**SNPGenie to measure nucleotide diversity**

From Wisco cat transmission paper author github:

* + The purpose of this script is to generate genewise nucleotide diversity (π) using the SNPGenie script and then to generate plots to visualize and compare genewise πS (synonymous diversity) and πN (nonsynonymous diversity).
  + Here's the syntax for running SNPGenie that I am using:
  + perl snpgenie.pl --vcfformat=4 --snpreport=path/to/SNPREPORT.vcf --fastafile=path/to/ref.fasta --gtffile=/path/to/GTF.gtf
* Need: sample vcf file, reference genome fasta and a gtf file
* Got wa1.fasta from the pipeline reference sequences, and we have a vcf file from each replicate of each sample, so I just picked the first replicate
* Gtf is an older format version of gff which we used—I tried converting our gffs to gtf but this didn’t work, so I went back to NCBI to download a gtf for WA1 and renamed it wa1.gtf <https://www.ncbi.nlm.nih.gov/assembly/GCA_009937905.1>
* Cloned snpgenie from github

git clone <https://github.com/chasewnelson/SNPGenie.git>

* For some reason vcfformat=4 didn’t work for me, it’s something about the heading, and I found in a forum that somebody else had the same issue, so I changed to vcfformat=2

perl snpgenie.pl --vcfformat=2 --snpreport=D47.wa1.bam.snv.vcf --fastafile=wa1.fasta --gtffile=wa1.gtf

* SNPgenie outputs a “population summary” table that is for the sample as a whole, and a “product results” table that breaks it down by gene product

**1.25.21 re-running with new vcf files output by the pipeline for all samples**

Copied over all the \_R vcf files

Cp /home/lbashor/for\_SARS2\_manuscript/for\_SARS2\_manuscript4/viral\_variant\_caller/results/vcf/\*\_R.wa1.bam.snv.vcf .

(And copied over Cat\_5 which doesn’t have a \_R file)

perl snpgenie.pl --vcfformat=2 --snpreport=vcfs\_1.25.21/Cat\_1\_R.vcf --fastafile=wa1.fasta --gtffile=wa1.gtf

error says ########################### CURRENTLY PROCESSING: ###########################

vcfs\_1.25.21/Passage\_3\_R\_wa1.bam.snv.vcf...

## WARNING: Conflicting SNP Report formats detected. Does the file extension match expectation? If not, please contact author. ## SNPGenie TERMINATED.

To fix this, I deleted the wa1.bam.snv part of the file, just leaving .vcf

perl snpgenie.pl --vcfformat=2 --snpreport=vcfs\_1.25.21/Cat\_4\_R.vcf --fastafile=wa1.fasta --gtffile=wa1.gtf

This command works ^

Ran them all individually, it takes maybe 15seconds, but it would be nice to find a command to run them all. Unfortunately it creates a SNPGenie Results folder each time, and I had to manually rename it after doing each sample

Main output files I’m interested in are population\_summary.txt and product\_results.txt

I put together all the population and product files into two big summary files, viewable in Excel

From the Wisco paper:

https://www.biorxiv.org/content/10.1101/2020.11.16.384917v2

To probe the evolutionary pressures shaping SARS-CoV-2 viruses within hosts, we first evaluated the proportion of variants shared between cats. Eighty-six percent of variants (42 out of 38 iSNVs and 11 indels) were found in a single cat (42/49), 8% of variants were found in 2-5 cats (4/49), and the remaining 6% of variants were found in all 6 cats (3/49).

Purifying selection, which acts to purge deleterious mutations from a population, is known to result in an excess of low-frequency variants. In contrast, positive selection results in the accumulation of intermediate- and high-frequency variation [36](https://www.biorxiv.org/content/10.1101/2020.11.16.384917v1.full#ref-36).

Next we compared nonsynonymous (πN) and synonymous (πS) pairwise nucleotide diversity to further evaluate the evolutionary forces shaping viral populations in index and contact animals [38](https://www.biorxiv.org/content/10.1101/2020.11.16.384917v1.full#ref-38). Broadly speaking, excess nonsynonymous polymorphism (πN/πS > 1) points toward diversifying or positive selection while excess synonymous polymorphism (πN/πS < 1) indicates purifying selection. When πN / πS, is approximately 1 genetic drift – stochastic changes in the frequency of viral genotypes over time – can be an important force shaping genetic diversity

One thing I’m confused about here is if you go into the literature, and to the main SNPGenie publication, they say this:

In general, π*N* = π*S* indicates neutrality, π*N* < π*S* indicates purifying selection and π*N* > π*S* may indicate positive selection favoring multiple amino acid changes ([Hughes, 1999](javascript:;)).

<https://academic.oup.com/bioinformatics/article/31/22/3709/241742>

This seems better because if piN or piS is 0, you’ll get 0 or undefined for piN/piS, and that won’t be as informative as comparing them directly

For example, there were a lot of times in our samples that piS was 0, but piN was a number, so this indicates positive selection, but we just get a divide by zero error

The other question I still have is how close do they have to be for neutrality? If piN/piS is 0.9, or 1.3, as it is for some of our samples, what does that mean?