The role of copy number polymorphism in shaping intraspecific structural variability of metabolic gene clusters in *Arabidopsis thaliana*

mgr inż. Małgorzata Marszałek-Zeńczak

The genomic DNA sequence is a characteristic feature of every living organism. Even the genomes of individuals within a species vary. Genetic variation ranges from single nucleotide polymorphisms (SNPs) to large structural variations (SVs). Unbalanced SVs, in which a fragment of a DNA sequence is lost or gained, are called copy number variations (CNV). Most CNVs do not have a significant impact on an individual's phenotype. However, some of them may have a deleterious effect e.g. associated with the development of disease, while others contribute to improved adaptation of the individual to environmental conditions. In eukaryotes, genes involved in a common metabolic pathway are usually dispersed throughout the genome. Nevertheless, recent investigations have identified metabolic gene clusters (MGCs) comprising non-homologous genes involved in a shared metabolic pathway. Both CNVs and MGCs are often found in dynamic regions of the genome such as centromere proximity or transposon-rich regions. Both these phenomena, although poorly understood, appear to have important implications for shaping plant genomes. In regions prone to structural rearrangements, the possibility of combining favorable sets of genes is greater than in the rest of the genome, thus promoting the formation of MGCs. CNVs and MGCs have many elements in common, and the contribution of CNVs to the formation and evolution of MGCs seems to be large and significant.

The primary objective of my research was to investigate intraspecific copy number variation in the model plant *Arabidopsis thaliana*, and to analyze to what extent CNVs affect the structure and stability of metabolic gene clusters. Using high-throughput next-generation sequencing data for more than 1,000 natural *A. thaliana* accessions, I developed a pipeline to integrate the results from seven different tools based on three main CNV detection methods: read depth, paired-end mapping and split read. I created a catalog of large indels (50-499 bp) and CNVs ($\geq$500 bp). I demonstrated that CNVs are important markers that can be used in population analyses and in genome-wide association study (GWAS). I then focused on the analysis of structural variation (mainly CNVs) in four MGCs, in the *A. thaliana* population. I observed significant diversity within these studied clusters. The marneral gene cluster appears to be fixed at the species level. The thalianol biosynthesis cluster exists in two versions. In this MGC, I identified an inversion, present in as many as 65% of the studied population, which resulted in a more compact central cluster. The compact version of the thalianol biosynthesis cluster was dominant and more conservative than the discontiguous version. The largest, highly variable and diverse in the population is the arabidiol/baruol biosynthesis cluster. In this cluster, I identified a large (21-27 kbp) genomic insertion present in about one-third of the analyzed population. This insertion introduced a new gene pair, *CYP705A2a-BARS2*, where *BARS2* was a non-reference gene encoding a previously uncharacterized oxidosqualene synthase (OSC) in the *A. thaliana* genome. GWAS indicated that accessions with this insertion displayed slower root growth dynamics and were associated with a warmer climate as opposed to accessions with the reference gene arrangement. In roots and leaves, the gene expression profile in the arabidiol/baruol cluster was different for accessions with and without the insertion. In addition, analysis of gene pairs: OSC- cytochrome P450 oxidase showed that

the gene pairs were more variable than their unpaired counterparts.

The findings of this research underscore the significant influence of genetic variability on the formation and shaping of MGCs. The high diversity of clusters in *A. thaliana* indicates their dynamic evolution while GWAS results confirm their possible role in phenotypic diversity and plant adaptation to environmental conditions. Understanding genetic variation and its impact on genome organization is crucial to gaining insights into their biological functions.