H. pseudoalbidus report

Illumina Reads QC

Georgios Koutsovoulos
University of Edinburgh

May 21, 2014

1 Objectives

- Check the quality of the Illumina reads.
- Make a preliminary assembly for each strain.
- Identify potential contamination of the Illumina libraries.

2 Conclusions

LSVM82 strain is different from the other strains. It appears that the library is contaminated and it is more obvious in Fig 2, where there are a lot of contigs at low coverage. Except LSVM82 strain

- Read Quality (Table 1)
  - Nothing problematic was found in the fastqc summary reports.
  - Between 82%-96% pairs were retained after trimming.
  - 84%-99% pairs were merged.
- Assembly Quality (Table 2)
  - Span of the assemblies between 50Mb-55Mb.
  - N50 is around 19kb-23kb.

- TAGC plots (Fig 1 - 24)
  - Contigs with ∼1000x coverage (depending on the strain) and hits to Ascomycota phylum have span ∼130kb and are similar to mitochondrial contigs.
  - The bulk of the Ascomycota contigs in the mid coverage range are either in one or two blobs.
  - There is a non annotated blob at lower GC.

3 ToDo

- Test different assemblers.
- Check the difference in strains that have different TAGC plots.
- Annotate the contigs of the low GC regions.

4 Raw Data

KW1 strain was sequenced by TGAC, and will be blessed as the reference strain. The Illumina library has an insert size of 500bp and each read is 151bp long.

A total of 23 *H. pseudoalbidus* strains were sequenced in Edinburgh Genomics. 3 strains (LSVM82, GIR-M-2, CBS122504) have read length of 250bp; while the remaining 20 have read length of 150bp. For the 20 strains with 150bp, 2 libraries were prepared, one with narrow and one with wide insert size distribution. In both cases most of the paired reads were overlapping, so the two libraries were pooled together.

FASTQC was used to check the raw sequence data.

5 Methods

The following pipeline was used for genome assembly and contamination check. Reads were trimmed and merged. Then, they were assembled with clc and the contigs were blasted against NT to create the TAGC plots.

5.1 Trimmomatic

Trimmomatic was used to trim low quality bases and remove adapter sequences. Standard Illumina adapter sequences were used. Program was run with the following parameters.

```
ILLUMINACLIP:adapters.fa:2:30:10
LEADING:3
TRAILING:3
SLIDINGWINDOW:4:20
MINLEN:51
```

5.2 PEAR

PEAR was used to merge overlapping reads. Default parameters were used.

5.3 CLC

CLC programs were used to assemble the reads into contigs and then map the reads back to the assembly. Assembler was run with default parameters Mapper was run with the following parameters.

```
-s 0.9
-l 0.9
```

5.4 BLAST+

BLAST+ was used to find similar sequences in the NT database. Program was run with the following parameters.

```
-evalue 1e-10
-max_target_seqs 1
-outfmt '6 qseqid staxids std'
```

5.5 TAGC plots

Various scripts were used to assign taxonomic clasification on contigs and create the TAGC plots.

Table 1: Read data for *H. pseudoalbidus*

Strain	Raw		Trim (%)		Pear (%)	Final	
	Reads	Bases	P	S		Reads	Bases
KW1	22,080,836	3,334,206,236	93.96	2.52	—	20,746,188	3,085,397,446
LSVM82	25,265,490	6,316,372,500	93.48	5.18	71.41	15,839,460	4,547,670,768
GIR-M-2	13,702,282	3,425,570,500	95.78	3.65	84.62	7,820,916	2,668,195,538
CBS122504	35,110,524	8,777,631,000	85.51	10.46	93.03	17,895,104	4,091,990,308
MIG-M-1-CTAB	32,800,620	4,920,093,000	82.13	1.74	98.14	14,293,854	1,752,005,433
LAN-M-1-Qiagen	30,732,236	4,609,835,400	86.47	1.82	96.89	14,261,787	1,934,133,311
FON-M-1-CTAB	30,668,060	4,600,209,000	86.92	1.74	97.98	14,131,516	1,954,694,431
CBS122191	29,749,066	4,462,359,900	88.15	1.68	96.47	14,075,505	2,028,423,995
CBS122507	20,471,310	3,070,696,500	90.20	1.85	92.80	10,276,650	1,516,023,611
CBS122503	21,004,890	3,150,733,500	90.43	1.92	95.15	10,361,776	1,560,260,269
CBS122505	20,061,902	3,009,285,300	91.51	1.68	93.67	10,098,357	1,548,353,521
2008-81-6	25,617,588	3,842,638,200	80.70	1.53	99.13	10,821,228	1,270,707,993
2008-148-4	22,105,298	3,315,794,700	82.61	1.52	98.89	9,569,321	1,177,321,097
2008-125-2	21,053,484	3,158,022,600	91.99	1.97	85.94	11,460,406	1,800,085,033
2008-152-4	29,253,918	4,388,087,700	82.61	1.59	98.62	12,718,308	1,592,638,079
2001--11-1	20,980,340	3,147,051,000	91.38	2.15	88.92	11,100,836	1,743,127,002
2008-139-1	23,260,226	3,489,033,900	92.09	1.95	90.92	12,136,318	1,930,811,857
2008-142-5	37,345,992	5,601,898,800	85.00	1.92	96.43	17,157,752	2,338,869,009
2009-86-3	30,301,038	4,545,155,700	85.82	1.84	96.84	13,970,473	1,900,658,705
2010-189-4	25,834,396	3,875,159,400	88.74	1.74	97.68	12,178,805	1,799,261,504
2010-189-5	30,773,158	4,615,973,700	87.88	1.93	94.19	14,900,958	2,152,055,137
2012-24-1	31,147,052	4,672,057,800	86.37	1.96	95.89	14,615,869	2,030,549,024
2012-38-2-2	39,073,990	5,861,098,500	86.71	2.06	94.86	18,619,126	2,663,252,903
2012-42-1-1	19,658,218	2,948,732,700	93.04	1.92	88.06	10,616,037	1,740,094,003

Table 2: CLC assembly statistics for *H. pseudoalbidus*

Strain	No. contigs	Span	L. contig	N50	GC(%)	N's span
KW1	3,865	55,703,897	269,150	40,287	41.8	427,287
LSVM82	30,659	70,733,244	115,161	11,679	39.3	3,951
GIR-M-2	7,082	50,980,250	117,023	19,799	43.8	859
CBS122504	6,600	55,043,853	157,317	23,363	42.2	876
MIG-M-1-CTAB	6,452	52,641,512	173,973	23,104	43.0	12,994
LAN-M-1-Qiagen	6,529	52,482,974	221,355	23,611	43.2	14,402
FON-M-1-CTAB	6,265	53,491,244	201,067	23,163	42.7	18,009
CBS122191	6,083	53,695,106	143,186	25,131	42.6	2,434
CBS122507	6,488	52,990,956	144,456	21,826	42.8	13,617
CBS122503	6,658	53,301,203	139,344	21,365	42.8	2,897
CBS122505	6,504	53,508,252	141,329	22,755	42.7	0
2008-81-6	7,397	51,321,948	136,402	19,696	43.3	4,225
2008-148-4	7,647	51,425,690	130,709	18,565	43.3	930
2008-125-2	6,368	53,805,655	142,933	23,521	42.6	820
2008-152-4	6,728	52,584,921	126,623	22,397	43.0	15,638
2001--11-1	6,320	53,328,487	141,326	22,797	42.7	16,751
2008-139-1	6,052	53,723,733	141,463	24,535	42.6	806
2008-142-5	6,047	53,710,115	137,987	25,765	42.6	13,354
2009-86-3	6,208	53,701,531	195,857	24,886	42.7	15,652
2010-189-4	6,365	53,453,592	241,296	24,044	42.7	915
2010-189-5	6,152	53,516,718	174,643	25,073	42.7	0
2012-24-1	6,219	53,422,785	176,316	24,024	42.7	15,533
2012-38-2-2	5,922	53,791,507	166,514	26,058	42.6	468
2012-42-1-1	6,655	53,153,132	214,916	22,132	42.8	13,616

