

PBIO/BIOL3240L: The Dynamic Genome
Fall 2007



Table of Contents

Syllabus

Grading Policy

Chapter 1: Essential Background and Experiment 1: mPing	1-64
A. Background	1-16
B. Blast Off: Bioinformatics	17-32
B. Exp. 1 Background	33-48
C. Exp. 1 Protocol	49-64
Chapter 2: Experiment 2: The Land of the Ancient Mariner	65-89
A. Background Mariners and Osmar	65-67
B. Background to Experiment	68-71
C. Experimental Protocol	72-84
D. Analysis of Data	85-89
Chapter 3: Experiment 3: Mining a Genome for Mariners	90-137
A. Mariners in Rice	90-109
B. Mariners in Maize	110-137
Chapter 4: DNA TEs to Retrotransposons	138-145
Chapter 5: Bioinformatics—Exp 5: Finding LTR Retros	146-161
Chapter 6: Bioinformatics—Exp 6: Finding Pack-MULEs	162-175
Chapter 7: Experiment 7: Using Degenerate PCR to find TEs	176-182

PBIO3240L: Transposable Elements of Style

- [Homepage](#)
- [Instructors](#)
- [Speakers](#)
- [Course Policies and Grading](#)

7

Fall 2007
Tu/Th 12:30-03:15 PM
Plant Sciences Building Room 1503A

Week 1: Introduction	
08/16	<p>Introduction to the (1) HHMI Facility, (2) Instructors, (3) HHMI Professors program and this course, (4) class website, and (5) grading policy and assignments.</p> <p>Student introductions.</p> <p>Log on to the computers and change your passwords. Set up your personal folder.</p> <p>Wessler - Scientific overview: (pdf of figures used in class)</p> <p>Williams - The narrative, fact and fiction; science writing and connections between science writing and literature.</p> <p>Open discussion of any issues.</p> <p>Assignment -</p> <p>Wessler - read the "Wet Bench Intro" pdf (Wet Bench Intro pdf, already in your class notebook)</p> <p>Williams - Begin to read <i>The Gold Bug Variations</i> (abbreviated <i>TGBVs</i>)</p>
Week 2	
08/21	<p>Wessler - Introduction to information transfer - DNA to RNA to protein (pdf of figures - handout provided)</p> <p>Wessler - Introduction to transposable elements and experiment #1 (handout provided)</p> <p>Yujun - Introduction to bioinformatics (pdf, already in your class notebook)</p> <p>Assignment -</p>

	<p>Wessler - Read and process all handouts given to you in class today.</p> <p>Williams - Continue reading <i>TGBV</i>. Check his blog for more</p>
08/23	<p>Han - Introduction to the Wet Bench Lab and lab techniques</p> <p>Wessler - Continued introduction to Expt 1</p> <p>Williams - Introduction to science writing</p> <p>Group discussion about lecture material</p> <p>Assignment -</p> <p>Williams - continue reading <i>TGBV</i>.</p>
Week 3	
08/28	<p>Today we begin Expt. 1. Yujun will work with separate groups in the lab while Professors Wessler and Williams continue with discussions of TEs and science writing. (Download pdf for expt 1, part 2, read before class)</p> <p>Assignments -</p> <p>Continue reading <i>TGBV</i> - you must be ready for an in-depth discussion of the first 100 pages or so by 9/4</p>
08/30	<p>Day 2 of Expt 1. At the end of today's labwork, you will send your DNA samples to the UGA DNA sequencing facility. In addition, in the lecture part you will learn about gene annotation - where genes start and stop. (Download pdf for expt 1, day 2) (Queries.doc)</p>
Week 4	
09/04	<p>Professor Williams will lead the first discussion of <i>The Gold Bug Variations</i> today, and this will take up the entire class period. Be prepared to talk cogently about plot, character, and the scientific importance of the story. As always, keep up with Professor Williams's blog for other assignments/class discussion topics.</p>
09/06	<p>Today you will analyze the sequences that come back from Expt 1 by learning how to analyze the DNA sequences you generated in lab. Wessler and Han will work with you to analyze your sequences. (handout - provided in class already)</p> <p>Also - here are the "rules" for your first lab report which will be due on September 13.</p> <p>Assignment:</p> <p>Williams - Read pages 128 to 266 of <i>TGBV</i> if you haven't gotten that far already.</p>

Week 5	
09/11	<p>Today you will begin experiment 2. Please read this handout before class and write questions to be discussed when your group is not in the laboratory (Experiment 2 background and day 1 protocol)</p> <p>NOTE: Lab report #1 is due before class on 9/13</p>
09/13	<p>Today you will continue experiment 2. Please read the handout before class (Expt 2 day 2, 3 - provided in class on 9/11)</p>
Week 6	
09/18	<p>Today you will finish the wet lab part of expt 2. Yujun will send your samples to the sequencing facility. (Expt 2 day 3 - revised p58-64)</p>
09/20	<p>Today we will receive and analyze your samples from the sequencing facility.</p>
Week 9 (Weeks 7,8 were skipped)	
10/9	<p>Class Cancelled</p>
10/11	<p>Wessler: Today we will begin Experiment 2. Here is the handout (pages 64-84)</p>
Week 10	
10/16	<p>Wessler: Experiment 3 - Work in teams to analyze your TE superfamily in the rice genome, focus on phylogenetic trees. (pages 85-90). (all rooted and unrooted trees pages 91-106)</p>
10/18	<p>Wessler: Experiment 3 continues. Work in teams to analyze your TE superfamily in the emerging maize genome sequence. Then, combine your maize and rice TEs and generate one tree. (combined rice and maize trees download here) (Tree review download here - how they are made, what they mean)</p>
Week 11	
10/23	<p>Audio conference with Dr. Sean Carroll (Tree-thinking article and quiz - download and look over)</p>
10/25	<p>Fall Break, No class (yeah!)</p>
Week 12	
10/30	<p>Introduction to Retrotransposons - begin experiment 4 (handout p 107-128).</p>
11/1	<p>Mining complete LTR retrotransposons from the genome - Expt 4, day 2 (NEW!! updated handout p129-136)</p>
Week 13	
11/6	<p>(1) The science of the Tangled Field - kernel genetics; (2) Expt 4, Day 3 - comparing your LTR retrotransposons, (3) Requirements for Lab Report #3 (LAB REPORT #3)</p>

	<u>ASSIGNMENT</u>
11/8	The science of McClintock (Figures from class) (McClintock's Nobel Lecture)
Week 14	
11/13	Nathaniel Comfort visit. Discussion of The Tangled Field and science writing.
11/15	Pack-MULEs experiment and audio conference with Dr. Ning Jiang (handout from class p 137-149). TE and genome evolution class presentation (PDF of PPT). Lab Report #3 due in class
Week 15	
11/20	
11/22	Thanksgiving - No Class (yeah!)
Week 16	
11/27	
11/29	
Week 17	
12/4	Friday Schedule - No Class (yeah!)
12/6	Last Day of Class - prepare for the symposium
12/8	Symposium: The TEs of the Maize Genome - with our featured speaker - Dr. Jeff Bennetzen (see his website) Combined 3240 lab manual (word.doc)

- [Homepage](#)
- [Instructors](#)
- [Speakers](#)
- [Course Policies and Grading](#)

PBIO3240L: Transposable Elements of Style

- [Homepage](#)
- [Syllabus](#)
- [Instructors](#)
- [Speakers](#)

Course Rules and Grading Policies

General rules

Because this is a lab class, and because we will be dividing you into teams for some experiments, class attendance is crucial. If you miss class, you let your team down. So absences should be only for illness or legitimately unavoidable reasons. Class participation is crucial. Especially in the writing part of this class, your grade will be affected by it. Someone who does exemplary work and doesn't participate in class discussions will have a hard time making the grade he or she wants. So come to class prepared to discuss your assignments.

Eager, engaged work in both the wet lab and bioinformatics lab is crucial. Teamwork isn't optional; it's mandatory.

You are responsible for checking our class web site and Professor Williams's class blog, which is linked to the web site. Important announcements may appear there, so get in the habit of checking. Professor Williams's blog is located at <http://transposable.blogspot.com>.

Course Grading Policy

Above all, we want you to use your minds and your imaginations in this class. We will be telling you all the time that connections exist between everything in your world if you look closely enough. While the idea of this course is combine science and art, we will be grading these parts separately, but each part reflects on the other.

Because this is a science lab class, your work on the science component will account for 60 percent of your grade. Obviously, the writing component will count for 40 percent. If you have questions during the semester about where you stand, feel free to contact either Wessler or Williams.

Science Grading Criteria (overall 60 points of 100)

Class participation: 20 points

Pop quizzes: 15 points (5 quizzes)

Lab reports: Experiments 1, 2, and 3 -- 5 points each (15 points total)

Experiment 4: 10 points (including presentation)

Writing Grading Policy (overall 40 points of 100)

Class participation: 20 points

First paper: 10 points

Second paper: 10 points

Specific Assignments

The science part

-You are responsible for all lecture material. Knowledge of this material will be assessed by participation in discussions and by in-class quizzes.

-Participation in class discussions is mandatory and is a major part of your class grade.

-You will be required to prepare 4 laboratory reports. The reports will be unlike others that you may have prepared for previous science classes as emphasis will be on discovery, observation, and in-depth analysis. The point of this exercise is to give you a better idea of how scientists think about their data.

The due dates and a brief description of content follow. You will be given more specific instructions at a later date.

Laboratory Report 1 - Due September 11.

Your first report about Experiment 1 will be done by yourself (not as part of a group as reports 3 and 4, see below). Experiment 1 will be completed on September 6.

Laboratory Report 2 - Due October 4.

Your second report about Experiment 2 will also be done by yourself. Experiment 2 will be completed on September 25.

Laboratory Report 3 - Due November XX (to be decided)

Your third report about Experiment 3 will be done as part of a research team.

Laboratory Report 4 - Due December XX (to be decided)

Your fourth report will be about Experiment 4 but will contain a comparative analysis of the data from Experiments 3 and 4 and will also be done as part of a research team. In addition, this report will form the basis of oral presentations which will be modeled after actual presentations at scientific meetings.

The writing part

In the writing part of class, you will be required to write two papers, one due at mid-term and the other on the last day of class. You will have no tests over writing assignments and discussion, but there *will* be in-class writing assignments occasionally that will be read but not graded. Professor Williams will take notes all semester on whether you are participating in class discussion, so don't be silent!

You have some choices in your long papers. Since we will be stressing science writing the first part of the semester and then edging into fiction and poetry toward the end, the assignments will reflect that.

Assignment One, due on Tuesday, October 9:

A paper of at least 10 pages announcing the discovery of the cure for a disease you invent, complete with descriptions of people involved and the science behind

the discovery and how transposable elements were crucial to that discovery. OR: A review of the science we have undertaken the first half the semester written in the third person, present tense.

Assignment Two, due on Tuesday, December 4:

A paper of at least 12 pages on one of the following subjects:

Write a first-person short story in which you are one of the following:

- A. The aged Barbara McClintock looking back on her life and her discovery of transposable elements.
- B. Rosalind Franklin near the end of her life thinking about her contributions to the discovery of the double helix.
- C. Yourself at age 50 looking back on how this class changed the direction of your life.
- D. Yourself as a transposable element, telling your autobiography, with accurate science and an invented narrative.
- E. The head of a secret research team after the U.S. becomes a totalitarian state. Scientists are forbidden to study genes or report anything on them. You write a secret memo to other scientists, explaining, with passion and facts, why genes and especially transposable elements must be studied.

If you have another idea for your final paper that fits the overall content of the class, it's possible that it could be approved by Professor Williams. Please discuss with him outside class.

An important note: Please don't worry that you can't do these assignments. You can and you will! And all Williams expects is your best effort. Fair enough?

Please note these important rules for your papers:

- A. Papers not turned in before the end of class on the due dates will drop one letter grade. After that, it's one letter grade a day for being turned in late.
- B. Papers must be double spaced and in 12-point Times Roman font, stapled.
- C. Papers may **not** be submitted by e-mail or on flash drives.
- D. Errors in spelling and grammar will hurt your grade. Revise and edit your work carefully!

- [Homepage](#)
- [Syllabus](#)
- [Instructors](#)
- [Speakers](#)

PBIO3240L
Wet-Bench Lab Manual



1. Essential Background information

1.1 The world of laboratory science

Remember the first time you sat at a laboratory bench in Middle School or High School? Maybe you stared at a beaker, absolutely sure you would break it, or perhaps your favorite part was sparking a Bunsen burner to life, ready to cook compounds and see what changes took place.

Our class this semester is fortunate enough to have its own lab, and we will be learning its mysteries and using it just as real scientists do—to bring the microscopic world into sharp focus. We will be discovering actual transposable elements, using a technique unavailable to scientists until the early 1980s. In other words, this isn't Chemistry Set science; it is real-world research, and these techniques will become as familiar and easy to you as driving a car.

Because we thrive on turning our daily activities into coherent narratives, science, from its beginnings, was filled with people who believed in Newton and people who believed in witches. It brought us "humors in the blood" and an "Earth-centered universe." And yet for the past two millennia we have created better and better tools and ideas to help us understand the world as it truly is. Science gave up on witches a long time ago. We know the blood doesn't have "humors," and no one believes our Solar System revolves around the Earth.

We still look at the stars with wonder and joy, but when it comes to the world at its smallest level, our progress was relatively slow until a man named Kary Mullis, in 1983, invented a technique called polymerase chain reaction or PCR. We will get to

a fuller description of PCR in a bit, but for right now, think of it as a kind of genetic photocopier, churning out millions of copies of genetic material for study. In molecular biology, everything is pre-PCR and post-PCR.

The study of transposable elements (TEs), of course, pre-dates PCR, and brilliant women and men used what techniques they had to infer the then-strange idea that some genes actually jump around and even comprise a startlingly large percentage of numerous genomes, including humans'.

The combination of new lab tools such as PCR and improved computers with rapidly increasing power allowed labs to see into the heart of transposable elements. Dr. Wessler's lab, for instance, has discovered numerous TEs. They published an important paper in the journal *Nature* in 2003 reporting the first active DNA TEs in rice. This was of particular interest because rice is the most important source of human calories.

Years ago, as a molecular genetics postdoctoral fellow at the Carnegie Institution in Baltimore, Wessler cloned transposons called *Ac* and *Ds* from maize and subsequently went on to study the effect of *Ac* and *Ds* elements inserting throughout the maize genome.

Since then, Wessler and her research group have played a key role in understanding the role of transposable elements (abbreviated TEs), which are incredible genetic entities that help in shaping both plant and animal (us!) genomes and creating the modified gene functions required for evolution.

1.2. 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 (like you and the person sitting next to you I class!). You might be surprised to know that corn is a *grass*, specifically a *cereal grass*. It 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 an almost ideal model genetic organism.

At the time it was known that maize had 10 chromosomes (this is the haploid number - maize, like us, is a diploid and maize cells actually have 2 sets of 10 chromosomes). The first thing of note that McClintock did as a scientist was to distinguish each of the 10 maize chromosomes under the microscope. This was the first time anyone was able to demonstrate that the chromosomes (of any organism) were distinct and recognizable as individuals. But as she studied the chromosomes from other maize strains, she found something peculiar. In one particular strain she discovered that chromosome 9 broke frequently and at one specific place or *locus* (Figure 1). 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*). Don't worry, this will be explained in class and is summarized in the textbook you have been given - Introduction to Genetic Analysis 8th Edition (abbreviated IGA8) - see pages 423-429. Going against the science of the day, McClintock theorized that *Ac* and *Ds* were actually mobile genetic elements (that is, pieces of DNA that could move from one

site in a chromosome - called a locus - to another chromosomal location. For example, *Ac* could move from chromosome 1 to chromosome 3.)

Science got a break, though. It turned out that rare kernels with different phenotypes could be derived from the original strain with frequent breaks at chromosome 9. One such phenotype was a rare colorless kernel containing pigmented spots. (**Figure 1**) McClintock, in a leap of ingenuity, realized that these spotted kernels could be a visual assay for the activity of these transposable elements (TEs)(also called mobile DNA, jumping genes and even junk DNA). Don't worry, we will go through this in class....

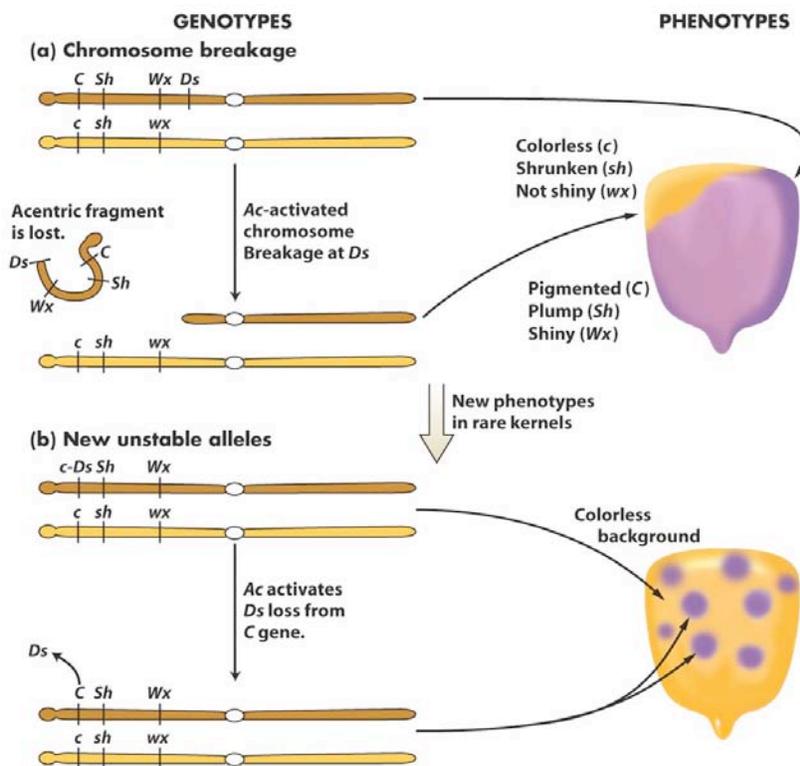
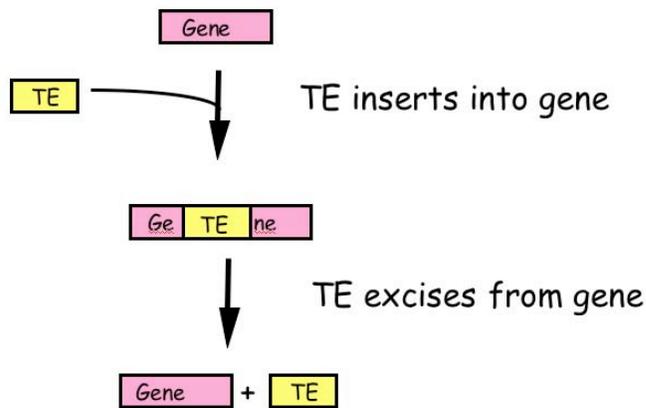


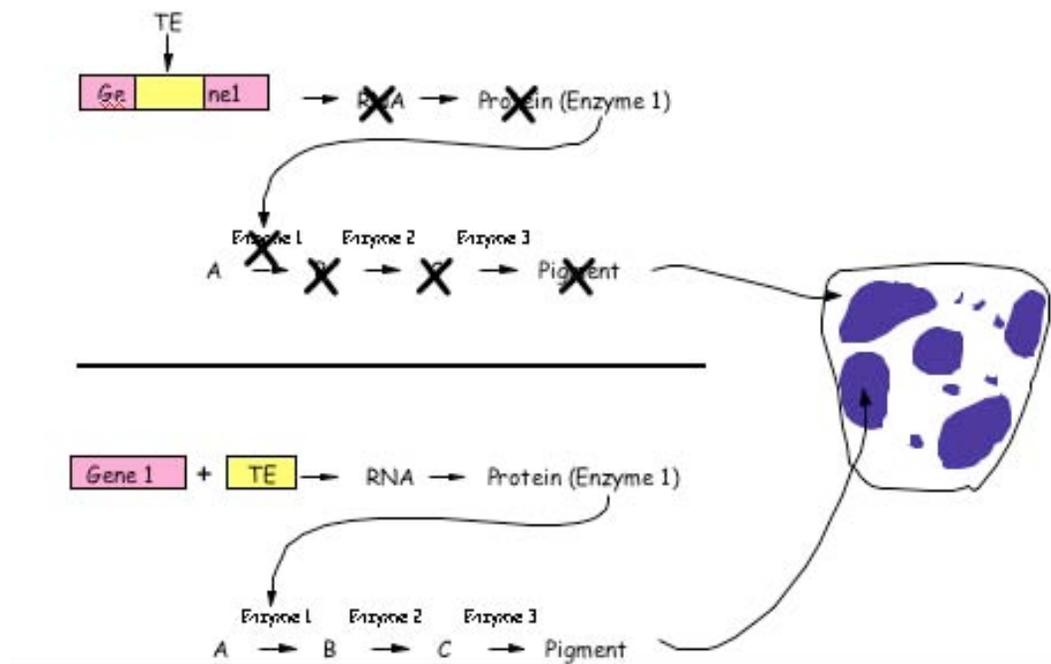
Figure 1 (13-3, p 426) New phenotypes in corn are produced through the movement of the *Ds* transposable element on chromosome 9. (a) A chromosome fragment is lost through breakage at the *Ds* locus. Recessive alleles on the homologous chromosome are expressed, producing the colorless sector in the kernel. (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. Her logic is summarized in the figure below....*



Furthermore, she provided the following explanation for what was going on with the spotted kernels:

Her discovery - highly simplified...



Finally, McClintock addressed the question of where these transposable elements were coming from. Meaning that TEs could be coming from a virus, they could be coming from the genome itself or they could be extraterrestrial!!! They could be coming from outerspace. Other experiments she conducted indicated that they were in fact "normal residents of our genomes". We will discuss how she figured this out in class. It is now known that TEs are in the genomes of all organisms including human. It is for this reason that McClintock was awarded the Nobel Prize in Medicine or Physiology in 1983. We will return to Barbara McClintock often during this course - especially when Nathaniel Comfort (the author of the McClintock biography *The Tangled Field*) visits our class.



Mutations - like those caused by TEs or by mutagens (like radiation, carcinogens etc) are normally eliminated from most populations because they decrease the fitness of an organism. (And why, thank goodness, there have been very few two-headed rattlesnakes.) The reason the spotted kernels were not eliminated is that corn is a crop plant and is subjected to artificial selection, not natural selection. And Native Americans thousands of years ago decided to grow and propagate these spotted kernels because they were unusual and they were considered sacred. Some even propagated maize strains with spotted kernels for use in religious ceremonies.

1.3. Viable Mutants, Morgan, Drosophila, and the Dawn of Genetics (see IGA8, Page 718 for more on Drosophila)

There is another reason why these mutants have hung around, and it is a concept that is essential for you to learn. They look different from regular corn, but they do not harm the plant and are thus called viable mutants. Such mutants have been very important in the history of genetics.

The American geneticist Thomas Hunt Morgan (1866-1945), for example, began early to use the common red-eyed fruit fly, *Drosophila melanogaster*, as a model organism. He and his students, starting in 1908, looked for a long time for a mutant and found a *white-eyed* fly. They then studied the segregation of this trait and, as luck would have it, found that it was sex linked. This led them to determine that particular genes were on particular chromosomes. When they saw how rare natural mutations are, they decided to use radiation and other methods to induce mutations. The issue, then? They needed lots of mutants and couldn't wait around to find naturally occurring ones.

For two long years, he found nothing useful, and in those early days of microscopy, Morgan was in danger of going blind just trying *see* the fruit flies, much less in identifying mutations. Finally a series of mutants appeared, and in 1910 Morgan noticed a white-eyed mutant male among the red-eyed wild types. When white-eyed flies were bred with a red-eyed female, their progeny were all red-eyed, while a second generation cross produced white-eyed males—a sex-linked recessive trait. Morgan called the gene that controlled this action *white*. Thus, a mutant helped unlock the clues to heritability in *Drosophila*, which are still used all over the world as a model organism, including just up the hill at the Life Sciences Building on the UGA campus.

Even before Morgan, though, the Austrian monk Gregor Mendel (1822-1884), known as the Father of Genetics, was using mutants to study heritability. Between 1856 and 1863, he cultivated and tested some 28,000 pea plants. His work, largely ignored in his own lifetime led to two generalizations, which later became known as *Mendel's Laws of Heredity* or *Mendelian inheritance*. He used the pea, a popular garden plant, because he was able to buy the seed. He then tested almost 30 mutant traits and found that seven (yellow, wrinkled, dwarf, and so forth) behaved nicely—segregating in the so-called Mendelian fashion.

What Morgan found with his white-eyed flies and Mendel discovered with his yellow, wrinkled peas was a visual assay—the same thing McClintock discovered with her maize kernels. And this semester, you will be using the same techniques—an ingenious system so that you can look at a plant and actually *see* the movement of TEs.

1.4. Transposable Elements: from genomic oddities to the single largest component of most genomes.....

What brings us to a crucial question that we must address before we go further: Where are most of the transposable elements in a genome? The answer is the first step toward complexity that we must solve. It turns out that the vast majority of all TEs are tucked away in hard-to-access places in the genome.

In the past decade, scientists have discovered more and more about the genetic content of plants and animals. Sean Carroll, an eloquent scientist who we will encounter later in the semester, explains it this way: “In just twenty years, the amount of DNA sequences in our databases has grown 40,000-fold, with the vast majority of that coming in this new century. To put that in perspective, in 1982 our total knowledge of DNA sequence from all living species amounted to fewer than one million characters. [This amount, Carroll explains, would fill up the pages of a normal sized book.] If all the text that we now have was printed into volumes and stacked, they would reach more than double the height of the 110-story Sears Tower in Chicago. This library of life is growing by more than 30 stories a year.” Actually, he wrote that a couple of years ago and now says that height would be about 8 to 10 times the *height of the Empire State Building*.

Along with all this knowledge, we have learned more and more about transposable elements. It turns out, for example, that the vast majority of TEs are between genes and in the noncoding regions of genes. The reason for this apparent preference is discussed in a later section.

While maize was a good place to start, rice, with its much smaller genome and the availability now of its entire genome, was a better direction for this work. (As of this semester, work on the determination of the maize genome sequence continues but is yet incomplete.)

Rice (*Oryza sativa*) has the smallest genome of all cereal grasses at 430 million base pairs. (By contrast, the maize genome is almost six times larger at 2,500 million base pairs.) The cereal grasses are the most significant source of calories for humans and include barley, maize, sorghum, oats, and wheat. About 40 percent of the rice genome comprises repetitive DNA that does not code for proteins. A significant fraction of this repetitive sequence appears to be TEs. The function of this so-called "junk DNA" has been a mystery, but the discovery of active transposons in rice provides insight into how TEs promote genome change.

Although genomes are comprised largely of transposable elements, virtually all of them are either inactive due to mutation (more on this in Experiment 3 where you will actually detect mutations in transposons) or turned off by the host (we will revisit this later in our narrative). Because the full genome sequence (including transposable elements) for rice is known, Wessler's group was able to use a computational approach to first identify potentially active transposable elements and then to test them for activity.

1.5. Active vs. Inactive TEs: From Genetics to Genomics

Inactive TEs cannot move from one site to another, and, as such, cannot increase their copy number in the genome. Over time, inactive elements accumulate mutations (point mutations and deletions) and, over millions of years, they literally disappear. In Experiment 3 you will have the opportunity to retrieve elements from the genome and you will be able to witness first-hand how TEs mutate.

In contrast, active DNA transposons move new copies of DNA to different places in the genome. As such, the copy number of active TEs can increase in a genome. This provided modern day biologists who study genomic sequence a clever way to identify active TEs in a genome. 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 an actively moving TE 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 repeat family with 50 copies, of which most were almost identical. Using this approach, the researchers found a repeated sequence of 430 base pairs that was identified as a *candidate* for an active DNA transposon. 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 (called "autonomous") in the rice genome that encodes the enzyme transposase necessary to enable itself and other related elements to move. To find this autonomous element, the researchers searched the rice genomic sequence for longer related elements.

It turned out that there are *different kinds of transposable elements*, and they can be grouped based on their mechanism of transposition. Class I mobile genetic elements, or retrotransposons, move in the genome by being transcribed to RNA and then back to DNA by the enzyme reverse transcriptase, while Class II mobile genetic elements move directly from one position to another within the genome using a transposase to "cut and paste" them within the genome. Our first two experiments in this class involve only Class II elements—we will return to Class I later.

Things were beginning to make sense. *Ac*, it was determined, is autonomous while *Ds* was non-autonomous. So classifying *mPing* began to get easier: it is nonautonomous!

The new findings showed that researchers can use a computational approach to identify active candidates from any genome when abundant DNA sequences are available.

Whoa! Hang on, Wessler and Williams. You're going too fast!

"No, no, no, no, no, no, you guys!" says Wessler, waving her arms. "This stuff is incredible, and you'll see it when we start doing it."

She's right. But you already know a lot about our first class laboratory experiment. We will be using a technique called Polymerase Chain Reaction or PCR (see next section) to analyze the excision of *mPing* from the leaves of a plant called *Arabidopsis*. We'll be using pipettes, centrifuges, flasks, prepared chemical compounds, something called gel electrophoresis, and we will even get to wear lab gloves and lab coats, which will not only make us look cool but keep us safe.

First, though, you need to learn a new language.

1:6. Learning to speak Laboratorian

Let's call the language we use at the wet-bench "Laboratorian." Just as we had to learn new "alphabets" and "words" for nucleotide and amino acid sequences, we have to learn the vocabulary of the laboratory. When you are fluent, you can practice speaking it with the others on your team. Later, you may want to try it on a friend or roommate just to impress them. (Well, it won't impress your roommate if she or he is majoring in biochemistry, but if the major is classics, you're safe.)

Almost everyone on South Campus speaks some version of Laboratorian. While there are dialects of the language, there are also two separate *voices*, just as English has active and passive. Laboratorian's voices are *Formal* and what we shall call, simply to be perverse, *Unformal*. Here is a use of Formal Laboratorian, taken from the aforementioned *Nature* paper from Dr. Wessler's lab:

"Structurally, MITEs are reminiscent of non-autonomous DNA (class 2) elements with their small size (<600 base pairs) and short (10-30 bp) terminal inverted repeats (TIRs). But their high copy number (up to 10,000 copies per family) and target-site preference (for TA or TAA) distinguish them from most previously described non-autonomous DNA elements. Non-autonomous elements, which make up a significant fraction of eukaryotic genomes, have been classified into families according to the transposase responsible for their mobility. But classifying MITEs in this way is problematic because no actively transposing MITE had been reported in any organism. Instead, the tens of thousands of plant MITEs have been classified into two superfamilies on the basis of the similarity of their TIRs and their target site duplication (TSD): *Tourist*-like MITEs and *Stowaway*-like MITEs. Much evidence links *Tourist* and *Stowaway* MITEs with two superfamilies of transposases, *PIF/Harbinger* and *Tc1/mariner*, respectively."

To scientists who can speak and write in Formal Laboratorian, that paragraph is no harder than reading the first sentence of *Moby-Dick*: "Call me Ishmael." But rest easy. You *don't* have to know any Formal Laboratorian for this course, though by December reading it will be far easier than it is right now. You probably won't be able to fake your way through an international science conference, but you can at least comprehend it, much as French suddenly looks familiar after you've learned Spanish.

You *will*, on the other hand, have to learn Unformal Laboratorian, and it's a snap. You will all be using it in no time! You will understand this sentence like a native-speaker:

"Hey, hand me the new PCR tubes so I can start to pipette my reagents—I need 10 microliters of that Extract-N-Amp PCR reaction mix first!"

1:7 The Vocabulary of Unformal Laboratorian

You're going to know a lot of these already, and many of you may have used most of them in biology labs. But don't worry if they seem, well, foreign. These definitions are from various sources, including *The New Penguin Dictionary of Science*, online science dictionaries, and Wikipedia. (As always, use Wikipedia with discretion and never use it as a source for a scholarly article, but it's amazingly reliable, unless you're checking on controversial current politicians and the like.)

1. **Arabidopsis**. This small flowering plant is a genus in the family *Brassicaceae*. It is related to cabbage and mustard. This genus is of great interest since it contains *Thale Cress* (*Arabidopsis thaliana*), one of the model organisms used for studying plant biology and the first plant to have its entire genome sequenced. Changes in the plant are easily observed, making it a very useful model. Note: You will become Arabidopsis farmers for this class, but *never* eat your crop.

2. **Polymerase Chain Reaction**. Known familiarly as PCR, this, as mentioned before, this is 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 including AIDS. The technique has revolutionized many areas of molecular biology—and won a Nobel Prize for Kary Mullis.

You'll be doing PCR in the lab this semester! Below is general—we get much more specific as you will see later on.

The reaction starts with a double-stranded DNA fragment with known end sequences, which is to be copied. Once we have that, here's what happens:

- A. The two DNA strands are separated by heating to 95 degrees Celsius.
- B. Two primers are added that have complementary base pairs to the end sequences in the DNA to be amplified. Cooling to about 50 degrees Celsius (50° C) allows the primer to anneal with each strand.
- C. DNA nucleotide triphosphates and the heat stable *Taq* polymerase become involved, and the temperature is raised to 72° C.
- D. New DNA strands are synthesized by *Taq* polymerase-incorporated DNA nucleotide triphosphates strands acting as templates.

E. The procedure is repeated with two new double strands.

Each cycle of heating and cooling doubles the number of copies of the DNA template, so that after 20 cycles, there are more than a million copies of the original DNA!

Our procedure won't look precisely like this, but you have the general idea. Much of it is done in a PCR machine anyway.

3. 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. Many laboratory techniques of biochemistry and molecular biology that involve DNA polymerases, such as DNA sequencing and PCR, require primers. The primers used for these techniques are usually short, chemically synthesized DNA molecules with a length about twenty nucleotides.

4. Base pair. The unit length of a double-stranded molecule. In molecular biology, two nucleotides on opposite complementary DNA or RNA strands that are connected by hydrogen bonds are called a base pair (often abbreviated bp). Adenine (A) forms a base pair with thymine (T), as does guanine (G) with cytosine (C) in DNA. In RNA, thymine is replaced by uracil (U). The size of an individual gene or an organism's entire genome is often measured in base pairs because DNA is usually double-stranded. Hence, the number of total base pairs is equal to the number of nucleotides in one of the strands (with the exception of non-coding single-stranded regions of telomeres.) The human genome is estimated to be about 3 billion base pairs long and to contain 20,000-25,000 distinct genes.

5. Microliter. "Micro" is a prefix meaning one-millionth of a base unit. So a microliter, a unit of measurement of which you will see a great deal, is one-millionth of a liter. It is designated by the symbol μ .

6. Pipette. A pipette (also spelled *pipet* or called a *pipettor*) is a laboratory instrument used to transport a measured volume of liquid. Most of you have probably seen these, even if you haven't used one. Pipettes are commonly used in chemistry and molecular biology research as well as medical tests. Pipettes come in several designs for various purposes with differing levels of accuracy and

precision, from single-piece glass pipettes to more complex adjustable or electronic pipettes. A pipette works by creating a vacuum above the liquid-holding chamber and selectively releasing this vacuum to draw up and dispense liquid. Pipettes that dispense between 1 and 1000 μ l are termed *micropipettes*, while *macropipettes* dispense a greater volume of liquid. We will be using adjustable micropipettes.

7. Centrifuge. A device for separating out particles from a suspension by rapidly spinning the suspension in a tube. Someone named Antonin Prandl invented the first centrifuge in order to separate cream from milk to make churning butter much easier! The load in a laboratory centrifuge must be carefully balanced—and we will do that. Small differences in mass of the load can result in a large force imbalance when the rotor is at high speed. This force imbalance strains the spindle and may result in damage to centrifuge or personal injury. Centrifuge rotors should never be touched while moving, because a spinning rotor can cause serious injury. Modern centrifuges generally have features that prevent accidental contact with a moving rotor. The centrifuge is your friend.

8. Gel electrophoresis. This is a laboratory technique used to separate mixtures of molecules such as proteins and nucleic acids in a suspension, by their charge-to-mass ratio. The mixture is added to an inert medium such as an agarose (which we will use) or acrylamide gel in an appropriate buffer solution and is subjected to an electric field. The charged molecules then move through the gel toward the appropriate electrode. Gel electrophoresis of fragments of DNA is routinely used to produce DNA fingerprints. The results can be analyzed quantitatively by visualizing the gel with UV light and a gel imaging device. The image is recorded with a computer-operated camera, and the migration of the band or spot of interest in the gel matrix is measured and compared against standard or markers loaded on the same gel. (Shorter molecules move faster and migrate further than longer ones.)

9. Agar. The gel we will be using in our gel electrophoresis is called “agarose gel,” meaning it is made with agar, which is a gelatinous substance chiefly used as a culture medium for microbiological work. As you probably guessed (just kidding), the word *agar* comes from the Malay word *agar-agar*, meaning *jelly*. Other uses are as a laxative, a vegetarian gelatin substitute, a thickener for soups, in jellies, ice cream and Japanese desserts such as anmitsu, and as a clarifying agent in brewing. It is an unbranched polysaccharide obtained from the cell walls of some species of red algae or seaweed. Agar polysaccharides serve as the primary structural

support for the algae's cell walls. Try dropping the term into dinner conversation with your family on Thanksgiving.

10. **Denaturation.** This involves the 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.

11. **Annealing.** This happens when DNA or RNA strands pair by hydrogen bonds to complementary strands, forming a double-stranded molecule—the double helix! (This is how we will use it, obviously.) The term is also used to describe the reformation (renaturation) of complementary strands that were separated by heat (thermally denatured).

12. **Extension.** This is simply enzymatically extending the primer sequence—copying DNA.

13. **Buffer solution.** A solution that resists changes in pH. Buffer solutions are used in experiments where a nearly constant pH is required, such as those involving an enzyme reaction such as DNA synthesis.

14. **Growth chamber.** A chamber for growing things. What did you think it would be?

15. ***mPing*.** A miniature, inverted-repeat transposable element from rice, *mPing* can transpose in rice or even in other plants like *Arabidopsis*.

There will be many more terms of Unformal Laboratorian that we will use, but these are the main ones that will allow you to speak the language with your classmates. On second thought, don't try any of this on Grandma at Thanksgiving.

Blast Off: Transposable Elements and Bioinformatics: Computer Protocols for Wessler-Williams Class

1:1 The world of bioinformatics

Welcome to the intriguing and amazing world of bioinformatics! In our class this semester, we will combine classical wet-bench science, creative writing, and the cutting edge of computer-assisted analysis of protein and nucleotide sequences. We will *what?* Yep, all this fits, and you will be learning how.

This guide will take you, step by step, through a relatively new way of examining how the biological world is put together at the cell level. Believe it or not, you will be discovering things, using computers, never before known and adding to the world's knowledge of transposable elements (TEs).

Learning to read and understand the world at its tiniest level has been an exciting challenge that has engaged scientists for many decades. It took a brilliant woman named Barbara McClintock to realize that our previous idea of genes as being somewhat like pearls on a string of chromosome was wrong. Instead of stable, unchanging biological units, some genes are, in fact, restless explorers, often finding a useful niche but more often inconsequential. McClintock's early work intrigued other scientists who lacked the tools, to be honest, to know exactly what was going on.

That changed as computers became stronger and stronger, part of our daily lives and an essential part of science. As researchers began to see how computer power could answer difficult questions, databases began to proliferate around the world. More and more, scientists could send their data to powerful computers with their resident programs and databases, and in a matter of seconds, astonishing information would return, sometimes confirming theories and other times exploding them.

But what is bioinformatics, really? Let's use information from a standard book on the subject: *Bioinformatics: Sequence and Genome Analysis*, by Dr. David W. Mount. We will paraphrase Dr. M first: *Bioinformatics is a way to organize biological data related to genomes, often to find ways of using this information in agriculture, pharmacology, and other commercial applications. Of course, we use it*

for much more—medical advances, crime detection, and pure research—answering basic questions about life itself in plants and animals.

You might also think of bioinformatics as code-breaking in a treasure hunt. In bioinformatics, whether we are looking at proteins or nucleotides, what we usually see is a series of letters strung together in seemingly endless lines. To find an analogy in English literature, let's take part of a well-known opening paragraph of a famous novel. This is how it would look if you saw it as we see sequences of letters that must be read in bioinformatics:

ITWASTHEBESTOFTIMESITWASTHEWORSTOFTIMESITWASTHEAGEOFWISDOMITWASTHEAGEOFFOOLISHNESSITWASTHEEPOCHOFBELIEFITWASTHEEPOCHOFINCRECULITYITWASTHESEASONOFFLIGHTITWASTHESEASONOFDARKNESSITWASTHESPRINGOFHOPEITWASTHEWINTEROFDESPAIR . . .

This of course, is the opening of *A Tale of Two Cities* by Charles Dickens, in all capital letters and with the punctuation removed. In a sense, it *is* a code, because we have to find a way to *arrange* it to understand it. In this case, it's a ridiculously easy cipher for most of us, isn't it? But what if it looked like this:

JUXBTUIFCFTUPGUJNFTIUXBTUIEXPSTUOGUJNFTJUXBTUIFBHFPGXJTEPN
JUXBTUIFBHFPGGPPMJTIOFTTJUXBTUIFEQPDIPGCFMJFGJUXBTUIFFQPDIP
GJODSFVEMJQZJUXBTUIFTFBTPOPGMJHIQJQXBTUIFTFBTPOPGEBSLMFTT
JUXBTUIFTQJSJMHPGIPQFJUXBSUIFXJMUFSPGEFTQBJS . . .

Could you read that? Or even better, what if it looked like this:

JUX BTU IFC FTU PGU JNF TIU XBT UIE XPS TUO GUJ NFT JUX BTU IFB HFP
GXJ TEP NJU XBT UIF BHF PGG PPM JTI OFT TJU XBT UIF EQP DIP GCF MJF
GJU XBT UIF FQP DIP GJO DSF EVM JQZ JUX BTU IFT FBT POP GMJ HIQ
JQX BTU IFT FBT POP GEB SLM FTT JUX BTU IFT QSJ MHP GIP QFJ UXB
SUI FXJ MUF SPG EFT QBJS . . .

Actually, it's the very easiest of substitution ciphers imaginable. You simply shift one letter in the alphabet for each letter above. I become J, T becomes U and so forth. In the second example, we just group them into threes to confuse things. A snap, right?

This is what we have to do in bioinformatics—learn the alphabet and then learn how to read it. That may sound daunting, but trust us, it's not as hard as it seems. For one thing, the alphabet for nucleotides makes the English alphabet look positively huge. The nucleotide alphabet has only four letters, each one representing one of the four bases that, in pairs, are the central part of DNA. You know these, of course: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T).

DNA, you will remember, is a fairly simple chemical, though its structure baffled scientists until 1953. This is a good place to quote from the textbook *Introduction to Genetic Analysis*. We know this is a good source because Dr. Wessler is a co-author! Here's how the book describes those crucial four bases to us:

"Two of the bases, adenine and guanine, have a double-ring structure characteristic of a type of chemical called a purine. The other two bases, cytosine and thymine, have a single-ring structure of a type called a pyrimidine. The chemical components of DNA are arranged into groups called nucleotides, each composed of a phosphate group, a deoxyribose sugar molecule, and any one of the four bases."

In double-stranded DNA, A pairs with T, and G pairs with C.

This is getting ahead of our story, though. While we can use bioinformatics to unravel the "code" of nucleotide sequences, we also use it to understand the linear sequence of amino acids in a protein chain. Why? As our *New Penguin Dictionary of Science* puts it, "Protein sequence data have been used to determine evolutionary relationships between organisms and to investigate the functioning of genes."

