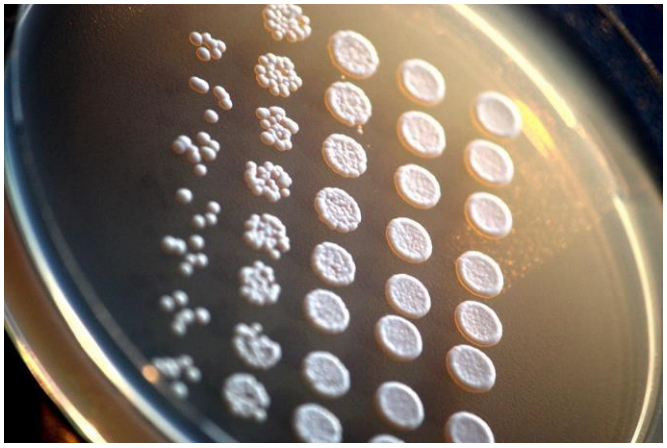


The Synthetic Yeast Project Sc2.0 developed synthetic yeast chromosomes with LoxPsym sites scattered throughout the chromosomes. By inducing the expression of Cre recombinase, new chromosome configurations can be generated in yeast cells. This population can then be screened for enhanced or altered phenotypes.

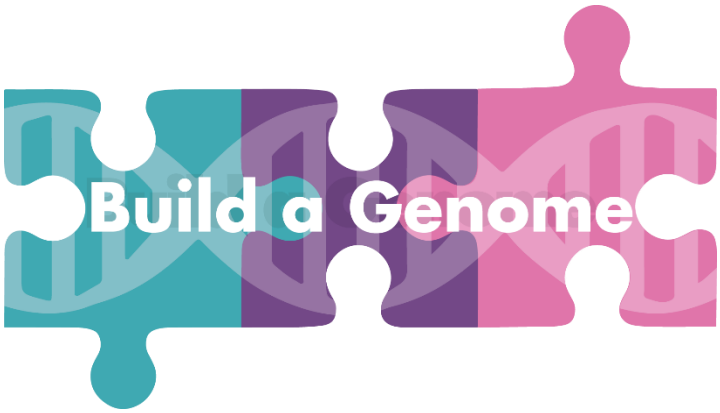
SCRaMbLE Lab Manual

v2

Synthetic Chromosome Recombination and Modification by LoxP-mediated Evolution

The Build-a-Genome Network

Mark Jones
John Mirarchi
Lisa Scheifele



Contents

Inducing Scramble and Selecting Cells

Inducing the Scramble system 3

Confirming Phenotypes

Measuring growth rates by serial dilution and plating 9

Extracting DNA and finding Genotype Changes

Extraction of Yeast DNA 12
Amplification of PCR Tags 14

Analysis of Illumina Sequencing Data

Error Correcton in HighThroughput Datasets 20
Identification of SNPs in Illumina Sequencing Data 27
Visualizing Sequence Variants 32

Supply List

Supply List 35
Equipment Needed 35
Media Recipes 36

Inducing the Scramble system

Starting Strain: Yeast cells containing synthetic chromosomes and the Cre plasmid (pRS413-CreEBD) (can also have one strain with the empty plasmid pRS413). The plasmid contains the Cre gene (remember that Cre induces the Scramble system) under the control of the estradiol promoter. In this experiment, you will add estradiol to the cells to induce the expression of Cre and therefore the Scramble system.

Preparation before lab: Inoculate a liquid culture in SC-Ura media and grow overnight at 30C (grow in an Erlenmeyer flask with at least 5 volumes of air per volume liquid). In the morning, measure the OD600, dilute to an OD of 0.2 into 200 ml SC-Ura media, and grow at 30 until mid-log phase.

Prepare selective plates. You want to pick a condition where most cells die or are very inhibited in their growth. 7% ethanol, 2mg/ml caffeine, a temperature of 40C, and 6% salt work well. Table 1 in this publication has additional options. The selective plates no longer need to be SC-Ura since the plasmid is no longer needed; instead, you can use YPD media. Typically, we use 150mm plates for selection since that gives a greater surface area and more colonies to select from.

In-lab:

1. Measure the optical density of the yeast culture at 600 nm, which gives you an estimate of the cell density.
    a. Transfer 1 ml of media to a cuvette. Place in the spectrophotometer and blank the instrument (this only needs to be done once for the class).
    b. Transfer 1 ml of your yeast culture to a cuvette. Place in spectrophotometer and record the optical density. If the density is greater than OD=1, dilute the cells and repeat the measurement. Remember to account for the dilution when entering the OD into the formula below
2. Use the formula $C_1V_1=C_2V_2$ to determine the volume of cells equal to an OD of 0.1

    (Concentration (OD as measured on spec)(Volume to be determined) = (0.1) (25 ml)

    Volume to be determined =_____ml
3. Into a sterile flask, add:
    a. The volume of cells that you calculated above
    b. 25 ml of SC-Ura liquid media
    c. 5 ul of beta-estradiol (1 uM final concentration; 5 mM stock).
4. Incubate the culture with shaking for 2-6 hours. (The cells are Scrambling now!) The time may need to be determined empirically-more time will produce more Scrambled cells but also more lethality.
5. Transfer 100 ul of the cell culture to a microcentrifuge tube.
6. Spin down the cells for 1 min at full speed. Remove the supernatant. Resuspend the culture in 1 ml sterile water.

7. Repeat step 7 two more times, ending with the cells resuspended in 1 ml sterile water.

8. Onto two selective plates, add 100 ul water. Into this drop, add 10 ul of cells from step 7. Spread the cells on the plates with glass beads (or a plastic spreader).

9. Discard beads and when the liquid has absorbed into the plates, incubate the plates upside down at 30C for 2 days.

Scramble Assignment

The Scramble system requires two steps: (1) introduction of the Scramble plasmids into yeast cells and selection of those plasmids using genetic markers, and then (2) induction of the Scramble system and selection of yeast cells with interesting properties. The questions below will address your understanding of those two steps. To answer these questions, use:

1. Use your class notes
2. The websites listed in the questions themselves
3. The Synthetic Yeast design paper
4. The Chapter 4: Working with Yeast article that was shared with you, and
5. The yeast transformation video: https://www.youtube.com/watch?v=izquCFtH5Ps

1. The yeast strain we are using has the genotype his3delta1 leu2delta0 met15delta0 ura3delta0. Go to: http://wiki.yeastgenome.org/index.php/Commonly_used_auxotrophic_markers and look up each of the four alleles and then complete the table below (you need to scroll all the way to the bottom of the page for the last 3 markers)

Allele	Is this marker a complete deletion of the gene?	If not, what portion of the gene is missing?
his3delta1		
leu2delta		
trp1delta		
ura3delta		

2. Each allele contains a hyperlink to the Saccharomyces Genome Database. Each of these alleles is an auxotrophic marker, meaning that yeast containing the allele requires a particular nutrient. Under "Name Description", determine what nutrient yeast containing each of the alleles requires.

