

The web links in this document may be broken. Please do not click on them.


PBIO/BIOL3250L: The Dynamic Genome
Spring 2009 

Table of Contents

Syllabus

Grading Policy

Chapter 1: Transposable Element Background and Experiment 1	1-28
A. Background to TEs	1-15
B. Background to mPing Experiment	16-21
C. Protocols for mPing Experiment	23-27
D. Discussion points	29
Chapter 2: Introduction to NCBI Website and Bioinformatics	31-47
A. PubMed	31-34
B. Blast	35-47
Chapter 3: Introduction to using yeast in lab	49-52
A. Background	49-51
B. Yeast Growth Curve	66-80
Chapter 4: TE Families, Phylogenetics, and Experiment 2	53-88
A. Background	53-62
B. What are phylogenetic trees?	63-69
C. Using TATE to construct trees	69-73
D. Background Experiment 2	75-78
E. Background on yeast transposition assay	79-83
F. Yeast transformation protocol	83-85
E. Excision assay protocol	86-88
Chapter 5: Introduction to Class 1 elements	89-106
A. Background	89-93
B. Finding LTR Retros: A bioinformatics experiment	94-106
Chapter 6: Analyzing excision events	109-124
A. Background	109
B. Protocol for footprint analysis	110-114
C. Analyzing DNA sequence	115-124
Chapter 7: Epigenetic Silencing of a TE in maize	124-137

A. Background to epigenetics	124-134	
B. Experimental Protocol	135-137	
Chapter 8: Distinguishing Wild from Domesticated Transposons	138-156	
A. Background	138-142	
B. Bioinformatics Experiment	143-156	
Chapter 9: Project		157-174
A. Using the Maize Genome Browser	157-162	
B. Finding captured genes	163-165	
C. Designing primers	167-174	
D. RNA Extraction	175-176	

Syllabus Spring 2009

Syllabus

Grading

Data

	Lecture Topic	Lab Work	Due dates	Downloads
Thursday, Jan 8	<ul style="list-style-type: none"> • Course mechanics • Course Intro (McClintock, TEs) • Making of Fittest 	<ul style="list-style-type: none"> • Safety • Pipetter skills • Notebook format 		<ul style="list-style-type: none"> • GFP 2008 Nobel Prize
Tuesday, Jan 13	Intro DNA Elements and Exp. 1	Experiment 1 <ul style="list-style-type: none"> • Make genomic DNA • Gel of DNA 	Notebook check	<ul style="list-style-type: none"> • Handout Pgs 1-27
Thursday, Jan 15	<ul style="list-style-type: none"> • Intro Exp. I • Making of Fittest Discussion 	Experiment 1 <ul style="list-style-type: none"> • View seedlings • mPing PCR 		<ul style="list-style-type: none"> • Sean Carrol YouTube video
Tuesday, Jan 20	<ul style="list-style-type: none"> • Bioinformatics intro (PubMed) 	Experiment 1 <ul style="list-style-type: none"> • Gel of PCR • Experiment 1 Data analysis 	Quiz 1	

<p>Thursday, Jan 22</p>	<ul style="list-style-type: none"> • Discussion with Sean Carroll • Bioinformatics (PubMed) 	<p>None</p>	<p>Quiz 2 take home due</p>	<p>Quiz 1 answers PubMed</p>
<p>Tuesday, Jan 27</p>	<ul style="list-style-type: none"> • Intro to Yeast manipulation 	<p>Yeast Skills</p> <ul style="list-style-type: none"> • Sterile Technique • Serial Dilution for Viable Counts 		
<p>Thursday, Jan 29</p>	<ul style="list-style-type: none"> • Intro to TE Families • Phylogenetics (Blast), <p>Greenhouse tour/ Plant arabidopsis seedlings</p>	<p>Yeast skills data analysis</p>	<p>Fri, Jan 30 Write up 1</p>	<p>Homework 1 Questions</p>
<p>Tuesday, Feb 3</p>	<ul style="list-style-type: none"> • Phylogenetics (TATE) • Intro to Exp 2 	<p>None</p>		<p>Chapter 4 Phylogenetics PPT</p>
<p>Thursday, Feb 5</p>	<p>Details on Transformation</p>	<p>Experiment 2 Yeast Transformation</p>		

<p>Tuesday, Feb 10</p>	<ul style="list-style-type: none"> • Review Exp 2 • Dr Hancock's presentation • Discuss step 2 of Exp 2 	<p>Experiment 2</p> <p>Start liquid cultures for excision assay</p>	<p>Wed, Feb 11.</p> <p>Homework 1</p> <p>Quiz 3</p>	<ul style="list-style-type: none"> • Ping Presentation • Exp 2 Presentation
<p>Thursday, Feb 12</p>	<p>Details on lab work (all class in lab)</p>	<p>Experiment 2 Plate for excision</p>		
<p>Tuesday, Feb 17</p>	<p>Class cancelled</p>	<p>Class cancelled</p>		
<p>Thursday, Feb 19</p>	<ul style="list-style-type: none"> • Intro Class 1 elements 			<p>Chapter 5</p>
<p>Tuesday, Feb 24</p>	<ul style="list-style-type: none"> • Dr. Shelley Schuster Bioindustry • Footprint analysis 	<ul style="list-style-type: none"> • Experiment 2 read plates • Footprint Assay <ul style="list-style-type: none"> ◦ Yeast prep ◦ PCR ◦ Pour gel for Thursday 		
<p>Thursday, Feb 26</p>	<p>Discuss Darwin's Surprise</p>	<p>Footprint Assay</p> <ul style="list-style-type: none"> • run gel • HpaI digest (overnight) • Cut bands, gel 		

		purify		
Tuesday, Mar 3	Review Mid-term	Footprint Assay <ul style="list-style-type: none"> • HpaI digest on gel • analyze sequence 	Handout	
Thursday, Mar 5	Mid-Term Exam			Report 2 requirements
Tuesday, Mar 10	Spring Break			
Thursday, Mar 12	Spring Break			
Tuesday, Mar 17	Talk about projects Host control of TEs	Dr. Damon Lisch visits		Lisch Lecture 1
Thursday, Mar 19	Host control of TEs	Dr. Damon Lisch visits		Lisch Lecture 2 Handout <ul style="list-style-type: none"> • Screen Shots PC • Screen Shots Mac • CoGe Tutorial Movie

				<ul style="list-style-type: none"> CoGe Tutorial
Tuesday, Mar 24	Project-Preliminary research		Mon, Mar 23 Homework 2 due Quiz 4	Homework 2
Thursday, Mar 26	Project-Preliminary research		Quiz 5	
Tuesday, Mar 31	Proposals submitted		Wed, Apr 1 Proposals Due	
Thursday, Apr 2			Fri., April 3 Exp 2 write up.	
Tuesday, Apr 7			Quiz 6	
Thursday, Apr 9			Fri., April 10 Exp 2 write up.	
Tuesday, Apr 14				
Thursday			Final Proposal due	

Thursday, Apr 16			due, Thurs. Apr. 16, noon.	
Tuesday, Apr 21				
Thursday, Apr 23			Quiz 7	
Tuesday, Apr 28	Final Presentation	Course Evaluation		
Thursday, May 7	Final Exam			

BIO/PBIO 3250L The Dynamic Genome
Spring 2009

Dr Susan Wessler and Dr Jim Burnette
Eleanor Kuntz, TA

Tuesday and Thursday 12:30-4:30

Course website: <http://www.dynamicgenome.org/classes/syllabusS09.html>

User: dynamicgenome

Password: tesjump

	Dr. Susan Wessler	Dr. Jim Burnette	Eleanor Kuntz
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	ekuntz@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 this course. You are expected to be prepared for each day, participate in all discussions, and ask a lot of questions. Fifteen percent of your grade is based on class participation.

Restrict cell phone/texting use and personal web browsing/e-mail to breaks. Cell phones should not be on your desk or lab bench at any other time. Do not use class time to work on assignments for other classes.

In the computer lab, place your backpacks in the cubby-holes.

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. Eating is permitted in the computer lab (room 1503A) but not the wet lab room 1606.

The syllabus and other handouts can be found on the website link above.

Assignment due dates

Assignment	Date
Notebook check	Tues., Jan. 13
Quiz 1	Tues., Jan. 20
Quiz 2: Questions for Dr. Carroll	Thurs., Jan. 22
Homework 1 (PubMed, Blast)	Mon., Jan. 26
Expt 1 write-up	Fri., Jan. 30
Homework 2 (Yeast viable counts)	Mon., Feb 2
Quiz 3	Thurs., Feb. 5
Homework 3 (Phylogenetics)	Mon., Feb. 9
Quiz 4	Thurs., Feb. 19
Homework 4 (Class 1 elements)	Mon., Mar. 2
Mid-Term	Thurs., Mar. 5
Project Proposal write-up	Mon., Mar. 23
Quiz 5	Thurs., Apr. 2
Expt 2& 3 write-up	Fri., Mar 27
Quiz 6	Thurs., Apr. 9
Homework 5 (TBD)	Wed., Apr. 15
Quiz 7	Thurs., Apr. 23
Final presentation	Tues., Apr. 28
Final Exam	Thurs., May 7

Notebook Checks: The first one is announced. Other checks will occur as necessary.

Quizzes: 30 min. or less and will cover material recently presented in class.

Homework assignments will be extensions of in-class lectures and exercises.

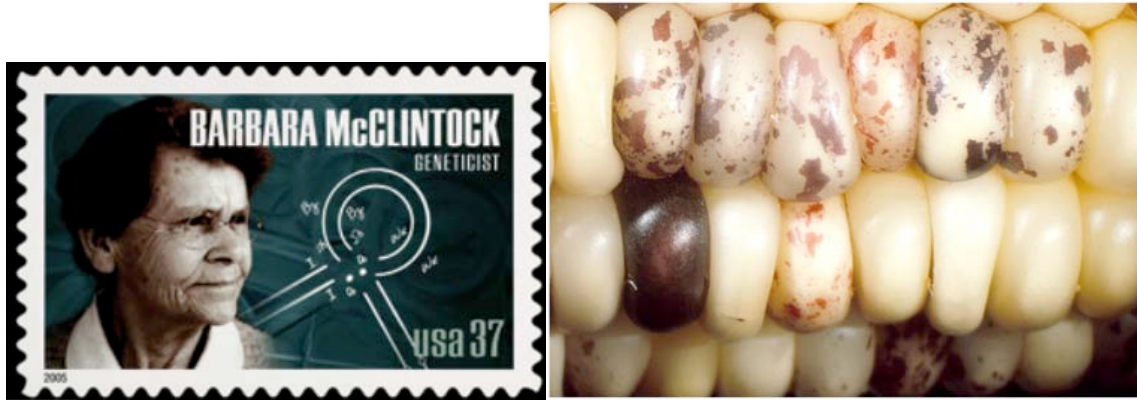
Experiment write-ups: Based on journal article format. Details will be provided for each write-up.

Grading Percentages: The final grade will be calculated using the following break down. Letter grades will be assigned using the standard plus/minus system (A=95-100, A-=90-94, B+=86-89, B=83-85, B-=80-82, C+=76-79, C=73-75, C-=70-72, D=60-69, F<60).

Notebook check average	5%
Quiz average	10%
Homework average	15%
Mid-Term	10%
Write-up average	25%
Final Presentation	10%
Final Exam	10%
Participation	<u>15%</u>
	100%

Chapter 1: Transposable Element Background and Introduction to Experiment 1

1.1. The Discovery of Transposable Elements



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 (recall - 2500 Mb) 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 but fascinating phenotypes. One such phenotype was a colorless kernel containing pigmented spots. This is summarized in **Figure 1b**.

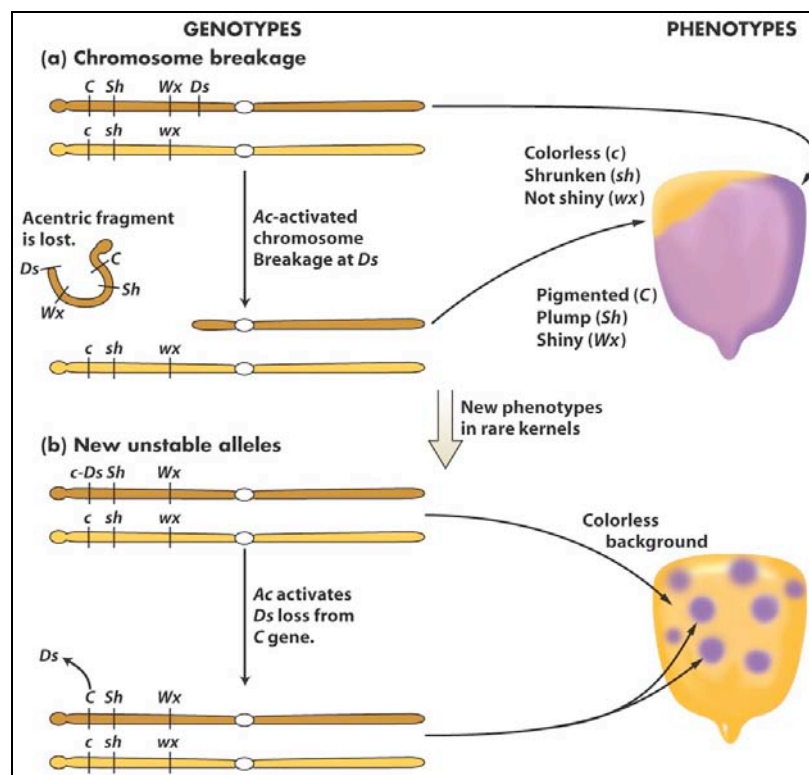


Figure 1. New phenotypes in corn are produced through the movement of the *Ds* transposable element on chromosome 9. (a) A chromosome fragment is lost through breakage at the *Ds* locus. Recessive alleles on the homologous chromosome are expressed, producing the colorless sector in the kernel. (b) Insertion 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 have learned about 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:

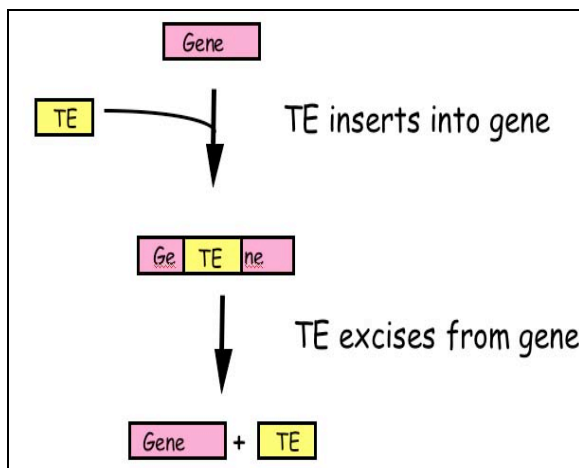
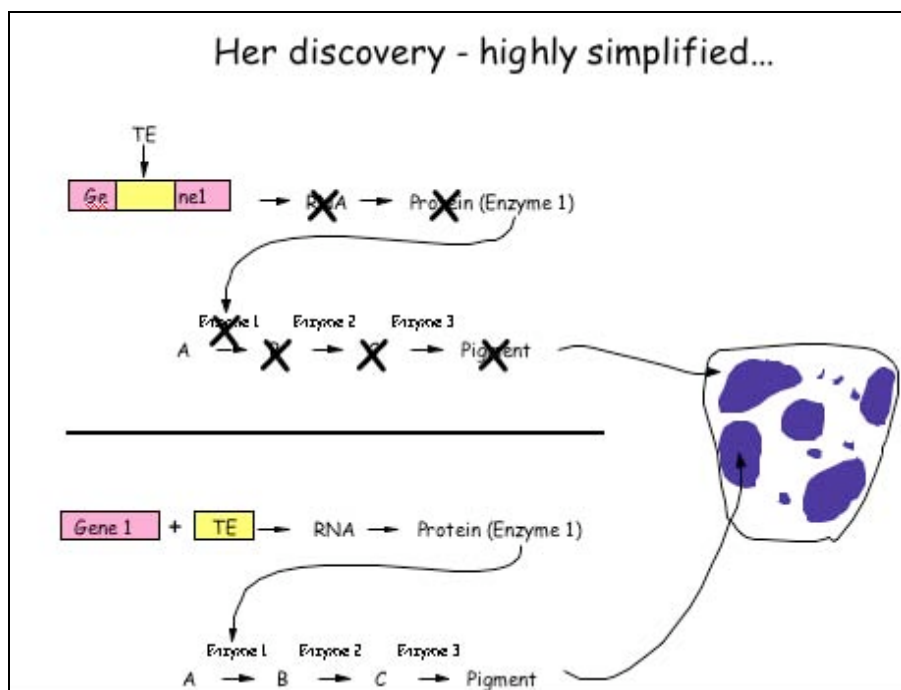


Figure 2: McClintock hypothesized that TEs were a source of "reversible" mutation. Their ability to transpose allowed them to excise from mutant genes leading to phenotypic



1.2 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.

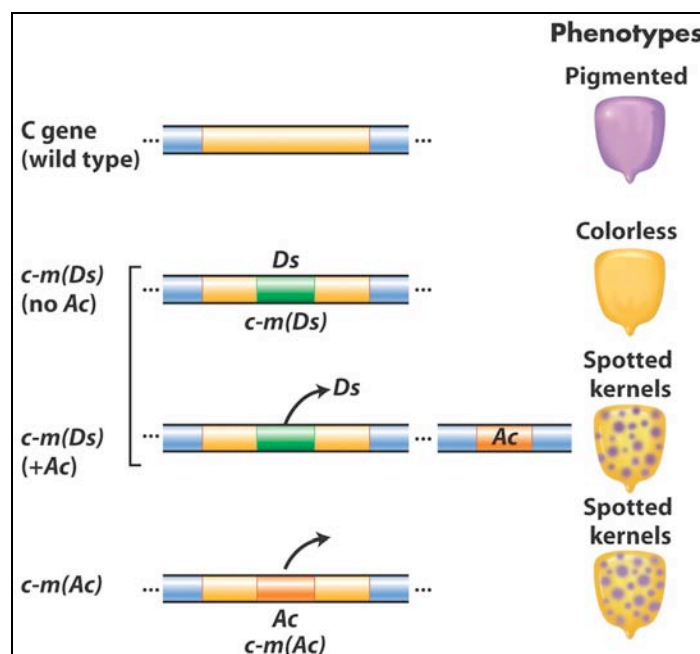


Figure 3 Summary of the main effects of transposable elements in corn. *Ac* and *Ds* are used as examples, acting on the *C* gene controlling pigment. In maize (but not many other organisms), normal alleles are capitalized and mutant alleles are written in lower case. In addition, McClintock designated 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. We will return to Barbara McClintock often during this course.

1.3 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.

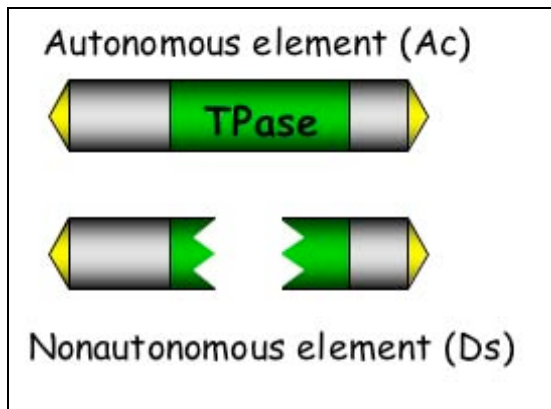


Figure 4: Molecular structure of Ac and Ds.

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. Figure 4 shows how this protein catalyzes the movement of Ac and Ds.

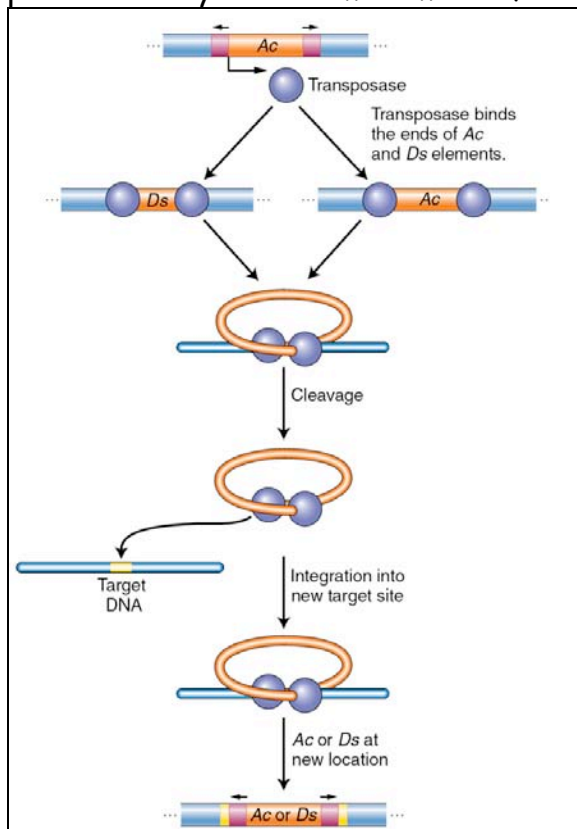


Figure 5 Activator transposase catalyzes excision and integration.

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 Figure 6.

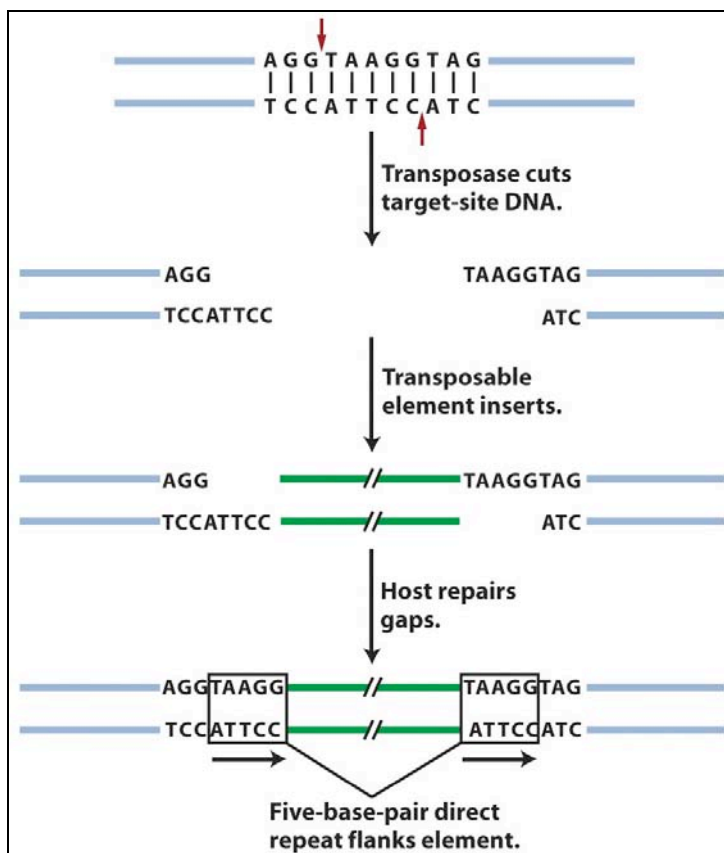


Figure 6: An inserted element is flanked by a short repeat. A short sequence of DNA is duplicated at the transposon insertion site. The recipient DNA is cleaved at staggered sites (a 5-bp staggered cut is shown), leading to the production of two copies of the five-base-pair sequence flanking the inserted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

1.4. 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, fruit fly (*Drosophila*), earthworm, dog, mouse, rat, chimp; plants - including *Arabidopsis thaliana*, rice, maize (corn) 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.

The elements you will be analyzing in experiment 1 are the Ping and mPing (for miniature Ping) elements - which were first identified by computational analysis of the rice genomic sequence (see page 14 below for how this was done). The figure below shows that Ping is the larger coding element like Ac. Unlike Ac Ping contains 2 genes (ORF1 + T_{pase}).

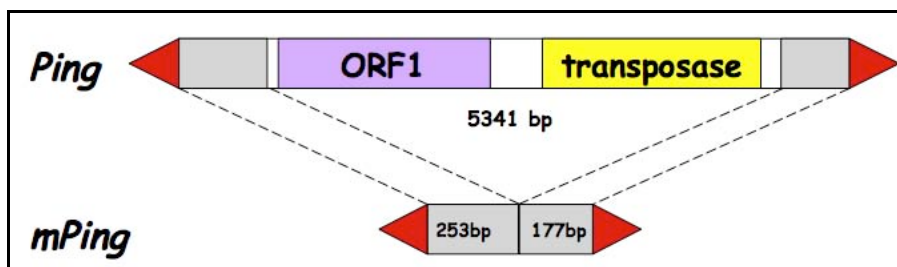


Figure 7: Ping encodes two genes: the transposase (T_{pase}) and ORF1 (open reading frame) (function unknown at this time). The red arrows are the terminal inverted repeats (14bp). mPing is 253bp + 177bp long.

As you can see, like Ds which is derived from Ac by a large deletion, mPing is derived from Ping by a large deletion. Our hypothesis is that Ping encodes a protein that binds to the ends of mPing and catalyzes its transposition.

So, this should be a snap, right? Let's just study an mPing element that is inserted into a rice gene and monitor its movement in the same way as McClintock did with spotted kernels (rice grains in this case). Well, unfortunately, we can't do that - because - like most TEs in the genome, mPing is not inserted into a rice gene - but rather - it is inserted between rice genes!

Let's step back a moment and go through the logic of the design for experiment #1 where you will be testing whether the Ping transposase can catalyze the movement of mPing.

1.5. Digression - how can organisms survive with so many TEs? Where are TEs located in the genome? (This section provides your first link between transposable elements and your assigned book "Making of the Fittest")

At this point we need to go up to 30,000 feet in order to understand a larger concept: the connections between TEs, evolution and natural selection. In short, the distribution of TEs in most genomes is due to the action of natural selection — the foundation for all modern biology. Most of you probably understand these concepts already. Just in case, here is a brief review...

There are three kinds of selection that will need to understand:

- *Negative selection
- *Neutral selection
- *Positive selection

It is important first to know something about natural selection itself. Here's a slightly edited version of its definition in Wikipedia: ". . . In the context of evolution, certain traits or alleles of a species may be subject to selection. Under selection, individuals with advantageous or 'adaptive' traits tend to be more successful than their peers reproductively—meaning they contribute more offspring to the succeeding generation than others do. When these traits have a genetic basis, selection can increase the prevalence of those traits, because offspring will inherit those traits from their parents."

Negative selection is the elimination of a deleterious trait from the population by natural selection. It is also called "purifying selection." In the context of TEs, insertions into genes are deleterious and, as such, are eliminated from the population. The word *elimination* in this case means that an individual with the TE insertion will either not be viable or will not be able to reproduce.

Neutral selection describes changes in the gene pool of a species that are a result of accumulated random neutral changes that do not convey any particular advantage to a species. Accordingly, neutral selection does not depend upon adaptation, fitness, and natural selection.

Positive selection occurs when a certain allele has a greater fitness than others, resulting in an increase in frequency of that allele. This process can continue until the entire population shares the fitter phenotype, then the allele is said to be

"fixed" in the population. An example of this is a TE insertion that affects a gene in some positive way that makes the organism more adaptive in a particular environment. Such a change would be incredibly rare, though, because there are thousands of genes in a genome where a TE can insert and most insertions in a gene are harmful. Think of a population where the climate has changed and become much drier. Increasing the expression of one particular gene in the genome might increase drought tolerance and allow an organism with such a "mutation" to survive. For a TE to insert into just that gene, in the right place so that it increases the expression of the gene, is extremely unlikely. However, when we think of probabilities it is important to keep in mind that there are lots of TEs in a genome, that there can be many individuals in a population and finally - evolution occurs over very long time periods - that's why it's called evolution, not revolution! This concept will be described in greater detail in "The Making of the Fittest".

So what does all this have to do with transposable elements?

Transposable elements can insert into all regions of the genome - in genes and between genes. However, if we look at an entire genome, we usually find most of the TEs between genes and in noncoding regions of a gene (e.g like introns). This is because insertions into genes have fallen victim to negative selection. In contrast TEs between genes remain for generations, hundreds of generations, because they are not harmful. Rather, they are usually neutral and may even be beneficial.

Most of the TEs in the genome are INACTIVE

This leads to a second point you need to remember: The vast majority of transposable elements in a genome are inactive (they can't move anymore). TEs can be inactivated in one of at least two ways—through mutation or through what is called "epigenetic silencing."

The mutation part is easier to understand. All DNA is susceptible to mutation - usually base pair changes or deletions. This happens (very rarely) when there is an error during replication and the wrong base is inserted - for example a G is put opposite T (instead of an A). This change could alter an amino acid in a protein. Mutation can also happen by "free radicals" - chemicals that accumulate in our cells and can damage our DNA. Finally, mutagens in our environment - like UV light or cigarette smoke - can damage our DNA.

There are dramatically different consequences of a mutation in a gene vs. in a TE. Stated simply, mutation in a gene is usually eliminated from the population by natural selection (negative selection), whereas mutation in a TE will be neutral and, as such, will persist in the population. Thus, unlike genes, TEs will accumulate mutations and become inactive. (NOTE - TES AND GENES SUSTAIN MUTATIONS AT THE SAME FREQUENCY. HOWEVER, IF YOU STUDY AN ORGANISMS GENOME, MOST OF THE GENES WILL BE ACTIVE WHILE MOST OF THE TES WILL HAVE SUSTAINED INACTIVATING MUTATIONS)

Epigenetic regulation of TEs (to be discussed later in the course in grueling detail). For now, suffice it to say that eukaryotic chromosomes exist in the nucleus as chromatin - an equal mixture of DNA and protein. The basic unit of chromatin is the nucleosome - about 180bp of DNA wound around a core of histone protein (shown as a ball in figure 8 below). Chromatin can be loosely organized (open chromatin, called euchromatin) or highly condensed (where nucleosomes are tightly packed, called heterochromatin). You can see both of these chromatin types in Figure 8. In most plant genomes such as maize, transposable elements are frequently clustered and associated with condensed chromatin. Genes in condensed chromatin cannot be expressed and are inactive. This is the fate of the vast majority of the TEs in a genome.

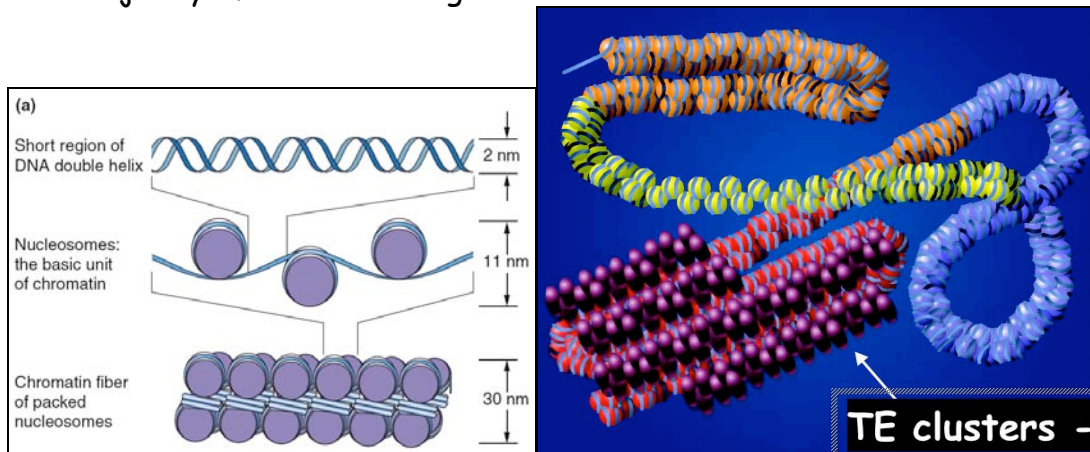


Figure 8. The nucleosome in decondensed and condensed chromatin.

(b) Chromatin structure varies along the length of a chromosome. The least-condensed chromatin (euchromatin) is shown in yellow, regions of intermediate condensation are in orange and blue, and heterochromatin coated with special proteins (purple) is in red.

So let's say it again: Most TEs in the genome of plants and animals are rendered inactive by mutation or by epigenetic silencing.

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



Let's re-state the problem we face here. *We need to create an experimental system that mimics the one used by McClintock with TEs inserted into pigment genes and expressed in the kernel.* For this experiment, we need to use a visual assay to test for movement of the rice mPing element 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 blue 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. The importance of the GFP reporter gene to modern science was evident when the 3 scientists responsible for its discovery and adaptation to the lab were awarded the 2008 Nobel Prize in Chemistry. You can learn more about this at this site:

http://nobelprize.org/nobel_prizes/chemistry/laureates/2008/press.html

In our experiments, the GFP reporter gene will substitute for the maize pigment gene. The mPing element has been engineered into the GFP gene so that it cannot produce fluorescent protein. If mPing excises the GFP gene will be able to function again.

1.7 *Arabidopsis thaliana*



In your previous biology classes you have certainly 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*).

1.8. *Agrobacterium tumefaciens*: introducing foreign DNA into 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 plants (Mary Dell Chilton, a postdoctoral associate at the University of Washington, and two other researchers

working in Germany named Marc Van Montagu and Jeff Schell). 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) that contains a fragment of DNA (called T-DNA). Proteins encoded by the Ti-plasmid facilitate the transfer of the T-DNA into plant cells and ultimately, insertion 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 DNA by Agrobacterium.

As shown in Figure 9, 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.

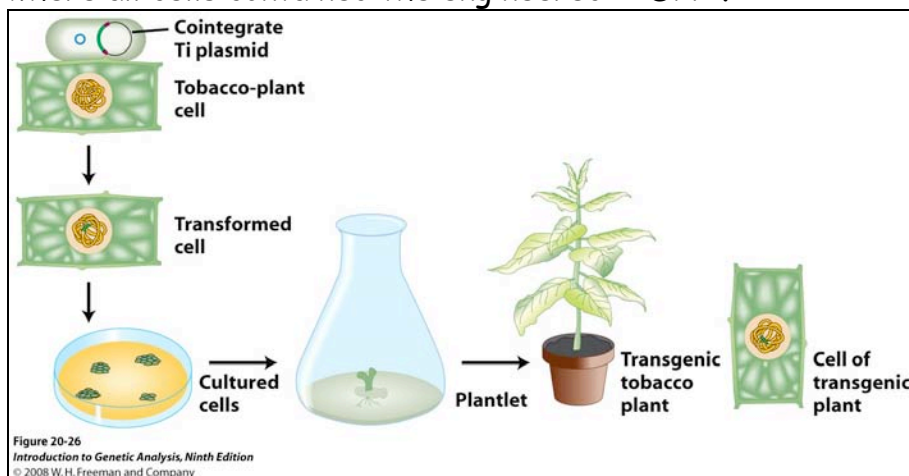


Figure 9:
Schematic of how
Agrobacterium
has been
exploited to
deliver foreign
DNA into plant
chromosomes.

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.

1.9 - Back to Ping and mPing: how they were discovered.

Isolating Ping and mPing from rice (refer back to figure 7, page 7): Geneticists had never isolated an active TE from rice like the Ac and Ds elements discovered by Barbara McClintock in maize. The logic used to isolate the first active rice TEs, Ping and mPing, is described.

Rice (*Oryza sativa*) has the smallest genome of all cereal grasses at 450 million base pairs (Mb). By contrast, the maize 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. As discussed above, most of the TEs in a genome are inactive due to mutation. 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.

To find an active TE in rice, researchers compared the publicly available genome sequence of rice to itself. This sounds confusing, but here is what it means: Scientists first used computers to compare the genome sequence of *Oryza sativa* (domesticated rice) to itself and identified several sequences that were repeated (called families or repeats). The repeat families were then analyzed (by computer again) to identify families that contained identical or almost identical sequences. The researchers reasoned that actively moving TEs should be represented by several identical or nearly identical copies in a genome. The reason for this is that when an element moves, an identical copy inserts elsewhere in the genome. Over millions of years these originally identical copies accumulate mutations (more on this later) and start looking different. By analyzing the genome this way, the researchers found a 430bp sequence with 50 nearly identical copies scattered across the 12 rice chromosomes. They named it "*mPing*" for "*miniature Ping*."

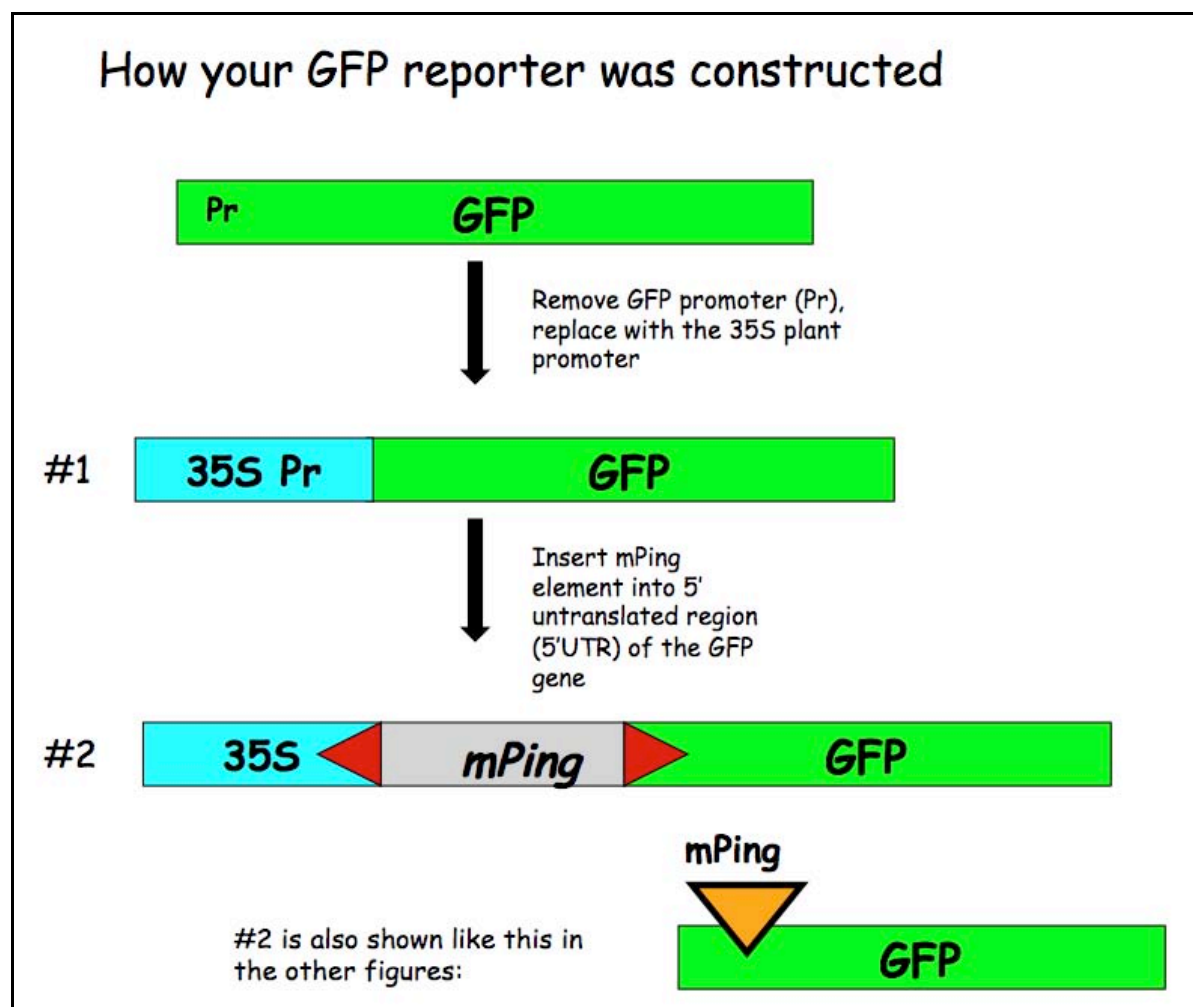
A note here about the precision of words that scientists use to describe experimental results. In this case, the researchers called *mPing* a "candidate" for an active transposon" and not simply an "active transposon." The reason is that computational analysis usually identifies sequences that must be tested further by experiments. In other words, finding identical copies of a TE in a genome is not sufficient evidence to conclude that *mPing* is in fact an active element. In Experiment 1 you will test whether *mPing* is actually able to move - right before your eyes.

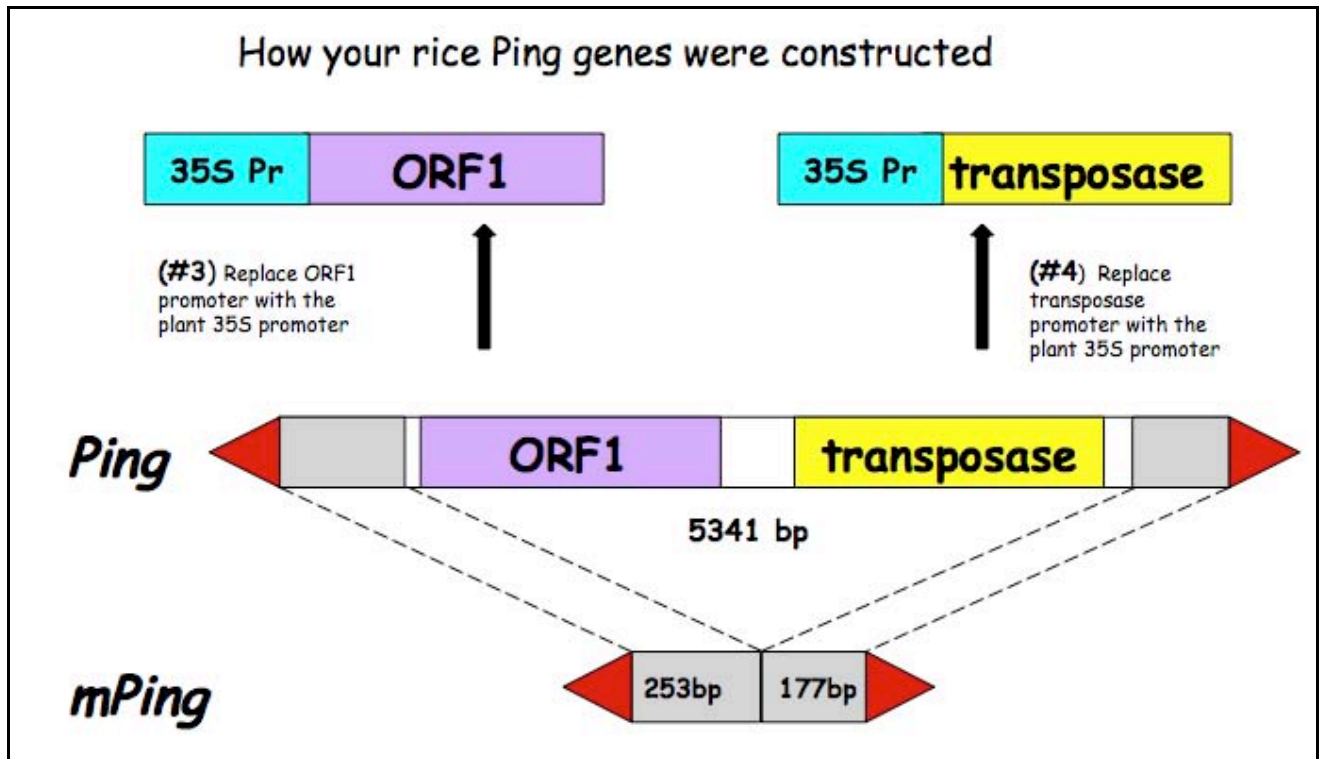
It was puzzling to understand how *mPing* could transpose because it is very small and does not code for any proteins and is thus unable to move on its own. The researchers reasoned that there must be a protein-encoding transposon in the rice genome that encodes the transposase necessary to enable itself and other related elements to move. To find this coding element, the researchers searched the rice genomic sequence for longer related elements. They found a candidate TE which they called *Ping* - that had the same ends as *mPing* but was much longer (~5000 bp) and contained two ORFs. One encodes the transposase gene and the second (called ORF1) is of unknown function (see Figure 7, page 7).

The purpose of your first experiment will be to test whether any or all of the proteins encoded by *Ping* can mobilize the *mPing* element. In other words, can either ORF1 or the transposase or both mobilize the *mPing* element.

1.10: Experiment 1: Design of the experiment and controls

Now up until this point Ping and mPing were considered active TE candidates, - as there was no evidence that these TEs were actually capable of moving around nor was there evidence that Ping produced a proteins that could catalyze the movement of mPing. Experimental evidence was necessary to move these elements from candidates to bona fide active TEs. To address these questions, transgenic Arabidopsis plants were generated by engineering T-DNA in the test tube and using *Agrobacterium tumefaciens* to deliver the following constructs into *Arabidopsis* plants. These are described in detail below.





In experiment #1 we are testing whether the Ping encoded protein(s) can catalyze the transposition of mPing. So, there are actually three questions we will be attempting to answer:

--Can ORF1 protein by itself excise (move) mPing?

--Can T_{pase} protein by itself move mPing?

--Can both proteins work together to move mPing?

To address these questions you will analyze mPing excision in transgenic Arabidopsis plants containing one or two of the following T-DNA constructs:

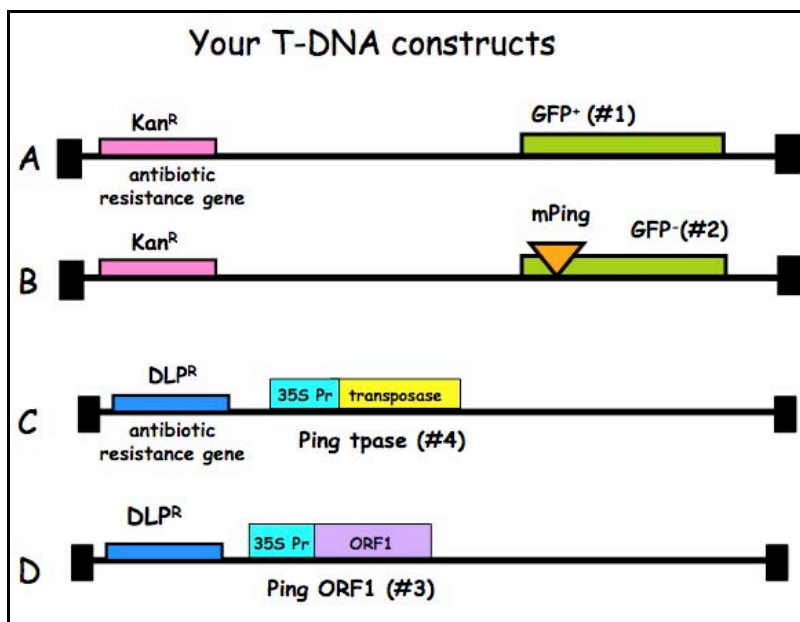


Figure 10:
The transgenic Arabidopsis plants used in this experiment contain one or two of the 4 T-DNA insertions in their genome.

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

(B) 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⁻). Note that the red color is due to chlorophyll fluorescence.

(C) Plants with this T-DNA are part of your experimental unknown.

(D) Plants with this T-DNA are also part of your experimental unknown.

(E) Not shown - NO T-DNA at all. This is the wild type control.

Note that A and B have the same antibiotic resistance gene and C and D share a different one.

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

You will be using the Polymerase chain reaction procedure in most of your experiments. For this reason, it is summarized below.

