Conceptual overview: Principal Components Analysis (PCA)

Julin Maloof

Updated April 22, 2020

Principal Components Analysis

motivation

- Often a data set consists of many different variables.
- Principal Components Analysis (PCA) provides a way to focus on the most important aspects of the data.
- Just as the name says, PCA determines the Principal Components of the data set.

Principal components in genomics

One major use of PCA in genomics is to simplify complex SNP data sets.

Consider a simple data set of two markers, M1 (A/G) and M2 (C/T). We can make a graphical representation of these markers by assigning numeric values to each genotype at each marker.

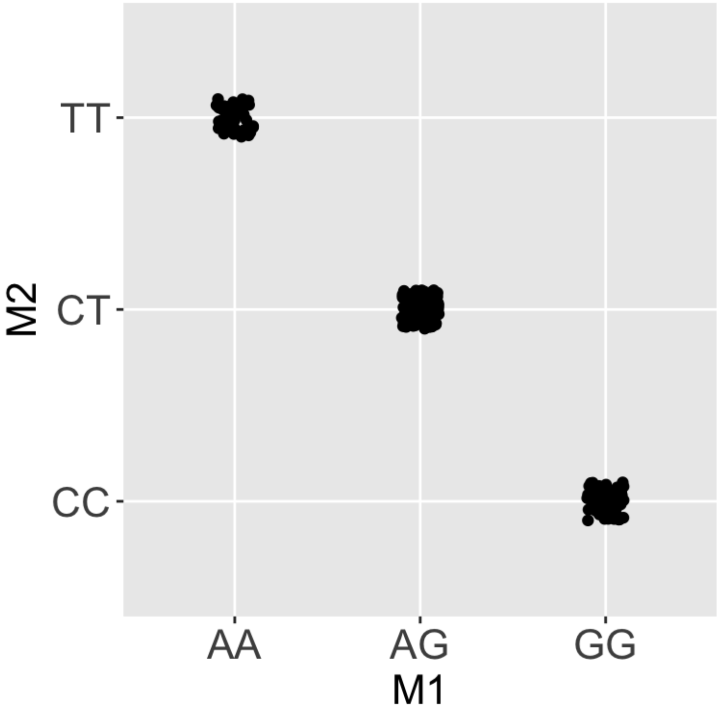
M1	M2
AA: 0	CC: 0
AG: 1	CT: 1
GG: 2	TT: 2

Principal components in genomics

We can plot each individual's genotypes on a 2D scatter plot:

M1	M2
AA: 0	CC: 0
AG: 1	CT: 1
GG: 2	TT: 2

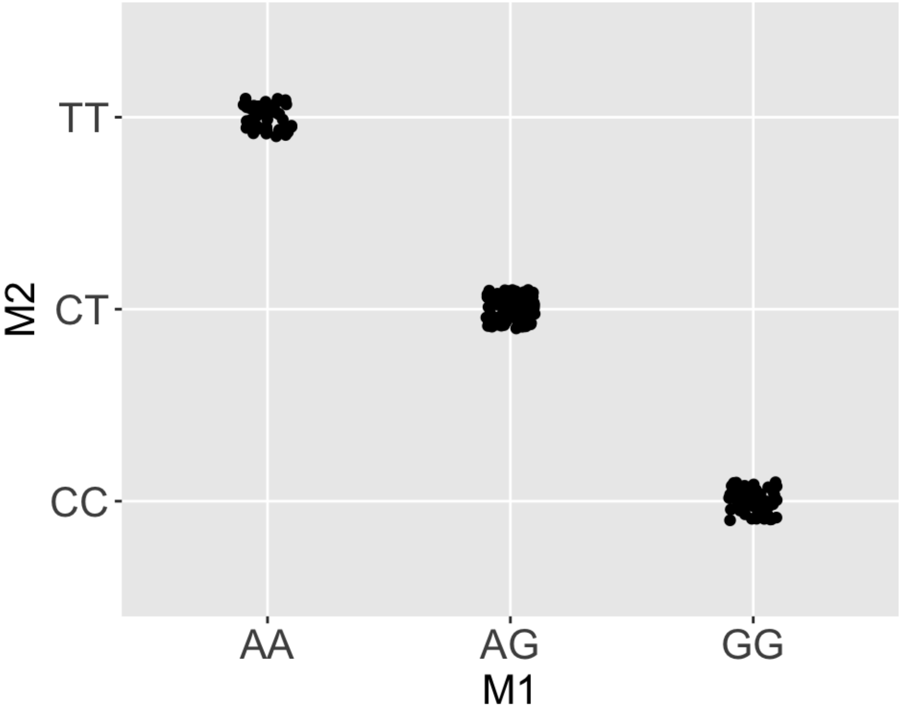
note: points are "jittered" as a visual aid.



