Module 19 Multivariate Analysis for Genetic data
Session 02: Biplots

Jan Graffelman

jan.graffelman@upc.edu

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

[2]Department of Biostatistics
University of Washington
Seattle, WA, USA

28th Summer Institute in Statistical Genetics (SISG 2023)

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Introduction

- Definition of a Biplot
- Biplot construction
- Interpretation rules
- Examples

What is a biplot?

A biplot is a powerful tool for the graphical exploration of multivariate data (e.g. pattern and outlier detection).

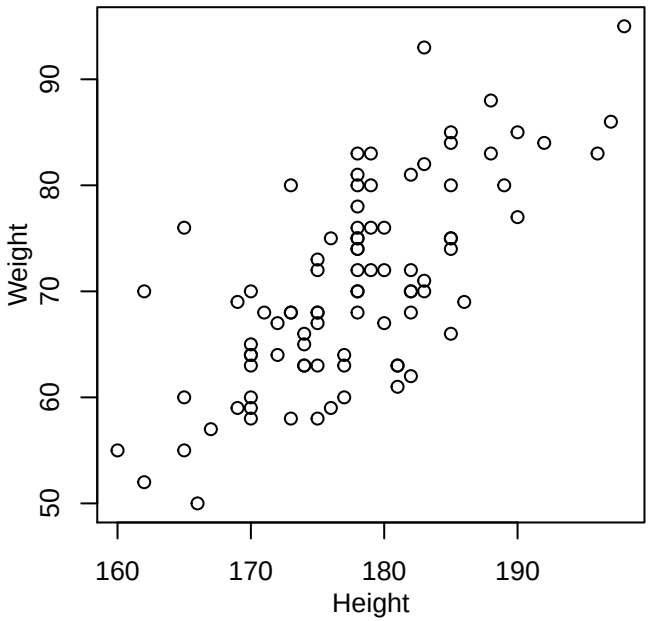
A biplot is a multivariate generalization of the scatter plot.

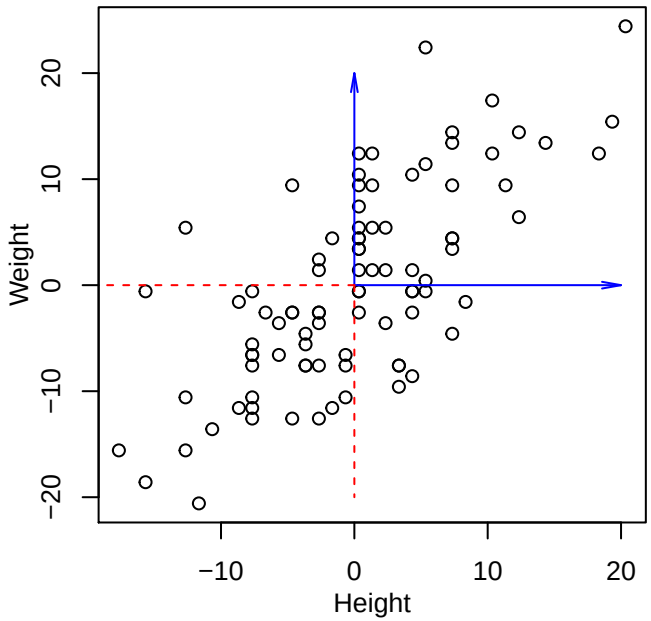
A biplot differs from a scatterplot in some ways:

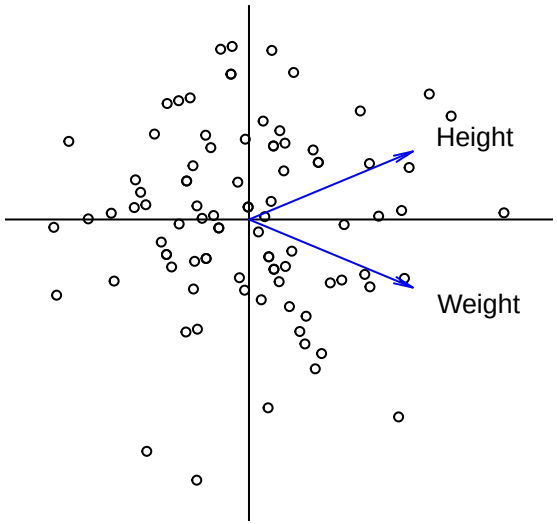
- It has typically more than 2 axes.
- The axes are generally not perpendicular.
- The representation of the data matrix is approximate, not exact.

