Welcome to the Plant Breeding and Genomics Webinar Series

Today’s Presenter: Dr. Candice Hansey
Presentation: http://www.extension.org/pages/60428

Host: Heather Merk
Technical Production: John McQueen

PBG home page: www.eXtension.org/plant_breeding_genomics
Sign up for PBG News: http://pbgworks.org
Please fill out the survey evaluation! (You will be contacted via email)

Watch past webinars and sign up for future webinars!
http://www.extension.org/pages/60426
How to Align Sequences

Presenter: Candice Hansey
hansey @msu.edu
Michigan State University
Overview

• Navigating NCBI to obtain sequences
• Using BLAST for sequence alignment
• Using other programs for specialized sequence alignment
• Next generation sequence alignment programs
Overview

• Navigating NCBI to obtain sequences
• Using BLAST for sequence alignment
• Using other programs for specialized sequence alignment
• Next generation sequence alignment programs
Goal

• Obtain sequence for the maize teosinte branched1 gene and then find organisms with orthologous genes
NCBI
Search specifying Zea Mays as the organism resulted in thousands of sequences
Results: 3

1. **tb1**
   - **Official Symbol**: tb1 and **Name**: teosinte branched1 (Zea mays)
   - **Other Aliases**: Z178A11.18
   - **Other Designations**: Transcription factor TEOSINTE BRANCHED 1; p-umc1082, teosinte branched protein 1; teosinte branched1 protein; umc1082
   - ID: 542361

2. **LOC100286105**
   - **teosinte branched1 protein** (Zea mays)
   - ID: 100286105

3. **LOC100282571**
   - **teosinte branched1 protein** (Zea mays)
   - ID: 100282571
Zea mays clone 494952 teosinte branched1 protein mRNA, complete cds

LOCUS       EUP75636
DEFINITION  Zea mays clone 494952 teosinte branched1 protein mRNA, complete cds.
ACCESSION  EUP75636
VERSION     EUP75636.1  GI:195656572
KEYWORDS    FLI, CUBH.
SOURCE      Zea mays
ORGANISM   Zea mays  
            Takahata; Viridiplantoceae; Streptophyta; Embryophyta; Tracheophyta;  
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; PACMAN  
            clade; Panicidoeae; Andropogoneae; Zea.
REFERENCE
AUTHORS     Alexandrov, N.N., Brooor, V.W., Fredin, S., Troukhin, N.E.,  
            Tatarinova, T.V., Zhang, H., Swaller, T.J., Lu, F.P., Bouch, J.,  
            Flavett, R.B. and Feldman, K.A.
TITLE       Insights into own genes derived from large-scale cDNA sequencing  
            Plant Mol. Biol. 69 (1-2), 179-194 (2009)
JOURNAL     Plant Mol. Biol.
PUBLISHER   182375398
REFERENCE
AUTHORS     Alexandrov, N.N., Brooor, V.W., Fredin, S., Troukhin, N.E.,  
            Tatarinova, T.V., Zhang, H., Swaller, T.J., Lu, F.P., Bouch, J.,  
            Flavett, R.B. and Feldman, K.A.
TITLE       Direct Submission  
            Submitted (04-Aug-2008) Corex, Inc., 1535 Rancho Conejo Blvd.,  
            Thousand Oaks, CA 91320, USA
FEATURES    Location/Qualifiers
            source 1..1024  
            /organism="Zea mays"  
            /mol_type="mRNA"  
            /db_xref="taxon:4873"  
            /clone="EUP75636"
            cds 48..854  
            /note="Similar to [NP_010663149.1]"  
            /codon_start=1  
            /product="teosinte branched1 protein"  
            /protein_id="EUP75636.1"  
            /db_xref="GI:195656573"  
            /translation="WDEIRKQVRPQGNNNIPFNNNNPFSIPDQLQNLNHEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
Zea mays clone 494952 teosinte branched1 protein mRNA, complete cds

GenBank: EU975636.1

>gi|195656572|gb|EU975636.1|  Zea mays clone 494952 teosinte branched1 protein mRNA, complete cds

