S288C is a well-studied strain of Saccharomyces Cerevisiae--brewer's yeast. This notebook explores whether the genes in S288C can be aligned with other yeast genomes (three of which are known to produce alcohol).

Conducto injects the following parameters:

- `datasets` : numbered or name genomes (see my_experiment.py)
- `dir` : where to look for information about which genes were found

It is expected that the previous pipeline nodes shelved a dict mapping genome names to data frames, one for each yeast species.

In [1]:
```
datasets = "[0,1,2,3,4,5]"
dir = "/conducto/data/pipeline"
```

In [2]:
```
# Command Line Parameters injected from papermill
dir = "/conducto/data/pipeline"
datasets = "[1, 2, 3]"
```

In [3]:
```
import json
from pathlib import Path
dataset_list = json.loads(datasets)
data_dir = Path(dir)
print("datasets:",dataset_list)
print("in location:", dir)
```

```
datasets: [1, 2, 3]
in location: /conducto/data/pipeline
```

Recover data into memory for analysis...

In [9]:
```
import shelve
import my_experiment

genomes = {}
for dataset in dataset_list:

    # coerce to genome name if not already there
    if not dataset in my_experiment.genome_files:
        dataset = my_experiment.genome_names[int(dataset - 1)]
    shelf_path = str(data_dir / dataset)

    # read it from disk
    with shelve.open(shelf_path) as shelf:
        genomes[dataset] = shelf['frame']

print(genomes.keys())
```

```
dict_keys(['s_cerevisiae', 'b_bruxellensis', 'z_kombuchaensis'])
```

In [12]:
```
sacc = genomes['s_cerevisiae']
brett = genomes['b_bruxellensis']
booch = genomes['z_kombuchaensis']
```

We're searching for S. Cerevisiae genes in S. Cerevisiae. This is our control variable. We expect relatively many genes to have been found

In [23]:
```
len(sacc.protein.unique())
```

Out[23]:
```
6700
```

But what about Brettanomyces Bruxellensis (shows up in sours and other funky beer) and Zygosaccharomyces Kombuchaensis (kombucha yeast)?

In [24]:
```
print({"Brett" : len(brett.protein.unique()),
       "Booch" : len(booch.protein.unique())})
```

```
{'Brett': 98, 'Booch': 253}
```

Ok so far so good. We expected those to be lower. Now let's look for genes we're familliar with. ADH1 / YOL086C is involved in alcohol production. All three of these yeasts are known to produce alcohol, so it's not crazy to expect that we found that gene thrice.

Here's the gene we're looking for:

In [122…
```
sacc[sacc['protein'] == 'YOL086C']['protein_desc'].values[0]
```

Out[122…
```
'YOL086C ADH1 SGDID:S000005446, Chr XV from 160594-159548, Genome Release 64-2-1, revers
e complement, Verified ORF, "Alcohol dehydrogenase; fermentative isozyme active as homo-
or heterotetramers; required for the reduction of acetaldehyde to ethanol, the last step
in the glycolytic pathway; ADH1 has a paralog, ADH5, that arose from the whole genome du
plication"'
```

In [129…
```
from textwrap import indent
def show_align(df, gene_id):
    hits = df[df['protein'] == gene_id]
    for row in hits[['species','locus','hsp']].itertuples():
        print(f"{row.species}:{row.locus}")
        print(indent(str(row.hsp), prefix="    "))
        print()

for strain in [sacc, booch, brett]:
    show_align(strain, "YOL086C")
```

```
Saccharomyces:S288C chromosome XV, complete sequence:160594,159548
    Score 1047 (1934 bits), expectation 0.0e+00, alignment length 1047
    Query:       1 ATGTCTATCCCAGAAACTCAAAAAGGTGTTATCTTCTACGAATCC...TAA 1047
                   |||||||||||||||||||||||||||||||||||||||||||...|||
    Sbjct:  160594 ATGTCTATCCCAGAAACTCAAAAAGGTGTTATCTTCTACGAATCC...TAA 159548

Saccharomyces:S288C chromosome XIII, complete sequence:874337,873291
    Score 696 (1286 bits), expectation 0.0e+00, alignment length 1047
    Query:       1 ATGTCTATCCCAGAAACTCAAAAAGGTGTTATCTTCTACGAATCC...TAA 1047
                   ||||||||  ||||||||||||||||   ||||||||||||||||...|||
    Sbjct:  874337 ATGTCTATTCCAGAAACTCAAAAAGCCATTATCTTCTACGAATCC...TAA 873291

Zygosaccharomyces:strain NRRL YB-4811 NODE_11_length_172022_cov_10.5353_ID_21, whole gen
ome shotgun sequence:118199,119240
    Score 403 (745 bits), expectation 0.0e+00, alignment length 1045
    Query:       6 TATCCCAGAAACTCAAAAAGGTGTTATCTTCTACGAATCCCACGG...TAA 1047
```

```
                  |||||||||||| || |||||| |||||||||||||  | |||||...|||
    Sbjct:  118199 TATCCCAGAAACCCAGAAAGGTATTATCTTCTACGAGCCTCACGG...TAA 119240
```

Hmm, I know that it's not uncommon for the same gene to show up in multiple locations in the genome--so the two Saccharomyces hits aren't that suprising.

And it would appear that Kombucha yeast also relies on ADH1 for alcohol production, but I'm supprised to find that it's not present in Brett. Does this mean that Brett produces alcohol via some alternate metabolic pathway? Or is my search perhaps too narrow?

The protein description on ADH1 mentions a paralog: ADH5 (whose systemic name is YBR145W)

In [130…

```
for strain in [sacc, booch, brett]:
    show_align(strain, "YBR145W")
```

```
Saccharomyces:S288C chromosome II, complete sequence:533762,534817
    Score 1056 (1951 bits), expectation 0.0e+00, alignment length 1056
    Query:       1 ATGCCTTCGCAAGTCATTCCTGAAAAACAAAAGGCTATTGTCTTT...TGA 1056
                   |||||||||||||||||||||||||||||||||||||||||||...|||
    Sbjct:  533762 ATGCCTTCGCAAGTCATTCCTGAAAAACAAAAGGCTATTGTCTTT...TGA 534817
```

I supose I shouldn't be suprised that I didn't find it, paralogs are transcriptions of a gene within the same species. I think what I'm after is an ortholog--which is genes for the same purpose but in different species. Not sure what conclusions to draw here.