Tutorial 4

Reconstruction of cDNAs using SAGE data

1. Introduction

In Tutorial 2, we covered cDNAs reconstruction from EST datasets using either DNA or protein seeds. At that time, we used protein seeds of 33 residues, which is a relatively long size. SAGE is a mainstream method to obtain a digital expression profile and to perform comparative expression studies. Once the SAGE tags are quantified, and differentially expressed genes identified, tag mapping becomes a crucial step. SAGE tags are classically composed by a CATG punctuation site followed by a 10-base sequence. LongSAGE, an improvement of the classic technique, generates tags of 21 bases (CATG plus 17 additional bases). As we will demonstrate here, GenSeed can efficiently reconstruct full-length cDNAs using the short 14-base SAGE tags, allowing for tag mapping from fragmentary EST data. In this tutorial, we will use real-life SAGE data of *Toxoplasma gondii*, a protozoan pathogen of medical and veterinary relevance.

2. Seed sequences

We will use a set of 10 SAGE tags obtained for *T. gondii*, as reported by Radke *et al*. (The transcriptome of *Toxoplasma gondii*. *BMC Biol.* **3**: 26, 2005). These tags were annotated and the whole 300,000-tag database is publicly available at the TgSAGEDB site (http://vmbmod10.msu.montana.edu/vmb/white-lab/newsage.htm).

The `/tutorial_4/seed` directory contains 10 files, each one containing a SAGE tag sequence derived from this database: files `tag1.fasta` to `tag10.fasta` (names were given arbitrarily):

3. Database

We will use a database of *Toxoplasma gondii* containing 129,421 unclustered ESTs. You can download such database from the NCBI (http://www.ncbi.nlm.nih.gov) choosing EST database (dbEST) and `txid5810[Organism:exp]` as query. The total number of sequences may be changed by the time you download it. The database file should be saved in the `/tutorial_4/db` directory. We assume from now on that the database file is named `T_gondii_EST.fasta`.

4. Running GenSeed

A comprehensive explanation on all GenSeed parameters is depicted in the "GenSeed - Quick Guide" document. Please refer to it if you need more information.

We will now reconstruct the cDNAs corresponding to each one of the SAGE tags.

Go to the `/tutorial_4/test` directory and type the command below:

```
genseed.pl -s ../seed/tag1.fasta -d ../db/T_gondii_EST.fasta -o
output_tag1 -b "-e 10 -b 50 -F F" -g no -i 90 -j 90
```

Now repeat the same command, but change the name of the seed file and the corresponding output file (`tag2.fasta` and `output_tag2`, `tag3.fasta` and `output_tag3`, etc.)

If everything works well, ten new subdirectories will be created:

```
output_tag1
output_tag2
output_tag3
output_tag4
output_tag5
output_tag6
output_tag7
output_tag8
output_tag9
output_tag10
```

5. Understanding GenSeed parameters

A comprehensive explanation on all *GenSeed* parameters is depicted in the "GenSeed - Quick Guide" document. Please refer to it if you need more information.

Shortly, the command line used above specifies the following parameters:

- `-s ../seed/name.fasta` – seed sequence file with path
- `-d ../db/name.fasta` - database file with path
- `-o output_name` - as the output directory name
- `-b "-e 10 -b 50 -F F"` – these are parameters for BLAST. We use parameters `-e 10` to reduce the expectation value because the seed sequence is only 14-base long and, as such, is expected to produce alignments with a high e-value. The parameter `-b 50`, limits the number of top positive hits that will be selected on each round for the assembly process. Because the database is relatively big and coverage is highly redundant, we decided to limit the number of selected reads to improve performance. The parameter `-F F` corresponds to the low complexity filter turned off.
- `–g no` – do not use the last assembly consensus sequence as a template in the final assembly step. We turned this option off to reduce computational load. You can try re-running the reconstruction with `-g yes` option and compare the processing time.
- `-i 90` – Minimum length of the alignment block of the first BLAST search, in percentage relative of the length of the seed sequence. GenSeed will only select reads and proceed with the sequence reconstruction process if the seed has found matches in the database presenting alignment blocks with a length equal or higher than 90% of its own length

- `-j -90` – Minimum identity percentage of the alignment block of the first BLAST search. The read(s) will only be accepted if the alignment blocks present a percentage identity equal to or higher than 90%.

