Gene duplication is a mutation type that has high potential to affect phenotypes. In addition to altering gene dosage, duplications may provide a substrate for further phenotypic change by allowing for rapid protein function divergence. However, not all gene duplications result in novel functional proteins; many duplicated genes lose function through pseudogenization. Here, we investigate the frequency of tandem gene duplication and the relative rates of the tandem duplicate fates in two inbred maize lines- B73 and PH207. Tandem duplicate clusters identified using the blast2raw script within the CoGe pipeline contain approximately 15% of the protein coding genes in B73 and approximately 23% of protein coding genes in PH207. Tandemly duplicated genes do not show an enrichment of transposable element sequence similarity with respect to genome wide averages when all classes of TEs are aggregated. However, LINE elements are highly over-represented in tandemly duplicated genes, and SINE elements are highly under-represented in tandemly duplicated genes, as compared to the genome wide average. We further test evolutionary rates hypotheses to compare the relative frequencies of pseudogenization, subfunctionalization, and divergent function of the tandem duplicates. This study represents the first genome-wide analysis of tandem duplicate evolution across multiple de novo genome assemblies within maize.

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 noco 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) How many tandem duplicate genes show evidence of functional divergence, at the DNA sequence or expression levels? 3) What is the age distribution of tandem duplications, and is there a relationship between their age and functional outcome? 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.

The B73v4 and PH207v1 reference sequences, including gene annotations, were downloaded from Phytozome V12 (link). Structural annotations of transposable elements in the B73 and PH207 reference sequences were provided by Michelle Stitzer. Ancestral states were estimated with the Sorghum bicolor and Oryza sativa reference sequences, available from Ensembl Plants (link).

Tandem duplicate genes were identified with the ‘blast2raw’ script from the CoGe pipeline (cite, link). A window size of 15 genes was used to identify potential tandem duplicates, with a C-value of 0.1. Tandem duplicate genes in regions that are syntenic with Sorghum bicolor were identified by comparisons of each maize reference sequence to the S. bicolor v3 reference sequence. Tandem duplicate genes in non-syntenic regions were identified by comparison of the B73 reference sequence to the PH207 reference sequence, using the same parameters. Quota align outputs were parsed with a custom Perl script to identify putative tandem duplicate gene IDs. The longest transcript from each putative tandem duplicate gene was chosen to represent the gene in evolutionary rates calculations and phylogenetic analyses. Putative tandem duplicate clusters and the longest transcripts for each gene are given in Table S\_.

CoGe identifies multiple genes in one genome as tandem duplicates if they match the same gene in another genome, but does not directly examine the sequence similarity in a single genome. Therefore, some tandem duplicate clusters identified with CoGe will not be true tandem duplicates; they will instead represent annotation errors in the genome assemblies. Tandem duplicate clusters were further filtered by similarity of their member genes. Full length coding sequences from each tandem duplicate gene cluster was translated into amino acid sequence, and iteratively aligned with Clustal Omega (cite). The alignment algorithm was run for a maximum of 10 iterations, using the full distance matrix in each iteration. The alignments were then backtranslated to nucleotides. Tandem duplicate clusters with less than 75% sequence identity and less than 50% coverage to each other were removed, because they are likely not true gene duplications.

Potential cassette duplications were identified by examining tiled clusters of tandem duplicate genes. Clusters of tandem duplicate genes were flagged as potential cassette duplications if the genomic region they occupy overlap, and are not nested. Potential cassettes were further filtered such that tandem duplicate clusters in cassettes were required to have membership numbers that are within two genes of each other. An illustration of the cassette identification procedure is given in Figure S\_.

Orthologous gene groups were identified with Orthofinder v. 1.1.5 (cite, link). Briefly, Orthofinder identifies orthologous groups of genes by reciprocal protein BLAST searches followed by MCL clustering. Orthofinder was run with the default E-value threshold of 0.001, and the DendroBLAST algorithm to infer orthologous groups. Orthofinder’s dependencies were satisfied with MCL v. 10.201 (cite), BLAST+ v. 2.2.29 (cite), MAFFT v. 7.305 (cite), FastTree v. 2.1.8 (cite), and DLCpar v. 1.0 (cite). The BLAST databases used by Orthofinder were generated from amino acid sequences of 13 publicly available grass reference sequences from Phytozome V12 and Ensembl Plants (Table S1\_), and from the longest transcripts for each gene model in the B73 and PH207 reference sequences.

The amino acid sequences of each member of each orthologous group were extracted, and aligned with Clustal-omega (cite.) Multiple sequence alignments were then back-translated to nucleotides. Maximum likelihood trees were generated from the back-translated alignments with RAxML (cite), using a HKY (cite) nucleotide substitution model. Synonymous (dS) and nonsynonymous (dN) substitution rates were estimated for each branch with PAML. Clade tests (cite) for differences in dN/dS among maize tandem duplicate genes with respect to sequenced grass species were performed with PAML.

Scripts to run Orthofinder, perform alignment back-translation, generate trees, and estimate dN/dS ratios are available at [GitHub Link].

Ages of tandem duplicate genes were estimated using the synonymous nucleotide substitution rate. Full coding sequence from each gene of each putative tandem duplicate cluster and an inferred ancestral gene sequence from Sorghum bicolor or Oryza sativa were translated to amino acids, then iteratively aligned with Clustal Omega (cite). Clustal Omega was run with a maximum of 10 iterations and with the full distance matrix used for each clustering iteration. Alignments were then backtranslated to nucleotides to compute synonymous and nonsynonymous substitution rates. Substitution rates were estimated with the ‘polydNdS’ program from the ‘analysis’ package (GitHub link).

We identified a total of 700 tandem duplicate gene clusters in B73, and 600 tandem duplicate gene clusters in PH207. These clusters contained \_\_\_ (\_\_% of protein coding genes in B73) and \_\_\_ (\_\_)% of protein coding genes in PH207) genes, respectively. Of the tandem duplicate clusters identified, 76 appear to be part of cassettes in B73, and 179 appear to be part of cassettes in PH207. Tandem duplicate genes were distributed randomly across the genome (Figure 1).

With our thresholds of 75% identity and 50% coverage between tandem duplicate genes, we identify 1,200 putative tandem duplicate gene clusters in B73 and 1,500 putative clusters in PH207 that appear to be the result of annotation errors. Genes from these clusters were either dramatically different in length, or were so dissimilar to each other at the nucleotide level to be unlikely to be derived from a recent duplication event.