**Protocol for pulling stat info from time-series papers**

Thank you all for your help with this project! As you’re all now authors on the paper, please share any ideas that come to you about how to improve our methods or about any interesting trends that you see emerging even if they aren’t written out in this protocol.

**Instructions for Phase 2 Paper Coding:**

1. Below, you’ll find a table with search terms (titled **Search Term Table**)along with the number of papers that contain that term and the things we’d like pulled out of the papers that are relevant to each search term. **Pick a search term** and put your name in the “Reviewer” column.

2. Navigate to the folder “Box/McNew Lab/collaboration/AmNat\_PaperScreening/SearchTermDatafiles” and **open the csv file specific to your search term in Excel**. The csv format cannot handle multiple sheets, so make sure that if you need to take additional notes outside of the spreadsheet, they go in a different document, preferably in the format of {search term}.docx i.e. effectivepopulation.docx.

3**. Create a column in the csv** for each of the relevant things in the “info to pull” section from the table on Page 4. Each of the relevant things are separated by a semicolon. To title the column, use the exact phrasing but excluding any info in parentheses and use ‘\_’ instead of a space. For instance, “measured in time series data (y/n)” becomes “measured\_in\_time\_series\_data”. Consistency in terms makes it really easy to combine these files later. Only use lowercase letters.

4. Since we know these terms can be grepped from the pdfs, you should be able to **open the pdf and cmd+f the term to quickly skim for relevant information**. **You may need to use synonyms and acronyms to capture all relevant instances in the document. If you aren’t familiar with synonyms, type into chatgpt something like: “as used in genomics, provide synonyms, acronyms, and associated symbols of the phrase ‘effective population’”.** If it is not apparent from the text of the file that the additional information was calculated and could be in the supplement, you do not need to search the supplement. However, you might need to if it’s suggested in the paper.

Every csv file has a pdf column that refers to the paper of interest. If the pdf system is annoying and you want DOIs, let me know and I can convert the pdf names to DOIs. All of the pdfs of the articles can be found here, and the name is consistent with that of the file:

Box/McNew Lab/collaboration/AmNat\_PaperScreening/pdfs\_all

**Notes on data to be pulled from each paper**

* Note that all 192 pdfs are listed on each CSV, but you only need to screen the ones relevant to the term. Every search of a paper should start with the measured\_in\_time\_series\_data column and, if the answer is no, you do not need to collect any additional information from that paper.
* If you come across a paper that should be excluded from all analyses (i.e. it doesn’t actually use genomic data in time series), email or text the name of that pdf to Danny.
* Please do not leave any box in your spreadsheet blank: if it is not in your paper or not relevant to your paper, please put “NA”.
* A full description of all available columns can be found above the table of terms. If one isn’t listed as required for your search term but it seems relevant, please add it! If you think of a new thing that needs to be added, either add it in ***bold, underlined, and italicized with your name and the date/time*** or email it to Danny.
* For any “y/n” column, use the full word “yes” or “no” since “n” could easily be a typo of “NA”.
* Always use lowercase in your column titles and coding for anything other than NA

**“Info to pull” explanations**

* first author
  + Last name of first author (so that we can confirm the proper pdf was examined)
* measured in time series data
  + Options: (yes/no)
* level of analysis
  + Options: (single/multiple\_spatial/multiple\_ temporal/multiple\_inferred)
  + Examples:
    - single – used time series to estimate total Ne of study population;
    - multiple\_spatial – used time series to estimate total Ne of multiple study populations;
    - multiple\_temporal – used time series to estimate Ne of historical and contemporary population(s))
    - multiple\_inferred – used time series from multiple samples to infer one or more historical patterns (see figure under “increase or decrease” for an example)
* hypothesized driver of change;
  + Start with most specific descriptor and then provide more and more broad terms, separating terms with a semicolon
  + Example:
    - oil spill; pollution; anthropogenic
* increase or decrease;
  + Options: (increase/decrease/none)
  + We’re often looking at change over time, which is the purpose of the increase\_or\_decrease column. This is binary and statistically significant. The options should be “increase” or “decrease” or “none” or “NA”. None indicates that they tested for change but found none, and NA means they did not look for a change. For some stats, it won’t be completely obvious because there will be inconsistent changes over time. For instance, effective population size could increase and decrease in their models. Here, first put a binary in response to the hypothesized driver of change and then a semicolon followed by a statement describing all the trends. For instance, one of the horse papers modeled effective population size as the following graph. Their hypothesized driver of change is captivity; domestication; anthropogenic and I’ve coded this as “decrease; increase in deep history followed by decrease, increase, decrease”
  + **A graph showing a line graph

