Using a Real-Time NGS Platform for Interpretative Analysis of Complex Plant Genomes

Shawn Quinn, CTO, Curio Genomics

Jacob Enk, R&D Manager, NGS Division, Arbor Biosciences

ABSTRACT

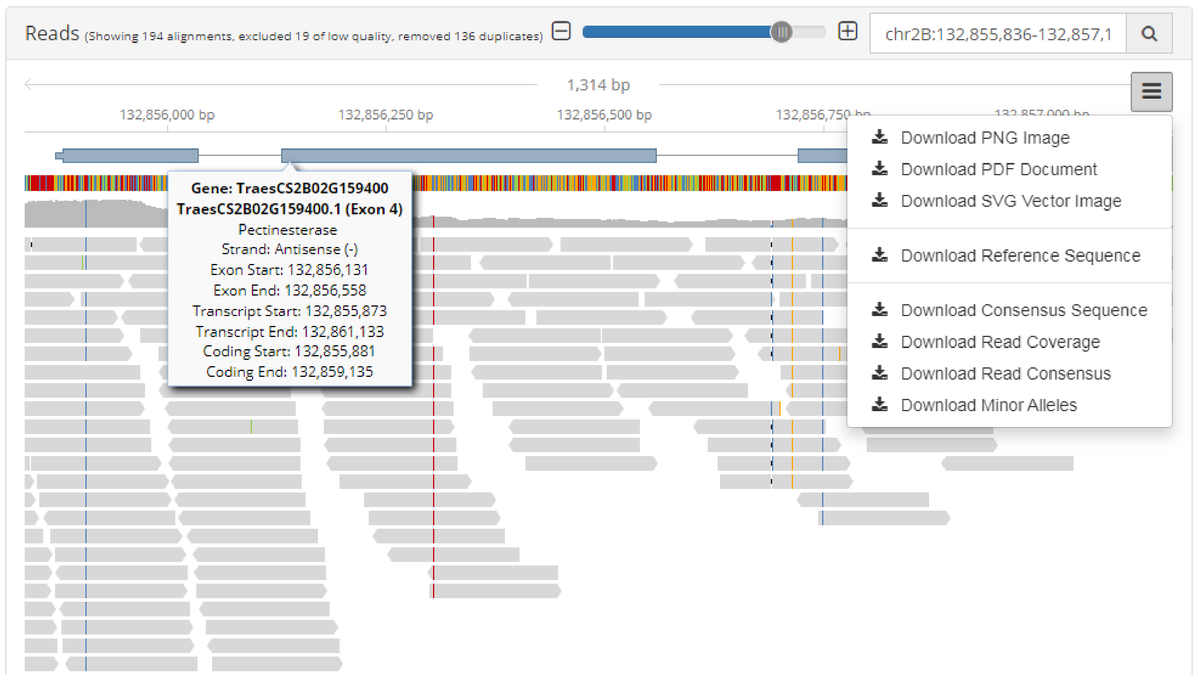
Crop genomes are often complex and analysis with traditional bioinformatics tools proves difficult. Leveraging decades of big data software development experience, we describe the utility of a fully scalable, real-time, NGS analysis platform for analyzing complex crop genomes. This advancement unleashes new opportunities for researchers seeking to address critical global challenges to develop more productive and resilient crops in a time of climate change and further population growth.

Among the crop genomes, none are more challenging to analyze than the Chinese Spring Wheat (Triticum aestivum) genome due to its large size, 85% repeating sequences, and allopolyploid nature. The publication of the wheat genome reference and related annotations by the International Wheat Genome Sequencing Consortium (IWGSC) has made the analysis of the wheat genome feasible, in theory.

Here we demonstrate how we overcame the challenges of read mapping (both for DNA-Seq and RNA-Seq libraries) and read alignment visualization when dealing with the "large chromosome" complexity of the wheat genome. Additionally, we show a novel approach to variant calling, coverage analysis, and gene expression calculation in hexaploid species. Including the dynamic incorporation of the IWGSC reference and annotation sets, we share several research examples developed using the Curio Genomics platform (www.curiogenomics.com). Based on collaboration with IWGSC members, we highlight powerful interpretive results and data visualizations, including an approach for filtering variants by predicted biological consequences.

