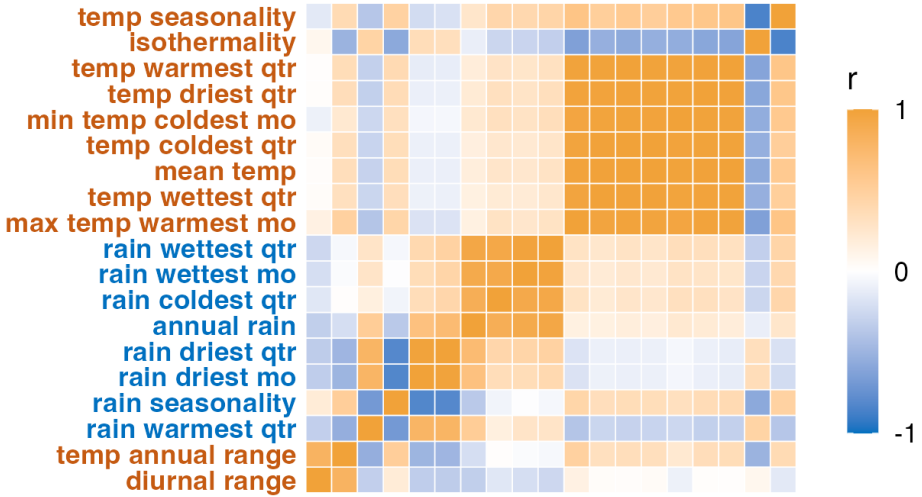


Dimension reduction

Principal component analysis, and its cousins

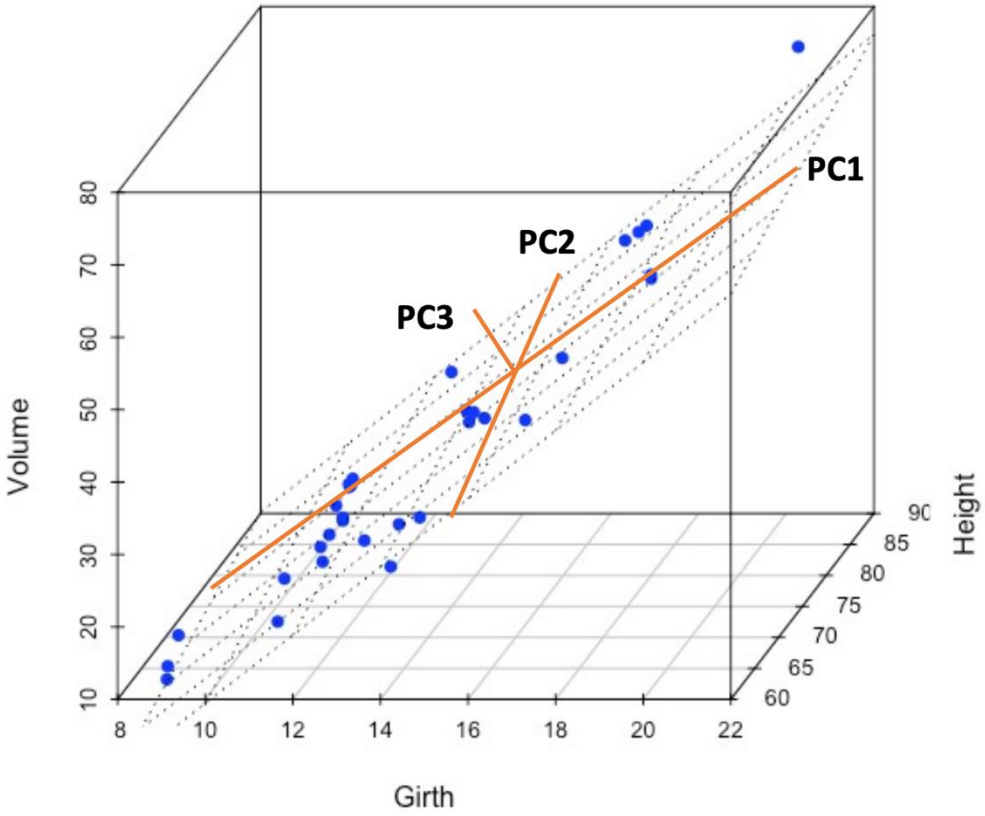
19 columns, only a few signals

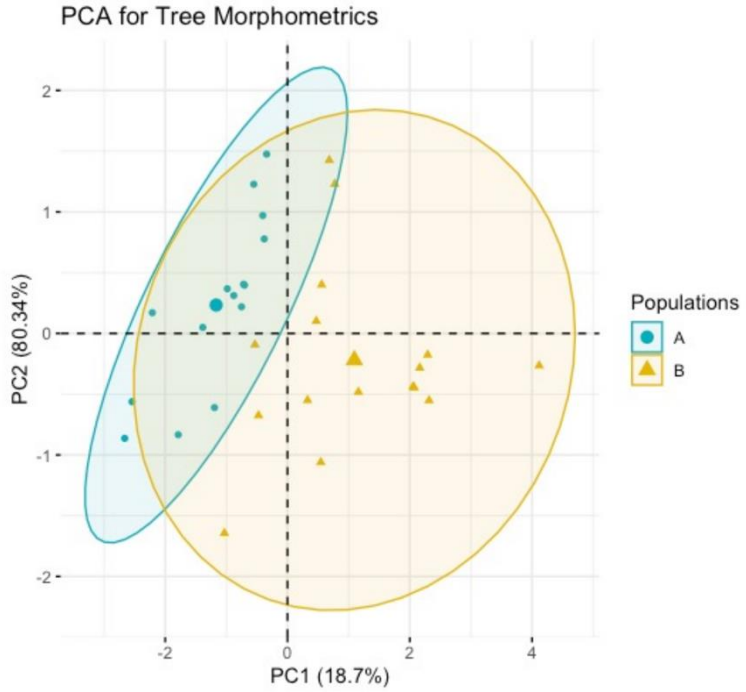



```
> m <- lm(Dry_matter ~ . - bio7, data = cassava)
> round(sort(car::vif(m), decreasing = TRUE)[1:3])
bio11 bio10  bio1
 7343  5276  4662
```

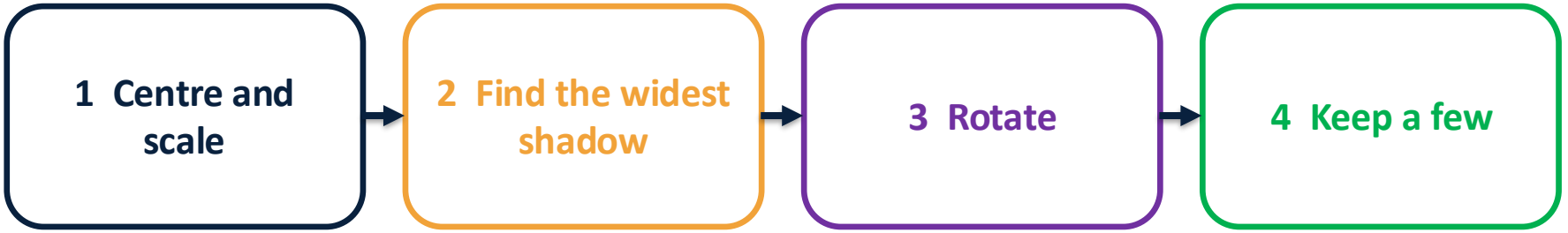
Blocks of near-copies: **temperature** and **rainfall**. **Can a few new columns keep the information?**