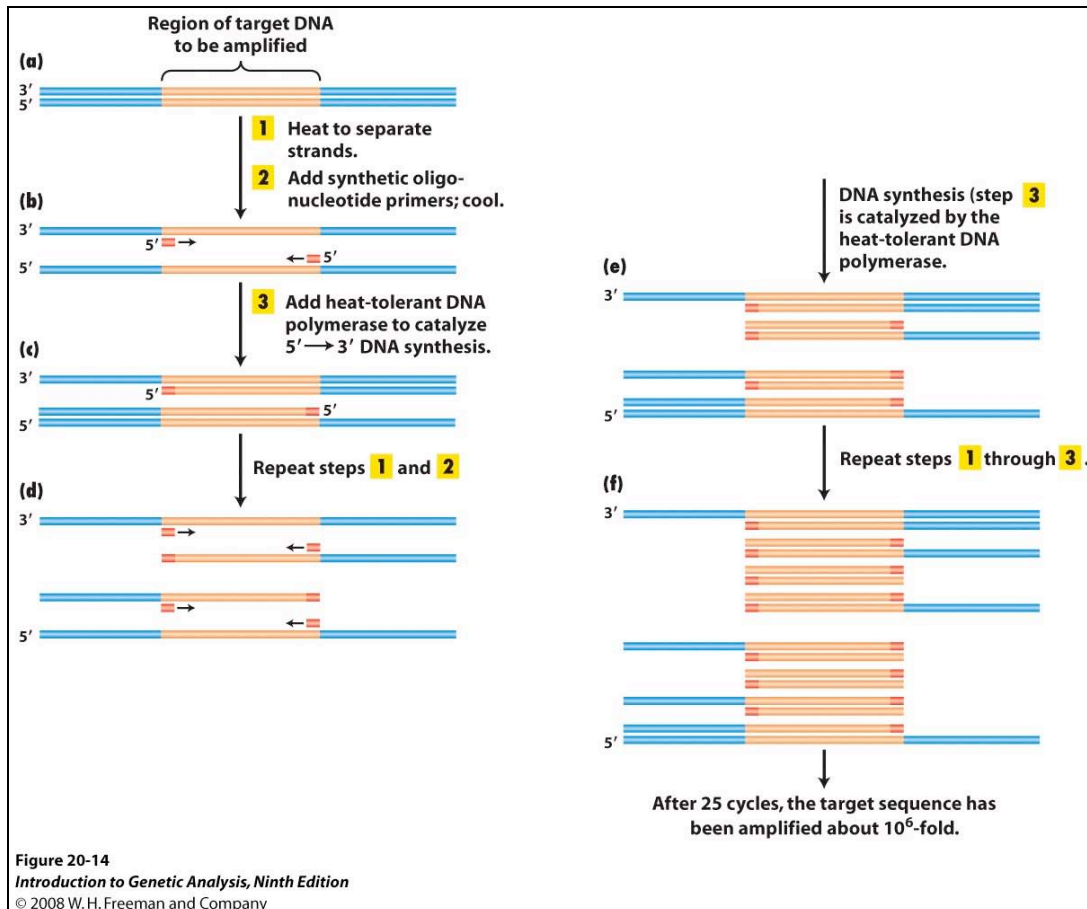


Figure 11: Details of PCR. See below.

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.

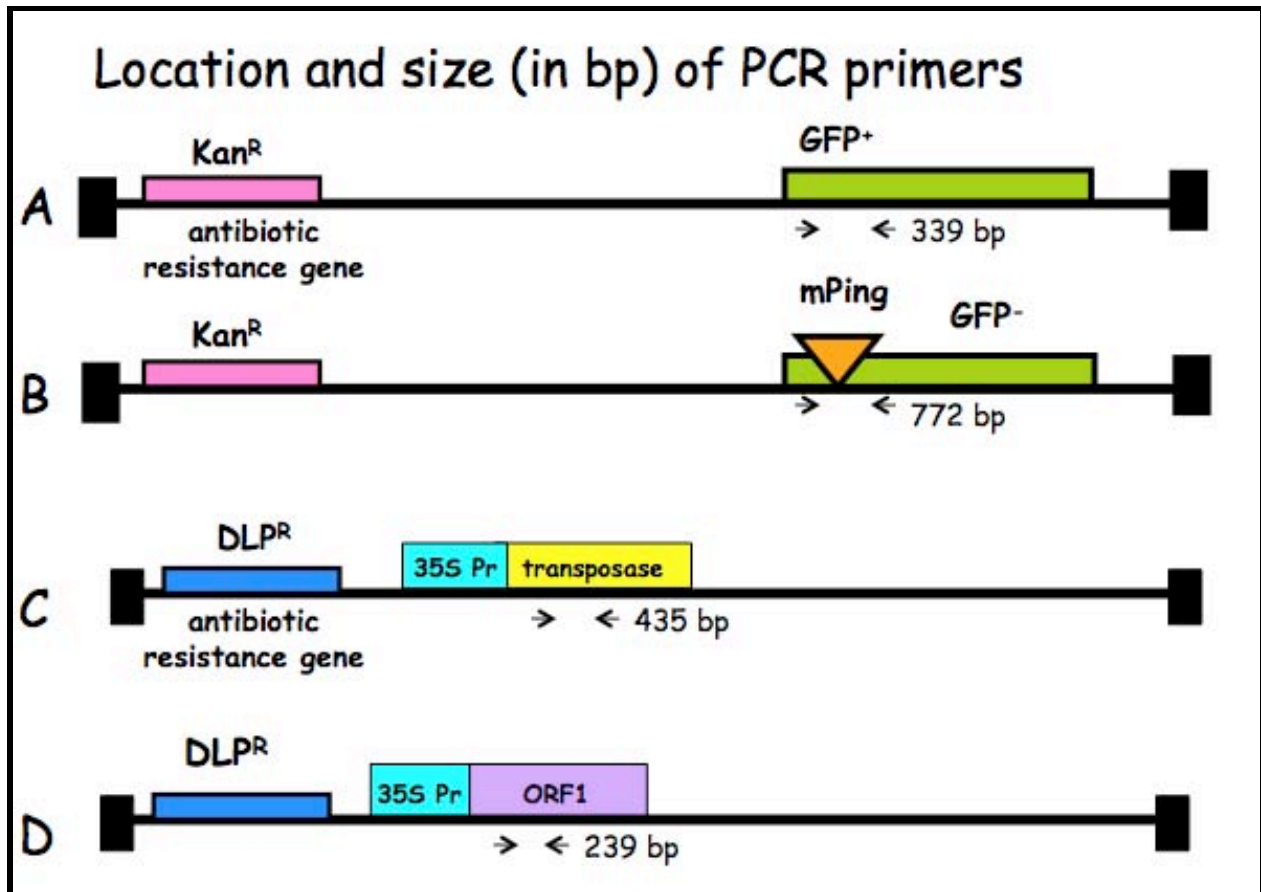
Annealing: when DNA or RNA strands pair by hydrogen bonds to complementary strands, forming a double-stranded molecule. The term is also used to describe the reformation (renaturation) of complementary strands that were separated by heat.

Extension: enzymatically extending the primer sequence—copying DNA.

1.12. A close look at the regions that will be amplified in Experiment 1

The regions to be amplified by PCR in Expt 1 are shown below as arrows to indicate the PCR primers and the direction of DNA synthesis.

Once you have grown your *Arabidopsis* seedlings, you're ready to isolate leaf DNA and to do PCR.



Experiment 1 Examination of mPing excision from *A. thaliana* leaf DNA

Overview: *In this experiment you will test the hypothesis that the Ping element of rice produces one or two proteins (transposase plus ORF1 protein) that can catalyze the excision the mPing element in the model plant Arabidopsis thaliana. In part 1, you will examine phenotypes of A. thaliana leaves, extract DNA from the leaves of three different strains and set up your PCR amplification.*

Protocol:

During this first day, you will extract DNA from the leaves of each strain. On day two you will view the leaves under the microscope and set up PCR.

You will work in groups of two. Each group will analyze a full set of plants.

I. Plating of Arabidopsis seeds.

One week before actually doing PCR with leaf DNA, the instructors will start growing the plants we will use in this experiment by plating Arabidopsis seeds on petri dishes containing the antibiotic kanamycin and MS salt media. Plate means more or less the same thing as plant, except in a petri dish. The plates were put into a growth chamber where they germinated for ~5 days. The reason that you will not be doing this part of the experiment is because it is very easy for a novice to contaminate the plates with bacteria and/or fungus.

II. Extract genomic DNA from seedlings.

Nucleic acids are extracted from the seedlings using a simple protocol. After you extract genomic DNA you will visualize it on an agarose gel. To save time, you need to prepare the gel first because it takes time for the gel to solidify after pouring.

1.5% Agarose Gel Protocol

1. Weigh out 1.5g agarose and add to flask.

2. Add 100 ml 1X TAE buffer (available in a big jug) to the flask with agarose.
(TAE = 40mM Tris acetate, 1mM EDTA pH 8.4)

3. Heat contents in the microwave until boiling (2-3 min). *Be very careful, as superheated liquids can boil over and burn you.*

4. Swirl to make sure that the agarose is completely melted.

5. Add 0.5 μ l of a stock solution of 10mg/ml ethidium bromide (EtBr). (This binds to the DNA allowing it to be visualized under UV light. *Do not let this stuff touch your skin.*)

6. Swirl again to mix and pour into a gel-casting stand with a comb (this will be demonstrated in the lab.) The gel should cool and solidify within 10-15 minutes at which time it is ready to place the gel in the electrophoresis apparatus and add enough TAE buffer to completely immerse the gel.

Now you will prepare the genomic DNA. We will walk through this protocol step-by-step in class.

Damon Lisch's All Natural (no organics) Genomic Miniprep
Modified for Arabidopsis seedlings

Materials list:

Extraction Buffer

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

1) Label 2 tubes for each plant. Set one set of tubes aside.

2) Harvest 5-6 seedlings. 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 1 ml of Extraction Buffer, and grind some more in the buffer.

4) Add 120 μ l of 10% SDS. Mix by inverting.

Prepare all samples to this step. Keep them on ice until all are ready for step 6.

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 10 minutes.

7) Spin for 5 minutes at top speed in microfuge. Squirt about 700 μ l of the supernatant through miracloth into second tube. (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. However, a really good prep (excellent grinding of tissue) should result in visible DNA at this stage.

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

10) Add 500 μ l of 70% ethanol and flick until the pellet comes off the bottom (for best washing results). Spin 3 min, then pipette off the ethanol with a P-1000. Suck off the rest of the ethanol with a P-20 pipette. Make sure the pellet stays in the tube! Let air dry in hood for around 30 minutes with the caps open.

11) Resuspend the DNA in 50 μ l water or TE. Let sit at RT for about 30 minutes, then mix by pipetting. Depending on amount of starting material may need to be diluted for PCR.

Visualize genomic DNA on an agarose gel:

12) Label new 1.5 ml tubes.

13) Put 15 μ l of DNA from step 11 into each tube.

14) Add 3 μ l of 6x loading dye to the tube. Tap the tube gently to mix.

15) Load all 18 μ l on your gel. Keep track of which sample went in which lane.

16) Load 7 μ l of DNA Ladder in one empty well.

17) Run the gel at 130 Volts for ~20 minutes.

18) Photograph the gel.

Thursday, January 15

III. Examine leaves under microscope.

You will view and photograph the seedling under a microscope using visible and UV light.

IV. Amplify genomic DNA using PCR.

Today you will amplify the DNA using 3 pairs of primers: one for GFP, one for ORF1, and one pair for T_pase. The primers for ORF1 and T_pase will be mixed and used in a single PCR reaction. This is called duplex PCR.

For each group, one person should set up the GFP PCR reaction and the other should set up the ORF1+T_pase reaction. Each person will have six DNA samples to analyze. We also need an additional negative control that will be water in place of DNA, giving you seven reactions in total. You need to make a master mix of eight reactions to make sure you have enough for the seven tubes.

1) Label a 1.5 ml tube with the primer set you are using: either GFP or ORF1+T_pase.

2) Label a strip of 0.2 ml PCR tubes. The instructors will show you where to label the tubes. The label may be rubbed off in the machine if you put it in the wrong place.

3) Mix the following in your tube using the volumes in the column labeled '8X (stands for 8-fold).' Keep this on ice.

	1X (μ l)	8X
2x Master Mix	12.5	100.0
H ₂ O	5.5	44.0
Forward Primer	1.0	8.0
Reverse Primer	1.0	8.0
DNA	5.0	-----
Total	25.0	160.0 μ l

The 2x Master Mix is supplied by a company (NEB) and contains Taq enzyme, buffer, and deoxynucleotide triphosphates (dNTPs) in a 2x concentration. This tube should be kept on ice to protect the enzyme from degradation.

4) Put 20.0 μ l of master mix in 7 of the PCR tubes.

5) Add 5.0 μ l of correct genomic DNA to the PCR tubes. Put 5 μ l of sterile water in tube 7.

6) Seal the tubes tightly with a strip of caps. Keep PCR tubes on ice until everyone is finished.

After everyone is done, your samples will be placed in a thermocycler or 'PCR machine' and cycled with the following conditions:

1 cycle for:	initial denaturation	94°C	3 min
30 cycles for:	denaturation	94°C	30 sec
	annealing	58°C	30 sec
	extension	72°C	1 min

[Note: "30 cycles" means all steps— denaturation, annealing, and extension— are repeated 30 times before going on to the next step]

1 cycle for:	final extension:	72°C	10 minutes
--------------	------------------	------	------------

7. After you finished setting up the PCR, you should pour a 1.5% agarose gel. See page 22. You will need one gel per group.

8. When the PCR finishes, add 5 μ l of 6x loading dye to each reaction. Load 20 μ l into each well of the gel. Remember to add a lane with 7 μ l of DNA ladder (molecular weight markers). Run the gel at 130 V for ~20 minutes.

Sequences of the Primers used:

GFP Primers (772 and/or 339 bp amplimers)

GFP-R 5'-AGA CGT TCC CAA CCA CGT CTT CAA AGC-3'

S35-F 5'-CCT CTC CAC TGA CAG AAA ATT TGT GC-3'

ORF 1 Primers (239 bp amplimer)

PING-ORF1-FOR 5'- CAC TGG TCA AGG TTG AAG TCA GCG ATC TCT G -3'

PING-ORF1-REV 5'- CAG CAT CCA TTT CGC TCT TGT CTT TCT CTG -3'

Tpase Primers (435 bp amplimer)

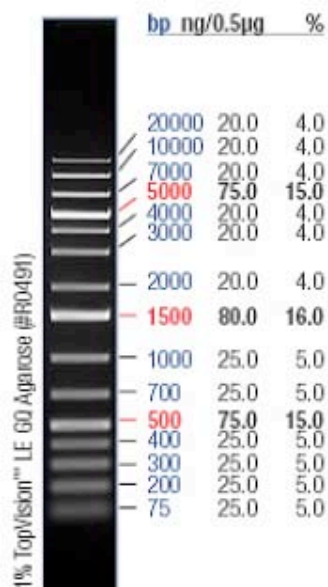
PING-TPase-For 5'- GGT ATG TTC GGT AGC ATT GAC TGT ATG CAT TGG C -3'

PING-TPase-REV 5'- GAA TCG ACG TTG TAG AAC ACC AAA TGC TCT CTC -3'

V. Gel analysis.

We will analyze the results of your PCR reactions and relate them to your observed leaf phenotypes.

O'GeneRuler 1 KB Plus DNA Ladder (Fermentas)



0.5 μ g/lane, 8cm length gel,
1XTAE, 7V/cm, 45min

Range

15 fragments (in bp): 20000, 10000, 7000, 5000, 4000, 3000, 2000, 1500, 1000, 700, 500, 400, 300, 200, 75.

Making of the Fittest: Concepts, ideas for discussion

P. 16 - "DNA contains an entirely new and different kind of information than what Darwin could have imagined...
"it is the decoding of the genes and genomes of other primates and mammals that enables us to interpret the
meaning of the human text"

The main ingredients of evolution - variation, selection time...

p 39 top The icefish.... May be a one-way trip

Chapter 2 - Math of evolution

p. 41 - the lottery example

p. 47 - hooded rat example - concept of QTL vs. saltation

p 57 - mutation lottery, erroneous concept that all mutations are bad

p 62 last sentence - continues to p. 64 - preexisting variation in populations

p 78 -79

p 85 - microbes exchange genes - horizontal selection

Chapter 4 - making new from old -

p 98-100 - from Molecules to trees - good intro to our next experiment. He mentions TEs and the figure on p
100 - TEs as markers.

P100 - use of TE (SINE) in evol studies

P104 - (also p 144- 145 and many other places) have them think about how TEs differ from opsins (and other
gene duplication events) re selection etc.

Chapter 5 - draw parallels with TEs as fossil genes but not quite.

P 136 -137 - fossil genes argue against design

P 139 - 140 - convergence

P 155-159 - the math of evolution again

P 162-163 - think of parallels with TEs

P 167 - comment on jello salad at a picnic (middle of page)

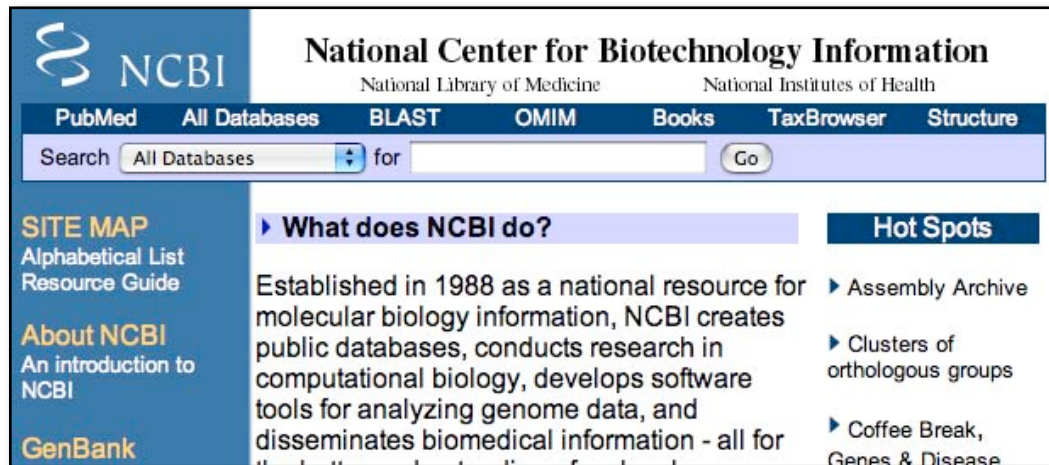
P 186 - "Wherever there is variation-in newts or snakes; parasites...the ongoing competition between predator
and prey, pathogen and host...leads to changes in the genetic makeup of populations."

TEs and host genomes to the list of conflicts

p 204 - regulatory vs. structural genes

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 also provides tools for searching and downloading the databases it maintains through the web portal NCBI Entrez. While the search tool for the literature database is PubMed, sequence is searched through a number of tools collectively called Blast. PubMed indexes 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. Blast searches on sequence databases that are often referred to as GenBank. There are three public repositories for sequence data: NCBI, DDBJ (Japan), and EMBL (Europe), and all share data on a nightly basis. Although the file formats and search tools may differ between the three repositories, they are essentially redundant.

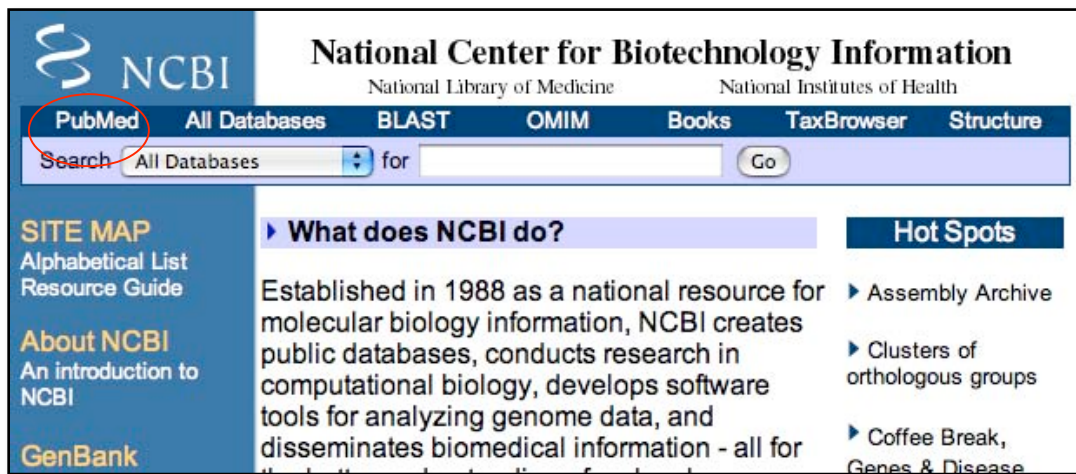
Spend some time browsing much of the NCBI Entrez portal. This is an amazing resource. There are plenty of help tutorials, free full-length (and slightly dated) textbooks, and a lot of interesting information.

Let's begin our tour by visiting the PubMed site and then move on to the BLAST site where we will be spending a great deal of our time today and in the rest of the course.

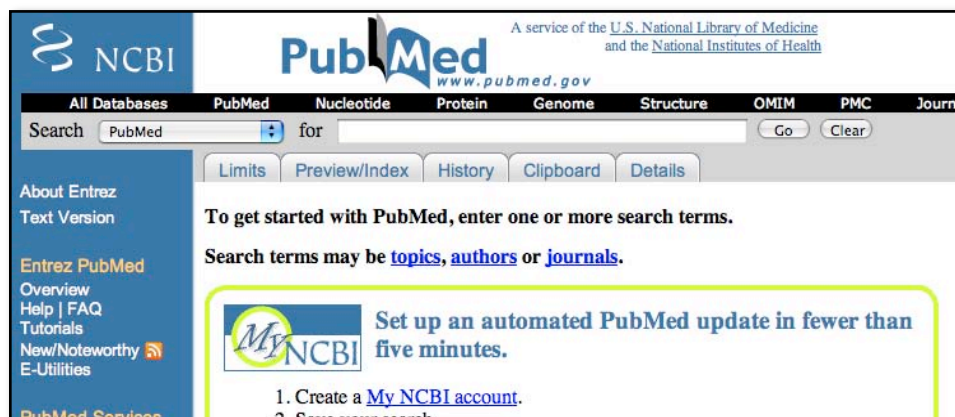
I. **PubMed**: Literature searches about a biological problem are very easy. PubMed makes the index available on its website with no access limitations. You can use PubMed (and Blast) from any computer and with any Internet connection. You are not required to be on the University network. To access many full-length articles and download PDF copies you will need to be using UGA's network.

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.

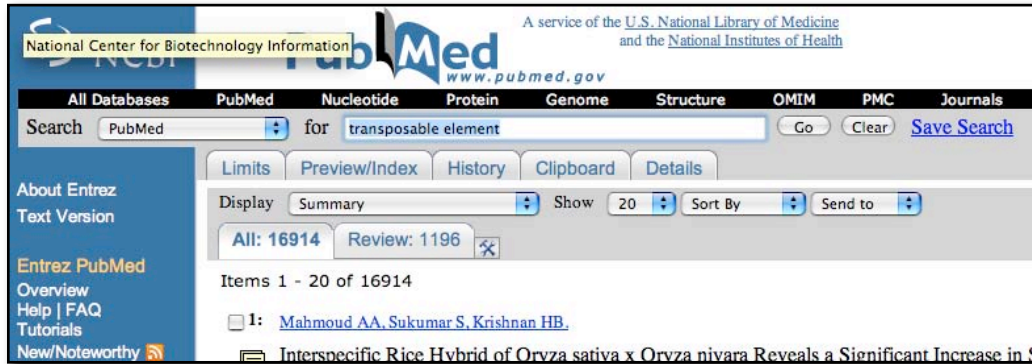


This is the PubMed homepage...

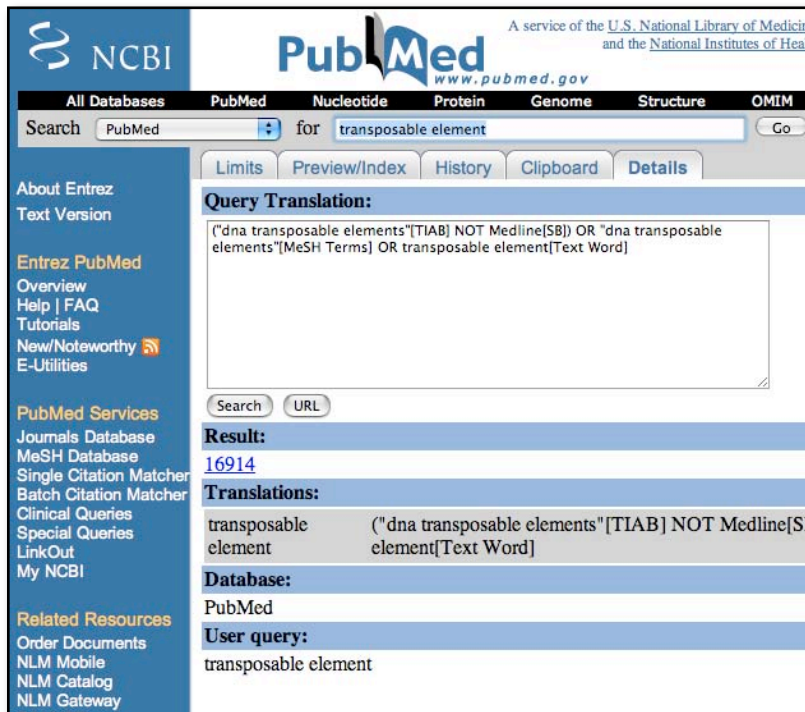


2. Think about the topic you want to search. You can use keywords, author last names, journal titles, publication year, or institution.

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

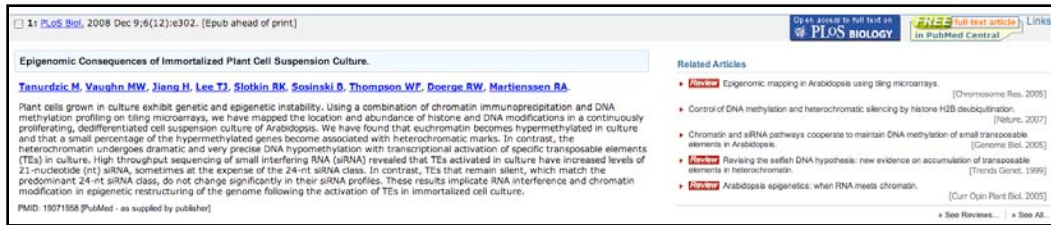
As you can see the simple query 'transposable element' is expanded into a more structured query by PubMed. MeSH is a controlled vocabulary for indexing PubMed. Curators at NCBI and journal editors assign these keywords based on suggestions by authors. Because of this query expansion it is a good idea to check the 'Details' tab whenever a search gives no results or unintended results.

4. The details of the results list like the one shown below will be discussed in class.

The screenshot shows a PubMed search results page. At the top, there are controls for 'Display' (set to 'Summary'), 'Show' (set to '20'), 'Sort By', and 'Send to'. Below these, it indicates 'All: 18215' and 'Review: 1314'. The main content area shows 'Items 1 - 20 of 18215' and a pagination bar for 'Page 1 of 911 Next'. The list of results includes:

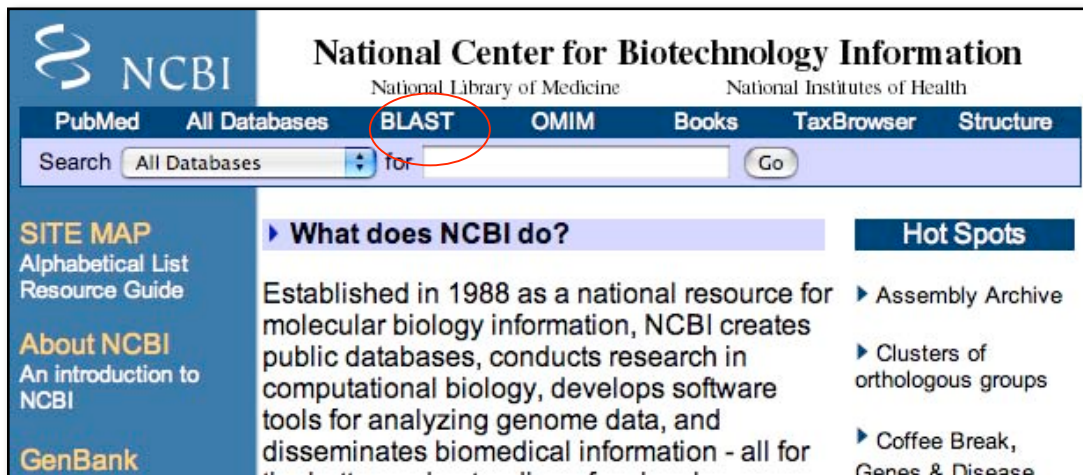
- 1: [Identification of r mutations conferring white flowers in the Japanese morning glory \(*Ipomoea nil*\).](#)
Hoshino A, Park KI, Iida S.
J Plant Res. 2008 Dec 16. [Epub ahead of print]
PMID: 19085046 [PubMed - as supplied by publisher]
[Related Articles](#)
- 2: [Transposable elements as genomic diseases.](#)
Wagner A.
Mol Biosyst. 2009 Jan;5(1):32-5. Epub 2008 Oct 27.
PMID: 19081928 [PubMed - in process]
[Related Articles](#)
- 3: [Epigenomic Consequences of Immortalized Plant Cell Suspension Culture.](#)
Tanurdzic M, Vaughn MW, Jiang H, Lee TJ, Slotkin RK, Sosinski B, Thompson WF, Doerge RW, Martienssen RA.
PLoS Biol. 2008 Dec 9;6(12):e302. [Epub ahead of print]
PMID: 19071958 [PubMed - as supplied by publisher]
[Related Articles](#) [Free article in PMC](#)
- 4: [Identification of a high frequency transposon induced by tissue culture, nDaiZ, a member of the hAT family in rice.](#)
Huang J, Zhang K, Shen Y, Huang Z, Li M, Tang D, Gu M, Cheng Z.
Genomics. 2008 Dec 8. [Epub ahead of print]
PMID: 19071208 [PubMed - as supplied by publisher]
[Related Articles](#)
- 5: [Efficient transposition of the Tol2 transposable element from a single-copy donor in zebrafish.](#)
Urasaki A, Asakawa K, Kawakami K.
Proc Natl Acad Sci U S A. 2008 Dec 16;105(50):19827-32. Epub 2008 Dec 5.
PMID: 19060204 [PubMed - in process]
[Related Articles](#)
- 6: [Transcriptome analysis in peripheral blood of humans exposed to environmental carcinogens: a promising new biomarker in environmental health studies](#)

5. Click on the underlined authors (in blue) and detailed information about the article is provided 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.



II. Introduction to **Blast**:

You will use Blast a lot this semester. 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 which will also be used frequently in this course:

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/BLAST Home

BLAST finds regions of similarity between biological sequences. [more...](#)

[Learn more](#) about how to use the new BLAST design

BLAST Assembled Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

- [Human](#)
- [Mouse](#)
- [Rat](#)
- [Arabidopsis thaliana](#)
- [Oryza sativa](#)
- [Bos taurus](#)
- [Danio rerio](#)
- [Drosophila melanogaster](#)
- [Gallus gallus](#)
- [Pan troglodytes](#)
- [Apis mellifera](#)

Basic BLAST

Choose a BLAST program to run.

- [nucleotide blast](#) Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontinuous megablast
- [protein blast](#) Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast
- [blastx](#) Search **protein** database using a **translated nucleotide** query
- [tblastn](#) Search **translated nucleotide** database using a **protein** query
- [tblastx](#) Search **translated nucleotide** database using a **translated nucleotide** query

There are six different versions of BLAST because you can use a nucleotide sequence or protein sequence to query nucleotide or protein databases. The different versions are summarized in the screenshot above. 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:

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

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

```
>mPing
ggccagtcac aatgggggtt tcaactggtgt gtcatgcaca ttaatatagg gtaagactga
ataaaaaatg attatttgca tgaatgggg atgagagaga aggaaagagt tcatcctgg
tgaactcgt cagcgtcgt tccaagtctt cggtaacaga gtgaaacccc cgttgaggcc
gattcgtttc attcaccgga tctcttgctt ccgctccgc cgtgcgacct ccgattctc
ccgcgcgcgc cggattttg ggtacaaatg atcccagcaa cttgtatcaa ttaaatgctt
tgcttagtct tggaaacgtc aaagtgaac ccctccactg tggggattgt tcataaaag
atctcatttg agagaagatg gtataatatt ttgggtagcc gtgcaatgac actagccatt
gtgactggcc
```

The screenshot shows the BLAST web interface. The 'Enter Query Sequence' section is active, with the sequence from the previous block pasted into the text area. The 'Job Title' field contains 'mPing'. The 'Query subrange' section is empty. The 'Database' section is not visible in this screenshot.

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.

The screenshot shows the 'Choose Search Set' section of the BLAST web interface. The 'Others (nr etc.)' radio button is selected, and the dropdown menu shows 'Nucleotide collection (nr/nt)'. The 'Database' section is highlighted in yellow. The 'Organism' and 'Entrez Query' sections are optional and empty.

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


Program Selection

Optimize for

Highly similar sequences (megablast)

More dissimilar sequences (discontiguous megablast)

Somewhat similar sequences (blastn)

Choose a BLAST algorithm 

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

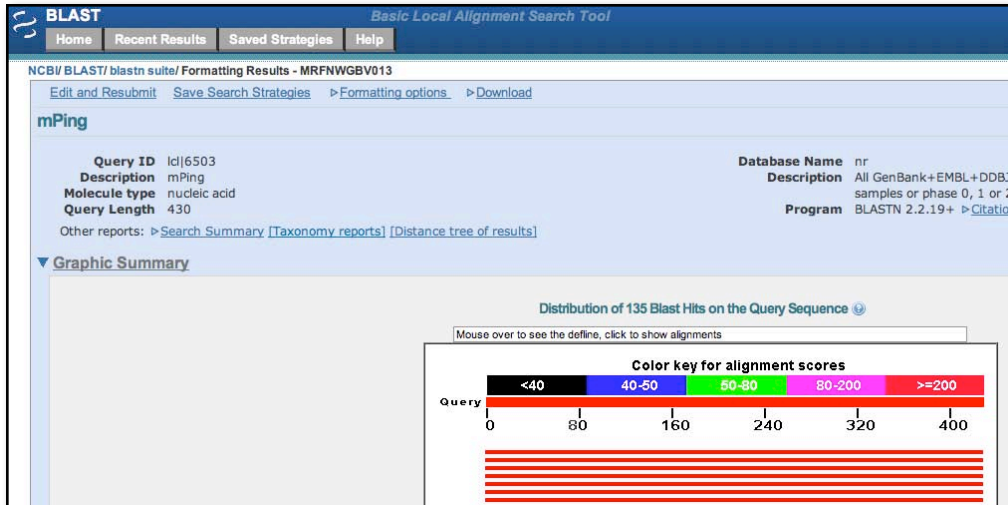
▸ [NCBI/ BLAST/ blastn/ Formatting Results - S9WZE65Y013](#) [\[Formatting options\]](#)

Job Title: lc|21740 (430 letters)

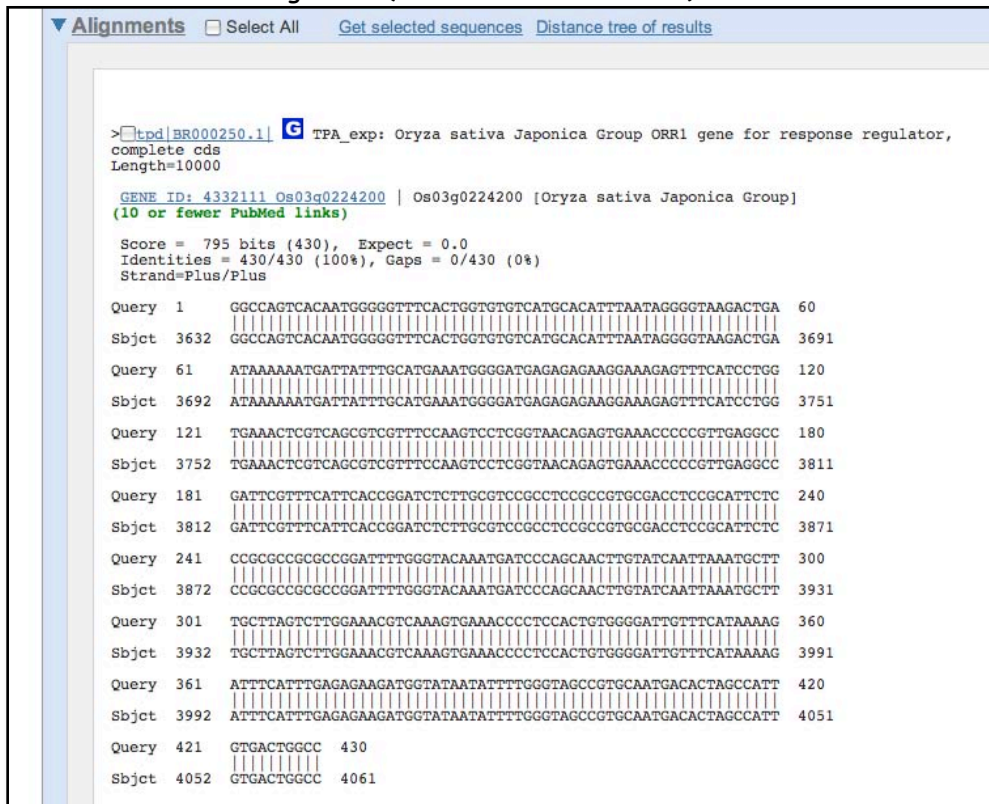
Request ID	S9WZE65Y013
Status	Searching
Submitted at	Wed Jan 9 11:18:54 2008
Current time	Wed Jan 9 11:18:56 2008
Time since submission	00:00:01

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)



A short discussion on how Blast works.

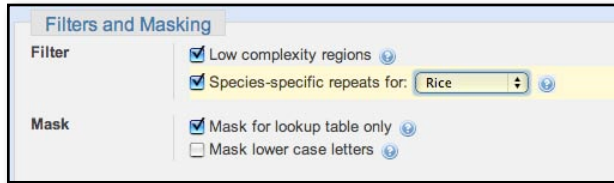
Blast takes the query sequence and divides it into "words" based on the word size parameter (the default is usually "fine"). For a megablast query the default (and minimum) is a size of 28. The algorithm then takes these "words" and runs them against a hash database where the large database is cut into 28 bp words. When an exact match occurs, the program attempts to extend the alignment in each direction on the full sequence. 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 that occur fairly frequently in a genome.

In class for nucleotide Blast

1. 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?



2. Choose one of the following sequences. What TE does the following sequence come from? How long is the element? What genomes is the sequence found in?

>TE1

```
tccatccc cctccctcc acagccgat tccccattcc caaacctaac cgtaggcgac
ggcggcggcg gcagcgacgg cggcggcggg ggtggcgggg ccggcgggtg cggcgccggc
ggaggccgat ggagctgtca tattggtagg cgcccgagcg gcagctagga agatgtcgcc
```

>TE2

```
ctccatccat ccataatat aagacgcacc cgtatttgaa gattaacctt taaacattt
gaccaacact tagttaatat aaaatatttt ttatttatta aaagttatat tattggattg
agatttaaat ttactttcat acggaataa ttttgttgct acaaacctta tagtatgtga
gaaattataa actaaatatt aattttggat ggag
```

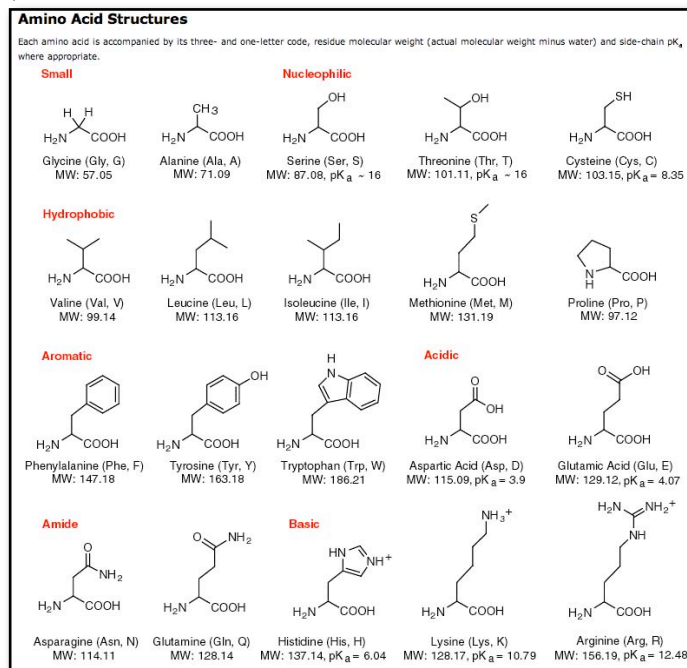
>TE3

```
cctcctctgt tcttcacctc tctgtctcta gttcttcctt tgctatgato tagtagtggt
acagtgggat cagattcggg ttttctgtgg ttgatgcggt caaattctct cgatccgact
ggtaaaatcc ctgtgctccg gtagacggct accctcgggt gcggctgctg taagtccagt
```

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→A, G→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.

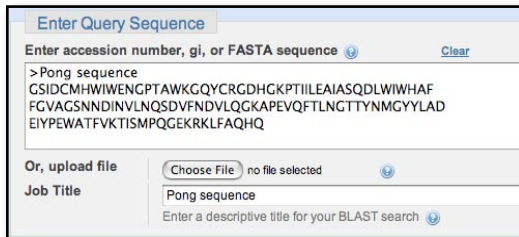
Common groupings of the amino acids (from

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

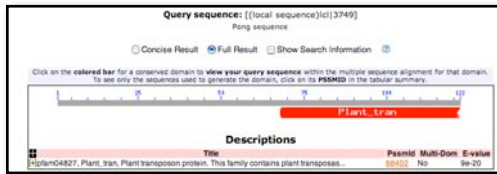
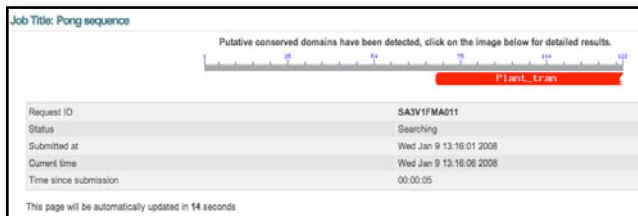
G, A, V, L, I, M	aliphatic (though some would not include G)
S, T, C	hydroxyl, sulfhydryl, polar
N, Q	amide side chains
F, W, Y	aromatic
H, K, R	basic
D, E	acidic

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.

```
>Ping Tpase
MSGNENQIPVSLLEDEFLAEDEIMDEIMDDVLHEMMVLLQSSIGDLEREAADHRLHPRKHIKRPREEAHQN
LVNDYFSENPLYPVSNIFRRRFRMYRPLFLRIVDALGQWSDYFTQVDAAGRQGLSPLQKCTAAIROLATG
SGADELDEYLKIGETTAMDAMKNFVKGIREVFGERYLRRPTVEDTERLLELGERRGFPGMFGSIDCMHWQ
WERCPTAWKGQFTRGDQKVPTLILEAVASHDLWIWHAFVGVAGSNNDINVLSRSTVFINELKGQAPRVQY
MVNGNQYNENGYFLADGIYPEWKVFAKSYRLPI TEKEKLYAQHQEGARKDIERAFVGLQRRFCILKRPARL
YDRGVLRDVVLCIILHNMIVEDEKEARLIEENLDLNEPSSSTVQAPEFSPDQHVPLERILEKDTSMRD
RLAHRRLKNDLVEHIWNKFGGGAHSSG
>Pong sequence
GSIDCMHWIWENGP TAWKGQYCRGDHGKPTIILEAIASQDLWIWHAF
FGVAGSNNDINVLNQSDVFNQDLQKKAPEVQFTLNGTTYNMGYLAD
EIYPEWATFVKTISMPQGEKRLFAQHQ
```



2. Run the Blast with all default parameters. The queue screen will report that it found a similarity between your query sequence and the Protein Family (PFam) database. This suggests that the query sequence came from a plant TE protein. Makes sense since Pong is a TE.



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 similar amino acids match a '+' is used.

```

▼ Alignments   Select All  Get selected sequences  Distance tree of results

>[gb]AAK92907.1| transposon protein, putative, ping/pong/SNOOPY sub-class [Oryza
sativa (japonica cultivar-group)]
Length=443

Score = 203 bits (517), Expect = 3e-51, Method: Composition-based stats.
Identities = 88/122 (72%), Positives = 103/122 (84%), Gaps = 0/122 (0%)

Query 1  GSIDCMHNIWENGPTANKGOYCRGDHGKPTIILEAIASQOLNIWHAFVGVAGSNNNDINVL  60
          GSIDCMHN WE  PTAW Q+ RGD+G PTIILEA+AS DL IWHAFVGVAGSNNNDINVL
Sbjct 158  GSIDCMHRRWKCPTANSQOFTRGDYGVPPTIILEAVASYDLRIWHAFVGVAGSNNNDINVL  217

Query 61  NQSDVFNQVQKAFVQFTLANGTYNMGYLADEIYPENAFVKTISMPQGEKRLFAQ  120
          NOS +F DVL+G AF+V+F++NG Y+ GYLA+ IYPENA FVK+I +FQ EK KL+AQ
Sbjct 218  NQSFLELDVLKGDAPQVKFVNGNEYSTCYLLANGIYPEWAAFVKSIIHLPQTEKHKLYAQ  277

Query 121 HQ 122
          +Q
Sbjct 278 YQ 279

```

In class Questions for protein blast.

1. Run a protein blast with the following sequence. What element does it come from? What is the function of the protein?

>TE1

```

MATNSWVRDRVINKLRQEPTLGATALKKFLEEKYKIN
ISYYVVVDGRQMALDEILGKWEDSFDAAYNFKAELERTSPGSIVEVDHVTV
DGKNHFSKMFVALKPCVDGFLNGCRPYLGIDSTVLTGKWRGQLASAIGIDG
HNWMFPVAYGVFESESTDNWAWFMDKLSAIGSPEGLVLSTDAGKGIDTAV
TRVFTNGVEHRECMRHLVKNFQKRFSGEVFERNLWPASRAYRQDIFESHYN
EMKEACPATEWIDNFHKHIWTRCQFSTLSKCDYVTNNAIETFNWIRHEK
SLPVVDLMDKIRQMIMERMMSVRKRLAVKLTGTILPSVMKSLYARSRDLYGK
LYSAHSHLGEIGGTGRDLKTRHTVDLNTRECSQRQVQTGIPCTHAIFLV
ISRRGLELEQFVDDCYSVATFKKAYAGHVVPMTDKSQWAKINVGFKLYPPL
LKRSAGRPSSRIKGMEEGGSGKRKYRCKRCGQFGHIKKTCEFPVADPSAP
PPAPPKPKRKRKRVKVVLSVQDPVIQVPTATVYVVNSKISFLCHTFAIFTLH
IYVDFHFFAGPHLQDHKHLG

```

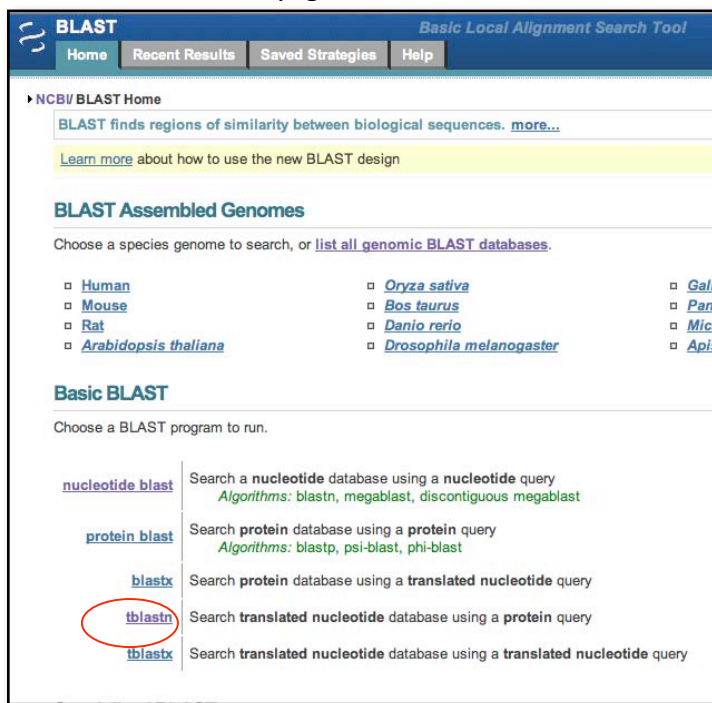
C. tblastn:

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 gene encoding 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 from above. Remember, this process compares a sequence of amino acids against sequences in existing genomes.

