


PBIO/BIOL3240L: The Dynamic Genome, Freshmen College Summer
Experience
Summer 2008 

Table of Contents

Syllabus

Grading Policy

Chapter 1: Introduction to NCBI Website and Bioinformatics	1-17
A. PubMed	1-5
B. Blast	6-17
Chapter 2: Transposable Element Background and Experiment 1	18-61
A. Background to TEs	18-25
B. Background to Osmar Experiment	26-33
C. Protocols for Osmar Experiment	34-52
D. DNA Sequence Analysis	52-61
Chapter 3: Getting a bit deeper into TEs	62-70
A. TE families	62-66
B. Class 1 Elements	67-70
Chapter 4: Identifying LTR Retros: A Bioinformatics Experiment	71-83
Chapter 5: Phylogenetics using TATE	84-100
A. Background	84
B. Using TATE	85-86
C. Understanding the results	87-100


Syllabus Summer 2008

Syllabus

Grading

Data

Course Links

	Lecture Topic and Material	Experiment Topic and Material	Handouts
Monday, June 30			
Tuesday, July 1			
Wednesday, July 2	Read <i>Making of the Fittest</i> , Review background in <i>Life</i>		
Thursday, July 3	Read <i>Making of the Fittest</i> , Review background in <i>Life</i>		
Friday, July 4		Independence Day	
Monday, June 7	<ul style="list-style-type: none"> • Review of DNA, RNA, and proteins. • TEs and <i>Making of Fittest</i> 	<ul style="list-style-type: none"> • Lab safety • Pipetting school 	<ul style="list-style-type: none"> • First Lecture Notes

Tuesday, July 8	<ul style="list-style-type: none"> • TEs and <i>Making of Fittest</i>, Group Discussion • PubMed 	<ul style="list-style-type: none"> • Agarose gels 	<ul style="list-style-type: none"> • Second Lecture
Wednesday, July 9	<ul style="list-style-type: none"> • NCBI Blast (pages 7-17) • Background Experiment I (pages 25-33) 	<ul style="list-style-type: none"> • Observe plants (page 33) 	<ul style="list-style-type: none"> • Blast handout
Thursday, July 10	<ul style="list-style-type: none"> • Background Experiment I • Background on PCR 	<ul style="list-style-type: none"> • Extract DNA (page 34-35) 	
Friday, July 11	Plant Biology Greenhouse Tour	<ul style="list-style-type: none"> • Setup PCR reactions 	<ul style="list-style-type: none"> • First Assignment
Monday, July 14	<ul style="list-style-type: none"> • RNA Elements 	<ul style="list-style-type: none"> • Gel of PCR • Purify and clone bands 	
Tuesday, July 15	<ul style="list-style-type: none"> • RNA Elements 	<ul style="list-style-type: none"> • Topo clone 	
Wednesday, July 16	<ul style="list-style-type: none"> • RNA Elements 	<ul style="list-style-type: none"> • Overnights 	<ul style="list-style-type: none"> • RNA Handout

<p>Thursday, July 17</p>	<ul style="list-style-type: none"> • Mid-term -- take home • 	<ul style="list-style-type: none"> • Mini-prep • Prepare DNA sequencing 	<ul style="list-style-type: none"> • Mid-Term
<p>Friday, July 18</p>	<ul style="list-style-type: none"> • Experimental Design • Primer Design 	<p>Isolate DNA</p> <p>Gel</p>	
<p>Monday, July 21</p>	<p>Finding whole RNA Elements (when time permits)</p>	<p>Projects</p>	
<p>Tuesday, July 22</p>		<p>Projects</p>	
<p>Wednesday, July 23</p>		<p>Projects</p>	
<p>Thursday, July 24</p>		<p>Projects</p>	<ul style="list-style-type: none"> • Assignment 2 due Thurs. July 24 (tonight). • Assignment 3 posted Friday
<p>Friday, July 25</p>		<p>Projects</p>	

Monday, July 28	Work on Posters		<ul style="list-style-type: none">• Poster template
Tuesday, July 29	Work on Posters		
Wednesday, July 30	Print Posters, Last Discussion		
Thursday, July 31	Poster Session		
Friday, July 25			

BIO/PBIO 3240L The Dynamic Genome
Syllabus Summer Mini-2 2008

Dr Susan Wessler and Dr Jim Burnette

Eunyoung Cho, TA

Monday - Friday 9:15 - 11:30

Course website: http://www.dynamicgenome.org/classes/summer_2008/

	Dr. Susan Wessler	Dr. Jim Burnette	Eunyoung Cho
Office	Plant Sciences 4510	Plant Sciences 1506	Plant Sciences 4505
Phone	706-542-1870	706-542-4581	706-542-1857
Hours	By appointment	By appointment	By appointment
E-mail	sue@plantbio.uga.edu	jburnette@plantbio.uga.edu	echo@plantbio.uga.edu

Attendance: We require 100% attendance and class participation. Any missed lab will be difficult to make up. If you know you will be absent for any class, make arrangements in advance with the instructor. Discuss unplanned absences immediately upon returning to class.

Class participation is a major part of a summer lab course. You are expected to be prepared for each day, participate in all discussions, and ask a lot of questions. Twenty percent of your grade is based on class participation.

For your safety, you must wear closed toe shoes (no flip-flops or sandals). Long shorts are permitted. Long hair should be pulled back away from the face for all labs. Goggles and gloves will be provided.

To be completed before the first day of class (Monday, July 7):

(i) *Making of the Fittest* by Sean Carroll. You will be given a study guide so that you can prepare for class discussion. The book will be provided by the instructors.

(ii) Review DNA, RNA, protein background information using a study guide that you will be given along with the first part of an Introductory Biology Textbook (also provided by the instructors at no cost to you).

Week 1:

Introduction to Transposable Elements, bioinformatics, and experiment 1.
Introduction to lab techniques.

Experiment 1: Analyzing active transposable elements in Arabidopsis.

Week 2:

Introduction to group projects.

Finish Experiment 1 and start group projects.

Week 3:

Group projects and discussions.

Week 4:

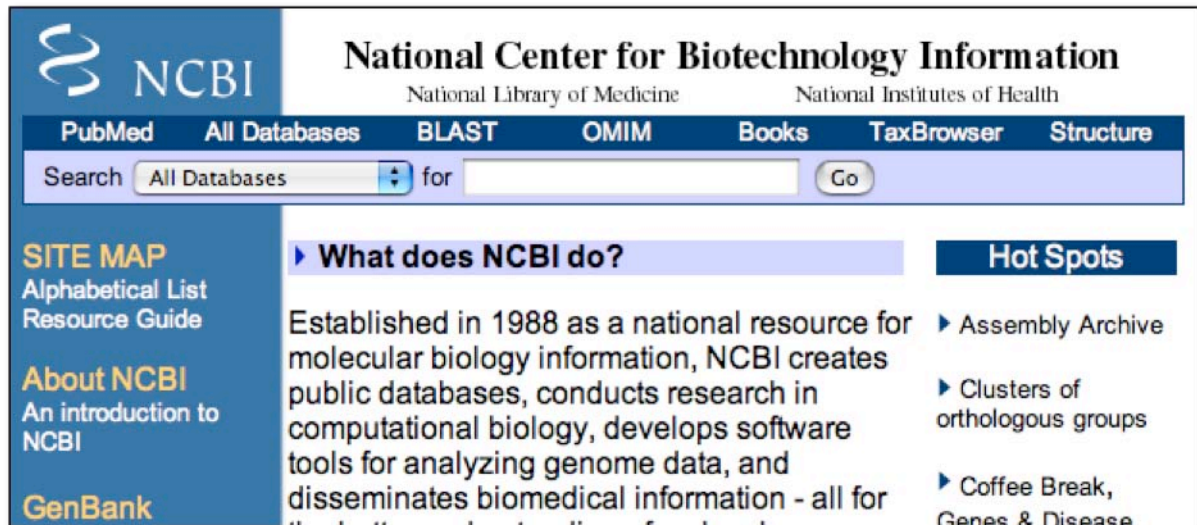
Finish group projects and make posters.

Grading:

Class Participation (includes attendance)	20%
1 hour mid-term (July 17 th)	10%
Project proposal	10%
Poster (July 31 st)	15%
Poster presentation	15%
Take home Assignments	20%
Quizzes (includes lab notebook checks)	<u>10%</u>
	100%

Introduction to the NCBI website: PubMed and Blast

Biological sequence data and journal articles are collected, indexed, and made available by the National Center for Biotechnology Information (NCBI). NCBI is a unit of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). Because it is a part of the NIH, the collections of sequence data and journal articles are available free to anyone at <http://www.ncbi.nlm.nih.gov/>. This is what the NCBI home page (currently) looks like....



NCBI provides tools for searching and downloading the databases it maintains through the web portal NCBI Entrez. PubMed is searched with text queries using the Entrez portal. PubMed is an index of thousands of biological journals going back as far as 1950. It also contains thousands of full-length articles in PDF format available for free download in a collection called PubMed Central.

NCBI also contains a collection of biological sequence databases. These are informally referred to as GenBank. The biological sequences are divided into DNA sequences (which includes RNA sequences) including GenBank proper, Protein sequence, and Genome Sequence. The sequence database collection is the result of collaborations between NCBI, DDBJ (Japan), and EMBL (Europe). Although the file formats and search tools may differ between the three repositories, they are essentially redundant at the data level. Most data in GenBank are in the public domain although some sequence data are patented.

Accessing GenBank sequence: GenBank sequence is usually accessed in one of two ways. A simple text search can be used to find sequences by name, author(s), and/or other supporting information. A more sophisticated search of GenBank uses a sequence query and a collection of tools called Blast. Blast will be described in detail later in this tutorial.

Other useful and interesting databases maintained by NCBI:

The Entrez portal also includes several databases that many people find useful. We will cover only one of these in this tutorial.

- TaxBrowser - This database provides taxonomic information on most extant and some extinct organisms. This is useful to explore the relationships between organisms. An easier to understand (but less complete) taxonomy web site is the Tree of Life website <http://www.tolweb.org/tree/>.
- Books - Books is a virtual library of out-of-print editions of books. Although out-of-print, many are still useful and all are free.
- OMIM - Online Mendelian Inheritance in Man. - OMIM is a database of articles on human genes associated with diseases and medical conditions. Each article is hand curated by people who read and summarize journal articles. This database is incredibly rich with information on human genes, diseases, and population genetics. An OMIM example will be covered later in this tutorial.

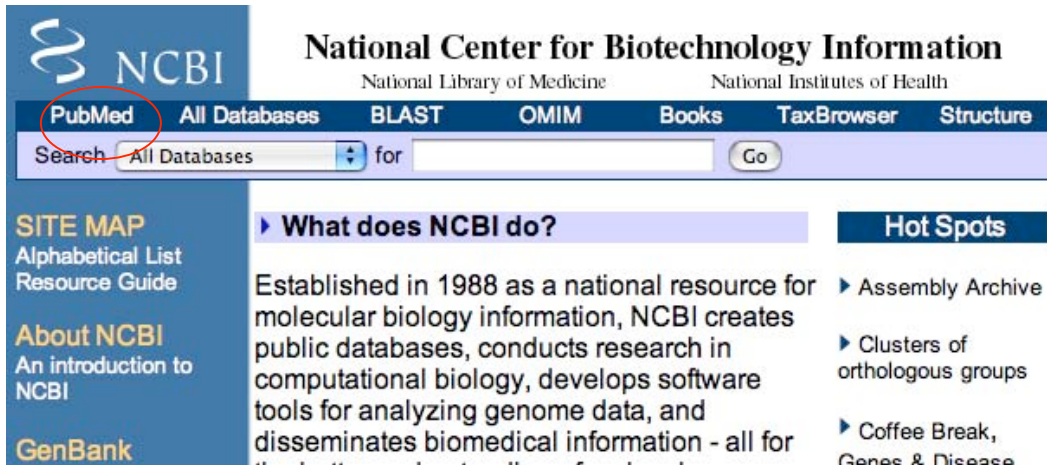
Let's begin our tour by visiting the PubMed site and then move on to the BLAST site where of our time in the rest of the course.

I. **PubMed**: Literature searches about a biological problem are very easy. PubMed makes the index available on its website with no access limitations. You can use PubMed (and Blast) from any computer or terminal with an Internet connection. (There is even a special access page for mobile phones.)

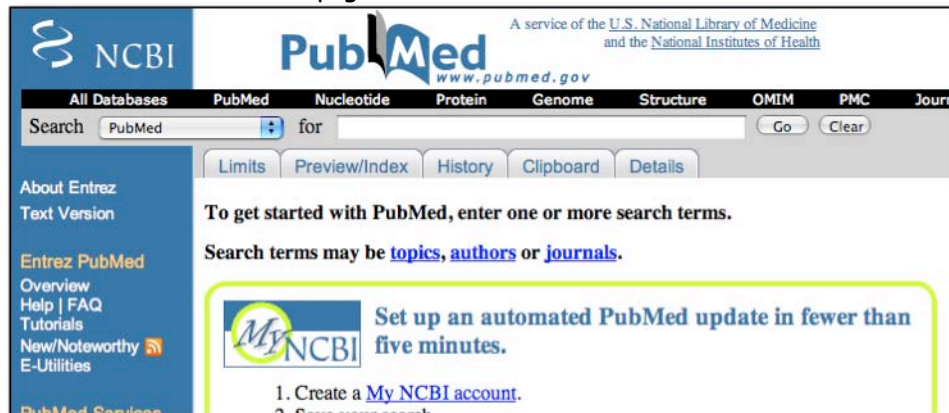
While searching the database and reading abstracts is free, accessing a full-length article may require a subscription with the article's publisher. Many universities have subscriptions and access is easy if you are on a university network. Many articles will be freely available either directly from the publisher or through PubMed Central.

Steps for a PubMed search:

1. Open NCBI in a web browser by going to the NCBI home page and click on PubMed in the bar. To get to the PubMed home page click on PubMed which is part of the main menu at the top.



This is the PubMed homepage...

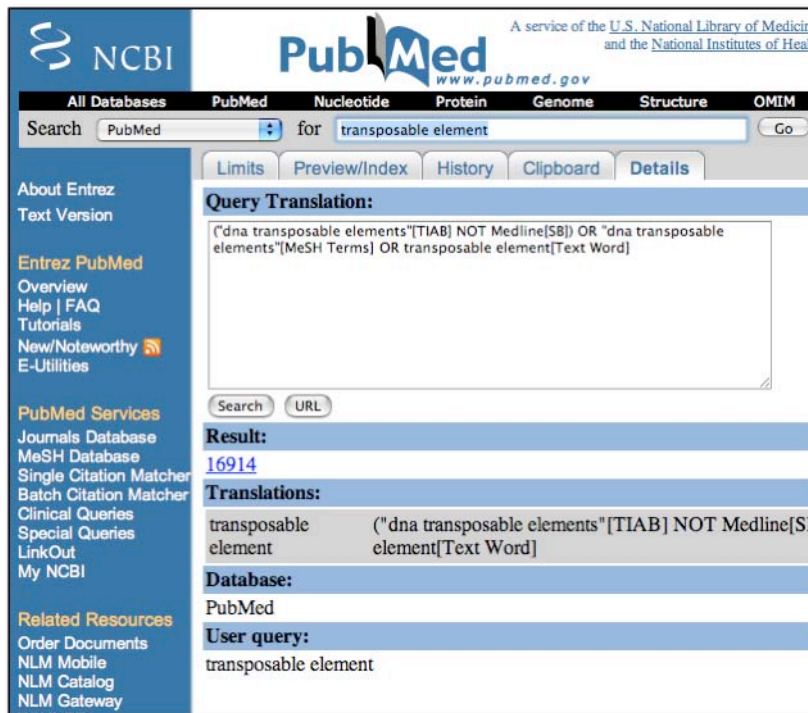


2. Think about the topic you want to search. You can use keywords, author last names, journal titles, publication year, or institution.

For the first search enter 'transposable element' and click 'Go.' The result of the search is shown below.



3. Before we discuss the results, click on the 'Details' tab. This will show you the details of the search that was performed by PubMed's search engine.



While you thought you were just searching "transposable element" PubMed was actually using these search terms (with added comments):

- | | |
|---|--|
| ("dna transposable elements"[TIAB] NOT Medline[SB]) | •Search titles and abstracts, not the medline subset |
| OR "dna transposable elements"[MeSH Terms] | •Seach Medline Subject Headings |
| OR transposable element[Text Word] | •Search all text |

As you can see, the simple query 'transposable element' is expanded into a more structured query by PubMed. MeSH is a controlled vocabulary for indexing PubMed. Curators at NCBI and journal editors assign these keywords based on suggestions by authors.

Because of this query expansion it is a good idea to check the 'Details' tab whenever a search gives no results or unexpected results.

4. Click on the Browser's Back Button. The icons and other details of the results list like the one shown below will be discussed in class.

Display Summary Show 20 Sort By Send to

All: 16914 Review: 1196

Items 1 - 20 of 16914 Page 1 of 846 Next

1: [Mahmoud AA, Sukumar S, Krishnan HB.](#)
Interspecific Rice Hybrid of *Oryza sativa* x *Oryza nivara* Reveals a Significant Increase in Seed Protein Content.
J Agric Food Chem. 2007 Dec 29; [Epub ahead of print]
PMID: 18163552 [PubMed - as supplied by publisher] Related Articles, Links

2: [Van K, Onoda S, Kim MY, Kim KD, Lee SH.](#)
Allelic variation of the Waxy gene in foxtail millet [*Setaria italica* (L.) P. Beauv.] by single nucleotide polymorphisms.
Mol Genet Genomics. 2007 Dec 19; [Epub ahead of print]
PMID: 18157676 [PubMed - as supplied by publisher] Related Articles, Links

3: [Mukherjee S, Chakraborty R.](#)
Conjugation potential and class 1 integron carriage of resident plasmids in river water copiotrophs.
Acta Microbiol Immunol Hung. 2007 Dec;54(4):379-97.
PMID: 18088011 [PubMed - indexed for MEDLINE] Related Articles, Links

4: [Fontanillas P, Hartl DL, Reuter M.](#)
Genome organization and gene expression shape the transposable element distribution in the *Drosophila melanogaster* euchromatin.
PLoS Genet. 2007 Nov 30;3(11):e210. Epub 2007 Oct 10.
PMID: 18081425 [PubMed - in process] Related Articles, Links

5: [Metcalfe CJ, Bulazel KV, Ferreri GC, Schroeder-Reiter E, Wanner G, Rens W, Obergfell C, Eldridge MD, O'Neill RJ.](#)
Genomic instability within centromeres of interspecific marsupial hybrids.
Genetics. 2007 Dec;177(4):2507-17.
PMID: 18073443 [PubMed - in process] Related Articles, Links

5. Click on one of the underlined authors (in blue). This will give you detailed information about the article including the abstract. The abstract provides a detailed summary of the paper. On the right are two icons. Clicking on either of those will take you to a download page for the full article. The Related Links section is also a useful area to help refine searches and will be discussed in class.

Display AbstractPlus Show 20 Sort By Send to

All: 1 Review: 0

1: [BMC Evol Biol.](#) 2007 Aug 29;7:152.

Full text free on... Links

Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*.

[Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA.](#)

Arizona Genomics Institute, Department of Plant Sciences, BIOS Institute, University of Arizona, Tucson, AZ 85721, USA. azuccolo@ag.arizona.edu

BACKGROUND: The genus *Oryza* is composed of 10 distinct genome types, 6 diploid and 4 polyploid, and includes the world's most important food crop - rice (*Oryza sativa* [AA]). Genome size variation in the *Oryza* is more than 3-fold and ranges from 357 Mbp in *Oryza glaberrima* [AA] to 1283 Mbp in the polyploid *Oryza ridleyi* [HHJJ]. Because repetitive elements are known to play a significant role in genome size variation, we constructed random sheared small insert genomic libraries from 12 representative *Oryza* species and conducted a comprehensive study of the repetitive element composition, distribution and phylogeny in this genus. Particular attention was paid to the role played by the most important classes of transposable elements (Long Terminal Repeats Retrotransposons, Long Interspersed Nuclear Elements, helitrons, DNA transposable elements) in shaping these genomes and in their contributing to genome size variation. **RESULTS:** We identified the elements primarily responsible for the most strikingly genome size variation in *Oryza*. We demonstrated how

Related Links

- Evolutionary history of *Oryza sativa* LTR retrotransposons: a preliminary survey of the rice genome sequ [BMC Genomics. 2004]
- Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in t [Plant J. 2007]
- Differential lineage-specific amplification of transposable elements is responsible for genome size variat [Genome Res. 2006]
- Long terminal repeat retrotransposons of *Oryza sativa*. [Genome Biol. 2002]
- Dasheng and RIRE2. A nonautonomous long terminal repeat element and its putative autonomous partner i [Plant Physiol. 2002]

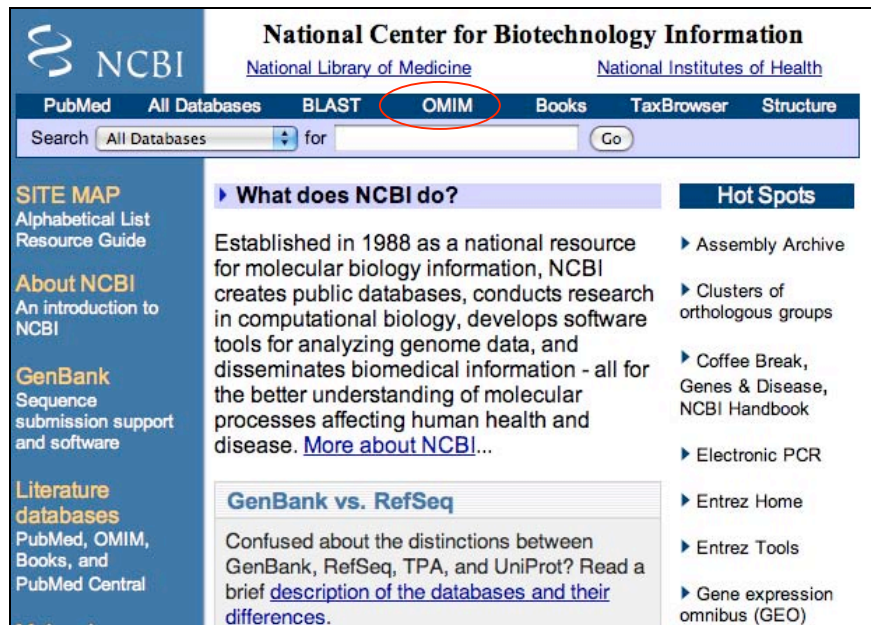
See all Related Articles...

6. Learning to search PubMed and all of the features takes time and practice. Research a topic that is interesting to you.

II. OMIM

This short introduction will get you acquainted with OMIM. With OMIM you can learn about a genetic disease, find examples for class and tests.

1. Click "OMIM" on the Entez Portal.



The screenshot shows the NCBI website interface. At the top, the NCBI logo is on the left, and the text "National Center for Biotechnology Information" is centered, with "National Library of Medicine" and "National Institutes of Health" below it. A navigation bar contains links for "PubMed", "All Databases", "BLAST", "OMIM" (circled in red), "Books", "TaxBrowser", and "Structure". Below the navigation bar is a search box with a dropdown menu set to "All Databases" and a "Go" button. On the left side, there is a "SITE MAP" section with links for "Alphabetical List Resource Guide", "About NCBI", "GenBank", and "Literature databases". The main content area features a "What does NCBI do?" section with a paragraph about NCBI's history and mission, and a "Hot Spots" section with a list of links including "Assembly Archive", "Clusters of orthologous groups", "Coffee Break, Genes & Disease, NCBI Handbook", "Electronic PCR", "Entrez Home", "Entrez Tools", and "Gene expression omnibus (GEO)".

2. Enter "transposable element" in the text box and Click "Go". You search OMIM with text queries similar to PubMed. If you wanted to search on a specific disease you would enter the disease name.

NCBI

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

All Databases PubMed Nucleotide Protein

Search OMIM for transposable element Go Clear

Limits Preview/Index History Clipboard Details

- Enter one or more search terms.
- Use **Limits** to restrict your search by search field, chromosome, and other criteria.
- Use **Index** to browse terms found in OMIM records.
- Use **History** to retrieve records from previous searches, or to combine searches.

Entrez

OMIM
Search OMIM
Search Gene Map
Search Morbid Map

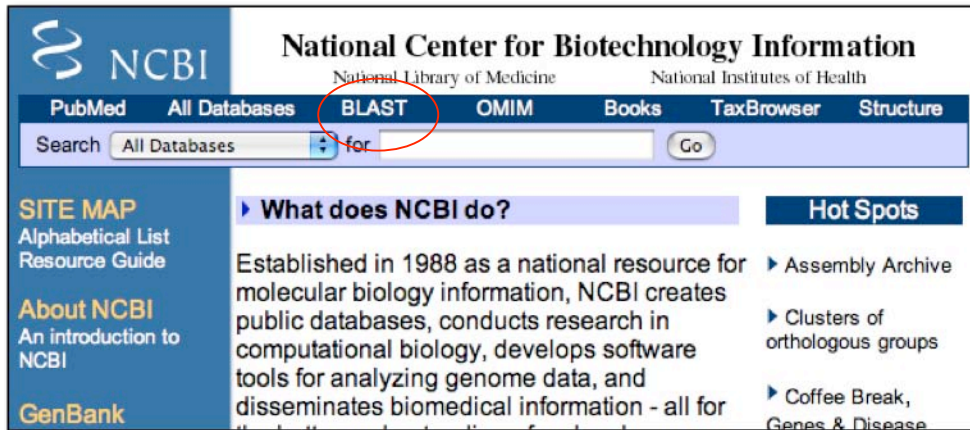
Help
OMIM Help
How to Link

OMIM™ - Online Mendelian Inheritance in Man™

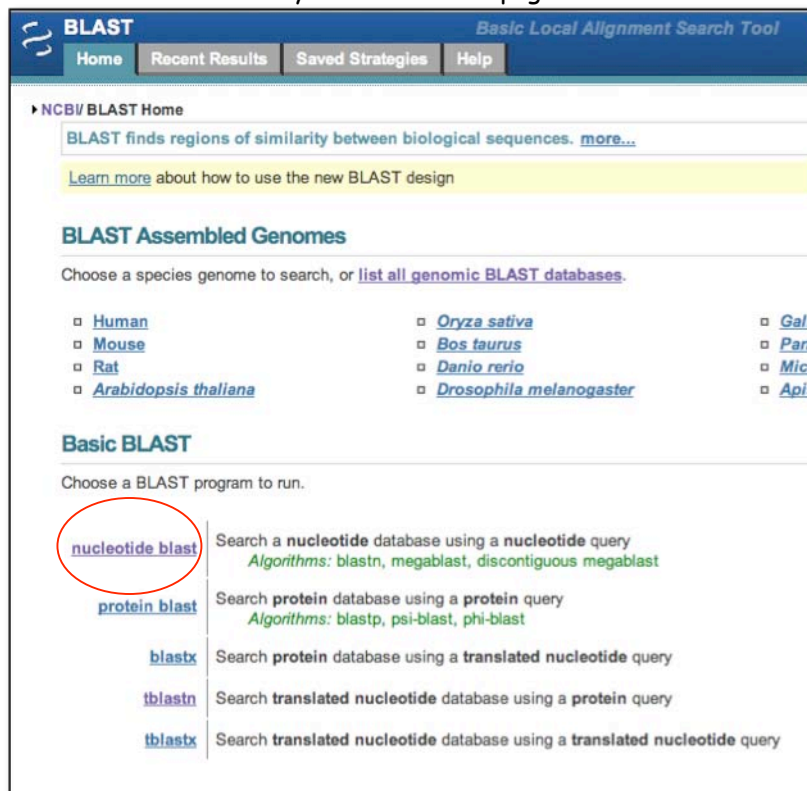
III. Introduction to **Blast**:

You will use Blast a lot in the workshop. It is the major biological sequence search tool for DNA, RNA, and protein databases. Whole genomes can be searched using Blast.

Access Blast by clicking on the Blast link on the NCBI home page.



The Blast link will take you to the Blast page and to the Basic Blast Menu.



There are six different versions of BLAST because you can use a nucleotide sequence or protein sequence to query nucleotide or protein sequence databases, This is summarized in

the screenshot above. Today, we will give three of these search tools a test drive: nucleotide blast, protein blast, and tblastn.

- A. **Nucleotide Blast:** This is the most straightforward type of search. You begin with a nucleotide sequence you want to know more about (the query) and "blast" it against a nucleotide database (the subject).

You can learn a lot about your query sequence with a blast including:

- Are there publications that already report information about this sequence (have you been "scooped")?
- Where is the sequence located in the genome (more on location in class)?
- Is the sequence found in genomes of closely related organisms?
- Does it code for a RNA and/or a protein? If so is anything known about its function?

- Select 'nucleotide blast.' Copy and paste the following sequence in the Query text window (Enter accession number....):

> Osmar5

```
tccatccc ccctccctcc acagcccgat tcccattcc caaacctaac cgtagggcac
ggcggcggcg gcagcgacgg cggcggcggg ggtggcgggg cggcggtgg cggcgccggc
ggaggccgat ggagctgtca tattggtagg cggccgagcg gcagctagga agatgtcgcc
```

Enter Query Sequence

Query subrange

Enter accession number, gi, or FASTA sequence

[Clear](#)

```
ggccagtcac aatgggggtt tcaactggtg gtcatgcaca ttaataggg gtaagactga
ataaaaaatg attatttga tgaatgggg atgagagaga aggaaagagt tcatcctgg
tgaaactcgt cagcgtcgtt tccaagtct cgtaacaga gtgaaacccc cgttgaggcc
gattcgttc attaccgga tctcttgcg ccgcctccg cgtgcgacct ccgattctc
ccgcgccg cggatttg ggtacaaatg atccagcaa cttgatcaa ttaaagtgtt
```

From
 To

Or, upload file

no file selected

Job Title
Enter a descriptive title for your BLAST search

2. Under "Choose Search Set" select "Others" and the drop down list changes to "Nucleotide Collection (nr/nt)." This is the complete non-redundant nucleotide database.