Allele	Name of gene	Nutrient Required
his3delta1		
leu2delta		
trp1delta		
ura3delta		

3. For the Scramble experiment, the Cre gene was cloned into a plasmid containing the URA3 gene as its selectable marker, transformed into yeast, and selected on SC-Ura medium. Based on your results above, what three other genes (selectable markers) could have been used instead of *URA3* and what medium would you use for selection?

4. The article, Chapter 4: Working with Yeast, describes different types of yeast media. If you were to make one of the types of ***selective*** media described above (SC-Ura, for example), would this be rich media or defined media? How do you know?

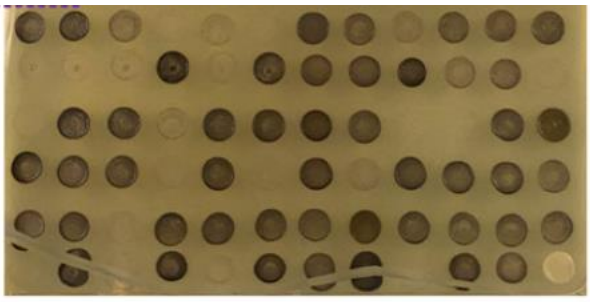
5. Why is it harder to introduce DNA into yeast cells, causing the heat shock step of transformation take 20 minutes for yeast cells but only 45 seconds for *E. coli* bacteria?

6. Where are the LoxP sites located in the yeast genome? What function do LoxP sites confer to the genome (in the presence of the Cre protein)?

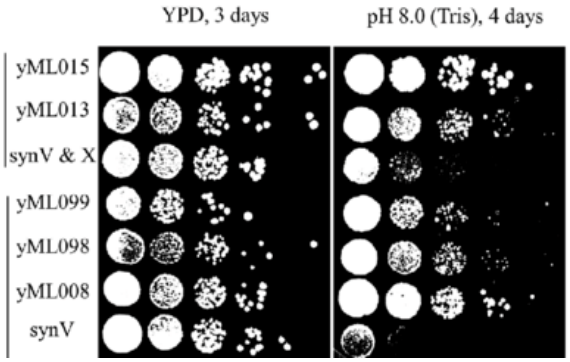
7. Watch the Transformation of Yeast video. What percent of genes that are the same between humans and yeast have an important role in human disease? This includes genes that are involved in cancer-what cellular process are many of these genes involved in (study of this process in yeast won the Nobel Prize in 2001)? Why are yeast also being used to screen new cancer drugs?

8. From the article, Chapter 4: Working with Yeast, describe 4 methods that are used to maintain sterility when working with yeast.

9. The figure below is the result of a Scramble experiment. The cells all express the pigment violacein (purple-black) and they all started off making the same amount of pigment. The Scramble experiment was performed multiple times and each spot represents the result of an independent Scramble experiment. Even though the experimental procedure was the same each time, why do the resulting cells produce different levels of pigment?



10. The figure below represents the results of a Scramble experiment. The strains labeled synV and syn V&X are the starting strains (not Scrambled) and they contain either synthetic chromosome 5 only or synthetic chromosomes 5 and 10. The strains are serially diluted and spotted onto rich media plates (YPD, left) and plates with alkaline media (pH 8.0, right).
    Strains yML015, yML013, yML099, yML098, and yML008 are five strains resulting from five independent Scramble experiments. They were diluted and plated in a similar manner.
    a) Which of the strains show the greatest improvement in its tolerance for alkaline conditions following the Scramble experiment? Describe how you can tell.
    b) When performing spot dilutions like in the experiment above, why is it important to plate the cells on both the desired condition (alkaline conditions, pH8.0) and also on rich media (YPD)? How would you interpret the results if the Scrambled strain grew significantly better on both YPD and on alkaline conditions?



11. The yeast genome was selected as the first synthetic genome because of the simplicity of the yeast genome. Think about the differences between the yeast and human genomes. What features or aspects of the human genome would make it much more complicated to implement the Scramble system and the process of Scrambling chromosomes?

Measuring growth rates by serial dilution and plating

Up to now we have Scrambled the synthetic yeast chromosomes and selected for our phenotype of interest. We selected this phenotype based on the size of colonies on the initial selection plates (high temperature, ethanol, or caffeine). Now we need to confirm that these larger colonies really do grow better on the selective conditions. We will do this by:

1. Selecting 11 colonies that you believe show increased growth in the selective conditions
2. Transferring an equal number of cells from each of the 11 colonies to one well of a cell culture dish. You also transfer cells that have not been Scrambled into one well of the cell culture dish.
3. Performing serial dilutions of those cells. This means that in each subsequent well of the cell culture dish, you will have progressively fewer cells
4. Transfer the diluted cells from each well onto a Petri dish containing media with the selective condition
5. Incubate the plates for 2 days and then assess the growth of the diluted cells (both the size and the number of colonies) relative to the non-Scrambled strains.

An example of serial dilutions of yeast cells to compare growth rate on selective conditions (size and number of cells)

Procedure

1. Obtain your yeast cells on the selective plate from last week. Also obtain a plate with non-scrambled cells.

2. Obtain a new 96-well plate. When working with the plate, keep the lid on as much as possible; remove the lid to add or pipette the yeast cells, but put the lid right back on when you're done.

3. Transfer 200 ul of sterile water to each of the wells in the first row of a 96-well plate.

4. Using a sterile toothpick, select a colony from the non-scrambled cells. Swirl the toothpick into the water in the first well of water in the 96-well plate to disperse all the yeast cells into the liquid.

5. Using a sterile toothpick, select one of the largest colonies from the scrambled cells on the selective plate. Swirl the toothpick into the water in the first well of water in the 96-well plate to disperse all the yeast cells into the liquid.

6. Repeat step 5 for 10 more colonies, resuspending each in a new well.

7. Determine the OD of the cells in each well by taking a reading on the plate reader at an absorbance of 600 nm.

8. For each sample, calculate the volume of cells needed to prepare 200 ul of culture at an OD of 0.5 by using the formula C1V1=C2V2. Use the formula below and then complete the table below it.

   (0.5 OD) (200 ul) = (OD measured from plate reader)(Volume of cells needed).

Well	1	2	3	4	5	6	7	8	9	10	11	12
Volume of cells needed (calculated above)												
Volume of water needed to achieve a total volume of 200 ul												

9. Obtain a new 96-well plate. Into the first row, transfer the volume of cells calculated above for each well (ie, you are transferring the calculate volume from well A1 on plate 1 to well A1 on plate 2, etc.)

