



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)

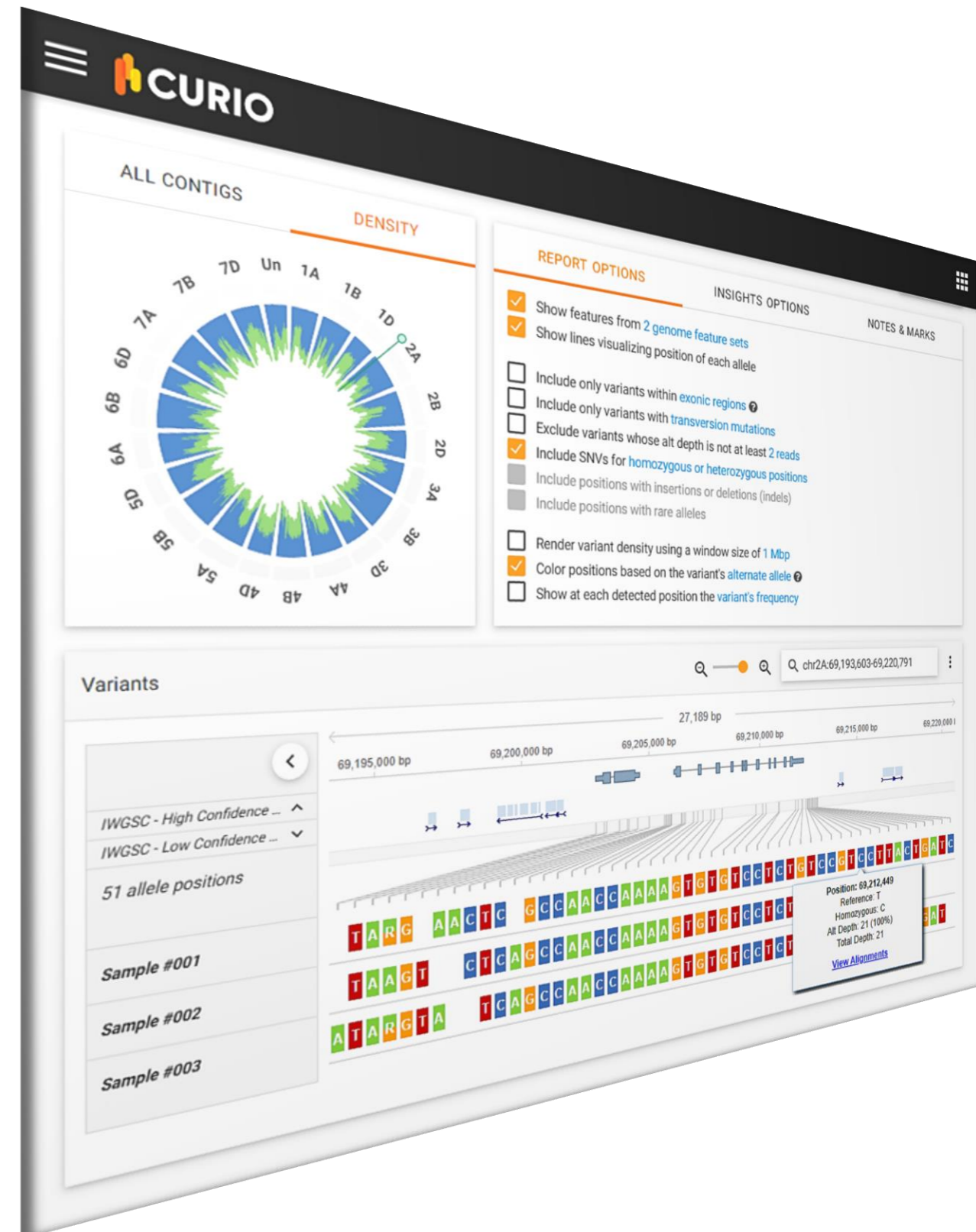
*Leveraging the IWGSC Gene Annotations
with Additional Cultivars in Curio*

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 both 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: TraesARI2A01G145700
- Translation table mapping to IWGSC RefSeq 1.1
gene/transcript IDs available at:
[https://galaxy-web-ipk-gatersleben.de/libraries/folders/F1cd8e2f6b131e891](https://galaxy-web.ipk-gatersleben.de/libraries/folders/F1cd8e2f6b131e891)



<https://10wheatgenomes.com/>



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: TraesARI2A03G00626940
- Translation table to IWGSC gene/transcript ids is not yet available



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

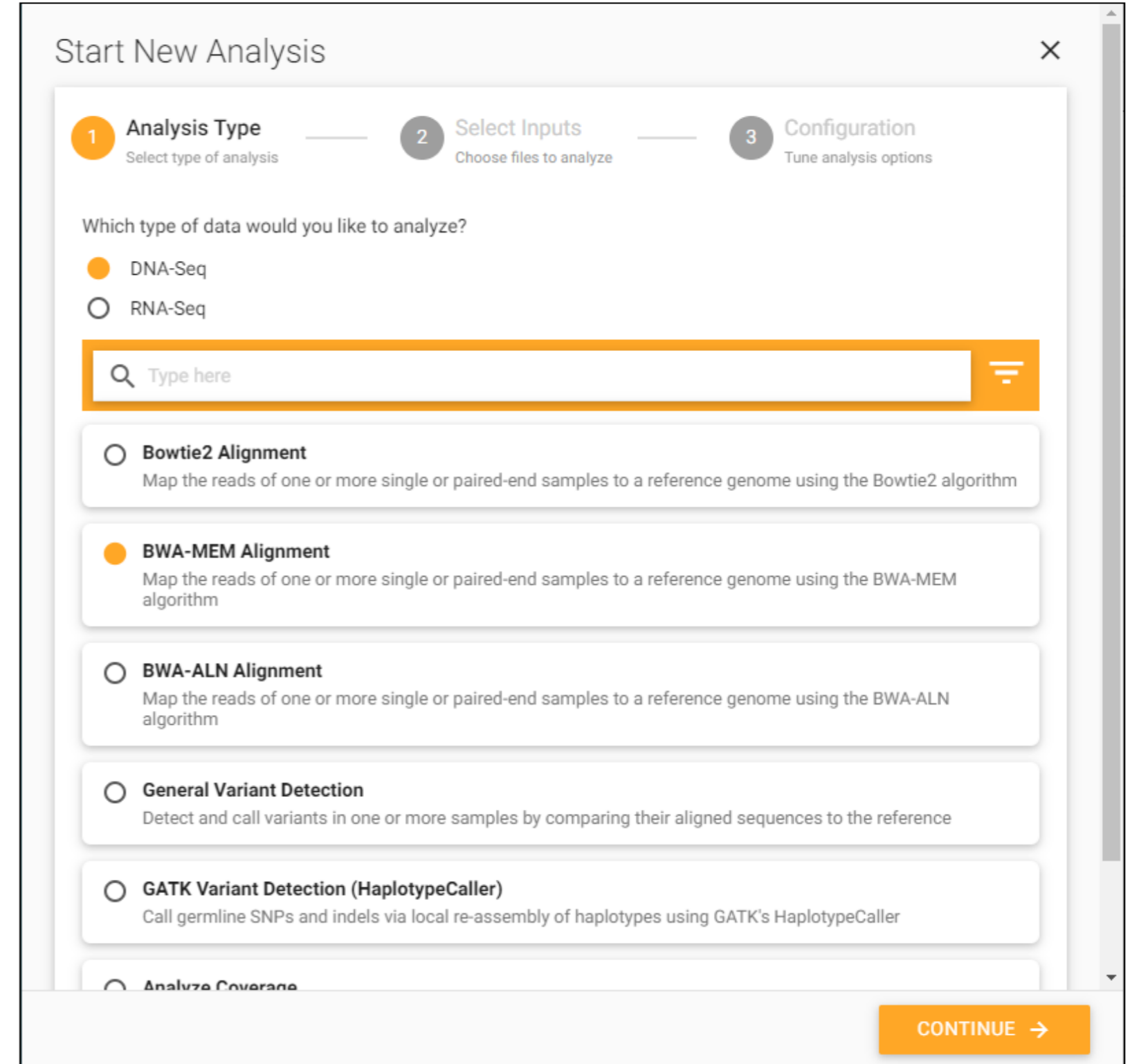
Wheat Cultivar	Original Version (#1)	Updated Version (#2)
ArinaLrFor	1,231,314 features	2,213,222 features
Jagger	1,221,309 features	2,110,846 features
Julius	1,217,757 features	2,289,641 features
Landmark	1,216,606 features	2,166,879 features
LongReach Lancer	1,225,483 features	2,079,609 features
Mace	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

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



Start New Analysis

1 Analysis Type Select type of analysis — 2 Select Inputs Choose files to analyze — 3 Configuration Tune analysis options

Which type of data would you like to analyze?

DNA-Seq
 RNA-Seq

🔍 Type here

Bowtie2 Alignment
Map the reads of one or more single or paired-end samples to a reference genome using the Bowtie2 algorithm

BWA-MEM Alignment
Map the reads of one or more single or paired-end samples to a reference genome using the BWA-MEM algorithm

BWA-ALN Alignment
Map the reads of one or more single or paired-end samples to a reference genome using the BWA-ALN algorithm

General Variant Detection
Detect and call variants in one or more samples by comparing their aligned sequences to the reference

GATK Variant Detection (HaplotypeCaller)
Call germline SNPs and indels via local re-assembly of haplotypes using GATK's HaplotypeCaller

Analyze Coverage

CONTINUE →

- 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

Start New Analysis
✕

✓ Analysis Type
Select type of analysis

✓ Select Inputs
Choose files to analyze

3 Configuration
Tune analysis options

Reference Assembly

Select one of the standard reference assemblies to align to, or use a custom reference. ?

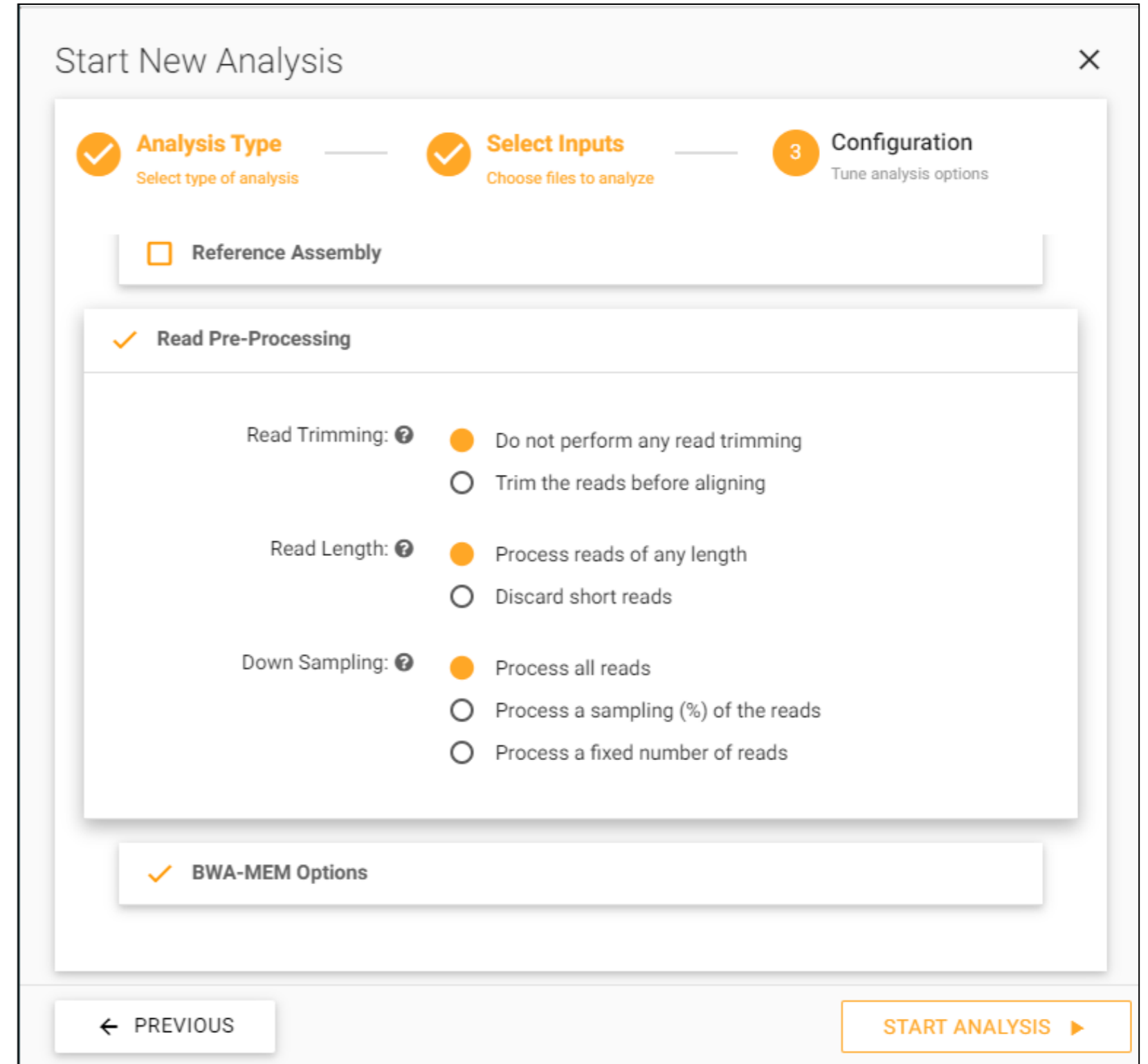
(Choose Reference Assembly)

