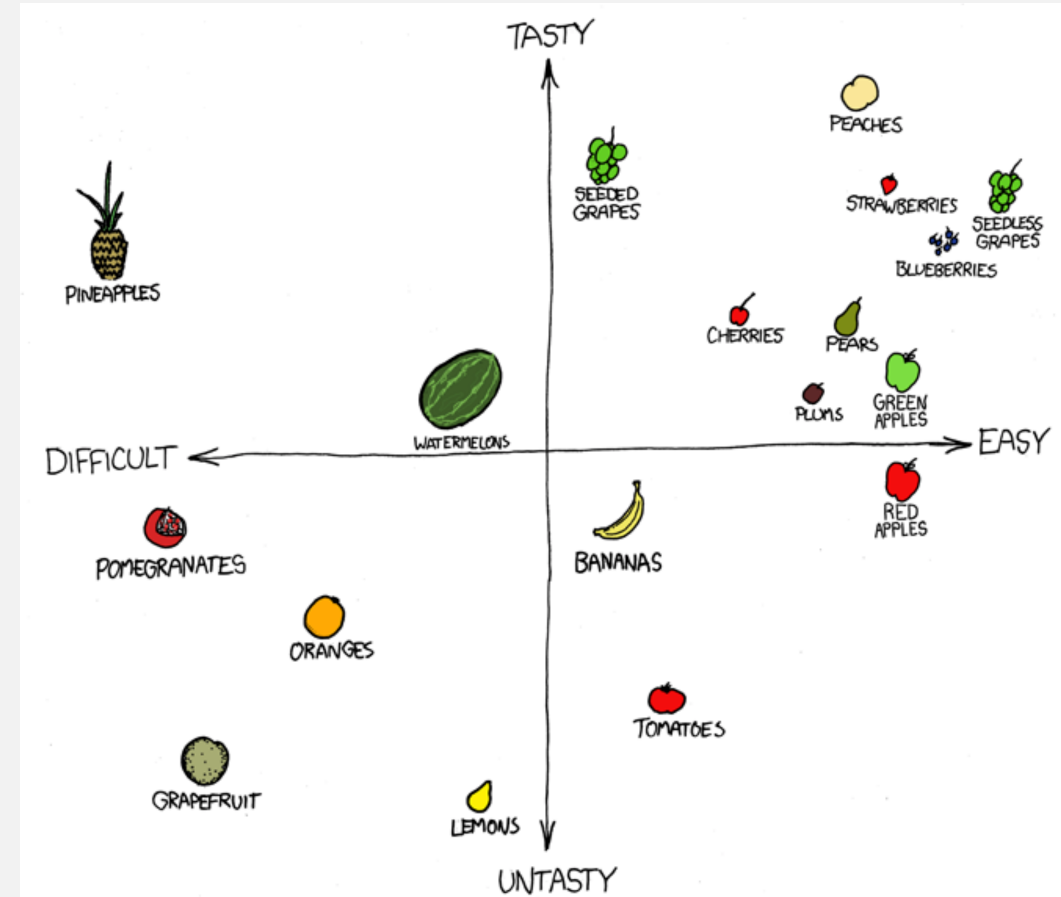# Dimensionality Reduction

Principal component analysis

Corrections for multiple tests

Biology 683

Heath Blackmon

What are multiple comparisons/testing

Horsebean	Linseed	Soybean	Sunflower	Meatmeal	Casein
179	205	189	205	179	205
216	198	186	198	216	198
165	212	175	212	165	212
183	218	193	218	183	218
169	210	149	210	169	210

$$FWER = 1-(1-alpha)^m$$

$$FWER = 1-(1-0.05)^{15}$$

45% chance of false positive

What are multiple comparisons/testing

Horsebean	Linseed	Soybean	Sunflower	Meatmeal	Casein
179	205	189	205	179	205
216	198	186	198	216	198
165	212	175	212	165	212
183	218	193	218	183	218
169	210	149	210	169	210

$$FWER = 1-(1-alpha)^m$$

$$FWER = 1-(1-0.05)^{15}$$

45% chance of false positive

Bonferroni Correction

With the Bonferroni correctio method we divide alpha by number of comparisons being completed.

$0.05/15 = 0.0033333$

$FWER = 1-(1-0.00333)^{15}$

5% chance of false positive

$FWER = 1-(1-0.05)^{15}$

45% chance of false positive

False Discovery Rate

An alternative approach is to decide what proportion of positive results we are ok with having be false positives. This is a common approach in genetic scans (GWAS). The math is bit more complicated but FDR approaches and Bonferroni methods are available in the p.adjust function.

Why do we do dimensional reduction?

