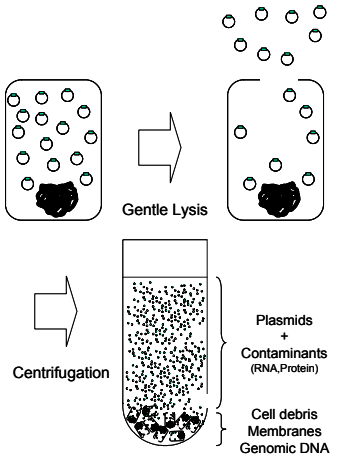
Recombinant DNA technology

1) Introduction

The various economic and public issues regarding genetic engineering are currently subject to considerable debate, but the technique is far more important for the fundamental biology of microorganisms, plants and animals then it is for crop improvement and applied biology. Currently, many plant processes are studied by manipulating genes and studying the consequences. In contrast to classical plant breeding where little is known about the exact nature of the mutation(s) that lead to a selected trait, genetic engineering means knowing exactly what you do on the molecular level, even though you may not be certain regarding the outcome. However, since you know what you have done to a gene, at least you can establish a causal lik, and that's what science is all about. In order to manipulate genes, they first need to be identified and cloned, and there must be some experimental data to predict which genes are likely to play a role in the process which is studied. But gene identification is an advanced topic and we first have to take a step back and understand the principles of basic genetic manipulation, before we can address how genes are cloned in the first place. The following pages will address this in very simple terms, but yet providing a good introduction to more detailed aspects of genetic engineering which you may learn in other courses.

2) Working with DNA, how to start?



First we need to extract DNA: The entire principle of genetic engineering and recombinant DNA technology relies on a very simple principle, the fact that bacteria like E. coli can contain small circular DNA molecules, termed plasmids, which replicate independently of the genome and are present in many copies in the cells. Because they are small, they are also soluble, in contrast to the genomic DNA, which is often called "chromosome", although it has little in common with eukaryotic chromosomes. The chromosome forms a large cluster, it is usually attached to membranes in the bacterium and it is much less soluble than a plasmid. Plasmids are present in many micro-organisms to provide extrachromosomal DNA that can be easily transferred from organism to organism. Plasmids encode a variety of abilities, like tolerance to certain antibiotics, heavy metals or even the ability to pathogenise another organism. In the laboratory, they are used by molecular biologists as cloning vehicles, because they can be easily extracted from E. coli without contaminating genomic DNA and they replicate fast, creating identical copies (= cloning). Laboratory plasmids are usually smaller, and have little left of their original ancestors, except for an origin of replication (see semester 1, DNA replication lectures) and an antibiotic resistance marker (often resistance to ampicilline).

If one starts with a single bacterium on a Petri dish, it grows into a small colony containing 10^6 identical individuals. Each of these bacteria will contain 100 or more identical copies of the same plasmid. This means, the small colony represents 10^8 copies of a plasmid. Unfortunately, this is still not enough to work with, so it is necessary to grow E. coli in a liquid culture. For this purpose, a single colony is used to inoculate a 3 ml culture and grown overnight on a shaker at 37°C to obtain a nice milky culture full of bacteria. This will then be 10^9 bacteria, and 10^{11} plasmids. With this starting material, it is possible to work.

The cells are centrifuged down in small tubes, the supernatant is discarded, and the pellet is then extracted using a variety of protocols which all follow one common principle. The cells are broken up gently, to release only soluble contents (including the plasmids, but not the genomic DNA). Further steps are then designed to remove contaminants, like proteins and RNA which are also

soluble, and then to concentrate the DNA by precipitation with alcohols and salt, followed by resuspension in a small volume of an aqueous buffer at slightly alcaline pH (7.5-8). One can then visualise the DNA on a gel. The detection limit of this method is approximately 10 nanograms (nano gram means 10^{-9} grams). For a plasmid of 3000 base pairs (we say 3kb), this means 3.000.000.000 copies ($3x10^9$), so if we have extracted potentially 10^{11} copies (less whatever remains with the cell debris in the pellet), we have enough to run a gel approximately 30 times and detect DNA bands each time. The DNA can be manipulated, cut into smaller pieces, individual fragments can be purified, and ligated together with other fragments, and new plasmids can be built (see later).

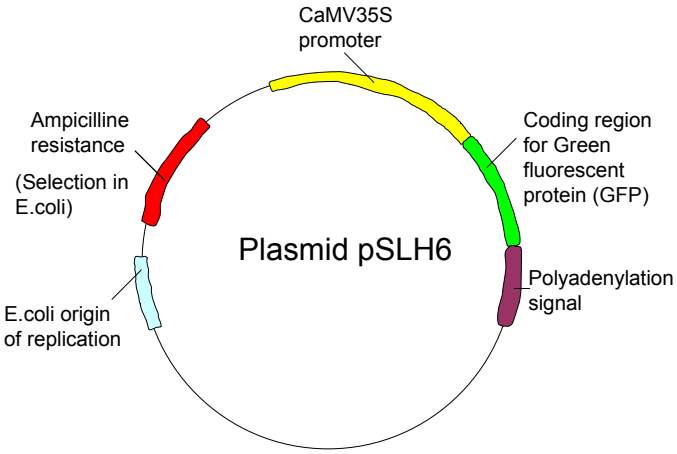
Transformation: E. coli can also be re-introduced into E.coli via transformation. This is a very important technique, because after modifying the DNA in a plasmid, it is necessary to make more copies of this new recombinant plasmid, and you can only make more copies of a plasmid by letting E. coli produce them for us. E. coli can thus be regarded as the manufacturing unit that makes plasmids for us. Transformation is based on the fact that E. coli cells bind to DNA, particularly when there are Calcium ions in the medium. We call these Calcium washed cells 'competent cells', and produce them using specific protocols. It involves growing a culture to the exponential phase and then washing it with ice cold Calcium buffer several times. They are kept on ice and allowed to bind to plasmid DNA for 15 minutes. Then, we subject the cells to a heat shock at 37°C for a few minutes. During the heat shock, cells will start taking up nutrients again, and it is believed that DNA is taken up in parallel to such transport processes. This is very inefficient, but approximately 1 in 100.000 (10^5) plasmids makes it into an E.coli. To get more than just one transformation event, you should use much more than 10^5 plasmids, more like 10^9 or 10^{10}, and you have to use enough bacteria too (usually also 10^9 at least). All these bacteria are then plated out on a special medium, with the purpose to let only those bacteria grow that have taken up the plasmid.

Resistance markers: How can we see that transformation is successful? Well, first of all, we introduce plasmids into an E. coli strain that does not yet contain a plasmid. The plasmid contains resistance markers, that allows the bacterium to grow on a certain type of antibiotic. Without the plasmid, the E. coli cells will die when the antibiotic is present. So if our plasmid contains a gene that encodes ampicilline resistance, we use competent E.coli which do not have such a plasmid and are susceptible to this antibiotic. After the heat shock, the 10^9 cells are plated out evenly on a plate containing ampicilline. Most of the bacteria will die, and a few thousand colonies are formed after an overnight incubation. Such colonies will then be ampicilline resistant, they each originated from a single bacterium that was lucky to receiv a plasmid, and you can use the colonies to inoculate a liquid culture, and then extract plasmid again. Whenever you run out of a plasmid preparation, just re-transform E.coli, grow op a culture and extract and purify plasmids again. But most importantly, it is possible to change a plasmid first, and then introduce it into E.coli to make a lot of this new plasmid. That is the basic cycle of events in standard recombinant DNA techniques.

Genetic engineering: The process of extracting a plasmid, changing it, transforming the changed molecule and extracting the modified plasmid again in large numbers is called "sub-cloning". It is the most basic form of genetic engineering. The term cloning is appropriate, because of the very large number of copies required to do this (to see the plasmid on a gel). We call it sub-cloning because the plasmid with which you start is already cloned. We say "the plasmid", but we mean 10^{11} identical copies (clones) of it, always remember this!
Oh, and did I say that sub-cloning is easy? Well, it is, if you compare it with cloning a gene from an animal or a plant for the first time. The reason for this is that the genes are present on the chromosome, they don't replicate like plasmids, and instead of dealing with 3kb, you have a genome of 3.000.000kb with 100.000 genes in it, and you don't know where your gene is. Getting such a gene into a plasmid, together with the 100.000 others, and finding the right one, is called cloning, and that is far more difficult and we will discuss this once you have understood what can go wrong with simple sub-cloning strategies. Now that you know how to extract a plasmid and re-introduce it into E.coli, let's discuss how we carry out a modification to a plasmid.

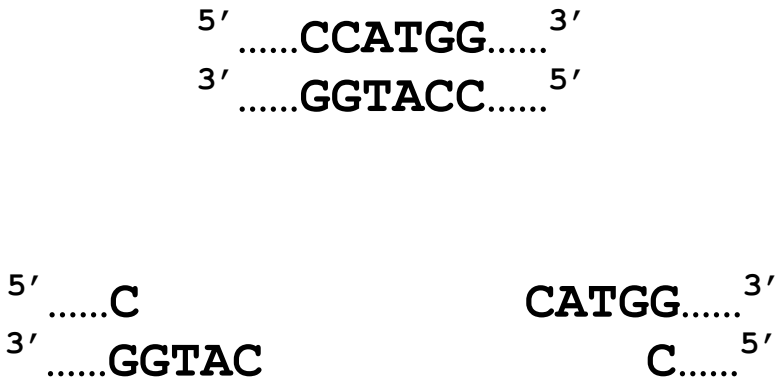
3) Simple subcloning

Structure of a plasmid: A typical laboratory plasmid is approximately 5 kb. That is because a gene of interest has already been cloned into the basic plasmid backbone (usually just less than 3 kb). The diagram shows a typical example of a plasmid (pSLH6) that carries a gene which we would like to re-construct. It has an origin of replication that works in E. coli and drives chromosome-independent replication to yield at least hundred copies in each E.coli cell. Then it carries a gene that codes for ampicilline resistance. These two plasmid elements are found in many plasmids, but our actual example contains a chimeric gene as well, our gene of interest. It has already been assembled from the three basic parts, a promoter, a coding region and a 3'end carrying a polyadenylation signal. This tells us that the gene is designed to work in a eukaryotic cell (yeast, fungi, protists, plants, animals).
Chimeric means hybrid, it has portions of genes from different origins, fused together in a way that is not found in nature (unless you consider human activity one of the natural processes on this planet, a semantical issue). The promoter is from a virus gene that codes for the 35S coat protein of cauliflower mosaic virus. The coding region is from a gene that encodes green fluorescent protein from a jellyfish, and the polyadenylation signal comes from a plant gene.The promoter is designed to work in cauliflower, which is infected by the virus, another example of an entity that can genetically engineer plant cells to manufacture more copies of itself. However, the promoter will also work in other plants, and because it is a very strong promoter, it is used to express certain genes. The coding region encodes the green fluorescent protein (GFP), a useful tool in modern cell biology, and like any other coding region, it starts with ATG, it ends with either TAG, TGA or TAA, and inbetween are the various codons that dictate the primary structure (= amino acid sequence) of the protein (see semester 1 or consult any genetics text book if this is difficult). The polyadenylation signal is essentially a piece of DNA that contains signals for the processing of the transcript to generate the final messenger RNA that can be transported out of the nucleus. These three building blocks build a eukaryotic gene, and because in this case they have been fused together from different sources, it is a chimeric gene.



Restriction sites: The plasmids contain a variety of restriction sites. To avoid confusion, only a few are shown here, and they are unique (which means the sites occur only once in the plasmid). Restriction sites are DNA sequences recognised by restriction enzymes. One or two types of them are normally present in bacteria, as part of a protection mechanism against viruses. When a virus (bacterial viruses are usually cause bacteriophages, or simply phages) injects its DNA into the bacterium, the DNA is cut by the restriction enzyme at those places where the sites occur, which will destroy the functionality of the DNA. The bacterial genomic DNA has much more sequence in it, and of course will also contain these sites, but to avoid that its own DNA is cut, the bacterium has specific enzymes, the so-called methylases, which will modify the sites so that they are no longer recognised and cut. This way, only foreign DNA will be cut into pieces. Scientists now use these restriction enzymes to cut DNA, in fact, you can buy tubes with restriction enzymes from companies these days. The E.coli strains used are mutants that don't produce restriction enzymes, so that our plasmids are not cut into pieces. Even methylase deficient strains exist to prevent methylation of a subset of restriction sites. So everything is quite artificial these days, but all the tools are found in nature.

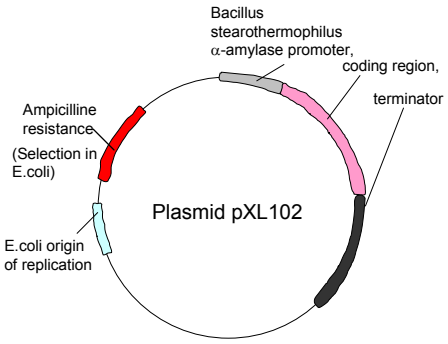
A typical restriction site is NcoI, as shown below:

