De novo transcriptome sequence assembly (454/Sanger ESTs)

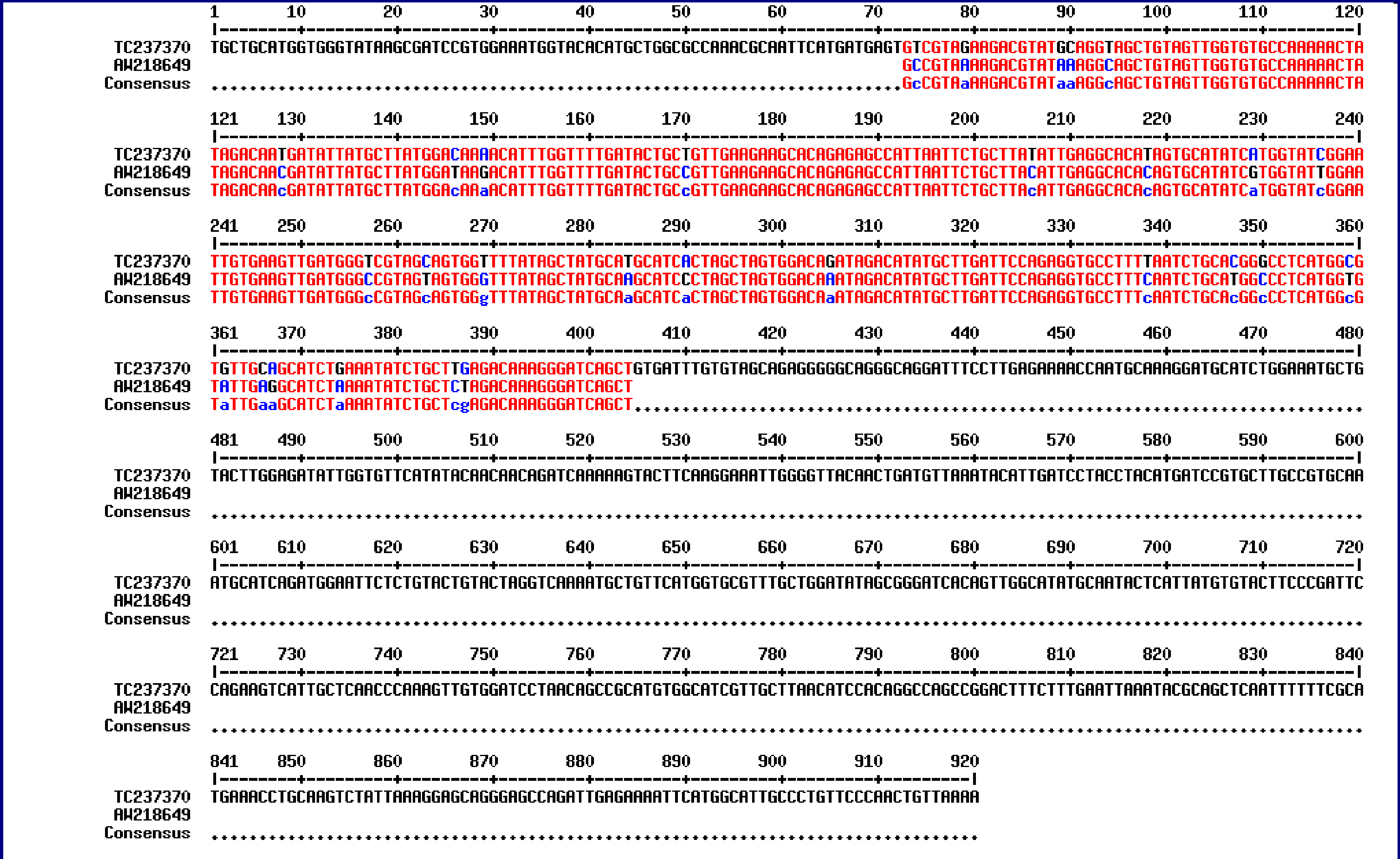
CAP3 (http://seq.cs.iastate.edu/cap3.html)
TGICL (http://compbio.dfci.harvard.edu/tgi/software/)
MIRA (http://www.chevreux.org/projects_mira.html)
Phrap (http://www.phrap.org)
Newbler (-cDNA)

Two major problems in existing EST assembly programs and unigene databases:

1) Large portion of different transcripts (mainly alternative spliced transcripts and paralogs) are incorrectly assembled into same transcripts – type I error
2) Large portion of nearly identical sequences are not assembled into one transcript – type II error

Example of type I assembly error (paralog)

In DFCI Tomato Gene Index, AW218649 is a member of TC237370.

