Exploring Gene Family Diversity with PANSCOPE and SLAC: Bridging Pan-Genomics and Manual Curation with Alignment Thumbnails

Zujic R1, Hastwell A1, Prabhuram D1, Udvardi M2, Ferguson B1

*rzujic@gmail.com*

1 School of Agriculture and Food Sustainability, The University of Queensland, St Lucia, AUS

2 Centre for Crop Science, Queensland Alliance for Agriculture and Food Innovation, St Lucia, AUS

Growing numbers of published plant genomes are fuelling the potential discovery of valuable intra-species gene variants. However, published gene-based pan-genomic data may inherit annotation assumptions that reduce their utility to gene family experts. This is particularly impactful for gene families that still involve human-aided sequence curation for sensitive detection, such as those encoding short peptides. Exceptional gene transcription patterns, such as narrow permissive conditions and/or low total expression may further obscure these genes from study in pan-genome gene sets guided by transcriptomic evidence.

Our bioinformatics toolset, PANSCOPE, is a generalised multi-genome, gene-centric search pipeline that aims to help bridge the gap between current pan-genome analysis methods, and manual curation workflows familiar to molecular geneticists. It demonstrates how a gene-structure-aware BLAST-based search pipeline can be integrated with a novel system for hit visualisation and interactive exploration. Combined, this facilitates sensitive, efficient and user-friendly exploration of a gene family query set’s sequence diversity across an arbitrary set of related genomes.

Of broader utility, we introduce Single Line Alignment with Context (SLAC) encoding. This system enables three aligned DNA sequences to be represented in a single or even thumbnail-scale text preview. The system is designed to communicate alignments of a hit sequence against a query gene’s genomic and coding sequences and holds potential to improve rapid human readability of gene DNA sequence variants and hit patterns.