We're not as fortunate in our "alphabet" for amino acids as we were for nucleotides. Still, we're dealing with only 20 standard amino acids, each one with its own alphabet letter. So it's *still* easier than the English alphabet. Since we will be studying transposable elements, let's look at the sequence for one called Pong. Here is a portion - about 1/5 - of its amino acid sequence:

```
GSIDCMHWIWENGPTAWKGQYCRGDH GKPTIILEAIASQDLWIWHAFFGVAGSNN
DINVLNQSDVFNQDLQ GKAPEVQFTLNGTTYNMGYLAD E IYPEWATFVK TISMPQG
EK RKLFAQH Q
```

Doesn't look much like a nucleotide sequence, does it? In fact, it looks more like our simple-substitution cipher for the opening of *A Tale of Two Cities*. But it's not all that hard to decipher, really. Each of the 20 standard amino acids has a shorthand designation with a single English letter. Here they are:

Alanine: A

Arginine: R

Asparagine: N

Aspartic acid: D

Cysteine: C

Glutamic acid: E

Glutamine: Q

Glycine: G

Histidine: H

Isoleucine: I

Leucine: L

Lysine: K

Methionine: M

Phenylalanine: F

Proline: P

Serine: S

Threonine: T

Tryptophan: W

Tyrosine: Y

Valine: V

Suddenly, the sequence for Pong starts to make sense, doesn't it? We might not know *exactly* what it means, but we can see what is there. (If you have time, check out what's known about all those amino acids, but we don't need to know much beyond what the letters stand for in our class.)

What bioinformatics does is to help us make sense of such a sequence, whether of proteins or nucleotides. It can do this in many ways, but the one with which we will be most concerned is in comparing sequences to other already-described and known sequences to see how they might be alike or different. And in doing so, we can learn a great deal about the "code" that only a few moments before looked so impenetrable to us.

Now that we know about bioinformatics, let's move on to how we use online programs and databases to understand our sequences.

1:2 The Sequence of Events for Sequences

If you think the development of protein sequencing methods is something quite new, you'd be quite wrong. Dr. Margaret Dayhoff and colleagues at the National Biomedical Research Foundation "were the first to assemble databases of these sequences into a protein sequence atlas in the 1960s," says Dr. Mount, who has decided to hang out with us and help out. Their collection was eventually known as the Protein Information Resource. This was pioneering work because Dayhoff and her collaborators "organized the proteins into families and superfamilies based on the degree of sequence similarity."

DNA-sequence databases, on the other hand, didn't come along until the early 1980s. Several research groups around the world began programs to collect and assemble data on nucleotide sequences. One, based in the United States, was (and is) called the National Center for Biotechnology Information (NCBI), and it's this resource that we will be using in our class this semester!

For a long time, before the creation of the Internet, the availability of information on these databases was, well, clunky, at best. You *could* get the information, but it was slow and difficult to obtain. With the advent of the World Wide Web and increasingly fast personal computers, however, all the information on these dazzling databases was at the touch of a computer keypad.

Even better, they weren't secret. In our treasure hunt, these clues are offered free, in fact, to whoever wants them. Dr. M, how would you describe all this?

"The idea behind these programs was to provide an easy-to-use interface with a flexible search procedure to the sequence databases using keywords searching on standardized entry fields," he tells us.

NCBI's first attempt at such a program was called *Entrez* (that's "enter" in French; not sure why it wasn't in English, but anyway!). This Center has moved well beyond Entrez, however, though you might see that term from time to time.

This is the bare bones of what you need to know about bioinformatics as a prelude to what we will do, but the field is huge and hugely complex. Dr. M's book, for instance, comprises almost 700 large, densely packed pages, and it probably weighs more than three pounds! No wonder he's decided to sit here and explain things for a while.

We will be discussing our bioinformatics work at length as we move along, so just remember that this brief workbook is meant to provide an overview and clear directions to work you will be doing during the whole semester.

2:1 NCBI and our class bioinformatics

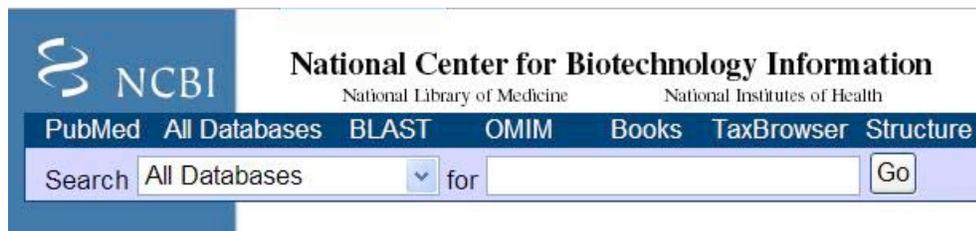
The NCBI has been around for almost 20 years and is a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH), and one of its focuses is computational molecular biology.

While we will be using the NCBI databases in our computer lab, please remember that you can access these (for free) from anywhere—your home, an Internet café, or using a Wi-Fi system such as UGA's. Bookmark the site: www.ncbi.nlm.nih.gov/ and feel free to play around and see what's available. You will be seeing it a lot this semester!

2:2 Procedures for identifying transposable element sequences

Okay. We're ready to begin.

1. Go to the NCBI web site (www.ncbi.nlm.nih.gov/) and take a look at the main page. You will see numerous links to the history of the NCBI, to the different programs available, and to much more. Hint: The site and most of the others you will use this semester are bookmarked on your lab computers.



2. Look at the top of the main page and find **BLAST**. This is where we start our identification of TE sequences. BLAST is an acronym for **Basic Local Alignment Search Tool**, and it finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases (you will learn all about these in class) and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

3. Click on **BLAST!** We are now underway with our bioinformatics analysis (or, if you will, our treasure hunt.) When you click on BLAST, you will be directed to a page that gives you a number of options. For now, let's skip the first section, called "Blast Assembled Genomes," and go to the second, "Basic BLAST." As you see, there are five BLAST programs.

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
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

4. Click on *tblastn*, the fourth choice down. As the explanation says, this lets us search translated nucleotide databases using a protein query. Don't worry too much right now about what this means, as it will become clear.

5. Yikes! What does this mean? Enter *query sequence*? You're right. There are a lot of empty blanks on this page, but it's not nearly as complicated as it looks. To begin with, let's review what a query sequence actually is. Remember, this process compares a sequence of amino acids or nucleotides against *sequences in existing genomes*.

► NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. [more...](#)

In doing this, we can find out how much our sample sequences are like already-known sequences. This can tell us about what our sequences resemble and what they may do. The sequence we enter to compare to existing genomic information is called a **query sequence**. We will be using sequences from real transposable elements, so let's start with one you've seen already, the sequence for the TE called Pong:

```
GSIDCMHWIWENGPTAWKGQYCRGDHGKPTIILEAIASQDLWIWHAFFGVAGSNN
DINVLNQSDVFNQVQLQKGAPEVQFTLNQTTYNMGYYLADEIYPEWATFVKTI SMPQG
EK RKLFAQH Q
```

(Make sure you don't copy and insert the first paragraph from *A Tale of Two Cities*—the programs will be warping binaries until the end of time!)

6. Copy, from the Word document on your computer desk top called "Queries" the sequence for Pong and then paste it into the panel at the top. This is the panel below the large heading "Enter Query Sequence" and a smaller one saying "Enter accession number, gi, or FASTA sequence." (Don't worry about what any of those mean right now.)

Enter Query Sequence

Enter accession number, gi, or FASTA sequence 🔍 Clear

```

GSIDCMHWIWENGP TAWKGYCRGDH GKPTIILEA IASQDLWIWHAFFGVAGSNNDINVLNQSDVFNDVL
QGRKPEVQFTLNGT TYNMGYYL ADEIYPEWATFVKTI SMPQGEKRRLFAQHQ

```

Query subrange 🔍

7. Ignore, for now, the rest of the panels down to the second section, which is headed "Choose Search Set."

Choose Search Set

Database 🔍

Organism Optional
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. 🔍

Entrez Query Optional
 Enter an Entrez query to limit search 🔍

8. For the first panel under "Choose Search Set," leave it on the default setting, which is "Nucleotide collection (nr/nt)"

Choose Search Set

Database 🔍

Organism Optional
 Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. 🔍

Entrez Query Optional
 Enter an Entrez query to limit search 🔍

9. The next panel, "Organism," deserves some explanation before we go on. What we select here defines the databases of genomic sequence from a particular organism that we BLAST *against*. So we are looking for sequences similar to Pong, remember? Well, we have lots of organisms from which to choose here. The ones on the main line are the usual suspects: "any," "human," "A. thaliana," "mouse," or "custom." "Any," "mouse," and "human" seem pretty clear. *A. thaliana* might not be familiar to you right now, but by the end of this course, you will be on extremely intimate terms with it! Its full name is *Arabidopsis thaliana*, and it is a tiny plant of the mustard family that is the warhorse for laboratory genetics experiments for two reasons: It has a very short life cycle, and its entire genome has been elucidated by researchers. "Custom" means "customized," and while this is usually a computer cue that only True Geeks should proceed, it's safe enough here.

10. **Click on custom.** Another panel will pop up underneath with this direction: "Enter organism common name, binomial, or tax id. Only top 20 taxa will be shown." This is an amazing panel! (And don't worry about April 15: "tax id" refers to "taxonomy.")

11. **Enter the organism to BLAST against.** 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. The first, oddly enough, is simply to type "rice," and it pops right up for you to click on. Its tax id, as you will see, is 4530. But let's be more Linnaean here and 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).

The screenshot shows the 'Choose Search Set' panel in the NCBI BLAST interface. The 'Database' is set to 'Nucleotide collection (nr/nt)'. The 'Organism' search box contains 'ory', and a dropdown menu shows a list of organisms. 'Oryza sativa (taxid:4530)' is highlighted. A 'BLAST' button is visible at the bottom left. To the right, there is a 'Show results in a new window' checkbox and a link to 'Algorithm parameters'.

12. **Ignore the final panel.** (This is the one called "Entrez query" and goes back to the early days of NCBI programs. We will pretend this is only for traditionalists or geezers and move silently on.)

13. **Go the bottom and click on BLAST!** You don't have to count down from ten, but you can if you want to. You have just entered the world of bioinformatics! Congratulations!

The screenshot shows the 'BLAST' button and search options in the NCBI BLAST interface. The 'BLAST' button is highlighted. The search options are 'Search database nr using Tblastn (search translated nucleotide database using a protein query)'. There is a checkbox for 'Show results in a new window' which is currently unchecked. A link to 'Algorithm parameters' is visible at the bottom.

14. **You'll see a "Job Panel" first.** This simply tells you that the program at NCBI is hard at work comparing your query and subject sequences to look for similarities.

In the case of Pong, it will tell you that it is working on this: Protein Sequence (122 letters). It will be on this panel while the program churns. How long it will take to give you results depends on many variables, but in most cases, it is less than a minute and in many only a few seconds. This is formidable computer power, since people all over the world are using it while you are!

• NCBI/ BLAST/ tblastn/ Formatting Results - BNA1Z2PJ012 [\[Formatting options\]](#)

Job Title: Protein sequence(122 letters)

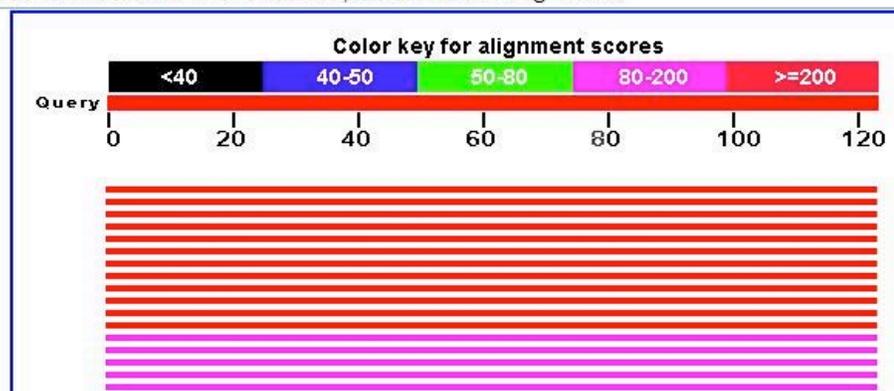
Request ID	BNA1Z2PJ012
Status	Searching
Submitted at	Thu Aug 9 01:38:07 2007
Current time	Thu Aug 9 01:38:12 2007
Time since submission	00:00:05

This page will be automatically updated in 13 seconds

15. Et voilà! You have hit the jackpot! Your search has found 212 "hits" in the *Oryza sativa* genome for the Pong sequence. But we need to let you in on a little secret. "Pong" is a transposable element in the rice genome! The contest was rigged! Still, it's a good place to start, and, after you learn to do this in your sleep (and you will), you can see exactly what's taking shape here. Not only will this page gives you a color panel of the range of your hits, but even better, it gives you sequences similar to your query sequence below that panel. These "hits" use percentages to rank what your query sequence is most like in the queried database. As you can see, Pong is everywhere!

Distribution of 212 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Pause for a Deep Breath

Let's take a break for a moment. We've accomplished a lot so far, so let's review. We have learned what sequences are, for proteins and for nucleotides, and why

they are important. We have learned the basics of bioinformatics and how scientists use this field to understand how organisms are built and what they do—from the smallest level of life up. We have also learned how to use one special part of the BLAST protocol on the NCBI web site.

And yet in our genomic treasure hunt, we aren't even close to the place where "X marks the spot." We have started on our journey, and we've done well so far, but what's a treasure hunt without more obstacles to overcome? So let's gather our "gear" and head on to uncharted territory.

16. The bad news: We can't use the information in the form we have it right now. The good news: Another program can! We must reformat it first, so we have to do two things in this regard:

First, we need to **copy the complete list of alignments**. Wait just a minute there, Wessler and Williams. What is an *alignment*? Let's ask Dr. Mount, who has his hand up, clearly dying to answer that question: "Alignment is the procedure of comparing two or more sequences by looking for a series of individual characters or character patterns that are in the same order in the sequences." Got that? Let's take the first alignment in our "Pong vs. rice" game:

Oryza sativa (japonica cultivar-group) genomic DNA, chromosome
2
Length=35954743

Score = 204 bits (520), Expect = 2e-52
Identities = 90/122 (73%), Positives = 104/122 (85%), Gaps = 0/122 (0%)
Frame = +2

Query 1 GSIDCMHWIWENGPTAWKGGYCRGDHKGKPTIILEAIASQDLWIWHAFFGVAGSNNDINVL 60
GS+DCMHW W+N P AWKGG+ RGD+G PTI+LEA+AS+DLWIWHAFFG AGSNNDINVL
Sbjct 15811853 GSVDCMHWEWQNCPVAWKGQFTRGDYGVPTIMLEAVASKDLWIWHAFFGAAGSNNDINVL 15812032

Query 61 NQSDVFNVDVLQGKAPEVQFTLNGTTYNMGYYLADEIYPEWATFVKTI SMPQG EKRLFAQ 120
+QS +F DVLQG+AP VQ+TLN + YNMGYYLAD IYPEWATF K+I PQ K KL+AQ
Sbjct 15812033 DQSPLFTDVLQGRAPPVQYTLNESDYNMGYYLADGIYPEWATFAKSIIRPQSAKHKLYAQ 15812212

Query 121 HQ 122
HQ
Sbjct 15812213 HQ 15812218

Score = 190 bits (483), Expect = 4e-48
Identities = 87/122 (71%), Positives = 103/122 (84%), Gaps = 0/122 (0%)
Frame = -3

You will see by looking at the Query sequence line (which starts on line 1 and then jumps to line 4) how similar it is to the Subject sequence line (which starts on line

3 and jumps to line 6). The middle line in each set of three lines, the one that looks like this:

```
GS+DCMHW W+N P AWKGQ+ RGD+G PTI+LEA+AS+DLWIWHAFFG AGSNNDINVL
```

... is something we don't have to worry about! It *does* have a purpose, but we don't need to know it. In other words, it's a red herring in the hunt for our genetic treasure map! We don't need it to reach the place where X marks the spot.

17. So, let's copy those alignments! Scroll down past the "Sequences producing significant alignments" to the section titled **Alignments**. Right beneath the block called **Get Selected Sequences**, you will see a caret, a box, and a hyperlink. Start copying at the caret (including it) and copy all the way down **until you see the Get Selected Sequences box again**. The last thing you copy will be a series of numbers at the end of the last "**Sbjct**" line. Now that you have this whole long block of alignments marked, **copy them**.

```
> dbj|AP008208.1| Oryza sativa (japonica cultivar-group) genomic DNA, chromosome
2
Length=35954743

Features flanking this part of subject sequence:
 5702 bp at 5' side: Os02q0466600
 4614 bp at 3' side: Os02q0467000

Score = 204 bits (520), Expect = 2e-52
Identities = 90/122 (73%), Positives = 104/122 (85%), Gaps = 0/122 (0%)
Frame = +2

Query 1          GSIDCMHWIWENGPTAWKGQYCRGDHGKPTIILEAIASQDLWIWHAFFGVAGSNNDINVL 60
                  GS+DCMHW W+N P AWKGQ+ RGD+G PTI+LEA+AS+DLWIWHAFFG AGSNNDINVL
Sbjct 15811853   GSVDCMHWIEWQNCVPAWKGQFTRGDYGVPTIMLEAVASKDLWIWHAFFGAAGSNNDINVL 15812032

Query 61         NQSDVFNVDVLOGKAPEVQFTLNGTTYNMGYLAD E IYPEWATFVK TISMPOGEKRKLF AQ 120
                  +QS +E DVLOG+AP VQ+TLN + YNMGYYLAD IYPEWATF K+I PQ K KL+AQ
Sbjct 15812033   DQSPFTDVLQGRAPPVQYTLNESDYNMGYLADGIYPEWATFAKSIIRPQSAKHKLYAQ 15812212

Query 121        HQ 122
                  HQ
Sbjct 15812213   HQ 15812218
```

18. Now we have to change to another web site! It is in your bookmarks and is called *Blast Editor*, and here it is:

<http://www.wunchiou.com/test/formatblast.html>. **Click on it.** This "scrip" was written by Wun Chiou, a chemistry teacher at the well-known Hi-Tech High-LA, a science high school in Los Angeles. The Wessler lab has welcomed students from Hi-Tech High to UGA for the past two summers, and Mr. Chiou wrote this scrip to help edit the sequence alignments so they can be further studied. As you can see, this is a simple site where your odds of screwing up are small because it only does two things!

Enter the BLAST result data to be formatted:

Query base pairs:

Normal format output
 Subject-only output

19. Paste the long series of alignments you just copied into the box at the top called "Enter the BLAST result data to be formatted." But before we go further, we must choose which format we want to use. As you can see, there are, as we said, only two, and they are:

- A. Normal format output
- B. Subject-only output

Think of these as "editing panels," because their only job is to edit your alignments for further work. For instance, in building phylogenetic trees—which we will learn—we will need an edited version of our alignments. So let's give it a whirl. (Before we go further, however, you can for now ignore the panel on this page called "Query base pairs." This is because you don't need it. Trust us.)

20. Click on the circle marked "Normal format output" and then in the area right below it, click on "Format the above text." This will pop up, in the "Output" window below, our Pong alignments without the line that, before, separated the Query and Subject lines from our original search. This line, as you recall is **Something Else You Don't Need to Know (SEYDNTK)**.

Enter the BLAST result data to be formatted:

```
>dbj|AP008208.1| Oryza sativa (japonica cultivar-group) genomic DNA, chromosome
2
Length=35954743

Features flanking this part of subject sequence:
 5702 bp at 5' side: Os02g0466600
 4614 bp at 3' side: Os02g0467000
```

21. **Et voilà, yet again!** Now you have a cleaned-up—edited—version of the Pong alignments. You should be pleased to have gotten this far along in your genetic

treasure hunt, and like any treasure hunter, you don't want to lose the map. So go to the output panel and *copy this again for safekeeping*.

Output:

```
>AP008208
Query 1          GSIDCMHWIWENGPTAWKQYCRGDH GKPTIILEAIASQDLWIWHAFFGVAGSNNNDINVL 60
Sbjct 15811853  GSVDCMHWEWQNCPPVAWKGFTRGDYGVPTIMLEAVASKDLWIWHAFFGAAGSNNNDINVL
15812032
Query 61        NQSDVFNVDVLQGGKAPEVQFTLN GTTYNMGYYLADEIYPEWATFVKTISMPQGEKRKLFAQ 120
Sbjct 15812033  DQSPLFTDVLQGRAPPVQYTLNESDYNMGYYLADGIYPEWATFAKSIIRPQSAKHKLYAQ
15812212
Query 121       HQ 122
Sbjct 15812213  HQ 15812218
```

22. Open Microsoft Word and create a new document. In this new document, paste the edited alignments you just copied. Great! Now, give it some kind of snazzy, unforgettable file name, such as "Normal Output Pong" or "Buried Treasure Map 1" and save it. This is mainly for documentary purposes in case you want to use these sequences again—and you may. After you have safely saved your alignments, you can minimize or close MS Word.

23. At this point, we're not yet through with our Blast Editor program, though. Bring it back up on your screen. It's right where you left it, of course, with the "Output" panel holding the "Normal format output" that you just saved to Word (this is the .doc file).

24. Go back to the top and click on the "Subject-only output" button. Now, click, yet again, on "Format the above text." At this point our Blast Editor does something even more amazing: It removes both our query sequences *and* the material between the query and subject lines (which is SEYDNTK, as you recall). What you get in the panel below this time is only our subjects. This is the format that we will use in building phylogenetic trees (you will do this in Expt 3, later in the course).

Output:

```
>AP008208_1
GSVDCMHWEWQNCPPVAWKGFTRGDYGVPTIMLEAVASKDLWIWHAFFGAAGSNNNDINVL
DQSPLFTDVLQGRAPPVQYTLNESDYNMGYYLADGIYPEWATFAKSIIRPQSAKHKLYAQ
HQ
>AP008208_2
GSIDCMHWRWEKPTTWRGQFTRGDYGVPTIILEAVATRDRLRIWHAFFGVAGSNNNDINVL
NQSLFLDVLKGEAPRVKFFVNGNEYNIGYYLADGIYLEWATFVKSI AAPQTENNKLYAQ
YQ
>AP008208_3
GSIDCMHWHWERCPPVAWKGFTRGDQKVPTIILEAVASHDLWIWHAFFGAAGSNNNDINVL
```

25. **Copy the subject-only text in the bottom panel for further use.** It would be a good idea to save this as a text file in MS Word also, perhaps with a name like "Subject-Only Output Pong" or "Important Clue 2"!

Pause for Another Deep Breath

2 Experiment 1: Does the rice mPing element transpose in the model plant Arabidopsis?

Dr. Guojun Yang, a postdoctoral associate in the Wessler laboratory, generated the Arabidopsis strains that you will be using. In fact, you will be repeating an experiment that appeared in a recent issue of the journal, Proceedings of the National Academy of Sciences (abbreviated PNAS).

Before you learn about the experimental protocols, it is necessary to learn several things so that you understand not just what you are going to do but why.

2.1: What transposable elements look like: *Ac* and *Ds*, *Ping* and *mPing*

(a) To the geneticist: You have already seen that Barbara McClintock discovered the transposable elements *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 things that geneticists figured out about TEs through observation of the plant phenotypes and by performing carefully designed crosses:

- that TEs could insert into a variety of corn genes - e.g. those involved in pigment production, starch biosynthesis, and early embryo development, to name but a few.
- that the TEs that inserted into genes were normal residents of the corn genome - they did not come from a viral infection or from outer space.
- that *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 non-autonomous.

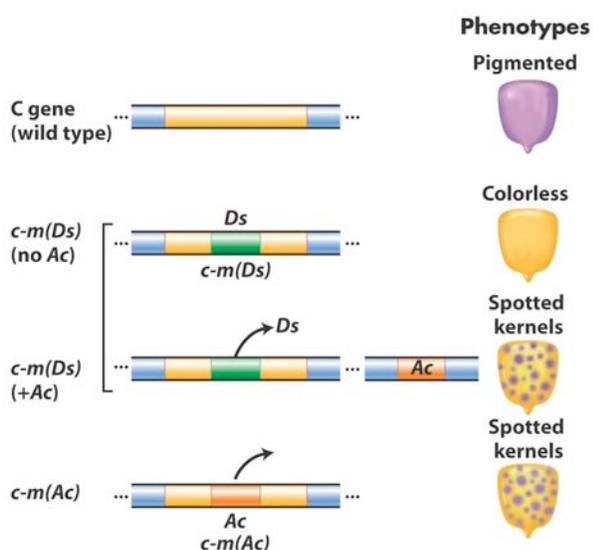


Figure 13-4 (IGA8)

What transposable elements look like:

(b) To the molecular biologist: With the advent of molecular cloning (the ability to use the bacteria *E.coli* as a factory to make large quantities of any DNA fragment from any organism) 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 figure showing what Ac and Ds look like at the DNA level....

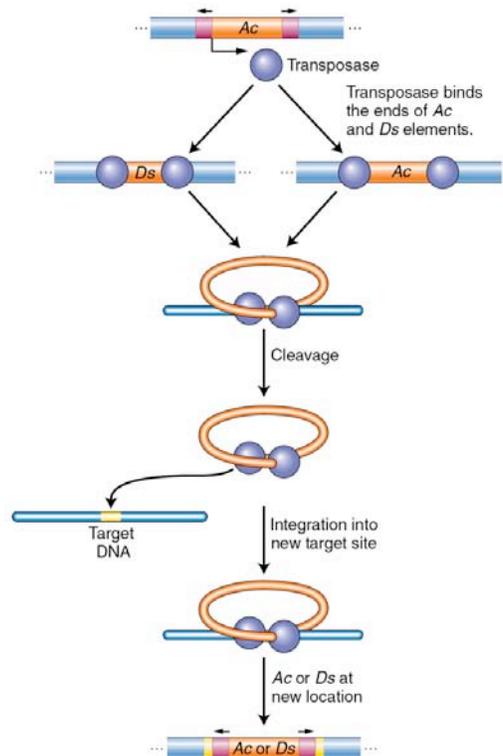
Autonomous element (Ac)



Nonautonomous element (Ds)

(TPase - the gene for the transposase enzyme)

Briefly, Ac contains a single gene - that encodes the transposase protein. The following figure shows how this protein catalyzes the movement of Ac and Ds. The transposase is said to catalyze the "transposition reaction".



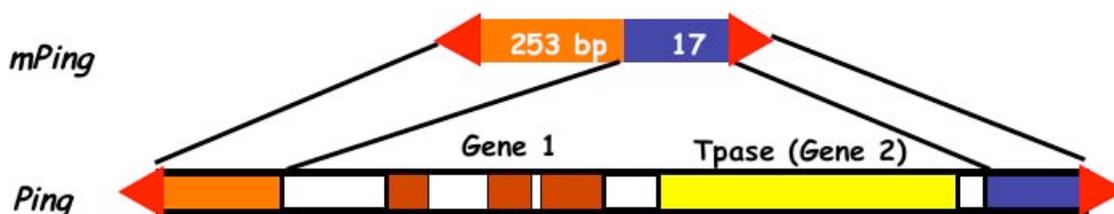
Like many other proteins, the transposase protein can multi-task. First, it is a DNA binding protein that is able to bind specifically to the ends of the Ac element. The protein also binds to the ends of Ds as it is identical to the Ac ends. Such "sequence-specific binding" is mediated by precise contacts 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.

What transposable elements look like:

(c) *To the bioinformaticist*. The Human Genome Project ushered in the genomics era which is characterized by the availability of increasing amounts of genomic sequence from a variety of plant and animal species (animals - including human, drosophila, the worm, dog, mouse, rat, chimp; plants - including *Arabidopsis thaliana*, rice, cottonwood (a tree)]. Most of these sequences are available via databases that you will learn how to access and integrate into your experimental analysis.

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 autonomous Ping and the non-autonomous mPing (for miniature Ping) elements - which were first identified by computational analysis of the rice genomic sequence.



In this experiment, we are going to test the hypothesis that the autonomous Ping element can produce a protein (or two proteins in this case) that can catalyze the transposition of the mPing element. 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. 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. Well, unfortunately, we can't do that - because - like most TEs in the genome, mPing is not inserted into a gene - but rather - it is inserted between genes.

2.2. A Digression - how can organisms survive with so many TEs? Where are TEs located in the genome?

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. 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."

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!

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.

So what does all this have to do with transposable elements? A lot, as it turns out.

Listen carefully: 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. 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.

2.3. Most of the TEs in the genome are INACTIVE.

This leads to a second point you must 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."

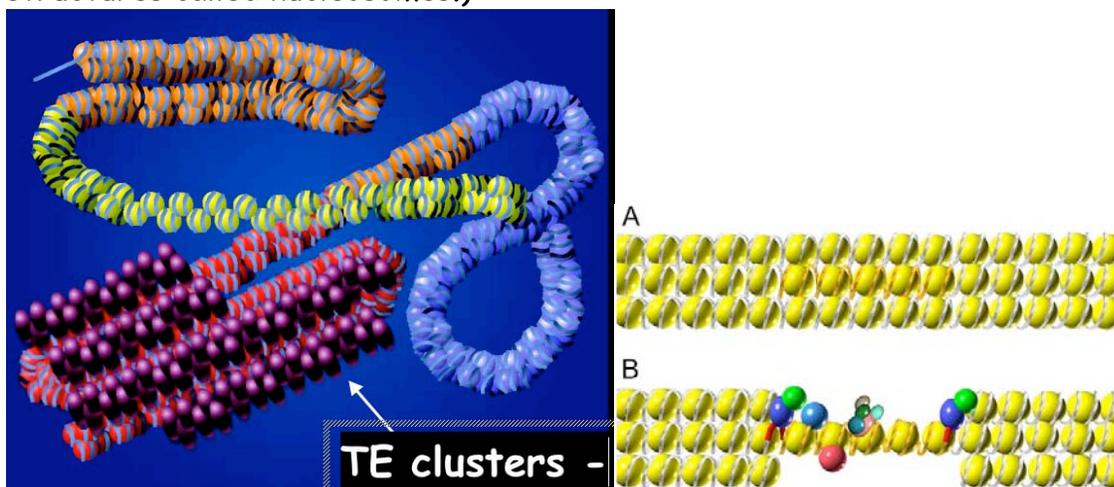
The mutation part is easier to understand (and will be discussed in more detail in Experiment 3). All DNA is susceptible to mutation - usually base pair changes or deletions. This happens (very rarely) when there is an error during replication and the wrong base is inserted - for example a G is put opposite 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 (put on that sunscreen!) 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 ORGANISM'S GENOME, MOST OF THE GENES WILL BE ACTIVE WHILE MOST OF THE TES WILL HAVE SUSTAINED INACTIVATING MUTATIONS)

Epigenetic regulation of TEs (another digression):

In chromosomes, DNA can be encased so securely by some proteins that other proteins cannot access the nucleic acid for transcription, and this process known as epigenetic silencing.)

One of the ways this occurs is through a phenomenon of wrapping and unwrapping of DNA around the core histones in the nucleus and referred to as *chromatin condensation*. (Histones are proteins that DNA tightly coils around to form structures called nucleosomes.)



Chromatin, an organized structure of nucleosomes in eukaryotic cells, governs diverse cellular processes including gene transcription, DNA replication, and DNA repair. Inappropriate gain or loss of chromatin structure plays a causal role in such process as carcinogenesis. For instance, just as genetic mutations in tumor-suppressor genes can cause cancer, epigenetic silencing of tumor suppressor genes can play a role in the development of a variety of tumors; however, in contrast to permanent genetic mutations, epigenetic silencing is reversible. Thus, understanding how epigenetic silencing is achieved in normal cells can help medical researchers to control inappropriate gain or loss of silencing of tumor suppressor genes in cancer cells and help to find novel therapeutics for prevention and treatment of tumors. (Thanks to the Mayo Clinic for this explanation.)

So let's say it again: Most transposable elements are rendered inactive by mutation or epigenetic silencing.

But remember, *most* doesn't mean *all*! We are still learning the language of TEs, and there is much more to learn—and you will be adding to our knowledge of them during your work this semester.

The next couple of sections will introduce us to “reporter genes”, the model plant *Arabidopsis thaliana*, and the bacteria *Agrobacterium tumefaciens* - all, believe it or not, are used in our first experiment....

2.4 A Visual Assay for the Movement of TEs



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*. With that in mind, we will press on.

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 is, to use resolutely unscientific language, *way cool*. It comes from the jellyfish *Aequorea victoria* and fluoresces green when exposed to blue light. (You will get a chance to see this during the semester!) It is widely used as a reporter gene in labs but has also been used for other strange purposes. Researchers bred Alba, a fluorescent bunny, using GFP, for purposes of "art and social commentary," though the point was probably lost on the rabbit. A company in California also tried the same thing with another species, promising to sell glow-in-the-dark zebra fish, but regulators banned the fish before they made it to the market.

Bunnies and fish aside, 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.

Actually, the way a GFP reporter works is a bit more complicated than we said above. There are, in fact, two general classes of reporter genes.

*In the first, a promoter from a cell is fused to the reporter gene so that the expression or transcription from the gene's promoter can be visualized. For example, if we isolate a mutant gene that leads to aberrant root growth, we could isolate the gene, identify its promoter, engineer a gene that has this promoter fused to GFP, transform the construct into the plant, and then look to see where the GFP is expressed. A lot of work to be sure, but if we did it right, the GFP should be expressed only in the roots. Amazing.

*In the second, a gene is disrupted by a transposable element so there is no GFP expression unless the TE excises. Also pretty amazing - and reminiscent of the "natural" visual assay - the spotted corn kernels that led McClintock to discover TEs.

GFP, as we have seen, is normally produced by the jellyfish *Aequorea victoria* which, when agitated, emits a bright green flash. Presumably a defense mechanism to blind attackers, the fluorescence is often observed in the wake of ships passing through ocean waters containing the jellyfish.

Scientists researching the GFP in jellyfish soon realized that if they could isolate the GFP gene and fuse its DNA coding sequence to those of other proteins whose expression or location is of interest, they would have an immensely valuable research tool, functioning as a reporter of gene expression *in vivo* over time.

Using GFP in plants, fluorescent-imaging microscopy can be used to track the expression and location of proteins and other microstructures within organisms as diverse as viruses and nematodes such as those with a destructive effect on agricultural crops.

2.5. *Arabidopsis thaliana*



Remember our discussion of model organisms? One of the features of a good model organism is that it can be easily transformed. And an easily transformed organism is one that can be easily analyzed. The ideal model organism would be one that is small, has a fast generation time, and whose genome (or a large part of it) is known.

That brings us to your new best friend, *Arabidopsis thaliana*. You have already been introduced, but it won't hurt to meet again: "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. Changes in the plant are easily observed, making it a very useful model." Being a model plant, *Arabidopsis* is easily transformed, which leads our narrative in another interesting direction.

Question: How in the heck does this system using GFP to study transposable elements work in the first place? To answer this question, we have to begin, strangely enough, with a soil-borne bacterium with a nifty little secret. In this next section you will see how a natural mechanism for inserting pathogen DNA into plant cells has been exploited to introduce all sorts of genes into plant cells.

BTW (by the way), a plant or any organism that contains foreign DNA introduced by a scientist in the lab is said to be "transformed" and is also called a "transgenic" organism.

2.6. The Little Bug That Could: *Agrobacterium tumefaciens*



The agriculture industry had a problem. A disease called Crown Gall was a problem among a large number of plant species. It was caused by something called *Agrobacterium tumefaciens*, and it caused serious problems in crop plants. Researchers needed to find out what made this bacterium tick and along the way, they discovered an amazing and, as it turns out, very useful trick the bacterium uses to get inside a plant.

What they discovered was that symptoms of the disease are caused by the insertion of a small segment of DNA (known as the T-DNA, for "transfer DNA") into the plant cell, which is incorporated at a semi-random location into the plant genome. The precise mechanism of how that happens is a bit too complicated to explain in its entirety here, but suffice it to say that scientists found out that the DNA transmission capabilities of *Agrobacterium* formed a near-perfect model system for artificially inserting foreign genes into plants.

In 1977, two groups acting independently discovered the gene transfer mechanism between *Agrobacterium* and plants: Mary Dell Chilton, a postdoctoral associate at the University of Washington, and two other researchers 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. The plasmid T-DNA that is transferred to the plant 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. This process has been performed using firefly luciferase gene to produce glowing plants.

The plasmid from *Agrobacterium* used to produce transgenic plants is called the Ti plasmid. The natural behavior of this plasmid makes it well suited to the role of a "vector" for plant genetic engineering. If the DNA of interest could be spliced into the T-DNA, then the whole package could be inserted in a stable state into a plant chromosome. This system has indeed been made to work in this way but with some necessary modifications.

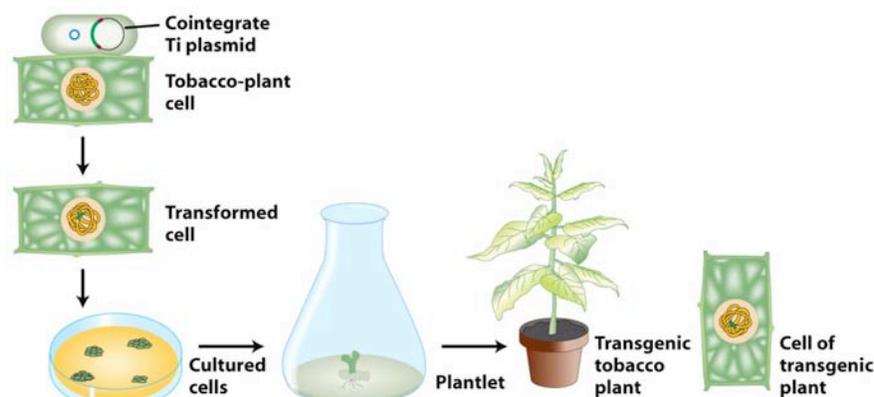


Figure 20-26
Introduction to Genetic Analysis, Ninth Edition
© 2008 W.H. Freeman and Company

There's another important fact, though. In addition to "your favorite gene" (YFG), the DNA transferred into the plant genome also contains an antibiotic resistance gene, which was engineered into the plant by scientists (see below).

2.7. Antibiotics for the World!

Actually, this section has nothing to do with human health, just as the last section was only tangentially about agriculture. But the use of antibiotic marker genes in understanding plant transformation is crucial to many labs, and to our class!

The techniques used for transferring a new gene into a plant are rather inefficient. Very few cells actually take up the gene of interest; when conditions are favorable, only some five cells in a thousand are genetically modified, meaning they have taken up the DNA in the solution and incorporated it into their genome. Most often this ratio is lower. In order to find the cells that have been successfully transformed, some kind of marker is needed.

To do this, the gene that will give the plant its new trait (gene of interest) is coupled with a marker gene, usually by putting the two genes next to each other on the same DNA molecule. Plant cells are then transformed with T-DNA containing both genes. The vast majority of these marker genes make the plant resistant to a particular antibiotic.

Plant cells that express an antibiotic resistance marker gene (ABR gene) are not harmed by the antibiotic. Treating the cells after the gene transfer with an antibiotic allows only the successfully transformed cells to survive. These cells also possess the gene of interest. Transgenic plants containing the gene of interest are then regenerated from these individual, successfully transformed cells. Although the marker gene serves no purpose after this procedure, it remains part of the genetically modified plant.

All this was done to get the transgenic Arabidopsis plants that we will be using in Experiments 1 and 2 this semester, and it was done in the Wessler lab. Everything you've read about—T-DNA, reporter genes, antibiotic genes and so forth—was inserted into the Arabidopsis genome by someone upstairs in a previous series of experiments.

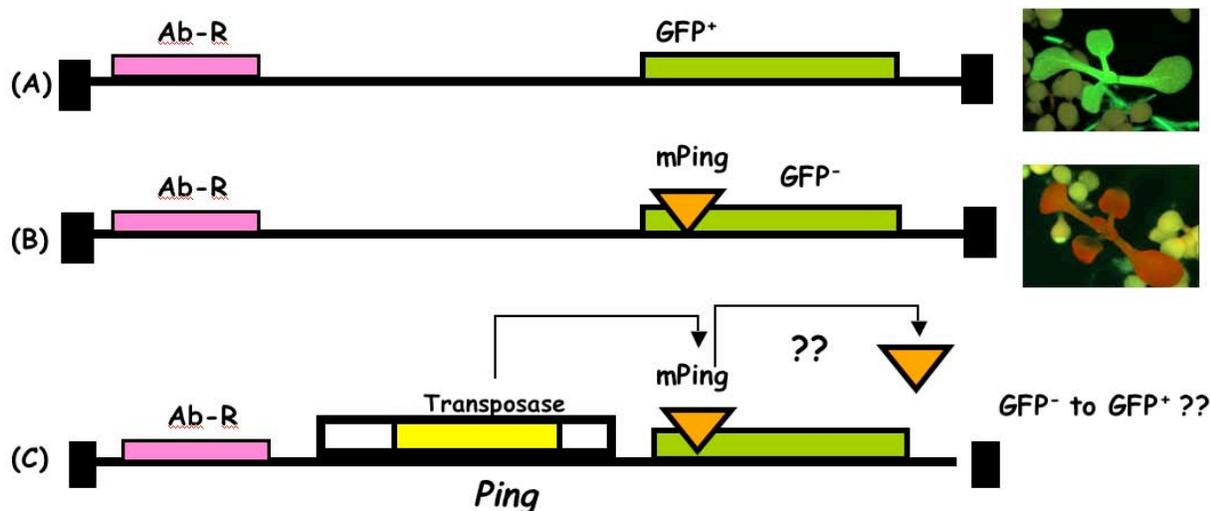
2.7 - Putting it all together

So, let's see. Geneticists had never isolated an active TE from rice like the Ac and Ds elements discovered by Barbara McClintock's in maize. To find an active TE in rice, the Wessler lab used computer analysis of the sequence of the entire rice genome to identify about 50 nearly identical elements in the rice genome. They called this element mPing. They concluded that mPing must be a nonautonomous element as it was too small (430bp) to carry a transposase gene (an average protein is about 500 amino acids - which would be encoded by a gene of at least 500 amino acids X 3 bp per amino acid = 1500bp). So, they hypothesized that the rice genome should contain another TE that contains the gene for the transposase necessary to move mPing. Again they went back to the drawing board (the computer actually) and searched this time for a likely autonomous element (analogous to the maize Ac). They found a candidate TE which they called Ping - that had the same ends as mPing but was much longer (~5000 bp) and contained a transposase gene (actually Ping encoded 2 genes but don't worry about that).

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 transposase that could move 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:



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

(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-*) and the red color is due to chlorophyll fluorescence.

(C) plants containing this T-DNA in their genome are the actual experiment. If our hypothesis is correct, then *Ping* will produce a transposase that will bind to the ends of *mPing* and catalyze its transposition out of the *GFP* gene restoring gene function.

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

However, we are not out of the woods yet. We need solid experimental evidence that *mPing* has actually excised in plants with T-DNA (C) but NOT in plants with T-

DNA (B). Experiment 1 is designed to determine just that - whether or not mPing has excised from the T-DNA.

2.8. PCR analysis of *A. thaliana* DNA

The PCR reaction is summarized below. The region to be amplified in expt 1 is shown below. Primers for your experiment are 22 nt (nucleotides) and are derived from the sequence of the GFP gene flanking the mPing insertion site.