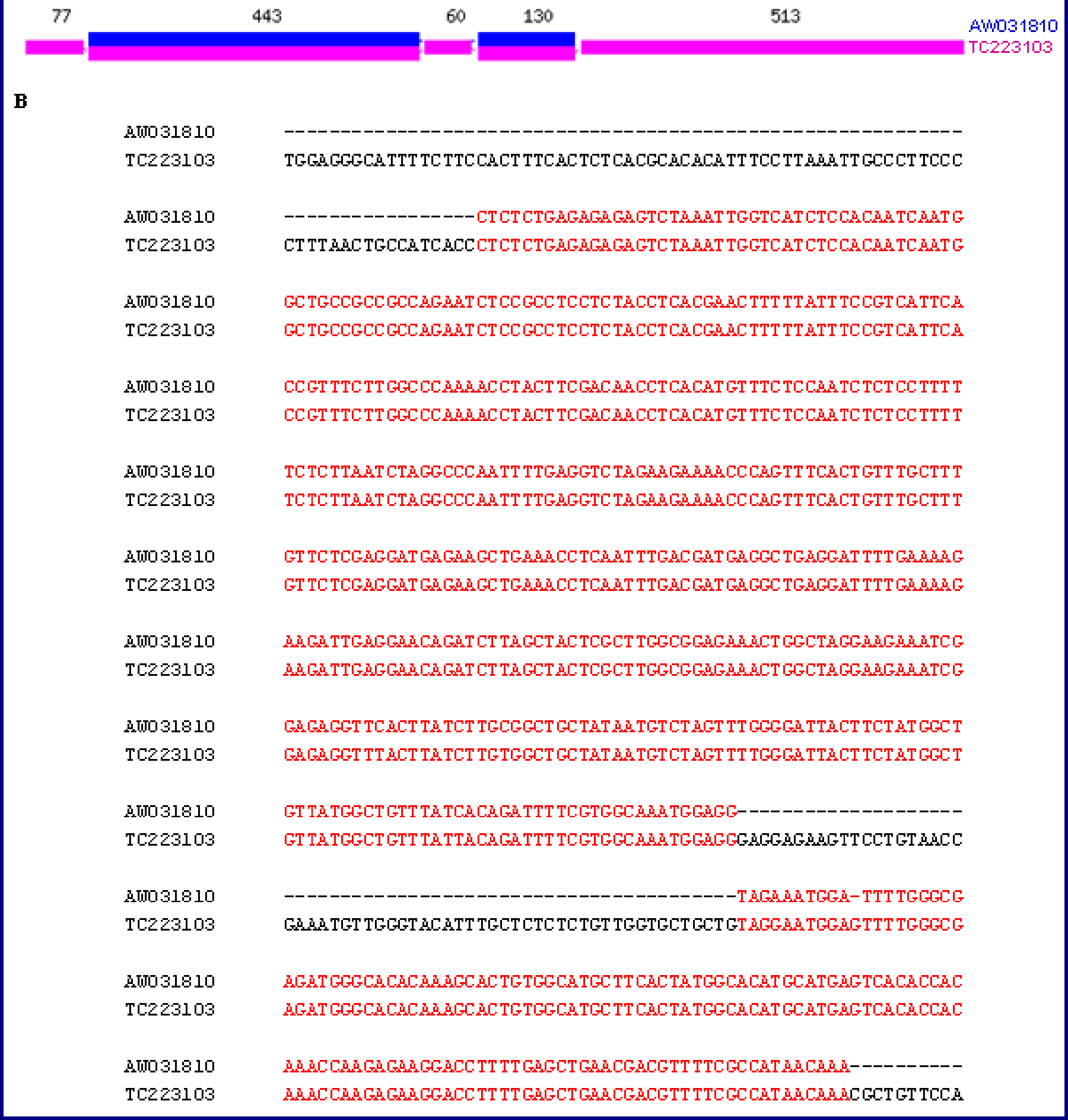


```
            1       10        20        30        40        50        60        70        80        90       100       110       120
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  TGCTGCATGGTGGGTATAAGCGATCCGTGGAAATGGTACACATGCTGGCGCCAAACGCAATTCATGATGAGTGTCGTAGAAGACGTATGCAGGTAGCTGTAGTTGGTGTGCCAAAAACTA
  AW218649                                                                          GCCGTAAAAGACGTATAAAGGCAGCTGTAGTTGGTGTGCCAAAAACTA
 Consensus  ........................................................................GcCGTAaAAGACGTATaaAGGcAGCTGTAGTTGGTGTGCCAAAAACTA

            121    130       140       150       160       170       180       190       200       210       220       230       240
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  TAGACAATGATATTATGCTTATGGACAAAACATTTGGTTTTGATACTGCTGTTGAAGAAGCACAGAGAGCCATTAATTCTGCTTATATTGAGGCACATAGTGCATATCATGGTATCGGAA
  AW218649  TAGACAACGATATTATGCTTATGGATAAGACATTTGGTTTTGATACTGCCGTTGAAGAAGCACAGAGAGCCATTAATTCTGCTTACATTGAGGCACACAGTGCATATCGTGGTATTGGAA
 Consensus  TAGACAAcGATATTATGCTTATGGAcAAaACATTTGGTTTTGATACTGCcGTTGAAGAAGCACAGAGAGCCATTAATTCTGCTTAcATTGAGGCACAcAGTGCATATCaTGGTATcGGAA

            241    250       260       270       280       290       300       310       320       330       340       350       360
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  TTGTGAAGTTGATGGGTCGTAGCAGTGGTTTTATAGCTATGCATGCATCACTAGCTAGTGGACAGATAGACATATGCTTGATTCCAGAGGTGCCTTTTAATCTGCACGGGCCTCATGGCG
  AW218649  TTGTGAAGTTGATGGGCCGTAGTAGTGGGTTTATAGCTATGCAAGCATCCCTAGCTAGTGGACAAATAGACATATGCTTGATTCCAGAGGTGCCTTTCAATCTGCATGGCCCTCATGGTG
 Consensus  TTGTGAAGTTGATGGGcCGTAGcAGTGGgTTTATAGCTATGCAaGCATCaCTAGCTAGTGGACAaATAGACATATGCTTGATTCCAGAGGTGCCTTTcAATCTGCAcGGcCCTCATGGcG

            361    370       380       390       400       410       420       430       440       450       460       470       480
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  TGTTGCAGCATCTGAAATATCTGCTTGAGACAAAGGGATCAGCTGTGATTTGTGTAGCAGAGGGGGCAGGGCAGGATTTCCTTGAGAAAACCAATGCAAAGGATGCATCTGGAAATGCTG
  AW218649  TATTGAGGCATCTAAAATATCTGCTCTAGACAAAGGGATCAGCT
 Consensus  TaTTGaaGCATCTaAAATATCTGCTcgAGACAAAGGGATCAGCT............................................................................

            481    490       500       510       520       530       540       550       560       570       580       590       600
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  TACTTGGAGATATTGGTGTTCATATACAACAACAGATCAAAAAGTACTTCAAGGAAATTGGGGTTACAACTGATGTTAAATACATTGATCCTACCTACATGATCCGTGCTTGCCGTGCAA
  AW218649
 Consensus  ........................................................................................................................

            601    610       620       630       640       650       660       670       680       690       700       710       720
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  ATGCATCAGATGGAATTCTCTGTACTGTACTAGGTCAAAATGCTGTTCATGGTGCGTTTGCTGGATATAGCGGGATCACAGTTGGCATATGCAATACTCATTATGTGTACTTCCCGATTC
  AW218649
 Consensus  ........................................................................................................................

            721    730       740       750       760       770       780       790       800       810       820       830       840
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  CAGAAGTCATTGCTCAACCCAAAGTTGTGGATCCTAACAGCCGCATGTGGCATCGTTGCTTAACATCCACAGGCCAGCCGGACTTTCTTTGAATTAAATACGCAGCTCAATTTTTTCGCA
  AW218649
 Consensus  ........................................................................................................................

            841    850       860       870       880       890       900       910       920
            |--------+---------+---------+---------+---------+---------+---------+---------|
  TC237370  TGAAACCTGCAAGTCTATTAAAGGAGCAGGGAGCCAGATTGAGAAAATTCATGGCATTGCCCTGTTCCCAACTGTTAAAA
  AW218649
 Consensus  ................................................................................
```

