How to Design Primers

Danielle A. Presgraves

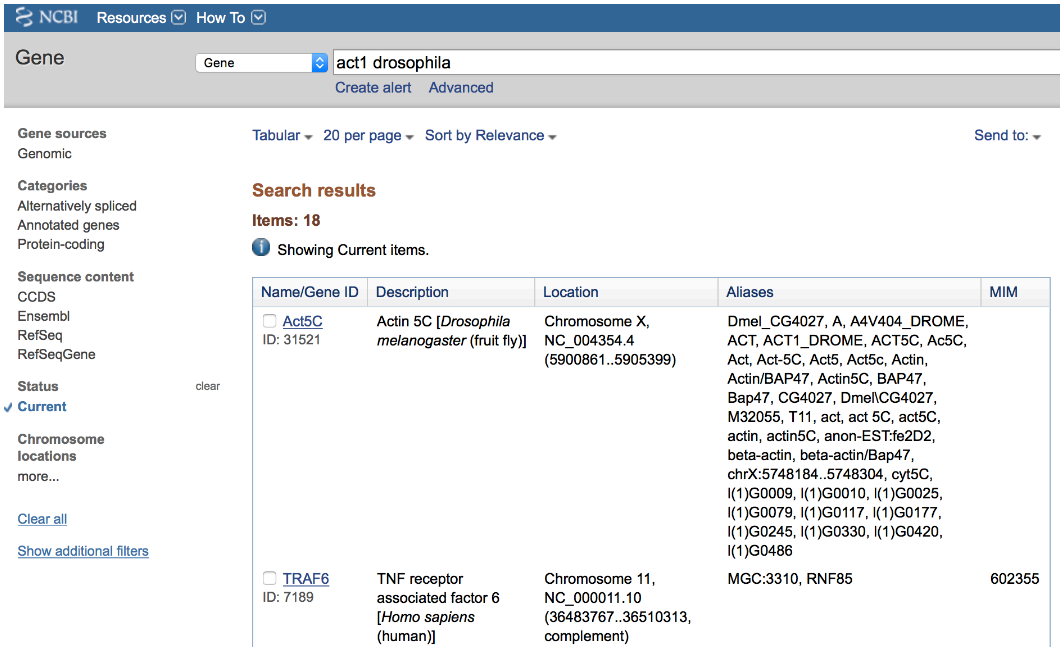
1 Introduction to Primer Design

During the two lab modules this week, we investigate how to amplify a particular genomic region and how to build a tree that will summarize evolutionary relationships. In order to accomplish these two goals, we will need to learn to do three major things:

1. Retrieve DNA sequences from an online database (in FASTA format). We learned to do this - and to how to BLAST sequence- in the first module/lab via ncbi or yeastgenome.com.
2. Align different sequences (ClustalΩ suite) - We also covered this in the first module.
3. Design diagnostic primers for the gene that you wish to amplify - We will learn to do this today

In today's lab, we are going to concentrate on how to design diagnostic primers. In order to do that, we will need to BLAST the gene you have been given. In the first module, we investigated the ncbi website and learned "how to BLAST". To remind you, the basic steps of a BLAST search involve the following pathway:
1. Go to the NCBI website: `http://www.ncbi.nlm.nih.gov`
2. Type your gene name into the Entrez search for the default "all databases" search
3. If your gene has a different name in the organism that are you interested in than in *S.cerevisiae*, limit your search to the GENE database - you can select it by clicking on the 'all databases' tab next to the search bar
4. Download the FASTA file of your gene in whatever organism you are using.
5. Run the BLAST program if you are interested in finding other sequences that are similar to your sequence

Now, back to the main question: how do we design primers? There are a number of considerations when designing a primer to amplify your given genes. The initial step involved in engineering primers, for both polymerase chain reaction (PCR) and sequencing, leverages the anti-parallel complementary nature of the double stranded target DNA.



Gene: act1 drosophila

Search results

Items: 18

Showing Current items.

Name/Gene ID	Description	Location	Aliases	MIM
□ Act5C ID: 31521	Actin 5C [*Drosophila melanogaster* (fruit fly)]	Chromosome X, NC_004354.4 (5900861..5905399)	Dmel_CG4027, A, A4V404_DROME, ACT, ACT1_DROME, ACT5C, Ac5C, Act, Act-5C, Act5, Act5c, Actin, Actin/BAP47, Actin5C, BAP47, Bap47, CG4027, Dmel\CG4027, M32055, T11, act, act 5C, act5C, actin, actin5C, anon-EST:fe2D2, beta-actin, beta-actin/Bap47, chrX:5748184..5748304, cyt5C, l(1)G0009, l(1)G0010, l(1)G0025, l(1)G0079, l(1)G0117, l(1)G0177, l(1)G0245, l(1)G0330, l(1)G0420, l(1)G0486	
□ TRAF6 ID: 7189	TNF receptor associated factor 6 [*Homo sapiens* (human)]	Chromosome 11, NC_000011.10 (36483767..36510313, complement)	MGC:3310, RNF85	602355