NAVIGATING SEQUENCES & IWGSC ANNOTATIONS

Browsing and visualizing aligned reads from samples of any size, dynamically incorporating IWGSC transcriptome and functional annotation information, and providing quick access to reference and consensus sequences via real-time cluster and database technology.



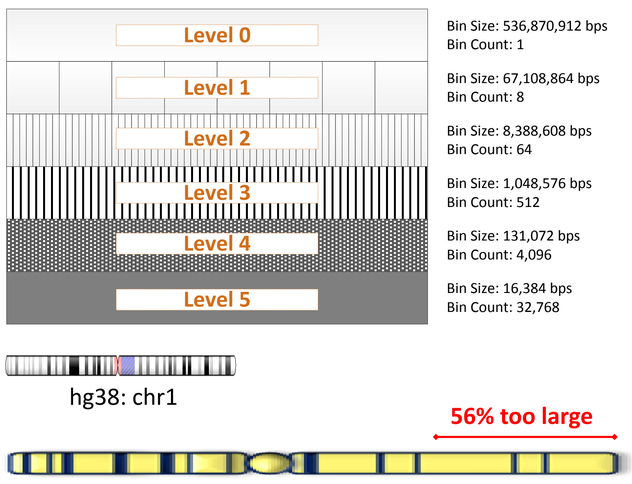
** (All screenshots shown are taken from the Curio Genomics platform)*

FEATURE INDEXING

The standard binning strategy used in many bioinformatics software tools is based on the SAM/BAM specification which only supports chromosomes up to 536 million base pairs in length. This causes a problem with many plant genomes given the larger size of the chromosomes, and therefore Curio leverages a dynamic binning strategy throughout the entire tool pipeline and data visualizations.

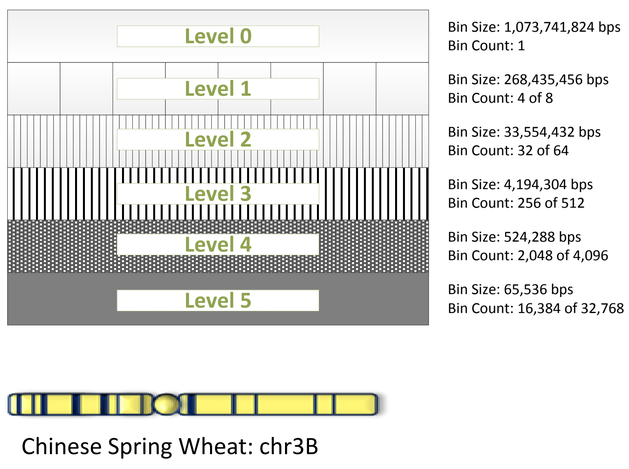
Standard Binning Strategy, Not Ideal

- Binning strategy used in BAM index files (i.e. a "BAI" file)
- 6 levels deep
- Max bin size: **536 million bases**
- Largest human chromosome (chr1): **249 million bases (fits fine)**
- Largest CSW chromosome (chr3B): **837 million bases (no good)**
- Forces split chromosome approach or causes numerous tool compatibility issues