- Barley (Hordeum vulgare) - Morex V2 Assembly
- Barley (Hordeum vulgare) - Morex V3 Assembly
- Chinese Spring Wheat (Triticum aestivum) - IWGSC 1.0 Assembly
- Chinese Spring Wheat (Triticum aestivum) - IWGSC 2.1 Assembly
- Human (Homo sapiens) - GRCh38 Assembly (hg38)
- Oat Diploid (Avena longiglumis) - CN58138 v1 Assembly
- Oat Hexaploid (Avena sativa) - Sang v1 Assembly
- Oat Tetraploid (Avena insularis) - BYU209 v1 Assembly
- Rye (Secale cereale) - Lo7 Assembly (IRGSC)
- Rye (Secale cereale) - Weining v1.0 Assembly (HAU)
- Svevo Wheat (Triticum turgidum) - IDWGSC v1 Assembly
- Triticum aestivum ArinalrFor (10+ Wheat Genomes, v3.0)
- Triticum aestivum Jagger (10+ Wheat Genomes, v1.1)
- Triticum aestivum Julius (10+ Wheat Genomes, v1.0)
- Triticum aestivum Lancer (10+ Wheat Genomes, v1.0)
- Triticum aestivum Landmark (10+ Wheat Genomes, v1.0)
- Triticum aestivum Mace (10+ Wheat Genomes, v1.0)
- Triticum aestivum Norin61 (10+ Wheat Genomes, v1.1)
- Triticum aestivum Stanley (10+ Wheat Genomes, v1.2)
- Triticum aestivum SY Mattis (10+ Wheat Genomes, v1.0)

← PREVIOUS

START ANALYSIS ▶

- 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



Start New Analysis

✓ Analysis Type
Select type of analysis

✓ Select Inputs
Choose files to analyze

3 Configuration
Tune analysis options

Reference Assembly

✓ Read Pre-Processing

Read Trimming: ? Do not perform any read trimming
 Trim the reads before aligning

Read Length: ? Process reads of any length
 Discard short reads

Down Sampling: ? Process all reads
 Process a sampling (%) of the reads
 Process a fixed number of reads

✓ BWA-MEM Options

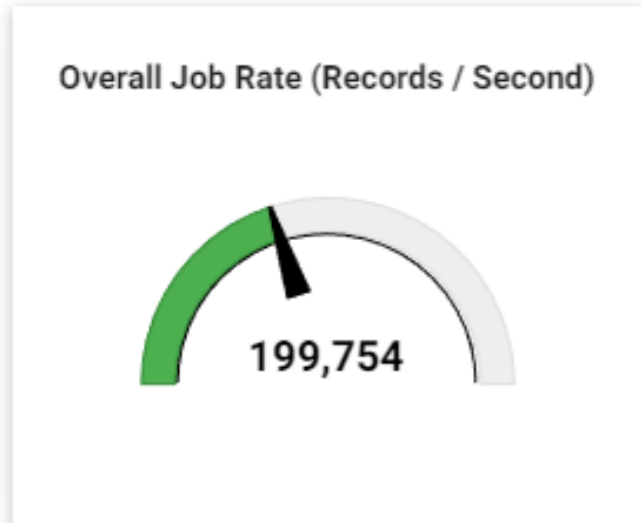
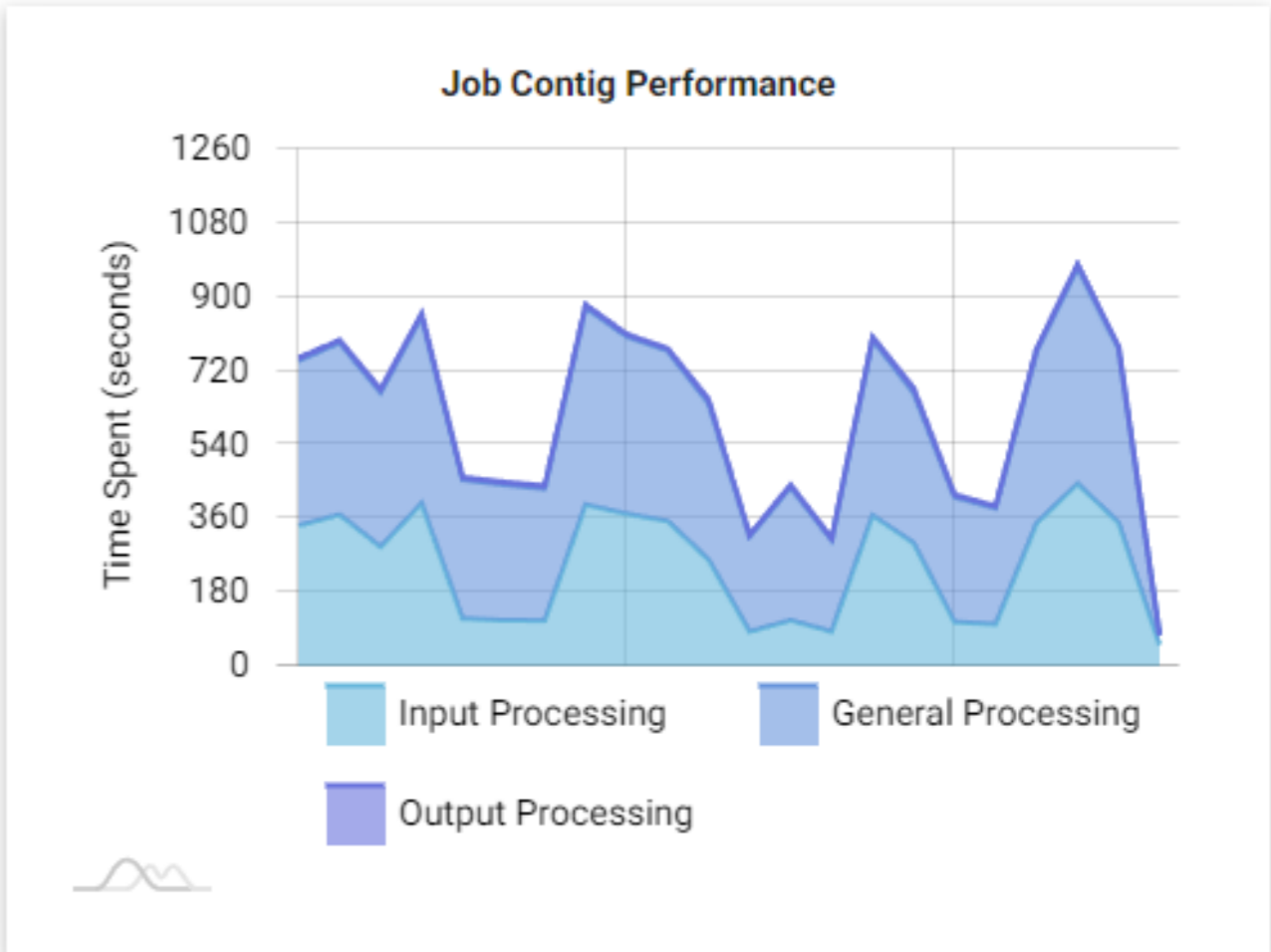
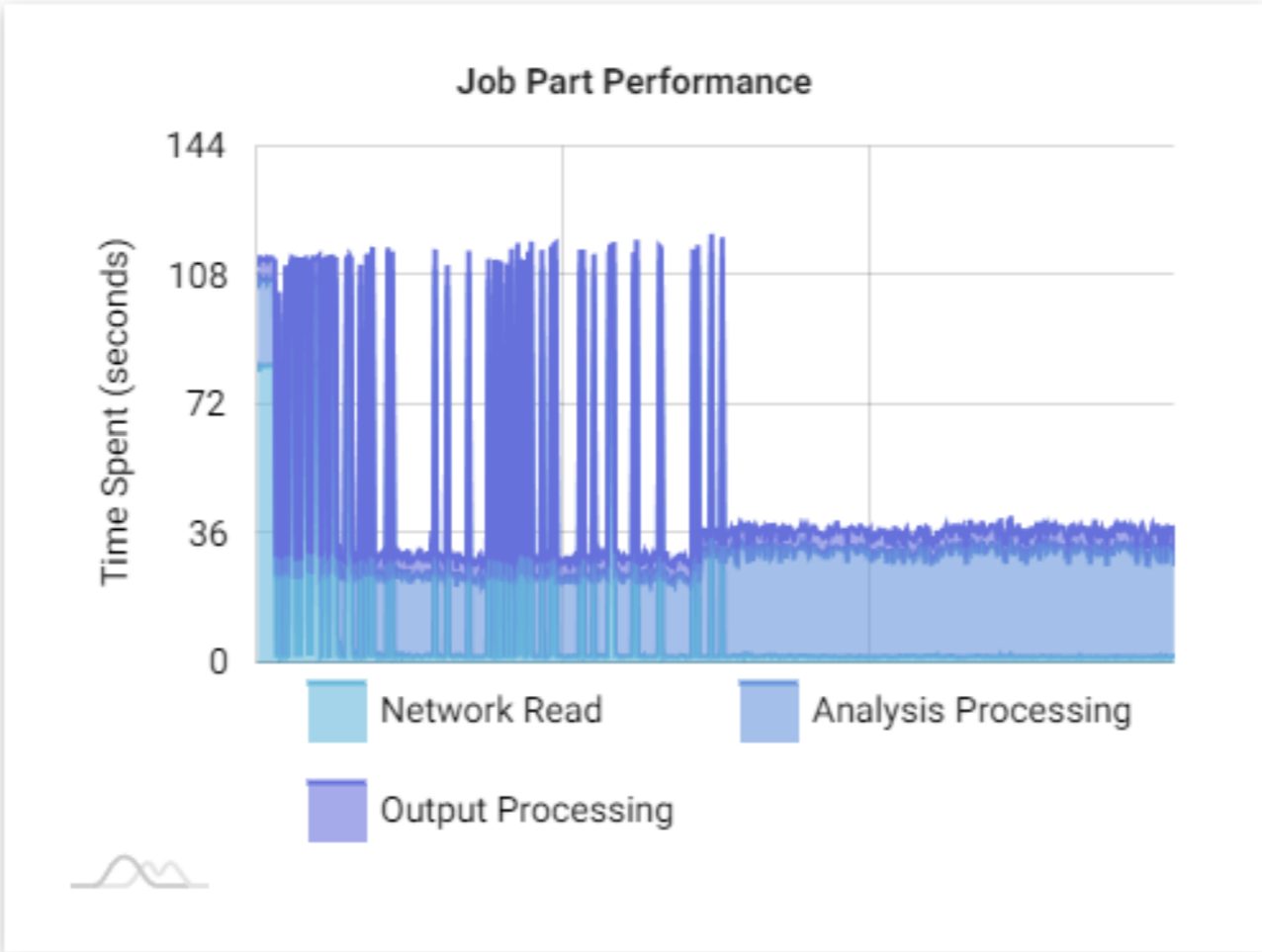
← PREVIOUS

