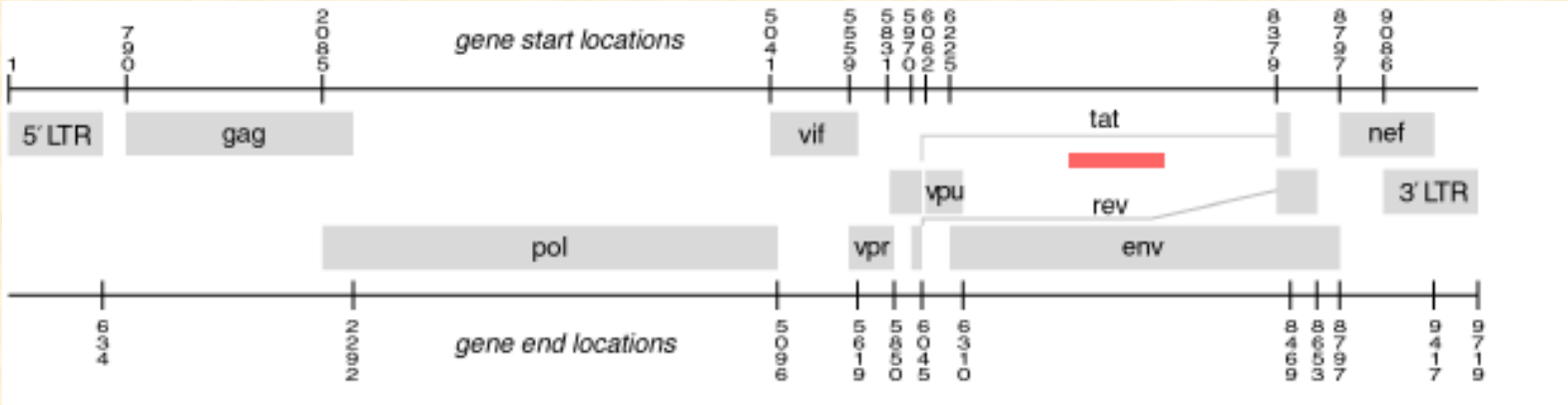
Modeling and Quantification of HIV-1 and SIV Intrahost Evolution

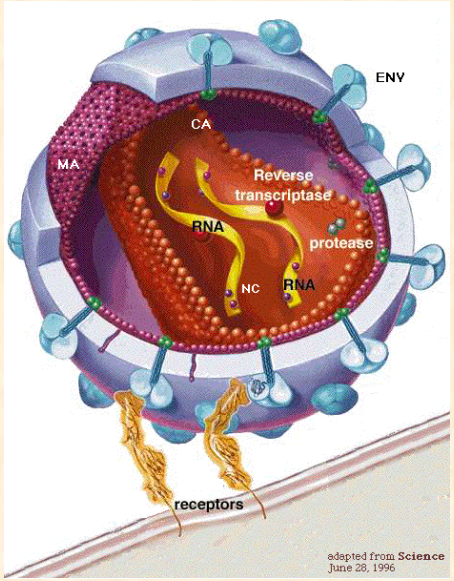
Ha Youn Lee

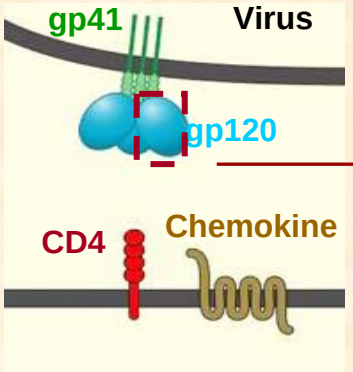
Department of Biostatistics and Computational Biology

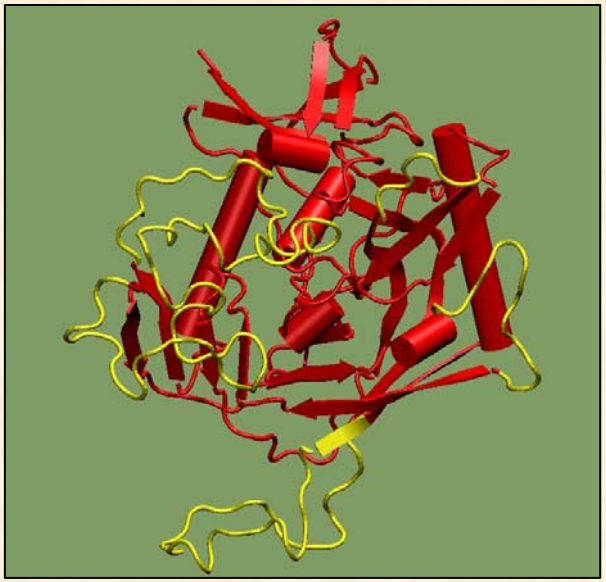
University of Rochester

HIV-1 Envelope Gene Evolution

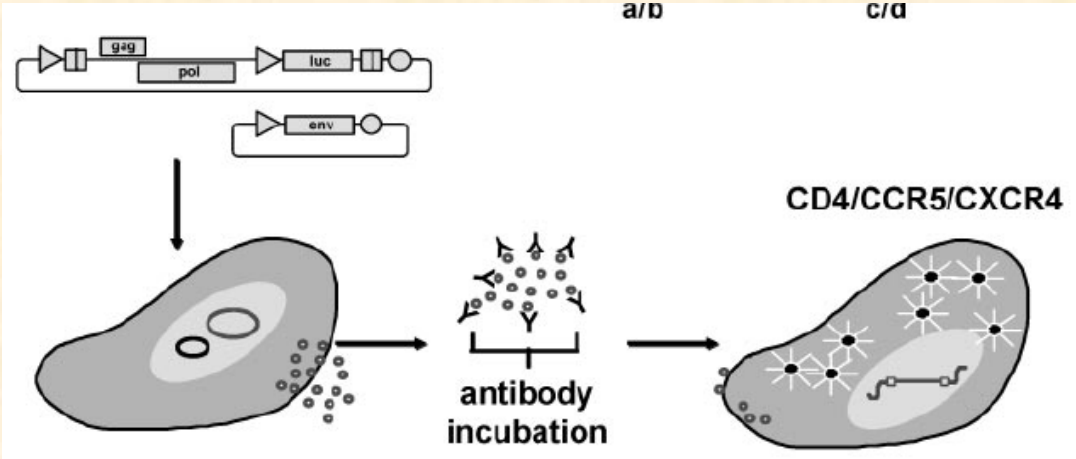




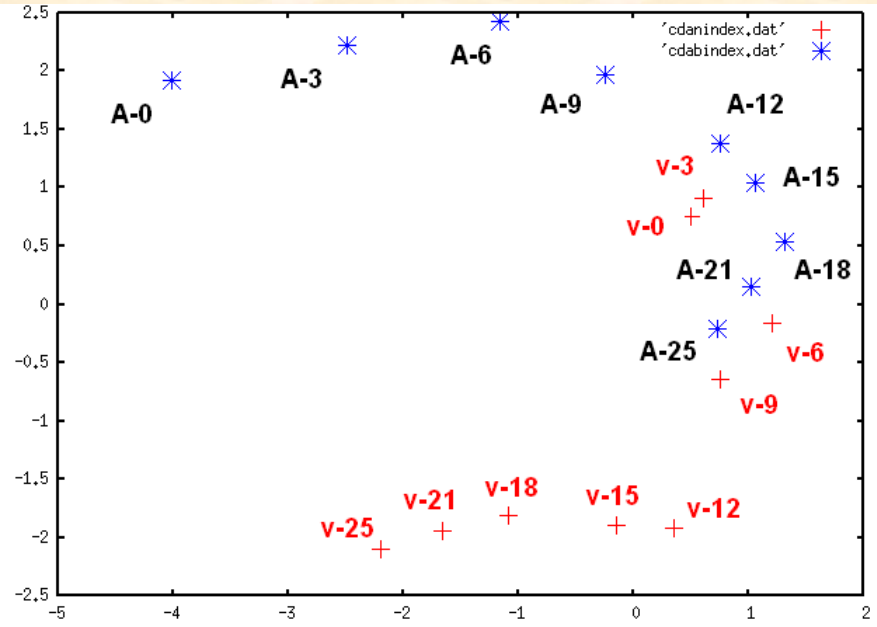




Antibody Neutralization Assay

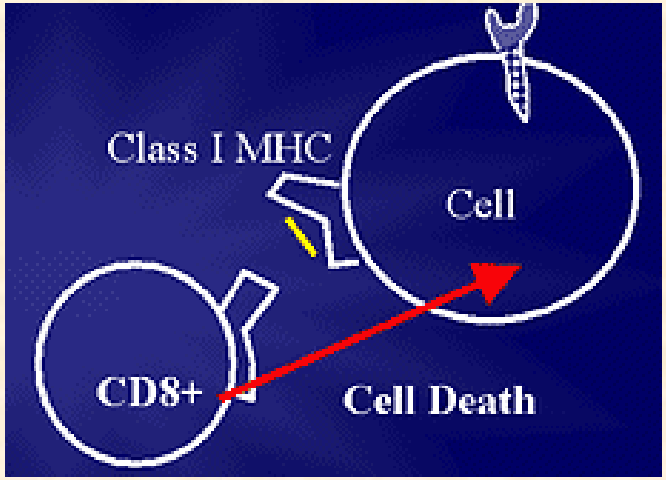


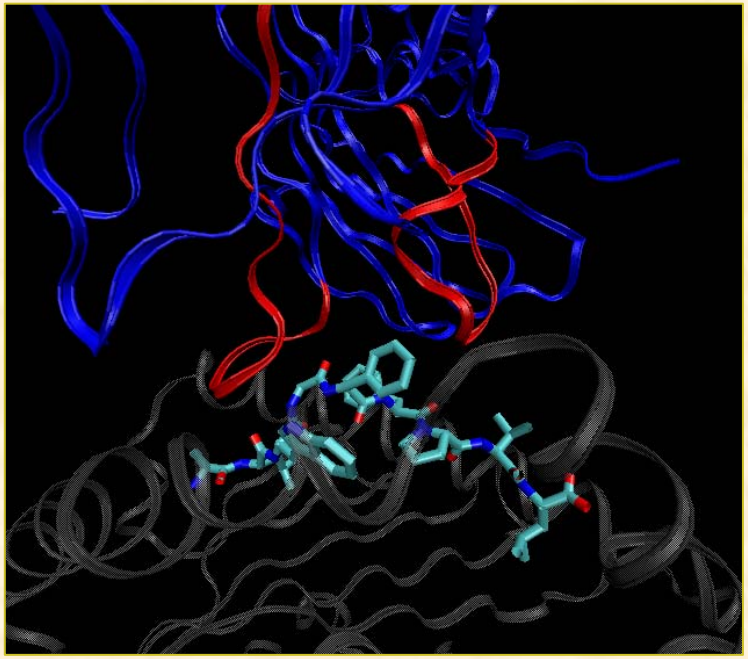
Virus, months	Plasma, months 0	3	6	9	12	15	18	21	25
0	**26**	219	675	1403	2670	2089	2190	2363	2411
3	29	**179**	1024	2151	3733	3152	2808	2953	3086
6	27	35	**78**	358	1769	1939	2247	3112	4345
9	36	67	82	**200**	795	1078	1371	2208	3375
12	19	48	36	64	**76**	166	556	937	1407
15	29	43	64	76	90	**119**	374	721	1234
18	42	65	61	152	117	134	**122**	289	526
21	41	66	82	84	85	113	78	**107**	296
25	42	62	56	62	85	77	55	61	**95**
Controls									
NL43	17	138	294	956	1172	953	1584	1868	2143
JRCSF	24	37	35	60	87	97	105	152	209
AMPHO	<10	32	14	13	14	13	<10	<10	31



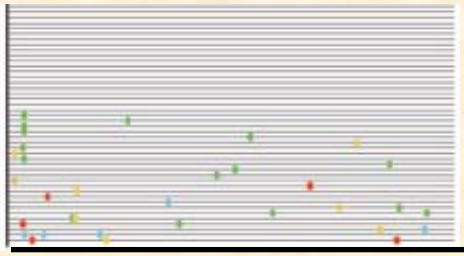
[Richman et al, PNAS (2003)]

CD8-TL Selections

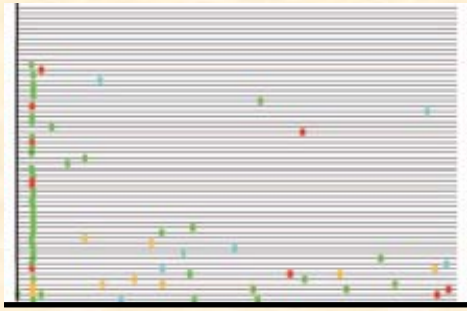




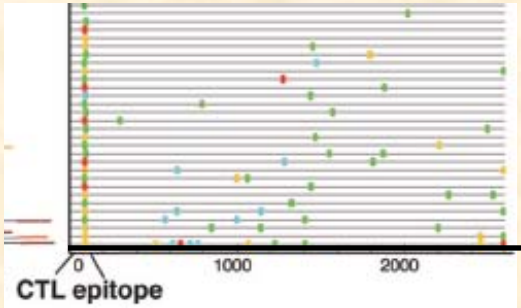
Viral Ramp-Up



+8 days



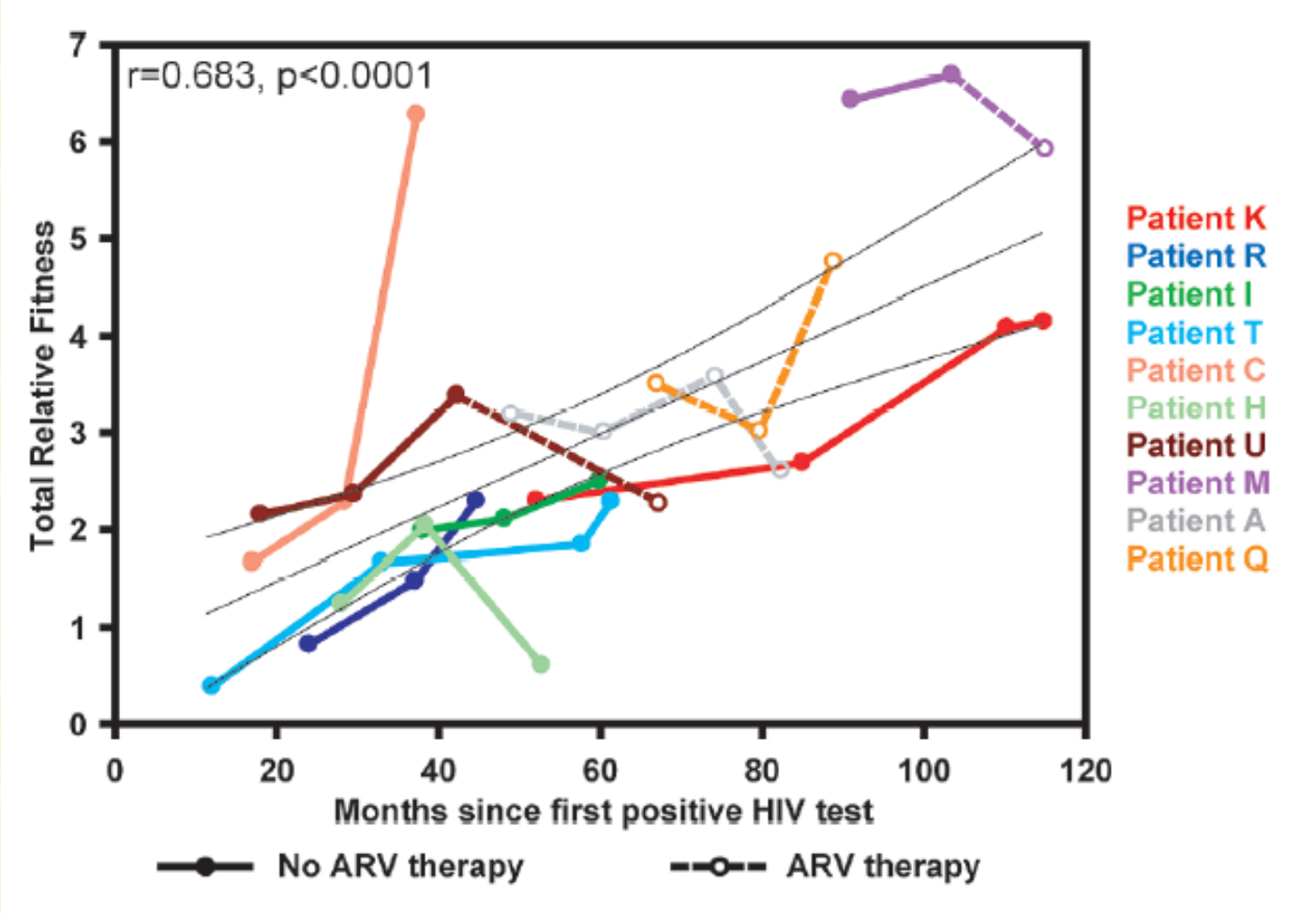
+15 days



+29 days

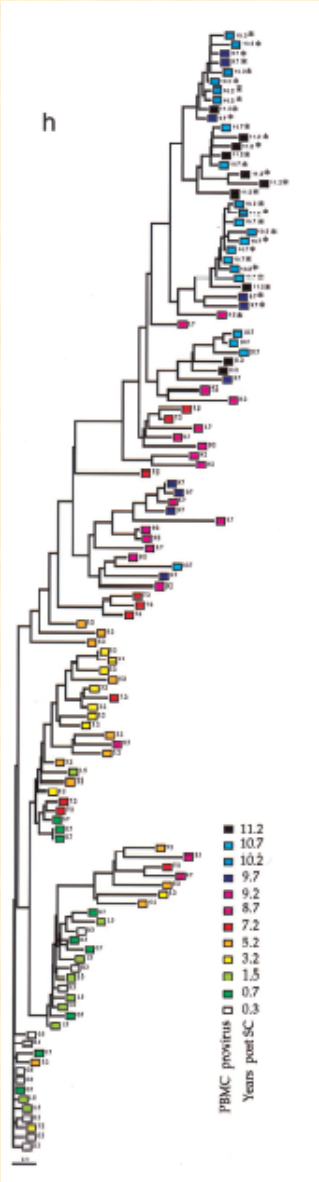
[Keele et al PNAS (2008)]

