CURIO

Alignment, Analysis, and Visualization of Wheat NGS Samples in Curio

Presented at the 2020 Plant and Animal Genome Conference

Author: Shawn Quinn (CTO, Curio Genomics)

Alignment, Analysis, and Visualization of Wheat NGS Samples in Curio

Presentation Overview

- ## What is Curio

- ## Sample Analysis with Chinese Spring Wheat

  - Read Alignment and Visualization

  - Handling Large Chromosomes

  - Mapping Algorithms & Coverage

  - Variant Analysis

  - Expression Analysis

- ## Looking Ahead

- ## Acknowledgements

What is Curio?

- Modern big data management and genomic analysis platform, fully web-based, collaboration ready

- Supports <u>both</u> bioinformatic processing and scientific interpretive analysis

- Provides 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 complex crop research solutions, including tetraploid and hexaploid wheat DNA-Seq and RNA-Seq analysis

Bread Wheat DNA-Seq:
Read Mapping and Navigation

CURIO

- Incorporates Chinese Spring Wheat (*Triticum aestivum*) reference assembly from the IWGSC

- Incorporates Chinese Spring Wheat (*Triticum aestivum*) reference assembly from the IWGSC

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

- Experiment with various alignment and read processing options without requiring any pipeline configuration

CURIO

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

Alignment Density

Display Options | Notes and Marks | Sharing

☑ Show gene and exon positions from IWGSC (HC)

☑ Show reference genome bases

☑ Show nucleotide coverage and highlight when differences exceed 20%

☑ Exclude reads and bases with quality under 90%

☐ Group reads by UMI (perfect matches), requiring consensus of 60%

☑ Remove read duplicates at the same alignment position and CIGAR

◉ Show individual read details

◯ Show read consensus sequence information

☑ Highlight mismatched bases

☐ Highlight read insertions

☐ Color forward and reverse reads differently

Reads (Showing 583 alignments of 777, excluded 1 of low quality, removed 615 duplicates)

chr1D:18,685,665-18,688,376

2,712 bp

18,685,750 bp   18,686,000 bp   18,686,250 bp   18,686,500 bp   18,686,750 bp   18,687,000 bp   18,687,250 bp   18,687,500 bp   18,687,750 bp   18,688,000 bp   18,688,250 bp

CURIO

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

- Adjust filters, PCR processing, etc. on the fly

CURIO

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

- Adjust filters, PCR processing, etc. on the fly

- Access both the transcript and functional annotations from the IWGSC inline

Alignment Density

Display Options | Notes and Marks | Sharing

☑ Show gene and exon positions from IWGSC (HC)

    IWGSC High Confidence Genes (1.1) ✔

    IWGSC Low Confidence Genes (1.1)

☑ Show reference genome bases

☑ Show nucleotide coverage and differences exceed 20%

☑ Exclude reads and bases with quality under 99.9%

☐ Group reads by UMI (perfect matches), requiring consensus of 60%

☑ Remove read duplicates at the same alignment position and CIGAR

◉ Show individual read details

    nsus sequence information

    hed bases

☐ Highlight read insertions

☐ Color forward and reverse reads differently

Reads (Showing 567 alignments of 740, excluded 40 of low quality, removed 613 duplicates)

chr1D:18,685,665-18,688,376

2,712 bp

18,685,750 bp | 18,686,000 bp | 18,686,250 bp | 18,686,500 bp | 18,686,750 bp | 18,687,000 bp | 18,687,250 bp | 18,687,500 bp | 18,687,750 bp | 18,688,000 bp | 18,688,250 bp

Gene: TraesCS1D02G038700
TraesCS1D02G038700.2 (Exon 2)
F-box family protein
Strand: Sense (+)
Exon Start: 18,686,645
Exon End: 18,688,162
Transcript Start: 18,685,856
Transcript End: 18,688,162
Coding Start: 18,685,867
Coding End: 18,687,841

CURIO

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

- Adjust filters, PCR processing, etc. on the fly

- Access both the transcript and functional annotations from the IWGSC inline

- Quickly search and navigate the genome using the IWGSC annotation data as well

Alignment Density

7D   Un   1A
7B   1B
7A   1D
6D   2A
6B   2B
6A   2D
5D   3A
5B   3D
5A 4D 4B 4A 3B

Display Options   Notes and Marks   Sharing

- ☑ Show gene and exon positions from IWGSC (HC)
- ☑ Show reference genome bases
- ☑ Show nucleotide coverage and highlight when differences exceed 20%
- ☑ Exclude reads and bases with quality under 99.9%
- ☐ Group reads by UMI (perfect matches), requiring consensus of 60%
- ☑ Remove read duplicates at the same alignment position and CIGAR

- ● Show individual read details
- ○ Show read consensus sequence information

- ☑ Highlight mismatched bases
- ☐ Highlight read insertions
- ☐ Color forward and reverse reads differently

Reads (Showing 83 alignments, excluded 17 of low quality, removed 49 duplicates)

⊟ ━━━━━━━━●━ ⊞   glutenin   🔍

377 bp

21,714,500 bp   21,714,600 bp   21,714,700 bp

TRAESCS1A02G008000 (chr1A)
low molecular weight **glutenin** subunit

TRAESCS1A02G010900 (chr1A)
low molecular weight **glutenin** subunit

TRAESCS1A02G039600 (chr1A)
high molecular weight **glutenin** subunit

TRAESCS1A02G317500 (chr1A)
high molecular weight **glutenin** subunit

TRAESCS1B02G011600 (chr1B)
low molecular weight **glutenin** subunit

CURIO

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

- Adjust filters, PCR processing, etc. on the fly

- Access both the transcript and functional annotations from the IWGSC inline

- Quickly search and navigate the genome using the IWGSC annotation data as well

- Conveniently access reference or consensus sequences, and read coverage details

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

- Adjust filters, PCR processing, etc. on the fly

- Access both the transcript and functional annotations from the IWGSC inline

- Quickly search and navigate the genome using the IWGSC annotation data as well

- Conveniently access reference or consensus sequences, and read coverage details

- Navigate and collaborate through annotated comments

Handling Large Chromosomes

Alignment, Analysis, and Visualization of Wheat NGS Samples in Curio / Presented at PAG 2020
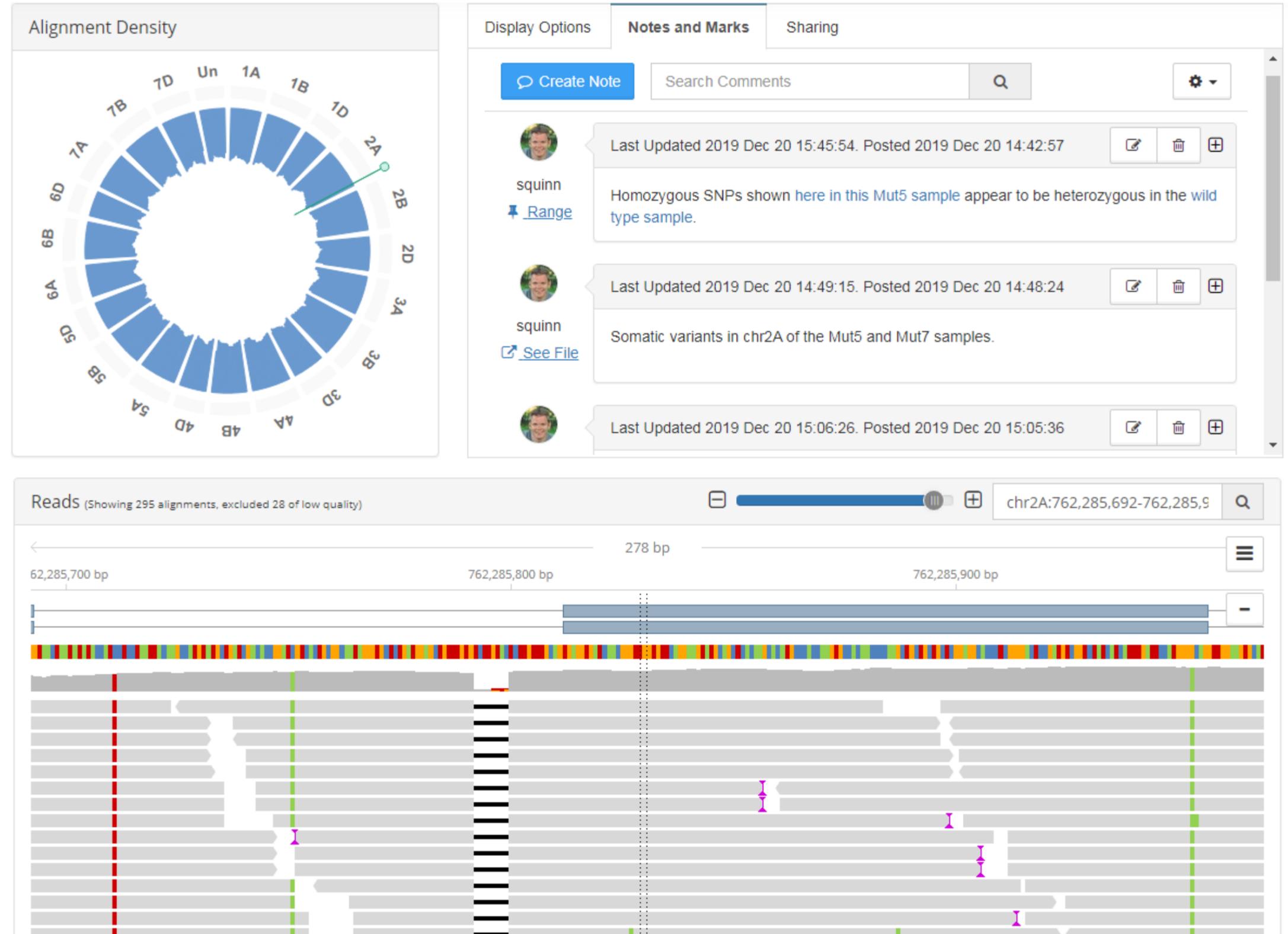
Standard Bin Strategy

- Binning strategy used in BAM index files (i.e. a "BAI" file)

