Leveraging the IWGSC Gene Annotations with Additional Cultivars in Curio

Presented at the 2023 Plant and Animal Genome Conference

Author: Shawn Quinn (CTO, Curio Genomics)

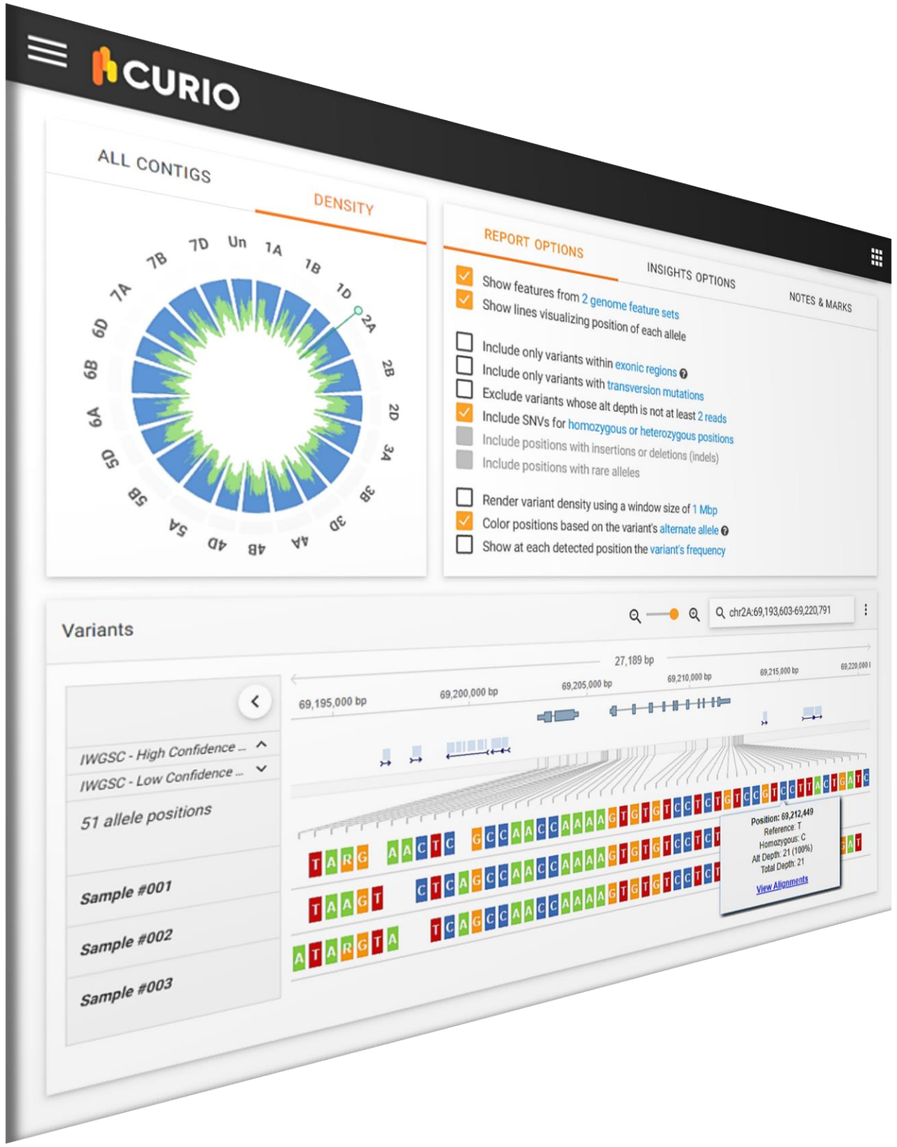
Leveraging the IWGSC Gene Annotations with Additional Cultivars in Curio

Presentation Overview

- What is Curio
- Leveraging the 10+ Wheat Genomes Project
  - Read Mapping/Alignment Impact
  - Impact on Variant Calling
  - Leveraging Projected IWGSC Structural and Functional Annotations
  - Sample Analysis Approach / Identifying Genes of Interest
- Looking Ahead
- Acknowledgements

What is Curio?

- Modern big data management and genomic analysis platform, fully web-based, collaboration ready
- Supports <u>both</u> bioinformatics processing and scientific interpretive analysis
- Massively scalable data processing and interactive data visualizations using real-time databases and clustering technologies
- Designed for extensibility to continuously support new analysis methods, data types, etc.
- Includes robust crop research analysis solutions, including tetraploid and hexaploid wheat DNA-Seq and RNA-Seq analysis



10+ Wheat Genomes Project

- 9 assemblies that include pseudomolecules (including one for the “unknown” chromosome)
- 5 other assemblies that are scaffold-level only
- Gene projections based on the IWGSC annotations
- Gene/Transcript IDs structured like: “Traes<Cultivar><Contig>01<Feature Id>”
- Example: TraesARI2A**01**G145700
- Translation table mapping to IWGSC RefSeq 1.1 gene/transcript IDs available at: https://galaxy-web.ipk-gatersleben.de/libraries/folders/F1cd8e2f6b131e891

https://10wheatgenomes.com/

Nine Available Cultivars w/ Pseudo Molecules




Wheat Cultivar	Region	Pedigree	Version
ArinaLrFor	Switzerland	Arina*3/Forno	3.0
Jagger	USA	KS-82-W-418/Stephens	1.1
Julius	Germany	Asketis/Drifter	1.0
Landmark	Canada	Unity/Waskada//Alsen/Superb	1.0
LongReach Lancer	Australia	VI184/Chara//Chara/3/Lang	1.0
Mace	Australia	Wyalkatchem/Stylet//Wyalkatchem	1.0
Norin 61	Japan	Fukuoka Komugi 18/Shinchunaga	1.1
Stanley	Canada	CDC Teal//EE8/Kenyon35//AC Barrie	1.2
SY Mattis	France	Apache/Intense	1.0





Ensembl Plants

- Provides curated versions of the same 14 (9 + 5) cultivars from the 10+ Wheat Genomes Project
- Assemblies include individual hundreds of thousands of scaffolds not assigned to pseudomolecules, instead of a single “unknown” pseudomolecule
- Updated de-novo gene annotations, which were processed by the PGSB (Plant Genome and Systems Biology)
- Gene/Transcript IDs structured like: “Traes<Cultivar><Contig>03<Feature Id>”
- Example: TraesARI2A**03**G00626940
- Translation table to IWGSC gene/transcript ids is not yet available

EnsemblPlants

https://plants.ensembl.org/Triticum_aestivum/Info/Strains

Projected and De-Novo Genes Annotations

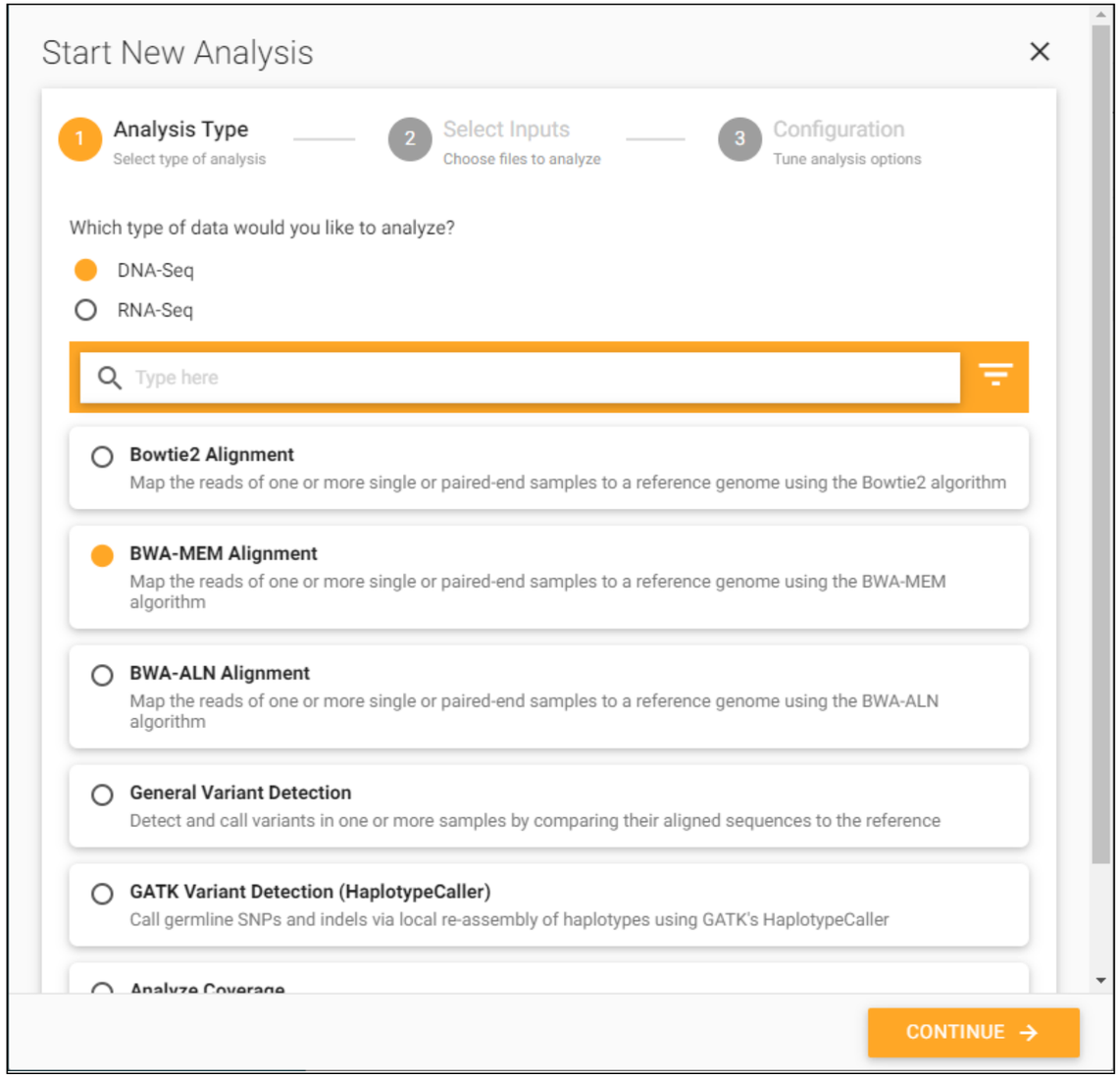
Wheat Cultivar	Original Version (#1)	Updated Version (#2)
ArinaLrFor	1,231,314 features	2,213,22 features
Jagger	1,221,309 features	2,110,846 features
Julius	1,217,757 features	2,289,641 features
Landmark	1,216,606 features	2,166,879 features
LongReach Lancer	1,225,483 features	2,079,609 features
Mace	1,223,663 features	2,085,425 features
Norin 61	1,219,068 features	2,188,731 features
Stanley	1,220,222 features	2,128,177 features
SY Mattis	1,230,268 features	2,138,635 features

#1: Original version based on projections of the IWGSC High Confidence structural annotations
#2: Updated de-novo gene annotations, as of the “release 54” version of the plants.ensembl.org build

Wheat DNA-Seq:
Read Mapping and Navigation

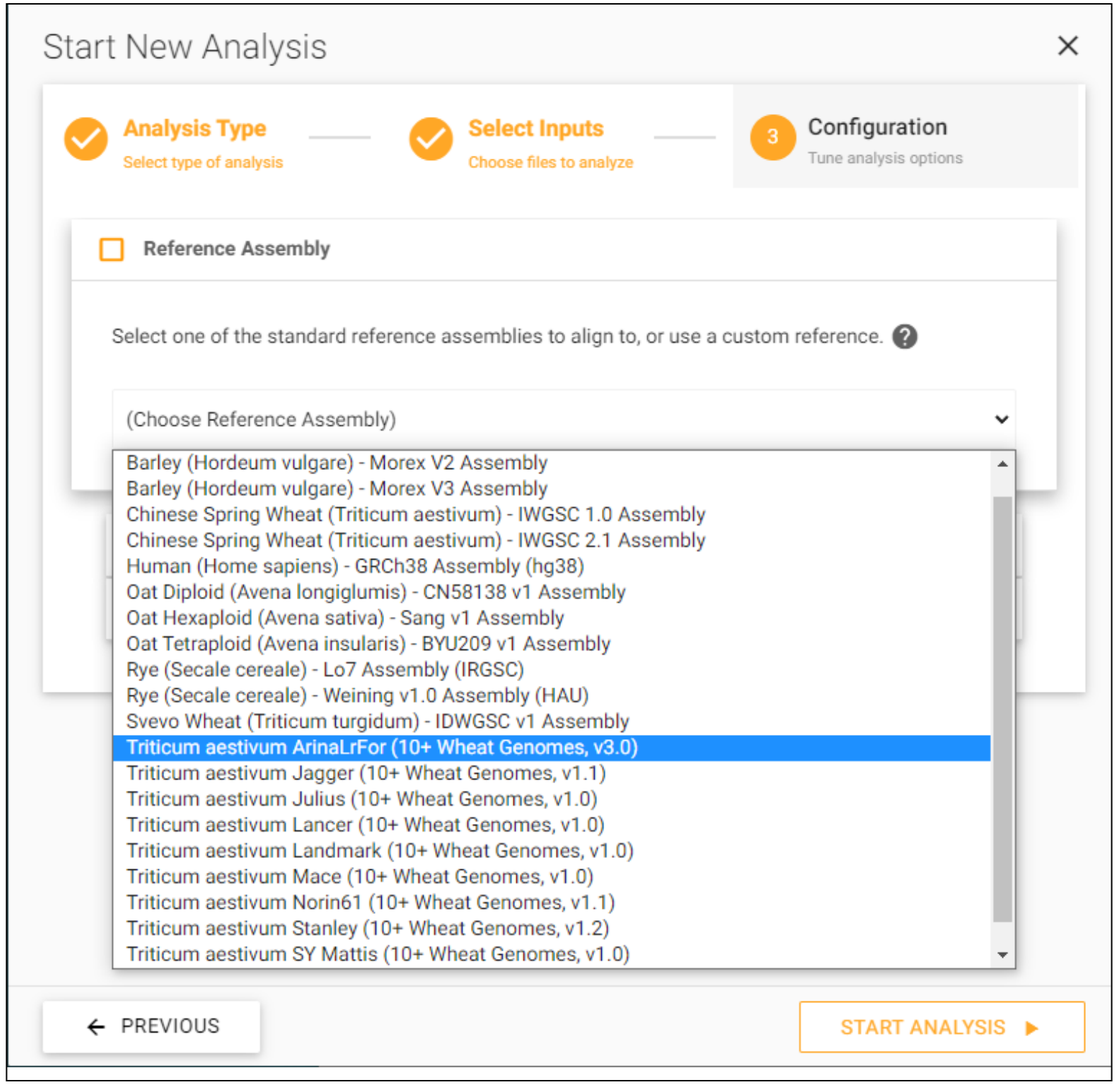
Wheat Read Mapping

- Multiple read mapping algorithms with pre-built indexes that are deployed and ready on computational cluster