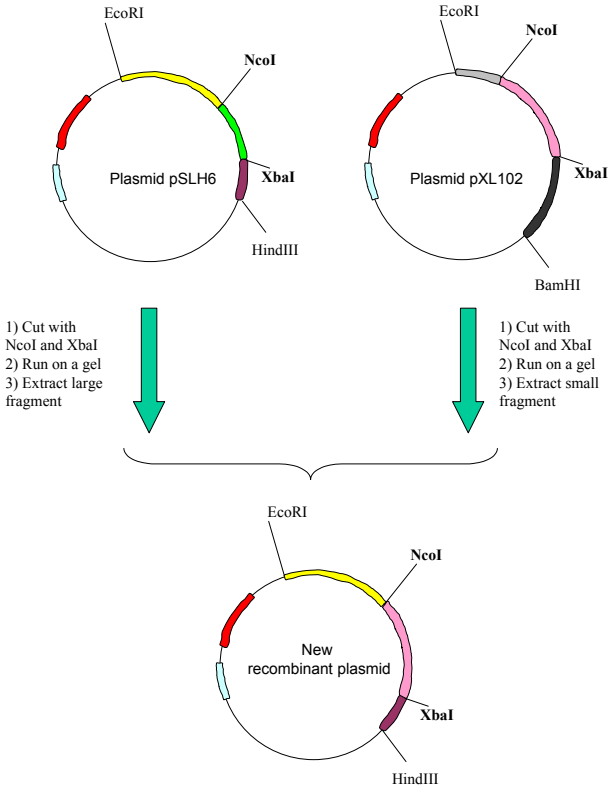


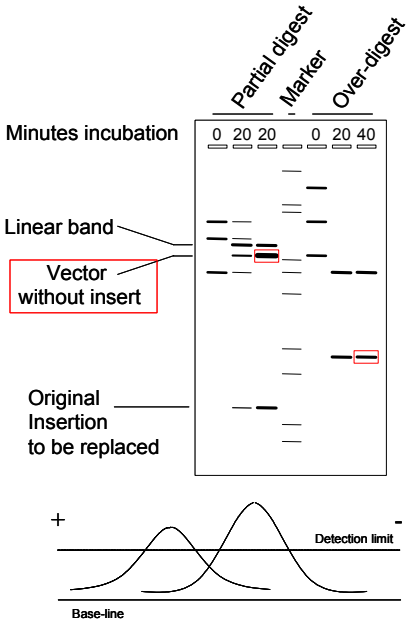
The arrows indicate where the enzyme cuts. After cutting, you get two 'sticky' ends, as we call them. NcoI is very popular, because it spans an ATG sequence, and is often engineered at the beginning of the coding region. Yes, your guess was right, it wasn't an accident, it is done on purpose and when you look at genes in nature, they don't always have this restriction site in this position, only by accident. The probability to have this site is one in 64, as each base pair has a probability of one in four to occur, and you need 3 base pairs in conjunction with the ATG to get the NcoI site. The plasmid shown has been built by researchers before, and the letters refer to the initials of the person who did it, the number is usually a running number. In this case, it is the 6th plasmid made by Sally L. Hanton, one of many graduate students who used to work with me in Leeds. Biologists like to use terms and create new names, like when someone clones a new darkness-induced gene, it could be *di1*, for instance…It is a pain to learn all these terms and new names, but think of it, if you were to clone a new gene after a year or two of hard research work, wouldn't you want to give it a name, so that people remember it (and perhaps also who was the person who achieved this absolutely stunning piece of work)? Let's start exploring how to reconstruct something in this plasmid, and be aware of the fact that this is easy compared to making a plasmid for the first time.
The probability to get an NcoI site anywhere is one divided by 4^6, one in 4096. This explains why small plasmids of a few thousand base pairs contain many unique restriction sites. This is also the very reason why it is possible to engineer small plasmids by cutting and pasting DNA, whereas it is not possible to do this with genomic DNA or with very large plasmids, such as the Ti plasmid of Agrobacterium tumefaciens. The other site is XbaI (TCTAGA), also very popular because it spans one of the three possible stop codons (TAG) of a coding region. When we write the sequence of restriction sites, we always write it from 5' to 3' direction (see semester 1 or consult a genetics textbook about 5'Phosphate and 3'OH ends) and we only write the upper strand because the lower one can be deduced from it. So the sequence of XbaI is TCTAGA (and spans a stopcodon: TAG). Cutting with XbaI also generates a sticky end, but it does not fit into the NcoI site (we say that it is incompatible). There are also restriction sites which cut in the centre and create the so-called "blunt" ends. EcoRV for instance, it cuts GATATC exactly in the middle. Blunt ends can also be formed when sticky ends are degraded by nucleases, or when DNA breaks randomly.

Subcloning: The coding region is flanked by NcoI and XbaI. If we now have another gene in which the coding region is also flanked by NcoI and XbaI, it is easy to cut this coding region out and exchange it in our plasmid. Such a second plasmid is shown as well (pXL102), obviously made by someone who has made a lot of plasmids… This plasmid contains a bacterial gene, encoding a thermostable enzyme α-amylase (digests starch). Bacterial genes don't contain polyadenylation

signals. Instead, they have a transcription stop signal, which stops the RNA polymerase from synthesising more RNA than necessary. If we would like to produce a thermostable α-amylase in plants, we have to make a new gene, because the bacterial gene will not work in plants (the promoter will not work and even if it would, the transcription terminator won't). However, the genetic code is conserved between bacteria and plants, so we can use the coding region to make the thermostable α-amylase, but we need a plant promoter and a plant polyadenylation signal. Therefore, we will replace the coding region for green fluorescent protein by that of α-amylase. The diagram shows how this is done. We cut both plasmids with NcoI and XbaI, and when digestion is complete, the plasmid fragments are separated on an agarose gel. This is done via an electric field, the DNA is negatively charged and will migrate through the gel towards the positive electrode. The bigger the DNA molecule, the more time it needs to do this, so you get a separation by size. It is now possible to cut out the appropriate bands and elute DNA from the gel pieces.







You will notice that in the gel example shown on the right, the plasmid pSLH6 was not completely cut after 40 minutes of digestion. You can see that in addition of the two fragment, you also get a linearised fragment that is the sume of the two fragment below. This would be bad and one would have to incubate longer until the linearised fragment is completely digested.
The large fragment contains the origin of replication and the ampicilline resistance, and we usually call this the ***'vector'***, because it is the replicating entity and could exist on its own in E.coli if it were circular. The small fragment cannot replicate, so it can never give rise to colonies on a plate, hence it is what it is, the ***'fragment'***. The fragment is in this case the small band from plasmid pXL102.

'Vector' and *'fragment'* can then be mixed and 'ligated' with the enzyme ligase. This enzyme can create covalent bonds between the 3'OH and the 5'phosphate group on the DNA ends (consult any genetics text book) with the consumption of ATP (which also has to be added to the reaction). This is done in a tube, and if you have done it the right way, you will get the recombinant plasmid shown at the bottom of the figure. In fact, you will get lots of them, because you do this type of reaction with 10^{10} molecules, and even if only some of them will assemble in the right way and ligate, it will still be lots of them (10^8 for instance). If this mixture is now transformed into competent E. coli (remember?), one or two thousand ligated plasmids will make it into the cells and start replicating, allowing the E. coli to grow on ampicilline and create colonies. Each colony will have one type of plasmid, and if you take one of these, grow up a liquid culture and extract plasmid, you can then check if the new plasmid is indeed what you wanted. You can use the restriction enzymes to cut the plasmid, and you get a characteristic pattern, which should allow you to distinguish it from the two parental plasmids. If this turns out to be right, you made it! You give it a new name, put a sticker on the tube, and tell your boss that you've got a new construct. All this takes just three days of half time work, piece of cake!

Is it really so easy? Not exactly, there are a few things that can go wrong, even in this simplest type of sub cloning work. There are a few things to remember as it will help to understand problems in the future. First of all, it should be noted that the sharp bands you get in a gel give the wrong impression that the molecules are perfectly separated. Every "band" is merely the peak of a distribution curve which is above the detection limit of the imaging device. This is because you have so many molecules and they all behave slightly differently, they vibrate through Brownian motion, and this is a random-walk in all directions that will show extreme cases as well as average behaviour, particularly when you have lots of individuals. Super-imposed on the general tendency to move directionally within the electric field, the molecules will diffuse in different directions. In fact, some of the molecules will be at the beginning and the bottom of the gel, so there is cross-contamination in each of the bands. They will be below the detection limit, but if this detection limit is $3x10^9$ molecules, you could still have millions in it. In other words, when you mix the two fragments you want, you will have the other two as well, albeit in much lower concentrations. This can give rise to other molecules, the so-called **by-products**.
Secondly, it is possible that you didn't cut the plasmids completely. Imagine that only one of the two enzymes cut 100%, whereas the other only did 90% of the work. 10% of all plasmids will have been just linearised. The linearised band is hardly different from the band the remains after cutting away the small fragment, so they are very close together as you can see on the figure. You may not see it, because it is below the detection limit, but it could be there. And because they are so close together, the cross-contamination will be quite serious. A linearised plasmid will have only one type of sticky end, and this can very easily self-ligate because it is a compatible sticky end. **Self-ligation** is an intra-molecular reaction, and you can imagine that the probability is much higher for this to occur, than the inter-molecular ligation reaction where two separate fragments have to get close to each other and then ligate. The self-ligation does not have to wait until the two ends are close to each other, they are by definition close because they are on the same molecule. So this is a very severe by-product that can occur with such high frequency that it prevents you from finding the right clone (needle in haystack).
The worst problem is that a small portion of the plasmid was not even cut at all in the first place. Plasmid DNA can exist in different forms (that's why uncut plasmids are never sharp bands), open linear, open circular, multimer and the so-called supercoil. In the supercoil, the DNA is folded up very tightly, and if you are unlucky, the restriction sites are buried in the centre of the macromolecule. The restriction enzyme will not reach it and it will never cut, no matter how long you wait. The supercoil runs at variable positions on the gel, and is usually not very far from the large vector fragment that contains the majority of the plasmid. To make it worse, it does not stain very well with Ethidium bromide (the chemical with which we stain DNA so that we can detect it with UV light). So you may not see it, but it could be present in your fragment preparation. Even if this is only 1% of the total, completely undetectable to the eye, it will mess up everything for the following reasons:

When competent E. coli is transformed with plasmid DNA, circular molecules (ligated molecules, or molecules that were never cut in the first place) will have a much higher chance to establish themselves and replicate than linear molecules. This is because the open ends of linear DNA are immediately recognised by aspecific nucleases which break down the DNA in a matter of seconds. This does not happen to circular DNA, and when it starts replicating, nobody can stop it any longer. The only way for linear DNA to maintain itself in E. coli is to be lucky and to be ligated by E. coli ligase before the nucleases reach it. This is very unlikely, and therefore we say that linear DNA 'gives no colonies'. But imagine you have some un-cut circular DNA in your ligations, this will create so many colonies that you will never find the recombinants.

The need for control ligations: These three examples illustrate that you can get by-products, and to control for this, we generally do some controles when we do the ligation. In the case of the example shown, it will be two controles:

Nr.	vector	fragment	water	10x buffer	ligase	Reason
1	1μl	-	17μl	2μl	-	test for uncut vector
2	1μl	-	16μl	2μl	1μl	test for partially cut vector
3	1μl	1μl	15μl	2μl	1μl	the real thing (hopefully most colonies are found here

The first two ligations are control ligations, only the third one will give rise to the desired recombinants (and all the by-products). In fact, the third ligation will give rise to the sum of all the colonies on plate 1 and 2, plus the additional ones from the desired recombination. If you have an extremely good vector preparation, you will only get colonies in plate 3, but this only happens in text books.
The result is usually as shown in the illustration. The example is from a real case in my laboratory from a visiting student who did this for the first time. It is important to check several colonies, make liquid cultures, extract DNA and then test them via restriction analysis.In our case, all of the colonies contained the right recombinants. One could have guessed that, because there were a lot more colonies on the last plate, but the student wanted to be sure because it was her first sub-cloning attempt. So you see that even though there are by-products, with good laboratory practice there are so few of them that the majority of recombinants on plate 3 is what we really wanted to get.
However, it can be different when you have more self-ligation (high number of colonies in plates 2 and 3) or when you have un-cut vector in the reaction (plenty of colonies in all 3 plates).

Typical problems Another example of things that can go wrong is that the sticky ends of the vector or fragment can degrade. Restriction enzymes are never 100% pure, so this is the reason why we try to avoid overdigestion. We do a time-course so that we see how the reaction proceeds. If the digest was complete after 20 minutes, and naturally still complete after 40 minutes, then we have no idea when the reaction was complete, anywhere inbetween 0 and 20 minutes incubation. We write in our notes that we have an overdigest. What happens if a vector is overdigested? The sticky ends might be degraded by unspecific nucleases that contaminate the commercial enzyme preps. Degraded ends become blunt ends, because the single stranded end is more susceptible than the double stranded DNA. Blunt ends are no longer incompatible, so they can self-ligate, and then again you will have loads of colonies on plate 2. What's worse, the blunt ends won't ligate with the sticky ends of the desired fragment. So for a variety of reasons, overdigests need to be avoided.
In the example above, the left plasmid was partially digested. When you can actually see a slightly larger linear band that is the singly cut vector, then you have to incubate longer or add more

enzyme until you don't see it any longer. And remember, even if you do not see evidence for a partial digest on the gel, the bands could simply be below the detection limit.

Ligation is chaos: The most important lesson perhaps is that in a sub-cloning strategy, at each step you loose molecules. You loose them when you do test runs on the gels, you loose them when you extract preparative bands from the gel, you loose during precipitations, and during ligations, a large number of molecules recombine in ways we don't want, such as dimerisation nof the correctly cut vector or dimerisation of the correctly isolated fragment. Only a fraction of the molecules recombines in the desired manner, and the only way in which we can test these individually is to amplify clones. Remember, one copy can't be detected, even millions of copies is not enough, we need billions. So we need to transform the ligation mixture into E.coli, get individual colonies that each contain one of the possible recombinants, amplify the individuals by growing up the colonies as liquid cultures, to isolate plasmid DNA and to test individual preps with restriction digests. For this you have to be able to read circular plasmid maps and calculate fragment sizes, and predict the various sizes of the fragments you expect from the desired recombinant, as well as those of the undesired by-products.

The real situation in the lab

Plate 1
few colonies

= evidence for some supercoil present (uncut vector)

Plate 2
more colonies

=evidence for self-ligation

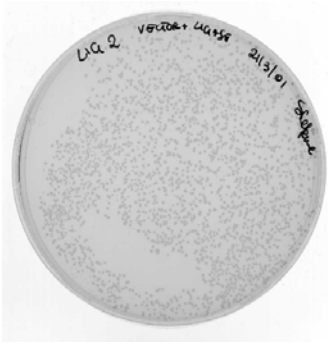
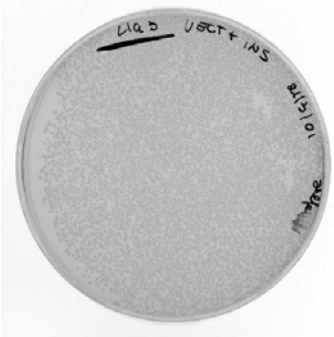


Plate 3
many more colonies

= success!
Most of the colonies come from recombinants

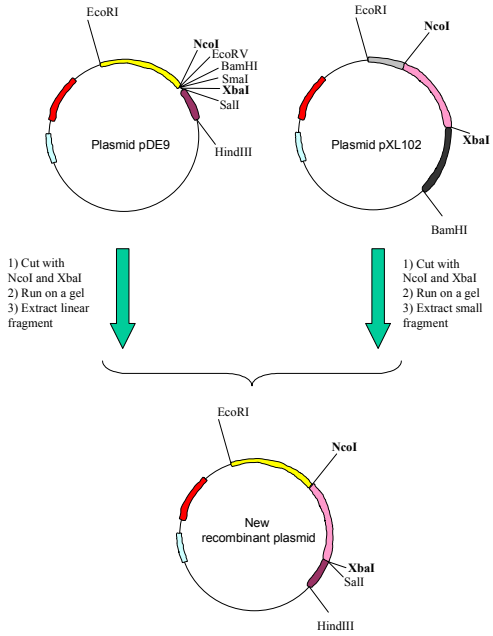


4) Towards genetic engineering

To appreciate the various issues related to subcloning, it is useful to have a look at a few other sub-cloning strategies, each exhibiting specific characteristics that require certain precautions. It is all about the art to understand what can go wrong, and how to control for it, and most importantly, how to minimize the problems in the first place. But we are still talking about the most basic recombinant DNA techniques.

Asymmetric cloning into a polylinker

The first example given was the simplest of all subcloning steps, we call it asymmetric cloning. In principle, the perfectly cut and purified vector should not be able to self-ligate, because the two ends are incompatible. The original fragment to be replaced cannot ligate either, because it has been removed during the gel-purification of the large fragment (in theory at least). The correct fragment is added in excess, should outcompete any remaining original fragment, and above all, it can only ligate in one orientation, the correct orientation. Yet there are a lot of things that can go wrong, as I discussed in the lectures.

