GIMAP IAP Haplotig Search Report

Erin Roberts

6/9/2020

Goal: Investigate presence of haplotigs in the genome annotation of *C. virginica* IAP and GIMAP gene families.

Methods

1. Compare sequence identity of all protein hits using CD-Hit at a level of 95% sequence identity
    - "global sequence identity" calculated as: number of identical amino acids or bases in alignment, divided by the full length of the shorter sequence
    - Code used in bluewaves: `module load CD-HIT/4.8.1-foss-2018b cd-hit -G 1 -c 1.0 -t 1 -i $F/BIR_IAP_HMMER_Interpro_XP_list_all.fa -o $F/BIR_IAP_HMMER_Interpro_XP_list_all_rm_dup.fa echo "done rm dup $(date)`
2. Merge CD-Hit protein sequence identity data with Jon Puritz data of mean coverage across all individuals.
3. Compare with Haplotig finder output (need to get software info from Jon) of genomic locations with haplotigs.
4. Identify protein clusters containing sequences from two different genes.
5. Compute average coverage within clusters, and then comparing between clusters.
6. Identify protein clusters with average coverage < 500 (about half of the coverage across most clusters).
7. Align nucleotide sequences from clusters and view alignments with CDS and haplotig information.

Load BED files from Jon

JP provided BED files for each of the two gene families, with mean coverage values averaged across all 90 individuals. Each gene has a coverage value (see below).

```
Cvir_GIMAP_meanCov <- read.table(file="/Users/erinroberts/Documents/PhD_Research/Chapter_1_Apoptosis Paper/Chapter_1_Apoptosis_Annotation_
                                 sep="\t", col.names = c("seqid","start","end","gene","meanCov"))
head(Cvir_GIMAP_meanCov)
```

```
##         seqid    start      end         gene   meanCov
## 1 NC_035783.1 40504479 40511614 LOC111129933  761.2769
## 2 NC_035783.1 41218181 41227834 LOC111130156 1112.3950
## 3 NC_035784.1 29147742 29149260 LOC111134644 1705.2812
## 4 NC_035784.1 89744459 89751167 LOC111132212  439.5429
## 5 NC_035785.1 30043993 30074271 LOC111102099  369.5091
## 6 NC_035786.1 13946434 14026414 LOC111103088  271.4175
```

```
Cvir_IAP_meanCov <- read.table(file="/Users/erinroberts/Documents/PhD_Research/Chapter_1_Apoptosis Paper/Chapter_1_Apoptosis_Annotation_Da
                               sep="\t", col.names = c("seqid","start","end","gene","meanCov"))
```

The data format for haplotigs file lists large regions in the genome where haplotigs were identified. All the counts are 0 (not sure what this means). Each identified haplotig encompasses many genes, not just a single gene.

```
Cvir_haplotigs <- read.table(file="/Users/erinroberts/Documents/PhD_Research/Chapter_1_Apoptosis Paper/Chapter_1_Apoptosis_Annotation_Data
                             sep="\t", skip= 1, col.names = c("seqid","start","end","counts","dataset"))
head(Cvir_haplotigs)
```

```
##         seqid    start      end counts  dataset
## 1 NC_035780.1 13598600 13674766      0 haplotig
## 2 NC_035780.1 13674865 13845210      0 haplotig
## 3 NC_035780.1 13845309 13849896      0 haplotig
## 4 NC_035780.1 13849995 13986879      0 haplotig
## 5 NC_035780.1 14404271 14736561      0 haplotig
## 6 NC_035780.1 14736660 14771157      0 haplotig
```

Investigate CD-hit results

Let's now review the results from CD-Hit and join the mean coverage information. CD-Hit software works by first clustering sequences by sequence similarity. Proteins in the cluster denoted by a * are the longest sequence in the cluster and are used as the reference sequence in the cluster. Sequences in the cluster denoted with a similarity percentage are that percentage identical to the sequence denoted with *.

Join the CD-Hit results for each family with mean coverage

In order to narrow down the CD-Hit clusters to view, we are only investigating the clusters where proteins were clustered across two different genes.

```
Cvir_GIMAP_meanCov_CD_Hit_95 <- left_join(AIG_seq_rm_dup_clstr6_dup_diff_gene_95_product_95 ,
                 Cvir_GIMAP_meanCov) %>% filter(Species =="Crassostrea_virginica")
Cvir_IAP_meanCov_CD_Hit_95 <- left_join(BIR_seq_rm_dup_clstr6_dup_diff_gene_product_95 ,
                  Cvir_IAP_meanCov) %>% filter(Species =="Crassostrea_virginica")
```

Join Gene Length with CD-Hit and Mean Coverage results

Ximing pointed out that haplotigs are likely 1 or more MB in length and are large sequences. Joining the nucleotide length of each to see if this is informative.

```
Cvir_GIMAP_meanCov_CD_Hit_95 <- left_join(Cvir_GIMAP_meanCov_CD_Hit_95, GIMAP_BED_name)
Cvir_IAP_meanCov_CD_Hit_95 <- left_join(Cvir_IAP_meanCov_CD_Hit_95, IAP_BED_name)

GIMAP_gene_length_aa <- AIG_seq_rm_dup_clstr6_NUC_95[,c("aa","gene")]
IAP_gene_length_aa   <- BIR_seq_rm_dup_clstr6_NUC_95[,c("aa","gene")]
colnames(GIMAP_gene_length_aa)[1] <- "gene_length"
colnames(IAP_gene_length_aa  )[1] <- "gene_length"

Cvir_GIMAP_meanCov_CD_Hit_95_length <- left_join(Cvir_GIMAP_meanCov_CD_Hit_95, GIMAP_gene_length_aa)
Cvir_IAP_meanCov_CD_Hit_95_length <- left_join(Cvir_IAP_meanCov_CD_Hit_95 , IAP_gene_length_aa)
Cvir_IAP_meanCov_CD_Hit_95_length <- unique(Cvir_IAP_meanCov_CD_Hit_95_length)

# Make unique for each gene
Cvir_GIMAP_meanCov_CD_Hit_95_length_unique <- Cvir_GIMAP_meanCov_CD_Hit_95_length %>% distinct(gene, .keep_all = TRUE)
Cvir_IAP_meanCov_CD_Hit_95_length_unique <-  Cvir_IAP_meanCov_CD_Hit_95_length %>% distinct(gene, .keep_all = TRUE)
```