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

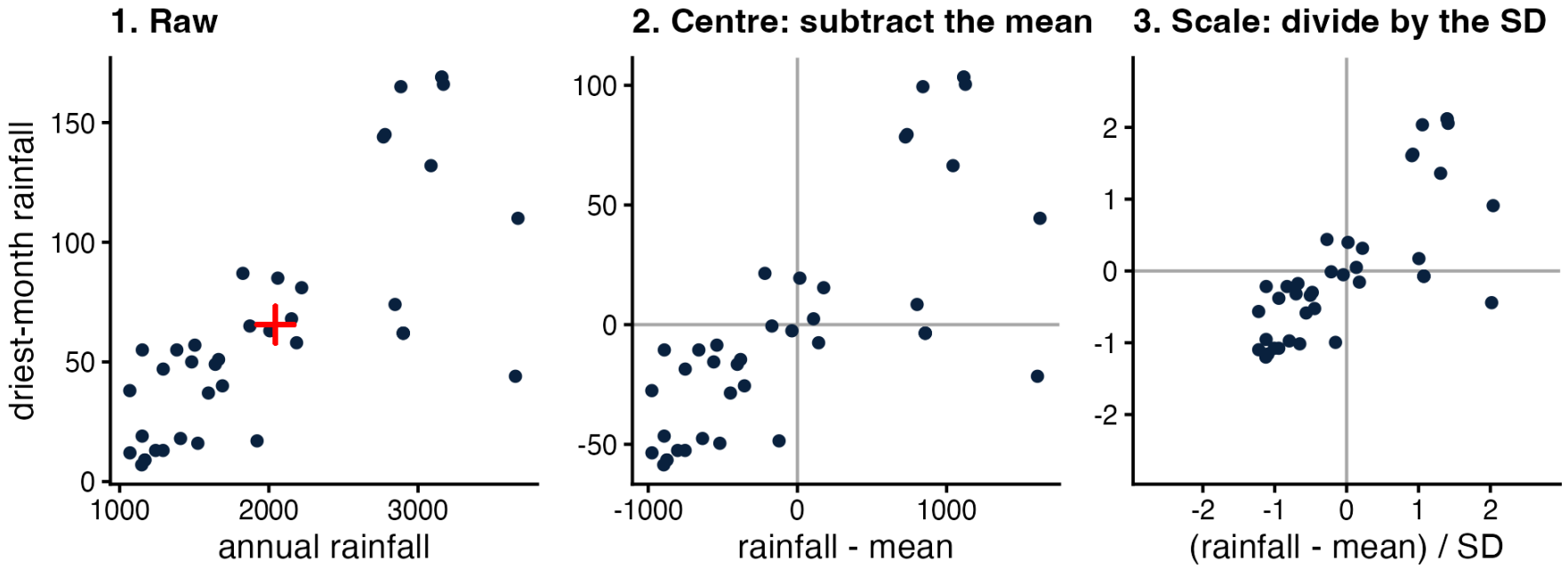




The recipe


1 Centre and scale
2 Find the widest shadow
3 Rotate
4 Keep a few

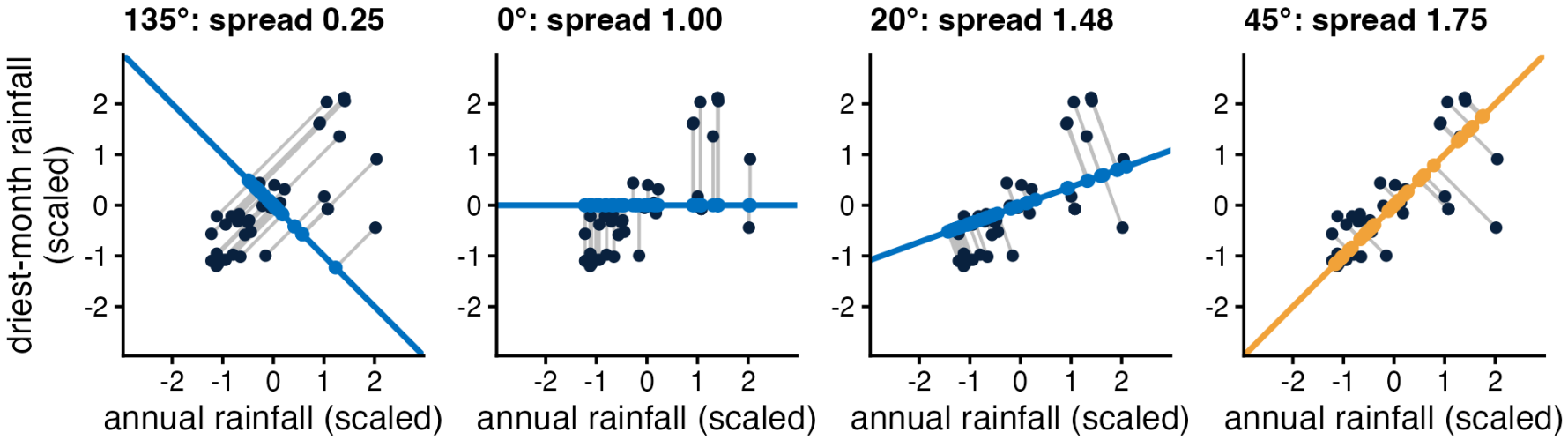
Step 1: centre and scale



Every variable ends with **mean 0, SD 1**: no unit can shout louder

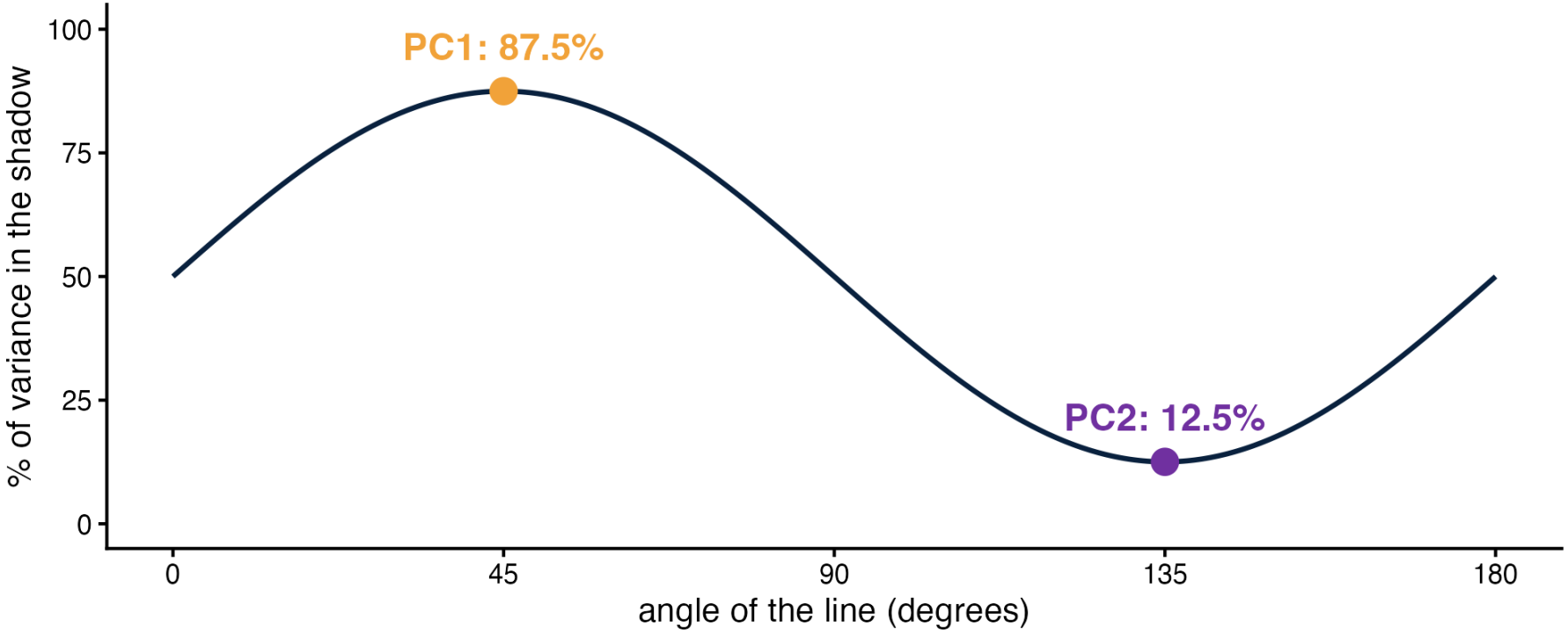
Step 2: spin a line, measure the shadow

Try it: PCA explorer



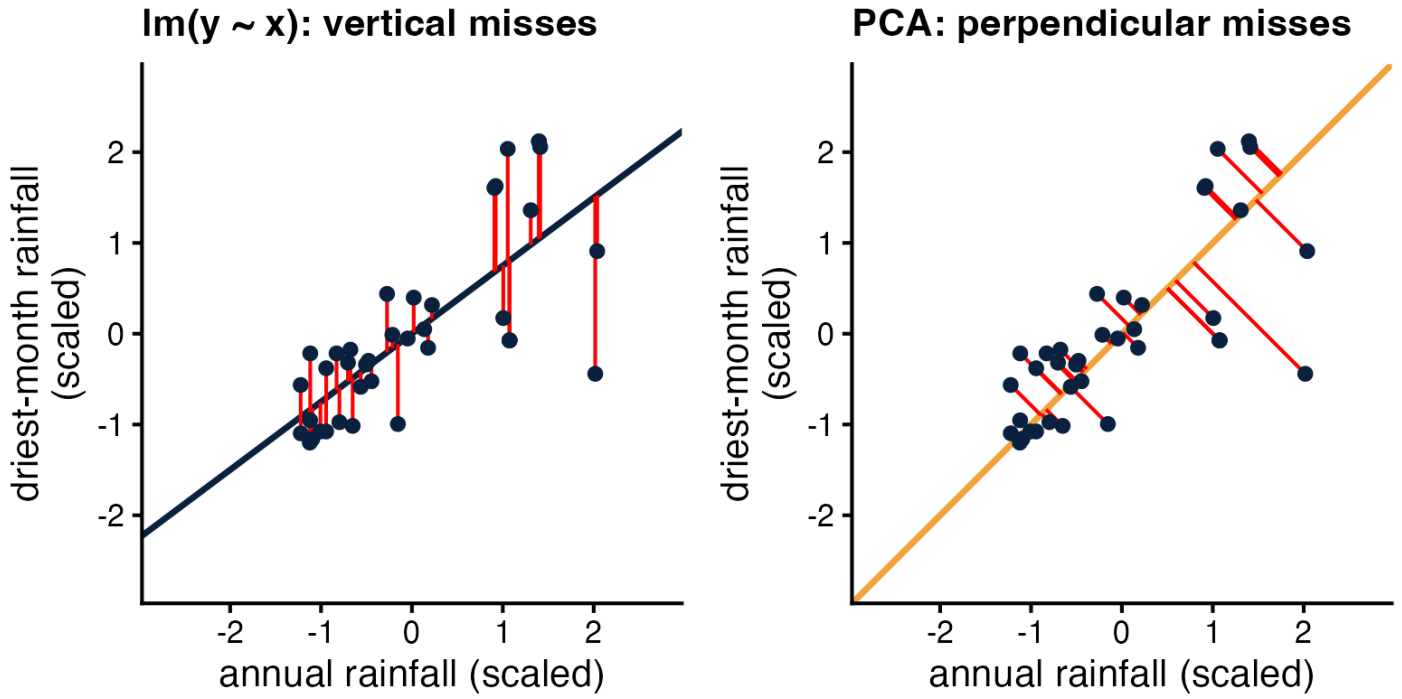
Drop every point onto the line. **Wider shadow = more of the variation kept.**

The widest shadow is PC1



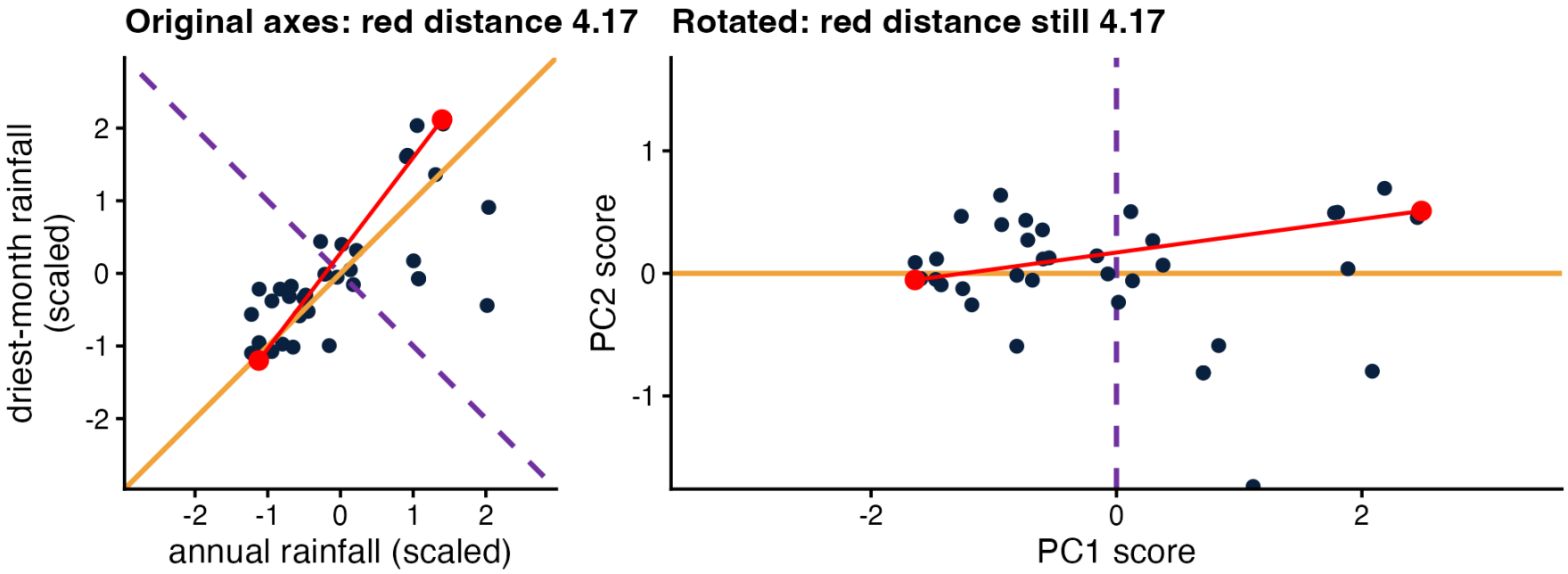
PC1: the angle with the most variance. **PC2**: at right angles, the rest.

Not a regression line



lm() predicts y. **PCA has no y**: both variables are treated the same.

Step 3: rotate, nothing is lost



Same points, new axes. **Scores** = each sample's position on the new axes.

Live poll

Multiple choice

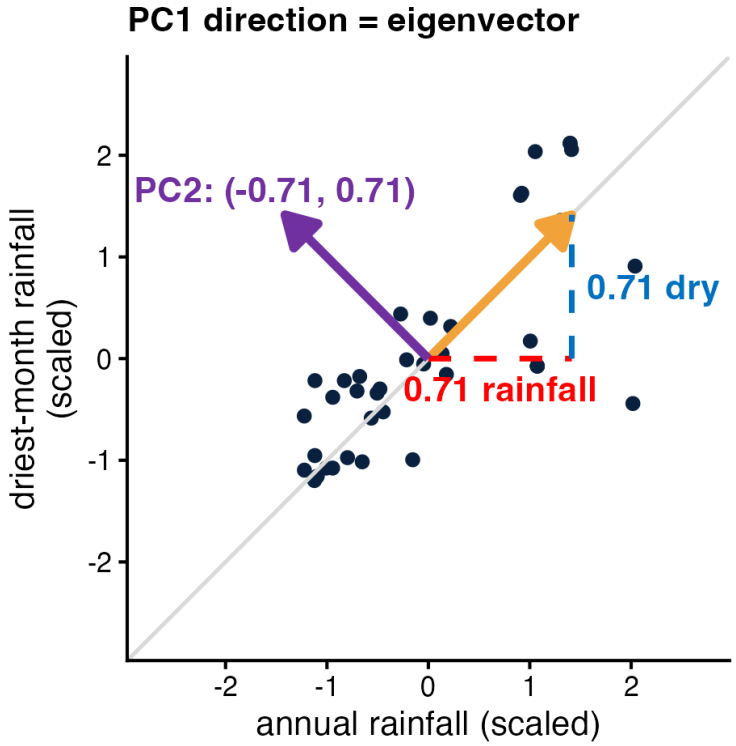
Iris has 4 variables, so 4 PCs. Keep all 4: what share of the variance do they explain together?