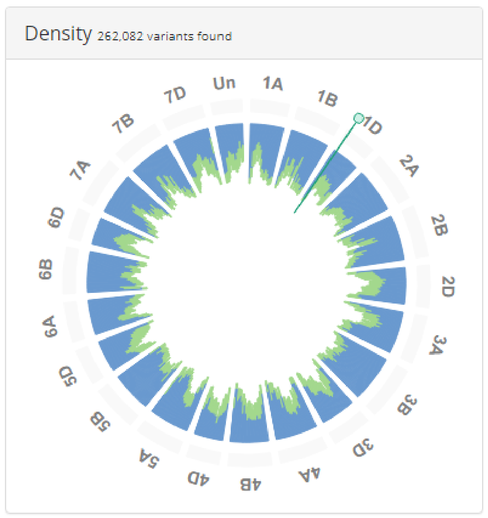
A Better Binning Strategy

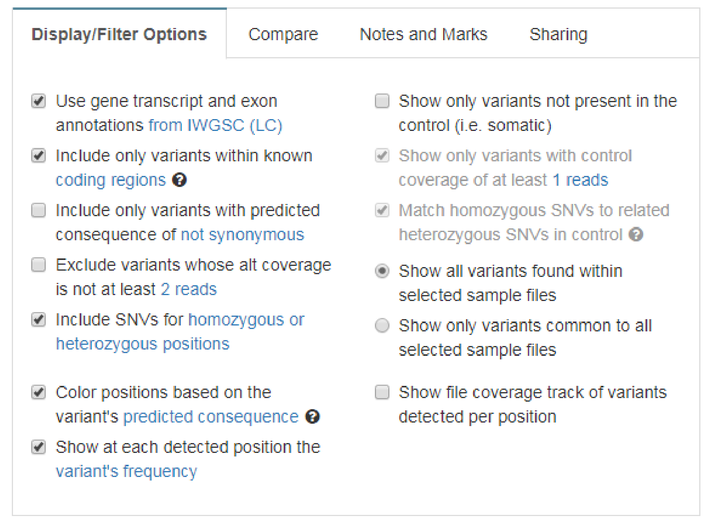
- Dynamic binning strategy based on species & assembly
- For wheat utilize a strategy based on the coordinate-sorted index specification with a minimum bit shift of 16
- Maintain 6 levels & use first half of each
- Utilize same number of bins per level
- Max bin size: **1 billion bases**
- Largest CSW chromosome (chr3B): **837 million bases (great fit)**

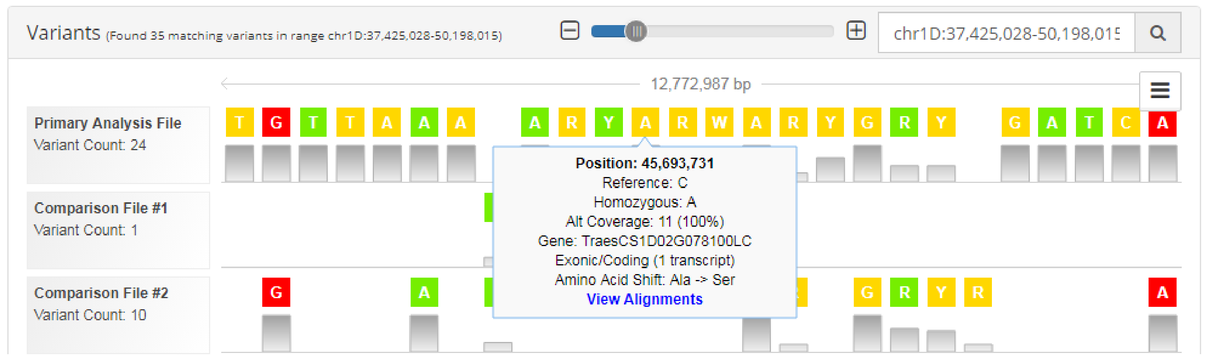


DNA-Seq: VARIANT ANALYSIS

Calling, navigating, comparing, and filtering of variants in multiple hexaploid wheat samples by predicted biological consequence leveraging IWGSC annotations - dynamically adjusted including choice of "High Confidence" or "Low Confidence" genes.

