Title: Fates of Tandem Duplications in the Maize Genome

Thomas J. Y. Kono1, Alex B. Brohammer1, Suzanne E. McGaugh2, and Candice N. Hirsch1

1: Department of Agronomy and Plant Genetics, University of Minnesota, 411 Borlaug Hall, 1991 Upper Buford Circle, Saint Paul, MN 55108

2: Department of Ecology, Evolution, and Behavior, University of Minnesota, 140 Gortner Laboratory, 1479 Gortner Avenue, Saint Paul, MN 55108

Corresponding Author:

Candice N. Hirsch  
411 Borlaug Hall  
1991 Upper Buford Circle  
Saint Paul, MN 55108  
USA  
Fax: 612-625-1268  
Email: cnhirsch@umn.edu

# Abstract

# Introduction

Gene duplications provide a mechanism through which functional novelty may arise. Many protein coding genes in eukaryotes are part of large families of genes with related function, consistent with origins in gene duplication (Rubin et al. 2000). Duplicated genes may either be tandem (proximal to each other), or dispersed (separated in the genome). Dispersed duplicate genes may functionally diverge because the two copies may be subject to different regulatory pressures. Conversely, genes that are tandemly duplicated may have the same regulatory elements, and thus may not diverge in the same way as dispersed duplicates.

When a gene duplicates, it potentially generates a phenotypic impact through alteration of gene product dosage. In many cases, the sudden change of gene product concentration has deleterious effects on the physiology of the organism, and will be selected against. In some cases, however, the increased gene expression may be beneficial, and there will be selection to maintain the duplication, e.g., tandem duplications conferring soybean cyst nematode resistance at Rhg1 (Cook et al. 2012).

In the long run, the fate of tandem duplicate genes is less straightforward the binary of retained as a beneficial allele, or lost as a deleterious variant. Mutations in the regulatory regions of tandem duplicates, or mutations in the coding sequence, may cause the genes to be expressed in different tissues, or may cause the genes to fulfill non-redundant functional roles. Several models such as the “duplication-degeneration-complementation” model or the “escape from adaptive conflict” model describe these scenarios as possible outcomes of tandem duplicates (Innan 2009).

Recent improvements in DNA sequencing technology allow for the resolution of duplicated sequences. Reads are now long enough to span repetitive elements, where shorter reads would either miss or collapse repetitive sequences. In maize, the latest B73 reference genome was assembled with long read sequencing, which allows for improved resolution of repetitive sequences than previous assemblies. Additionally, a *de novo* genome assembly of PH207, another maize inbred line, has become recently available. While the PH207 assembly was generated with short read sequencing, it provides an opportunity to examine the nature of gene duplication and loss in maize.

The major questions we aim to address in this study are 1) How many genes are tandemly duplicated in the B73 and PH207 reference genomes, and where do they occur? 2) What is the age distribution of tandem duplications? 3) Do tandem duplicate genes in maize show a different substitution rate from other maize genes, and from other grass genes? To answer these questions, we compared the B73 and PH207 reference genomes to the outgroup species Sorghum bicolor using the CoGe suite of scripts to identify putative tandem duplicates. We then use RNAseq data from six tissues and PAML to compare functional and expression divergence of tandem duplicates. Finally, we estimate the ages of tandem duplicates, and compare those to their estimated functional outcomes.

# Results

## Genomic Locations of Tandem Duplicate Genes

Filtering of SynMap-identified tandem duplications and comparison of adjacent gene pairwise similarity resulted in 1,758 tandem duplicate clusters in B73 and 1,467 tandem duplicate clusters in PH207. Most of the tandem duplicate clusters occur in regions of the genome that are syntenic with *Sorghum bicolor* (Table 1). The total number of annotated genes in tandem duplicate clusters is 4,448 (11.3%) in B73, and 3,788 (9.3%) in PH207. Cassette duplications, which are segmental duplications of multiple genes, are rare in these two genome assemblies: B73 has 50 inferred cassette duplications, and PH207 has 47 inferred cassette duplications.