AACCCTATTGCAAGCTGTTTTTGTCTTTATCGCCGCACGGATGTGCTGCTGCTTGAGGACGGGTA
CGCCGAGACATCTTTCCGCCCGAACAATGCTGCTGCTTCACCCGCAACAAAGGACACCTGGAGGGCCGGTTCC
CACACACCCGCCTGAGACAGACAGAGACGAGGCGCTGCTGAGGAAACAGCGGCTATATACCCGGTGTG
ACGACGCAGGGGCACCCGCAGCCACGGCGACGGGGGCGCTTCCGAGCCGACCAGCAAGAATCCGCAC
GCAGCAGGGGCTGCGGCACCGCCCGAGTGGTCTGGGCTCGGGGTCCGGCGCGGTGTTCTTTCGCGCCTGCA
GAGCCGGCTCGCTTCGACAAGGCCGCAAGACGTTGGAACTGTTGCTCTCTACCCAGTCAGGCAGCCCATCG
ACGGCCCTGCGAGCCCGGCGCGCGCCCGGAGGCCGGCTAGCAGCTGAGGCGCAGAGCCGGCTTCC
GACGTTGCTAGGGGCAAGAGCAGAGCCGAGTCTCTGAGCACTTTGCTGCTGCTGAGGACCTGAGAGAG
CCGCAGGACAGGAGGCAAGGGGAAACGGAAGAACAGAGGGCCCTGACGACGAGCCACGGGACACGGCTC
TGGAAGGACACGGCGCTGCGGCGGAGCTGCGCTGCTGATCTCAGTGGGGACGGCCACTACAGCAGCGGTG
ACGACGCGACGACGAGGACGGCGACAGGGGACGGGAAGTACTAATCCACAGGATTTGCTGCTGAGCTG
AGATGATGCGATGAGCAAGGAGGAAACGACGCGAAGTTGTTGTATGTAGATCGAGACGC
GTGATGATGCTTGGACCTCTTACACGCCGCTGAGAGATAACCTTCCATTACGCTCAGTAGATTATGCTGACT
GCTGGCACTGATGATCCATACCTCGATCTCTCACTGACATATAAATCTCAACTGCGCTAGCTTCTTA
TCTCATCAATCAAGATCCAAAGGGCCCAAAAAAAAAAAAAAAA
Overview

• Navigating NCBI to obtain sequences
• Using BLAST for sequence alignment
• Using other programs for specialized sequence alignment
• Next generation sequence alignment programs
BLAST

• BLAST = Basic Local Alignment Search Tool
BLAST

- BLAST = Basic Local Alignment Search Tool
- Used for searching a query sequence or sequences against a database or subject sequence
BLAST

• BLAST = Basic Local Alignment Search Tool
• Used for searching a query sequence or sequences against a database or subject sequence
• Uses a local alignment, which searches for regions in the query that locally align to subject sequences
BLAST

- BLAST = Basic Local Alignment Search Tool
- Used for searching a query sequence or sequences against a database or subject sequence
- Uses a local alignment, which searches for regions in the query that locally align to subject sequences
- BLAST uses word based heuristics to approximate the Smith-Waterman algorithm to find the near-optimal local alignments quickly, thus gaining speed over sensitivity.
BLAST

• BLAST = Basic Local Alignment Search Tool
• Used for searching a query sequence or sequences against a database or subject sequence
• Uses a local alignment, which searches for regions in the query that locally align to subject sequences
• BLAST uses word based heuristics to approximate the Smith-Waterman algorithm to find the near-optimal local alignments quickly, thus gaining speed over sensitivity.
• Two main versions: NCBI Blast and WU-BLAST
Local vs Global Alignments

Global  FTFTALILLAVAV
       F--TAL--LLA--AV

Local  FTFTALILL--AVAV
       --FTAL--LLAAV--

For sequences that are divergent, the optimal global alignment introduces gaps that can hide biologically relevant information such as motifs.
BLAST

• The process starts by searching the sequences for exact matches of small fixed length strings from the query called ‘words’.
BLAST

• The process starts by searching the sequences for exact matches of small fixed length strings from the query called ‘words’.
• These matches are the seeds for the local alignments.
BLAST

• The process starts by searching the sequences for exact matches of small fixed length strings from the query called ‘words’.
• These matches are the seeds for the local alignments.
• The next step is to extend the seed by aligning the sequence until a gap is found – mismatches are inserted here.
BLAST

• The process starts by searching the sequences for exact matches of small fixed length strings from the query called ‘words’.
• These matches are the seeds for the local alignments.
• The next step is to extend the seed by aligning the sequence until a gap is found – mismatches are inserted here.
• A gapped alignment is then performed using a modified Smith-Waterman algorithm – indels are added here.
The process starts by searching the sequences for exact matches of small fixed length strings from the query called ‘words’.

