication is a fast evolutionary process as early recognized by Darwin. Multiple independent domestication of various grass species makes them a suitable case study of key population genetic evolutionary processes like adaptation, drift, mutation and purifying selection.

Transposable Elements

The most significant contributor to changes in genome size in plants are Transposable Elements (TEs) [244]. They also form an explanation for the "C-value paradox" wherein genome size and organism complexity are usually uncoupled [184]. TEs were initially called as 'jumping genes' because of their ability to change positions in the genome. This ability to hop into and thereby potentially disrupt a gene can sometimes cause prominent phenotype changes even at a somatic level. Eye-catching examples of such changes include varied colored grains in a corn kernel and patchy or speckled pigmentation in flowers and leaves [81, 40]. The later led to their conceptual discovery by Barbara Mcclinktok. The verification of their molecular existence and unraveling of underlying mechanisms ensured her a place in scientific history. Since then, numerous types and strains of TEs have been discovered in nearly every species sequenced including plants. Nowadays TEs are largely discovered in genomes insilico, by sequence similarity searches with a library of TE sequences made for a particular clade, but specialized signature-based approaches also exist [137]. The MIPS (Munich Information for Protein Science) provides such TE libraries for many plant clades and as well provides a neat 'internet Protocol (IP) address' like nested classification of TEs [172]. Detection of TEs only gives a static picture but the exact molecular mechanisms of transposon jumping can be complicated, vary between different types of TEs [275]. They fall into two major classes [63], first being the class I TEs which use a 'Copy-and-paste' mechanism via an RNA intermediate (also called as retrotransposons). Second are the type II TEs which transpose using a 'cut-and-paste' mechanism (DNA transposons). But several subclassifications and variations exist within these two broad types [275].

Transposons are often called as 'selfish DNA' or 'genomic parasites' because of their ability to actively increase their copy number without contributing positively to fitness of the ensemble [177, 273]. Still they manage to survive, thrive and are an abundant source of genetic variation which can

affect genes in diverse ways [142]. Their abundance is the single-most potent factor which influences genome size [244, 34, 84]. For example, genome size of rice is ∼400Mb [187] compared to TE rich maize which is ∼2300Mb. Also impressive is their ability to change genome size in relatively short evolutionary timescale. A striking example of this is a 50% increase in genome size of zea *luxurians* compared to zea *mays* (maize) with only a divergence time of ∼140KYA [245]. These TE 'bursts' can have profound consequences not only at a species level but also higher [10]. Note that the exact mechanisms governing the abundance of TEs in a particular species is still an area of active research [242].

The tip of the TE iceberg

Looking at a colored speckled pattern on a leaf or flower or multicolored corn kernel one can imagine a TE insertion or excision in a pigment producing/regulating gene [61]. This insertion most probably happened in a stem cell and then transmitted to its progeny cells in the somatic tissue or the germline. The change being visible to the naked eye makes the detection effortless. Its local nature and commonness imply that it might not be drastically deleterious. But a multitude of such changes would be happening to other genes whose phenotype may not be so obvious. So on a phenotypic level TEs can cause large effect changes when present in or near genes, so much so that they have been used for generating artificial mutants and knockouts [126].

This ability to cause large genetic and phenotypic change can occasionally make them agents for adaptive change. Several adaptive effects of TEs have been documented and as such any striking and/or adaptive change caused by a TE is usually cherished and highlighted in the literature [260, 142]. Their contribution to gene birth, regulation and evolutionary innovation has also been proposed [273, 205] and shown in a some cases [11, 142]. But compared to their genomic abundance, known cases of TEs displaying adaptive effect form a minority and the evidence of their "general" utility in the genome is still unclear and actively researched [231, 195].

A strong mutagenic and phenotype altering potential would also imply copious deleterious effects of, and thereby purifying selection on TE insertions. Insertions resulting in dramatic and deleterious phenotype alternations such as insertions in coding regions of functional genes would be removed by natural selection and seldom seen in population genomic data [4, 3]. Nonetheless, the occasional longterm persistence and high abundance of TEs also imply that they exhibit neutral or nearly neutral effect and there is a role of life history and population genetic forces in maintaining them in genomes [242, 146, 16, 48]. The deleterious nature of TEs is in contrast to the nearly-neutral effects explained in the former section and the location of TE in the genome is crucial for a reconciliation.

Aside from purifying selection genomes have a few tricks up the sleeve to 'pro-actively' protect the genes from TE insertions which include methylation, chromatin organization and small interfering RNAs [214, 82]. These constitute an 'epigenetic immune system' for protecting against selfish elements by negatively affecting the ability of TEs to jump (called as 'silencing' of TEs) [228, 141]. Like mechanisms of TE jumping, the silencing mechanisms are also actively researched [232, 153]. Variants of NGS based technologies have accelerated the pace in this area by providing genome scale maps of chromatin configurations, methylation and expression levels of small RNAs and genes [9, 74].

My interest lied in studying the TE distribution in gene vicinity, which is more likely to be shaped by their deleterious effects via influencing gene expression. This presents a challenge because in contrast to TE insertion in protein coding regions, the effects of TEs upstream and downstream of genes are difficult to discern. A major reason is the scarcity of data on the location of promoters and cis-regulatory sites especially in plants [130]. These regions extend outwards from the transcription start and end sites and form a 'Grey zone' of the gene boundary. A second reason is the heterogeneity in the effects caused by TE insertion, making the outcome dependent on many factors. For example, the extent of damage to the cis-regulatory region would depend on the size and location of the TE insertion in relation to this region. The complexity and functional density of the upstream cis-regulatory landscape would in turn be the factor influencing the likelihood of a TE insertion. Since the most likely change an upstream TE insertion can make to a gene function, is to influence its expression, the resulting phenotypic change would not only depend on the magnitude of expression change but also gene specific properties like dosage, sensitivity to expression variation and overall effect on fitness. Studies have pointed to an indirect association between expression divergence and upstream TE abundance [99, 186]. Altered gene expression due to proximity to TEs was shown in wheat [117]. Multiple studies have also indicated the potential of TEs themselves to act as promoters for nearby genes, thereby conferring a new expression profile [166, 236, 37]. Epigenetic management of TEs via silencing adds an additional layer of complexity as the processes meant to suppress TEs can inadvertently suppress nearby gene expression thereby creating an indirect link between TEs and neighbouring gene expression [98, 41]. Methylation, which suppresses TEs has been strongly associated with repression of expression when present in the promoter region [41]. TEs were also linked to intraspecific variation in gene expression in *Arabidopsis* and the subset of TEs targeted by siRNAs were specifically found to be more distant from genes, presumably to avoid inadvertent gene silencing [266]. The cross connection between epigenetic silencing of TEs and gene expression was also shown in *Arabidopsis* by Hollister and Gaut, where gene expression was found to be negatively correlated with the density of nearby methylated TEs only and not non-methylated TEs [98]. An expedited removal of methylated TEs presumably due to their methylation affecting nearby gene expression was also shown [98]. Additionally a study found rice found that the methylation of downstream regions can repress transcription, even stronger than upstream regions [140].

Choosing of Maize

Amongst grasses, the foremost worldwide production is of maize. I decided to choose maize as a model organism for my dissertation as it has experienced all the aspects of grass evolution highlighted before. A complete high quality genome sequence with a chromosome level assembly was available since late 2009 [219] and maize is replete with a lot of functional genomics data like high throughput expression and genomewide epigenetic datasets (methylation) produced by a community of researchers. Ample amount of so called 'classical' data is also available which includes detailed functional studies of individual genes and a handful of direct genotype to phenotype associations [132]. The maize community is dynamic, vibrant and very open with regard to advice and data

sharing. Of particular interest was the availability of a major population resequencing dataset for maize, the hapmap2 project which not only sequenced modern maize lines but also their wild cousins (Teosinte), thus providing two contrasting datapoints for studying maize domestication [36].

WGD in maize is 'special'

Post divergence from the MRCA, grass species, in general have not encountered a well characterized species wide WGD event, except for maize. A maize specific WGD event was suspected early on by the maize community, but strong evidence based on sequence data was put forward by Gaut and Doebley [73] in 1997. Since then a substantial body of work was already been done on maize WGD, including its dating around 11-12 Mya [241]. I was particularly drawn towards WGD in maize as it is reasonably recent and happened after divergence from sorghum. Sorghum did not undergo a maize specific WGD thereby making it a perfect outgroup with a high quality genome sequence [182]. Older WGDs, for example at the root of the grass lineage make a less striking case as ample time has passed after WGD and most post WGD processes have diminished in strength and signal. Most older WGD are studied using interspecific measures like synonymous and non-synonymous substitution rates (Ka and Ks). The recent nature of maize WGD gives a rare opportunity to study post WGD evolution at a much recent and relevant timescale, particularly using intraspecific comparisons and patterns of polymorphism.

Whole genome sequence of maize not only reconfirmed the WGD but reported about uneven gene loss between duplicated regions and a gradual and still ongoing return of the number of genes to pre WGD level (Fractionation) [219, 272]. This diploidization post-WGD in maize was also associated with copious chromosomal rearrangements, breakages and inversions [272] when compared to related species like sorghum and rice (with no recent WGD) thus making them more representative of the ancestral state of gene order pre-WGD [219, 272]. Using sorghum as an outgroup, and with the maize WGS it became possible to map "which gene went where?" from its pre-WGD state. This catalyzed a string of following studies which explored the post WGD evolution and genome rearrangements at a higher resolution. Many immediate questions were addressed -a) How many genes retain their duplicate copy? b) What is the mechanism of gene loss? Do mutations render one copy non functional making it a pseudogene?. c) Are some genes more or less likely to loose a duplicate copy? d) Is gene loss a random process or some regions display a preference? Woodhouse et al. and schnable et al answered these questions [276, 217] by taking leverage from the availability of high quality genome sequence of both maize and sorghum. They first aligned duplicated copies in maize individually to their ortholog in sorghum and then created syntenic blocks of genes by using the location of gene in Sorghum and maize presuming that the ancestral (pre-WGD) gene arrangement would be similar to that of the sorghum. Approximately 20% of genes retained their duplicate copy which indicates a rapid and massive gene loss following WGD. This also explains why gene numbers in maize are still moderate [218]. The major mechanism for gene deletion was found not to be pseudogenization by mutations, but instead entire gene copies seem to be lost presumably by improper recombination [218]. Once the deletion removes a part of a duplicated gene, the other intact copy resists a deletion in it as there is no complementation by the first copy. The first copy is then

targeted for more progressive deletion events eventually vanishing away from the genome. The exact mechanism for such deletions is still unknown. These studies also reported a bias or non-randomness in the deletion process. Some syntenic blocks tend to loose more genes and were called the sensitive subgenome (maize2 subgenome) and the counterpart called as the dominant subgenome (maize1). The concept of subgenome at first seems to be counterintuitive and artificial as the very definition of a dominant subgenome is fewer deletion events. Also the subgenome can not be associated with individual progenator genomes which merged during the WGD as the definition does not imply this. But three lines of evidence imply biological relevance of subgenomes. First, maize1 subgenome genes were shown to have dominant(higher) expression when compared to their maize2 counterpart [217]. Second that well studied genes which have a known mutant phenotype associated are more likely to fall on the maize1 subgenome [215]. Thirdly, and most importantly, there is not difference between inherent 'neutral' deletion rate between subgenomes as regions in introns and TEs do not display higher deletion frequencies when present in maize2 [217]. This implies that, as the deletion rates are similar, a deletion in maize1 subgenome has a stronger effect on phenotype and is thereby selected against. In other words, differential purifying selection between duplicates can be invoked to explain the subgenome deletion bias and became the basis of the first part of my thesis work. Differences in purifying selection were assayed between duplicates and connected with expression differences in this work.

Gene dosage is well known factor which influences the probability of a gene to be retained post WGD. A proxy for gene dosage are the number of genetic or protein-protein interactions. The more interactions a gene has more tightly under selection it's dosage is. Transcription factors and multiprotein complexes are two functional categories that tend to have above average interaction partners. Maize WGD was no exception in this case with these functional categories showing higher post WGD gene retention probabilities [219]. Another study [105] quantified expression profile divergence of duplicates between two different leaf types of maize and analyzed long-term selection between ohnologs. It was shown that the purifying selection is weaker on one copy when the expression profile of duplicated pairs remains non-divergent across tissues, whereas, when the expression profile diverges in tissues, both copies are under strong purifying selection.

Another aspect of duplication lies in achieving mutational robustness via functional complementation by the duplicate copy. A study in yeast [86] showed that complementation occurs but decreases with sequence divergence between paralogs and the copy with higher dosage exhibits stronger deleterious effect when silenced. Schnable and Freeling [215] reported cases where a mutant phenotype is catalogued for a maize gene with has a duplicated copy present, thereby in these mutants the duplicate copy fails to complement. One part of this work was to connect purifying selection and dominant expression with an associated mutant phenotype.

Transposons in maize

Maize is of special historical significance when studying TEs [194]. Around $\sim$85% of the maize genome is composed of TEs [219]. This is a substantially large fraction compared to closely related Sorghum genome (divergence $\sim$12MYA [241]) [244, 96]. It is now accepted that the increase in

genome size of maize compared to recently diverged sorghum was primarily due to an explosion of TEs [128]. Since their discovery in maize an enormous volume of work has been conducted in maize by identification and characterization of various TE types and unraveling the molecular mechanisms involved in transposition [287, 143, 211, 129]. Continuing the tradition, even the approach for dating a retro-transposon insertion by inspecting the divergence between the terminal repeats was first demonstrated in maize with most TEs inserted ∼3MYA [210]. Maize also has an unusually high number of intact full length TEs [15], implying more active transposition events. It is common to find TE based transcripts in transcriptome surveys [258]. More than half of the genes have a TE in 1KB vicinity [219]. TEs also play a role in shaping the intraspecific diversity in maize intergenic regions [17, 264, 50]. Not only maize nucleotide diversity is on the higher end but also high diversity exists on a coarser scale where large chunks of genome display presence absence polymorphism (PAVs) [233, 240]. More than 3000 genes have been shown to display these PAVs [240]. The violation of genetic collinearity is rampant in non-coding regions [71] making whole sections display no sequence similarity within a population when seen relative to the distance from a gene. Intraspecific differences in transposons seems to be a major reason for this [264]. Increasing afford-ability of deep and long-read technologies combined with newly developed methods for calling TE insertion polymorphisms [56] would catalyze studies to detect these 'non-reference' transposons and catalog polymorphic transposon insertions in maize. Such studies in maize have just started appearing [271].

Several instances have been reported in maize where TEs have influenced the phenotype in an adaptive manner [236, 285, 152]. An upstream TE insertion in *NAC* was shown associated with drought tolerance by lowering the gene expression [152], this insertion was dated to be post domestication. An upstream TE insertion in the famous *tb1* genes acts as an enhancer of gene expression and is shown to be causal for the increased apical dominance in maize [236]. Another upstream TE insertion in the *ZmCCT* gene suppressed gene expression and consequently decreased photoperiod sensitivity thereby enabling flowering in long daylight conditions and was critical for maize adaptation to temperate climates [285]. This insertion has happened post domestication. Another TE upstream insertion near *ZmRap2.7* gene (which is involved in flowering time repression) is known to influence flowering time by regulating the gene expression [209]. Curiously enough, all these TE insertions are located upstream (sometimes as far as ∼70kb), are regulatory in nature and influence gene expression. Also both repression and enhancement of gene expression has been associated with TEs in these studies. Contrasting these pinpointed studies, activation of TEs detected by their expression also has been reported to be associated with conditions related to abiotic stress like salt and drought in maize [151, 50]. A large scale reprogramming of the transcriptome has also been reported in TE active lines [226]. High TE abundance and activity in maize would point to pervasive interaction between TEs and genes. Genes should be under persistent attack and nonfunctional genomic regions near genes constantly eroded away by transposition. TE free upstream regions of genes would mark the boundary where further transposition would alter gene expression and/or be deleterious. Thus the functional genomic regions under purifying selection might leave a 'footprint' in the background TE sequences. This footprint can be seen in genomic sequence of maize were genes form islands in the sea of TEs (Figure 2.5).

Figure 2.5: Repeats and Transposon track (black) and genes (blue) at two levels of resolution.

A very high fraction of TEs is methylated in maize, especially TEs away from genes [178]. A study used a clever technique involving filtering out the methylated DNA before sequencing and found that only around 7 percent of the repetitive DNA in maize is unmethylated [178]. Several genomewide epigenetic datasets have been generated recently [267, 55, 196, 75, 274]. Methylation was shown to be negatively correlated with gene expression in maize [55]. Context specific differences in methylation have been noted in maize wherein TEs proximal to genes have a different methylation (CHH context) vs (CG and CHG) for distant TEs [75]. These CHH islands were proposed as insulators between genes and the dense chromatin in maize.

Maize Domestication

Maize was domesticated from teosinte (a grass species) around ∼10000 years ago in present day Mexico [47, 43]. Maize also in appearance, strikingly different from teosinte [108] which made its sudden appearance a mystery as no wild cousins look similar to modern day maize, at least to the naked eye. But the similar chromosome structure and the ability to cross both teosinte and maize often producing fertile offspring lead George Beadle to promote the 'Teosinte Hypothesis' in 1930s `http://www.maizegenetics.net/genetics-of-domestication` [6]. Maize domestication contrasted the prevalent view that evolution is a slow process but maize community was fast acting in the identification of the type, nature and locations of the genetic changes introduced by domestication and this still continues today [46]. Earlier studies relied on molecular markers like isozymes and microsatellites to look for changes during domestication and indicated towards both multi-genic and single loci implicated in phenotypic differences [44, 44]. In all it turned out to be a mixed bag with both monogenic and multigenic loci and large and small effect changes involved. Longer linkage blocks in maize made the 'zooming in' to the exact gene quite difficult and it still poses a challenge

today but examples of single identified genetic change for the phenotypic change also exist. An example is the famous *tb1* gene which was implicated by early QTL studies in difference in branching patterns (teosinte forming multiple branches whereas maize has a single stem). Auxiliary genomic data showed maize specific higher expression of *tb1* in certain tissues pointing to cis-regulatory elements for the causal variants. Finally the difference was pinpointed to a transposon insertion in the cis-regulatory region which acts as an enhancer of gene expression [236].

Parallel to these are the studies investigating the population genetics aspect of the maize domestication and addressing questions like the magnitude of reduction in diversity, extent of post-differentiation gene flow, selective sweeps, timing of domestication, single v.s multiple domestication events, soft vs hard sweeps etc. Matusoka et.al provided microsatellite based evidence for a single domestication event and timing of around $\sim$9000 years [154]. Maize genetic diversity has been pointed towards being the outlier on the higher end in plant and animal kingdom [246, 282]. This became evident very early with traditional markers like isozymes and RFLPs [47, 43]. To study maize domestication, genetic diversity is best seen in relation to its undomesticated ancestor (teosinte) because domestication typically reduces diversity in comparison to the wild progenitor as few individuals are sampled and are founders to the domesticated population [66]. The loss of diversity is found to be uneven across the genome and regions conferring adaptive phenotypic changes in domestication experience a more severe loss due to both bottleneck and stronger selection [109]. The reporting of the magnitude of diversity reduction in maize due to domestication bottleneck has been heterogeneous amongst studies with both large and small losses reported [278, 247, 246, 57]. The major reason for this being the small part of genome surveyed and sample selection. This situation was ameliorated by the Maize hapmap2 project which involved a genome scale survey of several maize lines [36]. Hapmap2 covered a spectrum of maize evolution by sampling not only 60 modern improved maize inbred lines but also 23 landraces and 19 wild lines. The WILD samples were composed of teosinte which still grows in Mexico (a wild cousin of maize). The LANDRACEs are maize lines traditionally grown by farmers. The modern IMPROVED lines are also called as the 'elite lines' and are commercially produced. These three groups (WILD, LANDRACE and IMPROVED) contain snapshots of progression of maize before domestication to modern commercial breeding. The large sample size and a whole genome coverage helped to obtain a comprehensive estimate of both, the nucleotide diversity and the reduction in diversity in maize post domestication. A SNP based nucleotide-diversity value estimated as average pairwise differences (Π) of 0.0048 was reported for maize and a diversity reduction of about $\sim$17% when compared to teosinte [104]. This presents a contradiction as the massive reduction in diversity typically seen from domestication in most species is not seen in maize [57, 18], yet copious phenotypic changes are evident. This contradiction was noted long before the hapmap2 project by small scale surveys of polymorphisms. Eyre-Walker et al. [57] tackled the problem early on using 'coalescent simulations' and fitting simulated data to polymorphism seen in *Adh1* gene in maize. They reported that a diverse ancestral population of teosinte coupled with a bottleneck of small size and short duration can explain the observed patters. But the study involved a minuscule fraction of the genome. Two more aspects are relevant in this case, first is the out-crossing nature of maize with open pollination and ability to form fertile offspring with teosinte which would imply post

domestication gene flow between teosinte and maize. These were also addressed by Eyre-walker et al. as the study also predicted a small founder population for modern maize, a pervasive gene flow would have not resulted in this result. We have consistently thought about another aspect which is the heterogeneity of the breeding induced selection in maize wherein breeders have imposed a strong selection on certain loci but the rest of the genome was under lesser constraint. This aspect is strengthened by the heterogeneity in the polymorphism data with the nucleotide diversity displaying a high variance in maize [192]. On one hand there are genes devoid of polymorphism whereas on the other there are genes even displaying higher diversity than teosinte [278, 282]. A recent paper in maize reported that maize adaptation is largely shaped by 'softsweeps' [7] wherein the selected variant comes from standing genetic variation as opposed to a new mutation 'hardsweep'. This is relevant as the reduction in diversity is much milder for softsweeps [93].

The population genetics based bottom-up approach also yielded many candidate genes for domestication. For example, the hapmap2 project generated a list of 'domestication' and 'improvement' candidates [104] by using approaches involving scanning the genome for regions of strong local reduction in diversity (measured by extreme allele frequency distribution) [33]. Many studies have used these approaches to detect regions involved domestication and adaptive evolution in maize [111, 282]. During my PhD I was also involved in a genome scan based approach to detect loci involved in flowering time in two germplasm pools in maize [255].

Discussion on selection and diversity can not be complete without recombination and Linkage Disequilibrium (LD). LD is a measure of non-random association of alleles between sites [90], this association when caused by the proximity of the markers on a chromosome is called simply as 'linkage'. Recombination is one of the factors influencing LD as it delinks loci and minimizes the effect an allele on one site in the genome has on the neighboring site. Note that high LD also decreases the efficiency of selection, because selection at one site is affected by selection on other proximal sites (called as Hill-Robertson effect [94]). LD typically decreases with the distance between sites in the genome as the likelihood of a recombination event increases with distance. This rate of 'decay' of LD depends on many factors including recombination rate, mating system and demography (bottleneck) [67]. LD has large implications in shaping patterns of polymorphism by positive and purifying selection. A selective sweep will affect a large chunk of a genome if LD is high. Concurrently the removal of diversity by sweep can increase apparent LD in some cases [157]. A high LD can be a signature of selective sweep and this forms the basis of linkage based methods to detect adaptation [121]. High LD will also mean that the potentially deleterious alleles (eg non-synonymous) loci will affect nearby and neutral sites (eg synonymous). The neutral diversity also gets suppressed purely due to linkage with sites harboring deleterious alleles. This phenomena is called as 'background selection' [29]. LD has special significance in maize as domesticated species experience an increased LD because fewer allelic combinations make it out of bottleneck and domestication associated adaptation causes 'local bottlenecks' [67]. In maize both LD and recombination rate is shown to be population dependent [5, 67]. LD decay in maize has been shown to be rapid [246]. Wright et.al did a polymorphism survey of 774 genes and reported population recombination parameter (ρ which is inversely proportional to LD) in maize to be 17% of that in teosinte, thereby suggesting that the increase in LD was more

drastic compared to decrease in diversity [278]. Maize hapmap2 project also supported the results and reported a substantial increase in LD when compared to teosinte and an average haplotype in maize was $\sim$8Kb longer than in teosinte [104].

Objectives

The overall objective of this thesis was to understand the effects of purifying selection in shaping the maize genome. Three aspects were in this framework studied, namely, the recent WGD in maize, proximity of TEs to genes and domestication induced differences in selection strength.

Polymorphism and gene expression data were first obtained from extensive available whole genome datasets and used to assay strength of purifying selection and expression divergence between duplicates. In the first part of this thesis, I assess whether maize1 genes are under stronger purifying selection than maize2 genes. Dominant gene expression was then linked to stronger purifying selection by developing a new classification scheme for WGD duplicates based on gene expression rather than subgenomes. The extent of divergence between WGD duplicates was also quantified. The cases where a duplicate copy fails to complement a mutant phenotype were obtained from other studies, and this observation was explained in the context of my new classification scheme. Finally, mechanisms driving expression divergence were explored by looking at differences in TEs, splicing and methylation between duplicates.

The second part of this thesis involved identifying the different forces shaping the TE landscape in the vicinity of genes. The questions asked in this case were as follows: How close can a TE from a gene? What determines this distance? First, quantification of TEs in the gene vicinity (TEs upstream and downstream of a gene) was undertaken. Second, the role of purifying selection on the coding region, gene expression and regulatory complexity in shaping this landscape were explored. Regulatory complexity was assayed by tissue-specificity, gene ontology and TSS (Transcription start site) architecture. TE landscape in gene proximity was also obtained for sorghum and compared with maize to explore the influence of increased genome-wide TE abundance on this landscape.

In the third part of this work, the convenient division of individuals in wild, landrace and improved groups by the maize hapmap2 project was utilized to study differences in purifying selection likely to be influenced by domestication in maize. The abundance of shared and private SNPs between groups were obtained. Then purifying selection was assayed on SNPs by obtaining derived allele frequency, nature of SNP (synonymous vs non-synonymous), conservation at the site (SIFT score) and DoFE. Different scenarios were explored to explain the differences in abundance and strength of purifying selection for shared and private SNPs. Differences in strength of purifying selection were then inferred for these different groups. Differences in recombination events were also obtained for three groups, and purifying selection differences were assessed by bins of recombination. This analysis reveals the role of recombination in increasing the efficiency of selection. The differences between groups was then explained in the context of increased LD due to the domestication induced bottleneck.

In the future perspectives section I provide a general discussion of three aspects and suggest new avenues for future work.

Whole Genome Duplication in Maize

Materials and Methods

Obtaining SNP data

SNP data was obtained from from the maize hapmap2 project [36] which contains whole genome SNP data for 19 Teosinte, 23 Landraces and 60 Modern inbred lines. For this part of the thesis the data shown is from the teosinte lines only as they are the closest approximation to a panmictic population. VCF (Variant call format) file for maize hapmap2 SNPs was downloaded from the url (`data.iplantcollaborative.org/quickshare/e75bc315fc0f9fda/HapMapV2RefgenV220120328.vcf.gz`). The 19 teosinte lines contain 17 lines from subspecies parviglumis and two from subspecies mexicana. The lines belonging to mexicana where removed because it forms a different subspecies. Two more lines TIL04-TIP285:TEO and TIL06-TIP496:TEO were removed as they are similar lines to TIL04-TIP454 and TIL06-TIP260 already present and differing only in the generations of selfing. One more line TIL02 was removed due to low coverage. The resulting vcf containing 14 teosinte lines was annotated using the program snpEff [39]. The following snpEff annotations namely START_LOST, STOP_GAINED, FRAME_SHIFT, STOP_LOST were classified as Very Deleterious mutations (VDMs).

Calculating nucleotide diversity

The VCF file was converted to 'hapmap' format using a custom perl script. Variscan [259] was used with runmode 12 to get values for nucleotide diversity (estimated as the average pairwise difference between (Π)) for each SNP. SNPs with less than 80% of genotypes called were removed from the calculations. Nucleotide diversity reported was based on average pairwise differences (Π, see introduction). The per SNP diversity was added and the sum divided by the total number of sites for each category to obtain per basepair value of nucleotide diversity.

Calculating sequencing depth for genes

Since the calculations for nucleotide diversity can be influenced by the read coverage, it was calculated the for each retained gene. Hapmap2 BAM (binary alignment map) files for each line used in the analysis were downloaded from `mirrors.iplantcollaborative.org/download/iplant/home/shared/panzea/hapmap2/bam/`. Bedtools [189] was used to get the read depth at each position of genes. Since coverage varies at each position for different members of sequenced lines a 'mean coverage fraction' was calculated which is the sum of the fraction of individuals at each site which at least has one read aligned normalized by the length of the gene. Mean coverage fraction for UED-dominant genes was 0.70 vs 0.67 for BED-repressed, which indicates slightly low coverage for repressed genes. This might be caused by higher divergence or higher content of repeats or transposable elements. Since UED-repressed genes were reported in this study to have higher nucleotide diversity, the computation is an underestimate as we might miss SNPs in those genes due to the low coverage. Thus read coverage does not alter the direction of results, on the contrary makes them more conservative.

Calculating DoFE

DoFE was calculated separately for maize1 and maize2 retained genes. The Eyre-Walker and Keightley method as implemented in the software DoFE was used [59]. The software was downloaded from Adam Eyre-Walker's lab (`http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/Website/Software.html`). Non-synonymous SNPs were used as the deleterious class and the synonymous SNPs the neutral. The nature of the SNP (non-synonymous v.s synonymous) was obtained from snpEff output.

Obtaining Ka and Ks

The list of genes belonging to each subgenome with their Sorghum ortholog was downloaded from James Schnable's webpage (`skraelingmountain.com/datasets.php`). For each gene the splice variant with the longest protein coding sequence was chosen for further analysis. Ka and Ks values for maize-sorghum orthologs were obtained from ensemble biomart website (`plants.ensembl.org/biomart/martview/`) [124]. Ka and Ks values < 0 and > 0.5 were excluded from the analysis.

Obtaining SIFT scores

SIFT is an algorithm for predicting the deleterious effect of non-synonymous polymorphisms [170]. It is based on site conservation and nature of amino acid change caused by the SNP. It gives a score between 0 and 1, with 0 being intolerant and 1 being neutral. Typically a score between 0 and 0.05 is considered to be deleterious [223]. Sift scores for maize were downloaded from the sift4g website (`siftdb.org`).

Expression data

Expression data for maize inbred line B73 in the form of FPKM (Fragments per kilobase per million mapped reads) values for diverse tissues was downloaded from the qteller website (`qteller.com/qteller3/`). The qteller website gives expression for maize genes combining data from various studies. The tissues with expression data under peculiar local conditions (e.g. drought) were removed from analysis. Relative expression of ohnolog pairs was calculated and a two fold expression difference was used as a threshold to define dominance in expression. If both paralogs have expression $<$ 0.5 FPKM in a particular tissue that comparison was discarded. Combining of the data from different studies might cause some differences in expression to be due to differences in studies gathering the expression data. We would like to stand by this decision due to three reasons. Firstly and importantly only relative expression between two duplicate copies was compared and no comparison of expression values of genes across tissues was made. Secondly, it is important to sample many tissues to obtain accurate assignment of BED genes which is only possible by using expression data from different studies. Thirdly, FPKM is a normalized measure of expression which takes into account the sequencing depth. Note that a similar protocol is used for aligning reads and obtaining the FPKM values in the qteller website.

Gene ontology analysis

Gene ontology analysis was done using AgriGo tool [51]. "Single Enrichment Analysis" option of AgriGo was used with all maize genes with syntenic orthologs as background. This list was obtained from James Schnable's website (`skraelingmountain.com/datasets/grass_syntenic_orthologs.csv.zip`) [216]. A false Discovery Rate of < 0.05 was used for a given enrichment to be significant.

Upstream transposable elements

Information about the annotated repeats in maize AGPV2 assembly were obtained from (`ftp.maizesequence.org/release-5b/repeats/`). Most repeats were annotated as transposable elements. These files were combined and converted to a BED (browser extensible data) format. Bedtools [189] 'coverage' command was run to obtain fraction of 2kb upstream regions covered by repeats/transposable elements. The Bedtools 'closest' command was used to obtain the distance from the transcription start site to the nearest transposable element.

Obtaining methylation Data

Methylation data for maize line B73 was obtained from Eichten et al. [55]. They determined array based methylation states for non-repeat regions in the maize genome. This information was downloaded using the Genomaize website [62] as a bigwig format file which converted to bedGraph format using bigWigToBedGraph utility [120]. Methylation values for upstream regions of genes were obtained by using bedtools 'intersect' command.

Obtaining splicing data

The geneome wide data for known and novel splice variants for maize line B73 was obtained from Thatcher et.al [248].

