AB INITIO GENOME ANNOTATION REPORT

Project ID : xxxxxxx

Species : *Chenopodium quinoa*

Customer : xxxxxxxxxxxxxxxxxx

Company : xxxxxxxxxxxxxxx

ASSESSMENT OF THE INPUT GENOME ASSEMBLY

Input assembly Name	Quinoa_v1.1.fasta
Assembly size (bp)	1,385,456,844
Input assembly # of scaffolds	3,486
Input assembly N50 (bp)	3,846,917

STATISTICS OF REPEATS MASKED IN THE GENOME

Total genome masked	63.46 %
Class I TEs repeats	44.69 %
Class II TEs repeats	6.23 %
Low complexity repeats	0.20 %
Simple repeats	1.81 %

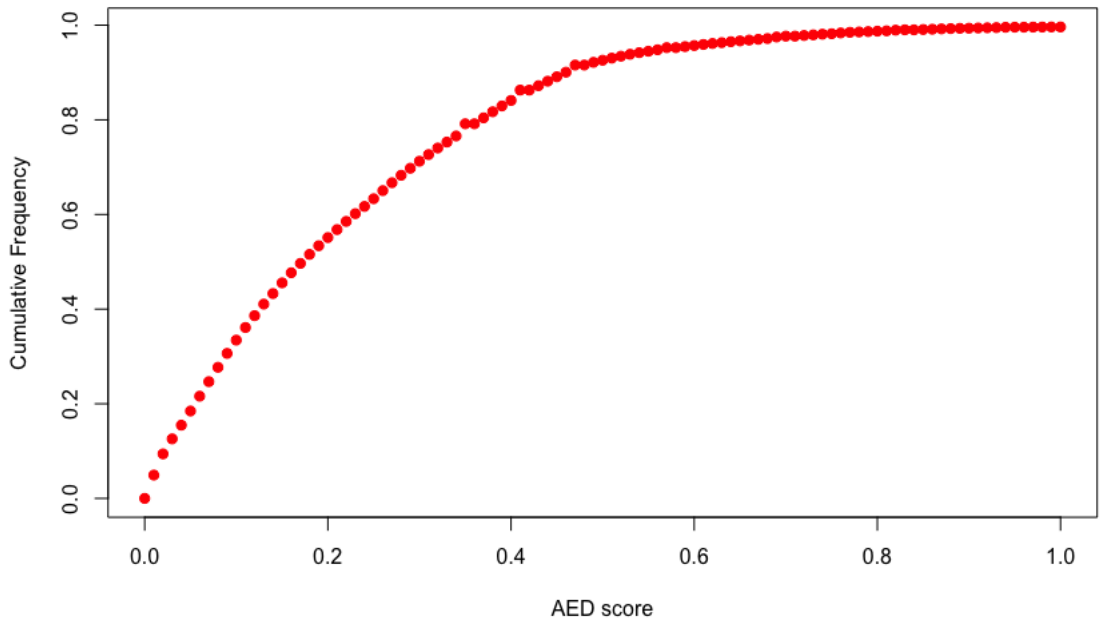
DE NOVO GENE PREDICTION METRICS

Total number of genes	44,776
Total coding region (bp)	57,064,233
Average length of genes (bp)	1,274
Largest gene (bp)	15,933
Number of single-exon genes	6,864

BUSCO ANALYSIS OF PREDICTED GENES

	Count	Percent
Complete single-copy BUSCOs	906	94.80 %
Complete duplicated BUSCOs	834	87.20%
Fragmented BUSCOs	24	2.50%
Missing BUSCOs	26	2.70%
Total BUSCO genes searched	956	100.00%