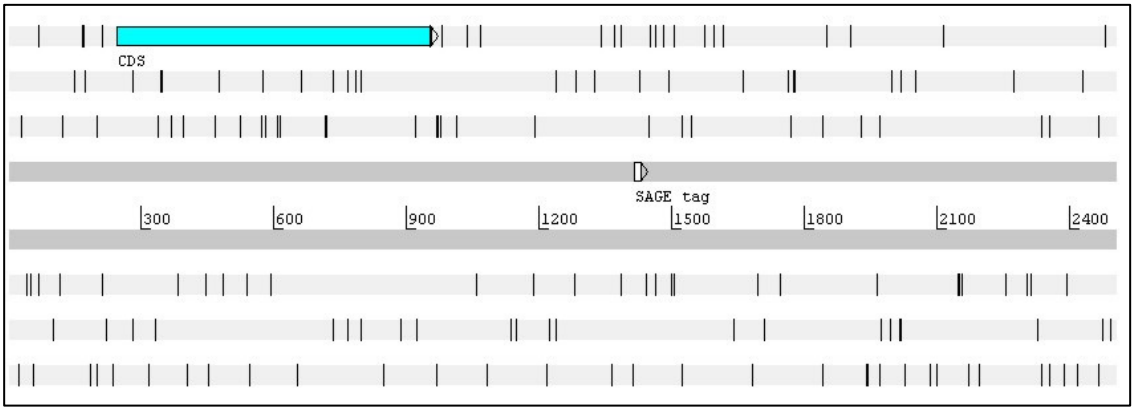
6. Inspecting GenSeed's output files

The table below shows the results for the 10 SAGE tags used as seeds. GenSeed succeeded to reconstruct the corresponding cDNAs with all seeds. The genes were identified using BLASTX similarity searches against nr database, and the results were in agreement with the annotated data from TgSAGEDB.

Seed ID	SAGE Sequence	cDNA size (bp)	cDNA product
tag1	CATGCCGACTGTGT	2503	gb|ABE69195.1| granule antigen protein GRA7 [Toxoplasma gondii]
tag 2	CATGACTCAGGTGC	2264	>gb|AAK20421.1|AF265362_1 glyceraldehyde-3-phosphate dehydrogenase [Toxoplasma gondii]
tag 3	CATGGTATCTCGAG	3115	>sp|P10878|TBB_TOXGO Tubulin beta chain (Beta-tubulin)
tag 4	CATGTAGGCGTAGT	2007	>gb|AAP41369.1|AF453384_1 gliding-associated protein 45 [Toxoplasma gondii]
tag 5	CATGGAAAATATAC	1442	>ref|XP_001349947.1| dihydrolipoamide succinyltransferase [Plasmodium falciparum 3D7]
tag 6	CATGGTAAACCAAT	2927	>sp|Q9XZD5|CATA_TOXGO Peroxisomal catalase
tag 7	CATGTGGCGAAAGT	2482	emb|CAB52368.1| microneme protein 5 [Eimeria tenella]
tag 8	CATGCAATCATCGC	1556	gb|AAM33361.1|AF509564_1 micronemal protein [Toxoplasma gondii]
tag 9	CATGTGGCTGCCTA	1887	gb|AAK26628.1|AF340229_1 surface antigen [Toxoplasma gondii]
tag 10	CATGAAACTTCCCT	2381	>gb|AAO65977.1| non-transmembrane antigen [Toxoplasma gondii]

Now let's have a look using an annotation viewer (Artemis) on some of the reconstructed sequences, and the BLASTX results of the reconstructed sequences against nr database.

