**Supporting Material for** John H. Boyle, Dino Martins, Paul M. Musili, Naomi E. Pierce (2019)Population Genomics and Demographic Sampling of the Ant-Plant *Vachellia drepanolobium* and Its Symbiotic Ants From Sites Across Its Range in East Africa.

**File S1: Sample sizes and climate data for each site in the population genetic survey**

**Table S1.1:** Sample sizes for population genomic analyses

|  |  |
| --- | --- |
| Site | Samples genotyped from each site for: |
| *V. drepanolobium*  | *C. mimosae* | *C. nigriceps* | *T. penzigi* |
| Suyian Ranch | 17 (12) | 15 (5) | 13 (5) | 15 (12) |
| Mpala R.C. | 15 (8) | 11 (4) | 13 (3) | 16 (12) |
| Kitengela | 10 (5) | 6 (4) | 12 (7) | 12 (6) |
| Ngong Hills1 | 23 (5) | 10 (8) | 9 (8) | 12 (6) |
| Isinya Road1 | 5 (0) | 7 (3) | 9 (7) | 5 (4) |
| Koriema | 18 (2) | 0 (0) | 9 (7) | 5 (3) |
| Mogotio | 16 (5) | 0 (0) | 14 (12) | 8 (7) |
| Mogotio-Koriema Road1 | 0 (0) | 0 (0) | 4 (3) | 3 (0) |
| Gilgil | 9 (0) | 0 (0) | 0 (0) | 5 (3) |
| Lake Naivasha2 | 10 (0) | 0 (0) | 0 (0) | 14 (4) |
| South Lake Road 12 | 0 (0) | 0 (0) | 6 (6) | 5 (1) |
| South Lake Road 22 | 0 (0) | 0 (0) | 4 (4) | 5 (4) |
| South Lake Road 32 | 2 (0) | 0 (0) | 6 (6) | 0 (0) |
| South Lake Road 42 | 4 (0) | 1 (1) | 6 (5) | 2 (0) |

Sample sizes for each site are those for Sample Set 1 and 2, with the sample size for Set 3 in parentheses. Samples sizes for *V. drepanolobium* represent individual trees. Samples sizes for the three ant species represent single workers, each taken from a different tree.

1 At Ngong Hills, Isinya Road, and the Mogotio-Koriema road, we sampled from separate sites within a short distance of each other. Since these transects had the same WorldClim data, we treated them as a single unit.
2 At Lake Naivasha, only *T. penzigi* was found at the site we surveyed. However, we found *T. penzigi*, *C. nigriceps*, and a single colony of *C. mimosae* at sites nearby along the South Lake Road. Since these sites were fairly close together, we treated them as a single site for most of the population genetic analyses. However, several of the South Lake Road transects had different WorldClim data, so we treated them as different sites for the purposes of distinguishing isolation-by-distance from isolation-by-environment using Multiple Matrix Regression with Randomization.

**File S2: Choosing stacks parameters.**

When building the Stacks catalogs of loci, we allowed two parameters to vary: the number of mismatches allowed between loci when processing an individual (-M parameter), and the number of mismatches allowed between loci when building the catalog (-n parameter). To present the outcomes, we use the individuals and run parameters of Data Set 1 to call SNPs. We then calculated average heterozygosity of the SNPs using the adegenet package version 2.0.1 (Jombart 2008, Jombart and Ahmed 2011) in R version 3.2.3 (R Core Team 2015).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | -M | -n | SNPs | Missing data | Heterozygosity |
| *V.* *drepanolobium* | 1 | 1 | 792 | 41% | 0.04 |
| *V.* *drepanolobium* | 2 | 1 | 872 | 41% | 0.04 |
| *V,* *drepanolobium* | 2 | 2 | 1027 | 41% | 0.04 |
| *V.* *drepanolobium* | 2 | 5 | 958 | 41% | 0.04 |
| *V.* *drepanolobium* | 5 | 1 | 885 | 41% | 0.06 |
| *C. mimosae* | 1 | 1 | 974 | 28% | 0.10 |
| *C. mimosae* | 2 | 1 | 1069 | 28% | 0.12 |
| *C. mimosae* | 2 | 2 | 1122 | 28% | 0.12 |
| *C. mimosae* | 2 | 5 | 1111 | 28% | 0.11 |
| *C. mimosae* | 5 | 1 | 1085 | 28% | 0.12 |
| *C. nigriceps* | 1 | 1 | 921 | 11% | 0.09 |
| *C. nigriceps* | 2 | 1 | 1038 | 10% | 0.09 |
| *C. nigriceps* | 2 | 2 | 1086 | 10% | 0.09 |
| *C. nigriceps* | 2 | 5 | 1068 | 10% | 0.09 |
| *C. nigriceps* | 5 | 1 | 1067 | 10% | 0.10 |
| *T. penzigi* | 1 | 1 | 1028 | 19% | 0.07 |
| *T. penzigi* | 2 | 1 | 1043 | 19% | 0.08 |
| *T. penzigi* | 2 | 2 | 1084 | 19% | 0.08 |
| *T. penzigi* | 2 | 5 | 1050 | 19% | 0.07 |
| *T. penzigi* | 5 | 1 | 1017 | 19% | 0.08 |

