

# Data mining *Chara vulgaris* (green algae) SRA reads for Cellulose synthase-like (Csl) genes



Tom Bean  
Brenda Pierson  
Steve Seydell  
Bill Wysocki

- The Csl genes encode enzymes involved for synthesis of celluloses and hemicelluloses which are important for biofuel production.
- The cellulose synthase superfamily has been classified into nine Csl families:  
Csl A, Csl B, Csl C, Csl D, Csl E, Csl F, Csl G, Csl H, Csl J  
and one CesA family.
- In 2009, Yin *et al.* identified Csl homologs in fully sequenced lower green algae. To continue his research, we were to expand this search in NGS 454 data available on NCBI for *Chara vulgaris*

# PROCEDURE OUTLINE

binary

*Chara vulgaris* SRA  
SRR099268 (68),  
SRR099269 (69)

Fastq-  
dump

FastQ

Fastq  
To 454

FastA

*C. vulgaris*  
SRA files

BMC Plant Biology  
Csl Protein  
Sequences

TBLASTN

Hits

68 -> 5438

69 -> 5404

~16

minutes

TEASTY

Hits

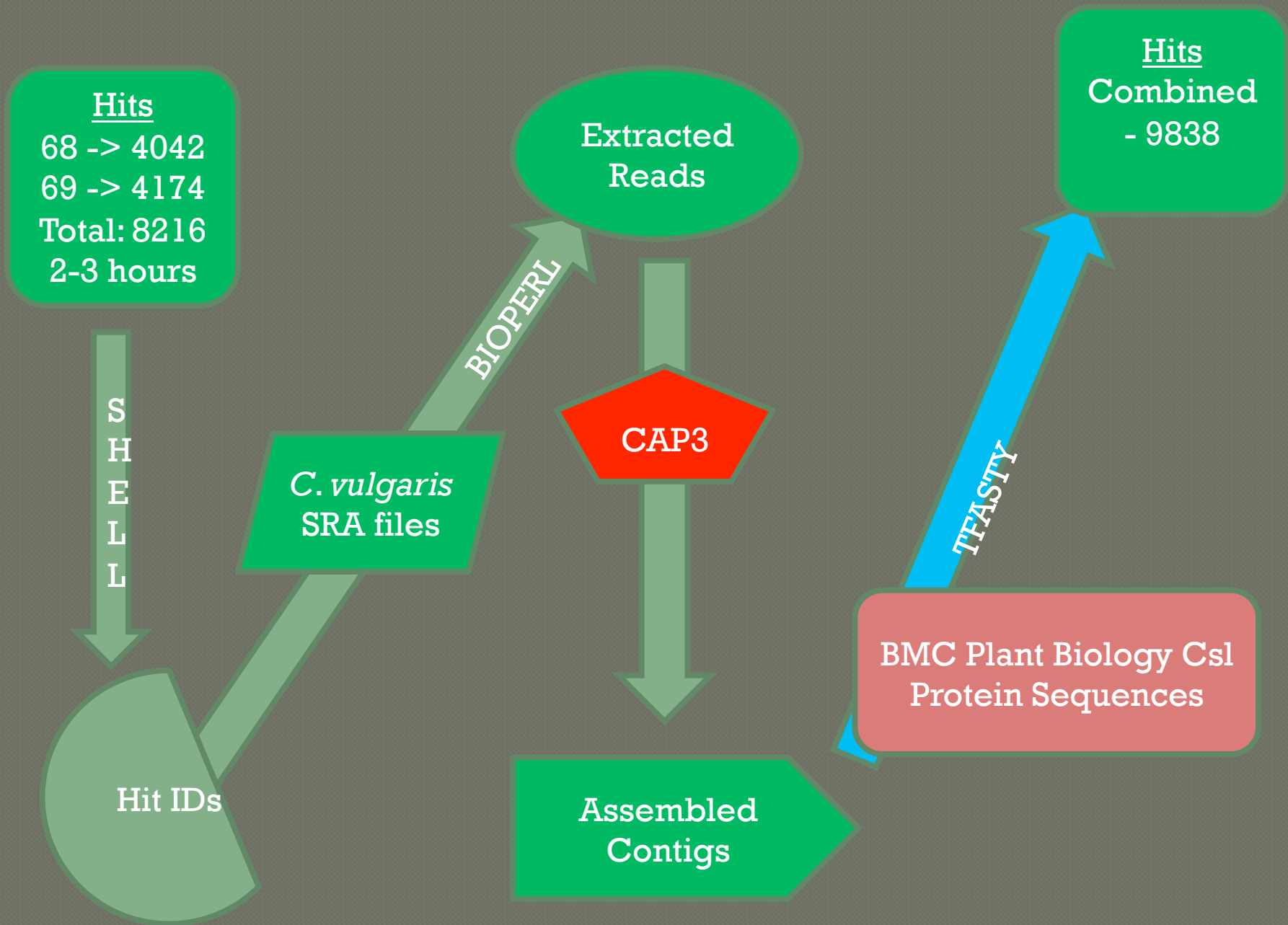
68 -> 4042

69 -> 4174

Total: 8216

2-3 hours

# PROCEDURE OUTLINE Cont.





binary

*Chara vulgaris* SRA  
SRR099268 (68),  
SRR099269 (69)

Fastq-  
dump

FastQ

Fastq  
To 454

FastA

*C. vulgaris*  
SRA files

BMC Plant Biology  
Csl Protein  
Sequences

TBLASTN

TEASTY

Hits

68 -> 5438

69 -> 5404

~16

minutes

Hits

68 -> 4042

69 -> 4174

Total: 8216

2-3 hours

# Download datasets:

1. NCBI SRA for *Chara vulgaris* using: `wget -q` &

<ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR099/SRR099268/>  
<ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR099/SRR099269/>

2. Csl proteins in BMC plant Biology 2009

<http://www.biomedcentral.com/qc/1471-2229/9/99/additional>

The screenshot shows the NCBI SRA (Sequence Read Archive) search results page for the query 'Chara vulgaris'. The page is displayed in a Firefox browser window. The search results show two runs: SRR099268 and SRR099269. A green arrow points to the table of results.

**Search Results Table:**

#	Run	# of Spots	# of Bases	Size
1.	<a href="#">SRR099268</a>	391,889	216.2M	<a href="#">458.8Mb</a>
2.	<a href="#">SRR099269</a>	348,466	199.3M	<a href="#">421.4Mb</a>

**Related information:** BioProject, BioSample, PubMed, Taxonomy

**Search details:** "Chara vulgaris"[Organism] OR Chara vulgaris[All Fields]

**Recent activity:** Chara vulgaris (1)

**Footer:** You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)

binary

*Chara vulgaris* SRA  
SRR099268 (68),  
SRR099269 (69)

Fastq-  
dump

FastQ

Fastq  
To 454

FastA

*C. vulgaris*  
SRA files

BMC Plant Biology  
Csl Protein  
Sequences

TBLASTN

TEASTY

Hits

68 -> 5438

69 -> 5404

~16

minutes

Hits

68 -> 4042

69 -> 4174

Total: 8216

2-3 hours

# Run the **fastq-dump** command to convert SRA format files to FASTQ format files

/home/mrupani/sratoolkit.2.1.16-centos\_linux64/bin/ fastq-dump SRR099268.sra

```
@glu: ~/project2
1]+ Stopped          metagenemark_predictions.faa | less (wd: ~/class/mar19)
2]- Running          wget -q ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR099/SRR099269/SRR099269.sra &
1003529@glu:~/project2$ rm index.html\?study\=SRP005673
1003529@glu:~/project2$ rm -rf ftp-trace.ncbi.nlm.nih.gov/
1003529@glu:~/project2$ ls
slproteinsequencesBMC SRR099268.sra SRR099269.sra
1003529@glu:~/project2$ jobs
1]+ Stopped          metagenemark_predictions.faa | less (wd: ~/class/mar19)
2]- Running          wget -q ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR099/SRR099269/SRR099269.sra &
1003529@glu:~/project2$ ls /home/mrupani/sratoolkit.2.1.16-centos_linux64/bin/
bi-dump          fastq-dump        illumina-load.2  nencvalid.2      refseq-load.2    sra-dbcc.2       srf-load.2       vdb-lock
bi-dump.2        fastq-dump.2.1.18 illumina-load.2  nencvalid.2.1.17 refseq-load.2.1.18 sra-dbcc.2.1.18  srf-load.2.1.17  vdb-lock.2
bi-dump.2.1.18   fastq-dump.2.1.18 illumina-load.2  nencvalid.2.1.17 refseq-load.2.1.18 sra-dbcc.2.1.18  srf-load.2.1.17  vdb-lock.2.1.17
bi-load          fastq-load.2.1.17 kar.2            pacbio-load.2    sff-dump.2       sra-pileup.1     vdb-copy.2       vdb-passwd
bi-load.2        fastq-load.2.1.17 kar.2.1.17      pacbio-load.2.1.17 sff-dump.2.1.17  sra-pileup.1.0.6 vdb-copy.2.1.17  vdb-passwd.1
bi-load.2.1.17   fastq-load.2.1.17 kdbmeta.2       rcexplain.2      sff-load.2       sra-stat.2       vdb-dump.2       vdb-unlock
lign-info        fastq-load.2.1.18 kdbmeta.2.1.17  rcexplain.2.1.17 sff-load.2.1.17  sra-stat.2.1.26 vdb-dump.2.1.17  vdb-unlock.2
lign-info.2      helicos-load.2    kdbmeta.2.1.17  rcexplain.2.1.17 sff-load.2.1.17  sra-stat.2.1.26 vdb-dump.2.1.17  vdb-unlock.2.1.1
lign-info.2.1.18 helicos-load.2.1.17 ncbi             refseq-load      sra-dbcc         srf-load         vdb-dump.2.1.17
sam-load         illumina-dump.2   nencvalid        refseq-load      sra-dbcc         srf-load         vdb-dump.2.1.17
sam-load.2       illumina-dump.2   nencvalid        refseq-load      sra-dbcc         srf-load         vdb-dump.2.1.17
sam-load.2.1.19 illumina-dump.2   nencvalid        refseq-load      sra-dbcc         srf-load         vdb-dump.2.1.17
1003529@glu:~/project2$ /home/mrupani/sratoolkit.2.1.16-centos_linux64/bin/fastq-dump

Usage:
/home/mrupani/sratoolkit.2.1.16-centos_linux64/bin/fastq-dump [options] <path [path...]>

Use option --help for more information

/home/mrupani/sratoolkit.2.1.16-centos_linux64/bin/fastq-dump : 2.1.18

1003529@glu:~/project2$ jobs
1]+ Stopped          metagenemark_predictions.faa | less (wd: ~/class/mar19)
2]- Running          wget -q ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR099/SRR099269/SRR099269.sra &
1003529@glu:~/project2$ /home/mrupani/sratoolkit.2.1.16-centos_linux64/bin/fastq-dump SRR099268.sra
Written 391889 spots for SRR099268.sra
Written 391889 spots total
1003529@glu:~/project2$ ls
slproteinsequencesBMC SRR099268.fastq SRR099268.sra SRR099269.sra
1003529@glu:~/project2$ less SRR099268.fastq
1003529@glu:~/project2$ jobs
1]+ Stopped          metagenemark_predictions.faa | less (wd: ~/class/mar19)
2]- Running          wget -q ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR099/SRR099269/SRR099269.sra &
1003529@glu:~/project2$
1003529@glu:~/project2$ jobs
1]+ Stopped          metagenemark_predictions.faa | less (wd: ~/class/mar19)
2]- Running          wget -q ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR099/SRR099269/SRR099269.sra &
1003529@glu:~/project2$
```

binary

*Chara vulgaris* SRA  
SRR099268 (68),  
SRR099269 (69)

