De novo assembly of transcriptome sequences

Zhangjun Fei
Boyce Thompson Institute for Plant Research
USDA Robert W. Holley Center for Agriculture and Health
Cornell University

Transcriptome sequencing

- Accelerating gene discovery and gene family expansion
- Accelerating genome annotation – identifying novel genes and gene models
- Identification of tissue/condition specific alternative splicing events
- Identification of transcript fusion events
- Building physical and genetic map (SNP and SSR marker identification – facilitating breeding)
- Gene expression and allele-specific analysis

RNA-seq

Problem of microarray

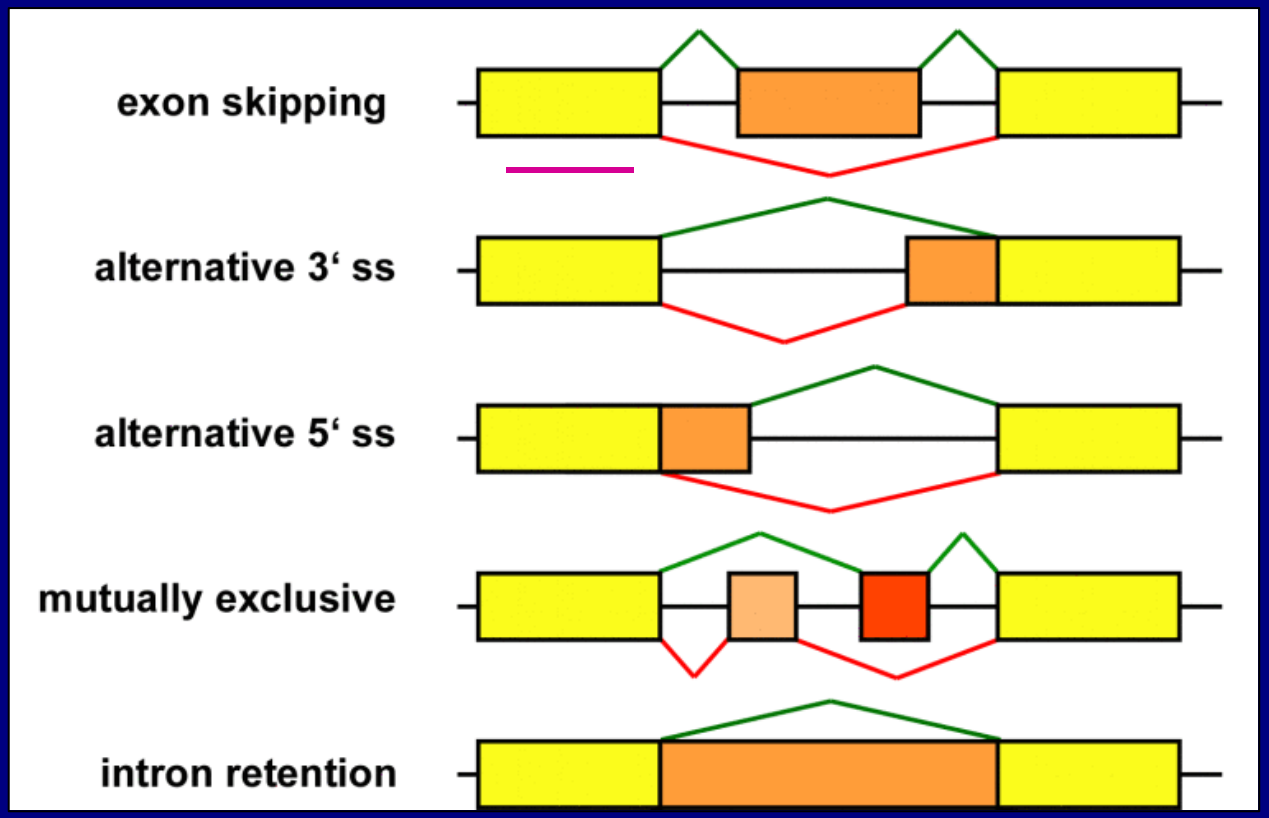
- Cross-hybridization
- Stable probe secondary structures
- high background (e.g., nonspecific hybridization)
- limited dynamic range (e.g., nonlinear and saturable hybridization kinetics)

RNA-seq (digital expression analysis)

- allow direct enumeration of transcript molecules
- Because counting statistics are well modeled by the Poisson distribution, they do not require repetition or standardization
- digital expression data are absolute so data can be directly compared across different experiments and laboratories without the need for extensive internal controls or other experimental manipulation
- provide open systems that allow detection of previously uncharacterized transcripts, as well as rare transcripts

RNA-seq

- Can’t distinguish the expression of alternative spliced transcripts
- Challenging in de novo transcriptome assembly
- Short reads (e.g. Solexa) requires whole genome sequences



RNA-seq

The beginning of the end for microarrays?

Jay Shendure

Two complementary approaches, both using next-generation sequencing, have successfully tackled the scale and the complexity of mammalian transcriptomes, at once revealing unprecedented detail and allowing better quantification.

For over a decade, DNA microarrays have provided a powerful approach to achieve parallel interrogation of biological systems at a genomic scale. But two new reports in this issue of *Nature Methods*[1,2] demonstrate that massively parallel DNA sequencing may be on its way to supplanting microarrays as the technology of choice for quantifying and annotating transcriptomes.

Since 2004, massively parallel DNA sequencing technologies have exploded onto the scene, offering dramatically lower per-base costs than had previously been possible with electrophoretic sequencing[3]. The two papers in this issue of *Nature Methods*[1,2] describe the application of next-generation sequencing to characterize several mouse

to the reproducibility of results between laboratories and across platforms.