1. Datasets are getting bigger and bigger
2. Understanding our data is often the new bottleneck
3. We can't think well beyond 3-4 dimensions
4. We can't illustrate well beyond 2-3 dimensions

What is principal component analysis

PCA is a dimensional reduction tool that takes many (possibly correlated) measurements and transforms it into a smaller set of uncorrelated measuerments.
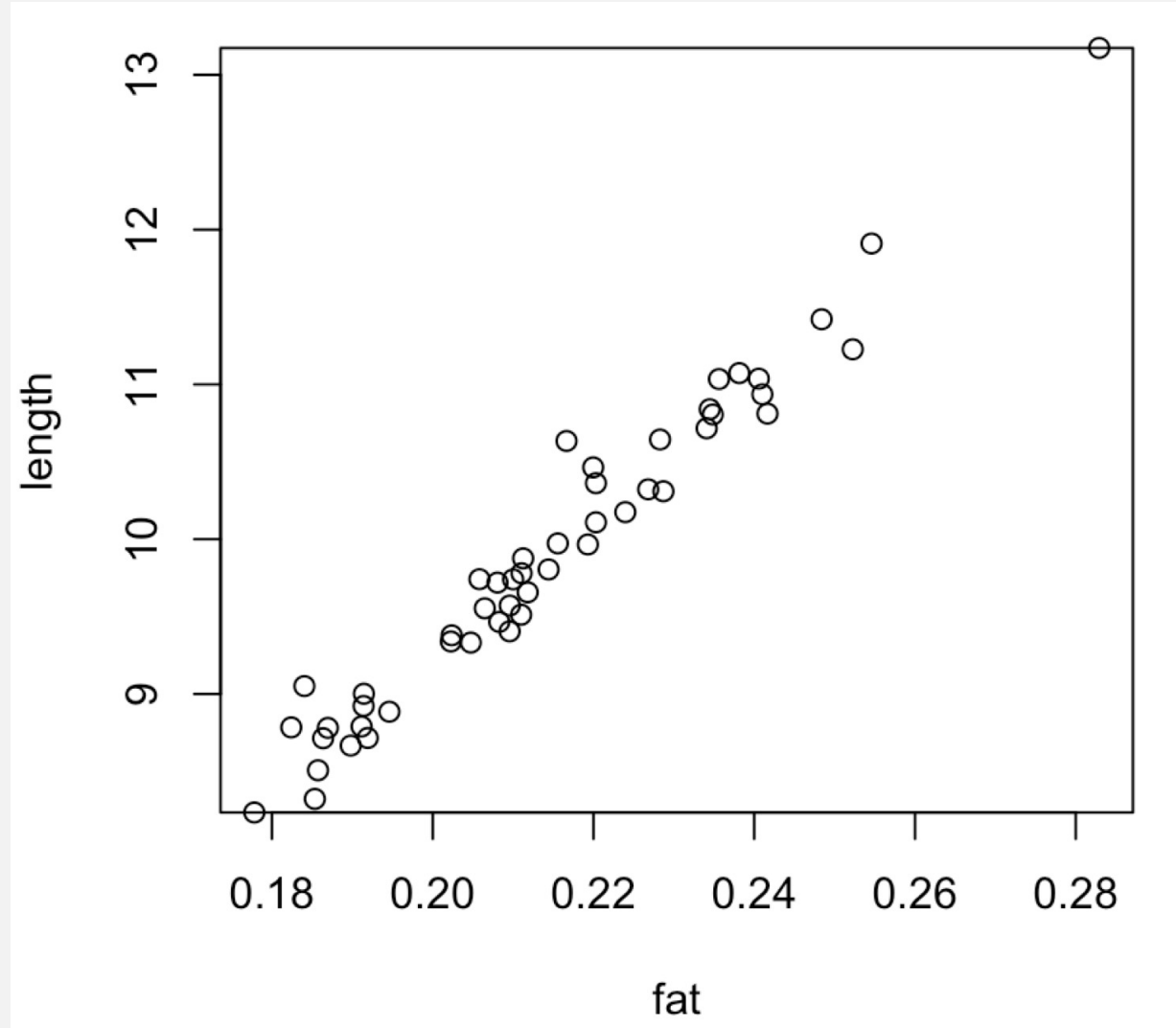
One dimensional data

