#044

Understanding the impact of structural variations on gene expression using pangenome graphs

***Gözde YILDIZ*[1]**

Silvia Zanini[1]
Matthias Frisch[1]
Amine Abbadi[2]
Rod Snowdon[1]
Agnieszka Golicz[1]

[1] Justus Liebig University, Gießen, Germany

[2] NPZ Innovation GmbH, Holtsee, Germany

Background:

Structural variations (SVs) are large genomic alterations including deletions, insertions, and duplications of DNA segments (>50 bp). Due to their size, they can have a greater impact on traits than single nucleotide polymorphisms (SNPs) and smaller InDels, leading to changes in gene expression, protein function, and cellular behaviour.

Brassica napus has an allotetraploid genome (2n = 38, AACC), and different accessions harbour extensive genomic variation including SVs. Some of the SVs have been shown to affect candidate genes associated with important agronomic traits. However, this extensive variation can lead to biased variant detection and gene expression quantification. Therefore, pangenome graphs, which capture species-wide genomic variation in a single data structure, provide an excellent framework for expression quantitative trait loci (eQTL) analyses, facilitating the association between SVs and gene expression.

Objective:

The primary aim of this project is to understand the impact of SVs on gene expression and transcriptional regulation using pangenome graphs to overcome the single reference bias in oilseed rape.

Methods:

To construct the pangenome graphs, we used 57 long-read datasets from Oxford Nanopore sequencing (>5x coverage) and ~100 short-read datasets from Illumina (>20x coverage). Long reads were mapped to the B. napus reference genome (Express617) using minimap2. SVs were identified using cuteSV. A final, non-redundant SV set was used for pangenome graph construction. SVs were genotyped from short reads using Paragraph and VG call. Spliced pangenome graph was constructed with VG autoindex using non-redundant SVs, SNPs, and available Express617 annotation. RNA-Seq reads were aligned to the spliced pangenome graph reference using VG mpmap. Graph-based expression quantification performed using RPVG was compared with linear reference-based quantification performed using HISAT2 and feature Counts. EQTL analysis was performed with matrixEQTL accounting for population structure and hidden variables.

Results:

Using Oxford Nanopore data for 57 winter oilseed rape genotypes we identified almost 100,000 structural variants including 46,428 deletions and 48,396 insertions, which were used for pangenome graph construction together with 1,975,171 SNPs identified from short-read Illumina data. Pangenome graphs were used for SV genotyping from short-read whole genome sequencing and gene expression quantification for 100 winter oilseed rape lines. Our analysis revealed that a substantial proportion of variants found in long reads could not be genotyped from short reads even using pangenome graph reference. We also found systematic differences between linear reference- and graph-based gene expression quantification. EQTL analysis revealed a subset of SVs associated with gene expression.

Conclusions:

In this study, we used pangenome graphs as a framework for eQTL analysis in oilseed rape. We characterized association between SVs and gene expression elucidating the impact of larger genomic variants on gene regulation.