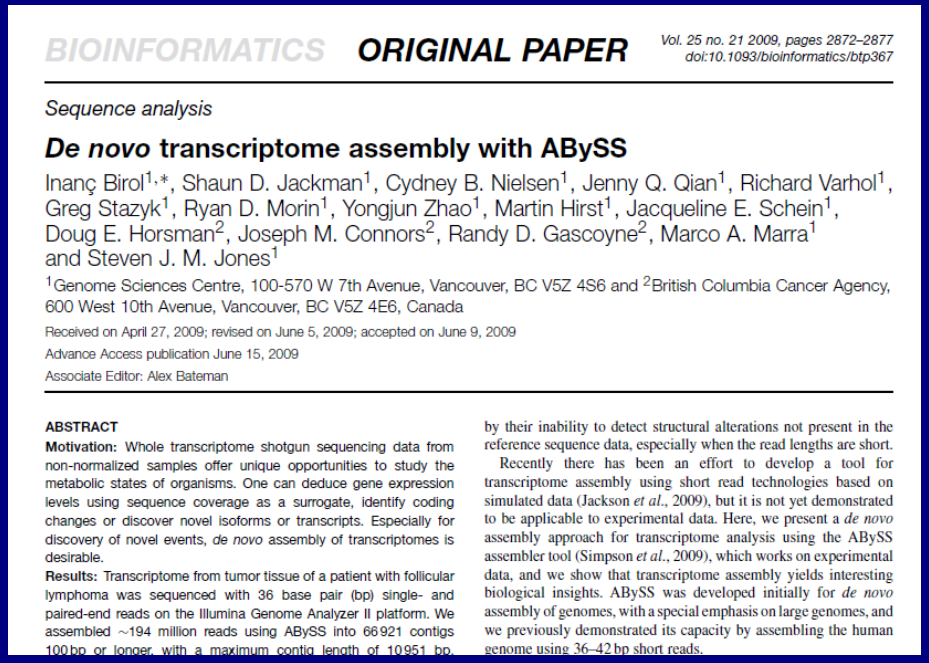
Of course, transcriptome sequencing by itself is nothing new. Sequencing of expressed sequence tags (ESTs)[7] provided an early means of discovering coding sequences in the absence of a reference genome and subsequently for annotation of transcriptional units. The high cost of deep EST sequencing motivated the development of serial analysis of gene expression (SAGE)[8], which lowered costs by minimizing the amount of information collected per transcript. Even with SAGE, however, the cost of transcriptome analysis with conventional sequencing remains high relative to that of microarray analysis. The introduction of next-generation sequencing technology into this area represents a major leap toward a leveling of the playing field. For example, tens of millions of independently derived sequencing tags can now be obtained at a cost similar to what tens of thousands used to cost.

The RNA-Seq approach also brings a qualitative and quantitative improvement to transcriptome analysis. For example, by tak-

Nature Method, July 2008

Transcriptome sequencing

- Organisms having a reference genome/transcriptome: currently dominated by Illumina/Solexa, followed by SOLiD and Roche/454
- Orphan organisms: currently dominated by Roche/454. No practical applications of Illumina/Solexa and SOLiD for de novo transcriptome sequencing have been published except the two methodology papers.



BIOINFORMATICS ORIGINAL PAPER
Vol. 25 no. 21 2009, pages 2872–2877
doi:10.1093/bioinformatics/btp367

Sequence analysis

De novo transcriptome assembly with ABySS

Inanç Birol[1,*], Shaun D. Jackman[1], Cydney B. Nielsen[1], Jenny Q. Qian[1], Richard Varhol[1], Greg Stazyk[1], Ryan D. Morin[1], Yongjun Zhao[1], Martin Hirst[1], Jacqueline E. Schein[1], Doug E. Horsman[2], Joseph M. Connors[2], Randy D. Gascoyne[2], Marco A. Marra[1] and Steven J. M. Jones[1]

[1]Genome Sciences Centre, 100-570 W 7th Avenue, Vancouver, BC V5Z 4S6 and [2]British Columbia Cancer Agency, 600 West 10th Avenue, Vancouver, BC V5Z 4E6, Canada

Received on April 27, 2009; revised on June 5, 2009; accepted on June 9, 2009

Advance Access publication June 15, 2009

Associate Editor: Alex Bateman

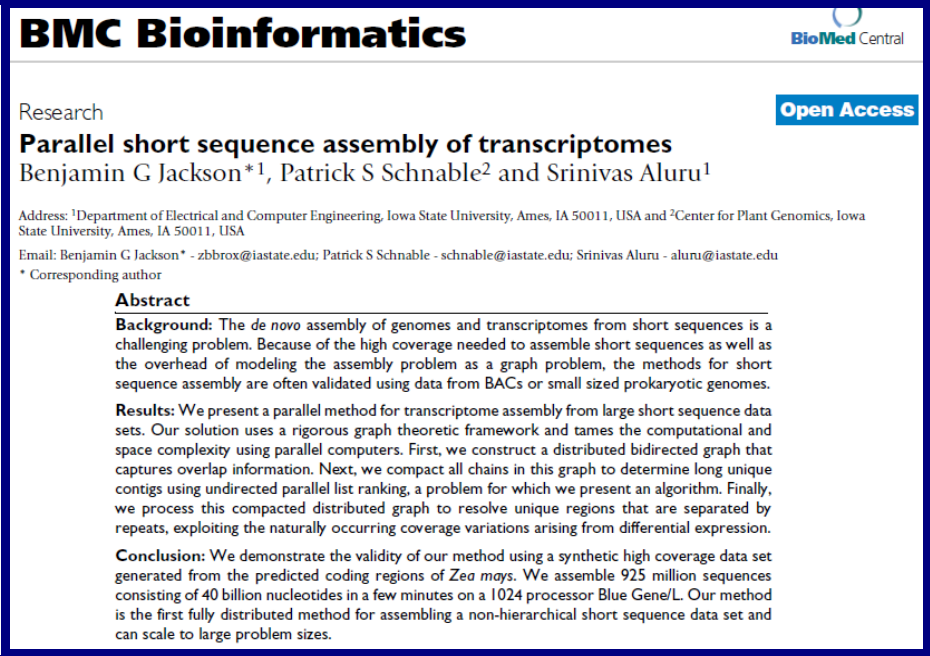
ABSTRACT

Motivation: Whole transcriptome shotgun sequencing data from non-normalized samples offer unique opportunities to study the metabolic states of organisms. One can deduce gene expression levels using sequence coverage as a surrogate, identify coding changes or discover novel isoforms or transcripts. Especially for discovery of novel events, *de novo* assembly of transcriptomes is desirable.

Results: Transcriptome from tumor tissue of a patient with follicular lymphoma was sequenced with 36 base pair (bp) single- and paired-end reads on the Illumina Genome Analyzer II platform. We assembled ~194 million reads using ABySS into 66 921 contigs 100 bp or longer, with a maximum contig length of 10 951 bp, by their inability to detect structural alterations not present in the reference sequence data, especially when the read lengths are short.

Recently there has been an effort to develop a tool for transcriptome assembly using short read technologies based on simulated data (Jackson *et al.*, 2009), but it is not yet demonstrated to be applicable to experimental data. Here, we present a *de novo* assembly approach for transcriptome analysis using the ABySS assembler tool (Simpson *et al.*, 2009), which works on experimental data, and we show that transcriptome assembly yields interesting biological insights. ABySS was developed initially for *de novo* assembly of genomes, with a special emphasis on large genomes, and we previously demonstrated its capacity by assembling the human genome using 36–42 bp short reads.

Human



BMC Bioinformatics

BioMed Central

Research

Open Access

Parallel short sequence assembly of transcriptomes

Benjamin G Jackson*[1], Patrick S Schnable[2] and Srinivas Aluru[1]

Address: [1]Department of Electrical and Computer Engineering, Iowa State University, Ames, IA 50011, USA and [2]Center for Plant Genomics, Iowa State University, Ames, IA 50011, USA

Email: Benjamin G Jackson* - zbbrox@iastate.edu; Patrick S Schnable - schnable@iastate.edu; Srinivas Aluru - aluru@iastate.edu

* Corresponding author

Abstract

Background: The *de novo* assembly of genomes and transcriptomes from short sequences is a challenging problem. Because of the high coverage needed to assemble short sequences as well as the overhead of modeling the assembly problem as a graph problem, the methods for short sequence assembly are often validated using data from BACs or small sized prokaryotic genomes.

