evaluating effects of missing data threshold and population assignment on PCA, t-SNE, and dxy

In [1]:
```
import ipyrad.analysis as ipa
import toyplot
import ipyparallel as ipp
import pandas as pd
import numpy as np
from sklearn.manifold import TSNE
```

In [2]:
```
# the path to your HDF5 formatted snps file
data = "../plate1/P_ni_6rm_v9_outfiles/P_ni_6rm_v9.snps.hdf5"
```

Let's define populations Western Inambari (Inam), Purus-Madeira (Puru), Jiparana-Guapore (JiGu), Jiparaná-Roosevelt (Mach), Roosevelt-Aripuana (Roar), Aripuana-Sucunduri (ArSu), Sucunduri-Tapajos (SuTa), and Western Para (Para).

These assignments are made only to color points in the PCA analysis below.

In [3]:
```
imap = {
#"ref": ["reference"],
"Inam": ["P_ni_A7862_In", "P_ni_A7911_In", "P_ni_A7928_In", "P_ni_T6243_
In",],
"Puru": ["P_ni_T5850_pu", "P_ni_T5940_pu", "P_ni_T5974_pu", "P_ni_T15938
_pu","P_ni_80034_pu", "P_ni_T3609_pu", "P_ni_T3611_pu",  "P_ni_T3817_pu"
, "P_ni_T4043_pu", "P_ni_T4051_pu", "P_ni_T4313_pu", "P_ni_T4404_pu"],
"JiGu": ["P_ni_T22153_jigu", "P_ni_T3261_jigu","P_ni_T15863_jigu", "P_ni
_T15868_jigu", "P_ni_T15871_jigu", "P_ni_A3255_jigu",],
"Mach": ["P_ni_T443_ma", "P_ni_T467_ma","P_ni_T369_ma","P_ni_J434_ma",
"P_ni_J461_ma", "P_ni_J462_ma", "P_ni_J485_ma", "P_ni_J210_ma", "P_ni_J2
27_ma", "P_ni_J260_ma", "P_ni_A2418_ma",  "P_ni_A542_ma",],
"Roar": ["P_ni_J684_roar", "P_ni_J724_roar","P_ni_J361_roar", "P_ni_J363
_roar", "P_ni_J371_roar", "P_ni_J373_roar", "P_ni_J381_roar", "P_ni_J385
_roar", "P_ni_J389_roar", "P_ni_J417_roar",],
"ArSu": ["P_ni_J551_arsu", "P_ni_J602_arsu", "P_ni_J603_arsu", "P_ni_J61
4_arsu", "P_ni_J617_arsu","P_ni_80555_arsu","P_ni_86072_arsu", "P_ni_806
84_arsu", "P_ni_80802_arsu", "P_ni_80874_arsu", "P_ni_85430_arsu",],
"SuTa": ["P_ni_T14543_suta",  "P_ni_T9076_suta","P_ni_T16698_suta","P_ni
_T10967_suta",  "P_ni_T11888_suta","P_ni_T10204_suta","P_ni_A15120_suta"
, "P_ni_77876_suta", "P_ni_78155_suta","P_ni_85721_suta",],
"Para": ["P_ni_T1642_pa", "P_ni_T18703_pa","P_ni_T12345_pa", "P_ni_T1285
4_pa","P_ni_T11193_pa", "P_ni_T11222_pa","P_ni_T10673_pa", "P_ni_T10940_
pa","P_ni_A7066_pa", "P_ni_A14342_pa", "P_ni_A15277_pa",]
}
# minimum % of samples that must be present in each SNP from each popula
tion: do 5 subsets to see robustness of impu
#Because there is a lot of missing data, we want to make sure that resul
ts are not biased due to imputing too many haplotypes
#sampling across different minmaps will help address this. If the data a
re more organized by population with more missing data, then this could
 be bias in the imputation from small sample sizes
minmap1 = {i: 0.5 for i in imap}
minmap2 = {i: 0.65 for i in imap}
minmap3 = {i: 0.75 for i in imap}
minmap4 = {i: 0.85 for i in imap}
minmap5 = {i: 0.95 for i in imap}
```

let's run PCA for different values of minmap to see how varying missing data affects our results

note we are assuming K=8 which is likely high, but k-means clustering will cluster samples independent of a priori geographic assignment

Here we are using k-means clustering to assign individuals to populations independently of our a priori geographic assignments in order to circumscribe populations from which to impute missing haplotypes.

The PCA plots can then be colored based on our geographic assignments above

In [4]:
```
#init pca object with input data and (optional) parameter options

pca = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap1,
    mincov=0.85,
    impute_method=8,
)
pca2 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap2,
    mincov=0.85,
    impute_method=8,
)
pca3 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap3,
    mincov=0.85,
    impute_method=8,
)
pca4 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap4,
    mincov=0.85,
    impute_method=8,
)
pca5 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap5,
    mincov=0.85,
    impute_method=8,
)
```

```
Kmeans clustering: iter=0, K=8, mincov=0.9, minmap={'global': 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1057178
Filtered (minmap): 991306
Filtered (combined): 1061426
Sites after filtering: 186262
Sites containing missing values: 166987 (89.65%)
Missing values in SNP matrix: 610212 (4.31%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 6.5%, 37.0%
{0: ['P_ni_T15938_pu', 'P_ni_T3611_pu', 'P_ni_T4051_pu', 'P_ni_T5974_p
u'], 1: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni
_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu',
'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_ar
su', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_a
rsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni
_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 2: ['P_ni_A14342
_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940
_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T1285
4_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_ni_T4043_pu', 'P_ni_T
5850_pu'], 4: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_n
i_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J361_roar', 'P_ni_J36
3_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J3
85_roar', 'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J434_ma', 'P_ni_J46
1_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_J724_roa
r', 'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T
22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T
467_ma'], 5: ['P_ni_T3609_pu', 'P_ni_T3817_pu', 'P_ni_T4313_pu', 'P_ni_
T4404_pu', 'P_ni_T5940_pu', 'P_ni_T6243_In'], 6: ['P_ni_A7862_In', 'P_n
i_A7911_In', 'P_ni_A7928_In'], 7: ['P_ni_80034_pu']}

Kmeans clustering: iter=1, K=8, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1039648
Filtered (minmap): 1170664
Filtered (combined): 1172205
Sites after filtering: 75483
Sites containing missing values: 56208 (74.46%)
Missing values in SNP matrix: 119740 (2.09%)
Imputation: 'sampled'; (0, 1, 2) = 57.7%, 4.6%, 37.7%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu',
'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu',
'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T1454
3_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_A7862_In',
'P_ni_A7911_In', 'P_ni_A7928_In'], 2: ['P_ni_A14342_pa', 'P_ni_A15277_p
a', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_p
a', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_p
a', 'P_ni_T18703_pa'], 3: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A5
42_ma', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J434_ma',
```

```
'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T15863_jigu', 'P_
ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_ji
gu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 4: ['P_ni_J361_ro
ar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_r
oar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J684_
roar', 'P_ni_J724_roar'], 5: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_
T3609_pu', 'P_ni_T3611_pu', 'P_ni_T3817_pu', 'P_ni_T4313_pu', 'P_ni_T44
04_pu', 'P_ni_T5974_pu', 'P_ni_T6243_In'], 6: ['P_ni_T4051_pu'], 7: ['P
_ni_T4043_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu']}

Kmeans clustering: iter=2, K=8, mincov=0.875, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1023011
Filtered (minmap): 1155616
Filtered (combined): 1157513
Sites after filtering: 90175
Sites containing missing values: 70900 (78.62%)
Missing values in SNP matrix: 162362 (2.37%)
Imputation: 'sampled'; (0, 1, 2) = 57.6%, 4.7%, 37.6%
{0: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_m
a', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J361_roar', 'P_ni_J363_roar',
'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar',
'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_
ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_J724_roar', 'P_ni_
T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jig
u', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'],
1: ['P_ni_80034_pu', 'P_ni_T15938_pu'], 2: ['P_ni_A14342_pa', 'P_ni_A15
277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11
193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1
642_pa', 'P_ni_T18703_pa'], 3: ['P_ni_T3609_pu', 'P_ni_T4043_pu', 'P_ni
_T4404_pu', 'P_ni_T5974_pu'], 4: ['P_ni_77876_suta', 'P_ni_78155_suta',
'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_ar
su', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15
120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_
J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta',
'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076
_suta'], 5: ['P_ni_T5850_pu', 'P_ni_T5940_pu'], 6: ['P_ni_A7862_In', 'P
_ni_A7911_In', 'P_ni_A7928_In'], 7: ['P_ni_T3611_pu', 'P_ni_T3817_pu',
'P_ni_T4051_pu', 'P_ni_T4313_pu', 'P_ni_T6243_In']}

Kmeans clustering: iter=3, K=8, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1007137
Filtered (minmap): 1170664
Filtered (combined): 1172205
Sites after filtering: 75483
Sites containing missing values: 56208 (74.46%)
Missing values in SNP matrix: 119740 (2.09%)
Imputation: 'sampled'; (0, 1, 2) = 57.7%, 4.6%, 37.8%
```

```
{0: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_
pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345
_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 1: ['P_ni_7
7876_suta', 'P_ni_86072_arsu', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_n
i_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T16698_suta'],
2: ['P_ni_80034_pu', 'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T3817_pu',
'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu', 'P_ni_T5850_pu', 'P_
ni_T5940_pu', 'P_ni_T5974_pu', 'P_ni_T6243_In'], 3: ['P_ni_A2418_ma',
'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_n
i_J260_ma', 'P_ni_J389_roar', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J46
2_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_J724_roar', 'P_ni_T15863
_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_
ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 4: ['P
_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_ars
u', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_A151
20_suta', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta',
'P_ni_T14543_suta', 'P_ni_T9076_suta'], 5: ['P_ni_A7862_In', 'P_ni_A791
1_In', 'P_ni_A7928_In'], 6: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_
J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni
_J417_roar'], 7: ['P_ni_T15938_pu', 'P_ni_T4404_pu']}

Kmeans clustering: iter=4, K=8, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 1153340
Filtered (combined): 1155296
Sites after filtering: 92392
Sites containing missing values: 73117 (79.14%)
Missing values in SNP matrix: 158227 (2.25%)
Imputation: 'sampled'; (0, 1, 2) = 57.4%, 4.9%, 37.7%
Kmeans clustering: iter=0, K=8, mincov=0.9, minmap={'global': 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1057178
Filtered (minmap): 991306
Filtered (combined): 1061426
Sites after filtering: 186262
Sites containing missing values: 166987 (89.65%)
Missing values in SNP matrix: 610212 (4.31%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 6.6%, 37.0%
{0: ['P_ni_80034_pu', 'P_ni_T3611_pu', 'P_ni_T3817_pu', 'P_ni_T4043_p
u', 'P_ni_T4313_pu', 'P_ni_T4404_pu', 'P_ni_T5940_pu', 'P_ni_T5974_p
u'], 1: ['P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227_
ma', 'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P
_ni_J485_ma', 'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jig
u', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_m
a', 'P_ni_T467_ma'], 2: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A706
6_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T112
22_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T187
03_pa'], 3: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu',
'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_ar
```

```
su', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J5
51_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J
617_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta',
'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 4: ['P_ni_A
2418_ma', 'P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J
373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_ni_
J417_roar', 'P_ni_J684_roar', 'P_ni_J724_roar'], 5: ['P_ni_A7862_In',
'P_ni_A7911_In', 'P_ni_A7928_In', 'P_ni_T3609_pu', 'P_ni_T4051_pu', 'P_
ni_T6243_In'], 6: ['P_ni_T5850_pu'], 7: ['P_ni_T15938_pu']}

Kmeans clustering: iter=1, K=8, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1039648
Filtered (minmap): 1163456
Filtered (combined): 1165194
Sites after filtering: 82494
Sites containing missing values: 63219 (76.63%)
Missing values in SNP matrix: 138141 (2.20%)
Imputation: 'sampled'; (0, 1, 2) = 58.0%, 4.8%, 37.1%
{0: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_r
oar', 'P_ni_J381_roar', 'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J724_
roar'], 1: ['P_ni_80034_pu', 'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T4
313_pu', 'P_ni_T5850_pu'], 2: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_n
i_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_n
i_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_n
i_T18703_pa'], 3: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_ar
su', 'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_854
30_arsu', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_
ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P
_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_su
ta', 'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 4: ['P
_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In', 'P_ni_T4043_pu', 'P_ni
_T4051_pu', 'P_ni_T4404_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu'], 5: ['P_
ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_
J227_ma', 'P_ni_J260_ma', 'P_ni_J385_roar', 'P_ni_J434_ma', 'P_ni_J461_
ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_T15863_jig
u', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T
3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 6: ['P_ni_
T15938_pu'], 7: ['P_ni_T3817_pu', 'P_ni_T6243_In']}

Kmeans clustering: iter=2, K=8, mincov=0.875, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1023011
Filtered (minmap): 1150350
Filtered (combined): 1152344
Sites after filtering: 95344
Sites containing missing values: 76069 (79.78%)
Missing values in SNP matrix: 185480 (2.56%)
Imputation: 'sampled'; (0, 1, 2) = 57.1%, 4.7%, 38.2%
```

```
{0: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_
pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345
_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 1: ['P_ni_A
2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227
_ma', 'P_ni_J260_ma', 'P_ni_J373_roar', 'P_ni_J389_roar', 'P_ni_J434_m
a', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar',
'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T2215
3_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_
ma'], 2: ['P_ni_T4043_pu', 'P_ni_T4313_pu'], 3: ['P_ni_77876_suta', 'P_
ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_ars
u', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_8607
2_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_
J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_
ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_s
uta', 'P_ni_T9076_suta'], 4: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni
_J371_roar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni_J417_roar', 'P_n
i_J724_roar'], 5: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In'],
6: ['P_ni_T15938_pu'], 7: ['P_ni_80034_pu', 'P_ni_T3609_pu', 'P_ni_T361
1_pu', 'P_ni_T3817_pu', 'P_ni_T4051_pu', 'P_ni_T4404_pu', 'P_ni_T5850_p
u', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_ni_T6243_In']}

Kmeans clustering: iter=3, K=8, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1007137
Filtered (minmap): 1148217
Filtered (combined): 1150280
Sites after filtering: 97408
Sites containing missing values: 78133 (80.21%)
Missing values in SNP matrix: 188332 (2.54%)
Imputation: 'sampled'; (0, 1, 2) = 57.7%, 5.0%, 37.3%
{0: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_m
a', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J385_roar', 'P_ni_J434_ma',
'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J724_roar', 'P_ni
_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jig
u', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 1: ['P_ni_A7911_I
n'], 2: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10
673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T1
2345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_
ni_T3611_pu', 'P_ni_T4051_pu', 'P_ni_T6243_In'], 4: ['P_ni_77876_suta',
'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_ar
su', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_860
72_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni
_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P
_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_
suta', 'P_ni_T9076_suta'], 5: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_n
i_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J389_roar', 'P_
ni_J417_roar', 'P_ni_J684_roar', 'P_ni_T15863_jigu'], 6: ['P_ni_80034_p
u', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3817_pu', 'P_ni_T4043_p
u', 'P_ni_T4313_pu', 'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu',
'P_ni_T5974_pu'], 7: ['P_ni_A7862_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=4, K=8, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
```

```
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 1145272
Filtered (combined): 1147377
Sites after filtering: 100311
Sites containing missing values: 81036 (80.78%)
Missing values in SNP matrix: 195454 (2.56%)
Imputation: 'sampled'; (0, 1, 2) = 57.9%, 4.9%, 37.3%
Kmeans clustering: iter=0, K=8, mincov=0.9, minmap={'global': 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1057178
Filtered (minmap): 991306
Filtered (combined): 1061426
Sites after filtering: 186262
Sites containing missing values: 166987 (89.65%)
Missing values in SNP matrix: 610212 (4.31%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 6.5%, 37.0%
{0: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_m
a', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J385_roar', 'P_ni_J434_ma',
'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni
_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jig
u', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'],
1: ['P_ni_T3609_pu'], 2: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A70
66_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11
222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18
703_pa'], 3: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu',
'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_ar
su', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J5
51_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J
617_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta',
'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 4: ['P_ni_T
15938_pu', 'P_ni_T3611_pu', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T40
51_pu', 'P_ni_T4313_pu', 'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_
pu', 'P_ni_T5974_pu', 'P_ni_T6243_In'], 5: ['P_ni_A7862_In', 'P_ni_A791
1_In', 'P_ni_A7928_In'], 6: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_
J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J389_roar', 'P_ni
_J417_roar', 'P_ni_J724_roar'], 7: ['P_ni_80034_pu']}

Kmeans clustering: iter=1, K=8, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1039648
Filtered (minmap): 1148163
Filtered (combined): 1150281
Sites after filtering: 97407
Sites containing missing values: 78132 (80.21%)
Missing values in SNP matrix: 192653 (2.60%)
Imputation: 'sampled'; (0, 1, 2) = 57.3%, 4.9%, 37.8%
```

```
{0: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_m
a', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J361_roar', 'P_ni_J363_roar',
'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar',
'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_
ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_J724_roar', 'P_ni_
T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jig
u', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'],
1: ['P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T4043_pu'], 2: ['P_ni_A1434
2_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T1094
0_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T128
54_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_ni_77876_suta', 'P_n
i_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_arsu',
'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_86072_ar
su', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603
_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T
10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_sut
a', 'P_ni_T9076_suta'], 4: ['P_ni_80034_pu', 'P_ni_T3817_pu', 'P_ni_T44
04_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu'], 5: ['P_ni_T
4051_pu', 'P_ni_T6243_In'], 6: ['P_ni_T15938_pu'], 7: ['P_ni_A7862_In',
'P_ni_A7911_In', 'P_ni_A7928_In', 'P_ni_T4313_pu']}

Kmeans clustering: iter=2, K=8, mincov=0.875, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1023011
Filtered (minmap): 1170664
Filtered (combined): 1172205
Sites after filtering: 75483
Sites containing missing values: 56208 (74.46%)
Missing values in SNP matrix: 119740 (2.09%)
Imputation: 'sampled'; (0, 1, 2) = 57.6%, 4.6%, 37.8%
{0: ['P_ni_A7862_In', 'P_ni_T4051_pu'], 1: ['P_ni_77876_suta', 'P_ni_78
155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_
ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_86072_ars
u', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603_
arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T1
0967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_suta',
'P_ni_T9076_suta'], 2: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066
_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T1122
2_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T1870
3_pa'], 3: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J
210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J385_roar', 'P_ni_J434_m
a', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_J724_roar',
'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T2215
3_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_
ma'], 4: ['P_ni_T15938_pu'], 5: ['P_ni_A7911_In', 'P_ni_A7928_In', 'P_n
i_T3611_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_ni_T6243_In'], 6: ['P
_ni_80034_pu', 'P_ni_T3609_pu', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni
_T4313_pu', 'P_ni_T4404_pu', 'P_ni_T5850_pu'], 7: ['P_ni_J361_roar', 'P
_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar',
'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J461_ma']}

Kmeans clustering: iter=3, K=8, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
```

```
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1007137
Filtered (minmap): 1145516
Filtered (combined): 1147617
Sites after filtering: 100071
Sites containing missing values: 80796 (80.74%)
Missing values in SNP matrix: 199816 (2.63%)
Imputation: 'sampled'; (0, 1, 2) = 57.7%, 4.8%, 37.5%
{0: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_m
a', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_
ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P
_ni_T22153_jigu', 'P_ni_T369_ma', 'P_ni_T467_ma'], 1: ['P_ni_T3817_pu',
'P_ni_T4051_pu', 'P_ni_T5974_pu', 'P_ni_T6243_In'], 2: ['P_ni_A14342_p
a', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_p
a', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_
pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_ni_77876_suta', 'P_ni_7
8155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P
_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_86072_ars
u', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603_
arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T1
0967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_suta',
'P_ni_T9076_suta'], 4: ['P_ni_T15938_pu', 'P_ni_T4404_pu'], 5: ['P_ni_J
361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_
J381_roar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni
_J684_roar', 'P_ni_J724_roar', 'P_ni_T15871_jigu', 'P_ni_T3261_jigu',
'P_ni_T443_ma'], 6: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_I
n'], 7: ['P_ni_80034_pu', 'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T4043
_pu', 'P_ni_T4313_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu']}

Kmeans clustering: iter=4, K=8, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 1146141
Filtered (combined): 1148231
Sites after filtering: 99457
Sites containing missing values: 80182 (80.62%)
Missing values in SNP matrix: 193086 (2.55%)
Imputation: 'sampled'; (0, 1, 2) = 57.9%, 4.8%, 37.3%
Kmeans clustering: iter=0, K=8, mincov=0.9, minmap={'global': 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1057178
Filtered (minmap): 991306
Filtered (combined): 1061426
Sites after filtering: 186262
Sites containing missing values: 166987 (89.65%)
Missing values in SNP matrix: 610212 (4.31%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 6.6%, 37.0%
```

```
{0: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_r
oar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_ni_J417_
roar', 'P_ni_J684_roar', 'P_ni_J724_roar', 'P_ni_T15863_jigu'], 1: ['P_
ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_
ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P
_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 2: ['P_ni_T3817_p
u', 'P_ni_T4043_pu', 'P_ni_T4404_pu', 'P_ni_T5974_pu'], 3: ['P_ni_77876
_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_
80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta',
'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_ars
u', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_
suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_n
i_T16698_suta', 'P_ni_T9076_suta'], 4: ['P_ni_A2418_ma', 'P_ni_A3255_ji
gu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P
_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T15
868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu',
'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 5: ['P_ni_80034_pu',
'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu', 'P_
ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T6243_In'], 6: ['P_ni_A7862_In',
'P_ni_A7911_In', 'P_ni_A7928_In'], 7: ['P_ni_T15938_pu']}

Kmeans clustering: iter=1, K=8, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1039648
Filtered (minmap): 1150901
Filtered (combined): 1152900
Sites after filtering: 94788
Sites containing missing values: 75513 (79.67%)
Missing values in SNP matrix: 174651 (2.42%)
Imputation: 'sampled'; (0, 1, 2) = 57.9%, 4.8%, 37.3%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu',
'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu',
'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T1454
3_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_A7862_In',
'P_ni_A7911_In', 'P_ni_A7928_In'], 2: ['P_ni_A2418_ma', 'P_ni_A3255_jig
u', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J724_roar', 'P_ni_T15863_jig
u', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T
3261_jigu', 'P_ni_T443_ma', 'P_ni_T467_ma'], 3: ['P_ni_A14342_pa', 'P_n
i_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_n
i_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_
ni_T1642_pa', 'P_ni_T18703_pa'], 4: ['P_ni_80034_pu', 'P_ni_T3609_pu',
'P_ni_T3611_pu', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_
ni_T4313_pu', 'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_
T5974_pu', 'P_ni_T6243_In'], 5: ['P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_
J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma'], 6: ['P_ni_T1
5938_pu'], 7: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P
_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni_J389_roar',
'P_ni_J417_roar', 'P_ni_J684_roar', 'P_ni_T369_ma']}

Kmeans clustering: iter=2, K=8, mincov=0.875, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
```

```
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1023011
Filtered (minmap): 1151649
Filtered (combined): 1153601
Sites after filtering: 94087
Sites containing missing values: 74812 (79.51%)
Missing values in SNP matrix: 169318 (2.37%)
Imputation: 'sampled'; (0, 1, 2) = 59.5%, 4.9%, 35.6%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4313_pu', 'P_ni_T4404_pu', 'P_ni_T5940_pu',
'P_ni_T6243_In'], 1: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_m
a', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J389_roar',
'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J
684_roar', 'P_ni_J724_roar', 'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P
_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_m
a', 'P_ni_T443_ma', 'P_ni_T467_ma'], 2: ['P_ni_A14342_pa', 'P_ni_A15277
_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193
_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642
_pa', 'P_ni_T18703_pa'], 3: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_n
i_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu',
'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_s
uta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_
arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_
T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076_sut
a'], 4: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In'], 5: ['P_ni_
T4043_pu', 'P_ni_T5974_pu'], 6: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P
_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar',
'P_ni_J417_roar'], 7: ['P_ni_T4051_pu', 'P_ni_T5850_pu']}

Kmeans clustering: iter=3, K=8, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 1007137
Filtered (minmap): 1153479
Filtered (combined): 1155425
Sites after filtering: 92263
Sites containing missing values: 72988 (79.11%)
Missing values in SNP matrix: 173389 (2.47%)
Imputation: 'sampled'; (0, 1, 2) = 57.6%, 4.8%, 37.6%
{0: ['P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T3817_pu', 'P_ni_T4051_p
u', 'P_ni_T4313_pu', 'P_ni_T4404_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu',
'P_ni_T6243_In'], 1: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_m
a', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J389_roar',
'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T
15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jig
u', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'],
2: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_p
a', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_
pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_ni_77
876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_
ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_sut
```