10. To these cells, add the volume of water calculated above to make a total volume of 200 ul.

11. Into each of the other wells in the dish, add 200 ul of water.

12. Using a P200 pipette set to 40 ul, pipette up and down 3-4X in well A1 and then transfer 40 ul of cells to the row beneath it (well B1). Pipette up and down 3-4X in well B1 and transfer 40 ul to well C1. Repeat all the way down to row H1.

13. Repeat step 12 for each column of cells (ie, transferring from A2 to B2 to C2…..)

14. Obtain 2 rectangular petri dishes with media. One should have rich media (YPD) and the other should have your media with your selective condition. Label the back of these plates with your initials and dates.

15. Obtain two paper templates containing the layouts of each plate. These will show you where on the smooth surface of the agar plate to place your cells. Place these under each of your dishes.

16. You should now have: your 96-well plate with diluted cells, your YPD plate, and your selective plate. Take 2 ul from the first well (A1) of diluted cells and pipette the liquid onto the

corresponding area of the YPD plate (A1). Take another 2 ul from the first well (A1) of diluted cells and pipette the liquid onto the corresponding area of the selective plate (A1).

17. Repeat step 16 for all wells of the 96-well plate.

18. Incubate the cells at 30C (or 40C if that is your selective condition) for 2 days.

Extraction of Yeast DNA

Now that we've Scrambled the yeast DNA and selected our phenotype of interest, we need to see which gene or region of the chromosome is responsible for our phenotype. To do this, we must extract the DNA from the yeast and amplify the PCR Tags. This will allow us to determine, for each section of the chromosome, whether the wild-type or synthetic DNA is present. Note that we will be checking localized regions along the synthetic chromosome; ideally, you would scan the entire chromosome using PCR Tags along the entire length. Today we will extract the DNA, and next week, we will amplify the PCR tags. The following week, we will run our PCR reactions on agarose gels.

This protocol uses the Promega Genomic DNA kit, but any other genomic DNA kit for yeast can be substituted.

1. Transfer 1ml of yeast cell culture into a 1.5ml microcentrifuge tube.

2. Centrifuge at 13,000–16,000 × *g* for 2 minutes to pellet the cells. Remove the supernatant.

3. Resuspend the cells thoroughly in 293µl of 50mM EDTA.

4. Add 7.5µl of Zymolyase enzyme, and gently pipet 4 times to mix.

5. Incubate the sample at 37°C for 60 minutes. During this time, the Zymolyase will digest the yeast cell wall. Cool to room temperature.

6. Centrifuge the sample at 13,000–16,000 × *g* for 2 minutes and then remove the supernatant.

7. Add 300µl of Nuclei Lysis Solution to the cell pellet and gently pipet to mix.

8. Add 100µl of Protein Precipitation Solution and vortex vigorously at high speed for 20 seconds.

9. Let the sample sit on ice for 5 minutes.

10. Centrifuge at 13,000–16,000 × *g* for 3 minutes.

11. Transfer the supernatant containing the DNA to a clean 1.5ml microcentrifuge tube.

> **Note:** Some supernatant may remain in the original tube containing the protein pellet. Leave this residual liquid in the tube to avoid contaminating the DNA solution with the precipitated protein.

12. Add 300µl of room temperature isopropanol and gently mix by inversion. You may see thread-like strands of DNA forming a visible mass.

13. Centrifuge at 13,000–16,000 × *g* for 2 minutes.

14. Carefully remove the supernatant with a P1000 pipet. Add 300µl of room temperature 70% ethanol and gently invert the tube several times to wash the DNA pellet.

15. Centrifuge at 13,000–16,000 × *g* for 2 minutes. Carefully remove the supernatant with a P1000 pipet.

16. Centrifuge at 13,000–16,000 × *g* for 30 sec. Carefully remove the supernatant with a P200 pipet. Allow the pellet to air-dry for 10–15 minutes by leaving the tube sitting open in a rack.

17. Add 50µl of DNA Rehydration Solution.

18. Add 1.5µl of RNase Solution to the purified DNA sample. Vortex the sample for 1 second. Centrifuge briefly in a microcentrifuge for 5 seconds to collect the liquid and incubate at 37°C for 15 minutes.

19. Nanodrop the samples to determine the concentration of DNA. Turn on the machine, use DNA Rehydration Solution to "blank" the reading, then add 1 ul of your DNA solution to the machine and collect the reading.

Concentration:_____ng/ul 260/280 : _____

Amplification of PCR Tags

For each strain of yeast, we will now verify that the strain contains both one copy of the wild-type chromosome and one copy of the synthetic chromosome, as expected. In amplifying these regions, you will also be able to determine whether there are any structural changes in these regions of the chromosomes that are the result of the Scramble process; these will be visible as larger, smaller, or missing DNA bands.

For this procedure, you will amplify three strains of DNA:
1. Wild-type (non-synthetic strain)
2. Synthetic strain prior to Scrambling
3. Your synthetic Scrambled strain.

Each strain of DNA will be amplified with the following primer sets:

1. YEL071W gene, wild-type primers
2. YEL071W gene, synthetic primers
3. YER188W gene, wild-type primers
4. YER188W gene, synthetic primers

Therefore, your reactions will be:

	Template DNA	Primer 1	Primer 2
1	Wild-type	YEL071W_F_WT	YEL071W_R_WT
2	Wild-type	YEL071W_F_SYN	YEL071W_R_SYN
3	Wild-type	YER188W_F_WT	YER188W_R_WT
4	Wild-type	YER188W_F_SYN	YER188W_F_SYN
5	Synthetic, non-Scrambled	YEL071W_F_WT	YEL071W_R_WT
6	Synthetic, non-Scrambled	YEL071W_F_SYN	YEL071W_R_SYN
7	Synthetic, non-Scrambled	YER188W_F_WT	YER188W_R_WT
8	Synthetic, non-Scrambled	YER188W_F_SYN	YER188W_F_SYN
9	Synthetic, Scrambled	YEL071W_F_WT	YEL071W_R_WT
10	Synthetic, Scrambled	YEL071W_F_SYN	YEL071W_R_SYN
11	Synthetic, Scrambled	YER188W_F_WT	YER188W_R_WT
12	Synthetic, Scrambled	YER188W_F_SYN	YER188W_F_SYN

Protocol:

1. Dilute your DNA to a concentration of 20 ng/ul by using the formula C1V1=C2V2.

2. Obtain one set of 12-tube PCR strip tubes. Distribute 12.5 ul of the 2X PCR master mix to each of your 12 PCR tubes.

3. To each tube, add 5 ul of the corresponding primer (1 uM concentration) according to the table above. It is highly recommended that you make a check mark next to each primer as you add it to the tube to keep yourself organized and to remember what was added.

