

SALS microsatellite population structure manuscript - IBD analysis documentation

The below information contains a complete record of the data sources, input file generation and structure, analysis and Rcode, output data and/or figures generated, and session information I used to complete the above title-referenced analysis for publication.

The following can also be compiled in RStudio from the file [SALSusat-IBD.Rnw](#)

Motivation: Presented below is a Mantel test [[Mantel, 1967](#)] to test for isolation-by-distance in saltmarsh sparrows; i.e., to evaluate the correlation of pairwise geographic distances among saltmarsh sparrow breeding marshes studied and the corresponding pairwise genetic distances of patch-level "populations" of individuals captured from each marsh. The output of the below analysis should be essentially identical to the old results given to me by Adrienne (see tab 'GGDvFST 13 loci MT' in the file '[SHARP IBD.xlsx](#)', dated July 31, 2016; or file '[IBD 13 loci.pdf](#)', from August 13, 2016; or slide/page 14 of file [SHARP - metapop gen PI retreat.pdf](#), from August 17, 2016). I have only rerun it here in order to use linearized F_{ST} as recommended by [Slatkin \[1995\]](#) and [Rousset \[1997\]](#); to make the final figure a vector format and cleaned up to be more suitable for publication; as well as to more fully document the data & analysis for posterity.

Add required packages:

```
> library(ggplot2)
> library(dplyr)
```

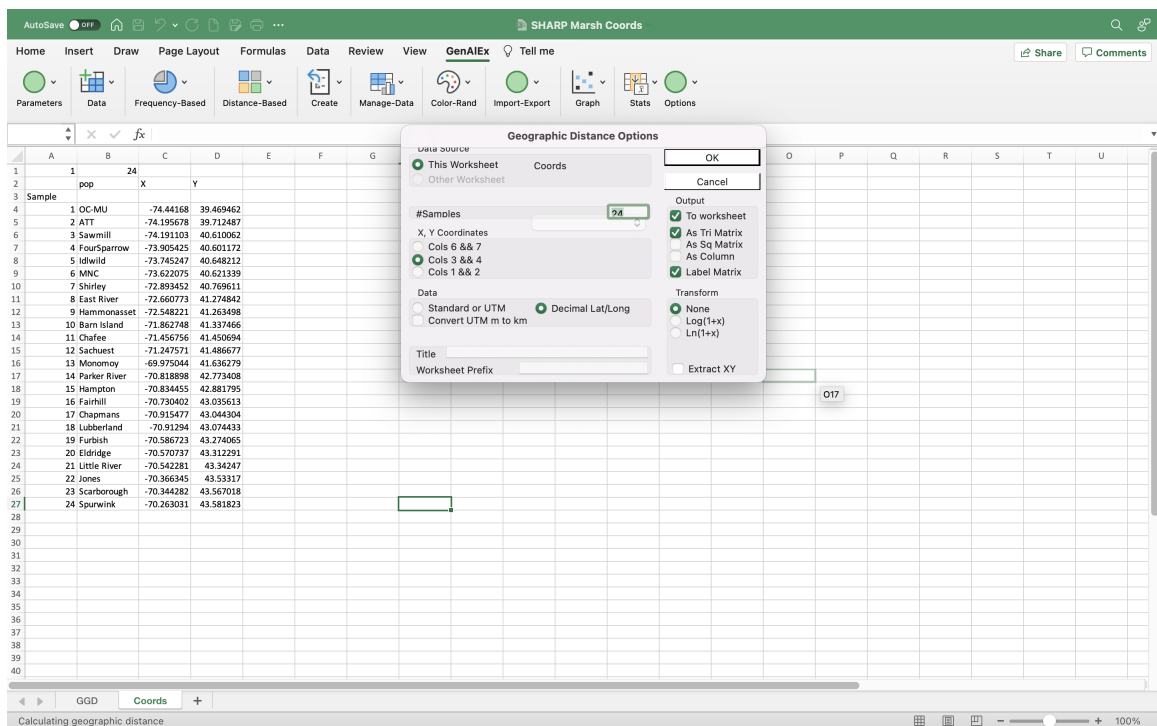
Read in data*:

```
> ARdata2 <- read.table("~/Documents/ModernSparrowGenomics/Copies of Associated
  ELN files/SALS usat ms/Data and Analyses/AR.txt", header = TRUE)
> ARdata <- read.table("~/Documents/ModernSparrowGenomics/SALSusat/LEF -
  Additional Manuscript Analyses_Notes/Genetic Diversity/Input/fstatalelic
  richness by popmelt.txt", header = TRUE)
```

***Documentation of data sources and structure:**

1.) File [GeoDist.txt](#).

