



GCA_050947715.1_AS...
File

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

Thought for 3m 40s >

Here are **all unique attribute keys** present in the attributes (9th) column of your file **GCA_050947715.1_ASM5094771v1_genomic.gff**. I parsed every non-comment feature line in the file.

Attribute keys (27 total)

chromosome, collection-date, country, Dbxref, end_range, gbkey, gene, gene_biotype, genome, go_component, go_function, go_process, ID, isolation-source, locus_tag, mol_type, Name, Note, Ontology_term, orig_protein_id, orig_transcript_id, Parent, partial, product, protein_id, start_range, strain.

For convenience, here's how often each key appears (out of **17,135** feature rows):

Attribute	Count
chromosome	7
collection-date	7
country	7
Dbxref	4402
end_range	8
gbkey	17135
gene	2562
gene_biotype	4114

Attribute	Count
genome	7
go_component	2662
go_function	2835
go_process	2827
ID	17135
isolation-source	7
locus_tag	17128
mol_type	7
Name	8516
Note	1193
Ontology_term	3399
orig_protein_id	8399
orig_transcript_id	12794
Parent	13014
partial	34
product	13014
protein_id	4395
start_range	8
strain	7

Notes

- The following keys often carry **multiple values** per feature (comma-separated or pipe-annotated): `Dbxref` , `Ontology_term` , `go_component` , `go_function` , `go_process` , `start_range` , `end_range` .
- Region-level metadata (e.g., `chromosome` , `genome` , `collection-date` , `country` , `mol_type` , `isolation-source` , `strain`) occurs on the 7 `region` features.

If you'd like, I can also export these keys (with counts, and an example value for each) to a CSV.



Tell me which species it is

Thought for 3m 7s >

It's ***Hanseniaspora uvarum*** (NCBI Taxonomy ID **29833**; synonym ***Kloeckera apiculata***).

Your GFF header includes:

shell

Copy code

```
##species ... id=29833
```

which maps to *Hanseniaspora uvarum*.

NCBI +1

(FYI: the region features also list `strain=AWRI5759_A6` in the attributes.)



What is the geographical origin of this genome?







Thought for 1m 35s >

Australia.

That’s taken directly from the GFF’s region features, which include `country=Australia` (repeated for each chromosome). For example:

python-repl Copy code

```
... region ... ID=CM116924.1:1..616678; ... country=Australia; collection-
```



Give me the number of chromosomes

Thought for 2m 4s >

7 chromosomes.

From the GFF’s `region` features (`genome=chromosome`), the chromosomes are labeled 1–7:

Chromosome	Accession	Length (bp)
1	CM116924.1	616,678
2	CM116925.1	971,298
3	CM116926.1	1,013,274
4	CM116927.1	1,293,909
5	CM116928.1	1,486,871
6	CM116929.1	1,575,313
7	CM116930.1	1,922,076



Give me the number of genes on each chromosome

Thought for 2m 21s >

Here are the **gene counts per chromosome** (counting GFF entries where the 3rd column is `gene`):

Chromosome	Genes	Protein-coding	tRNA
1	294	286	8
2	434	423	11
3	478	467	11
4	594	580	14
5	699	675	24
6	710	698	12
7	905	875	30
Total	4,114	4,004	110

Notes:

- Counts are grouped by the chromosome (from the `region` features) and include all features of type `gene`.
- Only two `gene_biotype` values appear in this GFF: `protein_coding` and `tRNA`.



List all vacuolar proteins and their protein_id

Thought for 5m 12s >

I scanned your GFF for **protein-coding CDS features** whose annotations indicate vacuolar localization/function (case-insensitive match of “vacuol” in any of: `product`, `Note`, `go_component`, `go_process`, or `go_function`).

This yields **423 unique proteins** (deduplicated by `protein_id`).

Download the full list (product → protein_id, with gene & locus_tag):

[Download CSV – 423 vacuole-associated proteins](#)

How many per chromosome (from the `region` → `chromosome` mapping):

- Chr1: 28
 - Chr2: 47
 - Chr3: 55
 - Chr4: 60
 - Chr5: 73
 - Chr6: 69
 - Chr7: 91
- Total: 423**

If you only want proteins whose product name explicitly contains “vacuol”

(6 proteins; shown here in full)

Gene	Locus tag	Product	protein_id
VPS4	ACO0OE_000020	Vacuolar protein sorting-associated protein 4	KAL6951512.1
VMA2	ACO0OE_000723	Vacuolar ATP synthase subunit B	KAL6951487.1
VPS1	ACO0OE_001531	vacuolar protein sorting-associated protein 1	KAL6949444.1
VTC1	ACO0OE_002138	vacuolar transporter chaperone	KAL6948115.1
VPS74	ACO0OE_003143	Vacuolar protein sorting-associated protein 74	KAL6946162.1

Gene	Locus tag	Product	protein_id
VMA21	ACO00E_004113	vacuolar ATPase assembly integral membrane protein Vma21	KAL6944203.1

Notes

- The CSV lists: `chromosome, locus_tag, gene, product, protein_id` (one row per unique `protein_id`), sorted by chromosome and genomic position.
- Many additional proteins in the CSV don't contain the word "vacuol" in the product line but are included because their GO annotations (component/process/function) reference the vacuole (e.g., "vacuolar membrane", "vacuole organization"). If you'd