Tandem duplicate genes appear to be most closely associated with annotated gene density in the genome assemblies (Figure 1A). The density of RNA TEs, density of DNA TEs, or subgenome assignments do not appear to explain tandem duplications in the maize genome. However, because our method for identifying tandem duplications in the maize genome is based on annotated genes, it is expected that annotated gene density is the genomic factor that is most closely associated with tandem duplications.

When tandem duplications are shared between B73 and PH207, they often have similar cluster sizes (Figure 1B). That is, when homologous genes are both part of tandem duplicate clusters in B73 and PH207, they are often part of duplicate clusters involving the same number of genes. This is consistent with most of the shared duplications being ancient, which predates the divergence of the B73 and PH207 lineages (Figure 2A).

## Estimated Dates of Tandem Duplications

Generally, tandem duplicate genes were inferred to have a bimodal distribution, with many ancient duplicates and many recent duplicates (Figure 2A). Duplications that occur in syntenic positions, and are shared between B73 and PH207 are mostly ancient duplications, with very few recent events. Both syntenic duplications that are private to B73 or PH207 and duplications in nonsyntenic positions show a large proportion that are inferred to have occurred recently in maize history.

One explanation for the large proportion of inferred recent duplications is the action of gene conversion. Gene conversion would cause tandem duplicate genes to have higher sequence similarity than non-recombining duplicates of the same age, and thus would bias estimates toward recent events. Gene conversion is known to be biased toward GC-rich regions. While tandem duplicate genes show higher GC content on average than genome-wide (Figure 2F), recent duplications have lower GC content than ancient duplications, on average (Figure 2E).

## Substitution Rates of Tandem Duplicate Genes

# Discussion

# Materials and Methods

## Tandem Duplicate Identification

Putative tandem duplicate clusters were identified by comparing the B73v4 and PH207v1 maize assemblies to the Sorghum bicolor v3.1 genome assembly with SynMap. The putative tandem duplicate clusters from SynMap were refined based on criteria that will be described below. Full details of the SynMap parameters can be found in [Brohammer et al. 2017]. Because SynMap identifies putative tandem duplications based on pairwise BLAST of one genome to another, we also compared adjacent genes within the maize genome assemblies.

The longest transcripts from adjacent maize genes were translated to amino acids, and aligned with Clustal-omega. A total of 10 iterations of refinement were used for alignment. The aligned sequences were back-translated to nucleotides. Pairwise similarity, down-weighted for the proportion of gaps opened during alignment, was calculated with the “compute” program from the “analysis” package. A distribution of adjusted pairwise similarities for adjacent genes in the B73v4 and PH207v1 assemblies is shown in Figure S\_. Adjacent genes and groups of genes within SynMap-identified tandem duplicate clusters with at least 0.3 adjusted pairwise similarity were retained for analysis.

## Duplication Date Estimation

The dates of tandem duplications were estimated with BEAST. The two subgenome homeologues, homolgous loci in B73 and PH207, any tandem duplicates of those genes, and their putative Sorghum ancestral gene were translated to amino acids, then aligned with Clustal-omega. The alignments were back-translated to nucleotides. Each gene alignment was analyzed with BEAST, with a GTR+Gamma nucleotide substitution model, estimated transition probabilities and equilibrium base frequencies, a random local clock to allow for branch-specific rate variation, and a monophyletic divergence between the maize subgenomes with a prior of ~N(11.9, 1) on the divergence date. The MCMC was run for 10,000,000 steps.

Resulting trees from the BEAST analysis were parsed to obtain the time to most recent common ancestor (TMRCA) between tandem duplicate genes. It should be noted that when a tandem duplication with the same estimated age is shared between both B73 and PH207, we assume that the duplication was a single event that occurred before the divergence of B73 and PH207. However, gene duplications likely do not follow the infinite sites mutational model, and identity by state does not necessarily imply identity by descent. Additionally, we assume that the tandem duplicates are evolving along truly separate trajectories, and that gene conversion among tandem duplicates is negligible.