4. To each tube, add 2.5 ul of template DNA according to the table above.

5. Cap your tubes (make sure they are on tightly!) and place them in the PCR thermocycler.

6. Make sure that the machine is set to the following program:

    94 deg, 30 sec
    30 cycles of 94 deg for 30 sec, 64 deg for 30 sec, 68 deg for 1 min
    68 deg, 5 min

Name:

Genome Databases Assignment

Step 1: DNA alignment: What portion of the chromosome are we amplifying?

The first thing that we would like to do is to determine what sequence in the yeast genome our DNA primers (PCR Tags) correspond to. This can be accomplished by performing an alignment. The sequences of our PCR tags are (remember there are two primers for each PCR reaction):

>PCR Tag 1
GTCAGGGTCCAATAAGAGACAC
>PCR Tag 2
CCACCTTGTGGTACTACTGCCA
>PCR Tag 3
ATGTCGGATCCAAGTTCTATTAAC
>PCR Tag 4
TTCTTCCTTCAACTTATATAGGTT

Align each of the PCR tags to the reference genome using BLAST and complete the table below.
Table 1.

	PCR Tag 1	PCR Tag 2	PCR Tag 3	PCR Tag 4
How many hits does your query retrieve?				
What is the score (bits) for your top hit?				
For your top hit, what is the probability the alignment would occur by random chance (the e-value)?				
Scroll down to look at the aligned sequences. What is the percent sequence identity between the search result and your query sequence?				
What chromosome is your sequence on?				
What are the chromosome coordinates of the sequence?				
For one of your PCR tags, go back and shorten your query sequence by three nucleotides. What does this do to the e-value of your BLAST hit(s) and why?				

Step 2: Annotation: What is the location of the sequence in the genome?

From your top alignment, click on "Genome Browser". This will give you a visual representation of the chromosome, including the sequence that your primer corresponds to. The yellow vertical line corresponds to the boundaries of your query sequence. Genes (open reading frames or ORFs) are colored red. There are many other genomic features that are indicated as well. Click on the name of the gene that corresponds to your DNA sequence. This will give you information about this gene.

Table 2.

	PCR tags 1 and 2	PCR tags 3 and 4
What is the name of this gene?		
What is the length of this gene?		
What is the position on the chromosome of this gene (starting and ending coordinates)?		
Is this gene encoded on the positive or negative strand of the DNA double helix?		
What is the systematic name (ID) of this ORF?		
What is the function of this gene?		

Step 3: Annotation: Learn more about your gene:

Click on the name of the gene, which will take you to the gene page where there is even more information, including functional information, links to information on gene and protein interactions, and perhaps a summary paragraph about the gene. Complete the table below for both of your DNA sequences.

Table 3.

	PCR Tags 1 and 2	PCR Tags 3 and 4
What is the closest gene just upstream of your gene?		
What is the closest gene just downstream of your gene?		
Does this gene make a protein? If so, what is its length in amino acids? About how many copies of this protein exist in the cell?		
Is this gene essential in the yeast cell? (You can tell this by looking in the Phenotype section; next to “null”, it should say either that the null mutation (deletion) is viable or inviable).		
Does this gene have a human homolog, and if so, what is it? (You can tell this by looking in the Phenotype section).		
Is the human homolog associated with any disease, and if so, which one (you can tell this by looking in the Disease section).		

Step 4: Gene Ontology

Watch the gene ontology video: https://www.youtube.com/watch?v=KzbX7tKeb5o&t=5s

Table 4.

	PCR Tags 1 and 2	PCR Tags 3 and 4
What is the molecular function(s) of your gene?		
What is the biological process that your gene is involved in?		
Where in the cell does your gene function?		

Analysis of Illumina Next-Gen Sequencing Data

Error Correcton in HighThroughput Datasets

Background

Next-generation sequencing has revolutionized both biological research and clinical medicine, with sequencing of entire human genomes being used to predict drug responsiveness and to diagnose disease. In contrast to traditional Sanger sequencing, next-generation sequencing datasets have shorter read lengths and higher error rates. This can create challenges for downstream analysis since even a small error rate will result in a large number of sequencing reads that contain errors due to the abundance of sequencing reads. Indeed, Illumina MiSeq data produces reads with an error rate of 0.1% (Glenn 2011), yet this corresponds to only ~85% of the 150 bp sequencing reads ($.999^{150}$) being error-free.

The biological sciences increasingly require the analysis of large datasets, and these highthroughput experiments require strict attention to quality control. We will work with raw sequencing reads from a eukaryotic genome sequencing project to identify sources or error, analyze the quality of the data set and then effectively employ error correction strategies.

Importing data into Galaxy

As an input, you will be given the raw sequencing data from a next-generation sequencing experiment using Illumina technology. Complete the following steps:

1. Go to the Galaxy website (https://usegalaxy.org/)
2. At the top of the website, click on "User" and then click on "Register" and sign up for an account.
3. Once on the Galaxy website, there is a panel on the right labeled "History". Click the plus (+) symbol to create a new history. Click on the name just below the search bar to rename your new history.
4. On the left, under the search bar in the “Tools” tab, click the button labeled “Upload Data” and a pop-up should appear. You can find files on your computer using the “Choose Local Files” button, or you can open your files tab separately and drag the files to the space labeled “Drop files here.”

Understanding the data format.

Your data is encoded in FASTQ format. Once your data file has uploaded, click on the eye icon, and your data should look like the data below:

```
    @HWUSI-EAS100R:6:73:941:1973#0/1
TCTTTGGGTTCTGGGGGGAGTATGGTCGCAAGGCTGAAACTTAAANGAATTG
@HWUSI-EAS100R:6:73:941:1973#0/1
hhhhhhhhfhhhhhhhhhbhbhhffZffacacchghgffehhf[[B]\___`
```

Here is how to interpret FASTQ format:

Line 1: This is the FASTQ header.

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end reads only)*

Line 2: This is the DNA sequence. N nucleotides represent ambiguous bases whose identity could not be determined with confidence.

