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(extrafont)
> library(ggplot2)
> library(vegan)
```

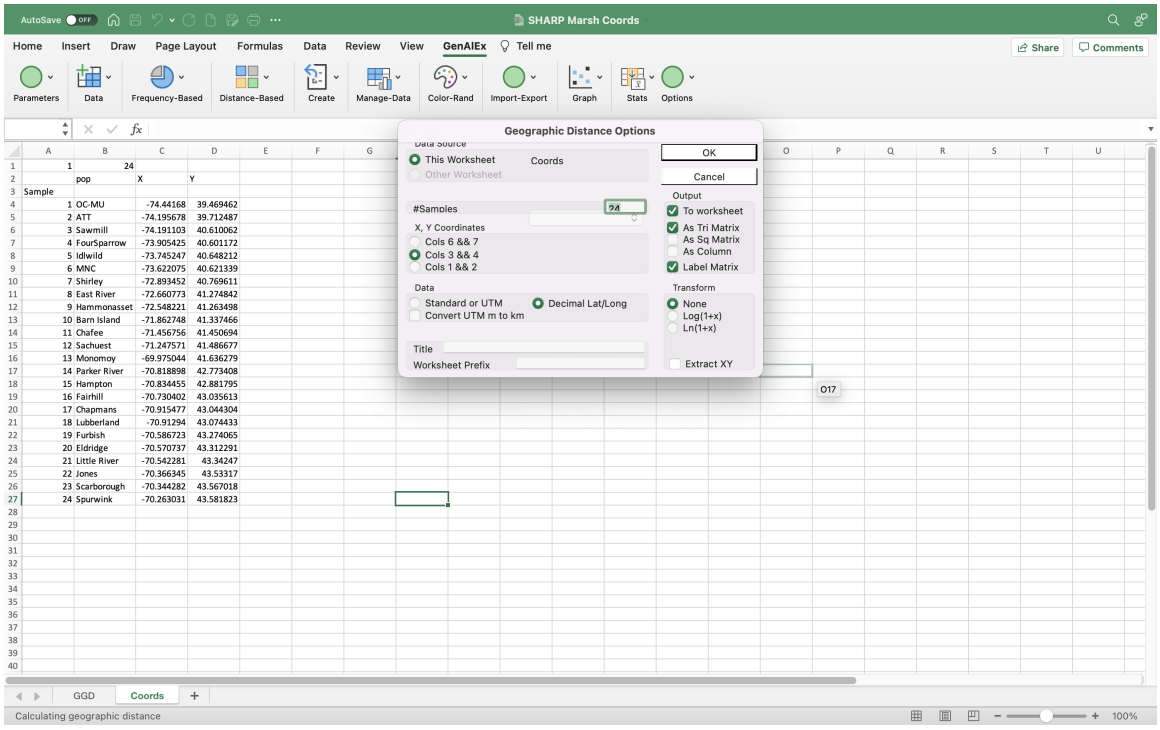
Read in data*:

```
> GeoDist <- read.table("~/Documents/ModernSparrowGenomics/Copies_of_Associated_
    ELN_files/SALS usat ms/Data and Analyses/GeoDist.txt", header = FALSE)
> GenDist <- read.table("~/Documents/ModernSparrowGenomics/SALSusat/LEF -
    Additional Manuscript Analyses_Notes/IBD and SA/FstatGenDist.txt", header =
    FALSE)
> GGDvFST13Datab <- read.table("~/Documents/ModernSparrowGenomics/SALSusat/LEF -
    Additional Manuscript Analyses_Notes/IBD and SA/IBD-GGDFST13LocidLinearizedFST
    .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(GeoDist, n = 5L)
```

```
     V1     V2     V3     V4     V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
1 28.312     NA     NA     NA     NA NA NA NA NA  NA  NA  NA  NA  NA  NA  NA
2 44.178 27.191     NA     NA     NA NA NA NA NA  NA  NA  NA  NA  NA  NA  NA
3 68.797 42.177 31.767     NA     NA NA NA NA NA  NA  NA  NA  NA  NA  NA  NA
4 85.362 57.740 49.591 17.871     NA NA NA NA NA  NA  NA  NA  NA  NA  NA  NA
5 97.706 69.659 63.273 31.513 13.721 NA NA NA NA  NA  NA  NA  NA  NA  NA  NA
  V17 V18 V19 V20 V21 V22 V23
1  NA  NA  NA  NA  NA  NA  NA
2  NA  NA  NA  NA  NA  NA  NA
3  NA  NA  NA  NA  NA  NA  NA
4  NA  NA  NA  NA  NA  NA  NA
5  NA  NA  NA  NA  NA  NA  NA
```

2.) File FstatGenDist.txt.
Source:
Fill in . . .

Data structure: Triangular matrix of F_{ST} values.
Data units: "F_{ST}" - a measure of genetic differentiation describing the variance in allele frequencies among subpopulations relative to the total population; range 0.0 - 1.0, where lower values indicate greater similarity among populations and higher values indicate greater differentiation among populations.
Data snippet:

```
> head(GenDist)
```

```
           V1          V2       V3       V4       V5       V6 V7 V8
1 -0.000499750          NA       NA       NA       NA       NA NA NA
2  0.009693053 0.011531459       NA       NA       NA       NA NA NA
```

```
3  0.035411058 0.042318115 0.03316458          NA          NA           NA NA NA
4  0.021033286 0.015228426 0.01677682 0.04525975          NA           NA NA NA
5  0.006846557 0.002807862 0.01358200 0.04668202 0.01306859           NA NA NA
6 -0.000499750 0.003512293 0.01358200 0.04832792 0.01698363 0.006339942 NA NA
  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23
1 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
2 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
3 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
4 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
5 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
6 NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

3.) File IBD-GGDFST13LocidLinearizedFST.txt.

Source:

Fill in . . .

Data structure: Rectangular matrix of pairwise geographic distance and genetic distance (F_{ST}) values.

X = GGD; values are listed in order from ()

Fill in . . .

Y = F_{ST}; values are listed in order from ()

Fill in . . .

Data units: Kilometers & F_{ST}

Data snippet:

```
> head(GGDvFST13Datab)

       X            Y
1 28.312 -0.000501748
2 44.178  0.009714463
3 27.191  0.011490532
4 68.797  0.035325299
5 42.177  0.042328980
6 31.767  0.033207282
```

Run Mantel test:

```
> MantelIBD <- mantel(GeoDist, GenDist, method="pearson", permutations=10000,
    strata = NULL, na.rm = TRUE, parallel = getOption("mc.cores"))
```

Printout of results:

```
> MantelIBD
```

```
Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = GeoDist, ydis = GenDist, method = "pearson", permutations = 10000,
        strata = NULL, na.rm = TRUE, parallel = getOption("mc.cores"))

Mantel statistic r: 0.2099
      Significance: 0.022298

Upper quantiles of permutations (null model):
  90%   95% 97.5%   99%
0.134 0.173 0.205 0.239
Permutation: free
Number of permutations: 10000
```

*As the significance value changes slightly in the thousandths place when the permutations are rerun (from 0.021 - 0.024 that I have seen), I truncated the reported P value to 2 significant figures.

Get intercept and slope values for drawing a best fit regression line:

```
> summary(lm(GGDvFST13Datab$Y ~ GGDvFST13Datab$X))$coefficients

                    Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)      1.257806e-02 1.109105e-03 11.340735 1.070253e-24
GGDvFST13Datab$X 1.664207e-05 4.532116e-06  3.672032 2.892680e-04
```

