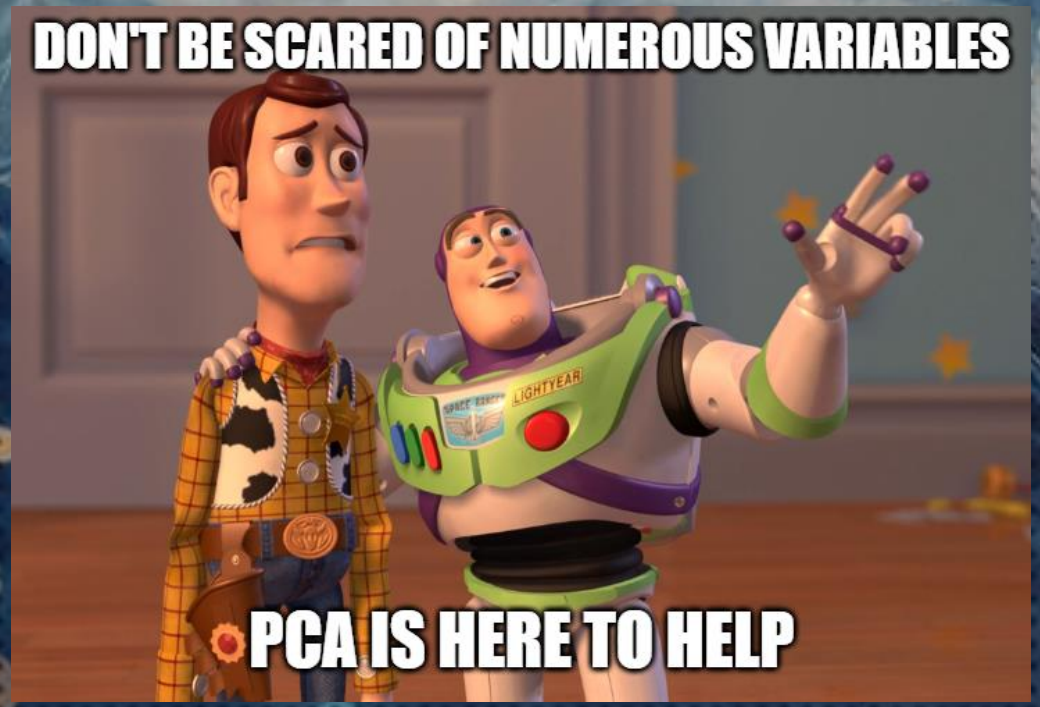
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

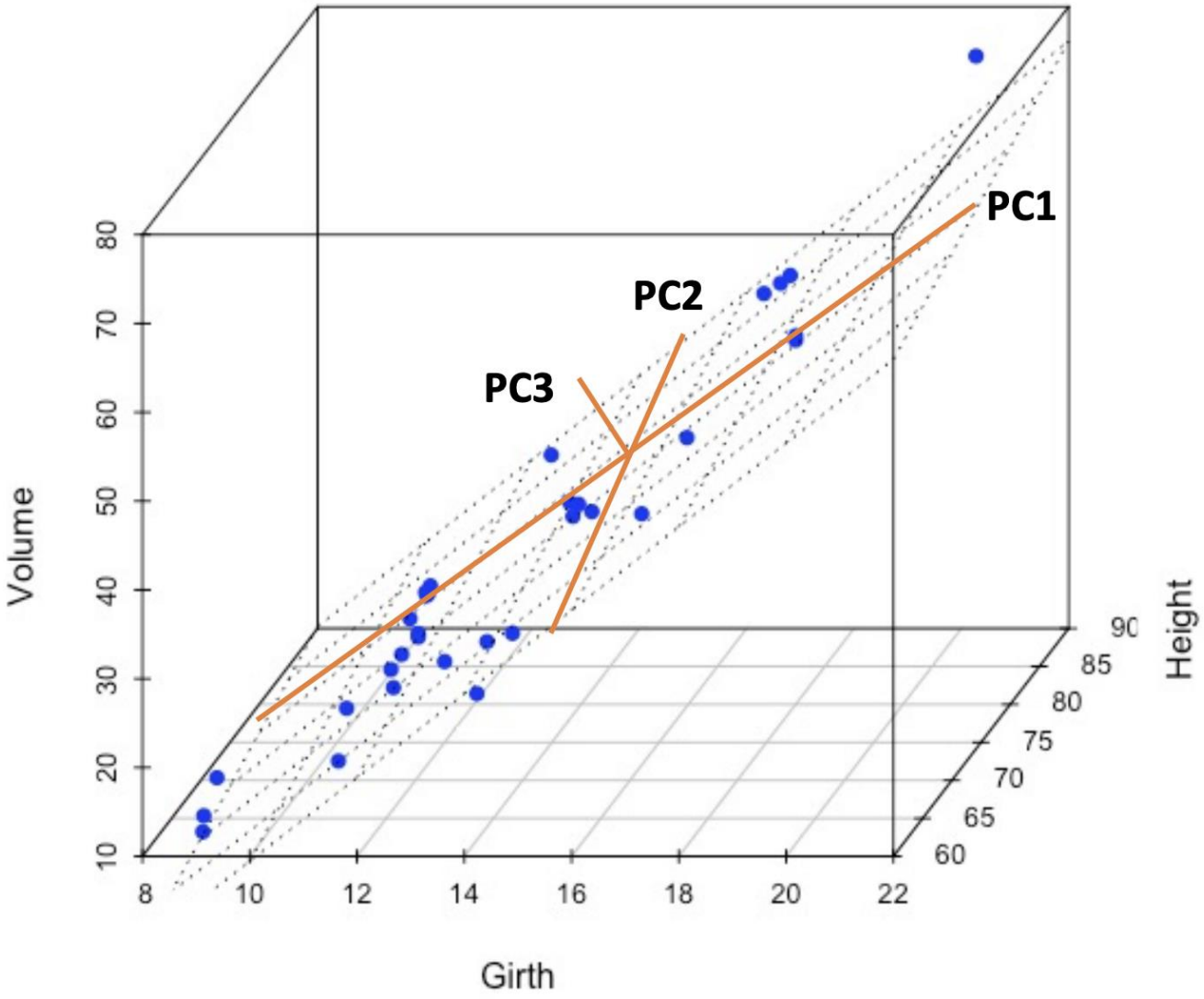
Dimension reduction:
Principal components analysis

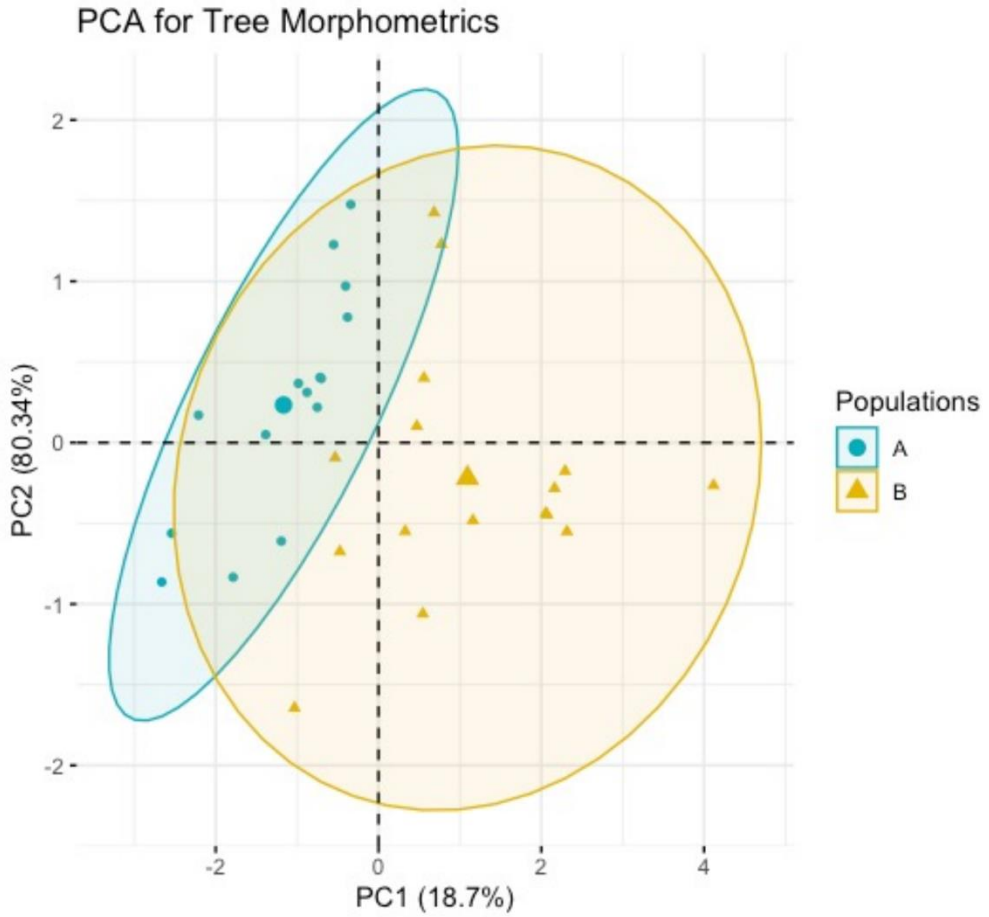


Terminology cheat sheet

1. **Variance**: A measure of the spread of a set of data points. In PCA, it is important because the first principal component is the direction along which the variance of the data is maximized.
2. **Covariance**: A measure of how much two random variables change together. In PCA, the covariance matrix expresses the correlation between the different variables.
3. **Eigenvalues and Eigenvectors**: These come from the covariance matrix. An eigenvector specifies a direction in the multidimensional space, while the eigenvalue corresponds to the magnitude that indicates the significance of the eigenvector (the amount of variance explained).
4. **Principal Components (PCs)**: New variables that are constructed as linear combinations of the initial variables. These are the eigenvectors of the covariance matrix, and they define the new space in which the data is now expressed.
5. **Dimensionality Reduction**: PCA reduces the dimensionality of the data set by retaining the principal components with the largest eigenvalues and ignoring those with smaller eigenvalues.
6. **Scree Plot**: A plot that shows the eigenvalues in a descending order versus the number of each eigenvalue. It's used to determine how many principal components should be retained.
7. **Loadings**: The weights applied to the original variables to obtain the principal components. They can be interpreted as the coefficients of the linear combinations that form the PCs.
8. **Orthogonal Transformation**: A type of linear transformation that preserves the geometric structure of the data, including distances and angles between points.
9. **Score**: The actual value of the principal components for each data point.

Dimension reduction (e.g., Principal component analysis)





Article

Current Biology

A New Perspective on Ecological Prediction Reveals Limits to Climate Adaptation in a Temperate Tree Species

Highlights

- NSC stores in *P. trichocarpa* are heritable and locally adapted to climate
- Despite species-wide evolutionary potential in storage, populations are at risk
- Northern populations lack the genomic variants to evolve with climate change
- Southern populations have genomic diversity but face intense selective pressures

Authors

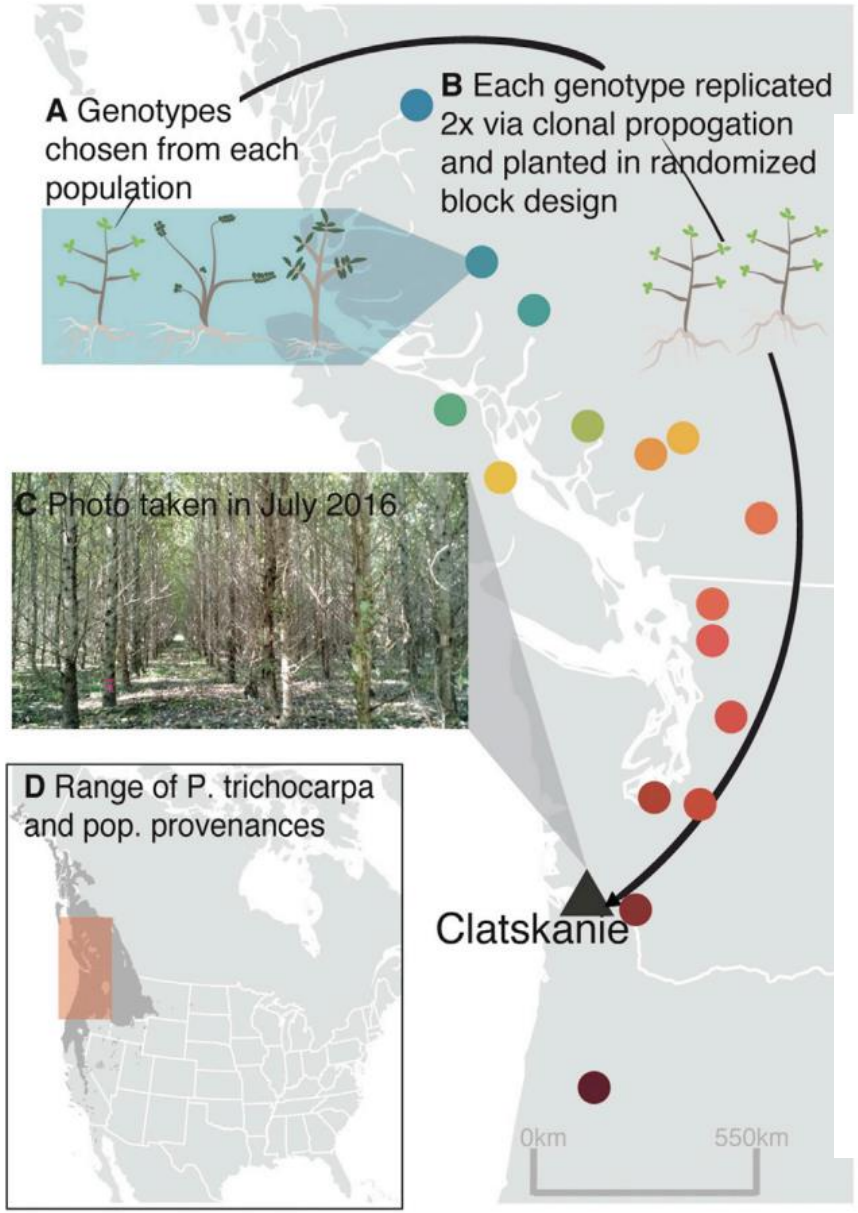
Meghan Blumstein, Andrew Richardson, David Weston, Jin Zhang, Wellington Muchero, Robin Hopkins