Figure 1: Gene Database

1.1 Genes that you will work on, 2019, are the ones that were disclosed in the first module to be attached to HMO1:

1. YGR1118W
2. YPR104C
3. YBR181C
4. YDR382W
5. YDL082W
6. YLR264W

Each group will design two primers within 500 bp of the 5‘ flank and 1 reverse primer within the 3‘ flank of the gene that they have been assigned. We will investigate the reasons behind these design specifications but, for now, we will state that we want the designed primer to be approximately 20 nt long and to have a T_m of approximately 52.5 $\circ$ Celsius. That particular T_m restriction has been chosen to allow everyone to use the same reactions (if everyone designed their primers to have different T_m, each reaction would have to be run individually which would be inefficient). As we work through this lab, we will also use the following website: `https://www.idtdna.com/calc/analyzer`.

1.2 A brief refresher on PCR

You may recall that the purpose of Polymerase Chain Reaction (PCR) is to make many copies (sometimes billions) of a specific segment of DNA. In fact, this technique is central to modern genetics; it has been reasonably credited with allowing the genomic revolution of the past 25 years to happen since prior to PCR amplifying genes was a time consuming process that involved expensively cloning bacterial vectors.

How does PCR work? It mimics many of the basic processes of DNA replication:
There are two primers, forward and reverse, that each attach to a single strand of the separated (by heat) target double stranded DNA at opposite ends of the gene of interest. Each primer, forward or reverse, possesses an available 3'-OH group to on which a DNA polymerase enzyme can add the 5'-Phosphate end of a free nucleotide to elongate the synthesized strand at the end of the primer. In general, PCR needs a mixture of dNTPs (nucleotide triphophates), *Taq* polymerase (which is stable and functional at the high temperature needed to separate double stranded target DNA) and some salts ($MgCl_2$). We see the process of PCR illustrated by Figure 1.

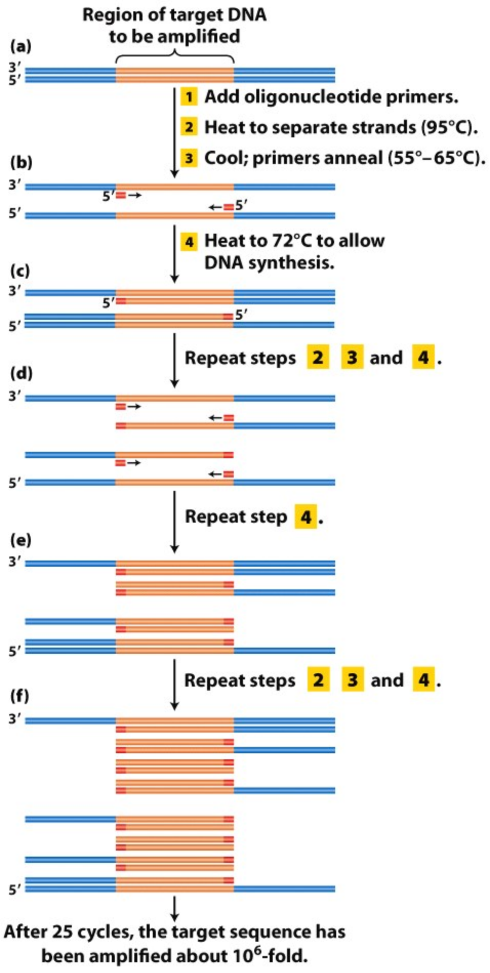

Region of target DNA to be amplified
(a)
3′
5′
1 Add oligonucleotide primers.
2 Heat to separate strands (95°C).
3 Cool; primers anneal (55°–65°C).
(b)
3′
5′
5′
5′
4 Heat to 72°C to allow DNA synthesis.
(c)
3′
5′
5′
5′
Repeat steps 2 3 and 4.
(d)
3′
5′
Repeat step 4.
(e)
3′
5′
Repeat steps 2 3 and 4.
(f)
3′
5′
After 25 cycles, the target sequence has been amplified about 10^6-fold.

Figure 10-3
Introduction to Genetic Analysis, Tenth Edition
© 2012 W. H. Freeman and Company

Figure 2: Visualizing PCR

1.3 Some important notational conventions

It is useful to remember some particular conventions of how we write out DNA sequences, since it can make the process of primer design much easier to visualize:

