Johnny Alexander Davila-Sandoval

2nd year PhD student

1. Title: **TCF-binding sites in coding sequences of annelid genomes**
2. Biological problem: Presence and number of TCF binding sites seem to be associated with activation of genes by the canonical Wnt pathway
3. Biological background:

Cell-cell communication is pivotal for cell-fate patterning and morphogenesis in the embryo (Gilbert & Barresi, 2018). In spite of an existing vast diversity of animals, a handful of pathways have been conserved across the kingdom, redeployed during different stages of animal development. Among them, the Wnt signaling pathway is one of the most well-studied due to its role in disease, patterning, cell fate determination, etc. Wnts are secreted signaling proteins that act as morphogens, inducing changes in gene expression depending on their concentration (Gilbert & Barresi, 2018).

One of the best characterized Wnt-based signaling pathways is Wnt-β-catenin (‘canonical’). The Wnt signal is transduced via binding of Wnt to a heterodimer of a Frizzled and Wnt co-receptor (LRP5/6) (Steinhart & Angers, 2018). This results in the recruitment of the cytoplasmic protein Dishevelled and Axin-GSK3 via interaction with Dishevelled. GSK3 then phosphorylates LRP5/6, which becomes more affin with Axin and in turn, titrates active destruction complex molecules. This causes the accumulation of β-catenin and its subsequent translocation into the nucleus. There, it acts as a transcriptional co-activator by binding TCF/LEF proteins.

Most invertebrates carry a single ortholog of TCF(T-cell factor)/LEF(Lymphoid enhancer factor) (Archbold et al., 2012). All TCF/LEF proteins contain a highly conserved HMG (high-mobility group) box and a small peptid motif of basic residues (basic tail) (Cadigan & Waterman, 2012). The consensus sequence that binds to the HMG box is 5’-YCTTTGATS-3’ (Atcha et al., 2007).

1. Data required to test the hypothesis:

-Multifasta file containing coding sequences (btained from NCBI)

1. Tools required:

-Biopython, Matplotlib, Pandas (Python)

-Geany or any other text editor

1. Approach:
2. Use of a platform that allows for the management of big data (thousands of sequences).

Genomic information contains thousands of sequences that can result in significant memory use. Biopython allows to parse fasta files (SeqIO) that can contain several thousands of sequences and iterate through the sequences.

1. Code that scans through sequences and look for the TCF-binding site.

The use of a ‘for’ loop allows to iterate over single records in a multifasta file. The code should identify and filter those sequences that contain TCF binding sites, and count the number of sites for further analysis.

1. Generation of a scatter plot for every multifasta file used for analysis.

A simple plot that should represent the size of the coding sequences and the frequency of the TCF binding site provides exploratory results.

1. Analytical results:

*Dinophilus gyrociliatus*:

Chart, scatter chart

Description automatically generated

*Helobdella robusta*

Chart, scatter chart

Description automatically generated

*Capitella teleta*

Chart, scatter chart

Description automatically generated

1. Discussion:

As expected, a majority of coding sequences contain only one TCF binding site, decreasing gradually to a maximum of 9 sites per sequence. For *D. gyrociliatus*, it was shown that four coding sequences contain more than 6 TCF-binding sites, with one coding sequence containing 9 TCF-binding sites. In contrast, *H. robusta* and *C. teleta* report a maximum of 4 and 6 sites, respectively, for 1 and 4 coding sequences.

It is also shown that some long sequences contain a higher number of binding sites, which could be due to chance more than being a potential target. Scanning in adjacent intergenic regions and introns could filter out those sequences that could be off-target. Also, annotating genes for all three species could help identify homolog genes containing binding sites and select better candidates as canonical Wnt pathway targets.

1. Follow-up proposal:
2. Extraction of intronic and intergenic regions from an annotated genome to look for TCF-binding sites.

As preliminary results, I looked for the presence of TCF-binding sites in coding sequences in two annelid genomes. Nonetheless, binding sites are more frequent in intergenic regions and introns. That requires a well-annotated genome, and I could expand it to other invertebrate animals.

1. Filtering of genes whose adjacent intergenic regions and introns contain TCF-binding sites.

Based on the data extracted from the genome, I will use the code generated to iterate over individual sequences and look for TCF binding sites in these regions. Then, I will proceed to filter genes based on this information and propose them as potential targets for the Wnt signaling pathway.

1. Generation of a code to search for any pattern and its occurrence in different genomic regions

The code generated should be able to extract intronic + intergenic regions for any GenBank file, pattern, and generate two scatterplots per species. In this way, I could explore other binding sites and potential targets for different signaling pathways.

References:

Archbold, H. C., Yang, Y. X., Chen, L., & Cadigan, K. M. (2012). How do they do Wnt they do?: Regulation of transcription by the Wnt/β-catenin pathway. *Acta Physiologica*, *204*(1), 74–109.

Atcha, F. A., Syed, A., Wu, B., Hoverter, N. P., Yokoyama, N. N., Ting, J.-H. T., Munguia, J. E., Mangalam, H. J., Marsh, J. L., & Waterman, M. L. (2007). A unique DNA binding domain converts T-cell factors into strong Wnt effectors. *Molecular and Cellular Biology*, *27*(23), 8352–8363.

Cadigan, K. M., & Waterman, M. L. (2012). TCF/LEFs and Wnt signaling in the nucleus. *Cold Spring Harbor Perspectives in Biology*, *4*(11), 1–22.

Gilbert, S. F., & Barresi, M. J. F. (2018). *Developmental biology* (11th editi). Sinauer Associates, Inc.

Steinhart, Z., & Angers, S. (2018). Wnt signaling in development and tissue homeostasis. *Development*, *145*(11), 1–8.