Principal components in genomics

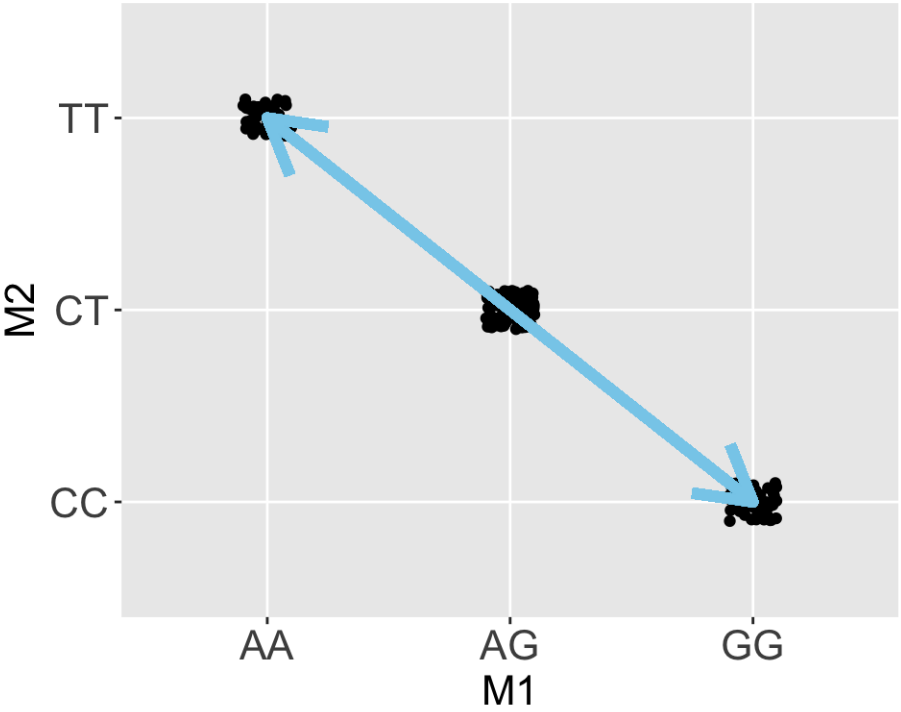
PCA identifies the vector through the data that contains the largest proportion of variance (i.e. the largest spread of data).

Where would you draw such a line here?



Principal components in genomics

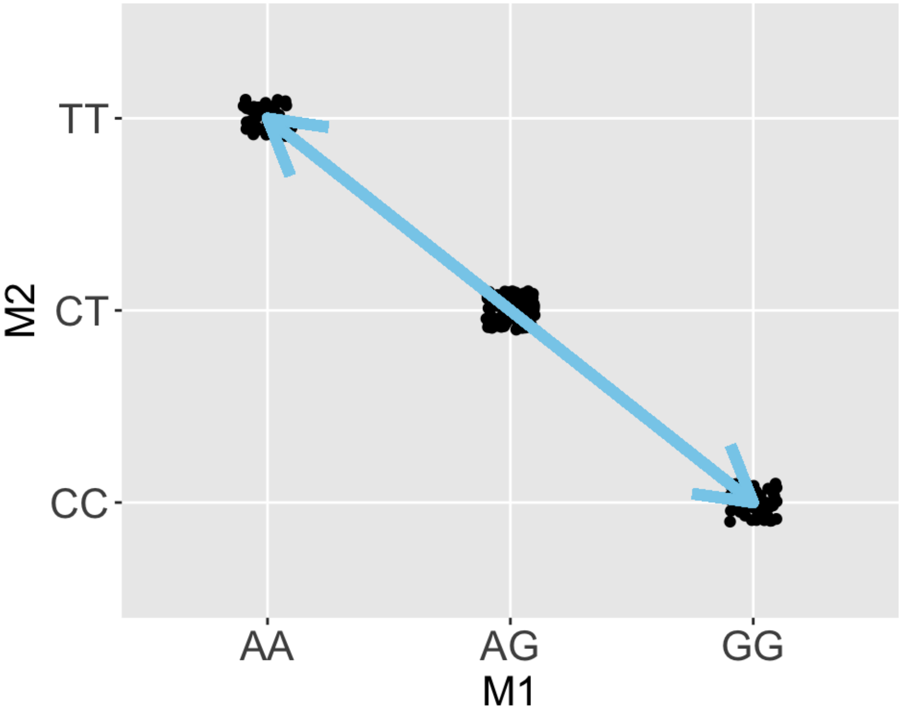
This vector represents the first principal component (PC1) and the contains the largest variance in the data:



Principal components in genomics

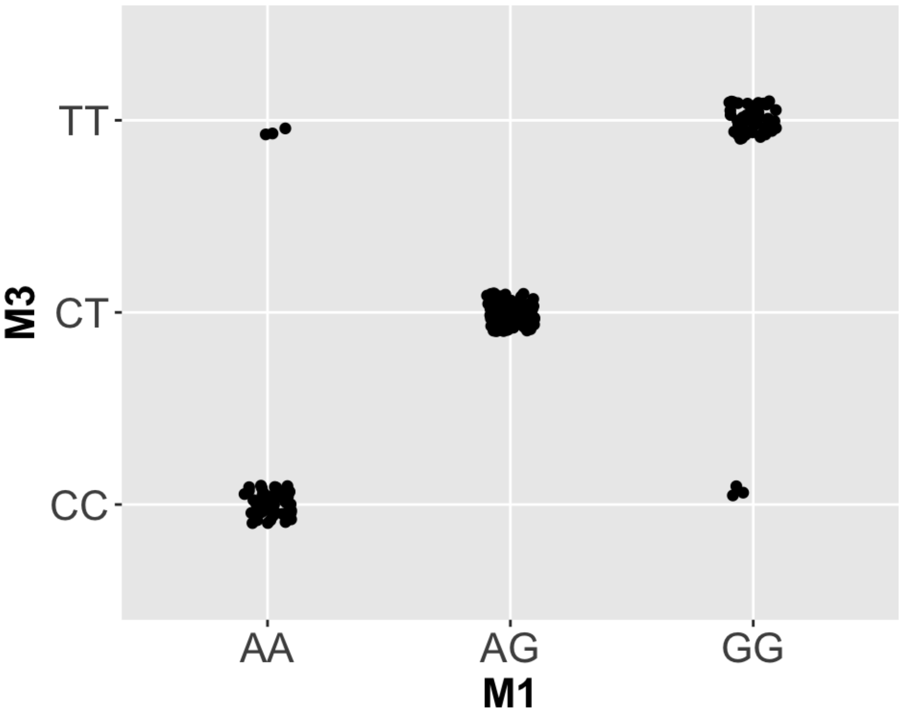
In this data set the second principal component contains no information.

Thus principal components has simplified a 2D data set to a single dimension.



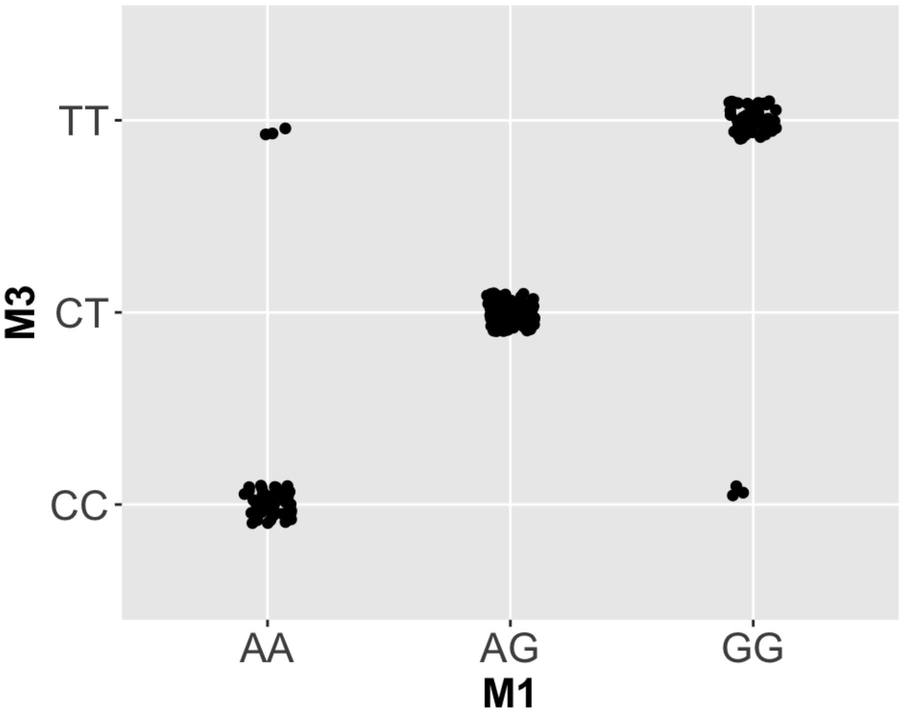
Principal components in genomics

Consider a new marker, M3:



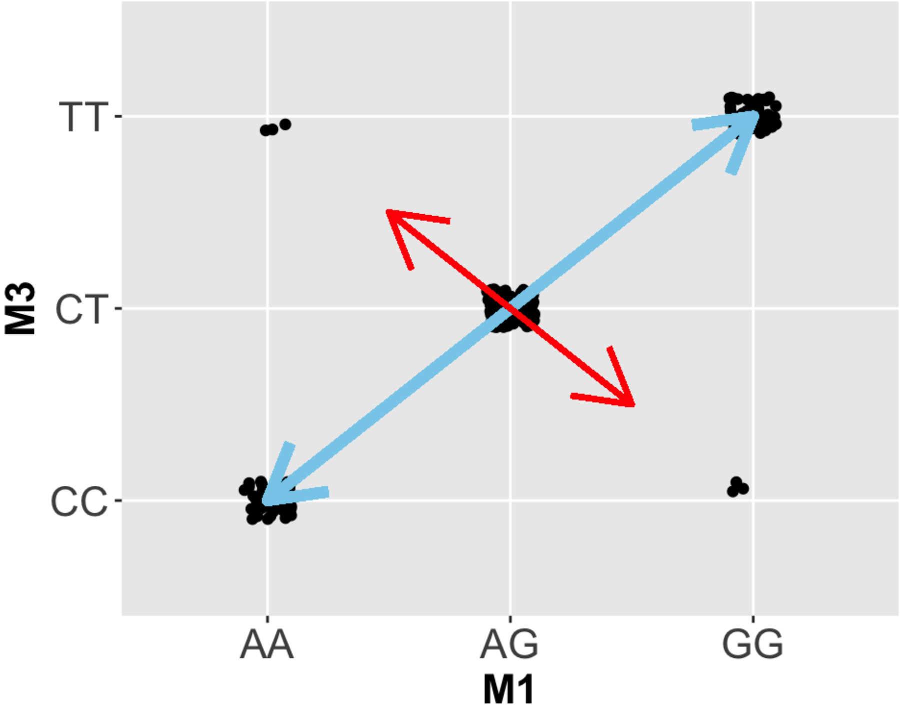
Principal components in genomics

Where are the first and second principal components here?