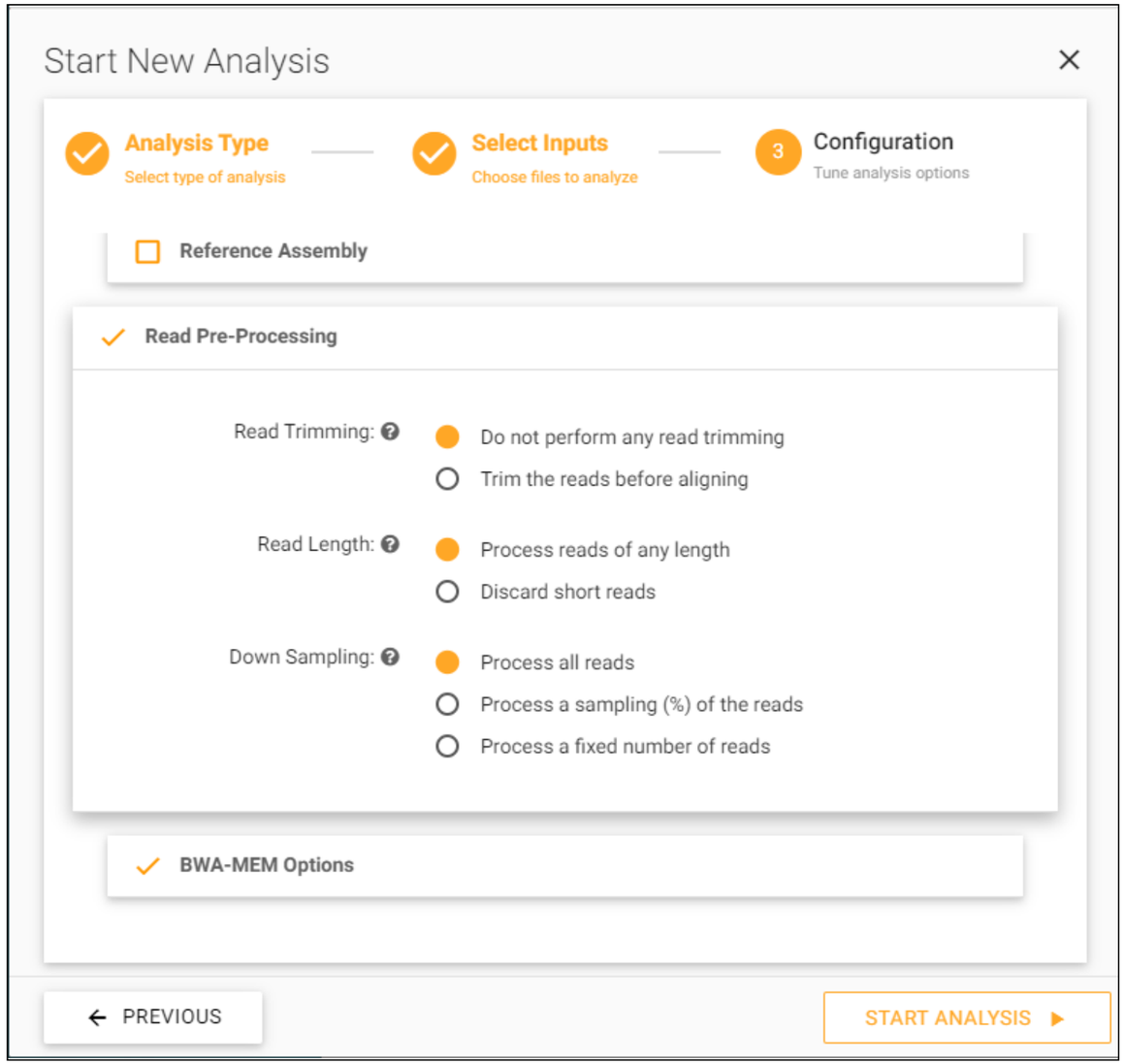
Wheat Read Mapping

- Multiple read mapping algorithms with pre-built indexes that are deployed and ready on computational cluster
- Conveniently curated and instantly available assemblies from the “10+ Wheat Genomes” project and others
- Automatic support for custom reference assemblies



Wheat Read Mapping

- Multiple read mapping algorithms with pre-built indexes that are deployed and ready on computational cluster
- Conveniently curated and instantly available assemblies from the “10+ Wheat Genomes” project and others
- Automatic support for custom reference assemblies
- Experiment with various alignment and read processing options without requiring any pipeline configuration

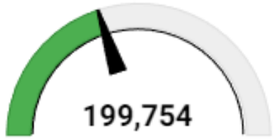


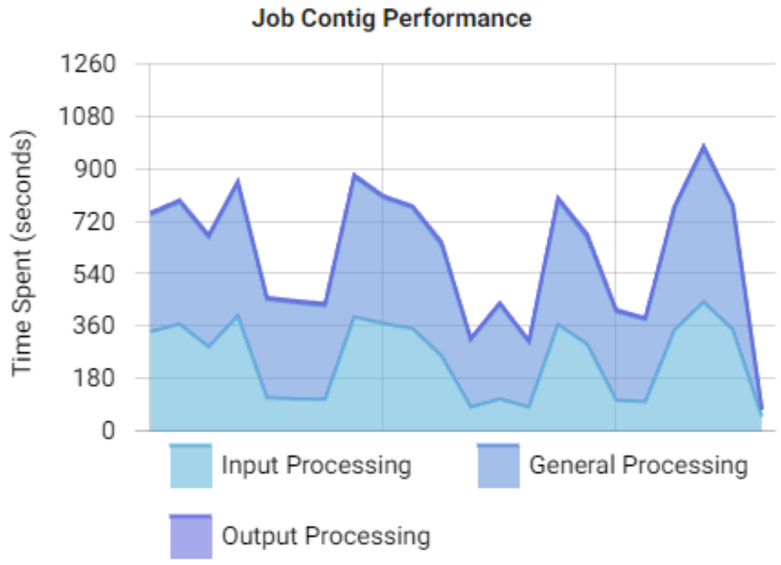
Parallel Read Mapping to ArinaLrFor Genome

Record Count:	**158,487,099 alignments**
Total Elapsed Clock Time Spent:	**13 minutes, 13 seconds**
Parallel Cluster Acceleration:	**3,364%**

Processing Rate (Records / Second)

Overall Job Rate (Records / Second)



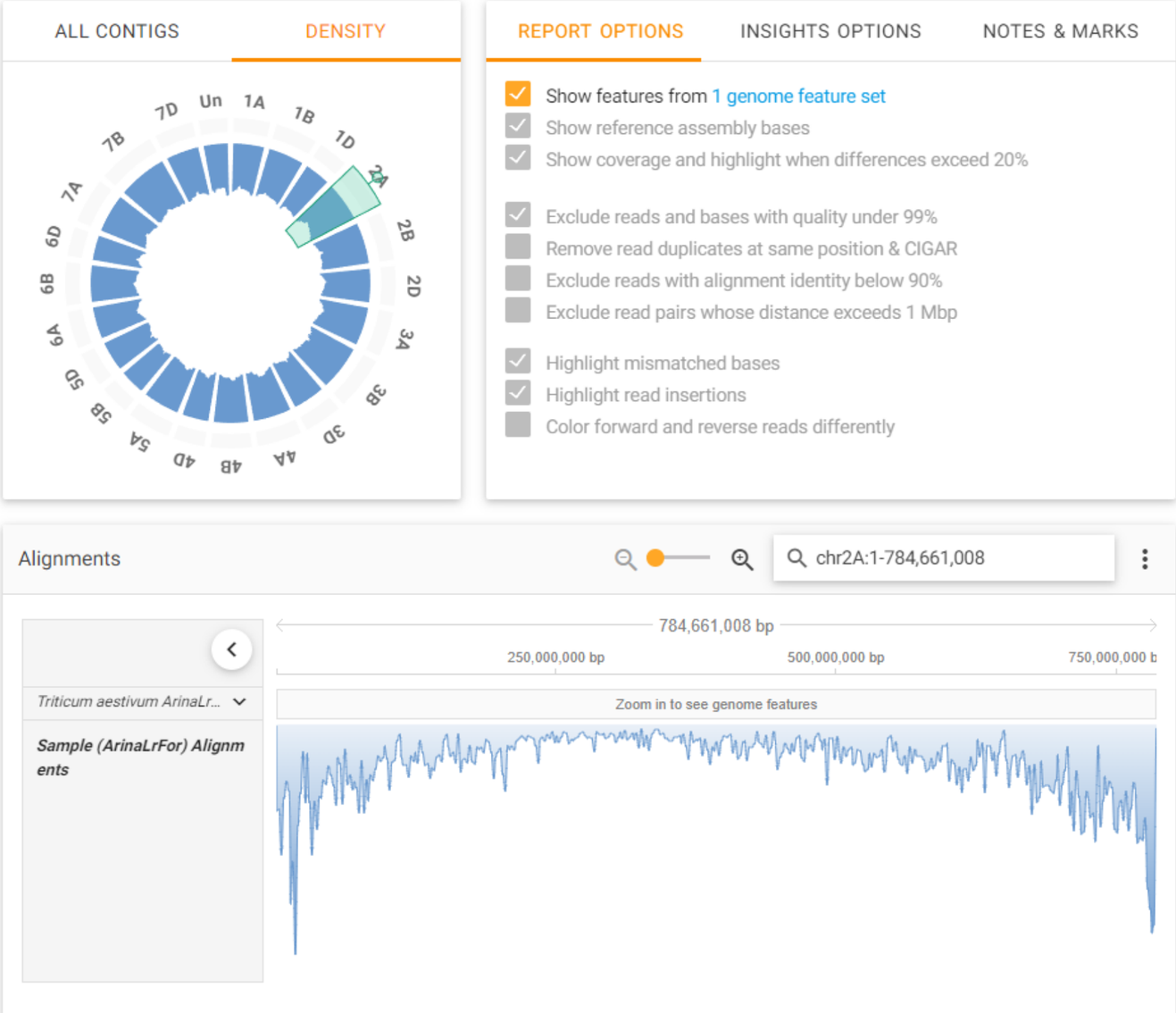


Number of Job Parts:	**601 parts**
Part Network Read Time Total:	**2 hours, 8 minutes**
Part Analysis Processing Time Total:	**4 hours, 23 minutes**
Part Output Processing Time Total:	**53 minutes, 10 seconds**
Total Compute Time for All Parts:	**7 hours, 24 minutes**

Number of Job Contigs:	**22 contigs**
Contig Input Processing Time Total:	**1 hour, 31 minutes**
Contig Record Processing Time Total:	**2 hours, 11 minutes**
Contig Output Processing Time Total:	**2 minutes, 28 seconds**
Total Compute Time for All Contigs:	**3 hours, 45 minutes**

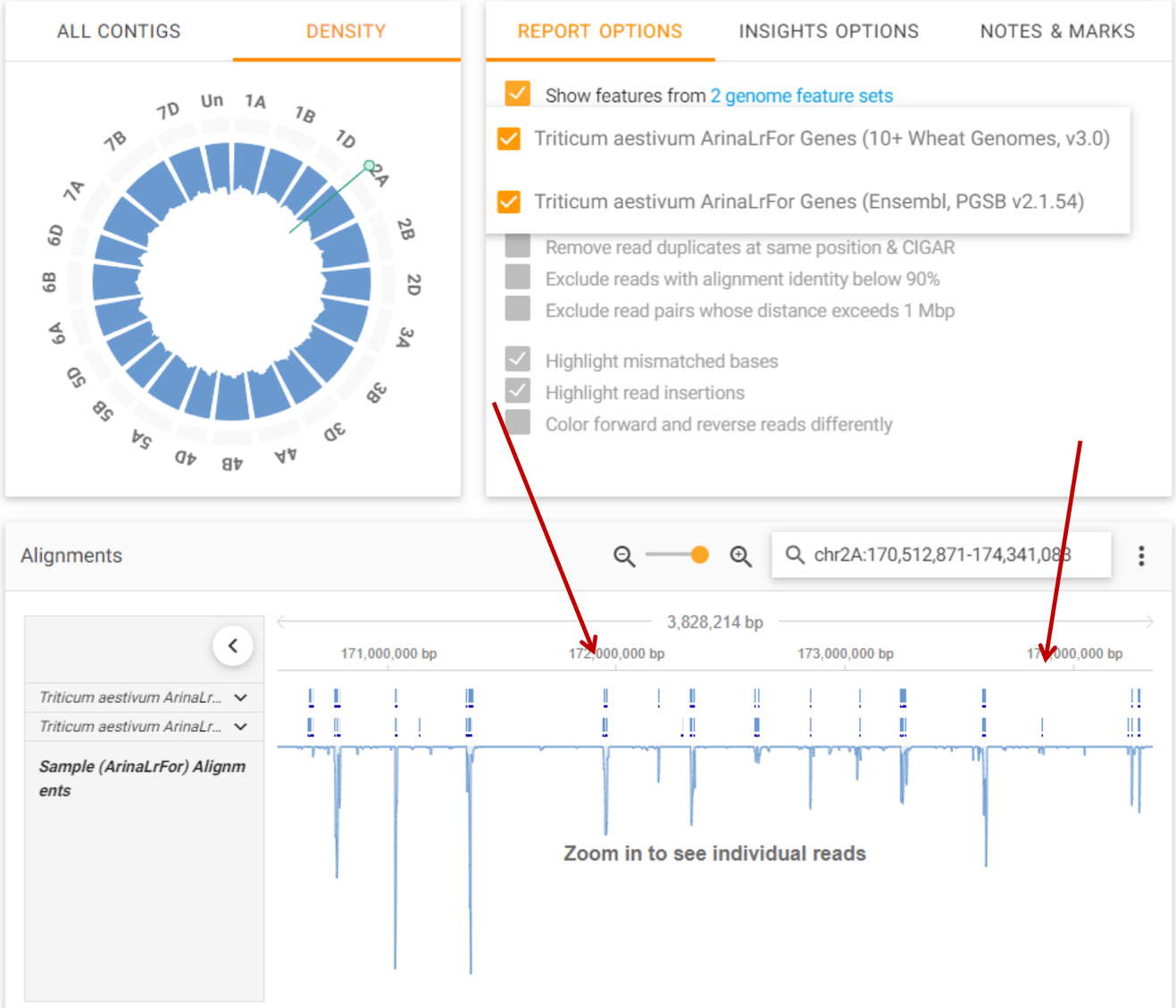
10+ Wheat Genomes: Alignment Visualization

- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome



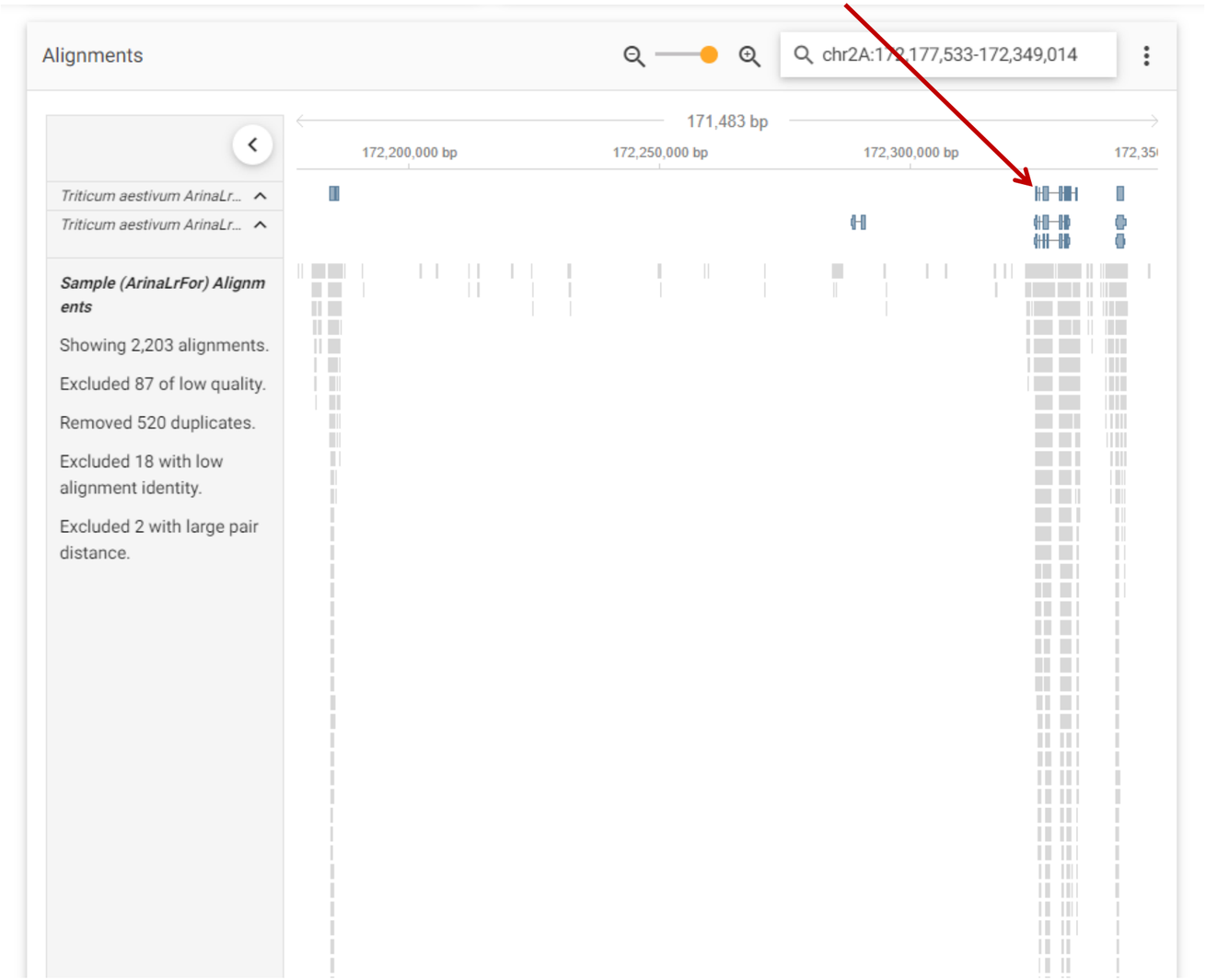
10+ Wheat Genomes: Alignment Visualization

- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes



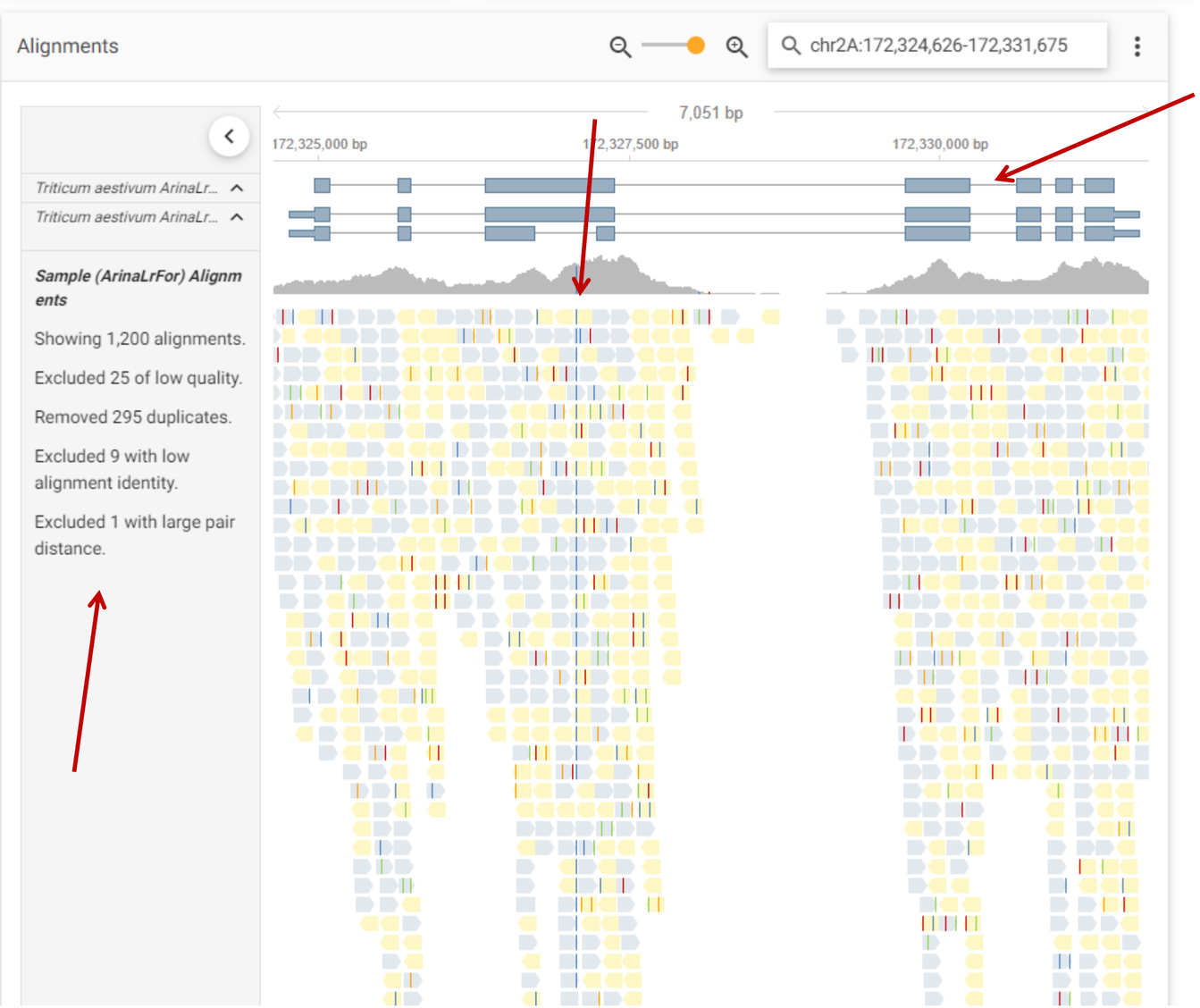
Bread Wheat: Alignment Visualization

- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes
- Quickly analyze areas of interest



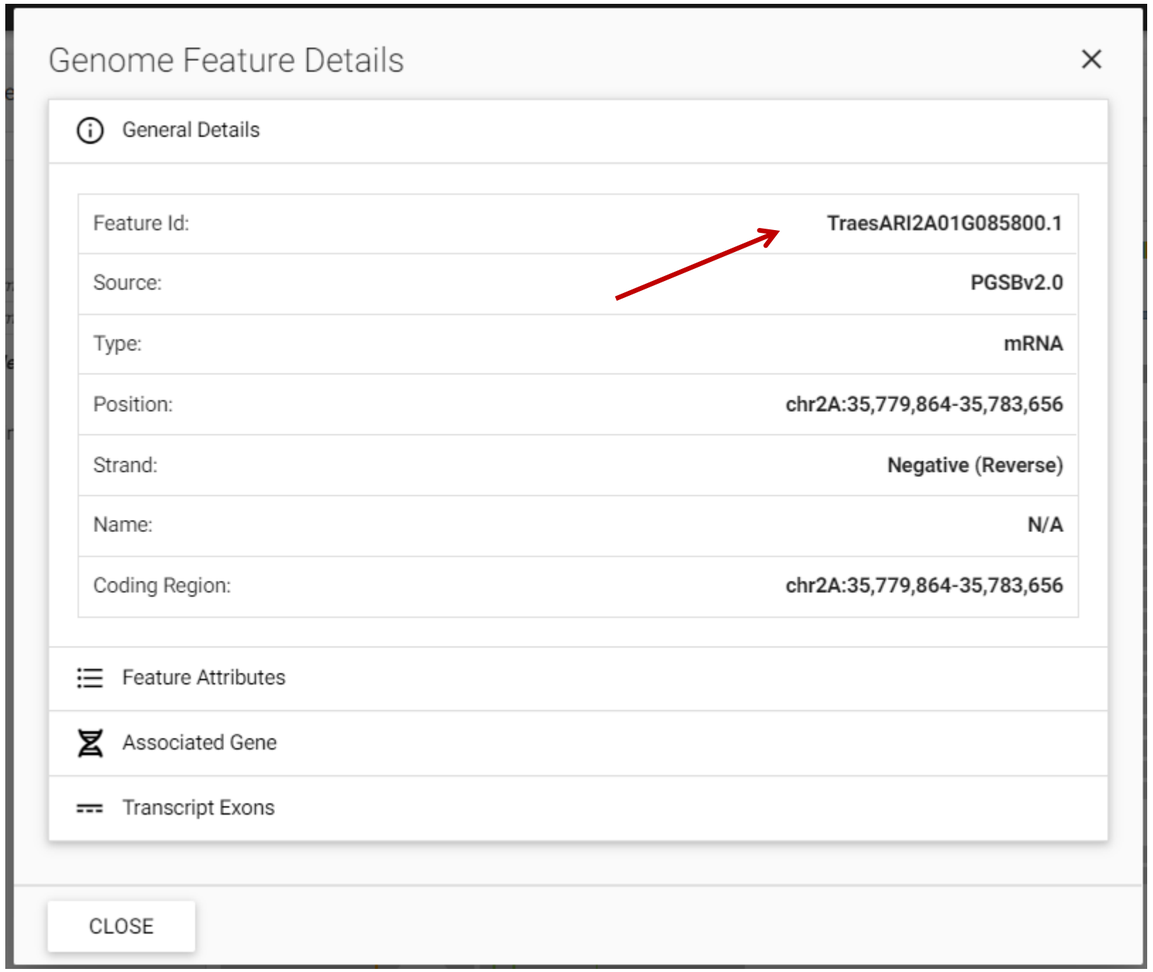
Bread Wheat: Alignment Visualization

- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes
- Quickly analyze areas of interest
- Dynamically visualize impact of filtering options
- Access gene transcript details