fish_ID	length
379	9.74
430	9.47
55	13.17
434	10.63
147	9.38
497	8.78
127	8.79
362	8.92
107	11.07
116	9.57
414	10.84
232	8.89
193	9.55
80	10.81
283	8.23
68	8.67
341	9.8
288	9.05
200	11.23
62	9.97
69	10.94
466	11.42
400	10.31
83	10.64
106	9.4
280	9
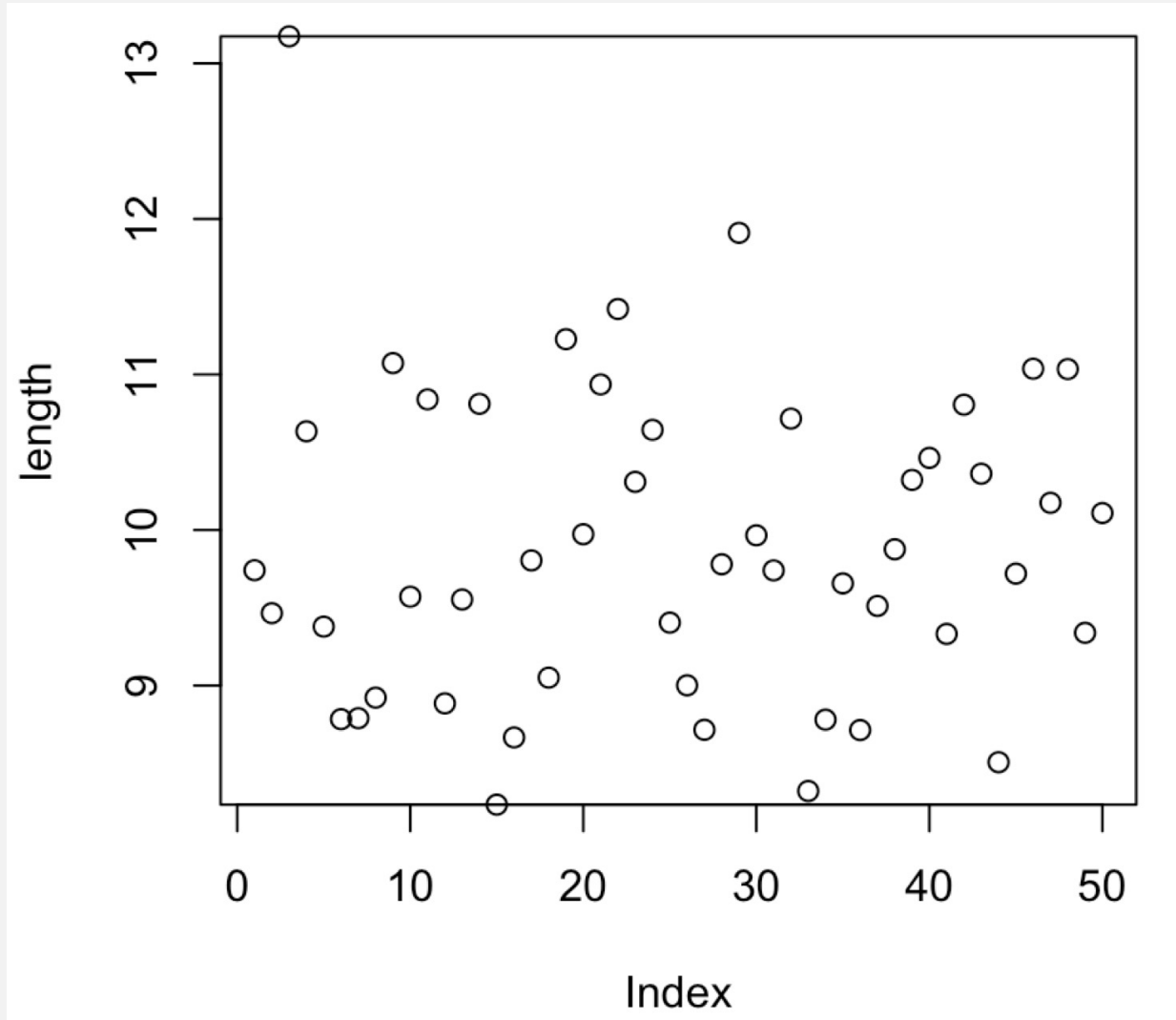
Two-dimensional data

fish_ID	length	fat
379	9.74	0.21
430	9.47	0.2
55	13.17	0.28
434	10.63	0.22
147	9.38	0.28
497	8.78	0.26
127	8.79	0.24
362	8.92	0.25
107	11.07	0.24
116	9.57	0.25
414	10.84	0.26
232	8.89	0.17
193	9.55	0.23
80	10.81	0.32
283	8.23	0.24
68	8.67	0.25
341	9.8	0.19
288	9.05	0.26
200	11.23	0.23
62	9.97	0.22
69	10.94	0.18
466	11.42	0.21
400	10.31	0.23
83	10.64	0.19