Scan with your phone

Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w7-all-pcs

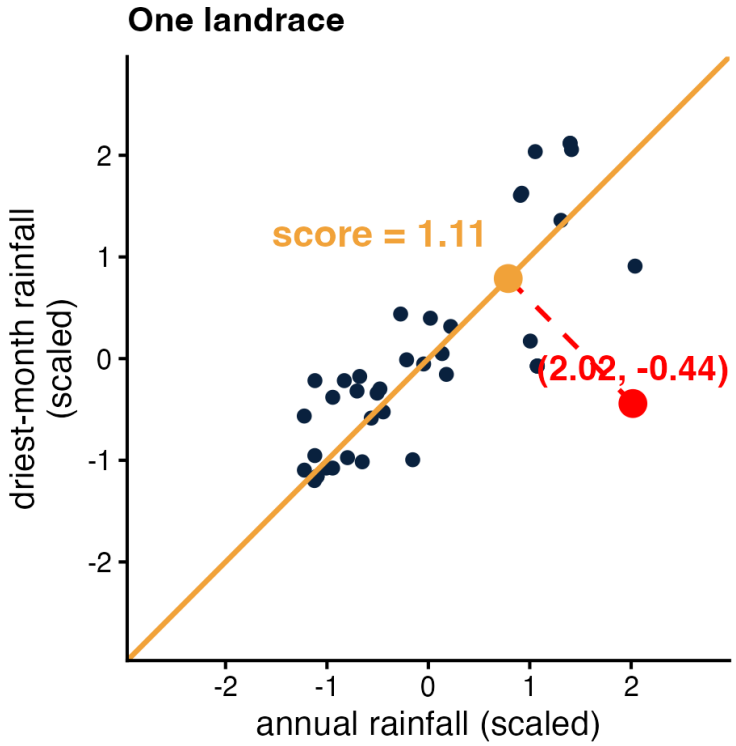
host screen

Eigenvector = the direction of PC1



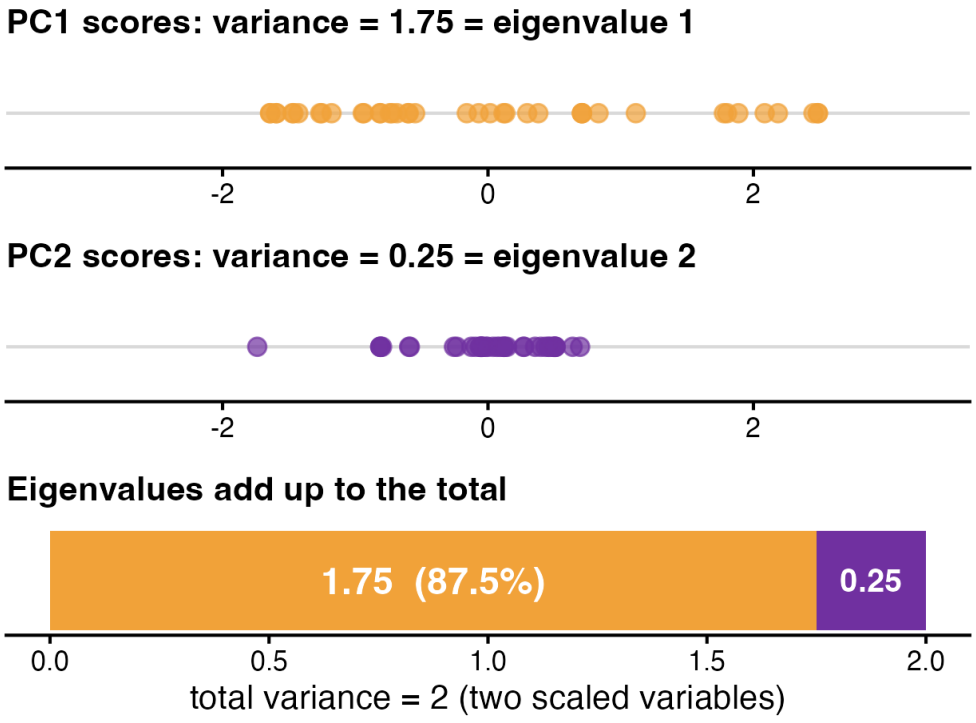
$$\text{PC1} = 0.71 \times \text{rain} + 0.71 \times \text{dry}$$

Score = one sample's position on PC1



$$\text{score} = 0.71 \times 2.02 + 0.71 \times (-0.44) = 1.11$$

Eigenvalue = the spread of the scores



$$\% \text{ PC1} = \frac{1.75}{1.75 + 0.25} = 87.5\%$$

In R: three pieces

```
> pca <- prcomp(rain2, scale. = TRUE)
> pca$sdev^2
[1] 1.7494834 0.2505166  ← eigenvalues
> pca$rotation
            PC1        PC2
rain  0.7071068 -0.7071068  ← eigenvectors = loadings
dry   0.7071068  0.7071068
> head(pca$x, 3)
              PC1         PC2
[1,]  0.01611942 -0.23548121  ← scores
[2,] -1.47323247 -0.04926863
[3,] -0.54889735  0.12682500
```

sdev^2 eigenvalues **rotation** loadings **x** scores

Which is which?

eigenvector / loadings	a direction: the recipe	`pca$rotation`
score	a sample's place on the new axis	`pca$x`
eigenvalue	spread of the scores	`pca$sdev^2`
% variance	eigenvalue / total	`summary(pca)`

Loadings as correlations

$$\text{cor}(\text{rain}, \text{PC1}) = \text{loading} \times \sqrt{\text{eigenvalue}}$$

$$= 0.71 \times \sqrt{1.75} = 0.94$$

scaled PCA: a loading times the PC's SD is a correlation

Scaled PCA: loading × SD of the PC = **correlation of the variable with the PC**

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

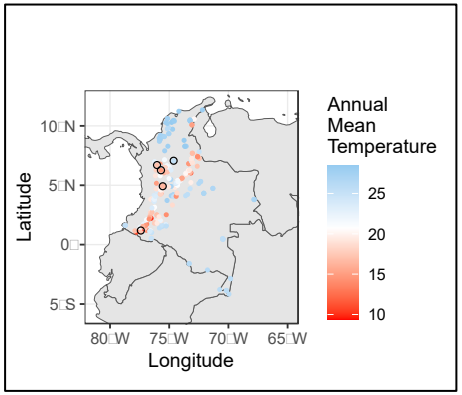
BIO15 = Precipitation Seasonality (Coefficient of Variation)

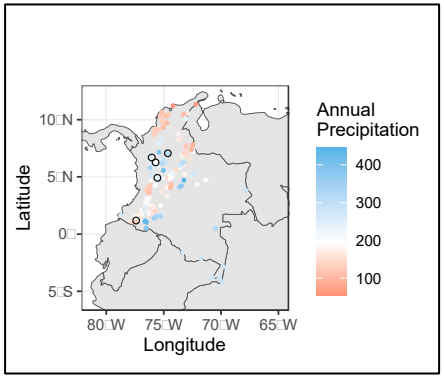
BIO16 = Precipitation of Wettest Quarter

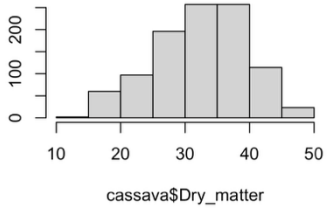
BIO17 = Precipitation of Driest Quarter

BIO18 = Precipitation of Warmest Quarter

BIO19 = Precipitation of Coldest Quarter





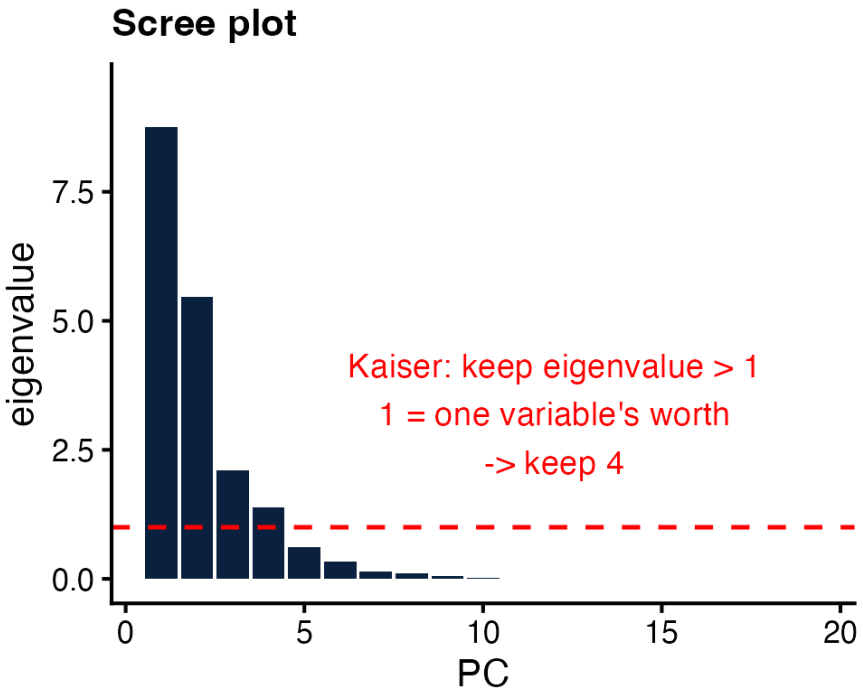


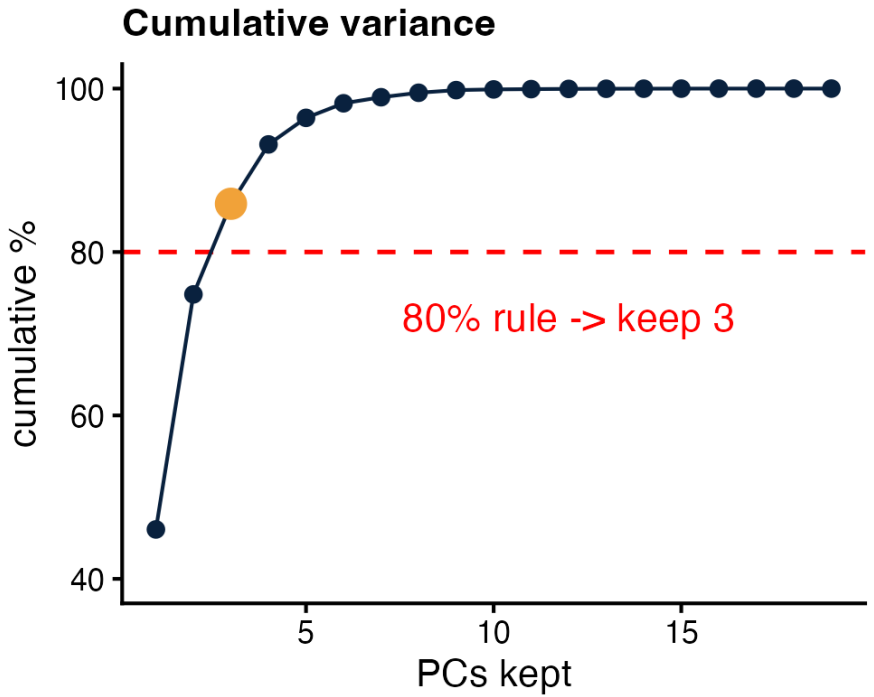


	bio1	bio2-bio18	bio19	Dry_matter
1	282	…	355	44.70
2	258		679	28.05
3	258		679	32.31
4	260		640	41.98
5	259		582	34.73
6	265		1143	28.18

+1000 more rows

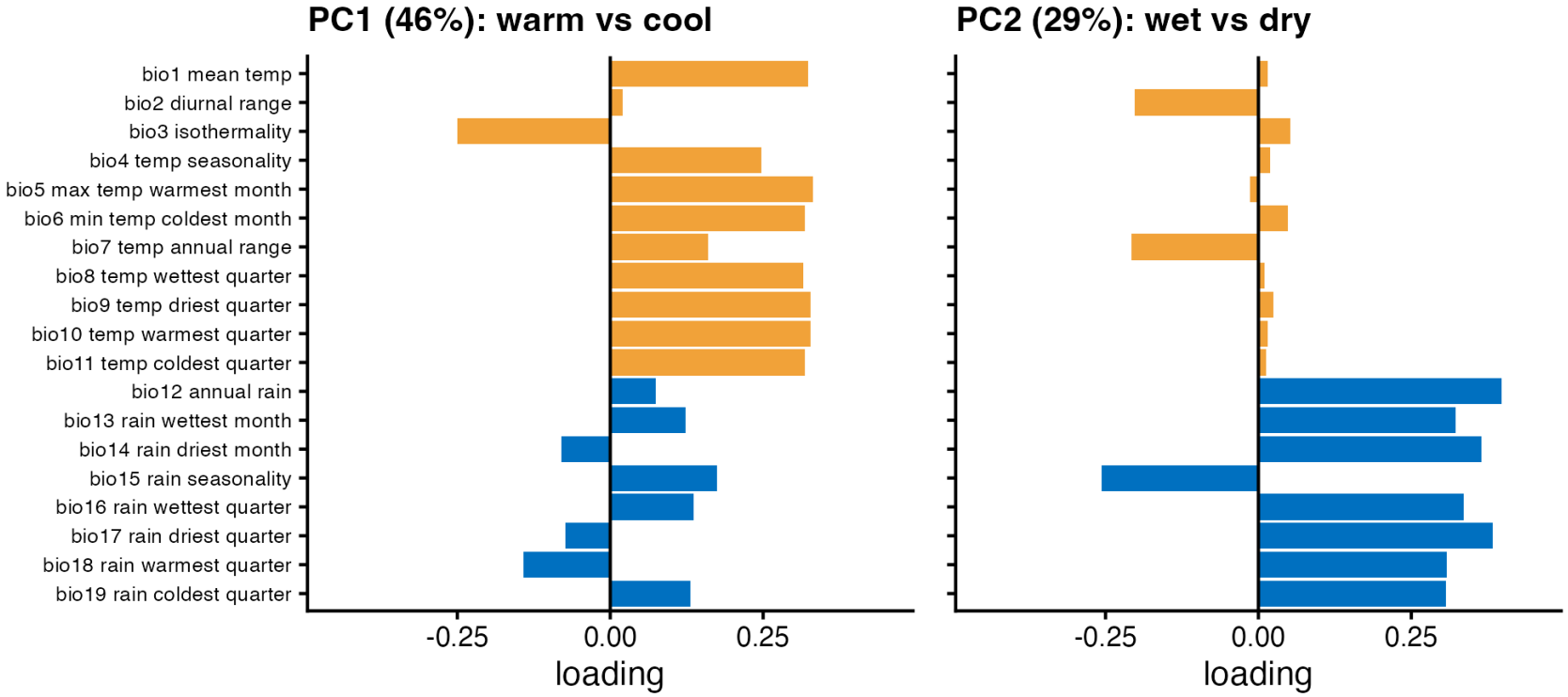
How many PCs to keep?