Choose Search Set

Database	<input type="radio"/> Human genomic + transcript <input type="radio"/> Mouse genomic + transcript <input checked="" type="radio"/> Others (nr etc.):
	<div style="border: 1px solid #ccc; padding: 2px; background-color: #fff;"> Nucleotide collection (nr/nt) </div>
Organism <small>Optional</small>	<input type="text" value="Enter organism name or id--completions will be suggested"/>
	<input type="text" value="Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown."/>
Entrez Query <small>Optional</small>	<input type="text" value="Enter an Entrez query to limit search"/>

3. The next section gives you three options for a nucleotide blast. Choose megablast (default) for now.

Program Selection

Optimize for	<input checked="" type="radio"/> Highly similar sequences (megablast)
	<input type="radio"/> More dissimilar sequences (discontiguous megablast)
	<input type="radio"/> Somewhat similar sequences (blastn)
	<input type="button" value="Choose a BLAST algorithm"/>

4. Select the "Blast" button. What you see below is called the queue page:

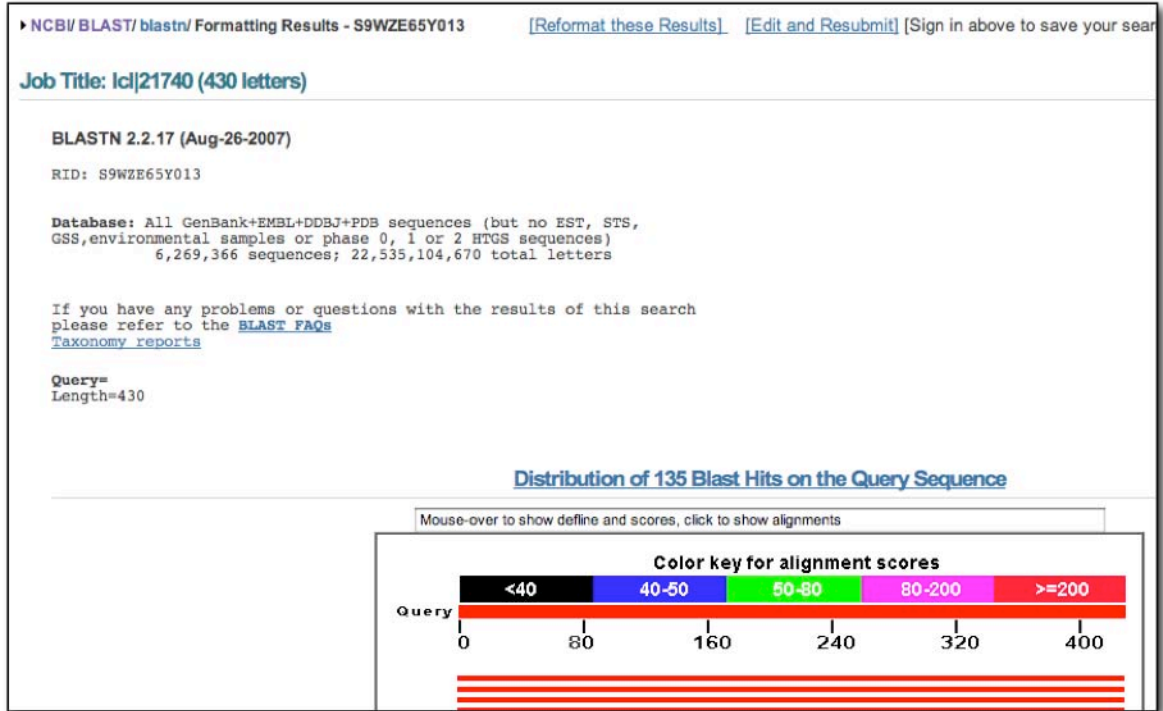
▶ [NCBI/ BLAST/ blastn/ Formatting Results - S9WZE65Y013](#) [\[Formatting options\]](#)

Job Title: lcl|21740 (430 letters)

Request ID	S9WZE65Y013
Status	Searching
Submitted at	Wed Jan 9 11:18:54 2008
Current time	Wed Jan 9 11:18:56 2008
Time since submission	00:00:01

This page will be automatically updated in **10** seconds

5. When your search is complete a results page will be presented. We will discuss this page in detail in class.



6. Details of the Alignment (to be discussed in class)

>tpd|BR000250.1| TPA_exp: Oryza sativa (japonica cultivar-group) ORR1 gene for response regulator, complete cds
Length=10000

GENE ID: 4332111 Os03g0224200 | Os03g0224200 [Oryza sativa Japonica Group]
(10 or fewer PubMed links)

Score = 795 bits (430), Expect = 0.0
Identities = 430/430 (100%), Gaps = 0/430 (0%)
Strand=Plus/Plus

```

Query 1      GGCCAGTCACAATGGGGTTTCACTGGTGTGTCATGCACATTTAATAGGGGTAAGACTGA 60
             |||
Sbjct 3632   GGCCAGTCACAATGGGGTTTCACTGGTGTGTCATGCACATTTAATAGGGGTAAGACTGA 3691

Query 61     ATAAAAAATGATTATTTGCATGAAATGGGGATGAGAGAGAAGGAAAGAGTTTCATCCTGG 120
             |||
Sbjct 3692   ATAAAAAATGATTATTTGCATGAAATGGGGATGAGAGAGAAGGAAAGAGTTTCATCCTGG 3751
Query 121    TGAAACTCGTCAGCGTCGTTTCCAAGTCCTCGGTAACAGAGTGAAACCCCGTTGAGGCC 180
             |||
Sbjct 3752    TGAAACTCGTCAGCGTCGTTTCCAAGTCCTCGGTAACAGAGTGAAACCCCGTTGAGGCC 3811

Query 181    GATTCGTTTCATTACCCGGATCTCTGCGTCCGCCTCCGCCGTGCGACCTCCGCATTCTC 240
             |||
Sbjct 3812    GATTCGTTTCATTACCCGGATCTCTGCGTCCGCCTCCGCCGTGCGACCTCCGCATTCTC 3871

Query 241    CCGCGCCGCGCCGGATTTTGGGTACAAATGATCCAGCAACTTGTATCAATTAATGCTT 300
             |||

```

```

Sbjct  3872  CCGCGCCGCGCCGGATTTTGGGTACAAATGATCCCAGCAACTTGTATCAATTAATGCTT  3931
Query  301    TGCTTAGTCTTGGAACGTCAAAGTGAAACCCCTCCACTGTGGGGATTGTTTCATAAAAG  360
        |||
Sbjct  3932  TGCTTAGTCTTGGAACGTCAAAGTGAAACCCCTCCACTGTGGGGATTGTTTCATAAAAG  3991
Query  361    ATTTCAATTTGAGAGAAGATGGTATAATATTTGGGTAGCCGTGCAATGACACTAGCCATT  420
        |||
Sbjct  3992  ATTTCAATTTGAGAGAAGATGGTATAATATTTGGGTAGCCGTGCAATGACACTAGCCATT  4051
Query  421    GTGACTGGCC  430
        |||
Sbjct  4052  GTGACTGGCC  4061

```

A short discussion on how Blast works.

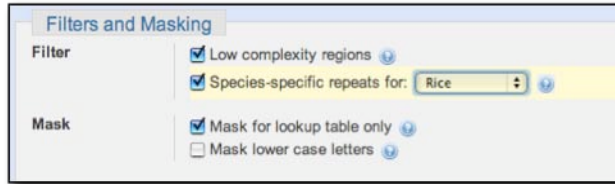
Blast takes the query sequence and divides it into "words" based on the word size parameter (the default is usually "fine"). For a megablast query the default (and minimum) is a size of 28. The algorithm then takes these "words" and runs them against a database. When an exact match occurs, the program attempts to extend the alignment in each direction. If the alignment extends then a score is calculated and as long as the score remains above a threshold the alignment continues. If a mismatch occurs the score decreases, but as long as the score remains above threshold the mismatch is allowed. Word size can be changed. Long word sizes increase stringency.

The threshold is determined by the Expect value in the "Algorithm Parameters" tab on the Blast page. The default Expect value is 10. This means that you expect to find 10 matches to your query in randomly generated sequence. Blast uses this value, the size of the query sequence, and the size of the database (called the search space) to calculate a threshold on 10 random matches and then reports only hits that score better than the random model. Lowering the Expect value increases the stringency of the search.

While extending the alignment, Blast may encounter a series of mismatched nucleotides. Blast will try to skip over the mismatch region (called opening a gap) to see if the alignment begins again. If the alignment begins again, Blast will continue. If the alignment does not begin again, the alignment process stops and Blast reports the hit. Opening a gap is penalized heavily. Extending a gap is also penalized. The process of opening gaps is necessary to allow for small insertion mutations (called indels for insertion deletion) that occur fairly frequently in a genome.

An important point for searches involving Transposable Elements: The ubiquitous low complexity filter.

Repeat the mega blast but with the following modification. Select the "Algorithm Parameters" and go to the Filters and Masking section. Check 'Species-specific repeats for:' and select Rice. Run the Blast. What happened?

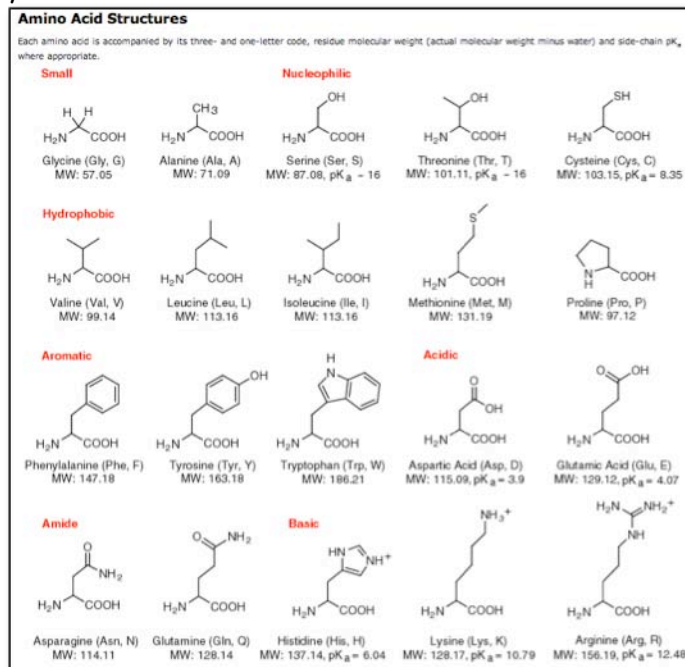


We will discuss low complexity, filtering, and masking.

B. Protein Blast:

A protein blast utilizes an amino acid sequence query from the user as the input and searches a protein database. This is often useful to determine whether the sequence already exists in the database or to predict the function of the predicted protein. The steps for submitting a query are similar to a nucleotide blast and the algorithm is essentially the same.

There is one key difference in the protein vs. nucleotide algorithm. When a nucleotide is compared to a nucleotide only matches between the same bases are allowed ($A \rightarrow A$, $G \rightarrow G$, etc). In contrast, some amino acids have similar chemical properties. For example asparagine (asp) and glutamine (glu) have the same functional group with glutamine having a slightly longer side chain due to an extra methyl group. Asp and glu are often interchangeable without detriment to protein function. The figure below groups the amino acids by functionality.



(www.neb.com)

To score similar amino acid matches, blast uses a look-up table called a BLOSUM matrix. This table contains all possible amino acid matches and a score to use for each. The default matrix is BLOSUM62.

Common groupings of the amino acids (from

<http://www.uky.edu/Classes/BIO/520/BIO520WWW/blosum62.htm>):

G,A,V,L,I, M aliphatic (though some would not include G)
 S,T,C hydroxyl, sulfhydryl, polar
 N,Q amide side chains
 F,W,Y aromatic
 H,K,R basic
 D,E acidic

1. Open a protein blast from the blast home page

(<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>), and choose protein blast.

Copy-and-paste this sequence into the window. If you also copy the top line preceding the amino acid sequence the search will be given a job title.

```
>Osmar5
SKDLTNIQRRGIYQLLLQKSKDGKLEKHTTRLVAQEFHVSIRTVQRIWKRKICHEQGI AVNVDSRKHGNSGR
KKVEIDL SVIAAIPLHQRRNIRSLA QALGVPKSTLHRWFKEGLIRRHNSLKPYLKEANKKERLQWCVSMLDPH
TLPNPKFIEMENIIHIDEKWFNASKKEKTFYLYPDEEPEYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGK
```

2. Run the Blast with all default parameters. The queue screen may report that it found a similarity between your query sequence and the Protein Family (PFam) database. No family will be detected for this search

Job Title: Osmar5	
No putative conserved domains have been detected	
Request ID	6JF07A6U012
Status	Searching
Submitted at	Mon Jun 30 16:03:19 2008
Current time	Mon Jun 30 16:03:34 2008
Time since submission	00:00:14
This page will be automatically updated in 9 seconds	

3. The results page is similar in organization to the nucleotide blast results page. Here is the first alignment reported. Note in this alignment that when two amino acids are identical at a position, the single letter is used in the consensus line. A '+' is used to indicate similar amino acids at a given position.

```
>emb|CAH66447.1| OSIGBa0145N07.3 [Oryza sativa (indica cultivar-group)]
Length=519
```

Score = 714 bits (1843), Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 336/382 (87%), Positives = 358/382 (93%), Gaps = 0/382 (0%)

Query	1	SKDLTNIQRRGIYQLLLQKSKDGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGIA	60
		SKDL N++RR IY LL+KS +GKLEK TT +VA+EFHVSIRTVQRIWKRAK+C EQGIA	
Sbjct	95	SKDLKNMERRAIYARLLEKSMNGKLEKDTTSIVAREFHVSIRTVQRIWKRAKVCREQGIA	154
Query	61	VNVDSRKHGNSGRKKVEIDLVSIAAIPHLQRRNIRSLAQALGVPKSTLHRWFKEGLIRRH	120
		VNVDSRKHG+SGRKKVE+DLS+IAAIPL Q+ NIRSLAQALGVPKSTLHRWFKEGLIRRH	
Sbjct	155	VNVDSRKHGSSGRKKVEVDLSLIAAIPLQOKSNIRSLAQALGVPKSTLHRWFKEGLIRRH	214
Query	121	SNSLKPYLKEANKKERLQWCVSMMLDPHTLPNNPKFIEMENIIHIDEKWFNASKKEKTFYL	180
		SNSLKPYLKEANKKERL+WCVSMMLDP TLPN PKFIEMENIIHIDEKWFN SKKEKTFYL	
Sbjct	215	SNSLKPYLKEANKKERLRWCVSMMLDPSTLPNRPKFIEMENIIHIDEKWFNGSKKEKTFYL	274
Query	181	YPDEEOPYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGKIGIWPFFTRKEPARRRSRN	240
		YPDEEOPYFT HNKNAIDKV FL+AVAKPR+DDEGNCTFDGKI IWPFF RKEPA+RRRSRN	
Sbjct	275	YPDEEOPYFTAHNKNAIDKVTFLA+AVAKPRFDDEGNCTFDGKICIWPFVVRKEPAQRRSRN	334
Query	241	RERGLVTKPIKVDRDTIRSFMISKVLP AIRACWPREDARKTIWIQQDNARTHLPIDDAQ	300
		RERGLVTKPIKVDR+TIRSFMISKVLP AIRACWPREDA KTIWIQQDNARTHLPI+D Q	
Sbjct	335	RERGLVTKPIKVDRNTIRSFMISKVLP AIRACWPREDAGKTIWIQQDNARTHLPINDEQ	394
Query	301	FGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRISRNMDIELIENVHKEYRD	360
		F VAVAQ+GLDIRLVNQPPNSPDMNCLDLGFFASLQSLT+NR SRNMDE+IENVHKEYRD	
Sbjct	395	FAVAVAQTGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTYNRTSRNMDEVIENVHKEYRD	454
Query	361	YNPNTLNRVFLTLQSCYIEVMR	382
		YNP TLNRVFLTLQ C+IE M+	
Sbjct	455	YNPTTLNRVFLTLQCCHIEAMK	476

C. tblastn:

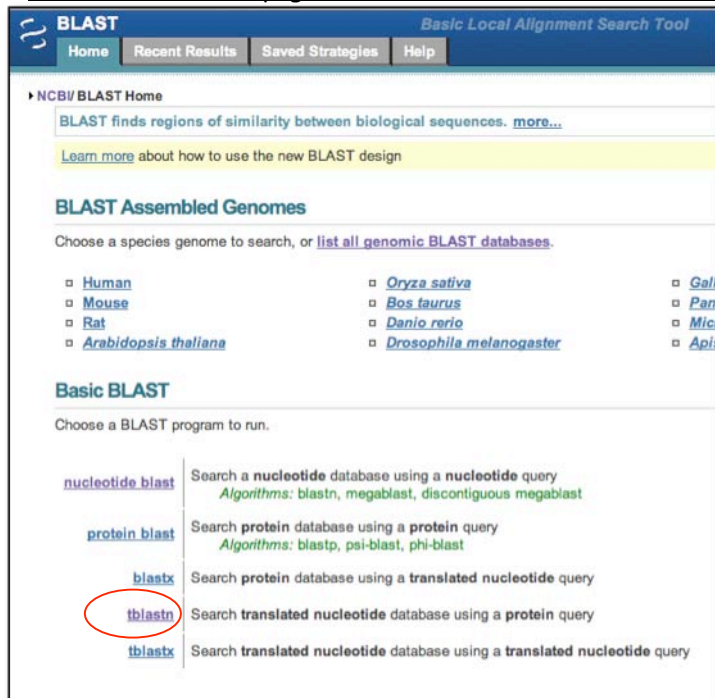
This type of blast takes a protein query sequence and blasts it against a nucleotide database. This is incredibly useful because:

1. it can find the location of the protein in a genome
2. it can find similar sequences in the genome
3. it can find similar sequences in related genomes

To search a nucleotide database with a protein query, the database must first be translated. NCBI stores the nucleotide databases translated in 6 frames.

Why 6 frames?

1. Start at the Blast page and click on *tblastn*, the fourth choice down.



2. Enter the query sequence. Remember, this process compares a sequence of amino acids against sequences in existing genomes.

>Osmar5

```
SKDLTNIQRRGIYQLLLQKSKDGGKLEKHTTRLVAQEFHVSIRTVQRIWKRKICHEQGI AVNVDSRKHGNSGR
KKVEIDL SVIAA IPLHQRRNIRSLA QALGVPKSTLHRWFKEGLIRRHNSNLKPYLKEANKKERLQWCVSMLDPH
TLPNNPKFIEMENIIHIDEKWFNASKKEKTFYLYPDEEPEYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGK
```

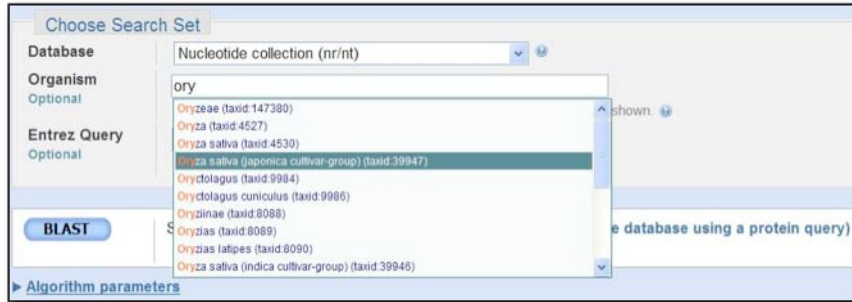
3. Now go down to the section called "Choose Search Set."

4. For the first panel under "Choose Search Set," leave it on the default setting, which is "Nucleotide collection (nr/nt)" nr: non-redundant, nt: nucleotide.

We're going to compare our query sequence to rice, specifically *Oryza sativa*, which is the taxonomic name for rice, one of two varieties of domesticated rice, the other being *Oryza glaberrima*, or African rice. There are two ways to enter "rice" on this panel. To find rice, you only have to type "Ory" and all the rice varieties pop up. You can also type 'rice.'

O. sativa is the third one down. Click on it!

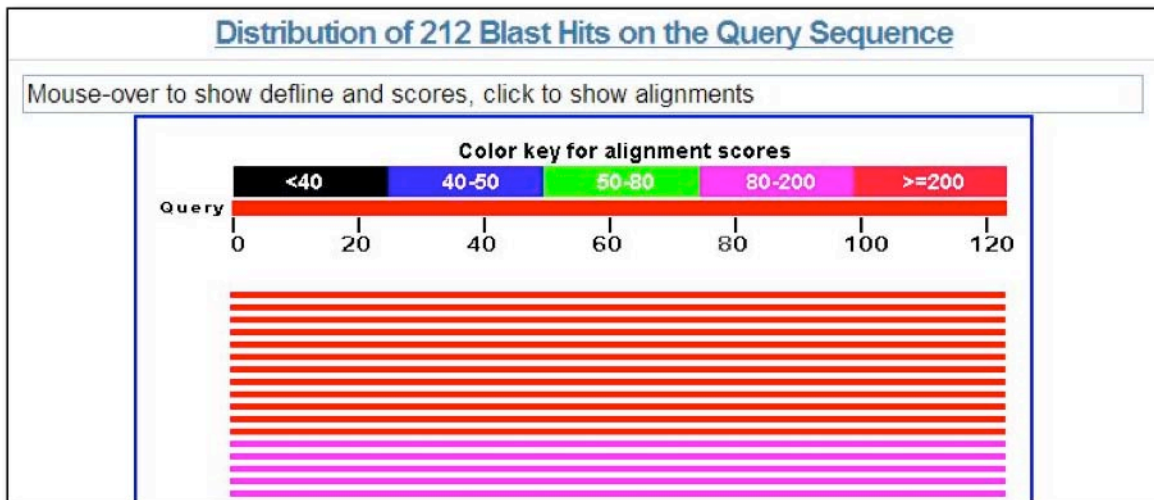
Now you see on your line this: *Oryza sativa* (taxid 4530).



5. Go the bottom and click on BLAST! The Algorithm parameters are similar to the nucleotide blast and protein blast search. They serve the same functions here.



6. **Results:** Will be discussed in class, but by now you should be able to read this page yourself.



Background

The Discovery of Transposable Elements in Maize by Barbara McClintock



It all began more than 60 years ago with a far-sighted scientist named Barbara McClintock who was studying the kernels of what we informally call "Indian corn." You know what it looks like—those ears with richly colored kernels that we associate with Thanksgiving and that we call maize.

Maize and corn are the same species. Maize is a grass that is taxonomically related to other familiar cereal grasses like barley, rice, wheat and sorghum. By the 1920s, researchers had found that maize kernels were ideal for genetic analysis because heritable traits such as kernel color and shape are so easy to visualize. The results of early studies on maize led to an understanding of chromosome behavior during meiosis and mitosis. As a result, by McClintock's time, maize was one of two model genetic organisms - the other being *Drosophila melanogaster* (the fruit fly).

As early as the 1920's it was known that maize had 10 chromosomes [this is the haploid number (n) - maize, is a diploid ($2n$) with 2 sets of 10 chromosomes]. In addition to being a superb geneticist, McClintock was one of the best cytologists in the world and her specialty was looking at whole chromosomes. Maize was ideal for this analysis because it has a large genome [~ 2500 Mb (million bases), about the same size as the human genome] and its chromosomes were easily visualized using a light microscope. The first thing of note that McClintock did as a scientist was to distinguish each of the 10 maize chromosomes of maize. This was the first time anyone

was able to demonstrate that the chromosomes (of any organism) were distinct and recognizable as individuals.

In the course of her studies of various maize strains, she noticed the phenotype shown below in **Figure 1a**. This phenotype is characteristic of chromosome breakage. While chromosome breakage is commonly observed in maize, it had not previously been observed at a single site (locus) in one chromosome. In one particular strain chromosome 9 broke frequently and at one specific place or *locus*. After considerable study, she found that the breakage was caused by the presence in the genome of two genetic factors. One she called *Ds* (for *Dissociation* -it caused the chromosome to "dissociate"), and it was located at the site of the break. But another genetic factor was needed to activate the breakage. McClintock called this one *Ac* (for *Activator*). Because she could not genetically map the position of *Ac* in the genome she hypothesized that it was capable of moving around (transposing). For example, *Ac* could move from chromosome 1 to chromosome 3.

As she followed the descendents of this strain, she identified rare kernels with fascinating phenotypes. One such phenotype was a colorless kernel containing pigmented spots. This is summarized in **Figure 1b**.

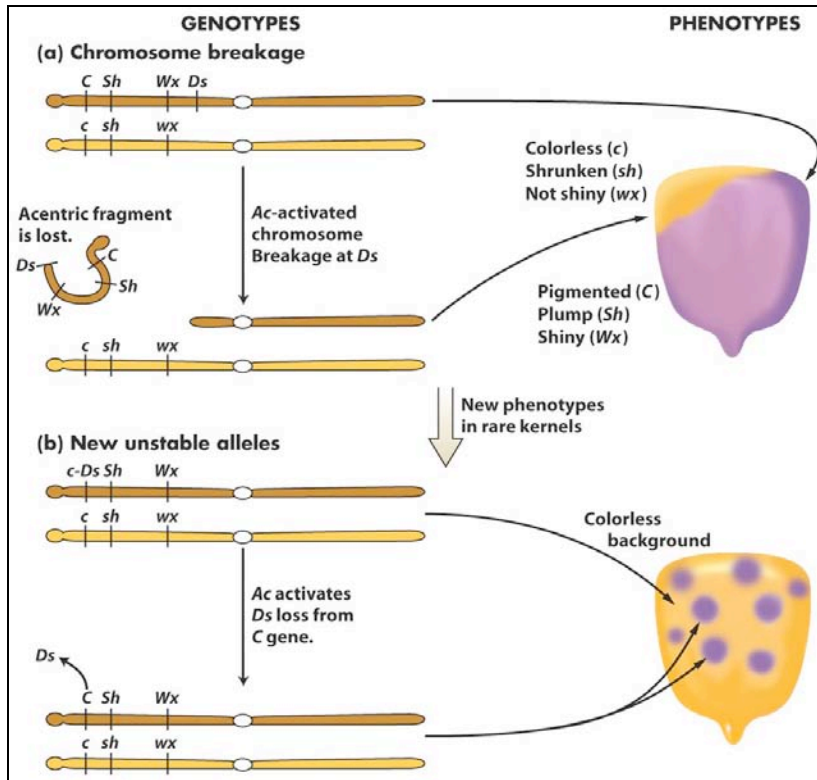
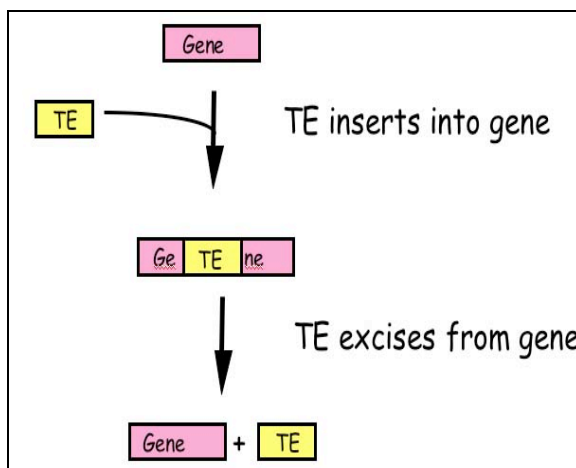


Figure 1. New phenotypes in corn are produced through the movement of the *Ds* transposable element on chromosome 9. (a) A chromosome fragment is lost through breakage at the *Ds* locus. Recessive alleles on the homologous chromosome are expressed, producing the colorless sector in the kernel. (b) Insertion of *Ds* in the *C* gene (top) creates colorless corn kernel cells. Excision of *Ds* from the *C* gene through the action of *Ac* in cells and their mitotic descendants allows color to be expressed again, producing the spotted phenotype.

What she soon knew conclusively was this: *The TEs that she was studying were inserting into the normal genes of maize and were causing mutations. What she had discovered was a different type of mutation - one that was caused by a transposable element and one that was reversible. This contrasts with other mutations that you are probably more familiar with like base pair changes and deletions that are essentially irreversible. Her logic is summarized in the figure below. Furthermore, she provided the following explanation for what was going on with the spotted kernels:*



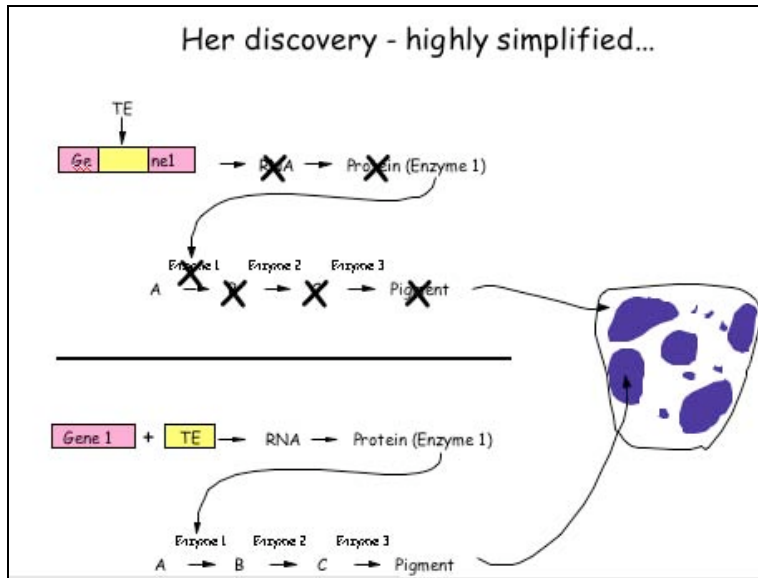


Figure 2: McClintock hypothesized that TEs were a source of “reversible” mutation. Their ability to transpose allowed them to excise from mutant genes leading to phenotypic

What DNA transposable elements look like to the geneticist (Ac, Ds)

As you have seen Barbara McClintock discovered the TEs Ac and Ds when she figured out that they were responsible for the spotted kernel phenotypes. She was a geneticist - and their main experimental tool is the genetic cross.

Here are some of the properties of Ac/Ds that McClintock figured out through observation of kernel phenotypes and by performing carefully designed crosses:

- (1) Ac and Ds could insert into a variety of genes - e.g. those involved in pigment production, starch biosynthesis, and early embryo development, to name but a few.
- (2) Ac and Ds were normal residents of the corn genome - they were not, for example, introduced into the genome by a virus.
- (3) Ds could not move without Ac in the genome, whereas Ac could move itself or Ds. Thus, Ac was called an autonomous element while Ds was called a non-autonomous element.

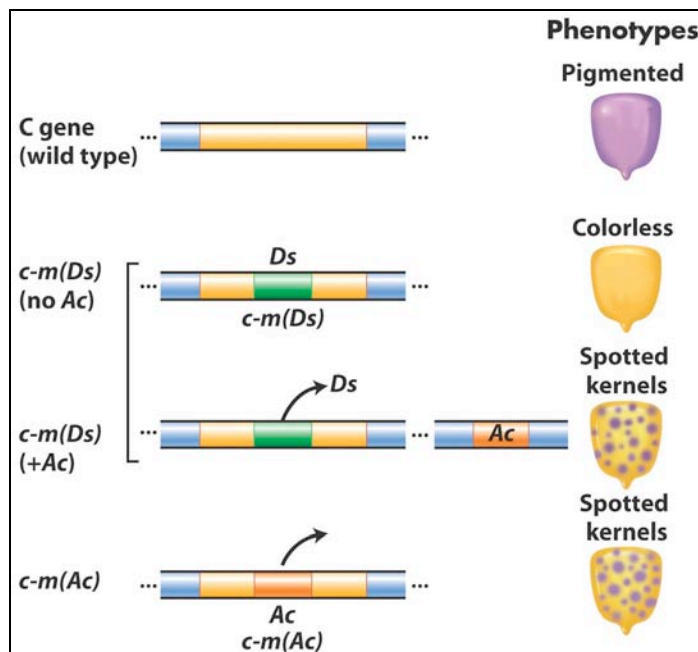


Figure 3 Summary of the main effects of transposable elements in corn. *Ac* and *Ds* are used as examples, acting on the *C* gene controlling pigment. In maize (but not many other organisms), normal alleles are capitalized and mutant alleles are written in lower case. In addition, McClintock designated alleles caused by the insertion of a TE as "mutable", m for short [e.g. c-m(Ds) or c-m(Ac)].