HIV-1 Competition Assay



[Troyer et al, J. Virol (2005)]

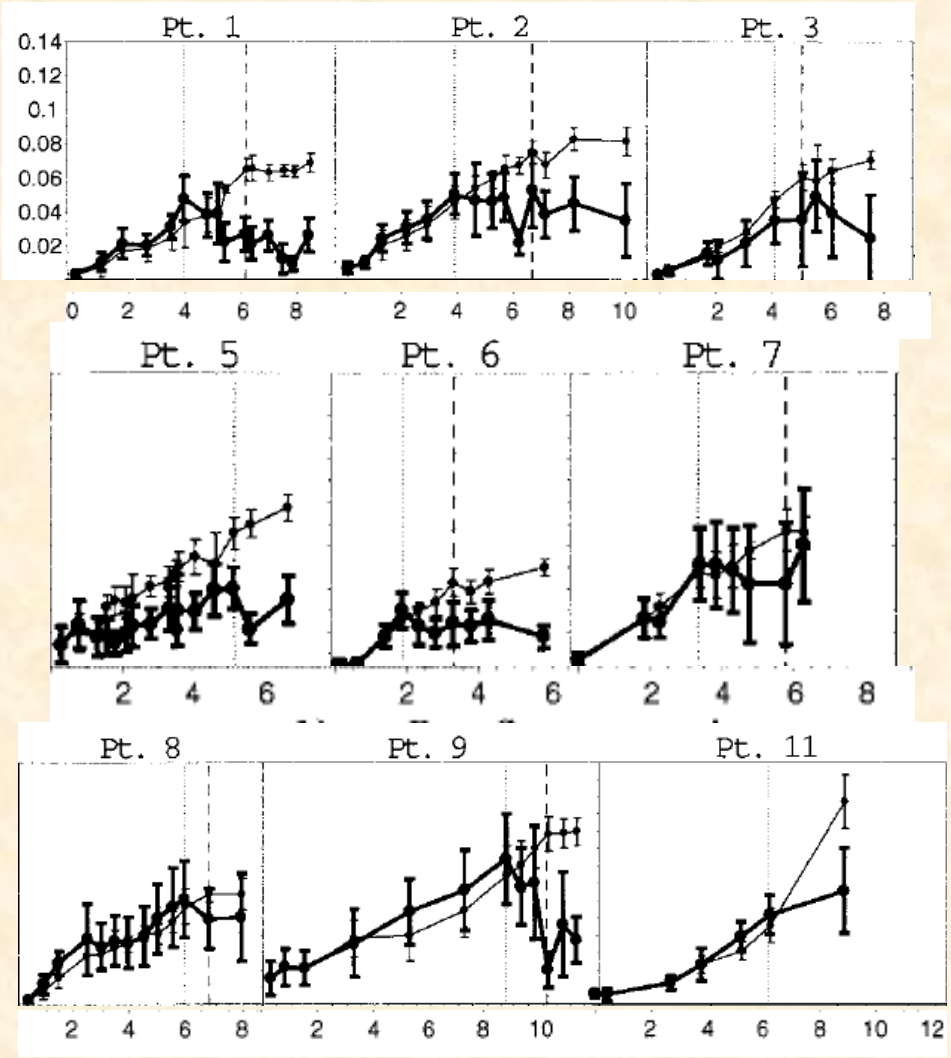
C2-V5 *env* of
9 patient data for 6-12 years



[Shankarappa J. Virol 1999]

Divergence

average distance
from the founder virus



Diversity

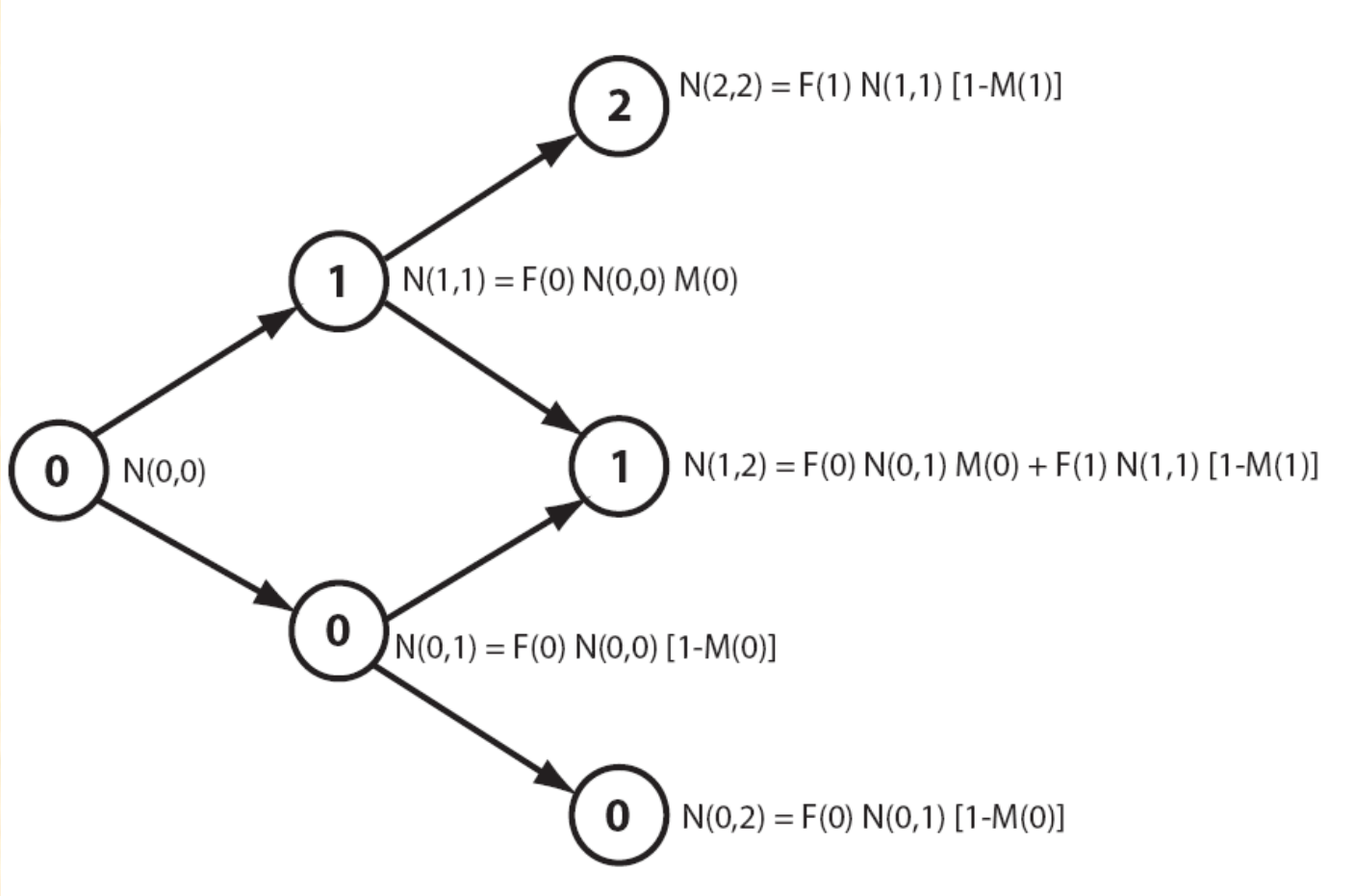
average distance among viruses
sampled at the same time point

Intrahost Sequence Evolution Model

Each virus is represented by the distance (d) from the founder strain

Fitness F(d)

Fraction of Mutant Offsprings M(d)



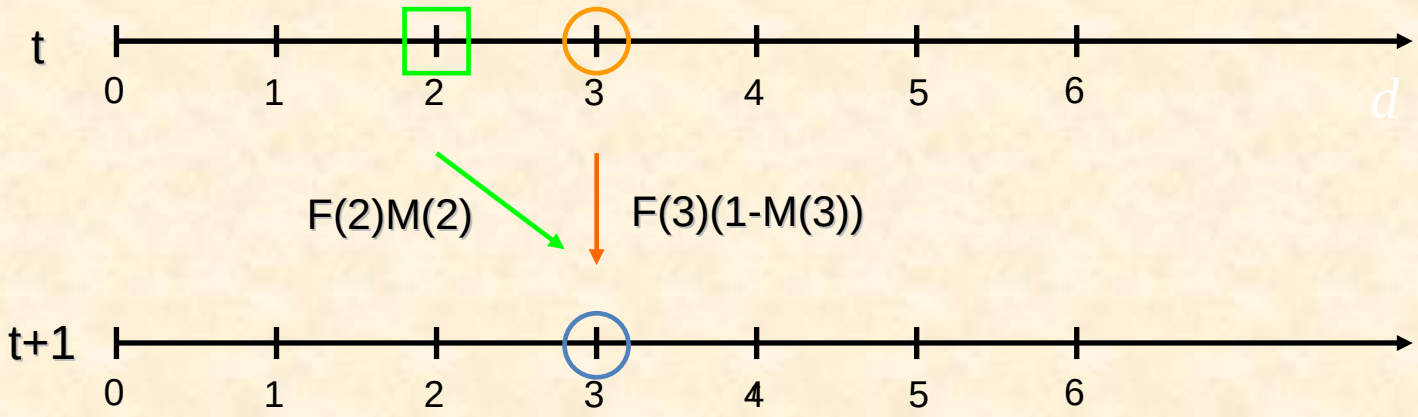
[Lee et al, PLoS Comput. Biol. (2008)]

Intrahost Sequence Evolution Model

Biased walk in one dimension in varying potential

Time unit: 1 month

Evolutionary Rate: 10^{-3} per site per month

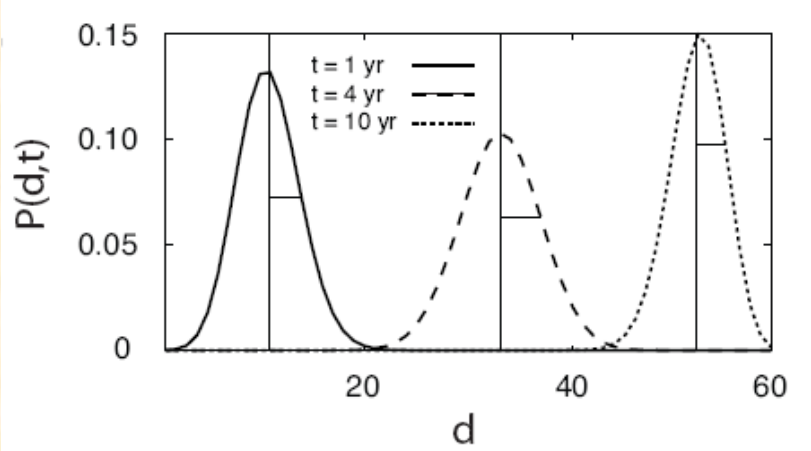


$$N(d,t) = F(d)N(d,t-1)(1-M(d)) + F(d-1)N(d-1,t-1)M(d-1)$$

$$P(d,t) = N(d,t) / \sum_{d=0}^{\infty} N(d,t)$$

$$D_{divergence}(t) = \sum_{n} n\ P(n,t)$$

$$D_{diversity}(t) = \sqrt{\sum_{n} n^2 P(n,t) - D_{divergence}(t)^2}$$



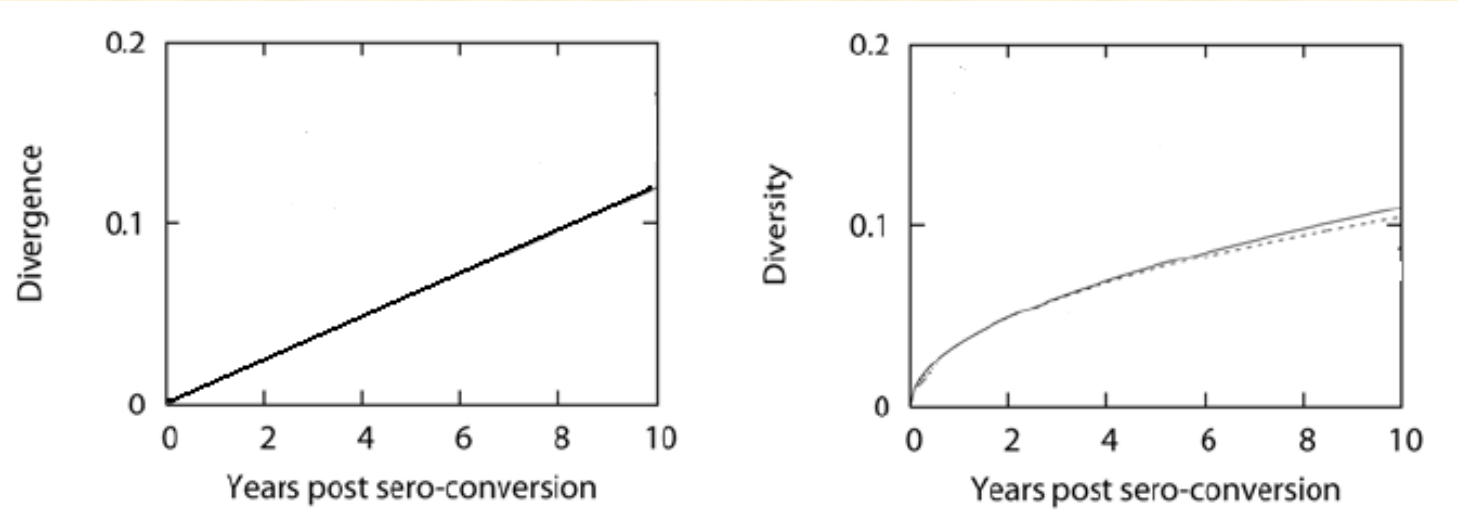
Constant Fitness and Constant Proportion of Mutant Offspring

