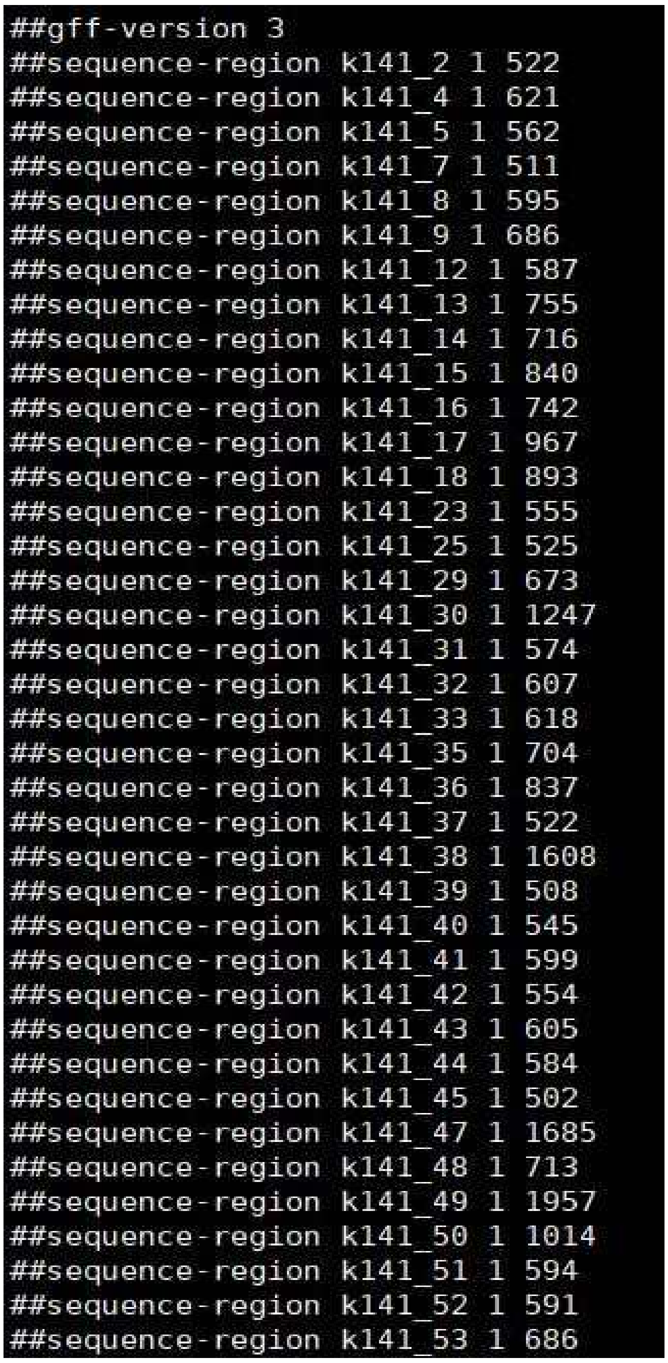
2021 2학기 과제 정리

20193852 문유빈

1) Goal : MGE(Mobile genetic element) & ARG(Antibiotic resistance gene) 둘 다 보유하는 ORF(Open reading frame) 찾기

2) Resources

(1) CF1A-114.gff -> Contig & ORF 파일



```
##gff-version 3
##sequence-region k141_2 1 522
##sequence-region k141_4 1 621
##sequence-region k141_5 1 562
##sequence-region k141_7 1 511
##sequence-region k141_8 1 595
##sequence-region k141_9 1 686
##sequence-region k141_12 1 587
##sequence-region k141_13 1 755
##sequence-region k141_14 1 716
##sequence-region k141_15 1 840
##sequence-region k141_16 1 742
##sequence-region k141_17 1 967
##sequence-region k141_18 1 893
##sequence-region k141_23 1 555
##sequence-region k141_25 1 525
##sequence-region k141_29 1 673
##sequence-region k141_30 1 1247
##sequence-region k141_31 1 574
##sequence-region k141_32 1 607
##sequence-region k141_33 1 618
##sequence-region k141_35 1 704
##sequence-region k141_36 1 837
##sequence-region k141_37 1 522
##sequence-region k141_38 1 1608
##sequence-region k141_39 1 508
##sequence-region k141_40 1 545
##sequence-region k141_41 1 599
##sequence-region k141_42 1 554
##sequence-region k141_43 1 605
##sequence-region k141_44 1 584
##sequence-region k141_45 1 502
##sequence-region k141_47 1 1685
##sequence-region k141_48 1 713
##sequence-region k141_49 1 1957
##sequence-region k141_50 1 1014
##sequence-region k141_51 1 594
##sequence-region k141_52 1 591
##sequence-region k141_53 1 686
```

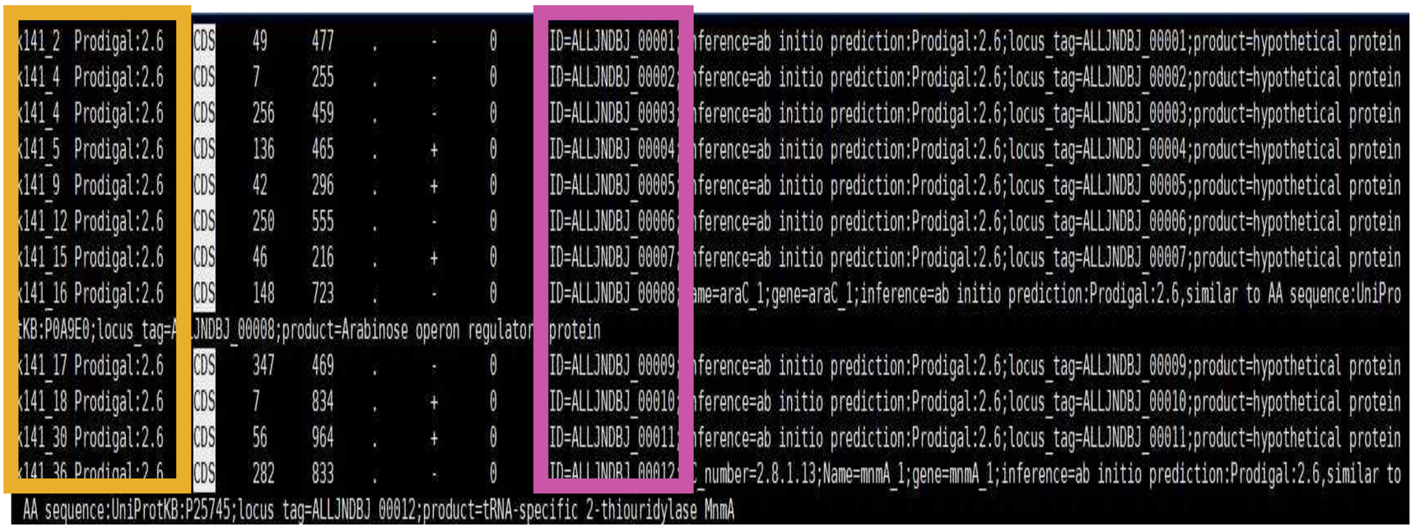
##sequence
필요 없는 것

/CDS 입력시

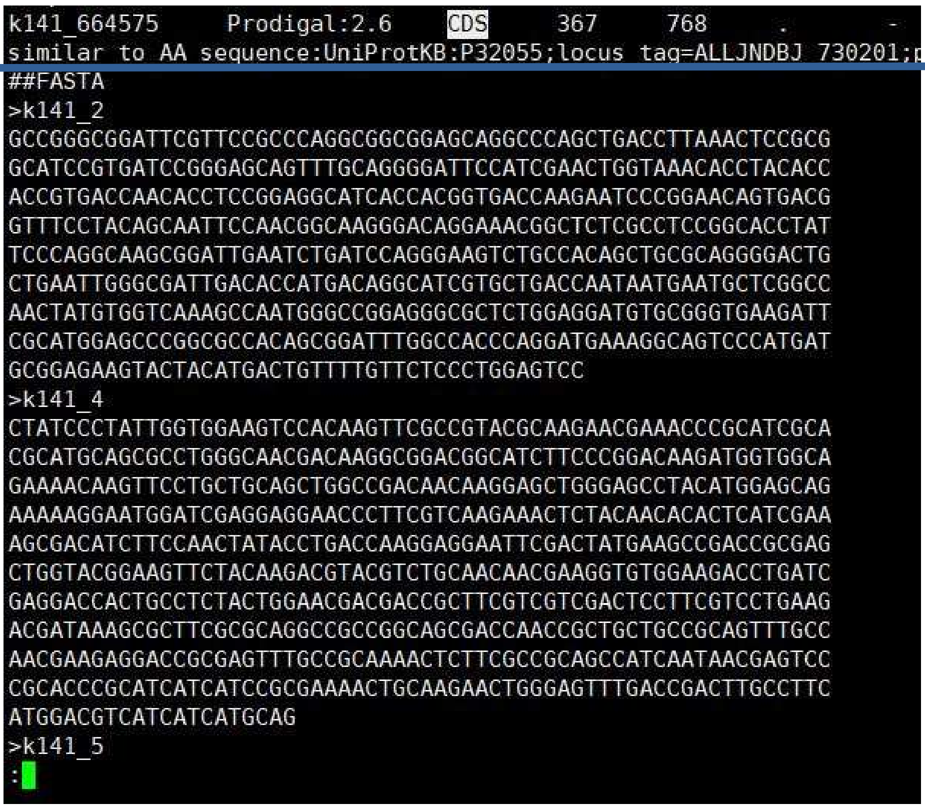
필요한 영역으로 바로 이동
(Contig-ORF)

