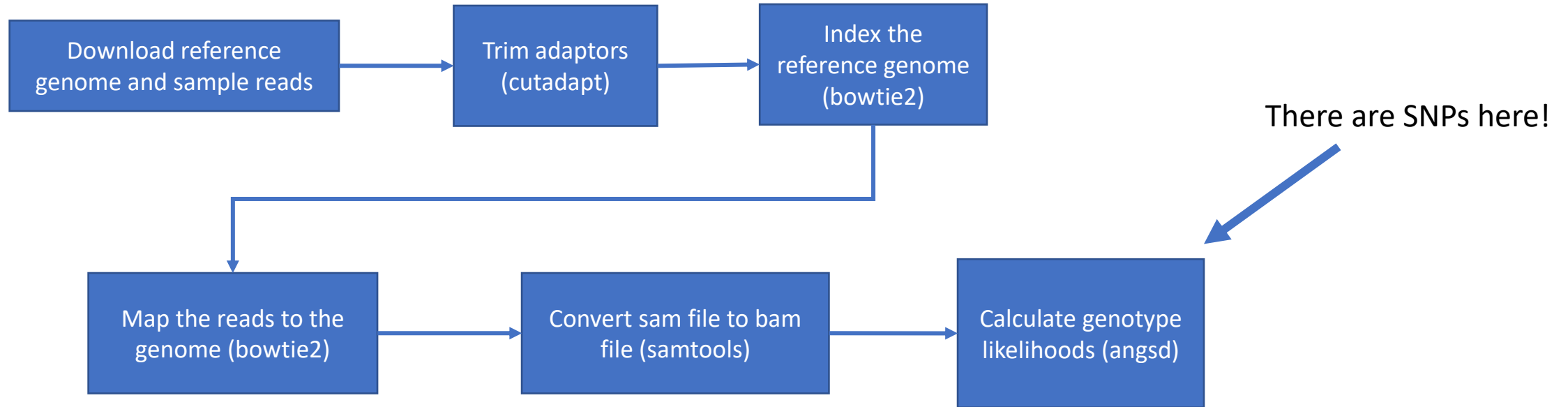




PCA for Population Genomics Data

Marine Genomics Week 6

From SNP identification



Now we can do things with those SNPs

Specifically, we can calculate allele frequencies and plot patterns of covariance between samples.