High dimensional data

fish_ID	length	fat	m3	m4	m5	m6	m7
379	9.74	0.21	9.74	0.21	0.27	1.45	2.97
430	9.47	0.2	9.47	0.2	4.28	0.83	1.57
55	13.17	0.28	13.17	0.28	7.01	1.25	4.59
434	10.63	0.22	10.63	0.22	15.33	0.89	2.09
147	9.38	0.28	9.38	0.28	5.77	0.77	2.03
497	8.78	0.26	8.78	0.26	18.81	0.61	1.39
127	8.79	0.24	8.79	0.24	8.92	1.24	2.62
362	8.92	0.25	8.92	0.25	0.44	0.98	2.19
107	11.07	0.24	11.07	0.24	16.89	0.84	2.24
116	9.57	0.25	9.57	0.25	13.55	1.26	3.02
414	10.84	0.26	10.84	0.26	11.23	0.2	0.57
232	8.89	0.17	8.89	0.17	2.02	0.94	1.42
193	9.55	0.23	9.55	0.23	5.24	1.05	2.32
80	10.81	0.32	10.81	0.32	6.04	0.79	2.73
283	8.23	0.24	8.23	0.24	0.12	0.07	0.13
68	8.67	0.25	8.67	0.25	2.76	0.41	0.89
341	9.8	0.19	9.8	0.19	9.73	1.59	2.97
288	9.05	0.26	9.05	0.26	15.31	0.49	1.16
200	11.23	0.23	11.23	0.23	3.04	1.26	3.24
62	9.97	0.22	9.97	0.22	13.14	0.28	0.61
69	10.94	0.18	10.94	0.18	5.23	0.14	0.28
466	11.42	0.21	11.42	0.21	6.11	0.3	0.71
400	10.31	0.23	10.31	0.23	2.78	0.43	1.02
83	10.64	0.19	10.64	0.19	10.93	0.86	1.73

What are our options?

Dimensionality reduction - PCA

The math behind - PCA

M measurements

fish_ID	length	fat	m3	m4	m5	m6	m7
379	9.74	0.21	9.74	0.21	0.27	1.45	2.97
430	9.47	0.2	9.47	0.2	4.28	0.83	1.57
55	13.17	0.28	13.17	0.28	7.01	1.25	4.59
434	10.63	0.22	10.63	0.22	15.33	0.89	2.09
147	9.38	0.28	9.38	0.28	5.77	0.77	2.03
497	8.78	0.26	8.78	0.26	18.81	0.61	1.39

N samples

= **X**

nxm matrix

PCA is an Eigendecomposition of $(X^T X)$

W = loadings
mxm matrix

$T = X \cdot W$

fish_ID	PC1	PC2	PC3	PC4	PC5	PC6	PC7
379	0.48401035	0.3204638	0.86441909	0.75599341	0.19841263	0.29382035	0.87356767
430	0.92864946	0.7321531	0.17546787	0.22461749	0.50242958	0.66343818	0.82559145
55	0.12804522	0.03920335	0.59362519	0.39215655	0.9883746	0.92951871	0.26687076
434	0.2062225	0.21152673	0.47047215	0.66007816	0.71549665	0.89006466	0.31692619
147	0.85635814	0.17733842	0.94029393	0.11666265	0.10601835	0.18949868	0.42276984
497	0.78104386	0.88445893	0.45355361	0.61294154	0.02441031	0.3948512	0.08124852

