Pipeline edits for split input

* Make some changes to config.yaml to specify number subsets to partition msmc\_clean.fof into

**Cause of flat lines???**

* When inspecting the msmc input files named msmc\_input\_\*.input files (derived from scaffold sequence files according to a search on ncbi db) which msmc\_clean.fof I notice that many of them are empty or have very few recorded variants
* Does the lack of called variants in whole contigs cause the MSMC to produce bad output? Probably.

QC

* Lots of mammal runs are producing “elbow-shaped” MSMC curves
  + Consider
    - How many multi-allelic sites get called in their data
      * Multi-allelic sites get removed from VCF by the bamCaller.py script before sending a VCF to tools like shapeit which compare runs to a reference panel (sounds like we do something like this)
        + Might want to compare how many multi-allelic sites are in used vs unused data
      * If there are many multi-allelic sites, are they at points of confusion for reads like in adapter regions which need to be trimmed by fastp? Idk if this is actually a problem, just an example
* Look a little more into the qc info
* Box plots for contig N50 and related stuff for used/unused data. Might show things better than the diagonal KDE plots on my pairplot
* Number of chromosomes possessed by an organism may affect how well their MSMC curves/parameters are estimated
  + Larger chromosome count implies the need for a larger number of parameters to be estimated
  + [Paper](https://link.springer.com/protocol/10.1007/978-1-0716-0199-0_7), Section 5.2
  + Bear paper: 64 time segments for ~37 chromosomes pairs
  + Dog paper: XX time segments for ~40 chromosomes
  + Human paper: XX time segments for 23 chromosomes
* We might want to try an MSMC on a human and try to match it to a paper as a kind of benchmark

Apps

* Focus on this more
* Consider clustering by certain windows of points (might already work with point clipping)
  + Force x and y min/max to focus on points considered for clustering (window of points that is actually clustered)
* Remeber the shape summary thing is something that I can do
* Associate curve clusters with
  + Phylogeny based things
    - Order/Family/taxa stuff
    - Geography based things
    - Latitude/location
    - Types of barriers
  + Creature size
  + Chromosome size
  + Any shared features
  + Environment
    - Temperature
    - Air composition
  + “Idiosyncratic vs group events” - Russ
* Beware of biases
* Russ mentioned something about random sampling or something for determining association between order/family distr in clusters
  + I said that the high use of Passeriformes or 1-2 samples

**Todos:**

* **I think tslearn is valid for my (63, 2) arrays since algo is valid for multidimensional cases**
  + [**Tslearn source code with documentation**](https://github.com/tslearn-team/tslearn/blob/ed745ba1401cd182a8c55642fb0f7ed2bb6f1e4b/tslearn/metrics/dtw_variants.py#L1144)
* **Learn how to import a model to quickly cluster curves**
* **Reshare Notebook and drive with Erik**
* **Make a table with Species name/id and associated cluster so Erik can take a peak**
* **Start creating my own MSMCs**
  + **Maybe try doing MSMCs on multiple sequences of the same organism**
* Draw DBAs
* Add in finished mammal data
  + Generate generation lengths
  + Color them differently
* Look more at DTW inertia for helping to define cluster spread
* Why are there loops in the DBA??? :-(((
* Consider how well DTW clustering works for characterizing Orders/Fams/Species vs Is there a biological significance behind these curves other than that they are of similar shapes?
  + They could share things like ancestry? Definitely not as well as things like the human samples in Ricky D’s paper
  + Is there any association between the features of curves (mins/maxes) and geographical location? (Might not be any)
  + Is there a way to determine how much divergence there is between different species (maybe in the form of a time series curve) while considering specific moments of their NE curves?
    - Is there association between NE curve height and divergence between 2 samples/species?
      * Is there a meaningfulness of Tajima’s D. for this as well?
* Consider rereading and looking over citations of the PSMC/MSMC paper which Erik sent a while ago
  + Do this for a better understanding of how this project uses coalescent theory
* Might also want to consider a function that classifies which samples fit in which clusters
* **Add in metric for quality of cluster (How much variation does the red line explain, think PCA)**
  + **Does TSlearn provide anything for this for DTW?**
  + **Silhouette coefficient analysis for assessing within cluster variation over number of clusters**
    - **This should produce an elbow curve like analysis to tell me the optimal number of clusters**
    - **With the optimal number of clusters, I should be able to list out the variation within each cluster**
      * **Draw out standard deviation bands?**
    - **Also play around with SD bands**
    - **Check out** [**other papers**](https://scholar.google.com/scholar?cites=2636100862607043646&as_sdt=2005&sciodt=0,5&hl=en) **which used tslearn package on Google Scholar**
    - **!!! Might need to modify tslearn DTW for 2d curves since it usually only works on X coords and not X coord, Y coord pairs !!!**
      * **NE is indexed by specific dates**
      * **TEST OUT: See if removing X values does anything to clusters. If it does then tslearn somehow accounts for x,y. Else if nothing changes, then I’ll need to add in this functionality to account for x, y myself**
      * [**Somewhat related paper**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5668684/)
      * **Might also consider just adding artificial points where points at all X positions are accounted for depending on where Y positions might be in steps**
        + **Might be related to the application of curve smoothing before clustering**
* **Consider coloring newly added curves Green to sanity check that they aren’t just shoved into one cluster**
* **Automate and make chart generation more clear (Especially with what data is considered)**
  + **Maybe even throw in reports**
* Erik mentioned a future with using Mammal samples
* Consider selection on sites and how they might affect
* some other methods like Multi-dimensional analysis things that Erik mentioned
* See how Orders/families are distributed among clusters
  + The clusters don’t seem to be good at separating Orders
* Make filters stricter (less likely to filter out curves) to preserve curves that show hints of