- 6 levels deep and maximum bin size of: **536 million bases**

- Largest human chromosome (chr1): **249 million bases**

- Largest CSW chromosome (chr3B): **837 million bases**

- Forces a split chromosome approach or causes various tool compatibility issues

Level 0	Bin Size: 536,870,912 bps Bin Count: 1
Level 1	Bin Size: 67,108,864 bps Bin Count: 8
Level 2	Bin Size: 8,388,608 bps Bin Count: 64
Level 3	Bin Size: 1,048,576 bps Bin Count: 512
Level 4	Bin Size: 131,072 bps Bin Count: 4,096
Level 5	Bin Size: 16,384 bps Bin Count: 32,768
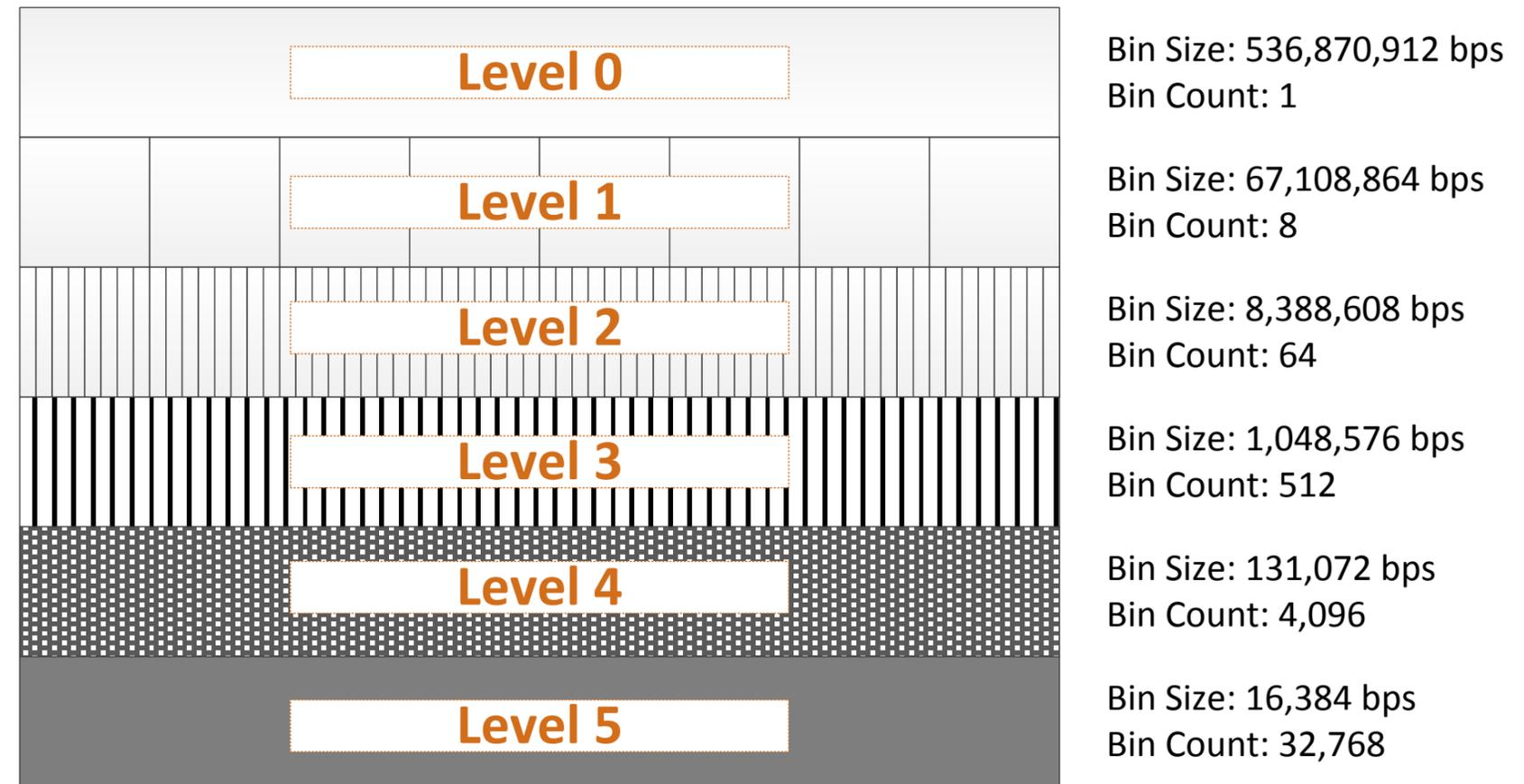
hg38: chr1

56% too large

Chinese Spring Wheat: chr3B

A Better Approach

- 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 of depth & the number of bins per level, but use the first half of each level

- Max bin size: **1 billion bases**

- Largest CSW chromosome (chr3B): **837 million bases**

Level 0
Level 1
Level 2
Level 3
Level 4
Level 5
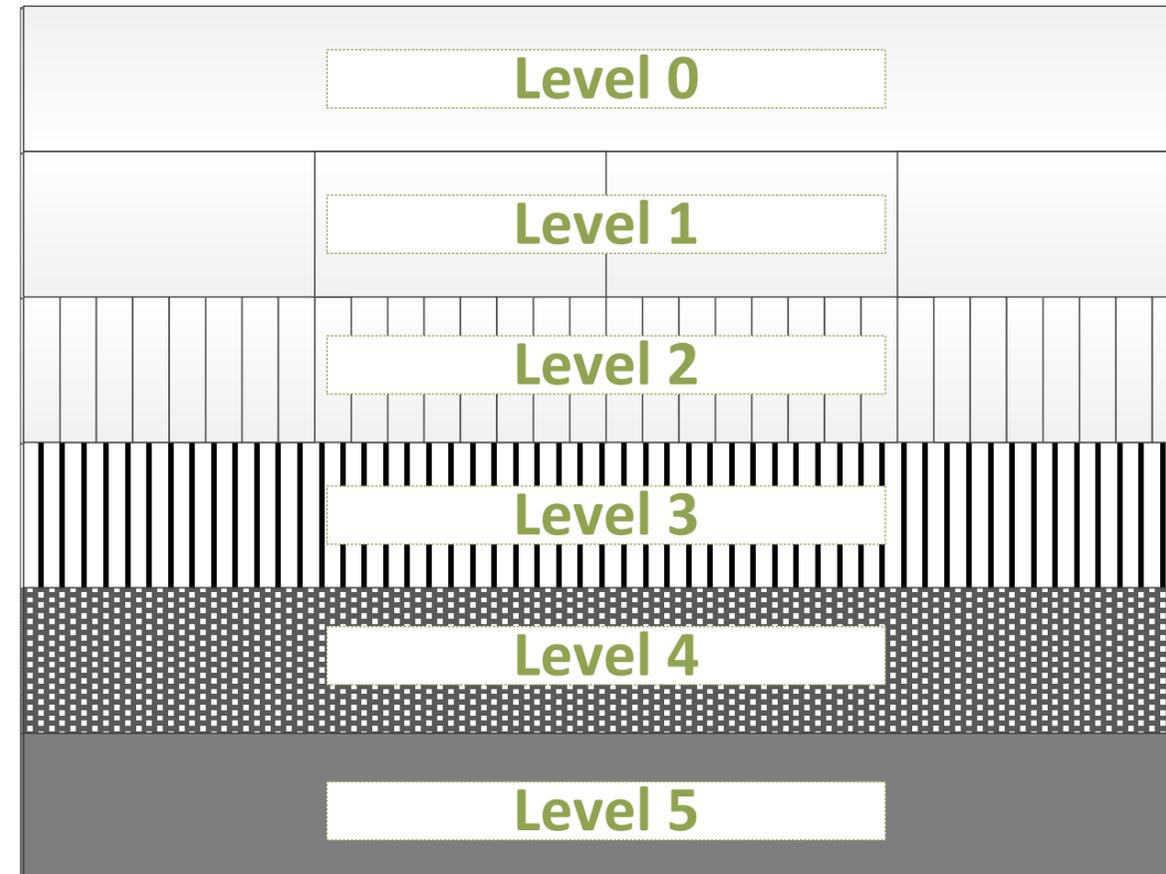
Bin Size: 1,073,741,824 bps
Bin Count: 1

Bin Size: 268,435,456 bps
Bin Count: 4 of 8

Bin Size: 33,554,432 bps
Bin Count: 32 of 64

Bin Size: 4,194,304 bps
Bin Count: 256 of 512

Bin Size: 524,288 bps
Bin Count: 2,048 of 4,096

Bin Size: 65,536 bps
Bin Count: 16,384 of 32,768

Chinese Spring Wheat: chr3B

DNA-Seq Bread Wheat:
Alignment Algorithm Coverage Impacts

Chinese Spring Wheat: Aligner Performance
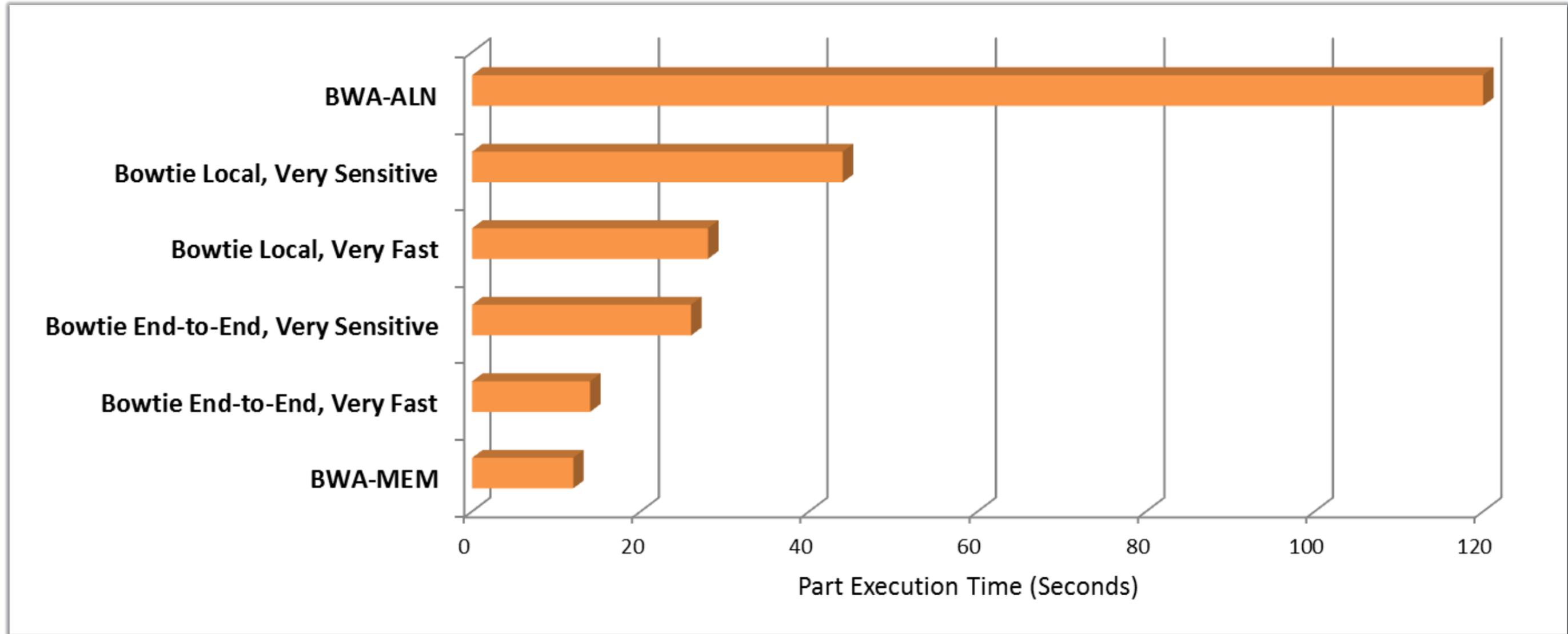
Note: the time needed for both the "bwa aln" and "bwa sampe" commands are included in the BWA-ALN metric shown above.