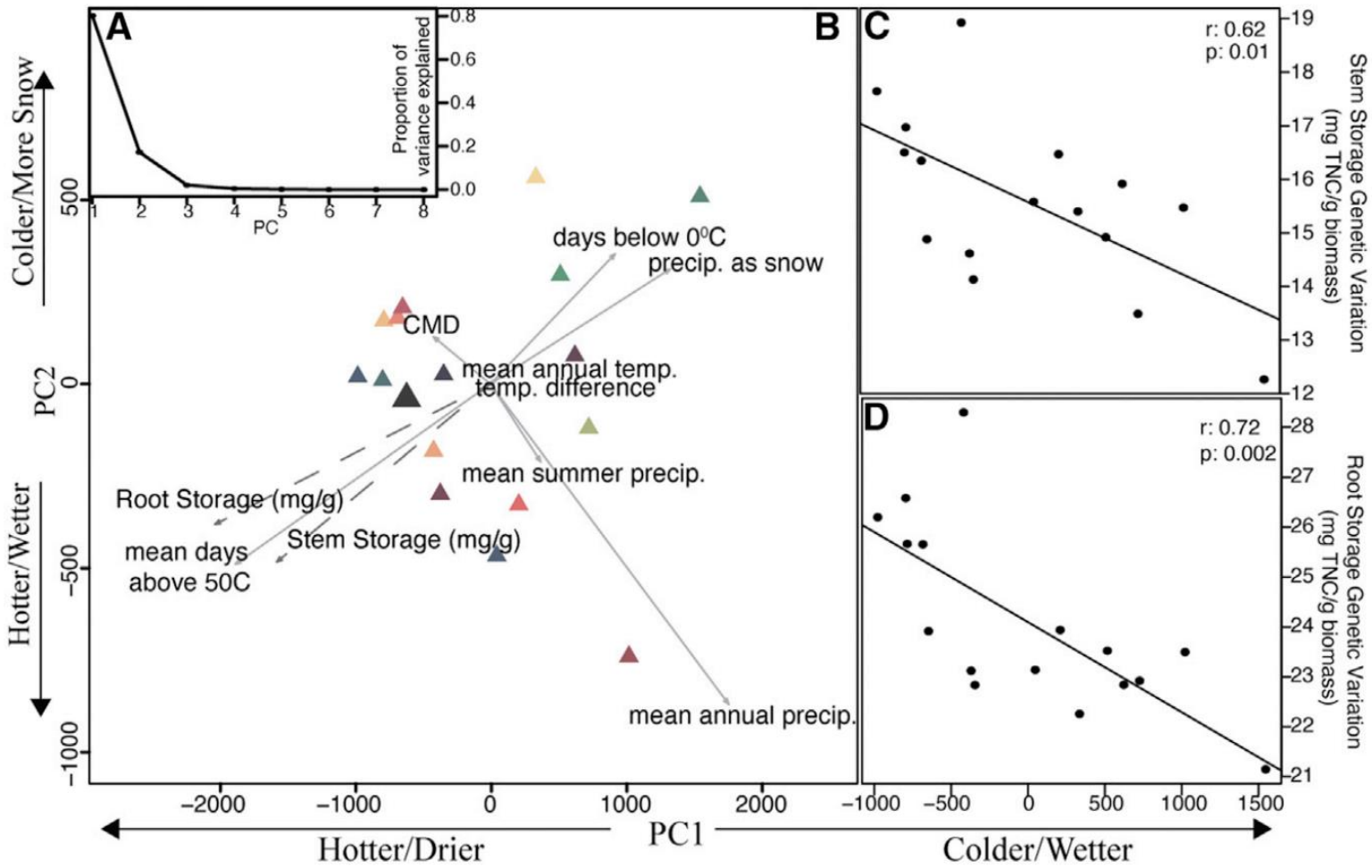
Correspondence

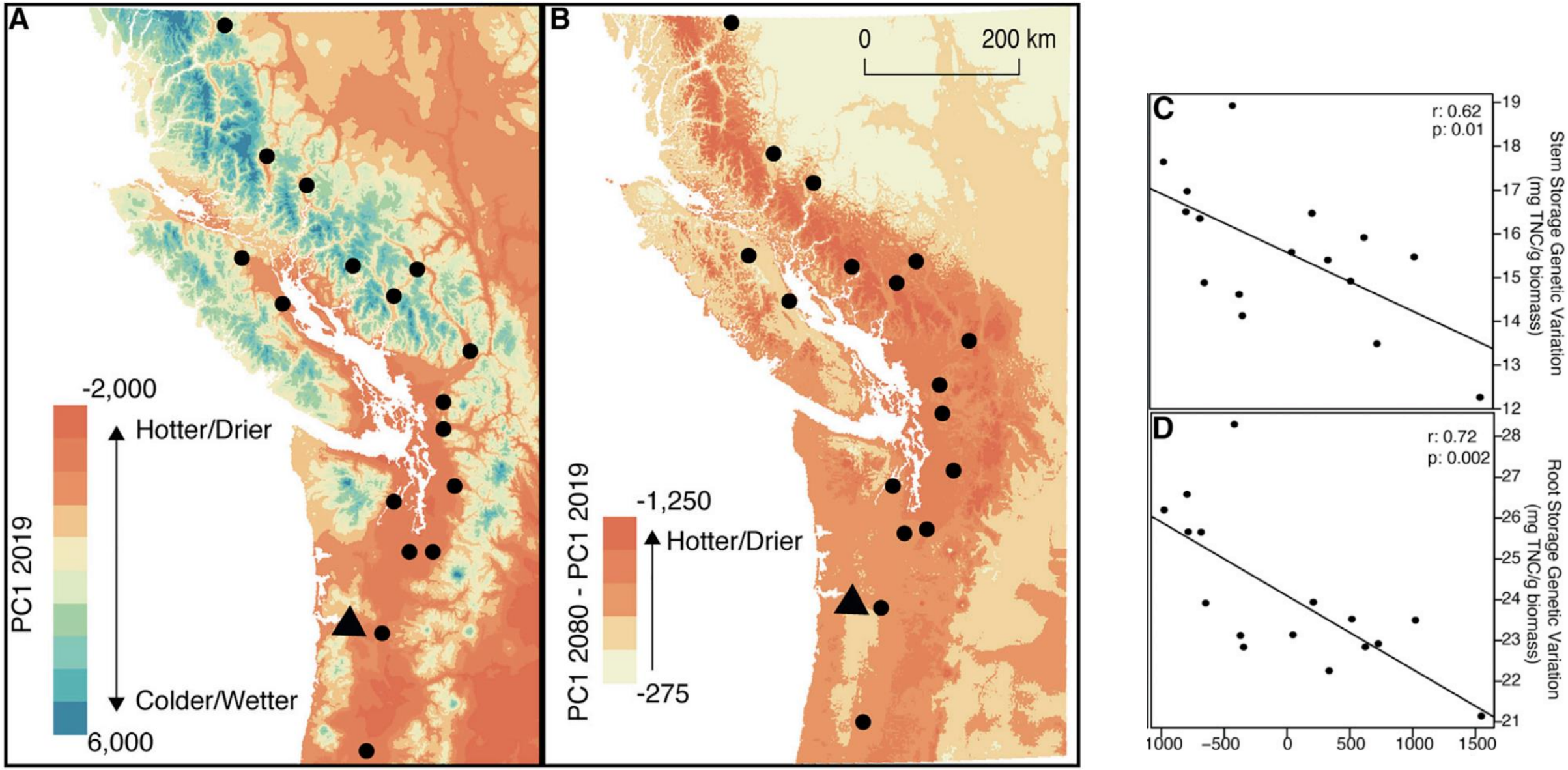
blumstein@fas.harvard.edu

In Brief

Blumstein et al. show variation in NSC storage in trees is heritable and locally adapted to climate. Despite species-wide evolutionary potential in storage, some populations are at risk. Northern populations lack genomic variants associated with high storage, while southern populations have these variants but face intense selective pressures.


A Genotypes chosen from each population
B Each genotype replicated 2x via clonal propogation and planted in randomized block design
C Photo taken in July 2016
D Range of P. trichocarpa and pop. provenances
Clatskanie
0km
550km


A
B
C
D
Proportion of variance explained
PC
Colder/More Snow
Hotter/Wetter
PC2
days below 0°C
precip. as snow
CMD
mean annual temp.
temp. difference
mean summer precip.
Root Storage (mg/g)
Stem Storage (mg/g)
mean days above 50C
mean annual precip.
Hotter/Drier
PC1
Colder/Wetter
r: 0.62
p: 0.01
Stem Storage Genetic Variation (mg TNC/g biomass)
r: 0.72
p: 0.002
Root Storage Genetic Variation (mg TNC/g biomass)


A
B
0
200 km
-2,000
Hotter/Drier
PC1 2019
Colder/Wetter
6,000
PC1 2080 - PC1 2019
-1,250
Hotter/Drier
-275
C
r: 0.62
p: 0.01
Stem Storage Genetic Variation
(mg TNC/g biomass)
19
18
17
16
15
14
13
12
D
r: 0.72
p: 0.002
Root Storage Genetic Variation
(mg TNC/g biomass)
28
27
26
25
24
23
22
21
1000
−500
0
500
1000
1500

Research

New Phytologist

Interactive effects of water limitation and elevated temperature on the physiology, development and fitness of diverse accessions of *Brachypodium distachyon*

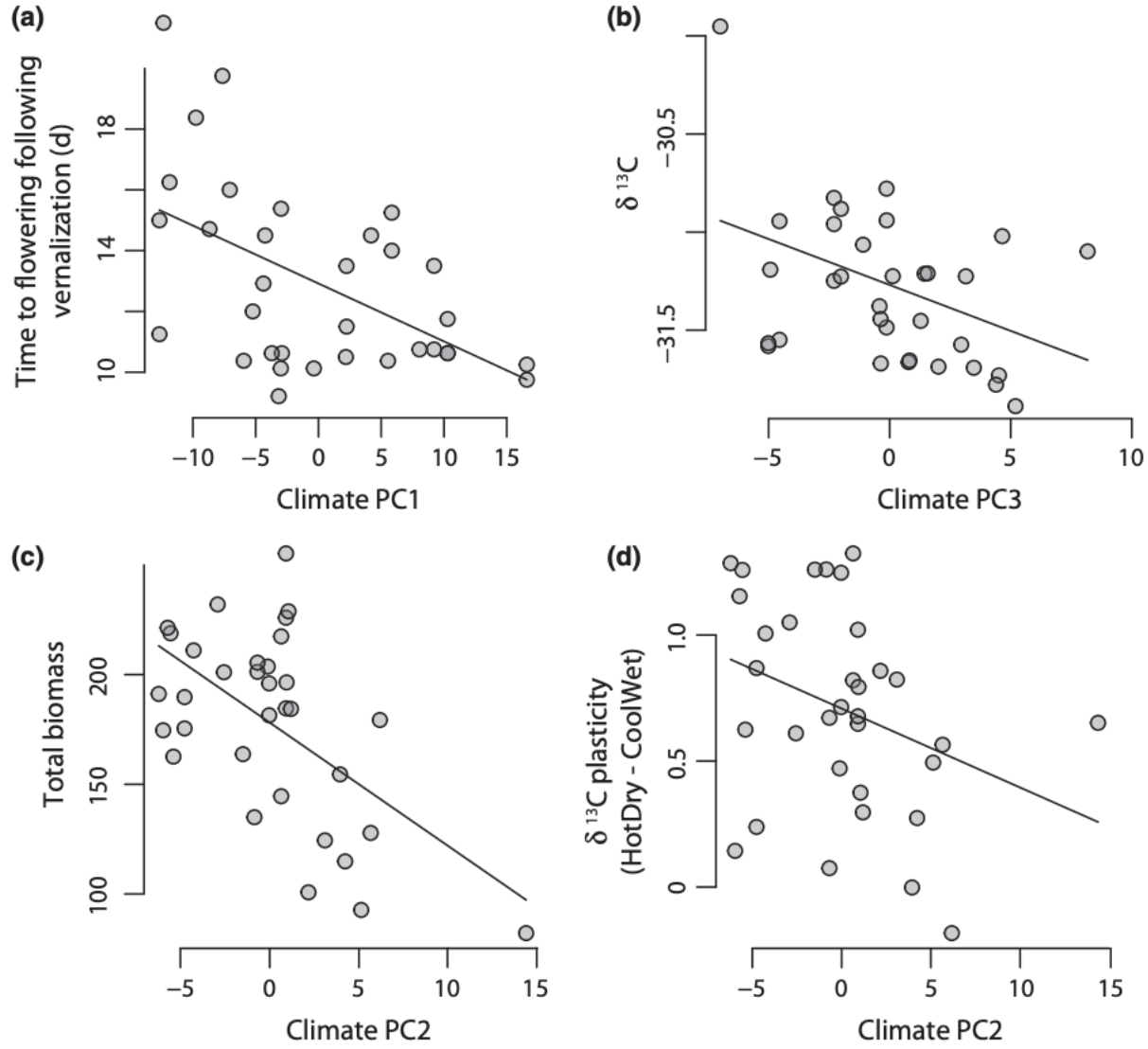
David L. Des Marais[1,4], Jesse R. Lasky[2], Paul E. Verslues[3], Trent Z. Chang[3] and Thomas E. Juenger[1]

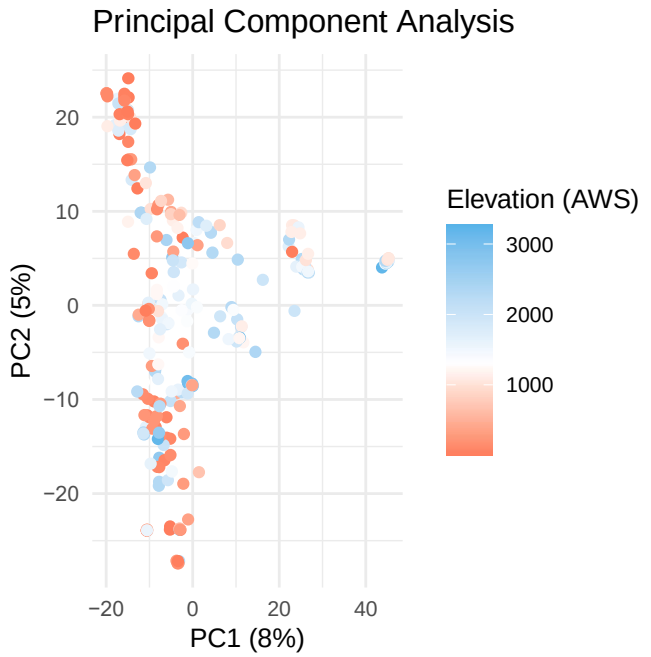
[1]Department of Integrative Biology and Institute for Cell and Molecular Biology, The University of Texas at Austin, Austin, TX 78712, USA; [2]Department of Biology, Pennsylvania State University, University Park, PA 16802, USA; [3]Institute of Plant and Microbial Biology, Academia Sinica, Taipei 11529, Taiwan; [4] Present address: Arnold Arboretum and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