Fastq-  
dump

FastQ

Fastq  
To454

FastA

*C. vulgaris*  
SRA files

BMC Plant Biology  
Csl Protein  
Sequences

TBLASTN

TEASTY

Hits

68 -> 5438

69 -> 5404

~16

minutes

Hits

68 -> 4042

69 -> 4174

Total: 8216

2-3 hours



# Run the **FastqTo454.pl** script to convert FASTQ format file to fasta + quality files

nohup perl /home/mrupani/ngs-qc/NGSQCToolkit\_v2.3/Format-converter/FastqTo454.pl -i SRR099268.fastq &

```
z1003529@glu: ~/project2
z1003529@glu:~$ ls/home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/
-bash: ls/home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/: Not a directo
ry
z1003529@glu:~$ ls/home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/
-bash: ls/home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/: Not a directo
ry
z1003529@glu:~$ ls /home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/
FastqTo454.pl      FastqToFastq.pl      SolexaFastqToIlluFastq.pl
FastqToFasta.pl   SangerFastqToIlluFastq.pl
z1003529@glu:~$ ls /home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/
FastqTo454.pl      FastqToFastq.pl      SolexaFastqToIlluFastq.pl
FastqToFasta.pl   SangerFastqToIlluFastq.pl
z1003529@glu:~$ cd projec2
-bash: cd: projec2: No such file or directory
z1003529@glu:~$ cd project2
z1003529@glu:~/project2$ ls
allSRR.contig          SRR099268.fastq_fna.nin
allSRR.contigs.BMC     SRR099268.fastq_fna.nsq
allSRR.contig.tfasty   SRR099268.fastq_qual
allSRR.contig.tfasty.fa SRR099268.hitid.fa
allSRR.hitid.fa        SRR099268.sra
allSRR.hitid.fa.cap.ace SRR099268.tblastn
allSRR.hitid.fa.cap.contigs SRR099268.tfasty36
allSRR.hitid.fa.cap.contigs.links SRR099268.tfasty36.hitid
allSRR.hitid.fa.cap.contigs.qual SRR099268.tfasty36.m8
allSRR.hitid.fa.cap.info SRR099269.fastq
allSRR.hitid.fa.cap.singlets SRR099269.fastq_fna
BMC                    SRR099269.fastq_fna.nhr
contigfile             SRR099269.fastq_fna.nin
CslproteinsequencesBMC SRR099269.fastq_fna.nsq
FastqTo454.pl          SRR099269.fastq_qual
formatdb.log           SRR099269.hitid.fa
get_sequence.pl        SRR099269.sra
hist1                  SRR099269.tblastn
nohup.out              SRR099269.tfasty36
SRR099268.fastq        SRR099269.tfasty36.hitid
SRR099268.fastq_fna    SRR099269.tfasty36.m8
SRR099268.fastq_fna.nhr top10_contigs
z1003529@glu:~/project2$ ls /home/mrupani/ngs-qc/NGSQCToolkit_v2.3/Format-converter/
FastqTo454.pl FastqToFasta.pl FastqToFastq.pl SangerFastqToIlluFastq.pl SolexaFastqToIlluFastq.pl
z1003529@glu:~/project2$
```



binary

*Chara vulgaris* SRA  
SRR099268 (68),  
SRR099269 (69)

Fastq-  
dump

FastQ

Fastq  
To 454

FastA

*C. vulgaris*  
SRA files

BMC Plant Biology  
Csl Protein  
Sequences

TFASTY

TBLASTN

Hits

68 -> 5438

69 -> 5404

~16

minutes

Hits

68 -> 4042

69 -> 4174

Total: 8216

2-3 hours

# TBLASTN

-Search translated nucleotide database using a protein query

formatdb -i SRR099268 -p F

time blastall -p blastn -i CslproteinsequencesBMC -d SRR099268.fastq\_fna -m 9 -o SRR099268.tblastn &. This yielded 5438 hits.

```
# TBLASTN 2.2.25 [Feb-01-2011]
# Query: AT2G21770.1|AT2G21770.1|cesA
# Database: SRR099268.fastq_fna
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
AT2G21770.1|AT2G21770.1|cesA SRR099268.100053 68.67 150 46 1 506 654 512 63 1e-71 237
AT2G21770.1|AT2G21770.1|cesA SRR099268.325348 68.52 108 34 0 550 657 454 131 1e-53 186
AT2G21770.1|AT2G21770.1|cesA SRR099268.77098 52.31 130 61 1 426 554 450 61 2e-36 137
AT2G21770.1|AT2G21770.1|cesA SRR099268.77098 60.00 20 8 0 553 572 65 6 2e-36 37.4
AT2G21770.1|AT2G21770.1|cesA SRR099268.77098 67.50 40 13 0 411 450 496 377 7e-11 64.3
AT2G21770.1|AT2G21770.1|cesA SRR099268.54011 43.21 162 85 6 458 612 7 480 6e-28 113
AT2G21770.1|AT2G21770.1|cesA SRR099268.111219 43.33 120 66 3 727 844 108 458 5e-20 90.9
AT2G21770.1|AT2G21770.1|cesA SRR099268.111219 37.50 88 55 0 775 862 248 511 4e-11 65.1
AT2G21770.1|AT2G21770.1|cesA SRR099268.284450 84.09 44 7 0 799 842 176 45 2e-16 80.5
AT2G21770.1|AT2G21770.1|cesA SRR099268.284450 32.28 127 80 4 702 822 483 109 1e-06 52.0
AT2G21770.1|AT2G21770.1|cesA SRR099268.187407 66.67 42 14 0 1027 1068 5 130 7e-12 70.1
AT2G21770.1|AT2G21770.1|cesA SRR099268.320890 48.21 56 28 1 1014 1068 109 276 2e-08 57.4
AT2G21770.1|AT2G21770.1|cesA SRR099268.48811 54.35 46 21 1 778 823 244 378 2e-07 54.3
AT2G21770.1|AT2G21770.1|cesA SRR099268.146293 51.28 39 18 1 1031 1068 111 227 3e-04 45.1
AT2G21770.1|AT2G21770.1|cesA SRR099268.146293 48.39 31 16 0 1014 1044 62 154 0.43 35.8
AT2G21770.1|AT2G21770.1|cesA SRR099268.357828 41.18 51 26 1 1022 1068 458 306 0.004 42.0
# TBLASTN 2.2.25 [Feb-01-2011]
# Query: AT2G25540.1|AT2G25540.1|cesA
# Database: SRR099268.fastq_fna
# Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
AT2G25540.1|AT2G25540.1|cesA SRR099268.100053 64.97 157 54 1 486 641 512 42 5e-69 229
AT2G25540.1|AT2G25540.1|cesA SRR099268.325348 55.40 139 62 0 530 668 454 38 8e-52 181
AT2G25540.1|AT2G25540.1|cesA SRR099268.77098 53.85 130 59 1 406 534 450 61 8e-38 141
AT2G25540.1|AT2G25540.1|cesA SRR099268.77098 65.00 40 14 0 391 430 496 377 1e-11 66.6
AT2G25540.1|AT2G25540.1|cesA SRR099268.77098 65.00 20 7 0 533 552 65 6 0.13 37.0
AT2G25540.1|AT2G25540.1|cesA SRR099268.54011 41.10 163 89 5 437 592 4 480 3e-26 108
AT2G25540.1|AT2G25540.1|cesA SRR099268.111219 41.22 131 77 2 693 823 72 458 1e-19 89.4
AT2G25540.1|AT2G25540.1|cesA SRR099268.111219 36.36 88 56 0 754 841 248 511 1e-10 63.9
AT2G25540.1|AT2G25540.1|cesA SRR099268.284450 75.00 44 11 0 778 821 176 45 4e-15 76.6
AT2G25540.1|AT2G25540.1|cesA SRR099268.284450 33.94 109 72 1 693 801 432 109 2e-08 57.4
AT2G25540.1|AT2G25540.1|cesA SRR099268.187407 76.19 42 10 0 1007 1048 5 130 3e-13 74.3
AT2G25540.1|AT2G25540.1|cesA SRR099268.320890 58.93 56 22 1 994 1048 109 276 1e-10 63.9
AT2G25540.1|AT2G25540.1|cesA SRR099268.48811 50.00 46 23 1 757 802 244 378 6e-07 52.8
AT2G25540.1|AT2G25540.1|cesA SRR099268.146293 55.00 40 17 1 1010 1048 108 227 5e-05 47.4
AT2G25540.1|AT2G25540.1|cesA SRR099268.146293 54.84 31 14 0 994 1024 62 154 0.020 39.7
AT2G25540.1|AT2G25540.1|cesA SRR099268.357828 49.02 51 22 1 1002 1048 458 306 0.001 43.9
AT2G25540.1|AT2G25540.1|cesA SRR099268.198581 29.76 84 53 3 792 869 219 467 2.2 33.5
AT2G25540.1|AT2G25540.1|cesA SRR099268.180144 32.39 71 44 2 606 672 95 304 2.9 33.1
AT2G25540.1|AT2G25540.1|cesA SRR099268.8503 30.59 85 59 3 86 349 116 3.2 33.1
AT2G25540.1|AT2G25540.1|cesA SRR099268.296672 30.99 71 45 2 606 672 99 308 4.7 32.3
# TBLASTN 2.2.25 [Feb-01-2011]
# Query: AT4G18780.1|AT4G18780.1|cesA
# Database: SRR099268.fastq_fna
```

# TBLASTN

formatdb -i SRR099269 -p F

Time blastall -p tblastn -i CslproteinsequencesBMC -d SRR099269.fastq\_fna -m 9 -o SRR099269.tblastn & This yielded **5404 hits**.