These matches are the seeds for the local alignments.

The next step is to extend the seed by aligning the sequence until a gap is found – mismatches are inserted here.

A gapped alignment is then performed using a modified Smith-Waterman algorithm – indels are added here.

Only results scoring above a threshold (expect value or e value) are reported back to the user.
BLAST Programs

• blastn - query DNA, subject DNA
• blastp - query Protein, subject Protein
• blastx - query DNA (6 frame translation) subject Protein
• tblastn - query protein - subject DNA (6 frame translation) - slow
• tblastx - query DNA (6 frame translation), subject DNA (6 frame translation) - very slow
Which Program Should You Use

What database do you have and how sensitive does your search has to be?

blastn, blastp – good for identifying sequences that are already in a database, finding local regions of similarity in closely related organisms.
Which Program Should You Use

What database do you have and how sensitive does your search has to be?

blastn, blastp – good for identifying sequences that are already in a database, finding local regions of similarity in closely related organisms.

blastx - Use when you have a nucleotide sequence with an unknown reading frame and/or sequencing errors that would lead to frame shifts or coding errors such as ESTs. More sensitive.
Which Program Should You Use

What database do you have and how sensitive does your search has to be?

blastn, blastp – good for identifying sequences that are already in a database, finding local regions of similarity in closely related organisms.

blastx - Use when you have a nucleotide sequence with an unknown reading frame and/or sequencing errors that would lead to frame shifts or coding errors such as ESTs. More sensitive.

tblastn- useful for finding homologs in sequence where the frame is unknown or sequencing errors are likely to be present such as ESTs and draft sequence.
Which Program Should You Use

What database do you have and how sensitive does your search have to be?

blastn, blastp – good for identifying sequences that are already in a database, finding local regions of similarity in closely related organisms.

blastx - Use when you have a nucleotide sequence with an unknown reading frame and/or sequencing errors that would lead to frame shifts or coding errors such as ESTs. More sensitive.

tblastn- useful for finding homologs in sequence where the frame is unknown or sequencing errors are likely to be present such as ESTs and draft sequence.

tblastx - used to detect novel ORFs/exons. Very Slow, use as the last resort.
Performing BLAST
Performing BLAST
Performing BLAST
Output from BLAST
## Descriptions

Legend for links to other resources: UniGene E GEO G Gene Structure Map Viewer PubChem BioAssay