Bowtie2 "End to End" vs "Local" Algorithm, Exome Coverage
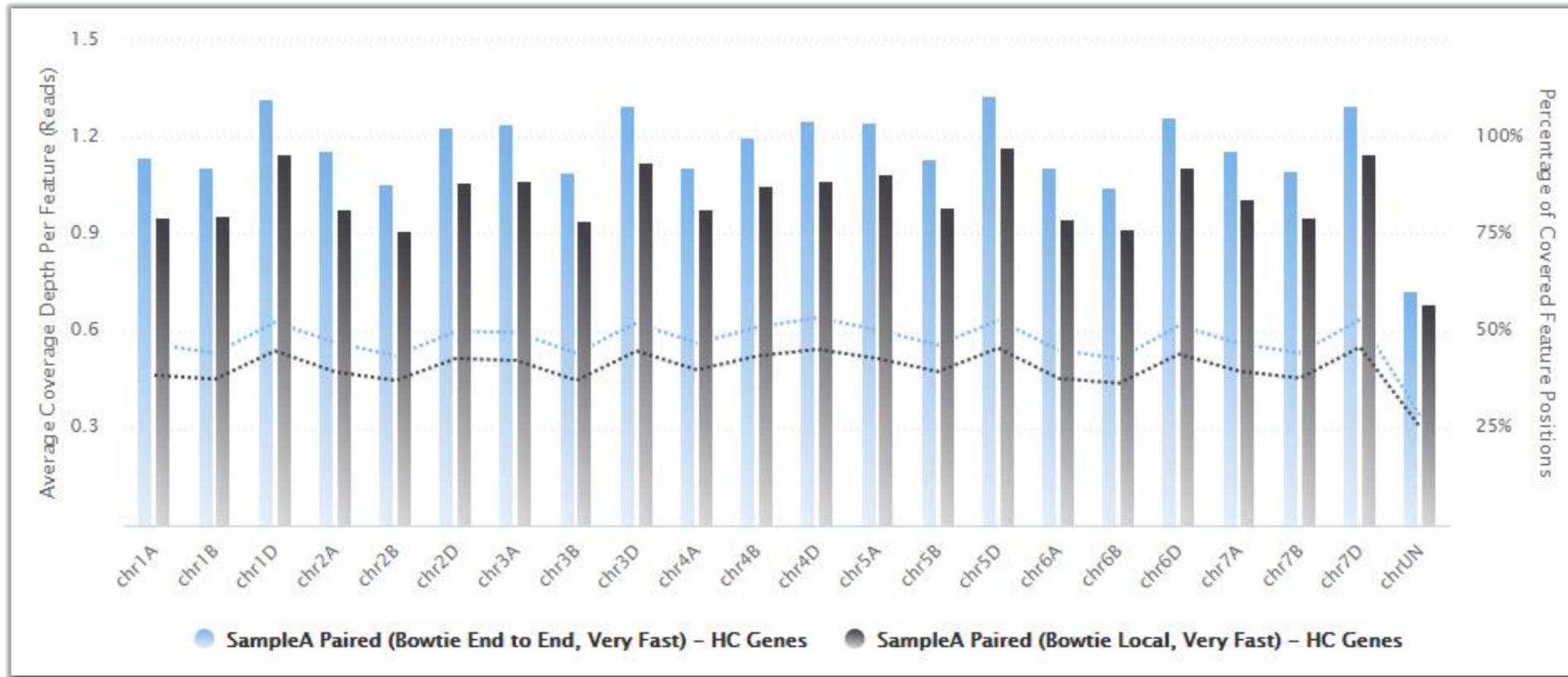
Average Exome Coverage Across all Chromosomes
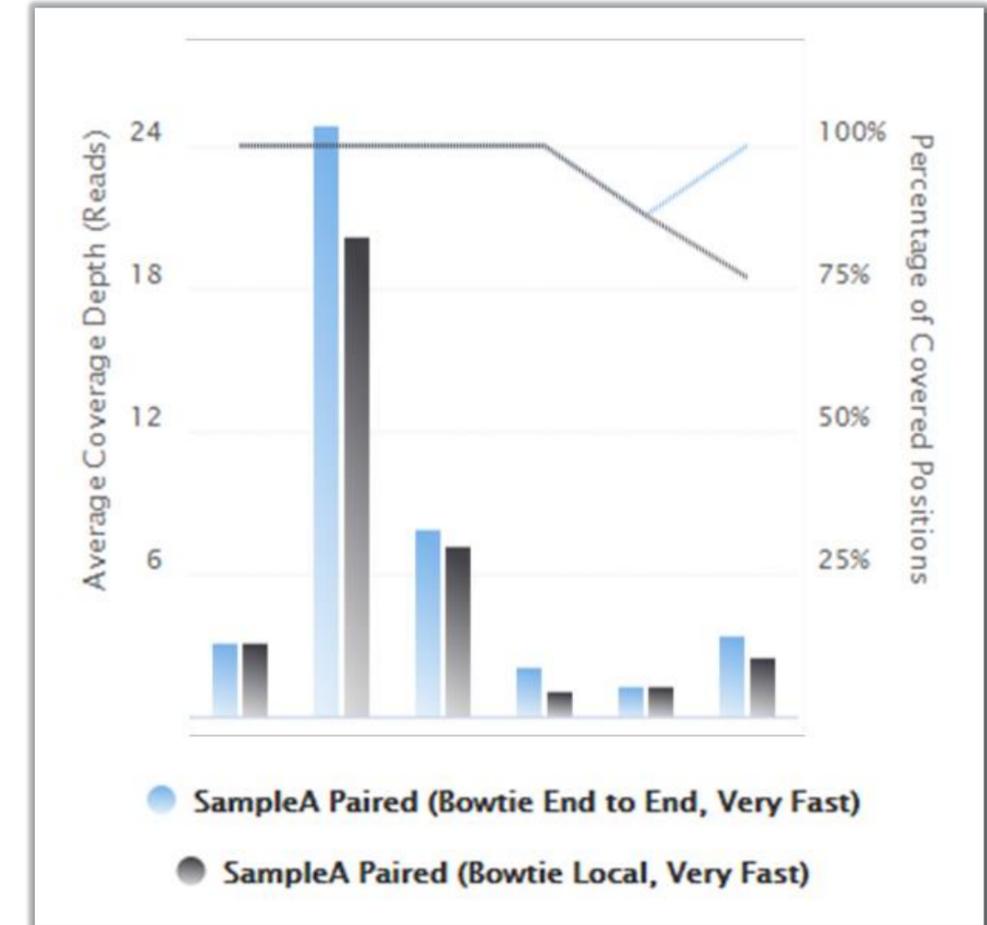
Exons of Single Gene

Note: L-lactate dehydrogenase gene
(TRAESCS1A02G238700) shown here

Bowtie2 "Very Sensitive" vs "Very Fast" Presets, Exome Coverage
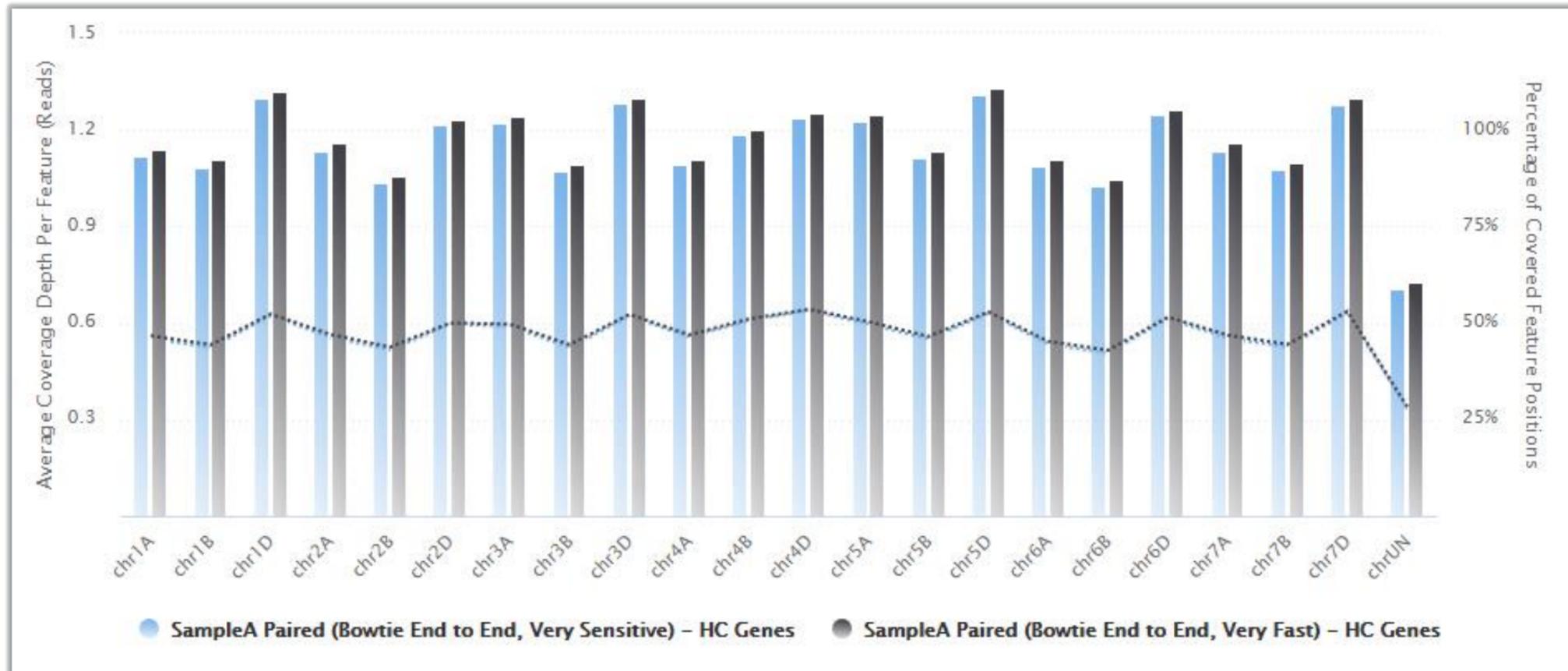
Average Exome Coverage Across all Chromosomes
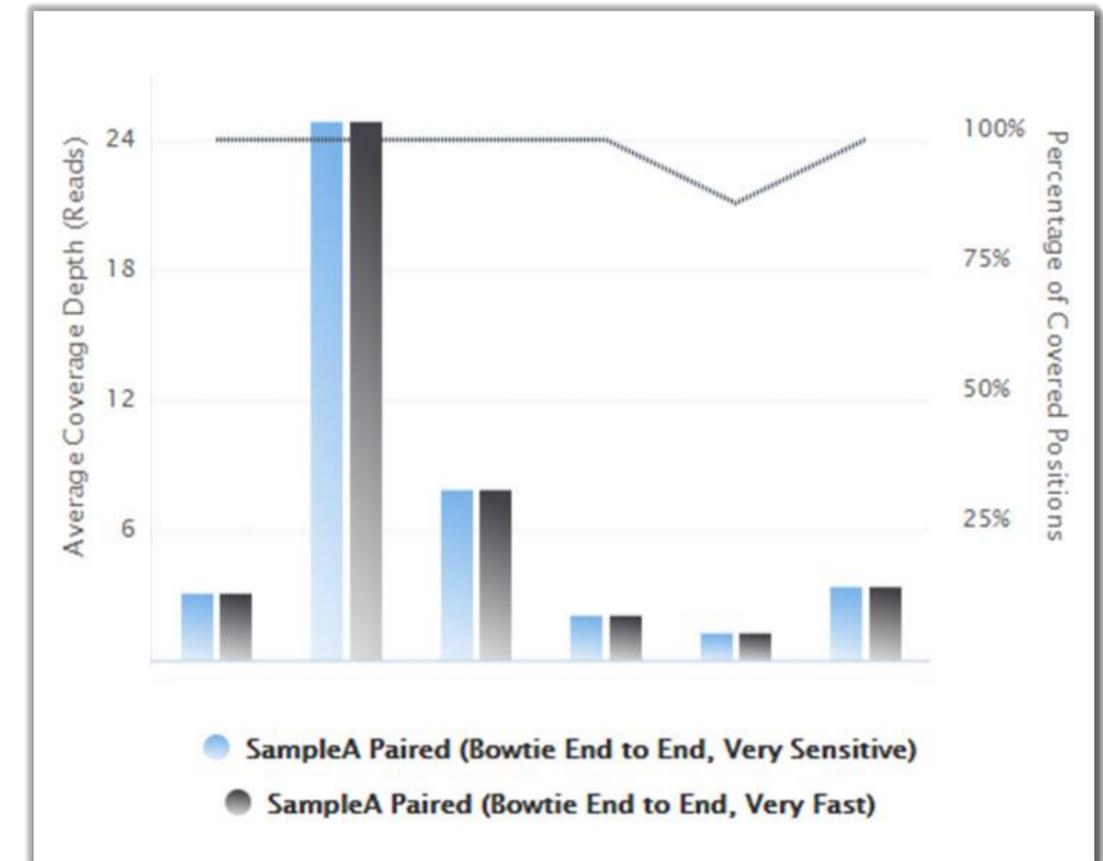
Exons of Single Gene

