```
🔧 Loaded environment variables from: .env
🤖 GFF Analysis AI Agent
📊 Model: claude-3-5-haiku-latest
🌐 Provider: Anthropic
🌡️ Temperature: 0.1
--------------------------------------------------
📝 System prompt loaded from package resources
🤖 Using Anthropic Claude model: claude-3-5-haiku-latest
🔒 Anthropic: file_read tool disabled for security
📋 Batch mode - Processing queries from: newbatch2.txt
--------------------------------------------------
📊 Found 3 queries to process


============================================================
Query 1/3: list all chromosomes in
/content/GCA_050947715.1_ASM5094771v1_genomic.gff
============================================================
I'll help you list all chromosomes from the specified GFF file. I'll use
the `get_chromosomes_info` function to retrieve this information.
Tool #1: get_chromosomes_info
The GFF file contains 7 chromosomes (or chromosome-like sequences):
1. CM116930.1
2. CM116925.1
3. CM116927.1
4. CM116929.1
5. CM116928.1
6. CM116926.1
7. CM116924.1

These appear to be accession numbers for different chromosomes or contigs
in the genome. Would you like me to provide any additional information
about these chromosomes, such as their lengths or feature distributions?
============================================================
Query 2/3: Retrieve all features associated with gene ACO0OE_002170,
including its chromosomal location.
============================================================
I'll help you retrieve the features associated with the gene
ACO0OE_002170. I'll break this down into multiple steps to get
comprehensive information:

1. First, I'll get the gene structure to understand its detailed
composition:
Tool #2: get_gene_structure
```

```
I noticed an error that the gene is not found directly. Let's try a few
alternative approaches:

2. I'll search for features with this gene ID:
Tool #3: search_features_by_attribute
3. Let's try a partial match:
Tool #4: search_features_by_attribute
Great! I found the gene. Let me break down the details for you:

Chromosomal Location:
- Chromosome: CM116928.1
- Start Position: 787,105
- End Position: 788,654
- Strand: Negative (-) strand
- Total Gene Length: 1,549 base pairs

Gene Features:
1. Gene Feature:
   - ID: gene-ACO0OE_002170
   - Biotype: protein_coding
   - Locus Tag: ACO0OE_002170

2. mRNA Feature:
   - ID: rna-ACO0OE_002170-T1_mrna
   - Product: hypothetical protein
   - Length: 1,549 base pairs

3. Exon Features (2 exons):
   a. First Exon:
      - Start: 788,101
      - End: 788,654
      - Length: 553 base pairs

   b. Second Exon:
      - Start: 787,105
      - End: 788,041
      - Length: 936 base pairs

Additional Observations:
- The gene encodes a hypothetical protein
- It is located on the negative strand of chromosome CM116928.1
- It has two exons with a small intron between them

Would you like me to retrieve any additional information about this gene,
such as its protein product or specific attributes?
============================================================
Query 3/3: For the mRNA rna-ACCO0OE_002170-T1_mrna, how many coding
sequences is has?
============================================================
I'll help you find the number of coding sequences (CDS) for the mRNA
rna-ACO0OE_002170-T1_mrna. I'll use the get_gene_structure function to
retrieve this information:
Tool #5: get_gene_structure
I'll modify the approach slightly by using the full feature ID:
Tool #6: get_gene_structure
Based on the results, the mRNA rna-ACO0OE_002170-T1_mrna has 2 Coding
Sequences (CDS):

1. First CDS:
   - ID: cds-KAL6948146.1_1
   - Start: 787,105
   - End: 788,041
   - Length: 936 base pairs

2. Second CDS:
   - ID: cds-KAL6948146.1
   - Start: 788,101
   - End: 788,654
   - Length: 553 base pairs

Additional interesting details:
- Both CDS regions are on the negative strand
- The protein is annotated as a "hypothetical protein"
- The CDS has several Gene Ontology (GO) terms related to hydrolase and
metabolic activities
- The protein has cross-references to InterPro and PFAM databases

So, to directly answer your question: This mRNA has 2 coding sequences
(CDS).
============================================================
📊 Batch Processing Summary
============================================================
✅ Successful: 3/3
❌ Failed: 0/3
============================================================




🔧 Loaded environment variables from: .env
🤖 GFF Analysis AI Agent
📊 Model: claude-3-5-haiku-latest
🌐 Provider: Anthropic
```

```
🌡️ Temperature: 0.1
--------------------------------------------------
📝 System prompt loaded from package resources
🤖 Using Anthropic Claude model: claude-3-5-haiku-latest
🔒 Anthropic: file_read tool disabled for security
📋 Batch mode - Processing queries from: batch6.txt
--------------------------------------------------
📊 Found 6 queries to process


============================================================
Query 1/6: I want a list of all attributes in
GCA_050947715.1_ASM5094771v1_genomic.gff