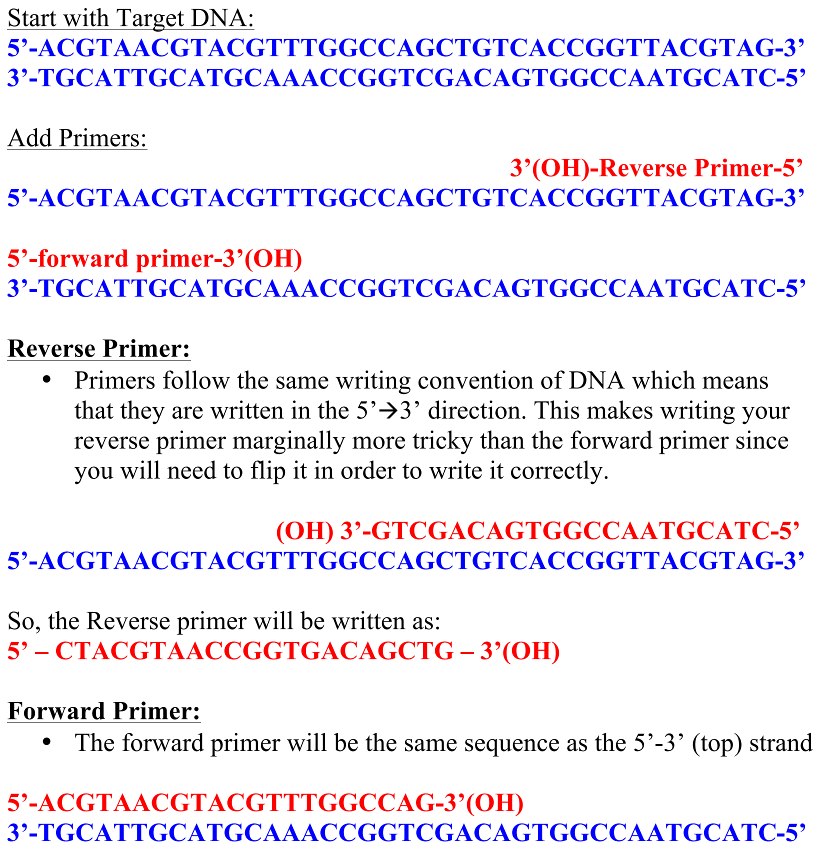
1. We always begin a sequence at the **5'** end: 5'-ATGCTGATCTTGGCCATCAATG-3' is complementary to the strand 5'-CATTGATGGCCAAGATCAGCAT -3'

2. We write the nucleotides in the following way: **A**=adenine, **C**=cytosine,**G**=guanine, **T**=thymine, **N**= any base, **R**=A or G, **Y**=C or T, **"-"** = gap

3. The A-T bond involves only two hydrogen bonds whereas the C-G bond involves three. This means that the A-T bond requires a lower temperature to break and is not as "strong" as the C-G bond. This fact has functional consequences regarding the optimal percentage of C-G compared to A-T when it comes to designing a primer that will securely bond to the intended locus.

4. Another thermodynamic constraint is summarized by the T_m which is the melting temperature measurement.

Speaking of visualization, there are some standard conventions about forward and reverse primers that are easiest to understand when diagrammed out (see figure 2).You can see that the notation of how we represent DNA strands means that you have to be a bit more careful with your reverse primer than with your forward primer.



Start with Target DNA:

5'-ACGTAACGTACGTTTGGCCAGCTGTCACCGGTTACGTAG-3'
3'-TGCATTGCATGCAAACCGGTCGACAGTGGCCAATGCATC-5'

Add Primers:

3'(OH)-Reverse Primer-5'
5'-ACGTAACGTACGTTTGGCCAGCTGTCACCGGTTACGTAG-3'

5'-forward primer-3'(OH)
3'-TGCATTGCATGCAAACCGGTCGACAGTGGCCAATGCATC-5'

Reverse Primer:

- Primers follow the same writing convention of DNA which means that they are written in the 5'→3' direction. This makes writing your reverse primer marginally more tricky than the forward primer since you will need to flip it in order to write it correctly.

(OH) 3'-GTCGACAGTGGCCAATGCATC-5'
5'-ACGTAACGTACGTTTGGCCAGCTGTCACCGGTTACGTAG-3'

So, the Reverse primer will be written as:
5' – CTACGTAACCGGTGACAGCTG – 3'(OH)

Forward Primer:

- The forward primer will be the same sequence as the 5'-3' (top) strand

5'-ACGTAACGTACGTTTGGCCAG-3'(OH)
3'-TGCATTGCATGCAAACCGGTCGACAGTGGCCAATGCATC-5'

Figure 3: Visualizing forward and reverse primers

1.4 Primer Design Considerations

Now we have a general cartoon of the process of creating/using primers. However, there are still a few BIG questions that we need to address about how to effectively design primers: **First, how do you determine the necessary sequence of your complementary primer when you don't even know the sequence of the locus/gene?**