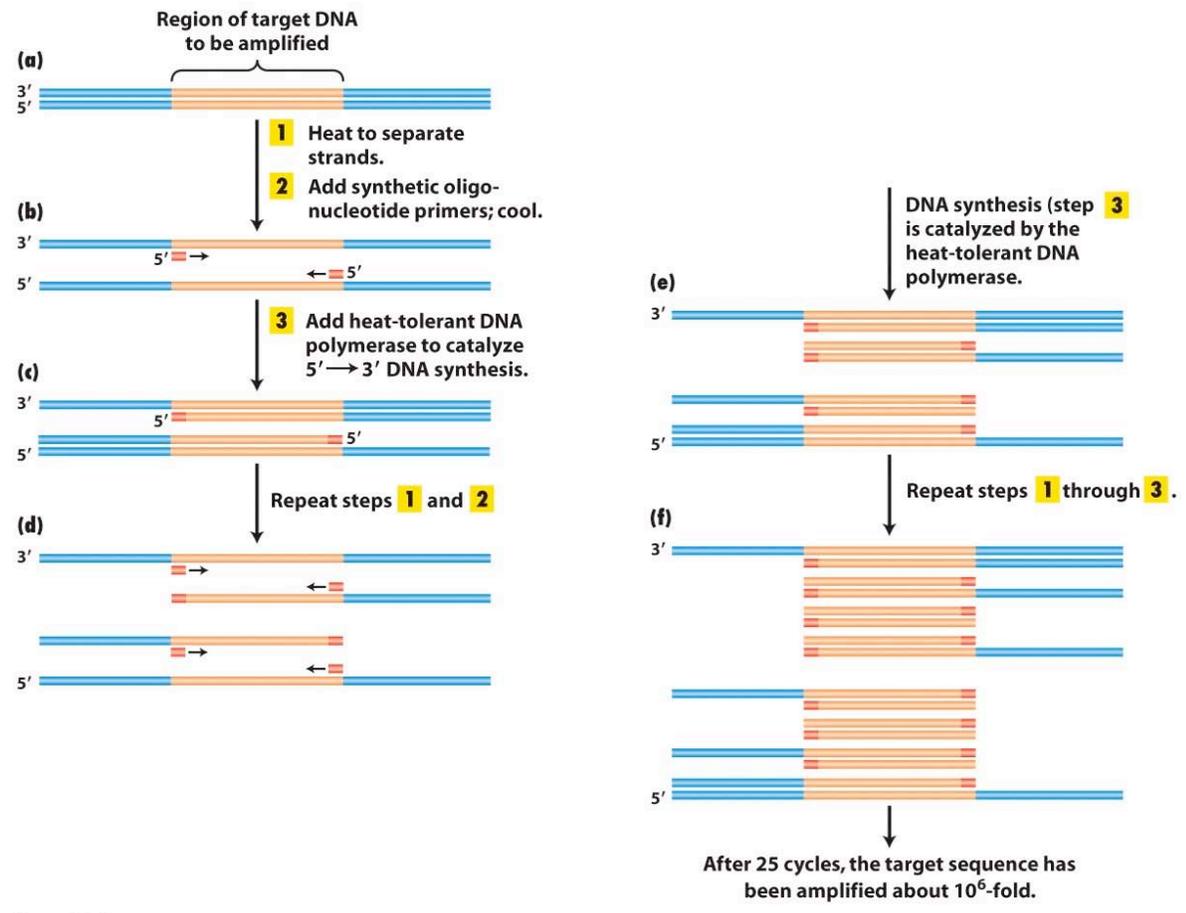
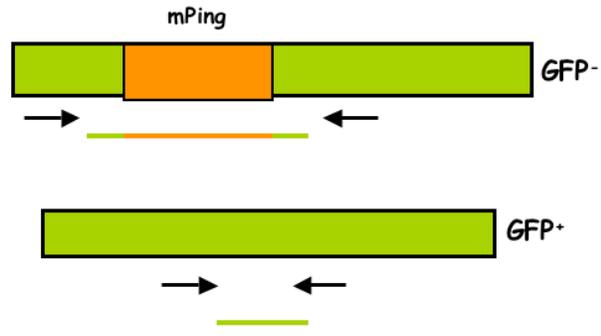


Figure 20-14
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company



2:9 Experiment One: Examination of mPing excision from *A.thaliana* leaf DNA by PCR: Day 1: Tuesday, August 28

Here it is: Your chance to become a Real Scientist. Everything before this is background. This is the real thing, and Wessler and Williams would like to thank several members of the Wessler Lab for making the material that you will use possible.

1. One week before actually doing PCR with leaf DNA, our TA, Yujun will start growing the plants we will use in this experiment by plating *Arabidopsis* seeds on agar 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. Yujun then put the agar plates 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. Yujun will, however, show you what he did.

2. Examining our seedlings

Once we have growth, we can examine the growing plants under the powerful (and expensive!) light microscope on the fourth floor of the Plant Sciences Building. Pay attention to what you see - and ask lots of questions, as we will be discussing the meaning of the beautiful patterns on the leaves.

3. Leaf PCR—The isolation of DNA from leaf tissue

Once you have grown your *Arabidopsis* seedlings, we're ready to look at the seedlings and isolate leaf DNA. There are two first steps: **extracting** the DNA from our *Arabidopsis* leaves and then **amplifying** it by using PCR. Here's how:

1. Prepare five 1.5 ml tubes

Label "WT" on the cover of one tube, "mP" (for mPing, no Ping) on another, "mP +P" (for mPing plus Ping) on two, and W (for water, negative control).

2. Add 100ul extraction solution to each tube

3. Pick 4 *Arabidopsis* seedlings (1 WT, 1 mPing and 2 mP+P), and put each of them into corresponding labeled tube. Make sure that the leaf is immersed in the extraction solution, vortex. Clean forceps between steps as shown in class.

4. Incubate the tubes at 95°C in the heat block for 10 minutes. Note that leaf tissues usually do not appear to be degraded after this treatment, though in fact it is. At the same time, prepare another four 1.5 ml tubes and make the PCR reaction mixture as following in each tube.

PCR:

5. Make a 5X PCR stock (5 times more concentrated than the final PCR mix) by combining the amount shown for each of the four ingredients into a single tube **marked "5X"** then briefly vortexing this to mix and aliquoting 16 ul into each of your 5 tubes:

		(X 5 stock)
Extract-N-Amp PCR reaction mix	10 ul	50 ul
forward primer (10pmol/ ul)	1 ul	5 ul
reverse primer (10pmol/ ul)	1 ul	5 ul
sterile H ₂ O	4 ul	20 ul

6. Take out the tubes, place them on ice, and add 100 ul of the dilution solution to each tube, and then vortex for a few seconds to mix well.

7. Apply 4 ul leaf DNA extract into each PCR reaction mixture except the fifth (control) tube. Put back on ice.

8. Yujun will transfer each PCR mixture to a 24-well plate and start PCR

1 cycle for:

initial denaturation 94°C 3 minutes

30 cycles for:

denaturation 94°C 30 seconds

annealing 58°C 30 seconds

extension 72°C 1 minutes

[Note: "30 cycles" means all steps—denaturation, annealing, and extension—are repeated 30 times before going on to the next step (look at pages 13-16 at the beginning of this handout for definitions of PCR terms).]

final extension: 72°C 10 minutes

PCR will be completed after you leave class. The PCR products will be stored in the frig (at 4°C) and will be waiting for you on Thursday.

2.10: Experiment #1, Day 2 (August 30) Electrophoresis & DNA fragment purification

We are now going to physically separate the PCR products in each tube by resolving them by electrophoresis on an agarose gel.

1. Making the gel (Yujun will do this part and the class will watch and ask questions):

2. Preparing your PCR samples (stored in the frig since Tuesday)

- Add 4 μ l loading dye to each of your PCR tubes
- Using your micropipette, load 5 μ l of the DNA size standards to one a well at the left end of the gel. Load 12 μ l of the PCR product into the wells.
- Attach the leads to the gel and set the power supply to run at 150V.
- After ~20 minutes, turn off the power, remove the gel tray from the apparatus, and take a picture of the gel for your records.

3. Purification of the PCR band from agarose gel (using the Qiagene Gel Extraction Kit) (Yujun will demonstrate):

--**Excise the gel slice** containing the desired band (both of the upper and lower bands of "mP+TP"). Gently slide the gel off the tray on to a sheet of Saran wrap covering the transilluminator (UV source). Put on protective face shield to prevent exposure to UV light. Turn on transilluminator and excise the gel slice containing the desired band with a new razor blade. Put the gel slices into 1.5 ml tubes.

--**Weigh the gel slice** in 1.5 ml 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.

--**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 vortexing the tube every 2-3 min during the incubation.

--**After the gel slice has dissolved completely, add 1 gel volume of isopropanol** to the sample and mix. For example, if the agarose gel slice is 100 mg,

add 100 μ l isopropanol. This step increases the yield of DNA fragments < 500 bp and > 4kb.

--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.

--Discard flow-through and place QIAquick column back in the same collection tube.

--Add 0.5 ml of buffer QG to QIAquick column and centrifuge for 1 min. Discard the flow through. This step is only required for directly sequencing.

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

--Discard the flow through and centrifuge for another 1 min at 13,000 rpm. *IMPORTANT: This spin is necessary to remove residual ethanol (Buffer PE).*

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

--To elute DNA from the column, add 30 μ l H₂O to the center of QIAquick membrane, leave column on bench for 2 min, and centrifuge the column for 1 min at 13Krpm.

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 (Yujun will do this).

2.11. Experiment #1 Day 3 (September 6, 2007) The Bioinformatics Part of Experiment #1: Analyzing your sequences

Background to DNA sequence analysis

Why sequence DNA?

The sequence of DNA encodes the necessary information for living things to survive and reproduce. Determining the sequence is therefore useful in 'pure' research into why and how organisms live, as well as in applied subjects. Because of the key nature of DNA to living things, knowledge of DNA sequence may come in useful in practically any biological research. For example, in medicine it can be used to identify, diagnose and potentially develop treatments for genetic diseases. Similarly, research into pathogens may lead to treatments for contagious diseases. Biotechnology is a burgeoning discipline, with the potential for many useful products and services.

How your DNA from Expt 1 was sequenced

DNA sequencing is the process of determining the nucleotide order of a given DNA fragment. Thus far, most DNA sequencing has been performed using the chain termination method developed by Frederick Sanger (who won a Nobel Prize for this discovery). The method is also called Sanger Sequencing. This technique involves the synthesis of copies of your input DNA by the enzyme DNA polymerase. However, one difference between this reaction and your 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.

In chain terminator sequencing DNA polymerase begins to synthesize DNA at a specific site by using a DNA primer - much like you used for PCR. This means that we need to notify the DNA sequencing facility of the sequence at the ends of the DNA we sent to them so that the sequencing reactions can be initiated. We will explain this in class.

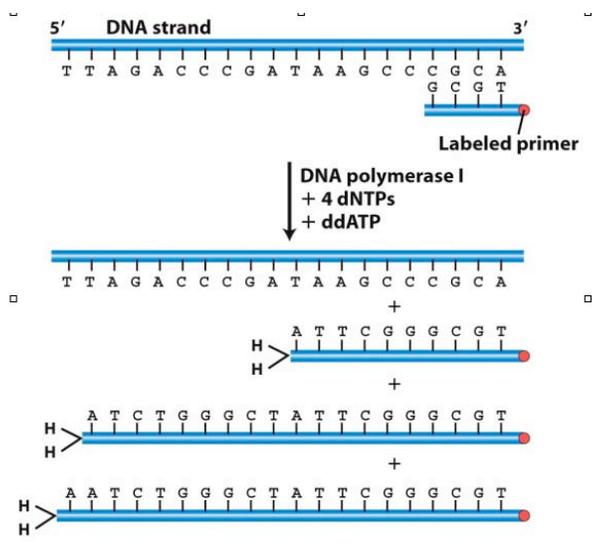


Fig.1 chain terminator sequencing (Sanger sequencing)

So, lets review - Yujun sent the DNA facility your DNA samples and the sequence needed for primers. They then took these samples, added 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).

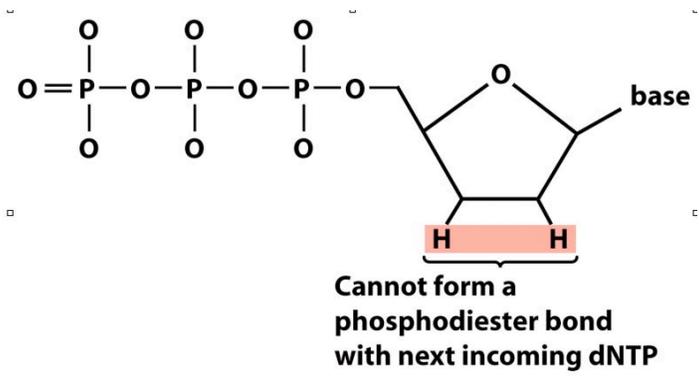


Fig. 2 A chain terminating nucleotide (di-deoxynucleotide).

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.

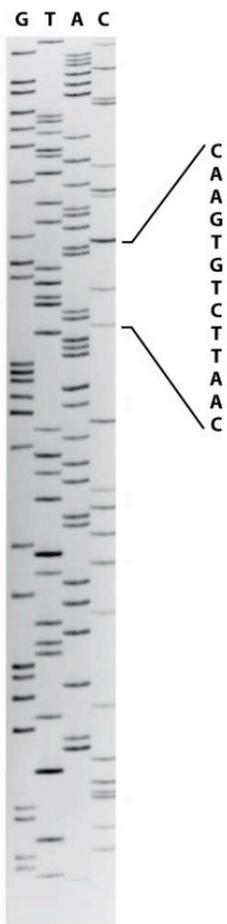
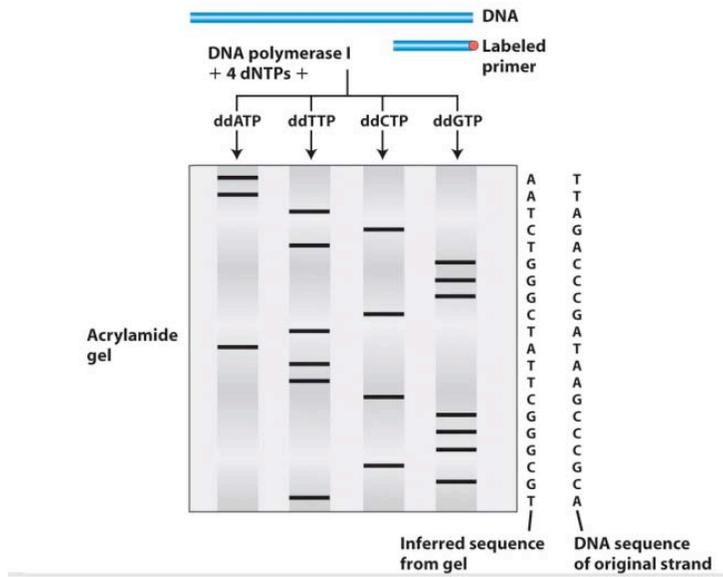


Fig.3 Part of a radioactively labeled sequencing gel

An alternative to the labeling of the primer is to label the terminators 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.

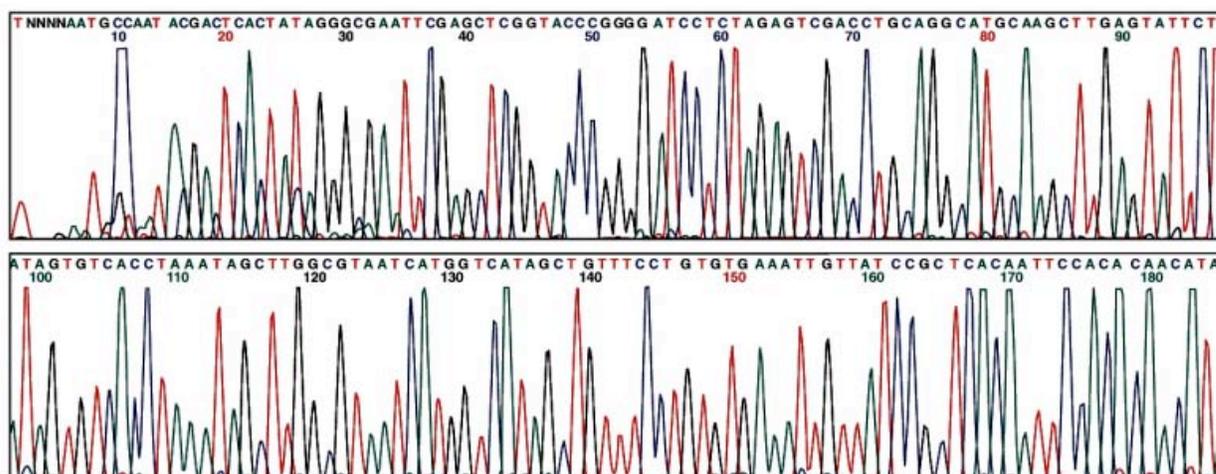


Fig. 4 Example of (the start of) a Sanger sequencing read. The four bases are detected using different fluorescent labels. These are detected and represented as 'peaks' of different colors, which can then be interpreted to determine the base sequence, shown at the top.

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.

Back from the Sequencing Facility*

Now you can open your computer and begin to analyze the sequences that have been returned from the sequencing facility. Like most experiments done for the first time, some of your sequences are not very pretty!

1. Your sequences have been downloaded to the class share file and we will explain how to find them.

The sequencing facility to which we submitted our DNA samples is:

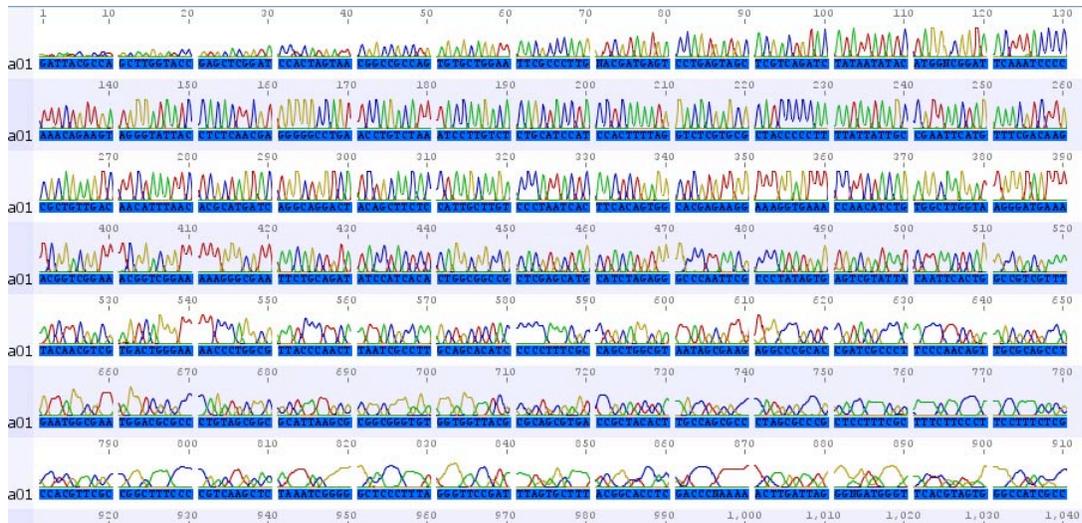
Integrated Biotechnology Laboratories
Riverbend Research Lab, Room 161
110 Riverbend Road
University of Georgia
(706) 542-6409

There are 2 files for each sequencing result, one chromatogram file and one text file. The text file contains the DNA sequence. The chromatogram file provides the quality information of sequencing result.

2. Open the upper and lower bands' text files. Check the sequences manually first. You will see many N's. This has to do with the quality of the sequencing reads. To understand what this means, we will have to open the chromatogram file....

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

Once opened, you will see curves in one of 4 colors representing A,G,C and T. Each nucleotide has a corresponding peak color. The higher/sharper the peak is, the more reliable is the corresponding nucleotide sequence. Very low peaks or twisted curves mean poor sequencing quality.



Usually, for each sequencing reaction, the high quality/reliable region is about 500 bps. Therefore, your lower band's sequence quality may be ok, but the upper band's sequence may be quite low as the sequence was read further from the primer. These sequences will be "trimmed" off when you perform the analysis.

4. In short, you will be comparing (Blasting) your sequences with two sets of sequences - either (i) reference sequences that we will provide or with (ii) the sequence databases that are at the NCBI site. For both comparisons, you will be using the Blast program at the NCBI site. The difference between what you already did at the Blast site and what you will do today is summarized below....

First, go to the NCBI site and click Blast -

The screenshot shows the top navigation bar of the NCBI website. On the left is the NCBI logo. To its right is the text "National Center for Biotechnology Information" with "National Library of Medicine" and "National Institutes of Health" below it. A dark blue navigation bar contains links for "PubMed", "All Databases", "BLAST", "OMIM", "Books", "TaxBrowser", and "Structure". Below this is a search bar with a dropdown menu set to "All Databases", a text input field, and a "Go" button.

Scroll to the bottom, click [Align two sequences using BLAST\(bl2seq\)](#).

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two sequences using BLAST (bl2seq)

There are two input windows on this page: Sequence 1 and Sequence 2. You can copy and paste your sequence and reference sequence into each window, leave all parameters as default, and then click Align at bottom.

The screenshot shows the "Align two sequences using BLAST (bl2seq)" interface. It features two identical input sections, "Sequence 1" and "Sequence 2". Each section has a text input field for "Enter accession, GI or sequence in FASTA format" with "from:" and "to:" labels and input boxes. Below each field is a "Browse..." button. At the bottom of the page are "Align" and "Clear Input" buttons.

An alignment example is showed below:

```
Score = 264 bits (137), Expect = 5e-67
Identities = 159/169 (94%), Gaps = 2/169 (1%)
Strand=Plus/Plus

Query  554  TAATCTAGAGGATCCAAGGAGATATAACANTGAAGACTAATCTTTTCTCTTTCATCT 613
      |||
Sbjct  126  TAATCTAGAGGATCCAAGGAGATATAACAATGAAGACTAATCTTTTCTCTTTCATCT 185

Query  614  TTTCACTTCTCCNATCATTATCCTCGGCCGAATTCAGTAAAGGAGAANAAC TTTCACTG 673
      |||
Sbjct  186  TTTCACTTCTCCTATCATTATCCTCGGCCGAATTCAGTAAAGGAGAAGAAC TTTCACTG 245

Query  674  GAGTTGCCC-ATTCTTG TNG-ANTAGATGGTGATGTTAATGGGCACAA 720
      |||
Sbjct  246  GAGTTGCCCAATTCTTGNTGAATTAGATGGNGATGTTAATGGGNACAA 294
```


**DNA SEQUENCE REQUEST FORM
TO 650 BASES**

DATE SUBMITTED _____ TIME SUBMITTED _____ WORK ORDER NO. _____

SUBMITTED BY _____ PHONE NO. _____ DATE _____

MAILING ADDRESS _____

PURCHASE ORDER OR GRANT NUMBER _____

GRANT NAME _____

E MAIL ADDRESS _____ PI _____

BILLING ADDRESS _____

PRICE: \$17.00 per reaction (one template/primer combination)

Dec. 2004

Plasmid or PCR? SS or DS	Template Size	GC or AT Rich	RUN #	Put A or S	TEMPLATE NAME Put "A" for Archived Put "S" for Submitted	Primer Tm if Known	Put A or S	PRIMER NAME Put "A" for Archived Put "S" for Submitted
			1					
			2					
			3					
			4					
			5					
			6					
			7					
			8					

COMMENTS:

A form like this was filled out with our order to sequence your DNA samples.

<http://www.ssf.uga.edu/assets/docs/650.pdf>

PBIO3240L First Laboratory Report: Due September 13.

Your lab report should be in the form of a narrative that addresses ALL of the following questions and should be no longer than 1000 words. This is an open book exam.

What was the rationale of the experiment? What hypothesis was being tested?

What is a visible marker?

Why was Arabidopsis used in this experiment?

What are mPing and Ping?

What is T-DNA? What is a visible marker?

What exactly are the green spots on the leaves of Arabidopsis?

Why are they of different size?

A well-designed experiment contains many "controls" - including a "positive" control and a "negative" control. Using the microscope, you looked at 3 Petri dishes containing 3 different strains of Arabidopsis. Then, in the laboratory you extracted DNA from each of these strains and performed PCR. These plants were:

(i) wildtype,

(ii) a strain with T-DNA with mPing inserted in GFP

(iii) a strain with T-DNA with mPing inserted into GFP and with the Ping element.

In addition, in the laboratory, you performed one PCR with just water.

Include a discussion of the overall design of this experiment by mentioning the significance of each of these samples in your narrative. Comment on what each of these strains looked like under the microscope and the result of each PCR (as deduced from your gel picture).

If you knew the sequence of another Arabidopsis gene (like the one encoding a particular enzyme), could you use the DNA that you extracted from the Arabidopsis leaves to isolate that gene? To do this, how would you modify the protocol for Experiment 1?

Why does the DNA go to the cathode during gel electrophoresis? What is the identity of the bands seen on your gel picture?

Instead of loading DNA on the agarose gel, what would happen if we loaded protein on the gel and performed electrophoresis (extra credit if you know this)?

With regard to the bioinformatic part of the experiment (the analysis of your DNA samples), address the following:

1. Did your results provide support for the hypothesis?
2. What is the significance of the chromatogram file data? How did this file allow you to see why bases are called N's?
3. Did all of the sequences in your sequence (text) file originally come from the rice genome? If not, where were they derived from? How would you figure this out using Blast if you did not know?

Please tell us how we could improve this experiment (not included in your word count but much appreciated).

Also not included in the word count- tell us anything else that is on your mind (about this experiment!).

3. Experiment 2: The Land of the Ancient Mariner

Now that you're becoming a pro with laboratory techniques, we're going to go on to our second experiment of the semester, and it deals with a different transposable element. This element belongs to the superfamily called Mariner, and we will be dealing more specifically with an element called Osmar, which is a member of the Mariner superfamily.

OK - we slipped in a few new words here -- which means that you will need to understand a few more concepts and terms dealing with transposon biology before we can launch into this experiment. First, most plant genomes contain different families of transposable elements. This concept is central to understanding what genomes are made of.

3.1 What is a TE family?

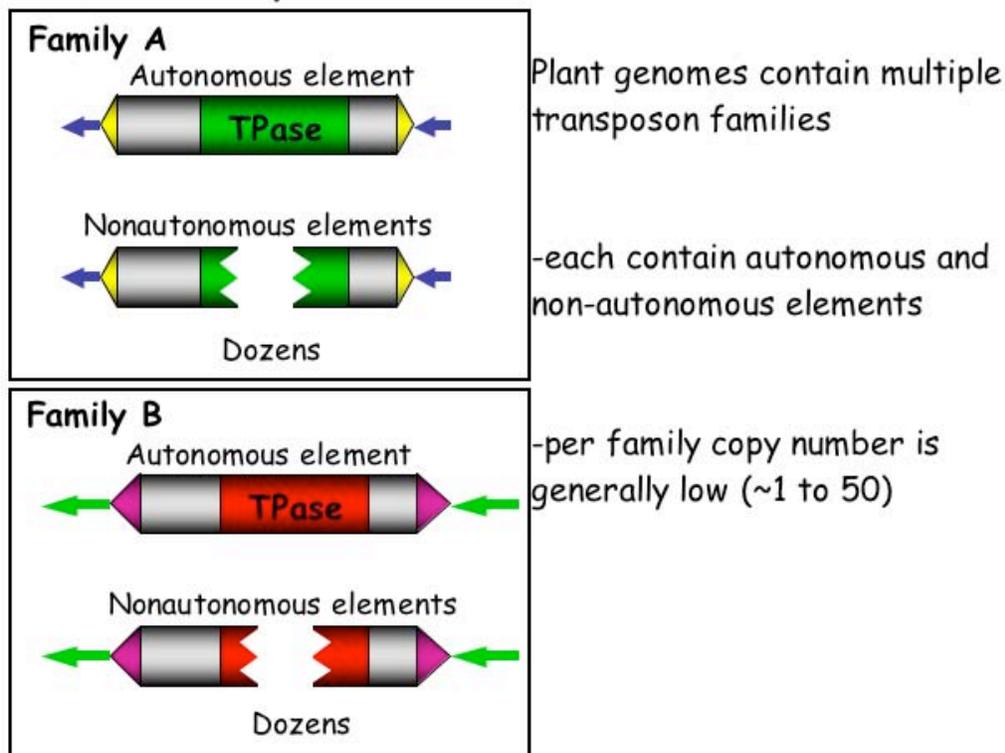
We have already been introduced to two TE families. One family contains the Ac and Ds elements while the second family contains Ping and mPing elements.

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 one or more autonomous element and many copies of nonautonomous elements (the maize genome has over 50 copies of Ds). The transposase encoded by the Ac element can mobilize both Ac and Ds elements. If there is no Ac element in the genome, all of the Ds elements will be "stuck" where they are - they will not be able to move elsewhere in the genome because there is no transposase to catalyze their movement.

Another TE family that you have had direct experience with is Ping/mPing. As you saw in the first experiment, the Ping transposase can mobilize the mPing element. Like the Ac/Ds family, there can be many mPing elements in the rice genome (While most strains have <50 mPing elements, some have over 1000!). If there is no Ping element in the genome, the cell cannot make transposase and all of the mPing elements are stuck and cannot move around.

A very important feature of TE families is that they are independent of each other. 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. The reason for this is quite simple. A transposase usually works by first binding to a specific DNA sequence near the ends of the element (as shown on page 18). 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.)



3.2. What is a transposable element superfamily?

Recall that the first TE discovered by McClintock was *Ds* - as a site of chromosome breakage in maize. She then showed that chromosomes only broke at *Ds* if a second genetic element, which she called *Ac*, was also present in the genome. Thus, *Ac/Ds* is the first family of transposable elements. McClintock then discovered a second TE family which she called *Spm* (for Suppressor-mutator - a long story!). The autonomous element in the *Spm* family is called *Spm* and the nonautonomous element is called *dSpm* (for defective-*Spm*). Thus, *Spm-dSpm* is the second family of transposons.

McClintock's discoveries resulted from genetic analyses of corn plants. After the discovery of TEs in maize, researchers working with other model organisms,

including *Antirrhinum majus* (a.k.a. snapdragon) *Drosophila melanogaster* (a.k.a. the fly) and *Caenorhabditis elegans* (a.k.a. the worm) also identified TEs through genetic studies. In the 1980's when it became possible to isolate specific genes, researchers isolated McClintock's Ac, Ds, Spm and dSpm elements and the elements from snapdragon (called Tam 1,2,3 etc), the fly (called P-elements, mariner elements and others) and the worm (called Tc1, 2 and 3 elements).

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

These similar transposases were subsequently organized into superfamilies. Fortunately, after all of the sequencing of genomes and comparisons of TEs, there are now known to be fewer than 10 superfamilies of transposases. Some superfamily names and elements and some members include: hAT (includes Ac, Tam3, P elements), CACTA (includes Spm, Tam1), PIF/Harbinger (includes Ping), Mutator and mariner. In this second experiment we will analyze the movement of a member of the mariner superfamily.

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

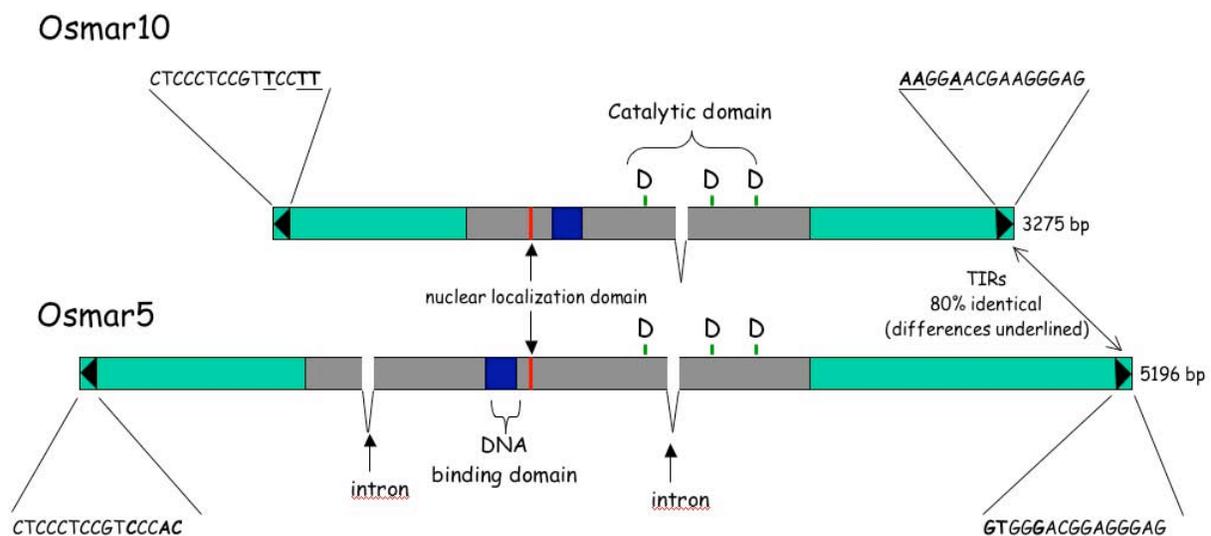
In short, many. First, members of most superfamilies are present in all plant genomes including maize and rice and are also present in most animal genomes. For example, the rice genome has mariner, PIF/Harbinger, hAT, CACTA and Mutator elements.

In addition, each superfamily in each genome contains many families. For example, the rice genome contains Ping and over 100 related but clearly different elements. Among these 100 elements are some that are very similar and others that are quite different. To understand the relationships between elements, their sequences can be organized and visualized as a family tree. The scientific term for such relationship maps is "phylogenetic trees". In experiment #3 you will learn how to construct your own phylogenetic tree from TE sequences that you will retrieve from the database. So... we will revisit trees in the background text for experiment 3.

3.4 Experiment #2: background and rationale

Overview: In this experiment you will be analyzing the movement of another rice transposable element family called Osmar and comparing it with the Ping/mPing family.

The mariner superfamily was named for the mariner element, which was isolated from the fly (*Drosophila*) and then from virtually all plant and animal genomes (even human). Members of the mariner superfamily from different species were given names that were derived from the species name. For example, members of the mariner superfamily from rice are called "Osmar" where the "Os" comes from *Oryza sativa*, and the "mar" comes from mariner. Pretty clever, huh? Computer analysis of the rice genome sequence reveals that there are about 40 Osmar elements and many more nonautonomous elements. These 40 elements are organized into about 25 families - which means that a few of the elements are nearly identical to one or two other Osmars in the genome but others are different enough to say that they are members of different families (with each family having only one autonomous element copy). In the figure below, the key features of two Osmar family members, Osmar 5 and 10 are compared. As you can see - while these elements share similar features [e.g. both transposase genes (the gray regions) have 3 domains - DNA binding, catalytic and nuclear localization] the size of the elements differ dramatically and even the sequences at the ends (the TIRs, triangles) are not exactly the same. We will discuss these differences in class.



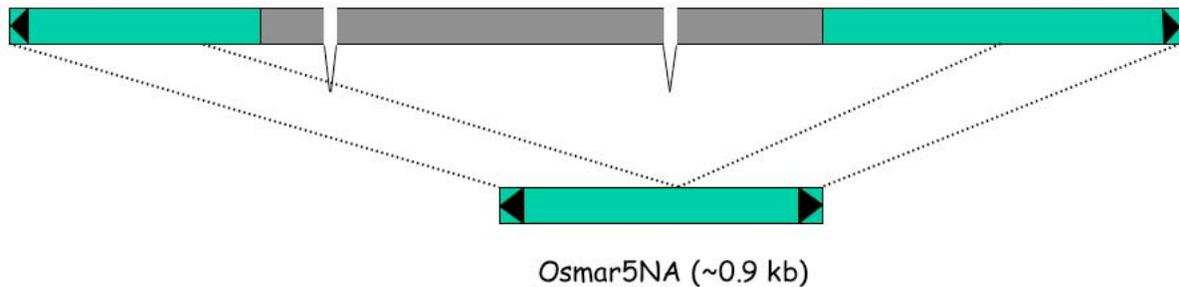
The nonautonomous element and transposase source in experiment 2:

Although the focus of experiment 2 is *Osmar5*, we are not actually using the whole element in this experiment.

Osmar5NA -

First, a nonautonomous version of *Osmar 5* (called *Osmar5NA*) does not exist in the rice genome so one had to be "constructed" in the test tube. That is shown below. As you can see, it is a precise deletion derivative of *Osmar5* and, as such, is analogous to *mPing* (which is a precise deletion derivative of *Ping*).

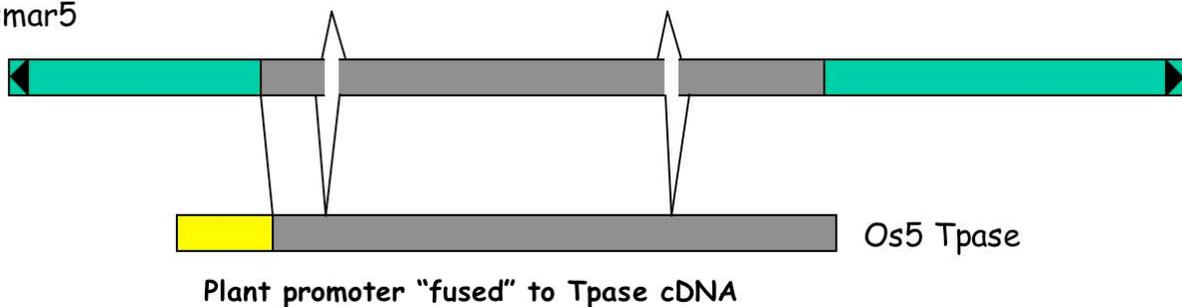
Osmar5 (~5.1 kb)



***Osmar5* Tpase - the transposase source - removing the introns**

Recall in experiment 1 that the transposase was provided by the transposase gene, which was part of the *Ping* element (see page 29). For reasons that will be described later in the course, it was necessary to remove the 2 introns from the *Osmar5* tpase gene and fuse the resulting DNA with a plant promoter, as shown...

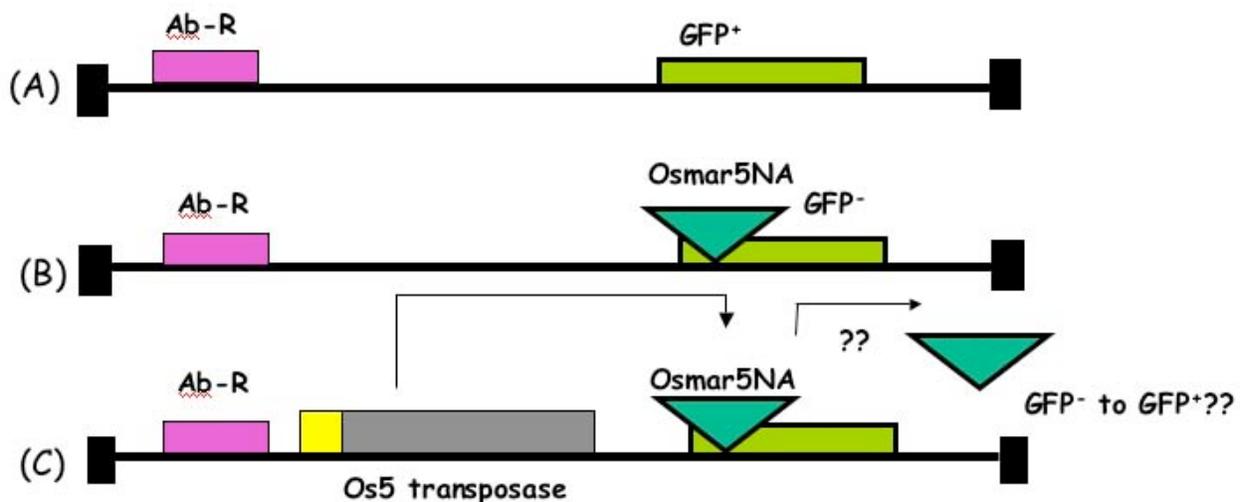
Osmar5



3.5. Putting it all together:

You should all be pros at this by now! Here you are going to be working with Arabidopsis strains that are different than the strains in Expt 1. There we were testing the hypothesis that the Ping element produced a transposase that could bind to the ends of the mPing element and catalyze its excision from the GFP gene. In Experiment 2 we are testing the transposase from a rice element that is a member of a different superfamily - the mariner element called Osmar5. Stated precisely (we need to do that for experiments), we are testing whether an artificial version of the Osmar5 transposase (with its introns removed and fused to a plant promoter - Os5 transposase) can catalyze the transposition of a nonautonomous version of the Osmar5 element that was also created in the test tube (Osmar5NA) by deleting the middle of the Osmar5 element. While that is a mouthful - it is summarized simply below...

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:



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

(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⁻) and the red color is due to chlorophyll fluorescence.

(C) plants containing this T-DNA in their genome are the actual experiment. If our hypothesis is correct, then, just as with Ping and mPing, the Osmar5 Tase will be able to bind to the ends of Osmar5NA and catalyze its transposition out of the GFP gene restoring gene function.

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

3.6: Experiment 2 Protocol Overview:

To make life easier, we're once again using our old friend, Arabidopsis, to help us on this second treasure hunt. Arabidopsis is, in fact, so useful that it's amazing that molecular biologists don't use it as a centerpiece at weddings! You'll be going back over some familiar ground and terms. *You will be repeating steps from experiment 1: looking under the microscope at transgenic Arabidopsis plantlets, extracting DNA from leaf tissue, doing PCR, running a gel and purifying the DNA from bands.* For this reason, we have included a shorthand version of the protocol from experiment 1 and noted where you will be using different biological materials.

The other major difference between this experiment and the last one occurs after the band is purified from the gel. Recall that last time this purified band of DNA was sent directly to the DNA sequencing facility and in return you received one DNA sequence per band. However, for experiment 2 we will be cloning the DNA purified from the gel, inserting (transforming) the DNA into a bacterium called *Escherichia coli* (*E. coli*), isolating individual transformed colonies, and then sending these colonies to the sequencing lab where they will purify the DNA from each colony and determine the relevant sequence in each bacterial colony. This will be explained in more detail below.

Expt 2 protocol: Day 1: September 11, 2007

Arabidopsis seed will be sterilized, plated and incubated as previously described. What will be different is the Arabidopsis strain you will be using will have the constructs shown above. In addition, Yujun will sterilize and plate the same seed as in Expt 1 so that you can compare the GFP spot patterns on the leaves.

1. Examining our seedlings - back up to the 4th floor.

DNA Extraction

1. Prepare four 1.5 ml tubes

Label the tubes on the lid: "WT", "Ona" (OsNA, no OsTpase), and 2 tubes with "ONA +OT" (OsNA plus OsmTpase).

2. Add 100ul extraction solution to each tube

3. Using forceps pick leaves from 4 Arabidopsis seedlings into 4 tubes (1 WT, 1 OsNA and 2 OsNA + OT). Make sure that the leaf is immersed in the extraction solution, vortex.

4. Incubate tubes for 10 minutes at 95°C in the heat block. During this 10 min, prepare five 1.5 ml tubes to be used for PCR (labeled: WT, OsNA, 2-ONA+OT and C for control, this will have no DNA) and make the PCR stock solution below.

PCR

Make a 5X PCR stock (5 times more concentrated than the final PCR mix) by combining the amount shown for each of the four ingredients into a **single tube marked "5X"** then briefly vortexing this to mix and aliquoting **16 ul into each of your 5 tubes:**

	(1X)	(5X stock)
Extract-N-Amp PCR reaction mix	10 ul	50 ul
Forward primer (10pmol/ ul)	1 ul	5 ul
Reverse primer (10pmol/ ul)	1 ul	5 ul
Sterile H ₂ O	4 ul	20 ul

5. Remove tubes from heating block and place on ice. Add 100 ul of the dilution buffer (provided) to each tube then vortex for a few seconds to mix well. Put back on ice.
6. Add 4 ul leaf of this diluted DNA solution into each tube except the fifth (control) tube, where you add 4 ul H₂O instead.
7. Yujun will transfer each of your PCR tubes into a 24-well plate and start PCR in the class PCR cyclor.

Expt 2, Day 2 - September 13, 2007

Gel electrophoresis and DNA purification from bands:

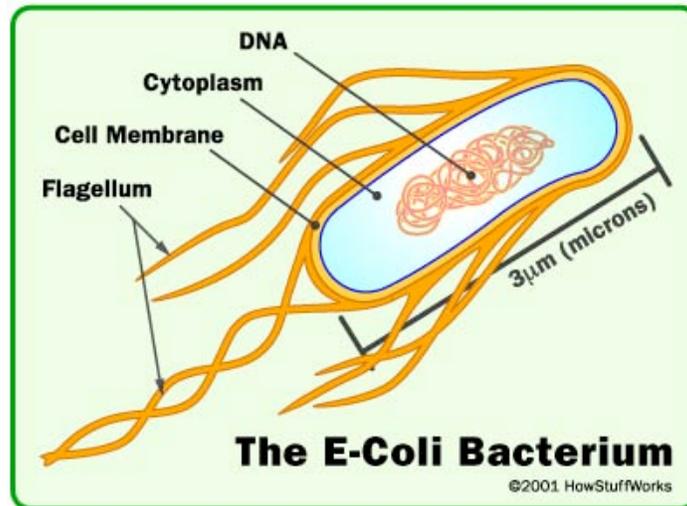
While you were home eating and sleeping, Yujun was busy running your PCR samples on gels and isolating the DNA from the bands. Just to show you that we are not kidding, he even photographed the gels as a record for your notebook. This was done so that you would be ready for the next part of the experiment which is different than expt 1 because the DNAs will not be sent directly to the sequencing facility

Rationale for changing the protocol from here on in....

In Experiment #1 we determined the sequence at the site of mPing excision from the GFP gene. To do this, we amplified this region by PCR, purified the excision band, sent this to the sequencing facility (along with the primer used for the sequencing reaction) and received one sequence back per band.