Results: We present a parallel method for transcriptome assembly from large short sequence data sets. Our solution uses a rigorous graph theoretic framework and tames the computational and space complexity using parallel computers. First, we construct a distributed bidirected graph that captures overlap information. Next, we compact all chains in this graph to determine long unique contigs using undirected parallel list ranking, a problem for which we present an algorithm. Finally, we process this compacted distributed graph to resolve unique regions that are separated by repeats, exploiting the naturally occurring coverage variations arising from differential expression.

Conclusion: We demonstrate the validity of our method using a synthetic high coverage data set generated from the predicted coding regions of *Zea mays*. We assemble 925 million sequences consisting of 40 billion nucleotides in a few minutes on a 1024 processor Blue Gene/L. Our method is the first fully distributed method for assembling a non-hierarchical short sequence data set and can scale to large problem sizes.

Maize

Transcriptome sequencing

De novo transcriptome sequencing using Solexa

- HUGE interest in the community (454 is expensive and relative low throughput)
- De novo assembly is very challenging

  - ABySS (http://www.bcgsc.ca/platform/bioinfo/software/abyss)

  - Oases (http://www.ebi.ac.uk/~zerbino/oases/; "Oases uploads a preliminary assembly produced by Velvet, and clusters the contigs into small groups, called loci. It then exploits the paired-end read and long read information, when available, to construct transcript isoforms")
- First generate reference transcriptome sequences using 454, then align Solexa reads to assembled transcriptomes

  - sequence a pool of **normalized samples of interest** (to maximize the transcriptome coverage)

De novo assembly using ABySS

Two lanes of Solexa single end reads
Total number of reads: 33,203,388
Length: 86bp

	k=35	k=40	k=45	k=50	k=55	k=60	k=64
total number of unigenes	360,396	309,340	206,016	164,364	139,185	118,449	105,695
total length of unigenes	50,404,118	51,424,489	46,945,884	43,392,695	40,418,375	36,597,960	33,334,698
average length of unigens	139.9	166.2	227.9	264	290.4	309	315.4
number of unigenes >100bp	123,782	126,825	126,390	119,490	118,137	107,051	98,681
total length of unigenes >100bp	37,339,837	40,224,718	41,287,983	40,062,768	38,907,733	35,747,930	32,792,490
average length of unigens > 100bp	301.7	317.2	326.7	335.3	329.3	333.9	332.3
N50 (for unigenes >100bp)	396	431	458	484	472	452	423
longest unigene	8,022	11,749	8,542	8,542	7,870	7,870	6,491
#reads used for assembly	31,718,000	31,391,037	29,211,754	28,002,369	26,841,445	25,560,960	24,354,015
#reads not used for assembly	1,458,393	1,754,664	3,890,721	3,760,861	4,823,146	5,960,921	7,092,160
% reads not used for assembly	4.40	5.29	11.75	11.84	15.23	18.91	22.55

Transcriptome sequence processing (454)

1. Remove low quality reads and regions

```
>F7FLTVN08JMFPP
25 26 30 10 30 50 44 33 18 42 25 22 42 25 30 30 27 38 19 30 29 25 28 30 42 25 34 22 5 30 29 28 27 29 34 22 5 43 25 30 27 22 28 41 26 9
28 50 50 49 36 16 29 39 24 27 28 50 37 26 7 30 30 26 25 30 5 34 22 5 50 50 46 36 22 5 39 24 50 50 41 26 4 27 43 25 29 30 29 27 30 48 38
24 30 29 40 23 22 30 27 30 29 40
>F7FLTVN08I7AKT
48 38 24 27 30 30 42 25 30 30 12 24 38 29 13 27 28 38 29 18 38 19 38 19 29 26 3 30 28 26 46 36 22 50 50 50 46 38 30 19 4 28 27 38 29 13
45 37 25 30 30 30 30 27 40 23 30 30 29 26 30 25 28 43 25 30 42 25 30 28 39 24 29 30 42 25 28 29 30 42 25 30 27 30 30 28 40 23 30 43 25
28 30 40 23 30 26 29 42 25 30 46 36 22 30 40 23 30 30 30 29 47 37 22 30 28 46 36 22 28 30 29 30 24 30 50 50 41 26 4 29 29 29 27 39 24
28 30 30 30 30 42 25 30 40 23 26 30 30 30 43 25 26 30 29 42 25 28 29 42 25 29 30 30 30 28 26 42 25 30 30 30 29 23 30 30 30 30 30 30 50
50 41 26 4 25 45 37 25 24 42 25 27 29 39 33 21 4 27 28 9 30 27 34 22 27 30 28 34 22 30 30 25 24 29 30 25 30 28 30 27 24 42 25 30 26 30
29 30 50 50 40 24 4 29 50 44 33 18 29 19 28 30 24 38 19 28 24 43 25 30 50 37 24 38 19 34 21 38 29 18 25
```

Lucy: http://lucy.sourceforge.net/ (can only process raw reads no less than 50 bp)

```
$ lucy test.fasta test.qul -o test_lucy.fasta test_lucy.qul -m 100 -e 0.01 0.01
$ awk -f zapping.awk test_lucy.fasta >test_highq.fasta
```

```
[feizj@localhost Desktop]$ lucy
Less Useful Chunks Yank (lucy) 1.19p, by Hui-Hsien Chou and Michael Holmes,
  with help from Granger, Anna, John and Terry Shea.
lucy: no input sequence file
usage: lucy
   [-pass_along min_value max_value med_value]
   [-range area1 area2 area3] [-alignment area1 area2 area3]
   [-vector vector_sequence_file splice_site_file]
   [-cdna [minimum_span maximum_error initial_search_range]] [-keep]
   [-size vector_tag_size] [-threshold vector_cutoff]
   [-minimum good_sequence_length] [-debug [filename]]
   [-output sequence_filename quality_filename]
   [-error max_avg_error max_error_at_ends]
   [-window window_size max_avg_error [window_size max_avg_error ...]]
   [-bracket window_size max_avg_error]
   [-quiet] [-inform_me] [-xtra cpu_threads]
   sequence_file quality_file [2nd_sequence_file]
```

Transcriptome sequence processing (454)

2. Remove adaptors and all possible contaminations (GenBank dbEST is full of contaminations)
rRNA, tRNA, vectors, chloroplast and mitochondrion DNAs, polyA/T, low complexity……