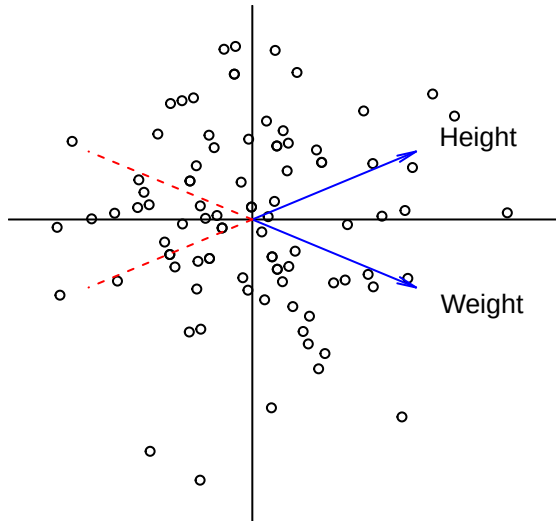
A biplot is a joint display of the rows and the columns a matrix that is optimal in a least squares sense.

(Bi)plots of student height and weight









Making a biplot

In order to make a biplot, the matrix to be represented needs to be factored

$$\mathbf{X}_{n\times p} = \mathbf{F}_{n\times r}\mathbf{G}_{r\times p}{}' \tag{1}$$

into a matrix of row markers ($\mathbf{F}$) and a matrix of column markers ($\mathbf{G}$).

Note that this factorization also exists in an ordinary scatter plot:

$$\mathbf{X}_{n\times 2} = \mathbf{X}_{n\times 2}\mathbf{I}_{2\times 2}$$

Often row markers are represented by dots and column markers by arrows, but not necessarily so.

Biplot uniqueness

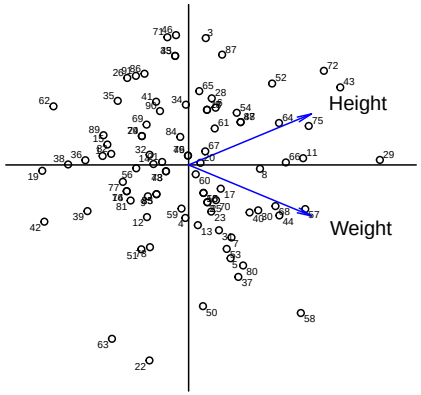
The factorization

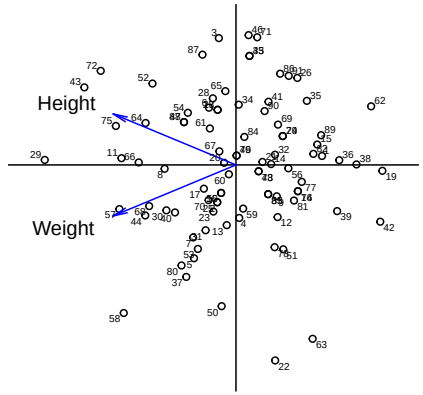
$$\mathbf{X}_{n\times p} = \mathbf{F}_{n\times r}\mathbf{G}_{r\times p}{}'$$

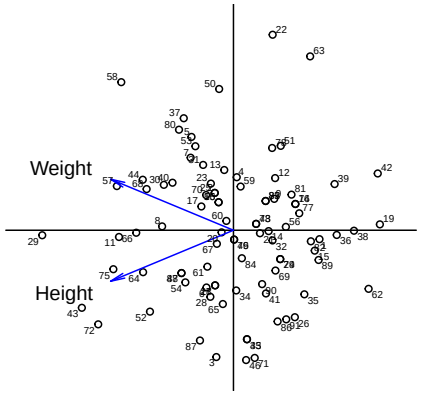
is not unique, because

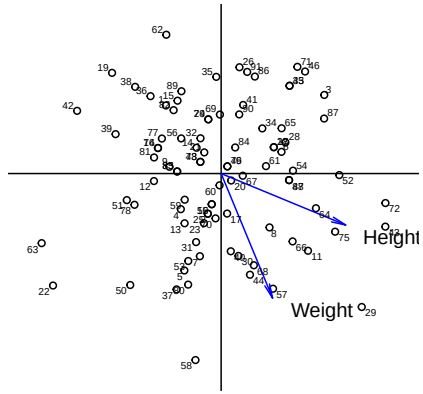
$$\mathbf{X}_{n\times p} = \mathbf{F}_{n\times r}\mathbf{T}\mathbf{T}^{-1}\mathbf{G}'_{r\times p} = \tilde{\mathbf{F}}_{n\times r}\tilde{\mathbf{G}}'_{r\times p}$$

where $\mathbf{T}$ is any non-singular linear transformation.









Biplots and SVD

The desired factorization can be obtained by the SVD of the matrix to be represented:

$$\mathbf{X} = \mathbf{UDV}' = \mathbf{FG}'$$

- The SVD guarantees that we obtain an approximation to $\mathbf{X}$ of given rank that is optimal in the least square sense.
- Depending on the particular multivariate method, $\mathbf{X}$ may be a matrix of quantitative variables, a correlation matrix, a contingency table, a matrix of regression coefficients, ...
- Different scalings are possible depending on how the singular values are dealt with. Different scalings have different geometrical properties.