============================================================
I'll help you retrieve the list of all attributes from the specified GFF
file. I'll use the `get_all_attributes` function to accomplish this.
Tool #1: get_all_attributes
Here's a breakdown of the attributes found in the GFF file:

1. gene_biotype
2. chromosome
3. genome
4. Dbxref
5. collection-date
6. go_component
7. Name
8. country
9. strain
10. ID
11. orig_protein_id
12. start_range
13. gbkey
14. end_range
15. go_function
16. mol_type
17. Note
18. orig_transcript_id
19. partial
20. Ontology_term
21. protein_id
22. isolation-source
23. go_process
24. product
25. gene
26. locus_tag
27. Parent
```

```
These attributes provide various types of information about the genomic
features in the file, including:
- Identifiers (ID, Name, locus_tag)
- Biological classifications (gene_biotype, mol_type)
- Ontology and functional annotations (go_component, go_function,
go_process)
- Metadata (collection-date, country, strain, isolation-source)
- Sequence-related information (protein_id, orig_protein_id,
orig_transcript_id)
- Structural information (Parent, start_range, end_range, partial)
- Additional notes and cross-references (Note, Dbxref, Ontology_term)

Is there anything specific you'd like to know about these attributes or
would you like me to help you explore them further?
============================================================
Query 2/6: Tell me which species it is
============================================================
I'll help you identify the species by using the `get_organism_info`
function with the accession number from the GFF file.
Tool #2: get_organism_info
The GFF file is for the species Hanseniaspora uvarum, which is a type of
budding yeast.

Some additional details:
- Scientific Name: Hanseniaspora uvarum
- Taxonomy ID: 29833
- Classification Group: Budding yeasts & allies

Hanseniaspora uvarum is a non-Saccharomyces yeast species that is often
found in various environments, including fruits, wine fermentation, and
other ecological niches. It's of interest in microbiology and fermentation
studies due to its unique characteristics and potential industrial
applications.

Would you like to know more about this organism or explore the genomic
features in the GFF file?
============================================================
Query 3/6: What is the geographical origin of this genome?
============================================================
I'll check the attributes we previously found to see if there's
information about the geographical origin. From the attributes list, I can
see there are two potentially relevant attributes: 'country' and
'isolation-source'. I'll use the `search_features_by_attribute` function
to retrieve this information.
Tool #3: search_features_by_attribute
I can provide you with detailed information about the geographical origin:
```

```
- Country of Origin: Australia
- Strain: AWRI5759_A6
- Collection Date: 2014
- Isolation Source: FOODON_00001071 (which seems to be a food-related
ontology term)

The genome was collected in Australia in 2014, likely from a food-related
source. The strain AWRI5759_A6 suggests it was collected by the Australian
Wine Research Institute (AWRI), which is known for studying yeasts
involved in wine fermentation.

Interestingly, the isolation source code FOODON_00001071 might represent a
specific type of food or food-related environment, though without
additional context, I can't specify exactly what that is.

Would you like me to elaborate on any of these details?
============================================================
Query 4/6: Give me the number of chromosomes
============================================================
I'll use the `get_chromosomes_info` function to retrieve the list of
chromosomes:
Tool #4: get_chromosomes_info
The genome has 7 chromosomes. Their specific identifiers are:
1. CM116924.1
2. CM116925.1
3. CM116926.1
4. CM116927.1
5. CM116928.1
6. CM116929.1
7. CM116930.1

Let me also provide some additional context about these chromosomes from
the previous query:

Chromosome | Length
-----------|-------
CM116924.1 | 616,678 bp
CM116925.1 | 971,298 bp
CM116926.1 | 1,013,274 bp
CM116927.1 | 1,293,909 bp
CM116928.1 | 1,486,871 bp
CM116929.1 | 1,575,313 bp
CM116930.1 | 1,922,076 bp

The chromosomes vary in size, with the smallest being 616,678 base pairs
and the largest being 1,922,076 base pairs. This is typical for yeast
genomes, which often have multiple chromosomes of varying sizes.
```

```
Would you like me to provide any further analysis of these chromosomes?
============================================================
Query 5/6: Give me the number of genes on each chromosome
============================================================
I'll use the `get_chromosome_summary` function to get the number of genes
on each chromosome:
Tool #5: get_chromosome_summary
Here's the breakdown of genes on each chromosome:

Chromosome | Chromosome Length | Number of Genes | Average Gene Length