START ANALYSIS ▶

Record Count:
158,487,099 alignments

Total Elapsed Clock Time Spent:
13 minutes, 13 seconds

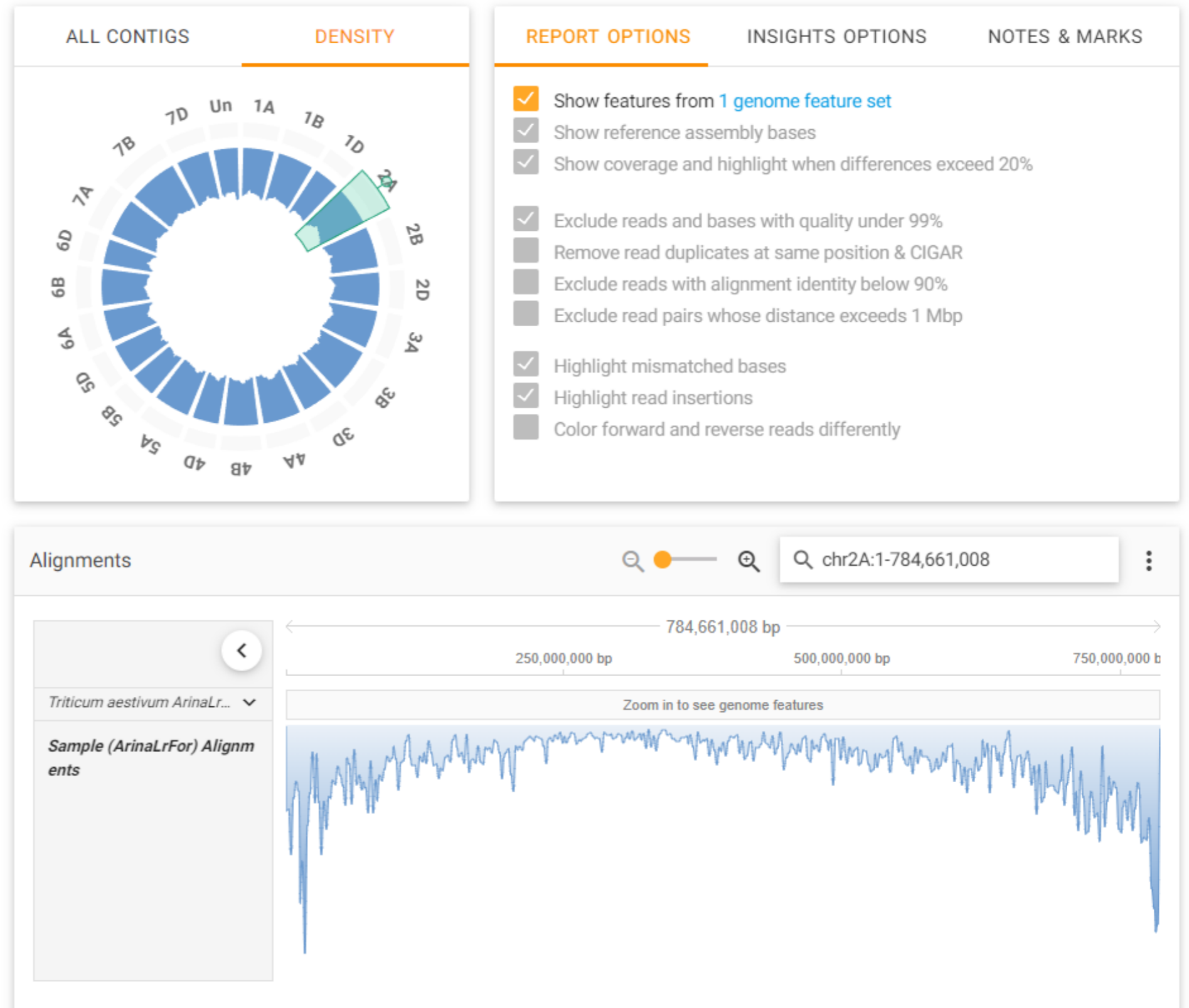
Parallel Cluster Acceleration:
3,364%



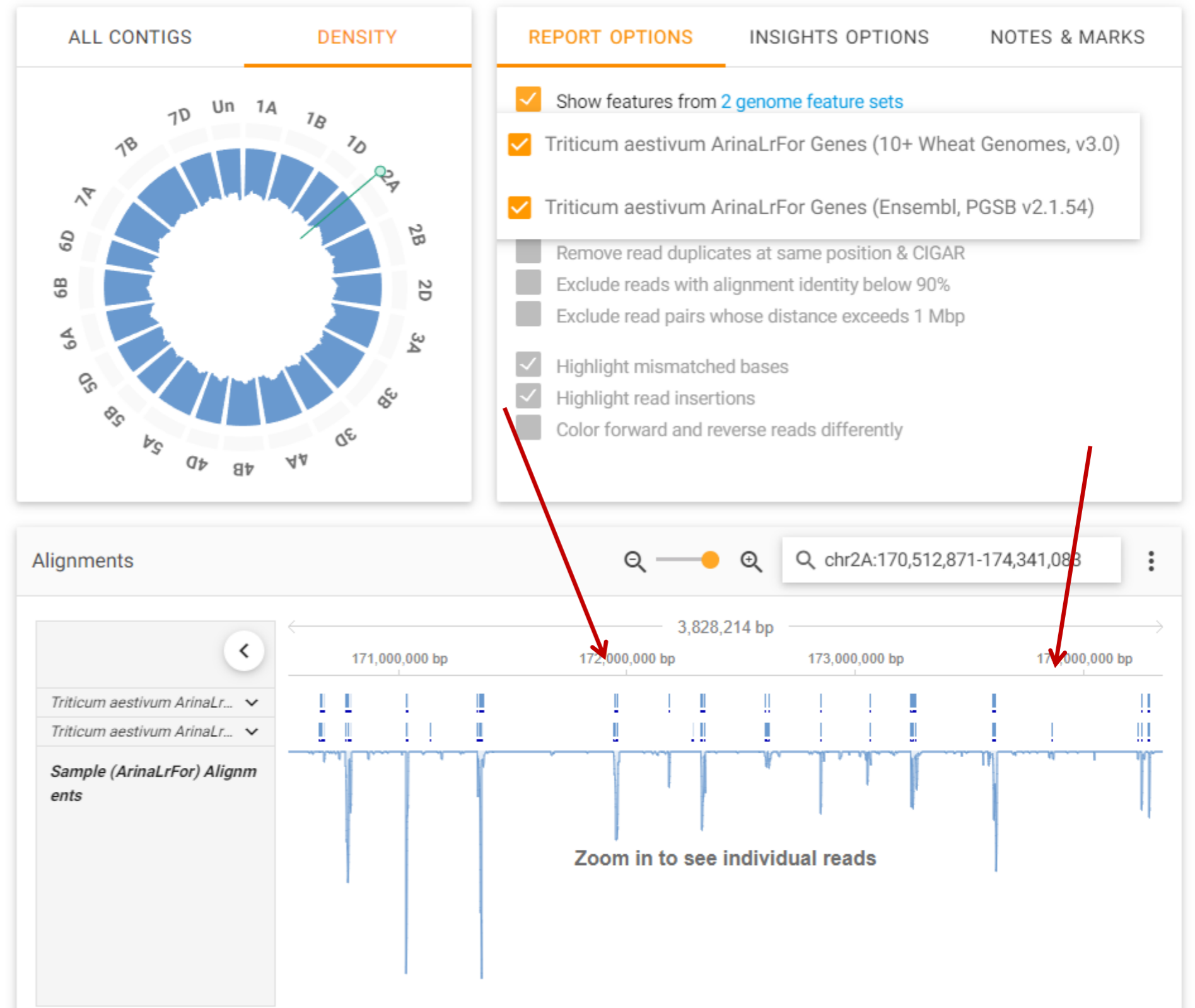
Number of Job Parts:	601 parts
Part Network Read Time Total:	2 hours, 8 minutes
Part Analysis Processing Time Total:	4 hours, 23 minutes
Part Output Processing Time Total:	53 minutes, 10 seconds
Total Compute Time for All Parts:	7 hours, 24 minutes

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

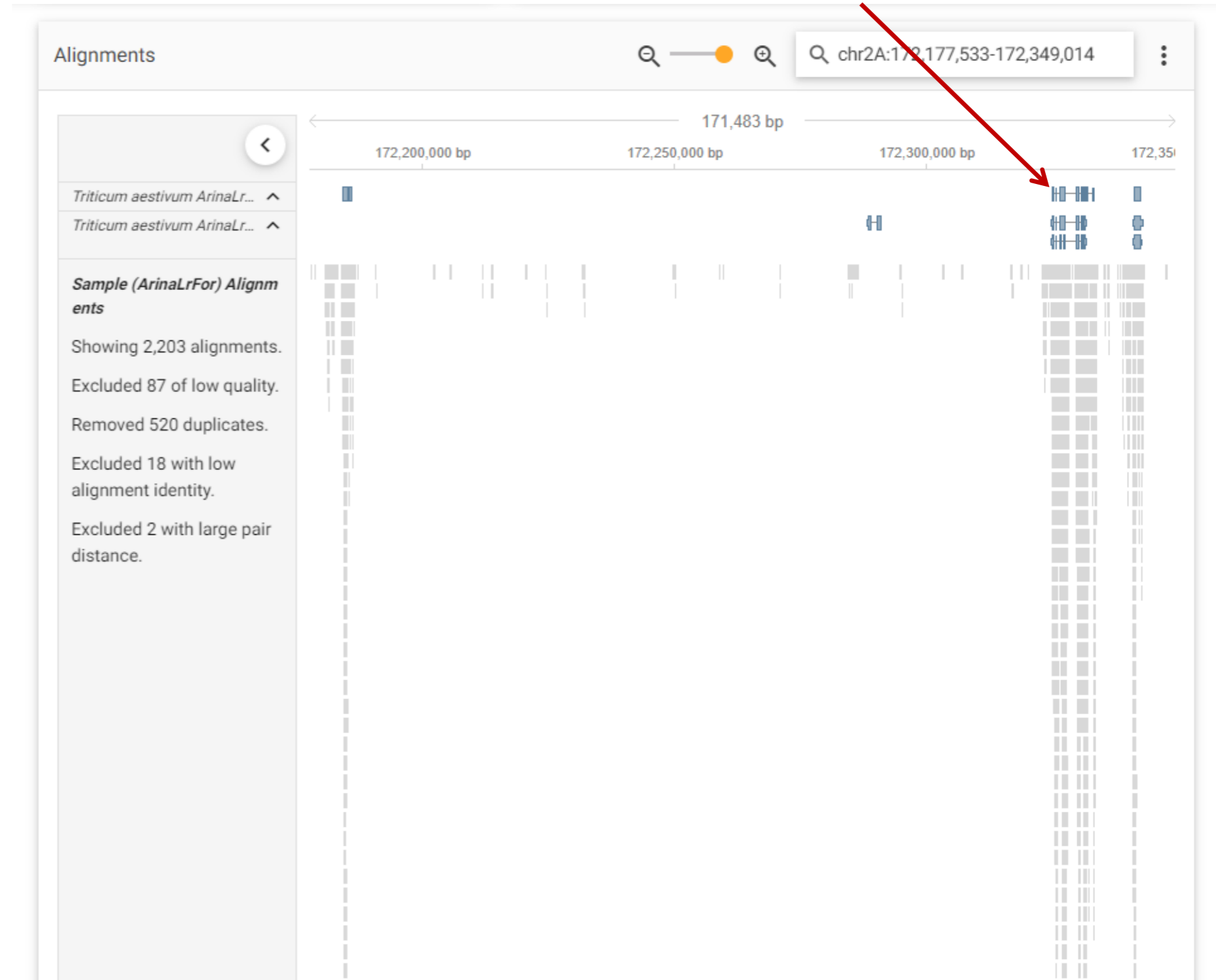
- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome



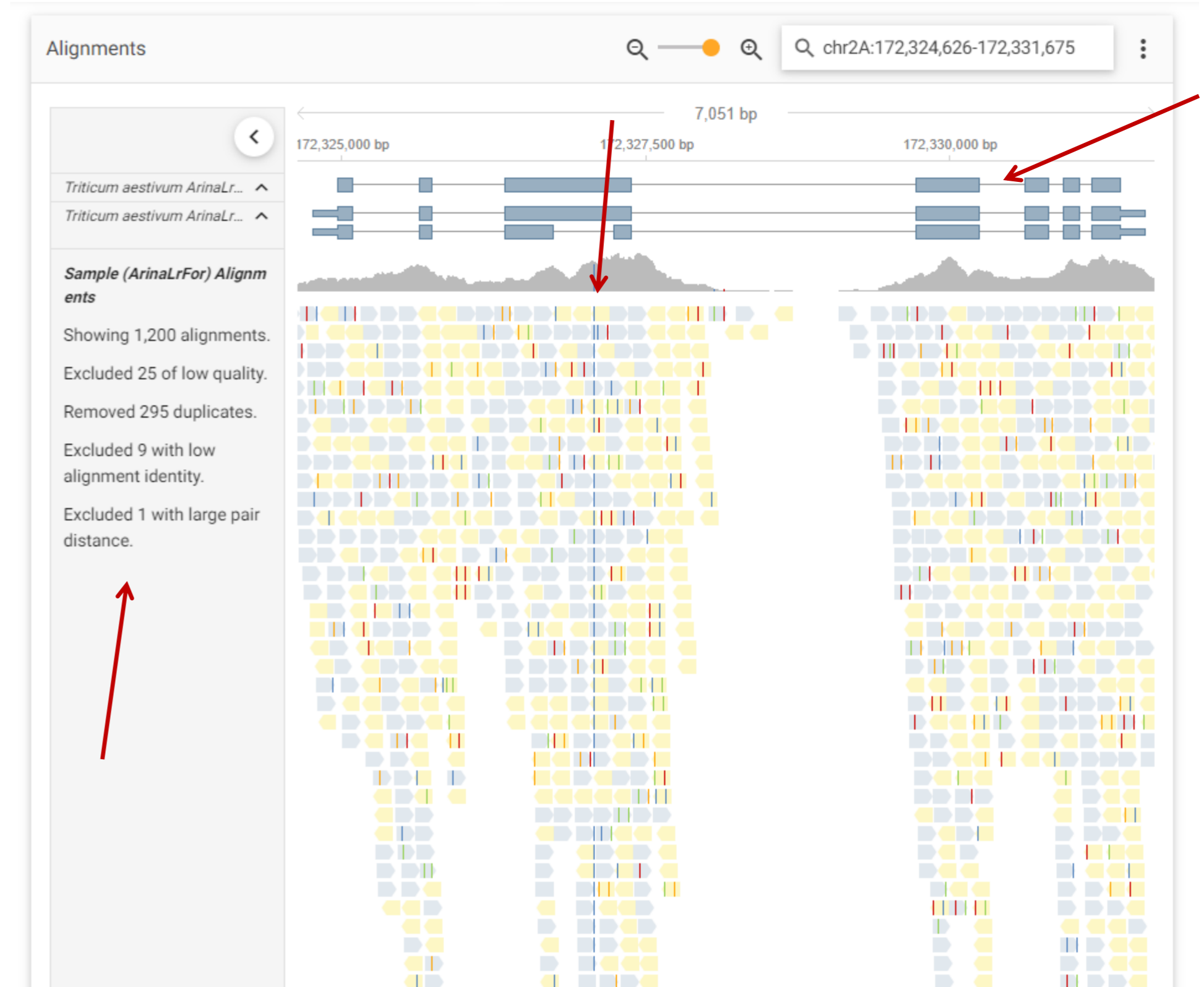
- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes



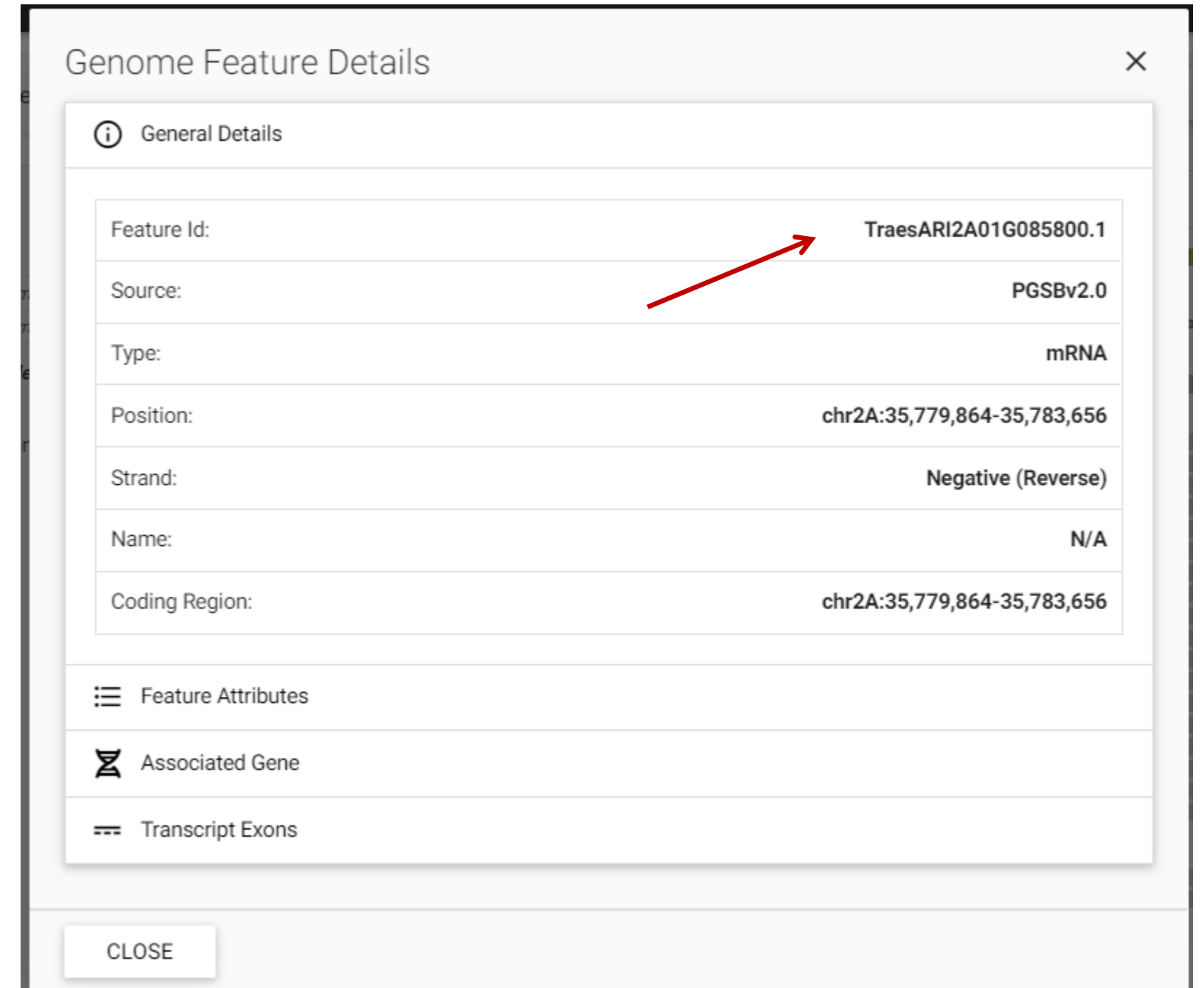
- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes
- Quickly analyze areas of interest



- 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



- 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”

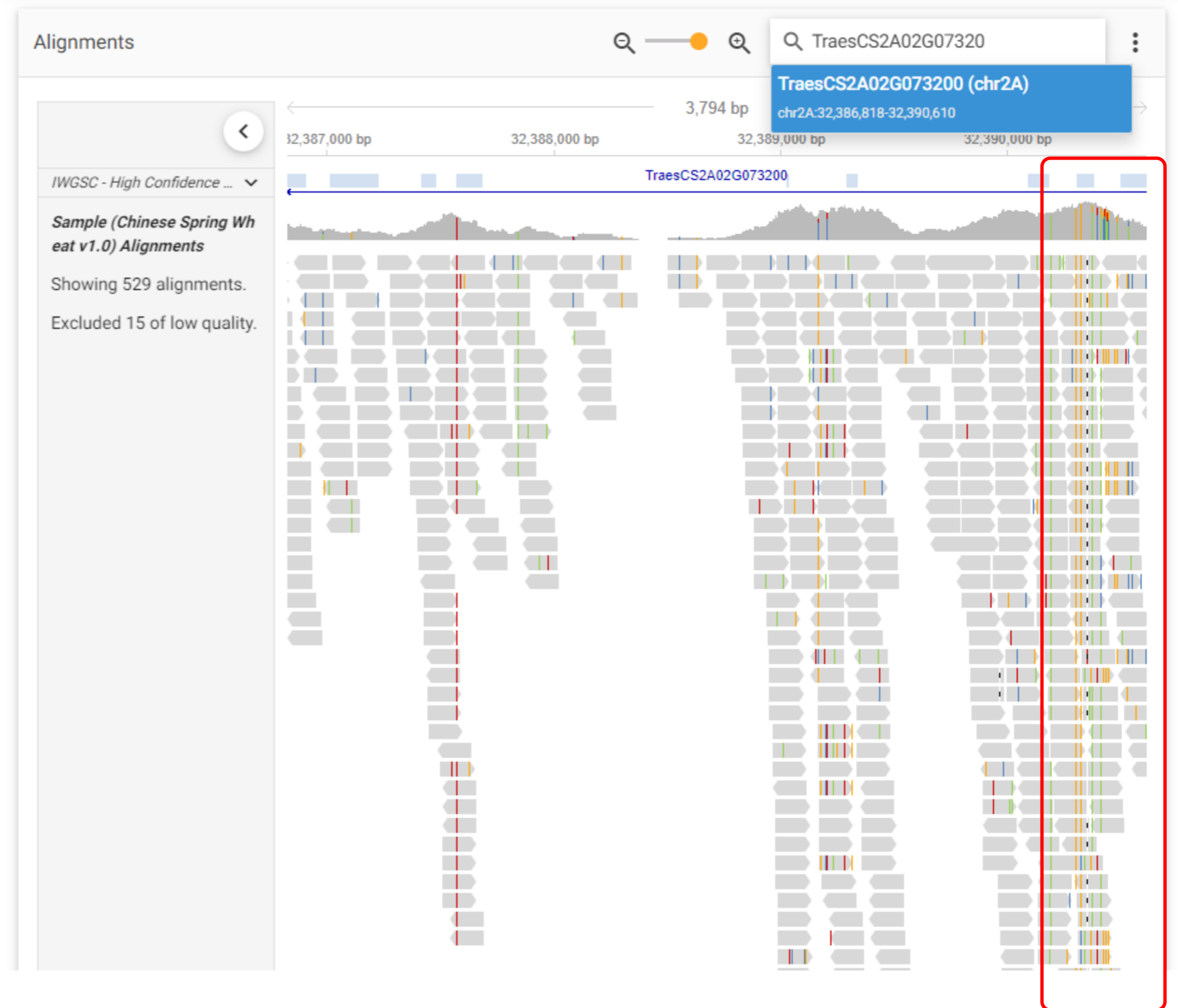


The screenshot shows a 'Genome Feature Details' window with a close button (X) in the top right corner. The window is divided into sections: 'General Details' (with an information icon), 'Feature Attributes', 'Associated Gene', and 'Transcript Exons'. The 'General Details' section contains a table with the following data:

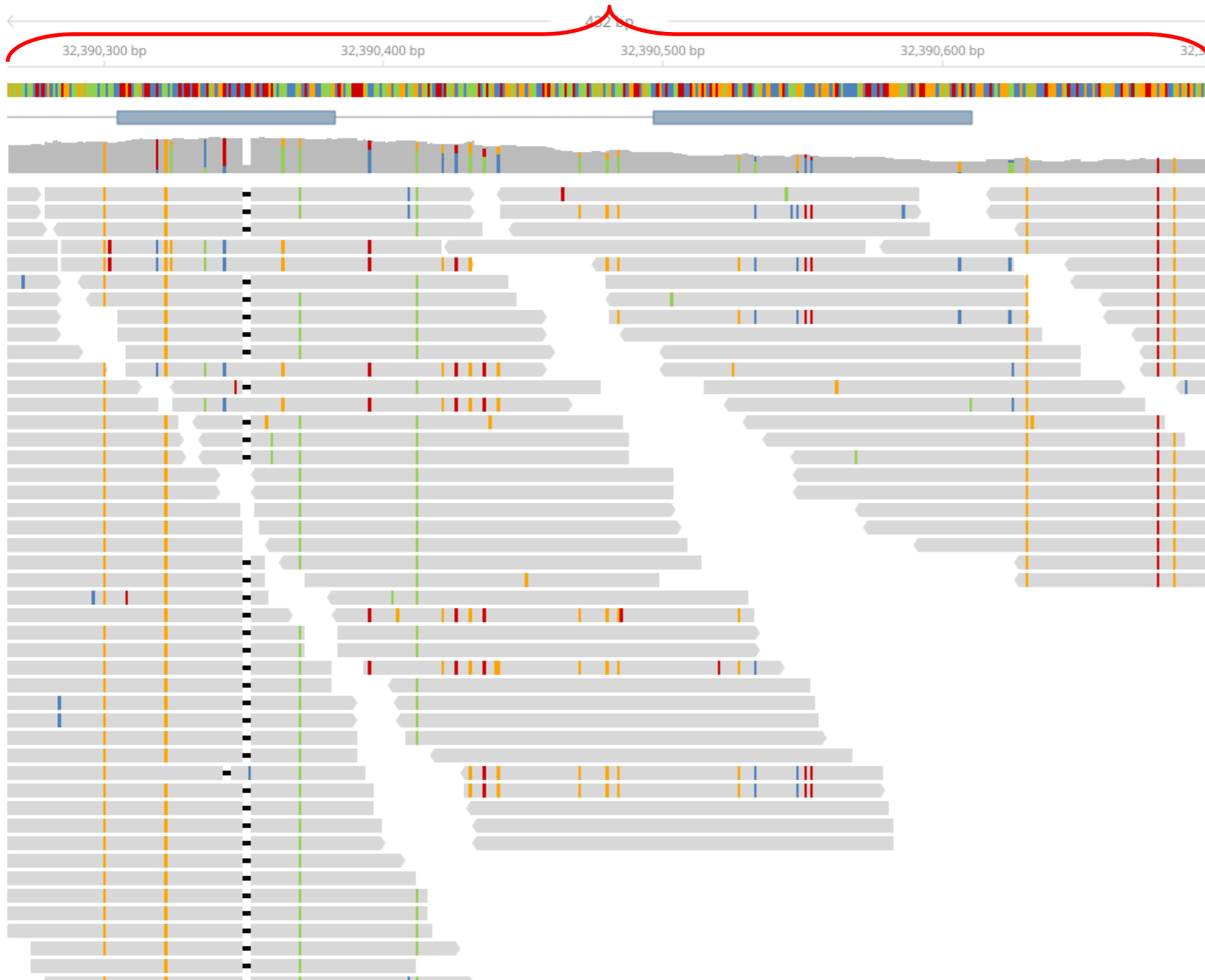
Feature Id:	TraesARI2A01G085800.1
Source:	PGSBv2.0
Type:	mRNA
Position:	chr2A:35,779,864-35,783,656
Strand:	Negative (Reverse)
Name:	N/A
Coding Region:	chr2A:35,779,864-35,783,656

A red arrow points to the 'Feature Id' field. Below the table are three expandable sections: 'Feature Attributes', 'Associated Gene', and 'Transcript Exons'. A 'CLOSE' button is located at the bottom left of the window.