Biplots and SVD

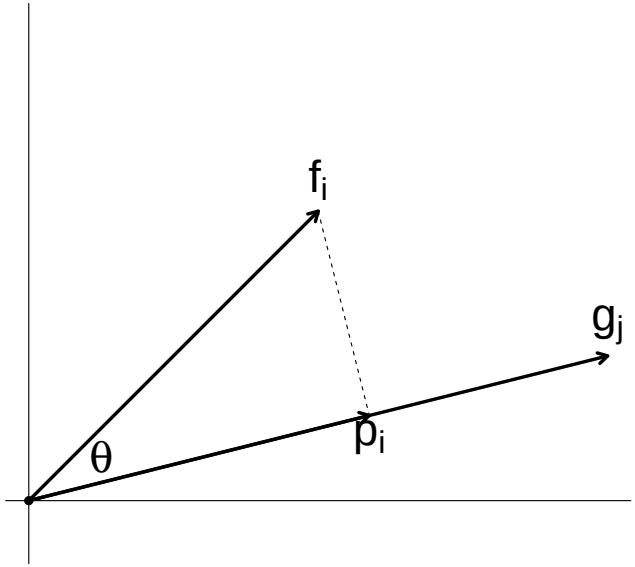
- The row and column markers can be defined in several ways, depending on what we do with the singular values.
- Some common definitions are
  - $\mathbf{F} = \mathbf{UD}$ and $\mathbf{G} = \mathbf{V}$
  - $\mathbf{F} = \mathbf{U}$ and $\mathbf{G} = \mathbf{VD}$
  - $\mathbf{F} = \mathbf{UD}^{1/2}$ and $\mathbf{G} = \mathbf{VD}^{1/2}$
- A general formulation for this indeterminacy is

$$\mathbf{F} = \mathbf{UD}^{\alpha} \quad \text{and} \quad \mathbf{G} = \mathbf{VD}^{1-\alpha} \quad \text{with } 0 \leq \alpha \leq 1.$$

- Due to the free choice of α and $\mathbf{T}$, for a given $\mathbf{X}$ infinitely many biplots can be constructed.

Biplots and scalar product (inner product)

In a biplot data values are approximated by the scalar product between two vectors.



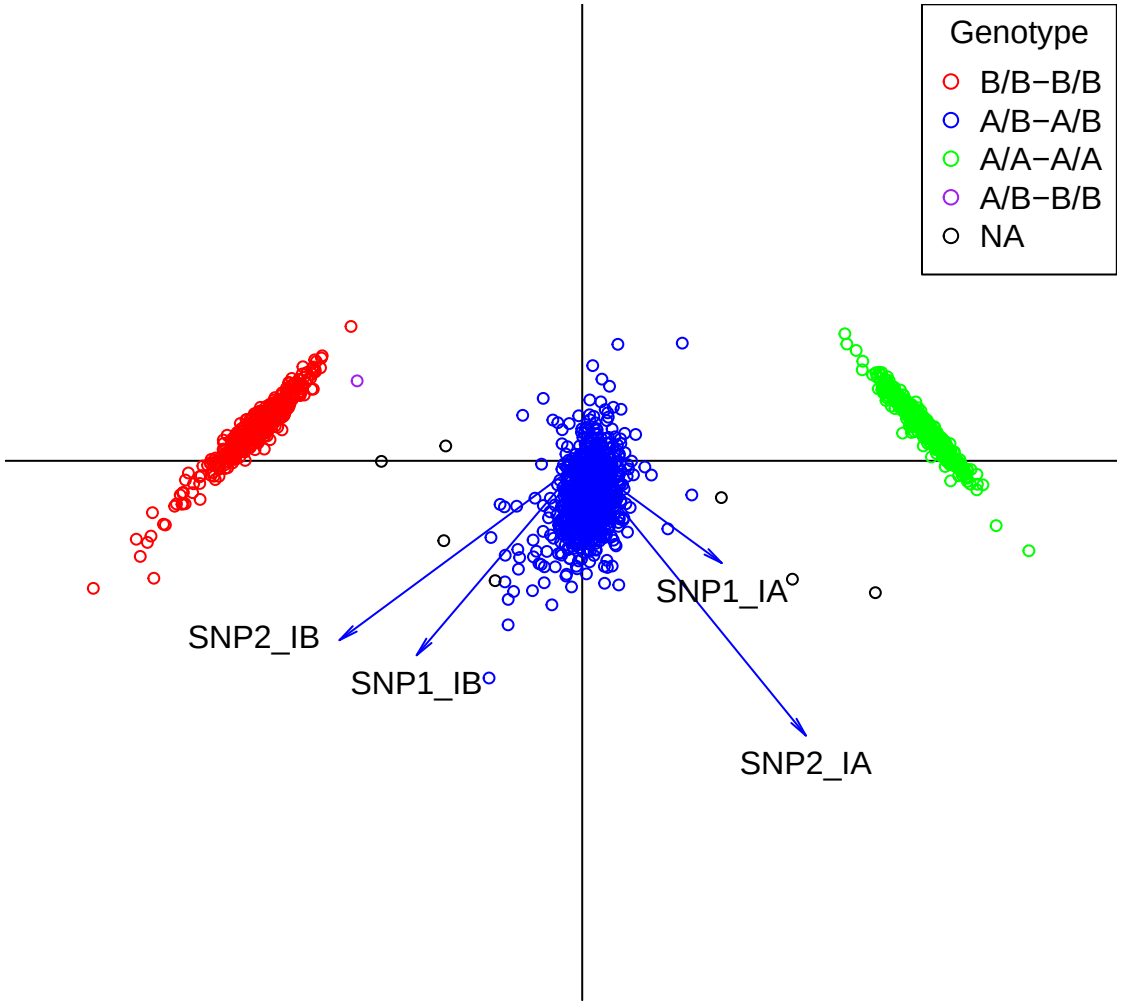
$$\cos\theta = \frac{\| \mathbf{p} \|}{\| \mathbf{f}_i \|} = \frac{\mathbf{f}_i' \mathbf{g}_j}{\| \mathbf{f}_i \| \| \mathbf{g}_j \|} \qquad \| \mathbf{p} \| = \frac{\mathbf{f}_i' \mathbf{g}_j}{\| \mathbf{g}_j \|} \qquad x_{ij} \approx \mathbf{f}_i' \mathbf{g}_j = \| \mathbf{p}_i \| \cdot \| \mathbf{g}_j \|$$