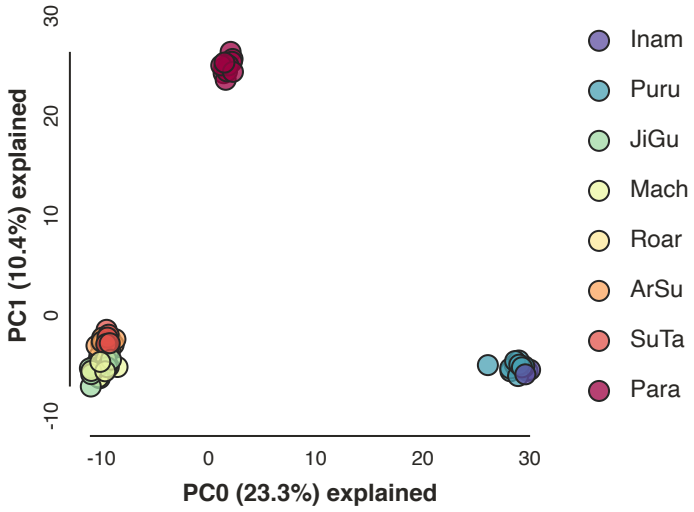
```
a', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602
_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10
204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta',
'P_ni_T16698_suta', 'P_ni_T9076_suta'], 4: ['P_ni_J361_roar', 'P_ni_J36
3_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J3
85_roar', 'P_ni_J417_roar', 'P_ni_J684_roar', 'P_ni_J724_roar'], 5: ['P
_ni_80034_pu', 'P_ni_T4043_pu', 'P_ni_T5850_pu'], 6: ['P_ni_T15938_p
u'], 7: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=4, K=8, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85, 5: 0.85, 6: 0.85, 7: 0.85}
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 1156545
Filtered (combined): 1158437
Sites after filtering: 89251
Sites containing missing values: 69976 (78.40%)
Missing values in SNP matrix: 158856 (2.34%)
Imputation: 'sampled'; (0, 1, 2) = 57.7%, 4.8%, 37.5%
```

In [5]:
```
# # run the PCA analysis
pca.run()
pca2.run()
pca3.run()
pca4.run()
pca5.run()
```

```
Subsampling SNPs: 14714/92392
Subsampling SNPs: 15779/100311
Subsampling SNPs: 15638/99457
Subsampling SNPs: 14291/89251
Subsampling SNPs: 3623/19275
```

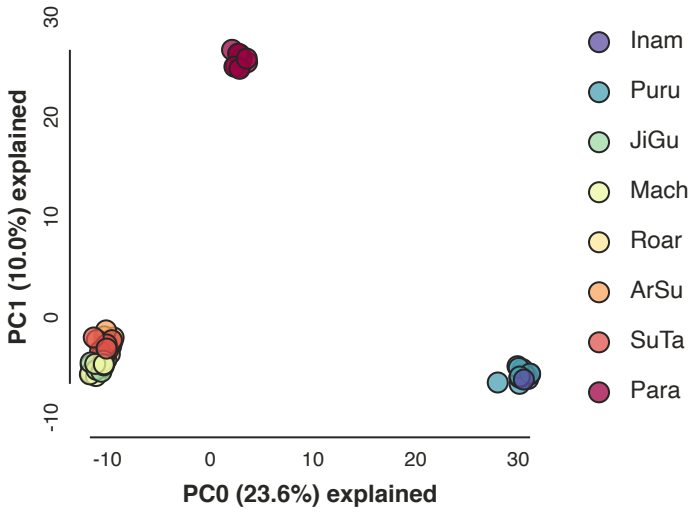
In [6]:
```
pca.draw()
```

Out[6]:
```
(<toyplot.canvas.Canvas at 0x2b8b942b7c50>,
 <toyplot.coordinates.Cartesian at 0x2b8b93bcb190>)
```



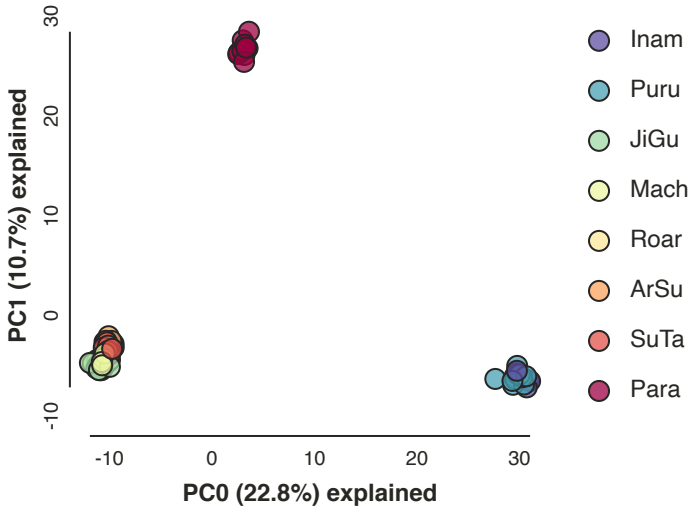
In [7]:
```
pca2.draw()
```

Out[7]:
```
(<toyplot.canvas.Canvas at 0x2b8b943a2750>,
 <toyplot.coordinates.Cartesian at 0x2b8b943a27d0>)
```



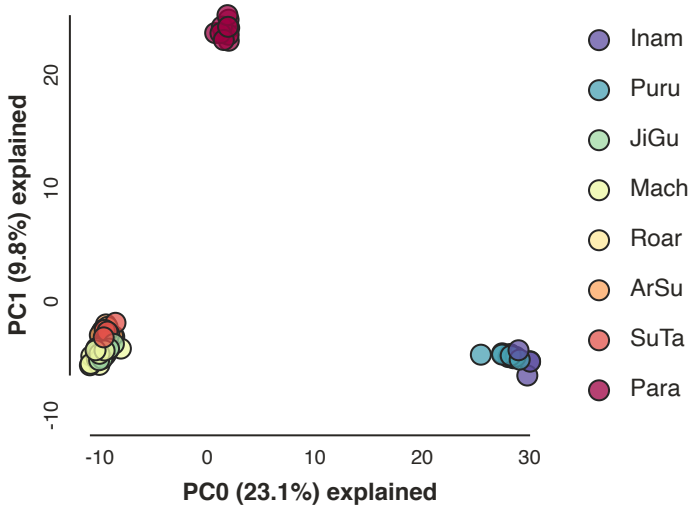
In [8]:
```
pca3.draw()
```

Out[8]:
```
(<toyplot.canvas.Canvas at 0x2b8b9493ffd0>,
 <toyplot.coordinates.Cartesian at 0x2b8b93a6ab90>)
```



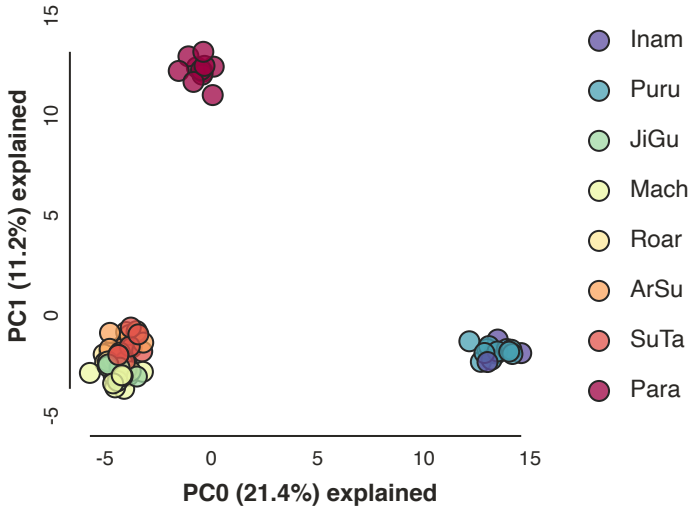
In [9]:
```
pca4.draw()
```

Out[9]:
```
(<toyplot.canvas.Canvas at 0x2b8b93a6ac50>,
 <toyplot.coordinates.Cartesian at 0x2b8b93eed710>)
```



In [10]:
```
pca5.draw()
```

Out[10]:
```
(<toyplot.canvas.Canvas at 0x2b8b93d02fd0>,
 <toyplot.coordinates.Cartesian at 0x2b8b93d02cd0>)
```



As you can see, for varying degrees of missing data, we generally get the same or similar results.

there are three to five clusters of points here that correspond pretty clearly to many river barriers and these are consistent among runs.