*References for File S2:*

Jombart, T. (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24:1403-1405.

Jombart, T., Ahmed, I. (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27:3070-3071.

R Core Team (2015) R: A language and environment for statistical computing. Vienna, Austria. <http://www.R-project.org>

**File S3: Our results are not biased by missing data.**

*Distribution of missing data*

Individuals varied widely in the amount missing data: most individuals had low levels of missing data, while a few individuals had considerably larger amounts (Figure S3.1). Fortunately, the individuals with larger amounts of missing data were quite evenly spread across populations; i.e., our results were not biased because some populations had disproportionate numbers of individuals with high missing data. To further test whether our results were affected by missing data, we also analyzed three data sets in parallel, each of which had different degrees of missing data (including no missing data in the final data set).

*DNA sequence alignment and base-calling*

As seen in Table S3.1, we successfully genotyped individuals of each species at between 120 and 3763 loci, depending on the species and the data set.

**Table S3.1:** results of RADseq sequencing and base-calling

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Species | Data set | Stacks -r 1 | Individuals | Loci | Missing data |
| *V. drepanolobium*  | 1 | 0.53 | 129 | 1027 | 41% |
| 2 | 0.6 | 129 | 400 | 35% |
| 3 | 1 | 37 | 120 | 0% |
| *C. mimosae* | 1 | 0.64 | 50 | 1122 | 28% |
| 2 | 0.6 | 50 | 1386 | 30% |
| 3 | 1 | 25 | 613 | 0% |
| *C. nigriceps* | 1 | 0.83 | 105 | 1086 | 10% |
| 2 | 0.6 | 105 | 3763 | 23% |
| 3 | 1 | 73 | 463 | 0% |
| *T. penzigi* | 1 | 0.71 | 107 | 1084 | 19% |
| 2 | 0.6 | 107 | 1655 | 24% |
| 3 | 1 | 62 | 231 | 0% |

1 The -r parameter in stacks denotes the minimum proportion of individuals in which a SNP must be found in order to be included in the data set.

*Population statistics*

The three data sets produced broadly similar results, despite differences in the individuals and SNPs making up each set (Tables S4.2, S4.3).

*Analysis of molecular variance*

As shown in Table S3.2, AMOVA tests revealed significant genetic variation partitioned both among our geographic populations (FCT), and also among sites within those populations (FSC). This was true for all species and all data sets.



**Figure S3.1:** Levels of missing data vary across individuals. Histogram bars are colored according to the number of individuals in that class from each site. MKR stands for Mogotio-Koriema Road. Although individuals vary in their degree of missing data, individuals with large amounts of missing data are mostly evenly distributed across sites.