Sequence identity between AW218649 and TC232370: 91.5%
AW218649 is aligned to tomato chromosome 4
TC237370 is aligned to tomato chromosome 11

Example of type I assembly error (alternative splicing)

In DFCI Tomato Gene Index, AW031810 is a member of TC223103

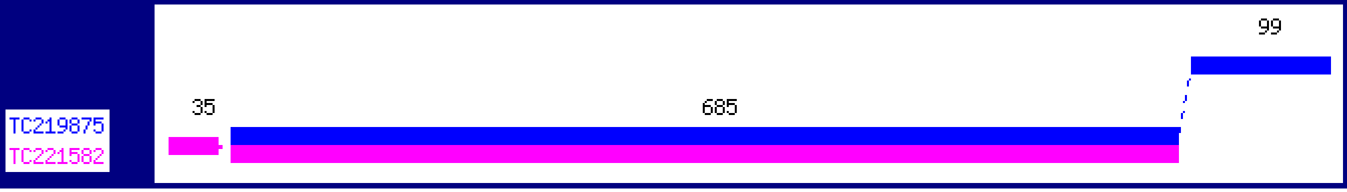


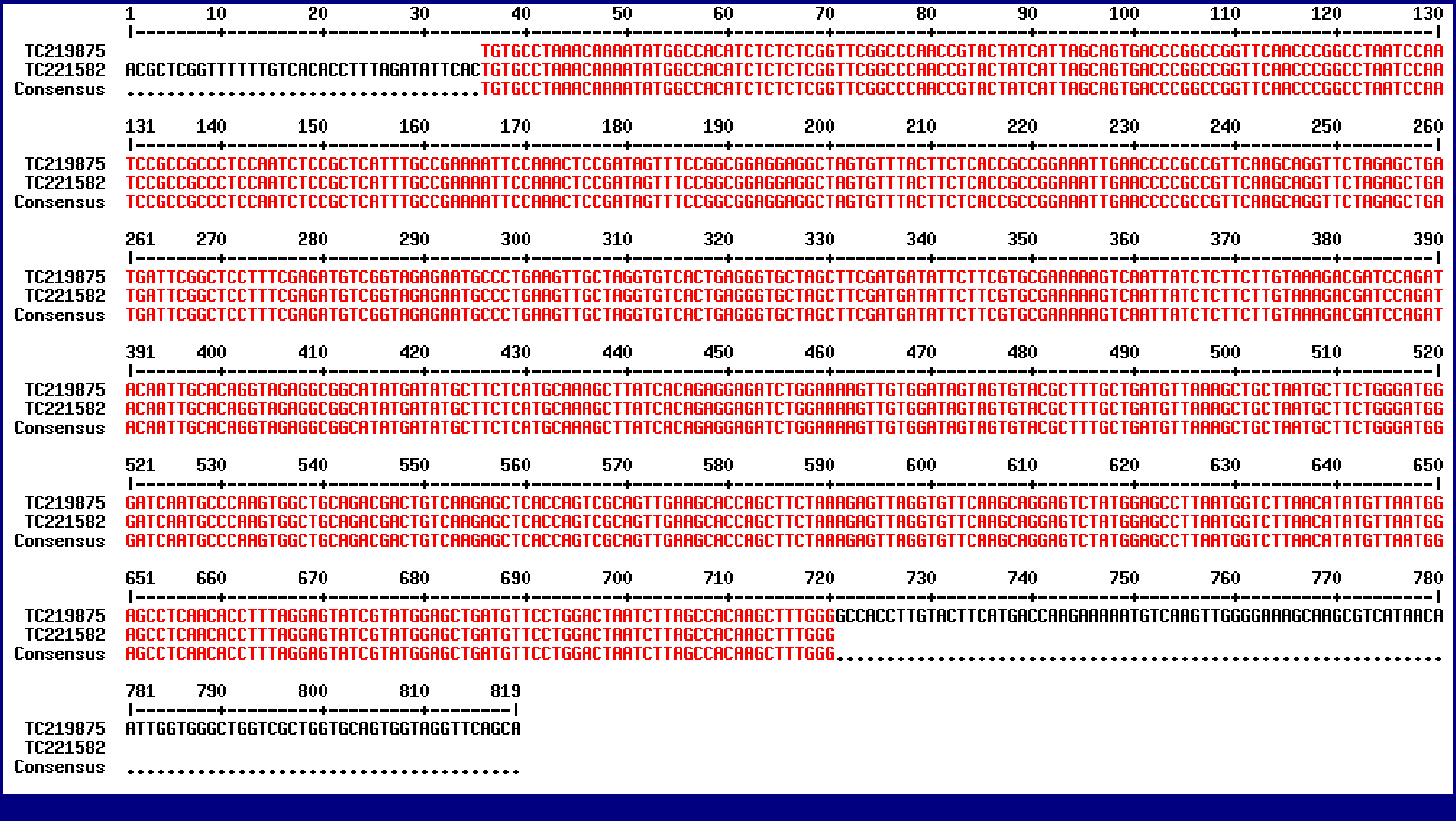
B

```
AW031810     ------------------------------------------------------------
TC223103     TGGAGGGCATTTTCTTCCACTTTCACTCTCACGCACACATTTCCTTAAATTGCCCTTCCC

AW031810     -----------------CTCTCTGAGAGAGAGTCTAAATTGGTCATCTCCACAATCAATG
TC223103     CTTTAACTGCCATCACCCTCTCTGAGAGAGAGTCTAAATTGGTCATCTCCACAATCAATG

AW031810     GCTGCCGCCGCCAGAATCTCCGCCTCCTCTACCTCACGAACTTTTTATTTCCGTCATTCA
TC223103     GCTGCCGCCGCCAGAATCTCCGCCTCCTCTACCTCACGAACTTTTTATTTCCGTCATTCA

AW031810     CCGTTTCTTGGCCCAAAACCTACTTCGACAACCTCACATGTTTCTCCAATCTCTCCTTTT
TC223103     CCGTTTCTTGGCCCAAAACCTACTTCGACAACCTCACATGTTTCTCCAATCTCTCCTTTT

AW031810     TCTCTTAATCTAGGCCCAATTTTGAGGTCTAGAAGAAAACCCAGTTTCACTGTTTGCTTT
TC223103     TCTCTTAATCTAGGCCCAATTTTGAGGTCTAGAAGAAAACCCAGTTTCACTGTTTGCTTT

AW031810     GTTCTCGAGGATGAGAAGCTGAAACCTCAATTTGACGATGAGGCTGAGGATTTTGAAAAG
TC223103     GTTCTCGAGGATGAGAAGCTGAAACCTCAATTTGACGATGAGGCTGAGGATTTTGAAAAG

AW031810     AAGATTGAGGAACAGATCTTAGCTACTCGCTTGGCGGAGAAACTGGCTAGGAAGAAATCG
TC223103     AAGATTGAGGAACAGATCTTAGCTACTCGCTTGGCGGAGAAACTGGCTAGGAAGAAATCG

AW031810     GAGAGGTTCACTTATCTTGCGGCTGCTATAATGTCTAGTTTGGGGATTACTTCTATGGCT
TC223103     GAGAGGTTTACTTATCTTGTGGCTGCTATAATGTCTAGTTTTGGGATTACTTCTATGGCT

AW031810     GTTATGGCTGTTTATCACAGATTTTCGTGGCAAATGGAGG--------------------
TC223103     GTTATGGCTGTTTATTACAGATTTTCGTGGCAAATGGAGGGAGGAGAAGTTCCTGTAACC

AW031810     -----------------------------------------TAGAAATGGA-TTTTGGCG
TC223103     GAAATGTTGGGTACATTTGCTCTCTCTGTTGGTGCTGCTGTAGGAATGGAGTTTTGGCG

AW031810     AGATGGGCACACAAAGCACTGTGGCATGCTTCACTATGGCACATGCATGAGTCACACCAC
TC223103     AGATGGGCACACAAAGCACTGTGGCATGCTTCACTATGGCACATGCATGAGTCACACCAC

AW031810     AAACCAAGAGAAGGACCTTTTGAGCTGAACGACGTTTTCGCCATAACAAA----------
TC223103     AAACCAAGAGAAGGACCTTTTGAGCTGAACGACGTTTTCGCCATAACAAACGCTGTTCCA
```

