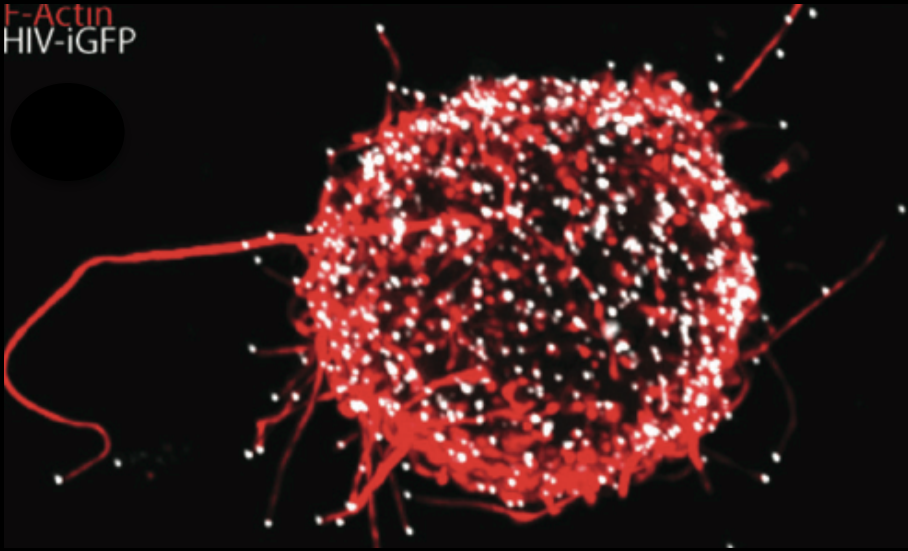
HOST-PATHOGEN CO-EVOLUTION THROUGH HIV-1 WHOLE GENOME ANALYSIS



Somdatta Sinha

Indian Institute of Science, Education
& Research Mohali, INDIA

International Visiting Research Fellow,
Peter Wall Institute of Advanced Studies,
Visiting Professor, Mathematics,
University of British Columbia, Canada

IDM 2018
Apr 16-18, 2018

HOST – PATHOGEN INTERACTION INVOLVES CONTINUOUS

‘ARMS RACE’

❖ Pathogens adapt to interact with their hosts to optimally utilize the host’ s resources and enhance their replicative fitness.

❖ Host mounts complex defense mechanisms against the pathogen attack.

❖ Host survival is required for the pathogen to propagate.

Identifying the selective constraints regulating host-pathogen adaptation is critical to the clinical management of any infection

The retrovirus HIV-1 is an obligate pathogen, which utilizes the host’s intracellular machinery for replication and growth.

- *High rates of mutation and frequent recombination events have led to* ***enormous genetic diversity*** *for HIV.*

- *High mutation rate, long incubation period inside the host, and the ability to adapt and evade the host’s defense mechanisms are* ***impediments to the development of a foolproof strategy for countering it****.*

HIV evolutionary processes continuously unfold, leaving a measurable footprint in viral gene sequences.

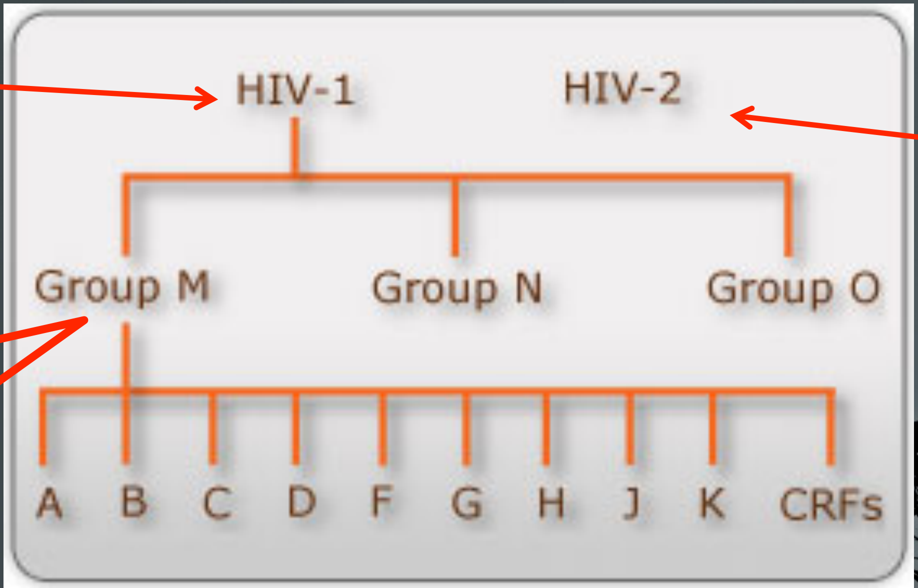
Different population genetic forces are at work both within and among hosts

Human Immunodeficiency Virus (HIV)

TYPES AND SUB-TYPES OF HIV

Of the two types of HIV, the Type 1 virus (HIV-1) is more infectious and causes higher mortality compared to Type 2.

Chimpanzee
Pan troglodytes

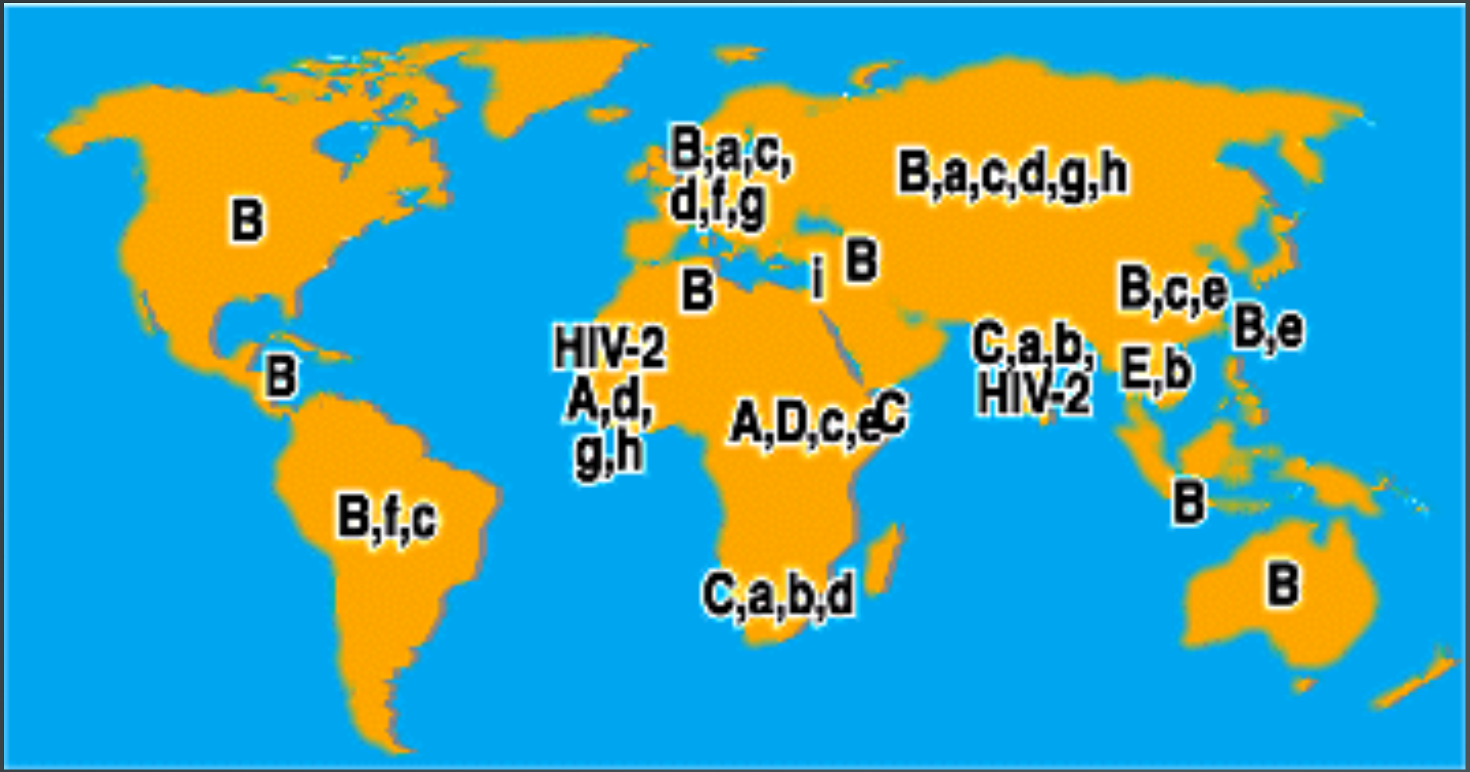


Sooty
Mangabey

25% to 35% sequence variation between subtypes

DISTRIBUTION OF HIV SUBTYPES

MAJOR SUBTYPES IN CAPITALS



Subtypes exhibit variable treatment response and differential selection of drug resistance mutations.

