



Selection in HIV-1 gp41 in acute infection

AIDS Vaccine Conference
2008

Gama Bandawe



gp41

- **gp120** and **gp41** have functionally distinct but additive roles in HIV infection and pathogenesis (Meissner et al, 2005)
- **gp120** mediates transmission via **CD4 binding** while **gp41** is essential for post CD4 binding events including **viral fusion** and **assembly** (Lambelet et al 2006; Martin et al. 1996; Mobley et al 1999; Charloreaux et al 2006; Dey et al 2006)
- **Post CD4** binding (and gp41 related) processes are among the most significant determinants of **replicative capacity** and **pathogenic potential** in any given strain (Hsu et al, 2003)

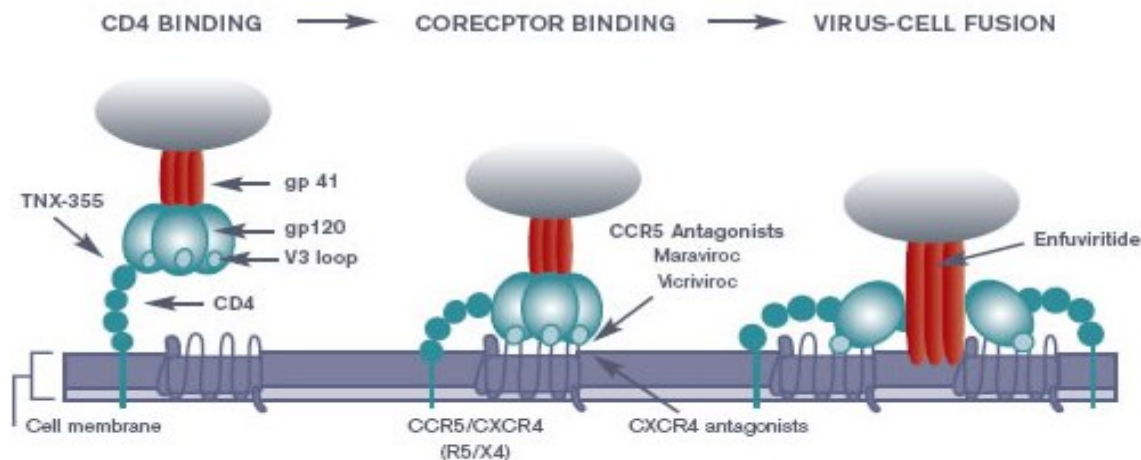
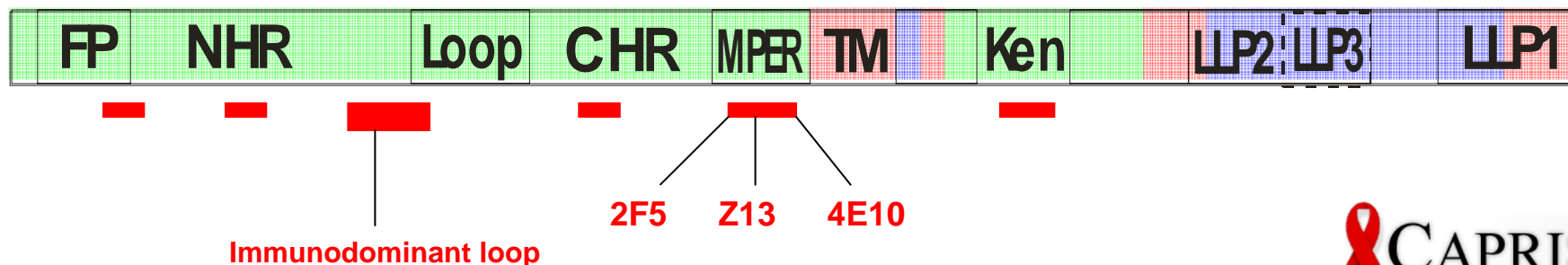


