Exploring Genome Data Using Correspondence Analysis

Fredj Tekaia
e-mail: tekaia@pasteur.fr

Bioinformatics and Genome Analyses
Institut Pasteur Tunis, Tunisia. September 18 – December 15, 2017.

Context

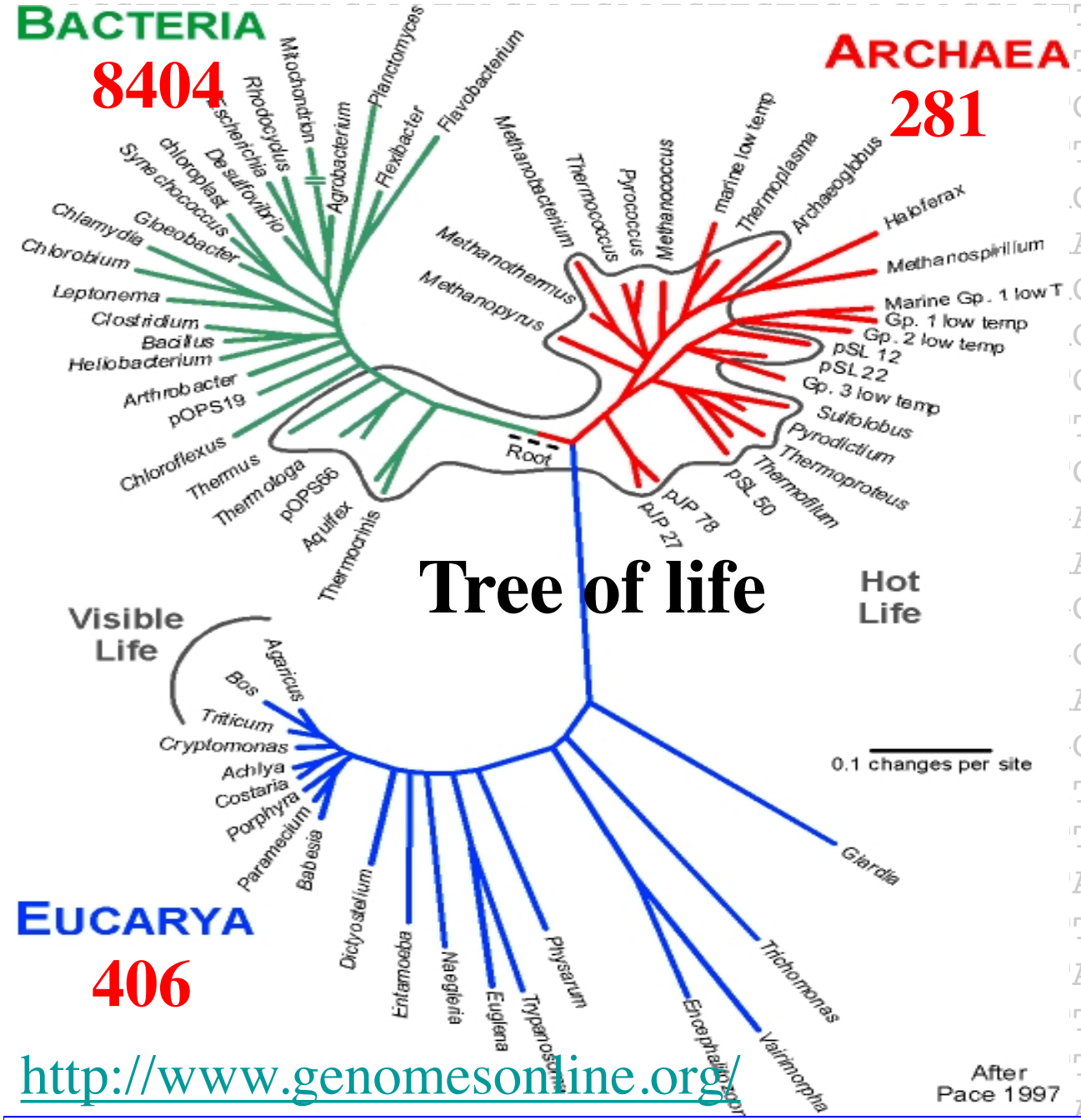
- Huge amounts of genome data are available
- Data are of multidimensional structure
- Big omics data

Need to:

- Synthesize the available data/Discover patterns
- Extract the general evolutionary (variation) trends from large data sets
- Discover universal/general trends that underlie these data

==> Necessity of multidimensional data analysis methods

22/11/2017



http://www.genomesonline.org/

Viruses: • **Completed: 3504** • **Permanent draft: 5008**

Complete sequenced genomes: 9091

- 8404 Bacteria
- 281 Archaea
- 406 Eukaryotes

Incomplete genomes: 15525

- 10740 Bacteria
- 292 Archaea
- 4493 Eukaryotes

Permanent Draft genomes: 75661

- 70632 Bacteria
- 781 Archaea
- 4248 Eukaryotes

Transciptomes: 75/15171

- 51/1026 Bacteria
- 0/160 Archaea
- 24/13985 Eukaryota

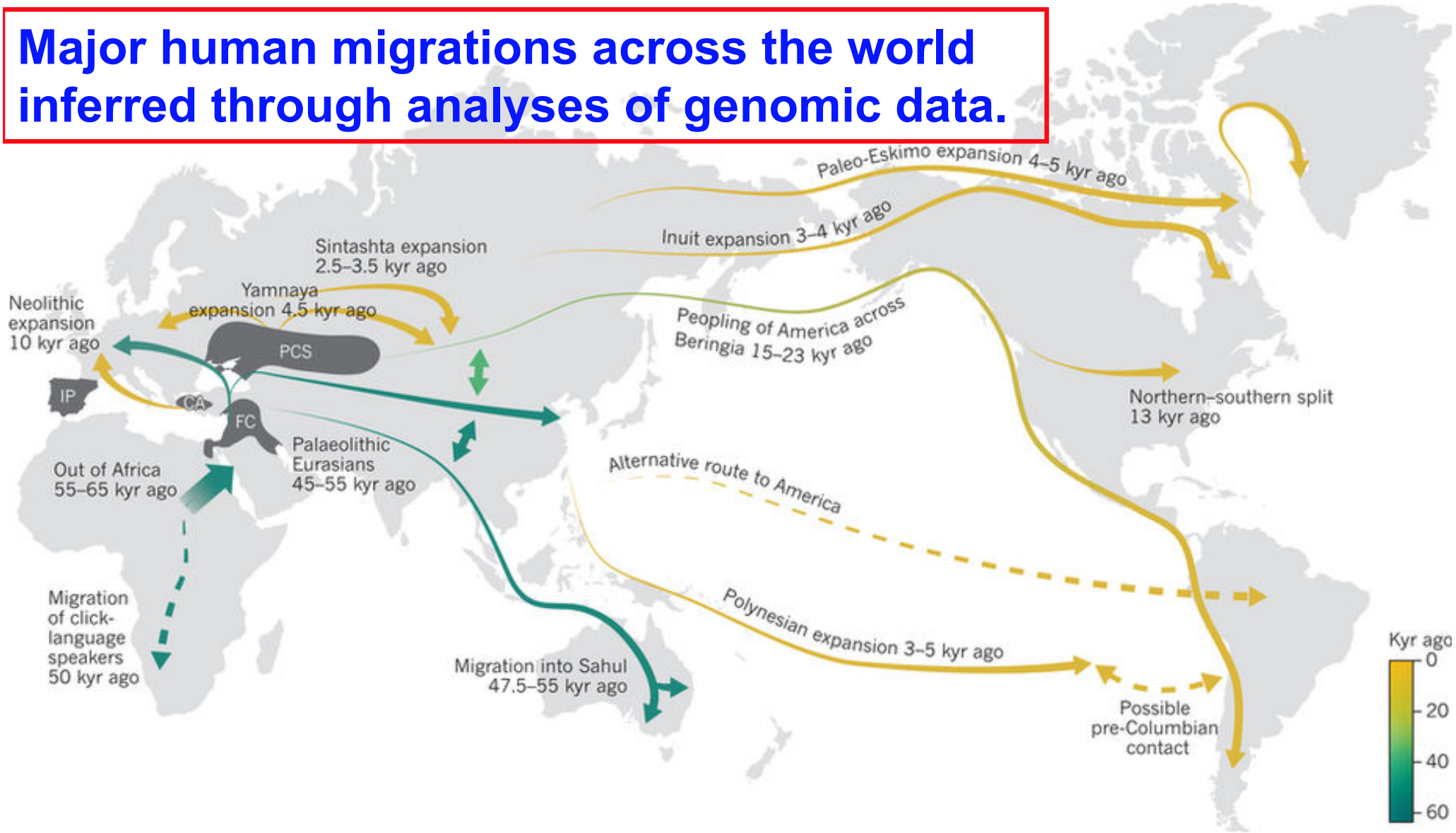
1000 Genomes Project

Has developed an open resource that now includes the genomic data for more than 2,500 individuals from 26 global populations across 5 continental regions, all analysed using whole-genome sequencing, deep exome sequencing and dense microarray genotyping approaches.

Key features were described of all 88 million genetic variants identified, while an extended analysis from the Structural Variation Analysis Group provides a more detailed account of larger and more complex variants.

Tracing the peopling of the world through genomics

Major human migrations across the world inferred through analyses of genomic data.



CA: Central Anatolia; FC: Fertile Crescent; IP: Iberian Peninsula; PCS: Pontic–Caspian steppe.

Nielson R et al. (2017). Tracing the peopling of the world through genomics. *Nature* 541, 302–310.

With the rapid expansion of genome sequences, multivariate analysis methods will be increasingly used for the discovery of important evolutionary trends intrinsically associated with the multidimensional structure of genomic data.

Correspondence analysis

Fundamentals of Correspondence Analysis can be summarized into 2 significant points:

- **Extract most significant information included in a data table**
- **Show relationships between observations and/or variables**

Plan

- Introduction to Correspondence Analysis

Examples:

- Species versus base compositions
- Species versus codons Compositions
- Species versus Amino Acids Compositions
- Genome Trees from Whole Proteome Comparisons
- Correspondence Analysis and Principal Component Analysis Methods
- Application to genotyping data
- Conclusion/References

Correspondence Analysis (CA) is an exploratory descriptive method designed to analyse two-way or multi-way data tables including measure of correspondence between rows and columns.

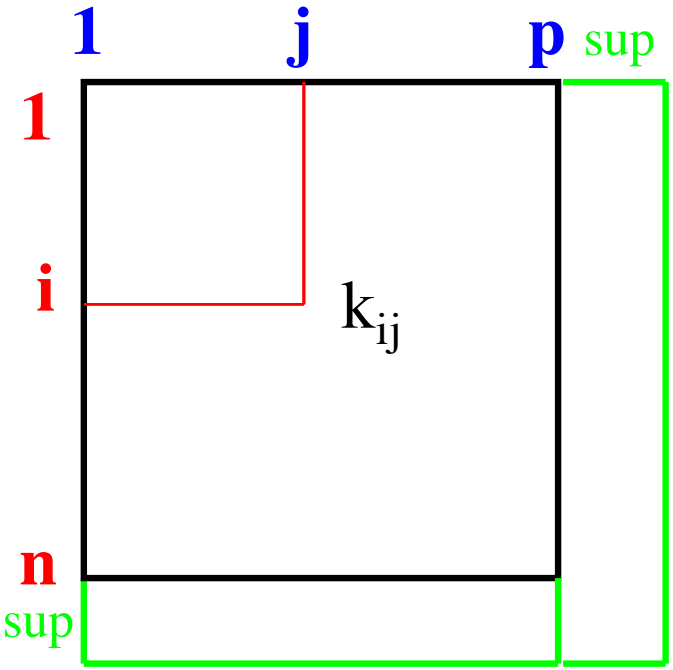
Aims at:

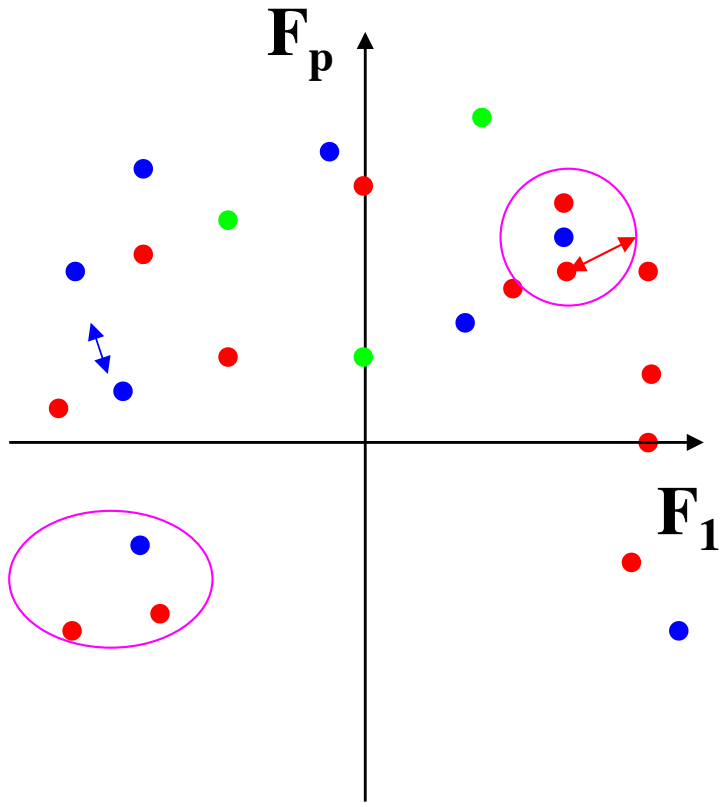
- **Exploring the relationships between rows, between columns and between both of them.**

- **Visual representation of the relationships between rows and columns of such numerical data tables.**

- **Each row and each column is depicted as a point in such a representation.**

Methodology





Matrice T

$k_{ij} > 0$

Correspondence Analysis

$$F_\alpha(i_s) = \lambda_\alpha^{-1/2} . \Sigma \{ f^{is}_j . G_\alpha(j) \; ; \; j=1,p \};$$

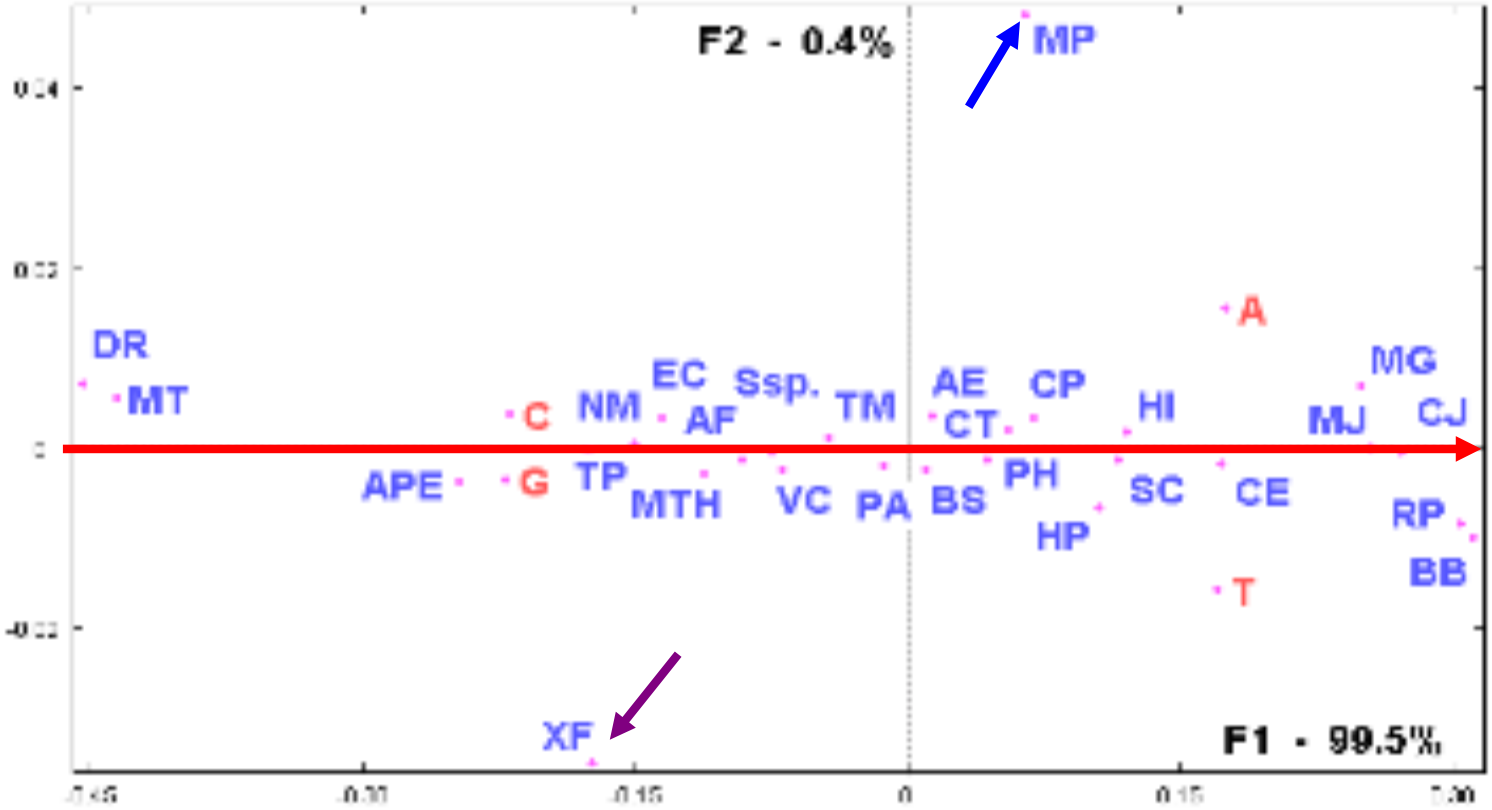
Jean-Paul Benzecri (Paris VI, 1973)

A Simple Example

A+C+G+T=100

Base composition (partial table)

org	A	C	G	T
D. radiodurans	**16.7**	33.3	33.2	**16.6**
M. tuberculosis	**17.1**	32.8	32.7	**17.1**
Aeropyrum pernix	**21.5**	28.3	27.9	**22.1**
Treponema pallidum	**23.5**	26.2	26.5	**23.6**
Xylella fastidiosa	**22.5**	24.9	27.6	**24.7**
N. meningitidis	**24.2**	25.5	25.9	**24.2**
Escherichia coli	**24.6**	25.4	25.3	**24.5**
M. thermoautotroph.	**25.0**	24.7	24.8	**25.3**
A. fulgidus	**25.6**	24.3	24.2	**25.8**
Synechosystis sp	**26.0**	23.8	23.8	**26.1**
Vibrio cholerae	**26.1**	23.6	23.8	**26.3**
T. maritima	**26.9**	22.7	23.4	**26.7**
Pyrococcus abyssi	**27.5**	22.4	22.2	**27.7**
Bacillus subtilis	**28.1**	21.8	21.7	**28.3**
Aquifex aeolicus	**28.4**	21.6	21.7	**28.1**
P. horikoshii	**28.9**	21.2	20.6	**29.1**
C. trachomatis	**29.4**	20.6	20.6	**29.2**
M. pneumoniae	**31.5**	20.0	20.8	**27.6**
C. pneumoniae	**29.8**	20.3	20.2	**29.5**
H. pylori	**30.3**	19.6	19.2	**30.8**
S. cerevisiae	**30.8**	19.1	19.1	**30.8**
H. influenzae	**31.0**	19.1	18.9	**30.8**
C. elegans	**31.6**	17.4	17.3	**31.6**
M. genitalium	**34.5**	15.7	15.9	**33.7**
M. jannaschii	**34.4**	15.5	15.8	**34.1**
C. jejuni	**34.8**	15.3	15.2	**34.6**
R. prowazekii	**35.3**	14.3	14.6	**35.6**
B. burgdorferi	**35.4**	14.3	14.2	**35.9**


F2 - 0.4%
MP
A
DR
MT
EC
Ssp.
AE
CP
HI
MG
MJ
CJ
C
NM
AF
TM
CT
APE
G
TP
MTH
VC
PA
BS
PH
SC
CE
RP
HP
BB
T
XF
F1 - 99.5%

A Simple Example

Base composition (partial table)

org	A	C	G	T	G+C
D. radiodurans	**16.7**	33.3	33.2	**16.6**	**66.5**
M. tuberculosis	**17.1**	32.8	32.7	**17.1**	**65.5**
Aeropyrum pernix	**21.5**	28.3	27.9	**22.1**	**56.2**
Treponema pallidum	**23.5**	26.2	26.5	**23.6**	**52.7**
Xylella fastidiosa	**22.5**	24.9	27.6	**24.7**	**52.5**
N. meningitidis	**24.2**	25.5	25.9	**24.2**	**51.4**
Escherichia coli	**24.6**	25.4	25.3	**24.5**	**50.7**
M. thermoautotroph.	**25.0**	24.7	24.8	**25.3**	**49.5**
A. fulgidus	**25.6**	24.3	24.2	**25.8**	**48.5**
Synechosystis sp	**26.0**	23.8	23.8	**26.1**	**47.6**
Vibrio cholerae	**26.1**	23.6	23.8	**26.3**	**47.4**
T. maritima	**26.9**	22.7	23.4	**26.7**	**46.1**
Pyrococcus abyssi	**27.5**	22.4	22.2	**27.7**	**44.6**
Bacillus subtilis	**28.1**	21.8	21.7	**28.3**	**43.5**
Aquifex aeolicus	**28.4**	21.6	21.7	**28.1**	**43.3**
P. horikoshii	**28.9**	21.2	20.6	**29.1**	**41.8**
C. trachomatis	**29.4**	20.6	20.6	**29.2**	**41.2**
M. pneumoniae	**31.5**	20.0	20.8	**27.6**	**40.8**
C. pneumoniae	**29.8**	20.3	20.2	**29.5**	**40.5**
H. pylori	**30.3**	19.6	19.2	**30.8**	**38.6**
S. cerevisiae	**30.8**	19.1	19.1	**30.8**	**38.2**
H. influenzae	**31.0**	19.1	18.9	**30.8**	**38.0**
C. elegans	**31.6**	17.4	17.3	**31.6**	**34.7**
M. genitalium	**34.5**	15.7	15.9	**33.7**	**31.6**
M. jannaschii	**34.4**	15.5	15.8	**34.1**	**31.3**
C. jejuni	**34.8**	15.3	15.2	**34.6**	**30.5**
R. prowazekii	**35.3**	14.3	14.6	**35.6**	**28.9**
B. burgdorferi	**35.4**	14.3	14.2	**35.9**	**28.5**

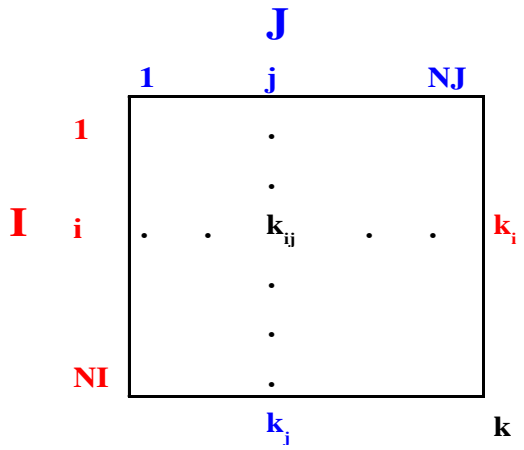
Raw data table

Variables, columns

Individuals

Rows, Lines

Observations

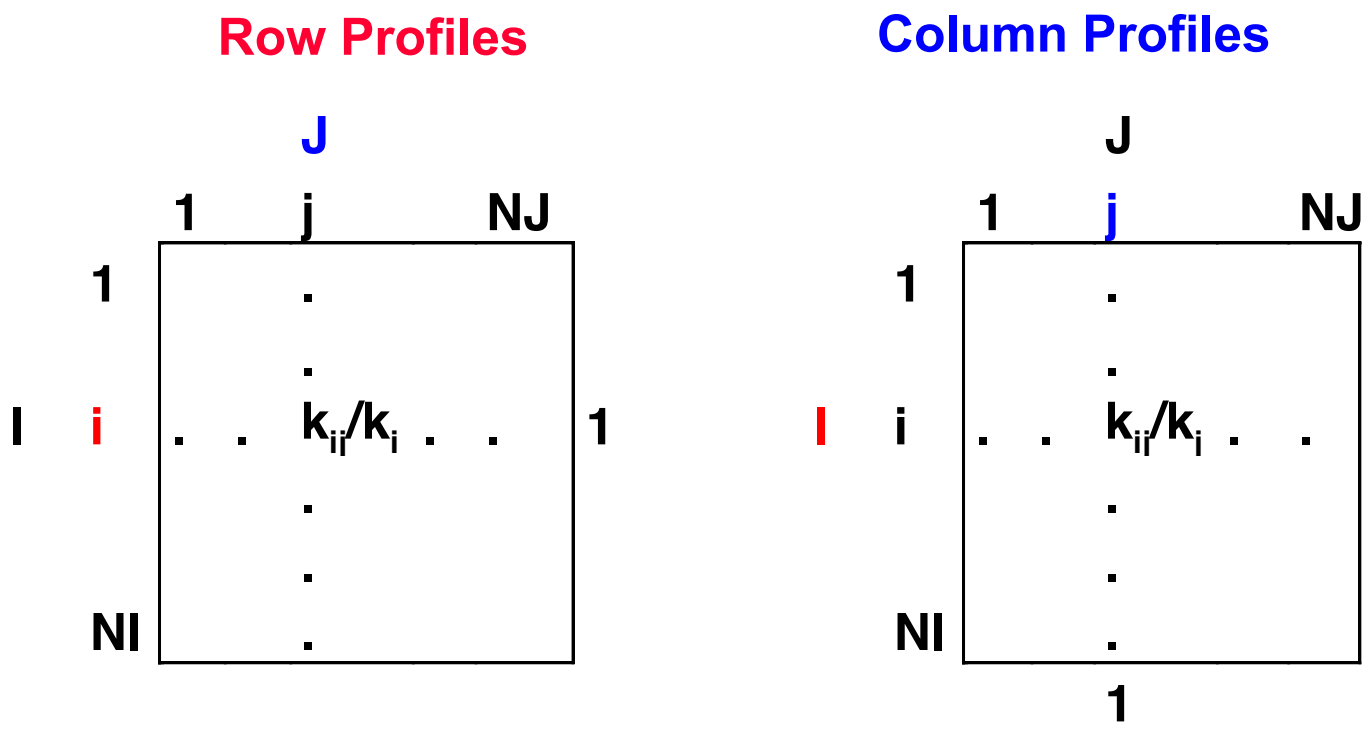


k_i : sum of the row *i*; k_j : sum of the column j and k : total sum of the table.

The sum of the row *i* is denoted k_i : $k_i = \sum\{k_{ij}; j=1, NJ\}$;
The sum of the column *j* is denoted k_j : $k_j = \sum\{k_{ij}; i=1, NI\}$;
The total sum of the data table k_{IJ} is denoted k : $k= \sum\{k_{ij}; i=1,NI; j=1, NJ\}$:
k can also be written : $k = \sum\{k_i; i=1, NI\} = \sum\{k_j; j=1, NJ\}$

These expressions define the type of data tables that can be submitted to correspondence analysis.
The main property of such data tables is that sums of lines should make sense.

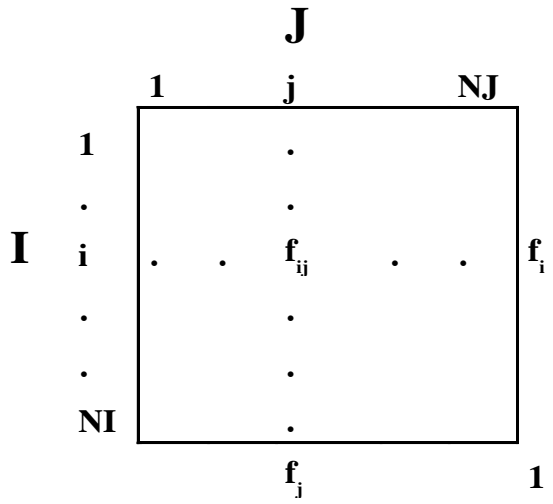
Profiles: rows, columns



The profile of i on J = { k_{i1}/k_i,....., k_{ij}/k_i , ..., k_{iNJ}/k_i};

The profile of j on I = { k_{1j}/k_j,....., k_{ij}/k_i , ..., k_{NIj}/k_j};

Frequencies



$f_{ij} = k_{ij}/k$ frequency associated with (i , j);

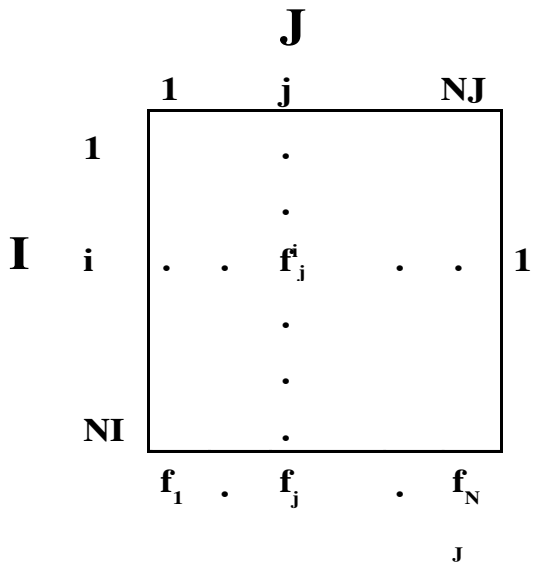
$f_i = k_i/k$ weight of the profile of i;

$f_j = k_j/k$ weight of the profile of j;

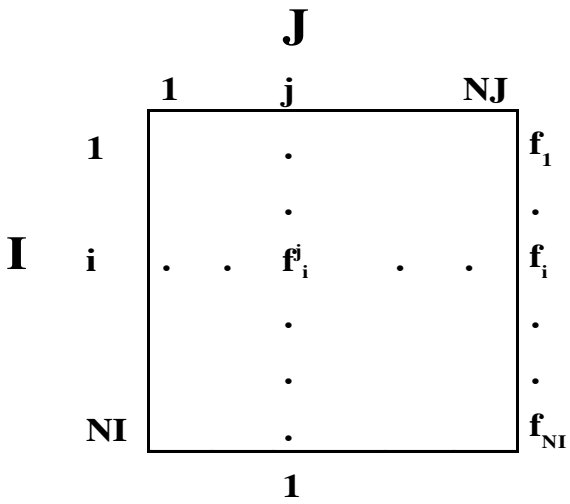
Are respectively weights of i and j relative to the total weight k of the data table.