= **T**

nxm matrix

```
pca <- prcomp(X)
pca$x
```

This is T a nxm matrix that has our PC scores. The first column of the matrix will be the one that captures the most variation.
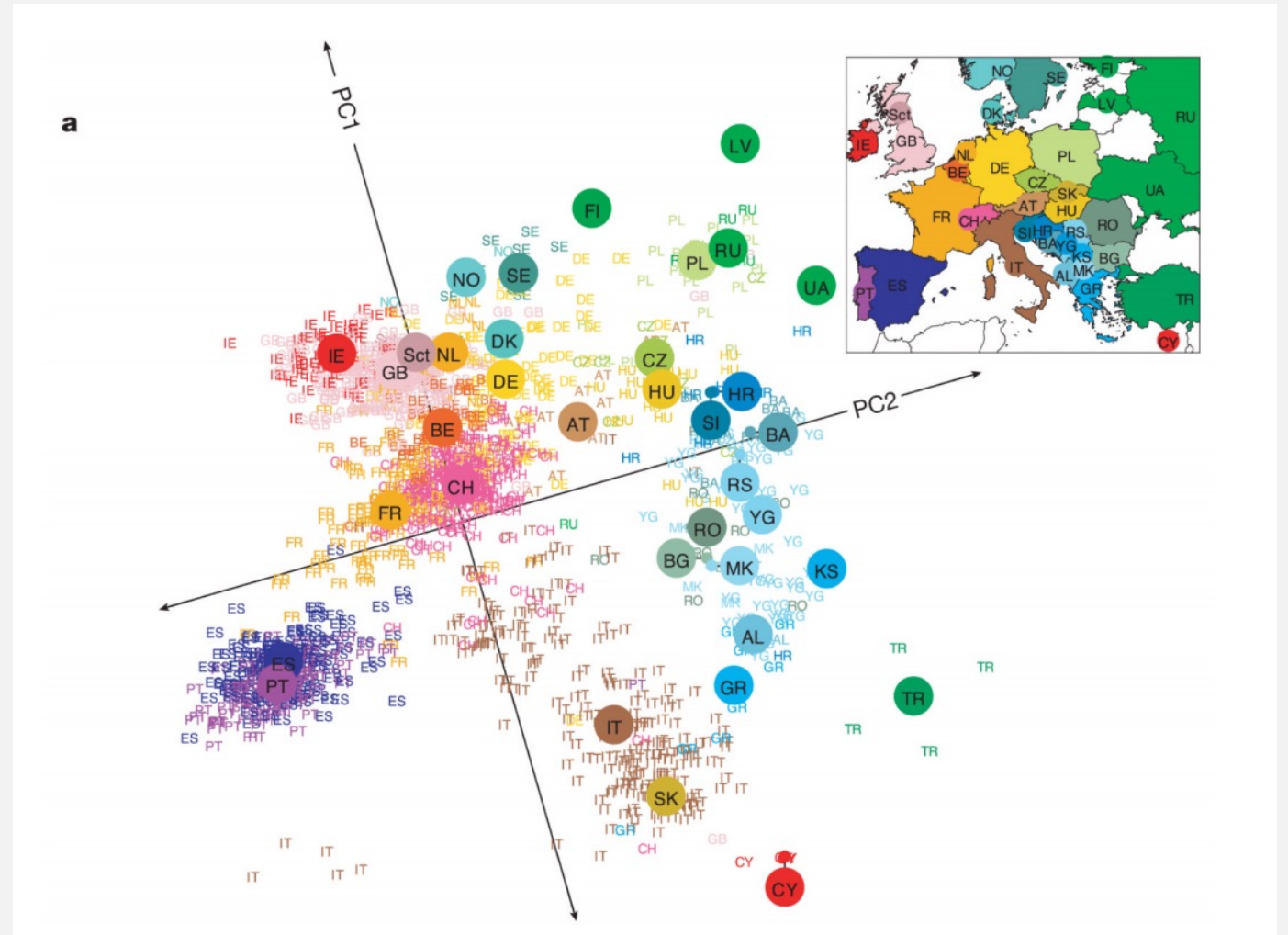
How do people use PCA

Input data: 500,000 SNP genotypes for 3000 Europeans.
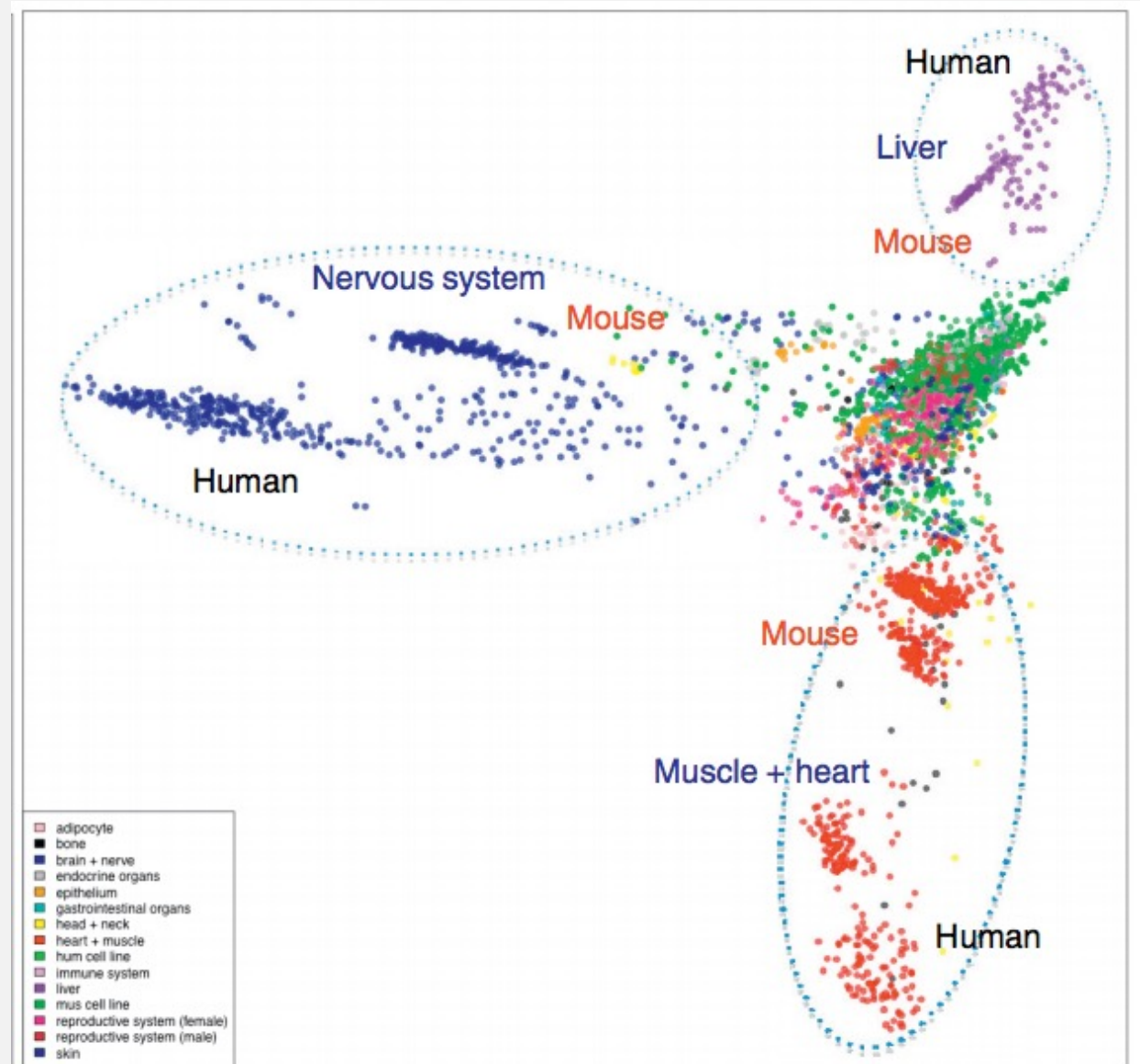
3000 rows
500,000 columns

