GBSBC

Genetics, Bioinformatics, & Systems Biology Colloquium

Genetics, Bioinformatics, & Systems Biology Colloquium

presents

Carl De Boer, PhD

University of British Columbia

Thursday, February 29th | **12PM** | **Leichtag Auditorium** | **Zoom**

LIVE IN-PERSON with Pizza

*Live stream via Zoom available

Deciphering Genome Regulation: How I Learned to Stop Worrying and Love Random Data

Gene expression is regulated by transcription factors that work together to read cis-regulatory DNA sequences. A primary aim of my group is to decipher the "cis-regulatory code" - the rules that cells use to determine when, where, and how much genes should be expressed. While cis-regulation has proven to be exceedingly complex, recent advances in our ability to query the activity of DNA, combined with Machine Learning have enabled significant progress towards deciphering this code. Here, I will describe a major ongoing effort to learn the cis-regulatory code by synthesizing and testing non-genomic DNA in extremely high throughput. In yeast, we measured 100 million random yeast promoter sequences in a reporter assay and used these data to train models that enabled us to uncover the mechanisms and query the evolution of cis-regulatory sequences. More recently, we demonstrated that random DNA appears to maintain its activity outside of reporter contexts, when tested in a much greater context of hundreds of kb of random DNA. Here, we put human DNA in yeast (which diverged ~1 billion years ago), and found that the yeast cell transcribes the human DNA extensively and in ways that resemble yeast genes, but in ways that are distinct from the regulation that occurs in human cells. Similarly, querying 200 kb sections of randomized DNA using state-of-the-art human gene expression neural network models revealed that these sequences are predicted to have similar levels of regulatory activity to our genomes. These surprising findings suggest a way of cracking the cis-regulatory code: we can train models by profiling the regulatory activities of non-genomic DNA sequences. Since random DNA is essentially in limitless supply, the amount of data we can generate with this approach has already surpassed the complexity of the human genome and will continue to grow. By using these expansive non-genomic data, we will create models that understand genome regulation without ever having seen genomic sequences. Unlike genome-trained models, these genome-free models would no chance of overfitting to genomic sequences and would thus be much more reliable for sequence interpretation and design. I will briefly touch on several ongoing efforts from my group to achieve this goal.

SPONSORED BY:

IGM Institute for Genomic Medicine | UC San Diego Bioinformatics and Systems Biology | CCMI Cancer Cell Map Initiative | UC San Diego School of Medicine

FOR MORE INFORMATION PLEASE VISIT GENOMIC.WEEBLY.COM