FREQUENCY GRAPH OF AED SCORES

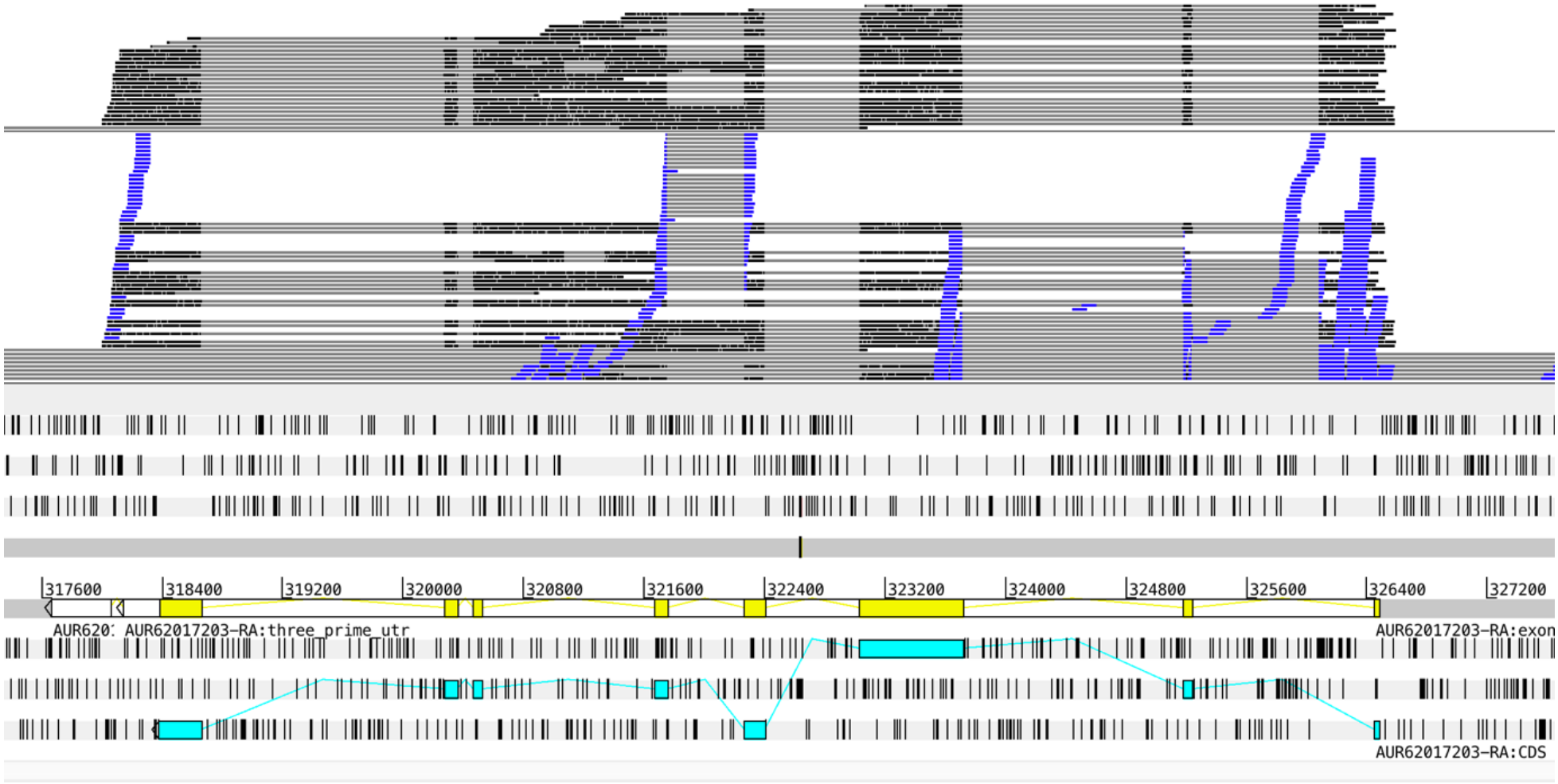


Annotation edit distance (AED) is a general measure of how well the predicted gene is supported by external evidence (UniProt protein and mRNA sequences). AED score ranges from 0 to 1 and a lower score represents more evidence support for the gene. AED is calculated for every gene. The AED cumulative frequency graph above provides an overview of the quality of the gene annotation.