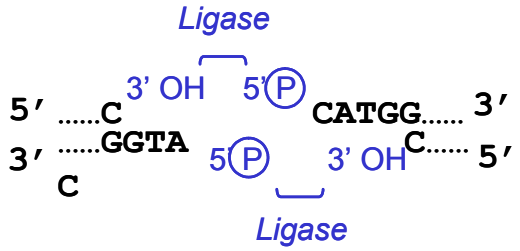


Cloning into a polylinker is similar if you use two different restriction enzymes within the polylinker, but instead of replacing another macroscopic coding region that can be seen as a real fragment on a gel, you just replace the tiny little bit of DNA present inbetween the sites of the polylinker. As you see from the diagram, the restriction sites are the same as in the first example, but this time there is no coding region to start with in the vector. Such a vector is called an expression vector, because it has a polylinker inbetween the promoter region and the 3'end, designed for insertion of all sorts of DNA fragments, and the polylinker is essentially there to give you a choice of restriction sites.
The issues surrounding the fragment from pXL102 are the same as before, it is the new vector we have to worry about. The problem here compared to the first example is that you will have absolutely no idea whether your plasmid was cut with both enzymes or only with one of them. The tiny piece of DNA inbetween the NcoI and the XbaI in the polylinker is so small that you cannot make it out on a gel, nor do you see a difference in the big vector between the linearised band (cut with one of the two enzymes) and the desired vector (slightly smaller band that remains after cutting with the second enzyme). When these two bands were close to each other when we took out a real coding region from pSLH6, this time the bands are of virtually identical size, there is no way you could separate them. So this gives us a much more serious problem because we cannot control for this while preparing the vector. In the previous example we were unsure if we had separated the fragment from the vector, and we were concerned about a tiny amount of supercoil or linearised band (cut with only one of the two enzymes) in our gel purified vector preparation. This time, we don't even know if we managed to cut it with both enzymes in the first place. In addition, we also have to worry more about the supercoil problem, because if some supercoil has the polylinker buried within the structure and is not accessible to the enzyme, it will affect immediately both enzymes simultaneously because they cut so close next to each other in the polylinker.
So this means that you have to be even more careful in designig your strategy. One way to deal with it is to start with NcoI first, check that it has cut (this is possible because you can tell the difference between the desired linearised vector and the complex mixture of supercoil, open circular and multimers you have in a plasmid preparation. Then you add the second enzyme (XbaI) and hope for the best. This second digest is called a "blind" digest, you can't tell if it works because of the reasons explained above. Then you carry out the rest of the procedure, and if you get a lot of colonies on plate 2, which means self ligation, you can blame it on the second enzyme. You then either re-cut the vector with XbaI, which is again a blind approach, or you decide to make the vector again and start with XbaI first, to make sure that it has cut. Then you add NcoI and do it blind, but at least you know from the previous experiment that it should have activity. Generally, you

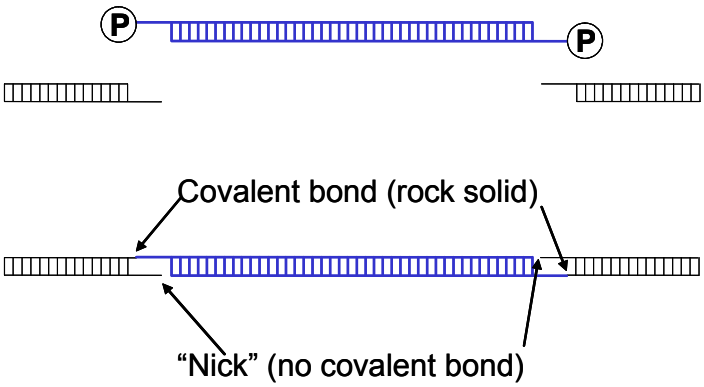
would start with the enzyme that gives you more problems, but you can guess this from common sense, it really is just molecular lego and if you put your mind to it, it makes sense. Of course it also helps to have some experience.

Dephosphorylation, a trick to prevent vector self-ligation

In addition to the blind digest problem, many restriction enzymes having an additional difficult to cut DNA when the restriction site is close to the end of a DNA strand. The enzyme sort of feels unhappy about the lack of DNA on one site of the restriction site and falls off before it cuts the site. This is a very frequent problem, particularly with polylinkers where all restriction sites are very close to each other and you should always worry about that. Together with the supercoil problem and the general issue about the blind second digest, this has led many researchers to prepare the vector to prevent self-ligation. This is not necessary if everything works fine, but because experience tells us that whatever can go wrong will go wrong, this is a prophylactic measure (prophylactic means "in anticipation of problems, just in case", like giving antibiotics to the chicken even if they are not ill). This is done using a dephosphorylation of the 5'ends with the enzyme phosphatase. Ligase cannot create a covalent bond between a 5'OH group and a 3'OH group, it requires the 5'phosphate group. By removing this in a dephosphorylation reaction, neither the tiny fragment nor the vector can ligate any longer, and this reduces self-ligation drastically. This does not prevent ligation of the desired fragment, because it still carries the 5'phosphate groups and can be ligated to the 3'OH groups of the vector. Of course this only happens on one of the strands each time, but this is enough to survive E.coli transformation until replication starts. These days, most molecular biologists will always dephosphorylate the vector, just in case, sort of a precaution, even in the case of the first example. Make a diagram for yourself and try to visualise the principle behind dephosphorylation and check why ligation of a long DNA fragment (30bp or more) will be stable even if only one of the two possible strands on each side are covalently bound. It has to do with the number of hydrogen bonds. Imagine you just align two compatible sticky ends with an overhang of 4 nucleotides, something that you get just before ligas can ligate it, it is only held together by hydrogen bonds, and there are two nicks on each strand. But the stability of the 4 base-pairs is insufficient to keep it together for longer than a few milliseconds at room temperature, most of the time they will be separated. Now if a long fragment is ligated, you also have two nicks, one on either side, and the plasmid is also just held together by hydrogen bonds, but because there are much more basepairs it is stable. If this molecule enters E.coli, the nicks get rapidly repaired by the combined action of kinases (to phosphorylate the 5'OH), and ligase. In fact, in nature DNA gets nicks all the time and they are constantly repaired. That's normal.

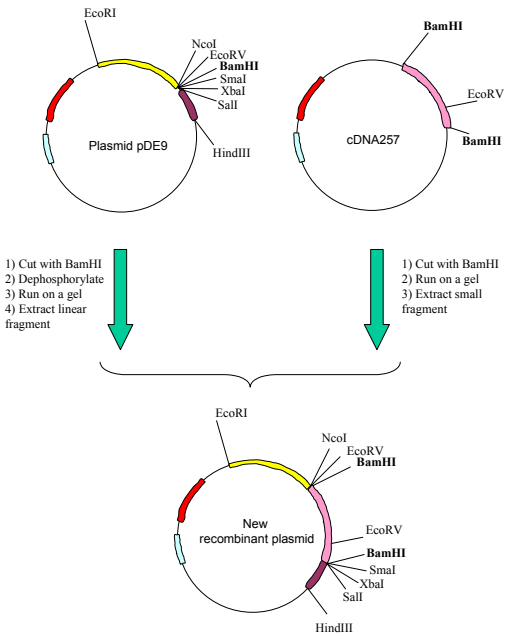


Here is the scheme for a longer fragment so that you can see for yourself:



It should be pointed out at this stage that the technique of dephosphorylation can also be used when you are replacing a real visible fragment but when you are too lazy to purify the vector on gel. Go back to the very first example where a purified DNA fragment is to replace another fragment within a vector. The dephosphorylation strategy is soo easy that lots of scientists including myself cheat and don't bother to purify the vector. If there is linearised singly cut plasmid and supercoiled plasmid close to the desired vector fragment on the gel, it is anyway extremely difficult to separate those without cross-contamination. So we simply dephosphorylate the vector and its original insert when they are still together. If both of these are dephosphorylated, the original insert can't ligate back and produce undesired parent plasmid. If you mix the desired fragment (which still has its phosphates) it will effectively outcompete the original insert and ligate because it still has its 5' phosphates.

Symmetric sub-cloning



The next example shows a much more difficult strategy, easier to explain, but more difficult to do in practice. As you see, this time you cut the vector only with one enzyme. Often, such a strategy is used for generating a genomic library, where you don't have a choice and generating asymmetric fragments is not possible because you don't know the sequence of the individual fragments. But let's first elaborate how it works for a single type of fragment cut from a plasmid.

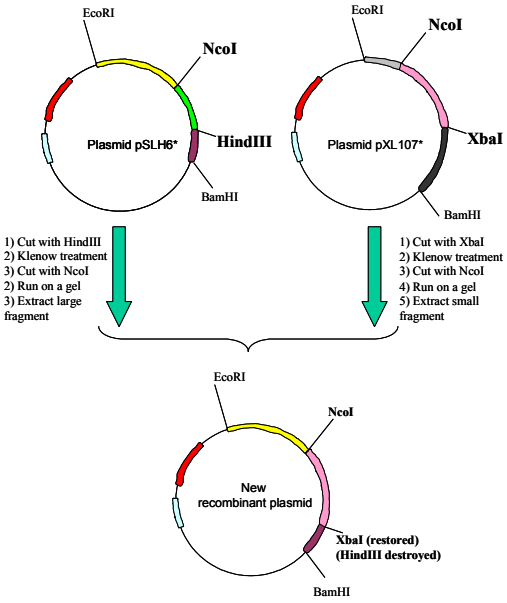
First of all, the same problems remain around the supercoiled vector, if you have some of this, it is a real pain and difficult to solve. The chance to encounter this problem is higher this time, because you cut with only one enzyme. Secondly, the vector will self-ligate by definition, because the ends are compatible. So this time you must dephosphorylate the vector to allow ligation of the fragment to be noticeable. Without reducing the high background of self-ligations, this would be impossible and finding the recombinant (vector with insertion) would be worse than finding a needle in a haystack. The self-ligation reaction is an intra-molecular reaction which has a much higher chance to occur compared to the inter-molecular reaction between the vector and the fragment. And unfortunately, the dephosphorylation reaction is also a blind step, you cannot control for it, you will only know after ligation, transformation and inspection of the plates if you have self-ligation or not (if dephosphorylation worked or not). And of course, like so many things in life, nothing is perfect and dephosphorylation will never be 100% complete.

The next problem is that your fragment can ligate in two different orientations. For the generation of libraries, this is usually not an issue, but if you want a certain construct, you have to design a strategy to check the orientation afterwards. Usually, another restriction site in the fragment (not exactly in the middle) and a unique site on one site of the polylinker in the vector can be used for this. And of course the fragment can multimerise and insert twice. In fact, you can get a lot of by-products, and it is important to realise this and to design strategies to distinguish between the various molecules that can emerge from the ligation. Look at the example given and think about the possibility that you have an insertion of two fragments together. With the sites on this vector, how would you control for this event when you check your minipreps? Thinking about this, you should now be able to see that by-products are also an issue for the two earlier examples of asymmetric cloning. This means that whenever you design a cloning strategy, you must be aware of all the by-products you can get after ligation, and you must be able to distinguish between these molecules by simple digests and pattern analysis on a gel.

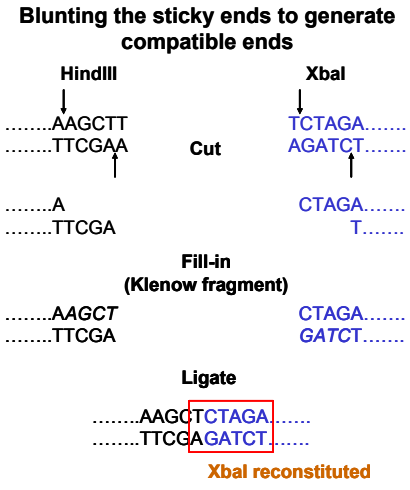
In conclusion, such a symmetric cloning step is a lot more difficult, and you need to get all steps right to make it work. And if you do, it does happen that you get 3 colonies on plate 1, 20 colonies on plate 2, and over 1000 colonies on plate 3.

Further tricks

Blunting ends: More elaborate strategies involve trimming the restriction ends, using PCR technology, site-directed mutagenesis etc.... The first of these is blunting sticky ends with Klenow fragment of DNA polymerase 1. This enzyme has the capability of filling up 5' overhanging single stranded DNA, like in NcoI sites or XbaI sites. For this you need the 4 nucleotides deoxyribonucleotides (dGTP, dATP, dTTP, dCTP). The upper strand acts as primer and Klenow simply elongates the upper strand by 4 nucleotides. Then the DNA is blunt ended. Now why would you do this? It seems a silly idea considering that blunt end ligations are less efficient because the DNA ends spend less time together than the sticky ends (no stabilizing hydrogen bonds). However, if you look at the scheme on the right you can see how it can help. If we want to exchange the coding region, then we can use the sticky NcoI sites because it is unique in both cases and in the right place (the start codon ATG). However, you see that the ends of the coding regions happen to be blessed with different restriction sites in the two cases, and they are incompatible, so they won't ligate.



By blunting those two sites, they become compatible again, and the ligation is still asymmetrical because one side of the fragment is sticky, and the other is blunt. So it can only ligate in one orientation. Ligation of bunted ends usually destroys the site, but in some cases one or even both of them restore after ligation. This is useful to know, because you can test the resulting construct better. For instance, ligating a blunted NcoI with a blunted BamHI site will see both sites restored in the ligation product. Check out yourself why this is (use Google).



The interesting point is that this technique gives you more options. It is also useful to notice that Klenow only fills up 5' overhangs. Some restriction sites cut differently, still generating sticky ends, but 3' overhangs instead. Examples of those are SphI, PstI…. In those cases, Klenow will not be able to synthesize DNA on the single stranded template, because DNA-synthesis only proceeds in the direction from 5' to 3' end. However, Klenow enzyme is capable of removing those 3' overhangs. This hydrolysis reaction (degradation) is much slower than the condensation reaction (biosynthesis), and you usually leave the reaction for longer, but it also works and you can make blunt ends in a similar manner. Also, you should realize that some restriction sites are cut in the centre and give rise to blunt ends to start with. Examples of those are SmaI and StuI. Those can only be ligated with another blunt end, but there is no specificity. So a SmaI site can ligate with a StuI site, but the resulting fusion doesn't cut with either of them. You can also ligate any blunt end with any Klenow-created blunt end.

Partial digests: If you have understood these points and are fully aware of all the things that can go wrong, how about this then: You would like to clone an NcoI-XbaI fragment into a vector cut with NcoI-XbaI, and you are lucky because you have made the vector already on a previous occasion, it worked well with one fragment and you have still some of it left in the freezer. It would be great to use it again, but you find that the second coding region has got an internal NcoI site in the coding region. So if you cut with NcoI and XbaI, you will not get one but two fragments. The chances that both of these ligate correctly in the vector are very slim, because the NcoI-NcoI fragment can ligate in two orientations, whilst the NcoI-XbaI fragment can close the

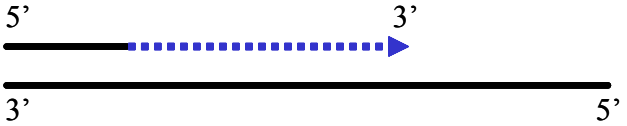
vector on its own. You will predominantly get the truncated shorter fragment and miss the beginning of your codinjg region. In those cases you carry out a partial digest. You first cut completely with XbaI alone, and when the band is linear, you add a 10-fold lower concentration of NcoI and do much shorter time-oints. Partial NcoI means some plasmids are cut twice with NcoI, some not at all, and some inbetween (either one or the other one is cut). So in addition to the linearised plasmid (cut with XbaI alone), you will get three other fragments, the largest of which is the intact coding region with the internal NcoI left untouched. This needs to be gel-purified and hurray, you can ligate it into your existing vector. Alternatively, don't use NcoI and use another vector, for instance one that uses ClaI as a site overlapping with the ATG.