Source: These data were generated in Genalex (verified using GenAlEx v. 6.51b2) based on the decimal degree coordinates for each marsh found on the 'Coords' tab in the [SHARP Marsh Coords.xlsx](#) file. I'm not sure who originally generated this file (dated July 29, 2016) or supplied the marsh coordinates, presumably Jen Walsh, Adrienne Kovach or maybe Bri Benvenuti. The output (GGD tab of above file) was generated using the Genalex options Distance -> Geographic... and options as shown below to calculate the pairwise distances according to the formula used by Genalex noted in the manual ([Genalex 6.502 Appendix 1](#)): "GenAlEx uses a modification of the Haversine Formula developed by R.W. Sinnott (Virtues of the Haversine (1984) Sky and Telescope 68,159) following computer code published online by Bob Chamberlain from JPL, NASA. (<http://www.usenet-replayer.com/faq/comp.infosystems.gis.html> still available on 12/12/12). Distances calculated via Lat/Long coordinators are returned in km." According to Wikipedia, [the Haversine formula](#) calculates great-circle distance, *not* Euclidean distance. This text file is a copy of the GGD tab, formatted for R (Genalex-specific formatting and column headers removed; blank fields in the upper triangle and self-pair distances of 0.000 changed to 'NA').



Calculating pairwise geographic distance from lat/long coordinates in Genalex

Data structure: Triangular matrix of pairwise geographic distance values.

Data units: Kilometers

Data snippet:

```
> head(ARdata, n = 5L)
```

```

      AR POP
1  3.609 OCMU
2  7.302 OCMU
3 10.490 OCMU
4   8.300 OCMU
5   8.305 OCMU
```

Run Mantel test:

```
> ARdata$POP <- factor(ARdata$POP, levels=c("OCMU", "ATT", "SAW", "FourSp", "IDL",
, "MNC", "SHIRLEY", "EASTRIVER", "HAMMO", "BARNI", "CHAFEE", "SACHUEST", "
MONOMOY", "PARKER", "HAMPTON", "FAIRHILL", "CHAPMANS", "LUB", "FURBISH", "
ELDRIDGE", "LITTLER", "JONES", "SCAR", "SPUR"))
> AR<- ARdata %>%
+   ggplot(aes(x=POP, y=AR, fill=POP)) +
+   geom_violin(width=1) +
+   geom_boxplot(width=0.1, color="grey", alpha=0.2) +
+   theme_bw() +
+   theme(
+     legend.position="none"
+   ) +
+   xlab("Breeding Marsh") + ylab("Allelic Richness") + theme(text = element_
text(family = "Constantia", color = "grey20", size=20))
> labels <- c("OCMU"="OC-Mullica", "ATT"="ATT", "SAW"="Sawmill", "FourSp"="Four
Sparrow", "IDL"="Idlewild", "MNC"="Marine Nature Center", "SHIRLEY"="Wertheim",
, "EASTRIVER"="East River", "HAMMO"="Hammonasset", "BARNI"="Barn Island", "
CHAFEE"="Chafee", "SACHUEST"="Sachuest", "MONOMOY"="Monomoy", "PARKER"="Parker
River", "HAMPTON"="Hampton", "FAIRHILL"="Fairhill", "CHAPMANS"="Chapman's
Landing", "LUB"="Lubberland Creek", "FURBISH"="Furbish", "ELDRIDGE"="Eldridge",
, "LITTLER"="Little River", "JONES"="Jones", "SCAR"="Scarborough", "SPUR"="
Spurwink")
> #ARFinal <- AR + theme(axis.text.x = element_text(angle = 90, vjust = 0.5,
hjust=1))+ scale_x_discrete(labels = labels)+scale_fill_manual(values=c("
midnightblue", "midnightblue", "midnightblue", "#741A36", "#741A36", "
midnightblue", "midnightblue", "midnightblue", "midnightblue", "midnightblue", "
midnightblue", "midnightblue", "midnightblue", "midnightblue", "midnightblue", "
midnightblue", "midnightblue", "midnightblue", "midnightblue", "midnightblue",
"#749D94", "midnightblue", "midnightblue", "#749D94"))
>
> ARFinal <- AR + theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust
=1))+ scale_x_discrete(labels = labels)+scale_fill_manual(values=c("
midnightblue", "midnightblue", "midnightblue", "midnightblue", "midnightblue",
"midnightblue", "midnightblue", "midnightblue", "midnightblue", "midnightblue",
"midnightblue", "midnightblue", "midnightblue", "midnightblue", "midnightblue",
"midnightblue", "midnightblue", "midnightblue", "midnightblue", "midnightblue",
"midnightblue", "midnightblue", "midnightblue", "midnightblue"))
```