$F(d) = f$ $\qquad$ $M(d) = m$

$$\frac{\partial P(d,t)}{\partial t} = -mP(d,t) + mP(d-1,t)$$

$$F(z,t) = \sum_{d=0}^{\infty} P(d,t) z^d$$

$$\longrightarrow \quad P(d,t) = e^{-mt}(mt)^d / d!$$



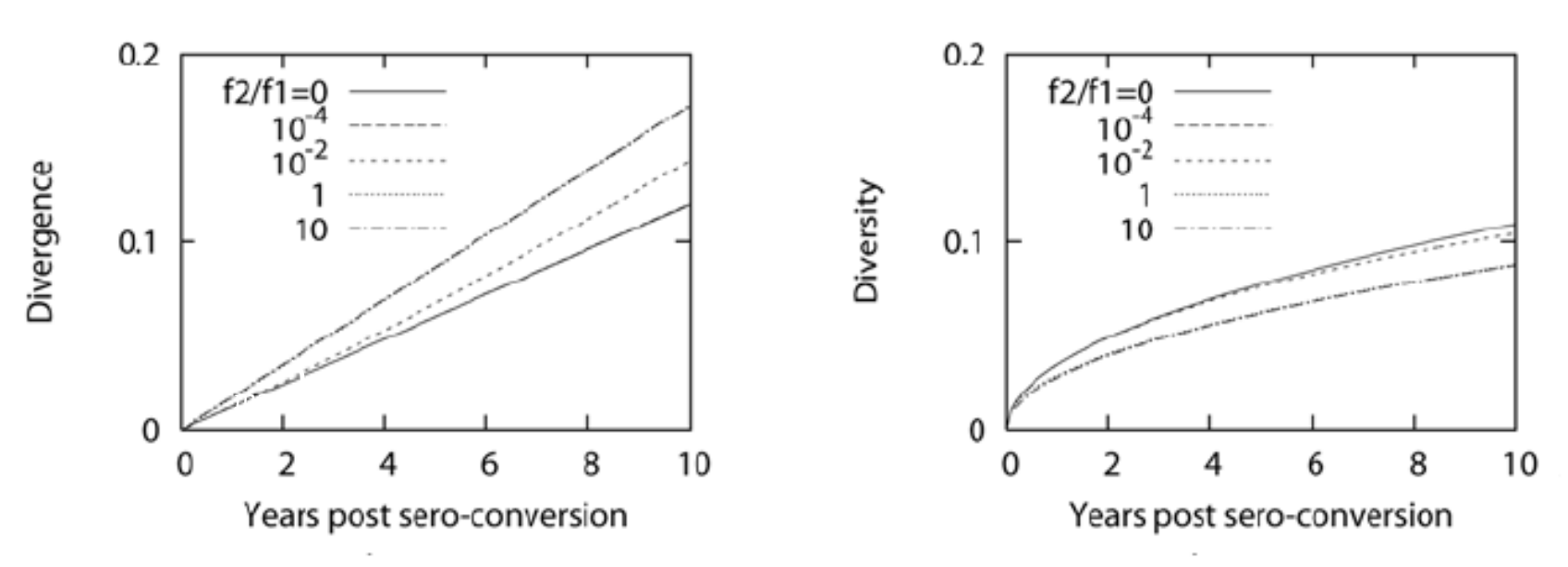
$$D_{divergence}(t) = mt$$

$$D_{diversity}(t) = \sqrt{mt}$$

Linear Increase of Fitness

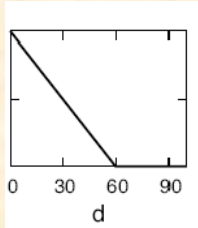
$$M(d) = m$$

$$F(d) = f_1 + f_2 d$$



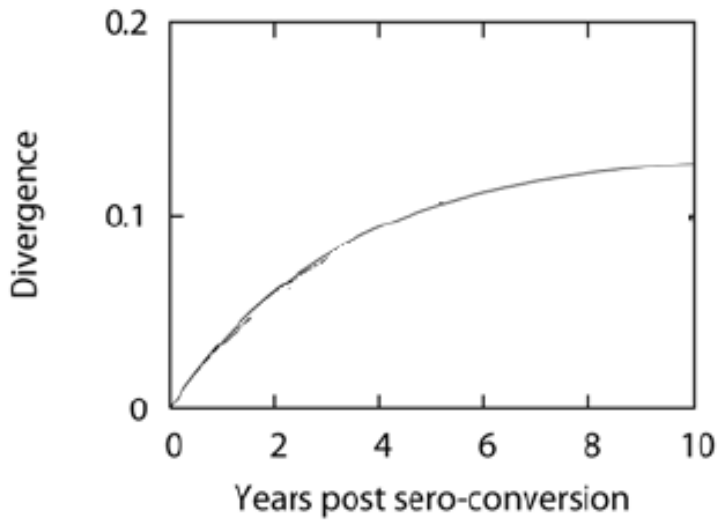
Constant Fitness
and Linear Decrease of Proportion of Mutant Offspring

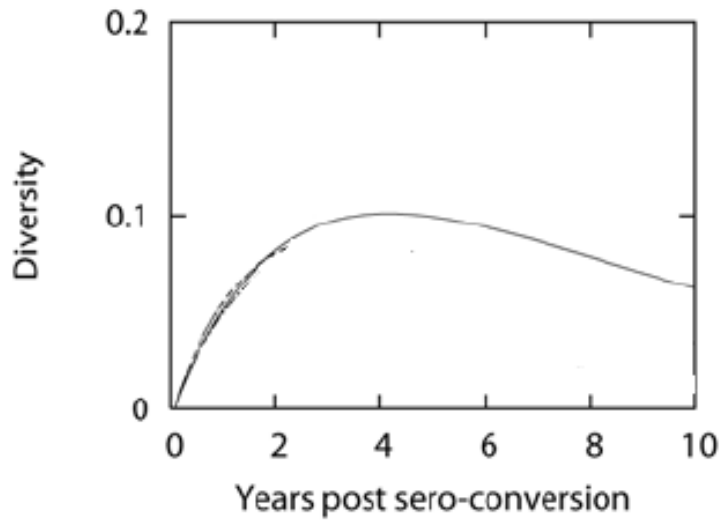
$$M(d) = (1 - d / d_{\max}) \text{ for } d \le d_{\max} \text{ and } M(d) = 0 \quad d > d_{\max}$$



$$\frac{\partial P(d,t)}{\partial t} = \left[\frac{d}{d_{\max}} - 1\right] P(d,t) + \left[1 - \frac{(d-1)}{d_{\max}}\right] P(d-1,t),$$

$$P(d,t) = d_{\max}!/[(d_{\max} - d)!d!][1 - e^{-t/d_{\max}}]^{d}[e^{-t/d_{\max}}]^{d_{\max}-d}$$

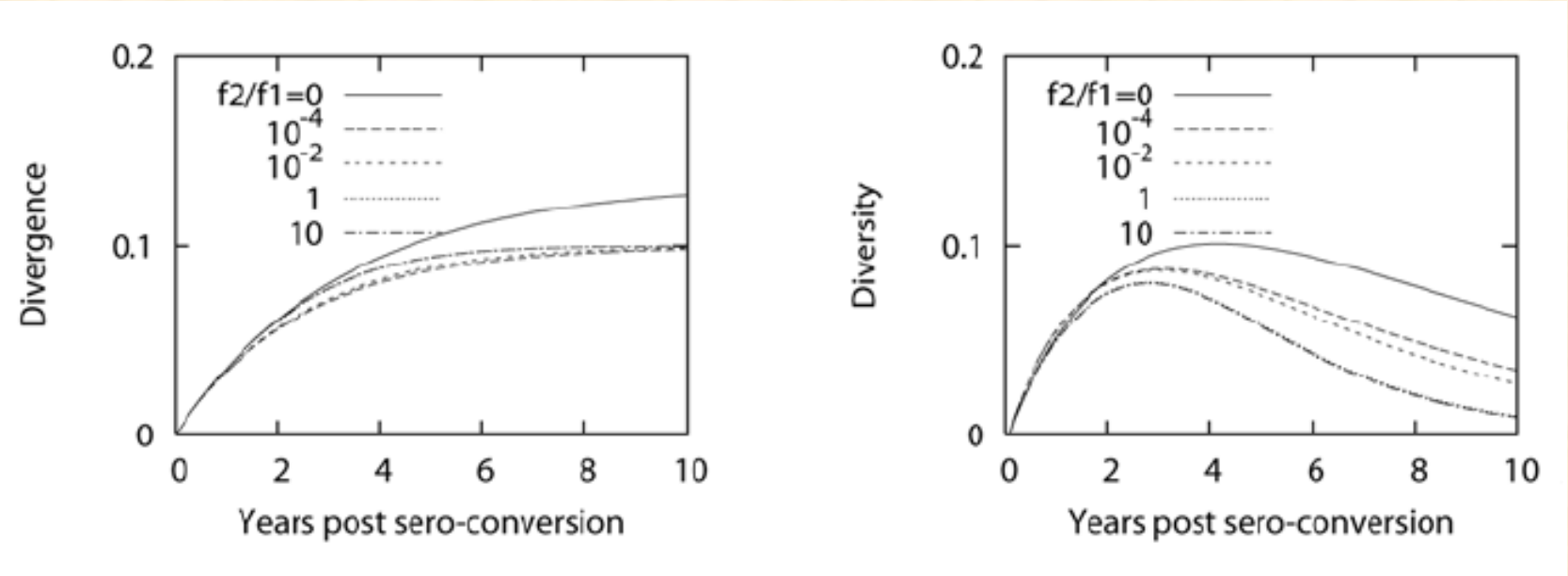




$$D_{divergence} = d_{\max}(1 - e^{-t/d_{\max}})$$

$$D_{diversity} = \sqrt{d_{\max}(1 - e^{-t/d_{\max}})e^{-t/d_{\max}}}$$

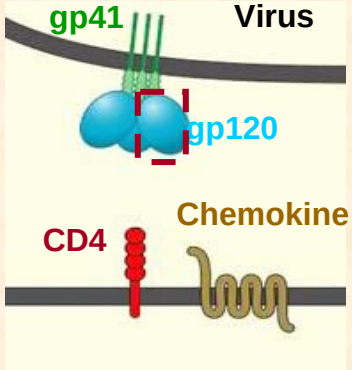
Linear Increase of Fitness
and Linear Decrease of Proportion of Mutant Offspring

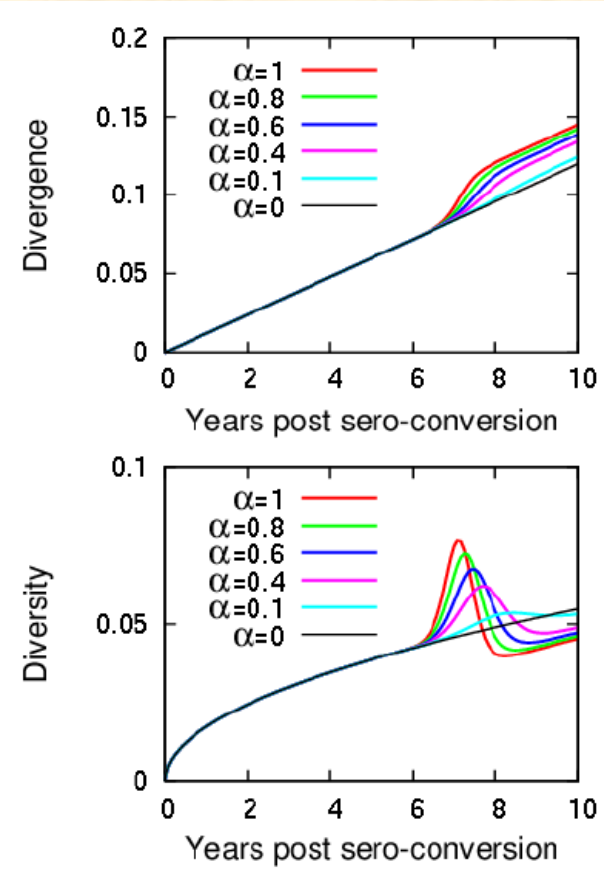


Emergence of More Fit Viruses (using CXCR4 coreceptors)

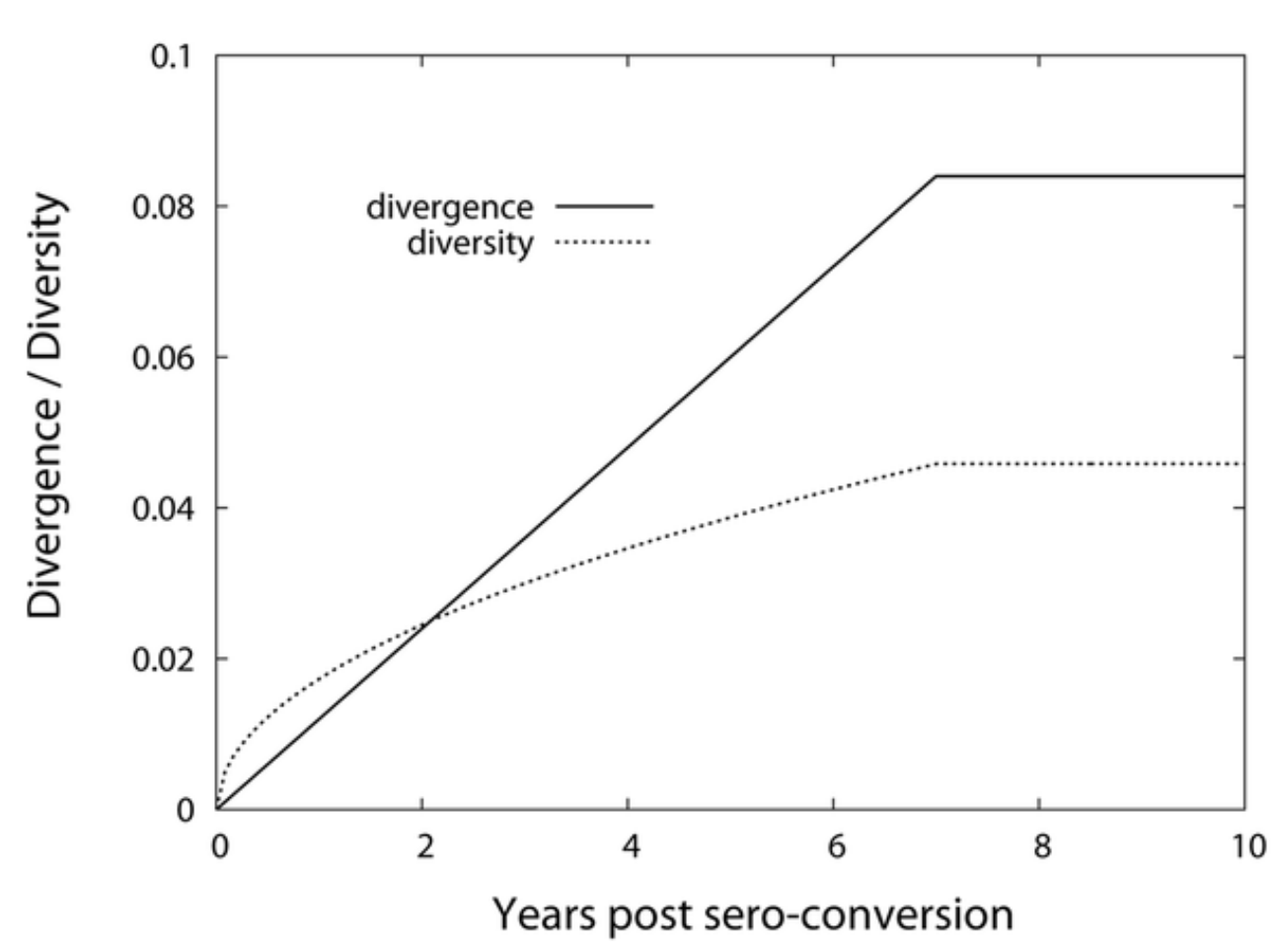
$$N(d,t) = F_{\text{high}}(d)\alpha N(d,t-1)[1-M(d)] + F[1-\alpha]N(d,t-1)[1-M(d)]$$
$$+F_{\text{high}}(d-1)\alpha N(d-1,t-1)M(d-1) + F[1-\alpha]N(d-1,t-1)M(d-1),$$

$$F_{\text{high}}(d) = F_{\text{high}} \quad \text{for } d \geq d_c \quad \text{and} \quad F_{high}(d) = F \quad \text{otherwise}$$