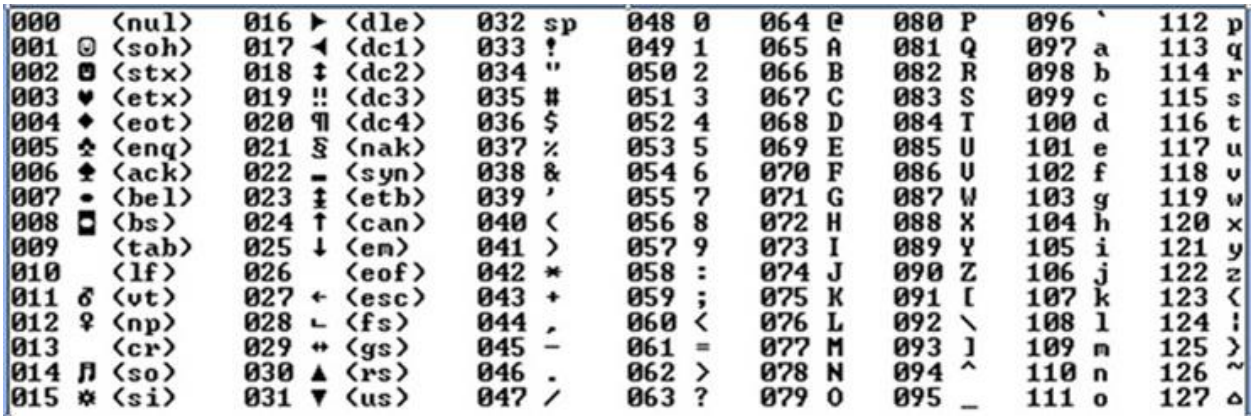
Line 3: This is a separator line (your header again).

Line 4: These characters represent the quality scores. There is one ASCII-encoded character for each nucleotide in your sequence in line 2. Each character can be converted into a probability that the nucleotide was identified in error (step 3 below).

1. **Converting ASCII characters to PHRED scores and then to error probabilities**. ASCII characters can be converted to a numerical score using the lookup table below.

   What is the numerical score for the ASCII character 'h'? _____

   These numerical scores can be converted into PHRED scores, which are the quality scores for DNA sequencing data with the PHRED score representing the quality of the sequencing data at that



000		(nul)	016	▶	(dle)	032	sp	048	0	064	@	080	P	096	`	112	p
001	☺	(soh)	017	◀	(dc1)	033	!	049	1	065	A	081	Q	097	a	113	q
002	☻	(stx)	018	↕	(dc2)	034	"	050	2	066	B	082	R	098	b	114	r
003	♥	(etx)	019	‼	(dc3)	035	#	051	3	067	C	083	S	099	c	115	s
004	♦	(eot)	020	¶	(dc4)	036	$	052	4	068	D	084	T	100	d	116	t
005	♣	(enq)	021	§	(nak)	037	%	053	5	069	E	085	U	101	e	117	u
006	♠	(ack)	022	▬	(syn)	038	&	054	6	070	F	086	V	102	f	118	v
007	•	(bel)	023	↨	(etb)	039	'	055	7	071	G	087	W	103	g	119	w
008	◘	(bs)	024	↑	(can)	040	(	056	8	072	H	088	X	104	h	120	x
009		(tab)	025	↓	(em)	041	)	057	9	073	I	089	Y	105	i	121	y
010		(lf)	026		(eof)	042	*	058	:	074	J	090	Z	106	j	122	z
011	♂	(vt)	027	←	(esc)	043	+	059	;	075	K	091	[	107	k	123	{
012	♀	(np)	028	∟	(fs)	044	,	060	<	076	L	092	\	108	l	124	¦
013		(cr)	029	↔	(gs)	045	-	061	=	077	M	093	]	109	m	125	}
014	♫	(so)	030	▲	(rs)	046	.	062	>	078	N	094	^	110	n	126	~
015	☼	(si)	031	▼	(us)	047	/	063	?	079	O	095	_	111	o	127	⌂

   position. Illumina data uses a PHRED33 encoding system, which means that numerical score above can be converted to a PHRED score by subtracting 33 from the numerical score above.

   What is the PHRED score for the ASCII character 'I'? _____

   The PHRED score is related to the accuracy of the sequencing data according to the following formula:

   PHRED score = $-10 \log_{10}(P_e)$ where P_e is the probability that the base was identified in error

   What is the probability of error for a nucleotide associated with the ASCII character 'I'? _____

Analysis of overall data quality

While we can manually convert ASCII characters into PHRED scores and error rates, Galaxy will perform this analysis for us, converting the ASCII score for each nucleotide to a PHRED score. To help us make sense of the large volume of data, we will use statistical analysis to summarize the quality of our data.

2. **Summary statistics to analyze PHRED scores**. Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on FASTQ Quality Control, then click on "FastQ Summary Statistics by column". Make sure that the proper FastQ data file is selected in the input window at the top and then click "Execute". Your output file should have many columns that look like the one below:

#column	count	min	max	sum	mean	Q1	med	Q3
1	4808294	2	40	167101114	34.7526823443	36.0	39.0	40.0

   column = column number (1 to 50 for a 50 nucleotide Illumina sequencing read)

count = number of bases found in this column (equal to the number of sequencing runs)
min = Lowest quality score **(PHRED)** value found in this column.
max = Highest quality score **(PHRED)** value found in this column.
sum = Sum of quality score**(PHRED)** values for this column.
mean = Mean quality score **(PHRED)** value for this column.
Q1 = 1st quartile quality score **(PHRED)**.
med = Median quality score **(PHRED)**.
Q3 = 3rd quartile quality score **(PHRED)**.

Copy and paste this data into an Excel spreadsheet.

What is the median PHRED score for your data in column 1?_____

What is the mean PHRED score for your data in column 1?_____

Using this information, what conclusions can you make about the quality of your data?

Analyzing sources of error.

We will now use the quality control tool to analyze sequencing reads individually and to analyze all of the sequencing reads one position at a time. This will help us to identify any systematic sources of error that we can then remove.

3. **FastQC quality control**. Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on FASTQ Quality Control, then click on "FastQC". Select as an input in the top bar your raw sequencing data in fastqsanger format. Click "Execute". This will output two new files, one labeled "RawData" and one labeled "Webpage".

4. Go to the History tab on the right side of your screen. Find the output files from step 3, then find the file labeled "Webpage". Click on the eye icon next to this file. Under "Summary" in the middle of the screen, click on "**Per Base Sequence Quality**". This will display a boxplot of the PHRED scores of your sequencing read one nucleotide at a time (It can be helpful to minimize the Tools tab by clicking on the arrow in the bottom left corner of the screen). For example, the first box and whiskers plot represents the summary data for the first nucleotide of all of the sequencing reads. The second box and whiskers plot represents the summary data for the second nucleotide of all of the sequencing reads. **Right click on the data to save as a JPEG.**

5. Next, click on "**Per Sequence Quality Scores**". This shows you the average sequence quality for each individual sequencing read. What would you estimate as the mean quality score of your sequencing reads? Are there any sequencing reads that contain a substantially higher or lower quality score? **Right click on the data to save as a JPEG.**

Correcting errors in the dataset.

Based upon your analysis of error sources above, we will perform error correction steps to improve the quality of the data and then we will perform FastQC quality control once again to see if there has been any improvement in the quality of the data.

In a large dataset, some sequencing reads will be of overall poor quality. If you have a problem with low quality sequencing reads (visible when you look at Per Sequence Quality Scores), perform filtering using step 6 below.