Using this procedure we would not be able to detect one type of experimental outcome. What if all of the PCR products in a band were not exactly the same? How could such a thing happen? What if the excision of the TE from the GFP gene was not always perfect and a few nucleotides were sometimes "left behind" from the element. Alternatively, what if the element excised and took a small piece of the gene with it or "scrambled" some of the sequences at the excision site? If any or all of these scenarios occurred, our one sequence per band would not be sufficient as your PCR products might be from several different excision events. So... let's say that the PCR band in fact contains many slightly different sequences. To figure out what those sequences are, we have to somehow analyze individual PCR products. Believe it or not, we can do this with the help of our bacteria friend, *E. coli*. Let's see how this is done.

Escherichia coli - the model bacterium

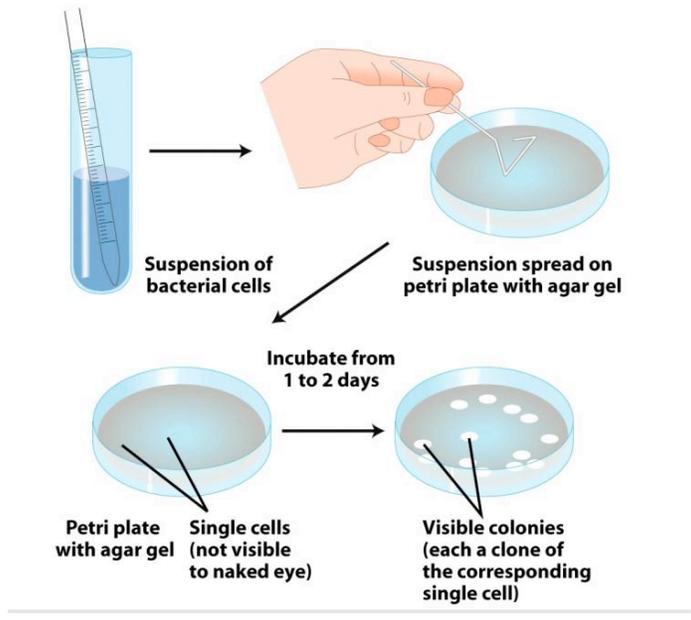
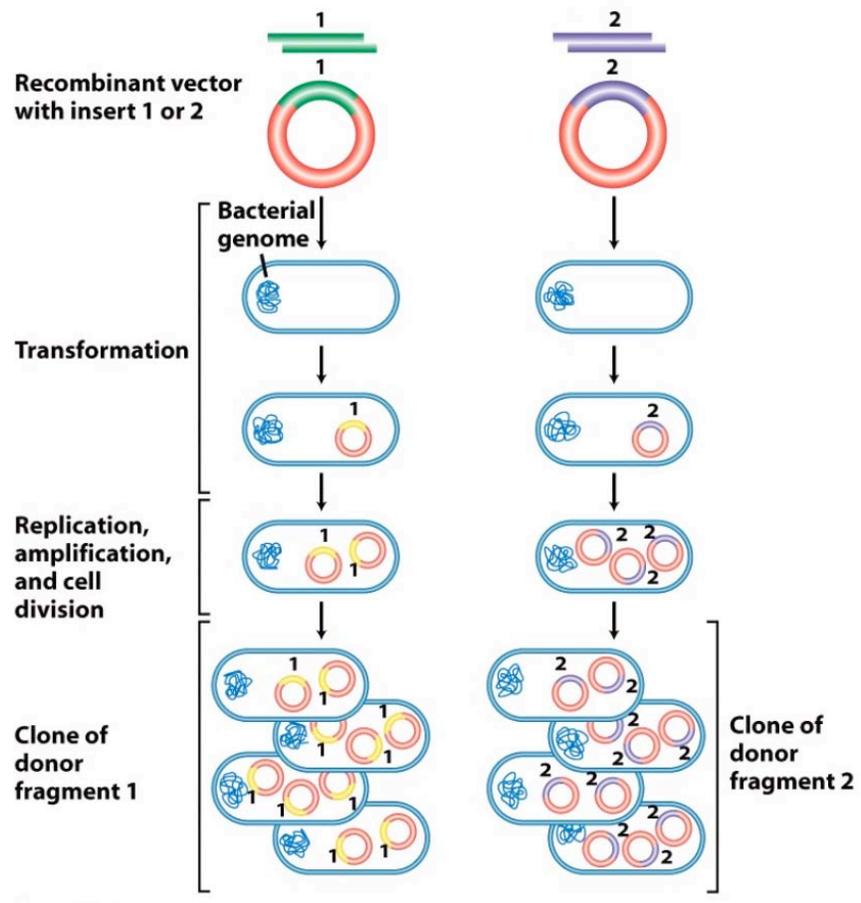


A quick word about *E. coli*. This bacterium is yet another workhorse for molecular biologists, because it has an easy-to-use structure, because it grows rapidly, it accepts foreign DNA, and it "allows" us to make lots of the "foreign" DNA. The strains used in the lab aren't quite like the ones in nature—which are called "wild type." This is a good thing, because wild type *E. coli* can be a nasty bug, causing all kinds of intestinal problems. (It's also a good bug, adapted to live in your digestive tract, and we all have trillions of them, so don't hate them, but don't turn your back on them, either.) Our lab strains still must be handled in a sterile environment, but Yujun will make sure you know the proper procedures, so don't worry! This goes on in labs all over the world every day, and anyway, pathogenic *E. coli* are different from the ones we will be using.

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

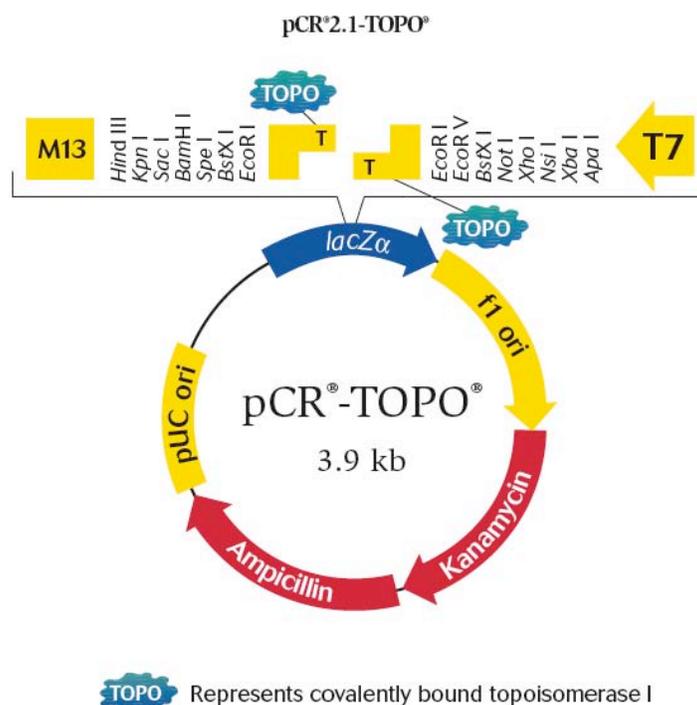
For this experiment, we will be using a patented system called TOPO Cloning from a company called Invitrogen. We will be ligating our PCR products into the TOPO plasmid (vector), transforming these plasmids into *E. coli*, picking *E. coli* colonies, isolating plasmid with our PCR inserts and then sending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

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



The TOPO vector:

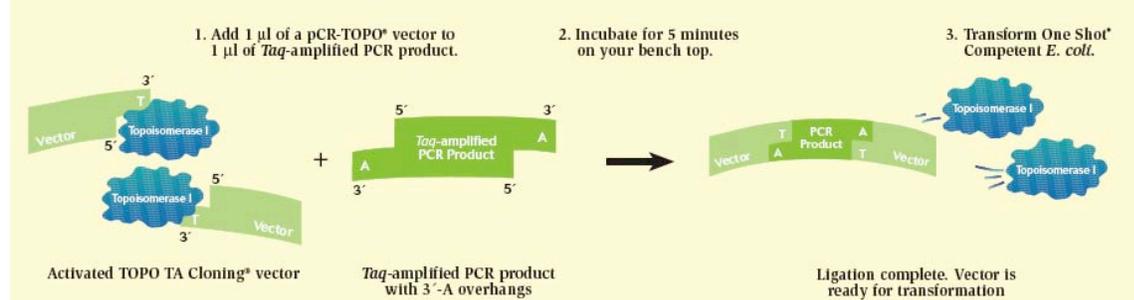
We will be using TOPO to clone the gel bands from the PCR of leaf DNA then we will insert this purified DNA into the TOPO vector, transform *E. coli*, let the transformed bugs grow overnight, purify plasmid from bacterial colonies and send these purified plasmids to the sequencing facility. The idea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



Direct ligation with TA Cloning[®] Technology

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

Figure 6 - How TOPO TA Cloning[®] works



Expt 2, Day 2 protocol (Thursday, Sept 13, 2007)

Materials (TA will have all this!)

PCR2.1 TOPO Vector - Invitrogen

Salt solution

1.5 ml centrifuge tubes

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

SOC medium (store at room temperature)

LB/Carb/X-Gal agar plates

Sterile glass beads

Bacterial waste container

Gloves

B. Protocol (Yujun will do the steps in red)

1. Place tube of Top-10 competent cells and PCR2.1 TOPO vector on ice to thaw
(Yujun will prepare these)

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

3 μ l gel purified PCR product

1 μ l salt solution

1 μ l PCR2.1 TOPO Vector (add last)

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

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

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

5. Incubate the tube on ice for 20 min. While waiting, prepare the LB selective plates, add X-gal and then put plates into 37 degree incubator to warm up. (Only transformed E.coli cells can grow on LB selective plate. Furthermore, if X-gal is

added, cells that have empty vectors will grow into blue colonies and they can be easily discerned from cells with "loaded" vectors, whose colonies would be white.)

6. Incubate in a water bath for 30 sec at 42°C. (This is called the heat shock - it is when DNA is actually taken up into the bacteria from the surrounding liquid)

7. Immediately place cells on ice for 1 min.

8. Add 250 µl SOC medium (keep sterile)

9. Incubate in a 37 °C shaker for 60 min

10. Label the selective plates. Pipet 100 µl of bacterial solution onto one selective plate (work quickly to keep the plates closed as much as possible). Pour 3-5 sterile glass beads onto the plates, cover and shake horizontally to spread the liquid. Dump the glass beads into the bacterial waste container.

11. Incubate the plates **overnight** in an incubator at 37°C
Yujun will take pictures of your plates for your lab notebooks.

Growing up your bacterial colonies

On Friday or sometime over the weekend, Yujun will pick bacterial colonies that have inserts (the white colonies) into test tubes containing liquid medium and grow them overnight as described below. He will also take a picture of each plate for your lab notebook.

Materials

LB/Carb liquid growth medium (for growing bacteria)

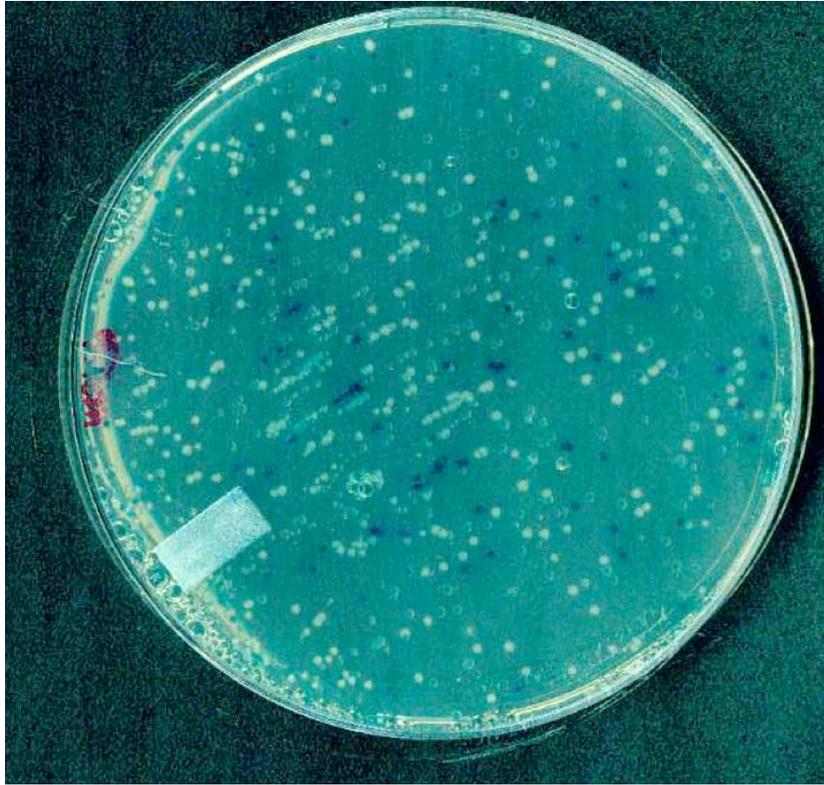
Sterile toothpicks

Protocol

1) Add ~4 ml of liquid growth medium (LB/Carb) into sterile test tubes

2) Using a sterile toothpick, touch a single white colony from the agar plate and drop the toothpick (one per tube) into the test tube.

3) Incubate the test tubes in the air shaker overnight at 37°C.



Wren's transformed *E. coli* (note the blue and white colonies and their relative numbers)

Expt 2. Day 3: Plasmid Purification from Bacteria (Mini-Prep) September 18

Materials

Buffers: P1, P2, N3, PE, and EB

Spin columns

1.5ml centrifuge tubes

agarose gel (+EtBr), TAE buffer, DNA loading buffer

Protocol (modified from the supplier's manual)

1. Transfer 1.0 ml of your *E. coli* sample from the overnight culture to a labeled 1.5ml centrifuge tube (put tip in bacterial waste container after use).

2. Cap and centrifuge for 3 min at 8,000 rpm. Decant (dump) supernatant into the bacterial waste.

3. Add 250 μ l **buffer P1** and vortex to re-suspend the pelleted bacterial cells. No cell clumps should be visible after re-suspension of the pellet.

4. Add 250 μ l **buffer P2** (lysis buffer - NaOH) and gently invert the tube 4-6 times to mix. Do not vortex.

5. Add 350 μ l **buffer N3** (high salt, neutralize) and invert the tube **immediately** but gently 4-6 times. The solution should become cloudy.

6. Centrifuge for 8 min at full speed in our table-top centrifuge. A compact white pellet will form.

7. Pipet ~ 800 μ l of the supernatants (not the white precipitate) from step 4 and apply to a labeled QIAprep spin column.

8. Centrifuge for 30 sec. Discard the flow-through.

9. Add 0.75 ml **PE buffer and centrifuge for 30 sec. Discard the flow-through.**

10. Centrifuge for an additional 1 min to remove residual buffer.

11. Transfer the QIAprep column to a clean labeled 1.5 ml centrifuge tube.

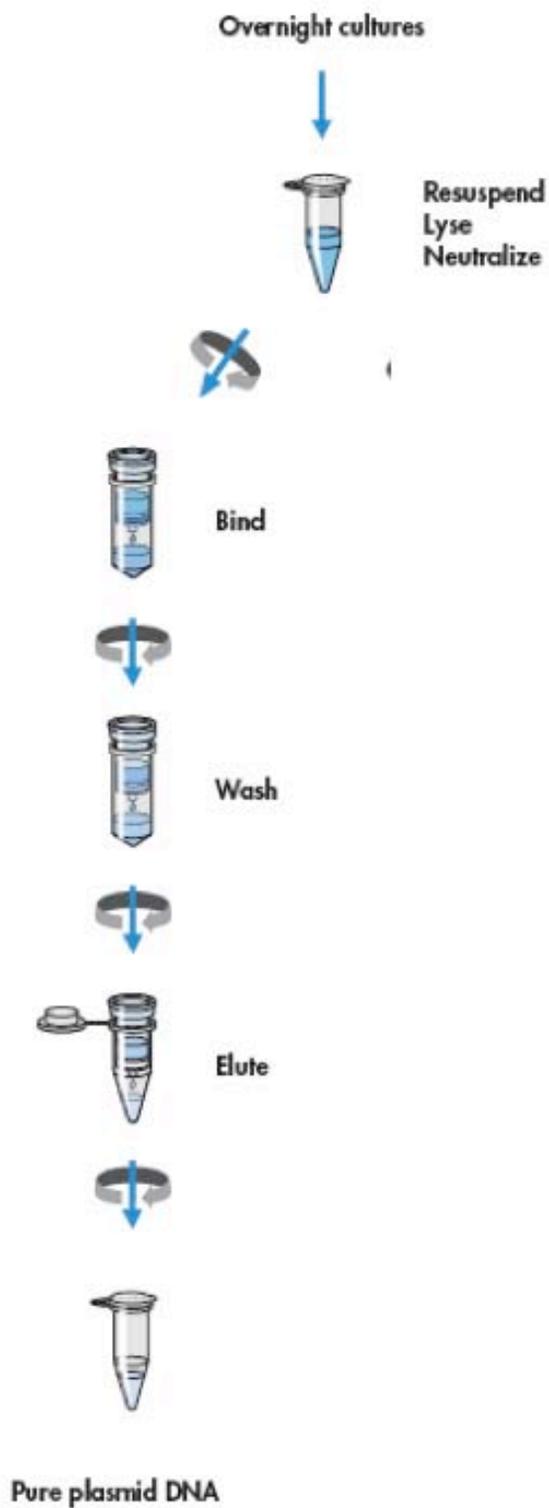
12. Add 100 μ l **EB buffer** to the center of each QIAprep spin column, let stand for 1 min, and centrifuge for 1 min.

13. Discard the column (plasmid DNA will be in the liquid at the bottom of the tube).

14. Run 5 μ l of the purified plasmid (the column flow-through) on an agarose gel to check the quality and quantity.

(This will be completed after class and Yujun will take a picture of the gels for your notebooks)

QIAprep Spin Procedure in microcentrifuges



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

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

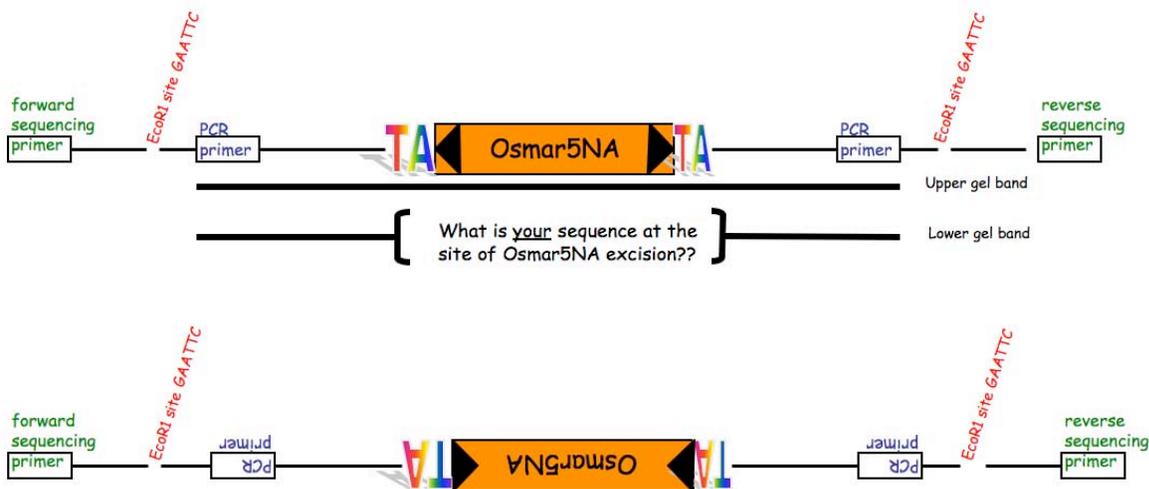
The lysate is neutralized and adjusted to high-salt binding conditions in one step by the addition of **Buffer N3**. The high salt concentration causes denatured proteins, chromosomal DNA, cellular debris, and SDS to precipitate, while the smaller plasmid DNA renatures correctly and stays in solution.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

Experiment 2: Bioinformatics for Thursday, September 20, 2007 (use the accompanying sequences on pages 65-66)

1. Find your samples by looking in the class share folder and the key
2. Open each folder and note good vs. bad sequences.
3. Open your best sequence - forward primer first
4. Cut and paste into word
5. Find and highlight in red the two EcoR1 sites that flank the PCR primers.
6. Find the PCR primers and highlight them in blue
7. Copy the sequences between and including the 2 PCR primers
8. Open Blast site, go to the "Align" program
9. Paste your sequence in the top box
10. Copy the upper band sequence between and including the 2 PCR primers and paste this in the bottom Align box - Hit Align
11. Compare the output and determine what sequences are missing in the lower band sequence. This will tell you what excised.
12. To see how this sequence differs from the original sequence (prior to OsmarNA insertion), clear the bottom window and compare your sequence to the vector sequence before insertion of Osmar5NA. In this way you will see if your excision produced a transposon footprint.
- 13 - To figure out whether your other two PCR products produced a footprint, open each of your two remaining sequence files. You will need to first determine the orientation that you PCR insert was ligated into the TOPO vector. Yujun will help you make this determination.
- 14 - Next, Yujun will show you how to compare the two strands of each of your PCR inserts.

Experiment 2: Information for Analyzing Your Sequences

During the cloning of our PCR products into the TOPO vector, the PCR DNA could be ligated into the vector in either of the two orientations shown above. The orientation must be taken into account when you interpret (annotate) your sequence files.



The EcoRI sites are shown in red bold; PCR primers are shown in blue bold; TA Target site duplication is shown in black bold and BIG; Osmar5 NA sequence is shown in peach bold lowercase; excision "footprints" are shown in purple bold.

```
>OS5_upper_band
...TCGGATCCACTAGTAACGGCCGCCAGTGTGCTGGAATTCGCCCTTCCTCTCCACTGACAGAAA
ATTTGTGCCCATTAACATCACCATCTAATTCAACAAGAATTGGGACAACTCCAGTGAAAAGTTC
TTCTCCTTTACTGAATTCGGCCGAGGATAATGATAGGAGAAGTAAAAGATGAGAAAGAGAAAA
AGATTAGTCTTCATTGTTATATCTCCTTGGATCCTActccctccgtcccacaaaacatgacgtt
ttaagggttagcagccaaaattagctggttggtgcaaaaatgaccaaattgtccccatgatttgatta
agctgtcatttacagcatttgtacatgcatccagattattctagagaagtttctgaaaccaca
gctcagtgccacgtgttaacgaattggcgccttagccacacggttgatacagggcaaaaccatc
attaacatattcaaaaattgaaatcaggtagggaaagattggggatcggcgaagggtggggg
atggagattggggatcggcgtggttgaggacgacggagagcgaaggatgggggacgactagaga
gaggataagatcggagtagtactagcgcaacaaataaaaacgcacttctttttctggttcacc
tccacgtatacggaggggcccaccacttctctctcgcagcacattttttctgggacaatccaggg
gcggtgaaacggcaggttttggtgggacggagggagTAGGATCCTCTAGAGTCCCCCGTGTTC
TCCAAATGAAATGAACTTCTTATATAGAGGAAGGGTCTTGCGAAGGATAGTGGGATTGTGCGT
CATCCCTTACGTCAGTGGAGATATCACATCAATCCACTTGCTTTGAAGACGTGGTTGGAACGTC
TAAGGGCGAATTCTGCAGATATCCATCACACTGGCGG...
```

These are the sequences where Osmar5 NA inserted. The TA dinucleotide that is duplicated upon insertion is **BIG**:

CCTCTCCACTGACAGAAAATTTGTGCCCATTAACATCACCATCTAATTCAACAAGAATTGGGAC
 AACCCAGTGAAAAGTTCTTCTCCTTTACT**GAATTC**GGCCGAGGATAATGATAGGAGAAGTGAA
 AAGATGAGAAAGAGAAAAAGATTAGTCTTCATTGTTATATCTCCTTGGATCC**TA**GGATCCTCTA
 GAGTCCCCCGTGTCTCTCCAAATGAAATGAACTTCCTTATATAGAGGAAGGGTCTTGCGAAGG
 ATAGTGGGATTGTGCGTCATCCCTTACGTCAGTGGAGATATCACATCAATCCACTTG**CTTTGAA**
GACGTGGTTGGAACGTCT

Complement of the above sequence:

AGACGTTCCAACCACGTCTTCAAAGCAAGTGGATTGATGTGATATCTCCACTGACGTAAGGGAT
 GACGCACAATCCACTATCCTTCGCAAGACCCTTCCTCTATATAAGGAAGTTCATTTTCAATTTGG
 AGAGAACACGGGGGACTCTAGAGGATCC**TA**GGATCCAAGGAGATATAACAATGAAGACTAATCT
 TTTTCTCTTTCTCATCTTTTCACTTCTCCTATCATTATCCTCGGCC**GAATTC**AGTAAAGGAGAA
 GAACTTTTCACTGGGGTTGTCCCAATTCTTGTGAATTAGATGGTGTGTTAAT**TGGGCACAAAT**
TTTCTGTCACTGGAGAGG

>OS5_lower_band1

...TCGGATCCACTAGTAACGGCCGCCAGTGTGCTG**GAATTC**GCCCTT
CCTCTCCACTGACAGAAAATTTGTGCCCATTAACATCACCATCTAATTCAACAAGAATTGGGAC
 AACCCAGTGAAAAGTTCTTCTCCTTTACT**GAATTC**GGCCGAGGATAATGATAGGAGAAGTGAA
 AAGATGAGAAAGAGAAAAAGATTAGTCTTCATTGTTATATCTCCTTGGATCC (?????) GGATC
 CTCTAGAGTCCCCCGTGTCTCTCCAAATGAAATGAACTTCCTTATATAGAGGAAGGGTCTTGC
 GAAGGATAGTGGGATTGTGCGTCATCCCTTACGTCAGTGGAGATATCACATCAATCCACTTG**CT**
TTGAAGACGTGGTTGGAACGTCT
 AAGGGC**GAATTC**TGCAGATATCCATCACACTGGCGG...

>OS5_lower_band2: footprint in pink

...TCGGATCCACTAGTAACGGCCGCCAGTGTGCTG**GAATTC**GCCCTT**CCTCTCCACTGACAGAAA**
ATTTGTGCCCATTAACATCACCATCTAATTCAACAAGAATTGGGACAACCTCCAGTGAAAAGTTC
 TTCTCCTTTACT**GAATTC**GGCCGAGGATAATGATAGGAGAAGTGAAAAGATGAGAAAGAGAAAA
 AGATTAGTCTTCATTGTTATATCTCCTTGGATCC**TACAGTA**GGATCCTCTAGAGTCCCCCGTGT
 TCTCTCCAAATGAAATGAACTTCCTTATATAGAGGAAGGGTCTTGCGAAGGATAGTGGGATTGT
 GCGTCATCCCTTACGTCAGTGGAGATATCACATCAATCCACTTG**CTTTGAAGACGTGGTTGGAA**
CGTCTAAAGGGC**GAATTC**TGCAGATATCCATCACACTGGCGG...

THESE IS YOUR FILE KEY

	1	2	3	4	5
A	Y	MG	C	CC	Y
B	Y	N	C	CC	Y
C	Y	N	R	CC	
D	I	N	R	JD	
E	I	E	R	JD	
F	I	E	W	JD	
G	MG	E	W	Y	
H	MG	C	W	Y	

Lab Report #2: Due Thursday, October 4.

As with Lab Report #1, incorporate answers to the following questions into whatever format you choose. Please keep the length under 2000 words.

1. Compare and contrast Osmar 5 and Ping. Discuss how they are similar and how they differ.
2. Discuss the differences in the sources of transposase and nonautonomous elements used in experiments 1 and 2.
3. Describe the different leaf patterns produced by the excision of mPing vs. Osmar5NA and what could be going on at the molecular level to explain the observed differences.
4. Aside from the use of different transgenic Arabidopsis plants, explain the major differences in the experimental protocols for experiments 1 and 2 and why the differences in the excision of mPing and Osmar5NA necessitated these changes. Include in your explanation what a library is, the role played by the TOPO vector and E.coli transformation.
5. Using labeled (annotated) DNA sequences, show your own "DNA footprint" (the sequence at the site of Osmar5NA excision) by showing the sequences before and after excision. Begin and end your sequences with the PCR primers and include about 30bp from the ends of Osmar5NA in the "before excision" sequence.
6. A few of your DNA sequences were "contaminants". Provide an explanation for when contamination could have been introduced and why we were able to determine that these sequences were contaminants.
8. The TOPO vector is a modified version of a naturally occurring plasmid. Describe the features of this vector that were exploited in Experiment 2.
9. You have formulated an hypothesis that if a transposase can bind to the ends of a nonautonomous element, it can catalyze excision of the nonautonomous element. You want to do an experiment to test that hypothesis based on the design of experiments 1 and 2.

Let's assume that the domains that make up proteins are like leggos and that a scientist can assemble any transposase that he/she wants by fusing together different domains in a test tube and then incorporating them into a T-DNA for delivery into a plant. Given this awesome power, design an experiment to test your hypothesis using *Arabidopsis* plants transformed with T-DNA containing all or part of the transposases from Osmar 5, Ping and the nonautonomous elements mPing and Osmar5NA. Of course you will need GFP with something stuck in it. Don't worry about the antibiotic resistance gene! To do this correctly, your transposase must have ALL relevant domains (we discussed this in class and it is in your notes).

Experiment #3: Analyzing all of the TEs in a genome: constructing Phylogenetic Trees

Overview: In this experiment you will begin to think like a genomicist as you learn to apply some of the bioinformatic techniques that are routinely used to compare many DNA sequences. Specifically, we will be "extracting" TE sequences from the rice and maize genome databases and comparing their relationships (within and between species) by building phylogenetic trees. While the genome sequencing projects of human, rice and Arabidopsis are essentially complete, the maize genome sequencing project is underway. As such, you will be among the first to analyze the TE content of maize, the organism where TEs were discovered by Barbara McClintock.

To do this experiment, we will divide the class into groups and each group will focus their analysis on a single TE family:

- Group 1: The CACTA Family - Ian and Martha
- Group 2: The Mutator Family - Renee and Wren
- Group 3: The PIF Family - Jordan and Cathy
- Group 4 - The hAT Family - Caroline and Erin

However, before we begin to explore our own families, Yujun and I will demonstrate how this will be done by using our old friend Osmar - the rice mariner elements.

But first, there is the all-important background section....

Introduction

Evolution through natural selection. It's the bedrock of all modern biology and incredibly beautiful to boot. Transposable elements are all about evolution, and indeed one of the reasons for our bioinformatics treasure hunt is to learn about the evolutionary significance of transposable elements.

When we look at the differences between our Query and Subject sequences, we are looking at variations and, as Dr. Mount tells us, these variations provide "an invaluable source of information for evolutionary biology." (Speaking of variations,

"invaluable" means the same thing as "valuable," just as "inflammable" means the same thing as "flammable." Interesting variation, no?) Dr. M is waving his hand to speak. *Sir?*

"With the wealth of sequence information becoming available, it is possible to track ancient genes, such as ribosomal RNA and some proteins, back through the tree of life and discover new organisms based on their sequence! Diverse genes may follow different evolutionary histories, reflecting the horizontal transfer of genetic material between species. Other types of phylogenetic analyses can be used to identify genes within a family that are related by evolutionary descent, called *orthologs*. Gene duplication events create two copies of a gene, called *paralogs*, and many such events can create a family of genes, each with a slightly altered or possibly new function."

Let's interrupt the Good Doctor for a moment. What he's talking about is the starting point for creating phylogenetic trees. You've seen these, of course. Sometimes we call them evolutionary trees, and if you've ever done any genealogy, you know what one looks like. (Though you wouldn't want anyone to confuse one of your great-grandmothers with Pong.) Go ahead, Dr. Mount:

"Once alignments have been produced and alignment scores found, the most closely related sequence pairs become apparent and may be placed in the outer branches of an evolutionary tree!" Excitable, isn't he? But he should be—this is extremely interesting stuff.

Bottom line: Using this process generates a predicted pattern of evolution for a particular gene.

We appreciate Dr. Mount's help, since he is busy with his primary research interest, applying bioinformatics and genome analysis to cancer research. He leads a computational group at the Arizona Cancer Center that is actively involved in using gene microarray data and other types of data as tools in cancer diagnosis and treatment by trying to predict biochemical and genetic changes that increase cancer risk or that cause cancer progression. See why this stuff is so important?

What Is a Phylogenetic Tree?

So what is a phylogenetic tree? Let's ask Susan Cates, who is a faculty lecturer and graduate laboratory coordinator in the department of biochemistry and cell biology

at Rice University. She writes for Connexions, which is an open-source web site where anyone can view and share educational material made of small knowledge chunks called modules that can be organized as courses, books, reports, and so forth. Anyone may view or contribute. It's at <http://cnx.org/>.

So, tell us, Ms. Cates: What the heck are these things?

"A phylogenetic tree is a graphical representation of the evolutionary relationship between taxonomic groups," she says. "The term *phylogeny* refers to the evolution or historical development of a plant or animal species or even a human tribe or similar group. Taxonomy is the system of classifying plants and animals by grouping them into categories according to their similarities.

"A phylogenetic tree is a specific type of cladogram where the branch lengths are proportional to the predicted or hypothetical evolutionary time between organisms or sequences. Cladograms are branched diagrams, similar in appearance to family trees, which illustrate patterns of relatedness where the branch lengths are not necessarily proportional to the evolutionary time between related organisms or sequences.

"Bioinformaticians produce cladograms representing relationships between sequences, either DNA sequences or amino acid sequences; however, cladograms can rely on many types of data to show the relatedness of species. In addition to sequence homology information, comparative embryology, fossil records, and comparative anatomy are all examples of the types of data used to classify species into phylogenic taxa. So, it is important to understand that the cladograms generated by bioinformatics tools are primarily based on *sequence data alone*. Given that, it is also true that sequence relatedness can be very powerful as a predictor of the relatedness of species."

Thanks to Susan Cates for stopping by! That seems pretty clear, doesn't it?

Figure 25.7 A phylogenetic tree for the Galapagos finches

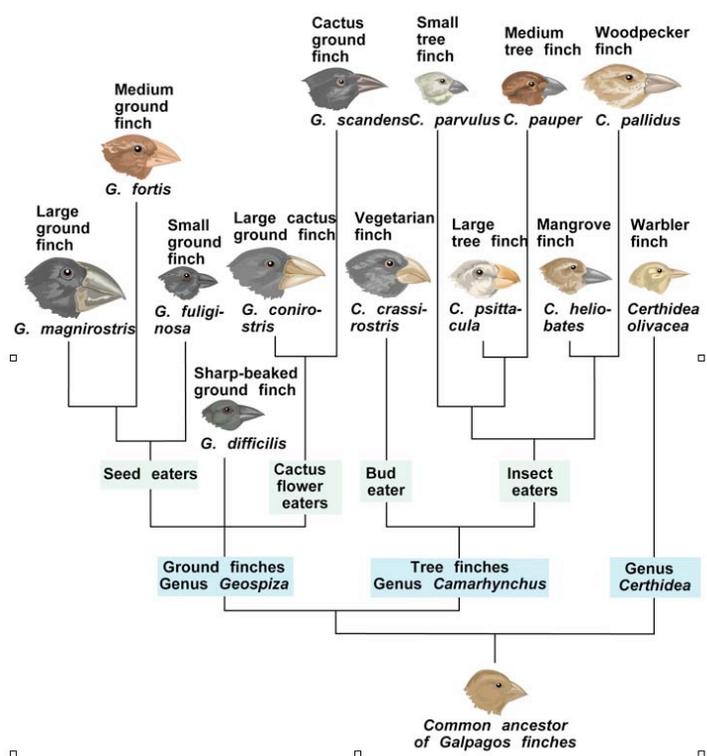


Figure - A tree of the Galapagos finches (from Introduction to Biology - Campbell) The data to construct a tree can be of many sorts. The tree above is based on comparative studies of body structures, especially beak shape and size and extensive field studies of reproductive isolation. Each branch point has meaning. The lowest branch on the right indicates that the warbler finch lineage diverged first. The next branch on the left diverged into ground and tree inhabiting species.

Experiment 3, Day 1, October 11, 2007

Today's objective: We will "mine" all of the osmar elements from the rice genome and use these data to generate a multiple alignment that will then be used to generate a phylogenetic tree. This tree will be a pictorial representation of the "evolutionary relationships" between the Osmar elements.

Step 1: Choose a query sequence (Osmar 5 transposase)

We start with our all important query sequence. In this case we will use part (383 aa) of the amino acid sequence of our old friend Osmar5:

```
SKDLTNIQRRGIYQLLLQKSKDGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGIA
VNVDSRKHGNSGRKKVEIDLSVIAAIPLHQRRNIRSLAQALGVPKSTLHRWFKEGLIRR
HSNSLKPYLKEANKKERLQWCVSMMLDPHTLPNNPKFIEMENIIHIDEKWFNASKKEKTF
YLYPDEEEPYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGKIGIWPFTTRKEPARR
RSRNRERGLVTKPIKVDRDTIRSFMISKVLP AIRACWPREDARKTIWIQQDNARTHLP
DDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRISRNMDELIENV
HKEYRDYNPNTLNRVFLTLQSCYIEVMRA
```

Step 2: Use Blast to identify other sequences related to the sequence of interest and download electronic files of those sequences.

Let's visit ncbi (<http://www.ncbi.nlm.nih.gov/>) again. Choose blast:

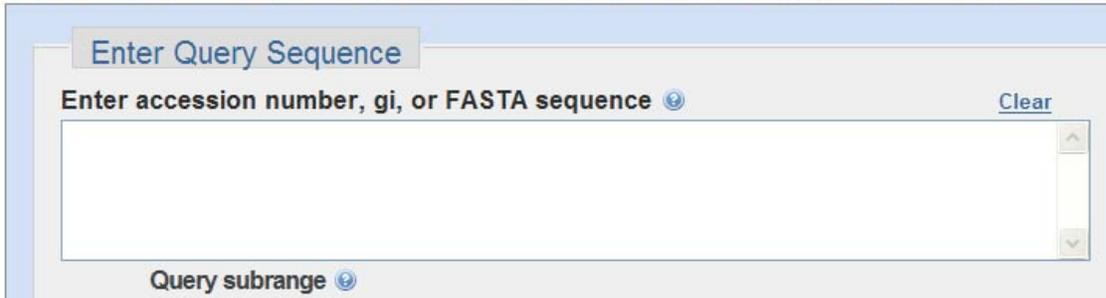
The image shows the NCBI search interface. At the top left is the NCBI logo. To its right is the text 'National Center for Biotechnology Information' with 'National Library of Medicine' and 'National Institutes of Health' below it. Below this is a navigation bar with links for 'PubMed', 'All Databases', 'BLAST', 'OMIM', 'Books', 'TaxBrowser', and 'Structure'. Below the navigation bar is a search box with the text 'Search All Databases' and a dropdown arrow, followed by the word 'for' and an empty search input field, and a 'Go' button.

Because our query is a protein sequence and the genome database is a DNA sequence. We need to select tblastn from the "Basic BLAST" panel:

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
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

Paste the osmar5 amino acid sequence into the window for the query sequence.

► NCBI/ BLAST/ tblastn: TBLASTN search translated nucleotide database using a protein query. [more...](#)

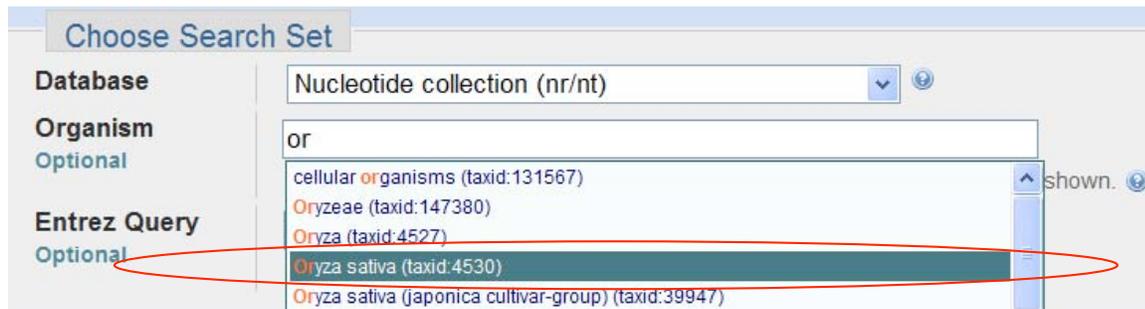


Enter Query Sequence

Enter accession number, gi, or FASTA sequence [Clear](#)

Query subrange [?](#)

Choose search set. It is *Oryza sativa* (taxid: 4530). You can type *Oryza* then select it from the list in the Organism window.



Choose Search Set

Database: Nucleotide collection (nr/nt) [?](#)

Organism: or

Optional

Entrez Query: [?](#)

Optional

- cellular organisms (taxid:131567)
- Oryzeae (taxid:147380)
- Oryza (taxid:4527)
- Oryza sativa (taxid:4530)**
- Oryza sativa (japonica cultivar-group) (taxid:39947)

shown. [?](#)

Then click blast.

Step 3: Fixing a serious problem with the blast output before the data can be used to create trees.

The reason we have a problem is because we used a protein sequence (part of the Osmar 5 transposase) to blast a DNA database (the rice genomic sequence) that was "computationally translated" (tblastn: translated nucleotide database using a protein query). The problem is that some parts of DNA sequences cannot be matched because they do not encode for protein - remember introns?? Introns are present in the genomic DNA and separate protein coding regions (exons).

Let's see how introns screw up our alignments...

```

Query 1          SKDLTNIQRRGIYQLLLQKSKDGGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGIA 60
Sbjct 9600061    SKDLTNIQRR IYQLLL KSKDGGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGI 9600240
                SKDLTNIQRRCIYQLLL*KSKDGGKLEKHTTRLVAQEFHVSIRTVQRIWKRAKICHEQGIT

Query 61         VNVDNRKHGNSGRKKVEIDLVSIAAIPHLQRRNIRSLAQALGVPKSTLHRWFKEGLIRRH 120
Sbjct 9600241    VNVDNRKHGNS RKKVEIDLVSIAAIPHLQR IRSLAQALGV KSTLHRWFKEGLIRRH 9600420
                VNVDNRKHGNSRRKKVEIDLVSIAAIPHLQRSTIRSLAQALGVSKSTLHRWFKEGLIRRH

Query 121        SNSLKPYLKEANKKERLQWCVSMPLDPTLPNNPKFIEMENIIHIDEKWFNASKKEKTFYL 180
Sbjct 9600421    SNSLKPYLKEANKKERLQWCVSMPLDPTLPNNPKFIEMENIIHIDEKWFNASKKEKTFYL 9600600
                SNSLKPYLKEANKKERLQWCVSMPLDPTLPNNPKFIEMENIIHIDEKWFNASKKEKTFYL

Query 181        YPDEEPEYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGKIGIWPFRK 231
Sbjct 9600601    YP E+EPYFTVHNKNAIDKVMFLSAVAKPRY DEGNCTFDGKIGIWPFRK 9600753
                YPGEKEPYFTVHNKNAIDKVMFLSAVAKPRYDDEGNCTFDGKIGIWPFRK

```

Features flanking this part of subject sequence:

[33069 bp at 5' side: Os01q0274800](#)
[11332 bp at 3' side: Os01q0275200](#)

Score = 296 bits (759), Expect(2) = 0.0
 Identities = 146/153 (95%), Positives = 149/153 (97%), Gaps = 0/153 (0%)
 Frame = +3

```

Query 231        KEPARRRSRNRERGLTVTKPIKVDRTIRSFMISKVLP AIRACWPREDARKTIWIQQDNA 290
Sbjct 9600840    QEPARRRSRNRERGLTVTKLIKVDRTIRSFMISKVLP AIRACWPREDARKTIWIQQDNA 9601019
                +EPA+RRSRNRERGLTVTK IKVDRTIRSFMISKVLP AIRACWPREDARKTIWIQQDNA

Query 291        RTHLPIDDAQFGVAVAQSGLDIRLVNQPNSPDMNCLDLGFFASLQSLTHNRISRNMDEL 350
Sbjct 9601020    RTHL IDDAQFGVAVAQ+GLDIRLVNQPNSPDMNCLDLGFFASL SLTHNRISRNMDEL 9601199
                RTHLTIDDAQFGVAVAQTGLDIRLVNQPNSPDMNCLDLGFFASL*SLTHNRISRNMDEL

Query 351        IENVHKEYRDYNPNTLNRVFLTLQSCYIEVMRA 383
Sbjct 9601200    IENVHKEYRDYNPNTLNRVFLTLQ CYIEVMRA 9601298
                IENVHKEYRDYNPNTLNRVFLTLQGCYIEVMRA

```

One way we can fix this is to "manually" remove the gap when we put our sequences into the blast editor (<http://www.wunchiou.com/test/formatblast.html>) that you were introduced to earlier in the course. You will need to first identify a separated alignments, and then paste both alignment paragraphs into the top

window of the blast editor. Then choose subject-only output and click "format the above text".