Tag1

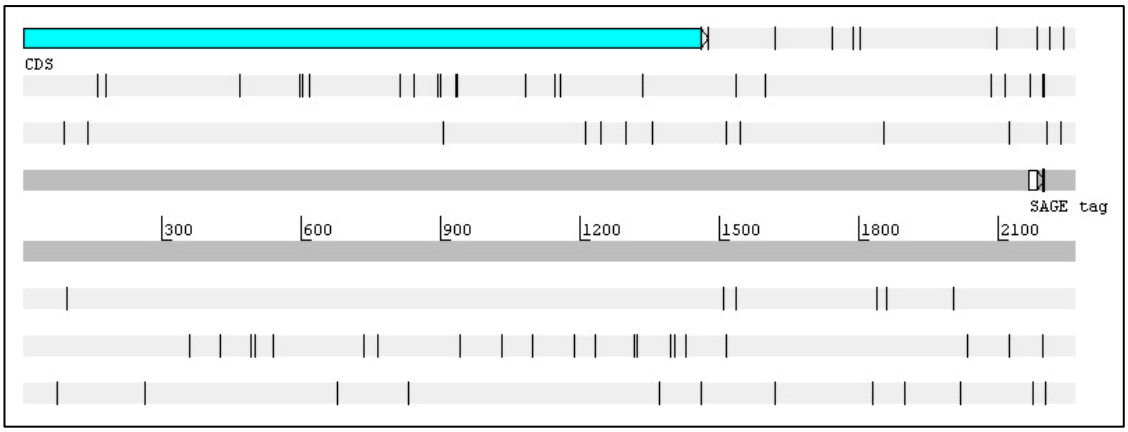


Screenshot of Artemis annotation editor, showing the reconstructed sequence using tag1 as a seed. The blue horizontal box represents the coding sequence and the white box the SAGE tag used as seed.

```
>gb|ABE69195.1| granule antigen protein GRA7 [Toxoplasma gondii]
          Length = 236

 Score =  360 bits (923), Expect = 3e-97
 Identities = 189/236 (80%), Positives = 189/236 (80%)
 Frame = +1

Query: 244 MARHAIFFXXXXXXXXXXXXXXXXXXXXSDDELMSRIRNSDFFDGQAPVDSLRPTNAGV 423
           MARHAIFF                    SDDELMSRIRNSDFFDGQAPVDSLRPTNAGV
Sbjct: 1   MARHAIFFALCVLGLVAAALPQFATAATASDDELMSRIRNSDFFDGQAPVDSLRPTNAGV 60

Query: 424 DSKGTDDHLTTSMDKASVESQLPRRXXXXXXXXXXXXXVHFRKRGVRSDAEVTDDNIYEEH 603
           DSKGTDDHLTTSMDKASVESQLPRR            VHFRKRGVRSDAEVTDDNIYEEH
Sbjct: 61  DSKGTDDHLTTSMDKASVESQLPRREPLETEPDEQEEVHFRKRGVRSDAEVTDDNIYEEH 120

Query: 604 TDRKVVPRKSEGKRSFKDLLKKLALPAVGMGASYFAADRILPELTEQQQTGEEPLTTGQN 783
           TDRKVVPRKSEGKRSFKDLLKKLALPAVGMGASYFAADRILPELTEQQQTGEEPLTTGQN
Sbjct: 121 TDRKVVPRKSEGKRSFKDLLKKLALPAVGMGASYFAADRILPELTEQQQTGEEPLTTGQN 180

Query: 784 VSTVXXXXXXXXXXXXXXMGLTRTYRHFSPRKNRSRQPALEQEVPESGKDGEDARQ 951
           VSTV              MGLTRTYRHFSPRKNRSRQPALEQEVPESGKDGEDARQ
Sbjct: 181 VSTVLGFAALAAAAAFLGMGLTRTYRHFSPRKNRSRQPALEQEVPESGKDGEDARQ 236
```

Sequence alignment of the best hit of sequence tag1 against nr database using BLASTX.