Example of type II assembly error

In DFCI Tomato Gene Index, two unigenes, TC219875 and TC221582, are identical





```
           1       10        20        30        40        50        60        70        80        90        100       110       120       130
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 TC219875                                     TGTGCCTAAACAAAATATGGCCACATCTCTCTCGGTTCGGCCCAACCGTACTATCATTAGCAGTGACCCGGCCGGTTCAACCCGGCCTAATCCAA
 TC221582  ACGCTCGGTTTTTTGTCACACCTTTAGATATTCACTGTGCCTAAACAAAATATGGCCACATCTCTCTCGGTTCGGCCCAACCGTACTATCATTAGCAGTGACCCGGCCGGTTCAACCCGGCCTAATCCAA
Consensus  ...................................TGTGCCTAAACAAAATATGGCCACATCTCTCTCGGTTCGGCCCAACCGTACTATCATTAGCAGTGACCCGGCCGGTTCAACCCGGCCTAATCCAA

           131    140       150       160       170       180       190       200       210       220       230       240       250       260
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 TC219875  TCCGCCGCCCTCCAATCTCCGCTCATTTGCCGAAAATTCCAAACTCCGATAGTTTCCGGCGGAGGAGGCTAGTGTTTACTTCTCACCGCCGGAAATTGAACCCCGCCGTTCAAGCAGGTTCTAGAGCTGA
 TC221582  TCCGCCGCCCTCCAATCTCCGCTCATTTGCCGAAAATTCCAAACTCCGATAGTTTCCGGCGGAGGAGGCTAGTGTTTACTTCTCACCGCCGGAAATTGAACCCCGCCGTTCAAGCAGGTTCTAGAGCTGA
Consensus  TCCGCCGCCCTCCAATCTCCGCTCATTTGCCGAAAATTCCAAACTCCGATAGTTTCCGGCGGAGGAGGCTAGTGTTTACTTCTCACCGCCGGAAATTGAACCCCGCCGTTCAAGCAGGTTCTAGAGCTGA

           261    270       280       290       300       310       320       330       340       350       360       370       380       390
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 TC219875  TGATTCGGCTCCTTTCGAGATGTCGGTAGAGAATGCCCTGAAGTTGCTAGGTGTCACTGAGGGTGCTAGCTTCGATGATATTCTTCGTGCGAAAAAGTCAATTATCTCTTCTTGTAAAGACGATCCAGAT
 TC221582  TGATTCGGCTCCTTTCGAGATGTCGGTAGAGAATGCCCTGAAGTTGCTAGGTGTCACTGAGGGTGCTAGCTTCGATGATATTCTTCGTGCGAAAAAGTCAATTATCTCTTCTTGTAAAGACGATCCAGAT
Consensus  TGATTCGGCTCCTTTCGAGATGTCGGTAGAGAATGCCCTGAAGTTGCTAGGTGTCACTGAGGGTGCTAGCTTCGATGATATTCTTCGTGCGAAAAAGTCAATTATCTCTTCTTGTAAAGACGATCCAGAT

           391    400       410       420       430       440       450       460       470       480       490       500       510       520