TEs are in all organisms: After her initial results were reported in the late 1940's, the scientific community thought that TEs were oddities and possibly restricted to maize and perhaps to a few other domesticated plant species. However, this proved not to be the case as in subsequent years TEs were discovered in the genomes of virtually all organisms from bacteria to plants

to human. It is for this reason that McClintock was awarded the Nobel Prize in Medicine or Physiology in 1983, almost 40 years after her discovery.

What transposable elements look like to the molecular biologist (Ac, Ds):

With the advent of molecular cloning biologists were able to isolate and sequence gene-sized fragments of DNA from the genomes of plants and animals. They say that a picture is worth a thousand words. So... here is a simplified figure showing what Ac and Ds look like at the DNA level.

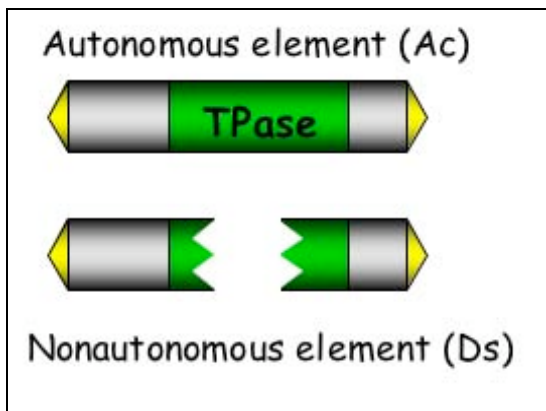


Figure 4: Molecular structure of Ac and Ds.

Ac: T_pase is the gene encoding the transposase enzyme, which is necessary for movement of both Ac and Ds.

Ds: Ds requires Ac for movement because it is a defective version of Ac where the T_pase gene has been deleted.

Yellow arrows at the ends are the terminal inverted repeats - this site where transposase binds and cuts the element out of the surrounding genomic DNA.

Ac contains a single gene - that encodes the transposase protein. Figure 5 shows how this protein catalyzes the movement of Ac and Ds.

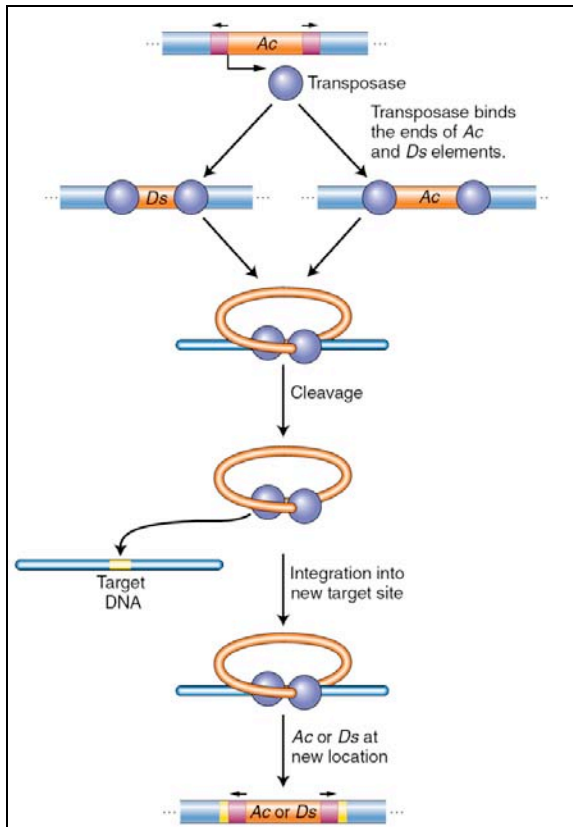


Figure 5 Ac transposase catalyzes excision and insertion (also called integration). The maize Ac element encodes a transposase that binds its own ends or those of a Ds element, excising the element, cleaving the target site, and allowing the element to insert elsewhere in the genome.

Like many other proteins, the transposase protein can multi-task. First, it is a DNA binding protein that is able to bind specifically to the ends of the Ac element. The protein also binds to the ends of Ds as it is identical to the Ac ends. Such "sequence-specific binding" is mediated by precise contacts between the amino acids of part of the transposase (called the binding domain) and the precise nucleotide sequences at the Ac (and Ds) ends. Second, it is an enzyme. Once bound, the two transposase molecules form a dimer (via protein-protein interactions) and another region of the transposase (called the catalytic domain) cuts the element out of the surrounding genomic DNA. The two transposase proteins bound to the TE then cuts the chromosome at another site (the target) in the host genome and the TE inserts.

Finally, for now at least, there is one other feature of TEs that needs to be introduced. This is the target site duplication (TSD) that is created during insertion of virtually all TEs. How it is generated is shown below in Fig 6.

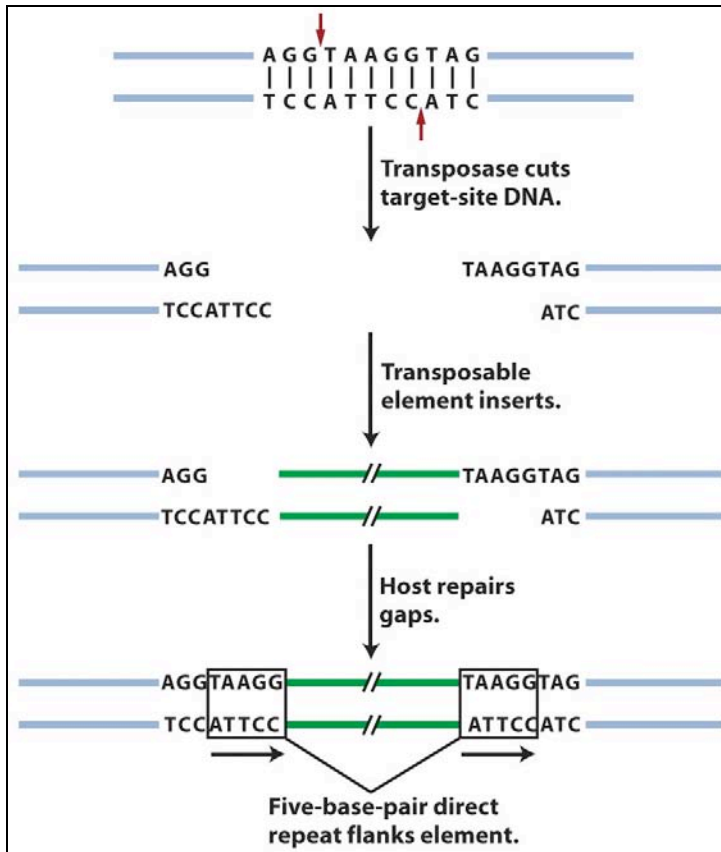


Figure 6: An inserted element is flanked by a short repeat. A short sequence of DNA is duplicated at the transposon insertion site. The recipient DNA is cleaved at staggered sites (a 5-bp staggered cut is shown), leading to the production of two copies of the five-base-pair sequence flanking the inserted element. This is called a target site duplication (TSD).

What transposable elements look like to the bioinformaticist

As you know, Human Genome Project ushered in the genomics era which is characterized by the availability of increasing amounts of genomic sequence from a variety of plant and animal species (animals - including human, drosophila, the worm, dog, mouse, rat, chimp; plants - including *Arabidopsis thaliana*, rice, cottonwood (a tree)]. For now, it is sufficient to know that TEs make up the vast majority of the DNA sequence databases and recognizing TEs in genomic sequence is usually the first step in the modern analysis of TEs.

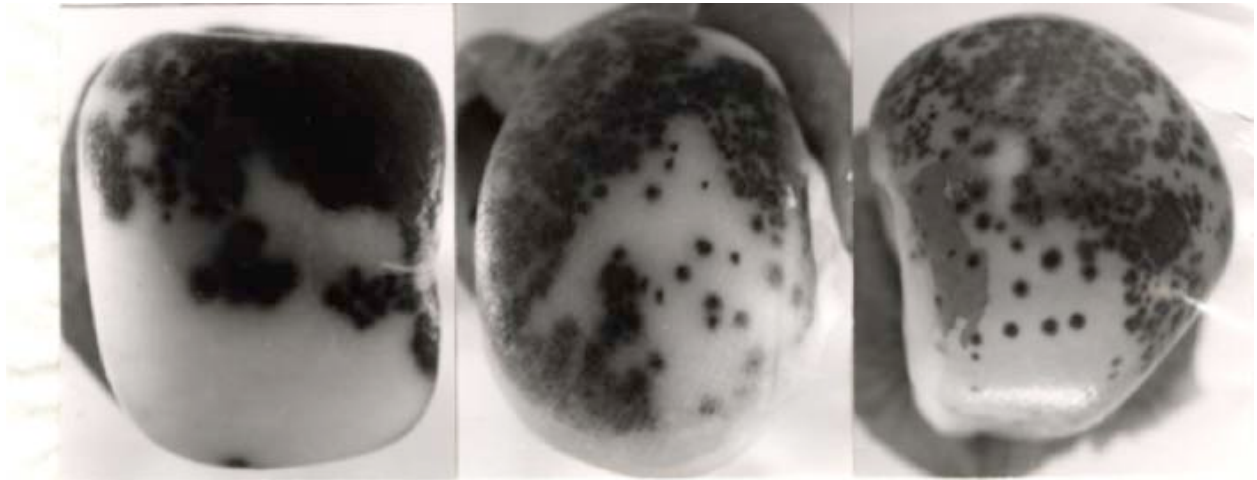
In the introduction to bioinformatics lesson, you will learn about how computers are used to annotate (give meaning to) genomic DNA and how to identify TEs. For many scientists, an experiment ends with the

identification of TEs in genomic sequence. They will write up their results and submit them to a journal for publication. However, for other scientists (like those in the Wessler lab), this is only the beginning. Identifying a TE in genomic DNA does not mean that this TE is still active (capable of moving around). In order to see if a TE sequence is active we have to do an experiment - which is precisely what you will be doing this week. However, you need to be introduced to a few things before we let you in the lab (!).

INTRODUCTION TO THE COMPONENTS OF THE EXPERIMENT:

Overview: In this experiment you will be using the same materials used by the Wessler lab to determine whether a TE sequence that was identified by computational analysis from the rice genome can function (that is, can it move from one place in the genome to another). This experiment will not be done in rice but in the model plant, the humble mustard weed, *Arabidopsis thaliana*.

A visual assay for the movement of rice TEs in *Arabidopsis thaliana*



The Wessler lab needed to create an experimental system that mimics the one used by McClintock with TEs inserted into pigment genes and expressed in the kernel. What was needed was a visual assay to test for movement of a rice TE in *Arabidopsis*.

Creating a visual phenotype: You know what a reporter is—someone who goes out, gathers facts, brings back information, and turns it into ordered and accessible information. Just so, scientists use so-called reporter genes to attach to another gene of interest in cell culture, animals, or plants. Certain genes are chosen as reporters because the characteristics they confer on organisms expressing them are easily identified and measured. Most reporter genes are enzymes that make a fluorescent or colored product or are fluorescent products themselves. Among the latter kind is one that is central to your work this semester, called Green Fluorescent Protein or GFP.

GFP comes from the jellyfish *Aequorea victoria* and fluoresces green when exposed to certain wavelengths of light. Researchers have found GFP extremely useful for an important reason: visualizing the presence of the gene doesn't require sacrificing the tissue to be studied. That is, GFP can be visualized in living organisms by using fluorescent-imaging microscopy.

In our experiments, the GFP reporter gene will substitute for the maize pigment gene. A rice TE (which is described below) has been engineered into the GFP gene so that it cannot produce fluorescent protein. If the TE excises from the GFP gene it will be able function again.

This is all done in *Arabidopsis* - so let's begin with some background on the botanical "lab rat".

Arabidopsis thaliana



In your previous classes you have probably discussed model organisms and their desirable features. Model organisms include *E. coli*, yeast (*Saccharomyces cerevisiae*), *Drosophila melanogaster*, *Caenorhabditis elegans* (a.k.a. the worm), mouse (*Mus musculus*), and *Arabidopsis thaliana*. Like the other model organisms, *A. thaliana* is easily transformed by foreign DNA and is small and has a relatively short generation time (~6 weeks). This small flowering plant is a genus in the family *Brassicaceae*. It is related to cabbage and mustard. *A. thaliana* is one of the model organisms used for studying plant biology and the first plant to have its entire genome sequenced (~125 Mb, about the same as *Drosophila*. The human genome is almost 20X bigger than this at ~2500 Mb).

***Agrobacterium tumefaciens*: introducing foreign DNA into plants (how scientists construct transgenic plants)**



A crown gall tumor. Infection by the bacterium *Agrobacterium tumefaciens* leads to the production of galls by many of plant species.

In 1977, two groups independently reported that crown gall is due to the transfer of a piece of DNA from *Agrobacterium* into plant cells. This resulted in the development of methods to alter *Agrobacterium* into an efficient delivery system for gene engineering in plants. In short, *Agrobacterium* contains a plasmid (the Ti-plasmid), which contains a fragment of DNA (called T-DNA). Proteins encoded by the Ti-plasmid facilitate the transfer of the bacterial T-DNA into plant cells and ultimately, insertion of the T-DNA into plant chromosomes. As such, the Ti-plasmid and its T-DNA is an ideal vehicle for genetic engineering. This is done by cloning a desired gene sequence into the T-DNA that will be inserted into the host (plant) DNA.

As shown in Figure 7, foreign DNA is inserted in the lab into the T-DNA (shown as the green DNA in the "cointegrate Ti plasmid below), which is then transformed into *Agrobacterium* which is then used to infect cultured tobacco cells. The Ti plasmid moved from the bacterial cell to the plant nucleus where it integrated into a plant chromosome. Tobacco cells can be easily grown into "transgenic" plants where all cells contained the engineered T-DNA.

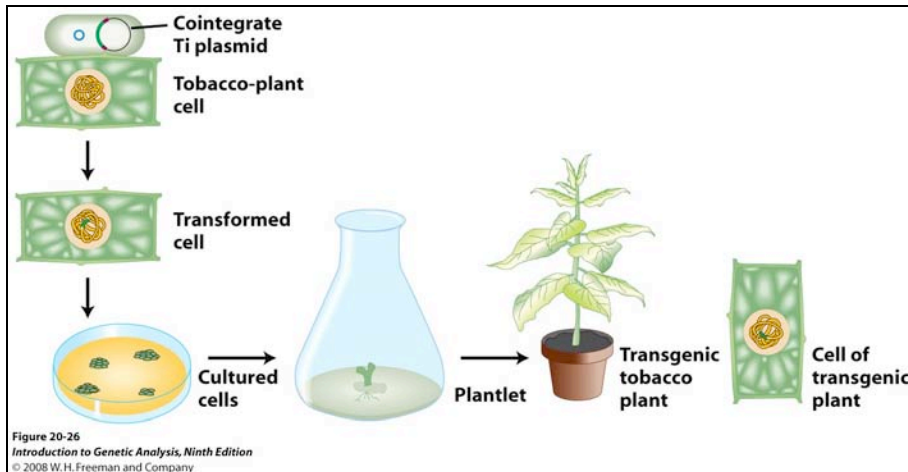


Figure 7: Schematic of how *Agrobacterium* has been exploited to deliver foreign DNA into plant chromosomes.

The foreign DNA inserted into the T-DNA included both a gene of interest and a "selectable" marker, in this case, an antibiotic resistance gene. This is necessary because the procedure for transferring a foreign DNA into a plant via *Agrobacterium*-mediated transformation is very inefficient. By using media/agar containing the antibiotic, only the cultured cells with the T-DNA in their chromosomes will be resistant to the antibiotic and able to grow.

Rice and the discovery of TE sequences in its genome

Rice (*Oryza sativa*) has the smallest genome of all cereal grasses at 450 million base pairs (Mb). By contrast, the maize (and human) genome is almost six times larger at 2500 Mb. About 40 percent of the rice genome comprises repetitive DNA and most of this is derived from TEs. Because the full genome sequence for rice is known, members of the Wessler lab were able to use a computational approach to identify TEs that were potentially active based solely on their sequence characteristics. Here is a figure showing the element that you will be analyzing.....

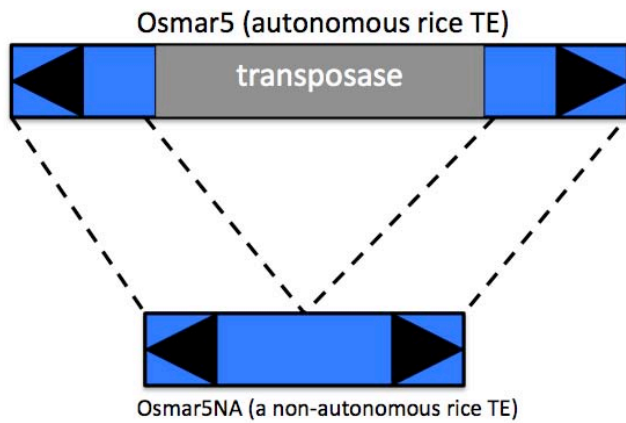


Figure 8: The Wessler lab identified Osmar5 using methods that you will learn in class. To do this experiment, they had to create a nonautonomous version of this element by deleting the middle part of the element. By analogy, the autonomous rice element is like Ac while the nonautonomous element (called Osmar5NA) is like Ds (see figures 3-5)

Design of the experiment and controls

Sequences identified in the rice genomes as TEs are called "candidates" because there is no evidence that these sequences are actually capable of moving around. Members of the Wessler lab (Dr. Guojun Yang to be exact) constructed transgenic Arabidopsis plants containing T-DNAs that looked like the following:

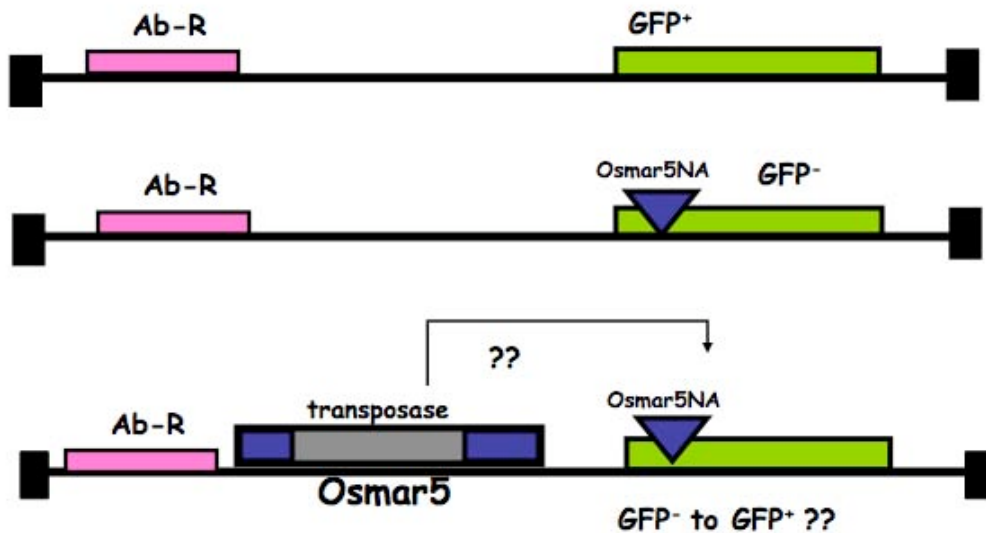


Figure 9: The transgenic Arabidopsis plants used in this experiment contain one of these 3 T-DNA insertions in their genome. Ab-R = the antibiotic resistance gene discussed above.

(TOP) Plants containing this T-DNA in their genome are the positive controls. These plants should be green under UV light because the GFP protein is produced (designated GFP⁺).

(MIDDLE) Plants containing this T-DNA in their genome are the negative control. These plants should be red under UV light because there is no GFP protein (designated GFP⁻) and the red color is due to chlorophyll fluorescence.

(BOTTOM) Plants containing this T-DNA in their genome are the actual experiment. If our hypothesis is correct, then the transposase encoded by *Osmar5* will produce a transposase that will bind to the ends of the non-autonomous TE (*Osmar5NA*) and catalyze its transposition out of the GFP gene restoring gene function.

(Not shown) Finally, plants designated as wild type (WT) do not have ANY T-DNA in their genome.

However, we are not out of the woods yet. We need solid experimental evidence that *Osmar5NA* has actually excised in plants with T-DNA (C) but NOT in plants with T-DNA (B). The experiment is designed to determine just that - whether or not rice TE NA has excised from the GFP gene and restored gene function in Arabidopsis.

PCR analysis of the T-DNA in *Arabidopsis thaliana* DNA

You will be using the polymerase chain reaction (PCR) procedure in this experiment. It is summarized below:

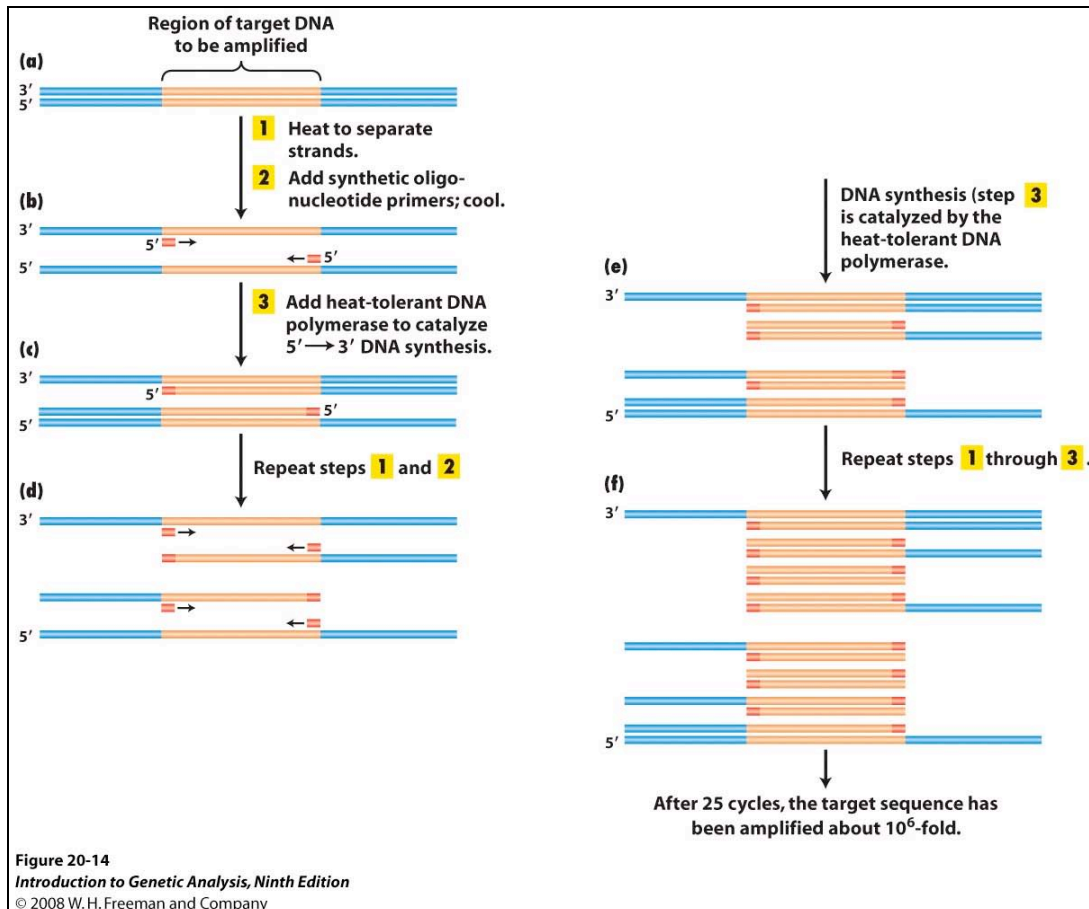


Figure 10:
Details of
PCR. See
below.

Polymerase Chain Reaction. Known familiarly as PCR: a technique enabling multiple copies to be made of sections of DNA molecules. It allows isolation and amplification of such sections from large heterogeneous mixtures of DNA such as whole chromosomes and has many diagnostic applications, for example in detecting genetic mutations and viral infections. The technique has revolutionized many areas of molecular biology—and won a Nobel Prize for Kary Mullis.

The reaction starts with a double-stranded DNA fragment. A part of it is to be copied.

A to B. The two DNA strands are separated (denatured) by heating to 95°Celsius (C).

B. After cooling, short oligonucleotide primers (see below) that are complementary to the ends of the region to be amplified anneal with each strand.

C. When the temperature is raised to 72° C the DNA polymerase (the heat-stable *Taq* polymerase) begins to catalyze DNA synthesis from the ends of the primer using the denatured DNA as template (the extension phase) and the nucleotide triphosphates that are in the test tube.

D, E, and F - The procedure is repeated beginning with denaturation then cooling, annealing, extension etc.

Oligonucleotide primer. A primer is a short nucleic acid strand or a related molecule that serves as a starting point for DNA replication. A primer is required because most DNA polymerases, enzymes that catalyze the replication of DNA, cannot copy one strand into another from scratch, but can only add to an existing strand of nucleotides. (In most natural DNA replication, the ultimate primer for DNA synthesis is a short strand of RNA. This RNA is produced by primase, and is later removed and replaced with DNA by a DNA polymerase.) The primers used for PCR are usually short, chemically synthesized DNA molecules with a length of about 20-30 nucleotides.

Denaturation: separation of the two DNA strands of a double helix by heating them to a very high temperature. This breaks the hydrogen bonds holding the double helix together.

Annealing: when DNA or RNA strands pair by hydrogen bonds to complementary strands, forming a double-stranded molecule. The term is also used to describe the reformation (renaturation) of complementary strands that were separated by heat.

Extension: enzymatically extending the primer sequence—copying DNA.

Details about your PCR....

The purpose of this experiment is two-fold:

- (1) to determine whether the *Osmar5NA* transposable element is able to excise from the *GFP* reporter gene with or without the transposase encoded by the *Osmar5* element.
- (2) to determine the DNA sequence at the putative site of *Osmar5NA* excision to see if excision is precise or imprecise and whether excision events are the same or different.

This figure shows where the PCR primers that you will use are located. After PCR, products will be separated on an agarose gel. Gel bands will be purified (isolated free of the agarose) and cloned into bacterial plasmids (see below).

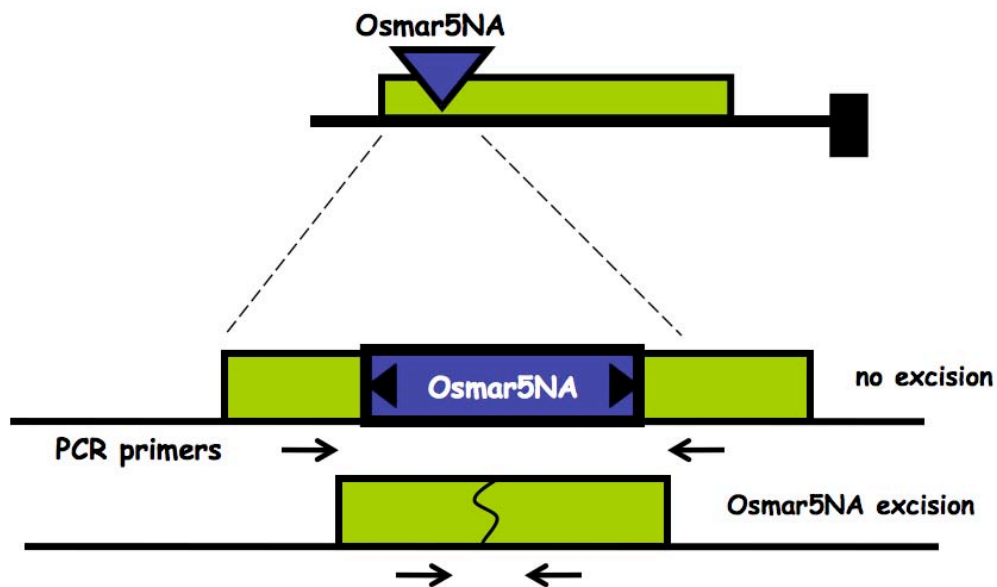


Figure 11: The details of the DNA that will be amplified in your PCRs. See Figure 9 for the T-DNAs that this came from. Green boxes represent the *GFP* gene.

Observing the Plants:

Before we do any molecular analysis we look at the seedlings with a microscope. We will use very strong blue light to excite the *GFP* to visualize the results of *Osmar5NA* activity. If present, the *GFP* will fluoresce green. We will take photographs of the leaves to document the results. You will need to annotate and add these to your notebook after class. Can you predict what the *GFP* pattern will look like for each type of seedling?

Extracting the DNA from the seedlings.

To perform molecular analysis, we will first isolate DNA.

Materials list:

Extraction Buffer (for 50 ml)

10% SDS

5M KOAC

100% Isopropanol

70% Ethanol

Ice Bucket with ice

Liquid nitrogen

65°C Heating block

sterile 1.5 mL tubes (2 for each prep)

sterile sticks for grinding

Miracloth cut into 10 cm squares.

- 1) Harvest 2-3 seedlings.
- 2) Grind tissue to a fine powder in a 1.5 mL tube dipped in liquid nitrogen.
Put a little liquid Nitrogen in a mortar and dip the end of the tube in it.
Grind the frozen tissue with a sterile stick.
- 3) Add 1000 μ l of Extraction Buffer, and grind some more in the buffer.
- 4) Add 120 μ l of 10% SDS. Mix by inverting.

If preparing multiple samples, do all to this step. Keep samples on ice until all are done.

- 5) Put at 65°C for 20 minutes.
- 6) Add 300 μ l 5M KOAc. Mix well by inverting several times (important!), then place on ice 5 minutes.

7) Spin for 5 minutes at top speed in microfuge. Squirt about 700 ml of the supernatant through miracloth (make small funnel, place tip directly onto the miracloth at the tip of the funnel and squirt through - do not allow the whole funnel to get soaked).

8) Add 600 μ l of isopropanol. Mix the contents thoroughly by inverting.

DNA precipitate may or may not be visible at this point; don't worry if you don't see much. A really good prep (excellent grinding of tissue) should result in visible DNA at this stage, however. Can put in the freezer for a while at this point, or proceed immediately to the next step

9) Spin for 5 minutes at top speed. Pour off supernatant.

10) Add 500 μ l of 70% ethanol and flick until the pellet comes off the bottom (for best washing results). Spin briefly, then pour off the ethanol. Suck off the rest of the ethanol with a pipet. Let air dry in hood for around 30 minutes.

11) Re-suspend the DNA in 100 μ l water or TE. Let sit at RT for awhile, mix by pipetting. Depending on amount of starting material may need to be diluted for PCR.

PCR reaction:Materials:

PCR Tubes

Reagents listed in table

Thermocycler

Ice bucket with ice.