Multiple fragment ligations: The previous example may have given the false impression that it is impossible to ligate two different fragments into a vector in one step. It is indeed a mess if not all of the fragments are asymmetrical. But if you have a vector cut with NcoI and XbaI, ready for ligation, you could easily mix an NcoI-BamHI fragment with a BamHI-XbaI fragment and ligate. Since the two fragments only fit in one way and also have only nice sticky ends, this works remarkably well. Often, researchers opt for a 2 fragment ligation just to avoid having to do a partial digest and work out the various combinations.

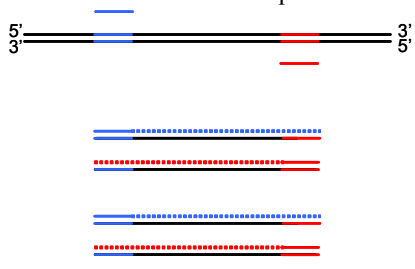
5) The polymerase chain reaction (PCR)

The chain reaction principle

Just like the filling up of 5' overhanging sticky ends using the Klenow enzyme, PCR is based on the synthesis of DNA, but this time using oligonucleotides as primers, single stranded DNA as template, and much more than just 4 nucleotides to fill up. The single stranded nature comes from a denaturation step at 92 degrees Celsius, after which oligonucleotides anneal at lower temperatures (40-65 degrees Celsius) with the single strands and act as primers for a thermostable DNA polymerase. Under theseconditions, the Klenow fragment wouldn't work very well, and it certainly denature at every denaturation step at 92 degrees. So one would have to add new enzyme at every cycle during the annealing step, which would not be very practical. Instead of using the Klenow, PCR makes use of thermostable DNA-polymerases from extreme thermophilic bacteria (I.e. Thermus aquaticus, Pyrococcus furiosus, Pyrodiction occultum….growing at extremely high temperatures, even above 100 degrees Celcius). These enzymes are fine during the annealing, work optimally around 70 for elongation (=DNA synthesis) and ven survives the denaturing steps, so one can place the tubes in a thermo-cycler and go for lunch. Modern thermostable DNA polymerases such as Pfu synthesize 500-1000 nucleotides each minute at approximately 70 degrees Celsius, so a conservative "save" assumption is that one has to count 2 minutes for 1kb. The "sense primer" is identical to the top strand of the template and will bind to the bottom strand after denaturation and re-annealing. This is indicated in the illustration below.



The antisense primer is identical to the bottom strand and will bind to the top strand, in other words everything foes in the opposite direction. If sense and antisense primers are at a reasonable distance, one can get a chain reaction because each newly synthesized strand will be able to serve as new template for the opposite primer. PCR amplifications can work for a few 100 bp up to over 10,000 bp (10kb). During the first reaction, newly synthesized fragments are of variable length, because the template is larger than the fragment that you want to amplify. The growing DNA strand will not stop when it reaches the position of the other primer, but it the polymerase will simply continue until the PCR machines switches from elongation temperature to denaturation temperature. If the elongation time is conservatively chosen, the vast majority of the amplified strands will thus be larger than what is needed, but they will always start with the oligonucleotides that were used as primers. In the second cycle, the original template is still there, and will give rise to the same slightly longer and variable amplification products. These variable products will amplify linearly over time (because the original template concentration does not change), so if you start with 1000 template DNAs, and you do 25 cycles, there will be 25000 of those



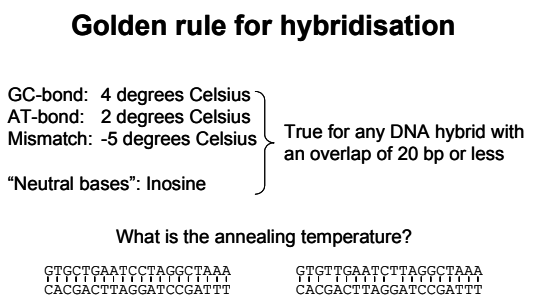
slightly longer fragments in the end. Not enough to see them on the gel. However, after the first cycle, these first round of amplification products will also be present and will be amplified by the opposite primers. In the second cycle, you will thus get correctly sized fragments for the first time, because the templates will stop at the oligonucleotide that was used for the first amplification (because the template stops there) and also start correctly with its own primer. These will now double in numbers at every cycle and give rise to the chain reaction, because they can act as template in each following reaction. So after 25 cycles, the correctly sized fragments will have grown 2^{23} times (because we need to ignore the first two cycles). 2^{23} is approximately 10,000,000, so if you start with the original 1000 copies, so you will get 10,000,000,000 = 10^{10} copies of the correctly sized bands, which is clearly detectable as a sharp band on a gel and immensely more abundant than the by-product. PCR reactions are often done with more than 25 cycles, and often with a very small amount of template, in principle even just one intact copy of template is enough for amplification.

PCR applications

As explained in the lectures, PCR can be used to amplify any fragment of sizes between 100bp and 10,000 bp. In practice fragments are usually between a few hundred to a few thousand bp. By choosing the oligonucleotides you can chose precisely where you want the DNA to begin and to end. You can add 5' trailers to the oligonucleotides to introduce restriction sites to help you subcloning them into a plasmid vector. Using the PCR assembly technique, you can introduce point mutations at any position, or create fusions and deletions at precise positions, without depending on pre-existing restriction sites. Please have a close look at slides nr 107-112 and try to fully understand the potential of the technique and the principles behind oligonucleotide design. It is not possible to imagine how we would do our work without PCR these days, so it is an absolutely transferable skill, from hard core cell biology and molecular biology down to population genetics, ecology, food quality control, detection of diseases, genetic fingerprinting and forensic science. Simply a must for applied Biology these days.

Checklist for PCR:

There are a few points that are worth to know, not really to learn by heart, but it is important that you know where to find this information when you need it, either for a more advanced course, your third year project or in the future when you get confronted with PCR reactions.



1) First of all, you need to understand the principle behind annealing temperatures. For oligonucleotides with sizes of 20 bp, you can apply the golden rule outlined in the diagram to the right. GC-bonds are stronger, so allow higher temperatures. AT bonds are weaker, because you have just 2 instead of 3 hydrogen bonds. A mismatch is very distorting, and so we count a negative temperature. For any given oligonucleotide of 20 bp, you simply calculate the sum of the various temperatures. If an oligonucleotide is longer, then this formula doesn't work any longer, and you have to calculate it for a window of 20bp, then move the window one nucleotide further, calculate again, etc… until you have covered the whole sequence. Then calculate the average. A computer programme does it for you these days. Work it out manually for the two examples given in the diagram, and note how the two mismatches negatively influence the annealing temperature (defined as the temperature at which 50% of the fragments are single stranded and 50% have annealed to te double stranded molecule at any given time-point). This temperature is usually used for annealing conditions. Note that the heat stable polymerase will also work (albeit a bit slower) at this annealing temperature, so the oligonucleotides will immediately get longer. When you raise the temperature to 70 degrees or slightly above (up to 73 degrees) for optimal elongation, the primer has already been extended and binds firmly enough so that it can be considered permanent. Of course ony until the next denaturation step at 90 degrees or above.

2) Second, you must realize that firm binding of the 3' end of the primer is important, because this is where DNA synthesis starts. So if you generate an oligonucleotide with a 5' primer, make sure that there is at least a good 20 bp of perfect match starting from the 3'end.

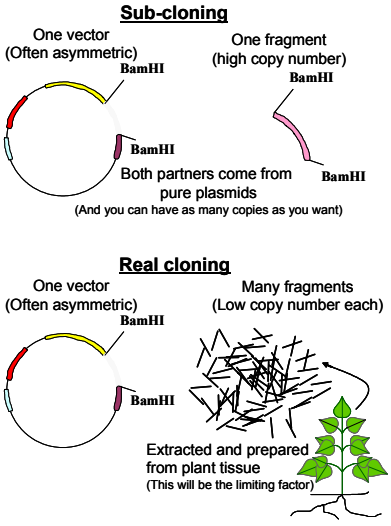
3) Third, avoid palindromic sequences at the very 3'end, to avoid primer-dimer formation. This is when oligonucleotides prime on themselves and give rise to a short undesired amplification product that will amplify better than your desired fragment (because it is so small) and will outcompete everything else. This must be avoided at all cost, once a primer-dimer is present in the lab, it can spoil all future reactions. Best is

not to get it in the first place, so if you are restricted to a certain position of the primer and by using 20bp you ed up with a 3' palindromic sequence, then simply make the oligonucleotide longer to avoid the palindrome at the extreme 3'end.
4) Fourth, oligonucleotides can also be degenerate, tht means that you use a mixture of oligonucleotides instead of a single oligonucleotides. Try to avoid excessive degeneracy, otherwise the PCR yield can be very limited and sometimes unspecific.
5) Fifth, PCR amplifications from plasmids is easy because they are already cloned pieces of DNA present in many copies, and the sequence complexity is limited because you have just a 5kb sequence. The chances the the oligonucleotides prime at the wrong place is very low. However, when cloning a specific fragment from genomic DNA extracted from cells of an organism, then the situation is different because the entire genome has many more combination of sequences. To avoid unspecific priming, we have to avoid that the polymerase works until the proper annealing of the primers takes place. When the ingredients of the reaction are pipetted together, the solutions are usually in tubes on ice. We can add the template, the primer, the reaction buffer, the 4 nucleotides, but we do not add the polymerase yet. This is because at this low temperature, the primers will bind randomly to various unrelated sequences, all they need is a match of 7-8 bases at the 3' end of the primer. If the polymerase were present, it would start extending those primers. The reaction will be slower than the usual 500-1000 nucleotides per minute because the temperature is sub-optimal, but it will synthesize perhaps one nucleotide each second. Nobody can pipet instantly, so we would get lots of unspecific products. The solution is simple, we add everything except the polymerase, put the tubes in the PCR machine, and start the programme. It starts with the denaturing step at 90 degrees or above. The first denaturing step is used to add the enzyme, because during the step the primers are definitely not bound to any template. So we pause the programme, the machine keeps the denaturing temperature, and then we quickly add the enzyme to each of the tubes. Once the enzyme is present in all tubes, the programme is re-started, and the reaction will first cool down to the annealing temperature (for instance 50 degrees Celsius) for a minute, then it goes up to the elongation temperature (70 – 73 degrees Celsius) for a time adequate for the size of the desired amplification product, and then it returns to the standard denaturation step (usually 20-30 seconds). You see that with this procedure (we call it "hot start"), you never have a lower temperature than the proper annealing temperature when the enzyme is present, ensuring the highest possible specificity. This may not be necessary for plasmid amplifications, but we do it routinely for all reactions, it can't do any harm.

6) Generation of libraries

The principle of making a library

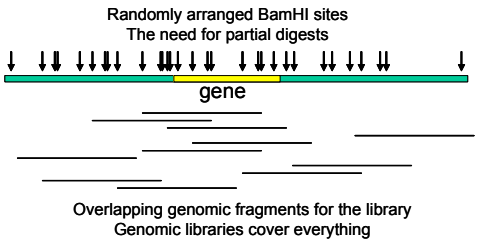
Now we will move on to the generation of libraries to clone new genes that have never been in a plasmid before. As a starting point, we use the last example of symmetric cloning using a BamHI site on the vector. On the left you see what sub-cloning means, it is a trivial exercise, because someone must have done a lot of work before you in order to create the vectors and the plasmids harboring the desired fragment. All you have to do is put it into a different plasmid. But still we have discussed that even simple subcloning of a single type of fragment into a vector can involve a lot of problems. Generating a genomic library means cloning several thousand fragments of DNA (each different) into a vector, representing an entire genome. Needless to say this this is a completely different dimension. First of all, one recombinant is not enough, you need 10 to 100 thousands of recombinants. Secondly, your starting material is not yet present in a vector, you have to get it directly from the organism. So you have no more copies of the genes than the number of cells you start with, and this is a lot less than for one of the typical subcloning reactions described earlier.



If you had to get everything right with the last example (symmetric subcloning), you now have to be absolutely perfect at all steps to make a library, particularly with the vector preparation. Essentially, you need a vector that does not self-ligate and gives no background at all, and ligates to every fragment that comes into contact. So you need a model ligation (get 3 colonies on plate 1, 20 colonies on plate 2, and several 1000

colonies on plate 3). This is not easy and you can't risk having a problem with the vector. You will spend a lot of time generating the collection of fragments (see below), so it is advisable to test your vector, before you ligate the genomic material. A test would involve checking self ligation and presence of uncut vector, and if it ligates to a compatible fragment (cut from any source) without any background in the controls, then you have a good vector.

Preparation of the fragments is also tricky. First of all, you must have enough genomic DNA so that each individual position on the genome is represented sufficiently frequent. With the E.coli transformation efficiency being generally below one event in 100.000, we need at least 100 times more copies than this, so 10^7 copies to start with. This means that each individual fragment will still be invisible on the gel. To think about this, imagine that an average plant cell is a cube of 50 micrometres side length with density 1 gram per cubic centimetre, and the efficiency of extracting genomic DNA is 10%, how many grams of plant tissue do you need to get 10^7 genomes? And soon you will understand that 10^7 will not be enough, so how much do you need when you want 10^9 genomes?



The biggest problem is representing the entire genome. Imagine, we cut the genomic DNA with BamHI, with a statistical cutting frequency of one in 4096, there will be a typical Gauss distribution of fragment sizes, ranging from a few base-pairs to very large fragments, with a maximum occurence frequency around 4096. Imagine you are unlucky, and your gene of interest has several BamHI sites, it will be cut into small pieces and you'll never find a complete clone. Therefore, we generally use a partial digest, starting with more genome copies, and using less enzyme or shorter incubations to cut only some of the sites. Then we select large fragments (10.000 to 20.000 base pairs) using sucrose gradient fractionation. All the small and very large fragments are eliminated, and this is the reason why we have to start with more genomes to compensate for this loss. To get a finer distribution of sizes, we even use a restriction enzyme that recognises a tetramer (GATC) instead of a hexamer (GGATCC). For example, Sau3AI cuts in the same way as BamHI, and thus produces BamHI compatible ends that can be ligated, but it only recognizes the core of the BamHI site (GATC), which occurs at much higher frequency on its own (every one in 256 base pairs). By doing an even more partial digest with Sau3A (to cut only a few % of the various sites) and collecting large fragments of 10.000 to 20.000 base pairs from sucrose gradients, you can be absolutely sure that one of the fragments generated spans the complete gene that you want to clone. Look at the diagram above to understand the principle. Oh, and yes, you still need 10^7 of each of those fragments in the end, so you have to start with a lot more copies of the genome (10^9) because the size selection will only yield 1-2 % of the total of course.