View GIMAP combined CD-Hit, gene length, mean coverage results

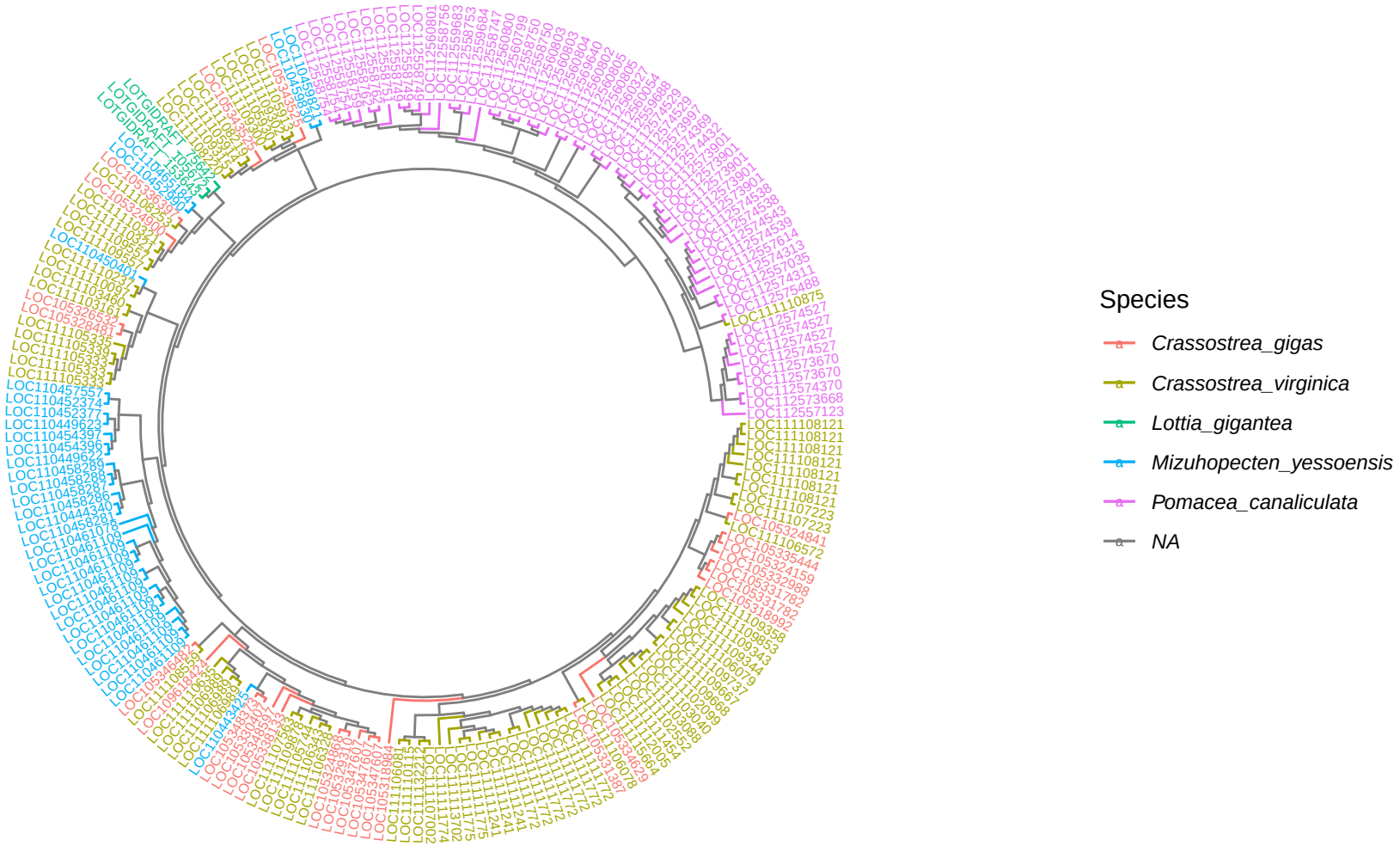
Lets first view the GIMAP gene family results.

cluster	prot_identity_stat	gene	product	meanCov	gene_length
Cluster 26	at 97.34%	LOC111108220	GTPase IMAP family member 7-like	945.1474	1961
Cluster 26	*	LOC111109315	uncharacterized protein LOC111109315	628.5355	5586
Cluster 36	*	LOC111109300	reticulocyte-binding protein 2 homolog a-like	560.4509	2300
Cluster 36	at 97.83%	LOC111105930	GTPase IMAP family member 4-like	784.2984	1528
Cluster 65	*	LOC111110237	uncharacterized protein LOC111110237	1005.7180	4805
Cluster 65	at 95.87%	LOC111110097	immune-associated nucleotide-binding protein 9-like	477.3209	3409
Cluster 95	*	LOC111106328	GTPase IMAP family member 4-like	1027.9801	5068
Cluster 95	at 98.55%	LOC111106343	GTPase IMAP family member 4-like	1756.0508	5096
Cluster 95	at 95.45%	LOC111105744	GTPase IMAP family member 4-like	1585.5985	6177
Cluster 219	*	LOC111110115	GTPase IMAP family member 4-like	557.6818	8257
Cluster 219	at 98.71%	LOC111106081	GTPase IMAP family member 4-like	316.2413	7929
Cluster 252	*	LOC111129933	uncharacterized protein LOC111129933 isoform X1	761.2769	7135
Cluster 252	at 99.56%	LOC111130156	uncharacterized protein LOC111130156 isoform X2	1112.3950	9653

Do any clusters with two genes with high protein simarilty have a large difference in coverage?

1. Cluster 65: genes LOC111110237, LOC111110097 have the largest difference in coverage within clusters
2. Cluster 219: gene LOC111110115, LOC111106081 genes have the lowest coverage compared to genes in other clusters

Review Gene Location on GIMAP tree



View IAP combined CD-Hit, gene length, mean coverage results

Lets view the IAP gene family results.