### Sequences producing significant alignments:

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Max score</th>
<th>Total score</th>
<th>Query coverage</th>
<th>E value</th>
<th>Max Ident</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_001158993.1</td>
<td>Zea mays teosinte branched1 protein (LOC100286105), mRNA &gt;gb</td>
<td>EU975630.3</td>
<td>1847</td>
<td>1847</td>
<td>100%</td>
<td>0.0</td>
<td>100%</td>
</tr>
<tr>
<td>EU975630.3</td>
<td>Zea mays clone 499966 teosinte branched1 protein mRNA, complete cds</td>
<td>1833</td>
<td>1833</td>
<td>100%</td>
<td>0.0</td>
<td>99%</td>
<td>UG</td>
</tr>
<tr>
<td>BT63457.1</td>
<td>Zea mays full-length cDNA clone ZM_BF010512 mRNA, complete cds</td>
<td>1793</td>
<td>1793</td>
<td>98%</td>
<td>0.0</td>
<td>99%</td>
<td>UG</td>
</tr>
<tr>
<td>XM_002244342.1</td>
<td>Sorghum bicolor hypothetical protein, mRNA</td>
<td>628</td>
<td>628</td>
<td>73%</td>
<td>2e-176</td>
<td>77%</td>
<td>UG</td>
</tr>
<tr>
<td>NM_001158979.1</td>
<td>Zea mays teosinte branched1 protein (LOC100282571), mRNA &gt;gb</td>
<td>EU961498.3</td>
<td>630</td>
<td>630</td>
<td>80%</td>
<td>9e-168</td>
<td>75%</td>
</tr>
<tr>
<td>NM_001166844.2</td>
<td>Oryza sativa Japonica Group Os09g0410500 (Os09g0410500) mRNA, complete cds</td>
<td>315</td>
<td>315</td>
<td>24%</td>
<td>5e-82</td>
<td>88%</td>
<td>UG</td>
</tr>
<tr>
<td>FL707402.1</td>
<td>Oryza sativa Japonica Group putative TCP transcription factor (DPL) mRNA, complete cds</td>
<td>315</td>
<td>315</td>
<td>24%</td>
<td>5e-82</td>
<td>88%</td>
<td>UG</td>
</tr>
<tr>
<td>AP005093.3</td>
<td>Oryza sativa Japonica Group genomic DNA, chromosome 9, BAC clone:O1J2454</td>
<td>315</td>
<td>315</td>
<td>24%</td>
<td>5e-82</td>
<td>88%</td>
<td>UG</td>
</tr>
<tr>
<td>HQ85763.1</td>
<td>Zea mays clone UT1660 TCP transcription factor partial cds</td>
<td>289</td>
<td>289</td>
<td>21%</td>
<td>2e-74</td>
<td>89%</td>
<td>UG</td>
</tr>
<tr>
<td>BT61091.1</td>
<td>Zea mays full-length cDNA clone ZM_BF010512 mRNA, complete cds</td>
<td>289</td>
<td>289</td>
<td>21%</td>
<td>2e-74</td>
<td>89%</td>
<td>UG</td>
</tr>
<tr>
<td>NM_001158146.1</td>
<td>Zea mays teosinte-blanched one (LOC100285252), mRNA &gt;gb</td>
<td>EU971416.1</td>
<td>289</td>
<td>289</td>
<td>21%</td>
<td>2e-74</td>
<td>89%</td>
</tr>
<tr>
<td>NM_001133610.1</td>
<td>Zea mays TCP transcription factor (LOC100282571), mRNA &gt;gb</td>
<td>BT033200.1</td>
<td>289</td>
<td>289</td>
<td>21%</td>
<td>2e-74</td>
<td>89%</td>
</tr>
<tr>
<td>XM_002244342.1</td>
<td>Sorghum bicolor hypothetical protein, mRNA</td>
<td>289</td>
<td>289</td>
<td>21%</td>
<td>2e-74</td>
<td>89%</td>
<td>UG</td>
</tr>
<tr>
<td>AP003066.3</td>
<td>Oryza sativa Japonica Group genomic DNA, chromosome 8, BAC clone:O1J2454</td>
<td>280</td>
<td>280</td>
<td>20%</td>
<td>1e-71</td>
<td>89%</td>
<td>UG</td>
</tr>
<tr>
<td>NM_001196613.1</td>
<td>Zea mays hypothetical protein LOC100501452 (LOC100501452), mRNA &gt;gb</td>
<td>EU961498.3</td>
<td>277</td>
<td>277</td>
<td>21%</td>
<td>1e-70</td>
<td>87%</td>
</tr>
<tr>
<td>AJC57202.2</td>
<td>Zea mays clone ZMM88B7_C14, complete sequence</td>
<td>259</td>
<td>259</td>
<td>25%</td>
<td>4e-65</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>NM_001057663.1</td>
<td>Oryza sativa Japonica Group Os03g0706500 (Os03g0706500) mRNA, complete cds</td>
<td>161</td>
<td>161</td>
<td>16%</td>
<td>7e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>ACO21775.10</td>
<td>Oryza sativa chromosome 3 BAC OSN5b00004G17 genomic sequence, complete</td>
<td>161</td>
<td>161</td>
<td>16%</td>
<td>7e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>AY896402.1</td>
<td>Oryza sativa (indica cultivar-group) teosinte-branched 1 (Tb1) gene, complete cds</td>
<td>161</td>
<td>161</td>
<td>16%</td>
<td>7e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>AY941215.1</td>
<td>Oryza sativa teosinte branched1 protein (tb1) mRNA, complete cds</td>
<td>161</td>
<td>161</td>
<td>16%</td>
<td>7e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>AK027033.1</td>
<td>Oryza sativa Japonica Group cDNA clone:002-121-G01, full insert sequence</td>
<td>161</td>
<td>161</td>
<td>16%</td>
<td>7e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>AK22143.1</td>
<td>Oryza sativa Japonica Group genomic DNA, chromosome 8, BAC clone:O1J2454</td>
<td>161</td>
<td>161</td>
<td>16%</td>
<td>7e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>ADK88343.1</td>
<td>Oryza sativa Japonica Group 0sTB1 gene for teosinte branched1 protein, partial cds</td>
<td>161</td>
<td>161</td>
<td>16%</td>
<td>7e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>AC145260.1</td>
<td>Panicum virgatum cDNA PV_Aba010-K10, complete sequence</td>
<td>158</td>
<td>158</td>
<td>15%</td>
<td>9e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
<tr>
<td>FF904156.2</td>
<td>Pennisetum glaucum isolate Tb103 teosinte-branched1 (Tb1) gene, complete cds</td>
<td>158</td>
<td>158</td>
<td>15%</td>
<td>9e-35</td>
<td>81%</td>
<td>UG</td>
</tr>
</tbody>
</table>
BLAST