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value
BAB33421.1	putative senescence-associated protein [Pisum sativum]	352	398	24%	1e-101
T02955	probable cytochrome P450 monooxygenase - maize (fragment)	160	342	21%	2e-78
ACR38454.1	unknown [Zea mays]	178	321	19%	4e-72
BAA10929.1	cytochrome P450 like_TBP [Nicotiana tabacum]	169	342	21%	8e-64
EEH50840.1	predicted protein [Micromonas pusilla CCMP1545]	211	259	11%	5e-60
ACR36970.1	unknown [Zea mays]	151	234	11%	2e-52
XP_001900327.1	Senescence-associated protein [Brugia malayi] >gb|EDP31077.1| Se	133	223	10%	6e-49
BAF01964.1	hypothetical protein [Arabidopsis thaliana]	169	335	14%	2e-39
ACA04850.1	senescence-associated protein [Picea abies]	168	275	13%	4e-39
ACA30301.1	putative senescence-associated protein [Cupressus sempervirens]	161	189	7%	5e-39
XP_001622003.1	hypothetical protein NEMVEDRAFT_v1g142908 [Nematostella vectens	111	188	9%	1e-38
ACJ85262.1	unknown [Medicago truncatula]	166	272	13%	2e-38
XP_665229.1	senescence-associated protein [Cryptosporidium hominis TU502] >g	108	186	9%	4e-38
XP_001786503.1	predicted protein [Physcomitrella patens subsp. patens] >gb|EDQ48	127	186	9%	5e-38
XP_001267665.1	hypothetical protein NFIA_061320 [Neosartorya fischeri NRRL 181] >	104	186	11%	6e-38
XP_002139698.1	senescence-associated protein [Cryptosporidium muris RN66] >ref|X	107	185	9%	8e-38
XP_002181474.1	predicted protein [Phaeodactylum tricornutum CCAP 1055/1] >gb|EE	103	185	9%	8e-38
XP_002163485.1	PREDICTED: similar to predicted protein [Hydra magnipapillata]	111	185	9%	1e-37
XP_002489117.1	hypothetical protein SORBIDRAFT_0057s002150 [Sorghum bicolor] :	98.6	185	9%	1e-37
XP_002338057.1	predicted protein [Populus trichocarpa] >gb|EEF07697.1| predicted p	93.2	185	7%	1e-37
AAR25995.1	putative senescence-associated protein [Pyrus communis]	93.2	185	7%	1e-37
ABO20851.1	putative senescence-associated protein [Lilium longiflorum]	152	184	7%	1e-37
BAF46313.1	putative senescence-associated protein [Ipomoea nil]	95.5	183	8%	3e-37

Arabidopsis 25S ribosomal RNA vs GenBank nr protein database

Transcriptome sequence processing (454)

seqclean (http://compbio.dfci.harvard.edu/tgi/software/)

```
[feizj@localhost seqclean]$ perl seqclean

 seqclean <seqfile> [-v <vecdbs>] [-s <screendbs>] [-r <reportfile>]
    [-o <outfasta>] [-n slicesize] [-c {<num_CPUs>|<PVM_nodefile>}]
    [-l <minlen>] [-N] [-A] [-L] [-x <min_pid>] [-y <min_vechitlen>]
    [-m <e-mail>]

Parameters

<seqfile>: sequence file to be analyzed (multi-FASTA)

       -c use the specified number of CPUs on local machine
         (default 1) or a list of PVM nodes in <PVM_nodefile>
       -n number of sequences taken at once in each
          search slice (default 2000)
       -v comma delimited list of sequence files
          to use for end-trimming of <seqfile> sequences
          (usually vector sequences)
       -l during cleaning, consider invalid the sequences sorter
          than <minlen> (default 100)
       -s comma delimited list of sequence files to use for
          screening <seqfile> sequences for contamination
          (mito/ribo or different species contamination)
       -r write the cleaning report into file <reportfile>
          (default: <seqfile>.cln)
       -o output the "cleaned" sequences to file <outfasta>
          (default: <seqfile>.clean)
       -x minimum percent identity for an alignemnt with
          a contaminant (default 96)
       -y minimum length of a terminal vector hit to be considered
          (>11, default 11)
       -N disable trimming of ends rich in Ns (undetermined bases)
       -M disable trashing of low quality sequences
       -A disable trimming of polyA/T tails
       -L disable low-complexity screening (dust)
       -I do not rebuild the cdb index file
       -m send e-mail notifications to <e-mail>
```

Low complexity sequences

```
AAAAAAAAAAATTTTTTTTTTTTTAAAAAAAAAAAAAAAAAGGGG
GGGCCCGCGGTTTTTTTAAAAAAAAAAAAAAAAAAAATTCCCCCC
CAAAAAAAAACCCCCC
```

All the adapter, vector, rRNA, UniVec ……databases need to be formatted so they can be searched by blast programs

```
$ formatdb -i database_name -p F
```

Run seqclean program

```
$ perl seqclean test_highq.fasta –c 2 -v adapter -s rRNA,
  tRNA,plastid,UniVec,E_coli
```

```
************************************************
Sequences analyzed:      9391
---------------------------------
                  valid:      9111  (325 trimmed)
                trashed:       280
************************************************
----= Trashing summary =------
             by 'adapter':        2
                by 'rRNA':      238
              by 'E_coli':        5
              by 'shortq':        1
                by 'dust':        1
             by 'plastid':       33
------------------------------
```

Transcriptome sequence processing (454)

E coli genome contamination

```
Query= F7FLTVN08JGQ91 (426 letters)

>NC_002655 Escherichia coli O157:H7 EDL933, complete genome. (Length = 5528445)

Identities = 424/426 (99%), Gaps = 1/426 (0%)

Query: 1       gcgcgaagttcaactattgttctgtggtgttctgttgcgtgttgacggacaaaattttgc 60
                |||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct: 4864420 gcgcgaagttcaactattgttctgtggtgttctgttgcgtgttgacgg-caaaattttgc 4864478

Query: 61      tggcgtaacatgcgcgcacgatcactctaagaggacattcgccttggacacacccagtag 120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4864479 tggcgtaacatgcgcgcacgatcactctaagaggacattcgccttggacacacccagtag 4864538

Query: 121     atactggctcactatcctgtcatccaggatcaactcctaaggctatccctttttgctgat 180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4864539 atactggctcactatcctgtcatccaggatcaactcctaaggctatccctttttgctgat 4864598

Query: 181     agccttagcggttgtcagcgacctcaatttttcccgtcgcgctgagtcaggctgtttaat 240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4864599 agccttagcggttgtcagcgacctcaatttttcccgtcgcgctgagtcaggctgtttaat 4864658

Query: 241     ggtctgaaacccaatttgtttctgtgtgcccactgaactgtccgatattttaagcattgg 300
                |||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct: 4864659 ggtctgaaacccaatttgtttctgtgtgcccaccgaactgtccgatattttaagcattgg 4864718

Query: 301     gagtcccggtcatgctgagcgcatttcaactggaaaataaccgactgacccggctggaag 360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4864719 gagtcccggtcatgctgagcgcatttcaactggaaaataaccgactgacccggctggaag 4864778

Query: 361     tcgaagagtcacaacccccttgtaaatgcagtatggattgatcttgtcgaaccggacgacg 420
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 4864779 tcgaagagtcacaacccccttgtaaatgcagtatggattgatcttgtcgaaccggacgacg 4864838

Query: 421     acgagc 426
                ||||||
Sbjct: 4864839 acgagc 4864844
```