Query 231 KEPARRRSRNRERGTLVTKPIKVDRDTIRSFMISKVLPPAIRACWPPREDARKTIWIQODNA 290
 +EPARRRSRNRERGTLVTKPIKVDRDTIRSFMISKVLPPAIRACWPPREDARKTIWIQODNA
 Sbjct 26759823 QEPARRRSRNRERGTLVTKPIKVDRDTIRSFMISKVLPPAIRACWPPREDARKTIWIQODNA
 26760002

Query 291 RTHLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRISRNMDEL 350
RTHLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRISRNMDEL

Query base pairs:

Normal format output
 Subject-only output

Output:

```
>AP008207_1
SKDLTNIQRREGIYQLLLQKSKDGKLEKHTTRLVAQEFHVS IHTVQRIWKRKICHEOGIA
VNVDSRKHGNSGRKVEIDLSVIAATPLHQRNRISLAQALGVPKSTLHRWFKEGLIRRH
SNSLKPYLKEANKKERLQWCVSMLDPHTLPNNPKFIEMENIIHIDEKWFNASKKEKTFYL
YPDEEEPYFTVHNKNAIDKVMFLSAVAKPRYDEGNCTFDGKIGIWPFTRK
QEPARRRSRNRERGTLVTKPIKVDRDTIRSFMISKVLPPAIRACWPPREDARKTIWIQODNA
RTHLPIDDAQFGVAVAQSGLDIRLVNQPPNSPDMNCLDLGFFASLQSLTHNRISRNMDEL
IENVHKEYRDNPNTLNRVFLTLQSCYIEVMRA
```

Now let's practice this task by collecting a few more sequences from the blast output and editing them in the blast edit program. What you will see is that this is tedious work? In fact, this step could take WEEKS to perform if we are analyzing large TE families (e.g. with over 1000 members in a genome).

Yujun to the rescue: How can you deal with them? Yujun has written a perl program for you to extract the disrupted "hits" from the blast output. This program automatically joins the separated hits, thus getting rid of introns and filtering out "low quality matches". So now things become easy: you can just copy the whole alignment result into one text file and put it into Yujun's folder in your computer. He will process your data and give you a fasta file that is ready for multiple alignment. The sequences in the fasta file should look like this:

```
>rice_mariner1
QHRRKDMTEEVTKQVYQALLKDNKNGKLGKGDTRRVADQFGVHIRSVQRLWKRGIQLTH
NIPVVVAASHKKGR-SGHKAIPLDLEQLRNIPLKQRMTIEDVSSKLGISSRVQRYLKKGL
LRRHSSSIKPYLTDANKKTRLKWCVDMIDRSLVGDPRFKDFFDYVFIDEKWFYLSQKSEK
YYLLPEEDEPHRTCKNKNYIPRFMFLCVCARPRFRNGECV-FDGKIGCFPLVTYEHAVRS
SQNRLRGELVIKPITSITRDVIRDFMVNKVLPPAIRAKWPREDVNKSIFIQQDNAPSHLKL
DDPDFCEAAREEGFDIRLVCVQPPNSPDFNTLDLGFFRAIQAIQYKKEAKTIKDLVPAV
VLQAFLEYSPWKANRMFVTLQTVLKEAMKVGGNKIKIPHMQKEKLEREDRLPLQISCEA
SLLAECT
```

>rice_mariner2

```
VLDNDKRRAIFDAMLVKARKGYLKGHESKEVSAKFSVPVRTVQRIWKKGKSCLDQGISVD
VASGRSR-CGRKKKVVDVSCLEDISILSRRTTIQDVATQLGVSTSKVYRMKKEGAIKRVSS
SLDPYLTDQNKIDRLKWCIEMLDPRSVPHNLVFKPLDFIFIDEKWFNITRKTVRYAAP
TASRRIRTIQNKNFIPKIMILTALARPRFDSNGNCIFDGGKIGCFPFVITYAAKRSSVNRP
AGTIEMKPIESITKEVIRSFMIKVLPAVRAKWPREDAGKPIYIQQDNARPHIAPDDRMF
CEAAKQDGFNIKLVCPANSPDLNVLDLGGFFNSIQSIQYKASATTTTEELVAIDRAFEDY
PVRLSNRIFLSLHACMREVIEVLGDNSYDLPHIKKGVLERQGRPLQLRCDAKSVNNANN
YL
```

>rice_mariner3

```
LTNPQRRAIYELLLTKSLDGYLEKGSSTRVVAEVENVSIRTVQRIWKRAQLCIAHGQVQIN
DSRKRYNCGRKKVEIDL SVVAAIPLRQRSTIRSLADALGVPKSTLHRLFKEGHLQRHSNS
LKPYLKEANKKERLRWCVGMLDHRTL PNNPKFIEMENIHIHIDEKWFNATKKDKTYYLHPL
EPEPYKTVQNKNAIEQVMFLSAVARPRFDDEGNCTFDGKIGIWPVFV
EPTQRSSRNREGLTVTKTI-KVDRD TMRSMISKVLP AIRACWPQEDARRTIFIQQDNA
RTHVPIDDE*FDVAVGQMGDLIRLVNQPPNSPDMNCLDLGFFASLQSLTSTRVSSNMEEL
IENIHKEYNDYNPNTLNRVFTLQSCYIEVMKASGGNKYKIPHMNERLEALGILPKVLC
CDHQLYERAVQLL
```

.....

Step 3: Using ClustalW to Generate multiple alignments of our osmars from the rice genome

Before creating our trees, we have to use yet another program to make a multiple alignment of the fasta file sequences. This program has the odd-sounding name of ClustalW. Let's move on to it.

1. Click on www.ebi.ac.uk/clustalw/. This takes us to the ClustalW web site.

The screenshot shows the ClustalW web interface. On the left is a navigation menu with categories like 'Help Index', 'General Help', 'Formats', 'Gaps', 'Matrix', 'References', 'ClustalW Help', 'ClustalW FAQ', 'Jalview Help', 'Scores Table', 'Alignment', 'Guide Tree', 'Colours', 'Similar Applications', and 'ClustalW Programmatic Access'. The main content area is titled 'ClustalW' and contains a description of the program and a 'Download Software' link. Below this is a configuration form with the following sections:

- YOUR EMAIL:** A text input field.
- ALIGNMENT TITLE:** A text input field containing 'Sequence'.
- RESULTS:** A dropdown menu set to 'interactive'.
- ALIGNMENT:** A dropdown menu set to 'full'.
- KTUP (WORD SIZE):** A dropdown menu set to 'def'.
- WINDOW LENGTH:** A dropdown menu set to 'def'.
- SCORE TYPE:** A dropdown menu set to 'percent'.
- TOPDIAG:** A dropdown menu set to 'def'.
- PAIRGAP:** A dropdown menu set to 'def'.
- MATRIX:** A dropdown menu set to 'def'.
- GAP OPEN:** A dropdown menu set to 'def'.
- END GAPS:** A dropdown menu set to 'def'.
- GAP EXTENSION:** A dropdown menu set to 'def'.
- GAP DISTANCES:** A dropdown menu set to 'def'.
- OUTPUT:**
 - OUTPUT FORMAT:** A dropdown menu set to 'aln w/numbers'.
 - OUTPUT ORDER:** A dropdown menu set to 'aligned'.
- PHYLOGENETIC TREE:**
 - TREE TYPE:** A dropdown menu set to 'none'.
 - CORRECT DIST.:** A dropdown menu set to 'off'.
 - IGNORE GAPS:** A dropdown menu set to 'off'.

At the bottom, there is a text area labeled 'Enter or Paste a set of Sequences in any supported format:' and a 'Help' button.

ClustalW is a general purpose multiple sequence alignment program for DNA or protein sequences. It “produces biologically meaningful multiple sequence alignments of divergent sequences.” It calculates the best match for the selected sequences, and lines them up so that the identities, similarities, and differences can be seen.

ClustalW is part of The European Bioinformatics Institute (EBI) and is a non-profit academic organization that forms part of the European Molecular Biology Laboratory. The EBI is a center for research and services in bioinformatics and the Institute manages databases of biological data including nucleic acid, protein sequences, and macromolecular structures.

2. Find the box called **Output format**, which is on the fourth line of panels down the page. There are a number of analytical tools at ClustalW, but this is the one we want. Before we do anything else, click on the pull-down panel beneath Output format, and select **gcg MSF**. This format is necessary for PAUP the software that we are going to use to draw the tree.

The screenshot shows the ClustalW web interface with various configuration options. The 'OUTPUT FORMAT' dropdown menu is open, showing the following options: 'aln w/numbers', 'aln w/numbers', 'aln wo/numbers', 'gcg MSF' (highlighted), 'phylip', 'pir', and 'gde'. Other visible options include 'YOUR EMAIL', 'ALIGNMENT TITLE' (Sequence), 'RESULTS' (interactive), 'ALIGNMENT' (full), 'KTUP (WORD SIZE)' (def), 'WINDOW LENGTH' (def), 'SCORE TYPE' (percent), 'TOPDIAG' (def), 'PAIRGAP' (def), 'MATRIX' (def), 'GAP OPEN' (def), 'END GAPS' (def), 'GAP EXTENSION' (def), 'GAP DISTANCES' (def), 'OUTPUT ORDER' (aligned), 'TREE TYPE' (none), 'CORRECT DIST.' (off), and 'IGNORE GAPS' (off). A 'Help' button is located in the bottom right corner.

3. Now, scroll down and paste your subject-only saved sequences into the panel at the bottom called “Enter or Paste a set of Sequences in any supported format.” In the bottom right of this panel, click Run.

Once it is finished let's look and see what it has done....

ClustalW Results

Results of search	
Number of sequences	93
Alignment score	4479167
Sequence format	Pearson
Sequence type	aa
ClustalW version	1.83
JalView	<input type="button" value="Start Jalview"/>
Output file	clustalw-20071011-00063831.output
Alignment file	clustalw-20071011-00063831.aln
Guide tree file	clustalw-20071011-00063831.dnd
Your input file	clustalw-20071011-00063831.input

To save a result file right-click the file link in the above table and choose "Save Target As".
If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets.

Here is what the alignment file contains...

```

Osmar_2      ...LIDEDRQ HVLDACFADS ENLK.LKR.. DTTTIVASLF NIKKSLVQSI
Osmar_39     ...LIDEDRQ HVLDACFADS ENLK.LKR.. DTTTIVASLF NIKKSLVQSI
Osmar_14     ...LIDEDRQ HVLDACFADS ENLK.LKR.. DTTTIVASLF NIKKSLVQSI
Osmar_96     .....VQSI
Osmar_21     .KYLPEAENK AIYGALLAST INGKLVDR.. DTTTIIATMV DVTRRVVQDI
Osmar_52     .KYLPEAENK AIYGALLAST INGKLVDR.. DTTTIIATMV DVTRRVVQDI
Osmar_53     .KYLPEAEKK AIYEALLAST INGKLADR.. DTTTIIATMV DVTRRVVQDI
Osmar_57     ...LPEAEKK AIYGALLAST INGKLADR.. DTTEIIAAMF DVTRRVVQDI
Osmar_74     ...LPEAEKK AIYGALLAST INGKLADR.. DTTEIIAAMF DVTRRVVQDI
Osmar_81     .KYLSDEQRQ DIYEALLAKS INGK.IER.. NATTIVANLF NVRRRVVQDL
Osmar_125    .KYLSDEQRQ DIYEALLAKS INGK.IER.. NATTIVANLF NVRRRVVQDL
Osmar_84     .KYLSDEQRQ DIYEALLAKS INGK.IER.. NAITIVANLF NVRRRVVQDL
Osmar_79     ...LSDEQRE DIYKALLAKS INGK.IER.. NVTAFVANLF NVRRRVVQDL
Osmar_1      .....EVVEMF NVKRGRVQDI
Osmar_22     .....EVVEMF NVKRGRVQDI
Osmar_16     .....KAT DIYIALVGKT TN.ILR...K KATTEVAEMF NVKRARVQDI
Osmar_47     .....KAT DIYIALVGKT TN.ILR...K KATTEVAEMF NVKRARVQDI
Osmar_117    .....IYIALVGKT TNRILR...K KATTEVAEMF NVKRARVQDI
Osmar_131    ....LTNIQR QIYIALVGKT TNGILR...K KATTEVAEMF NVKRARVQDF
Osmar_82     .KNLTNIQRQ EIYDALVKKT INGRLR...K KATTEVAEMF NIKRGRVQDI
Osmar_91     .KNLTNIQRQ EIYDALVKKT INGRLR...K KATTEVAEMF NIKRGRVQDI
Osmar_37     .KNLTNIQRQ EIYDALVKKT INGRLR...K KATTEVAEMF NIKRGRVQDI
Osmar_18     .....EVTAMF NVKRARVOAI

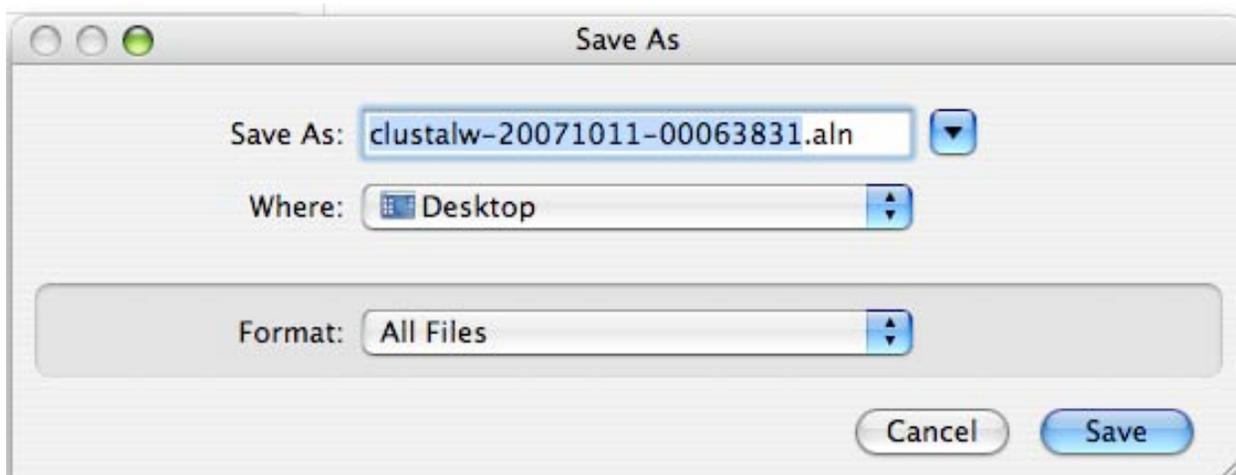
```

```

Osmar_2      IASELHVSKS TMHRSPFKD.. ..GKLCQESN TIKPLLKDE. ..NKEGHVRF
Osmar_39     IASELHVSKS TMHRSPFKD.. ..GKLCQESN TIKPLLKDE. ..NKEGHVRF
Osmar_14     IASELHVSKS TMHRSPFKD.. ..GKLCQESN TIKPLLKDE. ..NKEGHVRF
Osmar_96     IAQELYCSSS TGHRFFQE.. ..GKIRRRSN TVKPYLRDE. ..NKKARIQF
Osmar_21     LASALNASKS KVHRLVKE.. ..GALRRH.N SIKPYLKEA. ..NKKQRLKF
Osmar_52     LASALNASKS KVHRLVKE.. ..GALRRH.N SIKPYLKEA. ..NKKQRLKF
Osmar_53     LASALNASKS KVHRLVKE.. ..GALHRHSN RIKPYLKEA. ..NKKQRLKF
Osmar_57     LASALNASKS QVHRLVKE.. ..GALCRHSN SIKPYLKEA. ..NKKQRLKF
Osmar_74     LASALNASKS QVHRLVKE.. ..GALRRHSN SIKPYLKEA. ..NKKQRLKF
Osmar_81     LASALNVPKS TVHRAFKE.. ..GILLRHSN TLKPFLKDA. ..NKKLCLQF
Osmar_125    LASALNVPKS TVHRAFKE.. ..GILLRHSN TLKPFLKDA. ..NKKLCLQF
Osmar_84     LASALNVPKS TVHRAFKE.. ..GILLRHSK TLKPFLKDA. ..NKKLCLQF
Osmar_79     LASALNIPKS VVHRAFKE.. ..GILRRHSN TLKPFLKDA. ..NKKCRLQF
Osmar_1      FASALGIPKS TLHRMLKS.. ..GMLRRHSN TLKPLLKEE. ..NKK SRL.W
Osmar_22     FASALGIPKS TLHRMLKS.. ..GMLRRHSN TLKPLLKEE. ..NKK SRLQW
Osmar_16     FTSAGVPKS TLHRMLKE.. ..RILRRHSN TLKPLLKEE. ..NKR SRLRW
Osmar_47     FTSAGVPKS TLHRMLKE.. ..RILRRHSN TLKPLLKEE. ..NKR SRLRW
Osmar_117    FASAAGVPKS TLHRMLKE.. ..GILRRHSN TLKPLLKEE. ..NKR SRLRW
Osmar_131    FASAAGVPKS TLHRMLK.. ..G.LRRHSN TLKPLLKEE. ..NKR SRLRW
Osmar_82     FASAAGIPKS TLHRMLKE.. ..GLIRRHSN TLKPLLKEE. ..NKR SRLQW
Osmar_91     FASAAGIPKS TLHRMLKE.. ..GLIRRHSN TLKPLLKEE. ..NKR SRLQW
Osmar_37     FASAAGIPKS TLHRMLKE.. ..GLIRRHSN TLKPLLKEE. ..NKR SRLQW
Osmar_18     FASASGIPKS MLHRMLKE.. ..VLLRCHSN TLKPLLKEE. ..NKR SRLRW

```

Now let's save the alignment file. This is the data that will be used to make a tree...

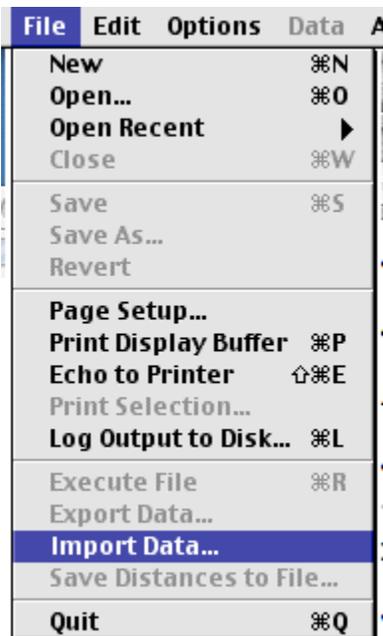


You're on your way to building a phylogenetic tree!

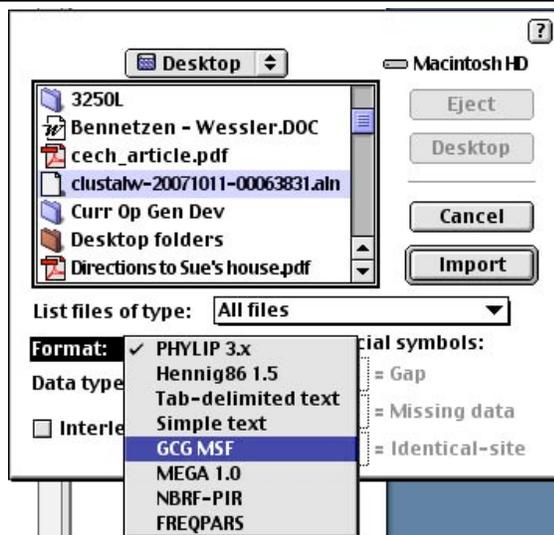
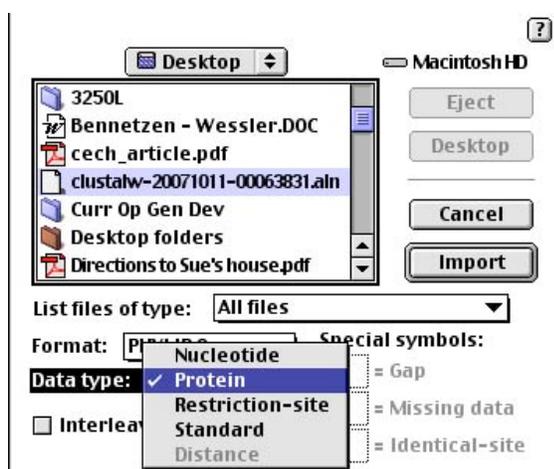
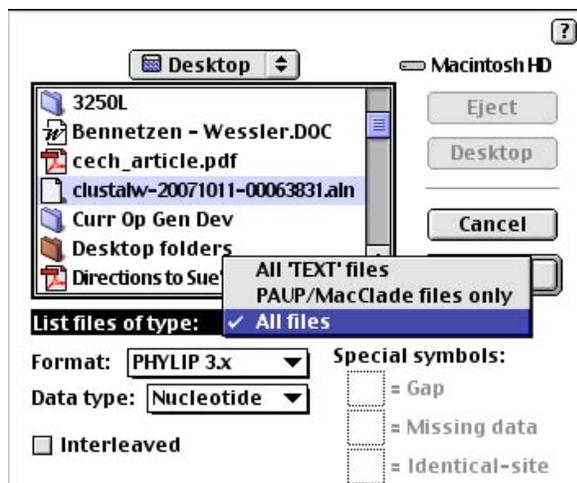
Step 4: Using PAUP to generate a tree with your mutiple alignment data -

Now that we have alignment information, we must access yet *another* program. This one is called **PAUP**, which is an acronym for Phylogenetic Analysis Using Parsimony. It is the most widely used software package for building evolutionary trees. (In addition, the PAUP manual, which goes with the site, serves as a comprehensive introduction to phylogenetic analysis for beginning researchers, as well as an important reference for experts in the field.)

Click on PAUP, which is also already in your computer's Dock or bookmarks. Select Import Data from File.

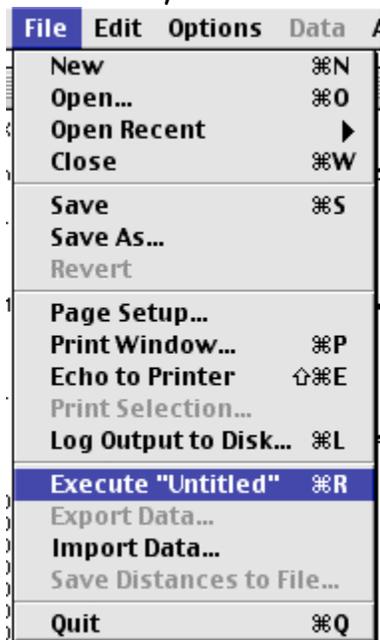


Select the file on the desktop that you just saved. You need to make the following choices in the window....



NOW CLICK IMPORT!

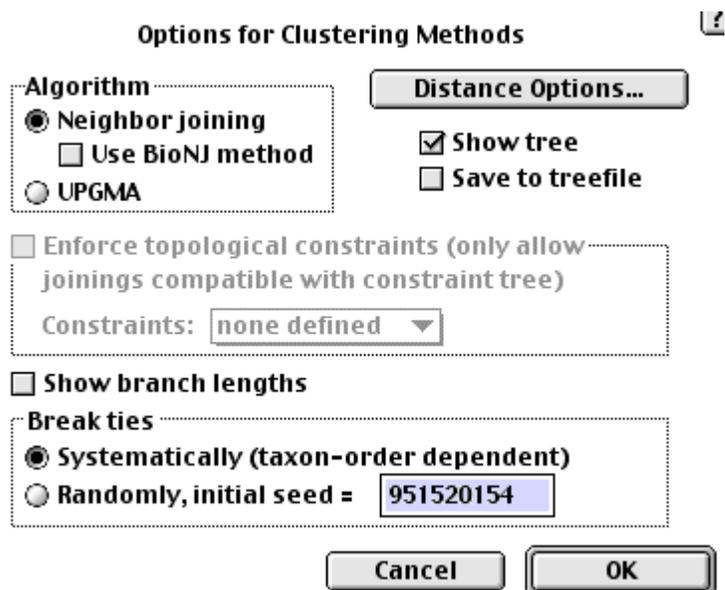
Now we can let PAUP draw the tree. Click "Execute Untitled" from "File". There will be a window pop up asking you whether to give it a name and save it. You can name it as what you like.



Now select "Neighbor Joining/UPGMA" in "Analysis" menu.



Click OK in the new window. Congratulations! You have your tree!

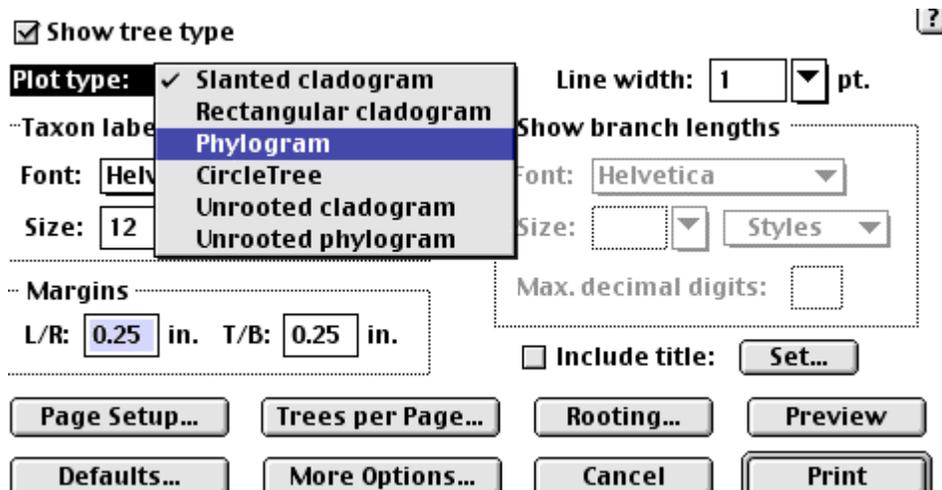


Step 5: Present the tree you produced to an intended audience.

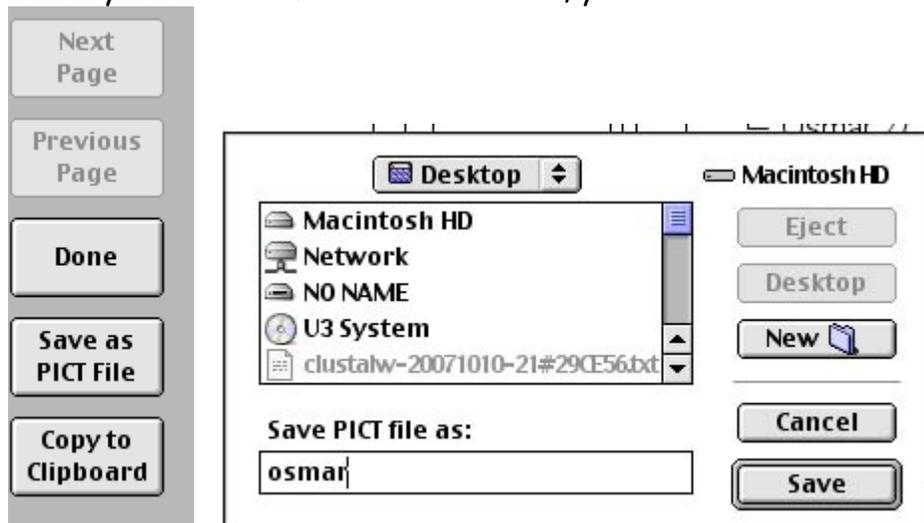
To save or print the tree, you can select "Print NJ tree..." in "Trees" menu.



Change the plot type to Phylogram. You can print it now. If you want to save it, click "preview".



When you are satisfied with the tree, you can click "save as PICT file" to save it.



Experiment #3 Day 2 - October 16, 2007
Mining your TE families from the rice and maize genomes
and generating phylogenetic trees

1. Copy your group's query sequence from Class share - Han.
2. Go to BLAST site.
 - Select tblastn,
 - Oryza sativa database
 - uncheck low complexity filterBLAST
3. To save your BLAST output file -
 - go to FIREFOX menu on top - "File" to "Save page as..."
 - use format - Web Page, complete
 - the file name should include the organism (Os), the element (e.g. CACTA) and .html, save to desktop
4. Put file into the class share folder of Han
5. There will be about 30 minutes for Yujun to work his magic on these files.

During this time we will discuss what you found out about the superfamilies PIF/Harbinger (Pong), CACTA (Spm/En), hAT (Ac) and Mutator (Robertson's).

6. When Yujun has finished processing your BLAST output, you can get your file from Class Share - Han - in a folder called rice TEs (output will be named "your file.aa").
7. Add your outgroup sequence manually to your ".aa" file in a Clustalw window. To do this...
 - First go to your original query doc (from #1 above) and copy your outgroup sequence
 - Open ClustalW and paste the outgroup sequence into the window
 - open your .aa file by dragging the icon onto Word in your dock
 - copy all and paste into the ClustalW window under your outgroup file
8. Before running ClustalW -

- change "output format" to gcgMSF
- click Run

9. ClustalW Results window

- go to "align file" and control/click to get a pop up menu
- select "Save link as..." to your desktop

10. close ClustalW

11. Open PAUP - immediately select "cancel" on pop up menu

12. Select from File menu - Import data

Make the following changes in the default menu

1. All text files to All files
2. Datatype from nucleotide to protein
3. Format from PHYLIP to gcgMF

Then select "your.aln" file from the Desktop

Click Import

13. Select from File menu - Execute Untitled

14. Save changed to untitled - Save

15. Save as - Oz_TE name_tree (this will be the PAUP input data file)

16. Select from "Analysis" menu - Distance ... then Neighbor Joining

Menu pops up - just click OK

17. Congrats - your "unrooted" tree should be displayed.

Print this tree - Go to "Trees" in menu - select print - select NJ tree

Change "plot type" to phologram - click preview

Print your unrooted tree.

18. Rooting your tree...

Go to "Trees" - select root trees

Window - click "rooting options"

Pick - Define outgroup

Window - Ingroup taxa - select your outgroup name

To outgroup - OK, then OK again, then Root!

19. Print out your rooted tree

Go to "trees" again - select print, NJ tree

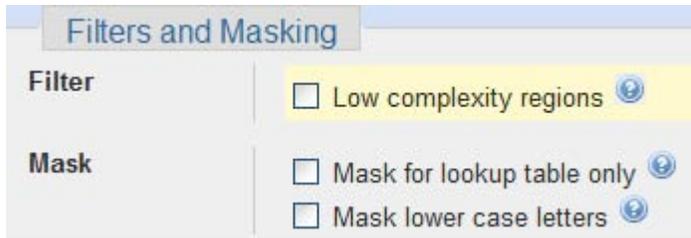
Preview, Print etc.

Now, you are going to mine data from the *Zea mays* genome database and use this to generate a maize tree...

1. Copy your group's query sequence from Class share - Han.
2. Changes in protocol in order to "mine" maize genome sequence...
 - select HTGS instead of nr/nt. The maize genome hasn't been fully sequenced and assembled. Most of the maize data is in the HTGS database. If you choose nr/nt as sbjct database, you will get much fewer hits (try this if you like).

After choosing HTGS, input "Zea" and select *Zea mays* as following:

Finally, turn off the filter for "low complexity regions" in hidden menu of "Algorithm parameters".



Now, go back to step 3 on page 85 and follow the same protocol. When you save your files, make sure to replace the "Os" with "Zm".

Experiment 3, Day 3: Combining the rice and maize output files and generating a single tree for your element family. October 18, 2007.

How to draw both maize TEs and rice TEs together on the same tree?

This is quite easy. You just need to combine the two outputs of maize and rice that you received from Yujun's program. To do this just use Word to open the two files separately and copy them into a third (combined) file. Don't forget to add the outgroup sequence if you want to draw a rooted tree.² Go to BLAST site.

- Select tblastn,
 - Oryza sativa database
 - uncheck low complexity filter
- BLAST

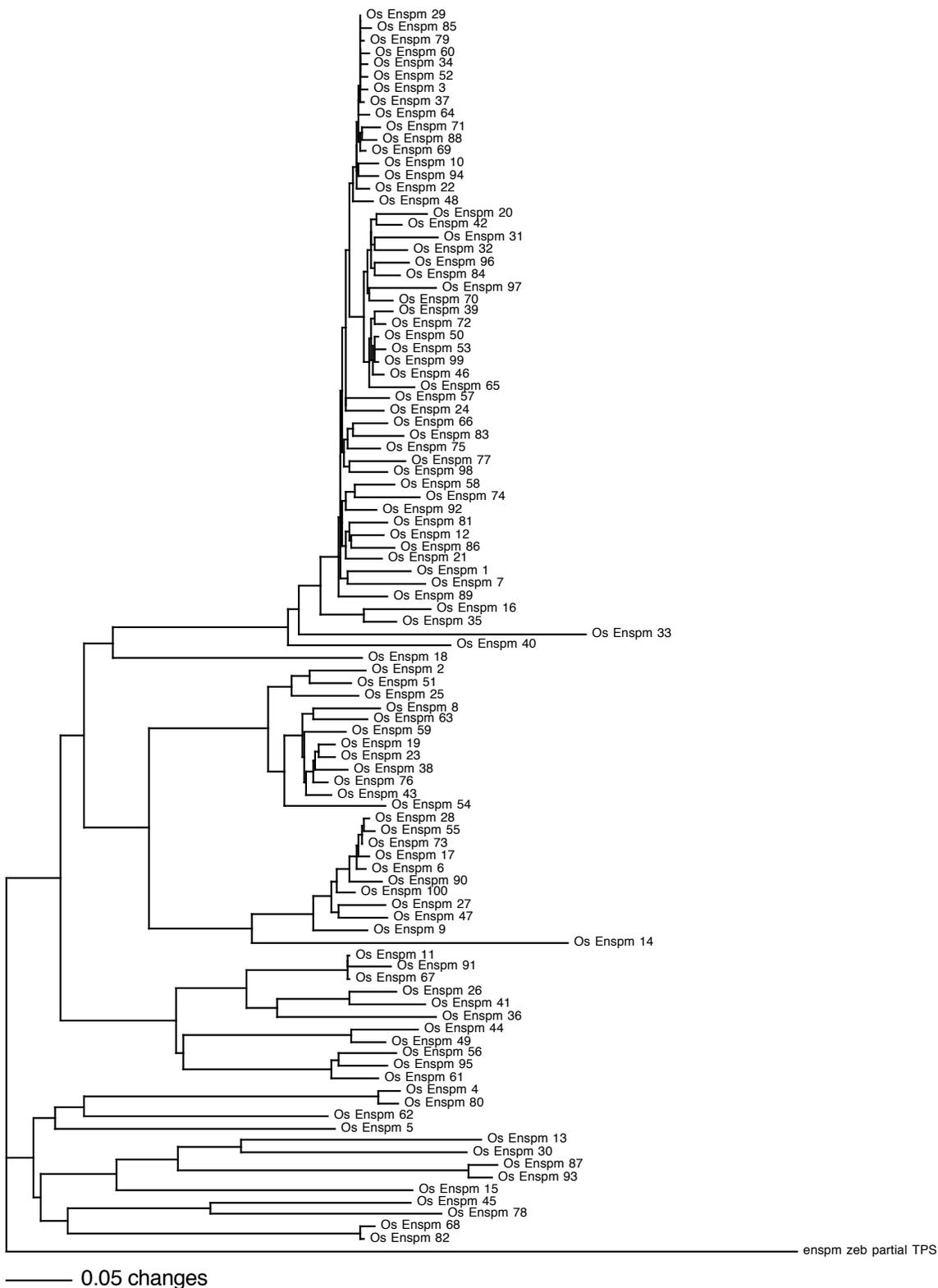
NJ



— 0.05 changes

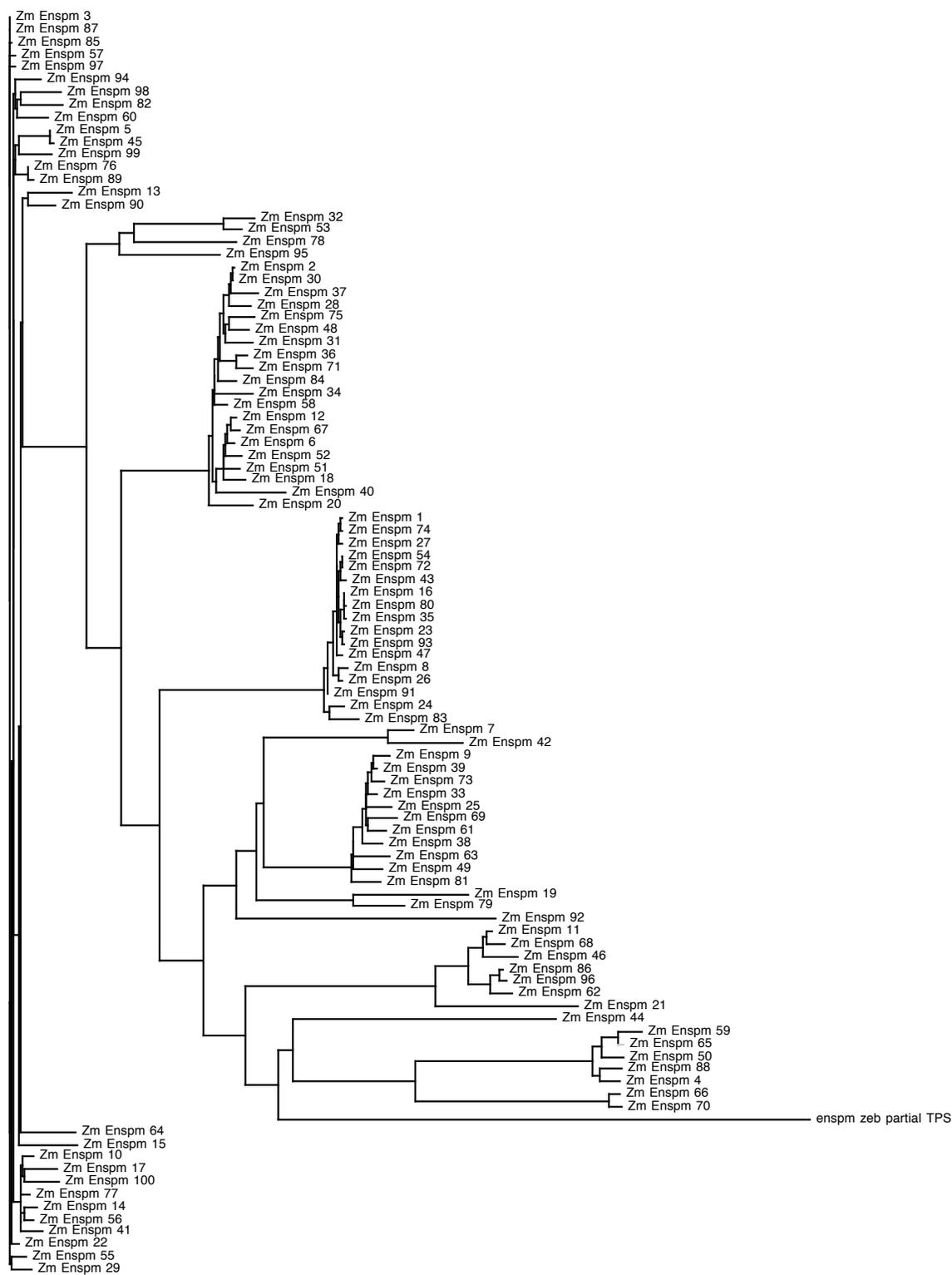
Os CACTA (unrooted)

NJ



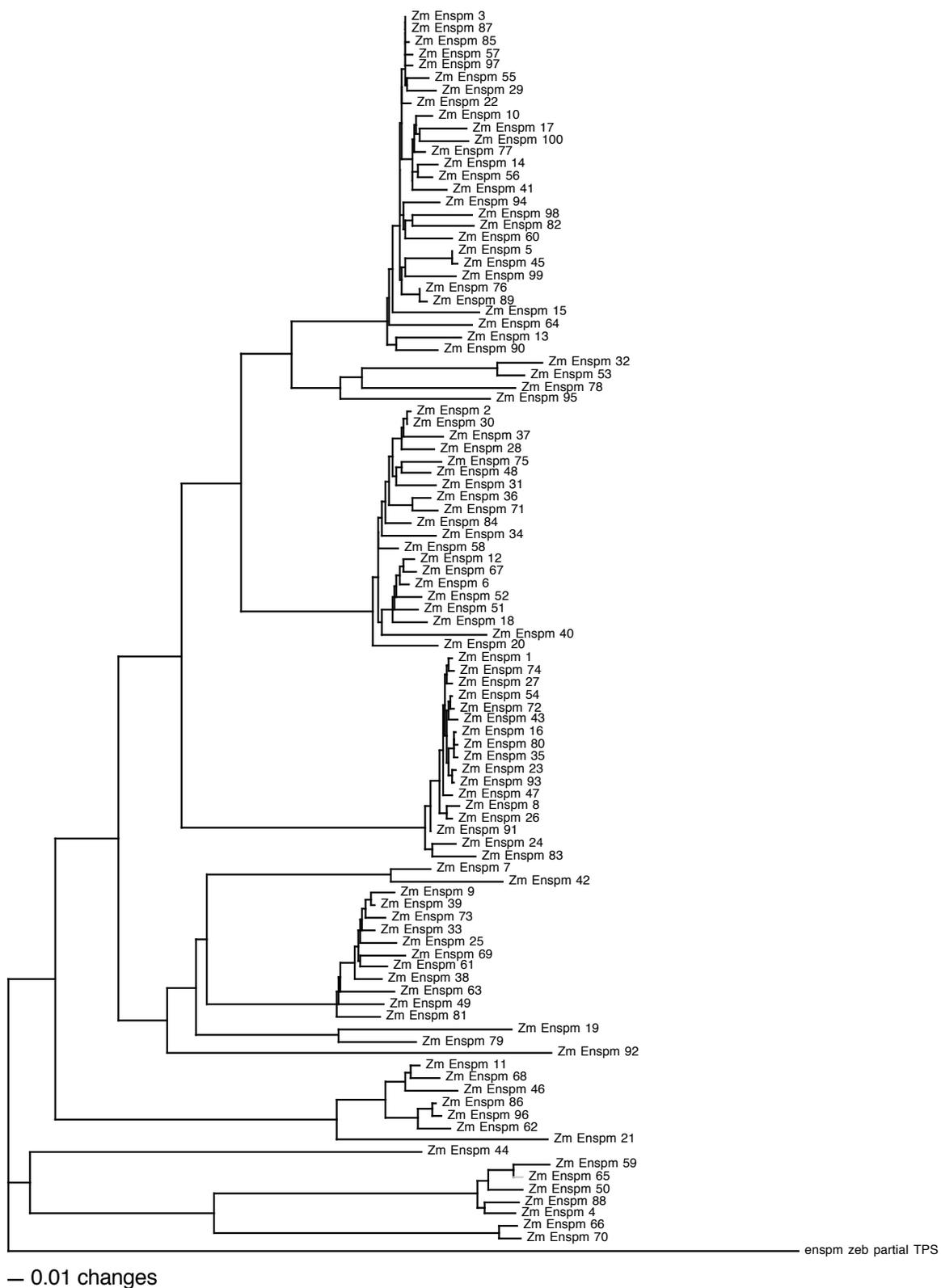
Os CACTA (rooted)