Contig name

ORF name



```
k141_2  Prodigal:2.6    CDS   49    477   .   -   0   ID=ALLJNDBJ_00001;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00001;product=hypothetical protein
k141_4  Prodigal:2.6    CDS   7     255   .   -   0   ID=ALLJNDBJ_00002;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00002;product=hypothetical protein
k141_4  Prodigal:2.6    CDS   256   459   .   -   0   ID=ALLJNDBJ_00003;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00003;product=hypothetical protein
k141_5  Prodigal:2.6    CDS   136   465   .   +   0   ID=ALLJNDBJ_00004;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00004;product=hypothetical protein
k141_9  Prodigal:2.6    CDS   42    296   .   +   0   ID=ALLJNDBJ_00005;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00005;product=hypothetical protein
k141_12 Prodigal:2.6    CDS   250   555   .   -   0   ID=ALLJNDBJ_00006;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00006;product=hypothetical protein
k141_15 Prodigal:2.6    CDS   46    216   .   +   0   ID=ALLJNDBJ_00007;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00007;product=hypothetical protein
k141_16 Prodigal:2.6    CDS   148   723   .   -   0   ID=ALLJNDBJ_00008;ame=araC_1;gene=araC_1;inference=ab initio prediction:Prodigal:2.6,similar to AA sequence:UniPro
tKB:P0A9E0;locus_tag=A LJNDBJ_00008;product=Arabinose operon regulator protein
k141_17 Prodigal:2.6    CDS   347   469   .   -   0   ID=ALLJNDBJ_00009;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00009;product=hypothetical protein
k141_18 Prodigal:2.6    CDS   7     834   .   +   0   ID=ALLJNDBJ_00010;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00010;product=hypothetical protein
k141_30 Prodigal:2.6    CDS   56    964   .   +   0   ID=ALLJNDBJ_00011;nference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00011;product=hypothetical protein
k141_36 Prodigal:2.6    CDS   282   833   .   -   0   ID=ALLJNDBJ_00012;C_number=2.8.1.13;Name=mnmA_1;gene=mnmA_1;inference=ab initio prediction:Prodigal:2.6,similar to
 AA sequence:UniProtKB:P25745;locus_tag=ALLJNDBJ_00012;product=tRNA-specific 2-thiouridylase MnmA
```



```
k141_664575     Prodigal:2.6    CDS     367     768     .       -
similar to AA sequence:UniProtKB:P32055;locus_tag=ALLJNDBJ_730201;p
##FASTA
>k141_2
GCCGGGCGGATTCGTTCCGCCCAGGCGGCGGAGCAGGCCCAGCTGACCTTAAACTCCGCG
GCATCCGTGATCCGGGAGCAGTTTGCAGGGGATTCCATCGAACTGGTAAACACCTACACC
ACCGTGACCAACACCTCCGGAGGCATCACCACGGTGACCAAGAATCCCGGAACAGTGACG
GTTTCCTACAGCAATTCCAACGGCAAGGGACAGGAAACGGCTCTCGCCTCCGGCACCTAT
TCCCAGGCAAGCGGATTGAATCTGATCCAGGGAAGTCTGCCACAGCTGCGCAGGGGACTG
CTGAATTGGGCGATTGACACCATGACAGGCATCGTGCTGACCAATAATGAATGCTCGGCC
AACTATGTGGTCAAAGCCAATGGGCCGGAGGGCGCTCTGGAGGATGTGCGGGTGAAGATT
CGCATGGAGCCCGGCGCCACAGCGGATTTGGCCACCCAGGATGAAAGGCAGTCCCATGAT
GCGGAGAAGTACTACATGACTGTTTTGTTCTCCCTGGAGTCC
>k141_4
CTATCCCTATTGGTGGAAGTCCACAAGTTCGCCGTACGCAAGAACGAAACCCGCATCGCA
CGCATGCAGCGCCTGGGCAACGACAAGGCGGACGGCATCTTCCCGGACAAGATGGTGGCA
GAAAACAAGTTCCTGCTGCAGCTGGCCGACAACAAGGAGCTGGGAGCCTACATGGAGCAG
AAAAAGGAATGGATCGAGGAGGAACCCTTCGTCAAGAAACTCTACAACACACTCATCGAA
AGCGACATCTTCCAACTATACCTGACCAAGGAGGAATTCGACTATGAAGCCGACCGCGAG
CTGGTACGGAAGTTCTACAAGACGTACGTCTGCAACAACGAAGGTGTGGAAGACCTGATC
GAGGACCACTGCCTCTACTGGAACGACGACCGCTTCGTCGTCGACTCCTTCGTCCTGAAG
ACGATAAAGCGCTTCGCGCAGGCCGCCGGCAGCGACCAACCGCTGCTGCCGCAGTTTGCC
AACGAAGAGGACCGCGAGTTTGCCGCAAAACTCTTCGCCGCAGCCATCAATAACGAGTCC
CGCACCCGCATCATCATCCGCGAAAACTGCAAGAACTGGGAGTTTGACCGACTTGCCTTC
ATGGACGTCATCATCATGCAG
>k141_5
:
```

