Lecture 19
EUKARYOTIC GENES AND GENOMES I

For the last several lectures we have been looking at how one can manipulate prokaryotic genomes and how prokaryotic genes are regulated. In the next several lectures we will be considering eukaryotic genes and genomes, and considering how model eukaryotic organisms are used to study eukaryotic gene function. During the course of the next six lectures we will think about genes and genomes of some commonly used model organisms, the yeast *Saccharomyces cerevisiae* and the mouse *Mus musculus.* But first let's look how the genes and genomes of these organisms compare to *E. coli* at one extreme, and humans at the other.

genome = DNA content of a complete haploid set of chromosomes
= DNA content of a gamete (sperm or egg)

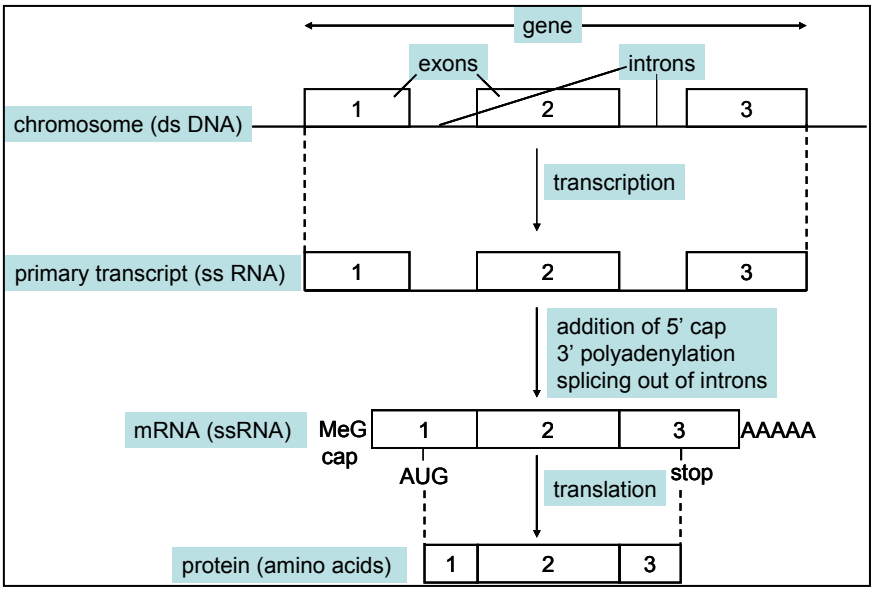
Species	Chromosomes	cM	DNA content/ haploid (Mb)	year sequence completed	genes/ haploid	genes have introns?
E. coli	1	N/A	5	1997	4,200	no
S. cerevisiae	16	4000	12	1997	5,800	rarely
C. elegans	6	300	100	1998	19,000	nearly all
D. melanogaster	4	280	180	2000	14,000	nearly all
M. musculus	20	1700	3000	2002 draft 2005 finished?	22,500?	nearly all
H. sapiens	23	3300	3000	2001 draft 2003 finished	22,500?	nearly all

Note: cM = centi Morgan = 1% recombination
Mb = megabase = 1 million base-pairs of DNA
Kb = kilobase = 1 thousand base-pairs of DNA

Let's think about the number of genes in an organism and the size of the organism's genome. The average protein is about 300 amino acids long, requiring 300 triplet codons, or roughly 1Kb of DNA. Thus it makes sense that to encode 4,200 genes *E. coli* requires a genome of 5 million base pairs. However, the human genome encodes about 22,500 proteins, and this should require a genome of lets say 25 million base pairs. Instead, humans have a genome that is ~ 3000 million base pairs, or ~ 3,000 Mb, i.e., ~ 3 billion base pairs. In other words, there is about 100-fold more DNA in the human genome than is required for encoding 22,500 proteins. What is it all doing? Some of it constitutes promoters upstream of each gene, some is structural DNA around centromeres

and telomeres (the end of chromosomes, some is simply intergenic regions (non-coding regions between genes) but much of it is present as **introns**.

What does it mean "Genes Have Introns". This represents one of the fundamental organizational differences between prokaryotic and eukaryotic genes. Eukaryotic genes turn out to be **int**errupted with long DNA sequences that do not encode for protein…these "**int**ervening sequences" are called **introns.**



The DNA segments that are ultimately **ex**pressed as protein, i.e., the DNA sequence that contains triplet codon information, are called **exons.** The intronic sequences are removed from the primary transcript by **splicing.**

A major consequence of this arrangement is the potential for **alternative splicing** to produce different proteins species from the same gene and primary transcript. This gives the potential for tremendous amplification of the complexity of mammals (and other eukaryotes) through many more thousands of possible proteins.