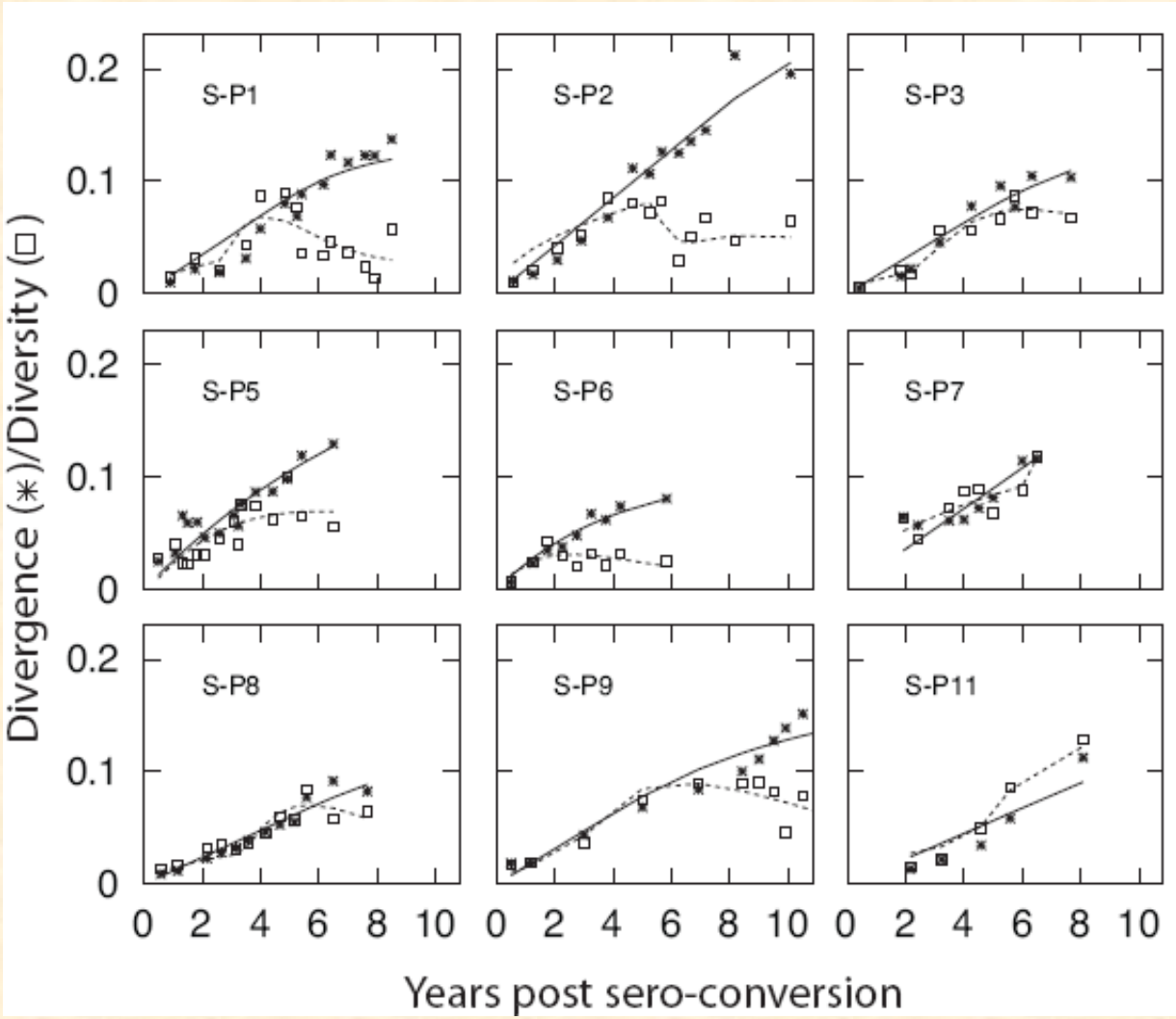


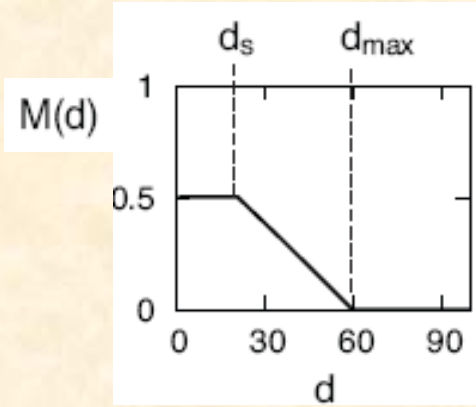
Time Dependent Proportion of Mutant Offsprings



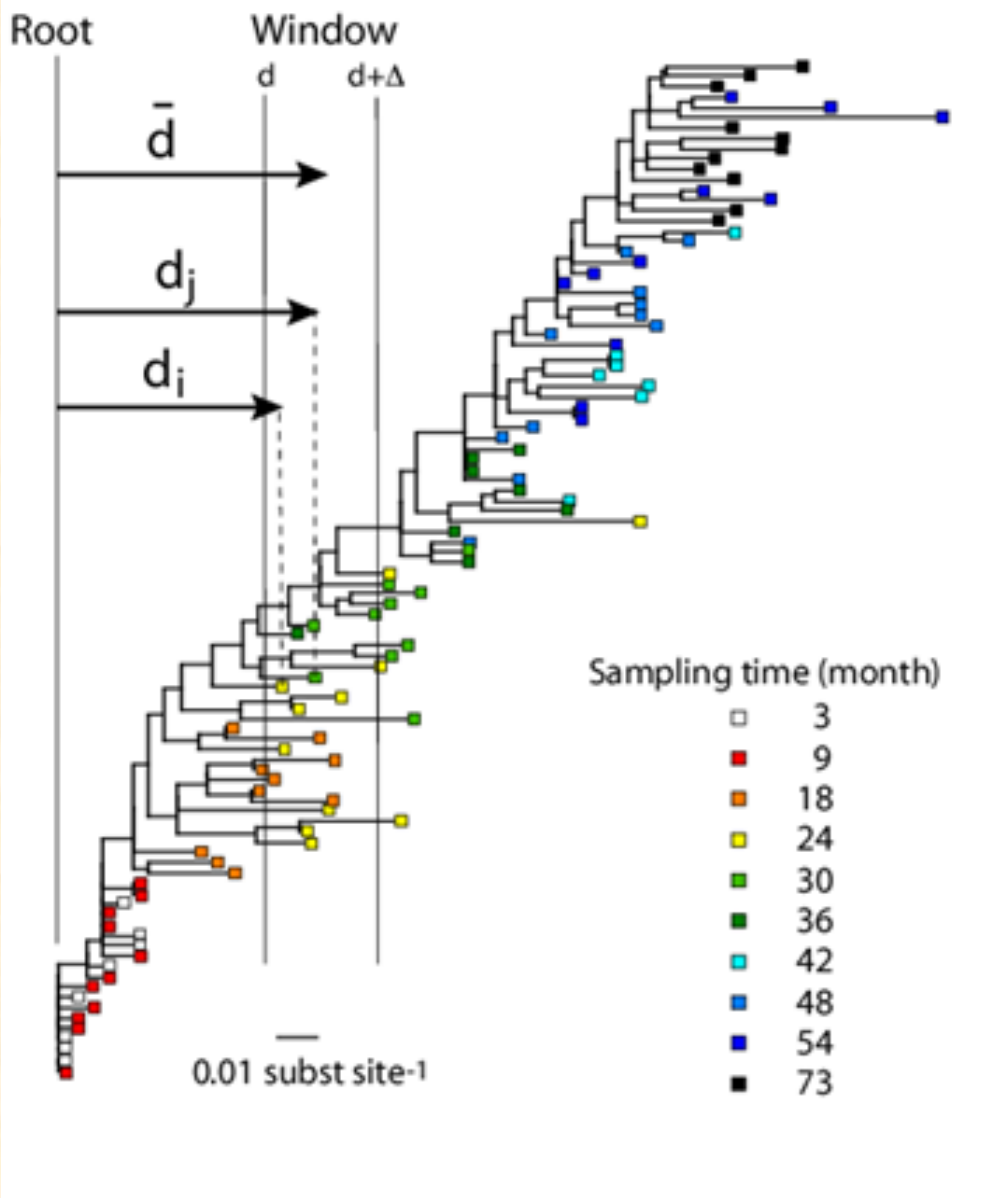
Fit to the Dynamics of Divergence and Diversity

$F(d) = f$



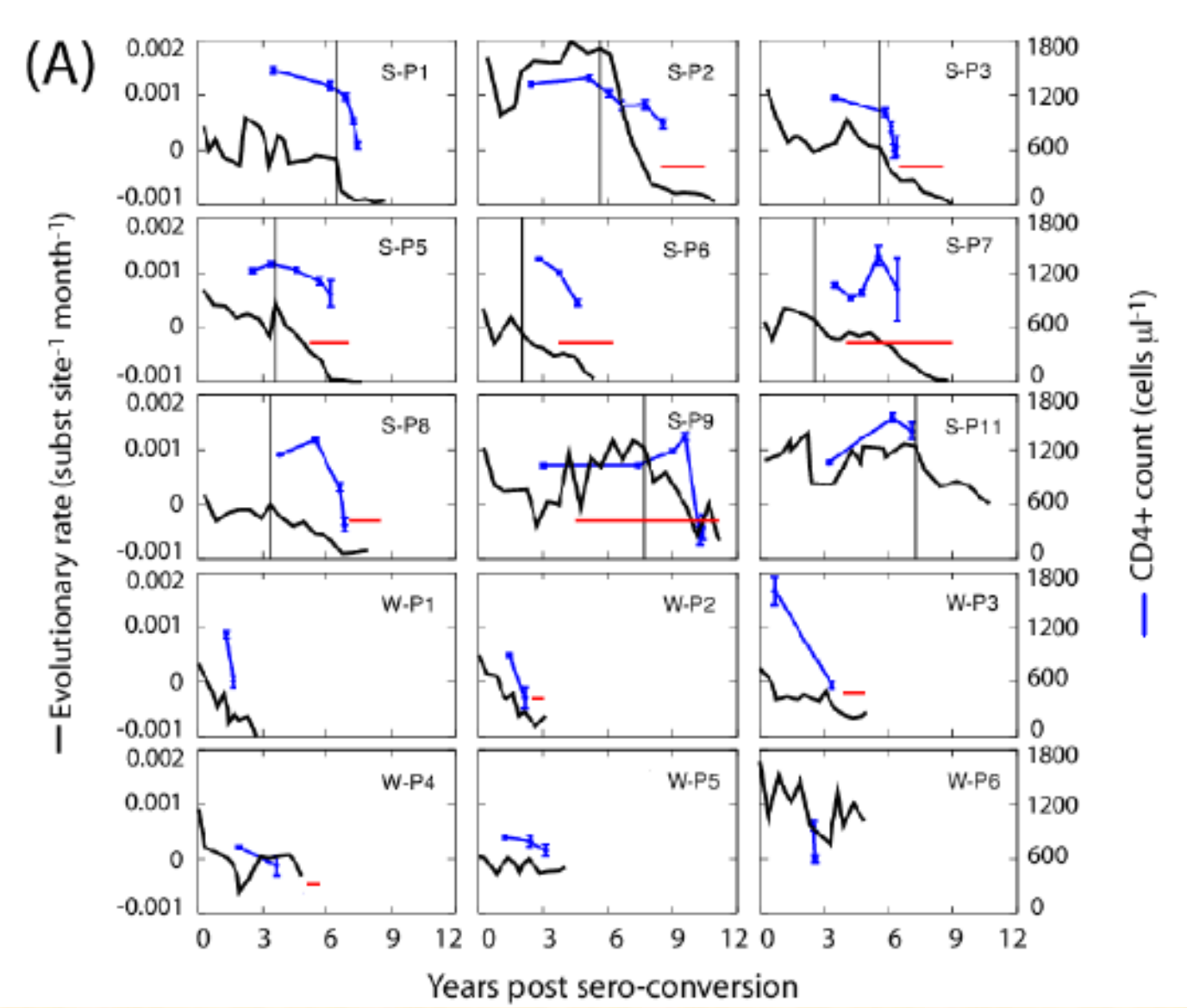


Measuring Evolutionary Rate



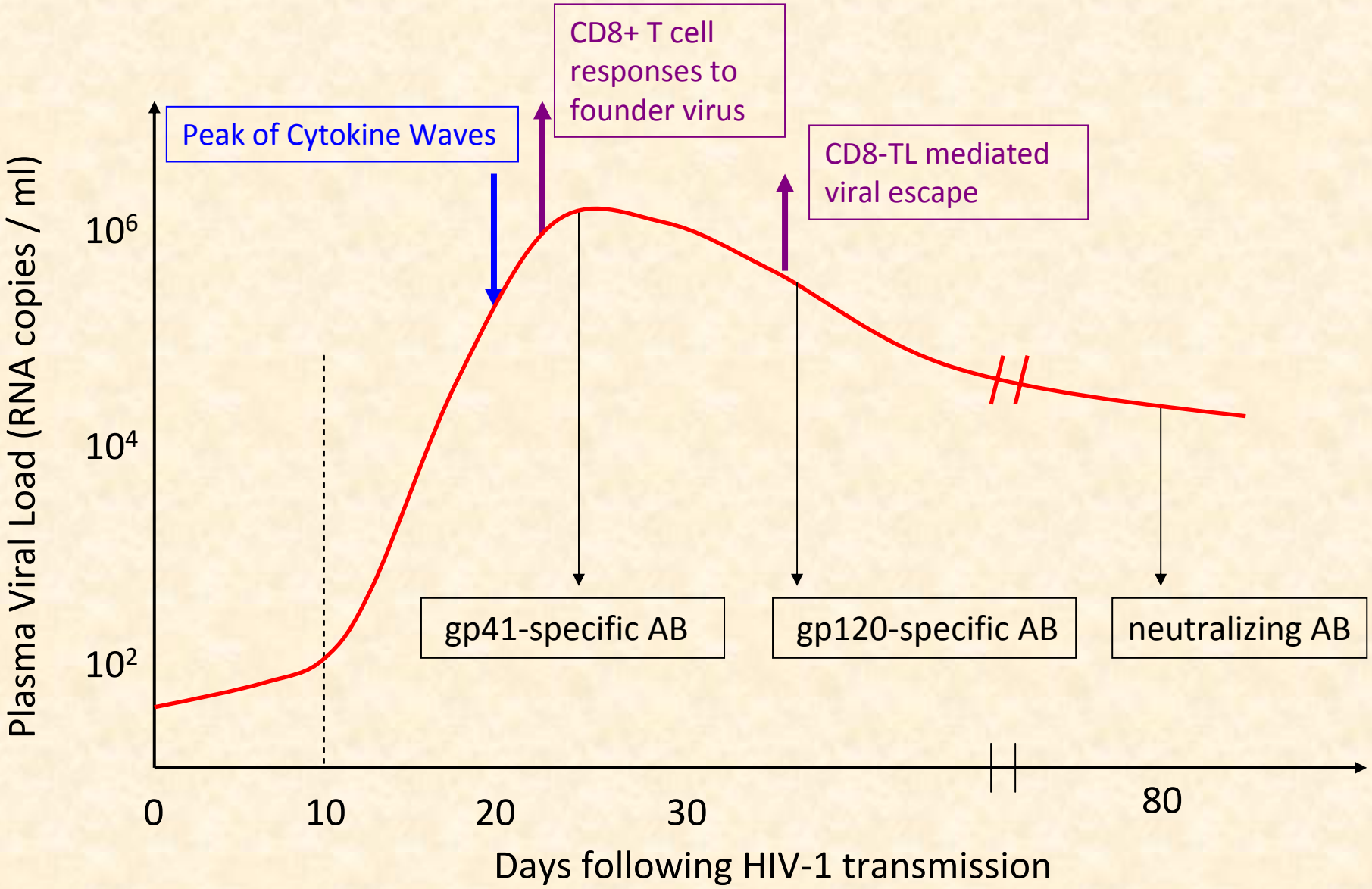
$$ER(\bar{d}) = \frac{\sum_{i,j(t_i \neq t_j)}^{d < d_i, d_j < d+\Delta} (d_j - d_i)/(t_j - t_i)}{\sum_{i,j(t_i \neq t_j)}^{d < d_i, d_j < d+\Delta} 1}$$

Dynamics of Evolutionary Rate



What is happening at the end stages of infection?

