Construction Pan-genome Graph for Peanut species

X. WANG*[1,2], J. VAUGHN [1,2], [1]Genomics and Bioinformatics Research Unit, USDA-ARS; University of Georgia, CAGT Building, 111 Riverbend Rd, Athens, GA 30602; [2] Genomics and Bioinformatics Research Unit, USDA-ARS; University of Georgia, CAGT Building, 111 Riverbend Rd, Athens, GA 30602.

For roughly three decades, bioinformatics has operated under a single-reference genomic paradigm. This approach was appropriate for individuals that are closely related to that single reference. As individuals diverge at the sequence level, fewer and fewer sequencing reads align correctly or at all to this reference, severely limiting modern genetic methods. In plant breeding, divergent material is often a major source of novel traits. Such exotic material can even belong to other genera. Indeed, the usage of exotic material may have been curtailed within the last few decades *not* because its agronomic value has lessened but because such germplasm confounds the ability to develop markers and understand their segregation patterns within a single-reference genomic paradigm.

Our recent work with peanut illustrates this point.

Biological sequence alignment seeks to create a data object that pairs every base in two sequences that were the same base in a common ancestor sequence. Chromosome-scale sequence alignments have a comparable goal, but this goal is often complicated by structural rearrangements and duplications. Duplications create ambiguous alignments because a base in one sequence should be matched to two bases in the other. Linear alignment methods must choose one match over another, and they do this based on gross collinearity with the surrounding sequence. In the final alignment, this ambiguity is lost.

Emerging graph data objects can represent these relationships without having to “choose” one representational over another. In this way, these graphs allow all ancestor relationships to be presented and all reads from any individual in the graph to be mapped entirely. While conceptually useful, such a representation is often difficult to adapt to genetic methods using meiotic recombination to associate subsequences in the genome with phenotypes of interest. Within a trait-associated locus, the evolutionary relationship between repeat elements that have expanded to hundreds of copies is fairly irrelevant, but the presence of such a repeat in a gene might be critical. By carrying such evolutionary relationships along in the graph structure, many current approaches are operational intractable and difficult to interpret genetically.

We feel the central utility of a graph is not that it elegantly reflects and compresses all structural variability but that it can support 100% full-length short-read alignments (ignoring contamination and chimeras). This is opposed to a linear reference, which can often support >90% read alignment but only by splitting reads. Such splits are difficult to interpret and often lead to false variation calls. Moreover, a single reference inevitably contains single-copy regions that have duplicated or triplicated in an unrelated sample. Short-read coverage of these regions is often very difficult to threshold given the general variability of true single-copy coverages. Such scenarios conspire to generate more false variation.

With this is mind, our goals were: 1) Find an optimum between breaking a chromosome-scale alignment into sub-alignments at positions of major structural change, 2) Retain positional stringency within those sub-alignments, 3) Retain coordinates of the underlying sequences, 4)

Maximize high-quality short-read mapping, 5) Visualize results relative to original sequence coordinates and multi-chromosome alignments. To achieve these goals, we first used a positional homology method for large-scale alignment that allows for rearrangement. These alignments are then hierarchically sorted based on underlying sequences in the alignment: sequences that are either community standards and/or have the highest quality can be prioritized by the user. Graph data objects are then generated from these alignments. In this way, we are able to account for effectively all structural variation when aligning reads, while still organizing the result in a manner that is applicable to further genetic interrogation.