Tag2

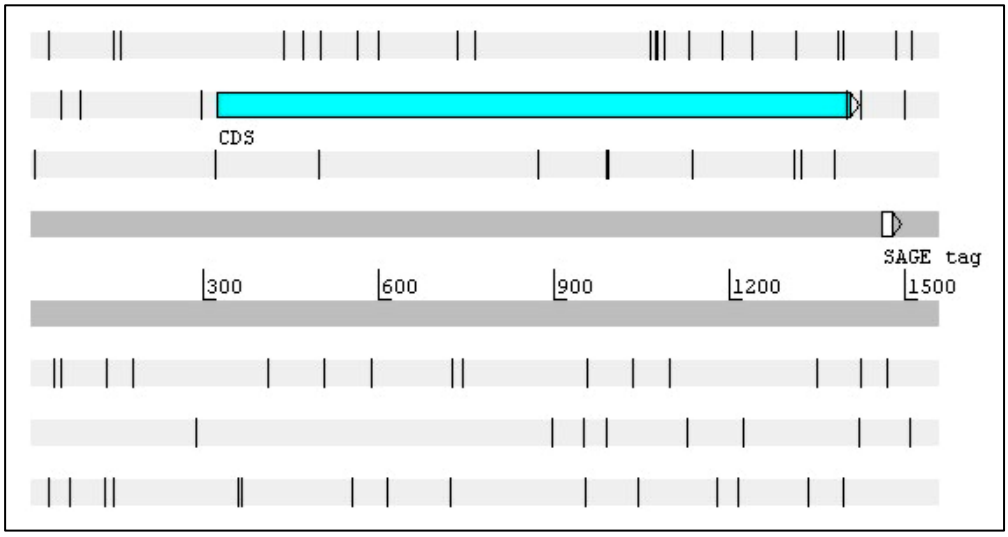


Screenshot of Artemis annotation editor, showing the reconstructed sequence using tag2 as a seed. The blue horizontal box represents the coding sequence and the white box the SAGE tag used as seed.

```
>gb|AAK20421.1|AF265362_1 glyceraldehyde-3-phosphate dehydrogenase
[Toxoplasma gondii]
          Length = 592

 Score =  916 bits (2368), Expect = 0.0
 Identities = 466/486 (95%), Positives = 466/486 (95%)
 Frame = +1

Query: 1    DGPAVLEHMARDGRGSASDLCSLVQTLATELLQTERDPRCAAAPTSEDRENIALTTEALL 180
             DGPAVLEHMARDGRGSASDLCSLVQTLATELLQTERDPRCAAAPTSEDRENIALTTEALL
Sbjct: 107  DGPAVLEHMARDGRGSASDLCSLVQTLATELLQTERDPRCAAAPTSEDRENIALTTEALL 166

Query: 181  TSAFGFLXXXXXXXXXXXXXXXXXXXXQNARSASIRSHRRNSFAPTGKRAVAPVPRVSPT 360
             TSAFGFL                    QNARSASIRSHRRNSFAPTGKRAVAPVPRVSPT
Sbjct: 167  TSAFGFLGPASPSATAATAATVASGTPQNARSASIRSHRRNSFAPTGKRAVAPVPRVSPT 226

Query: 361  GLFGLLGSSSEKASAPIRLGINGMGRIGRLVFRIAMSRPDVAVTHINCSMDPAYIAYMLK 540
             GLFGLLGSSSEKASAPIRLGINGMGRIGRLVFRIAMSRPDVAVTHINCSMDPAYIAYMLK
Sbjct: 227  GLFGLLGSSSEKASAPIRLGINGMGRIGRLVFRIAMSRPDVAVTHINCSMDPAYIAYMLK 286

Query: 541  YDSVHGKFDGEIVPTETSLIVNGQEVTISNTRDPEEIPWADKGADYVCESTGVFCTTEAA 720
             YDSVHGKFDGEIVPTETSLIVNGQEVTISNTRDPEEIPWADKGADYVCESTGVFCTTEAA
Sbjct: 287  YDSVHGKFDGEIVPTETSLIVNGQEVTISNTRDPEEIPWADKGADYVCESTGVFCTTEAA 346

Query: 721  AKHVNRPGGAKHAIISAPAKDETTPTLVVGVNAEQDYESSMKVVSCASCTTNGLAPLVKV 900
             AKHVNRPGGAKHAIISAPAKDETTPTLVVGVNAEQDYESSMKVVSCASCTTNGLAPLVKV
Sbjct: 347  AKHVNRPGGAKHAIISAPAKDETTPTLVVGVNAEQDYESSMKVVSCASCTTNGLAPLVKV 406

Query: 901  IDENFGLVEGLMTTVHAATGTQKVVDGTSKKDWRGGRAAAGNIIPSATGAAKAVARCLPH 1080
             IDENFGLVEGLMTTVHAATGTQKVVDGTSKKDWRGGRAAAGNIIPSATGAAKAVARCLPH
Sbjct: 407  IDENFGLVEGLMTTVHAATGTQKVVDGTSKKDWRGGRAAAGNIIPSATGAAKAVARCLPH 466

Query: 1081 MKGKLTGMAFRVPTLDVSVVDLTCRLNKSTTYEEIKKAVREASETYMRGIIGYTEEPIVS 1260
             MKGKLTGMAFRVPTLDVSVVDLTCRLNKSTTYEEIKKAVREASETYMRGIIGYTEEPIVS
Sbjct: 467  MKGKLTGMAFRVPTLDVSVVDLTCRLNKSTTYEEIKKAVREASETYMRGIIGYTEEPIVS 526

Query: 1261 QDIVGSQCSTVFDANAGIMLNPNFVKLVSWYDNEYAYSARLVDLIAVMAAKDGVVSPGTG 1440
             QDIVGSQCSTVFDANAGIMLNPNFVKLVSWYDNEYAYSARLVDLIAVMAAKDGVVSPGTG
Sbjct: 527  QDIVGSQCSTVFDANAGIMLNPNFVKLVSWYDNEYAYSARLVDLIAVMAAKDGVVSPGTG 586

Query: 1441 LDRRPF 1458
             LDRRPF
Sbjct: 587  LDRRPF 592
```