cluster	prot_identity_stat	gene	product	meanCov	gene_length
Cluster 5	at 99.78%	LOC111130310	baculoviral IAP repeat-containing protein 6-like isoform X3	710.95093	33169
Cluster 5	*	LOC111129365	baculoviral IAP repeat-containing protein 6-like isoform X1	658.66650	31922
Cluster 17	at 97.75%	LOC111100443	baculoviral IAP repeat-containing protein 2-like isoform X1	866.17023	5311
Cluster 17	*	LOC111100400	uncharacterized protein LOC111100400 isoform X1	1581.80383	15332
Cluster 30	at 96.32%	LOC111100408	baculoviral IAP repeat-containing protein 2-like	572.59491	29183
Cluster 30	*	LOC111100402	uncharacterized protein LOC111100402	679.60083	24362
Cluster 62	*	LOC111100470	baculoviral IAP repeat-containing protein 3-like	389.47702	13236
Cluster 62	at 97.09%	LOC111101689	baculoviral IAP repeat-containing protein 2-like	499.57413	16106
Cluster 118	*	LOC111100416	baculoviral IAP repeat-containing protein 7-like	823.04767	9310
Cluster 118	at 97.30%	LOC111100417	baculoviral IAP repeat-containing protein 7-like	921.41998	8689
Cluster 137	*	LOC111101864	baculoviral IAP repeat-containing protein 3-like	672.46875	6671
Cluster 137	at 95.75%	LOC111101018	baculoviral IAP repeat-containing protein 3-like	930.51135	6813
Cluster 153	*	LOC111104279	baculoviral IAP repeat-containing protein 7-A-like	775.09753	2369
Cluster 153	at 96.97%	LOC111105148	baculoviral IAP repeat-containing protein 7-B-like	558.19763	2606
Cluster 180	*	LOC111102451	putative inhibitor of apoptosis	543.89752	5094
Cluster 180	at 95.73%	LOC111105503	baculoviral IAP repeat-containing protein 2-like	714.34821	6238
Cluster 222	at 98.48%	LOC111122858	uncharacterized protein LOC111122858	593.17334	2700
Cluster 222	*	LOC111122723	uncharacterized protein LOC111122723	735.38690	2277
Cluster 265	*	LOC111103427	baculoviral IAP repeat-containing protein 3-like	1988.57214	9885
Cluster 265	at 95.02%	LOC111103428	baculoviral IAP repeat-containing protein 2-like	1585.04919	6520
Cluster 266	*	LOC111104637	putative inhibitor of apoptosis isoform X1	683.97888	6582
Cluster 266	at 98.75%	LOC111103790	putative inhibitor of apoptosis	995.14258	4811
Cluster 280	*	LOC111103392	baculoviral IAP repeat-containing protein 3-like isoform X1	492.13504	11552
Cluster 280	at 95.65%	LOC111103826	baculoviral IAP repeat-containing protein 2-like isoform X1	872.68256	14534
Cluster 282	*	LOC111104229	baculoviral IAP repeat-containing protein 8-like	513.64270	2292
Cluster 282	at 97.13%	LOC111103158	baculoviral IAP repeat-containing protein 8-like	1046.96606	2358
Cluster 315	*	LOC111105597	E3 ubiquitin-protein ligase XIAP-like	404.49841	4400
Cluster 315	at 100.00%	LOC111099688	E3 ubiquitin-protein ligase XIAP-like	616.43054	4397
Cluster 317	*	LOC111100396	E3 ubiquitin-protein ligase XIAP-like	576.50696	9884
Cluster 317	at 96.79%	LOC111102530	E3 ubiquitin-protein ligase XIAP-like	640.16211	9345
Cluster 328	*	LOC111132301	putative inhibitor of apoptosis	383.89096	587
Cluster 328	at 98.97%	LOC111114013	putative inhibitor of apoptosis	103.34412	587
Cluster 328	at 98.46%	LOC111103682	baculoviral IAP repeat-containing protein 3-like	103.06474	587
Cluster 328	at 98.46%	LOC111132489	putative inhibitor of apoptosis	890.39185	587

```
##  Cluster 328          at 99.49% LOC111132589                          putative inhibitor of apoptosis  223.18568         587
##  Cluster 328          at 99.49% LOC111102106                          putative inhibitor of apoptosis  148.71721         587
##  Cluster 328          at 99.49% LOC111114070                          putative inhibitor of apoptosis  113.52641         587
##  Cluster 338          at 98.61% LOC111116378 death-associated inhibitor of apoptosis 1-like  189.15207         434
##  Cluster 338                 * LOC111109152 death-associated inhibitor of apoptosis 1-like 1504.06763         503
##  Cluster 338          at 96.53% LOC111117137 death-associated inhibitor of apoptosis 1-like 1776.50464         434
##  Cluster 344                 * LOC111117856 death-associated inhibitor of apoptosis 1-like  996.02759         834
##  Cluster 344          at 95.65% LOC111116826 death-associated inhibitor of apoptosis 1-like  373.41403         855
##  Cluster 344          at 95.65% LOC111111659 death-associated inhibitor of apoptosis 1-like   84.94779         747
```

Do any clusters with two genes with high protein simarilty have a large difference in coverage?

- Cluster 17: LOC111100400 or LOC111100443 Cluster 17 has two genes with large differences in coverage. One has about half coverage and LOC111100400 is very long
- Cluster 280: LOC111103392 has about half coverage compared to other gene
- Cluster 282: LOC111104229 has half coverage as compared to the other gene in the cluster and very high sequence similarity
- cluster 328: Coverage is variable throughout cluster.
- cluster 338: Two have high coverage around 1500 and LOC111116378 has 189 coverage
- Cluster 344:LOC111111659 and LOC111116826 both have very low coverage compared to LOC111117856

RAxML tree is still run on the cluster so we can't review tree location.

Compare with Haplomerger results

Lets now investigate if these interesting gene clusters overlap with regions identified as haplotigs using JP software. Because the haplotig tool identified large regions, I am going to search for whether my cluster genes of interest within these regions. If both the gene start and gene end of my genes are inside any haplotig tool identified range, both of these columns are denoted with a "YES".

```
Cvir_GIMAP_meanCov_CD_Hit_95_length_unique$HM_found_start <- ifelse(sapply(Cvir_GIMAP_meanCov_CD_Hit_95_length_unique$start, function(p)
  any(Cvir_haplotigs$start <= p & Cvir_haplotigs$end >= p)),"YES", NA)
Cvir_GIMAP_meanCov_CD_Hit_95_length_unique$HM_found_end <- ifelse(sapply(Cvir_GIMAP_meanCov_CD_Hit_95_length_unique$end, function(p)
  any(Cvir_haplotigs$start <= p & Cvir_haplotigs$end >= p)),"YES", NA)

Cvir_IAP_meanCov_CD_Hit_95_length_unique$HM_found_start <- ifelse(sapply(Cvir_IAP_meanCov_CD_Hit_95_length_unique$start, function(p)
  any(Cvir_haplotigs$start <= p & Cvir_haplotigs$end >= p)),"YES", NA)
Cvir_IAP_meanCov_CD_Hit_95_length_unique$HM_found_end <- ifelse(sapply(Cvir_IAP_meanCov_CD_Hit_95_length_unique$end, function(p)
  any(Cvir_haplotigs$start <= p & Cvir_haplotigs$end >= p)),"YES", NA)
```