## Relative Rates Calculations

We compared the relative rates of sequence evolution of maize tandem duplicates with other grass genes by performing Clade model C (CMC) tests on tandem duplicates in orthologous gene groups. Orthologous gene groups were identified among publicly available grass genomes from Phytozome V12 and Ensembl Plants V34 using Orthofinder. The species, sources, and versions of the genomes used as Orthofinder input are shown in Table S\_. Orthofinder was run with the “dendroblast” orthologue search method, and default parameters for MCL clustering. Amino acid sequences from B73 and PH207 were kept in separate files. Orthologous gene groups that contain between 10 and 75 genes, contain maize tandem duplicates, and contain complete tandem duplicate clusters (i.e., a tandem duplicate cluster is not split among multiple orthologous groups) were retained for analysis.

Sequences for each orthologous group were aligned with clustal-omega, then back-translated to nucleotides using the CDS sequences as guides. Alignments were filtered to contain only sites with at most 50% gaps, because gaps greatly increase computation time and are not informative for substitution rates. Maximum likelihood trees were estimated from the filtered alignments with RAxML, using the default rapid hill-climbing search algorithm and a GTR+Gamma nucleotide substitution model.

Four models were fit to the filtered alignments and trees with the ‘codeml’ program in PAML. Model 1 marks maize genes and common ancestors of maize genes as different from other grass genes. Model 2 marks maize tandem duplicates and common ancestors of maize tandem duplicates as different from all other genes. Model 3 distinguishes maize tandem duplicates and common ancestors of maize tandem duplicates from maize genes, and maize genes as different from other grass genes. The null model treats every clade as evolving under the same constraint. The best-fitting model for each orthologous group was identified via a likelihood ratio test against the null model.

## Script Availability

Scripts to perform tandem duplicate identification, sequence alignment and back-translation, orthologous gene group identification, and relative rates comparisons are available at [GitHub link]. A detailed workflow document describing how to use the scripts and how to regenerate the results of this study is given at [link].

# Acknowledgments

# Tables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Genotype | Maize1 Clusters | Maize2 Clusters | Nonsyntenic Clusters | Total |
| B73 | 938 | 420 | 276 | 1,758 |
| PH207 | 691 | 316 | 248 | 1,467 |

Table 1: Counts of tandem duplicate clusters in syntenic and nonsyntenic positions, in B73 and PH207. PH207 has fewer inferred tandem duplicate genes than B73 because the repetitive nature of tandem duplications is better resolved in the B73 genome assembly.

# Figure Legends

Figure 1: Summaries of maize tandem gene duplications. A) Tandem duplications (purple ticks) associate most strongly with gene density (black line). RNA TEs (red line), DNA TEs (orange line), subgenome 1 (green shading), and subgenome 2 (blue shading) do not account for tandem gene duplication. The top panel shows B73 chr2, and the bottom panel shows PH207 chr2. B) Homologous tandem duplications between B73 and PH207 have similar cluster sizes.

Figure 2: Date estimates of maize tandem duplications. A) Distributions of syntenic and nonsyntenic duplication ages. Tandem duplications that occur in nonsyntenic positions and are private to either B73 or PH207 show an enrichment of recent duplicates, compared to shared syntenic duplications. B) Cassette duplications show a bimodal distribution of ages, suggesting that on average they are not arising via different mechanisms from tandem duplications generally. C) and D) Examples of tree estimates of tandem duplications. C) shows an ancient duplication and D) shows a recent duplication. Shaded colors show different tree topologies. E) Ancient (>=10MYA) and recent (<=2MYA) duplications have similar distributions of GC content. F) Genes in tandem duplicate clusters have higher GC content than genes genome-wide, on average.

# References