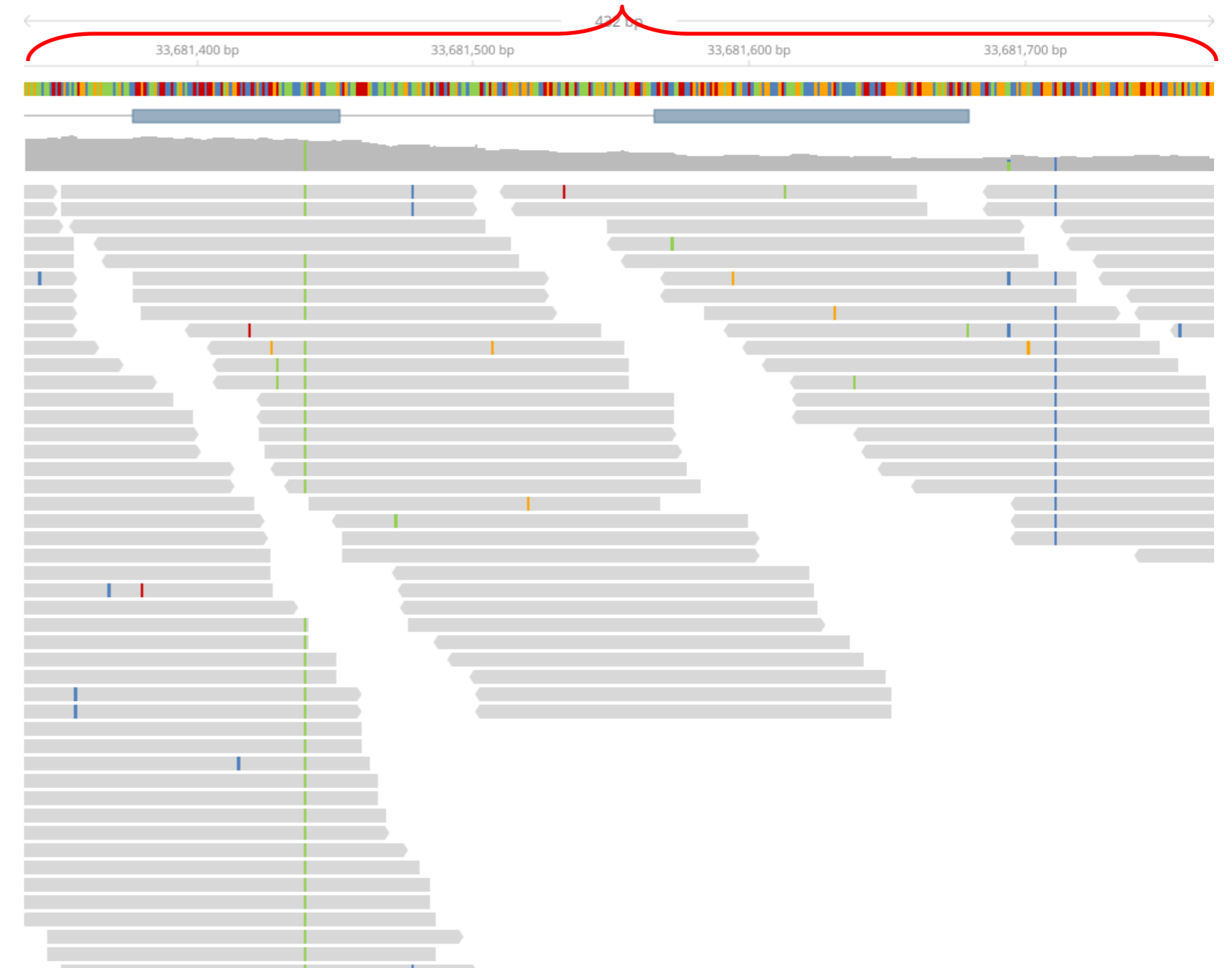
- Quickly browse and visualize aligned reads and depths from samples of any size, anywhere in the genome
- Reference projected & de-novo gene annotations from 10+ Wheat Genomes
- Quickly analyze areas of interest
- Dynamically visualize impact of filtering options
- Access gene transcript details
- Identified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



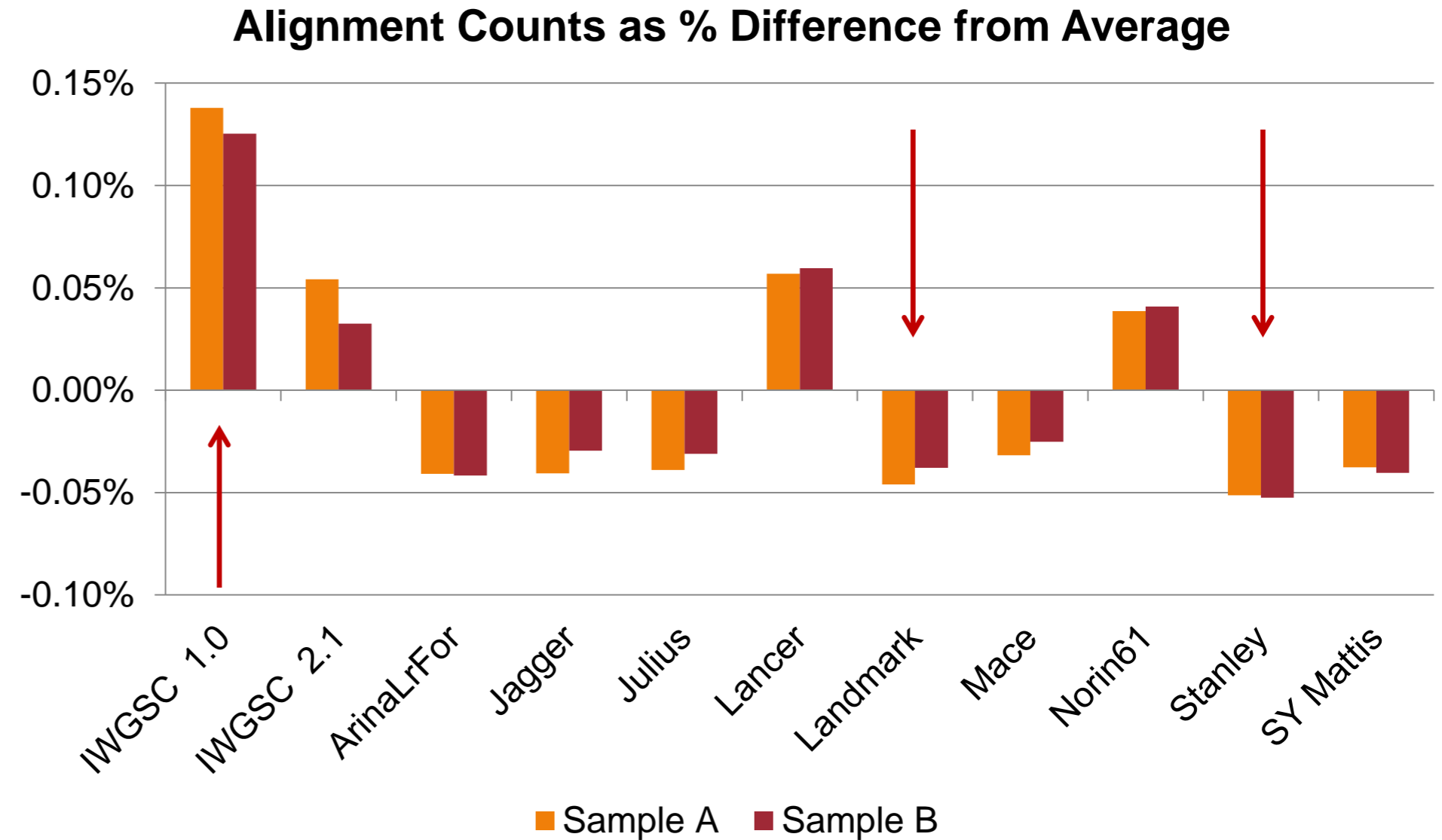
IWGSC “TraesCS2A02G073200” (Showing: chr2A:32,390,266-32,390,696)



Landmark “TraesLDM2A01G082900” (Equivalent: chr2A:33,681,338-33,681,768)



	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

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

Start New Analysis

1 Analysis Type — 2 Select Inputs — 3 Configuration

Select type of analysis — Choose files to analyze — Tune analysis options

Which type of data would you like to analyze?

DNA-Seq
 RNA-Seq

🔍 Type here

Bowtie2 Alignment
Map the reads of one or more single or paired-end samples to a reference genome using the Bowtie2 algorithm

BWA-MEM Alignment
Map the reads of one or more single or paired-end samples to a reference genome using the BWA-MEM algorithm

BWA-ALN Alignment
Map the reads of one or more single or paired-end samples to a reference genome using the BWA-ALN algorithm

General Variant Detection
Detect and call variants in one or more samples by comparing their aligned sequences to the reference

GATK Variant Detection (HaplotypeCaller)
Call germline SNPs and indels via local re-assembly of haplotypes using GATK's HaplotypeCaller

Analyze Coverage

CONTINUE →

- 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

Start New Analysis
✕

✓ Analysis Type
Select type of analysis

✓ Select Inputs
Choose files to analyze

3 Configuration
Tune analysis options

✓ Read Alignments Pre-Processing

De-duplication: ? Remove read duplicates at the same position and CIGAR

Min Mapping Quality: ? Phred: 30 (99.9%)

Min Alignment Identity: ? 90%

Max Pairing Distance: ? 1 Mbp

✓ GATK HaplotypeCaller Options

Base Call Quality: ? Phred: 10 (90%)

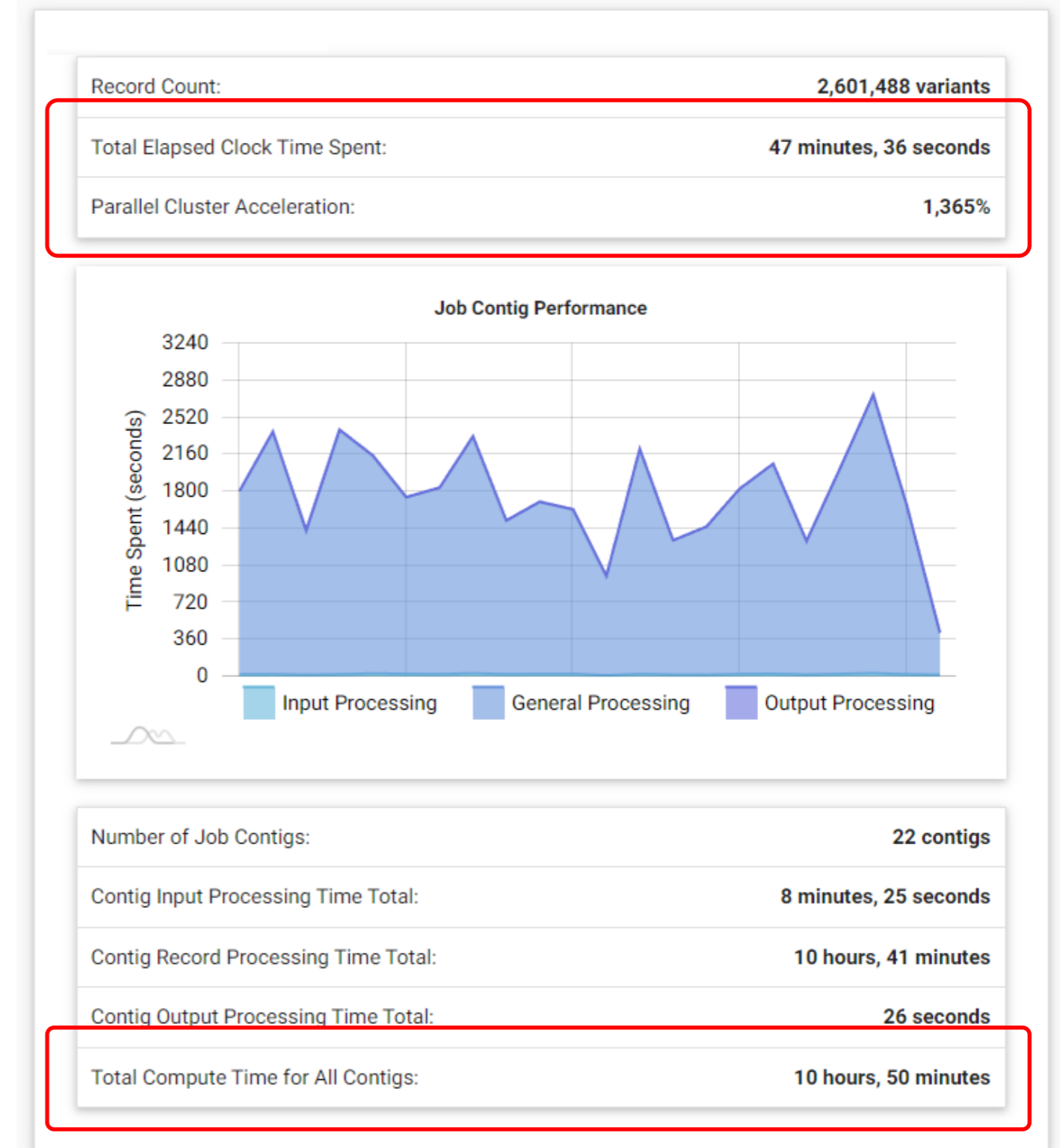
Minimum Confidence Threshold: ? Phred: 30 (99.9%)

Use Soft Clipped Bases: ? Include soft clipped bases of the reads in the analysis

PCR Indel Model: ?