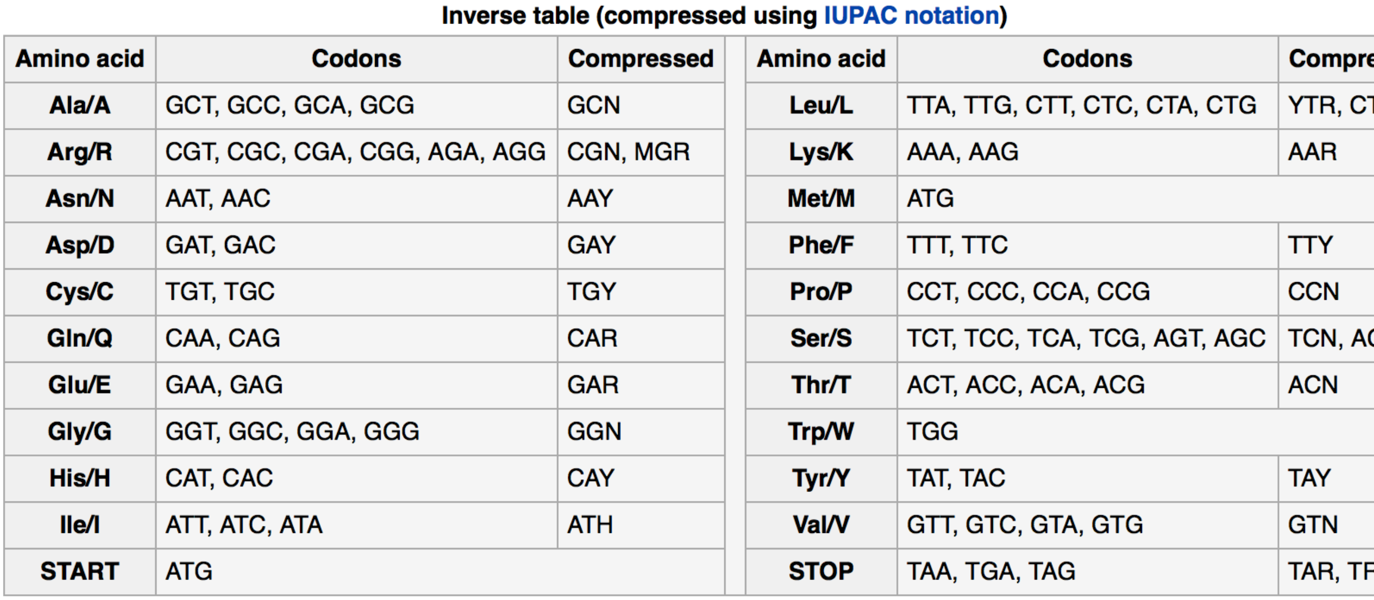
There are a couple of strategies depending on the organism that you are working on. Obviously, in this case, we are interested in *Saccharomyces cerevisiae*, a model organism, and you have been given the name of the gene(s) so you could look up the sequence of the gene by doing a BLAST search. For a model organism, the process of PCR primer design reasonably straight-forward and, besides the considerations outlined in the next section, mostly involve determining the complementary sequence flanking the gene and ensuring the correct polarity of the primers to allow PCR to amplify the gene of interest.

In the future, if you were working on a "non-model" organism and didn't have access to genomic resources, you could use deploy a strategy called "degenerate PCR" or "degenerate primers". Since we don't need to rely on the uncertainty of degenerate primers, I won't spend much time discussing them or the specifics of how to optimize them but they do offer an insight into the process of primer design and are worth at least a cursory understanding.

Remember how you learned in earlier biology courses that multiple triplet nucleotide codons can result in the same amino acid? For instance, the following DNA triplet codons: CGT, CGC, CGA, CGG, AGA and AGC are all translated into arginine. Unlike specific primers which will only bind to particular complementary nucleotides, degenerate primers are designed to allow some "wiggle room". Depending on your organism, the sequence of the gene of interest may be found in other related better-studied organisms whose sequence information is found in the NCBI database. You can use the amino acid sequence information and use a program such as CLUSTAL (or whatever your favourite/optimal program ends up being in the future) to align the gene from multiple related organisms. You can then use this information to design multiple primers that code for the same amino acids but with different triplet codons (hint: obviously, you want to try to avoid including highly degenerate amino acids such as arginine, serine, and leucine into the construction of your primers since they will give you too many possible primer sequences). The assumption, of course, is that at least one of these many primers will bind to your gene in your understudied organism and allow for amplification.

Lucky for you, *Saccharomyces* is a model organism that has a rich database of sequence information. This means that you will not be forced to use a degenerate primer but can, instead, opt for a primer that demonstrates specificity (meaning it will amplify only your segment of DNA or gene). Of course, it is important to ensure that your primers will only bind to the gene that you want them to bind to and not accidentally bind to other regions of the genome.