View GIMAP cluster overlap with haplotig results

"YES" in both columns indicates that these genes are indeed a region identified as a haplotig, while "NA" means the genes are not inside a region identified via software to contain haplotigs.

```
##      cluster prot_identity_stat         gene                                                product   meanCov gene_length
##   Cluster 26          at 97.34% LOC111108220                    GTPase IMAP family member 7-like  945.1474        1961
##   Cluster 26                  * LOC111109315                uncharacterized protein LOC111109315  628.5355        5586
##   Cluster 36                  * LOC111109300       reticulocyte-binding protein 2 homolog a-like  560.4509        2300
##   Cluster 36          at 97.83% LOC111105930                    GTPase IMAP family member 4-like  784.2984        1528
##   Cluster 65                  * LOC111110237                uncharacterized protein LOC111110237 1005.7180        4805
##   Cluster 65          at 95.87% LOC111110097 immune-associated nucleotide-binding protein 9-like  477.3209        3409
##   Cluster 95                  * LOC111106328                    GTPase IMAP family member 4-like 1027.9801        5068
##   Cluster 95          at 98.55% LOC111106343                    GTPase IMAP family member 4-like 1756.0508        5096
##   Cluster 95          at 95.45% LOC111105744                    GTPase IMAP family member 4-like 1585.5985        6177
##  Cluster 219                  * LOC111110115                    GTPase IMAP family member 4-like  557.6818        8257
##  Cluster 219          at 98.71% LOC111106081                    GTPase IMAP family member 4-like  316.2413        7929
##  Cluster 252                  * LOC111129933     uncharacterized protein LOC111129933 isoform X1  761.2769        7135
##  Cluster 252          at 99.56% LOC111130156     uncharacterized protein LOC111130156 isoform X2 1112.3950        9653
##  HM_found_start HM_found_end
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##            <NA>         <NA>
##            <NA>         <NA>
##             YES          YES
##            <NA>         <NA>
```

View IAP cluster overlap with haplotig results

```
##      cluster prot_identity_stat         gene                                                product    meanCov gene_length
##    Cluster 5          at 99.78% LOC111130310 baculoviral IAP repeat-containing protein 6-like isoform X3  710.95093       33169
##    Cluster 5                  * LOC111129365 baculoviral IAP repeat-containing protein 6-like isoform X1  658.66650       31922
##   Cluster 17          at 97.75% LOC111100443 baculoviral IAP repeat-containing protein 2-like isoform X1  866.17023        5311
##   Cluster 17                  * LOC111100400             uncharacterized protein LOC111100400 isoform X1 1581.80383       15332
##   Cluster 30          at 96.32% LOC111100408            baculoviral IAP repeat-containing protein 2-like  572.59491       29183
##   Cluster 30                  * LOC111100402                        uncharacterized protein LOC111100402  679.60083       24362
##   Cluster 62                  * LOC111100470            baculoviral IAP repeat-containing protein 3-like  389.47702       13236
##   Cluster 62          at 97.09% LOC111101689            baculoviral IAP repeat-containing protein 2-like  499.57413       16106
##  Cluster 118                  * LOC111100416            baculoviral IAP repeat-containing protein 7-like  823.04767        9310
##  Cluster 118          at 97.30% LOC111100417            baculoviral IAP repeat-containing protein 7-like  921.41998        8689
##  Cluster 137                  * LOC111101864            baculoviral IAP repeat-containing protein 3-like  672.46875        6671
##  Cluster 137          at 95.75% LOC111101018            baculoviral IAP repeat-containing protein 3-like  930.51135        6813
##  Cluster 153                  * LOC111104279          baculoviral IAP repeat-containing protein 7-A-like  775.09753        2369
##  Cluster 153          at 96.97% LOC111105148          baculoviral IAP repeat-containing protein 7-B-like  558.19763        2606
##  Cluster 180                  * LOC111102451                              putative inhibitor of apoptosis  543.89752        5094
##  Cluster 180          at 95.73% LOC111105503            baculoviral IAP repeat-containing protein 2-like  714.34821        6238
##  Cluster 222          at 98.48% LOC111122858                        uncharacterized protein LOC111122858  593.17334        2700
##  Cluster 222                  * LOC111122723                        uncharacterized protein LOC111122723  735.38690        2277
##  Cluster 265                  * LOC111103427            baculoviral IAP repeat-containing protein 3-like 1988.57214        9885
##  Cluster 265          at 95.02% LOC111103428            baculoviral IAP repeat-containing protein 2-like 1585.04919        6520
##  Cluster 266                  * LOC111104637                 putative inhibitor of apoptosis isoform X1  683.97888        6582
##  Cluster 266          at 98.75% LOC111103790                              putative inhibitor of apoptosis  995.14258        4811
##  Cluster 280                  * LOC111103392 baculoviral IAP repeat-containing protein 3-like isoform X1  492.13504       11552
##  Cluster 280          at 95.65% LOC111103826 baculoviral IAP repeat-containing protein 2-like isoform X1  872.68256       14534
##  Cluster 282                  * LOC111104229            baculoviral IAP repeat-containing protein 8-like  513.64270        2292
##  Cluster 282          at 97.13% LOC111103158            baculoviral IAP repeat-containing protein 8-like 1046.96606        2358
##  Cluster 315                  * LOC111105597                       E3 ubiquitin-protein ligase XIAP-like  404.49841        4400
##  Cluster 315         at 100.00% LOC111099688                       E3 ubiquitin-protein ligase XIAP-like  616.43054        4397
##  Cluster 317                  * LOC111100396                       E3 ubiquitin-protein ligase XIAP-like  576.50696        9884
##  Cluster 317          at 96.79% LOC111102530                       E3 ubiquitin-protein ligase XIAP-like  640.16211        9345
##  Cluster 328                  * LOC111132301                              putative inhibitor of apoptosis  383.89096         587
##  Cluster 328          at 98.97% LOC111114013                              putative inhibitor of apoptosis  103.34412         587
##  Cluster 328          at 98.46% LOC111103682            baculoviral IAP repeat-containing protein 3-like  103.06474         587
##  Cluster 328          at 98.46% LOC111132489                              putative inhibitor of apoptosis  890.39185         587
##  Cluster 328          at 99.49% LOC111132589                              putative inhibitor of apoptosis  223.18568         587
##  Cluster 328          at 99.49% LOC111102106                              putative inhibitor of apoptosis  148.71721         587
```

```
##  Cluster 328          at 99.49% LOC111114070                                putative inhibitor of apoptosis  113.52641         587
##  Cluster 338          at 98.61% LOC111116378          death-associated inhibitor of apoptosis 1-like  189.15207         434
##  Cluster 338                 * LOC111109152          death-associated inhibitor of apoptosis 1-like 1504.06763         503
##  Cluster 338          at 96.53% LOC111117137          death-associated inhibitor of apoptosis 1-like 1776.50464         434
##  Cluster 344                 * LOC111117856          death-associated inhibitor of apoptosis 1-like  996.02759         834
##  Cluster 344          at 95.65% LOC111116826          death-associated inhibitor of apoptosis 1-like  373.41403         855
##  Cluster 344          at 95.65% LOC111111659          death-associated inhibitor of apoptosis 1-like   84.94779         747
##  HM_found_start HM_found_end
##            <NA>         <NA>
##            <NA>          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##            <NA>         <NA>
##            <NA>         <NA>
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##            <NA>         <NA>
##            <NA>         <NA>
##             YES          YES
##            <NA>          YES
##             YES          YES
##            <NA>         <NA>
##            <NA>         <NA>
##             YES          YES
##             YES          YES
##            <NA>         <NA>
##             YES          YES
```

```
##            <NA>         <NA>
##             YES          YES
##            <NA>         <NA>
##            <NA>         <NA>
##             YES          YES
##            <NA>         <NA>
##             YES          YES
##             YES          YES
##             YES          YES
##             YES          YES
##            <NA>         <NA>
##             YES          YES
```