IT IS IMPORTANT TO BE ABLE TO CLASSIFY NEWLY EMERGING VARIANTS CORRECTLY FOR THERAPEUTIC INTERVENTIONS.

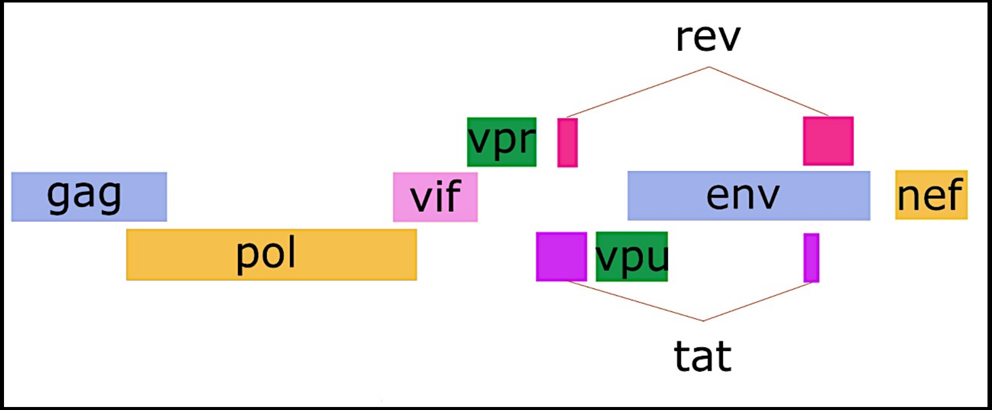
Persistence of high genetic diversity in HIV-1 strengthens the case for neutral forces/drift dominating population-level evolution

The seven stages of the HIV life cycle:

1) binding, 2) fusion, 3) reverse transcription,
4) integration, 5) replication, 6) assembly, 7) budding.

Nine HIV-1 genes

env, gag, pol, rev, tat,nef, vpr, vpu, vif



In CD4+ T- cells, HIV-1 completes its replication cycle in ~ 24 hrs.

No study was done on the population level variation in all nine HIV-1 genes

Study the evolutionary patterns of HIV-1 genes, with respect to its human host genome, to understand the mechanics of their adaptation

HIV-1 sequences downloaded from the HIV Sequence Database (*www.hiv.lanl.gov, Feb 2008*)

10,609 gene sequences from year 1983 to 2005 of various clades, subtypes and CRFs are analyzed.

1431 whole genome sequences for 23 years from different clades and geographical regions.

Genome length ranges from 8023 to 9859 base pairs.

For temporal analyses, the sequence data was further divided according to their year of extraction.

For each gene, 23 sets of gene sequences grouped based on their year of extraction.

HIV is an AT rich retrovirus, and a translational parasite on the human host.

How does it maintain the translation efficiency of its genes inside the GC3-rich host ?

Hypothesis:

The pattern of usage of codons should correlate with the host pattern.

**ANALYSIS OF
CODON USAGE PATTERN**

4 nucleotides (ATCG)

64 three-letter words (codons)

20 amino acids

Some amino acids are coded by multiple codons: **Degeneracy**

Codon Bias

Translational Optimization

CODONS

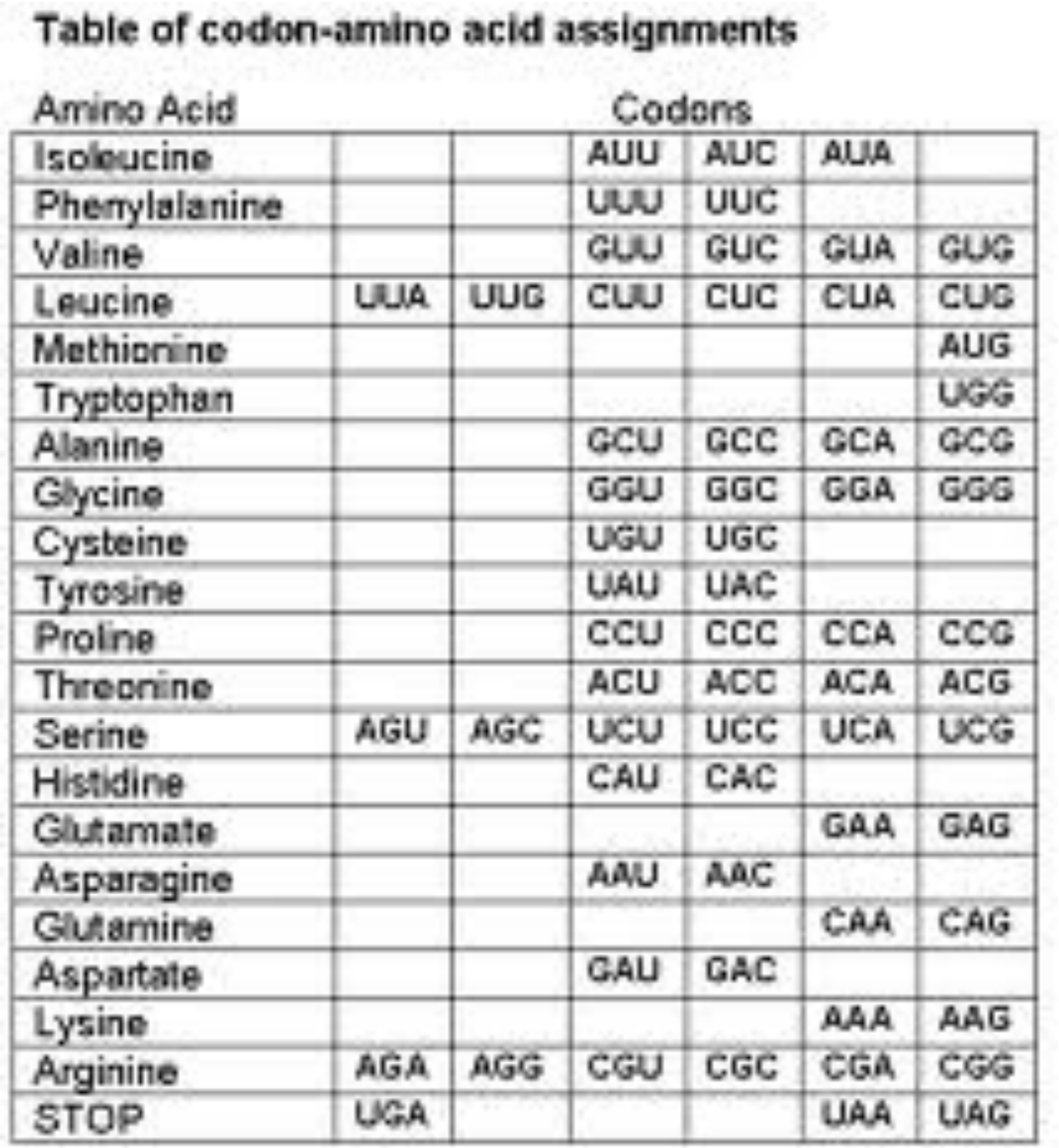
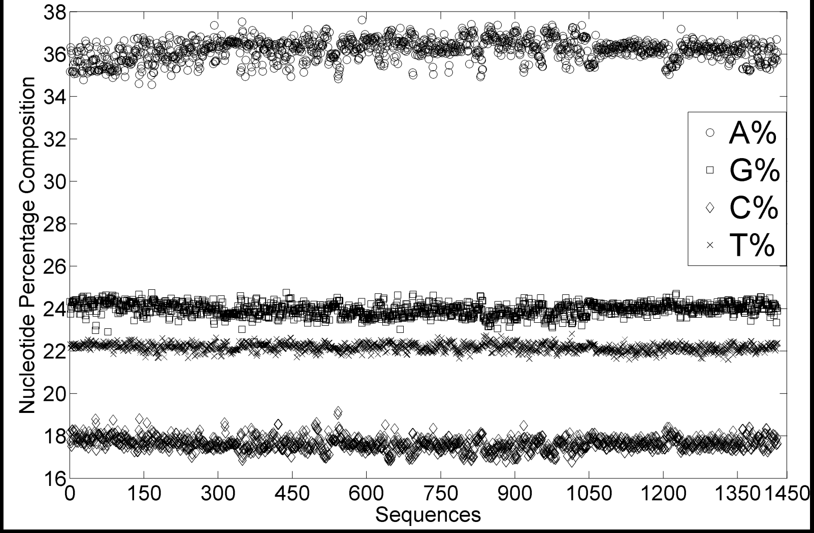