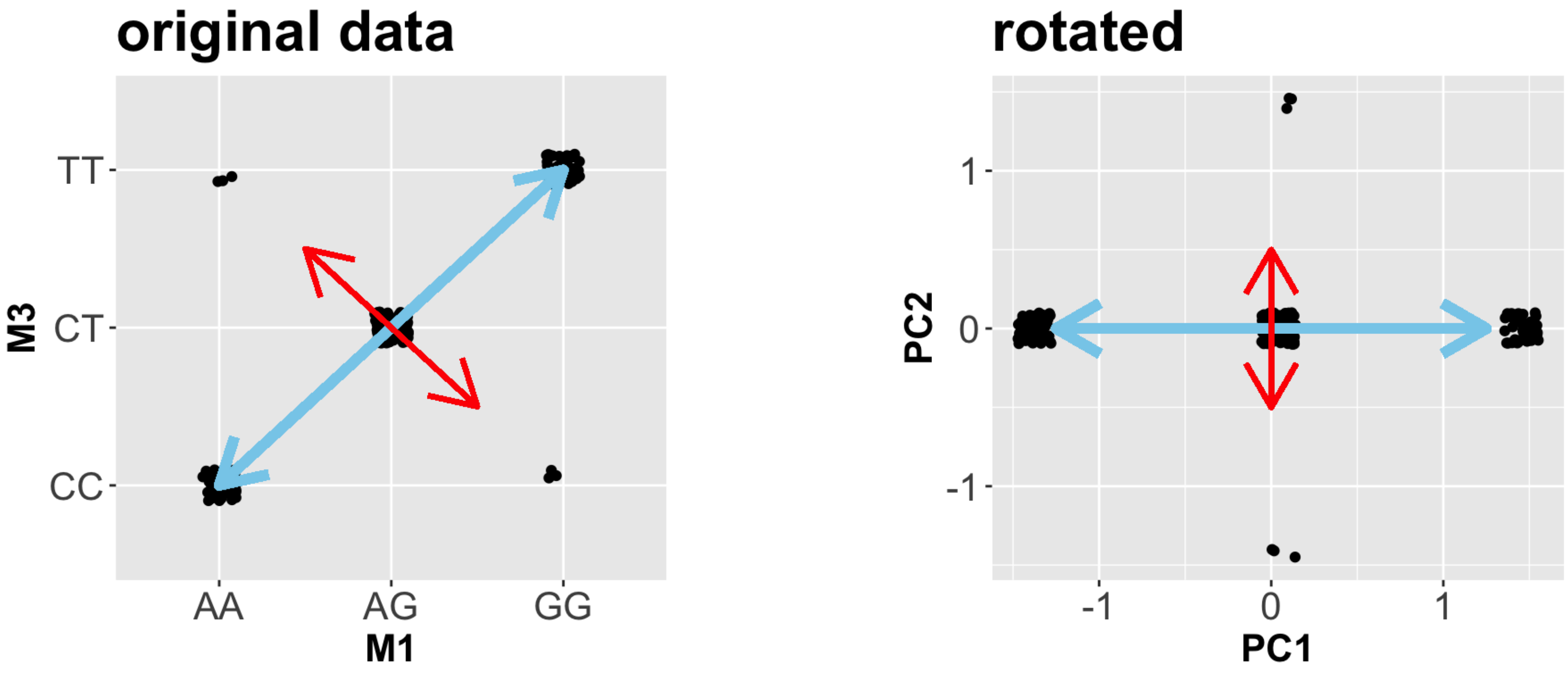
Principal components in genomics

Where are the first and second principal components here?



Principal components in genomics

We can rotate the data to align the plot with the principal components


original data

rotated

Now we have a single axis that represents the majority of the variation in the data, and a second axis that accounts for the remainder.

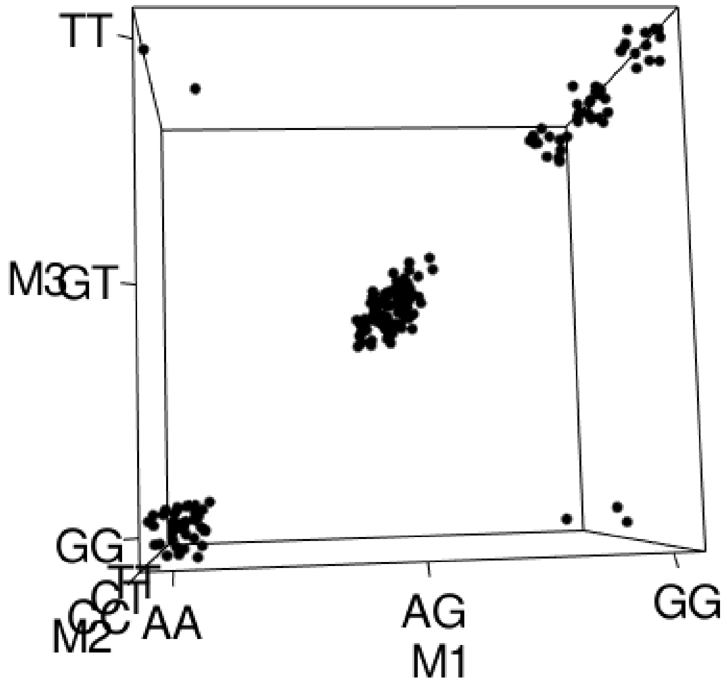
Three SNPs–First View

What if there are 3 SNPs?