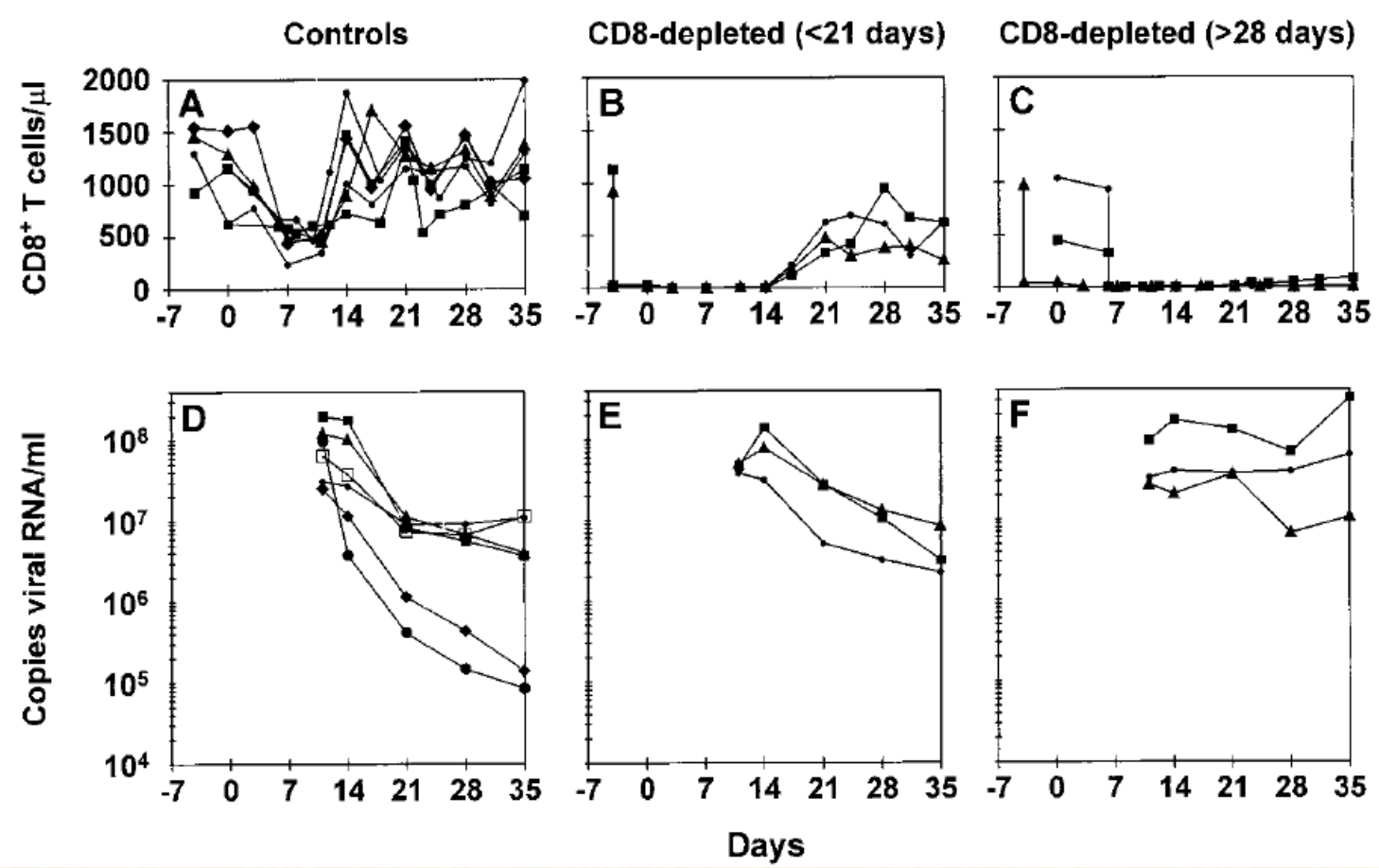
HIV-1 Transmission and Early Stage of Infection



[McMichael et al. Nat Rev Immunology 2010]

CD8+ T cells and acute SIV replication

- Effect of CD8+ lymphocyte depletion on control of SIV replication



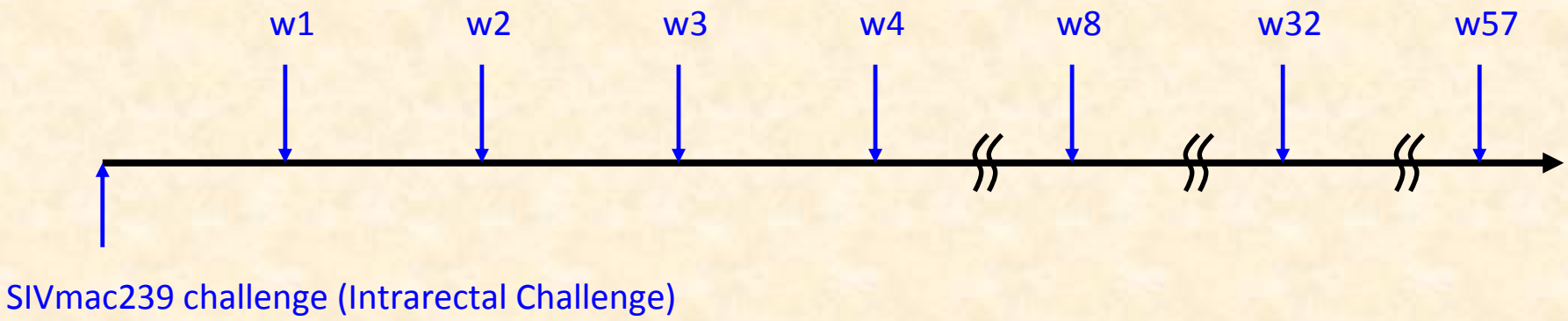
[Schmitz, Science 1999]

Study Design

[Bimber et al. J Virol 2009]

- 4 Mauritian cynomolgus macaques (cy0161, cy0162, cy0163, cy0165)

- 4 Indian rhesus macaques (rh2122, rh2124, rh2126, rh2127)

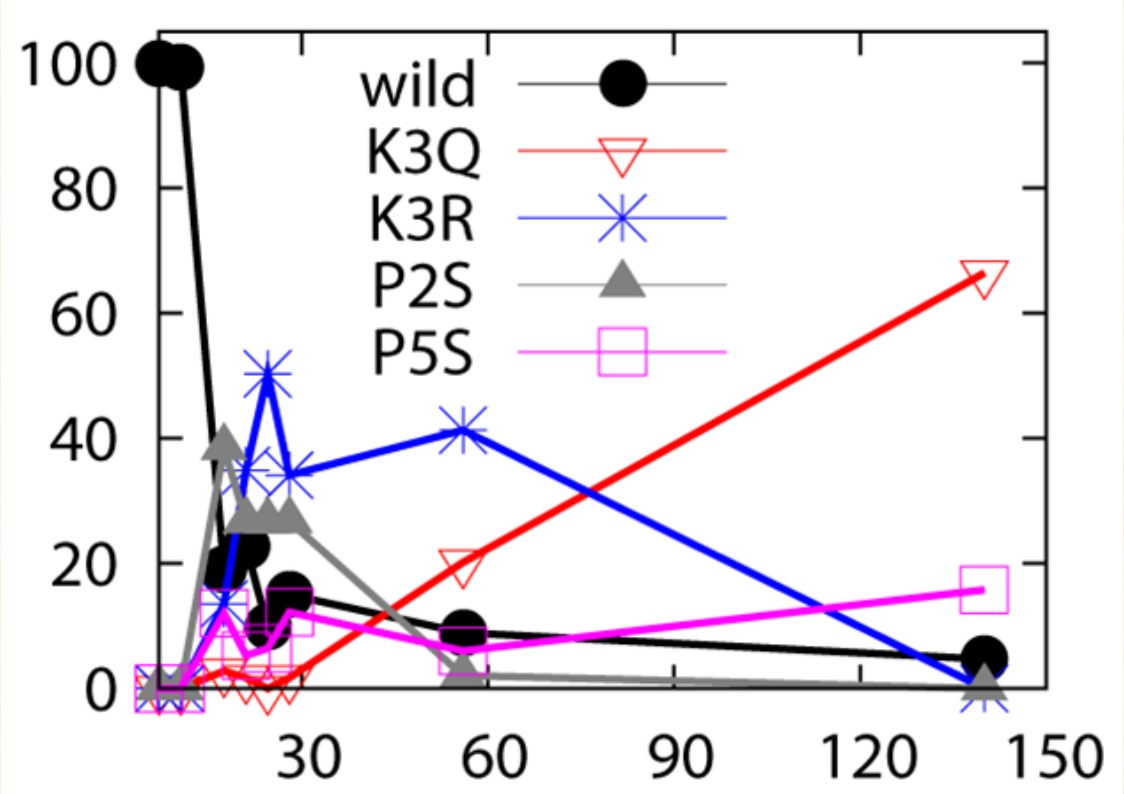


$Nef_{103\text{-}111}RM9$, $Tat_{28\text{-}35}SL8$, $Gag_{181\text{-}189}CM9$

(in average 1414 sequence reads were obtained using deep sequencing)

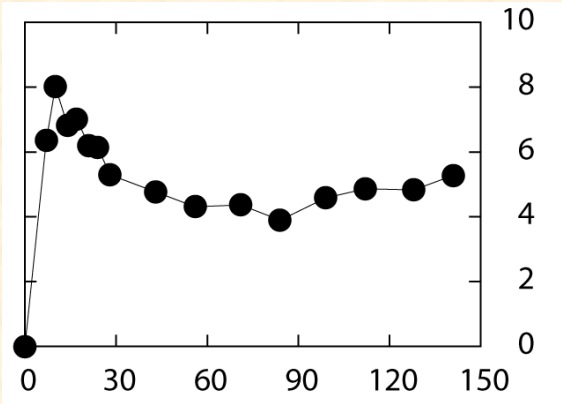
Mafa-A1*063 restricted $Nef_{103-111}$ RPKVPLRTM (RM9)

Proportion of Each Epitope Variant



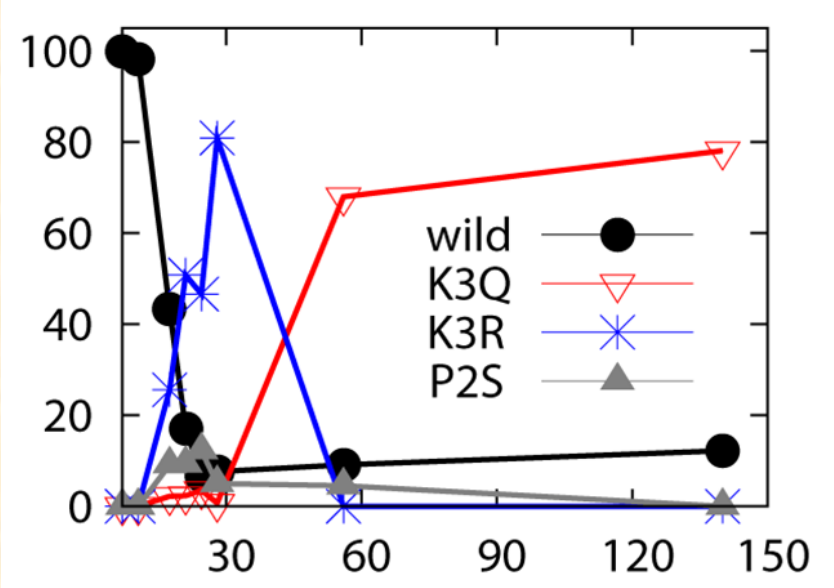
Days following SIV infection

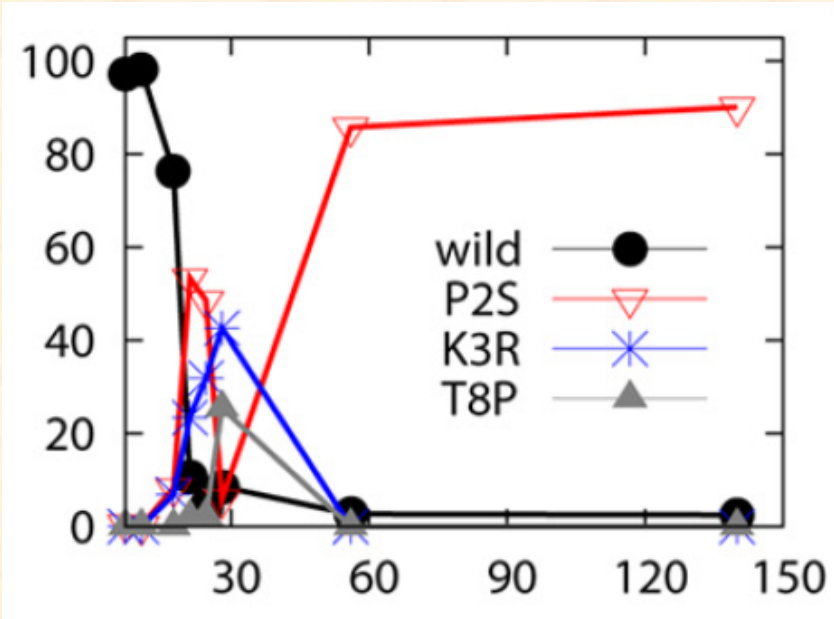
Viral Load

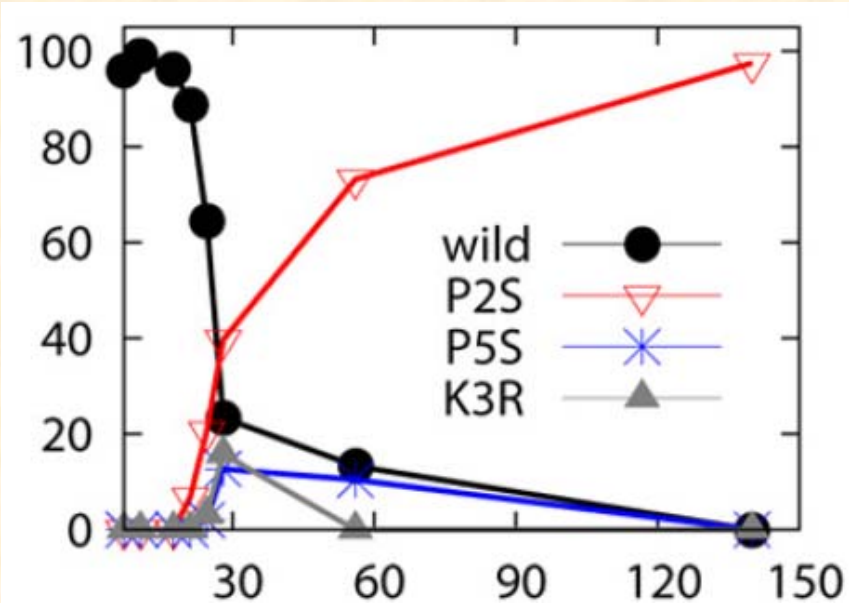


[Love et al. J Virol 2010]

Mafa-A1*063 restricted $Nef_{103-111}$ RPKVPLRTM (RM9)

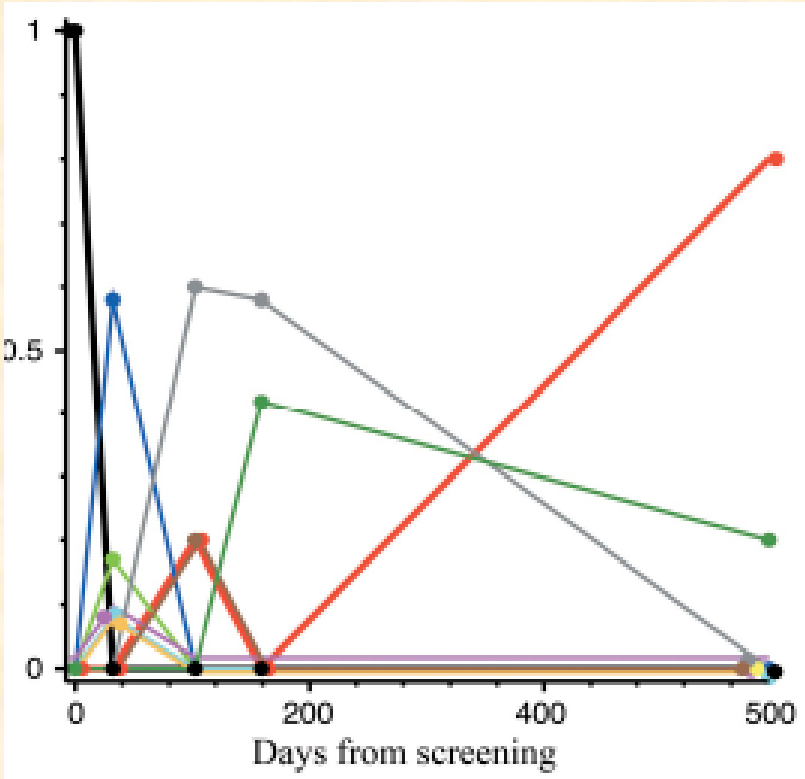


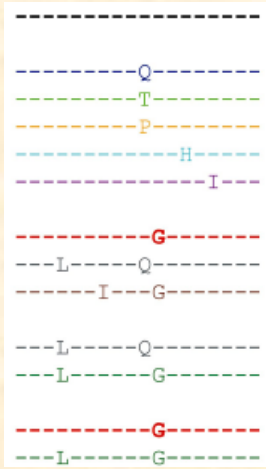




HIV-1 Escape Dynamics

KAALDLSHF, B*5701 $Nef_{73\text{-}90}$ epitope



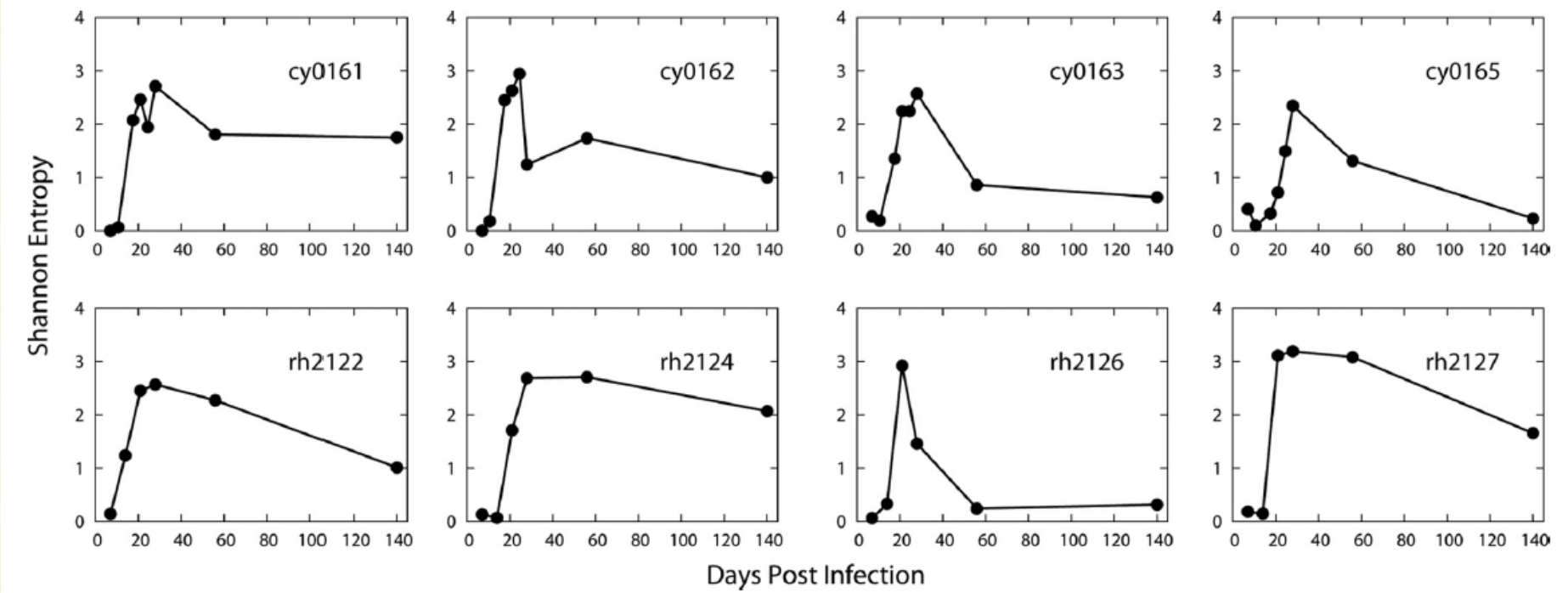


Within 42 days post the first screening

The rate of escape : 0.33 day^{-1} [0.22-0.42]