Note: L-lactate dehydrogenase gene
(TRAESCS1A02G238700) shown here

Alignment, Analysis, and Visualization of Wheat NGS Samples in Curio / Presented at PAG 2020

Aligner Algorithms, IWGSC "HC" Genes, Default Quality Filtering
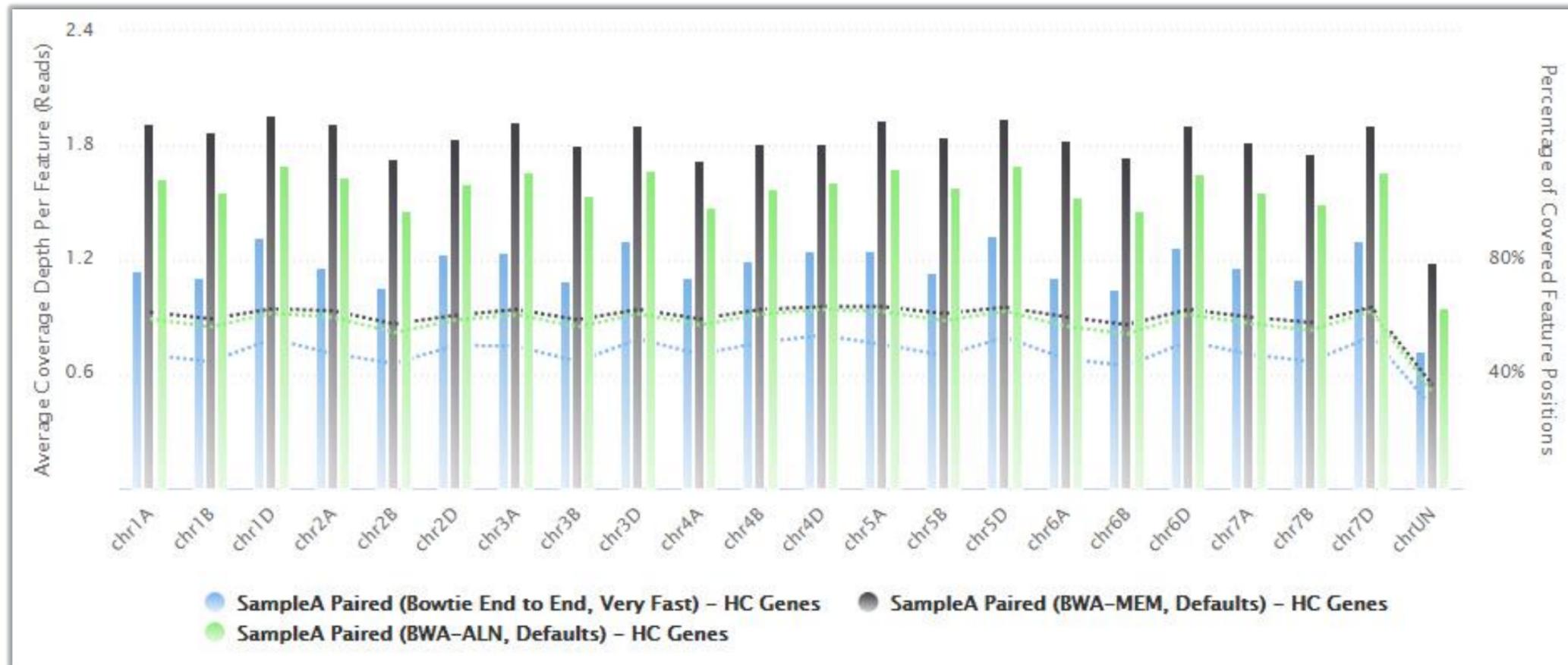
Average Exome Coverage Across all Chromosomes
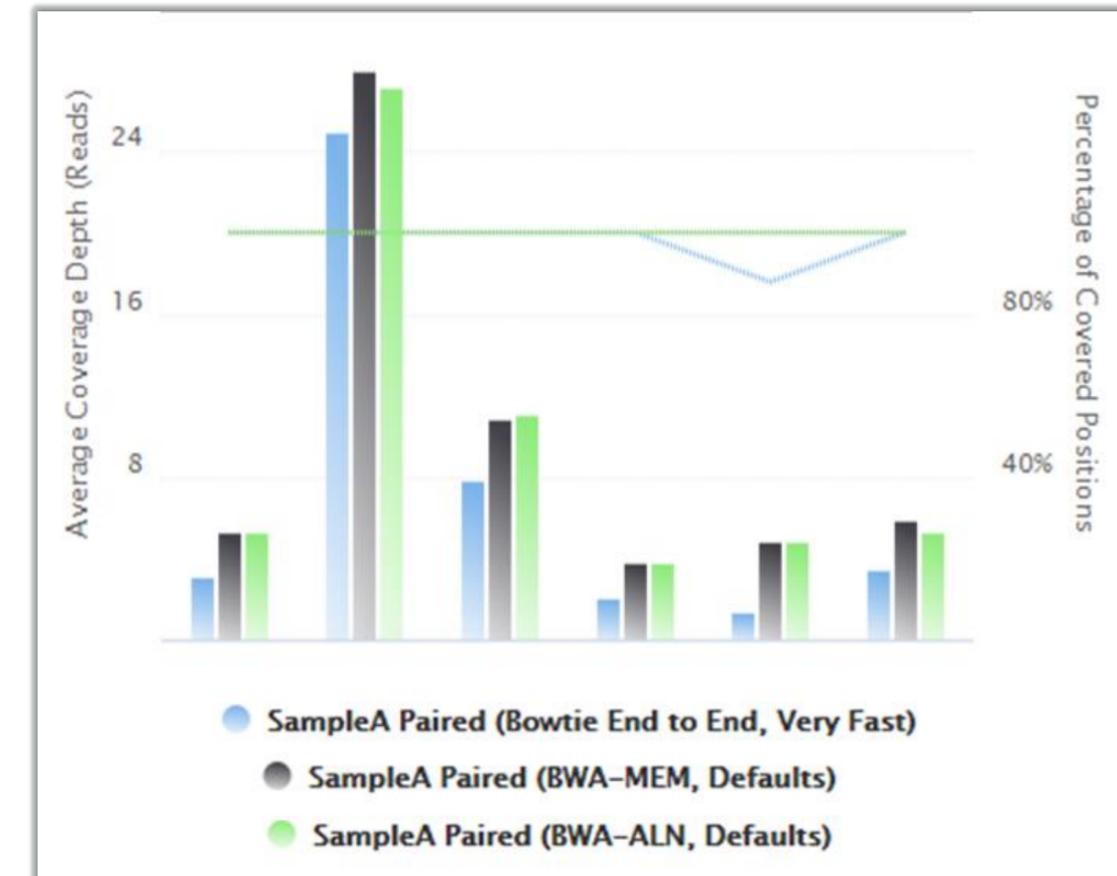
Exons of Single Gene

Note: L-lactate dehydrogenase gene
(TRAESCS1A02G238700) shown here

Alignment, Analysis, and Visualization of Wheat NGS Samples in Curio / Presented at PAG 2020

BWA-MEM Read Mapping Quality Impact on Coverage

Alignment, Analysis, and Visualization of Wheat NGS Samples in Curio / Presented at PAG 2020
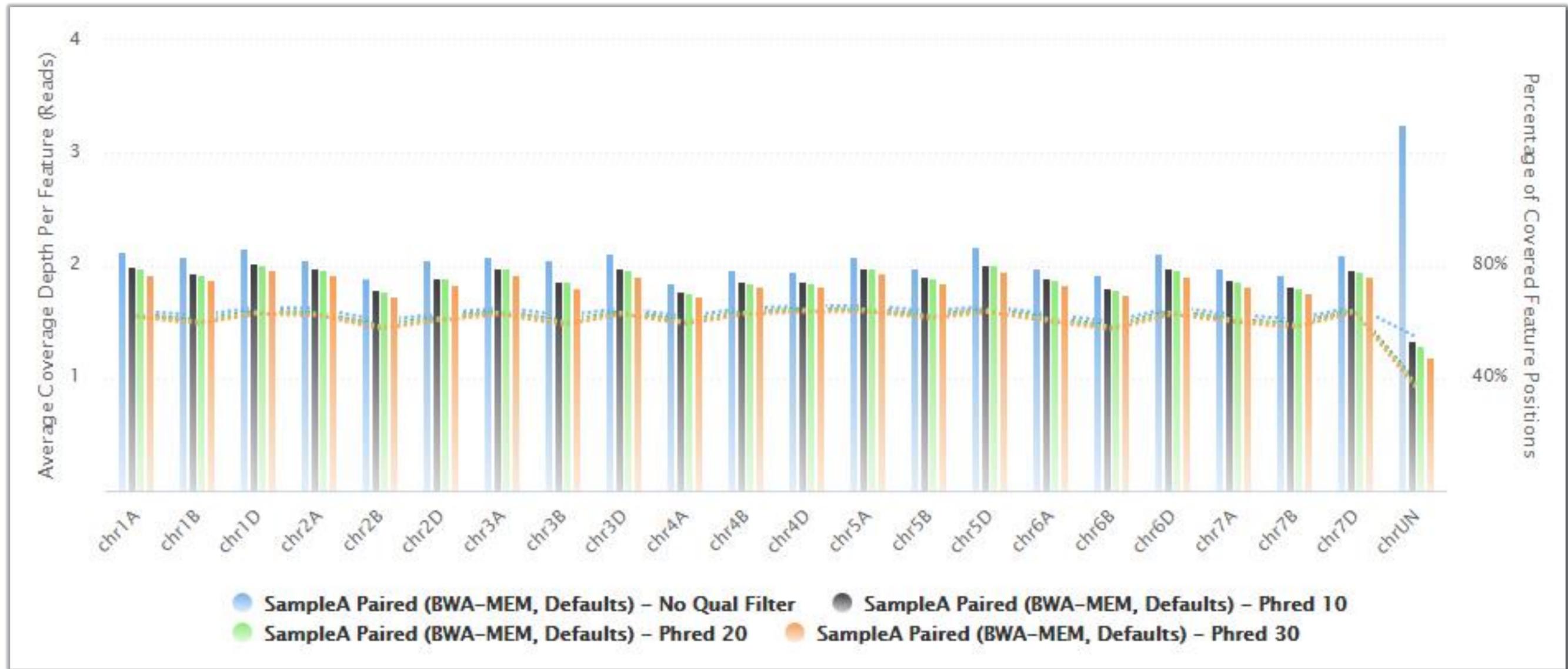
Bowtie2 Read Mapping Quality Impact on Coverage