Now we can write the PCA results to a file

In [11]:
```
# # store the PC axes as a dataframe
df4 = pd.DataFrame(pca4.pcaxes[0], index=pca4.names)

# # write the PC axes to a CSV file
df4.to_csv("P_ni_pca_85minmap_12Jan2022.csv")

# # show the first ten samples and the first 10 PC axes
df4.iloc[:10, :10].round(2)
```

Out[11]:

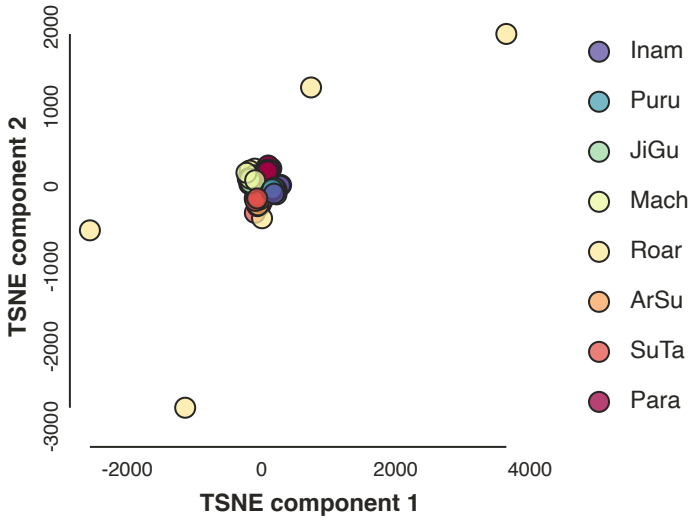
	0	1	2	3	4	5	6	7	8	9
P_ni_77876_suta	-9.92	-2.98	6.72	-1.46	-0.70	0.69	1.25	0.95	0.66	-1.11
P_ni_78155_suta	-8.94	-2.92	6.13	0.47	-1.23	0.71	-0.84	-0.14	1.33	-1.00
P_ni_80034_pu	28.49	-4.97	0.05	-4.44	3.70	9.14	18.91	-8.75	-7.13	-5.64
P_ni_80555_arsu	-9.44	-1.99	7.93	0.21	-1.61	0.07	1.39	0.51	-1.08	-0.07
P_ni_80684_arsu	-9.11	-2.73	9.12	-0.71	-1.00	-0.06	-0.54	2.00	0.22	0.58
P_ni_80802_arsu	-8.67	-3.07	10.27	-0.42	-0.19	-0.65	-1.99	1.30	1.31	-0.05
P_ni_80874_arsu	-9.88	-3.11	8.40	-0.46	-1.96	0.43	1.93	1.73	0.33	0.19
P_ni_85430_arsu	-9.83	-2.65	8.11	-0.32	-0.24	0.39	1.46	1.91	0.94	-0.03
P_ni_85721_suta	-9.82	-3.50	9.86	-0.00	-1.26	0.68	-0.24	0.48	-0.37	-0.40
P_ni_86072_arsu	-9.67	-2.81	8.81	-1.89	-2.32	0.28	1.72	0.16	1.03	-0.44

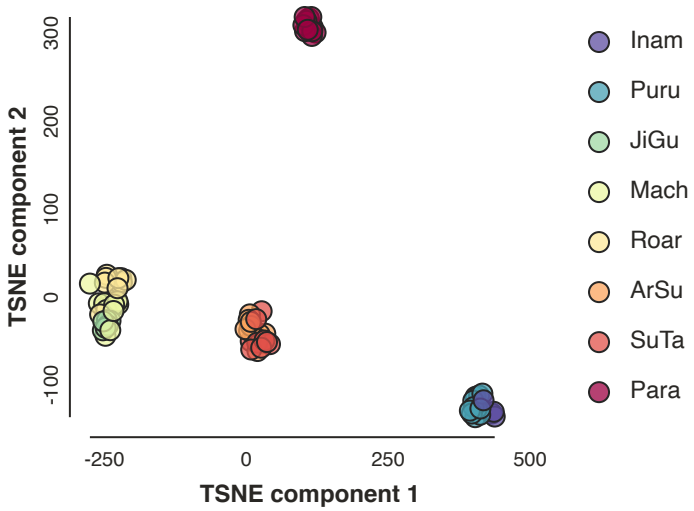
Now let's look at if and how t-SNE differs from PCA

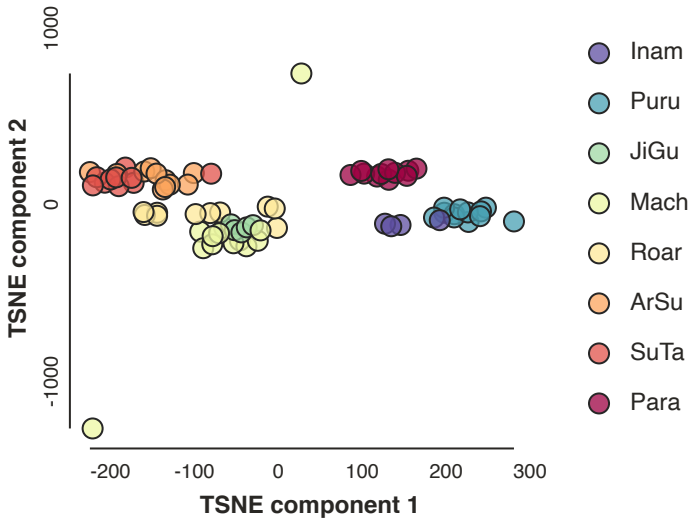
In [12]:
```
pca.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca2.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca3.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca4.run_tsne(subsample=True, perplexity=5.0, n_iter=10000000, seed=123)
pca5.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=223)
```

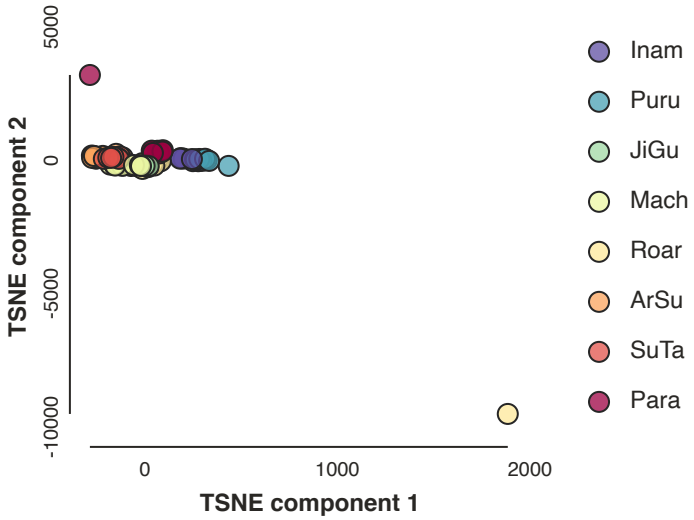
```
Subsampling SNPs: 14714/92392
Subsampling SNPs: 15779/100311
Subsampling SNPs: 15638/99457
Subsampling SNPs: 14291/89251
Subsampling SNPs: 3623/19275
```

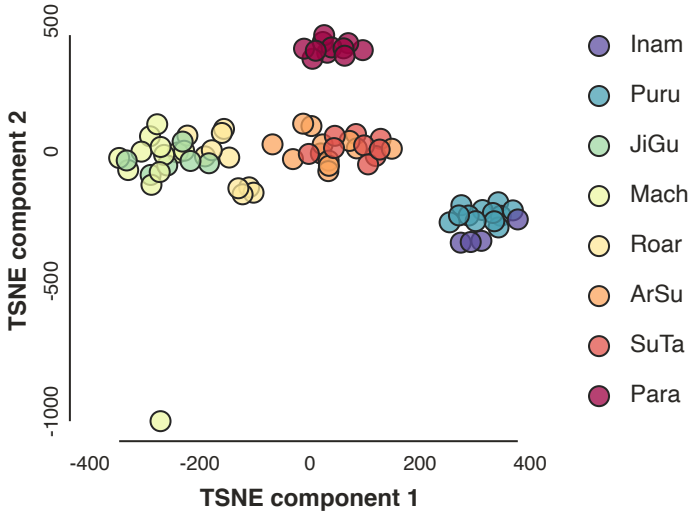
In [13]:
```
pca.draw();
pca2.draw();
pca3.draw();
pca4.draw();
pca5.draw();
```


TSNE component 2
2000
1000
0
-1000
-2000
-3000
-2000
0
2000
4000
TSNE component 1
Inam
Puru
JiGu
Mach
Roar
ArSu
SuTa
Para


TSNE component 2
300
200
100
0
-100
-250
0
250
500
TSNE component 1
Inam
Puru
JiGu
Mach
Roar
ArSu
SuTa
Para


TSNE component 2
1000
0
-1000
-200
-100
0
100
200
300
TSNE component 1
Inam
Puru
JiGu
Mach
Roar
ArSu
SuTa
Para





amazingly with t-sne we see significantly more clusters

note that these results are sensitive to values of perplexity and starting seed, so you can play with those parameters to get more interpretable results

Then we can again write the t-sne results to a file

In [14]:
```
# # store the PC axes as a dataframe
df4 = pd.DataFrame(pca4.pcaxes[0], index=pca4.names)

# # write the PC axes to a CSV file
df4.to_csv("P_ni_TSNE_85minmap_12Jan2022.csv")

# # show the first ten samples and the first 10 PC axes
df4.iloc[:10, :10].round(2)
```

Out[14]:

	0	1
P_ni_77876_suta	-174.49	56.78
P_ni_78155_suta	-251.73	71.05
P_ni_80034_pu	248.67	-39.43
P_ni_80555_arsu	-204.49	115.75
P_ni_80684_arsu	-205.79	69.92
P_ni_80802_arsu	-149.47	235.37
P_ni_80874_arsu	-115.74	102.75
P_ni_85430_arsu	-225.32	98.03
P_ni_85721_suta	-138.58	76.36
P_ni_86072_arsu	-218.82	158.18

Let's skip this for now, but the next section of code does 10,000 TSNE replicates for downstream analysis using randomply generated values for starting seed and "perplexity"

In [15]:
```
# !rm -r M_ru_TSNE
# !mkdir M_ru_TSNE
# import random
# for i in range(10000):
#     pca3.run_tsne(subsample=True, perplexity=random.randrange(3,8), n_
iter=100000, seed=random.randrange(100,9999))
#     df4 = pd.DataFrame(pca3.pcaxes[0], index=pca4.names)
#     df4.to_csv("./M_ru_TSNE/M_ru_TSNE_rep"+str(i)+".csv")
```

Now let's look at just the western clade, and here we assume K=2; again clustering is independent of geography

In [16]:
```
#RUN AGAIN WITH WESTERN POPULATIONS ONLY

imap = {
#"ref": ["reference"],
"Inam": ["P_ni_A7862_In", "P_ni_A7911_In", "P_ni_A7928_In", "P_ni_T6243_
In",],
"Puru": ["P_ni_T5850_pu", "P_ni_T5940_pu", "P_ni_T5974_pu", "P_ni_T15938
_pu","P_ni_80034_pu", "P_ni_T3609_pu", "P_ni_T3611_pu",  "P_ni_T3817_pu"
, "P_ni_T4043_pu", "P_ni_T4051_pu", "P_ni_T4313_pu", "P_ni_T4404_pu"],
#"JiGu": ["P_ni_T22153_jigu", "P_ni_T3261_jigu","P_ni_T15863_jigu", "P_n
i_T15868_jigu", "P_ni_T15871_jigu", "P_ni_A3255_jigu",],
#"Mach": ["P_ni_T443_ma", "P_ni_T467_ma","P_ni_T369_ma","P_ni_J434_ma",
 "P_ni_J461_ma", "P_ni_J462_ma", "P_ni_J485_ma", "P_ni_J210_ma", "P_ni_J
227_ma", "P_ni_J260_ma", "P_ni_A2418_ma",  "P_ni_A542_ma",],
#"Roar": ["P_ni_J684_roar", "P_ni_J724_roar","P_ni_J361_roar", "P_ni_J36
3_roar", "P_ni_J371_roar", "P_ni_J373_roar", "P_ni_J381_roar", "P_ni_J38
5_roar", "P_ni_J389_roar", "P_ni_J417_roar",],
#"ArSu": ["P_ni_J551_arsu", "P_ni_J602_arsu", "P_ni_J603_arsu", "P_ni_J6
14_arsu", "P_ni_J617_arsu","P_ni_80555_arsu","P_ni_86072_arsu", "P_ni_80
684_arsu", "P_ni_80802_arsu", "P_ni_80874_arsu", "P_ni_85430_arsu",],
#"SuTa": ["P_ni_T14543_suta",  "P_ni_T9076_suta","P_ni_T16698_suta","P_n
i_T10967_suta",  "P_ni_T11888_suta","P_ni_T10204_suta","P_ni_A15120_sut
a", "P_ni_77876_suta", "P_ni_78155_suta","P_ni_85721_suta",],
#"Para": ["P_ni_T1642_pa", "P_ni_T18703_pa","P_ni_T12345_pa", "P_ni_T128
54_pa","P_ni_T11193_pa", "P_ni_T11222_pa","P_ni_T10673_pa", "P_ni_T10940
_pa","P_ni_A7066_pa", "P_ni_A14342_pa", "P_ni_A15277_pa",]
}

# minimum % of samples that must be present in each SNP from each group
minmap1 = {i: 0.55 for i in imap}
minmap2 = {i: 0.65 for i in imap}
minmap3 = {i: 0.75 for i in imap}
minmap4 = {i: 0.85 for i in imap}
minmap5 = {i: 0.95 for i in imap}
```

In [17]:
```
# init pca object with input data and (optional) parameter options
pca = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap1,
    mincov=0.85,
    impute_method=2,
)
# init pca object with input data and (optional) parameter options
pca2 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap2,
    mincov=0.85,
    impute_method=2,
)
# init pca object with input data and (optional) parameter options
pca3 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap3,
    mincov=0.85,
    impute_method=2,
)
# init pca object with input data and (optional) parameter options
pca4 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap4,
    mincov=0.85,
    impute_method=2,
)
# init pca object with input data and (optional) parameter options
pca5 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap5,
    mincov=0.85,
    impute_method=2,
)
```

```
Kmeans clustering: iter=0, K=2, mincov=0.9, minmap={'global': 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1026233
Filtered (combined): 1084558
Sites after filtering: 163130
Sites containing missing values: 69446 (42.57%)
Missing values in SNP matrix: 69446 (2.66%)
Imputation: 'sampled'; (0, 1, 2) = 56.0%, 7.7%, 36.4%
{0: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In'], 1: ['P_ni_8003
4_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T3817_
pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu', 'P_ni_T4404_p
u', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_ni_T6243_I
n']}

Kmeans clustering: iter=1, K=2, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.4%, 36.2%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=2, K=2, mincov=0.875, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.3%, 7.6%, 36.1%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=3, K=2, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
```

```
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 7.6%, 36.0%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=4, K=2, mincov=0.85, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 7.4%, 36.2%
Kmeans clustering: iter=0, K=2, mincov=0.9, minmap={'global': 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1026233
Filtered (combined): 1084558
Sites after filtering: 163130
Sites containing missing values: 69446 (42.57%)
Missing values in SNP matrix: 69446 (2.66%)
Imputation: 'sampled'; (0, 1, 2) = 56.1%, 7.4%, 36.5%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=1, K=2, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.5%, 36.2%
{0: ['P_ni_T15938_pu'], 1: ['P_ni_80034_pu', 'P_ni_A7862_In', 'P_ni_A79
```

```
11_In', 'P_ni_A7928_In', 'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T3817_
pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu', 'P_ni_T4404_p
u', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_ni_T6243_I
n']}

Kmeans clustering: iter=2, K=2, mincov=0.875, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1036127
Filtered (combined): 1038602
Sites after filtering: 209086
Sites containing missing values: 115402 (55.19%)
Missing values in SNP matrix: 164751 (4.92%)
Imputation: 'sampled'; (0, 1, 2) = 55.5%, 7.6%, 36.9%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=3, K=2, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.5%, 36.1%
{0: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In'], 1: ['P_ni_8003
4_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T3817_
pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu', 'P_ni_T4404_p
u', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_ni_T6243_I
n']}

Kmeans clustering: iter=4, K=2, mincov=0.85, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.2%, 7.6%, 36.2%
Kmeans clustering: iter=0, K=2, mincov=0.9, minmap={'global': 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
```

```
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1026233
Filtered (combined): 1084558
Sites after filtering: 163130
Sites containing missing values: 69446 (42.57%)
Missing values in SNP matrix: 69446 (2.66%)
Imputation: 'sampled'; (0, 1, 2) = 56.2%, 7.5%, 36.4%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=1, K=2, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.4%, 36.2%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=2, K=2, mincov=0.875, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.5%, 36.1%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=3, K=2, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
```

```
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.3%, 7.5%, 36.2%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=4, K=2, mincov=0.85, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.2%, 7.6%, 36.2%
Kmeans clustering: iter=0, K=2, mincov=0.9, minmap={'global': 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1026233
Filtered (combined): 1084558
Sites after filtering: 163130
Sites containing missing values: 69446 (42.57%)
Missing values in SNP matrix: 69446 (2.66%)
Imputation: 'sampled'; (0, 1, 2) = 56.0%, 7.5%, 36.5%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=1, K=2, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1082723
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.5%, 36.1%
{0: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In'], 1: ['P_ni_8003
4_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_pu', 'P_ni_T3817_
pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu', 'P_ni_T4404_p
u', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_ni_T6243_I
n']}
```