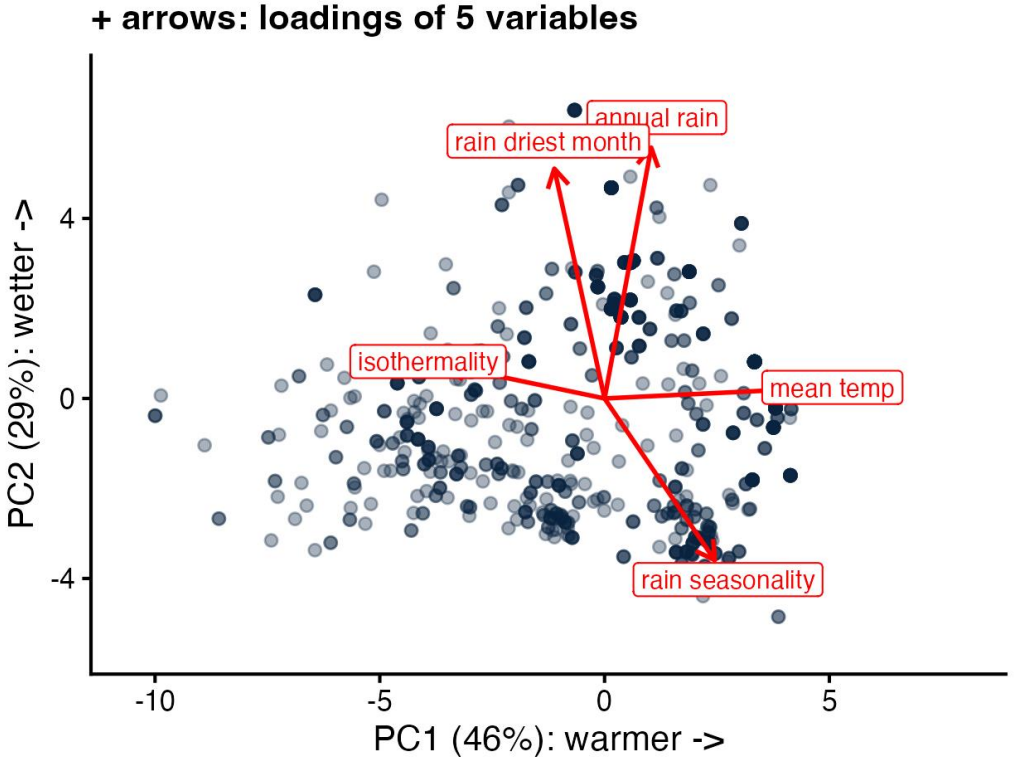
Elbow ≈ 3, eigenvalue > 1 keeps 4, 80% keeps 3. **No rule is the truth: check a few.**

Read the loadings to name each PC



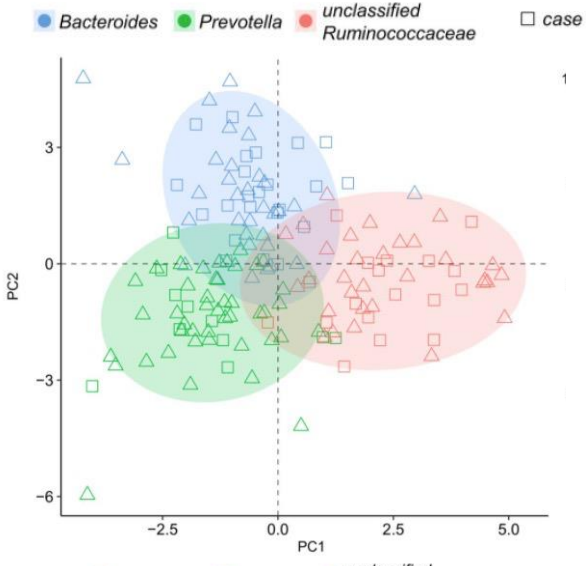
PC1 ≈ temperature. **PC2** ≈ rainfall. Name a PC by its big loadings.

Biplot: scores and loadings together



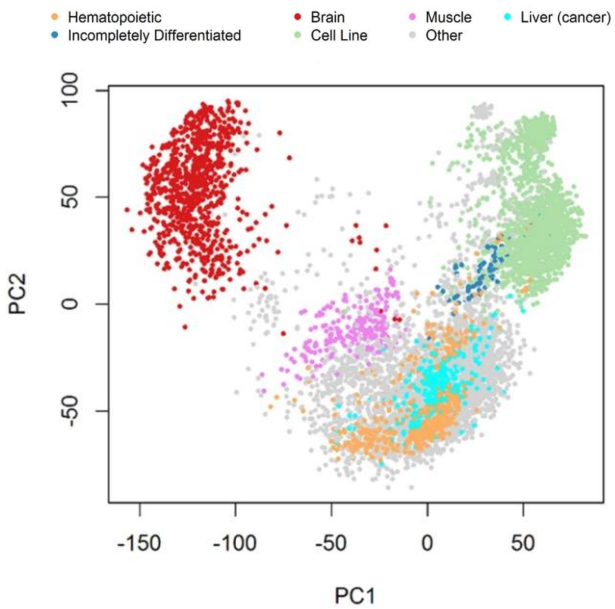
Dots: landraces (scores). **Arrows: variables (loadings).**

Example uses of PCA from Publications



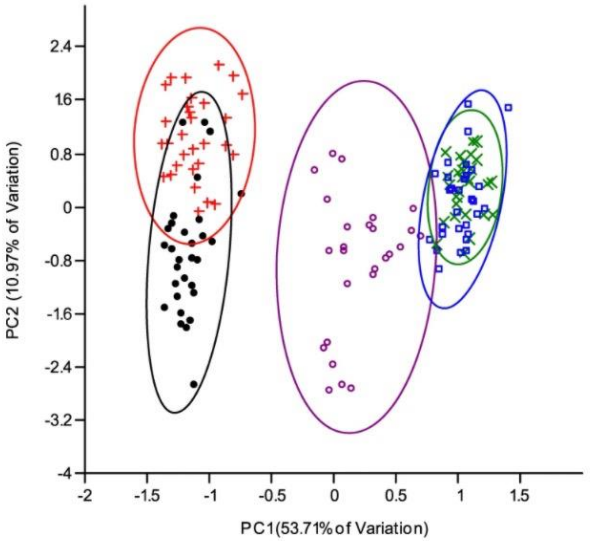
Hadizadeh et al. 2018. Gut. 67:778-779

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

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

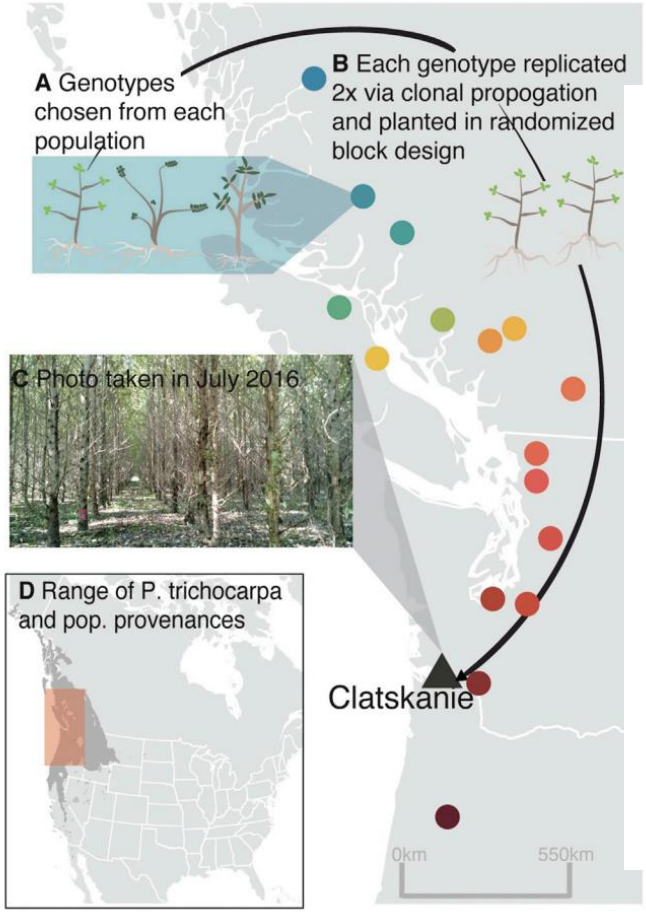
Meghan Blumstein, Andrew Richardson, David Weston, Jin Zhang, Wellington Muchero, Robin Hopkins