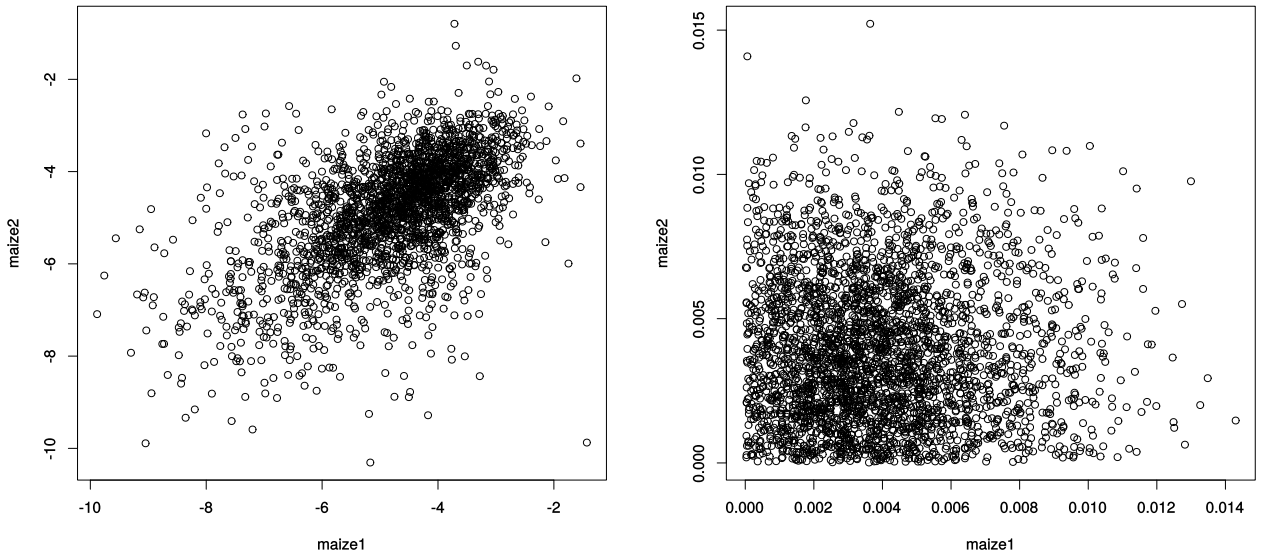
Statistical analysis

All statistical analyses were done using R (http://www.R-project.org). Since most parameters used in the study follow a non normal distribution, nonparametric tests were used for significance and medians reported instead of means. R package 'boot' (cran.r-project.org/web/packages/boot/index.html) was used to calculate confidence intervals for statistics but they were found to be very small and were not plotted. A visual inspection of the data for extreme values was done to ensure that statistical significance was not caused due to them.

Results

Nucleotide diversity between duplicates is correlated

Nucleotide diversity (estimated as per site average pairwise difference Π) was found to be correlated between duplicate gene pairs, (particularly strongly for introns) (Table 3.1, Figure 3.1). In addition, both non-synonymous (Ka) and synonymous divergence (Ks) calculated with respect to sorghum orthologs were also correlated (Spearmans rho 0.71 P<2.2e-16 and 0.73 P<2.2e-16 respectively). Interestingly, the diversity in upstream regions (2kb) was not correlated indicating a decoupled upstream evolution of the duplicates.



(a) Introns

(b) Upstream 2KB

Figure 3.1: (a) Correlation plot of logarithm of nucleotide diversity for introns for maize1 and maize2 gene pairs. (b) Correlation plot for nucleotide diversity of upstream (2KB) regions of duplicated pairs

Table 3.1: Nucleotide diversity between duplicate pairs is correlated for introns, synonymous and non-synonymous sites but not for upstream regions (*) P < 2.2e-16.

Region	Correlation R value for nucleotide diversity
Intron	0.56*
Synonymous	0.24*
Non-Synonymous	0.19*
gene	0.21*
Upstream	-0.01 (Not-significant)

Maize1 subgenome genes are under stronger purifying selection

Different measures of purifying selection indicated a consistent stronger purifying selection on the maize1 subgenome genes. The median Ka for maize1 retained copy was 0.039 vs 0.043 for maize2 retained copy (P=9e-6 Wilcoxon rank sum test). The median Ks was 0.195 vs 0.193 respectively (P=0.3). Occurrence of Very Deleterious Mutations (VDMs) between subgenomes was then looked into. Definition of VDMs include frame shift, stop-gained and stop lost. Out of 3228 retained gene pairs, 1003 in maize1 had a VDM in at least one sampled accession vs 1140 for maize2 subgenome (Chi Square P<0.0003). The per site nucleotide diversity for non-synonymous sites was found to be higher for maize2 genes compared to maize1 genes (0.0017 vs 0.0020 P=1E-6 Wilcoxon rank sum test), but the difference in synonymous diversity was not found to be significant, (0.0088 vs 0.0087 P=0.8 Wilcoxon rank sum test) indicating no inherent difference in mutation rate between subgenomes. SIFT scores [170] were further used which quantify the deleterious nature of a SNP based on the conservation at the site of the SNP. SIFT score < 0.01 was considered to be deleterious in nature. A gene with more than one SNP with score less than 0.01 was considered to be harboring deleterious mutations. Significantly more genes harboring deleterious mutations were found to be located in the maize2 subgenome (Table 3.2, Chi-Square P=5.5e-6). Distribution of Fitness Effects (DoFE) were then obtained for both maize subgenomes and are depicted in Figure 3.2. A higher fraction of mutations in highly deleterious class ($-$NeS > 100) for maize1 subgenome indicated of stronger purifying selection acting on it (Figure 3.2).

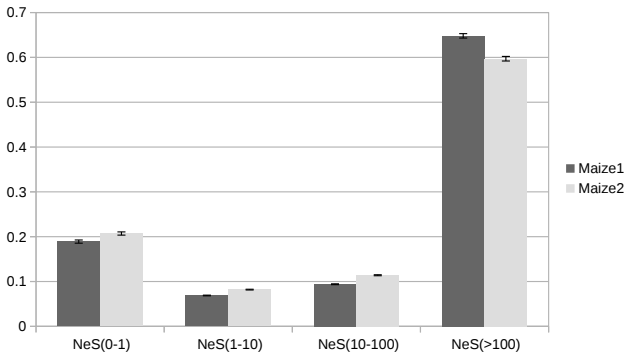


Figure 3.2: Distribution of fitness effects (DoFE) for WGD retained genes of two subgenomes.

Table 3.2: Number of genes classified as harboring deleterious (gene with > 1 SNP with a SIFT score < 0.01) and non-deleterious SNPs based on SIFT scores.

	Genes with > 1 deleterious mutations	Genes harboring < 2 deleterious mutations
Maize1	773	2185
Maize2	944	2056

Retained genes in maize show expected patterns of expression and Gene ontology (GO) enrichments

Studies have indicated that highly expressed genes tend to be preferentially retained as duplicates after WGD [27, 199, 2]. To test this in maize specific WGD, differences in the amount of gene expression (absolute dosage) between single genes vs 3228 gene pairs retained after WGD across 22 tissues were obtained (Figure 3.3). Retained genes showed consistently higher median expression in all tissues tested.

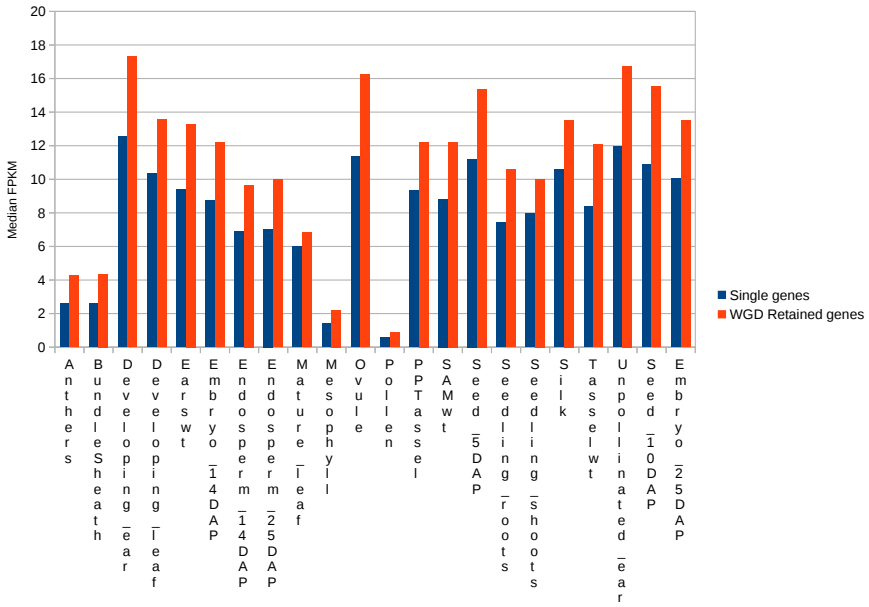


Figure 3.3: Median expression values (FPKM) for retained vs single copy genes for 22 tissues used in this analysis. Retained genes show consistent higher expression across all tissues compared to single copy genes. All comparisons are significant at P $<$ 10e-14 (Wilcoxon rank test) except for Mature Leaf which is significant at P=7e-5. All comparisons were also significant assuming a bonferroni correction.

In addition to absolute dosage, genes whose loss creates more dosage imbalance in regulatory and protein interaction networks and multi-protein complexes are more likely to be retained after WGD [279, 8, 13, 150]. Similar results were found in the gene ontology analysis of single copy vs WGD retained genes in maize. GO terms pertaining to regulation, transcription factors and macromolecular complexes were enriched in retained genes (Figures 3.4 and 3.5). For single genes only catalytic activity was enriched (Figure 3.6)

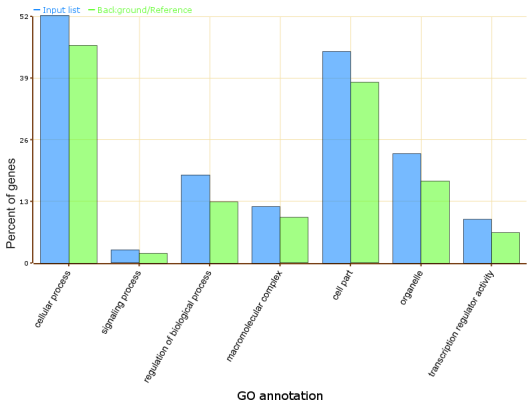


Figure 3.4: Significantly enriched top level Gene Ontology categories for WGD retained genes (FDR < 0.05). Background is maize genes with syntenic orthologs in other grass genomes.

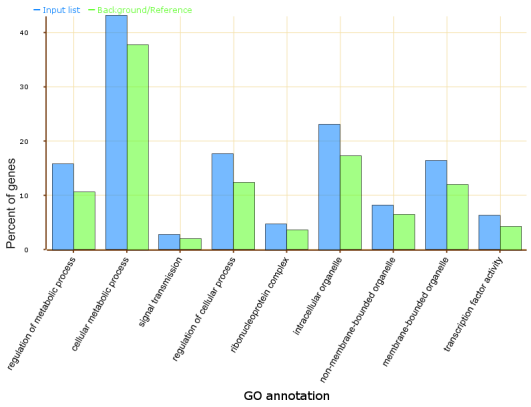


Figure 3.5: Significantly enriched second level Gene Ontology categories for WGD retained genes (FDR < 0.05). The background is composed of maize genes with syntenic orthologs in other grass genomes.

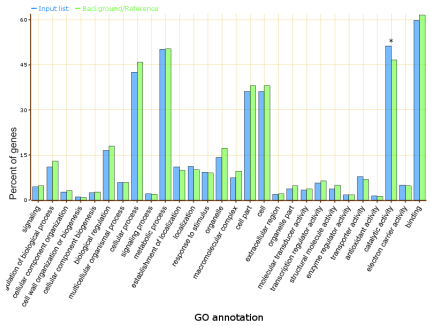


Figure 3.6: Gene Ontology categories for single copy genes (*)FDR$<$0.05. Only catalytic activity was found to be enriched in single genes. Background is composed of maize genes with syntenic orthologs in other grass genomes.

Classification of expression divergence between duplicates

The relative expression of duplicate pairs in various maize tissues was used to classify the expression divergence as unidirectional or bidirectional. Gene pairs with at least twofold difference were considered to be diverged in expression. Of 3228 retained gene pairs, 1641 pairs showed a unidirectional expression divergence (UED) where one member of the pair has a consistently higher expression in all tissues with a twofold difference (UED-dominant) compared to the counterpart (UED-repressed). 1517 pairs showed a bidirectional expression divergence (BED) where one copy displayed higher expression in one tissue and the other copy in another tissue. Thus close to 98% of the retained gene pairs seem to have diverged in expression or have divergent expression pattern across tissues. Given the recent nature of maize WGD the divergence in expression seems to have happened in a relatively short evolutionary timescale.

UED and BED genes form distinct subsets in GO enrichment

Separate GO enrichment analysis of UED and BED genes was done. Transcriptional regulation was found to be enriched for BED genes (Figure 3.7) and macromolecular complexes and structural molecule genes to be enriched for UED genes (Figure 3.8). A complete cross comparison of gene ontologies between these two datasets is available in the table 3.4. Comparing the gene expression of dominantly expressed UED and BED genes in each tissue revealed that UED-dominant genes have higher higher than the dominantly expressed BED gene in all except one tissue (Figure 3.9). Thus the overall expression of UED-dominant genes is higher than BED genes. Gene ontology enrichment of UED-Dominant/UED-repressed genes can be explained in the light of gene balance hypothesis where the maintenance of stoichiometric ratios between interacting partners is under strong selection [13] and is achieved by repression of one copy. For the transcription factors and regulators it is achieved by tissuewise subfunctionalization of duplicate genes. complexes.

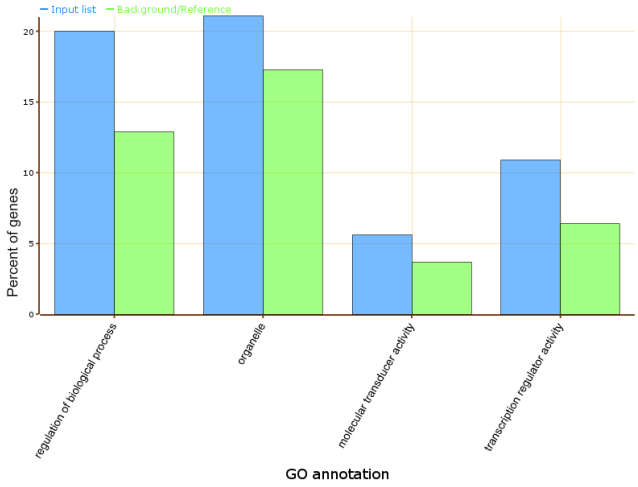


Figure 3.7: Significantly enriched top level Gene Ontology categories for BED genes (FDR < 0.05). The background is composed of maize genes with syntenic orthologs in other grass genomes.

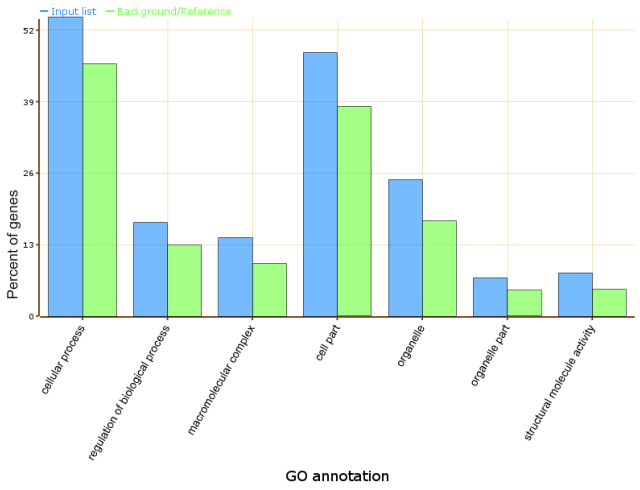


Figure 3.8: Significantly enriched top level Gene Ontology categories for UED genes (FDR < 0.05). The background is composed of maize genes with syntenic orthologs in other grass genomes. The "cellular process" and "cell part" include categories of proteosome, ribonucleoprotein complex, cytoskeleton organization, macromolecule localization (list not exhaustive).

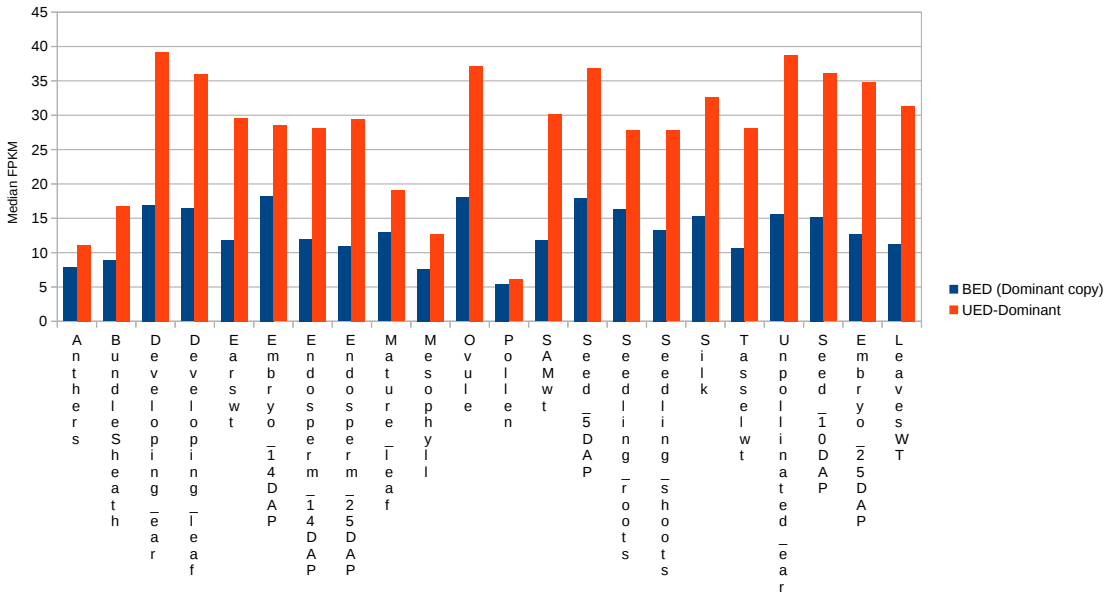


Figure 3.9: Gene expression (median FPKM) per tissue for dominantly expressed BED genes (blue) vs UED (red). UED genes have higher gene dosage compared to dominantly expressed BED genes. All comparisons significant at P<0.0005 except for Pollen which was not found to be significant. All comparisons except pollen were also significant assuming a bonferroni correction.

Increase in purifying selection from repressed to dominantly expressed genes

The ratio of non-synonymous to synonymous nucleotide diversity ($\Pi n/\Pi s$) progressively increases from UED-dominant to BED to UED-repressed genes (Figure 3.10) indicating gene dosage to be a strong determinant of purifying selection. Maize1 genes also exhibit lower ratio of non-synonymous to synonymous diversity than maize2 genes, but the effect is not as strong as for UED-dominant vs

UED-repressed genes (Figure 3.10). As nucleotide diversity data from teosinte lines was used in this work, while the expression data is from modern maize line B73, this analysis was repeated using nucleotide diversity data from modern inbred lines (of which B73 is a member) and qualitatively similar results were found (Figure 3.11).

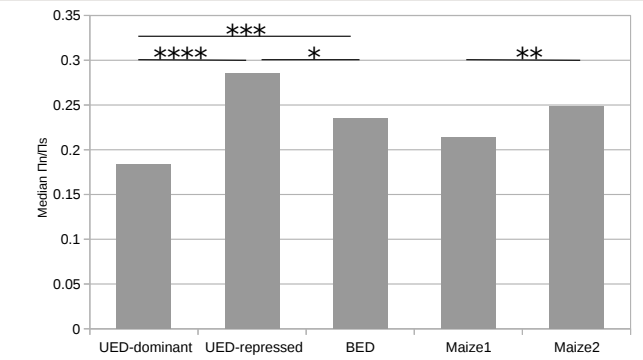


Figure 3.10: Median of ratio of non-synonymous to synonymous diversity $\Pi n/\Pi s$ compared between different datasets. Increase in strength of purifying selection from UED-repressed to tissuewise subfunctionalized (BED) to UED-dominant genes. P-values were calculated using Wilcoxon rank sum test (****)P<2.2e-16;(***)P=1.9E-15 ;(**)P=2e-6;(*)P=3e-4

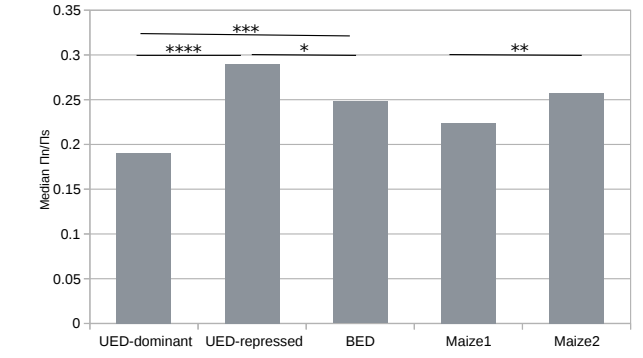


Figure 3.11: Median of ratio of non-synonymous to synonymous diversity Πn/Πs for 60 inbred lines compared between different datasets. Increase in purifying selection from UED-repressed to tissuewise subfunctionalized (BED) to UED-dominant genes. P-values were calculated using Wilcoxon rank sum test (****)P<2e-16;(***)P=2e-14;(**)P=2e-7;(*)P=5e-5

Stronger purifying selection on maize1 subgenome only exists for BED genes

The existance of subgenome dominance (maize1 genes having stronger purifying selection) was tested for expression based gene classes (UED-dominant, UED-repressed and BED). Increased purifying selection (measured by Πn/Πs) was not found in maize1 UED-dominant genes compared to maize2 UED-dominant genes and for maize1 UED-repressed genes compared to maize2 UED-repressed genes (Figure 3.12). Subgenome dominance in purifying selection thus largely originates from the dominant expression of UED-dominant genes compared to UED-repressed genes as there are more maize1 UED-dominant genes than maize2 (948 UED-dominant genes are maize1 compared to the expected value of 820 (1641/2)). Expression thus seems to be a dominant determinant of purifying

selection overriding the subgenome dominance for UED genes. The presence of subgenome dominance in purifying selection for BED genes is intriguing as shown by maize1 BED genes having significantly low ratio of non-synonymous to synonymous diversity (Figure 3.12). To dissect this further, the number of tissues in which maize1 and maize2 BED genes dominate in expression was calculated. Maize1 BED genes dominate the maize2 BED genes in expression in a larger number of tissues (Figure 3.13). The median number of tissues in which maize1 BED gene dominates in expression is 6 compared to 4 for a maize2 BED gene (P=1.836e-12 Wilcoxon rank sum test). Overall this suggests that dominant expression either consistently or in a larger number of tissues is a determinant for stronger purifying selection rather than subgenome location.

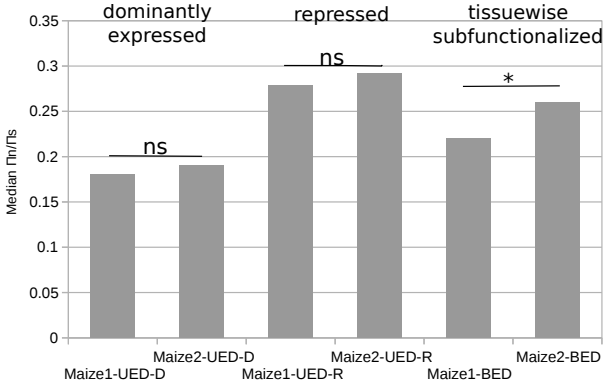


Figure 3.12: Ratio of nonsynonymous to synonymous nucleotide diversity (Πn/Πs) for maize subgenome 1 and 2 genes for different expression classifications. UED-Dominant (UED-D) and UED-repressed (UED-R). P-values were calculated using Wilcoxon rank sum test (*)P = 2.9e-4, (ns) not significant

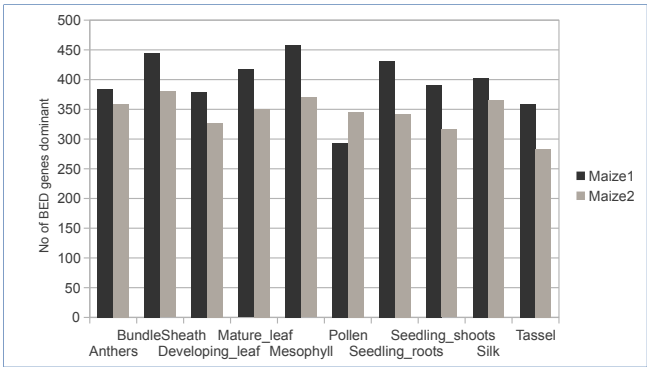


Figure 3.13: Number of maize1 and maize2 BED genes dominantly expressed in each tissue.

Upstream regions of repressed genes are enriched in TEs

Upstream transposable elements (TEs) have been proposed as a mechanism for generating differences in expression between paralogs [70]. To test this hypothesis in maize, the fraction of upstream 2kb region covered by annotated repeats and TEs was calculated. The mean(median) coverage for UED-repressed genes was 0.33(0.27) compared to the coverage of UED-dominant genes being 0.26(0.19)

(P=9.2e-13 Wilcoxon rank sum test) and BED genes 0.27(0.20) (P=2.9e-10). However the coverage difference between maize1 and maize2 was not found to be significant (0.28(0.2) vs 0.29(0.23)) (P=0.02) at P<0.01. Also, the nearest upstream distance between the annotated transcription start site and the upstream annotated TE was compared between different classes. The mean(median) upstream distance of an UED-repressed gene from an annotated repeat or TE was 908(480) base pairs (bp) compared to UED-dominant genes 1102(610)bp (P=1.3e-8 Wilcoxon rank sum test) and BED genes 1019(597)bp (P=4.8e-8) (Figure 3.14). The distribution of this distance was not found to be significantly different between UED-dominant and BED genes (P=0.24). Interestingly, this distance was also not found to be significant between maize1 and maize2 retained genes 1039(587) bp vs 975(564) bp (P=0.06 Wilcoxon rank sum test). Thus the repressed genes (UED-repressed) not only have higher fraction of upstream regions covered by TEs but also have the nearest distance to an upstream TE.

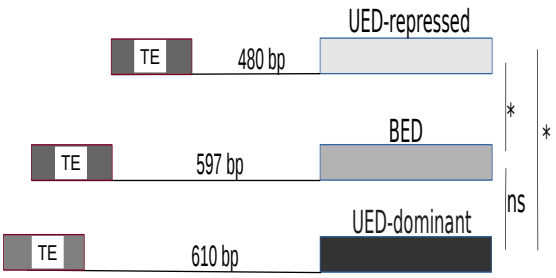


Figure 3.14: Median of the nearest upstream distance to a transposable element (TE) for different expression categories. (*)P<1e-7 Wilcoxon rank sum test;(ns) not significant

Genes displaying mutant phenotype are broadly expressed and are under purifying selection

Number of tissues with dominant expression and strength of purifying selection (ΠIn/ΠIs ratio) where obtained for cases where one copy of the WGD paralogous pair displays a mutant phenotype while the other copy has no visible mutant phenotype associated (thus the duplicate copy fails to complement). In all except two out of 15 such cases found, the member displaying the mutant phenotype is dominant in more tissues and has lower ratio of nonsynonymous to synonymous diversity (Table 3.3). This indicates that paralog displaying mutant phenotype is under stronger purifying selection than its counterpart and this is intertwined with the dominant expression of the copy displaying the phenotype.

Table 3.3: Table compares expression and ratio of non-synonymous to synonymous diversity for 15 paralogous gene pairs where only one paralog displays a mutant phenotype. The gene of the paralogous pair which displays a mutant phenotype generally shows dominant expression in larger number of tissues and has lower ratio of non-synonymous to synonymous diversity (barring two cases). (NA) Not available are the cases where no non-synonymous SNP was found in the gene making Π n/Π s zero.

Paralog	**Mutant Phenotype**	**Transcript id**	**Subgenome**	**No of tissues with dominant expression**	**Π n/Π s**
Paralog1	yes	GRMZM2G377215_T01	Maize1	21	0.108
Paralog2	no	GRMZM2G006830_T01	Maize2	0	0.548
Paralog1	yes	GRMZM2G060216_T02	Maize1	5	0.118
Paralog2	no	AC232238.2_FGT004	Maize2	3	0.271
Paralog1	yes	GRMZM2G307119_T01	Maize1	5	NA
Paralog2	no	GRMZM2G458437_T01	Maize2	0	0.354
Paralog1	yes	AC205703.4_FGT006	Maize1	4	0.158
Paralog2	no	GRMZM2G087323_T01	Maize2	3	0.11
Paralog1	yes	GRMZM2G042992_T01	Maize1	21	NA
Paralog2	no	GRMZM2G124115_T01	Maize2	0	0.647
Paralog1	yes	GRMZM2G160565_T02	Maize1	9	0.021
Paralog2	no	GRMZM2G003514_T01	Maize2	1	0.352
Paralog1	yes	GRMZM2G092542_T01	Maize1	2	0.132
Paralog2	no	AC149818.2_FGT009	Maize2	0	0.581
Paralog1	yes	AC233950.1_FGT002	Maize1	7	0.089
Paralog2	no	AC190734.2_FGT003	Maize2	0	0.663
Paralog1	yes	GRMZM2G036297_T01	Maize1	8	0.186
Paralog2	no	GRMZM2G058588_T01	Maize2	0	0.440
Paralog1	yes	GRMZM2G455809_T01	Maize1	8	0.049
Paralog2	no	GRMZM2G308351_T01	Maize2	1	0.141
Paralog1	yes	GRMZM2G104843_T01	Maize1	16	0.058
Paralog2	no	GRMZM2G070092_T01	Maize2	1	0.105
Paralog1	yes	GRMZM2G098813_T03	Maize2	6	NA
Paralog2	no	GRMZM2G180190_T01	Maize1	4	0.131
Paralog1	yes	GRMZM2G109987_T01	Maize2	3	0.036
Paralog2	no	GRMZM2G042250_T01	Maize1	0	0.008
Paralog1	yes	GRMZM2G039155_T05	Maize2	22	0.098
Paralog2	no	AC194174.3_FGT003	Maize1	0	1.47
Paralog1	yes	GRMZM2G307906_T01	Maize2	0	0.195
Paralog2	no	GRMZM2G100620_T02	Maize1	12	NA

UED-repressed genes have fewer splice variants

Thatcher et al. [248] used RNA-Seq libraries from multiple tissues of Mo17 and B73 and validated many known splice variants and discovered many new ones expanding the diversity of maize transcriptome. The number of splice variants (known+novel) was calculated for genes in each category (UED-dominant, UED-repressed and BED). UED-repressed genes had significantly fewer number of known and novel splice variants. The mean number of splice variants for UED-Dominant genes was 3.5 whereas for UED-repressed was 2.86 (Figure 3.15, P<2.2e-16 Wilcoxon rank test). However differences in number of splice variants between maize1 and maize2 genes was not found to be significant.

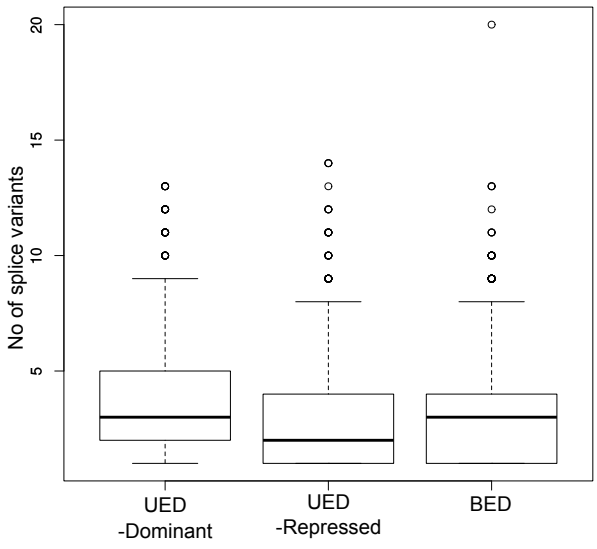


Figure 3.15: Boxplot of distribution of number of splice variants (known+novel) per gene for UED-repressed,UED-dominant and BED genes. Consistently repressed (UED-repressed) genes produce fewer splice variants.

Difference in methylation between repressed and dominant genes

A genomewide study of methylation in maize reported finding no differences in gene body methylation between maize1 and maize2 subgenomes [55]. Methylation values were obtained for the upstream regions of genes in out study. As expected, no significant differences in methylation was found between maize1 and maize2 genes (P=0.1447; Wilcoxon rank sum test). But interestingly, methylation differences were significant between expression categories with the UED-repressed genes displaying higher upstream methylation compared to UED-dominant with a median of -0.602 v.s -0.950 (P=3E-3; Wilcoxon rank sum test).

Discussion

The recent nature of WGD in maize and the availability of a comprehensive polymorphism dataset gives an unique opportunity to investigate subgenomes and evolution of duplicate genes at a population level. Genome scale correlations in nucleotide divergence and diversity were found between duplicate gene pairs (Table 3.1 and Figure 3.1) but no correlation for upstream (2KB) regions indicating that promoter evolution between duplicates is uncoupled. These correlations mean a general similarity in mutabilities, mutation rates and at least a component of protein structural and functional constraints to be similar amongst duplicates. Moreover both subgenomes share the same demography. These correlations are particularly striking given the $\sim$5-12 million years of post WGD divergent evolution and the volatile nature of maize genome with copious genome rearrangements and abundant transposon activity [219]. Differing recombination rates, intensities of positive and purifying selection would weaken these correlations. The retained genes in both subgenomes could be used as two identical samples for demographic analysis and extreme differences in nucleotide diversity between them can be distilled into a test for positive selection.

Studies have predicted maize1 genes to be under stronger purifying selection because of the dominant expression [217] and higher tendency of genes displaying mutant phenotypes to be located in maize1 subgenome [215]. This study shows that this difference is also reflected in the patterns of polymorphism. Maize1 genes exhibit less non-synonymous divergence and diversity compared to synonymous divergence and diversity. DoFE analysis shows an excess of large effect mutation class ($-\text{NeS} > 100$) in maize1 subgenome (Figure 3.2). Furthermore SIFT predictions show a significantly higher number of genes harboring deleterious polymorphisms in the maize2 subgenome (Table 3.2). The later part of the study shows that this effect largely arises from dominant expression of the maize1 gene copies.