1. Label 0.2 ml PCR Tubes:

A. WT

B. GFP

C. OSNA

D. OSNA/TP

E. Water

2. Label a 1.5 ml tube for the Master mix.3. Assemble the Master Mix using the table. Keep on ice.

[Stock]	[Final]	Volume to add (μ l)	Master Mix for 6 reactions (μ l)
Water	-----	4.2	25.2
2X Taq Mix	1X	10.0	60.0
10 μ M Primer A	0.2 μ M	0.4	2.4
10 μ M Primer B	0.2 μ M	0.4	2.4
		5.0	-----
Final Volume		20.0	90.0

4. Place 15 μ l of Master Mix in each PCR tube.5. Put the appropriate DNA or water into each tube.6. Keep on ice until placed in the PCR thermocycler.

Analyzing and purifying your PCR products: Gel Electrophoresis.

Overview - Today you will separate the products from your PCR amplification by agarose gel electrophoresis and then cut out and extract the DNA from one of the gel bands.

Prepare 1% agarose gel. Wear Gloves!

1. Prepare a gel rig with the proper comb.
2. Add 0.5 g agarose to a 500 ml flask. Add 50 ml 1x TAE buffer, swirl, and heat contents in the microwave until boiling (~2-3 min). Be careful!
3. Swirl to make sure it is completely melted and allow it to cool. Add ~0.5 μ l of a stock solution of 10mg/ml ethidium bromide (EtBr) (EtBr binds to DNA allowing it to be visualized under UV light. Don't let it touch your skin).
4. Swirl again to mix and pour into the gel rig. The gel should cool and solidify within 15-20 minutes at which time it is ready. Turn the gel 90 degrees and add enough TAE buffer to completely immerse.

Preparing gel samples, loading onto gel, electrophoresis:

1. Beginning with your PCR tubes from Tuesday, add 4 ul of 6X loading dye to each (change tip for each sample).
2. Load 10 ul of this mixture into the wells in the order: W, WT, "OS", "OS/TP", markers (to be provided).
3. Attach the leads and set the power supply to run at 150V. Remember "Run to the red."
4. Run for 20-30 minutes, then turn off the power and remove the gel tray.
5. Take a picture of the gel for your records. We will discuss the gel photos in class in detail.

Excise gel fragment

1. Weigh one 1.5 ml microcentrifuge tube and record.
2. Carefully slide gel off tray onto sheet of Saran wrap covering transilluminator
3. Put on protective face shield, turn on transilluminator.
4. Cut the gel slice (as small as possible) containing the band with a clean razor blade and put the slice into the tube that was weighed.

Gel Extraction

(Using the Qiagene Gel Extraction Kit, see detailed instructions at the end of this section):

1. Weigh the gel slice in the tube. Add 3 volumes of Buffer QG to 1 volume of gel (100mg ~ 100 μ l). For example, add 300 μ l of Buffer QG to each 100 mg of gel.
2. Incubate at 50°C for 10 min in a water bath (or until the gel slice has completely dissolved). To help dissolve gel, mix by inverting the tube every 2-3 min during the incubation.
3. After the gel slice has dissolved completely, add 1 gel volume of isopropanol to the sample and mix. For example, if the agarose gel slice is 100 mg, add 100 μ l isopropanol. This step increases the yield of DNA fragments < 500 bp and > 4kb.
4. To bind DNA to the column material, apply the sample to the QIAquick column and then spin at 13,000 rpm for 1 minute. The DNA is now in a high salt/non-polar solution. Under these conditions the DNA sticks to silica (the stuff in the column). The maximum volume of the column reservoir is 800 μ l. For sample volumes of more than 800 μ l, simply load again.
5. Discard flow-through and place QIAquick column back in the same collection tube.

6. Add 0.5 ml of buffer QG to QIAquick column and centrifuge for 1 min. Discard the flow through. This step is only required for direct sequencing (which is exactly what will be done with your samples).

7. To wash any impurities (EtBr and agarose), add 0.75 ml of Buffer PE to QIAquick column, let the column stand 3min and spin column at 13,000 rpm for 1min.

8. Discard the flow through and centrifuge for another 1 min at 13,000 rpm.

IMPORTANT: This spin is necessary to remove residual ethanol (which is present in Buffer PE).

9. Place QIAquick column in a clean 1.5 ml microcentrifuge tube.

10. To elute DNA from the column, add 30ul water to the center of QIAquick membrane, leave column on bench for 2 min, and centrifuge the column for 1 min at 13,000 rpm.

IMPORTANT: Ensure that the elution buffer is dispensed directly onto the QIAquick membrane for complete elution of bound DNA. The tube containing the eluted DNA will then be sent to the sequencing facility.

Product Information Sheet for QIAquick Gel Extraction Kit.
(<http://www1.qiagen.com/literature/handbooks/literature.aspx?id=1000252>)

The QIAquick Principle

The QIAquick system combines the convenience of spin-column technology with the selective binding properties of a uniquely designed silica membrane. Special buffers provided with each kit are optimized for efficient recovery of DNA and removal of contaminants in each specific application. DNA adsorbs to the silica membrane in the presence of high concentrations of salt while contaminants pass through the column. Impurities are efficiently washed away, and the pure DNA is eluted with Tris buffer or water (see page 17). QIAquick spin columns offer 3 handling options — as an alternative to processing the spin columns in a microcentrifuge, they can now also be used on any commercial vacuum manifold with luer connectors (e.g., QIAvac 6S or QIAvac 24 Plus with QIAvac Luer Adapters) or automated on the QIAcube.

Adsorption to QIAquick membrane — salt and pH dependence

The QIAquick silica membrane is uniquely adapted to purify DNA from both aqueous solutions and agarose gels, and up to 10 μg DNA can bind to each QIAquick column. The binding buffers in QIAquick Spin Kits provide the correct salt concentration and pH for adsorption of DNA to the QIAquick membrane. The adsorption of nucleic acids to silica surfaces occurs only in the presence of a high concentration of chaotropic salts [1], which modify the structure of water [2].

Adsorption of DNA to silica also depends on pH. Adsorption is typically 95% if the pH is ≤ 7.5 , and is reduced drastically at higher pH (Figure 1). If the loading mixture pH is >7.5 , the optimal pH for DNA binding can be obtained by adding a small volume of 3 M sodium acetate, pH 5.0.

pH Dependence of DNA Adsorption to QIAquick Membranes

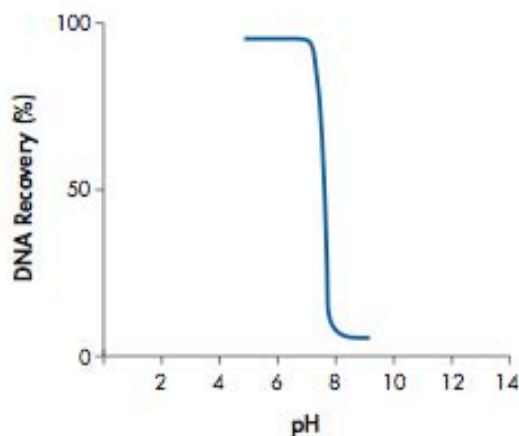


Figure 1 1 μg of a 2.9 kb DNA fragment was adsorbed at different pHs and eluted with Buffer EB (10 mM Tris-Cl, pH 8.5). The graph shows the percentage of DNA recovery, reflecting the relative adsorption efficiency, versus pH of adsorption.

QIAquick Gel Extraction Kit procedure, the color of the binding mixture allows easy visualization of any unsolubilized agarose, ensuring complete solubilization and maximum yields. The indicator dye does not interfere with DNA binding and is completely removed during the cleanup procedure. Buffers PBI and QG do not contain sodium iodide (NaI). Residual NaI may be difficult to remove from DNA samples, and reduces the efficiency of subsequent enzymatic reactions such as blunt-end ligation.

Washing

During the DNA adsorption step, unwanted primers and impurities, such as salts, enzymes, unincorporated nucleotides, agarose, dyes, ethidium bromide, oils, and detergents (e.g., DMSO, Tween® 20) do not bind to the silica membrane but flow through the column. Salts are quantitatively washed away by the ethanol-containing Buffer PE. Any residual Buffer PE, which may interfere with subsequent enzymatic reactions, is removed by an additional centrifugation step.

Elution in low-salt solutions

Elution efficiency is strongly dependent on the salt concentration and pH of the elution buffer. Contrary to adsorption, elution is most efficient under basic conditions and low salt concentrations. DNA is eluted with 50 or 30 µl of the provided Buffer EB (10 mM Tris-Cl, pH 8.5), or water. The maximum elution efficiency is achieved between pH 7.0 and 8.5. When using water to elute, make sure that the pH is within this range. In addition, DNA must be stored at -20°C when eluted with water since DNA may degrade in the absence of a buffering agent. Elution with TE buffer (10 mM Tris-Cl, 1 mM EDTA, pH 8.0) is possible, but not recommended because EDTA may inhibit subsequent enzymatic reactions.

DNA yield and concentration

DNA yield depends on the following three factors: the volume of elution buffer, how the buffer is applied to the column, and the incubation time of the buffer on the column. 100–200 µl of elution buffer completely covers the QIAquick membrane, ensuring maximum yield, even when not applied directly to the center of the membrane. Elution with ≤50 µl requires the buffer to be added directly to the center of the membrane, and if elution is done with the minimum recommended volume of 30 µl, an additional 1 minute incubation is required for optimal yield. DNA will be up to 1.7 times more concentrated if the QIAquick column is incubated for 1 minute with 30 µl of elution buffer, than if it is eluted in 50 µl without incubation (Figure 3, page 14).

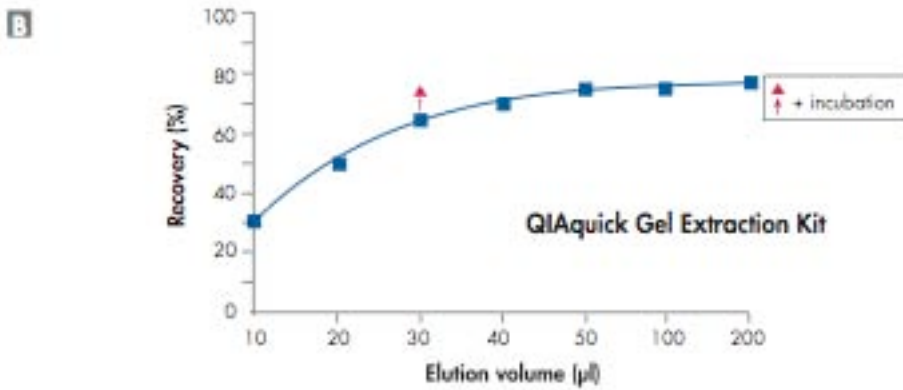


Figure 3 Effect of elution buffer volume on DNA yield for \square the QIAquick PCR Purification and QIAquick Nucleotide Removal Kit; \blacksquare the QIAquick Gel Extraction Kit. 5 µg of a 2.9 kb DNA fragment were purified and eluted with the indicated volumes of Buffer EB. 30 µl plus 1 minute incubation on the QIAquick column gives DNA yields similar to 50 µl without incubation, but at a concentration 1.7 times greater.

Agarose gel analysis of yield

Yields of DNA following cleanup can be determined by agarose gel analysis. Table 3 shows the total yield obtained following extraction of 1 µg or 0.5 µg starting DNA from an agarose gel with a recovery of 80% or 60% using the QIAquick Gel Extraction Kit. The corresponding amount of DNA in a 1 µl aliquot from 50 µl eluate is indicated. Quantities of DNA fragment corresponding to these 1 µl aliquots are shown on the agarose gel in Figure 4.

Table 3. Amount of DNA in 1 µl aliquots of a 50 µl eluate following QIAquick purification

Starting DNA	Recovery	Total yield (50 µl eluate)	Amount of DNA in 1 µl
1 µg	80%	0.8 µg	16 ng
	60%	0.6 µg	12 ng
0.5 µg	80%	0.4 µg	8 ng
	60%	0.3 µg	6 ng

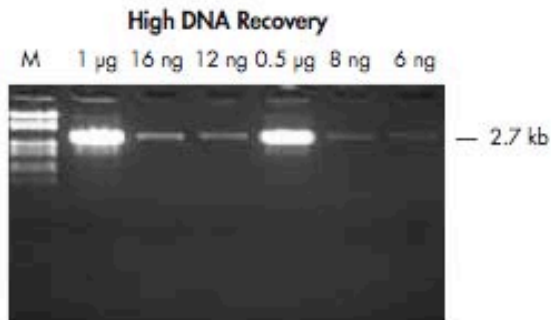


Figure 4 Quantities of purified 2.7 kb DNA fragment corresponding to 1/50 of the DNA obtained following purification from 1 µg or 0.5 µg starting DNA with a recovery of 80% or 60% (see Table 1). Samples were run on a 1% TAE agarose gel. M: lambda-EcoRI-HindIII markers.

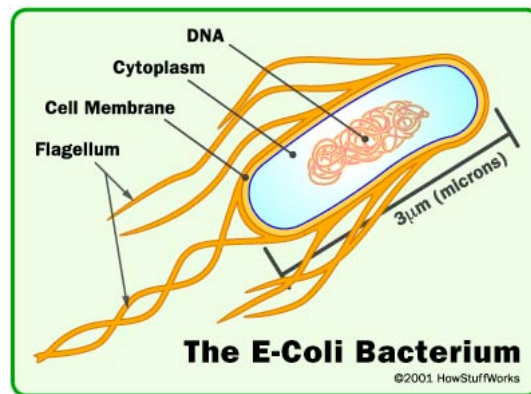
Ligation and transformation of *E. coli*, plate *E. coli*

Two ways to go at this point....

At this point in the experiment we could simply send the purified PCR products to the sequencing facility - PCR products can be directly sequenced in this way. However, what if all of the PCR products in a band were not exactly the same? How could such a thing happen? What if the excision of Osmar5NA from the GFP gene was not always perfect and a few nucleotides were sometimes "left behind" from the element. Alternatively, what if the element excised and took a small piece of the gene with it or "scrambled" some of the sequences at the excision site?

If any or all of these scenarios occurred, our one sequence per band would not be sufficient as your PCR products might be from several different excision events (slight differences like this would not be detected on the agarose gel. Can you figure out why?). So... let's say that the PCR band in fact contains many slightly different sequences. To figure out what those sequences are, we have to somehow analyze individual PCR products. We can do this with the help of *E. coli*. Let's see how this is done.

Escherichia coli - the model bacterium



A quick word about *E. coli*. This bacterium is yet another workhorse for molecular biologists, because it grows rapidly, can be easily transformed, it can be used as a living factory to make lots of plasmid DNA and its genome is completely sequenced.

Using E.coli to construct a "library" of DNA fragments

For this experiment, we will be using a patented system called TOPO Cloning from a company called Invitrogen. We will be ligating our PCR products into the TOPO plasmid (vector), transforming these plasmids into E.coli, picking E.coli colonies, isolating plasmid with our PCR inserts and then sending these to the sequencing facility. Here is some background to make all of this clearer.

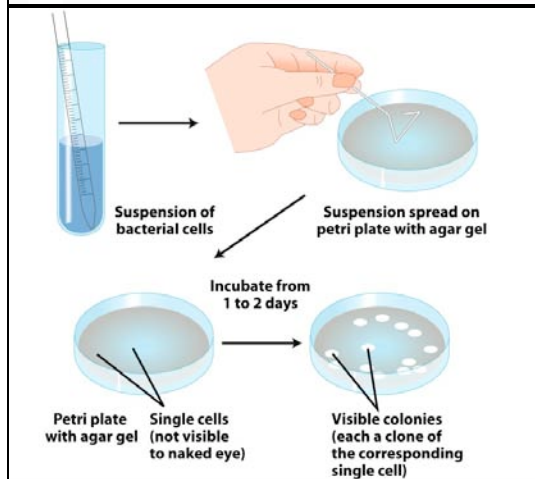
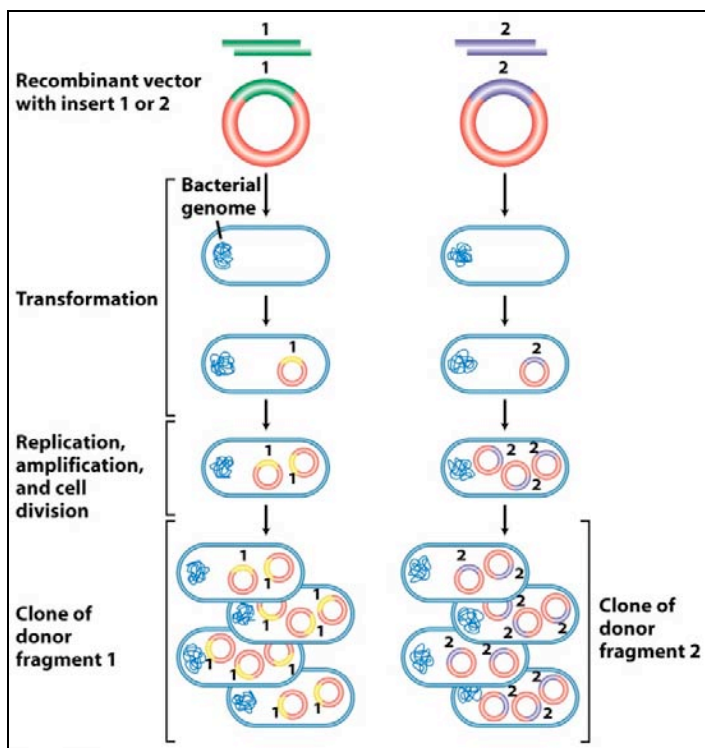
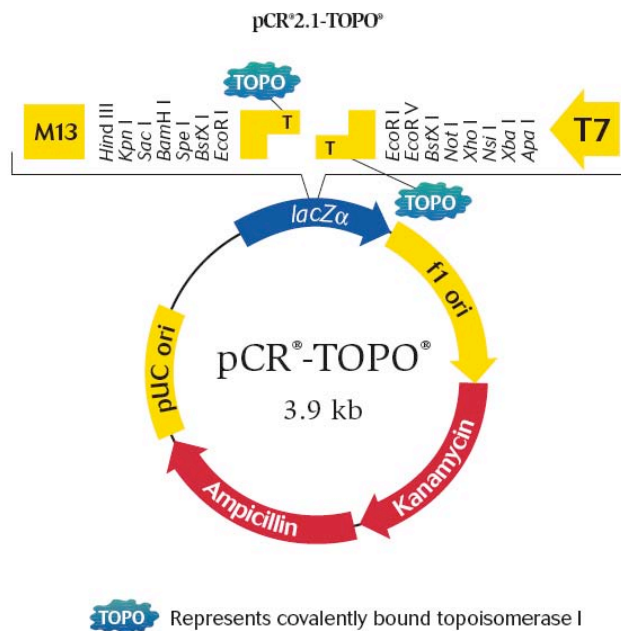


Figure 12: To make a library, single DNA fragments are ligated into a plasmid vector and then transformed into competent *E.coli*. Individual cells with a single plasmid and insert grow into a single colony, which is grown up and used for plasmid DNA isolation.

The TOPO vector:

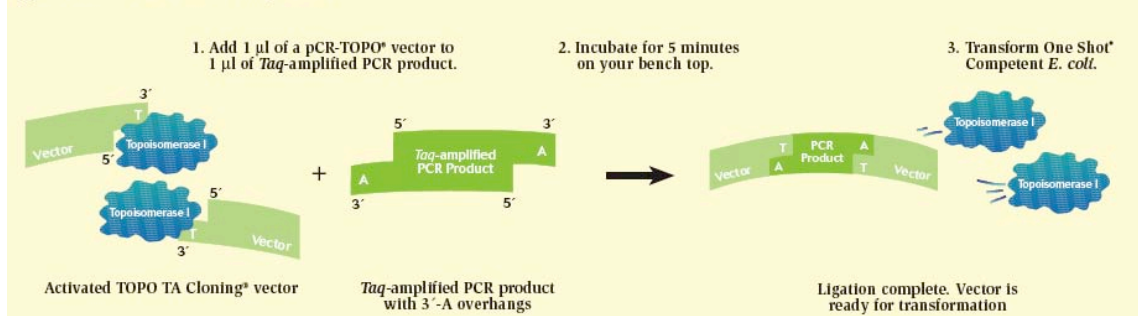
We will be using TOPO to clone the gel bands from the PCR of leaf DNA then we will insert this purified DNA into the TOPO vector, transform *E. coli*, let the transformed bugs grow overnight, purify plasmid from bacterial colonies and send these purified plasmids to the sequencing facility. The idea behind TOPO cloning, according to the company's web site, is "to effectively clone DNA produced by a particular method (in your case, PCR) and to enable specific downstream studies (in your case, DNA sequencing)."



Direct ligation with TA Cloning[®] Technology

The TA Cloning[®] technology makes it possible to easily clone PCR products produced by *Taq* polymerase. *Taq* has a terminal transferase activity that adds a single 3'-A overhang to each end of the PCR product. TOPO TA Cloning[®] vectors contain 3'-T overhangs that enable the direct ligation of *Taq*-amplified PCR products (Figure 6)(2,3).

Figure 6 - How TOPO TA Cloning[®] works



Ligation and transformation of E.coli, plate E.coli

Materials

PCR2.1 TOPO Vector - Invitrogen

Top-10 chemically competent E. coli cells (this will be explained in class)

SOC medium (store at room temperature)

LB/Carb/X-Gal agar plates

Sterile glass beads

Protocol

1. Place tube of Top-10 competent cells and PCR2.1 TOPO vector on ice to thaw

2. Add the following to a 1.5 ml centrifuge tube. Pipet gently and **do not** mix vigorously.

4 μ l gel purified PCR product

1 μ l salt solution

1 μ l PCR2.1 TOPO Vector (add last)

3. Incubate for 10 min at room temperature (on your benchtop)

From this point on you will be working with live E. coli bacteria. All contaminated tips, tubes, and plates must be disposed of properly (waste containers will be provided). Wash hands after handling.

4. Transfer 2 μ l of the incubated mixture to the tube containing Top-10 competent cells (keep on ice). Pipet gently and **do not** mix vigorously.

5. Incubate the tube on ice for 20 min. While waiting, label the LB selective plates and then place plates into 37 degree incubator to warm up. (Only transformed E.coli cells can grow on LB selective plate. This media contains S-Gal a galactose derivative. E. coli that take up the plasmid without an insert will make a enzyme that will convert S-Gal into a black pigment and the colony will turn black. If the plasmid has an insert, the enzyme will be disrupted and the colony will be the natural white color.)

6. Incubate in a water bath for 30 sec at 42°C. (This is called the heat shock - it is when DNA is actually taken up into the bacteria from the surrounding liquid)
7. Immediately place cells on ice for 1 min.
8. Add 250 μ l SOC medium (keep sterile).
9. Incubate in a 37 °C shaker for 60 min
10. Pipet 100 μ l of bacterial solution onto one selective plate (work quickly to keep the plates closed as much as possible). Pour 3-5 sterile glass beads onto the plates, cover and shake horizontally to spread the liquid.
11. Dump the glass beads into the bacterial waste container.
12. Incubate the plates **overnight** in an incubator at 37°C
Plates will be placed at 4°C after 14-20 hours of incubation.

Preparing an Overnight E. coli culture.

Take a picture of each plate for your lab notebook. Pick bacterial colonies that have inserts (the white colonies) into test tubes containing liquid medium and grow them overnight as described below.

Materials:

LB + Carb liquid growth medium (for growing bacteria)
Sterile disposable loops

Protocol:

1. Add 5 ml of liquid growth medium (LB/Carb) into sterile test tubes
2. Using a sterile loop touch a single white colony from the agar plate and swirl the loop in the liquid media. Use good sterile technique.
3. Incubate the test tubes in the spinner overnight at 37°C.

Mini-Prepping the culture to obtain large quantities of the plasmid for sequencing.

Materials:

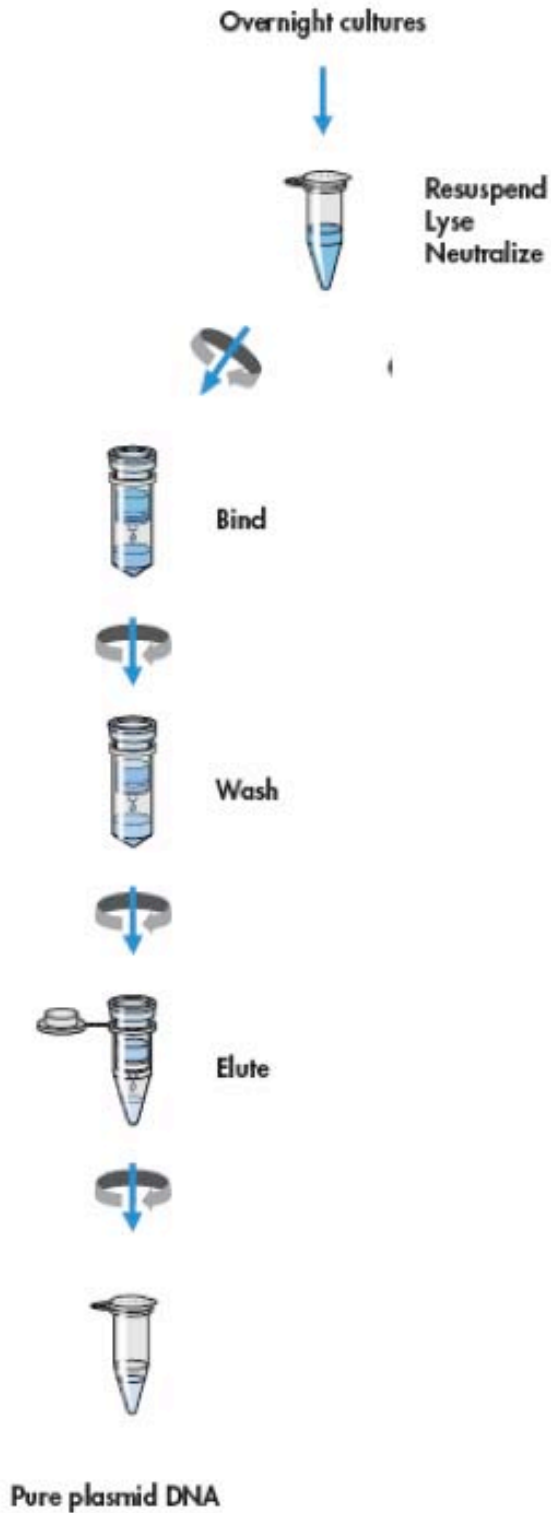
QIAprep Miniprep kit
Agarose gel (+EtBr), TAE buffer, 6X DNA loading buffer

Protocol:

1. Transfer 1.0 ml of your E. coli sample from the overnight culture to a labeled 1.5ml centrifuge tube (put tip in bacterial waste container after use).
2. Cap and centrifuge for 3 min at 8,000 rpm. Decant (dump) supernatant into the bacterial waste.
3. Transfer another 1.0 ml of the E. coli sample to the 1.5 ml tube and centrifuge again for 3 min. You will have a very large pellet of bacterial cells.
4. Add 250ul buffer P1 and vortex to re-suspend the pelleted bacterial cells. No cell clumps should be visible after re-suspension of the pellet. The solution will be a thick tan color.
5. Add 250ul buffer P2 and gently invert the tube 4-6 times to mix. Do not vortex. The solution will clear as the cells lyse. The solution will become mucilaginous.
6. Add 350ul buffer N3 and invert the tube immediately but gently 4-6 times. The solution should become cloudy.
7. Centrifuge for 10 min at 13,000 rpm. A compact white pellet will form.
8. Pipet 800ul of the supernatant (not the white precipitate) and apply to a labeled QIAprep spin column. The plasmid is in solution. The white pellet contains the chromosomal DNA, proteins, and lipids.

9. Centrifuge for 30 sec. Discard the flow-through. The plasmid is on the column.
 10. Add 0.75 ml PE buffer to the spin column and centrifuge for 1 minute. Discard the flow-through.
 11. Centrifuge for an additional 1 min to remove residual buffer.
 12. Transfer the QIAprep column to a clean labeled 1.5 ml centrifuge tube.
 13. Add 100ul EB to the center of each QIAprep spin column, let stand for 1 min, and centrifuge for 1 min. The plasmid is in the liquid flow through.
 14. Discard the column (plasmid DNA will be in the liquid at the bottom of the tube).
- Optional: Time permitting we will perform a restriction digest on the samples. This will be decided at class time and a protocol will be provided.
15. Run 10ul of the purified plasmid (the column flow-through) on an agarose gel to check the quality and quantity.
 16. Take a picture of the gel for your notebook. Based on the gel results, decide which plasmids will be sequenced and fill out the sequence request form.

QIAprep Spin Procedure in microcentrifuges



Preparation of cell lysates (this is a detailed summary of the “chemical logic” of the plasmid miniprep from the manufacturer)

Bacteria are lysed under alkaline conditions. After harvesting and resuspension, the bacterial cells are lysed in NaOH/SDS (**Buffer P2**) in the presence of RNase A. SDS solubilizes the phospholipid and protein components of the cell membrane, leading to lysis and release of the cell contents while the alkaline conditions denature the chromosomal and plasmid DNAs, as well as proteins. The optimized lysis time allows maximum release of plasmid DNA without release of chromosomal DNA, while minimizing the exposure of the plasmid to denaturing conditions. Long exposure to alkaline conditions may cause the plasmid to become irreversibly denatured.

The lysate is neutralized and adjusted to high-salt binding conditions in one step by the addition of **Buffer N3**. The high salt concentration causes denatured proteins, chromosomal DNA, cellular debris, and SDS to precipitate, while the smaller plasmid DNA renatures correctly and stays in solution. It is important that the solution is thoroughly and gently mixed to ensure complete precipitation. *To prevent contamination of plasmid DNA with chromosomal DNA, vigorous stirring and vortexing must be avoided during lysis. Separation of plasmid from chromosomal DNA is based on coprecipitation of the cell wall-bound chromosomal DNA with insoluble complexes containing salt, detergent, and protein. Plasmid DNA remains in the clear supernatant.* Vigorous treatment during the lysis procedure will shear the bacterial chromosome, leaving free chromosomal DNA fragments in the supernatant. Since chromosomal fragments are chemically indistinguishable from plasmid DNA under the conditions used, the two species will not be separated on QIAprep membrane and will elute under the same low-salt conditions. *Mixing during the lysis procedure must therefore be carried out by slow, gentle inversion of the tube.*

Analyzing your DNA sequences

How your DNA samples were sequenced

DNA sequencing is the process of determining the nucleotide order of a given DNA fragment. Most DNA sequencing is currently being performed using the chain termination method developed by Frederick Sanger. [Sanger is particularly notable as the only person to win two Nobel prizes in chemistry - his second in 1980 for developing this DNA sequencing method and his first in 1958 for determining the first amino acid sequence of a protein (insulin)]. His technique involves the synthesis of copies of your input DNA by the enzyme DNA polymerase. However, one difference between this reaction and PCR, for example, is the use of modified nucleotide substrates (in addition to the normal nucleotides), which cause synthesis to stop whenever they are incorporated. Hence the name: "chain termination".