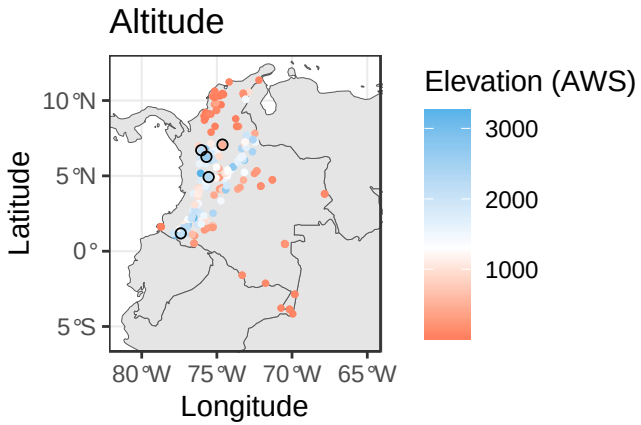
Table 1 Collection information for the natural accessions of *Brachypodium distachyon* used in the experiments

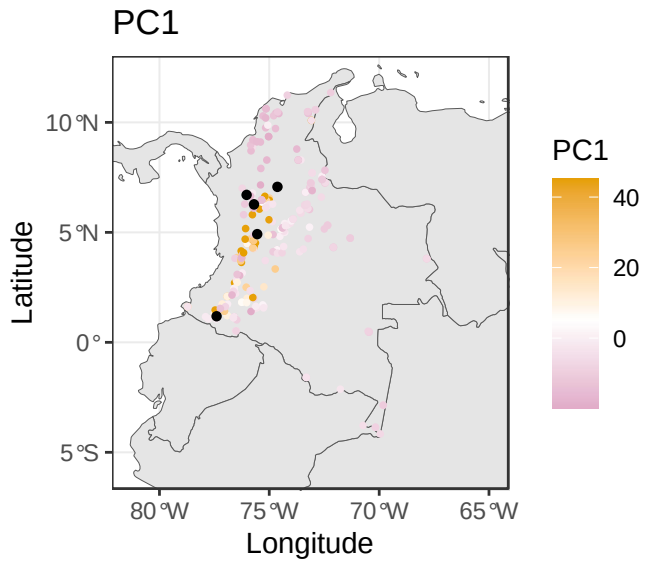
Line	Locale	Country	Latitude	Longitude
ABR2	Octon, Herault	France	43°39′18″N	3°18′12″E
ABR3	Huesca, Aisa	Spain	42.679°N	0.621°W
ABR4	Huesca, Aren	Spain	42°16′N	0°44′E
ABR5	Huesca, Jaca	Spain	42°33′N	0°33′W
ABR6	Navarra, Los Arcos	Spain	42.567°N	2.183°W
ABR7	Valladolid, Otero	Spain	39°59′57″N	4°30′46″W
ABR8	Siena, Italy	Italy	43°19′07″N	11°19′50″E
ABR9	Ljubljana	Slovenia	46°03′20″N	14°30′30″E
Adi-10	Adiyaman	Turkey	37°46′14.5″N	38°21′8.2″E
Adi-12	Adiyaman	Turkey	37°46′14.5″N	38°21′8.2″E
Adi-2	Adiyaman	Turkey	37°46′14.5″N	38°21′8.2″E
Bd1-1	Soma, Manisa	Turkey	39°11′N	27°37′E
Bd18-1	Kaman	Turkey	39°21′N	33°43′E
Bd2-3	Uncertain	Iraq		
Bd21	Near Irbil	Iraq	36°11′28″N	44°0′33″E
Bd21-3	Near Irbil	Iraq	36°11′28″N	44°0′33″E
Bd3-1	Uncertain	Iraq		
Bd30-1	Dilar, Spain	Spain	37°4′N	03°36′W
BdTR10c		Turkey	37°46′41.64″N	31°53′5.68″E
BdTR11g		Turkey	41°25′17.86″N	27°28′36.81″E
BdTR11i		Turkey	39°44′17.39″N	28°2′24.71″E
BdTR12c		Turkey	39°44′53.45″N	34°39′1.15″E
BdTR13A		Turkey	39°45′23.35″N	32°25′56.46″E
BdTR1i		Turkey	38°5′35.03″N	28°34′59.02″E
BdTR2b		Turkey	40°4′55.55″N	31°19′52.01″E
BdTR2g		Turkey	40°23′37.13″N	32°59′7.32″E
BdTR3c		Turkey	36°46′58.92″N	32°57′46.71″E
BdTR5i		Turkey	40°23′37.13″N	32°59′7.32″E
BdTR9k		Turkey	39°45′10.62″N	30°47′19.07″E
Bis-1	Bismil	Turkey	37°52′35.6″N	41°0′54.3″E
Gaz-8	Gaziantep	Turkey	37°7′39.8″N	37°23′26.9″E
Kah-1	Kahta	Turkey	37°44′2.3″N	38°32′0.2″E
Kah-5	Kahta	Turkey	37°44′2.3″N	38°32′0.2″E
Koz-1	Kozluk	Turkey	38°9′8.2.6″N	41°36′34.8″E
Koz-3	Kozluk	Turkey	38°9′8.2.6″N	41°36′34.8″E

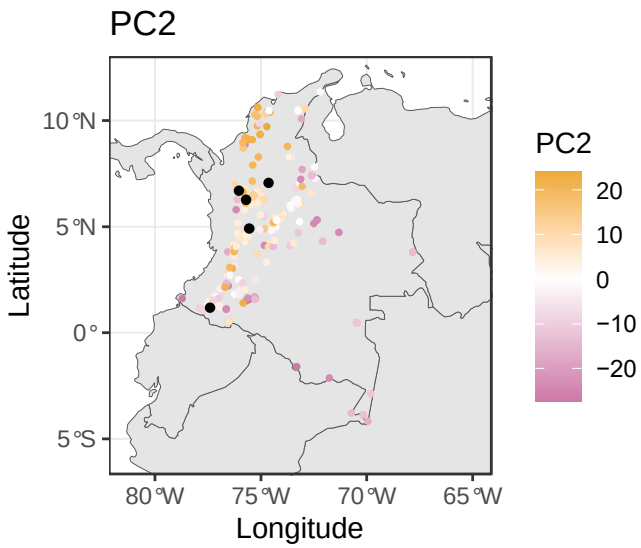
Data compiled from Jenkins *et al.* (2003), Vogel *et al.* (2009) and D. Garvin (pers. comm.).



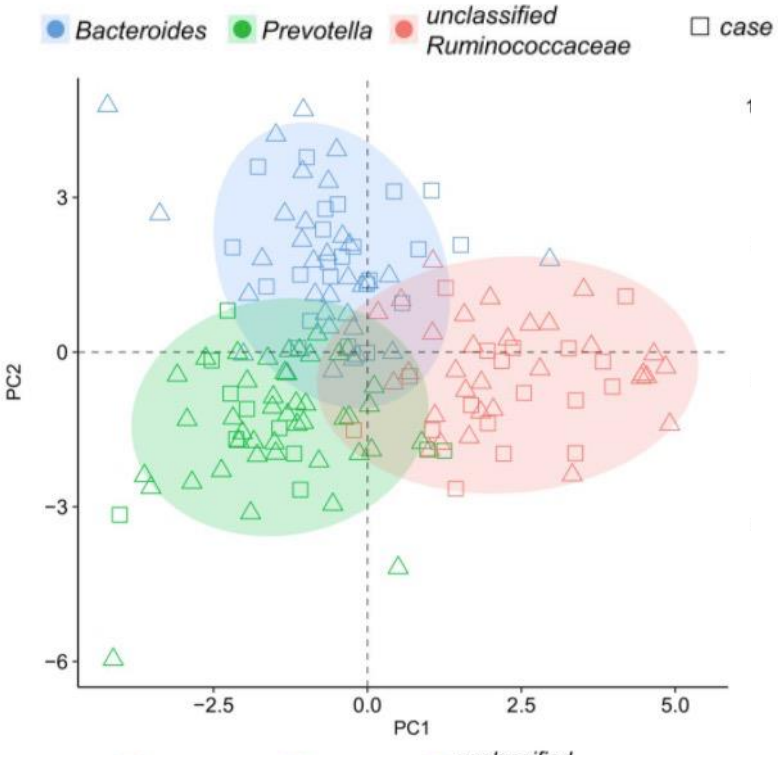

Principal Component Analysis
PC2 (5%)
PC1 (8%)
Elevation (AWS)
3000
2000
1000


Altitude
Latitude
Longitude
Elevation (AWS)
3000
2000
1000


PC1
Latitude
Longitude
PC1
40
20
0

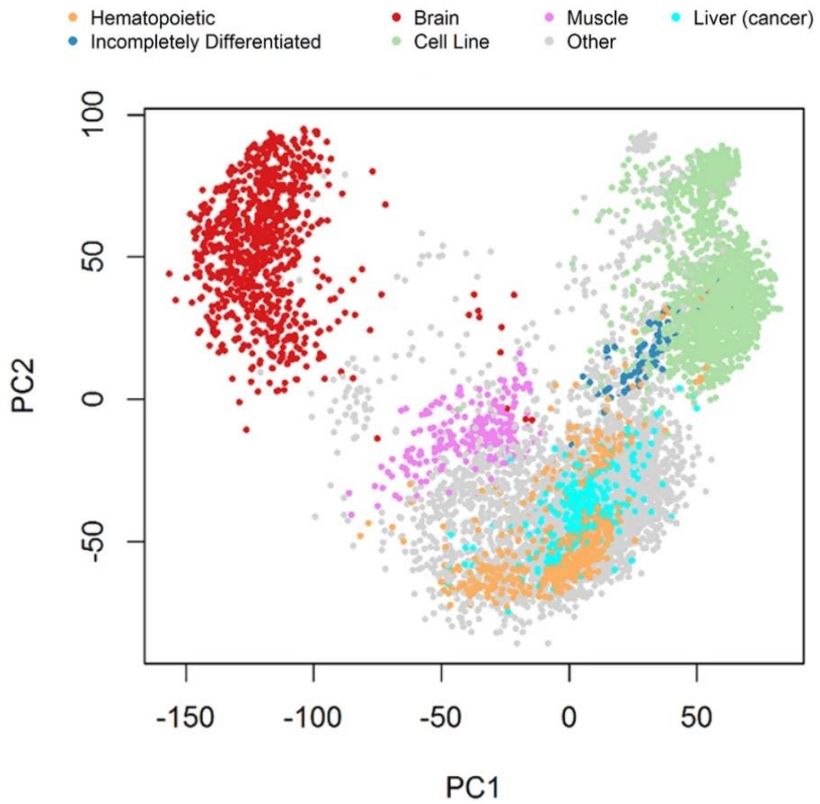

PC2
Latitude
Longitude
PC2
20
10
0
−10
−20

Example uses of PCA from Publications



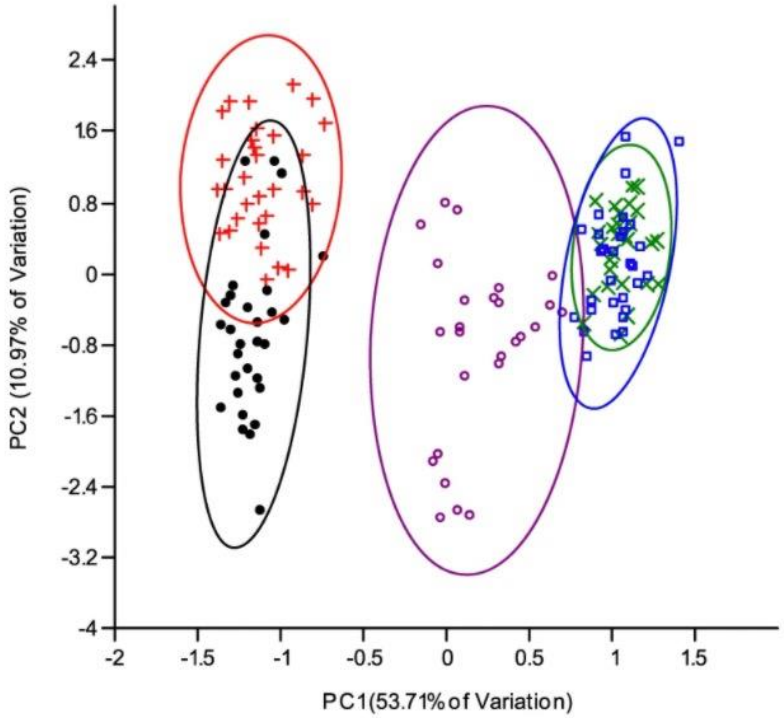
Hadizadeh et al. 2018. Gut. 67:778-779

Gut microbiome differs in humans with and without gut pain



Lenz et al. 2016 Sci Reports: 25696

Cell types show different gene expression



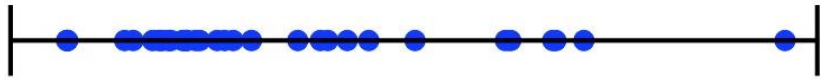
Okomoda et al. 2018 Sci Reports: 3827

Aquaculture fish stocks differ in morphology

Data Dimensionality

One-dimensional data

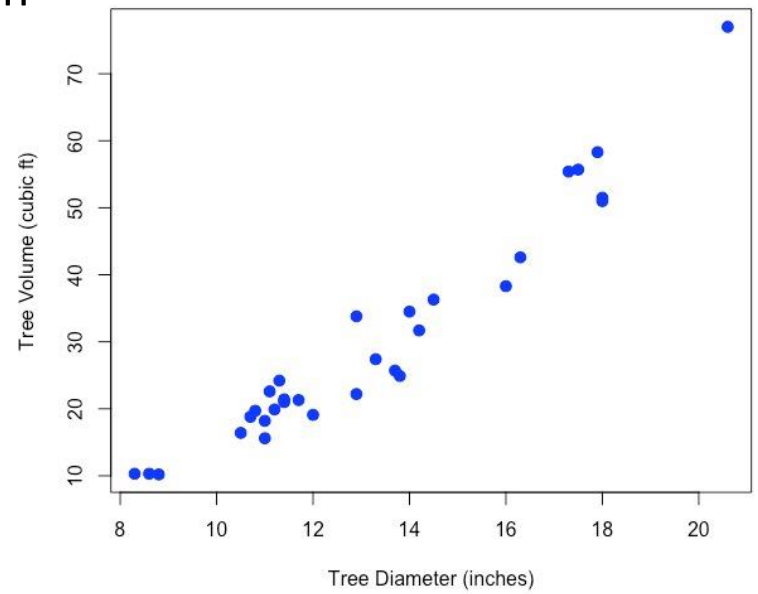
Points along a line



Tree Volume (cubic ft)