CURIO
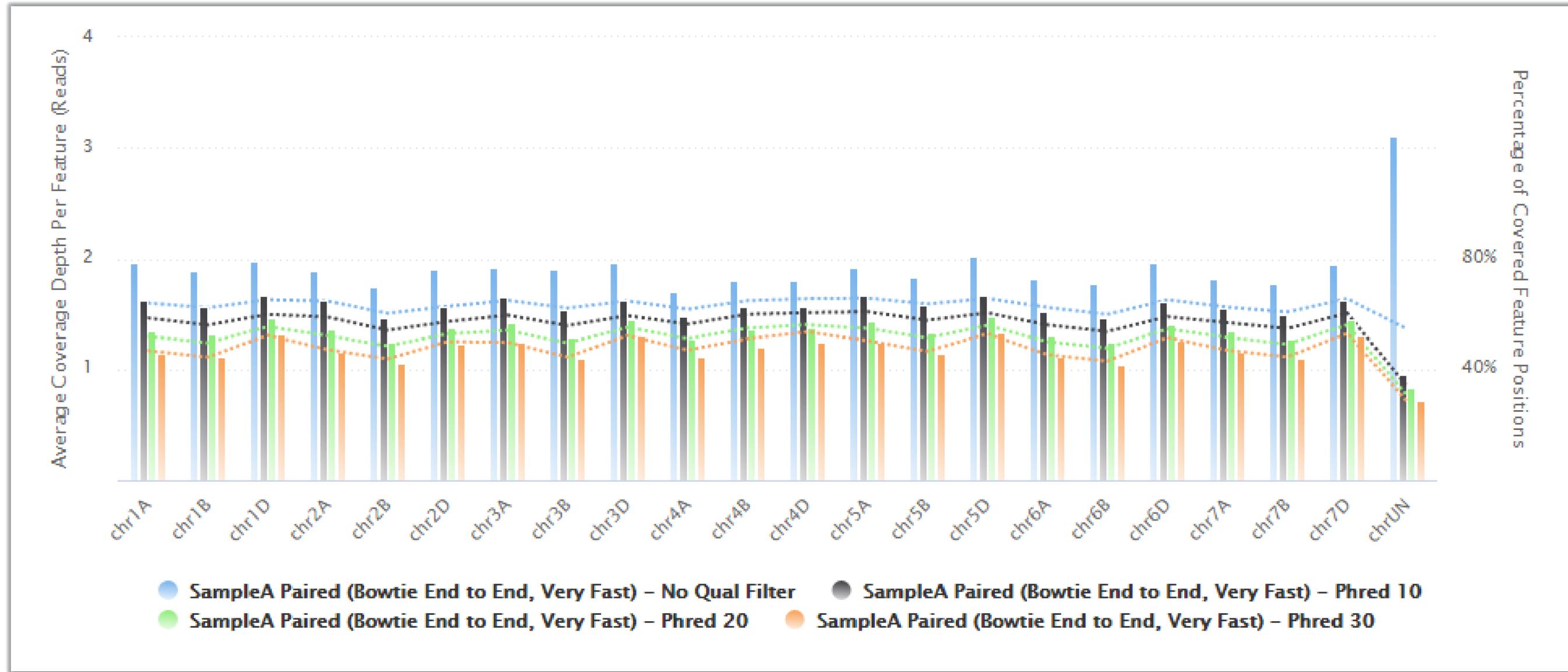
- BWA-MEM and Bowtie2 with no quality filter applied
- Visualizing low quality base calls within aligned reads

BWA-MEM / No Quality Filter
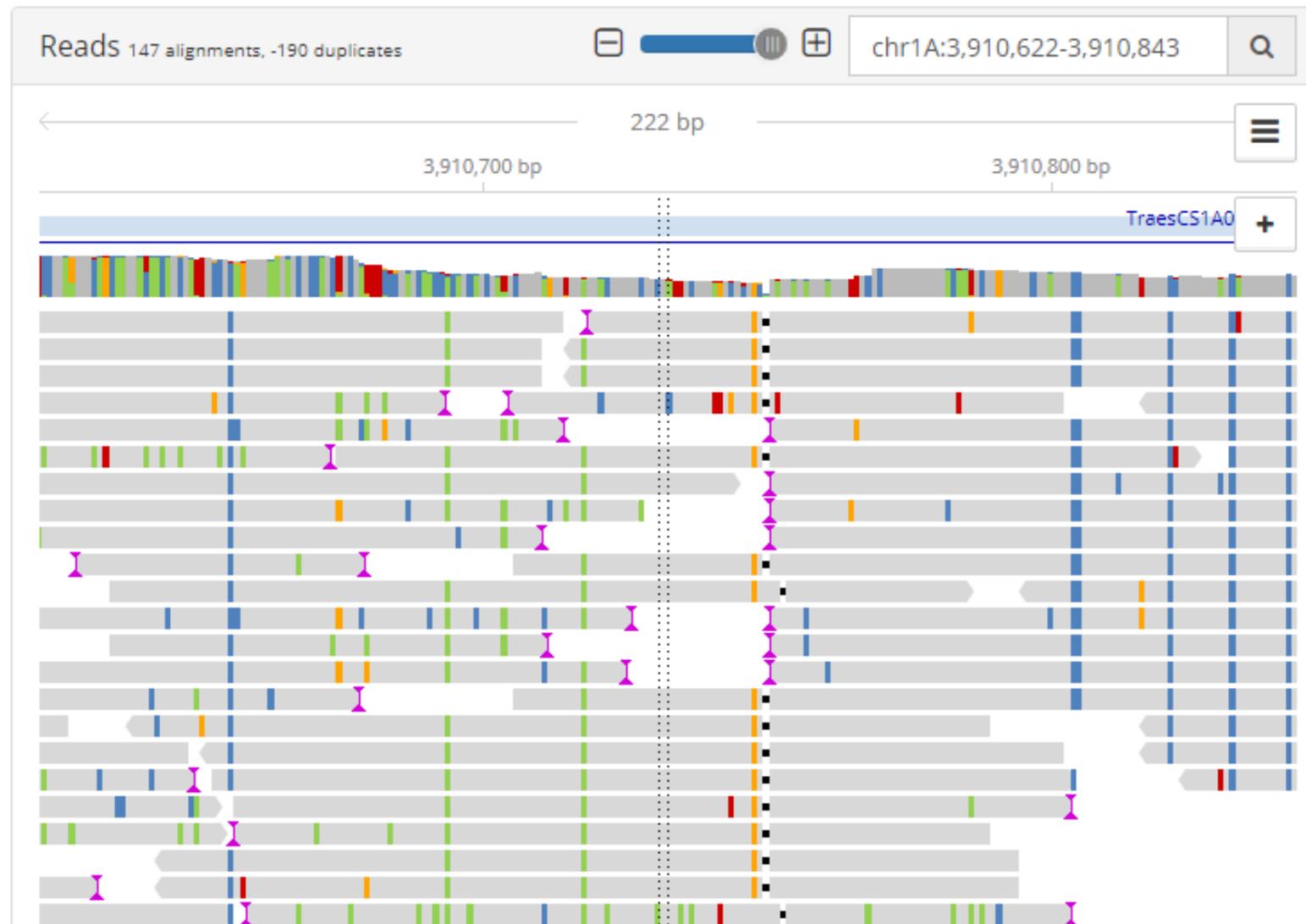
Bowtie2 / No Quality Filter

- BWA-MEM and Bowtie2 with a quality filter set to 99% (i.e. Phred 30)
- Visualizing impacts of different mapping algorithm approaches to alignment quality