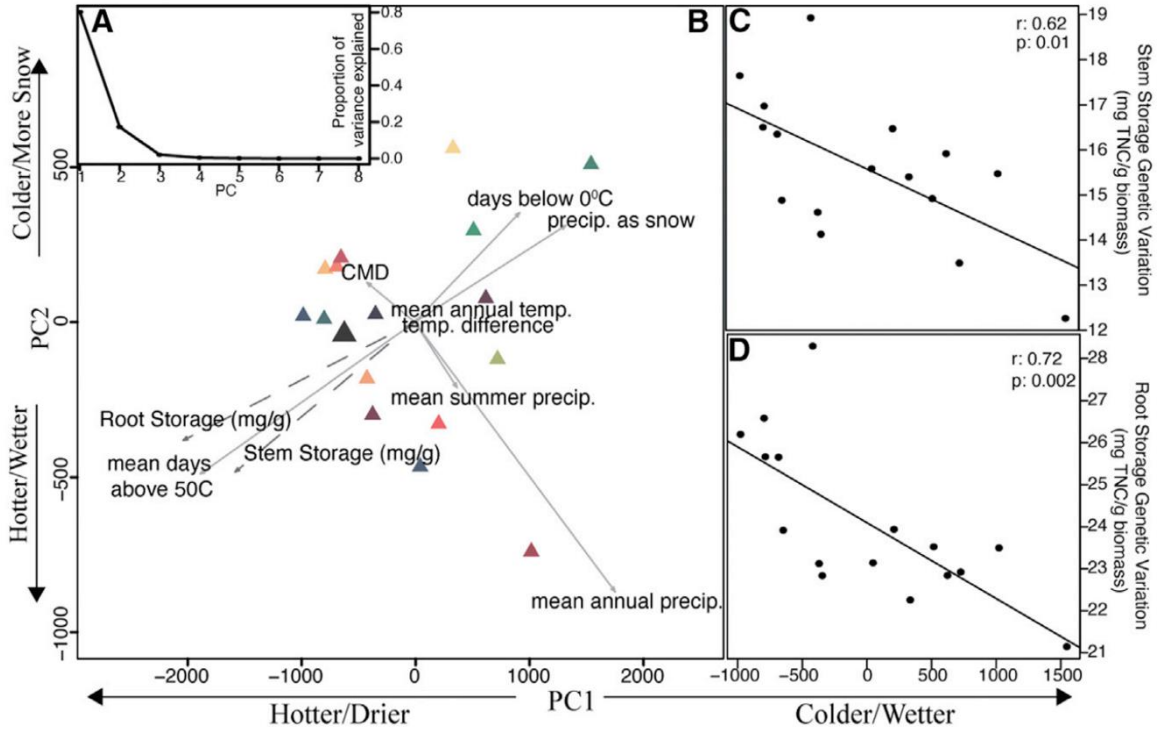
Correspondence

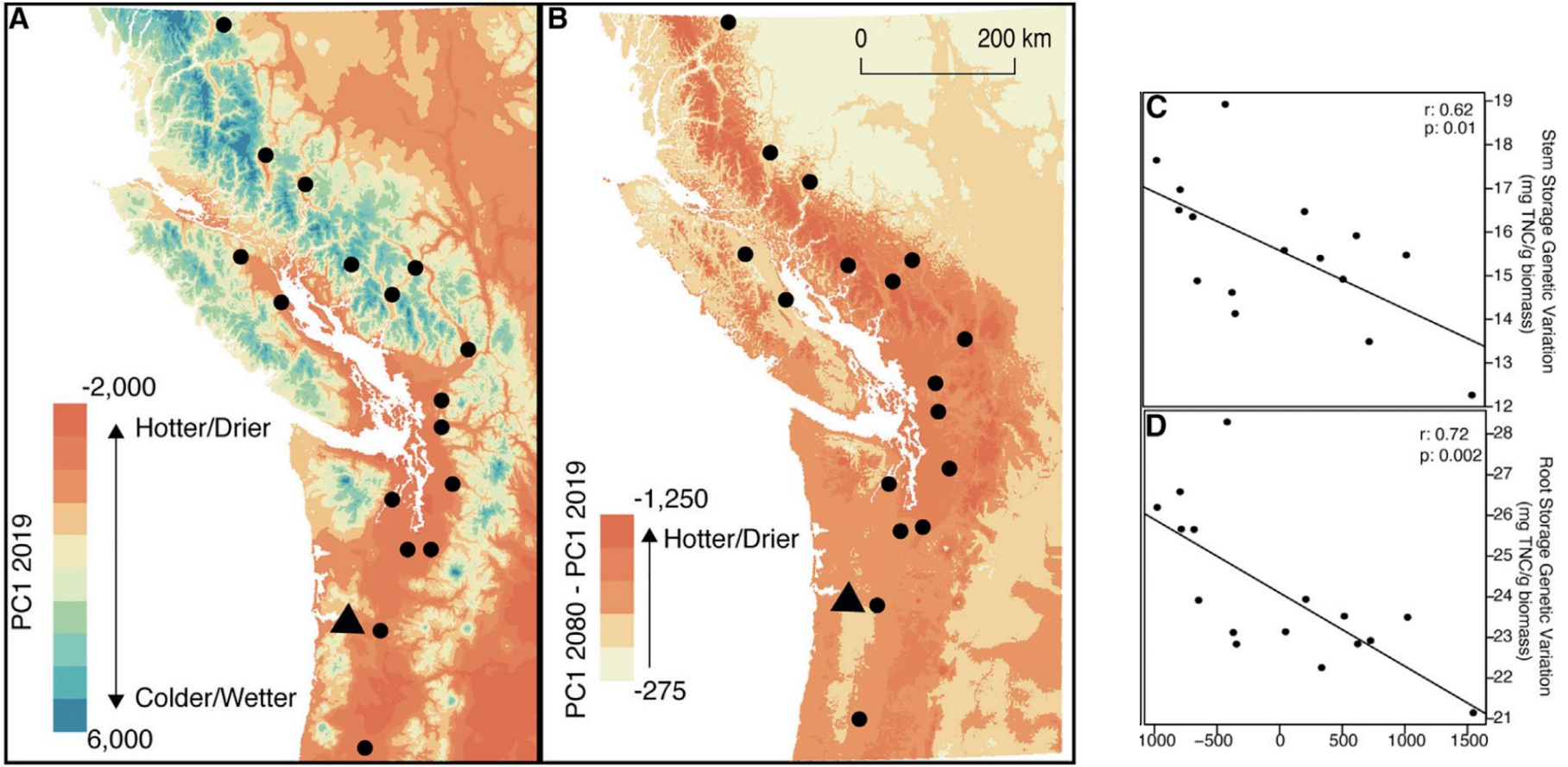
blumstein@fas.harvard.edu

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
PC1
Hotter/Drier
Colder/Wetter
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.
r: 0.62
p: 0.01
r: 0.72
p: 0.002
Stem Storage Genetic Variation (mg TNC/g biomass)
Root Storage Genetic Variation (mg TNC/g biomass)


A
-2,000
Hotter/Drier
PC1 2019
Colder/Wetter
6,000
B
0
200 km
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*

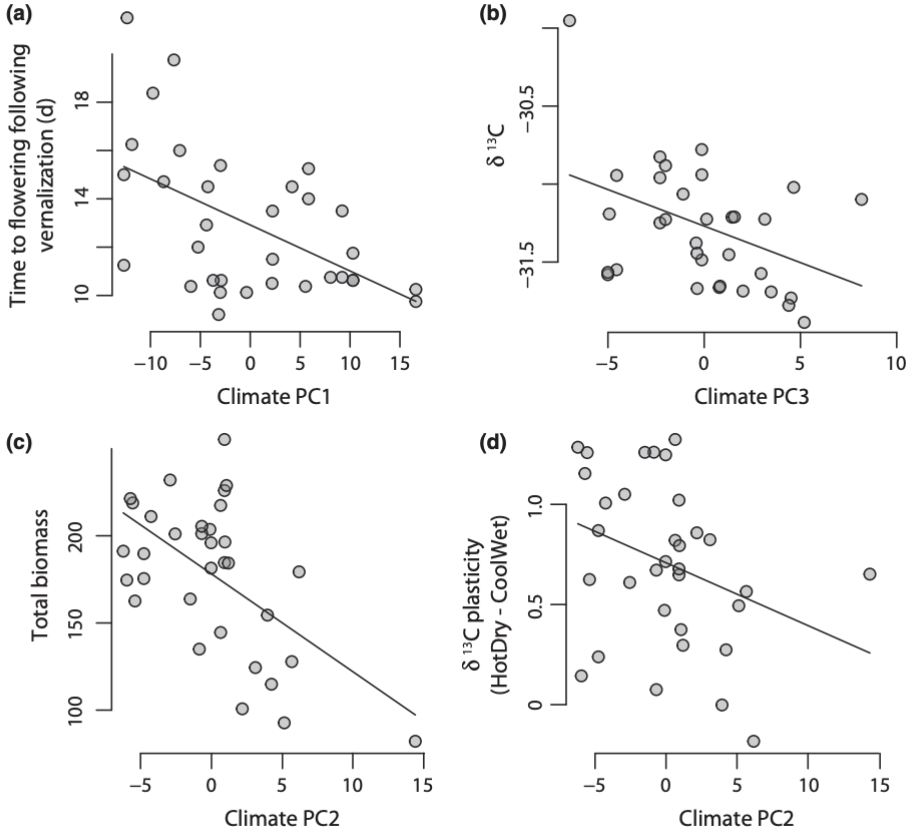
David L. Des Marais[1,4], Jesse R. Lasky[2], Paul E. Verslues[3], Trent Z. Chang[3] and Thomas E. Juenger[1]

[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

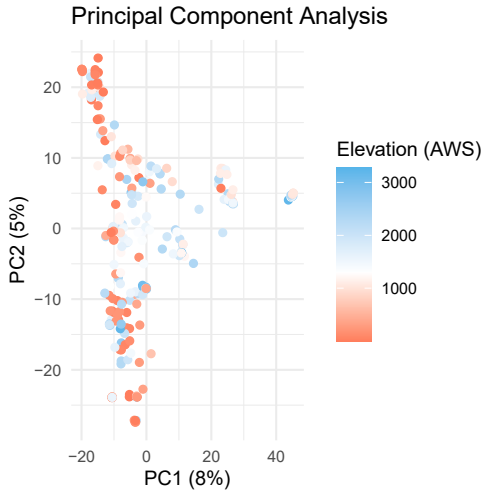
Table 1 Collection information for the natural accessions of *Brachypodium distachyon* used in the experiments

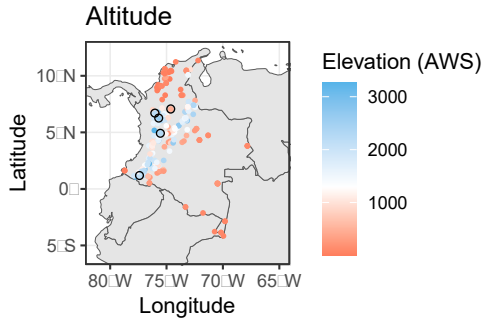
Line	Locale	Country	Latitude	Longitude
ABR2	Octon, Herault	France	43°39′18″N	3°18′12″E
ABR3	Huesca, Aisa	Spain	42.679°N	0.621°W
ABR4	Huesca, Aren	Spain	42°16′N	0°44′E
ABR5	Huesca, Jaca	Spain	42°33′N	0°33′W
ABR6	Navarra, Los Arcos	Spain	42.567°N	2.183°W
ABR7	Valladolid, Otero	Spain	39°59′57″N	4°30′46″W
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

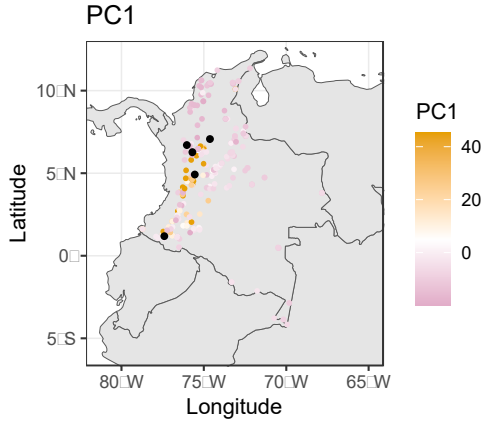
Data compiled from Jenkins *et al.* (2003), Vogel *et al.* (2009) and D. Garvin (pers. comm.).

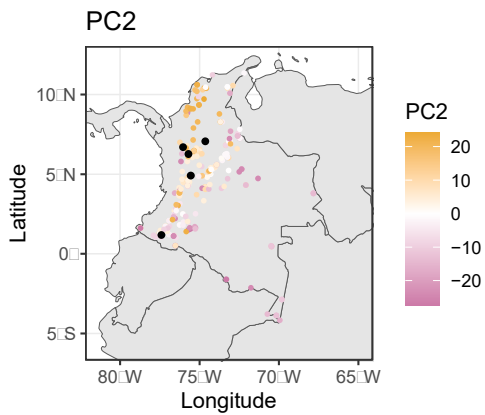




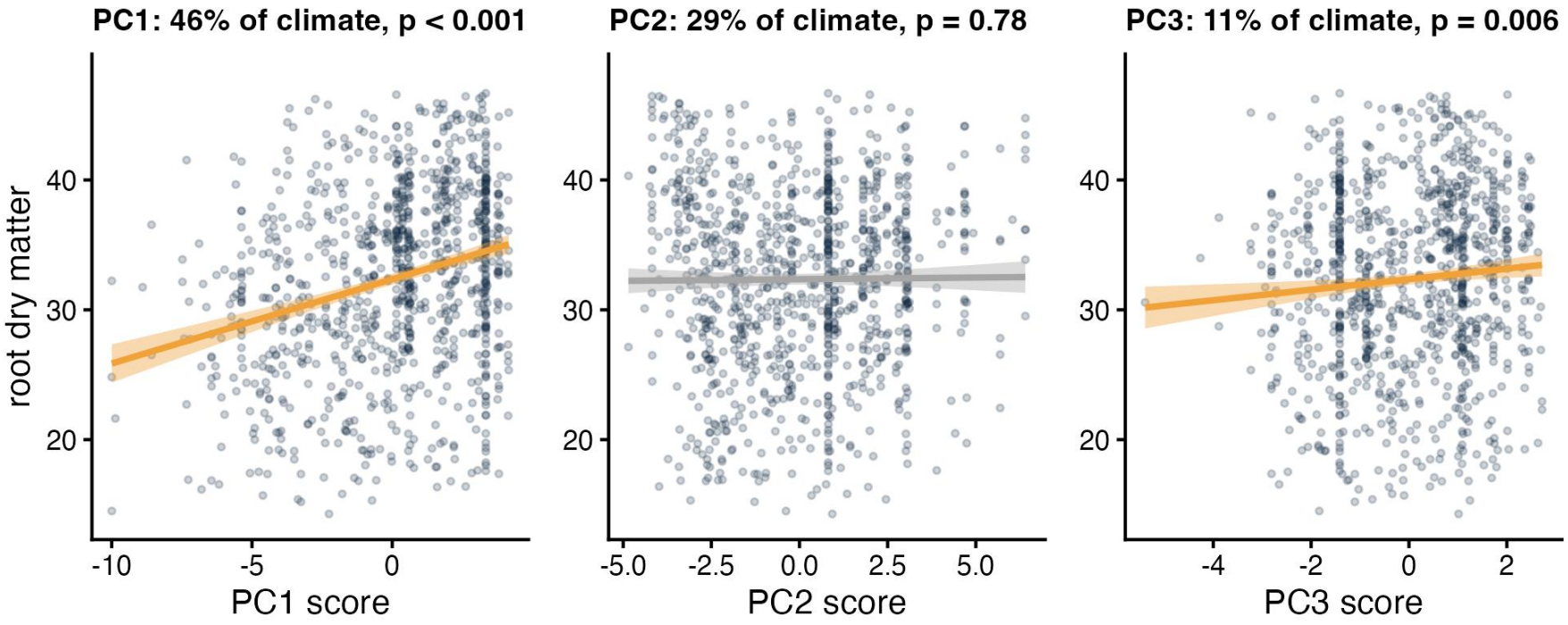

Principal Component Analysis
PC2 (5%)
PC1 (8%)
Elevation (AWS)
3000
2000
1000


Altitude
Latitude
Longitude
Elevation (AWS)
3000
2000
1000


PC1
Latitude
Longitude
PC1
40
20
0


PC2
Latitude
Longitude
PC2
20
10
0
−10
−20

PCs as predictors



PCs are uncorrelated: **VIF = 1**. **Lots of climate variance ≠ lots of effect on dry matter.**