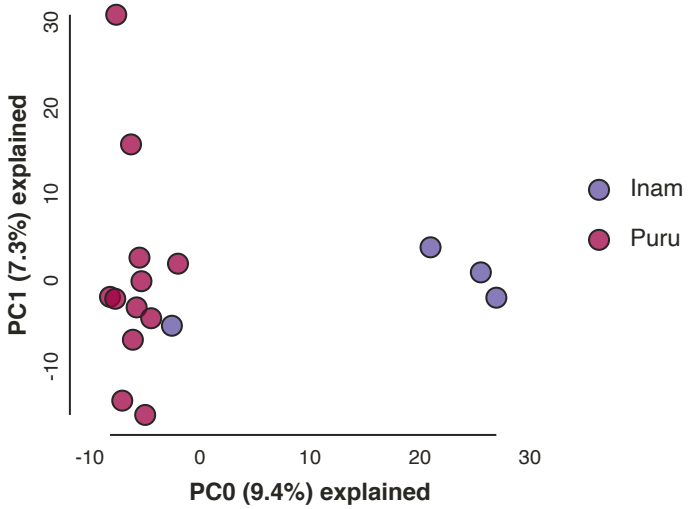
```
Kmeans clustering: iter=2, K=2, mincov=0.875, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.5%, 36.1%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=3, K=2, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 7.4%, 36.2%
{0: ['P_ni_80034_pu', 'P_ni_T15938_pu', 'P_ni_T3609_pu', 'P_ni_T3611_p
u', 'P_ni_T3817_pu', 'P_ni_T4043_pu', 'P_ni_T4051_pu', 'P_ni_T4313_pu',
'P_ni_T4404_pu', 'P_ni_T5850_pu', 'P_ni_T5940_pu', 'P_ni_T5974_pu', 'P_
ni_T6243_In'], 1: ['P_ni_A7862_In', 'P_ni_A7911_In', 'P_ni_A7928_In']}

Kmeans clustering: iter=4, K=2, mincov=0.85, minmap={0: 0.85, 1: 0.85}
Samples: 16
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 10854
Filtered (mincov): 1026233
Filtered (minmap): 1100021
Filtered (combined): 1101674
Sites after filtering: 146014
Sites containing missing values: 52330 (35.84%)
Missing values in SNP matrix: 52330 (2.24%)
Imputation: 'sampled'; (0, 1, 2) = 56.3%, 7.6%, 36.1%
```

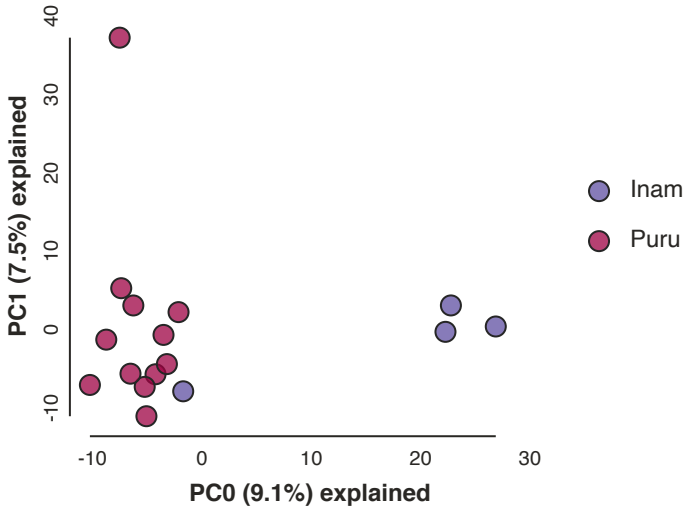
In [18]:
```
# run the PCA analysis
pca.run()
pca2.run()
pca3.run()
pca4.run()
pca5.run()
```

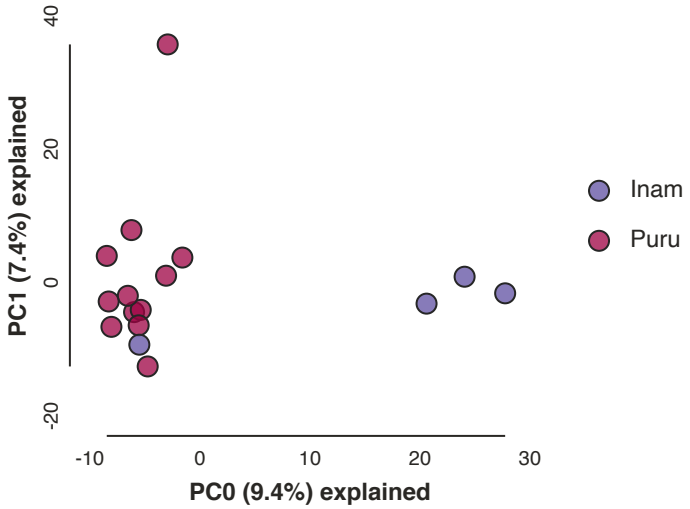
```
Subsampling SNPs: 22121/146014
Subsampling SNPs: 22121/146014
Subsampling SNPs: 22121/146014
Subsampling SNPs: 22121/146014
Subsampling SNPs: 15206/93684
```

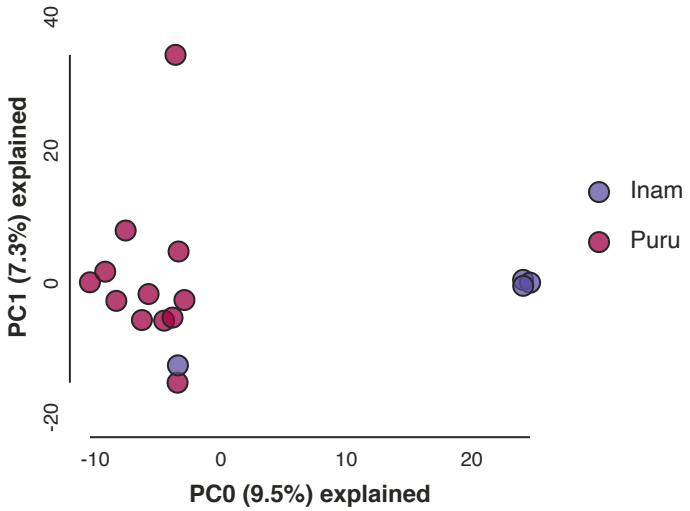
In [19]:
```
pca.draw()
pca2.draw()
pca3.draw()
pca4.draw()
pca5.draw()
```

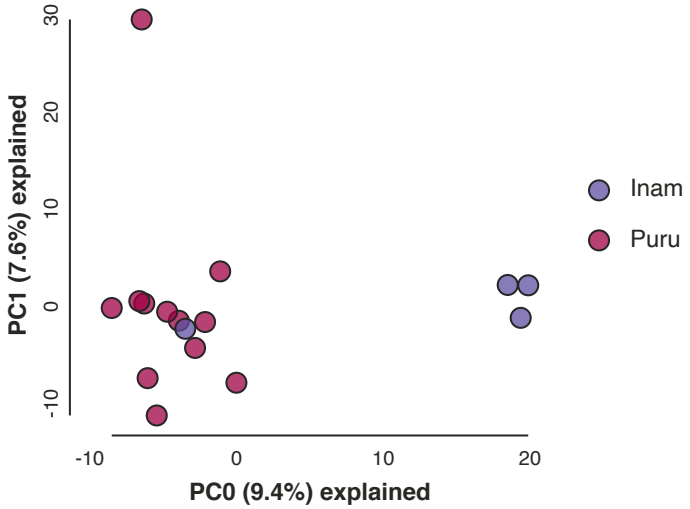
Out[19]: (<toyplot.canvas.Canvas at 0x2b8b93a7b990>,
<toyplot.coordinates.Cartesian at 0x2b8b93a7b610>)






PC1 (7.4%) explained
PC0 (9.4%) explained
Inam
Puru


PC1 (7.3%) explained
PC0 (9.5%) explained
Inam
Puru

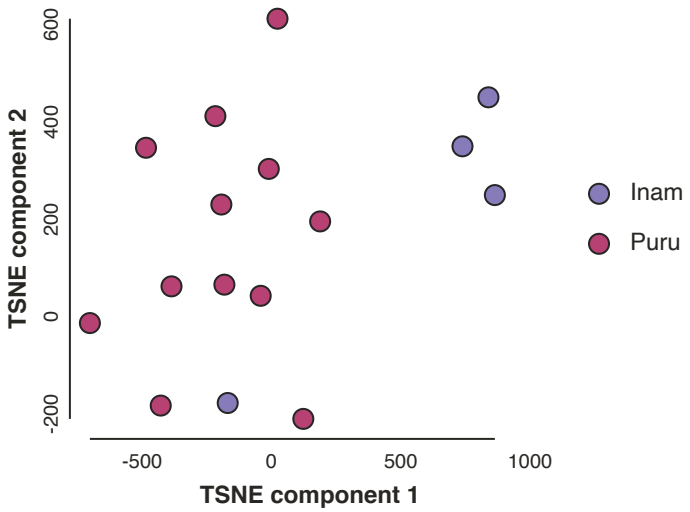

PC1 (7.6%) explained
PC0 (9.4%) explained
Inam
Puru

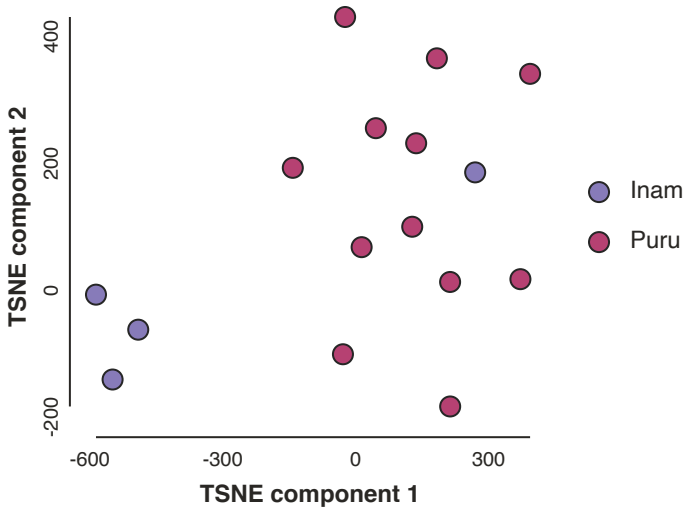
When we zoom in on one clade, we can see that there is even more structure than we initially thought

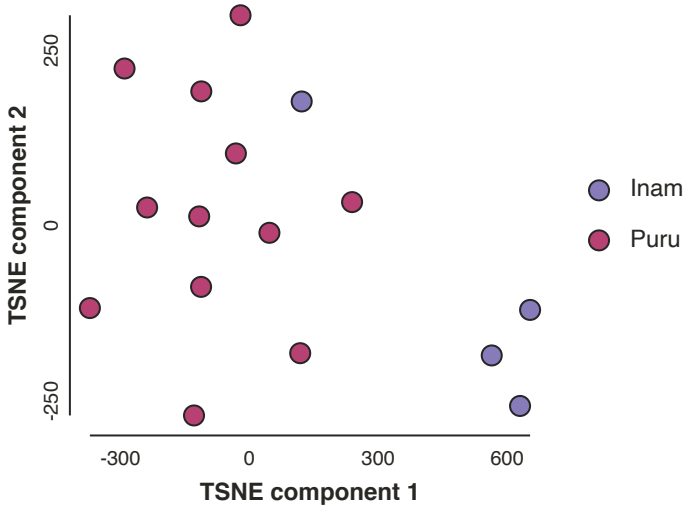
In [20]:
```
pca.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca2.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca3.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca4.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca5.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
```

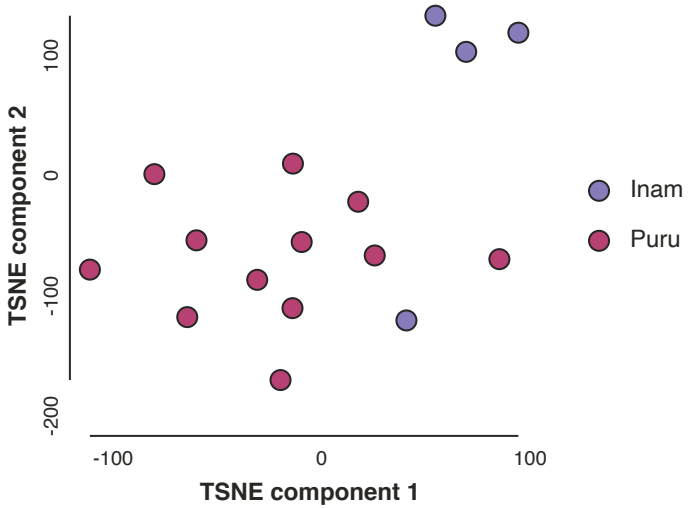
```
Subsampling SNPs: 22121/146014
Subsampling SNPs: 22121/146014
Subsampling SNPs: 22121/146014
Subsampling SNPs: 22121/146014
Subsampling SNPs: 15206/93684
```

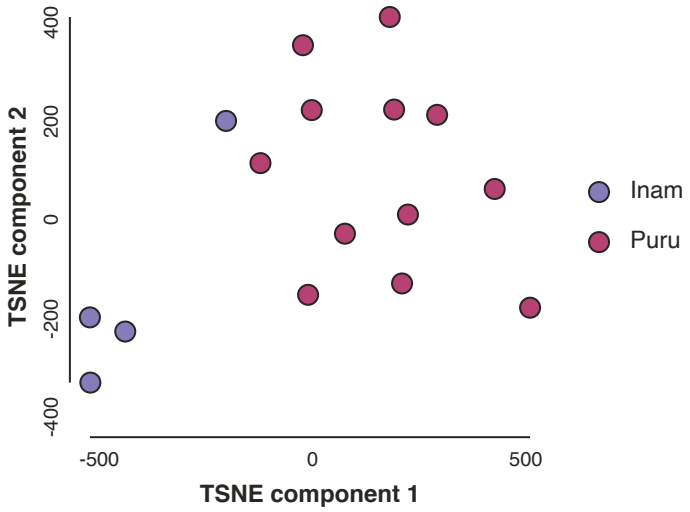
In [21]:
```
pca.draw();
pca2.draw();
pca3.draw();
pca4.draw();
pca5.draw();
```


TSNE component 2
600
400
200
0
-200
-500
0
500
1000
TSNE component 1
Inam
Puru


TSNE component 2
400
200
0
-200
-600
-300
0
300
TSNE component 1
Inam
Puru


TSNE component 2
250
0
-250
-300
0
300
600
TSNE component 1
Inam
Puru





Now let's look at just the eastern clade, and here we assume K=5; again clustering is independent of geography

In [22]:
```
#RUN AGAIN WITH EASTERN POPULATIONS ONLY

imap = {
#"ref": ["reference"],
#"Inam": ["P_ni_A7862_In", "P_ni_A7911_In", "P_ni_A7928_In", "P_ni_T6243
_In",],
#"Puru": ["P_ni_T5850_pu", "P_ni_T5940_pu", "P_ni_T5974_pu", "P_ni_T1593
8_pu","P_ni_80034_pu", "P_ni_T3609_pu", "P_ni_T3611_pu",  "P_ni_T3817_p
u", "P_ni_T4043_pu", "P_ni_T4051_pu", "P_ni_T4313_pu", "P_ni_T4404_pu"],
"JiGu": ["P_ni_T22153_jigu", "P_ni_T3261_jigu","P_ni_T15863_jigu", "P_ni
_T15868_jigu", "P_ni_T15871_jigu", "P_ni_A3255_jigu",],
"Mach": ["P_ni_T443_ma", "P_ni_T467_ma","P_ni_T369_ma","P_ni_J434_ma",
"P_ni_J461_ma", "P_ni_J462_ma", "P_ni_J485_ma", "P_ni_J210_ma", "P_ni_J2
27_ma", "P_ni_J260_ma", "P_ni_A2418_ma",  "P_ni_A542_ma",],
"Roar": ["P_ni_J684_roar", "P_ni_J724_roar","P_ni_J361_roar", "P_ni_J363
_roar", "P_ni_J371_roar", "P_ni_J373_roar", "P_ni_J381_roar", "P_ni_J385
_roar", "P_ni_J389_roar", "P_ni_J417_roar",],
"ArSu": ["P_ni_J551_arsu", "P_ni_J602_arsu", "P_ni_J603_arsu", "P_ni_J61
4_arsu", "P_ni_J617_arsu","P_ni_80555_arsu","P_ni_86072_arsu", "P_ni_806
84_arsu", "P_ni_80802_arsu", "P_ni_80874_arsu", "P_ni_85430_arsu",],
"SuTa": ["P_ni_T14543_suta",  "P_ni_T9076_suta","P_ni_T16698_suta","P_ni
_T10967_suta",  "P_ni_T11888_suta","P_ni_T10204_suta","P_ni_A15120_suta"
, "P_ni_77876_suta", "P_ni_78155_suta","P_ni_85721_suta",],
"Para": ["P_ni_T1642_pa", "P_ni_T18703_pa","P_ni_T12345_pa", "P_ni_T1285
4_pa","P_ni_T11193_pa", "P_ni_T11222_pa","P_ni_T10673_pa", "P_ni_T10940_
pa","P_ni_A7066_pa", "P_ni_A14342_pa", "P_ni_A15277_pa",]
}

# minimum % of samples that must be present in each SNP from each group
minmap1 = {i: 0.55 for i in imap}
minmap2 = {i: 0.65 for i in imap}
minmap3 = {i: 0.75 for i in imap}
minmap4 = {i: 0.85 for i in imap}
minmap5 = {i: 0.95 for i in imap}
```