Transcriptome sequence assembly (454)

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

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

1) Large portion of nearly identical sequences are not assembled into one unigene
2) Large portion of different transcripts (mainly alternative spliced transcripts) are incorrectly assembled into same unigenes

mis-assemblies


```
            1      10        20        30        40        50        60        70        80        90       100       110       120       130       140       150
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883    GGTTGAGGTATTCTTCATTCTGAACCCTCCTCCTCGTCGGCGAACGGCTGCTCTTCTTCTGCTATCTTTTGCTTCCAACTTACTCAATCGGCGGAGAAGCAGAATCCAGTGGCCGGCCGCCGGAATCTCAATTGAGCTGCGTAGTTAGAA
TC210672    GGTTGAGGTATTCTTCATTCTGAACCCTCCTCCTCGTCGGCGAACGGCTGCTCTTCTTCTGCTATCTTTTGCTTCCAACTTACTCAATCGGCGGAGAAGCAGAATCCAGTGGCCGGCCGCCGGAATCTCAATTGAGCTGCGTAGTTAGAA
Consensus   GGTTGAGGTATTCTTCATTCTGAACCCTCCTCCTCGTCGGCGAACGGCTGCTCTTCTTCTGCTATCTTTTGCTTCCAACTTACTCAATCGGCGGAGAAGCAGAATCCAGTGGCCGGCCGCCGGAATCTCAATTGAGCTGCGTAGTTAGAA

            151    160       170       180       190       200       210       220       230       240       250       260       270       280       290       300
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883    ATAAACTAGAGAACCATGGGGATCGTAGAGAGGATCAAAGAAATTGAAGCCGAGATGGCTCGTACCCAGAAAAATAAAGCAACTGAATATCATTTGGGTCAGCTGAAGGCTAAGATAGCAAAGCTGAGGACACAATTGTTAGAGCCCCCC
TC210672    ATAAACTAGAGAACCATGGGGATCGTAGAGAGGATCAAAGAAATTGAAGCCGAGATGGCTCGTACCCAGAAAAATAAAGCAACTGAATATCATTTGGGTCAGCTGAAGGCTAAGATAGCAAAGCTGAGGACACAATTGTTAGAGCCCCCC
Consensus   ATAAACTAGAGAACCATGGGGATCGTAGAGAGGATCAAAGAAATTGAAGCCGAGATGGCTCGTACCCAGAAAAATAAAGCAACTGAATATCATTTGGGTCAGCTGAAGGCTAAGATAGCAAAGCTGAGGACACAATTGTTAGAGCCCCCC

            301    310       320       330       340       350       360       370       380       390       400       410       420       430       440       450
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883    AAAGGTTCTAGTGGTGCTGGAGAAGGTTTCGAAGTTACAAAATTTGGCCATGGACGTGTTGCACTGATAGGATTTCCAAGTGTGGGAAAGTCTACACTCCTTACAATGTTGACAGGAACACATTCTGAAGCTGCATCATATGAGTTCACA
TC210672    AAAGGTTCTAGTGGTGCTGGAGAAGGTTTCGAAGTTACAAAATTTGGCCATGGACGTGTTGCACTGATAGGATTTCCAAGTGTGGGAAAGTCTACACTCCTTACAATGTTGACAGGAACACATTCTGAAGCTGCATCATATGAGTTCACA
Consensus   AAAGGTTCTAGTGGTGCTGGAGAAGGTTTCGAAGTTACAAAATTTGGCCATGGACGTGTTGCACTGATAGGATTTCCAAGTGTGGGAAAGTCTACACTCCTTACAATGTTGACAGGAACACATTCTGAAGCTGCATCATATGAGTTCACA

            451    460       470       480       490       500       510       520       530       540       550       560       570       580       590       600
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883    ACACTTACTTGCATCCCTGGTATTATCCATTACAATGATACTAAAATTCAATTGCTTGATCTTCCGGGAATCATTGAAGGTGCATCTGAAGGCAAGGGTCGTGGTAGGCAGGTCATTGCTGTTTCTAAGTCAGCGGACATTGTATTAATG
TC210672    ACACTTACTTGCATCCCTGGTATTATCCATTACAATGATACTAAAATTCAATTGCTTGATCTTCCGGGAATCATTGAAGGTGCATCTGAAGGCAAGGGTCGTGGTAGGCAGGTCATTGCTGTTTCTAAGTCAGCGGACATTGTATTAATG
Consensus   ACACTTACTTGCATCCCTGGTATTATCCATTACAATGATACTAAAATTCAATTGCTTGATCTTCCGGGAATCATTGAAGGTGCATCTGAAGGCAAGGGTCGTGGTAGGCAGGTCATTGCTGTTTCTAAGTCAGCGGACATTGTATTAATG

            601    610       620       630       640       650       660       670       680       690       700       710       720       730       740       750
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883    GTTCTCGATGCTTCAAAAAGTGAAGGCCATCGGCAAATATTGACAAAGGAGCTGGAAGCGGTGGGCCTGCGACTAAACAAGAAACCTCCTCAGATATACTTCAAGAAGAAAAAGACTGGTGGAATTTCTTTCAATAGCACTCTGCATCTG
TC210672    GTTCTCGATGCTTCAAAAAGTGAAGGCCATCGGCAAATATTGACAAAGGAGCTGGAAGCGGTGGGCCTGCGACTAAACAAGAAACCTCCTCAGATATACTTCAAGAAGAAAAAGACTGGTGGAATTTCTTTCAATAGCACTCTGCATCTG
Consensus   GTTCTCGATGCTTCAAAAAGTGAAGGCCATCGGCAAATATTGACAAAGGAGCTGGAAGCGGTGGGCCTGCGACTAAACAAGAAACCTCCTCAGATATACTTCAAGAAGAAAAAGACTGGTGGAATTTCTTTCAATAGCACTCTGCATCTG

            751    760       770       780       790       800       810       820       830       840       850       860       870       880       890       900
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883    ACACATGTGGATGAGAAGCTATGCTATCAAATACTGCATGAATACAAGATCCACAATGCTGAGGTGTTATTTCGTGAAGATGCCACAGTGGATGACCTTATTGATGTTATTGAGGGCAATCGTAAATACATGAAGTGCATATATGTCTAC
TC210672    ACACATGTGGATGAGAAGCTATGCTATCAAATACTGCATGAATACAAGATCCACAATGCTGAGGTGTTATTTCGTGAAGATGCCACAGTGGATGACCTTATTGATGTTATTGAGGGCAATCGTAAATACATGAAGTGCATATATGTCTAC
Consensus   ACACATGTGGATGAGAAGCTATGCTATCAAATACTGCATGAATACAAGATCCACAATGCTGAGGTGTTATTTCGTGAAGATGCCACAGTGGATGACCTTATTGATGTTATTGAGGGCAATCGTAAATACATGAAGTGCATATATGTCTAC

            901    910       920       930       940       950       960       970       980       990      1000      1010      1020      1030      1040      1050
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883    AACAAGA
TC210672    AACAAGATAGATGTTGTTGGTATTGATGATGTAGACAGATTAGCCAGACAGCCAAACTCTGTTGTCATCAGCTGCAACATGAAGCTGAATCTGGATAGATTGCTTGCAAAAATGTGGGAAGCGATGGGTCTTGTCAGAGTTTACACAAAG
Consensus   AACAAGA.....................................................................................................................................

            1051   1060      1070      1080      1090      1100      1110      1120      1130      1140      1150      1160      1170      1180      1190      1200
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883
TC210672    CCTCAAGGCCAGCAACCAGACTTCACAGATCCAGTGGTGCTTTCTGCTGATAGAGGTGGCTGTACGGTAGAAGATTTCTGTAATCACATACATCGGAGTCTTGTTAAGGATGTGAAGTATGTGTTGGTTTGGGGTACTAGTGCAAGGCAC
Consensus   .....................................................................................................................................

            1201   1210      1220      1230      1240      1250      1260      1270      1280      1290      1300      1310      1320      1330      1340      1350
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
TC191883
TC210672    TCCCCCCAGCATTGTGGCCTTGCTCAAATGCTTGAGGATGAAGATGTGGTGCAGATTGTTAAGAAAAAGGAGAAGGAAGATGGAGGAGGTAGAGGCCGATTCAAATCACATTCTAATGCTCCTTCTCGTATATCTGACCGTGAAAAGAAG
Consensus   .....................................................................................................................................

            1351   1360      1370      1380      1390      1400      1410      1420      1430      1440      1450      1460      1471
            |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+---------+|
TC191883
TC210672    GCTCCATTGAAGACCTGATTGAGCCAGATAGTGGCGCTGCTAGTTCTATATGCTGATCTGGTAGGATCCATGTGCCAGGTATTACTATTGAAGAATGAGCTATACAATGAGATAGCTAGAG
Consensus   ...................................................................................................
```