- For additional information on BLAST go to http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs
BLAST

• For additional information on BLAST go to http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs
• Web based BLAST interfaces are designed for low throughput searching
• Personalized databases can be generated and subsequent BLASTs performed using the command line
• Using the command line BLAST, multiple sequences can be aligned simultaneously
• For PC users I highly recommend getting Ubuntu
• This can be run as a virtual machine on your PC through programs such as VirtualBox https://www.virtualbox.org/
Overview

• Navigating NCBI to obtain sequences
• Using BLAST for sequence alignment
• Using other programs for specialized sequence alignment
• Next generation sequence alignment programs
EST and cDNA Alignment

Exonerate

a generic tool for sequence alignment

• Exonerate is a generic tool for pairwise sequence comparison (http://www.ebi.ac.uk/~guy/exonerate/) and comes grid ready with the ability to chunk files directly through exonerate
EST and cDNA Alignment

Exonerate

*a generic tool for sequence alignment*

- Exonerate is a generic tool for pairwise sequence comparison (http://www.ebi.ac.uk/~guy/exonerate/) and comes grid ready with the ability to chunk files directly through exonerate.
- Exonerate can be run with many different models for gapped and ungapped alignments.
EST and cDNA Alignment

Exonerate

*a generic tool for sequence alignment*

- Exonerate is a generic tool for pairwise sequence comparison (http://www.ebi.ac.uk/~guy/exonerate/) and comes grid ready with the ability to chunk files directly through exonerate
- Exonerate can be run with many different models for gapped and ungapped alignments
- The est2genom model can be used for alignment of EST sequences to genomic sequence and the cdna2genome model can be used to align full length cDNAs that can be flanked by UTRs
EST and cDNA Alignment

Exonerate

Exonerate is a generic tool for pairwise sequence comparison (http://www.ebi.ac.uk/~guy/exonerate/) and comes grid ready with the ability to chunk files directly through exonerate.

Exonerate can be run with many different models for gapped and ungapped alignments.

The est2genom model can be used for alignment of EST sequences to genomic sequence and the cdna2genome model can be used to align full length cDNAs that can be flanked by UTRs.

For additional models see the man page online at http://www.ebi.ac.uk/~guy/exonerate/exonerate.man.html or on the command line with the -h option

- exonerate -h
EST and cDNA Alignment

**Exonerate**

*a generic tool for sequence alignment*

- Command line example
  - `exonerate --query est_sequences.fasta --querytype DNA --target genome_sequence.fasta --targettype DNA --model est2genome --showalignment no --showvulgar no --showtargetgff no --ryo "%qi\t%ti\t%qab\t%qae\t%tab\t%tae\n" --minintron 10 --maxintron 1500 >exonerate_alignment.txt`
EST and cDNA Alignment

Exonerate

*a generic tool for sequence alignment*

• Command line example
  
  – exonerate --query est_sequences.fasta --querytype DNA --target genome_sequence.fasta --targettype DNA --model est2genome --showalignment no --showvulgar no --showtargetgff no --ryo “%qi\t%ti\t%qab\t%qae\t%tab\t%tae\n” --minintron 10 --maxintron 1500 >exonerate_alignment.txt

  – The output from this command will be a tab delimited file with the following columns for each alignment: query_id, target_id, query_start, query_end, target_start, target_end

  – Note: using this output format the positions are in interbase coordinates

• For complete list of options see the man page online at http://www.ebi.ac.uk/~guy/exonerate/exonerate.man.html or on the command line with the -h option
Whole Genome Alignment