-----------|------------------|----------------|--------------------
CM116924.1 | 616,678 bp       | 294            | 1,495.5 bp
CM116925.1 | 971,298 bp       | 434            | 1,631.19 bp
CM116926.1 | 1,013,274 bp     | 478            | 1,525.26 bp
CM116927.1 | 1,293,909 bp     | 594            | 1,577.15 bp
CM116928.1 | 1,486,871 bp     | 699            | 1,526.0 bp
CM116929.1 | 1,575,313 bp     | 710            | 1,562.77 bp
CM116930.1 | 1,922,076 bp     | 905            | 1,505.63 bp

Total number of genes: 4,114

Some interesting observations:
- The number of genes increases with chromosome size
- The largest chromosome (CM116930.1) has the most genes (905)
- The smallest chromosome (CM116924.1) has the fewest genes (294)
- The average gene length is relatively consistent across chromosomes,
ranging from about 1,495 to 1,631 base pairs

Would you like me to provide any further analysis of the gene
distribution?
============================================================
Query 6/6: List all vacuolar proteins and their protein_id
============================================================
I'll help you find the vacuolar proteins. I'll use a combination of
approaches to find these proteins:

1. First, I'll search for genes with GO terms related to vacuolar function
2. Then, I'll retrieve their protein IDs
Tool #6: search_genes_by_go_function_attribute
I'll try a broader search:
Tool #7: search_genes_by_go_function_attribute
I'll try a different approach by searching for features with "vacuolar" in
their attributes:
Tool #8: search_features_by_attribute
Based on the search results, here are the vacuolar proteins and their
protein IDs:
```

```
1. Vacuolar Protein Sorting-Associated Protein 4 (VPS4)
   - Protein ID: KAL6951512.1
   - Chromosome: CM116924.1
   - Gene Locus: ACO0OE_000020

2. Vacuolar ATP Synthase Subunit B (VMA2)
   - Protein ID: KAL6951487.1
   - Chromosome: CM116925.1
   - Gene Locus: ACO0OE_000723

3. Vacuolar Protein Sorting-Associated Protein 1 (VPS1)
   - Protein ID: KAL6949444.1
   - Chromosome: CM116927.1
   - Gene Locus: ACO0OE_001531

4. Vacuolar Transporter Chaperone (VTC1)
   - Protein ID: Not directly specified in the results
   - Chromosome: CM116928.1
   - Gene Locus: ACO0OE_002138

Note: There might be more vacuolar proteins in the genome, but these are
the ones that were found in this initial search. The search was based on
the "product" attribute containing the word "vacuolar".

Would you like me to perform a more comprehensive search or provide more
details about these proteins?
============================================================
📊 Batch Processing Summary
============================================================
✅ Successful: 6/6
❌ Failed: 0/6
============================================================



🔧 Loaded environment variables from: .env
🤖 GFF Analysis AI Agent
📊 Model: claude-3-5-haiku-latest
🌐 Provider: Anthropic
🌡️ Temperature: 0.1
--------------------------------------------------
📝 System prompt loaded from package resources
🤖 Using Anthropic Claude model: claude-3-5-haiku-latest
🔒 Anthropic: file_read tool disabled for security
📋 Batch mode - Processing queries from: newbatch.txt
--------------------------------------------------
```

```
📊 Found 2 queries to process


============================================================
Query 1/2: list all chromosomes in
/content/GCA_050947715.1_ASM5094771v1_genomic.gff
============================================================
I'll help you list all chromosomes from the specified GFF file. I'll use
the `get_chromosomes_info` function to retrieve this information.
Tool #1: get_chromosomes_info
The GFF file contains 7 chromosomes/sequences:
1. CM116928.1
2. CM116929.1
3. CM116925.1
4. CM116924.1
5. CM116926.1
6. CM116930.1
7. CM116927.1

These appear to be sequence accession numbers rather than traditional
chromosome names like "chr1" or "chromosome 1". This is common in many
genome assemblies, especially for bacteria or less-studied organisms.

Would you like me to provide more information about these chromosomes,
such as their lengths or feature distributions?
============================================================
Query 2/2: Identify intergenic regions longer than 2500bp on chromosome
CM116926.1 and show me the top 10 sorted by length.
============================================================
I'll use the `get_intergenic_regions` function to find intergenic regions
longer than 2500 base pairs on the specified chromosome. I'll set the
minimum length to 2500 and sort by length to get the top 10 longest
intergenic regions.
Tool #2: get_intergenic_regions
I'll present the top 10 intergenic regions sorted by length in a more
readable format:

1. Longest Intergenic Region:
   - Length: 5,720 bp
   - Position: 727,427 - 733,146
   - Upstream Gene: gene-ACO0OE_001069
   - Downstream Gene: gene-ACO0OE_001070

2. Second Longest:
   - Length: 5,137 bp
   - Position: 291,261 - 296,397
   - Upstream Gene: gene-ACO0OE_000856
```