Table of codon-amino acid assignments

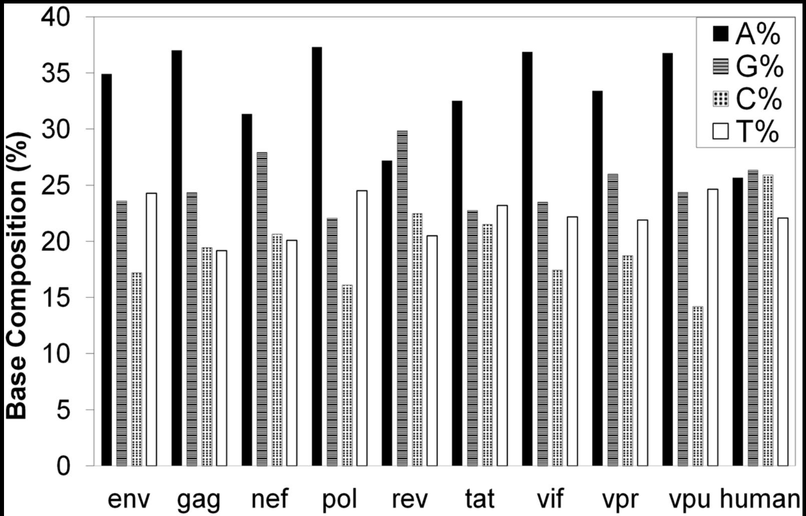
Amino Acid	Codons					
Isoleucine			AUU	AUC	AUA	
Phenylalanine			UUU	UUC		
Valine			GUU	GUC	GUA	GUG
Leucine	UUA	UUG	CUU	CUC	CUA	CUG
Methionine						AUG
Tryptophan						UGG
Alanine			GCU	GCC	GCA	GCG
Glycine			GGU	GGC	GGA	GGG
Cysteine			UGU	UGC		
Tyrosine			UAU	UAC		
Proline			CCU	CCC	CCA	CCG
Threonine			ACU	ACC	ACA	ACG
Serine	AGU	AGC	UCU	UCC	UCA	UCG
Histidine			CAU	CAC		
Glutamate					GAA	GAG
Asparagine			AAU	AAC		
Glutamine					CAA	CAG
Aspartate			GAU	GAC		
Lysine					AAA	AAG
Arginine	AGA	AGG	CGU	CGC	CGA	CGG
STOP	UGA				UAA	UAG

Base Composition of HIV-1 Whole Genome Sequences



HIV-1 genomes are A-rich (36%).

Base compositions has remained constant (23 years)



HIV-1 genes exhibit a similar A-richness (> 31%), ***except rev (27%), which is closer to the host (25.66%)***

Preference for G & C nucleotides at different codon positions

In absence of selection:

GC3 versus GC12 content for a gene lies along the diagonal.

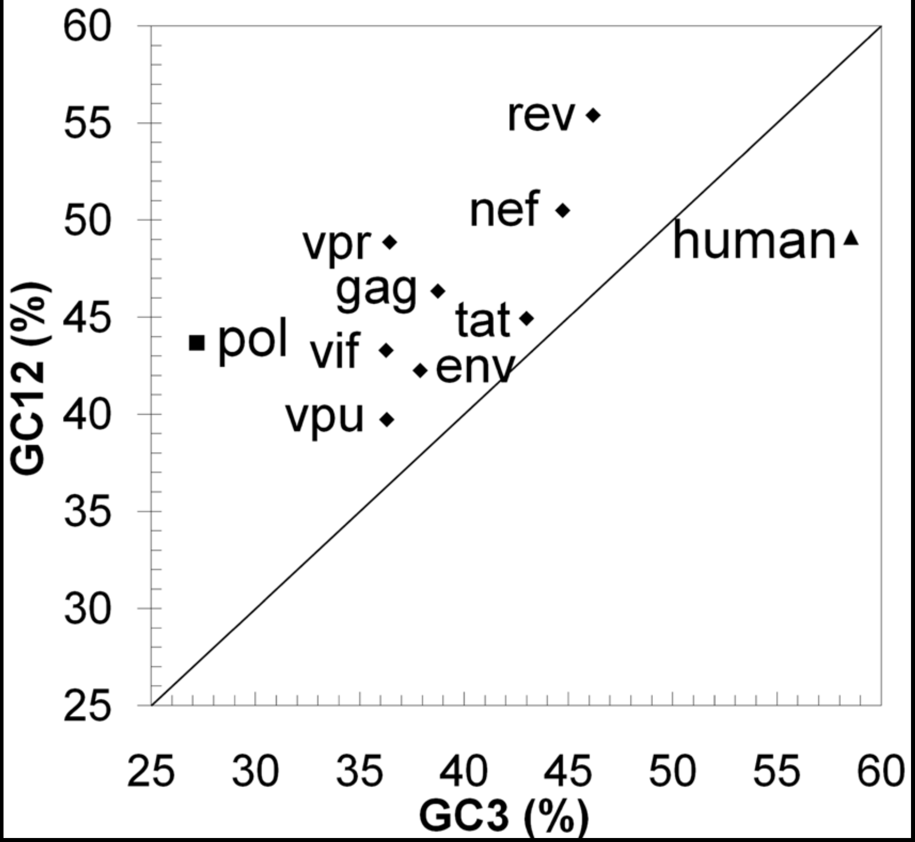
Deviation from the diagonal:

Indicative of selective constraints in modulating the position specific GC content.

Average GC3 content is high for human (58.55%).

For all the HIV-1 genes it is always lower than their GC12 content, with ***pol*** gene being the lowest.

Bias for AT3 codons, is even greater than the overall AT bias.



Codon Usage Analysis

- Codon usage data corresponding to each gene is calculated using CodonW (http://codonw.sourceforge.net).
- Base compositions and codon usage table for human retrieved from Codon Usage Table Database (http://www.kazusa.or.jp/codon/) derived from Genbank Release 160.0 (June 15, 2007).
- The start codon (AUG), UGG codon for tryptophan, and the three stop codons (UAA, UAG and UGA) not included in the analysis.

Codon usage data can be biased by number of synonymous codons for each amino acid, frequency of codons etc.

The normalized frequency values are

$$n_{ij} = \frac{x_{ij}}{x_{j\max}}$$

$\mathbf{n_{ij}}$ normalized value for ith codon & jth amino acid

$\mathbf{x_{ij}}$ frequency of ith codon for the jth amino acid

$\mathbf{X_{jmax}}$ frequency of the maximally used synonymous codon for the jth amino acid

(Suzuki, et al, FEBS Lett. 2005)

Raw Codon Frequency

Amino Acid	Codon	env	gag	nef	pol	rev	tat	vif	vpr	vpu
Phe	**UUU**	18682	10399	6225	25361	587	1314	3698	2559	729
	UUC	14989	6794	4590	13488	157	1187	110	1544	392
Leu	**UUA**	23604	21280	5800	39369	325	2266	6032	2871	5338
	UUG	22090	7248	1756	11397	1773	182	4537	947	2790
	CUU	13464	6349	2343	11740	5958	390	97	2150	2226
	CUC	14455	5373	1107	7454	3363	126	45	1338	173
	CUA	18233	7083	4804	20146	1727	2071	5038	2669	1460
	CUG	19602	4525	6190	11803	1913	429	4904	2754	1101