(Together both TBLASTN results took ~16 minutes)

```
TBLASTN 2.2.25 [Feb-01-2011]
Query: AT2G21770.1|AT2G21770.1|cesA
Database: SRR099269.fastq_fna
Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
2G21770.1|AT2G21770.1|cesA SRR099269.185742 67.54 114 37 0 759 872 343 2 1e-47 169
2G21770.1|AT2G21770.1|cesA SRR099269.123960 55.00 140 52 2 758 886 17 400 7e-43 155
2G21770.1|AT2G21770.1|cesA SRR099269.173430 47.20 125 64 4 259 381 108 470 5e-25 104
2G21770.1|AT2G21770.1|cesA SRR099269.225059 63.64 55 20 0 1014 1068 209 45 3e-24 87.4
2G21770.1|AT2G21770.1|cesA SRR099269.225059 42.19 64 33 4 919 978 510 322 3e-24 46.2
2G21770.1|AT2G21770.1|cesA SRR099269.308125 59.09 66 27 0 874 939 7 204 7e-22 95.9
2G21770.1|AT2G21770.1|cesA SRR099269.22664 61.76 68 26 0 308 375 257 460 5e-19 87.8
2G21770.1|AT2G21770.1|cesA SRR099269.22664 30.89 123 78 5 259 374 108 458 4e-05 47.8
2G21770.1|AT2G21770.1|cesA SRR099269.43157 63.64 55 20 0 1014 1068 174 338 9e-19 87.4
2G21770.1|AT2G21770.1|cesA SRR099269.197165 57.58 66 28 0 300 365 286 89 1e-17 84.0
2G21770.1|AT2G21770.1|cesA SRR099269.197165 47.50 40 20 1 354 392 125 6 0.004 41.6
2G21770.1|AT2G21770.1|cesA SRR099269.35237 35.88 131 83 3 200 329 439 101 1e-13 72.4
2G21770.1|AT2G21770.1|cesA SRR099269.165723 51.39 72 35 2 728 799 265 474 3e-12 67.8
2G21770.1|AT2G21770.1|cesA SRR099269.165723 81.25 16 3 0 639 654 27 74 6.8 32.0
2G21770.1|AT2G21770.1|cesA SRR099269.309892 65.71 35 12 0 1034 1068 25 129 2e-08 57.4
2G21770.1|AT2G21770.1|cesA SRR099269.94071 46.81 47 20 1 1019 1060 63 203 1e-05 51.2
2G21770.1|AT2G21770.1|cesA SRR099269.193507 45.24 42 23 1 1027 1068 9 131 0.002 42.4
2G21770.1|AT2G21770.1|cesA SRR099269.193507 70.59 17 5 0 1027 1043 5 55 0.60 35.0
2G21770.1|AT2G21770.1|cesA SRR099269.210560 45.24 42 23 1 1027 1068 9 131 0.002 42.4
2G21770.1|AT2G21770.1|cesA SRR099269.210560 70.59 17 5 0 1027 1043 5 55 0.62 35.0
2G21770.1|AT2G21770.1|cesA SRR099269.112083 36.11 36 23 0 114 149 428 535 8.0 32.0
TBLASTN 2.2.25 [Feb-01-2011]
Query: AT2G25540.1|AT2G25540.1|cesA
Database: SRR099269.fastq_fna
Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
2G25540.1|AT2G25540.1|cesA SRR099269.185742 66.67 114 38 0 738 851 343 2 6e-46 164
2G25540.1|AT2G25540.1|cesA SRR099269.123960 55.47 137 61 1 743 879 35 442 7e-41 150
2G25540.1|AT2G25540.1|cesA SRR099269.225059 64.94 77 24 1 975 1048 275 45 2e-31 105
2G25540.1|AT2G25540.1|cesA SRR099269.225059 52.83 53 21 2 907 958 477 322 2e-31 52.4
2G25540.1|AT2G25540.1|cesA SRR099269.308125 67.65 68 22 0 853 920 7 210 1e-25 106
2G25540.1|AT2G25540.1|cesA SRR099269.43157 61.05 95 34 1 957 1048 54 338 4e-25 105
2G25540.1|AT2G25540.1|cesA SRR099269.197165 45.87 109 59 3 237 345 367 89 2e-23 87.4
2G25540.1|AT2G25540.1|cesA SRR099269.197165 56.76 37 15 1 337 372 116 6 2e-23 43.5
2G25540.1|AT2G25540.1|cesA SRR099269.22664 68.63 51 16 0 288 338 257 409 1e-19 87.8
2G25540.1|AT2G25540.1|cesA SRR099269.22664 54.55 22 10 0 333 354 393 458 1e-19 30.4
2G25540.1|AT2G25540.1|cesA SRR099269.173430 40.65 123 73 2 239 361 108 470 7e-16 78.6
2G25540.1|AT2G25540.1|cesA SRR099269.165723 48.61 72 37 2 707 778 265 474 2e-11 65.9
2G25540.1|AT2G25540.1|cesA SRR099269.35237 29.32 133 94 4 177 309 448 101 4e-10 62.0
2G25540.1|AT2G25540.1|cesA SRR099269.309892 74.29 35 9 0 1014 1048 25 129 3e-09 59.3
2G25540.1|AT2G25540.1|cesA SRR099269.94071 64.86 37 13 0 1004 1040 93 203 4e-07 56.2
2G25540.1|AT2G25540.1|cesA SRR099269.193507 60.00 35 14 0 1014 1048 27 131 5e-04 43.9
2G25540.1|AT2G25540.1|cesA SRR099269.193507 70.59 17 5 0 1007 1023 5 55 0.30 35.8
2G25540.1|AT2G25540.1|cesA SRR099269.210560 60.00 35 14 0 1014 1048 27 131 6e-04 43.9
```

binary

*Chara vulgaris* SRA  
SRR099268 (68),  
SRR099269 (69)

Fastq-  
dump

FastQ

Fastq  
To 454

FastA

*C. vulgaris*  
SRA files

BMC Plant Biology  
Csl Protein  
Sequences

TBLASTN

Hits

68 -> 5438

69 -> 5404

~16

minutes

TEASTY

Hits

68 -> 4042

69 -> 4174

Total: 8216

2-3 hours



# TFASTY SRR099268

also compares a protein sequence to a DNA sequence database

**..better alignment with poor quality sequences**

`nohup time tfasty36 -m 8 CslproteinsequencesBCM SRR099268.fastq_fna > SRR099268.tfasty36.m8 &`

This gave 4042 hits.

```
TFASTY 36.3.5e Nov, 2012(preload8)
Query: AT2G21770.1|AT2G21770.1|cesA - 1089 aa
Database: SRR099268.fastq_fna
T2G21770.1|AT2G21770.1|cesA SRR099268.100053 67.95 156 49 2 506 660 512 46 8.8e-45 184.0
T2G21770.1|AT2G21770.1|cesA SRR099268.77098 63.19 163 59 3 411 572 496 6 4.1e-39 165.2
T2G21770.1|AT2G21770.1|cesA SRR099268.325348 62.25 151 54 4 542 691 479 32 3.8e-38 161.9
T2G21770.1|AT2G21770.1|cesA SRR099268.54011 59.04 166 66 7 458 621 7 507 5.9e-34 148.1
T2G21770.1|AT2G21770.1|cesA SRR099268.111219 57.35 136 57 4 727 862 108 511 1.4e-24 116.9
T2G21770.1|AT2G21770.1|cesA SRR099268.284450 53.64 151 65 8 697 842 498 45 2.5e-22 109.4
T2G21770.1|AT2G21770.1|cesA SRR099268.146293 56.36 55 24 1 1014 1068 62 227 1.2e-07 60.7
T2G21770.1|AT2G21770.1|cesA SRR099268.187407 66.67 42 14 0 1027 1068 5 130 5.2e-07 59.4
T2G21770.1|AT2G21770.1|cesA SRR099268.320890 46.91 81 39 7 992 1068 35 276 7e-06 54.8
T2G21770.1|AT2G21770.1|cesA SRR099268.48811 43.04 79 40 6 745 823 156 378 0.00017 50.2
T2G21770.1|AT2G21770.1|cesA SRR099268.174572 39.06 64 35 5 640 699 355 545 0.11 41.0
T2G21770.1|AT2G21770.1|cesA SRR099268.290286 47.22 36 19 0 640 675 353 460 0.18 40.1
T2G21770.1|AT2G21770.1|cesA SRR099268.298250 47.22 36 19 0 640 675 356 463 0.29 39.5
T2G21770.1|AT2G21770.1|cesA SRR099268.357828 45.10 51 24 4 1022 1068 458 306 0.34 39.4
T2G21770.1|AT2G21770.1|cesA SRR099268.13229 38.71 62 33 6 640 696 202 386 0.54 38.4
T2G21770.1|AT2G21770.1|cesA SRR099268.297900 37.50 56 32 4 620 675 304 461 1.3 37.3
T2G21770.1|AT2G21770.1|cesA SRR099268.276692 32.32 99 56 12 584 675 181 464 1.4 37.3
T2G21770.1|AT2G21770.1|cesA SRR099268.209607 48.72 39 20 1 640 678 355 470 1.5 37.0
T2G21770.1|AT2G21770.1|cesA SRR099268.305053 39.29 56 31 5 620 675 305 463 1.6 37.0
T2G21770.1|AT2G21770.1|cesA SRR099268.94031 50.00 36 18 1 640 675 221 327 1.6 37.0
T2G21770.1|AT2G21770.1|cesA SRR099268.102829 50.00 36 18 1 640 675 50 156 1.6 36.4
T2G21770.1|AT2G21770.1|cesA SRR099268.180597 50.00 36 18 1 640 675 357 463 1.8 36.8
T2G21770.1|AT2G21770.1|cesA SRR099268.247105 29.59 98 61 9 584 675 430 144 1.9 36.8
TFASTY 36.3.5e Nov, 2012(preload8)
Query: AT2G25540.1|AT2G25540.1|cesA - 1066 aa
Database: SRR099268.fastq_fna
T2G25540.1|AT2G25540.1|cesA SRR099268.100053 64.97 157 54 1 486 641 512 42 6.4e-48 194.4
T2G25540.1|AT2G25540.1|cesA SRR099268.77098 65.03 163 56 3 391 552 496 6 2.3e-45 185.9
T2G25540.1|AT2G25540.1|cesA SRR099268.54011 58.08 167 68 7 437 601 4 507 5.9e-36 154.7
T2G25540.1|AT2G25540.1|cesA SRR099268.325348 58.16 141 54 6 522 657 479 56 6.7e-36 154.4
T2G25540.1|AT2G25540.1|cesA SRR099268.111219 54.36 149 66 5 693 841 72 511 3e-28 129.1
T2G25540.1|AT2G25540.1|cesA SRR099268.284450 55.81 129 57 3 693 821 432 45 8.1e-26 121.0
T2G25540.1|AT2G25540.1|cesA SRR099268.146293 61.82 55 21 1 994 1048 62 227 1.4e-09 67.1
T2G25540.1|AT2G25540.1|cesA SRR099268.187407 76.19 42 10 0 1007 1048 5 130 9.2e-09 65.2
T2G25540.1|AT2G25540.1|cesA SRR099268.320890 53.09 81 34 5 972 1048 35 276 1e-08 64.2
T2G25540.1|AT2G25540.1|cesA SRR099268.48811 37.27 110 61 10 693 802 71 378 1.7e-05 53.4
T2G25540.1|AT2G25540.1|cesA SRR099268.357828 60.78 51 16 6 1002 1048 458 306 0.046 42.3
T2G25540.1|AT2G25540.1|cesA SRR099268.111594 29.03 93 64 3 372 463 187 461 0.24 39.8
T2G25540.1|AT2G25540.1|cesA SRR099268.179180 27.33 150 95 18 55 204 38 443 0.74 38.1
T2G25540.1|AT2G25540.1|cesA SRR099268.336085 26.28 156 107 12 69 219 58 512 0.76 38.1
T2G25540.1|AT2G25540.1|cesA SRR099268.180144 27.35 117 82 4 606 719 95 446 0.88 37.8
T2G25540.1|AT2G25540.1|cesA SRR099268.119068 24.54 163 108 18 64 224 14 462 0.89 37.9
T2G25540.1|AT2G25540.1|cesA SRR099268.153271 27.94 136 94 8 69 204 59 452 1.2 37.4
T2G25540.1|AT2G25540.1|cesA SRR099268.206401 24.82 141 99 9 66 204 14 421 1.5 37.1
SRR099268.tfasty36.m8
```

