**Finding a proper paper with the proper information**

-------------------------------------------------------------------------------------------------------------------------------

**UPDATES (v2):**

\_ An “Initial scanning” section is created to help scan for useful papers more quickly. The section combines sections 1 & 2 in v1 and has a new “Whole-genome sequencing” section.

\_ A completely revamped section for what to do after the initial scanning.

-------------------------------------------------------------------------------------------------------------------------------

When trying to find your next paper to analyze, it’s not always as easy as it seems so here are a few things that I look for in each paper.

1. Initial scanning:
   1. Year: **2010** or newer.
      1. 2008 or 2009 could also be helpful.
   2. Species: ***Saccharomyces cerevisiae*** or ***Escherichia coli***.
   3. **Whole-genome sequencing** (WGS): Look for words like “whole-genome sequencing”, “genome resequencing”, or “sequencing” (you get the idea) in the abstract. If you can’t find it in the abstract, go to “Methods & Material” (or a section of a similar nature) to look for WGS. If you can’t find WGS in either, there’s a high chance that paper isn’t useful.
2. Once you’ve done the initial scanning, check if the proper information is there:
   1. Pro tip: Some helpful keywords:
      1. “sequenc”: This checks for WGS and scans both “sequence…” & “sequencing”.
      2. “evolution”: This looks for the use of adaptive laboratory evolution.
      3. “figure” & “table”: If the publishing site doesn’t have a dedicated section for them, these helps you quickly scan the paper for useful info.
      4. “@”: This looks for the email of corresponding authors.
   2. Go to “Supplementary Material” and check data.
      1. Check for files with a **list of mutations**, which **genes** those mutations occurred in, which **Population** these genes were in, and **frequencies are a bonus**. If they do not have frequencies, then it usually means those genes listed were collected at a certain time point, indicating their presence.
         1. **At least 3 populations**. Must have **replicates** for each.
      2. We only consider experiments on **lab media** (M9, YPD, Davis etc.). Other environments (e.g. mice) sounds exciting but it’s almost impossible to determine (even approximately) how many duplications the bacteria would undergo per day/transfer.
      3. Also, we only consider mutations up until the point a **mutator** appears. Once a mutator appears, the number of mutations increase significantly, adding a layer of complexity which will not be included in this analysis.
      4. Subset your data by the different media (if more than 1 was used).
   3. If the SI doesn’t seem to have any data that works, then look through the paper, sometimes (not usually) they put useful tables in the paper.
      1. Certain papers/hosting sites have a “Figures & Tables” section.
   4. If you’re **missing any data**, i.e, the mutations don’t mention which genes occurred in them. Potential reasons:
      1. It might be in a different file/in the paper somewhere.
      2. It doesn’t exist at all.
      3. It would exist (i.e. the analysis shown would not have been possible without that piece of info), but just wasn’t included.
   5. If all else fails but the data and tables provided look so, so promising, **email the author** in hopes that they reply and provide the missing information.
      1. Find the corresponding author’s email. Usually, it is right under the title where they list the authors (look for an envelope icon), in the “Supplementary Information” section., or in a “Correspondence” section of its own.
3. Now that you have your information, we have to check how accurate it is.
4. I finally read the paper to make sure it’s actually what we want, it does whole genome sequencing, has more than 2 populations, it’s not a dissertation or a thesis paper and its microbial (a.k.a not viral).
5. At this point I look in the paper to see if the following information matches the information provided in the data:
   1. **Number of mutations**
   2. **Number of generations (if any)**
   3. **Number of populations**
   4. **Number of clones (if any)**
   5. **Any other information that is listed in the paper**
6. If the data matches everything the paper says, great, conduct the analysis. If not, try to figure out why it doesn’t add up.
7. Sometimes you simply won’t find an answer and it might be due to a mistake made by the author. Other times it’s a lack of information by the author, as in they might not include the synonymous mutations in their mutation count but won’t say that in their paper.
8. Overall, do as much sleuthing as you can, in the end of the day if you can’t figure out why it doesn’t match, you can either get another set of eyes to see if they can figure it out or email the author with your very specific questions.