Chain terminator sequencing (Sanger sequencing)

Your samples were sent to the sequencing facility at UGA along with information about the sequencing primer to be used (recall that DNA polymerase needs a primer to start DNA synthesis of a template strand). The reaction contains your DNA sample, the sequencing primer, DNA polymerase and a mixture of the 4 deoxynucleotides that are "spiked" with a small amount of a chain terminating nucleotide (also called dideoxy nucleotides, see below).

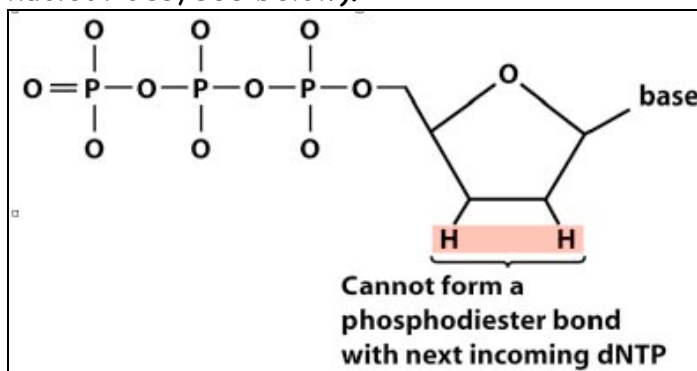
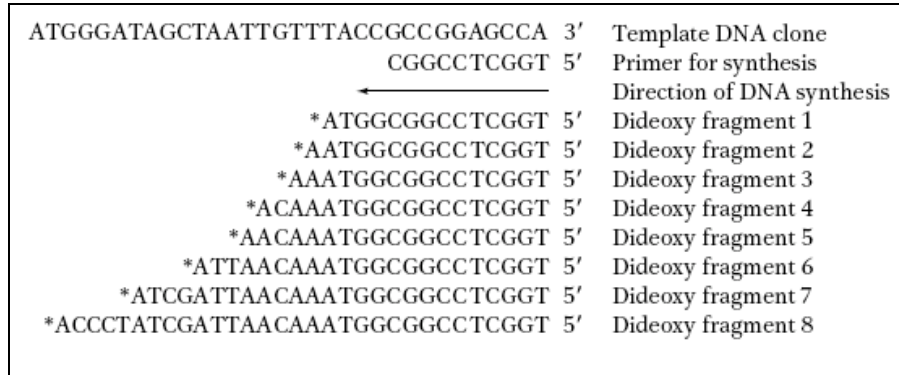


Figure 13. A chain-terminating nucleotide triphosphate (called a di-deoxynucleotide or ddNTP). Because it has a "H" instead of a "OH" at the 3' position, it is not a substrate for the addition of another NTP and DNA synthesis terminates.

Limited incorporation of the chain terminating nucleotide by the DNA polymerase results in a series of related DNA fragments that are terminated only at positions where that particular nucleotide is used.



The fragments are then size-separated by electrophoresis in a slab polyacrylamide gel, or more commonly now, in a narrow glass tube (capillary) filled with a viscous polymer.

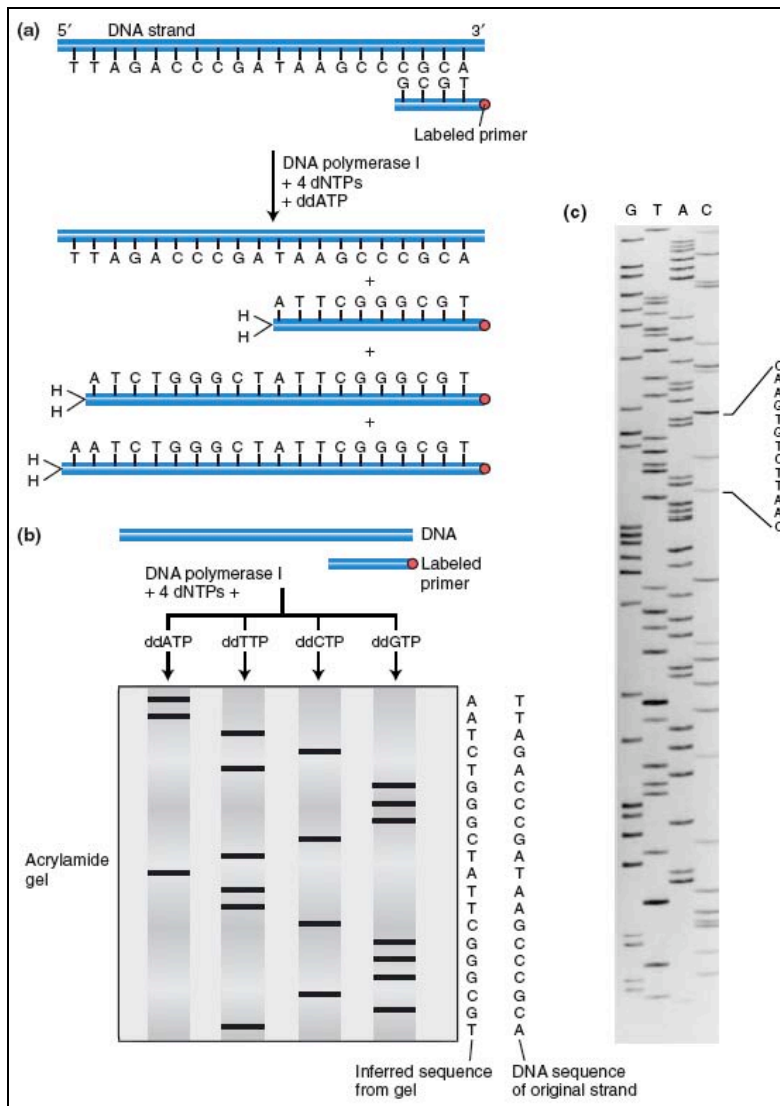


Figure 14. DNA is efficiently sequenced by including dideoxynucleotides among the nucleotides used to copy a DNA segment. (a) In this example, a labeled primer (designed from the flanking vector sequence) is used to initiate DNA synthesis. The addition of four different dideoxynucleotides (ddATP is shown here) randomly arrests synthesis. (b) The resulting fragments are separated electrophoretically and subjected to autoradiography. The inferred sequence is shown at the right. (c) Sanger sequencing gel.

Modifying DNA sequencing to automation: dye terminator sequencing
 (this is how your DNA samples will be sequenced)

An alternative to the labeling of the primer is to label the dideoxy nucleotides instead, commonly called 'dye terminator sequencing'. The major advantage of this approach is the complete sequencing set can be performed in a single reaction, rather than the four needed with the labeled-primer approach. This is accomplished by labeling each of the dideoxynucleotide chain-terminators with a separate fluorescent dye, which fluoresces at a different wavelength.

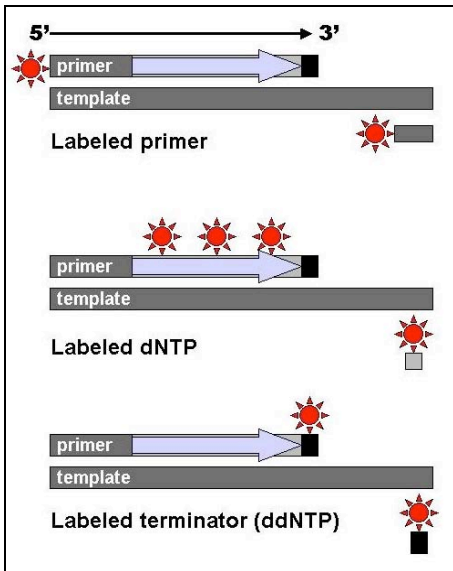


Figure 15: DNA fragments can be labeled by using a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

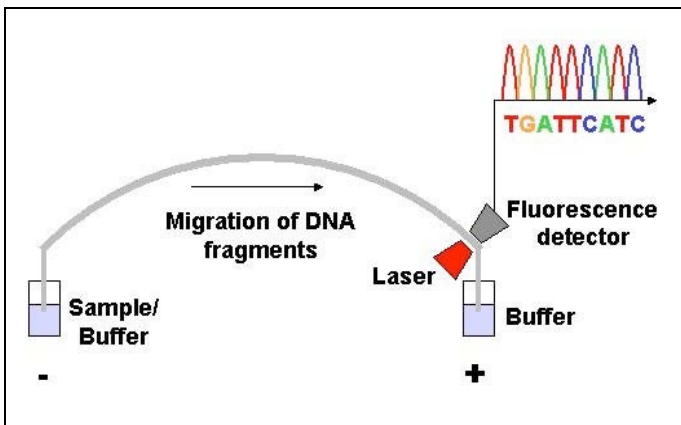


Figure 16. Modern automated DNA sequencing instruments (DNA sequencers) can sequence up to 384 fluorescently labeled samples in a single batch (run) and perform as many as 24 runs a day. However, automated DNA sequencers carry out only DNA size separation by capillary electrophoresis, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms (see Fig. 17, 18).

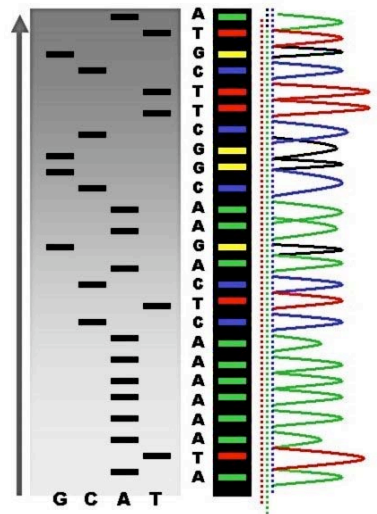


Figure 17: Sequencing by radioactive sequencing compared to fluorescent peaks

This method is now used for the vast majority of sequencing reactions, as it is both simpler and cheaper. The major reason for this is that the primers do not have to be separately labeled (which can be a significant expense for a single-use custom primer), although this is less of a concern with frequently used 'universal' primers.

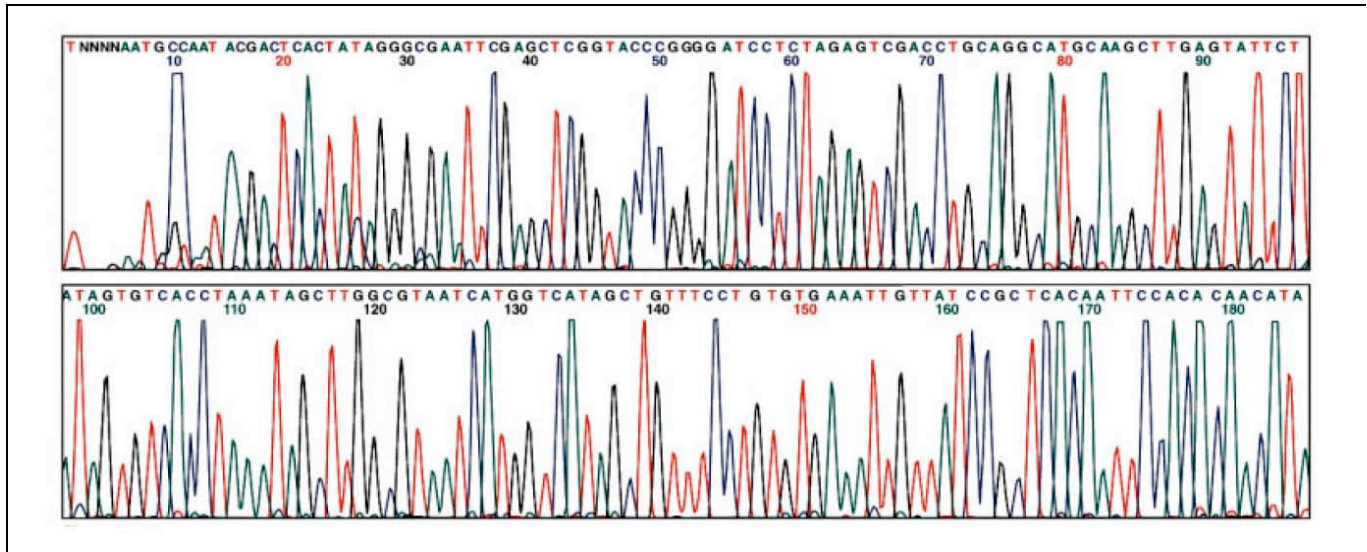


Figure 18. An example of a chromatogram file of a Sanger sequencing read. The four bases are detected using different fluorescent labels. These are detected and represented as 'peaks' of different colors, which can then be interpreted to determine the base sequence, shown at the top.

Information you will need to annotate your sequence

Your sequences have been uploaded to the class data web page and we will explain how to find them. Like most experiments done for the first time, some of your sequences may be not very pretty.

There are 2 files for each sequencing result, one chromatogram file (like Figure 18, above) and one text file. The text file contains the DNA sequence.

This figure should be useful in figuring out what your sequence means...

Information for Analyzing Your Sequences

Before we start, check the following two sequences:

Sequence 1:

5' ...CCTCTCCACTGACAGAAAATTTGTGCCATTAACATCACCATCTAATTCAACA...3'

Sequence 2:

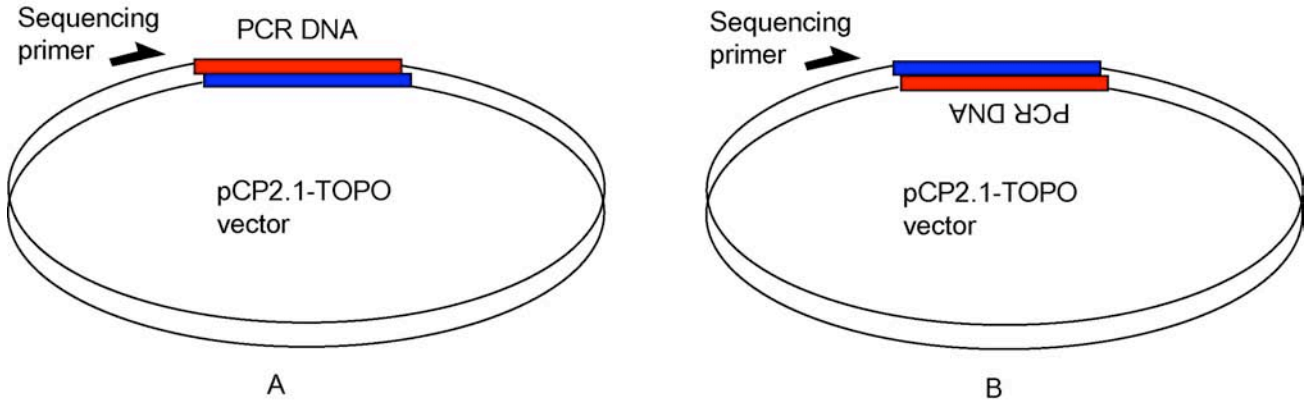
5' ...TGTTGAATTAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGG...3'

What is the difference between them? At first glance, they are very different from each other. However, they are complementary sequences of the same DNA.

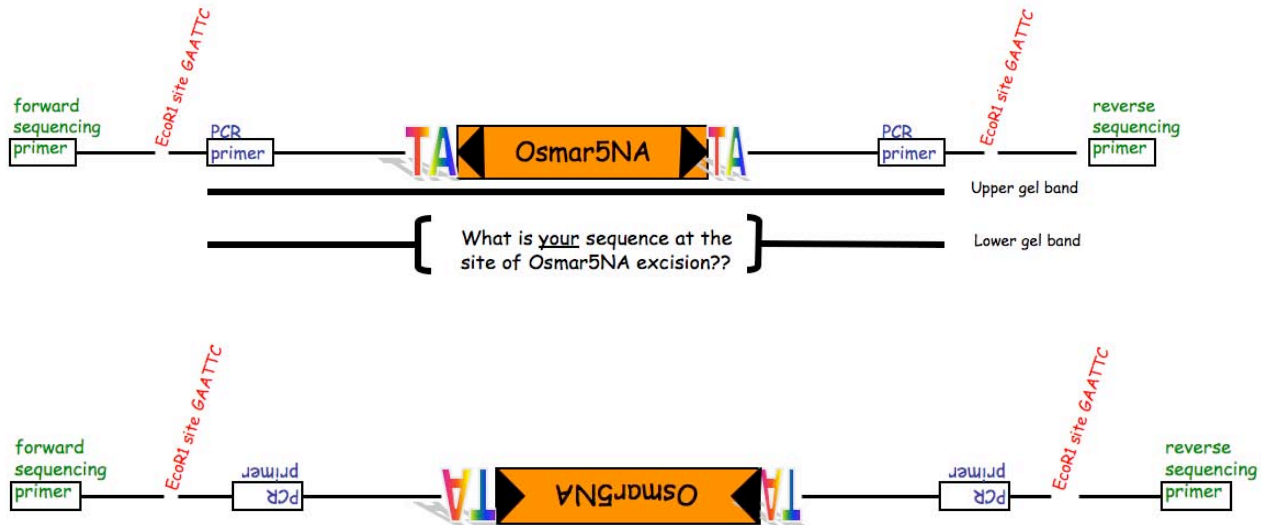
```

5' ...CCTCTCCACTGACAGAAAATTTGTGCCATTAACATCACCATCTAATTCAACA...3'
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
3' ...GGAGAGGTGACTGTC TTTTAAACACGGGTAATTGTAGTGGTAGATTAAGTTCT...5'
    
```

During the cloning of our PCR products into the TOPO vector, the PCR DNA could be ligated into the vector in either of the two orientations shown below:



Thus the sequencing result may be different even when using the same sequencing primer for the same inserted DNA. The orientation must be taken into account when you interpret (annotate) your sequence files.



Here are reference sequences you will need to analyze your sequences: The sequences have been annotated as follows: the EcoRI sites are shown in **red bold**; PCR primers are shown in **blue bold**; TA Target site duplication (TSD) is shown in **black bold and BIG**; Osmar5 NA sequence is shown in **peach bold lowercase**; NOTE: only one direction's sequences are shown.

>Raw DNA sequence (before the insertion of Osmar5 NA):
 ...**CCTCTCCACTGACAGAAAATTTGTGCCCA**TTAACATCACCATCTAATTCACAAGAAT
 TGGGACAACCCAGTGAAAAGTTCTTCTCCTTTACT**GAATTC**GGCCGAGGATAATGATA
 GGAGAAGTGAAAAGATGAGAAAGAGAAAAAGATTAGTCTTCATTGTTATATCTCCTTGG
 ATCC**TA**GGATCCTCTAGAGTCCCCCGTGTCTCTCCAAATGAAATGAACTTCCTTATAT
 AGAGGAAGGGTCTTGCGAAGGATAGTGGGATTGTGCGTCATCCCTTACGTCAGTGGAGA
 TATCACATCAATCCACTTG**CTTTGAAGACGTGGTTGGAACGTCT**...

>With the insertion of Osmar5 NA:
 ...**CCTCTCCACTGACAGAAAATTTGTGCCCA**TTAACATCACCATCTAATTCACAAGAAT
 TGGGACAACCTCCAGTGAAAAGTTCTTCTCCTTTACT**GAATTC**GGCCGAGGATAATGATA
 GGAGAAGTGAAAAGATGAGAAAGAGAAAAAGATTAGTCTTCATTGTTATATCTCCTTGG
 ATCC**TA**ctccctccgtcccacaaaacatgacgttttaagggttagcagccaaaattagct
gttgtgcaaaaatgaccaaattgtccccatgatttgattaagctgtcatttacagcattt
gtacatgcatccagattattctagagaagtttctgaaaccacagctcagtgccacgtg
ttaacgaattggcgccttagccacacggttgatacagggcaaaacatcattaacatat
tcaaaaatttgaaatcaggtagggaaagattggggatcggcgatggttgggggcgatgga
gattggggatcggcgtggttgaggacgacggagagcgatggatgggggcgactagaga
gaggataagatcggagtagtactagcgcaacaaataaaaacgcacttcttttcttggt
tcaacctccacgtatacggaggggcccaccacttctctctcgacgacatttttctggga
caatccaggggcggtgaaacggcaggtttt**gtgggacggaggagTA**GGATCCTCTAGA
 GTCCCCCGTGTCTCTCCAAATGAAATGAACTTCCTTATATAGAGGAAGGGTCTTGCGA

AGGATAGTGGGATTGTGCGTCATCCCTTACGTCAGTGGAGATATCACATCAATCCA**CTT**
GCTTTGAAGACGTGGTTGGAACGTCT...

Using multiple alignments to compare many similar sequences

Comparing many similar DNA sequences at once is not easy to do by eye. To do this you use a program called ClustalW2 (<http://www.ebi.ac.uk/Tools/clustalw2/>) or MUSCLE (<http://www.ebi.ac.uk/muscle/>). More details on multiple alignments will be given when we discuss tree building. For now, you will use multiple alignments to compare all the sequencing results from the class. The input for either program is a multiple Fasta file. Open the "class_multi_align.fasta" file to see the format of a multiple Fasta file.

1. You will be provided a fasta file of the class sequences (class_multi_align.fasta). Each sequence in the will be from the same primer and in the same direction.
2. Open the MUSCLE link and make the selections shown below: Add a Search Title, Select ClustalW2 Output format, and upload the file. Then select "Run."

EBI > Tools > Sequence Analysis

MUSCLE

MUSCLE stands for **M**ultiple **S**equences **C**omparison by **L**og-**E**xpectation. MUSCLE is claimed to achieve both better average accuracy and better speed than [ClustalW2](#) or [T-Coffee](#), depending on the chosen options.

[Download Software](#)

MUSCLE
multiple sequence alignment method with reduced time and space requirements complexity

RESULTS interactive	SEARCH TITLE Qsmar excision	YOUR EMAIL
OUTPUT FORMAT ClustalW2	OUTPUT TREE none	OUTPUT ORDER aligned

Enter or Paste a set of Sequences in any supported format: [Help](#)

Upload a file: [Choose File](#) class_multi_align.fasta [Run](#) [Reset](#)

3. The results will resemble this. Click the "Output file" link to retrieve your results (these will not be colored). For Experiment 2 the coloring is not necessary.

Muscle Results

Results of search	
Number of sequences	4
Sequence type	DNA
Muscle version	MUSCLE v3.6 by Robert C. Edgar
Max length	740
Average length	502
Output file	muscle-20080207-15294644.output (clustalv)
Your input file	muscle-20080207-15294644.input

Alignment

MUSCLE (3.6) multiple sequence alignment

```

seq      -----GGGTTG-----AGGGATAGCCTCTCCACCC
-----GGNTCTA--TTCACA--GAATTGGGAC-ACTCCAGTAAAAGTCTTCTCCT
-----CTCGGCATCTAATTCAACA--GAATTGGGAC-ACTCCAGTAAAAGTCTTCTCCT
GTGNCNCGGCNTCTAATTCACAAGGAATGGGACAACTCCAGTAAAAGTCTTCTCCT
          *  *
          *  *

seq      -----AAGCGSCCGAGAACCTGCGTGCAATCCAATCATTGTTCAATCATGNGAAA----
TTACTGAATTCGGCCGAGGATAATGATAGGAGAA-----GTGAAAAGATGAGAAAGAGA
TTACTGAATTCGGCCGAGGATAATGATAGGAGAA-----GTGAAAAGATGAGAAAGAGA
TTACTGAATTCGGCCGAGGATAATGATAGGAGAA-----GTGAAAAGATGAGAAAGAGA
          *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *  *
  
```

4. Examine the results. There should be almost complete alignment between the sequences except at the excision site. What variation do you see there?

5. **Homework assignment:** Prepare a multiple alignment of all the sequences obtained by the class. Include any other sequence you think is relevant in the alignment. Annotate the EcoRI site(s), the TSD, and the PCR priming sites. Describe the types of modifications left by the excision of OsmaRNA. For the annotation you can use a computer program or print the alignment and annotate with pen or pencil.

Week 2: Getting a bit deeper into TEs

The genomes of plants and animals contain different families of transposable elements. This concept is central to understanding what genomes are made of.

What is a TE family?

We have already been introduced to two TE families. One family contains the Ac and Ds elements while the second family contains Osmar5 and Osmar5A elements.

In functional terms, a TE family contains all the elements that can be mobilized by a particular transposase. A TE family usually contains autonomous elements (e.g. Ac, Osmar5) and nonautonomous elements (e.g. Ds, Osmar5NA) elements. When we analyze the DNA sequence of entire genomes we often find one or more autonomous elements and many copies of nonautonomous elements (the maize genome has over 50 copies of Ds). The transposase encoded by the Ac element can mobilize both Ac and Ds elements. If there is no Ac element in the genome, all of the Ds elements will be “stuck” where they are - they will not be able to move elsewhere in the genome because there is no transposase to catalyze their movement. The same is true for Osmar5 and Osm5NA in rice – Osm5NA will be stuck in place if Osmar5 is not in the genome.

A very important feature of TE families is that each family is independent. In practical terms this means that the Ac transposase cannot mobilize Osmar5 or Osm5NA elements and, similarly, the Osmar5 transposase cannot mobilize Ac or Ds elements. Or, as shown in Figure 19, the transposase from family A cannot move the elements in family B.

The reason for this is quite simple. A transposase usually works by first binding to a specific DNA sequence near the ends of the element (as shown in Fig 5, on page 22). The Ac transposase first binds to a specific sequence of nucleotides that is only near the ends of Ac and Ds elements while the Osmar5 transposase binds to a specific sequence that is only near the ends of Osmar5 and Osm5NA elements. (Recall that in addition to catalyzing chemical reactions, proteins can also bind to DNA. Transposases are proteins that do both: bind to DNA and then catalyze the transposition reaction.)

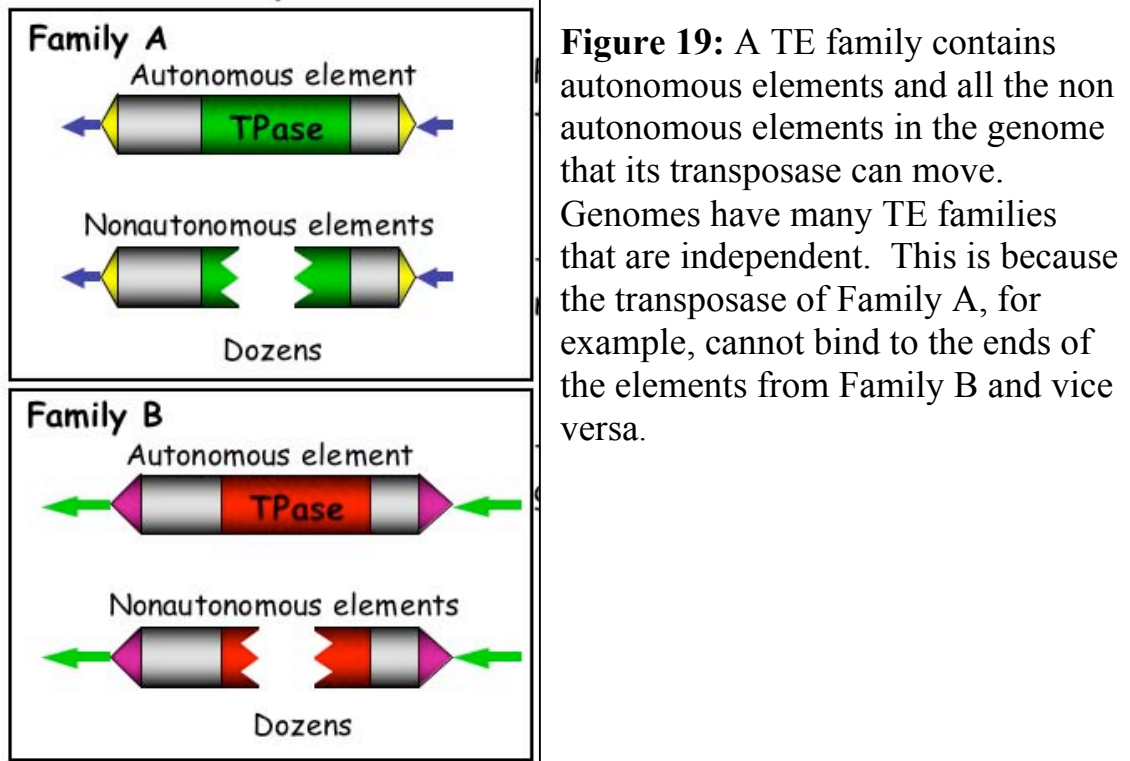


Figure 19: A TE family contains autonomous elements and all the non autonomous elements in the genome that its transposase can move. Genomes have many TE families that are independent. This is because the transposase of Family A, for example, cannot bind to the ends of the elements from Family B and vice versa.

What is a TE superfamily?

After McClintock discovered Ac and Ds she then discovered a second TE family which she called Spm (for Suppressor-mutator - a long story!). The autonomous element in this family is called Spm and the nonautonomous element is called dSpm (for defective-Spm). Thus, Spm-dSpm is another family of transposons.

McClintock's discoveries resulted from genetic analyses of corn plants. After the discovery of TEs in maize, researchers working with other model organisms, including *Antirrhinum majus* (a.k.a. snapdragon) *Drosophila melanogaster* (a.k.a. the fly) and *Caenorhabditis elegans* (a.k.a. the worm) also identified TEs through genetic studies. In the 1980's when it became possible to isolate specific genes, researchers isolated McClintock's Ac, Ds, Spm and dSpm elements and the elements from snapdragon (called Tam 1,2,3 etc), the fly (called P-elements, mariner elements and others) and the worm (called Tc1, 2 and 3 elements).

When the DNA sequences of these elements were determined and compared (by computer analysis), researchers were surprised to find that the transposases encoded by some of the elements from different species, even from different kingdoms (animal vs. plant), were similar. For example, the transposase from the maize Ac element was similar to the transposases of Tam3 from snapdragon and the P element from the fly, while the transposases of the mariner (fly) and Tc1 (worm) elements were similar.

These similar transposases were subsequently organized into superfamilies. Fortunately, after all of the sequencing of genomes and comparisons of TEs, there are now known to

be fewer than 10 superfamilies of transposases. Some superfamily names and elements and some members include: hAT (includes Ac, Tam3, P elements), CACTA (includes Spm, Tam1), PIF/Harbinger, Mutator and mariner. The distribution of some of the superfamilies across the tree of life is summarized in Figure 20.