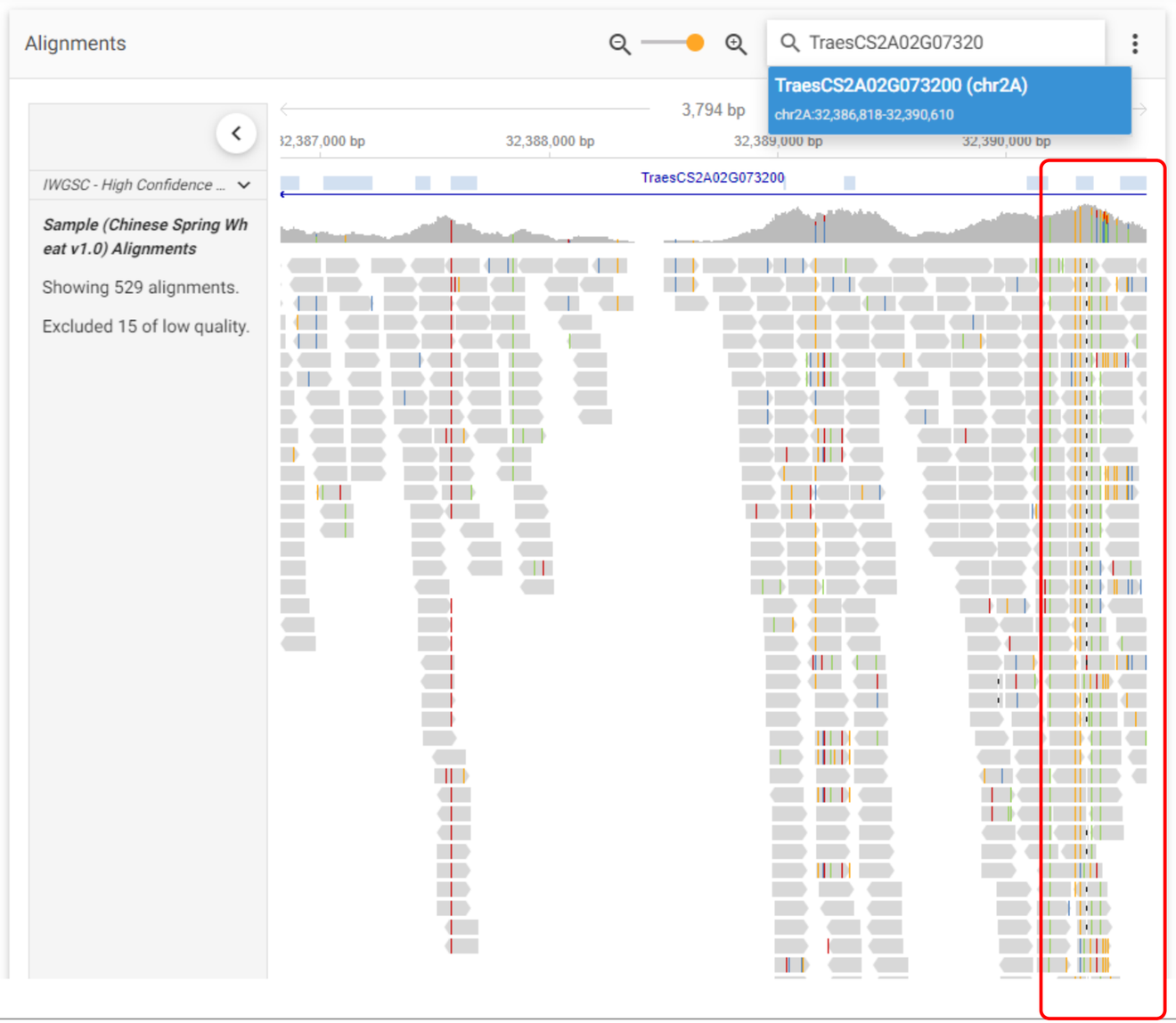
Bread Wheat: Alignment Visualization

- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes
- Quickly analyze areas of interest
- Dynamically visualize impact of filtering options
- Access gene transcript details
- Identified in ArinaLrFor as gene id “TraesARI2A01G085800”



Bread Wheat: Alignment Visualization

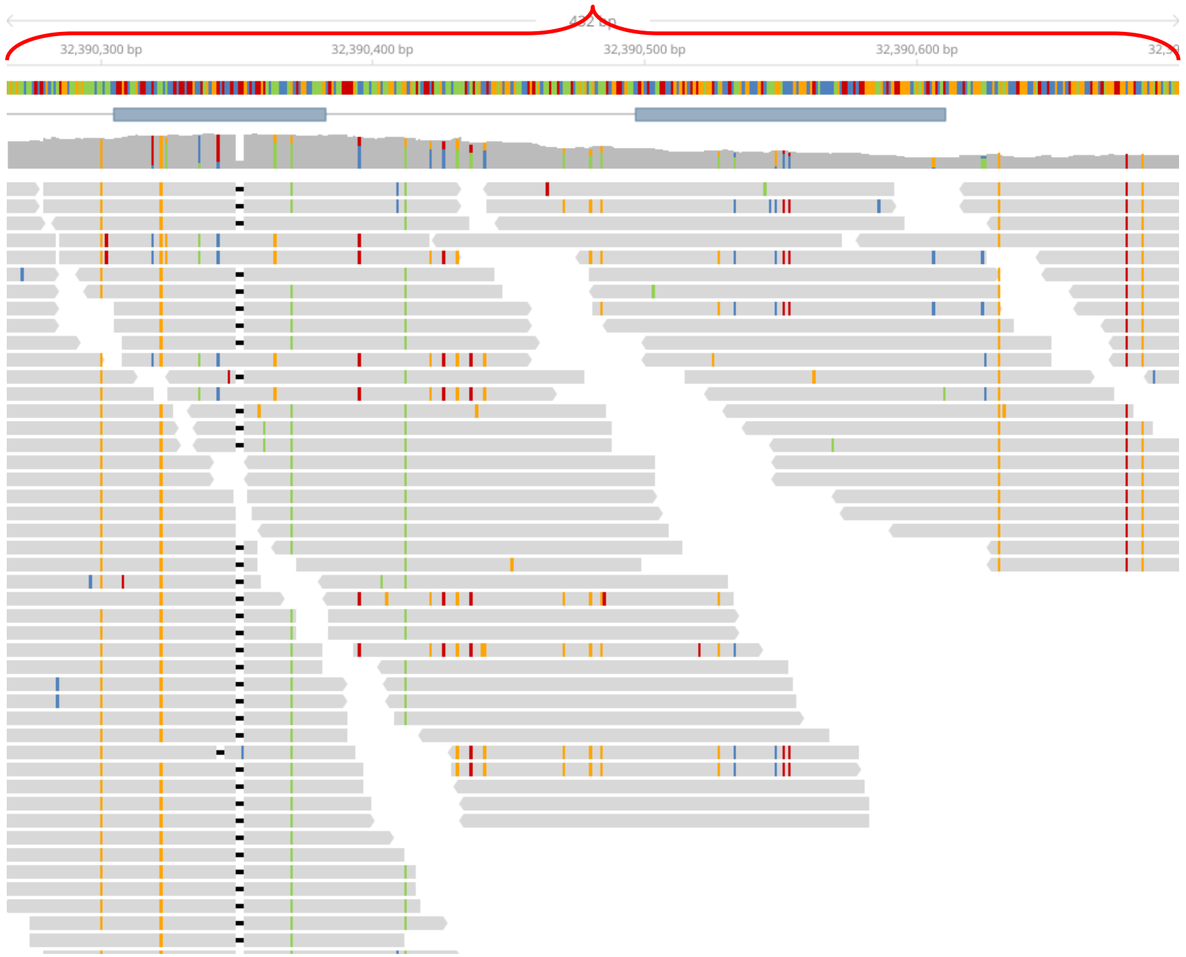
- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes
- Quickly analyze areas of interest
- Dynamically visualize impact of filtering options
- Access gene transcript details
- Identified in ArinaLrFor as gene id “TraesARI2A01G085800”
- IWGSC equivalent: TraesCS2A02G073200
- Landmark equivalent: TraesLDM2A01G082900



Comparing “10+ Wheat Genomes” Read Mapping

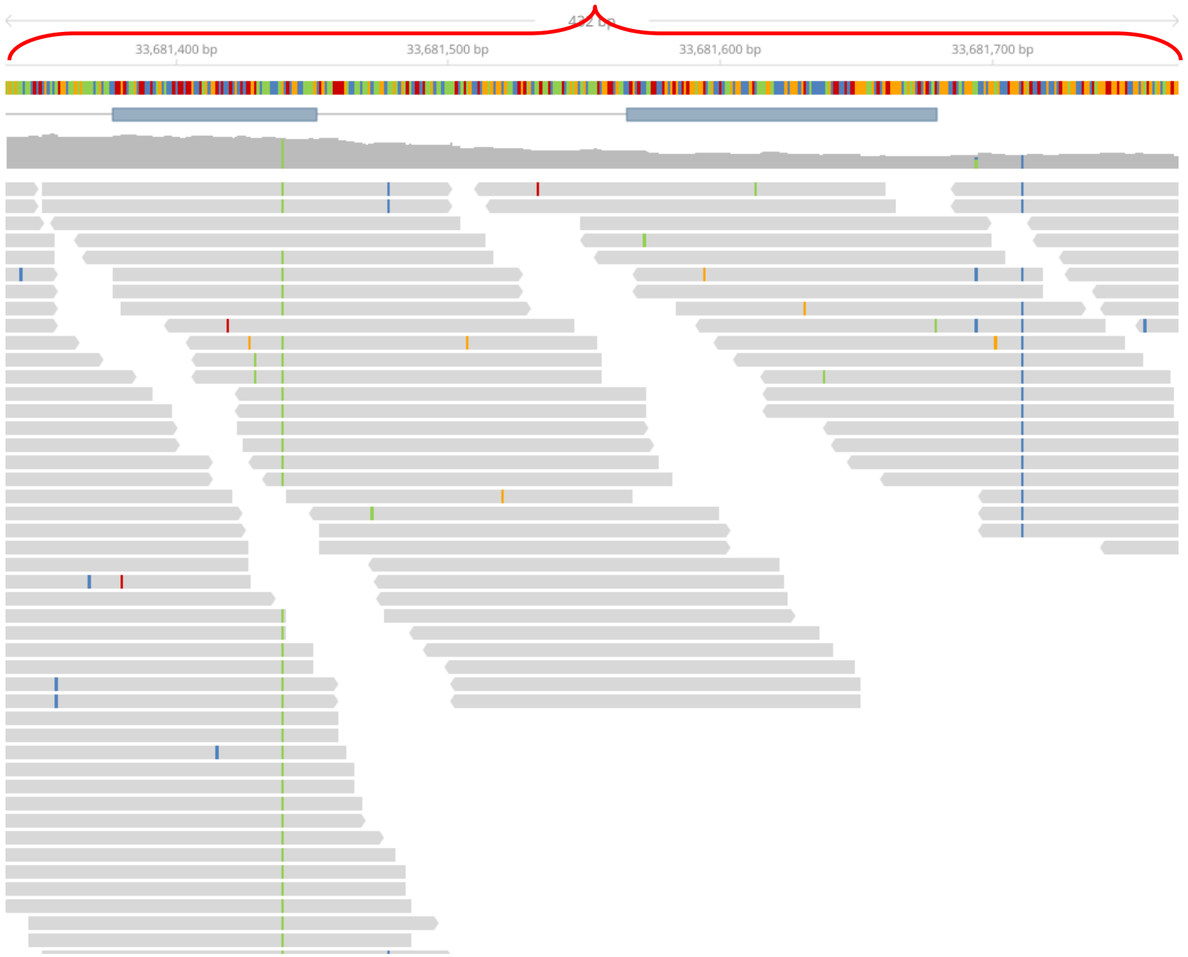
IWGSC “TraesCS2A02G073200”

(Showing: chr2A:32,390,266-32,390,696)



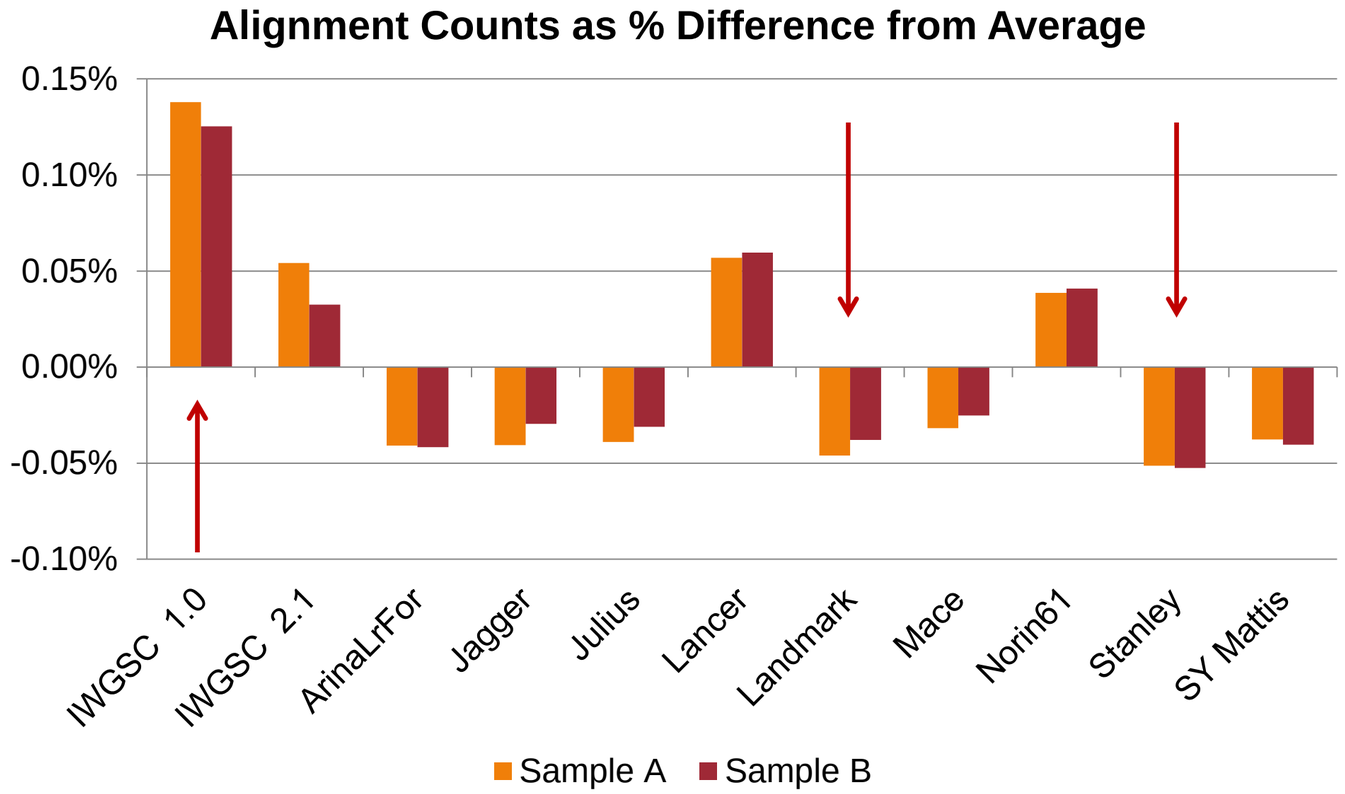
Landmark “TraesLDM2A01G082900”