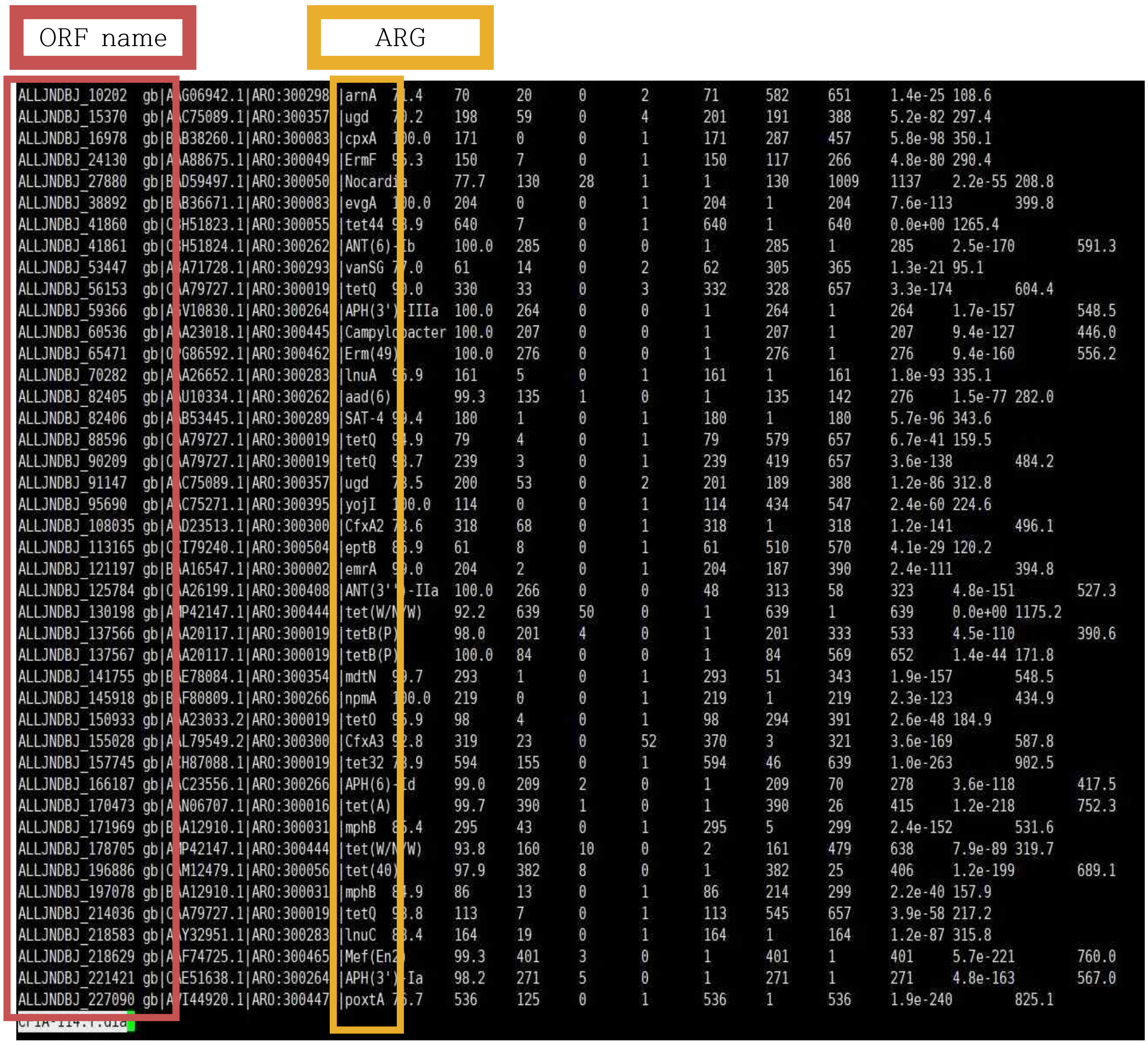
##FASTA
이후에는 필요 없는 것

즉 전체에서
##sequence
필요한 것
##FASTA

* 하나의 contig안에 여러개의 ORF 있을 수 있음
* /CDS, /##FASTA처럼 입력 해 원하는 위치로 이동

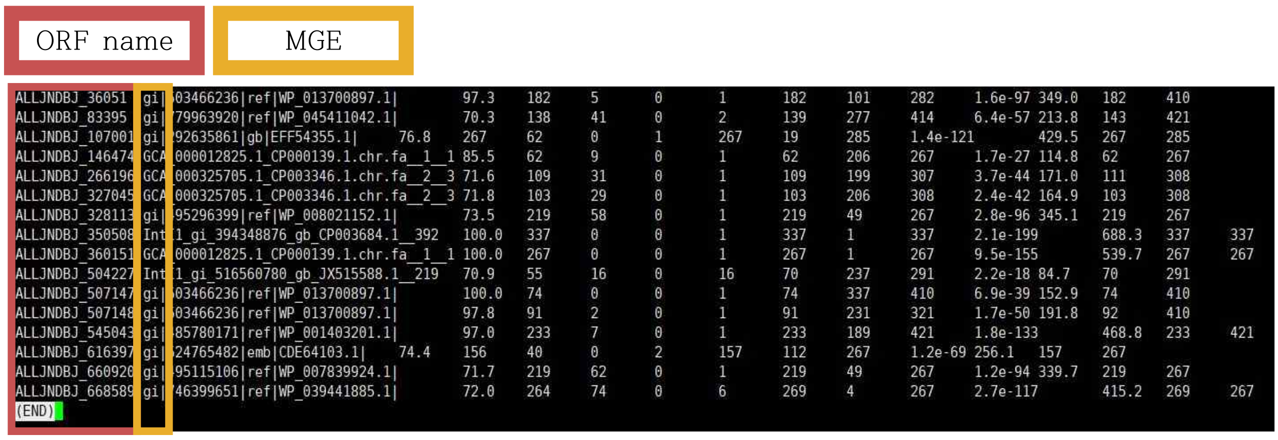
** 전체에서 ##sequence ~ ##FASTA 사이 contig – orf name 나온 것만 필요함

