Video Tutorial 12.1: Depicting transcription factor binding sites

When you look at the cells that make up an individual you see there are cells of all different shapes, sizes, structures and functions. Yet they all contain the same genome; their DNA sequence is basically identical. So how are there so many different cells when they have the same genomic DNA?

Well this is in large part due to expression of different sets of genes in different cells. The genome of an individual contains all the genes for that individual, but only a specific subset of all these genes are used, are expressed, in different cells.

And there are many proteins in cells whose role it is to control the expression of other genes, to make it possible to turn sets of genes on or off in different cell types. These proteins are called transcription factors since they control the expression of other genes, by regulating the process of gene transcription.

Different cell types contain different combinations of transcription factors so that different sets of genes are transcribed depending on which transcription factors are present.

Here we have a diagram of a typical eukaryotic gene, and we see that the transcription factors bind to stretches of DNA referred to as regulatory regions, or cis-regulatory modules, CRMs. When the transcription factors bind to DNA in the CRM regions they will interact with other proteins to regulate the DNA polymerase enzyme, shown here in green, that drives transcription, turning gene expression on or off by either activating or repressing transcription.

We often indicate CRMs and the sites where the transcription factors bind to the DNA with color or highlighted regions as in this figure where the blue, green, purple and red regions indicate transcription factor binding sites and the correspondingly blue, green purple and red colored shapes indicate the transcription factor proteins bound to the DNA.

Now lets look at a specific example. This figure shows a zoomed in view of the regulatory region of the *Eat-4* gene from the nematode worm, *C. elegans*. And when we look at this regulatory region in a specific neuron, a type of taste receptor cell called ASE, we find that expression of the *Eat-4* gene in this cell is controlled by transcription factors that include CEH-36, shown here by the blue shapes, and CHE-1, shown here by the purple shape. The specific stretches of DNA that these transcription factor proteins bind to are indicated by the segments of the line shown in corresponding shades of blue and purple. But remember that these colored regions of the line are just representations of double-stranded DNA with a sequence of nucleotide bases of As, Ts, Gs and Cs. If we focus in on the DNA segment where the CHE-1 transcription factor is bound we would find that the purple line indicates a specific stretch of DNA with this sequence: GAAACC.

Since DNA is double-stranded the bottom strand of the sequence is the reverse complement of the top, with G paired with C and A paired with T. This allows us to use shorthand and refer to the entire sequence using just the bases of the top strand - GAAACC. But we can deduce what the bottom strand sequence must be using the rules of DNA base pairing.

By convention, for the shorthand we take the top five prime to three prime sequence of the DNA when the gene is oriented such that transcription is proceeding to make a transcript in that same five prime to three prime direction. This is the DNA sequence to which CHE-1 binds in the cis-regulatory module of the *Eat-4* gene. The CHE-1 transcription factor protein has a specific conformation and shape, enabling it to recognize and bind to this specific sequence, which we can write in shorthand like this.

But CHE-1 doesn't just participate in regulating expression of this *Eat-4* gene; it regulates other genes too. So we can ask when CHE-1 binds to the regulatory regions of other genes what sequence is it recognizing and binding to? Is it always GAAACC or does it vary? If it varies which bases change? All of them or just some?

Here we have some examples of DNA sequences that CHE-1 has been found to bind to. There are 6 bases involved we can see immediately that the first base is always G and the second and third are always A. But the fourth is an A half of the time and a G the rest of the time. The fifth is always a C, but then the sixth is usually C but occasionally G. Analyzing the sequences in this way allows us to summarize and share information about what DNA sequences the CHE-1 transcription factor protein binds to.

We can present this information in a diagram like this, called a sequence logo. This logo is simply a diagrammatic representation of what we just reviewed. We created this logo for our example using just six known sequences but this is usually done with a larger number.

In sequence logos the convention for a position like this last one where the base can vary, is to put the most commonly occurring base on the top like this C and the least frequent variant on the bottom like this G.

Note that these logos give letters different heights depending on how important they are for the binding of the transcription factor. The height of the letters is kind of proportional to the number of times that particular letter is found at that particular position in the known binding sites. But it is not simply proportional. If it were simply proportional the logo would look like this, on the right here – where the total height at each position is the same. But instead the most important bases at specific positions, the ones that don't vary, are given a higher score, a higher total height. If the base at a position varies, the total height is lower and the relative heights of the possible bases found at that position are related to how frequently each one is found.

This type of logo on the left has some advantages. For example you can see right away that these are the positions of the key bases that enable the CHE-1 transcription factor to recognize and bind to this DNA sequence. If one of these essential bases in the sequence varies, then it's a bust. The CHE-1 transcription factor is not going to bind to it. These bases at these positions are crucial for CHE-1 to bind the DNA. And you can see that more easily from these weighted logos, like the one shown here, than from strictly proportional representations.

Sequence logos like these can help us visualize the consensus sequence of the binding site for a particular transcription factor. A consensus sequence is a single sequence that at each position shows the base that is most commonly found at that position in all the known binding sites being analyzed. But the sequence logos convey more information than just a consensus sequence. They show the relative importance of

specific sequences in specific positions and they indicate all the possible variants found in a particular position.

Also don't forget when looking at these logos that we're reading just the top strand of a double-stranded DNA molecule. The transcription factor protein itself is binding to the full double-stranded DNA molecule.

And transcription factors don't usually bind to *every* occurrence of their binding site sequence in the genome. Other factors, such as interactions with other proteins and/or the accessibility of a stretch of DNA can also influence whether or not a transcription factor will bind.

Sequence logos like this have now been compiled to describe the binding sites for many different transcription factors. Here are just a few examples. These sequence logos provide a helpful way to describe the DNA sequences that specific transcription factors recognize and bind. And because different transcription factors recognize and bind to different sequences they bind to different places in the genome and therefore regulate different sets of genes. In this way, it is the particular combination of transcription factors present in each cell that controls which genes in our genomes are being expressed in each of the many different types of cells that make up an individual.

References:

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2015). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016 44: D110-D115.

Schneider TD; Stephens RM (1990). "Sequence Logos: A New Way to Display Consensus Sequences". Nucleic Acids Res. 18 (20): 6097–6100. PMC 332411 Freely accessible. PMID 2172928

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D91-4

Image credits:

Squamous epithelial cells courtesy of Alex_bollo/CC BY-SA 3.0

Nerve cell from Lee, WCA et al (2006). Dynamic Remodeling of Dendritic Arbors n GABAergic Interneurons of Adult Visual Cortex. PLOS Biology Vol4 No2 e29.

Red blood cells courtesy of John Alan Elson/CC BY-SA 3.0

Muscle cells courtesy of Rollroboter/CC BY-SA 3.0

Columnar epithelial cells courtesy of Dr. Glen H. Kageyama, California State Polytechnic, Pomona.