Normalized Codon Frequency

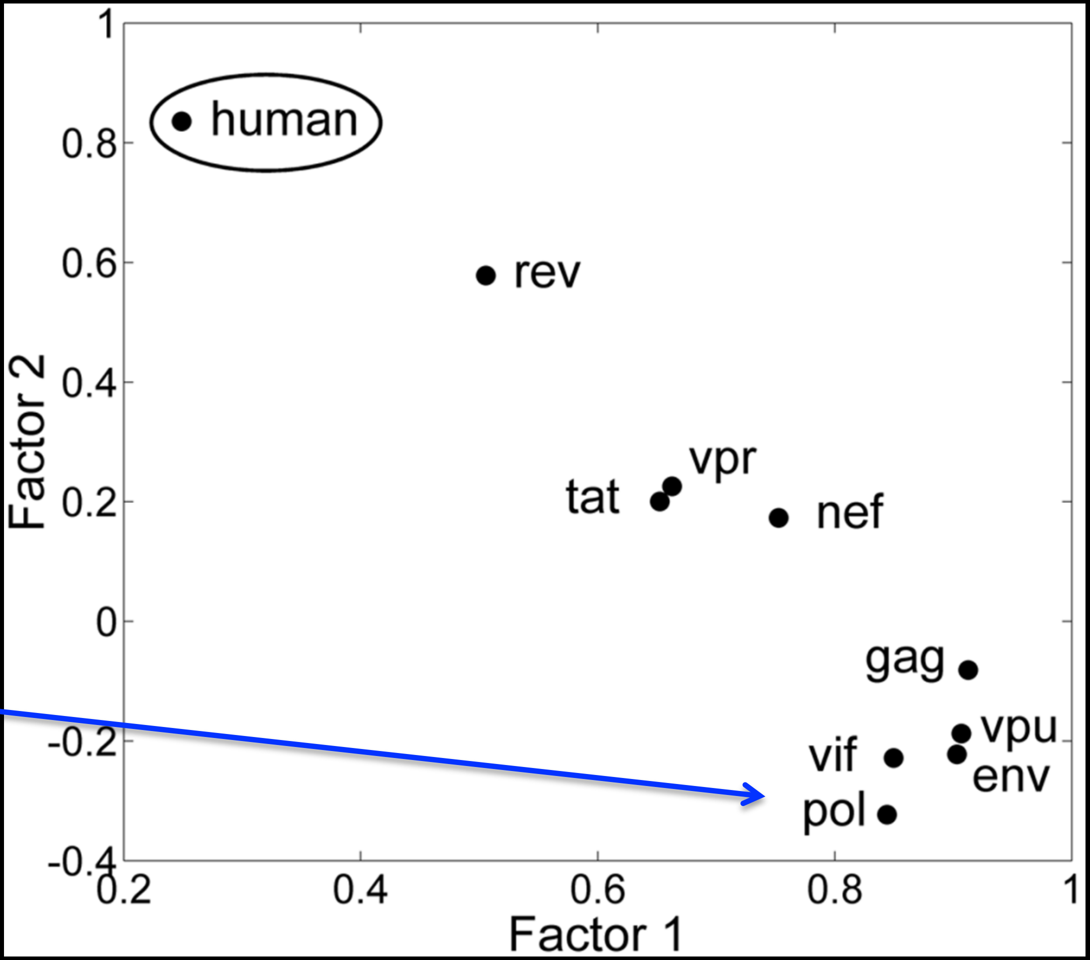
Amino Acid	Codon	env	gag	nef	pol	rev	tat	vif	vpr	vpu	human
Phe	**UUU**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.866
	UUC	0.802	0.653	0.737	0.532	0.267	0.903	0.030	0.603	0.538	**1.000**
Leu	**UUA**	**1.000**	**1.000**	0.937	**1.000**	0.055	**1.000**	**1.000**	**1.000**	**1.000**	0.193
	UUG	0.936	0.341	0.284	0.289	0.298	0.080	0.752	0.330	0.523	0.326
	CUU	0.570	0.298	0.379	0.298	**1.000**	0.172	0.016	0.749	0.417	0.333
	CUC	0.612	0.252	0.179	0.189	0.564	0.056	0.007	0.466	0.032	0.494
	CUA	0.772	0.333	0.776	0.512	0.290	0.914	0.835	0.930	0.274	0.180
	CUG	0.830	0.213	**1.000**	0.300	0.321	0.189	0.813	0.959	0.206	**1.000**

FACTOR ANALYSIS

(normalized codon usage data of HIV-1 genes and human)

Clear presence of a cline among the four regulatory genes (*nef*, *rev*, *tat* and *vpr*)

Three structural genes (*env*, *gag*, *pol*) and two regulatory genes (*vif*, *vpu*) cluster away from human



Principal Component Analysis (PCA)

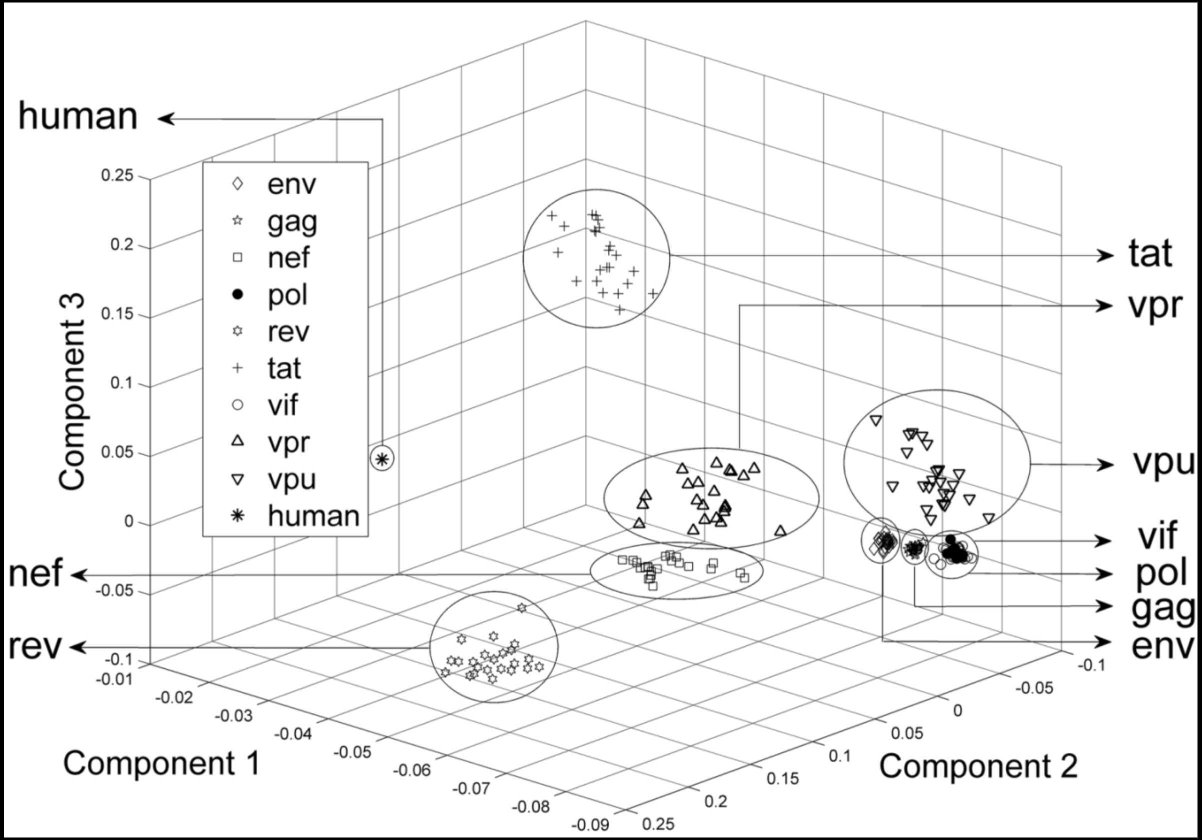
PCA was performed on the 207 variables with 59 observations each corresponding to the degenerate codons

