**Strain report edits**

For examples of reports:

Raw output file: <https://www.dropbox.com/s/m65nk3qaf6e201p/Merged_strain_results.tsv?dl=0>

Modified report: <https://www.dropbox.com/s/9wpv2qcm0py7pfl/Merged_strain_results.tsv.xlsx?dl=0>

* + Sort the data according to TP2 cluster, then TP1 cluster, then Strain
  + Increase cluster sizes by +1 so that cluster size of 0 → 1
* Solves denominator problem for cluster growth
* Solves problem with non-existent bubbles in bubble graph
  + Moving columns
* Moved V to P
* Moved V-W to S
* Moved X to V
* Moved T to Y
* Moved Y-AE to O

**Data adjustments based on four main scenarios:**

* + Note: cluster sizes are based on +1 adjustment
  + Type I: TP1 > 2, TP2 > 2, TP1 = TP2
  + Type II: TP1 > 2, TP2 > 2, TP2 > TP1
  + Type III: TP1 < 3, TP2 > 2
  + Type IV: TP1 < 3, TP2 < 3

*Type I modifications:* TP1 > 2, TP2 >2, TP1 = TP2

* + None required
  + we can use the ECC stats for TP1 & TP2
  + we can use the cluster averages for TP1 & TP2

*Type II modifications:* TP1 > 2, TP2 > 2, TP2 > TP1

* + Main problem is that the novel strains in TP2 don’t have TP1 data
  + TP1, no modification required
  + TP2: no change for strains also in TP1; for novel strains in TP2: in TP1, needs to have the cluster size and ECC stats from the TP1 strains that they cluster with in TP2, need to have the TP1 cluster number

*Type III modifications:* TP1 < 3, TP2 > 2

* + Main problem is that TP1 cluster doesn’t have ECC stats, impacts the change vector calculation; also, if TP1 = 0 then cluster size for bubble plot & the cluster growth have no data (no bubble for TP1 & “Inf” growth rate)
  + TP1 needs to have a size of 1 (+ 1 adjustment for every cluster) so that there is a bubble of size 1 and so that the denominator is not 0 for cluster growth
  + TP1 needs to have ECC of (1,1) so that we have reference for change vector

*Type IV modifications:* TP1 < 3, TP2 < 3

* + Main problem is that TP1 and TP2 are both small and do not have ECC stats since they are singletons or non-existent
  + Force TP1 and TP2 ECC stats to blanks
  + Filter these strains prior to analysis & give a ECC blanks
  + Eventually, include in analysis but maybe do not include them in EpiMatrix calculation

**Strain filtering criteria:** this would allow us to simply upload a file with the cleaned metadata and to be able to perform the rest of the analysis by filtering for strains of interest & perhaps sub-sampling them to make sure that we can handle the data size

* + Has defined lineage information
  + Has province-level data
  + Has defined date information
  + Is within the temporal window (TP0 to TP2)
  + Is in non-singleton cluster
  + Is in geographical area of interest (e.g. “country” or “region” to include)
  + Perhaps subsample based on the maximum number of strains that we can handle and the total size of the strains of interest.

**Cluster Report Edits (TBD)**