• MUMmer – designed for rapid alignment of entire genomes
• NUCmer – designed for alignment of contigs to another set of contigs or a genome
• PROmer – designed for alignment of species too divergent for DNA alignment using six-frame translation of both input sequences
• http://mummer.sourceforge.net/
Whole Genome Alignment

• mummer can handle multiple reference and query sequences
• Command line usage
  – mummer [options] <reference-file> <query-files>
  – mummerplot [options] <match file>
• Example command line
  – mummer –mum –b –c genotype1.fasta genotype2.fasta > output.mums
    • -mum finds all maximal unique matches, -b will compute forward and reverse complement matches, -c reports all match positions relative to the forward strand
  – mummerplot –x “[0,275287]” –y “[0,265111]” –png output.mums
    • -x sets the x-axis range in the plot, -y sets the y-axis range in the plot, --png outputs the plot in .png format
• Additional options are available in the manual or by using the -h option for both mummer and mummerplot
• Examples for NUCmer and PROmer are available at http://mummer.sourceforge.net/examples/
Whole Genome Alignment

Forward MUMs are red and reverse MUMs are green. Dots on a line with a slope = 1 are unchanged, and dots on a line with a slope = -1 are inverted.

http://mummer.sourceforge.net/examples/
Short Sequence Alignment

• Vmatch is a tool that is ideal for aligning short sequences such as probes, primers, and SNPs with short context sequence
• This program requires a license
• http://www.vmatch.de/
Overview

• Navigating NCBI to obtain sequences
• Using BLAST for sequence alignment
• Using other programs for specialized sequence alignment
• Next generation sequence alignment programs
High Throughput Sequencing Platforms

- Illumina HiSeq 1000 and HiSeq 2000
- Illumina Genome Analyzer IIx
- Life Sciences/Roche 454 pyrosequencing
- Pacific Biosciences
- Ion Torrent
High Throughput Sequencing

• HiSeq 2000
• Highly parallel sequencing by synthesis
• Single and paired-end reads between 50 bp and 100 bp
• 187 million single end or 374 million paired-end reads per lane
• High error rate in the 3’ end
Quality scores are in ASCII characters that are converted to Phred scores. These scores provide a likelihood that the base was called incorrectly.

10 – 1 in 10 chance the base call is incorrect
20 – 1 in 100 chance the base call is incorrect
30 – 1 in 1000 chance the base call is incorrect

Suggest checking the quality of the reads prior to performing sequence alignments.
Read Quality with the FASTX-Toolkit

Introduction
The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

Next-Generation sequencing machines usually produce FASTA or FASTQ files, containing multiple short-reads sequences (possibly with quality information).

The main processing of such FASTA/FASTQ files is mapping (aka aligning) the sequences to reference genomes or other databases using specialized programs. Example of such mapping programs are: Blat, SHRIMP, LastZ, MAQ and many many others.

However, it is sometimes more productive to preprocess the FASTA/FASTQ files before mapping the sequences to the genome - manipulating the sequences to produce better mapping results.

The FASTX-Toolkit tools perform some of these preprocessing tasks.

http://hannonlab.cshl.edu/fastx_toolkit/
Read Quality with the FASTX-Toolkit

**FASTX Statistics**

```bash
$ fastx_quality_stats -h
usage: fastx_quality_stats [-h] [-i INFILE] [-o OUTFILE]
version 0.0.6 (C) 2008 by Assaf Gordon (gordon@csil.edu)
[-h] = This helpful help screen.
[-i INFILE] = FASTA/Q input file. default is STDIN.
[-o OUTFILE] = TEXT output file. default is STDOUT.

The output TEXT file will have the following fields (one row per column):
column = column number (1 to 36 for a 36-cycles read solexa file)
count = number of bases found in this column.
min = Lowest quality score value found in this column.
max = Highest quality score value found in this column.
sum = Sum of quality score values for this column.
mean = Mean quality score value for this column.
Q1 = 1st quartile quality score.
med = Median quality score.
Q3 = 3rd quartile quality score.
IQR = Inter-Quartile range (Q3-Q1).
LW = 'Left-Whisker' value (for boxplotting).
RW = 'Right-Whisker' value (for boxplotting).
A_Count = Count of 'A' nucleotides found in this column.
C_Count = Count of 'C' nucleotides found in this column.
G_Count = Count of 'G' nucleotides found in this column.
T_Count = Count of 'T' nucleotides found in this column.
N_Count = Count of 'N' nucleotides found in this column.
max-count = max. number of bases (in all cycles)
```