HIV-1 gene sequences from year 1983 to 2005

9 HIV-1 genes separates into 9 different clusters

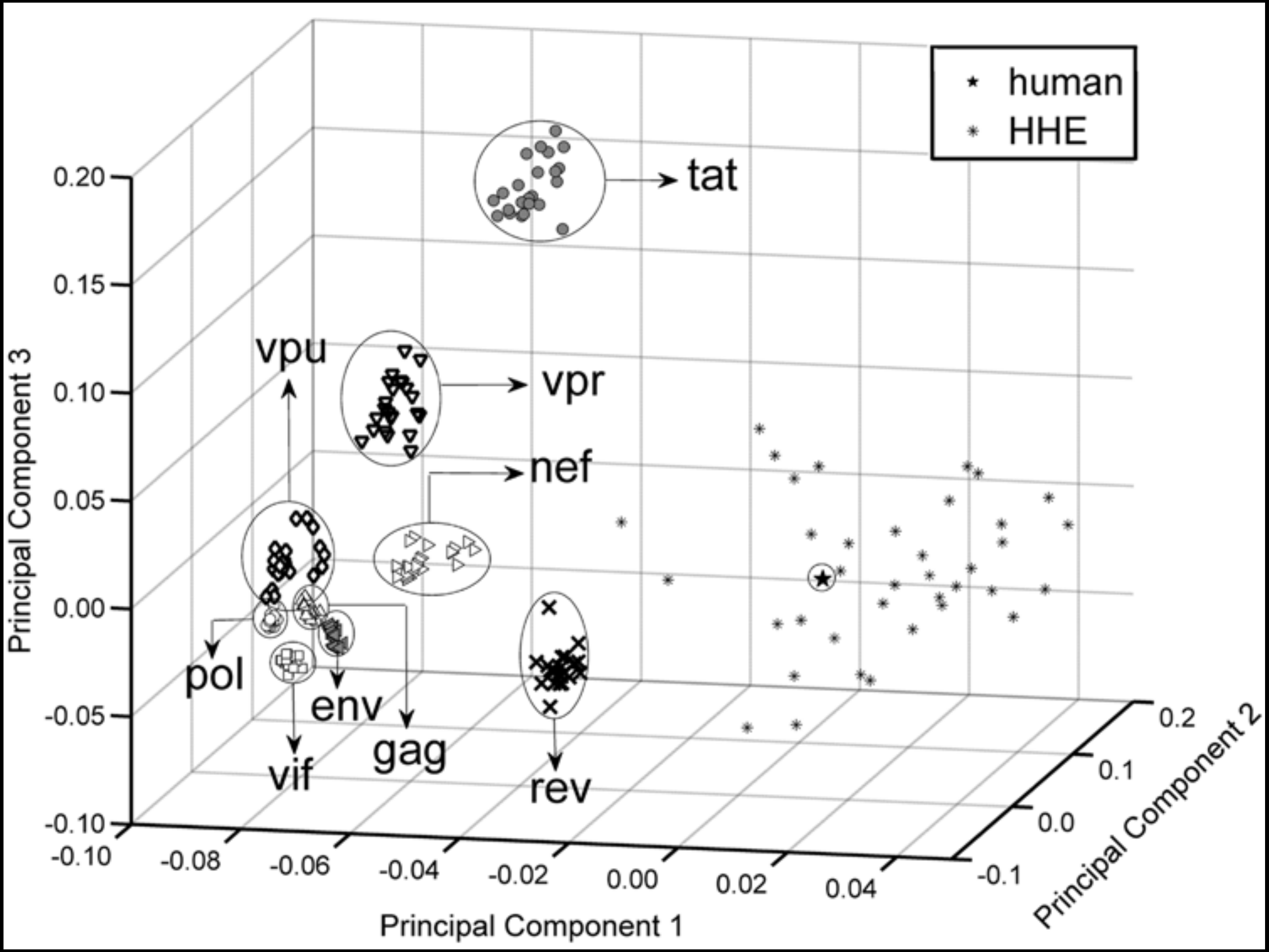
Each cluster contains 23 data points for 23 years

Structural and Regulatory genes show differential cluster compactness



First 3 components in the PCA account for 76.29 % & first 6 components for 90.69 % of variance of the original data

PCA for all HIV-1 genes, average human and Human Highly Expressed genes (HHE)

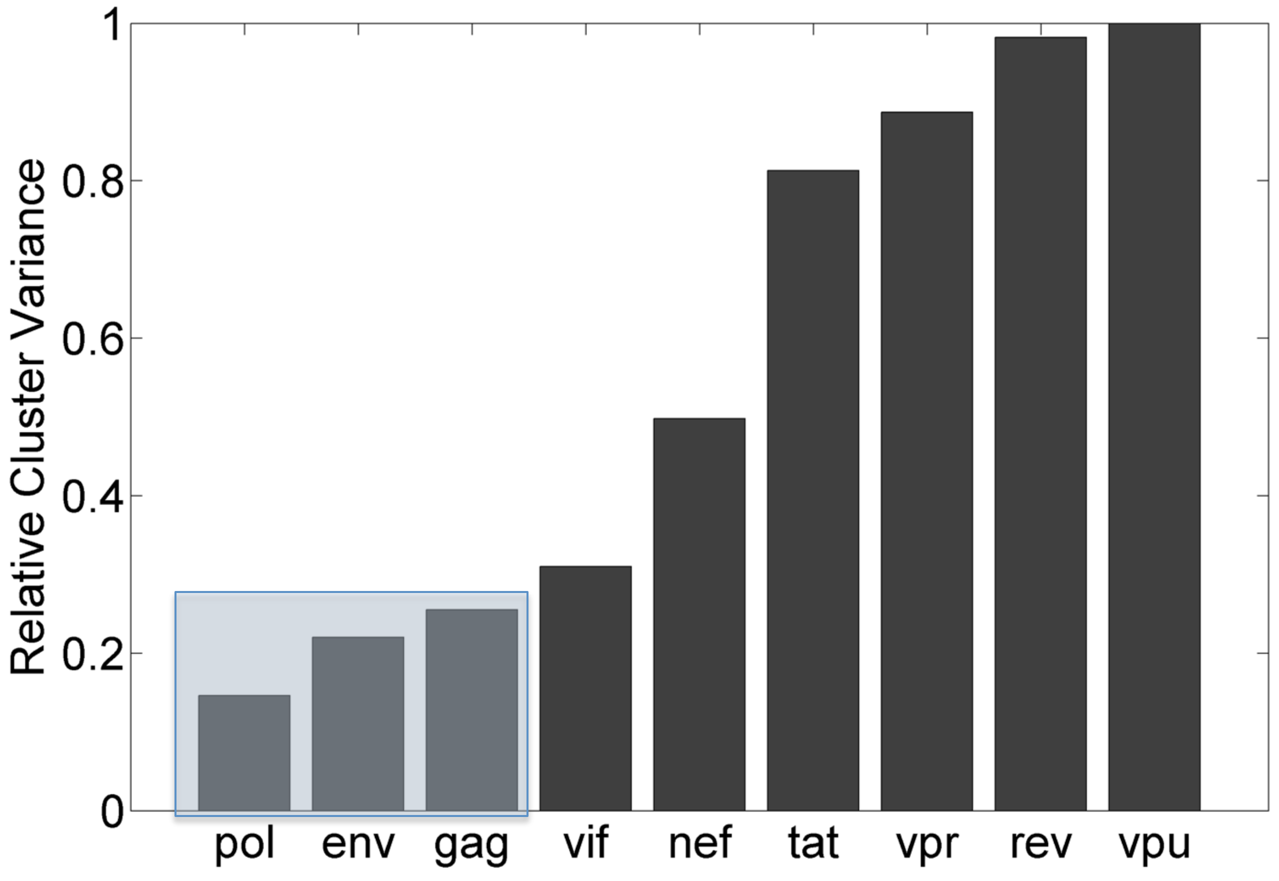


Cluster Compactness

He, et al (2003)

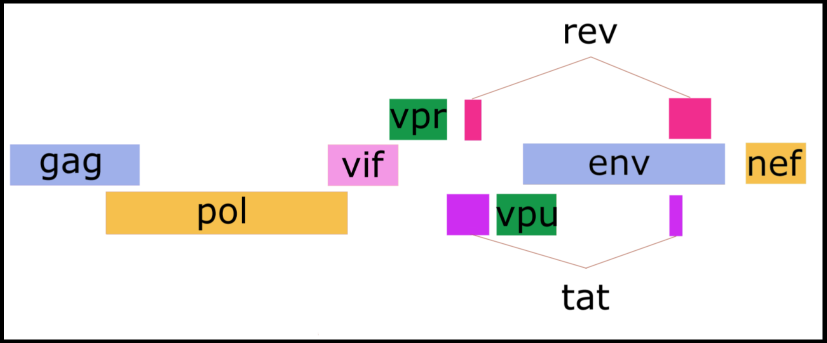
$$v = \sqrt{\frac{1}{T}\sum_{i=1}^{T} d^2(x_i, \overline{x_i})}$$

