# How does DNA sequence specify gene expression in a timecourse of fungal growth?

**Expert**: Edward Wallace (University of Edinburgh – Cell Biology)
**Session**: Session 2 (6 July – 7 August)

Cells contain thousands of genes, whose "expression" dynamically changes to ensure that they have the right amounts of the proteins that they need to thrive and grow in a dynamic environment. Each gene can be thought of as a region of DNA sequence (modelled as a string of A,C,G,T's), some of which is transcribed into RNA, and then some of this messenger RNA is translated into protein. A fundamental task in molecular biology is to understand how gene sequence specifies the amounts of RNA and protein in any given condition. Statistically, this can be thought of as a high-dimensional model fitting problem where the input is strings and the output is numerical variables or timecourses of gene expression; since it is high-dimensional, dimension reduction methods such as feature selection and clustering are crucial [Eisen 1998].

This project asks what sequence features best explain gene expression patterns in the fungal pathogen Cryptococcus neoformans as it "wakes up", resuming growth after a period of quiescence. Cryptococcus is a major pathogen of immunocompromised humans [May 2016], and the goal of this experiment is to understand the initial stages of infection when the cells "wake up" in a human lung.

The output dataset for this project consists of RNA abundance measurements of thousands of genes over a timecourse, in two growth media at two temperatures, and with two replicates. We will provide this data in integer counts per gene, as well as normalised continuous "transcripts per million" values and log2-fold change "differential expression". The input data consist of DNA sequences for each gene, divided up into distinct regions known to have different functions: for example, the "promoter" sequence mostly regulates RNA production but the "3 prime untranslated region" regulates RNA decay. We will also provide the counts of short k-mers, which are like "words" of DNA sequence, because many factors that regulate gene expression are known to recognize short conserved motifs. Since there are huge differences in output RNA abundances between growth points, we seek statistical models that explain this variation in terms of sequence features such as k-mers or motifs [Bailey 2015, Cheng 2017].

The ideal project outcome would be a statistical model using a small number of sequence features to predict gene expression patterns. For example, one feature might correspond to a transcription factor that promotes expression of a particular group of genes. In the future, the Wallace lab will genetically engineer Cryptococcus cells to test the most promising hypotheses.

This project is an example of generic problem, and common questions include:

- Is there a good low-dimensional description of expression patterns across conditions (e.g. by using Principal Components Analysis)?
- Are there clusters of genes with similar patterns of differential gene expression in the considered conditions (time, growth media, temperature)?
- What kind of pattern of gene expression is each cluster (discuss plots)?
- How to select a small number of significant features (k-mers or motifs) that are associated with the considered design of experiment (e.g. using a Bayesian linear model)?
- How well do k-mers or motifs predict gene expression in different conditions (time, growth media, temperature), or cluster membership, or associated principal components?
- How do we account for redundancy in the feature space? For example, the 5-mer TTTTT is contained in multiple 6-mers, ATTTTT, CTTTTT, ..., TTTTTG, TTTTTT.

Useful courses: Biomedical Data Science, Bayesian Data analysis

References:
[1] Bailey TL., et al., 2015, The MEME Suite. Nucleic Acids Res. http://doi.org/10.1093/nar/gkv416
[2] Cheng J, et al., 2017. Cis-regulatory elements explain most of the mRNA stability variation across genes in yeast. RNA. http://doi.org/10.1261/rna.062224.117

[3] Eisen, MB., et al., 1998, Cluster analysis and display of genome-wide expression patterns, Proc. Nat. Acad. Sci. https://www.pnas.org/content/95/25/14863
[4] May, R., et al. Cryptococcus: from environmental saprophyte to global pathogen. Nat Rev Microbiol 14, 106–117 (2016) https://doi.org/10.1038/nrmicro.2015.6

[5] A.Sánchez & M. Carme Ruíz de Villa (2008) A Tutorial Review of Microarray Data Analysis (Sections 2,5,6) http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/slides/A_Tutorial_Review_of_Microarray_data_Analysis_17-06-08.pdf

[6] Statistical analysis of gene expression microarray data, edited by Terry Speed, CRC Press, 2003