BWA-MEM / Phred 30 or Higher
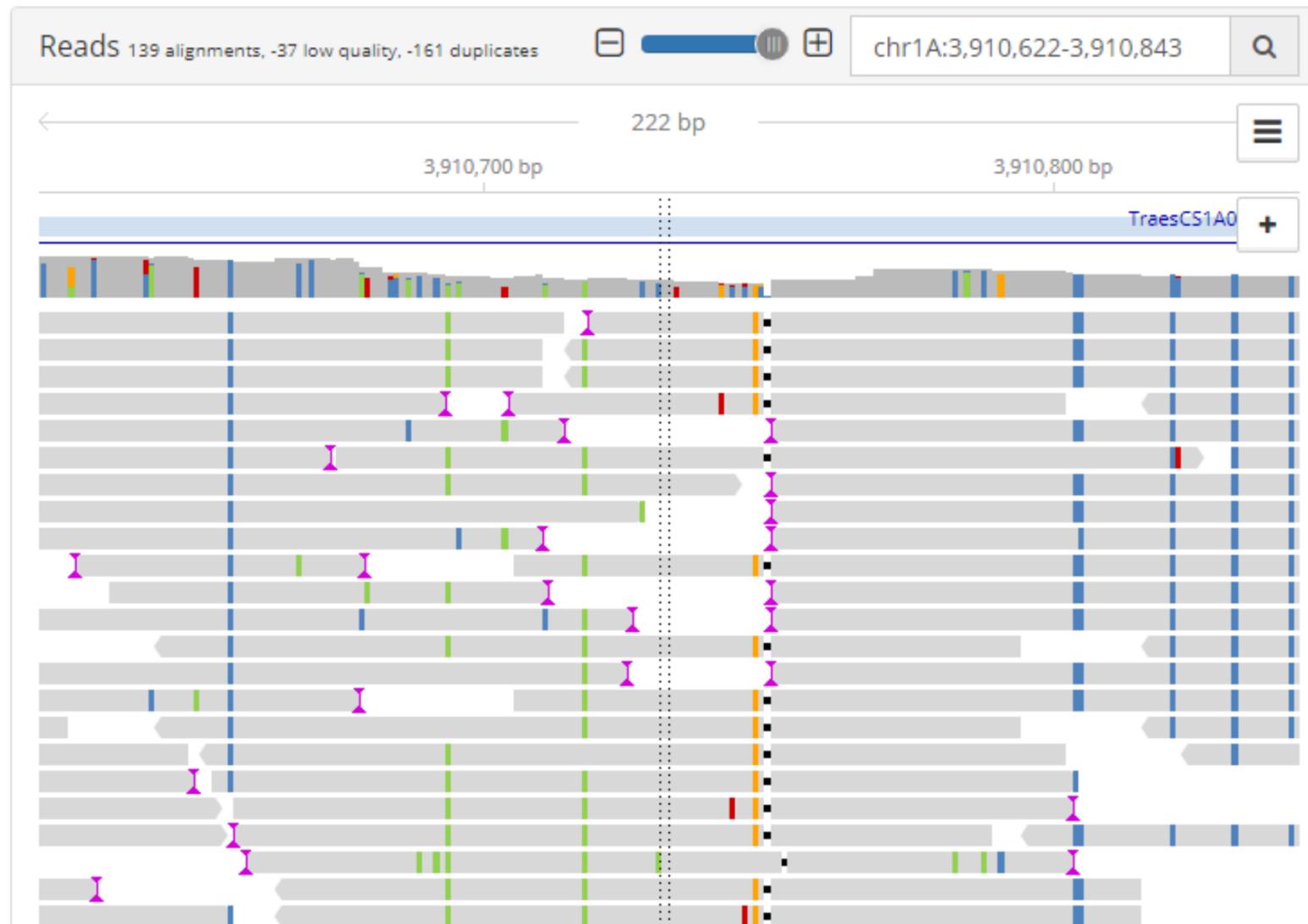
Bowtie2 / Phred 30 or Higher
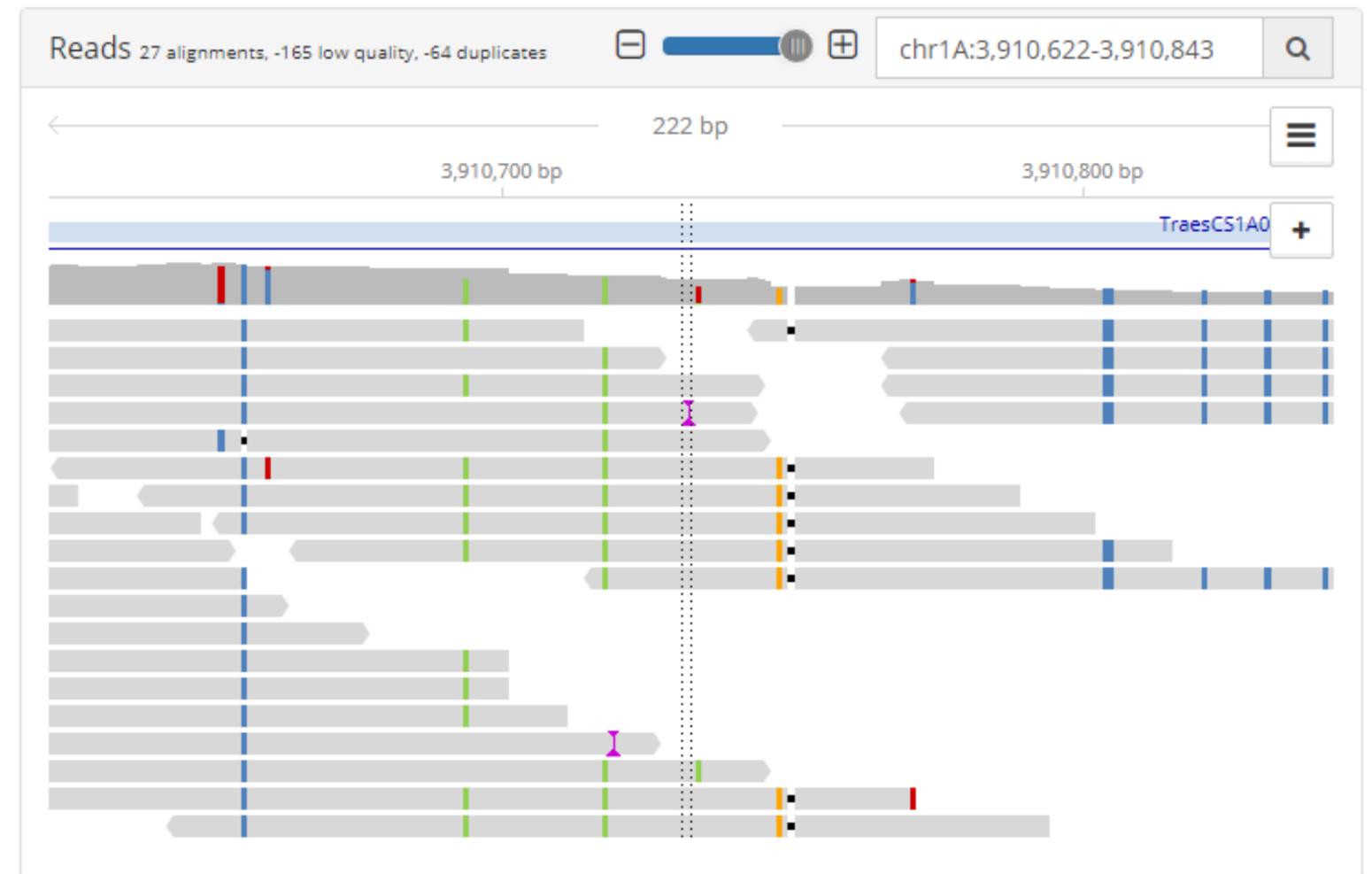
DNA-Seq Bread Wheat:
Variant Analysis

- Call variants leveraging the Chinese Spring Wheat reference assembly and related annotations from the IWGSC

- Operate on one or multiple samples of any read depth simultaneously

- Conveniently adjust for sensitivity and specificity

CURIO

- Browse and visualize variants across the entire CSW genome

Variant Analysis CultivarE-WT - BWA-Aln (Analyzed 2019-12-20 19:53 UTC) - High Sensitivity / HC Only