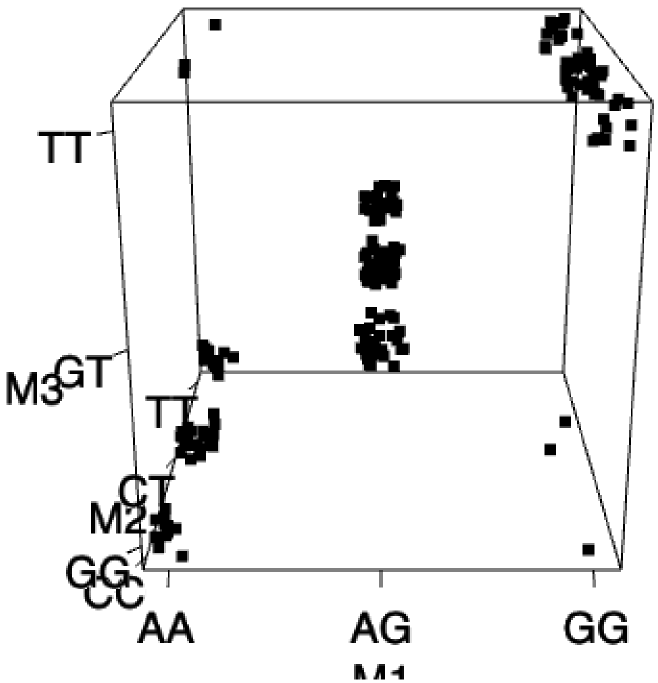
Now we have 3 dimensions

In this view it appears that most of the variance in along a single vector.

first view



Three SNPs–live 3D view



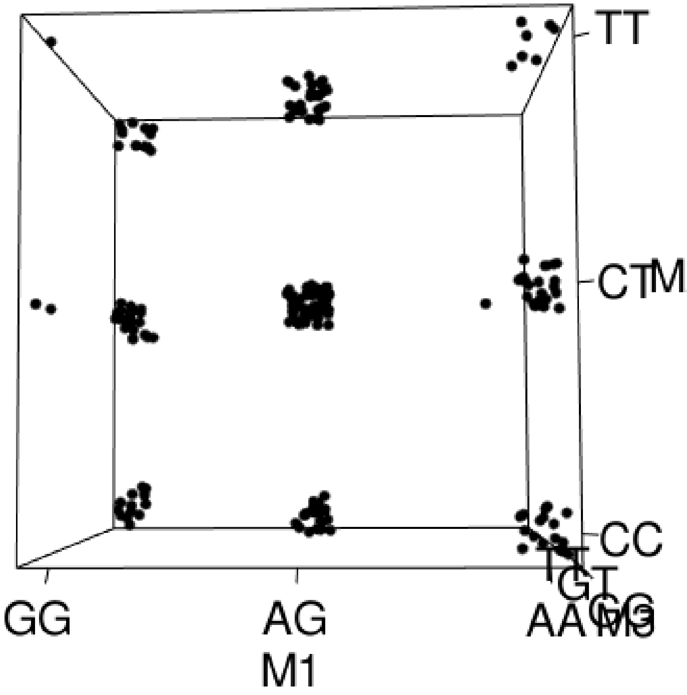
demo live rotation of data cube

Three SNPs–Second View

Changing rotation alters our interpretation of the data.

Now we see that we could draw 2 principal components that each would capture a fair bit of variance