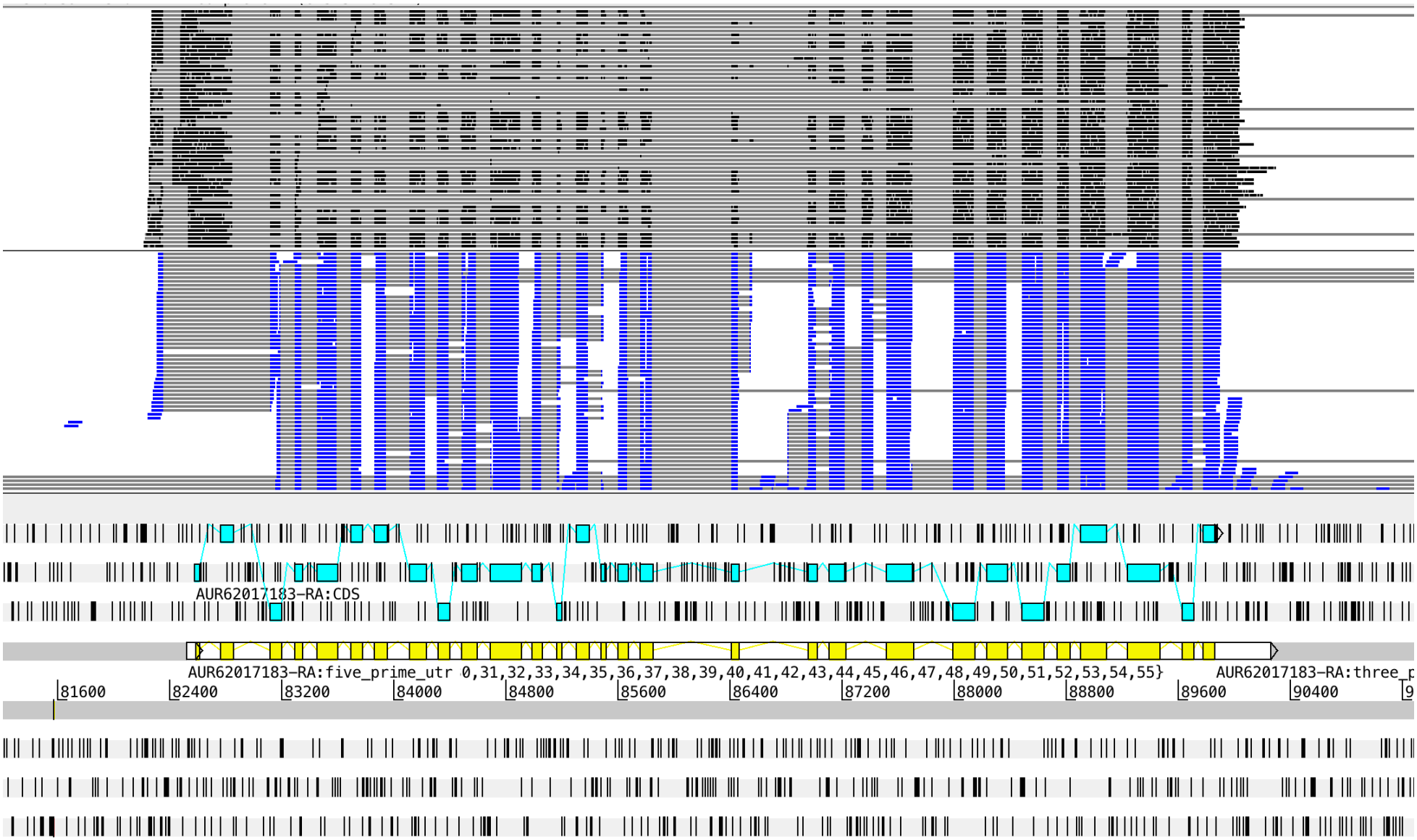
LIST OF GENES IDENTIFIED FOR MANUAL CURRATION

Customer genes of interest	Matched genes from annotation
Chromatin remodelling protein	EBS AUR62017203
Puromycin aminopeptidase	AUR62 017183
BHLH - Saponin pathway	AUR62017204
Beta Amyrin - Saponin pathway	AUR62 025693
Histone acetyltransferase	HAC12 AUR62001296