Live coding: PCA



`L12 PCA intro.R`

Iris, then cassava

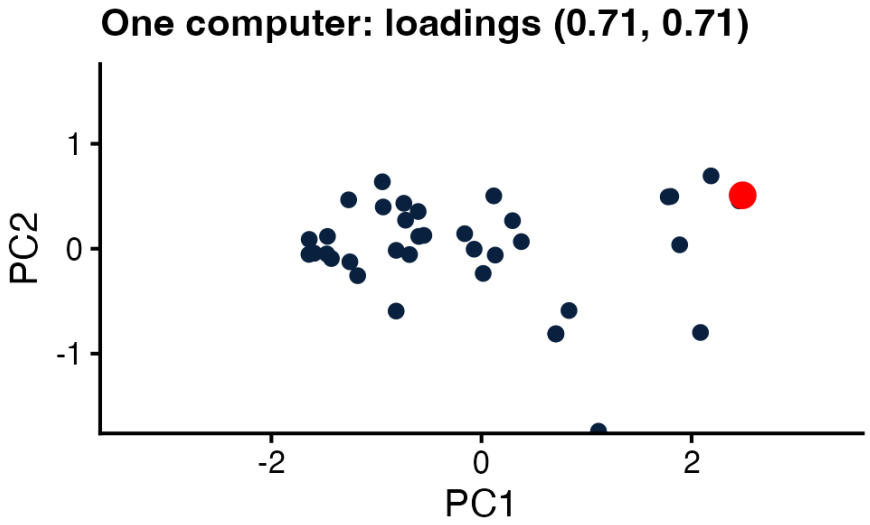
prcomp(scale. = TRUE)

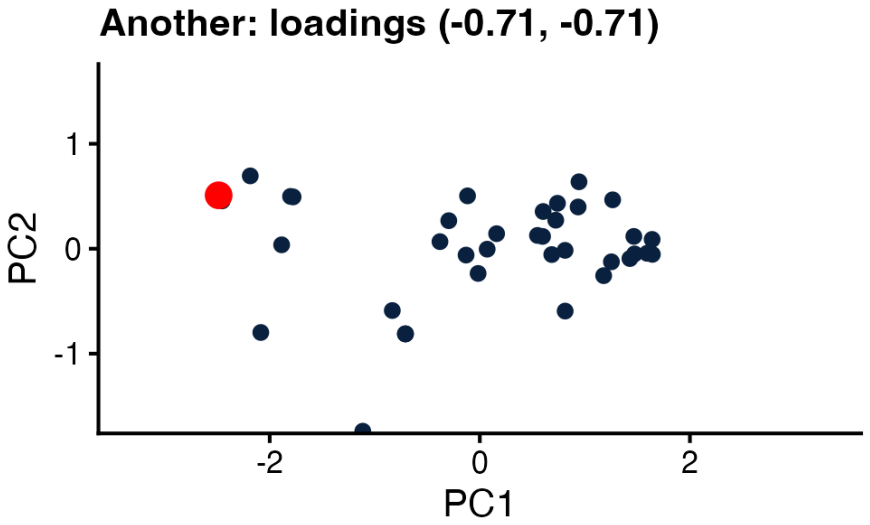
summary(), scree plot

$rotation, $x, $sdev^2

PCs in lm()

The sign is arbitrary





Flip a PC and nothing changes. **Describe it by its variables, not by + or –.**

Live poll

Multiple choice

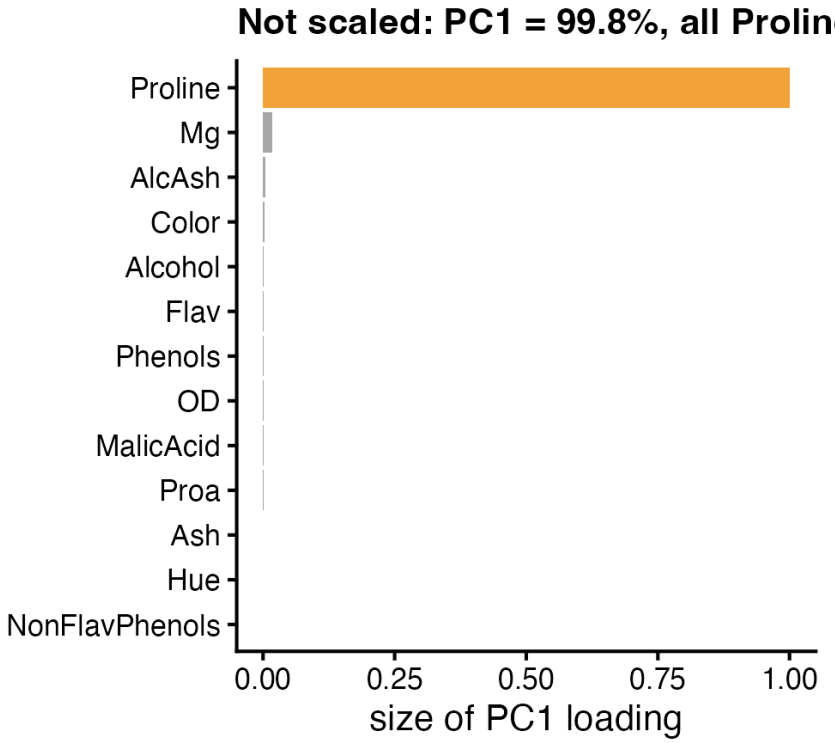
Wine data, PCA WITHOUT scaling: which chemical takes over PC1?

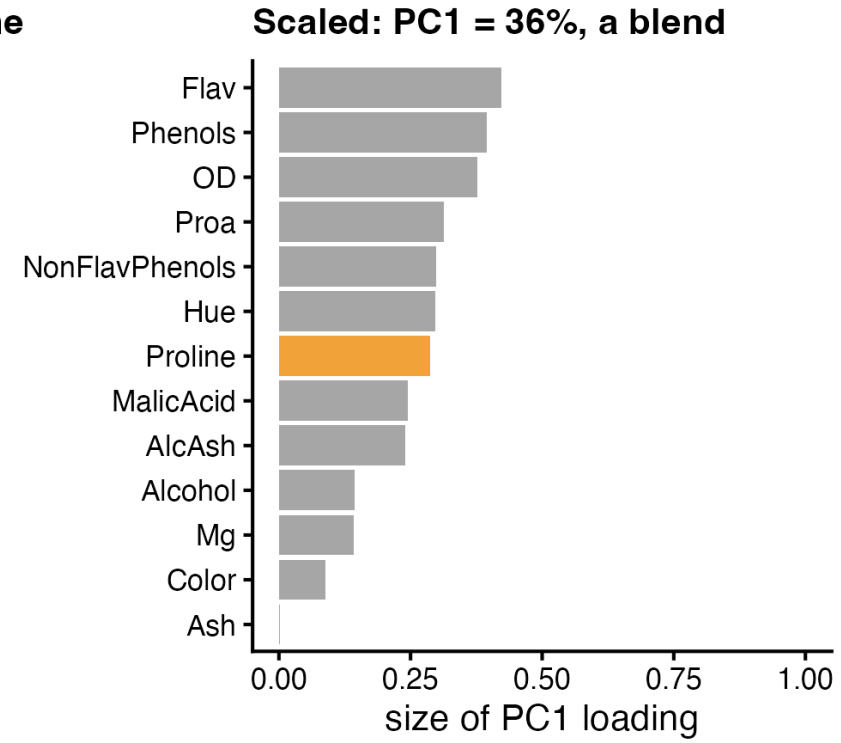
Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w7-scale](greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w7-scale)

[host screen]