NJ



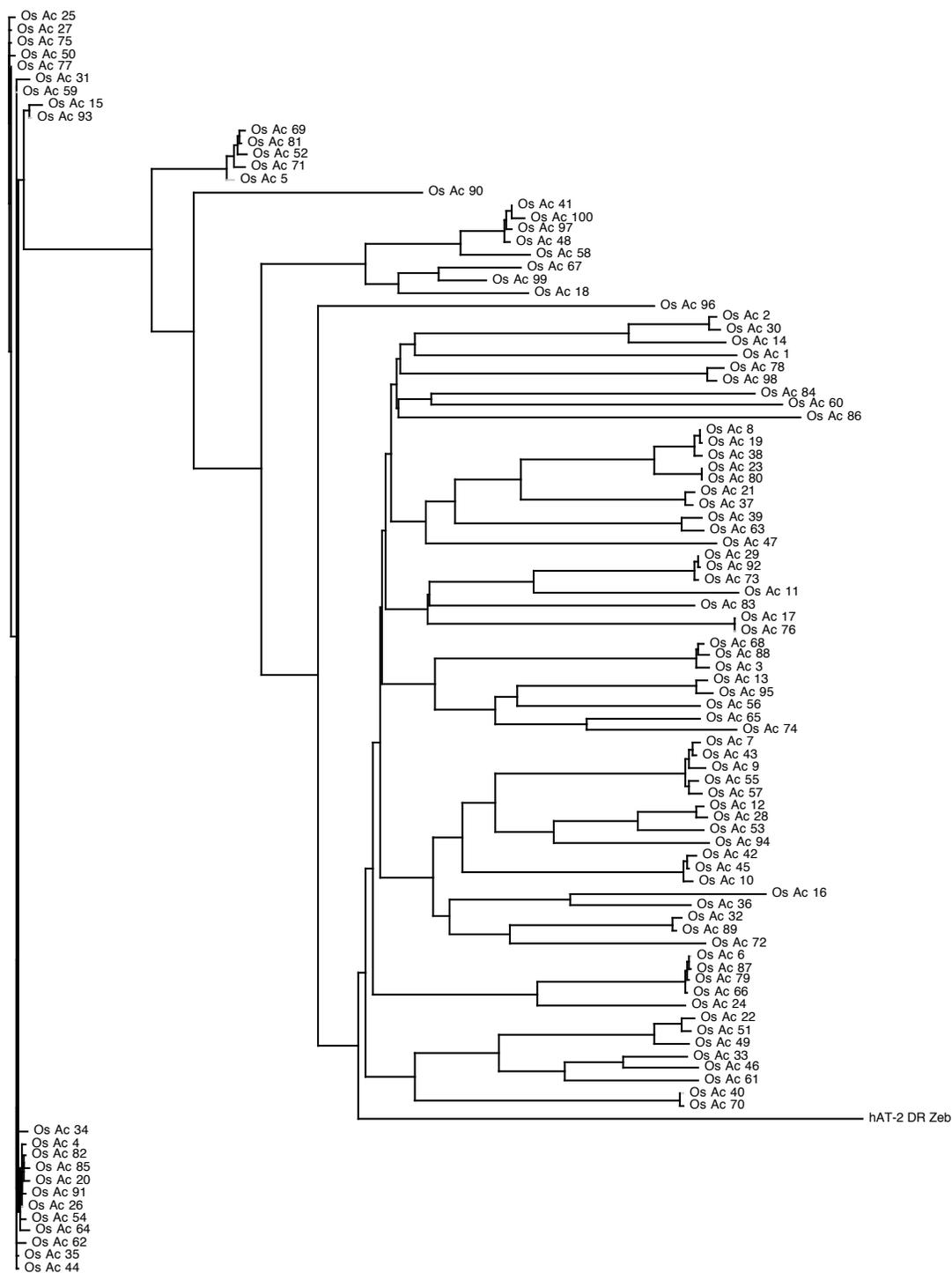
Zm CACTA (unrooted)

NJ



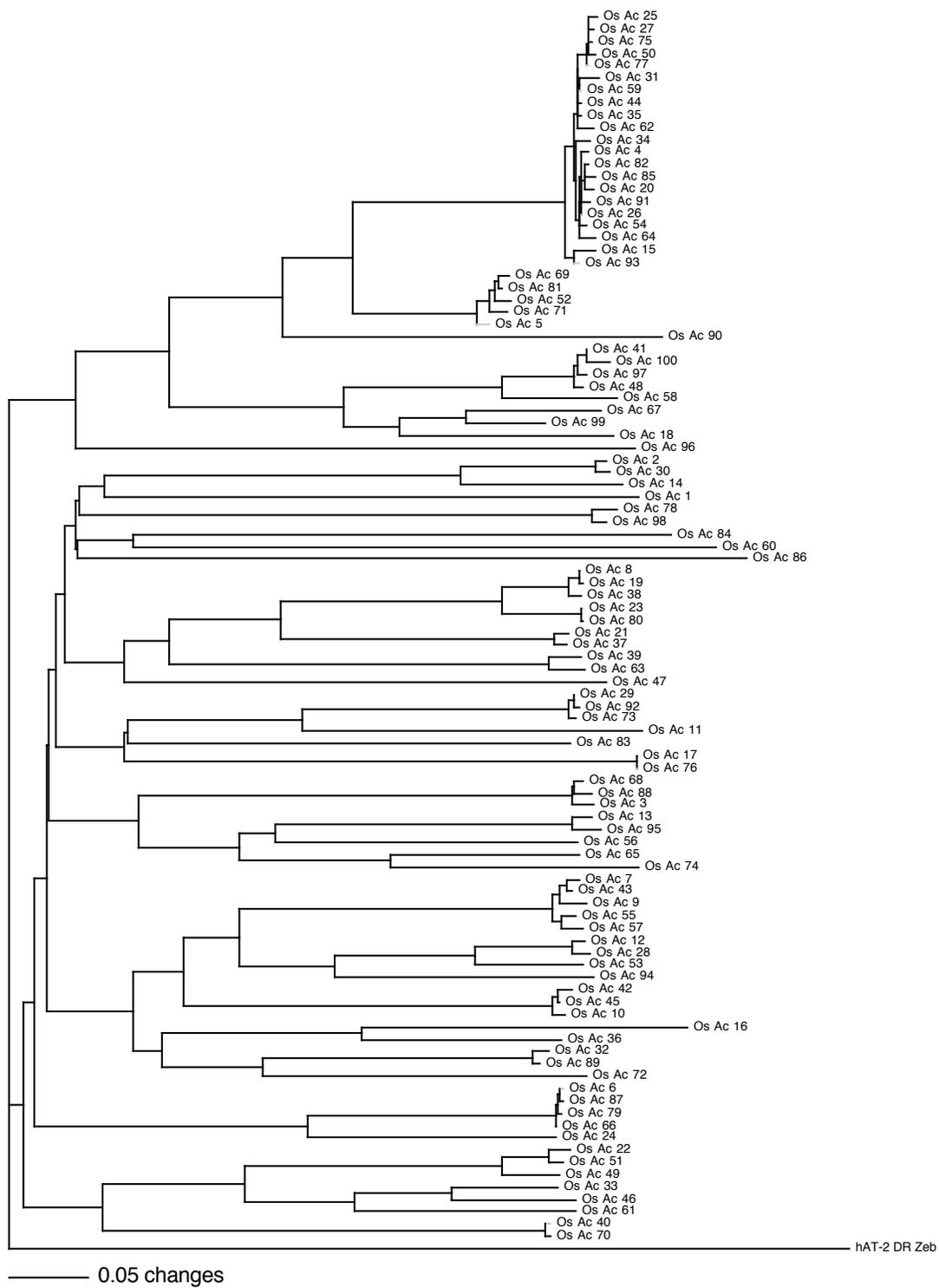
Zm CACTA (rooted)

NJ

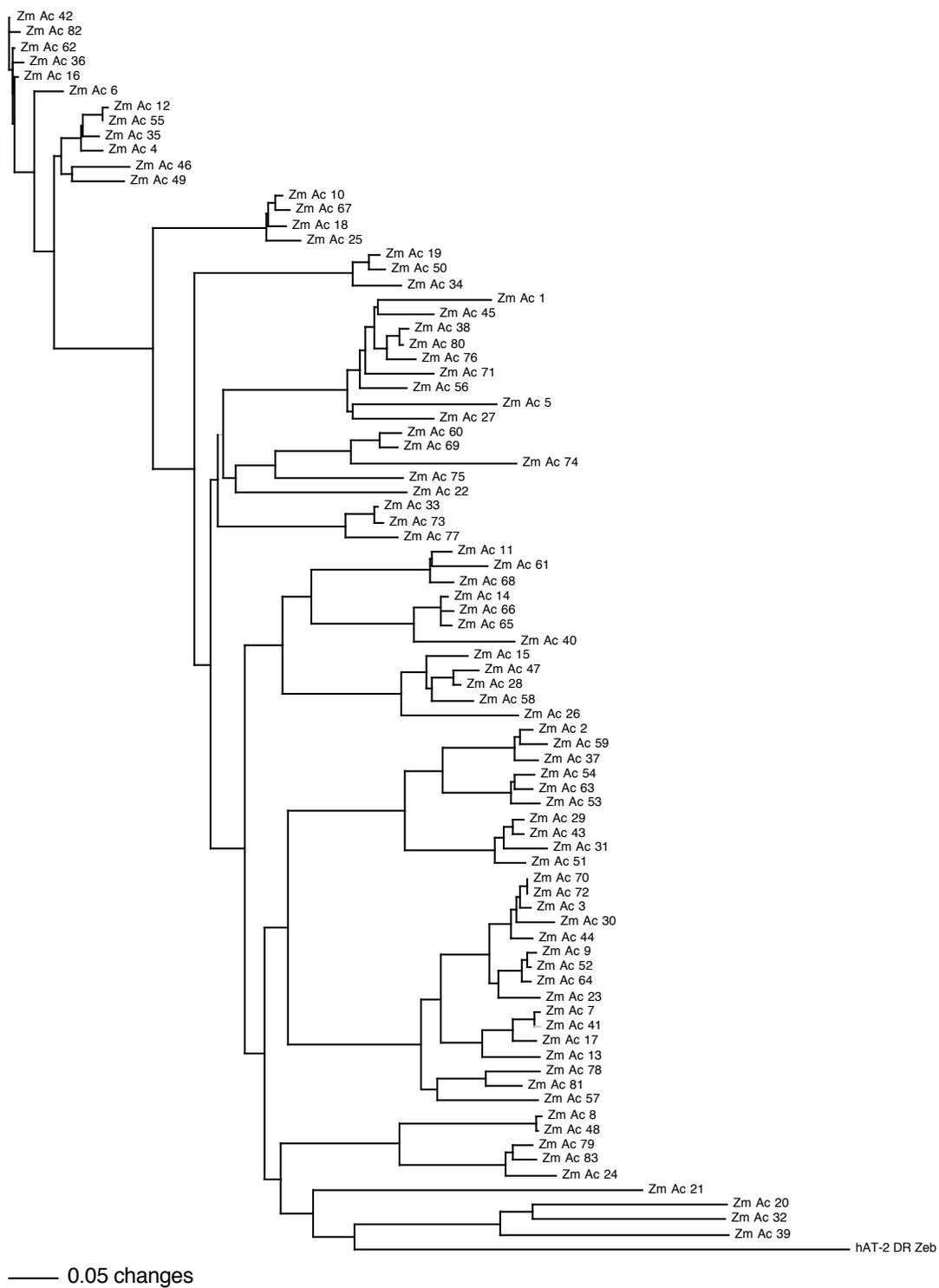


Os hAT (unrooted)

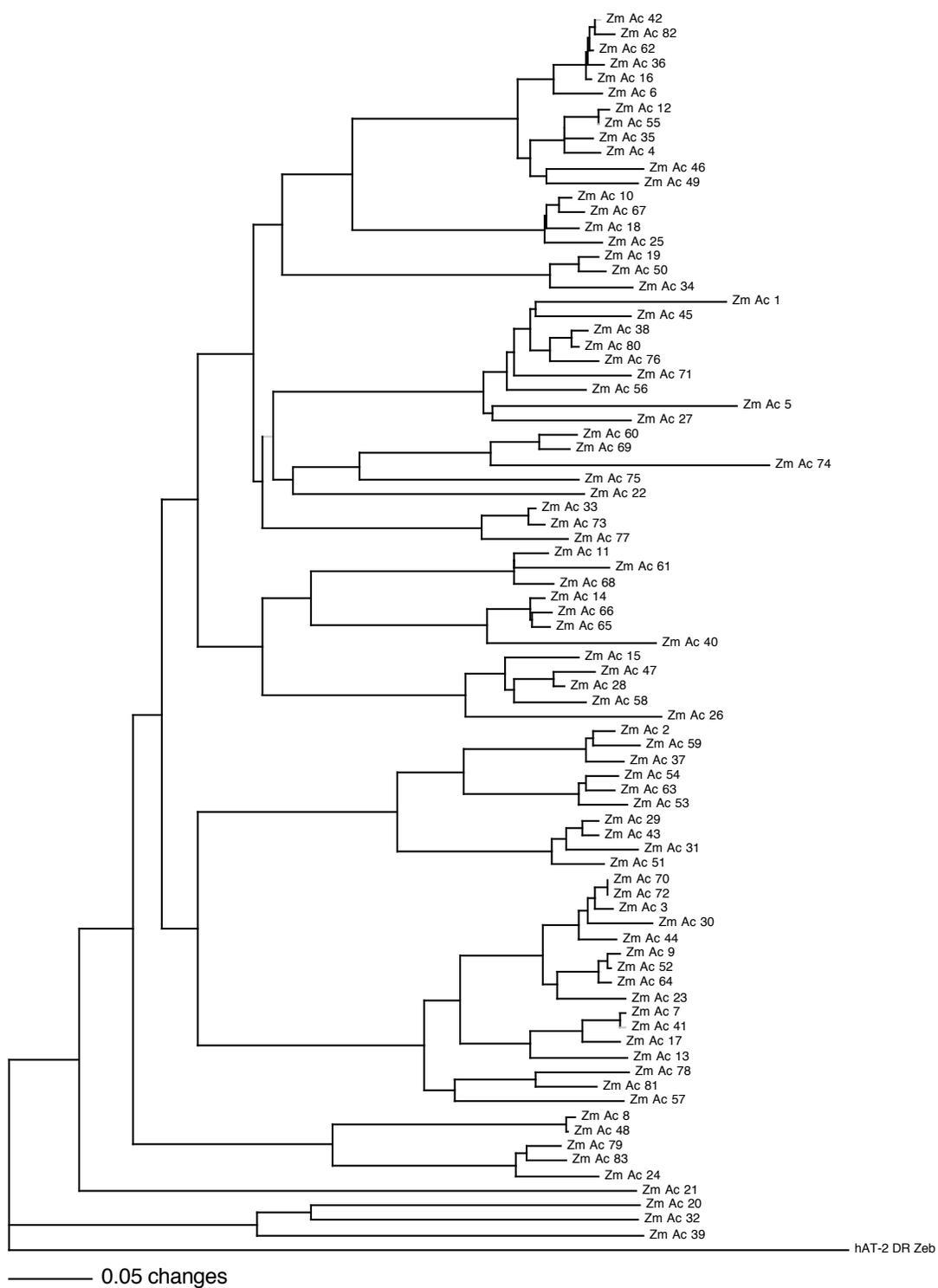
NJ



Os hAT (rooted)

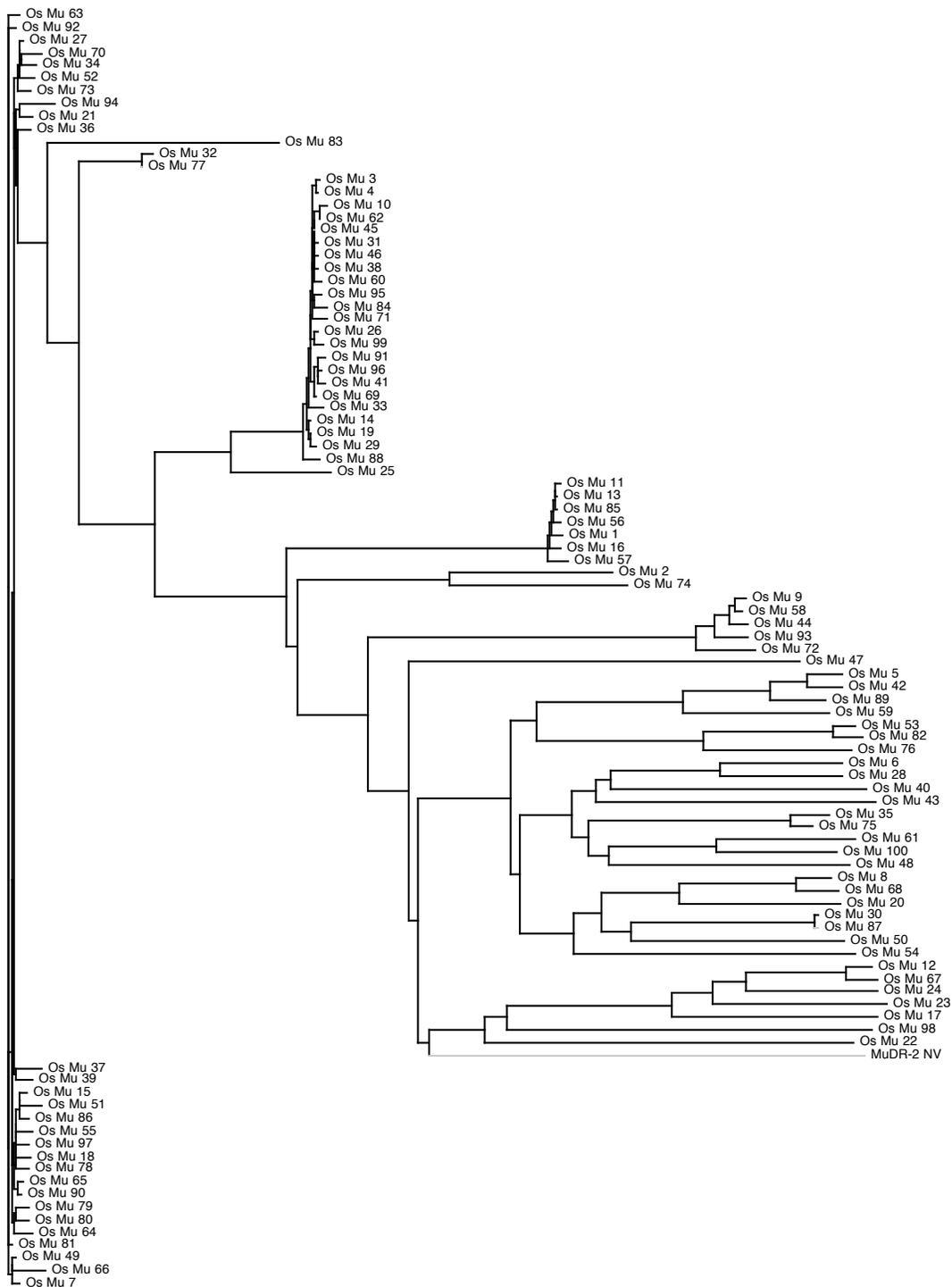


Zm hAT (unrooted)



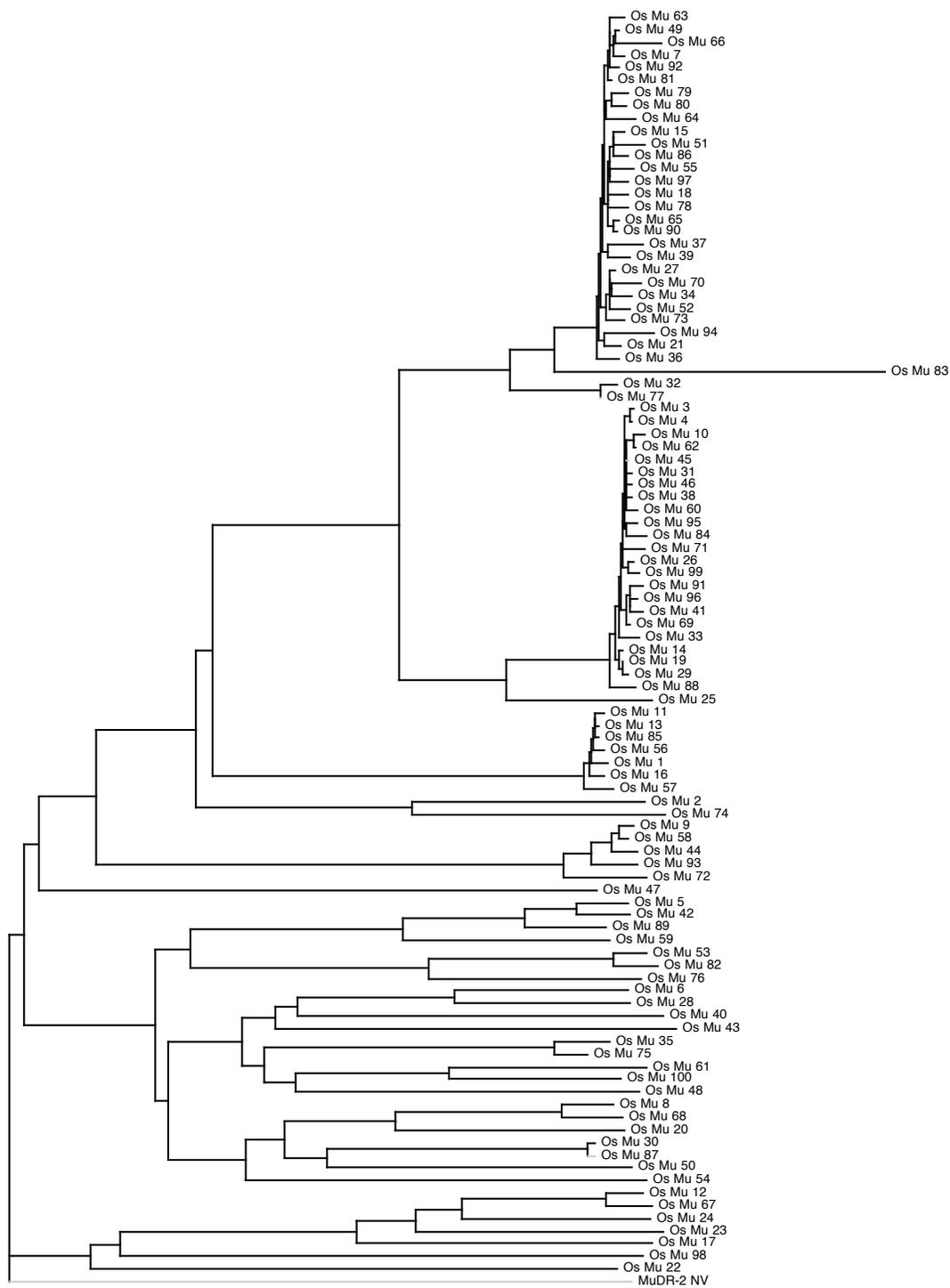
Zm hAT (rooted)

NJ

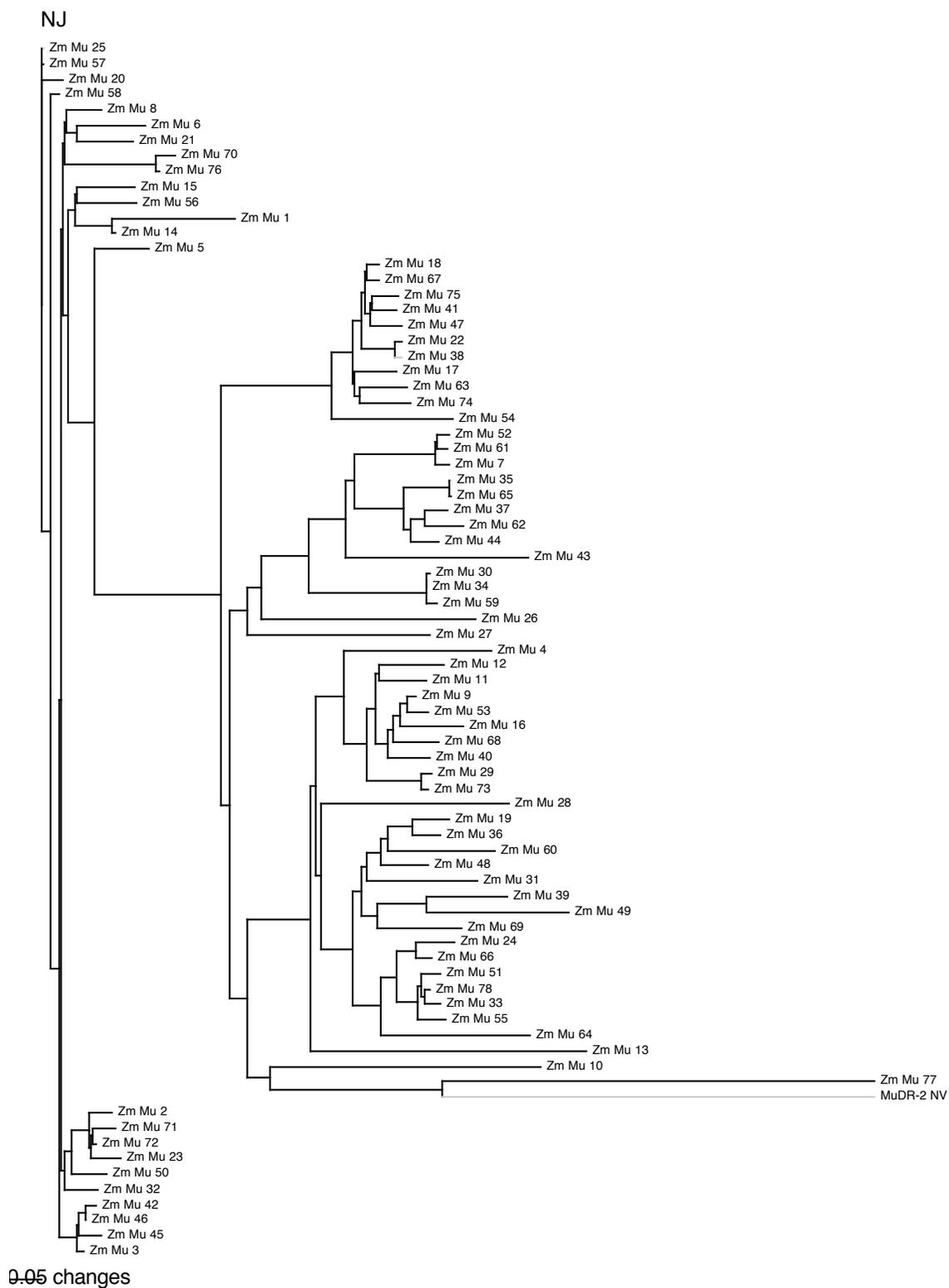


Os Mutator (unrooted)

NJ



Os Mutator (rooted)



NJ



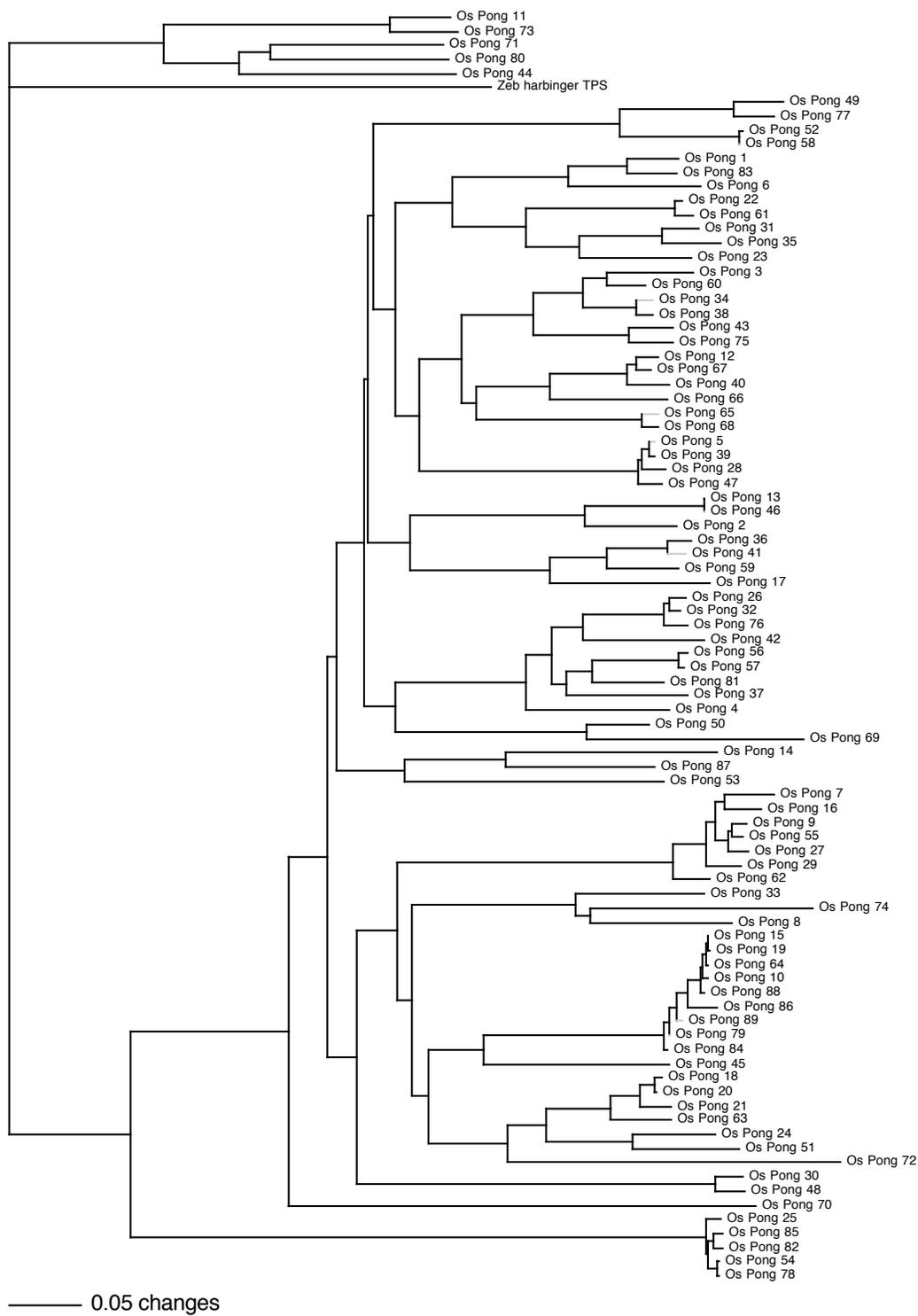
Zm Mutator (rooted)

NJ



Os PIF (unrooted)

NJ



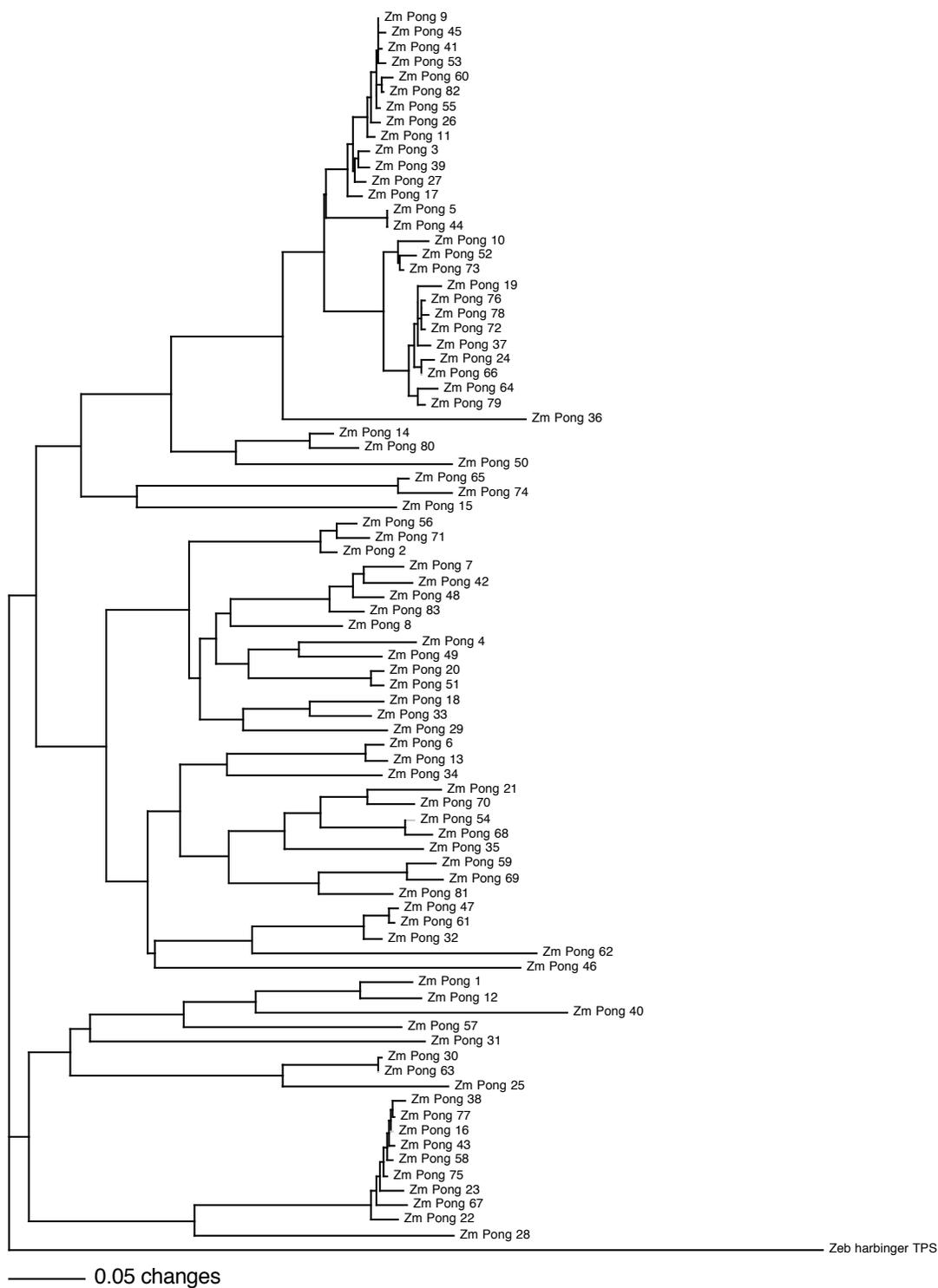
Os PIF (rooted)

NJ



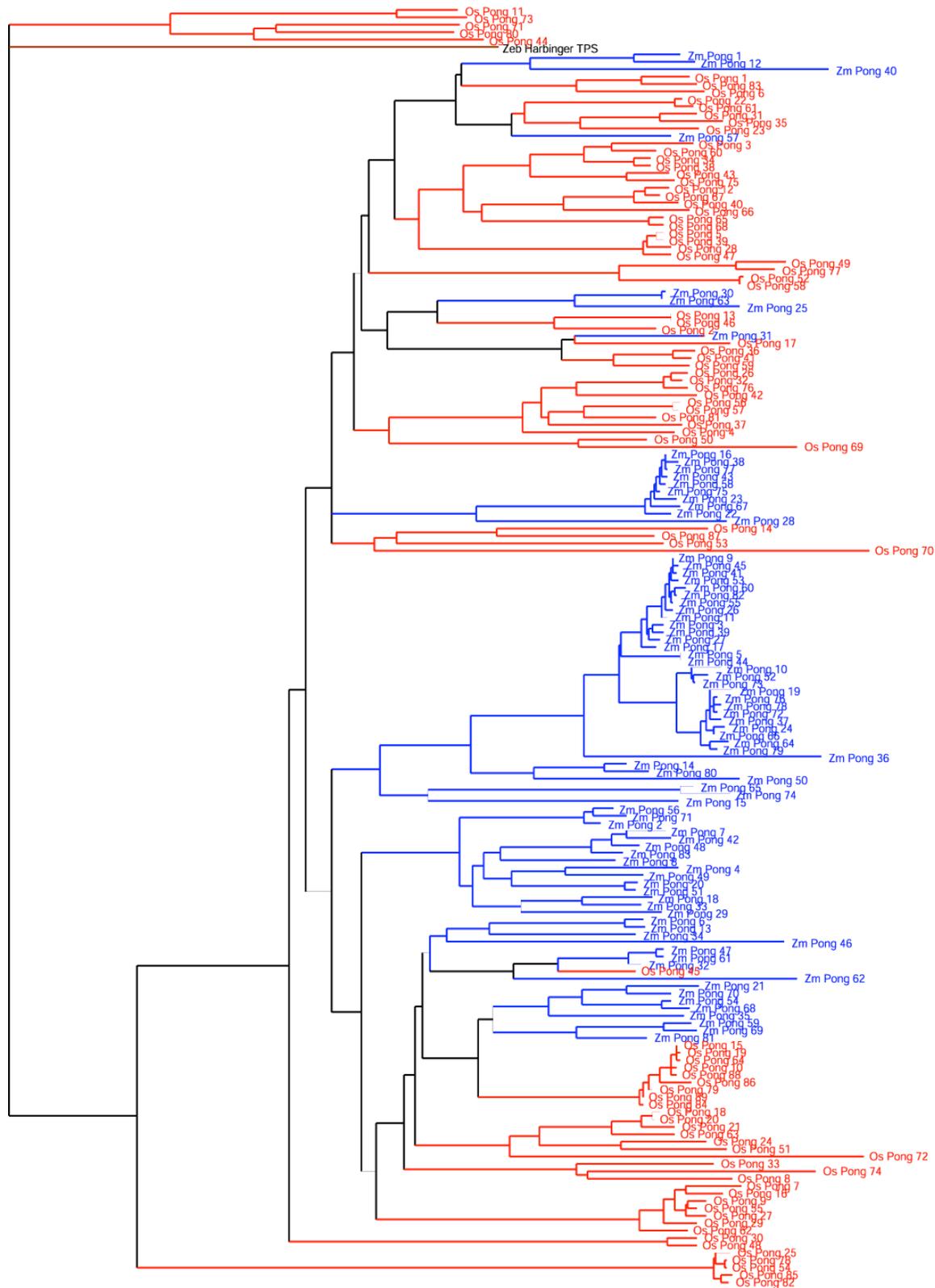
Zm PIF (unrooted)

NJ



Zm PIF (rooted)

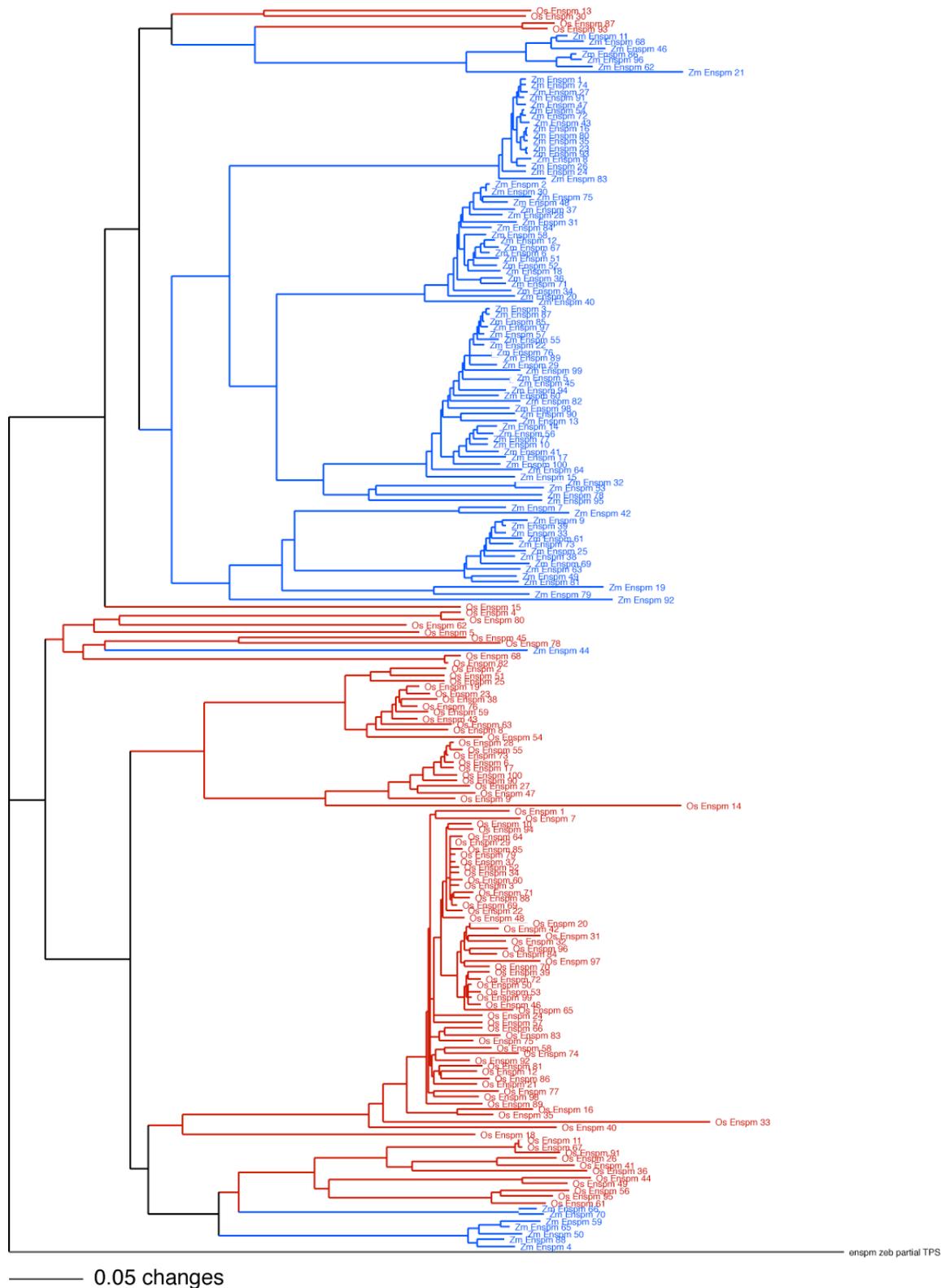
PIF (*Oryza sativa* & *Zea mays*)
NJ



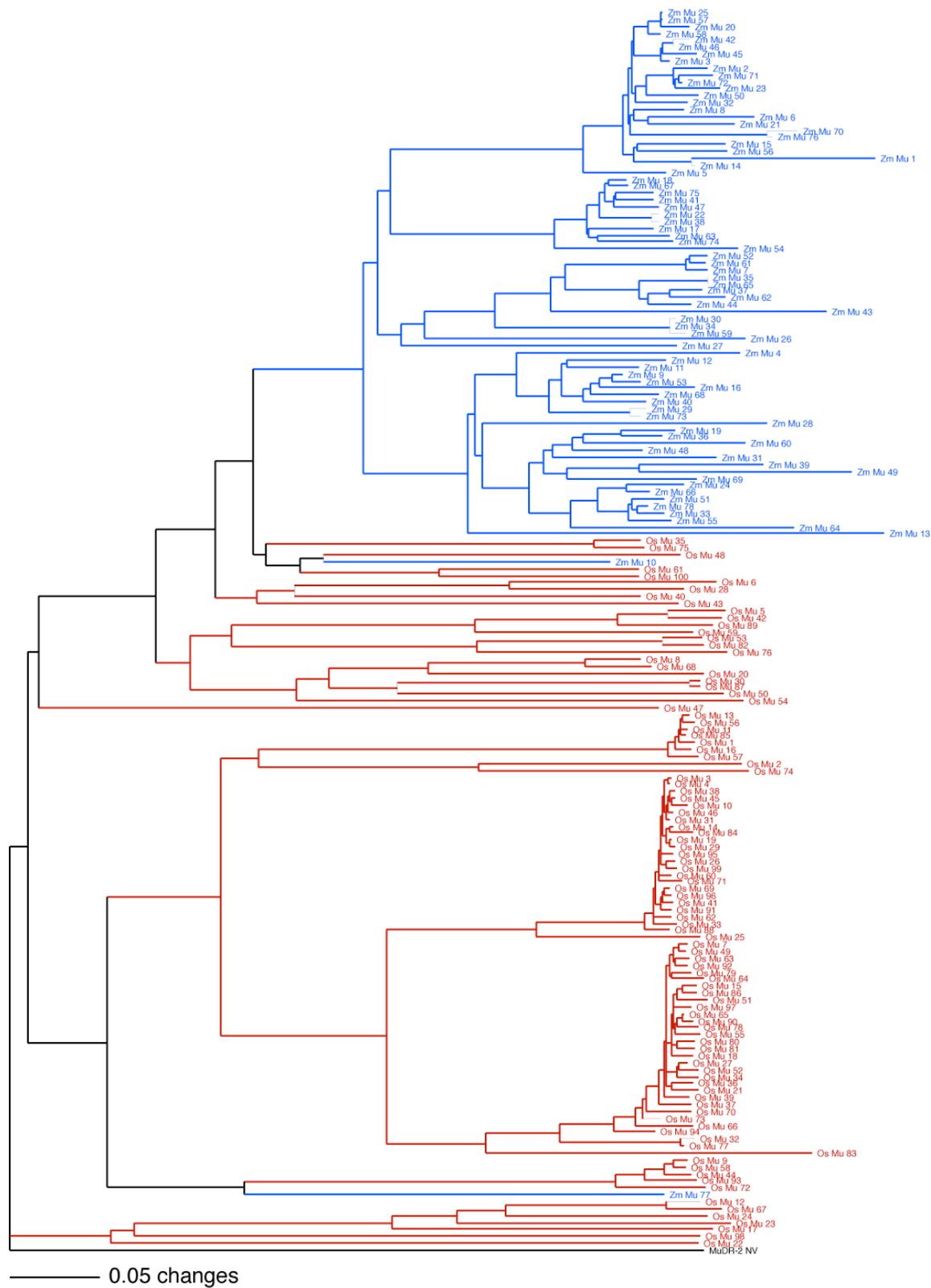
— 0.05 changes

CACTA (*Oryza sativa* & *Zea mays*)

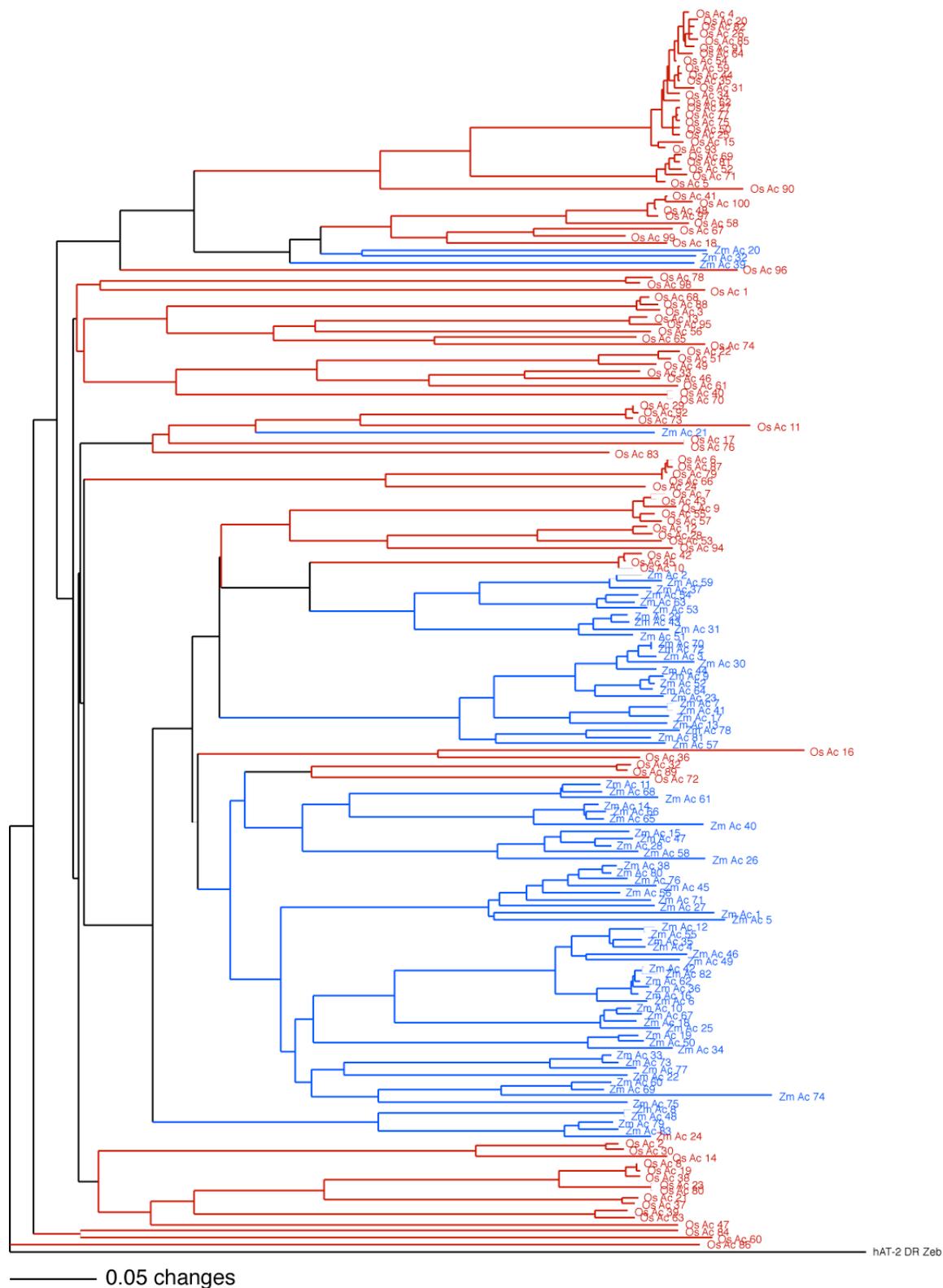
NJ



Mutator (*Oryza sativa* & *Zea mays*)
NJ



hAT (*Oryza sativa* & *Zea mays*)
NJ



Lab Report #3: TE Analysis: Applying and Integrating Evolutionary Concepts
Hand in both hardcopy and disk copy at the beginning of class on November 15

Despite what I said in class, work on this report independently; not with your group partner. Sorry!