>Pong sequence

```

GSIDCMHIWENGPTAWKGQYCRGDHGKPTIILEAIASQDLWIWHAF
FGVAGSNNDINVLNQSDVFNVDVLQGKAPEVQFTLNGTTYNMGYYLAD
EIYPEWATFVKTI SMPQGEKRRLFAQHQ

```

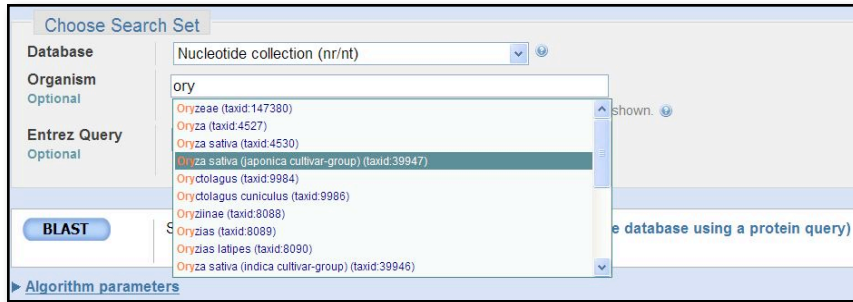
3. Now go down to the section called "Choose Search Set."

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 name for Asian 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. You can also enter *Oryza sativa*. As you see, you only have to type "Ory" and all the rice varieties pop up.

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

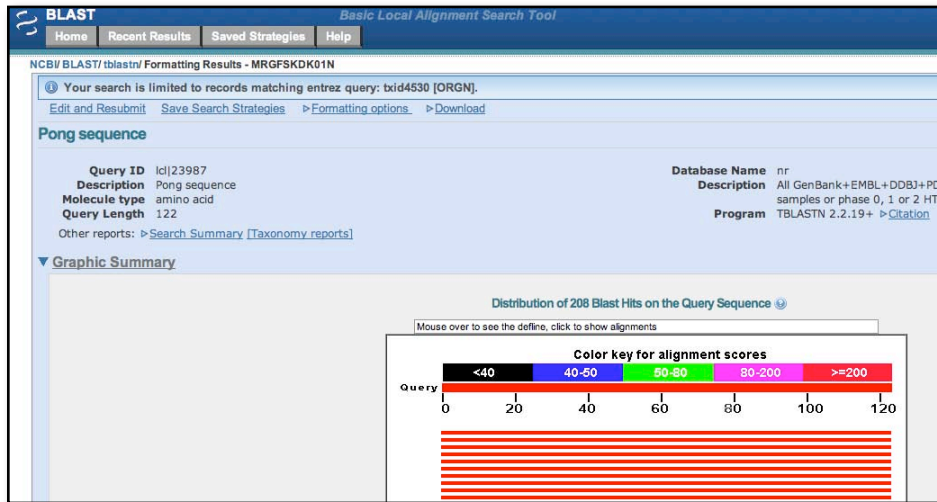
Now you see on your line this: *Oryza sativa* (taxid 4530).



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.



Chapter 3. Introduction to using yeast in the lab

You will use yeast for the next two experiments. Yeast, like *Arabidopsis* is a model organism. Yeast has features that make it well suited to use in a molecular biology laboratory for fast experiments. Yeast is a single-celled eukaryote that can live as a haploid or diploid, that is with one or two copies of its genome. We will use yeast in its haploid state because it makes genetic manipulation very simple. Yeast also has a short generation time of 90 minutes. This means that can grow from a single cell to a colony on a petri dish in about two days or saturate a liquid culture in the same amount of time. In contrast, it takes about a week or more to germinate *Arabidopsis* seedlings. Because yeast are small, we can grow millions on a single plate. This gives us the opportunity to screen for very rare transposition events using fewer resources.



Figure 1: A Yeast cell with a bud.

When working with yeast you must have very good sterile technique. Antibiotics are rarely used in yeast work and bacteria and other molds love to grow on yeast media. The rich medium YPD contains a digest of yeast, a lot of dextrose (glucose), and salts and supports the growth of a lot of things. You only want yeast on your plate!

Because yeast can grow to very high numbers in liquid culture we need a way to accurately count the number of living yeast. Shown below in Figure 2 is a growth curve for yeast. (or any organism that doubles at each generation) The curve is generated by counting the number of organisms in a culture over time.

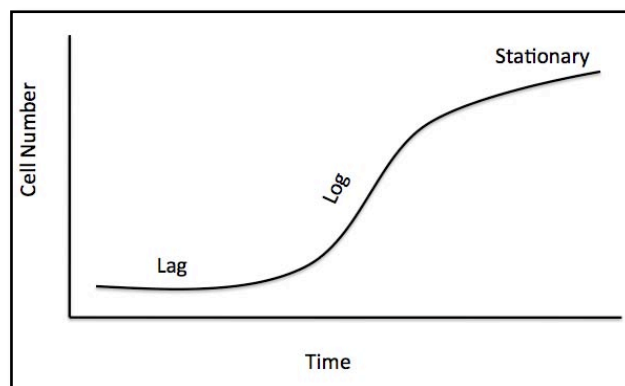


Figure 2: A standard growth curve showing the three growth phases of culture growth.

There are three parts to a growth curve: lag phase, log phase, and stationary phase. Typically we want to work with yeast when they are in the log phase. During this time, yeast are dividing at maximal rate (around 90 minutes/generation) and are metabolically active. A log phase growth culture has between 10^7 to 10^8 cells/ml. During lag phase cells are adjusting to the new media and are not as metabolically active as during log phase. At stationary phase the cells have run out of nutrients and may become quiescent (where they carry out normal metabolism but do not divide).

In this lab we will determine the population size of a culture using a viable count. This type of count will detect only living cells. Other methods use light scattering to determine population size but this number is inflated because it can also detect dead cells and debris.

A viable count simply means putting a small sample of culture on a rich media plate (YPD) that supports the growth of all yeast regardless of genetic background. After two days in the incubator at 30°C you simply count the colonies on the plate. The complication comes from the fact that a small sample of a log phase culture can have thousands of cells making it impossible to count them. To plate only a countable number of cells, you must first dilute the culture sample several times. This is called serial dilution.

Let's assume that a mid-log phase culture has 1×10^7 cells/ml. A $100 \mu\text{l}$ sample ($1/10$ of a milliliter, the usual amount of culture put on a petri dish) will have 1 million cells (1×10^7 cells/ml * 1×10^{-1} ml = 1×10^6 cells). The target number is 100 cells on a plate that will become 100 colonies that we can easily count. To achieve this we need to do a series of dilutions. Looking at Figure 3 below you can see that if you do a 10 fold dilution by putting 1 ml of culture into 9 ml of water you will have 1×10^6 cells/ml and 1×10^5 cells in $100 \mu\text{l}$. This is still too many. You will need to do a series of four 10-fold dilutions to get to the desired amount.

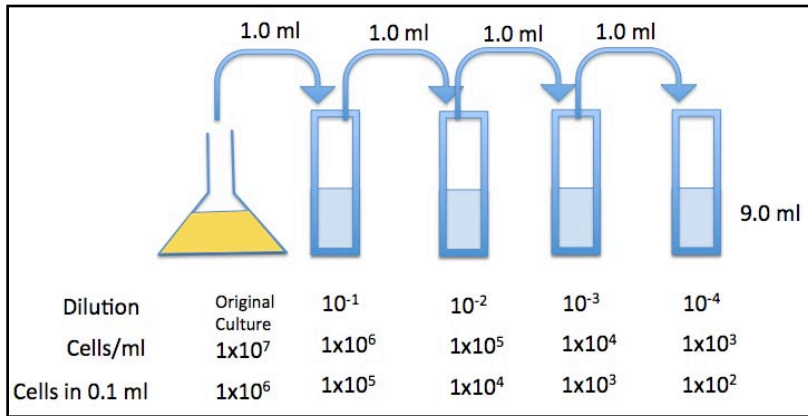


Figure 3: An example of serial dilution using 10-fold dilutions.

To simplify the experimental procedure, we often do 100-fold dilutions in the lab. To do this you would place 0.1 ml of culture into 9.9 ml of water. This time, the first dilution is a 100-fold or 1×10^{-2} dilution. Fill in the missing values in Figure 4 to determine how many dilutions you need to do so that you can plate 100 cells.

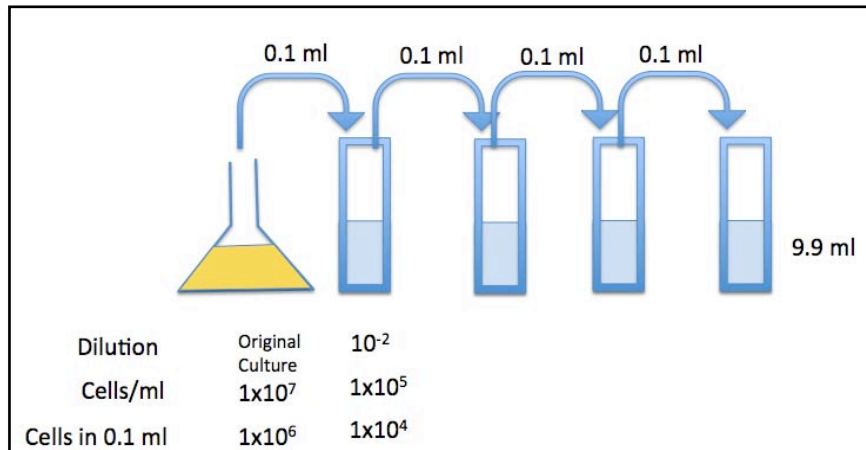


Figure 4: 100-fold dilutions. You need to fill out the missing numbers to determine how many dilutions are necessary.

During this lab several yeast cultures will be available for you to do serial dilutions and plate for viable count. You need to do three viable counts. This will provide you with the opportunity to practice your sterile technique. It will also give you practice in working with yeast and working with many tubes and different volumes. Label everything very well. All of these techniques will be very important when we start Experiment 2.

If you have forgotten how to do calculations with powers of ten, please read this web page: <http://www.astronomynotes.com/mathrev/s3.htm>.

1. Label 2 glass tubes: 10^{-2} and 10^{-4} . Put 9.9 ml of sterile H_2O in each tube
2. Label 2 YPD plates with the culture name and 10^{-2} or 10^{-4} plus your initials and the date. Pour 5-6 glass beads on each plate
3. Pipette 100 μ l of culture into the 10^{-2} tube
4. Gently vortex tube (use setting 7-8 on vortex)
5. Pipette 100 μ l from the 10^{-2} tube into the 10^{-4} tube, vortex
6. Vortex the 10^{-2} tube and pipette 100 μ l of the liquid onto the corresponding 10^{-2} plate
7. Repeat for the 10^{-4} tube
8. Gently shake plates to distribute the liquid evenly across the agar.
9. Incubate 'upside down' for 2 days at $30^\circ C$

Repeat with 2 more cultures.

Chapter 4: From a single element to all of the elements in the genome

In Experiment 1 you worked with the autonomous Ping and nonautonomous mPing elements. It turns out that the rice genome can contain up to 1000 copies of mPing and from 0 up to seven Ping elements (this depends on the rice strain). Ping and all of the mPing elements in the rice genome make up a TE family.

The genomes of plants and animals contain many 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 (from maize) contains the Ac and Ds elements while the second family (from rice) contains Ping and mPing 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, Ping) and nonautonomous elements (e.g. Ds, mPing) elements. When we analyze the DNA sequence of entire genomes we often find that a family contains several elements including one or more autonomous elements and many copies of nonautonomous elements (the maize genome has over 50 copies of Ds and, as mentioned above, some rice genomes have up to 1000 copies of mPing).

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 Ping and mPing in rice - mPing will be stuck in place if Ping 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 Ping or mPing elements and, similarly, the Ping transposase cannot mobilize Ac or Ds elements. Or, as shown in **Figure 1**, the transposase from family A cannot move the elements in family B. The reason for this is simple. A transposase works by first binding to a specific DNA sequence near the ends of the element (as shown in **Fig 5**, on page 5) called the Terminal Inverted Repeat or TIR. The Ac transposase first binds to a specific sequence of nucleotides that is only near the ends of Ac and Ds elements while the Ping transposase binds to a specific sequence that is only near the ends

of Ping and mPing 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.)

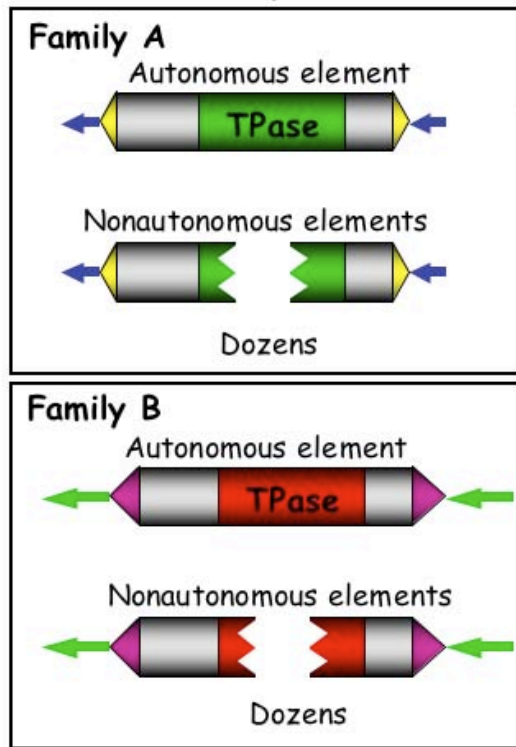


Figure 1: A TE family contains autonomous elements and all the nonautonomous elements in the genome that its transposase can move. Genomes have many TE families that are independent. This is because the transposase of Family A, for example, cannot bind to the ends of the elements from Family B and vice versa.

What is a TE superfamily?

After Barbara McClintock discovered *Ac* and *Ds* (in the 1940's) 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 TE family.

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 fruit 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 *Tc*1, 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 amino acid sequence of the transposase from the maize *Ac* element was similar to the amino acid sequences of 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 2**.

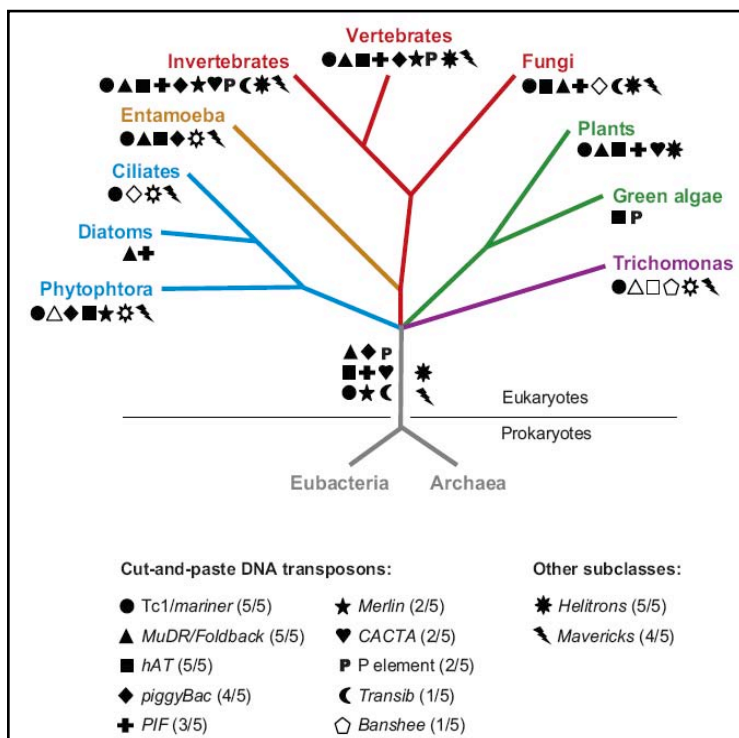


Figure 2. Distribution of the major groups of DNA transposons across the eukaryotic tree of life. The tree depicts 4 of the 5 "supergroups" of eukaryotes where DNA transposons have been detected. The occurrence of each TE superfamily is denoted by a different symbol. (Feschotte ·Pritham *Annu. Rev. Genet.* 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, rice and Arabidopsis, and are also present in most animal genomes (**Figure 2**). For example, the rice genome has Mariner, PIF/Harbinger (Ping), hAT (Ac/Ds), CACTA and Mutator elements. In addition, each superfamily usually contains many families in one genome.

Structural features shared by superfamily members:

Before you can study a TE superfamily, we need to look closely at the structural features of transposable elements in more detail because these features are usually shared within a superfamily.

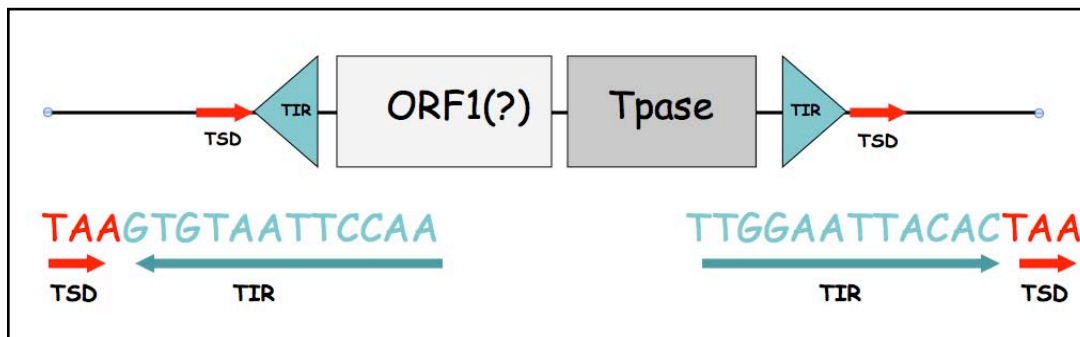


Figure 3: Structural features of transposable elements that are shared by superfamily members. TSD = target site duplication, TIR = terminal inverted repeat, Tase = transposase gene that is present in all autonomous elements, ORF1 - a second gene that is only encoded by members of the Ping/PIF/Harbinger superfamily

The terminal inverted repeat (TIR):

In the figure above, the sequence of the blue triangles is shown. Look closely and you will see that the sequence of the right TIR is the reverse-complement of the left TIR. These sequences help define a TE family because they are bound by TPase produced by a family member. While all members of a TE family have identical or near identical TIRs, the TIRs of superfamily members (elements from different species) are usually similar but not identical. In addition, the length of the TIR can vary. For example, the length of the Ping TIR is 15bp while the length of the Ac TIR is 11bp.

Target site duplication (also on page 6)

The target site duplication (TSD) is a direct repeat sequence that flanks the TIR. It is generated during the insertion of virtually all TEs into genomic DNA. How it is formed is shown below.

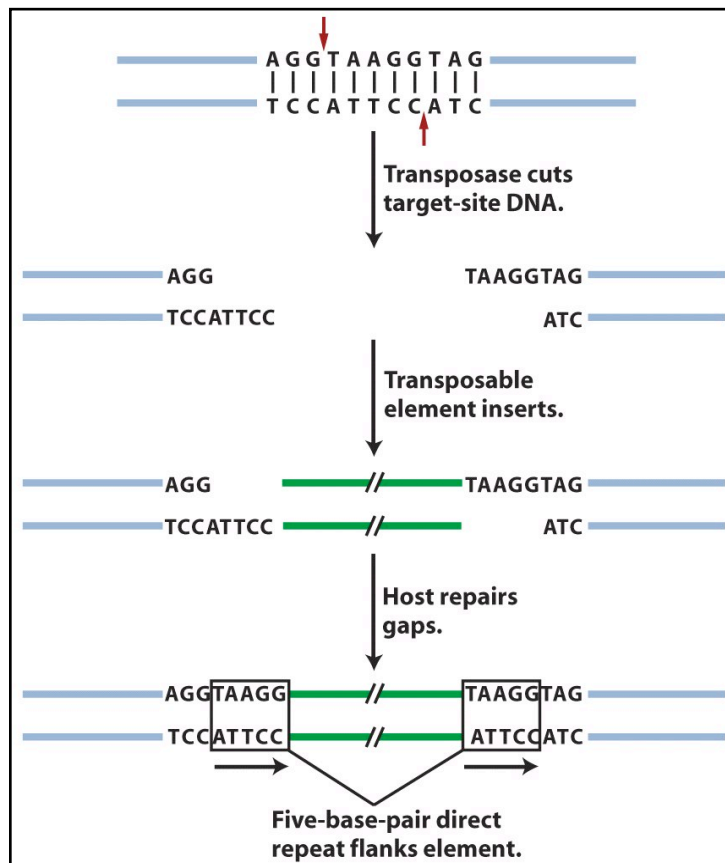


Figure 4: An inserted element is flanked by a short repeat. A short sequence of DNA is duplicated at the transposon insertion site. The recipient DNA is cleaved at staggered sites (a 5-bp staggered cut is shown), leading to the production of two copies of the five-base-pair sequence flanking the inserted element.

The length of the TSD, but usually not the sequence, is a common feature of a TE superfamily. For example, members of the hAT superfamily (*Ac*, *Tam* etc) all have an 8bp TSD, while members of the Mutator family have a 9bp TSD. *Ping* has a 3 bp TSD that is almost always TAA or TTA.

The Transposase (Tpase) gene:

The sequence of the TPase is also characteristic of a superfamily. In fact, the tpase sequence (or part of it) is THE feature used to define superfamilies. Later in the course you will use the sequence of the TPase and the sequence of the TIR to find *Ping* family members in rice and in other plant species.

Using computational analysis to find all elements related to Ping in rice and other genomes:

You can identify all TEs related to Ping in rice because the whole genome of rice has been sequenced. To do this one performs a Blast search using either the DNA sequence of the whole element or the protein sequence of the TPase. Using the whole element as query would retrieve only very related elements. To explore the diversity of the superfamily (in the rice genome or in other sequenced genomes) you would use the amino acid sequence of the TPase protein or part of the sequence. The Blast results in either case would be numerous and determining relatedness of the elements is impossible from a Blast output. To analyze the relationships between large numbers of related DNA sequences we use phylogenetic trees. These trees are similar to the species trees you have seen in other classes.

You have learned about sequence alignments using a single query of a large database. The result is many 'hits' that must be compared to each other in order to determine which sequences are most closely related. This process is called multiple alignment and there are several computer programs for this task. Once you have a multiple alignment a different software program is used to construct a phylogenetic tree. The process of generating a tree can be time consuming and tedious. Luckily for us Yujun Han (a graduate student in the Wessler lab) streamlined this process by creating a single web-based bioinformatics pipeline called TATE. You will learn all about TATE during class and you will use it often during the rest of the semester.

A Brief Look at a Phylogenetic Tree of Ping-like elements in rice

Look at the phylogenetic tree of the elements related to Ping in rice in **Figure 5**. The red arrow is pointing to THE Ping element. This tree was constructed using a part of the TPase amino acid sequence. Shown beside the tree is the DNA structure of each element. We will discuss this tree in class and point out its features and key terms (also later in this section). You will be making lots of your own trees in this class. This is just meant as an introduction and overview.

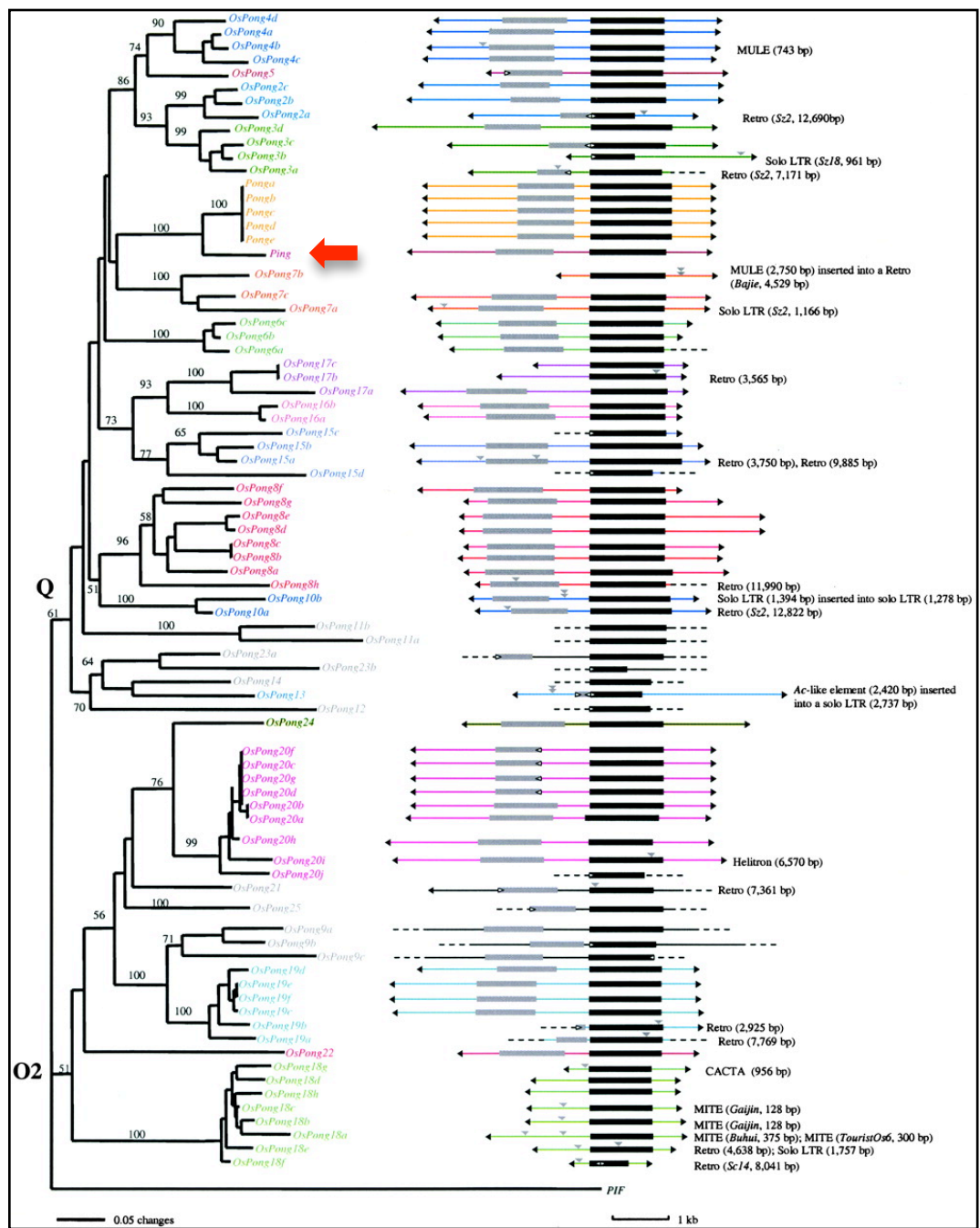
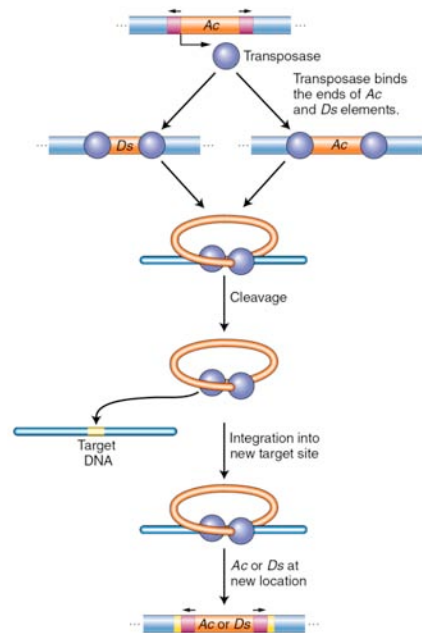


Figure 5: A phylogenetic tree of the relationships between Ping/Pong-like elements in the rice genome. The structure of each element is shown at the right with the transpose gene drawn as a black box, ORF1 drawn as a gray box, and arrows for the terminal inverted repeats (TIRs). Ping is indicated with a red arrow. Zhang et al. (2004) *Genetics*. 166: 971-986.

How do DNA Transposons make duplicate copies when they transpose?

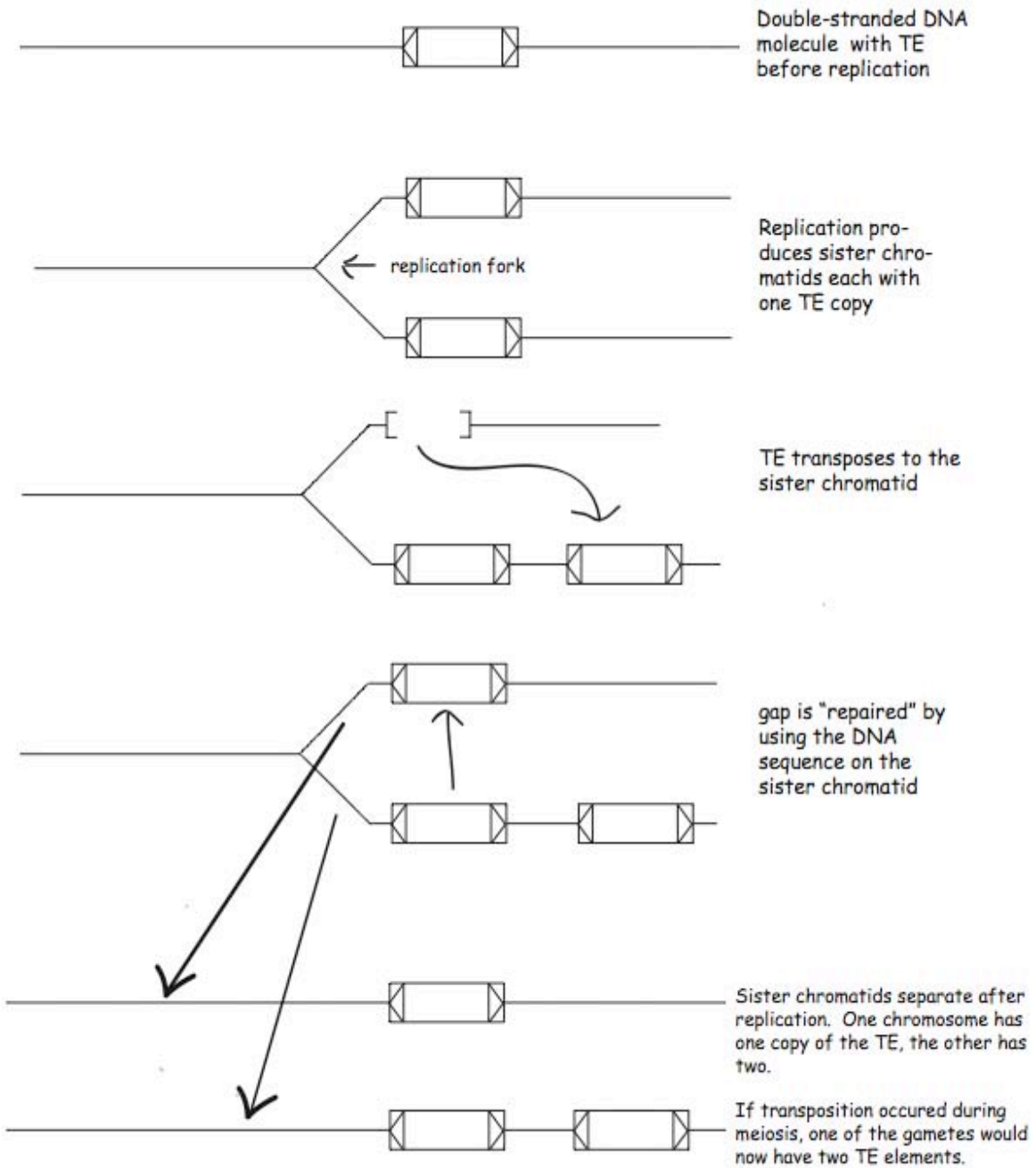
To understand how to interpret the phylogenetic trees of TEs that you will generate, it is important to understand how DNA elements increase their copy numbers in the genome. In short, we need to know how all the TE sequences arose that you identify in genomes. The mechanism of TE transposition was first discussed on page 5 figure 5. The relevant figure is "duplicated" below....



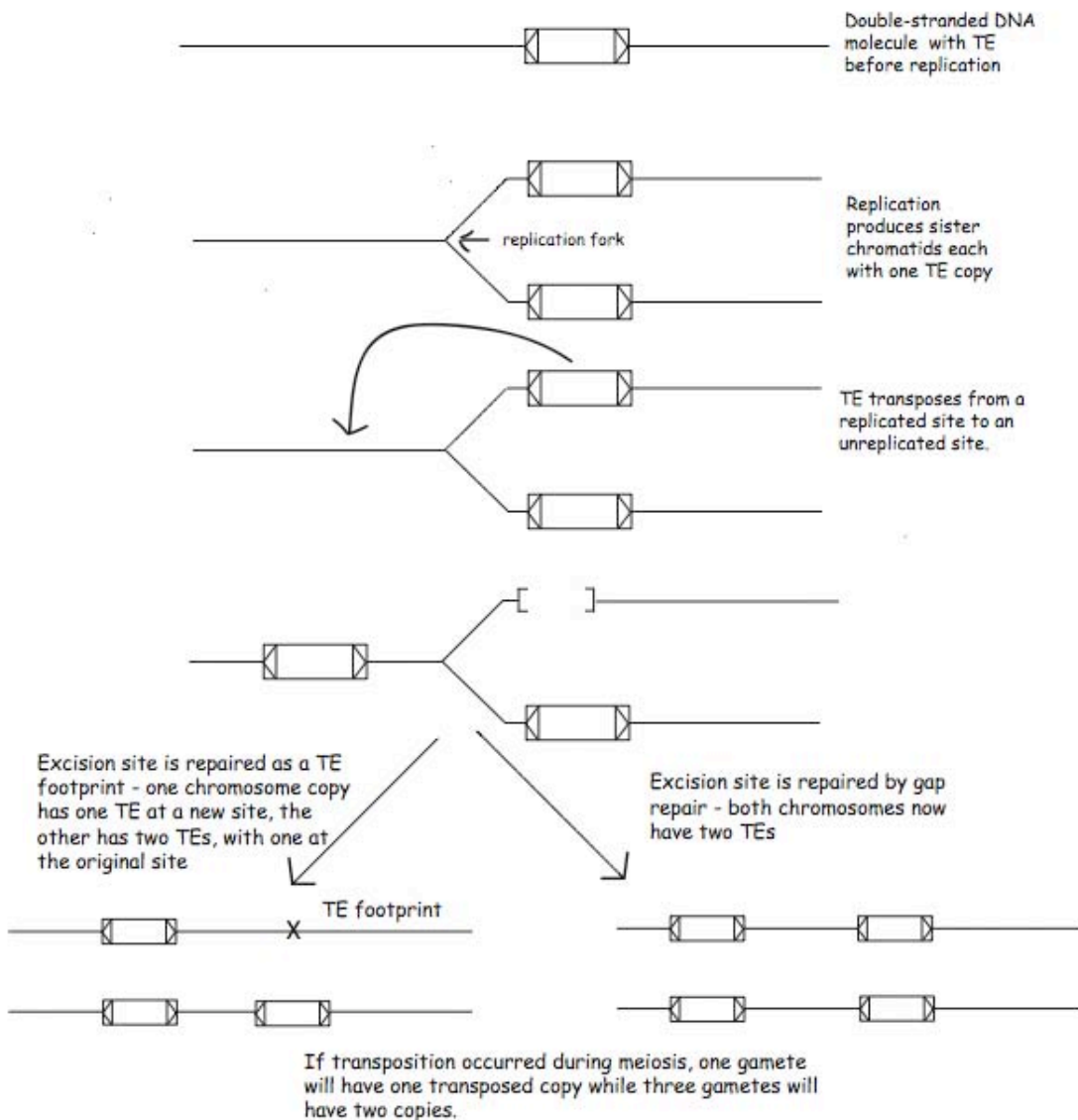
This figure shows that an autonomous element (in this case, Ac) encodes a transposase protein that binds to the ends of both itself and non-autonomous elements in its family (in this case, Ds) and catalyzes both element excision and reinsertion. As such, the element itself is the intermediate in transposition. Stated in another way, class 2 elements move via a DNA intermediate.

However, this figure does not explain how class 2 elements like Ac and Ping can increase their copy number during transposition. According to the above figure, Ac and Ds elements move from one site in the genome to another without making a duplicate copy. DNA transposons can also make duplicate copies when transposition occurs during DNA replication. The two ways they can do this are shown below:

1. Gap repair using the sister chromatid to repair the excision site....



2. Transposition from a replicated site to an unreplicated site, which is then replicated:



What you should note is that no matter which way a class 2 element moves, the element and its duplicate(s) are identical. Over time (evolutionary time that is), the element sequences mutate independently (e.g. by errors introduced during DNA replication). Elements accumulate mutations over time (they diverge). Thus, the extent of sequence divergence between elements is a measure of the time since duplication. In the figure you also see the term - transposon footprint. This will be discussed briefly in the class but in much more detail later as it will be the basis of at least one experiment you will be doing.

Back to phylogenetic trees

What are phylogenetic trees?

Here we have a graphical representation of a phylogenetic tree. Notice the terms and what they refer to.

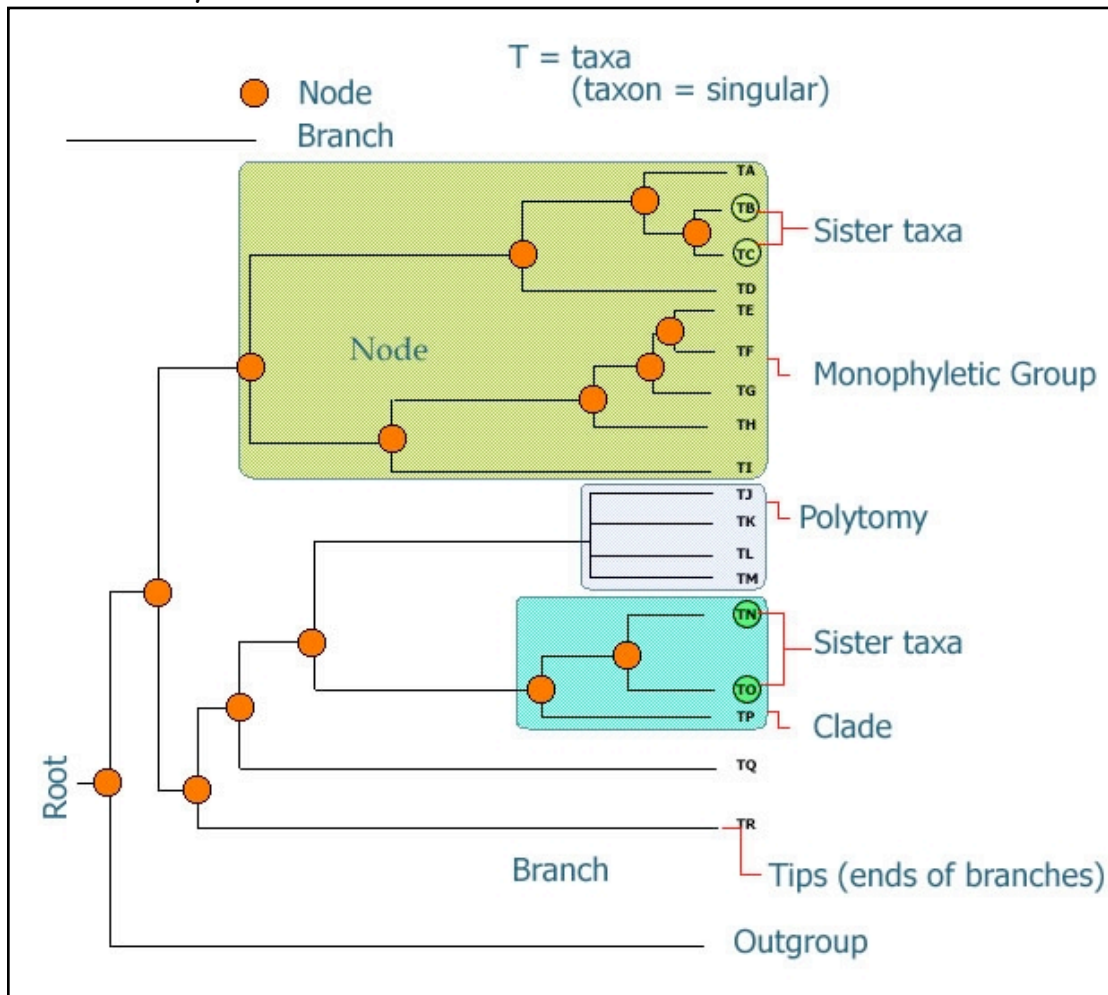


Figure 6: This figure shows a graphical representation of a phylogeny. The important features of the phylogenetic tree are

Before you can interpret a tree you need to understand some terminology. The tips of a tree represent the sampled individuals. These units are called taxa (taxon = singular). We use the term taxa to refer to any level of organization or any named group of organisms. A taxon can represent all individuals in a defined species, a single individual, or a specific amino acid or nucleotide sequence. In a species tree these represent the living organisms that were sampled to reconstruct the phylogeny. In **figure 6** our species are labeled taxa A-R. Other types of trees can

be made using specific genes or gene families, or in our case these input taxa represent the TE sequences that are obtained from a database (as in **Figure 5** on page 59). The individual members of the tree are placed on horizontal lines called branches and branches intersect to form nodes. A node represents the common ancestor (in this case the last common sequence) shared by all members that branch from that node. In figure 6 the nodes are labeled with orange dots. Ancestral node sequences are inferred based on the extant (existing) sequences. These nodes represent what the last common ancestor of that group 'looked like' to the best of our knowledge. We can infer the sequence of the nodes using the information we have from the tips. The ancestral sequence is a best guess based on the available data.

When looking at a tree we are able to visualize the relatedness of the individual members that make it up. Individuals that are placed next to each other on the tree (they are connected by only one node) are called sister taxa. In our case the sister taxa on our transposon trees represent the sequences with the highest level of sequence identity (they are most similar to each other).

All members that arise from the same node are said to be in a clade (also called lineages). If all members of a group occur in the same clade the group is said to be monophyletic. In **Figure 5**, osPong7b, osPong7c and osPong7a form a monophyletic group because they all share a common ancestor, which is represented by the node that they all branch from in the tree. If all members of a defined group are not included in a single clade then the group is considered either polyphyletic or paraphyletic. **Figure 7** shows us the distinction between these two states. In the polyphyletic situation all members of the group do not share a most recent common ancestor. In the paraphyletic case some but not all of the descendants from a most common recent ancestor are included in the group.

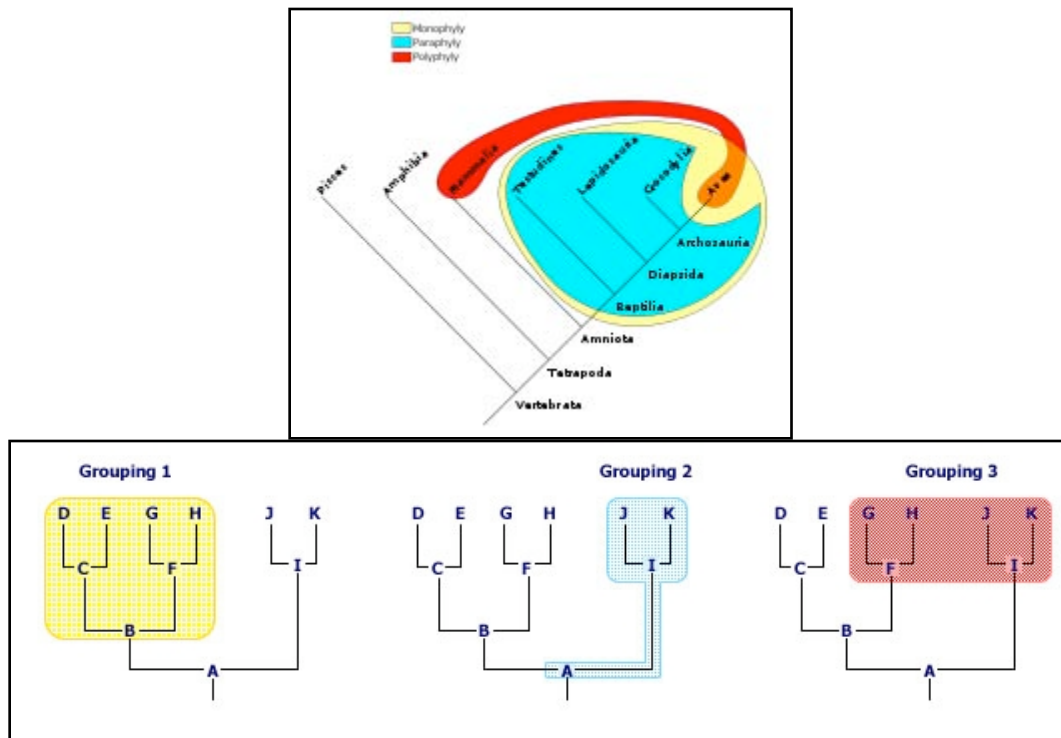


Figure 7: Monophyletic groups are highlighted in yellow, paraphyletic groups are highlighted in blue, and polyphyletic groups are highlighted in red. The tree of the vertebrates gives us an example of a monophyletic group, the sauropsids, a paraphyletic group, the reptiles, and a polyphyletic group, the warm-blooded animals.

In some trees the actual length of the branches connecting the TEs (or species) represent the number of base pair changes over time. So, long branches represent many changes while short branches represent few changes. Branches where more than one tip emerges from one node, are called polytomies. If we find polytomies in our transposable element trees, we can assume that the elements placed in the polytomy were very recently active, as all of the sequences are virtually (or are exactly) identical. An example of this can be seen in a magnified section of Figure 5 from page 59. This example is presented below in **Figure 8**. In the case of the Ponga, Pongb, Pongc, Pongd, and Ponge clade the branches were too short to draw because these are virtually identical copies in the rice genome, and thus they are represented as a polytomy. We can use this information to design new experiments.

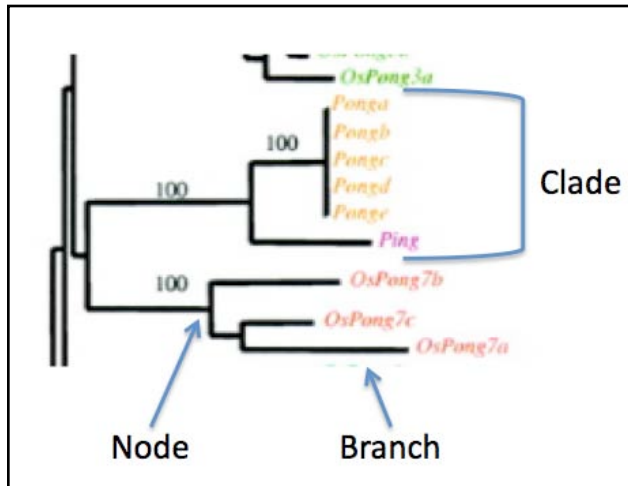


Figure 8. A magnified view of Fig 5 that includes the elements most closely related to the Ping element in the rice genome.

How are phylogenetic trees constructed?

Generally, trees are constructed by identifying shared derived characters, also known as synapomorphies. These characters can be, morphological (e.g beak dimensions or the presence of a hinged jaw), developmental (e.g presence of a developmental stage such as gastrulation), or DNA or amino acid sequences. In our case we will use the transposase amino acid sequence as the basis for our comparisons.