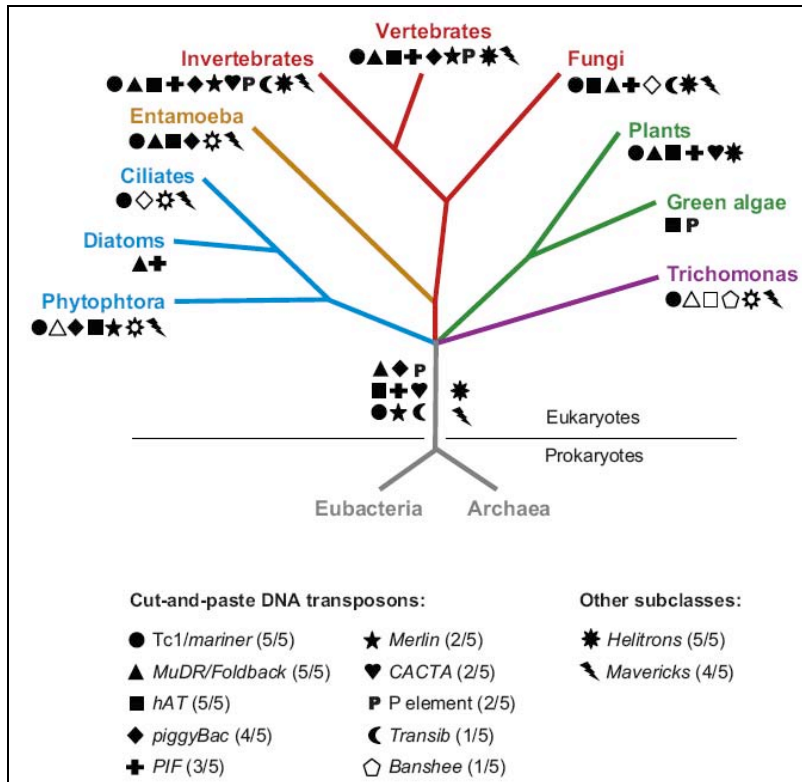


Figure 20. Distribution of the major groups of DNA transposons across the eukaryotic tree of life. The tree depicts 4 of the 5 “supergroups” of eukaryotes where DNA transposons have been detected. The occurrence of each superfamily/subclass of DNA transposons is denoted by a different symbol. (*Feschotte ·Pritham Annu. Rev. Genet.* 2007.41:331-68).

How many families and superfamilies can an organism have in its genome?

In short, many. First, members of most superfamilies are present in all plant genomes including maize and rice and are also present in most animal genomes (Figure 20). For example, the rice genome has *mariner*, *PIF*/Harbinger, *hAT*, *CACTA* and *Mutator* elements. In addition, each superfamily usually contains many families in one genome. Let’s look more closely at the *Osmar* element in rice.

Some background about Osmar

Osmar5 is a member of the mariner superfamily of DNA transposons which is widespread in plant and animal genomes (see Figure 20, where the mariner superfamily is also called Tc1/mariner and its distribution is indicated by a filled circle). Members of the mariner superfamily from different species were given names that were derived from the species name. "Osmar" stands for **Oryza sativa mariner** - *Oryza sativa* is the scientific name for rice and mariner is the name of the superfamily of elements. The rice genome contains Osmar5 and over 40 related but clearly different elements. Among these 40 elements are some that are very similar and others that are quite different. To understand the relationships between elements, their sequences can be organized and visualized as a family tree. The scientific term for such relationship maps is "phylogenetic trees". A tree of the Osmar elements in the rice genome is shown below in Figure 21. Later in this course you will learn how to construct your own phylogenetic tree from TE sequences that you will retrieve from the database. So... we will revisit trees a bit later.

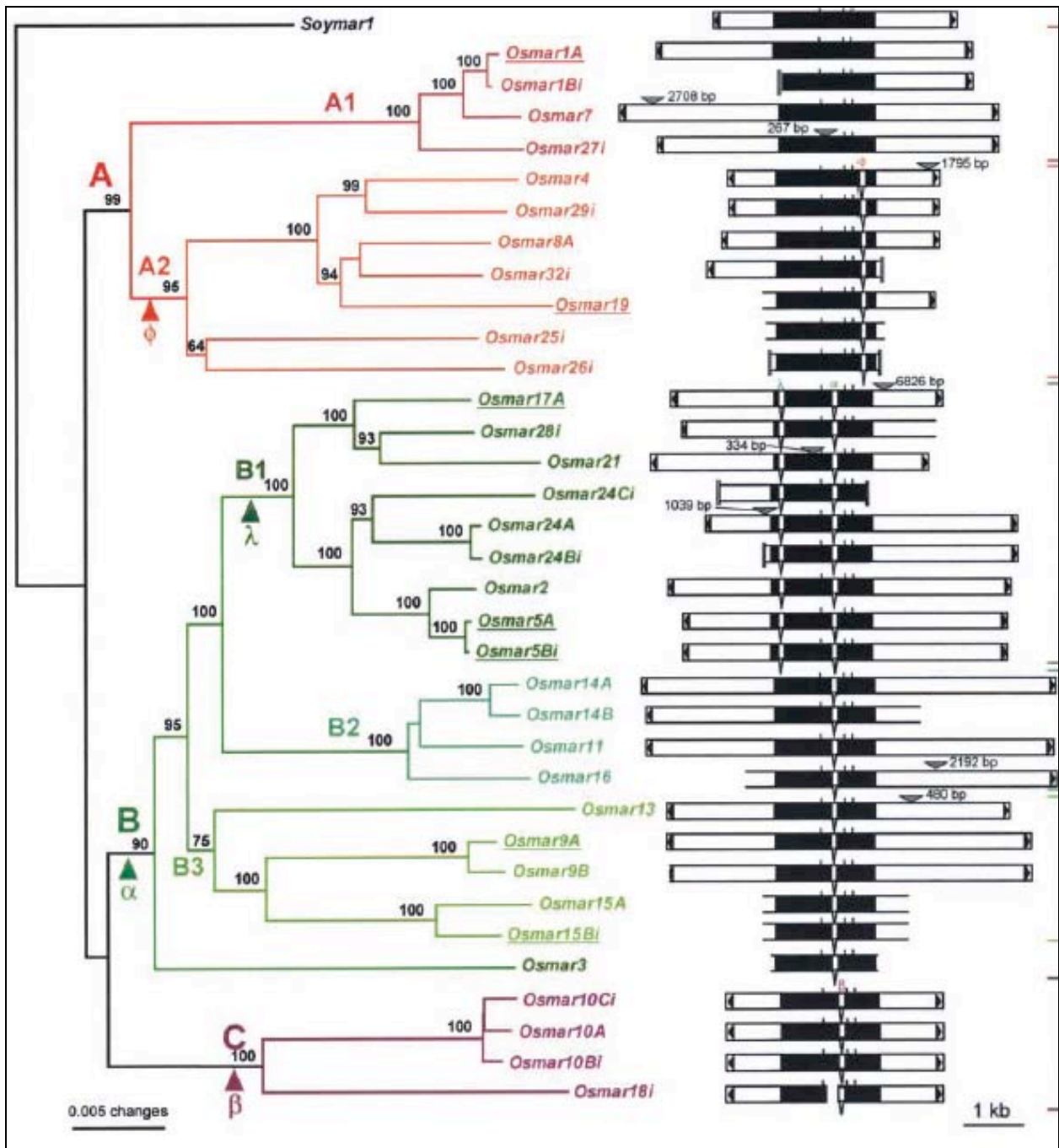


Figure 21: A phylogenetic tree of the relationships between Osmar elements in the rice genome. The outgroup was the mariner element *Soymar1* from soybean. (You will learn about outgroups later in the course.) The structure of each element is shown at the right with the transposase gene drawn as a black box interrupted by introns (as white gaps), and arrows for the terminal inverted repeats (TIRs).

An Introduction to the Other Transposable Element Class - Retrotransposons

The genomes of higher organisms contain hundreds of thousands, even millions of TEs. All of these elements are actually divided into 2 classes and up until now you have learned about one of the two classes. As you have seen, the elements discovered by McClintock are now known to be present in most eukaryotes and are grouped into several superfamilies. These elements are moved by transposases, have inverted repeats at their ends and transpose by excising from one place in the genome and reinserting somewhere else. Because the element itself moves from one site to another they are also called DNA elements. This seems like a “duh” statement – how else would an element move? Well, there is another way – many TEs move via an RNA copy. Such elements are called RNA or “retro” elements – and they are the focus of this section.

Retrotransposons (a.k.a. RNA elements, retroelements, class 1 elements)

What you see in the figure below is region of the chromosome (DNA) that contains a retrotransposon. Like any gene, this retrotransposon can be transcribed by RNA polymerase into a mRNA copy. Unlike most mRNAs (which are translated into protein by the ribosome), the mRNA copy of the retrotransposon serves as a template for the synthesis of a double stranded DNA copy of the element which then inserts at another site in the genome. The key enzyme involved is reverse transcriptase - the most abundant gene in the world!

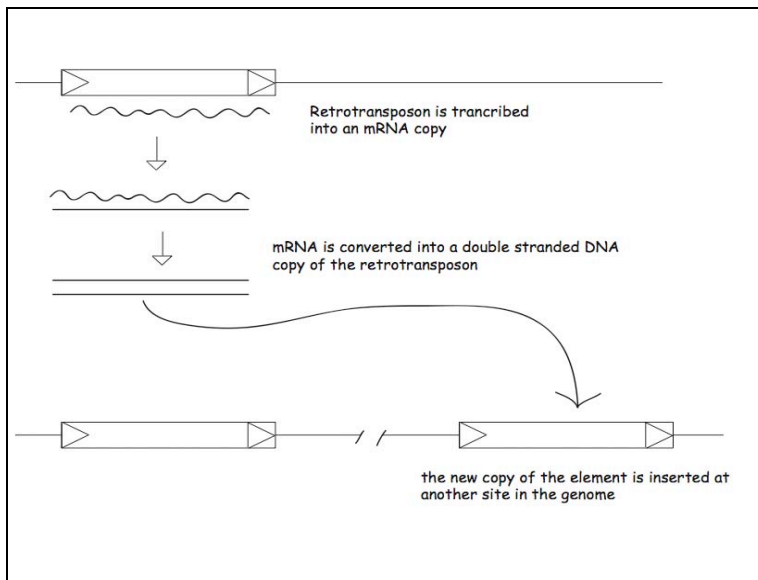


Figure 22. Like any gene, retrotransposons are transcribed by RNA polymerase. However, the RNA transcript can be “reverse transcribed” by the enzyme reverse transcriptase into a double stranded DNA copy that can insert elsewhere in the genome. Unlike DNA elements, retrotransposons do not excise from one site and insert elsewhere. Rather the RNA copy is made into DNA which then inserts elsewhere in the genome.

Class 1 elements are said to retrotranspose (retrotransposition) while class 2 elements transpose (transposition).

Three features of retrotransposition differ from that of transposition:

- (1) the transposition intermediate is the mRNA copy of the element. This feature is true for all class 2 elements.
- (2) like genes, a class 2 element can serve as template for many mRNA transcripts. Because each transcript has the potential to be converted into a new element, one element can produce many new elements. Class 2 elements are thus like printing presses that can potentially produce many new elements in the host genome.
- (3) once inserted, retrotransposons do not excise. Because they transpose through an RNA intermediate, the DNA copy of the element does not excise like DNA elements.

The structure of LTR retrotransposons is strikingly similar to a common pathogenic agent and a cause of some cancers - retroviruses.

How a retrovirus moves:

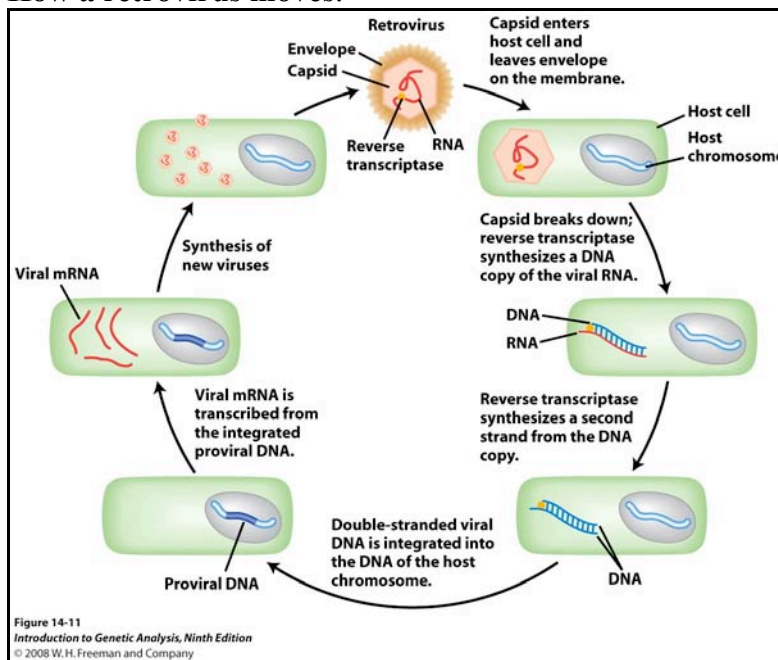


Figure 23: A retrovirus has a single stranded RNA genome that is reverse transcribed by the viral encoded reverse transcriptase into double stranded DNA. The DNA copy can integrate into the host genome (the provirus) where it can replicate along with the host genome. To make new viruses, the provirus is transcribed into mRNA which has two roles – (1) it is translated into viral proteins, (2) it is packaged into viral particles which can leave the cell and infect other cells starting the cycle all over again.

How a LTR retrotransposon moves:

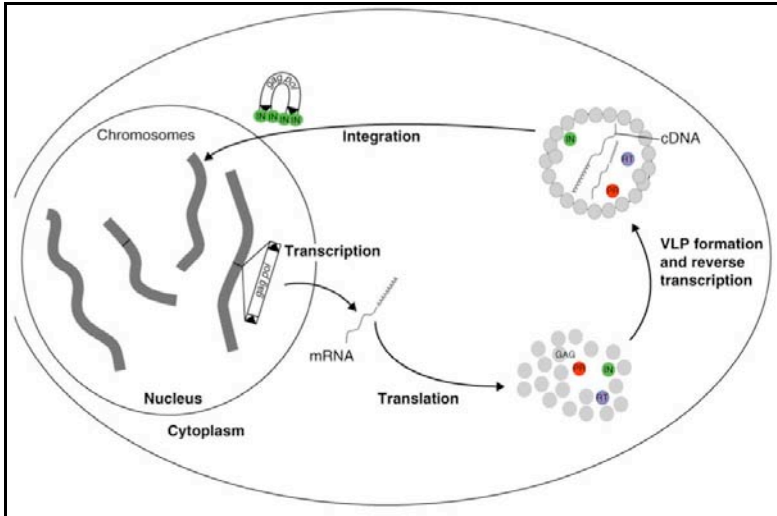


Figure 24: An LTR retrotransposon is transcribed into mRNA that serves two roles (1) it is translated into proteins needed for retrotransposition including reverse transcriptase (pol) and proteins needed to make the VLP where the mRNA is reverse transcribed into complementary DNA (cDNA) which then can integrate into the host genome. Unlike a retrovirus, DNA copies of LTR retrotransposons cannot leave the cell; they can only integrate in the genome.

The structures and gene content of LTR retrotransposons and retroviruses are very similar:

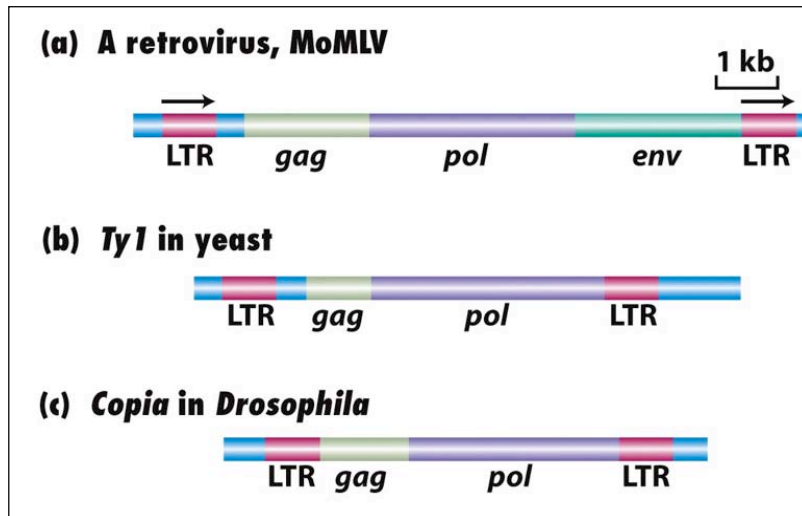


Figure 25: LTR = long terminal repeat, pol = reverse transcriptase gene. Ty1 (b) and Copia (c) are LTR retrotransposons. Note that retrotransposons do not have an “env” gene. This gene encodes the viral envelop and is required for retroviruses to leave the cell – something that LTR retrotransposons cannot do.

There are two key features of LTR retrotransposons (and retroviruses)– the LTR and the reverse transcriptase gene.

LTR: stand for long terminal repeat. Recall that DNA transposons had inverted terminal repeats (TIRs)? Well, LTRs are at the end of retrotransposons. Fortunately, LTRs are easier to understand than TIRs because the same sequence is repeated at the ends (that is why the arrows are pointing in the same direction in the figure above). LTRs are usually

much longer than TIRs. LTRs usually range from 50 bp (base pair) all the way up to 2000bp or even more!

pol: is the shorthand designation for the reverse transcriptase gene, which encodes reverse transcriptase, which is a polymerase enzyme. Recall that DNA polymerase is used for DNA replication (synthesizes DNA from a DNA template) while RNA polymerase is used for transcription (synthesizes RNA from a DNA template). Well, reverse transcriptase synthesizes DNA from a RNA template (it is the reverse of transcription).

An interesting fact to wow your friends with –

pol is the most abundant gene in the world! We can say this because retrotransposons make up the majority of TEs in most genomes. Of the ~50% of the human genome that is derived from TEs, most of this is retrotransposon. Only about 3% of the human genome is derived from DNA transposons.

Identifying LTR Retrotransposons in Genomic Sequences – a Bioinformatic Experiment

Objective: *To find and characterize LTR retrotransposons in genomic sequence. We will use a query (below) from the reverse transcriptase domain to “mine” related elements in the rice genome. This information will be used to “venture out” into the wilds of the sequence surrounding our Blast hits to identify the telltale structure that distinguishes this element type - long terminal repeats (LTRs). Knowledge of the LTR positions will allow us to define a complete element (the LTRs and all the sequence in between).*

Before reading this, review pages 12 - 17 in your course notes. This will provide information on protein blast and tblastn. These two types of blast use protein sequence as the query. In order to find divergent sequences it is useful to use tblastn. Pages 12 to 17 will explain why.

Step 1: As with all of our bioinformatics experiments we start with a query sequence....

(partial reverse transcriptase Copia; LTR retrotransposon in rice):

>SZ-55

```
GGLGERVTRNRSYELVNSAFVASFEPKNVCHALSDENWVNAMHEELENFERNKVWSLVE
PPLGFNVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRI
LLAFAASKGFKLFQMDVKSAFLNGVIEEEVYVKQPPGFENPKFPNHVFKLEKALYGLKQ
APRAWYERLKTFFLLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIFGSSHALVAQF
SDVMSREFEMSMGELTFFLGLQIKQTKEGIFVHQTKEYSKELLKKFDMADCKPIATPMA
TTSSLGPDEEDGEVDQREYRSMIGSLLYLTASRPDIHFSVCLCARFQASPRTSHRQAVK
RIFRYI
```

Which is used in a tblastn search for related elements in the rice genome...

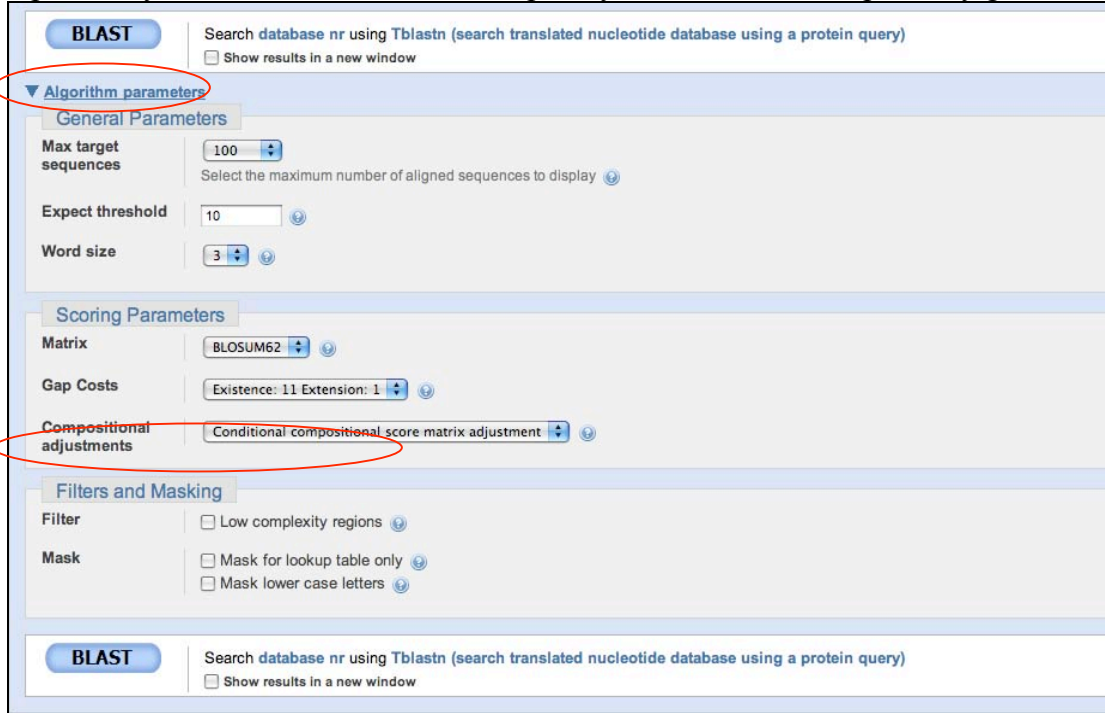
Blast search (<http://www.ncbi.nlm.nih.gov/BLAST/tblastn>)

Database: Nucleotide collection (nr/nt)

Organism: *Oryza sativa* (taxid:4530)

The screenshot shows the 'Choose Search Set' section of the NCBI BLAST search interface. It includes three main input areas: 'Database' with a dropdown menu set to 'Nucleotide collection (nr/nt)', 'Organism' with a text box containing 'Oryza sativa (taxid:4530)' and a note 'Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.', and 'Entrez Query' with an empty text box and a note 'Enter an Entrez query to limit search'.

Uncheck “low complexity filter” in “Algorithm parameters” menu. (Low complexity filtering removes common repeats found in REPbase. TEs are considered common repeats. Try the search with the low complexity filter on. You will probably get zero hits.



Click “BLAST”

Step 2: Defining the ends of the element (finding LTRs and TSDs): retrieving a BAC that contains one of your blast hits.

Let’s go back to your tblastn result. First, pick a hit that comes from a BAC, not from a longer contig (like a pseudomolecule) or from an EST or mRNA. We will explain why in class:

Sequences producing significant alignments:	Score (Bits)	E Value
dbj AP008209.1 Oryza sativa (japonica cultivar-group) genomi...	723	0.0
gb AC092559.4 Oryza sativa chromosome 3 BAC OSJNBb0096M04 ge...	723	0.0
gb AC107224.2 Oryza sativa (japonica cultivar-group) chromos...	721	0.0
emb AL606652.4 Oryza sativa genomic DNA, chromosome 4, BAC c...	721	0.0
dbj AP008210.1 Oryza sativa (japonica cultivar-group) genomi...	721	0.0
gb AC137696.2 Genomic sequence for Oryza sativa, Nipponbare ...	721	0.0
dbj AP008207.1 Oryza sativa (japonica cultivar-group) genomi...	720	0.0
dbj AP002538.2 Oryza sativa (japonica cultivar-group) genomi...	720	0.0
dbj AP008215.1 Oryza sativa (japonica cultivar-group) genomi...	718	0.0
dbj AP006849.2 Oryza sativa (japonica cultivar-group) genomi...	718	0.0

Click the score on this line to see the details of the blast hit:

```

> [emb|AL606652.4] [D] Oryza sativa genomic DNA, chromosome 4, BAC clone: OSJNBb0004A17,
complete sequence
Length=159894

Score = 721 bits (1862), Expect = 0.0
Identities = 359/360 (99%), Positives = 360/360 (100%), Gaps = 0/360 (0%)
Frame = -1

Query 1      GGLGERVTRNRSYELVNSAFVASFEPKNVCHALSDENWVNAMHEELENFERNKVWSLVEP 60
Sbjct 17511  GGLGERVTRNRSYELVNSAFVASFEPKNVCHALSDENWVNAMHEELENFERNKVWSLVEP 17332

Query 61     PLGFNVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILL 120
Sbjct 17331  PLGFNVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILL 17152

Query 121    AFAASKGFKLFQMDVKSAFLNGVIEEEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPR 180
Sbjct 17151  AFAASKGFKLFQMDVKSAFLNGVIEEEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPR 16972

Query 181    AWYERLKTFLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIFGGSSHALVAQFSDVM 240
Sbjct 16971  AWYERLKTFLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIFGGSSHALVAQFSDVM 16792

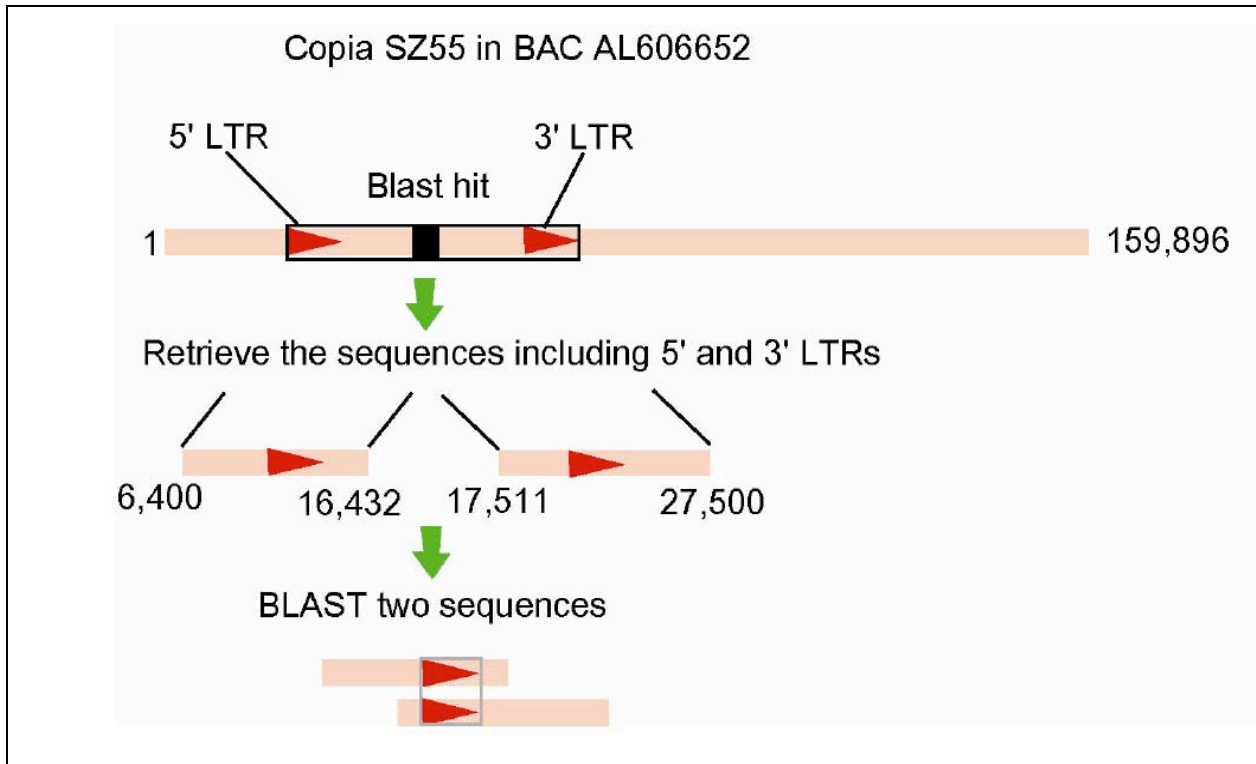
Query 241    SREFEMSMGELTFFLGLQIKQTKEGIFVHQTKYSKELLKFFDMADCKPIATPMATTSSL 300
Sbjct 16791  SREFEMSMGELTFFLGLQIKQTKEGIFVHQTKYSKELLKFFDMADCKPIATPMATTSSL 16612

Query 301    GPDEDGEEVDQREYRSMIGSLLYLTA SRPDIHFSVCLCARFQASPRTS HRQAVK RIFRYI 360
Sbjct 16611  GPDEDGEEVDQREYRSMIGSLLYLTA SRPDIHFSVCLCARFQASPRTS HRQAVK RIFRYI 16432

```

The sbjct is in the "minus" direction (see Frame = -1) meaning that the hit reads in the opposite direction as the numbering of the BAC sequence in the database. The BAC is 159,894 bp long and this hit begins at position 16432 and ends at position 17511. Write these numbers down. Now we know where the reverse transcriptase is in this BAC. Our goal is to determine the complete copia element, but first we have to retrieve the whole BAC sequence and use this to figure out the element ends.

We can make an educated guess as to position of the complete element on this BAC by taking into account the following considerations: (i) LTRs are at the end of this element, (ii) most LTR retrotransposons are no longer than 15KB, and (iii) the RT domain is usually near the middle of the complete element. Thus, our RT hit should be less than 10kb from each end of the element. To precisely identify the LTRs, we need to retrieve the BAC sequences containing the so-called 5' and 3' LTRs and compare them using "[BLAST 2 SEQUENCES](#)". Here is a visual of our search strategy...



(NOTE that unlike in your search, you will not know where the arrows are that represent the 5' and 3' LTR. That is the objective of this protocol)

Step 3: Retrieving the complete BAC sequence

Click the BAC's name: emb|AL606652.4

```
> emb|AL606652.4 ■ Oryza sativa genomic DNA, chromosome 4, BAC clone: OSJNBb0004A17,
complete sequence
Length=159894

Score = 721 bits (1862), Expect = 0.0
Identities = 359/360 (99%), Positives = 360/360 (100%), Gaps = 0/360 (0%)
Frame = -1
```


A new webpage will show up. This page contains all of the information about this BAC including its complete sequence - yes - all 159,896 plus bases. You will use this later.

NCBI Nucleotide Protein Genome

Search Nucleotide for [] Go Clear

Limits Preview/Index History

Display GenBank Show 5 Send to Hide: sequence all but gene, CDS and mRNA features

Range: from begin to end Reverse complemented strand Features: + Refresh

1: [AL606652](#). Reports *Oryza sativa* geno...[gi:70663936]

[Comment](#) [Features](#) [Sequence](#)

LOCUS AL606652 159894 bp DNA linear PLN 08-JUL-2005

DEFINITION *Oryza sativa* genomic DNA, chromosome 4, BAC clone: OSJNBb0004A17, complete sequence.

ACCESSION AL606652

VERSION AL606652.4 GI:70663936

KEYWORDS HTG.

SOURCE *Oryza sativa* (japonica cultivar-group)

ORGANISM [Oryza sativa \(japonica cultivar-group\)](#)
Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; BEP clade; Ehrhartoideae; Oryzaceae; Oryza.

REFERENCE 1

AUTHORS Feng, Q., Zhang, Y., Hao, P., Wang, S., Fu, G., Huang, Y., Li, Y., Zhu, J.,

Scroll down to the bottom of the page to view the BAC sequence.