Principal Component Analyses

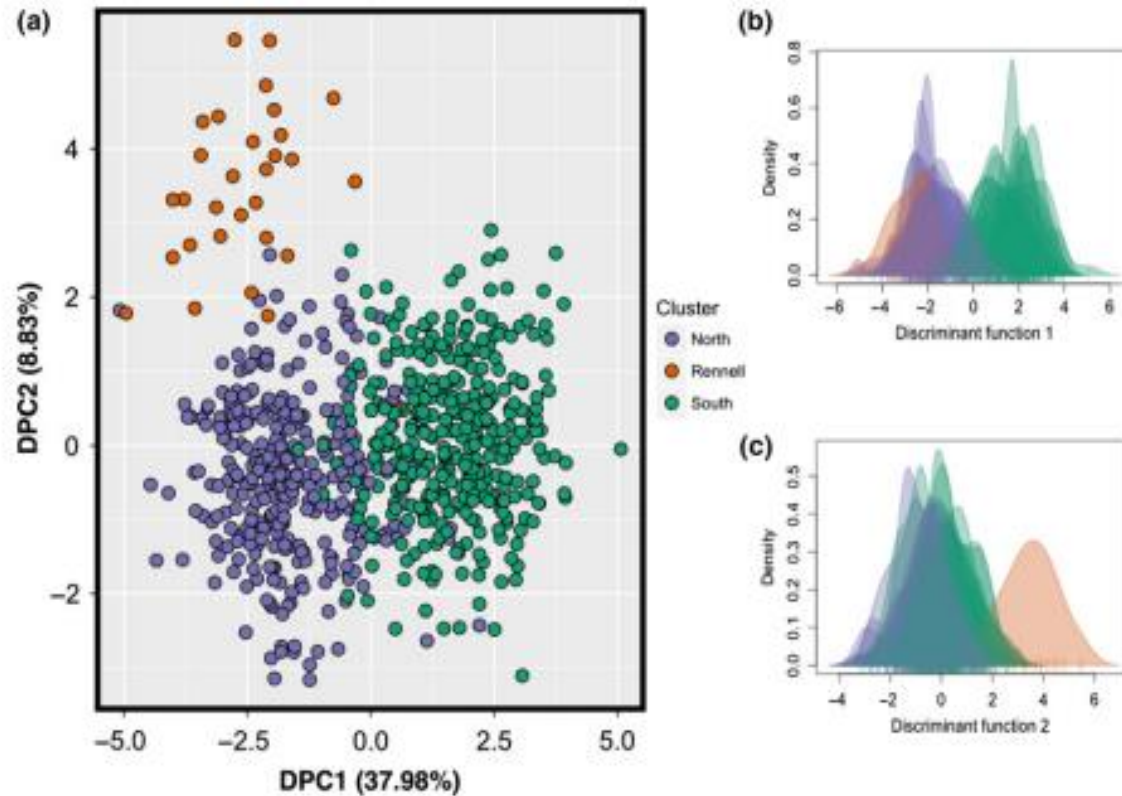


FIGURE 3 (a) Scatterplot showing the genetic clusters identified by DAPC with a priori sampling location information and density plots for (b) the first and (c) the second discriminant axes [Colour figure can be viewed at wileyonlinelibrary.com]

What is a PCA?

A method to reduce the dimensionality of large datasets

In population genetics data is typically:

Many SNPs for many individuals

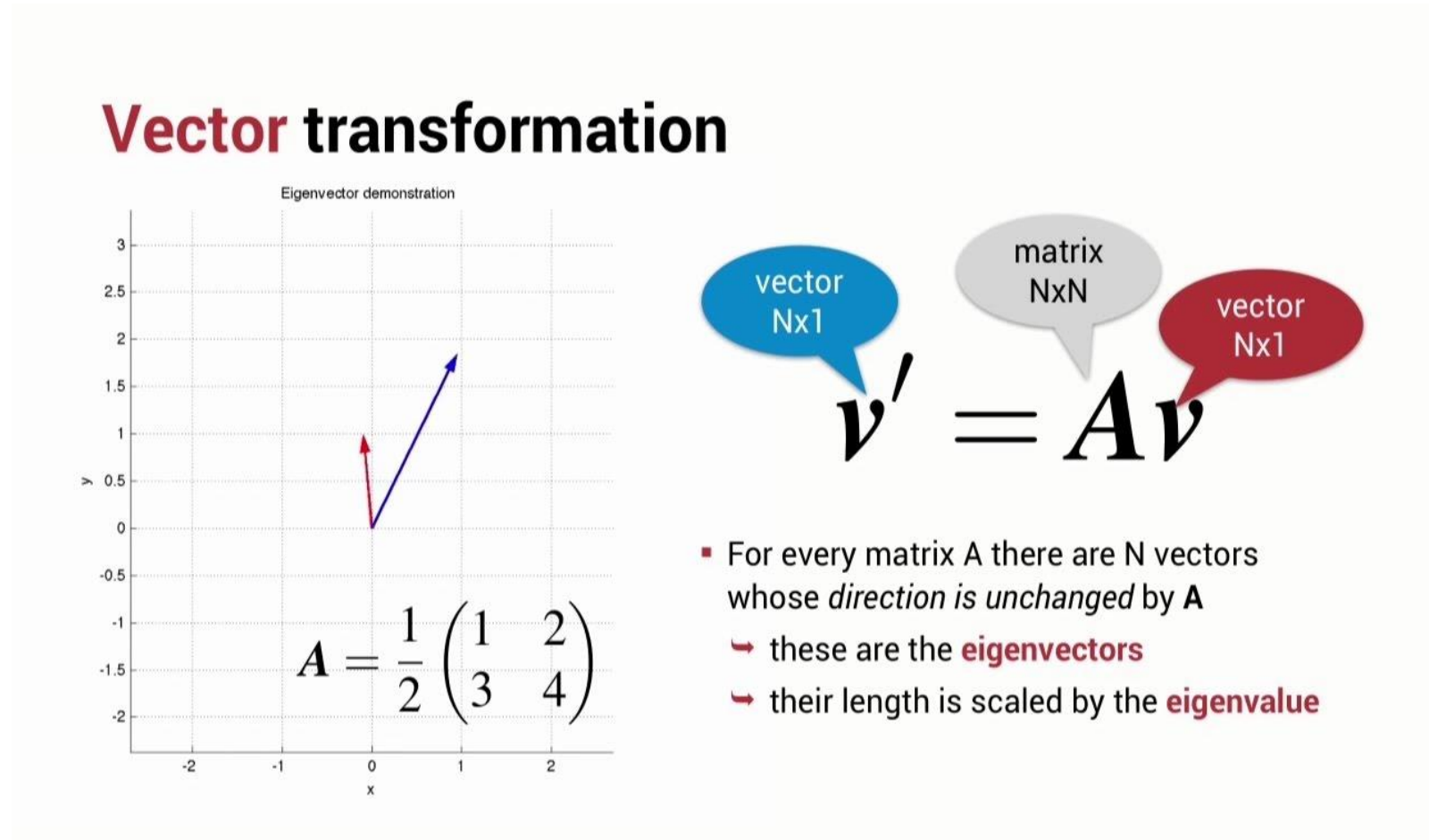
Steps involved in running a PCA

- Data standardization
 - PCA is sensitive to variances in the data
 - For example, if one site is much more variable than any other
 - Calculate a covariance matrix where (x, y, z) are individual samples
$$\begin{bmatrix} Cov(x, x) & Cov(x, y) & Cov(x, z) \\ Cov(y, x) & Cov(y, y) & Cov(y, z) \\ Cov(z, x) & Cov(z, y) & Cov(z, z) \end{bmatrix}$$
 - Compute the eigen vectors and eigen values of the covariance matrix
 - This is where we identify the principal components of the data!

Eigenvalues and eigenvectors

Basically:

An eigenvector points in the direction it is stretched by the transformation and the eigenvalue is how much it was stretched.

