**Aciurina in NM comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **MAL** | **SNM** | **TES** | **RGG** | **ABQ** |
| **Atrix** |  | 3 | 3 | 3 | 3 |
| **Abige** | 3 | 3 | 3 | 3 |  |
| **Anota** | 3 |  | 3 |  |  |

With these samples we want to know:

How similar are Atrix and Abige?

(Phylogeny, basic sequence similarity)

Phylogeny of Atrix, Abige, Anota

How much genetic exchange do Atrix and Abige have?

(abba/baba, fst, etc.)

How does this compare to the Atrix-Anota and Abige-Anota interbreeding?

Does the pattern change when species are in sympatry?

Structure plot of Atrix, Abige, Anota – what is the optimum k?

Do we need 4 species to perform any of these tests?

If so, the 4th lineage can be the 3 Aopac

MAC:  
The choice of the minimum minor allele count (min.mac) in your VCF filtering depends on the specific goals of your analysis and the characteristics of your dataset. Here are some considerations for choosing the appropriate min.mac value:

Data Quality and Stringency: If you want to be more conservative and retain only high-quality variants, you might choose a higher min.mac value (e.g., min.mac = 2 or min.mac = 3). This can help filter out variants with low allele counts, which could be due to sequencing errors or other artifacts.

Retaining Information: On the other hand, if you want to retain a larger number of variants to capture more genetic diversity, you might choose a lower min.mac value (e.g., min.mac = 1). This is more permissive and keeps more variants, even if they are less common.

Sample Size: The choice of min.mac can also be influenced by the size of your sample. In smaller sample sizes, it's common to have lower allele counts for rare variants. So, for larger datasets, you might consider a higher min.mac value.

Research Objectives: Consider your research goals. If you are studying rare genetic variants that are biologically important, you might choose a lower min.mac. If you are concerned about false positives, a higher min.mac may be more appropriate.

Quality Control: It's a good practice to assess the quality of your sequencing data and understand the characteristics of your variants. You can visualize allele frequencies, quality scores, and other metrics to make an informed decision about the appropriate min.mac value.

In summary, there is no one-size-fits-all answer, and the choice of min.mac depends on the specifics of your analysis. It's common to try different values, assess the impact on your results, and choose the one that best aligns with your research goals and quality control criteria. If you're unsure, starting with min.mac = 2 is a reasonable choice, and you can adjust it based on your data exploration and analysis needs.

Trevor’s methods and filtering parameters:

Dried leaf material from each plant was ground into powder form using a Qiagen TissueLyser and DNA was extracted using Qiagen DNeasy Plant kits (Qiagen, Valencia CA). Reduced representation libraries were prepared using a double-digest restriction-site associated DNA sequencing (ddRADseq) method (Parchman et al., 2012; Peterson et al., 2012). The restriction endonucleases EcoRI and MseI were used to digest genomic DNA and uniquely barcoded Illumina adaptors were subsequently ligated to the resulting fragments using T4 DNA ligase (New England Biolabs, Ipswich MA). Barcoded fragments were PCR amplified with Iproof DNA polymerase (BioRad, Hercules CA), and fragments ranging from 350 to 450 bp were size selected using a Pippin Prep quantitative electrophoresis unit (Sage Science, Inc). Single-end sequencing (100 bp read lengths) was performed using two lanes on an Illumina HiSeq 4000 platform at the University of Wisconsin.

Sequences representing potential contaminants (E. coli and PhiX) or various Illumina associated oligos were detected and discarded using a pipeline of Perl and bash scripts (http://github.com/ncgr/tapioca). A custom Perl script was then used to correct barcode sequencing errors, to trim cut site and barcoded oligo associated bases, and to demultiplex reads by individual. As a reference genome is not available for E. nauseosa or a close relative, we used a de novo assembly of unique reads to create a consensus reference of genomic regions sampled by our reduced representation approach (reference assembly, hereafter). Parameters for this step were optimized using shell scripts and documentation provided for dDocent (Puritz et al., 2014) (cutoffs: individual = 8, coverage = 5; clustering similarity: -c .92), and the reference assembly was produced using cd-hit-est (L. Fu et al., 2012). Demultiplexed reads for each individual were mapped to the reference assembly using bwa-mem v0.7.17 (Li, 2013). Sequence variants were identified using samtools v1.9 and bcftools v1.9 (Li et al., 2009), and subsequent filtering was completed using vcftools v0.1.16 (Danecek et al., 2011). **We retained only biallelic loci covered by reads present in at least 70% of individuals, thinned to one locus per contig to reduce effects of linkage disequilibrium and sequencing error. Additionally, individuals missing data for greater than 40% of the loci were removed before further downstream analyses. After additional filtering with custom Python scripts, we retained loci with minor allele frequency (MAF) ≥ 0.02, mean depth across individuals ≥ 3 or < 25, and alternate allele call quality ≥ 750. We removed loci with excessive coverage and only retained loci with two alleles present to ameliorate genotyping bias from the potential misassembly of paralogous genomic regions (Hapke & Thiele, 2016; McKinney et al., 2018). Lastly, we retained loci with FIS > –0.5, as misassembly of paralogous genomic regions can lead to abnormal heterozygosity (Hohenlohe et al., 2013; McKinney et al., 2017). To detect duplicated, clonal, or highly related individuals, we quantified relatedness among all pairs of individuals within each population using vcftools employing the cutoff criteria provided by the king method (Manichaikul et al., 2010; Turner et al., 2022).**