GENETIC VARIATION

Resolving the roles of structural variants

> nearly half of the SVs affecting coding sequences were associated with differential gene expression

Structural variants (SVs) such as large deletions, insertions, duplications and rearrangements heavily influence crop traits but are underrepresented in high-throughput short-read sequencing data. Two recent studies use novel sequencing tools to characterize SVs in crop species and explore their phenotypic consequences.

In their study, Alonge, Wang et al. used long-read sequencing to sequence a panel of 100 tomato accessions, including wild, domesticated and modern species, aligned sequencing reads to a reference genome to identify SVs and merged SV data from all accessions to produce a 'panSV' genome accounting for 238,490 SVs of >30 bp. Almost half of the SVs overlapped with genes or regulatory sequences.

The authors performed RNA sequencing (RNA-seq) on tissue from 23 of the accessions that together represented 44,358 gene-associated SVs and found nearly half of the SVs affecting coding sequences were associated with differential gene expression. To test the hypothesis that the panSV genome may reveal SVs linked to quantitative trait loci (QTLs) not detected in previous genome-wide association studies (GWAS), the authors investigated a QTL involved in the metabolism of guaiacol — a volatile compound associated with an undesirable flavour — which has been linked to the genes *E8*, *NSGT1* and *NSGT2*, although its genetic basis is unclear. Using a set of de novo genome assemblies made from short- and long-read sequencing data, the authors found novel SVs in the coding sequences of these genes that resolved into five distinct haplotypes. Haplotypes with non-functional *NSGT1* coding sequences were associated with guaiacol accumulation.

Credit: Burke's Backyard/Alamy

Further investigation of *fw3.2*, a QTL associated with fruit weight and previously linked to an SNP in the promoter of the cytochrome P450 gene *SIKLUH*, revealed a 50 kbp tandem duplication containing two copies of *SIKLUH* in some accessions. Through a combination of CRISPR–Cas9 and genetic crosses, the authors showed higher copy numbers of *SIKLUH* were associated with increased fruit weight, indicating that copy number variation — rather than an SNP — controls *SIKLUH* expression. These findings suggest pan-genome approaches can be used to resolve complex haplotypes missed by GWAS.

In a concurrent study, Liu et al. sequenced 2,898 globally sourced soybean accessions from wild soybeans, landraces and cultivars, mapped these to a reference genome and sorted the species phylogenetically based on SNP analysis and geographical distribution. The authors chose 26 accessions that best represented the variation in soybean species, sequenced these using single-molecule real-time sequencing and assembled sequences de novo, identifying >750,000 SVs.

To investigate whether changes in gene structure result in the phenotypic diversity seen across soybean strains, the authors performed a GWAS investigating the genetic basis of seed luster using their SV data and identified a 10 kb SV on chromosome 15 as critical for regulating seed luster. Further investigations into *E3*, a gene associated with flowering, identified large insertions and deletions that accounted for multiple haplotypes linked to different flowering phenotypes. Deletions in parts of *E3* and the related *SoyZH13_19G210600* gene were associated with transcription read-through of these genes and expression of a novel transcript. These data confirmed that SVs influence agronomic phenotypes and suggest that SVs can induce gene fusion events.

Using RNA-seq data taken from their 26 soybean accessions, the authors aimed to correlate gene expression with the presence of SVs. They identified a large deletion in the promoter region of *SoyZH13_14G179600*, a gene associated with a QTL for iron deficiency. Accessions could be categorized into two distinct groups (Hap-1 and Hap-2) based on whether they had this deletion. These haplotypes were differentiated by the expression of *SoyZH13_14G179600* and geographical location, which suggests that genetic divergence of this gene contributed to soybean adaptation of iron uptake. These data show that studying pan-genomes can help identify functional, evolutionarily relevant SVs.

Taken together, these studies reveal the impact of SVs on crop genomes and phenotypes. The platforms used in these studies will allow further insights into the functional and evolutionary relevance of SVs.

Joseph Willson

ORIGINAL ARTICLES Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* https://doi.org/10.1016/j.cell.2020.05.021 (2020) | Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* https://doi.org/10.1016/j.cell.2020.05.023 (2020)