Sequence alignment of the best hit of sequence tag2 against nr database using BLASTX.

Tag8

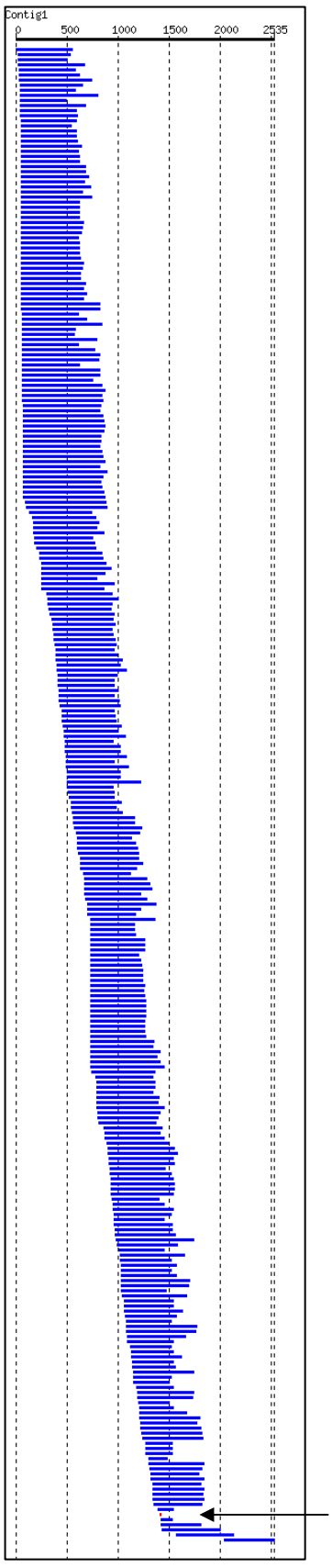


Screenshot of Artemis annotation editor, showing the reconstructed sequence using tag8 as a seed. The blue horizontal box represents the coding sequence and the white box the SAGE tag used as seed.

```
>gb|AAM33361.1|AF509564_1 micronemal protein [Toxoplasma gondii]
 emb|CAB56644.1| MIC3 microneme protein [Toxoplasma gondii]
          Length = 359

 Score =  732 bits (1890), Expect = 0.0
 Identities = 342/359 (95%), Positives = 342/359 (95%)
 Frame = +2

Query: 323  MRGGTSALLHALTFSGAVWMCTPAEALPIQKSVQLGSFDKVVPSREVVSESLAPSFAVTE 502
            MRGGTSALLHALTFSGAVWMCTPAEALPIQKSVQLGSFDKVVPSREVVSESLAPSFAVTE
Sbjct: 1    MRGGTSALLHALTFSGAVWMCTPAEALPIQKSVQLGSFDKVVPSREVVSESLAPSFAVTE 60

Query: 503  THSSVQSPSKQETQLCAISSEGKPCRNRQLHTDNGYFIGASCPKSACCSKTMCGPGGCGE 682
            THSSVQSPSKQETQLCAISSEGKPCRNRQLHTDNGYFIGASCPKSACCSKTMCGPGGCGE
Sbjct: 61   THSSVQSPSKQETQLCAISSEGKPCRNRQLHTDNGYFIGASCPKSACCSKTMCGPGGCGE 120

Query: 683  FCSSNWIFCSSSLIYHPDKSYGGDCSCEKQGHRCDKNAECVENLDAGGGVHCKCKDGFVG 862
            FCSSNWIFCSSSLIYHPDKSYGGDCSCEKQGHRCDKNAECVENLDAGGGVHCKCKDGFVG