Figure adapted from Doms R. et al Genes Dev. 2000. 14:2677-2688

Gp41 in vaccines

- Gp41 is **conserved** relative to gp120 with fewer indels
- It is one of the **earliest** targets of the immune response
- Several regions of gp41 elicit **Ab reactivity**.
- Of the **6** known **broadly neutralizing** Abs **3** are found in **gp41** (2F5, Z13 and 4E10)



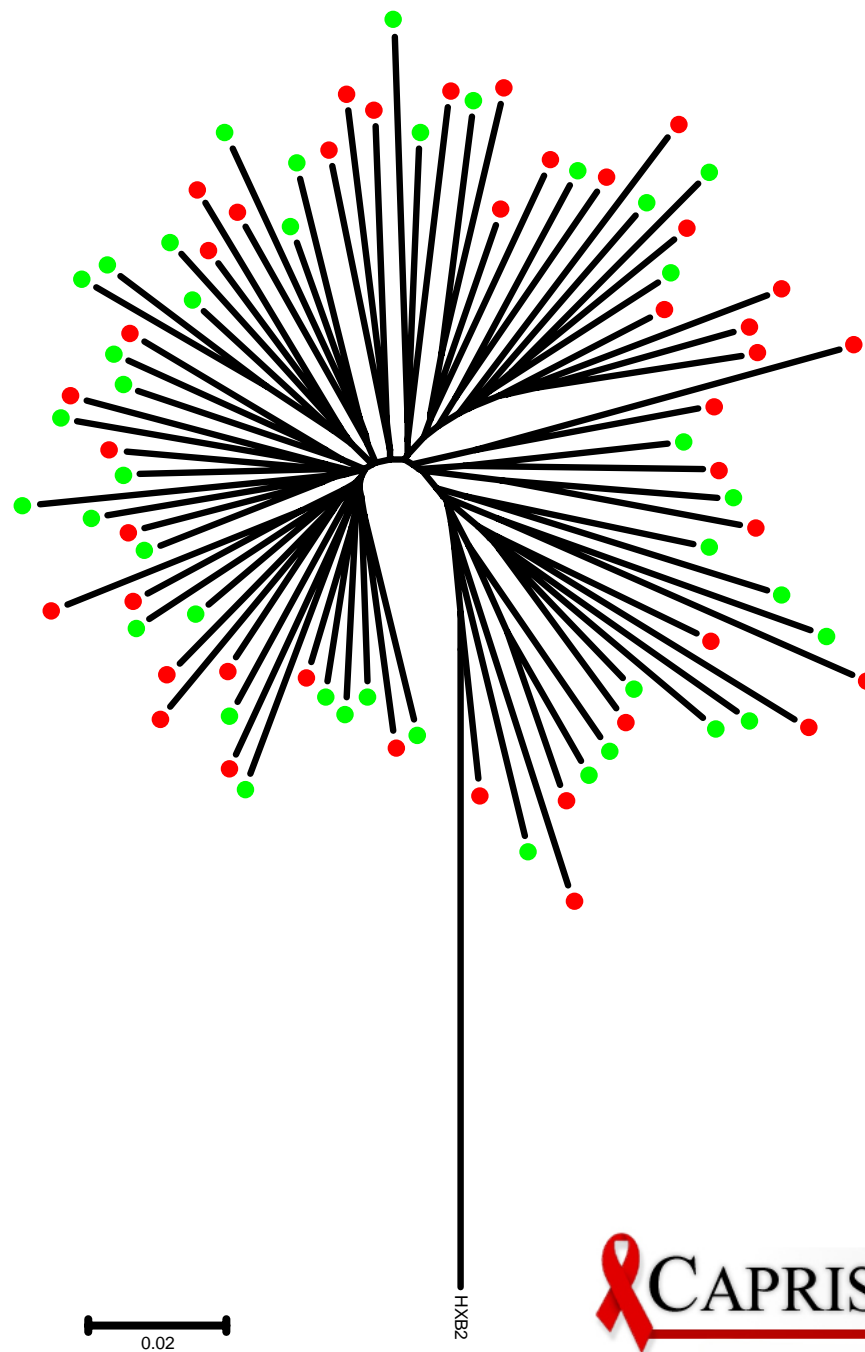
Study aims

- Is it possible to detect **signals** associated with **transmission** separate from host specific adaptation?
- **Which sites** are associated with this sort of adaptation?