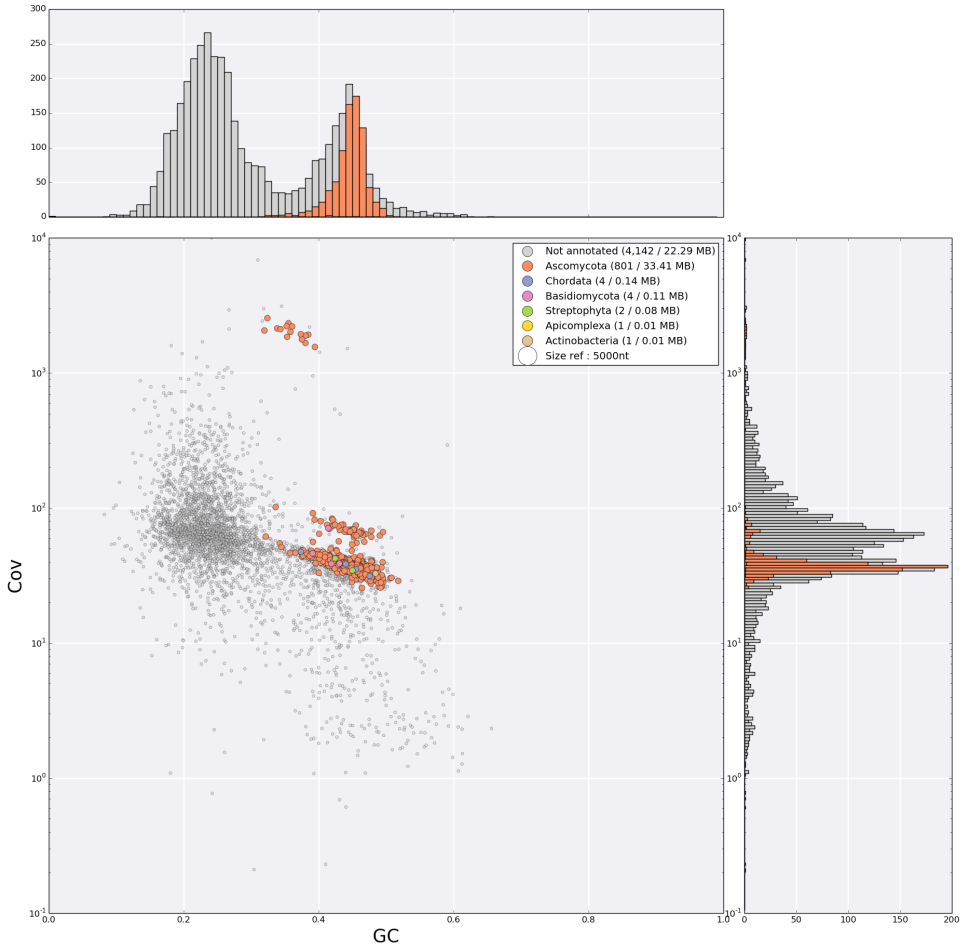

Figure 1: KW1 TAGC plot

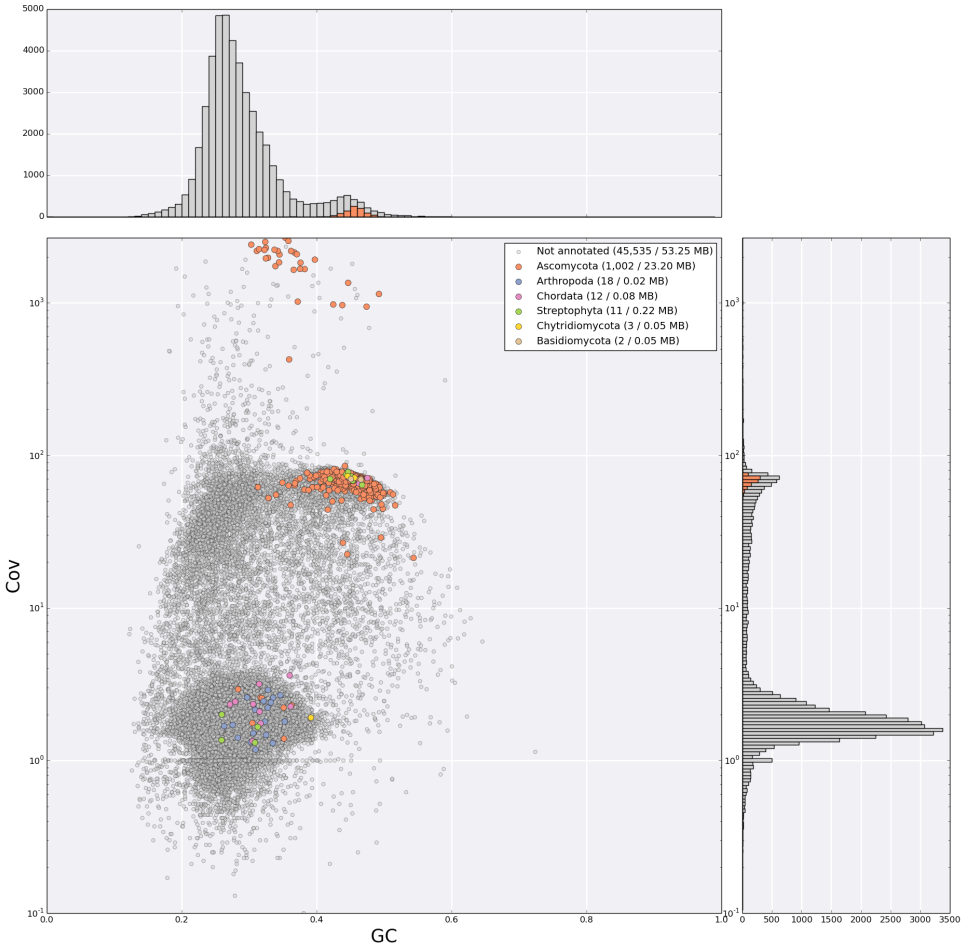

Figure 2: LSVM82 TAGC plot

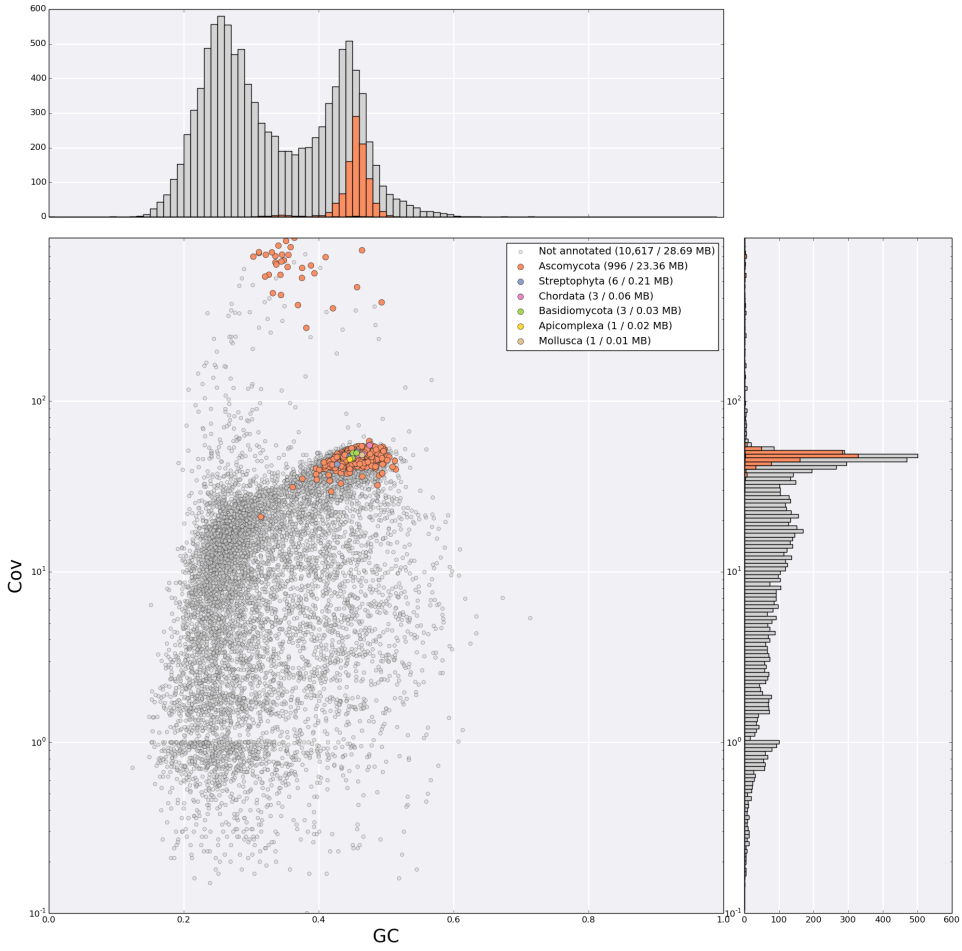

Figure 3: GIR-M-2 TAGC plot

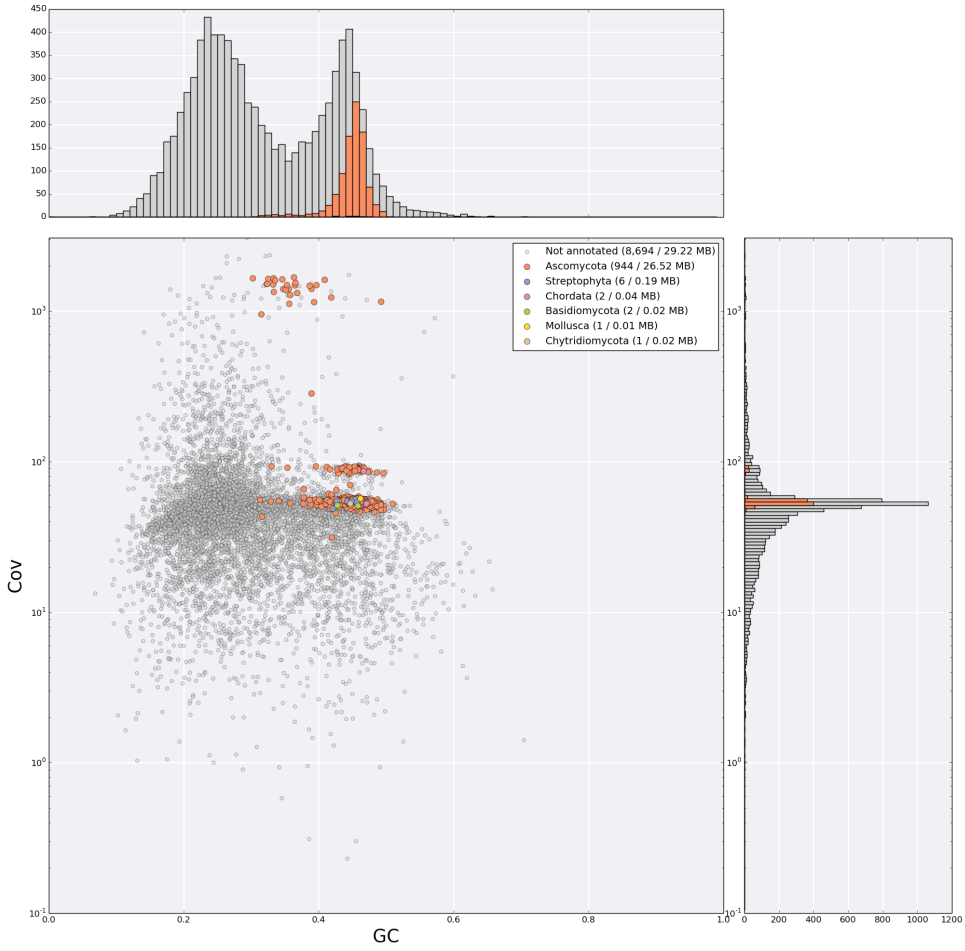