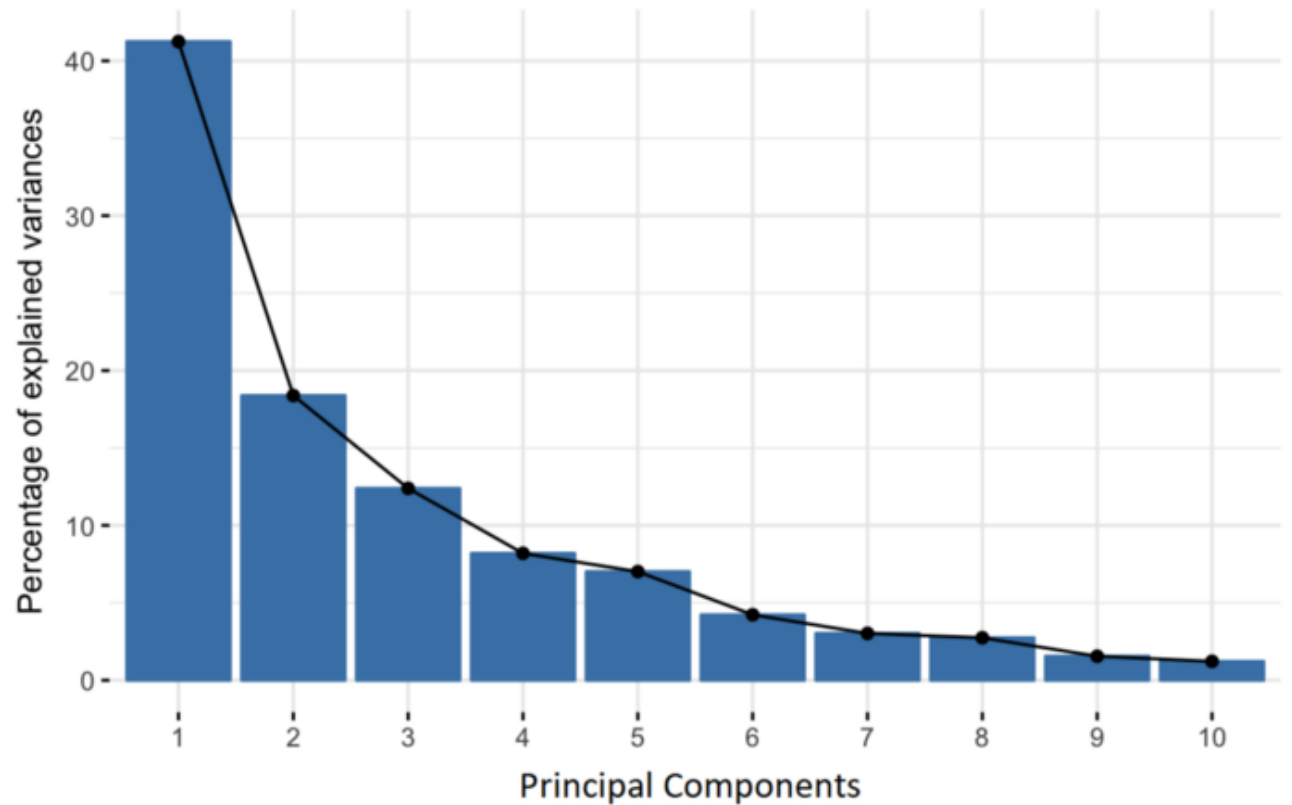


What is a principal component?

New variables constructed from a mixture of our initial variable.

A PCA will return to you the same dimension of data you gave it.

- For 10 individuals genotyped at 10 loci you will get 10 principal components
- But most of the information should be in the first few PCs




What data are we using this week?

ORIGINAL ARTICLE

WILEY **MOLECULAR ECOLOGY**

Asymmetric oceanographic processes mediate connectivity and population genetic structure, as revealed by RADseq, in a highly dispersive marine invertebrate (*Parastichopus californicus*)

Amanda Xuereb¹  | Laura Benestan² | Éric Normandeau² | Rémi M. Daigle¹ | Janelle M. R. Curtis³ | Louis Bernatchez² | Marie-Josée Fortin¹

We're using this subset:

15 individuals from 7 populations

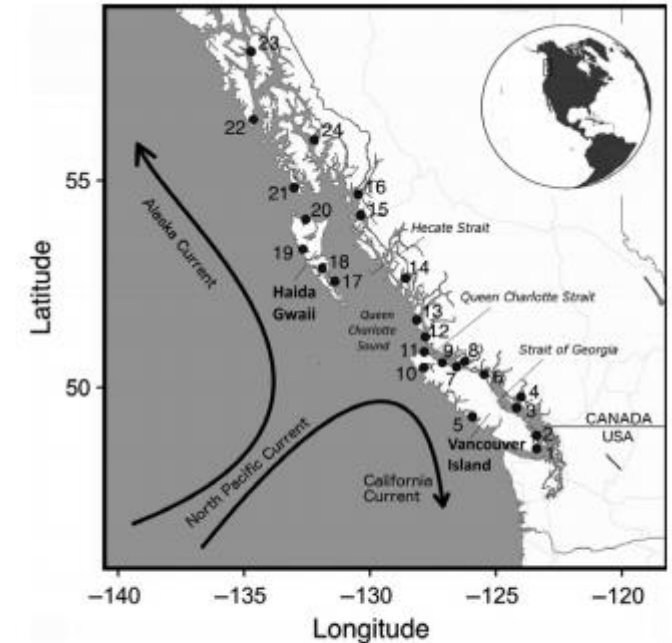
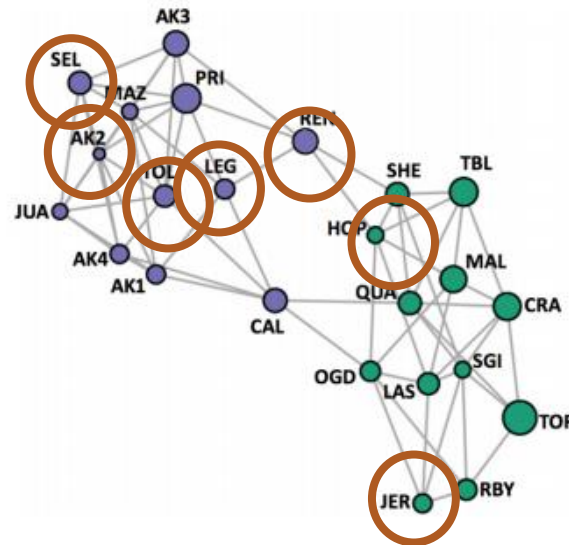


FIGURE 1 Map of sampling locations in coastal British Columbia (1–20) and southeastern Alaska (21–24). Site labels correspond with numbers in Table 1

Caveats

Our methods differ from those used in this paper, primarily for convenience and time (ours).

What they did prior to making the PCA plot

Many filtering steps

Separated neutral SNPs from “selected” SNPs