Two overlapping unigenes were not assembled in Tomato Gene Index (TGI)

mis-assemblies


```
AI773661  ------------------------------------------------------------------------------------------GCAAA
TC194329  ATAAAAATGTATTAAGAGGAATTAGTATAAAACAAAATAAATTACCAAGCATAACCTTGTTGTAAAGATTCAGTCTCTAAATAGAGATTAATTTTGCAAA

AI773661  ACAGATAACACATTCAAAAATCAGCAATATGATGCTGGAGTGGGTAAGAAAAACAGAGTTGAAATAAGAAAATCCAACAAGATACTATTAATGGTAAAAG
TC194329  ACAGATAACACATTCAAAAATCAGCAATATGATGCTGGAGTGTGTAAGAGAAACAGAGTTGAAATAAGAAAATCCAACAAGATACTATTAATGGTAAAAG

AI773661  TAGGTTGGCAAATACTTCTAATAGAAACTCTATCAATGTGCATTCAAACACCAGCACTTCCTCAAGCCTTGTAGCCTCTGCTATTTCTCTTTCCTCTTGC
TC194329  TAGGTTGGCAAATACTTCTAATAGAAACTCTATCAATGTGCATTCAAACACCAGCACTTCCTCAAGCCTTGTAGCCTCTGCTATTTCTCTTTCCTCTTGC

AI773661  TCGCTCAAACTTCCTTCCCTTTGCGCGAACATAAAGCTTTGTGTGGCTGTGTGGGACACCAAGTGCTTTACCAAAGTGCTTAACGGCCTCCCTTGAGTTC
TC194329  TCGCTCAAACTTCCTTCCCTTTGCGCGAACATAAGGCTTTGTGTGGCTGTGTGGGACACCAGGTGCTTTACCAAAGTGCTTAACGGCCTCCCTTGAGTTC

AI773661  TTTGGACCCCTAAGCAAAAC----------------------------------------------------------------------------
TC194329  TTTGGACCCCTAAGCAAAACCTGATCAAGAAAGCATTACAAAACAAATGACTAGCCAATGTTACCTCTACCATGAATACATAATCATTTCCTATGCAGGA

AI773661  ----------------------------------------------------------------------------------------------
TC194329  CGAGATTCAAACCACAATTCTAAAAACACCTCGGTTTGTTATTATATACTAGTGAAGAGCACTACAGCATATTCTCTAATTATTCACTTGATTGATGAAA

AI773661  ------------------------------------------------AGTGTTCTGCCCAAGAGGGGCTCTGAGAGCAAGCTGATCAAAGGTCAAACATTC
TC194329  GAAAAGGCATAAAAATGCAAACTGATGAGAGAGAGAAGTTCCGTACAGTGTTCTGCCCAAGAGGGGCTCTGAGAGCAAGCTGATCAAAGGTCAAACATTC

AI773661  TCCTCCAGCCTTCTCAATCCTAGCTCTAGCTGTTTCCGTGAATCTCAATGCAGTAACCTTGATTTTTGGGACTTCATAAGCTCGAACATCATCGGTAACA
TC194329  TCCTCCAGCCTTCTCAATCCTAGCTCTAGCTGTTTCCGTGAATCTCAATGCAGTAACCTTGATTTTTGGGACTTCATAAGCTCGAACATCATCGGTAACA

AI773661  GTCCCAACAACAACAGCAATATTGTCCTCCTTTCCAGTCATGTAAGTAACCAAACGTGATAGAGACAATGGAGCTTTATTGATCTTGCTCATGAAGAG--
TC194329  GTCCCAACAACAACAGCAATATTGTCCTCCTTTCCAGTCATGTAAGTAACCAAACGTGATAGAGACAATGGAGCTTTATTGATCTTGCTCATGAAGAGTC

AI773661  ----------------------------------------------------------------------------------------------
TC194329  TCTTCAGTATCACAGCATTGAACTTACTACCAGTCCTCCGTGATAGAAATCGGTACAACTTGACGAGAAGCTTGAGATAAACATCGTCGGATTTTGGTGC

AI773661  ----------------------------------------------------------------------------------------------
TC194329  AATGCGCTTAGTCTTTTTGGACTTACCTCCGGCAACTAGATCGATACCCATGATGCTCCGCCTGCTGCTTCTACTTTCCTCCGCCGCCGCCTGCTGCTGA

AI773661  ------------------------
TC194329  TAGGTTTTTCTGTTCGGAGAATGG
```

In Tomato Gene Index (TGI), AI773661 is a member of TC194329.