Figure 4: CBS122504 TAGC plot

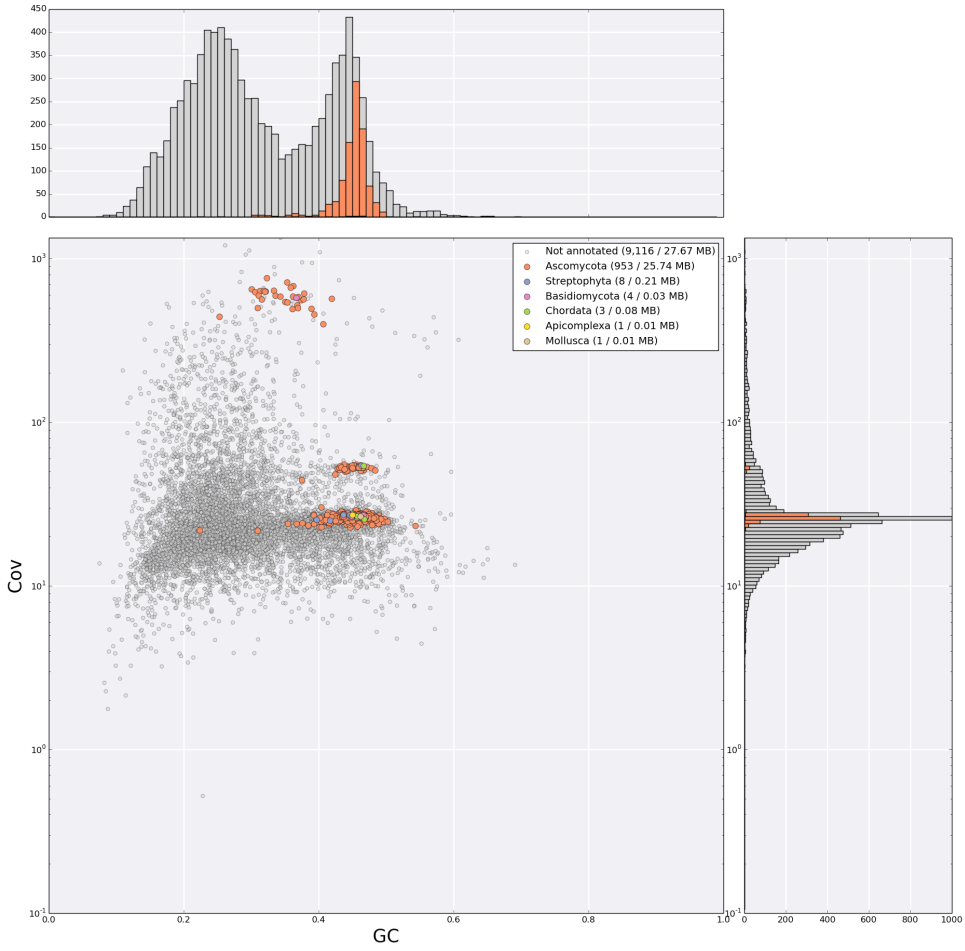

Figure 5: MIG-M-1-CTAB TAGC plot

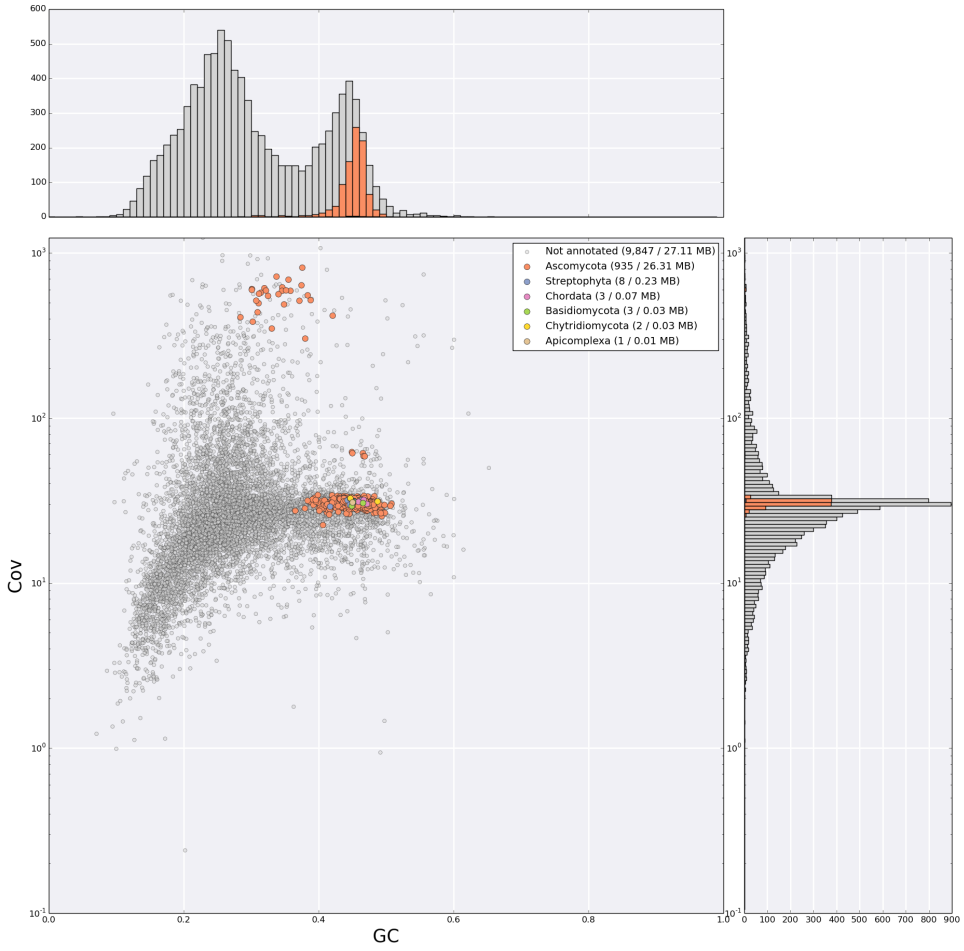

Figure 6: LAN-M-1-Qiagen TAGC plot

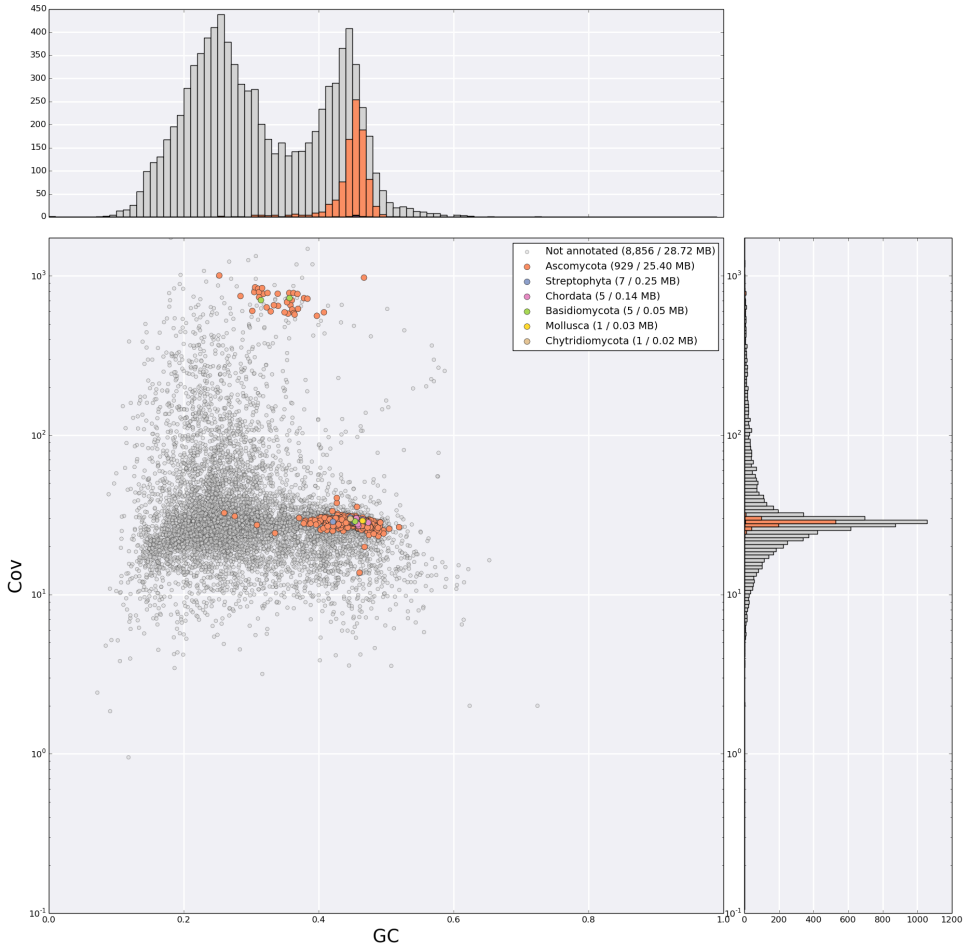

Figure 7: FON-M-1-CTAB TAGC plot