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 TC219875  ACAATTGCACAGGTAGAGGCGGCATATGATATGCTTCTCATGCAAAGCTTATCACAGAGGAGATCTGGAAAAGTTGTGGATAGTAGTGTACGCTTTGCTGATGTTAAAGCTGCTAATGCTTCTGGGATGG
 TC221582  ACAATTGCACAGGTAGAGGCGGCATATGATATGCTTCTCATGCAAAGCTTATCACAGAGGAGATCTGGAAAAGTTGTGGATAGTAGTGTACGCTTTGCTGATGTTAAAGCTGCTAATGCTTCTGGGATGG
Consensus  ACAATTGCACAGGTAGAGGCGGCATATGATATGCTTCTCATGCAAAGCTTATCACAGAGGAGATCTGGAAAAGTTGTGGATAGTAGTGTACGCTTTGCTGATGTTAAAGCTGCTAATGCTTCTGGGATGG

           521    530       540       550       560       570       580       590       600       610       620       630       640       650
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 TC219875  GATCAATGCCCAAGTGGCTGCAGACGACTGTCAAGAGCTCACCAGTCGCAGTTGAAGCACCAGCTTCTAAAGAGTTAGGTGTTCAAGCAGGAGTCTATGGAGCCTTAATGGTCTTAACATATGTTAATGG
 TC221582  GATCAATGCCCAAGTGGCTGCAGACGACTGTCAAGAGCTCACCAGTCGCAGTTGAAGCACCAGCTTCTAAAGAGTTAGGTGTTCAAGCAGGAGTCTATGGAGCCTTAATGGTCTTAACATATGTTAATGG
Consensus  GATCAATGCCCAAGTGGCTGCAGACGACTGTCAAGAGCTCACCAGTCGCAGTTGAAGCACCAGCTTCTAAAGAGTTAGGTGTTCAAGCAGGAGTCTATGGAGCCTTAATGGTCTTAACATATGTTAATGG

           651    660       670       680       690       700       710       720       730       740       750       760       770       780
           |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
 TC219875  AGCCTCAACACCTTTAGGAGTATCGTATGGAGCTGATGTTCCTGGACTAATCTTAGCCACAAGCTTTGGGGCCACCTTGTACTTCATGACCAAGAAAAATGTCAAGTTGGGGAAAGCAAGCGTCATAACA
 TC221582  AGCCTCAACACCTTTAGGAGTATCGTATGGAGCTGATGTTCCTGGACTAATCTTAGCCACAAGCTTTGGG
Consensus  AGCCTCAACACCTTTAGGAGTATCGTATGGAGCTGATGTTCCTGGACTAATCTTAGCCACAAGCTTTGGG............................................................

           781    790       800       810      819
           |--------+---------+---------+--------|
 TC219875  ATTGGTGGGCTGGTCGCTGGTGCAGTGGTAGGTTCAGCA
 TC221582
Consensus  .......................................
```

iAssembler

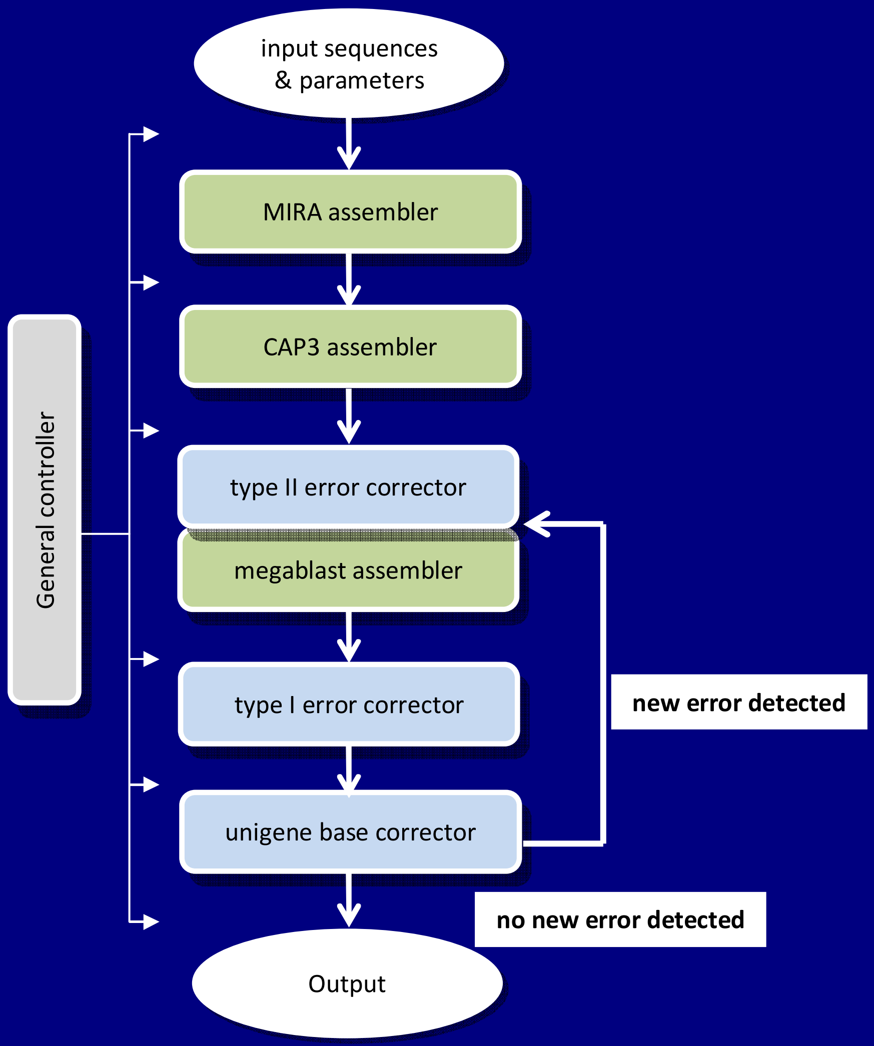
http://bioinfo.bti.cornell.edu/tool/iAssembler/