Profiles, weights

Row profiles



Column profiles



$f^i_j = k_{ij}/k_i = f_{ij}/f_i$ j^{th} coordinate of the profile of i;

$f^j_i = k_{ij}/k_j = f_{ij}/f_j$ i^{th} coordinate of the profile of j;

$f^i_J = \{f^i_j = f_{ij}/f_i\ ,\ j \in J\}$ is the profile of i on J; (f^i_J, f_i)

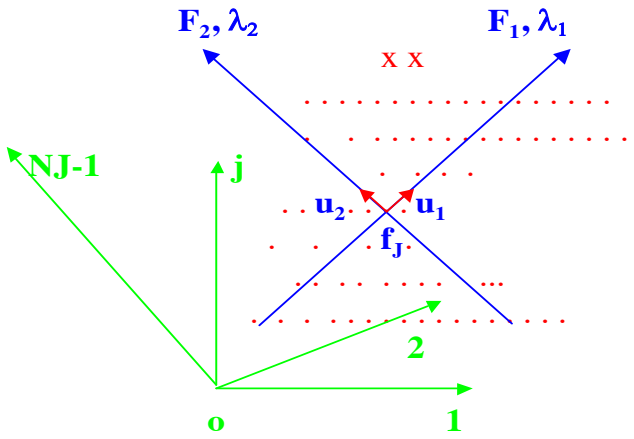
$f^j_I = \{f^j_i = f_{ij}/f_j\ ,\ i \in I\}$ is the profile of j on I; (f^j_I, f_j)

Note the symmetry between i and j.

$f_i = \sum\{f_{ij}\ , j \in J\}$ and $f_j = \sum\{f_{ij}\ , i \in I\}$;weights of i and j.

Graphical representation of the set I: the $N_J(I)$ cloud

$\mathbf{N_J(I)=\{(f^i_J,f_i), i \in I\}}$, the cloud of individual profiles associated with their corresponding weights f_i.



(f^i_J,f_i)

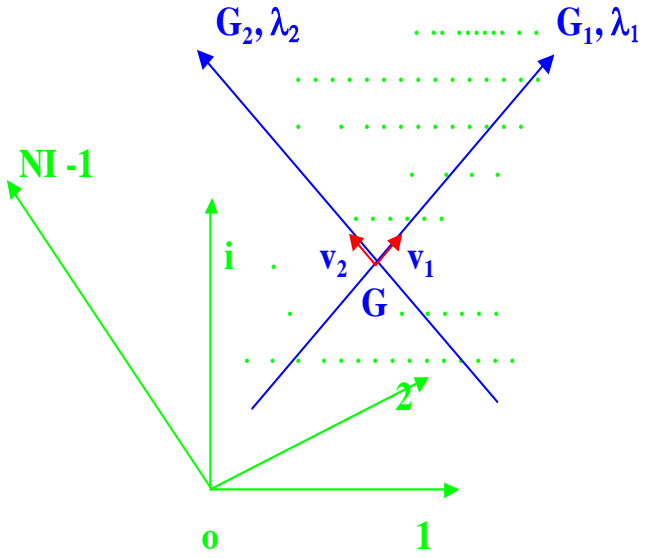
The barycenter of this cloud of points is :

$\mathbf{f_J = \{f_1, ..., f_j, ..., f_{NJ}\}}$

Orthogonal system; λ_α : proportion relative to the whole information

Graphical representation of the set J: the $N_I(J)$ cloud

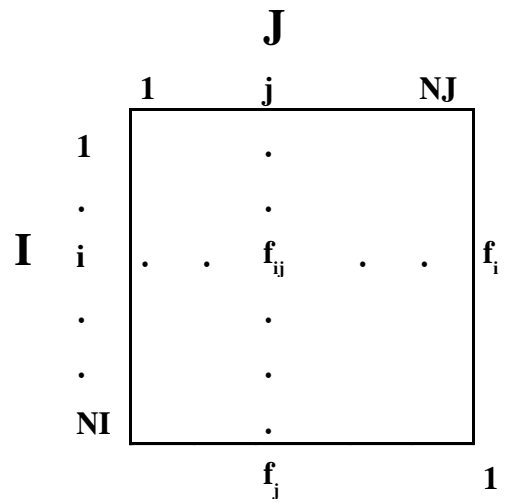
$\mathbf{N_I(J)=\{(f^j_I,f_j), j \in J\}}$, the cloud of variable profiles associated with their corresponding weights f_j.



$\mathbf{(f^j_I,f_j)}$

This cloud of points has f_I as barycenter : $f_I = \{f_1, ..., f_j, ..., f_{NI}\}$.

Formally considering:



$f_{ij} = k_{ij}/k$ frequency associated with (i , j);

$f_i = k_i/k$ weight of the profile of i;

$f_j = k_j/k$ weight of the profile of j;

- a symmetric matrix S is derived with elements:

$s_{ij} = (f_{ij} - f_i.f_j).(f_i f_j)^{1/2}$

- $S = U\Lambda V^t$ where $U^tU = V^tV = VV^t = I_{dentity}$.

Singular Value Decomposition

- U is the orthonormalized eigenvectors (denoted $F_1, F_2, F_3, \ldots, F_\alpha, \ldots$) associated with the largest eigenvalues of SS^t.

- V is the orthonormalized eigenvectors (denoted $G_1, G_2, G_3, \ldots, G_\alpha, \ldots$) of S^tS

- Λ is a diagonal matrix of non negative square roots of the eigenvalues of S^tS denoted λ_α and sorted from largest to smallest

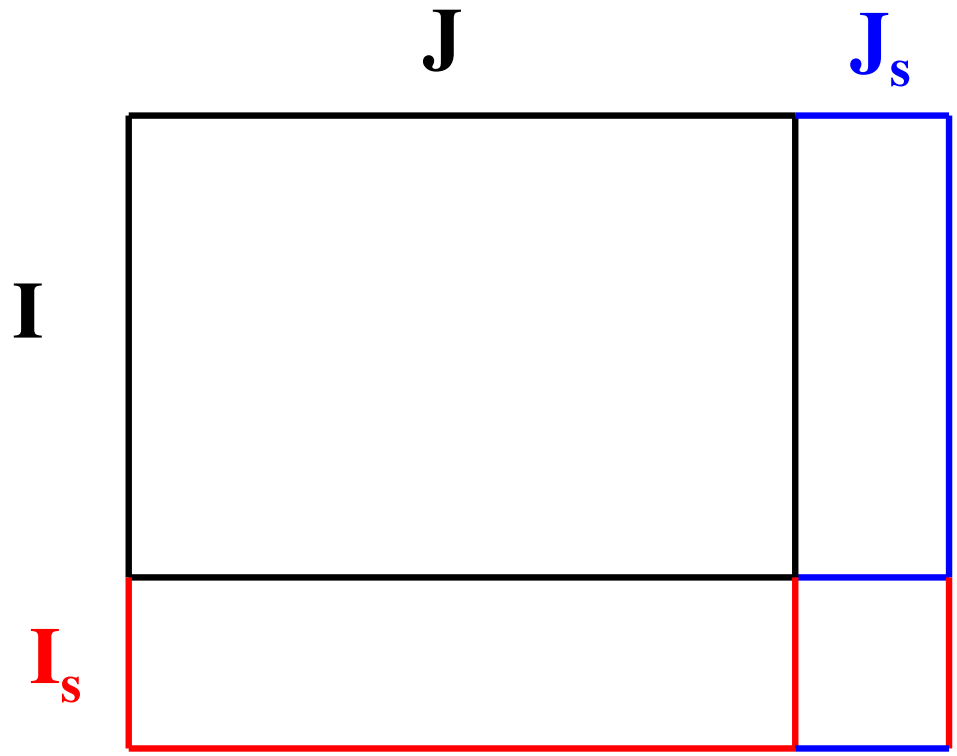
Transition formulae - Barycentric principal

One of the most interesting properties of CA is the ability to express the coordinates of a given individual *i* as a function of the coordinates of the set of the considered variables and vice versa:

- $F_\alpha(i) = \lambda_\alpha^{-1/2} . \sum\{f_j^i G_\alpha(j) \; ; \; j=1, p\}$

- $G_\alpha(j) = \lambda_\alpha^{-1/2} . \sum\{f_i^j F_\alpha(i) \; ; \; i=1, n\}$

Supplementary elements



Given a supplementary individual i_s, its profile on the variable set is denoted : f^{is}_J.

$$F_\alpha(i_s) = \lambda_\alpha^{-1/2} . \Sigma\{f^{is}_j . G_\alpha(j) \; ; \; j=1,p\}$$

A supplementary variable j_s has a profile f^{js}_I.

$$G_\alpha(j_s) = \lambda_\alpha^{-1/2} . \Sigma\{f^{js}_i . F_\alpha(i) \; ; \; i=1,n\}.$$

=> relationships between supp elements and principal elements

Number of factors = min (NI , NJ) - 1.

Where NI is the number of rows and NJ the number of columns

Relative importance of each Factor

Example: 91 yeast/fungal species vs aa composition

FACTOR	EIGEN VALUE	PERCENT.	PECENT. CUMUL.	
F1	0.0184	88.83	88.83	**********
F2	0.0005	2.58	91.41	***
F3	0.0004	2.17	93.58	**
F4	0.0003	1.54	95.12	**
F5	0.0002	0.94	96.06	*
F6	0.0002	0.83	96.89	*
F7	0.0002	0.77	97.66	*
F8	0.0001	0.52	98.19	*
F9	0.0001	0.45	98.64	*
F10	0.0001	0.31	98.95	*
F11	0.0001	0.28	99.22	*
F12	0.0000	0.17	99.40	*
F13	0.0000	0.16	99.55	*
F14	0.0000	0.12	99.67	*
F15	0.0000	0.11	99.78	*
F16	0.0000	0.08	99.86	*
F17	0.0000	0.06	99.92	*
F18	0.0000	0.05	99.96	*
F19	0.0000	0.04	100.00	*

Factor	Eigen value	PERCENT.	PERCENT. CUMULATED	
F1	0.0521	65.82	65.82	**********
F2	0.0051	6.49	72.30	********
F3	0.0048	6.06	78.36	********
F4	0.0030	3.79	82.15	*****
F5	0.0020	2.55	84.70	****
F6	0.0016	2.06	86.76	***
F7	0.0009	1.14	87.90	**
F8	0.0007	0.84	88.74	**
F9	0.0006	0.79	89.53	*
F10	0.0006	0.72	90.25	*
F11	0.0005	0.66	90.91	*
F12	0.0005	0.63	91.54	*
13	0.0005	0.62	92.15	*
14	0.0005	0.58	92.74	*
15	0.0004	0.47	93.21	*
16	0.0004	0.45	93.65	*
17	0.0003	0.39	94.05	*
18	0.0003	0.37	94.42	*

92 yeast/fungal species shared orthologs

Coordinates - Contributions - Correlations

Example: 91 yeast/fungal species vs aa composition

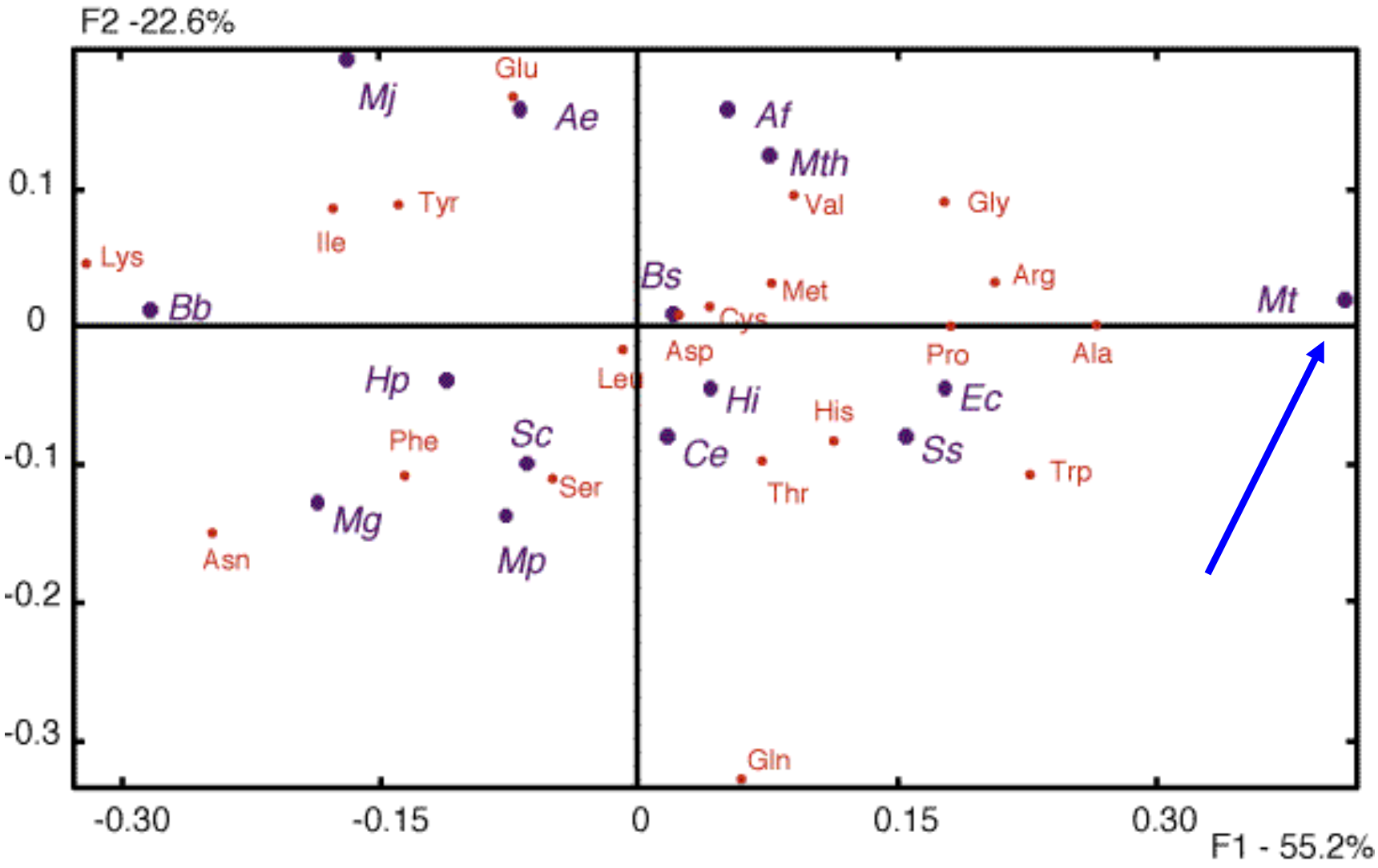
VARIABLES			COORDINATES					CONTRIBUTIONS					COSINUS**2				
IDEN	REL.W	DISTO	F1	F2	F3	F4	F5	C1	C2	C3	C4	C5	CO1	CO2	CO3	CO4	CO5
A - Ala	7.28	0.06	0.24	-0.02	-0.02	-0.01	0.02	22.0	3.7	6.6	3.5	18.1	0.97	0.00	0.01	0.00	0.01
R - Arg	5.41	0.04	0.18	0.02	0.03	-0.03	0.00	9.8	3.5	8.7	13.1	0.2	0.92	0.01	0.02	0.02	0.00
N - Asn	4.65	0.07	-0.26	-0.06	0.02	0.01	0.01	17.5	26.7	2.6	0.4	3.2	0.94	0.04	0.00	0.00	0.00
D - Asp	5.78	0.00	-0.02	-0.01	-0.01	0.01	0.00	0.1	1.3	1.4	3.8	0.4	0.23	0.08	0.07	0.14	0.01
C - Cys	1.22	0.01	-0.01	0.07	0.03	-0.02	0.04	0.0	12.6	2.2	0.9	8.9	0.01	0.56	0.08	0.02	0.14
Q - Gln	4.03	0.00	0.00	-0.01	-0.02	0.02	0.00	0.0	1.3	3.8	7.1	0.1	0.00	0.05	0.13	0.17	0.00
E - Glu	6.44	0.00	-0.03	0.02	0.01	0.01	-0.02	0.4	5.4	1.1	0.7	16.7	0.43	0.17	0.03	0.01	0.20
G - Gly	6.07	0.02	0.15	-0.01	0.01	0.02	-0.01	7.4	0.8	2.9	9.3	2.4	0.93	0.00	0.01	0.02	0.00
H - His	2.29	0.00	0.05	0.00	0.00	-0.01	0.01	0.4	0.1	0.0	1.1	1.0	0.64	0.01	0.00	0.03	0.02
I - Ile	5.61	0.04	-0.19	0.00	0.04	-0.02	0.01	11.1	0.1	18.1	4.7	6.1	0.93	0.00	0.04	0.01	0.01
L - Leu	9.20	0.00	-0.03	0.02	-0.01	-0.01	0.00	0.6	5.6	3.8	1.2	0.4	0.52	0.15	0.09	0.02	0.00
K - Lys	5.92	0.04	-0.21	0.01	0.00	0.00	-0.01	13.6	0.7	0.2	0.0	6.6	0.97	0.00	0.00	0.00	0.01
M - Met	2.09	0.00	0.02	0.03	0.02	0.01	0.02	0.1	3.1	1.3	1.1	4.0	0.10	0.18	0.06	0.04	0.09
F - Phe	4.01	0.01	-0.11	0.02	-0.01	0.00	0.00	2.5	4.3	0.8	0.0	0.1	0.88	0.04	0.01	0.00	0.00
P - Pro	5.31	0.03	0.18	-0.02	0.03	0.00	-0.03	9.2	5.6	13.1	0.1	18.6	0.92	0.02	0.03	0.00	0.02
S - Ser	8.61	0.00	-0.03	-0.02	-0.03	-0.03	-0.01	0.5	5.0	13.1	32.5	7.0	0.30	0.08	0.17	0.31	0.04
T - Thr	5.82	0.00	0.02	-0.03	0.00	0.02	0.01	0.1	7.9	0.0	6.7	3.1	0.11	0.30	0.00	0.15	0.04
W - Trp	1.23	0.03	0.15	0.02	0.06	0.02	0.02	1.5	0.8	8.3	1.1	2.4	0.78	0.01	0.11	0.01	0.01
Y - Tyr	3.04	0.02	-0.13	0.03	0.00	0.01	0.01	2.8	3.9	0.0	1.9	0.5	0.90	0.04	0.00	0.01	0.00
V - Val	6.00	0.00	0.04	0.03	-0.03	0.02	0.00	0.6	7.7	12.0	10.7	0.4	0.43	0.16	0.20	0.13	0.00

INDIVIDUALS			COORDINATES					CONTRIBUTIONS					COSINUS**2				
IDENT	REL.W	DISTO	F1	F2	F3	F4	F5	C1	C2	C3	C4	C5	CO1	CO2	CO3	CO4	CO5
SACE	1.10	0.02	-0.15	0.00	0.00	0.00	0.00	1.3	0.0	0.0	0.1	0.0	0.97	0.00	0.00	0.00	0.00
SAAR	1.10	0.02	-0.14	0.00	0.00	0.00	0.00	1.2	0.0	0.0	0.0	0.1	0.97	0.00	0.00	0.00	0.00
NACA	1.10	0.03	-0.16	-0.01	0.01	0.01	0.00	1.4	0.3	0.3	0.3	0.1	0.96	0.01	0.01	0.00	0.00
CAGL	1.10	0.02	-0.14	0.00	0.00	0.01	0.00	1.1	0.0	0.0	0.7	0.0	0.95	0.00	0.00	0.01	0.00
NADE	1.10	0.02	-0.13	0.01	0.00	0.02	0.00	1.0	0.1	0.0	1.0	0.1	0.93	0.00	0.00	0.02	0.00
PIST	1.10	0.02	-0.13	-0.01	-0.02	0.00	-0.01	1.0	0.1	1.1	0.0	0.5	0.90	0.00	0.03	0.00	0.01
PISO	1.10	0.02	-0.13	0.00	-0.02	-0.03	-0.02	1.0	0.0	0.6	2.3	2.6	0.86	0.00	0.01	0.04	0.02
LOEL	1.10	0.02	-0.12	-0.03	-0.02	0.02	0.00	0.9	1.8	0.8	2.0	0.0	0.82	0.05	0.02	0.03	0.00
CAOR	1.10	0.02	-0.13	-0.01	-0.01	0.01	-0.01	1.0	0.4	0.4	0.6	0.5	0.93	0.01	0.01	0.01	0.01
CAPA	1.10	0.02	-0.13	-0.02	-0.03	0.03	-0.01	1.0	0.9	2.7	2.5	0.2	0.82	0.02	0.06	0.04	0.00
CATE	1.10	0.02	-0.13	0.01	-0.01	0.01	0.00	1.0	0.4	0.5	0.5	0.1	0.89	0.01	0.01	0.01	0.00
TEBL	1.10	0.06	-0.23	-0.08	0.02	-0.02	0.02	3.1	13.2	1.5	0.8	2.4	0.85	0.11	0.01	0.00	0.01
CATR	1.10	0.03	-0.17	-0.03	0.01	0.02	-0.01	1.7	1.5	0.1	0.8	0.6	0.94	0.02	0.00	0.01	0.00
CAAL	1.10	0.04	-0.18	-0.04	0.01	0.01	0.00	2.0	3.3	0.1	0.6	0.0	0.93	0.04	0.00	0.00	0.00

.....

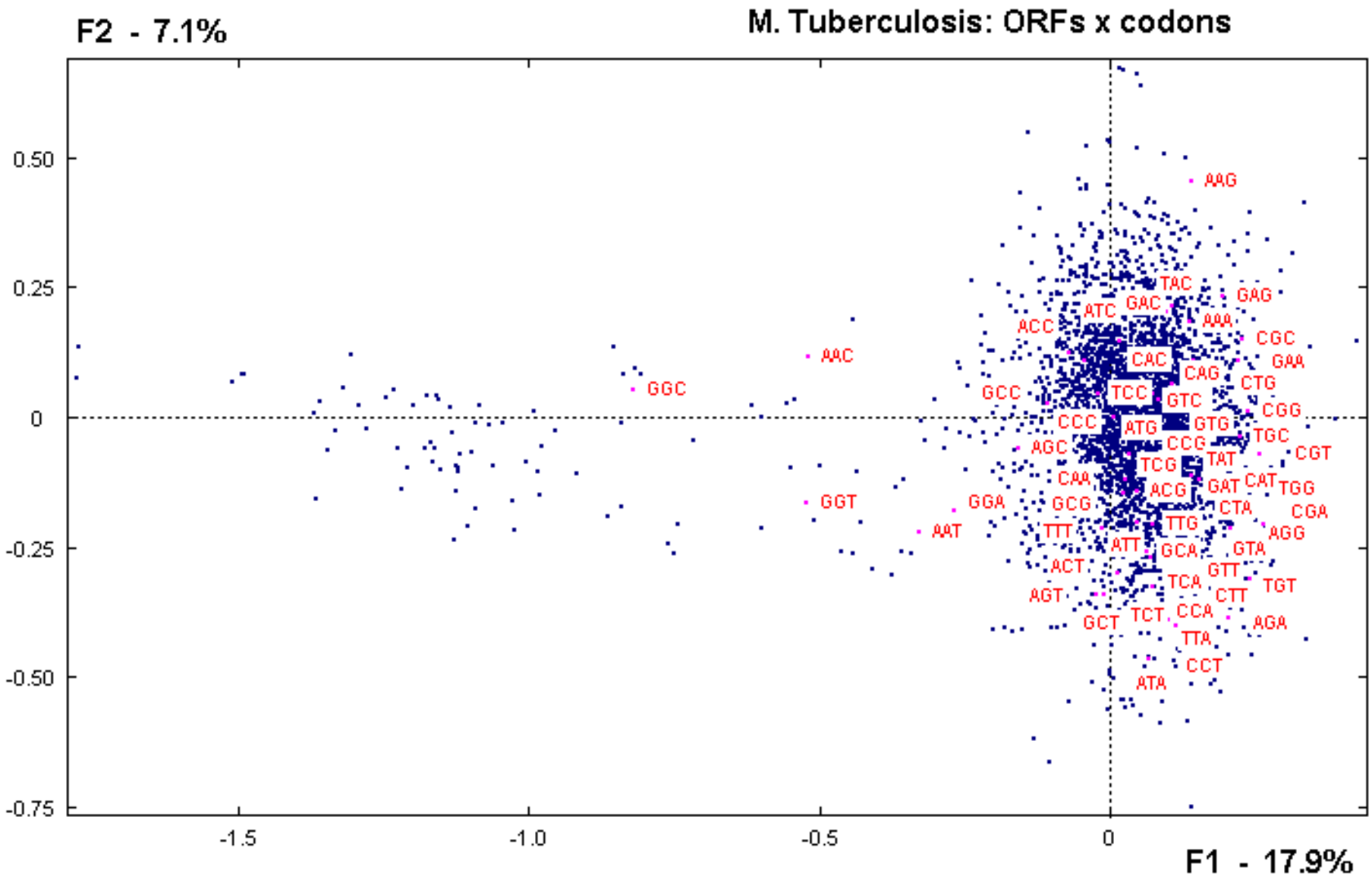
Examples

Amino-acids composition of the first published proteomes



Cole ST, Brosch R, Parkhill J, et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998 Jun 11;393(6685):537-44.

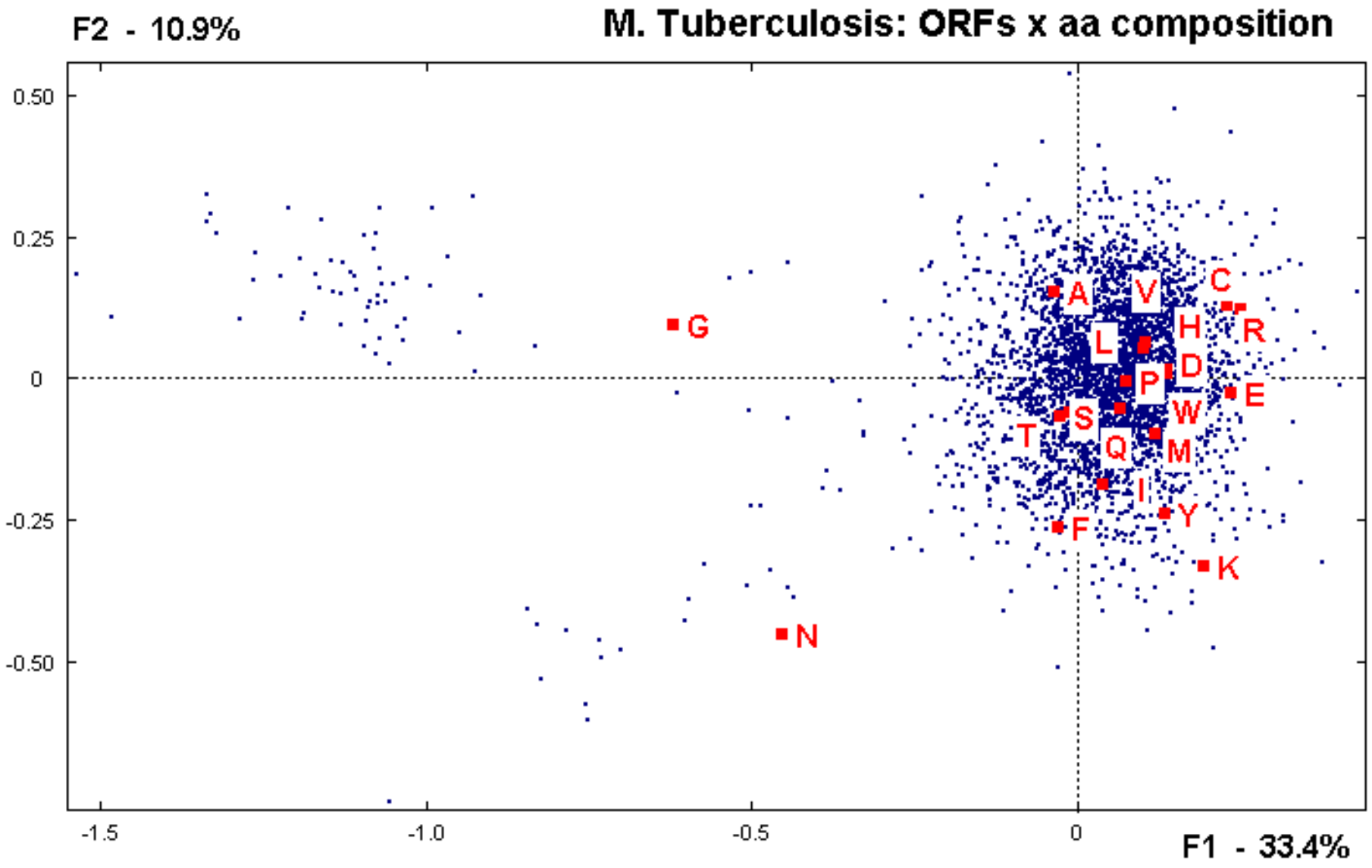
Codons composition of Mycobacterium tuberculosis total genes



3996 gene sequences vs 64 codons

https://webext.pasteur.fr/tekaia/caongenomes.html

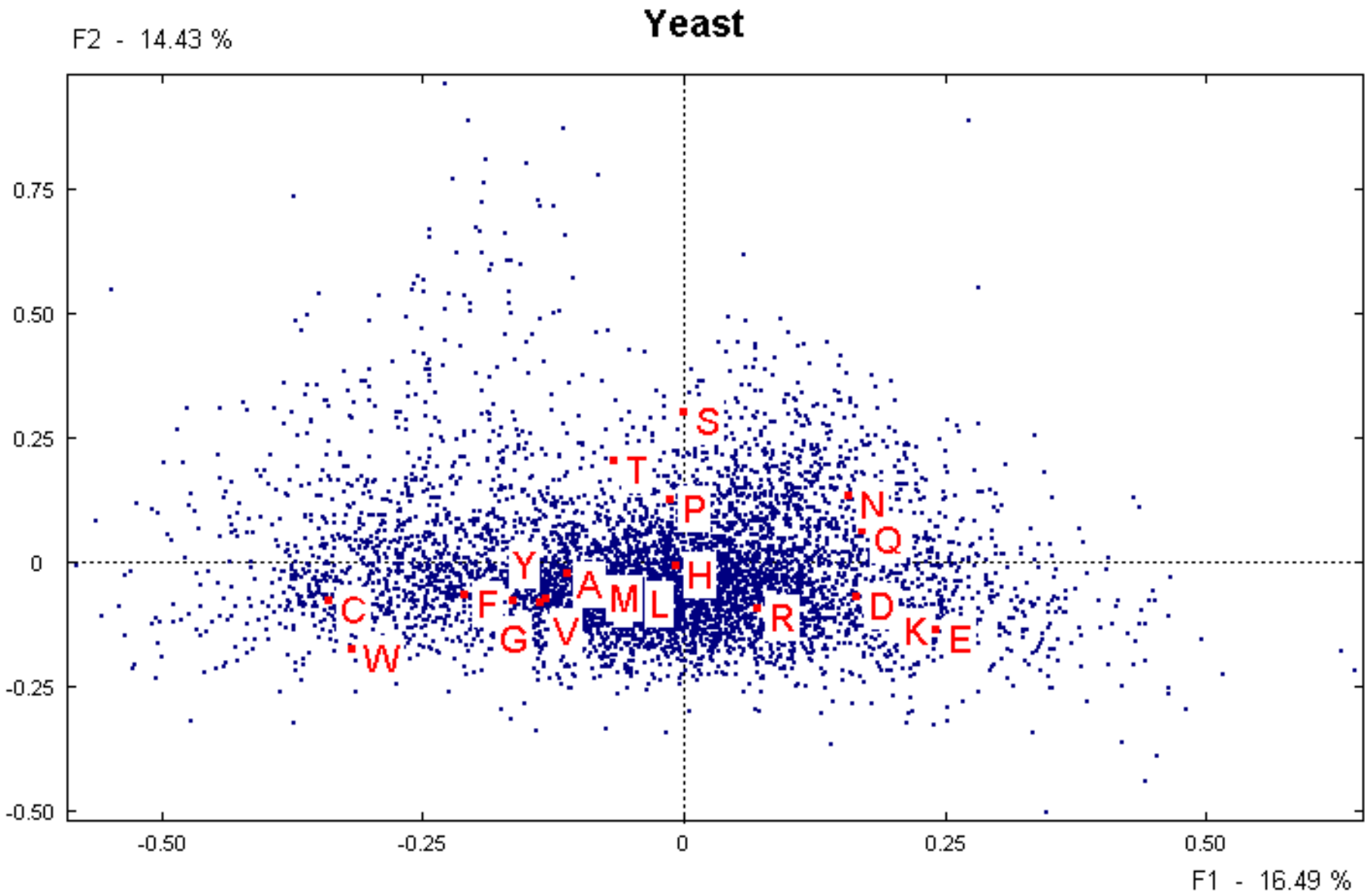
aa composition of Mycobacterium tuberculosis total proteins