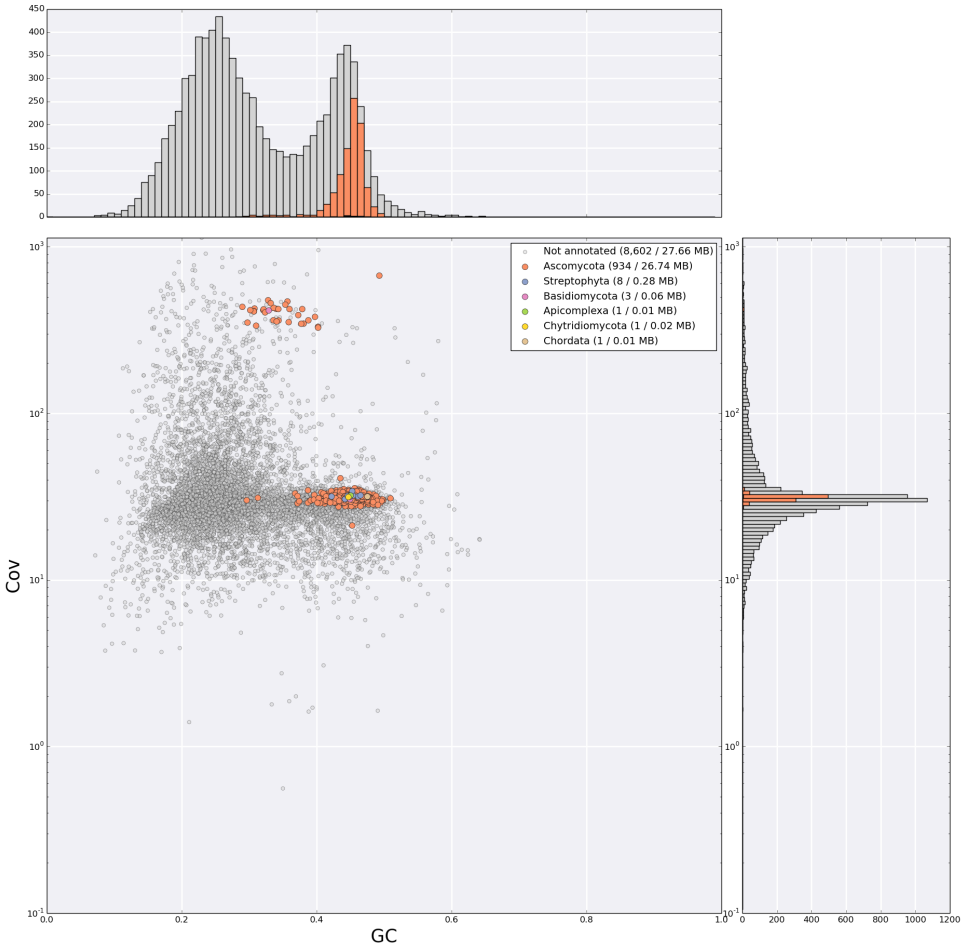

Figure 8: CBS122191 TAGC plot

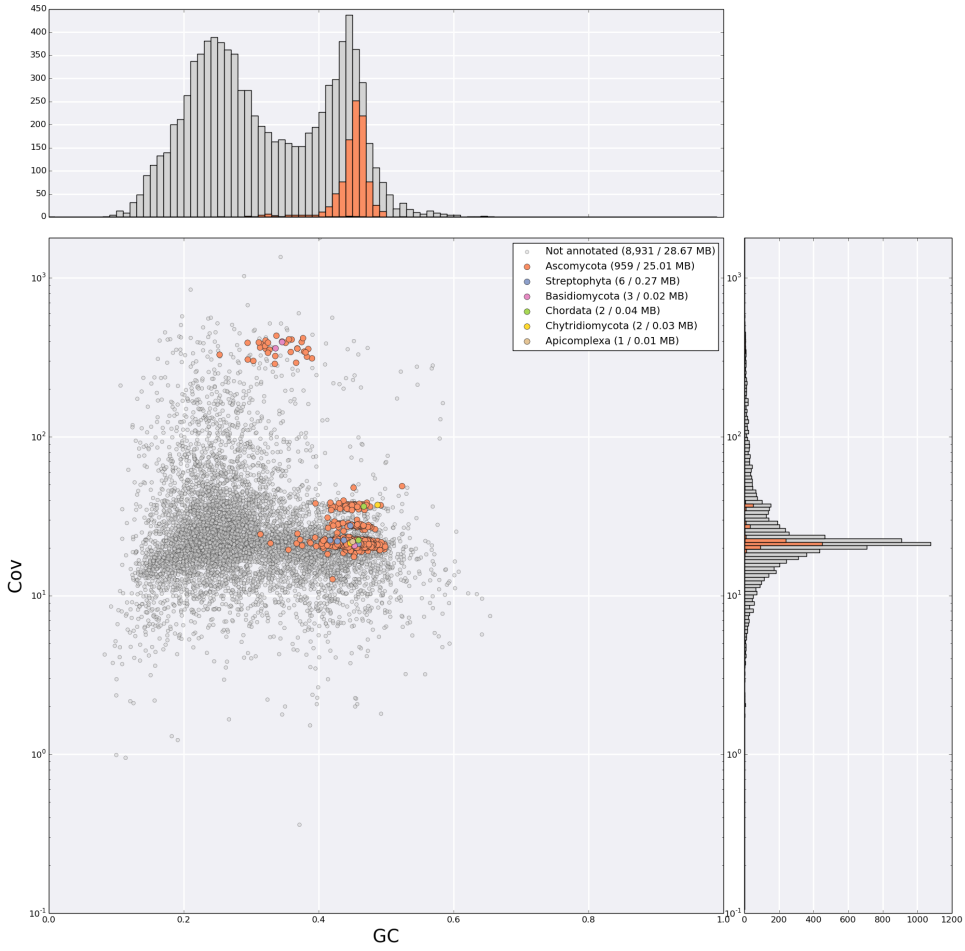

Figure 9: CBS122507 TAGC plot

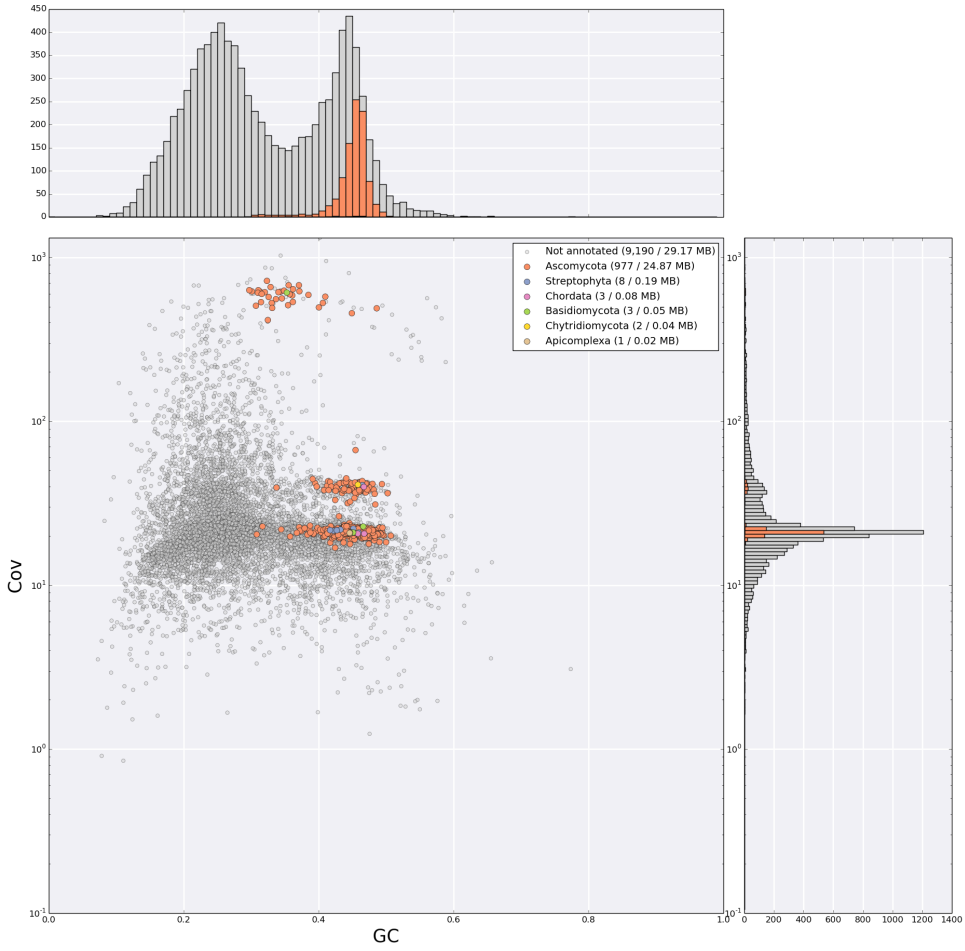

Figure 10: CBS122503 TAGC plot

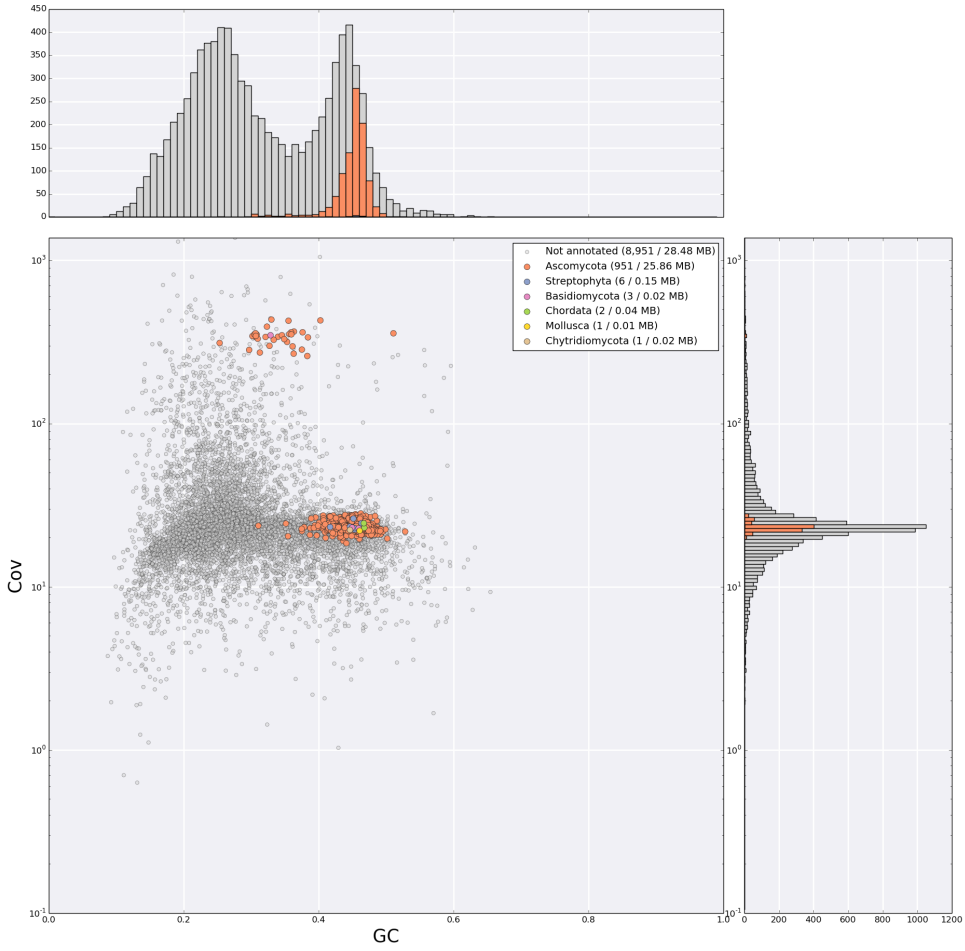