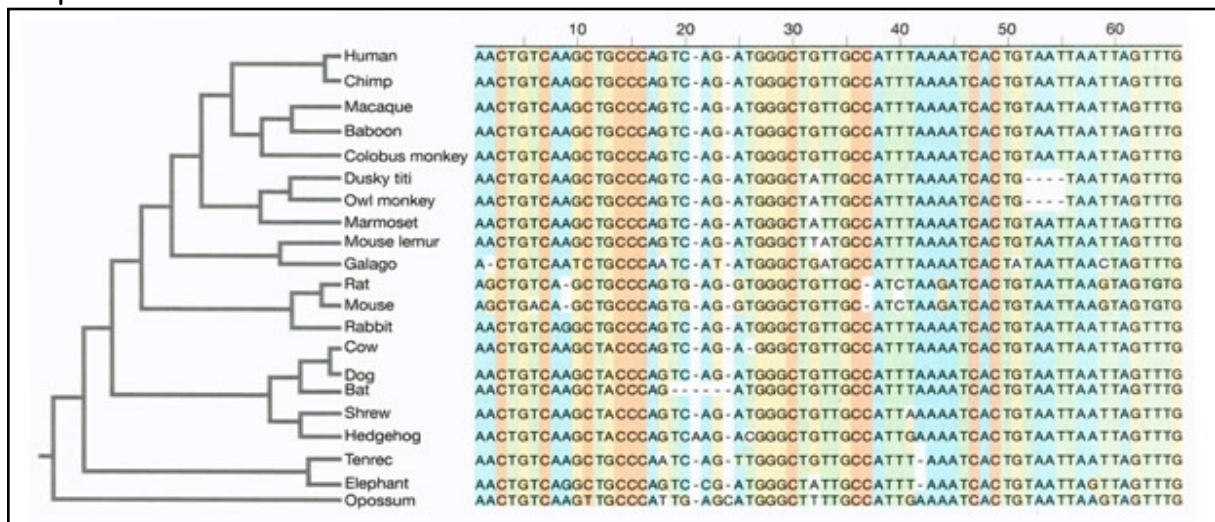


Figure 9: Sequence alignments show the relationships between the mammals. The nucleotide sequences alignment visually demonstrates nucleotide similarities and differences while also showing the presence of gaps in the sequences. The level of similarity between the sequences is used to reconstruct the phylogeny.

When reconstructing a phylogeny we first collect our data and assign similarity between the individuals based on how many characters differ between them. In our example we would place the specific sequence for each individual into a table, with each individual's sequence in a separate row. The nucleotides (or amino acids) in these rows are then aligned with one another such that each position in the alignment is counted as a character. Any deletions, insertions or base pair differences between the individual's sequences are highlighted by the alignment. See **Figure 9** above for an example of this process. Once the alignments are complete, all pair wise comparisons of the sequences are made. What this means is that each sequence is used as a starting point and is compared to all other sequences present in the alignment. The differences between the sequences (point mutations, deletions and insertions) are noted and the cumulative numbers of changes between sequences are used to generate a value describing how similar the sequences are to one another. From these comparisons a distance matrix is constructed. The more similar a sequence is to another sequence the lower the distance. A sequence compared to itself would have a distance of 0, as all the characters (nucleotides or amino acids) are the same. The more differences we see between any particular comparison, the higher the distance value. Once the distance matrix is generated based on all of the pair wise comparisons a tree can be drawn.

Although this sounds simple, it is not. If we look at 2 taxa there is 1 possible tree, 3 taxa there are 3 possible trees, 4 taxa there are 15 possible trees, 5 taxa there are 105 possible trees. Once we get up to even the modest number of 10 taxa there are 34,459,425 possible trees. If we want to look at 20 taxa there are 8,200,794,532,637,891,559,375 possible trees. Even with the best computers available we cannot efficiently investigate and evaluate the likelihood of all possible trees for any reasonable data set (more than 15 individuals/sequences in the study). Our distance matrix from our multiple alignments can rule out many of these possible trees as impossible given the data, but there are still many trees that 'fit' the data.

In order to pick the best tree, programs use complex algorithms to find the tree(s) that require the fewest number of changes to explain each step in the tree. A tree with the least number of steps or changes needed to explain the relationships between the taxa is the most parsimonious tree. Parsimony simply defined just means 'less is better'. In other words the path that requires the fewest changes is the most likely answer. There are many different approaches to

generate the best tree. The most common methods include: neighbor-joining, Bayesian, and maximum likelihood methods. If you are interested, we can go into more detail about the specifics of these methods.

Trees can be rooted or un-rooted. In a rooted tree we have chosen one of the taxa (e.g. one sequence) to be the most ancestral. In a species tree you would use a moderately distantly related species as an outgroup to root the tree. For instance if you wanted to resolve the relationships between the cereals (maize, rice, sorghum, millet, rye, oats etc.) you may chose another monocot that is not in the same group, such as a lily as your outgroup. You would not want to choose *Arabidopsis thaliana*, the mustard weed, as the outgroup because it is TOO distant. *A. thaliana* belongs to the other major class of flowering plants, the dicots, making it too distant to be a reliable outgroup. Outgroups are used to define what is ancestral to all taxa under consideration. The outgroup acts as an anchor, giving the tree an evolutionary framework and orienting the tree.

When looking at a tree you will notice that there are numbers on the branches, these represent what we call bootstrap values. This is a confidence level indicator of how probable that clade is based on the data available. If a clade has a bootstrap value of 100 we can be very confident that this relationship is accurately pictured in the tree. If the bootstrap value is 60 we have less confidence in this portion of the tree. The bootstrap value is analogous to a p-value or confidence interval in statistics.

Bootstrap values are generated as follows. Let's say that we have 100 individuals in our data set. We first use all the samples to generate the best fit tree. Once we have the best fit tree we take a sub-set of the original data set, 50 individuals, and re-run the program generating a new tree. The new tree generated from the smaller sub-sample is compared to the original tree generated from all the data. The original tree is evaluated by counting the number of times the same groupings are generated in the sub-sampled data sets. If the same relationships are seen again and again then we have more confidence in their biological reality. A value of 100 indicates that the clade was generated every time the data was sampled. A value of 60 indicates that that particular clade was found 60% of the time that the data was sampled. With a bootstrap value of 60, we say that this clade would not be "well supported". This process of sub-sampling is done over and over again using a different random set of 50 individuals each time. Typically 100 to 200 bootstrap replicates are used to estimate tree reliability. The more often the

same clades are constructed using different subsamples, the higher the bootstrap value, and the more confident we are that the relationships are represented accurately. As you might imagine, generating so many trees is an enormous task that would not be possible without computers.

Using TATE to construct phylogenetic trees

Now that you have an understanding of what a phylogenetic tree is you need to learn how to construct them. We will use the TATE pipeline to construct trees. There are four steps to constructing a tree:

1. Obtain the sequences you want to compare. This is usually done with a BLAST search.
2. Reformat the Blast results into a machine friendly format called FASTA. Put together 'split hits' in the blast result.
3. Generate a multiple alignment with the sequences.
4. Construct the tree.

TATE combines all of these steps into one web page. TATE really shines at step 2 and saves you hours of busy work. The whole process using TATE requires 2-3 minutes. The results of each step are presented to you in a unified web page.

For experiment 2 we are interested in learning about close relatives of Ping. How many copies are in the rice genome? Are any of these elements active? We also want to learn more about the function of the Tases, but more on that later.

To explore the relationship of Ping to other Pong elements you will construct a phylogenetic tree similar to the one in figure 4. We will start with a protein query and use tblastn to search the rice genome. The protein query will be the TPase amino acid sequence.

```
>Ping
MSGNENQIPVSLLEFLAEDEIMDEIMDDVLHEMMVLLQSSIGDLEREAADHRLHPRKHIKRPREEAHQNLVNDYFSENPLYPSNIFRRRFRMYRPLFLRIVDALGQWSDYFTQRVDAAGRQGLSPLQKCTAAIRQLATGSGADELDEYLNKIGETTAMDAMKNFVKGIREVFGERYLRRPTVEDTERLLELGERRGFPGMFGSIDCMHWQWERCPTAWKGFTRGDQKVPTLILEAVASHDLWIWHAFVGVAGSNNDINVLSRSTVFINELKGQAPRVQYMVNGNQYNEGYFLADGIYPEWKVFAKSYRLPITEKEKLYAQHQEGARKDIERAFVGLQRRFCILKRPARLYDRGVLRDVLGCIILHNMIVEDEKEARLIEENLDLNEPSSSTVQAPEFSPDQHVPLERILEKDTSMRDLAHRRLKNDLVEHIWNKFGGGAHSSG
```

We will use TATE to construct the tree.

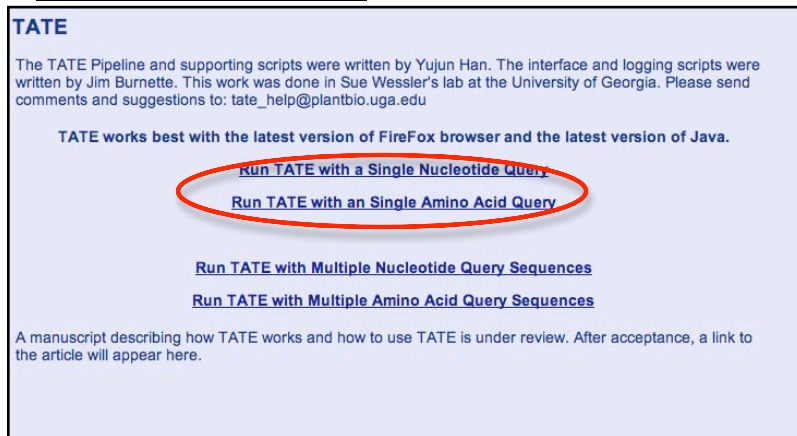
1. Open the TATE web page:

http://tate.iplantcollaborative.org/tate_research_index.html

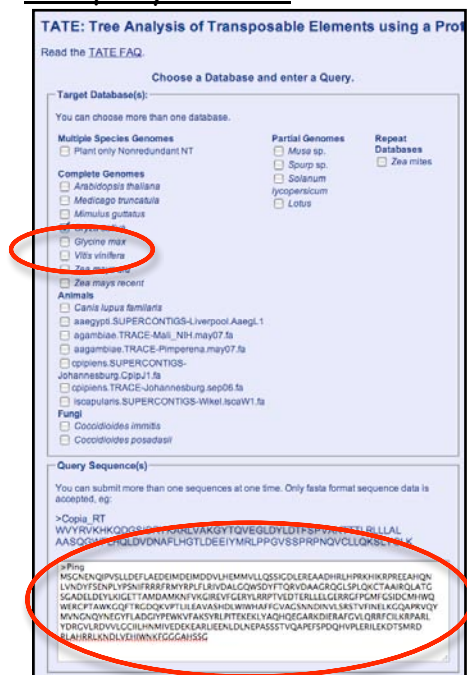
username: tate

password: collaborate

2. Select the second link



3. Select the *Oryza sativa* database and copy and paste the Ping TPase sequence in the query window.



4. Below the Query Sequence window you need to copy the Ping TPase sequence again in the "Other Sequences" Window. Click on "Enter Other Sequences" to get the window. This step is necessary to have Ping included in your tree.

Other Sequence(s)

You can submit additional sequence(s) that you want to include into the tree. These sequences are not used in the Blast query, but are added to the fasta file before the multiple alignment.

[Enter Other Sequences](#)

Only fasta format sequence data is accepted.

```
>Ping
MSCNENQIPVLLDEFLAEDIMDEIMDDVLHEMMVLQSSIGDLEREAADHRLHPRKHKRPREAHQV
LVNDYFSENPLYPNIFRRFRMYRPLFRIVDALGQWSDYFTQRVDAAGRQQLSPLQKCTAAIRQLATG
SGADELDEYKJIGETTAMDAMKNFKGRVDFGRLRRPTVEDTERLELGERRFPMPFGSDCMHWQ
WERCFATWKKQFTRODQKVPFTLILEAVASHDLWVHAFQVACSNDDIVLSRSTVFNELLKGGQAPRVQY
MVNCGNQYECYFLADGIYFENKVFVAKSYRLPITEKELYAQHQEGARKDIERAFGLQRFFCLKRPARL
YDRGVLRDVLVGLCILHNMVIEDEKARLIEENLDLNEPSSSTVQAFSPDQHVPLERILEKDTSMRD
RLAHIRLKNLDLVEHWKFKGGGAHSSG
```

5. In the right-hand you should fill out the Run Name and Notes textboxes. This will help you document your trees later.

Steps of TATE and optional settings

1. Blast:
blastn
[Modify Blast Parameters](#)

2. PHI:
Re-formatting of the Blast Results.
[Modify PHI Parameters](#)

3. Multiple Alignment:
Multiple Alignment of the Blast Results
[Modify MUSCLE Parameters](#)

4. Phylogenetic Tree:
Tree Calculation
[Modify TreeReST Parameters](#)

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:

Notes

Run TATE
Choose a stopping point:
 → → →

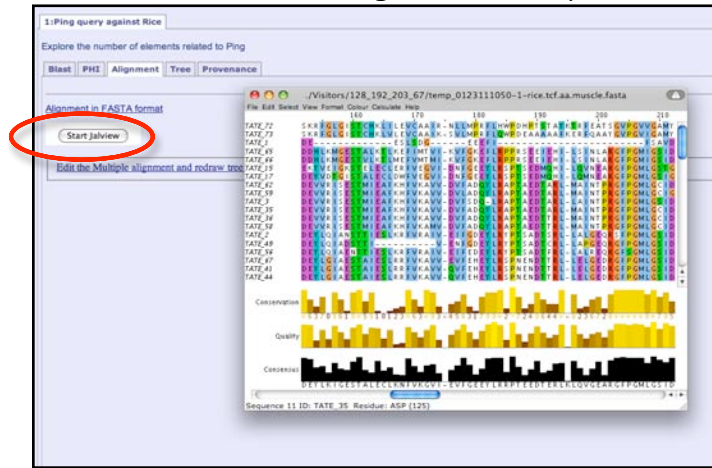
The other sections in this column allow you to modify the programs that perform each step of TATE. For now the defaults are fine. Later we will discuss when you need to modify these parameters.

6. Click the **Tree** Button.

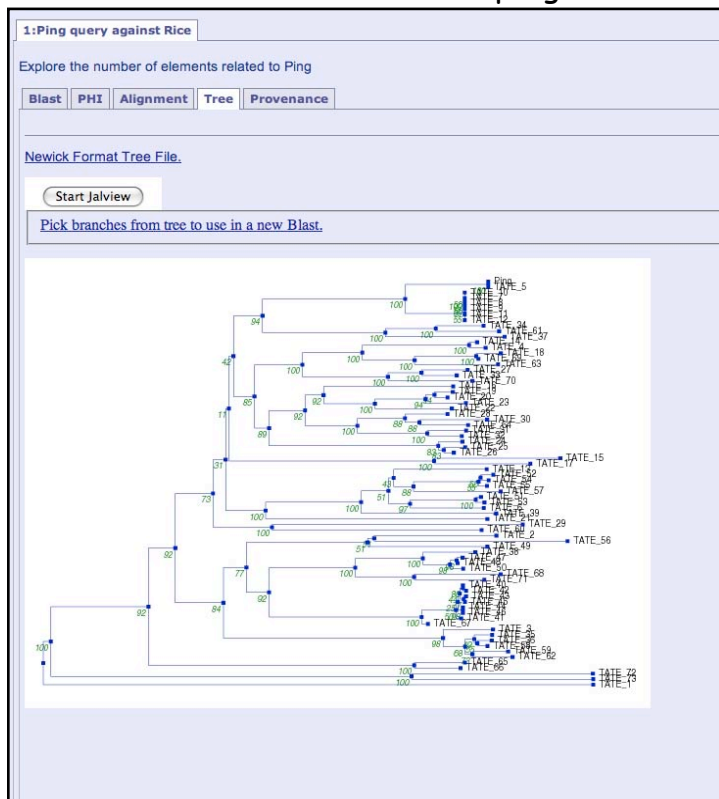
Your query will be submitted. Be patient it will take a couple of minutes.

7. The results of the TATE run are organized in a set of tabs. Each step gets a tab. There is a fifth tab called Provenance. You can download the files generated by TATE on this tab. If you included a Run Name and Note you will also see them.

10. Multiple Alignment tab. The multiple alignment program (called Muscle) uses the FASTA file generated by PHI. To see the alignment you need to click on the "Start Jalview" button. The alignment will open in a new window.



11. You will find the tree constructed from the multiple alignment on the tree tab. The tree was constructed with a program called TreeBest.



If you compare this tree with the tree in figure 4 you will see they are very similar. There are some differences. The tree in figure 4 is rooted. That means a

very distantly related TE called PIF was included. This provides an ancestral node for the rest of the elements. The tree generated by TATE is unrooted. While rooting can help analyze a tree, it is not necessary for our purposes.

Another difference is that there are more members of the Ponga-Ponge clade. This is probably due to the fact you searched a more recent version of the rice genome sequence than was available in 2004.

Dr. Nathan Hancock cloned several Pong elements closely related to Ping while he was a postdoc in Dr. Wessler's lab. He has provided these cloned elements to the class for you to test. These elements are Pong, Pong2c, Pong7c and Pong8b. He also provided us with nonautonomous versions of the elements so we can test them in the yeast excision assay. Unfortunately the names on the tree generated by TATE do not correspond to these names. How would you repeat the TATE search to find out where these elements are on the tree? Does your new tree look like the one in figure 4?

TPase protein sequence of cloned Pong-like elements.

>Pong8b

```
MSSKSPHQSSSEDDSSSSDYLEELILEEINDPMEAEIEDEIEAQLQAQMQAQQTGHSNRRGGYKRRYINR
DYQDDHNRLFAKYYSNDPLYTDDQFRRRFRMRKHLFLHIVEALGIWSPYFRLRRDAFGKVGLSPLQKCTA
AIRMLAYGTTPADLMDETFGVAESTAMECMINFVQGVRRHIFGQOYLKRPNEQDIQCLLOQGEAHGFFGILG
SLDCMHWEWQNCPPVAWKQGFTRGDYGVPTIMLEAVASADLWFHAFFGAAGSNNDINVLDQSPLFTAVLQ
GRAPSVQFTVNGTEYNGYLLADNIYPEWAFAKSIITRQSDKAKLYAQRQESARKDVERAFVGLQKRWA
IIRHPARLWERDELADIMYACIILHNMIVEDKRDDYDIPDDNTYEQSQSSVQLAGLDHGPIGHFAEVLDA
DMNIRDRTTHRRLKSDLMEHIWQYGGQQQN
```

>Pong2c

```
MSDSFSYSSNSDDLDPSKVLDKYISEQNVLGSFASRIIEKMKGRFGAGRLKRQGGTIKTIIRRDHVDASH
RLVADYFAEHPLYPERMFRTRFRMHKPLFLRIVEALSQWSPYFTQRGDCSGHTSLSPLOKCTAALRMLAY
GTPADALDEYLGKSTALECLEMFSRGIIVVFGGTYLRRPTREDVEHILHVNESRGGFFGMLGSDCMHW
RWESCPRAWRGQFTRGDYKVPTIILEAVASHDLWIWHAFVAGSNNDINVLNQSPFLDTRVGEAPRVH
YYVNGEYNHGYLLDGIYPEWAVFQKTIPLPQIEKHKLYAEHQEGARKDVERAFVGLQARFNI VRRSAK
KWKRSIGNTMLACVILHNMIVEDEGEDAICDDLNRIPRTSIVLPPEVTSGGNPCRFDVLSRKAARIR
SMHTQLKTDLIEHIWNRFRNMORA
```

>Pong7c

```
MHLFLILHLTTIYCVNPFLLQLNTINCVNPFDPDIFMCVYIYPLAHTFSPSSAFHILSSLFTLPSLI
LNTMSNQSDGDSPTHDDSLDEVSSIDPMDLYPLDEISNILGLADHVVAELKSEVEALQDMRPTROSGPR
RYVDRPYEESKHGLLKDYFVQNPVYNDTTFRRFRMRKHLFLRIVEALGQWDKYFTLRMDALNRPGLSPL
KKCTSAICQLGNGSPADQLDEYLNIGDSTTVECLKMFVKGVIEVFVGAEYLRRPMVQDVERLVQIGERRGF
PGMLGSDCMHWHWEKCPVAWKEMYTRGNQGVPTVILEAVASHDRWIWHAFVAGSNNDINVLNQSPFL
VQQLRGEQPQVQYHVNGRQYNTGYLADGIYPEWAVFVKSIRHPQSEKHKLFKAKHQEGKWKDVECAFGIL
QSRFSILKRPARLYDQGDLENIMLACIILHNMIVEDEKDIEQLPLDLNETPSTLTV
```

>Pong

```
MQSLAISLLSETHSLFSHTKTSSLLSLLFLSSSKMSEQNTDGSQVPVNLDEFLEAEDEIIDDLLTEATV
VVQSTIEGLQNEASDRHRHPRKHIKRPREEAHQQLVNDYFSENPLYPSKIFRRRFRMSRPLFLRIVEALG
QWSVYFTQRVDVAVNRKGLSPLQKCTAAIRQLATGSGADELDEYLIKGETTAMEAMKNFVKGLQDVFGERY
LRRPTMEDTERLLQLGEKRGFPGMFGSIDCMHWHWERCPVAWKQGFTRGDQKVP TLILEAVASHDLWIWH
AFFGAAGSNNDINVLNQSTVFIKELKGQAPRVQYVNGNQYNTGYFLADGIYPEWAVFVKSIRLPNTEKE
KLYADMQEGARKDIERAFVGLQRRFCILKRPARLYDRGVLRDVVLACIILHNMIVEDEKETRIIEEDL
NVPSSSTVQEPFESPEQNTPFDRVLEKDISIRDRAAHNRLKKDLVEHIWKNFGGAAHRTGN
```

>Ping

```
MSGNENQIPVSLLEDEFLEAEDEIMDEIMDDVLHEMMVLLQSSIGDLEREADHRLHPRKHIKRPREEAHQ
```

LVNDYFSENPLYPSNIFRRRFRMYRPLFLRIVDALGQWSDYFTQRVDAAGROGLSPLQKCTAAIRQLATG
 SGADELDEYLKIGETTAMDAMKNFVKGIREVFGERYLRRPTVEDTERLLELGERRGFPGMFGSIDCMHWQ
 WERCPTAWKGFTRGDQKVPTLILEAVASHDLWIWHAFFGVAGSNNDINVLSRSTVFINELKGQAPRVQY
 MVNGNQYNEGYFLADGIYPEWKVFAKSYRLPITEKEKLYAQHQEGARKDIERAFVGLQRRFCILKRPARL
 YDRGVLRDVLGCIILHNMIVEDEKEARLIEENLDLNEPASSSTVQAPEFSPDQHVPLERILEKDTSMRD
 RLARRRLKNDLVEHIWNKFGGGAHSSG

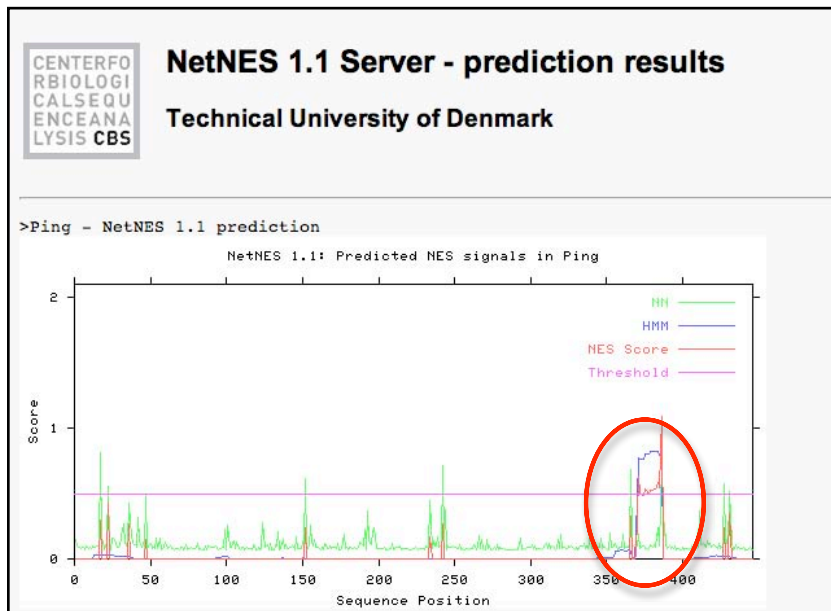
In Experiment 2 you will be testing whether these three new Pong elements are active. In addition, you will be retesting Ping and Pong that are already known to be active (so they are controls). In Experiment 2 we will not be using Arabidopsis because it would take all semester just to transform the plants. You would never get to do any interesting biology! Instead, we will use a yeast transposition assay. You can transform yeast in an afternoon, pick transformed colonies in four days, and start the excision assay that day. In just two weeks we will know which if any of these elements is active.

Dr. Nathan Hancock was a postdoctoral associate in the Wessler lab and his project was to develop a transposition assay for Ping and Ping in yeast. After months and months of trying without success, Nathan noticed something interesting about the sequence of the Ping TPase, something he thought might be the reason why he could not get the assay to work. He noticed that there was a nuclear export signal (NES) in the amino acid sequence of the TPase. When present in a protein, a NES leads to the export of the protein out of the nucleus.

This observation provided a possible reason for why the assay was not working. TPase is required in the nucleus to excise and move TEs. However, if Ping was actively exported out of the nucleus, it would not be able to catalyze transposition. So Nathan hypothesized that mutating the NES would cause an increase in the number of excision events. To do this, Nathan "destroyed" the NES by changing two leucines to two alanines. He saw that this mutant TPase was able to catalyze transposition with a higher frequency. He chose these leucines because they were previously shown to be the major component of the NES. This brings us to the second question being addressed in Experiment 2: Do the TEs that you will be testing, (Pong, Pong2c, 7c, and 8b) have identifiable NESs?

To do begin to address this question we first need to determine the location of the predicted NES in Ping. This is done using the NET NES website:
<http://www.cbs.dtu.dk/services/NetNES/>. Copy and paste the Ping protein sequence in the textbox. Click Submit.

This program compares a sequence, in this case Ping, to experimentally verified NESs. The output is a plot of the prediction score. When the score is above the threshold (the pink horizontal line in the graph), your sequence has a predicted NES.



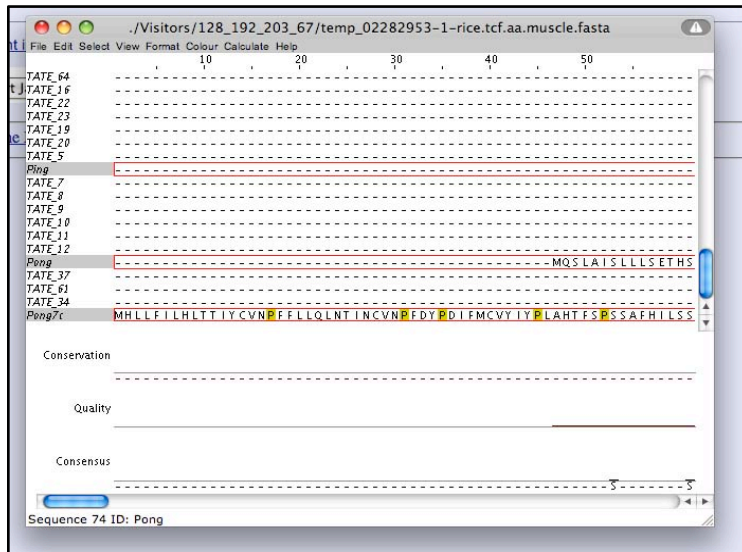
Now, how would you compare the other Pong sequences to Ping to see if the predicted NES is conserved?

If the sequence is not conserved, how could you go about analyzing the sequence to see if there is a putative NES, but in a different location?

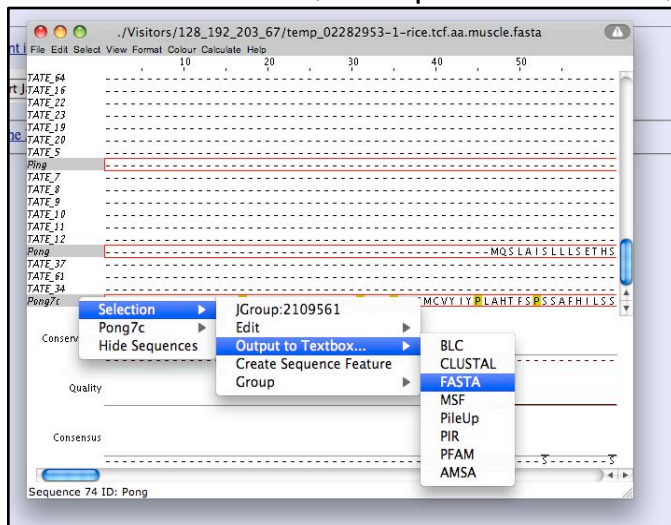
Based on these types of analyses, Nathan generated NES mutants of Pong2c and Pong7c. He called them LALA mutants.

To view a subalignment of selected sequences follow these steps. These instructions apply to the Mac keyboard and mouse.

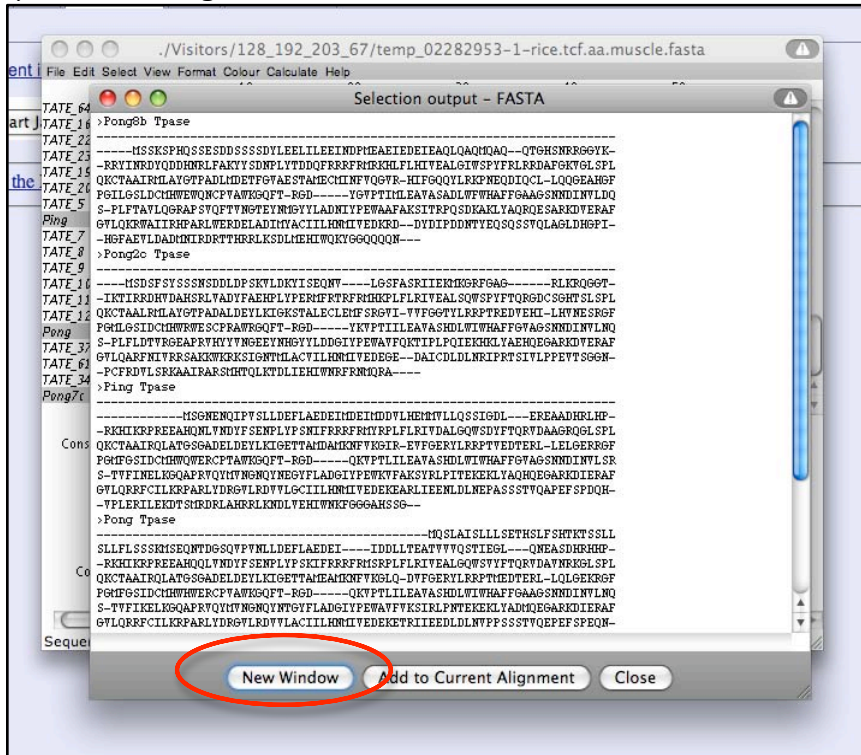
1. Go to the Alignment tab and start Jalview.
2. Use "Control + left click" to select individual sequences.



3. "Command + left click" on one of the selected sequences. Go through the menus and select "Selection," "Output to Textbox...," and "FASTA".



4. A new window will open with the selected sequences. Click on "New Window" to open a new alignment viewer.



Yeast Transposition Assay

Before you can fully understand Experiment 2 you need to first learn some yeast genetics. Unlike Arabidopsis, where you can use antibiotics to select for transformants, you use genetic complementation to select for transformed yeast.

It is easy to get yeast to take up foreign DNA (a process called transformation). If the foreign DNA is a plasmid that contains a yeast centromere then it will be stably maintained from generation to generation (as the cells divide). We will use plasmids to introduce genes and transposable elements from rice and other organisms into yeast. These genes will be expressed using yeast promoters. In many cases the expressed gene will make a functional protein and we can study its function in the relatively simple yeast. The process of transformation and gene expression can take as little as seven days. Compare this to the months required to transform a plant!

The experiments we will be doing require three plasmids: a reporter plasmid, a TPase plasmid, and an ORF1 plasmid.

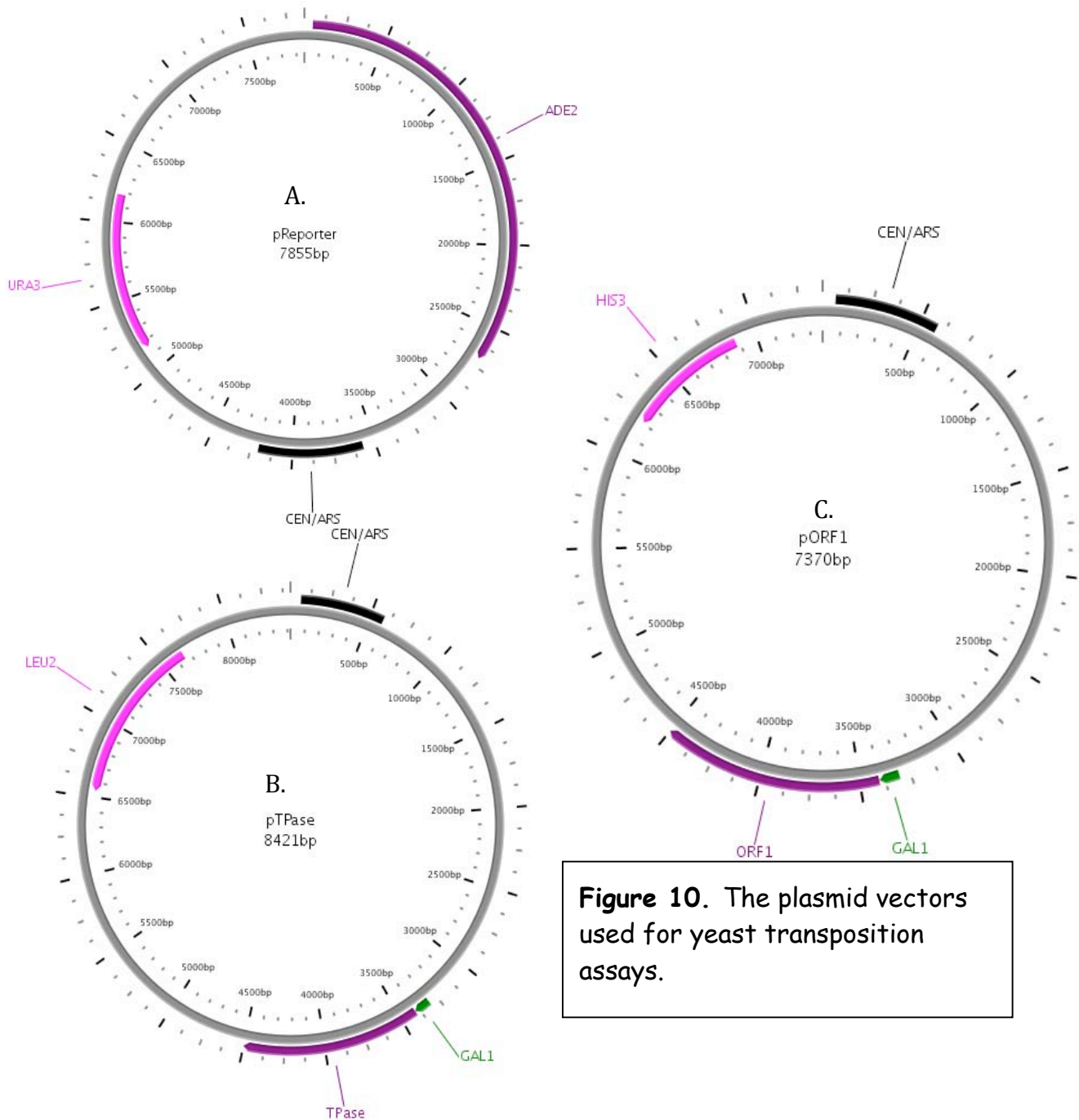


Figure 10. The plasmid vectors used for yeast transposition assays.

Each of these plasmids has a *CEN/ARS* sequence (black box in figure 10). This sequence is a yeast centromere and is required so that each daughter cell receives one plasmid following cell division. *ARS* stands for Autonomous Replication Sequence and is the site where DNA polymerase binds and initiates plasmid replication. The *CEN/ARS* sequence is required to maintain the plasmid(s) through many cell generations.

Genetic complementation in yeast - a substitute for antibiotic selection

Recall that in our first experiment the *Arabidopsis* plants you analyzed had been transformed (previously in the Wessler lab) with T-DNAs that contained an antibiotic resistance gene (see Figure 10, page 18). In this way, one could grow *Arabidopsis* seedlings on large petri plates containing the antibiotic and only plants that had the T-DNA (with the antibiotic resistance gene and other genes such as GFP and transposase) could grow. As mentioned in class, most wild type yeast can grow on plates with antibiotics. So we need to come up with another "selectable" marker whereby only yeast transformed with a plasmid that has this gene will be able to grow on plates (recall that transformation is very inefficient - less than .01% of cells in a test tube are competent to take up the plasmid from the media). In this regard, the most important genetic concept to understand for Experiment 2 and for the future experiments is complementation. For experiments with yeast, most complementation involves nutritional deficiencies. Yeast can take up nutrients such as vitamins, amino acids and nucleotides from the growth media. Yeast is also able to synthesize some amino acids and all nucleotides if they are not available in the growth media. For example if wild type yeast is placed in a medium that lacks uracil it will produce enzymes to synthesize uracil from precursor compounds. This will allow the yeast to grow and multiply. If there is a mutant allele for any of the uracil biosynthetic enzymes, the yeast strain is said to be a uracil auxotroph: it cannot synthesize its own uracil and it is designated Ura^- . If the Ura^- yeast is placed in rich medium such as YPD (contains all nutrients) the yeast will be able to grow. If the Ura^- strain is placed in medium that lacks uracil it will not be able to grow.

A common Ura^- strain used in the lab cannot synthesize uracil because it has a mutant *URA3* gene. We can take advantage of this mutant by introducing a plasmid that contains a functional *URA3* gene (shown in pink on pReporter, figure 10A). We can guarantee that the plasmid will be taken up and maintained by always growing the yeast in media lacking uracil. If the yeast is grown in medium with uracil there is a pretty good chance that it will lose the plasmid (there is no "selection pressure" to keep it - why make something if you can get it for free). The *URA3* gene on the plasmid is said to complement the *ura3* mutant gene in the chromosome. This works to our advantage because any other gene we put on the plasmid will also be maintained in the cell (just like the other genes that were on the T-DNA with the antibiotic resistance gene).

Because you will be transforming multiple plasmids into a single yeast colony, you will need multiple nutritional markers to complement (much like in Expt 1 where different T-DNAs had to have different antibiotic resistance genes). To this end, you will be working with a yeast strain that is mutant for *leu2* and *his3*. Yeast that takes up pTPase and pORF1 will be able to grow on media lacking leucine and histidine because pTPase has a *LEU2* gene and pORF1 has a *HIS3* gene, pink boxes figure 10B and C.

Yeast gene and protein nomenclature.

It is very useful to learn the nomenclature for yeast gene and protein names. Almost all gene names are three letters with a number designation. This number does not indicate any order to a pathway. The gene *URA3* will be used as an example:

- *URA3*--Wild Type of functional allele--all capital letters, italicized

- *ura3*--Mutant allele--all lower-case letters, italicized

This name will be followed by an allele designation, e.g., *ura3-1*.

- Ura3--Protein name--First letter capital, not italicized

- Ura--The phenotype of a *ura3* mutant--First letter capital, not italicized, no number

The different reporter in this experiment: ADE2 instead of GFP

To detect transposition in yeast we need a reporter system similar to the system you used in Arabidopsis with GFP. The difference is that the 'report' is the ability to grow and divide. Here's how it works. The yeast stain you will use is an *ade2* mutant. *ADE2* is part of a pathway that synthesizes adenine. The pReporter plasmid (figure 6A) has an *ADE2* gene on it, but we put the nonautonomous TE inside the *ADE2*. This makes the gene nonfunctional and it will not complement the *ade2* chromosomal mutation. When TPase and ORF1 proteins are present the TE may excise and repair the *ADE2* gene on the plasmid. If this happens then the yeast will be able to grow on media lacking adenine. In this way we will recover only yeast where the TE has excised from *ADE2*. This is called genetic selection. We select for *ADE2* reversion and this indicates TE activity. This is a very powerful system because transposition may happen in only 1 in a million cells.

Yeast that are mutant for *ade2* turn pink. This is due to a build up of the substrate for the *Ade2* enzyme. This substrate is turned into a pink pigment due to some

secondary biochemical mechanism. This is useful because you can easily tell whether or not the yeast strain is transformed with the pReporter plasmid.

The first step in Experiment 2 is to introduce the three plasmids into the yeast strain DG2523. The genetic background of this strain is *ade2-1*, *his3-1*, *leu2-1*, and *ura3-52*. All other genes are assumed to be wild type. That is, they have normal function.

After we have made the yeast competent (able to take up foreign DNA), you will plate the yeast on a selective medium called DOB(dex), csm -his, -leu, -ura. DOB means the media is defined. This means that every component was weighed out including the amino acids, vitamins, minerals, and nucleotides. This is different than the rich media YPD, which is an extract of yeast itself. The dex part means the carbon source is dextrose (glucose). This media has all 20 amino acids EXCEPT histidine and leucine, and all nucleotides EXCEPT uracil. Yeast must synthesize these three nutrients to grow on this medium. Only if a yeast cell takes up all three plasmids will it grow on these plates.

Transformation Protocol

Before you come to class, the instructors will have prepared a mid-log phase culture of DG2523. Each group will transform two sets of TEs. You will choose your two sets in class during the pre-lab discussion. Because there are five different TEs and six groups, you will share data at the end. Together we will design a table for your lab notebook to help you keep track of the different transformations you will do. Be prepared with the protocol already written into your lab notebook. The table will be added during class.

USE VERY GOOD STERILE TECHNIQUE.

Reagents:

100 mM lithium acetate (LiOAc)

PEG solution

Plasmids

Single stranded DNA pre-warmed to 50°C

Sterile H₂O

42°C water bath

1. Pellet cells 25 ml of culture in 50 ml Falcon tube for 1 min at 4000 rpm.
2. Pour off supernatant and resuspend in 20 ml sterile water.
3. Pellet cells again for 1 minute at 4000 rpm.
4. Pour off supernatant and resuspend in 0.5 ml 100 mM LiOAc.
5. Transfer yeast to a 1.5 ml tube.
6. Pellet cells and remove supernatant using a pipetter.
7. Resuspend in 225 μ l of 100 mM LiOAc.
8. Label 1.5 ml tubes.
9. Add 5.8 μ l salmon sperm DNA to each tube. (This solution is very viscous.)
10. Add 1.0 μ l of vectors to the appropriate tube.
 - a. pReproter
 - b. pORF1
 - c. pTPase
11. Mix yeast from step 7 by gentle vortex on setting 6. Add 50 μ l of yeast to each tube.
12. Briefly vortex tube on speed 6.
13. Add 400 μ l PEG. (This solution is very viscous.)
14. Briefly vortex tube on speed 6.
15. Heat shock at 42°C for 45 min.
16. Label plates. Pour 5-6 glass beads on each plate.
17. Pellet cells for 30 sec. at 7000 rpm.

18. Completely pipette off supernatant. Resuspend yeast in 100 μ l sterile water.

19. Plate all 100 μ l on DOB(dex) csm -his, -leu, -ura.

20. Incubate at 30°C 'upside' down for 3-5 days.

Tuesday, February 10, 2009 Excision Assay, Step 1.

Today you will start the excision assay. The first step is a simple one. To detect excision events you will need to plate a very large number of cells (around 10^7 - 10^8). To obtain these cell numbers you will set up liquid cultures of the transformed yeast cells.

Materials

Glass culture tubes

DOB(dex) csm -his, -leu, -ura liquid medium

Sterile loops

1. Label 3 glass tubes for each transformation you did and your initials.
2. Pipette 5 ml of DOB(dex) csm -his, -leu, -ura medium into each tube.
3. Using a sterile loop touch one colony on the transformation plate and transfer it to a glass tube.
4. Repeat for the other two tubes using a different colony each time.
5. Place the tubes in the 'spinner.' The tubes will be incubated for 2 days.

Thursday, February 12, 2009, Excision Assay, Step 2

Today you will plate the cells that grew in the liquid culture onto large petri dishes that have DOB(gal) csm -his, -leu, -ura. These plates have galactose instead of dextrose in them. The galactose will activate the *GAL1* promoter of the TPase plasmid and this will result in TPase expression. In order for the *GAL1* promoter to be fully activated, all dextrose from the liquid media must be washed away. Today you will pellet the cells from the culture media and resuspend the pellet in water to wash away any remaining medium. The cells are pelleted a second time and resuspended in 500 μ l of water. This total volume will be plated on the large inducing plates.

A second part of the experiment requires you to do a viable count on each culture. You will take 100 μ l from the wash step and dilute it into 9.9 ml of water. This

dilution will be set aside until you finish washing and plating each culture. After all of the cultures are plated, finish the dilutions and plating for the viable counts.

1. Label DOB(gal) CM-ade, -ura, -leu large Plates with the same labels you used on the glass culture tubes. Pour ~10 glass beads on each plate.

2. Label one plastic culture tube for each glass culture tube. Pour yeast culture from glass tubes to plastic ones. Vortex each culture before transferring it.

3. Label 2 glass tubes for each culture and designate one 10^{-2} and the other 10^{-4} . Put 9900 μ l (9.9 ml) sterile water in each.

4. Pellet cells at 4000 rpm for 1 minute.

5. Pour off supernatant carefully. Resuspend in 5 ml sterile water. Vortex to resuspend the pellet.

6. Transfer 100 μ l of each culture to the appropriate dilution tube. Set aside and finish dilution series after plating the cells in the plastic tube.

7. Pellet cells in plastic tube at 4000 rpm for 1 min.

8. Pour off supernatant carefully.

9. Resuspend in 500 μ l of sterile water.

10. Plate all 500 μ l on the correct plate. **Do not discard glass beads yet.**

11. Keep plates '**right-side up**' in 30°C incubator if the plate hasn't absorbed all of the water.

12. Finish dilution series to 10^{-4} . Vortex the first dilution tube and put 100 μ l of 10^{-2} dilution into the 10^{-4} dilution tube.

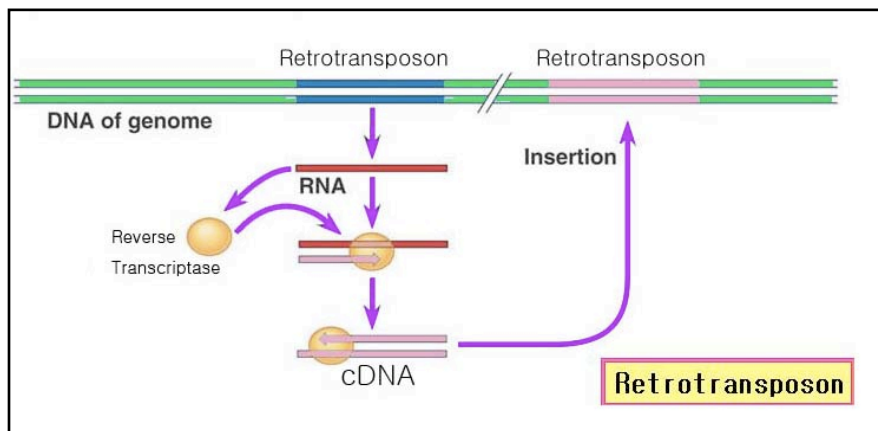
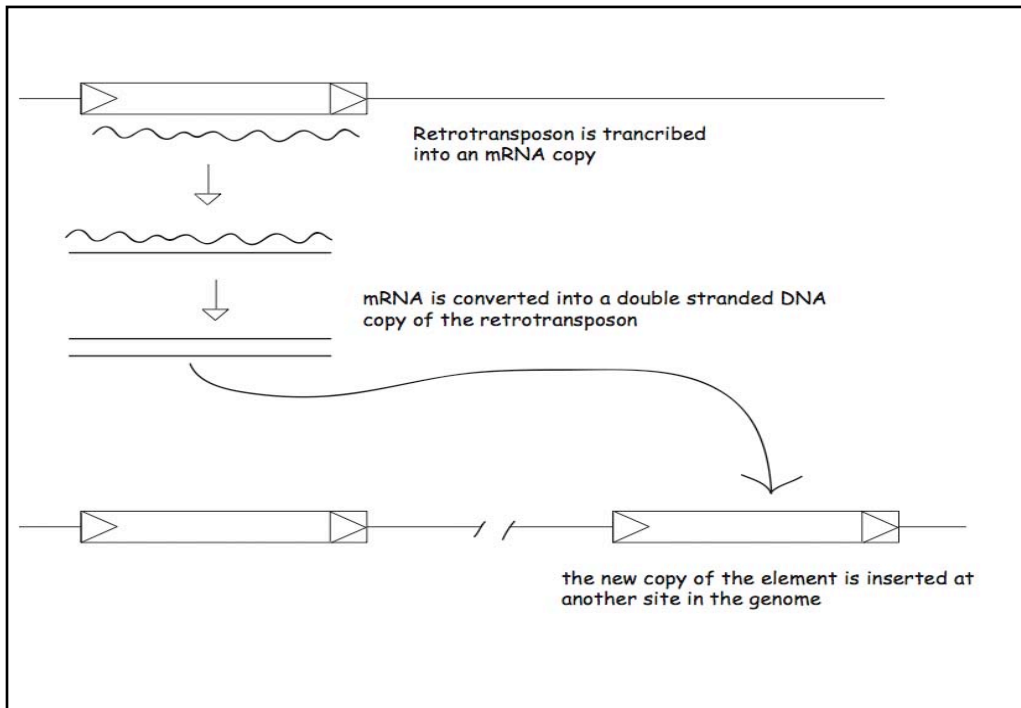
13. Vortex the tube and plate 100 μ l of the 10^{-4} dilution on YPD.