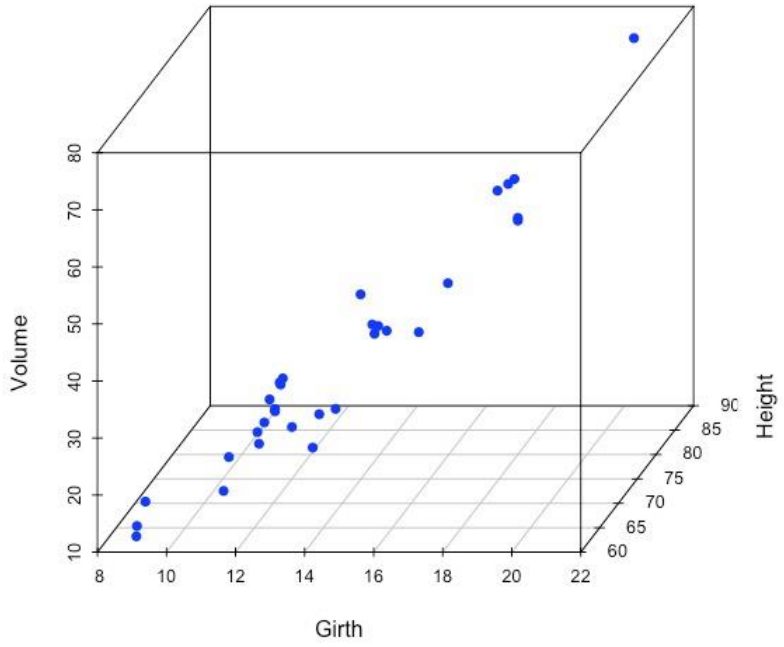
Two-dimensional data

Relationship between two variables
X, Y graph



Three-dimensional data

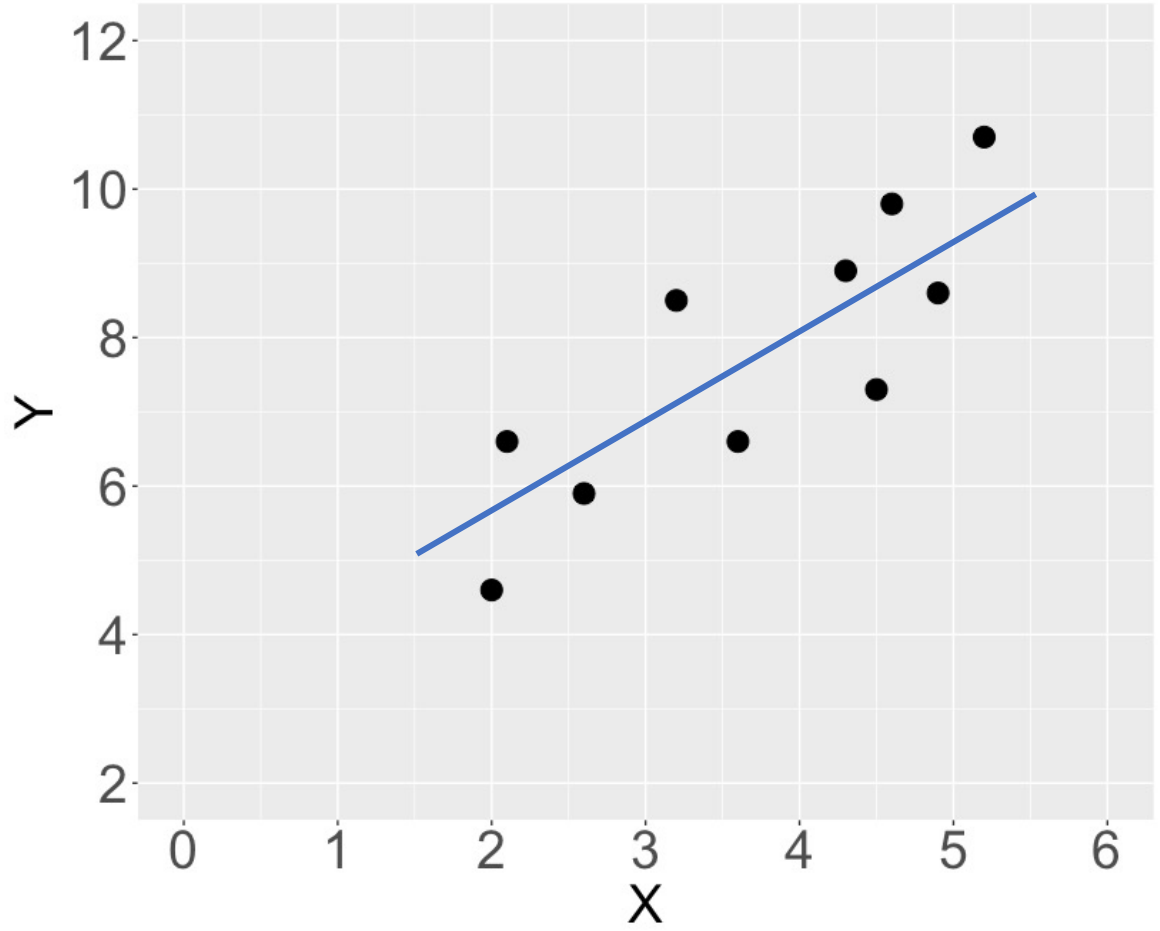
Relationship between 3 variables
X, Y, Z graph



Data can (and usually does) exist in many more than 3 dimensions, but we are unable to visualize it using a graph

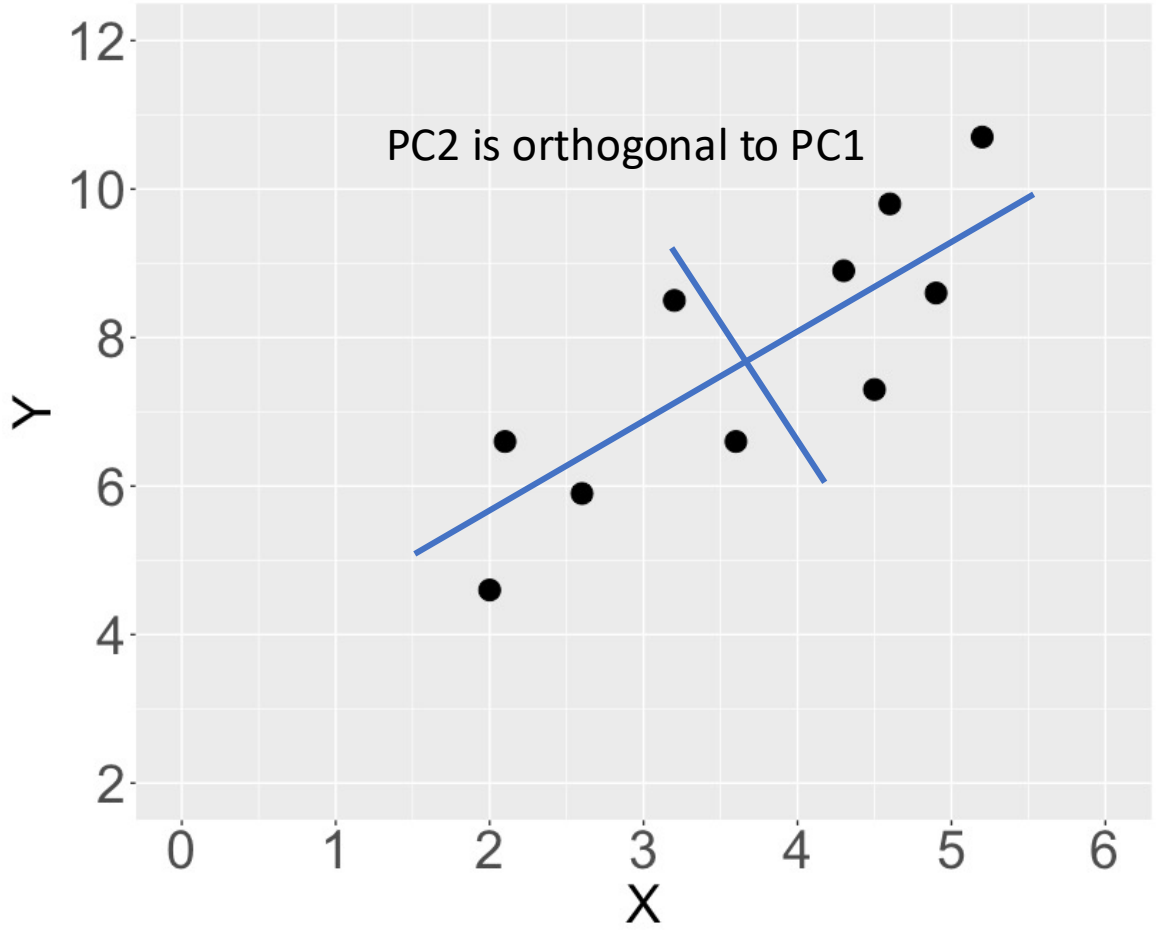
The Basic Concepts of PCA

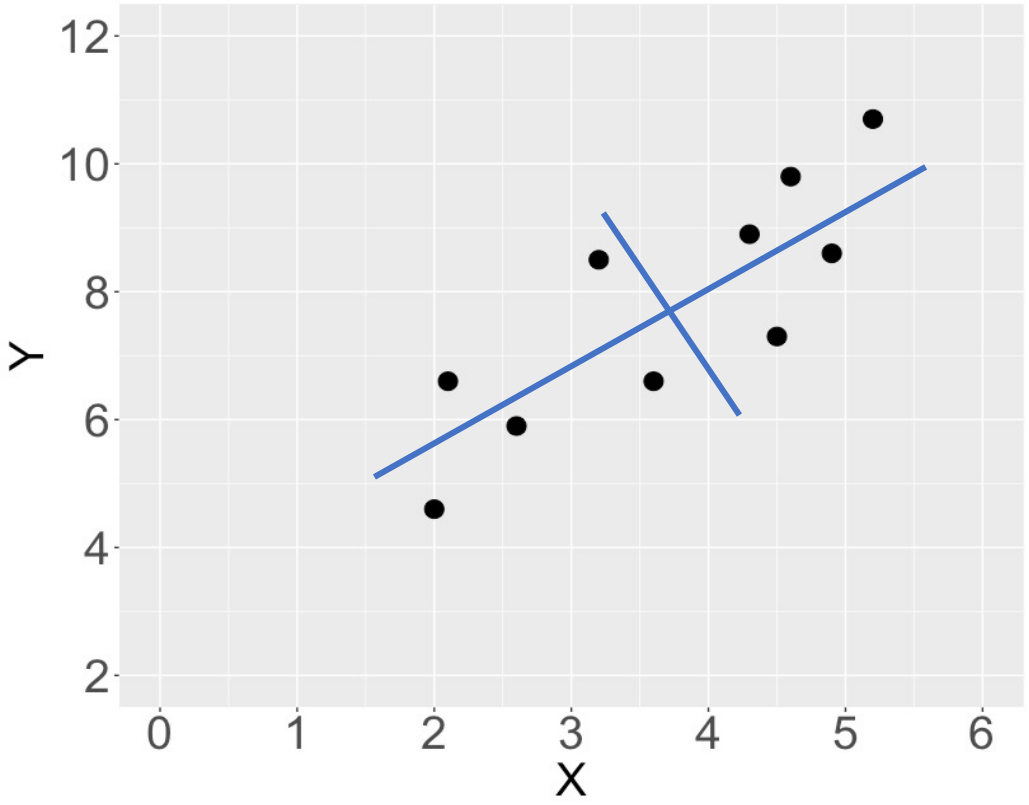
- Draw a line that passes through center of the cloud of data points that explains the maximum amount of variation in the data
- This is the 1st Principal Component



The Basic Concept

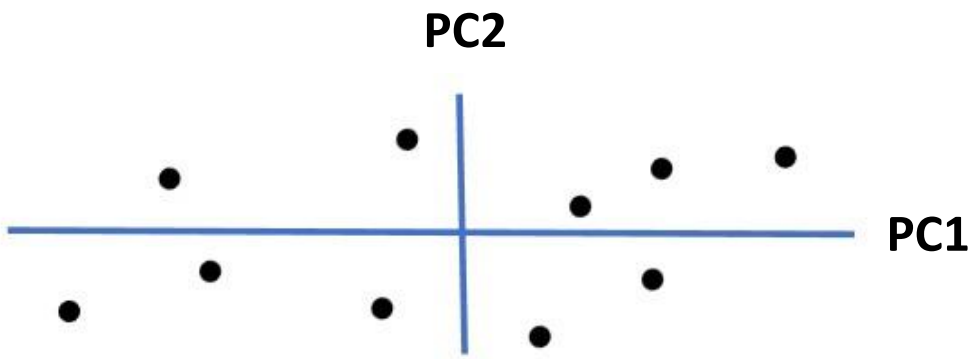
- Add a second line that is perpendicular to the first line
- This is the 2nd Principal Component
- This axis explains the remaining variation BUT it explains less variation than the first line





ROTATE

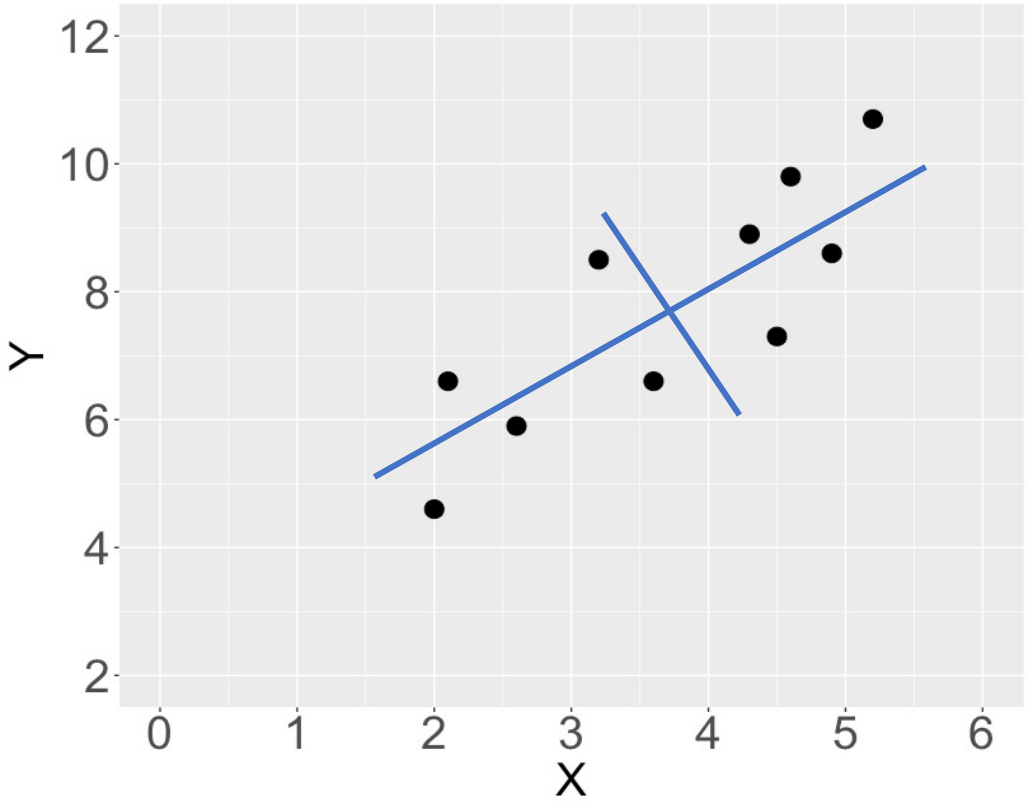
The lines we drew become the new x and y-axis for the data