Figure 11: CBS122505 TAGC plot

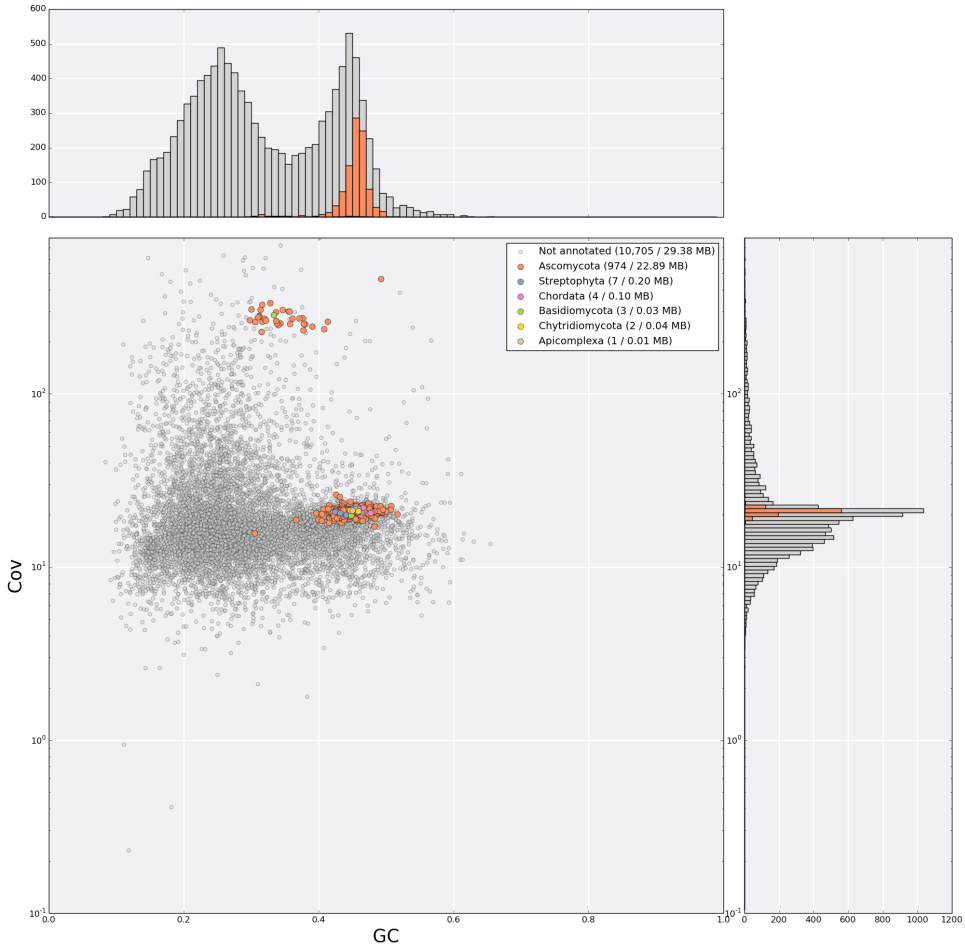

Figure 12: 2008-81-6 TAGC plot

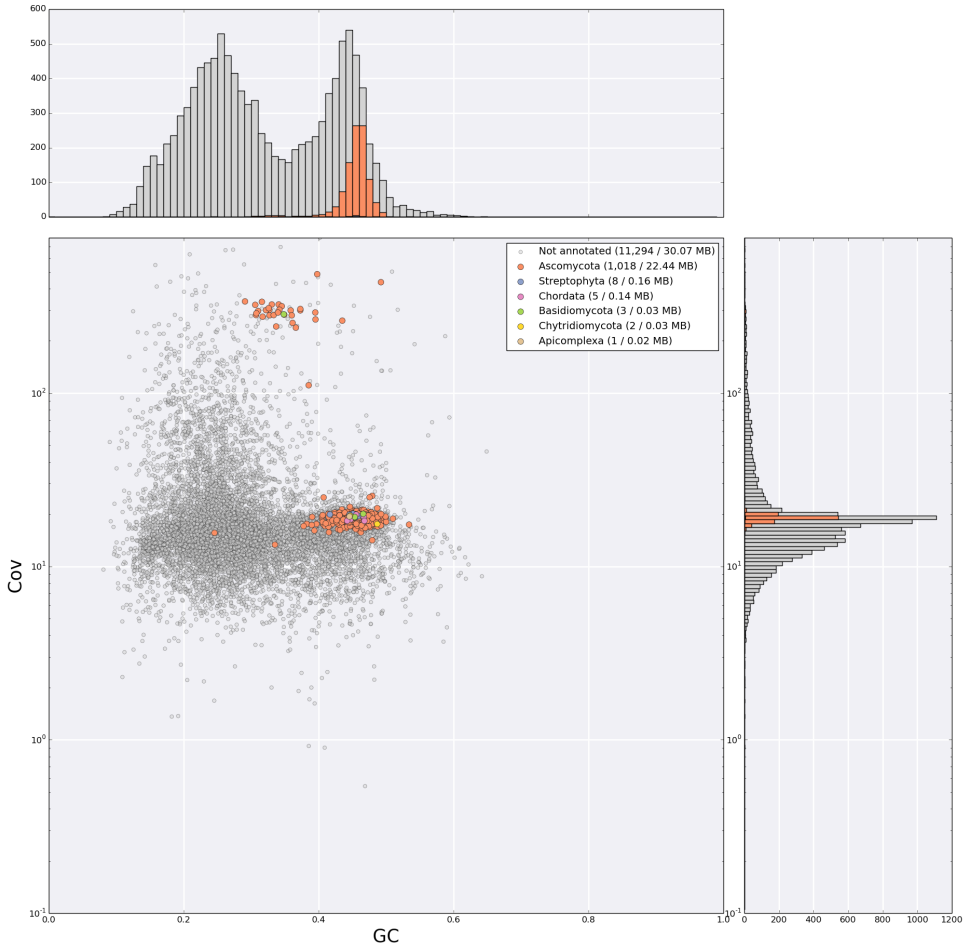

Figure 13: 2008-148-4 TAGC plot

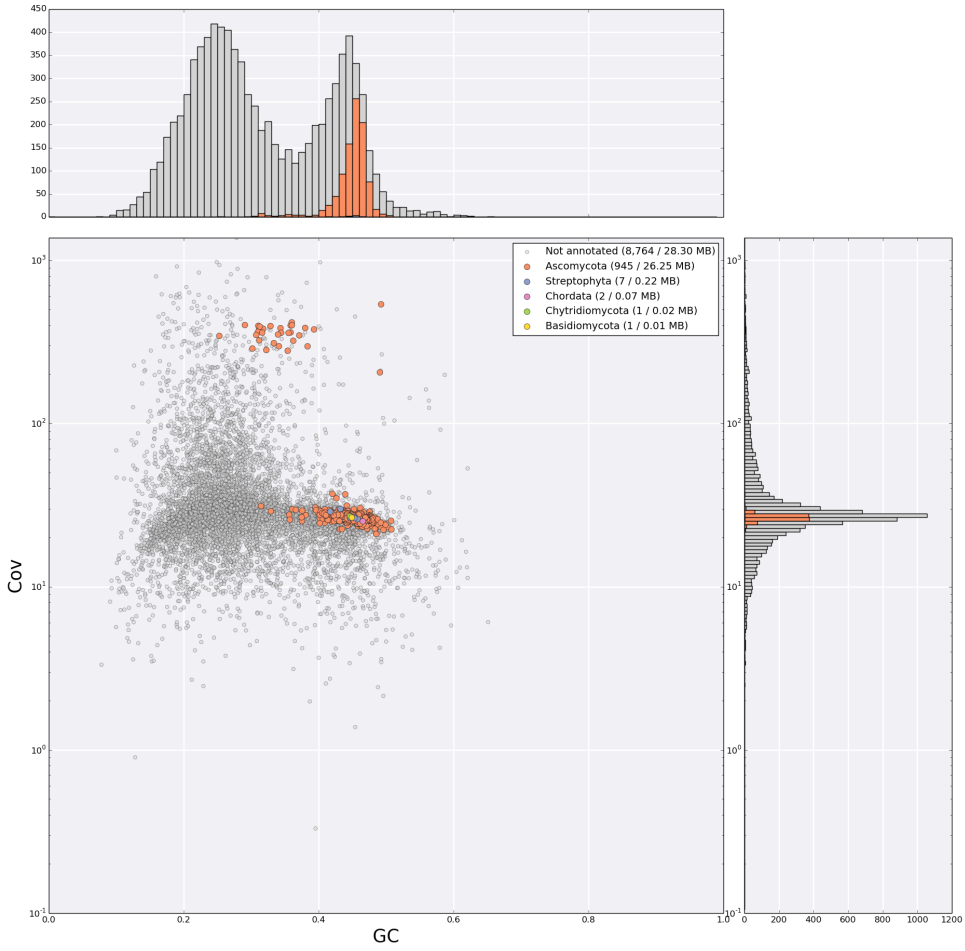

Figure 14: 2008-125-2 TAGC plot