Overall, for both gene families, most genes identified in these clusters are inside ranges called as haplotigs.

Questions and Observations:

1. What is the expected coverage for each gene? We estimated from these results the average coverage for genes within clusters is around 1000.
2. What is the best way to identify clusters containing haplotigs?
    - Two potential methods:
    - 1) Investigate clusters where one gene has "normal" coverage and the other gene has half normal (what I did above in my results for each family).
    - 2) Take the average of mean coverage and investigate clusters where the average gene coverage across the cluster is about half of what we would expect (<500)
3. Is the haplotig finding tool over-assigning haplotigs?

Moving forward we decided to take approach 2.2 above to narrow clusters to investigate, since clusters identified with strategy 1.2 may just caused by rare genes across populations.

Average the mean gene coverage within clusters

Calculate mean coverage within gene clusters

```
Cvir_GIMAP_meanCov_CD_Hit_95_length_unique_mean <- Cvir_GIMAP_meanCov_CD_Hit_95_length_unique %>% group_by(cluster) %>%
  mutate(mean_Cov_clstr = mean(meanCov))
Cvir_IAP_meanCov_CD_Hit_95_length_unique_mean <- Cvir_IAP_meanCov_CD_Hit_95_length_unique %>% group_by(cluster) %>%
  mutate(mean_Cov_clstr = mean(meanCov))
```

View GIMAP mean coverage results

```
## # A tibble: 13 x 9
## # Groups:   cluster [6]
##    cluster   prot_identity_st~ gene     product                    meanCov gene_length HM_found_start HM_found_end mean_Cov_clstr
##    <chr>     <chr>             <chr>    <chr>                        <dbl> <chr>       <chr>          <chr>                  <dbl>
##  1 Cluster ~ at 97.34%         LOC1111~ GTPase IMAP family member ~    945. 1961        YES            YES                     787.
##  2 Cluster ~ *                 LOC1111~ uncharacterized protein LO~    629. 5586        YES            YES                     787.
##  3 Cluster ~ *                 LOC1111~ reticulocyte-binding prote~    560. 2300        YES            YES                     672.
##  4 Cluster ~ at 97.83%         LOC1111~ GTPase IMAP family member ~    784. 1528        YES            YES                     672.
##  5 Cluster ~ *                 LOC1111~ uncharacterized protein LO~   1006. 4805        YES            YES                     742.
##  6 Cluster ~ at 95.87%         LOC1111~ immune-associated nucleoti~    477. 3409        YES            YES                     742.
##  7 Cluster ~ *                 LOC1111~ GTPase IMAP family member ~   1028. 5068        YES            YES                    1457.
##  8 Cluster ~ at 98.55%         LOC1111~ GTPase IMAP family member ~   1756. 5096        YES            YES                    1457.
##  9 Cluster ~ at 95.45%         LOC1111~ GTPase IMAP family member ~   1586. 6177        YES            YES                    1457.
## 10 Cluster ~ *                 LOC1111~ GTPase IMAP family member ~    558. 8257        <NA>           <NA>                    437.
## 11 Cluster ~ at 98.71%         LOC1111~ GTPase IMAP family member ~    316. 7929        <NA>           <NA>                    437.
## 12 Cluster ~ *                 LOC1111~ uncharacterized protein LO~    761. 7135        YES            YES                     937.
## 13 Cluster ~ at 99.56%         LOC1111~ uncharacterized protein LO~   1112. 9653        <NA>           <NA>                    937.
```

GIMAP results:

- Cluster 291: Mean coverage of 436. Inspected the nucleotide sequence alignment of these genes LOC111110115, LOC111106081 below.

View IAP mean coverage results

```
## # A tibble: 43 x 9
## # Groups:   cluster [18]
##    cluster   prot_identity_st~ gene     product                     meanCov gene_length HM_found_start HM_found_end mean_Cov_clstr
##    <chr>     <chr>             <chr>    <chr>                         <dbl> <chr>       <chr>          <chr>                  <dbl>
##  1 Cluster 5 at 99.78%         LOC1111~ baculoviral IAP repeat-con~    711. 33169       <NA>           <NA>                    685.
##  2 Cluster 5 *                 LOC1111~ baculoviral IAP repeat-con~    659. 31922       <NA>           YES                     685.
##  3 Cluster ~ at 97.75%         LOC1111~ baculoviral IAP repeat-con~    866. 5311        YES            YES                    1224.
##  4 Cluster ~ *                 LOC1111~ uncharacterized protein LO~   1582. 15332       YES            YES                    1224.
##  5 Cluster ~ at 96.32%         LOC1111~ baculoviral IAP repeat-con~    573. 29183       YES            YES                     626.
##  6 Cluster ~ *                 LOC1111~ uncharacterized protein LO~    680. 24362       YES            YES                     626.
##  7 Cluster ~ *                 LOC1111~ baculoviral IAP repeat-con~    389. 13236       YES            YES                     445.
##  8 Cluster ~ at 97.09%         LOC1111~ baculoviral IAP repeat-con~    500. 16106       YES            YES                     445.
##  9 Cluster ~ *                 LOC1111~ baculoviral IAP repeat-con~    823. 9310        YES            YES                     872.
## 10 Cluster ~ at 97.30%         LOC1111~ baculoviral IAP repeat-con~    921. 8689        YES            YES                     872.
## # ... with 33 more rows
```

IAP results:

- Cluster 62: mean coverage of 444. Includes LOC111100470 and LOC111101689
- Cluster 328: mean coverage across cluster of 280. Includes LOC111132301 LOC111114013, LOC111103682, LOC111132489, LOC111132589, LOC111102106, LOC111114070
- Cluster 344: mean coverage across cluster 484. Includes LOC111117856, LOC111116826, LOC111111659

Inspect Gene Clusters

Now let's investigate these clusters more closely. The genes identified in each cluster were aligned using MAFFT with default settings and visualized in Unipro UGENE.