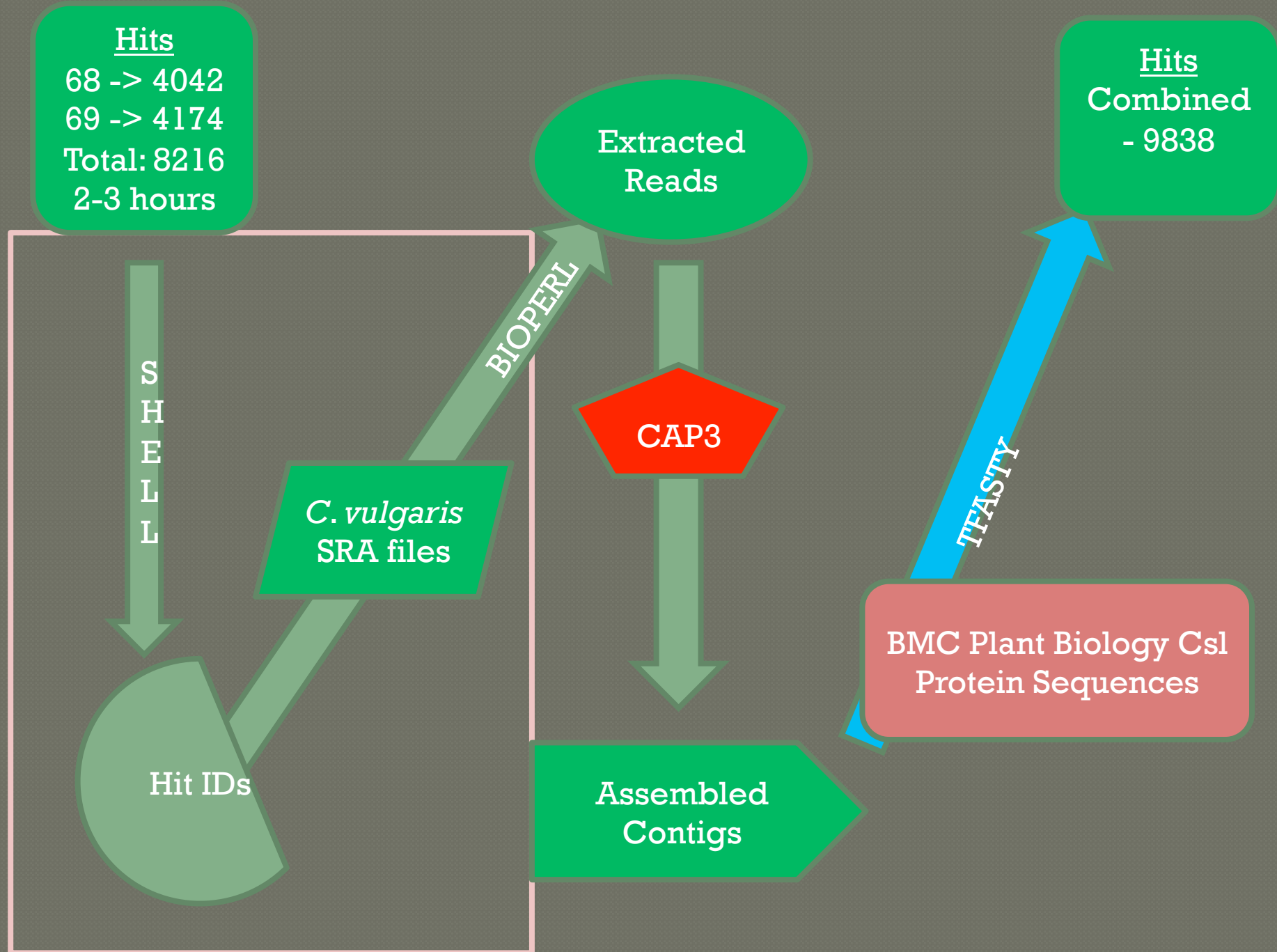
# TFASTY SRR099269

nohup time tfasty36 -m 8 CslproteinsequencesBCM SRR099269.fastq\_fna >  
SRR099269.tfasty36.m8 &. It retrieved 4174 hits.  
(Took ~2 or 3 hours)

```
TBLASTN 2.2.25 [Feb-01-2011]
Query: AT2G21770.1|AT2G21770.1|cesA
Database: SRR099269.fastq_fna
Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
2G21770.1|AT2G21770.1|cesA SRR099269.185742 67.54 114 37 0 759 872 343 2 1e-47 169
2G21770.1|AT2G21770.1|cesA SRR099269.123960 55.00 140 52 2 758 886 17 400 7e-43 155
2G21770.1|AT2G21770.1|cesA SRR099269.173430 47.20 125 64 4 259 381 108 470 5e-25 104
2G21770.1|AT2G21770.1|cesA SRR099269.225059 63.64 55 20 0 1014 1068 209 45 3e-24 87.4
2G21770.1|AT2G21770.1|cesA SRR099269.225059 42.19 64 33 4 919 978 510 322 3e-24 46.2
2G21770.1|AT2G21770.1|cesA SRR099269.308125 59.09 66 27 0 874 939 7 204 7e-22 95.9
2G21770.1|AT2G21770.1|cesA SRR099269.22664 61.76 68 26 0 308 375 257 460 5e-19 87.8
2G21770.1|AT2G21770.1|cesA SRR099269.22664 30.89 123 78 5 259 374 108 458 4e-05 47.8
2G21770.1|AT2G21770.1|cesA SRR099269.43157 63.64 55 20 0 1014 1068 174 338 9e-19 87.4
2G21770.1|AT2G21770.1|cesA SRR099269.197165 57.58 66 28 0 300 365 286 89 1e-17 84.0
2G21770.1|AT2G21770.1|cesA SRR099269.197165 47.50 40 20 1 354 392 125 6 0.004 41.6
2G21770.1|AT2G21770.1|cesA SRR099269.35237 35.88 131 83 3 200 329 439 101 1e-13 72.4
2G21770.1|AT2G21770.1|cesA SRR099269.165723 51.39 72 35 2 728 799 265 474 3e-12 67.8
2G21770.1|AT2G21770.1|cesA SRR099269.165723 81.25 16 3 0 639 654 27 74 6.8 32.0
2G21770.1|AT2G21770.1|cesA SRR099269.309892 65.71 35 12 0 1034 1068 25 129 2e-08 57.4
2G21770.1|AT2G21770.1|cesA SRR099269.94071 46.81 47 20 1 1019 1060 63 203 1e-05 51.2
2G21770.1|AT2G21770.1|cesA SRR099269.193507 45.24 42 23 1 1027 1068 9 131 0.002 42.4
2G21770.1|AT2G21770.1|cesA SRR099269.193507 70.59 17 5 0 1027 1043 5 55 0.60 35.0
2G21770.1|AT2G21770.1|cesA SRR099269.210560 45.24 42 23 1 1027 1068 9 131 0.002 42.4
2G21770.1|AT2G21770.1|cesA SRR099269.210560 70.59 17 5 0 1027 1043 5 55 0.62 35.0
2G21770.1|AT2G21770.1|cesA SRR099269.112083 36.11 36 23 0 114 149 428 535 8.0 32.0
TBLASTN 2.2.25 [Feb-01-2011]
Query: AT2G25540.1|AT2G25540.1|cesA
Database: SRR099269.fastq_fna
Fields: Query id, Subject id, % identity, alignment length, mismatches, gap openings, q. start, q. end, s. start, s. end, e-value, bit score
2G25540.1|AT2G25540.1|cesA SRR099269.185742 66.67 114 38 0 738 851 343 2 6e-46 164
2G25540.1|AT2G25540.1|cesA SRR099269.123960 55.47 137 61 1 743 879 35 442 7e-41 150
2G25540.1|AT2G25540.1|cesA SRR099269.225059 64.94 77 24 1 975 1048 275 45 2e-31 105
2G25540.1|AT2G25540.1|cesA SRR099269.225059 52.83 53 24 2 907 958 477 322 2e-31 52.4
2G25540.1|AT2G25540.1|cesA SRR099269.308125 67.65 68 22 0 853 920 7 210 1e-25 106
2G25540.1|AT2G25540.1|cesA SRR099269.43157 61.05 95 34 1 957 1048 54 338 4e-25 105
2G25540.1|AT2G25540.1|cesA SRR099269.197165 45.87 109 59 3 237 345 367 89 2e-23 87.4
2G25540.1|AT2G25540.1|cesA SRR099269.197165 56.76 37 15 1 337 372 116 6 2e-23 43.5
2G25540.1|AT2G25540.1|cesA SRR099269.22664 68.63 51 16 0 288 338 257 409 1e-19 87.8
2G25540.1|AT2G25540.1|cesA SRR099269.22664 54.55 22 10 0 333 354 393 458 1e-19 30.4
2G25540.1|AT2G25540.1|cesA SRR099269.173430 40.65 123 73 2 239 361 108 470 7e-16 78.6
2G25540.1|AT2G25540.1|cesA SRR099269.165723 48.61 72 37 2 707 778 265 474 2e-11 65.9
2G25540.1|AT2G25540.1|cesA SRR099269.35237 29.32 133 94 4 177 309 448 101 4e-10 62.0
2G25540.1|AT2G25540.1|cesA SRR099269.309892 74.29 35 9 0 1014 1048 25 129 3e-09 59.3
2G25540.1|AT2G25540.1|cesA SRR099269.94071 64.86 37 13 0 1004 1040 93 203 4e-07 56.2
2G25540.1|AT2G25540.1|cesA SRR099269.193507 60.00 35 14 0 1014 1048 27 131 5e-04 43.9
2G25540.1|AT2G25540.1|cesA SRR099269.193507 70.59 17 5 0 1007 1023 5 55 0.30 35.8
2G25540.1|AT2G25540.1|cesA SRR099269.210560 60.00 35 14 0 1014 1048 27 131 6e-04 43.9
```