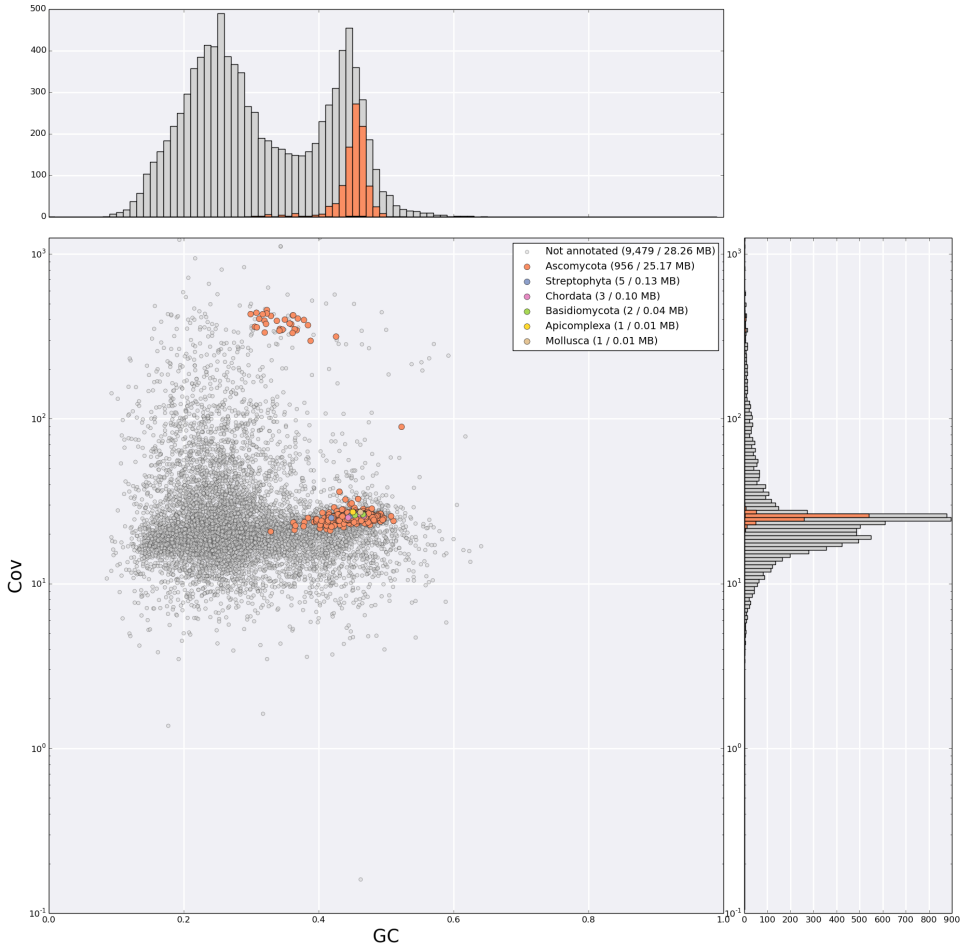

Figure 15: 2008-152-4 TAGC plot

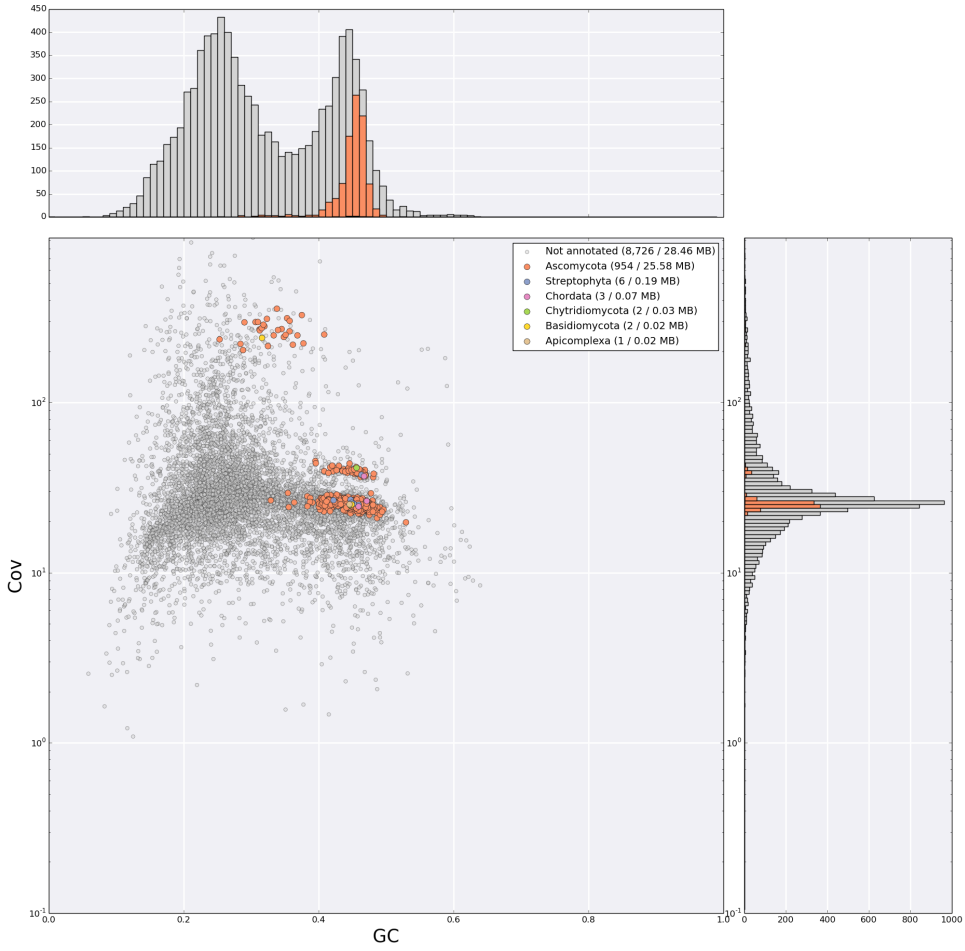

Figure 16: 2001–11-1 TAGC plot

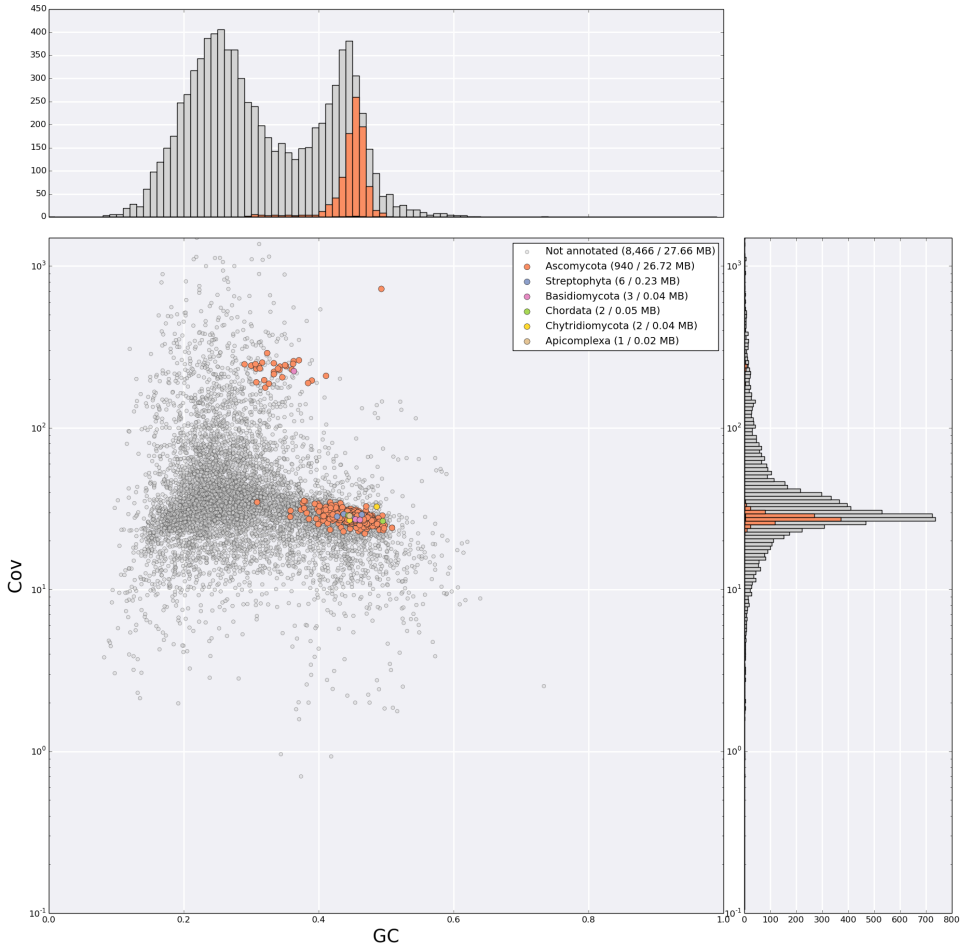

Figure 17: 2008-139-1 TAGC plot