6. Go to the Tools tab on the left side of your screen. In the search bar, type in "**Filter by quality**" and scroll down the search results until you find the one that just says **Filter by quality**. For "Input FASTQ file", select your raw sequencing data from step 1. For your "quality cut-off value", use a PHRED score of 13. For the "Percent of bases in sequence that must have quality equal to / higher than cut-off value", set the value to 100. Click "Execute".

As sequencing reads get extended, the sequencing quality becomes poorer. If you have a problem with low quality data toward the end of your sequencing read (visible when you look at Per Base Sequence Quality), perform trimming using step 7 below.

7. Go to the Tools tab on the left side of your screen. In the search bar, type in "**Trim sequences**" and scroll down the search results until you find the one that just says **Trim sequences**. For "Input File", select your raw sequencing data from step 1. For "First base to keep", use 1. For "Last base to keep", use 44. Click "Execute". [Note: the sequences that you analyze for your paper will be much longer than these sequences and you will want to keep more than 44 bases!]

Analysis of error correction

Now that we have performed error correction on our sequencing data, we would like to look at the summary statistics for our sequencing data again to see what kind of an improvement we have made.

1. **Summary statistics to analyze PHRED scores**. Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on FASTQ Quality Control, then click on "FastQ Summary Statistics". Make sure that you select the data file following error correction in the input window and then click "Execute". Your output file should have many columns that look like the one below:

#column	count	min	max	sum	mean	Q1	med	Q3
1	4808294	2	40	167101114	34.7526823443	36.0	39.0	40.0

What is the median PHRED score for your data in column 1?_____

What is the mean PHRED score for your data in column 1?_____

What conclusions can you make about the quality of your data following error correction? **Copy and paste this data into your Excel spreadsheet.**

2. **FastQC quality control**. **FastQC quality control**. Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on FASTQ Quality Control, then click on "FastQC". Select as an input your filtered data. Click "Execute".

3. Click on "**Per Base Sequence Quality**" and look at the graph. How has your sequencing data changed as a result of error correction? **Right click on the data to save as a JPEG.**

4. Click on "**Per Sequence Quality Scores**" and look at the graph. How has your sequencing data changed as a result of error correction? **Right click on the data to save as a JPEG.**

Include in the data section of your paper:

1. An Excel table containing the output from the Summary Statistics analysis for your data. Include analysis of data prior to error correction and following error correction. For each corrected dataset, make sure that the error correction parameters are noted in the legend for the table.

2. Histograms showing the Per Sequence Quality Scores prior to prior to error correction and following error correction. For each corrected dataset, make sure that the cutoff values that you used for error correction are noted. Make sure to include figure legends.

3. Boxplots showing the Per Base Quality Scores prior to error correction and following error correction. For each corrected dataset, make sure that the error correction parameters (which nucleotides were trimmed) are noted. Make sure to include figure legends.

Things to discuss in your paper:

1. Why is it important that sequencing data includes a numerical measure of error in addition to the nucleotide sequence? How did you evaluate the overall quality of your original dataset? What piece(s) of data did you use and how did you interpret it/them?

2. How did you identify source(s) of error in your sequencing data? What piece(s) of data did you use and how did you interpret it/them? Where (based upon the chemistry of the DNA sequencing reactions) did these error(s) originate?

3. How did you determine what error correction parameters to use? Why did you select these values as opposed to stricter or more lenient error correction parameters?

4. How did your sequence improve based on the error correction steps that you used? What pieces of data justify your analysis?

References

Bentley, DR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 456(7218):53-9

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy Team. 2010. Manipulation of FASTQ data with Galaxy. Bioinformatics. 26(14):1783-5

Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, and Peter M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants Nucl. Acids Res. (2010) 38 (6): 1767-177.1

Glenn TC. 2011. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 11(5):759-69.

Skrabanek L. Using Galaxy for NGS Analyses. 2012. Accessed from: http://chagall.med.cornell.edu/galaxy/GalaxyWorkshopNotes.pdf

www.en.wikipedia.org/wiki/FASTQ_format

Identification of SNPs in Illumina Sequencing Data

Goal

To identify any SNPs in our sequenced genomes that are not present in the reference yeast genome.

Note: Each of these steps follows the preceding step in order, so make sure that the input for each of these steps is the output file from the immediately preceding step.

Aligning data to the reference yeast genome

First, we need to take each of our sequencing reads and determine what position in the reference genome they correspond to.

8. **Mapping data to the reference genome**. To align your data to the reference genome, go to the Tools tab on the left side of your screen. In the search bar type "**Bowtie2**" and in the search results click on the first one labeled **Bowtie2**.

   Under "Is this single or paired library" at the top, select "Paired-end"
   Select the data from your two paired end reads as the inputs
   Scroll all the way down to "Do you want to tweak SAM/BAM Options?" and select Yes. Just below that, under "Would you like the output to be a SAM file" set the switch to Yes.
   Click "Execute".

The output of the Bowtie2 mapping program is a SAM file. The Sequence Alignment/Map (SAM) format is a generic nucleotide alignment format that describes the alignment of sequencing reads to a reference sequence.

Filtering out reads that do not map to the reference genome.

The SAM file now needs to be filtered so that we select only reads that are proper paired-end reads and that also map to a unique position to the reference genome (ie. we want to filter out sequencing reads that map simultaneously to multiple positions in the reference genome).

1. Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on SAM/BAM, and click on **Filter SAM**, which should be all the way at the bottom of the list.
   Click "Add New Flag"
   Select Flag 1, set "Type" as "Read is paired", and set the state to "Yes"
   Click "Add New Flag"
   Select Flag 2, set "Type" as "Read is mapped in proper pair", then "Yes"
   Click "Add New Flag"
   Select Flag 3, set "Type" as "Read is unmapped", set the state to "No"
   Click "Execute".

2. **Converting file formats**. We must now convert the SAM file to a binary BAM file. To change the file format, go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on SAM/BAM, then scroll down the list and click on **SAM to BAM.** Make sure the SAM file you just filtered is the one selected.

Click "Execute".

3. **Determining how many of our reads map to the reference genome**. One of the important metrics that describes our sequencing data is the percentage of sequencing reads that map uniquely to the reference genome. Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on SAM/BAM, and click on **Samtools flagstat**. Make sure your converted BAM file from step 2 is selected. Click "Execute".

Once the file turns green, view the data using the eye icon.

Record the number of sequences that map to the reference. View your original paired-end sequencing data using the eye icon and record the original number of sequences. **Calculate and record the percentage of sequences that map to the reference.**

Identifying SNPs

To identify SNPs, we need to line up all the sequencing reads at each position (a pileup file), determine the consensus nucleotide at each position, and then identify any nucleotides that do not match the reference genome.