--While I prefer that you answer the questions below in a clear format, you are free to include your answers in some form of narrative.

--READ THE QUESTIONS CAREFULLY AND ANSWER ALL PARTS

1. About your TE superfamily:

Describe the important and interesting features of your assigned TE. Include in your answer: What distinguishes it from other TE superfamilies? What is its species distribution (in general terms - e.g. "plant" is sufficient)? What are its structural characteristics (e.g. length of TIR, TSD, anything else). Has it been used by scientists in some sort of applied way?

2. About your phylogenetic trees:

For this section you will need:

-- the phylogenetic trees you generated in experiment #3.

--to consult parts of the review article - Phylogeny for the Faint of Heart. Pay special attention to the definitions of homology, outgroup (to root a tree), orthologues vs. paralogues (also written as orthologs, paralog), horizontal vs. vertical transmission.

--a general understanding of how class 2 elements duplicate (class notes pages 108 to 109)

Print out copies of your maize, rice, and combined trees of your TE superfamily and use them to address the following:

--Compare the 4 trees - maize, rice - with and without outgroups - What is an outgroup and why is it used? How did the inclusion of an outgroup in your maize vs. rice trees change the shape (topology) of each tree?

--Compare the 2 trees - maize, rice with outgroup - Describe key similarities and differences between your maize and rice trees and what these might mean.

--Now look at your unified tree (maize + rice TEs with outgroup)

- Describe the key features of this tree.

-Note examples (if any) where elements in the rice genome are more closely related to elements in the maize genome than they are to other elements in the rice genome (or vice versa). Provide an explanation for how this may have happened during the evolution of both organisms from a common ancestor (Maize and rice are both cereal grass species that arose from a common ancestor about 50 million years ago.).

--Include other interesting features that did not fit into the questions above.

3. About TEs and "The Making of the Fittest":

For this section, refer to Sean's definition of immortal genes (chapter 3, p 79) and fossil genes (chapter 5, p. 123 bottom) and additional descriptions in these chapters.

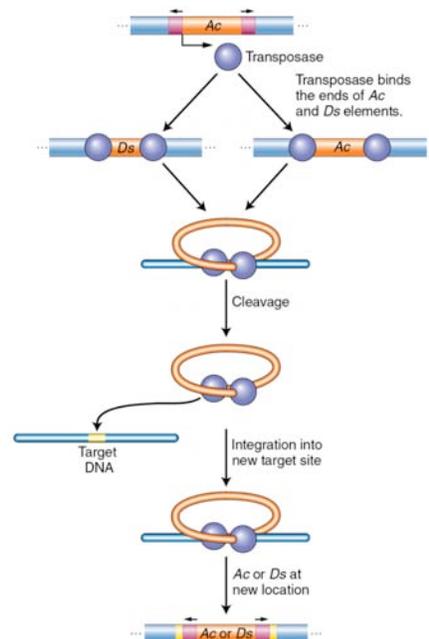
Are the transposases of TEs examples of immortal genes or fossil genes or both? Provide support for your answer. Include in your answer an explanation of what positive and negative selection has to do with TEs and whether these concepts apply to TEs in the same ways as they apply to the host genes.

Transitioning From Experiment 3 to Experiment 4: From DNA Transposons to Retrotransposons to the Meaning of Phylogenetic Trees Composed of Transposable Elements

As we head into the home stretch of the course, it is time (finally) to introduce the other major TE class - retroelements. With only two classes, TEs are usually divided into the Class 2 DNA transposons (the hATs, mariners etc) and the Class 1 retro (RNA) elements (which are also subdivided into families, see below). The primary feature that distinguishes the two classes is their mechanism of transposition. These are described and compared below along with information on how TEs make duplicate copies and increase their copy number in the genome.

How do DNA Transposons (Class 2 elements) make duplicate copies when they transpose?

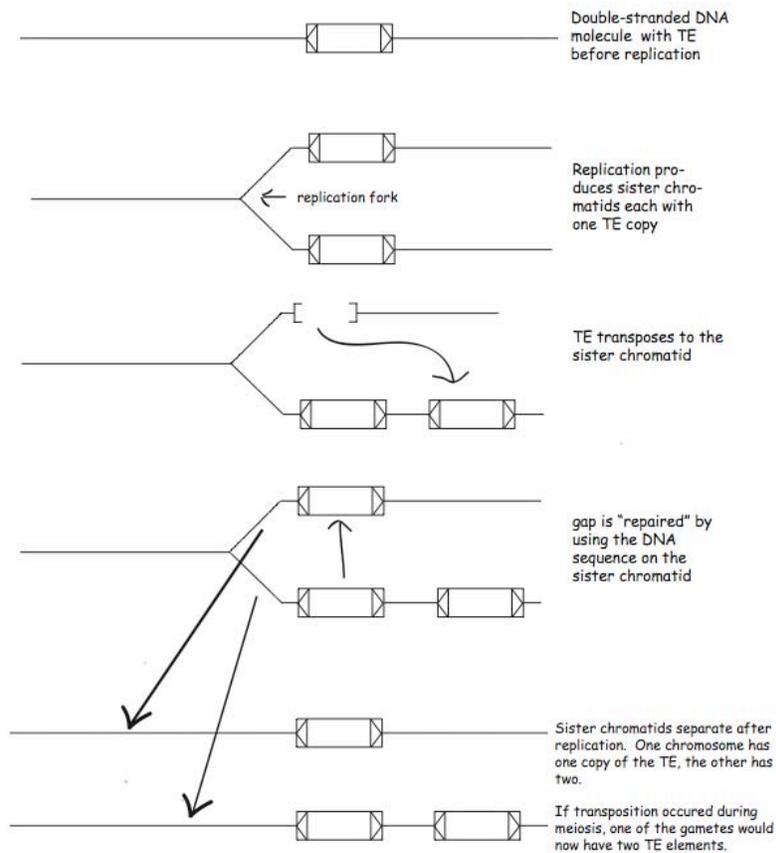
To understand how to interpret the phylogenetic trees of TEs that you generated in Experiment 3, it is important to understand how DNA elements increase their copy numbers in the genome. In short, we need to know how all the TE sequences arose that you mined with Blast, aligned with ClustalW and constructed trees with PAUP. The mechanism of class 2 element transposition was first discussed on page 18. The relevant figure is "duplicated" below....



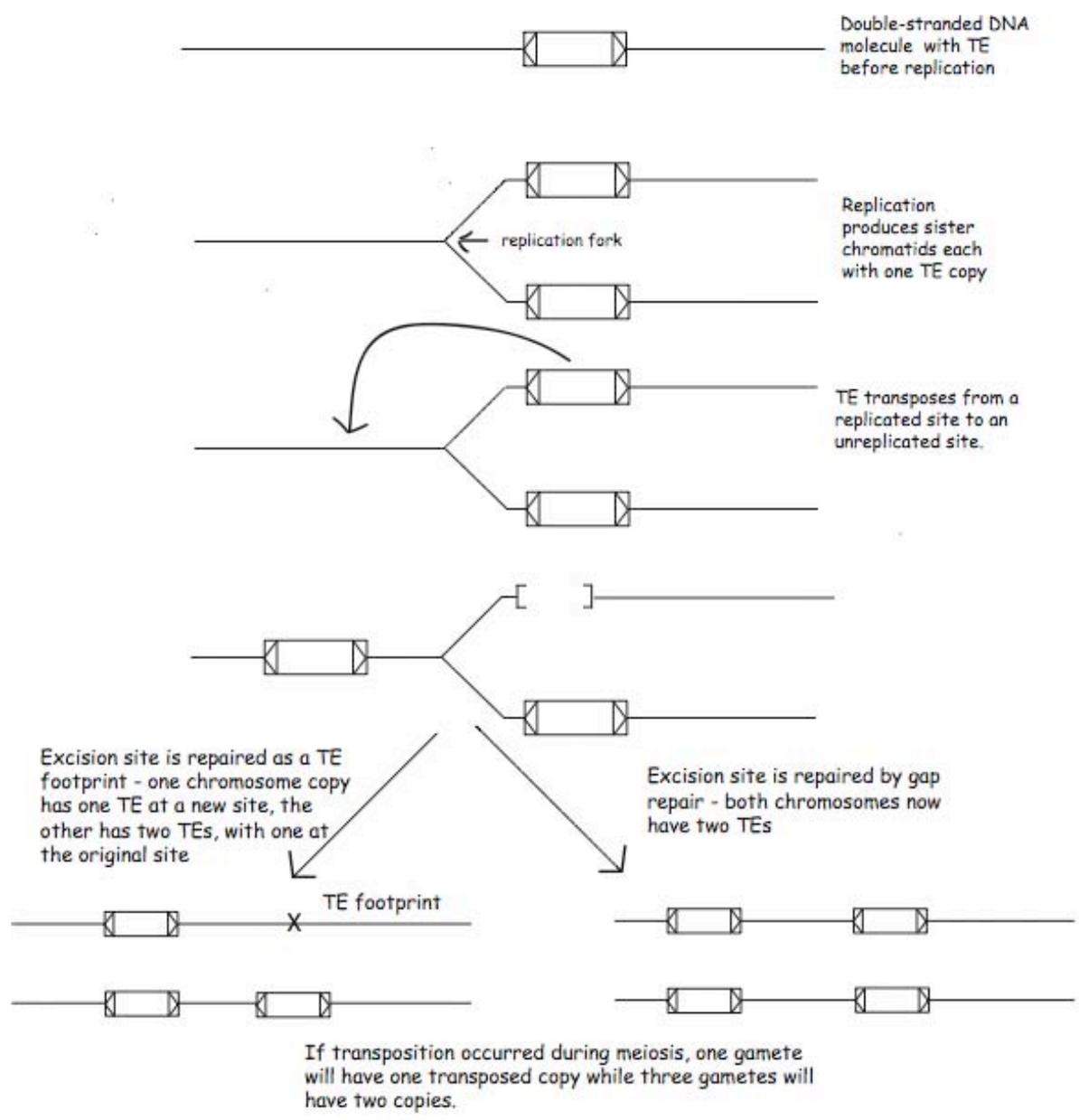
This figure shows that an autonomous element (in this case, *Ac*) encodes a transposase protein that binds to the ends of both itself and non-autonomous elements in its family (in this case, *Ds*) and catalyzes both element excision and reinsertion. As such, the element itself is the intermediate in 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* 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. In class last week, you saw that DNA transposons 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:

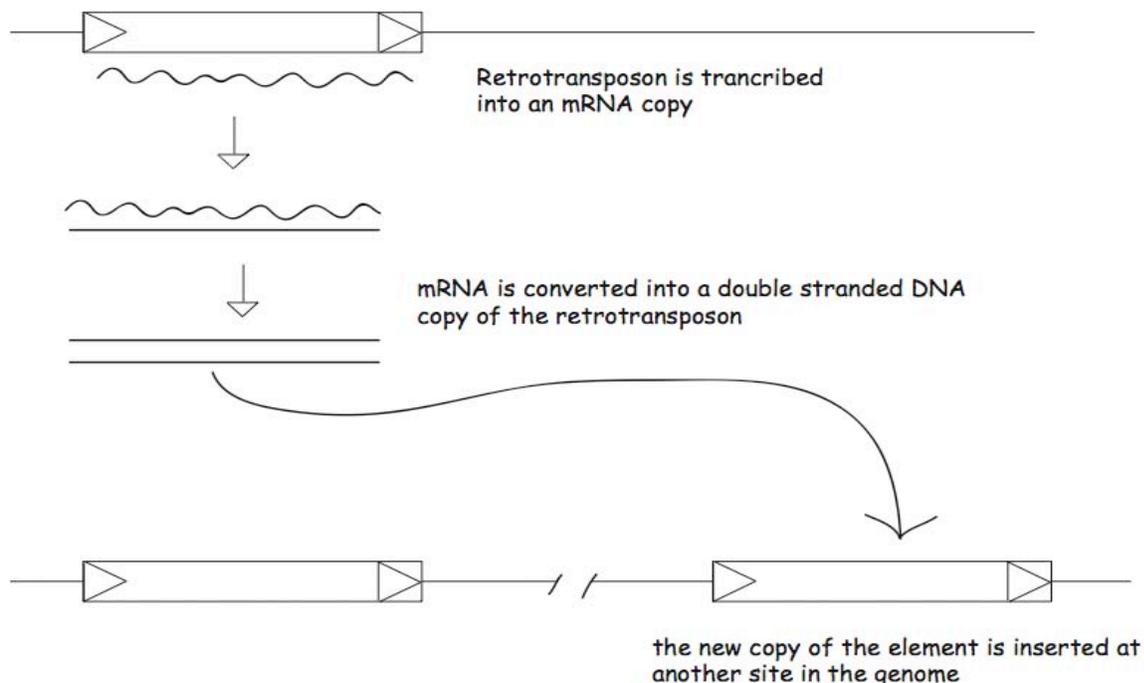


What you should note is that no matter which way a class 2 element moves, the element and its duplicate(s) are identical. Over time (evolutionary time that is), the element sequences mutate independently (e.g. by errors introduced during

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.

How do Class 1 Retroelements 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 figure 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.



Three features of this mode of transposition differ from that of DNA transposons:

- (1) the transposition intermediate is the mRNA copy of the element. This feature is true for all class 2 elements including the retrotransposon shown here.
- (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):

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. We will learn more about the structure of Class 2 elements in the sections to follow.

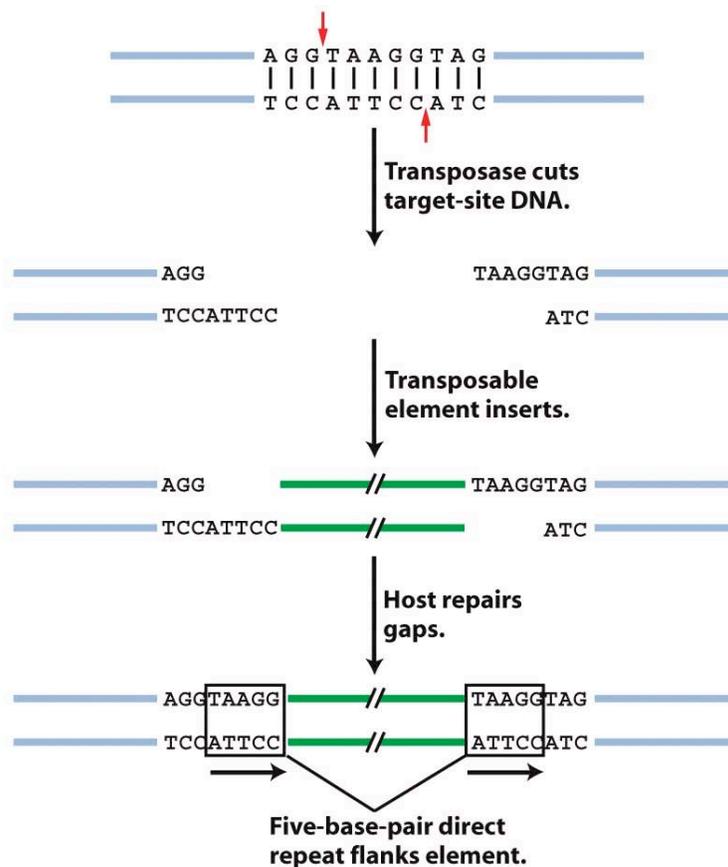


Figure 14-8
Introduction to Genetic Analysis, Ninth Edition
 © 2008 W. H. Freeman and Company

LTR retrotransposons: a class 2 element that is very similar to retroviruses:
 There are many types of class 2 elements. For now we will focus on LTR retrotransposons - which are the most abundant TE type in many eukaryotes including yeast, plants and insects. Their structure and mode of transposition are strikingly similar to a common pathogenic agent and a cause of some cancers - retroviruses.

Life cycle of a typical retrovirus:

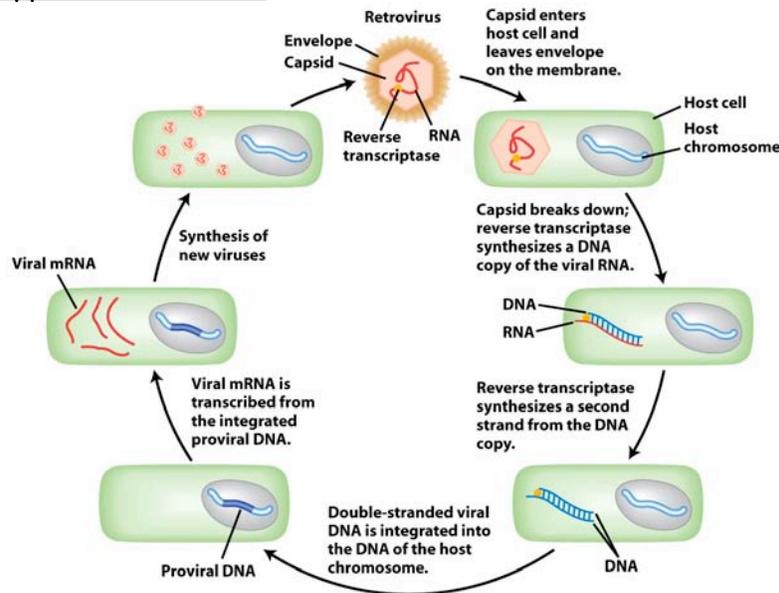


Figure 14-11
 Introduction to Genetic Analysis, Ninth Edition
 © 2008 W.H. Freeman and Company

Life cycle of a typical LTR retrotransposon:

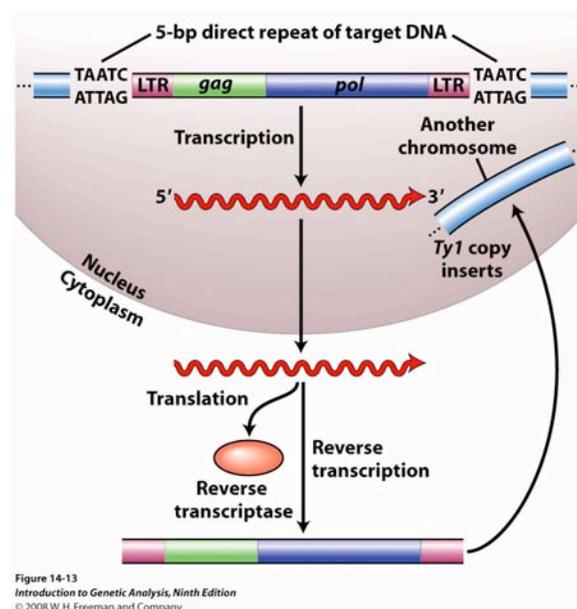
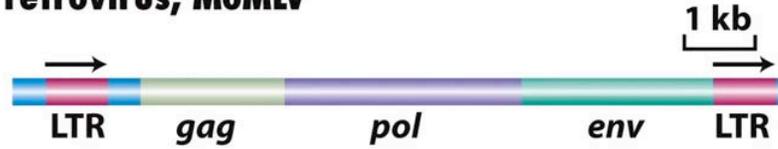


Figure 14-13
 Introduction to Genetic Analysis, Ninth Edition
 © 2008 W.H. Freeman and Company

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

(a) A retrovirus, MoMLV



(b) *Ty1* in yeast



(c) *Copia* in *Drosophila*



Now let's start looking for these elements in the genome in Experiment 4....

Experiment 4, Day 1: October 29, 2007
Introduction to Retrotransposons - Mining Complete Copia LTR
Retrotransposons from the Rice Genome

Today's objective: Yujun and I will introduce you to a very different type of element, class 2 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 in two ways, one old and one new. The old stuff - we will generate phylogenetic trees. The new stuff - we will "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.

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
NVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILLFAASKGF
KLFQMDVKS AFLNGVIEEEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPRAWYERLKTFFLQ
NGFEMGAVDKTLF'TLHSGIDFLLVQIYVDDIIIFGGSSHALVAQFSDVMSREFEMSMMGELTFFL
GLQIKQTKEGIFVHQTKYSKELLKKFDMADCKPIATPMATTSSLGPDDEGEVDQREYRSMIGS
LLYLTASRPDIHFSVCLCARFQASPRTSHRQAVKRIFRYI
```

Which is used in a blast search for related elements...

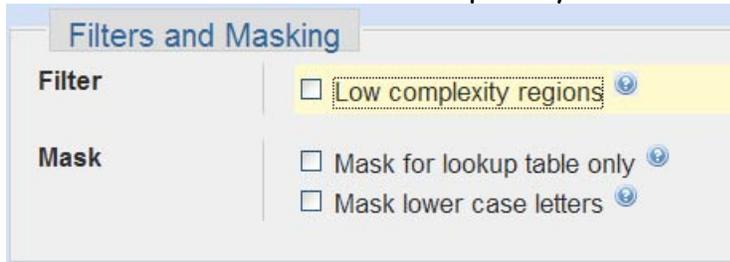
Blast search (<http://www.ncbi.nlm.nih.gov/BLAST/> tblastn)

Database: Nucleotide collection (nr/nt)

Organism: Oryza sativa (taxid:4530)

Choose Search Set	
Database	Nucleotide collection (nr/nt) <input type="button" value="v"/> <input type="button" value="i"/>
Organism Optional	Oryza sativa (taxid:4530) <input type="text"/> Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. <input type="button" value="i"/>
Entrez Query Optional	<input type="text"/> Enter an Entrez query to limit search <input type="button" value="i"/>

Remember to uncheck "low complexity filter" in "Algorithm parameters" menu.



Click "BLAST"

Step 2. Saving your Blast hits to generate the multiple alignment needed to build a phylogenetic tree:

Save all of your blast results as one text file and copy it Yujun's (Han) folder in the class share folder. Yujun will process it with his program and put the results in the class share folder for future use (at the end of the class period). To build rooted tree, you can use the following sequence, which is a copia reverse transcriptase in *Drosophila*.

```
>Drosophila_copia_frogger (Outgroup)
MLLAVAAEKDLHMHQIDISNAYLNSDLEEDVYLKQPKNYVDKENPGKVLKLQKAIYGLKQSERL
WNDALNEVLQNMGFKRKNEACLYYKKQONGFSYIAVYVDDLIIISPKESDIEDIKGSIATKFD
MKDGGQLRYFLGMEISRKGQTGPIKLCQKRYIENLLRRYGMQSCRLVGTFFDPPGYESGCTNEKC
AKVNLTHFQSLIGSLMYLAVVSRPDILHSVSKLSQRNTDPHHEDEAAAKHVLRYLCGTINLSII
YMKTGELVKEFADADWANDKVDRKSYSGYAFLMAGSAFSWGSSKQSVIAQSSTEAEYIALSTAA
KEAVFLRRLLOEMGWFDKGPLKLLCDNLSASSIAKNP INHKRTKHIDVRYHFIRDKVNKNEIIV
EYVNTQNNVA
```

Step 3: 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. For an example, the following BAC: AL606652:

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

Click the score in this line and you will 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 GGLGERVTRNRSYELVNSAFVASFEPKNVCHALSDENWVWVAMHEELENFERNKVWSLVEP 60
GGLGERVTRNRSYELVNSAFVASFEPKNVCHALSDENWVWVAMHEELENFERNKVWSLVEP
Sbjct 17511 GGLGERVTRNRSYELVNSAFVASFEPKNVCHALSDENWVWVAMHEELENFERNKVWSLVEP 17332

Query 61 PLGFNVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILL 120
PLGFNVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILL
Sbjct 17331 PLGFNVIGTKWVFKNKLGEDGSIVRNKARLVAQGFTQVEGLDFEETFAPVARLEAIRILL 17152

Query 121 AFAASKGFKLFQMDVKSFAFLNGVIEEBEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPR 180
AFAASKGFKLFQMDVKSFAFLNGVIEEBEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPR
Sbjct 17151 AFAASKGFKLFQMDVKSFAFLNGVIEEBEVYVKQPPGFENPKFPNHVFKLEKALYGLKQAPR 16972

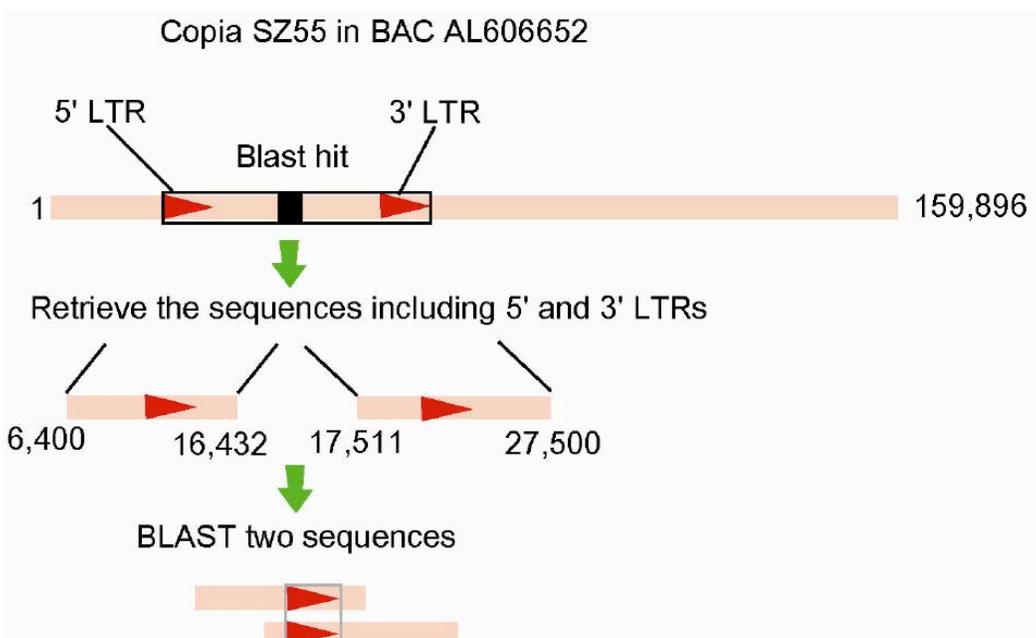
Query 181 AWYERLKTFFLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIIFGGSSHALVAQFSDVM 240
AWYERLKTFFLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIIFGGSSHALVAQFSDVM
Sbjct 16971 AWYERLKTFFLQNGFEMGAVDKTLFTLHSGIDFLLVQIYVDDIIIFGGSSHALVAQFSDVM 16792

Query 241 SREFEMSMMGELTFFLGLQIKQTKEGIFVHQTKYSKELLKCFDMADCKPIATPMATTSSL 300
SREFEMSMMGELTFFLGLQIKQTKEGIFVHQTKYSKELLKCFDMADCKPIATPMATTSSL
Sbjct 16791 SREFEMSMMGELTFFLGLQIKQTKEGIFVHQTKYSKELLKCFDMADCKPIATPMATTSSL 16612

Query 301 GPDEDGEEVDQREYRSMIGSLLYLTA SRPDIHFSVCLCARFQASPR TSHRQAVKRIFRYI 360
GPDEDGEEVDQREYRSMIGSLLYLTA SRPDIHFSVCLCARFQASPR TSHRQAVKRIFRYI
Sbjct 16611 GPDEDGEEVDQREYRSMIGSLLYLTA SRPDIHFSVCLCARFQASPR TSHRQAVKRIFRYI 16432
```

The sbjct is in the "minus" direction (see Frame = -1) meaning that the hit reads in the opposite direction as the BAC sequence is numbered 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 the visual of our strategy...



Step 4: Retrieving the BAC

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

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

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

A new webpage will show up. This page contains all of the information about this BAC including its complete sequence - yes - all 159,000 plus bases.

The screenshot shows the NCBI Nucleotide search interface. The search bar contains 'Nucleotide' and 'AL606652'. The results section shows a single entry for 'AL606652. Reports Oryza sativa geno...[gi:70663936]'. Below the entry, there are links for 'Comment', 'Features', and 'Sequence'. The 'Features' section is expanded, showing details for the LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS, SOURCE, ORGANISM, REFERENCE, and AUTHORS.

NCBI Nucleotide Protein Genome

Search Nucleotide for [AL606652] Go Clear

Limits Preview/Index History

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

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

1: AL606652. Reports Oryza sativa geno...[gi:70663936]

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

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

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

ACCESSION AL606652

VERSION AL606652.4 GI:70663936

KEYWORDS HTG.

SOURCE Oryza sativa (japonica cultivar-group)

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

REFERENCE 1

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

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

ORIGIN

```

1 gaattctttc aatgtttct tcaactttag caactgtctc ctttgagacc tgatggccag
61 ccttatcaaa gactgcataa ctgtaacaga atcaattgac agagttgatg taagaatcaa
121 caaggattgt gcggatcggg aaagaaaagc gtaagatcaa gagctaaaag attacctttc
181 taaatcatga tcatacagaa cagagttgtc gccactagtg cgatataatt tcagcccaag
241 atagccaatc aatggtgcca aggagttctc attacaaacg ccgtggagcc tgaatTTTTG
301 gaatgaacag taagtaagct tgtatgaaca gaatctaaag tgaatttctc acactaacia
361 ttcagggtga gactgaccat gaggctccca tatcaattgg gcatccgaaa gagtaatcgg
421 tatggacacg accgcctacg cgatctctgg actccaaaac agtcacctca aacgaagcat
481 tggagagagc acgggctgct gcaacccttg aaattcccc accgatcagc atgacgggatg
541 gaggcgaagc acattgcctc tcaatggctg 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 include the LTRs and Blasting them against each other: First, we need to change the default format from Genbank to FASTA so that we can use the sequence in the blast program. To do this use the Display menu to choose the FASTA format:

NCBI Nucleotide search interface. The 'Display' dropdown menu is open, showing options: GenBank, GenBank(Full), INSDSeq XML, **FASTA**, TinySeq XML, ASN.1, XML, Graph, GI List, Brief, Summary, and ACCESSION. A red arrow points to the 'FASTA' option. Below the menu, the search results for 'Oryza sativa' are visible, including the accession number 'AL606652' and a description: 'Oryza sativa genomic DNA, chromosome 4, BAC clone: OSJNBb0004A17, sequence.' The 'Range' field is currently empty, and the 'Refresh' button is visible.

Next, we will retrieve and save the regions presumed to contain the 5' LTR and the 3' LTR. In our example, to get the sequence including the 5'LTR, we input "6400" and "16400" into "Range" windows and click "Refresh".

NCBI Nucleotide search interface. The 'Display' dropdown menu is set to 'FASTA'. The 'Range' field is filled with '6400' and '16400'. The 'Refresh' button is highlighted. The search results for 'Oryza sativa' are visible below the menu.

Copy the FASTA sequence...

NCBI Nucleotide search interface. The 'Display' dropdown menu is set to 'FASTA'. The 'Range' field is filled with '6400' and '16432'. The 'Show whole sequence' button is highlighted. The FASTA sequence is displayed below the search results:

```
>gi|70663936:6400-16432 Oryza sativa genomic DNA, chromosome 4, BAC clone: OSJNBb0004A17, comple
GGCGTTCCTCCCTCGGGCCACCGCCGCCCGCCGCGCGCGCGGAGCCGAGCCGAGCCGCGTGC
GTGGAGAGCCGAGAGAGACACAGCGAGAGATTGGCTCGAGCCGAGGGGAGCCGATGATGCGGTGGATT
GGCGCCCGGGAGCCCGGTTTTATAGCGATGCCCTTTTGGGTTTTGACCCCTCTTTTTTTGGGTTAT
TTCGAAACACGGCCGGCTCGTAGATCGAGCCGGCTTGAACAGCCGAGCTGGATGGGCACCTGTCCATGTG
GTCCATCCACTCTCACTCTCCCGTCGATGCGAGATTTTTTGTAAATATTTTTAGAACGATGCGAGATT
CTTGGCTCTGTGAATTCACACACCAACCAAAATACACCGGCAATGTTAAAAAAAATGTGATGCTGAA
AATCCTGTGTGGAGTGTGCAATGTGCGATGATGTTTACGGGTAACCTTACAAAACGTGTG
GGAAGTGTAAATGATATATATATATATATATATATATATATATATATATATATATATATATATAT
ACAAAATTTATGTAACCTATATATATATATATATATATATATATATATATATATATATATATATAT
AGATCAAGTATTTCAAACACTCAAATTTAATATGAAATATCACATTGTTAAATATCTAGCAGTACCG
TTATGATAAAGTAAATATAAAGTGTAGTTTTTACTTGGTGGAGACATAAAATTTTATGATAAAT
TTATGCTAAATTTGTTGTTGGTATAAATTTAGTCTAATATATACCTTATGATACATTGATTA
TATATAGTTTTCTATAATTTAATTTTATGAAATTTACCGGTGATATATCATCAGTGTGTGATGAGCC
```

...and paste it into the top window of "BLAST 2 SEQUENCES"
(<http://www.ncbi.nlm.nih.gov/BLAST/bl2seq/wblast2.cgi>).

Sequence 1

Enter accession, GI or sequence in FASTA format from: to:

```

CTTACGAGAGGACCAAGAGACCAAAGAAGTGCCAAGAAAATGACAAGTGCCAGAAGTAC
TTTTGCGATCA
ATCTTACACCCAGCAAAATCAGCATCTGAAAAAGCTCGAACTGAAAGGGCAGAAGAGCA
GGAGTACAAA
TACCGTACTCAAGTGTAGACTTG

```

or upload FASTA file

Similarly, to retrieve the sequence including the 3' LTR, you just need to input "17500" and "27500" into the "Range" windows and click "Refresh".

NCBI Nucleotide

Search for

Display Show Send to

Range: from to Reverse complemented strand

Copy and paste the output into the lower window of the "BLAST 2 SEquences".

Sequence 2

Enter accession, GI or sequence in FASTA format from: to:

```

TATGACTTATG
AAAAATGAAGAAAATAAGAAAGTAGTCATACGAATTCAATTGAATGAAATTCCTAAA
CATACCCCACA
AAGGAACAATATCCGTTTGTGTTGAGCCAGTTACAGAGCCTCCAAGCGCTTTCACAAAA
GG

```

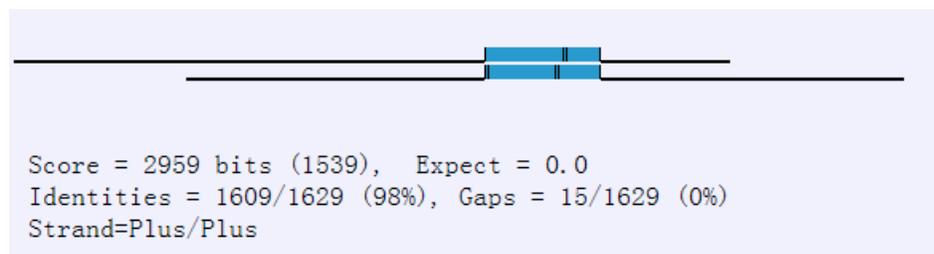
or upload FASTA file

Note: uncheck "Filter" option before clicking Align.

Parameters used in [BLASTN](#) program only:

Match: Mismatch:
 Open gap and extension gap penalties
 gap x_dropoff expect word size [Filter](#)

The result should look like this:



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

Step 5: Retrieving the LTR sequence from this alignment and reformatting.
 Copy the complete blast hit.

```

Query 6789 CTAGAAGTGAGGTGAGAAAGAGAGAGCAAAACTCATCATCGCAAAGTTCAAATTGCAAGC 6848
          |||
Sbjct 4373 CTAGAAGTGAGGTGAGAAAGAGAGAGCAAAACTCATCATCGCAAAGTTCAAATTGCAAGC 4432

Query 6849 GGAATTTAAATTGCGGAATTTAAATGGACAAGGCAAAAATGAAATCCTTCAAATCATTTC 6908
          |||
Sbjct 4433 GGAATTTAAATTGCGGAATTTAAATGGACAAGGCAAAAATGAAATCCTTCAAATCATTTC 4492

Query 6909 ATTTATAGGTGATGCAAATAAACCCTCAACTAGGAGCAAACATACACCTTCAGAGGAAC 6968
          |||
Sbjct 4493 ATTTATAGGTGATGCAAATAAACCCTCAACTAGGAGCAAACATACACCTTCAGAGGAAC 4552

Query 6969 ATTAACACAAACTTAAATCTCTCGGACAAACACACTCCAAACTAATCCTAATACAAAAG 7028
          |||
Sbjct 4553 ATTAACACAAACTTAAATCTCTCGGACAAACACACTCCAAACTAATCCTAATACAAAAG 4612

Query 7029 CCTCTCGGGCAAACACACTCCAAACTCACACGGAAACTCTCTCACCAGCATCTCAAAT 7088
          |||
Sbjct 4613 CCTCTCGGGCAAACACACTCCAAACTCACACGGAAACTCTCTCACCAGCATCTCAAAT 4672

```

To "extract" one copy of this sequence to use as a query, we need to format this hit at Wun Chiou's website....

(<http://www.wunchiou.com/test/formatblast.html>).

Paste the blast hit into the top window, select "Subject-only output" and click "format the above text". The LTR sequence should be in the lower window.

Enter the BLAST result data to be formatted:

```

Sbjct  5719  |||...|||
CTTCTTGCACATCCGGTACTTGTCAATTTCCCATCACAAAAGAACTATAACCACACATGG  5778

Query  8193  TTTCACAAT  8201
      |||...|||
Sbjct  5779  TTTCACAAT  5787
  
```

Query base pairs:

Normal format output

Subject-only output

Copy the LTR sequence from the "Output" window and read on....

Output:

```

TGAAAGACCAAGAACAGCTATAGAGGG-----GGGGGTGAATATAGCAATTCAAAT
CTTGCCCCCGAAAATACTCATCAAGCCGGATTCTCAAATCCTTACTAGAAATCGCGGCT
ATTAGAGAAGCCGGATCTAGAAAAGAAGAGAGAAAAAGAAGAGAAAAGGAATFCCCGAAA
CTAGAGGAGGAAAGAGAAAAGGAATCCCGAAACTAGAAGTGAGGTGAGAAAGAGAGAGCA
AAACTCATCATCGCAAAGTTCAAATGCAAGCGGAATTTAAATTGCGGAATTTAAATGGA
CAAGGCAAAAATGAAATCCTTCAAATCATTTCATTTATAGGTGATGCAAAAATAACCGCTC
AACTAGGAGCAAACATACACCTTCAGAGGAACATTAACACAAACTTAAAATCTCTCGGAC
AAACACACTCAAACATAATCCTAATACAAAAGCCTCTCGGGCAAACACACTCAAACCTCA
CACGGAACCTCTCTACCGAGCATCTCAAATGATTACCAAAGGAGCAACCTCCACCCT
TGCATCCATCTCTTATTATAGCCTAAGACCCCTAAGACATTGTCTCAAATACCCCTAG
GGCGAAACCCTAACTCAGAACAGATCTGGTCCATCCATTGTTCTTACTCAAAGGAAA
AGCTCCAGATGATTGCCACCTCATCGATCCGAACCCTCACATGGTCTTCTTGCTGATGAA
TCCGCGTGCTCTCCGTCTTGACGAATCCAAACTTGCCACGTTGCCTAGCCGCGTCTGCG
  
```

Step 6: Retrieving the entire *Copia* element from the BAC sequence.

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

To do this we need to go to "BLAST two sequences" again and paste the LTR sequence in the top window and input the BAC name "[AL606652](#)" in the lower window. Uncheck the "Filter" option.

Open gap and extension gap penalties
 gap x_dropoff [expect](#) word size [Filter](#)

Sequence 1
 Enter accession, GI or sequence in FASTA format from: to:
 CAAACAACCGTTGTCTTGGCACAGATGTCGCAACCTGACCAACGTTAGTCCACACACAC
 A
 CTTCTTGACATCCGGTACTTGTCAATTTCCCATCACAAAAGAACTATAACCACACATG
 G
 TTTCACAAT

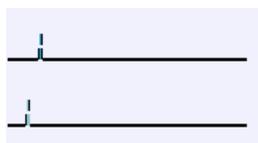
or upload FASTA file

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

or upload FASTA file

Click "Align".

Because each element has two LTRs, there should be two hits in the blast result.
 This is designated as follows...



The first hit is (middle part not shown):

```

Score = 3111 bits (1618), Expect = 0.0
Identities = 1618/1618 (100%), Gaps = 0/1618 (0%)
Strand=Plus/Plus

Query 1      TGAAAGACCAAGAACAGCTATAGAGGGGGGGTGAATATAGCAATTCAAATCTTGCCCCC 60
          |||
Sbjct 21669  TGAAAGACCAAGAACAGCTATAGAGGGGGGGTGAATATAGCAATTCAAATCTTGCCCCC 21728

.....
.....
.....

```

```

Query 1561 TCCGGTACTTGTCAATTTCCCATCACAAAAGAACTATAACCACACATGGTTTCACAAT 1618
          |||
Sbjct 23229 TCCGGTACTTGTCAATTTCCCATCACAAAAGAACTATAACCACACATGGTTTCACAAT 23286

```

The second hit is (middle part not shown):

```

Score = 2959 bits (1539), Expect = 0.0
Identities = 1609/1629 (98%), Gaps = 15/1629 (0%)
Strand=Plus/Plus

```

```

Query 1 TGAAAGACCAAGAACAGCTATAGAGGGG-----GGGGTGAATATAGCAATTCAAAT 51
          |||
Sbjct 12976 TGAAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGGGGGTGAATATAGCAATTCAAAT 13035

```

```

Query 1550 CTTCTTGCACATCCGGTACTTGTCAATTTCCCATCACAAAAGAACTATAACCACACATGG 1609
          |||
Sbjct 14532 CTTCTTGCACATCCGGTACTTGTCAATTTCCCATCACAAAAGAACTATAACCACACATGG 14591

Query 1610 TTTCACAAT 1618
          |||
Sbjct 14592 TTTCACAAT 14600

```

Write down the lowest and highest numbers - "12976 and 23286". These define the location of this complete element on BAC AL606652.

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

NCBI Nucleotide search interface. Search criteria: Nucleotide for AL606652. Range: from 12976 to 23286. Display: FASTA. Show: 5. Send to: [dropdown]. Reverse complemented strand: [checkbox]. Refresh button is present.