Absolute gene dosage was shown to be a major determinant for gene retention after WGD in maize. Genes retained after WGD in maize exhibit a higher median expression than single genes across many tissues (Figure 3.3). This is consistent with studies from diverse organisms [2, 32, 27]. In addition to dosage, typically networked proteins like transcription factors and macromolecular complexes are more likely to be retained after a WGD [150, 279, 8]. We find similar trends in gene ontology analysis of retained genes (Figure 3.4 and 3.5) indicating that established features behind gene retention after WGD are also working in maize.

There exist two major mechanisms for ohnolog retention. First the dosage balance model in which relative dosage of interacting genes is under purifying selection and deletion of one member of the pair creates stoichiometric imbalance in the interaction network. The second is the subfunctionalization model where ancestral function is subdivided between both members. Relative dosage between paralogous pairs in various tissues was used to partition the genes in two major categories consistent with both models. The UED (unidirectional expression divergence) gene category where one paralog has consistently dominant expression (UED-dominant) across all tissues with divergent expression compared to the other paralog (UED-repressed). The second category being bidirectional expression divergence (BED) for which both paralogs have alternatively dominant and repressed expression but

in different tissues. The UED genes comply with the dosage balance model where dosage of one member is consistently reduced ameliorating the dosage constraints and making the repressed gene dispensable. The BED genes comply with the tissuewise subfunctionalization model where both paralogs perform the function albeit in different tissues. Around $\sim$98% of the duplicate pairs could be classified indicating that sorting of duplicate genes by expression is almost complete and most of the duplicate gene pairs have either diverged in expression consistently or in a tissuewise manner. Such rapid divergence in gene expression between duplicates has been observed for yeast and humans [85, 91] and can even be established in a few generations after the formation of an allotetraploid [65]. These classifications also show strikingly different gene ontology enrichments (Table 3.4). UED genes tend to be part of macromolecular complexes and structural molecules (Figure 3.8) whereas, BED genes are more likely to be involved in transcription regulation (Figure 3.7). These results indicate that tissuewise subfunctionalization seems to largely influence the regulatory network while consistent repression influences the protein-protein interaction network. The agreement between UED genes and relative dosage model could be further be tested by overlaying the protein-protein interaction network on UED genes and checking if members of the duplicated pathways divide equally between UED-repressed and UED-dominant. The differences indicate that tissue specific subfunctionalization is a major mechanism for retention of regulatory genes and suppressing the expression of one copy across many tissues is the mechanism for retention of genes involved in macromolecular complexes.

The UED-dominant genes show the most purifying selection followed by BED genes and then the UED-repressed genes (Figure 3.10), indicating that absolute gene dosage is positively correlated with the strength of purifying selection. Since the strength of purifying selection for tissuewise subfunctionalized (BED) genes seem to be intermediate between UED-dominant and UED-repressed we checked if the dosage of dominantly expressed BED genes in a particular tissue is less than dosage of the UED-dominant genes. This was found to be true. Dominantly expressed BED genes have lower expression than BED-dominant genes in all tissues assayed except pollen (Figure 3.9). Overall, genes with higher dosage appear to suppress one copy consistently whereas, genes with lesser gene dosage evolve tissue specific regulatory patterns.

The subgenome dominance in purifying selection is not significant for UED genes and only exists for BED genes (Figure 3.12). In other words, when the genes are classified as UED-dominant and UED-repressed the strength of purifying selection does not appear to be different between maize1 and maize2 genes in these categories. Thus the purifying selection results from the dominant expression and maize1 genes appear statistically to be under stronger purifying selection because there are more UED-dominant maize1 genes. The tissuewise subfunctionalized genes (BED) genes still show a difference in purifying selection between maize1 and maize2 genes (Figure 3.12). However, for BED genes a maize1 BED gene is more likely to be dominantly expressed in larger number of tissues compared to its maize2 counterpart (Figure 3.13). Thus the factors determining the strength of purifying selection are not only the dominant expression but also the number of tissues in which the gene is dominant and expression dominance overrides the subgenome dominance. It could be possible that the broadly expressed copy retains the original expression pattern whereas other copy

gains a new tissuewise expression profile or loses some part of the ancestral expression profile. It is hard to distinguish between such subfunctionalization and neofunctionalization unless the ancestral state of expression is known.

The WGD paralog genes pairs of which one member is associated with a mutant phenotype present an enigmatic case wherein the duplicated copy does not seem to complement the function. However this work shows that the gene of the pair with a mutant phenotype is generally dominant in a larger number of tissues and has lower ratio of non-synonymous to synonymous diversity (Table 3.3) indicative of stronger purifying selection acting on it. It can be suggested that the distribution of function amongst paralogs is asymmetric with one paralog handling larger component of the function. This explains why the duplicate copy fails to complement.

Nevertheless the mechanism behind the subgenome dominance is still not understood. Dominant expression can explain that the deletion of a nondominantly expressed ohnolog incurs a lower fitness cost due to relaxed dosage constraints. However the reason why maize1 subgenome has more dominantly expressed genes is yet to be answered. Epigenetic effects like chromatin modification and DNA methylation have been proposed as underlying mechanisms but have not been proven. Reduced expression of genes caused by upstream transposable elements (TEs) can cause a subgenome dominance if the two progenitor genomes in an allotetraploid have differing number of TEs at start [70]. A part of this hypothesis was tested by associating upstream TEs with the different expression categories and an enrichment of TEs in upstream regions of repressed genes (UED-repressed) was found (Figure 3.14). Repression of these TEs could inadvertently cause the repression of nearby genes. Note though that a causality can not be established from this analysis and upstream transposon abundance can as well be a consequence of low expression and reduced purifying selection. A recent study reported an unexpected correlation between expression of nearby genes which extends to more than 100Kb [77], which points to mechanism generating subgenomes.

The study also reports significantly higher number of splice variants for UED-dominant genes when compared to UED-repressed genes indicating that repression in expression might be partly caused by loss of one or more splice variants. Loss of splice variants soon after duplication has been documented in humans [237] and it could be a part of subfunctionalization process. Divergence in splicing has also been shown for Arabidopsis ohnologs [288].

Overall, this part of the thesis concludes that majority of ohnolog pairs from the maize WGD have diverged in expression either consistently or in a tissuewise manner. This divergence seems to prevent functional complementation by the duplicate copy. Relative and absolute gene dosage and the number of tissues in which a gene is dominant is an important determinant of purifying selection. Transcription regulators are more likely to develop tissue wise subfunctionalization while macromolecular complexes tend to suppress one duplicate copy consistently. An expression based classification of duplicates is found to be more biologically relevant than subgenomes. Finally, upstream divergence and TEs form a possible mechanism for the expression divergence.

Appendix

Table 3.4: Cross comparison of gene ontologies for BED and UED genes. First entry is is for BED genes and second for UED. The table was generated using Agri-Go cross comparison of gene ontolgoies (SEACOMPARE option). Please see the next page for the table. First entry in comparison is BED (ID:458762418) and second entry is UED (ID:133289173).

[Suppress this table]

GO Information				CM		ID:458762418		ID:133289173	
No	GO Term	Onto	Description	1	2	FDR	Num	FDR	Num
1	GO:0009889	P	regulation of biosynthetic process			2.5e-14	283	0.04	231
2	GO:0010556	P	regulation of macromolecule biosynthetic process			2.5e-14	283	0.04	231
3	GO:0031326	P	regulation of cellular biosynthetic process			2.5e-14	283	0.04	231
4	GO:0045449	P	regulation of transcription			2.5e-14	283	---	---
5	GO:0010468	P	regulation of gene expression			2.5e-14	283	---	---
6	GO:0019219	P	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process			2.5e-14	283	0.04	231
7	GO:0051171	P	regulation of nitrogen compound metabolic process			2.5e-14	283	0.04	231
8	GO:0080090	P	regulation of primary metabolic process			5.1e-14	290	0.0031	252
9	GO:0031323	P	regulation of cellular metabolic process			7.4e-14	283	---	---
10	GO:0060255	P	regulation of macromolecule metabolic process			9.3e-14	290	0.0034	253
11	GO:0019222	P	regulation of metabolic process			2.7e-13	290	0.0041	254
12	GO:0006350	P	transcription			6.9e-13	287	---	---
13	GO:0050794	P	regulation of cellular process			1.9e-10	321	0.037	287
14	GO:0050789	P	regulation of biological process			5.9e-10	329	0.0039	309
15	GO:0065007	P	biological regulation			9.9e-09	433	---	---
16	GO:0006355	P	regulation of transcription, DNA-dependent			1.9e-08	180	0.04	158
17	GO:0051252	P	regulation of RNA metabolic process			2e-08	180	0.04	158
18	GO:0006351	P	transcription, DNA-dependent			1.2e-07	182	0.035	165
19	GO:0032774	P	RNA biosynthetic process			1.2e-07	182	0.036	165
20	GO:0034645	P	cellular macromolecule biosynthetic process			1.3e-06	360	0.00013	380
21	GO:0009059	P	macromolecule biosynthetic process			1.4e-06	360	0.00013	380
22	GO:0044249	P	cellular biosynthetic process			1.9e-06	441	0.0015	456
23	GO:0009058	P	biosynthetic process			2.1e-06	462	0.0045	471
24	GO:0010467	P	gene expression			6e-05	341	0.00012	378
25	GO:0006139	P	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process			9.8e-05	334	0.036	339
26	GO:0006807	P	nitrogen compound metabolic process			0.00025	379	---	---
27	GO:0009628	P	response to abiotic stimulus			0.0015	137	---	---
28	GO:0042592	P	homeostatic process			0.0016	149	---	---
29	GO:0009266	P	response to temperature stimulus			0.002	130	---	---
30	GO:0009409	P	response to cold			0.0024	129	---	---
31	GO:0048871	P	multicellular organismal homeostasis			0.0024	129	---	---
32	GO:0050826	P	response to freezing			0.0024	129	---	---
33	GO:0042309	P	homoiothermy			0.0024	129	---	---
34	GO:0016070	P	RNA metabolic process			0.0024	190	0.015	202
35	GO:0001659	P	temperature homeostasis			0.0024	129	---	---
36	GO:0065008	P	regulation of biological quality			0.003	152	---	---
37	GO:0044260	P	cellular macromolecule metabolic process			0.022	569	0.00044	656
38	GO:0032501	P	multicellular organismal process			0.046	134	---	---
39	GO:0044237	P	cellular metabolic process			0.046	705	0.01	791
40	GO:0030528	F	transcription regulator activity			9e-07	179	---	---
41	GO:0003677	F	DNA binding			6.8e-05	292	0.00072	310
42	GO:0003700	F	transcription factor activity			7.7e-05	120	---	---
43	GO:0050825	F	ice binding			0.011	129	---	---
44	GO:0050824	F	water binding			0.011	129	---	---
45	GO:0043565	F	sequence-specific DNA binding			0.015	79	---	---
46	GO:0060089	F	molecular transducer activity			0.015	92	---	---
47	GO:0004871	F	signal transducer activity			0.015	92	---	---
48	GO:0004879	F	ligand-dependent nuclear receptor activity			0.015	41	---	---
49	GO:0008270	F	zinc ion binding			0.017	197	---	---
50	GO:0004872	F	receptor activity			0.026	70	---	---
51	GO:0005634	C	nucleus			3.5e-05	232	1.1e-05	249
52	GO:0043231	C	intracellular membrane-bounded organelle			0.0014	264	4.6e-06	301

53	GO:0043227	C	membrane-bounded organelle	0.0014	264	3.1e-06	304
54	GO:0043229	C	intracellular organelle	0.0081	347	1e-11	451
55	GO:0043226	C	organelle	0.0081	347	1e-11	451
56	GO:0006412	P	translation	---	---	6.8e-05	119
57	GO:0046907	P	intracellular transport	---	---	0.00012	57
58	GO:0006886	P	intracellular protein transport	---	---	0.00013	49
59	GO:0034613	P	cellular protein localization	---	---	0.00013	49
60	GO:0070727	P	cellular macromolecule localization	---	---	0.00013	49
61	GO:0051641	P	cellular localization	---	---	0.00013	67
62	GO:0006996	P	organelle organization	---	---	0.00013	55
63	GO:0051649	P	establishment of localization in cell	---	---	0.00014	65
64	GO:0009987	P	cellular process	---	---	0.00022	990
65	GO:0051246	P	regulation of protein metabolic process	---	---	0.00044	21
66	GO:0044267	P	cellular protein metabolic process	---	---	0.00046	354
67	GO:0006325	P	chromatin organization	---	---	0.00076	36
68	GO:0008104	P	protein localization	---	---	0.0017	64
69	GO:0051276	P	chromosome organization	---	---	0.0018	36
70	GO:0045184	P	establishment of protein localization	---	---	0.0019	62
71	GO:0015031	P	protein transport	---	---	0.0019	62
72	GO:0043170	P	macromolecule metabolic process	---	---	0.0045	710
73	GO:0006333	P	chromatin assembly or disassembly	---	---	0.0092	29
74	GO:0019538	P	protein metabolic process	---	---	0.011	403
75	GO:0016043	P	cellular component organization	---	---	0.016	78
76	GO:0033036	P	macromolecule localization	---	---	0.021	67
77	GO:0033365	P	protein localization in organelle	---	---	0.029	12
78	GO:0034728	P	nucleosome organization	---	---	0.04	25
79	GO:0006270	P	DNA replication initiation	---	---	0.04	8
80	GO:0006275	P	regulation of DNA replication	---	---	0.04	5
81	GO:0065004	P	protein-DNA complex assembly	---	---	0.04	25
82	GO:0031497	P	chromatin assembly	---	---	0.04	25
83	GO:0017038	P	protein import	---	---	0.04	8
84	GO:0006334	P	nucleosome assembly	---	---	0.04	25
85	GO:0034622	P	cellular macromolecular complex assembly	---	---	0.04	36
86	GO:0034621	P	cellular macromolecular complex subunit organization	---	---	0.043	38
87	GO:0006323	P	DNA packaging	---	---	0.045	25
88	GO:0003735	F	structural constituent of ribosome	---	---	9e-08	103
89	GO:0005198	F	structural molecule activity	---	---	9e-08	143
90	GO:0003676	F	nucleic acid binding	---	---	0.00072	423
91	GO:0004365	F	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity	---	---	0.031	5
92	GO:0008943	F	glyceraldehyde-3-phosphate dehydrogenase activity	---	---	0.031	5
93	GO:0042803	F	protein homodimerization activity	---	---	0.031	7
94	GO:0019787	F	small conjugating protein ligase activity	---	---	0.031	31
95	GO:0016881	F	acid-amino acid ligase activity	---	---	0.049	33
96	GO:0032991	C	macromolecular complex	---	---	1e-11	260
97	GO:0044424	C	intracellular part	---	---	3e-11	540
98	GO:0005622	C	intracellular	---	---	4.9e-11	631
99	GO:0043232	C	intracellular non-membrane-bounded organelle	---	---	3.3e-10	183
100	GO:0043228	C	non-membrane-bounded organelle	---	---	3.3e-10	183
101	GO:0030529	C	ribonucleoprotein complex	---	---	1.8e-08	111
102	GO:0005840	C	ribosome	---	---	1.8e-08	103
103	GO:0044444	C	cytoplasmic part	---	---	7.6e-07	175
104	GO:0044422	C	organelle part	---	---	1e-06	128
105	GO:0044446	C	intracellular organelle part	---	---	1e-06	128
106	GO:0044464	C	cell part	---	---	1e-06	872
107	GO:0005623	C	cell	---	---	1e-06	872
108	GO:0005737	C	cytoplasm	---	---	1.7e-05	211
109	GO:0031967	C	organelle envelope	---	---	0.00053	34
110	GO:0043234	C	protein complex	---	---	0.00071	127
111	GO:0031975	C	envelope	---	---	0.00081	37
112	GO:0031968	C	organelle outer membrane	---	---	0.0032	9
113	GO:0015934	C	large ribosomal subunit	---	---	0.0065	9
114	GO:0019867	C	outer membrane	---	---	0.0067	10
115	GO:0033279	C	ribosomal subunit	---	---	0.0091	15
116	GO:0031090	C	organelle membrane	---	---	0.012	37

117	GO:0031966	C	mitochondrial membrane			---	---	0.012	24
118	GO:0044425	C	membrane part			---	---	0.013	185
119	GO:0009538	C	photosystem I reaction center			---	---	0.015	5
120	GO:0000502	C	proteasome complex			---	---	0.022	11
121	GO:0005740	C	mitochondrial envelope			---	---	0.023	26
122	GO:0000786	C	nucleosome			---	---	0.024	24
123	GO:0000785	C	chromatin			---	---	0.026	31
124	GO:0005741	C	mitochondrial outer membrane			---	---	0.027	7
125	GO:0032993	C	protein-DNA complex			---	---	0.027	24
126	GO:0044428	C	nuclear part			---	---	0.028	21

Transposable Elements near genes

Material and Methods

Getting TE information

The information about repeats and TEs in maize genome version AGPv2 [219] was obtained in 'gff' format from `http://ftp.maizesequence.org/release-5b/repeats/`. The 'gff' files were converted to a 'BED' format file. Entries with completely overlapping coordinates were merged with bedtools and entries with length <20bp were removed. For sorghum TE information was obtained from 'PGSB PlantsDB' website [172] by the URL `ftp://ftpmips.helmholtz-muenchen.de/plants/sorghum/repeats/MIPS_Sb_repeat_annotation_July08.gff3.gz`. The 'gff' files were converted to a 'BED' format file. Entries with completely overlapping coordinates were merged with bedtools and entries with length <20bp were removed in this case as well.

Selection of Genes

The information about maize genes was obtained in 'gff' format from `ftp.maizesequence.org/release-5b/ZmB73_5b_FGS_info.txt`. Two datasets of genes are available, a 'working gene set (WGS)' which comprises of all gene predictions including computationally predicted genes and its subset called 'filtered gene set (FGS)' which excludes low confidence predictions [219]. Only genes in the FGS were chosen for analysis. Out of several transcripts for a gene, the one marked as 'canonical' was chosen which was also found typically to be the largest for maize. Additionally, genes annotated as 'transposable elements' and without cDNA evidence were excluded. The information about sorghum genes was obtained by installing 'integrated genome browser' [171] and downloading the BED file for sorghum genes from the installation. For sorghum the longest protein coding transcript was chosen. Genes classified as 'transposable element' and genes falling in plastid and mitochondria were removed.

BED files with interval of 1kb upstream and downstream from the annotated TSS and TES of the chosen genes were prepared. Entries in these BED files which overlapped with any other gene in the WGS were removed. The rationale for this was to exclude any possible genic overlap, even with a low confidence gene. After this 25239 genes remained for upstream and 28239 genes for downstream. This list of genes for maize also formed the background (All genes) for all gene ontology (GO) enrichment analysis in this work. The list of syntenic orthologs for maize and sorghum were obtained from James Schanable's website `www.skraelingmountain.com/datasets.php` [216]. TE coverage (the number of basepairs in 1KB region annotated as a TE) was calculated by using bedtools "coverage" command [190].

Coordinate Conversion

All analysis in the study was done on Maize genome version 2 (AGPv2). The coordinates of data when available in version 3 were ported to AGPv2 using CrossMap [290]. The files needed for Crossmap to work for maize were downloaded from the `ftp://ftp.ensemblgenomes.org/pub/plants/release-30/assembly_chain/zea_mays/`.

Obtaining Ka and Ks

Ka and Ks values were obtained from the Ensemble Biomart website for maize-sorghum orthologs (`plants.ensembl.org/biomart/martview/`). A static file is available at `ftp.ensemblgenomes.org/pub/plants/release-27/mysql/plants_mart_27/zmays_eg_gene_homolog_sbicolor_eg_dm.txt.gz`. Ka and Ks values < 0 and > 0.5 and Ka/Ks > 1 were excluded from analysis.

TSS type for maize genes

The TSS (transcription start site) type classification data (sharp v.s broad) were obtained by a personal request from Mejia-guerra et.al [158].

Expression data

Microarray based expression data for maize was obtained by study from Sekhon et.al [221] via PlexDb [42] (experiment identifier ZM37) which also ported the data to AGPv2. Tissue specificity index (tau) was calculated from this data with a custom script using method described here [283]. RNA-Seq based expression data was obtained from the qteller website `qteller.com`.

Gene ontology analysis

Gene ontology analysis was done using AgriGo tool [51]. "Single Enrichment Analysis" option of AgriGo was used with the background "All genes" (see 'selection of genes' above). A false Discovery Rate of < 0.05 was used for a given enrichment to be significant. Transcription factor genes in maize were downloaded from PlantTFDB [112].

Results

TE abundance patterns in upstream and downstream of genes

The genes in the maize 'filtered gene set' which had no overlap with any other gene in the 1KB upstream and downstream region were chosen for analysis (See methods). This was done as any overlap with another gene might leave fewer room for TEs to jump in and the TE distribution around genes will get biased for fewer TEs. 28239 genes had a 'clear' 1KB upstream region with no other gene overlapping compared to 25888 genes downstream. Fewer number of gene to gene overlaps in upstream regions could be due to presence of regulatory elements in upstream of genes. TE coverage was defined as the fraction of 1Kb region covered by TEs. 1KB distance was chosen because a study in maize showed the influence of TEs on nearby gene expression diminishes strongly beyond 1kB [151]. The mean(median) coverage of TEs in upstream 1Kb of a gene was 0.350(0.297). The coverage for downstream 1Kb was found to be 0.396(0.339). This difference was found to be significant (P $<$ 2E-16; Wilcoxon rank sum test), thus on an average there are $\sim$45 more TE annotated base pairs in downstream 1Kb as compared to upstream. Both these are very low compared to genome wide mean TE coverage of $\sim$0.850. The histogram plot showing distribution of upstream and downstream coverage shows majority of genes are in low coverage ranges (Figure 4.1). The upstream and downstream TE coverages were weekly correlated (Spearman's rho 0.14 P$<$2E-16) (Figure 4.2).

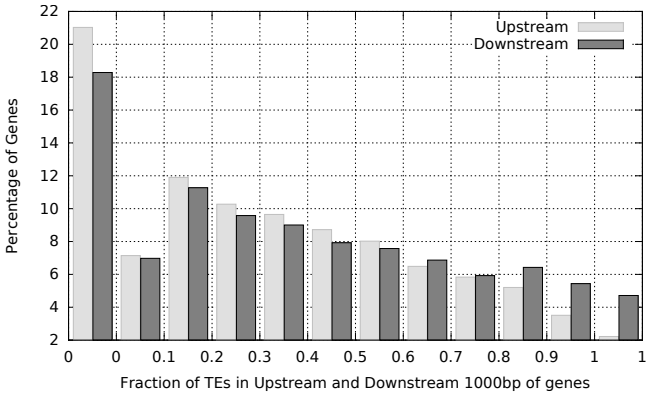


Figure 4.1: Histogram of percentage of genes in given upstream 1KB TE coverage range

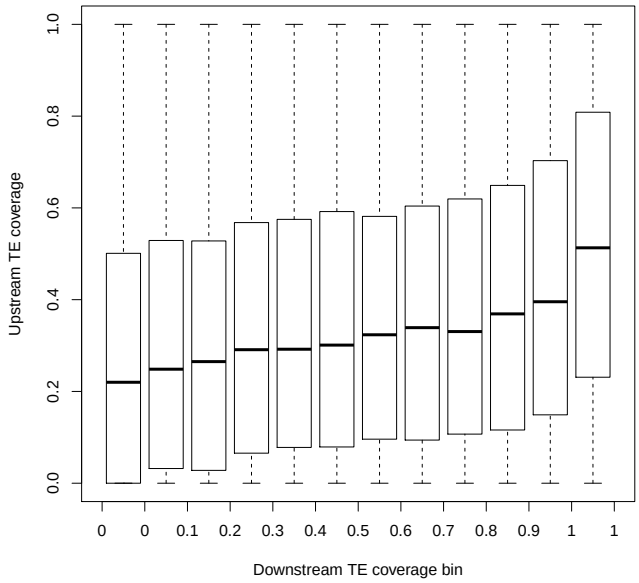


Figure 4.2: Boxplot of upstream TE coverage in corresponding downstream TE coverage bin. Correlation was calculated using non binned data.

TE coverage with distance from genes

Upstream TEs are expected to be rare in the immediate proximity of the genes due to presence of core promoter elements which are typically ∼100bp flanking the transcription start site (TSS) and function as transcription initiation elements [158, 115]. As coverage gives an average value of TE basepairs over 1KB and does not factor in completely the distance of TE from the gene start, the percentage of genes with TEs was plotted in relation to increasing upstream and downstream distance from the gene (Figure 4.3). For upstream the annotated 'transcription start site' (TSS) was used as a starting point and for downstream 'transcription end site' (TES) was used. TE abundance increases with increasing upstream and downstream distance from the TSS/TES (Figure 4.3). Upstream regions show a consistent depletion of TEs compared to the downstream but the difference decreases with increasing distance (Figure 4.3). To check distance from the gene at which the TE coverage reaches the genomic background (∼80-85% sites have a TE) and the upstream and downstream coverage are similar, a 5Kb upstream and downstream distance was used (Figure 4.4). It takes approximately ∼5kb to reach the genomewide TE coverage of ∼80-85% (Figure 4.4).

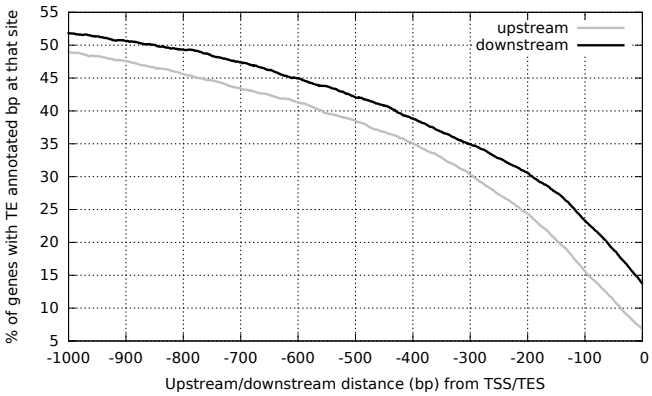


Figure 4.3: Percentage of genes with Upstream/Downstream (Grey/Black) TE basepair shown in relation to increasing upstream(-)/downstream(+) distance from the gene. Upstream distance was calculated from the TSS (transcription start site) and downstream from the TES (transcription end site).

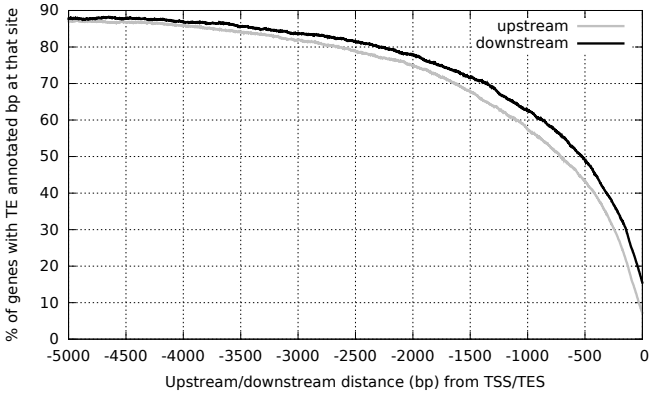


Figure 4.4: Percentage of genes with Upstream/Downstream (Grey/Black) TE basepair shown in relation to increasing upstream(-)/downstream(+) distance from the gene. A distance of 5kb upstream and downstream from the gene is shown.

TE coverage and purifying selection on genes

The median(mean) Ka/Ks for genes in first quartile (Q1) of upstream TE coverage was 0.192(0.233) compared to the fourth quartile (Q4) which was 0.297(0.260) (P=4E-11;Wilcoxon rank sum test). The values for downstream coverage were (Q1) 0.192(0.241) vs (Q4) 0.206(0.252) (P=0.006, Wilcoxon rank sum test). Purifying selection as measured by Ka/Ks was thus slightly weaker at higher ranges of upstream coverage. However the evidence for a consistent decrease in Ka/Ks with increased TE coverage was found to be weak for upstream regions (Figure 4.5) and non-existent for downstream regions.

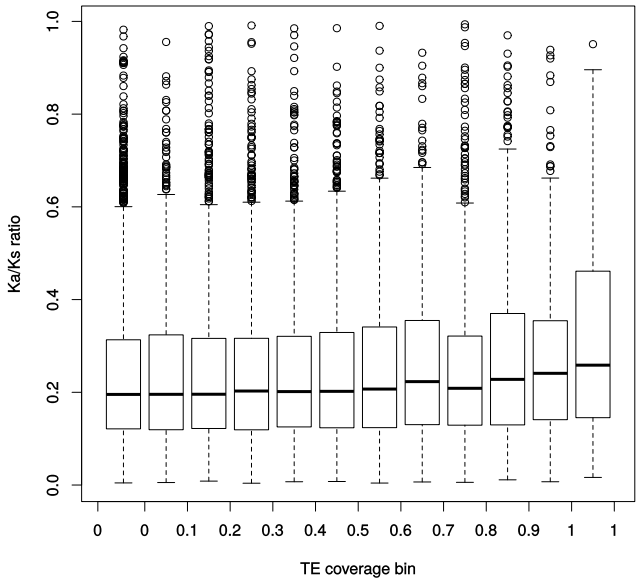


Figure 4.5: Boxplots of Ka/Ks ratio in different TE coverage bins

TE coverage and gene expression

Boxplots of gene expression (measured as FPKM) were made for each upstream and downstream TE coverage bin for six tissues namely Bundle Sheath, Leaf, Pollen, Silk, Tassel and Ear, represented in Figures 4.6, 4.7, 4.8, 4.9, 4.10 and 4.11 respectively. Please note that only genes with FPKM$>=$1 were included in this analysis. Although no strong effect of expression on TE coverage could be seen, at low TE coverage, gene expression was found to increase with TE coverage in some of the tissues analyzed (for silk, tassel and ears in case of upstream TE coverage and for Bundle Sheath, Mature Leaf, Silk, Tassel and Ears for downstream TE coverage). The patterns tend to reverse at high TE coverages ($>$0.6) where expression showed a decrease with increasing TE coverage for upstream and downstream. Surprisingly this non-monotonic relation seemed to be stronger for downstream TE coverage ranges.

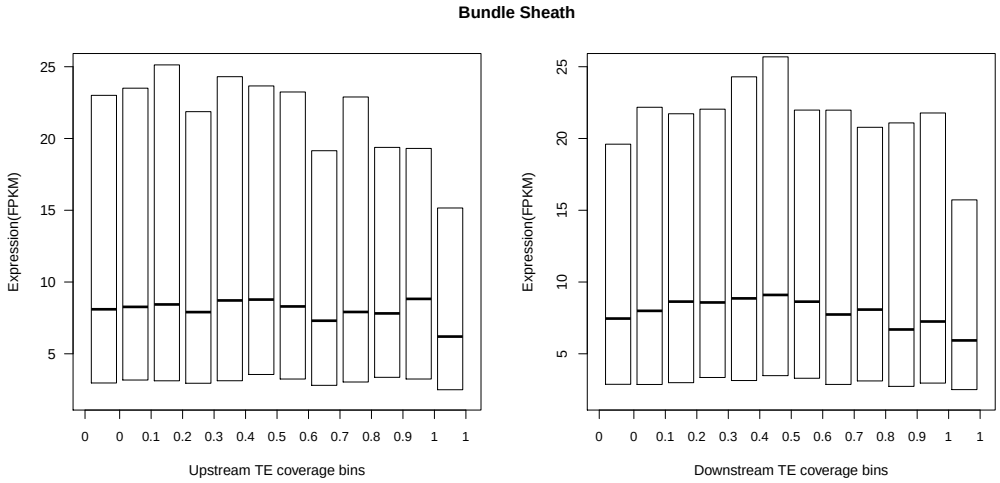


Figure 4.6: Boxplots for expression values for genes binned by TE coverage for tissue bundle sheath

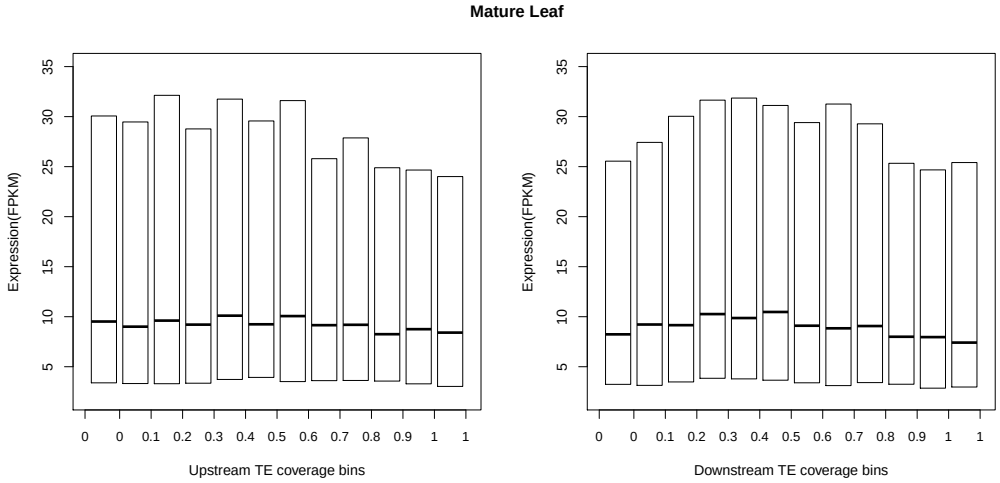


Figure 4.7: Boxplots for expression values for genes binned by TE coverage for tissue mature leaf.