3996 prot sequences vs 20 aa

https://webext.pasteur.fr/tekaia/caongenomes.html

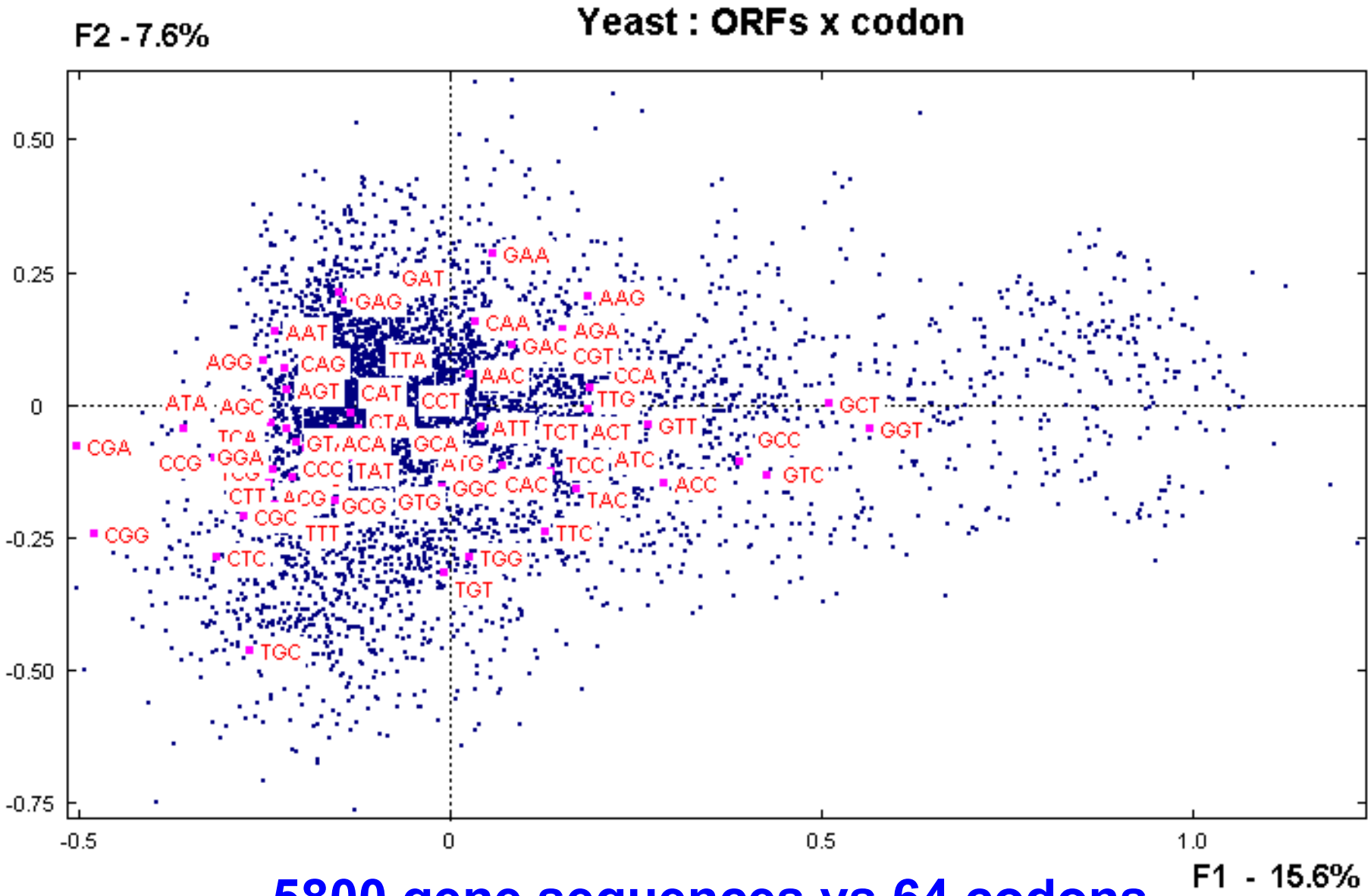
aa composition of *Saccharomyces cerevisiae* total proteins



5800 prot sequences vs 20 aa

https://webext.pasteur.fr/tekaia/caongenomes.html

Codons composition of *Saccharomyces cerevisiae* total genes



5800 gene sequences vs 64 codons

https://webext.pasteur.fr/tekaia/caongenomes.html

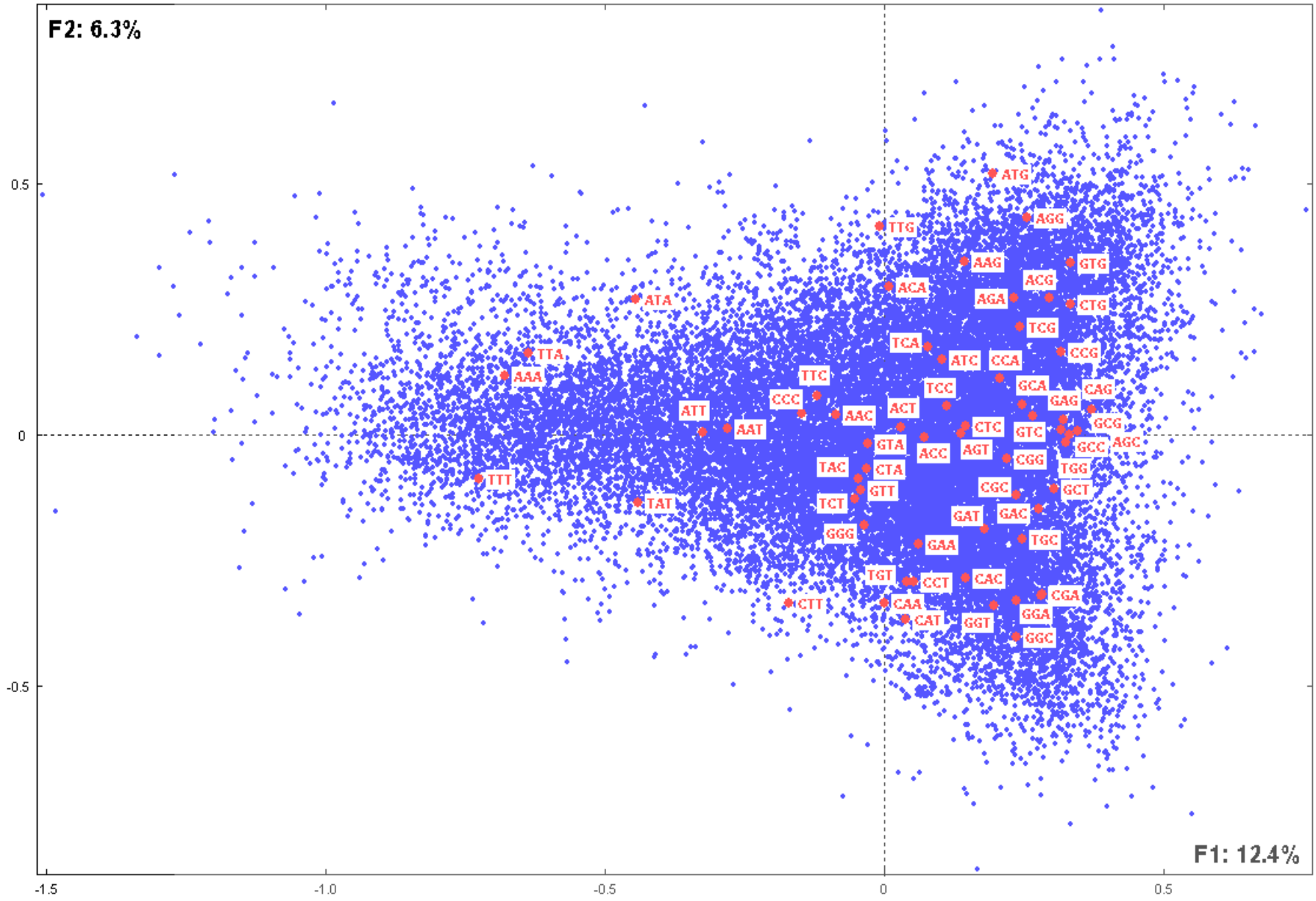
Species versus aa x codons compositions

ID	CDs/Prot	Size (Mb)	GC%	species name
CRFA	11646	40.25	56.3	Crithidia_fasciculata
LEBR3	8567	35.21	57	Leishmania_braziliensis_MHOM/BR/75/M2903
LEBR4	8357	NF		Leishmania_braziliensis_MHOM/BR/75/M2904
LEDO	8083	32.4	59.0	Leishmania_donovani_BPK282A1
LEIN	8239	NF		Leishmania_infantum_JPCM5
LEFR	8400	32.86	59.7	Leishmania_major_strain_Friedlin
LEME	8250	32	59	Leishmania_mexicana_MHOM/GT/2001/U1103
LETA	8452	NF		Leishmania_tarentolae_Parrot-TarII
TRBL	8833	NF		Trypanosoma_brucei_Lister_strain_427
TRBT	11567	NF		Trypanosoma_brucei_TREU927
TRBG	9895	35		Trypanosoma_brucei_gambiense_DAL972
TRCO	13148	15.31	47.6	Trypanosoma_congolense_IL3000
TRCBE	10342	38.06	50.9	Trypanosoma_cruzi_CL_Brener_Es-like
TRCBNE	10834	Draft		Trypanosoma_cruzi_CL_Brener_Non-Es-like
TRCS	10876	Draft		Trypanosoma_cruzi_Sylvio_X10/1
TRCM	10228	34.23	51	Trypanosoma_cruzi_marinkellei_strain_B7
TREV	10111	NF		Trypanosoma_evansi_strain_STIB_805
TRGR	10591	20.93	53.9	Trypanosoma_grayi_ANR4
TRVI	11885	13.99	53.6	Trypanosoma_vivax_Y486
PhmRNA	16265			Phlebotomus mRNA
PhmRNA2	44129			Phlebotomus mRNA

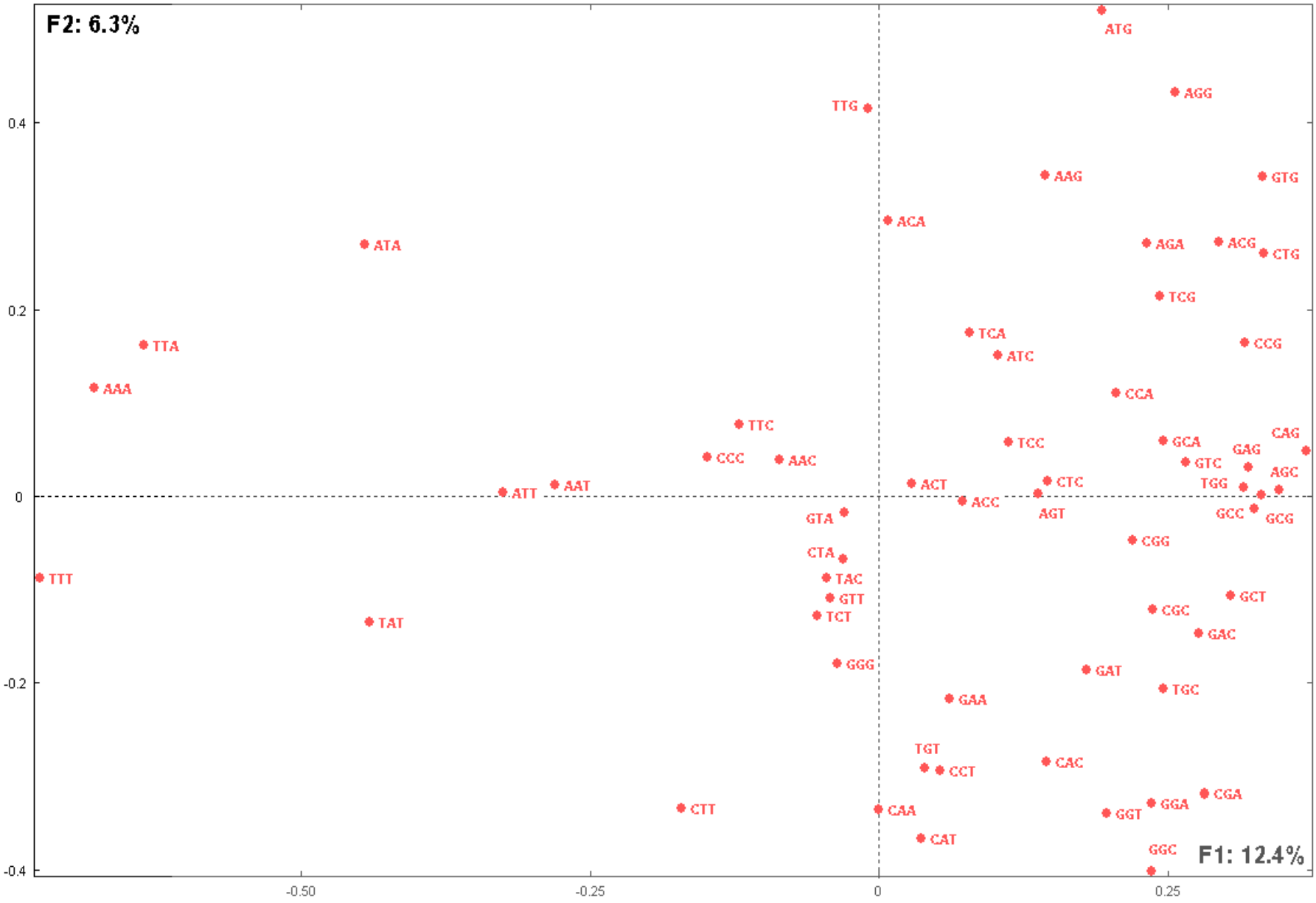
ORF	TTT	TTC	TTA	TTG	CTT	CTC	CTA	CTG	ATT
	ATC	ATA	ATG	GTT	GTC	GTA	GTG	TCT	TCC
	TCA	TCG	CCT	CCC	CCA	CCG	ACT	ACC	ACA
	ACG	GCT	GCC	GCA	GCG	TAT	TAC	CAT	CAC
	CAA	CAG	AAT	AAC	AAA	AAG	GAT	GAC	GAA
	GAG	TGT	TGC	TGG	CGT	CGC	CGA	CGG	AGT
	AGC	AGA	AGG	GGT	GGC	GGA	GGG		
JC5936	2.2	4.3	0.	0.9	0.4	2.6	0.	5.2	0.9
	2.6	0.4	3.9	0.9	0.4	0.	6.0	0.	1.3
	1.3	0.9	2.2	1.7	2.6	0.9	0.9	0.9	0.
	1.3	1.3	1.7	0.9	0.4	0.4	2.2	2.2	1.3
	2.2	2.2	2.2	1.3	1.7	5.2	5.2	2.2	1.3
	4.7	0.9	1.7	1.7	0.9	1.3	0.	1.3	0.9
	0.9	1.7	0.9	1.7	0.9	1.7	0.9		
JK973632		4.1	1.9	1.9	3.0	1.9	0.4	1.1	1.1
	1.5	1.5	1.1	2.6	0.4	0.7	2.2	0.7	1.9
	0.4	1.1	0.7	0.7	1.1	1.1	0.	1.9	1.1
	1.1	0.4	0.	0.	1.5	0.4	2.6	1.9	1.9
	0.	2.6	1.9	2.2	2.6	9.0	2.6	2.6	0.4
	4.9	0.	1.9	3.4	1.9	0.4	0.7	2.2	0.4
	0.7	1.9	3.0	2.2	1.5	0.7	3.4	1.1	

.....

Phlebotomus mRNA : 44129 CDSs vs 64 codons



Phlebotomus mRNA : 44129 CDSs vs 64 codons

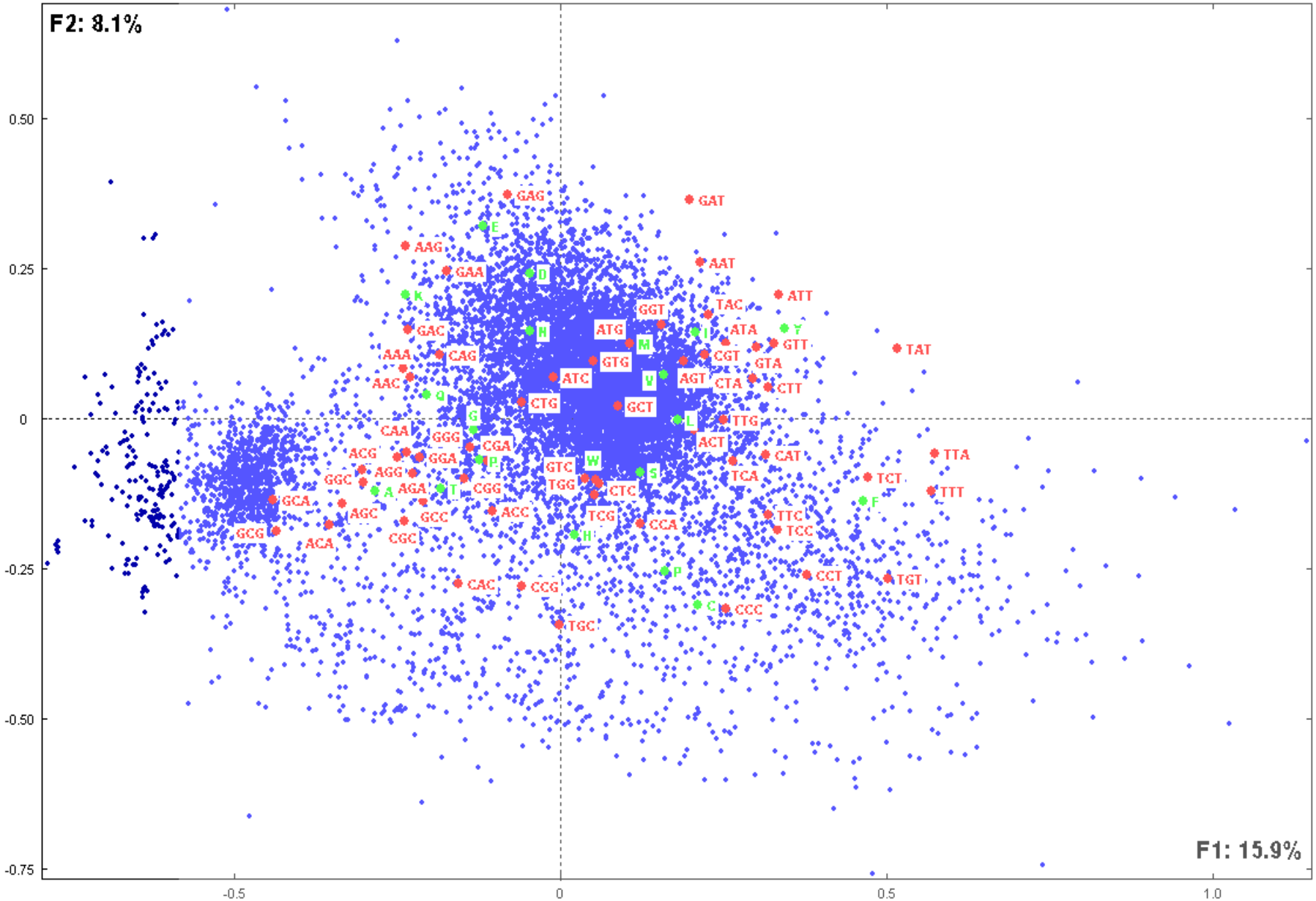


Phlebotomus mRNA: Codons distribution

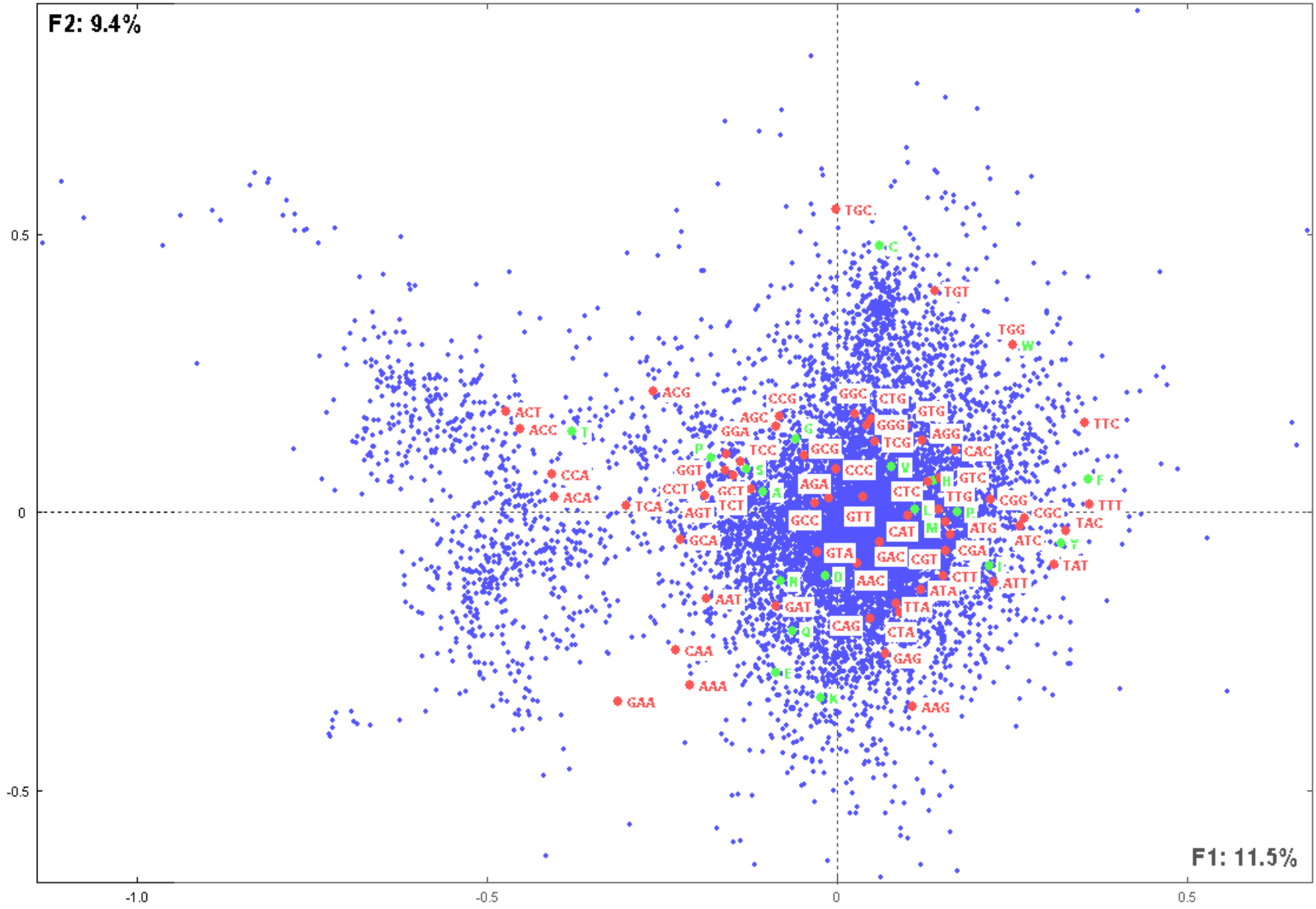
Spec	A	GCT	GCC	GCA	GCG	R	CGT	CGC	CGA
	CGG	AGA	AGG	N	AAT	AAC	D	GAT	GAC
	C	TGT	TGC	Q	CAA	CAG	E	GAA	GAG
	G	GGT	GGC	GGA	GGG	H	CAT	CAC	I
	ATT	ATC	ATA	L	TTA	TTG	CTT	CTC	CTA
	CTG	K	AAA	AAG	M	ATG	F	TTT	TTC
	P	CCT	CCC	CCA	CCG	S	TCT	TCC	TCA
	TCG	AGT	AGC	T	ACT	ACC	ACA	ACG	W
	TGG	Y	TAT	TAC	V	GTT	GTC	GTA	GTG
TvY486_0100040	6.2	1.8	1.5	1.8	1.1	9.2	5.1	1.8	
	1.1	1.1	0.	0.	3.7	1.8	1.8	5.9	3.3
	2.6	5.1	3.3	1.8	2.9	1.8	1.1	4.8	3.7
	1.1	4.4	2.9	0.4	0.4	0.7	1.5	1.1	0.4
	4.8	3.3	0.4	1.1	10.3	1.1	2.9	4.0	0.7
	0.7	0.7	3.3	1.5	1.8	1.5	1.5	5.9	3.7
	2.2	5.9	2.6	1.5	1.5	0.4	11.0	4.4	1.8
	1.5	1.5	1.1	0.7	2.2	1.5	0.	0.4	0.4
	1.8	1.8	2.6	1.1	1.5	7.3	4.8	0.	1.1
	1.5								
TvY486_1010940	5.3	1.6	2.0	0.8	0.8	7.0	1.6	0.8	
	0.4	0.	2.0	2.0	7.4	6.1	1.2	4.9	3.3
	1.6	0.8	0.8	0.	3.3	2.0	1.2	5.7	4.1
	1.6	5.3	1.2	1.2	2.5	0.4	2.5	2.0	0.4
	5.7	3.3	1.2	1.2	4.5	1.6	0.4	0.	0.4
	1.2	0.8	7.0	4.9	2.0	3.3	3.3	3.7	2.9
	0.8	4.9	2.0	0.8	1.6	0.4	6.6	1.2	0.4
	1.6	0.8	1.2	1.2	7.8	2.5	0.8	3.7	0.8
	0.4	0.4	2.9	1.2	1.6	11.1	5.7	1.2	1.2
	2.9								

.....

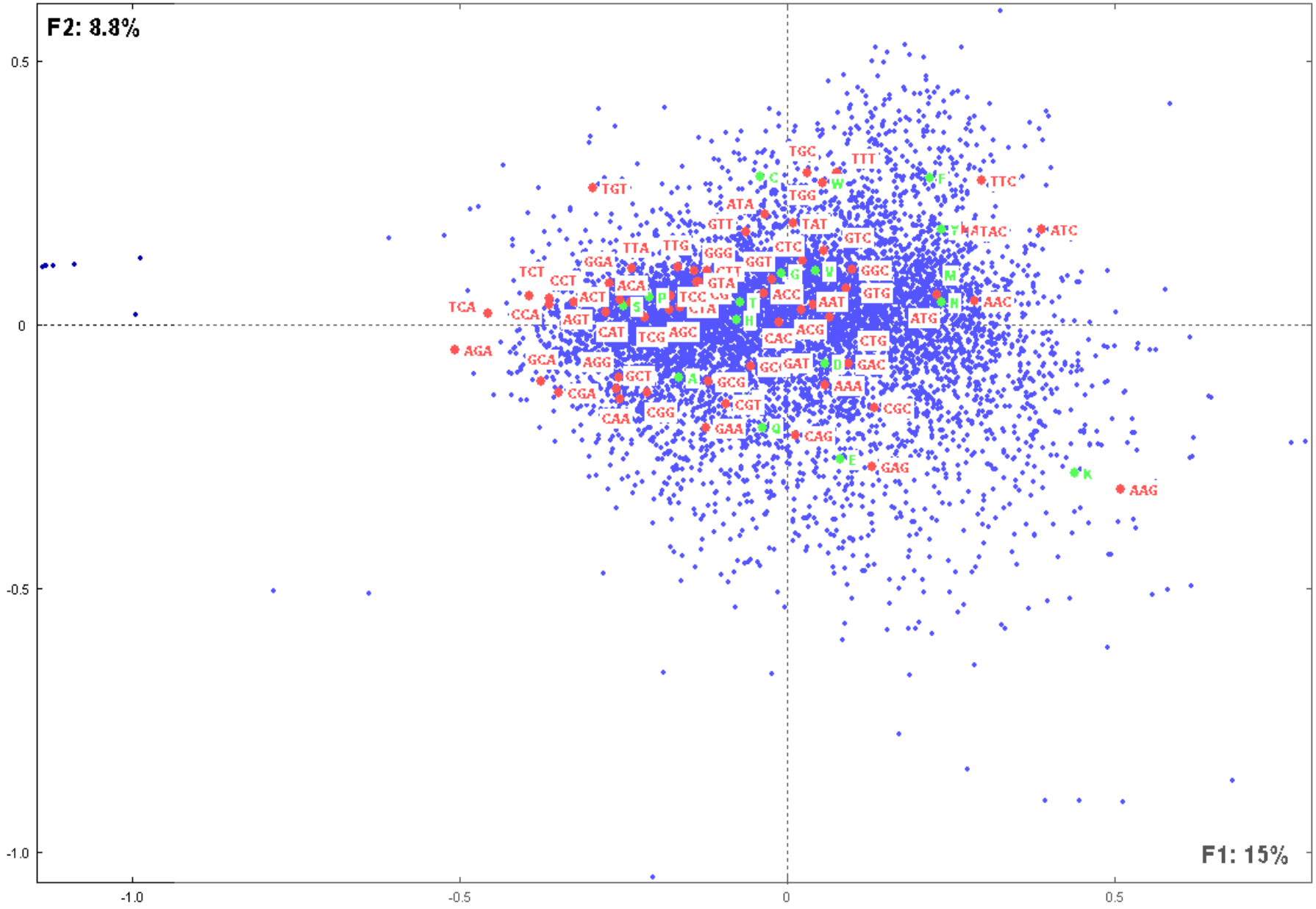
Prot/CDS vs 20 aa x 64 codons



Trypanosoma_vivax_Y486 (11885 vs 84)



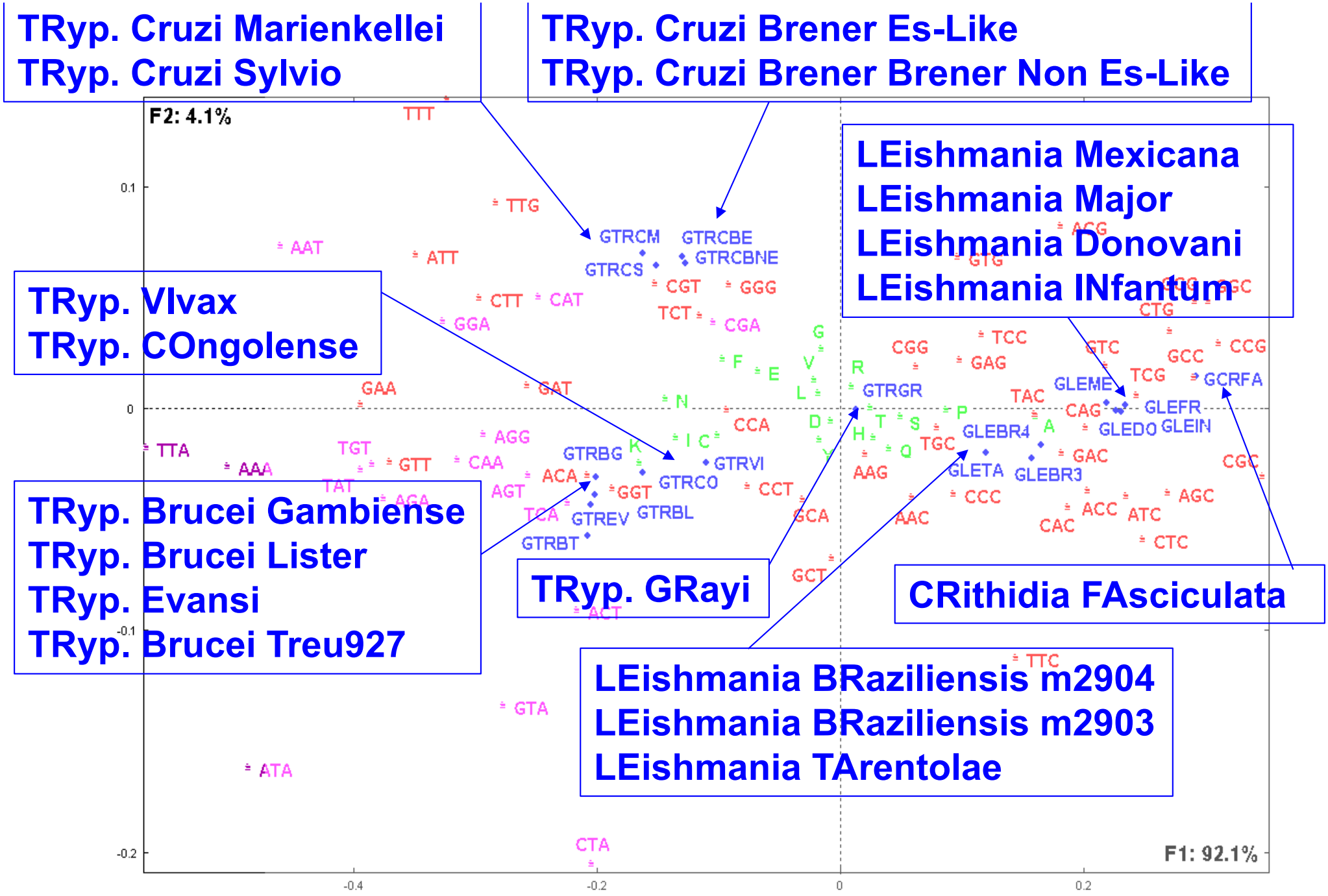
Trypanosoma_cruzi_CL_Brener 10342 cs 84



Leishmania_mexicana 8252 cs 84

All species vs aa & codons compositions

19 species vs 20 aa vs 64 codons



All species vs aa vs codons

EXAMPLES:

Amino acid composition

Evolution of Proteomes: Signatures and Trends in Amino Acid Compositions

- Are there specific properties associated with lifestyles and with phylogeny?