[Goonetilleke et al. JEM 2009]

Epitope Diversity Dynamics



First 6 months vs. Whole Infection Period (3-10 years)

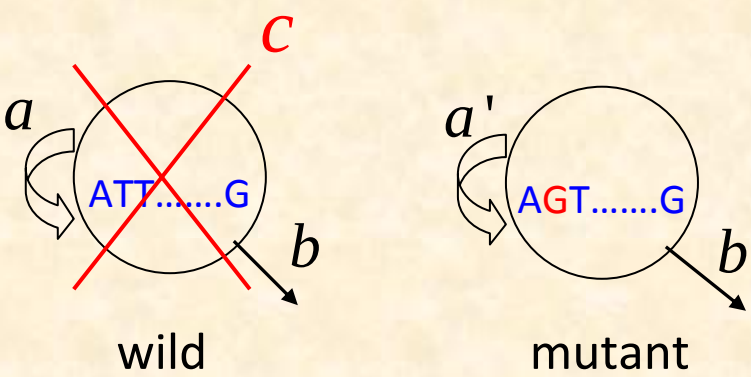
Changes in Selection Strength?

Mathematical model for viral escape

Model Equations

$$\frac{dw(t)}{dt} = aw(t) - bw(t) - cw(t)$$

$$\frac{dm(t)}{dt} = a'm(t) - bm(t)$$

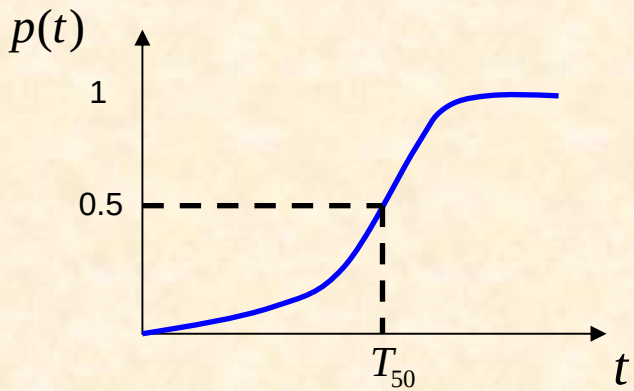


[Asquith, PLoS Biol 2006]

[Asquith, PNAS 2007]

Model Solution

$$p(t) = \frac{m(t)}{w(t)+m(t)} = \frac{1}{1+g\exp(-kt)}$$



Timing of escape : $T_{50} = \log(g)/k$

Rate of escape = Rate of CTL Killing – Fitness Cost

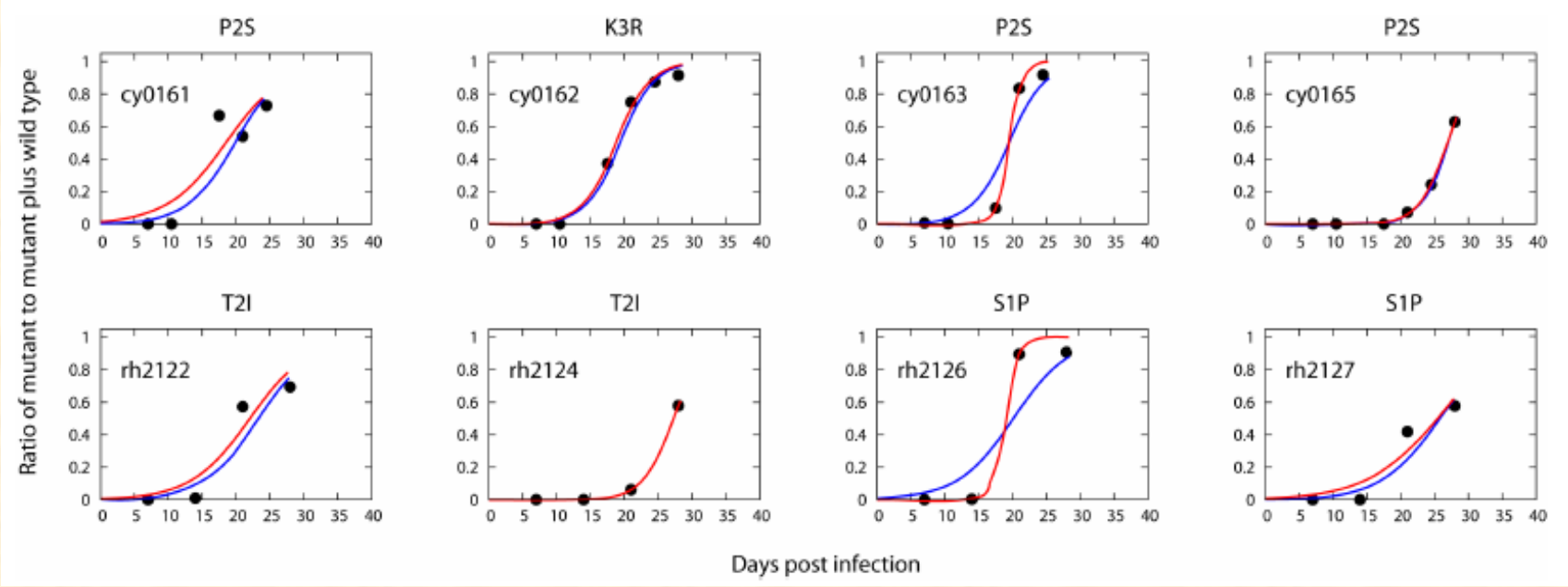
$$k = c - (a - a')$$

Fit the model to acute SIV escape dynamics

$$p(t) = \frac{m(t)}{w(t)+m(t)} = \frac{1}{1+g\exp(-kt)}$$

RM9

SL8



Nonlinear Least Squares

Beta Regression : data comes from a Beta distribution
with mean E(t)=1/[1+gexp(-kt)] and variance E(t)[1-E(t)]/(1+phi)

[Ferrari and Cribari-Neto, J. Appl. Stat. 2004]

Least Square vs. Beta Regression

Nonlinear Least Squares

	First Acute Escape
Timing of First Escape	22.1 (± 3.78) days
Rate of First Escape	0.36 (± 0.36) day^{-1}

Beta Regression

	First Acute Escape
Timing of First Escape	22.7 (± 3.51) days
Rate of First Escape	0.33 (± 0.097)day^{-1}

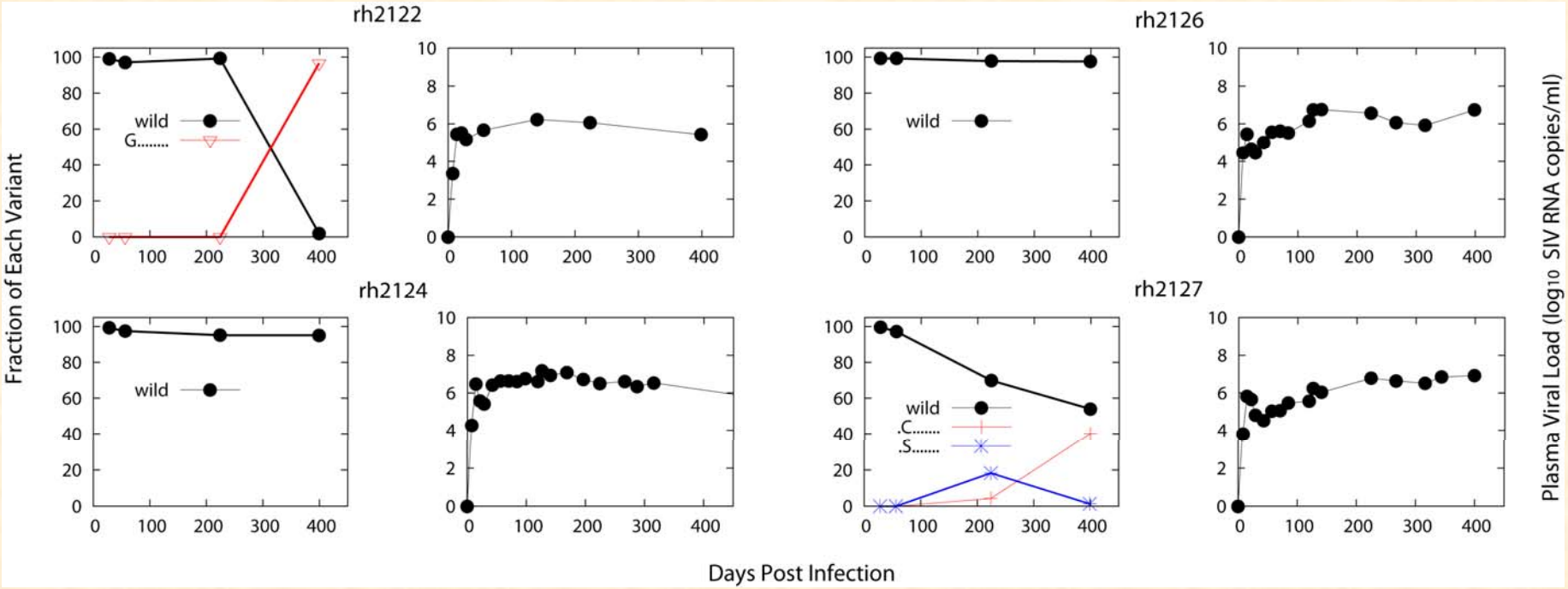
Comparison to other studies

Our Estimate

Rate of Escape 0.36 (± 0.36) day^{-1}

SIV		
Study	Disease Stage	Rate of Escape (day^{-1})
Fernandez et al.	Acute	0.93
Ganusov et al.	Acute	0.31 [0.03-0.58]
Loh et al.	Acute	0.31 [0.1-0.52]
HIV		
Goonetilleke et al.	Acute	0.33 [0.22-0.42]

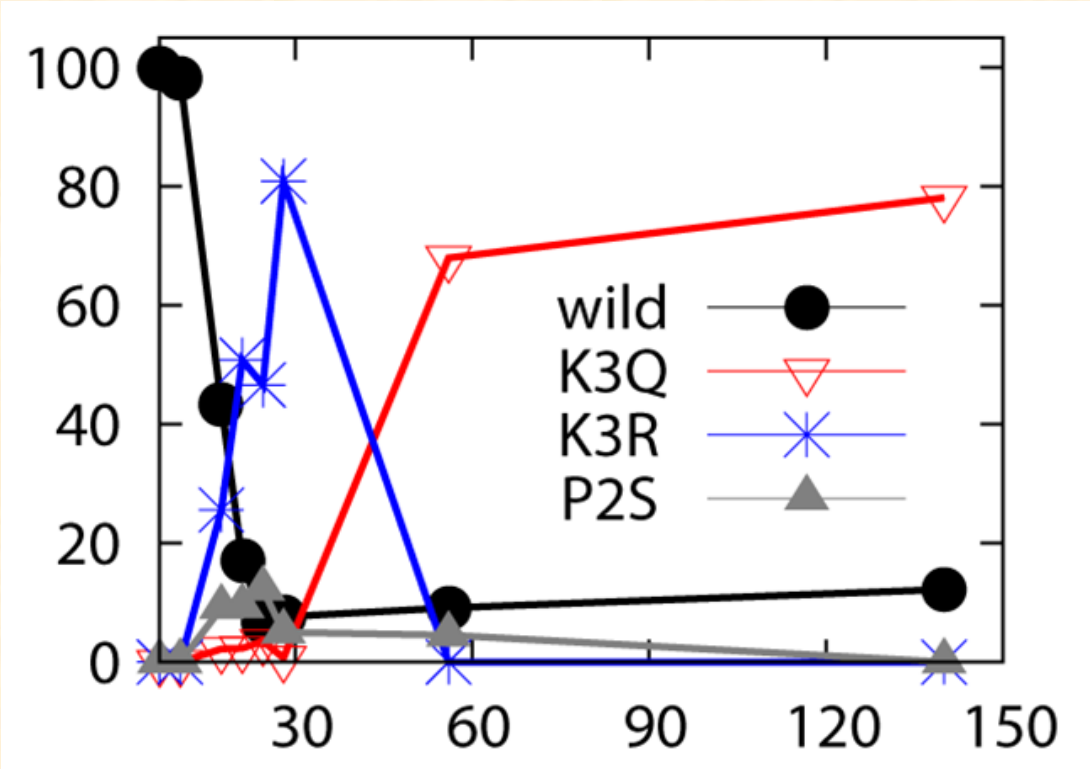
Escape dynamics of $Gag_{181-189}$CM9 epitope



Timing of escape : 419.6 day

Rate of escape : 0.014 day^{-1}

Persistence of Founder/Wild-type Epitope

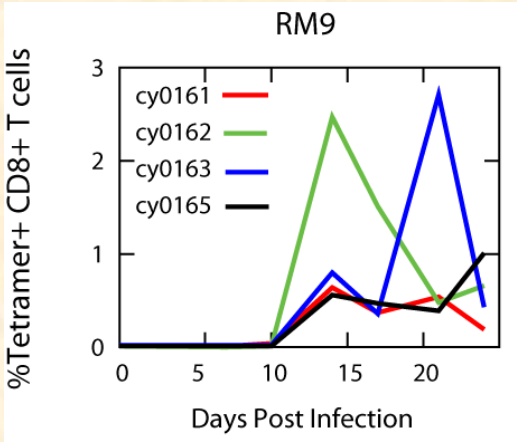


At 140 days post infection

i) Wild-type (transmitted) epitopes persist
as a minor population in every animal (mean 5.7%).

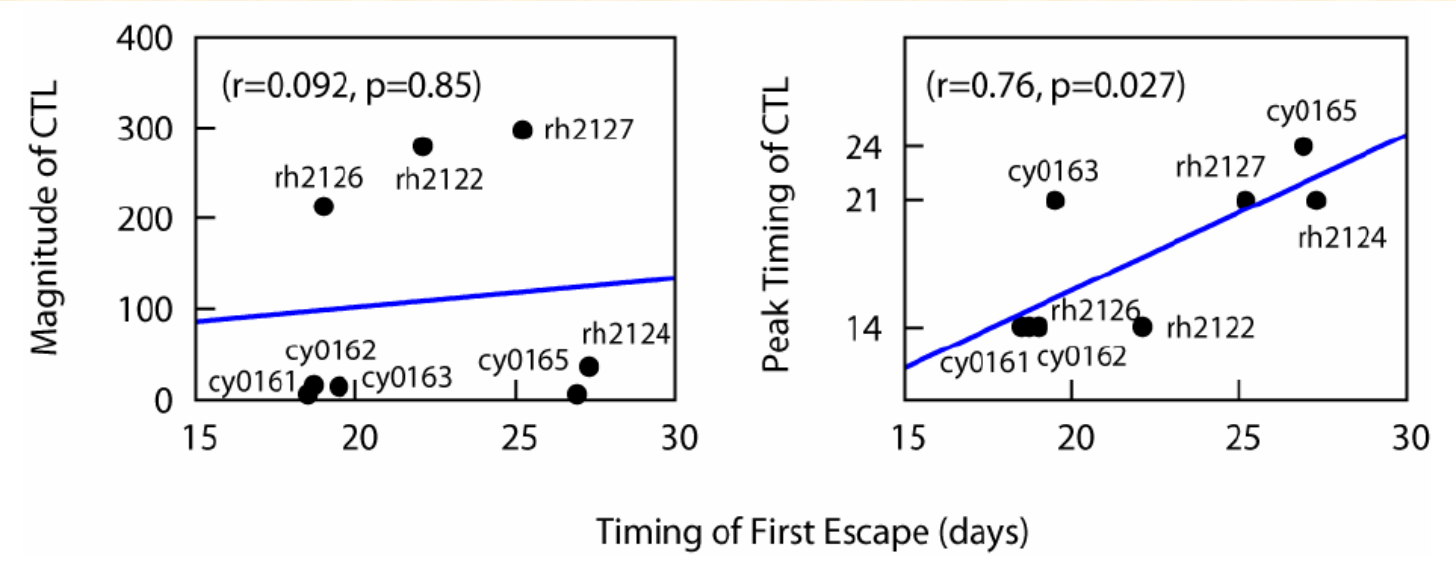
ii) 56% of acute transient epitopes are completely cleared.