**Table S3.2:** Population genetic summary statistics for each species.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *V. drepanolobium*  | *C. mimosae* | *C. nigriceps* | *T. penzigi* |
| Stat.1 | Data set | NE | SE | NW | SW | NE | SE | NE | SE | NW | SW | NE | SE | NW | SW |
| HO | 1 | 0.0024 | 0.0013 | 0.0009 | 0.0007 | 0.0018 | 0.0020 | 0.0014 | 0.0026 | 0.0011 | 0.0018 | 0.0011 | 0.0014 | 0.0005 | 0.0015 |
| 2 | 0.0021 | 0.0012 | 0.0009 | 0.0006 | 0.0019 | 0.0021 | 0.0014 | 0.0026 | 0.0012 | 0.0020 | 0.0013 | 0.0014 | 0.0006 | 0.0015 |
| 3 | 0.0018 | 0.0012 | 0.0010 | \* | 0.0016 | 0.0015 | 0.0011 | 0.0021 | 0.0009 | 0.0014 | 0.0006 | 0.0018 | 0.0005 | 0.0016 |
| HE | 1 | 0.0032 | 0.0036 | 0.0031 | 0.0039 | 0.0028 | 0.0030 | 0.0018 | 0.0029 | 0.0014 | 0.0021 | 0.0013 | 0.0024 | 0.0007 | 0.0028 |
| 2 | 0.0026 | 0.0031 | 0.0027 | 0.0032 | 0.0031 | 0.0032 | 0.0018 | 0.0032 | 0.0016 | 0.0024 | 0.0014 | 0.0024 | 0.0007 | 0.0029 |
| 3 | 0.0021 | 0.0021 | 0.0016 | \* | 0.0018 | 0.0017 | 0.0011 | 0.0023 | 0.0010 | 0.0015 | 0.0006 | 0.0021 | 0.0005 | 0.0024 |
| θπ | 1 | 0.0033 | 0.0037 | 0.0032 | 0.0041 | 0.0029 | 0.0030 | 0.0018 | 0.0030 | 0.0014 | 0.0022 | 0.0013 | 0.0025 | 0.0007 | 0.0029 |
| 2 | 0.0027 | 0.0031 | 0.0027 | 0.0034 | 0.0031 | 0.0033 | 0.0019 | 0.0033 | 0.0016 | 0.0025 | 0.0014 | 0.0025 | 0.0008 | 0.0030 |
| 3 | 0.0022 | 0.0022 | 0.0017 | \* | 0.0019 | 0.0018 | 0.0012 | 0.0023 | 0.0010 | 0.0015 | 0.0007 | 0.0021 | 0.0005 | 0.0025 |
| FSC | 1 | 0.104 \*\* | 0.125 \*\* | 0.031 \*\* | 0.227 \*\* |
| 2 | 0.097 \*\* | 0.126 \*\* | 0.0120 \*\* | 0.235 \*\* |
| 3 | 0.083 \*\* (*p* = 0.001) | 0.203 \*\* | 0.017 \*\* | 0.237 \*\* |
| FCT | 1 | 0.154 \*\* | 0.108 \*\* | 0.538 \*\* | 0.482 \*\* |
| 2 | 0.137 \*\* | 0.101 \*\* | 0.478 \*\* (*p* < 0.001) | 0.482 \*\* |
| 3 | 0.212 \*\* | 0.089 \*\* | 0.512 \*\* | 0.546 \*\* |

1 FSC corresponds to the proportion of genetic variation within each of the four geographically-determined populations which is partitioned among sites; FCT corresponds to the proportion of total genetic variation partitioned among the four populations. After each is a *p-*value from an AMOVA indicating whether the partitioning of genetic variation is significantly greater than zero.
\* No *V. drepanolobium* individuals from the SW population had high enough coverage to be included in data set 3.

\*\* All FSC and FCT values were significantly greater than zero, with *p* < 0.00001, except where noted.