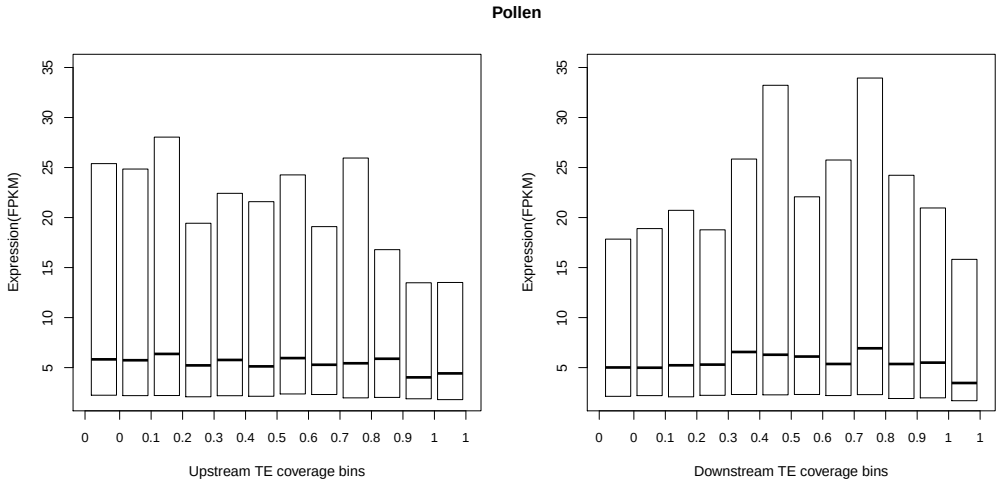


Figure 4.8: Boxplots for expression values for genes binned by TE coverage for tissue Pollen.

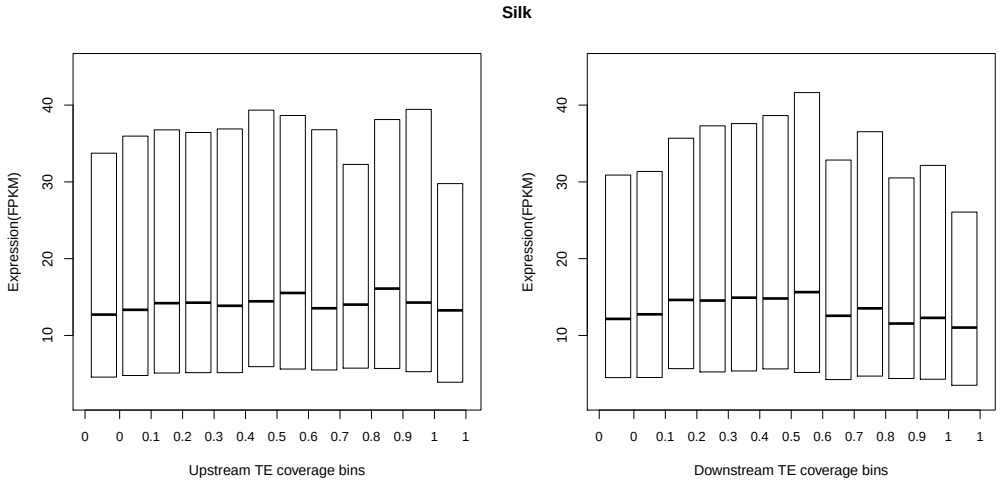


Figure 4.9: Boxplots for expression values for genes binned by TE coverage for tissue Silk.

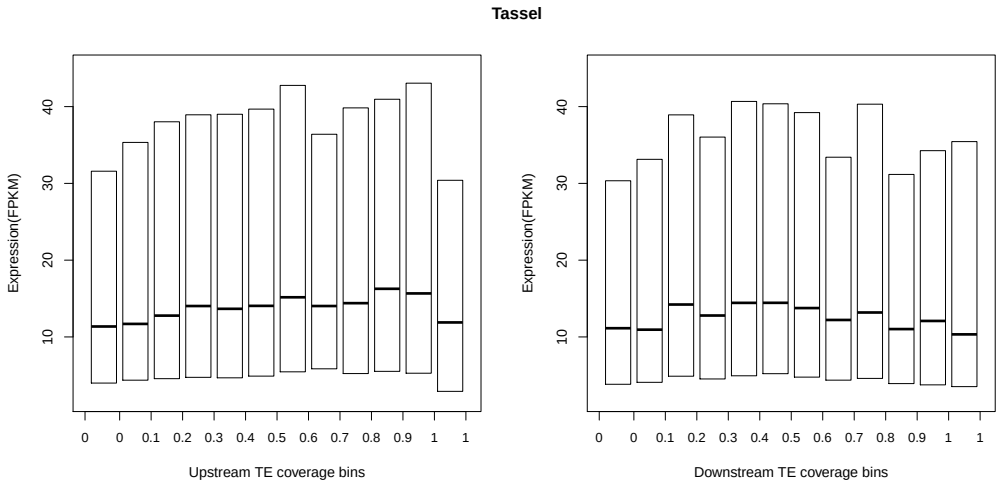


Figure 4.10: Boxplots for expression values for genes binned by TE coverage for tissue Tassel.

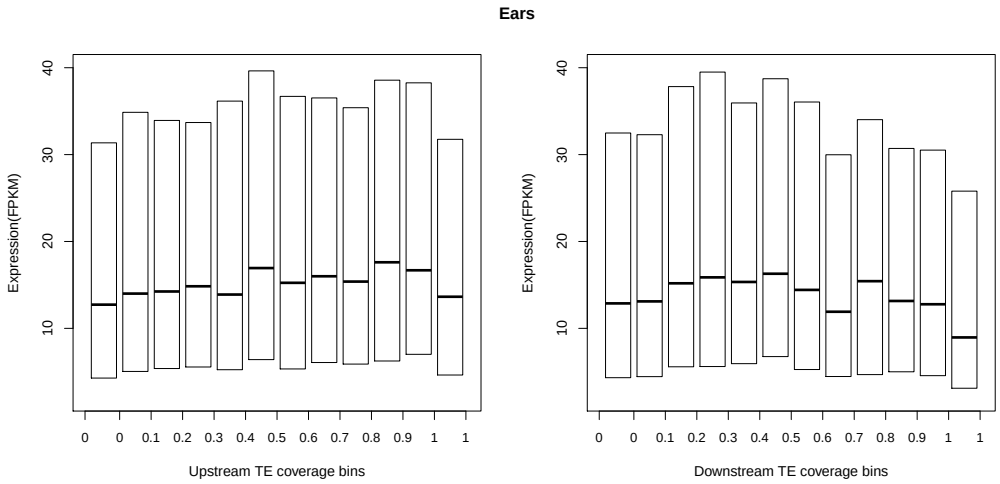


Figure 4.11: Boxplots for expression values for genes binned by TE coverage for tissue Ear.

Gene Ontology (GO) Enrichment

Biological regulation was significantly enriched in genes (FDR<0.05) in low TE coverage category (upstream TE coverage <0.1 and downstream TE coverage <0.1) for both upstream and downstream TE coverages (Figure 4.12 and 4.13), though downstream regions showed a much weaker enrichment. Since gene ontology gives a broader view of the function, TE coverage of transcription factors (TFs) was specifically compared against background group of genes (see methods). Transcription factor genes were chosen because they are strongly associated with regulation and are comprehensively annotated in maize [112]. TF genes had significantly lower (P<2E-16;Wilcoxon rank sum test) TE coverage for both upstream and downstream compared to all genes with mean(median) of 0.236(0.152) and 0.273(0.179) for upstream and downstream (Figure 4.14). The percentage of genes with a TE annotated basepair at each site with increasing outward distance from the gene measured from TSS/TES for both upstream and downstream regions is shown in Figure 4.15 where two classes of genes are compared namely TFs and all genes. TFs have fewer TEs even at larger distances from the gene (Figure 4.15).

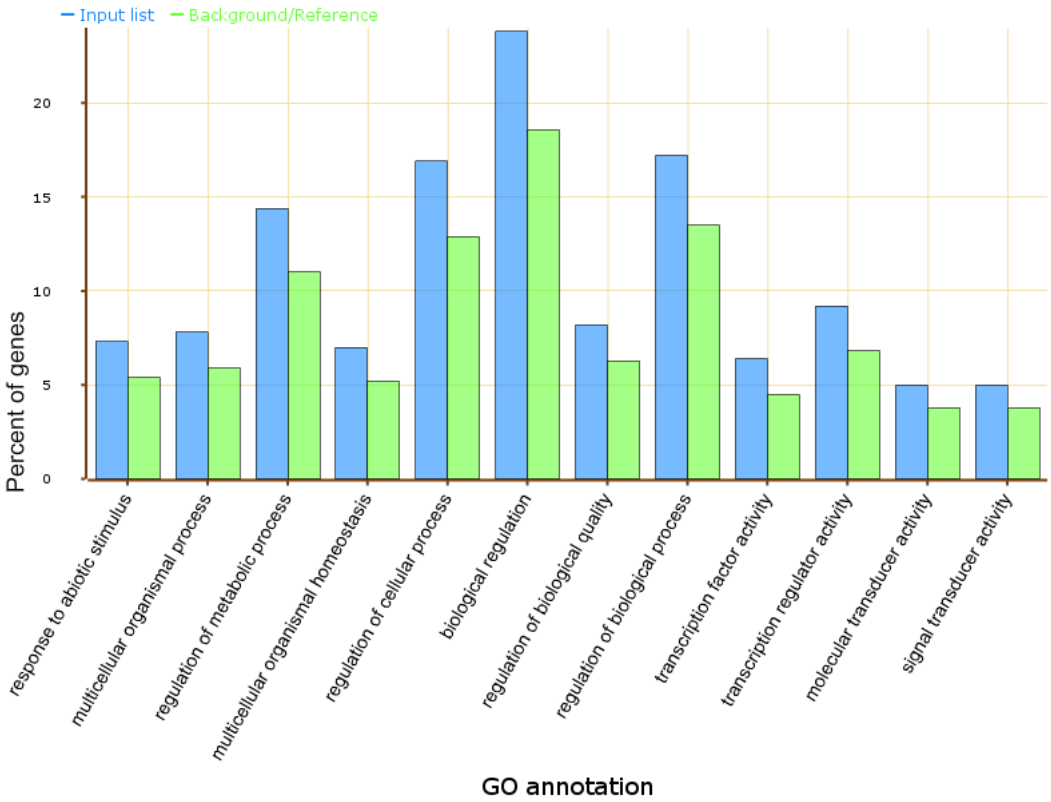


Figure 4.12: GO categories displaying significant (FDR<0.05) enrichment for genes with low upstream TE coverage. The percentage of genes in the input (blue) and (background) is given in y-axis with x-axis giving the name of the GO category.

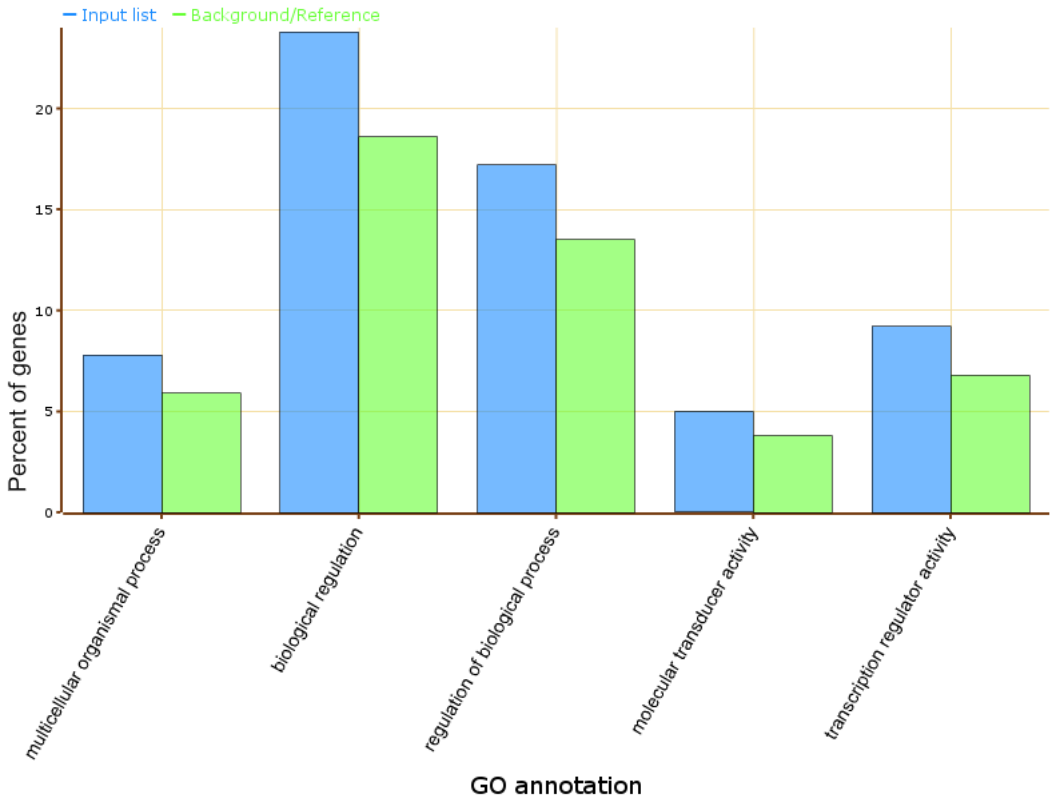
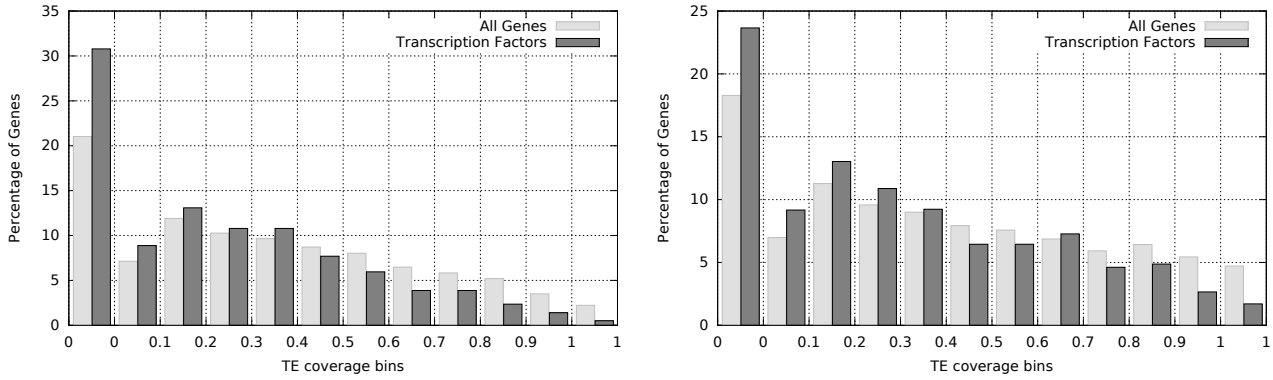


Figure 4.13: GO categories displaying significant (FDR<0.05) enrichment for genes with low downstream TE coverage. The percentage of genes in the input (blue) and (background) is given in y-axis with x-axis giving the name of the GO category.



(a) Upstream

(b) Downstream

Figure 4.14: Fraction of genes in each TE coverage bin is shown for two sets of genes namely transcription factors (dark Grey) and all genes (light Grey)(see methods). (a) Upstream TE coverage (b) Downstream TE coverage.

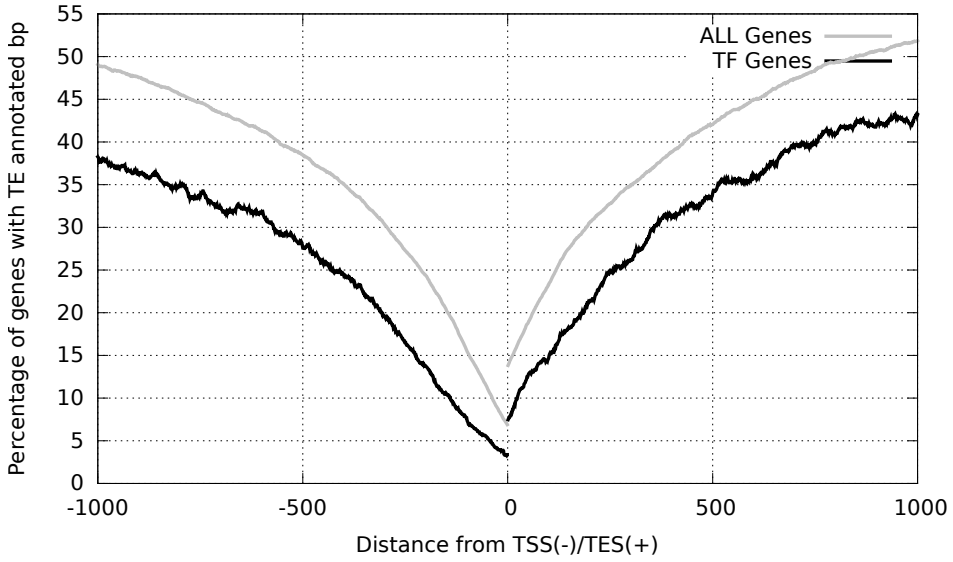
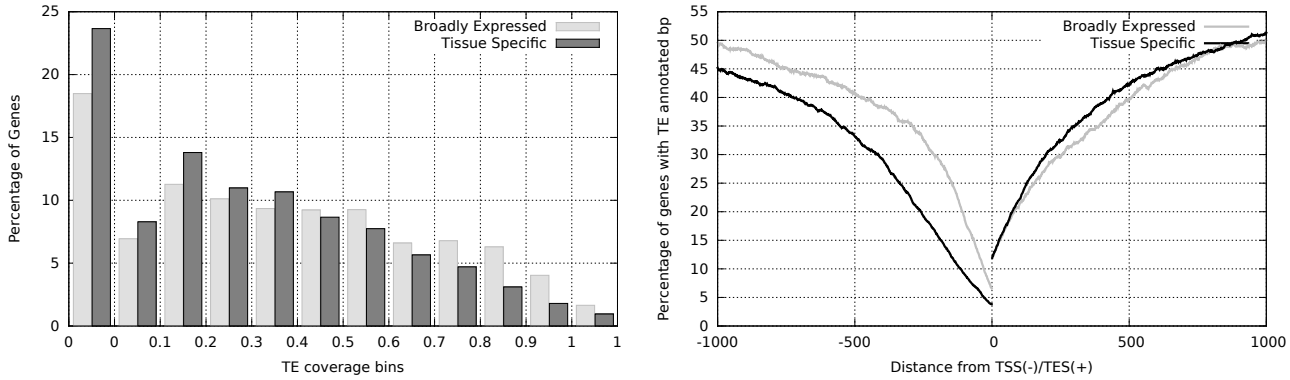


Figure 4.15: Percentage of genes which contain a TE annotated basepair at a given distance from the gene. The distances are in basepairs and were measured from TSS/TES of genes for Upstream/Downstream (negative/positive) regions. Distances are shown for two classes of genes namely transcriptions factors (TF genes) and all genes.

TE coverage and expression breadth

The TE coverage of Tissue-specific (TS) genes was compared against Broadly Expressed (BE) genes. Genes were classified as tissue-specific (TS) or broadly expressed (BE) based on the tissue specificity index (tau) (See methods). The mean(median) upstream TE coverage for TS genes was 0.293(0.235) vs 0.373(0.328) for BE genes (P<2E-16; Wilcoxon rank sum test). The downstream coverage however displayed a reversed pattern 0.393(0.345) vs 0.374(0.320) for TS and HS respectively although at a much lower significance (P=0.01; Wilcoxon rank sum test) (Figure 4.16).



(a) Upstream TE coverage bins

(b) Upstream and Downstream

Figure 4.16: (a) Percentage of genes in each upstream TE coverage bin for Broadly expressed (BE) vs Tissue specific (TS) genes. The difference between two categories significant (P<2E-16; Wilcoxon rank sum test) for both plots. (b) Percentage of genes with a TE basepair at a given upstream/downstream (-/+) distance from gene start/end for two categories.

TEs and TSS Type

Genes often have multiple Transcription Start Sites (TSS) which constitute a TSS cluster and each TSS is linked to a separate promoter [158]. TSS are usually determined by CAGE(Cap Analysis of Gene Expression) studies [222] which take advantage of the fact that 5' end of a mature mRNA has a chemical cap. Special library is made from transcript fragments which contain the chemical cap (called as tags). The tags are then mapped to the genome and the start of the tag in the alignment with respect to the gene marks the TSS. The TSS clusters are categorized into broad and sharp based on the density of TSS sites in a cluster [158]. Sharp TSS genes were found to have lower TE coverage in upstream regions with mean(median) of 0.296(0.237) compared to 0.329(0.281) for broad genes (P<1E-5; Wilcoxon rank sum test). However, TE coverage difference between downstream regions was not found to be significant between sharp and broad genes with mean(median) of 0.291(0.206) and 0.285(0.213)(P=0.9; Wilcoxon rank sum test). The percentage of genes with TE basepair plotted with respect to increasing upstream/downstream distance for broad and sharp TSS genes is given in Figure 4.17.

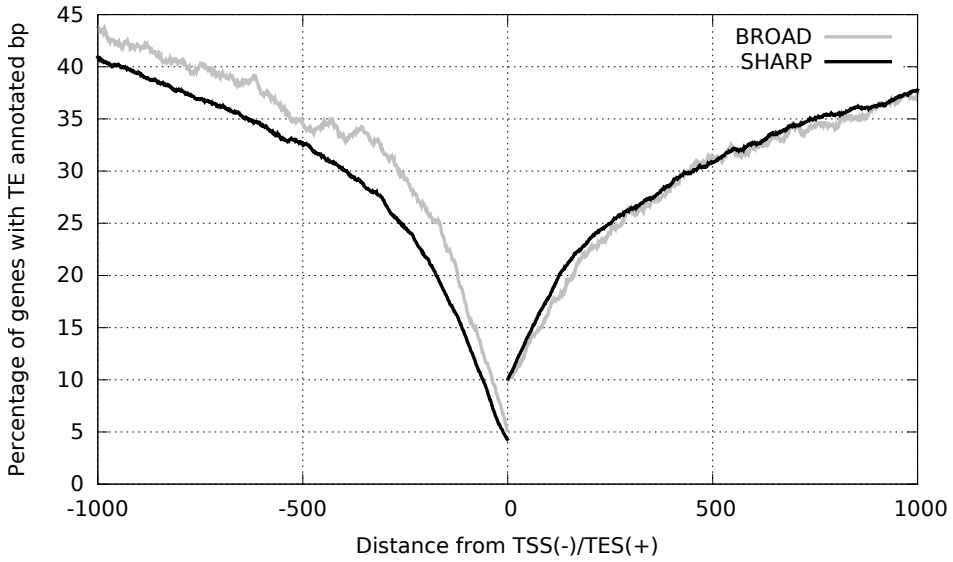


Figure 4.17: Percentage of genes with a TE basepair at a given upstream/downstream (-/+) distance from gene start/end for two categories (broad (Grey) and sharp (black)).

Comparing TEs in Maize and Sorghum

TE content of sorghum genome is fewer than that of maize, even though the split is relatively recent (∼11MYA) [241]. To see if genome wide TE content influences the cis-TE content for genes, upstream TE coverage was compared between maize and sorghum ortholog pairs. Since most transcripts in sorghum do not have an annotated TSS (Transcription start site) assigned, TE coverage was calculated for 1KB upstream and downstream regions from the CDSS (CDS start site). CDSS was chosen as it can be relatively reliably assigned in genome annotation. TE coverage for maize was slightly but significantly higher than sorghum with mean(median) of 0.213(0.139) vs 0.193(0.128) (P<2E-16; Wilcoxon rank sum test). TE coverage between maize and sorghum in upstream regions

was also correlated (Spearmans rho 0.18 ;P<2E-16).

In order to check for differences between maize and sorghum for TE coverage with increasing distance from the gene, the upstream TE coverage for 1,2,3 and 5KB for a gene in maize was subtracted from the corresponding TE coverage of syntenic ortholog in sorghum. A histogram plot for the difference is depicted in Figure 4.18, The x-axis gives the difference in the number of TE basepairs between maize and sorghum. Negative values mean that maize has less TE basepairs than sorghum and positive values the contrary. The distribution is centered on zero for 1KB TE coverage, indicating that a majority of genes have not experienced drastic differences in TE content between maize and sorghum in the upstream 1KB region. But with increasing distance from the gene a maize gene has more likely to be near a TE. No significant gene ontology enrichment were found for genes in extreme ends of the distributions.

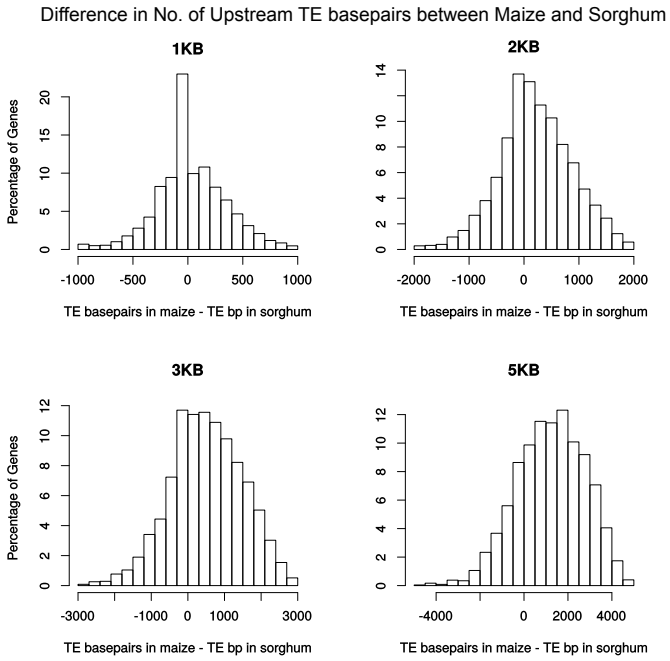


Figure 4.18: Histogram plot for the difference in number of TE basepairs between maize and sorghum. Negative values imply that maize has less TE basepairs than sorghum and positive values the contrary.

Discussion

Maize is particularly suited for studying factors shaping the TE abundance proximal to genes due to ample available genomic data and a very high TE content. Here different factors were tested for their role in shaping this landscape.

For studying upstream/downstream TEs unhindered by the overlaps between genes, only genes with other gene overlapping in 1Kb upstream and downstream region were chosen. Fewer gene to gene overlaps in upstream of genes (28239 genes had no other gene overlapping in upstream 1KB compared to 25888 genes for downstream) compared to downstream indicates that the cis-regulatory upstream DNA has a role in shaping the distribution of distance between genes. A study reported results along these lines in *Caenorhabditis elegans* and *Drosophila melanogaster* wherein intergenic distance was shown to be non-random and positively correlated with the regulatory complexity of the gene [169]. They proposed that this influence is much stronger for compact genomes with high DNA deletion rates. The genes in these genomes get closer to one another by erosion of non-functional non-coding DNA. The functional noncoding DNA resists deletion and then becomes a major factor shaping the boundary between genes. The genes with more functional non-coding DNA then show longer intergenic distance with other genes. Although maize genome is far less compact but genic regions form islands in a vast sea of TEs and hetrochromatin [139, 201], so a similar effect is thereby expected. The TEs will jump into non-functional non-noncoding DNA and the functional non-coding DNA will then mark the boundary between TEs and genes. Gene density has also been shown to be negatively correlated with TE abundance in *Arabidopsis* [277] indicating that purifying selection against TE insertion in functional DNA shapes the TE abundance in and near genes.

We report a lower TE coverage for 1KB upstream genic regions than downstream (mean 350bp v.s 396 TE basepairs in 1KB upstream/downstream) (Figure 4.1). This is expected as the canonical biological knowledge states the regulatory DNA to be upstream of a gene which would lead to fewer TE insertions in that region. Upstream and downstream TE coverage was also found to be correlated (Figure 4.2). A region specific rate of TE insertion could explain this. For example, a high recombination rate in a region would increase the efficiency of selection against TE insertions. A study in soybean found fewer LTR insertions in regions with higher recombination rate [252], but this was not found to be the case in arabidopsis [277]. Since upstream/downstream 1KB TE coverage only gives an average TE content and not the variation in TE content with the distance from the gene, TE coverage was plotted with increasing outward distance from upstream/downstream of genes (Figure 4.3). Very few genes ($<$5%) have TEs in the immediate vicinity in upstream regions and the percentage of TE associated basepairs increases with distance from genes, for both upstream and downstream regions (Figure 4.3). As the core promoter usually flanks the TSS (transcription start site) where a TE insertion could have more deleterious effect thereby fewer TE insertions are seen in immediate vicinity of the gene. Also the difference between upstream and downstream regions gets progressively smaller with distance but it takes $\sim$5KB to reach the genomic background of $\sim$85% TE coverage. One reason for this could point to regulatory elements located far away from the genes where a TE insertion would be deleterious. An extreme example in maize is of the *Vgt1* locus associated with the

flowering time which is 70kb upstream from the effected gene [209]. But for downstream regions the reasons are not clear.

The relation between strength of purifying selection on a gene (measured by Ka/Ks ratio) and upstream TE coverage was found to be weakly negative (Figure 4.5). Although a study in maize linked reduced purifying selection in one of the post WGD duplicate copy with higher upstream TE abundance but on a shorter evolutionary timescale [186]. A study in humans reported a negative relation between rate of evolution (Ka/Ks ratio) and TE insertions inside genes, but only for tissue specific genes [113]. However, the effect seen in maize is weak and upstream TEs do not seem to be a major determinant of rate of protein evolution in maize.

The relation between TE coverage (upstream and downstream) and absolute gene expression (FPKM) was weak, although in some tissues a non-monotonic relation was seen wherein expression is low initially at zero TE coverage, then increases and then starts to decrease at higher ranges of coverage and finally at very high TE coverage is the lowest (Figure 4.6, 4.7, 4.8, 4.9 and 4.10). A universal non-monotonic relation between gene expression and compactness (measured by gene length) has been reported [24] where gene length first increases with gene expression, peaks and then decreases for highly expressed genes. The initial increase in length with expression was proposed to be associated with gain of regulatory elements to increase expression, and the selection for efficient expression was then proposed for smaller lengths of highly expressed genes [24]. One reason for expression not being strongly dependent on TE coverage could be the lack of methylation status for the TEs. Methylation of TEs could negatively affect the nearby gene expression (see introduction) and Hollister and Gaut showed that gene expression in arabidopsis was negatively correlated with methylated upstream TEs but not with the unmethylated ones [98]. Chromatin state and methylation status are crucial variables to explore further. Chromatin states tend to be different for actively transcribed genes and TEs [213, 96]. The open chromatin associated with active transcription is a preferred location for some class of TEs to jump [96]. This could generate a positive correlation with increasing neighboring TEs with gene expression. At very high TE coverages however, TEs are more likely to influence the core promoter regions and TE suppression by methylation would reduce the proximal gene expression. The non-monotonic nature of this relation would weaken the Ka/Ks v.s TE coverage association, as the rate of protein evolution in plants (Ka/Ks) is negatively correlated with absolute expression [197, 284].

The GO analysis pointed a strong tendency of genes involved in regulation to be depleted in TEs for both upstream and downstream regions. This observation was reported in mammals for full length TEs [162]. The transposon free regions which maintained this status in vertebrate evolution were also associated with regulation [224]. Our study shown that the negative relation between TEs and regulatory genes is also true in maize, and one reason explaining the observation is higher content of functional cis-regulatory elements reflected by higher conservation of upstream regions of genes with complex regulation [135]. This would make TE insertions in these regions deleterious and thus selected against. The extension of the analysis to transcription factors (TFs) showed a strong depletion of TEs in upstream and downstream regions of TFs supporting the conclusion as TFs have been shown to display higher conservation of upstream regions [110]. Note that the TEs inside genes are also not

favored for regulators and transcription factors in mammals [225] and this observation can easily be tested in maize.

The results from the GO analysis prompted analysis of TE coverage in relation to expression breadth; tissue specificity (TS) v.s broader expression (BE). This is because TS genes display a higher cis-conservation compared to housekeeping genes in mammals [60]. This may come as surprising at first, but in order to coordinate the expression across tissues a given gene has to be under regulation and thereby enriched in regulatory elements. Another observation in mammals reported housekeeping genes to be enriched in simple sequence repeats (SSR) in the 5'UTR regions when compared to tissue specific genes [133]. Tissue-specific (TS) genes displayed a significantly lower TE coverage in the upstream region (Figure 4.16) adding weight to purifying selection against insertion in cis regulatory elements to be a dominant force shaping TE abundance patterns upstream of genes.