Timing of first escape & CT8-TL response



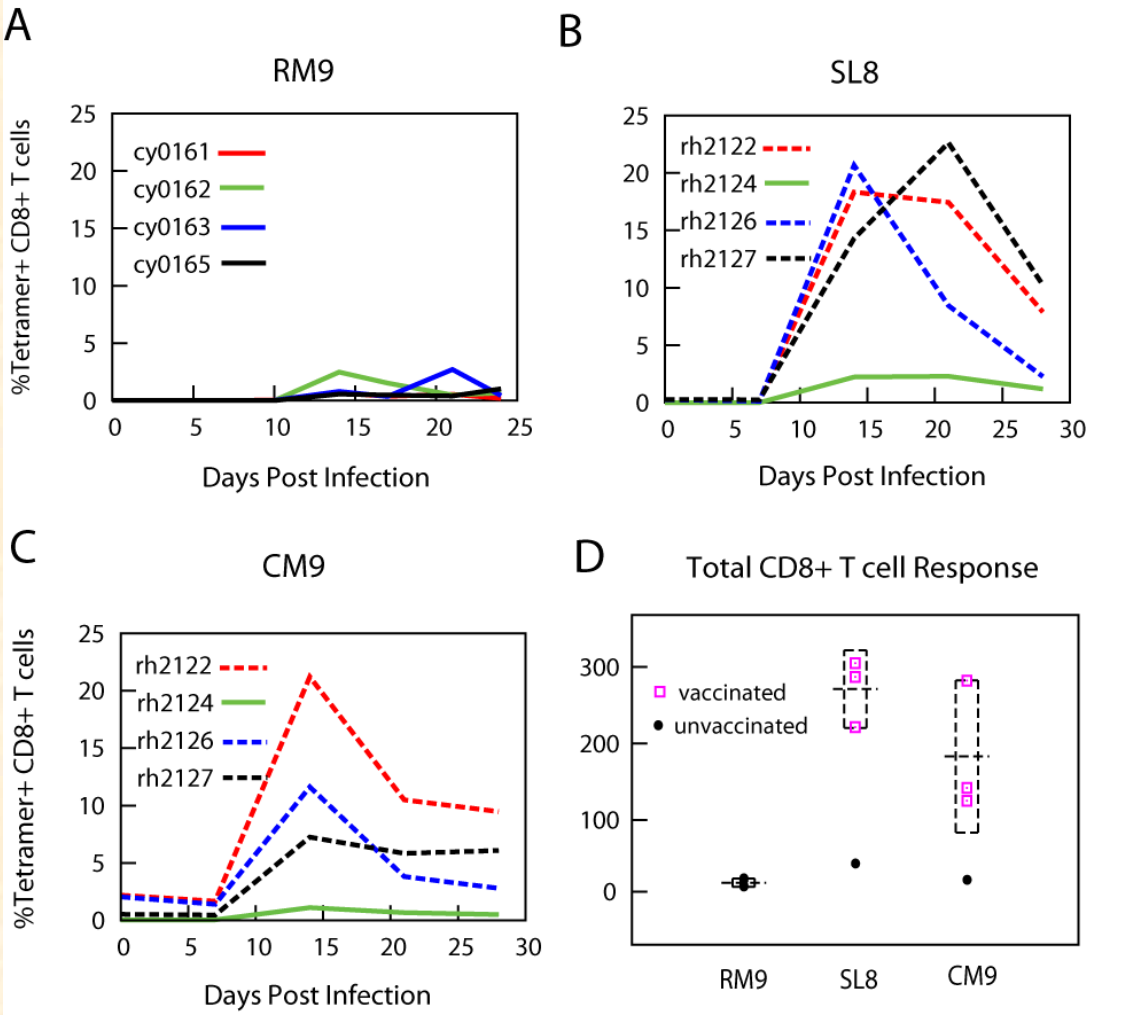
area

peak timing



Rather than the magnitude of CT8-TL response but peak timing of CT8-TL response plays a critical role of controlling timing of first escape

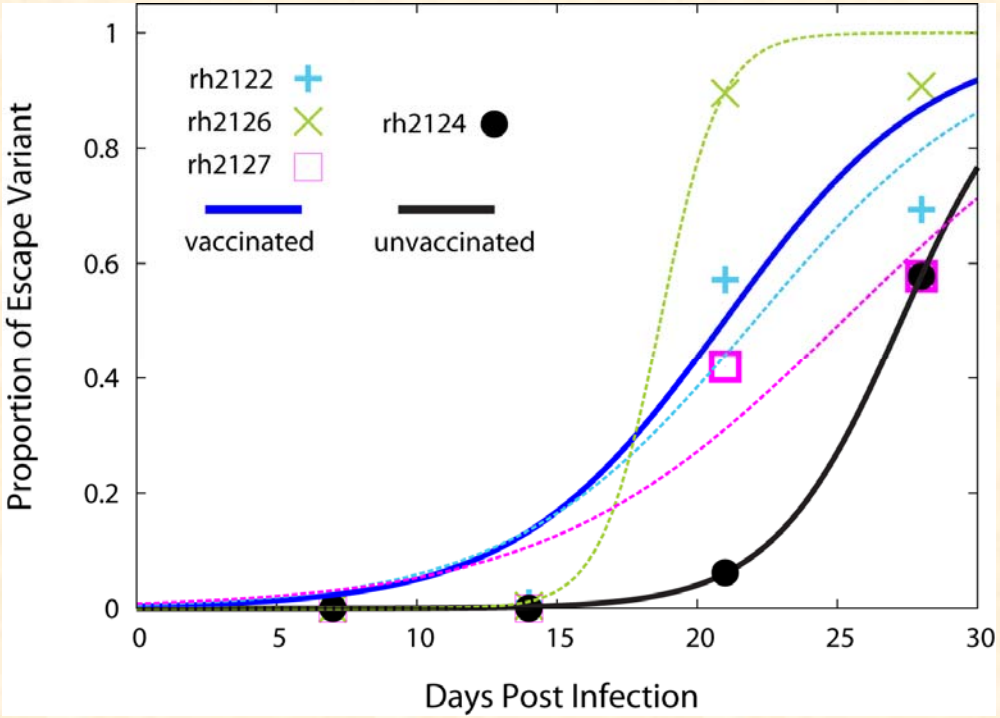
Vaccine induced greater level of CTL response



magnitude of CTL response (area under the curve) : 17 fold difference

[Vogel et al. J Virol 2003]

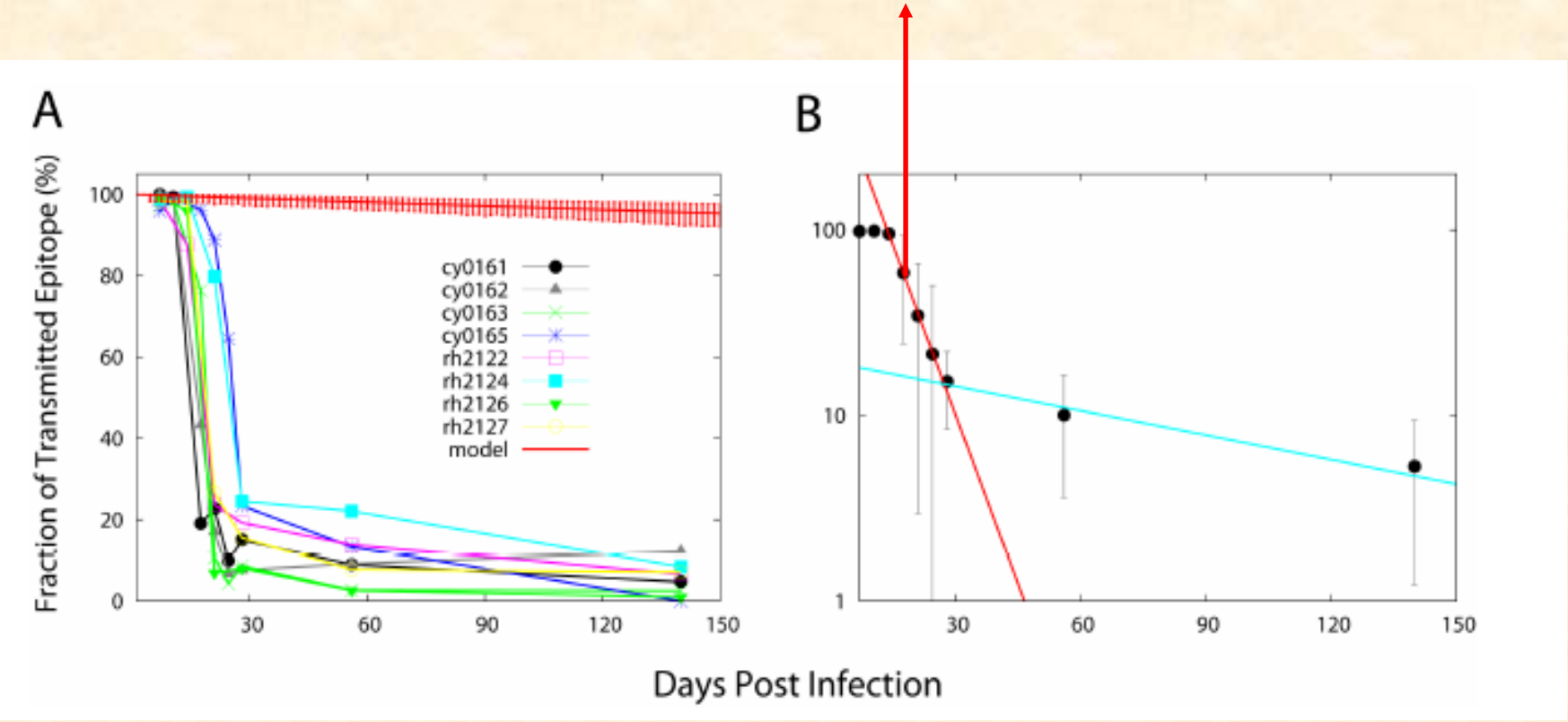
CTL Vaccination induced earlier escape in immunodominant SL8 epitope



	Vaccinated Animals (rh2122, rh2126, rh2127)	Unvaccinated Animal (rh2124)
Timing of First Escape	21.0 [18.4 : 23.6] days	27.3 (± 0.02) days
Rate of First Escape	0.27 [0.09 : 0.44] days	0.44 (± 0.0055) days

Kinetics of loss of the wild-type epitope

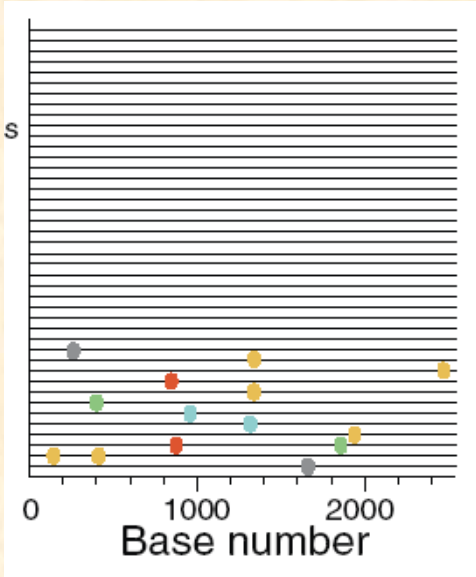
438 times faster than predicted by the model equation



Strong selection pressure by CD8-TL targeting transmitted (founder) epitope

Questions

majority : 70-80% from 102 subtype B+69 subtype C
acutely infected subjects



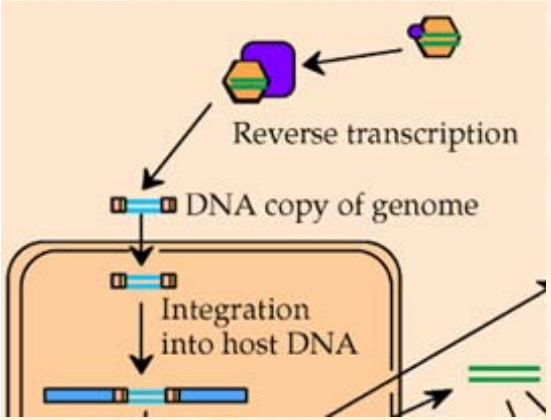
i) Single Strain Infection, One founder (transmitted) virus?

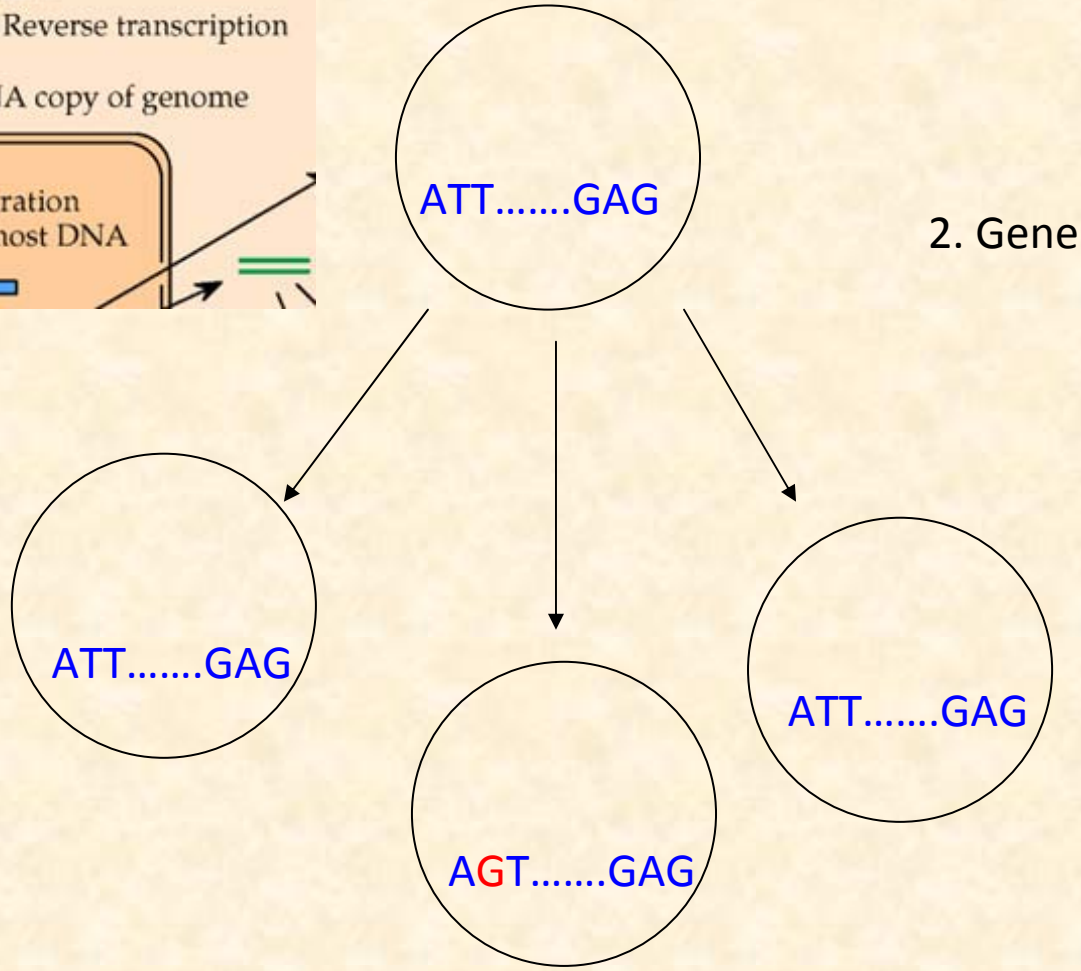
ii) Estimation of days post infection

iii) Identification of Transmitted Virus (Vaccine Design)

[Keele, PNAS 2008] [Lee, J Theo. Biol 2009]

Modeling Evolution of HIV-1 Proviral DNAs





1. Single Cycle HIV-1 Mutation Rate

[Mansky, JVI (1995)]

2.16×10^{-5} base substitutions per base per cycle

2. Generation Time

[Perelson, Science (1996)]

[Markowitz, JVI (2003)]