second view

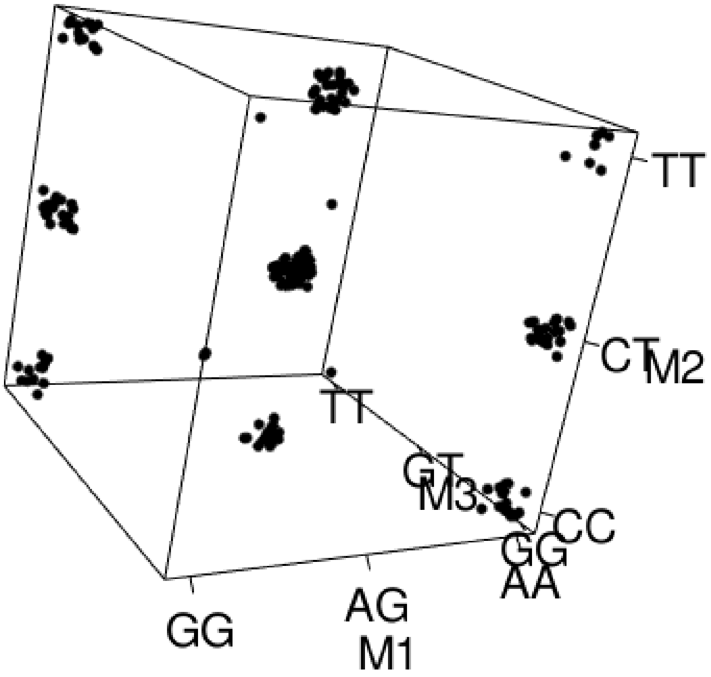


Three SNPs–Third View

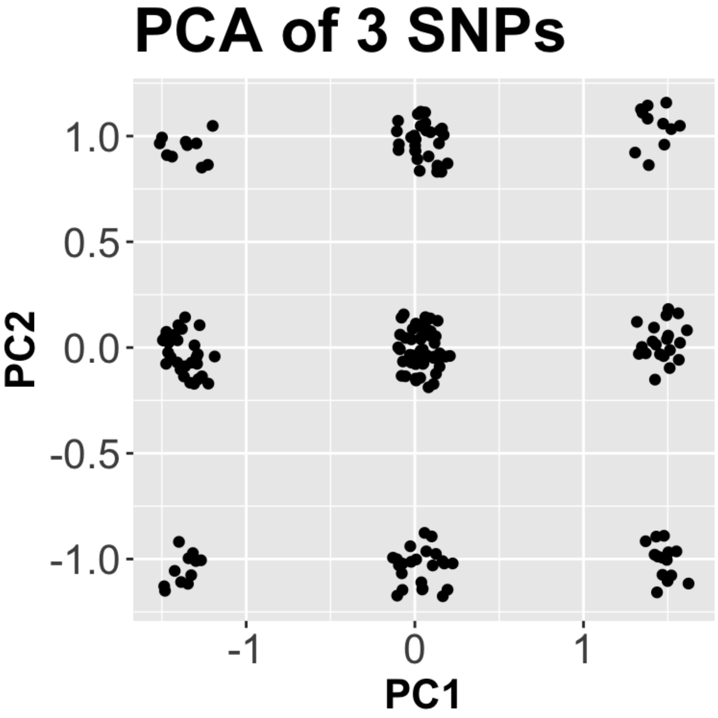
Changing rotation alters our interpretation of the data.

This rotation shows a third axis of variation.

third view



PCA analysis of 3 SNPs

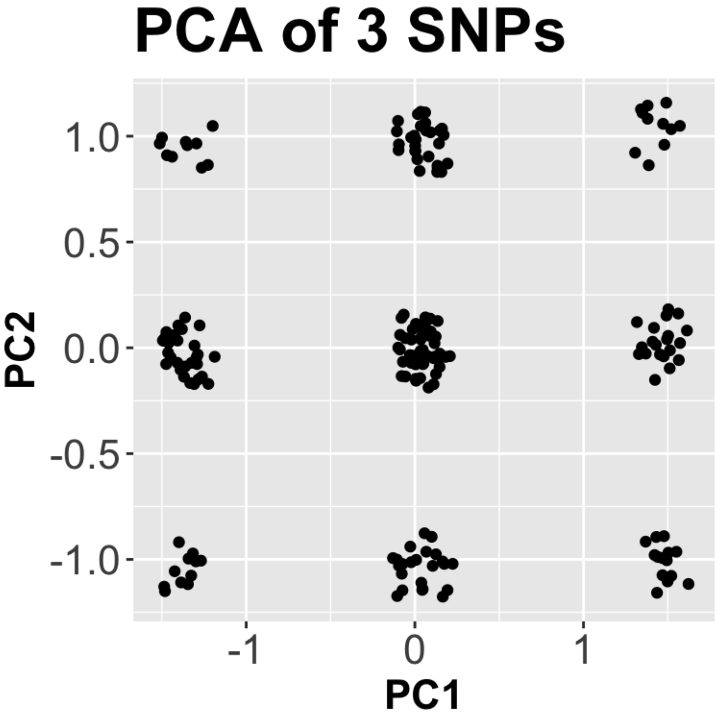


	PC1	PC2	PC3
M1	-0.71	0.02	0.71
M2	0.01	-1.00	0.04
M3	-0.71	-0.04	-0.71

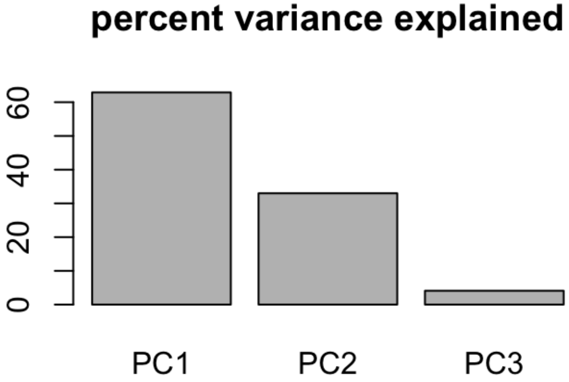
- PC1 captures co-variation at M1 and M3
- PC2 captures variation at M2
- PC3 captures opposite variation at M1 and M3

What do these PCs represent?