- What are the underlying evolutionary trends?

Tekaia, F. and Yeramian, E. (2006). Evolution of Proteomes: Fundamental signatures and global trends in amino acid composition. *BMC Genomics*. 7:307.

Amino Acid composition of 208 proteomes

including:

- **20 hyperthermophiles (HTH) (OGT >60° C up to 120° C),**
- **7 thermophiles (TH) (OGT >50° C up to 60° C),**
- **8 psychrophiles (PSYC) (OGT: -10° C, up to 15° C),**
- **173 mesophiles (BMES) including 53 eukaryotes (EUK)**

Data table: 222 (208 + 14 sup) vs 23(20 aa + pol, char, hyd)

Tekaia, F. and Yeramian, E. (2006). Evolution of Proteomes: Fundamental signatures and global trends in amino acid composition. *BMC Genomics*. 7:307.

Amino Acid composition

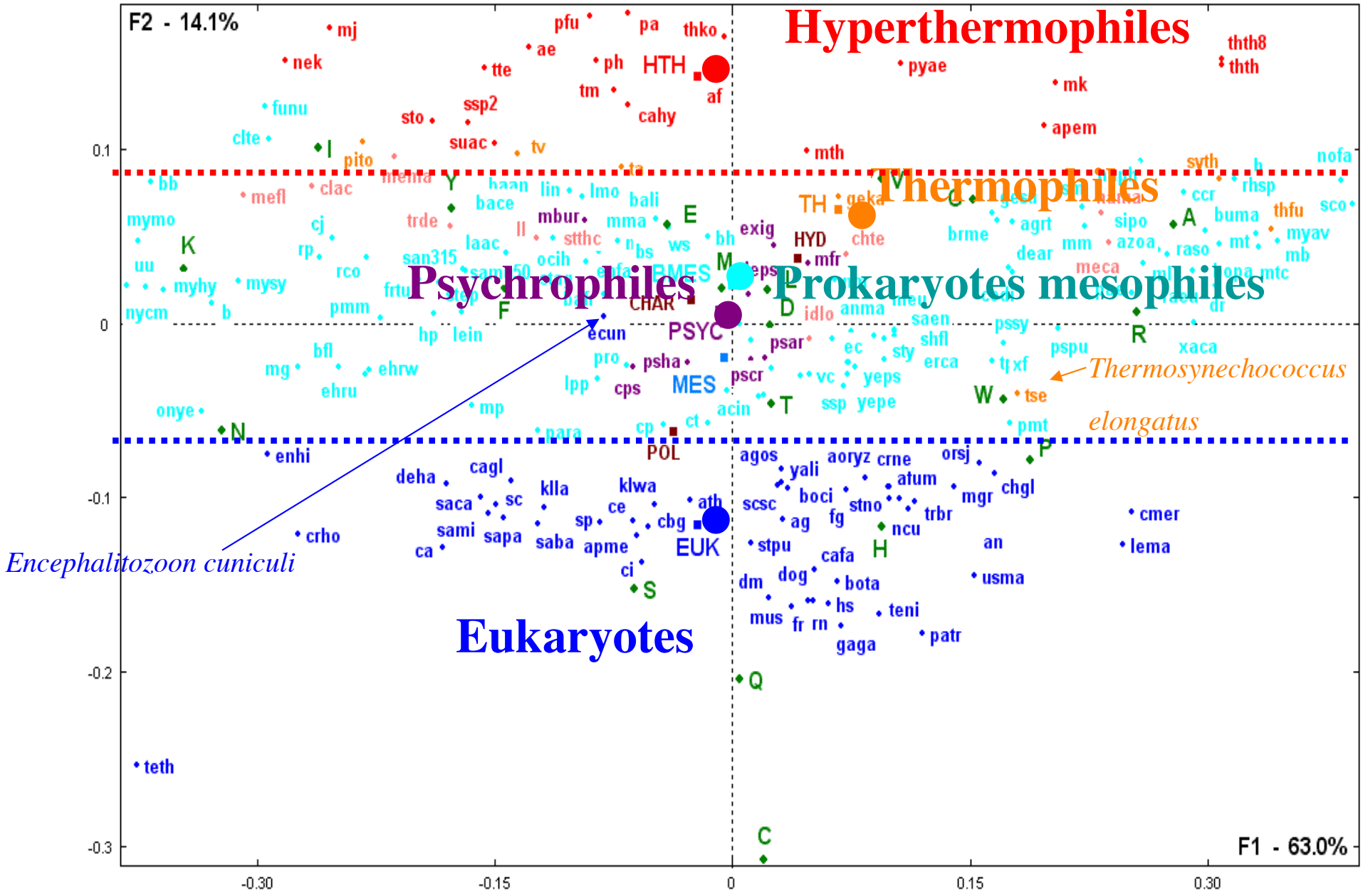
208

org	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	char	pol	hyd	PC
sc	5.5	4.4	6.1	5.8	1.3	3.9	6.6	5.0	2.1	6.6	9.6	7.4	2.1	4.5	4.3	9.0	5.8	1.0	3.3	5.6	26.3	34.4	39.1	8.1
sp	6.3	4.8	5.2	5.3	1.5	3.8	6.5	5.0	2.3	6.1	9.8	6.4	2.1	4.6	4.7	9.4	5.6	1.1	3.4	6.0	25.3	33.9	40.7	8.6
ncu	8.7	6.2	3.7	5.6	1.1	4.3	6.5	7.2	2.5	4.4	8.4	5.1	2.2	3.4	6.5	8.3	6.1	1.4	2.6	6.0	25.8	33.3	40.8	7.5
ca	4.9	3.7	6.7	5.7	1.2	4.4	6.2	5.0	2.1	7.1	9.3	7.2	1.9	4.5	4.5	9.3	6.2	1.0	3.5	5.5	25.	36.2	38.7	11.2
mgr	9.4	6.6	3.5	5.7	1.3	4.1	5.9	7.4	2.3	4.4	8.5	4.8	2.2	3.5	6.3	8.0	5.9	1.5	2.5	6.2	25.3	32.7	42.0	7.4
fg	8.2	5.8	3.9	5.9	1.3	4.0	6.2	6.7	2.4	5.1	8.7	5.1	2.3	3.8	5.9	8.1	6.1	1.5	2.8	6.1	25.4	32.9	41.6	7.5
an	8.6	6.2	3.7	5.6	1.2	4.0	6.2	6.8	2.4	5.0	9.2	4.6	2.0	3.7	6.0	8.4	6.0	1.5	2.9	6.1	24.9	32.9	42.0	8
ecun	5.0	6.7	3.9	5.5	2.0	2.3	8.1	6.5	1.9	6.7	9.5	7.1	3.0	4.8	3.4	8.0	4.1	0.8	3.6	7.0	29.3	30.4	40.2	1.1

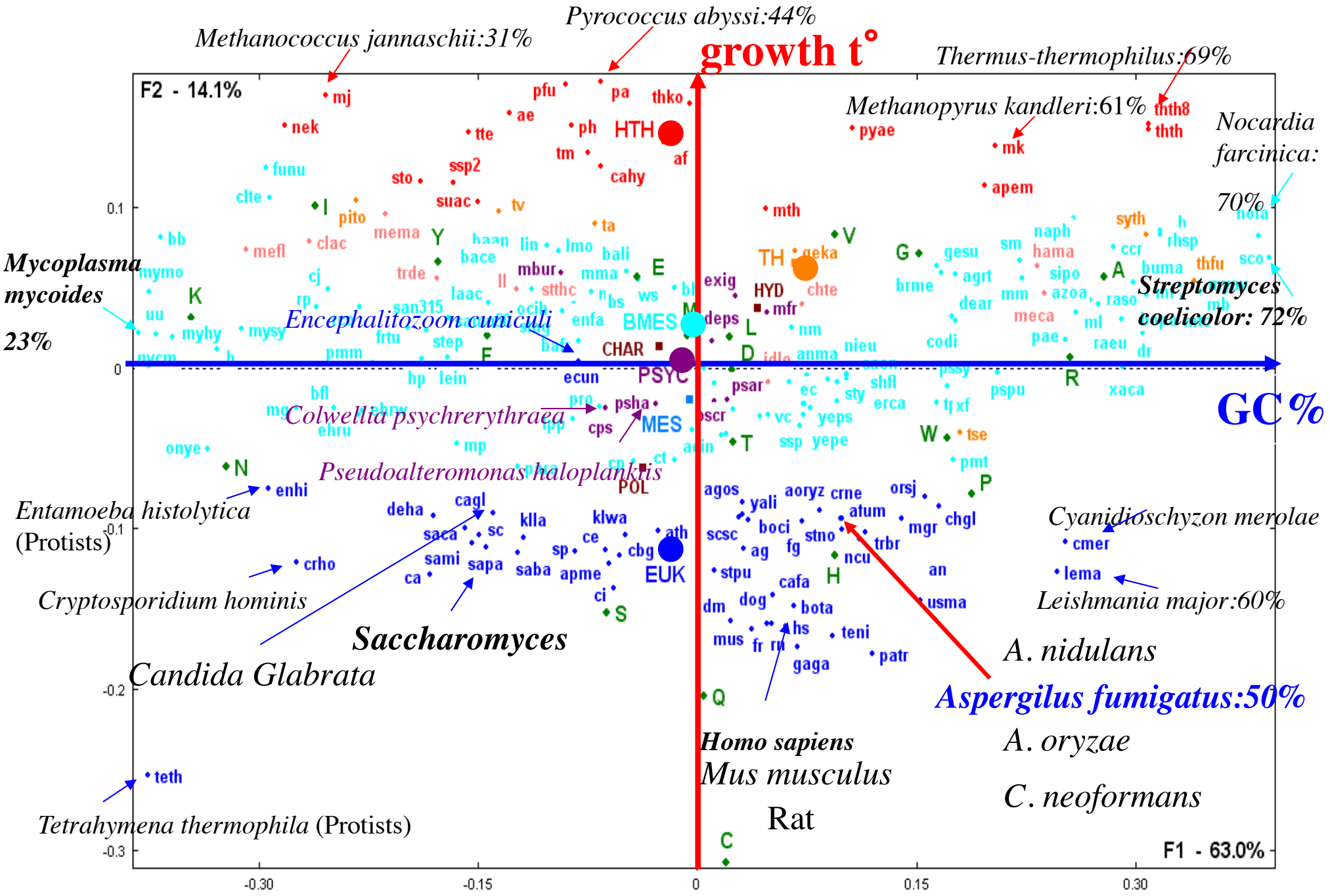
13

HTH	7.4	5.8	3.5	4.7	0.8	2.0	8.3	7.4	1.6	7.4	10.6	7.0	2.2	4.2	4.5	5.2	4.4	1.1	3.9	8.0	27.4	27.0	45.4	-0.4
TH	9.0	6.3	3.6	5.3	0.8	3.1	6.4	7.5	1.9	7.0	9.9	4.7	2.6	4.0	4.7	6.1	5.1	1.2	3.6	7.4	24.6	29.7	45.6	5.2
PSYC	8.4	4.6	4.3	5.7	1.1	4.0	6.3	6.9	2.2	7.2	9.9	5.5	2.7	4.1	3.9	6.5	5.8	1.1	3.2	6.9	24.2	31.8	44.0	7.6
BMES	8.6	5.1	4.4	5.4	1.0	3.8	6.3	7.0	2.1	6.9	10.2	5.8	2.3	4.3	4.1	6.2	5.4	1.1	3.2	6.9	24.6	30.9	44.4	6.3
EUK	6.9	5.4	4.9	5.4	1.7	4.2	6.6	6.0	2.4	5.6	9.3	6.1	2.2	4.0	5.2	8.4	5.6	1.2	3.1	6.0	25.9	33.8	40.2	7.9
SPEC	7.6	6.1	4.8	5.1	1.8	4.0	6.3	6.1	2.5	4.9	8.8	5.7	2.2	3.6	5.7	8.8	5.8	1.2	2.9	6.0	25.8	34.2	39.9	8.4
A	6.7	5.4	4.8	5.4	1.2	2.6	7.8	6.3	1.8	7.3	9.6	6.9	2.3	4.1	4.0	6.7	5.0	1.1	3.9	7.1	27.2	30.5	42.2	3.3
B	9.4	5.8	4.1	5.4	1.0	4.1	6.0	7.3	2.1	5.6	10.1	5.0	2.2	3.9	4.7	6.6	5.5	1.4	3.0	6.8	24.3	31.5	44.0	7.2
E	6.9	5.7	4.4	5.3	2.0	4.6	6.6	6.0	2.6	4.8	9.1	5.8	2.2	3.8	5.8	8.7	5.7	1.2	2.9	5.9	26.0	34.2	39.7	8.2
EA	6.8	5.7	4.5	5.5	1.8	4.1	6.8	5.8	2.4	5.7	9.6	6.5	2.3	4.0	4.6	7.6	5.5	1.1	3.2	6.5	26.8	32.4	40.7	5.6
EB	7.4	5.5	4.3	5.5	1.5	4.0	6.3	6.7	2.5	5.4	9.5	5.4	2.2	4.1	5.4	7.7	5.6	1.3	3.1	6.5	25.1	33.	41.8	7.9
AB	8.6	5.3	3.9	5.0	1.0	3.3	6.3	7.1	1.9	7.0	10.7	5.5	2.4	4.5	4.2	6.2	5.1	1.3	3.3	7.3	24.0	29.8	46.0	5.8
EAB	8.1	5.4	4.0	5.4	1.3	3.8	6.6	7.0	2.2	6.1	9.9	5.7	2.4	4.1	4.6	6.9	5.5	1.2	3.0	7.0	25.2	31.4	43.3	6.2

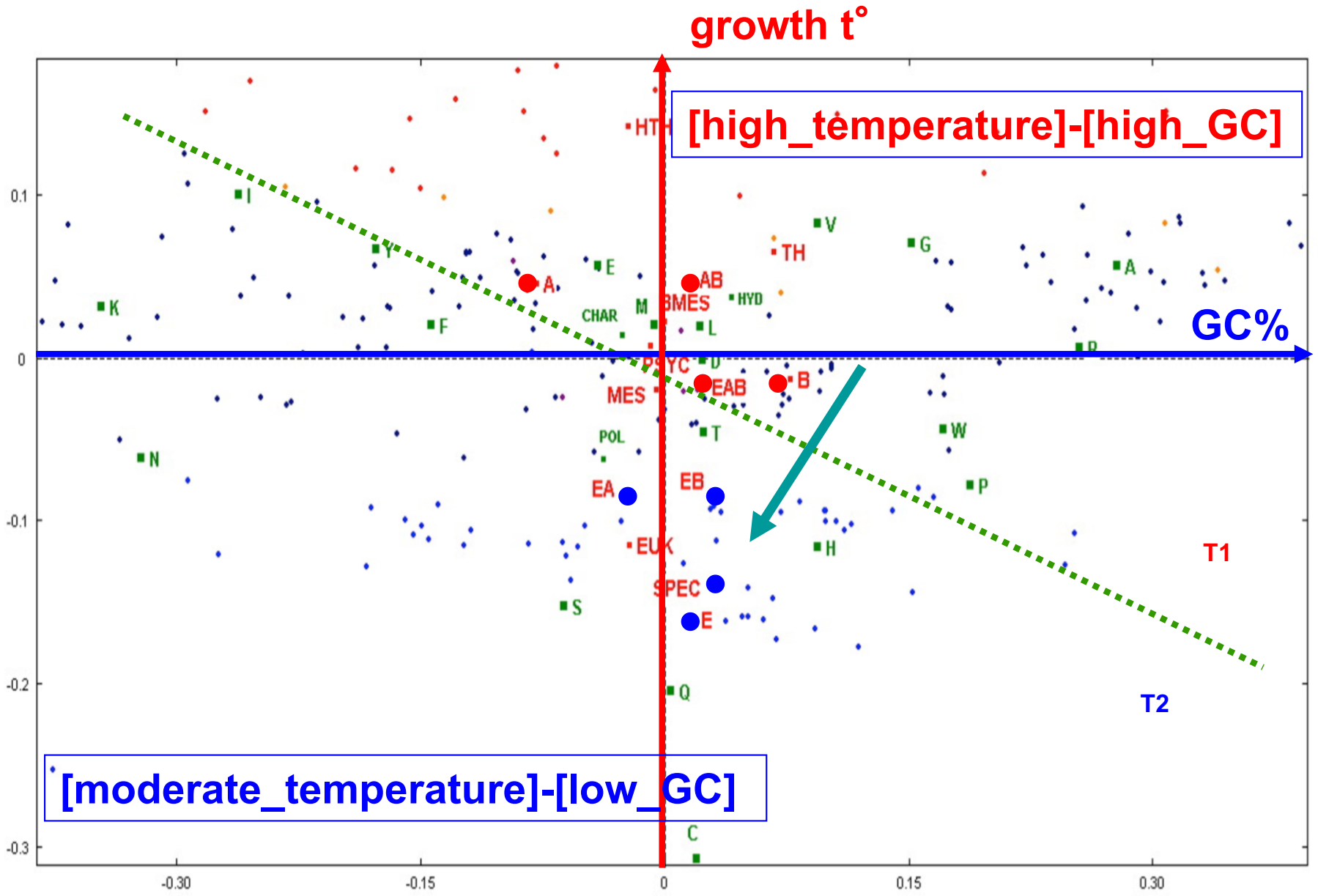
Correspondence Analysis was used to explore relationships between species and amino acids.



Tekaia, F. and Yeramian, E. (2006). *BMC Genomics*. 7:307



Tekaia, F. and Yeramian, E. (2006). *BMC Genomics*. 7:307


growth t°
[high_temperature]-[high_GC]
GC%
T1
T2
[moderate_temperature]-[low_GC]
HTH
AB
A
E
TH
V
G
A
I
Y
K
F
CHAR
M
BMES
HYD
L
D
PSYC
EAB
B
MES
T
W
N
POL
EA
EB
P
EUK
H
SPEC
S
E
Q
C
0.1
0
-0.1
-0.2
-0.3
-0.30
-0.15
0
0.15
0.30

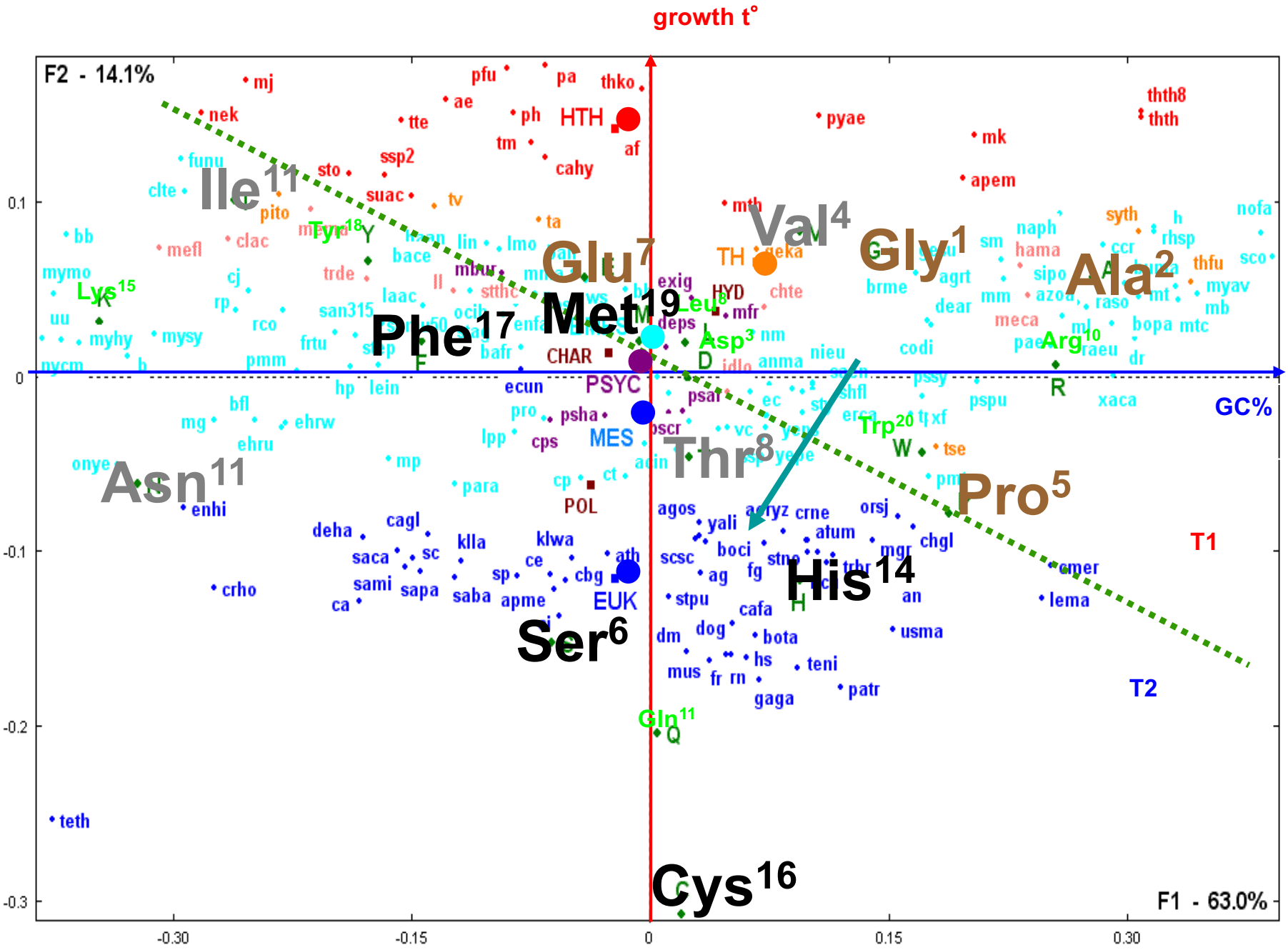
Trifonov, E.N. 2004. The triplet code from first principles. *J. Biomol. Struct. & Dyn.* **22**: 1-11.

• A consensus chronology of amino acids is built on the basis of 60 different criteria each offering certain temporal order.

• After several steps of filtering the chronology vectors are averaged resulting in the consensus order:

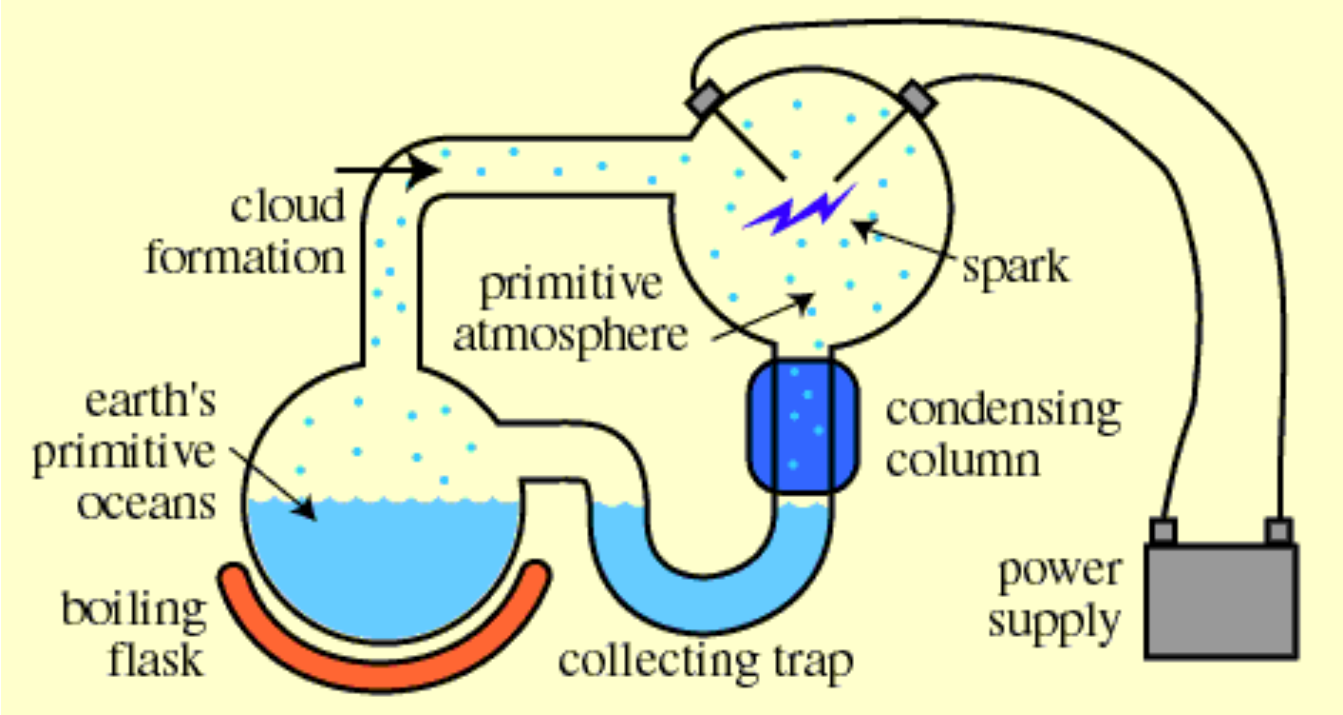
G[1] (Gly), A[2] (Ala), D[3] (Asp), V[4] (Val), P[5] (Pro), S[6] (Ser), E[7] (Glu), (L[8] (Leu), T[8] (Thr)), R[10] (Arg), (I[11] (Ile), Q[11] (Gln), N[11] (Asn)), H[14] (His), K[15] (Lys), C[16] (Cys), F[17] (Phe), Y[18] (Tyr), M[19] (Met), W[20] (Trp).

• It reveals two important features: the amino acids synthesized in imitation experiments of S. Miller appeared first, while the amino acids associated with codon capture events came last.



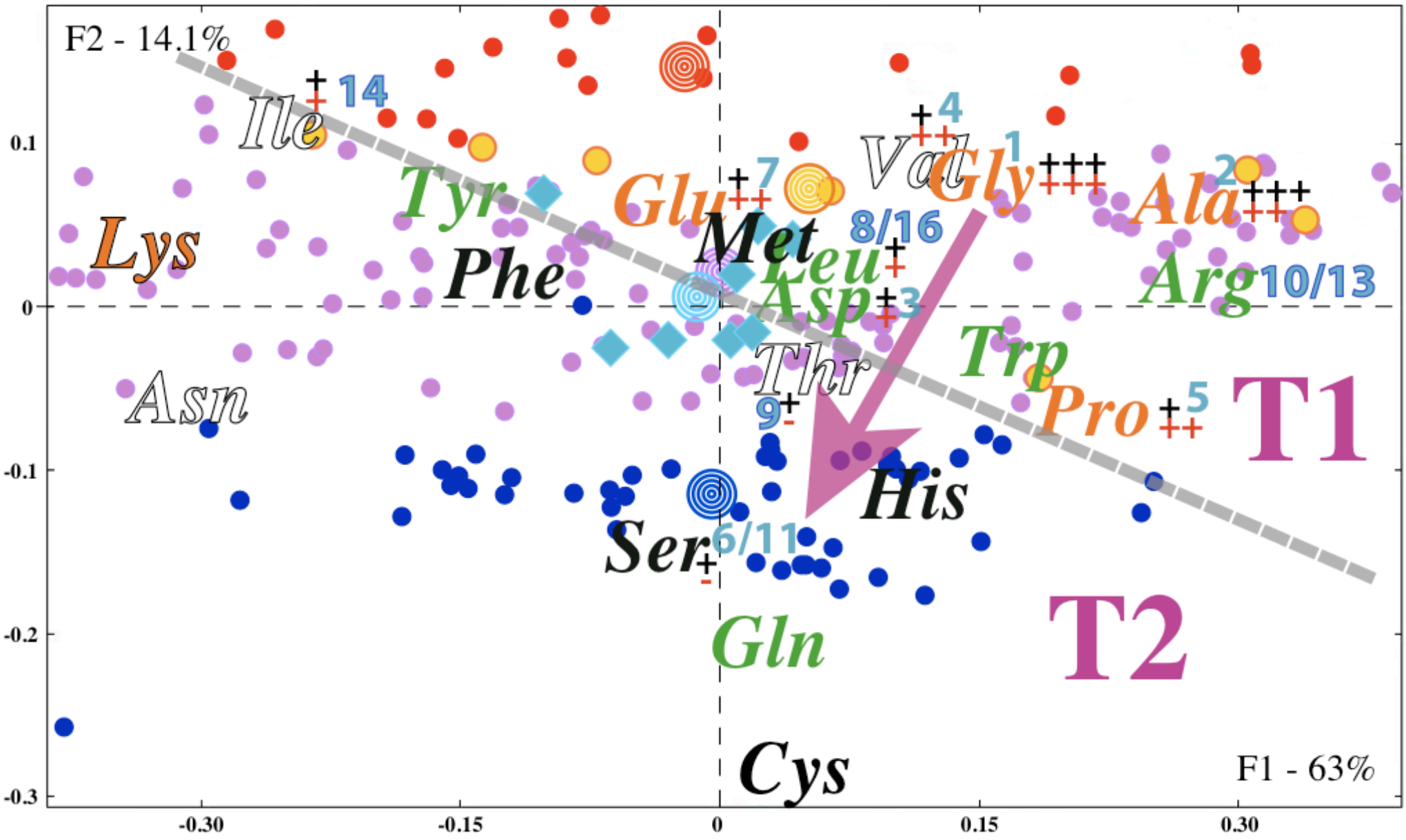
Trifonov, E.N. 2004. The triplet code from first principles. *J. Biomol. Struct. & Dyn.* **22**: 1-11.