A recent study generated a genomewide catalog of transcription start site (TSS) locations for maize genes [158]. Genes typically have multiple TSS sites which make a TSS cluster and the corresponding promoters make a promoter cluster [291]. TSS detection is important for getting accurate gene start coordinate and promoter identification [158]. TSS clusters have been classified as broad and sharp depending on the length of the region covered by CAGE tags and number of CAGE tags by statistics measuring the spread of the TSS cluster [191]. The broad and sharp classification has been shown to be biologically relevant with sharp promoters enriched in regulatory and developmental genes whereas broad promoters enriched in housekeeping genes [25, 100]. Maize was reported to have predominance of sharp TSS clusters when compared to arabidopsis (66 to 80% v.s 36%)[158]. It was hypothesized that sharp TSS clusters could be a result of more TE content in the genome where the transcription initiation is under selection to be more pinpointed and accurate [158]. This prompted to study the TE coverage in broad v.s sharp TSS cluster genes. Upstream TE coverage was found to be significantly lower in sharp TSS cluster genes when compared to broad (Figure 4.17) but no significant difference was found for downstream TE coverage. Since TS genes have been shown to be enriched in sharp TSS clusters [158] disentanglement of determinants of TE coverage would need more analysis. TE coordinates in the genome would be needed to looked in relation to TSS clusters to gain more insights into the reason for this observation. TSS cluster data information in Sorghum would be invaluable to test the influence of high TE content on TSS cluster shapes.

TE content is lower in sorghum compared to maize (60% vs 85%) [174]. This enrichment is also seen in TE content near genes and is distance dependent (Figure 4.18). TE abundance was calculated for 1KB to 5KB upstream for both maize and sorghum gene ortholog pairs and for higher distances the distribution of difference in TE coverage between maize and sorghum (maize TE coverage - sorghum TE coverage) is skewed towards positive values (Figure 4.18) indicating to an increase in TE content in the maize gene when compared to the sorghum ortholog. This points to more interactions between TEs and the regulatory apparatus of genes in maize. The correlation of upstream TE coverage between maize and sorghum orthologs indicates a long term persistence of factors influencing the upstream TE landscape, although no significant gene ontology terms were identified in genes with extreme differences in TE coverage.

We propose purifying selection on cis regulatory elements as a dominant factor shaping the upstream TE landscape in maize. The amount of cis regulatory DNA was measured by indirect proxies like gene ontology, tissue specificity and the TSS architecture. The proposed hypothesis of detection of cis-regulatory DNA by its transposon free 'footprint' does work on a coarse level with genes with higher conservation or amount of cis-regulatory displaying a depletion of TEs in upstream region. Although, fine scale discovery of functional elements seem not to be possible by TE based footprinting. It is interesting to note that most relevant studies in this regard were done in mammals. Earlier access to genomes and datasets could be one reason. But studies have shown plants to have fewer noncoding sites under selection lowering the amount of cis-regulatory noncoding DNA [145, 101]. We show that purifying selection against TE insertions makes these elements visible in the genome. This study included only the reference maize genome (B73) but increasing number of studies are now exploring the intraspecific landscape of TEs or so called 'Non Reference TEs' [252, 271]. A preference of insertions in or near genes is indicated [271] presumably due to TE specific insertion preferences or chromatin states. Our study predicts fewer such insertions near regulatory and tissuespecifc genes. Alternatively any such insertion could be given a preferential scrutiny for phenotype altering effects.

Genes involved in regulation also displayed lower TE coverage in downstream regions and the relation of TE coverage and gene expression was stronger in downstream regions. We hypothesize that read-through transcription could be a factor influencing purifying selection on TE insertions in downstream regions. In other words, if gene transcription does not end at the designated stop site (TES) and then downstream TEs could get transcribed. A study in mammals utilized differing tendencies of TEs to get transcribed and and reported that TEs more prone to transcription are less likely to be seen near genes [163]. Given the availability of deep RNA-seq datasets in maize, events of readthrough transcription can be detected including the proximal TEs which get transcribed. Also downstream epigenetic state needs a closer scrutiny in relation to TEs. A study in rice reported a strong influence of downstream methylation in repression of gene expression which would make downstream TE insertions deleterious [140]. TE insertions in 3'UTR regions have been implicated in loss of epigenetic silencing in arabidopsis [116].

It would also be worthwhile to compare TE content inside genes in relation to TEs near genes. Since selection is stronger in housekeeping genes [101], a TE insertion inside them would be deleterious. Conversely, since these genes have smaller cis regulatory DNA, TE insertions would be more commonly accepted in nearby regions. We find selection against insertion in cis-regulatory regions as a dominant factor shaping the TE landscape upstream of genes. We recommend that studies involving TEs and their interaction with epigenetic mechanisms (methylation and chromatin states) should consider regulatory complexity of the gene as one factor. The effect of downstream TEs need more studies and closer scrutiny to explore their effects, particularly on gene expression.

Selection Post Domestication

Materials and Methods

Obtaining SNP data

SNP data was obtained from from the maize hapmap2 project [36] which contains whole genome SNP data for 19 Teosinte(wild lines), 23 Landraces and 60 Modern inbred lines. VCF (Variant call format) file for maize hapmap2 SNPs was downloaded from the url `data.iplantcollaborative.org/quickshare/e75bc315fc0f9fda/HapMapV2RefgenV220120328.vcf.gz`. The 19 teosinte lines contain 17 lines from subspecies parviglumis and 2 from subspecies mexicana. The lines belonging to mexicana where removed because it forms a different subspecies. Two more lines TIL04-TIP285:TEO and TIL06-TIP496:TEO were removed as they are similar lines to TIL04-TIP454 and TIL06-TIP260 already present and differing only in the generations of selfing. One more line TIL02 was removed due to low coverage.

Calculating population genetics statistics

The VCF file was converted to ”hapmap” format using a custom perl script. Variscan [259] was used with runmode 12 to get values for nucleotide diversity (estimated as the average pairwise difference between (Π)) for each SNP. SNPs with less than 50% of genotypes called were removed from the calculations by setting the 'NumNuc' parameter of variscan to the number of individuals sampled. The per SNP diversity was added separately for each non-synonymous and synonymous SNP and the sum divided by the total number of sites for that category. The total number of synonymous and non-synonymous sites were obtained by modifying the Bioperl [234] module Bio::Align:DNAStatistics. For genic statistics, variscan was run with a 'block data file' listing the coordinates of the gene to get the per base pair Π and Tajima'D values for each gene. For sliding window based statistics, variscan was run in sliding window mode 'SlidingWindow=1'. For display in the genome browser a sliding window of 100Kb with a 10Kb slide was chosen. The genomewide statistics reported were for 10kb window with no slide.

Obtaining Derived allele state

The list maize syntenic orthologs with sorghum was downloaded from James Schnable's website url: `http://skraelingmountain.com/datasets/grass_syntenic_orthologs.csv.zip` [216]. For each gene, the splice variant annotated as 'canonical' was chosen for further analysis. The sorghum-Maize ortholog protein pairs were aligned using clustalw [250] and the amino acids of in the alignment were replaced by corresponding codons using software pal2nal [239]. These alignments were parsed using a custom perl script and the sorghum nucleotide extracted for SNP coordinates. The allele of the SNP which matched the sorghum was assumed to be the ancestral allele.

Calculating recombination events

As the sequenced maize/teosinte lines in hapmap2 project were inbred and contained few heterozygous SNPs, the genotypes along a chromosome for each sample essentially represented a chromosome-long haplotype. Thereby no phasing was required and recombination events could be identified using a 'four gamete test' [103], this number, also called as 'Rm' is an underestimate of the recombination rate, as recombination events not marked by SNPs can not be detected. The software 'RecMin' [165] was used, which calculates a variant of 'Rm' called as 'Rh'. The values of 'Rh' were found in our case to be positively correlated with sample size as also shown by Meyers and Griffiths [165]. As the number of samples differed between three groups (WILD, LANDRACE, IMPROVED), a same number of samples (N=16) were randomly chosen from each of the groups. This sampling was done three times and the results were found to be consistent across all three random samples, so the results from one sample are only shown here. RecMin needs a haplotype for each sample for the region analyzed. For this first a genome sequence was created for each sample by using the reference sequence (B73) and replacing the reference nucleotide with the alternate allele for each SNP. For example, if the reference contained an A at a particular site and genotype at the site in sample was G/G then the A was replaced by a G to create the sample specific genome fasta file. The genic regions in each of these fasta files were extracted and formed the haplotypes used as RecMin input. RecMin was then run with default parameters.

Calculating DoFE

Eyre-Walker and Keightley's method for calculating DoFE as implemented in software DoFE [59] was used. Software was downloaded from Adam Eyre-Walker's lab `http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/Website/Software.html`. DoFE needs a frequency spectrum for two categories which in the case were non-synonymous and synonymous SNPs and total number of sites in each category. Since the sample size varies for the SNPs due to missing calls 13, 18 and 48 alleles were sampled randomly for each SNP for WILD, LANDRACE and IMPROVED groups. SNPs with less than the set number of alleles were ignored. Also required are the number of fixed differences and sites in each category which were calculated by maize sorghum alignment of orthologs using bioperl [234] module Bio::Align:DNAStatistics. These numbers were reaffirmed by similar computation with DNAsp software [206] for randomly chosen genes.

Obtaining SIFT scores

SIFT is an algorithm for predicting the deleterious effect of non-synonymous polymorphisms [170]. It is based on site conservation and nature of amino acid change caused by the SNP. It gives a score between 0 and 1, with 0 being intolerant and 1 being neutral. Typically a score between 0 and 0.05 is considered to be deleterious [223]. Sift scores for maize were downloaded from the sift4g website (`siftdb.org`).

Results

Installation and configuration of Genome Browser

The torrential downpour of data created by population scale sequencing projects needs tools to for analysis and visualization, specially in the context of existing genome annotation. UCSC genome browser is a popular and time tested tool used primarily by the animal research community [120]. The official version of this browser does not support plant genomes. As a first part of my work I installed a configured clone of UCSC genome browser with common plant genomes like maize, sorghum and tomato (Figure 5.1). The scattered gene annotation information for these plant genomes were obtained from various sources and compiled for display in the browser. Figure 5.1 displays chromosome 1 of maize and values of TajimasD [243] for wild(teosinte), landrace and improved maize lines. As expected (see introduction), modern improved lines display higher number of genomic regions with negative values. Other custom genomes being sequenced in the lab were also added. This tool became not only became an integral part of my work but is also used by collaborators and colleagues.

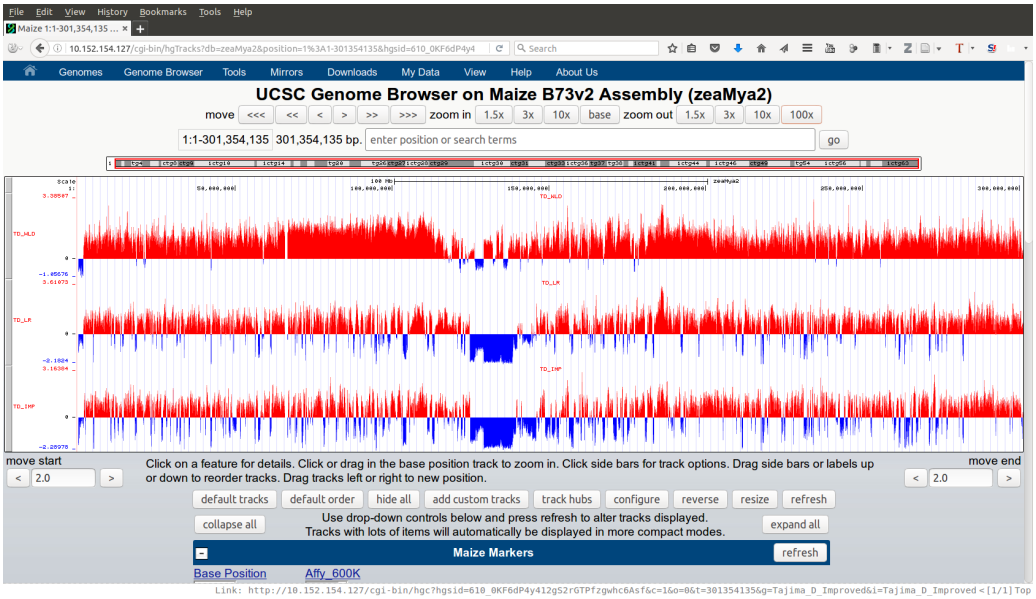


Figure 5.1: Locally installed version of UCSC genome browser configured for maize genome. Displaying TajimasD over entire maize chromosome 1 for teosinte, landraces and modern inbred lines.

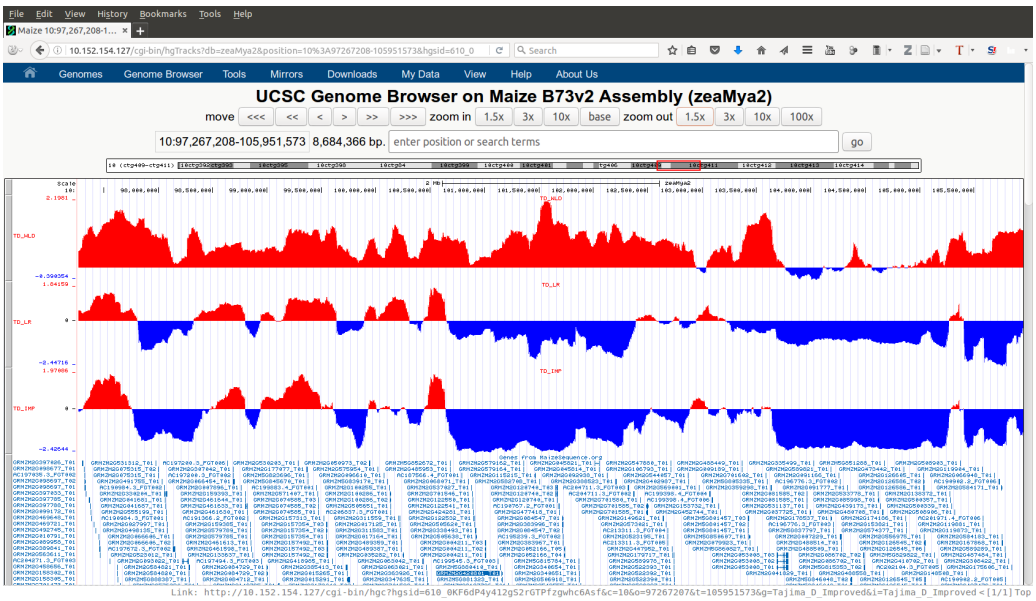


Figure 5.2: Zoomed in region of chromosome 10 displaying a massive selective sweep first reported by Tian et al. [251]. Also seen in bottom are the genes falling in this sweep, further information about the genome including its GO categories can be seen by clicking on the gene id.

Genomewide diversity in maize line groups

Median of genomewide persite nucleotide diversity (Π) and Tajima's D were calculated for 10kb non-overlapping windows and are given for the groups are given in Table 5.1. A consistent decrease in both is seen from wild to landrace to improved lines. Progressive decrease in Tajima's D can also be seen in the browser graphic (Figure 5.1). Qualitatively similar trends were also reported by the hapmap2 project [104] and are expected following the history of maize domestication and improvement (see introduction). Persite diversity for genes and TajimasD are shown in Table 5.2. The non-synonymous and synonymous diversity and the ratio of them are displayed in table 5.3. Πn/Πs increases from wild to landrace to improved although only two comparisons were significant namely, WILD v.s IMPROVED (P<2E-9; Wilcoxon rank sum test) and LANDRACE v.s IMPROVED (P<1E-8; Wilcoxon rank sum test).

Table 5.1: Median of Nucleotide Diversity (Π) and Tajima's D calculated for 10Kb windows over the genome for different groups.

Statistic	Wild	Landrace	Improved
Median Π(persite)	0.0018	0.0017	0.0015
Median Tajima'sD	0.12	-0.14	-0.58

Table 5.2: Median(mean) genic diversity and Tajima'sD.

median(mean)	Wild	Landrace	Improved
Π	0.00545(0.00602)	0.00455(0.00519)	0.00452(0.00516)
Tajima'sD	0.37(0.36)	0.47(0.37)	0.46(0.32)

Table 5.3: Median(mean) Πn, Πs and Πn/Πs

Median(mean)	Wild	Landrace	Improved
Non-Synonymous(Πn)	0.0016(0.0022)	0.0013(0.0019)	0.0015(0.0021)
Synonymous (Πs)	0.008(0.0092)	0.0063(0.0077)	0.0071(0.0083)
(Πn/Πs)	0.207(0.323)	0.209(0.353)	0.223(0.412)

Shared polymorphisms in different populations

Polymorphisms were defined as shared when a SNP segregates (Allele frequency >0 and <1) in atleast two groups. As there were three groups (WILD, LANDRACE and IMPROVED), the sharing status was represented as a 3 digit code ([0,1][0,1][0,1]) for the three groups respectively with a 1 indicating segregating and 0 indicating non-segregating for the three populations. For example, 110 would mean that the SNP segregates in both WILD and LANDRACE groups but not in IMPROVED. The number/percentage of SNPs shared between different groups over the entire genome is displayed in Figure 5.3. SNPs with a code of '111' (i.e segregating in all 3 groups) form the largest category with 36% of total SNPs (Figure 5.3).

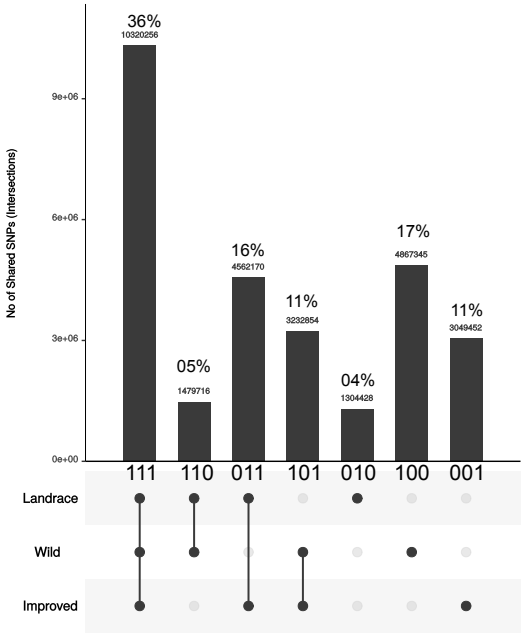


Figure 5.3: Number of SNPs segregating in different groups. X-axis also gives the status of SNP encoded as Segregating (1) and Non-Segregating(0) in Wild, Landraces and Improved lines respectively. For example 111 means that the SNP is segregating in all three populations.

Derived allele frequencies

Derived allele frequency (DAF) gives an indication of the age of the SNP and the selection acting on it [227]. DAF was obtained for coding SNPs and is divided into two categories (synonymous and non-synonymous). The boxplots of derived allele frequencies in different groups and for different classes of SNPs (according to their sharing status) are displayed in Figure 5.4, Figure 5.5 and

Figure 5.6 for WILD, LANDRACE and IMPROVED groups respectively. Boxplots were made for non-synonymous and synonymous SNPs separately. Synonymous SNPs display a higher DAF when compared to corresponding DAF for non-synonymous SNPs. The 111 category displays highest median DAF in all three groups. Just to display this more clearly the DAF of 111 SNPs in all three groups is separately plotted in Figure 5.7.

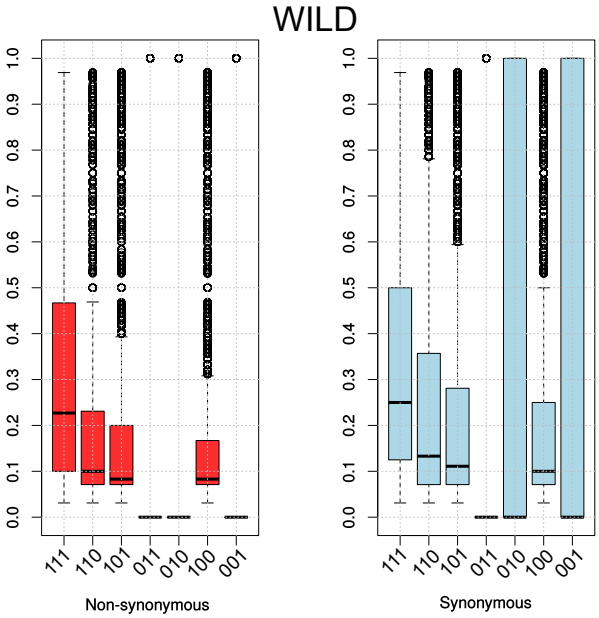


Figure 5.4: Boxplots of derived allele frequency for classes of shared polymorphisms in WILD group. Non-synonymous SNPs in red and synonymous in blue.

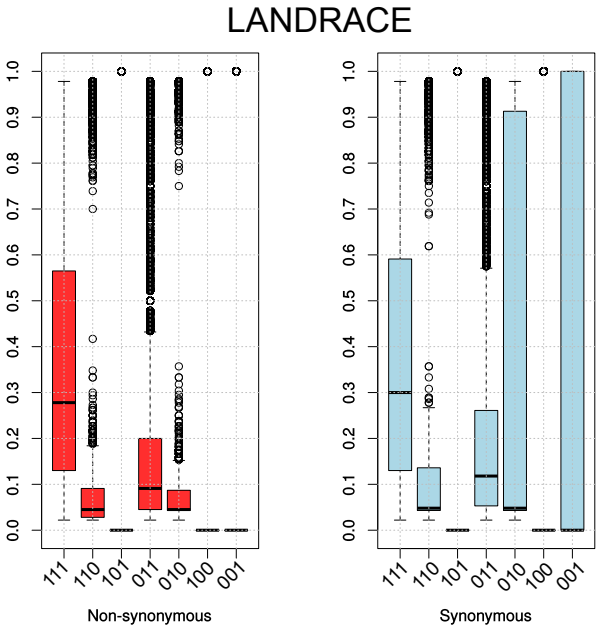


Figure 5.5: Boxplots of derived allele frequency for classes of shared polymorphisms in LANDRACE group. Non-synonymous SNPs in red and synonymous in blue.

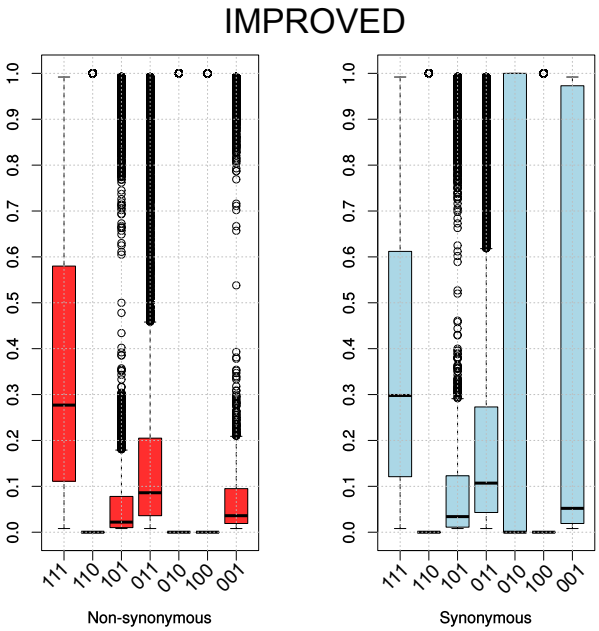


Figure 5.6: Boxplots of derived allele frequency for classes of shared polymorphisms in IMPROVED group. Non-synonymous SNPs in red and synonymous in blue.

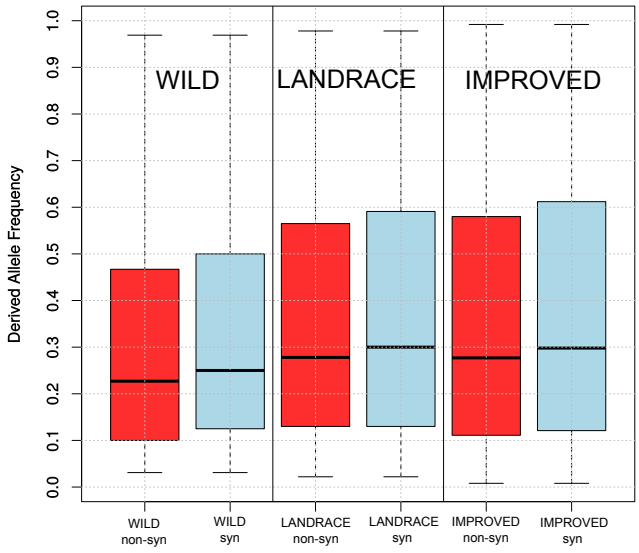
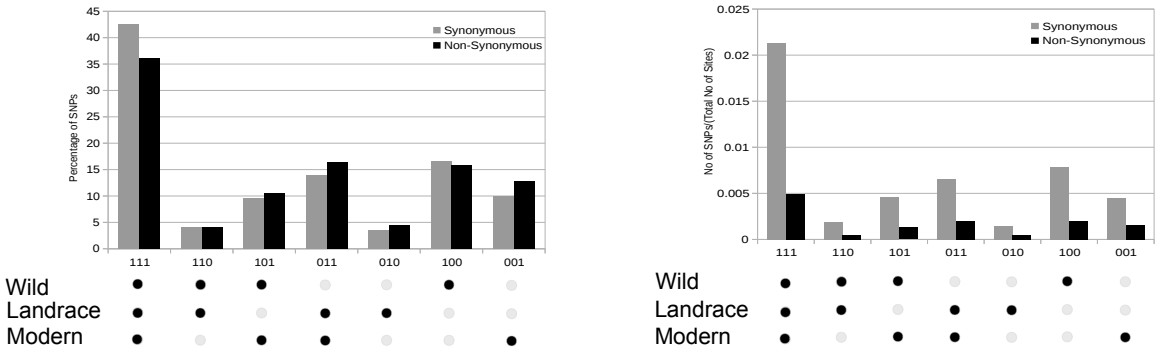


Figure 5.7: Boxplots for derived allele frequency for class '111' (Shared in all three groups). syn(Synonymous,blue), non-syn(Non-Synonymous,red)

Shared polymorphisms and Purifying selection

The percentage of coding SNPs for both categories in displayed in Figure 5.8a. The category '111' shows the most difference between synonymous and non-synonymous SNPs. Since the total number of non-synonymous sites in the genome are more than the synonymous sites, the number of SNPs were divided by the total number of non-synonymous and synonymous sites. This gives the abundance of

SNPs in proportion to the number of potential sites in the genome and is displayed in Figure 5.8b. In order to further examine the proportion of shared SNPs with respect to the deleterious nature of the SNP, SIFT scores were obtained for non-synonymous SNPs (See Methods). Non-synonymous SNPs were classified as deleterious (SIFT <0.01) and benign (SIFT >0.01). The percentage of SIFT-deleterious and SIFT-benign SNPs in each category of sharing is displayed in Figure 5.9.



(a) (b)

Figure 5.8: (a)Fraction of synonymous and Non-Synonymous coding SNPs segregating in different groups. (b)Fraction of synonymous and non-synonymous coding SNPs divided by the total number of sites in different groups.

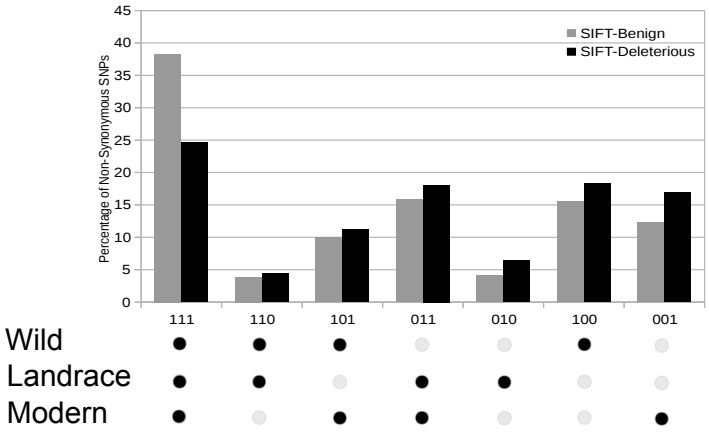


Figure 5.9: Percentage of Non-synonymous SNPs divided in two categories benign and deleterious according to SIFT score. X-axis gives the status of SNP encoded as Segregating (1) and Non-Segregating(0) in Wild, Landraces and Improved lines respectively. For example 111 means the SNP is segregating in all three populations.

Differences between purifying selection between groups

The strength of purifying selection was measured by DoFE between three groups. A progressive decrease in fraction of mutations in highly deleterious class (Ns>100) can be seen for WILD, LANDRACE and IMPROVED lines (Figure 5.10). Also displayed is α which is the fraction of sites under positive selection.

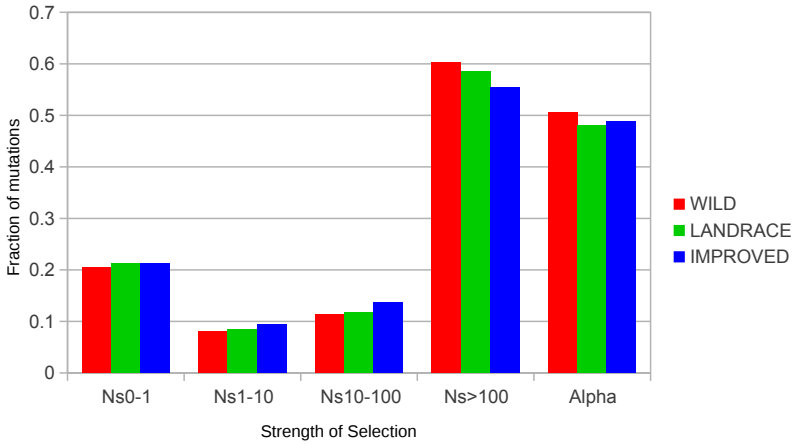


Figure 5.10: Distribution of fitness effects (DoFE) for three groups.

Differences in Number of Recombination Events

The number of recombination events in the history of the sample (Rh) were calculated for canonical splice variants for the three groups separately. Only genes with a minimum 10 SNPs were only used in the analysis. The Rh per gene was normalized by dividing by the number of segregating sites in a gene giving a normalized value (Rh-norm). The mean(median) for WILD, LANDRACE, and IMPROVED was 0.0026(0.0032), 0.0019(0.0014) and 0.0017(0.0012). All three pairwise comparisons were significant (P<2E-16;Wilcoxon rank sum test). The density plots for all three groups are represented in Figure 5.11. As could be expected, WILD lines displayed a higher number of recombination events.

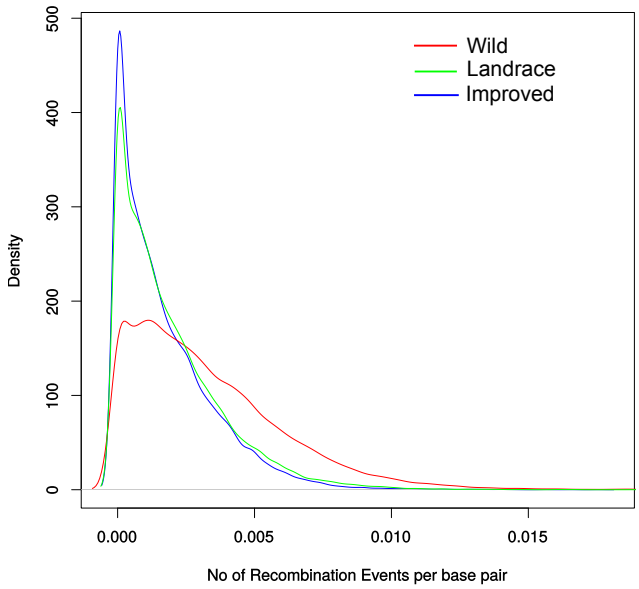


Figure 5.11: Density plots for recombination events per base pair (Rh-norm) for three sample groups.

Recombination and Purifying selection

In order to explore the link between recombination and purifying selection in each group, the Rh-norm values for each group were binned according to quartile value in 5 bins (0-5) with bin zero genes having no recombination event (Rh-norm is 0) and bin five with genes in fourth quartile of Rh-norm values. Purifying selection was measured by the ratio of non-synonymous to synonymous diversity (Πn/Πs). Boxplots for Πn/Πs, Πn and Πs values in each recombination bin are displayed in Figure 5.12. Πn/Πs decrease with increasing Rh-norm but the trend levels off at higher recombination bins. Both Πn and Πs increase with Rh-norm Figure 5.12. These results are consistent with results shown in drosophila by Campos et.al [22], outlining the role of recombination in increasing the efficiency of purifying selection.

