Abstract

Genomic variation in the model plant *Arabidopsis thaliana* has been extensively used to understand evolutionary processes in natural populations, mainly focusing on single nucleotide polymorphisms (SNPs). Conversely, structural variation has been largely ignored in spite of its potential to dramatically affect phenotype. Here in the first part of my thesis, I investigated general patterns of structural variations and indels among 1301 diverse *A. thaliana* accessions from Morocco, Madeira, Europe, Asia and North America using Illumina sequencing data. I validated the SVs identified and used these in the downstream analyses. I show evidence for strong purifying selection on presence absence variants (PAVs) in genes, in particular for housekeeping genes and homeobox genes, and I found an excess of PAVs in defense-related genes (R-genes, secondary metabolites) and F-box genes. This implies the presence of a 'core' genome underlying basic cellular processes and a 'flexible' genome that includes genes that may be important in spatially or temporally varying selection. Further, I find an excess of intermediate frequency PAVs in defense response genes in nearly all populations studied, consistent with a history of balancing selection on this class of genes. Then, I used environmental variables as phenotypes and showed that PAVs can assist to SNPs to identify possible functional variants.

After looking at the global pattern of the SVs, in the second part of my thesis I focused on characterizing the structural genetic variation that accumulated in the lineage of *Arabidopsis thaliana* that colonized the Cape Verde Islands from Morocco. In agreement with unpublished SNP-based analyses, I found a dramatic difference in the number of SVs between Morocco and Cape Verde such that among Cape Verdean accessions there was dramatically less variation than among Moroccan accessions. With the help of long nanopore reads, I identified two large SVs including a 100kb inversion on chromosome 1 and a 21kb tandem duplication event (repeated 5 times) on chromosome 2. These two large SVs overlap with genes related to flowering time, seed dormancy, and stress response (light, heat, drought and toxins). Then, I used Cvi-0 SVs to identify candidate functional variants underlying quantitative trait variation based on 129 traits of 47 mapping studies using the Cvi-0 x Ler-0 RIL population and found possible functional variants overlapping with genes involved in seed dormancy, drought stress, flower development and timing of flowering.

Lastly, I successfully assembled two chloroplasts haplotypes of Cvi-0 accession and I identified a 17 kb inversion between two haplotypes which overlaps with two inverted regions and a small single copy region.