ORIGIN

```

1 gaattctttc aaatgtttct tcaactttag caactgtctc ctttgagacc tgatggccag
61 ccttatcaaa gactgcataa ctgtaacaga atcaattgac agagttgatg taagaatcaa
121 caaggattgt gcggatcggg aaagaaaagc gtaagatcaa gagctaaaag attacctttc
181 taaatcatga tcatacagaa cagagttgtc gccactagtg cgatataatt tcagcccaag
241 atagccaatc aatggtgcca aggagttctc attacaaaac ccgtaggagc tgaatttttg
301 gaatgaacag taagtaagct tgtatgaaca gaatctaaag tgaatttttc acactaacia
361 ttcagggtga gactgacatc gaggctccca tatcaattgg gcatccgaaa gagtaatcgg
421 tatggacacg accgcctacg cgatctctgg actccaaaac agtcacctca aacgaagcat
481 tggagagagc acgggctgct gcaacccttg aaattcccc accgatcagc atgacgggatg
541 gaggcgaagc acattgcctc tcaatggctc gaagcaagag gcctgaacia aaaatgtttt
601 ttactgtcag gtatgtgaat cataagagag agaaatcacg ttgaacatca agctcactaa
661 tctacataat actgtagata cccaagttac caactaacta accaatttgt acccaactag
721 aattataaat tctaataatc ttgtaaaatc taaagtgtga tgatcacctt ccctatgtgg

```

Step 4: Retrieving only the sequences that you estimate should include the LTRs and Blasting them against each other:

Open a blast2seq page: <http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi>

a. Type the accession number of the BAC into the Sequence 1 text field, AL606652 and enter the limits from 6400 to 16432.

b. Type the accession number of the BAC into the Sequence 2 text field, AL606652 and enter the limits from 17511 to 27000.

Program: **blastn** Matrix: **Not Applicable**

Parameters used in **BLASTN** program only:
Reward for a match: 1 **Penalty for a mismatch:** -2

Use **Mega BLAST** Strand option: **Both strands** View option: **Standard**
 Masking character option: **X for protein, n for nucleotide** Masking color option: **Black**
 Show CDS translation

Open gap: 5 and extension gap: 2 penalties
 gap x_dropoff: 50 **expect**: 10.0 word size: 11 **Filter** **Align**

Sequence 1
 Enter accession, GI or sequence in FASTA format **from:** 6400 **to:** 16432
 AL606652

or upload FASTA file **Choose File** no file selected

Sequence 2
 Enter accession, GI or sequence in FASTA format **from:** 17511 **to:** 27000
 AL606652

or upload FASTA file **Choose File** no file selected

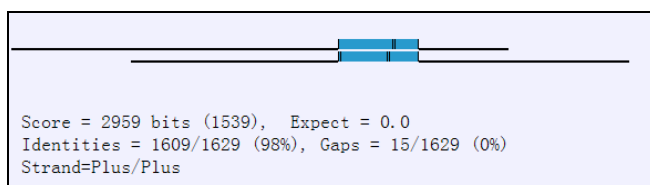
Align **Clear Input**

Note: uncheck “Filter” option before clicking Align.

Parameters used in **BLASTN** program only:
Match: 1 **Mismatch:** -2

Open gap: 5 and extension gap: 2 penalties
 gap x_dropoff: 30 **expect**: 10.0000 word size: 11 **Filter** **Align** **Clear Input**

The result should look like this:



Black lines are query (top) and sbjct (bottom); blue bars stand for matched regions and the small black bars in the blue are gaps in the matched sequence.

Step 5: Retrieving the entire Copia element from the BAC sequence.

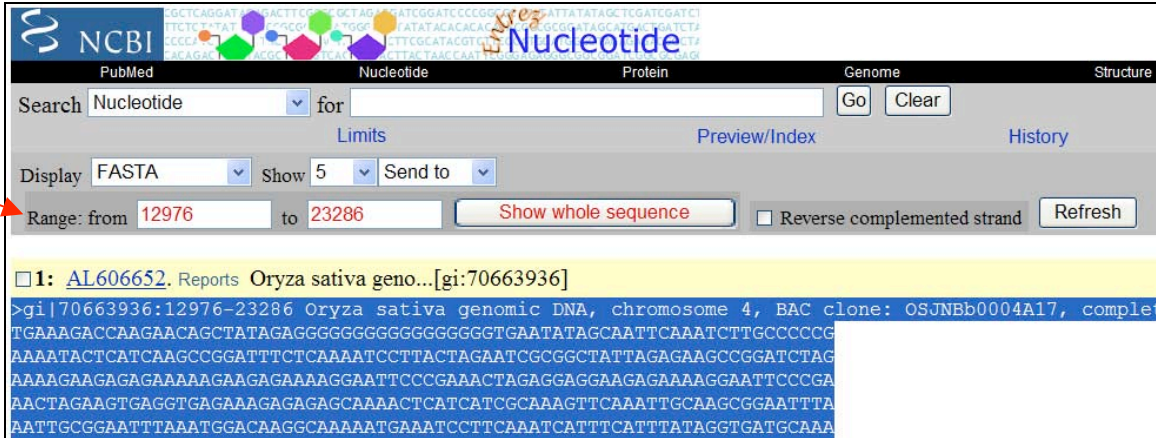
In order to retrieve the complete element, we can use our LTR sequence to Blast the entire BAC that it came from. When we do this, the LTR should match in two places along the length of the BAC that define the two ends of the complete element (remember LTR = long terminal repeats).

Query	12976	TGAAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGGGGGTGAATATAGCAATTCAAAT	13035
Sbjct	21669	TGAAAGACCAAGAACAGCTATAGAGGG-----GCGGGTGAATATAGCAATTCAAAT	21719
Query	13036	CTTGCCCCGAAAATACTCATCAAGCCGGATTTCTCAAAATCCTTACTAGAAATCGCGGCT	13095
Sbjct	21720	CTTGCCCCGAAAATACTCATCAAGCCGGATTTCTCAAAATCCTTACTAGAAATCGCGGCT	21779
Query	13096	ATTAGAGAAGCCGGATCTAGAAAAGAAGAGAGAAAAGAAGAGAAAAGGAATTCCCGAAA	13155
Sbjct	21780	ATTAGAGAAGCCGGATCTAGAAAAGAAGAGAGAAAAGAAGAGAAAAGGAATTCCCGAAA	21839
Query	13156	CTAGAGGAGGAAGAGAAAAGGAATTCCCGAAACTAGAAGTGAGGTGAGAAAAGAGAGAGCA	13215
Sbjct	21840	CTAGAGGAGGAAGAGAAAAGGAATTCCCGAAACTAGAAGTGAGGTGAGAAAAGAGAGAGCA	21899
Query	13216	AAACTCATCATCGCAAAGTTCAAATTGCAAGCGGAATTTAAATTTGCGGAATTTAAATGGA	13275
Sbjct	21900	AAACTCATCATCGCAAAGTTCAAATTGCAAGCGGAATTTAAATTTGCGGAATTTAAATGGA	21959
Query	13276	CAAGGCAAAAATGAAATCCTTCAAATCATTTTATAGGTGATGCAAAAATAACCGCTC	13335
Sbjct	21960	CAAGGCAAAAATGAAATCCTTCAAATCATTTTATAGGTGATGCAAAAATAACCGCTC	22019
Query	13336	AACTAGGAGCAAACATACACCTTCAGAGGAACATTAACACAAACTTAAATCTCTCGGAC	13395
Sbjct	22020	AACTAGGAGCAAACATACACCTTCAGAGGAACATTAACACAAACTTAAATCTCTCGGAC	22079
Query	13396	AAACACACTCCAAACTAATCCTAATACAAAAGCCTCTCGGGCAAACACACTCCAAACTCA	13455
Sbjct	22080	AAACACACTCCAAACTAATCCTAATACAAAAGCCTCTCGGGCAAACACACTCCAAACTCA	22139
Query	13456	CACGGAAACTCTCTCACCGAGCATCTCAAATGATTACCAAAGGAGCAACCTCCACCCT	13515
Sbjct	22140	CACGGAAACTCTCTCACCGAGCATCTCAAATGATTACCAAAGGAGCAACCTCCACCCT	22199
Query	13516	TGCATCCATCTCTCTATTTATAGCCTAAGACCCCTAAGACATTTTCTCAAATACCCCTAG	13575
Sbjct	22200	TGCATCCATCTCTCTATTTATAGCCTAAGACCCCTAAGACATTTGCTCAAATACCCCTAG	22259
Query	13576	GGCGAAACCCCTAACTCAGAACAGATCTGGTCCATCCATTGTTCTTCTACTCAAAGGAAA	13635
Sbjct	22260	GGCGAAACCCCTAACTCAGAACAGATCTGGTCCATCCATTGTTCTTCTACTCAAAGGAAA	22319
Query	13636	AGCTCCAGATGATTGCCACCTCATCGATCCGAACCTCACATGGTCTTCTTGCTGATGAA	13695
Sbjct	22320	AGCTCCAGATGATTGCCACCTCATCGATCCGAACCTCACATGGTCTTCTTGCTGATGAA	22379
Query	13696	TCCGCGTGCTCTCCGTCTTGACGAATCAAACCTGCCACGTTGCCAGCCGCGTCTGCG	13755
Sbjct	22380	TCCGCGTGCTCTCCGTCTTGACGAATCAAACCTGCCACGTTGCCAGCCGCGTCTGCG	22439
Query	13756	CGTTCCGTCTCCTTTTCTCAGCGCGGCTCGCCAGCGCCCAACCAGCCCGAAGCCGCCA	13815
Sbjct	22440	CGTTCCGTCTCCTTTTCTCAGCGCGGCTCGCCAGCGCCCAACCAGCCCGAAGCCGCCA	22499

Query	13816	CGCGTCCCCTGAGCCGCTCCC	CGCGCTGCGCTTGCAAAATCGCGTGTGGGTCCC	CGCGCTCC	13875
Sbjct	22500	CGCGTCCCCTGAGCCGCTCCC	CGCGCTGCGCTTGCAAAATCGCGTGTGGGTCCC	CGCGCTCC	22559
Query	13876	TGCACCCGCTCCGATCGAGT	CACGCGAAACGGGGTTGCCGCGTACGGTTTTTCCGCGC		13935
Sbjct	22560	TGCACCCGCTCCGATCGAGT	CACGCGAAACGGGGTTGCCGCGTACGGTTTTTCCGCGC		22619
Query	13936	GCGTCCTTGCGCATGCAAGCT	TGCTTGCAC--TCTGAGCCCATGCGCCATGGGCCGCGCT		13993
Sbjct	22620	GCGTCCTTGCGCATGCAAGC	-TGCCTGCACCTCCTGAGCCCATGCGCCATGGGCCGCGCT		22678
Query	13994	CTTGGGCCGCTCGCTGCTTGG	ACGATGCAAGGCCTGCCGAGCCGAGCCATTACCATTCTTG		14053
Sbjct	22679	CTTGGGCCGCTCGCTGCTTGG	ACGATGCAAGGCCTGCCGAGCCGAGCCATTACCATTCTTG		22737
Query	14054	GGCCACGTGGAACGGCTGGATT	GG--TTGGCCTCCCTGCCAACCAATCACAGCGCACATGC		14112
Sbjct	22738	GGCCACGTGGAACGGCTGGATT	GG--TTGGCCTCCCTGCCAACCAATCACAGCGCACATGC		22797
Query	14113	ACAGCTAGCTAG--TTGACTTTT	CCACCGAGCCATGTTAGTAGCAACCAGTACAGTGCAAG		14171
Sbjct	22798	ACAGCTAGCTAGCTTACTTTT	CCACCGAGCCATGCTAGTAGCAACCAGTACAGTGCAAG		22857
Query	14172	CTCCTCCTTGACACAAGTACAG	TACGTGTACATGCATGTATGCTACCTACAGCAAGTACTG		14231
Sbjct	22858	CTCCTCCTTGACACAAGTACAG	TACGTGTACATGCATGTATGCTACCTACAGCAAGTACTG		22917
Query	14232	TAGCAGCAATGCACTTGCACAGT	CCCTTCTGATTTCTTCGCGAATCCGATGCTTGCACAC		14291
Sbjct	22918	TAGCAGCAATGCACTTGCACAGT	CCCTTCTGATTTCTTCGCGAATCCGATGCTTGCACAC		22977
Query	14292	TTGGCCTTGTGAAGCCTGTTG	CAAAGACCTTTTCACACGGTGTTCGTCCACCGTGTGCAA		14351
Sbjct	22978	TTGGCCTTGTGAAGCCTGTTG	CAAAGACCTTTTCACACGGTGTTCGTCCACCGTGTGCAA		23037
Query	14352	CCTTGTGTCCAATCTTGTCA	CCCCGGCATCCTTGATCGCTTTGGACCTCAACTCCTCCCTG		14411
Sbjct	23038	CCTTGTGTCCAATCTTGTCA	CCCCGGCATCCTTGATCGCTTTGGACCTCAACTCCTCCCTG		23097
Query	14412	AGTCTAGTCCCGATCCGCCGTT	GACCAAGATCGACCCCGATCACCTGCACACACATGAAC		14471
Sbjct	23098	AGTCTAGTCCCGATCCGCCGTT	GACCAAGATCGACCCCGATCACCTGCACACACATGAAC		23157
Query	14472	CAAACAACCGTTGTCTTGGC	CACAGATGTCGCAACCTGACCAACGTTAGTCCACACACACA		14531
Sbjct	23158	CAAACAACCGTTGTCTTGGC	CACAGATGTCGCAACCTGACCAACGTTAGTCCACACACACA		23217
Query	14532	CTTCTTGCACATCCGGTACTT	TGCAATTTCCCATCACAAAAGAACTATAACCACACATGG		14591
Sbjct	23218	CTTCTTGCACATCCGGTACTT	TGCAATTTCCCATCACAAAAGAACTATAACCACACATGG		23277
Query	14592	TTTCACAAT	14600		
Sbjct	23278	TTTCACAAT	23286		

Write down the lowest and highest numbers - “12,976 and 23,286”. These define the location of this complete element on BAC AL606652. One LTR is 12,976 to 14,600 (green ovals) and the other is 21,669 to 23,286 (red ovals).

To retrieve the complete element sequence we simply go back to the webpage of BAC AL606652 and input the start and end locations into the “Range” windows as follows:

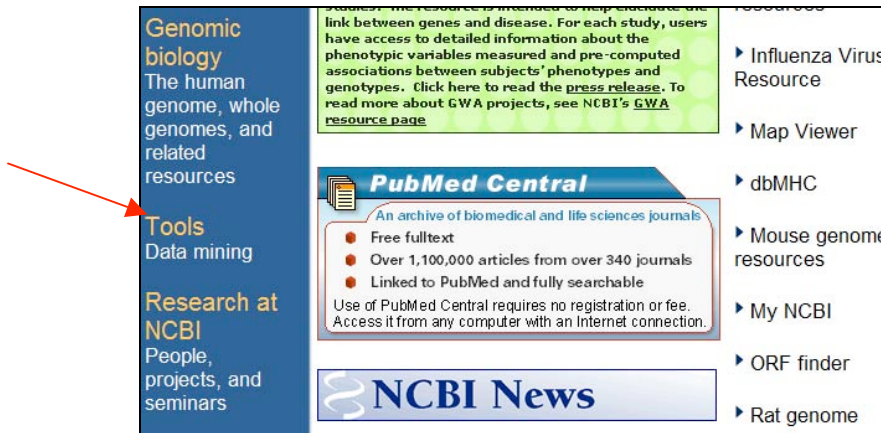


Click Refresh.

Save this sequence as a word file and call it something like complete element. You will need it later.

Step 6: Finding the element-encoded open reading frames (ORFs):

Go to the homepage of NCBI and find the “Tools” at the left. Click it.



From the new webpage, find “ORF finder” in the left part. Click it.

Map Viewer
Interactive chromosome viewer

Model Maker
View evidence used to build a gene model

ORF finder
Open reading frames

Organism Specific Resources
Bee, Cat, Chicken, Cow, etc.

how comes in several types including PSI-BLAST, PPI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

BLINK - ("BLAST Link") displays the results of BLAST searches that have been done for every protein sequence in the Entrez Proteins data domain.

CD Search - search the Conserved Domain Database with Reverse Position Specific BLAST.

CDART - when given a protein query sequence, CDART displays the functional domains that make up the protein and lists proteins with similar domain architectures.

Open Mass Spectrometry Search Algorithm (OMSSA) - The OMSSA search service allows proteomics researchers to submit the mass spectra of peptides and proteins for identification. OMSSA then compares these mass spectra to theoretical ions generated from data libraries of known protein sequences and ranks the results using a score derived from classical hypothesis testing.

TaxPlot - a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. To use TaxPlot, one selects a reference genome to which two other genomes are compared. Pre-computed

Now, paste your saved copia sequence into the sequence input window and click "OrfFind".

NCBI **ORF Finder (Open Reading Frame Finder)**

PubMed Entrez BLAST

NCBI
Tools for data mining
GenBank sequence submission support and software
FTP site download data and software

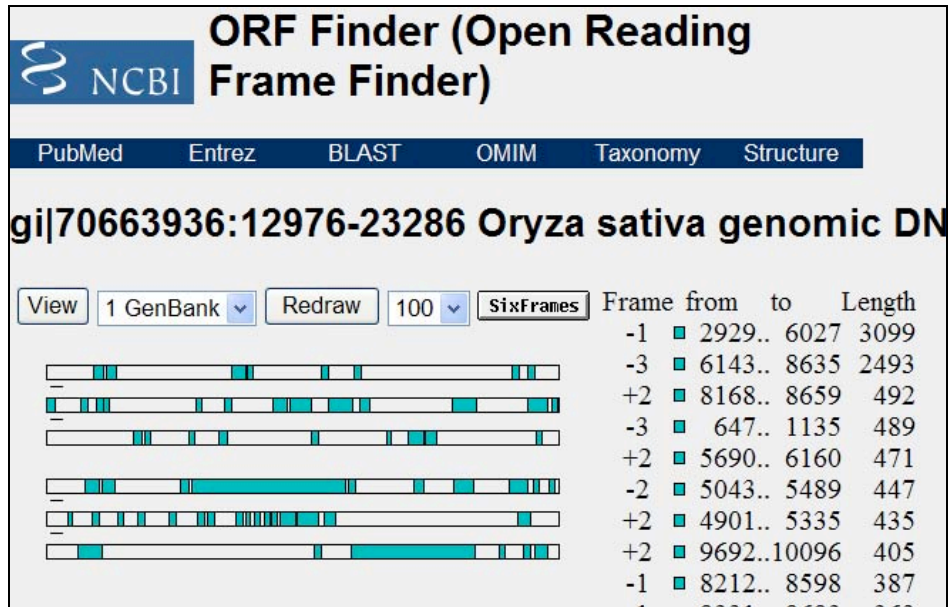
The ORF Finder (Open Reading Frame Finder) is a graphical analysis already in the database. This tool identifies all open reading frames using the standard or alternate against the sequence database using the WWW BLAST server. The C with the Sequin sequence submission software.

Enter GI or ACCESSION **OrfFind**

or sequence in FASTA format

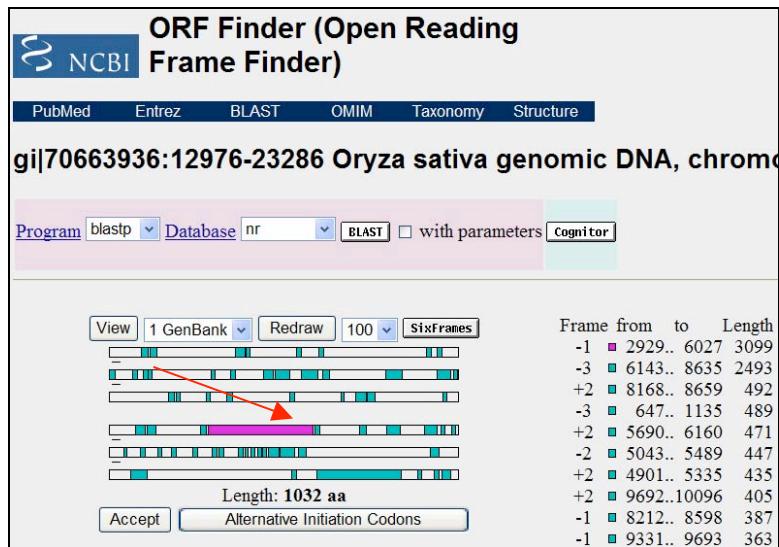
```
>gi|70663936:12976-23286 Oryza sativa genomic DNA,
chromosome 4, BAC clone: OSJNb0004A17, complete
sequence
TGAAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGTGAATATAGCAAT
TCAAATCTTGCCCCCG
AAAATACTCATCAAGCCGGATTTCAAAATCCTTACTAGAATCGCGGCTATTA
GAGAAGCCGGATCTAG
AAAAGAAGAGAAAAAGAGAAAAGGAATCCCGAACTAGAGGAGGAAGA
```

The result will look something like this:



The colored bars are the predicted ORFs. To see what they represent, you can either click on the regions in the bar itself or click on the match in the list at the right.

Note: usually the longer the ORF, the more reliable the information. Let's click the longest one and see what happens.



Select the ORF you're interested in. Click "BLAST" at the top of the new page.

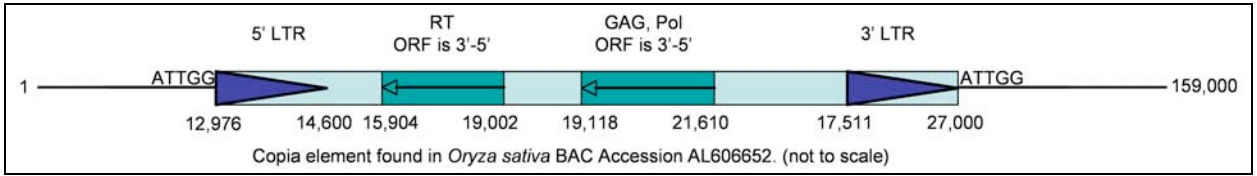
A new page will pop up. Just click “view report”.

You can see details of the blast result when it is done.

Sequences producing significant alignments:	(Bits)	Value
gb ABF94836.1 retrotransposon protein, putative, unclassifie...	2057	0.0
gb ABF93543.1 retrotransposon protein, putative, unclassifie...	2057	0.0
gb AAN60494.1 Putative Zea mays retrotransposon Opie-2 [Oryz...	2057	0.0
emb CAE03600.2 OSJNBb0004A17.2 [Oryza sativa (japonica cultivar	2057	0.0
gb AAO37957.1 putative gag-pol polyprotein [Oryza sativa (ja...	2051	0.0
gb AAW57789.1 putative polyprotein [Oryza sativa (japonica cult	1905	0.0
ref NP_001061216.1 Os08g0201800 [Oryza sativa (japonica cult...	1444	0.0
gb AAP53706.1 retrotransposon protein, putative, unclassifie...	1430	0.0
gb ABA93940.1 retrotransposon protein, putative, Tyl-copia s...	1384	0.0
gb AAT85178.1 putative polyprotein [Oryza sativa (japonica c...	1375	0.0
gb ABF97694.1 retrotransposon protein, putative, unclassifie...	1347	0.0

The longest ORF appears to be the reverse transcriptase. Click on the next longest ORF on the ORF Finder page. It is the gag, pol, env ORF. The importance of these will be discussed in class.

You can now fully annotate the Copia element you retrieved from the BAC by using a diagram like this....



An Introduction to Phylogenetic Tree Construction using TATE

One aspect of bioinformatics is determining how sequences are related. These relationships can be determined using phylogenetic trees. In transposable element research phylogenetic trees can be used to show:

1. Elements that are related by descent.
2. Predict active elements.
3. Predict new elements.
4. Cluster elements into groups.

In this exercise you will learn to use TATE to construct phylogenetic trees.

Steps for tree construction:

1. **Sequence assembly.** The first step in building a phylogenetic tree is assembling the sequences to be included in the tree. In TATE you will use a TE sequence as a query and blast it against a target database. TATE uses a 'local' blast search instead of NCBI Blast for speed.
2. **Re-formatting and cleaning the Blast results.** Blast is very good at finding relevant hits statically, but it knows little about biology. The Blast results must be modified to allow for introns in the hits. Also the Blast output must be re-formatted into fasta type output. Before TATE this was a laborious step.
3. **Multiple Alignment.** Before a tree can be constructed the sequences must be ordered in terms of relatedness. A multiple alignment program does this by doing all possible pairwise alignments and produces an output file in fasta format with the sequences in order of relatedness to each other. Currently TATE uses a program called Muscle to produce the multiple alignment.
4. **Tree Construction.** The phylogenetic tree is calculated from the multiple alignment. A tree building program considers many trees that fit the data in an attempt to re-construct the past. TATE uses TreeBeSt a phylogenetic program that produces the one tree that is the most likely reconstruction of evolution of the sequences provided.

We will go over each step in TATE. We will discuss in detail parts of TATE that you need to pay attention to, parts that are modifiable, and discuss what how to use the results.

1. First TATE Session.

In the first TATE Session we will use a part of the transposase gene from Osmar5. This is the same element we have been working with in the lab. We will go step-by-step through TATE with this example.

1 The query sequences is Osmar5:

>Osmar5

```
SKDLTNIQRRGIYQLLLQKSKDGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGIAVNVDSRKHGNS
GRKKVEIDL SVIAAIPLHQRRNIRSLAQALGVPKSTLHRWFKEGLIRRHSNSLKPYLKEANKKERLQWCVS
MLDPHTLPNNPKFIEMENIIHIDEKWFNASKKEKTFYLYPDEEPEYFTVHNKNAIDKVMFLSAVAKPRYDD
EGNCTFDGKIGIWPFFTRKEPARRRSRNRERGLTVTKPIKVDRTIRSFMI SKVLP AIRACWPREDARKTIW
IQQDNARHTLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRI SRNMDELIENVH
KEYRDYNPNTLNRVFLTLQSCYIEVMR
```

2. Open the TATE website using

http://tate.iplantcollaborative.org/tate_index.html and Click "Run TATE."

Enter the username 'tate' and password 'collaborate' if necessary. Opening the web form starts a TATE 'Session.' The form will look like this:

The first web page can be bookmarked. Instructions for using TATE and links to the most recent TATE version will be found on the page.

3. Choose 'Oryza sativa' as your target database.

Target Database(s):
You can choose more than one database.

Multiple Species Genomes **Partial Genomes**
 Plant only Nonredundant NT *Musa* sp.

Complete Genomes
 Arabidopsis thaliana
 Oryza sativa
 Zea mays

4 Paste your sequence into the Query Sequence window. Include the comment line '>Osmar5'

>Osmar5

```
SKDLTNIQRRGIYQLLLQKSKDGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGIAVNVDSRKHGNS
GRKKVEIDL SVIAAIPLHQRRNIRSLAQALGVPKSTLHRWFKEGLIRRHSNSLKPYLKEANKKERLQWCVS
MLDPHTLPNNPKFIEMENI IHI DEKWFNASKKEKTFYLYPDEEEEPYFTVHNKNAIDKVMFLSAVAKPRYDD
EGNCTFDGKIGIWPFFTRKEPARRRSRNRERGTLVTKPIKVDRTIIRSFMISKVLP AIRACWPREDARKTIW
IQQDNARTHLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRI SRNMDELIENVH
KEYRDYNPNTLNRVFLTLQSCYIEVMR
```

Query Sequence(s)
You can submit more than one sequences at one time. Only fasta format sequence data is accepted, eg:

```
>Copia_RT
WYRVKHKQDGSIDRYKARLVAKGYTQVEGLDYLDTFSPVAKTTTLRLLAL
AASQGWFLHQLDVDNAFLHGTLDDEIYMRLPPGVSSPRPNQVCLLQKSLYGLK
```

```
>Osmar5
SKDLTNIQRRGIYQLLLQKSKDGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGIAVNVDSRKHGNSG
```

4 Enter a Run Name and Run Notes. During a session you can do multiple runs of TATE. By using meaningful Run Names you will be able to easily identify each run. The Run Notes are for you to document your work in TATE.



Run Name and Notes

Enter a Run Name that identifies this run. Each subsequent run will have its own name so you can easily identify it on the output page. The Run Name will appear on a tab.

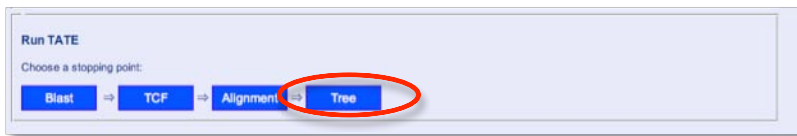
The Notes allow you to keep track of what you did for each run. These will appear in the output window.

Name: Dsmar5 query OS database

Notes

Lecture Example 1. Complete Run.

5 Click Tree. Clicking "Tree" will start the first run of the session. TATE can stop at several points. For today's class let's do the whole thing. Click only once. The run will take a few minutes. The first run of TATE will be with the default parameters. Later we will discuss when you may want to alter the parameters.

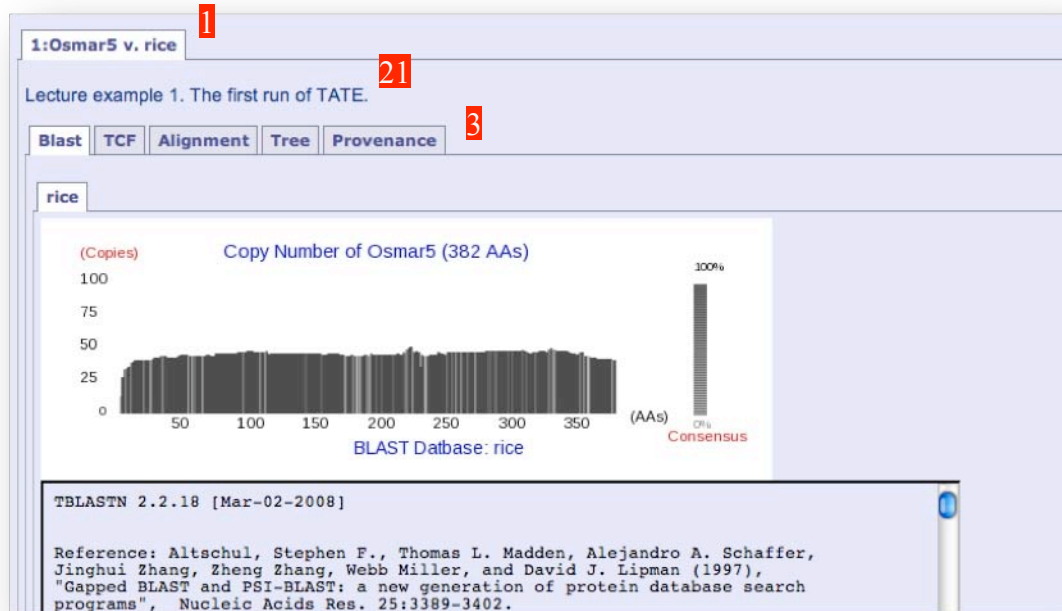


Run TATE

Choose a stopping point:

Blast ⇒ TCF ⇒ Alignment ⇒ **Tree**

6 Output screen of a TATE Run. The output will open with the Blast results presented first.



1. Top Tab is the Run with the Run Name as the label. Each subsequent run will get a new Tab.
2. The Run Notes.
3. The output from each step of TATE is contained in its own tab. Clicking on a Tab will reveal the output. Each Tab will be discussed.
- 3 **Blast Tab.** The results of the Blast are presented here. In the scrollable window you can see the typical Blast output. The figure is a diagrammatic representation of the blast hits and will be described in detail in class.