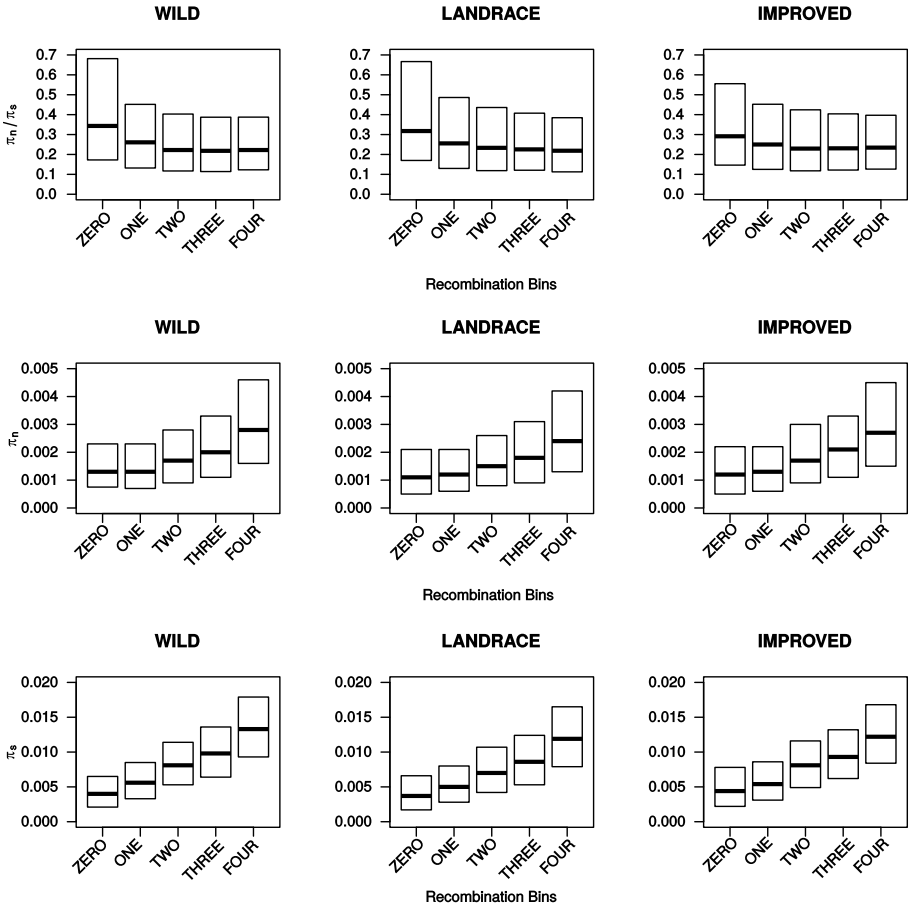
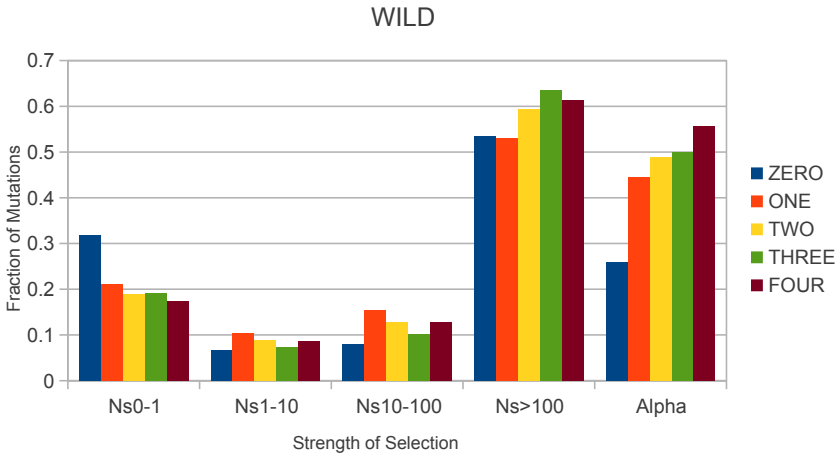


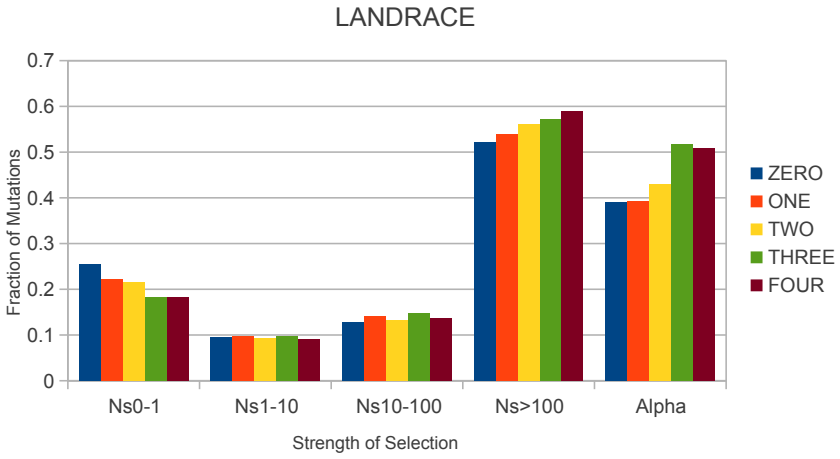
Figure 5.12: Boxplots for Πn/Πs,Πn and Πs in relation to bins (zero to four) based on increasing number of recombination events. Data is displayed for three groups WILD, LANDRACE and IMPROVED.

Recombination and DoFE

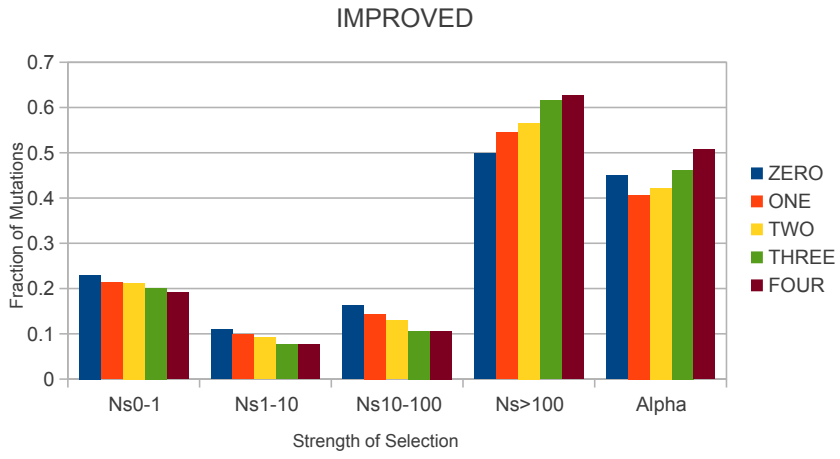
The DoFE and α was obtained for genes in each of the recombination bins for the three groups and is displayed in Figure 5.13a, 5.13b and 5.13c. Strength of purifying selection increases with recombination as displayed by progressively larger fraction of genes in higher -NeS class in all three groups. The fraction of sites under positive selection α also increases with increasing recombination.



(a)



(b)



(c)

Figure 5.13: Distribution of fitness effects in different recombination bins in three groups. (a)WILD (b)LANDRACE (c)IMPROVED

Discussion

A large fraction of polymorphism is shared between maize and teosinte (Figure 5.3) and a striking 36% of polymorphisms are seen segregating in all three groups over the whole genome. Given the high nucleotide diversity in maize [246, 282] and a very mild reduction [36] when compared to the wild progenitor population (see introduction) this is expected. The nature of domestication induced bottleneck (strong but short) and high diversity in the progenitor have been indicated as the possible reasons [57]. Post domestication geneflow from teosinte specially into landraces due to outcrossing nature and viable offsprings with teosinte also could have contributed to the high diversity and increased the number of shared SNPs. These processes could have prevented the diversity loss even if a single domestication event for maize is the general consensus. The selection of the individuals should make a difference, particularly in the case of LANDRACE and IMPROVED lines as the deviation from the 'panmictic' behavior would mean individual subgroups within samples and uneven sampling of diversity (especially for SNPs in low frequencies). A newer version of hapmap2, which recently became available (hapmap3 [21]) which includes a staggering 916 lines should ameliorate the sampling issue to an extent. One of the unexpected observations in Figure 5.3 is that the '110' category has less than half the number of SNPs compared to '101' category. More gene-flow between WILD and LANDRACE groups is expected compared to WILD and IMPROVED, so more number of SNPs should be seen in '110' category vs '101' category. This is due to the physical proximity of teosinte to sampled landraces (in terms of geographical location) and open pollination [104]. Adding to this trend is the number of unique SNPs in WILD, LANDRACE and IMPROVED lines (100, 010 and 001) with 17, 4 and 11 percent of the total SNPs respectively (Figure 5.3). More unique SNPs in WILD sample is expected due to high diversity and a panmictic population. One technical aspect to note is that the reference genome used for mapping and SNP calling was of the maize line B73, a modern inbred line and a member of the IMPROVED group. Thereby a 'reference bias' [230] exists whereby comparatively fewer sequencing reads map with the increasing distance of the sample from the reference due to divergence. Reference bias can lead to an underestimation of the number of SNPs in distant samples. This does not seem to be case as WILD samples should be more affected due to much higher divergence to IMPROVED when compared to LANDRACES, but 100 has the highest number of SNPs amongst category of private SNPs. Also qualitatively similar fractions of SNPs were obtained when only coding SNPs gathered from maize genes with syntenic orthologs in Sorghum were used (Figure 5.8a). A frequent introgression from WILD to IMPROVED lines could explain a higher fraction of '101' SNPs compared to '110' but it is unlikely because the origin of IMPROVED lines is very recent (in the order of decades) and is much controlled and documented.

The expected age of an allele is strongly related to its frequency [227]. Shared SNPs represent pre-domestication polymorphism which is still segregating, thereby at-least 9K years old [154] as the likelihood of a mutation striking at a site independently in two groups is supposed to be negligible. Bottlenecks and selection influence the allele frequency and so shared SNPs represent a good case where these effects can be studied by looking at their allele frequency in two groups (WILD and IMPROVED). Polarizing the SNP infers the direction of the mutation which led to the SNP and

identifies the new (derived) and the old allele (ancestral). The boxplots for derived allele frequency (DAF) in three groups are presented in Figure 5.4, 5.5 and 5.6 for non-synonymous and synonymous SNPs. The subset '111' displays the highest DAF in all cases. This is expected given the '111' class predominantly represents old pre-domestication SNPs. Although, an old pre-domestication SNP can sometimes not be categorized as '111 due to fixation of the SNP in one or more of the groups. Also the variance of the '111' SNPs frequency is increasing after domestication showing the expected effect of bottleneck (Figure 5.7). In all comparisons non-synonymous SNPs have a lower median frequency than synonymous indicating the effects of purifying selection throttling the allele frequency ascent. SNPs unique to each group are likely to be caused by young mutations reflected in their lower average DAF. The '110' and '101' SNPs have a lower median DAF than '111' SNPs in the WILD group (Figure 5.4). The reason for this is not clear as their DAF was expected by us to be close to '111' in WILD group because they also represent shared polymorphism and thereby expected to be old. Also the DAF for '101' and '110' in LANDRACE and IMPROVED is closer to unique SNPs ('100', '010' and '001') (Figures 5.5 and 5.6). A relatively recent gene flow by introgression could explain this but more investigation is needed including identifying the distribution of these SNPs along the genome. The DAF for '011' category is higher than the unique SNPs ('010' and '001') which is consistent with their origin in the LANDRACE and then passing on to IMPROVED lines. Although, it should be noted that the unique SNPs ('100', '010' and '001') could be caused by new mutations or by fixation/removal of a shared SNP in all but one groups. For unique SNPs, the difference in synonymous and non-synonymous SNPs is striking. For example, the category '001' should not segregate in the WILD which can be seen in Figure 5.4. Most non-synonymous SNPs in this case have DAF of zero in WILD but for synonymous SNPs fixation seems to have occurred in many cases, which makes DAF 1 and thereby the width of the barplot for synonymous SNPs covers range from zero to one.

Purifying selection tends to suppress deleterious variants. As the shared SNPs comprise largely of older polymorphisms purifying selection must have acted longer on them and the surviving SNPs must be enriched in benign variants. This was already reflected in the DAF plots (discussed earlier) with non-synonymous SNPs having a lower average DAF. To explore this further, first the fraction of non-synonymous and synonymous SNPs was examined. The '111' SNPs show an enrichment of synonymous SNPs (Figure 5.8a and 5.8b) indicating that purifying selection has removed non-synonymous disproportionately more from this class. The '100' SNPs which are new SNPs in WILD also show an ever slight enrichment of synonymous SNPs contrasting with '001' which shows an enrichment of non-synonymous SNPs. When non-synonymous SNPs were classified as benign and deleterious, all categories of shared SNPs showed a higher fraction of deleterious non-synonymous SNPs except '111' which is enriched in the benign category (Figure 5.9). In addition strength of purifying selection between different groups was assayed using DoFE (see introduction) (Figure 5.10). LANDRACE and IMPROVED groups show a fewer fraction of SNPs in the highly deleterious category indicating of weaker purifying selection. These differences consistently point towards purifying selection to be stronger in the WILD. Although a recent report indicates a stronger purifying in maize on young alleles and a $\sim$3 fold increase in the population size compared to ancestral teosinte before

domestication [7]. Thereby more analysis is needed to conclusively prove that domestication has lowered the strength of purifying selection in modern maize.

The strong increase in LD post domestication in maize [278, 104] is bound to make selected and neutral sites less independent of one another. This effect would be seen for both purifying and positive selection via background selection and hitchhiking respectively (see introduction). Campos et.al [22] studied the relation between recombination and diversity comprehensively in Drosophila. They binned genes w.r.t number of recombination events and assayed for differences in purifying selection between bins. They showed that recombination is positively correlated with for the strength of purifying selection. An increase in α (which is the fraction of sites under positive selection) was also observed in higher recombination bins. Maize presented a fitting case to test these predictions. A strong difference in the number of recombination events between groups is depicted in Figure 5.11. This difference provides a contrast to test the effects of linked selection. An increase in strength of selection as measured by a decreasing ratio of non-synonymous to synonymous diversity (Πn/Πs) is depicted in Figure 5.12. Recombination has also shown to be positively correlated with diversity and although both Πn and Πs increase with number of recombination events (Figure 5.11), a direct relation could not be assumed due to the way in which recombination events were calculated. As detecting recombination itself depends on diversity and availability of markers a positive relation between recombination and diversity is technical. Independent measurements of recombination rate from experimental work or studying regions in the genome which are known to have differing recombination (e.g centromeres vs telomeres) would be needed to test the prediction in maize. Nevertheless the biological effect of recombination on increasing diversity has been documented in Drosophila and humans. Stronger purifying selection in high recombination bins was also supported by DoFE (Figure 5.13a, 5.13b, 5.13c for WILD, LANDRACE and IMPROVED).

Recombination also influences adaptation in a positive way by delinking the adaptive variant with proximal deleterious variants. An increase in α (which is the fraction of sites under positive selection) can be seen with increasing recombination in all three groups (Figure 5.13a, b and c). Also α is slightly higher for the WILD group (Figure 5.10), implying a higher fraction of sites under positive selection in WILD group. This seems confusing as higher α is intuitively expected in the LANDRACES and IMPROVED lines due to selective breeding for enhancements of selected traits. We have three explanations, first is that the divergence data was taken from Sorghum so α measured in the maize lineage is the result of adaptation post maize-sorghum divergence ($\sim$11MYA) and not just post domestication ($\sim$9KYA). Secondly, the divergence was calculated using the reference maize genome (B73) (which is in the group IMPROVED) and more analysis is needed to see the effect it has on the determination of α. Hapmap2 also sequenced gamma grass genome (*Tripsacum*) which can be used to obtain divergence (divergence from maize around $\sim$45KYA [95]). Thirdly, adaptive events must have occurred in teosinte lineage after the maize domestication event but studies are highly biased towards detecting adaptation in maize. Also higher recombination in teosinte would mean a faster dilution of the sweep signal.

Future Perspectives

Purifying selection plays an important role in shaping the genome of a species, Although adaptive evolution often in the focus, especially for domesticated species, purifying selection is a pervasive force [168, 58]. Maize is a classical study organism with a legacy of many important discoveries in the field of genetics. In parallel, maize is equally important for applied research and its implications. In terms of data throughput there is ample and growing amount of genomewide datasets which include a high quality genome, polymorphism surveys, gene expression in varying conditions and tissues and epigenetic, proteomic and metabolic studies. Although maize is well studied, ceaseless surprises are thrown and many inviting and unanswered questions remain. The very origin of maize was a mystery (see introduction). Heterosis is an aspect of maize with has received a lot of applied and empirical scrutiny but the details of mechanisms still evade us. In the current era of high-throughput genomics, maize has continued the tradition of expecting the unexpected. An example is the massive presence absence variations (PAVs) where two maize lines can differ often in thousands of genes [240]. Also intergenic regions are often found to be nonhomologus between two maize lines due to transposon activity [50, 264]. But maize still maintains integrity as a species despite the stormy nature of the genome. Three aspects of maize became interesting to me to explore the role of purifying selection in maize.

Maize is the only species of the grass family with a recent and well studied WGD event. Subgenomes are an important aspect of WGDs in order to classify genomic regions post WGD and have also been identified in maize, although exact mechanism of their origin are unknown. The first part of this work showed differences in purifying selection between subgenomes, and the dominant subgenome (maize1) was shown to be under stronger purifying selection. Dominant gene expression was then shown to be main reason behind it. An expression based classification of WGD duplicates was developed which better explained the observed differences. This also agrees with the general observation of purifying selection to be expression dependent in plants. Subgenome based classification can be difficult to grasp and first and may sometimes seem circular. Significant differences have been shown between subgenomes in expression, phenotype and selection, but in general the effects have been weak. Follow up studies have failed to report any differences in transposon content, methylation and interaction networks between subgenomes [186, 55, 138]. The expression based classification developed in this work has shown promises in this regard. For example the TE content, methylation and splicing were found to be significantly different between expression categories. A recent study found no subgenome bias in co-expression networks [138]. Such an analysis could be conducted on the expression based classification to check if dominant expression is related to more interactions. Since the post WGD gene deletion in maize still continues as seen in modern inbred lines, expression based classification can be used to check if deleted genes have a preference for repression in expression. Expression based classification can be overlayed with protein interaction and metabolic network available for maize [208] to check if the dominance/repression in expression is coordinated along pathways. This work also revealed repression of expression to explain the lack of complementation of phenotype by the duplicate copy. Dominant expression can be used for screening/modification of phenotypes. A study of genomewide trait prediction in maize reported a higher predictive ability in maize1 subgenome regions [220]. Dominant expression of genes can be tested as a factor in the trait

prediction. As adaptive evolution was not covered in this work, the relation between expression classification of duplicates and positive selection still has not been explored. Dominant expression of a gene was associated with a visible phenotype when mutated indicating that the dominantly expressed duplicate copy performs a larger component of function. It is likely that dominantly expressed duplicates were more likely modified in maize domestication and improvement. Expression based classification can also be used to explore the mechanisms behind subgenome formation. One approach in this regard would be look at the proximity of genes with dominant expression along the genome. This would help to identify if there is a clustering of dominantly expressed genes. Regions with high number of dominantly expressed genes could then be assayed for possible mechanisms like chromatin and methylation states. A study in humans reported clustering of gene expression changes which could even be observed in regions exceeding 100KB [77]. Expression data from Sorghum could be used to 'polarize' gene expression to obtain ancestral expression state and the direction of expression change between duplicates determined. The work also reported strong correlations in divergence and diversity between WGD duplicates which are unexplained. These correlations were stronger for introns indicative of neutral processes involved. Population genetics simulations involving duplicate genes under various scenarios could be one way to explore the forces acting on the duplicates. Since duplicates have similar lengths and 'functional density' (proportion of sites contributing to function) they present a good study-set for assessing the effect of these factors on diversity. The two subgenomes also share same demography so they can be utilized as two replicates in demographic studies. The outliers from the correlations can be used to assay the factors like adaptive evolution or neofunctionalization and differences in breeding induced selection between duplicates.

A substantial ($\sim$85%) portion of maize genome are transposons (TEs). Maize experienced a lineage specific increase in TE content post divergence from sorghum ($\sim$11MYA). TE insertions often cause non coding regions to display no homology when compared between two maize lines. This work reported significantly more upstream TE content in post WGD duplicates displaying repression in expression. I suggest to build alignment of upstream regions of post WGD duplicates to identify cis-regulatory regions. TE insertions in the cis regions could be one way to look for the influence of TEs in repression of expression. Available whole genome methylation datasets could identify if repression of TEs causes repression of the duplicate copy with higher TE content. The work identified the amount of cis-regulatory upstream regions as a dominant factor shaping the TE abundance upstream of genes. Tissue-specificity was used as a proxy for selection on cis-regulatory elements. This is because being tissue specific involves complex regulation in order to keep the different expression states across tissues. Although only one measure of tissue-specificity was used (tau), but several measures exist [127] and can be tested to identify the measure which better captures the regulatory complexity in maize. Differences in expression across time points has been used to develop a new measure of regulatory complexity [238] which as been associated with conservation of upstream region. The measure could be modified for tissues instead of timepoints and the conservation of upstream regions be replaced with TE abundance measures. Simple measures like fraction 1KB region covered by TEs were used to quantify TE abundance. Since the effect of a TE insertion also depends on the distance of TE in relation to the gene, a new measure which combines both the length of TEs

and the distance of the TE would be more sensitive to study the effect and responsible factors for TE abundance. Type (family) and orientation of TE in relation to direction of transcription were ignored in this study but could be additional factors to be tested. DNA methylation status of TEs is a crucial factor when studying their effect on gene expression. Although multiple genomewide methylation datasets have been published for maize, the data available is usually raw sequencing reads and generally methylation calls are not provided. One reason might be the lack of popular and standardized file formats for representing methylation data. Development and general acceptance of VCF like format for methylation data is important for efficient use and reuse of these datasets. Cis-regulatory regions of genes are generally not included when studying plant TEs. This work highlighted their importance in shaping the TE landscape. An approach which combines cis-regulation, TEs, gene expression and methylation status will give a comprehensive view of TE effects on gene expression. TE insertions upstream of genes with robust and shorter cis-regulatory elements like housekeeping genes or genes involved in catalytic activity would be expected to be neutral and display higher methylation levels. In contrast, TE insertions upstream of genes with more comprehensive regulation and elaborate cis elements would be deleterious and with low methylation levels. An extension to more functional aspects of methylation like imprinting could also be made in an attempt to approach the question of why some genes are regulated by methylation. In this case methylation can act on the cis elements to provide an additional layer of regulation. If true then imprinted genes are expected to display longer cis-regulatory regions. Downstream TEs need more scrutiny as isolated cases of them influencing gene expression have been reported [140, 116]. This work reported their effect on gene expression and their abundance, which is although higher than upstream regions is still much smaller than over the whole genome. Deep paired end RNA sequencing datasets could be used to test the existence and prevalence of readthrough transcription where these TEs get inadvertently transcribed. The extent of readthrough transcription can then be tested for correlation with gene expression and DNA methylation in downstream regions.

Maize diversity is on the higher end in plants, this is particularly striking because domestication has not caused the strong diversity reduction usually seen in other species (see introduction). The third part of this work quantifies differences in purifying selection between teosinte and modern maize. The first approach was to test the differences between shared and private SNPs between groups (WILD, LANDRACE and IMPROVED). Shared SNPs were shown to be older and under reduced purifying selection. The distribution and age of shared SNPs along a genome could be used to detect recent introgression. The work also showed shared SNPs to be depleted in deleterious variants which could make them good neutral markers. Since a large fraction of SNPs is shared, an outlier based for detecting genomic regions with reduction/fixation in the number of shared SNPs could be used to detect maize specific adaptive evolution or differing strength of purifying selection. Next, recombination bins were used to test the strength of purifying selection as a function of recombination. Both purifying and positive selection was shown to be positively correlated with the number of recombination events. This is in agreement with the theoretical predictions. WGD duplicated genes but with differing number of recombination events in groups can to used to check for differences in purifying selection as it controls for gene length and functional density. Combining of shared poly-

morphisms and recombination bins could be used to test for the prediction of a diversity increase due to recombination. Gene ontology analysis of genes in extreme recombination bins in different populations could provide new biological insights specially if the enriched functional categories vary between different groups (WILD, LANDRACE and IMPROVED). For example, pathogen resistance genes are predicted to be recombination hotspots [38]. A recent study reported strong purifying selection in maize and an effective population size $\sim$3 times larger than ancestral teostine [7]. Since this work reports a decrease in the strength of purifying selection in modern maize, more analysis with a larger number of samples is needed. Hapmap3 with more than 900 maize lines now provides a more comprehensive dataset [21]. The value of α (fraction of sites under positive selection) was found to be much higher than reported for plants [80, 101]. This needs further scrutiny as older studies in maize were conducted involved fewer loci. A study in sunflower did report higher values of α but it was positively correlated with effective population size [235]. A key question for further studies in teosinte and modern maize is then to investigate if positive selection (α) in modern maize is dating from teosinte or reflects selection by humans for domestication. Overall a recalculation of α in different plant species is also needed given the increased availability of genomewide polymorphism coverage datasets.

Acknowledgements

Acknowledgements sometimes seem clichéd and templated if seen externally. But for the person having gone through the endeavor and being aided by many people who cross paths along the journey, they have a special meaning. I would like to first thank Aurelien for giving me a second chance to pursue my interests and also rebuild my life. This implied a lot of trust being put in me. He has been, in general, encouraging to people who first might seem unconventional but posses a love and enthusiasm for science. He has also been aware and understanding of the fact that research is seldom if ever, a straight line between two points but a complicated maize of paths where the final outcome is unknown until the end and thereby accommodating of delays due to complications in work and personal life. The most challenging aspect of doctoral work is the faith one has to sustain, in the favorable outcome of the project, for a long time. He kept the faith for me and after every discussion I felt renewed and energized. He was never even slightly hesitant to discuss unfavorable or negative outcomes and always managed to see and show me the positive side. Also, he let me take initiative and supported me in the entire process and in the end let me take the credit.

The atmosphere in the lab has been extremely cozy, warm and lively with a lot of fun, events and scientific and non-scientific exchange which made my stay here one of the most memorable time of my life. Over the many years, I had the pleasure of meeting some amazing and bright people who joined the lab. Silke who is our team assistant provided a constant forcefield protecting against bureaucracy and paperwork. She has been incredibly forgiving of my carelessness in official paperwork and went out of the way to help me. Daniela handled all the experimental work in the lab. She has been a true friend and I always found her concerned and helping. She also help make my life very much easier outside work and helped me adjusting to life in Germany. I will always remember her spirited personality and conversations. Hanna, helped me a lot with her amazing skills with R and Latex. She patiently listened to my incoherent ramblings about work and always had excellent advice and perspective to give. With her help, I honed my earlier nearly non-existent baking skills. I will always remember how I could drop in her office anytime with a problem and always went out with a solution be it technical, scientific or personal. I was always assured an excellent advice in scientific, professional and career matters by Remco. His sharp skills in observation and analysis were also accompanied by a witty sense of humor which could add light to any conversation. I will always remember how he was always welcoming with an open office door, which I exploited often for conversations which started with science but then drifted lightyears away.

The maize community has reputation of being very open with regard to advice and data sharing. I personally witnessed this in my interactions with researchers in the maize breeding department headed

by Chris-Carolin Schön. I often dropped by for advice and always felt welcome. I was amazed by the knowledge, experience and insights of Eva and Sandra about maize genomics and breeding and touched by their generosity in sharing it with me. I admired the ability of Sandra to examine and analyze scientific details with microscopic precision. She helped me a lot with scientific, technical and literary aspects of my work. I will always remember the marathon discussions with her about maize and miscellaneous and her generous personality and I also feel immensely fortunate to have her as my colleague and friend.

I would like to thanks Georg Haberer for accepting to be my mentor for my doctoral work and John Parsch to be in my thesis committee. Their foresight and advice regarding scientific and technical aspects of work were insightful and helped me avoid many pitfalls. The discussions during meetings were always constructive and encouraging.

During the course of my PhD, I was fortunate to have made friends outside work. I specially thank Daniel, Ramona, Chris and Rahul. I found Daniel to be both knowledgeable and wise and his advice from technical to existential issues had a strong influence in my life. He is also a true embodiment of the proverb 'A friend is need is a friend indeed'. Ramona and Chris brought so much light in my life. I realized that the German ideal of quality v.s quantity also applies to personal sphere and friendships. I looked up to and admired the ability of Rahul to tackle problems and adapt to any situation and learned a lot from him.

Finally, I can not put in words my gratitude to my parents and my brother. Your care, concern and sacrifice for me has been unconditional. All that is good in me is only because of you.

Bibliography

[1] Keith L. Adams, Richard Cronn, Ryan Percifield, and Jonathan F. Wendel. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8):4649–4654, April 2003.

[2] Jean-Marc Aury, Olivier Jaillon, Laurent Duret, Benjamin Noel, Claire Jubin, Betina M. Porcel, Béatrice Ségurens, Vincent Daubin, Véronique Anthouard, Nathalie Aiach, Olivier Arnaiz, Alain Billaut, Janine Beisson, Isabelle Blanc, Khaled Bouhouche, Francisco Câmara, Sandra Duharcourt, Roderic Guigo, Delphine Gogendeau, Michael Katinka, Anne-Marie Keller, Roland Kissmehl, Catherine Klotz, France Koll, Anne Le Mouël, Gersende Lepère, Sophie Malinsky, Mariusz Nowacki, Jacek K. Nowak, Helmut Plattner, Julie Poulain, Françoise Ruiz, Vincent Serrano, Marek Zagulski, Philippe Dessen, Mireille Bétermier, Jean Weissenbach, Claude Scarpelli, Vincent Schächter, Linda Sperling, Eric Meyer, Jean Cohen, and Patrick Wincker. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature*, 444(7116):171–178, November 2006.

[3] Maite G. Barrón, Anna-Sophie Fiston-Lavier, Dmitri A. Petrov, and Josefa González. Population Genomics of Transposable Elements in Drosophila. *Annual Review of Genetics*, 48(1):561–581, 2014.

[4] Carolina Bartolomé, Xulio Maside, and Brian Charlesworth. On the Abundance and Distribution of Transposable Elements in the Genome of Drosophila melanogaster. *Molecular Biology and Evolution*, 19(6):926–937, June 2002.

[5] Eva Bauer, Matthieu Falque, Hildrun Walter, Cyril Bauland, Christian Camisan, Laura Campo, Nina Meyer, Nicolas Ranc, Renaud Rincent, Wolfgang Schipprack, Thomas Altmann, Pascal Flament, Albrecht E. Melchinger, Monica Menz, Jesús Moreno-González, Milena Ouzunova, Pedro Revilla, Alain Charcosset, Olivier C. Martin, and Chris-Carolin Schön. Intraspecific variation of recombination rate in maize. *Genome Biology*, 14:R103, 2013.

[6] G. W. Beadle. Teosinte and the Origin of Maize. *Journal of Heredity*, 30(6):245–247, June 1939.

[7] Timothy M. Beissinger, Li Wang, Kate Crosby, Arun Durvasula, Matthew B. Hufford, and Jeffrey Ross-Ibarra. Recent demography drives changes in linked selection across the maize genome. *Nature Plants*, 2:16084, 2016.

[8] Michaël Bekaert, Patrick P. Edger, J. Chris Pires, and Gavin C. Conant. Two-Phase Resolution of Polyploidy in the Arabidopsis Metabolic Network Gives Rise to Relative and Absolute Dosage Constraints. *The Plant Cell Online*, 23(5):1719–1728, May 2011. 00027.

[9] Oliver Bell, Vijay K. Tiwari, Nicolas H. Thomä, and Dirk Schübeler. Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, 12(8):554–564, August 2011.

[10] A. Belyayev. Bursts of transposable elements as an evolutionary driving force. *Journal of Evolutionary Biology*, 27(12):2573–2584, December 2014.

[11] Jeffrey L. Bennetzen and Hao Wang. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology*, 65(1):505–530, 2014.

[12] Jeffrey L. Bennetzin and Michael Freeling. Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends in Genetics*, 9(8):259–261, August 1993.

[13] James A. Birchler and Reiner A. Veitia. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proceedings of the National Academy of Sciences*, 109(37):14746–14753, September 2012.

[14] Guillaume Blanc and Kenneth H. Wolfe. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *The Plant Cell*, 16(7):1667–1678, July 2004.

[15] Alexandros Bousios, Evangelia Minga, Nikoleta Kalitsou, Maria Pantermali, Aphrodite Tsaballa, and Nikos Darzentas. MASiVEdb: the Sirevirus Plant Retrotransposon Database. *BMC Genomics*, 13:158, 2012.

[16] Thibaud S. Boutin, Arnaud Le Rouzic, and Pierre Capy. How does selfing affect the dynamics of selfish transposable elements? *Mobile DNA*, 3:5, 2012.

[17] Stephan Brunner, Kevin Fengler, Michele Morgante, Scott Tingey, and Antoni Rafalski. Evolution of DNA Sequence Nonhomologies among Maize Inbreds. *The Plant Cell*, 17(2):343–360, February 2005.

[18] E. S. Buckler, J. M. Thornsberry, and S. Kresovich. Molecular diversity, structure and domestication of grasses. *Genetical Research*, 77(3):213–218, June 2001.

[19] Richard J. A. Buggs, Srikar Chamala, Wei Wu, Lu Gao, Gregory D. May, Patrick S. Schnable, Douglas E. Soltis, Pamela S. Soltis, and W. Brad Barbazuk. Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular Ecology*, 19 Suppl 1:132–146, March 2010.

[20] Robert Bukowski, Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, Dawen Xu, Bicheng Yang, Chuanxiao Xie, Longjiang Fan, Shibin Gao, Xun Xu, Gengyun Zhang, Yingrui Li, Yinping Jiao, John Doebley, Jeffrey Ross-Ibarra, Vince Buffalo, Cinta M. Romay, Edward S. Buckler, Yunbi Xu, Jinsheng Lai, Doreen Ware, and Qi Sun. Construction of the third generation Zea mays haplotype map. *bioRxiv*, page 026963, September 2016.

[21] Robert Bukowski, Xiaosen Guo, Yanli Lu, Cheng Zou, Bing He, Zhengqin Rong, Bo Wang, Dawen Xu, Bicheng Yang, Chuanxiao Xie, Longjiang Fan, Shibin Gao, Xun Xu, Gengyun Zhang, Yingrui Li, Yinping Jiao, John Doebley, Jeffrey Ross-Ibarra, Vince Buffalo, Edward S. Buckler, Yunbi Xu, Jinsheng Lai, Doreen Ware, and Qi Sun. Construction of the third generation Zea mays haplotype map. *bioRxiv*, page 026963, September 2015.