14. Pour off the glass beads from the large plates. Return to the incubator **up-side down** for 14 days. We will check the plates often during the next 14 days.

Chapter 5: Introduction to Class 1 elements: LTR Retrotransposons - the most abundant TEs in plant genomes.

How do Class 1 Retroelements (retro)transpose and make duplicate copies?

Relax! The way class 1 elements make duplicate copies is easy when compared to the duplication of class 2 elements. As you can see in the figures below, 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!



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.

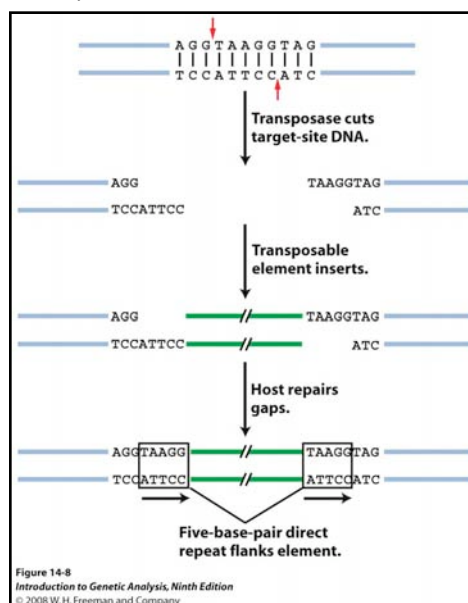
(2) like genes, a class 2 element can serve as template for many mRNA transcripts. Because each transcript has the potential to be converted into a new element, one element can produce many new elements. Class 2 elements are thus like printing presses that can potentially produce many new elements in 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.

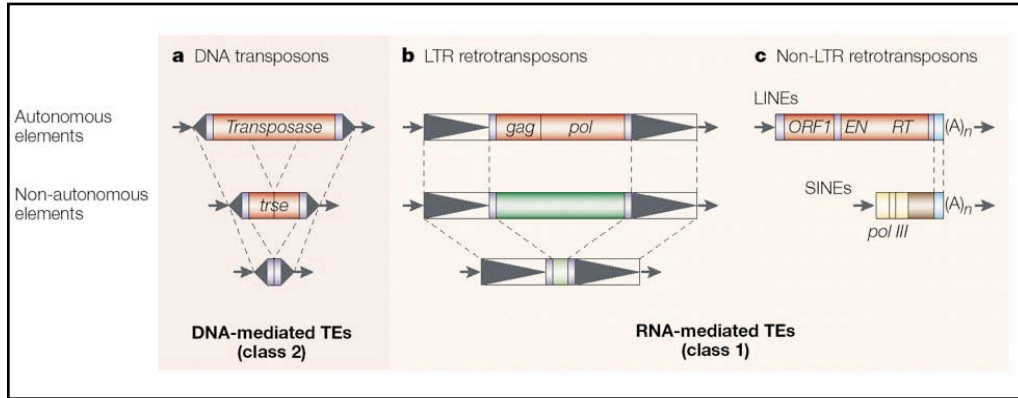
The mechanism that generates Target Site Duplications (TSDs):

(this is repeated from an earlier section as you will need to review it)

One other thing - both class 1 and class 2 elements are flanked by a target site duplication (TSD). The figure below shows how TSDs are generated during the insertion of a class 2 element. Recall that for class 2 elements, the reaction is catalyzed by the transposase. For class 1 retro elements, insertion is mediated by an enzyme called an integrase. Virtually all class 1 elements have a 5 bp TSD which means that the enzyme that cleaves the target site makes a 5 bp staggered cut in the double stranded substrate.



Both TE classes have autonomous and nonautonomous elements. Class 1 elements are divided into 2 broad types - LTR retrotransposons and Non-LTR retrotransposons. We will focus on the LTR-retros in this class as they are the most abundant TE in most characterized plant genomes:



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

Life cycle of a typical retrovirus:

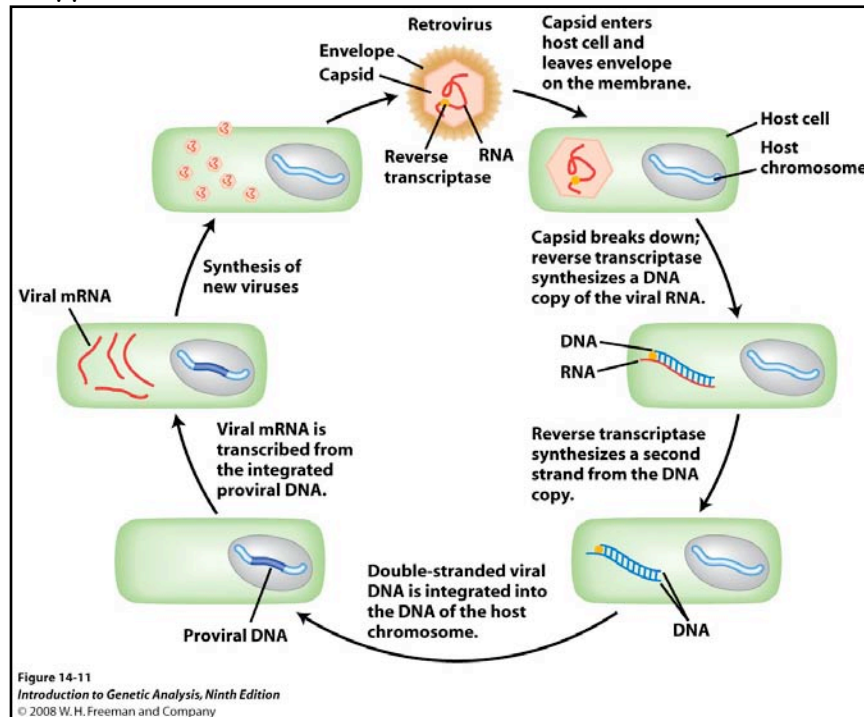
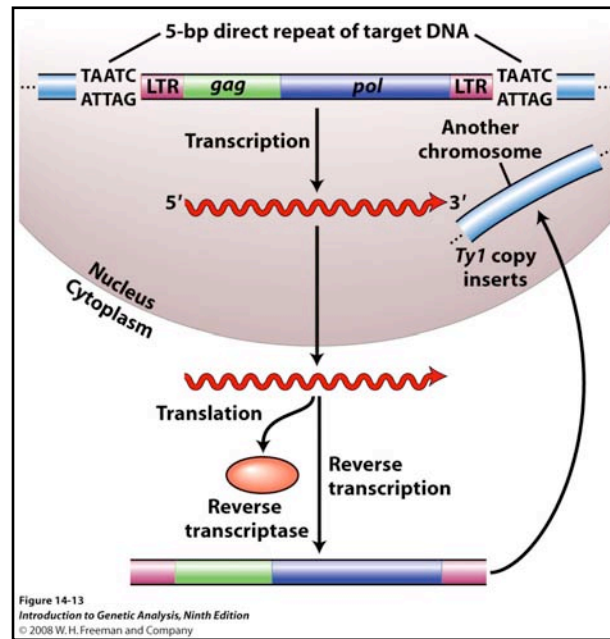
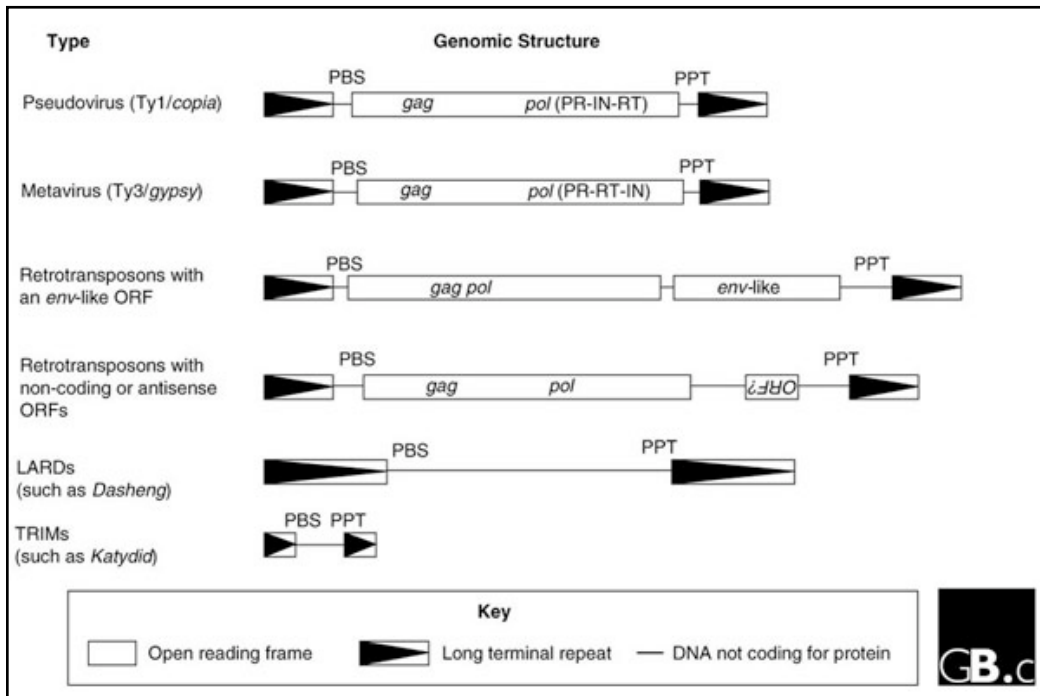


Figure 14-11
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

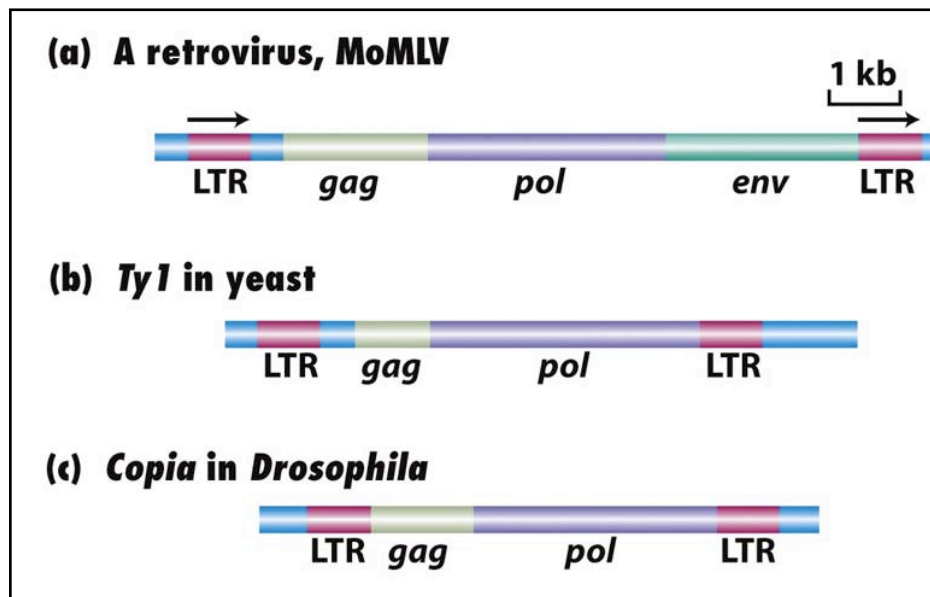
Retrotransposition of LTR retrotransposons:



LTR retrotransposons are also divided into different types...



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



Identifying complete elements: LTR Retrotransposons

Today's objective: Today you will learn how to find and characterize class 1 LTR retrotransposons (see introductory material on the prior pages). 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), which will allow us to identify all of its open reading frames (ORFs) and determine what they encode.

Before reading this, review pages 35-47 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 45-47 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

```
GGLGERVTRNRSYELVNSAFVASFEPKNVCHALSDENWVNAMHEELENFERNKVWVSLVEPPLGF
NVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILLAFASKGF
KLFQMDVKSFAFLNGVIEEEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPRAWYERLKTFLLO
NGFEMGAVDKTLFLLHSGIDFLLVQIYVDDIIIFGGSSHALVAQFSDVMSREFEMSMMGELTFFL
GLQIKQTKEGIFVHQTKYSKELLKKFDMADCKPIATPMATTSSSLGPDDEDGEEVDQREYRSMIGS
LLYLTA SRPDIHFSVCLCARFQASPR TSHRQAVKRIFRYI
```

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 a 'Database' dropdown menu set to 'Nucleotide collection (nr/nt)', an 'Organism' dropdown menu set to 'Oryza sativa (taxid:4530)', and an 'Entrez Query' text input field. There are also small icons for help and search options next to the dropdowns.

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...	749	0.0
qb AC092559.4	Oryza sativa chromosome 3 BAC OSJNBb0096M04 ge...	749	0.0
qb AC107224.2	Oryza sativa Japonica Group chromosome 3 clone...	748	0.0
emb AL606652.4	Oryza sativa genomic DNA, chromosome 4, BAC c...	748	0.0
dbj AP008210.1	Oryza sativa (japonica cultivar-group) genomi...	748	0.0
qb AC137696.2	Genomic sequence for Oryza sativa, Nipponbare ...	748	0.0
dbj AP008207.1	Oryza sativa (japonica cultivar-group) genomi...	748	0.0
dbj AP002538.2	Oryza sativa Japonica Group genomic DNA, chro...	748	0.0
dbj AP008208.1	Oryza sativa (japonica cultivar-group) genomi...	744	0.0
dbj AP005476.3	Oryza sativa Japonica Group genomic DNA, chro...	744	0.0

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

```
> 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

Query 1      GGLGERVTRNRSYELVNSAFVASFEPKKNVCHALSNDENWVNMHEELENFERNKVWVSLVEP 60
Sbjct 17511   GGLGERVTRNRSYELVNSAFVASFEPKKNVCHALSNDENWVNMHEELENFERNKVWVSLVEP 17332

Query 61     PLGFNVIGTKWVFNKKNLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILL 120
Sbjct 17331   PLGFNVIGTKWVFNKKNLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILL 17152

Query 121    AFAASKGFKLFQMDVKSAFLNGVIEEEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPR 180
Sbjct 17151   AFAASKGFKLFQMDVKSAFLNGVIEEEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPR 16972

Query 181    AWYERLKTFLQLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIFGGSSHALVAQFSDVM 240
Sbjct 16971   AWYERLKTFLQLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIFGGSSHALVAQFSDVM 16792

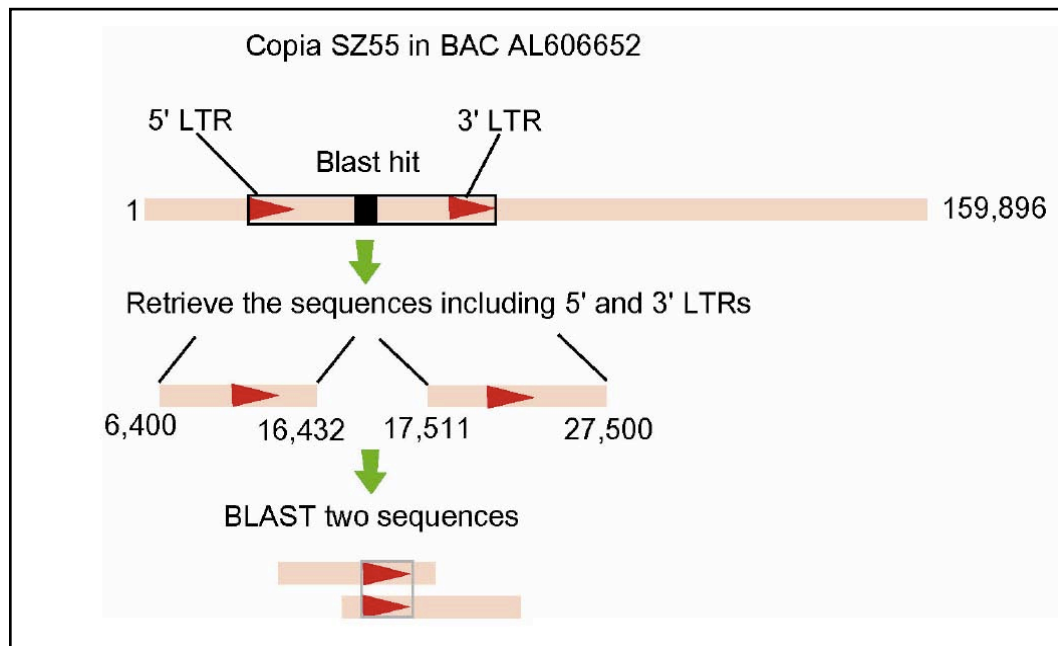
Query 241    SREFEMSMGELTFFLGLQIKQTKEGIFVHQTKYSKELLKKFDMADCKPIATPMATTSSL 300
Sbjct 16791   SREFEMSMGELTFFLGLQIKQTKEGIFVHQTKYSKELLKKFDMADCKPIATPMATTSSL 16612

Query 301    GPDEEDGEVDQREYRSMIGSLLYLTA SRPDIHFSVCLCARFQASPRTSHRQAVKRFRI 360
Sbjct 16611   GPDEEDGEVDQREYRSMIGSLLYLTA SRPDIHFSVCLCARFQASPRTSHRQAVKRFRI 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

"Command 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 159,894 bases. You will use this later.

NCBI Nucleotide

Search Nucleotide for [] Go Clear

Limits Preview/Index History Clipboard Details

Format: GenBank FASTA Graphics More Formats Download Save Links

GenBank: AL606652.4

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

Change Region Shown
Customize View

[Pick Primers](#)
Design and test primers for this sequence using Primer-BLAST.

Recent Activity
Turn Off Clear

- AL606652.4 (1) Nucleotide
- Oryza sativa Japonica Group chromosome 3 clone OSJNBa0009C08, complete sequence
- AL606652:Oryza sativa gen...

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 Group
 ORGANISM [Oryza sativa Japonica 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., et al.

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 gcggtacggt aaagaaaagc gtaagatcaa gagctaaaag attacctttc
  181 taaatcatga tcatacagaa cagagttgtc gccactagtg cgatataatt tcagcccaag
  241 atagccaatc aatggtgcc aaggattctc attacaaaac ccgtggagcc tgaatttttg
  301 gaatgaacag taagtaagct tgtatgaaca gaatctaaag tgaattttc acactaacia
  361 ttcagggtga gactgacat gaggtctcca tatcaattgg gcatccgaaa gagtaatcgg
  421 tatggacacg accgcctacg cgatctctgg actccaaaac agtcacctca aacgaagcat
  481 tggagagagc acgggctgct gcaaccctg aaattcccc accgatcacg atgacggatg
  541 gaggggaagc acattgcctc tcaatggtcg gaagcaagag gcctgaacaa 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 blastn page: <http://www.ncbi.nlm.nih.gov/blast/>.

a. Click the Blast 2 Sequences check box.

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.

The screenshot shows the NCBI BLAST interface for the 'blastn' suite. The 'Enter Query Sequence' section has a text input field containing 'AL606652' and a 'Query subrange' section with 'From' set to 6400 and 'To' set to 16432. Below this, there is an 'Or, upload file' section with a 'Choose File' button and 'no file selected' text. The 'Job Title' field contains 'AL606652:Oryza sativa genomic DNA, chromosome...'. A checkbox labeled 'Blast 2 sequences' is checked. The 'Enter Subject Sequence' section has a text input field containing 'AL606652' and a 'Subject subrange' section with 'From' set to 17511 and 'To' set to 27000. Another 'Or, upload file' section is visible at the bottom.

The results of the blast2seq will look like this.

Query	12976	TGAAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGGGTGAATATAGCAATTCAAAT	13035
Sbjct	21669	TGAAAGACCAAGAACAGCTATAGAGGG-----GGGGTGAATATAGCAATTCAAAT	21719
Query	13036	CTTGCCCCGAAAATACTCATCAAGCCGGATTTCTCAAAATCCTTACTAGAAATCGCGGCT	13095
Sbjct	21720	CTTGCCCCGAAAATACTCATCAAGCCGGATTTCTCAAAATCCTTACTAGAAATCGCGGCT	21779
Query	13096	ATTAGAGAAGCCGGATCTAGAAAAGAAGAGAGAAAAAGAAGAGAAAAGGAATTCGCCAAA	13155
Sbjct	21780	ATTAGAGAAGCCGGATCTAGAAAAGAAGAGAGAAAAAGAAGAGAAAAGGAATTCGCCAAA	21839
Query	13156	CTAGAGGAGGAAGAGAAAAGGAATTCGCCAAACTAGAAAGTGAAGTGAAGAGAGAGCA	13215
Sbjct	21840	CTAGAGGAGGAAGAGAAAAGGAATTCGCCAAACTAGAAAGTGAAGTGAAGAGAGAGCA	21899
Query	13216	AAACTCATCATCGCAAAGTTCAAATTGCAAGCCGAATTTAAATTGCGGAATTTAAATGGA	13275
Sbjct	21900	AAACTCATCATCGCAAAGTTCAAATTGCAAGCCGAATTTAAATTGCGGAATTTAAATGGA	21959
Query	13276	CAAGGCAAAATGAAATCCTTCAAATCATTTCATTTATAGGTGATGCAAAATAACCGCTC	13335
Sbjct	21960	CAAGGCAAAATGAAATCCTTCAAATCATTTCATTTATAGGTGATGCAAAATAACCGCTC	22019
Query	13336	AACTAGGAGCAAACATACACCTTCAGAGGAACATTAACACAAACTTAAAATCTCTCGGAC	13395
Sbjct	22020	AACTAGGAGCAAACATACACCTTCAGAGGAACATTAACACAAACTTAAAATCTCTCGGAC	22079
Query	13396	AAACACACTCCAAACTAATCCTAATACAAAAGCCTCTCGGGCAAACACACTCCAAACTCA	13455
Sbjct	22080	AAACACACTCCAAACTAATCCTAATACAAAAGCCTCTCGGGCAAACACACTCCAAACTCA	22139
Query	13456	CACGAAACTCTCTCACCGAGCATCTCAAATGATTACCAAAGGAGCAACCTCCACCCT	13515
Sbjct	22140	CACGAAACTCTCTCACCGAGCATCTCAAATGATTACCAAAGGAGCAACCTCCACCCT	22199
Query	13516	TGCATCCATCTCTCTATTTATAGCCTAAGACCCCTAAGACATTTCTCAAATACCCCTAG	13575
Sbjct	22200	TGCATCCATCTCTCTATTTATAGCCTAAGACCCCTAAGACATTTCTCAAATACCCCTAG	22259
Query	13576	GGCGAAACCCTAACTCAGAACAGATCTGGTCCATCCATTGTTCCCTTCTACTCAAAGGAAA	13635
Sbjct	22260	GGCGAAACCCTAACTCAGAACAGATCTGGTCCATCCATTGTTCCCTTCTACTCAAAGGAAA	22319
Query	13636	AGCTCCAGATGATTGCCACCTCATCGATCCGAACCTCACATGGTCTTCTTGCTGATGAA	13695
Sbjct	22320	AGCTCCAGATGATTGCCACCTCATCGATCCGAACCTCACATGGTCTTCTTGCTGATGAA	22379
Query	13696	TCCGCGTGCTCTCCGTCTTGACGAATCCAAACTTGCCACGTTGCCAGCCGCGTCTGCG	13755
Sbjct	22380	TCCGCGTGCTCTCCGTCTTGACGAATCCAAACTTGCCACGTTGCCAGCCGCGTCTGCG	22439
Query	13756	CGTTCCGTCTCCTTTTCTCAGCGCGCTCGCCAGCGCCCAACCAGCCGAAGCCGCCA	13815
Sbjct	22440	CGTTCCGTCTCCTTTTCTCAGCGCGCTCGCCAGCGCCCAACCAGCCGAAGCCGCCA	22499

Query	13816	CGCGTCCCGCTGAGCCGCTCCCGCGCGCGCTTGCAAAATCGCGTGTGGGTCCCGCGCTCC	13875
Sbjct	22500	CGCGTCCCGCTGAGCCGCTCCCGCGCGCGCTTGCAAAATCGCGTGTGGGTCCCGCGCTCC	22559
Query	13876	TGCACCCGCCTCCGATCGAGTCACGCGAAACGGGGGTTGCCGCGTACGGTTTTTCCGCGC	13935
Sbjct	22560	TGCACCCGCCTCCGATCGAGTCACGCGAAACGGGGGTTGCCGCGTACGGTTTTTCCGCGC	22619
Query	13936	GCGTCCTTGCGCATGCAAGCTTGCTTGCAC--TCTGAGCCCATGCGCCATGGGCCGCCGT	13993
Sbjct	22620	GCGTCCTTGCGCATGCAAGC-TGCCTGCACCTCCTGAGCCCATGCGCCATGGGCCGCCGT	22678
Query	13994	CTTGGGCCGTGCTGCTTGGACGATGCAAGGCCTGCCGAGCCGAGCCATTACCATCTTG	14053
Sbjct	22679	CTTGGGCCGTGCTGCTTGGACGATGC-AGGCCTGCCGAGCCGAGCCATTACCATCTTG	22737
Query	14054	GGCCACGTGGAACGGCTGGATTGG-TTGGCCTCCCTGCCAACCAATCACAGCGCACATGC	14112
Sbjct	22738	GGCCACGTGGAACGGCTGGATTGGCTTGGCCTCCCTGCCAACCAATCACAGCGCACATGC	22797
Query	14113	ACAGCTAGCTAG-TTGACTTTTCCACCGAGCCATGTTAGTAGCAACCAGTACAGTGCAAG	14171
Sbjct	22798	ACAGCTAGCTAGCTTGACTTTTCCACCGAGCCATGCTAGTAGCAACCAGTACAGTGCAAG	22857
Query	14172	CTCCTCCTTGACACAAGTACAGTACGTGTACATGCATGTATGCTACCTACAGCAAGTACTG	14231
Sbjct	22858	CTCCTCCTTGACACAAGTACAGTACGTGTACATGCATGTATGCTACCTACAGCAAGTACTG	22917
Query	14232	TAGCAGCAATGCACCTGCACAGTCCCTTCTGATTTCTTCGCGAATCCGATGCTTGCACAC	14291
Sbjct	22918	TAGCAGCAATGCACCTGCACAGTCCCTTCTGATTTCTTCGCGAATCCGATGCTTGCACAC	22977
Query	14292	TTGGCCTTGTGAAGCCTGTTGCAAAGACCTTTTCACACGGTGTTCGTCCACCCTGTGCAA	14351
Sbjct	22978	TTGGCCTTGTGAAGCCTGTTGCAAAGACCTTTTCACACGGTGTTCGTCCACCCTGTGCAA	23037
Query	14352	CCTTGTGTCCAATCTTGTCAACCCGGCATCCTTGATCGCTTTGGACCTCAACTCCTCCCTG	14411
Sbjct	23038	CCTTGTGTCCAATCTTGTCAACCCGGCATCCTTGATCGCTTTGGACCTCAACTCCTCCCTG	23097
Query	14412	AGTCTAGTCCCAGTCCGCGCTTGACCAAGATCGACCCCGATCACCTGCACACACATGAAC	14471
Sbjct	23098	AGTCTAGTCCCAGTCCGCGCTTGACCAAGATCGACCCCGATCACCTGCACACACATGAAC	23157
Query	14472	CAAACAACCGTTGTCTTGGCACAGATGTCGCAACCTGACCAACGTTAGTCCACACACACA	14531
Sbjct	23158	CAAACAACCGTTGTCTTGGCACAGATGTCGCAACCTGACCAACGTTAGTCCACACACACA	23217
Query	14532	CTTCTTGCACATCCGGTACTTGTCAATTTCCCATCACAAAAGAAGTATAACCACACATGG	14591
Sbjct	23218	CTTCTTGCACATCCGGTACTTGTCAATTTCCCATCACAAAAGAAGTATAACCACACATGG	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.

Click on the tan arrow "Change Region Shown" and input the start and end locations into the windows as follows:

NCBI Nucleotide

Search Nucleotide for [Go] [Clear]

Format: GenBank FASTA Graphics More Formats

GenBank: AL606652.4

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

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 Group

ORGANISM Oryza sativa Japonica Group

Change Region Shown

Whole sequence

Selected Region

from: 12976 to: 23286

Update View

Customize View

Pick Primers

Design and test primers for this sequence using Primer-BLAST.

Click Update View.

In the next window click FASTA.

NCBI Nucleotide

Search Nucleotide for [Go] [Clear]

Format: GenBank FASTA Graphics More Formats

GenBank: AL606652.4

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

Showing 10.31kb region from base 12976 to 23286.

>gi|70663936:12976-23286 Oryza sativa genomic DNA, chromosome 4, BAC clone: OSJNBb0004A17, complete sequence

TGAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGGGTGAATATAGCAATTCAAATCTTGCCCCCGG

AAATACCTCATCAGCCGGATTCTCAAATCCTTACTAGAAATCGCGGCTATTAGAGAAGCCGGATCTAG

AAAGAGAGAGAAAAGAGAGAGAAAGCAATTCGCCAACTAGAGAGGAAAGAGAAAAGGAATTCCCGA

ACTAGAAAGTGAAGTGAAGAGAGAGAGCAAACTCATCGCAAGTTCAAATTCAGAGCGAAATTTA

AAATCCCGAATTTAAATGCACAGGCCAAAATGAAATCCTCAAATCATTTCATTTATAGGTGATGCAAA

ATAACCCCTCACTAGGAGCAAACTACACCTTCAGAGGAACATTAACACAACTTAAATCTCTCGGAC

AAACACACTCCAACTAATCTTAATACAAAAGCCTCTCGGGCAACACACTCCAACTCACACGGAAACT

Change Region Shown

Whole sequence

Selected Region

from: 12976 to: 23286

Update View

Customize View

Pick Primers

Design and test primers for this sequence using Primer-BLAST.

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

Step 6: Retrieving and identifying the TSD: You now have before you the sequence of a full-length rice copia element. We still need to identify the target site duplication (TSD). (See page 85). For most LTR retrotransposons, the TSD is 5bp.

To find the TSD simply examine the 5 nucleotides flanking the element you just annotated. Remember these are direct repeats.

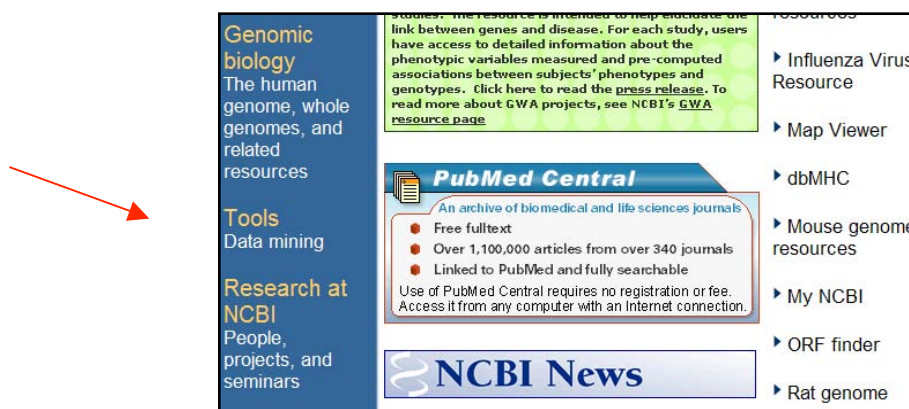
Go to the full BAC accession AL606652. Limit the sequence region to 12956 to 12986 and click Refresh.

Scroll all the way down and copy the sequence to a Word file. Visually compare this sequence to the 5' end of the LTR sequence. The 3' end of this sequence should overlap the 5' LTR by 10 nt. Write down the 5 nucleotides next to the LTR. It should be 'ATTGG.'

Repeat on the other end by selecting sequence from 23276 to 23306. You should find the 'ATTGG' present in this sequence also. You have identified the target site duplication.

Step 7: 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, PHI-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

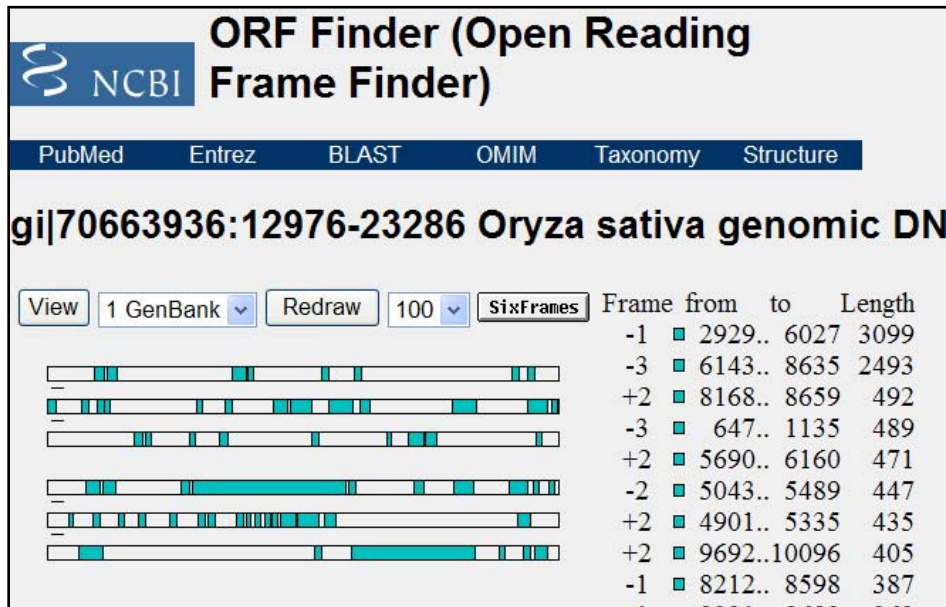
The ORF Finder (Open Reading Frame Finder) is a graphical analysis already in the database. This tool identifies all open reading frames using the standard or alternate against the sequence database using the WWW BLAST server. The C with the Sequin sequence submission software.

Enter GI or ACCESSION OrfFind Clear

or sequence in FASTA format

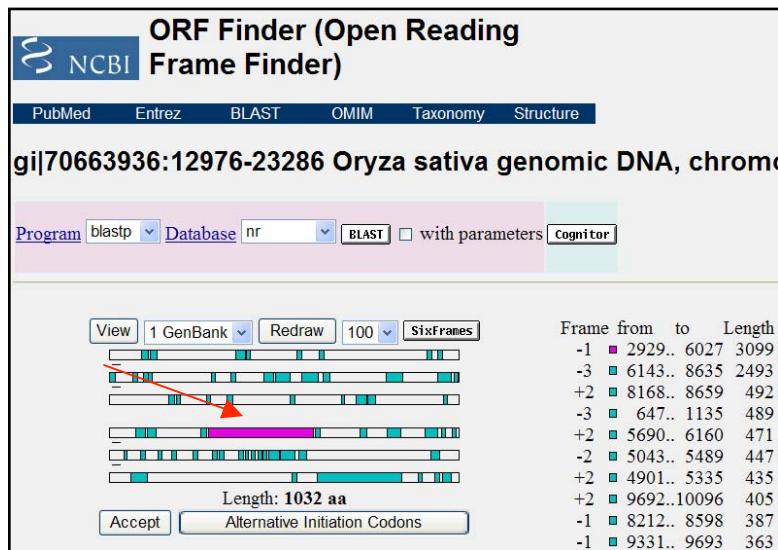
```
>gi|70663936:12976-23286 Oryza sativa genomic DNA,
chromosome 4, BAC clone: OSJNBb0004A17, complete
sequence
TGAAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGTGAATATAGCAAT
TCAAACTTTGCCCCCG
AAAATACTCATCAAGCCGGATTTCTCAAAATCCTTACTAGAATCGGGCTATTA
GAGAAGCCGGATCTAG
AAAAGAAGAGAGAAAAAGAAAGAAAAGGAATTCCTCCGAACTAGAGGAGGAAGA
```


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.

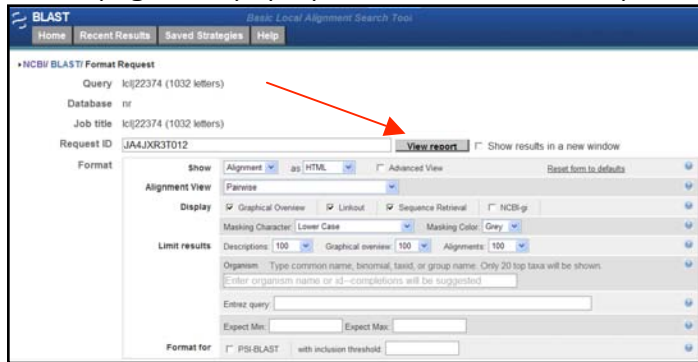
Note: usually the longer the ORF, the more reliable the information. Let's click the longest one and see what happens.



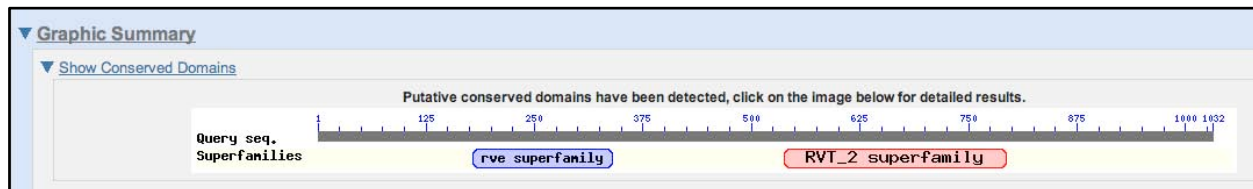
Select the ORF you're interested in. Click "BLAST" at the top of the new page.



A new page will pop up. Just click "view report".



The Graphic Summary tells you it is an reverse transcriptase superfamily member.

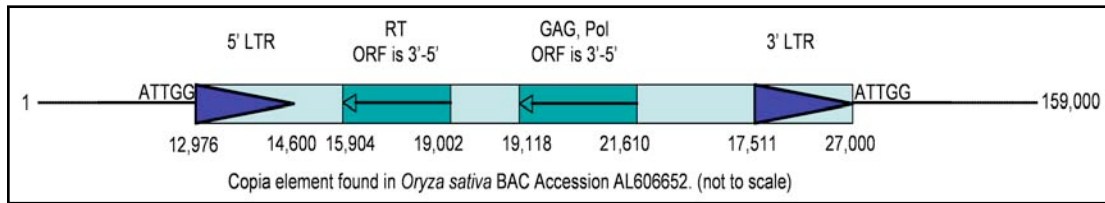


You can see details of the blast result when it is done.

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....



Chapter 6: Analyzing Excision Events.

The final part of Experiment 2 is analysis of the excision events. This involves two steps. The first is to count the number of colonies on the plates and compare them to the viable count. The second step is to analyze the excision site in *ADE2* using PCR and DNA sequencing. We do this for two reasons. First we need to show that reversion to *Ade*⁺ was due to excision of the TE. Second we want to know if the *ADE2* was repaired correctly, or if a "footprint" was created during repair of the excision site (Figure 1).

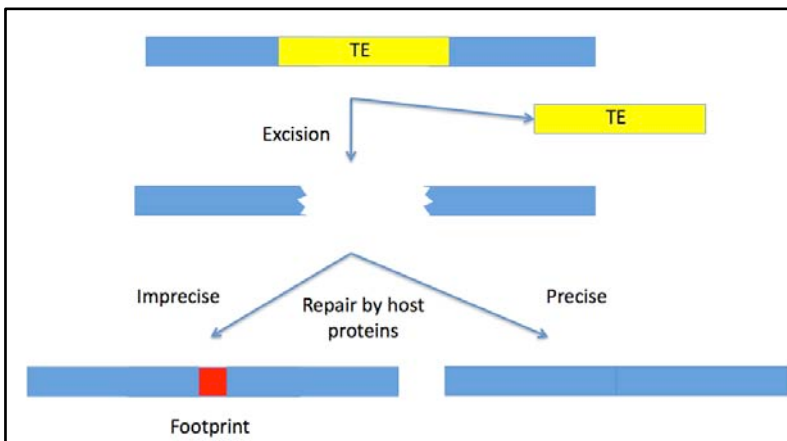


Figure 1. TE excision. The host must repair when a TE excises the gap left in the chromosome. The repair can be either be repaired precisely (on right) or imprecisely (on left). If the repair is imprecise a footprint sequence (shown in red is left behind).

For experiment 2 the TE was placed in the *Hpa*I restriction site of *ADE2*. If excision is repaired precisely then the *Hpa*I site will be regenerated. We can easily screen for excisions by first using PCR with primers cF and cR (Figure 2). The PCR can be digested with *Hpa*I enzyme. We can also sequence the PCR product.

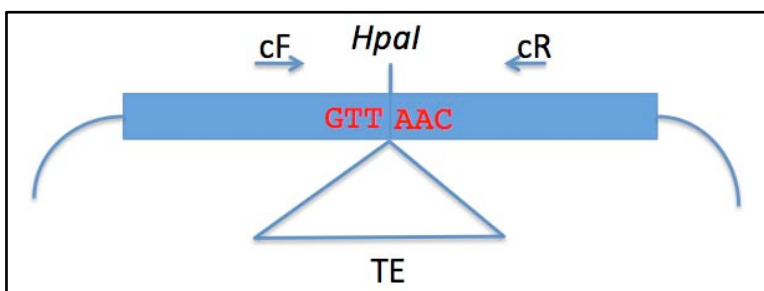


Figure 2. Part of pReporter showing the *ADE2* gene (blue box) with a TE inserted at the *Hpa*I site. The PCR primers (cF and cR) used in this experiment are

The size of the amplicon with the TE inserted is 785 bp and without you will get a 348 bp product.

Protocol for Footprint Analysis

Step 1. Tuesday, February 17, 2009 PCR Amplify across the TE insertion site in ADE2.

Materials

Zymolyase solution

PCR Master Mix

Primers

ADE2-cF

ADE2-cR

Sterile Water

1.5 ml tubes

PCR tubes

Ice

Each student will analyze the colonies from the following plates:

Ping LALA	1
Pong LALA or WT	1
Pong/mPing	2
<u>Osma</u>	<u>2</u>
Total analyzed	7

For each colony being analyzed:

1. Label a 1.5 ml tube. Pipette 20 μ l of Zymolyase solution in it. Zymolyase is an enzyme that degrades the chitin from cell wall of yeast. The yeast will then burst allowing the plasmid to go into solution.
2. Using a yellow tip, scrape some of the yeast off the plate and put it in the zymolyase.
3. Set these tubes aside at room temp while you prepare the PCR reaction.

PCR Reaction.

1. Label a strip of PCR tubes. 1-7 will contain yeast, 8 is the negative control.

2. Label a 1.5 ml tube.

3. Mix the following reagents in the 1.5 ml tube using the volumes in the shaded column.

	1X (μ l)	9X
2x Master Mix	25.0	225.0
H ₂ O	18.0	162.0
Forward Primer	1.0	9.0
Reverse Primer	1.0	9.0
DNA	5.0	-----
Total	50.0	405.0 μ l

Keep this mix on ice!

3. Pipette 45.0 μ l into each PCR tube.

4. Add 5.0 μ l of the yeast DNA to each tube. Add 5.0 μ l water to tube 8.

5. Seal the PCR tubes well. Give strip to instructor.

6. Pour a 1.5% agarose gel.

Step 2. Thursday, February 19, 2009**Run Gel of PCR reaction**

1. Label eight 1.5 mL tubes.

2. Pipette 5 μ l 6x loading dye in each tube.

3. Pipette 20 μ l of PCR reaction into the correct tube.

4. Load gel. Do not forget to include a DNA ladder.

***HpaI* digest of PCR reaction.**

1. Label a 1.5 mL tube "HpaI mix."

2. To *HpaI* mix tube mix the following in shaded column:

	1X (μ l)	9X
H ₂ O	6.5	58.5
10X Buffer 4	3.0	27.0
<i>HpaI</i> enzyme	0.5	4.5

Keep this mix on ice!

3. Pipette 10 μ l of *HpaI* mix to each PCR tube. Keep tubes on ice until you are finished.

4. Place tubes at 37°C for overnight.

5. Pour a 1.5% gel.

Step 3: Gel Extraction and DNA Sequencing.**Excise gel fragment**

1. Label and weigh seven 1.5 ml tubes 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 QIAquick Gel Extraction Kit):

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.
 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. To wash any impurities (EtBr and agarose), add 0.75 ml of Buffer PE to QIAquick column and spin column at 13,000 rpm for 1min.
 7. 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).
8. Place QIAquick column in a clean 1.5 ml microcentrifuge tube.

9. To elute DNA from the column, add 30ul water to the center of QIAquick membrane, leave column on bench for 1 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.

DNA Sequencing

1. The DNA concentration will be measured with a Nanodrop DNA analyzer.
2. The DNA will be mixed with primer cF.
3. Each sample will be entered into a spreadsheet.
4. Samples will be shipped overnight to GeneWiz.

DNA Sequence Analysis

DNA sequence analysis will be done in class. Sequences will be provided that you can use to compare to the sequencing results.

Step 3. Tuesday, March 3, 2009

Run Gel of HpaI digest

1. Pipette 5 μ l 6x loading dye in each tube.
2. Load gel. Do not forget to include a DNA ladder.

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).

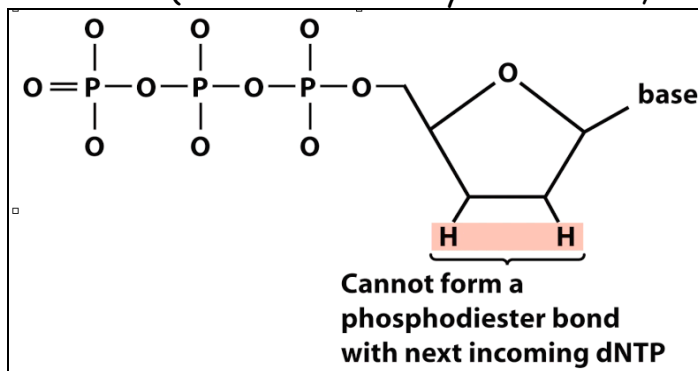
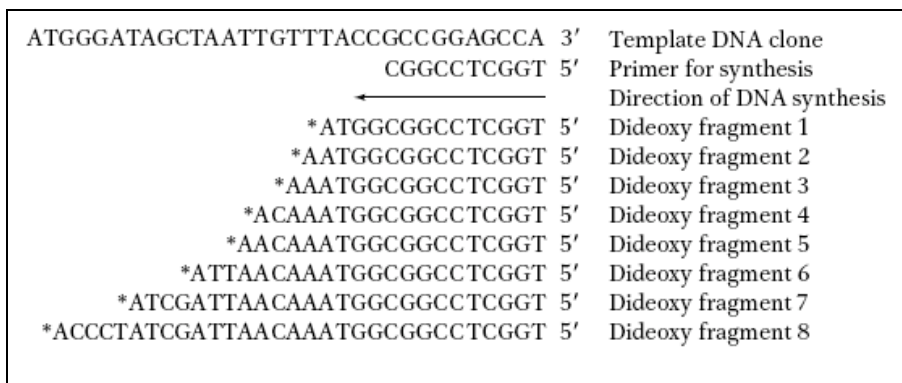


Figure 12. A chain-terminating nucleotide triphosphate (called a di-deoxynucleotide or ddNTP). Because it has a "H" instead of a "OH" at the 3' position, it is not a substrate for the addition of another NTP and DNA synthesis terminates.

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.

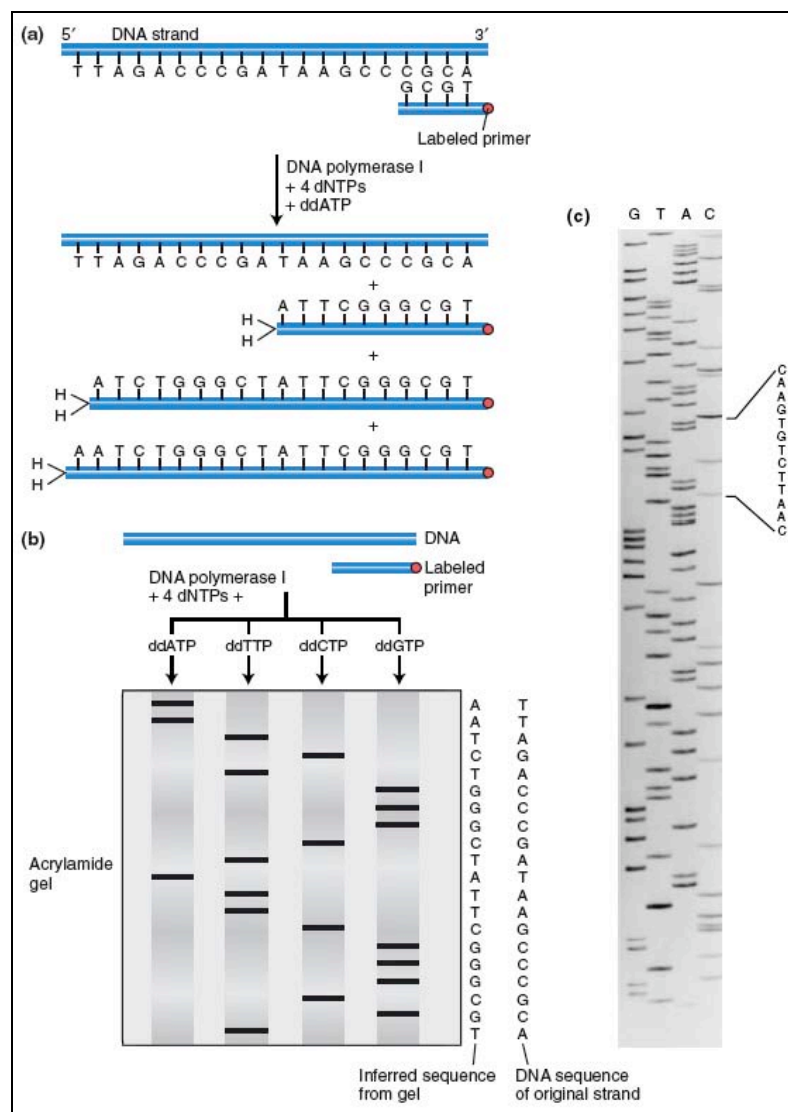


Figure 13. DNA is efficiently sequenced by including dideoxynucleotides among the nucleotides used to copy a DNA segment. (a) In this example, a labeled primer (designed from the flanking vector sequence) is used to initiate DNA synthesis. The addition of four different dideoxynucleotides (ddATP is shown 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.

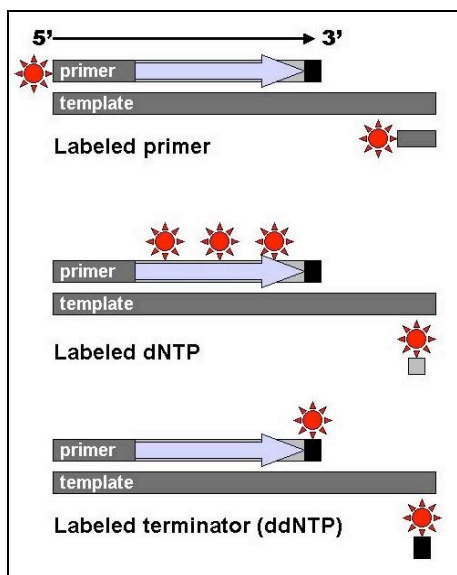


Figure 14: DNA fragments can be labeled by using a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

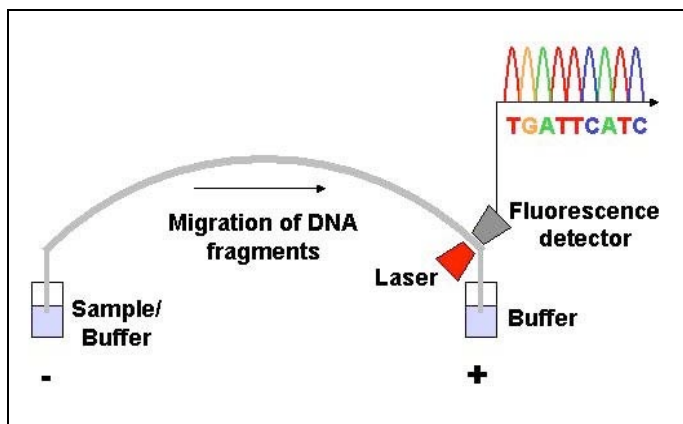


Figure 15. Modern automated DNA sequencing instruments (DNA sequencers) can sequence up to 384 fluorescently labelled samples in a single batch (run) and perform as many as 24 runs a day. However, automated DNA sequencers carry out only DNA size separation by capillary electrophoresis, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms (see Fig. 16, 17).

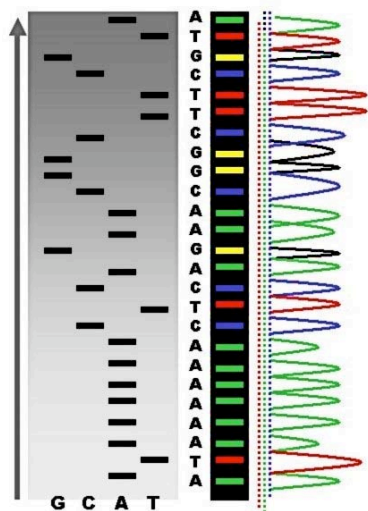


Figure 16: Sequence ladder by radioactive sequencing compared to fluorescent peaks

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.

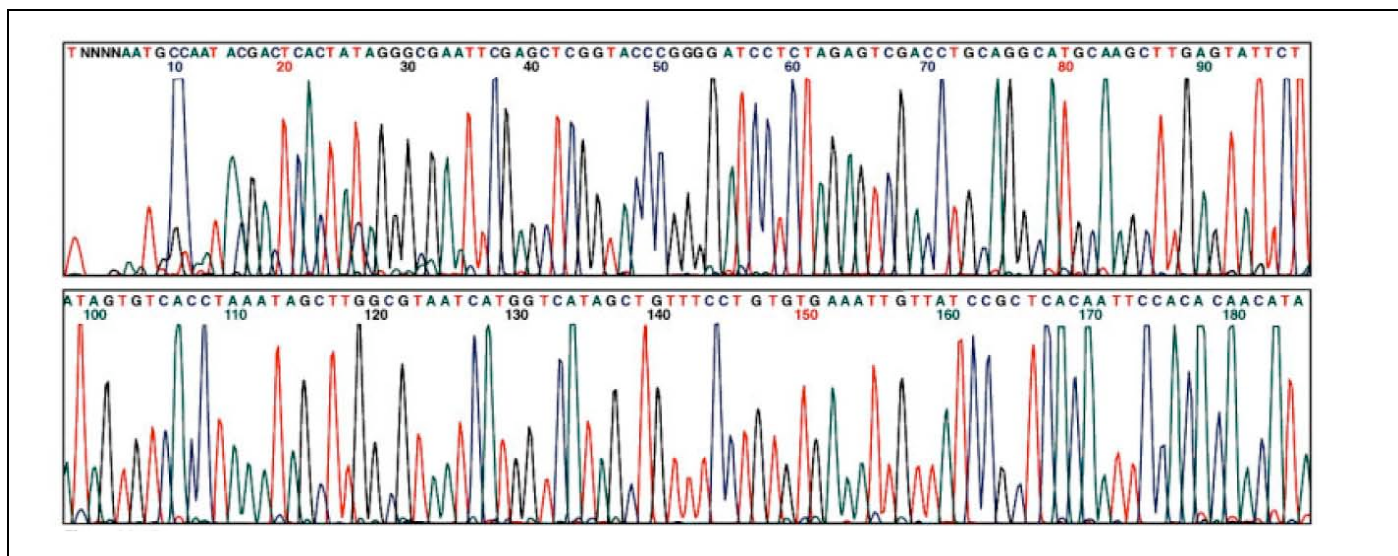


Figure 17. An example of a chromatogram file of a Sanger sequencing read. The four bases are detected using different fluorescent labels. These are detected and represented as 'peaks' of different colors, which can then be interpreted to determine the base sequence, shown at the top.

Information you will need to annotate your sequence

Your sequences have been downloaded to the class share file 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 17, above) and one text file. The text file contains the DNA sequence.

Here is the information you need to analyze your sequence...

The sequence of the forward and reverse primers used for PCR.....

cF: 5' - GGGTTTTCCATTTCGTCTTGAAGTCGAGGAC - 3'

cR: 5' - ACATTCCCACACCAAATATACCACAACCGGG - 3'

If these primers were used to amplify a pReproter plasmid that just had the ADE2 gene (no TE insertion), the sequence of that fragment would be the following.

Note that the sequence of the forward primer and the complement of the reverse primer are color-coded above and below. Furthermore, the ADE2 sequence is in blue as in figure 2 above. The *HpaI* site is shown in red. The sequencing reaction was primed with cF.

> DNA sequence of the sequence (before the insertion of the TE):