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 thtat nearly identical sequences are not assembled
- Further quality checking:

  1) comparing unigene sequences against themselves to identify nearly identical sequences

  2) aligning EST sequences to their corresponding unigene sequences to identify mis-assembled ESTs
- Mis-assemblies were corrected automatically by the program

iAssembler

- Test data on CAC (test.fna): 9391 Roche/454 sequences with average length of 307.8bp and total of 2.89Mb
- Cap3 assembly

```
$ bin/cap3 test.fna -p 95 -o 30 -y 30
```

- MIRA assembly

```
$ bin/mira -project=mira1 -fasta=test.fna -job=denovo,est,normal,sanger -notraceinfo -
GENERAL:kcim=yes,not=1 -CO:fnicpst=yes -
CL:cpat=no,pec=no,pvlc=no,qc=no,bsqc=no,mbc=no,emlc=yes,mlcr=0,smlc=0,emrc=yes,mrcr
=0,smrc=0 -AL:bip=5,bmin=10,mrs=93,mo=29 -AS:mrl=30,bdq=30 -SK:mmhr=6
```

- iAssembler assembly

```
$ perl iAssembler.pl -i test.fna
```

	cap3	MIRA	iAssembler
time (seconds)	128	48	121
number of unigenes	6211	5937	5386
number of nearly identical sequence pairs (98%)	559	471	0

```
                1       10        20        30        40        50        60        70        80        90       100  105
                |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
F7FLTVN08JDOUJ  GTAGATGAAACGATCTCGTCGACGAGTTCTACGTTTTTTCTCTTCGCTAACGATAGGACTTACGAAGAAATCGTAAACCGCATCGTCGTCGAAGATTTTCTCGTA
F7FLTVN08JJH7Z  GTAGATGAAACGATCTCGTCGACGAGTTCTACGTTTTTTCTCTTCGCTAACGATAGGACTTACGAAGAAATCGTAAACCGCATCGTCGTCGAAGATTTTCTCGTA
      Consensus GTAGATGAAACGATCTCGTCGACGAGTTCTACGTTTTTTCTCTTCGCTAACGATAGGACTTACGAAGAAATCGTAAACCGCATCGTCGTCGAAGATTTTCTCGTA

                106     115       125       135       145       155       165       175       185       195       205  210
                |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
F7FLTVN08JDOUJ  TTTTTTCGAGATCTGCTGTGGTTTGTTCTAGTGTTTCGTTGGATTTAGCTACGTCGGCGAGAGCGTTCGCGTAGCTACCGGCGGCGGAATCTGCCATTTTTGCGC
F7FLTVN08JJH7Z  TTTTTTCGAGATCTGCTGTGGTTTGTTCTAGTGTTTCGTTGGATTTAGCTACGTCGGCGAGAGCGTTCGCGTAGCTACCGGCGGCGGAATCTGCCATTTTTGCGC
      Consensus TTTTTTCGAGATCTGCTGTGGTTTGTTCTAGTGTTTCGTTGGATTTAGCTACGTCGGCGAGAGCGTTCGCGTAGCTACCGGCGGCGGAATCTGCCATTTTTGCGC

                211     220       230       240       250       260       270       280       290       300       310  315
                |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
F7FLTVN08JDOUJ  CGAGTACACCGCGGCCGCGGCCGCGGCGGGAGGTGCGTTTGGATCGGAGTTTGAGTGTGAGTTTTGGGAGCTTGAGACCGCCGGAGTAGGATAGCTTTGCGACGG
F7FLTVN08JJH7Z  CGAGTACACCGCGGCCGCGGCCGCGGCGGGAGGTGCGTTTGGATCGGAGTTTGAGTGTGAGTTTTGGGAGCTTGAGACCGCCGGAGTAGGATAGCTTTGCGACGG
      Consensus CGAGTACACCGCGGCCGCGGCCGCGGCGGGAGGTGCGTTTGGATCGGAGTTTGAGTGTGAGTTTTGGGAGCTTGAGACCGCCGGAGTAGGATAGCTTTGCGACGG

                316     325       335       345       355       365       375       385       395       405       415
                |--------+---------+---------+---------+---------+---------+---------+---------+---------+---------|
F7FLTVN08JDOUJ  TTCCGTTGATGAGTTGAGTCGGCGGCGGTGAACGGGATT
F7FLTVN08JJH7Z  TTCCGTTGATGAGTTGAGTCGGCGGCGGTGAACGGGATTGGAAAGTTATCGGAGTTTGTTGTAGAGCCGCATTGATGAAAAATCAAATCTCTATGCTCGA
      Consensus TTCCGTTGATGAGTTGAGTCGGCGGCGGTGAACGGGATT...........................................................
```

iAssembler

parameters

```
[feizj@localhost iAssembler-1.0b.x32]$ perl iAssembler.pl

VERSION: v1.0-beta
USAGE:
        Perl  iAssembler.pl  [parameters]

        Input parameters
        -i      [String]        Name of the input sequence file in FASTA format (required)
        -q      [String]        Name of the quality file in FASTA format (default: none)
        -z      [String]        Name of the parameter configuration file (default: none)

        Assembly parameters
        -b      [String]        BLAST program used for clustering and alignments of ESTs
                                to their corresponding unigenes (megablast or blastn;
                                default = megablast)
        -a      [Integer]       number of CPUs used for blast program (default = 1)
        -e      [Integer]       maximum length of end clips (0~100; default = 30)
        -h      [Integer]       minimum overlap length (>=30; default = 30)
        -x      [Integer]       minimum percent identity for sequence clustering (95~99;
                                default = 97)
        -p      [Integer]       minimum percent identify for sequence assembly (95~100;
                                default = 95)

        Output parameters
        -u      [String]        prefix used for IDs of the assembled unigenes (default = UN)
                                iAssembler names the resulted unigenes with a prefix and
                                trailing numbers, e.g., UN00001
        -l      [Integer]       length of the trailing numbers in unigene IDs
                                (>= default; defalut = number characters of the maximum number
                                assigned to unigenes)
        -s      [Integer]       start number of unigene ID trailing number (>= 1; default = 1)
        -o      [String]        Name of the output directory (default = "input file name" +
                                "_output")
```

iAssembler

Output files

1. unigene_seq - unigene sequence file in FASTA format
2. contig_member - a tab-delimited txt file containing unigenes and their corresponding EST members.

```
1 UN0001  F7FLTVN08JRY0A  F7FLTVN08I5YLB  F7FLTVN08JLZ2L  F7FLTVN08JNO1H  F7FLTVN08JPS8A
  F7FLTVN08JB699  F7FLTVN08I7CEB  F7FLTVN08JRQX7  F7FLTVN08I4JGP  F7FLTVN08JWJ00
  F7FLTVN08I3PRQ  F7FLTVN08JCVFW  F7FLTVN08JDD3E  F7FLTVN08JKWD5  F7FLTVN08JRT4K
2 UN0002  F7FLTVN08JRSNN  F7FLTVN08I9Y8W  F7FLTVN08JGFS0  F7FLTVN08JT26G  F7FLTVN08JFU5L
  F7FLTVN08JUVZQ  F7FLTVN08I5H03  F7FLTVN08JQW3L
3 UN0003  F7FLTVN08I8FHT  F7FLTVN08JSC4T
```

3. unigene_mp - a tab-delimited txt file containing the mapping details of EST members to their corresponding unigenes

F7FLTVN08JKWD5	286	UN0001	468	1	286	7	292	1	98.61
F7FLTVN08JRT4K	233	UN0001	468	1	233	7	238	1	98.72
F7FLTVN08JRSNN	453	UN0002	867	5	453	320	768	1	100.00
F7FLTVN08I9Y8W	370	UN0002	867	1	370	501	867	-1	99.19
F7FLTVN08JGFS0	171	UN0002	867	1	167	253	419	-1	100.00

4. member_position_stat - A tab-delimited file containing the summary statistics of aligning ESTs to their corresponding unigenes.