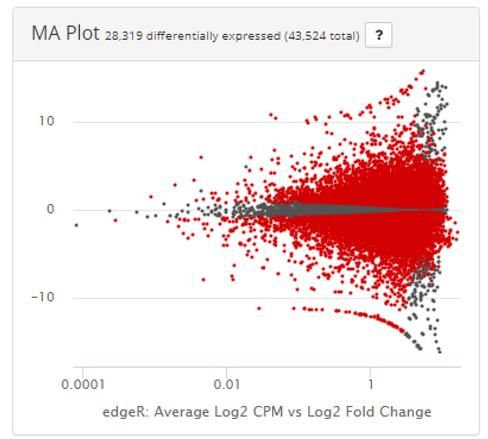


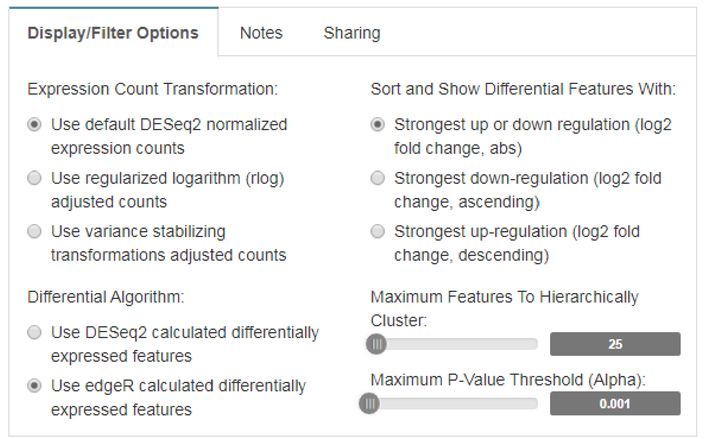


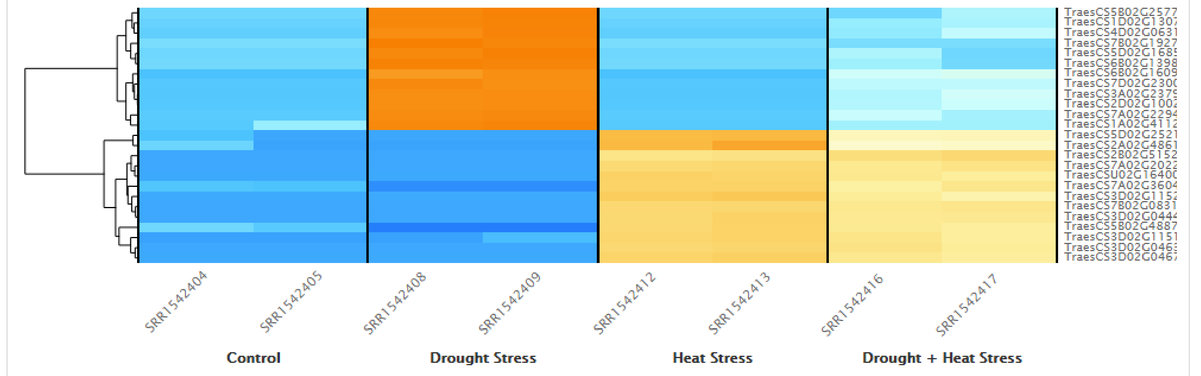


RNA-Seq: EXPRESSION ANALYSIS

Calculating gene/transcript/exon expression and analyzing for differentially expressed genes in hexaploid wheat samples – dynamically comparing algorithms such as edgeR and DESeq2 on the fly.

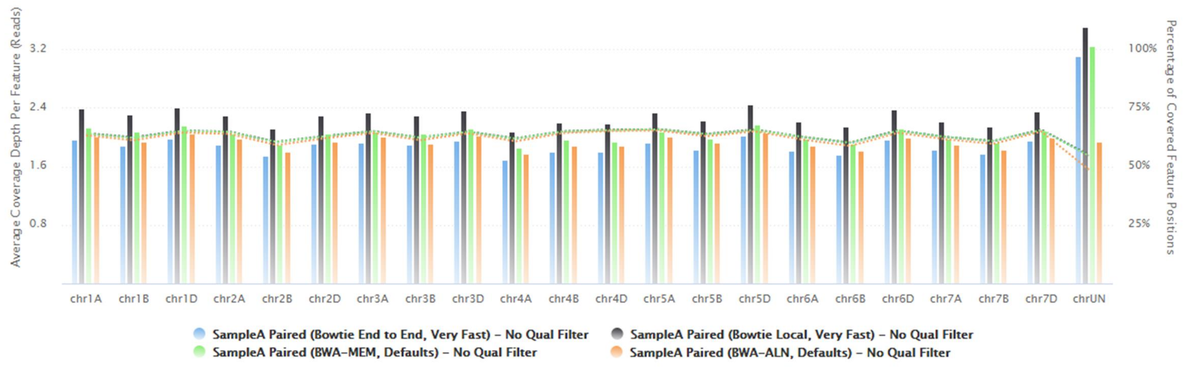






READ MAPPING / COVERAGE / QUALITY ANALYSIS

Measuring library / kit exome coverage, analyzing alignment algorithm impacts, and accounting for mapping quality in allopolyploid organisms.



ACKNOWLEDGEMENTS

Special thanks to collaborators at:

- International Wheat Genome Sequencing Consortium (IWGSC)
- French National Institute for Agricultural Research (INRA)
- John Innes Centre (JIC)
- University of Adelaide, Plant Genomics Centre

Expression analysis utilizes FASTQ data from: Liu Z, et al. Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (*Triticum aestivum* L.) BMC Plant Biol. 2015;15:152. doi: 10.1186/s12870-015-0511-8.