(Equivalent: chr2A:33,681,338-33,681,768)



Read Alignment Across “10+ Wheat Genomes”

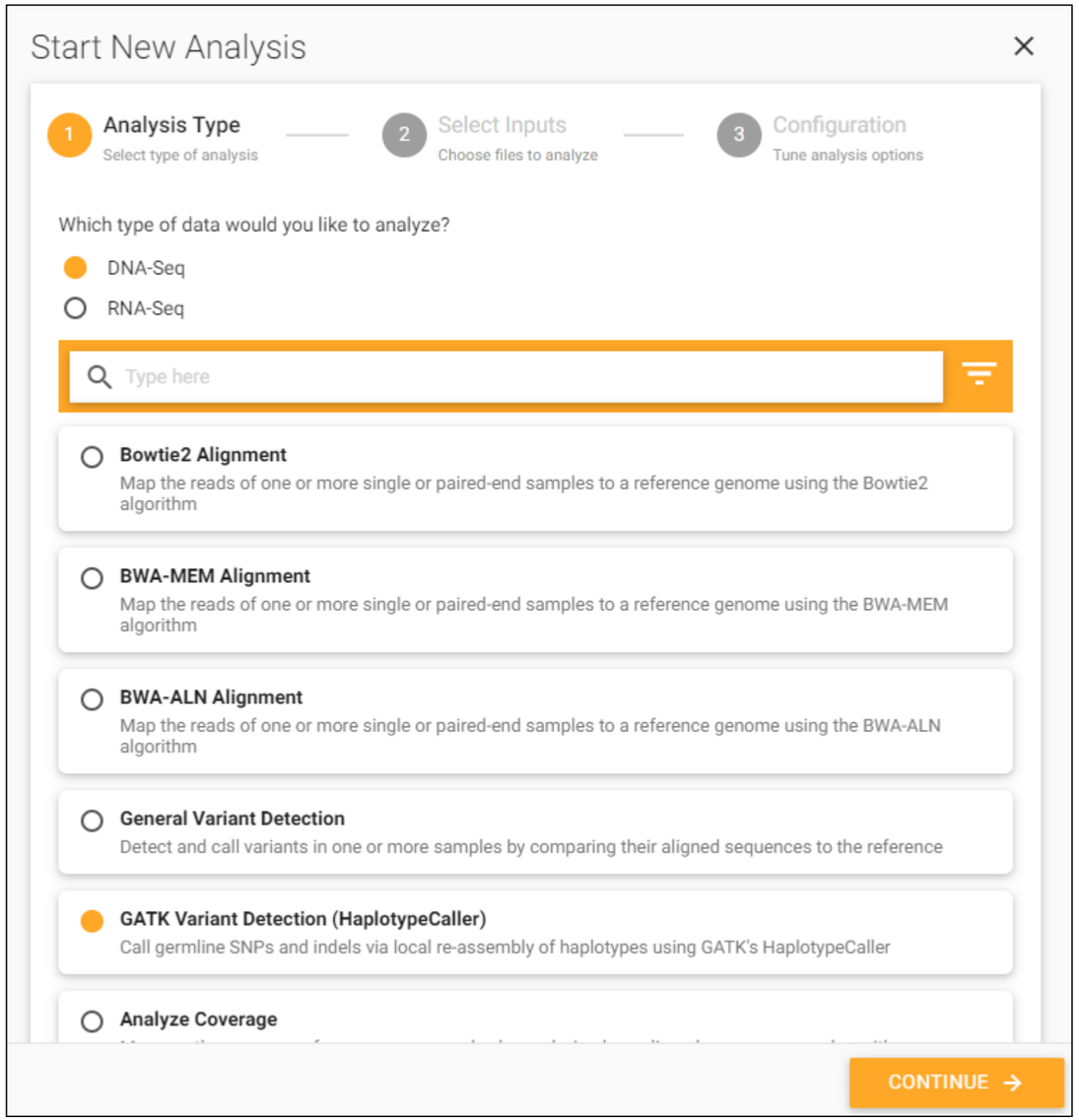
	Sample A	Sample B
IWGSC 1.0	130,401,964	158,730,060
IWGSC 2.1	130,292,935	158,582,884
ArinaLrFor	130,169,124	158,465,361
Jagger	130,169,417	158,484,503
Julius	130,171,553	158,482,155
Lancer	130,296,404	158,625,929
Landmark	130,162,411	158,471,316
Mace	130,180,915	158,491,452
Norin61	130,272,621	158,596,217
Stanley	130,155,509	158,448,214
SY Mattis	130,173,317	158,467,435
Raw Reads:	**129,567,060**	**157,465,504**



Variant Calling & Filtering Using 10+ Wheat Genomes Assemblies

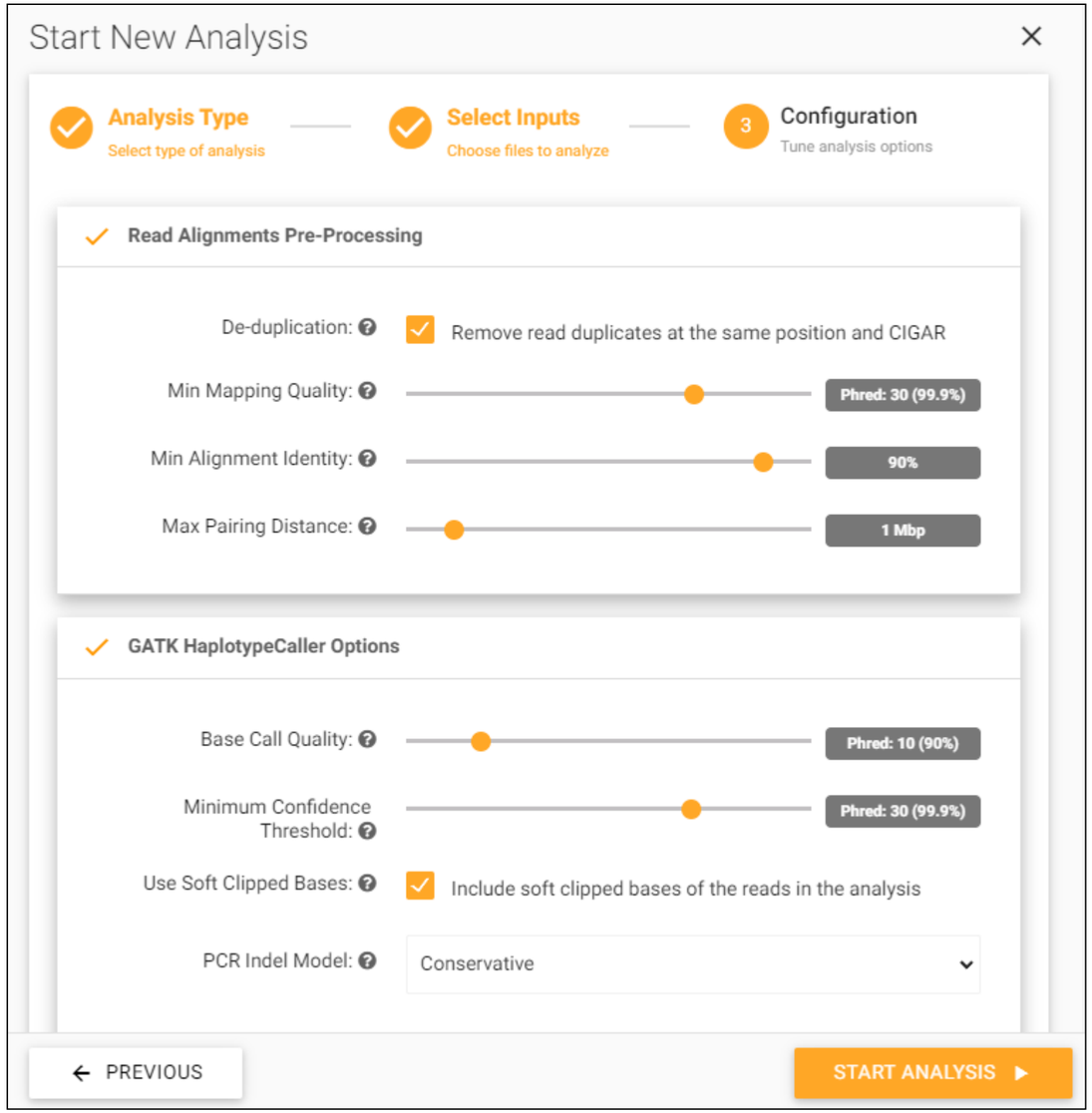
Wheat Variant Calling

- Various algorithms available to call variants, including INDELS & reference alleles
- Alleles automatically called leveraging associated 10+ Wheat Genomes reference assemblies



Wheat Variant Calling

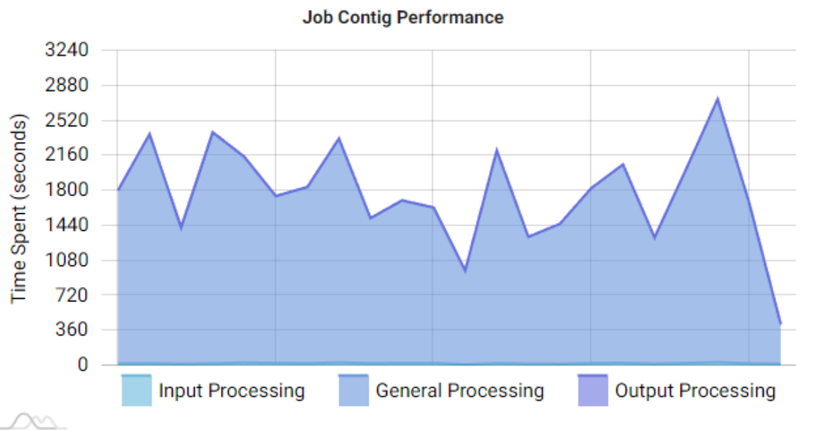
- Various algorithms available to call variants, including INDELS & reference alleles
- Alleles automatically called leveraging associated 10+ Wheat Genomes reference assemblies
- Experimentally adjust alignment pre-processing options and tune for desired specificity & sensitivity
- Convenient and highly-performant integration with the industry-standard GATK Haplotype Caller



Wheat Variant Calling

- Various algorithms available to call variants, including INDELS & reference alleles
- Alleles automatically called leveraging associated 10+ Wheat Genomes reference assemblies
- Experimentally adjust alignment pre-processing options and tune for desired specificity & sensitivity
- Convenient and highly-performant integration with the industry-standard GATK Haplotype Caller
- Parallel processing technology efficiently processes even massive individual samples or large batches of multiple samples

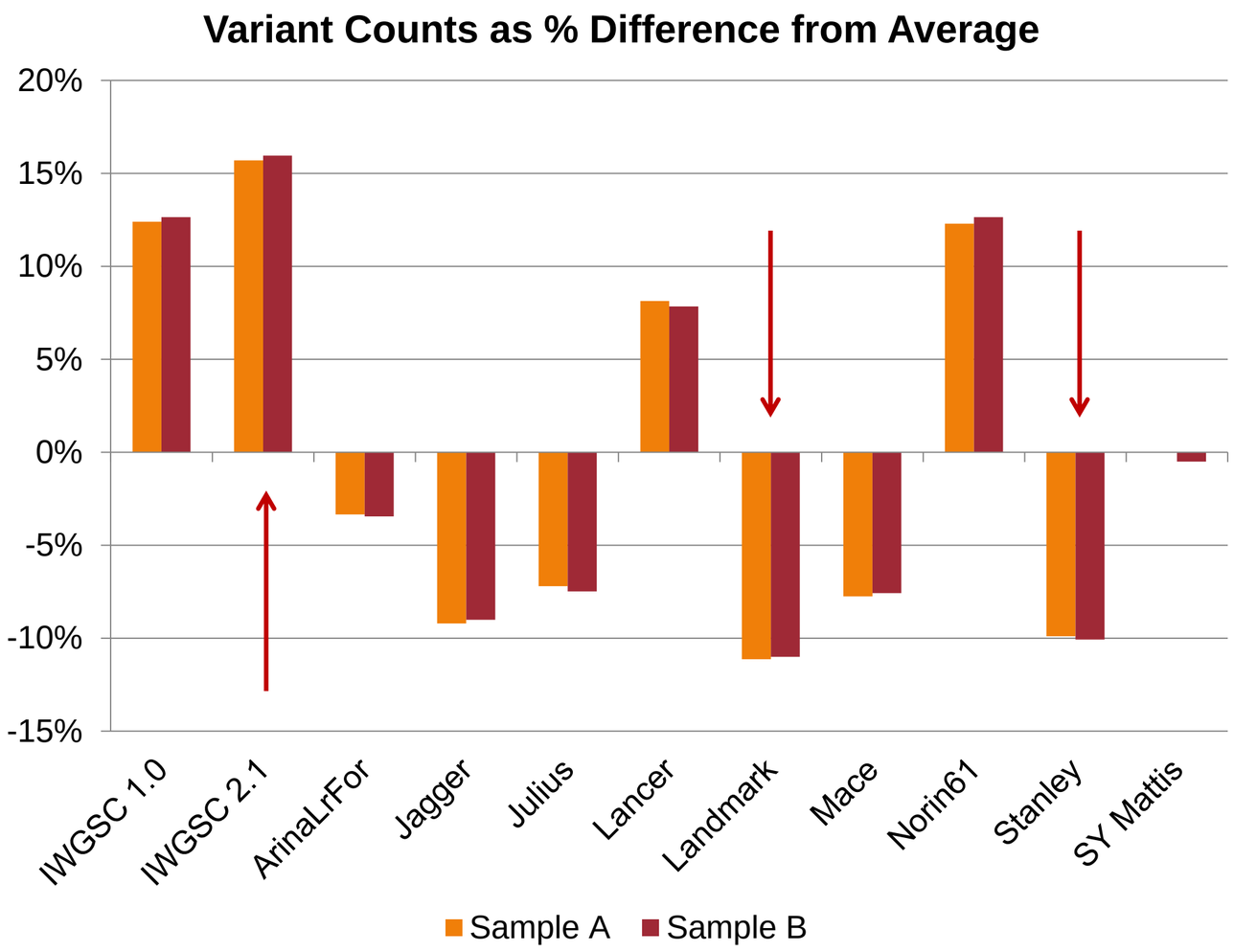
Record Count:	2,601,488 variants
Total Elapsed Clock Time Spent:	47 minutes, 36 seconds
Parallel Cluster Acceleration:	1,365%



