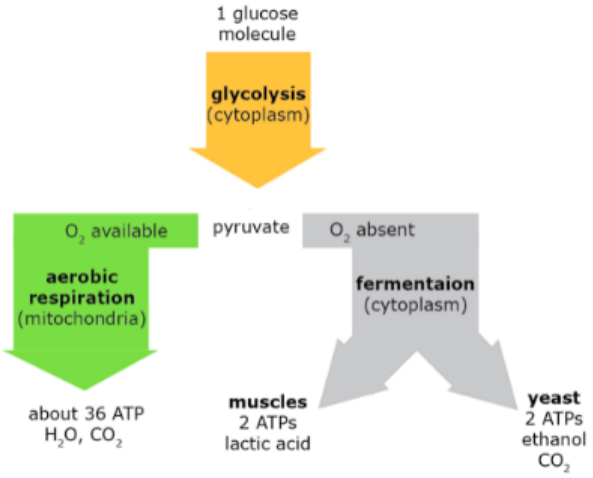
Differential RNA expression analysis

New bioinformatics skills covered: splice-junction aware alignment,
guided transcript assembly, differential expression analysis

In this project we will study changes that happen in yeast cells before or during fermentation. Yeasts, one of the most important eukaryotic model organisms, were domesticated more than 6,000 years ago. They are still widely used in biotechnology. They are facultative anaerobes, which means they can switch their metabolism depending on the environmental conditions. When there is plenty of glucose and oxygen available, the yeast cells can use both to create significant amounts of ATP, the main cellular currency. They do this through *aerobic respiration* in the mitochondria (just like we do). However, when there is an oxygen shortage, they switch to **fermentation**, the process of converting sugars to acids, gases, or alcohol.

In some bacteria (such as *Lactobacilli* that are used in the production of yogurt and cheese), and in our own muscle cells, fermentation converts sugar into lactic acid and 2 ATP molecules. (That's the same lactic acid that makes your muscles sore after working out!) In yeast, however, sugars are converted into ethanol and carbon dioxide. In actuality, the process is making energy in the form of ATP and leaving ethanol and CO_2 as a by-product. We use this carbon dioxide in baking as a leavening agent, which means the CO_2 causes the dough to expand or rise as the gas forms pockets or bubbles.



This week, we will see in-detail the changes in RNA expression during the fermentation process.

1. **Input data:**

We will explore how RNA expression levels change as yeast undergo fermentation to make bread rise. There are two replicates of RNA-seq data from yeast before and during fermentation, and our goal is to find out if the yeast express different genes during fermentation than they do under normal growth.

yeast reads:
SRR941816: fermentation 0 minutes replicate 1
ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941816/SRR941816.fastq.gz (413 Mb)
SRR941817: fermentation 0 minutes replicate 2
ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941817/SRR941817.fastq.gz (455 Mb)
SRR941818: fermentation 30 minutes replicate 1
ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941818/SRR941818.fastq.gz (79.3 Mb)
SRR941819: fermentation 30 minutes replicate 2
ftp.sra.ebi.ac.uk/vol1/fastq/SRR941/SRR941819/SRR941819.fastq.gz (282 Mb)

As a reference genome we will use *Saccharomyces cerevisiae*, in the genome database at NCBI. Make sure you have strain S288c and assembly R64. Download the reference genome in FASTA format and annotation in GFF format.

reference genome file:
ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.fna.gz

annotation file:
ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_R64_genomic.gff.gz

2. Analysis Pipeline

There are plenty of different tools aiming to solve this task. Hereafter you can find a pipeline describing HISAT2 + deseq2, but you can try any aligner and DE count package of your choice. It's worth checking out kallisto + sleuth (kallisto is based on an interesting algorithmic approach of *pseudoalignment*. And it's blazingly fast.)

For deseq2 you can find here two scripts that can be useful for analysis (you also can try to write your own).

a). Aligning with HISAT2
build genome index:
run hisat2-build
hisat2-build <reference.fasta> <genome index>

run hisat2 in single-end mode:
hisat2 -p [number of threads] -x [path to HISAT2 index] -U [FASTQ file] | samtools sort > out.bam
If you ran out of memory in this step, let me know - I'll share with you precomputed BAM

files.

b) Quantifying with featureCounts

featureCounts can not work with GFF files. We need to convert the GFF file to GTF format. For this purpose we will use **gffread**.

Install gffread:
conda install gffread

Convert from GFF to GTF:
gffread <input GFF> -T -o <output GTF>

Run the feature counts program:

featureCounts -g gene_id -a <annotation file> -o <output file>
<input file(s) - SAM/BAM>

We don't need all columns from featureCounts output file for further analysis, so let's simplify it.

Simplify the counts:
cat <output file from featureCounts> | cut -f 1,7-10 > simple_counts.txt

c) Find differentially expressed genes with Deseq2

calculate metrics:
cat simple_counts.txt | R -f deseq2.r

This script generates following files:

1) **result.txt** will contain calculated metrics for our genes
2) **norm-matrix-deseq2.txt** will contain normalised counts that we will use in visualisation

draw heatmap:
cat norm-matrix-deseq2.txt | R -f draw-heatmap.r
Look at the output.pdf file. What do you see?

3. Result Interpretation

In the **result.txt** file genes are sorted by adjusted p-values. So let's take the first 50 genes from this file using linux **head** utility and keep only the first column (gene names) using linux **cut** program:
head -n 50 result.txt | cut -f 1 | cut -d "-" -f 2 > genes.txt

Use gene ontology terms to get a sense of what these genes are doing

Gene ontology (GO) terms are curated keywords that describe the function of individual genes and can help researchers understand what role they may be playing. The Saccharomyces Genome Database, maintained by Stanford, contains all of the GO terms associated with yeast.

Go to http://www.yeastgenome.org/cgi-bin/GO/goSlimMapper.pl
For your top 50 differentially expressed genes:
- in step 1 press “Choose file” and upload **genes.txt**
- in step 2, select “Yeast GO-Slim: Process”
- in step 3, make sure “SELECT ALL TERMS” is highlighted. Press “Search”
- Try to interpret these results

For your lab report:

Intro: Provide biological motivation and some background on differential expression analysis and its purpose.

Methods: Briefly explain the raw data, reference data, tools you used and their parameters.

Results: For each step in the Methods section, list the outcome of that step.

Discussion: Interpret your results and focus on the following:
- How many genes and GO terms changed before and after fermentation in bread dough?
- Pick one process each from your upregulated AND downregulated GO results, and hypothesise how they might be a part of actual changes in yeast metabolism or cellular state during the switch from respiration to fermentation in bread dough.

Good luck!