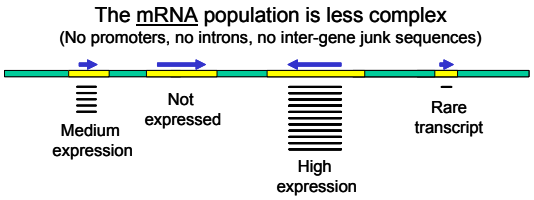
Once you have done all that, the fragments are fine and tested on a gel, the vector has been cut with BamHI, dephosphorylated and tested rigurously, you can start. Because you are generating a library, you need bigger Petri dishes and more of them to get enough colonies separated. It is advisable to transform only a very small portion of the transformation mixture in order to get an idea of the total number of colonies you will get (the titre of the library). Needless to say that you use your very best competent E.coli cells for transformation of a library. When you have done that, you test 10 random colonies, prepare a liquid culture and make minipreps. If most of the preparations show vectors with inserts, and the titre of the library is high, like 5000 or more, deduced from the number of colonies you have, the amount of ligation mix used for the transformation and the amount of it left, then you have a good library. Good for a small genome like that of a bacterium (see further down). Now, the library can be screened using a variety of methods.

The advantages of phage libraries

Before going into screening, let's first discuss a bit more about the types of libraries you can generate and which method is suitable for which occasion. First of all, the library preparation described above is really a special case and just applies to very small genomes or sub-genomic regions (see later). If you want to make a genomic library from a plant, you need a lot more recombinants than just a few thousand. Most higher plants have a genome size of 1000 E.coli equivalents, which means you need a few million recombinants. Due to the low efficiency of E-coli plasmid transformations (one in 100.000), it is impossible to achieve such libraries with plasmids. This is the reason why genomic libraries from plants or animals are usually generated in bacteriophage Lambda. To make a long story short, this phage is designed to introduce DNA into E.coli, it does so very efficiently, a phage attached to E.coli will almost always manage to transform E.coli. Secondly, a completely infective phage can be assembled *in vitro* using the ligated DNA and the phage protein

components in what is termed an *in vitro* packaging reaction (you can buy this nowadays). This is also very efficient, and as much as 10% of the ligated phage DNA can be packaged into an infective capsid.
In other words, phage vectors allow you to represent a far higher proportion of the ligation mix in E.coli at much higher efficiency compared to transformation of naked ligated plasmid DNA. Therefore, it is possible to generate genomic libraries with 10^6-10^7 individuals, even more than that. The screening of such a library is done using the same principles as that of a plasmid library in E.coli, but instead of colonies, you get plaques. To plate out a phage library, you mix an E.coli culture with the packaged phage mix (think of it as a sophisticated ligation mix), mix it with a molten agarose solution in bacterial growth medium, and then pour the mix onto a normal petri dish with solid medium. The agarose solidifies, and then the E.coli (some of which with phage attached) will be immobilised in a thin agarose layer on top of the agar plate. E coli cells will start dividing, but remain more or less where they are because of their size. At the same time infected E.coli cells (only very few in the populations) will lyse and release phage. The latter are much smaller and will diffuse much faster. Diffusion of the phage populations will bring them in contact with non-transfected dividing E.coli. The phages quickly expand into all directions and give rise to a a zone of destruction, whereas the surrounding area becomes more and more filled with bacteria and becomes turbid. After a few hours, E.coli will stop dividing due to the depletion of the medium, and from this moment on phage can no longer infect E.coli (because it requires an active E.coli to complete the infection cycle and lyse the cells). The result is a turbid plate with small spots (plaques) that represent the circular devastated areas of lysed E.coli which originate from one phage attached to one bacterium. The plaques are more transparent and can be easily appreciated when the plate is placed on a light box. Every plaque contains a vast number of phage (several hundred times more than the number of lysed E.coli cells in the plaque, easily 10^7-10^8. Phage will continue to diffuse in all directions even though the plaque gives the wrong impression of a confined zone. To limit diffusion to a minimum, the plates are kept at 4°C. This does not stop diffusion completely, and you must get on with the screening quickly, otherwise the individual plaques are too diffuse to allow picking of a concentrated plaque and to achieve a certain degree of enrichment.

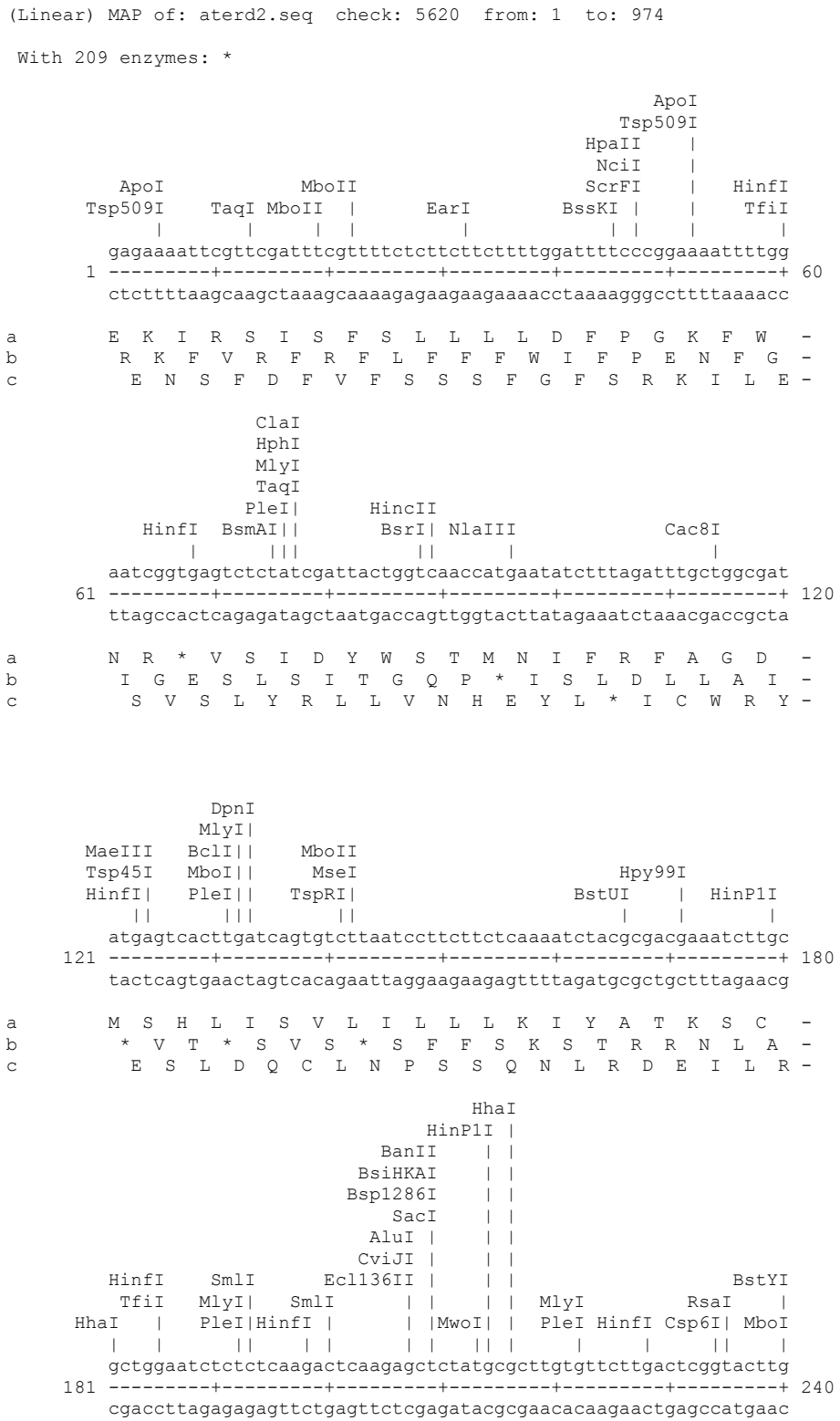
cDNA library or genomic library?
A genomic library can be very complex, particularly when the organism exhibits a high genome size. Many plants have genomes of the same complexity as humans, and this means screening at least a million plaques or more to find a genomic clone. Most of the DNA in a higher organism does not encode for proteins, but represents introns or spacer DNA. But of course, you can be sure that every gene is present because regardless of the cell type, the genome remains constant. The other issue is that you have to use a genomic library if you want to get hold of a promoter sequence. If you are interested in coding regions, or if you think your gene of interest is highly expressed in a certain tissue which would facilitate cloning, then you can also generate a cDNA library. This will represent the entire mRNA population of the cells in the tissue from which you have extracted material. If a certain mRNA represents several % of the entire population, because the corresponding gene has a strong promoter and is highly expressed, then you only have to screen a few thousand colonies or plaques to get hold of it. Of course if the gene is not expressed, it won't be in the library at all. Secondly, analysing a cDNA clone is far less complicated than analysing a genomic clone, because you have to do less sequencing and it is easier to identify the reading frame (region encoding the protein).

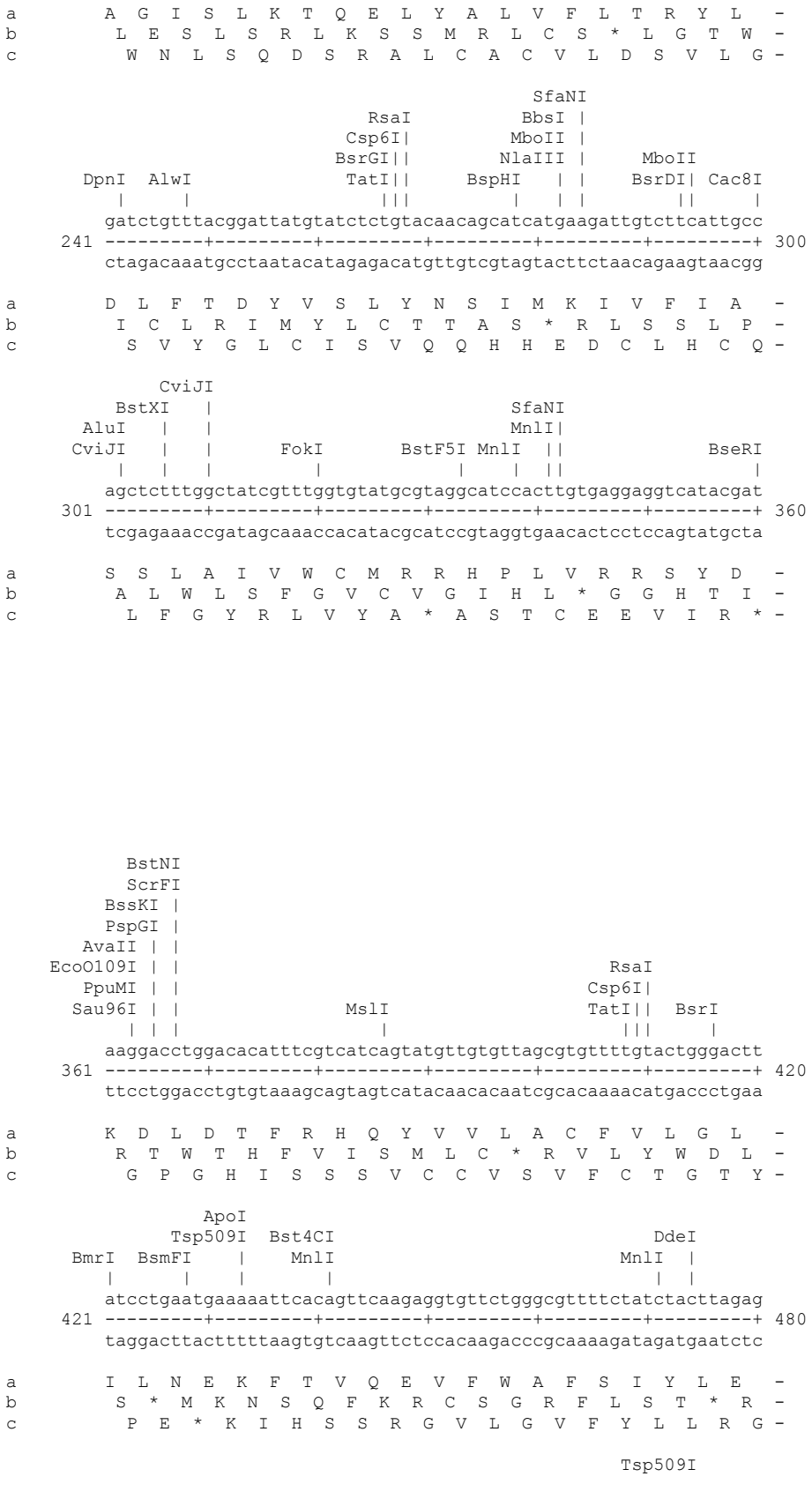


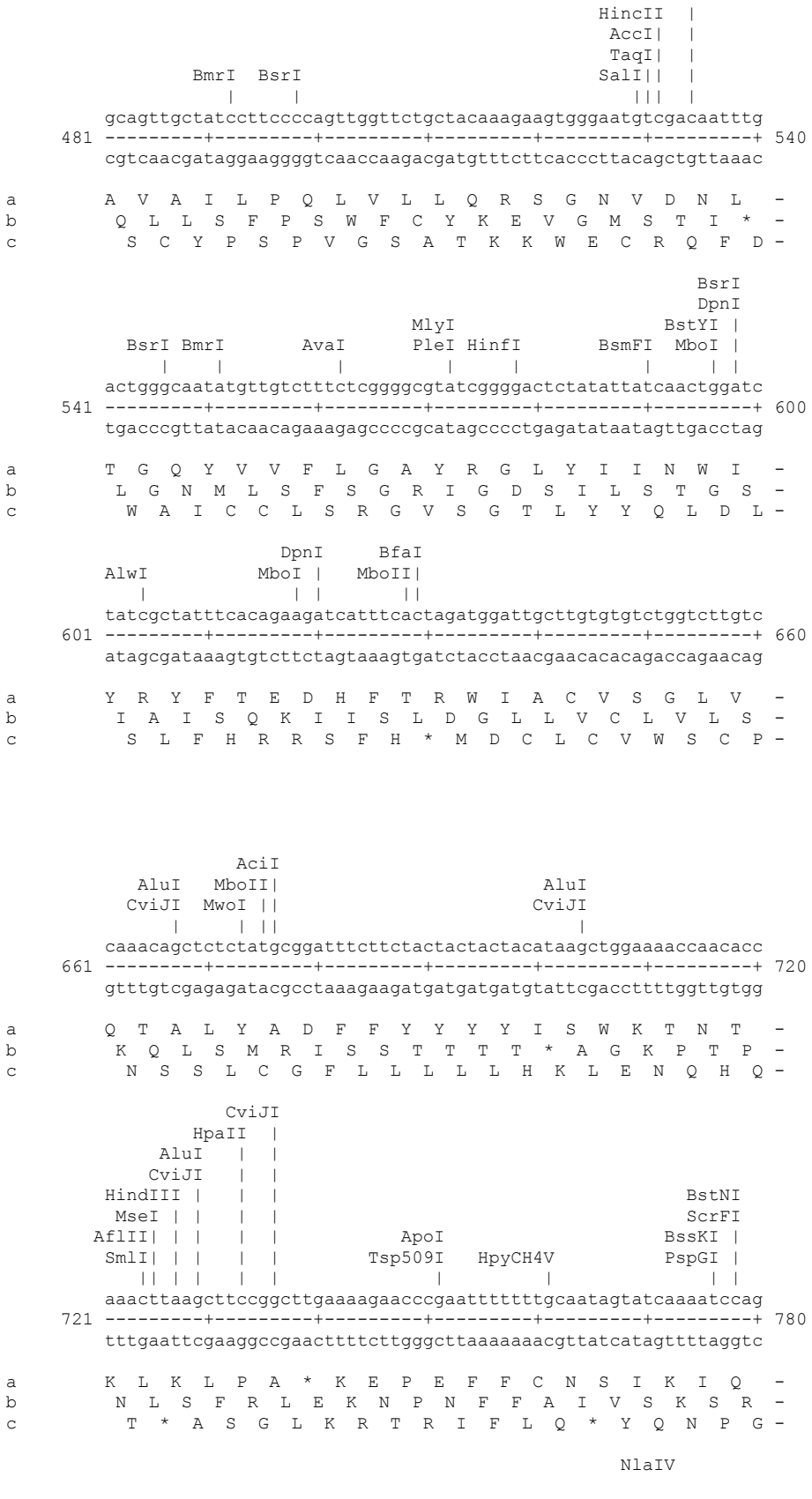
Genes that are not expressed will not be present in the library at all.