More Options ▾

Density 393,609 variants found
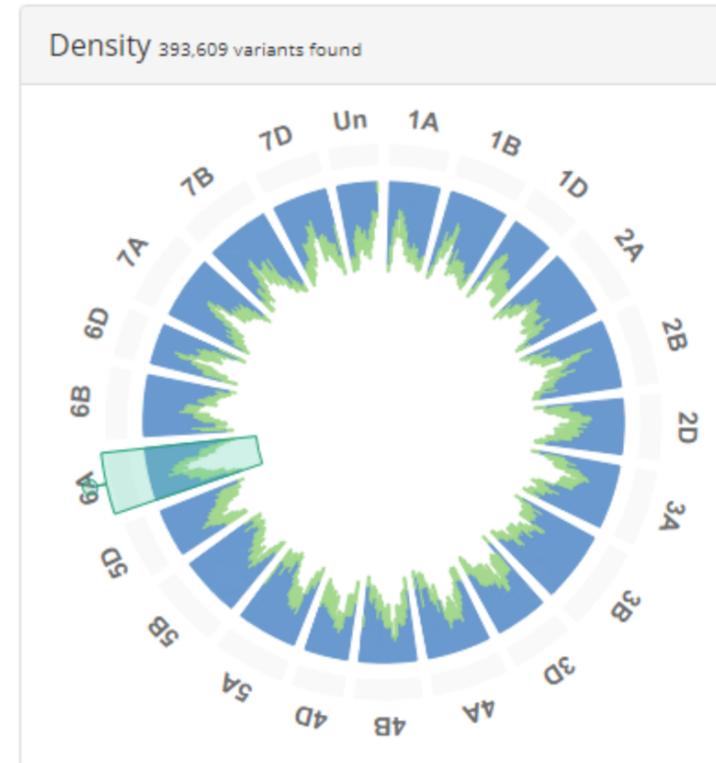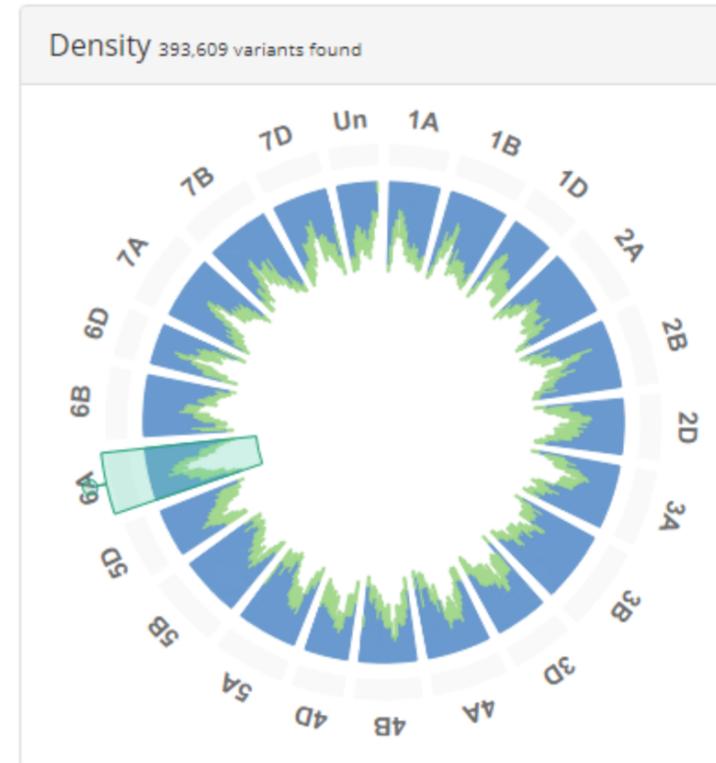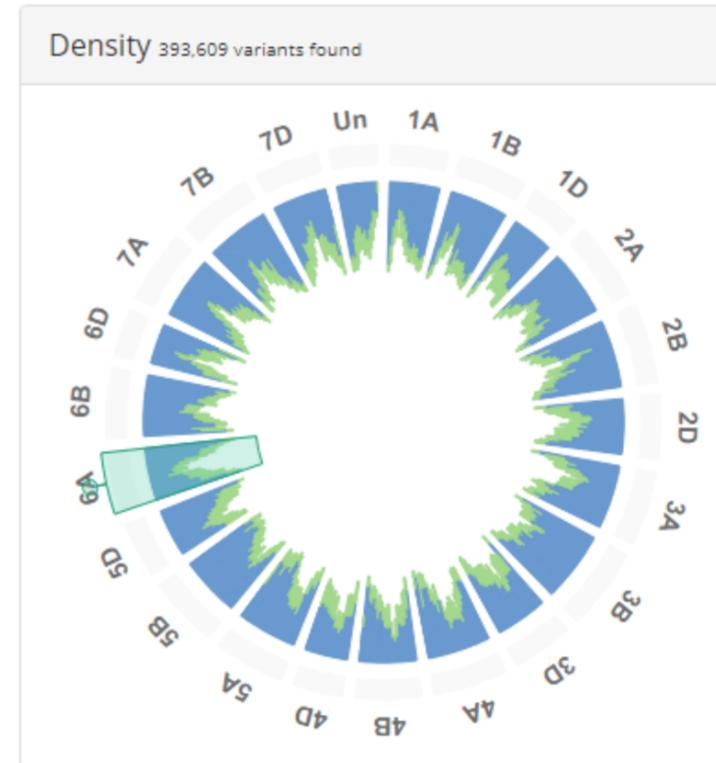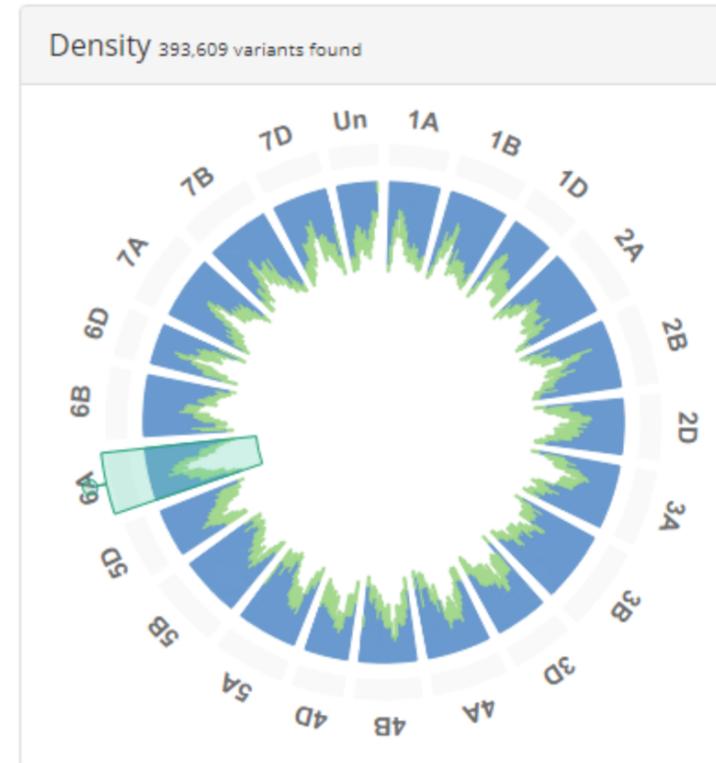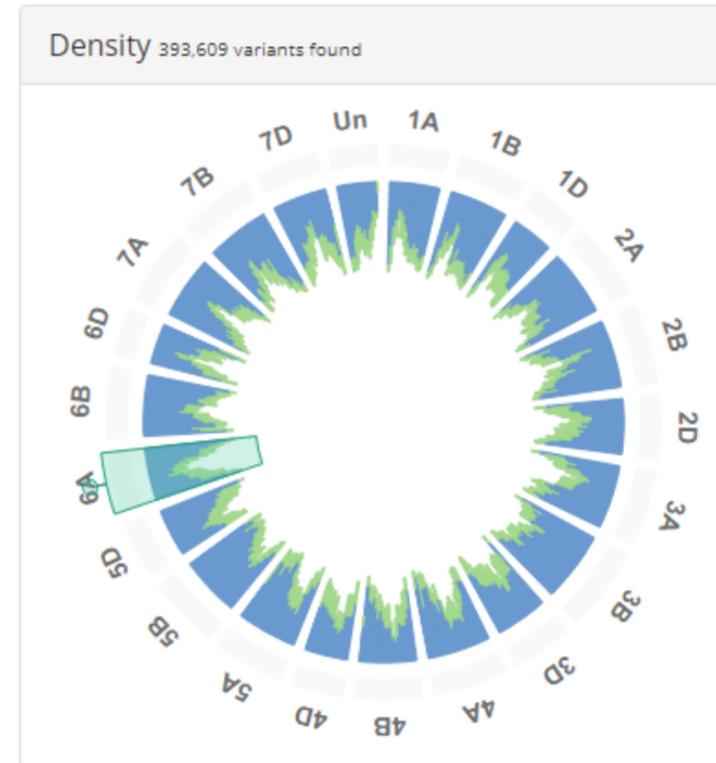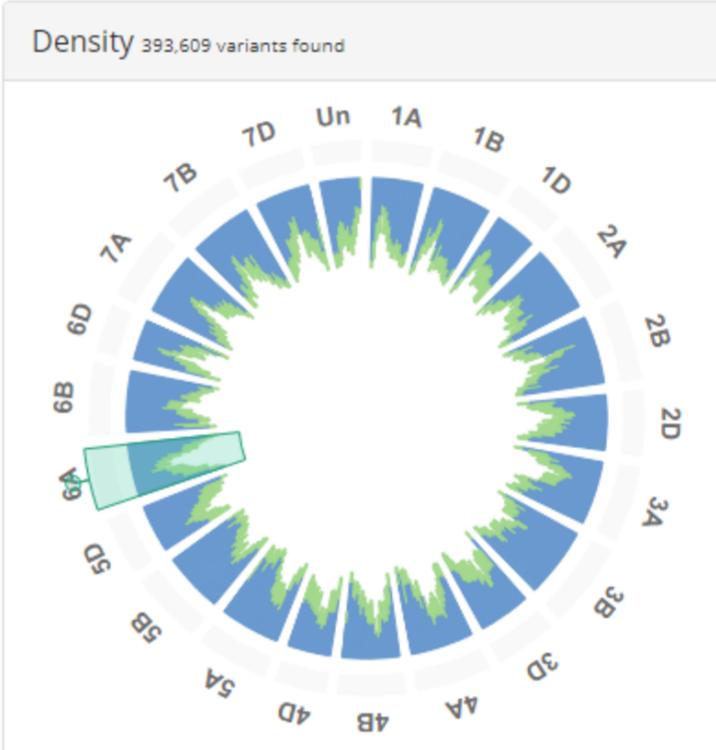
Display/Filter Options | Compare | Notes and Marks | Sharing

☑ Use gene transcript and exon annotations from IWGSC (HC)

☐ Include only variants within known coding regions ❓

☐ Include only variants with predicted consequence of missense

☑ Exclude variants whose alt coverage is not at least 5 reads

☑ Include SNVs for homozygous or heterozygous positions

☐ Include positions with rare alleles

☑ Color positions based on the variant's alternate allele ❓

☑ Show at each detected position the variant's coverage

☐ Show only variants not present in the control (i.e. somatic)

☑ Show only variants with control coverage of at least 5 reads

☑ Match homozygous SNVs to related heterozygous SNVs in control ❓

◉ Show all variants found within selected sample files

○ Show only variants common to all selected sample files

☐ Show file coverage track of variants detected per position

Variants (Found 18,248 matching variants in range chr6A:1-618,079,261)

⊖ ●━━━━━ ⊕ | chr6A:1-618,079,261 | 🔍

618,079,260 bp

Primary Analysis File
Variant Count: 18,248

- Browse and visualize variants across the entire CSW genome

- Simultaneously compare and visualize multiple samples

- Browse and visualize variants across the entire CSW genome

- Simultaneously compare and visualize multiple samples

- Use one or more samples as a control to exclude variants present in the wild type

- Browse and visualize variants across the entire CSW genome

- Simultaneously compare and visualize multiple samples

- Use one or more samples as a control to exclude variants present in the wild type

- Focus on genomic regions of interest based on IWGSC gene transcript annotations

- Browse and visualize variants across the entire CSW genome

- Simultaneously compare and visualize multiple samples

- Use one or more samples as a control to exclude variants present in the wild type

- Focus on genomic regions of interest based on IWGSC gene transcript annotations

- Access predicted biological consequences based on calculated amino acid shifts

- Browse and visualize variants across the entire CSW genome

- Simultaneously compare and visualize multiple samples

- Use one or more samples as a control to exclude variants present in the wild type

- Focus on genomic regions of interest based on IWGSC gene transcript annotations

- Access predicted biological consequences based on calculated amino acid shifts

- Interactive visualizations provide deeper access to the information behind each call

Bread Wheat RNA-Seq:
Read Mapping and Visualization

CURIO

- Conveniently select to align to either the full Chinese Spring Wheat genome or a transcriptome

CURIO

- Conveniently select to align to either the full Chinese Spring Wheat genome or a transcriptome

- Transcriptome is dynamically assembled on demand using selected annotation set from the IWGSC

CURIO

- Conveniently select to align to either the full Chinese Spring Wheat genome or a transcriptome

- Transcriptome is dynamically assembled on demand using selected annotation set from the IWGSC

- Quickly visualize and navigate RNA-Seq data from samples of any size

Alignment Density

Display Options | Notes and Marks | Sharing

- ☑ Show gene and exon positions from IWGSC (HC)
- ☑ Show reference genome bases
- ☑ Show nucleotide coverage and highlight when differences exceed 20%
- ☑ Exclude reads and bases with quality under 99%
- ☐ Group reads by UMI (perfect matches), requiring consensus of 60%
- ☐ Remove read duplicates at the same alignment position and CIGAR

- ⦿ Show individual read details
- ○ Show read consensus sequence information

- ☑ Highlight mismatched bases
- ☑ Highlight read insertions
- ☐ Color forward and reverse reads differently

Reads (Showing 466 alignments of 1,736, excluded 1,040 of low quality)

chr2B:158,616,522-158,619,4

2,889 bp

00 bp | 158,617,000 bp | 158,617,500 bp | 158,618,000 bp | 158,618,500 bp | 158,619,000 bp

TraesCS2B02G183300

CURIO

- Conveniently select to align to either the full Chinese Spring Wheat genome or a transcriptome

- Transcriptome is dynamically assembled on demand using selected annotation set from the IWGSC