- iterative assemblies (assembly of assemblies) using MIRA and CAP3 (four cycles of MIRA followed by one cycle of CAP3) – reduce errors that nearly identical sequences are not assembled
- Further assembly error identification

  1) comparing unigene sequences against themselves to identify nearly identical sequences (type II errors)

  2) aligning EST sequences to their corresponding unigene sequences to identify mis-assembled ESTs (type I errors)
- Both type I and II assembly errors are corrected automatically by the program
- Unigene base errors are then corrected based on the resulting SAM files

Workflow of iAssembler



Correct unigene base errors: Iterative assemblies can result in loss of accuracy in unigene base call. iAssembler reassign each individual unigene base sequence according to the SAM output file which contains detailed alignment information of individual ESTs and their corresponding unigenes. The most frequent base in the specific position will be assigned to that position in the unigene.

iAssembler performance

Test datasets

1. Tomato Sanger ESTs: 362,445 with average length of 579 bp
2. Olive 454 ESTs: 246,993 with average length of 196 bp

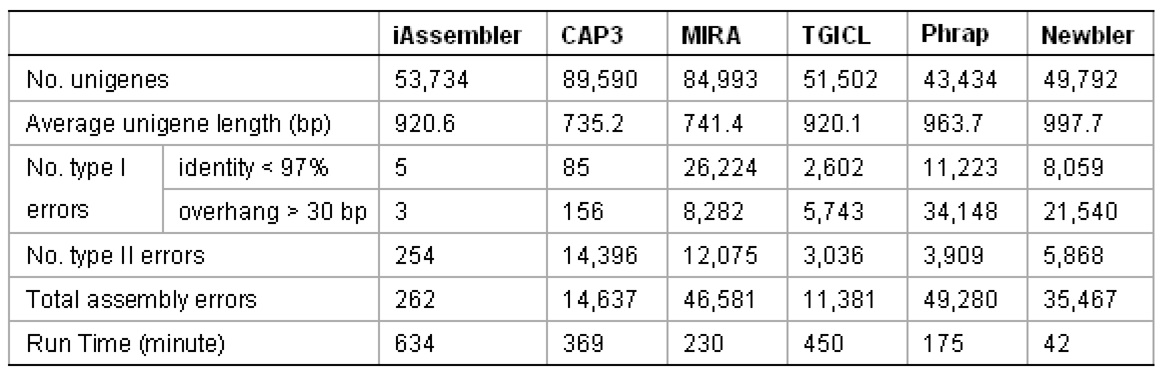
Parameters

Percent identity - 97, maximum overhang - 40, minimum overlap - 30

Program	**Command and parameters**
iAssembler	iAssembler.pl -i input_est -h 40 -e 30 -p 97 -d -o output
CAP3	cap3 input_est -o 40 -y 30 -p 97 -f 6 -s 251
TGICL	tgicl -F input_est -l 40 -v 30 -p 97
MIRA (olive)	mira -project=project -fasta=input_est -job=denovo,est,normal,454 -notraceinfo -GE:not=1 454_SETTINGS -LR:wqf=no -AS:epoq=no:mrl=30 COMMON_SETTINGS -AS:nop=4 -SK:not=1:pr=97 -CL:pec=no 454_SETTINGS -AL:mo=40:mrs=97
MIRA (tomato)	mira -project=project -fasta=input_est -job=denovo,est,normal,sanger -notraceinfo -GE:not=1 SANGER_SETTINGS -LR:wqf=no -AS:epoq=no:mrl=30 COMMON_SETTINGS -AS:nop=4 -SK:not=1:pr=97 -CL:pec=no SANGER_SETTINGS -AL:mo=40:mrs=97
Phrap	phrap input_est –ace