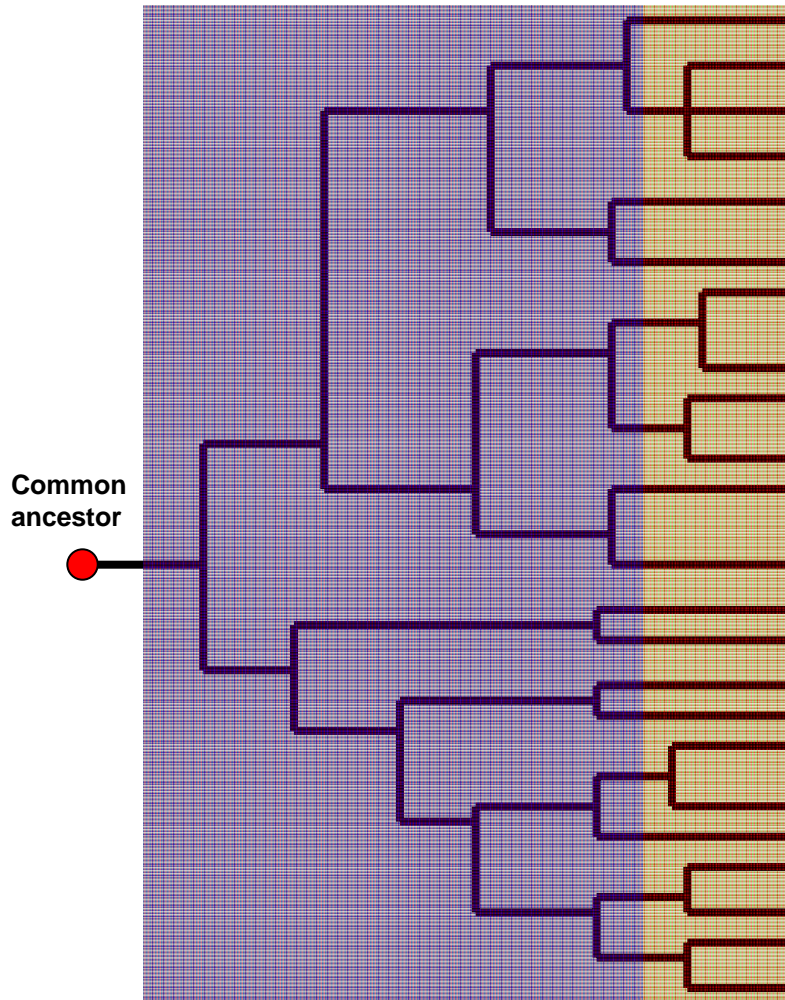
Subtype C datasets

40 CAPRISA enrolment gp41 sequences sampled at a mean 40 days post infection. This is the **Acute/early** Infection (**AI**) dataset

40 **chronic** HIV-1 subtype C sequences matched for gender and geographic location from the Kiepiela et al (2004) study. CD4 counts <200 and viral loads >200000 were excluded. This constituted the chronic infection (**CI**) dataset



Methods of inference of selection



Single Likelihood Counting

SLC

Direct counting of substitutions along phylogeny from ancestral state
Using substitution rates on a site by site basis
Internal branches of the likelihood tree are considered

Uses likelihood ratio tests
No assumption of distribution of rates
This removes the effect of recent and possibly transient SLC
Most substitutions (on terminal branches)

Good for large datasets
Computationally intense
therefore detects population level signals of selection
Lacks power with small or low diversity datasets

All methods are able to accommodate recombination by creating separate trees between potential breakpoints

Adaptive Evolution Server @ Datamonkey.org - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Refresh Mail Print Word Pad Notepad