New axes are called PC1 and PC2

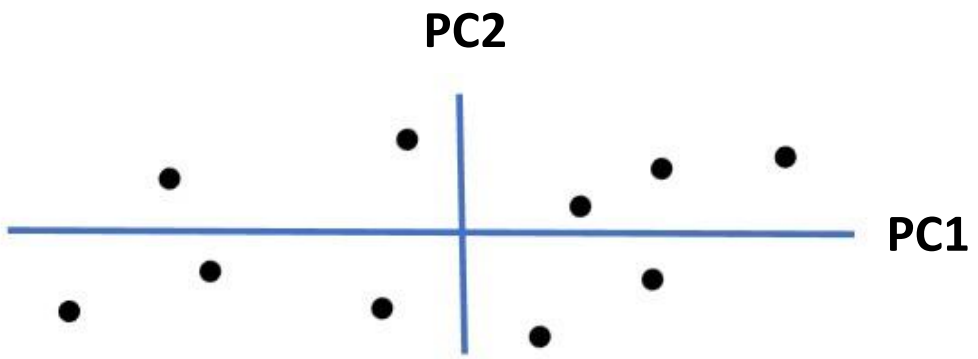
Notice how the spatial relationships among the data are unchanged after rotation

The two axes are uncorrelated because they are perpendicular



ROTATE

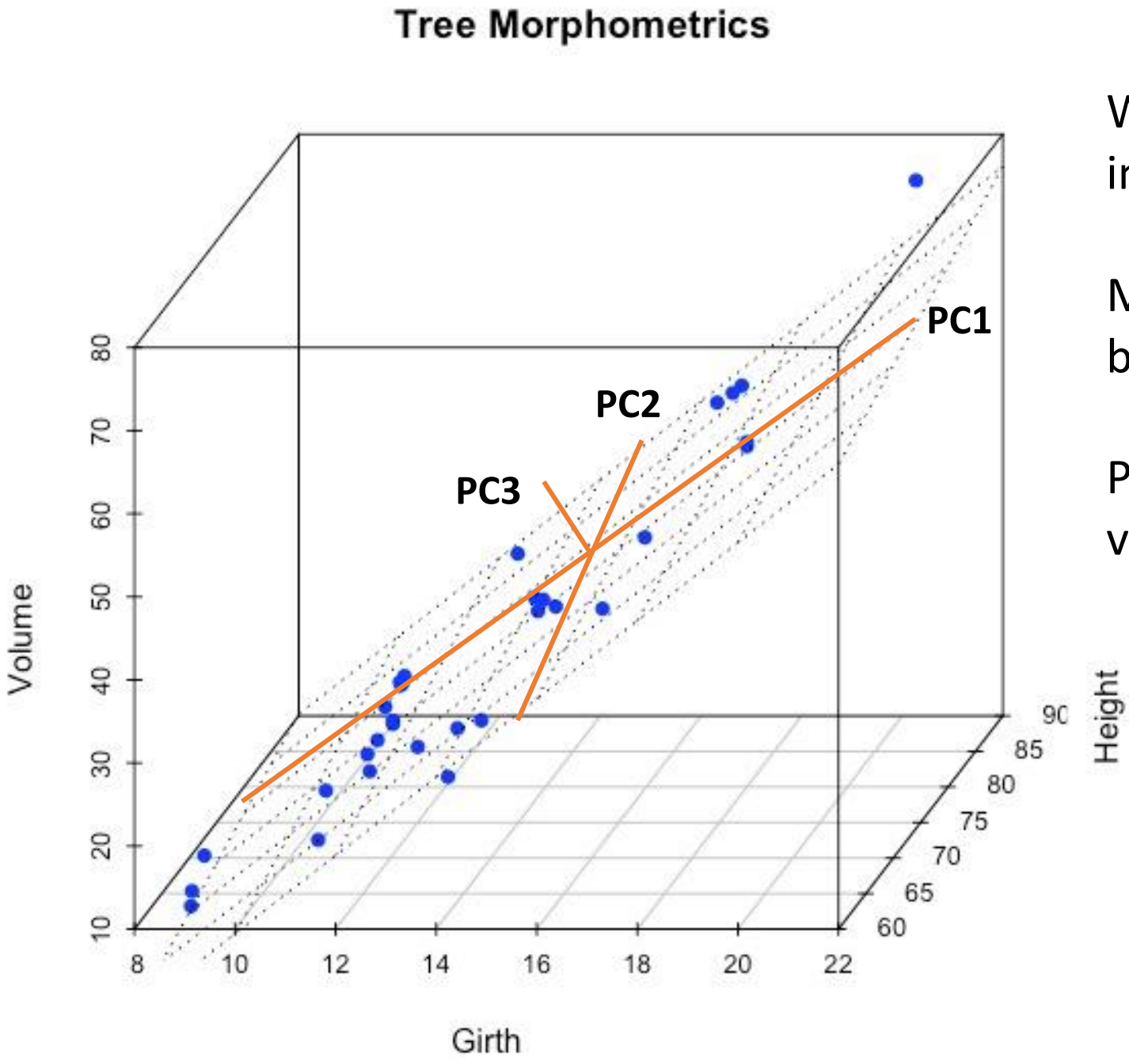
The lines we drew become the new x and y-axis for the data



New axes are called PC1 and PC2

Notice how the spatial relationships among the data are unchanged after rotation
The two axes are uncorrelated because they are perpendicular

PCA continues this process to make as many new axes as there are variables in the data
Each axis is orthogonal and statistically independent (aka uncorrelated with other axes)
Each subsequent axis explains less variation than the one before



We can also imagine this process in 3 dimensions

Most of the variation is explained by the plane of PC1 and PC2

PC3 explains very little of the variation

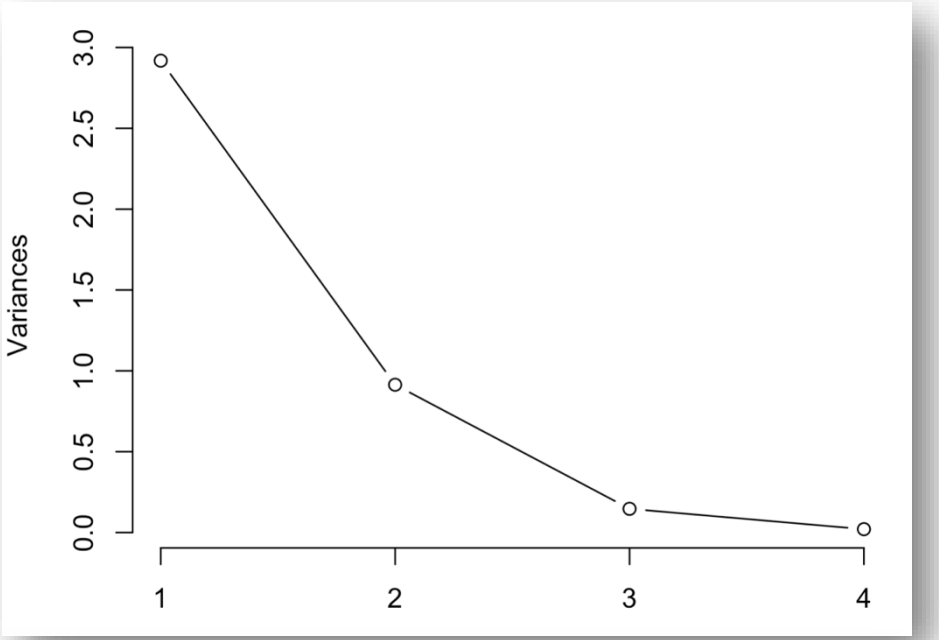
Amount of Variation Explained by New Axes

- New axes are called Principal Components or PCs
- Eigenvalue: summarizes how much variation each axis explains
- PC1 will always have the largest eigenvalue and the eigenvalues will decrease with each additional PC

Principal Component	Eigenvalue	% Variance Explained	Cumulative % Variance
PC1	2.41	80.34%	80.34%
PC2	0.56	18.72%	99.06%
PC3	0.03	0.94%	100%

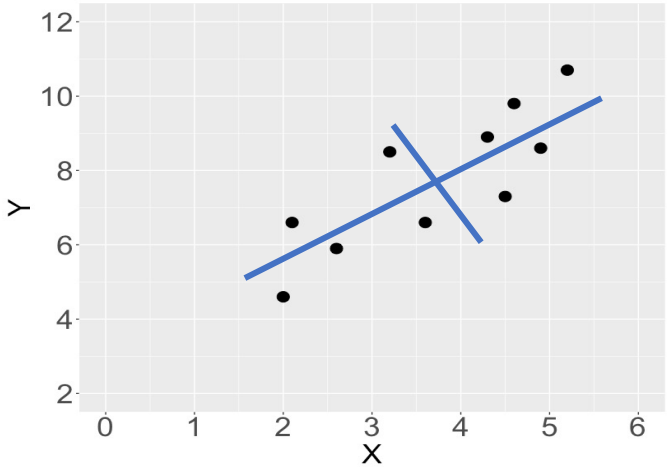
- The total variation explained by the PCA can be found by summing together all the eigenvalues
- We can calculate percentage of variation explained by each PC:
  = (eigenvalue for a PC / sum of all eigenvalues)*100

Example from Iris dataset in R:

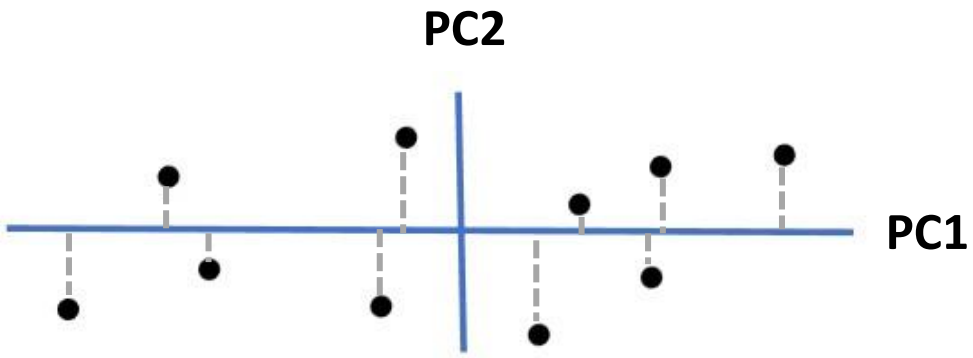



```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

Each PC is a Linear Combination of Variables



We can find the value for each point along PC1. It is a combination of the variables X and Y



Each Principal Component is a new variable that is linear combination of all the variables in the original data

Each original variable is given a loading (a coefficient) that describes its relationship with the new PC

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots a_{1p}X_p$$

Y_1 is the first principal component

$X_1, X_2, \ldots X_p$ are the variables in the original data

$a_{11}, a_{12}, \ldots a_{1p}$ are the loadings or coefficients for the variables on the PC

Each PC is a Linear Combination of Variables

For PC1

$$a_1 = \begin{bmatrix} 0.61 \\ 0.49 \\ 0.62 \end{bmatrix}$$

Tree Girth

Tree Height

Tree Volume

Loadings or coefficients are stored in eigenvectors

This example shows the eigenvectors for first 2 Principal Components for the tree morphometric data

Size of the loading reflects the strength of the relationship between variable and the PC

For PC2

$$a_2 = \begin{bmatrix} -0.41 \\ 0.87 \\ -0.28 \end{bmatrix}$$

Tree Girth

Tree Height

Tree Volume

You can think of the loading as a measurement of how much a particular variable contributed to building the new PC

The sign of a loading indicates whether a variable and a PC are positively or negatively correlated

Generate Scores or Coordinates for Each Sample on the New Axes

- The position of each data point in the new coordinate system created by the PCA is called the PC Score
- To find the PC score for any sample:
  - Multiply the loading by the value of the corresponding variable measurement for the sample
  - Sum all the products together

For PC1

$$a_1 = \begin{bmatrix} 0.61 \\ 0.49 \\ 0.62 \end{bmatrix}$$

Tree Girth

Tree Height

Tree Volume

For Tree Sample 1

standardized data * loading

Girth: -1.57685421 * 0.61 = -0.9618811

Height: -0.9416472 * 0.49 = -0.4614071

Volume: -1.20885469 * 0.62 = -0.7494899

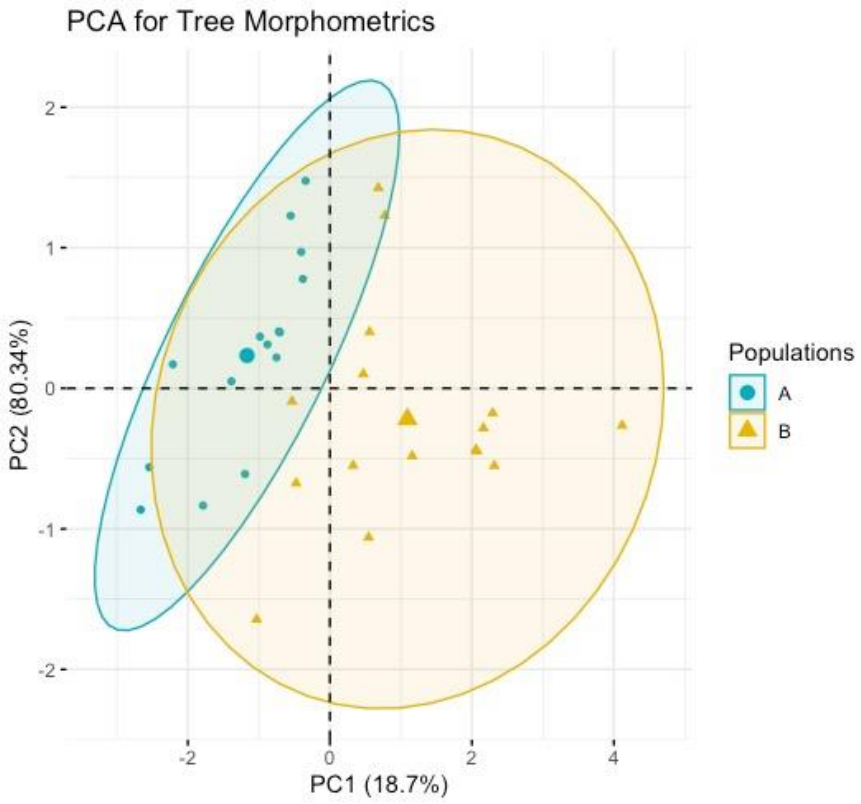
-2.172778

-2.17 is the PC1 score for sample 1

(Does not need to be calculated: you can extract this directly from R output from prcomp() function):

pca$x

Generate Scores or Coordinates for Each Sample on the New Axes



We can use the PC scores to plot the data in fewer dimensions -> this is a very useful exploratory tool

These plots are often included in publications

The points on the plots can be colored to highlight patterns in the data

We can also use the PC scores for hypothesis testing… more on this in later lectures

- For this example: one possibility is a t-test to compare the mean PC1 score of Population A to the mean PC1 score of Population B

“Iris” dataset in R