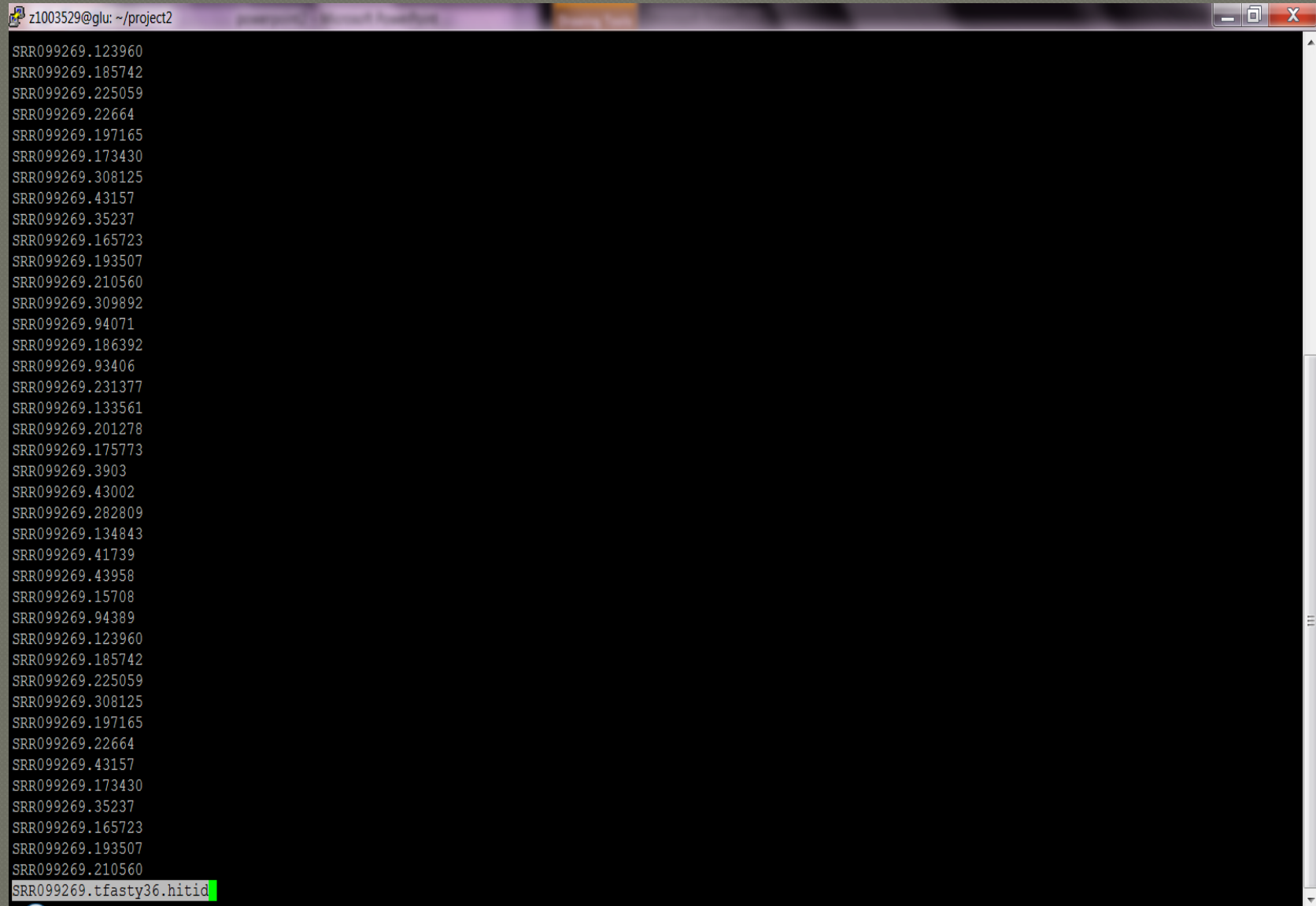
SRR099269.tblastn





```
Less SRR099268.tfasty36.m8 | cut-f2 | grep 'SRR' | grep -v '#' | less >  
SRR099268.tfasty36.hitid
```

```
Less SRR099269.tfasty36.m8 | cut-f2 | grep 'SRR' | grep -v '#' | less >  
SRR099269.tfasty36.hitid
```



A terminal window titled 'z1003529@glu: ~/project2' displays a list of SRR099269 accession numbers. The list starts with SRR099269.123960 and ends with SRR099269.210560, followed by the filename SRR099269.tfasty36.hitid. The window has a standard Linux terminal interface with a title bar and window controls.

```
SRR099269.123960  
SRR099269.185742  
SRR099269.225059  
SRR099269.22664  
SRR099269.197165  
SRR099269.173430  
SRR099269.308125  
SRR099269.43157  
SRR099269.35237  
SRR099269.165723  
SRR099269.193507  
SRR099269.210560  
SRR099269.309892  
SRR099269.94071  
SRR099269.186392  
SRR099269.93406  
SRR099269.231377  
SRR099269.133561  
SRR099269.201278  
SRR099269.175773  
SRR099269.3903  
SRR099269.43002  
SRR099269.282809  
SRR099269.134843  
SRR099269.41739  
SRR099269.43958  
SRR099269.15708  
SRR099269.94389  
SRR099269.123960  
SRR099269.185742  
SRR099269.225059  
SRR099269.308125  
SRR099269.197165  
SRR099269.22664  
SRR099269.43157  
SRR099269.173430  
SRR099269.35237  
SRR099269.165723  
SRR099269.193507  
SRR099269.210560  
SRR099269.tfasty36.hitid
```

## Perl script (with bioperl modules for taking hit ids and the original databases as the input files to extract the fasta sequences

perl get\_sequence.pl SRR099268.tfasty36.hitid SRR099268.fastq\_fna SRR099268.hitid.fa

```
get_sequence.pl ✕
#!/usr/bin/perl -w

# get_sequence

# Author: Steven Seydell
# Date: 4/13/2013

# This program will read in a file of unsorted sequence IDs, and locate the
# fasta format sequences in a sequence library. The library need not be
# sorted.
#
# Usage: perl get_sequence.pl <input file> <database file> <output file>

# Use the BioPerl SeqIO library
use Bio::SeqIO;

# Read in the list of hit ids and store them in a hash
open(ID,$ARGV[0]);
while(<ID>){
    chomp $_;
    $id_hash{$_}=1;
}

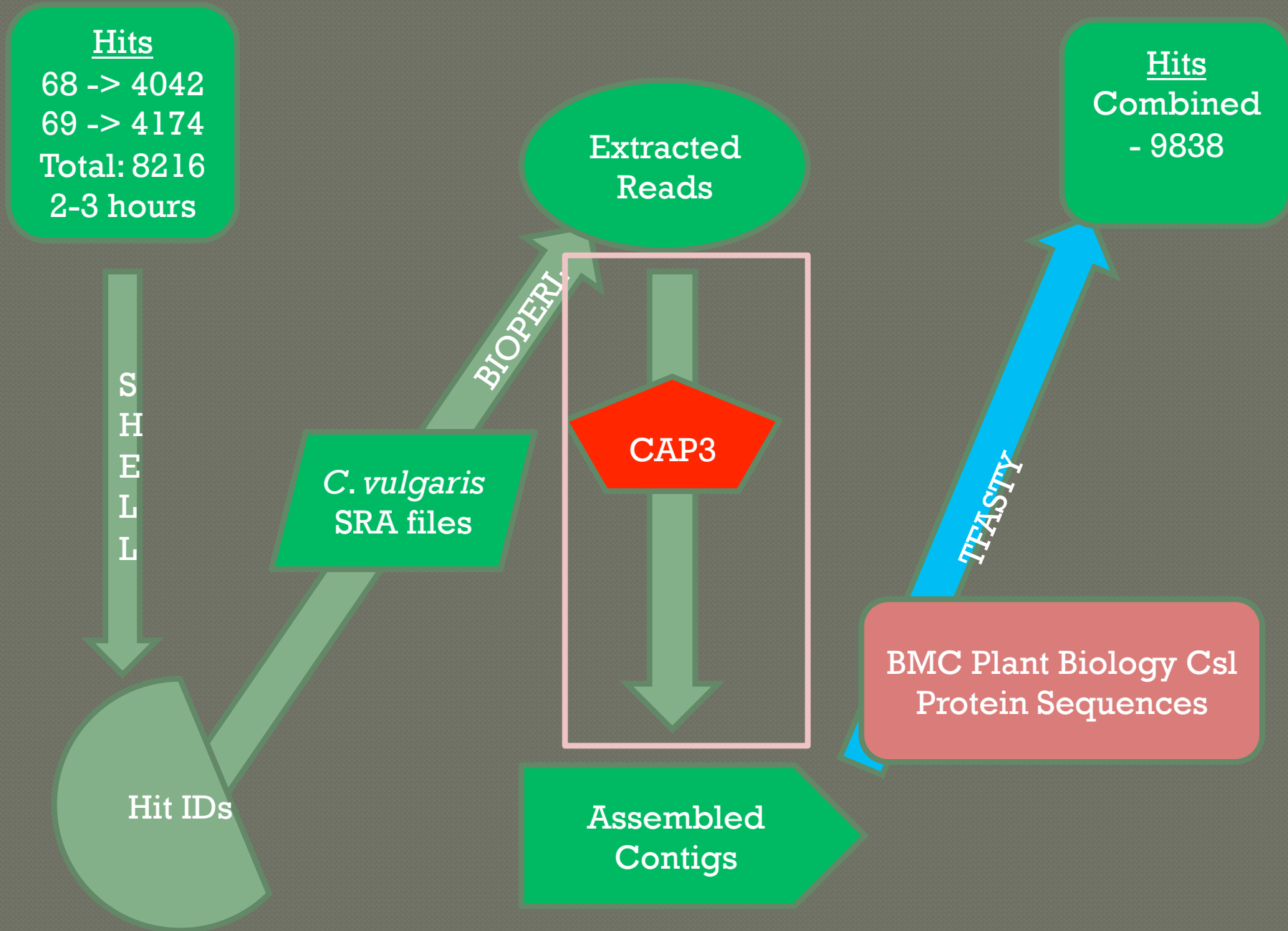
# open the database and output files
$db=Bio::SeqIO->new(-file=>$ARGV[1], -format=>"fasta");
$out_file = Bio::SeqIO->new(-file => ">$ARGV[2]", -format => "fasta" );

# Loop through each sequence in the database looking for matches
while($seq=$db->next_seq){
    if(defined $id_hash{$seq->id}){
        # print the sequence out to the output file
        $out_file->write_seq($seq);
    }
}
```