GIMAP Cluster 219

Zooming in on GIMAP cluster 219

```
## # A tibble: 2 x 7
## # Groups:   cluster [1]
```

```
##   seqid          start      end prot_identity_stat cluster     meanCov mean_Cov_clstr
##   <chr>          <int>    <int> <chr>              <chr>         <dbl>          <dbl>
## 1 NC_035787.1 71422293 71430550 *                  Cluster 219    558.           437.
## 2 NC_035787.1 69469805 69477734 at 98.71%          Cluster 219    316.           437.
```

Both sequences are on the same chromosome. Lets take a look now at several sections of nucleotide alignment of these two genes where there are sequence insertions.

Figure 1: GIMAP 219 section 1 insertion alignment.

Overall most of the nucleotide sequences have complete identity, except for several sections with insertions or deletions.

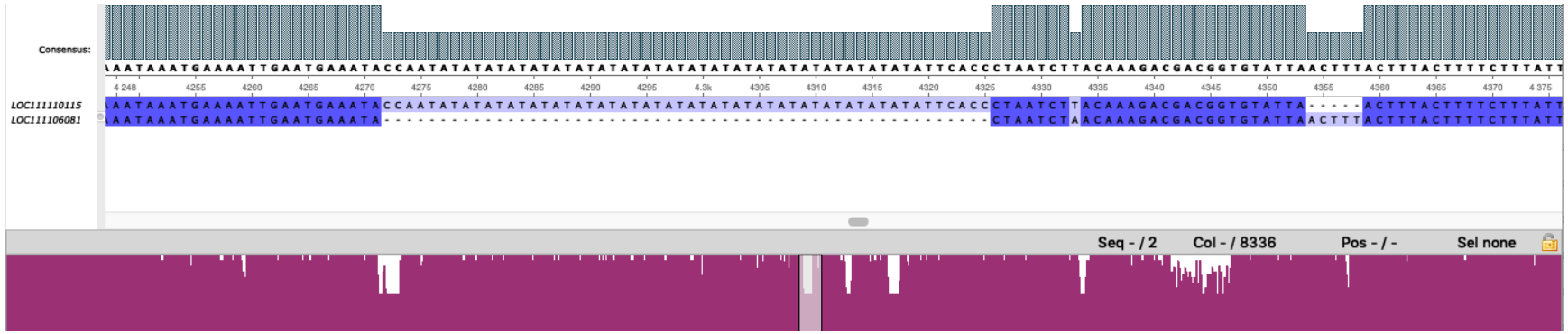


Figure 2: GIMAP 219 section 2 insertion alignment.

Figure 3: GIMAP 219 section 3 insertion alignment.

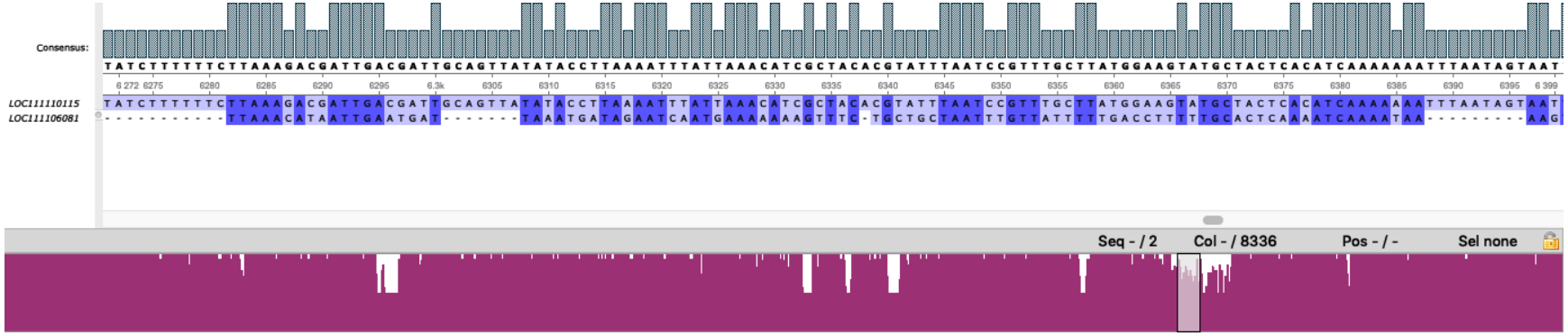


Figure 4: GIMAP 219 section 4.

Cluster 219 Conclusion: The two genes in the GIMAP protein cluster 219 should be collapsed into one based on high sequence identity and low coverage compared to other clusters where genes are similar.

IAP Cluster 62

Zooming in on IAP cluster 62

```
## # A tibble: 2 x 7
## # Groups:   cluster [1]
##   seqid          start      end prot_identity_stat cluster    meanCov mean_Cov_clstr
##   <chr>          <int>    <int> <chr>              <chr>        <dbl>          <dbl>
## 1 NC_035785.1 50080117 50093353 *                  Cluster 62    389.           445.
## 2 NC_035785.1 26008111 26024217 at 97.09%          Cluster 62    500.           445.
```

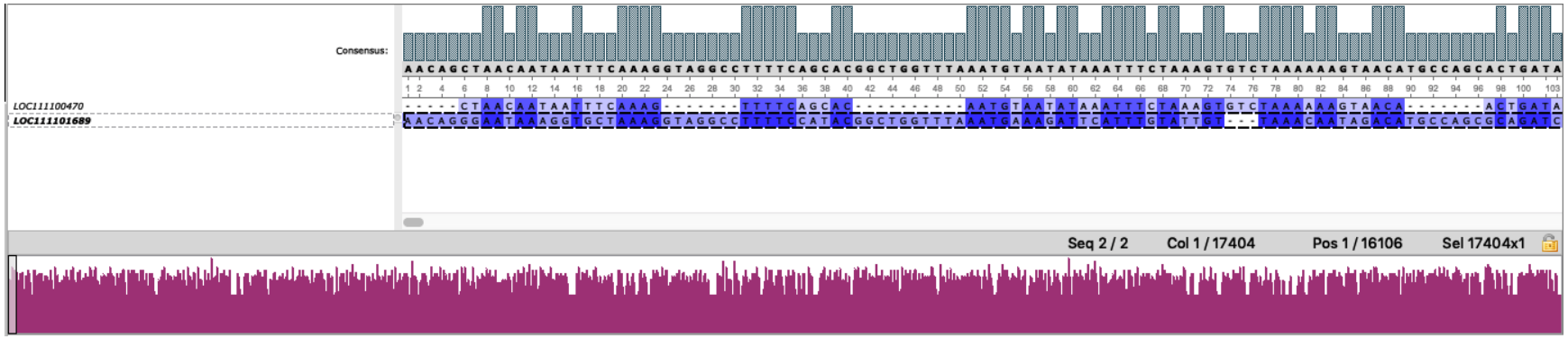


Figure 5: IAP cluster 62 alignment.