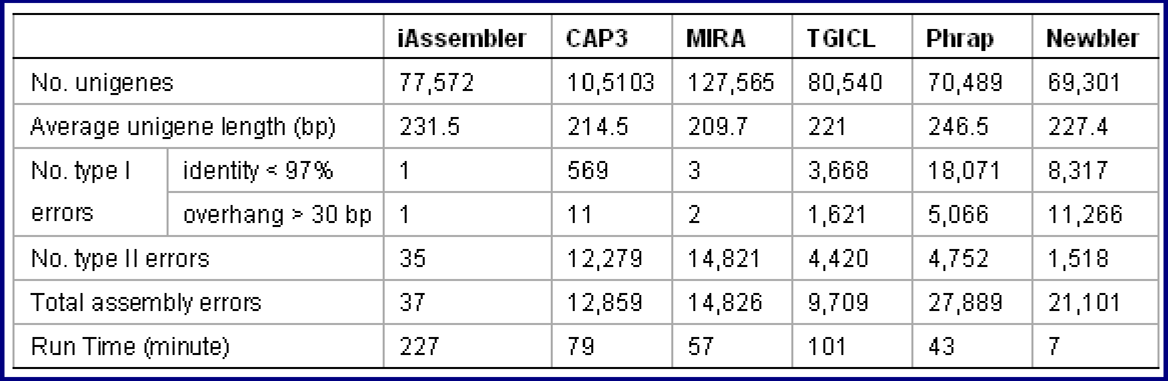
iAssembler performance

Tomato



		iAssembler	CAP3	MIRA	TGICL	Phrap	Newbler
No. unigenes		53,734	89,590	84,993	51,502	43,434	49,792
Average unigene length (bp)		920.6	735.2	741.4	920.1	963.7	997.7
No. type I errors	identity < 97%	5	85	26,224	2,602	11,223	8,059
	overhang > 30 bp	3	156	8,282	5,743	34,148	21,540
No. type II errors		254	14,396	12,075	3,036	3,909	5,868
Total assembly errors		262	14,637	46,581	11,381	49,280	35,467
Run Time (minute)		634	369	230	450	175	42

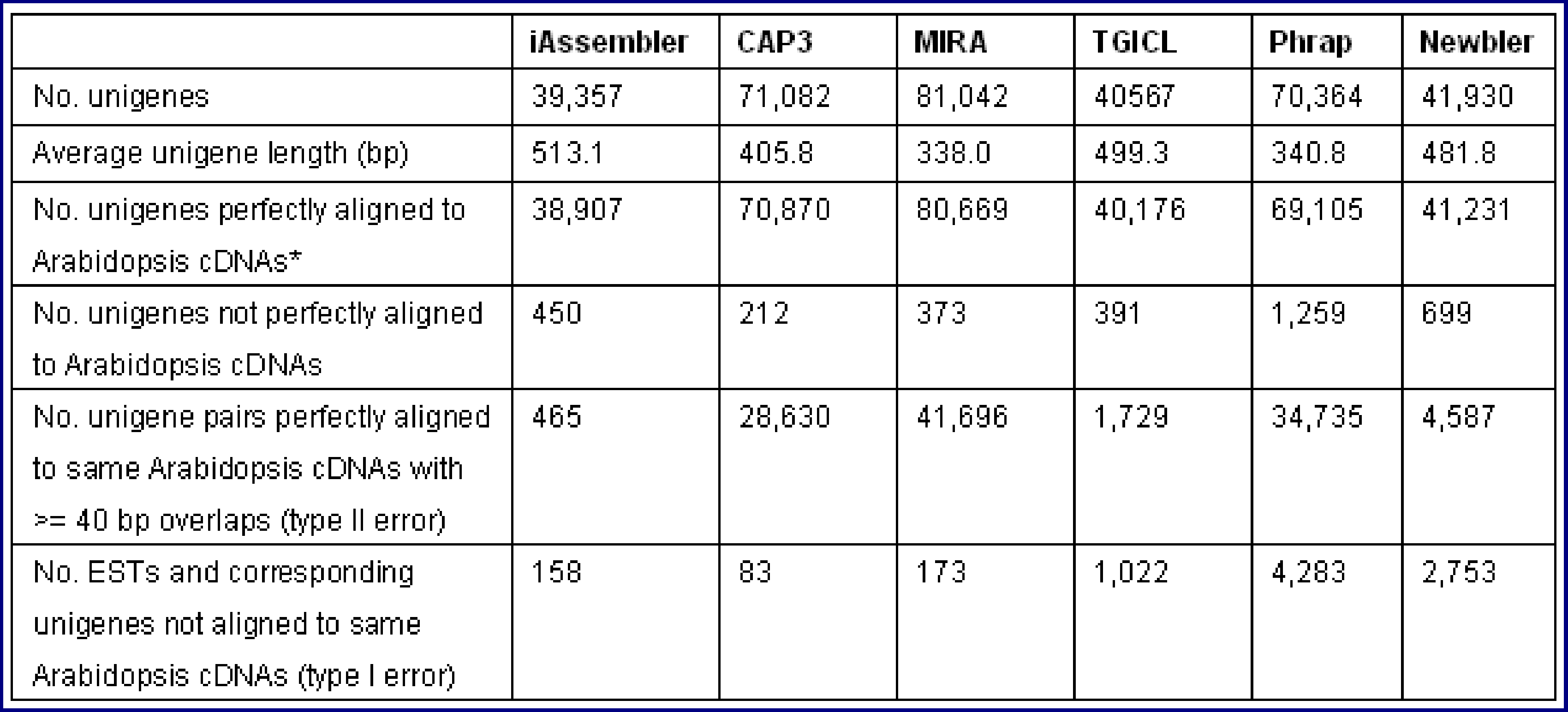
Olive



		iAssembler	CAP3	MIRA	TGICL	Phrap	Newbler
No. unigenes		77,572	10,5103	127,565	80,540	70,489	69,301
Average unigene length (bp)		231.5	214.5	209.7	221	246.5	227.4
No. type I errors	identity < 97%	1	569	3	3,668	18,071	8,317
	overhang > 30 bp	1	11	2	1,621	5,066	11,266
No. type II errors		35	12,279	14,821	4,420	4,752	1,518
Total assembly errors		37	12,859	14,826	9,709	27,889	21,101
Run Time (minute)		227	79	57	101	43	7