Printout of results:

```
> #anova(ARdata)
```

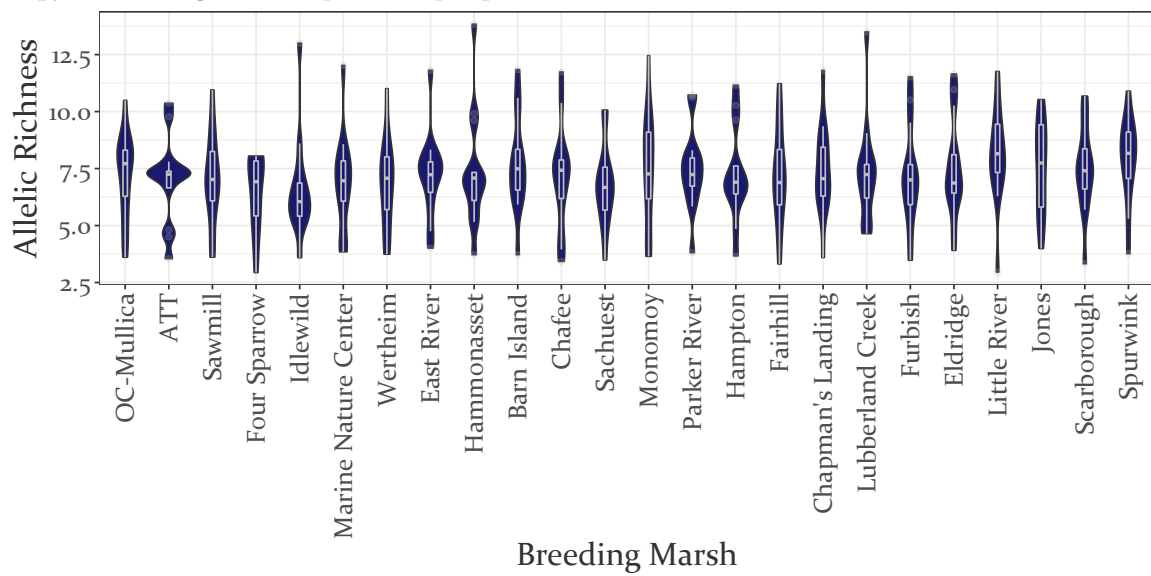
Format final figure for publication:

```
>
```

Save final figure:

```
> ggsave(file = "/Users/Lindsey/Documents/ModernSparrowGenomics/Copies of
Associated ELN files/SALS usat ms/Final Figures/ARViolinplotFinal.pdf", plot =
ARFinal, device = cairo_pdf, width = 10, height = 5, units = "in")
```

Copy of resulting [ARViolinplotFinal.pdf](#) plot:



Record of Session Info:

```
> sessionInfo()

R version 4.0.3 (2020-10-10)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Big Sur 10.16

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods   base

other attached packages:
[1] dplyr_1.0.2  ggplot2_3.3.5

loaded via a namespace (and not attached):
 [1] rstudioapi_0.13  magrittr_2.0.2  tidyselect_1.1.0 munsell_0.5.0
 [5] colorspace_2.0-3 R6_2.5.1        rlang_1.0.1      fansi_1.0.2
 [9] tools_4.0.3      grid_4.0.3      gtable_0.3.0     utf8_1.2.2
[13] cli_3.2.0        withr_2.4.3     ellipsis_0.3.2   digest_0.6.29
[17] tibble_3.1.6     lifecycle_1.0.1 crayon_1.5.0     purrr_0.3.4
[21] farver_2.1.0     vctrs_0.3.8     glue_1.4.2       labeling_0.4.2
[25] compiler_4.0.3   pillar_1.7.0    generics_0.1.2   scales_1.1.1
[29] pkgconfig_2.0.3
```

References

- Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27 (2 Part 1):209–220, 1967. ISSN 0008-5472.
- F. Rousset. Genetic differentiation and estimation of gene flow from f-statistics under isolation by distance. *Genetics*, 145(4):1219–1228, 1997. ISSN 0016-6731 (Print). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9093870.
- M. Slatkin. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1): 457–62, 1995. ISSN 0016-6731 (Print). URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7705646.