Len/%ID	100-99	99-98	98-97	97-96	96-95	95-94	94-93	93-92	92-91	91-90	<90
000-100	0	0	0	0	0	0	0	0	0	0	0
100-200	1008	166	58	14	0	0	0	0	0	0	0
200-300	2002	337	137	23	0	0	0	0	0	0	0
300-400	3490	567	132	31	2	0	0	0	0	0	0
400-500	937	143	47	7	0	0	0	0	0	0	0
>500	9	1	0	0	0	0	0	0	0	0	0

SAM and ACE format outputs are in plan

iAssembler

2,442,651 (454) + 362,445 (Sanger) = 2,805,096 reads (740 Mb)

Strategy 1:

- Using the whole 2,805,096 ESTs as the input file, assemble them using iAssembler on a single CPU
- Total time spent: 23 days and 2 hours
- Around 2/3 of the time was spent on the first cycle of MIRA
- Maximum memory: ~50G
- MIRA supports multi-threads

Strategy 2:

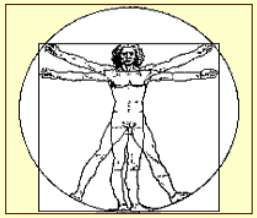
- Split the input files into 10 small files, assemble each file using iAssembler (5 hours for each)
- Combine the resulted unigenes, assemble them using iAssembler (43 hours)
- Remap members to unigenes, check and correct mis-assemblies (39 hours)
- Total time spent: ~90 hours (4 days)

transcriptome assembly

# unigene:	349,005
# contig:	147,708
# singleton:	201,297

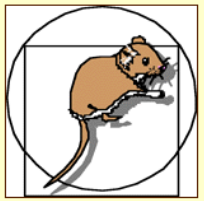
Release 14.0
(June 25, 2009)

Input Sequences	
ESTs	1456654
ETs	136359
Output Sequences	
TC sequences	34155
singleton ESTs	170849
singleton ETs	28496
Total unique: 233500	



Release 17.0
(July 28, 2006)

Input Sequences	
ESTs	7233257
HTs	234976
Output Sequences	
THC sequences	328301
singleton ESTs	736049
singleton HTs	19585
Total unique: 1083935	

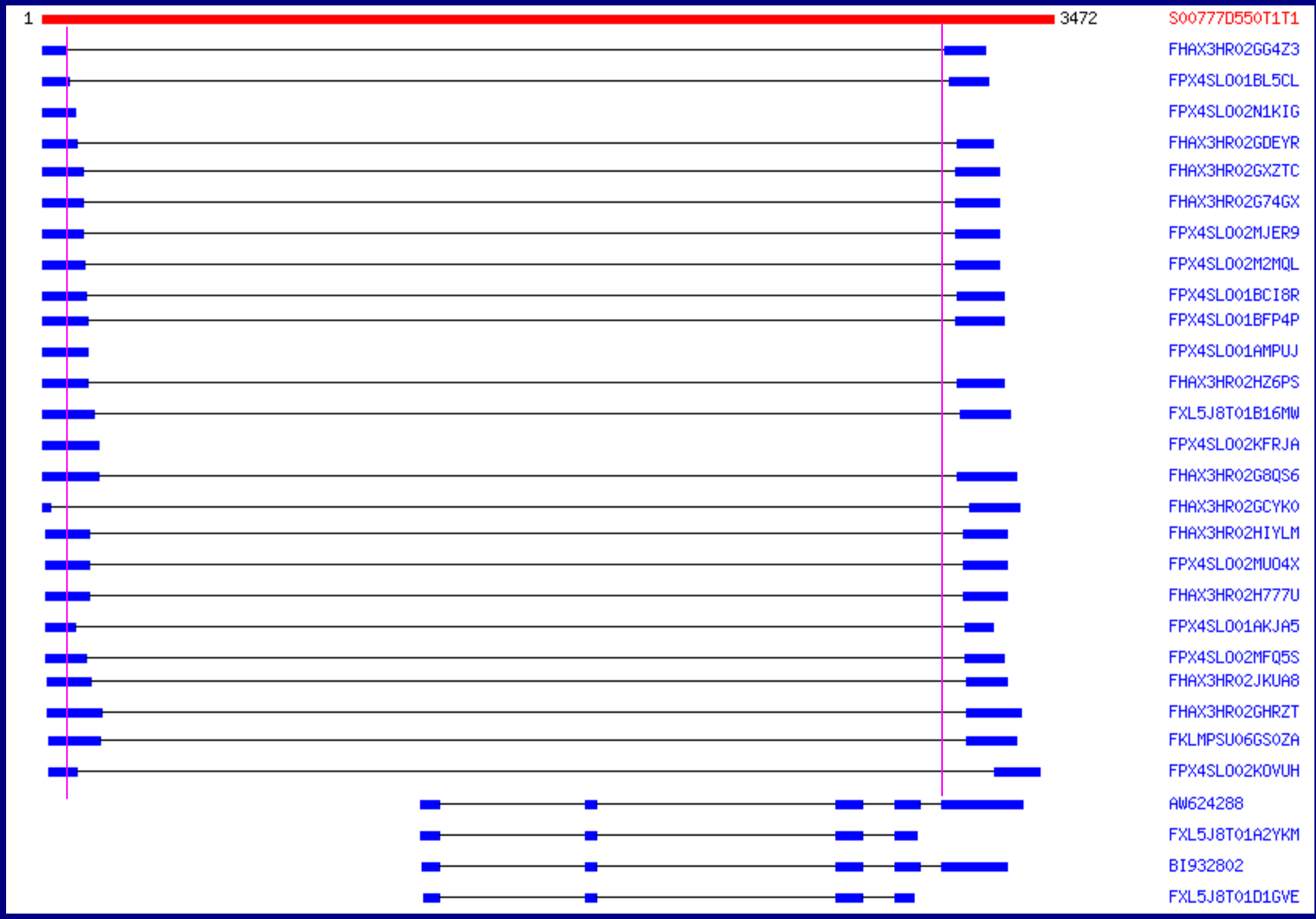


Release 16.0
(July 27, 2006)

Input Sequences	
ESTs	4591294
ETs	173495
Output Sequences	
TC sequences	210249
singleton ESTs	769704
singleton ETs	10684
Total unique: 990637	

Arabidopsis, human and mouse gene indices

Alternative splicing



Fragments can’t be connected