In [23]:
```
# init pca object with input data and (optional) parameter options
pca = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap1,
    mincov=0.85,
    impute_method=5,
)
# init pca object with input data and (optional) parameter options
pca2 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap2,
    mincov=0.85,
    impute_method=5,
)
# init pca object with input data and (optional) parameter options
pca3 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap3,
    mincov=0.85,
    impute_method=5,
)
# init pca object with input data and (optional) parameter options
pca4 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap4,
    mincov=0.85,
    impute_method=5,
)
# init pca object with input data and (optional) parameter options
pca5 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap5,
    mincov=0.85,
    impute_method=5,
)
```

```
Kmeans clustering: iter=0, K=5, mincov=0.9, minmap={'global': 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 948281
Filtered (combined): 1017570
Sites after filtering: 230118
Sites containing missing values: 200550 (87.15%)
Missing values in SNP matrix: 633362 (4.59%)
Imputation: 'sampled'; (0, 1, 2) = 56.1%, 5.2%, 38.6%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu',
'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu',
'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T1454
3_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_A14342_pa',
'P_ni_A15277_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa',
'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa',
'P_ni_T18703_pa'], 2: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_J210_m
a', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_
ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P
_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_m
a', 'P_ni_T443_ma', 'P_ni_T467_ma'], 3: ['P_ni_A542_ma', 'P_ni_J361_roa
r', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_ro
ar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J684_r
oar', 'P_ni_J724_roar'], 4: ['P_ni_A7066_pa']}

Kmeans clustering: iter=1, K=5, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 1153812
Filtered (combined): 1155361
Sites after filtering: 92327
Sites containing missing values: 62759 (67.97%)
Missing values in SNP matrix: 142117 (2.57%)
Imputation: 'sampled'; (0, 1, 2) = 55.9%, 4.5%, 39.6%
{0: ['P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma'], 1: ['P_ni_A14342_
pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_
pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854
_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 2: ['P_ni_J361_roar', 'P_ni_J
363_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_
J389_roar', 'P_ni_J417_roar'], 3: ['P_ni_77876_suta', 'P_ni_78155_sut
a', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_8087
4_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni
_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P
_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_sut
a', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T
9076_suta'], 4: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P
_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J385_roar', 'P_ni_J
485_ma', 'P_ni_J684_roar', 'P_ni_J724_roar', 'P_ni_T15863_jigu', 'P_ni_
T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jig
```

```
u', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma']}

Kmeans clustering: iter=2, K=5, mincov=0.875, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 990949
Filtered (minmap): 1069680
Filtered (combined): 1072160
Sites after filtering: 175528
Sites containing missing values: 145960 (83.15%)
Missing values in SNP matrix: 378765 (3.60%)
Imputation: 'sampled'; (0, 1, 2) = 56.3%, 4.4%, 39.3%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J602_arsu',
'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_sut
a', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T
16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_J373_roar', 'P_ni_J381_roa
r', 'P_ni_J385_roar', 'P_ni_J551_arsu', 'P_ni_J724_roar', 'P_ni_T15868_
jigu'], 2: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J
210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J461_m
a', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_T15863_jig
u', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T3
69_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 3: ['P_ni_J361_roar', 'P_ni_J3
63_roar', 'P_ni_J371_roar', 'P_ni_J389_roar', 'P_ni_J417_roar'], 4: ['P
_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P
_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa',
'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa']}

Kmeans clustering: iter=3, K=5, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 968397
Filtered (minmap): 1091705
Filtered (combined): 1093841
Sites after filtering: 153847
Sites containing missing values: 124279 (80.78%)
Missing values in SNP matrix: 293209 (3.18%)
Imputation: 'sampled'; (0, 1, 2) = 57.4%, 4.5%, 38.1%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J602_arsu',
'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_sut
a', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T
16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_A14342_pa', 'P_ni_A15277_p
a', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_p
a', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_p
a', 'P_ni_T18703_pa'], 2: ['P_ni_J373_roar', 'P_ni_J381_roar'], 3: ['P_
ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_
J227_ma', 'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_m
a', 'P_ni_J485_ma', 'P_ni_J684_roar', 'P_ni_T15863_jigu', 'P_ni_T15868_
```

```
jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni
_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 4: ['P_ni_J361_roar', 'P_ni
_J363_roar', 'P_ni_J371_roar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_n
i_J417_roar', 'P_ni_J551_arsu', 'P_ni_J724_roar']}

Kmeans clustering: iter=4, K=5, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 948281
Filtered (minmap): 1083392
Filtered (combined): 1085688
Sites after filtering: 162000
Sites containing missing values: 132432 (81.75%)
Missing values in SNP matrix: 326278 (3.36%)
Imputation: 'sampled'; (0, 1, 2) = 57.5%, 4.6%, 37.9%
Kmeans clustering: iter=0, K=5, mincov=0.9, minmap={'global': 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 948281
Filtered (combined): 1017570
Sites after filtering: 230118
Sites containing missing values: 200550 (87.15%)
Missing values in SNP matrix: 633362 (4.59%)
Imputation: 'sampled'; (0, 1, 2) = 56.1%, 5.2%, 38.7%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu',
'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu',
'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T1454
3_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_A14342_pa',
'P_ni_A15277_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa',
'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa',
'P_ni_T18703_pa'], 2: ['P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_r
oar', 'P_ni_J684_roar', 'P_ni_J724_roar'], 3: ['P_ni_A2418_ma', 'P_ni_A
3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_
ma', 'P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J389_r
oar', 'P_ni_J417_roar', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma',
'P_ni_J485_ma', 'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_ji
gu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_
ma', 'P_ni_T467_ma'], 4: ['P_ni_A7066_pa']}

Kmeans clustering: iter=1, K=5, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 1163607
Filtered (combined): 1164976
Sites after filtering: 82712
```

```
Sites containing missing values: 53144 (64.25%)
Missing values in SNP matrix: 112681 (2.27%)
Imputation: 'sampled'; (0, 1, 2) = 56.3%, 4.7%, 39.0%
{0: ['P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar'], 1: ['P_ni_A
14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T
10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_
T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 2: ['P_ni_77876_suta',
'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_ar
su', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_860
72_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni
_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P
_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_
suta', 'P_ni_T9076_suta'], 3: ['P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni
_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J389_roar', 'P_ni_J434
_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J684_roar',
'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261
_jigu', 'P_ni_T369_ma', 'P_ni_T467_ma'], 4: ['P_ni_A2418_ma', 'P_ni_J36
1_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J417_roar', 'P_ni_J7
24_roar', 'P_ni_T15863_jigu', 'P_ni_T443_ma']}

Kmeans clustering: iter=2, K=5, mincov=0.875, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 990949
Filtered (minmap): 1091214
Filtered (combined): 1093412
Sites after filtering: 154276
Sites containing missing values: 124708 (80.83%)
Missing values in SNP matrix: 295783 (3.20%)
Imputation: 'sampled'; (0, 1, 2) = 57.3%, 4.5%, 38.3%
{0: ['P_ni_A3255_jigu', 'P_ni_J373_roar', 'P_ni_J684_roar', 'P_ni_T1586
3_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P
_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 1:
['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_a
rsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85
721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_
ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P
_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_
suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 2: ['P_ni_A14342_pa', 'P
_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P
_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa',
'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_ni_A2418_ma', 'P_ni_J361_roa
r', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J381_roar', 'P_ni_J385_ro
ar', 'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni_J724_roar'], 4: ['P_ni_A
542_ma', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J434_m
a', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma']}

Kmeans clustering: iter=3, K=5, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 968397
```

```
Filtered (minmap): 1074987
Filtered (combined): 1077428
Sites after filtering: 170260
Sites containing missing values: 140692 (82.63%)
Missing values in SNP matrix: 351729 (3.44%)
Imputation: 'sampled'; (0, 1, 2) = 56.8%, 4.4%, 38.8%
{0: ['P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227_ma',
'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J
485_ma', 'P_ni_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P
_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P
_ni_T467_ma'], 1: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa',
'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa',
'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'],
2: ['P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_ar
su'], 3: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_n
i_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu',
'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_ar
su', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_
T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 4: ['P_ni_A2418_m
a', 'P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_ro
ar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_ni_J417_r
oar', 'P_ni_J684_roar', 'P_ni_J724_roar']}

Kmeans clustering: iter=4, K=5, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 948281
Filtered (minmap): 1093881
Filtered (combined): 1095987
Sites after filtering: 151701
Sites containing missing values: 122133 (80.51%)
Missing values in SNP matrix: 270694 (2.97%)
Imputation: 'sampled'; (0, 1, 2) = 56.7%, 4.4%, 38.9%
Kmeans clustering: iter=0, K=5, mincov=0.9, minmap={'global': 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 948281
Filtered (combined): 1017570
Sites after filtering: 230118
Sites containing missing values: 200550 (87.15%)
Missing values in SNP matrix: 633362 (4.59%)
Imputation: 'sampled'; (0, 1, 2) = 56.2%, 5.2%, 38.6%
{0: ['P_ni_A7066_pa'], 1: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_ni_J3
71_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni_J
389_roar', 'P_ni_J417_roar', 'P_ni_J684_roar'], 2: ['P_ni_77876_suta',
'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_ar
su', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_860
72_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni
_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P
_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_
suta', 'P_ni_T9076_suta'], 3: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_n
```

```
i_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_
ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 4:
['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P
_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J46
2_ma', 'P_ni_J485_ma', 'P_ni_J724_roar', 'P_ni_T15863_jigu', 'P_ni_T158
68_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P
_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma']}

Kmeans clustering: iter=1, K=5, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 1153679
Filtered (combined): 1155236
Sites after filtering: 92452
Sites containing missing values: 62884 (68.02%)
Missing values in SNP matrix: 142548 (2.57%)
Imputation: 'sampled'; (0, 1, 2) = 55.7%, 4.7%, 39.6%
{0: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_
pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345
_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 1: ['P_ni_A
542_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J361_roar', 'P_ni_J363_r
oar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_
roar', 'P_ni_J417_roar', 'P_ni_J461_ma', 'P_ni_J684_roar', 'P_ni_J724_r
oar', 'P_ni_T15863_jigu', 'P_ni_T369_ma', 'P_ni_T467_ma'], 2: ['P_ni_77
876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_
ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_sut
a', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602
_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10
204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta',
'P_ni_T16698_suta', 'P_ni_T9076_suta'], 3: ['P_ni_J434_ma', 'P_ni_J462_
ma', 'P_ni_J485_ma'], 4: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_J21
0_ma', 'P_ni_J389_roar', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_
T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_T443_ma']}

Kmeans clustering: iter=2, K=5, mincov=0.875, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 990949
Filtered (minmap): 1062989
Filtered (combined): 1065541
Sites after filtering: 182147
Sites containing missing values: 152579 (83.77%)
Missing values in SNP matrix: 410699 (3.76%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 4.4%, 39.2%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu',
'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu',
'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T1454
3_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_J373_roar',
```

```
'P_ni_J381_roar', 'P_ni_J389_roar', 'P_ni_J684_roar', 'P_ni_J724_roa
r'], 2: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10
673_pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T1
2345_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_
ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_
J227_ma', 'P_ni_J260_ma', 'P_ni_J385_roar', 'P_ni_J434_ma', 'P_ni_J461_
ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T15863_jigu', 'P_ni_T15868_j
igu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P_ni_
T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 4: ['P_ni_J361_roar', 'P_ni_
J363_roar', 'P_ni_J371_roar', 'P_ni_J417_roar']}

Kmeans clustering: iter=3, K=5, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 968397
Filtered (minmap): 1091057
Filtered (combined): 1093198
Sites after filtering: 154490
Sites containing missing values: 124922 (80.86%)
Missing values in SNP matrix: 295665 (3.19%)
Imputation: 'sampled'; (0, 1, 2) = 57.5%, 4.5%, 38.0%
{0: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_
pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345
_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 1: ['P_ni_J
361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J385_roar', 'P_ni_
J417_roar', 'P_ni_T443_ma'], 2: ['P_ni_77876_suta', 'P_ni_78155_suta',
'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_ar
su', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15
120_suta', 'P_ni_J551_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_
J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta',
'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076
_suta'], 3: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_
J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J389_roar', 'P_ni_J434_
ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T15863_jig
u', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T
3261_jigu', 'P_ni_T369_ma', 'P_ni_T467_ma'], 4: ['P_ni_J373_roar', 'P_n
i_J381_roar', 'P_ni_J684_roar', 'P_ni_J724_roar']}

Kmeans clustering: iter=4, K=5, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 948281
Filtered (minmap): 1093337
Filtered (combined): 1095434
Sites after filtering: 152254
Sites containing missing values: 122686 (80.58%)
Missing values in SNP matrix: 287185 (3.14%)
Imputation: 'sampled'; (0, 1, 2) = 57.5%, 4.4%, 38.1%
Kmeans clustering: iter=0, K=5, mincov=0.9, minmap={'global': 0.85}
Samples: 60
Sites before filtering: 1247688
```

```
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 948281
Filtered (combined): 1017570
Sites after filtering: 230118
Sites containing missing values: 200550 (87.15%)
Missing values in SNP matrix: 633362 (4.59%)
Imputation: 'sampled'; (0, 1, 2) = 56.2%, 5.2%, 38.6%
{0: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_806
84_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_n
i_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu',
'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu',
'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T1454
3_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 1: ['P_ni_A14342_pa',
'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa',
'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa',
'P_ni_T1642_pa', 'P_ni_T18703_pa'], 2: ['P_ni_A2418_ma', 'P_ni_A3255_ji
gu', 'P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P
_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar', 'P_ni_J389_roar',
'P_ni_J461_ma', 'P_ni_J684_roar', 'P_ni_J724_roar', 'P_ni_T15863_jigu',
'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261
_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'], 3: ['P_ni_J361
_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J417_roar'], 4: ['P_n
i_J434_ma', 'P_ni_J462_ma', 'P_ni_J485_ma']}

Kmeans clustering: iter=1, K=5, mincov=0.8875, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 1014236
Filtered (minmap): 1059879
Filtered (combined): 1065499
Sites after filtering: 182189
Sites containing missing values: 152621 (83.77%)
Missing values in SNP matrix: 407004 (3.72%)
Imputation: 'sampled'; (0, 1, 2) = 56.4%, 4.5%, 39.1%
{0: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma', 'P_ni_J210_m
a', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J389_roar', 'P_ni_J434_ma',
'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_J724_roar', 'P_ni
_T15863_jigu', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jig
u', 'P_ni_T3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma'],
1: ['P_ni_78155_suta', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_r
oar'], 2: ['P_ni_77876_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_
ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_sut
a', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J551_arsu', 'P_ni_J602
_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu', 'P_ni_T10
204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta', 'P_ni_T14543_suta',
'P_ni_T16698_suta', 'P_ni_T9076_suta'], 3: ['P_ni_A14342_pa', 'P_ni_A15
277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T10940_pa', 'P_ni_T11
193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_T12854_pa', 'P_ni_T1
642_pa', 'P_ni_T18703_pa'], 4: ['P_ni_J361_roar', 'P_ni_J363_roar', 'P_
ni_J371_roar', 'P_ni_J417_roar', 'P_ni_J684_roar']}

Kmeans clustering: iter=2, K=5, mincov=0.875, minmap={0: 0.85, 1: 0.85,
```

```
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 990949
Filtered (minmap): 1088261
Filtered (combined): 1090427
Sites after filtering: 157261
Sites containing missing values: 127693 (81.20%)
Missing values in SNP matrix: 314346 (3.33%)
Imputation: 'sampled'; (0, 1, 2) = 57.3%, 4.5%, 38.3%
{0: ['P_ni_A542_ma', 'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P
_ni_J434_ma', 'P_ni_J461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T46
7_ma'], 1: ['P_ni_77876_suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P
_ni_80684_arsu', 'P_ni_80802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_ars
u', 'P_ni_85721_suta', 'P_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_J55
1_arsu', 'P_ni_J602_arsu', 'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J6
17_arsu', 'P_ni_T10204_suta', 'P_ni_T10967_suta', 'P_ni_T11888_suta',
'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni_T9076_suta'], 2: ['P_ni_A
14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_pa', 'P_ni_T
10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345_pa', 'P_ni_
T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 3: ['P_ni_A2418_ma', 'P
_ni_A3255_jigu', 'P_ni_J373_roar', 'P_ni_J381_roar', 'P_ni_J385_roar',
'P_ni_J389_roar', 'P_ni_J684_roar', 'P_ni_J724_roar', 'P_ni_T15863_jig
u', 'P_ni_T15868_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T
3261_jigu', 'P_ni_T369_ma', 'P_ni_T443_ma'], 4: ['P_ni_J361_roar', 'P_n
i_J363_roar', 'P_ni_J371_roar', 'P_ni_J417_roar']}

Kmeans clustering: iter=3, K=5, mincov=0.8625, minmap={0: 0.85, 1: 0.8
5, 2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 968397
Filtered (minmap): 1065649
Filtered (combined): 1068208
Sites after filtering: 179480
Sites containing missing values: 149912 (83.53%)
Missing values in SNP matrix: 393776 (3.66%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 4.5%, 39.0%
{0: ['P_ni_A14342_pa', 'P_ni_A15277_pa', 'P_ni_A7066_pa', 'P_ni_T10673_
pa', 'P_ni_T10940_pa', 'P_ni_T11193_pa', 'P_ni_T11222_pa', 'P_ni_T12345
_pa', 'P_ni_T12854_pa', 'P_ni_T1642_pa', 'P_ni_T18703_pa'], 1: ['P_ni_J
361_roar', 'P_ni_J363_roar', 'P_ni_J371_roar', 'P_ni_J373_roar', 'P_ni_
J381_roar', 'P_ni_J385_roar', 'P_ni_J389_roar', 'P_ni_J417_roar', 'P_ni
_J684_roar', 'P_ni_J724_roar'], 2: ['P_ni_J551_arsu', 'P_ni_J602_arsu',
'P_ni_J603_arsu', 'P_ni_J614_arsu', 'P_ni_J617_arsu'], 3: ['P_ni_77876_
suta', 'P_ni_78155_suta', 'P_ni_80555_arsu', 'P_ni_80684_arsu', 'P_ni_8
0802_arsu', 'P_ni_80874_arsu', 'P_ni_85430_arsu', 'P_ni_85721_suta', 'P
_ni_86072_arsu', 'P_ni_A15120_suta', 'P_ni_T10204_suta', 'P_ni_T10967_s
uta', 'P_ni_T11888_suta', 'P_ni_T14543_suta', 'P_ni_T16698_suta', 'P_ni
_T9076_suta'], 4: ['P_ni_A2418_ma', 'P_ni_A3255_jigu', 'P_ni_A542_ma',
'P_ni_J210_ma', 'P_ni_J227_ma', 'P_ni_J260_ma', 'P_ni_J434_ma', 'P_ni_J
461_ma', 'P_ni_J462_ma', 'P_ni_J485_ma', 'P_ni_T15863_jigu', 'P_ni_T158
68_jigu', 'P_ni_T15871_jigu', 'P_ni_T22153_jigu', 'P_ni_T3261_jigu', 'P
```