What are PC1 and PC2?

How do people use PCA

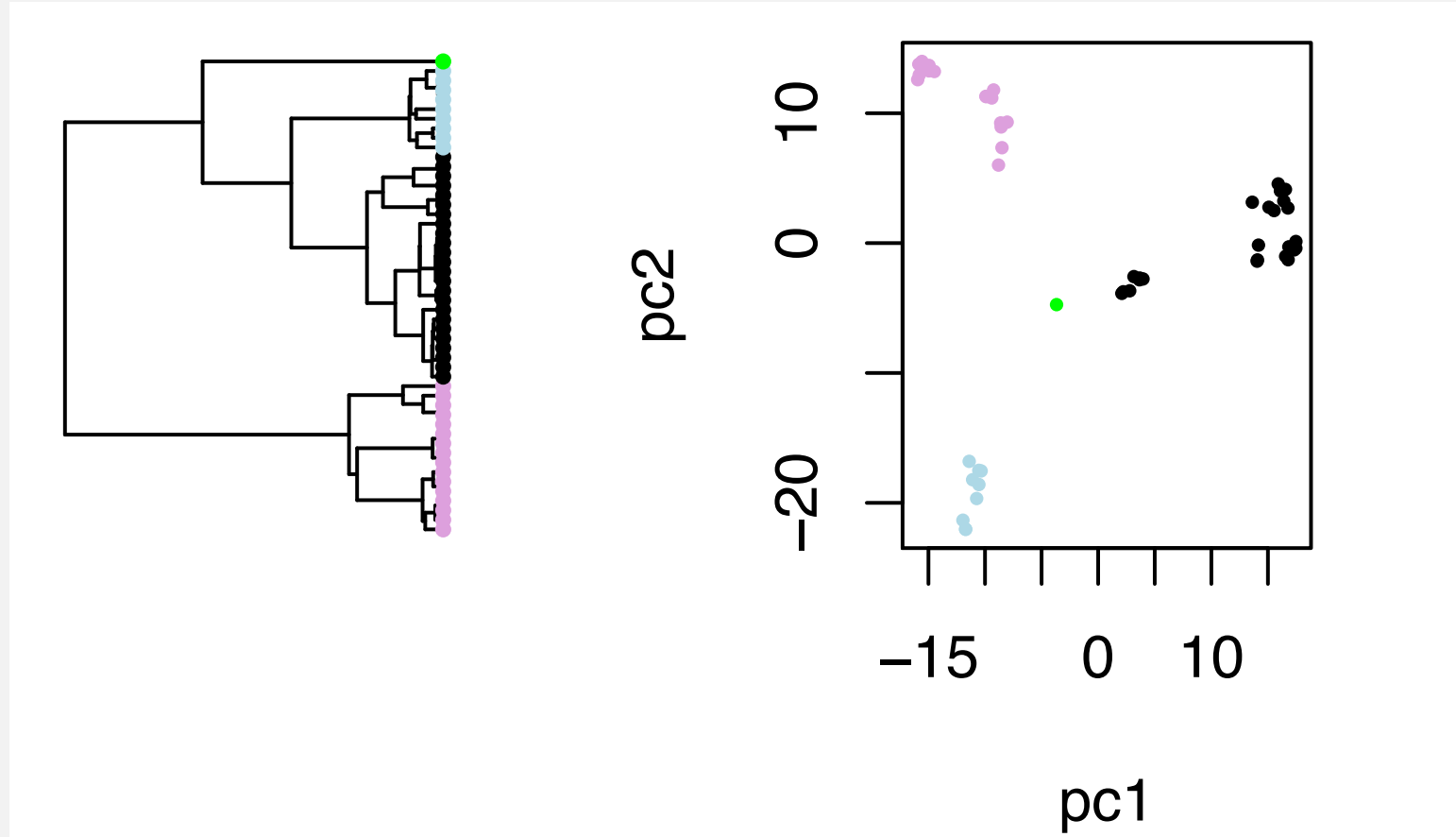Input data:
Expression level for 1000s of genes
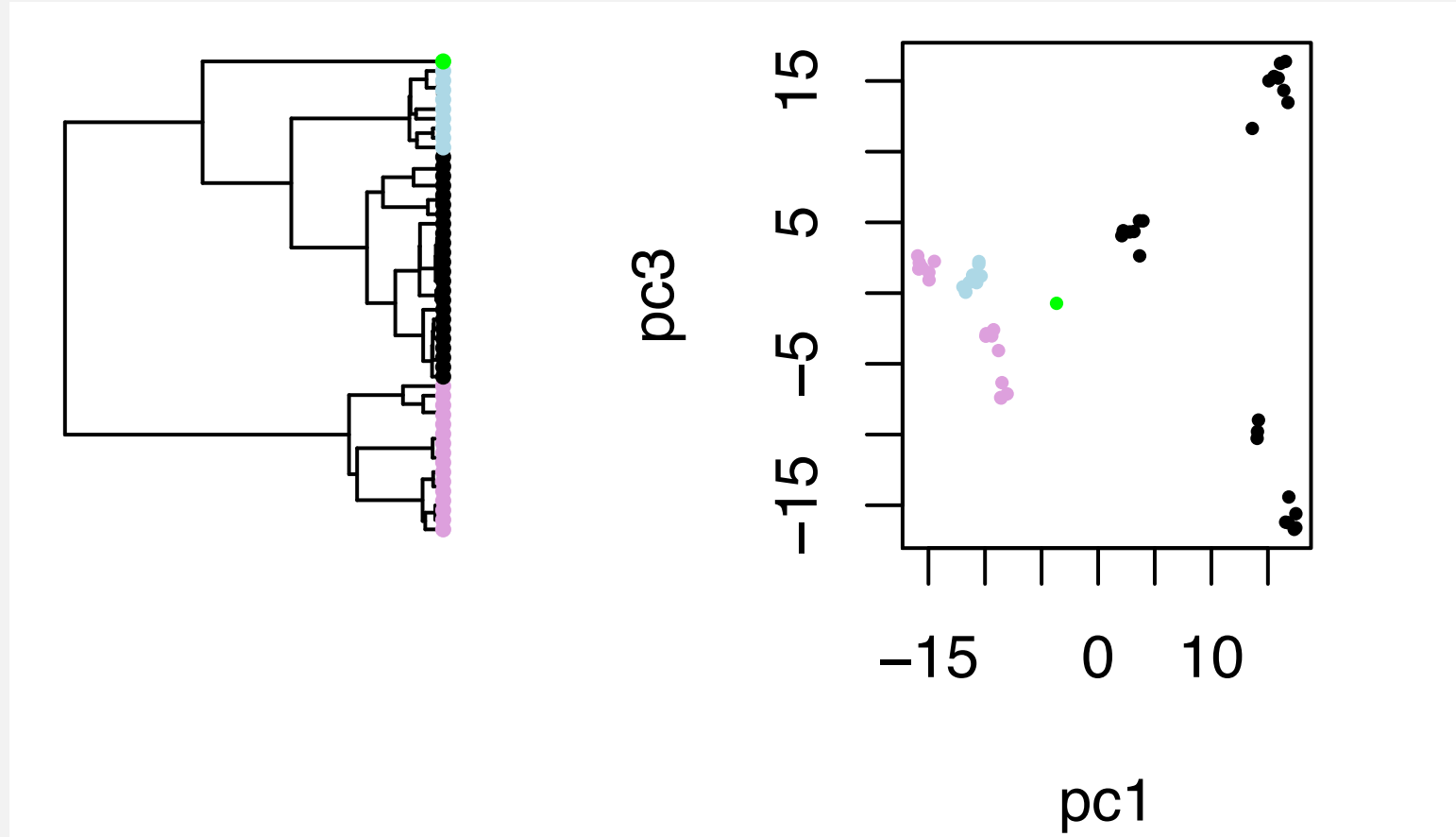In 100s of cells (color indicates type of cell)

What are the PCs here?