7. Click on the TCF tab. The second step of TATE, called TCF, will clean up and re-filter the Blast results. Finally TCF re-formats the results into a computer friendly fasta format.

1:Osmar5 v. rice

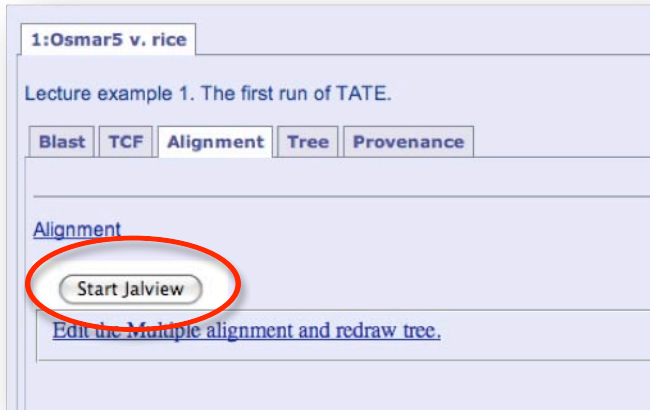
Lecture example 1. The first run of TATE.

Blast TCF Alignment Tree Provenance

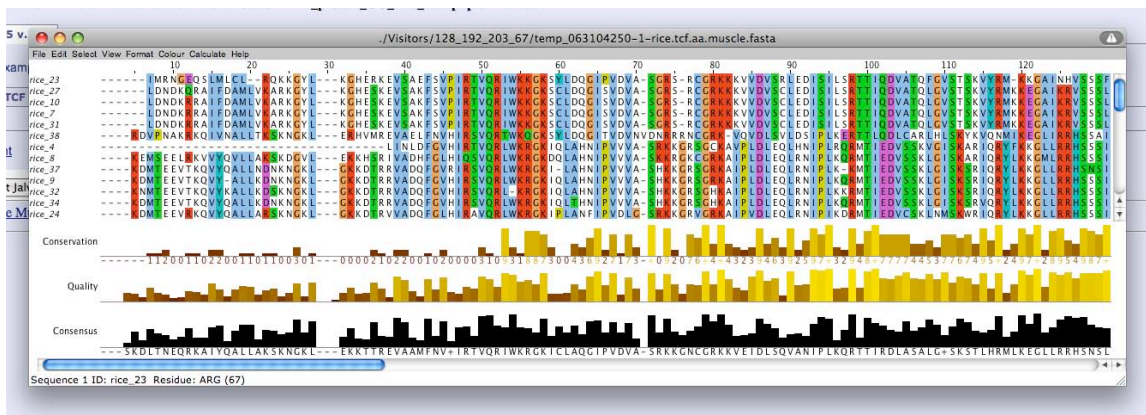
rice

```
>rice 1 Query:Osmar5 Sbjct:jap_ch8 Length:335 Location:(22818136 - 22819239) Direction:
LWTKVKKCQAAGIKVDFTSKKSKKCCRKRVRDRDWSQVATIPLNQRRTTIRNLASALNIPKSVVHRAFKEGILRRHSNTLKPFLKDA
>rice 2 Query:Osmar5 Sbjct:jap_ch8 Length:375 Location:(15579484 - 15580705) Direction:
KYLSDERQODIYEALLAKSINGKIERNATTIVANLNFVRRRVQDLWTKVKKCQAAGIKVDITSKSKKCGWKRVRDRDWSQAATI
>rice 3 Query:Osmar5 Sbjct:jap_ch8 Length:337 Location:(22816987 - 22818096) Direction:
QDLWTKVKKCQAAGIKVDFTSKKSKKCCRKRVRDRDWSQVATIPLNQRRTTIRNLASALNIPKSVVHRAFKEGILRRHSNTLKPFLK
>rice 4 Query:Osmar5 Sbjct:jap_ch5 Length:315 Location:(18054440 - 18055388) Direction:
LINLDFGVHIRTQRLWKRCKIQLAHNIPVVVASRKKGRSGCKAVPLDLEQLHNIPLRQRMTIEDVSSKVGISKARIQRYFKKGI
>rice 5 Query:Osmar5 Sbjct:jap_ch5 Length:381 Location:(4693866 - 4695100) Direction:
SKDLTNIQRRGIYQLLQKSKDKLEKHTTRLVAQEPHVSIRTQRIWKRKAKICHEQGITVNVDSRKHGNSGRKKVEIDLVSIAJ
>rice 6 Query:Osmar5 Sbjct:jap_ch5 Length:380 Location:(27703955 - 27705186) Direction:
ARELTNPQRRSIYELLLTKSLDGYLEKGSTRVVAEVPVNSIRTQRIWKRKALCIAQGVQVNVDSRKRYNCGRKKVEIDLVSIAJ
>rice 7 Query:Osmar5 Sbjct:jap_ch5 Length:377 Location:(4886040 - 4887173) Direction:
LDNDKRRAIFDAMLVKARKGYLKGHEKESKFSVPVIRTQRIWKKGKSCLDQGISVDVASGR-SRCCRKKKVVVDSVCLEDISI
>rice 8 Query:Osmar5 Sbjct:jap_ch6 Length:381 Location:(14378610 - 14379833) Direction:
KEMSELRKVVYQVLLAKSKDGVLEKHSRIVADHFLHIQSVQRLWKRCKDQLAHNIPVVVASRKKRCKCRKAIPLLDLEQLRNI
>rice 9 Query:Osmar5 Sbjct:jap_ch6 Length:388 Location:(18998387 - 18999612) Direction:
KDMTEVTKQVY*ALLKDNKNGKLGKKDTRRVADQFGVHIRSVQRLWKRCKIQLAHNIPVVVASHKKGRSGRKAIPLLDLEQLRNI
>rice 10 Query:Osmar5 Sbjct:jap_ch6 Length:374 Location:(21936403 - 21937516) Direction:
LDNDKRRAIFDAMLVKARKGYLKGHEKESKFSVPVIRTQRIWKKGKSCLDQGISVDVASGR-SRCCRKKKVVVDSVCLEDISI
>rice 11 Query:Osmar5 Sbjct:jap_ch6 Length:385 Location:(13657855 - 13659099) Direction:
NKNLTKIQRQIYAALTGKTNNGILRKRKNATTEVAAMFNVRARVQAIWRRVKQCRAQCIPIIDVISRKKKNCGRKKKEINLTI
>rice 12 Query:Osmar5 Sbjct:japo_ch2 Length:306 Location:(12263220 - 12264223) Direction:
KKLFPDPSVIKDVPLGQRHSIRDLANALHMAKATLFRRLKEGLFRRHTNAIKPTLTEDNMKARVHFCIQMLDS*SIPTDPTFKSI
>rice 13 Query:Osmar5 Sbjct:japo_ch2 Length:333 Location:(17542825 - 17543908) Direction:
LIDEDRQHVLDACFADSENKLRDRTTIVASLEFNKSLVQSIWRKAKHCHAEQVPLDLTSKKE-KCGRHRVVDLSLVPPTIPI
>rice 14 Query:Osmar5 Sbjct:jap_ch4 Length:374 Location:(27341384 - 27342573) Direction:
KRAIYALCLERSDPMKKEGVTKSVATDMGVPRVV*RVWRHGQI---GGGIEAFESKKNKCGRKKLSPNDAIKDVPLRQRRT
>rice 15 Query:Osmar5 Sbjct:jap_ch4 Length:380 Location:(14652518 - 14653753) Direction:
KEMSDLRKLVFPQTLVRSKNGKLGKDDTSIVAAQFGLGIQSVQRLWKRCKIQLANSIPVVVSSLKKGVRGRRKKIPVDLEALRSJ
>rice 16 Query:Osmar5 Sbjct:jap_ch4 Length:380 Location:(24939896 - 24941094) Direction:
LSDKERYAVYIALHAKSKGRLEKDDTTKVAEYFNVGIVIQRIWKHAREQVALGLKVDVDRNRKTRCCGPNKMEIDLKSIATIFI
>rice 17 Query:Osmar5 Sbjct:jap_ch4 Length:380 Location:(15273846 - 15275081) Direction:
KEMSDLRKLVSRLLARSKNGKLGKDDTSIVAAQFGLGIQSVQRLWKRCKIQLANSIPVVVSSLKKGVRGRRKKIPVDLEALRSJ
>rice 18 Query:Osmar5 Sbjct:jap_ch4 Length:381 Location:(19288923 - 19290154) Direction:
SRDLKNNR*AIYARLLEKSMNEKLEKDDTSIVAREPHVSIRTQRIWKKAKVCREQGIANVNVDSRKHGSSGRKKVEIDLVSIAJ
>rice 19 Query:Osmar5 Sbjct:jap_ch1 Length:381 Location:(26759044 - 26760278) Direction:
```

8. Click on the Alignment tab. After the TCF program runs, TATE runs a multiple alignment program called Muscle. The multiple alignment output is in fasta format and not easy to read. To view the alignment click the Jalview button. (If this button is not visible, click the "Tree" tab and then the "Alignment" tab again.)



The multiple alignment will be discussed in detail in class.



9. Click on the "Tree" tab to see the tree.

data were manipulated. TATE helps with this by producing a log file that contains every input and output for the programs, the sequences you provided, and the Run name and notes. The log can be viewed in raw form by clicking the link on the Provenance Tab. TATE also provides an archive of all files produced by the run. This archive can be downloaded and kept as documentation of the tree. The archive will also contain the image files and in future will be used to re-create a TATE session.



2. Details of TATE function and output.

1. Return to the 'Blast' Tab.

To use the Blast output in other applications you must first modify the results to account for the following:

1. Remove low scoring hits.
2. Paste together hits that are separated by introns.

It is likely that there are introns in the transposase gene. If you look at the example of a Blast result the Query matches a region of chromosome 1 from 26,759,044 to 26,759,736. The match then resumes from 26,759,823 to 26,760,278. To use this 'hit' in the multiple alignment, the complete sequence minus introns must be used.

```
>jap_ch1 ref|NC_008394.1|:1-43261740 Oryza sativa (japonica cultivar-group)
      genomic DNA, chromosome 1
      Length = 43261740

Score = 483 bits (1242), Expect(2) = 0.0, Method: Compositional matrix adjust.
Identities = 230/231 (99%), Positives = 230/231 (99%)
Frame = +1

Query: 1      SKDLTNIQRRGIVQLLLQKSKDGKLEKHTTLVAQEFHYSIRTVQRIWKRAKICHEQGIA 60
              SKDLTNIQRRGIVQLLLQKSKDGKLEKHTTLVAQEFHYSI  TVQRIWKRAKICHEQGIA
Sbjct: 26759044 SKDLTNIQRRGIVQLLLQKSKDGKLEKHTTLVAQEFHYSIHTVQRIWKRAKICHEQGIA 26759223

Query: 61      VNVDSRAKHGNSGRKKVEIDL SVIARAIPLHQRRANIRSLAQAALGVPKSTLHAWFKEGLIRAH 120
              VNVDSRAKHGNSGRKKVEIDL SVIARAIPLHQRRANIRSLAQAALGVPKSTLHAWFKEGLIRAH
Sbjct: 26759224 VNVDSRAKHGNSGRKKVEIDL SVIARAIPLHQRRANIRSLAQAALGVPKSTLHAWFKEGLIRAH 26759403

Query: 121     SNSLKPVLKEANKKERLQWCVSM LDPHTLPNNPKF IEMENI IHI DEKWFNASKKEKTFYL 180
              SNSLKPVLKEANKKERLQWCVSM LDPHTLPNNPKF IEMENI IHI DEKWFNASKKEKTFYL
Sbjct: 26759404 SNSLKPVLKEANKKERLQWCVSM LDPHTLPNNPKF IEMENI IHI DEKWFNASKKEKTFYL 26759583

Query: 181     YPDEEFPYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGKIGIWPFTAK 231
              YPDEEFPYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGKIGIWPFTAK
Sbjct: 26759584 YPDEEFPYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGKIGIWPFTAK 26759736

Score = 317 bits (813), Expect(2) = 0.0, Method: Compositional matrix adjust.
Identities = 151/152 (99%), Positives = 152/152 (100%)
Frame = +3

Query: 231     KPPARRASRNREGTLVTKPIKVDRDTIRSFMISKVLPARACWPREDARKTIWIQQDNA 290
              KPPARRASRNREGTLVTKPIKVDRDTIRSFMISKVLPARACWPREDARKTIWIQQDNA
Sbjct: 26759823 KPPARRASRNREGTLVTKPIKVDRDTIRSFMISKVLPARACWPREDARKTIWIQQDNA 26760003

Query: 291     RTHLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRSRNMDL 350
              RTHLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRSRNMDL
Sbjct: 26760003 RTHLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRSRNMDL 26760182

Query: 351     IENVHKEYRDVNPNTLN RVFLTQSCYIEVMR 382
              IENVHKEYRDVNPNTLN RVFLTQSCYIEVMR
Sbjct: 26760183 IENVHKEYRDVNPNTLN RVFLTQSCYIEVMR 26760278
```

3. Remove the Blast alignment formatting and create a fasta file containing only the hits.

TCF does these three things for you. See the TCF Tab for the resulting fasta file that was used for the multiple alignment.

2. Click the 'Alignment' Tab.

Here we will discuss how a multiple alignment is generated and what the output means. We will also discuss when it is appropriate to edit the alignment.

The multiple alignment can be edited to remove large gaps, low similarity columns, and sequences that are much too short. Once an alignment is edited the tree can be re-calculated by TATE. We will discuss how to edit an alignment using Jalview. After editing the alignment you can click the link "Edit the Multiple alignment and redraw tree." A new form will be visible:

This is the first run of the session.

Blast TCF Alignment Tree Provenance

Alignment

Start Jalview

[Edit the Multiple alignment and redraw tree.](#)

Edit the alignment using the following steps.

1. Open the Jalview viewer using the button above. (Skip if the viewer is already open.)
2. Edit the alignment.
 - o Delete sequence: Draw a box around the sequence and use the 'Delete' key.
 - o Delete entire sequence: select name and hit the 'Delete' key.
 - o Insert or remove gap: Shift+click (on Mac and PC) and drag the sequence to the right to open or extend a gap or left to shorten or close.
3. Click back on this browser window
4. Click the 'Paste edited alignment' button below to paste the into the textbox.
5. Click the 'Draw Tree' button. A new browser window will appear.

On a Mac you must use Safari or Firefox for this to work. IE and Firefox will work on Windows.

Paste edited alignment.

Enter a Run Name that identifies this run. Each subsequent run will have its own name so you can easily identify it on the output page. The Run Name will appear on a tab.

The Notes allow you to keep track of what you did for each run. These will appear in the output window.

Name:

Notes

Draw Tree

Clicking on the "Paste edited alignment" will fill in the text box with the edited alignment from Jalview. Enter a Run Name and Notes and click 'Draw Tree'. A new tab will appear with the new tree.

3. Tree Tab. Once you have a tree that can be your stopping point. Or, as discussed below, you can select branches of the tree as TATE input for new searches. We will discuss what your tree means.

3. Troubleshooting TATE.

As you can see TATE is a complex bioinformatics pipeline that simplifies a great deal of phylogenetic tree construction. But there can be times that the simplification works against you. This section will point out the most frequent reasons TATE fails and how to overcome them. If you follow these troubleshooting steps and still get no results then you are left with two possibilities: 1. The starting query is no good or actually has no similar sequences in the databases or 2. You have encountered a bug. In both cases contact Jim Burnette at jburnette@plantbio.uga.edu with a detailed message including the query. Please include the archive of the session (on the Provenance tab) in the e-mail.

For an example of when TATE fails let's walk through two examples.

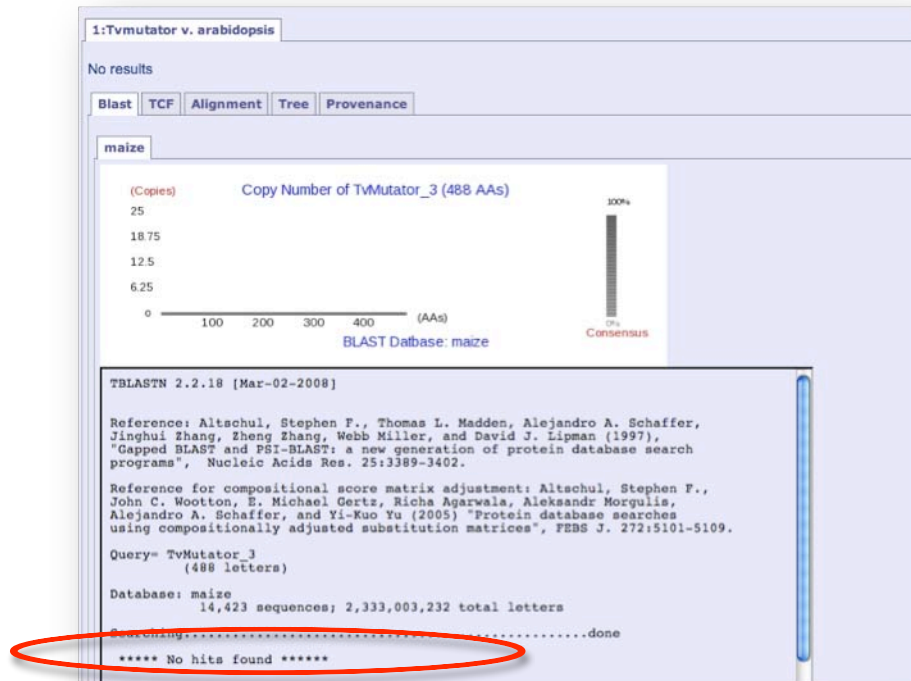
1. Start a new TATE session.

2. Enter this query.

```
>Tvmutator
MQHPPDPDPCEVDQAIGYLEYSRFVIAAELHGNIRFVKDPQKLSIGTGVLIIYHRCEYEGC
PAGFKFIKNFDNYIFKSANLTYIHSGPPPOHKNTPTSGTYRAWIKKFLMNHGSPLNATQE
VNKTLEIPKDHTCIHMTMKQTAINQLKYYLQOKDLMRSLPDISLNQFOYLQNYVNQWEDH
HDDLIIYFDIQGEDTDFDKLVFIYSDNDMISQIHEKPPYHLDSTFKLI IHGFPFYVLAT
KFANTHSIPLCYFIIYPDNSENISFCLSKYFETHTTEPEFIMSDCALNIFNGIRNSFPEC
NIFWCALHVIRALKKNLSKINDEEIRSEVEKFMNILCYYRDCTEEDAAMKYKEHIIDKIQ
DQLEFNQYFTRQWDIHKQOWIAAAGPNELTVVNNVSESLFKKIKYHDFGCVKNQRIDVVFV
KNLLEEVPANFYFRIKNDLLQTFIPSI RIREPRLTDYKTKLKREVQSRLYQVLFVQNO
EANLNPLR
```

3. Enter Run Name and Notes. Choose "Tree" as the stopping point.

In this case there were no Blast results. Let's look at the Blast parameters to understand why.



Failure of Blast to produce hits is usually due to the very stringent default parameter values. Modifying these will help.

- Expectation value (E) (1×10^{-10}), circled in red in Figure 11. The default is restrictive to high scoring hits. Raise the parameter to 1×10^{-5} or higher if the default is too restrictive or if too few hits are being returned. You may need to raise this value if you use a small database or short query sequence.

- Filter query sequence, circled in yellow in Figure 3. For TEs this should be set to 'No,' which is the default. For all other types of searches, you may want to change this to mask low complexity in the query sequence.

The screenshot shows the 'tblastn' command-line interface. The following parameters are visible:

- tblastn
- Modify Blast Parameters
- Expectation value (E): 1e-10
- Show descriptions for how many hits? 100
- Show how many alignments? 500
- Filter query sequence (DUST with blastn, SEG with others): Yes No
- If you need other parameters please input here:

Figure 1 Blast parameters

The following example will demonstrate what happens if there are Blast results but no TCF output.

1. Start a TATE session and enter this query:

```
>Osat_PIF1
MAGGTGSGGAQRGKRREETEAVREREVAELTEGHGDGESSWRREAAALHRSTMRRGSRRRKAATGRQMGKSGTRRSRRC
SRCVAAAAGVKANGDRRRRKAAGRUVVWRGGGSGGGGEGEKMGGRACPCGSEANGGRGVAESSEVRPEATTGVAEDNGSG
PGGGDRGRGGRQRPAGAPATWASGARTAPIFGGEMGEKVEEGEDVGWTT SARARERGGARGKRRRGVAQAAMAVGVGR
REDLQDLESTAILLGFDDLQEKYWRGMNSLIRKSKDEEDEEII MFWLPALYLLT SNGGIEKVRHTSSQYSEEKLRNI
LEGHEKNCLVAFRMEPNIFRAIVTYLRTEHLLRDTRGITVEEKLGHFLYMISHNASYEDLQHEFHHSGETIHRHIKAVF
KVIPSLTYRFIKQTRRVETHWKISTDQLFFPYFQNC LGAIDGTHVPITISQDLQAPYRNKGTLSQNVMLVCDFDLNFL
FIPSGWEGSATDARVLR SAMLKGFNVPQ GKYYLVDGGYANTPSFLAPYRGVRYHLKEFGRGQQRPRNYKELFNHRHAIL
RNHIERAIGVLKKRFPILKVGTHHR IKNOVKIPVATVVFHNLIRMLNGDEGWLNHQGSNISPEQFIDVPEGDDEYSNDV
MSLNSQVDDGNAQQCLEEGHFHDCKRLYTTQVVKCKRLYTTLHNCVNLETKFCLGWKISISTPSVTS LTLDNPMGIVV
LKDMKSLVRASVRLNRQWPHDDFDARDLRNYLWTL SGIENLKFYCGRRKLTIQNSLQWCPKFFNLVSLTLGHWC LHGNF
YTLIVFLQNSPRLEKLT LILGNDHWKTFEASIGDKLDEFSFTCEHLSMVKVRCSEDDPMLWQKRFGRLPKLWWRGRKLD
DGVAELTKSGLGLSPLAAAIFPSSDLRLSPDRRHQLVKLRVAMHSFNAPLRAMPHAL TNPNEKRGTQLGSTVFTWC GS
LLRKP DGCTPVAISSWTQILCLWPLKSHRDF TSSSLPSDLIQWTGLLGPSTGLAHC
```

2. Select the *Oryza sativa* database.

3. Enter Run Name and Notes and click Tree. This query will produce Blasts results, but not a tree. When this happens it usually means that no Blast hit past the TCF stage of TATE. We will use this example to explore TCF and the parameters it takes. We will discuss this in detail in class.

2. TCF:
Re-formatting of the Blast Results.

[Modify TCF Parameters](#)

These parameters affect the re-formatting of the Blast results for subsequent steps. Any introns are removed and split hits are "spliced" together.

Minimal BLAST match: 50 **1**

Minimal Length of Intron: 10 **2**

Minimal Matched percentage of Query: 0.8 **3**

Maximal Number in output: 100

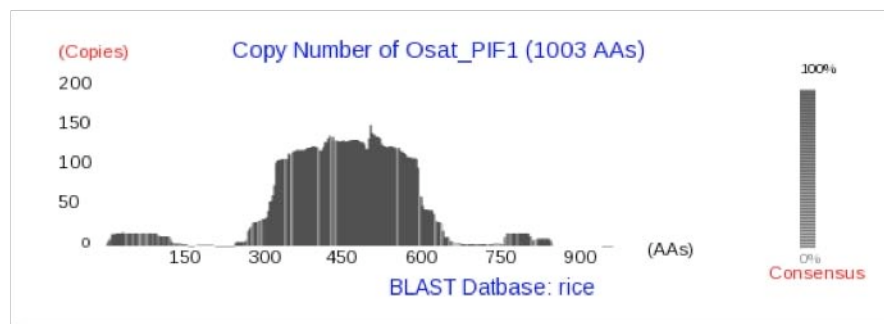
Remove Duplicate hits? Yes No

Understanding the parameters of TCF are the most critical to a TATE run. If a query fails to produce a tree it is usually due to the stringency of the default TCF parameters. Refer to the red numbers in Figure 12.

1. Minimal Blast Match: In a hit alignment, there must be at least this number of matches between query and sequence. Look at the Blast result to see how many matches there are in the alignments. Lower this number if necessary.

2. **Minimal Length of Intron:** A gap within a Blast hit may represent an indel or an intron. If it is an indel the gap should be left in the hit, but if it is an intron, the gap should be removed. Indels in Blast hits are usually short so TATE sets the minimum length of an intron to be removed at 10. Any gap less than 10 is left in the hit. This is arbitrary and can be modified. Introns less than 10 definitely exist and indels of greater than 10 can occur.
3. **Minimal matched percentage of query:** After TCF removes introns in hits and splices together split hits it re-calculates the percent match to the query. If the hit is greater than the supplied value it will be kept. This value may need to be lowered if the hits to the query are short relative to the query size.

For example, in the figure below there are many hits, but few if any are greater than 80% the length of the query. Most are below 50%. In this case the best results would be obtained by limiting the Blast query to the region between 300 and 600. You may need to also lower the Minimal Matched Percentage and the Minimal Blast Match.



If TATE produces Blast and TCF output, but no alignment and/or tree the most likely explanation is a weird or unexpected character in the fasta files. If you have problems at these steps, contact Jim Burnette and include the archive of the run (found on the provenance tab) in the e-mail.

Query sequences for DNA TE families.

Use these queries for exploring the TE content of genomes.

```
>Copia_RT
WVYRVKHKQDGSIDRYKARLVAKGYTQVEGLDYLDTFSPVAKTTTLRLLLAL
AASQGWFLHQLDVDNAFLHGTLDEEIYMRLPPGVSSPRPNQVCLLOKSLYGLK
```

```
>gypsy_RT
RLVINYKPLNQALCWIRYPIPNKKDLLARLHDAKVFSKFDKSGFWQIQLOEK
DRYKTAFTVPPFGQYEWNVMPFGLKNAPSEFQRIMNEIFNPYSKFTIVYIDDVL
IFSQTLDQHFHKLNTFISVIKRNGLAVSKTKVSLFQTKIRFLGH
```

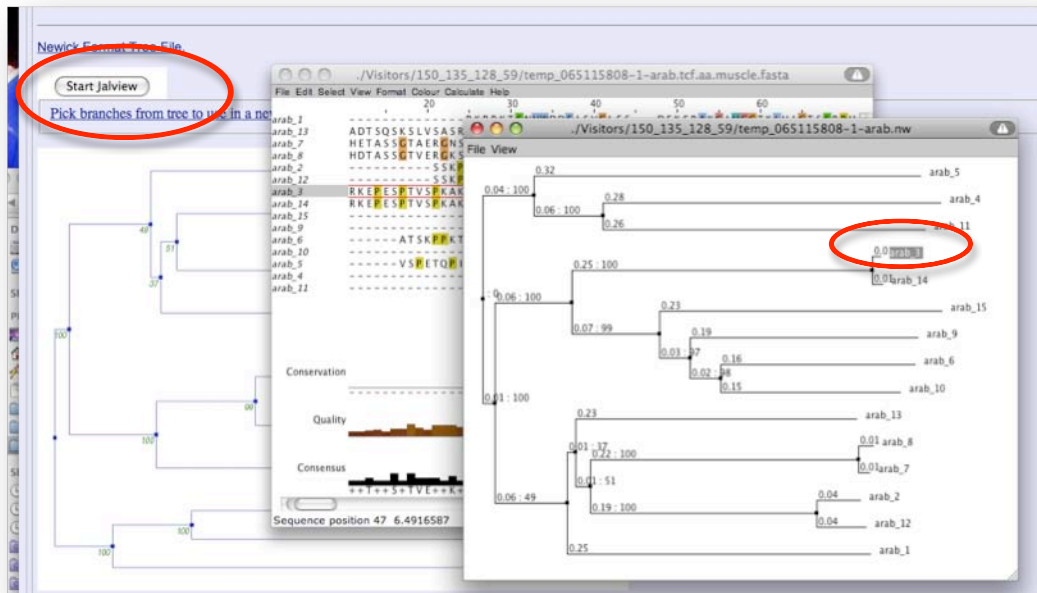
These queries should work for all genomes, but may not pull all family members out from the genome. This tutorial will show you how to pull out members closely related to the query and then re-query to get many more family members.

1. Start a TATE session.
2. Choose a query and a database. Enter Run Name and Notes. Now it is critical that you start good documentation.
3. Choose 'Tree' as the stopping point.
4. Inspect the output of TATE from Blast to the Tree. Did you get a tree? If not, can you troubleshoot the problem? Repeat the TATE search with modified parameters if necessary.

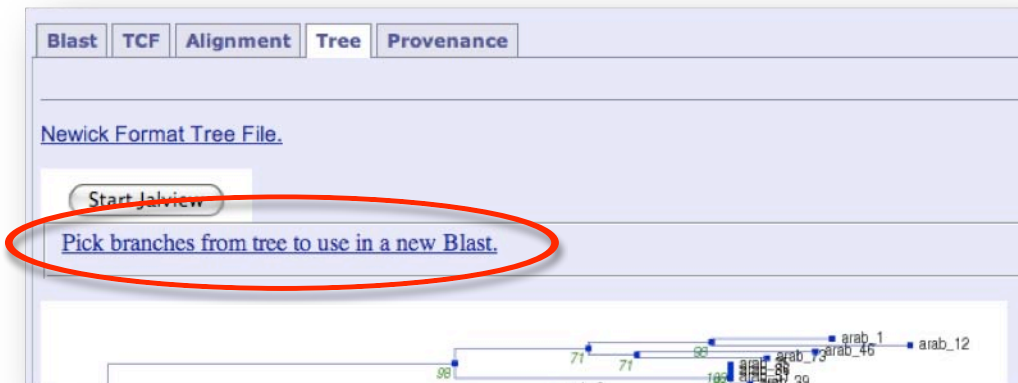
The tree may not have many branches on it. This is because the Blast parameters are set very stringently. You could increase the expect value and repeat the search to get more hits. The other option is to pick a hit from the tree and use that as a Blast query. This is the better option because you are now using a TE family member from the species you choose to query. Here is how TATE helps you do this.

5. Open the Tree in Jalview by clicking the 'Start Jalview' button in the 'Tree' tab.

6. In the Tree window of Jalview you can select a branch by clicking on the sequence name.



7. Click the "Pick branches from tree to use in a new Blast" to open a new form.



8. Choose the Database and click the "Paste Selected Sequences" button.



9. Enter a Run Name and Notes and click "Tree" as the stopping point.

10. TATE will produce a new tree in a second Tab.