4. **Generating a pileup file**. Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on SAM/BAM, then scroll down the list and click on **Generate pileup**.
   Keep all the defaults, but for "Call consensus according to MAQ", click "Yes".
   Click "Execute" and then click the eye icon to view the data.
   Each line in the file includes information on:
   Column 1: The chromosome
   Column 2: The nucleotide position on the chromosome.

Column 3: The reference base at that position
Column 4: The consensus base from the sequencing data at that position
Column 5: The average quality score for each read covering that position
Column 8: The coverage (number of sequencing reads) at that position

5. **Convert file format**. Click the pencil icon next to the file from step 4. At the top, under the "Datatype" tab, change the file format to "pileup" if it is not already selected. Click "Save".

6. **Examining coverage**. To see how well the various positions in the genome are covered, Go to the Tools tab on the left side of your screen. Under Statistics and Visualization, click on Statistics, then scroll down the list and click on **Summary Statistics**. Make sure your converted file from step 5 is selected.
   For "column or expression", enter c8. Click "Execute" and then click the eye icon once the file has turned green.
   Record the mean coverage for the genome.

7. **Compare the consensus to the reference genome to identify variants (SNPs).** We now need to filter the pileup file so that we have a file that contains only nucleotides that vary from the reference genome.
   Go to the Tools tab on the left side of your screen. Under Genomic File Manipulation, click on SAM/BAM, then click on **Filter pileup**.
   Under "which contains" select "Pileup with ten columns (with consensus)"
   For "Do not consider read bases with quality lower than", enter 20
   For "Do not report positions with coverage lower than", enter 6
   For "Only report variants" select Yes
   For "Convert coordinates to intervals" select Yes
   For "Print total number of differences" select Yes
   For "Print quality and base strings" select No
   Click "Execute"

This will produce a file with only those nucleotides that vary from the reference

Separating true SNPs from sequencing errors

If you click on the eye icon to view the filtered pileup file, you will notice that for most of the positions in the genome, the consensus base is identical to the reference base (ie, it does not appear to be a SNP!) This is because the file contains all of positions in the genome where any sequencing read differs from the reference, even if the consensus base at that position is identical to the reference. We can distinguish between sequencing errors (where the variant base is present in few sequencing reads) and true SNPs by filtering out variant bases that are not present in >80% of the sequencing reads at that position.

8. **For each variant, calculate the percent of the sequencing reads that support that variant.**

   In your filtered pileup file, column 15 represents the total number of reads at each position and column 14 represents the number of sequencing reads that vary from the reference at that position.

   Go to the Tools tab on the left side of your screen. Under General Text Tools, click on Text

Manipulation, then click on **Compute**.
For "Add expression", enter (c15/c14)*100
For "Round result" enter No.
Click "Execute"
This will add a column (column 16) which tells you the percent of sequencing reads supporting each variant.

9. **Only keep variants that are present in 80% or more of the sequencing reads.**

Go to the Tools tab on the left side of your screen. Under General Text Tools, click on Filter and Sort, then click on **Filter**.
For "With following condition", enter c16>80
For "Number of header lines to skip", enter 0
Click "Execute".
This produces a file that only contains SNPs that are present in >80% of the sequencing reads.

Identify variants that are unique to the evolved (scrambled) strain

The original strains that we used for our scramble experiments (the unscrambled strains) are not the same strain of yeast that serves as the reference genome. Therefore, some of the variants that we have identified will be present in the starting unscrambled strain and some will be unique to the scrambled strain. We want to remove the SNPs that are present in the unscrambled strain so that we are left only with the SNPs that are unique to the scrambled strain.

10. **Identify variants that are unique to the scrambled strain and not present in the unscrambled strain.**

Go to the Tools tab on the left side of your screen. Under General Text Tools, click on Join, Subtract, and Group, then click on **Compare two datasets**.
For the first dataset, select your filtered data from step 9 on page 4.
For the second data set, you must upload a file for SNPs in an unscrambled genome.
This data should come from your original yeast strain containing synthetic chromosomes that have not been subjected to scramble.
Compare your two RO strain datasets using column 2 (the chromosome coordinates) to find Nonmatching rows of first dataset.

Click the disk icon to download your data in a format that can be opened in Excel.

The result of all of your analysis should be a file containing SNPs that:

- are supported by properly mate-paired sequencing reads (step 1, page 2)
- are supported by sequencing reads that map to unique positions in the genome (step 1, page 2)
- are supported by sequencing reads of quality >20 (step 7, page 3)
- are at positions that have been sequenced >6 times (step 7, page 3)
- are present in >80% of the sequencing reads at that position (step 9, page 4)
- are unique to the evolved (scrambled) strain (step 10, page 5)

Things to include in your report:

A table including:

- The total number of initial sequencing reads (from Step 3, page 2)
- The percent of unique reads that map to the genome (from Step 3, page 2)
- The mean genome coverage (from Step 6, page 3)

A second table that includes data downloaded from step 10, page 5:

- The chromosome number (Column A)
- The numerical position of each SNP (Columns B and C)
- The base (A, T, C, or G) in the reference genome (Column D)
- The base in the sequenced genome (Column E)
- The percent of sequencing reads that support each variant (column P)
- You do not need to include columns F through O

References

http://blog.avadis-ngs.com/2012/04/elegant-exact-string-match-using-bwt-2/

Skrabanek L. Using Galaxy for NGS Analyses. 2012. Accessed from: http://chagall.med.cornell.edu/galaxy/GalaxyWorkshopNotes.pdf

Visualizing Sequence Variants

Using Galaxy, you identified SNPs in evolved yeast cells. Now we want to:

- Verify that those SNPs are real and not just sequencing errors
- Determine where in the genome (in what gene, for example) the SNPs are located and whether they alter the protein coding sequence

To accomplish this, we will examine the raw data (the individual sequencing reads) aligned to the reference yeast genome using the Integrated Genome Viewer (IGV).

Open the Genome Viewer

1. Go to http://www.broadinstitute.org/igv/ and download the genome viewer.
   Double click the IGV2.1 icon to open the Integrated Genome Viewer browser.

2. In the top left corner from the pull-down menu, select *S. cerevisiae* (SacCer3) as your genome.

Loading the Data

There are two ways to load data into IGV:

1) **Download your BAM files and open them in IGV**. To open a BAM file, IGV needs both the BAM file itself with the .bam label, and an Index file of the exact name with the label .bai

- In Galaxy, click on your BAM file. In the grey bar below the file name there are several icons. Click on the first one that looks like a floppy disc labeled Download. It should give you the option to download either the dataset (the BAM file) or the BAM Index file. Download both files and move them to their own folder on your computer.
- In IGV, go to File, then Load from File. Find the files you downloaded from Galaxy, select your BAM file and click "Open". It may take a few seconds for the data to load.

