# pixy (Korunes and Samuk, 2021)

* v1.2.7
* Documentation - <https://pixy.readthedocs.io/en/latest/>
* Manuscript - <https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13326>

## Notes

* Pixy does not use the filter column in vcf files. The input vcf file must be hard filtered
* In the pixy manuscript, they filtered their vcf files by the following:
  + DP > =10, GQ >= 40|RGQ >= 40

The goal is to compare sequence divergence between introgressed segments and segments confidently called for the species tree. So, I calculated the divergence for each introgression tract and each species tree tract and plotted the divergence distributions for both.

The data look normally distributed, and the variances are equal. However, the data sets are different sizes. More segments are confidently called for the species tree and these segments are much larger on average.

For the species tree divergence distribution, my thought was to create 1,000 replicate data sets by randomly permuting the same number and size of introgressed segments among the intervals called for the species tree.

Then I would calculate the mean divergence for the segments within each replicate and have a distribution of 1,000 mean divergence values, which I could compare back to the distribution of introgression tract divergence values.

\*\*Why calculate P1-P4 dXY? Introgression from P2 into P3 will decrease P1-P3 divergence in introgressed regions.

# Analysis

1. Generate invariant sites vcf
2. pixy.sh – Calculate dXY for each interval in the introgressed tracts interval file (ten\_kb\_tracts.bed)
   1. Calls bootstrap\_pixy.py – Get the distribution of mean dXY for 1,000 resampled datasets with replacement. Resample the introgression tracts with replacement, recording the dXY. Create 1,000 replicate data sets the same size as the original introgressed tracts interval file.
3. Create\_dxy\_null\_distribution.py – Randomly permute the same number and size of introgressed intervals into the regions confidently called for the species tree. Calculate dXY for all intervals. Create 1,000 replicate datasets and get a distribution of mean dXY.
4. Divergence\_analysis.R – Compare distributions of introgressed regions and species tree regions. Run t-test and visualize.

## Variant Filtering

1. Separate variant and invariant records
   1. We must do this because invariant sites don’t have a lot of the attributes that GATK recommends filtering by (e.g., GQ). Sometimes the invariant sites have a QUAL score, other times, they do not. I’m not sure why this is. You cannot filter invariant sites by QUAL, or you’ll lose a lot of data. Instead, we can filter based on a metric called reference genotype confidence (RGQ) - "Unconditional reference genotype confidence, encoded as a phred quality -10\*log10 p(genotype call is wrong)"
2. Filter variants by GATK recommendations and select only those passing all filters.
3. Merge the filtered variant file with the invariant file
4. Set individual genotypes with low depth, genotype quality, or reference genotype confidence to missing. Remove all the records where both samples are missing genotypes.
   1. bcftools filter -S . -e "FMT/RGQ<30 | FMT/DP<10 | FMT/GQ<30" input\_vcf | bcftools filter -e 'F\_MISSING > 0.5' > output\_vcf

# bcftools Notes

* Hard filter variants with low quality
  + bcftools filter -e 'QUAL<30' input\_vcf > output\_vcf
* Set individual samples to missing (./.) if depth or genotype quality is low.
  + bcftools filter -S . -e 'FMT/DP<3 | FMT/GQ<20'