1. Run bioinfo pipeline with the 16 bivalves to get gene lists
    a. Pull sequence FASTA from NCBI
    b. Run blast and filter for gamete development
    c. use SPID and uniprot mapping to get GO information
    d. multiple species may hit for each sequence entry, filter for best hit by bitscore
    e. Will be left with list of genes for that species
        i. includes seq id, SPID, gene name, protein name, GO info, eval, bitscore
    f. Repeat for each species

-> Gene list ->

2. Will retrieve fasta file for each gene for each species
    a. `rentrez` r package to do via R: https://cran.r-project.org/web/packages/rentrez/vignettes/rentrez_tutorial.html
    b. Can do as function / loop
    c. Sub directories for each species? -- if so, how will I create gene trees? If not, how will I differentiate different sequences?
        i. Could differentiate vs naming scheme?

-> FASTA file for each gene ->

3. Align sequences to prep for gene tree creation
    a. Using MUSCLE / MSA

-> aligned sequences ->

4. **Create gene trees**
    a. https://rdrr.io/cran/ape/man/write.tree.html
    b. APE:: write.tree()
        i. Can write as function to loop through all aligned fasta files within a subdirectory
        ii. Raxml - IQtree -- loop through aligned sequences to get ML tree

-> .nwk tree files ->

5. Run *treedist::robinsonfould*s() to get RF TD **OR** skip step 4 and do sequence based distance estimation (see R packages below)
    a. RF:
        i. Input: tree (.nwk)
        ii. Output: tree distance matrix with RF score
    ~~*b. Seq. Based:*~~
        ~~*i. Input: aligned sequences*~~
        ~~*ii. Output: sequence distance matrix*~~

-> distance matrix ->

6. Scatterplot of distance matrix

7. Threshold -- top ~20 outliers (RF > 8/9?) become **candidate genes**
    a. Indicates duplication, convergent evolution, trait loss, selection for/against… etc
    b. Determine significant outliers (may be more or less than 20)

** at this point, how would I tie this back to presence/ absence for individual species? It seems this workflow does not lend itself to 'keeping tabs' on the species as we go?

8. Will map these candidate genes onto known/accepted phylogenetic relationships of the 16 species
9. Create matrix of sexual systems and gene presence / absence

Questions

- **Organization of aligned fasta files -- best practice to set up for easy query when come time to make trees and/or distance matrix**
    - **How then, once we get our distance matrix, can we tie individual data points back to individuals ?**
- **Which distance metric is best to use: sequence, FD, pairwise evolutionary distance using ML (built in to phagorn)**

Potential limitations

- Cannot control for sex the genome was taken from

Helpful papers:

https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2023.1102250/full#h3

Remarkably similar study, but uses marker genes and is on a prokaryote

Use case of RF distance

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-07011-0

General info on RF distance

R packages:

Alignment

- MSA (multi sequence alignment)
    - https://academic.oup.com/bioinformatics/article/31/24/3997/197486
- APE:: clustal()
    - https://rdrr.io/cran/ape/man/clustal.html

Tree creation

- Phagorn :: treedist() or rfdist()
    - https://rdrr.io/cran/phangorn/man/treedist.html

Distance Matrix Creation

- Phagorn:: dist.ml() or dist.p()
    - https://cran.r-project.org/web/packages/phangorn/vignettes/Trees.html
    - https://rdrr.io/cran/phangorn/man/dist.p.html
- Decipher:: distancematrix()
    - https://rdrr.io/bioc/DECIPHER/man/DistanceMatrix.html

- *APE::dist.DNA()*
    - *https://rdrr.io/cran/ape/man/dist.dna.html*