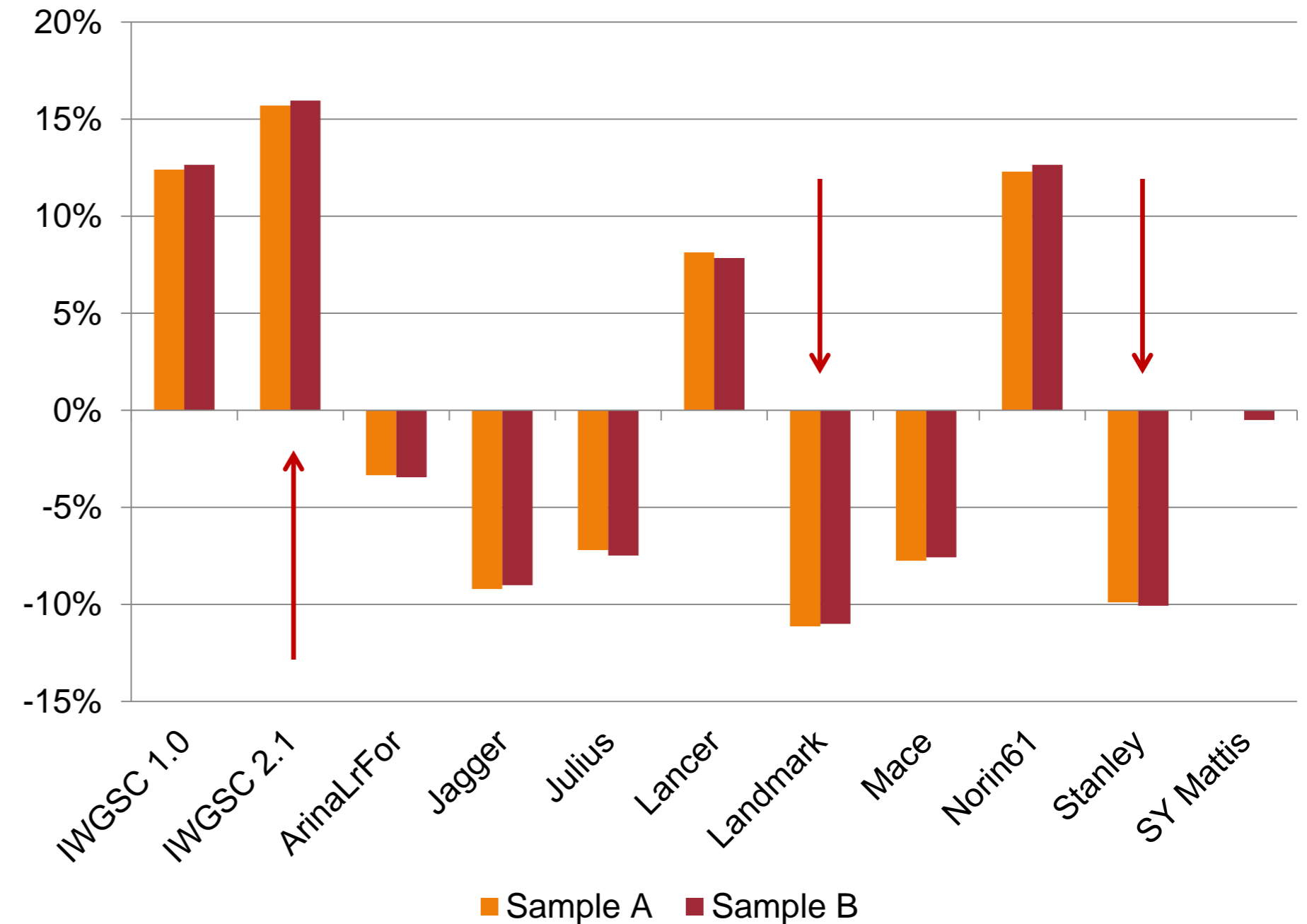
← PREVIOUS
START ANALYSIS ▶

- 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



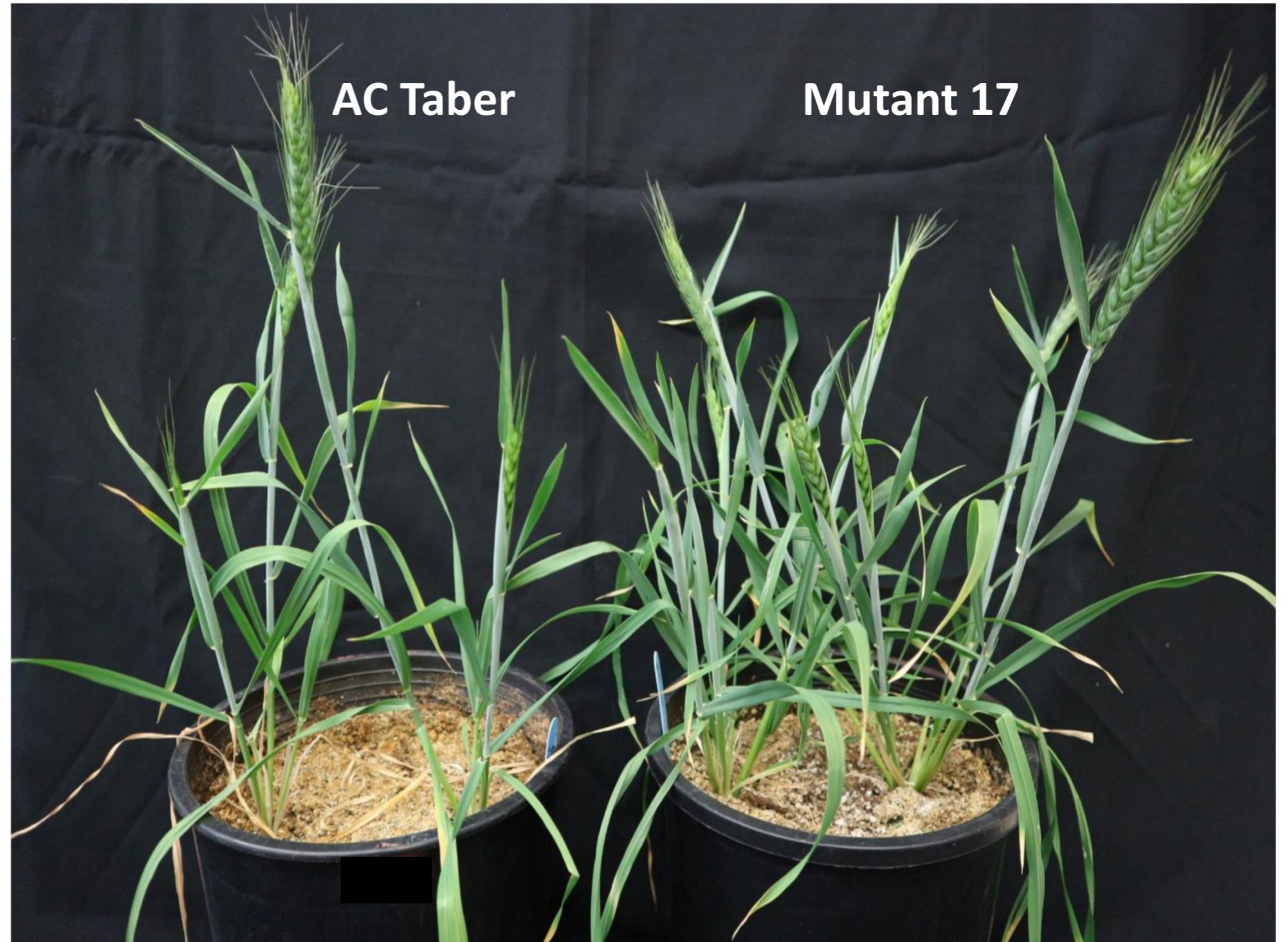
	Sample A	Sample B
IWGSC 1.0	2,705,646	3,035,134
IWGSC 2.1	2,785,165	3,124,243
ArinaLrFor	2,326,655	2,601,488
Jagger	2,185,424	2,451,528
Julius	2,233,777	2,492,644
Lancer	2,603,096	2,905,634
Landmark	2,139,171	2,397,919
Mace	2,220,472	2,490,270
Norin61	2,703,332	3,035,071
Stanley	2,168,964	2,422,944
SY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

Variant Counts as % Difference from Average



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)

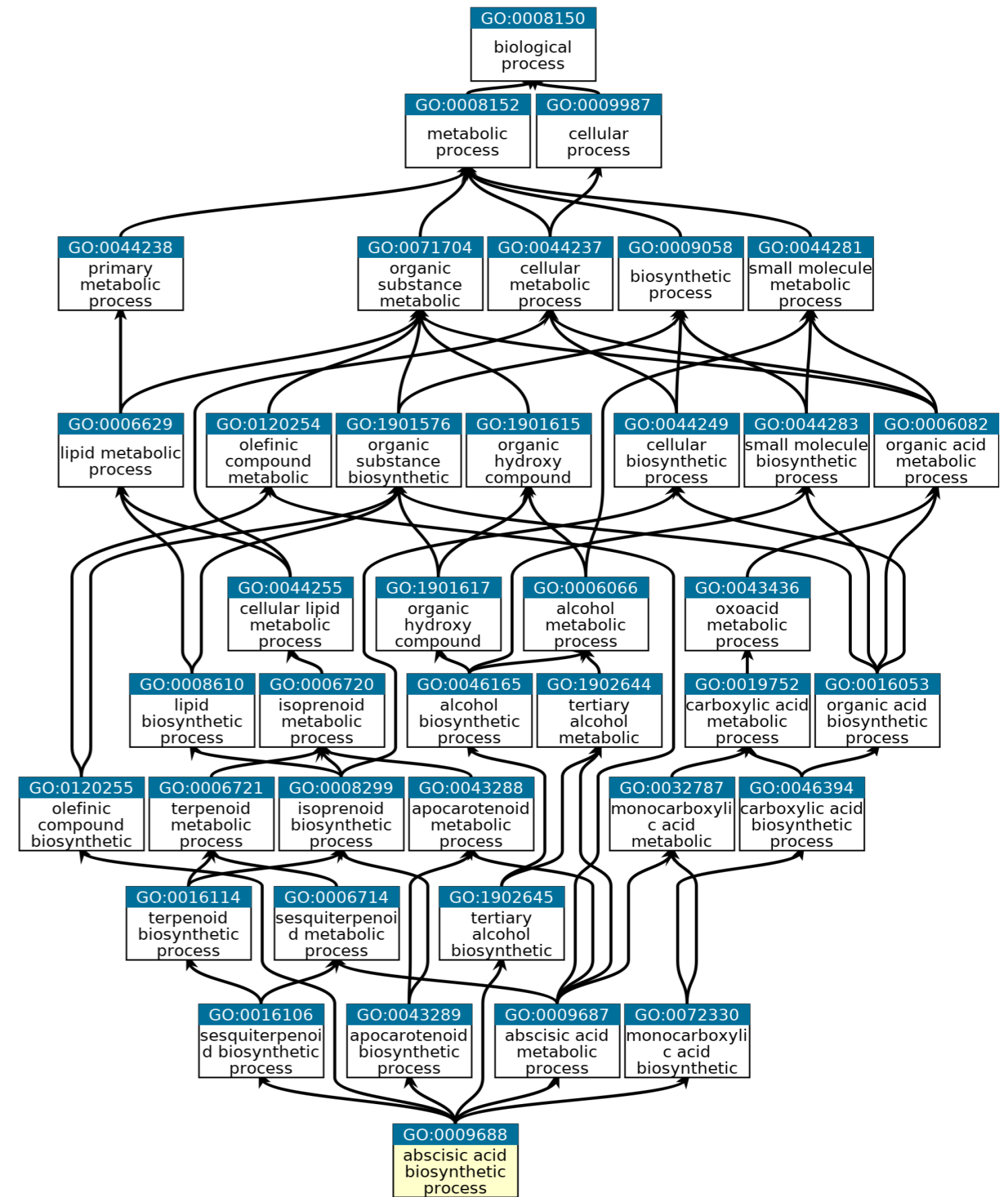
Leveraging IWGSC Functional Annotations

- 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-yl)-3-methylpenta-2, 4-dienoic acid.