General interpretation rules

- It's very important to control the aspect ratio of the biplot.
- The precise interpretation depends on the matrix that is represented and on the chosen scaling.
- In general, the origin represents a mean vector.
- Orthogonal projection of points (or vectors) onto vectors approximate the entries of the matrix of interest.
- Distances between points reflect some chosen measure of similarity of the corresponding observations.

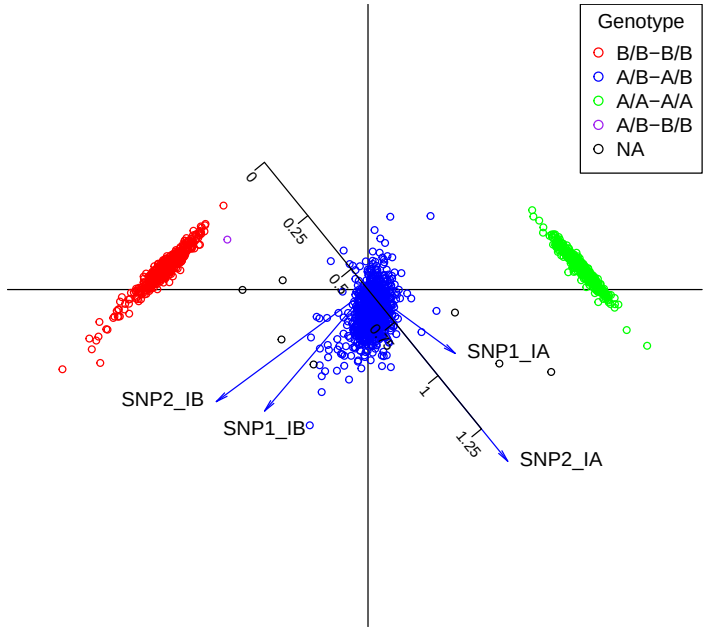
Example of a biplot

Biplot of the centered allele intensity data.



Biplot calibration

- Biplot calibration refers to drawing scales with tickmarks and labels along the biplot vectors.
- Calibration facilitates biplot interpretation and data recovery.
- Calibrated scales are usually omitted to avoid overcrowded plots.
- Calibration can be used to highlight a variable of particular interest.



Graffelman, J. and van Eeuwijk, F. (2005) Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. *Biometrical Journal* **47**(6): 863-879. doi: 10.1002/bimj.200510177.

Supplementary information

- Once a biplot has been made, additional observations or variables can be added by projection.
- Such observations are called supplementary points or variables, or inactive points or variables,
- Coordinates for supplementary points/variables can be obtained by regression onto the biplot axes.
- This principle is widely used in genetics for ancestry research: project individuals of unknown ancestry onto a biplot of genetic variants made for a set of known populations.