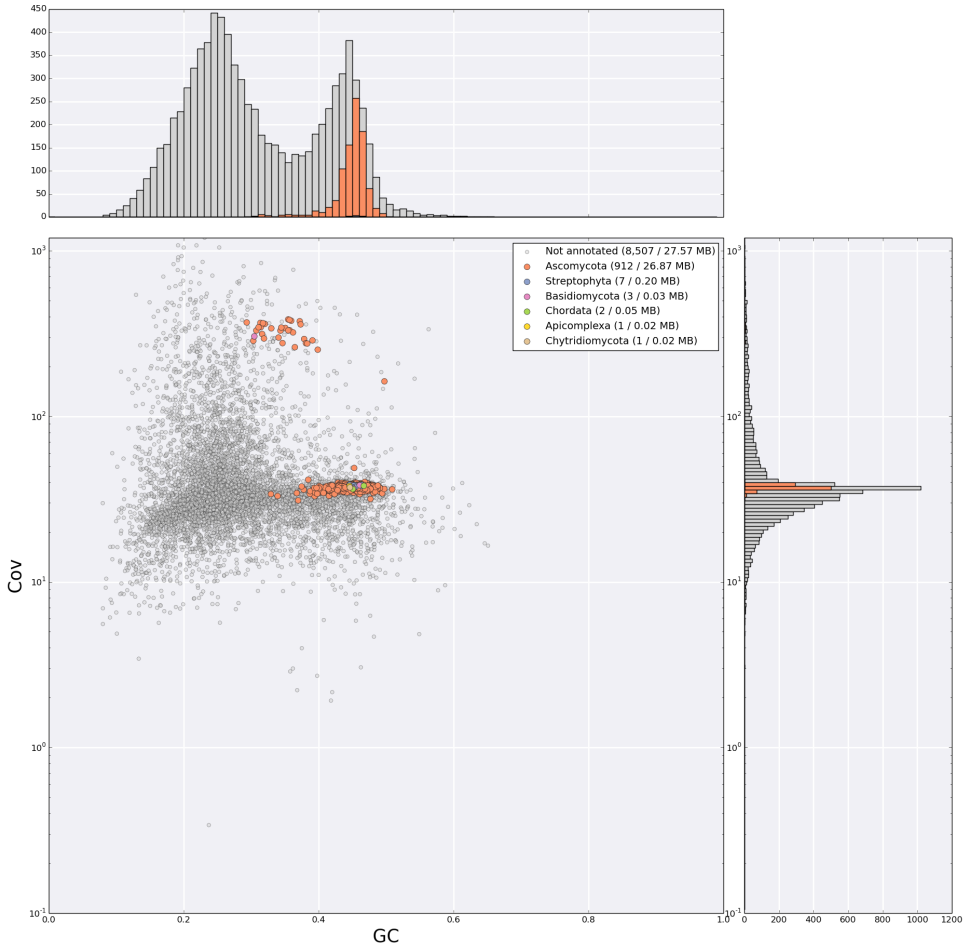

Figure 18: 2008-142-5 TAGC plot

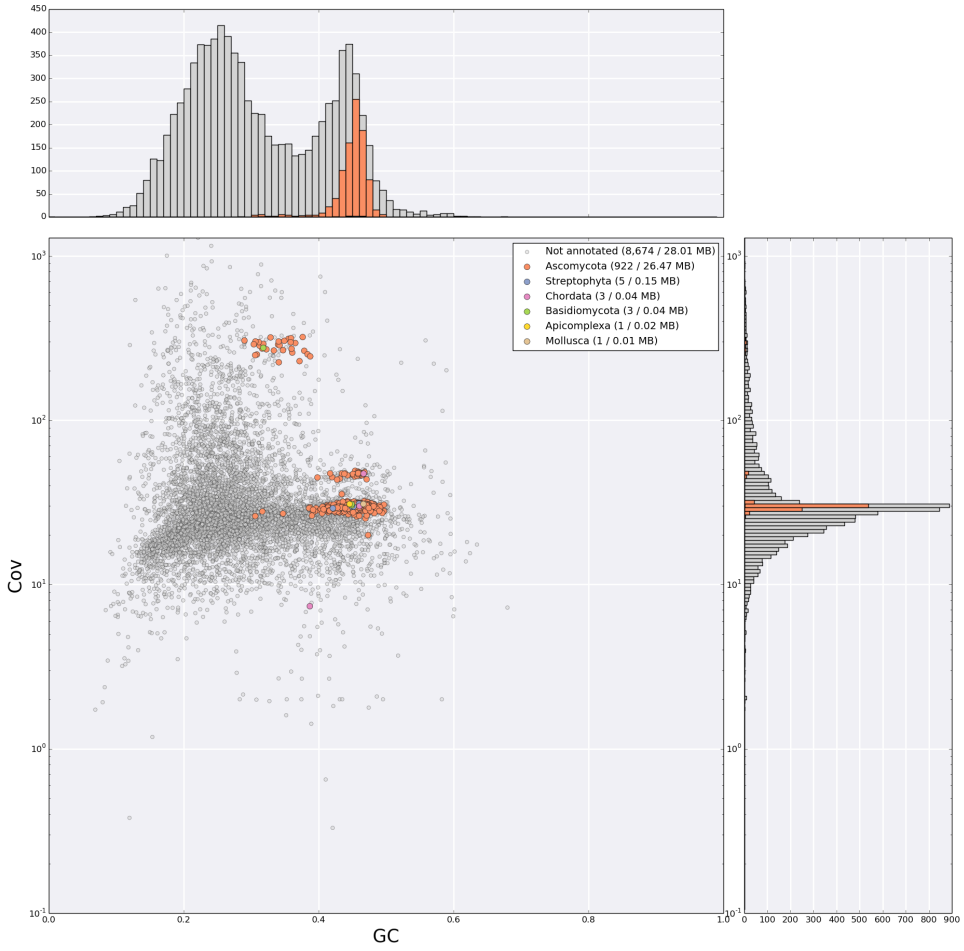

Figure 19: 2009-86-3 TAGC plot

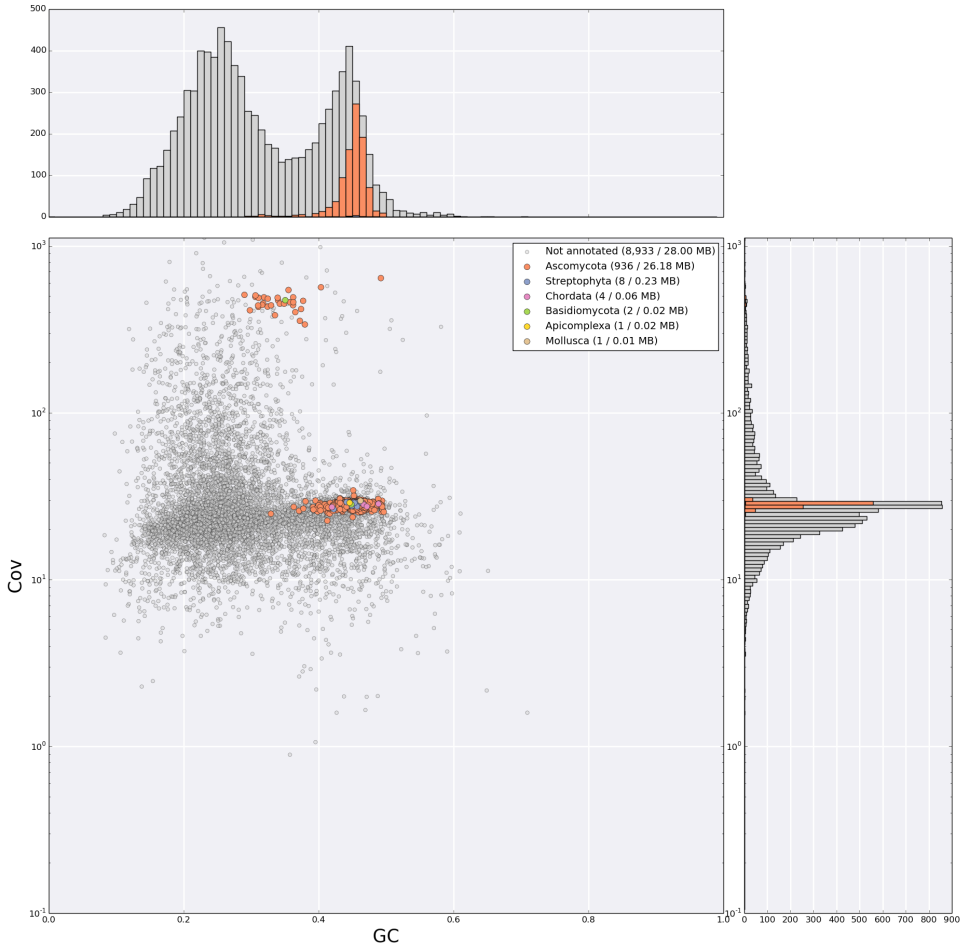

Figure 20: 2010-189-4 TAGC plot

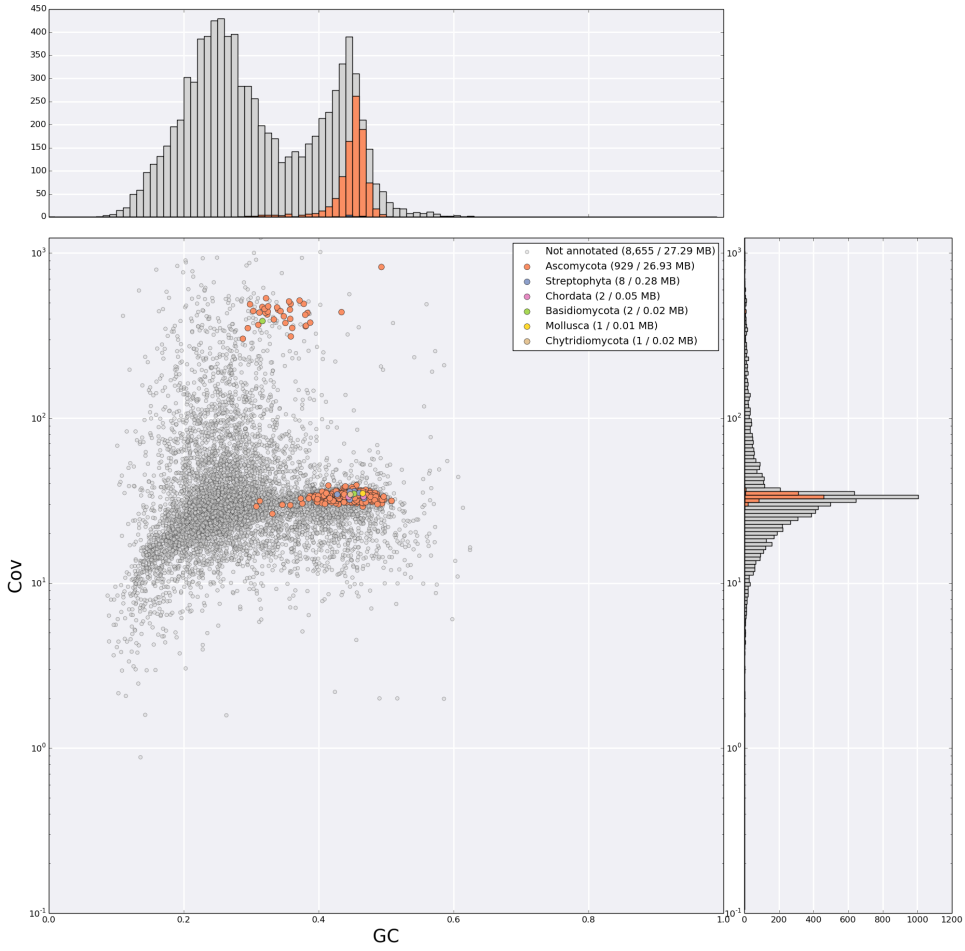