(2) CF1A-114.f.dia -> ORF & ARG 파일



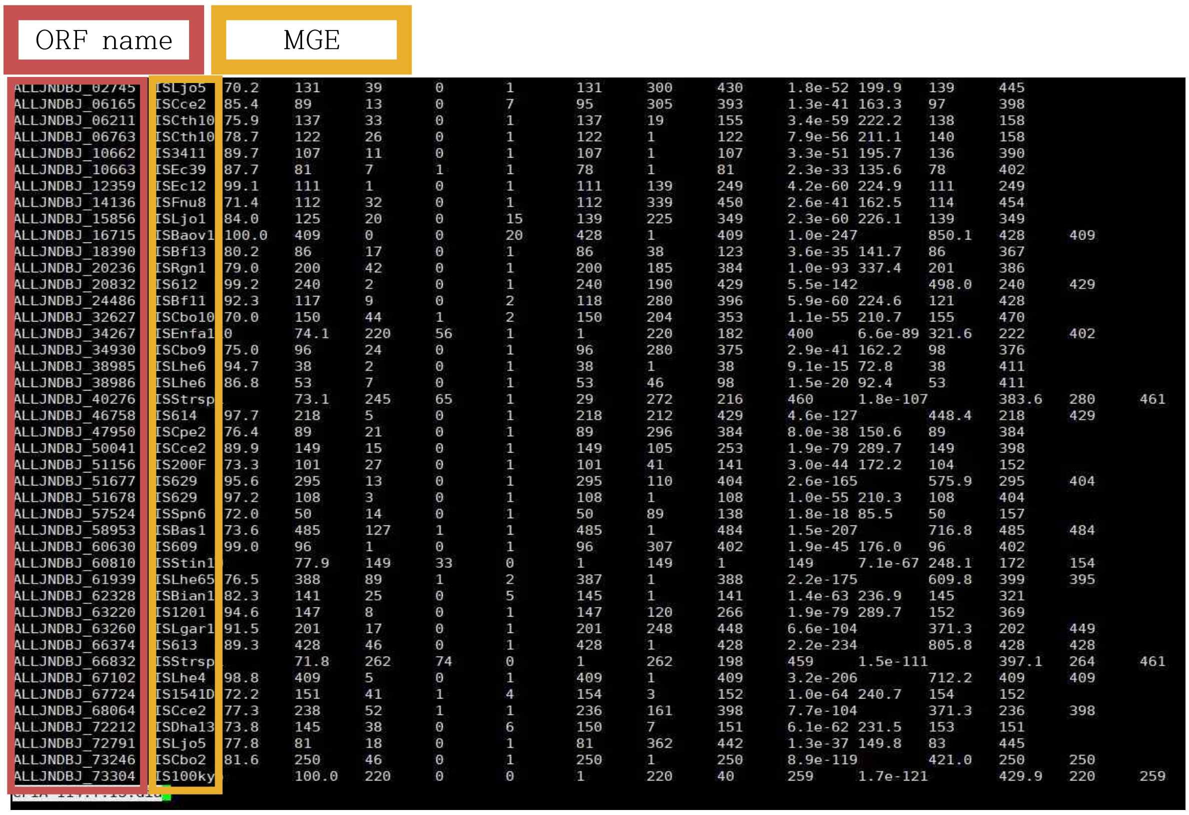
```
ALLJNDBJ_10202  gb|A G06942.1|ARO:300298 |arnA  71.4    70      20      0       2       71      582     651     1.4e-25 108.6
ALLJNDBJ_15370  gb|A C75089.1|ARO:300357 |ugd   7 .2    198     59      0       4       201     191     388     5.2e-82 297.4
ALLJNDBJ_16978  gb|B B38260.1|ARO:300083 |cpxA  1 0.0   171     0       0       1       171     287     457     5.8e-98 350.1
ALLJNDBJ_24130  gb|A A88675.1|ARO:300049 |ErmF  9 .3    150     7       0       1       150     117     266     4.8e-80 290.4
ALLJNDBJ_27880  gb|B D59497.1|ARO:300050 |Nocardi       77.7    130     28      1       1       130     1009    1137    2.2e-55 208.8
ALLJNDBJ_38892  gb|B B36671.1|ARO:300083 |evgA  1 0.0   204     0       0       1       204     1       204     7.6e-113        399.8
ALLJNDBJ_41860  gb|C H51823.1|ARO:300055 |tet44 9 .9    640     7       0       1       640     1       640     0.0e+00 1265.4
ALLJNDBJ_41861  gb|C H51824.1|ARO:300262 |ANT(6)-Ib     100.0   285     0       0       1       285     1       285     2.5e-170        591.3
ALLJNDBJ_53447  gb|A A71728.1|ARO:300293 |vanSG 7 .0    61      14      0       2       62      305     365     1.3e-21 95.1
ALLJNDBJ_56153  gb|C A79727.1|ARO:300019 |tetQ  9 .0    330     33      0       3       332     328     657     3.3e-174        604.4
ALLJNDBJ_59366  gb|A V10830.1|ARO:300264 |APH(3') IIIa  100.0   264     0       0       1       264     1       264     1.7e-157        548.5
ALLJNDBJ_60536  gb|A A23018.1|ARO:300445 |Campyl bacter 100.0   207     0       0       1       207     1       207     9.4e-127        446.0
ALLJNDBJ_65471  gb|0 G86592.1|ARO:300462 |Erm(49)       100.0   276     0       0       1       276     1       276     9.4e-160        556.2
ALLJNDBJ_70282  gb|A A26652.1|ARO:300283 |lnuA  9 .9    161     5       0       1       161     1       161     1.8e-93 335.1
ALLJNDBJ_82405  gb|A U10334.1|ARO:300262 |aad(6)        99.3    135     1       0       1       135     142     276     1.5e-77 282.0
ALLJNDBJ_82406  gb|A B53445.1|ARO:300289 |SAT-4 9 .4    180     1       0       1       180     1       180     5.7e-96 343.6
ALLJNDBJ_88596  gb|C A79727.1|ARO:300019 |tetQ  9 .9    79      4       0       1       79      579     657     6.7e-41 159.5
ALLJNDBJ_90209  gb|C A79727.1|ARO:300019 |tetQ  9 .7    239     3       0       1       239     419     657     3.6e-138        484.2
ALLJNDBJ_91147  gb|A C75089.1|ARO:300357 |ugd   7 .5    200     53      0       2       201     189     388     1.2e-86 312.8
ALLJNDBJ_95690  gb|A C75271.1|ARO:300395 |yojI  1 0.0   114     0       0       1       114     434     547     2.4e-60 224.6
ALLJNDBJ_108035 gb|A D23513.1|ARO:300300 |CfxA2 7 .6    318     68      0       1       318     1       318     1.2e-141        496.1
ALLJNDBJ_113165 gb|C I79240.1|ARO:300504 |eptB  8 .9    61      8       0       1       61      510     570     4.1e-29 120.2
ALLJNDBJ_121197 gb|B A16547.1|ARO:300002 |emrA  9 .0    204     2       0       1       204     187     390     2.4e-111        394.8
ALLJNDBJ_125784 gb|C A26199.1|ARO:300408 |ANT(3'' -IIa  100.0   266     0       0       48      313     58      323     4.8e-151        527.3
ALLJNDBJ_130198 gb|A P42147.1|ARO:300444 |tet(W/N W)    92.2    639     50      0       1       639     1       639     0.0e+00 1175.2
ALLJNDBJ_137566 gb|A A20117.1|ARO:300019 |tetB(P)       98.0    201     4       0       1       201     333     533     4.5e-110        390.6
ALLJNDBJ_137567 gb|A A20117.1|ARO:300019 |tetB(P)       100.0   84      0       0       1       84      569     652     1.4e-44 171.8
ALLJNDBJ_141755 gb|B E78084.1|ARO:300354 |mdtN  9 .7    293     1       0       1       293     51      343     1.9e-157        548.5
ALLJNDBJ_145918 gb|B F80809.1|ARO:300266 |npmA  1 0.0   219     0       0       1       219     1       219     2.3e-123        434.9
ALLJNDBJ_150933 gb|A A23033.2|ARO:300019 |tetO  9 .9    98      4       0       1       98      294     391     2.6e-48 184.9
ALLJNDBJ_155028 gb|A L79549.2|ARO:300300 |CfxA3 9 .8    319     23      0       52      370     3       321     3.6e-169        587.8
ALLJNDBJ_157745 gb|A H87088.1|ARO:300019 |tet32 7 .9    594     155     0       1       594     46      639     1.0e-263        902.5
ALLJNDBJ_166187 gb|A C23556.1|ARO:300266 |APH(6) Id     99.0    209     2       0       1       209     70      278     3.6e-118        417.5
ALLJNDBJ_170473 gb|A N06707.1|ARO:300016 |tet(A)        99.7    390     1       0       1       390     26      415     1.2e-218        752.3
ALLJNDBJ_171969 gb|B A12910.1|ARO:300031 |mphB  8 .4    295     43      0       1       295     5       299     2.4e-152        531.6
ALLJNDBJ_178705 gb|A P42147.1|ARO:300444 |tet(W/N W)    93.8    160     10      0       2       161     479     638     7.9e-89 319.7
ALLJNDBJ_196886 gb|C M12479.1|ARO:300056 |tet(40)       97.9    382     8       0       1       382     25      406     1.2e-199        689.1
ALLJNDBJ_197078 gb|B A12910.1|ARO:300031 |mphB  8 .9    86      13      0       1       86      214     299     2.2e-40 157.9
ALLJNDBJ_214036 gb|C A79727.1|ARO:300019 |tetQ  9 .8    113     7       0       1       113     545     657     3.9e-58 217.2
ALLJNDBJ_218583 gb|A Y32951.1|ARO:300283 |lnuC  8 .4    164     19      0       1       164     1       164     1.2e-87 315.8
ALLJNDBJ_218629 gb|A F74725.1|ARO:300465 |Mef(En2)      99.3    401     3       0       1       401     1       401     5.7e-221        760.0
ALLJNDBJ_221421 gb|C E51638.1|ARO:300264 |APH(3') Ia    98.2    271     5       0       1       271     1       271     4.8e-163        567.0
ALLJNDBJ_227090 gb|A I44920.1|ARO:300447 |poxtA 7 .7    536     125     0       1       536     1       536     1.9e-240        825.1
CF1A-114.f.dia
```

(3) CF1A-114.f.Int.dia -> ORF & MGE(Int)



```
ALLJNDBJ_36051  gi| 03466236|ref|WP_013700897.1|        97.3    182     5       0       1       182     101     282     1.6e-97 349.0   182     410
ALLJNDBJ_83395  gi| 79963920|ref|WP_045411042.1|        70.3    138     41      0       2       139     277     414     6.4e-57 213.8   143     421
ALLJNDBJ_107001 gi| 92635861|gb|EFF54355.1|     76.8    267     62      0       1       267     19      285     1.4e-121        429.5   267     285
ALLJNDBJ_146474 GCA 000012825.1_CP000139.1.chr.fa__1__1 85.5    62      9       0       1       62      206     267     1.7e-27 114.8   62      267
ALLJNDBJ_266196 GCA 000325705.1_CP003346.1.chr.fa__2__3 71.6    109     31      0       1       109     199     307     3.7e-44 171.0   111     308
ALLJNDBJ_327045 GCA 000325705.1_CP003346.1.chr.fa__2__3 71.8    103     29      0       1       103     206     308     2.4e-42 164.9   103     308
ALLJNDBJ_328113 gi| 95296399|ref|WP_008021152.1|        73.5    219     58      0       1       219     49      267     2.8e-96 345.1   219     267
ALLJNDBJ_350508 Int 1_gi_394348876_gb_CP003684.1__392   100.0   337     0       0       1       337     1       337     2.1e-199        688.3   337     337
ALLJNDBJ_360151 GCA 000012825.1_CP000139.1.chr.fa__1__1 100.0   267     0       0       1       267     1       267     9.5e-155        539.7   267     267
ALLJNDBJ_504227 Int 1_gi_516560780_gb_JX515588.1__219   70.9    55      16      0       16      70      237     291     2.2e-18 84.7    70      291
ALLJNDBJ_507147 gi| 03466236|ref|WP_013700897.1|        100.0   74      0       0       1       74      337     410     6.9e-39 152.9   74      410
ALLJNDBJ_507148 gi| 03466236|ref|WP_013700897.1|        97.8    91      2       0       1       91      231     321     1.7e-50 191.8   92      410
ALLJNDBJ_545043 gi| 85780171|ref|WP_001403201.1|        97.0    233     7       0       1       233     189     421     1.8e-133        468.8   233     421
ALLJNDBJ_616397 gi| 24765482|emb|CDE64103.1|    74.4    156     40      0       2       157     112     267     1.2e-69 256.1   157     267
ALLJNDBJ_660920 gi| 95115106|ref|WP_007839924.1|        71.7    219     62      0       1       219     49      267     1.2e-94 339.7   219     267
ALLJNDBJ_668589 gi| 46399651|ref|WP_039441885.1|        72.0    264     74      0       6       269     4       267     2.7e-117        415.2   269     267
(END)
```