Even when you want *fairly* specific primers, it is possible to not want complete specificity. After all, what happens if you want primers that will amplify a gene that is polymorphic in



Inverse table (compressed using IUPAC notation)

Amino acid	Codons	Compressed	Amino acid	Codons	Compre
Ala/A	GCT, GCC, GCA, GCG	GCN	**Leu/L**	TTA, TTG, CTT, CTC, CTA, CTG	YTR, CT
Arg/R	CGT, CGC, CGA, CGG, AGA, AGG	CGN, MGR	**Lys/K**	AAA, AAG	AAR
Asn/N	AAT, AAC	AAY	**Met/M**	ATG	
Asp/D	GAT, GAC	GAY	**Phe/F**	TTT, TTC	TTY
Cys/C	TGT, TGC	TGY	**Pro/P**	CCT, CCC, CCA, CCG	CCN
Gln/Q	CAA, CAG	CAR	**Ser/S**	TCT, TCC, TCA, TCG, AGT, AGC	TCN, AC
Glu/E	GAA, GAG	GAR	**Thr/T**	ACT, ACC, ACA, ACG	ACN
Gly/G	GGT, GGC, GGA, GGG	GGN	**Trp/W**	TGG	
His/H	CAT, CAC	CAY	**Tyr/Y**	TAT, TAC	TAY
Ile/I	ATT, ATC, ATA	ATH	**Val/V**	GTT, GTC, GTA, GTG	GTN
START	ATG		**STOP**	TAA, TGA, TAG	TAR, TR

Figure 4: Degeneracy of Codons in DNA - stolen from wikipedia

the population under study? What about wanting to use the same set of primers to study the same gene between species? It is possible that you may want to use a highly related set of primers that differ at one or more nucleotides and thus allow for some latitude in the complementarity of the target sequence.

How many nucleotides long should a primer be in order to ensure 100% specificity?

It is generally agreed that the optimal length of primers should be approximately 18-22 nucleotides. What would be the minimum length of primer to be specific to your particular gene? We can tackle that question with a bit of cartoon scenario: The genome of *Saccharomyces cerevisiae* is 12,495,682 nucleotides long. We will assume that all nucleotides are present at 25% (not a correct assumption, by the way).

Let's start with a primer of length 8, 5'-ATGCATAC-3'. This octomer would find a complement, due to random chance, $(0.25^8) * 12,495,682$ almost 191 times.

So, lets add 2 more nucleotides. If you had a 10 nucleotide primer for the following made-up sequence: 5'-ATGCATACGC-3', you might expect, by chance, that this decamer would be present $(0.25^10) * 12,495,682$ almost 12 times. How many nucleotides would we need to have in a primer to ensure that it was present just once by chance?

$$(0.25^n) = 1/12,495,682$$

We then take the *log* of both sides and divide by *log*(0.25) to get:

$$\begin{aligned} n * log(0.25) &= log(1/12,495,682) \\ n &= log(1/12,495,682)/log(0.25) \\ n &= 11.787 \end{aligned}$$

Of course, to be completely careful we may wish to add some nucleotides to the primer length especially if it is possible, based on your specific sequence, to end the primer with two C and Gs since they will form stronger bonds to the single stranded target DNA than A and Ts would (you don't usually want more than 2CG or they could bond to each other instead of to the target DNA). We don't want to make the primer too long; the 18-22 nucleotides length is a trade-off; after all, shouldn't the primer be as long as possible to ensure that it only binds to the intended gene? Well, no - the primer still needs to be short enough that the primers anneal to the intended target gene instead of binding to each other under the heat required for PCR.

What other considerations go into designing primers? As already mentioned, there is an entire website dedicated to designing primers specifically for *Saccharomyces*: `http://www.yeastgenome.org/help/analyze/design-primers` This website lists the following criteria - and gives its default values -when designing primers for either PCR or for sequencing. Of course, some of these criteria will influence how far away from the target gene sequence start site you design your primer to bind.

Melting Temperature - T_m is the temperature at which half of the double stranded DNA is single stranded. This means that at temperatures $> T_m$, the DNA is single stranded and when temperature of the reaction is $< T_m$, the DNA is double stranded. In order for the primers to bind to the target DNA, the reaction temperature must be approximately 5 degrees $<$ than the T_m of the primers which should, in turn, have similar T_m. The T_m is dependent on sequence content according to several approximate formulas including the "Wallace formula": $T_m = 2(A + T) + 4(C + G)$.

A/T and C/G composition - As already mentioned, A-T nucleotides possess two hydrogen bonds whereas C-G possess three. Besides determining the melting temperature of the PCR reaction since the T_m governs the point at which primers will bind to the denatured single stranded target DNA, a primer should aim to be approximately 50% A/T and 50% C/G. Primers should also avoid long stretches of either C/G or A/T so that they don't accidentally create opportunities to anneal to themselves and result in primer dimers or hairpin loop structures.

***Taq* polymerase** - the forward and reverse primers should be approximately 400-1000 bp apart. This is partially a result of the ability of the special *Taq* polymerase we use during the PCR reaction. *Taq* polymerase still works at the high temperatures needed to separate the target DNA, it's optimal temperature range is between 70-75 degrees, but it only amplifies regions that are 1000-2000bp long.