Supplementary information and regression

From

$$\mathbf{X} = \mathbf{UDV}' = \mathbf{FG}'$$

it follows that

$$\mathbf{G} = \mathbf{X}'\mathbf{F}\left(\mathbf{F}'\mathbf{F}\right)^{-1}$$

and

$$\mathbf{F} = \mathbf{XG}\left(\mathbf{G}'\mathbf{G}\right)^{-1}$$

- Golden rule: row and column markers are regression coefficients (always!).
- Substitute supplementary data for $\mathbf{X}$ to compute the markers.

Graffelman, J. and Aluja-Banet, T. (2003) Optimal representation of supplementary variables in biplots from principal component analysis and correspondence analysis. *Biometrical Journal* **45**(4): 491-509. doi: 10.1002/bimj.200390027

Biplots in multivariate analysis

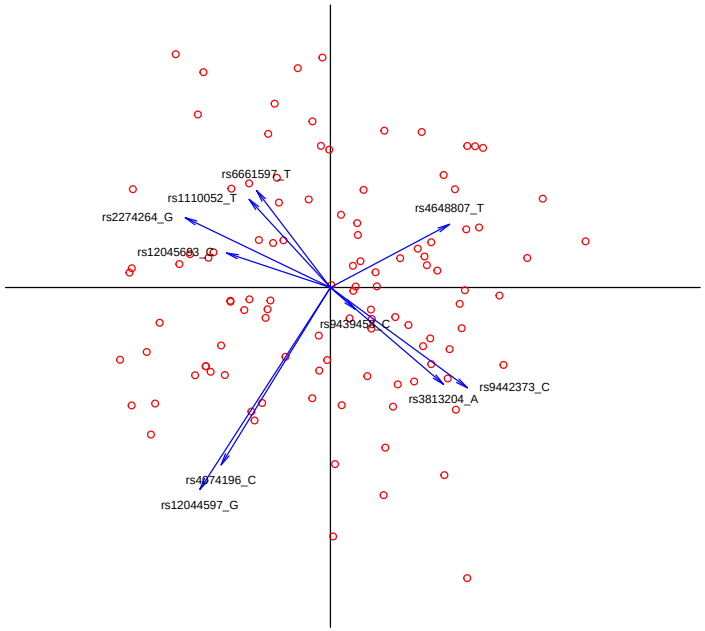
- Most classical multivariate methods are based on the singular value decomposition
- With all these methods biplots can be constructed
- Multivariate methods differ with respect to the matrix or matrices they approximate.

Biplots for some matrices and the related methods

Matrix		Method
$\mathbf{X}$	[raw data]	Principal component analysis (PCA)
		Scatter plot + regression
$\mathbf{X}_c$	[centered data]	Principal component analysis (PCA)
		Scatter plot + regression
$\mathbf{X}_s$	[standardized data]	Principal component analysis (PCA)
		Scatter plot + regression
$\mathbf{X}_{co}$	[compositional data]	Log-ratio principal component analysis (LR-PCA)
$\mathbf{S}$	[covariances]	Principal component analysis (PCA)
$\mathbf{R}$	[correlations]	Principal component analysis (PCA)
		Factor analysis (FA)
$\mathbf{N}$	[cross table; profiles]	Correspondence analysis (CA)
$\mathbf{M}$	[table of group means]	Linear discriminant analysis (LDA)
$\mathbf{R}_{xy}$	[correlations between sets]	Canonical correlation analysis (CCA)
$\mathbf{B}_{xy}$	[regression coefficients]	Redundancy analysis (RDA)
$\vdots$	$\vdots$	$\vdots$

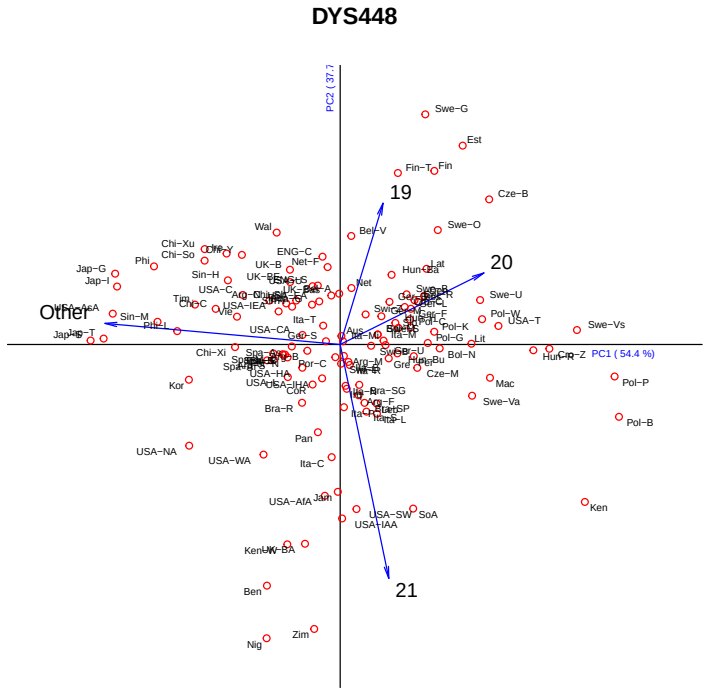
Biplot of **X** (PCA): 10 SNPs of the CHD sample of the 1,000G project