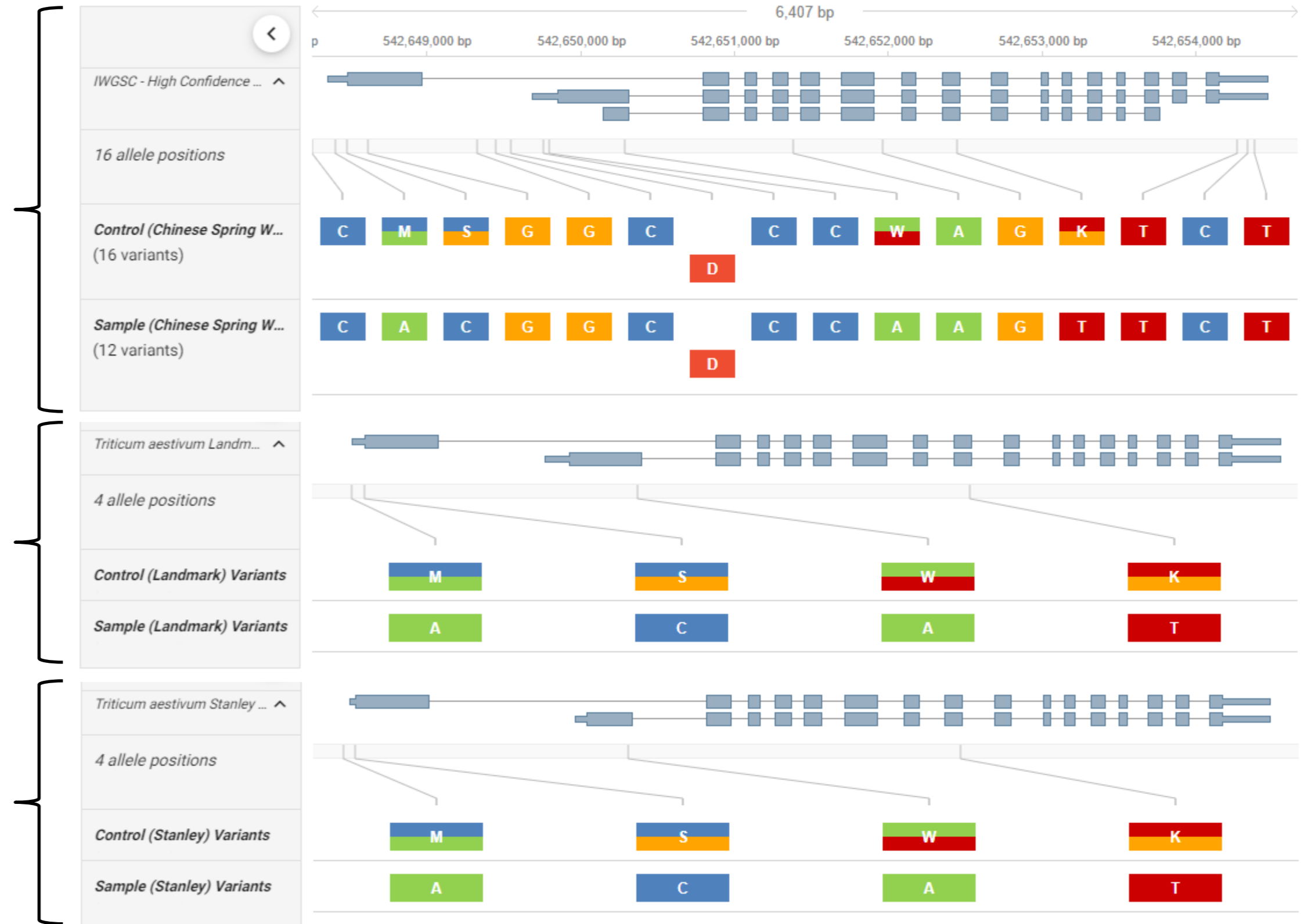


Source: <https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0009688>

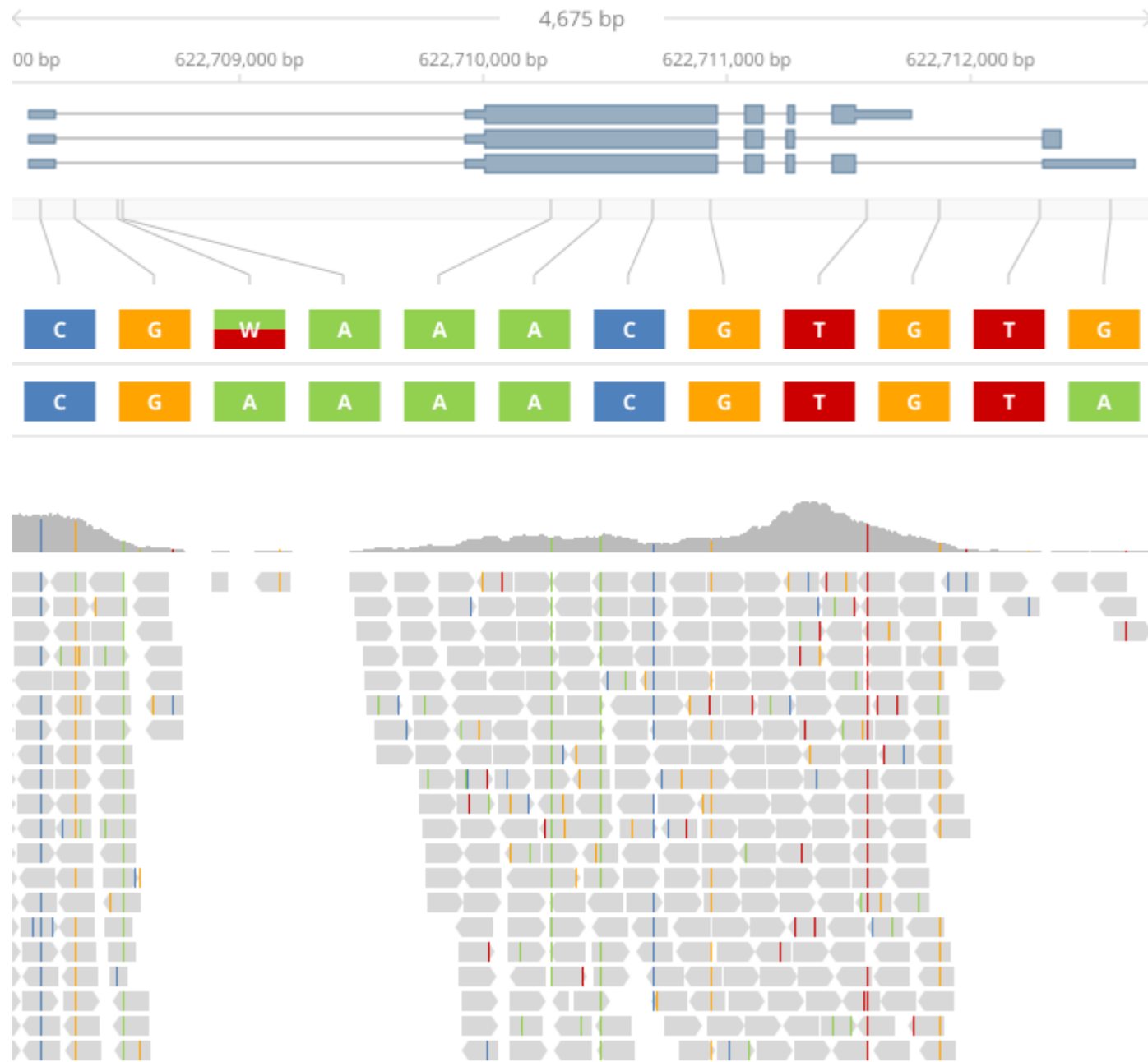
Alleles identified in
IWGSC gene:
TraesCS2A02G317000

Alleles identified in
Landmark gene:
TraesLDM2A03G00726620

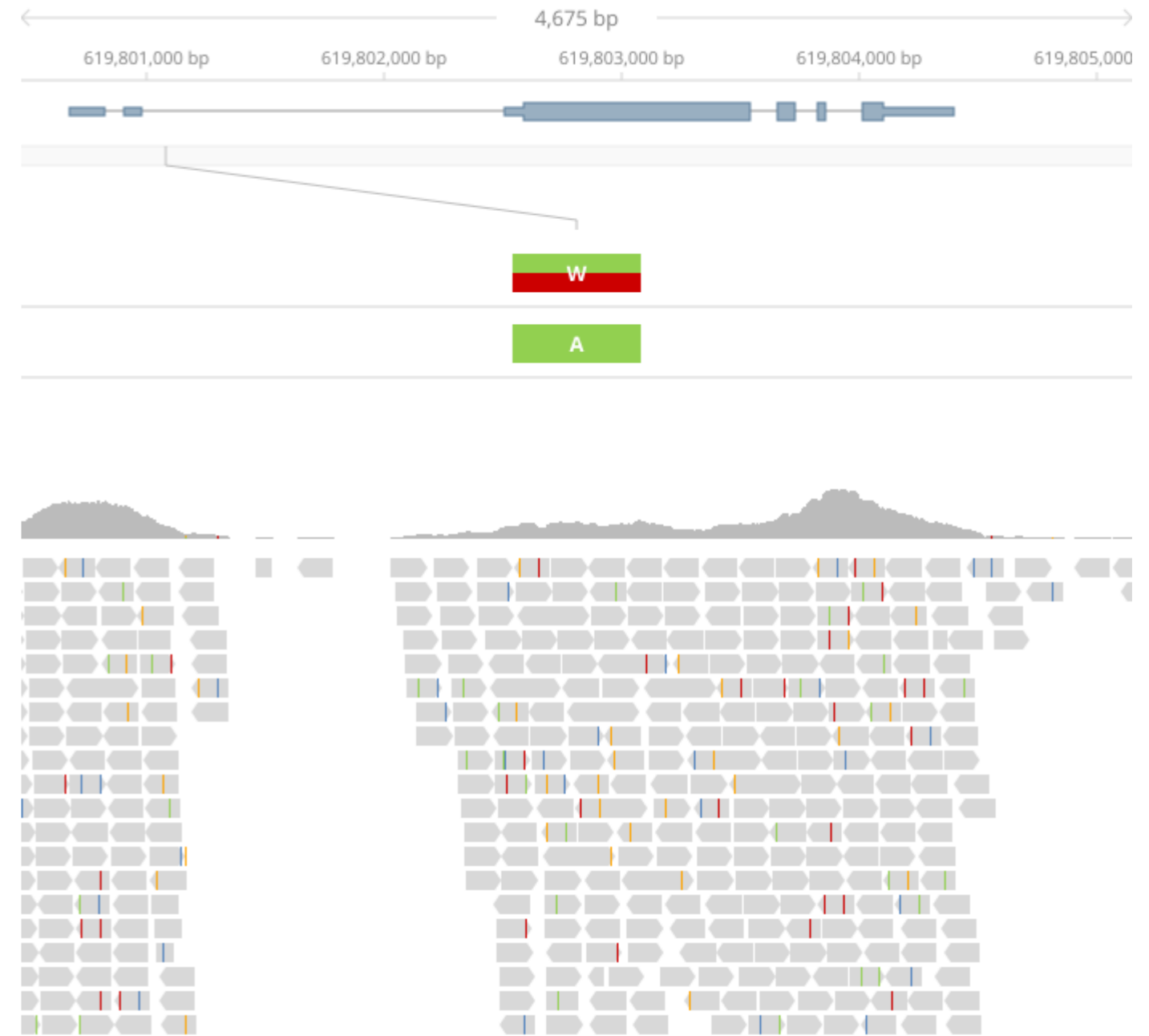
Alleles identified in
Stanley gene:
TraesSTA2A03G00722440

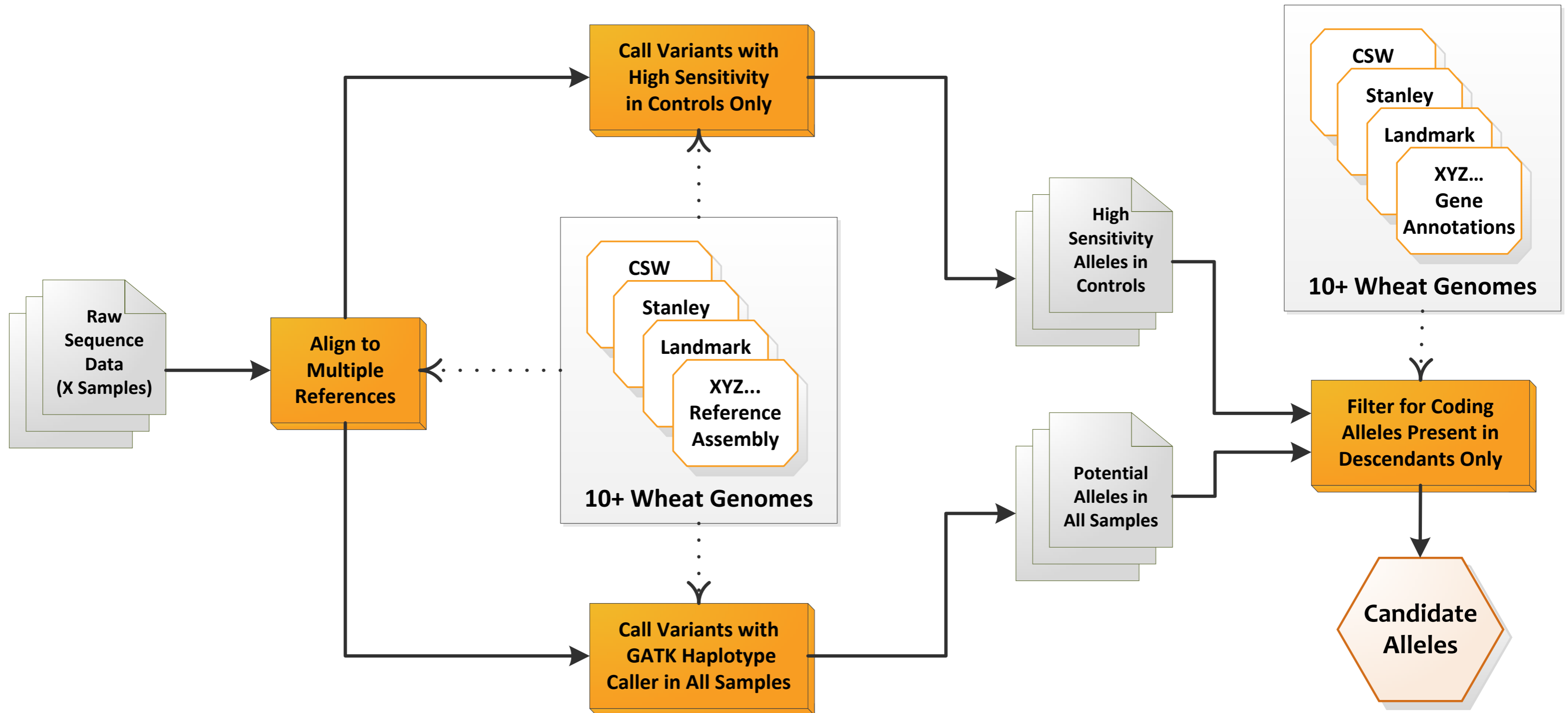


Alleles and Mapped Reads Using IWGSC (Showing Gene: TraesCS3A02G371900)

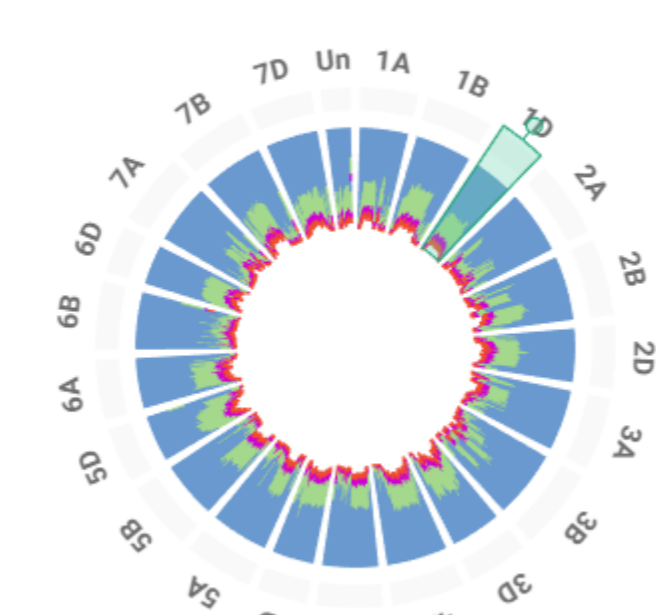


Same Samples Mapped to Landmark (Equivalent Gene: TraesLDM3A03G01467370)