Note that lower eukaryotes such as the yeast *S. cerevisiae* only have ~ 5% of their genes interrupted by introns, but for multicellular organisms, like humans, >90% of all genes are interrupted by anywhere between 2 and 60 introns, but most genes have between 5 and 12 introns.

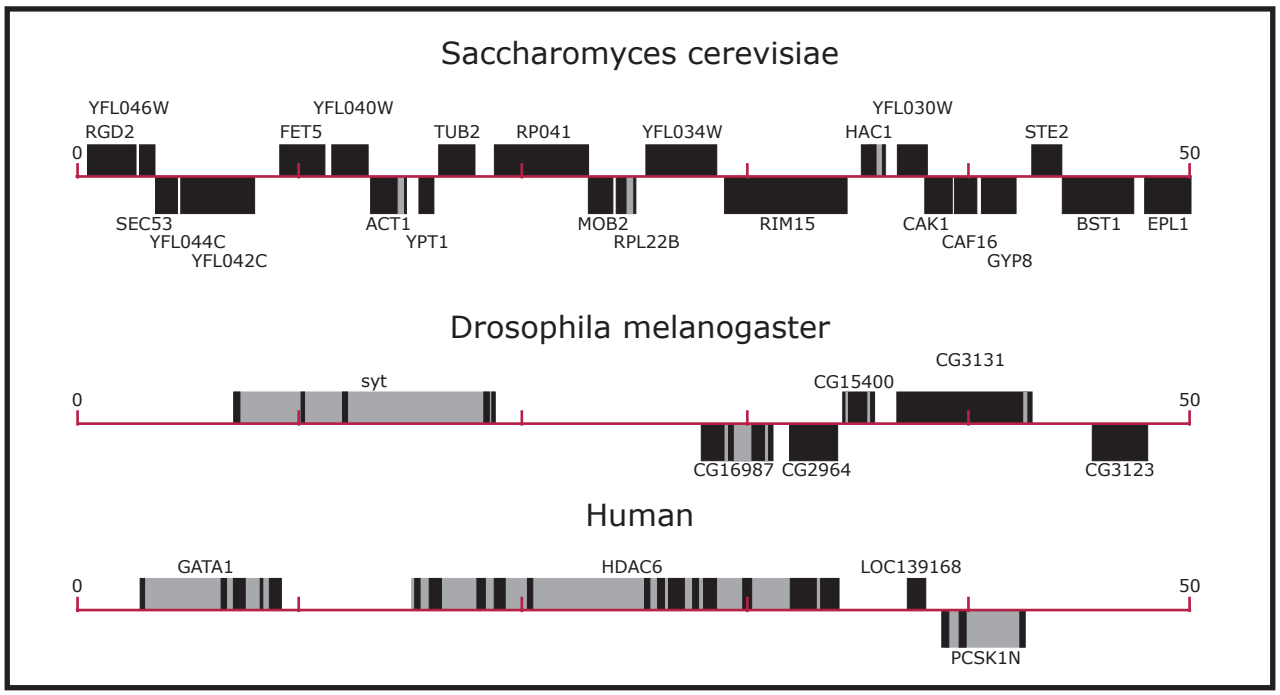
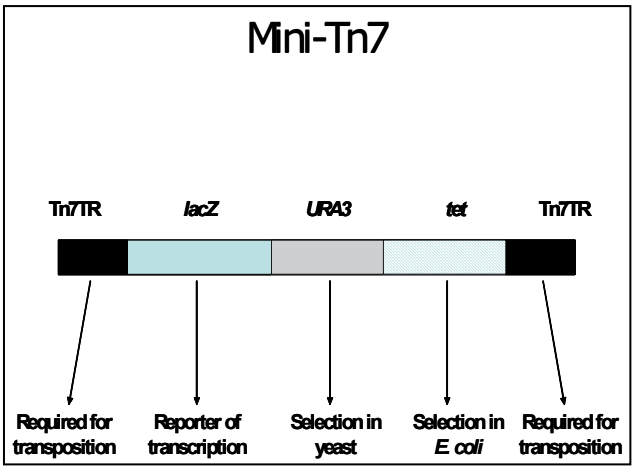


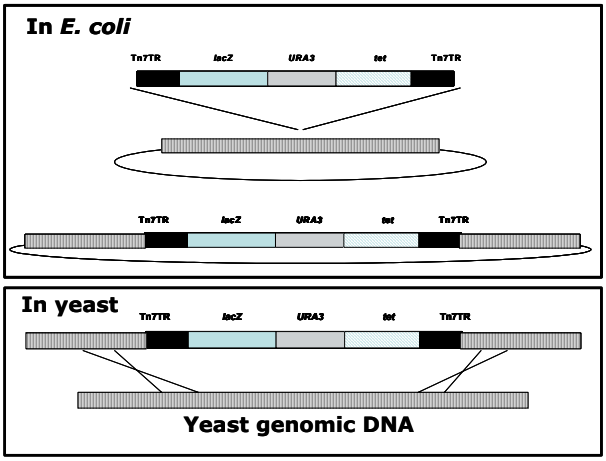
Figure by MIT OCW.