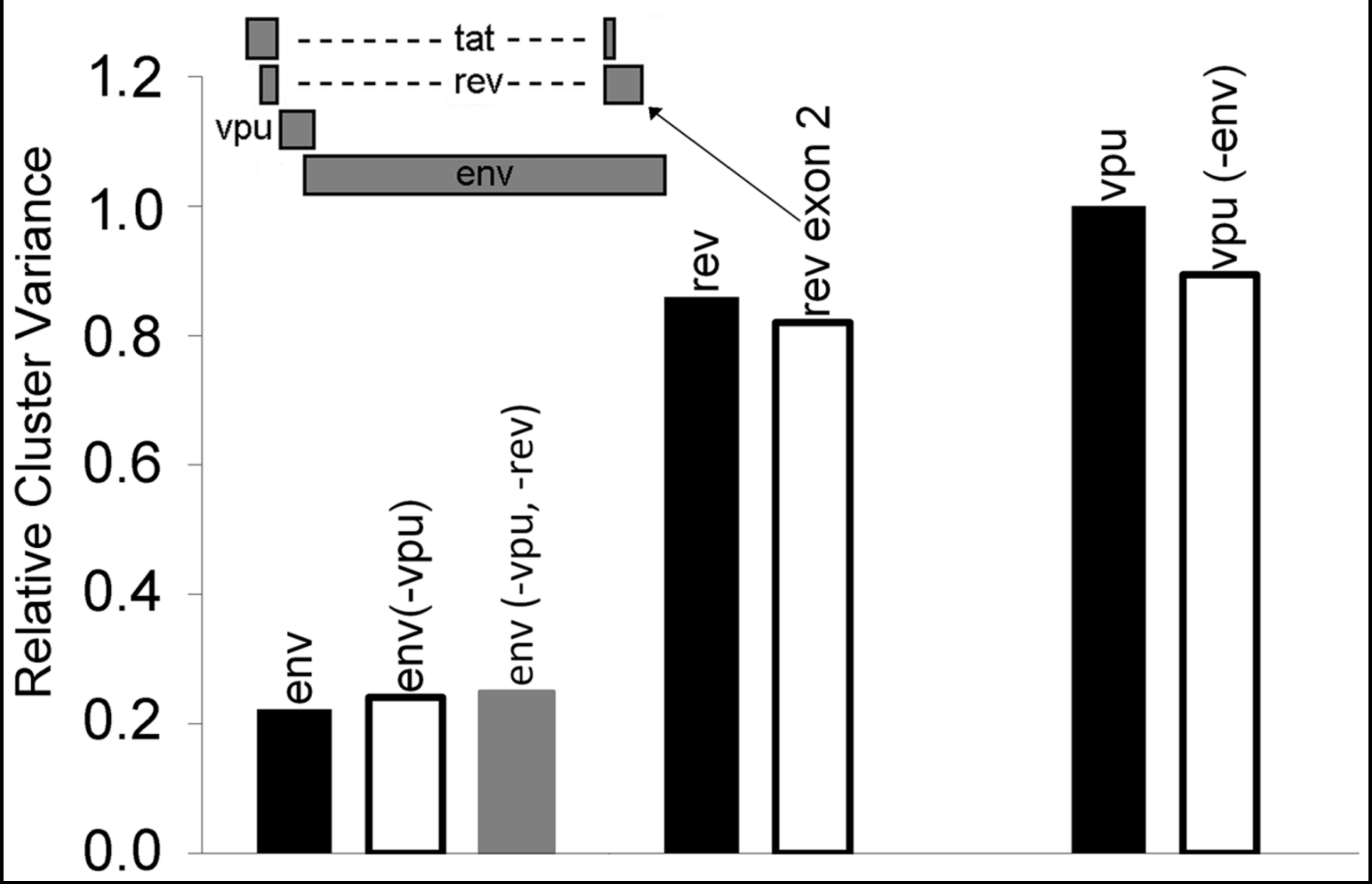
v = variance, T = number of time points,
d = Euclidean metric,
$\mathbf{x_i}$ = value at the ith time point in a cluster
$\overline{x_i}$ = mean value for a cluster



Structural and Regulatory genes show differential cluster compactness

Overlapping Genes and their effect on Cluster Variance





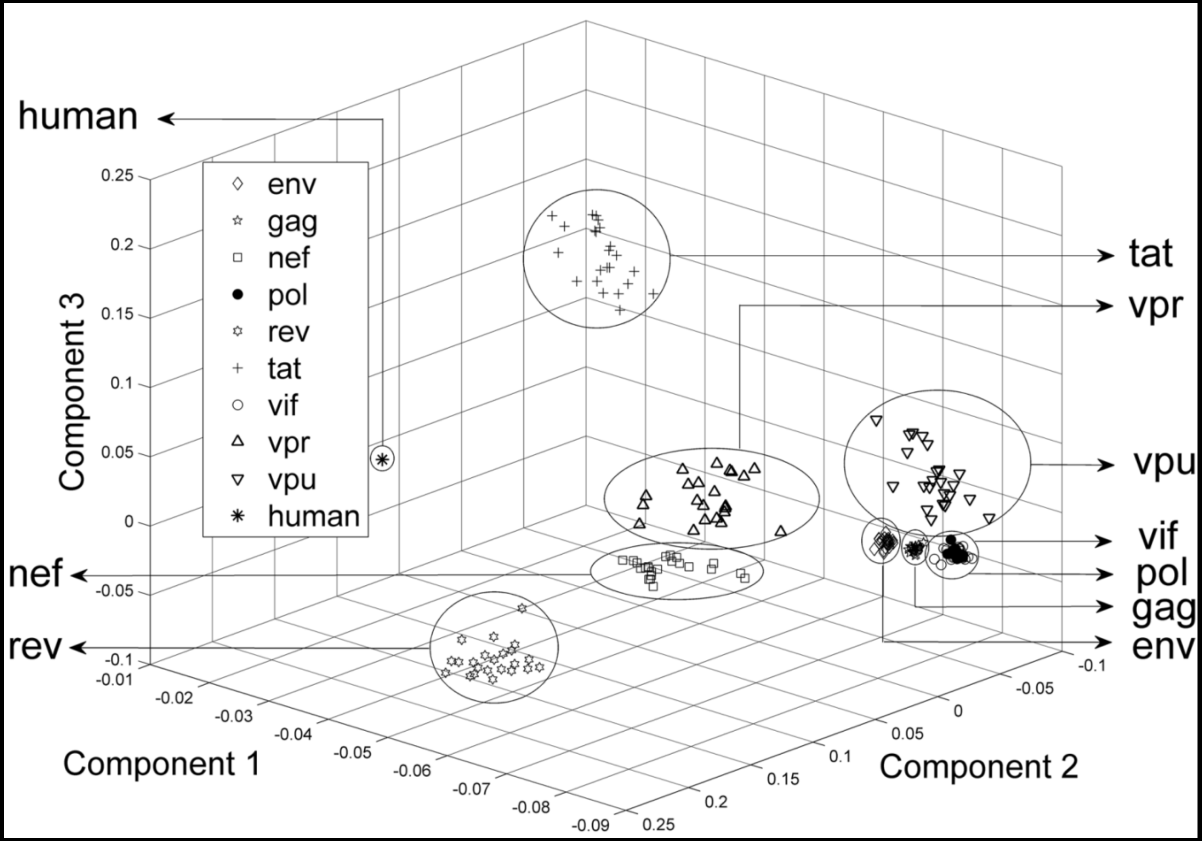
Is there a temporal pattern in the variability observed for the genes in the PCA plot ?