ID	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
NA17962	2	0	0	2	2	0	0	1	0	2
NA17965	1	0	1	2	1	2	2	0	0	1
NA17966	1	1	1	1	0	1	1	2	0	2
NA17967	0	2	1	1	2	1	0	0	0	1
NA17968	1	0	1	1	1	2	1	2	1	2
NA17969	2	2	0	1	1	2	1	1	1	2
NA17970	1	1	2	0	2	0	2	0	0	1
NA17972	1	2	2	2	2	0	1	0	2	0
NA17974	0	2	2	0	2	1	2	1	1	1
NA17975	2	0	0	0	2	1	2	2	2	0
NA17976	0	2	2	1	1	0	0	1	2	0
NA17977	1	0	1	2	1	2	2	1	1	1
NA17978	2	1	1	1	1	1	1	2	2	1
NA17979	1	0	2	1	0	2	0	0	1	0
NA17980	0	0	0	2	0	0	0	2	1	0
NA17981	1	0	0	0	1	2	2	1	1	1
NA17982	1	1	2	1	2	2	0	1	1	0
NA17983	1	0	1	0	2	0	0	2	1	0
NA17986	1	0	0	0	1	2	2	1	1	1
NA17987	0	1	2	0	2	1	2	0	0	1
NA17988	2	1	0	1	0	1	0	0	2	2
NA17989	1	1	1	1	2	1	2	1	1	1
NA17990	1	2	2	1	1	1	2	0	1	1
NA17993	0	2	1	1	1	2	2	0	0	0
NA17995	0	1	1	1	0	1	0	1	1	0
NA17996	1	1	1	1	1	0	1	1	1	1
NA17997	0	1	1	2	1	2	2	2	0	1
NA17998	1	0	0	2	2	1	1	2	0	1
NA17999	2	2	1	0	1	1	0	1	1	1
NA18101	1	1	2	0	1	2	0	2	0	1
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.



Biplot of $\mathbf{X}_{co}$ (LR-PCA) Allele frequencies of Y-STR DYS448 over 129 populations worldwide

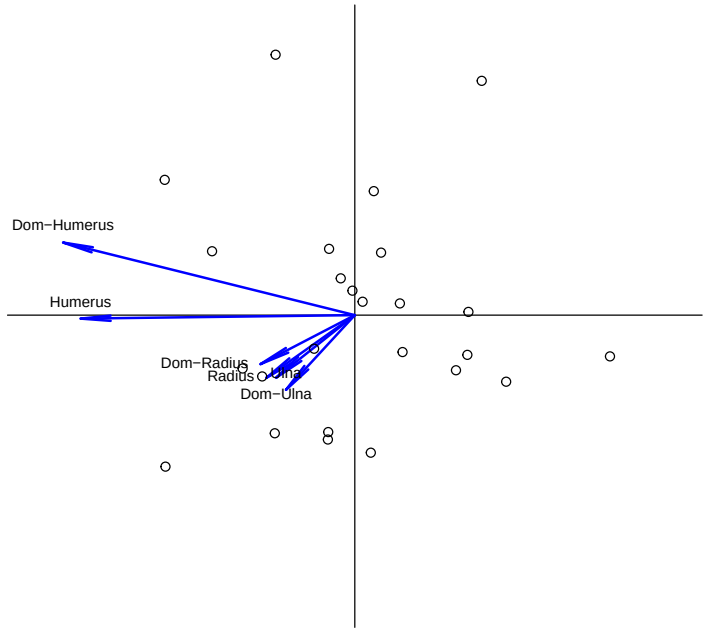
Sample	19	20	21	other
Arg-B	0.40	0.34	0.11	0.15
Arg-F	0.42	0.31	0.18	0.08
Arg-M	0.45	0.34	0.12	0.10
Arg-N	0.46	0.28	0.06	0.20
Arg-S	0.50	0.20	0.12	0.18
Aus	0.42	0.39	0.09	0.10
Bel-A	0.56	0.30	0.06	0.09
Bel-V	0.60	0.30	0.03	0.07
Ben	0.08	0.14	0.63	0.16
Bol-M	0.43	0.41	0.09	0.07
Bol-N	0.29	0.59	0.09	0.04
Bos	0.44	0.46	0.06	0.04
Bra-R	0.43	0.23	0.20	0.15
Bra-SG	0.57	0.20	0.16	0.07
Bra-SP	0.45	0.28	0.20	0.07
Chi-B	0.37	0.33	0.11	0.20
Chi-C	0.44	0.19	0.06	0.31
Chi-Sh	0.32	0.45	0.06	0.17
Chi-So	0.57	0.13	0.03	0.27
Chi-Xi	0.07	0.58	0.04	0.32
Chi-Xu	0.37	0.26	0.03	0.35
Chi-Y	0.31	0.39	0.03	0.28
CoR	0.45	0.25	0.16	0.13
Cro-C	0.43	0.46	0.06	0.05
Cro-Z	0.35	0.55	0.09	0.01
Cze-B	0.35	0.62	0.01	0.01
Cze-M	0.31	0.52	0.12	0.05
Den	0.41	0.50	0.05	0.04
ENG-C	0.57	0.30	0.04	0.10
ENG-S	0.58	0.26	0.05	0.11
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Wal	0.72	0.14	0.03	0.11
Zim	0.05	0.11	0.76	0.07