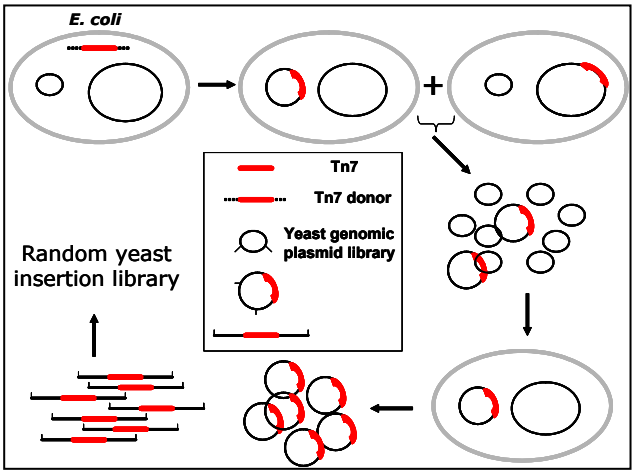
Gene Regulation in Yeast

In the next few lectures we will consider how eukaryotic genes and genomes can be manipulated and studied, and we will begin with an example of examining how genes are regulated in *S. cerevisiae.* First, let's figure out how to use some neat genetics to identify some regulated genes, and in the next lecture we will figure out how one can use genetics to dissect the mechanism of that regulation.

Characterizing function and regulation of *S. cerevisiae* genes: We are going to combine a few neat genetic tools that you learned about in Prof. Kaiser's lectures for this, namely a **library** of yeast genomic fragments cloned into a bacterial **plasmid**, a modified **transposon (mini-Tn7)**, and the ***lacZ*** gene embedded within the transposon. In this experiment the *lacZ* gene is going to be used as a **reporter** for transcriptional activity of yeast genes.





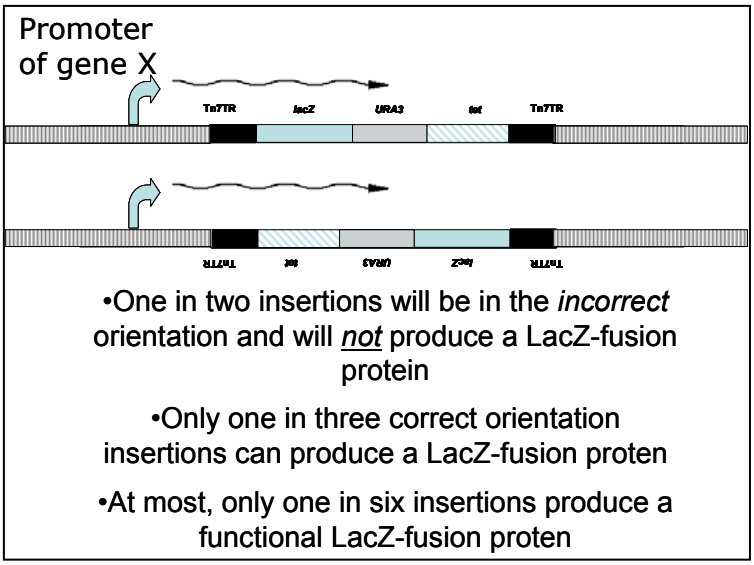


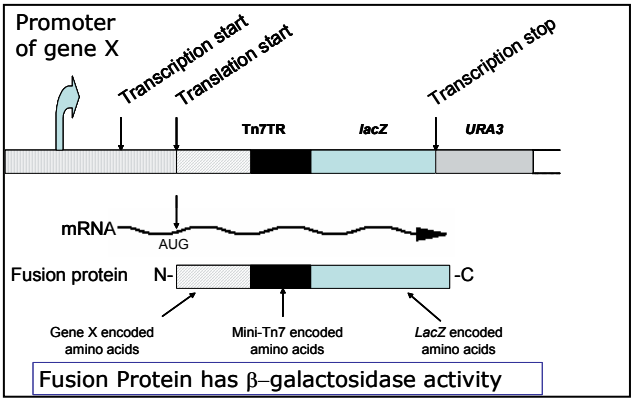
The mini-Tn7 is introduced into a population of *E.coli* that harbor a plasmid library of the *S. cerevisiae* genome; i.e., each *E. coli* cell is home to a plasmid that contains a different segment of the *S. cerevisiae* genome, such that the whole geneome is represented many times over in this population of *E. coli*. The mini-Tn7 is allowed to transpose by integrating into either the plasmid DNA or the bacterial DNA; the original DNA that carries the mini-Tn7 can not replicate, but cells that have integrated the mini-Tn7 into the plasmid or *E.* coli chromosome are selected as Tetracycline resistant colonies. Plasmid DNA is purified from these transformants and retransformed into tetracycline sensitive *E. coli*; the resulting tetracycline resistant bacteria harbor only plasmids that have an integrated mini-Tn7 transposon. Plasmid is isolated

from these cells and the yeast genomic fragments are isolated by digestion with an appropriate restriction enzyme.