HIV-1 gene sequences from year 1983 to 2005

9 HIV-1 genes separates into 9 different clusters

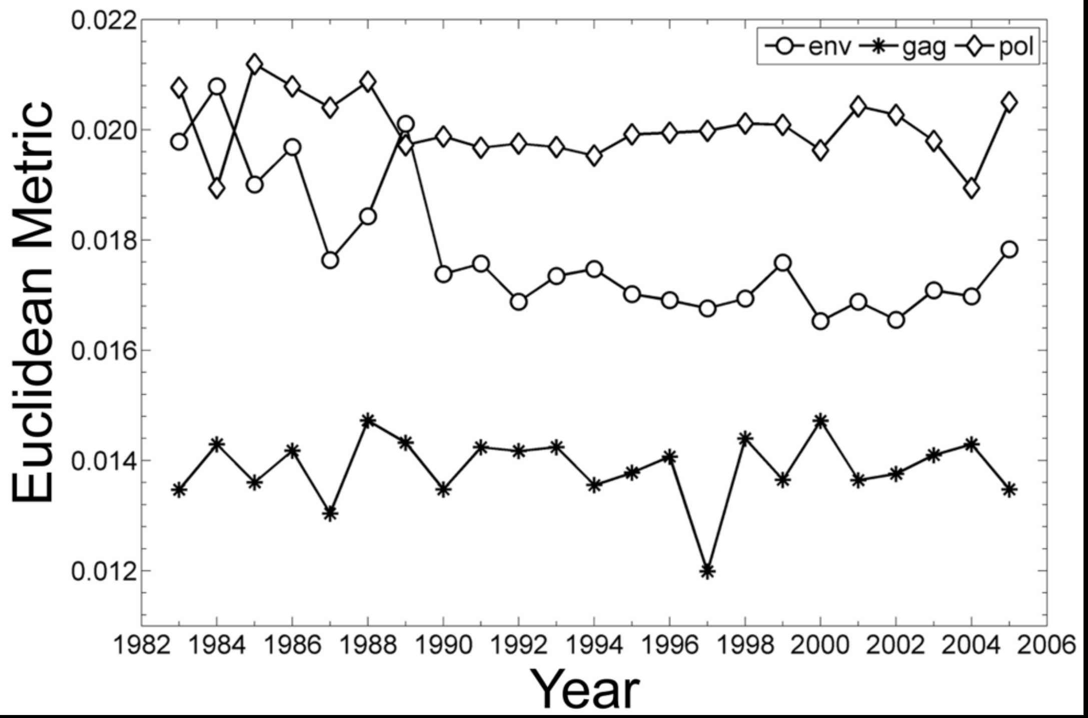
Each cluster contains 23 data points for 23 years

First 6 components account for 90.69 % of variance of original data



$$\text{distance} = \sum_{i=1}^{6} \sqrt{(PC_i^{HIV\ gene} - PC_i^{human})^2}$$

Euclidean metric calculated using the first 6 principal components between each point in a cluster in PCA & average human point

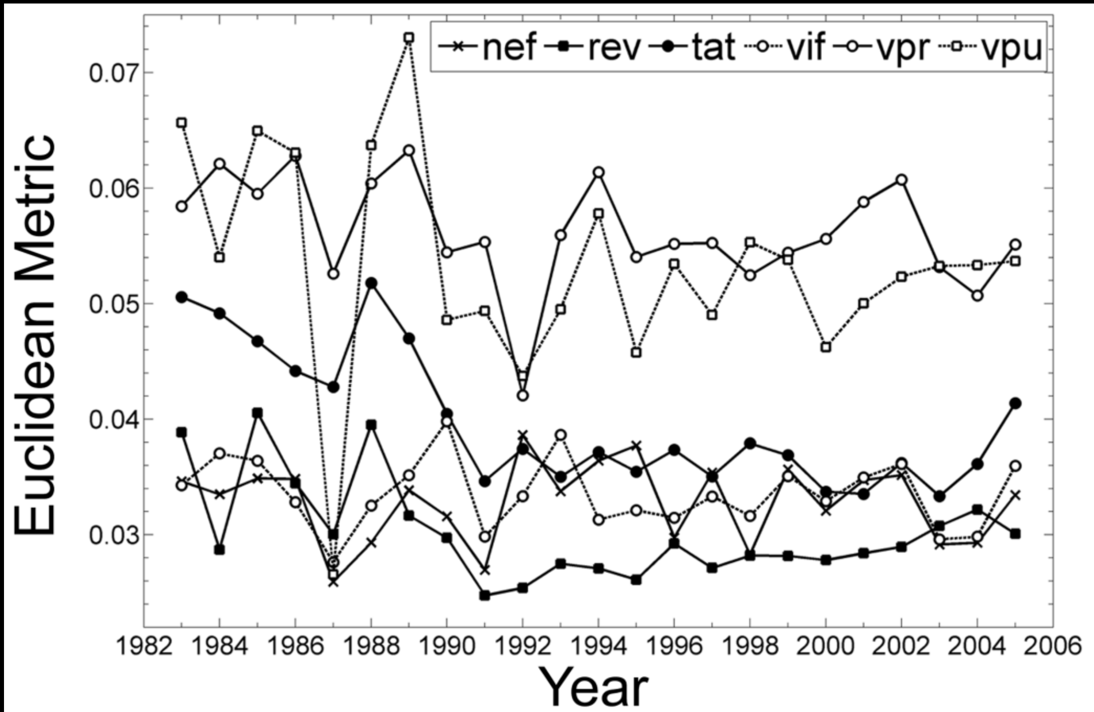


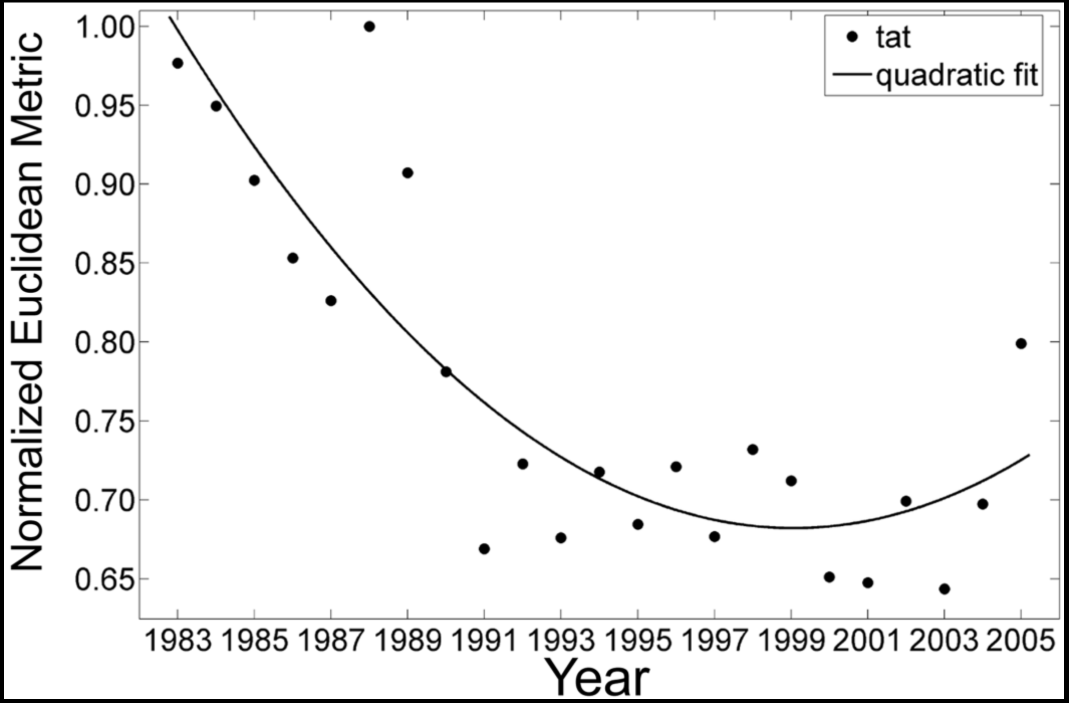
Structural genes (*env*, *gag*, and *pol*) exhibit lower fluctuations and do not show any clear temporal trend over years

Regulatory genes *rev*, *tat*, *vpr*, and *vpu* exhibit decreasing trend with time

Kendall' s rank correlation coefficient (τ) for the genes

tat	$\tau = -0.70, p < 10^{-6}$
vpu	$\tau = -0.47, p < 10^{-3}$
vpr	$\tau = -0.31, p < 0.019$
rev	$\tau = -0.29, p < 0.028$



