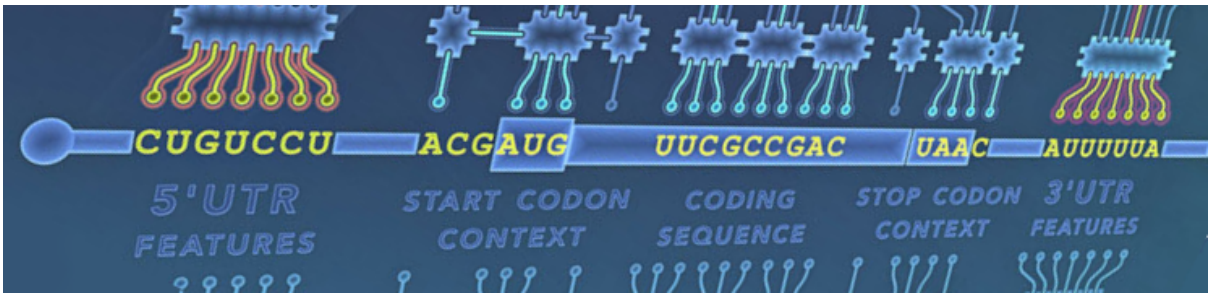

Learning the cis-regulatory code of simple eukaryotes from 300 million years of sequence variation (Master-thesis)

We (Benz lab, Gagneur lab) are looking for a motivated student of biology/molecular biotechnology/bioprocess engineering/biochemistry to assist in developing and carrying out a large, multi-species investigation of regulatory genomics.

The Project:

Deciphering the impact of variation in non-coding regulatory regions of the genome on gene expression is a central challenge in genetics. This goal has not been achieved for any organism but important progress has been made by leveraging genome-wide omics datasets (RNA-Seq, ChIP-Seq, ATAC-Seq and quantitative proteomics, see e.g. [1]) and artificial intelligence, particularly deep learning ([2–4], see [5] for a review). A key challenge, however, is the limited amount of available sequence data with high variation, as most omics experiments are done in the same few model organisms. As such, deep learning models are only aware of a tiny fraction of the vast space of possible gene regulatory sequences.

We propose to move beyond these limitations by collecting genome-wide multi-omics data (RNA-Seq, ATAC-Seq and Mass-spectrometry based proteomics) from diverse fungal species in a variety of growth conditions. Based on these experiments, we will train a deep learning model to predict these omics observations from sequence. By jointly training on many species, our model can leverage both *conservation* and *variation* between species, to better learn the effect of individual regulatory elements on gene expression. We will use existing data, which leverages expression differences between closely related yeast strains and massively parallel reporter assays, to evaluate the model's ability to predict the impact of sequence changes on expression [6,7]. Our ultimate goal is to crack the regulatory code of simple eukaryotes.

Your Tasks:

- Growing cultures of diverse yeasts and filamentous fungi in the Benz lab
- Preparation of samples for NGS (RNA/ATAC-Seq) and Mass-spectrometry assays
- Depending on your interest and prior programming experience, contributing to the bioinformatics processing of the data, as well as the development, implementation and evaluation of the deep learning model (Gagneur lab)

Requirements:

- Currently studying biology, biochemistry or a related subject at the MSc level
- Strong interest for regulatory genomics
- Prior experience with bioinformatics methods and machine learning is helpful, but not required

To apply, send a short motivation letter, your transcript of records, and your envisioned start and end date to: benz@hfm.tum.de (Benz) and jobs-gagneurlab@in.tum.de (Gagneur)

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74.

2. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12: 931–934.

3. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet. 2021;53: 354–366.

4. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18: 1196–1203.

5. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20: 389–403.

6. Renganaath K, Cheung R, Day L, Kosuri S, Kruglyak L, Albert FW. Systematic identification of -regulatory variants that cause gene expression differences in a yeast cross. Elife. 2020;9. doi:10.7554/eLife.62669

7. Shih C-H, Fay J. Cis-regulatory variants affect gene expression dynamics in yeast. Elife. 2021;10. doi:10.7554/eLife.68469