Biplot of **S** (PCA) Mineral content measurements in bones

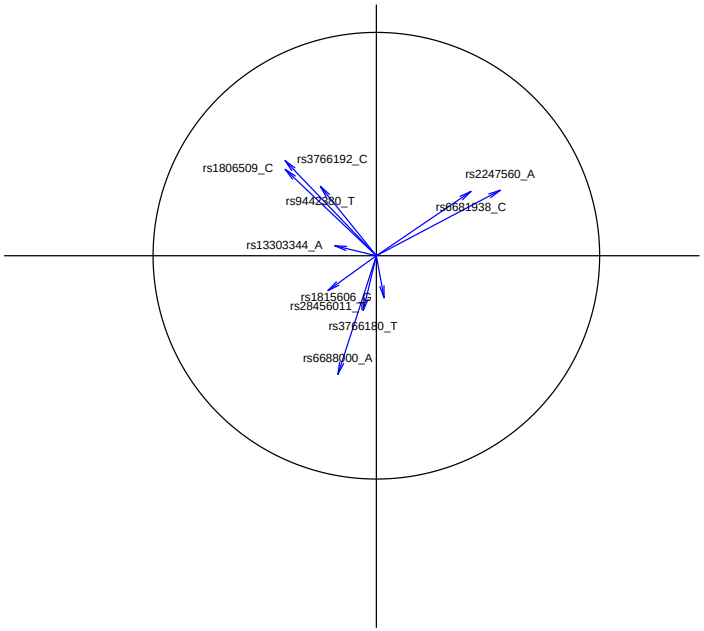
	Dom-Radius	Radius	Dom-Humerus	Humerus	Dom-Ulna	Ulna
1	1.103	1.052	2.139	2.238	0.873	0.872
2	0.842	0.859	1.873	1.741	0.590	0.744
3	0.925	0.873	1.887	1.809	0.767	0.713
4	0.857	0.744	1.739	1.547	0.706	0.674
5	0.795	0.809	1.734	1.715	0.549	0.654
6	0.787	0.779	1.509	1.474	0.782	0.571
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

	Dom-Rad	Radius	Dom-Hum	Humerus	Dom-Ulna	Ulna
Dom-Rad	0.013	0.010	0.022	0.020	0.009	0.008
Radius	0.010	0.011	0.019	0.021	0.009	0.009
Dom-Hum	0.022	0.019	0.080	0.067	0.017	0.013
Humerus	0.020	0.021	0.067	0.069	0.018	0.017
Dom-Ulna	0.009	0.009	0.017	0.018	0.012	0.008
Ulna	0.008	0.009	0.013	0.017	0.008	0.011



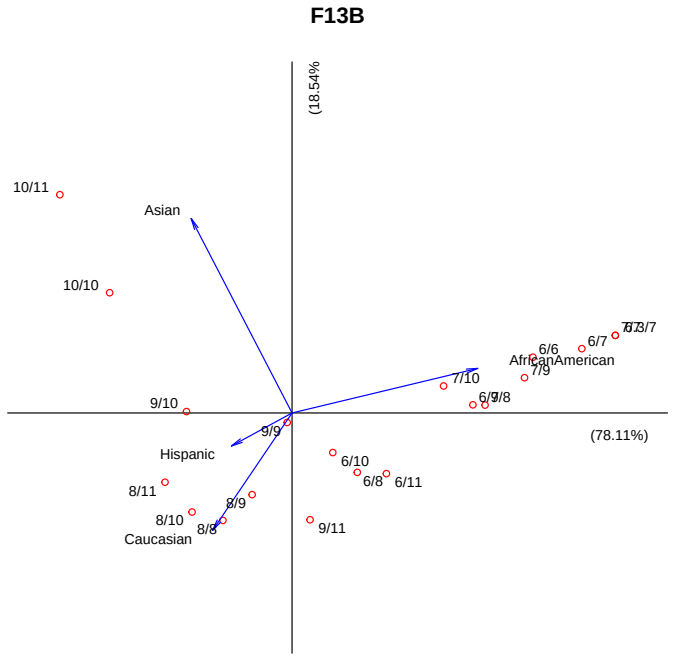
Biplot of **R** (PCA): 10 SNPs of the CHD sample of the 1,000G project