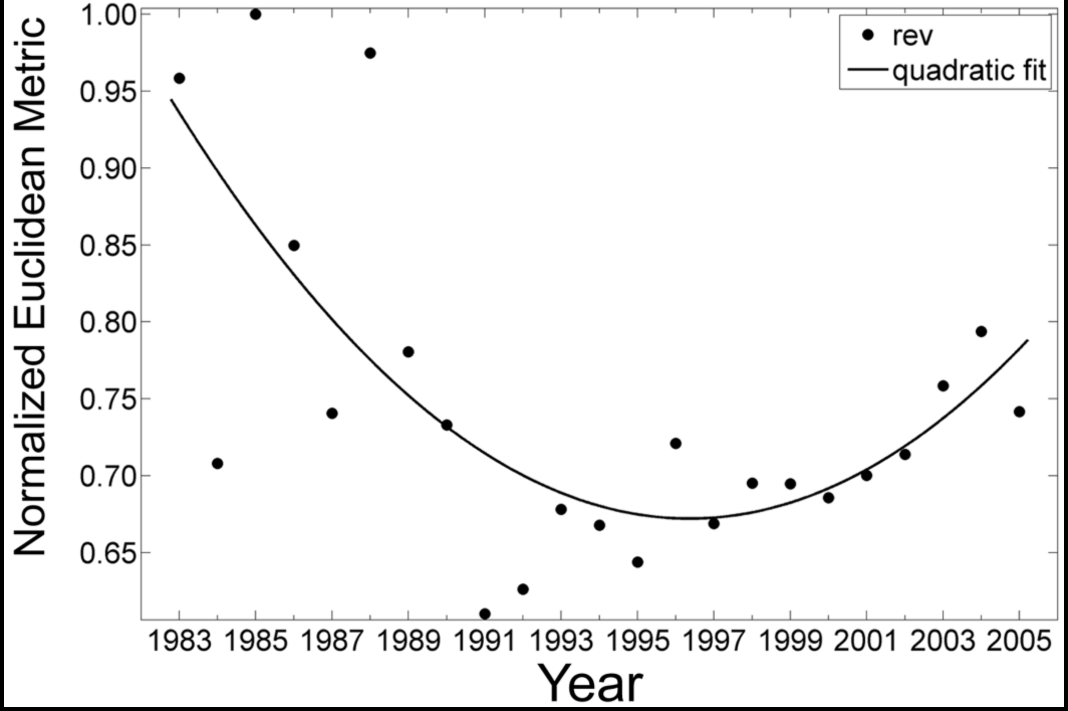


Quadratic fit for ***tat*** (R² = 0.75) shows a clear negative slope for the first 19 years

Reversal of the negative slope after 2000

Quadratic fit for ***rev*** (R² = 0.51) shows a clear negative slope for the first 15 years

Reversal of the negative slope after 1997

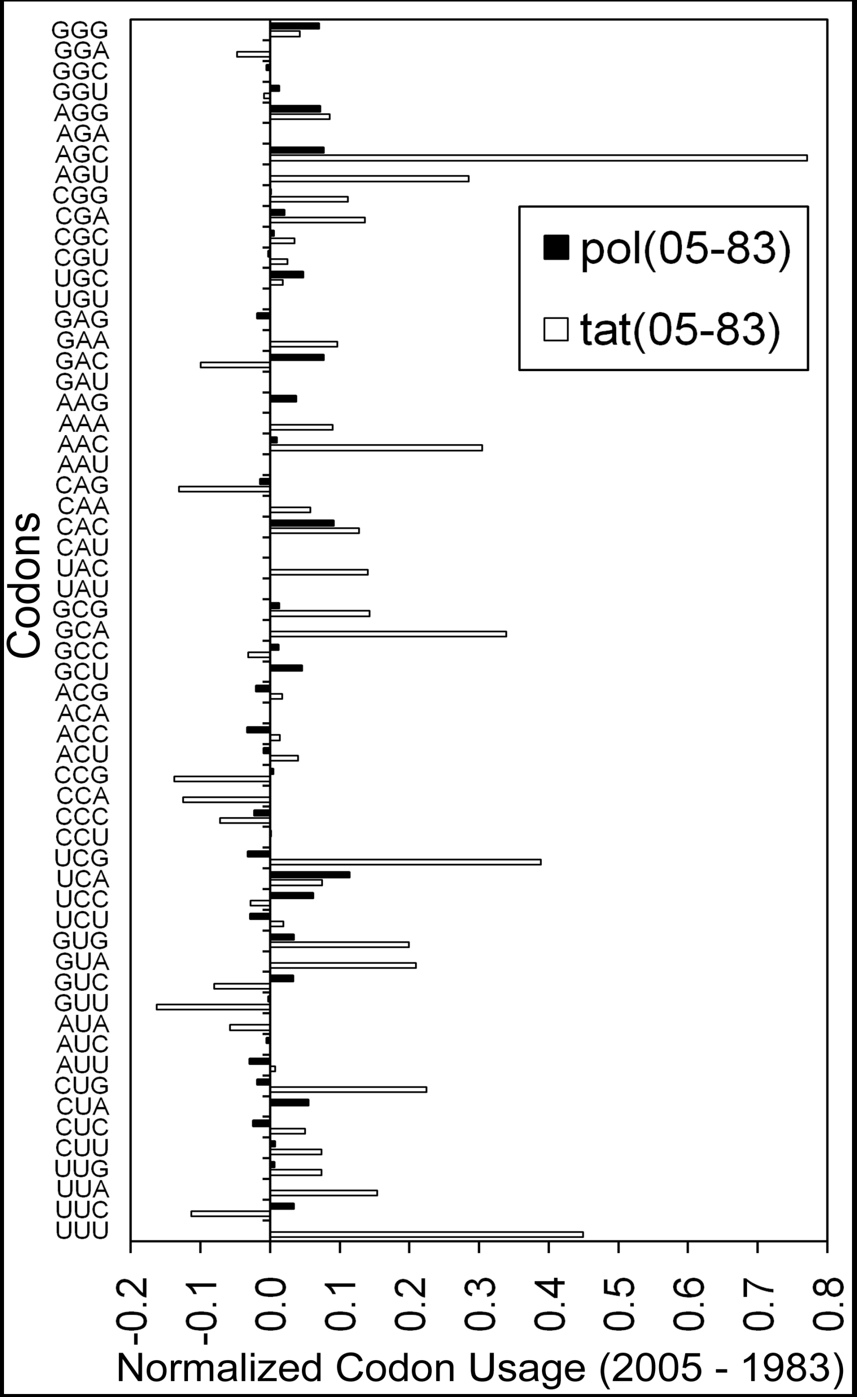


Codon Based Analysis

All genes show synonymous variations in more than 66% of the codons

Regulatory gene *tat* exhibits larger changes in codon frequencies (white bars) **compared to structural gene *pol*** (black bars).

The codons exhibiting more than 10% change are many in the regulatory genes (20 for *tat*), **compared to structural genes** (1 for *pol*)



CONCLUSIONS

- ***Using multivariate statistics on synonymous codon usage values of the nine genes of HIV-1 and its human host,*** **we show presence of temporal change in the regulatory genes of HIV-1 towards host-preferred codons**

- ***Structural genes (env, gag and pol) do not show any temporal pattern***

- ***Regulatory genes show codon usage pattern correlating with the host over time***

- **Synonymous nucleotide changes over time can act as a weak selective force to aid in evolutionary adaptation and differential synonymous codon usage pattern is a method to regulate translation (*Nat Rev Genet 2006; 2011*)**

Differential host-specific adaptation of codon usage patterns in some pathogen genes indicate positive translational selection *(and not by drift alone)* **during inter-host transmission.**

Implications in vaccine development

This study points towards the regulatory genes (*rev, tat*) being likely candidates for developing therapies, and may help in rationalizing design of a more robust and widely applicable HIV therapy

Aridaman Pandit

Utrecht Medical School, The Netherlands

Bette Korber & Tanmoy Bhattacharya

Los Alamos Natl. Lab. and Santa Fe Institute, USA

Partha P. Majumdar,

NIBMG & Indian Statistical Institute Kolkata, India

Peter Wall Institute of Advanced Studies

International Visiting Research Scholar

Leah Keshet, Maths

PETER WALL
INSTITUTE FOR ADVANCED STUDIES
THE UNIVERSITY OF BRITISH COLUMBIA VANCOUVER

THANK YOU