**FASTQ Quality Chart**

```bash
$ fastx_quality_boxplot_graph.sh -h
Usage: /usr/local/bin/fastx_quality_boxplot_graph.sh [-i INPUT.TXT] [-t TITLE] [-p] [-o OUTPUT]
[-i INPUT.TXT] = Input file. Should be the output of "solexa_quality_statistics" program.
[-o OUTPUT] = Output file name. default is STDOUT.
[-t TITLE] = Title (usually the solexa file name) - will be plotted on the graph.
```

>fastx_quality_stats -i sequence_file.fastq -o stats.txt

>fastx_quality_boxplot_graph.sh -i stats.txt -t sample1 -o quality.png
Read Quality with the FASTX-Toolkit

Bad Sequence

Good Sequence
High Throughput Sequence Alignment

• Traditional sequence alignment algorithms cannot be scaled to align millions of reads
High Throughput Sequence Alignment

- Traditional sequence alignment algorithms cannot be scaled to align millions of reads
- Utilize genome indexing such as Burrows-Wheeler for ultrafast and memory-efficient alignment programs
High Throughput Sequence Alignment

• Traditional sequence alignment algorithms cannot be scaled to align millions of reads
• Utilize genome indexing such as Burrows-Wheeler for ultrafast and memory-efficient alignment programs
• Next generation sequence alignment algorithms are rapidly evolving to accommodate the increasing sequence throughput
Tuxedo Suite

• Bowtie – fast and quality aware short read aligner for aligning DNA and RNA sequence reads (http://bowtie-bio.sourceforge.net/index.shtml)
• TopHat – fast, splice junction mapper for RNA-Seq reads built on the Bowtie aligner (http://tophat.cbcb.umd.edu/)
• Cufflinks – assembles transcripts, estimates their abundances, and test for differential expression and regulation using the alignments from Bowtie and TopHat
Bowtie

• Available for Windows, Mac OS X, Linux, and Solaris
Bowtie

• Available for Windows, Mac OS X, Linux, and Solaris
• Not a general-purpose alignment tool like MUMmer, BLAST, or Vmatch
• Ideal for aligning short reads to large genomes
Bowtie

- Available for Windows, Mac OS X, Linux, and Solaris
- Not a general-purpose alignment tool like MUMmer, BLAST, or Vmatch
- Ideal for aligning short reads to large genomes
- Forms the basis for TopHat, Cufflinks, Crossbow, and Myrna
Bowtie

• Available for Windows, Mac OS X, Linux, and Solaris
• Not a general-purpose alignment tool like MUMmer, BLAST, or Vmatch
• Ideal for aligning short reads to large genomes
• Forms the basis for TopHat, Cufflinks, Crossbow, and Myrna
• Online manual (http://bowtie-bio.sourceforge.net/manual.shtml) is very helpful to understand the options that are available
Bowtie

• Bowtie Index
  – To print a help screen with the optional parameters
    • bowtie-build –h
  – Command line usage
    • bowtie-build [options]* <reference_in> <ebwt_base>
  – Example command line
    • bowtie-build chr1.fa,chr2.fa,chr3.fa,chr4.fa test_index
    • Additional options can be used to improve performance
Bowtie

• Bowtie Index
  – To print a help screen with the optional parameters
    • bowtie-build –h
  – Command line usage
    • bowtie-build [options]* <reference_in> <ebwt_base>
  – Example command line
    • bowtie-build chr1.fa,chr2.fa,chr3.fa,chr4.fa test_index
    • Additional options can be used to improve performance

• Alignment with Bowtie
  – To print a help screen with the optional parameters
    • bowtie –h
  – Command line usage
    • bowtie [options]* <ebwt> {-1 <m1> -2 <m2> | --12 <r> | <s>} [hit]
  – Example command line for single end reads with our test_index
    • Bowtie test_index --solexa1.3-quals test_index sequence_file.fastq > output_file
Bowtie