Miller/Urey Experiment: 1953



• By the 1950s, scientists were in hot pursuit of the origin of life. Around the world, the scientific community was examining what kind of environment would be needed to allow life to begin.
• In 1953, Stanley L. Miller and Harold C. Urey, working at the University of Chicago, conducted an experiment which would change the approach of scientific investigation into the origin of life.
•Miller took molecules which were believed to represent the major components of the early Earth's atmosphere and put them into a closed system

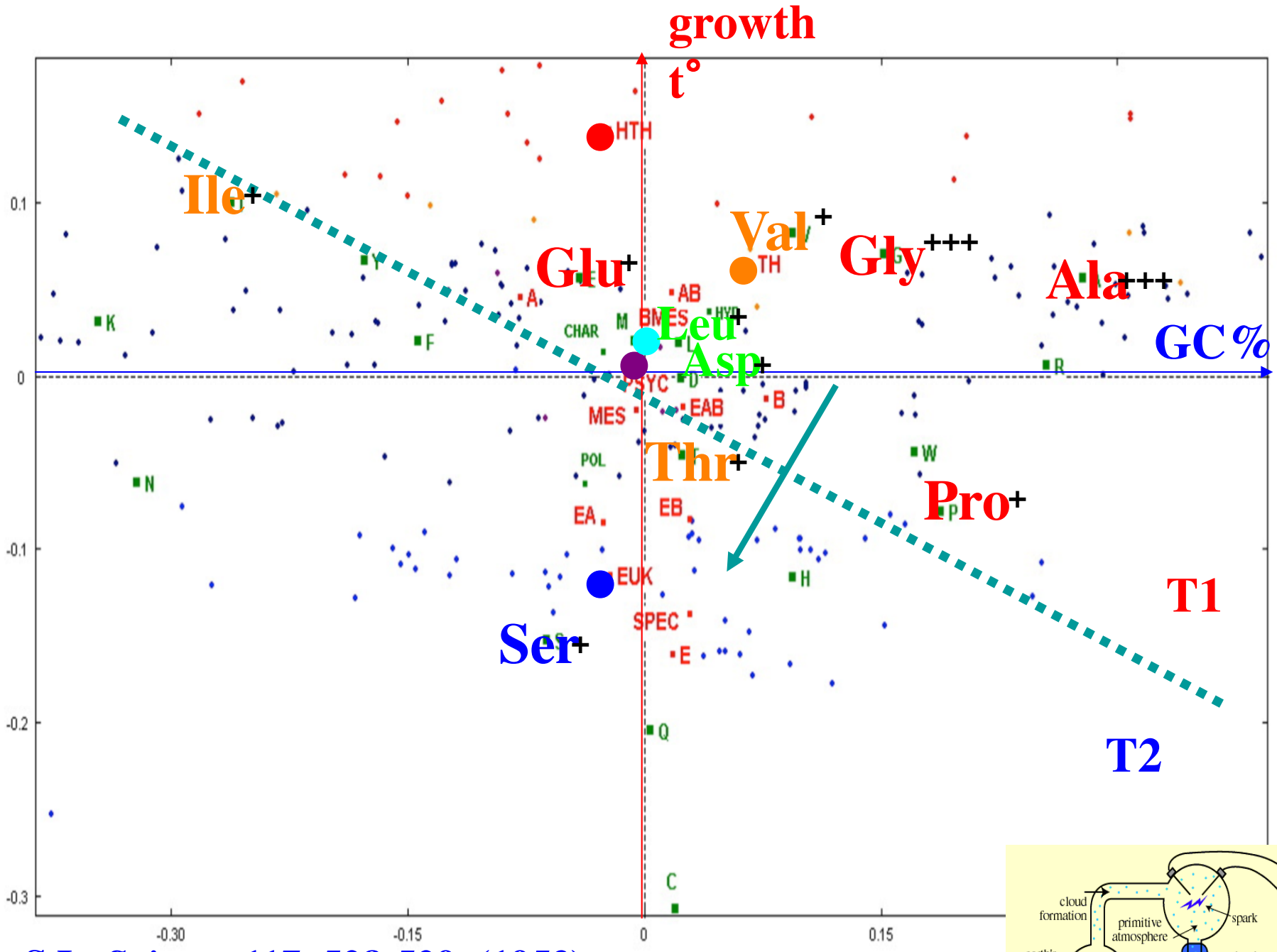
• Miller's experiment showed that organic compounds such as amino acids, which are essential to cellular life, could be made easily under the conditions that scientists believed to be present on the early earth.

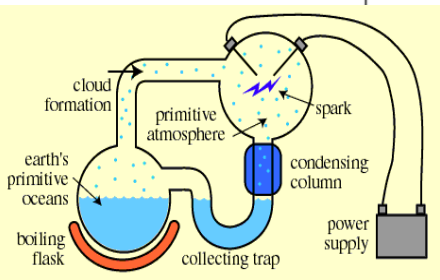


Tekaia, F. and Yeramian, E. (2006). Evolution of Proteomes: Fundamental signatures and global trends in amino acid composition. *BMC Genomics*. 7:307.

Miller, S.L. *Science* 117, 528-529. (1953)

Cronin, J.R. and Pizzarello, S. (1983).





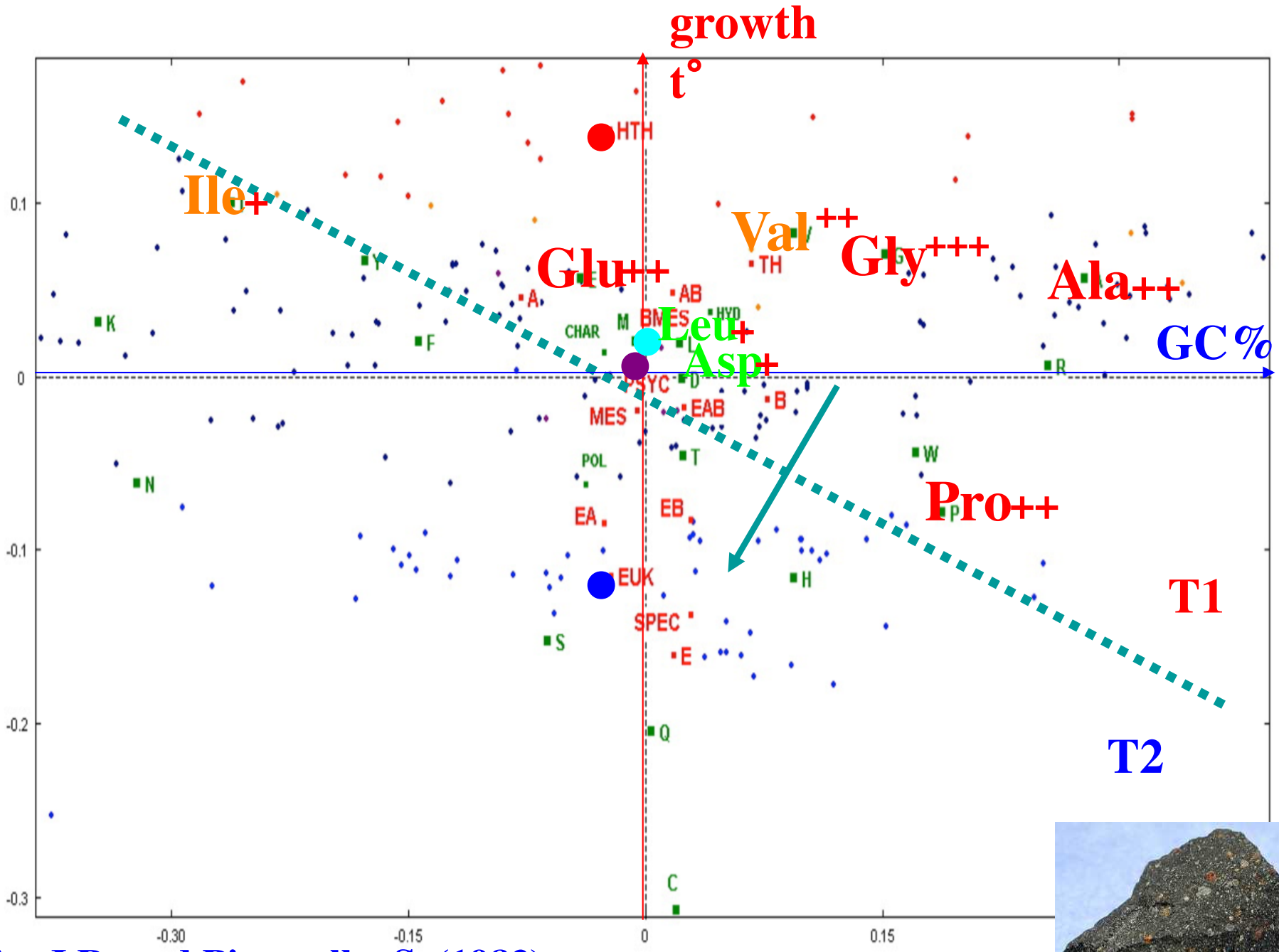
Miller, S.L. *Science* 117, 528-529. (1953)

Production of aa under possible primitive earth conditions.

Murchison meteorite

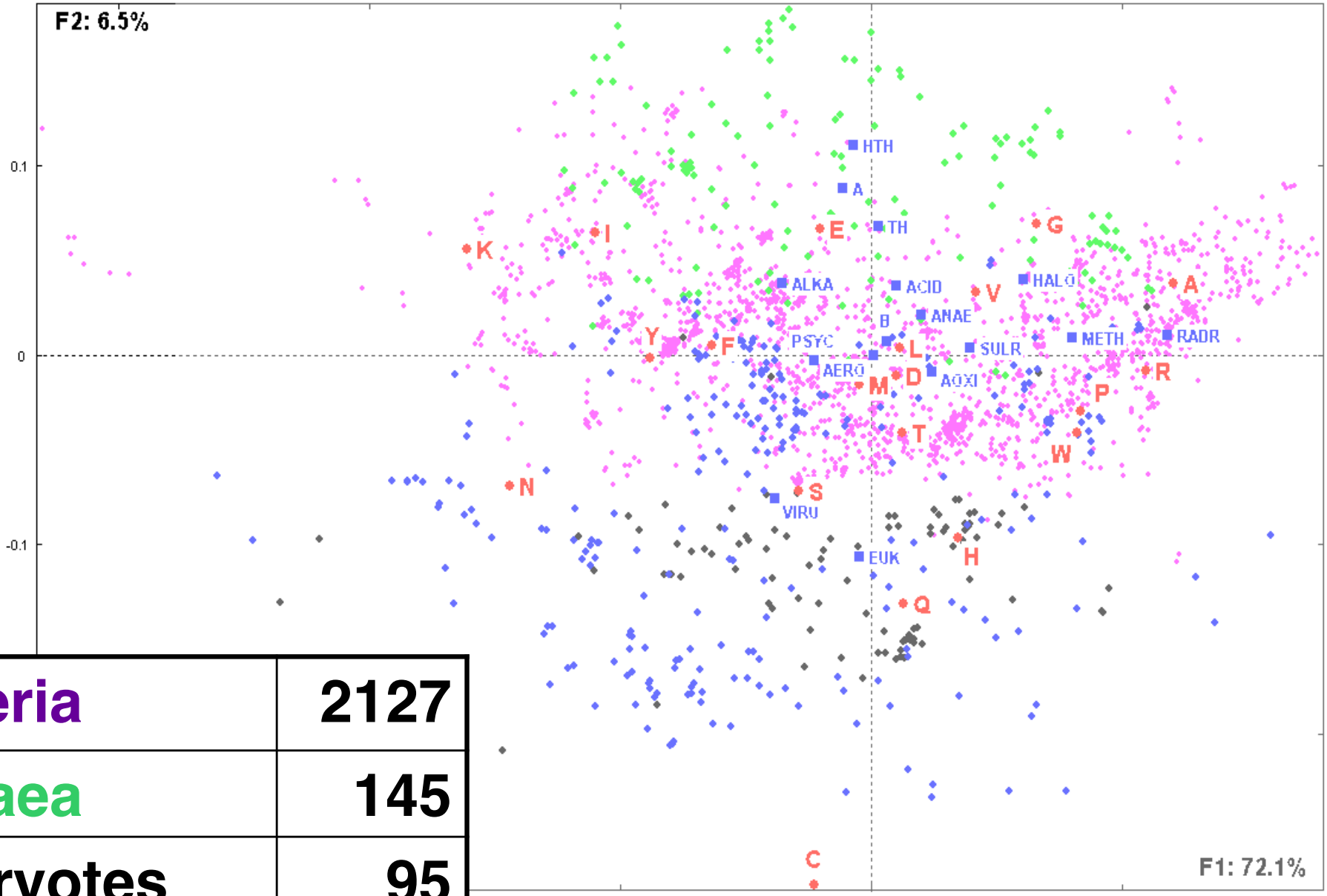
The Murchison meteorite fall occurred on September 28, 1969 over Murchison, Australia. Over 100 kilograms of this meteorite have been found. This meteorite is of possible cometary origin due to its high water content of 12%.

An abundance of amino acids found within this meteorite has led to intense study by researchers as to its origins. More than 92 different amino acids have been identified within the Murchison meteorite to date. Nineteen of these are found on Earth. The remaining amino acids have no apparent terrestrial source.



Cronin, J.R. and Pizzarello, S. (1983).
Amino acids in meteorites. *Adv Space Res.* 3: 5-18.

Murchison meteorite 28-09-1969



Bacteria	**2127**
Archaea	**145**
Eukaryotes	**95**
Giant Viruses	**273**

Example:

43 yeast / 48 fungal species versus aa compositions

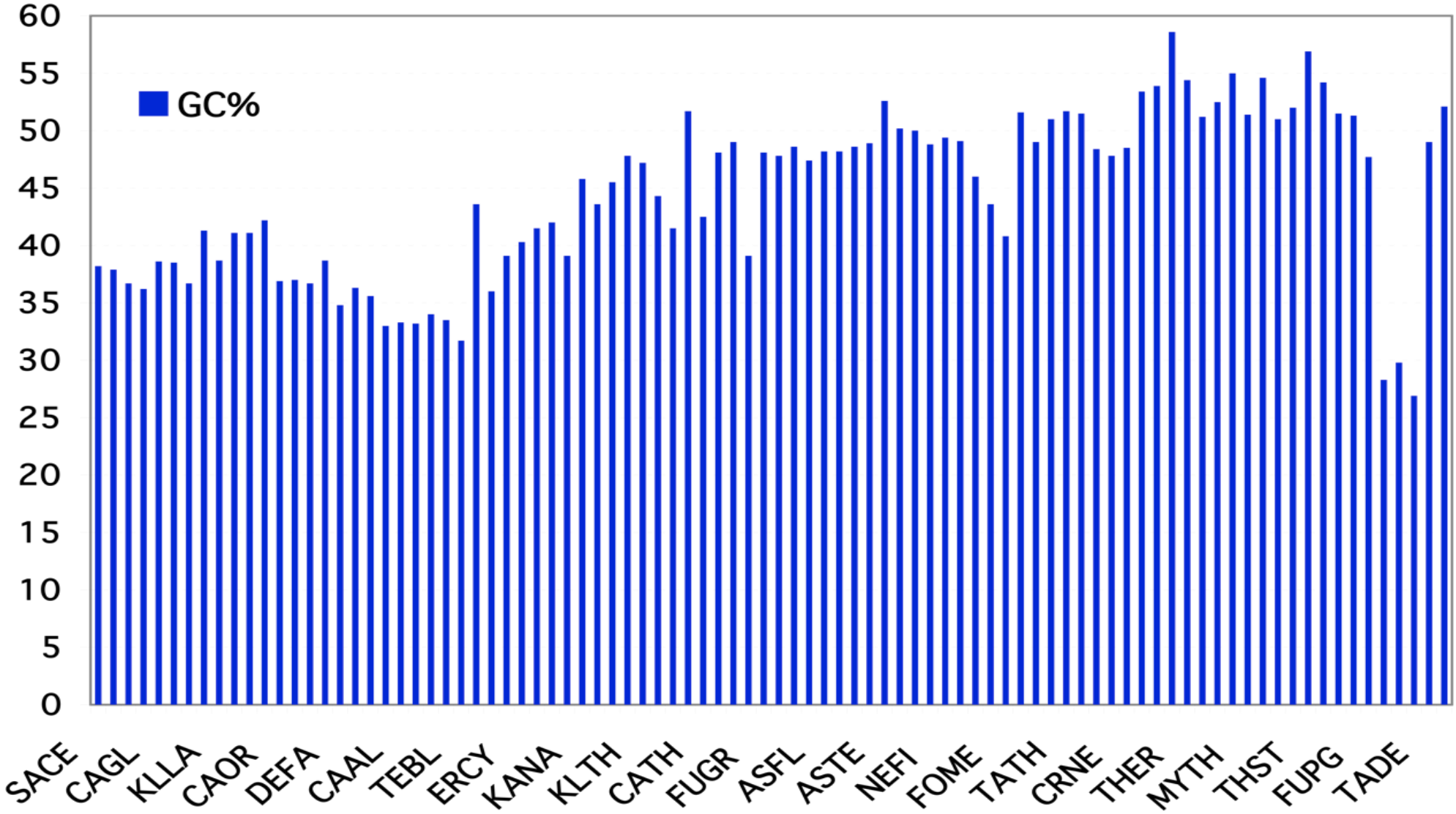
IDENT	#Prots	Size (Mb)	GC%	Species
SACE	5769	12.2	38.2	Saccharomyces_cerevisiae [1]
SAAR	5527	11.6195	37.9	Saccharomyces_arboricola [2]
NACA	5592	11.2195	36.7	Naumovozyma_castellii_CBS_4309 [3]
KAZA	5378	11.13	36.2	Kazachstania_africana_CBS_2517 [3]
CAGL	5204	12.3182	38.6	Candida_glabrata [4]
NADE	5112	10.9691	38.5	Nakaseomyces_delphensis [5]
DECA	6219	11.76	36.7	Debaryomyces_carsonii [6]
PISO	11175	21.4596	41.3	Pichia_sorbitophila [7]
KLLA	5083	10.6891	38.7	Kluyveromyces_Lactis [4]
PIPA	5040	9.2163	41.1	Pichia_pastoris_GS115 [8]
PIST	5816	15.4411	41.1	Pichia_Stipidis [9]
CATE	6985	10.75	42.2	Candida_tenuis [10]
CAOR	5677	12.6594	36.9	Candida_orthopsilosis [11]
SPPA	5983	13.1821	37.0	Spathaspora_passalidarum_NRRL_Y-27907 [10]
LOEL	5796	15.4	36.7	Lodderomyces_elongisporus [12]
CAPA	5817	12.9984	38.7	Candida_parapsilosis [12]
DEFA	6182	12.00	34.8	Debaryomyces_fabryi [6]
DEHA	6272	12.2	36.3	Debaryomyces_hansenii [4]
DETY	6747	12.40	35.6	Debaryomyces_tyrocola [6]
CATR	6258	14.5798	33.0	Candida_Tropicalis [12]
CAAL	6112	14.4176	33.3	Candida_albicans_WO-1 [13]
CADU	5983	14.6184	33.2	Candida_dubliniensis_CD36_uid38659 [14]
STAM	5790	0	0	Starmera_amethionina [6]
NADA	5772	13.5275	34.0	Naumovozyma_dairenensis_CBS_421 [3]
TEPH	5250	12.1	33.5	Tetrapisispora_phaffii_CBS_4417 [3]
TEBL	5388	14.0486	31.7	Tetrapisispora_blattae [3]
CAGU	5920	10.61	43.6	Candida_guilliermondii [12]
SCPO	5142	12.6	36.0	Schizosaccharomyces_pombe [15]
DEBR	5255	13.0582	39.1	Dekkera_Bruxellensis_STO5_12_22 [16]
ERCY	4434	9.6694	40.3	Eremothecium_cymbalariae_DBVPG_7215 [17]
SAKL	5306	11.3458	41.5	Saccharomyces_kluyveri [18]
KLTH	5103	10.3928	47.2	Kluyveromyces_thermotolerans [18]
CALU	5936	12.1148	44.3	Candida_lusitaniae [12]
SCJA	5167	11.7332	41.5	Schizosaccharomyces_japonicus_yfs275_5 [22]
ERGO	4718	9.0957	51.7	Eremothecium_gossypii_(AGOS) [23]
CATH	11703	28.1975	42.5	Calcariporiella_thermophila [56]
ARAD	6152	11.8046	48.1	Arxula_adeninivorans [24]
YALI	6434	20.5029	49.0	Yarrowia_lipolytica [4]
BOCI	16389	42.6630	39.1	Botrytis_cinerea [25]
FUGR	13321	36.3130	48.1	Fusarium_graminearum [26]
GIZE	11578	36.2585	47.8	Gibberella_zeae_PH-1_uid243 [26]
FUVE	14195	41.1043	48.6	Fusarium_verticillioides [26]
FUOX	17608	57.7206	47.4	Fusarium_oxysporum [27]
ASFL	12587	36.7902	48.2	Aspergillus_flavus [28]
ASOR	12063	37.0886	48.2	Aspergillus_oryzae [29]
PECH	11396	31.3410	48.6	Penicillium_chrysogenum [30]
PERU	12790	32.2237	48.9	Penicillium_rubens [31]
ASTE	10406	29.3312	52.6	Aspergillus_terreus [32]
ASNG	8592	34.0066	50.2	Aspergillus_niger [33]
ASNI	9410	29.7113	50.0	Aspergillus_nidulans [34]
ASFU	9630	29.3849	48.8	Aspergillus_fumigatus [35]
NEFI	10407	32.5517	49.4	Neosartorya_fischeri [36]
ASCL	9120	27.8594	49.1	Aspergillus_clavatus [36]
COIM	9910	28.9479	46.0	Coccidioides_immitis_RS [37]
PABR	8390	29.9525	43.6	Paracoccidioides_brasiliensis [38]
FOME	11338	63.3544	40.8	Fomitiporia_mediterranea_MF3-22 [39]
COCI	13544	36.2944	51.6	Coprinus_cinereus [40]
THAU	10450	31.4823	49.0	Thermoascus_aurantiacus [56]
THLA	8133	19.9438	51.0	Thermomyces_lanuginosus [56]
TATH	7920	19.8875	51.7	Talaromyces_thermophilus [56]
POAN	10219	33.7760	51.5	Podospora_anserina_S_mat+ [41]
NECR	9822	40.4631	48.4	Neurospora_crassa [42]

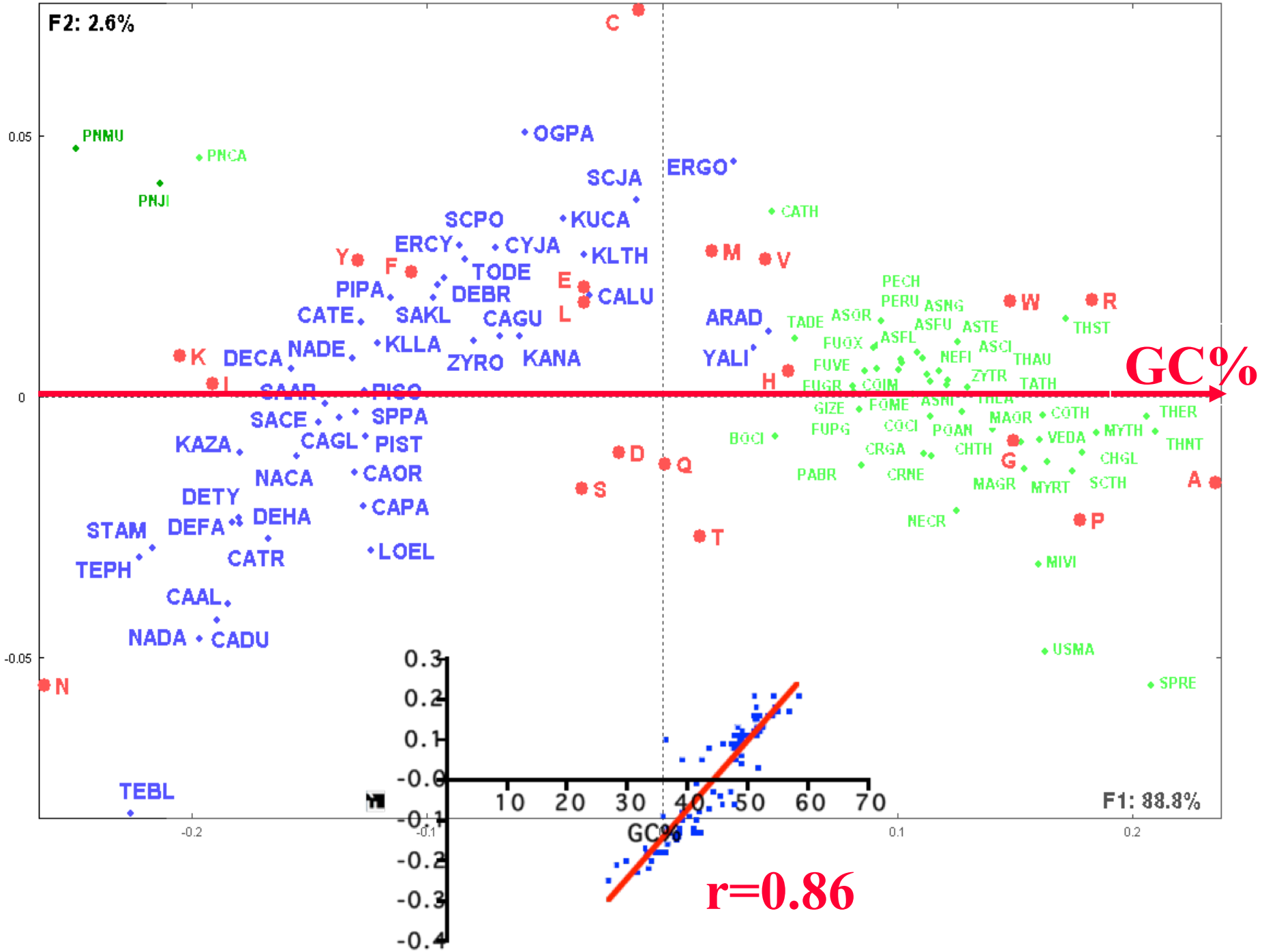
43 yeast and 48 fungal species

91 species vs 20 aa

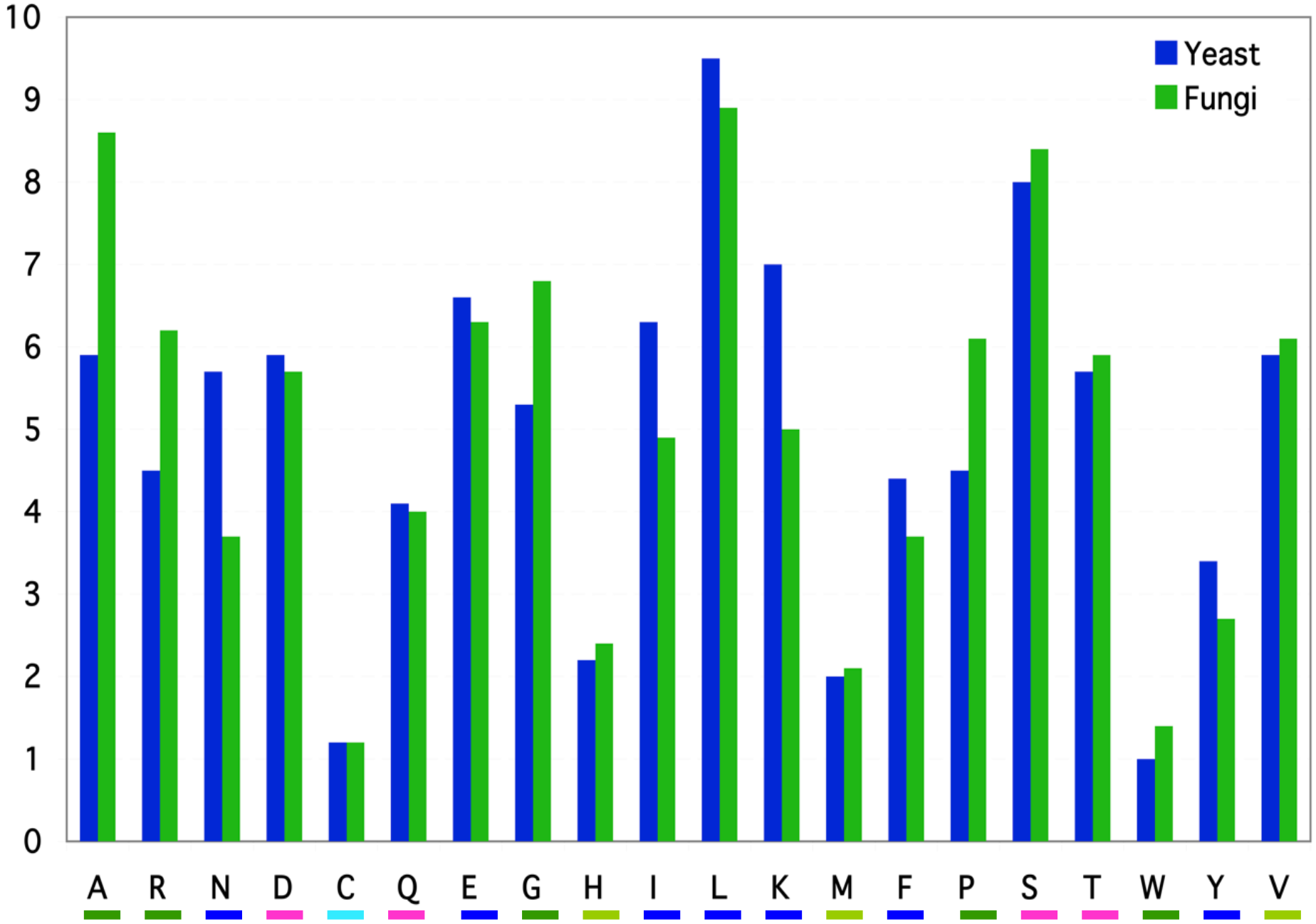
THER	9815	36.9196	54.4	Thielavia_terrestris [47]
THNT	9204	40.6623	51.2	Thielavia_antarctica [57]
CHTH	8280	28.3147	52.5	Chaetomium_thermophilum_ATTC1651 [48]
SCTH	10945	29.3248	55.0	Scytalidium_thermophilum [56]
MYTH	9099	38.7442	51.4	Myceliophthora_thermophila_ATCC_42464 [47]
CHGL	11124	34.8869	54.6	Chaetomium_globosum [58]
COTH	10644	33.3614	51.0	Corynascus_thermophilus [56]
MYRT	8635	31.6872	52.0	Myriococcum_thermophilum [56]
THST	10387	29.5796	56.9	Thermomyces_stellatus [56]
VEDA	10535	33.9000	54.2	Verticillium_alfalfae [49]
MAOR	12755	41.0278	51.5	Magnaporthe_oryzae [50]
MAGR	11054	41.6955	51.3	Magnaporthe_grisea [51]
FUPG	12447	36.9329	47.7	Fusarium_pseudograminearum_CS3096 [52]
PNJI	3520	8.1799	28.3	Pneumocystis_jirovecii [53]
PNCA	6874	6.3	29.8	Pneumocystis_carinii [59]
PNMU	3838	7.4514	26.9	Pneumocystsis_murina [60]
TADE	4663	13.7735	49.0	Taphrina_deformans_JCM_22205 [54]
ZYTR	10931	39.6863	52.1	Zymoseptoria_tritici [55]