Other options included in the website's algorithm include: Location of start and stop codons in a gene (obviously to amplify or sequence the gene you will want to choose a primer that attached upstream of the start of the gene), primer composition (AT/CG ratio), primer melting temperature and primer annealing. The website also gives optimal values for these criteria based on complex formula (links are listed on the site), is explicit about default values that you may change if you wish and returns a ranking of optimal primers. **Okay but now: how do we actually design primers?** The following steps are highlighted:

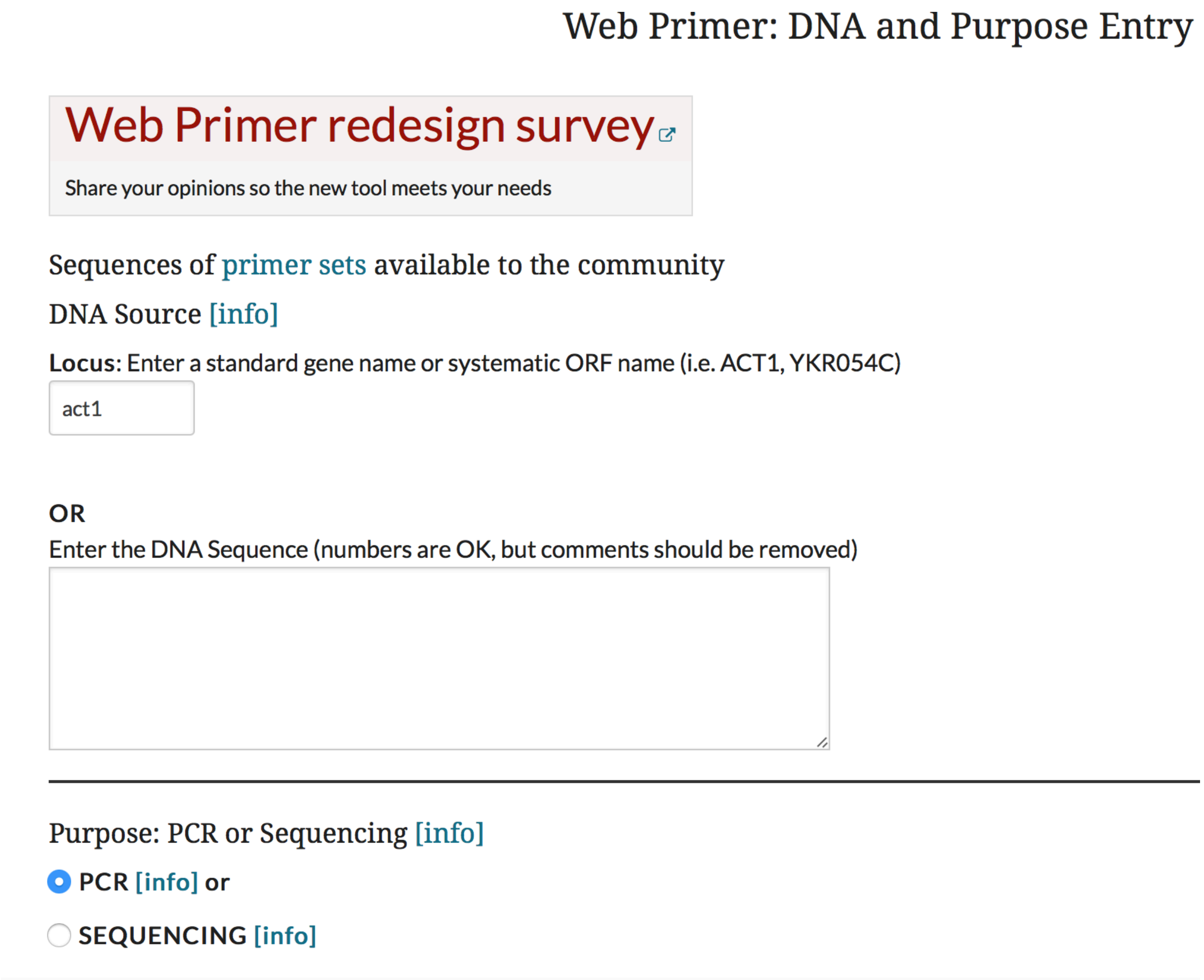
- You have been given two genes to sequence and amplify; enter Locus Name into the box found here: `http://www.yeastgenome.org/cgi-bin/web-primer`
- You will need to use primers that are specific for the PCR process. Remember that the criteria discussed above. There are particular default values for the distance upstream from the start and stop codons of the named locus.

1.5 Worked Example

We will take the act1 open reading frame (also called YFL039C) which produces a protein called actin. You can find a summary of the ORF here: `http://www.yeastgenome.org/`

`locus/S000001855/sequence`. For a summary of your own gene, type the name you have been given into the search bar. This will return an overview of your particular ORF including information such as where it is located (chromosome I, II, etc), how long it is and what function it has (if known). If you want more detailed knowledge about your particular sequence you can simply click on the “Sequence Details” tab. This will take you to a page that gives you any introns and CDS included in the sequence file.