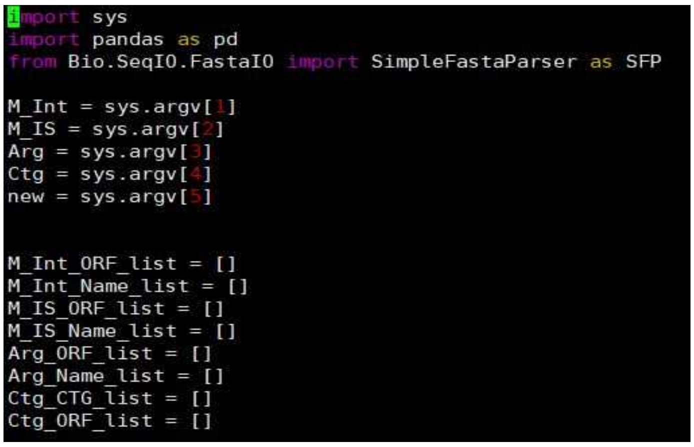
(4) CF1A-114.f.IS.dia -> ORF & MGE(IS)



```
ALLJNDBJ_02745  ISLjo5  70.2    131     39      0       1       131     300     430     1.8e-52 199.9   139     445
ALLJNDBJ_06165  ISCce2  85.4    89      13      0       7       95      305     393     1.3e-41 163.3   97      398
ALLJNDBJ_06211  ISCth10 75.9    137     33      0       1       137     19      155     3.4e-59 222.2   138     158
ALLJNDBJ_06763  ISCth10 78.7    122     26      0       1       122     1       122     7.9e-56 211.1   140     158
ALLJNDBJ_10662  IS3411  89.7    107     11      0       1       107     1       107     3.3e-51 195.7   136     390
ALLJNDBJ_10663  ISEc39  87.7    81      7       1       1       78      1       81      2.3e-33 135.6   78      402
ALLJNDBJ_12359  ISEc12  99.1    111     1       0       1       111     139     249     4.2e-60 224.9   111     249
ALLJNDBJ_14136  ISFnu8  71.4    112     32      0       1       112     339     450     2.6e-41 162.5   114     454
ALLJNDBJ_15856  ISLjo1  84.0    125     20      0       15      139     225     349     2.3e-60 226.1   139     349
ALLJNDBJ_16715  ISBaov1 100.0   409     0       0       20      428     1       409     1.0e-247        850.1   428     409
ALLJNDBJ_18390  ISBf13  80.2    86      17      0       1       86      38      123     3.6e-35 141.7   86      367
ALLJNDBJ_20236  ISRgn1  79.0    200     42      0       1       200     185     384     1.0e-93 337.4   201     386
ALLJNDBJ_20832  IS612   99.2    240     2       0       1       240     190     429     5.5e-142        498.0   240     429
ALLJNDBJ_24486  ISBf11  92.3    117     9       0       2       118     280     396     5.9e-60 224.6   121     428
ALLJNDBJ_32627  ISCbo10 70.0    150     44      1       2       150     204     353     1.1e-55 210.7   155     470
ALLJNDBJ_34267  ISEnfa1 0       74.1    220     56      1       1       220     182     400     6.6e-89 321.6   222     402
ALLJNDBJ_34930  ISCbo9  75.0    96      24      0       1       96      280     375     2.9e-41 162.2   98      376
ALLJNDBJ_38985  ISLhe6  94.7    38      2       0       1       38      1       38      9.1e-15 72.8    38      411
ALLJNDBJ_38986  ISLhe6  86.8    53      7       0       1       53      46      98      1.5e-20 92.4    53      411
ALLJNDBJ_40276  ISStrsp         73.1    245     65      1       29      272     216     460     1.8e-107        383.6   280     461
ALLJNDBJ_46758  IS614   97.7    218     5       0       1       218     212     429     4.6e-127        448.4   218     429
ALLJNDBJ_47950  ISCpe2  76.4    89      21      0       1       89      296     384     8.0e-38 150.6   89      384
ALLJNDBJ_50041  ISCce2  89.9    149     15      0       1       149     105     253     1.9e-79 289.7   149     398
ALLJNDBJ_51156  IS200F  73.3    101     27      0       1       101     41      141     3.0e-44 172.2   104     152
ALLJNDBJ_51677  IS629   95.6    295     13      0       1       295     110     404     2.6e-165        575.9   295     404
ALLJNDBJ_51678  IS629   97.2    108     3       0       1       108     1       108     1.0e-55 210.3   108     404
ALLJNDBJ_57524  ISSpn6  72.0    50      14      0       1       50      89      138     1.8e-18 85.5    50      157
ALLJNDBJ_58953  ISBas1  73.6    485     127     1       1       485     1       484     1.5e-207        716.8   485     484
ALLJNDBJ_60630  IS609   99.0    96      1       0       1       96      307     402     1.9e-45 176.0   96      402
ALLJNDBJ_60810  ISStin1         77.9    149     33      0       1       149     1       149     7.1e-67 248.1   172     154
ALLJNDBJ_61939  ISLhe65 76.5    388     89      1       2       387     1       388     2.2e-175        609.8   399     395
ALLJNDBJ_62328  ISBian1 82.3    141     25      0       5       145     1       141     1.4e-63 236.9   145     321
ALLJNDBJ_63220  IS1201  94.6    147     8       0       1       147     120     266     1.9e-79 289.7   152     369
ALLJNDBJ_63260  ISLgar1 91.5    201     17      0       1       201     248     448     6.6e-104        371.3   202     449
ALLJNDBJ_66374  IS613   89.3    428     46      0       1       428     1       428     2.2e-234        805.8   428     428
ALLJNDBJ_66832  ISStrsp         71.8    262     74      0       1       262     198     459     1.5e-111        397.1   264     461
ALLJNDBJ_67102  ISLhe4  98.8    409     5       0       1       409     1       409     3.2e-206        712.2   409     409
ALLJNDBJ_67724  IS1541D 72.2    151     41      1       4       154     3       152     1.0e-64 240.7   154     152
ALLJNDBJ_68064  ISCce2  77.3    238     52      1       1       236     161     398     7.7e-104        371.3   236     398
ALLJNDBJ_72212  ISDha13 73.8    145     38      0       6       150     7       151     6.1e-62 231.5   153     151
ALLJNDBJ_72791  ISLjo5  77.8    81      18      0       1       81      362     442     1.3e-37 149.8   83      445
ALLJNDBJ_73246  ISCbo2  81.6    250     46      0       1       250     1       250     8.9e-119        421.0   250     250
ALLJNDBJ_73304  IS100ky         100.0   220     0       0       1       220     40      259     1.7e-121        429.9   220     259
```

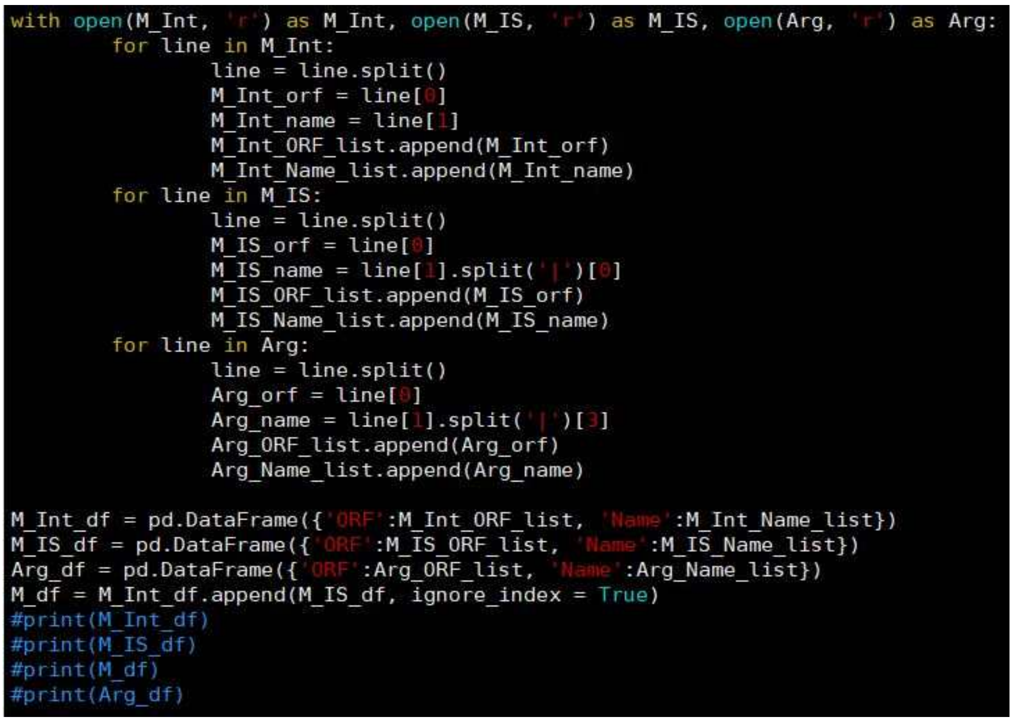
3) test01.py -> 추출

(1) import, 필요한 자료 구조 생성



```python
import sys
import pandas as pd
from Bio.SeqIO.FastaIO import SimpleFastaParser as SFP

M_Int = sys.argv[1]
M_IS = sys.argv[2]
Arg = sys.argv[3]
Ctg = sys.argv[4]
new = sys.argv[5]


M_Int_ORF_list = []
M_Int_Name_list = []
M_IS_ORF_list = []
M_IS_Name_list = []
Arg_ORF_list = []
Arg_Name_list = []
Ctg_CTG_list = []
Ctg_ORF_list = []
```