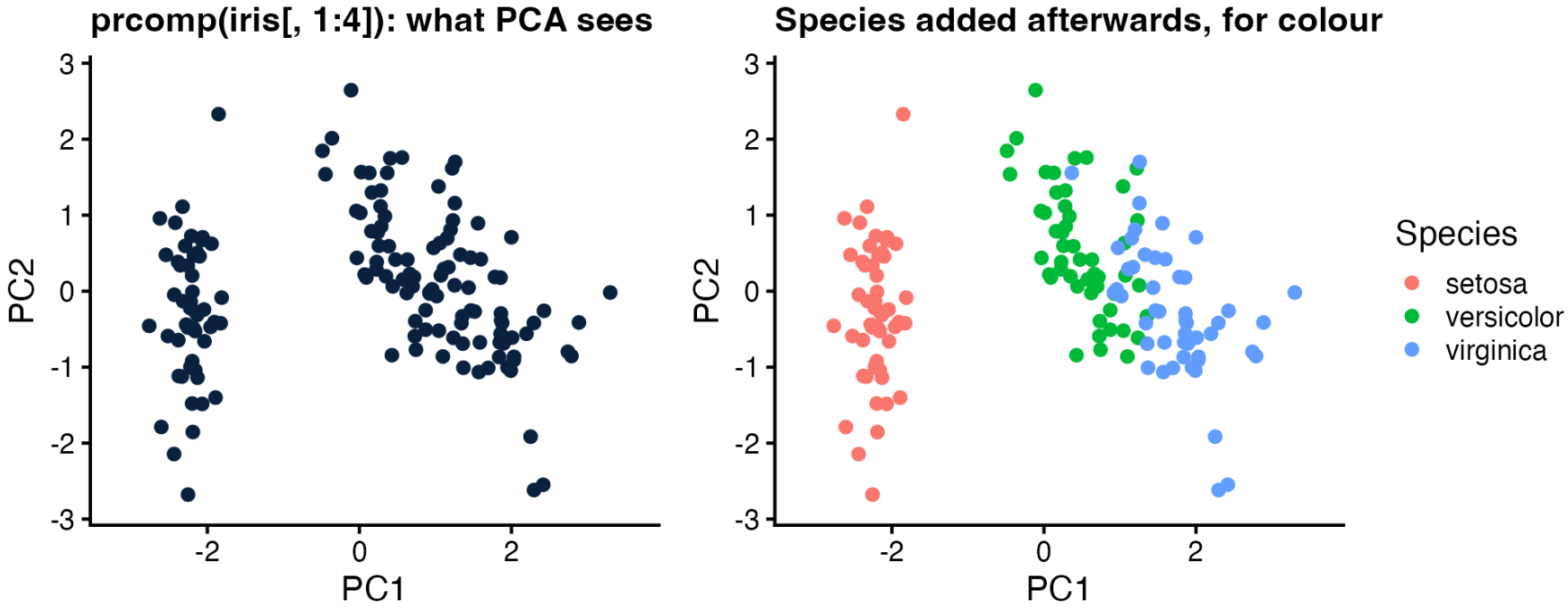
Scale first, or units win





Different units → **prcomp(x, scale. = TRUE)**

PCA never sees the groups

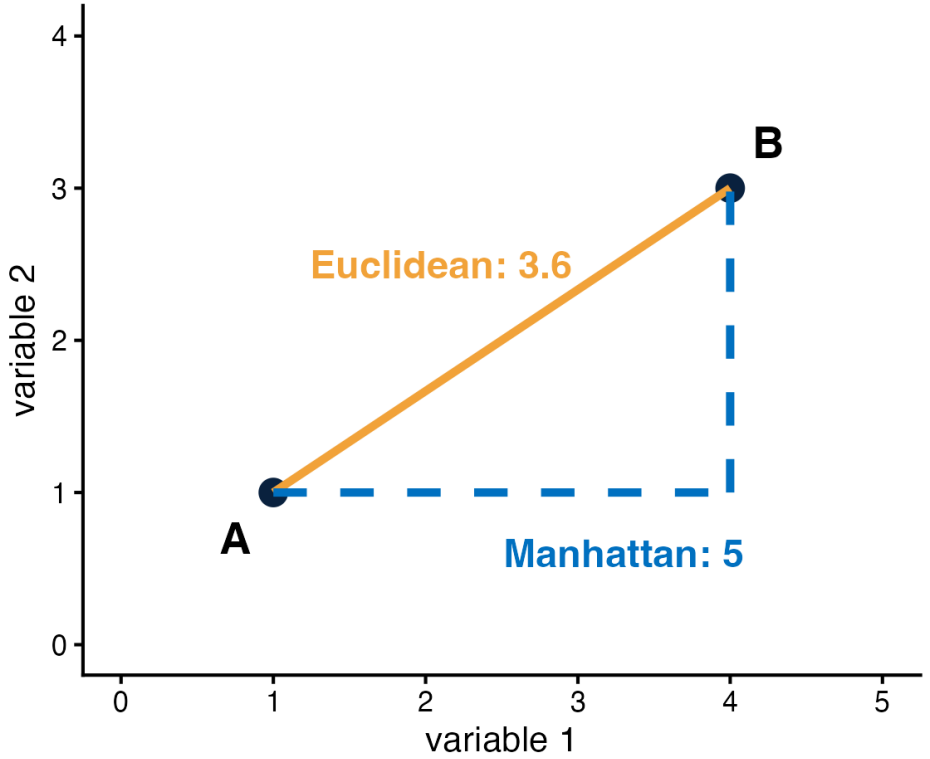


PCA is **unsupervised**: groups or a response are added afterwards.

Where PCA trips people up

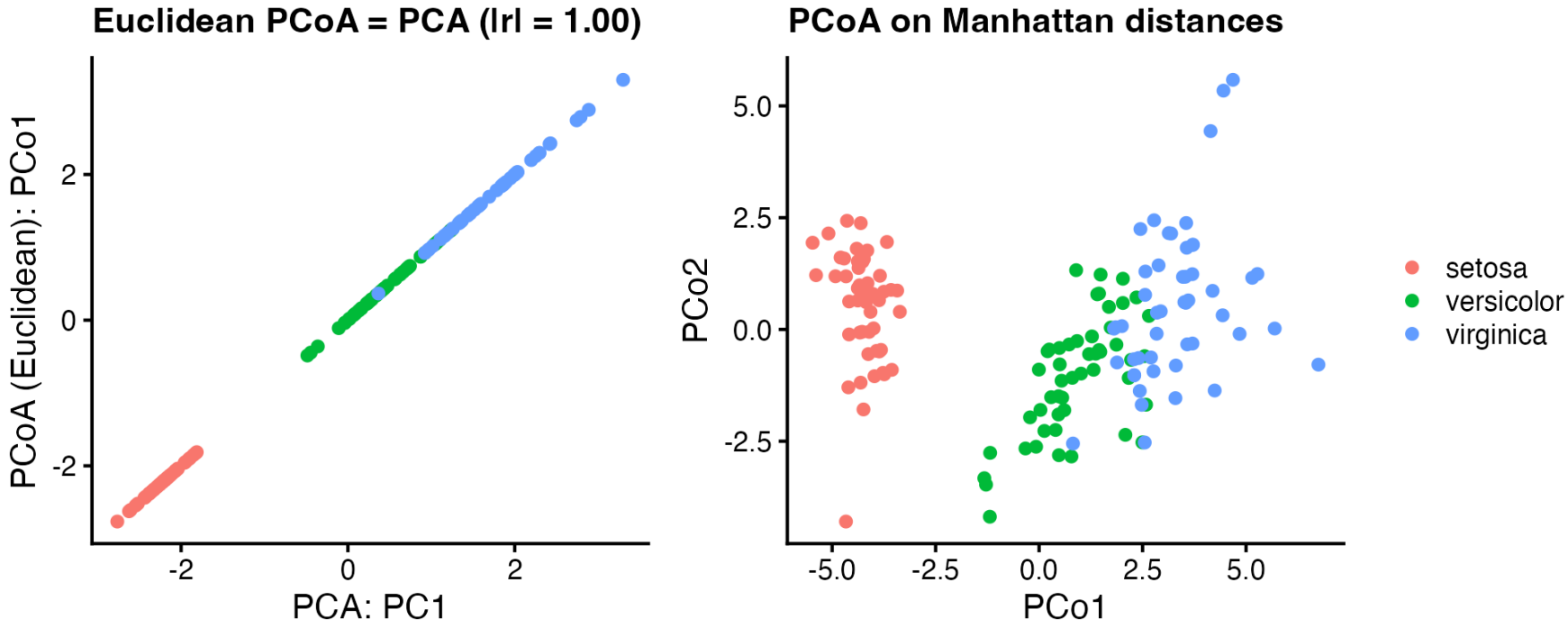
- Loadings describe variables; scores describe samples
- Signs flip: read the variables, not the sign
- Scale when units differ
- % variance is not importance for your y
- PCA is not a regression and sees no groups
- All PCs together = 100%: nothing lost until you drop some

Distance: how far apart are two samples?



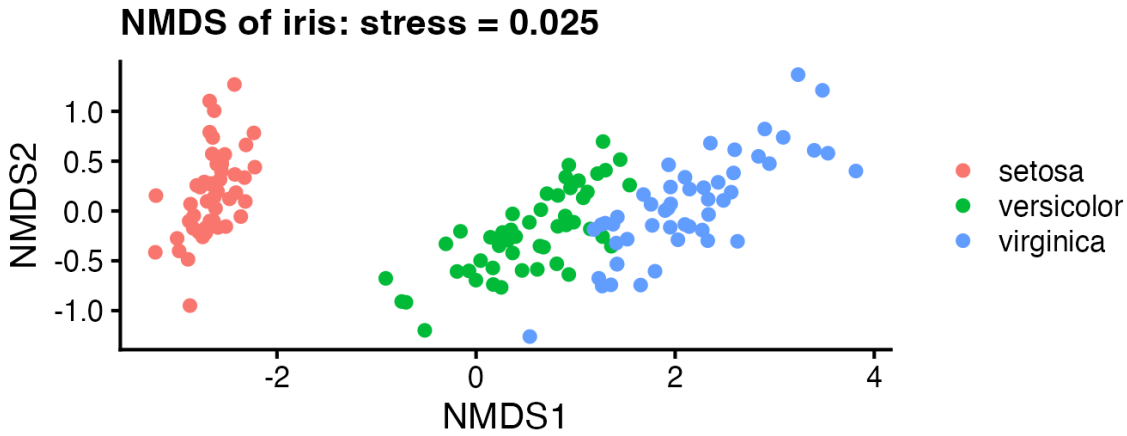
Straight line or city blocks: **the choice of distance changes the map**

PCoA: start from the distances



Any distance matrix in, a map out. **Euclidean PCoA = PCA.**

NMDS: keep the order of the distances



Only ranks matter; **always report the stress** (lower = better fit)

Live poll

Multiple choice

20,000 genes measured in 15,000 single cells. You want a picture of the cell types. Which method?

Scan with your phone

Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w7-method

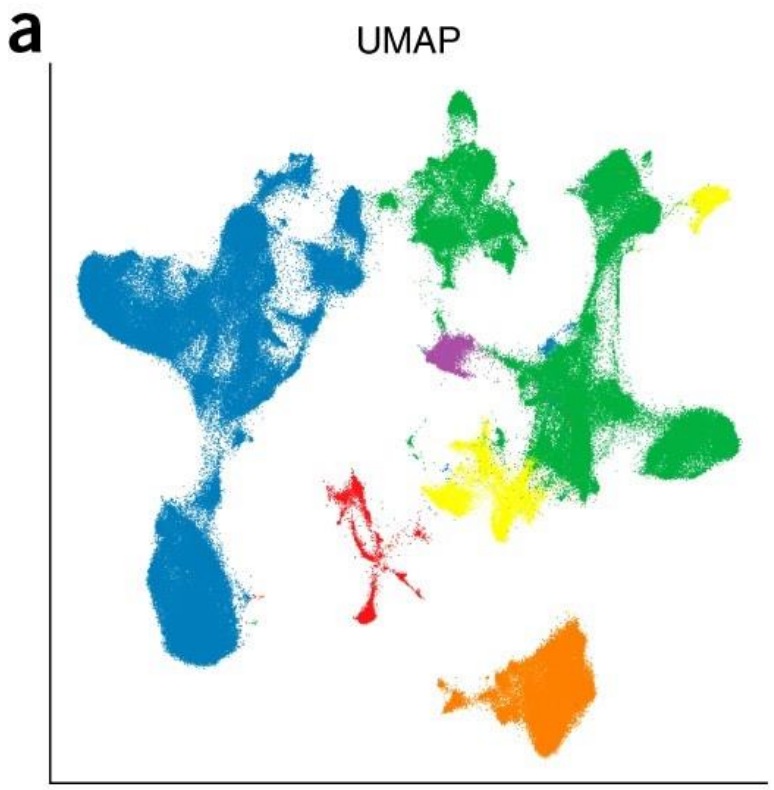
host screen

Gene expression
(30,000+ columns)

Cell type

Cells
(100,000's
rows)

Data



https://www.nature.com/articles/nbt.4314/figures/1

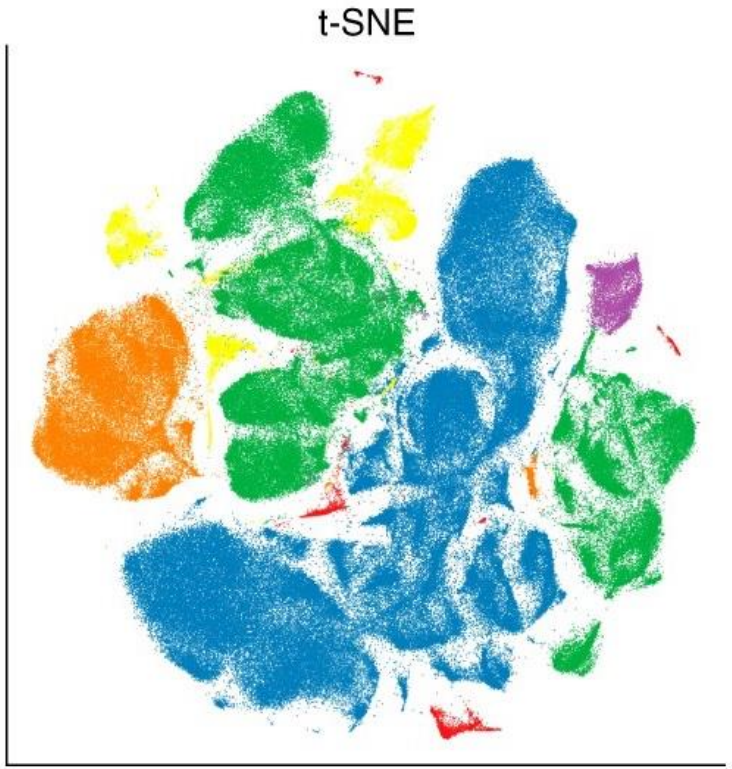
Cell types
Contaminant (including B) CD4 T CD8 T MAIT NK/ILC γδ T

Gene expression
(30,000+ columns)