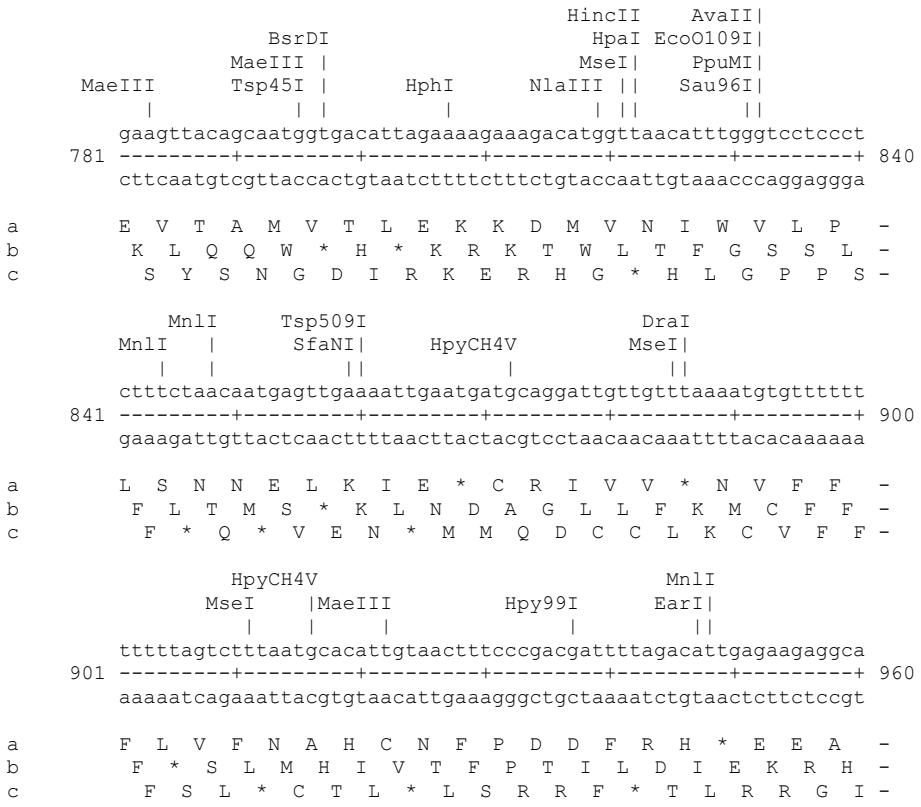
The exercise below hopefully illustrates how easy it is to identify the coding region and the right frame in a typical cDNA clone after sequencing. You simply look for a long region that codes for amino acids only and not stop codons (underlined). Stop-codos have a high occurrence frequence in a random DNA sequence, 3 triplets of a total combination of 64 triplets are stop-codons, so roughly every 60 nucleotides on the DNA you should find a stop-codon by accident. If you have a frame without stopcodons for hundreds of bases, then you have a high chance that this is the coding region, it really doesn't happen by accident. Now you only have to look at the beginning and the end and see if it starts with ATG (coding for methionine) and ends with either (TAA, TAG or TGA). Done! Well, perhaps it is a bit more complicated the first time you do it, but come on, have a go at it, give it your best short.

Here is an example of a computer generated map of a cDNA clone. Go ahead and find the coding region (amino acids are given in single letter code, * means stopcodon).




a A G I S L K T Q E L Y A L V F L T R Y L -
b L E S L S R L K S S M R L C S * L G T W -
c W N L S Q D S R A L C A C V L D S V L G -
SfaNI
RsaI BbsI
Csp6I| MboII
BsrGI|| NlaIII | MboII
DpnI AlwI TatI|| BspHI | | BsrDI| Cac8I
gatctgtttacggattatgtatctctgtacaacagcatcatgaagattgtcttcattgcc
241 ctagacaaatgcctaatacatagagacatgttgtcgtagtacttctaacagaagtaacgg 300
a D L F T D Y V S L Y N S I M K I V F I A -
b I C L R I M Y L C T T A S * R L S S L P -
c S V Y G L C I S V Q Q H H E D C L H C Q -
CviJI
BstXI SfaNI
AluI MnlI|
CviJI FokI BstF5I MnlI BseRI
agctctttggctatcgtttggtgtatgcgtaggcatccacttgtgaggaggtcatacgat
301 tcgagaaaccgatagcaaaccacatacgcatccgtaggtgaacactcctccagtatgcta 360
a S S L A I V W C M R R H P L V R R S Y D -
b A L W L S F G V C V G I H L * G G H T I -
c L F G Y R L V Y A * A S T C E E V I R * -
BstNI
ScrFI
BssKI
PspGI
AvaII
EcoO109I RsaI
PpuMI Csp6I|
Sau96I MslI TatI|| BsrI
aaggacctggacacatttcgtcatcagtatgttgtgttagcgtgttttgtactgggactt
361 ttcctggacctgtgtaaagcagtagtcatacaacacaatcgcacaaaacatgaccctgaa 420
a K D L D T F R H Q Y V V L A C F V L G L -
b R T W T H F V I S M L C * R V L Y W D L -
c G P G H I S S S V C C V S V F C T G T Y -
ApoI
Tsp509I Bst4CI DdeI
BmrI BsmFI MnlI MnlI
atcctgaatgaaaaattcacagttcaagaggtgttctgggcgttttctatctacttagag
421 taggacttacttttaagtgtcaagttctccacaagacccgcaaaagatagatgaatctc 480
a I L N E K F T V Q E V F W A F S I Y L E -
b S * M K N S Q F K R C S G R F L S T * R -
c P E * K I H S S R G V L G V F Y L L R G -
Tsp509I


HincII
AccI
TaqI
BmrI BsrI
SalI
gcagttgctatccttccccagttggttctgctacaaagaagtgggaatgtcgacaatttg
481
cgtcaacgataggaaggggtcaaccaagacgatgtttcttcacccttacagctgttaaac
a A V A I L P Q L V L L Q R S G N V D N L -
b Q L L S F P S W F C Y K E V G M S T I * -
c S C Y P S P V G S A T K K W E C R Q F D -
BsrI
DpnI
MlyI
BstYI
BsrI BmrI AvaI PleI HinfI BsmFI MboI
actgggcaatatgttgtctttctcggggcgtatcggggactctatattatcaactggatc
541
tgacccgttatacaacagaaagagccccgcatagcccctgagatataatagttgacctag
a T G Q Y V V F L G A Y R G L Y I I N W I -
b L G N M L S F S G R I G D S I L S T G S -
c W A I C C L S R G V S G T L Y Y Q L D L -
DpnI
BfaI
AlwI
MboI
MboII
tatcgctatttcacagaagatcatttcactagatggattgcttgtgtgtctggtcttgtc
601
atagcgataaagtgtcttctagtaaagtgatctacctaacgaacacacagaccagaacag
a Y R Y F T E D H F T R W I A C V S G L V -
b I A I S Q K I I S L D G L L V C L V L S -
c S L F H R R S F H * M D C L C V W S C P -
AciI
AluI
MboII
CviJI
MwoI
AluI
CviJI
caaacagctctctatgcggatttcttctactactactacataagctggaaaaccaacacc
661
gtttgtcgagagatacgcctaaagaagatgatgatgatgtattcgacctttttggttgtgg
a Q T A L Y A D F F Y Y Y Y I S W K T N T -
b K Q L S M R I S S T T T T * A G K P T P -
c N S S L C G F L L L L L H K L E N Q H Q -
CviJI
HpaII
AluI
CviJI
HindIII
MseI
AflII
SmlI
BstNI
ScrFI
ApoI
BssKI
Tsp509I
HpyCH4V
PspGI
aaacttaagcttccggcttgaaaagaacccgaatttttttgcaatagtatcaaaatccag
721
tttgaattcgaaggccgaacttttcttgggcttaaaaaaaacgttatcatagttttaggtc
a K L K L P A * K E P E F F C N S I K I Q -
b N L S F R L E K N P N F F A I V S K S R -
c T * A S G L K R T R I F L Q * Y Q N P G -
NlaIV



Enzymes that do cut:

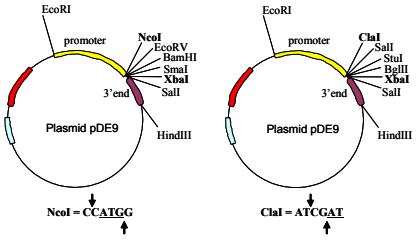
AccI	AciI	AflII	AluI	AlwI	ApoI	AvaI	AvaII
BanII	BbsI	BclI	BfaI	BmrI	BseRI	BsiHKAI	BsmAI
BsmFI	Bsp1286I	BspHI	BsrI	BsrDI	BsrGI	BssKI	Bst4CI
BstF5I	BstNI	BstUI	BstXI	BstYI	Cac8I	ClaI	Csp6I
CviJI	DdeI	DpnI	DraI	EarI	Ecl136II	EcoO109I	FokI
HhaI	HinP1I	HincII	HindIII	HinfI	HpaI	HpaII	HphI
Hpy99I	HpyCH4V	MaeIII	MboI	MboII	MlyI	MnlI	MseI
MslI	MwoI	NciI	NlaIII	NlaIV	PleI	PpuMI	PspGI
RsaI	SacI	SalI	Sau96I	ScrFI	SfaNI	SmlI	TaqI
TatI	TfiI	Tsp45I	Tsp509I	TspRI			

Enzymes that do not cut:

AatII	Acc65I	AclI	AfeI	AflIII	AgeI	AhdI	AlwNI
ApaI	ApaLI	AscI	AseI	AvrII	BaeI	BamHI	BanI
BbeI	BbvI	BbvCI	BcgI	BciVI	BglI	BglII	BlpI
BplI	BpmI	Bpu10I	BsaI	BsaAI	BsaBI	BsaHI	BsaJI

BsaWI	BseMII	BseSI	BsgI	BsiEI	BsiWI	BslI	BsmI
BsmBI	BspEI	BspMI	BsrBI	BsrFI	BssHII	BssSI	BstAPI
BstBI	BstEII	BstZ17I	Bsu36I	BtgI	BtrI	BtsI	DraIII
DrdI	EaeI	EagI	EciI	Eco57I	EcoNI	EcoRI	EcoRV
FauI	Fnu4HI	FseI	FspI	HaeII	HaeIII	HgaI	Hpy188I
KasI	KpnI	MaeII	MfeI	MluI	MscI	MspA1I	NaeI
NarI	NcoI	NdeI	NgoMIV	NheI	NotI	NruI	NsiI
NspI	PacI	PciI	PflMI	PmeI	PmlI	Ppu10I	PshAI
PsiI	PspOMI	PstI	PvuI	PvuII	RsrII	SacII	SanDI
SapI	SbfI	ScaI	SexAI	SfcI	SfiI	SfoI	SgfI
SgrAI	SmaI	SnaBI	SpeI	SphI	SrfI	SspI	StuI
StyI	SwaI	TaiI	TseI	Tth111I	XbaI	XcmI	XhoI
XmaI	XmnI						

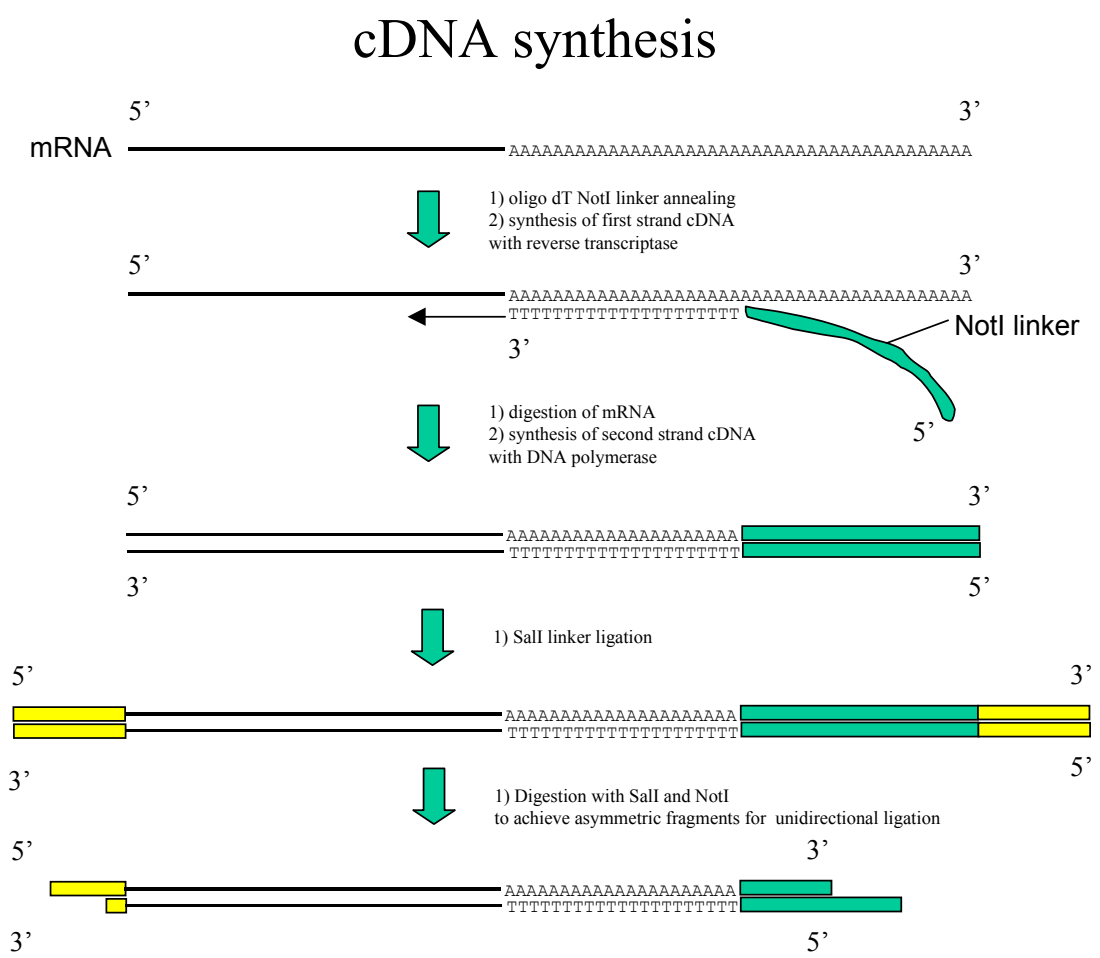
Not really difficult, right? This is the type of map you work with in the laboratory, it gives you the three possible reading frames and a restriction map. You have learned in semester 1 how to play with such sequences on the computer. With these skills, it should be absolutely no problem to find the correct coding region for this gene. Use common sense, and use the internet. If you are a Biology, Biochemistry or Microbiology stydent, you must be able to do this already or learn it now at the latest. Also for joint honors degrees there is no harm in acquiring some basic skills to generate these maps on the computers and get a feeling for basic bioinformatics programs available, you might use PCR amplifications to check for pathogens in food, or to test the origin of a soft cheese (to check the identity of the fungus). Food quality control uses more and more molecular techniques, far beyond the Ames test to check for mutagens….