(2) MGE, ARG에서 필요한 부분만 받아오기



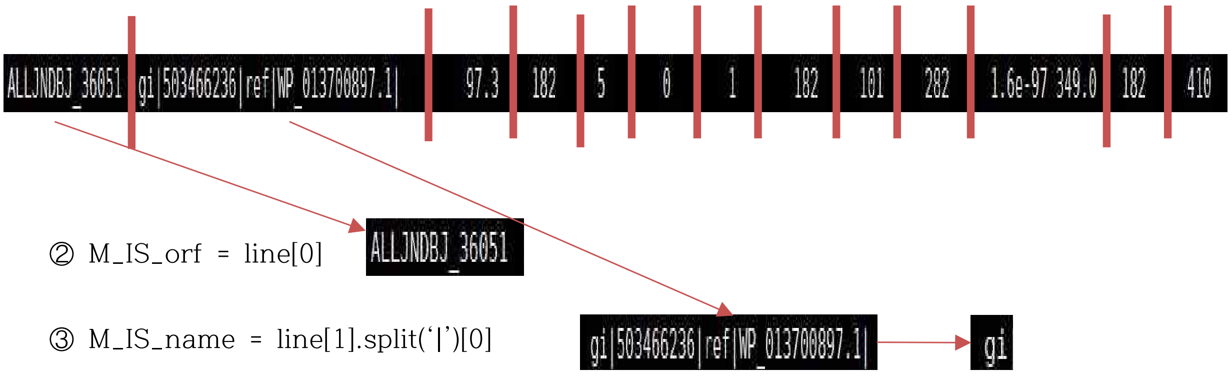
```
with open(M_Int, 'r') as M_Int, open(M_IS, 'r') as M_IS, open(Arg, 'r') as Arg:
        for line in M_Int:
                line = line.split()
                M_Int_orf = line[0]
                M_Int_name = line[1]
                M_Int_ORF_list.append(M_Int_orf)
                M_Int_Name_list.append(M_Int_name)
        for line in M_IS:
                line = line.split()
                M_IS_orf = line[0]
                M_IS_name = line[1].split('|')[0]
                M_IS_ORF_list.append(M_IS_orf)
                M_IS_Name_list.append(M_IS_name)
        for line in Arg:
                line = line.split()
                Arg_orf = line[0]
                Arg_name = line[1].split('|')[3]
                Arg_ORF_list.append(Arg_orf)
                Arg_Name_list.append(Arg_name)

M_Int_df = pd.DataFrame({'ORF':M_Int_ORF_list, 'Name':M_Int_Name_list})
M_IS_df = pd.DataFrame({'ORF':M_IS_ORF_list, 'Name':M_IS_Name_list})
Arg_df = pd.DataFrame({'ORF':Arg_ORF_list, 'Name':Arg_Name_list})
M_df = M_Int_df.append(M_IS_df, ignore_index = True)
#print(M_Int_df)
#print(M_IS_df)
#print(M_df)
#print(Arg_df)
```

필요한 ORF, GENE NAME 만 받기

Dataframe에 옮기기

ex) M_IS (**M_Int 파일이어야 하는데... 잘못 적은 듯..)

① line = line.split() -> 라인마다 공백으로 split



ALLJNDBJ_36051 gi|503466236|ref|WP_013700897.1| 97.3 182 5 0 1 182 101 282 1.6e-97 349.0 182 410
ALLJNDBJ_36051
gi|503466236|ref|WP_013700897.1|
gi

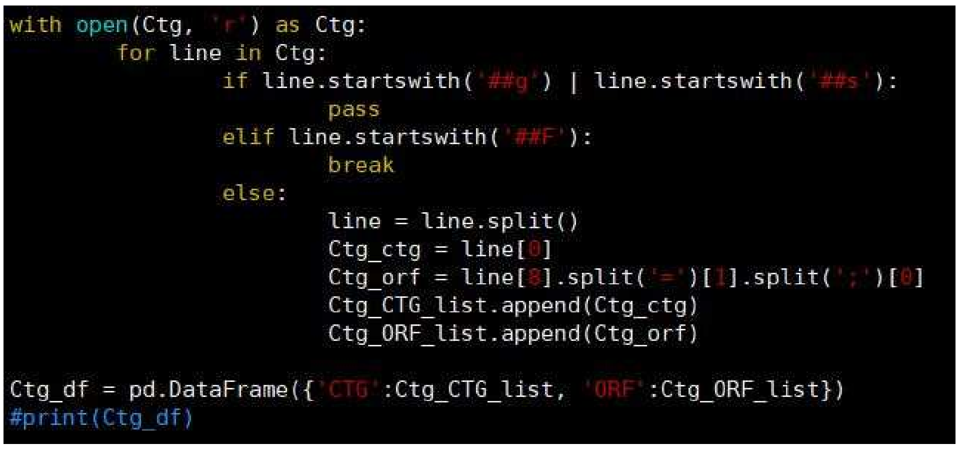
② M_IS_orf = line[0]

③ M_IS_name = line[1].split('|')[0]

④ M_IS_ORF_list.append(M_IS_orf) => ORF 어펜드

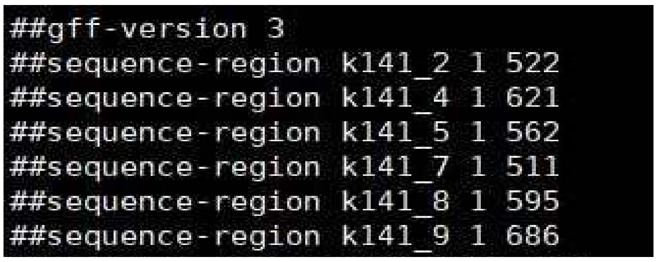
⑤ M_IS_NAME_list.append(M_IS_name) => MGE name 어펜드

(3) Contig-ORF 추출



```
with open(Ctg, 'r') as Ctg:
        for line in Ctg:
                if line.startswith('##g') | line.startswith('##s'):
                        pass
                elif line.startswith('##F'):
                        break
                else:
                        line = line.split()
                        Ctg_ctg = line[0]
                        Ctg_orf = line[8].split('=')[1].split(';')[0]
                        Ctg_CTG_list.append(Ctg_ctg)
                        Ctg_ORF_list.append(Ctg_orf)

Ctg_df = pd.DataFrame({'CTG':Ctg_CTG_list, 'ORF':Ctg_ORF_list})
#print(Ctg_df)
```

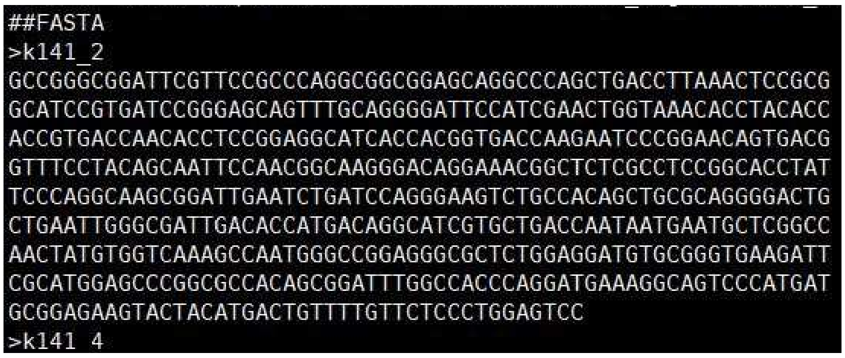
① '##g', '##s'로 시작하면 pass



```
##gff-version 3
##sequence-region k141_2 1 522
##sequence-region k141_4 1 621
##sequence-region k141_5 1 562
##sequence-region k141_7 1 511
##sequence-region k141_8 1 595
##sequence-region k141_9 1 686
```