[22] José L. Campos, Daniel L. Halligan, Penelope R. Haddrill, and Brian Charlesworth. The relation between recombination rate and patterns of molecular evolution and variation in Drosophila melanogaster. *Molecular Biology and Evolution*, page msu056, January 2014.

[23] Cristian Cañestro. Two Rounds of Whole-Genome Duplication: Evidence and Impact on the Evolution of Vertebrate Innovations. In Pamela S. Soltis and Douglas E. Soltis, editors, *Polyploidy and Genome Evolution*, pages 309–339. Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-31442-1_16.

[24] Liran Carmel and Eugene V. Koonin. A Universal Nonmonotonic Relationship between Gene Compactness and Expression Levels in Multicellular Eukaryotes. *Genome Biology and Evolution*, 1:382–390, 2009.

[25] Piero Carninci, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A. M. Semple, Martin S. Taylor, Pär G. Engström, Martin C. Frith, Alistair R. R. Forrest, Wynand B. Alkema, Sin Lam Tan, Charles Plessy, Rimantas Kodzius, Timothy Ravasi, Takeya Kasukawa, Shiro Fukuda, Mutsumi Kanamori-Katayama, Yayoi Kitazume, Hideya Kawaji, Chikatoshi Kai, Mari Nakamura, Hideaki Konno, Kenji Nakano, Salim Mottagui-Tabar, Peter Arner, Alessandra Chesi, Stefano Gustincich, Francesca Persichetti, Harukazu Suzuki, Sean M. Grimmond, Christine A. Wells, Valerio Orlando, Claes Wahlestedt, Edison T. Liu, Matthias Harbers, Jun Kawai, Vladimir B. Bajic, David A. Hume, and Yoshihide Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38(6):626–635, June 2006.

[26] Colin R. Cavanagh, Shiaoman Chao, Shichen Wang, Bevan Emma Huang, Stuart Stephen, Seifollah Kiani, Kerrie Forrest, Cyrille Saintenac, Gina L. Brown-Guedira, Alina Akhunova, Deven See, Guihua Bai, Michael Pumphrey, Luxmi Tomar, Debbie Wong, Stephan Kong, Matthew Reynolds, Marta Lopez da Silva, Harold Bockelman, Luther Talbert, James A. Anderson, Susanne Dreisigacker, Stephen Baenziger, Arron Carter, Viktor Korzun, Peter Laurent Morrell, Jorge Dubcovsky, Matthew K. Morell, Mark E. Sorrells, Matthew J. Hayden, and Eduard Akhunov. Genome-wide comparative diversity uncovers multiple targets of selection for

improvement in hexaploid wheat landraces and cultivars. *Proceedings of the National Academy of Sciences*, 110(20):8057–8062, May 2013.

[27] Frédéric JJ Chain, Jonathan Dushoff, and Ben J. Evans. The odds of duplicate gene persistence after polyploidization. *BMC Genomics*, 12(1):599, December 2011.

[28] Peter L. Chang, Brian P. Dilkes, Michelle McMahon, Luca Comai, and Sergey V. Nuzhdin. Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. *Genome Biology*, 11(12):R125, December 2010.

[29] B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, August 1993.

[30] Brian Charlesworth. Molecular population genomics: a short history. *Genetics Research*, 92(5-6):397–411, December 2010.

[31] Bhupendra Chaudhary, Lex Flagel, Robert M. Stupar, Joshua A. Udall, Neetu Verma, Nathan M. Springer, and Jonathan F. Wendel. Reciprocal Silencing, Transcriptional Bias and Functional Divergence of Homeologs in Polyploid Cotton (Gossypium). *Genetics*, 182(2):503–517, June 2009.

[32] Eric Ch Chen and David Sankoff. Gene expression and fractionation resistance. *BMC genomics*, 15(Suppl 6):S19, October 2014.

[33] Hua Chen, Nick Patterson, and David Reich. Population differentiation as a test for selective sweeps. *Genome Research*, 20(3):393–402, March 2010.

[34] Benoît Chénais, Aurore Caruso, Sophie Hiard, and Nathalie Casse. The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*, 509(1):7–15, November 2012.

[35] Feng Cheng, Jian Wu, Lu Fang, Silong Sun, Bo Liu, Ke Lin, Guusje Bonnema, and Xiaowu Wang. Biased Gene Fractionation and Dominant Gene Expression among the Subgenomes of Brassica rapa. *PLoS ONE*, 7(5):e36442, May 2012.

[36] Jer-Ming Chia, Chi Song, Peter J. Bradbury, Denise Costich, Natalia de Leon, John Doebley, Robert J. Elshire, Brandon Gaut, Laura Geller, Jeffrey C. Glaubitz, Michael Gore, Kate E. Guill, Jim Holland, Matthew B. Hufford, Jinsheng Lai, Meng Li, Xin Liu, Yanli Lu, Richard McCombie, Rebecca Nelson, Jesse Poland, Boddupalli M. Prasanna, Tanja Pyhäjärvi, Tingzhao Rong, Rajandeep S. Sekhon, Qi Sun, Maud I. Tenaillon, Feng Tian, Jun Wang, Xun Xu, Zhiwu Zhang, Shawn M. Kaeppler, Jeffrey Ross-Ibarra, Michael D. McMullen, Edward S. Buckler, Gengyun Zhang, Yunbi Xu, and Doreen Ware. Maize HapMap2 identifies extant variation from a genome in flux. *Nature Genetics*, 44(7):803–807, July 2012.

[37] Li-Wei Chiu, Xiangjun Zhou, Sarah Burke, Xianli Wu, Ronald L. Prior, and Li Li. The Purple Cauliflower Arises from Activation of a MYB Transcription Factor. *Plant Physiology*, 154(3):1470–1480, November 2010.

[38] Kyuha Choi, Carsten Reinhard, Heïdi Serra, Piotr A. Ziolkowski, Charles J. Underwood, Xiaohui Zhao, Thomas J. Hardcastle, Nataliya E. Yelina, Catherine Griffin, Matthew Jackson, Christine Mézard, Gil McVean, Gregory P. Copenhaver, and Ian R. Henderson. Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes. *PLOS Genet*, 12(7):e1006179, July 2016.

[39] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, June 2012.

[40] Michael T. Clegg and Mary L. Durbin. Flower color variation: A model for the experimental study of evolution. *Proceedings of the National Academy of Sciences*, 97(13):7016–7023, June 2000.

[41] Michela Curradi, Annalisa Izzo, Gianfranco Badaracco, and Nicoletta Landsberger. Molecular Mechanisms of Gene Silencing Mediated by DNA Methylation. *Molecular and Cellular Biology*, 22(9):3157–3173, May 2002.

[42] Sudhansu Dash, John Van Hemert, Lu Hong, Roger P. Wise, and Julie A. Dickerson. PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Research*, 40(Database issue):D1194–D1201, January 2012.

[43] J. Doebley, W. Renfroe, and A. Blanton. Restriction Site Variation in the Zea Chloroplast Genome. *Genetics*, 117(1):139–147, September 1987.

[44] J Doebley, A Stec, J Wendel, and M Edwards. Genetic and morphological analysis of a maize-teosinte F2 population: implications for the origin of maize. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9888–9892, December 1990.

[45] John Doebley. Isozymic Evidence and the Evolution of Crop Plants. In Douglas E. Soltis, Pamela S. Soltis, and Theodore R. Dudley, editors, *Isozymes in Plant Biology*, pages 165–191. Springer Netherlands, 1989. DOI: 10.1007/978-94-009-1840-5_9.

[46] John Doebley. The Genetics of Maize Evolution. *Annual Review of Genetics*, 38(1):37–59, 2004.

[47] John F. Doebley, M. Goodman, and Charles W. Stuber. Isoenzymatic Variation in Zea (Gramineae). *Systematic Botany*, 9(2):203–218, 1984.

[48] Elie S. Dolgin and Brian Charlesworth. The Effects of Recombination Rate on the Distribution and Abundance of Transposable Elements. *Genetics*, 178(4):2169–2177, April 2008.

[49] Yang Dong and Yin-Zheng Wang. Seed shattering: from models to crops. *Frontiers in Plant Science*, 6, June 2015.

[50] Hugo K. Dooner and Limei He. Maize Genome Structure Variation: Interplay between Retrotransposon Polymorphisms and Genic Recombination. *The Plant Cell*, 20(2):249–258, February 2008.

[51] Zhou Du, Xin Zhou, Yi Ling, Zhenhai Zhang, and Zhen Su. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research*, 38(suppl 2):W64–W70, July 2010.

[52] Jill M. Duarte, Liying Cui, P. Kerr Wall, Qing Zhang, Xiaohong Zhang, Jim Leebens-Mack, Hong Ma, Naomi Altman, and Claude W. dePamphilis. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Molecular Biology and Evolution*, 23(2):469–478, February 2006.

[53] Nancy A. Eckardt. Grass Genome Evolution. *The Plant Cell*, 20(1):3–4, January 2008.

[54] Michael D. Edgerton. Increasing Crop Productivity to Meet Global Needs for Feed, Food, and Fuel. *Plant Physiology*, 149(1):7–13, January 2009.

[55] Steve R. Eichten, Ruth A. Swanson-Wagner, James C. Schnable, Amanda J. Waters, Peter J. Hermanson, Sanzhen Liu, Cheng-Ting Yeh, Yi Jia, Karla Gendler, Michael Freeling, Patrick S. Schnable, Matthew W. Vaughn, and Nathan M. Springer. Heritable Epigenetic Variation among Maize Inbreds. *PLOS Genet*, 7(11):e1002372, November 2011.

[56] Adam D. Ewing. Transposable element detection from whole genome sequence data. *Mobile DNA*, 6:24, 2015.

[57] Adam Eyre-Walker, Rebecca L. Gaut, Holly Hilton, Dawn L. Feldman, and Brandon S. Gaut. Investigation of the bottleneck leading to the domestication of maize. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8):4441–4446, April 1998.

[58] Adam Eyre-Walker and Peter D. Keightley. The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8(8):610–618, August 2007. 00331.

[59] Adam Eyre-Walker and Peter D. Keightley. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change. *Molecular Biology and Evolution*, 26(9):2097–2108, September 2009.

[60] Domènec Farré, Nicolás Bellora, Loris Mularoni, Xavier Messeguer, and M. Mar Albà. Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biology*, 8:R140, 2007.

[61] Nina V. Fedoroff. The Discovery of Transposition. In Nina V. Fedoroff, editor, *Plant Transposons and Genome Dynamics in Evolution*, pages 3–13. Wiley-Blackwell, 2013.

[62] Justin A. Fincher, Daniel L. Vera, Diana D. Hughes, Karen M. McGinnis, Jonathan H. Dennis, and Hank W. Bass. Genome-Wide Prediction of Nucleosome Occupancy in Maize Reveals Plant Chromatin Structural Features at Genes and Other Elements at Multiple Scales. *Plant Physiology*, 162(2):1127–1141, June 2013.

[63] David J. Finnegan. Eukaryotic transposable elements and genome evolution. *Trends in Genetics*, 5:103–107, January 1989.

[64] Lex E. Flagel and Jonathan F. Wendel. Gene duplication and evolutionary novelty in plants. *New Phytologist*, 183(3):557–564, August 2009.

[65] Lex E. Flagel and Jonathan F. Wendel. Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *New Phytologist*, 186(1):184–193, 2010.

[66] Sherry A. Flint-Garcia. Genetics and Consequences of Crop Domestication. *Journal of Agricultural and Food Chemistry*, 61(35):8267–8276, September 2013.

[67] Sherry A. Flint-Garcia, Jeffry M. Thornsberry, Edward S, and and Buckler IV. Structure of Linkage Disequilibrium in Plants. *Annual Review of Plant Biology*, 54(1):357–374, 2003.

[68] Allan Force, Michael Lynch, F. Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics*, 151(4):1531–1545, April 1999.

[69] Jakob Fredslund, Lene H. Madsen, Birgit K. Hougaard, Niels Sandal, Jens Stougaard, David Bertioli, and Leif Schauser. GeMprospector—online design of cross-species genetic marker candidates in legumes and grasses. *Nucleic Acids Research*, 34(Web Server issue):W670–W675, July 2006.

[70] Michael Freeling, Margaret R Woodhouse, Shabarinath Subramaniam, Gina Turco, Damon Lisch, and James C Schnable. Fractionation mutagenesis and similar consequences of mechanisms removing dispensable or less-expressed DNA in plants. *Current Opinion in Plant Biology*, 15(2):131–139, April 2012.

[71] Huihua Fu and Hugo K. Dooner. Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences*, 99(14):9573–9578, July 2002.

[72] Olivier Garsmeur, James C. Schnable, Ana Almeida, Cyril Jourda, Angélique D'Hont, and Michael Freeling. Two Evolutionarily Distinct Classes of Paleopolyploidy. *Molecular Biology and Evolution*, 31(2):448–454, February 2014.

[73] Brandon S. Gaut and John F. Doebley. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences*, 94(13):6809–6814, June 1997.

[74] Mary Gehring. Prodigious plant methylomes. *Genome Biology*, 17:197, 2016.

[75] Jonathan I. Gent, Nathanael A. Ellis, Lin Guo, Alex E. Harkess, Yingyin Yao, Xiaoyu Zhang, and R. Kelly Dawe. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome Research*, 23(4):628–637, April 2013.

[76] Paul Gepts. Crop Domestication as a Long-Term Selection Experiment. In Jules Janick, editor, *Plant Breeding Reviews*, pages 1–44. John Wiley & Sons, Inc., 2003.

[77] Avazeh T. Ghanbarian and Laurence D. Hurst. Neighboring genes show correlated evolution in gene expression. *Molecular Biology and Evolution*, page msv053, March 2015.

[78] Sylvain Glémin and Thomas Bataillon. A comparative view of the evolution of grasses under domestication. *The New Phytologist*, 183(2):273–290, 2009.

[79] Isabel Gordo and Brian Charlesworth. Genetic linkage and molecular evolution. *Current Biology*, 11(17):R684–R686, September 2001.

[80] Toni I. Gossmann, Bao-Hua Song, Aaron J. Windsor, Thomas Mitchell-Olds, Christopher J. Dixon, Maxim V. Kapralov, Dmitry A. Filatov, and Adam Eyre-Walker. Genome Wide Analyses Reveal Little Evidence for Adaptive Evolution in Many Plant Species. *Molecular Biology and Evolution*, 27(8):1822–1832, August 2010.

[81] Anthony J. F. Griffiths, editor. *An introduction to genetic analysis*. W.H. Freeman, New York, 7th ed edition, 2000.

[82] Liza Gross. Transposon Silencing Keeps Jumping Genes in Their Place. *PLoS Biology*, 4(10), October 2006.

[83] C. E. Grover, J. P. Gallagher, E. P. Szadkowski, M. J. Yoo, L. E. Flagel, and J. F. Wendel. Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist*, 196(4):966–971, December 2012.

[84] Corrinne E. Grover and Jonathan F. Wendel. Recent Insights into Mechanisms of Genome Size Change in Plants. *Journal of Botany*, 2010:e382732, May 2010.

[85] Zhenglong Gu, Dan Nicolae, Henry H-S. Lu, and Wen-Hsiung Li. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics*, 18(12):609–613, December 2002.

[86] Zhenglong Gu, Lars M. Steinmetz, Xun Gu, Curt Scharfe, Ronald W. Davis, and Wen-Hsiung Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(6918):63–66, January 2003.

[87] Rama R. Gullapalli, Ketaki V. Desai, Lucas Santana-Santos, Jeffrey A. Kant, and Michael J. Becich. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *Journal of Pathology Informatics*, 3, October 2012.

[88] Karl Hammer. Das Domestikationssyndrom. *Die Kulturpflanze*, 32(1):11–34, June 1984.

[89] Jack R. Harlan, J. M. J. de Wet, and E. Glen Price. Comparative Evolution of Cereals. *Evolution*, 27(2):311–325, 1973.

[90] Daniel L. Hartl and Andrew G. Clark. *Principles of population genetics*. Sinauer Associates, Sunderland, Mass, 4th ed edition, 2007.

[91] Xionglei He and Jianzhi Zhang. Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics*, 169(2):1157–1164, February 2005.

[92] Robert J. Henry. Next-generation sequencing for understanding and accelerating crop domestication. *Briefings in Functional Genomics*, 11(1):51–56, January 2012.

[93] Joachim Hermisson and Pleuni S. Pennings. Soft Sweeps. *Genetics*, 169(4):2335–2352, April 2005.

[94] W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genetical Research*, 8(3):269–294, December 1966.

[95] Holly Hilton and Brandon S. Gaut. Speciation and Domestication in Maize and Its Wild Relatives: Evidence From the Globulin-1 Gene. *Genetics*, 150(2):863–872, October 1998.

[96] Cory D. Hirsch and Nathan M. Springer. Transposable element influences on gene expression in plants. *Biochimica Et Biophysica Acta*, May 2016.

[97] Federico G. Hoffmann, Juan C. Opazo, and Jay F. Storz. Whole-Genome Duplications Spurred the Functional Diversification of the Globin Gene Superfamily in Vertebrates. *Molecular Biology and Evolution*, page msr207, September 2011.

[98] Jesse D. Hollister and Brandon S. Gaut. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Research*, 19(8):1419–1428, August 2009.

[99] Jesse D. Hollister, Lisa M. Smith, Ya-Long Guo, Felix Ott, Detlef Weigel, and Brandon S. Gaut. Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. *Proceedings of the National Academy of Sciences*, 108(6):2322–2327, February 2011.

[100] Roger A. Hoskins, Jane M. Landolin, James B. Brown, Jeremy E. Sandler, Hazuki Takahashi, Timo Lassmann, Charles Yu, Benjamin W. Booth, Dayu Zhang, Kenneth H. Wan, Li Yang, Nathan Boley, Justen Andrews, Thomas C. Kaufman, Brenton R. Graveley, Peter J. Bickel, Piero Carninci, Joseph W. Carlson, and Susan E. Celniker. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Research*, 21(2):182–192, February 2011.

[101] Josh Hough, Robert J. Williamson, and Stephen I. Wright. Patterns of Selection in Plant Genomes. *Annual Review of Ecology, Evolution, and Systematics*, 44(1):31–49, 2013.

[102] Xuehui Huang, Nori Kurata, Xinghua Wei, Zi-Xuan Wang, Ahong Wang, Qiang Zhao, Yan Zhao, Kunyan Liu, Hengyun Lu, Wenjun Li, Yunli Guo, Yiqi Lu, Congcong Zhou, Danlin Fan, Qijun Weng, Chuanrang Zhu, Tao Huang, Lei Zhang, Yongchun Wang, Lei Feng, Hiroyasu Furuumi, Takahiko Kubo, Toshie Miyabayashi, Xiaoping Yuan, Qun Xu, Guojun Dong, Qilin Zhan, Canyang Li, Asao Fujiyama, Atsushi Toyoda, Tingting Lu, Qi Feng, Qian Qian, Jiayang Li, and Bin Han. A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 490(7421):497–501, October 2012.

[103] Richard R. Hudson and Norman L. Kaplan. Statistical Properties of the Number of Recombination Events in the History of a Sample of DNA Sequences. *Genetics*, 111(1):147–164, September 1985.

[104] Matthew B. Hufford, Xun Xu, Joost van Heerwaarden, Tanja Pyhäjärvi, Jer-Ming Chia, Reed A. Cartwright, Robert J. Elshire, Jeffrey C. Glaubitz, Kate E. Guill, Shawn M. Kaeppler, Jinsheng Lai, Peter L. Morrell, Laura M. Shannon, Chi Song, Nathan M. Springer, Ruth A. Swanson-Wagner, Peter Tiffin, Jun Wang, Gengyun Zhang, John Doebley, Michael D. McMullen, Doreen Ware, Edward S. Buckler, Shuang Yang, and Jeffrey Ross-Ibarra. Comparative population genomics of maize domestication and improvement. *Nature Genetics*, 44(7):808–811, July 2012.

[105] Thomas E. Hughes, Jane A. Langdale, and Steven Kelly. The impact of widespread regulatory neofunctionalization on homeolog gene evolution following whole-genome duplication in maize. *Genome Research*, 24(8):1348–1355, August 2014.

[106] Lukasz Huminiecki and Gavin C. Conant. Polyploidy and the Evolution of Complex Traits. *International Journal of Evolutionary Biology*, 2012:e292068, July 2012.

[107] Harriet V. Hunt, Kay Denyer, Len C. Packman, Martin K. Jones, and Christopher J. Howe. Molecular Basis of the Waxy Endosperm Starch Phenotype in Broomcorn Millet (Panicum miliaceum L.). *Molecular Biology and Evolution*, 27(7):1478–1494, July 2010.

[108] H. H. Iltis. From teosinte to maize: the catastrophic sexual transmutation. *Science (New York, N.Y.)*, 222(4626):886–894, November 1983.

[109] Hideki Innan and Yuseob Kim. Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences of the United States of America*, 101(29):10667–10672, July 2004.

[110] Hisakazu Iwama and Takashi Gojobori. Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49):17156–17161, December 2004.

[111] Yinping Jiao, Hainan Zhao, Longhui Ren, Weibin Song, Biao Zeng, Jinjie Guo, Baobao Wang, Zhipeng Liu, Jing Chen, Wei Li, Mei Zhang, Shaojun Xie, and Jinsheng Lai. Genome-wide genetic changes during modern breeding of maize. *Nature Genetics*, 44(7):812–815, July 2012.

[112] Jinpu Jin, He Zhang, Lei Kong, Ge Gao, and Jingchu Luo. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research*, 42(D1):D1182–D1187, January 2014.

[113] Ping Jin, Sheng Qin, Xi Chen, Yumei Song, Jesse Li-Ling, Xiaofeng Xu, and Fei Ma. Evolutionary rate of human tissue-specific genes are related with transposable element insertions. *Genetica*, 140(10-12):513–523, December 2012.

[114] Deborah A. Johnson and Michael A. Thomas. The Monosaccharide Transporter Gene Family in Arabidopsis and Rice: A History of Duplications, Adaptive Evolution, and Functional Divergence. *Molecular Biology and Evolution*, 24(11):2412–2423, November 2007.

[115] Tamar Juven-Gershon and James T. Kadonaga. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology*, 339(2):225–229, March 2010.

[116] Tina Kabelitz and Isabel Bäurle. Get the jump – Do 3'UTRs protect transposable elements from silencing? *Mobile Genetic Elements*, 5(4):51–54, July 2015.

[117] Khalil Kashkush, Moshe Feldman, and Avraham A. Levy. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genetics*, 33(1):102–106, January 2003.

[118] Karin S. Kassahn, Vinh T. Dang, Simon J. Wilkins, Andrew C. Perkins, and Mark A. Ragan. Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Research*, 19(8):1404–1418, August 2009.

[119] Elizabeth A. Kellogg and C. Robin Buell. Splendor in the Grasses. *Plant Physiology*, 149(1):1–3, January 2009.

[120] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, June 2002.

[121] Yuseob Kim and Rasmus Nielsen. Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3):1513–1524, July 2004.

[122] Motoo Kimura. On the Probability of Fixation of Mutant Genes in a Population. *Genetics*, 47(6):713–719, June 1962.

[123] Motoo Kimura. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. *Genetics*, 61(4):893–903, April 1969.

[124] Rhoda J. Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011:bar030, January 2011.

[125] Thomas J. Y. Kono, Fengli Fu, Mohsen Mohammadi, Paul J. Hoffman, Chaochih Liu, Robert M. Stupar, Kevin P. Smith, Peter Tiffin, Justin C. Fay, and Peter L. Morrell. The role of deleterious substitutions in crop genomes. *bioRxiv*, page 033175, May 2016.

[126] Patrick J. Krysan, Jeffery C. Young, and Michael R. Sussman. T-DNA as an Insertional Mutagen in Arabidopsis. *The Plant Cell*, 11(12):2283–2290, December 1999.

[127] Nadezda Kryuchkova-Mostacci and Marc Robinson-Rechavi. A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, page bbw008, February 2016.

[128] Amar Kumar and and Jeffrey L. Bennetzen. Plant Retrotransposons. *Annual Review of Genetics*, 33(1):479–532, 1999.

[129] Shailesh K. Lal, Nikolaos Georgelis, and Curtis L. Hannah. Helitrons: Their Impact on Maize Genome Evolution and Diversity. In Jeffrey L. Bennetzen and Sarah Hake, editors, *Handbook of Maize*, pages 329–339. Springer New York, 2009. DOI: 10.1007/978-0-387-77863-1_16.

[130] Amanda K. Lane, Chad E. Niederhuth, Lexiang Ji, and Robert J. Schmitz. pENCODE: A Plant Encyclopedia of DNA Elements. *Annual Review of Genetics*, 48(1):49–70, 2014.

[131] Richard J. Langham, Justine Walsh, Molly Dunn, Cynthia Ko, Stephen A. Goff, and Michael Freeling. Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics*, 166(2):935–945, February 2004.

[132] Carolyn J. Lawrence, Qunfeng Dong, Mary L. Polacco, Trent E. Seigfried, and Volker Brendel. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Research*, 32(Database issue):D393–397, January 2004.

[133] Mark J. Lawson and Liqing Zhang. Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region. *Gene*, 407(1–2):54–62, January 2008.

[134] Elizabeth A. Lee and William F. Tracy. Modern Maize Breeding. In Jeffrey L. Bennetzen and Sarah Hake, editors, *Handbook of Maize*, pages 141–160. Springer New York, 2009. DOI: 10.1007/978-0-387-77863-1_7.

[135] Soohyun Lee, Isaac Kohane, and Simon Kasif. Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics*, 6:168, 2005.

[136] Tae-Ho Lee, Haibao Tang, Xiyin Wang, and Andrew H. Paterson. PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Research*, 41(D1):D1152–D1158, January 2013.

[137] E. Lerat. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, 104(6):520–533, June 2010.

[138] Lin Li, Roman Briskine, Robert Schaefer, Patrick S. Schnable, Chad L. Myers, Lex E. Flagel, Nathan M. Springer, and Gary J. Muehlbauer. Co-expression network analysis of duplicate genes in maize (Zea mays L.) reveals no subgenome bias. *BMC Genomics*, 17:875, 2016.

[139] Qing Li, Jonathan I. Gent, Greg Zynda, Jawon Song, Irina Makarevitch, Cory D. Hirsch, Candice N. Hirsch, R. Kelly Dawe, Thelma F. Madzima, Karen M. McGinnis, Damon Lisch, Robert J. Schmitz, Matthew W. Vaughn, and Nathan M. Springer. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. *Proceedings of the National Academy of Sciences*, 112(47):14728–14733, November 2015.

[140] Xin Li, Jingde Zhu, Fengyi Hu, Song Ge, Mingzhi Ye, Hui Xiang, Guojie Zhang, Xiaoming Zheng, Hongyu Zhang, Shilai Zhang, Qiong Li, Ruibang Luo, Chang Yu, Jian Yu, Jingfeng Sun, Xiaoyu Zou, Xiaofeng Cao, Xianfa Xie, Jun Wang, and Wen Wang. Single-base resolution maps of cultivated and wild rice methylomes and regulatory roles of DNA methylation in plant gene expression. *BMC Genomics*, 13:300, 2012.

[141] Damon Lisch. Epigenetic Regulation of Transposable Elements in Plants. *Annual Review of Plant Biology*, 60(1):43–66, 2009.

[142] Damon Lisch. How important are transposons for plant evolution? *Nature Reviews Genetics*, 14(1):49–61, January 2013.

[143] Damon Lisch and Ning Jiang. Mutator and MULE transposons. In Jeffrey L. Bennetzen and Sarah Hake, editors, *Handbook of Maize*, pages 277–306. Springer New York, 2009. DOI: 10.1007/978-0-387-77863-1_14.

[144] Zhenlan Liu and Keith L. Adams. Expression Partitioning between Genes Duplicated by Polyploidy under Abiotic Stress and during Organ Development. *Current Biology*, 17(19):1669–1674, October 2007.

[145] Steven Lockton and Brandon S. Gaut. Plant conserved non-coding sequences and paralogue evolution. *Trends in Genetics*, 21(1):60–65, January 2005.

[146] Steven Lockton, Jeffrey Ross-Ibarra, and Brandon S. Gaut. Demography and weak selection drive patterns of transposable element diversity in natural populations of Arabidopsis lyrata. *Proceedings of the National Academy of Sciences*, 105(37):13965–13970, September 2008.

[147] Jian Lu, Tian Tang, Hua Tang, Jianzi Huang, Suhua Shi, and Chung-I. Wu. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends in Genetics*, 22(3):126–131, March 2006.

[148] Gordon Luikart, Phillip R. England, David Tallmon, Steve Jordan, and Pierre Taberlet. The power and promise of population genomics: from genotyping to genome typing. *Nature Reviews Genetics*, 4(12):981–994, December 2003.

[149] Eric Lyons, Brent Pedersen, Josh Kane, and Michael Freeling. The Value of Nonmodel Genomes and an Example Using SynMap Within CoGe to Dissect the Hexaploidy that Predates the Rosids. *Tropical Plant Biology*, 1(3-4):181–190, October 2008.

[150] Steven Maere, Stefanie De Bodt, Jeroen Raes, Tineke Casneuf, Marc Van Montagu, Martin Kuiper, and Yves Van de Peer. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5454–5459, April 2005.

[151] Irina Makarevitch, Amanda J. Waters, Patrick T. West, Michelle Stitzer, Candice N. Hirsch, Jeffrey Ross-Ibarra, and Nathan M. Springer. Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLOS Genet*, 11(1):e1004915, January 2015.

[152] Hude Mao, Hongwei Wang, Shengxue Liu, Zhigang Li, Xiaohong Yang, Jianbing Yan, Jiansheng Li, Lam-Son Phan Tran, and Feng Qin. A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. *Nature Communications*, 6:8326, September 2015.

[153] Arturo Marí-Ordóñez, Antonin Marchais, Mathilde Etcheverry, Antoine Martin, Vincent Colot, and Olivier Voinnet. Reconstructing de novo silencing of an active plant retrotransposon. *Nature Genetics*, 45(9):1029–1039, September 2013.

[154] Yoshihiro Matsuoka, Yves Vigouroux, Major M. Goodman, Jesus Sanchez G, Edward Buckler, and John Doebley. A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences*, 99(9):6080–6084, April 2002.

[155] John Maynard and John Haigh. The hitch-hiking effect of a favourable gene. *Genetics Research*, 89(5-6):391–403, December 2007.

[156] Casey L. McGrath, Jean-Francois Gout, Parul Johri, Thomas G. Doak, and Michael Lynch. Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research*, 24(10):1665–1675, October 2014.

[157] Gil McVean. The Structure of Linkage Disequilibrium Around a Selective Sweep. *Genetics*, 175(3):1395–1406, March 2007.

[158] María Katherine Mejía-Guerra, Wei Li, Narmer F. Galeano, Mabel Vidal, John Gray, Andrea I. Doseff, and Erich Grotewold. Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with Predominance of Sharp Transcription Initiation Sites. *The Plant Cell*, 27(12):3309–3320, December 2015.

[159] Sofiane Mezmouk and Jeffrey Ross-Ibarra. The Pattern and Distribution of Deleterious Mutations in Maize. *G3: Genes|Genomes|Genetics*, 4(1):163–171, January 2014.

[160] Yoko Mizuta, Yoshiaki Harushima, and Nori Kurata. Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. *Proceedings of the National Academy of Sciences*, 107(47):20417–20422, November 2010.

[161] G. Moore, K. M. Devos, Z. Wang, and M. D. Gale. Cereal Genome Evolution: Grasses, line up and form a circle. *Current Biology*, 5(7):737–739, July 1995.

[162] Hussein Mortada, Cristina Vieira, and Emmanuelle Lerat. Genes Devoid of Full-Length Transposable Element Insertions are Involved in Development and in the Regulation of Transcription in Human and Closely Related Species. *Journal of Molecular Evolution*, 71(3):180–191, August 2010.

[163] Tobias Mourier and Eske Willerslev. Does Selection against Transcriptional Interference Shape Retroelement-Free Regions in Mammalian Genomes? *PLOS ONE*, 3(11):e3760, November 2008.

[164] Florent Murat, Rongzhi Zhang, Sébastien Guizard, Raphael Flores, Alix Armero, Caroline Pont, Delphine Steinbach, Hadi Quesneville, Richard Cooke, and Jerome Salse. Shared Subgenome Dominance Following Polyploidization Explains Grass Genome Evolutionary Plasticity from a Seven Protochromosome Ancestor with 16k Protogenes. *Genome Biology and Evolution*, 6(1):12–33, January 2014.

[165] Simon R. Myers and Robert C. Griffiths. Bounds on the Minimum Number of Recombination Events in a Sample History. *Genetics*, 163(1):375–394, January 2003.

[166] Ken Naito, Feng Zhang, Takuji Tsukiyama, Hiroki Saito, C. Nathan Hancock, Aaron O. Richardson, Yutaka Okumoto, Takatoshi Tanisaka, and Susan R. Wessler. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*, 461(7267):1130–1134, October 2009.

[167] M Nei and W H Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5269–5273, October 1979.

[168] Masatoshi Nei, Yoshiyuki Suzuki, and Masafumi Nozawa. The Neutral Theory of Molecular Evolution in the Genomic Era. *Annual Review of Genomics and Human Genetics*, 11(1):265–289, 2010.