Another training exercise before we move on again: Imagine you have two possible expression vectors, with different polylinkers. They both contain XbaI as a unique site. But in one case, NcoI is in line with the ATG of the coding region that normally belongs to the promoter on the plasmid, and in the other case it's ClaI. Both sites overlap fully or partially with the ATG, so we like to use them for sub-cloning cloning purposes. Have a look at the two plasmids, have a look at the cDNA sequence above, and when you have identified the coding region, design two primers for PCR amplification with all the understanding you have gathered from the text above and my previous lectures. The aim is to clone the full coding region (that means from ATG to stop-codon) into one of these two vectors. Choose the most suitable expression vector, and then design the PCR amplification and subcloning strategy. I want to see an acceptable strategy, two acceptable primers and an explanation of why you have chosen one of the two plasmids. Take your time, the problem is not trivial.

Back to the libraries now. Imagine you have a new genomic clone that was just sequenced and you are analyzing it, you would find it extremely hard to identify the coding region, because the exons will be separated by several introns (which don't contain coding regions and can often be bigger than the exons). Even if you want to clone a promoter, you might first want to clone the corresponding cDNA, identify the 5' end of the coding region and then use this area to generate a probe to screen the genomic library. This helps to get immediately to the point and prevents you from wasting time identifying tedious intron-exon structures. Having sequenced a cDNA and identifying the 5'untranslated end (the region of the transcript just before the start codon ATG), you can use this as a probe and avoid all these introns.
So how do we synthesize cDNA? The diagram illustrates one example, but there are many more. The main thing is that you use the polyA tail of the transcripts to distinguish and purify them from the far more abundant ribosomal RNA. Then you hybridise an oligo dT25-mer (a small single stranded oligonucleotide containing T only) to the polyA tails, use it as a primer and then synthesize first strand copy DNA (cDNA) using the enzyme reverse transcriptase. That's a polymerase like any other, except that it can use RNA as template and synthesizes DNA. It comes from some viruses that such things for living. So the reverse transciptase synthesizes the so called "first strand" DNA, and then it is easy to synthesize the second strand with a normal DNA polymerase like Klenow, using random hexanucleotides as primers for instance. Then there are various tricks to generate sticky ends on the resulting DNA fragments, one way is illustrated in the diagram. It is these days very popular to ligate cDNA fragments in a unidirectional manner, because it

increases the efficiency of fragment insertion and limits vector self-ligation (remember the advantage of asymmetric subcloning over symmetric subcloning). The other advantage is that you know the orientation of the cDNA and it is easier to identify the reading frame.
What do you have to remember about this? Essential is to realise that cDNA libraries only work if you know that your gene is expressed in the material from which you extract RNA. Secondly, you won't find the promoter region in any cDNA clone, nor any introns. The scheme below is cool to know, particularly if you understand it, and it counts as extra reading. So how does it work? It is based on the restriction enzyme NotI, which recognizes an octameric sequence. The statistical occurrence frequency of such an octamer is 16 times lower than that of a standard hexa-cutting enzyme like the ones w'[ve been discussing earlier. Since this is added to the oligo-dT as a 5' trailer, it will be at the end (=the 3'ends) of the cDNAs. Very few cDNAs will be cut internally by Not1, and so we have a high chance to get unidirectional insertion into the vector.



7) Screening libraries for new clones

Screening with a radioactively labelled DNA probe

Let us imagine you have a portion of a gene cloned, for example a PCR fragment with just a portion of the coding region, and you would like to clone also the rest of the coding region and perhaps even the promoter. This is a typical situation where you would screen a library to find a homologous clone and hope that the cloned fragment of DNA contains more material than present on your first clone.

Transfer to membrane: The screening of libraries, whether they are E.coli colonies on a plate or plaques of phage on a plate, is based on a simple principle, the fact that proteins and DNA have a tendency to stick to nylon or nitrocellulose membranes. The first step is to place a membrane carefully onto the plate, wait a few seconds, and then lift the membrane off, turn it round and place it onto a soft filter paper. The side facing the plate surface must be up (not facing the paper sheet). Usually, you have several plates to handle, so you just leave the membranes to dry after lifting them. Either E.coli cells or phage will have transferred to the membrane and will be present at the same position relative to the original plate. The latter is called the 'master' plate, because it has colonies or plaques left (not everything is transferred) which are still alive (or infective in case of phage) and you have to go back to them once you have screened the membranes (the screening kills everything) to pick your individuals. The master plates are kept in a save place at 4°C until the screening is done, this may take a few days and has to be done as soon as possible.
The difference between plaques and E.coli colonies is that plaques are easier transferred because of a higher number of smaller, more diffusable phage compared to E.coli in a colony. Also, the transfer by diffusion onto the plate is more quantitative and reproducible, which allows you to put one membrane on the plate for a few seconds, lift it off, and then place another filter on the plate for a bit longer, lift it off etc... You can easily do this 4 to 5 times, each time allowing for a bit more time to compensate for the depletion of the plaque (but don't worry, there are enough left on the master plate to pick them afterwards). These replicas can be useful when you want to screen the library with different probes (see differential screening). Replicas are difficult to make with E.coli colonies, because the transfer of a colony to the membrane is far less reproducible and usually much more of it sticks to the first filter. You need to keep the rest on the master plate to make sure you have living individuals by the time you finish the screening of the membranes. So E.coli colonies are usually just screened once, whereas phage plaques can be screened several times and are much easier to handle. The other difference is that diffusion is usually not a big problem with E.coli colonies, because you can see them and they don't expand while waiting in the fridge. In contrast, phage are invisible to the eye, they may have formed a plaque, but by the time you finish the screening, they will have diffused around and all you can pick is an enriched population containing also phage from other plaques nearby. This means you have to dilute the phage, plate out again, and screen again to purify the phage, and this has to be 3-4 times. With E.coli, the first screen usually gives you a clone, at least if you have a bit of experience with it. So each of the methods has advantages and disadvantages.

Denaturation, neutralisation, washing and baking: After lifting membranes, the next step is to extract the DNA out of the E.coli or the phage without lossing the confined nature of the replicated colony or plaque. This is done by placing the membranes (transferred material still facing upwards) onto a series of moist paper filters for a few minutes each. The first one contains an alcaline denaturating solution to brake up the cells or phage. The second contains a neutralisation solution to bring the pH back to almost neutral. The last one is a buffered washing solution to remove the excess salt. Finally, the membranes are placed onto a dry filter paper and are dried, first at room temperature, and then in an oven at 80°C under vacuum for 2 hours . The complete procedure will release the DNA from the cells or phage, which will then stick to the filters almost instantaneously but becomes only permanently attached when the filter is dried and baked. During the denaturing, neutralisation and washing steps, cells, phage, or released DNA can still diffuse, and often gives rise to 'tailing'. This is not a problem, in fact it often does the trick in telling a real signal from a background spot.

Hybridisation: After baking, you can screen the membranes with a radioactively labelled probe. This usually a piece of DNA in which one of the 4 bases is radioactively labelled (see diagram). Due to the complementarity of the probe with the target DNA, it will bind with high affinity to such DNA but not to unrelated DNA. The DNA on the membranes has been denatured by the treatment described above and

represents single stranded DNA sticking to the membrane in a permanent fashion. Hybridisation with the probe involves placing the membranes into a buffer solution together with the probe, and allowing the probe to seek out homologous regions at a relative high temperature (often 65°C). At this temperature, only homologous DNA will be bound, and the stringency of the screening is dependent on the temperature and the specific properties of the buffer. One also has to make sure that the probe doesn't bind to the membrane itself, as our extracted DNA from the colonies did in the first place. For this purpose, the hybridisation buffer contains a number of other ingredients to block all binding sites on the membrane. One of these is an excess of fish-sperm DNA, think of it as random DNA that fills all positions on the filters (of course you don't use this if you want to clone a gene from that fish). Therefore, we usually incubate the membrane with this buffer for one or two hours before adding the probe, this step is called pre-hybridisation. After adding the probe, the membranes are left to incubate for a longer period, to allow binding of the probe to the desired target DNA. Then, the hybridisation buffer (with the probe) is recovered and stored for further use, and the filters are washed with buffer alone. These washing steps remove any probe left that swims around and isn't really attached to the target DNA. Sometimes, the probe also binds to other unrelated DNA, containing accidental regions of similarity. But by increasing the temperature slightly (for eaxmple to 68°C), you can remove this as well, because the probe will bind strongest to the real target DNA. After these various washing steps, you will end up with a membrane which is hardly radioactive, and you can check this with a Geiger counter. When you are satisfied (hardly any clicks, only in some areas a tiny bit), you can then place the membrane in a cassette, cover it with cling film, and then place a sensitive film on it (in a darkroom) and expose. Usually, after a few hours or sometimes days, you can develop the film and you see a few (3-4, sometimes up to 20) dark spots on the combined exposures. So only very few of these thousands or perhaps millions of colonies or plaques contained the gene of interest. Consult the 'Maniatis' for any further details (only if you are interested), but don't get carried away and focus on the principles only. The idea of this part of the course is to have enough knowledge of the techniques so that you can understand the various types of cloning strategies that we will discuss.

Purification: Once you have identified a spot or a few, you can go back to the master plate (Remember? The one you left in the cold room), place the autoradiogram over the plate and pick the corresponding area. This is either a defined E.coli colony or plaque, or sometimes a zone of several colonies/plaques. This requires marking of the membrane during lifting to fix the orientaion, and you also have to remember that the membrane is turned round and thus shows a mirror image. However, the developped film is transparent, and you can look at it either way. You have to work it out how to place it over the master plate to find the right corresponding zone. Think about this, and if you can't come up with a way to do this, consult the Maniatis or send me an e-mail.
Due to the diffusion problem of the plaques, you have to do several purifying steps. To make a long story short, you cut out the corresponding zone of the master plate, place it into a tube with a buffer, the phage will diffuse into the liquid, and then you do serial dilutions and plate out in the same way as you plated out in the original master plate, just that you select an appropriate dilution to have a bigger distance between the plaques. This should be no problem because your first screen should have resulted in a strong enrichment so that you can have far less individuals on a plate. And then you have to do the whole procedure again, and this is where the old hybridisation solution can come in handy. The bigger distance means that this time you have less diffusion from other plaques by the time you pick the plaque, but in general this step has to be done at least 3 times and sometimes more. This (rather tedious) procedure is called purification and you only stop once every single plaque on the plate gives rise to a radioactively marked spot on the membrane. Then you can be sure that you have only the desired type of phage in your sample, and from that moment on you can extract DNA from the phage, cut out the insert, subclone it into a plasmid, sequence it, work with it etc... In case of screening a plasmid library in E.coli, usually you can pick a defined colony and start working with it straight away, extracting plasmid, sequencing etc...

Other screening methods

There are a vast number of other screening methods, some do not even involve a probe of any kind but are based on genetic approaches. The latter will be discussed in year two when we deal with the model organism Arabidopsis and compare it with yeast or bacteria. Besides using radioactive DNA probes, you can also use antibodies as the probe. This can only be done with cDNA libraries, and only if you use an E.coli expression vector. The principle is based on the fact that the genetic code in eukaryotes and prokaryotes is the same, and

that as long as you don't have introns (which E.coli couldn't splice out), you can get translation of eukaryotic coding regions in E.coli if they are placed under the transcriptional control of an E.coli promoter. The diagram above on unidirectional cDNA ligation also illustrates that controlling the right orientation is of great advantage. A cDNA in the reversed orientation with respect to the promoter wouldn't give you any protein and certaily not the one you want.

Expression libraries can be screened either with DNA probes or with antibodies. This is because the E.coli will have the plasmid or the phage, but it will also have produced the corresponding protein. You can either use a membrane lifting protocol that favours DNA extraction and fixing to the membrane, or you can use a method to favour protein binding. The properties of the membrane can also be optimised for either purpose. Finally, screening of expression libraries using antibodies follows the same principles as before. You have the filter in a liquid that contains blocking agents. These are random proteins to block any other area of the membrane to prevent non-specific binding of the antibody (which is also a protein) to the membrane. For example, Sainsbury's Marvel (dry low fat milk powder) is used, but not if you want to clone a gene from milk producing glands... After blocking the membrane (this is the equivalent of the pre-hybridisation step with a DNA probe), you will add the antibody and allow for some time for binding to the right antigen. Spots decorated with those antibodies can then be visualised by allowing the binding of a secondary antibody which recognises the first one and which is linked to a colour or light generating enzyme, to allow detection. At the end of the day, you will have a filter with dark blue spots or a film with dark dots, which you have to position over the master plate, pick the right zone, purify if necessary, until you have a pure clone. All these steps are principally the same as for the screening with a DNA probe.

Screening expression libraries is then useful when you have purified a protein, you have generated antibodies, but you now want to clone the gene. The antibodies will replace the DNA probe, which you can't use because you don't have any DNA yet. In year 2, we will discuss how even this can be circumvented, but for now it is useful to know that a good antiserum can also be used as a probe to identify a gene in a library.

8) Cloning novel genes, a bit more difficult!!