~ 2 days

3. Basic Reproductive Ratio

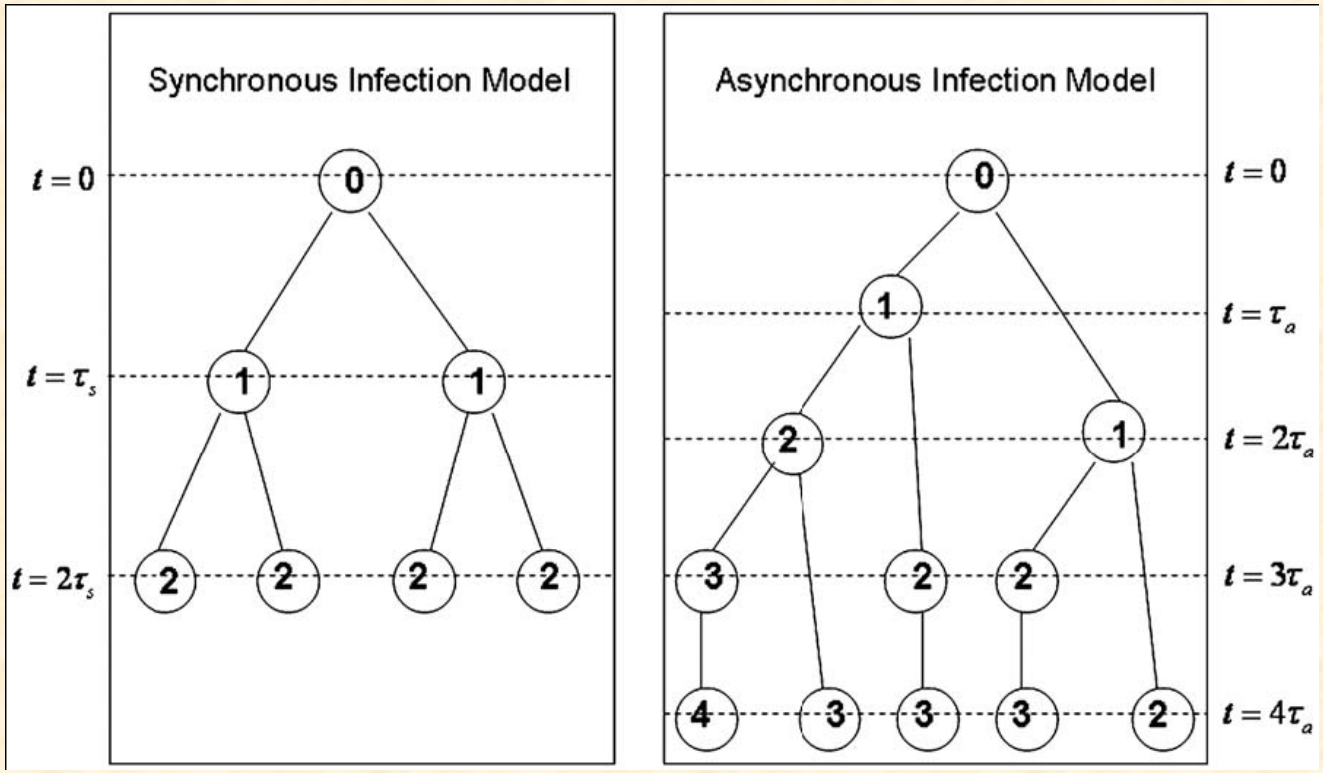
[Nowak, JVI (1997)]

[Stafford, J. Theo. Biol. (2000)]

~ 6

[Lee et al. J Theo. Biol 2009]

Synchronous and Asynchronous Infection Models



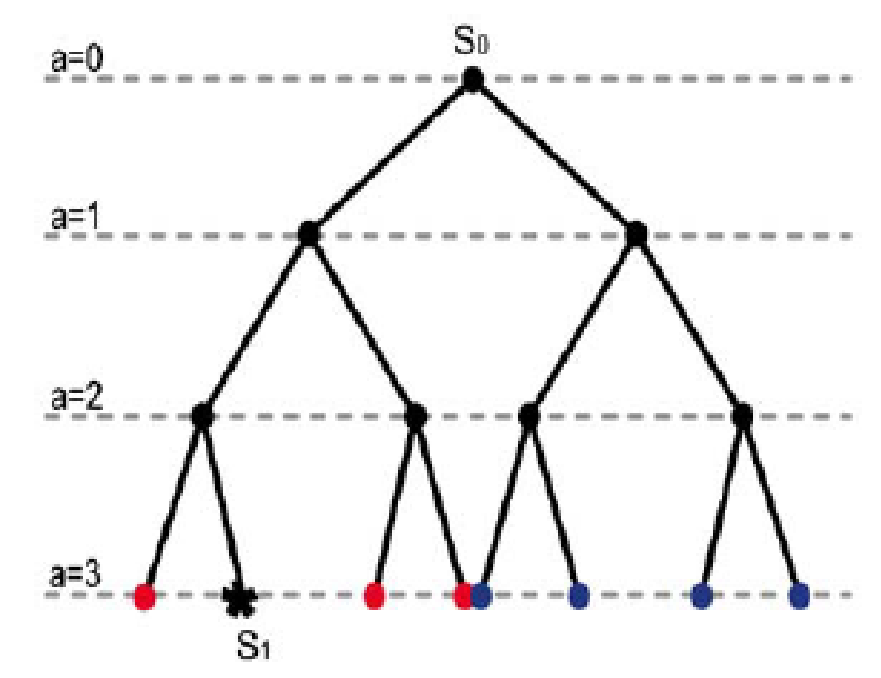
Probability
of having n substitutions

$$P(mutations = n) = \binom{N_B}{n} \varepsilon^n (1-\varepsilon)^{N_B - n} \equiv Binom(n; N_B, \varepsilon).$$

$$P(mutations = n | gen. = 2) = \sum_{k=0}^{n} Binom(n-k; N_B, \varepsilon) Binom(k; N_B, \varepsilon)$$
$$= Binom(n; 2N_B, \varepsilon).$$

$$P(mutations = n | gen. = a) = Binom(n; aN_B, \varepsilon).$$

Probability of Coalescence



The probability of two sequences (at generation a) coalescing at generation m

$$\frac{R_0^{a-m} - 1}{R_0^a - 1}$$

At generation 40, the probability of coalescence
at generation 1 is 0.167,
at generation 2 is 0.028,
at generation 3 or higher is <0.005.

Genetic Diversity : Hamming Distance

The probability that at least one site has mutated more than once

$$P = \frac{1}{2} N_B \frac{t}{\tau_a} (\frac{t}{\tau_a} - 1) \varepsilon^2$$

$N_B = 2800$ $\varepsilon = 2.16 \times 10^{-5}$

1. 30 days post infection $P = 4.8 \times 10^{-4}$
2. 100 days post infection $P = 5.6 \times 10^{-3}$

Hamming Distance Distribution

Synchronous Infection Model

$$
\begin{aligned}
P(HD[s_1, s_2] = d|a) &= \sum_{k=0}^{d} P(HD_0 = k|a) P(HD_0 = d-k|a) \\
&= \sum_{k=0}^{d} Binom(k; aN_B, \varepsilon) Binom(d-k; aN_B, \varepsilon) \\
&= Binom(d; 2aN_B, \varepsilon).
\end{aligned}
$$

Asynchronous Infection Model

$$P(HD_0 = d|n) = \frac{1}{F_n} \sum_{a=\lceil n/2 \rceil}^{n} \binom{a}{n-a} \binom{aN_B}{d} \left(\frac{\gamma}{\alpha}\right)^{n-a} \varepsilon^d (1-\varepsilon)^{aN_B - d}, \quad F_n = (1+\varphi)^{n+1} - (1-\varphi)^{n+1} / 2^{n+1} \varphi \quad \varphi = \sqrt{1 + 8/R_0}$$

Poisson Distribution

$$P(HD = d, t) = e^{-\lambda(t)} \frac{\lambda(t)^d}{d!}$$

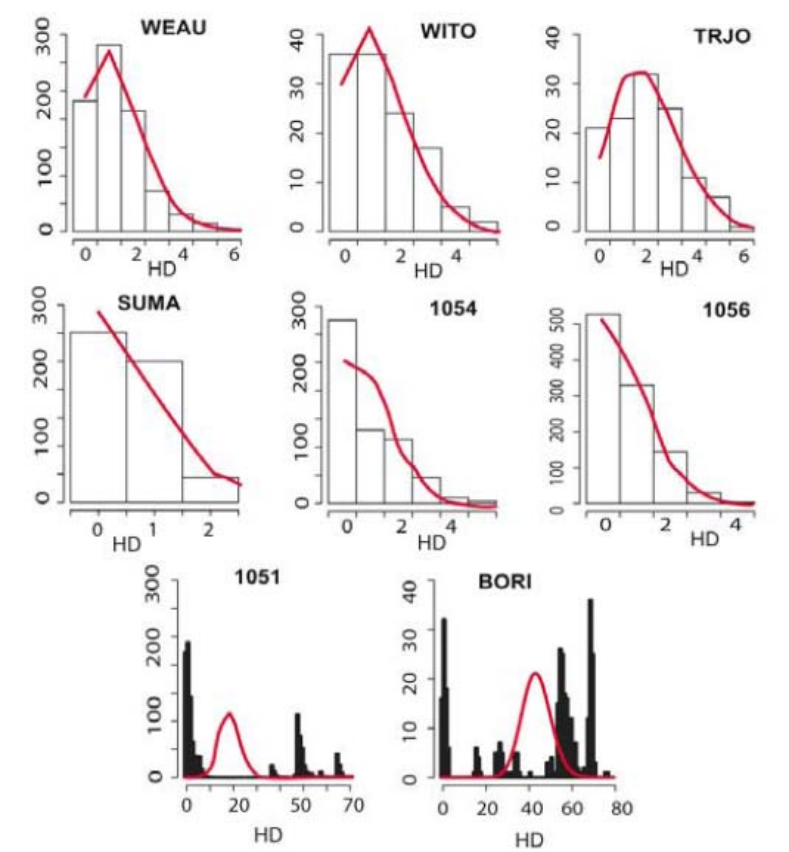
$$\lambda(t) = 2\varepsilon N_B \left(\left\lfloor \frac{t}{\tau_a} \right\rfloor \frac{1+\varphi}{2\varphi} + \frac{1-\varphi}{2\varphi^2} \right)$$

Hamming Distance Distribution

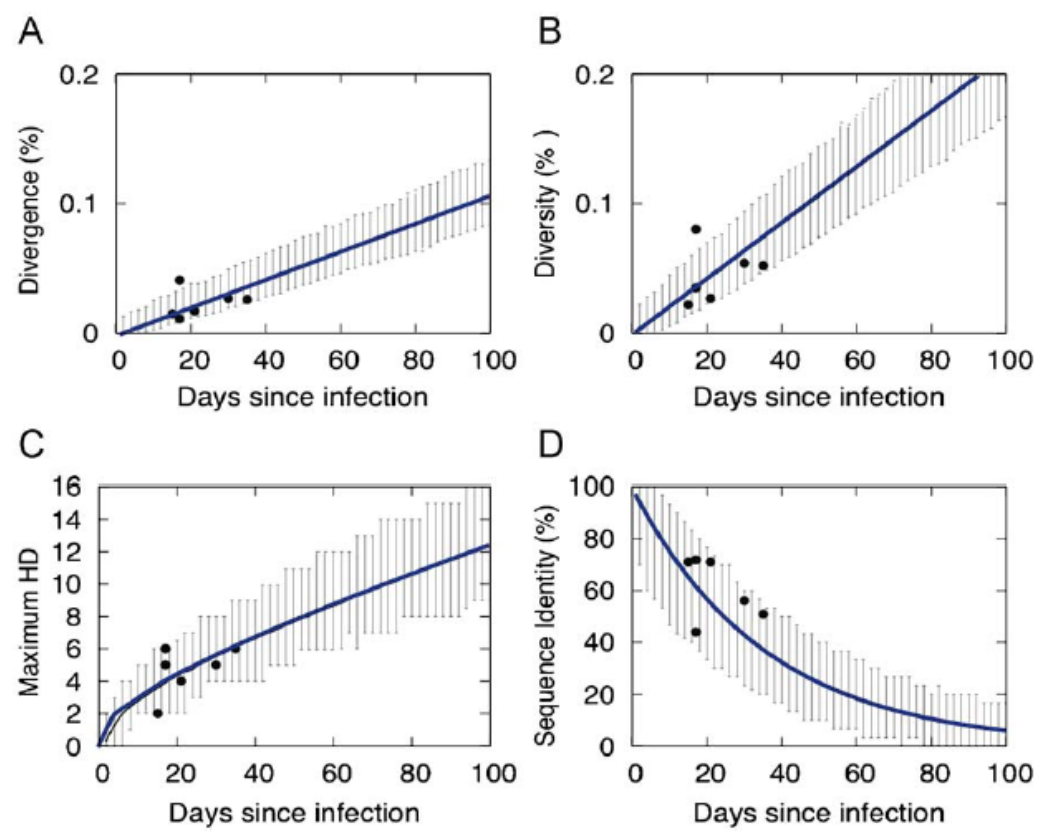
Subject	Gender	Fiebig stage	Number of sequences[a]	Base length	Viral load (copies/ml)	CD4 count (cells/ml)	GenBank accession number
WEAU0575	M	II	43	2581	216,415	714	EU577344–EU577387
WITO4160	M	II	16	2550	325,064	247	EU577388–EU577403
TRJO4551	M	II	16	2574	8,121,951	336	EU577101–EU577118
SUMA0874	M	II	35	2568	939,260	760	EU577039–EU577073
1054	M	II	36	2562	320,000	NA	EU575244–EU575282
1056	M	II	46	2589	140,000	NA	EU575283–EU575328
1051	F	III	50	2619	280,000	NA	EU575134–EU575183
BORI0637	M	II	29	2607	2,400,000	902	EU576274–EU576302

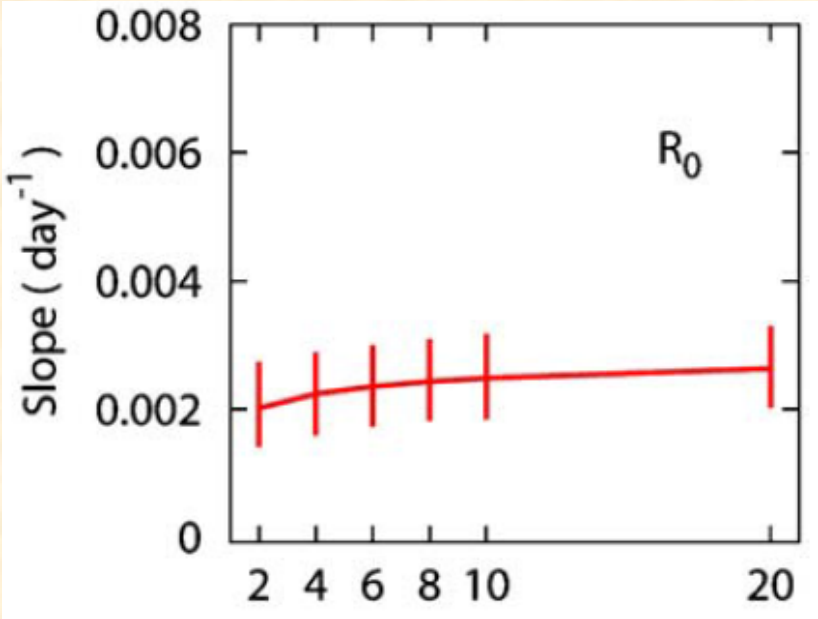
Poisson Distribution

$$P(HD = d, t) = e^{-\lambda(t)} \frac{\lambda(t)^d}{d!}$$



Dynamics


A
Divergence (%)
0.2
0.1
0
0 20 40 60 80 100
Days since infection
B
Diversity (%)
0.2
0.1
0
0 20 40 60 80 100
Days since infection
C
Maximum HD
16 14 12 10 8 6 4 2 0
0 20 40 60 80 100
Days since infection
D
Sequence Identity (%)
100 80 60 40 20 0
0 20 40 60 80 100
Days since infection


Slope (day^{-1})
0.008
0.006
0.004
0.002
0
R_0
2 4 6 8 10 20

Collaborations

Los Alamos National Lab
Alan Perelson
Thomas Leitner
Elena Giorgi
Bette Korber

Colonge Univeristy
Su-Chan Park

University of Alabama, Birmingham
George Shaw

University of Rochester
Tanzy Love
Sally Thurston
Sung Yong Park
Michael Keefer
Stephen Dewhurst

University of Wisconsin-Madison
Benjamin N. Bimber
Benjamin J. Burwitz
Matthew Reynolds
David Watkins
David O'Connor

Funded by

NIH R01-AI083115
NIH T32-ES007271
NIH U of R DCFAR (P30AI078498)