Address <http://www.datamonkey.org/> Go Links

ANALYZE YOUR DATA HOME HELP CITATIONS JOB QUEUE STATS HYPHY PACKAGE DATAMONKEYS WEBCOMIC



DATAMONKEY

RAPID DETECTION OF POSITIVE SELECTION

a Web-Server of the HyPhy Package

Welcome to the free public server for detecting signatures of positive and negative selection from coding sequence alignments using state-of-the-art statistical models. This service is brought to you by the viral evolution group at the Antiviral Research Center of the University of California, San Diego. The methods and software tools are developed and maintained by [Sergei L. Kosakovsky Pond](#), [Simon Frost](#) and [Art Poon](#).

March 17th, 2008: [Spidermonkey](#) - a new tool for discovering evolutionary dependancies between sites in a coding alignment using Bayesian Graphical Models (BGMs) has been added to Datamonkey. The SpidermonkeyBGM analysis option is available for non-recombinant coding alignments of up to 150 sequences.

Datamonkey.org can help you answer the following questions ([publications citing datamonkey.org](#)):

Which codon sites are under positive or negative selection?

Three different codon-based maximum likelihood methods, [SLAC](#), [FEL](#) and [REL](#), can be used estimate the dN/dS (also known as Ka/Ks or ω) ratio at every codon in the alignment. An exhaustive discussion of each approach can be found in the [methodology paper](#). All methods can also take [recombination into account](#). This is done by screening the sequences for recombination breakpoints, identifying non-recombinant regions [GARD tool](#) and allowing each to have its own phylogenetic tree.

Is there evidence of selection in my alignment?

The [PARRIS](#) method, developed by [Konrad Scheffler and colleagues](#), extends traditional codon-based likelihood ratio tests to detect if a proportion of sites in the alignment evolve with $dN/dS > 1$. The method takes recombination and synonymous rate variation into account.

Which codon sites are under positive or negative selection at the population level?

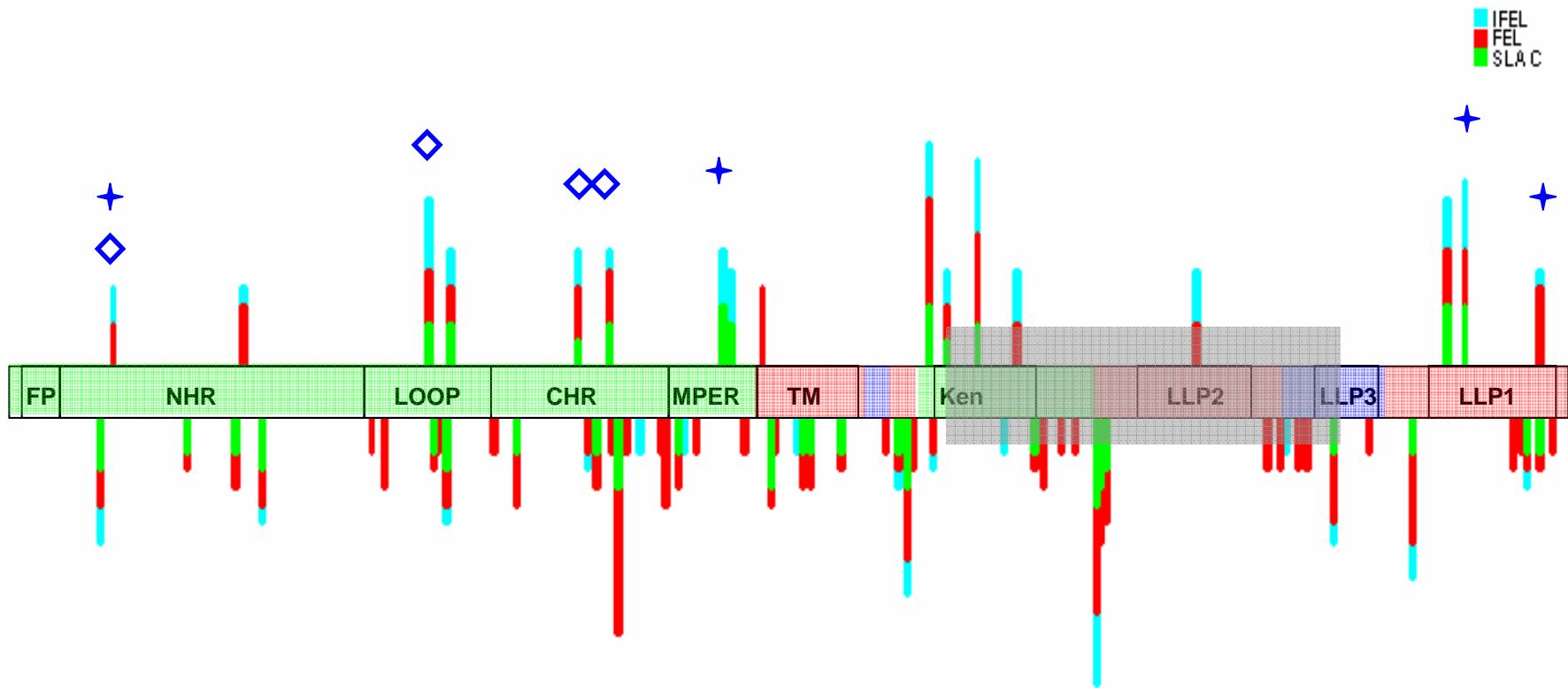
The codon-based maximum likelihood [FEL](#) method can investigate whether sequences sampled from a population (e.g. viral sequences from different hosts) have been subject to selective pressure at the population level (i.e. along internal branches). A discussion of the method and its