첫 부분(필요 x)
pass

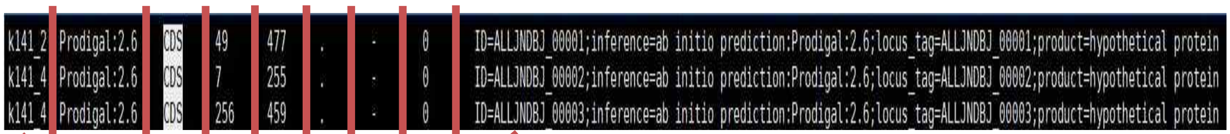
② '##F'로 시작하면 break



```
##FASTA
>k141_2
GCCGGGCGGATTCGTTCCGCCCAGGCGGCGGAGCAGGCCCAGCTGACCTTAAACTCCGCG
GCATCCGTGATCCGGGAGCAGTTTGCAGGGGATTCCATCGAACTGGTAAACACCTACACC
ACCGTGACCAACACCTCCGGAGGCATCACCACGGTGACCAAGAATCCCGGAACAGTGACG
GTTTCCTACAGCAATTCCAACGGCAAGGGACAGGAAACGGCTCTCGCCTCCGGCACCTAT
TCCCAGGCAAGCGGATTGAATCTGATCCAGGGAAGTCTGCCACAGCTGCGCAGGGGACTG
CTGAATTGGGCGATTGACACCATGACAGGCATCGTGCTGACCAATAATGAATGCTCGGCC
AACTATGTGGTCAAAGCCAATGGGCCGGAGGGCGCTCTGGAGGATGTGCGGGTGAAGATT
CGCATGGAGCCCGGCGCCACAGCGGATTTGGCCACCCAGGATGAAAGGCAGTCCCATGAT
GCGGAGAAGTACTACATGACTGTTTTGTTCTCCCTGGAGTCC
>k141_4
```

끝 부분(필요 x)
더 이상 읽을 필요 x
break

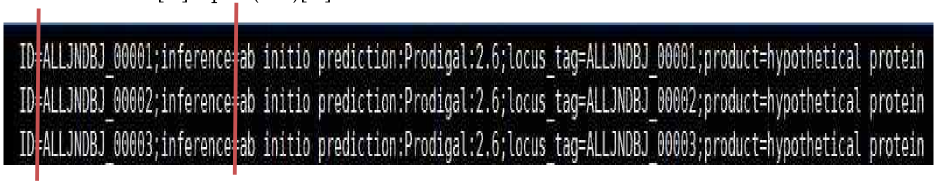
③ line = line.split(), Ctg_ctg = line[0], Ctg_orf = line[8].split('=')[1].split(';')[0]



```
k141_2 Prodigal:2.6 CDS 49  477 . - 0 ID=ALLJNDBJ_00001;inference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00001;product=hypothetical protein
k141_4 Prodigal:2.6 CDS 7   255 . - 0 ID=ALLJNDBJ_00002;inference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00002;product=hypothetical protein
k141_4 Prodigal:2.6 CDS 256 459 . - 0 ID=ALLJNDBJ_00003;inference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00003;product=hypothetical protein
```

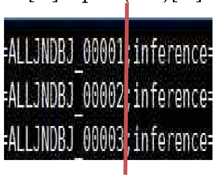
Ctg_ctg = line[0]   Ctg_orf = line[8]

line[8].split('=')[1]

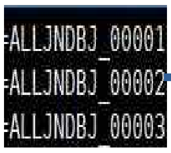


```
ID=ALLJNDBJ_00001;inference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00001;product=hypothetical protein
ID=ALLJNDBJ_00002;inference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00002;product=hypothetical protein
ID=ALLJNDBJ_00003;inference=ab initio prediction:Prodigal:2.6;locus_tag=ALLJNDBJ_00003;product=hypothetical protein
```

line[8].split('=')[1].split(';')[0]



```
=ALLJNDBJ_00001;inference=
=ALLJNDBJ_00002;inference=
=ALLJNDBJ_00003;inference=
```



```
=ALLJNDBJ_00001
=ALLJNDBJ_00002
=ALLJNDBJ_00003
```

결과물
->ORF만 추출

(4) merge

① 현재까지 생성된 데이터 프레임 3개

M_df

ORF	Name
.	.
.	.
.	.

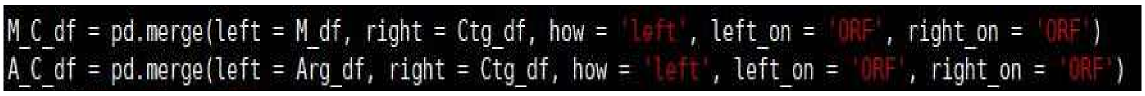
Arg_df

ORF	Name
.	.
.	.
.	.

Ctg_df

CTG	ORF
.	.
.	.
.	.

② 데이터 프레임 합치기



```
M_C_df = pd.merge(left = M_df, right = Ctg_df, how = 'left', left_on = 'ORF', right_on = 'ORF')
A_C_df = pd.merge(left = Arg_df, right = Ctg_df, how = 'left', left_on = 'ORF', right_on = 'ORF')
```

=> M_df + Ctg_df = M_C_df

ORF	Name
.	.
.	.
.	.

+

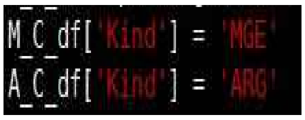
CTG	ORF
.	.
.	.
.	.

=

M	ORF	CTG
.	.	.
.	.	.
.	.	.

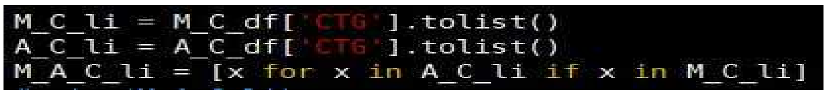
how = ‘left’ -> 왼쪽 데이터 프레임(M_df)을 기준으로 병합
** 두 데이터 프레임은 ORF열을 통해 병합됨
** 왼쪽이 기준으로 MGE가 있는 ORF에 해당하는 contig만 알아냄

③ Kind 적기



```
M_C_df['Kind'] = 'MGE'
A_C_df['Kind'] = 'ARG'
```

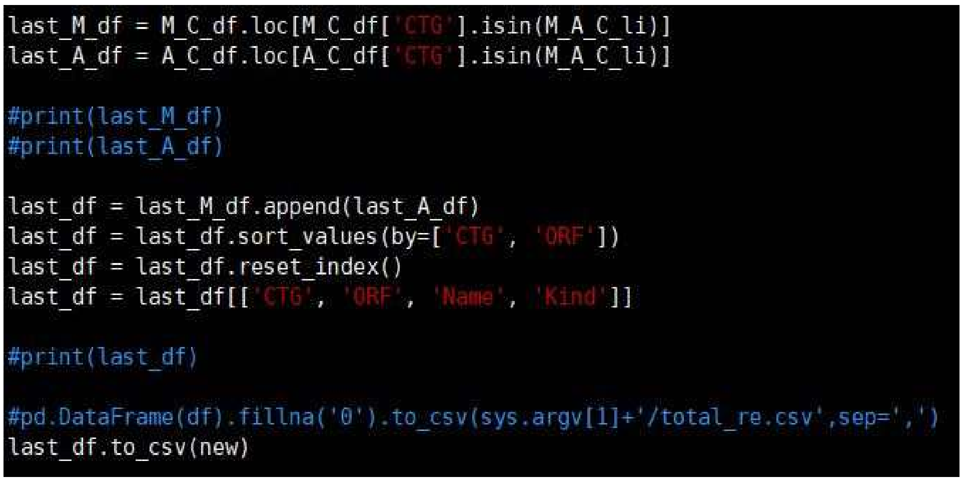
(4) list comprehension



```
M_C_li = M_C_df['CTG'].tolist()
A_C_li = A_C_df['CTG'].tolist()
M_A_C_li = [x for x in A_C_li if x in M_C_li]
```

M_A_C_li = [x for x in A_C_li if x in M_C_li]
=> 만약 A_C_li의 x가 M_C_li에도 있으면 x를 M_A_C_li 리스트에 삽입한다.
** A_C_li가 M_C_li보다 양이 많기 때문에 A_C_li의 x로 설정