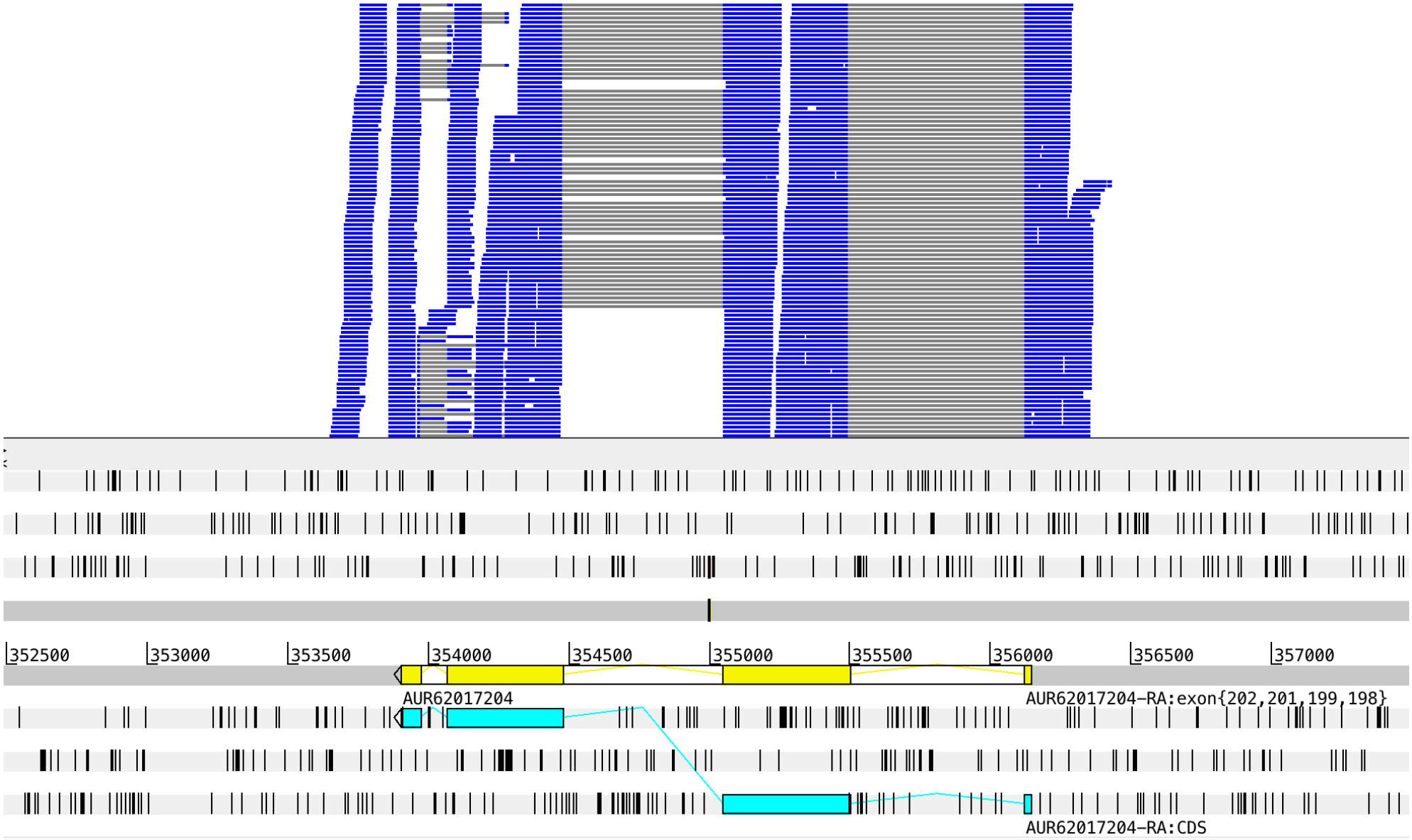
VISUALISATION OF GENE STRUCTURE FOR AUR62017203



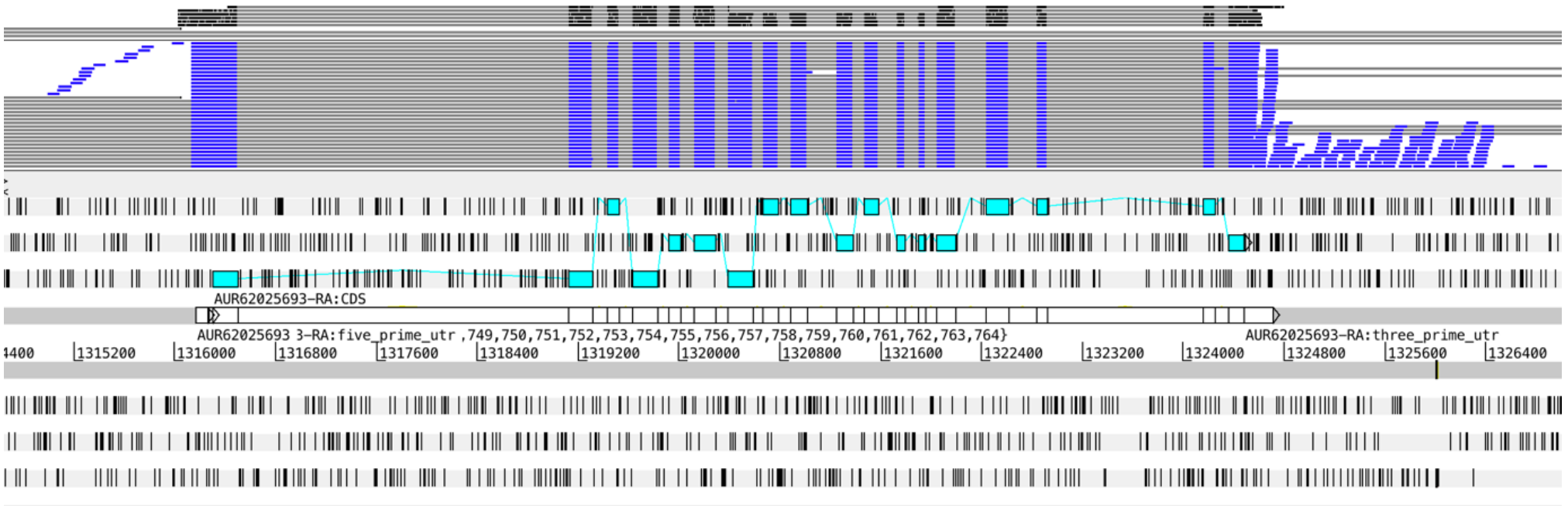
VISUALISATION OF GENE STRUCTURE FOR AUR62017183