Cell type

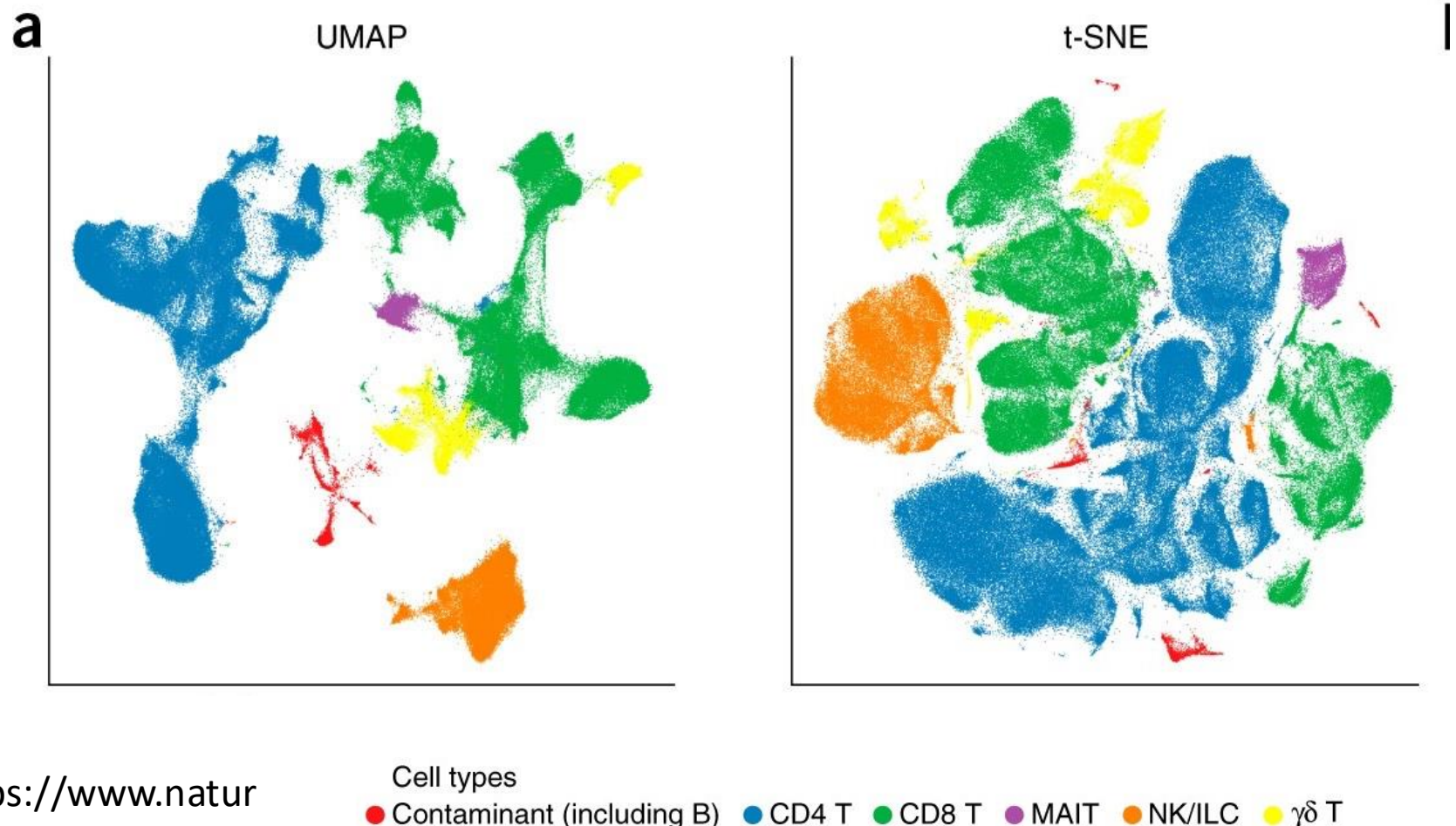
Cells
(100,000's
rows)

Data



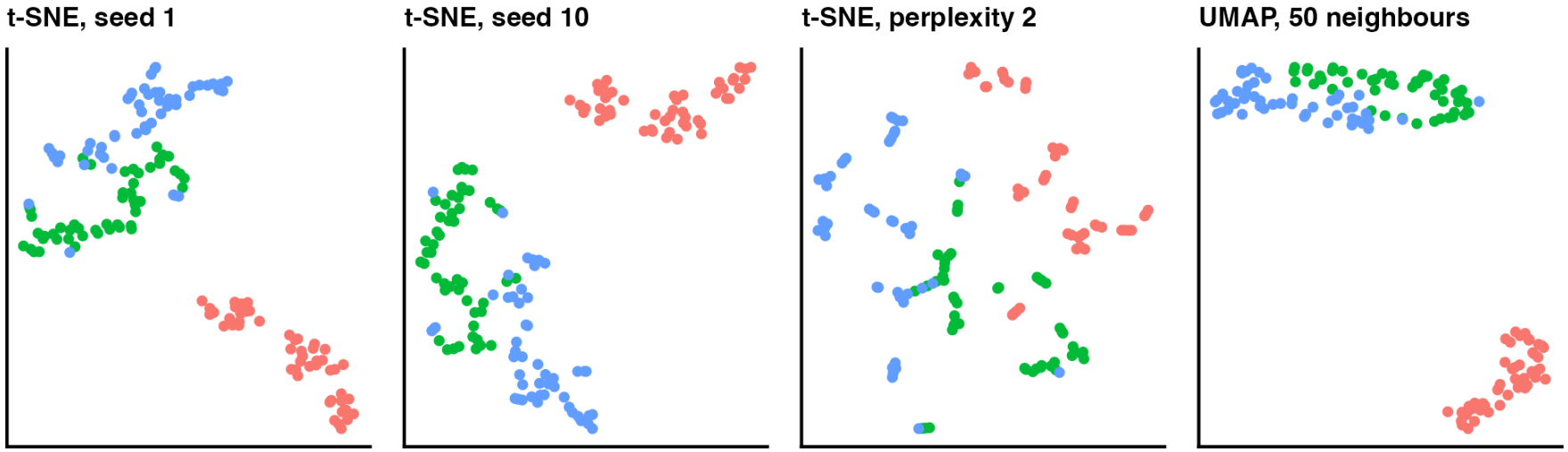
https://www.nature.com/articles/nbt.4314/figures/1





https://www.nature.com/articles/nbt.4314/figures/1

UMAP and t-SNE: settings change the picture



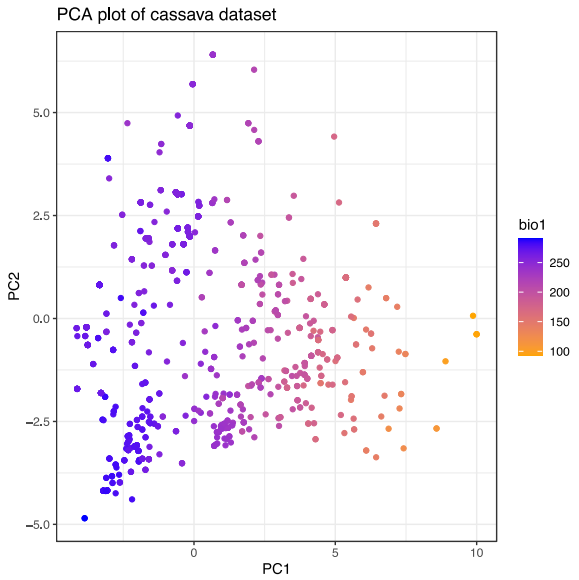
Same iris data every time. **Clusters: maybe. Distances and axes: meaningless.**

PCA vs UMAP vs t-SNE example

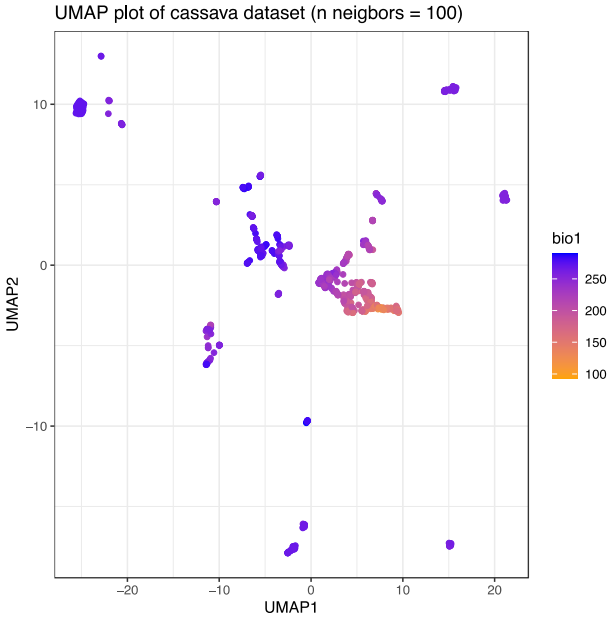
Cassava dataset: points colored by Mean Annual Temperature

You will notice that UMAP and t-SNE are biased toward identify clusters

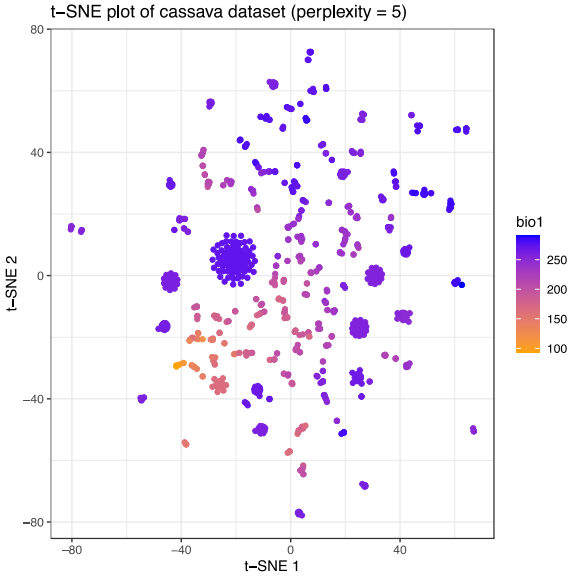
PCA



UMAP



t-SNE



Caution!

MAKING SENSE OF BIG DATA

Why you should not rely on t-SNE, UMAP or TriMAP

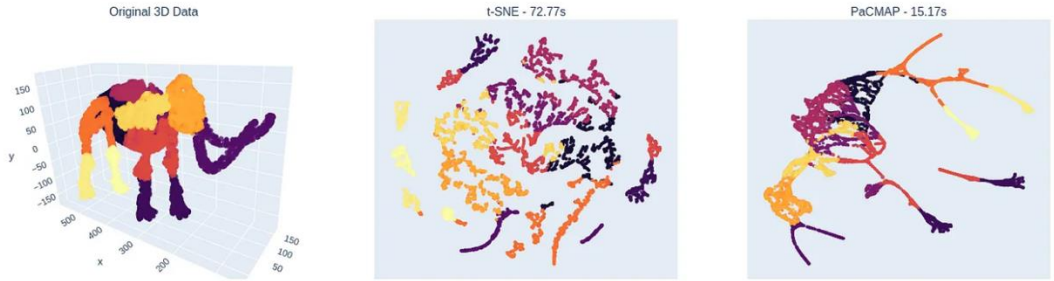
Use PaCMAP instead for increased interpretability & speed

Mathias Gruber · Follow
Published in Towards Data Science · 8 min read · Apr 7, 2021

The specious art of single-cell genomics

Tara Chari, Lior Pachter

Published: August 17, 2023 • https://doi.org/10.1371/journal.pcbi.1011288

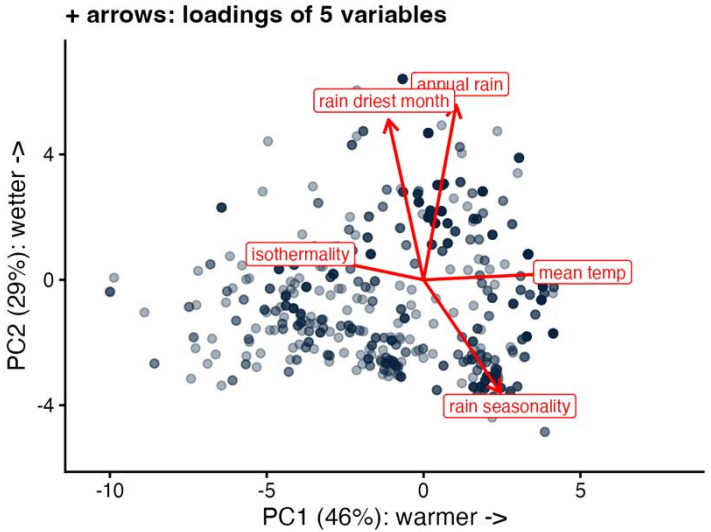


Plot created by author

Which method, when?

PCA	continuous variables	interpretable axes; reuse scores
PCoA	any distance matrix	genetic distance, Bray-Curtis
NMDS	any distance matrix	communities with many zeros; report stress
UMAP / t-SNE	huge data	pictures only

Problem Set 7



Wine: scale, PCs to keep, loadings, scores

Cassava: PCs as predictors

Choose a method

Due Monday Nov 16, 1:10 PM