Figure 21: 2010-189-5 TAGC plot

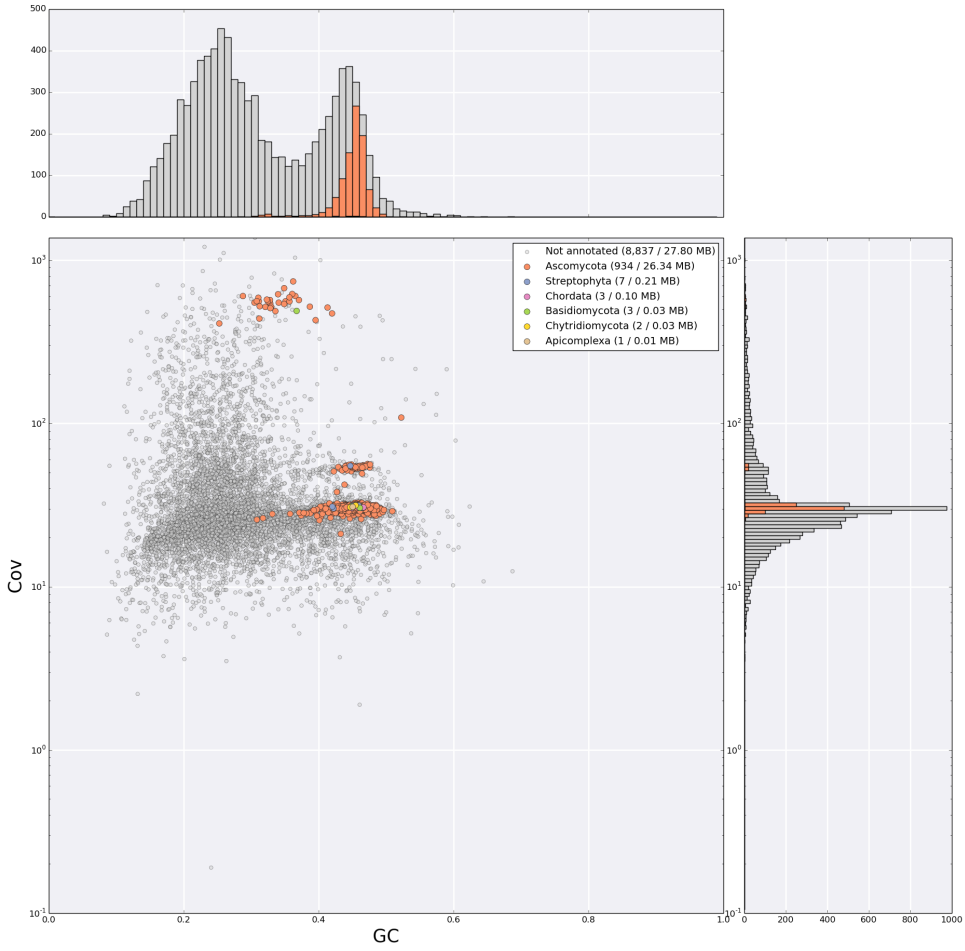

Figure 22: 2012-24-1 TAGC plot

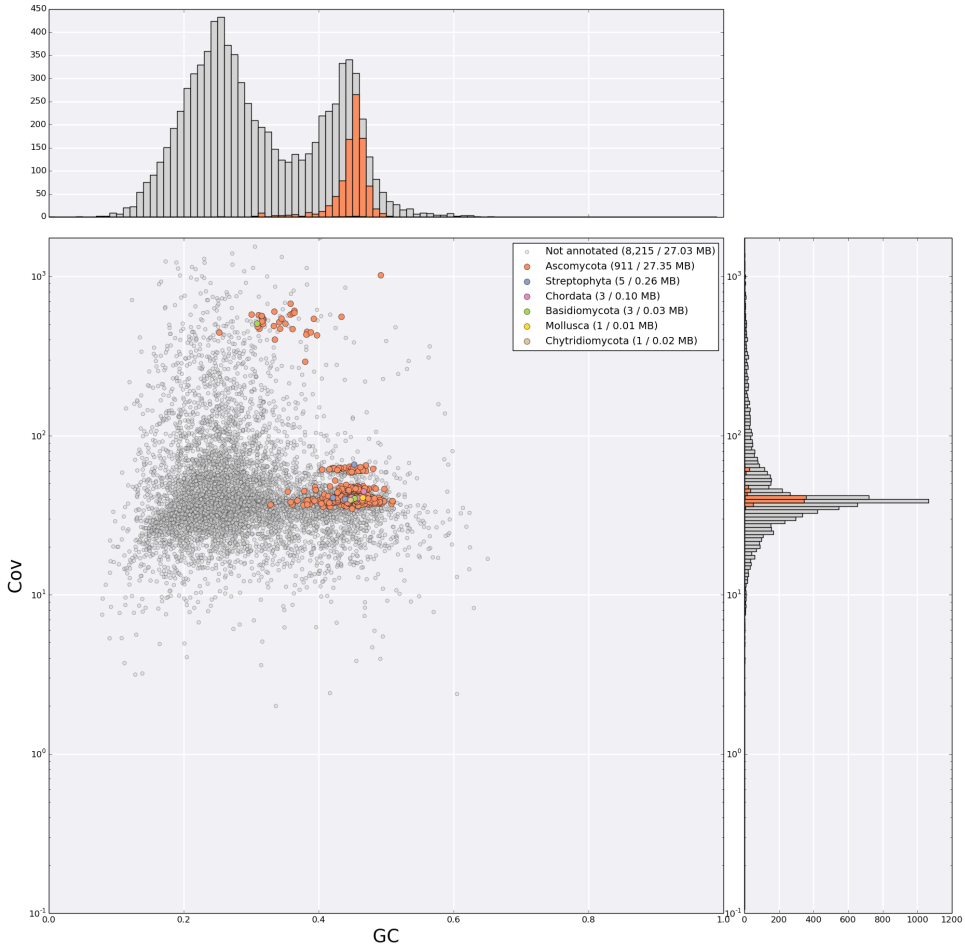

Figure 23: 2012-38-2-2 TAGC plot

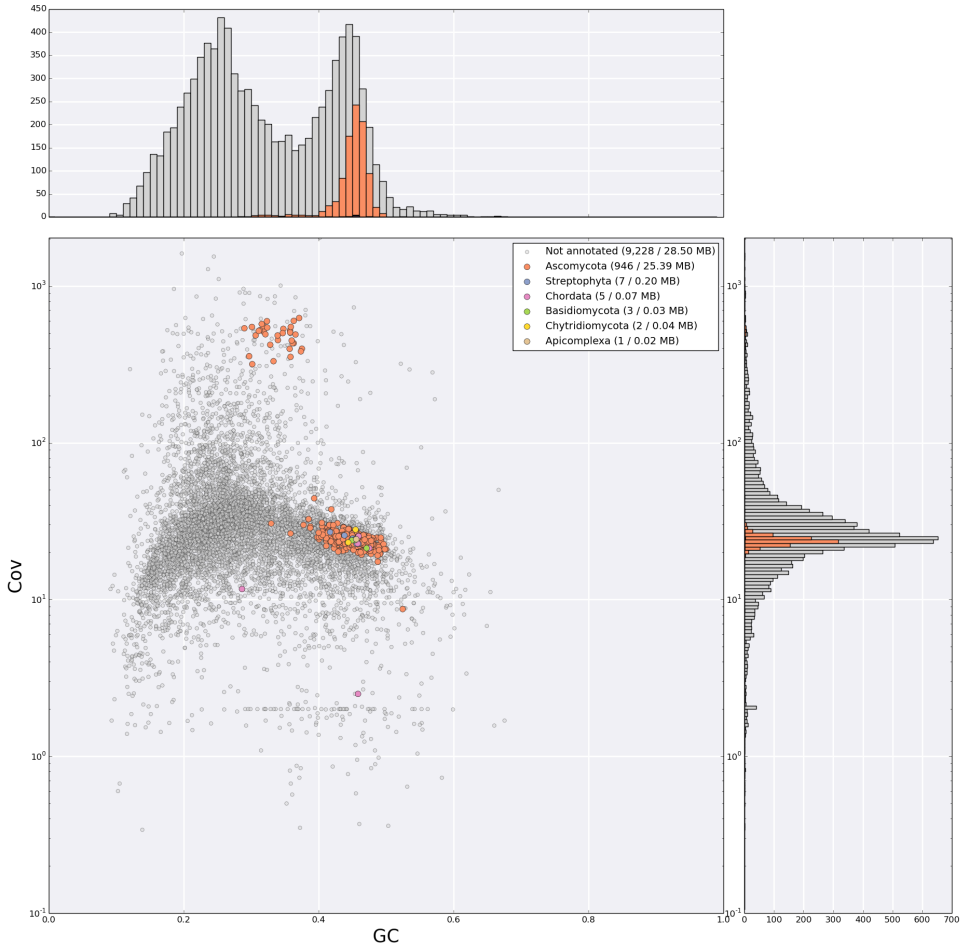

Figure 24: 2012-42-1-1 TAGC plot