GC%	Number	Mean	Std	Min	Max
Yeast	42	39.7	4.9	31.7	51.7
Fungi	48	48.6	6.4	26.9	58.6

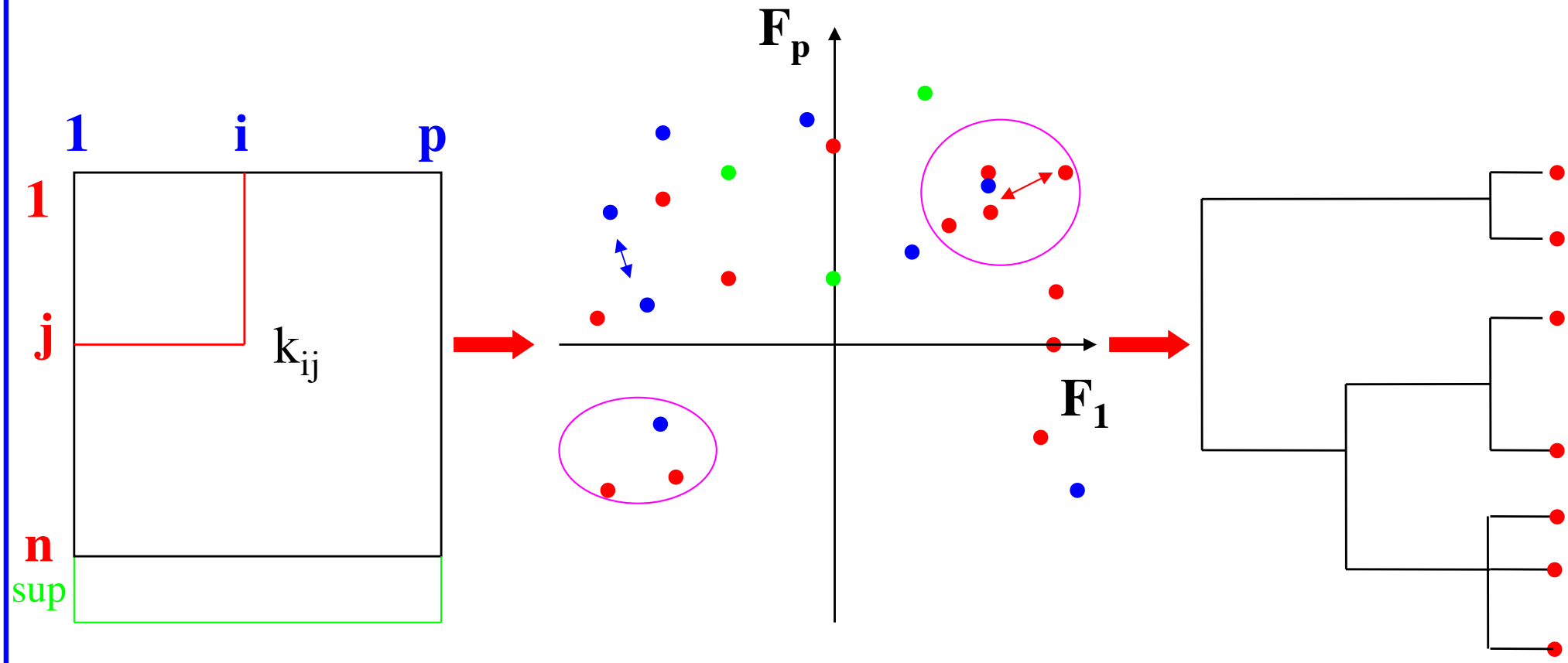



F2: 2.6%
F1: 33.8%
GC%
r=0.86
C
OGPA
PNMU
PNCA
PNJI
ERGO
SCJA
CATH
SCPO
KUCA
ERCY
CYJA
KLTH
M
V
Y
F
TODE
E
PIPA
DEBR
CALU
CATE
SAKL
CAGU
L
ARAD
TADE
ASOR
PECH
PERU
ASNG
W
R
ASFU
ASTE
THST
FUGX
ASFL
ASCI
K
DECA
NADE
KLLA
ZYRO
KANA
YALI
NEFI
THAU
FUVE
ZYTR
H
I
SAAR
PISC
FUGR
COIN
TATH
GIZE
FOME
ASHI
THLA
MAGR
COTH
THER
SACE
SPPA
FUPG
COCI
POAN
BOCI
VEDA
MYTH
THNT
KAZA
CAGL
PIST
CRGA
CHTH
G
CHGL
D
Q
A
NACA
CAOR
PABR
CRNE
MAGR
MYRT
SCTH
S
DETY
CAPA
DEHA
STAM
DEFA
NECR
P
T
CATR
LOEL
TEPH
MIVI
CAAL
CADU
NADA
USMA
SPRE
N
TEBL
0.05
0
-0.05
-0.2
-0.1
0.1
0.2
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.4
10
20
30
40
50
60
70

Distribution of amino acids in Yeast and Fungi



Methodology



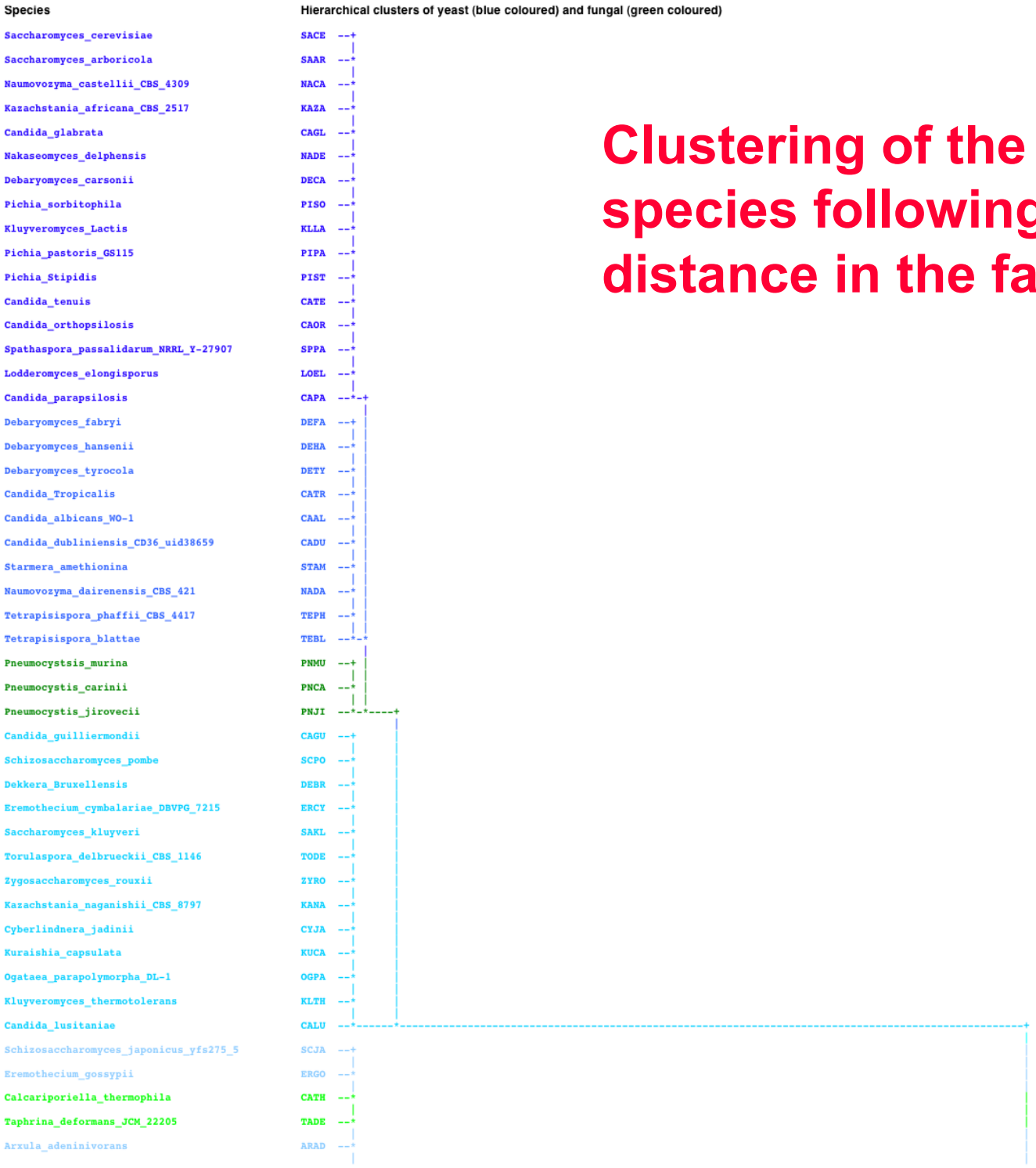
Matrice T

$k_{ij} > 0$

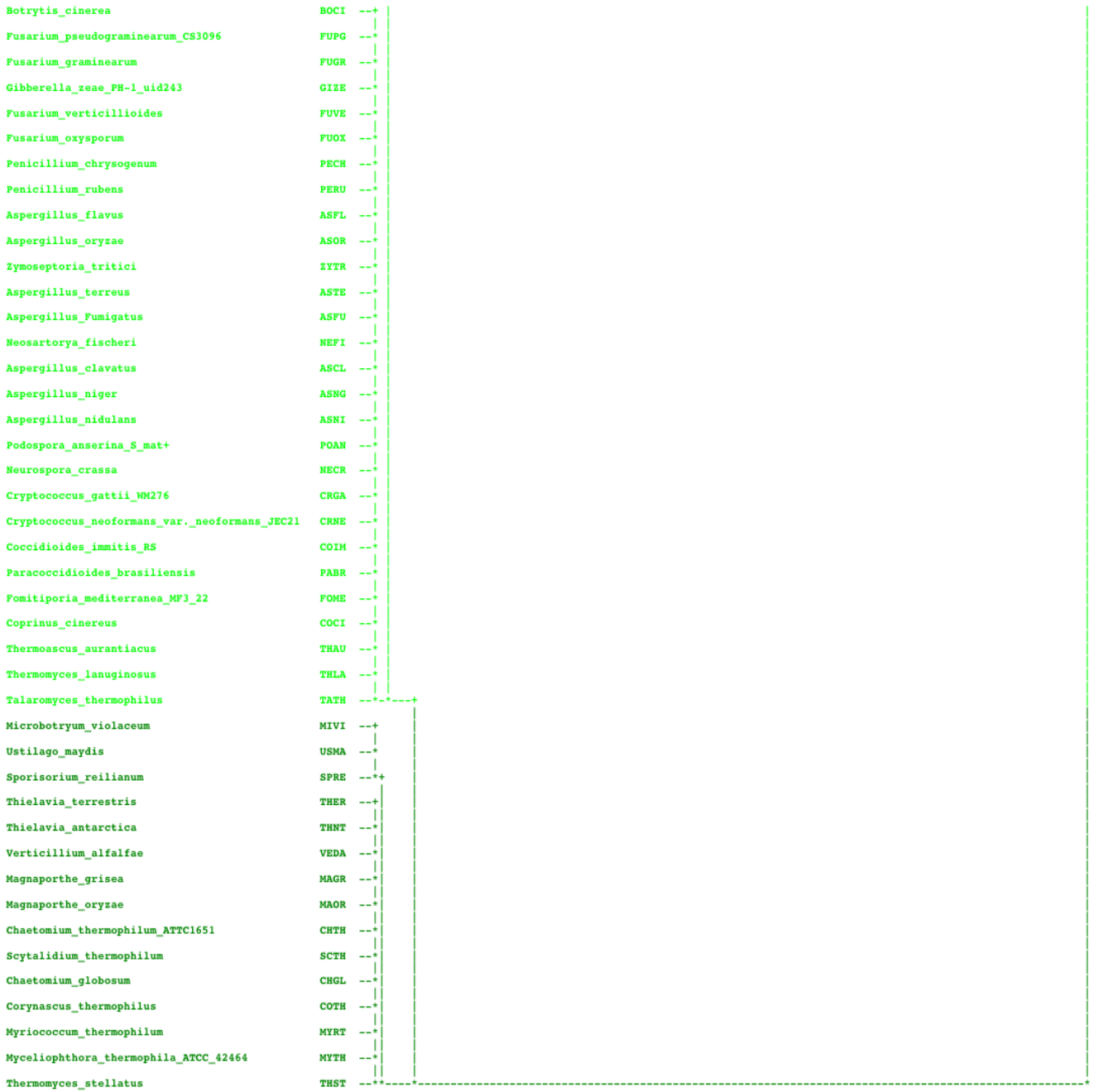
Correspondence Analysis

Classification

- orthogonal system;
- use of euclidean distance;

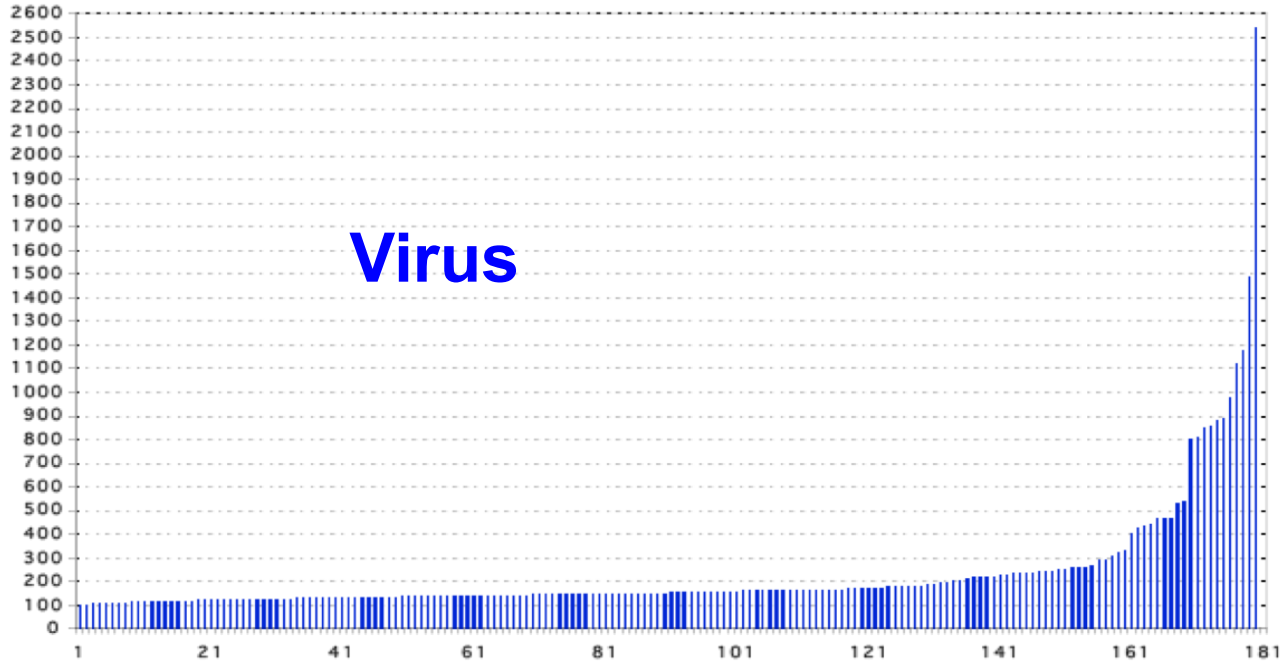


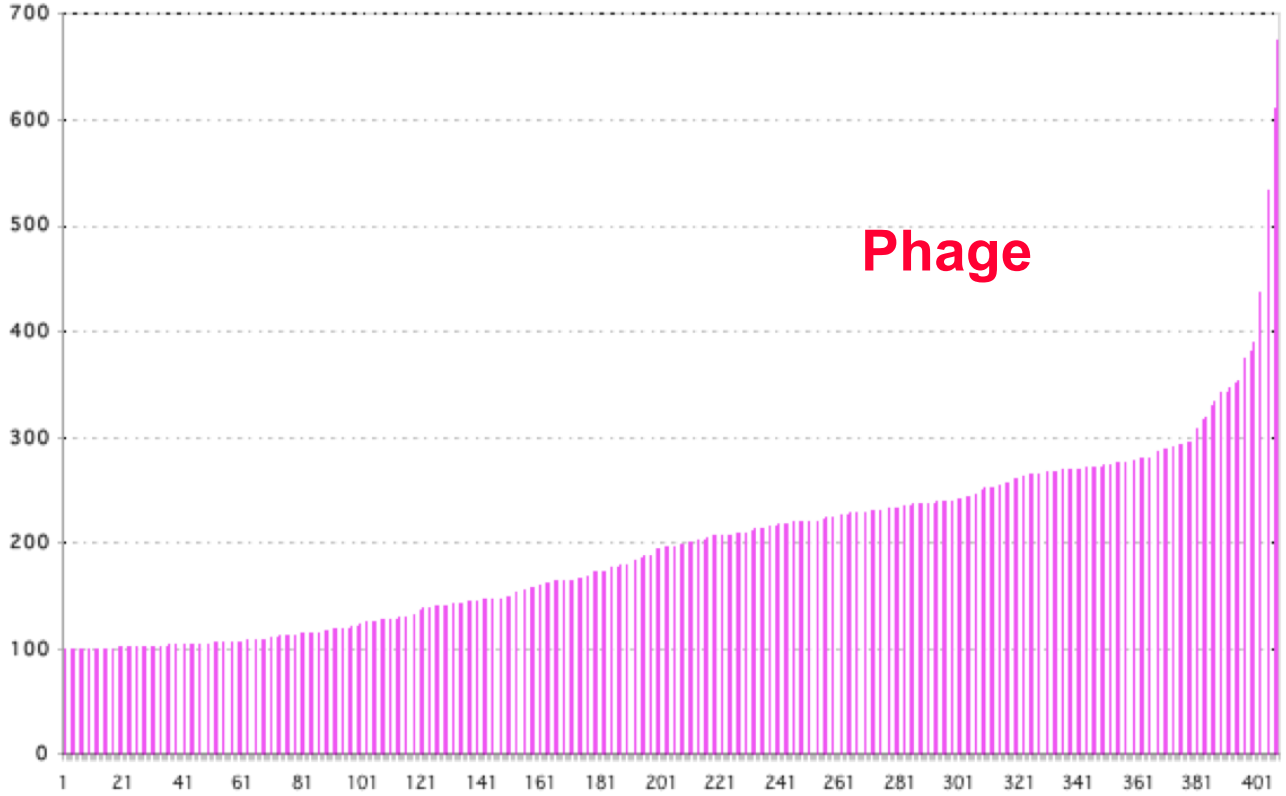
Clustering of the yeast and fungal species following their Euclien distance in the factorial space.


Botrytis_cinerea BOCI
Fusarium_pseudograminearum_CS3096 FUPG
Fusarium_graminearum FUGR
Gibberella_zeae_PH-1_uid243 GIZE
Fusarium_verticillioides FUVE
Fusarium_oxysporum FUOX
Penicillium_chrysogenum PECH
Penicillium_rubens PERU
Aspergillus_flavus ASFL
Aspergillus_oryzae ASOR
Zymoseptoria_tritici ZYTR
Aspergillus_terreus ASTE
Aspergillus_Fumigatus ASFU
Neosartorya_fischeri NEFI
Aspergillus_clavatus ASCL
Aspergillus_niger ASNG
Aspergillus_nidulans ASNI
Podospora_anserina_S_mat+ POAN
Neurospora_crassa NECR
Cryptococcus_gattii_WM276 CRGA
Cryptococcus_neoformans_var._neoformans_JEC21 CRNE
Coccidioides_immitis_RS COIM
Paracoccidioides_brasiliensis PABR
Fomitiporia_mediterranea_MF3_22 FOME
Coprinus_cinereus COCI
Thermoascus_aurantiacus THAU
Thermomyces_lanuginosus THLA
Talaromyces_thermophilus TATH
Microbotryum_violaceum MIVI
Ustilago_maydis USMA
Sporisorium_reilianum SPRE
Thielavia_terrestris THER
Thielavia_antarctica THNT
Verticillium_alfalfae VEDA
Magnaporthe_grisea MAGR
Magnaporthe_oryzae MAOR
Chaetomium_thermophilum_ATTC1651 CHTH
Scytalidium_thermophilum SCTH
Chaetomium_globosum CHGL
Corynascus_thermophilus COTH
Myriococcum_thermophilum MYRT
Myceliophthora_thermophila_ATCC_42464 MYTH
Thermomyces_stellatus THST

Example:

Large Viruses & phages versus aa compositions





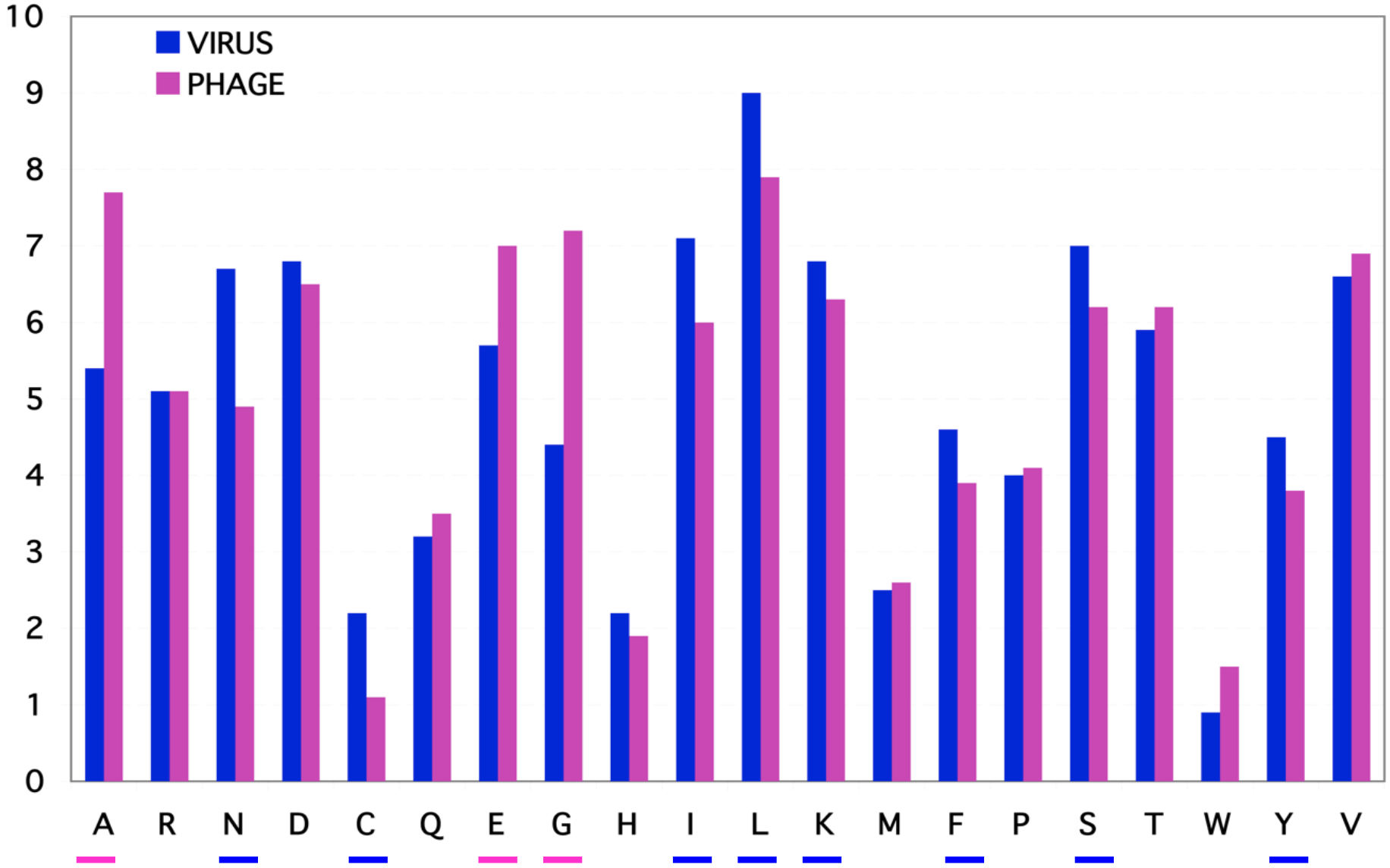
- **181 viruses**
- **407 phages**

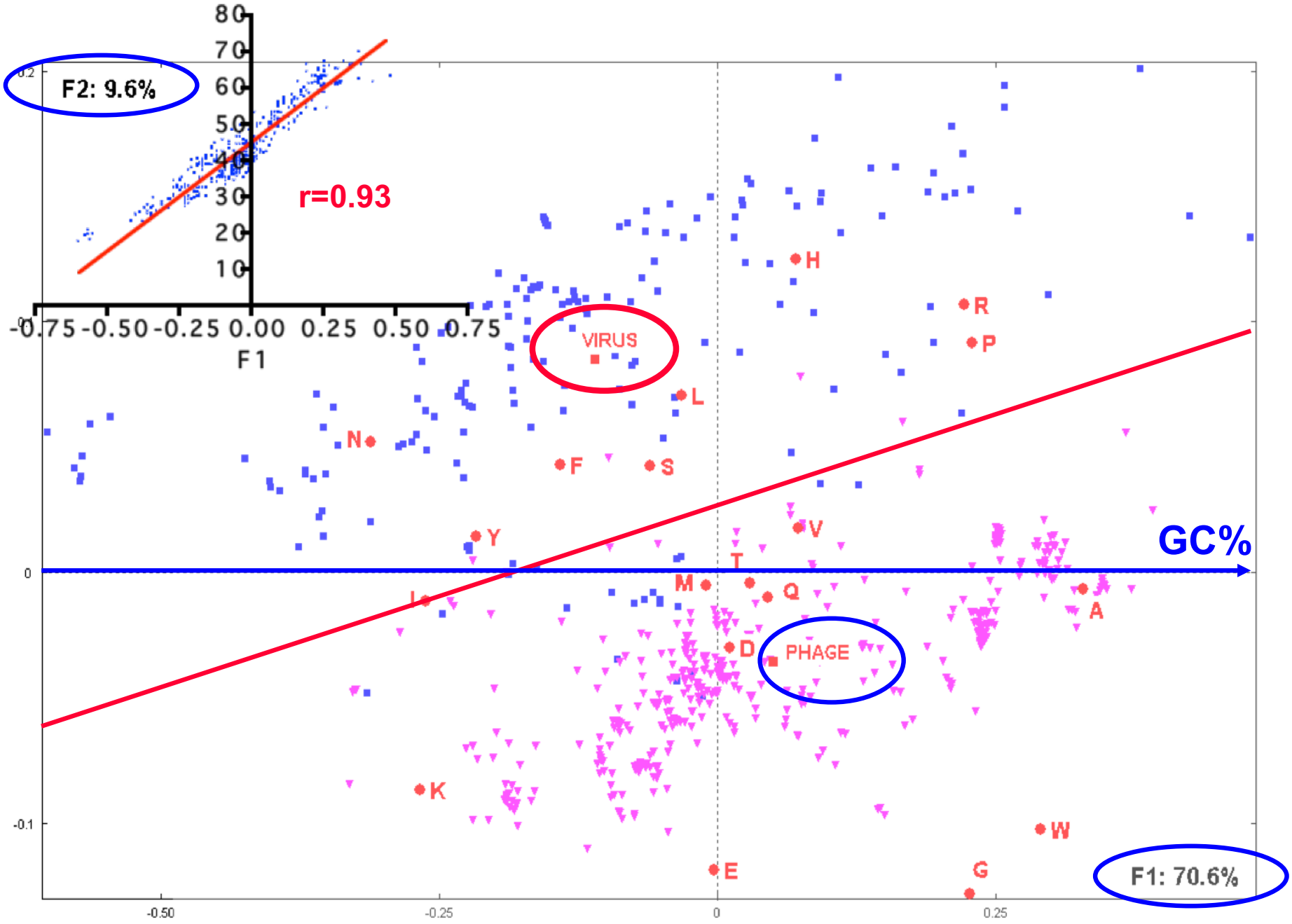
including more than 100 ORF products

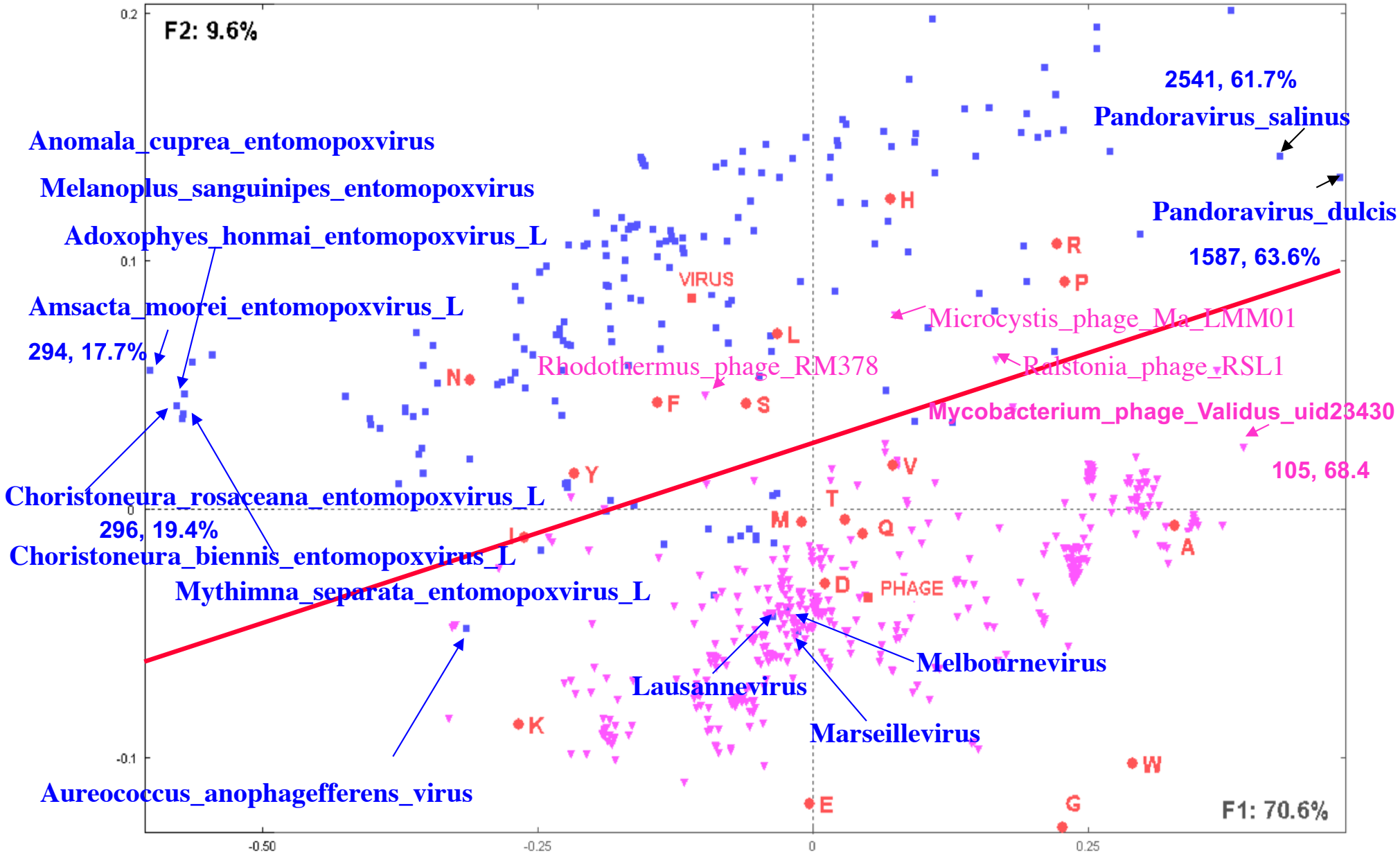
Distribution of the number of proteins in Viruses and Phages

Table: 588 vs 20 aa composition

Spec	Number	Mean GC%	STD	Min GC%	Max GC%
Virus	181	40.8	10.8	17.7	66.6
Phages	407	46.9	11.7	26.0	70.0




F2: 9.6%
r=0.93
F1
VIRUS
PHAGE
GC%
F1: 70.6%
H
R
P
L
N
F
S
Y
V
T
M
Q
A
I
D
K
W
E
G


F2: 9.6%
F1: 70.6%
0.2
0.1
0
-0.1
-0.50
-0.25
0
0.25
2541, 61.7%
Pandoravirus_salinus
Pandoravirus_dulcis
1587, 63.6%
Anomala_cuprea_entomopoxvirus
Melanoplus_sanguinipes_entomopoxvirus
Adoxophyes_honmai_entomopoxvirus_L
Amsacta_moorei_entomopoxvirus_L
294, 17.7%
Choristoneura_rosaceana_entomopoxvirus_L
296, 19.4%
Choristoneura_biennis_entomopoxvirus_L
Mythimna_separata_entomopoxvirus_L
Aureococcus_anophagefferens_virus
VIRUS
PHAGE
Microcystis_phage_Ma_LMM01
Rhodothermus_phage_RM378
Ralstonia_phage_RSL1
Mycobacterium_phage_Validus_uid23430
105, 68.4
Lausannevirus
Melbournevirus
Marseillevirus
H
R
P
L
N
F
S
Y
V
T
M
Q
A
D
K
W
E
G

Example

91 Yeast/fungal species according to their shared orthologs

Tekaia F, Lazcano A, Dujon B. The genomic tree as revealed from whole proteome comparisons.*Genome Res.* 1999 Jun;9(6):550-7.