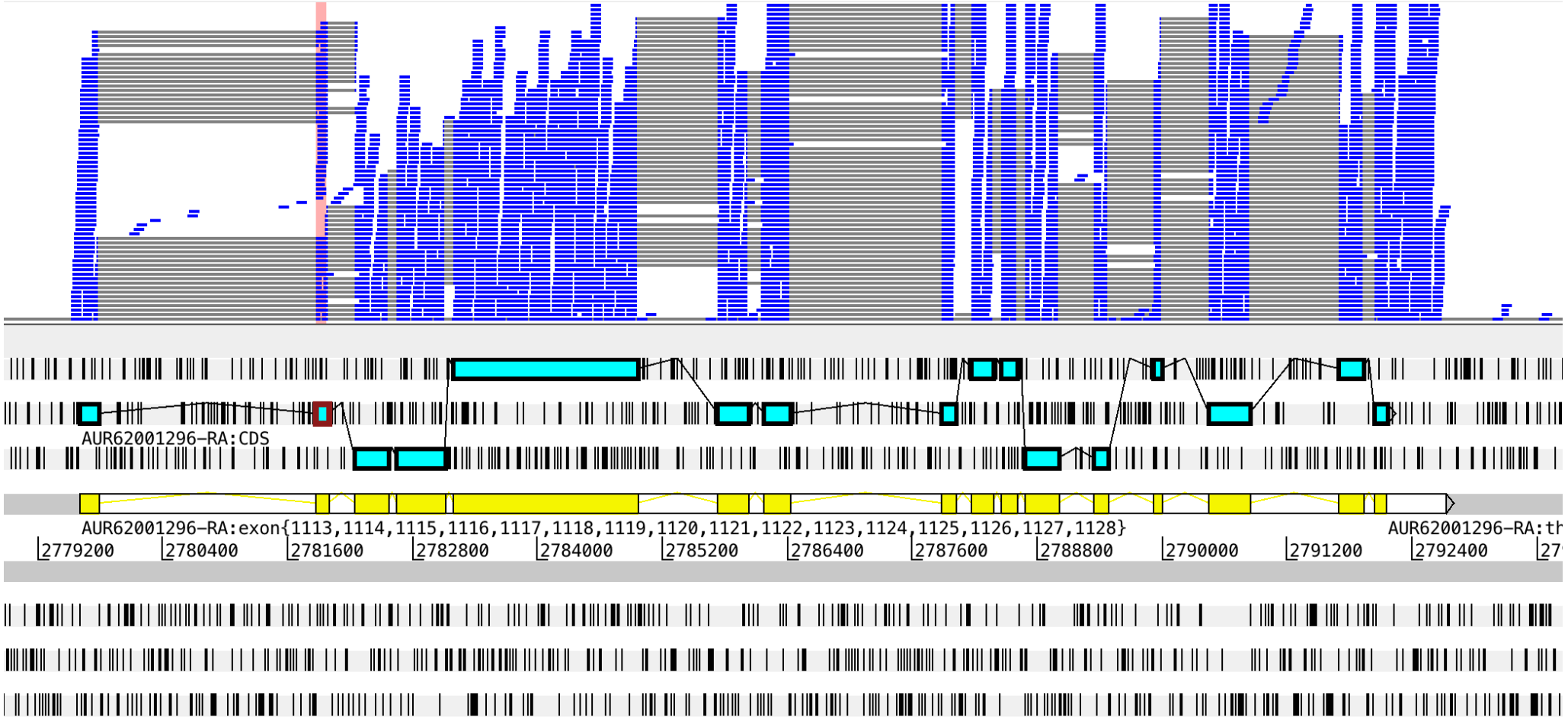
VISUALISATION OF GENE STRUCTURE FOR AUR62017204



VISUALISATION OF GENE STRUCTURE FOR AUR62025693



VISUALISATION OF GENE STRUCTURE FOR AUR62001296



LIST OF FILES AVAILABLE

1. Soft-masked genome assembly file
2. Genome annotation file in GFF3 format
3. Genes CDS sequences in fasta format
4. Genes peptide sequence in fasta format
5. Gene prediction matrix in text format
6. Bam alignment files from RNAseq and ISOseq data (if data is provided)