Per

# Extracted SRA Reads in FASTA format combined from both of the data sets SRR099268 and SRR099269

```
z1003529@glu: ~/project2
>SRR099269.307 FRX18GR02Q71LT length=676
TCAGAGGACTGCAGATCGGCGAGAGTGTGTGCAGGCACCTGCAACTAGGAGATGTCAAC
GNGCANACGGATGGCAGATAGTTGTGTGTACGCAGAGGTACAACGTGTGCGTCGACACG
TAGCTCGGCCCGCGAAGGTTAAAGGACCTCGACGAAAGTACGTACGGGTGCGAAGGAAGG
TTGTTGTGGAACNGGACCTCCTTCCGACGGTAACGTAGGAAGGTACTTCGACCGTACC
GGCGAAGCCACGGGAAGGGAAGTTTGGCTGACGTACGTAGTAGTAGTAGTCGTAGACGTA
CGTACGTAACTACGTAGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT
ACGGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGT
CGACCGACCGACGACGACGACGACGACGACGACGACGACGACGACGACGACGACGACGACG
TACGACGACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTAC
GTCGTGTCGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
TACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACGTACG
TACGTACGTACGTACG
>SRR099269.531 FRX18GR02P8GM2 length=544
TCAGAAGCAGTGGTATCAACGCAGAGTACCGGGGGCTGTCTTTCTTTTACTCACGTC
GGTGAGATAATCCGACCGCGCGCGCGAGTTTCGGGATTCGAGCGGACGTAACGCCA
CCATGAGGGCGAAGTGAAGAAGAAGCGTATGAGGCGCTTAAGAGGAAGCGCAGAAAGA
TGCGCCAGAGATCCAAGTAGCGTGCACGCCTAGGGAATGGATTGGGTGATGGAGAGGGG
GTGGTTCGGGTTTGGGAGAGGAGTGAGGTGCTCCCTTGTGCTGGGAAGGGGAATGGGAGG
AAGAGCCCATAGGGAGATGGCAAAAGCATTTGCTATACGGATATACGGGGTGCTTCCTTT
TTGGCAAATGCACCTCTTCATTGCTCATTCTATTAACAATTTGTGTAATGTAAATGATATA
AGCCATAAAAAAAAAAAAAAAAAAAAAAAAAAATACTNTAGTACGTAGAACTACCAC
TGCTTCTAGAGTACGTACGACCGAAAGGCCAACACGAGGGAGTAGGNGNNGNNNNNNNN
NNNN
>SRR099269.1240 FRX18GR02QK3T6 length=347
TCAGAAGCAGTGGTATCAACGCAGAGTACGGGCGAGTACGTAGGATCCGGCGAGGTTTG
AAGGAGAAGAGAAACGGGAGAGAAATGGAATTTGACGACAAAAGCAGAGAGAGAGGGAGAG
AGAGGAGGAAGGACGGGCGAGTGAAGTTGAGGGTCAGAGGAGGAAGGTTCTCGACAATT
TCTCGGCAGACACTCGCAGGCGGTATACACTTTCGATAATTGTATATCTTTTCCGATT
TTCTTAAATAAAAAAAAAATTAAGTACCGGAAAAAAAGTAAAACTACTGCGTTGATACC
ACTGCTTCTGAGACTGCCAAGGCACACAGGGGATAGGNNNNNNNNNN
>SRR099269.1885 FRX18GR02TRKHH length=295
TCAGAAGCAGTGGTATCAACGCAGAGTACGGGGTCGTGAGAGAAGAGAGAGAGAGAGAG
GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAAGAAGAAG
AAGAAGAAGAAGAAGAAGAGACGACGACGACGACGACGACGACGACGACGACGACGACG
CGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
ACGACGACGACGACGAGACGACGACGACGACGACGACGACGACGACGACGACGACGACG
>SRR099269.3903 FRX18GR02SQVOQ length=538
TCAGAAGCGTGGTATCAACGCAGAGTACGGGGCTCCTTCGTCTCCACCGTTGACACTTC
GTCTATCTAGTTCCACGTTCCCGACGTCGTTCTGCTAGGGGTTGGTATCGACATCAGCAG
SRR099269.hitid.fa
```



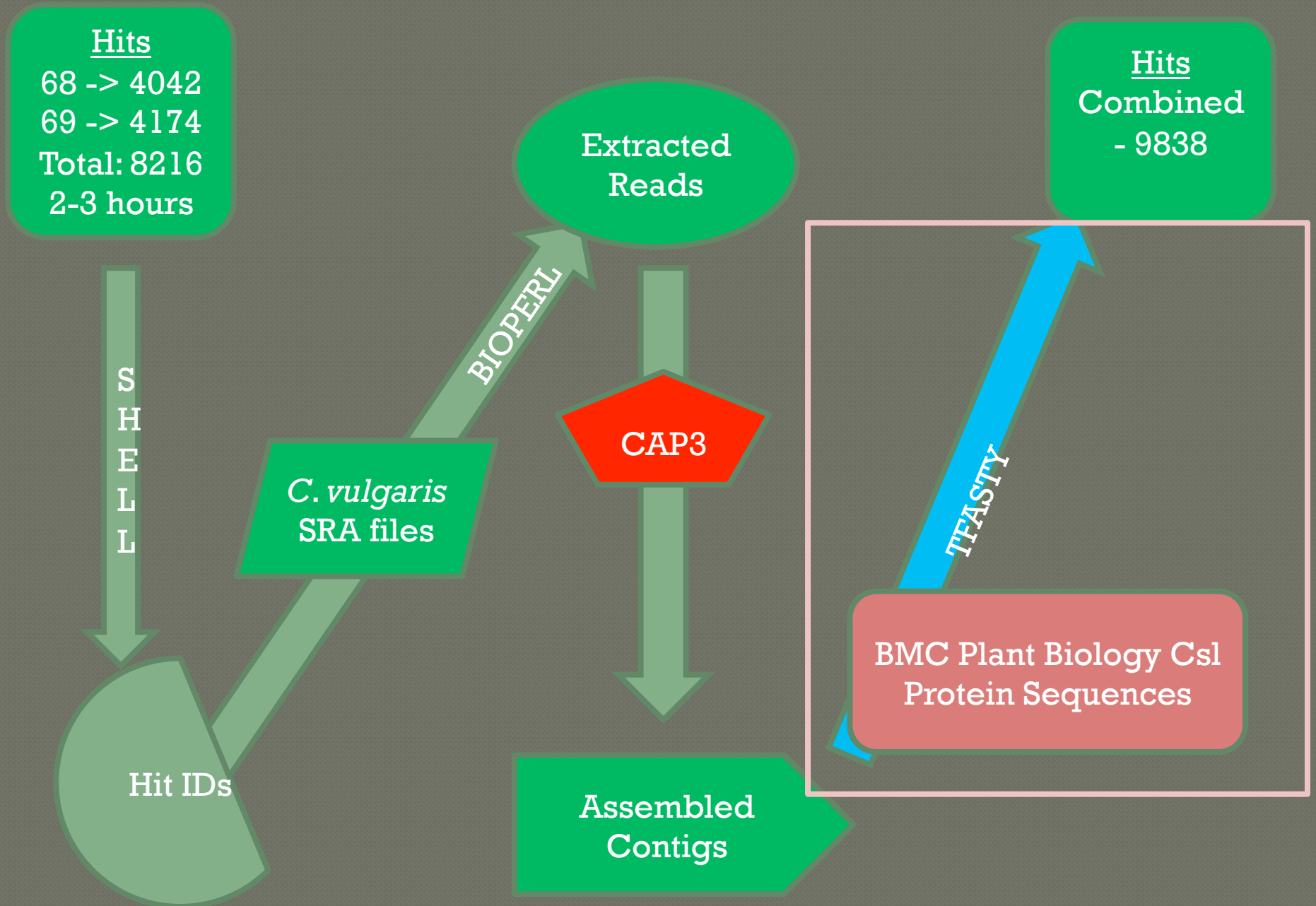
# Cap3- sequence (contig) assembly program

cap3 allSRR.hitid.fa -o 60 (coverage ..how much overlap) -p 97 (percentage..of similarity in the overlap) > allSRR.hitid.fa.contigs