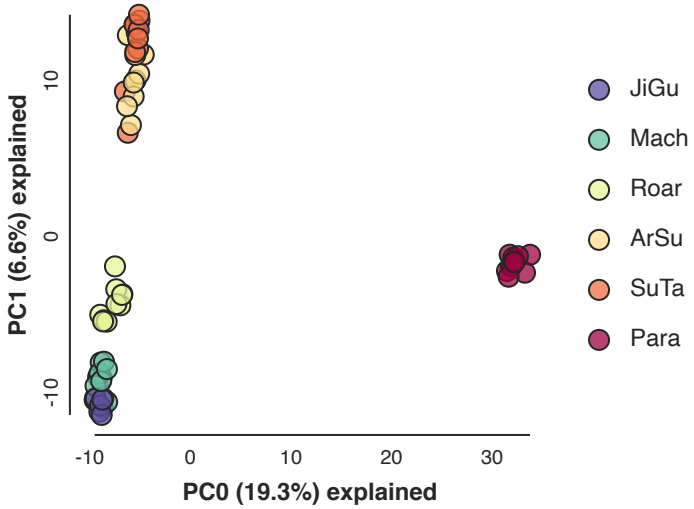
```
_ni_T369_ma', 'P_ni_T443_ma', 'P_ni_T467_ma']}

Kmeans clustering: iter=4, K=5, mincov=0.85, minmap={0: 0.85, 1: 0.85,
2: 0.85, 3: 0.85, 4: 0.85}
Samples: 60
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 20282
Filtered (mincov): 948281
Filtered (minmap): 1088873
Filtered (combined): 1091062
Sites after filtering: 156626
Sites containing missing values: 127058 (81.12%)
Missing values in SNP matrix: 290475 (3.09%)
Imputation: 'sampled'; (0, 1, 2) = 56.5%, 4.3%, 39.2%
```

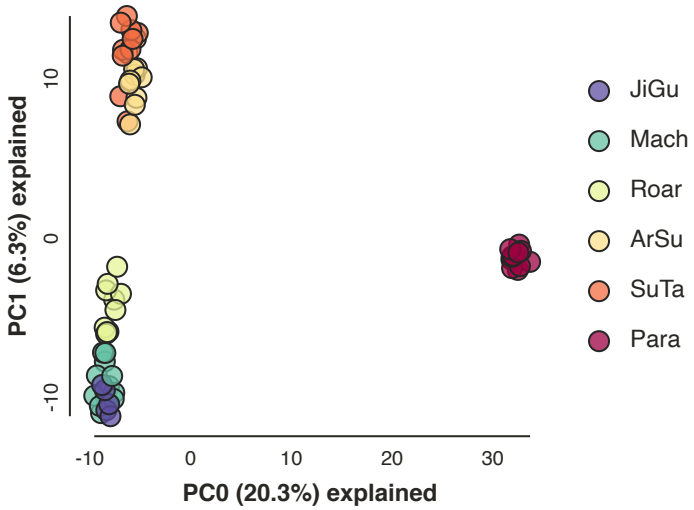
In [24]:
```
# run the PCA analysis
pca.run()
pca2.run()
pca3.run()
pca4.run()
pca5.run()
```

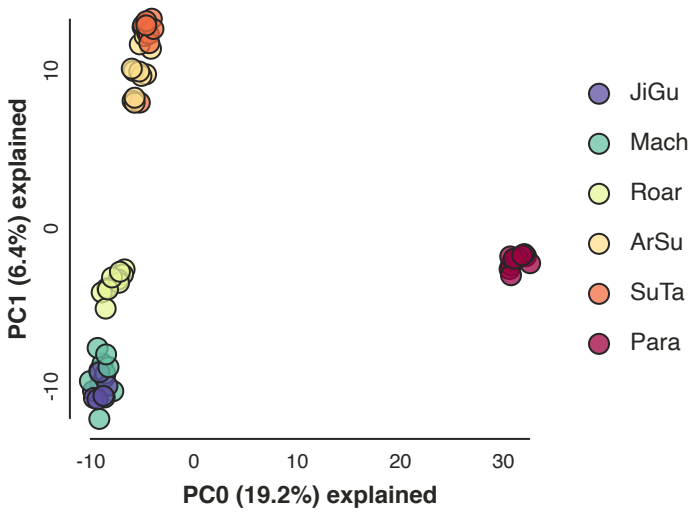
```
Subsampling SNPs: 22341/162000
Subsampling SNPs: 21557/151701
Subsampling SNPs: 21307/152254
Subsampling SNPs: 22315/156626
Subsampling SNPs: 5126/29568
```

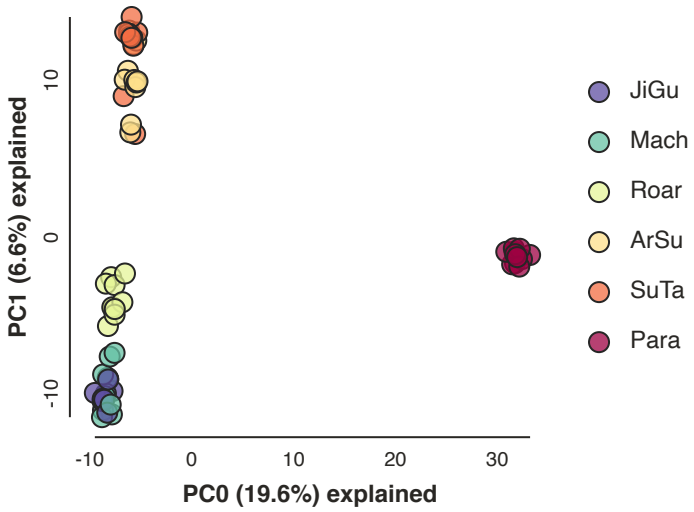
In [25]:
```
pca.draw()
pca2.draw()
pca3.draw()
pca4.draw()
pca5.draw()
```

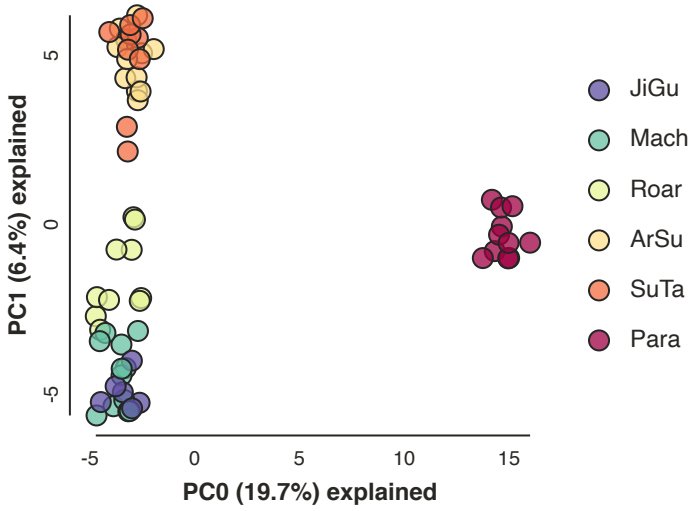
Out[25]: (<toyplot.canvas.Canvas at 0x2b8b94714f10>,
<toyplot.coordinates.Cartesian at 0x2b8b9471f390>)






PC1 (6.4%) explained
10
0
-10
-10
0
10
20
30
PC0 (19.2%) explained
JiGu
Mach
Roar
ArSu
SuTa
Para

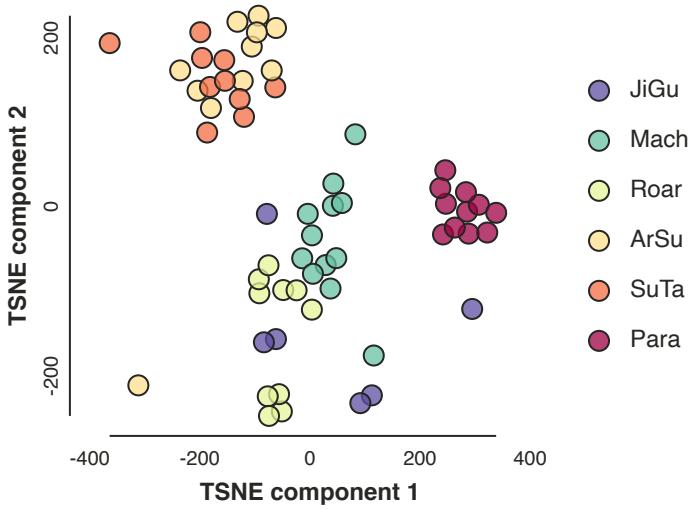

PC1 (6.6%) explained
10
0
-10
-10
0
10
20
30
PC0 (19.6%) explained
JiGu
Mach
Roar
ArSu
SuTa
Para

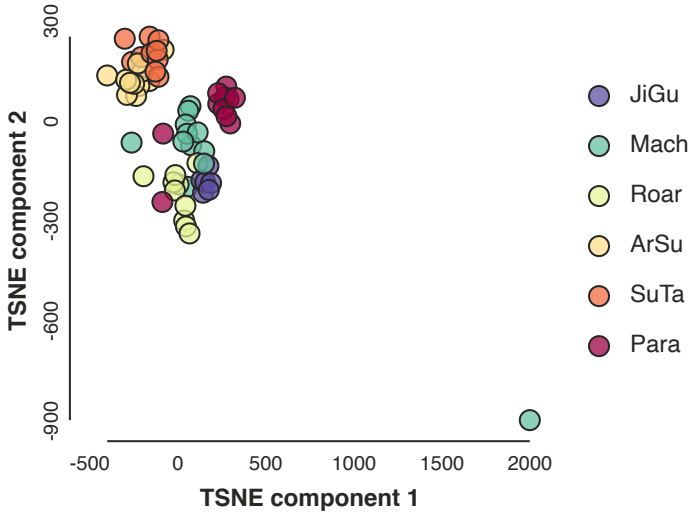

PC1 (6.4%) explained
5
0
-5
-5
0
5
10
15
PC0 (19.7%) explained
JiGu
Mach
Roar
ArSu
SuTa
Para

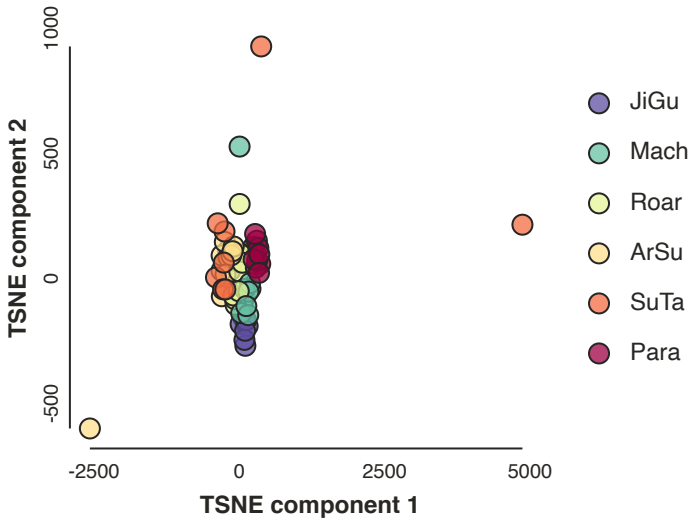
In [26]:
```
pca.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca2.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca3.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca4.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
pca5.run_tsne(subsample=True, perplexity=5.0, n_iter=1000000, seed=123)
```

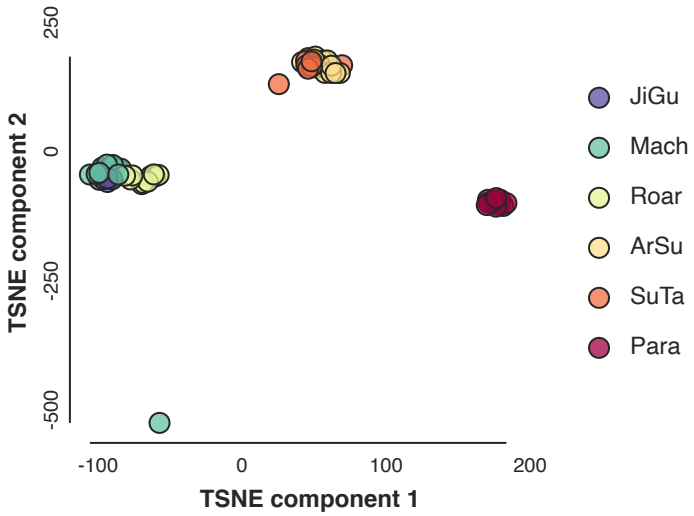
```
Subsampling SNPs: 22341/162000
Subsampling SNPs: 21557/151701
Subsampling SNPs: 21307/152254
Subsampling SNPs: 22315/156626
Subsampling SNPs: 5126/29568
```

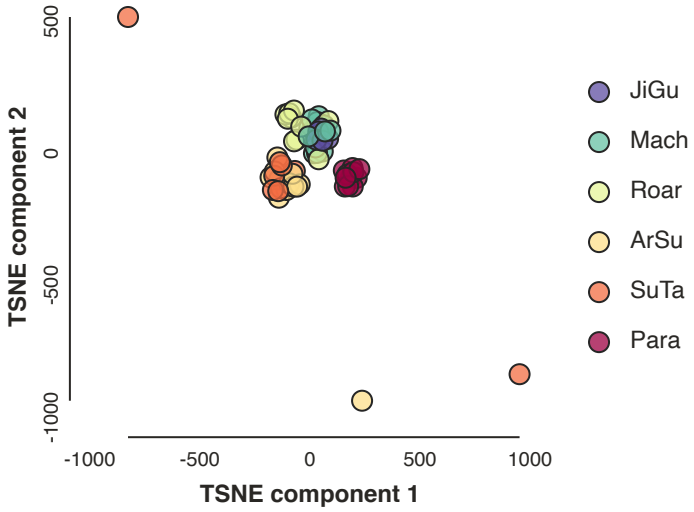
In [27]:
```
pca.draw();
pca2.draw();
pca3.draw();
pca4.draw();
pca5.draw();
```


TSNE component 2
200
0
-200
-400
-200
0
200
400
TSNE component 1
JiGu
Mach
Roar
ArSu
SuTa
Para


TSNE component 2
300
0
-300
-600
-900
-500
0
500
1000
1500
2000
TSNE component 1
JiGu
Mach
Roar
ArSu
SuTa
Para