PCA analysis of 3 SNPs



How much variation is explained by each PC?



PC1 and PC2 capture almost all of the variance. We have converted our 3D data set into a 2D data set

PCA: many dimensions

- As you have seen in these examples, each SNP column can be considered a dimension of data.
- In the Li et al. paper there are 650,000 SNPs = 650,000 dimensions!
- By applying PCA the data is partioned such that the most informative aspects are represnted in the first PCs

PCA Summary

- Genomics data is typically highly dimensional.
- There is often redundancy in the data.
- PCA allows rotation and rescaling of the data so that we can focus on a smaller set of variables that contain the majority of the information.
- PCA enables 2D visualization of multi-dimensional datasets (for example by plotting the first and second PCs against one another).

PCA vs MDS

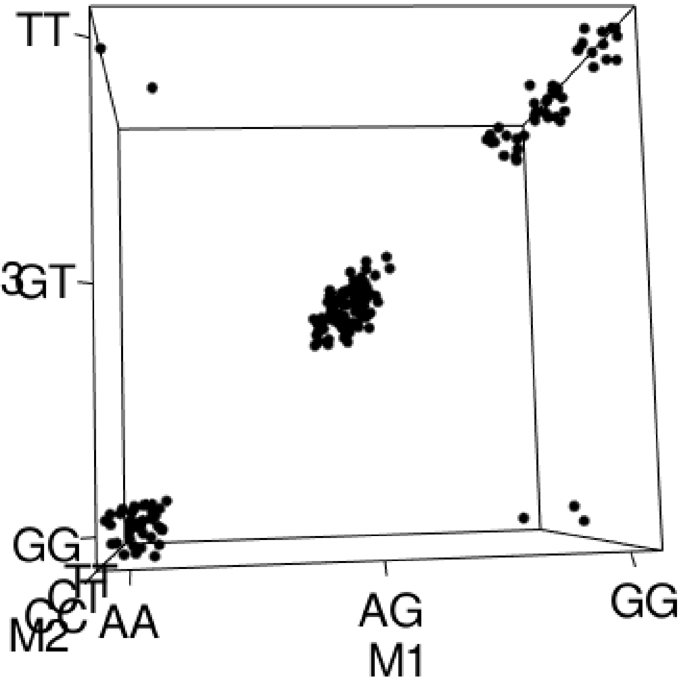
A related technique is multi-dimensional scaling (MDS).

- In PCA the final number of components is the same as the starting number of dimensions, but the information has been shifted towards a fewer number of dimensions.
- In MDS the data is rescaled and rotated to project it into a fixed number of dimensions (typically 2).

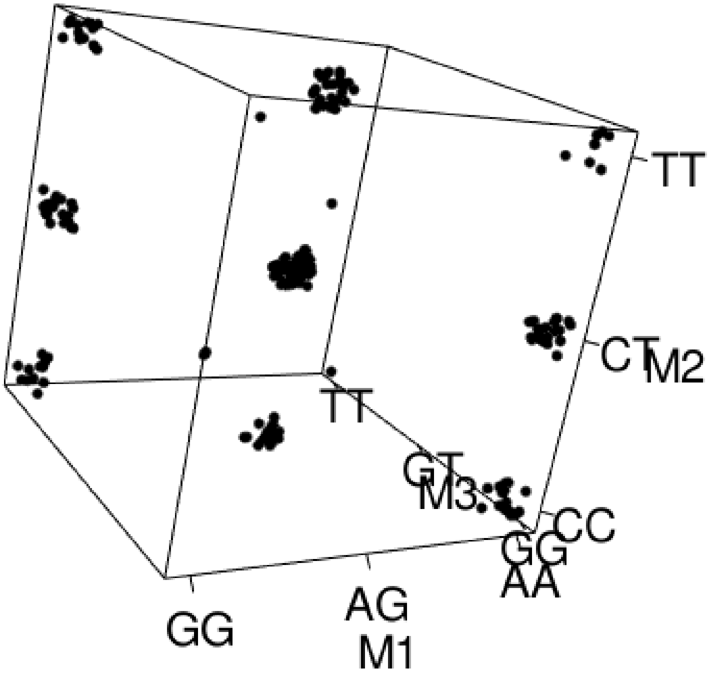
MDS

Determines the optimal projection to display the data in 2D

poor rotation



good rotation


TT
CT
M2
TT
GT
M3
CC
GG
AA
GG
AG
M1