```
iris_pca<-prcomp(iris[,-5], scale=T)
pcs<-data.frame(iris_pca$x)
```

```
head(iris)
```

All observations are remapped to PCA coordinate space

```
head(pcs)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4

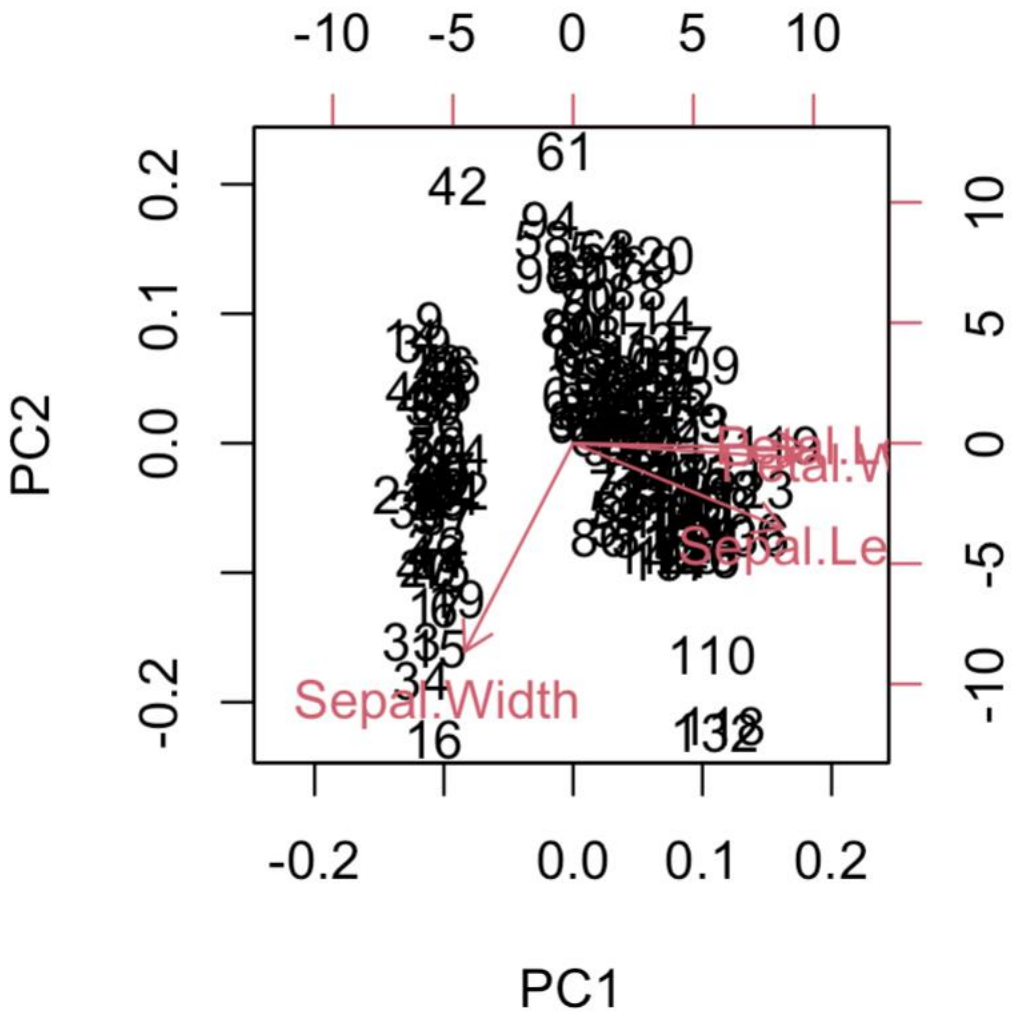
… +150 rows

	PC1	PC2	PC3	PC4
1	-2.257141	-0.4784238	0.12727962	0.024087508
2	-2.074013	0.6718827	0.23382552	0.102662845
3	-2.356335	0.3407664	-0.04405390	0.028282305
4	-2.291707	0.5953999	-0.09098530	-0.065735340
5	-2.381863	-0.6446757	-0.01568565	-0.035802870
6	-2.068701	-1.4842053	-0.02687825	0.006586116

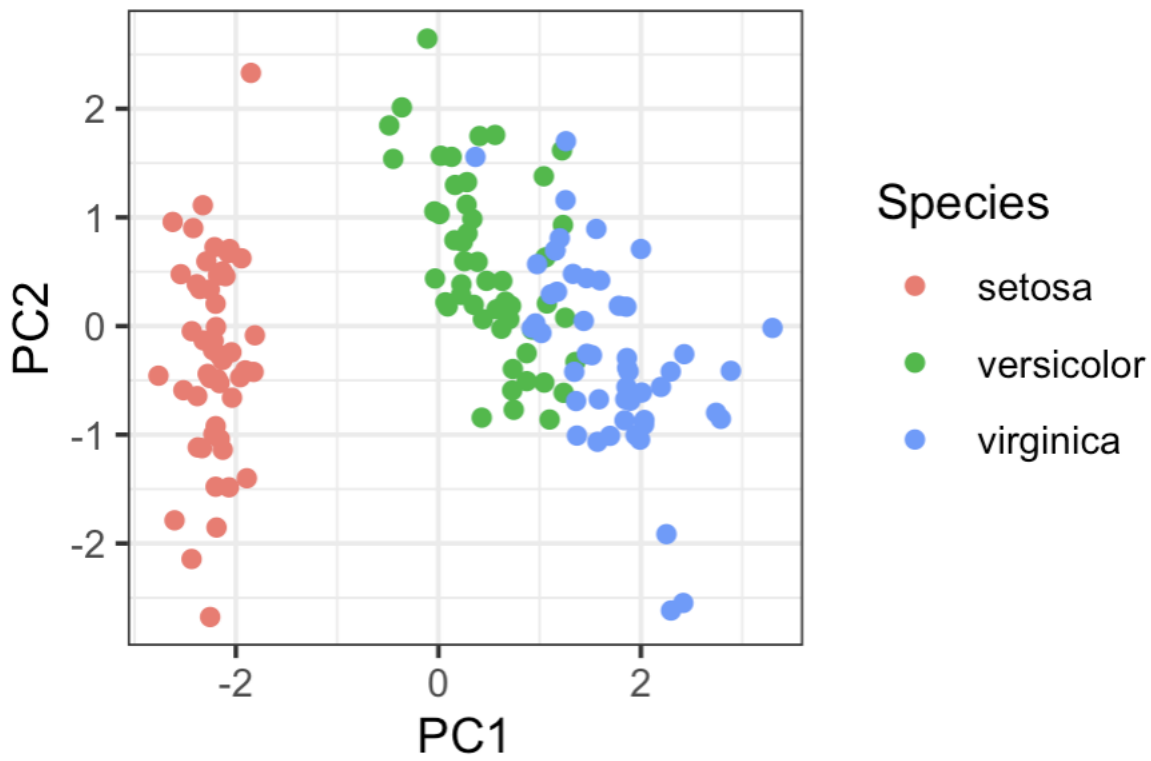
… +150 rows

```r
iris_pca<-prcomp(iris[,-5], scale=T)

biplot(iris_pca)
```




```r
pcs<-data.frame(iris_pca$x)
pcs$Species<-iris$Species
ggplot(pcs, aes(x=PC1, y=PC2, col=Species))+
  geom_point()+
  theme_bw()
```



```
iris_pca<-prcomp(iris[,-5], scale=T)
```

```
# loadings: relationship between variables and PCs
iris_pca$rotation
```

	PC1	PC2	PC3	PC4
Sepal.Length	0.5210659	-0.37741762	0.7195664	0.2612863
Sepal.Width	-0.2693474	-0.92329566	-0.2443818	-0.1235096
Petal.Length	0.5804131	-0.02449161	-0.1421264	-0.8014492
Petal.Width	0.5648565	-0.06694199	-0.6342727	0.5235971

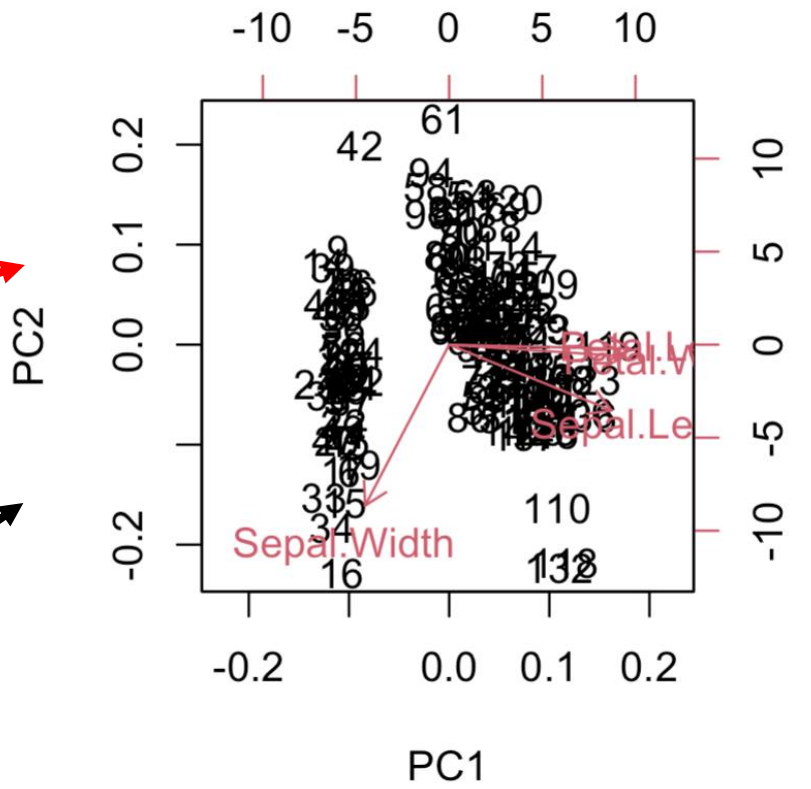
```
# scores: relationship between observations and PCs
iris_pca$x
```

	PC1	PC2	PC3	PC4
[1,]	-2.25714118	-0.478423832	0.127279624	0.024087508
[2,]	-2.07401302	0.671882687	0.233825517	0.102662845
[3,]	-2.35633511	0.340766425	-0.044053900	0.028282305
[4,]	-2.29170679	0.595399863	-0.090985297	-0.065735340
[5,]	-2.38186270	-0.644675659	-0.015685647	-0.035802870
[6,]	-2.06870061	-1.484205297	-0.026878250	0.006586116
[7,]	-2.43586845	-0.047485118	-0.334350297	-0.036652767
[8,]	-2.22539189	-0.222403002	0.088399352	-0.024529919
[9,]	-2.32684533	1.111603700	-0.144592465	-0.026769540
[10,]	-2.17703491	0.467447569	0.252918268	-0.039766068

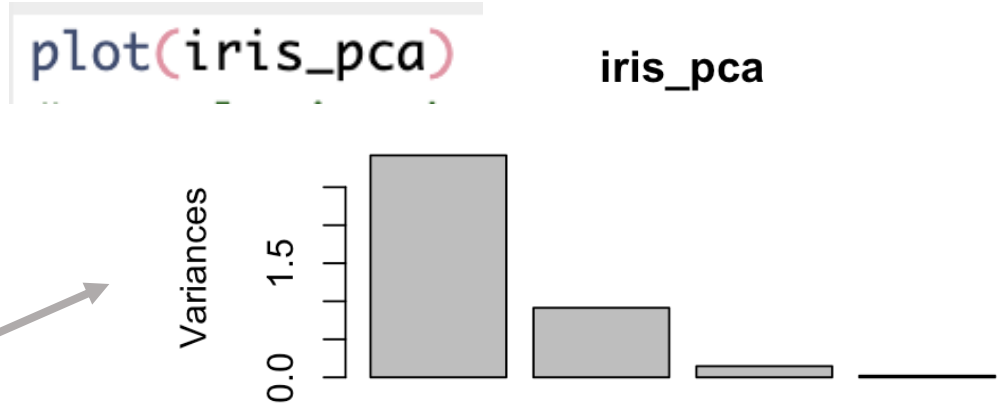
```
# eigenvalues: variance explained by PCs
iris_pca$sdev^2
```

```
[1] 2.91849782 0.91403047 0.14675688 0.02071484
```

```
biplot(iris_pca)
```




```
plot(iris_pca)
```



Things to Keep in Mind

Original data needs to contain correlated variables for a PCA to be useful

PCA is highly sensitive to differences in units or scale

- If variables have the same units and are on similar scales: okay to use variance-covariance matrix
- If variables have different units or are on different scales: use the correlation matrix, which contains centered and scaled data

Failure to standardize data leads the PCA to be dominated by the one or two variables that have the largest measurements

- Model becomes uninformative
- 1-2 variables will be the only variables that load on the PCs
- PCA only describes differences between samples using those 1-2 variables rather than using many or all of the variables in the data

Which Variables have the Largest Effect on the New Axes?