- Additional Options:
- Quality aware MAQ-like mode is the default, Bowtie can also be run in a non-quality aware SOAP-like mode
  - Bowtie test_index --solexa1.3-quals -v 2 test_index sequence_file.fastq > output_file
  - -v 2 is specifying that there can only be two mismatches in an alignment
- -a report all valid alignments
- -k 2 report up to 2 valid alignments
- --best report the best alignment
- -m 2 do not report any alignments for reads with greater than 2 reportable alignments
- -S print alignments in SAM format which is compatible with SAMtools for subsequent variant calling and manipulation of the alignments
- Many other options available
TopHat

- Available for Linux and OS X
TopHat

• Available for Linux and OS X
• Built on Bowtie and uses the same genome index
TopHat

- Available for Linux and OS X
- Built on Bowtie and uses the same genome index
- Used for alignment of RNA-Seq reads to a genome
TopHat

- Available for Linux and OS X
- Built on Bowtie and uses the same genome index
- Used for alignment of RNA-Seq reads to a genome
- Optimized for paired-end, Illumina sequence reads >70bp
TopHat

- Available for Linux and OS X
- Built on Bowtie and uses the same genome index
- Used for alignment of RNA-Seq reads to a genome
- Optimized for paired-end, Illumina sequence reads >70bp
- Online manual (http://tophat.cbcb.umd.edu/manual.html) is very helpful to understand the options that are available
TopHat

Map reads to whole genome with Bowtie

Collect initially unmappable reads

Assemble consensus of covered regions

Generate possible splices between neighboring exons

Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend
TopHat

• **Alignment with TopHat**
  
  – To print a help screen with the optional parameters
    • `tophat --h`
  
  – Command line usage
    • `tophat [options]* <index_base> <reads1_1[,...,readsN_1]> [reads1_2,...readsN_2]`
  
  – Example command line for single end reads with our test_index from before
    • `Tophat --o output_directory --solexa1.3-quals test_index sequence_file.fastq`
  
  – Additional options
    • `--max-intron size` (default is 500kb)
    • `--min-intron size` (default is 70bp)
    • `--max-multihits` (default is 40)
    • `--mate-inner-dis` (required for paired-end alignment mode)
    • `--num_threads N` (runs on N CPUs)
• Wiggle tracks can be generated from the TopHat output file accepted_hits.bam
• Convert to SAM file
  – samtools view –h –o accepted_hits.sam accepted_hits.bam
• Generate wiggle file
  – wiggles accepted_hits.sam coverage.wig
• Wiggle tracks are viewable in programs such as Integrated Genome Browser (http://bioviz.org/igb/)
Additional Alignment Programs

How to map billions of short reads onto genomes

Cole Trapnell & Steven L Saizberg

Mapping the vast quantities of short sequence fragments produced by next-generation sequencing platforms is a challenge. What programs are available and how do they work?

Nature Biotechnology 27, 455-457 (2009)

<table>
<thead>
<tr>
<th>Program</th>
<th>Website</th>
<th>Open source?</th>
<th>Handles ABI color space?</th>
<th>Maximum read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowtie</td>
<td><a href="http://bowtie.ccbcb.umd.edu">http://bowtie.ccbcb.umd.edu</a></td>
<td>Yes</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>BWA</td>
<td><a href="http://maq.sourceforge.net/bwa-man.shtml">http://maq.sourceforge.net/bwa-man.shtml</a></td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>Maq</td>
<td><a href="http://maq.sourceforge.net">http://maq.sourceforge.net</a></td>
<td>Yes</td>
<td>Yes</td>
<td>127</td>
</tr>
<tr>
<td>Mosaik</td>
<td><a href="http://bioinformatics.bc.edu/marthlab/Mosaik">http://bioinformatics.bc.edu/marthlab/Mosaik</a></td>
<td>No</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>Novoalign</td>
<td><a href="http://www.novocraft.com">http://www.novocraft.com</a></td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>SOAP2</td>
<td><a href="http://soap.genomics.org.cn">http://soap.genomics.org.cn</a></td>
<td>Yes</td>
<td>No</td>
<td>60</td>
</tr>
<tr>
<td>ZOOM</td>
<td><a href="http://www.bioinfor.com">http://www.bioinfor.com</a></td>
<td>No</td>
<td>Yes</td>
<td>240</td>
</tr>
</tbody>
</table>
Valuable Resources
Questions