[169] Craig E Nelson, Bradley M Hersh, and Sean B Carroll. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biology*, 5(4):R25, 2004.

[170] P. C. Ng and S. Henikoff. Predicting deleterious amino acid substitutions. *Genome Research*, 11(5):863–874, May 2001.

[171] John W. Nicol, Gregg A. Helt, Steven G. Blanchard, Archana Raja, and Ann E. Loraine. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731, October 2009.

[172] Thomas Nussbaumer, Mihaela M. Martis, Stephan K. Roessner, Matthias Pfeifer, Kai C. Bader, Sapna Sharma, Heidrun Gundlach, and Manuel Spannagl. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Research*, 41(Database issue):D1144–1151, January 2013.

[173] Susumu Ohno. *Evolution by gene duplication*. Allen & Unwin; Springer-Verlag, London, New York, 1970.

[174] Keith R. Oliver, Jen A. McComb, and Wayne K. Greene. Transposable Elements: Powerful Contributors to Angiosperm Evolution and Diversity. *Genome Biology and Evolution*, 5(10):1886–1901, October 2013.

[175] Kenneth M. Olsen and Jonathan F. Wendel. Crop plants as models for understanding plant adaptation and diversification. *Frontiers in Plant Science*, 4, 2013.

[176] F. P. O'Mara. The role of grasslands in food security and climate change. *Annals of Botany*, page mcs209, September 2012.

[177] L. E. Orgel and F. H. Crick. Selfish DNA: the ultimate parasite. *Nature*, 284(5757):604–607, April 1980.

[178] Lance E. Palmer, Pablo D. Rabinowicz, Andrew L. O'Shaughnessy, Vivekanand S. Balija, Lidia U. Nascimento, Sujit Dike, Melissa de la Bastide, Robert A. Martienssen, and W. Richard McCombie. Maize Genome Sequencing by Methylation Filtration. *Science*, 302(5653):2115–2117, December 2003.

[179] Isobel AP Parkin, Chushin Koh, Haibao Tang, Stephen J. Robinson, Sateesh Kagale, Wayne E. Clarke, Chris D. Town, John Nixon, Vivek Krishnakumar, Shelby L. Bidwell, France Denoeud, Harry Belcram, Matthew G. Links, Jérémy Just, Carling Clarke, Tricia Bender, Terry Huebert, Annaliese S. Mason, J. C. Pires, Guy Barker, Jonathan Moore, Peter G. Walley, Sahana Manoli, Jacqueline Batley, David Edwards, Matthew N. Nelson, Xiyin Wang, Andrew H. Paterson, Graham King, Ian Bancroft, Boulos Chalhoub, and Andrew G. Sharpe. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome Biology*, 15(6):R77, June 2014.

[180] A. H. Paterson, J. E. Bowers, and B. A. Chapman. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9903–9908, June 2004.

[181] Andrew H. Paterson. Polyploidy, evolutionary opportunity, and crop adaptation. *Genetica*, 123(1-2):191–196, February 2005.

[182] Andrew H. Paterson, John E. Bowers, Rémy Bruggmann, Inna Dubchak, Jane Grimwood, Heidrun Gundlach, Georg Haberer, Uffe Hellsten, Therese Mitros, Alexander Poliakov, Jeremy Schmutz, Manuel Spannagl, Haibao Tang, Xiyin Wang, Thomas Wicker, Arvind K. Bharti, Jarrod Chapman, F. Alex Feltus, Udo Gowik, Igor V. Grigoriev, Eric Lyons, Christopher A. Maher, Mihaela Martis, Apurva Narechania, Robert P. Otillar, Bryan W. Penning, Asaf A. Salamov, Yu Wang, Lifang Zhang, Nicholas C. Carpita, Michael Freeling, Alan R. Gingle, C. Thomas Hash, Beat Keller, Patricia Klein, Stephen Kresovich, Maureen C. McCann, Ray Ming, Daniel G. Peterson, Mehboob ur Rahman, Doreen Ware, Peter Westhoff, Klaus F. X. Mayer, Joachim Messing, and Daniel S. Rokhsar. The Sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551–556, January 2009.

[183] Andrew H. Paterson, Yann-Rong Lin, Zhikang Li, Keith F. Schertz, John F. Doebley, Shannon R. M. Pinson, Sin-Chieh Liu, James W. Stansel, and James E. Irvine. Convergent Domestication of Cereal Crops by Independent Mutations at Corresponding Genetic Loci. *Science*, 269(5231):1714–1718, September 1995.

[184] D. A. Petrov. Evolution of genome size: new approaches to an old problem. *Trends in genetics: TIG*, 17(1):23–28, January 2001.

[185] Romain Philippe, Brigitte Courtois, Kenneth L. McNally, Pierre Mournet, Redouane El-Malki, Marie Christine Le Paslier, Denis Fabre, Claire Billot, Dominique Brunel, Jean-Christophe Glaszmann, and Dominique This. Structure, allelic diversity and selection of Asr genes, candidate for drought tolerance, in Oryza sativa L. and wild relatives. *Theoretical and Applied Genetics*, 121(4):769–787, August 2010.

[186] Saurabh D. Pophaly and Aurélien Tellier. Population Level Purifying Selection and Gene Expression Shape Subgenome Evolution in Maize. *Molecular Biology and Evolution*, 32(12):3226–3235, December 2015.

[187] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*, 436(7052):793–800, August 2005.

[188] Michael D. Purugganan and Dorian Q. Fuller. The nature of selection during plant domestication. *Nature*, 457(7231):843–848, February 2009.

[189] Aaron R. Quinlan. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., 2002.

[190] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[191] Elizabeth A. Rach, Hsiang-Yu Yuan, William H. Majoros, Pavel Tomancak, and Uwe Ohler. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biology*, 10:R73, 2009.

[192] Antoni Rafalski and Evgueni Ananiev. Genetic Diversity, Linkage Disequilibrium and Association Mapping. In Jeffrey L. Bennetzen and Sarah Hake, editors, *Handbook of Maize*, pages 201–219. Springer New York, 2009. DOI: 10.1007/978-0-387-77863-1_10.

[193] Ryan A. Rapp, Joshua A. Udall, and Jonathan F. Wendel. Genomic expression dominance in allopolyploids. *BMC Biology*, 7:18, 2009.

[194] Sandeep Ravindran. Barbara McClintock and the discovery of jumping genes. *Proceedings of the National Academy of Sciences*, 109(50):20198–20199, December 2012.

[195] Nirmala Arul Rayan, Ricardo C. H. del Rosario, and Shyam Prabhakar. Massive contribution of transposable elements to mammalian regulatory sequences. *Seminars in Cell & Developmental Biology*, 57:51–56, September 2016.

[196] Michael Regulski, Zhenyuan Lu, Jude Kendall, Mark T. A. Donoghue, Jon Reinders, Victor Llaca, Stephane Deschamps, Andrew Smith, Dan Levy, W. Richard McCombie, Scott Tingey, Antoni Rafalski, James Hicks, Doreen Ware, and Robert A. Martienssen. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Research*, 23(10):1651–1662, October 2013.

[197] Sébastien Renaut, Christopher J. Grassa, Brook T. Moyers, Nolan C. Kane, and Loren H. Rieseberg. The Population Genomics of Sunflowers and Genomic Determinants of Protein Evolution Revealed by RNAseq. *Biology*, 1(3):575–596, October 2012.

[198] Simon Renny-Byfield, Joseph P. Gallagher, Corrinne E. Grover, Emmanuel Szadkowski, Justin T. Page, Joshua A. Udall, Xiyin Wang, Andrew H. Paterson, and Jonathan F. Wendel. Ancient Gene Duplicates in Gossypium (Cotton) Exhibit Near-Complete Expression Divergence. *Genome Biology and Evolution*, 6(3):559–571, March 2014.

[199] Simon Renny-Byfield, Lei Gong, Joseph P. Gallagher, and Jonathan F. Wendel. Persistence of sub-genomes in paleopolyploid cotton after 60 million years of evolution. *Molecular Biology and Evolution*, page msv001, January 2015.

[200] Eli Rodgers-Melnick, Shrinivasrao P. Mane, Palitha Dharmawardhana, Gancho T. Slavov, Oswald R. Crasta, Steven H. Strauss, Amy M. Brunner, and Stephen P. DiFazio. Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus. *Genome Research*, 22(1):95–105, January 2012.

[201] Eli Rodgers-Melnick, Daniel L. Vera, Hank W. Bass, and Edward S. Buckler. Open chromatin reveals the functional maize genome. *Proceedings of the National Academy of Sciences*, 113(22):E3177–E3184, May 2016.

[202] Jeffrey Ross-Ibarra, Peter L. Morrell, and Brandon S. Gaut. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences*, 104(suppl 1):8641–8648, May 2007.

[203] Christian Roth, Shruti Rastogi, Lars Arvestad, Katharina Dittmar, Sara Light, Diana Ekman, and David A. Liberles. Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 308B(1):58–73, January 2007.

[204] Anne Roulin, Paul L. Auer, Marc Libault, Jessica Schlueter, Andrew Farmer, Greg May, Gary Stacey, Rebecca W. Doerge, and Scott A. Jackson. The fate of duplicated genes in a polyploid plant genome. *The Plant Journal*, 73(1):143–153, January 2013.

[205] Scott W Roy. The origin of recent introns: transposons? *Genome Biology*, 5(12):251, 2004.

[206] Julio Rozas. DNA Sequence Polymorphism Analysis Using DnaSP. In David Posada, editor, *Bioinformatics for DNA Sequence Analysis*, number 537 in Methods in Molecular Biology, pages 337–350. Humana Press, January 2009. DOI: 10.1007/978-1-59745-251-9_17.

[207] Claudia A. M. Russo and Carolina M. Voloch. Beads and Dice in a Genetic Drift Exercise. *Evolution: Education and Outreach*, 5(3):494–500, September 2012.

[208] Rajib Saha, Patrick F. Suthers, and Costas D. Maranas. Zea mays iRS1563: A Comprehensive Genome-Scale Metabolic Reconstruction of Maize Metabolism. *PLoS ONE*, 6(7):e21784, July 2011. 00044.

[209] Silvio Salvi, Giorgio Sponza, Michele Morgante, Dwight Tomes, Xiaomu Niu, Kevin A. Fengler, Robert Meeley, Evgueni V. Ananiev, Sergei Svitashev, Edward Bruggemann, Bailin Li, Christine F. Hainey, Slobodanka Radovic, Giusi Zaina, J.-Antoni Rafalski, Scott V. Tingey, Guo-Hua Miao, Ronald L. Phillips, and Roberto Tuberosa. Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences*, 104(27):11376–11381, July 2007.

[210] Phillip SanMiguel, Brandon S. Gaut, Alexander Tikhonov, Yuko Nakajima, and Jeffrey L. Bennetzen. The paleontology of intergene retrotransposons of maize. *Nature Genetics*, 20(1):43–45, September 1998.

[211] Phillip SanMiguel and Clémentine Vitte. The LTR-Retrotransposons of Maize. In Jeffrey L. Bennetzen and Sarah Hake, editors, *Handbook of Maize*, pages 307–327. Springer New York, 2009. DOI: 10.1007/978-0-387-77863-1_15.

[212] Manon Savard, Mark Nesbitt, and Martin K. Jones. The role of wild grasses in subsistence and sedentism: new evidence from the northern Fertile Crescent. *World Archaeology*, 38(2):179–196, June 2006.

[213] Hidetoshi Saze and Tetsuji Kakutani. Differentiation of epigenetic modifications between transposons and genes. *Current Opinion in Plant Biology*, 14(1):81–87, February 2011.

[214] Hidetoshi Saze, Kazuo Tsugane, Tatsuo Kanno, and Taisuke Nishimura. DNA Methylation in Plants: Relationship to Small RNAs and Histone Modifications, and Functions in Transposon Inactivation. *Plant and Cell Physiology*, 53(5):766–784, May 2012.

[215] James C. Schnable and Michael Freeling. Genes Identified by Visible Mutant Phenotypes Show Increased Bias toward One of Two Subgenomes of Maize. *PLOS ONE*, 6(3):e17855, March 2011.

[216] James C. Schnable, Michael Freeling, and Eric Lyons. Genome-Wide Analysis of Syntenic Gene Deletion in the Grasses. *Genome Biology and Evolution*, 4(3):265–277, January 2012.

[217] James C. Schnable, Nathan M. Springer, and Michael Freeling. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences*, 108(10):4069–4074, March 2011.

[218] James C. Schnable, Xiaowu Wang, J. Chris Pires, and Michael Freeling. Escape from preferential retention following repeated whole genome duplications in plants. *Plant Genetics and Genomics*, 3:94, 2012.

[219] Patrick S. Schnable, Doreen Ware, Robert S. Fulton, Joshua C. Stein, Fusheng Wei, Shiran Pasternak, Chengzhi Liang, Jianwei Zhang, Lucinda Fulton, Tina A. Graves, Patrick Minx, Amy Denise Reily, Laura Courtney, Scott S. Kruchowski, Chad Tomlinson, Cindy Strong, Kim Delehaunty, Catrina Fronick, Bill Courtney, Susan M. Rock, Eddie Belter, Feiyu Du, Kyung Kim, Rachel M. Abbott, Marc Cotton, Andy Levy, Pamela Marchetto, Kerri Ochoa, Stephanie M. Jackson, Barbara Gillam, Weizu Chen, Le Yan, Jamey Higginbotham, Marco Cardenas, Jason Waligorski, Elizabeth Applebaum, Lindsey Phelps, Jason Falcone, Krishna Kanchi, Thynn Thane, Adam Scimone, Nay Thane, Jessica Henke, Tom Wang, Jessica Ruppert, Neha Shah, Kelsi Rotter, Jennifer Hodges, Elizabeth Ingenthron, Matt Cordes, Sara Kohlberg, Jennifer Sgro, Brandon Delgado, Kelly Mead, Asif Chinwalla, Shawn Leonard, Kevin Crouse, Kristi Collura, Dave Kudrna, Jennifer Currie, Ruifeng He, Angelina Angelova, Shanmugam Rajasekar, Teri Mueller, Rene Lomeli, Gabriel Scara, Ara Ko, Krista Delaney, Marina Wissotski, Georgina Lopez, David Campos, Michele Braidotti, Elizabeth Ashley, Wolfgang Golser, HyeRan Kim, Seunghee Lee, Jinke Lin, Zeljko Dujmic, Woojin Kim, Jayson Talag, Andrea Zuccolo, Chuanzhu Fan, Aswathy Sebastian, Melissa Kramer, Lori Spiegel, Lidia Nascimento, Theresa Zutavern, Beth Miller, Claude Ambroise, Stephanie Muller, Will Spooner, Apurva Narechania, Liya Ren, Sharon Wei, Sunita Kumari, Ben Faga, Michael J. Levy, Linda McMahan, Peter Van Buren, Matthew W. Vaughn, Kai Ying, Cheng-Ting Yeh, Scott J. Emrich, Yi Jia, Ananth Kalyanaraman, An-Ping Hsia, W. Brad Barbazuk, Regina S. Baucom, Thomas P. Brutnell, Nicholas C. Carpita, Cristian Chaparro, Jer-Ming Chia, Jean-Marc Deragon, James C. Estill, Yan Fu, Jeffrey A. Jeddeloh, Yujun Han, Hyeran Lee, Pinghua Li, Damon R. Lisch, Sanzhen Liu, Zhijie Liu, Dawn Holligan Nagel, Maureen C. McCann, Phillip SanMiguel, Alan M. Myers, Dan Nettleton, John Nguyen, Bryan W. Penning, Lalit Ponnala, Kevin L. Schneider, David C. Schwartz, Anupma Sharma, Carol Soderlund, Nathan M.

Springer, Qi Sun, Hao Wang, Michael Waterman, Richard Westerman, Thomas K. Wolfgruber, Lixing Yang, Yeisoo Yu, Lifang Zhang, Shiguo Zhou, Qihui Zhu, Jeffrey L. Bennetzen, R. Kelly Dawe, Jiming Jiang, Ning Jiang, Gernot G. Presting, Susan R. Wessler, Srinivas Aluru, Robert A. Martienssen, Sandra W. Clifton, W. Richard McCombie, Rod A. Wing, and Richard K. Wilson. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*, 326(5956):1112–1115, November 2009.

[220] CC Schön, V Wimmer, and C Lehermeier. Efficiency of Variable Selection in Genome-Wide Prediction for Traits of Different Genetic Architecture. In *10th World Congress of Genetics Applied to Livestock Production*, 2014.

[221] Rajandeep S. Sekhon, Haining Lin, Kevin L. Childs, Candice N. Hansey, C. Robin Buell, Natalia de Leon, and Shawn M. Kaeppler. Genome-wide atlas of transcription during maize development. *The Plant Journal: For Cell and Molecular Biology*, 66(4):553–563, May 2011.

[222] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa, Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, December 2003.

[223] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, 40(Web Server issue):W452–W457, July 2012.

[224] Cas Simons, Igor V. Makunin, Michael Pheasant, and John S. Mattick. Maintenance of transposon-free regions throughout vertebrate evolution. *BMC Genomics*, 8:470, 2007.

[225] Manuela Sironi, Giorgia Menozzi, Giacomo P. Comi, Matteo Cereda, Rachele Cagliani, Nereo Bresolin, and Uberto Pozzoli. Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. *Genome Biology*, 7:R120, 2006.

[226] David S. Skibbe, John F. Fernandes, Katalin F. Medzihradszky, Alma L. Burlingame, and Virginia Walbot. Mutator transposon activity reprograms the transcriptomes and proteomes of developing maize anthers. *The Plant Journal*, 59(4):622–633, August 2009.

[227] M Slatkin. Allele age and a test for selection on rare alleles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355(1403):1663–1668, November 2000.

[228] R. Keith Slotkin and Robert Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4):272–285, April 2007.

[229] J. M. Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23(1):23–35, February 1974.

[230] Vitor Sousa and Jody Hey. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, 14(6):404–414, June 2013.

[231] Flávio S. J. de Souza, Lucía F. Franchini, and Marcelo Rubinstein. Exaptation of Transposable Elements into Novel Cis-Regulatory Elements: Is the Evidence Always Strong? *Molecular Biology and Evolution*, 30(6):1239–1251, June 2013.

[232] Nathan M. Springer, Damon Lisch, and Qing Li. Creating Order from Chaos: Epigenome Dynamics in Plants with Complex Genomes. *The Plant Cell*, 28(2):314–325, February 2016.

[233] Nathan M. Springer, Kai Ying, Yan Fu, Tieming Ji, Cheng-Ting Yeh, Yi Jia, Wei Wu, Todd Richmond, Jacob Kitzman, Heidi Rosenbaum, A. Leonardo Iniguez, W. Brad Barbazuk, Jeffrey A. Jeddeloh, Dan Nettleton, and Patrick S. Schnable. Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLOS Genet*, 5(11):e1000734, November 2009.

[234] Jason E. Stajich, David Block, Kris Boulez, Steven E. Brenner, Stephen A. Chervitz, Chris Dagdigian, Georg Fuellen, James G. R. Gilbert, Ian Korf, Hilmar Lapp, Heikki Lehväslaiho, Chad Matsalla, Chris J. Mungall, Brian I. Osborne, Matthew R. Pocock, Peter Schattner, Martin Senger, Lincoln D. Stein, Elia Stupka, Mark D. Wilkinson, and Ewan Birney. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, 12(10):1611–1618, October 2002.

[235] Jared L. Strasburg, Nolan C. Kane, Andrew R. Raduski, Aurélie Bonin, Richard Michelmore, and Loren H. Rieseberg. Effective Population Size Is Positively Correlated with Levels of Adaptive Divergence among Annual Sunflowers. *Molecular Biology and Evolution*, 28(5):1569–1580, May 2011.

[236] Anthony Studer, Qiong Zhao, Jeffrey Ross-Ibarra, and John Doebley. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nature Genetics*, 43(11):1160–1163, November 2011.

[237] Zhixi Su, Jianmin Wang, Jun Yu, Xiaoqiu Huang, and Xun Gu. Evolution of alternative splicing after gene duplication. *Genome Research*, 16(2):182–189, February 2006.

[238] Xiaoliang Sun, Yong Zou, Victoria Nikiforova, Jürgen Kurths, and Dirk Walther. The complexity of gene expression dynamics revealed by permutation entropy. *BMC Bioinformatics*, 11:607, 2010.

[239] Mikita Suyama, David Torrents, and Peer Bork. PAL2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(Web Server issue):W609–W612, July 2006.

[240] Ruth A. Swanson-Wagner, Steven R. Eichten, Sunita Kumari, Peter Tiffin, Joshua C. Stein, Doreen Ware, and Nathan M. Springer. Pervasive gene content variation and copy number

variation in maize and its undomesticated progenitor. *Genome Research*, 20(12):1689–1699, December 2010.

[241] Zuzana Swigoňová, Jinsheng Lai, Jianxin Ma, Wusirika Ramakrishna, Victor Llaca, Jeffrey L. Bennetzen, and Joachim Messing. Close Split of Sorghum and Maize Genome Progenitors. *Genome Research*, 14(10a):1916–1923, October 2004.

[242] Amir Szitenberg, Soyeon Cha, Charles H. Opperman, David M. Bird, Mark L. Blaxter, and David H. Lunt. Genetic drift, not life history or RNAi, determine long term evolution of transposable elements. *Genome Biology and Evolution*, page evw208, August 2016.

[243] F. Tajima. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, 123(3):585–595, November 1989.

[244] Maud I. Tenaillon, Jesse D. Hollister, and Brandon S. Gaut. A triptych of the evolution of plant transposable elements. *Trends in Plant Science*, 15(8):471–478, August 2010.

[245] Maud I. Tenaillon, Matthew B. Hufford, Brandon S. Gaut, and Jeffrey Ross-Ibarra. Genome Size and Transposable Element Content as Determined by High-Throughput Sequencing in Maize and Zea luxurians. *Genome Biology and Evolution*, 3:219–229, January 2011.

[246] Maud I. Tenaillon, Mark C. Sawkins, Anthony D. Long, Rebecca L. Gaut, John F. Doebley, and Brandon S. Gaut. Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). *Proceedings of the National Academy of Sciences*, 98(16):9161–9166, July 2001.

[247] Maud I. Tenaillon, Jana U'Ren, Olivier Tenaillon, and Brandon S. Gaut. Selection Versus Demography: A Multilocus Investigation of the Domestication Process in Maize. *Molecular Biology and Evolution*, 21(7):1214–1225, July 2004.

[248] Shawn R. Thatcher, Wengang Zhou, April Leonard, Bing-Bing Wang, Mary Beatty, Gina Zastrow-Hayes, Xiangyu Zhao, Andy Baumgarten, and Bailin Li. Genome-Wide Analysis of Alternative Splicing in Zea mays: Landscape and Genetic Regulation. *The Plant Cell Online*, page tpc.114.130773, September 2014.

[249] Brian C. Thomas, Brent Pedersen, and Michael Freeling. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Research*, 16(7):934–946, July 2006.

[250] J D Thompson, D G Higgins, and T J Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, November 1994.

[251] Feng Tian, Natalie M. Stevens, and Edward S. Buckler. Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proceedings of the National Academy of Sciences*, 106(Supplement 1):9979–9986, June 2009.

[252] Zhixi Tian, Meixia Zhao, Maoyun She, Jianchang Du, Steven B. Cannon, Xin Liu, Xun Xu, Xinpeng Qi, Man-Wah Li, Hon-Ming Lam, and Jianxin Ma. Genome-Wide Characterization of Nonreference Transposons Reveals Evolutionary Propensities of Transposons in Soybean. *The Plant Cell*, 24(11):4422–4436, November 2012.

[253] Itay Tirosh and Naama Barkai. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biology*, 8(4):R50, April 2007.

[254] Joshua A. Udall and Jonathan F. Wendel. Polyploidy and Crop Improvement. *Crop Science*, 46(Supplement_1):S–3–S–14, November 2006.

[255] Sandra Unterseer, Saurabh D. Pophaly, Regina Peis, Peter Westermeier, Manfred Mayer, Michael A. Seidel, Georg Haberer, Klaus F. X. Mayer, Bernardo Ordas, Hubert Pausch, Aurélien Tellier, Eva Bauer, and Chris-Carolin Schön. A comprehensive study of the genomic differentiation between temperate Dent and Flint maize. *Genome Biology*, 17:137, 2016.

[256] Kevin Vanneste, Steven Maere, and Yves Van de Peer. Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution. *Phil. Trans. R. Soc. B*, 369(1648):20130353, August 2014.

[257] D. A. Vaughan, E. Balázs, and J. S. Heslop-Harrison. From Crop Domestication to Super-domestication. *Annals of Botany*, 100(5):893–901, October 2007.

[258] Carlos M. Vicient. Transcriptional activity of transposable elements in maize. *BMC Genomics*, 11:601, 2010.

[259] Albert J. Vilella, Angel Blanco-Garcia, Stephan Hutter, and Julio Rozas. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*, 21(11):2791–2793, June 2005.

[260] Clémentine Vitte, Margaux-Alison Fustier, Karine Alix, and Maud I. Tenaillon. The bright side of transposons in crop evolution. *Briefings in Functional Genomics*, page elu002, March 2014.

[261] Joseph J. Vitti, Sharon R. Grossman, and Pardis C. Sabeti. Detecting Natural Selection in Genomic Data. *Annual Review of Genetics*, 47(1):97–120, 2013.

[262] Huai Wang, Anthony J. Studer, Qiong Zhao, Robert Meeley, and John F. Doebley. Evidence That the Origin of Naked Kernels During Maize Domestication Was Caused by a Single Amino Acid Substitution in tga1. *Genetics*, 200(3):965–974, July 2015.

[263] Jianlin Wang, Lu Tian, Hyeon-Se Lee, Ning E. Wei, Hongmei Jiang, Brian Watson, Andreas Madlung, Thomas C. Osborn, R. W. Doerge, Luca Comai, and Z. Jeffrey Chen. Genomewide Nonadditive Gene Regulation in Arabidopsis Allotetraploids. *Genetics*, 172(1):507–517, January 2006.

[264] Qinghua Wang and Hugo K. Dooner. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proceedings of the National Academy of Sciences*, 103(47):17644–17649, November 2006.

[265] SuiKang Wang, YouHuang Bai, ChenJia Shen, YunRong Wu, SaiNa Zhang, DeAn Jiang, Tom J. Guilfoyle, Ming Chen, and YanHua Qi. Auxin-related gene families in abiotic stress response in Sorghum bicolor. *Functional & Integrative Genomics*, 10(4):533–546, November 2010.

[266] Xi Wang, Detlef Weigel, and Lisa M. Smith. Transposon Variants and Their Effects on Gene Expression in Arabidopsis. *PLOS Genet*, 9(2):e1003255, February 2013.

[267] Xiangfeng Wang, Axel A. Elling, Xueyong Li, Ning Li, Zhiyu Peng, Guangming He, Hui Sun, Yijun Qi, X. Shirley Liu, and Xing Wang Deng. Genome-Wide and Organ-Specific Landscapes of Epigenetic Modifications and Their Relationships to mRNA and Small RNA Transcriptomes in Maize. *The Plant Cell*, 21(4):1053–1069, April 2009.

[268] Xiyin Wang, Udo Gowik, Haibao Tang, John E. Bowers, Peter Westhoff, and Andrew H. Paterson. Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biology*, 10:R68, 2009.

[269] Xiyin Wang, Haibao Tang, and Andrew H. Paterson. Seventy Million Years of Concerted Evolution of a Homoeologous Chromosome Pair, in Parallel, in Major Poaceae Lineages. *The Plant Cell*, 23(1):27–37, January 2011.

[270] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, April 1975.

[271] Bin Wei, Hanmei Liu, Xin Liu, Qianlin Xiao, Yongbin Wang, Junjie Zhang, Yufeng Hu, Yinghong Liu, Guowu Yu, and Yubi Huang. Genome-wide characterization of non-reference transposons in crops suggests non-random insertion. *BMC Genomics*, 17:536, 2016.

[272] Fusheng Wei, Ed Coe, William Nelson, Arvind K. Bharti, Fred Engler, Ed Butler, HyeRan Kim, Jose Luis Goicoechea, Mingsheng Chen, Seunghee Lee, Galina Fuks, Hector Sanchez-Villeda, Steven Schroeder, Zhiwei Fang, Michael McMullen, Georgia Davis, John E. Bowers, Andrew H. Paterson, Mary Schaeffer, Jack Gardiner, Karen Cone, Joachim Messing, Carol Soderlund, and Rod A. Wing. Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History. *PLOS Genet*, 3(7):e123, July 2007.

[273] John H. Werren. Selfish genetic elements, genetic conflict, and evolutionary innovation. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10863–10870, June 2011.

[274] Patrick T. West, Qing Li, Lexiang Ji, Steven R. Eichten, Jawon Song, Matthew W. Vaughn, Robert J. Schmitz, and Nathan M. Springer. Genomic Distribution of H3k9me2 and DNA Methylation in a Maize Genome. *PLOS ONE*, 9(8):e105267, August 2014.

[275] Thomas Wicker, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, Philippe Leroy, Michele Morgante, Olivier Panaud, Etienne Paux, Phillip SanMiguel, and Alan H. Schulman. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982, December 2007.

[276] Margaret R. Woodhouse, James C. Schnable, Brent S. Pedersen, Eric Lyons, Damon Lisch, Shabarinath Subramaniam, and Michael Freeling. Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs. *PLOS Biol*, 8(6):e1000409, June 2010.

[277] Stephen I. Wright, Newton Agrawal, and Thomas E. Bureau. Effects of Recombination Rate and Gene Density on Transposable Element Distributions in Arabidopsis thaliana. *Genome Research*, 13(8):1897–1903, August 2003.

[278] Stephen I. Wright, Irie Vroh Bi, Steve G. Schroeder, Masanori Yamasaki, John F. Doebley, Michael D. McMullen, and Brandon S. Gaut. The Effects of Artificial Selection on the Maize Genome. *Science*, 308(5726):1310–1314, May 2005.

[279] Xudong Wu and Xiaoquan Qi. Genes encoding hub and bottleneck enzymes of the Arabidopsis metabolic network preferentially retain homeologs through whole genome duplication. *BMC Evolutionary Biology*, 10(1):145, May 2010. 00014.

[280] Yufeng Wu, Zhengge Zhu, Ligeng Ma, and Mingsheng Chen. The Preferential Retention of Starch Synthesis Genes Reveals the Impact of Whole-Genome Duplication on Grass Evolution. *Molecular Biology and Evolution*, 25(6):1003–1006, June 2008.

[281] Hongyan Xing, Ramesh N. Pudake, Ganggang Guo, Guofang Xing, Zhaorong Hu, Yirong Zhang, Qixin Sun, and Zhongfu Ni. Genome-wide identification and expression profiling of auxin response factor (ARF) gene family in maize. *BMC Genomics*, 12:178, 2011.

[282] Masanori Yamasaki, Maud I. Tenaillon, Irie Vroh Bi, Steve G. Schroeder, Hector Sanchez-Villeda, John F. Doebley, Brandon S. Gaut, and Michael D. McMullen. A Large-Scale Screen for Artificial Selection in Maize Identifies Candidate Agronomic Loci for Domestication and Crop Improvement. *The Plant Cell*, 17(11):2859–2872, November 2005.

[283] Itai Yanai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, Shirley Horn-Saban, Marilyn Safran, Eytan Domany, Doron Lancet, and Orit Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659, March 2005.

[284] Liang Yang and Brandon S. Gaut. Factors that contribute to variation in evolutionary rate among Arabidopsis genes. *Molecular Biology and Evolution*, page msr058, March 2011.

[285] Qin Yang, Zhi Li, Wenqiang Li, Lixia Ku, Chao Wang, Jianrong Ye, Kun Li, Ning Yang, Yipu Li, Tao Zhong, Jiansheng Li, Yanhui Chen, Jianbing Yan, Xiaohong Yang, and Mingliang Xu. CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the postdomestication spread of maize. *Proceedings of the National Academy of Sciences*, 110(42):16969–16974, October 2013.

[286] Ziheng Yang and Joseph P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12):496–503, December 2000.

[287] Jianbo Zhang, Thomas Peterson, and Peter A. Peterson. Transposons Ac/Ds, En/Spmand their Relatives in Maize. In Jeffrey L. Bennetzen and Sarah Hake, editors, *Handbook of Maize*, pages 251–276. Springer New York, 2009. DOI: 10.1007/978-0-387-77863-1_13.

[288] Peter G. Zhang, Suzanne Z. Huang, Anne-Laure Pin, and Keith L. Adams. Extensive Divergence in Alternative Splicing Patterns after Gene and Genome Duplication During the Evolutionary History of Arabidopsis. *Molecular Biology and Evolution*, 27(7):1686–1697, July 2010.

[289] Zengcui Zhang, Harry Belcram, Piotr Gornicki, Mathieu Charles, Jérémy Just, Cécile Huneau, Ghislaine Magdelenat, Arnaud Couloux, Sylvie Samain, Bikram S. Gill, Jack B. Rasmussen, Valérie Barbe, Justin D. Faris, and Boulos Chalhoub. Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proceedings of the National Academy of Sciences*, 108(46):18737–18742, November 2011.

[290] Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, April 2014.

[291] Xiaobei Zhao, Eivind Valen, Brian J. Parker, and Albin Sandelin. Systematic Clustering of Transcription Start Site Landscapes. *PLOS ONE*, 6(8):e23409, August 2011.