Done

start | Inbox - Microsoft Out... | 3 Internet Explorer | www.datamonkey.or... | Program Files

Test of the methods

- Are the methods used **reliable**?
- How do sites detected by these methods **compare** to sites described previously?



7 We used 8 datasets of identified purifying selection was very pervasive and often gave stronger signal from different subtypes and CRFs selection in multiple subtypes



Rev exon 2 overlapping reading frame



external



membrane



internal



Choisy et al 2003



Travers et al 2005

Hypothesis

- Given the largely **overlapping phylogenetic history**, the two **subtype C** datasets should share more selection signals than other subtypes
- sites where the selection signal is strongest in the **terminal branches (intra-host adaptation)** should be, less conserved between the **AI** and **CI** datasets than those detectable throughout the tree
- Differential selection detected in terminal branches of **AI** and **CI** datasets should be indicative of **different selection pressures** related to these stages.

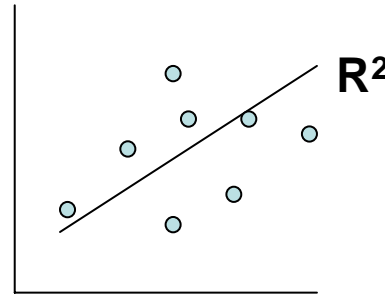
Normalized dN/dS

AI	CI	Sub A
0.0024	0.0006	0.0051
0.0006	0.0071	0.0468
0.0071	0.0034	0.0321
0.0034	0.0294	0.0006
0.0294	0.0051	0.0071

All signals

AI	CI	Sub A
0.0024	0.0189	0.0013
0.0006	0.0376	0.0291
0.0071	0.0051	0.0059
0.0034	0.0468	0.0107
0.0294	0.0321	0.0084

Subset of signals



Regression analysis

(1-R)

	AI	CI	Sub A	Sub B
AI	0			
CI	0.0006	0		
Sub A	0.0071	0.0051	0	
Sub B	0.0034	0.0468	0.0107	0
Sub D	0.0294	0.0321	0.0084	0.0321