- Quickly visualize and navigate RNA-Seq data from samples of any size

- Reference IWGSC transcript and functional annotations on the fly
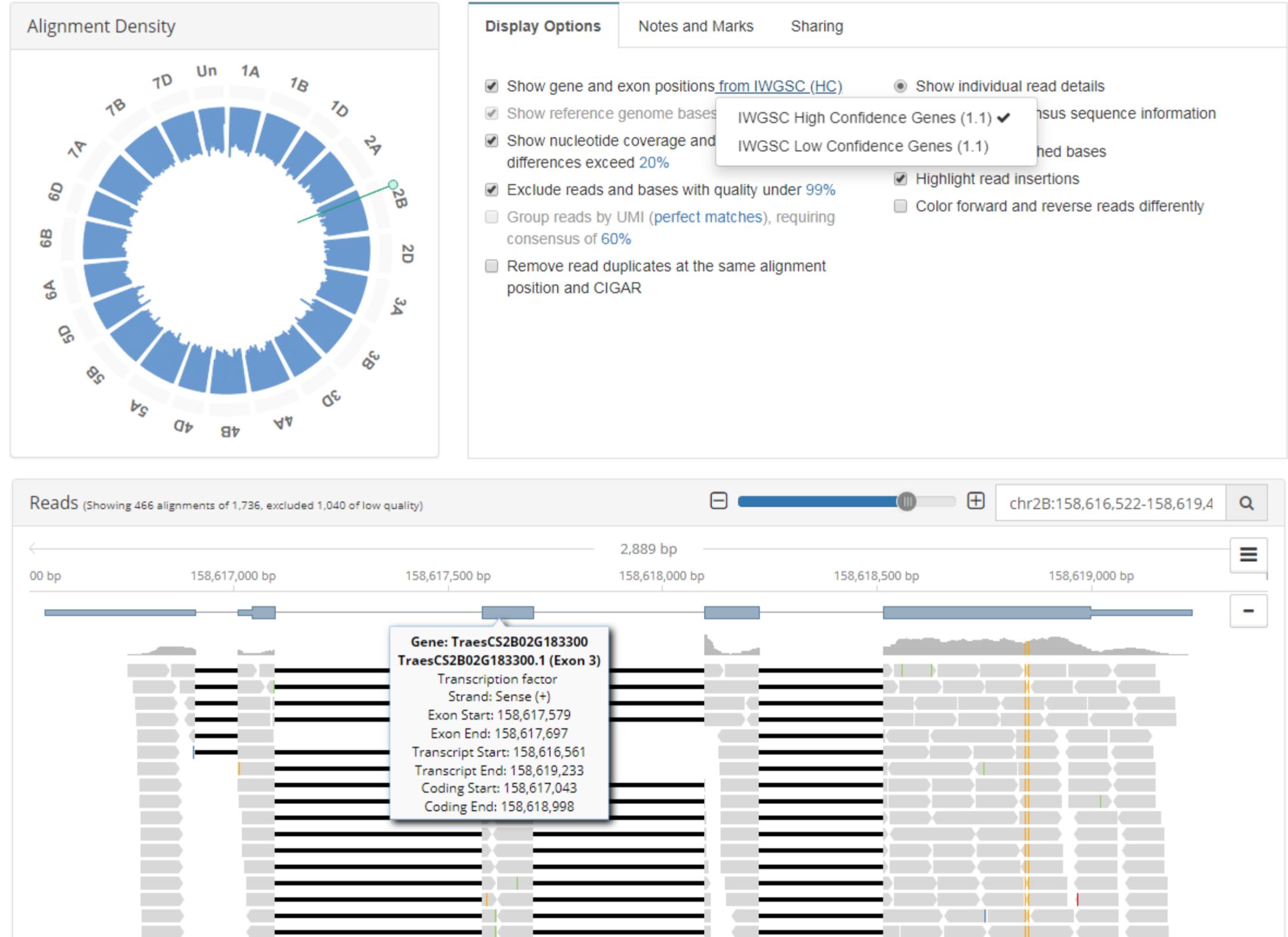
Bread Wheat RNA-Seq: Expression Analysis

CURIO

- Analyze feature expression utilizing the Chinese Spring Wheat reference assembly

Dashboard / All Projects / Project Details

Project Details

More Options ▼

General Information

Project Name:

Species:

General Notes:

Start Multiple Expression Analyses ✕

Genome Feature Set: ❓ | IWGSC - High Confidence Genes - Chinese Spring Wheat (v1.1) ▼

Feature to Count: ❓ | Gene ▼

Exon Matching: ❓ | ⦿ Count only reads that overlap with exons of the gene
| ◯ Count reads that overlap introns or exons of the gene

Multi Feature Overlap: ❓ | ☑ Allow reads to count towards all features they overlap with

De-duplication: ❓ | ☐ Remove potential read duplicates at the same position and CIGAR

Minimum Quality: ❓ | [slider] Phred: 20 (99%)

Cancel   ▶ Start Expression Analysis

All Files   Raw Sequences

⚙ Available Actions ▼   🗑 Delete Files

☑	File Name	Reads Processed	▼ Last Updated	
☑	SRR1542407.1 Paired (Aligned 2019-12-13 21:40 UTC)	193,694,736	2019-12-16 17:27	📊 View Alignments ▼
☑	SRR1542404.1 Paired (Aligned 2019-12-13 21:40 UTC)	198,975,170	2019-12-16 13:05	📊 View Alignments ▼
☑	SRR1542405.1 Paired (Aligned 2019-12-13 21:40 UTC)	184,418,714	2019-12-16 11:02	📊 View Alignments ▼
☑	SRR1542406.1 Paired (Aligned 2019-12-13 21:40 UTC)	184,129,219	2019-12-16 07:01	📊 View Alignments ▼
☑	SRR1542408.1 Paired (Aligned 2019-12-13 21:40 UTC)	165,312,410	2019-12-16 02:07	📊 View Alignments ▼
☑	SRR1542409.1 Paired (Aligned 2019-12-12 19:53 UTC)	203,941,764	2019-12-14 08:25	📊 View Alignments ▼

CURIO

- Analyze feature expression utilizing the Chinese Spring Wheat reference assembly

- Leverage either the IWGSC "High Confidence" or "Low Confidence" annotation sets

CURIO

- Analyze feature expression utilizing the Chinese Spring Wheat reference assembly

- Leverage either the IWGSC "High Confidence" or "Low Confidence" annotation sets

- Measure expression levels of genes, transcripts, or individual exons

- Analyze feature expression utilizing the Chinese Spring Wheat reference assembly

- Leverage either the IWGSC "High Confidence" or "Low Confidence" annotation sets

- Measure expression levels of genes, transcripts, or individual exons

- Group samples for a differential expression analysis...

- Analyze feature expression utilizing the Chinese Spring Wheat reference assembly

- Leverage either the IWGSC "High Confidence" or "Low Confidence" annotation sets

- Measure expression levels of genes, transcripts, or individual exons

- Group samples for a differential expression analysis...

- ...and conveniently tune industry standard 'R' based algorithms

CURIO

- Visualize expression levels utilizing transcriptome-based alignments leveraging IWGSC annotation sets

Bread Wheat: Differential Expression

- Visualize expression levels utilizing transcriptome-based alignments leveraging IWGSC annotation sets

- Validate sample distances through principal component analysis

CURIO

- Visualize expression levels utilizing transcriptome-based alignments leveraging IWGSC annotation sets

- Validate sample distances through principal component analysis

- Highlight up or down regulated genes through interactive heat maps

CURIO

- Visualize expression levels utilizing transcriptome-based alignments leveraging IWGSC annotation sets

- Validate sample distances through principal component analysis

- Highlight up or down regulated genes through interactive heat maps

- Compare and contrast different industry standard algorithms in real-time

In Conclusion

Looking Ahead

- Incorporation of the recently released IWGSC 2.0 reference assembly

- Addition of the soon-to-be released IWGSC 2.0 transcript and functional annotations

- Consideration of other standard hexaploid and tetraploid wheat assemblies and related annotation sets

- Support for custom reference assemblies

- Additional crop-research-specific analysis types and interpretive visualizations
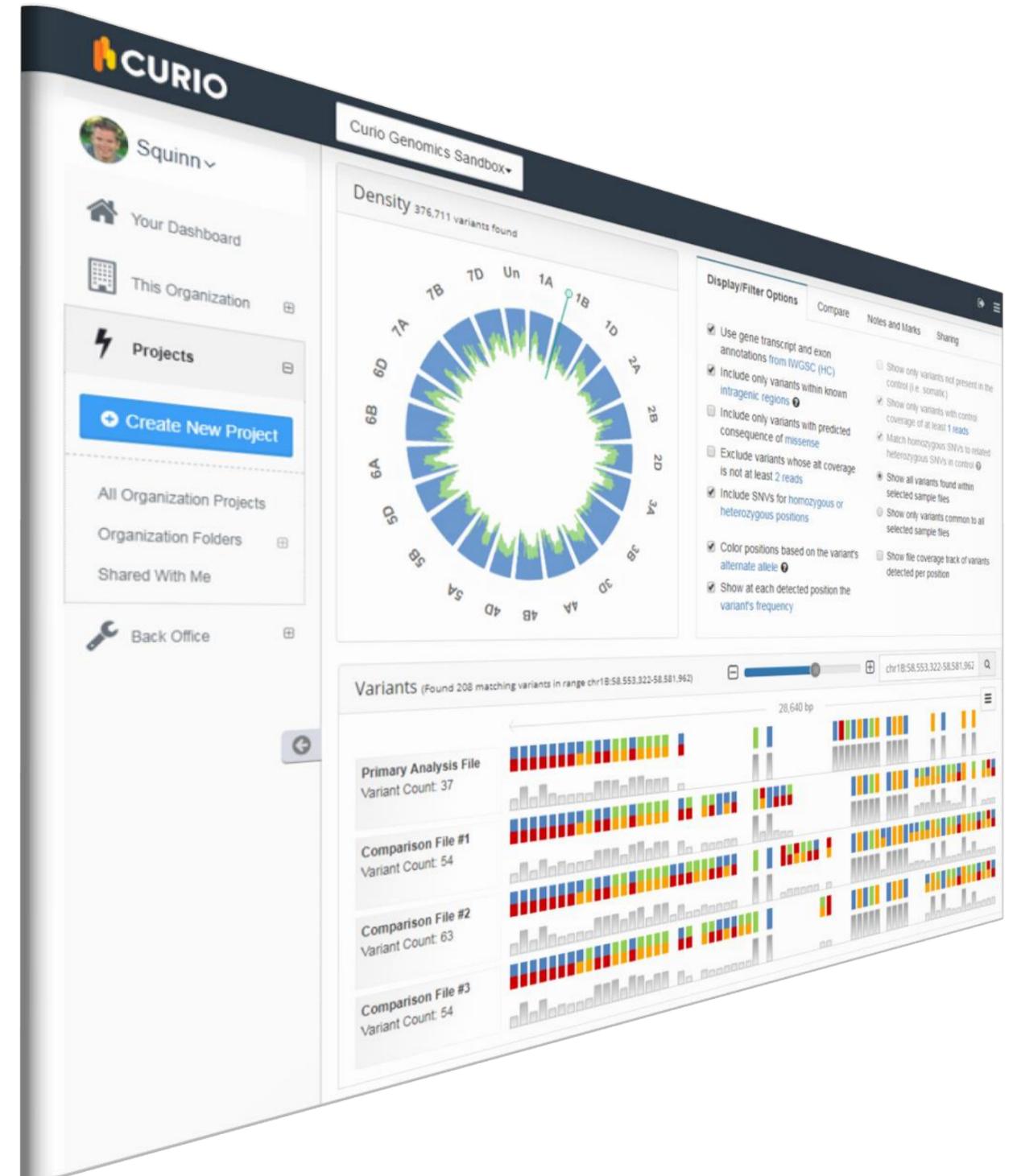
Acknowledgements

Special thanks to collaborators at:

- Arbor Biosciences

- John Innes Centre (JIC)

- French National Institute for Agricultural Research (INRA)

- University of Adelaide, Plant Genomics Centre

And the collective efforts of the:

- International Wheat Genome Sequencing Consortium (IWGSC)

Differential expression analysis utilized FASTQ files published as part of the following study:

- 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.

Q&A

Alignment, Analysis, and Visualization of Wheat NGS Samples in Curio / Presented at PAG 2020