

ABSTRACT

GENE CONTENT EVOLUTION IN PLANT GENOMES: STUDIES OF WHOLE GENOME DUPLICATION, INTERGENIC TRANSCRIPTION AND EXPRESSION EVOLUTION IN BRASSICACEAE AND POACEAE SPECIES

By

Gaurav Dilip Moghe

Phenomena that create new genes and influence their diversification are important contributors to evolutionary novelty in living organisms. My research has focused on addressing the following questions regarding such phenomena in plants. First, what are the patterns of evolution of duplicate genes derived via whole genome duplication (WGD)? Second, do transcripts originating from intergenic regions constitute novel genes? Third, how do expression patterns of orthologous genes evolve in plants?

I have addressed these questions using comparative genomic and transcriptomic analyses of species in the Brassicaceae and Poaceae families.

To understand the evolution of WGD derived duplicate genes, we sequenced and annotated the genome of wild radish (*Raphanus raphanistrum*), a Brassicaceae species which experienced a whole genome triplication (WGT) event ~24-29 million years ago. Through comparative genomic analyses of sequenced Brassicaceae species, I found that most WGT duplicate genes were lost over time. Duplicates that are still retained were found to undergo sequence and expression level divergence.

Interestingly, while duplicate copies tend to diverge in expression level, one of the copies tends to maintain its original expression state in the tissue studied. Furthermore, duplicates that are retained in extant species tend to have higher expression levels,

greater expression breadth, higher network connectivity and tend to be involved in functions such as transcription factor activity, stress response and development.

Functional diversification of such duplicates can assist in evolution of novel characters in plants post WGD.

To understand the nature of intergenic transcription, I analyzed multiple transcriptome datasets in *Arabidopsis thaliana* as well as in species of the Poaceae family. My results suggest that plant genomes do not show any evidence of pervasive intergenic transcription. Although thousands of intergenic transcripts can be found in each species, most of these transcripts have low breadths of expression, tend not to be conserved within or between species and show a significant bias in being located very close to genes or in open chromatin regions. My results suggest that most intergenic transcripts may be associated with transcription of the neighboring genes or may be produced as a result of noisy transcription. Properties of intergenic transcripts identified in my research will be useful in distinguishing functionally relevant transcripts from noise.

To understand expression evolution, I analyzed patterns of evolution of orthologous genes between Poaceae species and found that sequence divergence is strongly associated with level and breadth of expression, and very weakly with expression divergence. Both sequence and expression evolution were found to be constrained for genes involved in core biological processes such as metabolism, transcription, photosynthesis and transport.

Overall, the results of this research are broadly applicable to the field of gene annotation and increase our understanding of evolution of gene content in plant

genomes.