ALL CONTIGS DENSITY



REPORT OPTIONS

- Show features from [2 genome feature sets](#)
- Show lines visualization
 - Triticum aestivum Landmark Genes (10+ Wheat Genomes, v1.0)
 - Triticum aestivum Landmark Genes (Ensembl, PGSB v2.1.54)
- Include only variant: homozygous positions
- Include only variant: heterozygous positions
- Exclude variants whose alt depth is not at least 5 reads
- Include SNVs for homozygous positions
- Include positions with insertions or deletions (indels)
- Include positions with rare alleles
- Render variant density using a window size of 1 Mbp
- Color positions based on the variant's alternate allele ?
- Show at each detected position the variant's coverage

INSIGHTS OPTIONS

NOTES & MARKS

ALL CONTIGS DENSITY

REPORT OPTIONS
INSIGHTS OPTIONS
NOTES & MARKS

Multi-Sample Analysis:

- Show only variants where all samples have min coverage of 5 reads
- Show only variants not present in the control (e.g. somatic) ←
- Show only variants with control coverage of at least 5 reads ←
- Show all variants found within selected sample files
- Show only variants in at least 100% of selected sample files
- Match homozygous SNVs to related heterozygous SNVs ←
- Show reference alleles in samples with no detected variant

Items to Compare
Control

- Control (Landmark) Variants Control ▾ ←
- Control (Landmark) Variants - High Sensitivity Control ▾
- Control (Landmark) Variants - Medium Sensitivity Sample ▾
- Sample (Landmark) Variants Sample ▾
- Sample (Landmark) Variants - Sensitive Sample ▾

Variants chr1D:1-499,102,108

← 50,000,000 bp 100,000,000 bp 150,000,000 bp 200,000,000 bp 250,000,000 bp 300,000,000 bp 350,000,000 bp 400,000,000 bp 450,000,000 bp 500,000,000 bp →

499,102,108 bp

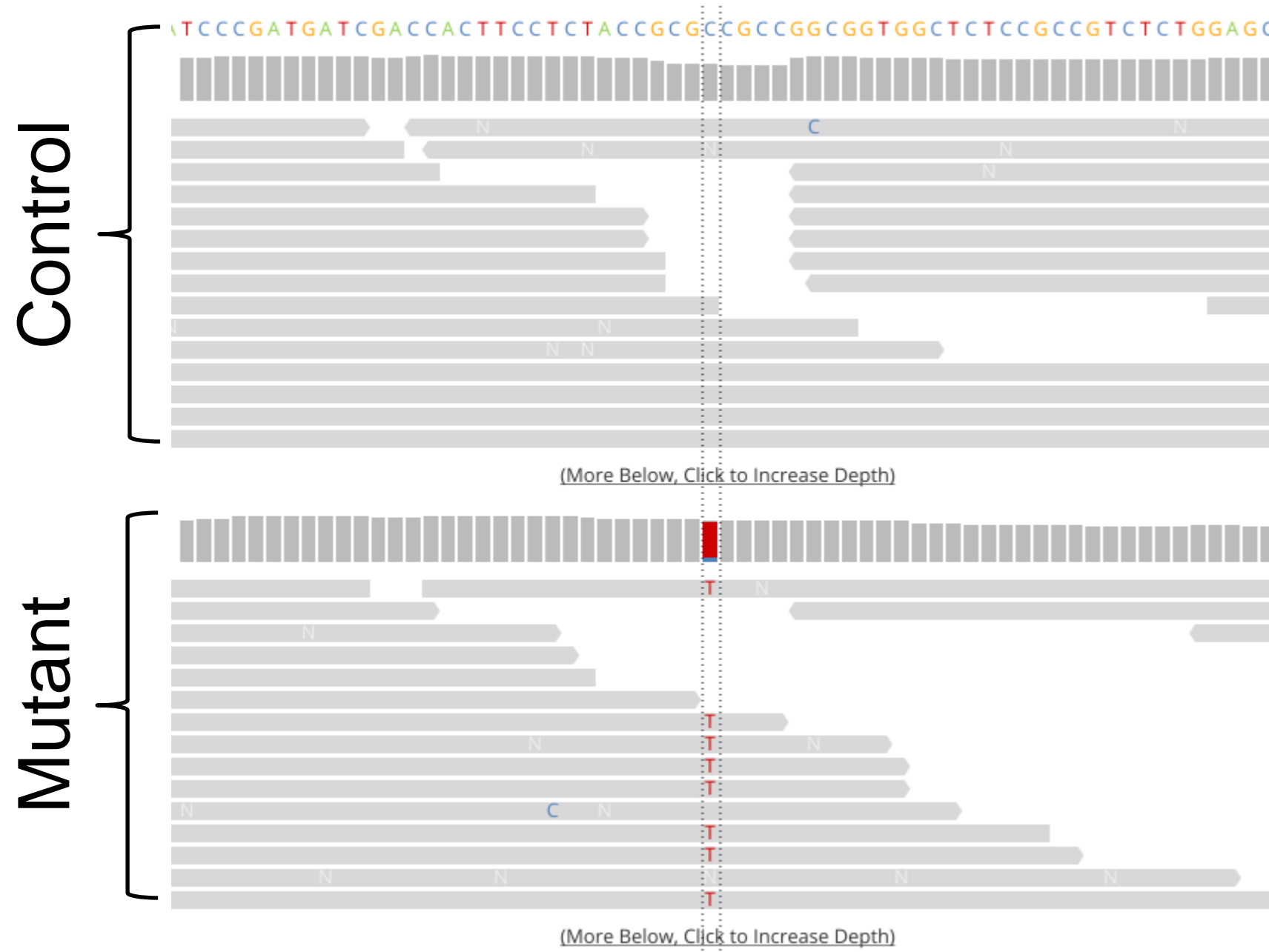
Zoom in to see genome features

Zoom in to see genome features

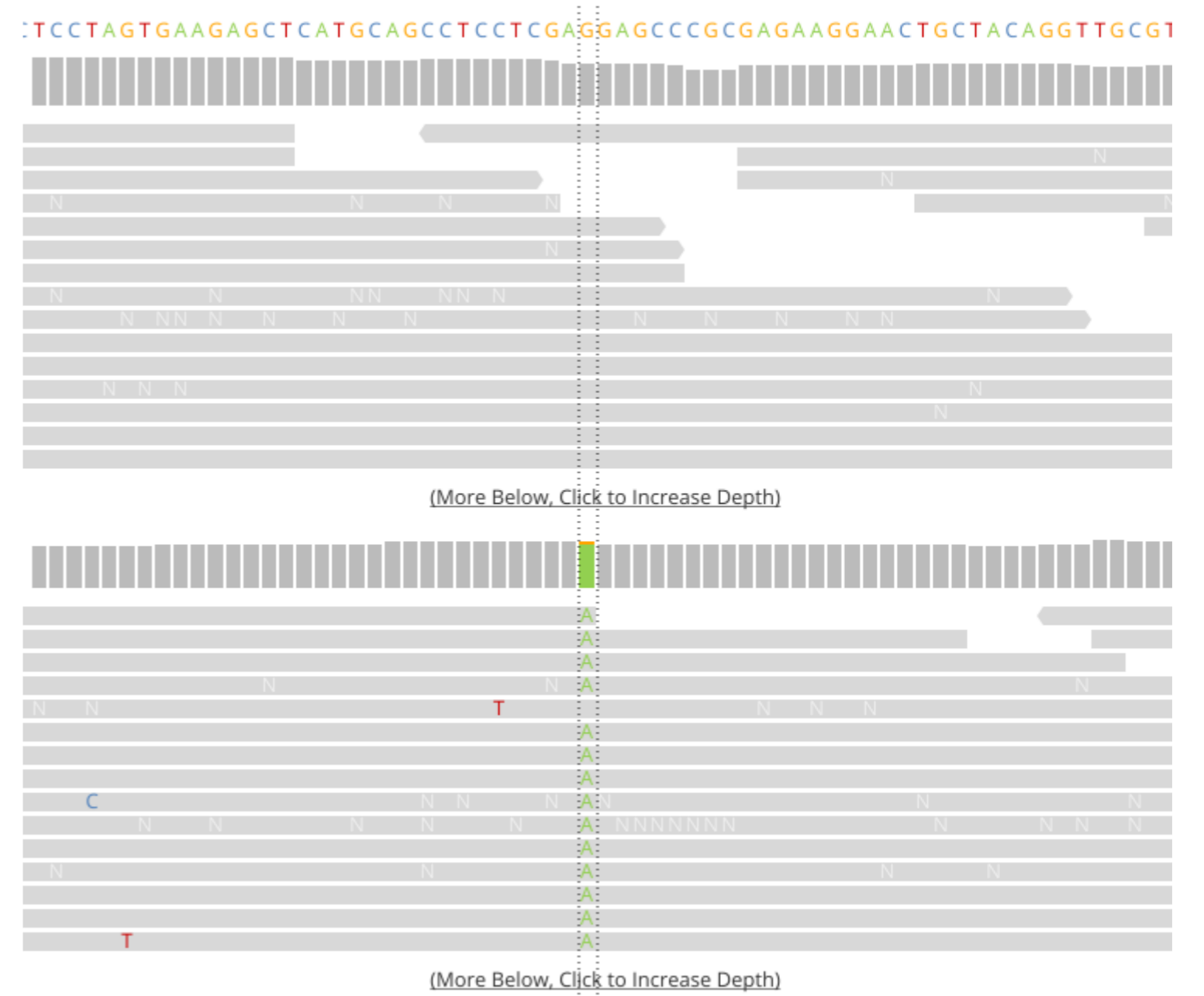
6 allele positions

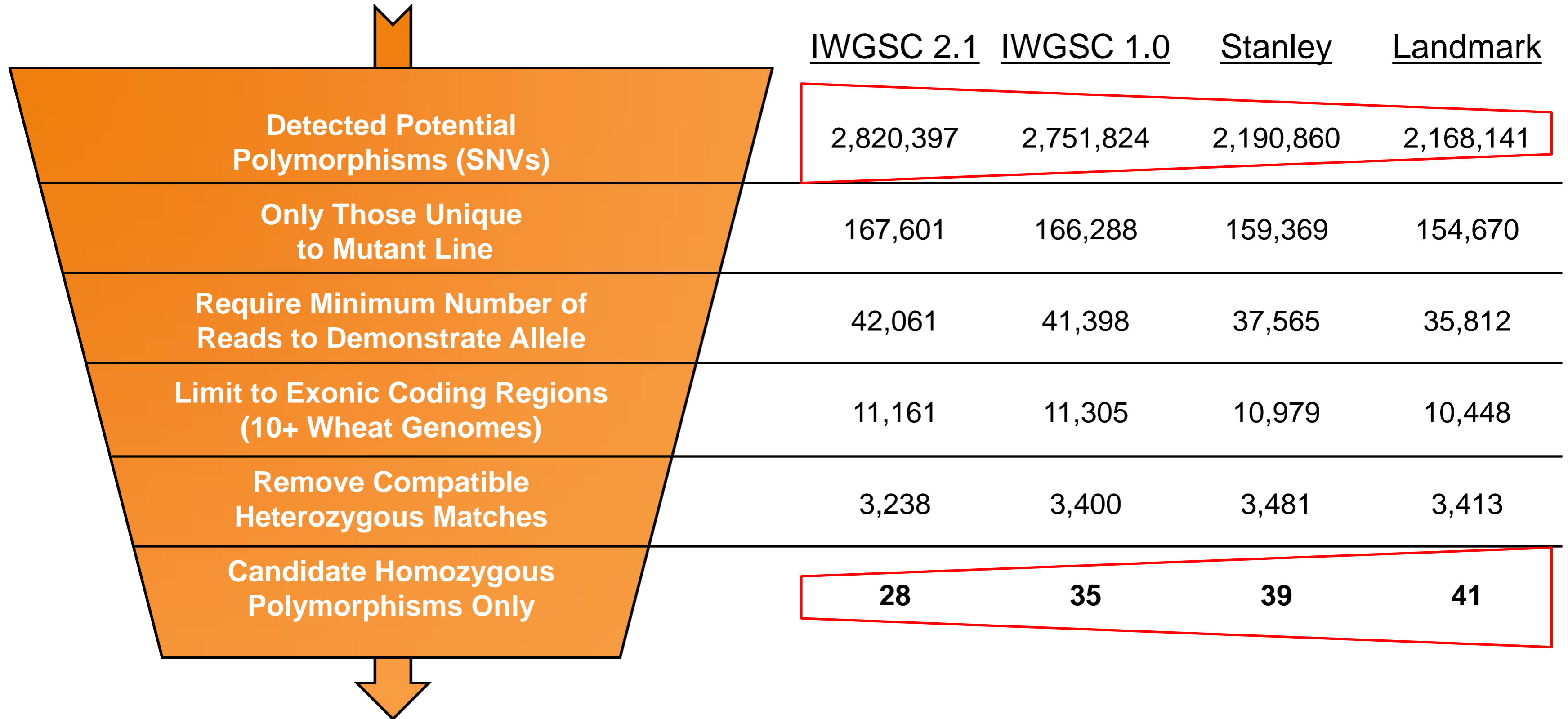
Sample (Landmark) Variants
(6 variants)

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



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





In Conclusion

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