Number of Job Contigs:	22 contigs
Contig Input Processing Time Total:	8 minutes, 25 seconds
Contig Record Processing Time Total:	10 hours, 41 minutes
Contig Output Processing Time Total:	26 seconds
Total Compute Time for All Contigs:	10 hours, 50 minutes

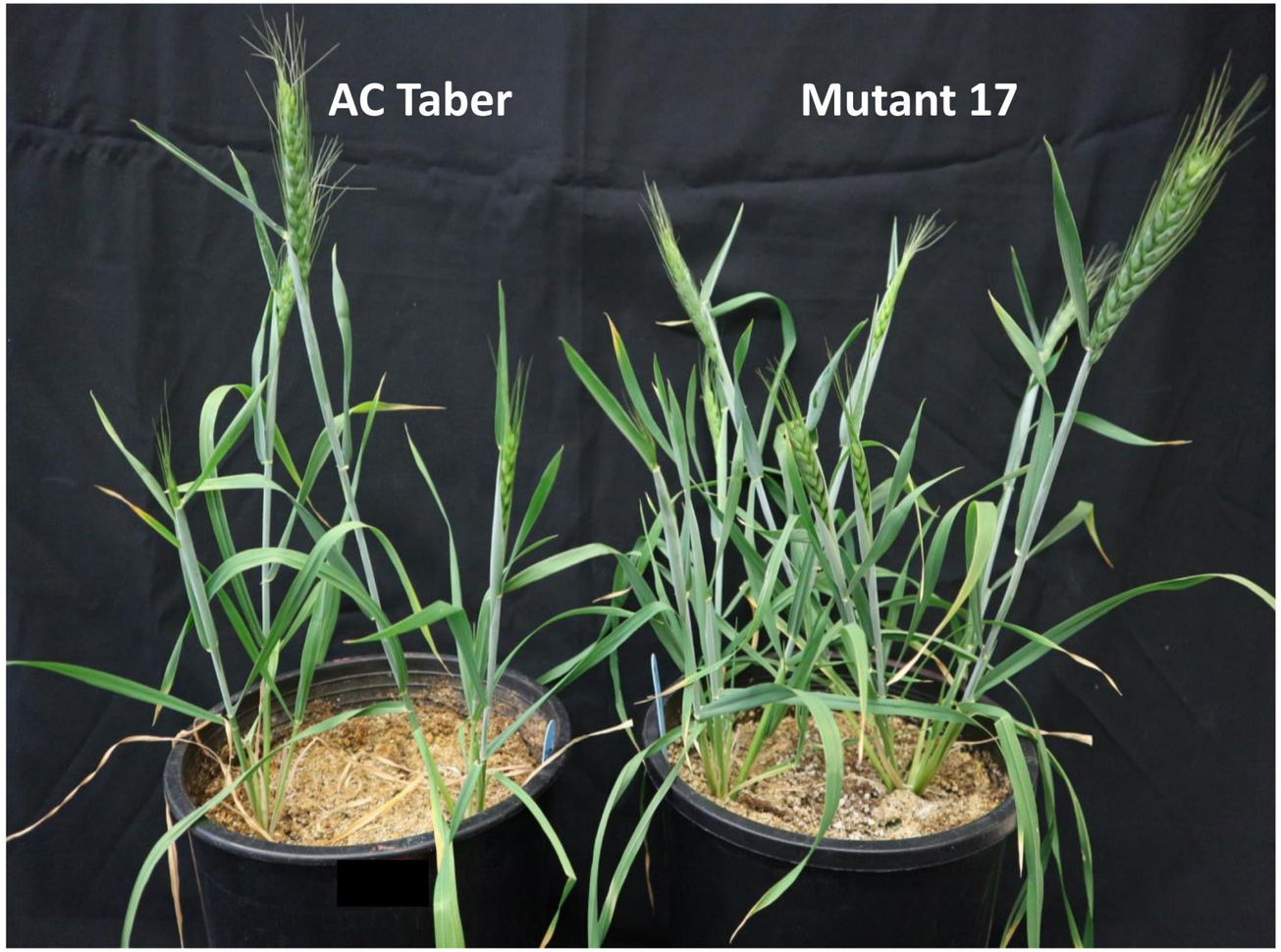
Variant Calling Across 10+ Wheat Genomes

	Sample A	Sample B
IWGSC 1.0	2,705,646	3,035,134
IWGSC 2.1	2,785,165	3,124,243
ArinaLrFor	2,326,655	2,601,488
Jagger	2,185,424	2,451,528
Julius	2,233,777	2,492,644
Lancer	2,603,096	2,905,634
Landmark	2,139,171	2,397,919
Mace	2,220,472	2,490,270
Norin61	2,703,332	3,035,071
Stanley	2,168,964	2,422,944
SY Mattis	2,406,907	2,680,842



Identifying Key Mutations

- AAFC project used mutagenesis to try and identify genotype of more drought tolerant wheat lines
- Under drought stress mutant shows:
  - Higher biomass
  - Stress adaptation features
  - Comparable spike length
  - Increased number of spikelets
  - Higher seed weight
- Abscisic acid pathways direct drought response signaling



Source: "Genomics-Driven Development of Drought Tolerant Wheat Cultivars", (Presented at PAG 2022, Neha Vaid)

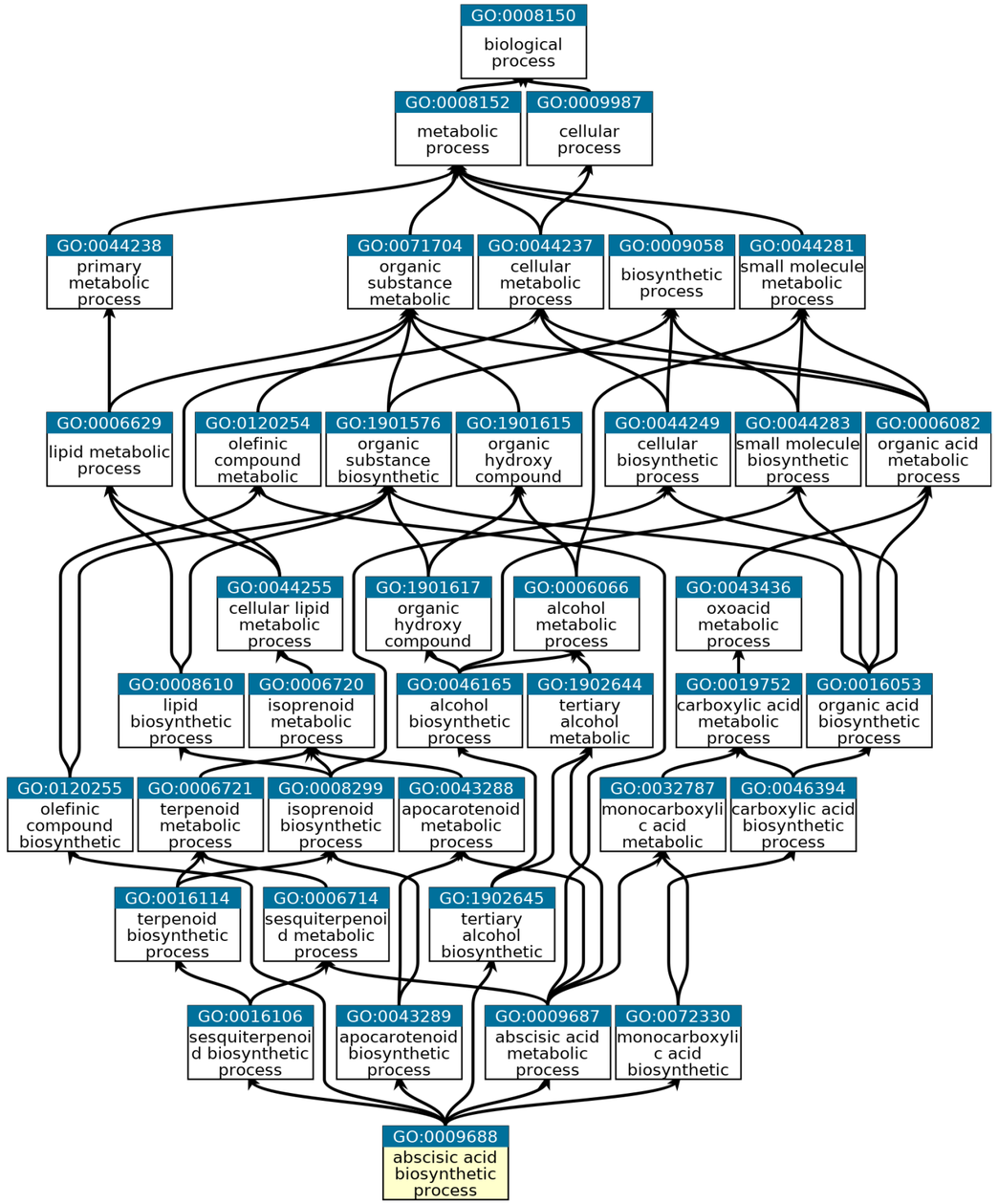
Leveraging IWGSC Functional Annotations

- 78 genes (85 transcripts) reference abscisic acid (ABA) within their IWGSC functional descriptions
- Gene names are based on the IWGSC 1.0 naming convention, which is easy to translate to the IWGSC 1.1 convention
- Translation table from the “10+ Wheat Genome” project can then be used to identify equivalent genes
- Example of translating the IWGSC gene with name “TraesCS2A01G317000”

Genome	Gene
IWGSC 1.0	TraesCS2A01G317000
IWGSC 1.1	TraesCS2A02G317000
ArinaLrFor	TraesARI2A01G344100
Jagger	TraesJAG2A01G345500
Julius	TraesJUL2A01G344300
Landmark	TraesLDM2A01G344000
LongReach Lancer	TraesLAC2A01G344600
Mace	TraesMAC2A01G334600
Norin 61	TraesNOR2A01G347400
Stanley	TraesSTA2A01G333700
SY Mattis	TraesSYM2A01G334400

Gene Ontology for Example Gene

- IWGSC annotations show TraesCS2A02G317000 associated with ‘GO:0009688’
- Ancestor tree for this gene ontology shown
- GO:0009688, biological process: **abscisic acid biosynthetic process**
- GO:0009688 definition:
  - The chemical reactions and pathways resulting in the formation of abscisic acid, 5-(1-hydroxy-2,6,6,trimethyl-4-oxocyclohex-2-en-1-y1)-3-methylpenta-2, 4-dienoic acid.



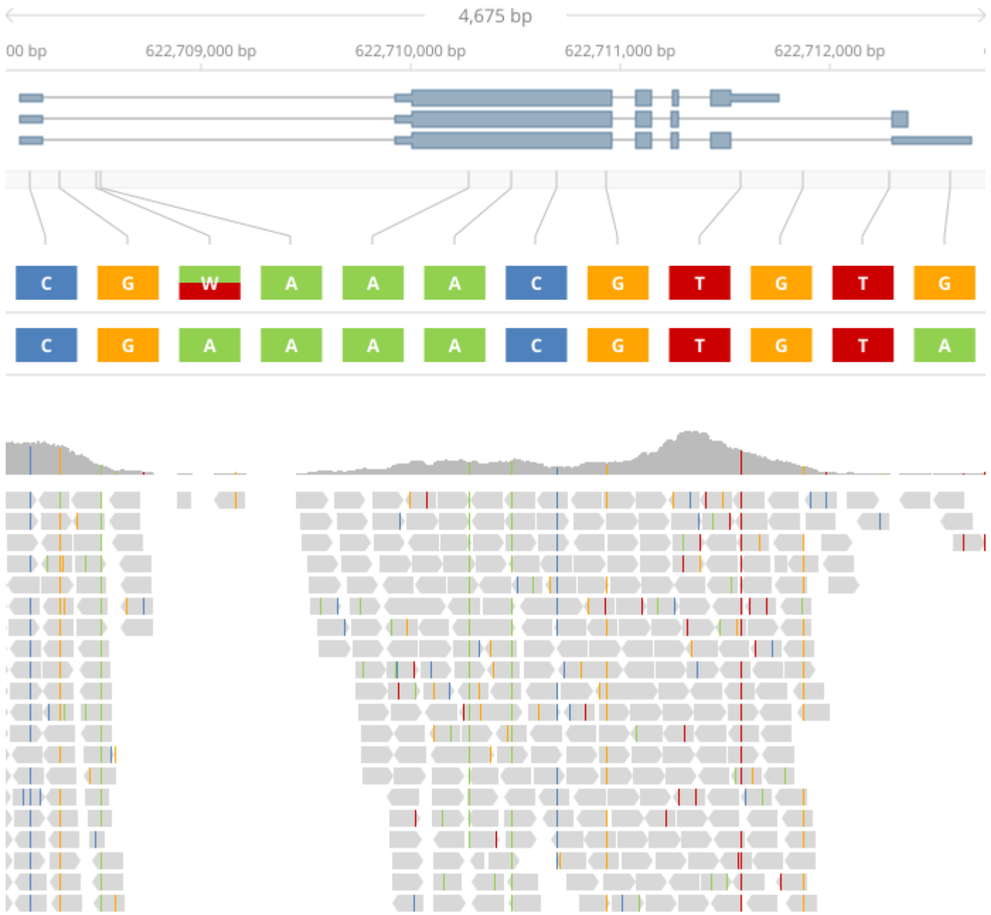
Source: https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0009688