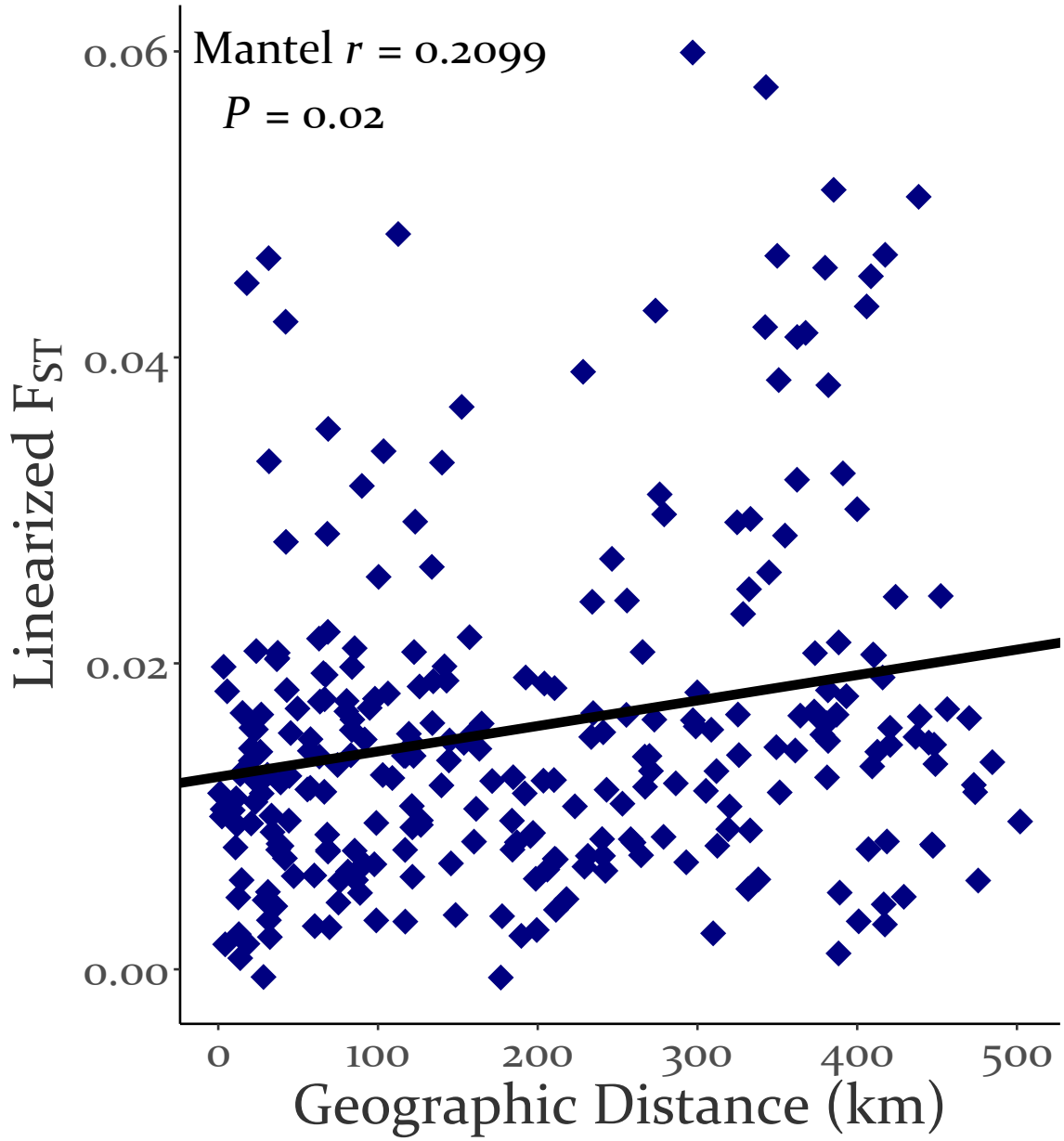
Format final figure for publication:

```
> eq1 = paste0("Mantel ~italic(r) == ", 0.2099)
> eq2 = paste0("~italic(P) == ", 0.02)
> IBD<-ggplot(GGDvFST13Datab, aes(x=X, y=Y)) + geom_point(size=5, shape=18, color
    = "navy") + xlab("Geographic Distance (km)") + ylab((y=expression("Linearized
    F"[S][T]))) + annotate(geom="text", x=50, y=0.056, family = "Constantia",
    label=eq2, size=7, parse=TRUE) + annotate(geom="text", x=95, y=0.06, family =
    "Constantia", label=eq1, size=7, parse=TRUE)
> IBD2 <- IBD + geom_abline(intercept = 1.257806e-02, slope = 1.664207e-05, size
    =2)+theme_bw()+theme(panel.grid.major = element_blank(), panel.grid.minor =
    element_blank(), panel.border = element_blank(), axis.line = element_line(
    colour = "black"))+theme(text = element_text(family = "Constantia", color = "
    grey20",size=25))
```

Save final figure:

```
> ggsave(file = "/Users/Lindsey/Documents/ModernSparrowGenomics/Copies_of_
    Associated_ELN_files/SALS usat ms/Final Figures/IBDplotFinal.pdf", plot = IBD2
    , device = cairo_pdf, width = 6.5, units = "in")
```

Copy of resulting IBDplotFinal.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] vegan_2.5-7     lattice_0.20-41 permute_0.9-5   ggplot2_3.3.5
[5] extrafont_0.17

loaded via a namespace (and not attached):
 [1] pillar_1.7.0     compiler_4.0.3   tools_4.0.3      digest_0.6.29
 [5] lifecycle_1.0.1  tibble_3.1.6     gtable_0.3.0     nlme_3.1-149
 [9] mgcv_1.8-33      pkgconfig_2.0.3  rlang_1.0.1      Matrix_1.2-18
[13] cli_3.2.0        rstudioapi_0.13  parallel_4.0.3   Rttf2pt1_1.3.9
[17] withr_2.4.3      dplyr_1.0.2      cluster_2.1.0    generics_0.1.2
[21] vctrs_0.3.8      grid_4.0.3       tidyselect_1.1.0 glue_1.4.2
[25] R6_2.5.1         fansi_1.0.2      farver_2.1.0     purrr_0.3.4
[29] extrafontdb_1.0  magrittr_2.0.2   scales_1.1.1     ellipsis_0.3.2
[33] MASS_7.3-53      splines_4.0.3    colorspace_2.0-3 labeling_0.4.2
[37] utf8_1.2.2       munsell_0.5.0    crayon_1.5.0
```

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.