```

>gi|70663936:12976-23286 Oryza sativa genomic DNA, chromosome 4, BAC clone: OSJNBb0004A17, complete
TGAAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGGGGGTGAATATAGCAATTCAAATCTTGCCCCCG
AAAATACTCATCAAGCCGGATTCTCAAAAATCCTTACTAGAATCGCGGCTATTAGAGAAGCCGGATCTAG
AAAAGAAGAGAGAAAAAGAGAAAAGGAATTCCCGAAACTAGAGGAGGAAGAGAAAAGGAATCCCGA
AACTAGAAGTGAAGTGAGAAAGAGAGAGCAAACTCATCATCGCAAAGTTCAAATTGCAAGCGGAATTTA
AATGCGGAATTTAAATGGACAAGGCAAAAATGAAATCCTTCAAATCATTTTATAGGTGATGCAAA

```

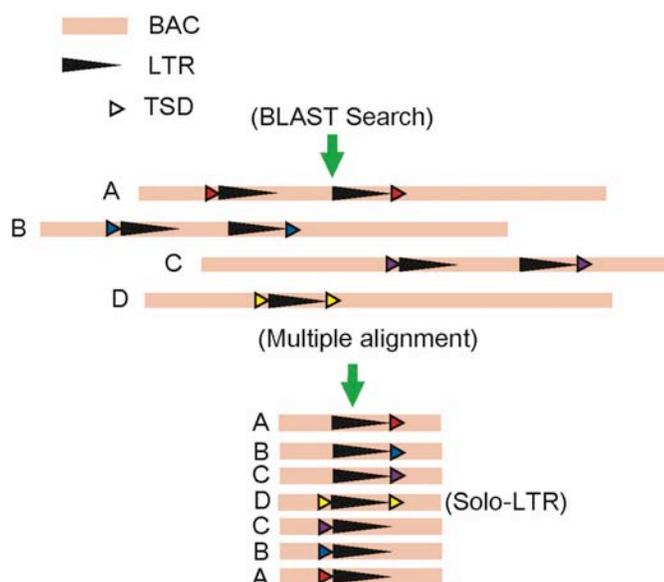
Click Refresh.

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

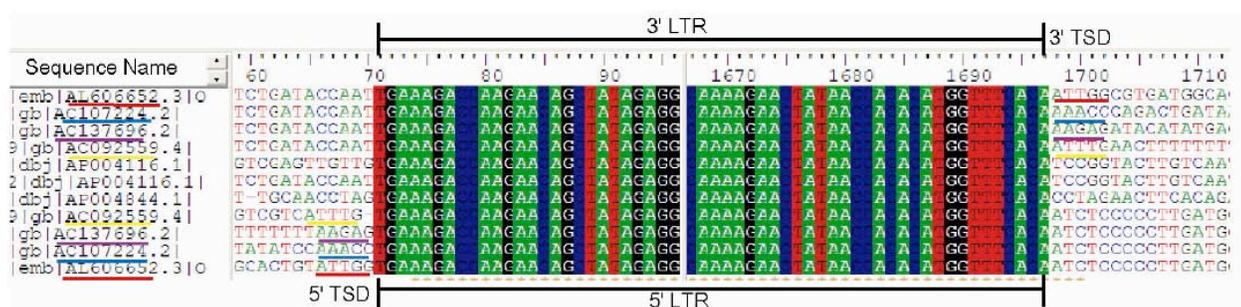
Step 7: Retrieving and identifying the TSD: Yujun to the rescue!

You now have before you the sequence of a full-length rice copia element. Hold your excitement because we still need to identify the target site duplication (TSD). Remember that? (See page 111 if you have forgotten what this is and how it is generated during the insertion of any TE). For virtually all LTR retrotransposons, the TSD is 5bp.

How do we do that if we have effectively cut out the complete element from the rest of the BAC?? It turns out that this can be tedious and inefficient. So, as in previous situation, Yujun has come to the rescue and has written a short program to find TSDs. All you need to do is copy your LTR sequences in a word file and save it in the class share folder. Yujun will use his program to do blast searches to find all copies of this LTR in the rice database including 60bps of flanking sequences. This is then used as an input file to generate multiple alignments with ClustalW. Here is the flowchart of his program:



From the alignment result, you can easily find the TSD as follows:



Step 8: Finding the element-encoded open reading frames (ORFs):

Go to the homepage of ncbi and find the "Tools" at the left. Click it.

The screenshot shows the NCBI homepage. On the left, a blue navigation menu contains the following items: 'Genomic biology' (The human genome, whole genomes, and related resources), 'Tools' (Data mining), and 'Research at NCBI' (People, projects, and seminars). A red arrow points to the 'Tools' link. The main content area includes a 'PubMed Central' banner with the text 'An archive of biomedical and life sciences journals' and a list of features: 'Free fulltext', 'Over 1,100,000 articles from over 340 journals', and 'Linked to PubMed and fully searchable'. Below this is a 'NCBI News' banner. On the right side, there is a list of resources: 'Influenza Virus Resource', 'Map Viewer', 'dbMHC', 'Mouse genome resources', 'My NCBI', 'ORF finder', and 'Rat genome'.

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

The screenshot shows the NCBI 'Tools' page. On the left, a blue navigation menu contains the following items: 'Map Viewer', 'Interactive chromosome viewer', 'Model Maker', 'View evidence used to build a gene model', 'ORF finder', 'Open reading frames', and 'Organism Specific Resources' (Bee, Cat, Chicken, Cow, etc.). A red arrow points to the 'ORF finder' link. The main content area contains several tool descriptions: '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.', and '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 alterr 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: OSJNb0004A17, complete
sequence
TGAAAGACCAAGAACAGCTATAGAGGGGGGGGGGGGGGGTGAATATAGCAAT
TCAAATCTTGCCCCG
AAAATACTCATCAAGCCGGATTTCTCAAATCCTTACTAGAATCGCGGCTATTA
GAGAAGCCGGATCTAG
AAAAGAAGAGAGAAAAGAAGAGAAAAGGAATCCCGAAACTAGAGGAGGAAGA
```

The result will look something like this:

NCBI ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

gi|70663936:12976-23286 Oryza sativa genomic DN

View 1 GenBank Redraw 100 SixFrames

Frame	from	to	Length
-1	2929..	6027	3099
-3	6143..	8635	2493
+2	8168..	8659	492
-3	647..	1135	489
+2	5690..	6160	471
-2	5043..	5489	447
+2	4901..	5335	435
+2	9692..	10096	405
-1	8212..	8598	387

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.

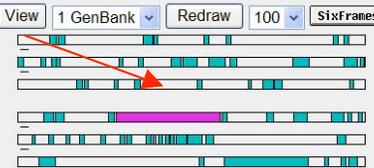
ORF Finder (Open Reading Frame Finder)

PubMed Entrez **BLAST** OMIM Taxonomy Structure

gi|70663936:12976-23286 *Oryza sativa* genomic DNA, chromo

Program **blastp** Database **nr** **BLAST** with parameters **Cognitor**

View 1 GenBank Redraw 100 SixFrames



Length: 1032 aa

Accept Alternative Initiation Codons

Frame	from	to	Length
-1	2929..	6027	3099
-3	6143..	8635	2493
+2	8168..	8659	492
-3	647..	1135	489
+2	5690..	6160	471
-2	5043..	5489	447
+2	4901..	5335	435
+2	9692..	10096	405
-1	8212..	8598	387
-1	9331..	9693	363

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

ORF Finder (Open Reading Frame Finder)

PubMed Entrez **BLAST** OMIM Taxonomy Structure

gi|70663936:12976-23286 *Oryza sativa* genomic DNA,

Program **blastp** Database **nr** **BLAST** with parameters **Cognitor**

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

BLAST Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ Format Request

Query **lc|22374 (1032 letters)**

Database **nr**

Job title **lc|22374 (1032 letters)**

Request ID **JA4JXR3T012** **View report** Show results in a new window

Format

Show Alignment as HTML Advanced View [Reset form to defaults](#)

Alignment View Pairwise

Display Graphical Overview Linkout Sequence Retrieval NCBI-gi

Masking Character: Lower Case Masking Color: Grey

Limit results Descriptions: 100 Graphical overview: 100 Alignments: 100

Organism Type common name, binomial, taxid, or group name. Only 20 top taxa will be shown.
Enter organism name or id--completions will be suggested

Entrez query:

Expect Min: Expect Max:

Format for PSI-BLAST with inclusion threshold:

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

Experiment #5: Pack-MULEs - DNA transposons that reorganize the genome by capturing fragments of host genes. November 14, 2007

Overview: In this experiment you will learn how to search for and analyze Pack-MULEs - a very special TE that was discovered in the Wessler lab by Dr. Ning Jiang, who is now a Professor at Michigan State University. In addition to putting together the protocol for this experiment, Dr. Jiang will join us in a short Skype call where she will tell us "the narrative" behind her discovery of these elements and answer your questions.

In this experiment you will follow these 5 steps:

Step 1: Use Pack-MULE TIR sequences as query to identify other elements in the rice genome (Blast).

Step 2: Identify complete Pack-MULE elements (using a modification of the protocol used to identify complete LTR retrotransposon).

Step 3: Analyzing the complete Pack-MULE

Step 4: Annotate the internal sequence of your Pack MULE.

Step 5: Find the rice gene that was the source of the captured fragment.

Step 1. (Blast search) Use Pack-MULE TIR sequences as query to identify other elements in the rice genome:

Pack MULE TIRs Query:

>Os0037

```
GGAAAAAGTACACCGAAGGTCCCTCAACTTGTCATCGAGTTACAAAATCG
TCCTCCAACCGCAAAACCAGATACATGGCGTCCCTCAACTTACAAAACCG
TTCACATTAGGTCTTCGGTGGTTTTGACCCGGTTTTATCCGACGTGGC
GGCTGAGTCAGCGTGGGACCCACGTGGGCCCCACATGTCAGGATGCCACG
TCATCTCTCTTTCCCTCTCCCTTCCCTCCTCTCTCTCTCACTTC
TCTCCTCTCT
```

>Os0053

```
GGCAAATTTTGCTACAGGACACTGTGACTTTGCGGTTTTAGCTGTAGGAC
ACCGCCCGAAGTCACTTTTGGGGAAAACACTCCCTAAAGTTGGTAATTT
GCTGGTGGACACCGCGCCATTAAAATAATAATTTCCGGTTGAAGAGGAG
AGAGAAATCGCGTGAAATGTCAAAAATGCCCTTGGACCCACATGTCAGCT
CTCCTATCTCTTTCTCTCTATCCCTCCTTC
```

>Os0949

```
GGAATAAGTTCACCTGCCGTCCCTCAACTTTACGTCGAGTTTGTATGACA
TCCCTTATCTCCAATACCAGAAATCTTCAACCCCTAAACTATACAAAACCG
TGCAATTTAGGTCTATAGCAGTATGGATCTCTGGTTTTCGCTGACGTGGC
ATCCTAGTTAGCTAAAAAAAAAAAAAAAAAATATGGGGCCACATGTAAGTGA
GAAGAAAAAAGTTTTTTC
```

BLAST search. Go to the NCBI homepage click on BLAST and choose nucleotide blast.

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
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

Copy and paste each of the TIR sequences in "Enter Query Sequence".

Enter Query Sequence

Enter accession number, gi, or FASTA sequence Clear

```
>Os0037
GGAAAAAGTACACCGAAGGTCCCTCAACTTGTCATCGAGTTACAAAATCG
TCCTCCAACCGCAAACAGATAACATGGCGTCCCTCAACTTACAAAACCG
TTCACATTAGGTCCTTCGGTGGTTTGGACCCCGGTTTATCCGACGTGGC
GGCTGAGTCAGCGTGGGACCCACGTGGGCCCCACATGTCAGGATGCCACG
```

Query subrange

Select database "Nucleotide collection (nr/nt)" and choose *Oryza sativa* (taxid: 4530) as Organism, using "highly similar sequences" option.

Choose Search Set

Database
 Human genomic + transcript Mouse genomic + transcript Others (nr etc.):
 Nucleotide collection (nr/nt)

Organism
 Optional
 Oryza sativa (taxid:4530)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query
 Optional
Enter an Entrez query to limit search

Program Selection

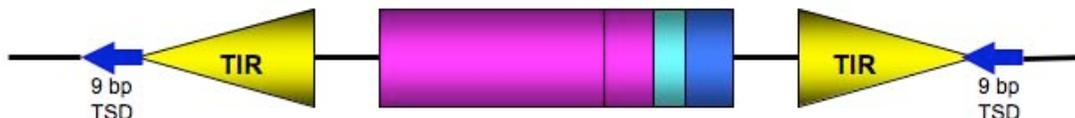
Optimize for
 Highly similar sequences (megablast)
 More dissimilar sequences (discontinuous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm

Click "BLAST" and wait for results. Multiple hits will show up in the window.

IMPORTANT NOTE: Only pay attention to the hits derived from BAC sequences (accession number starts with AP, AC, AL, BX, but exclude the pseudomolecules with numbers like AP008XXX)

Step 2. Identifying a complete Pack-MULE.

The goal is to find a BAC that has 2 blast hits near each other (less than 3 kb between them) but in opposite orientation. What this means is that one of your TIR hits will be the same orientation as your query and will be designated as "strand Plus/Plus, while the second TIR nearby will be in the opposite orientation "Strand = Plus/Minus".



This is illustrated below where the **orientation** is circled in red; the **location numbers** are circled in blue: from "143162" to "144199", therefore the distance is "144199 - 143162 = 1037 (bps)".

```
>dbj|AP002866.1| Oryza sativa (japonica cultivar-group) genomic DNA, chromosome
1, PAC clone:P0410E01
Length=166753

Sort alignments for this sub
E value  Score  Percent id
Query start position  Subj

Score = 291 bits (157), Expect = 6e-77
Identities = 229/262 (87%), Gaps = 12/262 (4%)
Strand=Plus/Plus
Query 1      GGAAAAAGTACACCGAAGGTCCCTCAACTTGTTCATCGAGTTACAAAATCGTCTCCCAACC 60
Sbjct 143162  GGAAAAAGTACACCGAAGGTCCCTCAACTTGTTCATCGGGATAAAAAAGCGTCTCCGAACC 143221
Query 61     GCAAAAACAGATACATGGCGTCCCTCAACT-TACAAAACCGTTACACATTAGTTCCTTCGG 119
Sbjct 143222  GCAAAAACAGATATATGGGGTCCCTTAACTATACAAAAACCGATCACCCGAGATCCTTCGG 143281
Query 120    TGGTTTTGACCCCGGTTTTA-TCCGACGTGGCGGCTGAGTCAGCGTGGGACCCACGTGGG 178
Sbjct 143282  TGGTTTTGACTCCAGTTTGGTCT-ACGGGGCGGCTGAGTCAGCGTGGGACCCACGTGGG 143340
Query 179    CCCCACATGTCAGGATG-CCACGTCACTCTCTTTCCCTC--CTCT--CCCTTCCTCCT 233
Sbjct 143341  TCCCACATGTCAGG-TGTCCAGTCATCTCTCTTTCCCTCTTTCTCTTTCCCTTCCTCC- 143398
Query 234    CCTCTCT-CTCTCACTTCTCTC 254
Sbjct 143399  CCTCTCTGCTCTCTCT-CTCTC 143419

Score = 255 bits (138), Expect = 2e-66
Identities = 220/258 (85%), Gaps = 12/258 (4%)
Strand=Plus/Minus
Query 2      GAAAAAGTACACCGAAGGTCCCTCAACTTGTTCATCGAGTTACAAAATCGTCTCCCAACCG 61
Sbjct 144199  GAAAAAGTACACCGAAGGTCCCTCAACTTGTTCATCGGGATA-AAAACGTCCTTGAACCG 144141
Query 62     CAAAACAGATACATGGCGTCCCTCAACT-TACAAAACCGTTACACATTAGGTCCTTCGGT 120
Sbjct 144140  CAAAACAGATATACGGGGTCCCTTAAATATATAAAAACCGGTCACCCGAGGTCCTTCGGT 144081
Query 121    GGTTTTGACCCCGGTTTT-ATCCGACGTGGCGGCTGAGTCAGCGTGGGACCCACGTGGGC 179
Sbjct 144080  GGTTTTGACTCCGGTTTTGGT-CTACATGGCGGCTGAGTCAGCGTGGGACCCACGTGGGT 144022
Query 180    CCCACATGTCAGGATG-CCACGTCACTCTCTTTCCCTCCT--CT--CCCTTCCTCCTC 234
Sbjct 144021  CCCACATGTCAGG-TGTCCACATCATCTCTTTACTCTCTTATCTTTCCCTTCCTCC-C 143964
Query 235    CTCTCTCTCTCACTTCTC 252
Sbjct 143963  CTCTATCTCTCTCT-CTC 143947
```

Step 3: Analyzing the complete Pack-MULE

Determining TIR similarity:

Recall that for an LTR retrotransposon, the LTRs of a single element are identical when the element inserts and that they "drift" (accumulate mutations) over (evolutionary) time. The same is true for DNA transposons, that is, the TIRs are identical when the element inserts. Thus we can distinguish old insertion events from recent ones by comparing the TIRs of a single element. This is how you do this...

Go to BLAST and select "Align two sequences using BLAST (bl2seq)"

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two sequences using BLAST (bl2seq)

--put the BAC name into the windows "Sequence 1" and "Sequence 2"

--put the location numbers of both TIRs into the windows of "from" and "to"

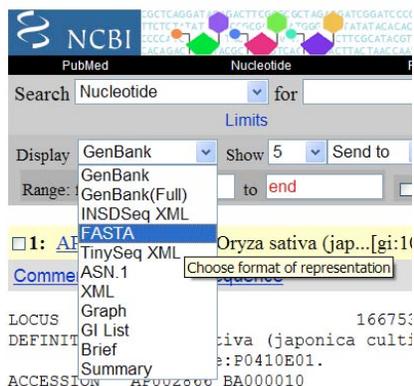
--uncheck the "Filter" option and click "Align".

gap_x_dropoff expect word size Filter

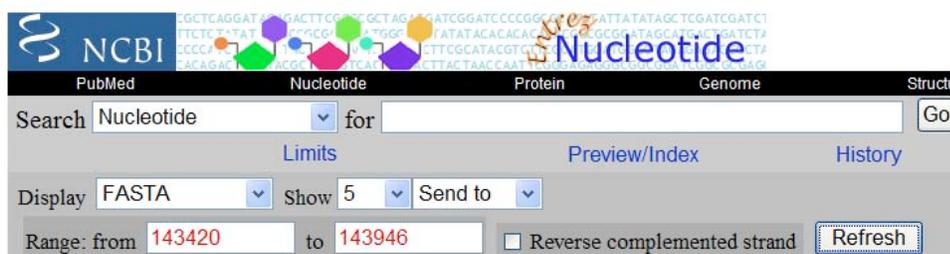
Sequence 1
 Enter accession, GI or sequence in FASTA format from: to:
 AP002866
 or upload FASTA file

Sequence 2
 Enter accession, GI or sequence in FASTA format from: to:
 AP002866
 or upload FASTA file

Select "FASTA" format in "Display" menu.



Input the beginning (143420) and end locations (143946) into the "from" and "to" windows and click "refresh".



Copy and save the internal sequence.

```

1: AP002866. Reports Oryza sativa (jap...[gi:10179051]
>gi|10179051:143420-143946 Oryza sativa (japonica cultivar-group) genomic
GCAGGCAGGGCCGATGGTGAACGGGTTGAGCAGCGGCACTGCAACGGCGGGTGGCCGCGCACACAGCC
ATGGACGACGTTGAAGTGGTGAAGCGTACTTATCGGCCAGTACTACAACGGGCTGGTGCTAGAGCTCGTGG
TACACCGCTCTCGCCCGCCAACGCCGCTCTCCGCAGCACCAGCACCAGCCAGTCATAGTACTAGCCGAT
AAGTGCACCTACCACTTCAACGCCGTCATGGCTGTGTGTCCTGGTGGCTTCGGCTCATGGACTTCGCCG
GCGTTGGTCTTACGACCGCGCGTCCGGTGTCTCCCCGCTGAGCGGTCGACGAGGCGCGTGGCCTCGG
CGAGGTCGTCGCGGGCGATGAGGCGGCAGAGCCTGGCATTGGGGGACGACGCAATGGGGGCTCGGGCCT
CGTGGCGGCAGCGGCGAGGTAGGCGTGGTGTGCGCGGCGGAGTCGCTGTGCGCCCGGACCACCGTCGCA
GTGCCGCTGCTCAACCTGCTCACCATCGGCCCTGCC
  
```

Step 5: Find the rice gene that was the source of the captured fragment

To do this you need to Blast the internal sequence you have saved against the rice database to find all related sequences. However, we do not want ALL of the related sequences because some will be our original Pack-MULE while other will be additional copies of the Pack-MULE in the rice genome. The procedures below will remove these sequences so that you will be left with the rice gene.

Go to BLAST webpage again and select "nucleotide blast"

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast</i>
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

--Paste the internal sequence of your Pack-MULE into the query window
--select nr database and choose *Oryza sativa* (taxid:4530).

Choose Search Set

Database
 Human genomic + transcript Mouse genomic + transcript Others (nr etc.):

Organism
Optional

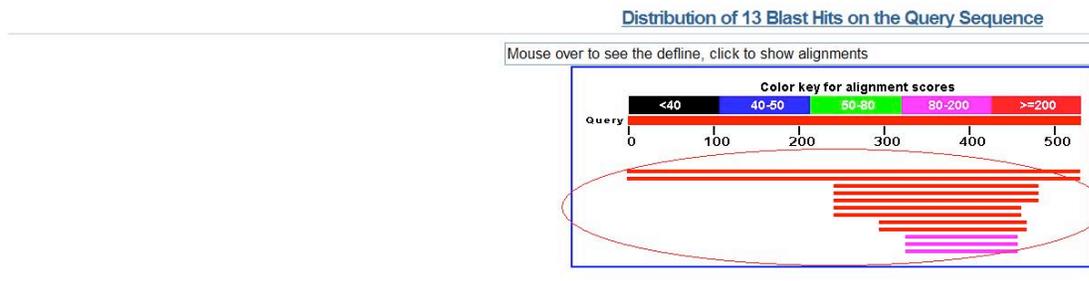
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query
Optional

Enter an Entrez query to limit search

Click BLAST.

Multiple hits will appear on the result page. To find out the detailed information of each hit, you can either left-click the "color bar" on the figure or the "Max score" in the table as follows:



[Distance tree of results](#) NEW

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score
AP008207.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 1	974	1134
AP002866.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 1, P	974	974
CR855046.1	Oryza sativa genomic DNA, chromosome 4, BAC clone: OSIGBa012530	272	272
AP008210.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 4	272	272
AL731597.3	Oryza sativa genomic DNA, chromosome 4, BAC clone: OSJNBa002330	272	272
AC148759.2	Oryza sativa Japonica Group chromosome 11 clone OSJNBa0030K09, c	254	254
AP008217.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 11	254	254
AP008208.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 2	239	239
AP005536.3	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 2, B	239	239
NM_001050986.1	Oryza sativa (japonica cultivar-group) Os01q0783100 (Os01q0783100)	159	159
AP003368.2	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 1, B	159	159
AK067606.1	Oryza sativa (japonica cultivar-group) cDNA clone:J013112M07, full ins	159	159

NOTE: Ignore matches with an E value lower than -10 (low score hits) or with 100% similarity (self hits). You should also exclude "mRNA" or "cDNA", and pseudomolecules (with number **AP008XXX**). Only pay attention to the hits derived from BACs (Accession number starts with AP, AC, AL, BX).

For example, after clicking the first hit's score (AP008207), we will see the alignment result. It is from a pseudomolecule (chrom 1, length=43261740) and it is also a self hit (100% identity), therefore we should skip this hit.

[db|AP008207.1](#) [Oryza sativa \(japonica cultivar-group\) genomic DNA, chromosome 1](#)
length=43261740

Sort alignments for this subject
E value [Score](#) [Percent ident](#)
[Query start position](#) [Subject](#)

Features flanking this part of subject sequence:
827 bp at 5' side: [Os01q0615300](#)
6538 bp at 3' side: [Os01q0615500](#)

Score = 974 bits (527), Expect = 0.0
Identities = 527/527 (100%), Gaps = 0/527 (0%)
Strand=Plus/Plus

```
Query 1 GCAGGCAGGGCCGATGGTGAACGGGTTGAGCAGCGGCCTGCAACGGCGCGGTGGCCGC 60
      |||
Sbjct 24397516 GCAGGCAGGGCCGATGGTGAACGGGTTGAGCAGCGGCCTGCAACGGCGCGGTGGCCGC 24397575
```



```

> emb|AL731597.3|OSJN00237 ■ Oryza sativa genomic DNA, chromosome 4, BAC clone:
complete sequence
Length=175555

Score = 272 bits (147), Expect = 5e-71
Identities = 223/256 (87%), Gaps = 19/256 (7%)
Strand=Plus/Minus

Query 241  GGCTGTGTGTGCCTGGTGCTTCGGCTCATGGACTTCGCCGGCGTTGGTCTTCA----- 293
          |||
Sbjct 81907 GGCTGTGTGCGCCTGGTGCTTCGACTCACGGACTTCGCCGGCGTTGGTCTTACGGCGGT 81848

Query 294  --CGACGGCGGGCGTCGGGTGTCTCCCCGCGTGAGCGGTTCGACGAGGCGCGTGGCCTCGGC 351
          |||
Sbjct 81847  GGCGACGGCGGGCGTCGGGTGCCTCCCCGCGTGAGCGGTTCGACGAGGCGCGCGGCCTCGGC 81788

Query 352  GAGGTCGTTCGCGGGCGATG-AGGCGGCAGA----G--CCT-GG-CATTGGGGGACGACGC 402
          |||
Sbjct 81787  GAGGTCGTTCGTGGGCGATGCAGGCGGGCGACCTCGTGCCTCGGTGTTGGGGGACGACGC 81728

Query 403  AATGGGGGCCTCGGGCCTCGTGGCGGCAGCGGCGAGGTAGGCGTGGTGTGCGCGGCGGCA 462
          |||
Sbjct 81727  GCTGGGGGCCTCGGGCCTCGTGGCGGCAGCGGCGAGGTGGGCGTGGTGTGCGCGGCGGCC 81668

Query 463  GTCGCTGTGCGCCGCCG 478
          |||
Sbjct 81667  GTCGCGAG-CGCCGCCG 81653

```

Go to the webpage of BLAST and select "Align two sequences using BLAST"

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector contamination](#) (vecscreen)
- [Align](#) two sequences using BLAST (bl2seq)

Now it is your turn.

Since most Pack-MULEs are very complex. Dr. Jiang has picked several good examples for us to use in this experiment.

To do this, we have divided the class into 3 groups and assigned the following BAC to each group.

Group 1:

Renee

Martha

BAC **AC096855** (Chr03)

Group 2:

Cathy

Ian

Wren

BAC **AP003682** (Chr06)

Group 3:

Erin

Caroline

Jordan

BAC **AL732641** (Chr12)

First, each group will use the Pack-MULE TIRs as query to perform a BLAST search as shown at the beginning of today's handout, then pick your assigned BAC from the blast result and continue with the steps in the handout. Good luck!

Experiment #6: Touchdown PCR using Degenerate Primers November 20, 2007

"The identification of novel members of gene families by PCR using degenerate primers has been considered more of an art than a science." Michael Koelle

Overview: In previous experiment, you learned how to find TEs in genomic sequence databases through BLAST searches. However, if only partial genome sequence is available (e.g. for maize), only some of the TEs in a genome can be identified. In today's experiment you will perform PCR with degenerate primers in order to isolate parts of TE family member that may not be in the database. You will be able to submit your sequence to GenBank and become a (very small) part of science history!

There are 6 main steps in this experiment:

Step 1: Design degenerate primers.

Step 2: Touchdown PCR.

Step 3: Gel electrophoresis and DNA purification from bands.

Step 4: TOPO TA cloning.

Step 5: Plasmid purification from bacteria

Step 6: Sequencing and sequence analysis

Because of time constraints, you will only be doing steps 2 and 6. Your highly skilled TA, Yujun will handle the other steps. Because you have already performed steps 3, 4, 5 and 6 in prior experiments, we only need to go over what is meant by degenerate primers and touchdown PCR.

Step 1: Degenerate primers - what they are and how they are designed

Rationale: Think about what we need to do - we are going to use PCR to amplify DNA that has yet to be sequenced. In the past when we have done PCR (for the mPing and Osmar5NA excision sites in the GFP reporter gene) we have used PCR primers whose sequences were derived from existing sequences - that is - the GFP gene flanking the TE insertion sites. Those primers were 22 nt long and each was only a single sequence. Here you will see how we design primers to amplify many different but related DNA templates. The important point to keep in mind is that a degenerate primer is actually a collection of many related primers.

Designing degenerate primers: The basic idea is to first use multiple alignments of sequences from the database to identify two conserved domains and then to design degenerate PCR primers based on the consensus sequences of these regions such that PCR products extend from one conserved domain to the other. This strategy is diagrammed in the figure below....

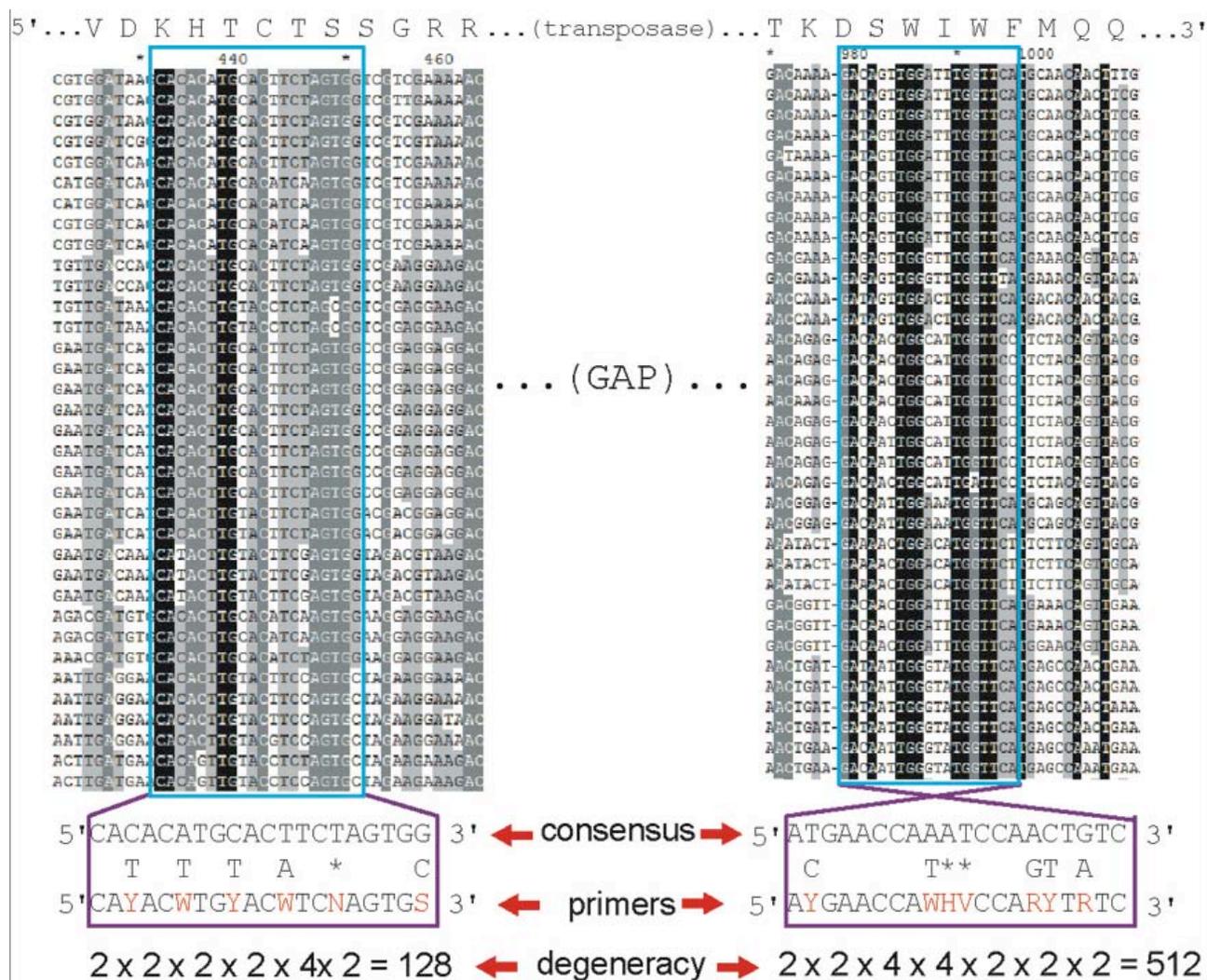


Figure - The conserved amino acid residues in two domains of the transposase are shown. Below are the multiple alignments derived from the nucleotide sequences of individual TEs from, for example, a particular organism (like maize). These alignments are used to derive a consensus sequence. Note the some positions have a poor consensus and in these cases (marked by an *) 3 or 4 nucleotides are used in the primer).

Standard MixBase Definitions:

(Use this to understand the primers and the degeneracy in the previous figure)

Definition	MixBase	degeneracy
R	A, G	2
Y	C, T	2
M	A, C	2
K	G, T	2
S	C, G	2
W	A, T	2
H	A, C, T	3
B	C, G, T	3
V	A, C, G	3
D	A, G, T	3
N	A, C, G, T	4

Step 2: Touchdown PCR: The use of degenerate primers requires modifications in the PCR cycling conditions. This modification is called Touchdown PCR.

One side effect of using degenerate primers for PCR is a higher probability of annealing to the wrong sequence (because the degenerate primer may not be a perfect match with the template sequence). Touchdown PCR is a modification of conventional PCR that is designed to reduce the occurrence of nonspecific amplification. To do this we use an annealing temperature that is higher than optimum in early PCR cycles. The annealing temperature is decreased by 1°C every cycle or every second cycle until a specified or 'touchdown' annealing temperature is reached. The touchdown temperature is then used for the remaining number of cycles. This allows for the enrichment of the correct product over any non-specific product.

Groups:

CACTA: Renee, Wren

PIF/Harbinger: Ian, Cathy

hAT: Erin, Caroline

Mutator: Martha, Jordan

For each group:

1. Prepare two 1.5 ml tubes: label (on the cover) both with your TE's name and one with a T (teosinte) and the other with M (maize).
2. Add the following to each tube:

Mg ²⁺	5	ul
buffer	5	ul
dNTP	5	ul
primer_L	2.5	ul
primer_R	2.5	ul
DNA	2.5	ul
RT-TAG	0.5	ul
water	27	ul
(total:	50	ul)

Mix thoroughly and give the tube to Yujun. PCR will be completed after you leave class. The PCR products will be stored in the frig (at 4°C) and Yujun will take care of the following steps. Yujun will set the PCR program as following for you:

1 cycle for: Initial Denaturation 94°C 3 minutes

1 cycle for: Denaturation 94°C 1 minute
Annealing 57°C 1 minute
Extension 72°C 1 minute

1 cycle for: Denaturation 94°C 1 minute
Annealing 56°C 1 minute
Extension 72°C 1 minute

1 cycle for: Denaturation 94°C 1 minute
Annealing 55°C 1 minute
Extension 72°C 1 minute

1 cycle for: Denaturation 94°C 1 minute
Annealing 54°C 1 minute
Extension 72°C 1 minute

1 cycle for: Denaturation 94°C 1 minute
Annealing 53°C 1 minute
Extension 72°C 1 minute

30 cycles for: Denaturation 94°C 1 minute
Annealing 52°C 1 minute
Extension 72°C 1 minutes

Final extension: 72°C 10 minutes

You will get the sequencing results to analyze next week.

Experiment #6.2: Data analysis (November 20, 2007)

Overview: In our previous experiment, we have set up Touchdown PCR reactions using degenerate PCR primers of each of your TE families and either Maize or Teosinte genomic DNA as templates. Today each group will draw two phylogenetic trees using the knowledge you have learned in experiment 3. One tree is based on the trimmed old protein sequences (please see figure 1), and the other is based on the combination data set of both the trimmed old data and the new sequencing data that were from your degenerate PCR reactions. Both two sets of data were prepared for you. You can start with them directly with clustalw.

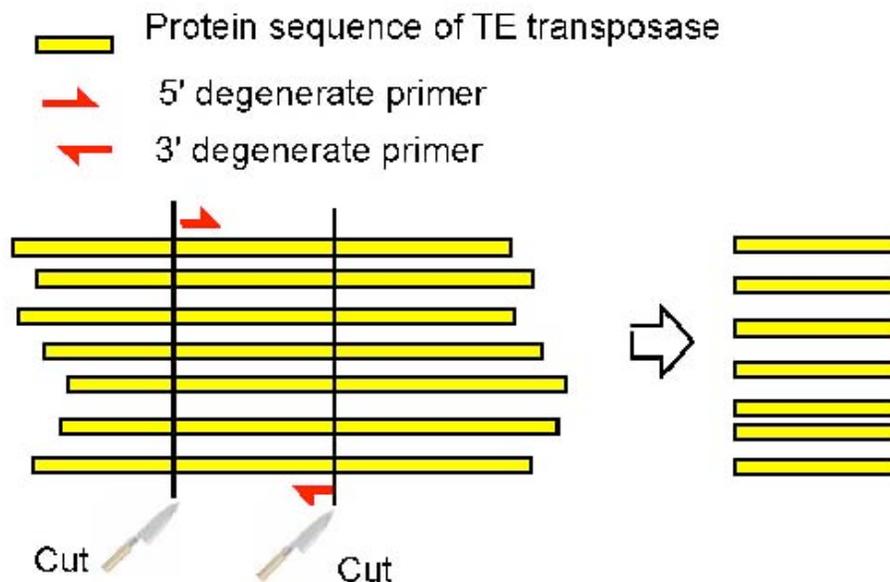


Figure 1. The protein sequences that you have used for building phylogenetic trees in Experiment2 need to be trimmed based on the location of the degenerate primers. The reason is if we use unmatched sequences to do multiple alignment then build the tree, there will be significant difference between the branch length between the long and short sequences, although they may share high similarity.

The data that you will start with was prepared as follows:

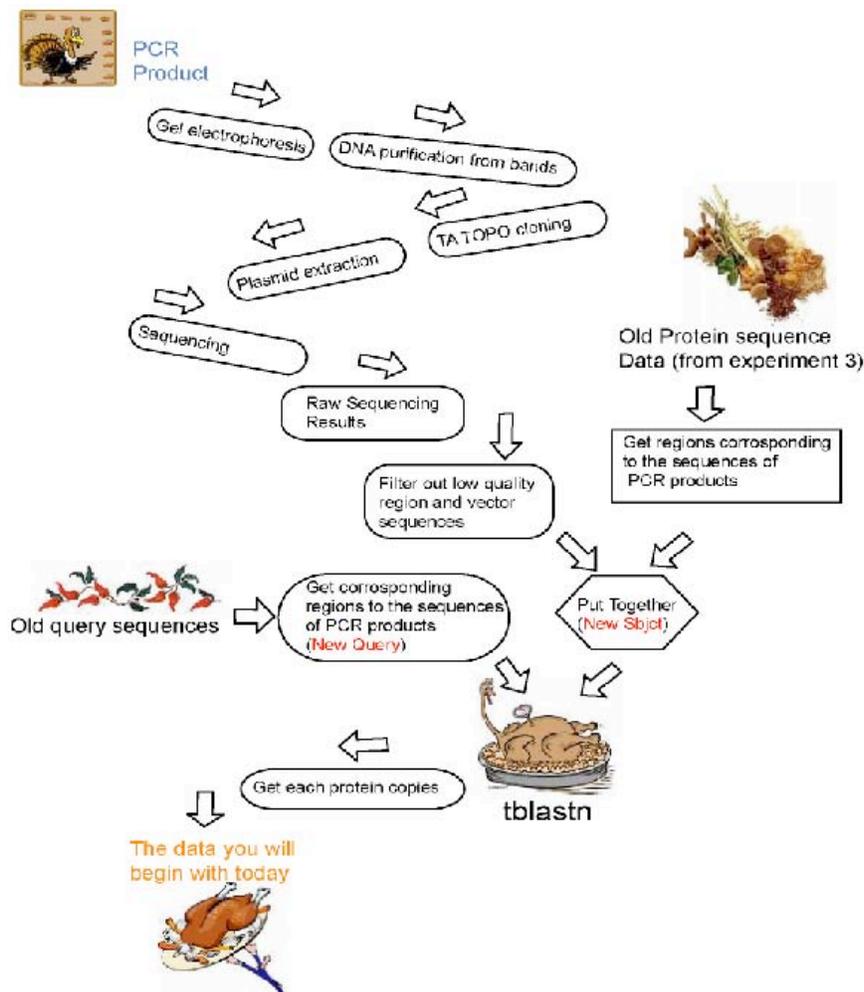


Figure 2. What happened to your PCR products.

This experiment has 4 major steps:

Step 1: Multiple alignment using Clustalw.

Step 2: Drawing trees using PAUP.

Step 3: Tree analysis.

After your trees are constructed (with and without PCR data), they need to be combined. In the combined tree, the names of PCR sequences that were from Maize DNA begin with "Maize", while the sequences from teosinte begin with "Teosinte", and the remaining (old) sequence names begin with "ZM". You need to answer the following questions:

Step 4: Data submission.

Yujun will show you how to submit your PCR sequencing results to ncbi.