Alleles identified in IWGSC gene:
TraesCS2A02G317000

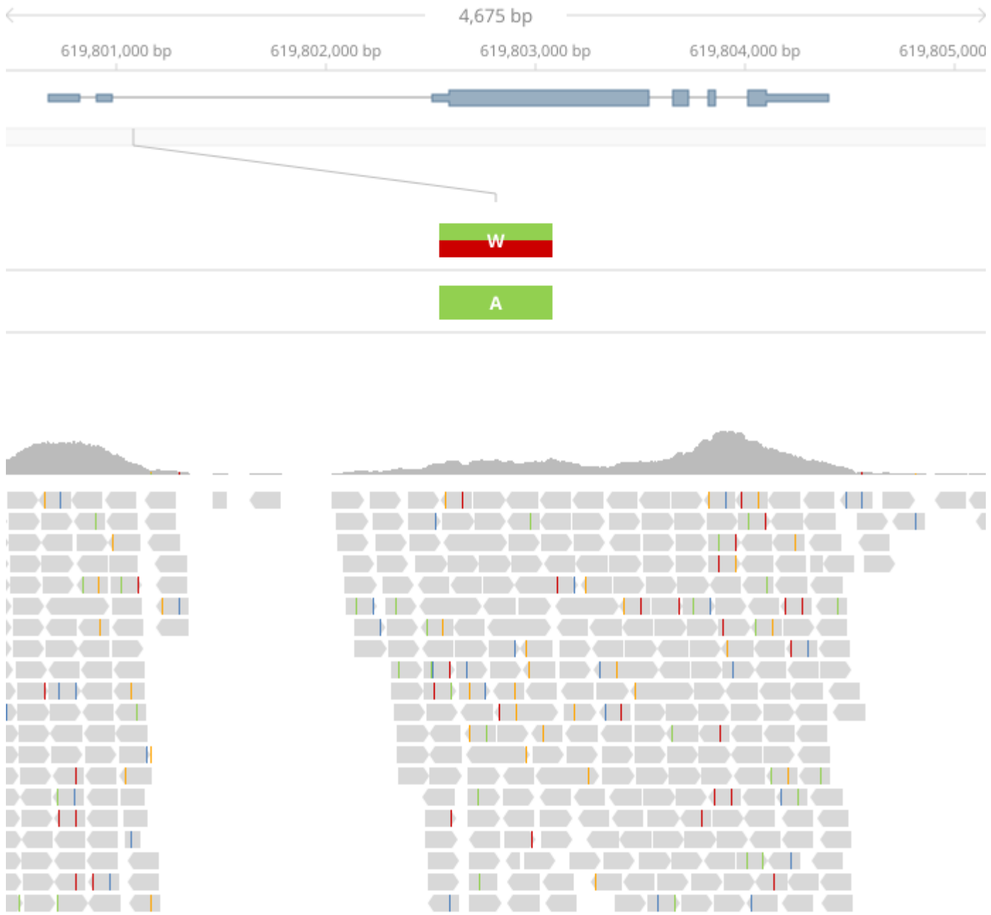
Alleles identified in Landmark gene:
TraesLDM2A03G00726620

Alleles identified in Stanley gene:
TraesSTA2A03G00722440

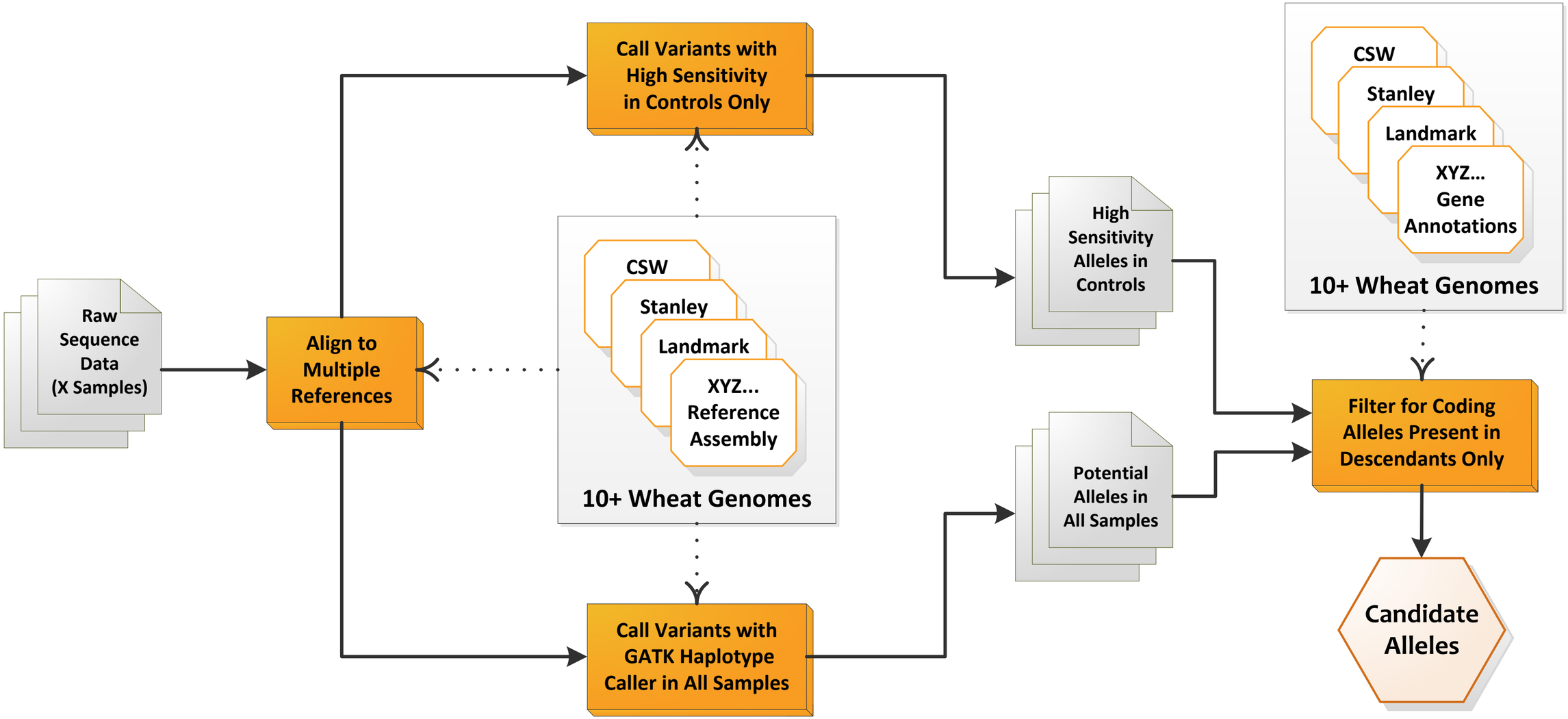
Alleles and Mapped Reads Using IWGSC
(Showing Gene: TraesCS3A02G371900)



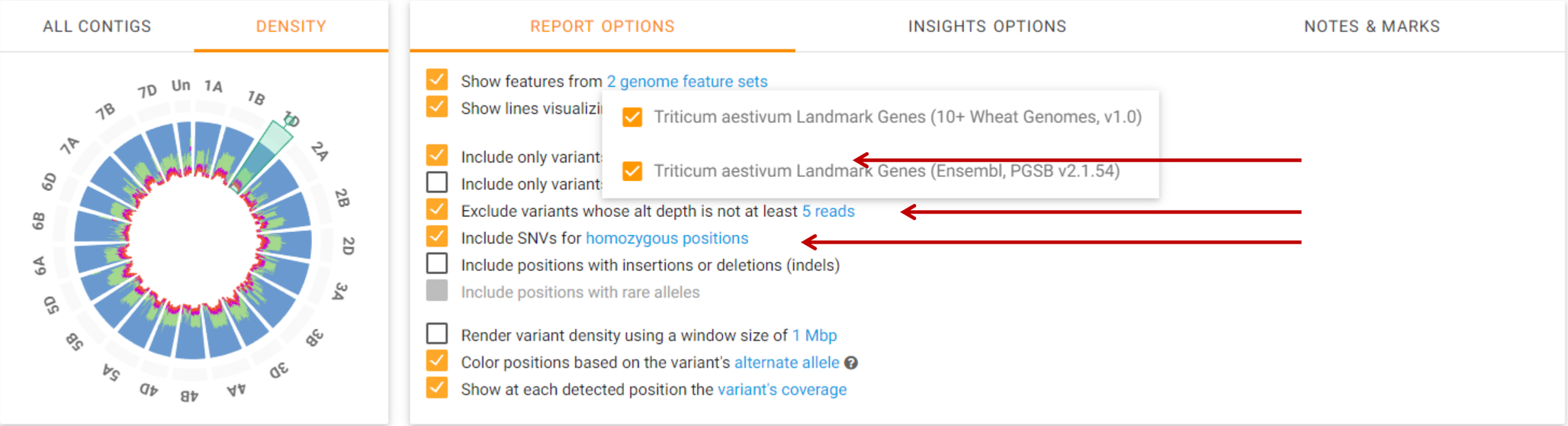
Same Samples Mapped to Landmark
(Equivalent Gene: TraesLDM3A03G01467370)