iAssembler performance

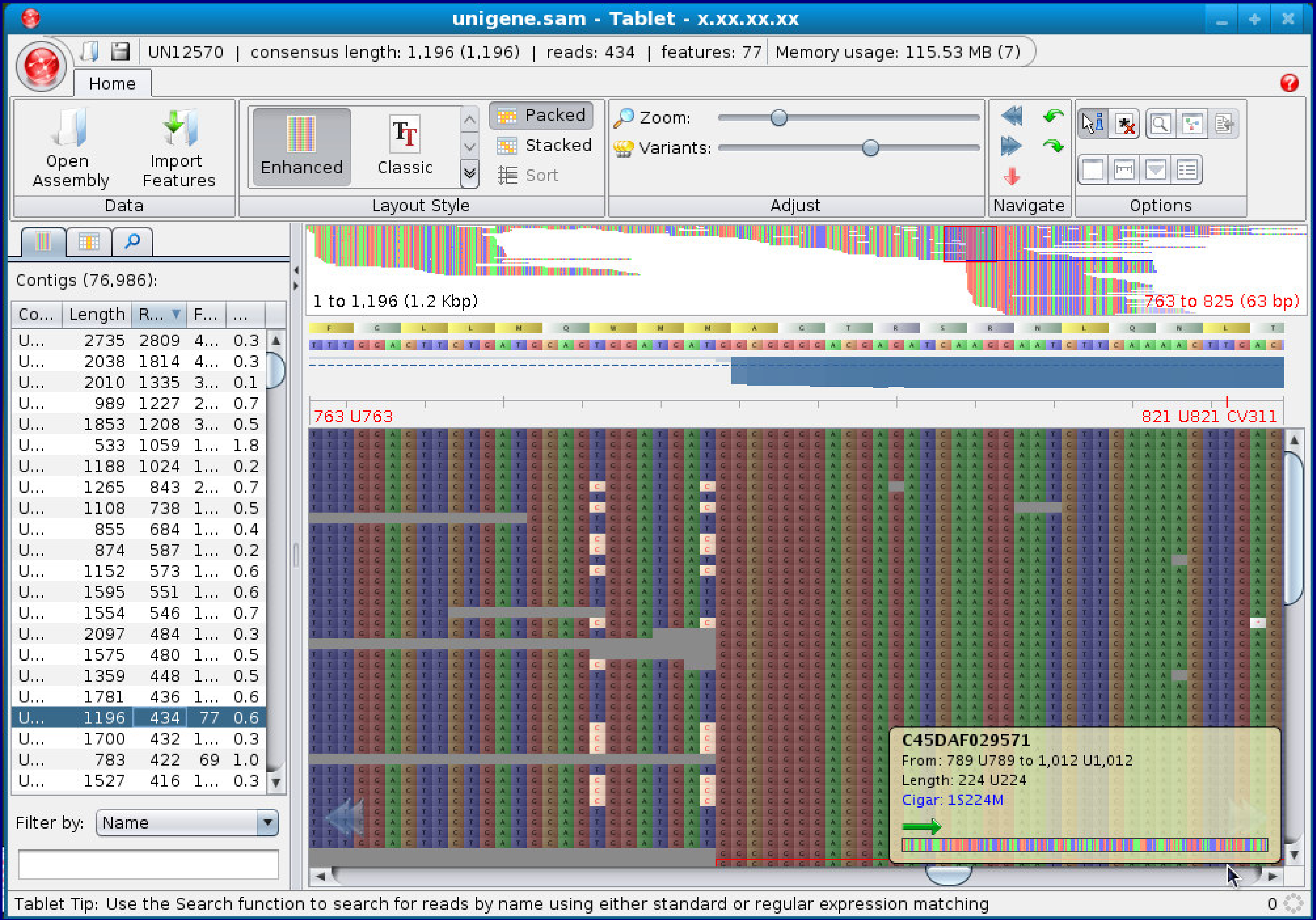
A curated Arabidopsis EST dataset, which only contain ESTs that can be perfectly aligned to the TAIR10 cDNAs



	iAssembler	CAP3	MIRA	TGICL	Phrap	Newbler
No. unigenes	39,357	71,082	81,042	40567	70,364	41,930
Average unigene length (bp)	513.1	405.8	338.0	499.3	340.8	481.8
No. unigenes perfectly aligned to Arabidopsis cDNAs*	38,907	70,870	80,669	40,176	69,105	41,231
No. unigenes not perfectly aligned to Arabidopsis cDNAs	450	212	373	391	1,259	699
No. unigene pairs perfectly aligned to same Arabidopsis cDNAs with >= 40 bp overlaps (type II error)	465	28,630	41,696	1,729	34,735	4,587
No. ESTs and corresponding unigenes not aligned to same Arabidopsis cDNAs (type I error)	158	83	173	1,022	4,283	2,753

perfectly aligned means that the sequences were aligned to Arabidopsis cDNAs in their entire lengths

iAssembler - SAM (Sequence Alignment/Map) format output



unigene.sam - Tablet - x.xx.xx.xx

UN12570 | consensus length: 1,196 (1,196) | reads: 434 | features: 77 | Memory usage: 115.53 MB (7)

Contigs (76,986):

Co...	Length	R...	F...	...
U...	2735	2809	4...	0.3
U...	2038	1814	4...	0.3
U...	2010	1335	3...	0.1
U...	989	1227	2...	0.7
U...	1853	1208	3...	0.5
U...	533	1059	1...	1.8
U...	1188	1024	1...	0.2
U...	1265	843	2...	0.7
U...	1108	738	1...	0.5
U...	855	684	1...	0.4
U...	874	587	1...	0.2
U...	1152	573	1...	0.6
U...	1595	551	1...	0.6
U...	1554	546	1...	0.7
U...	2097	484	1...	0.3
U...	1575	480	1...	0.5
U...	1359	448	1...	0.5
U...	1781	436	1...	0.6
U...	1196	434	77	0.6
U...	1700	432	1...	0.3
U...	783	422	69	1.0
U...	1527	416	1...	0.3

Filter by: Name

1 to 1,196 (1.2 Kbp)

763 to 825 (63 bp)

763 U763

821 U821 CV311

C45DAF029571
From: 789 U789 to 1,012 U1,012
Length: 224 U224
Cigar: 1S224M

Tablet Tip: Use the Search function to search for reads by name using either standard or regular expression matching