Video Tutorial 11.1: Locating horizontally acquired gene clusters

A plasmid is a piece of DNA that replicates separately from the chromosome. Plasmids are most commonly found in bacteria and usually circular. Figure 11-4 of your textbook is a schematic map of one plasmid, pSB102. Features of the sequence have been marked and color-coded to illustrate the number and nature of genes the plasmid contains as well as the functional categories of these genes. Some of these functions are essential to the plasmid functions such as replication and maintenance. Genes encoding these core functions are colored yellow and those for conjugative transfer are colored green. Others are accessory functions, such as resistance to inorganic mercury, which is normally toxic to bacterial cells, as seen in this plasmid pSB102.

This video looks at another plasmid, known as R100. We can then make inferences about how this piece of DNA might have come together, based on the locations of various genes and the signatures found in the its sequence.

Plasmid R100 was isolated in the 1950s. The plasmid carries mercury resistance genes like pSB102, but it also confers resistance to the antibiotics chloramphenicol, tetracycline, as well as those in the aminoglycoside class. Multidrug resistance plasmids are commonplace today, thanks to selective pressure from widespread antibiotic use.

How does a single relatively small ring of DNA come to carry so many genes encoding resistance to drugs used in medicine? Let's take a look.

This figure is a schematic map of R100. The plasmid is actually circular but it is presented here in linear form so that we can look at parts of it closely. Each gene – real or hypothesized - is indicated by an arrow. The light blue arrows are represent hypothetical genes for which the function is unknown. As you can see here, there are a large number of hypothetical genes!

The plasmid is composed of double-stranded DNA. Arrows going in one direction represent genes on one strand. Genes on the complementary strand are depicted as arrows going in the opposite direction. You will probably notice a few clusters of genes on the same strand that are so close together that no promoter sequence could lie between them. These clusters are most likely operons containing genes that are expressed as a single messenger RNA transcript.

Certain genes on the plasmid are necessary for the plasmid to replicate and partition itself. Here they are colored yellow. These replication genes are among those considered core to the plasmid and are probably ancestral. Other core genes are those that encode conjugative transfer (here colored green). These are the genes that make the plasmid self-transmissible. They encode the conjugative pilus and other factors necessary to mediate horizontal DNA transfer. In contrast, other genes are non-essential and at least some of these accessory genes were acquired horizontally and integrated into the plasmid more recently than the core genes. Antibiotic resistance genes, here colored magenta and genes encoding resistance to mercury, colored red, are examples of accessory genes.

According to Erwin Chargaff's rule, the number of Gs and Cs in a double stranded molecule will always be equal as will the number of Ts and As. However, the relative proportion of the DNA that is composed of Gs and Cs, compared to the proportion that is composed of As and Ts varies among organisms. Different organisms have different preferred codons and this is one determinant of how G+C (or GC) rich a genome can be. The average GC content is remarkably conserved among species. For *E. coli*, it is about 50% whilst for *Mycobacterium* spp it is above 60%. *Streptococcus* species are relatively AT rich with a GC content under 40%. The differing GC contents are in part determined by codon usage preferences in different organisms. For example, in the GC rich *Mycobacterium tuberculosis* genomes, phenylalanine codons are predominantly UUC, whilst in the much more AT rich genome of *Streptococcus pneumoniae*, they are mostly UUU.

When DNA has been acquired via horizontal transfer recently in evolutionary time, it tends to have the GC content and other molecular signatures of the donor species. Therefore portions of DNA in the *E. coli* R100 plasmid that came from a different species might have different GC contents.

Here again is the R100 sequence presented in linear form. You can see the gene arrows up here, as well as the nucleotide sequence down here.

In the first 100 bases, beginning at nucleotide 1, there are 48 Gs and Cs and 52 As and Ts. Therefore the GC content is 48%. For the 100 bases beginning at nucleotide 2, the GC content is 47% and beginning at nucleotide 3, it is 47%. We can continue to compute the GC content for each 100 nucleotide sequence beginning at nucleotide n, ultimately getting continuous data from the sliding window we are using. We can actually use the data obtained from the sliding window to plot the GC content along the length of the R100 plasmid sequence. Up here, we have a plot of GC content for a sliding window of 100 bases. What you can see is that for many parts of the plasmid the GC content is around 50%. However take a look at this cluster of tetracycline resistance genes and you will see that the GC content plots are much lower. The mercury resistance genes have higher than average GC content.

You'll notice that there are often sets of genes with similar GC content. Very often these sets have a shared function, such as with tetracycline or mercury resistance for example, and we can hypothesize that they were acquired horizontally in chunks. The mercury resistance cluster is flanked by several genes colored royal blue, which we have used to mark transposases and other genes that may be involved in mobility. This might be one way that the horizontally acquired genes became integrated into the ancestor of plasmid R100. Note that many antibiotic resistance genes, colored in magenta, are also flanked by mobility genes.

GC content plots and the locations of transposases and insertion sequences are just two of the analyses that computational biologists can use to determine whether a fragment of DNA on a plasmid is likely to be ancestral or accessory.

Let's now look at a circular map of plasmid R100. It is color coded in the same way as the linear one and you can better see how the antibiotic resistance genes in magenta and mercury resistance genes in red are often flanked by royal blue mobility genes. Inside the gene map, I have now inserted a GC content plot, demonstrating that many of these genes have GC contents significantly higher or lower than the plasmid's core.

For each potentially horizontally acquired cluster, the story is slightly different; it is probable that a variety of gene acquisition and gene loss events built plasmid R100. Quite a few of these events can be inferred by studying patterns in the DNA.

This video was narrated by Iruka N. Okeke and recorded and edited by Charles Woodard and Upma Singh at Haverford College, PA, USA. We are grateful to Kate Heston, Rachel Hoang and Philip Meneely for helpful suggestions. We thank Pathogen Informatics, The Wellcome Trust Sanger Institute for permission to use Artemis for this instructional video.

The video used unannotated (FASTA) and annotated (Genbank) records for plasmid R100 (Genbank accession number AP000342) retrieved from the NCBI Genbank database http://www.ncbi.nlm.nih.gov/ on 11 April 2015. The sequences were viewed in Artemis, a genome browser available from http://www.sanger.ac.uk/resources/software/artemis/

Codon usage information was obtained from the Codon Usage Database at http://www.kazusa.or.jp/codon/