So now we have a library of yeast genomic fragments each of which has the transposon inserted; these genomic fragments can be transformed into *S. cerevisiae* cells that are *ura3*-. Each Ura+ transformant colony will have recombined a Tn7 transposon-containing genomic DNA into its genome. This essentially gives us a **library of yeast with transposons randomly integrated into it genome.**

Note that the *lacZ* gene in the transposon does not carry its own transcription or a translation start site, but if the transposon inserts in the correct orientation downstream of a yeast gene promoter, and in the correct triplet codon reading frame, the *lacZ* gene comes under the control of that promoter and when transcription is activated from that promoter a LacZ-fusion protein is expressed, and most LacZ-fusion proteins display robust **β-galactosidase activity**.





Yeast cells expressing **β-galactosidase** activity can easily be detected by growth in the presence of **5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside**, better known as **X-gal**. LacZ cleaves X-gal to release a chemical moiety that has a brilliant blue color…and so the colonies turn bright blue!

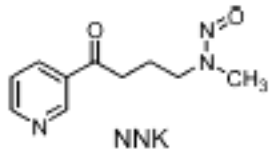
There are at least two useful things to come out of such a collection of yeast strains:

(1) Any transposon that integrated into a gene will essentially disrupt that gene and is likely to cause a null mutation.

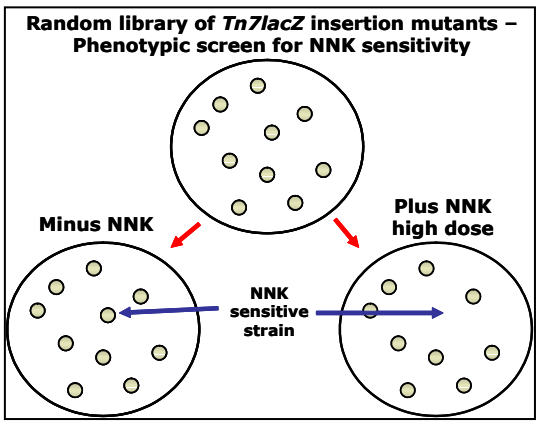
(2) For transposons that integrate into a yeast gene such that the *lacZ* gene is in frame with the genes coding region, the level of β-galactosidase activity in these cells therefore becomes a **reporter** for the transcription of that gene.

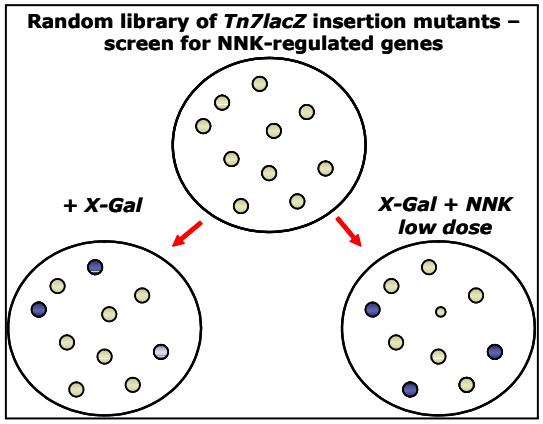
Here are just two examples of how such a library can be used: (1) to identify genes that protect cells against a DNA damaging agent that causes cancer; lets take the example of one of the many many compounds found in tobacco smoke; and (2) to identify genes whose transcription is up-regulated in response to being exposed to this tobacco smoke chemical.

The chemical we'll use as an example is 4-(Methylnitrosoamino)-1-(3-pyridyl)-1-butanone (NNK). The yeast random insertion library is first plated out so that individual cells give rise to a colony; these colonies are then replicated onto test plates. To screen the library for genes that protect against the cell killing that can be induced by NNK the colonies are replica plated onto agar medium that either does or does not contain a high dose of NNK. To screen the library for genes that are transcriptionally regulated in the presence of this nasty carcinogenic compound, the colonies are replica plated onto agar medium containing either X-gal alone or X-gal plus a low dose of NNK.



NNK





Interesting colonies can be retrieved from the master plate for further study and for identification (and subsequent cloning) of the gene responsible for the interesting phenotype.

Once we have identified a gene that is transcriptionally up or down regulated in response to an environmental change, how can we use genetics to figure out how regulation is achieved. This is the topic of the next lecture.