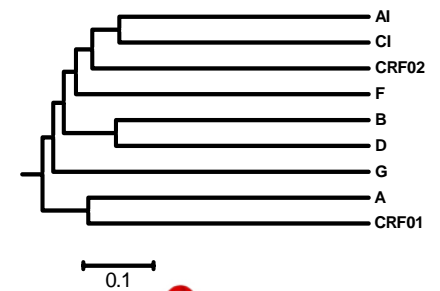
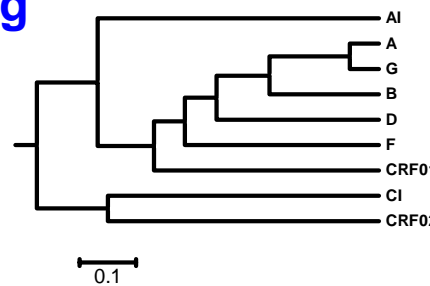
Matrix or R-1 values



UPGMA or NJ clustering algorithm

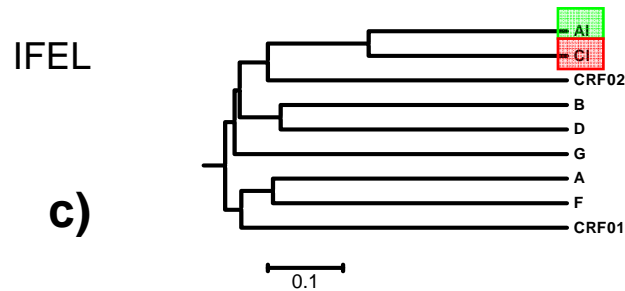
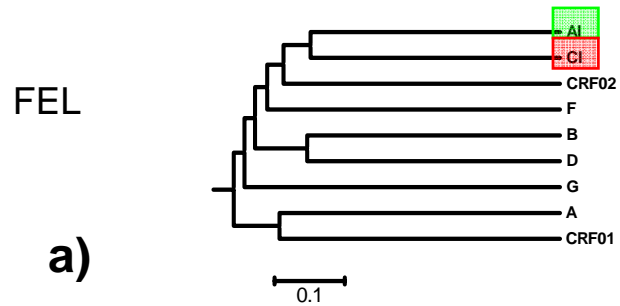


For all 9 datasets we took dN/dS ratios from sites with any significant evidence ($p < 0.05$) of either positive or negative selection. We then used UPGMA or NJ clustering algorithms to join and visualize the relatedness of between datasets from all datasets.

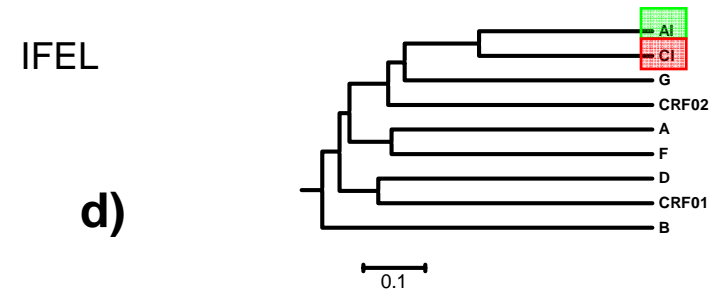
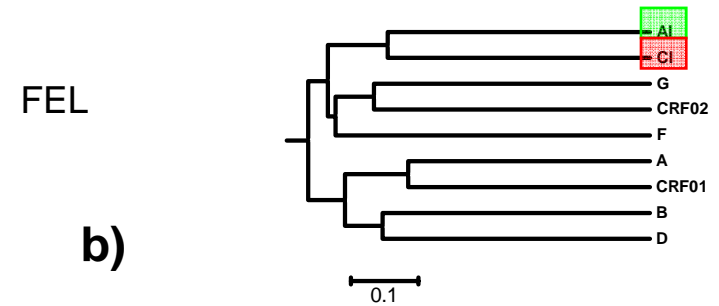


