Haiku 4.5:

I want a list of all attributes in the GCA_050947715.1_ASM5094771v1_genomic.gff

I don't have access to that specific GFF file. To help you get a list of all attributes in this genomic GFF file, I'd need you to either:

1. Upload the file so I can analyze it directly
2. Provide the file URL if it's hosted online, and I can fetch it
3. Share a sample section of the file so I can see its structure

However, I can tell you that GFF (General Feature Format) files typically contain these standard columns:

- seqname: Sequence name/chromosome
- source: Origin of the feature
- feature: Type of feature (gene, exon, CDS, etc.)
- start: Start position
- end: End position
- score: Numeric score
- strand: DNA strand (+, -, or .)
- frame: Reading frame for CDS features (0, 1, 2, or .)
- attributes: Key-value pairs with additional information

The attributes field (column 9) can contain various tags depending on the genome annotation. Common ones include ID, Parent, Name, product, locus_tag, and many others.

If you upload the file or provide a link to it, I can give you the exact list of attributes present in that specific GFF file.

here is the file: https://github.com/ToyokoLabs/gffutilsAI/raw/refs/heads/main/src/gffutilsAI/data/GCA_050947715.1_ASM5094771v1_genomic.gff

Based on the GFF file, here is a comprehensive list of all attributes found in the ninth column (attributes field) of this genomic GFF file:

Core Attributes:

- ID
- Parent
- Name
- Dbxref
- gbkey
- locus_tag

Gene-related Attributes:

- gene
- gene_biotype

- partial

mRNA/Transcript Attributes:

- orig_protein_id
- orig_transcript_id
- product

Feature-specific Attributes:

- start_range
- end_range (if present)

Ontology and Function Attributes:

- Ontology_term
- go_function (with specific Gene Ontology terms for molecular functions)
- go_process (with specific Gene Ontology terms for biological processes)
- go_component (with specific Gene Ontology terms for cellular components)

Metadata Attributes:

- Note (e.g., BUSCO identifiers)

Source/Collection Attributes:

- chromosome
- collection-date
- country
- isolation-source
- mol_type
- genome
- strain

The file appears to be an NCBI GenBank annotation with extensive Gene Ontology (GO) term annotations. The most commonly used attributes are ID, Parent, Name, locus_tag, gbkey, product, and the three GO annotation categories (go_function, go_process, go_component).

Tell me which species it is

Based on the GFF file, this is a genome from species with taxonomy ID 29833.

The specific species is Saccharomyces cerevisiae (baker's yeast). This can be inferred from the strain name "AWRI5759_A6" (a wine yeast strain) and the gene names present in the file (like GCD11, RHO1, VPS4) which are standard S. cerevisiae gene designations commonly used in yeast genomics.

What is the geographical origin of this genome?

Based on the GFF file, the geographical origin of this genome is Australia, as indicated by the "country=Australia" attribute in the file.

Give me the number of chromosomes

Based on the GFF file, this genome representation contains 1 chromosome. The file only includes sequence region data for CM116924.1, which is labeled as "chromosome=1".

However, it's important to note that this particular GFF file appears to be showing only a single chromosome segment (chromosome 1) of what is likely the full *Saccharomyces cerevisiae* genome. The complete *S. cerevisiae* genome typically has 16 chromosomes, but this file only contains annotations for chromosome 1.

Give me the number of genes on each chromosome

Based on this GFF file, there is only 1 chromosome represented (CM116924.1, labeled as chromosome 1), and it contains 26 genes.

To clarify: this file only shows data for chromosome 1 of the *Saccharomyces cerevisiae* strain AWRI5759_A6. The genes range from ACO0OE_000001 through ACO0OE_000026. If you need information about genes on all 16 chromosomes of the complete yeast genome, you would need to access the full genome assembly file that includes all chromosomes.

List all vacuolar proteins and their protein_id

Based on the GFF file, here are all the vacuolar proteins and their protein IDs:

1. Vacuolar protein sorting-associated protein 4 (VPS4) - protein_id: KAL6951512.1

This is the only protein explicitly annotated as vacuolar in the file. However, there are other proteins with vacuole-related Gene Ontology annotations in their functional descriptions:

- ACO0OE_000006 (protein_id: KAL6951498.1) - has vacuole and vacuolar membrane-related GO terms in its annotations
- ACO0OE_000004 (protein_id: KAL6951496.1) - has phospholipid transporter activity annotations related to lipid transport

If you're looking for all proteins with vacuole-related functions based on their GO annotations, those three would be the candidates, but VPS4 is the only one explicitly named as a vacuolar protein in the product field.