```
GGGTTTTCCATTTCGTCTTGAAGTCGAGGACTTTGGCATACGATGGAAGAGGTAACCTTCGTTGTA
AAGAATAAGGAAATGATTCGGAAGCTTTGGAAGTACTGAAGGATCGTCCTTTGTACGCCGAAA
AATGGGCACCATTTACTAAAGAATTAGCAGTCATGATTGTGAGGTCTGTTAACGGTTTAGTGTT
TTCTTACCCAATTGTAGAGACTATCCACAAGGACAATATTTGTGACTTATGTTATGCGCCTGCT
AGAGTTCCGACTCCGTTCAACTTAAGGCGAAGTTGTTGGCAGAAAATGCAATCAAATCTTTTC
CCGTTGTGGTATATTTGGTGTGGAAATGT
```

Protocol for analyzing your sequence...

Open the text file for your sequence. Check the sequences manually first. You may see many N's. This has to do with the quality of the sequencing read. To understand what this means, we will have to open the chromatogram file, since the chromatogram file provides the quality information of the sequencing result.

You have to open the chromatogram file with a program called "4 peaks"
 (<http://mekentosj.com/4peaks/>). We will show you how to do this in class. See
 Figure 18 for an example.

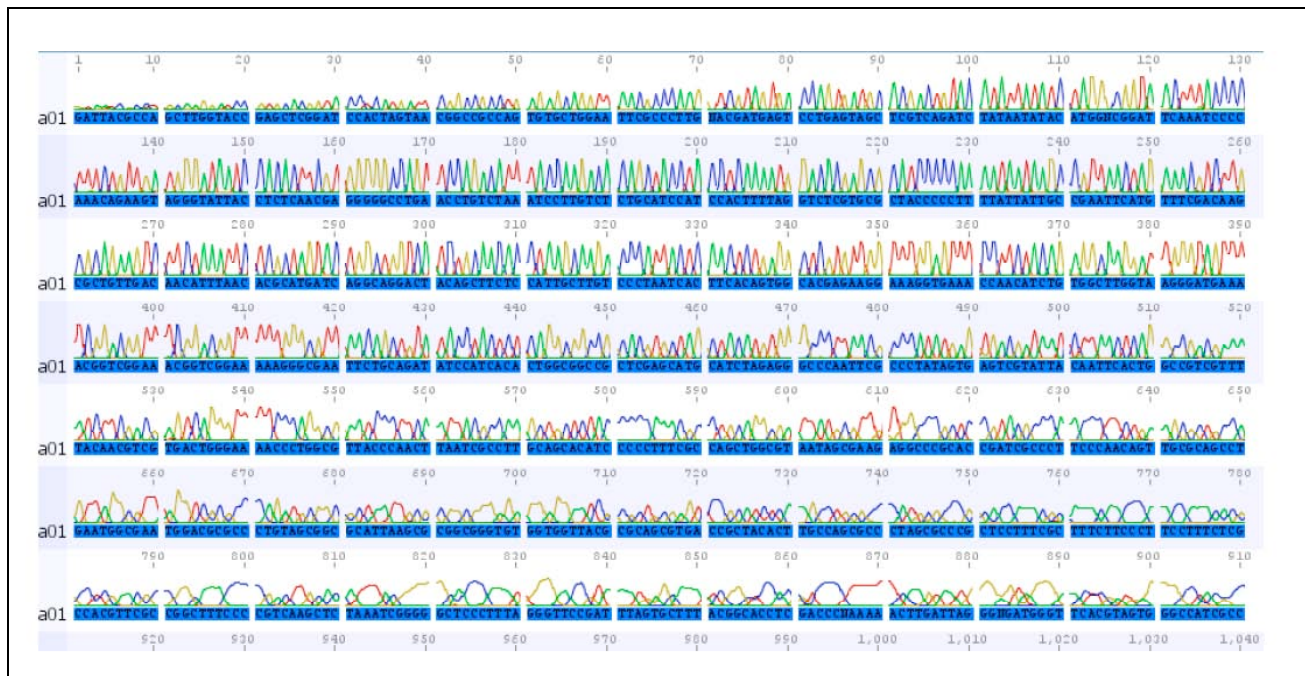


Figure 18. Example of a chromatogram file. Each nucleotide (A,G,C, T) is drawn in one of 4 colors. Higher/sharper peak means strong signal, and the corresponding nucleotide sequence is highly reliable. Very low peaks or twisted curves mean poor sequencing quality.

Analyzing Your Sequences

There are two sets of sequencing results to analyze. The first is for the Ping and Pong excision events you processed in class. The second set is for a different TE superfamily called *OSmar* for *Oryza sativa* Mariner. The *OSmar* experiment was conducted similarly to your Ping experiment.

Using Multiple Alignment to analyze your results.

You will use a program called Muscle that performs multiple alignments.

Step 1. Open the Muscle web page:

<http://www.ebi.ac.uk/Tools/muscle/index.html>

The screenshot shows the EBI Muscle web interface. The page title is "MUSCLE" and it is part of the "Tools > Sequence Analysis" section. The main content area contains a description of MUSCLE and a "Download Software" link. The form includes several sections: "RESULTS" with a dropdown set to "interactive", "SEARCH TITLE" with a text input "Sequence", "YOUR EMAIL" with an empty text input, "OUTPUT FORMAT" with a dropdown set to "ClustalW2" (circled in red), "OUTPUT TREE" with a dropdown set to "none", and "OUTPUT ORDER" with a dropdown set to "aligned". Below these is a large text area for "Enter or Paste a set of Sequences in any supported format:" and a "Help" button. At the bottom, there is an "Upload a file:" section with a "Choose File" button and "no file selected" text, and "Run" and "Reset" buttons.

Step 2. Change the Output Format to ClustalW2.

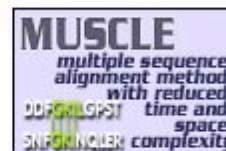
Step 3. Copy your sequences into the Text Window. Use Fasta format. Click Run.

EBI > Tools > Sequence Analysis

MUSCLE

MUSCLE stands for **M**ultiple **S**equence **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](#)



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

Enter or Paste a set of Sequences in any supported format:

[Help](#)

```
>4741-4-cF_D01.ab1
NNNNNNATGGANAGGTAACCTTCGTTGTAAAGAATAAGGAAATGATTCCGGAAGCTTTGGAAGTACTGAAGG
ATCGTCCTT
TGACGCCGAAAAATGGCCACCATTTACTAAAGAATTAGCAGTCATGATTGTGAGGTCTGTTTACTCCGAGG
TAAACGGT
TTAGTGTTCCTTACCCAATTGTAGAGACTATCCACAAGGACAATATTTGTGACTTATGTTATGCCCTGCTA
GAGTTCC
GGACTCCGTTCAACTTAAGGCCAAGTTGTTGCCAGAAAATGCAATCAAATCTTTCCCGTTGTGGTATATTT
GGTGTGG
AAATGAAGN
```

Upload a file: no file selected

Run

Reset

The Results will look similar to this:

MUSCLE Results

Results of search	
Number of sequences	5
Sequence type	DNA
Muscle version	MUSCLE v3.7 by Robert C. Edgar
Max length	437
Average length	355
Output file	muscle-20090303-17165098.output (clustalw)
Your input file	muscle-20090303-17165098.input

Alignment

MUSCLE (3.7) multiple sequence alignment

```

ade2          GGGTTTTCCATTCGTCTTGAAGTCGAGGACTTTGGCATAACGATGGAAGAGGTAACCTTCGT
4741-2-cF_B01.ab1 -----GNNNGATGG-ANAGGTAACCTTCGT
4741-4-cF_D01.ab1 -----NNNNNATGG-ANAGGTAACCTTCGT
4741-3-cF_C01.ab1 -----NNNNNATGG-ANAGGTAACCTTCGT
4741-1-cF_A01.ab1 -----TNNNNNNCNATGG-ANAGGTAACCTTCGT
                                     *****

ade2          TGTAAGAATAAGGAAATGATTCGGAAGCTTTGGAAGTACTGAAGGATCGTCCTTTGTA
4741-2-cF_B01.ab1 TGTAAGAATAAGGAAATGATTCGGAAGCTTTGGAAGTACTGAAGGATCGTCCTTTGTA
4741-4-cF_D01.ab1 TGTAAGAATAAGGAAATGATTCGGAAGCTTTGGAAGTACTGAAGGATCGTCCTTTGTA
4741-3-cF_C01.ab1 TGTAAGAATAAGGAAATGATTCGGAAGCTTTGGAAGTACTGAAGGATCGTCCTTTGTA
4741-1-cF_A01.ab1 TGTAAGAATAAGGAAATGATTCGGAAGCTTTGGAAGTACTGAAGGATCGTCCTTTGTA
*****

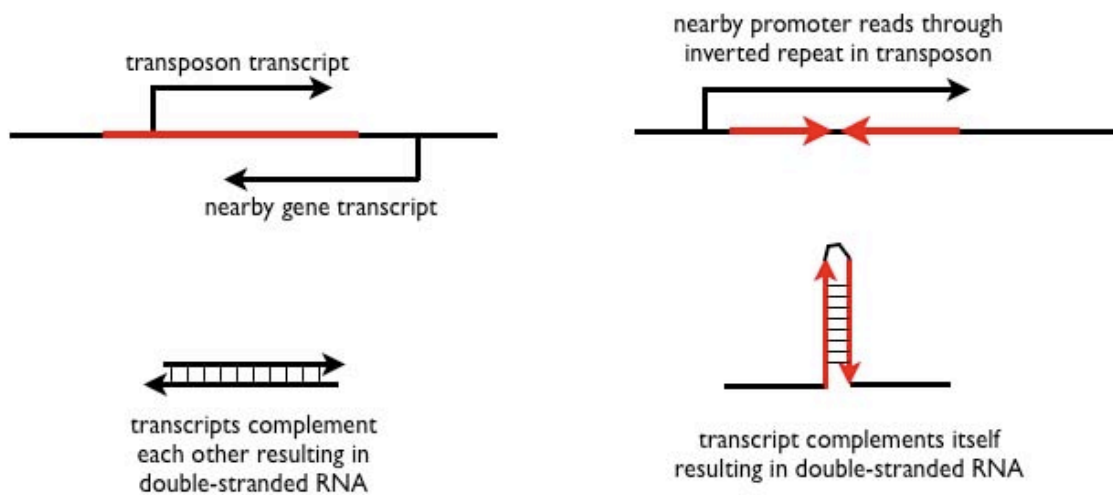
ade2          CGCCGAAAAATGGGCACCATTACTAAAGAATTAGCAGTCATGATTGTGAGGTCTGTTT--
4741-2-cF_B01.ab1 CGCCGAAAAATGGGCACCATTACTAAAGAATTAGCAGTCATGATTGTGAGGTCTGTTTAA
4741-4-cF_D01.ab1 CGCCGAAAAATGGGCACCATTACTAAAGAATTAGCAGTCATGATTGTGAGGTCTGTTTAA
4741-3-cF_C01.ab1 CGCCGAAAAATGGGCACCATTACTAAAGAATTAGCAGTCATGATTGTGAGGTCTGTTTAA
4741-1-cF_A01.ab1 CGCCGAAAAATGGGCACCATTACTAAAGAATTAGCAGTCATGATTGTGAGGTCTGTTTAA
*****

ade2          -----AACGGTTAGTGTPTTCTTACCCAATTGTAGAGACTATCCACAAGGACAA
4741-2-cF_B01.ab1 CTCCGAGTAAACGGTTTAGTGTPTTCTTACCCAATTGTAGAGACTATCCACAAGGACAA
4741-4-cF_D01.ab1 CTCCGAGTAAACGGTTTAGTGTPTTCTTACCCAATTGTAGAGACTATCCACAAGGACAA
4741-3-cF_C01.ab1 CCCCAGTAAACGGTTTAGTGTPTTCTTACCCAATTGTAGAGACTATCCACAAGGACAA
4741-1-cF_A01.ab1 CTCCGAGTAAACGGTTTAGTGTPTTCTTACCCAATTGTAGAGACTATCCACAAGGACAA
*****
    