**Table S3.3:** Pairwise FST between populations for each species

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Comparison | Data Set | *V. drepanolobium*  | *C. mimosae* | *C. nigriceps* | *T. penzigi* |
| NE-SE | 1 | 0.057 | 0.150 | 0.461 | 0.397 |
| 2 | 0.038 | 0.139 | 0.354 | 0.400 |
| 3 | 0.063 | 0.156 | 0.467 | 0.398 |
| NE-SW | 1 | -0.008 |  | 0.406 | 0.629 |
| 2 | -0.022 | 0.197 | 0.638 |
| 3 | \* | 0.417 | 0.684 |
| NE-NW | 1 | 0.345 | 0.720 | 0.722 |
| 2 | 0.321 | 0.651 | 0.711 |
| 3 | 0.320 | 0.743 | 0.819 |
| SE-SW | 1 | -0.011 | 0.308 | 0.337 |
| 2 | -0.021 | 0.284 | 0.347 |
| 3 | \* | 0.280 | 0.325 |
| SE-NW | 1 | 0.276 | 0.678 | 0.581 |
| 2 | 0.251 | 0.658 | 0.583 |
| 3 | 0.387 | 0.660 | 0.616 |
| SW-NW | 1 | 0.205 | 0.521 | 0.694 |
| 2 | 0.182 | 0.498 | 0.703 |
| 3 | \* | 0.493 | 0.731 |

\* None of the *V. drepanolobium* individuals from the SW population had high enough coverage to be included in data set 3, and thus there are no pairwise comparison between the SW population and the others.

*Genetic clustering analysis*

For *V. drepanolobium*, Admixture cross-validation identified the best number of clusters to be 2 for all three data sets (Figure S3.2). All three data sets show separation between the sites in the Northern Rift and the sites in the Northern Highlands, with all the individuals in the Northern Highlands having ancestry from one cluster, and all of the individuals from the Northern Rift having ancestry from a second cluster. However, in the Southern Highlands and Southern Rift sites, individuals with ancestry from both clusters are common (Figure S3.3).

 For *C. mimosae*, cross-validation indicated that the best number of clusters was 2 for data set 1, and 1 cluster for data sets 2 and 3, although the difference in cross-validation error between *k* = 1 and *k* = 2 was small for all three data sets (Figure S3.2). We present the results for data sets 2 and 3 using *k* = 2 in Figure S3.3 in order to show that the clustering pattern was similar across data sets for this *k* value.

*C. nigriceps* and *T. penzigi* show much stronger associations between genetic clustering and geography. Across all data sets for both species, cross-validation indicated that the best number of clusters was 4, and each of these clusters dominated a single region, and was rarely, if ever, found outside that region. The exception *T. penzigi*, ancestry from both genetic clusters was found in the Ngong Hills site in all three data sets.

We found no evidence that trees assigned to different genetic clusters were associated with particular ant occupants (multinomial exact tests, *p* > 0.1 for all); this equally the case for all three data sets.



**Figure S3.2:** Admixture cross-validation produced similar results across data sets. For *V. drepanolobium*, *C. nigriceps*, and *T. penzigi*, the same number of clusters was chosen across all three data sets. For *C. mimosae*, the best cluster number was 2 for data set 1, and 1 for data sets 2 and 3.



**Figure S3.3:** Admixture results were similar across the three data sets. Each vertical bar represents a single individual tree or ant, colored according to its proportion of ancestry in a particular cluster. The horizontal position of bars within sites for data set 3 does not correspond to the positions for data sets 1 and 2. For *C. mimosae*, cross-validation indicated that the optimal number of clusters was 1 in data sets 2 and 3; we present the results for *k* = 2 for those two data sets here instead. 1: Mpala RC, 2: Suyian, 3: Isinya Road, 4: Kitengela, 5: Ngong Hills, 6: Lake Naivasha, 7: Gilgil, 8: South Lake Road sites (combined), 9: Koriema, 10: Mogotio-Koriema Road, 11: Mogotio.

*Isolation by distance and environment*

For both *V. drepanolobium* and *C. mimosae*, we found no evidence of isolation by distance or isolation by environment in any of the data sets (all *p*-values > 0.05). In the case of *C. nigriceps*, we found evidence of both isolation by distance and isolation by environment (all *p-*values < 0.05), except for data set 3, in which we only found evidence for isolation by distance. For *T. penzigi*, in data sets 1 and 2, we found evidence of isolation by distance (*p* < 0.01 in both), but no evidence of isolation by environment (*p* > 0.05 in both). We no evidence of isolation by environment (*p* = 0.06) or distance (*p* =0.2) in data set 3.