Procedure

- **Pair-wise blastp comparisons of the 91 proteomes**

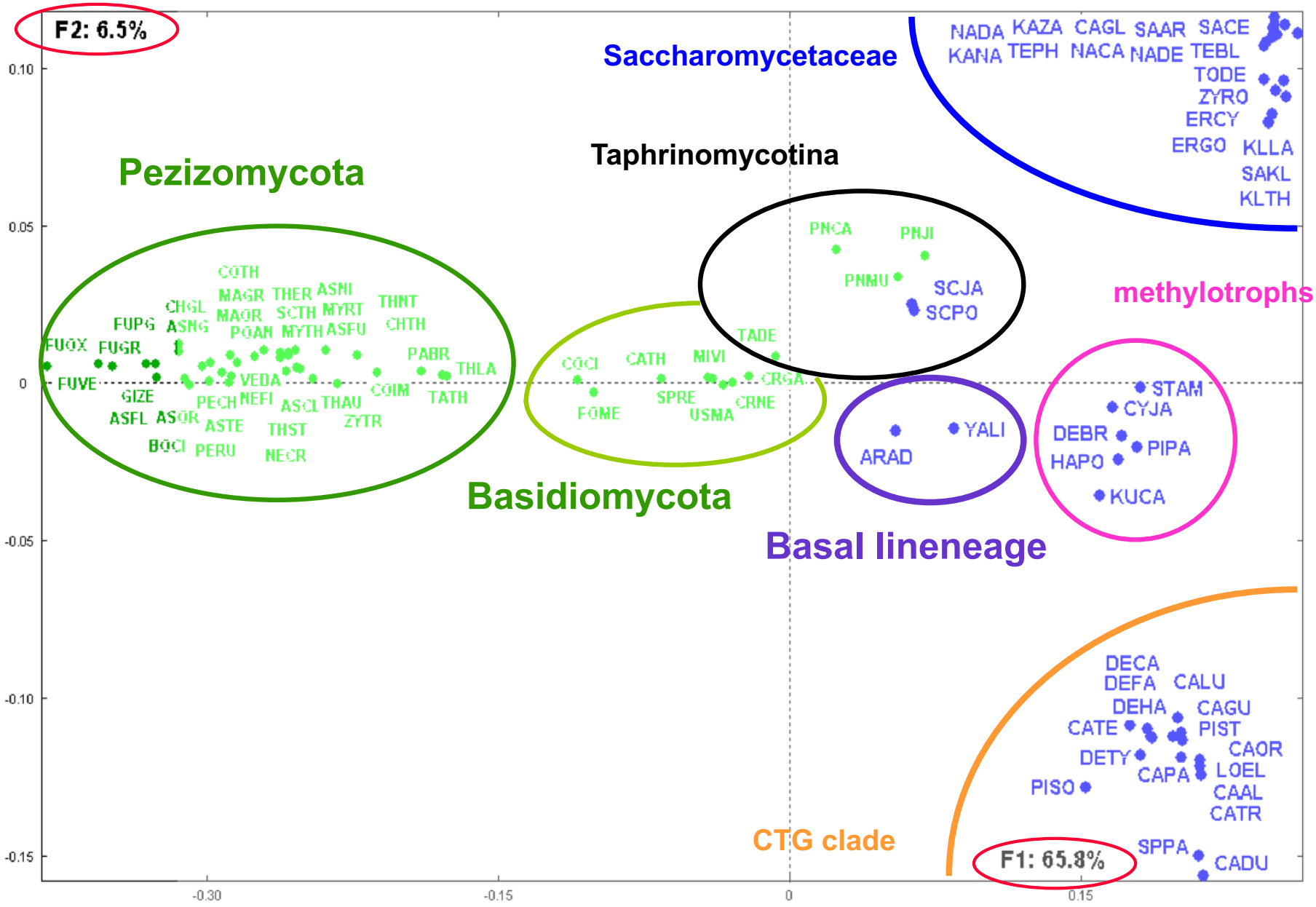
- **Determination of Reciprocal Best Hits between all pairs of species - considered as shared orthologs**

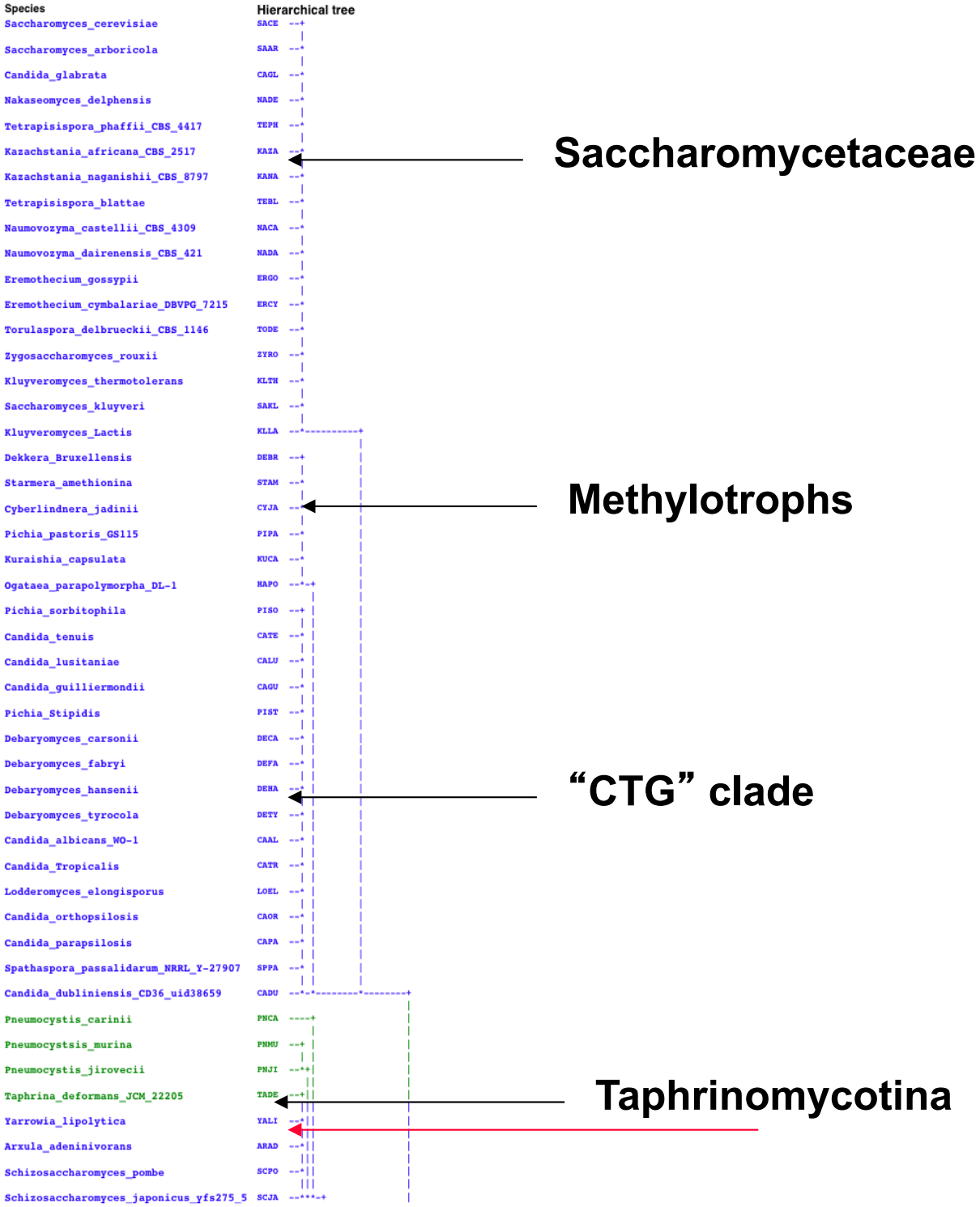
- $T_{ij} = 100 * s_{ij}/(n_i + n_j)$

s_{ij} is the number of shared orthologs between species i and j

n_i and n_j are respectively the total number of proteins in i and in j.

- **The matrix T of dimensions 91x91 is submitted to Correspondence Analysis.**


F2: 6.5%
F1: 65.8%
Saccharomycetaceae
Taphrinomycotina
Pezizomycota
Basidiomycota
methylotrophs
Basal lineneage
CTG clade
NADA KAZA CAGL SAAR SACE
KANA TEPH NACA NADE TEBL
TODE
ZYRO
ERCY
ERGO KLLA
SAKL
KLTH
PNCA PNJI
PNMU
SCJA
SCPO
TADE
CGCI CATH MIVI CRGA
FOME SPRE CRNE
USMA
ARAD YALI
STAM
CYJA
DEBR PIPA
HAPO
KUCA
DECA
DEFA CALU
DEHA CAGU
CATE PIST
CAOR
DETY
CAPA LOEL
PISO
CAAL
CATR
SPPA CADU
COTH
MAGR THER ASNI
CHGL MAGR SCTH MYRT THNT
FUPG ASNG POAN MYTH ASFU CHTH
FUGX FUGR PABR
THLA
FUVE VEDA
GIZE PECH NEFI ASCL THAU CGIM TATH
ASFL ASGR ASTE THST ZYTR
BOCI PERU NECR
0.10
0.05
0
-0.05
-0.10
-0.15
-0.30
-0.15
0
0.15



Saccharomycetaceae

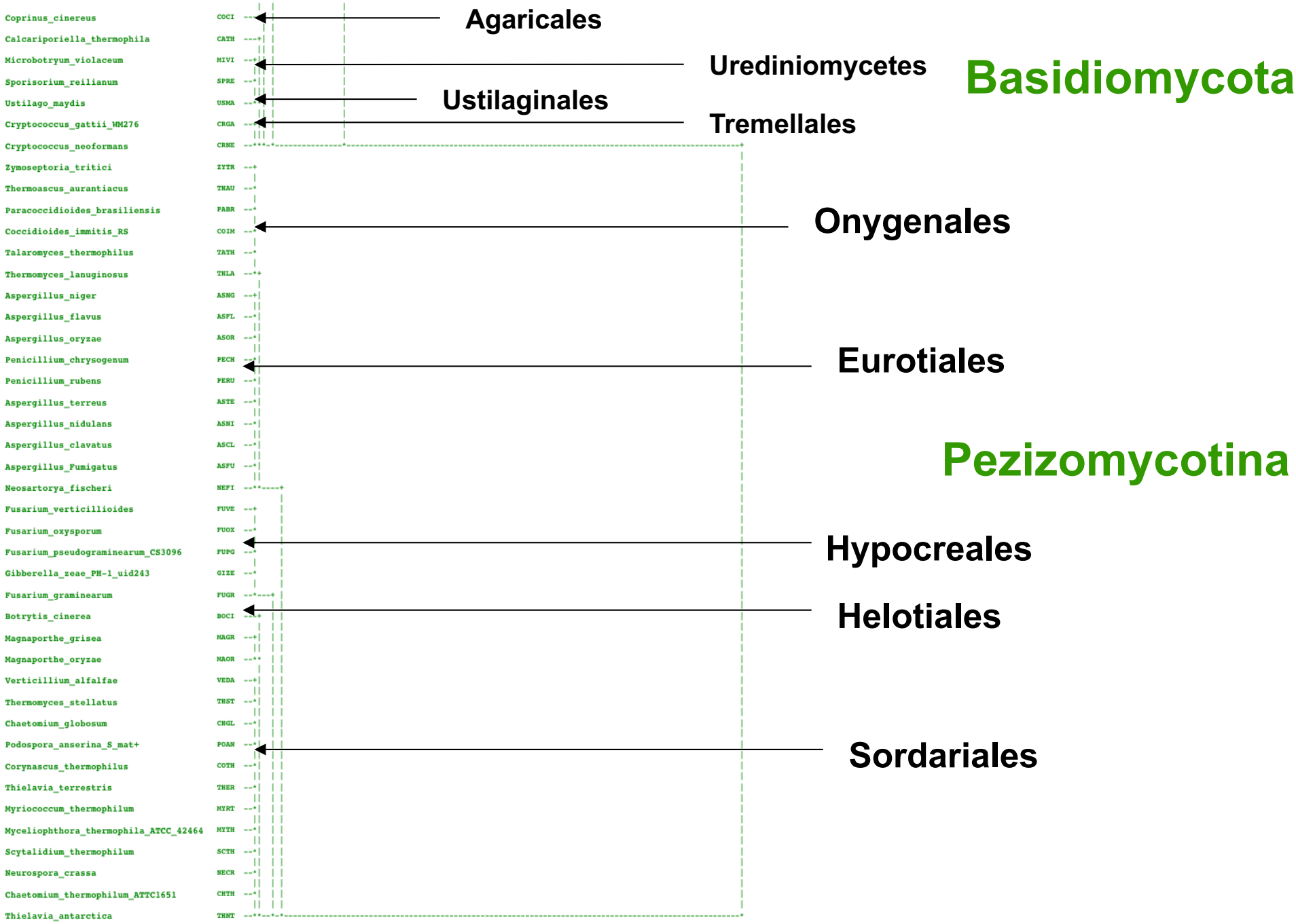
Methylotrophs

“CTG” clade

Taphrinomycotina

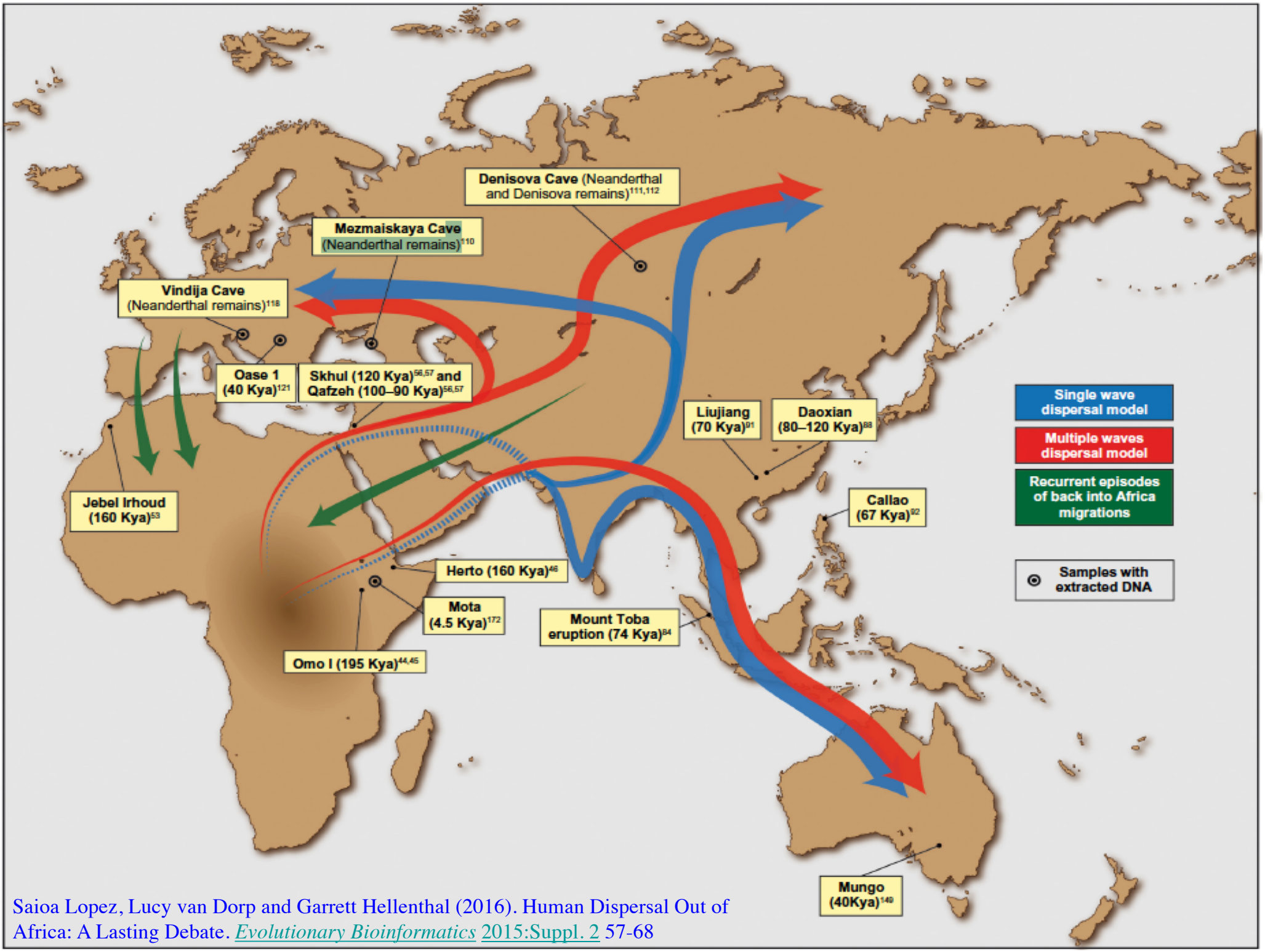
Clustering of the species using Euclien distances between species in the factorial space.

Saccharomycotina


Coprinus_cinereus COCI
Calcariporiella_thermophila CATH
Microbotryum_violaceum MIVI
Sporisorium_reilianum SPRE
Ustilago_maydis USMA
Cryptococcus_gattii_WM276 CRGA
Cryptococcus_neoformans CRNE
Zymoseptoria_tritici ZYTR
Thermoascus_aurantiacus THAU
Paracoccidioides_brasiliensis PABR
Coccidioides_immitis_RS COIM
Talaromyces_thermophilus TATH
Thermomyces_lanuginosus THLA
Aspergillus_niger ASNG
Aspergillus_flavus ASFL
Aspergillus_oryzae ASOR
Penicillium_chrysogenum PECH
Penicillium_rubens PERU
Aspergillus_terreus ASTE
Aspergillus_nidulans ASNI
Aspergillus_clavatus ASCL
Aspergillus_Fumigatus ASFU
Neosartorya_fischeri NEFI
Fusarium_verticillioides FUVE
Fusarium_oxysporum FUOX
Fusarium_pseudograminearum_CS3096 FUPG
Gibberella_zeae_PH-1_uid243 GIZE
Fusarium_graminearum FUGR
Botrytis_cinerea BOCI
Magnaporthe_grisea MAGR
Magnaporthe_oryzae MAOR
Verticillium_alfalfae VEDA
Thermomyces_stellatus THST
Chaetomium_globosum CHGL
Podospora_anserina_S_mat+ POAN
Corynascus_thermophilus COTH
Thielavia_terrestris THER
Myriococcum_thermophilum MYRT
Myceliophthora_thermophila_ATCC_42464 MYTH
Scytalidium_thermophilum SCTH
Neurospora_crassa NECR
Chaetomium_thermophilum_ATTC1651 CHTH
Thielavia_antarctica THNT
Agaricales
Urediniomycetes
Ustilaginales
Tremellales
Basidiomycota
Onygenales
Eurotiales
Pezizomycotina
Hypocreales
Helotiales
Sordariales

Example:

Exploring genotyping data using Correspondence Analysis as compared to Principal Component Analysis



Saioa Lopez, Lucy van Dorp and Garrett Hellenthal (2016). Human Dispersal Out of Africa: A Lasting Debate. *Evolutionary Bioinformatics* 2015:Suppl. 2 57-68

Fu Q, Li H, Moorjani P, et al. *Nature*. 2014;514(7523):445-9.

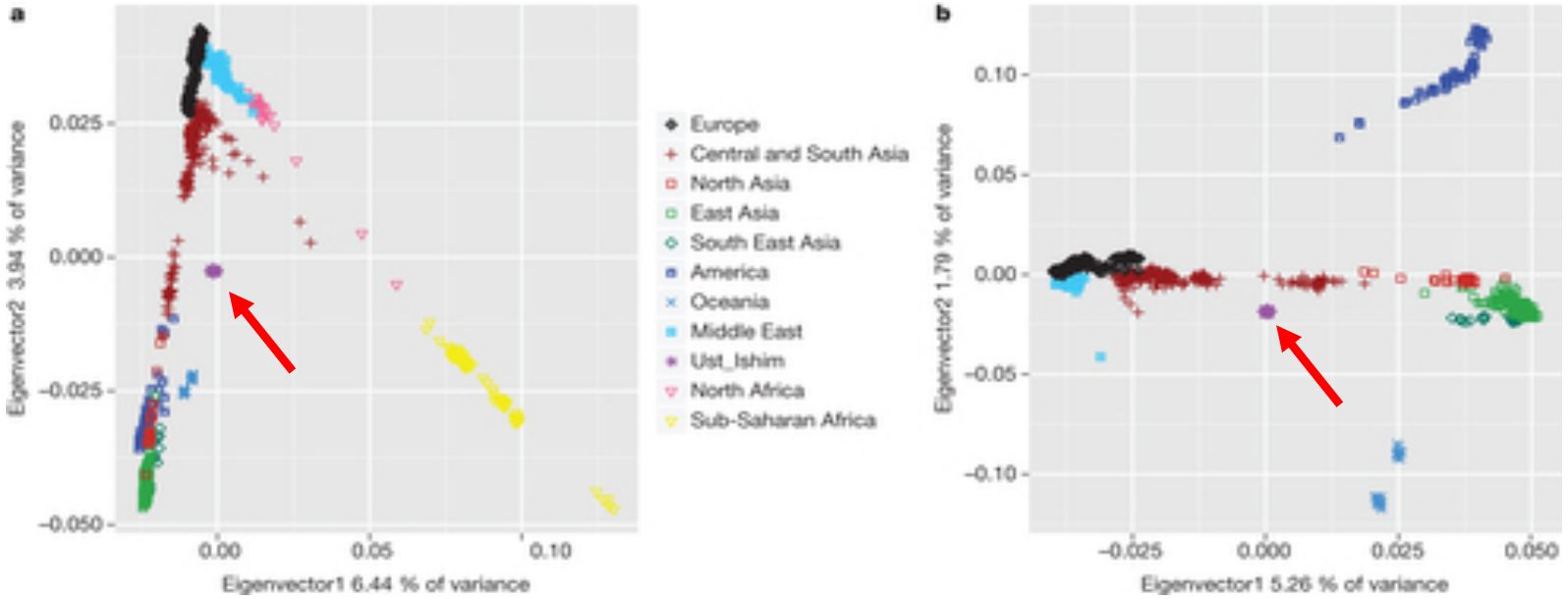
Genome sequence of a 45,000-year-old modern human from western Siberia

Considered 53 present-day human populations

Present-day Humans Populations	Present-day Humans Populations
Tujia (China)	MbutiPygmy
Dai (China)	Oroqen (Mongolia – China)
Daur (China)	Pathan (Pashtun)
Surui (Brazil)	She (Fuji – China)
Uygur (China)	Tu (Mongoe – China)
Xibo(China)	Hezhen (China)
Yakut (Skha, Russia)	Mozabite
Yoruba (West Africa)	Bedouin
Cambodian	Italian
Druze	Kalash (Nuristan – Pakistan)
HanNChina	Orcadian (Orkney –Scotland)
Mandenka (Senegal)	Pima (indigenous americans)
Maya	San (South Africa)
Tuscan	Sardinian
Hazara (Persian Afghan)	Adygei (Caucasus)
Sindhi (Pakistan)	BiakaPygmy
Yi (China)	Han
Balochi (Baloshistan)	Naxi (China)
French	Russian
Karitiana (Brazil)	Brahui (Pakistan)
Lahu (Vietnam-China)	Miao (China)
Burusho (Pakistan)	Basque
Colombian	BantuSouthAfrica
Papuan	Melanesian
Mongola	Japanese
Palestinian	BantuKenya
Makrani (Pakistan)	

- 922 present-day humans are considered from these populations
- Described by their genotyping data.

Genome sequence of a 45,000-year-old modern human from western Siberia



Principal Component Analysis exploring the relationships of Ust-Ishim to present-day humans.