	509	192	380	606	180	011	938	000	560	344
rs1806509	1.00	0.35	0.14	-0.13	0.09	-0.00	-0.04	-0.06	-0.14	0.17
rs3766192	0.35	1.00	0.21	-0.12	0.03	-0.02	-0.09	-0.12	-0.19	-0.08
rs9442380	0.14	0.21	1.00	0.11	-0.16	0.08	-0.00	-0.01	-0.05	0.09
rs1815606	-0.13	-0.12	0.11	1.00	-0.10	-0.19	-0.20	-0.00	-0.05	-0.07
rs3766180	0.09	0.03	-0.16	-0.10	1.00	0.09	0.10	0.27	-0.02	0.02
rs28456011	-0.00	-0.02	0.08	-0.19	0.09	1.00	-0.06	0.10	-0.13	0.16
rs6681938	-0.04	-0.09	-0.00	-0.20	0.10	-0.06	1.00	-0.07	0.45	-0.03
rs6688000	-0.06	-0.12	-0.01	-0.00	0.27	0.10	-0.07	1.00	-0.24	-0.00
rs2247560	-0.14	-0.19	-0.05	-0.05	-0.02	-0.13	0.45	-0.24	1.00	-0.06
rs13303344	0.17	-0.08	0.09	-0.07	0.02	0.16	-0.03	-0.00	-0.06	1.00



Biplot of **N** (CA) NIST STR F13B genotypes by ethnicity

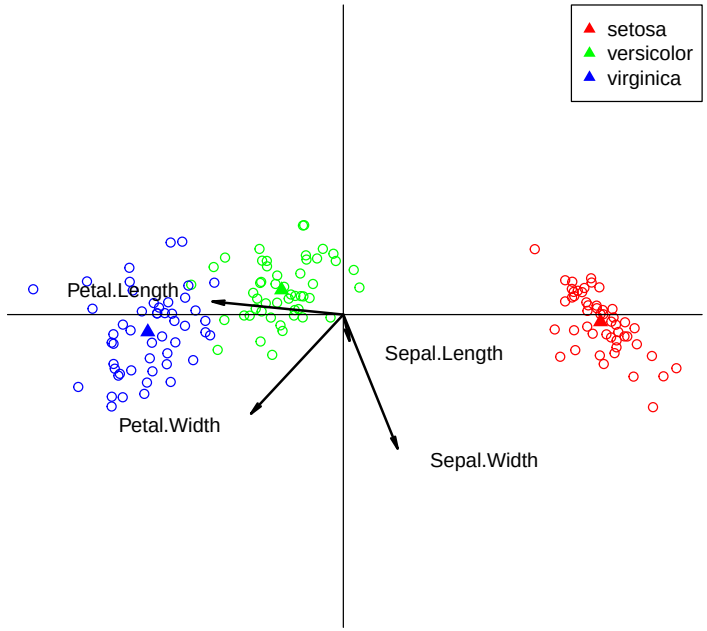
F13B	African American	Asian	Caucasian	Hispanic
10/10	5	54	58	48
10/11	0	1	1	0
6.3/7	1	0	0	0
6/10	31	1	22	27
6/11	1	0	1	0
6/6	43	1	4	4
6/7	50	0	1	3
6/8	23	0	22	8
6/9	57	0	15	11
7/10	16	1	5	3
7/7	7	0	0	0
7/8	10	0	3	1
7/9	24	0	3	3
8/10	10	5	64	32
8/11	0	0	0	1
8/8	6	1	24	9
8/9	19	2	41	28
9/10	21	26	73	49
9/11	1	0	2	0
9/9	17	5	22	9



Biplot of **M** (LDA) Fisher's classical Iris data

	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
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Group	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.01	3.43	1.46	0.25
versicolor	5.94	2.77	4.26	1.33
virginica	6.59	2.97	5.55	2.03



References

- Gabriel, K.R. (1971) The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3) pp. 453-467.
- Gower, J.C. & Hand, D.J. (1996) *Biplots*, Chapman & Hall, London.
- Greenacre, M.J. (2010) *Biplots in Practice*, Rubes Editorial.
- Gower, J., Lubbe, S. & le Roux, N. (2011) *Understanding Biplots*, Wiley, Chicester, UK.