- Size and sign of loading reflects the relationship between the variables and the PCA axes
- To get Pearson correlations between variables and PCs:

  Correlation = loadings * sqrt (eigenvalue for the PC)

- Correlations range from -1 to 1 -> they are easy to interpret
- The correlations show the same patterns as the loadings

	PC1	PC2	PC3
Girth	0.6085705	-0.4099013	0.67942837
Height	0.4891267	0.8680065	0.08555556
Volume	0.6248176	-0.2802600	-0.72873681

Loadings

	PC1	PC2	PC3
Girth	0.9448143	-0.3071847	0.11385787
Height	0.7593761	0.6504939	0.01433731
Volume	0.9700381	-0.2100300	-0.12212093

Correlations

Case study: Cassava climate adaptation

BIO1 = Annual Mean Temperature

BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))

BIO3 = Isothermality (BIO2/BIO7) (×100)

BIO4 = Temperature Seasonality (standard deviation ×100)

BIO5 = Max Temperature of Warmest Month

BIO6 = Min Temperature of Coldest Month

BIO7 = Temperature Annual Range (BIO5-BIO6)

BIO8 = Mean Temperature of Wettest Quarter

BIO9 = Mean Temperature of Driest Quarter

BIO10 = Mean Temperature of Warmest Quarter

BIO11 = Mean Temperature of Coldest Quarter

BIO12 = Annual Precipitation

BIO13 = Precipitation of Wettest Month

BIO14 = Precipitation of Driest Month

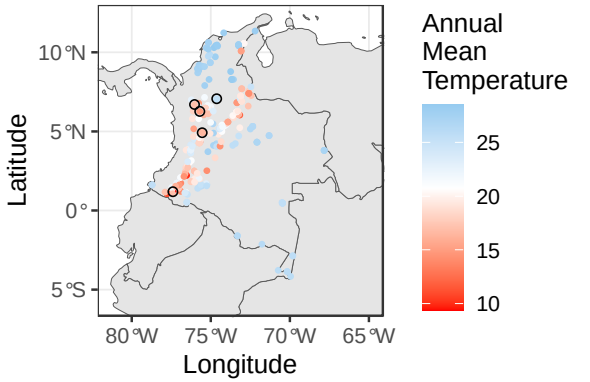
BIO15 = Precipitation Seasonality (Coefficient of Variation)

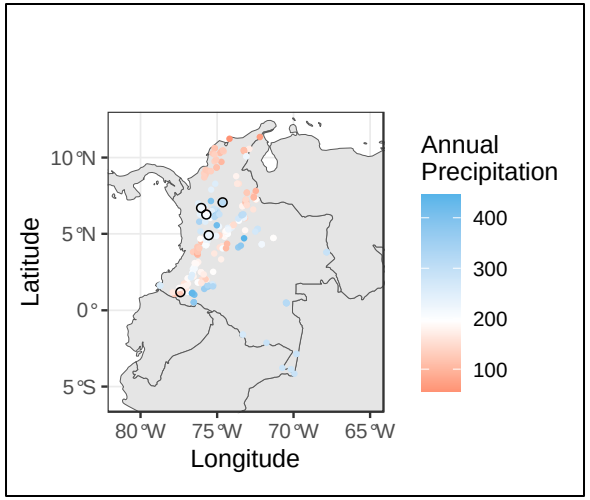
BIO16 = Precipitation of Wettest Quarter

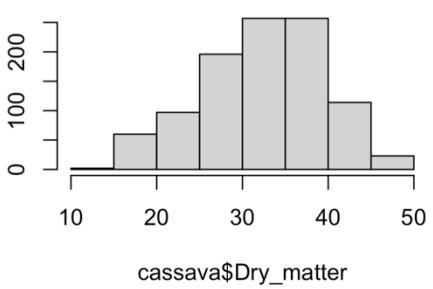
BIO17 = Precipitation of Driest Quarter

BIO18 = Precipitation of Warmest Quarter

BIO19 = Precipitation of Coldest Quarter







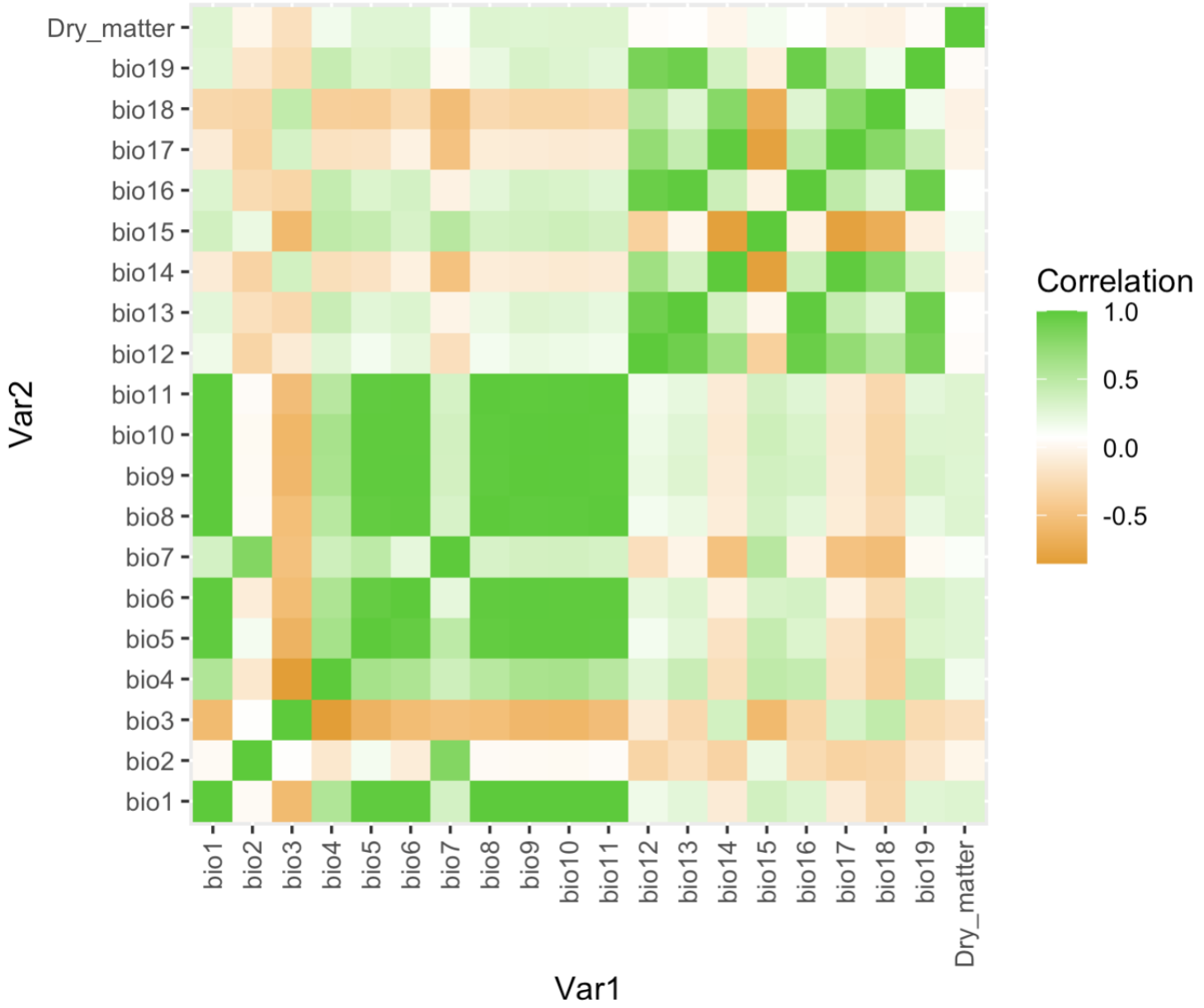
	bio1	bio2-bio18	bio19	Dry_matter
1	282	…	355	44.70
2	258		679	28.05
3	258		679	32.31
4	260		640	41.98
5	259		582	34.73
6	265		1143	28.18

+1000 more rows

Lots of Collinearity in the data

	bio1	bio2-bio18	bio19	Dry_matter
1	282	…	355	44.70
2	258		679	28.05
3	258		679	32.31
4	260		640	41.98
5	259		582	34.73
6	265		1143	28.18

+1000 more rows



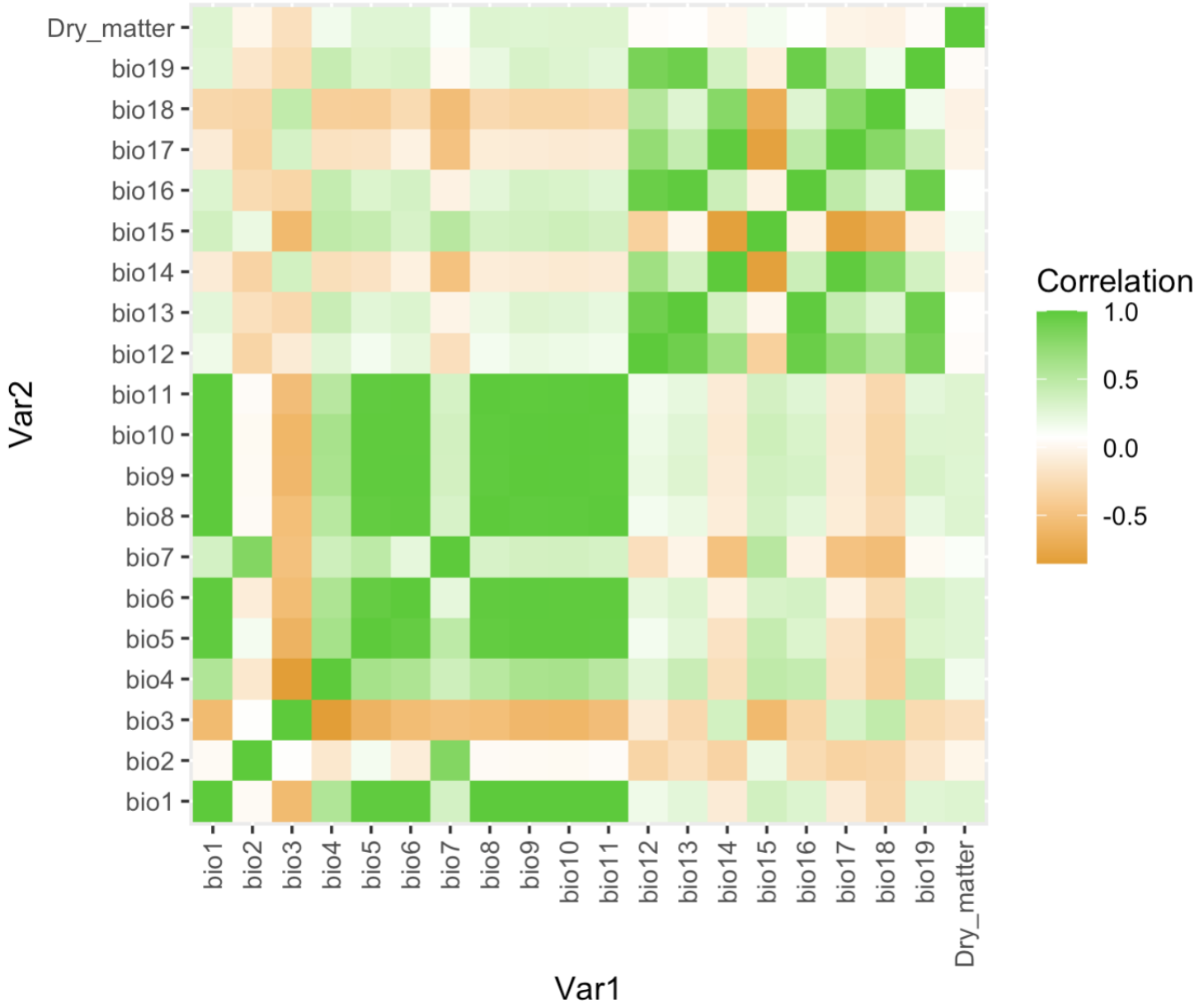
Lots of Collinearity in the data

```
Call:
lm(formula = Dry_matter ~ ., data = cassava)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1777  -4.1371   0.6319   4.8324  15.3503

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.614e+01  3.858e+01  -0.418  0.67575
bio1         1.101e+00  3.609e-01   3.051  0.00234 **
bio2        -4.464e-01  3.515e-01  -1.270  0.20439
bio3         3.705e-01  4.407e-01   0.841  0.40070
bio4        -1.175e-02  1.054e-02  -1.114  0.26535
bio5         2.322e-01  3.171e-01   0.732  0.46430
bio6        -5.666e-01  3.033e-01  -1.868  0.06206 .
bio7                NA         NA      NA       NA
bio8        -9.527e-02  1.179e-01  -0.808  0.41935
bio9        -9.476e-02  1.260e-01  -0.752  0.45220
bio10        6.496e-02  3.694e-01   0.176  0.86044
bio11       -6.053e-01  4.690e-01  -1.291  0.19714
bio12        2.224e-03  2.622e-03   0.848  0.39652
bio13       -2.177e-02  1.481e-02  -1.470  0.14184
bio14        9.115e-02  3.263e-02   2.793  0.00532 **
bio15        1.401e-01  4.674e-02   2.999  0.00278 **
bio16        4.167e-03  8.318e-03   0.501  0.61650
bio17       -2.451e-02  1.264e-02  -1.940  0.05271 .
bio18       -7.802e-04  2.357e-03  -0.331  0.74076
bio19        2.842e-04  2.995e-03   0.095  0.92442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.586 on 987 degrees of freedom
Multiple R-squared:  0.1391,    Adjusted R-squared:  0.1234
F-statistic: 8.859 on 18 and 987 DF,  p-value: < 2.2e-16
```