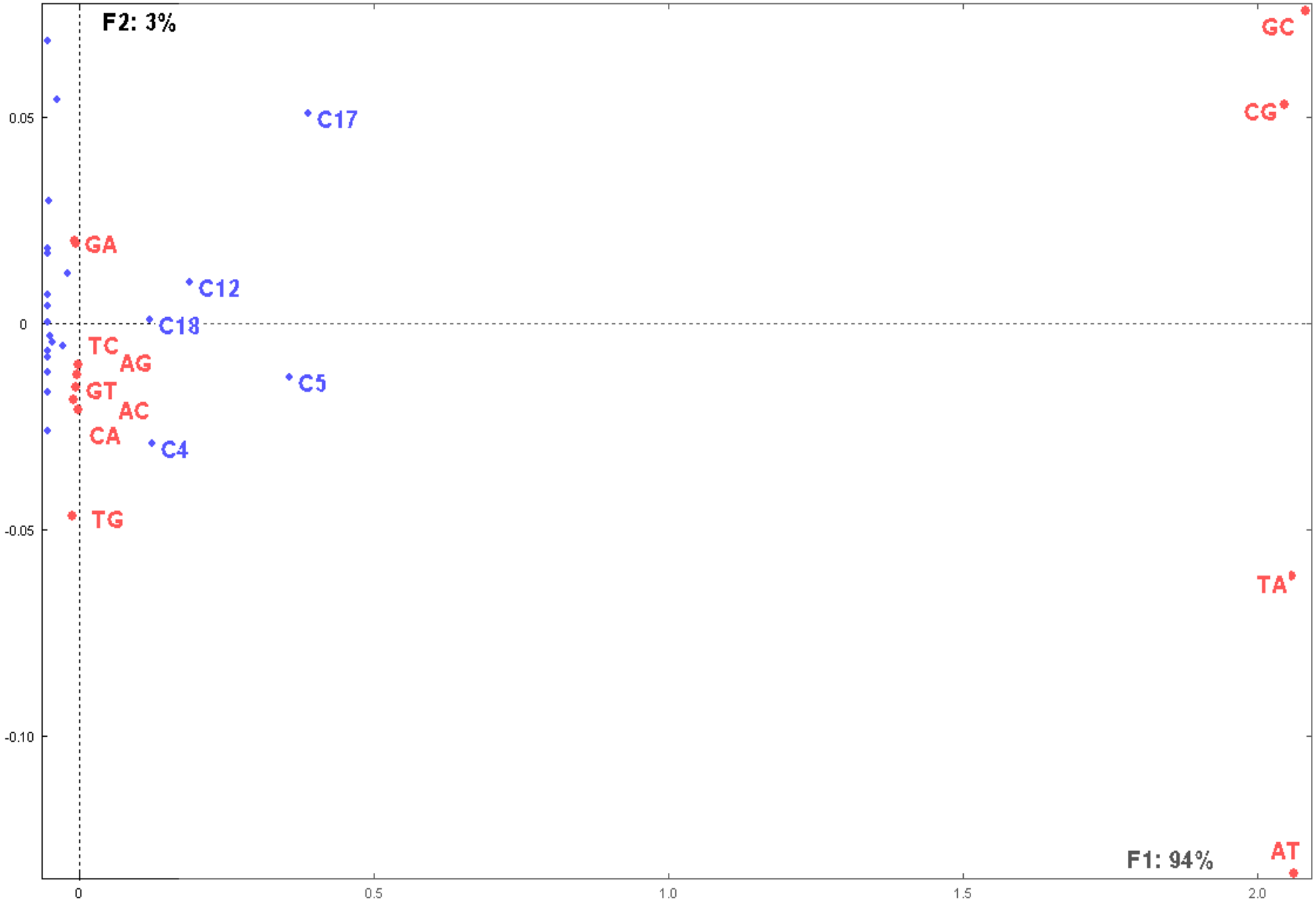
- **The Ust-Ishim individual clusters with present-day Eurasian rather than with Africans.**
- **Note that the genetic diversity is not shown on these plots. The genetic diversity is assumed to explain the displayed distribution.**

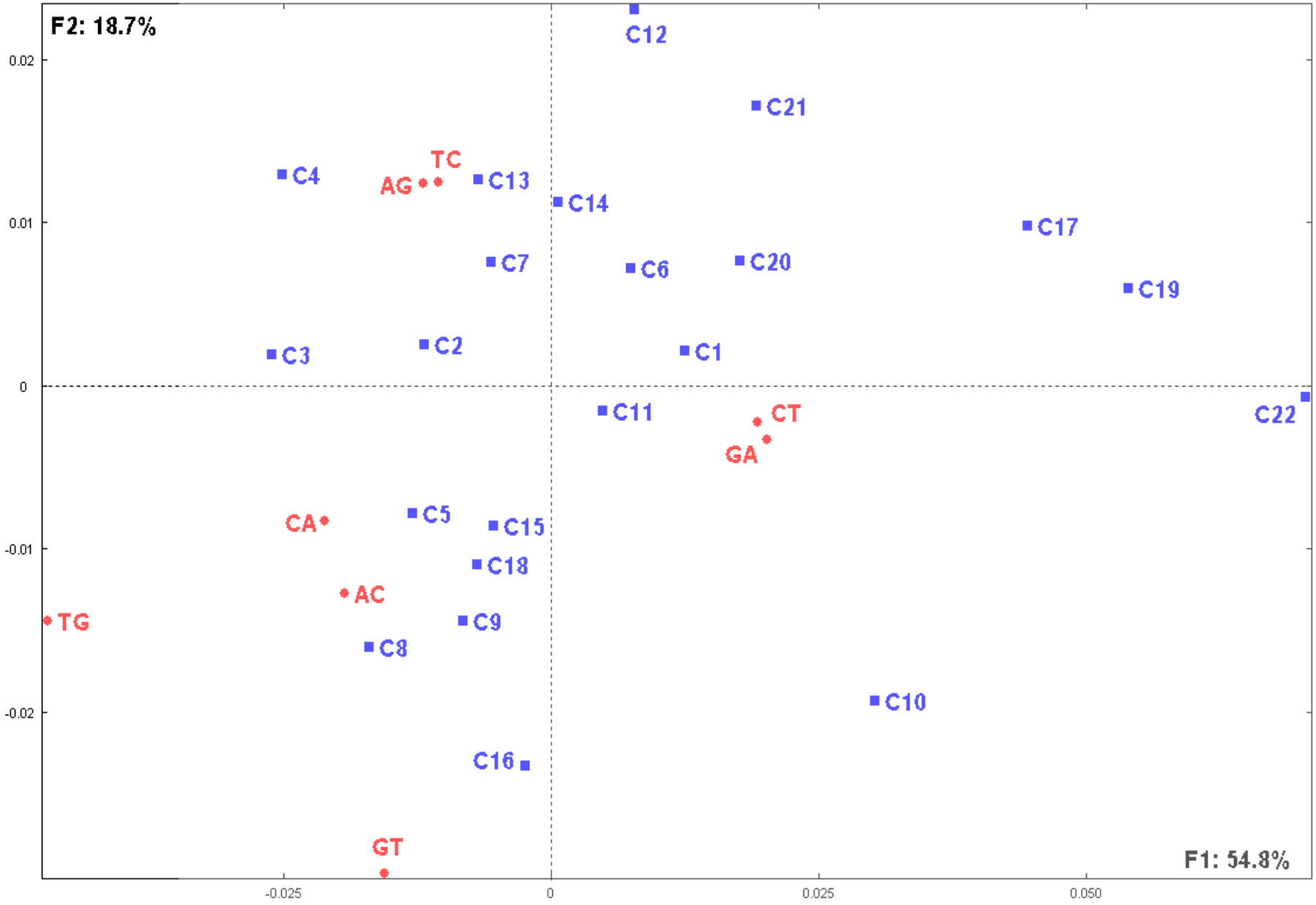
Fu Q, Li H, Moorjani P, et al. Nature. 2014;514(7523):445-9. Genome sequence of a 45,000-year-old modern human from western Siberia

Use of Correspondence Analysis for the same data set (thanks to Qiaomei Fu)

Chromosomes vs SNPs in 922 present-day humans

Chr	Size	AC	AG	AT	CA	CG	CT	GA	GC	GT	TA	TC	TG	Total
C1	224999719	1533	8965	14	1863	27	8869	10678	23	2440	19	6480	1984	42895
C2	237712649	1642	10145	0	2144	1	9437	11717	0	2725	1	7162	2396	47370
C3	194704827	1437	8429	1	1814	0	7774	9542	0	2279	1	6056	2041	39374
C4	187297063	917	5729	54	1204	86	5172	6458	55	1481	64	4101	1379	26700
C5	177702766	779	4660	91	1029	167	4339	5388	132	1329	113	3281	1062	22370
C6	167273993	1113	6951	0	1489	0	6634	8250	0	1827	0	4934	1550	32748
C7	154952424	932	5674	7	1182	14	5278	6531	10	1507	6	4003	1271	26415
C8	142612826	1010	5756	0	1268	0	5516	6846	0	1642	0	4147	1468	27653
C9	120312298	846	4798	0	1060	0	4596	5732	0	1307	0	3327	1196	22862
C10	131624737	1043	5869	0	1250	2	6103	7543	2	1714	1	4282	1360	29169
C11	131130853	936	5628	0	1270	0	5440	6859	0	1514	1	4074	1315	27037
C12	130303534	627	3863	39	809	89	3671	4501	62	948	52	2781	835	18277
C13	95559980	645	3625	0	789	0	3400	4268	0	893	0	2615	848	17083
C14	88290585	650	3890	0	792	0	3611	4517	0	1024	0	2757	842	18083
C15	81341915	636	3558	1	802	3	3515	4272	3	1020	1	2673	847	17331
C16	78884754	643	3700	2	851	1	3578	4401	1	1062	0	2506	870	17615
C17	77800220	273	1665	31	363	74	1742	2108	56	427	42	1228	352	8361
C18	74656155	461	2652	14	579	49	2559	3267	37	750	25	1972	672	13037
C19	55785651	263	1651	0	316	4	1710	2095	2	438	1	1177	342	7999
C20	59505254	481	3152	0	671	0	3139	3751	0	806	0	2212	722	14934
C21	34171998	286	1651	0	324	0	1569	1972	0	392	0	1136	360	7690
C22	34893953	222	1487	0	320	0	1510	2036	0	388	0	1038	304	7305
Total		17375	103498	254	22189	517	99162	122732	383	27913	327	73942	24016	492308


F2: 3%
F1: 94%
GC
CG
TA
AT
GA
TC
AG
GT
AC
CA
TG
C17
C12
C18
C5
C4
0.05
0
-0.05
-0.10
0
0.5
1.0
1.5
2.0


F2: 18.7%
F1: 54.8%
0.02
0.01
0
-0.01
-0.02
-0.025
0
0.025
0.050
C12
C21
C4
TC
AG
C13
C14
C17
C7
C6
C20
C19
C3
C2
C1
C11
CT
GA
C22
CA
C5
C15
C18
AC
TG
C9
C8
C10
C16
GT

Data table

Chromosome

SNP

0 : zero copy of reference allele,
1 : one copy of reference allele,
2 : two copies of reference allele.

IND	1AC0	1AC1	1AC2	1AG0	1AG1	1AG2	1AT0	1AT2	1CA0	1CA1	1CA2	1CG0	1CG2	....	22TG0	22TG1	22TG2
HGDP00784	310	375	846	1662	1981	5309	14	0	280	414	1161	27	0	...	47	62	194
HGDP00785	323	358	850	1708	1905	5330	14	0	309	38							191
HGDP00786	302	393	838	1700	2008	5251	14	0	314	41							192
HGDP00787	318	361	853	1823	1829	5305	14	0	312	35							204
.....	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Ust_Ishim	288	392	853	1514	2306	5151	0	14	289	429	1154	0	27	...	45	73	186

Final data table:

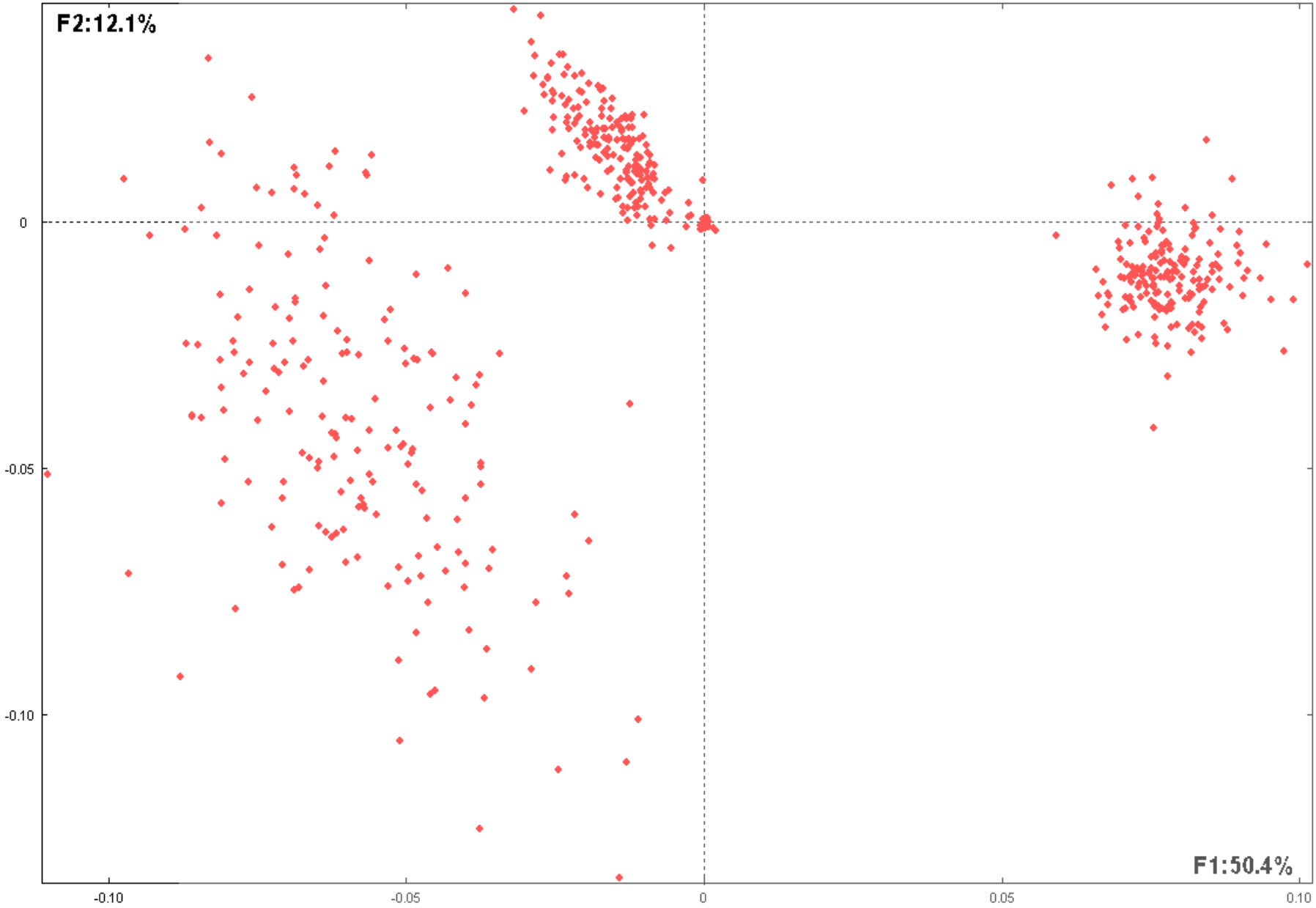
976 vs 622

622 columns corresponding to non-null SNPs modalities

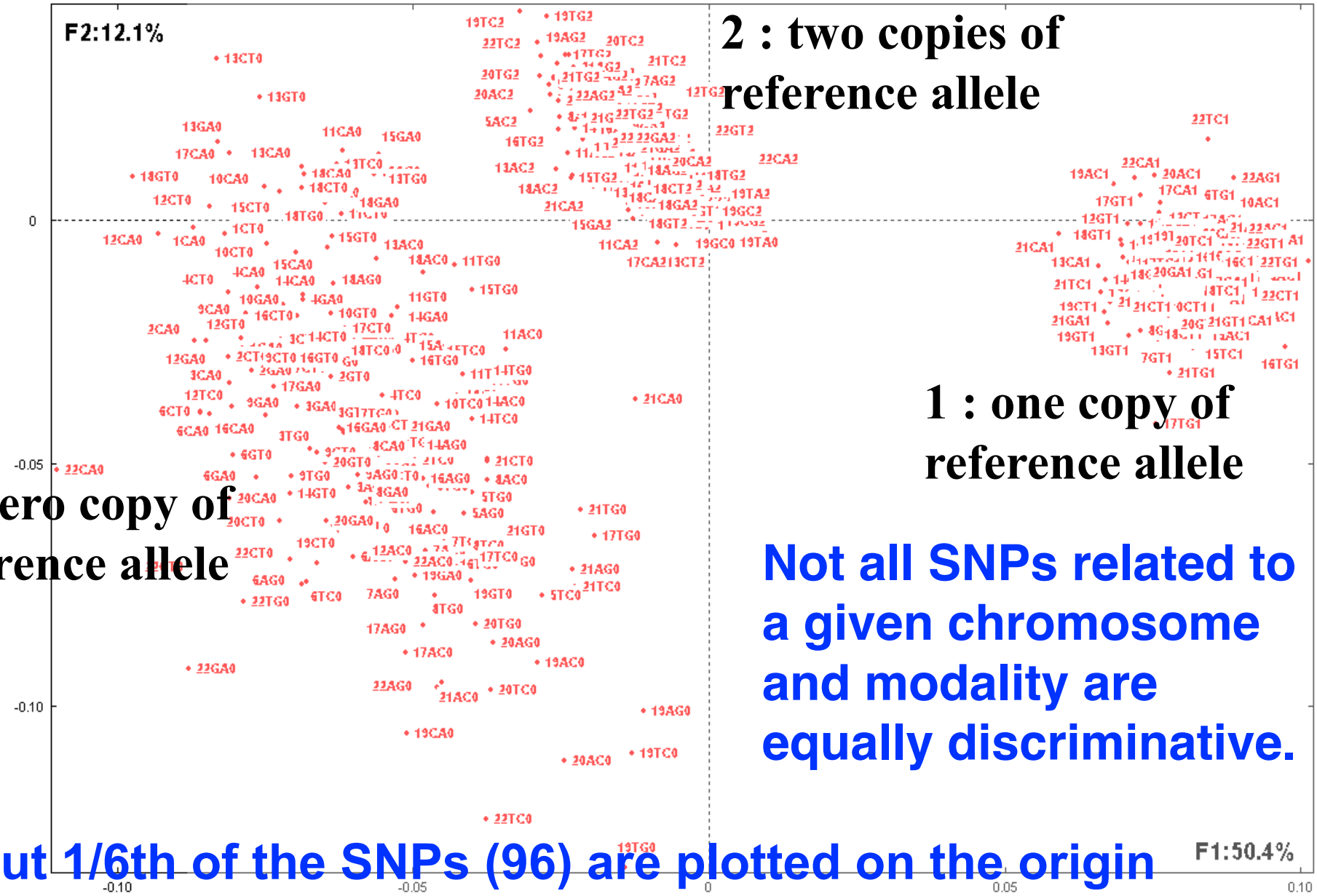
T_{ij} = Number of SNPs associated with its (chromosome number and modality j) present in individual i.

- 922 Present-day humans from 53 different populations
- Supplementary individuals: *Ust-Ishim* + 53 distinct populations.

Distribution of the 622 SNPs modalities



Distribution of the 622 SNPs modalities



2 : two copies of reference allele

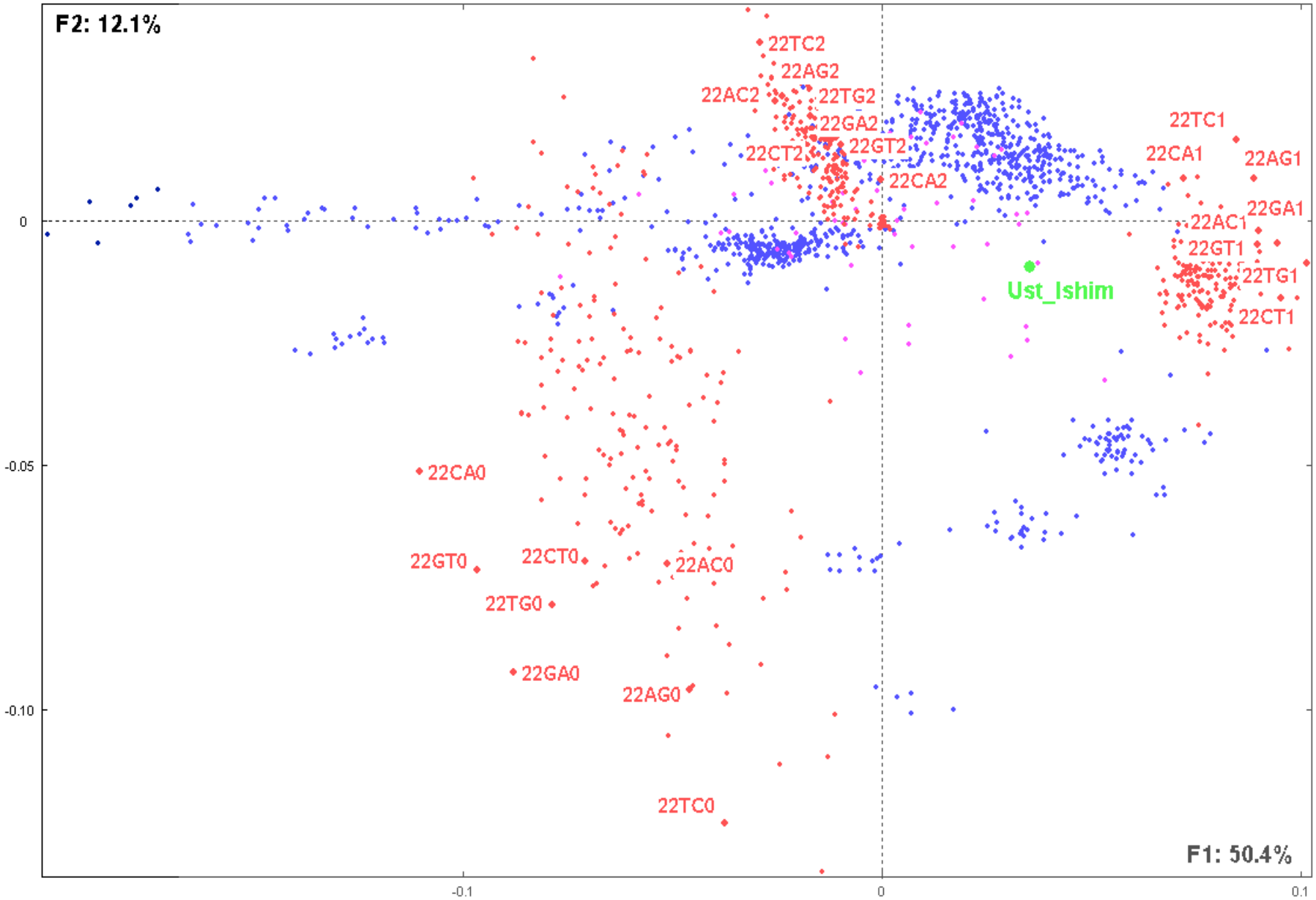
1 : one copy of reference allele

0 : zero copy of reference allele

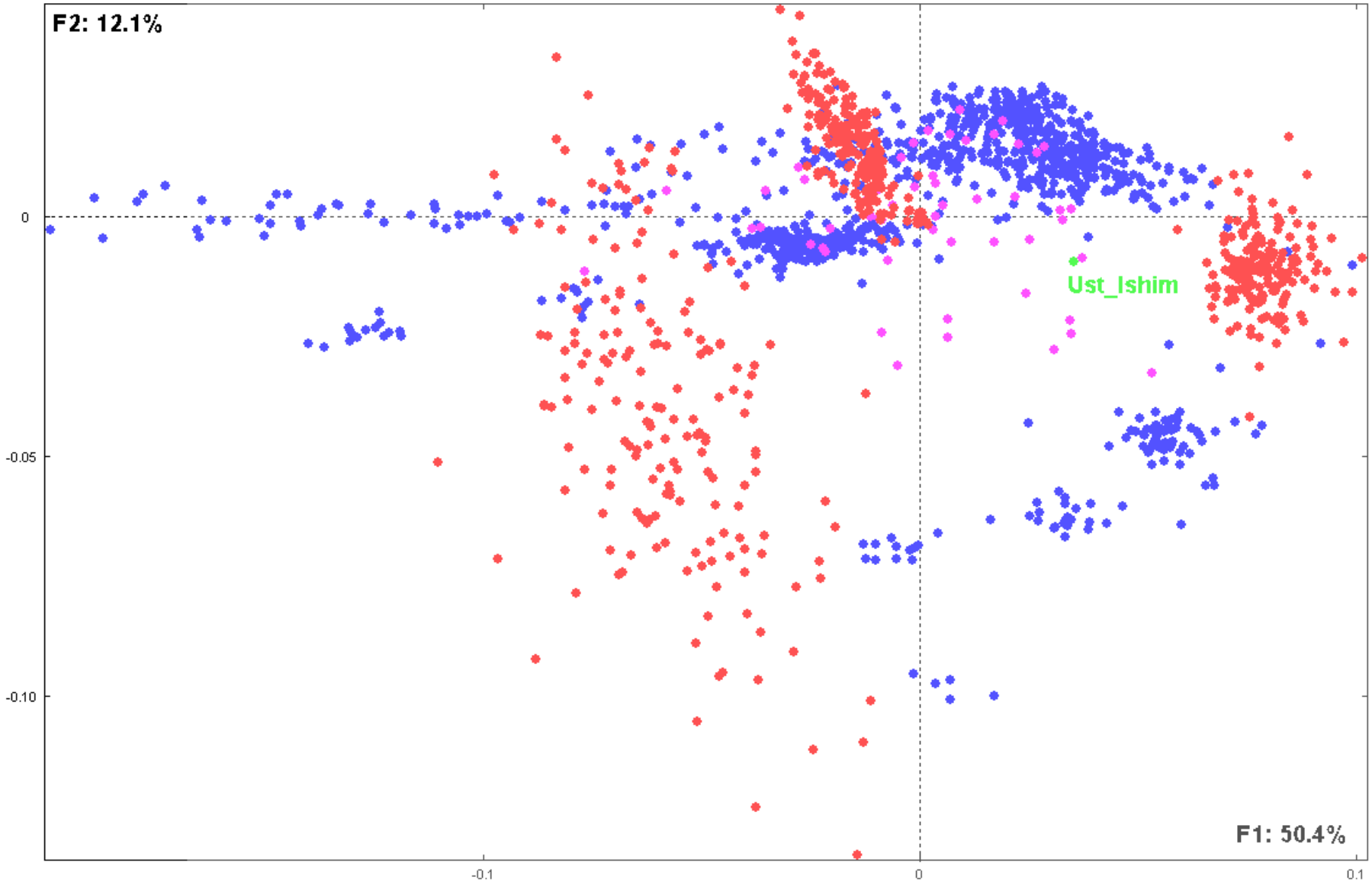
Not all SNPs related to a given chromosome and modality are equally discriminative.

About 1/6th of the SNPs (96) are plotted on the origin meaning that they are equally distributed in all considered present-day humans or showing weak values.

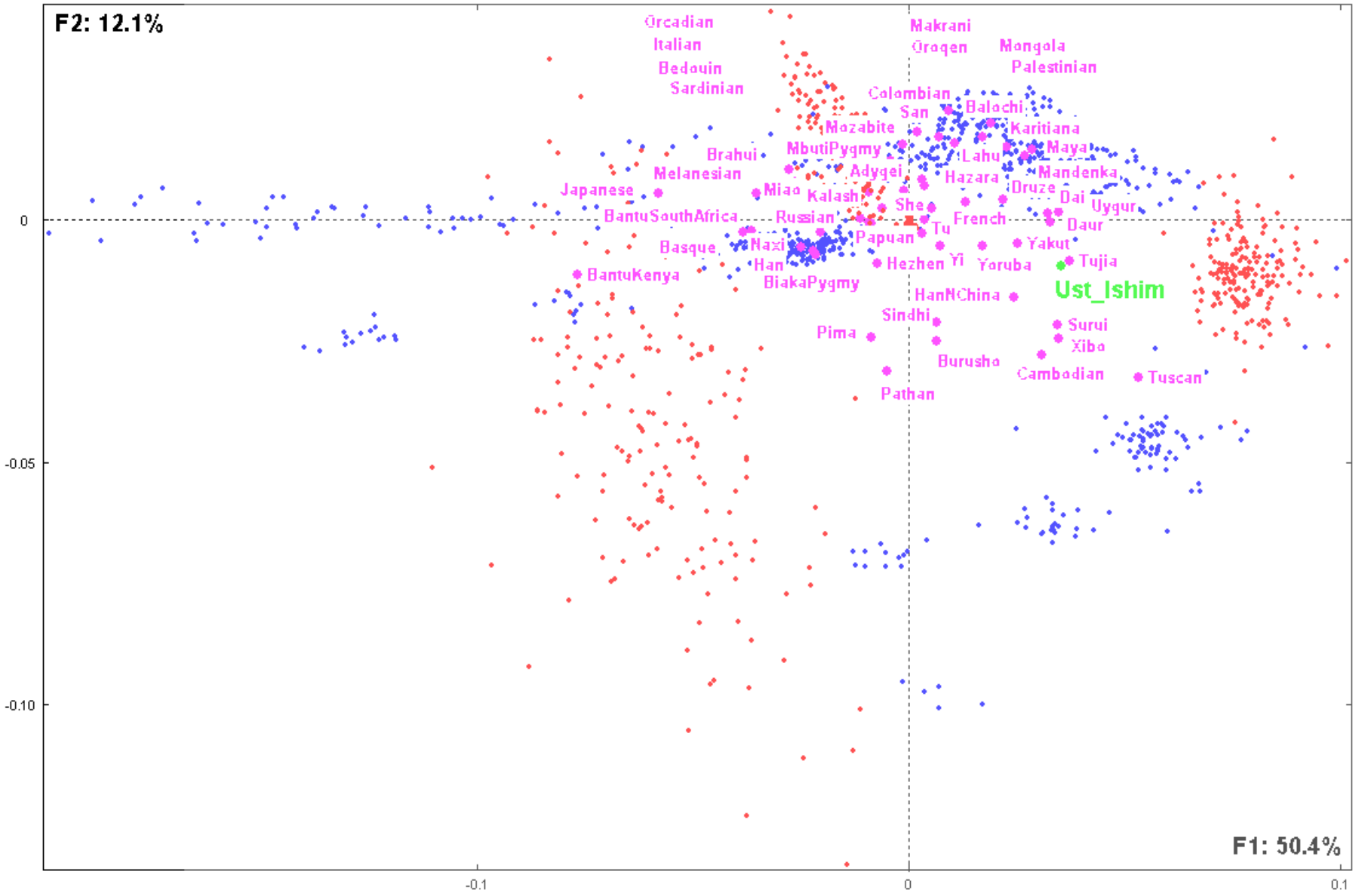
Example: distribution of SNPs related to chromosome 22



Distribution of present-day humans (blue) together with SNPs modalities (red) and position of Ust_Ishim individual



Distribution of representatives of the 53 considered populations and the Ust_Ishim individual



Present-day Humans	Distance^2 to Ust_Ishim
Tujia (China)	0.0001
Dai (China)	0.0003
Daur (China)	0.0003
Surui (Brazil) *	0.0003
Uygur (China)	0.0003
Xibo(China)	0.0003
Yakut (Skha, Russia) *	0.0003
Yoruba (West Africa) *	0.0005
Cambodian	0.0006
Druze	0.0006
HanNChina	0.0006
Mandenka (Senegal)	0.0006
Maya	0.0006
Tuscan	0.0006
Hazara (Persian Afghan)	0.0011
Sindhi (Pakistan)	0.0011
Yi (China)	0.0011
Balochi (Baloshistan)	0.0014
French	0.0014
Karitiana (Brazil)	0.0014
Lahu (Vietnam-China)	0.0014
Burusho (Pakistan)	0.0015
Colombian	0.0018
Papuan	0.0018
Mongola	0.0019
Palestinian	0.0019
Makrani (Pakistan)	0.0021
MbutiPygmy	0.0021
Oroqen (Mongolia – China)	0.0021
Pathan (Pashtun)	0.0021
She (Fuji – China)	0.0021
Tu (Mongoe – China)	0.0021
Hezhen (China)	0.0026
Mozabite	0.0026
Bedouin	0.0027
Italian	0.0027
Kalash (Nuristan – Pakistan)	0.0027
Orcadian (Orkney –Scotland)	0.0027
Pima (indigenous americans)	0.0027
San (South Africa)	0.0027

Distance of Ust_Ishim with each of the considered present-day population

FU et al Conclusions:

- Ust' Ishim individual clusters with non-Africans rather than Africans.
- The Ust' Ishim genome shares more derived alleles with present-day people from East Asia than with present-day Europeans.

Note nevertheless the proximity with some non Asian humans and even African

Use of CA in genotype data analyses is much more informative than with Principal Component Analysis.

Correspondence Analysis output results are richer :

- **Distribution and clustering of SNPs with corresponding chromosomes and number of associated copies.**

- **Not all SNPs are equally discriminative with regard to the present-day human individuals**

- **Distribution and clustering of present day humans**

and most importantly

- **The connection of present-day humans with their associated chromosomes, SNPs, number of copies allowing a much finer interpretation of the analysed data than PCA.**

Correspondence Analysis and Principal Component Analysis

What differenciate CA from PCA:

- **Essentially CA deals with profiles (of rows and of columns) instead of raw data;**
- **The use of the χ^2 distance or the distributional distance ;**
- **The symmetry between individuals (or rows) and variables (or columns) ;**
- **Duality: the transition formulae that allows in the factorial space, the calculation of row coordinates as a function of column coordinates and vice versa;**
- **The display of individuals together with the variables on the same factorial space ;**
- **The ability of introducing supplementary individuals or variables.**

Concluding notes

- **Correspondence Analysis is a descriptive multivariate data analysis method.**

- **It allows to synthesize information included in a large data table by constructing an orthogonal system of axes and by displaying observations and variables on a reduced number of factors.**

- **Planar graphical representations of observations and of variables allow salient relationships to be easily detected.**

Reference

- **Tekaia F. 2016.**
Exploring Genome Data Using Correspondence Analysis.
***Bioinformatics and Biology Insights*. 2016: 10 59-72.**

http://la-press.com/article.php?article_id=5675

- **Benzecri J.P. 1973. *L'analyse des données Vol. 2*: *L'analyse des correspondances* . Dunod, Paris.**

- **Murtagh, F. 2005. Correspondence Analysis And Data Coding With Java And R. ed . Chapman & Hall/CRC . 248 p. ISBN : 1584885289.**

Thank You