2) **Upload the files directly to IGV from Galaxy**. We have found that some people experience problems when trying to open the files within IGV itself; this method can circumvent those issues by taking them directly from Galaxy

-In Galaxy, click on your BAM file. Below the grey toolbar, there are several display options available. Make sure you have IGV downloaded to your computer, then find the option labeled "display with IGV". Next to it there are two options: local and S. cerevisiae (sacCer3). Click on the S. cerevisiae link, and it should download a file to your computer. Open this file (Your computer may ask you to download / update Java for this to work). This should open up IGV automatically and display the reads you input.

Once you have opened the file in IGV, you can now view your reads against the S. cerevisiae (sacCer3) genome. To look at an individual chromosome, like chromosome I, go to the top of the viewer, where you will have several drop-down menus. In the middle drop-down menu, select chrI; this will take you to chromosome I of yeast.

Visualizing the Data

Just below the drop-down menus, you will see a ruler. Just above the ruler is a double-sided arrow that will indicate the length of chromosome that is visible in the viewer (it should start at 229kb, the size of yeast chromosome 1). To zoom in on the sequencing data, click on the ruler and drag the mouse across the region that you want to zoom in on. You should suddenly see four new features:

- A **data track** (a gray track that looks like a mountain range). Hover your mouse over this region and you will note that at each position, you will get data for the Total Count (the number of sequencing reads that include that particular base), as well as the number and percentage of sequencing reads that have either and A, C, T, G or N (undetermined nucleotide) at that position.

- An **alignment track** consisting of many rows of thick gray arrows that represent individual sequencing reads. Note that some arrows point to the left and others point to the right. Hover over one of the tracks and the viewer will give you information about that sequencing read, including the Phred quality of the sequencing read.

- A multicolored **feature track** (red, blue, green, and brown) that represents the nucleotide sequence of the reference yeast strain. Keep zooming in if necessary to see the letters of the sequence. Below the nucleotide sequence will the amino acid sequence in all three possible reading frames (only one of which is the correct reading frame for this particular protein).

- A blue line that represents the **amino acid sequence** of the reference yeast strain. Note that not all nucleotides code for amino acids-the proteins are represented by the thicker blue lines, with white arrows indicating the direction of the coding sequence. Keep zooming in if necessary to see the letters of the amino acid sequence.

To zoom back out, hover over the gray alignment tracks that represent the individual sequencing reads, then hold down the Alt key and press the minus sign.

There is a great 5-minute video on Youtube from IGV's channel explaining these features further: https://www.youtube.com/watch?v=E_G8z_2gTYM

Verify all mutations

Using Galaxy, you produced a list of SNPs in evolved strains. You now want to use IGV to verify those mutations and determine whether they have altered the protein.

3. You can go directly to the mutation of interest by typing its genomic location (chromosome and nucleotides) into the white box at the top center.

4. Confirm each mutation as being authentic only if it is present in >80% of the sequencing reads (remember that you can determine this by hovering over the data track).

5. At the bottom you will see the sequence of the amino acids so that you can determine what change (if any) occurred at the protein level. Use the codon table on the next page to determine whether there is a change to the amino acid.

Note: when determining changes to the amino acid, make sure that the gene sequence corresponds to one of the three amino acid sequences immediately above it. If the two do not agree, your gene may be on the Crick strand of the DNA. Click the arrow to the right of the word "Sequence" and this will change the amino acid translation to the Crick strand. One of the three translations should now match the protein sequence at the bottom. **Note that your nucleotide sequence should now be read from right to left!!**

TABLE

64 RNA Codons and the Corresponding 20 Amino Acids

	U		C		A		G	
U	UUU	Phe (F)	UCU	Ser (S)	UAU	Tyr (Y)	UGU	Cys (C)
	UUC		UCC		UAC		UGC	
	UUA	Leu (L)	UCA		UAA	Ocher	UGA	Opal
	UUG		UCG		UAG	Amber	UGG	Trp(W)
C	CUU	Leu (L)	CCU	Pro (P)	CAU	His (H)	CGU	Arg (R)
	CUC		CCC		CAC		CGC	
	CUA		CCA		CAA	Gln (Q)	CGA	
	CUG		CCG		CAG		CGG	
A	AUU	Ile (I)	ACU	Thr (T)	AAU	Asn (N)	AGU	Ser (S)
	AUC		ACC		AAC		AGC	
	AUA		ACG		AAA	Lys (K)	AGA	Arg (R)
	AUG	Met(M)	ACG		AAG		AGG	
G	GUU	Val (V)	GCU	Ala (A)	GAU	Asp (D)	GGU	Gly (G)
	GUC		GCC		GAU		GGC	
	GUA		GCA		GAA	Glu (E)	GGA	
	GUG		GCG		GAG		GGG	

Supply List

Empty plasmid (pRS413)		Obtain from network	
PCR tag primers		Sequences in Scramble lab manual	
WT DNA and synthetic DNA		Extract from cells using Promega Genomic DNA kit	
Spectrophotometer cuvettes		VWR	97000-586
150 mm plates		VWR	25384-326
Glass plating beads		VWR	76019-982
96-well plates (2 per student)		VWR	82050-652
OmniTray (2 per student)		VWR	62409-600
PCR strip tubes		VWR	53509-306
Beta estradiol		Sigma Aldrich	E8875
Promega Wizard genomic DNA kit		Promega	A1120
2X PCR master mix		NEB	M0482
Zymolyase 20T		Amsbio	120491-1
Bacto Agar		VWR	90000-760
SC-Ura liquid media			
	YNB w/o ammonium sulfate, w/o amino acids	VWR	90002-646
	Ammonium sulfate		
	Dextrose		
	SC-Ura dropout mix	Sunrise science	
YPD media		VWR	76470-208
96-well paper templates			
EDTA (50mM solution)			
Isopropanol and ethanol			

Equipment Needed

1. 30C shaker
2. 30C incubator
3. Spectrophotometer to measure cell density
4. Plate reader for absorbance (alternatively can use the cell counter setting on a NanoDrop spectrophotometer or can use a regular spectrophotometer with a larger volume)
5. NanoDrop spectrophotometer to measure DNA concentration
6. Thermocycler

Recipes

SC-Ura media:

YNB w/o ammonium sulfate, w/o amino acids	1.7 g
Ammonium sulfate	5.0 g
SC-Ura dropout mix	2.0 g

Combine ingredients above in a total volume of 500 ml diH2O.

Dextrose	20 g

Bring dextrose to 500 ml volume with diH2O.

Autoclave YNB/AS/AA and dextrose separately, then combine to make 1L media.

YPD media:

Weigh 50g powder and bring up to 1000 ml with diH2O. Autoclave.

YPD plates:

Weigh 50g powder and bring up to 500 ml with diH2O.
Weigh 20g agar and bring up to 500 ml with diH2O.
Autoclave the two bottles separately, then combine and pour plates.