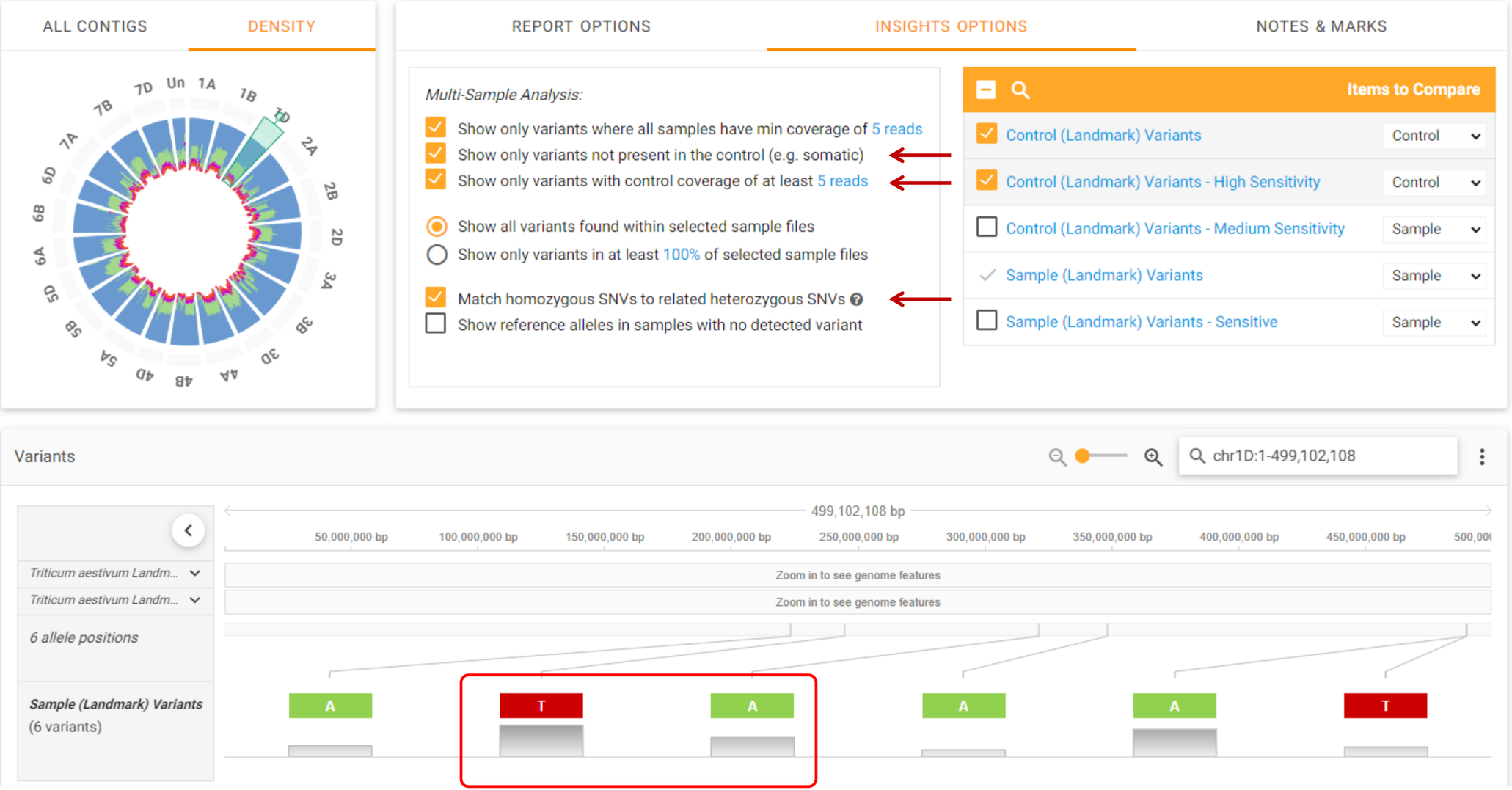
Candidate Alleles Analysis Approach



Candidate Allele Filtering

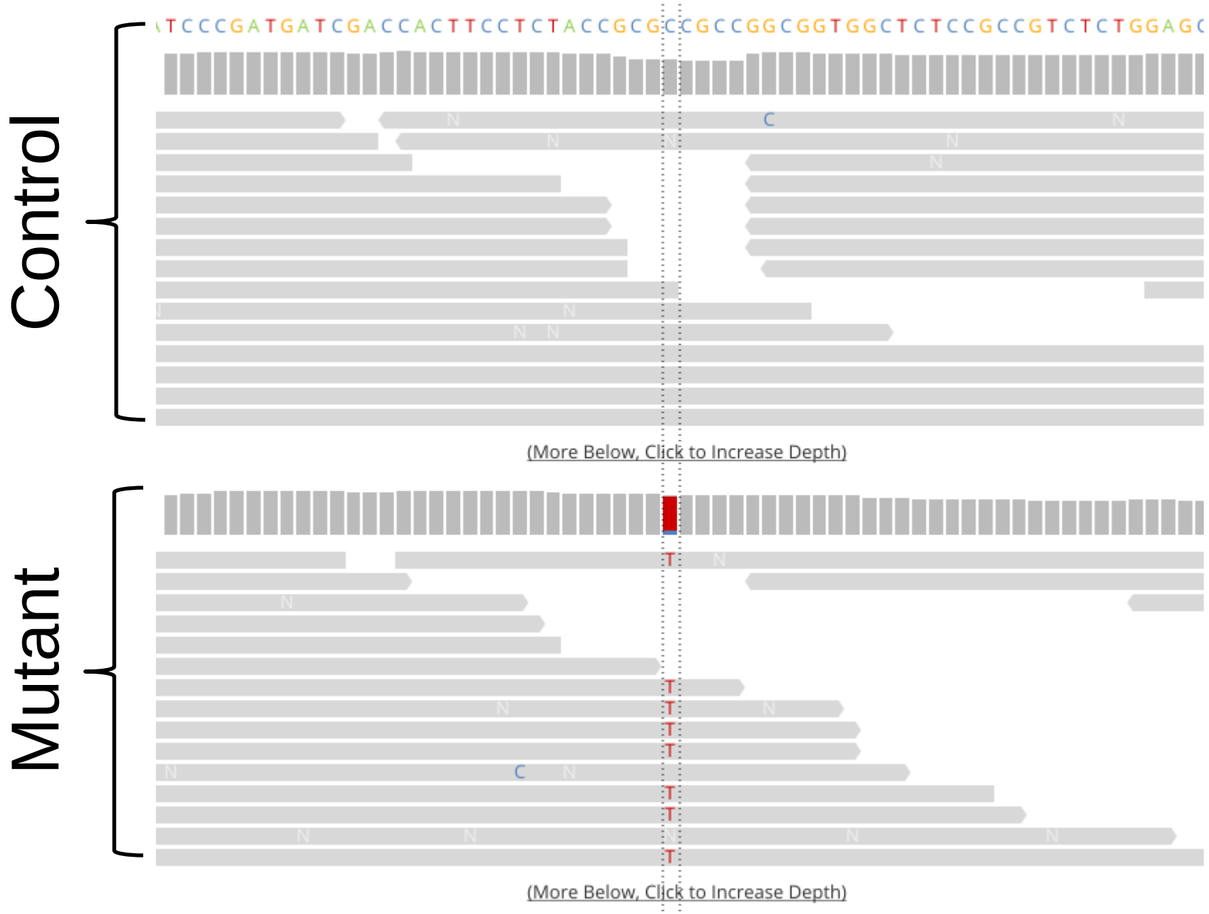


Candidate Allele Filtering

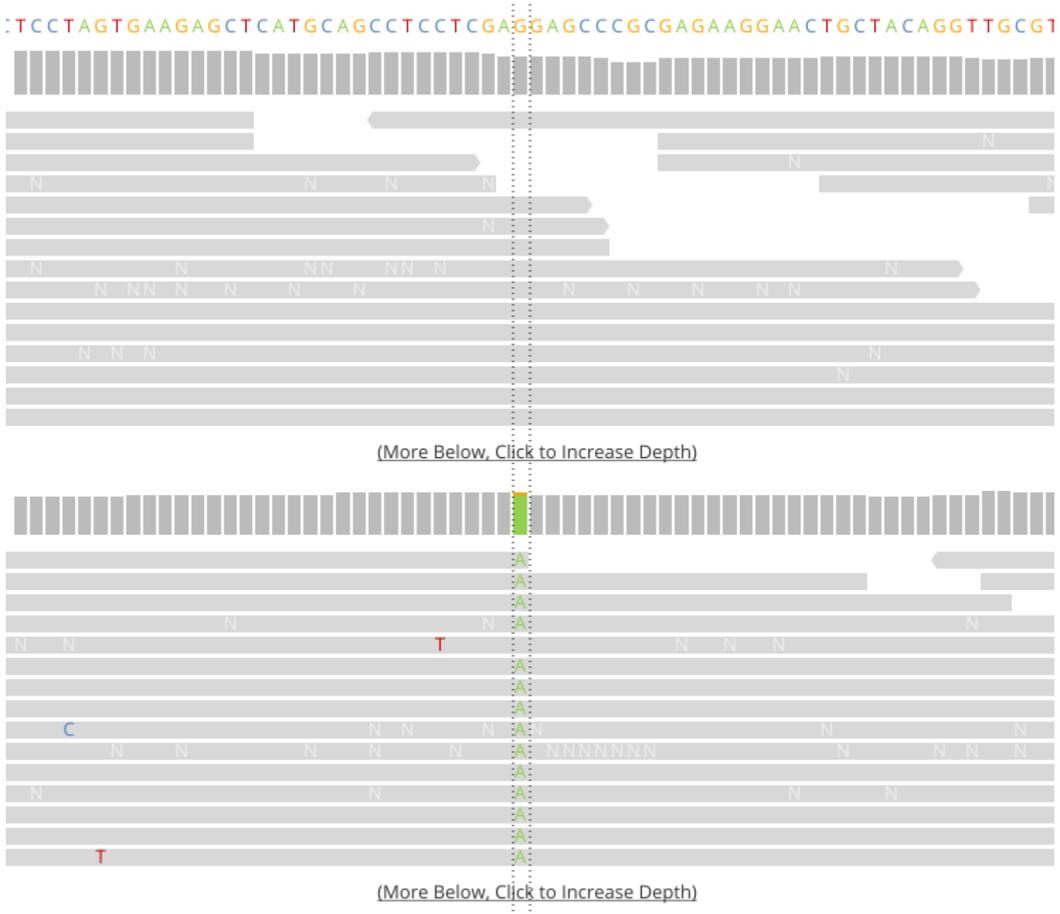


Candidate Allele Filtering

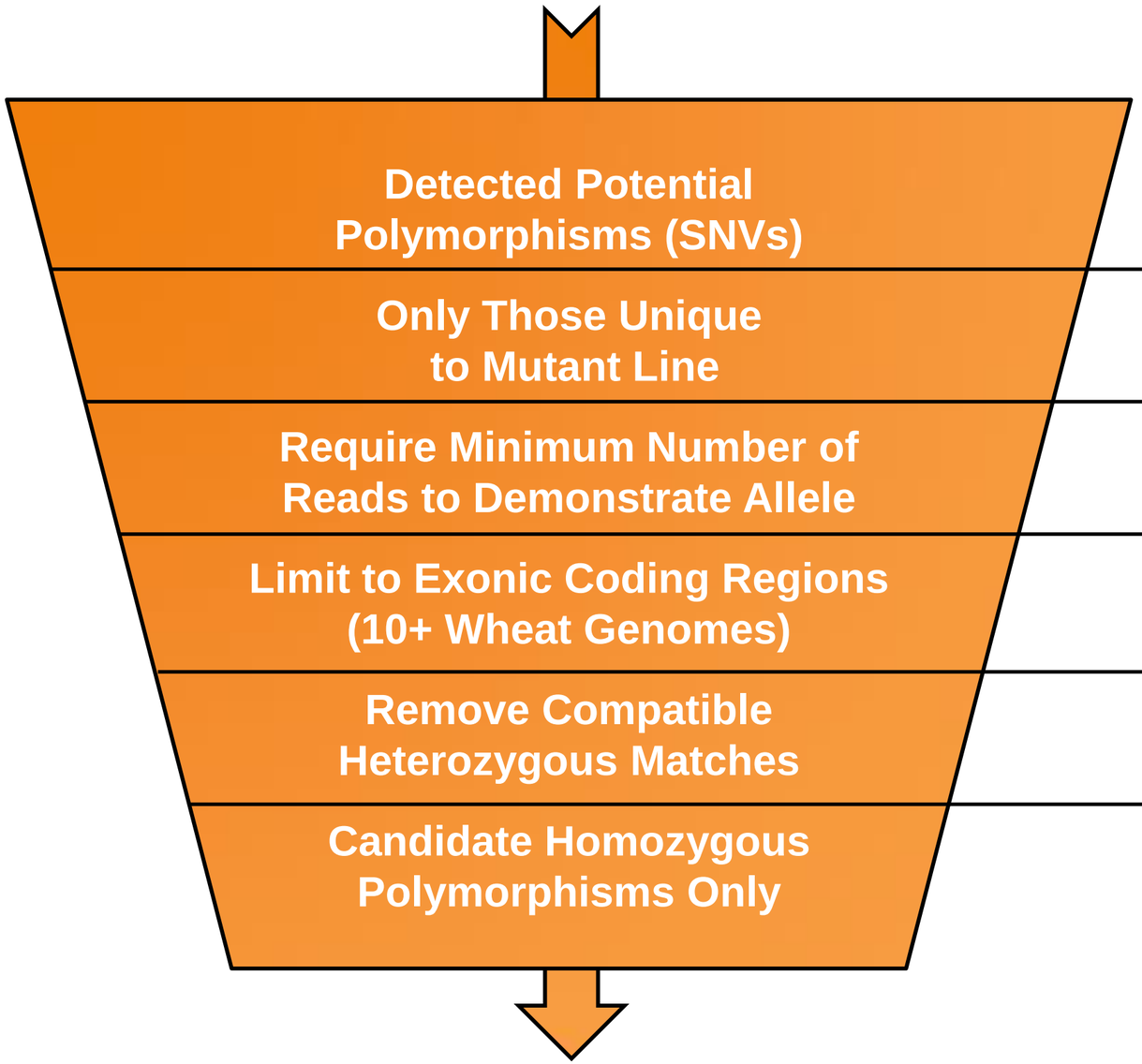
Gene: TraesLDM1D01G192000
(chr1D:244,524,468-244,524,532)



Gene: TraesLDM1D01G256200
(chr1D:321,109,370-321,109,434)



Candidate Allele Filtering (SNV Counts)

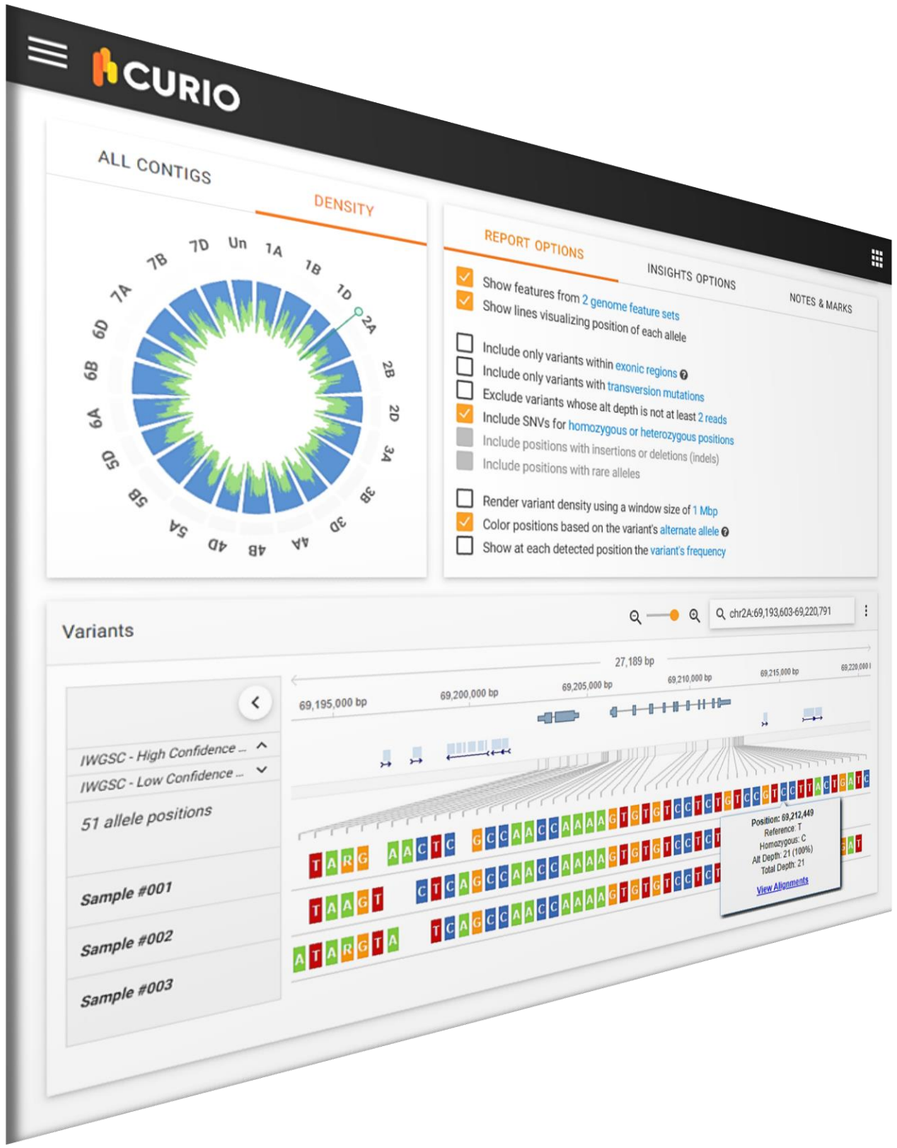


	IWGSC 2.1	IWGSC 1.0	Stanley	Landmark
Detected Potential Polymorphisms (SNVs)	2,820,397	2,751,824	2,190,860	2,168,141
Only Those Unique to Mutant Line	167,601	166,288	159,369	154,670
Require Minimum Number of Reads to Demonstrate Allele	42,061	41,398	37,565	35,812
Limit to Exonic Coding Regions (10+ Wheat Genomes)	11,161	11,305	10,979	10,448
Remove Compatible Heterozygous Matches	3,238	3,400	3,481	3,413
Candidate Homozygous Polymorphisms Only	**28**	**35**	**39**	**41**

In Conclusion

Looking Ahead

- Continued incorporation of new public reference assemblies, gene annotations, and variant call sets
- Enabling pangenome analysis, graph construction, alignment, variant calling , and visualization of sample data
- Enabling haplotype-based analysis approaches, both genome wide and gene based
- Incorporation of other data types, such as SNP array data
- Additional crop-research-specific analysis types and interpretive visualizations



Acknowledgements

Special thanks to collaborators with the Agriculture and Agri-Food Canada (AAFC) project:

- Marcus Samuel, PhD (Professor), University of Calgary
- Raju Soolanayakanahally, PhD (Research Scientist), AAFC, Saskatoon
- Sateesh Kagale (Research Officer), National Research Council, Saskatoon
- Neha Vaid, PhD (Research Associate), University of Calgary / AAFC, Saskatoon

And the collective efforts of the:

- International Wheat Genome Sequencing Consortium (IWGSC)
- 10+ Wheat Genomes Project

Q&A