Less  
allSRR.hitid.fa.contigs

```
>Contig1
TCAGAAGCAGTGGTATCAACGCAGAGTACGGGGGGGCTGATTGAGTGGTTCGTCCACGTCTG
GTAATATCTCCAGCGCCATCTCCAGCCATCTCCCGCGCCATCCCCAGCGCCATCTCCAGC
GTTACCTCCAGCGCCATCTCCAGCGCCATCTTCAGCGCCATCTCCAACGCCATCTTCAGC
GCCATCTCCAGCGCAATCTCCGGCTCCATCTCTAGGGACATCTTCGGCGCCATTTCCAGC
GCTACAGAAGGAGGTACACCTGTTTCTGCGTACGACCGCTCATCAGGCTGACAGTAAGAG
GTCATCGGGCAAGCGTGGTGCCTGCTGATCGTTGCAGCCATCTCCGTGATAGCTGAGGTT
AATCGAAGTGTGACGGGAGCGGAGGTCTCTTCTTGCTGTATTACAGGGAGAGAGGGCCA
AAACTACTTACCGCGAAATGGCTCGTTTCGAGGGGAGATCAGGTGGACTTGCAAGGAGCCG
>Contig2
TCAGAAGCAGTGGTATCAACGCAGAGTACGGGGGGGCTGATTGAGTGGTTCGTCCACGTCTG
GTAATATCTCCAGCGCCATCTCCAGCCATCTCCCGCGCCATCCCCAGCGCCATCTCCAGC
GTTACCTCCAGCGCCATCTCCAGCGCCATCTTCAGCGCCATCTCCAACGCCATCTTCAGC
GCCATCTCCAGCGCAATCTCCGGCTCCATCTCTAGGGACATCTTCGGCGCCATTTCCAGC
GCTACAGAAGGAGGTACACCTGTTTCTGCGTACGACCGCTCATCAGGCTGACAGTAAGAG
GTCATCGGGCAAGCGTGGTGCCTGCTGATCGTTGCAGCCATCTCCGTGATAGCTGAGGTT
AATCGAAGTGTGACGGGAGCGGAGGTCTCTTCTTGCTGTATTACAGGGGAGAGGGAGGG
GCCAAAATACTTACCGCGAAATGGCTCGTTTCGAGGGGAAGATCAGGTGGACTTGCAAG
GAGGCCTG
>Contig3
TCAGTGAGGGATAGCAGGTGGGGGAGCCTGTGCCGCTGGACGAGGAGCTAGGGGGGAAGGT
GCAGGCCTCCTTGCAAGTCCACCTGATCTTCCCCTCGAACGAGCCATTTTCGCGGTAAGT
AGTTTTGGCCCTCTCTCCCCTGTAATACAGCAAGAGACCTCCGCTCCCGTGACACT
TCGATTAACCTCAGCTATCACGGAGATGGCTGCAACGATCAGCAGGCACCACTGTGCC
GATAACCTCTTACTGTCAGCCTGATGAGCGGTCTGTACGCAGAAACAGGTGTACCTCCTTC
TGTAGCGCTGGAAATGGCGCCGAAGATGTCCTAGAGATGGAGCCGGAGATTGCGCTGGA
GATGGCGCTGAAGATGGCGTTGGAGATGGCGCTGAAGATGGCGCTGGAGATGGCGCTGGA
GGTAACGCTGGAGATGGCGCTGGGGATGGCGCGGGAGATGGCTGGAGATGGCGCTGGAGA
TATTACCG
>Contig4
TCAGAAGCAGTGGTATCAACGCAGAGTACGGGGGGCTTTTGGTTGGGGGAGCGAGTGGGCC
GATCGACAGACGGGCGGGTTTGTGTCGTCGTCGTCGTCGTTGTCCCTGTTGTGGTACCTT
TCGGTCTTGTGTTGCCGGGCTTTTGCATTTCCGCGGCAAGAAAGTGTAGCAGCCATGGCGT
CGGAGAAGAAGCAGTTGAACGTGATGAGGGAGATCAAAGTTCAGAAGCTTGTCTGAACA
TTTCCGTTGGAGAGAGTGGCGATAGGCTGACCCGTGCTTCCAAGGTGCTGGAGCAACTGA
GTGGCCAATCTCCCGTGTCTCGAAGGCCAGATACACCGTGCGATCCTTCGGTATTCGAA
GGAATGAGAAGATCGCCTGCTATGTACCGTGAGAGGGGAGAAAGCACTCCAGCTGCTTG
AGAGCGGACTGAAGGTGAAGGAATACGAATTGCTGCGCAGGAACCTTCAGCCAGACTGGAT
GTTTGGTTTGGATCAGACACATCGATTTGGGAATTAAGTATGAC
>Contig5
TCAGAAGCAGTGGTATCAACGCAGAGTACGGGGGGCTTTTGGTTGGGGGAGCGAGTGGG
CCGATCGACAGACGGGCGGGTTTGTGTCGTCGTCGTCGTCGTTGTCCCTGTTGTGGTACCTT
CGGTCTTGTGTTGCCGGGCTTTTGCATTTCCGCGGCAAGAAAGTGTAGCAGCCATGGCGTC
GGAGAAGAAGCAGTTGAACGTGATGAGGGAGATCAAAGTTCAGAAGCTTGTCTGAACAT
TTCCGTTGGAGAGAGTGGCGATAGGCTGACCCGTGCTTCCAAGGTGCTGGAGCAACTGAG
TGGCCAATCTCCCGTGTCTCGAAGGCCAGATACACCGTGCGATCCTTCGGTATTCGAAG
GAATGAGAAGATCGCCTGCTATGTACCGTGAGAGGGGAGAAAGCACTCCAGCTGCTTGA
GAGCGGACTGAAGGTGAAGGAATACGAATTGCTGCGCAGGAACCTTCAGCCAGACTGGATG
TTTGTGTTGAACCAAAACACATCGATTTGGGAATTAAGTA
>Contig6
allSRR.hitid.fa.cap.contigs
```





tfasty -m 8 CslproteinsequencesBMC allSRR.hitid.fa.contigs > allSRR.contigs.BMC

Total hits 9,838

```
# TFASTY 36.3.5e Nov, 2012(preload8)
# Query: AT2G21770.1|AT2G21770.1|cesA - 1089 aa
# Database: allSRR.hitid.fa.cap.contigs
AT2G21770.1|AT2G21770.1|cesA   Contig37   52.46   305    122    25     506    804    871    6      5.9e-60 222.2
AT2G21770.1|AT2G21770.1|cesA   Contig42   51.59   157    68     12     919   1068    20    485    6.4e-33 132.9
AT2G21770.1|AT2G21770.1|cesA   Contig26   38.33    60    32      6     640    694    355    533    3.4e-06 42.9
AT2G21770.1|AT2G21770.1|cesA   Contig17   50.00    36    18      1     640    675    355    461    7e-06   42.1
AT2G21770.1|AT2G21770.1|cesA   Contig10   25.00    56    38      4     641    693    102    266    9e-05   37.6
AT2G21770.1|AT2G21770.1|cesA   Contig39   32.84    67    36     10     43    107    442    262    0.0016  34.1
AT2G21770.1|AT2G21770.1|cesA   Contig31   29.41    34    23      1     670    703    424    326    0.0035  32.8
AT2G21770.1|AT2G21770.1|cesA   Contig29   37.04    27    11      6     666    690    345    413    0.0064  31.8
AT2G21770.1|AT2G21770.1|cesA   Contig8 27.63   76     46     9    431    505    67    270    0.0071  31.8
AT2G21770.1|AT2G21770.1|cesA   Contig7 34.67   75     47     5    602    675    256    478    0.0074  32.3
AT2G21770.1|AT2G21770.1|cesA   Contig38   34.78    23    15      0     779    801    75     7    0.0091  31.6
AT2G21770.1|AT2G21770.1|cesA   Contig30   30.88    68    45      3     609    675    470    269    0.0099  31.3
AT2G21770.1|AT2G21770.1|cesA   Contig44   35.29    34    21      2     670    703    195    294    0.024   30.1
AT2G21770.1|AT2G21770.1|cesA   Contig15   29.63    81    45     13     602    675    497    269    0.034   29.6
AT2G21770.1|AT2G21770.1|cesA   Contig33   20.90   134    96     10     458    586    58    444    0.091   28.1
AT2G21770.1|AT2G21770.1|cesA   Contig43   31.75    63    39      6      61    123    519    345    0.095   28.3
AT2G21770.1|AT2G21770.1|cesA   Contig9 47.62   21     10     2    301    321    525    467    0.13    27.8
AT2G21770.1|AT2G21770.1|cesA   Contig25   66.67     6     2     0     666    671    227    210    0.17    27.3
AT2G21770.1|AT2G21770.1|cesA   Contig6 24.56   57     42     1    659    715    60    227    0.27    26.6
AT2G21770.1|AT2G21770.1|cesA   Contig5 26.21   103    68     9    642    736    10    319    0.27    26.6
AT2G21770.1|AT2G21770.1|cesA   Contig4 31.71   41     21     8     56    89    126    247    0.27    26.6
AT2G21770.1|AT2G21770.1|cesA   Contig10   36.36    33    21      0    170    202    320    222    0.53    25.1
AT2G21770.1|AT2G21770.1|cesA   Contig6 24.44   45     27     8     61    105    191    79    0.63    25.3
AT2G21770.1|AT2G21770.1|cesA   Contig9 42.42   33     15     5     76    104    522    619    0.64    25.6
AT2G21770.1|AT2G21770.1|cesA   Contig9 19.79   96     64    13    420    509    167    433    0.54    25.8
AT2G21770.1|AT2G21770.1|cesA   Contig9 35.71   14      9     0    919    932    68    109    0.54    25.8
AT2G21770.1|AT2G21770.1|cesA   Contig30   24.19    62    45      3      4     65    295    473    0.76    25.1
AT2G21770.1|AT2G21770.1|cesA   Contig43   29.21    89    62      3    939   1026    220    486    0.9     25.1
AT2G21770.1|AT2G21770.1|cesA   Contig12   45.83    24    13      1     85    108    417    487    0.95    24.8
AT2G21770.1|AT2G21770.1|cesA   Contig28   25.49    51    33      6    653    703    346    482    1.2     24.3
AT2G21770.1|AT2G21770.1|cesA   Contig41   25.00    60    42      5    321    377    302    125    1.2     24.6
AT2G21770.1|AT2G21770.1|cesA   Contig41   66.67     6     2     0    663    668    361    344    1       24.8
AT2G21770.1|AT2G21770.1|cesA   Contig22   42.11    38    22      2     11     48    556    443    1.5     24.3
AT2G21770.1|AT2G21770.1|cesA   Contig22   24.14    58    41      3    141    198    296    132    1.1     24.8
AT2G21770.1|AT2G21770.1|cesA   Contig13   27.59    58    34      9    820    869    758    586    1.6     24.6
AT2G21770.1|AT2G21770.1|cesA   Contig13   29.03    31    22      0    487    517    414    322    0.56    26.1
AT2G21770.1|AT2G21770.1|cesA   Contig13   37.04    27    17      1    792    818    150    71    1.1     25.1
AT2G21770.1|AT2G21770.1|cesA   Contig13   36.67    30    18      1    227    256    320    234    1.1     25.1
AT2G21770.1|AT2G21770.1|cesA   Contig32   23.58   123    92      4    264    384    369    3    1.7     24.1
AT2G21770.1|AT2G21770.1|cesA   Contig32   38.89    18    11      0    132    149    522    469    1.4     24.3
AT2G21770.1|AT2G21770.1|cesA   Contig5 24.44   45     27     8     61    105    192    80    1.8     23.8
AT2G21770.1|AT2G21770.1|cesA   Contig5 45.45   22     12     0    266    287    398    333    0.38    26.1
AT2G21770.1|AT2G21770.1|cesA   Contig4 26.67   45     26     8     61    105    193    81    1.8     23.8
AT2G21770.1|AT2G21770.1|cesA   Contig4 45.45   22     12     0    266    287    399    334    0.38    26.1
AT2G21770.1|AT2G21770.1|cesA   Contig46   15.00    60    48      4    321    377    333    511    1.8     24.1
AT2G21770.1|AT2G21770.1|cesA   Contig19   17.39    46    38      0    662    707    216    79     2       23.6
# TFASTY 36.3.5e Nov, 2012(preload8)
# Query: AT2G25540.1|AT2G25540.1|cesA - 1066 aa
allSRR.contigs.BMC
```