Input data:
Radseq data (genotypes at 100s of loci)
For a large number of species or strains
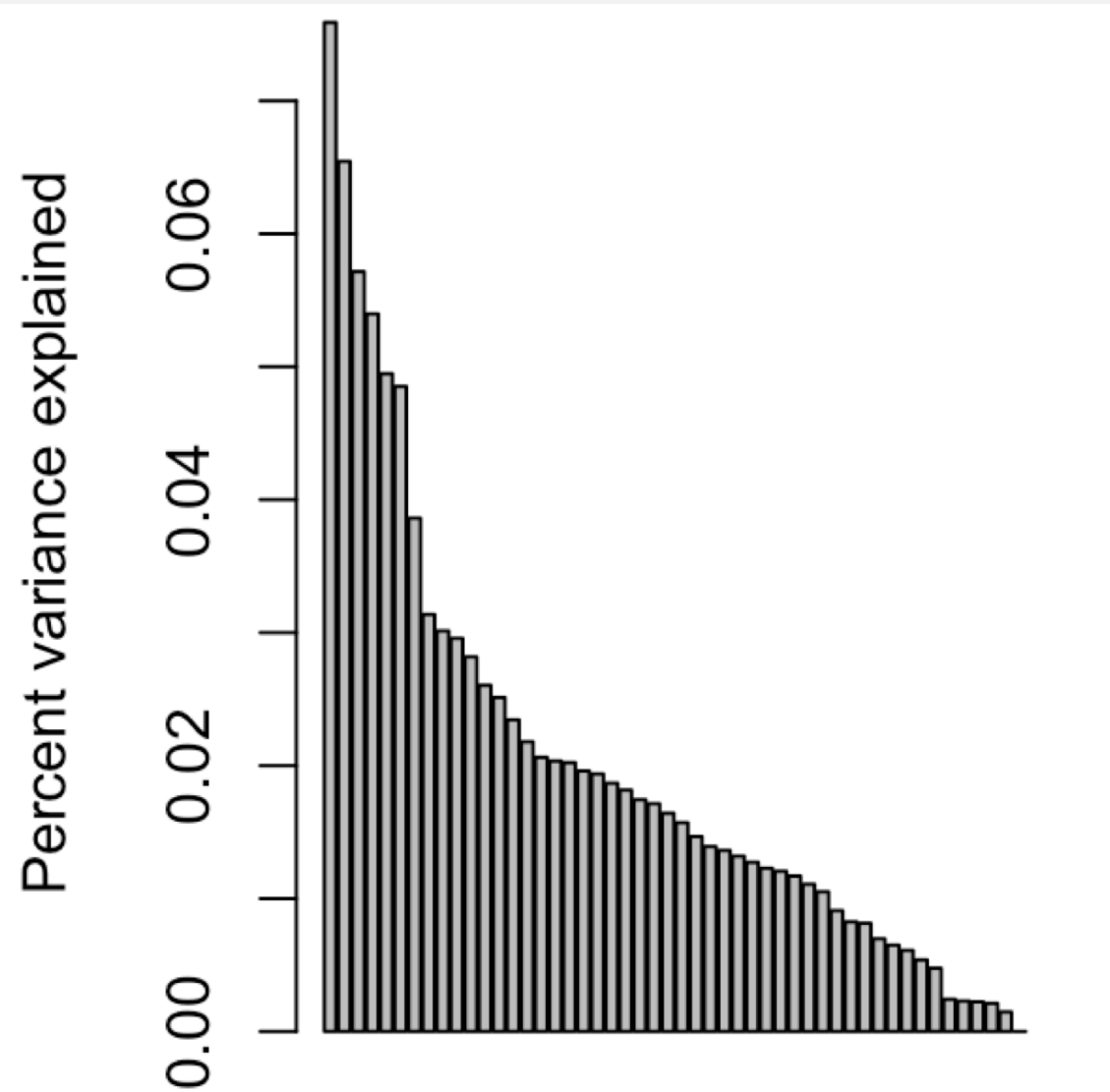
Input data:
Radseq data (genotypes at 100s of loci)
For a large number of species or strains

Scree Plot: A plot that illustrates the proportion of total variance that is captured by each principal component.

Steep means you can greatly reduce dimensionality without losing information

Alternatives

Discriminate function analysis: This is similar to PCA but you assign groupings to the data first and the discriminating components best parse your assigned groups from one another.

Example in Rmarkdown

Standard Packages

stats – this is part of the base install and has the function `prcomp`

New Packages

car – `data.ellipse` function

FactoMineR – `PCA` function