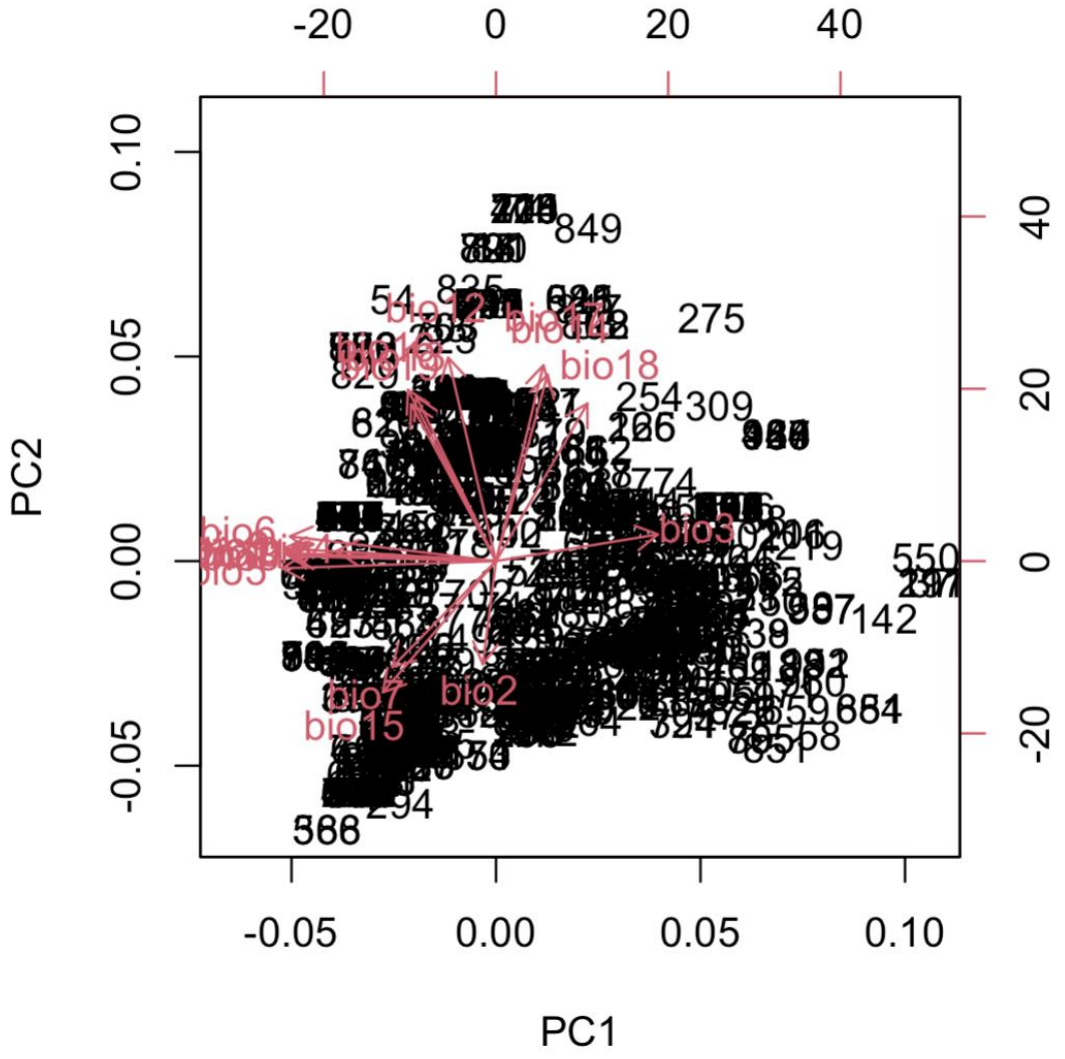


Lots of Collinearity in the data

	bio1	bio2-bio18	bio19	Dry_matter
1	282	…	355	44.70
2	258		679	28.05
3	258		679	32.31
4	260		640	41.98
5	259		582	34.73
6	265		1143	28.18

+1000 more rows

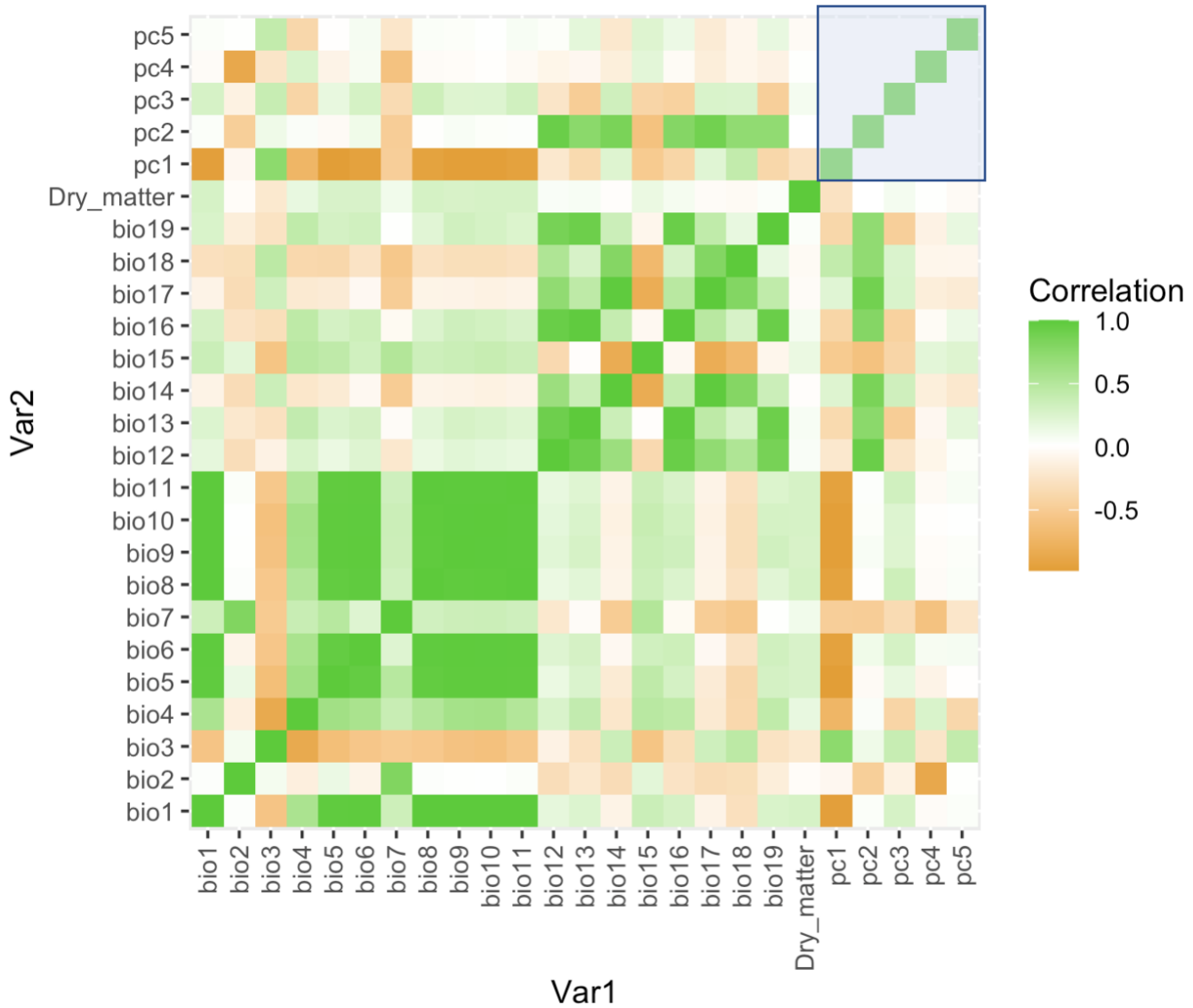
PCA ---------->>>>



We can add PC coordinates back to the observation data

```
'data.frame':   1006 obs. of  25 variables:
 $ bio1      : int  282 258 258 260 259 265 280 257 257 260 ...
 $ bio2      : int  118 91 91 94 93 96 115 94 93 96 ...
 $ bio3      : int  81 87 87 86 85 75 78 81 80 82 ...
 $ bio4      : int  588 376 376 401 399 935 518 562 594 512 ...
 $ bio5      : int  358 309 309 312 311 336 355 320 320 324 ...
 $ bio6      : int  214 205 205 203 202 208 208 205 205 208 ...
 $ bio7      : int  144 104 104 109 109 128 147 115 115 116 ...
 $ bio8      : int  278 259 259 261 260 257 277 253 253 253 ...
 $ bio9      : int  280 255 255 256 253 273 277 263 263 265 ...
 $ bio10     : int  288 262 262 264 262 278 287 263 263 265 ...
 $ bio11     : int  273 252 252 254 253 253 273 248 247 252 ...
 $ bio12     : int  1296 2885 2885 2836 2779 2833 1020 3236 3260 3670 ...
 $ bio13     : int  200 294 294 332 314 453 140 407 403 467 ...
 $ bio14     : int  16 165 165 158 145 14 12 111 105 110 ...
 $ bio15     : int  59 19 19 22 22 61 53 37 37 38 ...
 $ bio16     : int  511 855 855 889 880 1209 398 1178 1187 1380 ...
 $ bio17     : int  65 509 509 509 505 105 66 395 391 442 ...
 $ bio18     : int  181 751 751 672 681 164 258 395 391 442 ...
 $ bio19     : int  355 679 679 640 582 1143 245 1030 1055 1231 ...
 $ Dry_matter: num  44.7 28.1 32.3 42 34.7 ...
 $ pc1       : num  -3.102 0.151 0.151 -0.22 -0.148 ...
 $ pc2       : num  -3.87 2.48 2.48 2.21 1.99 ...
 $ pc3       : num  0.517 2.445 2.445 2.018 2.013 ...
 $ pc4       : num  -1.8223 0.131 0.131 -0.2021 -0.0148 ...
 $ pc5       : num  -0.578 -0.338 -0.338 -0.408 -0.541 ...
```

Resulting PCs are uncorrelated with each other… but correlated with the original predictors and with Dry_matter



Using PCs as variables in linear models

```
Call:
lm(formula = Dry_matter ~ ., data = cassava)

Residuals:
     Min       1Q   Median       3Q      Max
-20.1777  -4.1371   0.6319   4.8324  15.3503

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.614e+01  3.858e+01  -0.418  0.67575
bio1         1.101e+00  3.609e-01   3.051  0.00234 **
bio2        -4.464e-01  3.515e-01  -1.270  0.20439
bio3         3.705e-01  4.407e-01   0.841  0.40070
bio4        -1.175e-02  1.054e-02  -1.114  0.26535
bio5         2.322e-01  3.171e-01   0.732  0.46430
bio6        -5.666e-01  3.033e-01  -1.868  0.06206 .
bio7                NA         NA      NA       NA
bio8        -9.527e-02  1.179e-01  -0.808  0.41935
bio9        -9.476e-02  1.260e-01  -0.752  0.45220
bio10        6.496e-02  3.694e-01   0.176  0.86044
bio11       -6.053e-01  4.690e-01  -1.291  0.19714
bio12        2.224e-03  2.622e-03   0.848  0.39652
bio13       -2.177e-02  1.481e-02  -1.470  0.14184
bio14        9.115e-02  3.263e-02   2.793  0.00532 **
bio15        1.401e-01  4.674e-02   2.999  0.00278 **
bio16        4.167e-03  8.318e-03   0.501  0.61650
bio17       -2.451e-02  1.264e-02  -1.940  0.05271 .
bio18       -7.802e-04  2.357e-03  -0.331  0.74076
bio19        2.842e-04  2.995e-03   0.095  0.92442
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.586 on 987 degrees of freedom
Multiple R-squared:  0.1391,    Adjusted R-squared:  0.1234
F-statistic: 8.859 on 18 and 987 DF,  p-value: < 2.2e-16
```

PCA of climate variables

Use PC1 as predictor

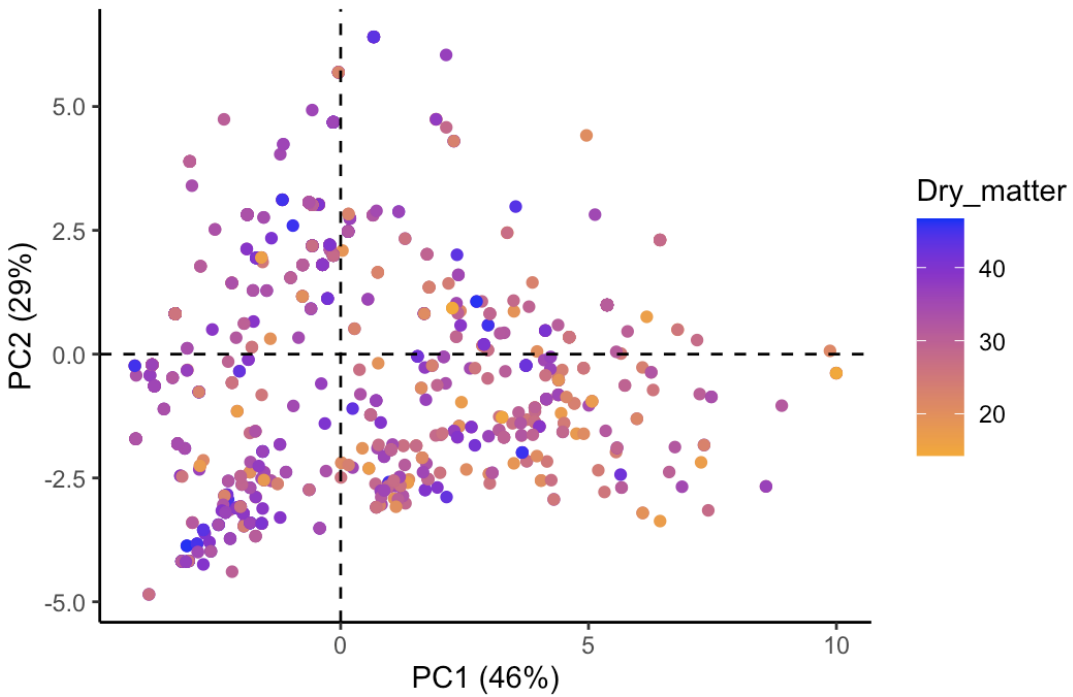
```
Call:
lm(formula = Dry_matter ~ pc1, data = cassava)

Residuals:
     Min       1Q   Median       3Q      Max
-17.6697  -4.2522   0.3768   4.7670  15.6517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.3587     0.2134 151.603   <2e-16 ***
pc1          -0.6496     0.0722  -8.997   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.77 on 1004 degrees of freedom
Multiple R-squared:  0.07461,   Adjusted R-squared:  0.07369
F-statistic: 80.95 on 1 and 1004 DF,  p-value: < 2.2e-16
```

```r
#Making PCA plots with ggplot2
ggplot(cassava, aes(x=pc1, y=pc2, col=Dry_matter))+
  geom_point()+
  theme_classic()+
  coord_fixed(ratio = 1) +
  geom_hline(yintercept = 0, linetype="dashed")+
  geom_vline(xintercept = 0, linetype="dashed")+
  scale_color_gradient(low="orange",high="blue")+
  xlab(paste0("PC1 (", round(pca_climate$sdev[1]^2 / sum(pca_climate$sdev^2) * 100), "%)")) +
  ylab(paste0("PC2 (", round(pca_climate$sdev[2]^2 / sum(pca_climate$sdev^2) * 100), "%)"))
```




```r
# plot loadings for each predictor
loadings <- as.data.frame(pca_climate$rotation)

library(ggrepel)
ggplot(data = loadings, aes(x = PC1, y = PC2, label = rownames(loadings))) +theme_classic()+
  geom_segment(aes(x=0, y=0, xend=PC1, yend=PC2), col="gray") +
  geom_text_repel(size = 4, col="blue") +
  coord_fixed(ratio = 1) +
  xlab(paste0("PC1 (", round(pca_climate$sdev[1]^2 / sum(pca_climate$sdev^2) * 100), "%)")) +
  ylab(paste0("PC2 (", round(pca_climate$sdev[2]^2 / sum(pca_climate$sdev^2) * 100), "%)"))
```