06. Comparing samples and plotting

Erik Dassi & Davide Bressan

Genomics Technologies Lab

Where are we ?

1. *QC*

2. *Alignment*

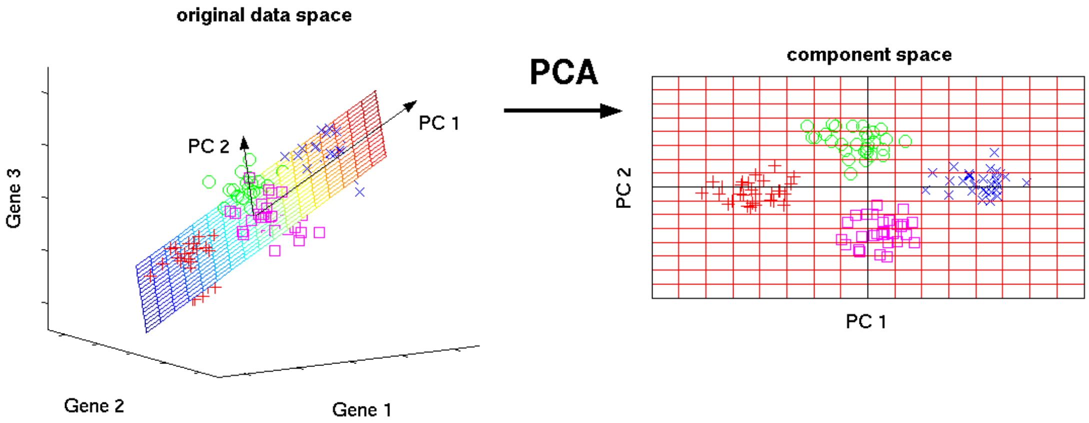
3. *Variant calling*

4. *Variant annotation:* you have obtained a VCF-format file containing all the variants identified in your cell line along with their functional annotation

5. **Comparing samples to identify similarities/differences**

PCA

Principal Component Analysis



- Technique that allows to simplify the input dataset by **reducing the number of features** within it

- Each observation (*sample*) can be represented in an **n**-dimensional space but **not** all these **n** dimensions (*genes*) are interesting and informative

- PCA identifies a **small set of informative dimensions**, i.e. that are able to **capture the most variance** (*variability*) in the input dataset

PCA

- The new dimensions, called ***principal components***, are linear combinations of the input dataset features (genes)

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

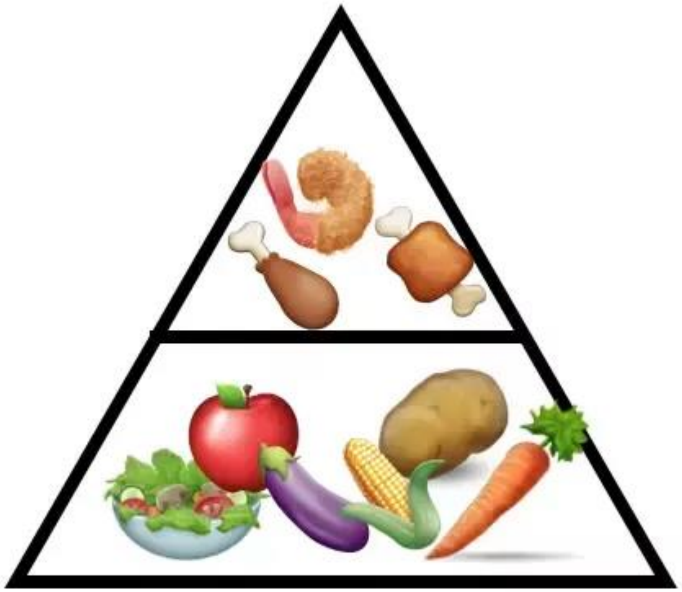
Xi is the observations vector of feature (gene) **i**
Φi1 are the weights of each feature (gene) in principal component **1**