```

Chapter 7: Tracking epigenetic silencing of a transposon in maize.

Overview: As you have seen in class, transposons can reach very high copy numbers, to the extent that some genomes are mostly transposon. However, the hosts are not without resources to counter these selfish elements. Over the past few years, we have become aware of an ancient immune system whose function is to recognize and silence transposons and invading viruses. It does so by recognizing particular forms of RNA that are produced by transposons but not by most host genes. One

Two ways to get double stranded RNA from a transposon

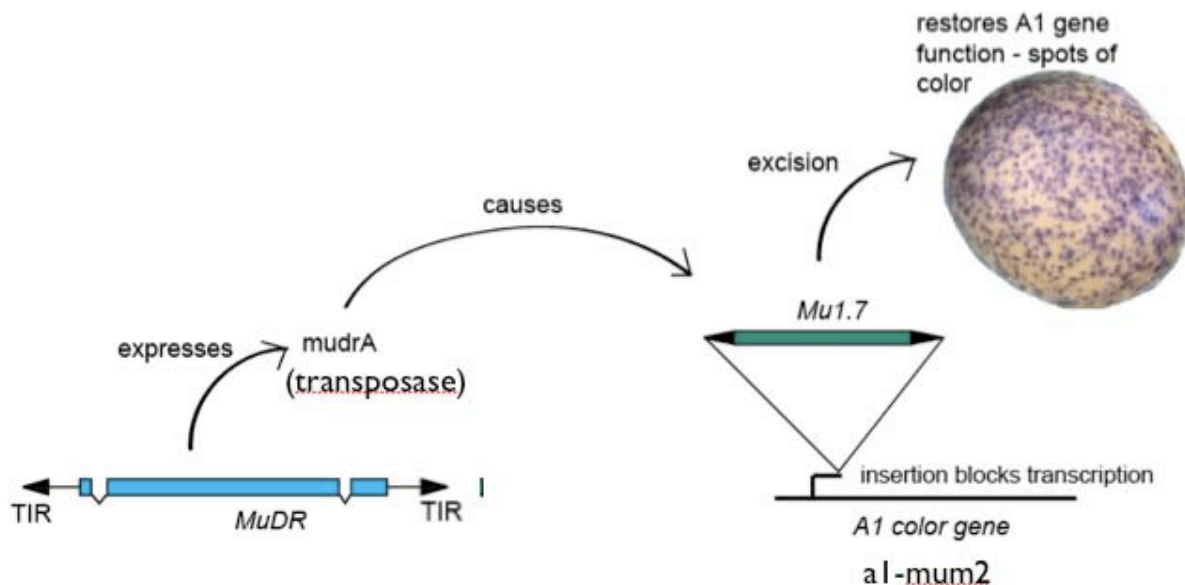


powerful trigger is double-stranded RNA (dsRNA). Transposons are particularly prone to produce double stranded RNA. They move from place to the place within the genome and can cause a variety of genomic rearrangements - events that can produce aberrant transcripts, including antisense and hairpin RNAs such as the ones portrayed above. Antisense transcripts can anneal with mRNAs to produce dsRNA. Hairpin RNAs are stretches of RNA where one portion of the molecule is complementary to another part of the same molecule. Such RNAs can be produced if transcription proceeds through an inverted repeat, a common feature of some transposons (see above figure).

Once dsRNA is detected, it triggers a cascade of events that lead to the production of small interfering RNAs (siRNAs) that are used to target all

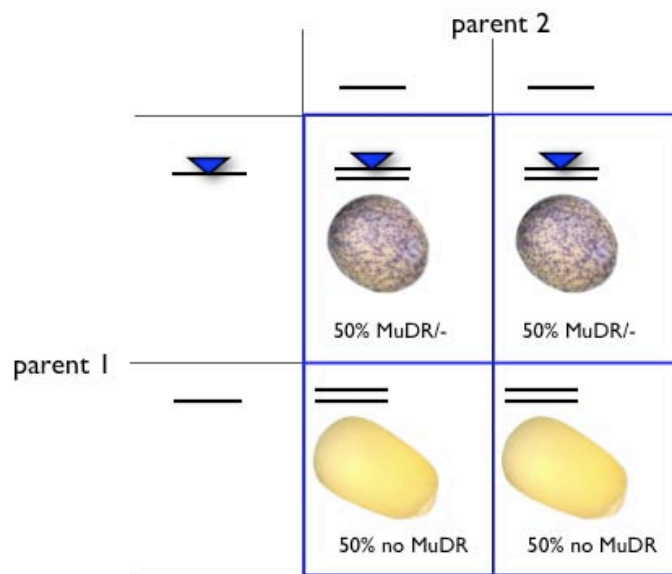
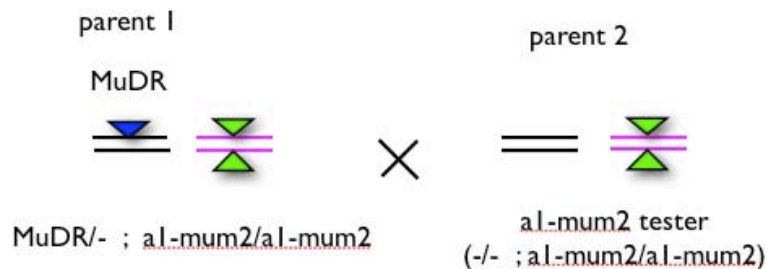
complementary RNAs in the cell for degradation. As you will see in class, this is an effective way to eliminate all transcripts from a given transposon, and without transcript, autonomous transposons have no way of producing the enzymes they need to transpose. In this process, the dsRNA produced by a given transposon acts as an “antigen” that triggers an immune response. However, the system is actually smarter than that, because it also includes a system that “remembers” who the transposon was, even after the trigger is lost. This memory system involves modification of cytosine nucleotides by the addition of a methyl group (DNA methylation), as well as modification of histones (which can change chromatin density). These modifications force the transposon to produce a form of aberrant RNA that continues to trigger silencing, even after the initial trigger is gone. These modifications will be described in more detail in class on tuesday.

Today we will be examining the genetic segregation of a DNA sequence that acts as a trigger of transposon silencing in maize. The focus of our experiments will be members of the Class II superfamily Mutator. The Mutator system was discovered in maize in the 1970s. It got its name



because it is highly mutagenic and prone to increase its copy number rapidly. The system is composed of non-autonomous elements (or several types, called Mu1 through Mu8) and an autonomous element, MuDR. The non-autonomous elements only jump when MuDR is present. In our

example, a non-autonomous element is inserted into a gene required to make color in the maize seed. When MuDR is present, the non-autonomous element (Mu1.7 in this case) jumps out of the color gene late during development of the seed. The result is small spots of revertant (wild-type) colored tissue on a mutant background. These spots of color serve as a powerful assay for the presence of an active MuDR element in the genome of that plant. Further, the frequency of spotting can indicate how active MuDR is in a given kernel. We will use this fact to track transposon inactivation in our families.



Classical genetic analysis:

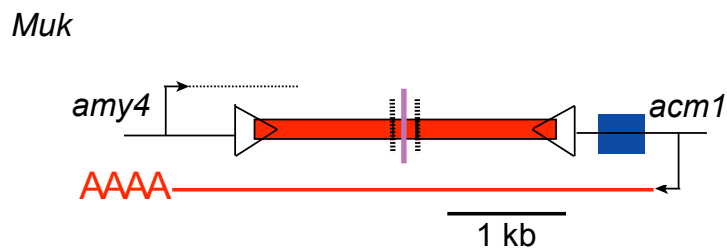
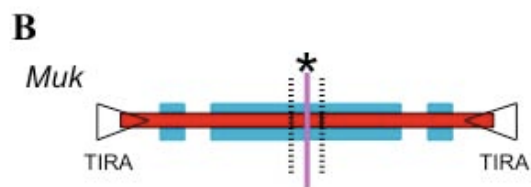
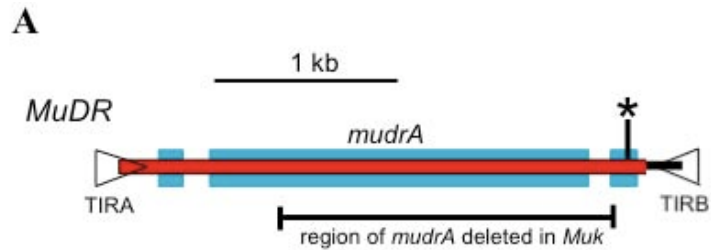
In order to understand our experiment we need to understand basic classical genetic analysis. The figure to the right illustrates a cross. At the top, each parent carries two homologous chromosomes; one from each parent. These are indicated by the parallel lines. MuDR is indicated by the triangle inserted into one of the lines. This plant is heterozygous for the insertion, since only one homologous chromosome has it. Below the cross are the progeny, with the expected genotypes laid out in a grid. To the left of the blue box are the possible genotypes in the egg of the female. Because of the principle of



random assortment, each product of meiosis has an equal probability of getting one or the other homologous chromosome. Thus, half of the eggs in the female in this cross will get MuDR and half will not. Above the blue box are the expected genotypes of the pollen.

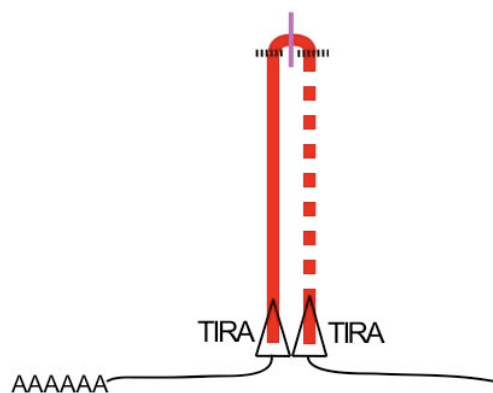
Note that the male parent only contributes pollen that lacks MuDR. Inside the blue box are the expected progeny classes, which we get by combining the various egg and pollen genotypes. As we can see, when a plant that is heterozygous for a single MuDR element is crossed to a plant that doesn't carry MuDR, half the progeny will carry MuDR and half don't, because half of the eggs receive MuDR and half do not.

The result is an ear that segregates 50% spotted kernels. As we will see below, plants grown from these kernels can be genotyped for the presence or absence of MuDR.

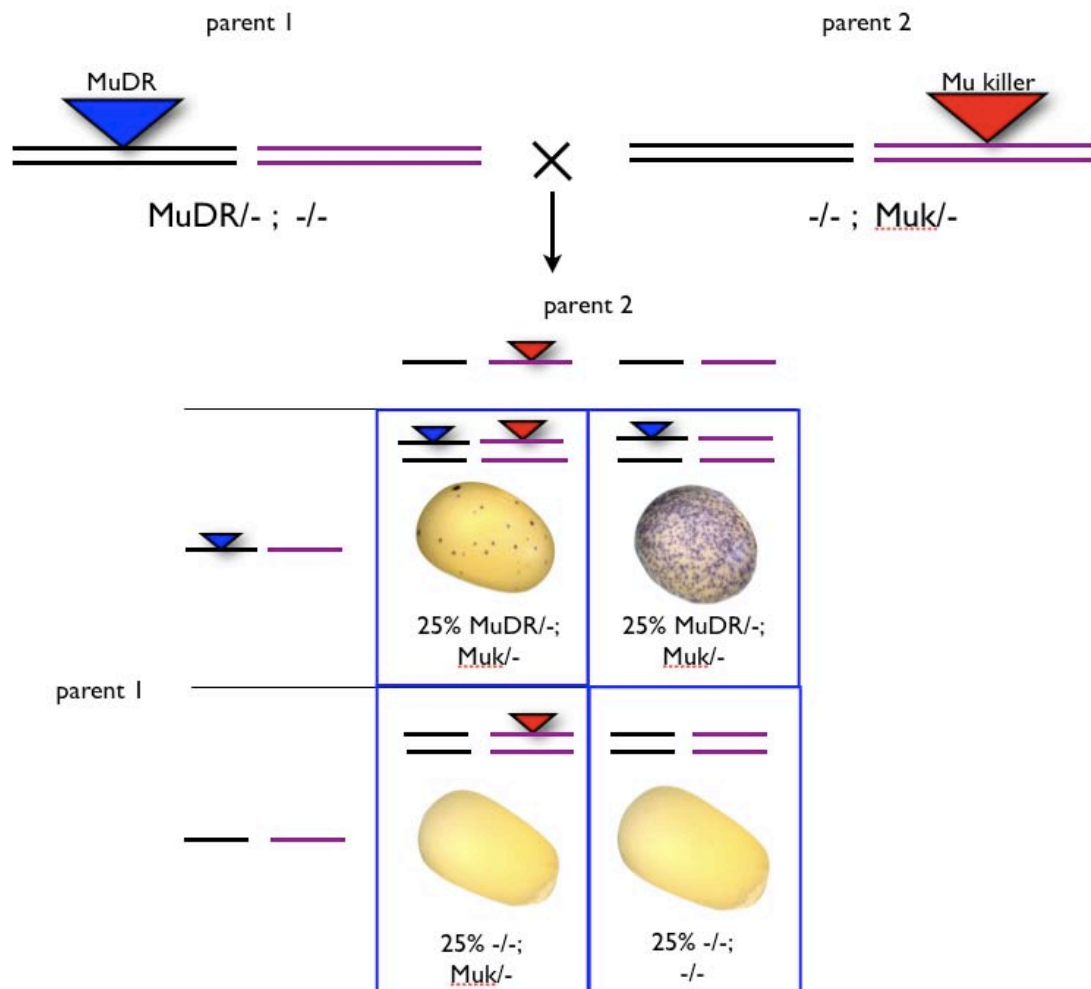


Epigenetic silencing of the Mutator System:

Silencing of the Mutator transposon system can be triggered by a rearranged version of MuDR called Mu killer (Muk). Mu killer is a version of MuDR in which half the element has been duplicated and inverted relative to itself, resulting in a mirror image, similar to the TIRs that are associated with many transposons, only much longer. Transcription from a nearby promoter (the *acm1* gene) results in transcription all the way through this mirror image of part of a MuDR



element. The resulting transcript has a hairpin - RNA sequences that compliment each other, just like the two strands of DNA compliment each other. In the figure to the right, the transcript produced by Mu killer is illustrated. Transcription proceeds from a flanking promoter, through the mirror image version of MuDR and is polyadenylated (AAAA) at a flanking site. Polyadenylation permits export of this RNA to the cytoplasm. The resulting RNA transcript includes parts of two flanking genes as well a the MuDR hairpin. The double stranded RNA portion of this transcript (which corresponds to MuDR) is processed into siRNAs. As will be discussed in class, this double-stranded RNA is processed by a dicer into siRNAs, which cause subsequent cleavage of normal MuDR transcripts and eventual



transcriptional silencing of MuDR elements. When a plant carrying MuDR is crossed to a plant that is heterozygous for Muk, the result is an ear that segregates for MuDR and Muk. The figure above shows the genetic

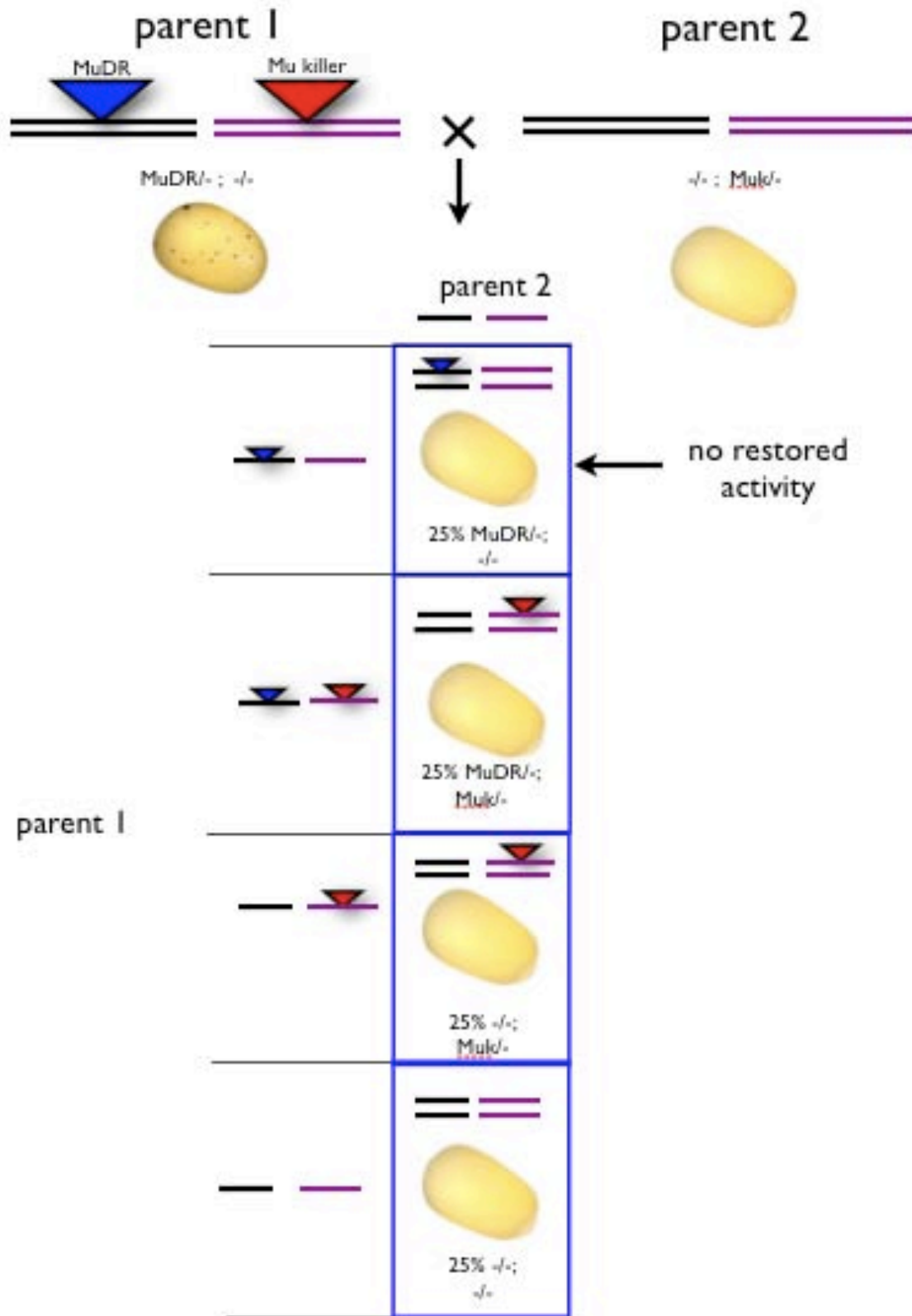
composition of the two parents and the resulting progeny based on genetic principles. One parent carries MuDR as a heterozygote, and the other carries Mu killer as a heterozygote. Note that the two elements are on different chromosomes, as indicated by the black and purple lines. When the female parent undergoes meiosis, half the eggs carry MuDR, and half lack it; none of them carry Muk. Similarly when the male parent undergoes meiosis, half the pollen carry Muk, and half lack it. Each of four



possible combinations are possible in the progeny. Thus, the resulting progeny kernels segregate 25% MuDR with Muk, 25% with MuDR without Muk, 25% Muk alone and 25% neither MuDR nor Muk. The resulting ear looks like the picture to the right. Notice the weakly spotted kernels. These are individuals that inherited both MuDR and Muk. In these kernels, MuDR transcript is being degraded because Muk produces the trigger, or antigen, that causes dicer-mediated degradation of MuDR transcript.

What happens next is even more interesting, because it shows that the plant can remember that an element has been targeted for silencing in a previous generation. The cross described above produces a class of weakly spotted kernels that carry MuDR and Mu killer. When plants grown from these kernels are crossed to a tester (a plant with neither MuDR nor Muk), nearly all of the resulting kernels are non-spotted, even the plants with MuDR and without Muk.

The cross is illustrated on the next page. A plant carrying MuDR and Muk is crossed to a plant lacking both MuDR and Muk. Note that genetic segregation of these elements should (and does) produce a class of progeny that carries MuDR but that lacks Muk. The female parent can produce four kinds of eggs, representing each combination of MuDR and Mu killer. The tester produces one kind of pollen. Thus, there are four different kinds of progeny, each with an equal probability. Note the class of progeny that carries MuDR but lacks Mu killer. If MuDR is effected only when it is exposed to Muk, then we would expect that once Muk is lost, MuDR should regain activity. But it does not, as is indicated by the fact that



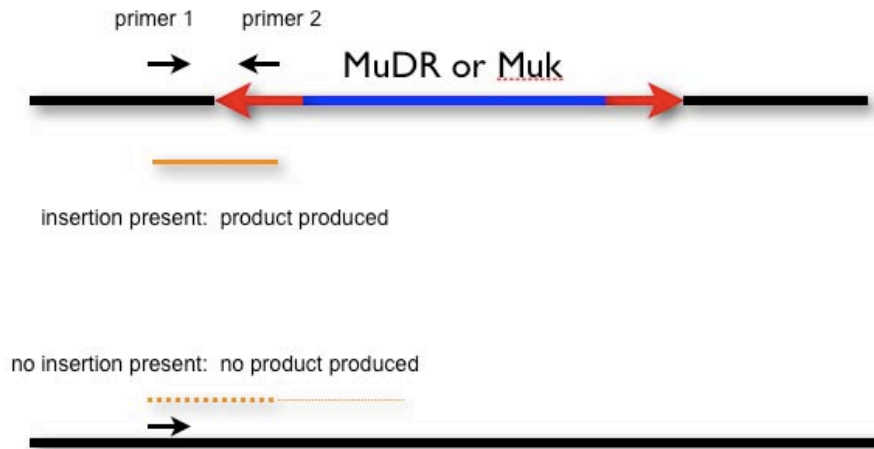
the kernel is not spotted. The progeny of these plants can be propagated indefinitely, and MuDR never wakes up again. Thus the genome “remembers” that MuDR was silenced, even after the trigger, or antigen, Muk, is lost. It’s important to note, however, that the DNA sequence of the MuDR element is identical to when it was active. Thus, the element is only sleeping, not dead. One note on classical genetics. Although it can be confusing, it is also remarkably powerful because it can give clean elegant answers to very complicated problems. Indeed, geneticists use the acronym “APOG” to describe how useful it can be - ask Damon what it stands for.

Experimental Protocol:

In this laboratory, we will examine the genetic segregation of MuDR and Muk to examine the effects of Muk on MuDR. In order to do this we will first need to extract DNA from sprouts grown from corn kernels. The extraction protocol is provided on a separate hand-out. The protocol is relatively straight-forward. First we will grind up the tissue in liquid nitrogen, which makes grinding easier and breaks up the cells and cell walls. Next, we will add a buffer, basically salty water EDTA, which inhibits enzyme activity (we don’t want our DNA chewed up by nucleases). Then we will add SDS, which is a concentrated surfactant, or soap, which pops open cell and nuclear membranes (that’s why “antibacterial” is such a rip off - soap is plenty antibacterial all by itself). Then we heat to 65° C. This dissociates protein complexes and allows everything to mix well. Then we add potassium acetate. This causes the proteins, but not the DNA, to precipitate out of solution. After chilling for a few minutes (to aid in protein precipitation) we spin the mix in a centrifuge, which separates the protein and assorted cell debris from the liquid that still contains the DNA. Then we suck off the liquid, leaving behind the debris, and add it to a new tube. Then we add an equal volume of isopropanol. This causes the DNA to precipitate out of solution. Depending on the amount of DNA, we may see fine strands of it at this stage (this is the fun part). Then we spin the DNA down to the bottom of the tube, pour off the liquid, and re-suspend the DNA in water. Now it’s ready for analysis.

After extracting the DNA, we want to find out who has MuDR and who has Muk. To do this we will use PCR, which can specifically identify MuDR and Muk. It can do this because each of these elements is at a unique position in the maize genome. Thus, although the sequences of MuDR and Muk

are identical, sequences *flanking* those elements are unique. We will use PCR primers specific to MuDR at a particular position (p1) and primers specific to Mu killer. In each case, one primer will be specific to the transposon and one will be specific to the DNA sequence



into which the transposon is inserted. Amplification will only work if the element (MuDR or Muk) is present at a particular position. If the element is not at that position, or is missing altogether, there will be no product.

Step 1: DNA extraction. First we will obtain leaf tissue from plants grown from kernels with the various genotypes described above. Then we will extract the DNA using the protocol provided. After the DNA extraction we will have about 50µl of DNA in a series of 1.5 ml centrifuge tubes.

Label each tube with a name for the particular family being examined (GH2 or GH3 in this case), the class of kernel (H for heavily spotted, W for weakly spotted and P for pale, or no spots) and a number for the individual plants. Thus, a given tube would be labeled, for instance, GH2 H1. Put your initials on the side of the tube as well. Remember that none of the results will make sense later if you don't label the tubes clearly. If you think you might have contaminated a tip, go ahead and throw it away.

Step 2: PCR genotyping.

As described above, we are interested in using PCR to genotype each individual from families segregating for MuDR and Mu killer. First we will examine the progeny of the following cross:

MuDR(p1)^{-/-} ; ^{-/-} x ^{-/-} ; Muk^{-/-}.

This is a cross between a plant that is heterozygous for MuDR at a particular position (p1) with a plant that is heterozygous for Mu killer. Seeds derived from the above cross have been separated into classes for you based on excision frequency (heavy, weak, pale). As we have seen, Muk acts on MuDR to reduce its activity, resulting in weakly spotted kernels. Thus, we will expect to find that plants grown from kernels with many spots will carry MuDR without Muk, and plants grown from weakly spotted kernels will carry both MuDR with Muk. Pale kernels should lack MuDR, but half of them should have Muk. We have grown sprouts from heavily spotted kernels, weakly spotted kernels and pale kernels from this cross.

Next we will examine progeny of the cross:

MuDR(p1)/-; Muk/- x tester (-/- ; -/-)

Recall that the plant carrying MuDR(p1) and Muk (MuDR(p1)/-; Muk/-) was derived from the cross between MuDR(p1) and Muk. This plant was grown from one of the weakly spotted kernels. The plant was then crossed to a tester that lacked both MuDR and Mu killer. Here, nearly all of the progeny kernels were pale. However, as described earlier, 25% of the progeny of this cross should carry MuDR without Muk. As before, we will genotype for MuDR(p1) and Mu killer. If there is a class of kernels that carries MuDR without Mu killer and that remained pale (no MuDR activity), then we will conclude that the genome is remembering to keep MuDR inactive even after Mu killer has segregated away.

As was done in a prior experiment, you will be resolving PCR products by agarose gel electrophoresis.

To perform the PCR, we need to dilute our DNA ten-fold. This is because PCR is very sensitive, and it actually works better if the DNA is less concentrated.

Prepare a new tube for each sample. Add 45µl of water to each tube. Then add 5µl of the concentrated DNA to each of the new tubes. Mark the new tubes the same way as the old tubes, but add the word “dilute” to the tube so that you can tell the difference.

Because we are interested in the genetic segregation of MuDR and Mu killer in both of the families, we will use PCR primers specific to each on all of the samples. In addition to the samples you prepare today, additional samples from the same families have already been prepared, so that we have enough data to make conclusions.

Important controls:

As you have learned in this course, ALL experiments must include controls; nothing makes sense without them. In this case, we will include POSITIVE controls in the form of samples that we know should amplify with a given primer pair, and a NEGATIVE control which includes all of the reagents EXCEPT DNA.

For this experiment we will use two pairs of primers:

to genotype for MuDR(p1):

Ex1: ACATCCACGCTGTCTCAGCC

RLTIR2: ATGTCGACCCCTAGAGCA

RLTIR2 is a primer in the end of all MuDR elements. Ex1 is a primer in the sequence flanking MuDR(p1). Successful amplification from any given sample will indicate that MuDR(p1) is present in this individual.

amplification conditions:

- 1: 94° 5 minutes initial melting step
 - 2: 94° 30 sec melting step
 - 3: 57° 45 sec annealing step
 - 4: 72° 45 sec extension step
- repeat step 2-4 35 times amplifications
- 5: 72° 10 minutes final
 - 7: soak at 4°

Mu killer:

12-4R3: CGGTATGGCGGCAGTGACA

TIRAR: AGGAGAGACGGTGACAAGAGGAGTA

TIRA is a primer in the end of all MuDR elements (remember that Mu killer is a rearranged MuDR element). 12-4R3 is a primer in the sequence

flanking MuDR(p1). Successful amplification from any given sample will indicate that Mu killer is present in this individual.

amplification conditions:

- 1: 94° 5 minutes
- 2: 94° 30 sec
- 3: 60° 45 sec
- 4: 72° 1 min
- repeat step 2-4 35 times
- 5: 72° 10 minutes
- 7: soak at 4°

Each PCR reaction will include:

Primer 1:	1.0 μ l
Primer 2:	1.0 μ l
dNTP:	1.0 μ l
10X buffer:	5.0 μ l
Taq polymerase:	0.5 μ l
Water:	40.5 μ l
DNA sample:	1.0 μ l

Note that you can make a cocktail of all of the ingredients except the DNA and then aliquot everything but the DNA into a series of PCR tubes. For a cocktail of ten samples, you want to make enough for 11 to correct for imperfect pipetting:

	each sample	cocktail
Primer 1:	1.0 μ l	11.0
Primer 2:	1.0 μ l	11.0
dNTP:	1.0 μ l	11.0
10X buffer:	5.0 μ l	55.0
Taq polymerase:	0.5 μ l	5.5
Water:	40.5 μ l	445.5

Aliquot 49 μ l of the cocktail into each PCR tube, then add 1 μ l of each DNA sample to each tube as well (be sure to check and make sure you actually sucked up one μ l). Now you are ready to PCR.

As you have learned in this course, PCR depends on using the proper annealing temperature (which is different for each primer pair) and extension time (which varies depending on the length of the product).

After amplification, we will remove 15 μ l, add loading dye, and resolve by electrophoresis on an agarose gel.

With luck, our results should look something like this:



Chapter 8: Distinguishing wild from domesticated transposons

Overview: Although transposons are largely selfish, they can sometimes be co-opted to serve a useful function. This process has been called “molecular domestication”. What was a purely selfish gene, whose function was exclusively to make more copies of itself takes on a role in the normal functioning of the host. As a consequence the transposon loses the ability to transpose, and gains new functions. After this shift, the transposon is no longer recognized as “non-self” by the host genome; unlike their “wild” relatives, these genes are not epigenetically silenced. An example is the FAR1 gene in *Arabidopsis*. Although it was isolated in a mutant screen, it is in fact derived from a MULE transposase. The best and final evidence for domestication is a mutant phenotype, like that observed in the *far1* mutant. In this laboratory, we learn how to distinguish between wild and domesticated transposons by looking at differences in targeting by the hosts silencing machinery, expression of the elements, and conservation of the domesticated elements.

In this laboratory we will use the following web sites:

<http://synteny.cnr.berkeley.edu/CoGe/> This is a web site that allows rapid comparisons between regions of related genomic sequence, from individual genes to large regions of chromosomes.

<http://mpss.udel.edu/rice/> This is a web site that shows data for expression of normal transcripts and small RNAs associated with silencing. It can show single gene data or regions of chromosomes.

In addition, we can use <http://sundarlab.ucdavis.edu/smrnas/>, a second data base for small RNAs, and <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>, which can allow us to compare our genes with all other DNA sequences in the data base.

Background: As you have already learned, transposons are rapidly evolving relative to the rest of the genome. In any given genome, vast numbers of elements are competing for space. Some element families rapidly bloom into populations of thousands of elements, while others die off. In any given genome, there are thousands of copies of transposase and reverse transcriptase genes. Although primarily devoted to increasing the copy number of their cognate elements, the enzymatic machinery encoded by transposons represents an opportunity for the host.

Transposase, for instance, is involved in DNA recognition and protein-protein interaction. Not surprisingly, these functions can be, and have been, co-opted by the host. That is, they have taken on a function that increases the fitness of the host, rather than that of the transposon. They move from being “wild” to being “domesticated”.

In the last laboratory, you saw how genomes recognize and epigenetically inactivate transposons. Any given genome contains vast numbers of sequences that are maintained in a silenced state. To do this, small interfering RNAs (siRNAs), DNA methylation and histone modifications are used to keep transposons from becoming reactivated. Analysis of the sequences of databases of siRNAs is a convenient way to tell which sequences the genome thinks are transposons. Domesticated transposons are no longer recognized as such, and so we can use siRNAs to distinguish between wild and domesticated transposons.

We can also use positional stability to distinguish between wild and domesticated elements. This is because transposons must move or they will die. Selection favors elements that can transpose. Any single element that stays at the same place long enough accumulates mutations and, eventually, will no longer be able to transpose. Thus, an active transposon family can be distinguished from a dead family by the number of elements that are very similar to each other that are not in the same position in related organisms. The converse is true of domesticated elements. These elements, like any other gene, stay at the same position because selection at the level of the host favors continued function. They don't have to move to live, because the host needs them where they are. Thus, they no longer need to retain features such as TIRs, and they tend to stay in the same place relative to other genes in related species

To summarize:

Wild:

- Replicates frequently (or dies slowly)
- rapidly evolving
- targeted by siRNAs
- expresses primarily under stress or in the germline
- DNA usually methylated
- Usually present at different positions in related species
- selfish

Domesticated:

- Does not replicate

- evolves more slowly
- not targeted by siRNAs
- expresses in many tissues
- DNA not methylated
- Often present at the same position in related species
- useful

Today we will use a series of clues to pick out which elements in a given family are domesticated and which are wild. We will examine MULEs (Mu-like elements), a Class II DNA type element present in all plants, because a number of domestication events have already been detected (FAR1 for instance). We will examine phylogenies (which show degrees of relatedness and, thus, evidence of recent transposition), comparative positions, expression characteristics and siRNA targeting to fish out examples of domestication. Some of the sequences you will be examining have not been analyzed in this way before, so you will be the first one to know whether or not a given element is wild or domestic. Happy hunting!

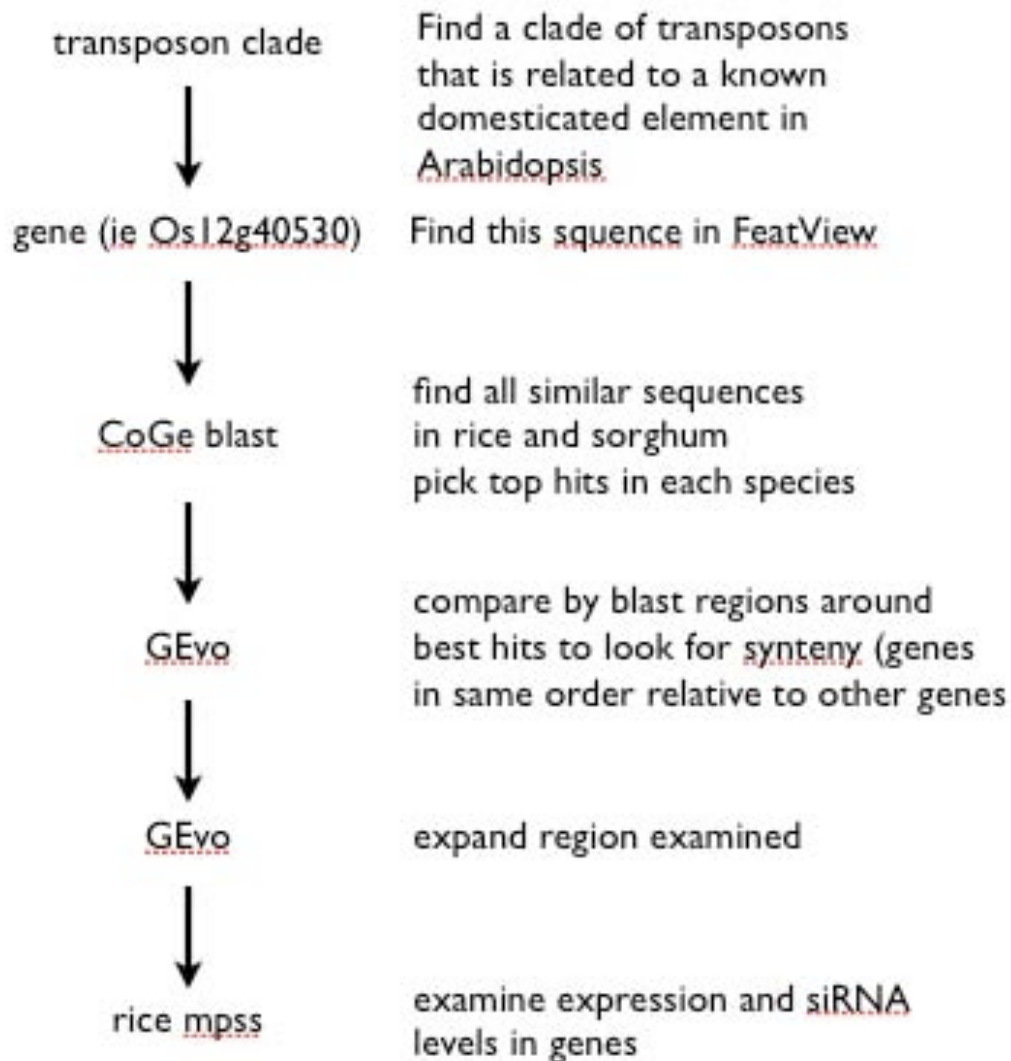
Experimental protocol:

The first thing to do is to identify domestication candidates. We will be using data from a recent paper detailing the phylogeny and expression characteristics (transcription) of MULEs in rice because little is known about transposon domestication in this species, but it has relatives of Arabidopsis MULEs. The first thing we will do is to examine a very nice figure from the rice paper. The whole thing is presented on the next page. A couple of things to notice: On the left is a phylogram showing how the different elements are related. To the right is a line of boxes that are more or less filled in. These represent expression characteristics in various tissues. On the right are rice accession numbers (i.e. Osxxxx). These are your reference numbers for each gene. Notice that some phylogenetic groups, or clades, are producing RNA in many tissues (darker boxes), while some are only producing RNA in a few or no tissues. Note also, that two of the clades that are expressing are actually related to elements in Arabidopsis that are known to be domesticated. Note also that there is a third clade in rice that is expressing but that has not been identified as being domesticated. This is the group we will focus on.



So as to not spoil the surprise, we will walk through the protocol for a member of one of the clades that is related to a domesticated MULE in Arabidopsis (the MUSTANG - MuG- clade). You will use the same protocols to test your rice genes. The red circles indicate buttons to push, in order if they are numbered.

This is a flow-chart to orient you. Each step will be explained in the following pages.



Go to the following web site: <http://synten.cnr.berkeley.edu/CoGe/> then log in as public After logging in, choose FeatView.

Access Point	Description
OrganismView	Search for organisms, get an overview of their genomic make-up, and visualize them using a dynamic, interactive genome browser.
CoGeBlast	Blast sequences against any number of organisms in CoGe.
FeatView	Find and display information about a genomic feature (e.g. gene).
	Compare any two genomes to identify regions of

Type in the number of the rice gene you want to check (get this from the phylogram), and hit “search”. We will use Os12g40530.

Welcome, Damon Lisch Sign-out Home

FeatView Feature Viewer

Feature Selection

Feature Name: Add wildcard to side(s) of the name.

Feature Annotation: Add wildcard to side(s) of the annotation.

Note: wildcards may slow search substantially

Feature Type:

Organism: Name: Description:

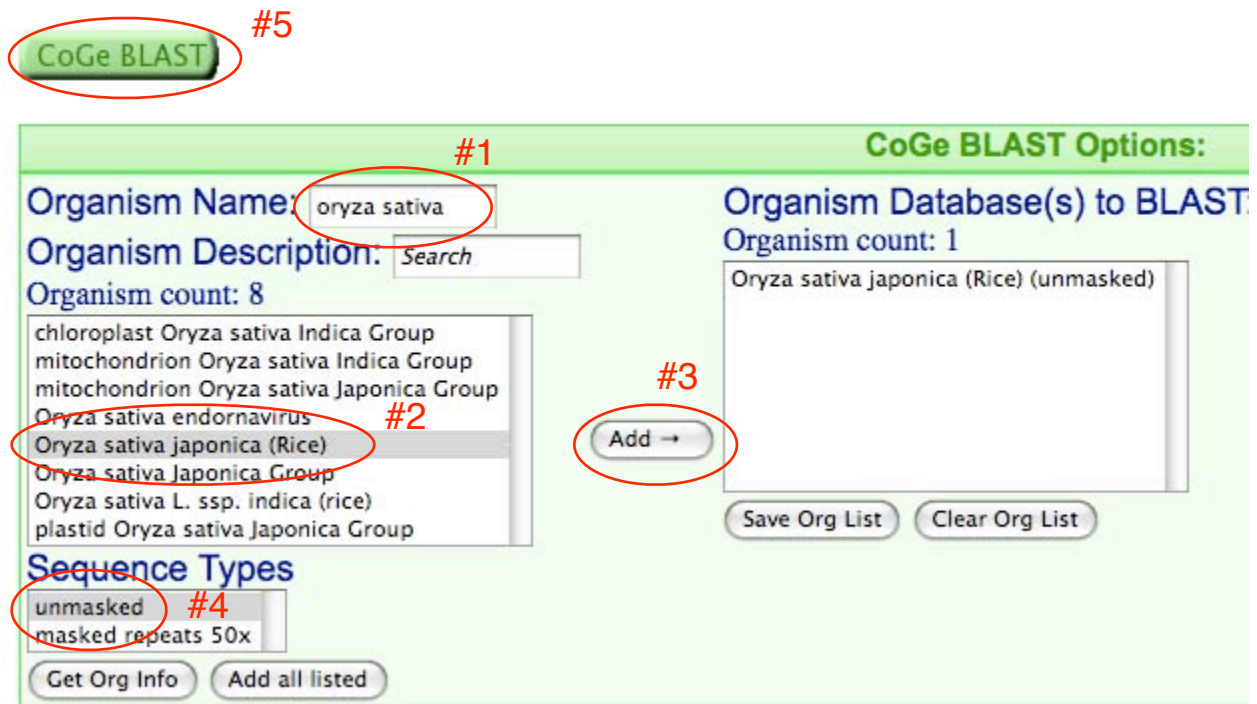
Organism count: 6671

All Listed Organisms
 dwarf virus strain Wheat
 Abaca bunchy top virus (ABTV)

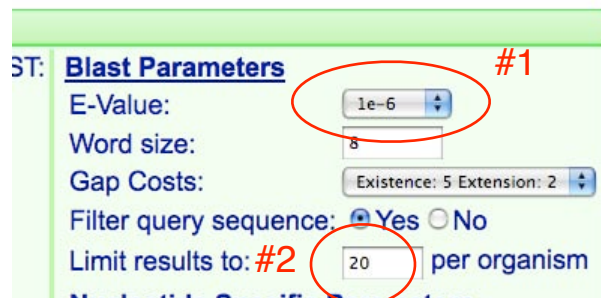
This will take you to a page with a lot of information about this particular gene. At the bottom of the page is a button marked “Blast”. Hit this.

Length: 2250
 Location: Chr 12 25045984-25048233(1)
 Dataset: TIGR masked
 Organism: Oryza sativa japonica (Rice)
 DNA content: GC: 53.11% AT: 46.89%
 Wobble content: GC: 57.47% AT: 42.53%

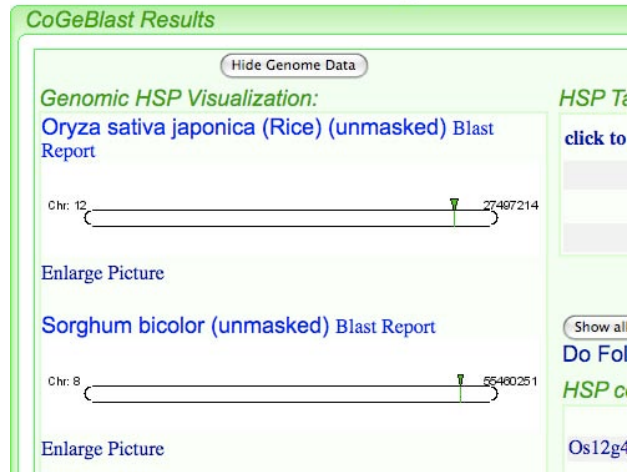
This will take you to a blast interface, where you can use the sequence of the gene in question to blast against selected data bases. In this case, because we want to compare rice and sorghum, we will select these two organisms. To do this, we will type in “oryza sativa” in the indicated box (#1), select oryza sativa japonica from the pull down list (#2) and click on the “add” button (#3).



Be sure that you are searching the unmasked data base (this includes transposons as well as genes). After rice, select “sorghum bicolor” in the same way. At this point you also have the opportunity to change the blast parameters. On the right side of the screen, change the E-value to 1e-6 (#1) and limit results to 20 per organisms(#2). This will give more specificity for your search and limit the number of less significant blast hits. Now we’re ready to blast against rice and sorghum with the sequence we’ve selected. Click on the CoGe BLAST button (#5).



The resulting page will show you your hits. On the left is a graphic representation of the hits on the chromosomes within each species. The triangles localize the hits to a particular positions on each chromosome.



On the right of the screen is information on the nature of each HSP (High Scoring Pair of blast hits). In this case, we only have one hit in each species, which makes things simple, but sometimes you'll see many more hits. In those cases you'll just select the highest two or three from each species. One piece of information is particularly useful, and that is the gene model name associated with each hit. To see those, hit the "show all features" button (#1). This will bring up a series of gene model names on the far right of the screen.

HSP Table: Show HSP Table column display options?

click to sort →	Query Seq	Org	Chr	Position	
#2 <input type="checkbox"/>	Os12g40530	Oryza sativa japonica (Rice) (unmasked)	12	25045984	1
<input type="checkbox"/>	Os12g40530	Sorghum bicolor (unmasked)	8	51246916	1
<input type="checkbox"/>	Os12g40530	Sorghum bicolor (unmasked)	8	51246585	2

#1

Do Following to Checked Features: #3

#2 HSP count: Show Table?

Query Seq Oryza sativa japonica (Rice) (unmasked) Sorghum bicolor (u

Now we will select the hits we want to compare. To do this, just click on the boxes (#2) of the hits. Note that sometimes there are more than one HSP to the same gene model. In this case, only hit one of the boxes. Now we

are ready to compare the selected hits. Hit the “Send to GEvo” button (#3). This will take us to a page where we can blast the selected hits and sequences around them to each other to look for additional similarities.

Options:

Go **Clear** Number of sequences: 2 **Add** **Remove**

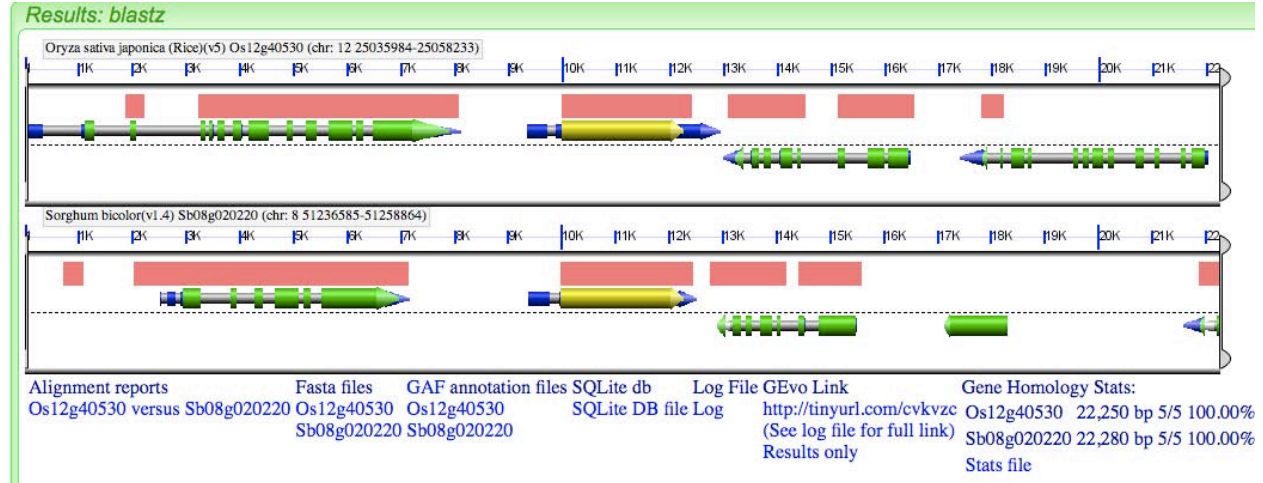
Sequence Submission:

Sequence 1: CoGe Database Name
 Name: Os12g40530
 Oryza sativa japonica (Rice): chr12.xml (TIGR, v5, unmasked) (4)
 (CDS) Chr:12 25045984-25048233 (3)
 Left sequence: 10000
 Right sequence: 10000
 Get Sequence
 Sequence 1 Options:

Sequence 2: CoGe Database Name
 Name: Sb08g020220
 Sorghum bicolor: Sorbi1_assembly_scaffolds.fasta (JGI, v1.4, unmasked) (4)
 (CDS) Chr:8 51246585-51248864 (5)
 Left sequence: 10000
 Right sequence: 10000
 Get Sequence
 Sequence 2 Options:

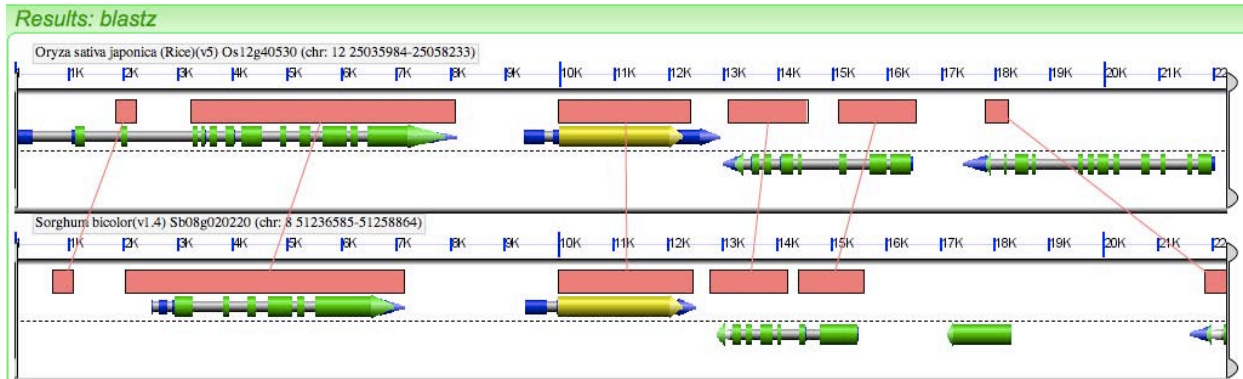
Apply distance to all CoGe submissions? Pad CoGe Sequences with additional sequence: 0
 Open all sequence option menus

The default settings are fine for now, so all we have to do is to hit the “Go” button to make a first pass comparison. Note that the default compares the two selected gene models as well as 10,000 base pairs on either side of them. The output looks like this:



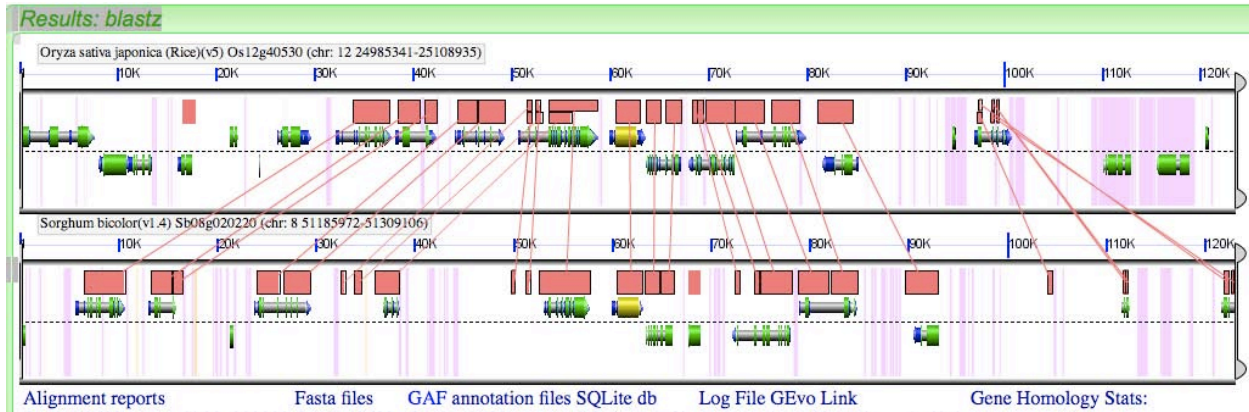
The regions around our genes have been blasted against each other lined up on top of each other, with rice on the top and sorghum on the bottom. The gene models that we started with are colored yellow - other gene models are colored green. The blast hits are portrayed as red boxes above the gene model lines. Note that most of the hits correspond to genes, as expected because sequences that are not coding drift and are therefore no longer similar, but exons are subjected to selection and therefore remain

more similar. If you click on one of the red blocks, a line will be drawn between that block and the corresponding block in the other sequence. This makes it easier to orient ourselves.

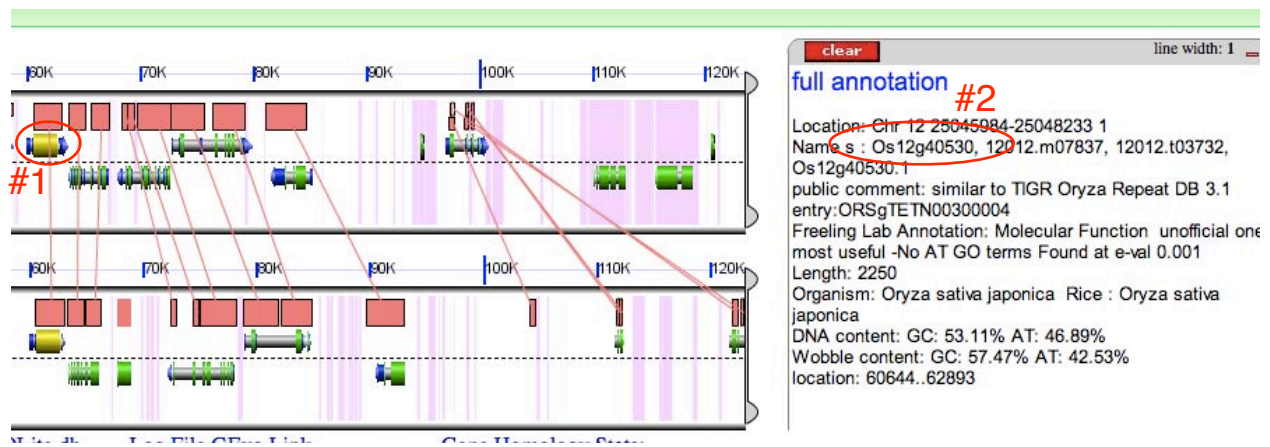


If we want, we can zoom out to compare larger regions of chromosomes. To do this, click on the “Apply distance to all CoGe submissions” button

(#1), change the sequence value to 60000 (#2). We also want to screen out repetitive sequences that will make comparisons more difficult. To do this, change the data bases from “unmasked” to 50x masked (#3). Now hit go (#4). This will lead to a new blast comparison screen. The sequences that are present in more than 50 copies in a given genome are masked out a pink regions - these are all almost certainly wild transposons (use the unmasked data bases to see that they are never in order, or “colinear”). Note that over a wide distance there are gene models that are very similar and in the same order. Note also that our gene is one of those genes. Since this gene was derived from a transposase, we can conclude that this gene is almost certainly a domesticated transposon.



Now let's take a closer look at this gene. First we can pull up the sequence by clicking on the yellow rice element (#1). This will pull up some information in the upper right hand corner of the screen. Highlight and copy the gene name (Os12g4053 (#2)).



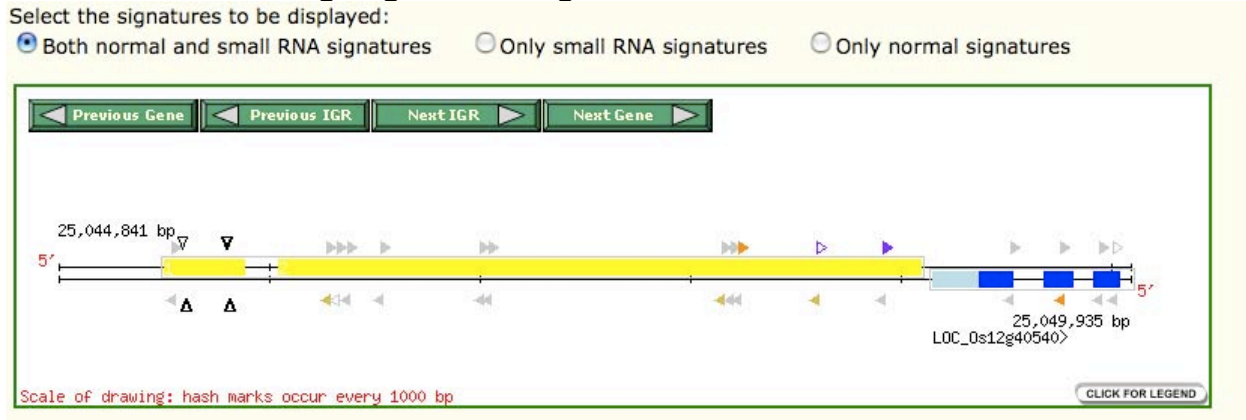
Now we are going to go to a new web site, where we can examine expression characteristics and accumulation of small RNAs (if any) associated with this gene. Open up a new window and go to the following site: <http://mpss.udel.edu/rice/>. Past in the name of the gene into the indicated window. Be sure to include "LOC_" at the beginning of the gene name (#). Now hit "Get Data" (#2).

Next in the box below, enter specific information to retrieve map or data: #1

Protein entry code (e.g. [LOC_Os03g63450](#))
Using an old ID? [TIGR Version Converter](#)

#2

This will bring us to a page that shows us a graphical representation of our gene, along with RNAs that have been detected associated with that gene and the surrounding region. Our gene looks like this:



The color-coded triangles are the data for RNAs. The methodology used (mpss) sequences tens or hundreds of thousands of short chunks of RNA. By size selecting the RNA before hand, it can also distinguish between RNAs from gene transcripts from small RNAs associated with silencing (as discussed in the previous lecture). The color horizontal triangles are associated with normal transcripts; the black vertical triangles represent small RNAs. Note that this gene is expressed, and that it has only a few small RNAs associated with it, but that it has been classified as a potential transposon. Now let's look at the region around this gene.

LOC_Os12g40530	OSJNBa0001B02	12	w	protein coding gene	25,045,341	25,048,935
----------------	---------------	----	---	---------------------	------------	------------

TIGR predicted function: transposon protein, putative, Mutator sub-class, expressed
 This gene is shown in yellow because it was identified as a potential transposon-related sequence by the [TIGR Oryza Repeat Database](#).

No splice variant annotated for this gene.

[Link to BLASTP results](#)

[Link to Rice vs. Arabidopsis BLASTP results](#)

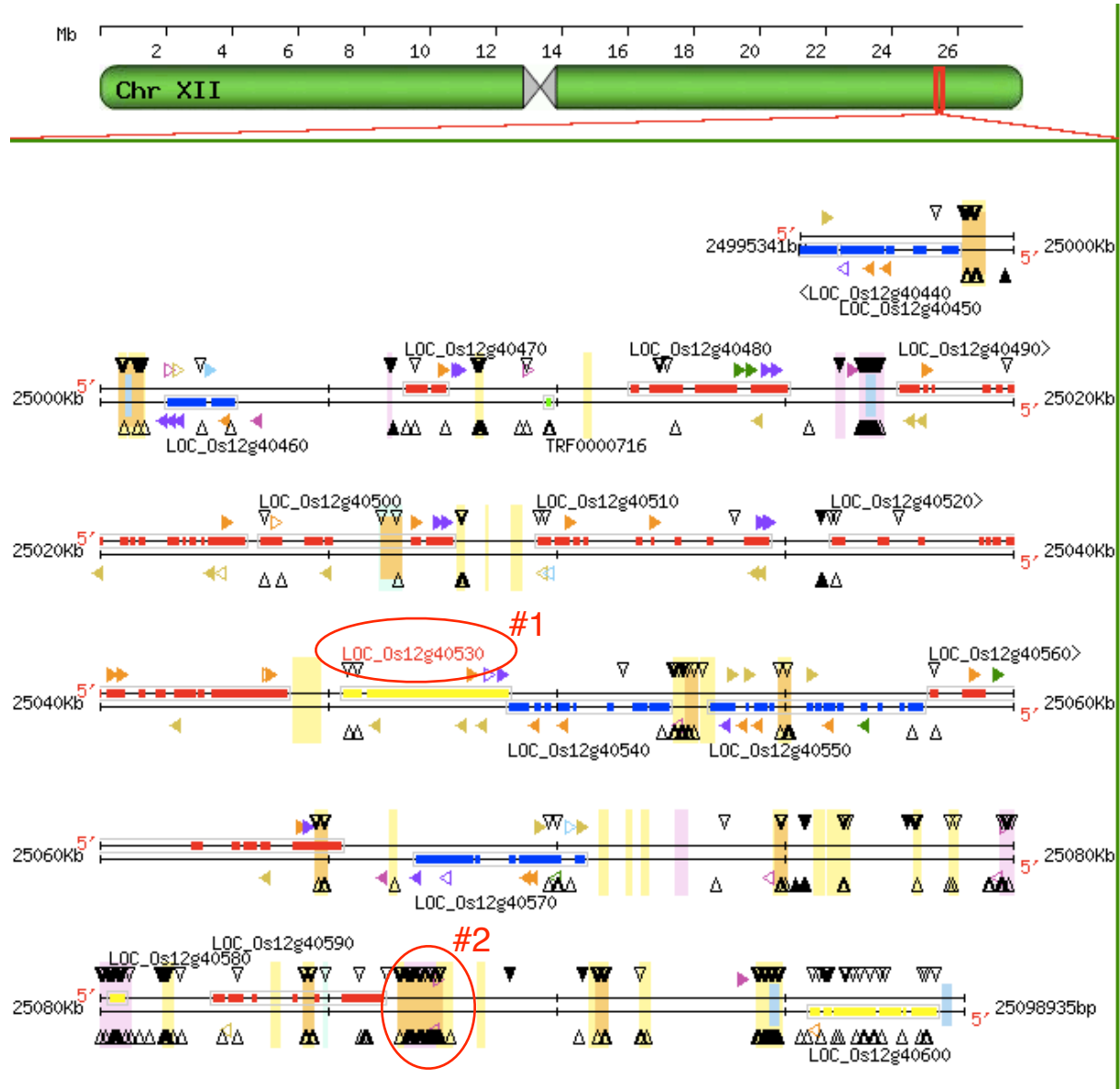
Link to the following for this gene: [TIGR Version Converter](#) [TIGR](#) [OsGDB](#)

[Link to chromosome viewer for region around this gene](#)

[Link to 20 bp signature database for this gene](#)

[Extract sequence for this gene](#)

Click on “Link to chromosome viewer for region around this gene”. This will take you to a a more broad view. In this case, it looks like this:



Our gene is highlighted in red (#1). The transposons in this region are clearly indicated by dense regions with many small RNAs and no normal transcript (#2).

Now let's look at a wild relative of this gene. We're going to use Os02g48930. As before, we pull up the gene using FeatView and blast it against the sorghum and rice genomes. This time, using a 10^{-6} cutoff, we see only hits to rice, suggesting that there is not strong conservation between this gene and those in sorghum. Not only that, but you should notice that there are several hits to different genes at various positions in rice that are nearly identical - this is typical of transposons that are actively duplicating themselves, but not of

domesticated transposons, which have not way of making duplicate copies of themselves.

CoGeBlast Results

Hide Genome Data

Hide Table/Feature Data

Genomic HSP Visualization:

Oryza sativa japonica (Rice) (unmasked) Blast Report

Enlarge Picture

No Hits: Sorghum bicolor (unmasked)

Data Download: HSP Data, Query HSP FASTA File, Subject HSP FASTA File, Alignment File

Analysis Files: SQLite DB file, Blast file for Oryza sativa japonica (Rice) (unmasked), Blast file for Sorghum bicolor (unmasked)

Log File Log

HSP Table: Show HSP Table column display options?

click to sort →	Query Seq	Chr	Position	HSP#	Perc ID	Quality	Closest Genomic Feature
<input checked="" type="checkbox"/>	Os02g48930	2	29921055	1	99.2%	39.9%	Os02g48930
<input checked="" type="checkbox"/>	Os02g48930	7	24423884	2	98.3%	39.5%	Os07g40760
<input type="checkbox"/>	Os02g48930	2	29919373	3	100%	38.1%	Os02g48930
<input type="checkbox"/>	Os02g48930	7	24425567	4	98.8%	37.7%	Os07g40760
<input type="checkbox"/>	Os02g48930	4	16202528	5	96.8%	38.9%	Os04g27730
<input type="checkbox"/>	Os02g48930	4	16200845	6	98.1%	37.5%	Os04g27730
<input type="checkbox"/>	Os02g48930	2	29919680	7	100%	9.5%	Os02g48930
<input type="checkbox"/>	Os02g48930	7	24425259	8	97.4%	9.3%	Os07g40760
<input type="checkbox"/>	Os02g48930	4	16201153	9	97.4%	9.3%	Os04g27730
<input type="checkbox"/>	Os02g48930	2	29919969	10	100%	7.7%	Os02g48930

Select All Select None

Do Following to Checked Features: Send to GEvo Go

Let's go ahead and check off the the genes with high homology to our sequence and blast them against each other, as we did before. The result looks like this:

Results: blastz

Alignment reports Fasta files GAF annotation files SQLite db Log File GEvo Link Gene Homology Stats:

Not surprisingly, since these are two genes in the same genome, these genes are not syntenous. They are, however, nearly identical. To show this, click on one of the larger pink blocks and look at the information panel in the upper right

hand corner of the screen.

The screenshot shows a window titled "full annotation" with a "clear" button and "line width: 1" control. The text inside the window is as follows:

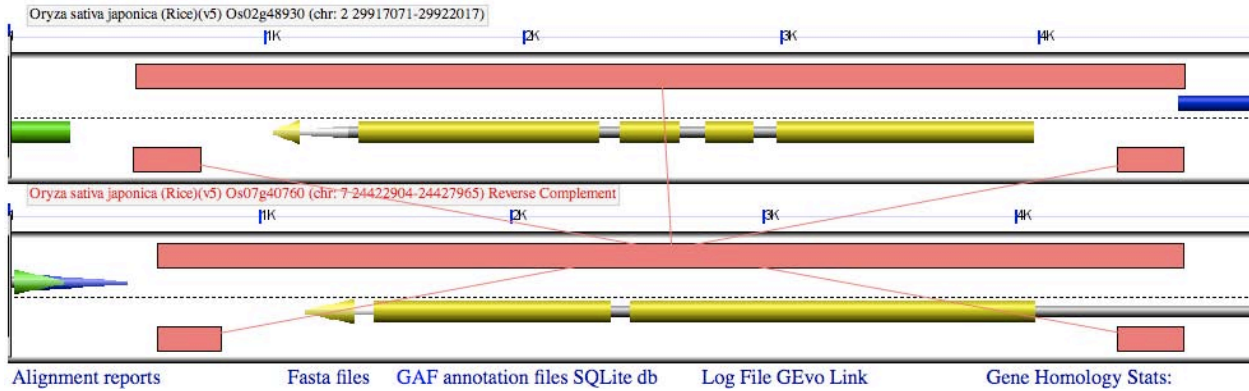
```

HSP: 1  Os02g48930-Os07g40760  reverse complement
Location: 9476-13542  ++
Match: 3998
Length: 4071
Identity: 98.30
E_val: N/A
Score: 373092
Sequence:
CAAGGGGAAAATCCAAATACCCCCCTACAAGTCACTGCTC
TTTCTACTCTCCCCCTACAACCTCAATATTGTTCAAACACT
ACACATAAGTTAATTTCTCTTCCAATTTCCCACCCTATTCGG
TTTTACGCTTTCTCTGATGCGCTCAATTTCTCGTTATCGTG
    
```

This gives you information about the hit (it says “reverse complement” because I reversed one of the sequences. Not that these sequences are 98.3% identical over 4071 base pairs. Finally, we can zoom in using the bars on either side of the genes. This can be useful because what the computer that defined these genes thinks and what is reality are not always the same thing. To do this, click and hold a bar and drag it nearer to the region of homology (sequences covered by pink hits).

The screenshot shows a genomic browser interface with two tracks. The top track is labeled "Oryza sativa japonica (Rice)(v5) Os02g48930 (chr: 2 29908096-29931038)" and the bottom track is "Oryza sativa japonica (Rice)(v5) Os07g40760 (chr: 7 24411050-24436782) Reverse Complement". Both tracks show gene models with exons as boxes and introns as lines with arrows. Two red ovals highlight specific regions of homology between the two tracks. At the bottom, there are links for "Alignment reports", "Fasta files", "GAF annotation files", "SQLite db", "Log File", "GEvo Link", and "Gene Homology Stats".

Now hit “Go” again. We are now only comparing the sequences between the bars for each region of the genome. The result looks like this:



The shorter blocks of sequence are inverted repeats - we can tell because they match the complementary sequence at the other end of the transposon. Now, as before, let's past the name of our gene into the mpss data base:

Select the signatures to be displayed:

Both normal and small RNA signatures
 Only small RNA signatures
 Only normal signatures

◀ Previous Gene
◀ Previous IGR
Next IGR ▶
Next Gene ▶

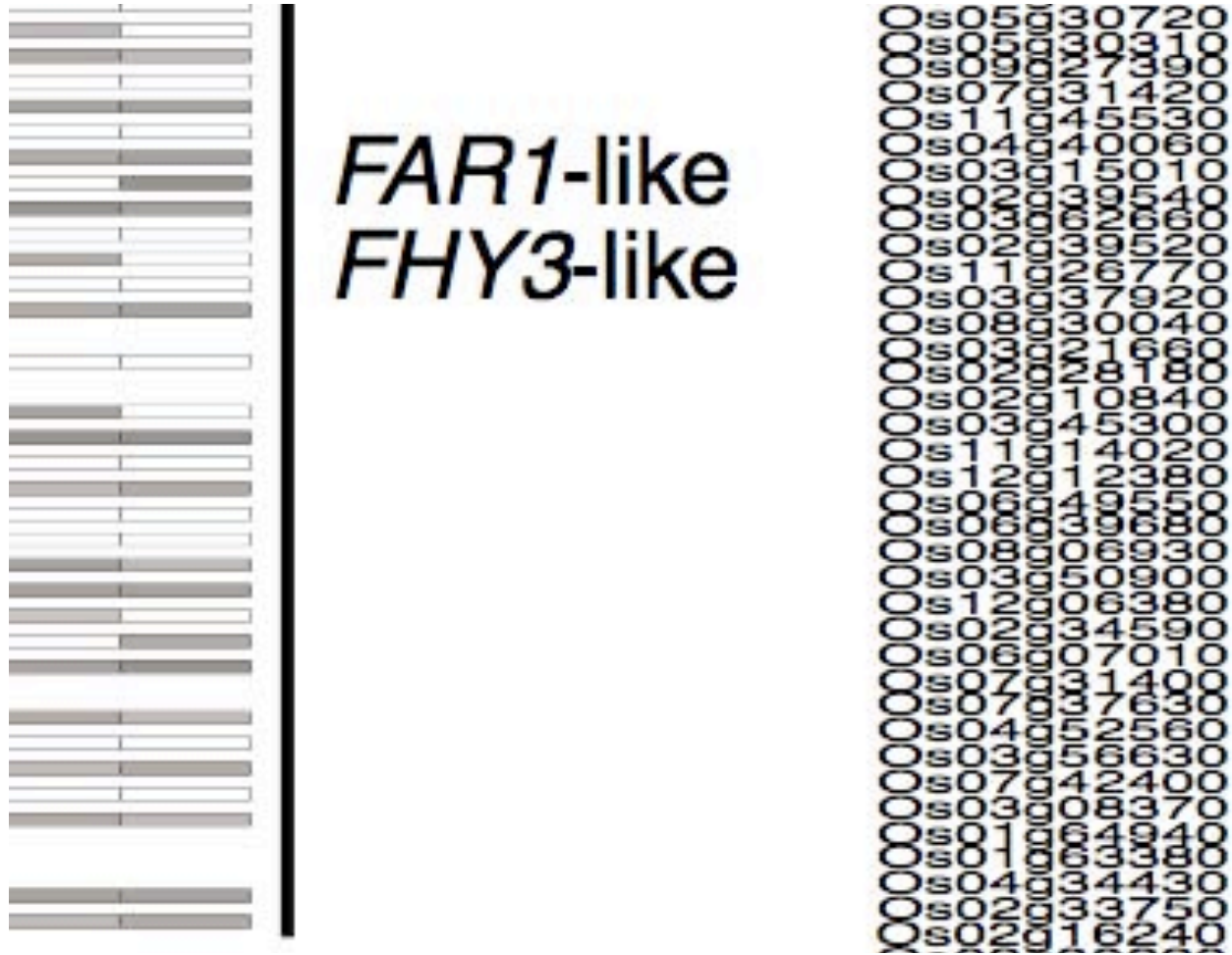
Scale of drawing: hash marks occur every 1000 bp CLICK FOR LEGEND

Note that the gene has low or no expression (horizontal colored triangles, and at least a few small RNA hits (black vertical triangles). Not as deeply silenced as some genes, but not terribly active. Together, these data suggest that this transposon, and all nearly identical copies of it, are “wild”

Conclusion:

These are the basic protocols for finding domesticated transposons. Your assignment is to take a look at the phylogram and choose candidates that are likely to be domesticated. Then use the tools you've learned to determine whether or not you were right. The candidate clades are as follows:

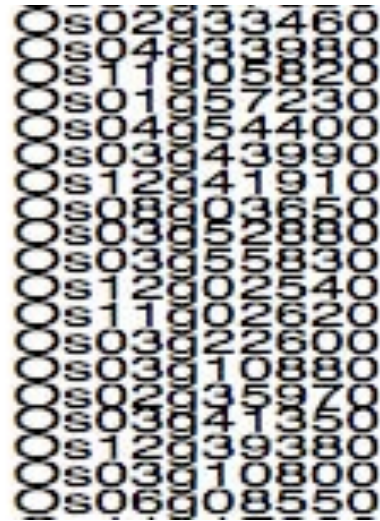
These are related to a gene in Arabidopsis that is known to have been domesticated:



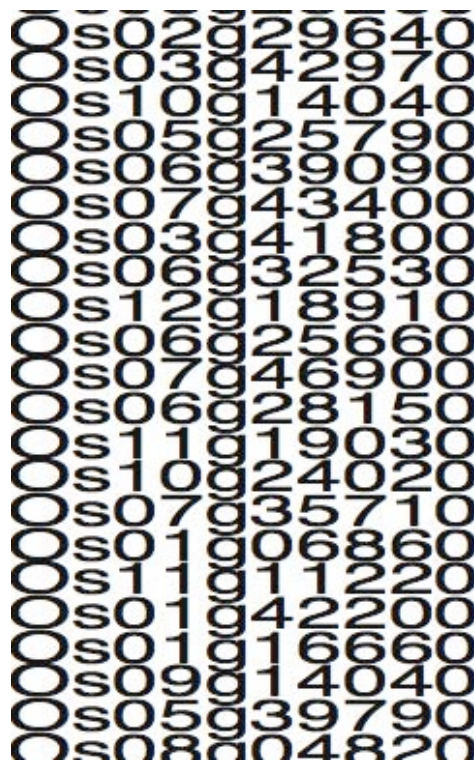
This is a second clade in rise that is similar to the MUSTANG elements in Arabidopsis, some of which may have been domesticated the one we looked at is indicated:



Finally, this is a mystery clade. Although it has the same characteristics as MUG and FAR1-like genes, it has not been characterized as being domesticated



Finally, this is a group of related MULEs from rice that do not express at high levels and are more likely to be wild. You can use them for comparison.



A final note: Be sure to make notes of the genes that you test. Right down which gene you are looking at, and what the results of your analysis are. Include whether or not it is in a syntenous region and whether or not it has small RNAs. It is also useful to take screen shots as you go along so that you can retrace your steps later if you want to.

Steps to annotating transposable elements

Using the Maize Genome Browser at maizesequence.org

The maize genome is very large and has many features: genes, simple repeats, transposable elements, and many other things. To view all of the features in a coherent way genome browsers were developed. CoGe the browser that you used for synteny analysis is one example. The maize genome project chose the Ensembl browser.

Open the browser website <http://maizesequence.org/index.html>.

There are two main entries into the genome browser: **1.** Using a Blast search and **2.** Clicking on a region of the chromosome. We will use a Blast search as our entry point.

Click on Blast located in the Left-hand bar. 1. Copy-and-paste the Pack-MULE sequence in the Query window and select the 2. 'Sequenced Clones' for the database.