Now, we will use the built in primer design feature of this website. Under the “Analyze” pull down menu, select “Design Primers”. You can then enter the locus name you have been given into the window.



Web Primer: DNA and Purpose Entry

Web Primer redesign survey

Share your opinions so the new tool meets your needs

Sequences of primer sets available to the community

DNA Source [info]

Locus: Enter a standard gene name or systematic ORF name (i.e. ACT1, YKR054C)

act1

OR

Enter the DNA Sequence (numbers are OK, but comments should be removed)

Purpose: PCR or Sequencing [info]

● **PCR** [info] **or**

○ **SEQUENCING** [info]

Initially, use the default settings of the primer design. In the case of act1, it gives the following output:

```
Choosing primers near act1 for PCR

There were 32 forward primers in the valid range for GC content.
There were 64 reverse primers in the valid range for GC content.
There were 3 forward primers in the valid range for melting temperature.
There were 29 reverse primers in the valid range for melting temperature.
There were 3 forward primers with valid self anneal values.
There were 29 reverse primers with valid self anneal values.
There were 3 forward primers with valid self end anneal values.
There were 28 reverse primers with valid self end anneal values.
There were 84 pairs of valid primers.

This is the BEST pair of primers
List of all valid pairs of primers
```

Scroll your mouse over the ‘BEST’ and you will be rewarded with your ideal primers under the default criteria. You can also scroll across the ‘all’ and you will be given a table with all the primer pairs that fit the criteria.

Question 1: What are the sequences of your forward and reverse primers?

Question 2: Compare the fragment that will be amplified by these primers to your FASTA file sequence for your gene. At what position (defined by the numbers in the sequence of your FASTA file) will the amplification start?

Remember that the resources available on model organism websites such as yeastgenome and flybase are usually available through the central clearing house of **NCBI**. Primer design is no exception. Once you are on the ncbi website, you can choose **’Sequence Analysis’** from the side menu or just type in: `http://www.ncbi.nlm.nih.gov/guide/sequence-analysis/`. Under the **’tools’** section, you will find the option for **Primer-BLAST** - the tools are listed in alphabetical order. By clicking on **Primer-BLAST**, you will be able to conduct a primer design similar to the one you have already conducted on yeastgenome but with somewhat different specified default values.

The yeastgenome website includes an alignment program which allows you to align multiple strains of *Saccharomyces*. It can be useful to see how much variation there is between strains since that will govern your decision to design multiple primers or not (if there is a lot of variation, you will need to design at least two primers for the different variant species for the same gene). The reference strain is called **S288C**. Under the ’Resources’ section of your ORFs overview, you will find a link titled “S288C vs. other strains”. You can click on the ”strain alignment” to open up a window that includes an alignment of the ORF with 42 other strains as well as a tree produced from the program “ClustalW’. You are able to control whether you prefer to built an alignment (and subsequent tree) from protein sequencing or from nucleotide sequencing.

Question 3: Do you expect that the alignment based on protein sequence will produce a different tree than the alignment produced from nucleotide alignment? Why or why not?

Finally, we can check to ensure that the primers we have designed will not form any of the problematic dimers or hairpin structures and thus reduce the amount of amplification possible with PCR. You can check this by using the website `https://www.idtdna.com/calc/analyzer`. You want to ensure that the hairpin structures are not formed in the range of the T_m for the annealing reaction.

Lastly, you will want to use this website, `http://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi?LINK_LOC=BlastHome`, to run a search against your primers and make sure that they are specific to one genomic region.

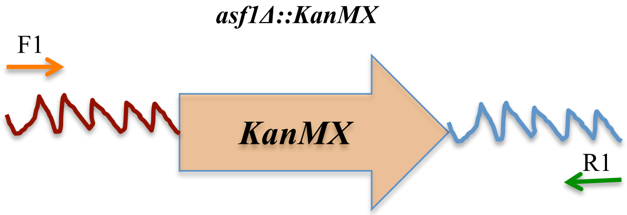
1.6 Designing primers for your very own assigned gene!

We are now going to amplify a KanMX gene that is present in a special strain where it is attached to the promoter and the 3' flanking from one of your fourteen genes-of-interest. This KanMX cassette fragment will be used to replace the target gene via homologous recombination. This results in a genome with your gene deleted, since it has been replaced by a marker gene, KanMX. We will use two forward primers and one reverse primer and we will run a gel to ensure that our recombinant has replaced the ORF. The basic process is outlined in the following diagram:

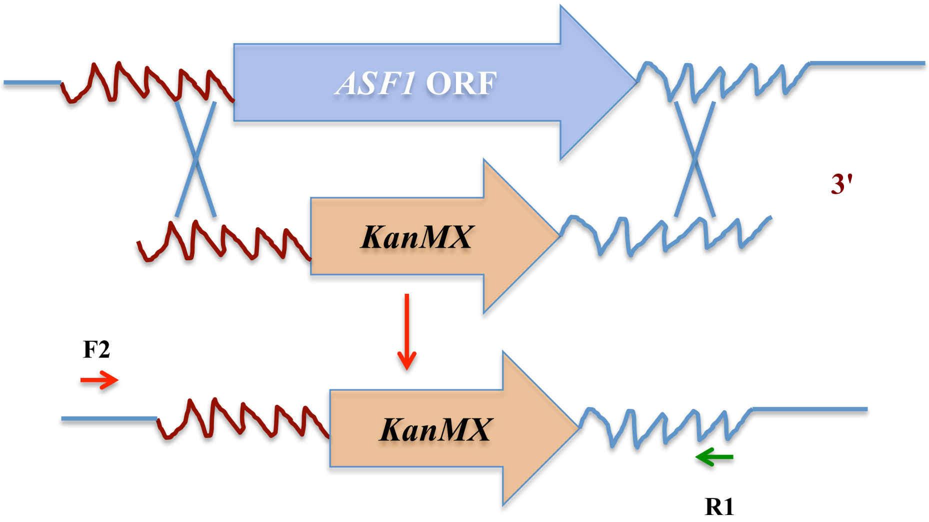
Replacement of your ORF (gene of interest, e.g., *ASF1*) with the *KanMX* marker:

The *KanMX* cassette confers resistance to the antibiotic geneticin (aka G418), and is widely used as a marker gene to replace (aka knock out) a gene of interest. For example, replacing the open reading frame (ORF) of *ASF1* gene results in a locus referred to as *asf1Δ::KanMX* at which the 5' and 3' flanking sequence *of ASF1* ORF is linked to *KanMX.* Researchers have individually knocked all nonessential yeast gene (~4000 in total) with *KanMX* creating a KO yeast library. This library provides a convenient resource of KO templates (genomic DNA) from which one can use PCR to amplify a KO cassette (e.g., *asf1Δ::KanMX*), and use it to knock out the gene of interest (*ASF1*) in another strain as follows.

Step 1:
Amplification *asf1Δ::KanMX* cassette using forward and reverse primers (F1 and R1) homologous to 5' and 3' flanking sequence of *ASF1*. The DNA template would be genomic DNA from the *ASF1* KO strain in the KO library.

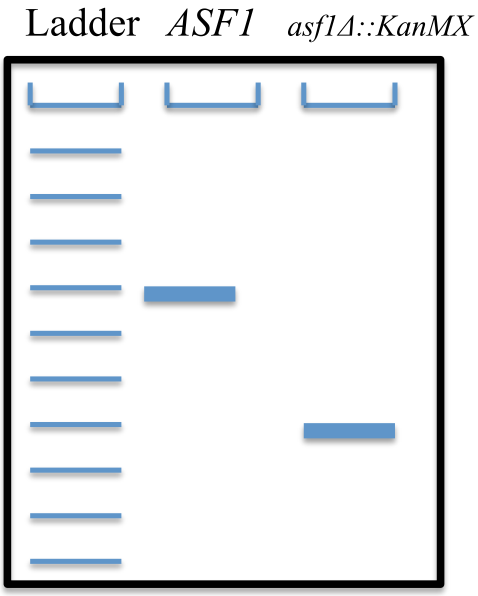


Step 2:
Transform strain YBIO268-I with the PCR-amplified *asf1Δ::KanMX* sequence. Homologous recombination between *asf1Δ::KanMX* sequence and the 5' and 3' flanks of *ASF1* in BIO268-I would result in the replacement of *ASF1* ORF with KanMX.



Step 3:

To test the existence of the *asf1Δ::KanMX* in the transformant, we will perform PCR with primer R1 and a 3rd primer F2 that is upstream from the F1 primer, using genomic DNA from the transformant as template. If the transformation has worked and caused the *ASF1* ORF to be replaced with *KanMX*, we expect that there will be length differences between the PCR fragment that result from non-transformed (still containing the *ASF1* ORF) strain and the transformant. *Gel electrophoresis of the PCR product can be done to confirm the expected length of PCR product from the transfromant in comparison to the expected length of PCR product from the parental strain.*



With all of the above outlined features of good primers, we will want to ensure that our three primers have the following attributes:

1. approximately 20 nucleotides long
2. approximately 50% of GC content
3. the two forward primers should be non-overlapping and fall within 300bp - 500 bp of the ORF
4. the one reverse primer should be within 300bp-500 bp of the ORF
5. the last 5 bases of the forward primers should have a “GC” clamp of 2-3 G or C nucleotides
6. Tm of approximately 52.5 degrees Celsius so that everyone is within the same range and we can run all of the reactions together simultaneously.

We can then return to the primer design website: `http://www.yeastgenome.org/cgi-bin/web-primer`, cut and paste our gene of interest and pick out two non-overlapping forward primers and one reverse primer. We can then return to the following website to analyze the properties of our primers: `https://www.idtdna.com/calc/analyzer`.