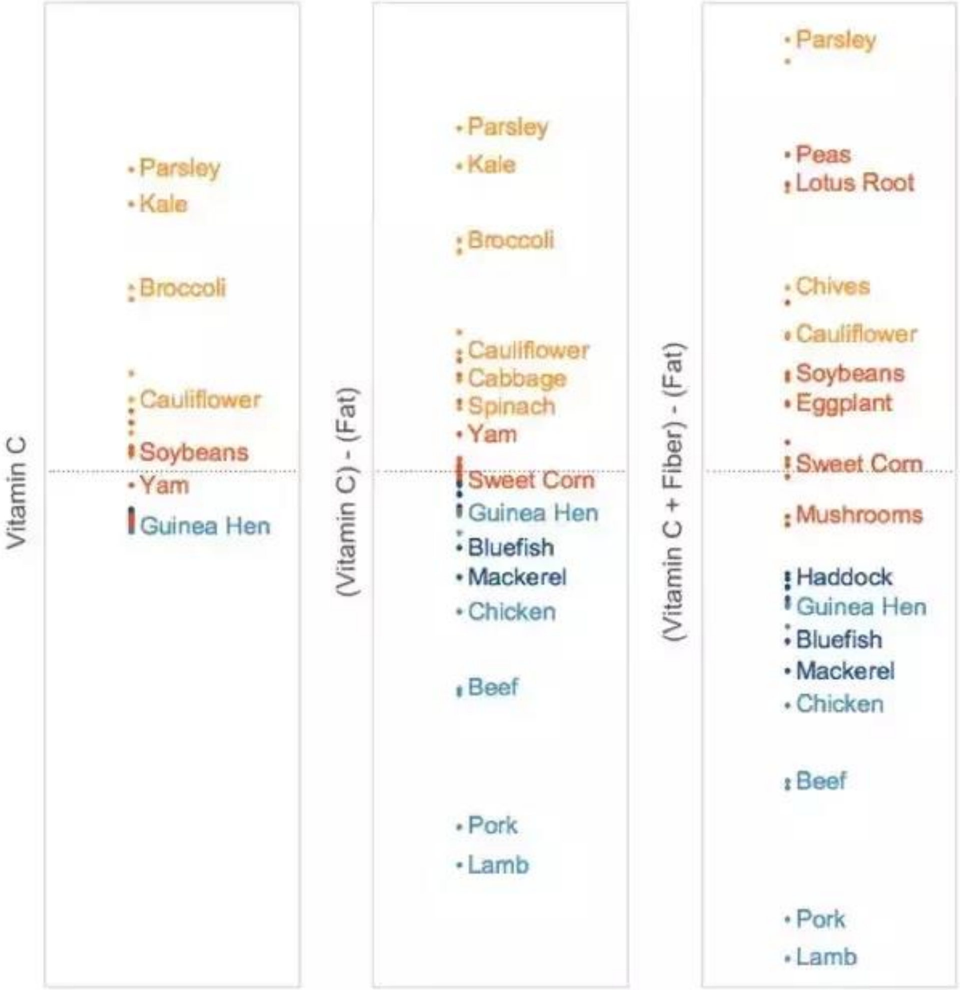
- The first principal component is the one with the **biggest variance**, so the first dimension will be the one ***separating the samples the most***

- The second principal component will be the one with the **highest variance** among the linear combinations that are **independent** (*orthogonal*) of the first principal component

PCA

What is the best way to differentiate food items ?