Sbjct: 121  FCSSNWIFCSSSLIYHPDKSYGGDCSCEKQGHRCDKNAECVENLDAGGGVHCKCKDGFVG 180

Query: 863  TGLTCSEDPCSKRGNAKCGPNGTCIVVDSVSYTCTCGDGETLVNLPEGGQGCKRTGCHAF 1042
            TGLTCSEDPCSKRGNAKCGPNGTCIVVDSVSYTCTCGDGETLVNLPEGGQGCKRTGCHAF
Sbjct: 181  TGLTCSEDPCSKRGNAKCGPNGTCIVVDSVSYTCTCGDGETLVNLPEGGQGCKRTGCHAF 240

Query: 1043 RENCSPGRCIDDASHENGYTCECPTGYSREVTSKAEESCVEGVEVTLAEKCEKEFGISAS 1222
            RENCSPGRCIDDASHENGYTCECPTGYSREVTSKAEESCVEGVEVTLAEKCEKEFGISAS
Sbjct: 241  RENCSPGRCIDDASHENGYTCECPTGYSREVTSKAEESCVEGVEVTLAEKCEKEFGISAS 300

Query: 1223 SCKCDNGYSGSASATXXXXXXXXXXXXXXXXXMNIVFKCPSGYHPRYHAHTVTCEKIKQ 1399
            SCKCDNGYSGSASAT                 MNIVFKCPSGYHPRYHAHTVTCEKIKQ
Sbjct: 301  SCKCDNGYSGSASATSHHGKGESGSEGSLSEKMNIVFKCPSGYHPRYHAHTVTCEKIKQ 359
```

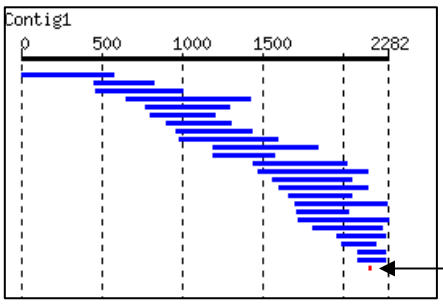
Sequence alignment of the best hit of sequence tag8 against nr database using BLASTX.

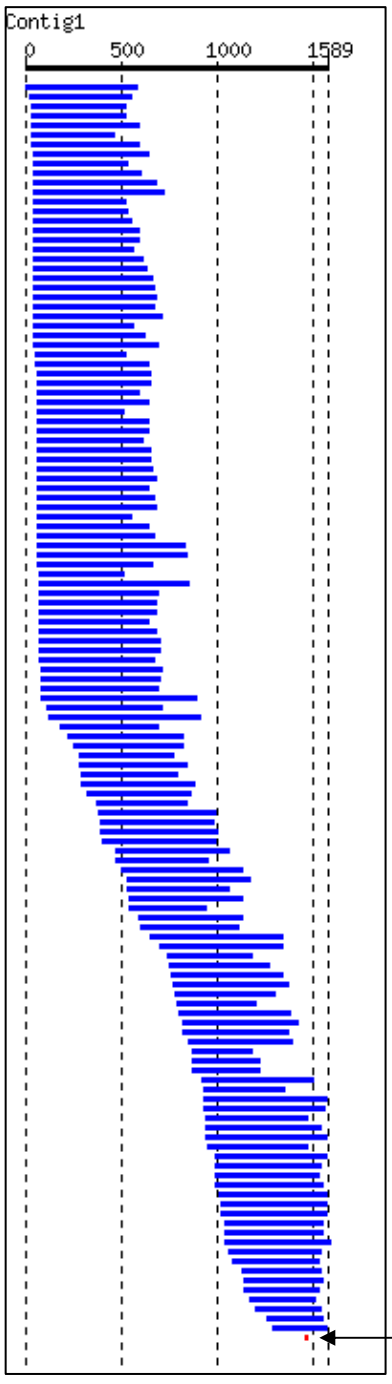
Let's now visualize the assembly of each gene by opening the `report.html` files of each directory. The figures below represent the assemblies, with each horizontal blue bar representing a read. The arrows, when present, point to the seed (red bars).



Tag1

Tag2





Tag8