TSNE component 2
1000
500
0
-500
-2500
0
2500
5000
TSNE component 1
JiGu
Mach
Roar
ArSu
SuTa
Para





Now let's see if the results hold up to a different clustering algorithm so I will group samples into "populations" based on results from the structure analysis to see if PCA and t-SNE look the same. Here we use "sample" to impute from a priori defined populations

In [28]:
```
imap = {
#"ref": ["reference"],
"Inam": ["P_ni_A7862_In", "P_ni_A7911_In", "P_ni_A7928_In", ],
"Puru": ["P_ni_T6243_In","P_ni_T5850_pu", "P_ni_T5940_pu", "P_ni_T5974_p
u", "P_ni_T15938_pu","P_ni_80034_pu", "P_ni_T3609_pu", "P_ni_T3611_pu",
"P_ni_T3817_pu", "P_ni_T4043_pu", "P_ni_T4051_pu", "P_ni_T4313_pu", "P_n
i_T4404_pu","P_ni_T22153_jigu"],
"Rondonia": [ "P_ni_T3261_jigu","P_ni_T15863_jigu", "P_ni_T15868_jigu",
"P_ni_T15871_jigu", "P_ni_A3255_jigu","P_ni_T443_ma", "P_ni_T467_ma","P_
ni_T369_ma","P_ni_J434_ma", "P_ni_J461_ma", "P_ni_J462_ma", "P_ni_J485_m
a", "P_ni_J210_ma", "P_ni_J227_ma", "P_ni_J260_ma", "P_ni_A2418_ma",  "P
_ni_A542_ma","P_ni_J684_roar", "P_ni_J724_roar","P_ni_J361_roar", "P_ni_
J363_roar", "P_ni_J371_roar", "P_ni_J373_roar", "P_ni_J381_roar", "P_ni_
J385_roar", "P_ni_J389_roar", "P_ni_J417_roar","P_ni_J551_arsu", "P_ni_J
602_arsu", "P_ni_J603_arsu", "P_ni_J614_arsu", "P_ni_J617_arsu","P_ni_80
555_arsu","P_ni_86072_arsu", "P_ni_80684_arsu", "P_ni_80802_arsu", "P_ni
_80874_arsu", "P_ni_85430_arsu","P_ni_T14543_suta",  "P_ni_T9076_suta",
"P_ni_T16698_suta","P_ni_T10967_suta",  "P_ni_T11888_suta","P_ni_T10204_
suta","P_ni_A15120_suta", "P_ni_77876_suta", "P_ni_78155_suta","P_ni_857
21_suta",],
"Para": ["P_ni_T1642_pa", "P_ni_T18703_pa","P_ni_T12345_pa", "P_ni_T1285
4_pa","P_ni_T11193_pa", "P_ni_T11222_pa","P_ni_T10673_pa", "P_ni_T10940_
pa","P_ni_A7066_pa", "P_ni_A14342_pa", "P_ni_A15277_pa",]
}

# minimum % of samples that must be present in each SNP from each group
minmap1 = {i: 0.55 for i in imap}
minmap2 = {i: 0.65 for i in imap}
minmap3 = {i: 0.75 for i in imap}
minmap4 = {i: 0.85 for i in imap}
minmap5 = {i: 0.95 for i in imap}
```

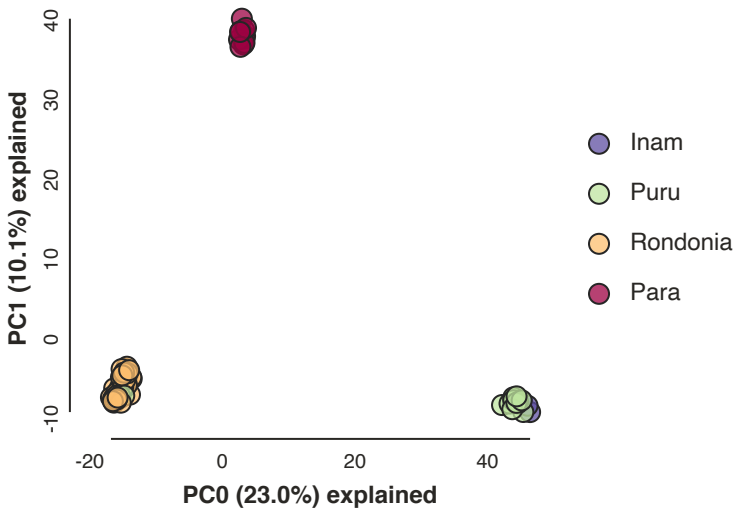
In [29]:
```
# init pca object with input data and (optional) parameter options
pca = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap1,
    mincov=0.85,
    impute_method="sample",
)
# init pca object with input data and (optional) parameter options
pca2 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap2,
    mincov=0.85,
    impute_method="sample",
)
# init pca object with input data and (optional) parameter options
pca3 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap3,
    mincov=0.85,
    impute_method="sample",
)
# init pca object with input data and (optional) parameter options
pca4 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap4,
    mincov=0.85,
    impute_method="sample",
)
# init pca object with input data and (optional) parameter options
pca5 = ipa.pca(
    data=data,
    imap=imap,
    minmap=minmap5,
    mincov=0.85,
    impute_method="sample",
)
```

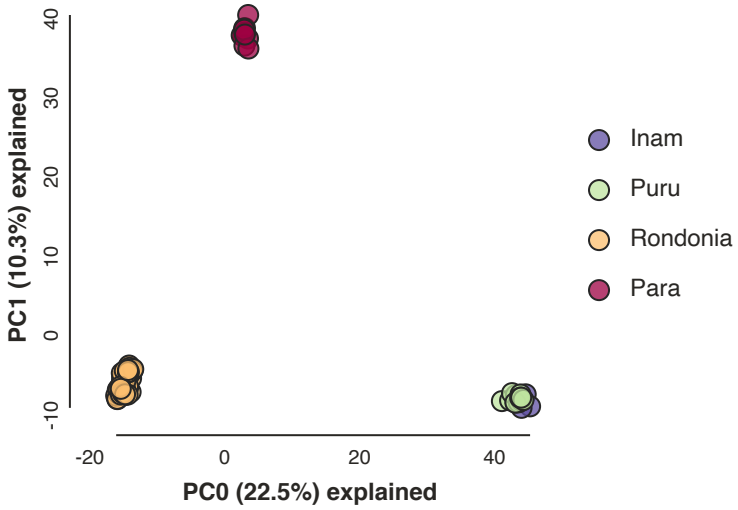
```
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 906778
Filtered (combined): 1007669
Sites after filtering: 240019
Sites containing missing values: 220744 (91.97%)
Missing values in SNP matrix: 1127425 (6.18%)
Imputation: 'sampled'; (0, 1, 2) = 56.6%, 5.6%, 37.8%
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 964634
Filtered (combined): 1015600
Sites after filtering: 232088
Sites containing missing values: 212813 (91.69%)
Missing values in SNP matrix: 1055182 (5.98%)
Imputation: 'sampled'; (0, 1, 2) = 56.7%, 5.5%, 37.8%
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 1075782
Filtered (combined): 1085438
Sites after filtering: 162250
Sites containing missing values: 142975 (88.12%)
Missing values in SNP matrix: 574431 (4.66%)
Imputation: 'sampled'; (0, 1, 2) = 57.0%, 5.5%, 37.5%
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 1115492
Filtered (combined): 1118338
Sites after filtering: 129350
Sites containing missing values: 110075 (85.10%)
Missing values in SNP matrix: 348662 (3.55%)
Imputation: 'sampled'; (0, 1, 2) = 57.0%, 5.4%, 37.6%
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 991306
Filtered (minmap): 1210885
Filtered (combined): 1211587
Sites after filtering: 36101
Sites containing missing values: 16826 (46.61%)
Missing values in SNP matrix: 22113 (0.81%)
Imputation: 'sampled'; (0, 1, 2) = 57.8%, 4.6%, 37.6%
```

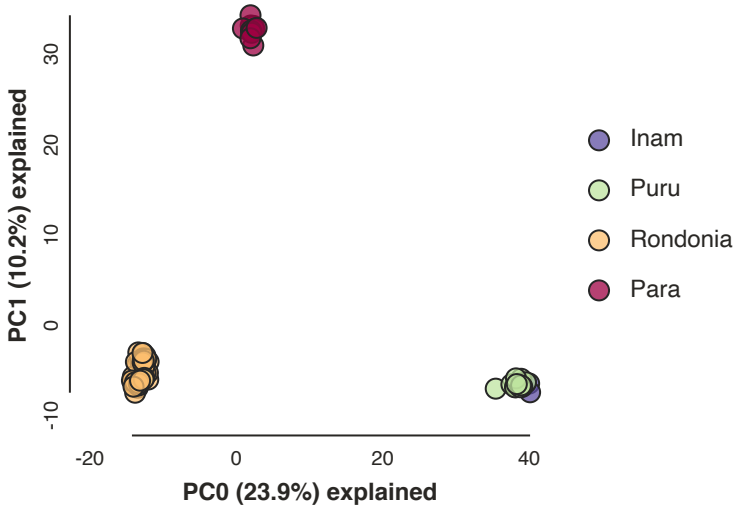
In [30]:
```
# run the PCA analysis
pca.run()
pca2.run()
pca3.run()
pca4.run()
pca5.run()
```

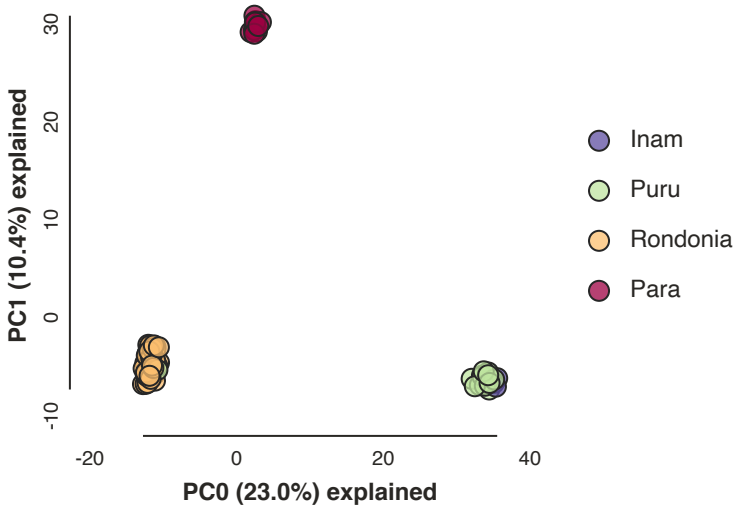
```
Subsampling SNPs: 32610/240019
Subsampling SNPs: 31800/232088
Subsampling SNPs: 23662/162250
Subsampling SNPs: 19387/129350
Subsampling SNPs: 6182/36101
```

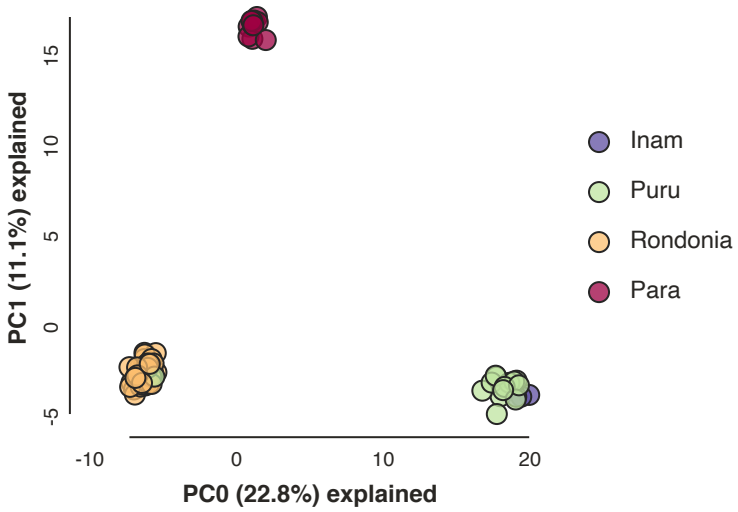
In [31]:
```
pca.draw();
pca2.draw();
pca3.draw();
pca4.draw();
pca5.draw();
```


PC1 (10.1%) explained
PC0 (23.0%) explained
Inam
Puru
Rondonia
Para


PC1 (10.3%) explained
PC0 (22.5%) explained
Inam
Puru
Rondonia
Para


PC1 (10.2%) explained
PC0 (23.9%) explained
Inam
Puru
Rondonia
Para

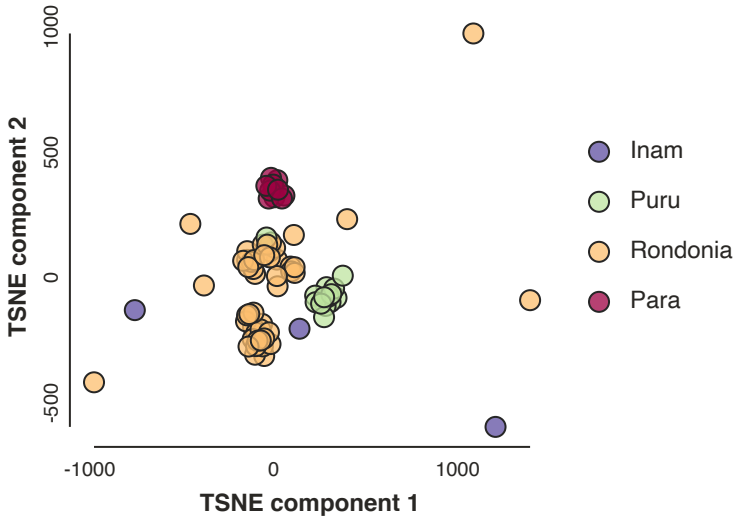


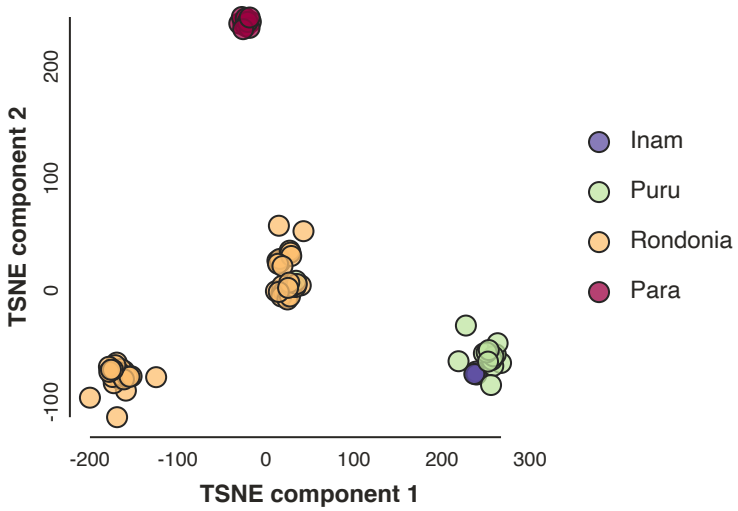


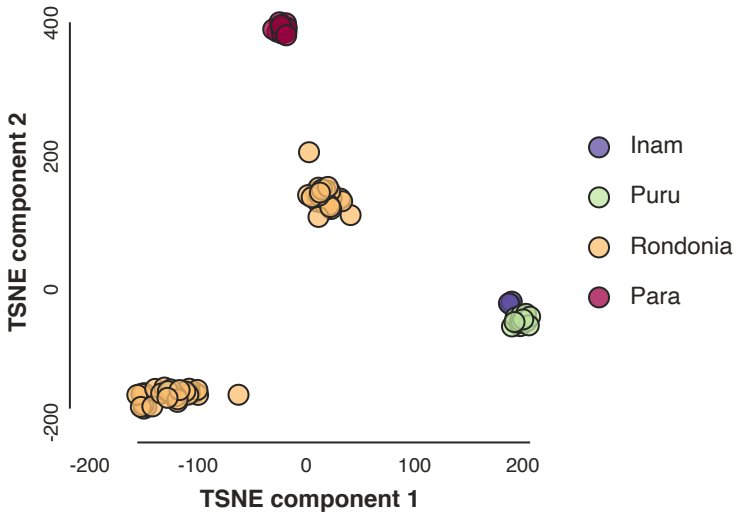
In [32]:
```
pca.run_tsne(subsample=True, perplexity=4.0, n_iter=1000000, seed=223)
pca2.run_tsne(subsample=True, perplexity=4.0, n_iter=1000000, seed=123)
pca3.run_tsne(subsample=True, perplexity=4.0, n_iter=1000000, seed=223)
pca4.run_tsne(subsample=True, perplexity=4.0, n_iter=1000000, seed=123)
pca5.run_tsne(subsample=True, perplexity=4.0, n_iter=1000000, seed=223)
```