   - Downstream Gene: gene-ACO0OE_000857

3. Third Longest:
   - Length: 4,110 bp
   - Position: 333,786 - 337,895
   - Upstream Gene: gene-ACO0OE_000876
   - Downstream Gene: gene-ACO0OE_000877

4. Fourth Longest:
   - Length: 3,736 bp
   - Position: 60,056 - 63,791
   - Upstream Gene: gene-ACO0OE_000748
   - Downstream Gene: gene-ACO0OE_000749

5. Fifth Longest:
   - Length: 3,229 bp
   - Position: 317,463 - 320,691
   - Upstream Gene: gene-ACO0OE_000868
   - Downstream Gene: gene-ACO0OE_000869

6. Sixth Longest:
   - Length: 3,205 bp
   - Position: 1,006,950 - 1,010,154
   - Upstream Gene: gene-ACO0OE_001205
   - Downstream Gene: gene-ACO0OE_001206

7. Seventh Longest:
   - Length: 3,032 bp
   - Position: 883,803 - 886,834
   - Upstream Gene: gene-ACO0OE_001147
   - Downstream Gene: gene-ACO0OE_001148

8. Eighth Longest:
   - Length: 3,001 bp
   - Position: 678,924 - 681,924
   - Upstream Gene: gene-ACO0OE_001045
   - Downstream Gene: gene-ACO0OE_001046

9. Ninth Longest:
   - Length: 2,824 bp
   - Position: 377,590 - 380,413
   - Upstream Gene: gene-ACO0OE_000896
   - Downstream Gene: gene-ACO0OE_000897

10. Tenth Longest:
    - Length: 2,736 bp
    - Position: 401,266 - 404,001

```
    - Upstream Gene: gene-ACO0OE_000905
    - Downstream Gene: gene-ACO0OE_000906

These intergenic regions represent significant gaps between genes on
chromosome CM116926.1. The longest intergenic region is 5,720 base pairs
long, located between genes ACO0OE_001069 and ACO0OE_001070.

Would you like me to provide any additional analysis or information about
these intergenic regions?
============================================================
📊 Batch Processing Summary
============================================================
✅ Successful: 2/2
❌ Failed: 0/2
============================================================
```