plausibility (curvier curves are better curves)

* + Also consider N50 somehow when throwing stuff out
* Consider adding Order/Family distribution to the Cluster Distr of Kmean
  + Something like a barchart breakdown of each bar’s Orders/Families
* Are the species curves which were filtered out predominantly belong to a particular order or family?
  + Might relate to some of the findings in the [paper](https://www.nature.com/articles/s41586-020-2873-9#Sec2) like the abundance of TEs/LINEs (repeats) from Piciformes/Bucerotiformes
  + Might even be something with lineage-specific sequences

Notes:

* Detecting selection at ORFs? Something that Jakob was talking about
  + Virion?
* Counting allele frequencies with a tree?
* Probability of ongoing transmission?
* Non trivial problem: Figuring out how selection occurs and how it shapes a phylogeny through time
  + Another non trivial problem are massive tree/lineage bifurcation events
    - A factor might be epistasis which may have been seen by some Flu strain in the 1970s - mentioned by Russ
* Does the order in which curves can be sorted within their own cluster (sorted/ordered by similarity from most to least similar) have any meaning?
  + Can it be interpreted as an evolving process?
    - I think it likely doesn’t for now since the only thing these curves share in common are similarly inferred demographic histories
* [**Effective population size and patterns of molecular evolution and variation**](https://www.nature.com/articles/nrg2526)
  + Extreme inbreeding (100% selfing) causes Ne to be multiplied by a factor approaching ½, halving the mating population Ne
  + Bottlenecks of Ne have been observed in populations moving out of Africa
  + Paper claims that studying Ne is useful for “designing conservation or selective breeding programmes, and for interpreting data on DNA sequence variation and evolution”
  + Theoretical results of NE aren’t accurate reflections of population sizes, but do indicate trends and evolutionary processes
    - Evolutionary processes are better understood when considering selection (diffusion equations)
  + Efficacy of selection in reduced when Ne is less than N when compared with the Wright-Fisher population of size N

What could have caused MSMC to output bad curves given scaffold data? What is up with the scaffold data specifically?

* Look into quality in supplementary info of Dense sampling paper
* Study some coalescent theory
* [Supplementary Info](https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-020-2873-9/MediaObjects/41586_2020_2873_MOESM1_ESM.pdf)
  + NCBI and Out genomes were
    - Contained in
      * [B10K db](https://b10k.genomics.cn/species.html)
      * [NCBI Project](https://www.ncbi.nlm.nih.gov/nuccore/?term=PRJNA545868)
    - Contig N50 > 5 kb
    - Scaffold N50 < 30 kb
    - Tot Asm length > 0.9 Gb as average bird genome is ~1.2 Gb