```
>Pack-MULE-1
GAGAAAATTGGATTTATGCCATTATGAAACTTAGGATTCGCTCAAATGCC
ATTATGGGAACACGCTTCGCTAAGAGACCACTATCAAACGTTGTCTTACA
GACAAAATGCCATTTAGCGGTTATTTGAAACTTAACCTTATTTGATTAT
ATTGGCAGCTAGTGACCGGACACTTTTGCCCTATCAAATCTGTTCCGCCG
TTCTCCCTCCCCCGCTCTGCCGCCGCTCAGCTCTAGATCAAAACTGCTA
AGCTCCGCCATTACAGGAAAGTCTATTTTACGGCACAGTTTAAAAACAC
AGTTTGCACGTTTTCAGAACACAGTTTCAAACGAACACAAATTCAGAAC
ACAGTCGGCAGACGCTAACACGCGATGGGATACGGCGAACACCGAGCAGC
AGGTATCTAAACTAAAACCTTTGGCAGCACTCTTAACCAGTCAACAACAGA
TCGCCAACCATTTCGATTTTCACACAAATGTAATCCTACTGCAAGTGCAACA
```

```
GCCACCTTTAGAGGCGTTCATTAGCAAACACTAGAAAACTGCAAATGGC
CATATTCATATCGAACCAATATGCCACTAGTTGAATCACGTAAGAGCAG
ACCAATATGCCACTAGTTGAAGAAACAATCAAATCGAACCAAAGTGGAGA
TCTTTACAAATTAATTCGAGATAATACTAGGCAAACCAATTCATATCA
TAGTTTCATTCAATAGGATGGGAGAGCAGGCGATTCGAGCTGTTGATTGG
GAATCTAACCGAGCTGCTCACCTGCAGATTGGAATGGATCGATGGATGG
ACCGCGGACGCCCGACGCCGATTGGATCTCCTCCTCCTTGGGCGGCTGC
TAACTGGGTTGGGATGGGACAGGAGCCTTCCTCGCCACAGCCGCCGTGCC
GTGCTGCTCCTCTTCTCCGAGACGCTGGGAGCCCCGATTGGAGACGCTG
GAAGACGGGAGGCGGAGAGGTATGACGTGACGGAGTTTCGTCTGCGAACT
CCACCGGCGCGTGAACCTTCAGCTCGGCCCTGCAGGAGACGCTGGGAGC
TTCGGCTGCGGCGGGGCAGGAGCCGGTTCGTTAGGCCGAGCAAGCGGGTGC
GGTCCGGATCACCGGGGCACTGGGGGAGGGCCAGCGGCGGTGGGACTG
CCAACGGAGGCGTGAGCTACCCGATGTGCCAGGTGGATGACTGCCAAGCG
GATCTGACCAGCGCCCGCGGTGGAGCTCGCCGAGCGTCCGTGGACCGCC
GCTGGTGGAGGGAGAGCTCAAGCACCGGTGGAGGAGAGCGACGTCGCCG
TCGGTTTTGGTGGAGGGACAGAGGATAGAGAAGAGAGCGCCGTCGGGTT
TGAGATGCCGGCGCTTTTGGTTCGAGGGGCAAAAGTGTCCCCTCATTTGC
TGCCAAATAATCAATAAGGTTAAGTTTCAAAAACTCTCTAAATGTCA
TTTTGTCTATAAGACAACGTTTGATAATGGTTTCTTAGCTAAGCGTGTTC
TTATAATGGCATTAAAGCGAATCCTAAGTTTCATAATGACATAAATCTAA
TTTTCTC
```

New Setup Configure Results Display Refresh Help

Summary

- ▶ setup
ⓘ Not yet initialised
- ▶ configure
ⓘ Not yet initialised
- ▶ results
ⓘ Not yet initialised
- ▶ display
ⓘ Not yet initialised

Enter the Query Sequence

Either Paste sequences (max 10 sequences) in FASTA or plain text:

1 >Pack-MULE-1
GACAAAATGGATTTATGCCATTATGAACTTAGGATTCGCTCAAATGCC
ATTATGGGAACACGCTTCGCTAAGAGACCACTATCAAACGTTCTTACA
GACAAAATGCCATTAGGCGGTTATTGAACTTAACCTATTTGATTAT
ATTGGCAGCTAGTGACCGGACACTTTGCCCTATCAAATCTGTTCCGCC

Or Upload a file containing one or more FASTA sequences
Browse...

Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)
Retrieve

Or Enter an existing ticket ID:
Retrieve

queries

dna queries
 peptide queries

Select the databases to search against

dna database
 peptide database

Sequenced Clones (RepeatMasked) 2
Filtered Gene Set
Sequenced Clones
Sequenced Clones (RepeatMasked)
Working Gene Set
Configure Run >>

Select the Search Tool

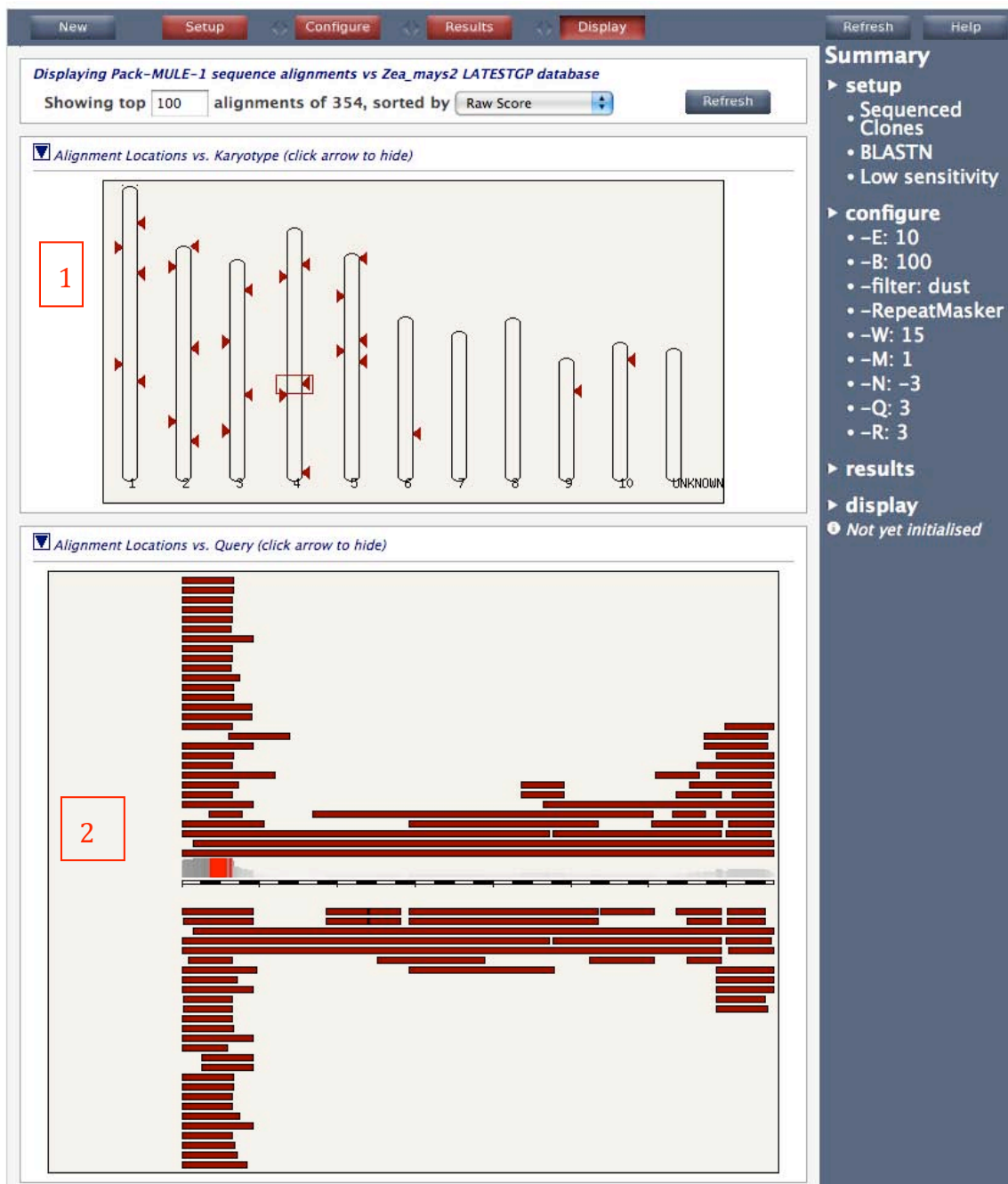
BLASTN
BLASTP
BLASTX

Search sensitivity:
Optimise search parameters to find the following alignments
Near-exact matches

Notes on the Databases:

- a. *Filtered Gene Set*: A set of 40,000 genes predicted by gene prediction software but with Pseudogenes removed. This set is useful if you are looking to see if your TE is inserted into a gene or if you are looking for the parental gene for a captured gene segment.
- b. *Sequenced Clones*: This is the genome sequence.
- c. *Sequence Clones (RepeatMasked)*: This is the genome sequence with the repeats like TEs masked or hidden. Not very useful for TE research!
- d. *Working Gene Set*: Another set of maize genes but before filtering.

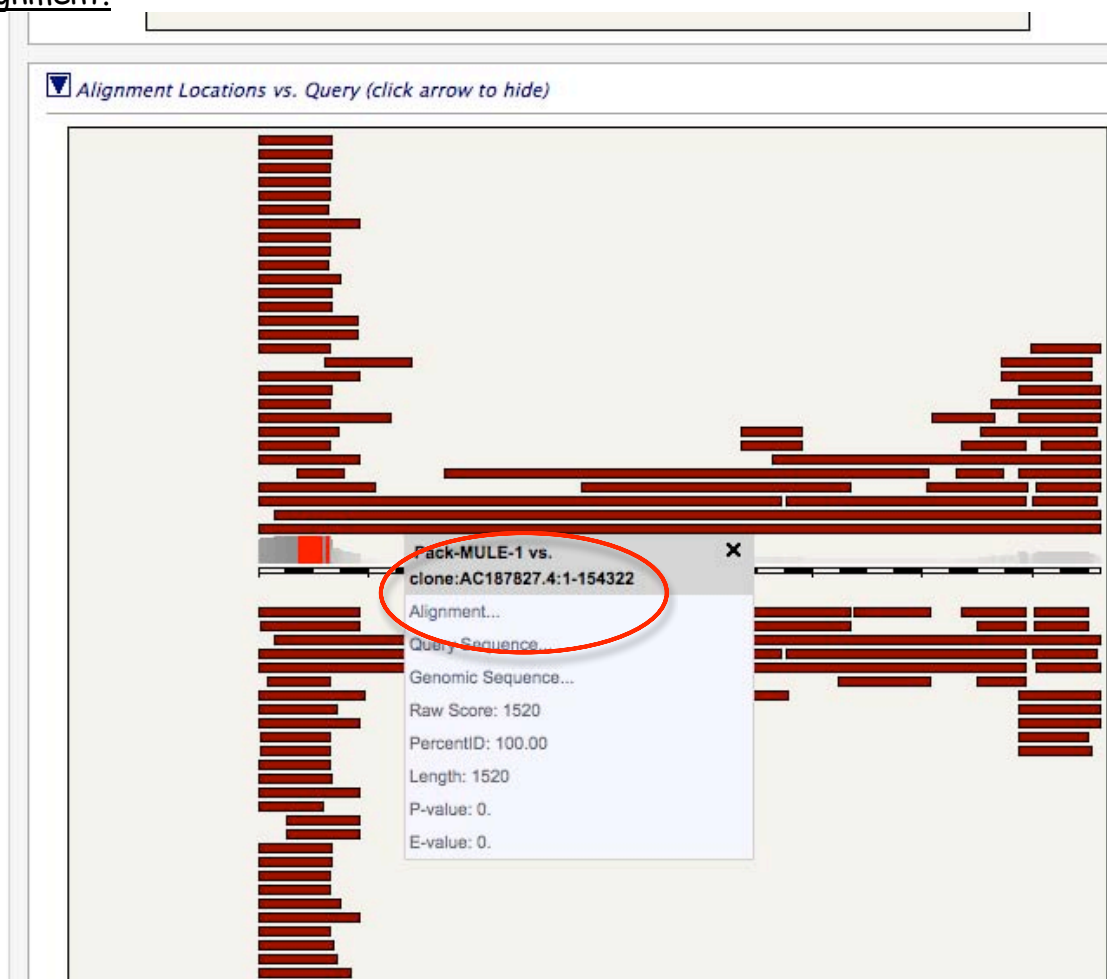
The Blast results are shown graphically in two windows.



1. Blast hits by chromosome. Not very useful for our purposes.

2. Blast hits based on the query sequence. The query is shown in the center as a black/white bar. The Blast hits (subjects) are shown on the top and bottom of the query bar indicating the strand of the hit. The hits closest to the center bar have the lower (better) e-values. In this case we are looking for hits to the full element. There is only one just above the black/white bar.

Click on that hit and a contextual menu opens. You will see that the accession the hit is on is AC187827.4 Click on Alignment and new tab will open with the alignment.



To get to the chromosome view for that hit we need to first add accession numbers to the Alignment Summary. To do this, 1 select Name in the Subject window and 2. click refresh. Your screen should look like this:

Alignment Summary (click arrow to hide)

Select rows to include in table, and type of sort (Multiple options supported)

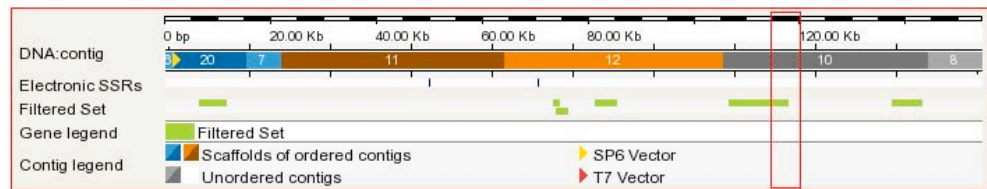
Query: Subject: Clone: Contig: Stats: Sort By:

Buttons: (labeled 2)

Links	Query	Subject	Stats					
	Start	End	Ori	Name	Score	E-val	%ID	Length
[A] [S] [G] [C]	1	1520	+	AC187827.4:1-154322	1520	0.	100.00	1520
[A] [S] [G] [C]	30	1520	+	AC225359.3:1-194219	1344	0.	97.52	1492
[A] [S] [G] [C]	30	1520	-	AC191697.3:1-175461	1344	0.	97.52	1492
[A] [S] [G] [C]	1	1387	-	AC199058.2:1-171383	1239	0.	97.27	1391
[A] [S] [G] [C]	1	946	+	AC186507.4:1-213179	853	0.	97.47	949
[A] [S] [G] [C]	1	946	-	AC185635.4:1-201275	853	0.	97.47	949
[A] [S] [G] [C]	338	1212	+	AC190792.2:1-180902	776	0.	97.05	880
[A] [S] [G] [C]	930	1520	+	AC194635.2:1-154090	461	7.7e-312	94.44	593
[A] [S] [G] [C]	583	1071	-	AC187142.3:1-163048	446	7.5e-255	97.76	490
[A] [S] [G] [C]	583	1071	-	AC213040.3:1-192701	442	2.1e-252	97.55	490
[A] [S] [G] [C]	583	1071	+	AC213040.3:1-192701	438	5.2e-250	97.35	490
[A] [S] [G] [C]	953	1387	+	AC186507.4:1-213179	377	0.	96.57	437
[A] [S] [G] [C]	953	1387	-	AC185635.4:1-201275	377	0.	96.57	437
[A] [S] [G] [C]	583	958	-	AC213032.3:1-189701	306	6.9e-184	95.24	378
[A] [S] [G] [C]	504	780	-	AC209440.3:1-206241	245	1.1e-132	97.11	277
[A] [S] [G] [C]	1	213	+	AC190792.2:1-180902	189	0.	97.18	213
[A] [S] [G] [C]	1	241	+	AC196048.2:1-180601	165	8.3e-121	92.12	241
[A] [S] [G] [C]	1	183	-	AC203051.3:1-188409	159	1.1e-82	96.72	183

3. Click the [C] in the first line. This will open a new tab with the Contig Viewer.

This view will be explained in class. There is a large left margin for you to make notes.



Unfortunately the blast hit is not visible in this view. You will need to go to the alignment tab and get the start and stop positions of the alignment. You should take a screen shot of this page, print it, and draw in your element. Later you can make a fancy PowerPoint figure.

Identifying captured genes and the parental location.

There are several steps:

1. Find captured gene regions.
2. Determine function.
3. Use captured gene blocks to find parental genes.
4. Align Pack-MULE to BAC containing parental genes.
5. Annotate element.

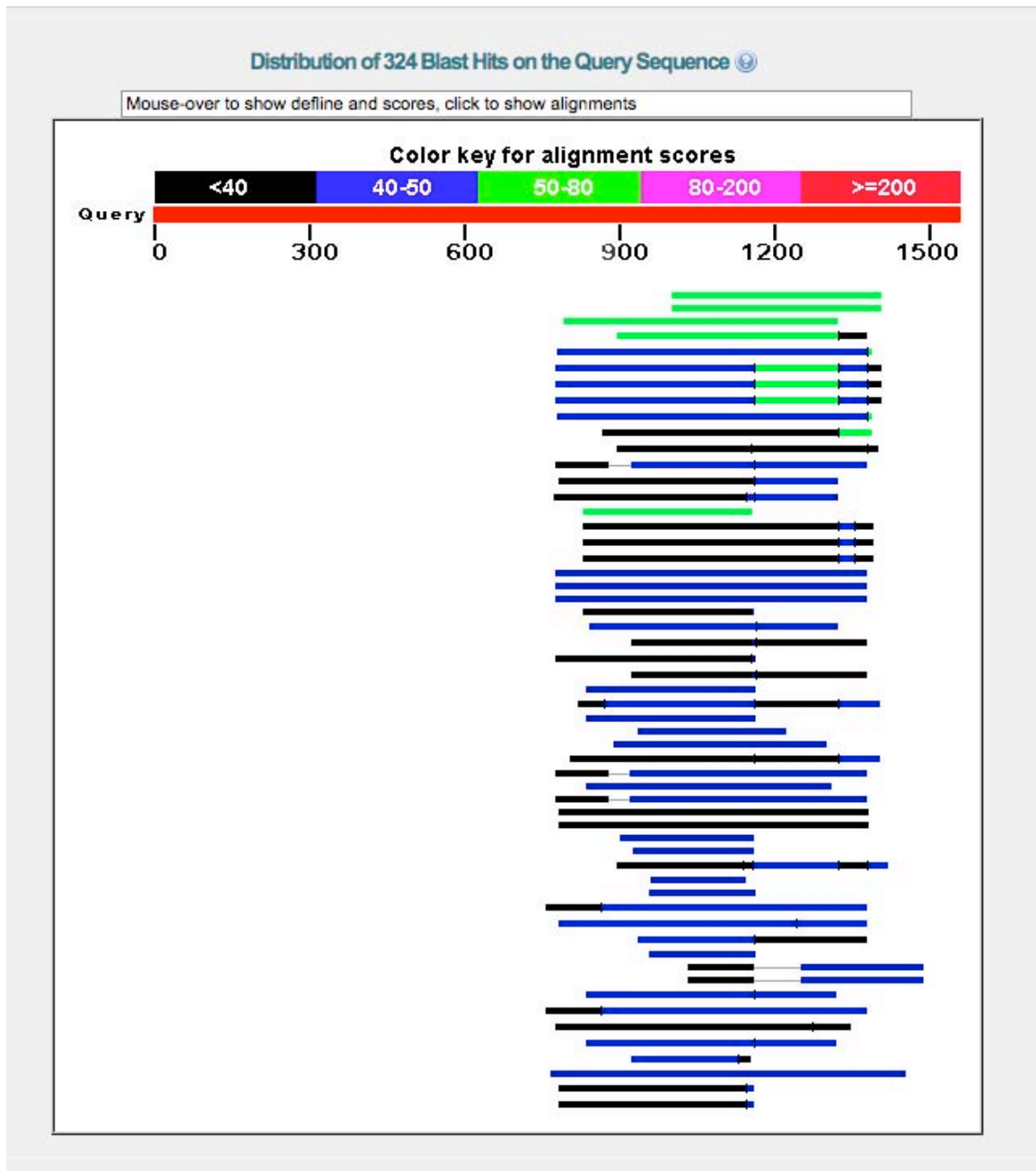
Step 1: Find captured gene regions.

Use blastx to find captured gene regions. Go to the NCBI Blast Website and choose blastx. This flavor of blast translates the query sequence and runs it against a protein database. This will give you a good indication of where captured genes are in a TE.

Open the blast website <http://blast.ncbi.nlm.nih.gov/Blast.cgi> and select blastx.

1. Enter the query sequence and 2. limit the search to *Oryza sativa*. We use the rice database as it is more complete than maize and experience has shown that this gives the best results.

The screenshot shows the NCBI BLAST/blastx web interface. The 'Enter Query Sequence' section is active, showing a text input field containing a FASTA sequence: `>AC187827_00116449_00118005` followed by several lines of nucleotide sequence. A red box labeled '1' highlights the sequence text. Below the sequence input are fields for 'Genetic code' (set to 'Standard (1)'), 'Job Title' (set to 'AC187827_00116449_00118005'), and a checkbox for 'Align two or more sequences'. The 'Choose Search Set' section is also visible, with the 'Database' set to 'Non-redundant protein sequences (nr)' and the 'Organism' dropdown set to 'Oryza sativa (taxid:4530)'. A red box labeled '2' highlights the 'Oryza sativa' selection. At the bottom, there is a 'BLAST' button and a checkbox for 'Show results in a new window'.



Hover the cursor over the alignment bar and sometimes you will be able to see a potential function.

From the top hit copy and paste the Query sequence from that alignment into Word:

```

> ref|NP\_001062227.1| UG Os08g0513700 [Oryza sativa (japonica cultivar-group)]
  dbj|BAF24141.1| G Os08g0513700 [Oryza sativa (japonica cultivar-group)]
  Length=1104

  GENE ID: 4346023 Os08g0513700 | Os08g0513700 [Oryza sativa Japonica Group]
  (10 or fewer PubMed links)

  Score = 66.2 bits (160), Expect = 2e-10
  Identities = 58/165 (35%), Positives = 74/165 (44%), Gaps = 32/165 (19%)
  Frame = +1

  Query  1003  TGALNLQLGLQEDAGS-----FGCGGAGAGR*AEQAGAVR-----IT  1113
             +C LNLQLGL+EDA +                A   Q  VR      +
  Sbjct  106    SGGLNLQLGLREDAATPMDVSPAATTVSSSPSPPASSAPAQEPVVRPSKRVRSGSPGSAS  165

  Query  1114  GGTGGGPGGGGTANGGVSYPMCQVDDCQADLTSA----RRWSSPSVRG---PPLVEGELQ  1272
             GG GGG GGG +  GG SYPMCQVDDC+ADLT+A   RR      + G   LV  ++Q
  Sbjct  166    GGGGGGGGGNSGGGGGSYPMCQVDDCRADLTNAKDYHRRHKVCEIHGKTTKALVGNQM  225

  Query  1273  APVEESDVAVGFGGGDRG*RRRERRV*DAGAFGSRGKSVPSLAAN  1407
             ++                D G R  RRR+  AG   R K+ P+  A+
  Sbjct  226    RFCQQCSRPHPLSEFDEGKRSCRRL--AGHNRRRRKTPQPTDVAS  268
    
```

You will now have (the '-'s and '*' are fine):

```

TGALNLQLGLQEDAGS-----FGCGGAGAGR*AEQAGAVR-----IT
GGTGGGPGGGGTANGGVSYPMCQVDDCQADLTSA----RRWSSPSVRG---PPLVEGELQ
APVEESDVAVGFGGGDRG*RRRERRV*DAGAFGSRGKSVPSLAAN
    
```

You can use this sequence in the maize genome browser to find the parental gene. More on that next time.

A. Obtain the DNA sequence of the genomic region around your TE.

1. Blast the sequence in the Maize Genome Browser (pages 157-161). In the "Alignment Summary" click on the [G] link for the genomic region.

2. Change the limits to 500. This will give you the TE sequence with 500 nucleotides added to the beginning and the end. Click "Update."

Query location : Hip1_51 191 to 7774 (+)
 Database location : clone:RC200552.4:1-228875 153375 to 160958 (+)
 Genomic location : 10 121916329 to 121923912 (+)

Alignment score : 7584
 E-value : 0.
 Alignment length : 7584
 Percentage identity: 100.00

5' Flanking sequence (bp)
 3' Flanking sequence (bp)

Coordinate system
 Orientation
 Alignment markup
 Feature markup
 Line numbering

THIS STYLE: Location of selected alignment
THIS STYLE: Location of other alignments
THIS STYLE: Location of Exons

```
>fpc_pseudomolecule:BAC_clones:10:121915829:121924412:1
CGATTTATTTTGCCTAGGCCATCGCCCAAGTGGTCTCTATATATATACACGTTAGAAGG
TACCAAAATAGTATGGTCTATATCAAACGATTACTGGTCGGTCGAAATTTATCTGGTT
TTATAATTAATCTATTTACTGGTATAATTTGATAGTTATAGTTATTTAAACTTAAATTT
CAAATACGGCCATTTTCGAAAACGATATTGTCTTTAGTAAATGGTACGGTGACGAGACCG
GAGGATTTTCGTTACTGTTTTATCGCTACAGGCAGCTCTCCATTACCGGAGGCAGA
ACCGCAGAAATCTCTACTACTTATTAAGTAAGCAATAGTAGTCTGCCCTCTGACAAGTCT
GCCGTTCTGCCGTTCTGCCCTCTGAACCCGTTCTGCTCAGAACCGCGCCAAAAAATCTT
GTTCTTGCAACCAGAGACATTAGTCCGAGCAACAACAACACACCAACGTTGCTGAGC
AACAACCAACACACCAACGCTTATAAGTCAAGTCAATGCTCTGGGTGGCCGGTATG
GCACTAATCAAATAATACCATAGCAGTGCCTGTGTTGAGCGGTGTATGCGGCCCAACGA
TGGTGCGGCCATTCTTACGCCCATCTATACGGCCTGCCCTTACACCCTAGGCTAC
GGCGCTGCCGCTCCACTTAGCGACGGCGCTCTGCAGGATGCCACGACCTCGACCTC
CTCCCCACGACACGGCCTCAACCGCTCCGCGAGCTCCGECATTTACGCACACGACTTC
AACCTCGCCGATTCAAGCGCACGACGACTTCAACGGCGCGCAGGCTCAGGCTCCGCGAC
GGCCGCCACCTCGCCGAGAGCGGCGTCCCCGGGATCAAGCTAGGTTCCGCGTCTGCTTC
TCCACGGATTCAATTGGCTCCGAGAGGACACCGTCCGCGCTCCAGGTACCGACCGAT
CCAAGCACCAACACCGCAATGGGGAGGTCACGACCCCGCTCTCTCCCGCT
CTCCTCCGACGCTGGCTCTCCCGCACTTCCCTCCCGCTCGCGGACGCGCGCGCGG
CGCGTGGACCTGCTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTTCACTT
```

3. Copy-and-paste the sequence into a word document. Save it.

4. Draw a sketch of your element and label the new coordinates. The TE sequence begins at 501 and ends at 501 + the length of the element.

B. Primer Design using the web program Primer 3 Plus

1. Open the Primer3Plus website: <http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi>.

1. Give the search a name in "Sequence Id."
2. Copy the genomic sequence into the large textbox.
3. In "Targets" enter 470, 50. This means that the PCR product must include the nucleotides 470-520. Since nucleotide 501 is the start of the TE, these primers will flank the end of the TE.

Primer3Plus
pick primers from a DNA sequence

[Primer3Manager](#) [Help](#)
[About](#) [Source Code](#)

Task: Detection Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified.

Main **General Settings** **Advanced Settings** **Internal Oligo** **Penalty Weights** **Sequence Quality**

1 **Sequence Id:** >Hip1_51 genomic +-50C

2 **Paste source sequence below** Or upload sequence file: no file selected

```
CGATTTAAATTGCCTAGGCCATCGCCCAAGTGGTCTATATATATACACGTTAGAAGG
TAGCAAAATAGTATGGTCTATATCAAACGATTACTGGTCGGTCGAAAATTTATCTGGTT
TATAAATAATCTATTTACTGGTATAATTGATAGTTATAGTTATTTAAACTTAAATTT
CAATACGGCCATTTTCGAAAACGATATTGCTTTAGTAAATGGTACGGTGACGAGACCG
GAGGATTTTCGTTACTGTTTTTATCGCTACAGGCAGCTCTCCATTACCGGAGGCAGA
ACCGCAGAAATCTACTACTTTATAAGTAAGCAATAGTAGTCTGCCTCTGACAAGTCT
GCCGTTCTGCCGTTCTGCCTTGAACCCGTTCTGCTCAGAACCAGCCGCAAAAAATTTCT
GTTCTTGCAACAGACATAGTCCGAGCAACAACAACAACAACAACAACAACAACAACA
AACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAACA
ACACTAATCAAAATACCATAGCAGTGCCTGTGTTTCGAGCGGTGTATGCGGCCCAACGA
TGGTCGGGCCCATTTCTCAGCCATCCTATACCGCGCTGCCTTCTACACCTAGGCTAC
GGCGCTTCGGCTCCACTAGCGAGCGGCTCTGCAGGATGCCACGACCTCGACCTC
```

3 **Excluded Regions:** < >
Targets: [470,50]
Included Region: { }

Pick left primer or use left primer below. Pick hybridization probe (internal oligo) or use oligo below. Pick right primer or use right primer below (5'->3' on opposite strand).

2. Click on the "General Settings" Tab.

Primer3Plus
pick primers from a DNA sequence

Primer3Manager Help
About Source Code

Task: Detection

Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified.

Main General Settings **Advanced Settings** Internal Oligo Penalty Weights Sequence Quality

Product Size Ranges | 150-250 100-300 301-400 401-500 501-600 601-700 701-850 851-1000

Primer Size Min: 18 Opt: 20 Max: 27
 Primer Tm Min: 57.0 Opt: 60.0 Max: 63.0 Max Tm Difference: 100.0
 Primer GC% Min: 20.0 Opt: Max: 80.0 Fix the 5 prime end of the primer
 Concentration of monovalent cations: 50.0 Annealing Oligo Concentration: 50.0
 Concentration of divalent cations: 0.0 Concentration of dNTPs: 0.0

Mispriming/Repeat Library: NONE

Load and Save
 Please select special settings here: Default (use Activate Settings button to load the selected settings)
 To upload or save a settings file from your local computer, choose here:
 no file selected

3. In the field for "Product Size Ranges" delete the 150-200, and 100-300 ranges. These are too small for standard PCR. The smallest product we want is around 300 bp. Leave the rest of the settings alone.

Primer3Plus
pick primers from a DNA sequence

Primer3Manager Help
About Source Code

Task: Detection

Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified.

Main **General Settings** **Advanced Settings** Internal Oligo Penalty Weights Sequence Quality

Product Size Ranges | 301-400 401-500 501-600 601-700 701-850 851-1000

Primer Size Min: 18 Opt: 20 Max: 27
 Primer Tm Min: 57.0 Opt: 60.0 Max: 63.0 Max Tm Difference: 100.0
 Primer GC% Min: 20.0 Opt: Max: 80.0 Fix the 5 prime end of the primer
 Concentration of monovalent cations: 50.0 Annealing Oligo Concentration: 50.0
 Concentration of divalent cations: 0.0 Concentration of dNTPs: 0.0

Mispriming/Repeat Library: NONE

Load and Save
 Please select special settings here: Default (use Activate Settings button to load the selected settings)
 To upload or save a settings file from your local computer, choose here:
 no file selected

4. Click on the "Advanced Settings" Tab. Set the GC Clamp to 1. This will require that the 3' nucleotide be a G or C.

Primer3Plus
pick primers from a DNA sequence

Primer3Manager Help
About Source Code

Task: Detection Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified. **Pick Primers** Reset Form

Main **General Settings** Advanced Settings Internal Oligo Penalty Weights Sequence Quality

Max Poly-X: 5 Table of thermodynamic parameters: Breslauer et al. 1986
 Max #N's: 0 Salt correction formula: Schildkraut and Lifson 1965
CG Clamp: 1
 Number To Return: 5
 Max Self Complementarity: 8.00 Max 3' Self Complementarity: 3.00
 Max Repeat Mispriming: 12.00 Max 3' Stability: 9.0
 Max Template Mispriming: 12.00 Pair Max Repeat Mispriming: 24.00
 Left Primer Acronym: F Internal Oligo Acronym: IN
 Right Primer Acronym: R Primer Name Spacer: _

5. Click the green "Pick Primers" button.

6. Results come back quickly. The 'best' primer pair is listed at the top. Other good pairs are listed further down the page. The first pair is good and would be the ones ordered. The "Left" or "Forward" primer is highlighted in purple and the "Right" or "Reverse" primer is highlighted in yellow.

Primer3Plus
pick primers from a DNA sequence

Primer3Manager Help
About Source Code

< Back

Pair 1:
 Left Primer 1: >Hip1_51 genomic -500_F
 Sequence: CCGGAGATTTCGTTACTG
 Start: 238 Length: 20 bp Tm: 59.6 °C GC: 50.0 % ANY: 4.0 SELF: 1.0
 Right Primer 1: >Hip1_51 genomic -500_R
 Sequence: TAGGATGGCTGAAGAATGG
 Start: 629 Length: 20 bp Tm: 60.0 °C GC: 50.0 % ANY: 2.0 SELF: 0.0
 Product Size: 392 bp Pair Any: 4.0 Pair End: 0.0

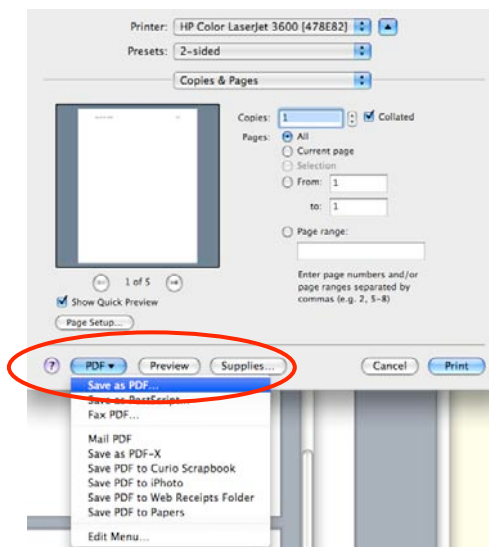
Send to Primer3Manager Reset Form

1	CGATTATTTT	TGCCTAGGCC	ATCGCCCCAA	GTGGTCTCTA	TATATATACA
51	CCTTAGAAGG	TACCAAAATA	GTATGGCTTA	TATCAAACCG	ATTACTGGTC
101	GCTCGAAAT	TATCTGGT	TATAAATA	TCTATTACT	GGTATAATTT
151	GATAGTTATA	GTTATTTAAA	ACTTAAATTT	CAAAACGGC	CATTTTCGAA
201	AACGATATTC	TCTTTAGTAA	ATGGTACGGT	GACGAGACGG	GAGGATTTTC
251	GTACTGTGTT	TATCCCTAC	AGGCAGCTCT	CTCCATTCAC	CGGAGGCAGA
301	ACCGCAGAAA	TCTCTACTAC	TATTAAGTA	AGCAATAGTA	GTCTGCCCTCT
351	GACAAGTCT	CGCCTCTGCG	CGTCTGGCT	CTCAACCCGT	TCTGCTCAGA
401	ACCGCCCAA	AAAATTCCTT	GTCTTCGAA	CCAGACGAT	TACTCCGAGC
451	AACAACGAC	ACACACCAAC	GTGTCTGAGC	AACAACGAC	ACACACCAAC
501	GCTTATAGT	CACGTTTCAA	TGCTCTGGGT	GGCCGGTATG	GCACATAACA
551	AATAATACCA	TAGCAGTGCC	TGTGTTGAG	CGGTGTATGC	GGCCCAACGA
601	TGCTCGGCC	CATTCCTCAG	CCCATTCCTAT	ACCGCGCTG	CCTTCTACAC
651	CCTAGGCTAC	GGCCCTGCC	GCCTCCACTC	TAGCAGCGCC	GCTCTCGAGG
701	ATGCCACGA	CCTCGACCTC	CTCCCCACG	ACCACGGCCT	CAACCCGCTC
751	GCCAGCTCCG	CCATTTACGC	ACACGACTTC	AACCTCGCCC	GATTCAGCGC
801	ACCAGCACT	TCAACGGCC	CGAGCTCAG	CGTCCGCGAC	GGCCGCAACG

Typically primers are designed in pairs. A good pair of primers will meet these criteria:

1. Each primer should be 18-25 base pairs long.
2. The 3' nucleotide of the primer should be a *G* or *C*. This is called the *G/C* clamp.
3. The *G/C* content should be around 50%.
4. The melting temperature (T_m) of each primer should be close to 60°C. The difference between the T_m s of a primer pair should 5°C or less.
5. Neither primer should form a hairpin.
6. Primers should not be self-complementary.
7. Partners should not complement.
8. All primer sequences are reported in 5' to 3' direction.
9. The product size should be between 200 and 1000 nt for Taq DNA Polymerase.

Save this document. On a mac you can create a PDF by selecting File->Print. In the Print dialog box select save as PDF. Select a location save the file.



Before ordering your primer, use the annotated genomic region for the element to make sure that no primers are in transposable elements that surround your element.

Primer 3 rarely fails to find good pairs. However, if Primer3 cannot find primers you can use the statistics table on the results page to modify and repeat the search. The statistics tell you how many primers were considered and why they were rejected. This information can be used to modify the parameters of Primer 3.

Statistics:	
Left Primer:	considered 13422, GC content failed 567, GC clamp failed 1853, low tm 4667, high tm 3141, high end compl 15, long poly-x seq 69,high 3' stability 222, ok 2888
Right Primer:	considered 13442, GC content failed 567, GC clamp failed 1855, low tm 4549, high tm 3300, high end compl 3, long poly-x seq 74,high 3' stability 224,high template mispriming score 1, ok 2869
Primer Pair:	considered 1492, unacceptable product size 1475, ok 17

7. To order the primer, go to the course Data webpage. There you will notice a short form to fill out. You must enter all fields of the form for each primer.

Primer Order Form

Enter each primer separately. All fields

* Required

Student Name *

Primer Name *
Do not use spaces. End primer with '_F'

Primer Sequence *

Comments *
Short description of purpose.

Target *
Gene or TE primer targets

Powered by [Google Docs](#)

[Terms of Service](#) - [Additional Terms](#)

8. Next you will design primers that span the element. To do this you will need to "remove" the element from the genomic sequence. You can use DNA Extracter: <http://www.molecularworkshop.com/programs/xtrct001.html> to do this. Open the webpage and copy in the genomic sequence. Enter 1-500 for the beginning of the sequence and the coordinates of the last 500 nt for the end. Click "Submit Data."

DNA EXTRACTER (ver 1.01)

Extracts a subsequence from a larger sequence.

last modified 16/10/01

Sequence Name:

Enter (or paste) sequence below:

1

```
CGATTGATTTGCTAGGCCATCGCCCCAAGTGGTCTCTATATATATACACGTTAGAAGG
TACAAAATAGTATGGTCTATATCAAAACGATTACTGGTCGGTCGAAAATTTATCTGGTT
TTAATAATTAATCTATTTACTGGTATAATTTGATAGTTATAGTTATTTAAAACCTAAATTT
CAATACGGCCATTTTCGAAAACGATATTTGCTTTAGTAAATGGTACGGTGACGAGACCG
GATGATTTTCCTTACTGTTTTTATCGCTACAGGCAGCTCTTCATTCACCGGAGGCAGA
ACCCAGAAATCTCTACTACTTATTAAGTAAGCAATAGTAGTCTGCCTCTGACAAGTTCT
GCCGACTGCGGTTCTGCCTCTGAACCCGTTCTGCTCAGAACCGGCCAAAAAATCTT
```

The coordinates of a desired contiguous subsequence are the first and last nucleotide
12-1948
If you wish to extract and join noncontiguous segments (ie: an ORF from genomic I
1-337,365-894,907-1899

2

Enter coordinates of subsequence:

1-500, 8084-8503

3

Submit Data Clear Fields

9. The new sequence is at the very bottom of the results page. Copy this into a word document and save it. This sequence represents an "Empty Site."

10. Repeat Primer3Plus to design primers that flank the insertion site. This time we will tell it to use the forward primer you picked earlier. After pasting the "Empty Site" sequence into the sequence window. Choose the same Target Region. Finally, paste the Forward primer from before into the box as shown.

Primer3Plus pick primers from a DNA sequence		Primer3Manager	Help
		About	Source Code
Task: Detection	Select primer pairs to detect the given template sequence. Optionally targets and included/excluded regions can be specified.		Pick Primers Reset Form
Main	General Settings	Advanced Settings	Internal Oligo
Sequence Id: >Hip1_51 genomic -500			
Paste source sequence below		Or upload sequence file: Choose File no file selected Upload File	
<pre>CGATTATTTTCCTAGGCCATCGCCCAAGTGGTCTCTATATATACACGTTAGAGGTACCAAAATAGTATGGTCTA TATCAAACGATTACTGGTCGGTCGAAAATTTATCTGGTTTTATAATTAATCTATTACTGGTATAATTTGATAGTTATA GTTATTTAAAACTTAAATTTCAAATACGGCCATTTTCGAAAACGATTTGTCTTTAGTAAATGGTACGGTGACGAGACCG GAGGATTTTCGTTACTGTTTTATCGCTACAGGCAGCTCTCTCCATTACCGGAGGCAGAACCCGAGAAATCTCTACTAC TTATTAAGTAAGCAATAGTAGTCTGCTCTGACAAGTCTGCGCTTCTGCGGTTCTGCGCTGAAACCGTTCTGCTCAGA ACCGGCCAAAAAATCTTGTCTTGCACCAGAGACATTAGTCCGAGCAACAAACACACACCAACCAACGCTCTGAGC AACAACCAACACACCAACATTTTTGTTTTTAAATCGATGTAGTGTGTTTTGCGTTACAGTCCATTGTGGCTAATGGG GGCGGCACATGTAGTGTGTAGTCAATCGTGAATTTAATAGTAACIAGCTCTGCAGTTGATTTGTTTATGAATGTGTTG GTTGGACCTATCAATTGGATGTTTTGTTGGCTACATCACATATCTTTTAAAGTTTACTTTGTGTGGTCTGATGTTGTAT TTATAAAGTGCAGGTTTCATGTTTTAACTCCCGTGGCAACGCACGGGCATATACCTAGTAAAAAATATGGAGTTGACAAAA AAAGCTGGGCAGGACCGAGGACACGGACAGGGCACGTGATCCAAGATTCGAAGGTGCGTCCGGCGTGTCTGCGGCTCGT TGCCCTCCCTGCTGGCTGCGTAGATGACGGTCTTTGC</pre>			
Mark selected region: <> { } Clear		Save Sequence	
Excluded Regions: < >			
Targets: [470,50]			
Included Region: { }			
<input checked="" type="checkbox"/> Pick left primer or use left primer below.	<input type="checkbox"/> Pick hybridization probe (internal oligo) or use oligo below.	<input checked="" type="checkbox"/> Pick right primer or use right primer below (5'->3' on opposite strand).	
CCGGAGGATTTTCGTTACTG			

Order only the new reverse primer.

RNA Extraction and Reverse Transcription from Maize tissue

Extracting RNA is similar to extracting DNA except that RNA is very unstable. Also, we are not concerned with RNA contamination in a DNA sample, but DNA contamination in an RNA sample is bad. RNA is unstable at temperatures above room temperature due to 2' OH group on the ribose. The 2' OH will attack the 3' Phosphate backbone and result in breaking the RNA backbone. RNA is also unstable due to the abundance to RNase enzymes that are everywhere. For RNA work we use RNase Free tips, tubes, and reagents. Even autoclaved, double distilled, filtered water is not considered RNase Free.

The kits we use contain RNase blockers. For RNA extraction the reagents contain guanidine salt, which denatures proteins. Guanidine salts are very irritating so do not get these in your eyes. The reverse transcription kit contains a protein called RNasin that a very effective inhibitor of RNases.

RNase A and RNase H are the two RNases that you need to know about. RNaseA is very abundant, highly stable in all conditions (included surviving autoclaving). It is an endonuclease that cleaves between C and U residues. The second RNase is RNaseH. This enzyme is actually a subunit of reverse transcriptase and degrades the RNA partner of a RNA-DNA hybrid. You will purposefully add RNaseH at the end of the reverse transcriptase reaction.

You need to be very careful and work quickly when working with RNA. You will wear gloves to protect your samples from RNases on your hands. You will use tips, reagents, and water that are RNase free. RNA samples must be kept on ice and are stored at -80°C.

RNA Extraction - Prep two different maize samples.

1. Get a piece of maize leaf about 3 cm (1 inch) in length.
2. Place the leaf in a mortar and grind with Liquid Nitrogen. Before the nitrogen completely evaporates pour into a 1.5 ml tube. Add more Nitrogen if necessary. Do not close the cap. Not all of the tissue will transfer.

3. Add 450 μ l RLT. Vortex. (Repeat 1 and 2 for next sample. Continue when all samples are at this stage.)
4. Pipette lysate onto a QiAshredder spin column (lilac colored column). Spin 2 minutes at full speed. This will be very viscous.
5. Transfer flow through to a new 1.5 μ l tube. Do not disturb the pellet.
6. Add 225 μ l Ethanol.
7. Pipette sample to RNeasy column (pink). Spin 15 secs at 8,000 rpm.
8. Empty the collection tube. Add 700 μ l RW1 to column. Spin 15 sec. at 8,000 rpm.
9. Place column in new collection tube. Add 500 μ l RPE to column. Spin 15 sec. at 8,000 rpm.
10. Empty collection tube. Add 500 μ l RPE to column. Spin **2 min** at 8,000 rpm.
11. Place Column into a new 1.5 ml tube. Spin **1 min** at **full speed**.
12. Place Column into the RNase free 1.5 ml tube. Place 30 μ l RNase free water on column. Spin 1 min at 8,000 rpm. Keep the RNA on ice from now on.

RNA is stored long term at -80°C .

Checking the RNA quality.

1. Place 5 μ l of RNA into a new tube. Add 10 μ l of water and 2 μ l of loading dye. Load a 1.5% gel. Include a DNA ladder.