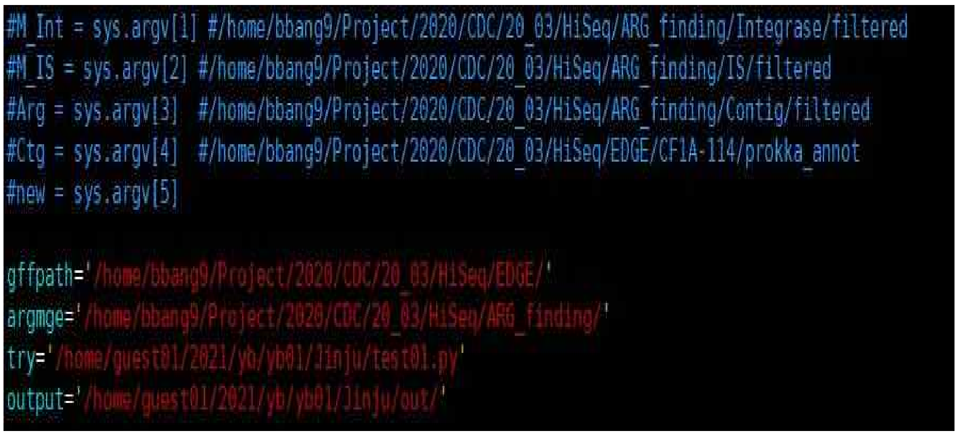
(5) 최종 + 저장



```
last_M_df = M_C_df.loc[M_C_df['CTG'].isin(M_A_C_li)]
last_A_df = A_C_df.loc[A_C_df['CTG'].isin(M_A_C_li)]

#print(last_M_df)
#print(last_A_df)

last_df = last_M_df.append(last_A_df)
last_df = last_df.sort_values(by=['CTG', 'ORF'])
last_df = last_df.reset_index()
last_df = last_df[['CTG', 'ORF', 'Name', 'Kind']]

#print(last_df)

#pd.DataFrame(df).fillna('0').to_csv(sys.argv[1]+'/total_re.csv',sep=',')
last_df.to_csv(new)
```

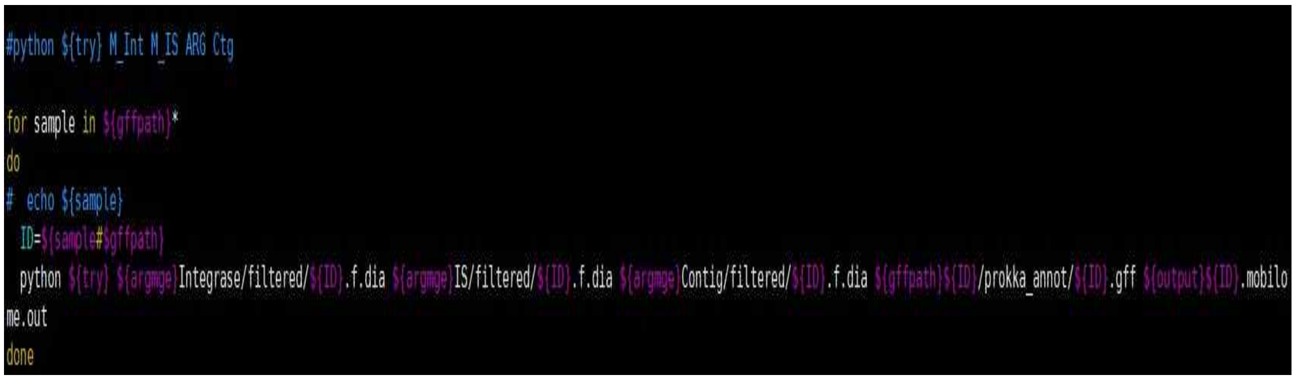
** 참고 : 진주언니가 만든 파일 -> 3inc_ctg.py

4) find_ctg.sh



```
#M_Int = sys.argv[1] #/home/bbang9/Project/2020/CDC/20_03/HiSeq/ARG_finding/Integrase/filtered
#M_IS = sys.argv[2] #/home/bbang9/Project/2020/CDC/20_03/HiSeq/ARG_finding/IS/filtered
#Arg = sys.argv[3]  #/home/bbang9/Project/2020/CDC/20_03/HiSeq/ARG_finding/Contig/filtered
#Ctg = sys.argv[4]  #/home/bbang9/Project/2020/CDC/20_03/HiSeq/EDGE/CF1A-114/prokka_annot
#new = sys.argv[5]

gffpath='/home/bbang9/Project/2020/CDC/20_03/HiSeq/EDGE/'
argmge='/home/bbang9/Project/2020/CDC/20_03/HiSeq/ARG_finding/'
try='/home/guest01/2021/yb/yb01/Jinju/test01.py'
output='/home/guest01/2021/yb/yb01/Jinju/out/'
```

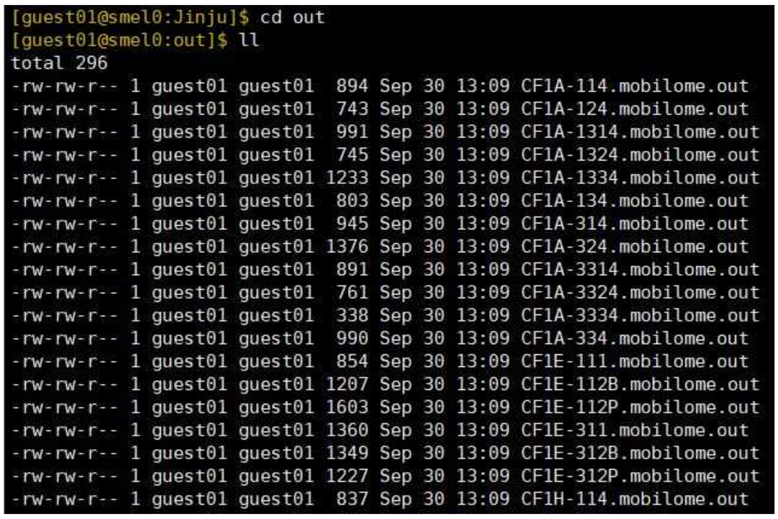
패스 설정



```
#python ${try} M_Int M_IS ARG Ctg

for sample in ${gffpath}*
do
#  echo ${sample}
  ID=${sample#$gffpath}
  python ${try} ${argmge}Integrase/filtered/${ID}.f.dia ${argmge}IS/filtered/${ID}.f.dia ${argmge}Contig/filtered/${ID}.f.dia ${gffpath}${ID}/prokka_annot/${ID}.gff ${output}${ID}.mobilo
me.out
done
```

리눅스 쉘 스크립트
find_ctg.sh

5) output



```
[guest01@smel0:Jinju]$ cd out
[guest01@smel0:out]$ ll
total 296
-rw-rw-r-- 1 guest01 guest01  894 Sep 30 13:09 CF1A-114.mobilome.out
-rw-rw-r-- 1 guest01 guest01  743 Sep 30 13:09 CF1A-124.mobilome.out
-rw-rw-r-- 1 guest01 guest01  991 Sep 30 13:09 CF1A-1314.mobilome.out
-rw-rw-r-- 1 guest01 guest01  745 Sep 30 13:09 CF1A-1324.mobilome.out
-rw-rw-r-- 1 guest01 guest01 1233 Sep 30 13:09 CF1A-1334.mobilome.out
-rw-rw-r-- 1 guest01 guest01  803 Sep 30 13:09 CF1A-134.mobilome.out
-rw-rw-r-- 1 guest01 guest01  945 Sep 30 13:09 CF1A-314.mobilome.out
-rw-rw-r-- 1 guest01 guest01 1376 Sep 30 13:09 CF1A-324.mobilome.out
-rw-rw-r-- 1 guest01 guest01  891 Sep 30 13:09 CF1A-3314.mobilome.out
-rw-rw-r-- 1 guest01 guest01  761 Sep 30 13:09 CF1A-3324.mobilome.out
-rw-rw-r-- 1 guest01 guest01  338 Sep 30 13:09 CF1A-3334.mobilome.out
-rw-rw-r-- 1 guest01 guest01  990 Sep 30 13:09 CF1A-334.mobilome.out
-rw-rw-r-- 1 guest01 guest01  854 Sep 30 13:09 CF1E-111.mobilome.out
-rw-rw-r-- 1 guest01 guest01 1207 Sep 30 13:09 CF1E-112B.mobilome.out
-rw-rw-r-- 1 guest01 guest01 1603 Sep 30 13:09 CF1E-112P.mobilome.out
-rw-rw-r-- 1 guest01 guest01 1360 Sep 30 13:09 CF1E-311.mobilome.out
-rw-rw-r-- 1 guest01 guest01 1349 Sep 30 13:09 CF1E-312B.mobilome.out
-rw-rw-r-- 1 guest01 guest01 1227 Sep 30 13:09 CF1E-312P.mobilome.out
-rw-rw-r-- 1 guest01 guest01  837 Sep 30 13:09 CF1H-114.mobilome.out
```