Cluster 62 Conclusion: The two genes in the IAP protein cluster 62 should not be collapsed because they have low nucleotide level sequence identity and likely are truly different genes despite having very low coverage.

IAP Cluster 328

Zooming in on IAP cluster 328

```
## # A tibble: 7 x 6
```

```
## # Groups:   cluster [1]
##   seqid       gene         prot_identity_stat cluster     meanCov mean_Cov_clstr
##   <chr>       <chr>        <chr>              <chr>         <dbl>          <dbl>
## 1 NC_035784.1 LOC111132301 *                  Cluster 328    384.           281.
## 2 NC_035788.1 LOC111114013 at 98.97%          Cluster 328    103.           281.
## 3 NC_035786.1 LOC111103682 at 98.46%          Cluster 328    103.           281.
## 4 NC_035784.1 LOC111132489 at 98.46%          Cluster 328    890.           281.
## 5 NC_035784.1 LOC111132589 at 99.49%          Cluster 328    223.           281.
## 6 NC_035785.1 LOC111102106 at 99.49%          Cluster 328    149.           281.
## 7 NC_035788.1 LOC111114070 at 99.49%          Cluster 328    114.           281.
```

These 7 genes found on different chromosomes have an odd distribution of coverage, with two genes appearing to be "real" beacuse of higher relative coverage, and the other 5 appearing to be artifacts.

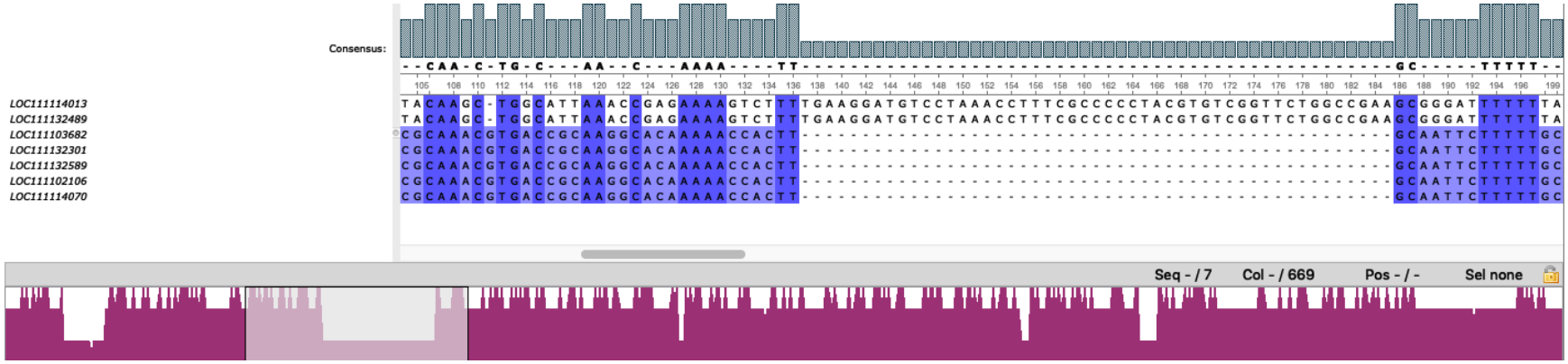


Figure 6: IAP cluster 328 section 1 alignment.

Cluster 328 Conclusion: The two genes on top of the alignment (LOC111132489 and LOC111114013) form a cluster together and are most similar and should be collapsed into one gene. One has an individual coverage of 890 while the other has coverage of 103. The five other genes are very similar in nucleotide sequence and all have relatively low coverage, though LOC111132301 has the highest relatively. These five genes should also be collapsed together.

IAP Cluster 344

Zooming in on IAP cluster 344

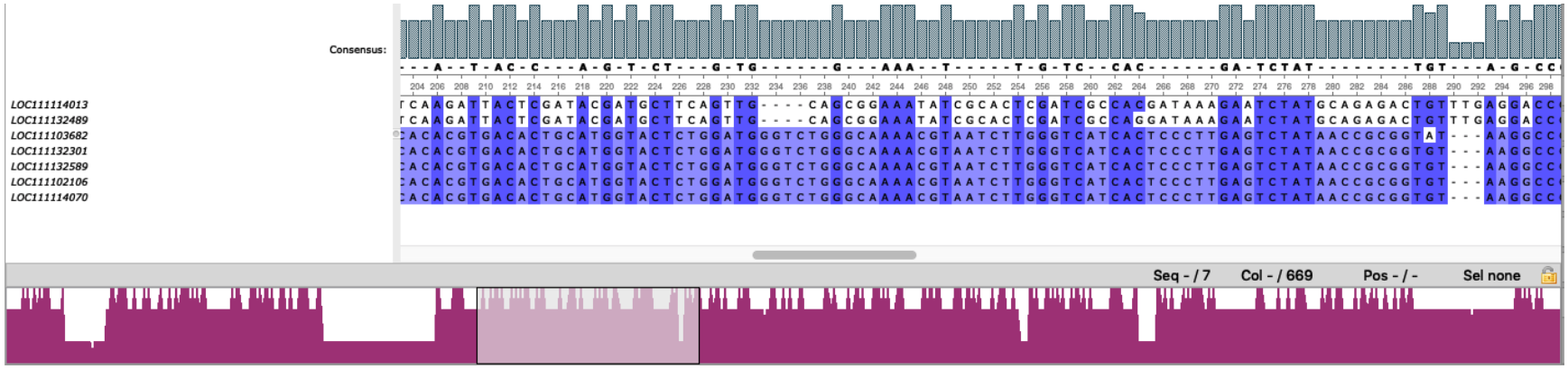


Figure 7: IAP cluster 328 section 2 alignment.