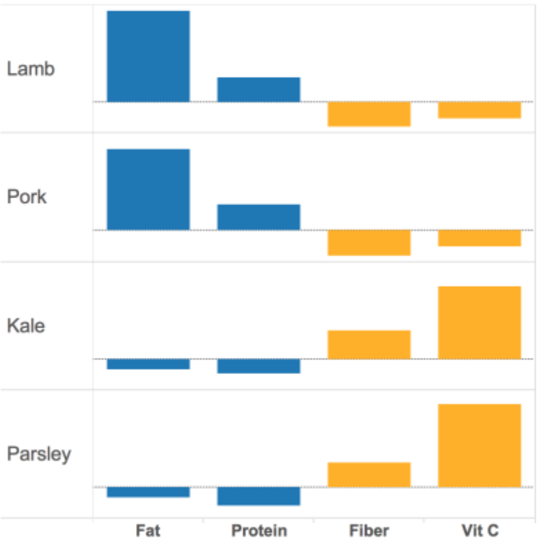


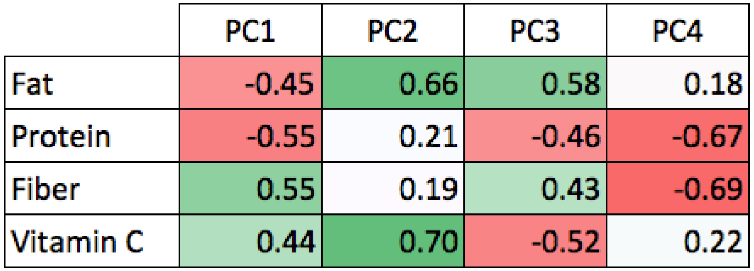


https://algobeans.com/2016/06/15/principal-component-analysis-tutorial/

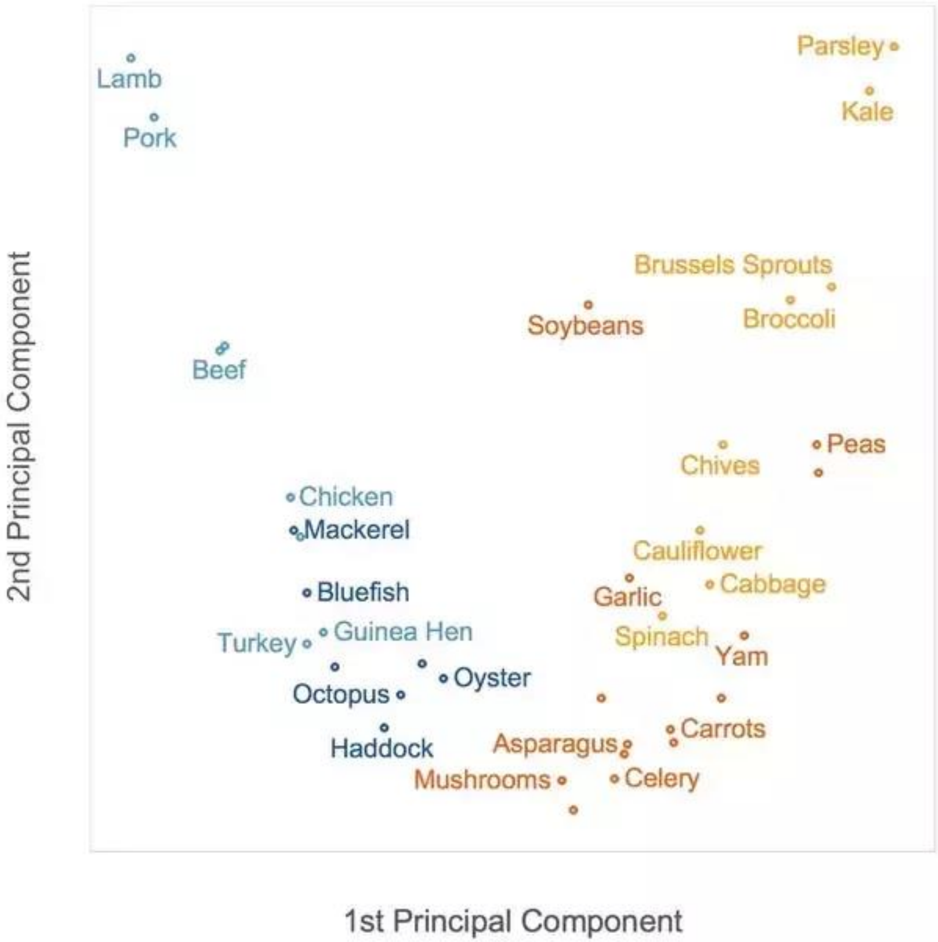
PCA

What is the best way to differentiate food items ?





	PC1	PC2	PC3	PC4
Fat	-0.45	0.66	0.58	0.18
Protein	-0.55	0.21	-0.46	-0.67
Fiber	0.55	0.19	0.43	-0.69
Vitamin C	0.44	0.70	-0.52	0.22



https://algobeans.com/2016/06/15/principal-component-analysis-tutorial/

1. Preparing the merged VCF

We must first correct the VCFs from Mutect to be compliant with the format specifications:

```
./correctAllVCFinFolder.sh
```

Then, we need to merge the variants for each sample into a single VCF:

```
bcftools merge *corrected.vcf.gz -o
AllSamplesMerged_chrM.vcf.gz

bcftools index AllSamplesMerged.vcf.gz
```

2. Computing the PCA

- We’re going to use the *SNPRelate* R package through *RStudio*
- Let’s open a new project into our working folder

```
# load the package
library("SNPRelate")

# open the VCF file and convert it to GDS format
vcf_file <- "AllSamplesMerged_chrM.vcf.gz"
snpgdsVCF2GDS(vcf_file, "AllSamplesMerged_chrM.gds", method="biallelic.only")

# load the GDS file
genofile <- snpgdsOpen("AllSamplesMerged_chrM.gds")

# compute the PCA
variantsPca <- snpgdsPCA(genofile, autosome.only=F)
```

3. Plotting the PCA

- Let’s set a different color for each sample
- We also write the name of the sample next to the corresponding point
- We add some jitter to space the labels

```
# choose 10 colors, one per sample
jColors <- c('chartreuse3', 'cornflowerblue', 'darkgoldenrod1’,
         'peachpuff3', 'mediumorchid2', 'turquoise3', 'wheat4’,
         'slategray2','slategray2','slategray2’)

# compute a small random position offset for each sample (jitter)
x_jitter = jitter(variantsPca$eigenvect[,1], .1)
y_jitter = jitter(variantsPca$eigenvect[,2], .1)

# plot the PCA
plot(x_jitter, y_jitter, col=jColors, xlab = "PC1", ylab = "PC2", pch=16)

# plot the labels of each sample
text(x_jitter, y_jitter, labels=variantsPca$sample.id,
        pos=runif(10, min=1, max=4))
```

What should you do?

1. Compute and plot the PCA
2. Look at the plot to understand the relationships between samples
3. Use the VCF to analyze which variants are in common / different between interesting sample pairs from the PCA

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	A549-MITO-NXT	A549-ROI	Cal51-MITO-NXT	Cal51-ROI	HCT116-MITO-NXT	HCT116-ROI	Nalm6-MITO-NXT	Nalm6-ROI	RPE-MITO-NXT	RPE-ROI
chrM	152	.	T	C	.	PASS	…	…	./.:.:.:.:.:.:.	./.:.:.:.:.:.:.	./.:.:.:.:.:.:.	./.:.:.:.:.:.:.	./.:.:.:.:.:.:.	./.:.:.:.:.:.:.	./.:.:.:.:.:.:.	./.:.:.:.:.:.:.	0/1:0,843:0.998:843:0,395:0,423:0,0,322,521	0/1:0,338:0.997:338:0,0:0,307:0,0,66,272

Questions ?