# Results



## Method Identified

## Genes

tblastn SRR099268

5438

tblastn SRR099269

5404

tfasty SRR099268

4042

tfasty SRR099269

4174

tfasty+cap3 allSRR's

9838

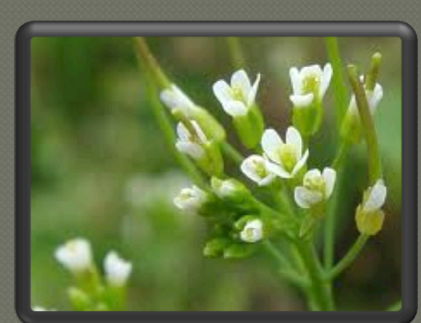
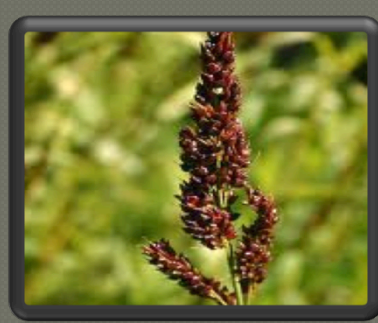
# The cap3 sequence assembly program

1. Uses forward-reverse constraints to correct assembly errors and link contigs.
2. Automatic clipping of 5' and 3' poor regions of reads.
3. Takes longer sequences of higher errors and produce more accurate consensus sequences.
4. Makes use of a large number of forward-reverse constraints to locate and correct errors in layout of sequence reads. This capability allows CAP3 to address assembly errors due to repeats.
5. The alignment method in CAP3 is very tolerable of reads of high sequencing errors.

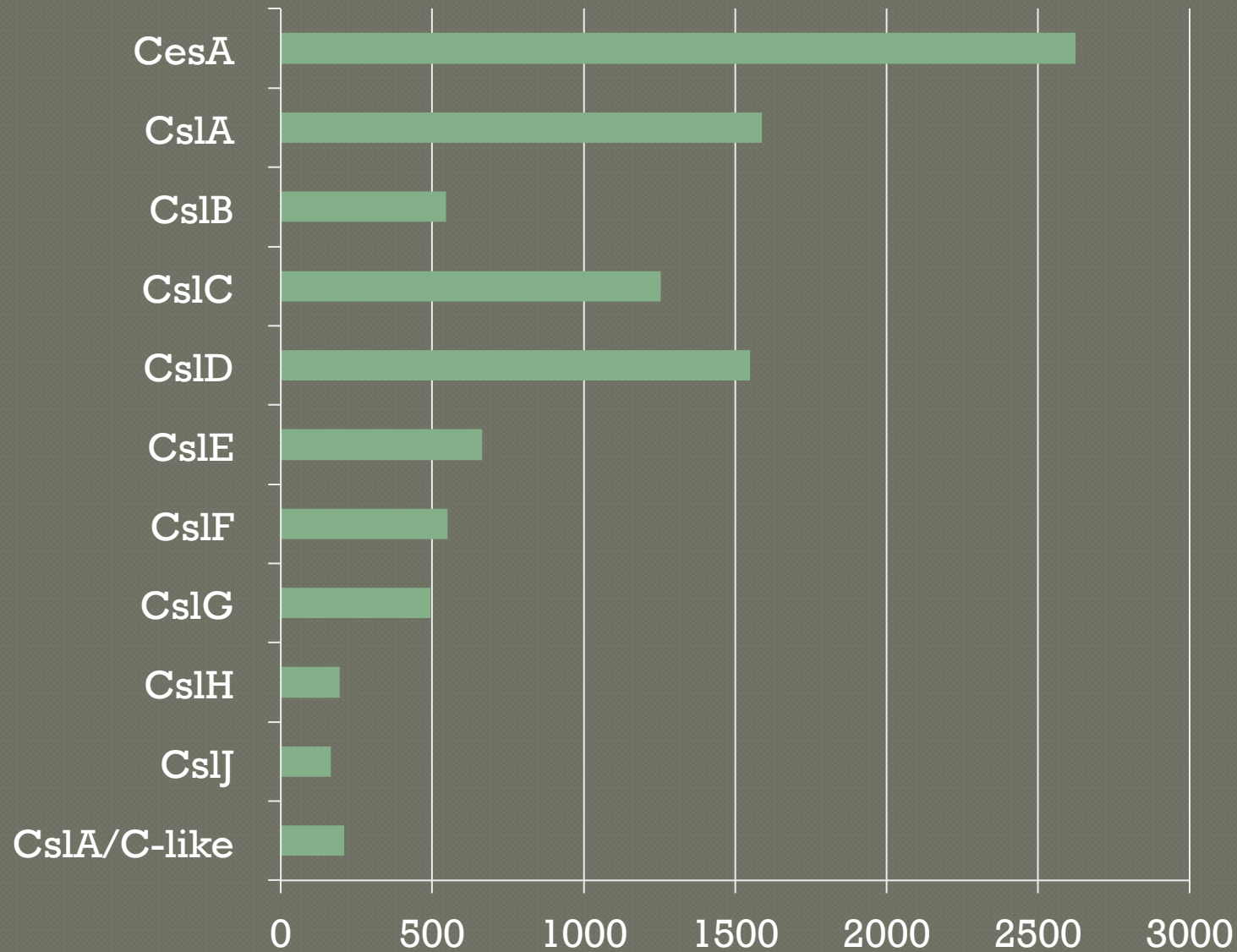


# Top 10 hit accessions using cap3 & tfasty

Index	Accession	Organism	Occurrences	Gene
1	os_37806 LOC_Os06g39970.1	Oryza sativa	67	CesA
2	sb_8735 e_gw1.2.5910.1	Sorghum bicolor	65	CsIA
3	AT4G15290.1 AT4G15290.1	Arabidopsis thaliana	57	CsIB
4	vv_13999 GSVIVP00019341001	Vitis vinifera	57	CesA
5	os_53309 LOC_Os09g39920.1	Oryza sativa	56	CsIA
6	AT4G13410.1 AT4G13410.1	Arabidopsis thaliana	55	CsIA
7	sb_8936 e_gw1.2.257.1	Sorghum bicolor	55	CsIC
8	AT1G02730.1 AT1G02730.1	Arabidopsis thaliana	54	CsID
9	sb_6547 estExt_Genewise1.C_chr_210573	Sorghum bicolor	52	CsIA
10	sb_12937 estExt_Genewise1Plus.C_chr_21690	Sorghum bicolor	51	CesA



## Distribution of 9,838 Csl/CesA genes in *Chara vulgaris* NGS 454 data





# Bioinformatics is the answer to our questions

- 1. In 2009, Yin *et al.* identified Csl homologs in fully sequenced lower green algae. To continue his research, we were to expand this search in NGS 454 data available on NCBI for *Chara vulgaris*
- 2. The 454 data was available because Wodnoick *et al.* in 2011 were doing research to determine the Origin of land plants: Do conjugating green algae hold the key?
- It has been widely accepted that Streptophytes (*Chara vulgaris*) are the closest living relative to the land plants.
- However, in this 2011 study, they found that the **Zygnematales** are most closely related to land plants.

# Future Bioinformatics Project

- Since, we identified every gene in cellulose synthase superfamily in *Chara vulgaris*, it would be very interesting to data mine the Zygnematales to see if the numbers of Csl and CesA genes would increase.
- Most Zygnematales live in freshwater, and form an important component of the algal scum that grows on or near plants, rocks, and various debris.

NCBI Resources How To

SRA SRA zygnematales Search SRA

Save search Limits Advanced

Display Settings: Summary

Results: 2

454 sequencing of *Spirogyra pratensis* transcriptome fragment library

1. 4 LS454 (454 GS FLX) runs: 614,139 spots, 164.4M bases, 326.2Mb downloads  
Accession: SRX017045

454 sequencing of *Coleochaete orbicularis* transcriptome fragment library

2. 1 LS454 (454 GS 20) run: 354,659 spots, 185.6M bases, 397.8Mb downloads  
Accession: SRX017046

Send to: Filter your results:

All (2)

access: Controlled (0)

access: Public (2)

aligned data (0)

source: DNA (0)

source: metagenomic

source: RNA (2)



# References

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990)  
**Basic local alignment search tool.** *J. Mol. Bio.* **215**: 403-410.

<http://etutorials.org/Misc/blast/Part+III+Practice/Chapter+9.+BLAST+Protocols/9.4+TBLASTN+Protocols/>

Huang, X. and Madan, A. (1999) **CAP3: A DNA sequence assembly program.**  
*Genome Res.*, **9**, 868-877.

Patel RK, Jain M (2012) **NGS QC Toolkit for Quality Control of Next Generation Sequencing Data.** PLoS ONE 7(2): e30619. doi:10.1371/journal.pone.0030619

Sabina Wodnick, Henner Brinkmann, Gernot Glockner, Andrew J. Heidel,  
Herve Philippe, Michael Melkonian, Burkhard Becker **Origin of land plants:  
Do conjugating green algae hold the key?** *BMC Evol Biol.* 201111: 104.

SRA Knowledge Base [internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-. **Using the SRA Toolkit.** Available from:  
<http://www.ncbi.nlm.nih.gov/books/NBK56560/>

W.R. Pearson and D.J. Lipman (1988), “**Improved Tools for Biological Sequence Analysis**”, *PNAS* **85**:244-2448.

Yanbin Yin, Jinling Hauang, Ying Xu **The Cellulose synthase superfamily in fully sequenced plants and algae.** *BMC Plant Biology* 2009, **9**:99