UPGMA dendograms of regression analysis of all selection signals

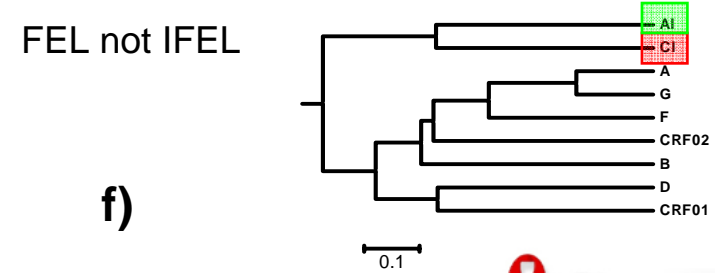
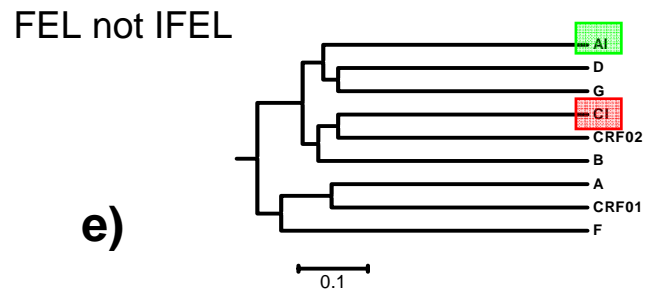
Purifying selection signals



Diversifying selection signals



Differences detected only on terminal branches



sites evolving under significantly different selection between AI and CI

- We took all sites where there was any **significant** evidence of different types of selection happening between **AI** or **CI**
- And only sites where these differences were detectable by **FEL and not IFEL** (where the difference was in the terminal branch)
- We confirmed differences in selection using **PARRIS**, a **random effects likelihood** test that assigns a **posterior probability** of the **type of selection** at that site

sites evolving under significantly differential selection between AI and CI

Codon position (HXB2 gp160)	Normalised dN/dS (SLAC, FEL, IFEL) ^a			PARRIS	Gp41 domain
	AI	CI			
43 (554)	-1.62, -0.36, -0.37	0.50, 0.1, 0	no	N-heptad repeat	
48 (559)	-3.43, -0.71, -0.7	0.43, 0.1, 0	no	N-heptad repeat	
105 (616)b	-1.62, -0.36, -0.37	0.49, 0.03, 0	YES	Loop region (glyc site)	
163 (674)b	-0.07, -0.04, 0.48	2.88, 0.19, 0.44	YES	MPER (4E10 epitope)	
300 (804)b	0.22, 0.01, -0.04	-1.62, -0.16, -0.18	YES	LLP3	
309 (813)b	-0.24, -0.05, -0.04	2.14, 0.19, 0.24	YES	LLP3	

mutations at site **43** are associated with **resistance to** **SI** **diffusion** **through** **the** **envelope** **trimer** **core** **during** **fusion**
 Site 43 is a member of the N-glycosylation site 43, which is a highly conserved site in the gp120 glycoprotein and is essential for fusion of the envelope trimer core during fusion

Conclusions

- We show that variations in the **selective pressures** acting on viruses during the acute and chronic stages of infections are **detectable** by **comparing** sequences sampled during these infection phases
- Our analysis reveals clear **differences** in the distributions of sites evolving under positive and negative selection in **chronic** and **acute** subtype C infections
- 4 of 6 the gp41 residues evolving under stronger **purifying selection** during **acute infection** are involved in **fusion** or **transmission** related functions
- This is suggestive of a **selective processes** such as a **transmission sieve**

Conclusions (cont'd)

- This may be a **useful** means of identifying **viral genetic features** that are important for **transmission** or **early infection** and thus relevant to rational **vaccine** or **microbicide** design
- We aim to further to establish the **biological** significance of identified sites **experimentally**

Acknowledgements

- Darren Martin
- Florette Treurnicht
- Zenda Woodman
- CAPRISA Acute Infection Study team and participants
- CAPRISA sequence assembly Pipeline developers and technicians
- Sergei Kosakovsky Pond
- CBIO@UCT
- Carolyn Williamson