    Description automatically generated with medium confidence**
* method of estimating;
  + Example: Wright-Fisher Markov-Chain
* spacetime or time
  + Options: (spacetime/time)
  + Are they directly comparing time series samples within one population (time)? Or are they comparing fst between spatially segregated pops over time (spacetime)
* sliding window vs entire genome
  + Options: (sliding window/entire genome/both)
* raw stats (if present)
  + Example: Past (1888-1909): 23.1-36.9; Present (1996-2015): 75.6-112.6
  + This column should be a bit of a mess, but should still be legible. If you can’t legibly paste it in the raw stats cell, write “see notes document” and put the raw stats in a file title {search term}.docx i.e. effectivepopulation.docx.
  + Please provide all relevant information here, including sample sizes, statistical terms, and p-values if reported. **The** **goal is to pull information that’s comparable among studies, not to extract every number from the paper. Look for commonalities among your paper group, i.e., if you’re looking at Fst or Tajimas D, try and pull relevant Fst and D statistics from the papers. Try looking through a few papers and seeing if there’s any common threads among them in what they report. We’re particularly interested in numbers that support qualitative statements like whether something increased or decreased.**
* candidate gene list (different spreadsheet)
  + Options: (yes/no/NA)
  + If your paper discovered or otherwise dealt with a list of candidate genes, please put those in the CandidateGenes.csv file. The column “new\_or\_confirming” refers to the way that the paper dealt with candidate genes. If they uncovered a list of candidate genes through their time series analyses, code this column as “new”. If they used a previously curated list of candidate genes to inform their analyses, code this column as “confirming”. The “confirmed\_yn” column only refers to genes coded as “confirming” in the previous column; those coded as “new” should be coded as “NA” here. If the paper used a curated list of genes and tested if they responded to a selective pressure, this column is used to state which ones did respond “yes” and which ones didn’t “no”
* Please add a final column called “help\_needed” if you run into issues that require a second opinion. Include details about what needs to be double checked.

**Search Term Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **# of papers** | **Reviewer & status** | **Info to pull** |
| effective population | 78 | Danny / DONE | measured in time series data; level of analysis; hypothesized driver of change; increase or decrease; method of estimating; raw stats (if present) |
| inbreeding | 41 | Ally / DONE | measured in time series data; level of analysis; hypothesized driver of change; increase or decrease; method of estimating; raw stats (if present) |
| population structure  genetic structure  spatial structure  genomic structure | 74  42  8  7 | Henrey / done | measured in time series data; level of analysis; hypothesized driver of change; increase or decrease; method of estimating; raw stats (if present) |
| gene flow | 55 | Isabella / DONE | measured in time series data; level of analysis; hypothesized driver of change; increase or decrease; method of estimating; raw stats (if present) |
| Selection  balancing selection  purifying selection  positive selection  selection coefficient  gene ontology | 171  21  30  66  39  29 | Danny / DONE | measured in time series data; level of analysis; hypothesized driver of change; method of estimating; candidate gene list (different spreadsheet); raw stats (if present) |
| drift | 108 | Sabrina / DONE | measured in time series data; level of analysis; hypothesized driver of change; method of estimating; increase/decrease; raw stats (if present) |
| nucleotide diversity | 39 | Sabrina  (DONE!) | measured in time series data; level of analysis; increase/decrease; hypothesized driver of change; method of estimating; raw stats (if present) |
| fixed  *Note: Only relevant if used in the context of “fixed SNP/haplotype/allele” etc., otherwise NA in measured\_in\_time\_series\_data” column* | 73 | Ally DONE | measured in time series data; level of analysis; method of estimating; hypothesized driver of change; candidate gene list (different spreadsheet) ; raw stats (if present) |
| fst | 50 | Henrey (done) | measured in time series data; level of analysis; hypothesized driver of change; spacetime or time; sliding window vs entire genome; hypothesized driver of change; candidate gene list (different spreadsheet); raw stats (if present) |
| Tajima | 24 | Ally (complete) | measured in time series data; level of analysis; hypothesized driver of change; sliding window vs entire genome; hypothesized driver of change; candidate gene list (different spreadsheet); raw stats (if present) |
| molecular clock | 21 | Ally (done) | measured in time series data; level of analysis; hypothesized driver of change; increase/decrease; method of estimating; raw stats (if present) |
| parallel |  | Sabrina (done) | Evidence (yes/no/partial) snp\_evidence, gene/protein evidence |