The two screening methods above are the simplest of cases, because you have a specific probe to identify the gene. Simple, because having a DNA probe means that you have already cloned a portion of the gene and you just want to find a bigger clone. Having an antibody is also quite good, because it means you have a pure protein (to generate antibodies). A pure protein provides an enormous amount of information, as we will see later, not just because you can make an antibody. One further example is that you have a clone from one organism (for example yeast) and you now would like to find the homologous gene in plants. Homologous means that it carries out the same function. This can be done using a co-called heterologous probe. This is a radiactive DNA probe generated from the yeast clone, and you hope that it exhibits enough conserved domains of high homology to allow binding of the probe to the corresponding areas in the plant gene.

```
scerd2.pep x aterd2.pep

       1 MNPFRILGDLSHLTSILILIHNIKTTRYIEGISFKTQTLYALVFITRYLD 50
         || ||  ||:||| |:|||:  |  |:   ||| ||| ||||||:|||||
       1 MNIFRFAGDMSHLISVLILLLKIYATKSCAGISLKTQELYALVFLTRYLD 50

      51 LLTFHWVSLYNALMKIFFIVSTAYIVVLLQGSKRTNTIAYNEMLMHDTFK 100
         | |  :|||||.:||| || |.  ||  :.        .|.. |  |||:
      51 LFT.DYVSLYNSIMKIVFIASSLAIVWCMRRHPLVRR.SYDKDL..DTFR 96

     101 IQHLLIGSALMSVFFHHKFTFLELAWSFSVWLESVAILPQLYMLSKGGKT 150
          |:..:   .: .  . |||  |. |.||::||.|||||| :| : |
      97 HQYVVLACFVLGLILNEKFTVQEVFWAFSIYLEAVAILPQLVLLQRSGNV 146

     151 RSLTVHYIFAMGLYRALYIPNWIWRYSTEDKKLDKIAFFAGLLQTLLYSD 200
          .||  |:  :| || ||| |||:|| |||     ||  .||.|| ||.|
     147 DNLTGQYVVFLGAYRGLYIINWIYRYFTEDHFTRWIACVSGLVQTALYAD 196

     201 FFYIYYTKVIRGKGFKLPK 219
         ||| ||         |||
     197 FFYYYYISWKTNTKLKLPA 215
```

The diagram above illustrates a comparison between a yeast protein and the corresponding plant protein. Those areas of homology will also have homology on the DNA level, but this will be less due to the degeneracy of the genetic code. This means that if you have a good antibody to the yeast protein, you can use this tool to screen an expression library and have a higher chance of success. Try to think about the various issues regarding the use of heterologous probes and think about what can go wrong. If you can picture the difficulties, you have already solved half the problem. But still, even these more tricky cloning strategies are still easy, because you have a probe to start with. It may not be as good as a homologous probe, but it is still a probe. But what will you do when you have no probe at all? When you have no specific information about the structure of the gene?
The following subheadings are designed to give you an idea about the various strategies used to clone new plant genes. This is extremely important, because it is the first step in understanding plant functions, and many research papers describing new principles start by introducing a newly cloned gene. It is difficult to understand such articles without some general knowledge about cloning strategies. Many of this will be applicable to other organisms as well, but I will use specific examples from the plant kingdom to illustrate the various techniques.

Strategies based on differential screening

Differential screening was one of the first techniques used to screen libraries without a specific probe, and it is still used these days and has recently enjoyed further popularity due to the recent development of the

microarray technology (although this is still mainly used in model organisms, not so much in crops). Differential screennng is based on the hypothesis that 1) genes are usually expressed during a physiological condition in which they carry out a vital function, and 2) they are often transcriptionally inactive when they are not needed. This is a typical way of thinking for a molecular biologist, genes are induced when they are needed. The so-called house-keeping genes are always active, because they support fundamental processes in all cell types, but other genes are only active in certain specialised tissues (tissue specific expression) or under certain conditions (for example induced by light, low temperature, pathogen stress etc....).

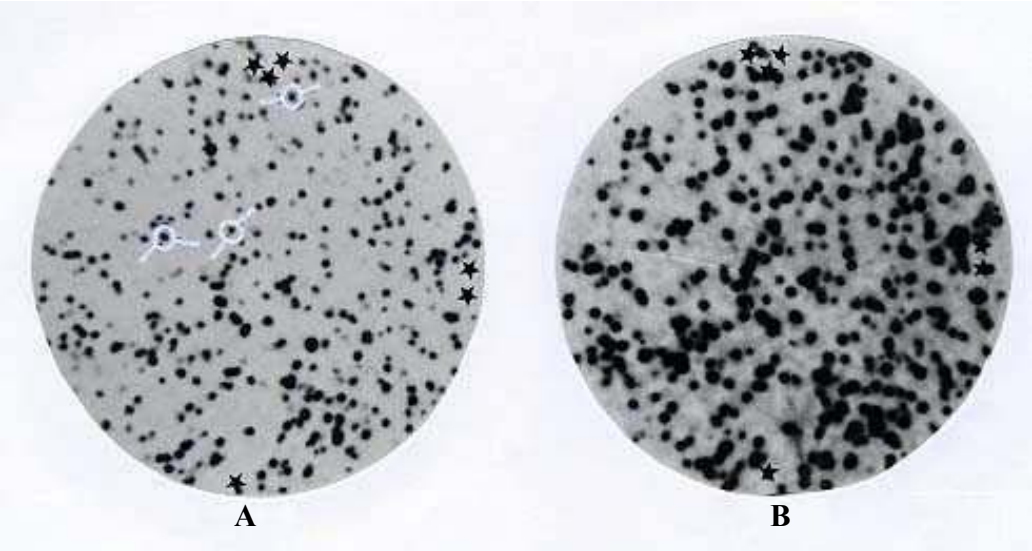
Imagine we want to understand the molecular mechanisms that control the process of cold acclimation. This is a process by which plants of temperate regions adapt during a period of low temperature (one day at 5°C) and become frost-resistant. To illustrate this, if you grow *Arabidopsis thaliana* plants at 20°C and suddenly place them at -5°C. they die from frost damage! However, if you first place them for 12-24 hours at a moderately low, but non-freezing temperature of 5°C, and then at -5°C, they happily survive freezing without any sign of stress. In the 60ties and 70ties, scientists tried to understand this phenomenon by studying lipid compositions, protein profiles, levels of certain ions, all sorts of physiological experiments. In the 80ties, the molecular approach emerged and scientists tried to understand these processes in terms of genes. It was postulated that during the acclimation process at 5°C, those genes involved in the process must have been induced. Secondly, since the plants are not able to survive rapid freezing when grown at 20°C, it was postulated that those genes must be inactive under normal conditions (20°C).

This differential gene expression hypothesis has been the foundation for many successful cloning strategies. mRNAs corresponding to cold-acclimation genes were thought to be present only in plants grown at 5°C and absent in those grown at 20°C. This biochemical difference was exploited to screen a library, either genomic or cDNA. For this purpose, several replica membranes were lifted from the plates with phage plaques. So that's one library, screened with different probes. The library was made with cDNA from mRNA extracted from plants grown at 5°C, because that's when we assume the genes of interest are expressed. Then, to screen the library differentially, one probe was generated from mRNA extracted from plants grown at 20°C, the other probe was generated from mRNA extracted from plants grown at 5°C. A total transcriptome probe is generated simply by using the polyA tail of mRNA as a tool. Since all transcripts have this polyA tail in plants, one can separate mRNA on oligodT cellulose columns, or using magnetic beads covered with oligodT. This procedure is used when you want to generate a cDNA library, and we will now use it to generate a probe. After elution of mRNA (we call it polyA+), an oligodT primer is annealed to the polyA tails, and a first strand cDNA synthesis is done as before, but now one of the four nucleotides (for example dATP) is radioactively marked with ^{32}P. In practice, it is often a mixture of cold and hot dATP to get a compromise between efficient first strand synthesis and incorporation of radioactivity.

You thus generate a complex mixture of different probes, representing the entire mRNA population (transcriptome). mRNAs which are abundant will be over-represented in the probe, whereas transcripts of low abundance (coding for receptors, transcription factors, signal transducing molecules) will be underrepresented. This is not perfect, but in principle every gene should be highlighted by the probe if it is expressed, some strongly, and some weakly. When the library is screened, you will thus get signals for all clones containing a coding region of a gene expressed under the specific condition and represented in the sample from which you extracted the RNA. Some signals will be very strong, others very weak. If you use a genomic library, some plaques will not give any signal at all, because they don't carry coding regions. This illustrates clearly why a cDNA library can be better, because it will only contain clones carrying coding regions. Of course then you have to make a cDNA library from the material in which the gene of interest is expressed. So in case you have missed this, I repeat again that you must make a cDNA library from RNA extracted from plants grown at 5°C.

To carry out a differential screening procedure, we now generate a transcriptome probe for RNA from plants grown at 5°C and also for plants grown at 20°C. Now we have two mixed probes, and we will screen the same library with those two probes. For this purpose, we must lift membranes from the plates, as before, just that this time we need to do it twice, to get a replica membrane (an indentical copy of the first membrane). Now we follow the screening procedure as before, but we do everything twice in parallel, one screening with one probe, and one screening with the other. Needless to say that you have to mark the membranes, so that you can position them over each other, and also over the master plate (remember, the one with the living material kept safely at 4°C ??). After autoradiography, you then compare the signals of the two screenings and look for spots only highlighted by the 5°C probe and not the 20°C probe. This may sound difficult, imagining the comparison of 100.000 stars in the sky which leaps to mind here, but in fact it is easier than you

may think. The illustration below shows that you can easily recognise constellations of signals, and then spot differences. Plate A is probed with the probe that represents the condition of interest (plants grown at 5°C), whereas plate B is probed with the transcriptome of plants grown at 20°C . Try to do this yourself with the example given. You may find more than those specifically enriched in A. In fact, you should also find the opposite, spots which are weaker or disappear in A compared to B.



Once such a differential spot is found, you pick that area of the master plate and start the phage purification process. Again, it is the same as before, but you have to do everything twice again using replicas and the two probes. Everytime you look for 5°C specific spots, until you have a plate in which all spots are seen with 5°C only and never with the 20 °C probe. Yes, it is a lot of work, it takes several weeks to complete such a screening programme, and you wouldn't do it just with one differential spot, you would try to get as many as you can handle, hoping for more than just one gene involved in the process. But it is worth it, because nobody will have cloned those genes before and you will be able to publish several papers (after a bit more work, like sequencing, expression analysis etc...).

So what do you do next? You purify phage DNA, cut out the insert, ligate it into a plasmid vector and generate more material to work with. The insert needs to be sequenced, the coding regions has to be identified, and you have to check if it is complete. If you can't find an ATG codon in frame with the coding region, the cDNA is partial. It is advisable to check the size of the transcript by doing a Northern blot, suing the cloned cDNA as a probe, and compare it with the size of the clone. If the latter is much shorter, you can now screen the library again, this time with a specific probe. Hopefully, you can find one that is full-length, representing the entire coding region. If you want the promoter you can use the cDNA to screen a genomic library. Usually, you would make a probe against the 5'end only. This will increase the chance to find the promoter region and avoids tedious sequencing of intro-exon structures. You can then use the entire array of molecular biology techniques to study gene expression (promoter-reporter fusions, Northern analysis), protein localisation (generation of antibodies, fluorescence or electron microscopy) and prediction of the gene function (database searches, has a gene with similar properties been cloned from another organism? Educated guesses about what it might do during cold acclimation?) followed by functional experiments (overexpression, antisense-inhibition, gene knockouts etc). Functional experiments usually address the following questions: 1) Will prevention of gene induction during cold-acclimation prevent the acquiring of frost resistance (is the gene necessary?). 2) Can constitutive overexpression of a cold acclimation gene mediate constitutive resistance to freezing, even without pre-incubation at 5°C (Is the gene sufficient?)? 3) Is a gene knockout lethal? A research paper will usually contain the original cloning strategy, followed by a few

of these experimental approaches. Such experiments are more interesting than the original cloning work, but can't be covered in these lectures yet. A few examples are listed at the end of the notes on cloning, but you will hear a lot more about these techniques when you choose to focus on molecular biology in level 2. If such research is done in a seed company, then mostly this work will be kept in house, and it might be patented if it works. But you can probably guess that just because a gene is induced at 5°C, there is no guarantee that it plays any role at all during cold-acclimation. It could simply be a false positive, perhaps the mRNA stability is higher at 5°C, giving rise to higher transcript levels, and you think that it is induced, but the promoter hasn't actually changed its activity at all. mRNA levels are determined by synthesis and stability. Also, there are plenty of cool genes that are constitutively expressed, and the regulatory function lies within the protein structure, changing shape and activity in response to environmental stimuli. Not everything that is induced is exciting, and not everything that is exciting has to be induced.

Limitations of differential screening

As outlined in the previous paragraph, transcriptome probes are biased due to different concentrations of individual mRNAs in the population. Not surprisingly, few genes encoding receptors or other regulatory molecules have been cloned this way. Also, it is rather simplistic to assume that every gene involved in cold acclimation has to be induced during incubation at 5°C. Why should a mechanism to detect low temperature in the first place have to be induced by the very process it is meant to detect? Indeed, it should not be induced at all, it should be there all the time to detect whatever it is meant to detect, low temperature in our case here, or light, the presence of pathogens etc... One could even imagine the opposite, because many receptors are known to be down-regulated if the signal that stimulates the pathway is persistent, and often this occurs via a reduction in receptor numbers. So differential screening is not suitable in all cases.
Also, even if a gene is induced during a specific condition, if it is not reasonably well expressed, you won't find it because the signal is simply too weak. Then, you also have all sorts of intermediate situations, in addition to the black and white on-off situation. Some genes are just a bit higher expressed during certain conditions, sometimes you can also get the opposite, and a gene is down-regulated when you apply the condition. However, the latter can be a source of information too, and many differential screening approaches are now designed to find down-regulated genes.

Differential screening and tissue specific gene expression

Tissue specific gene expression means that a gene is only expressed in specialised cell types of the plant, for example the secretory zone of the stigma, or the tapetum layer of the anthers, or the roots only, but not anywhere else. Differential screening is essentially done as before, just that this time the difference is the tissue type, not the physiological condition. This also illustrates that in the former exmple on cold acclimation, you need to be sure the plant material is as homogeneous as possible, with only the temperature being the different factor. One difference is also that studying tissue specificity means screening a library with more than just one probe. A typical experiment would be screening a library with a transcriptome probe from 1) the complete plant, 2) leaves, 3) roots, 4) flowers etc... This is where multiple replica membranes come in handy. You have to be sure your library contains everything, so you have to make the library from the complete plant if it is a cDNA library. A genomic library will have everything by definition, and a lot of extra junk. One can also screen differentially with 1) flower material and 2) the entire plant except the flowers. This will create a better black and white picture and has allowed scientists to easily identify plaques corresponding to a flower specific gene. The example of nuclear male sterility from the lectures on case studies illustrates the value of a tapetum-specific promoter. Think of what you could do if you had a potato-tuber-specific promoter?

Conclusion:
The possibilities are endless and only limited by our own imagination and the inherent problems with this technique associated with low abundant transcripts. But the biggest problem we have currently is that scientists still need to come up with a GM crop that makes a real difference, sweeter strawberries or crunchier tomatoes are not going to be enough to turn the tide. The best current example is the "Golden rice", and even that has not been made available yet. Plenty of work to do for you…..and above all, never loose contact with reality. If you cannot explain your work to a friend in the pub, you need to work on your skills in promoting the public understanding of science.