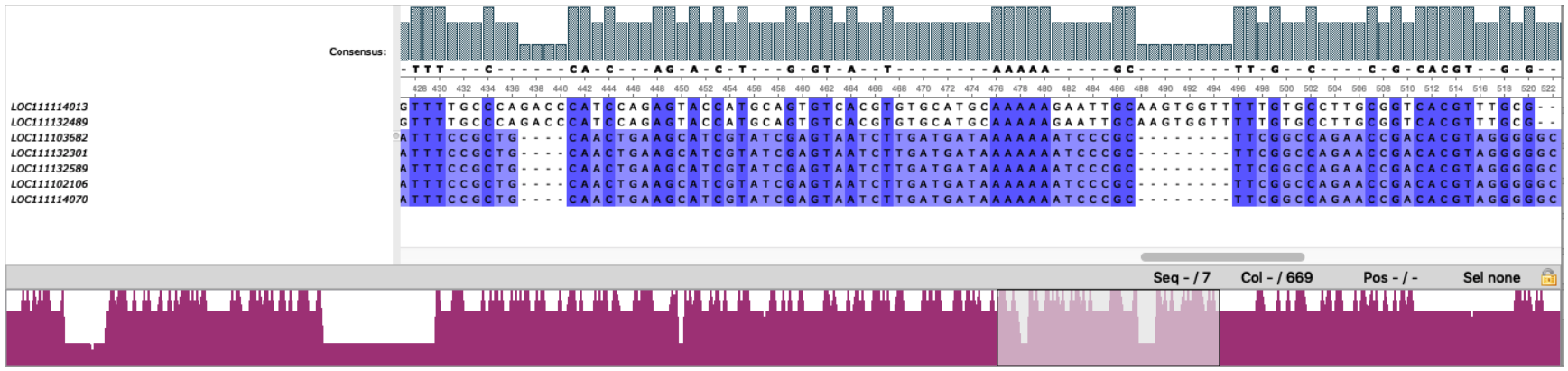


Figure 8: IAP cluster 328 section 3 alignment.

```
## # A tibble: 3 x 8
## # Groups:   cluster [1]
##   seqid          start      end gene         prot_identity_stat cluster     meanCov mean_Cov_clstr
##   <chr>          <int>    <int> <chr>        <chr>              <chr>         <dbl>          <dbl>
## 1 NC_035789.1 29989809 29990643 LOC111117856 *                  Cluster 344   996.           485.
## 2 NC_035789.1 29706284 29707139 LOC111116826 at 95.65%          Cluster 344   373.           485.
## 3 NC_035788.1  8160844  8161591 LOC111111659 at 95.65%          Cluster 344    84.9          485.
```

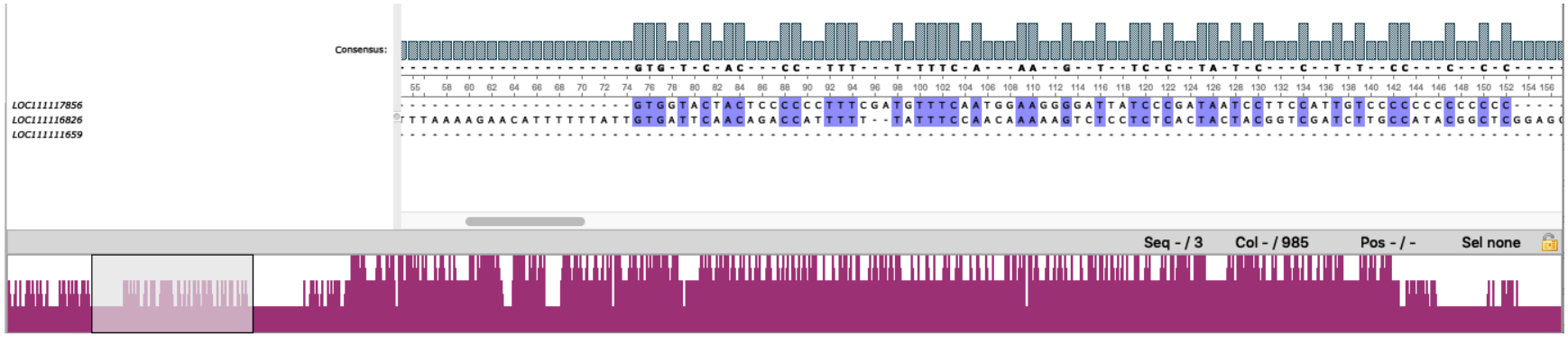


Figure 9: IAP cluster 344 section 1 alignment.

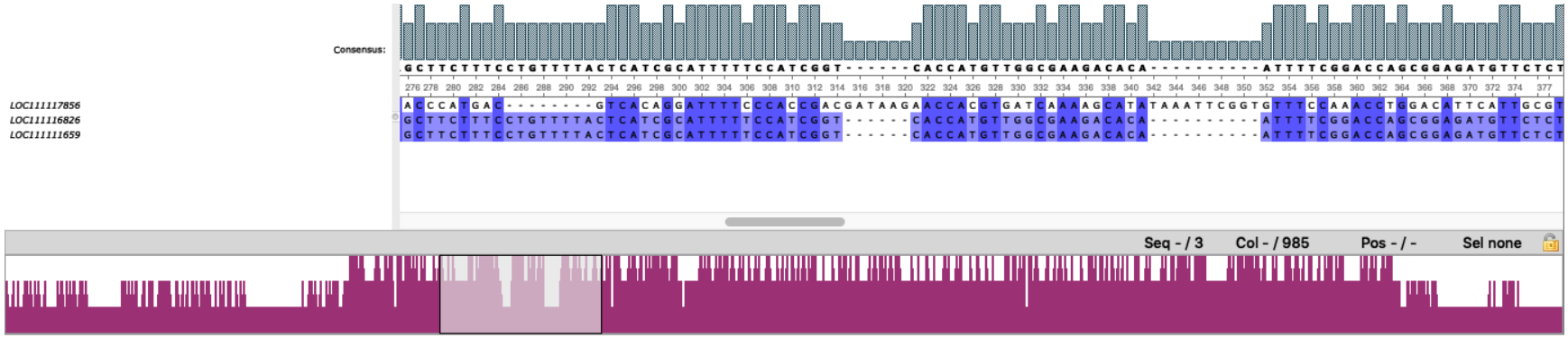


Figure 10: IAP cluster 344 section 2 alignment.

Notice that in this cluster, two genes on the same chromosome, while one is on separate chromosome. The two sequences with the greatest similarity in gene sequence are LOC111116826 and LOC111111659. These two genes should be collapsed into 1.

Final Additional Conclusions

1. If the methods used here to confirm haplotigs seem valid, the haplotig identifying tool may be over assigning haplotigs.