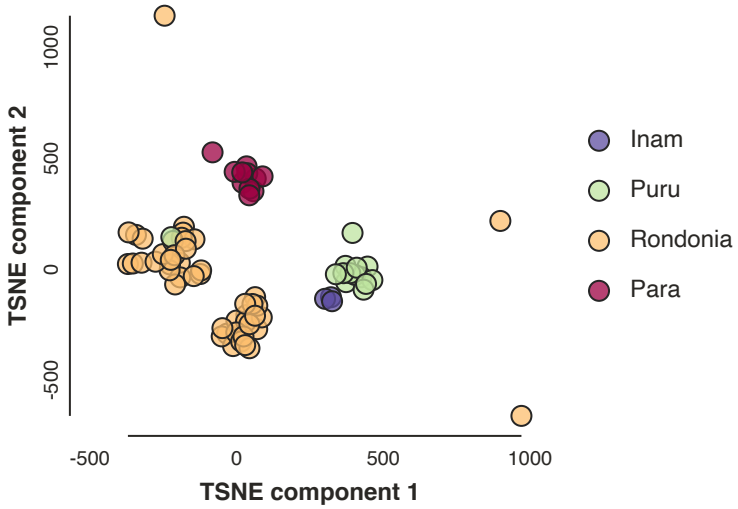
```
Subsampling SNPs: 32610/240019
Subsampling SNPs: 31800/232088
Subsampling SNPs: 23662/162250
Subsampling SNPs: 19387/129350
Subsampling SNPs: 6182/36101
```

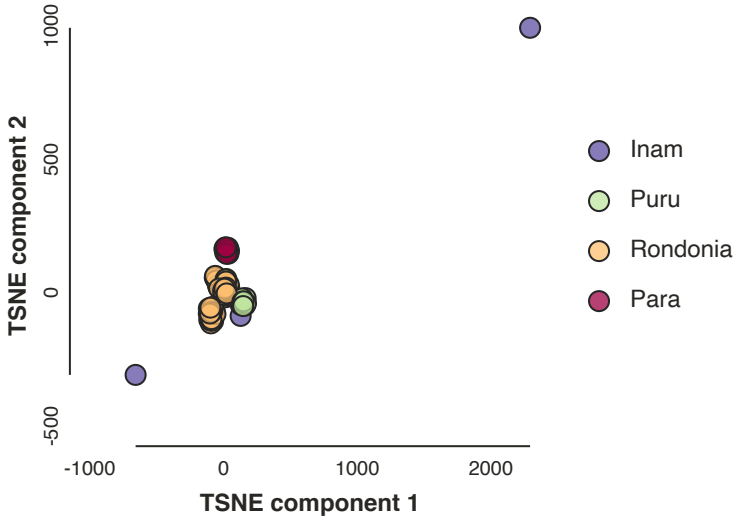
In [33]:
```
pca.draw();
pca2.draw();
pca3.draw();
pca4.draw();
pca5.draw();
```


TSNE component 2
1000
500
0
-500
-1000
0
1000
TSNE component 1
Inam
Puru
Rondonia
Para


TSNE component 2
200
100
0
-100
-200
-100
0
100
200
300
TSNE component 1
Inam
Puru
Rondonia
Para


TSNE component 2
400
200
0
-200
-200
-100
0
100
200
TSNE component 1
Inam
Puru
Rondonia
Para





Amazingly, the results look rather similar to the initial results of imputation by popuilations defined by the k-means clustering analysis

(You can place your cursor on a point to see the label)

Genetic Distances (dxy)

Now let's generate distance matrices for EEMs and see if imputation effects measures of distance. First look at imputation results based on structure assignments, and then look at imputation results based on a priori assignments

In [46]:
```
imap = {
#"ref": ["reference"],
"Inam": ["P_ni_A7862_In", "P_ni_A7911_In", "P_ni_A7928_In", ],
"Puru": ["P_ni_T6243_In","P_ni_T5850_pu", "P_ni_T5940_pu", "P_ni_T5974_pu", "P_ni_T15938_pu","P_ni_80034_pu", "P_ni_T3609_pu", "P_ni_T3611_pu", "P_ni_T3817_pu", "P_ni_T4043_pu", "P_ni_T4051_pu", "P_ni_T4313_pu", "P_ni_T4404_pu","P_ni_T22153_jigu"],
"Rondonia": [ "P_ni_T3261_jigu","P_ni_T15863_jigu", "P_ni_T15868_jigu", "P_ni_T15871_jigu", "P_ni_A3255_jigu","P_ni_T443_ma", "P_ni_T467_ma","P_ni_T369_ma","P_ni_J434_ma", "P_ni_J461_ma", "P_ni_J462_ma", "P_ni_J485_ma", "P_ni_J210_ma", "P_ni_J227_ma", "P_ni_J260_ma", "P_ni_A2418_ma",  "P_ni_A542_ma","P_ni_J684_roar", "P_ni_J724_roar","P_ni_J361_roar", "P_ni_J363_roar", "P_ni_J371_roar", "P_ni_J373_roar", "P_ni_J381_roar", "P_ni_J385_roar", "P_ni_J389_roar", "P_ni_J417_roar","P_ni_J551_arsu", "P_ni_J602_arsu", "P_ni_J603_arsu", "P_ni_J614_arsu", "P_ni_J617_arsu","P_ni_80555_arsu","P_ni_86072_arsu", "P_ni_80684_arsu", "P_ni_80802_arsu", "P_ni_80874_arsu", "P_ni_85430_arsu","P_ni_T14543_suta",  "P_ni_T9076_suta", "P_ni_T16698_suta","P_ni_T10967_suta",  "P_ni_T11888_suta","P_ni_T10204_suta","P_ni_A15120_suta", "P_ni_77876_suta", "P_ni_78155_suta","P_ni_85721_suta",],
"Para": ["P_ni_T1642_pa", "P_ni_T18703_pa","P_ni_T12345_pa", "P_ni_T12854_pa","P_ni_T11193_pa", "P_ni_T11222_pa","P_ni_T10673_pa", "P_ni_T10940_pa","P_ni_A7066_pa", "P_ni_A14342_pa", "P_ni_A15277_pa",]
}


minmap4 = {i: 0.5 for i in imap}
```

In [47]:
```
# load the snp data into distance tool with arguments
from ipyrad.analysis.distance import Distance
dist = Distance(
    data=data,
    imap=imap,
    minmap=minmap4,
    mincov=0.5,
    impute_method="sample",
    subsample_snps=False,
)
dist.run()
```

```
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 678269
Filtered (minmap): 878334
Filtered (combined): 887235
Sites after filtering: 360453
Sites containing missing values: 341178 (94.65%)
Missing values in SNP matrix: 3343773 (12.21%)
Imputation: 'sampled'; (0, 1, 2) = 55.7%, 5.9%, 38.4%
```

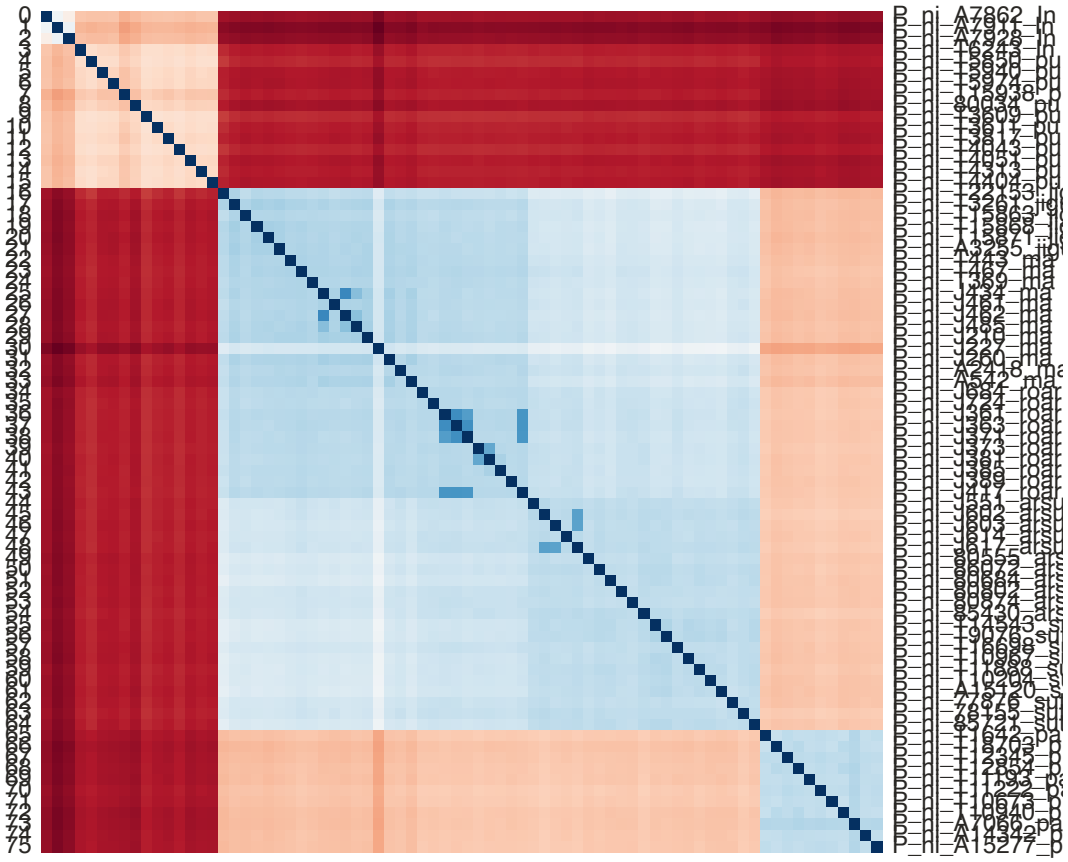
In [48]:
```
# save to a CSV file
dist.dists.to_csv("P_ni_distances_12Jan2022.csv")

# save to a CSV file with no labels (eems style)
dist.dists.to_csv(
    "P_ni_distances_eems_12Jan2022.csv",
    header=None,
    index=False,
    sep=" ",
)
```

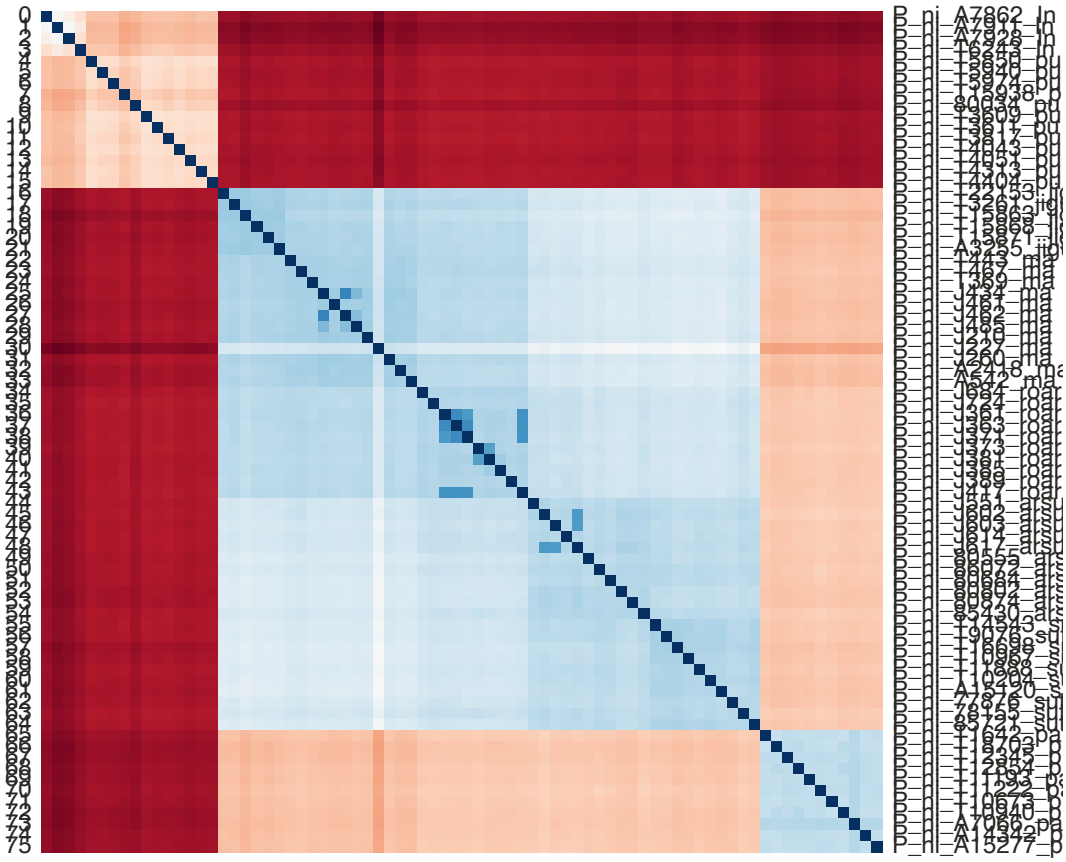
In [49]:
```
imap = {
#"ref": ["reference"],
"Inam": ["P_ni_A7862_In", "P_ni_A7911_In", "P_ni_A7928_In", "P_ni_T6243_
In",],
"Puru": ["P_ni_T5850_pu", "P_ni_T5940_pu", "P_ni_T5974_pu", "P_ni_T15938
_pu","P_ni_80034_pu", "P_ni_T3609_pu", "P_ni_T3611_pu",  "P_ni_T3817_pu"
, "P_ni_T4043_pu", "P_ni_T4051_pu", "P_ni_T4313_pu", "P_ni_T4404_pu"],
"JiGu": ["P_ni_T22153_jigu", "P_ni_T3261_jigu","P_ni_T15863_jigu", "P_ni
_T15868_jigu", "P_ni_T15871_jigu", "P_ni_A3255_jigu",],
"Mach": ["P_ni_T443_ma", "P_ni_T467_ma","P_ni_T369_ma","P_ni_J434_ma",
"P_ni_J461_ma", "P_ni_J462_ma", "P_ni_J485_ma", "P_ni_J210_ma", "P_ni_J2
27_ma", "P_ni_J260_ma", "P_ni_A2418_ma",  "P_ni_A542_ma",],
"Roar": ["P_ni_J684_roar", "P_ni_J724_roar","P_ni_J361_roar", "P_ni_J363
_roar", "P_ni_J371_roar", "P_ni_J373_roar", "P_ni_J381_roar", "P_ni_J385
_roar", "P_ni_J389_roar", "P_ni_J417_roar",],
"ArSu": ["P_ni_J551_arsu", "P_ni_J602_arsu", "P_ni_J603_arsu", "P_ni_J61
4_arsu", "P_ni_J617_arsu","P_ni_80555_arsu","P_ni_86072_arsu", "P_ni_806
84_arsu", "P_ni_80802_arsu", "P_ni_80874_arsu", "P_ni_85430_arsu",],
"SuTa": ["P_ni_T14543_suta",  "P_ni_T9076_suta","P_ni_T16698_suta","P_ni
_T10967_suta",  "P_ni_T11888_suta","P_ni_T10204_suta","P_ni_A15120_suta"
, "P_ni_77876_suta", "P_ni_78155_suta","P_ni_85721_suta",],
"Para": ["P_ni_T1642_pa", "P_ni_T18703_pa","P_ni_T12345_pa", "P_ni_T1285
4_pa","P_ni_T11193_pa", "P_ni_T11222_pa","P_ni_T10673_pa", "P_ni_T10940_
pa","P_ni_A7066_pa", "P_ni_A14342_pa", "P_ni_A15277_pa",]
}

minmap4 = {i: 0.5 for i in imap}
```

In [50]:
```
# load the snp data into distance tool with arguments
from ipyrad.analysis.distance import Distance
dist2 = Distance(
    data=data,
    imap=imap,
    minmap=minmap4,
    mincov=0.5,
    impute_method="sample",
    subsample_snps=False,
)
dist2.run()
```

```
Samples: 76
Sites before filtering: 1247688
Filtered (indels): 0
Filtered (bi-allel): 27379
Filtered (mincov): 678269
Filtered (minmap): 879565
Filtered (combined): 888415
Sites after filtering: 359273
Sites containing missing values: 339998 (94.63%)
Missing values in SNP matrix: 3141505 (11.51%)
Imputation: 'sampled'; (0, 1, 2) = 56.0%, 5.4%, 38.6%
```

In [51]:
```
# get list of concatenated names from each group
ordered_names = []
for group in dist.imap.values():
    ordered_names += group

# reorder matrix to match name order
ordered_matrix = dist.dists[ordered_names].T[ordered_names]

toyplot.matrix(
    ordered_matrix,
    bshow=False,
    tshow=False,
    rlocator=toyplot.locator.Explicit(
        range(len(ordered_names)),
        ordered_names,
));

# get list of concatenated names from each group
ordered_names = []
for group in dist2.imap.values():
    ordered_names += group

# reorder matrix to match name order
ordered_matrix = dist2.dists[ordered_names].T[ordered_names]

toyplot.matrix(
    ordered_matrix,
    bshow=False,
    tshow=False,
    rlocator=toyplot.locator.Explicit(
        range(len(ordered_names)),
        ordered_names,
));
```





top=structure assignments, bottom=a priori assignments and you can see they are nearly identical

In [ ]:

In [ ]: