

Microsoft® Research

Faculty Summit 2010

Guarujá, Brasil | May 12 – 14 | In collaboration with FAPESP

Microsoft® Research

Faculty Summit 2010

Guarujá, Brasil | May 12 – 14 | In collaboration with FAPESP

Fighting HIV with Machine Learning and HPC

David Heckerman
Distinguished Scientist
Senior Director of e Science
Microsoft Research

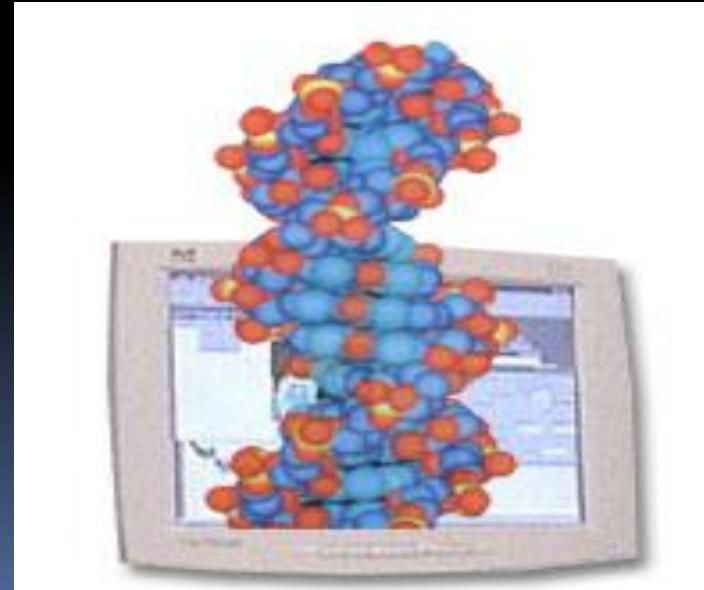
Fighting HIV with Machine Learning and High Performance Computing

David Heckerman
eScience Group, Microsoft Research

Really? Why?

The convergence of computer science and biology

- DNA is a programming language and a computation device



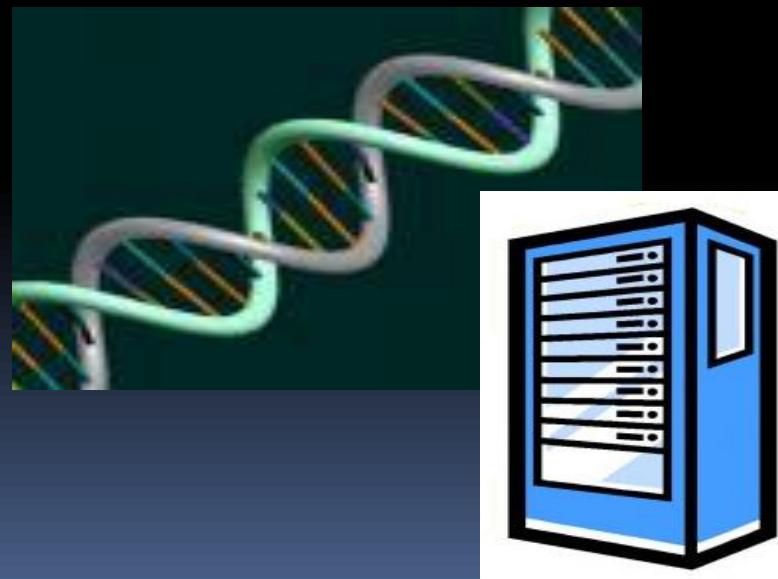
The convergence of computer science and biology

- Drinking from the fire hose or ...



The convergence of computer science and biology

- Striking similarities in concepts that can be shared both ways



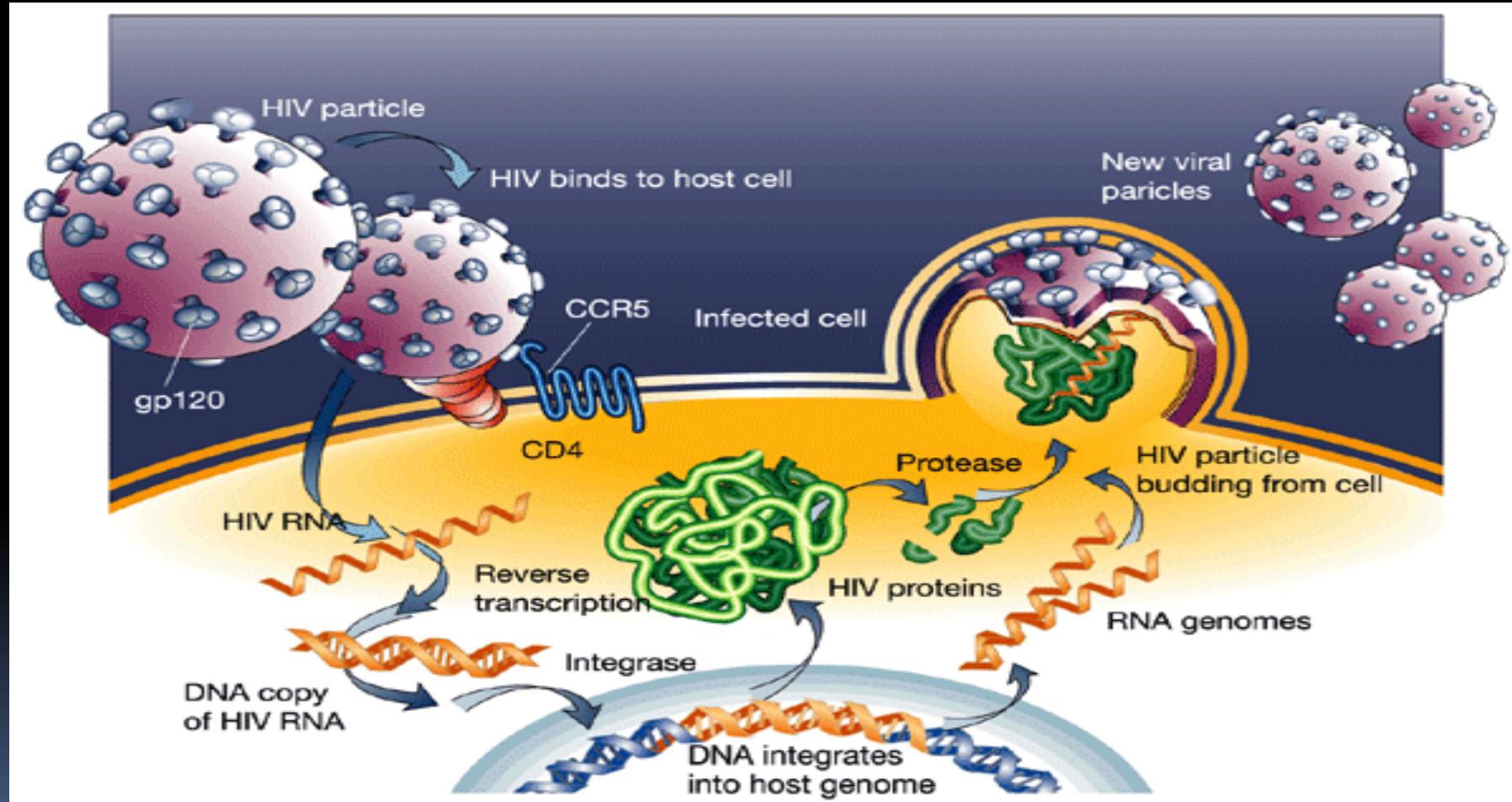
Fighting HIV with machine learning (aka statistics) and high performance computing

- HIV & immunology 101
- PhyloD.net: A tool for studying HIV
- Important discoveries toward a vaccine and possible treatments

HIV is the virus that causes AIDS

- AIDS kills 5,000 people every day
- Drugs work fairly well but are expensive and need to be taken regularly
- Vaccine is perhaps the best hope for developing countries

HIV Lifecycle



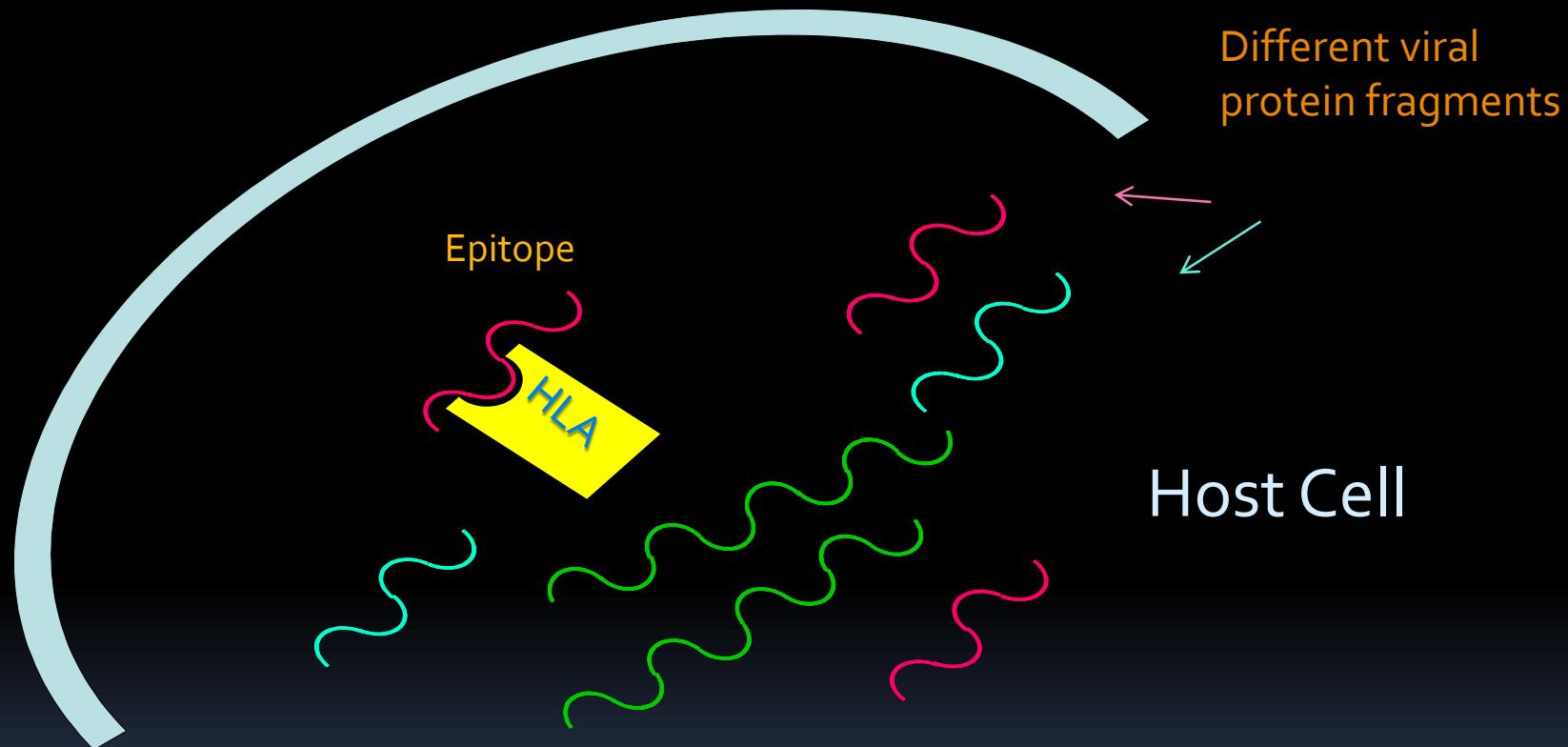
Our immune system fights viral infections

- Innate (e.g., natural killer cells)
- Adaptive
 - Antibodies (humoral arm)
 - T cells (cellular arm)

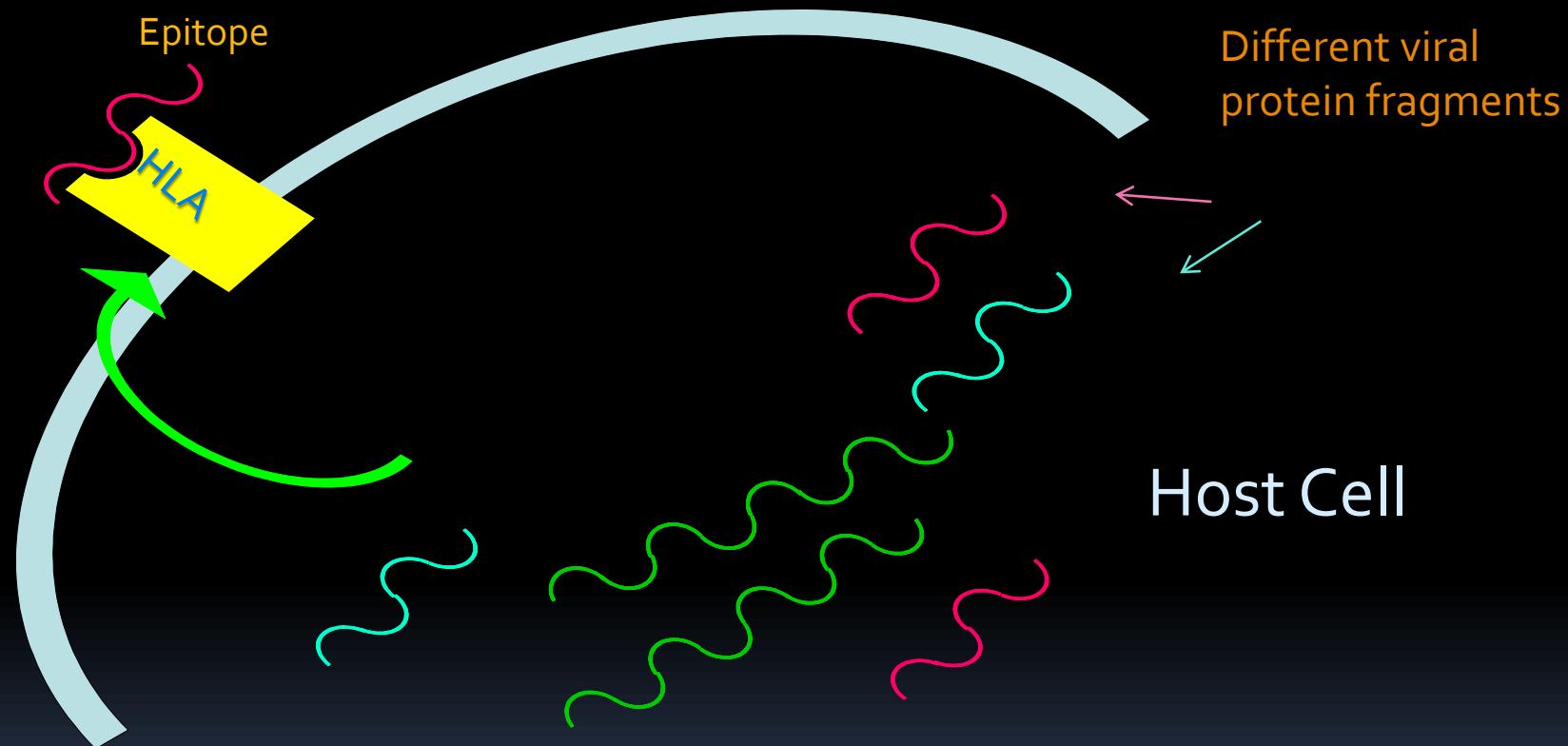


Vaccines pretrain the adaptive response thereby generating a stronger response that prevents infection or at least keeps the virus under control

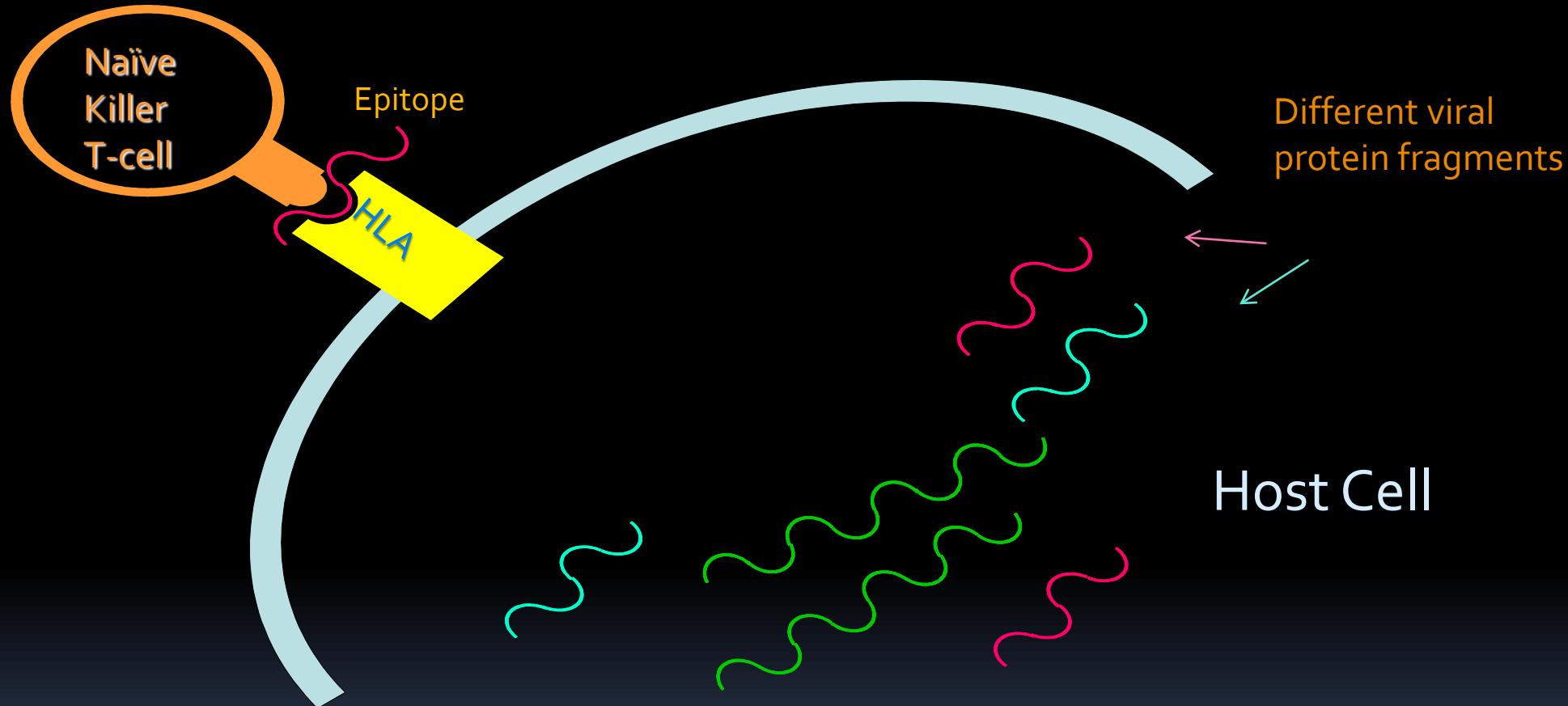
T-cells



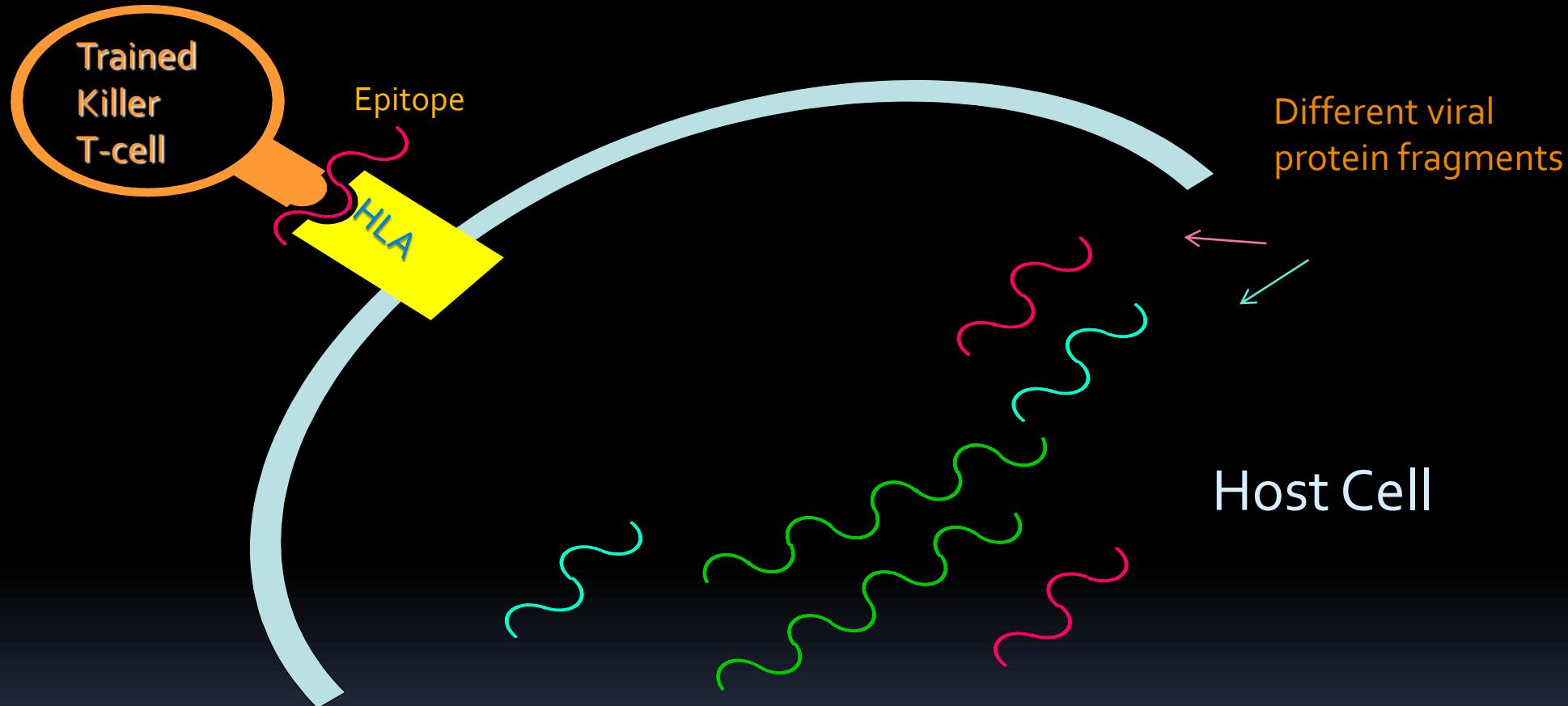
T-cells



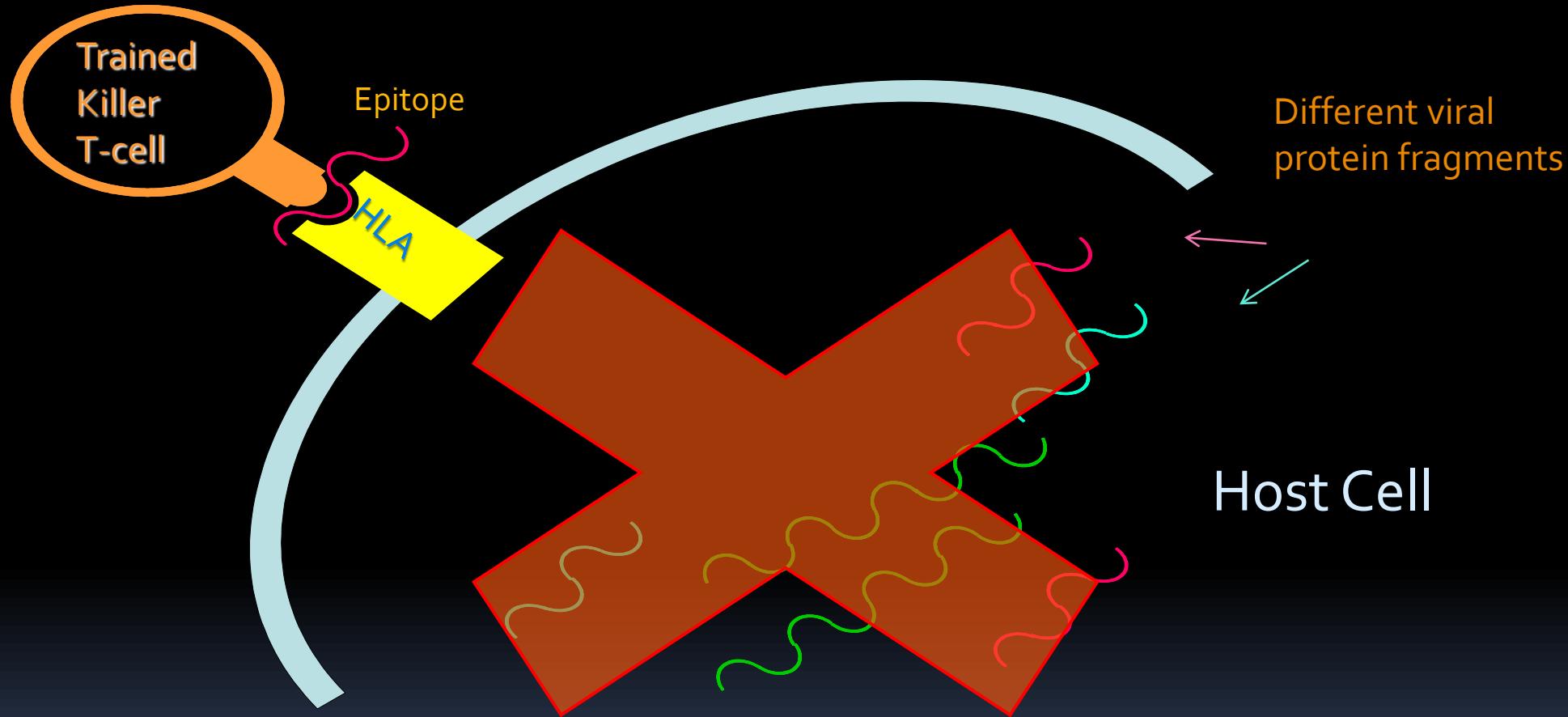
T-cells



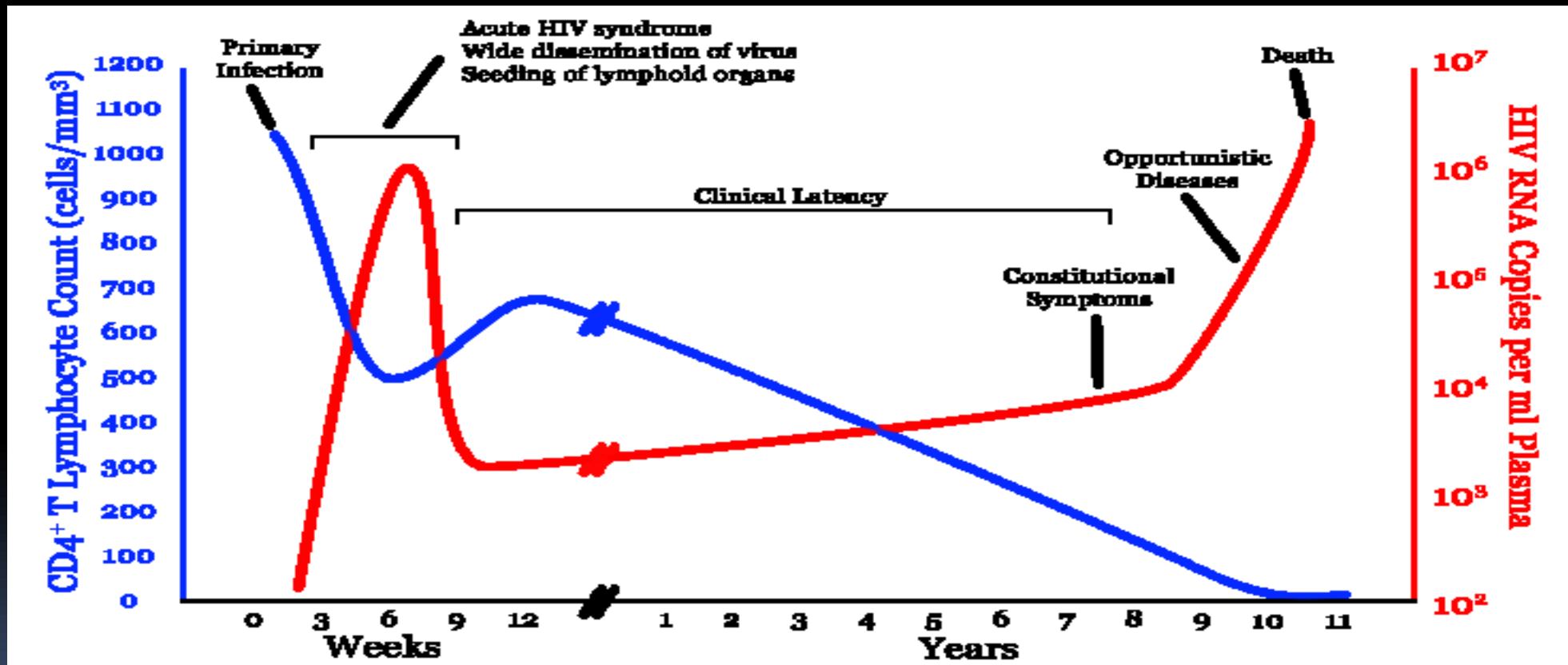
T-cells

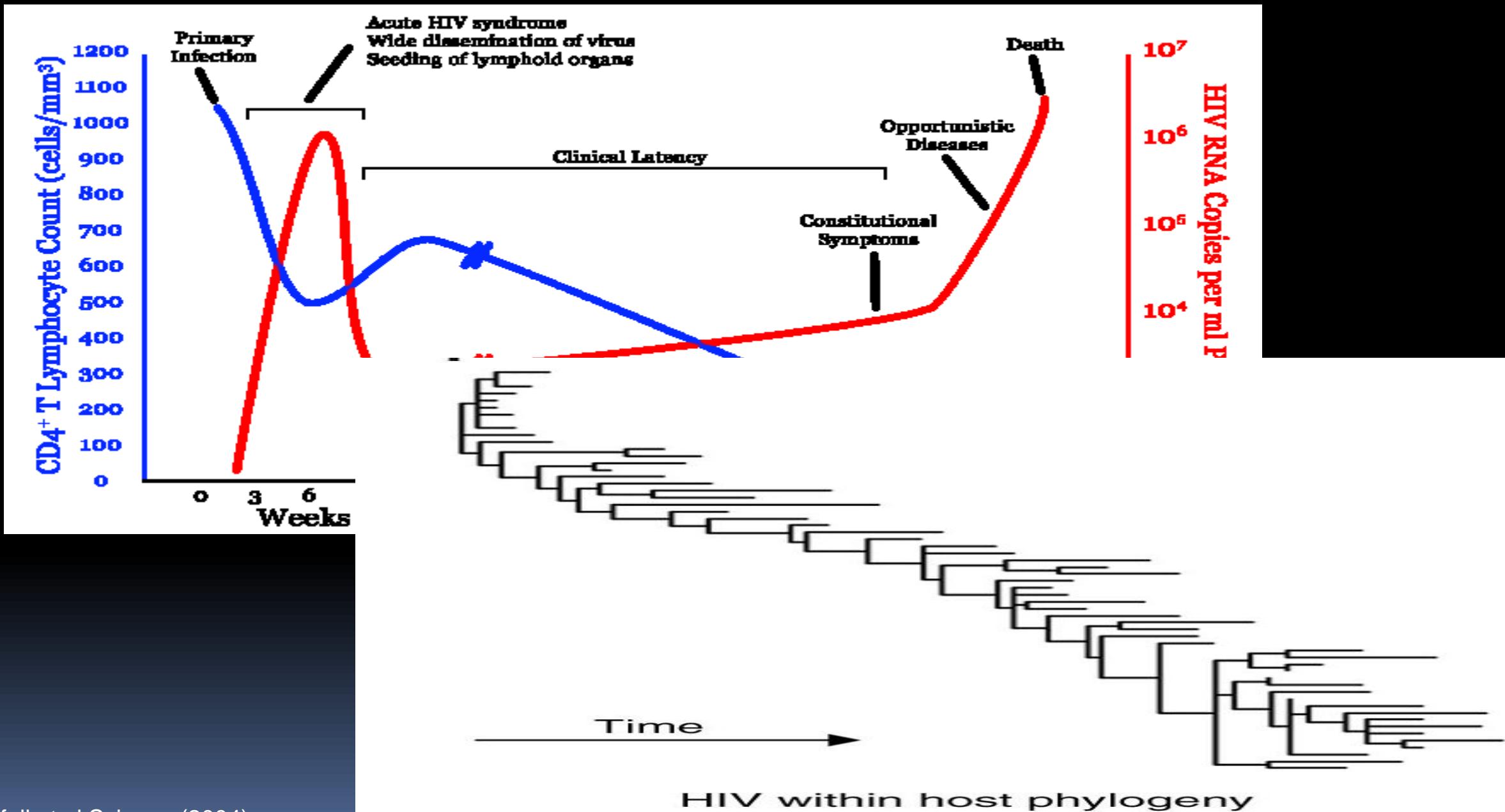


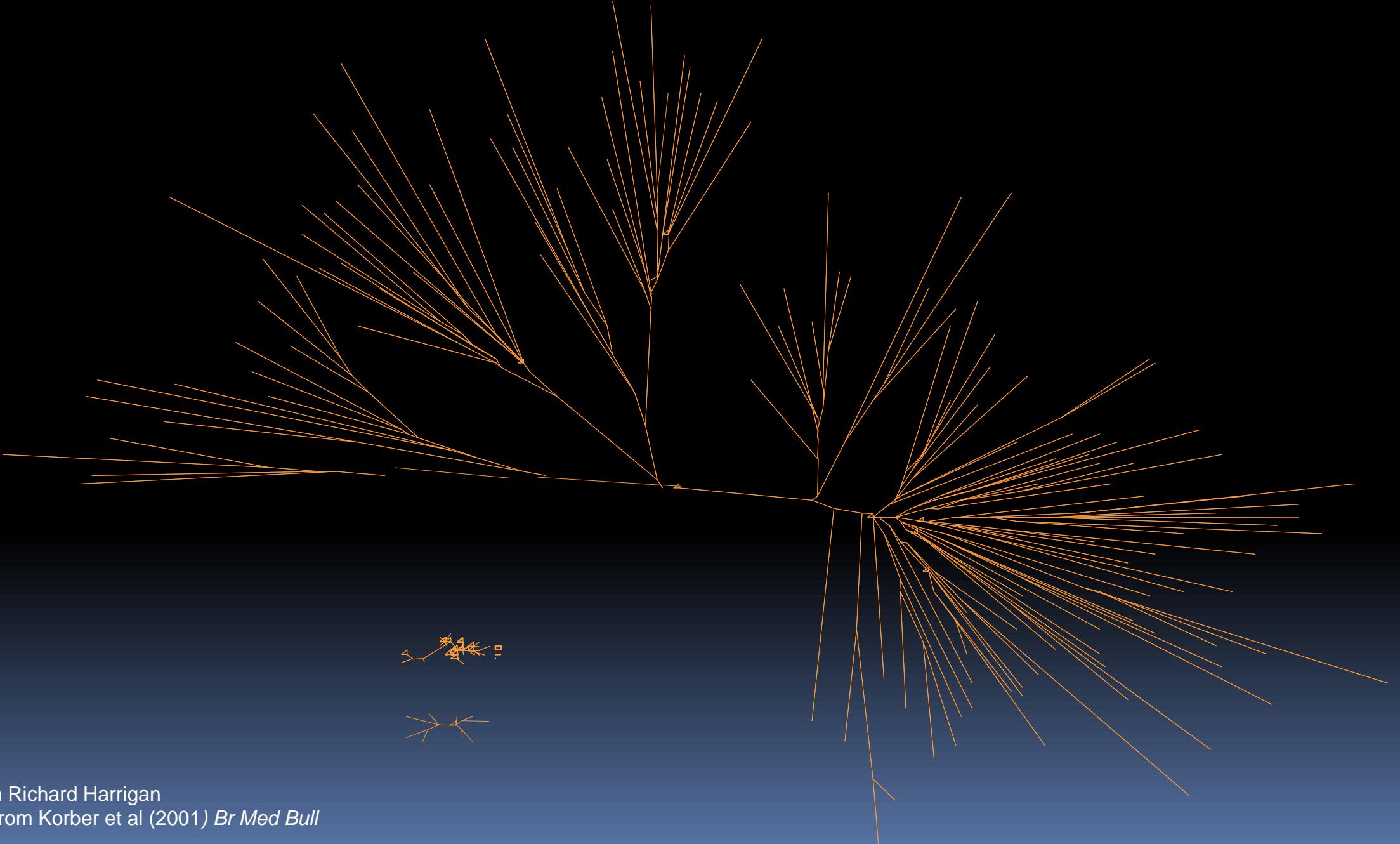
T-cells



Progression to AIDS

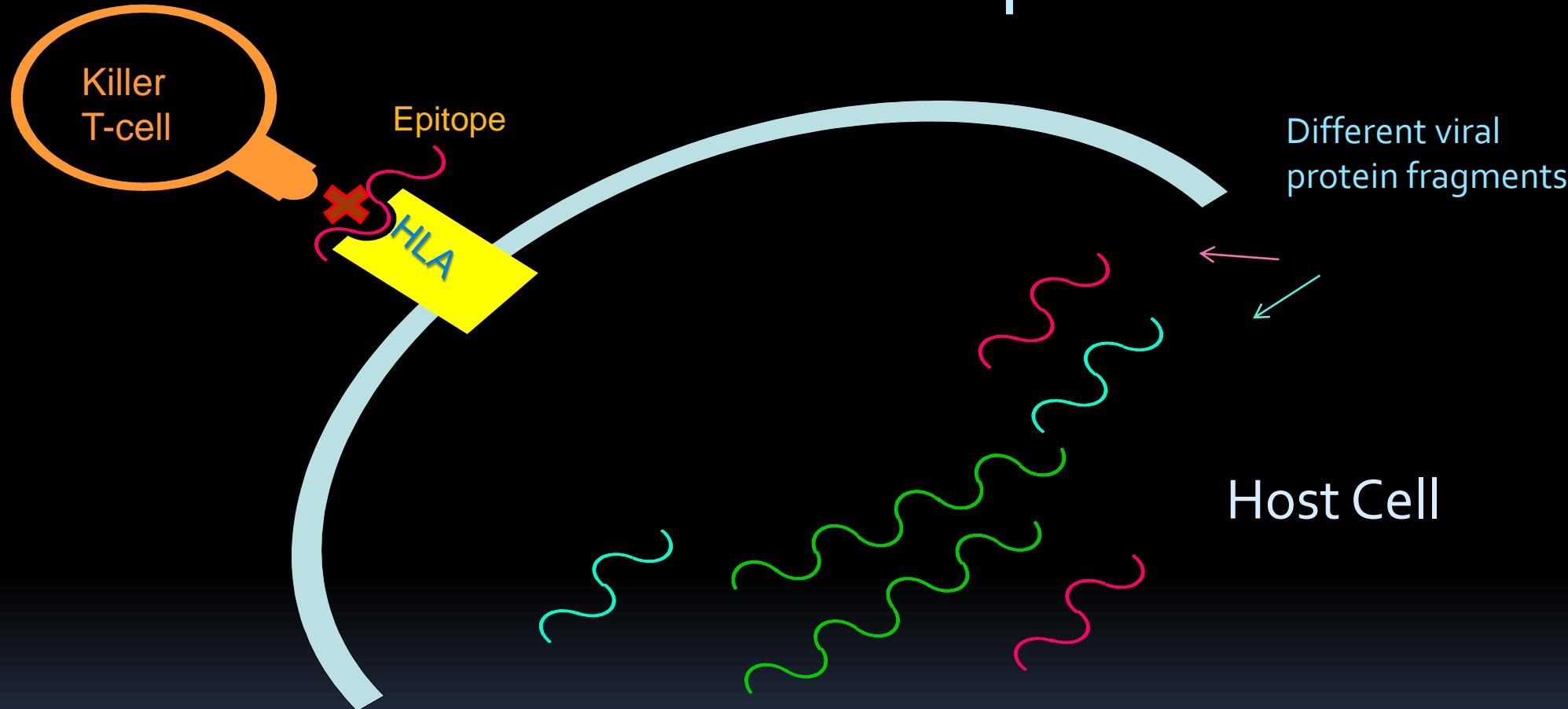


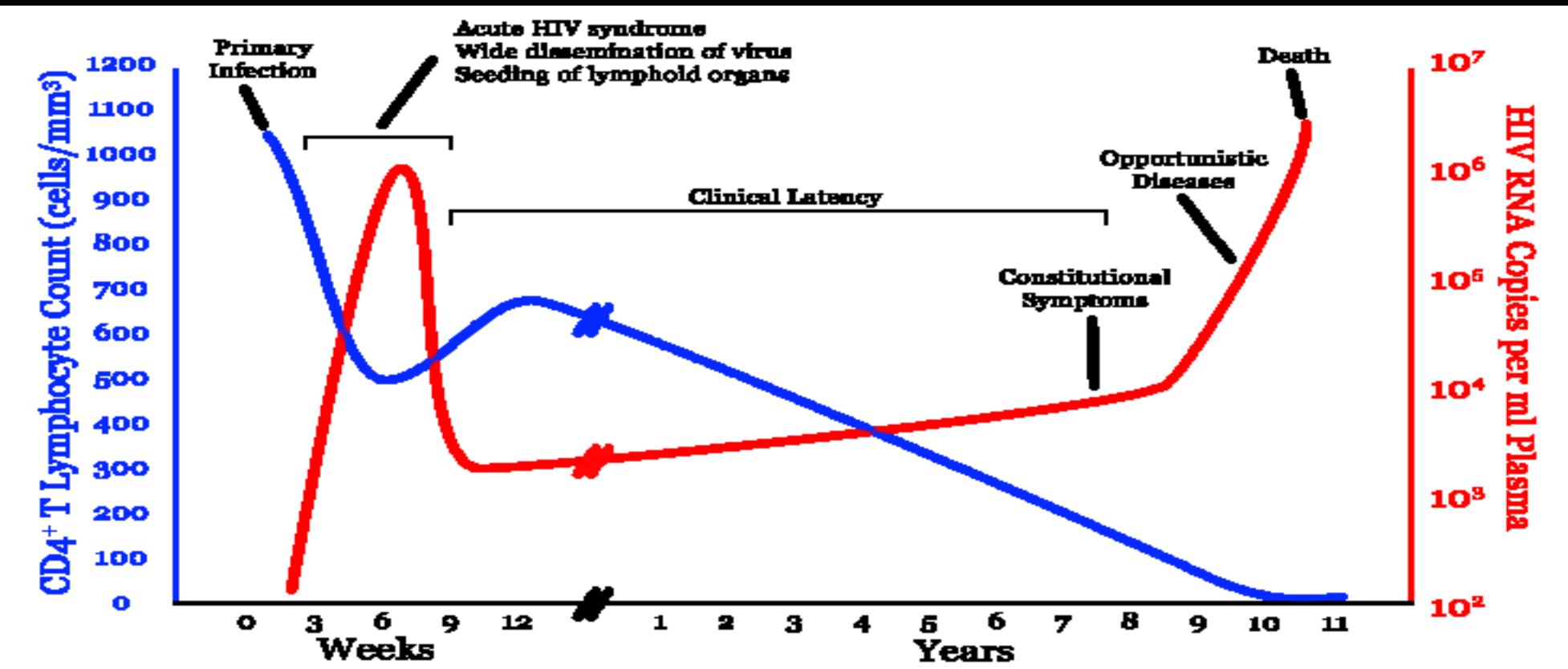




Slide from Richard Harrigan
Adapted from Korber et al (2001) *Br Med Bull*

Mechanisms of T-cell Escape





Spam filtering

The cover of the April 2005 issue of *Scientific American* features a large, bold title "STOPPING SPAM" in black letters against a background of a dense field of green grass. In the foreground, there is a circular inset showing a wooden mailbox post with a sign that says "@ mail". Below the title, the authors' names are listed: "By Joshua Goodman, David Heckerman and Robert Rounthwaite". At the bottom left, there is a small column of text: "For years, 100 million e-mail users, like the ones at left, have been plagued by spam. Computer scientists think they've cracked a way to filter out all incoming junk e-mail to protect what gets through to you." The bottom right corner of the cover has the number "42" and the word "SCIENTIFIC AMERICAN".

What can be done to stanch the flood of junk e-mail messages?

In 1978, the first spam e-mail—a plug from a marketing representative at Digital Equipment Corporation for the new DECsystem-20 computer—was dispatched to about 400 people on the Arpanet. Today such correspondence in the form of unsolicited commercial solicitations constitutes more than two-thirds of all e-mail transmitted over the Internet, accounting for billions of messages every day. For a third of all e-mail users, about 80 percent of the messages received are spam. Recently spam has become more threatening with the proliferation of so-called phishing attacks—fake e-mails that look like they are from people or institutions you trust but that are actually sent by crooks to steal your credit-card numbers or other personal information. Phishing attacks cost approximately \$1.2 billion a year, according to a 2004 Gartner Research study.

The phenomenon of spam afflicts more than just e-mail. Inside what routes, called “inboxes,” that pretend to be humans and attempt to convince people to click on links to buy and to give away personal information (IMI) scores with friends—called spIMI—a viral spam program. Phishers have upgraded by “folk spammers” who degrade Internet search engine operations by adding misleading links to sites that distract the user ratings of Web sites and links.

The suffocating effect of spam sometimes seems likely to undermine, if not wreck, Internet communications as we have come to know them. The reality, however, is not so bleak. Several techniques for intercepting spam and discovering spammers have been invented, and more are on the way. The methods we shall discuss focus on junk e-mail, but many of them could apply to other incarnations of spam as well. No one of these will be a magic cure, but combinations—if employed by enough of us—could work wonders. It is not unrealistic to hope for a day when our e-mail boxes will once again be nearly spam-free.

Insidious E-mails:

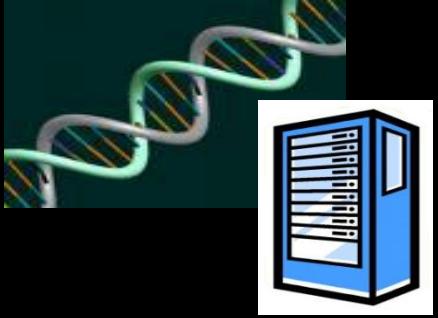
THE COST OF SPAM TO THE FRAUDULENT e-mail results directly from low costs. Spammers find spam is exceedingly cheap to distribute. It is not altogether free, though. We estimate that a message costs about one fourth cents of a cent to send. At these rate, rates a spammer can earn only \$1 per sale and still make a profit, even if the response rate is as low as one in 100,000. Hence, although very few e-mail users ever buy anything advertised in spam, all of us suffer because of those who do.

One of the most infuriating aspects of spam is that it changes continually to adapt to new attempts to stop it. Each time software engineers attack spam in some way, spammers find a way around their methods. This spammers’ rise has led to a continuous coevolution of the two, which has resulted in ever increasing sophistication on both sides.

42 SCIENTIFIC AMERICAN APRIL 2005

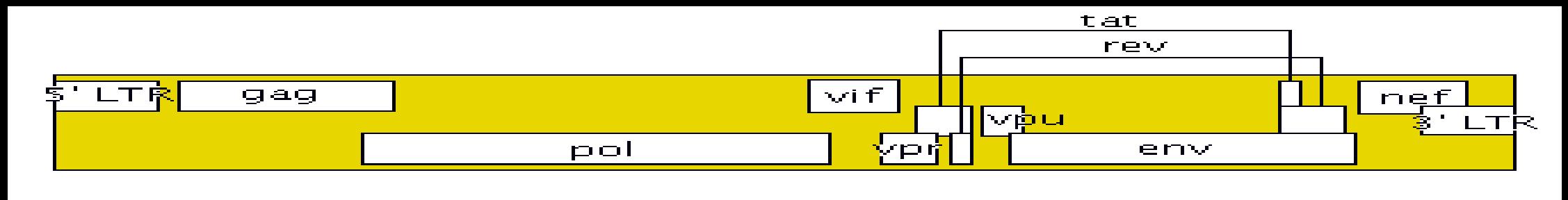
43 SCIENTIFIC AMERICAN

Goodman, Heckerman, & Rounthwaite
Scientific American, April 2005

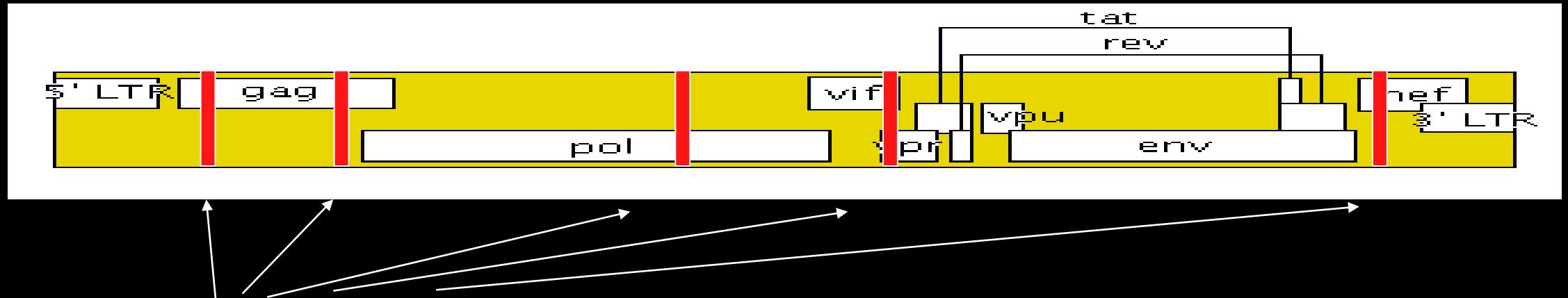


From spam filters to vaccine design

- Immune system attacks HIV: filter blocks spam
- Problem: Spammers “mutate” their emails
- Solution: Spam mail can’t be arbitrary – they are trying to sell you something
 - Look for disguised product name (e.g., “ViAGRA”)
 - Follow the money
- What is HIV’s Achilles’ heel?



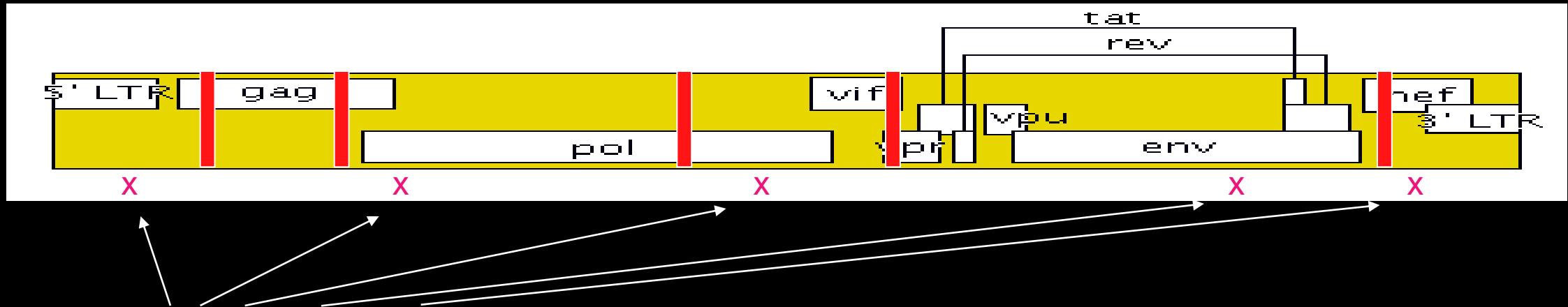
Hypothesis: Certain parts of HIV are critical to its function



If HIV mutates within these epitopes, it becomes less or non-functional

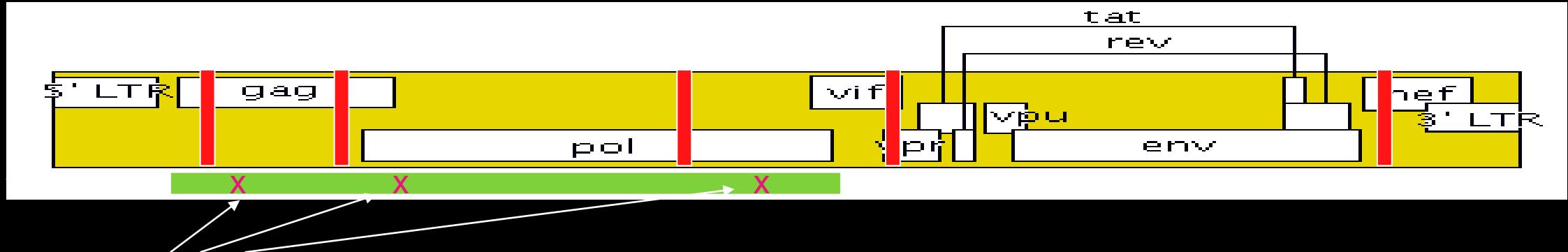
Suggests a vaccine design...

A design for an HIV vaccine



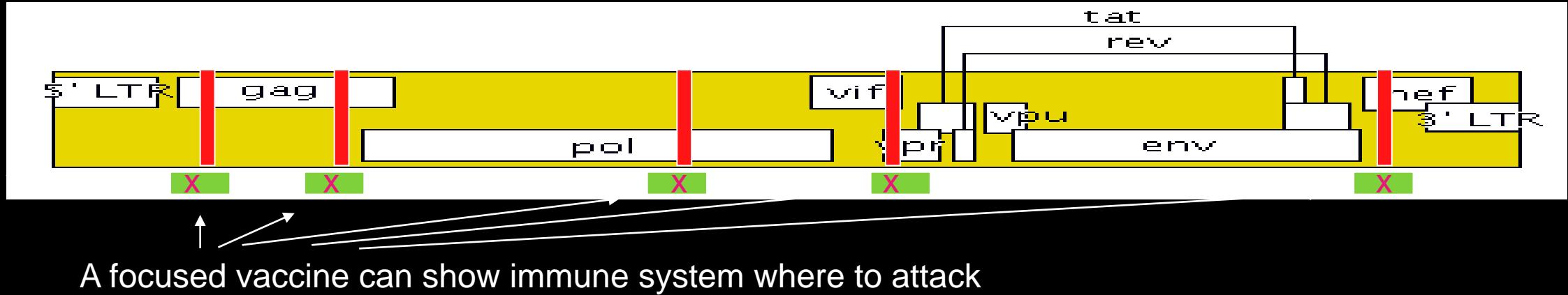
Left to its own devices, our immune system attacks at random epitopes

A design for an HIV vaccine



A “whole protein” vaccine does little to help the situation (explains failure of Merck vaccine)

A design for an HIV vaccine



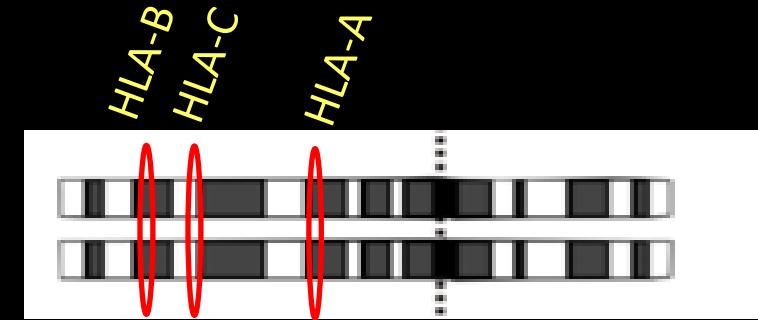
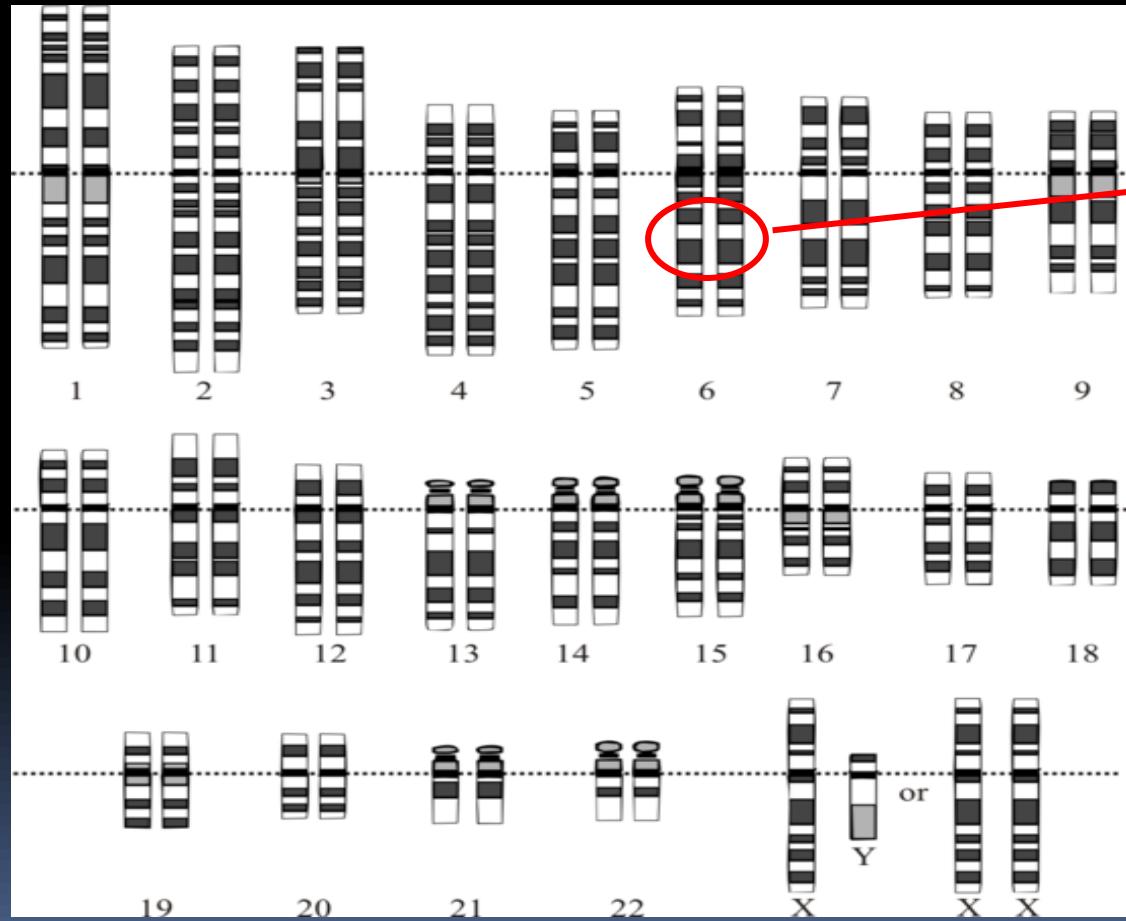
A design for an HIV vaccine

- Accumulating evidence for hypothesis and identifying these “protective epitopes”
 - Brute force testing of known epitopes (Walker and Pereyra)
- Continuing to search for these protective epitopes with more clever methods

Fighting HIV with machine learning (aka statistics) and high performance computing

- HIV & immunology 101
- PhyloD.net: A tool for studying HIV
- Important discoveries toward cures/vaccines

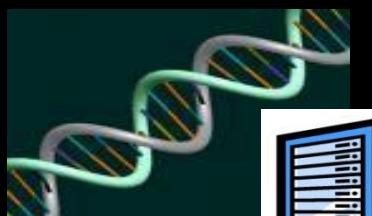
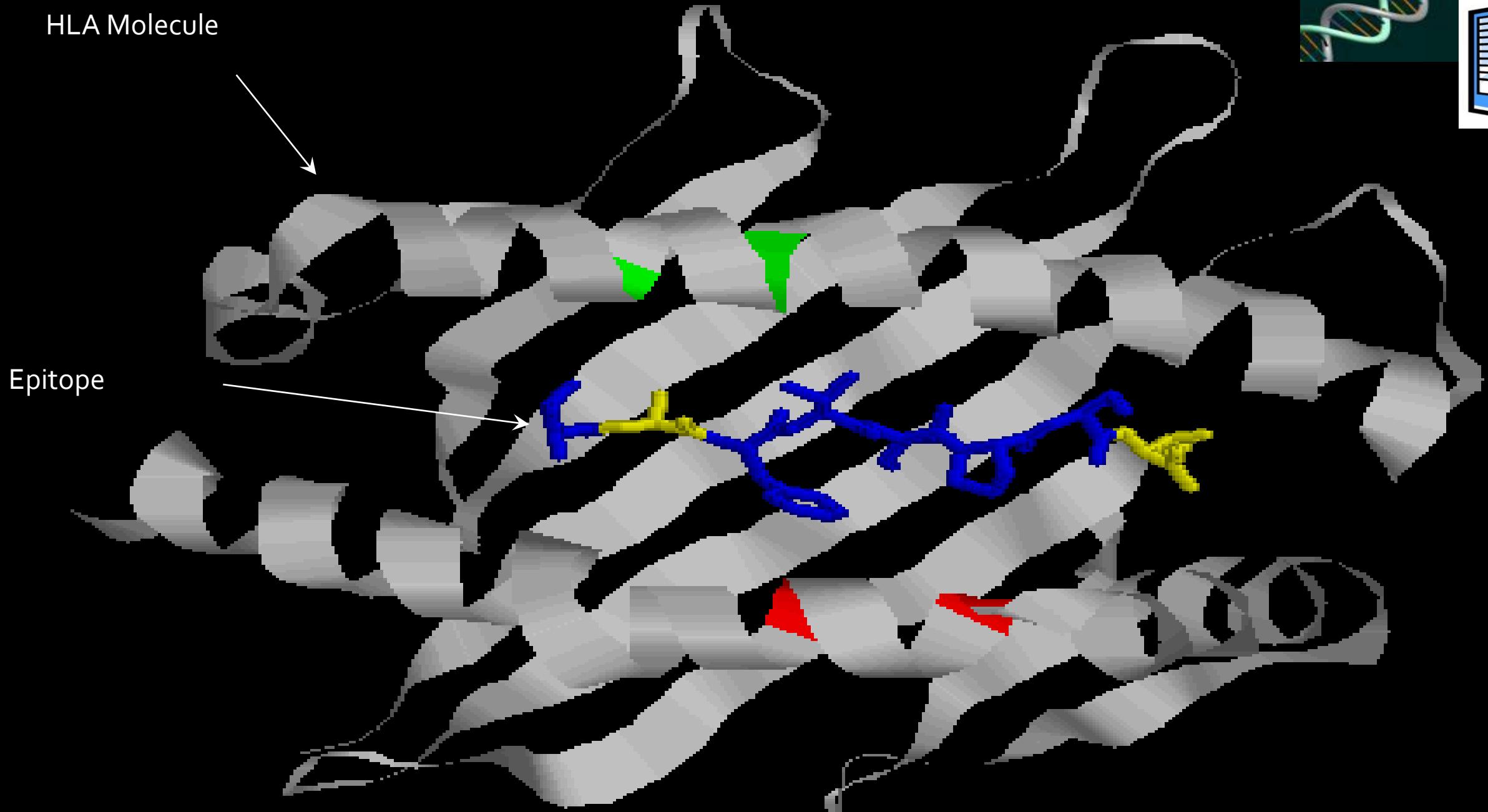
HLA variability



- Each person has up to 6 different HLA types: (2 'A', 2 'B', 2 'C')
- HLA region is most variable region of DNA--rare for two people to have the same HLA types

Epitope variability

HLA Molecule



Use this variability to search for epitopes and the HLAs that attack them

Example:

HLA B57

HIV protein

...PPGQMREPRGSDIAGT**T**STLQE**Q**I**G**W**M**TSNPP**P**VG**E**IYKR**W**I**L**GLNKIV...



...PPGQMREPRGSDIAGT**T****S****N**LQE**Q**I**G**W**M**TSNPP**P**VG**E**IYKR**W**I**L**GLNKIV...

Straightforward approach

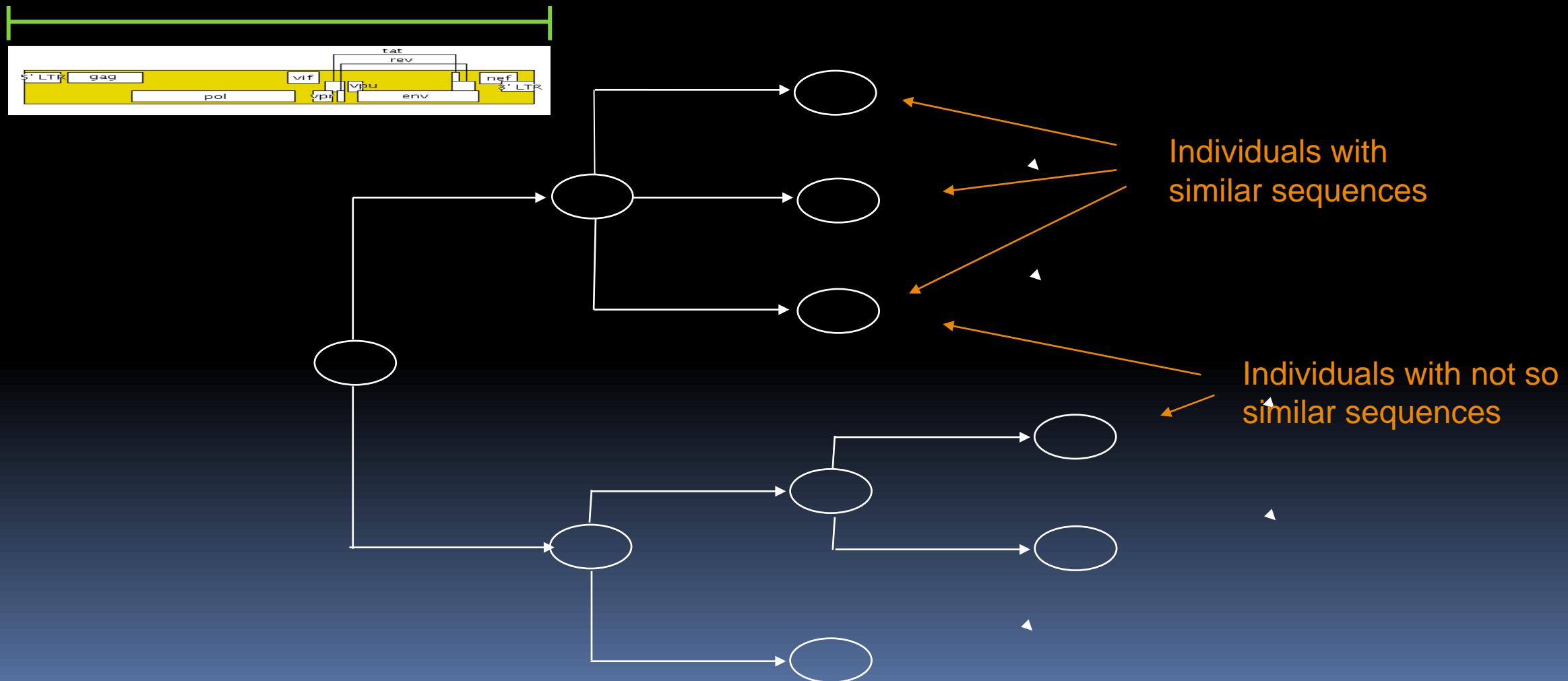
- Sequence someone's HIV when they first get infected and re-sequence every month or so
- Expensive, difficult to find subjects early infection (lucky to find 100 in a year)

Our approach: PhyloD.net

- Take a single snapshot of a person's HIV
- Use the phylogeny of sequences among individuals to **infer** the infecting sequence
- Requires machine learning algorithms and high-performance computing
- Bottom line: much less expensive and can get data from thousands of subjects

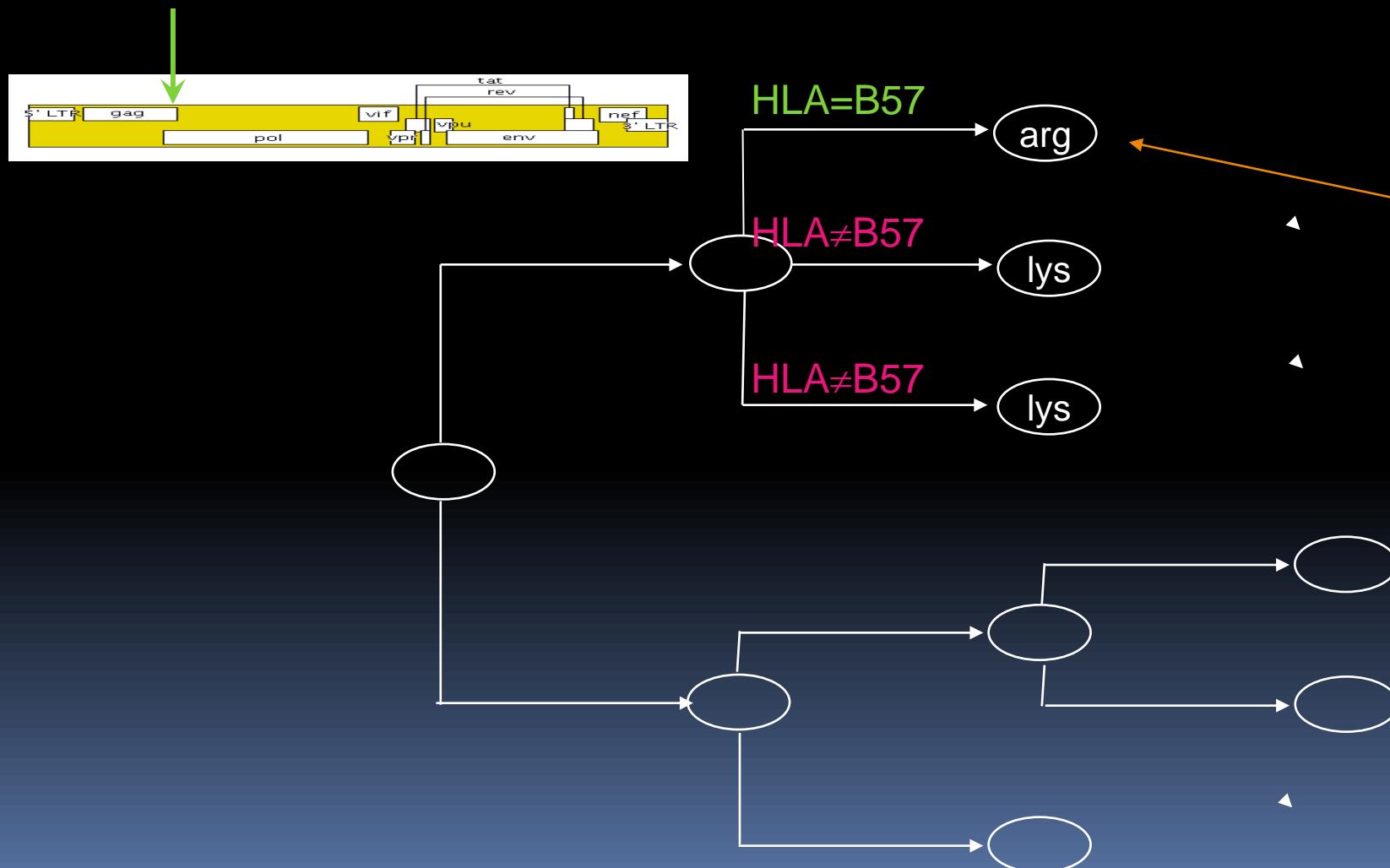
PhyloD.Net: Basic idea

Phylogeny of HIV sequences



PhyloD.Net: Basic idea

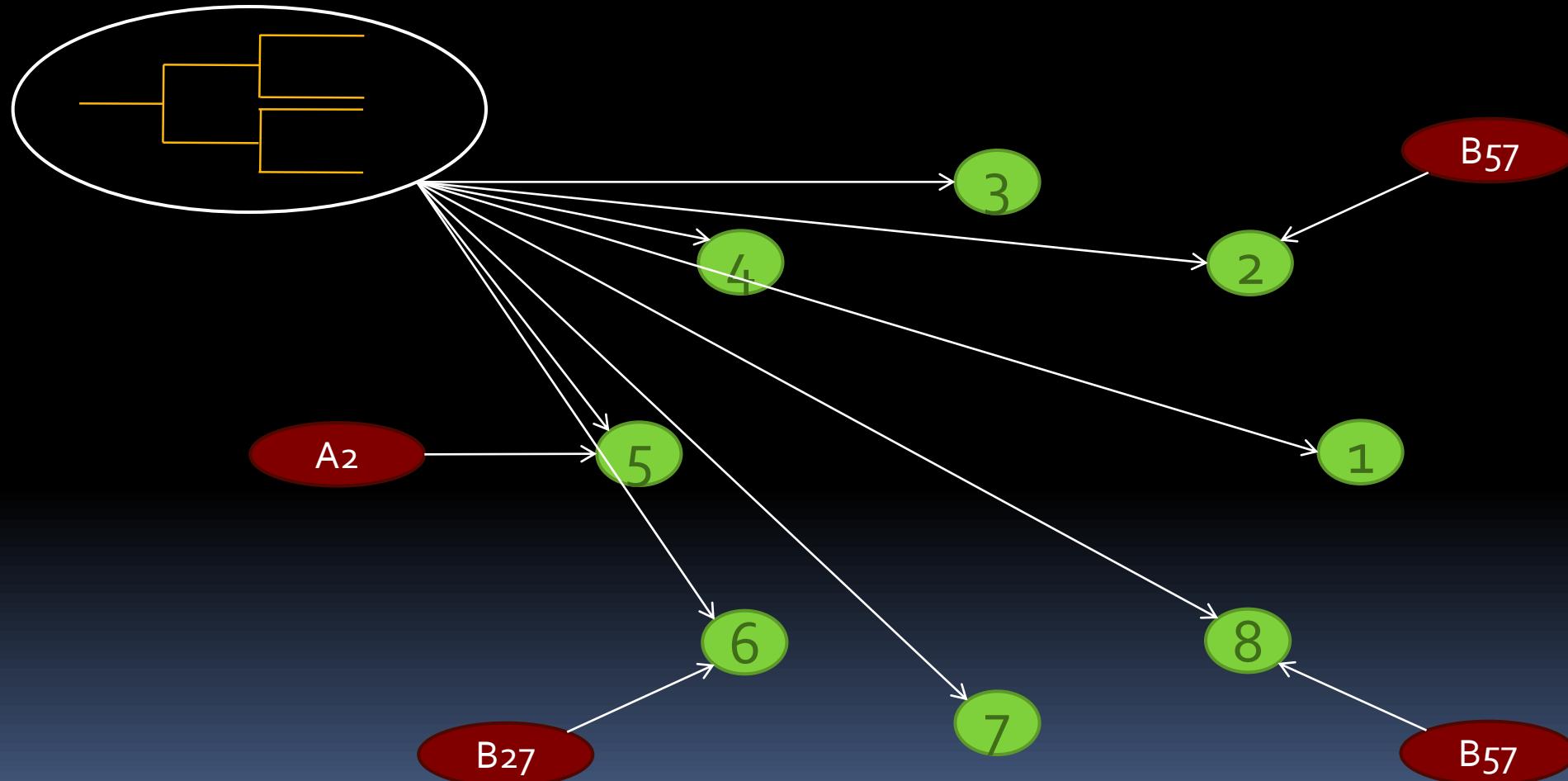
Focus on single position



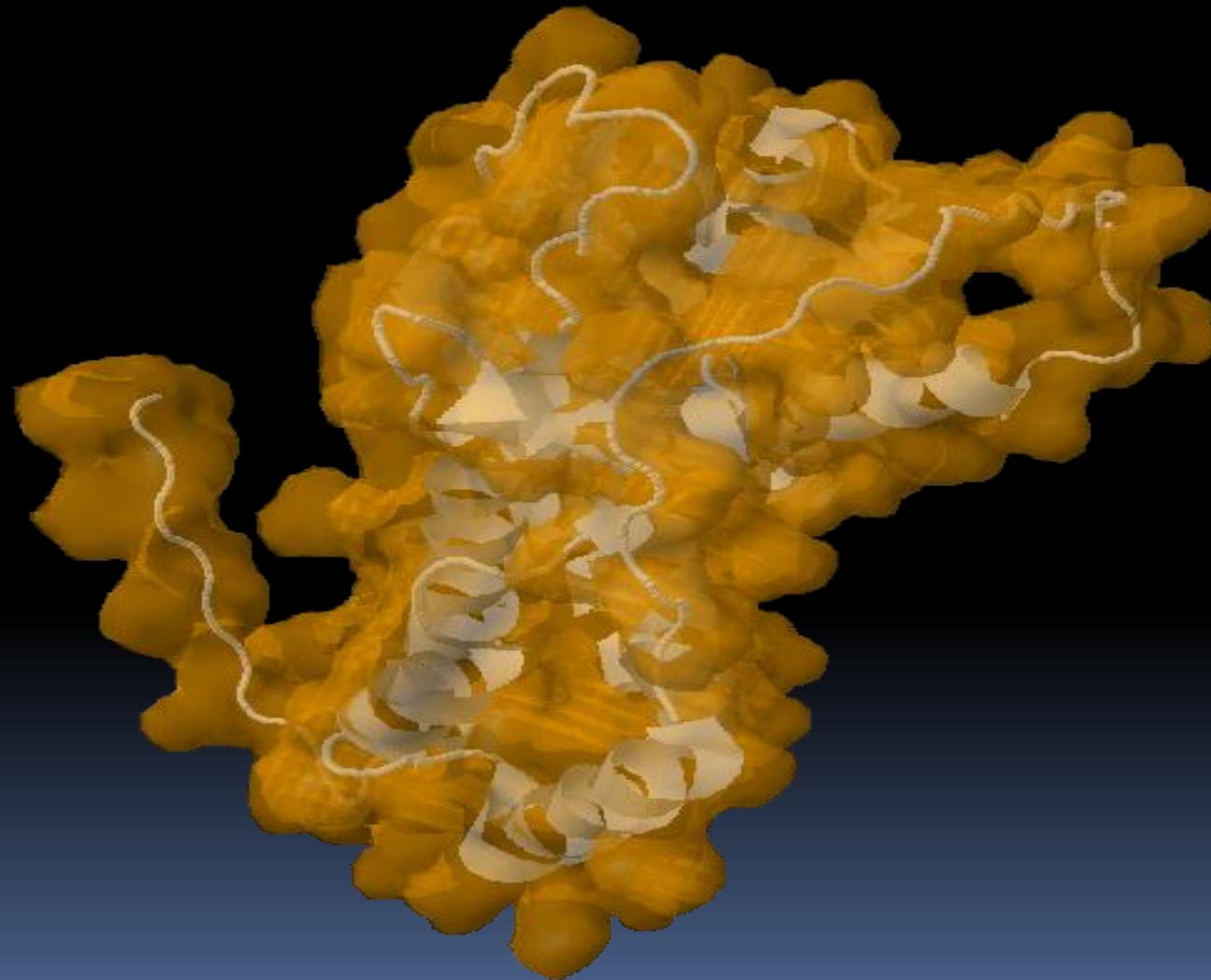
Likely that this individual
was infected with lys,
which then mutated to arg
due to HLA=B57

Multiple positions, multiple HLAs

Science 2007



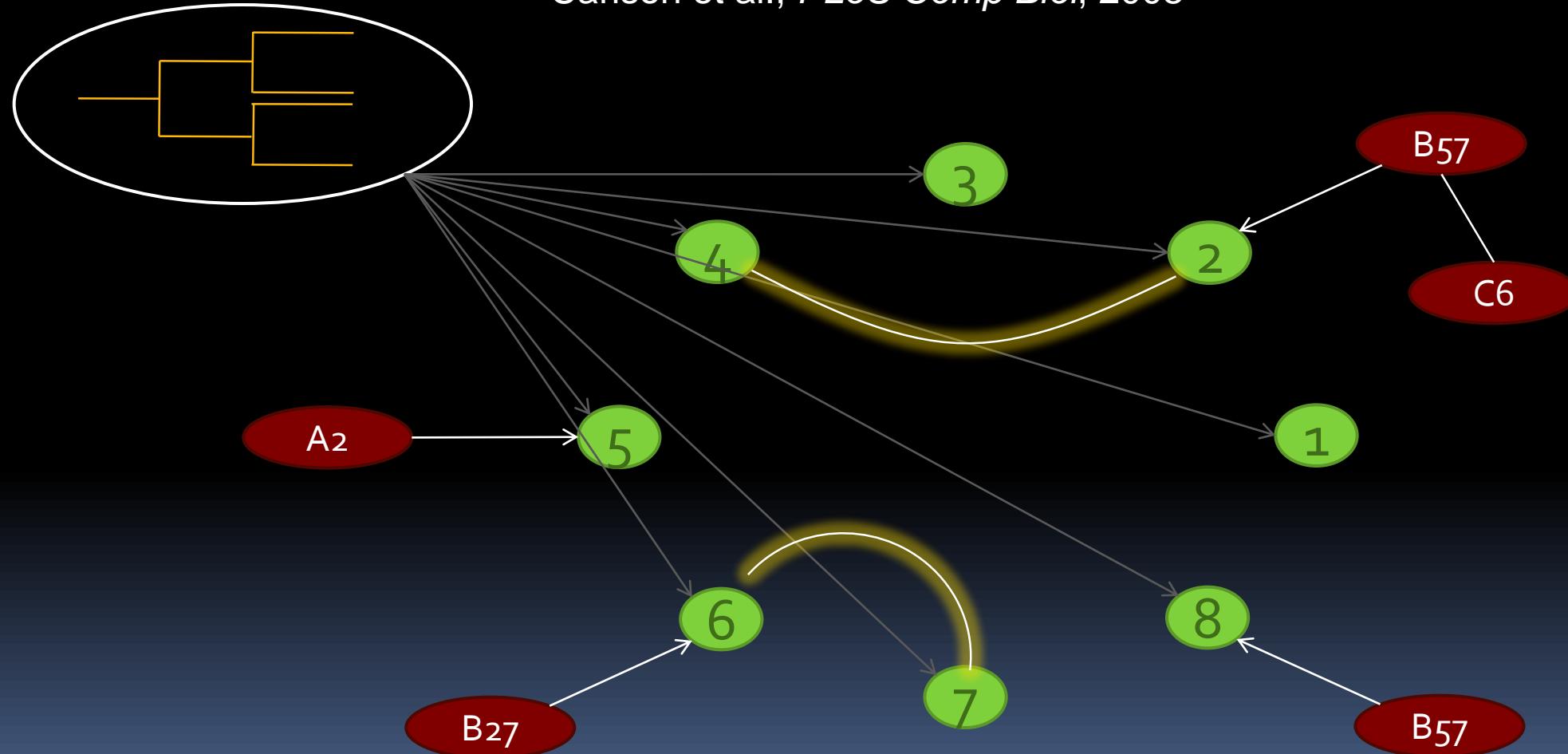
Mutations are not independent

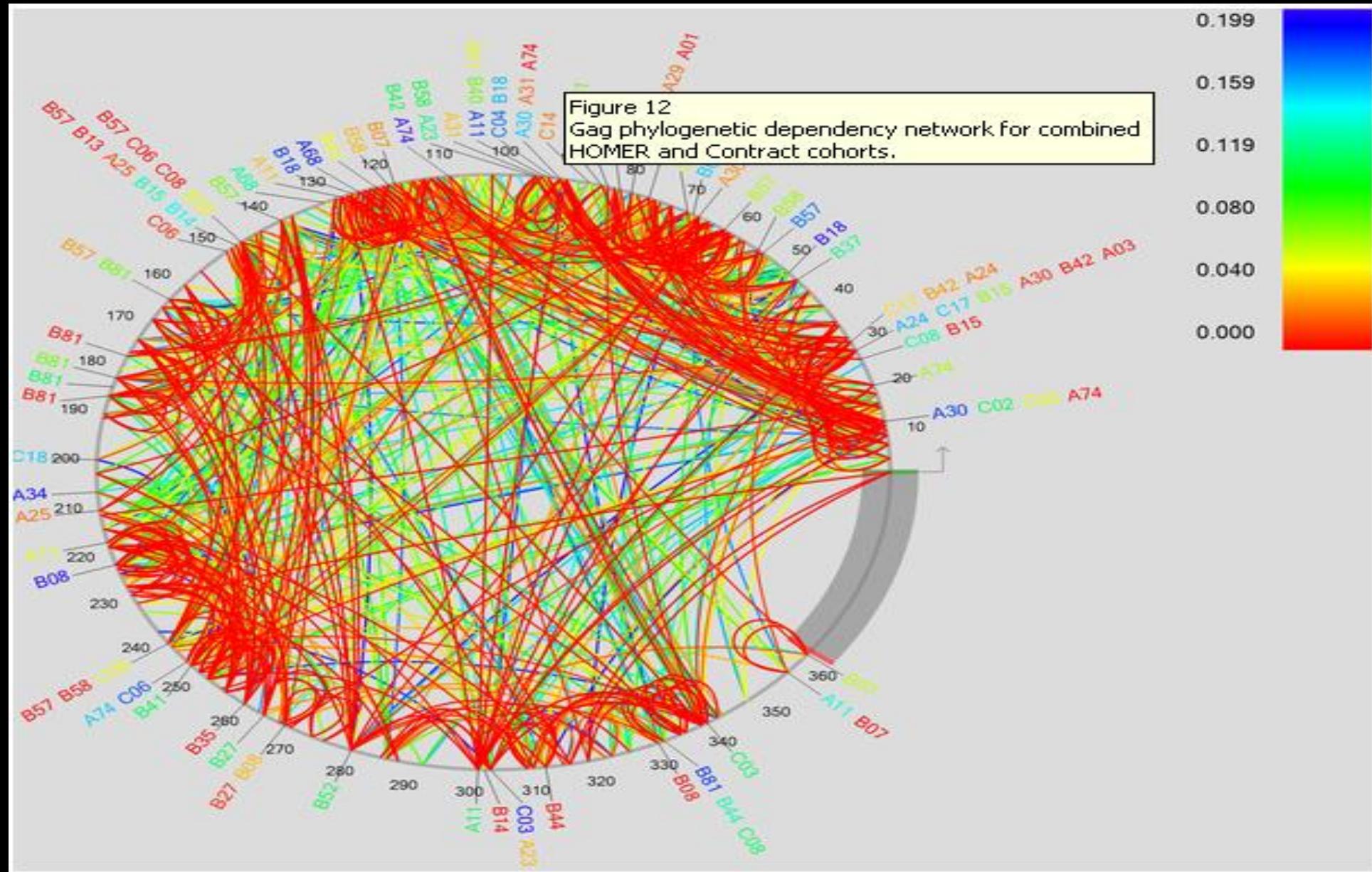


Jmol

Covariation Effects

Carlson et al., *PLoS Comp Biol*, 2008

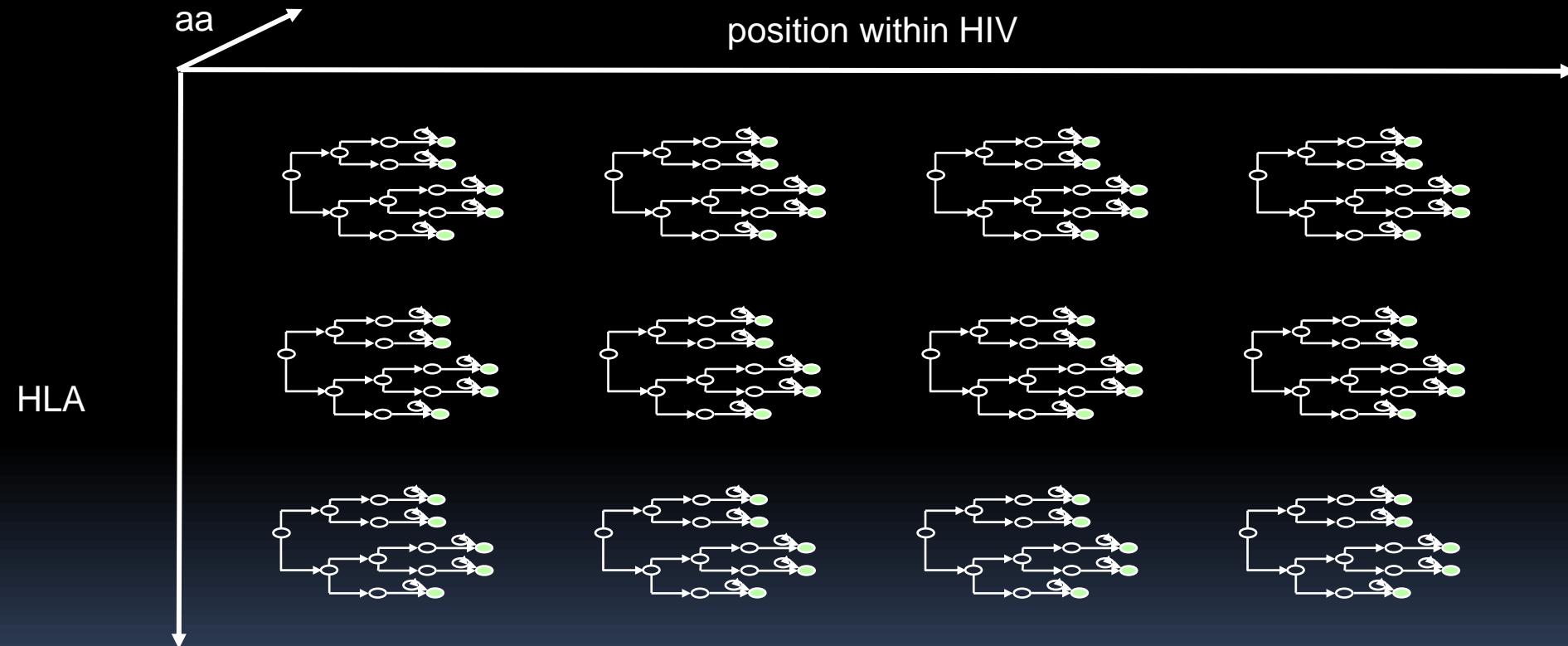




PhyloD on cover of *PLoS Comp Bio*, Nov 2008

High performance computing a must

Hundreds of thousands to millions of tests



Fortunately, the computations are pleasantly parallel

Fighting HIV with machine learning (aka statistics) and high performance computing

- HIV & immunology 101
- PhyloD.net: A tool for studying HIV
- Important discoveries toward cures/vaccines

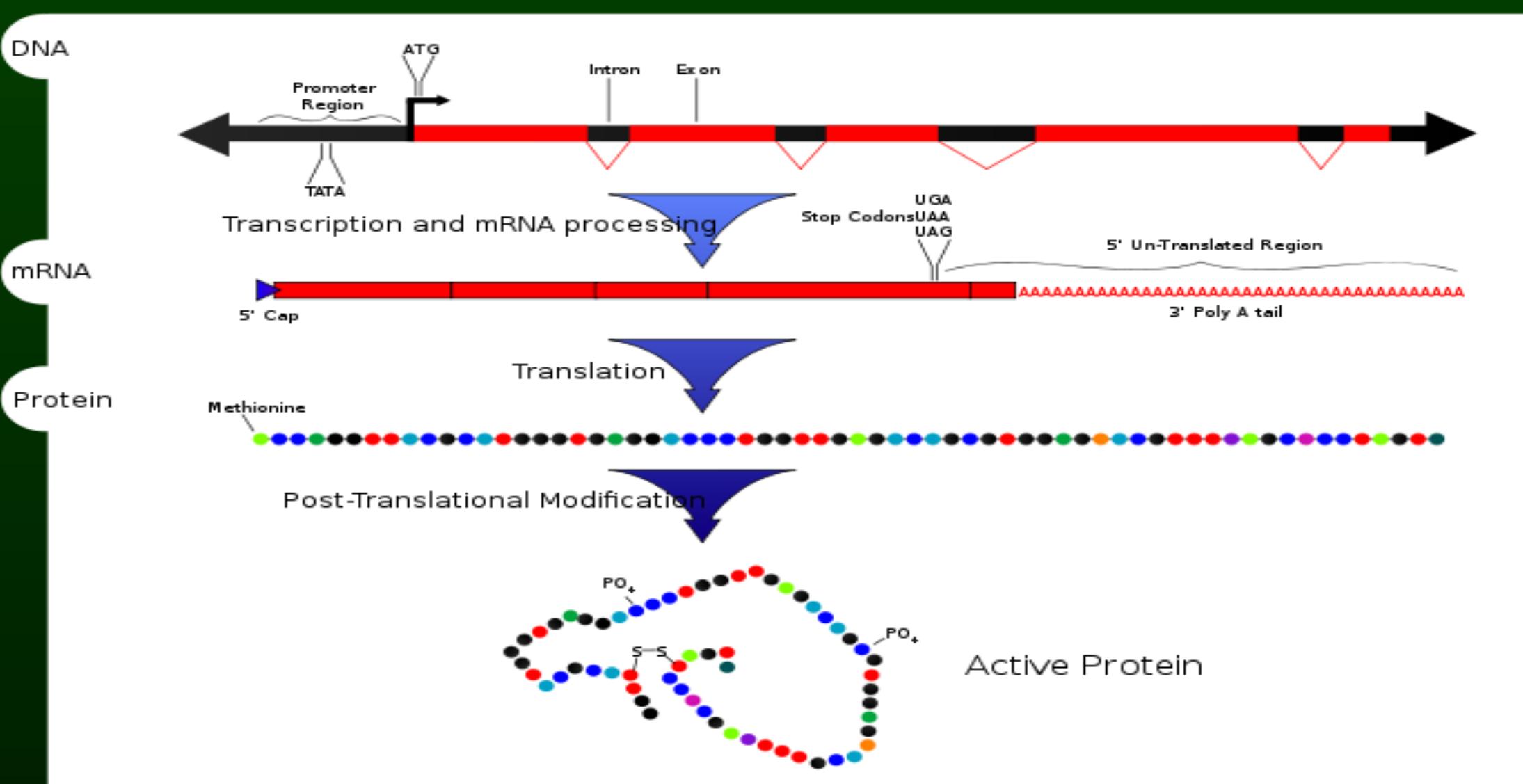
PhyloD.net publications

- M. John, D. Heckerman, I. James, L. Park, J. Carlson, A. Chopra, S. Gaudieri, D. Nolan, D. Haas, S. Riddler, R. Haubrich and S. Mallal. Adaptive interactions between HLA and HIV-1: [Highly divergent selection imposed by HLA class I molecules with common supertype motifs](#). *J. Immunology*, in press.
- A. Bansal, J. Carlson, J. Yan, O. Akinsiku, M. Schaefer, S. Sabbaj, A. Bet, D. Levy, S. Heath, J. Tang, R. Kaslow, B. Walker, T. Ndungu, P. Goulder, D. Heckerman, E. Hunter, and P. Goepfert. [CD8 T cell response and evolutionary pressure to HIV-1 cryptic epitopes derived from antisense transcription](#). *J. Exp. Med.*, 10.1084/J. Exp. Med.. 20092060, January 2010.
- C. Berger, J. Carlson, C. Brumme, K. Hartman, Z. Brumme, L. Henry, P. Rosato, A. Piechocka-Trocha, M. Brockman, P. Harrigan, D. Heckerman, D. Kaufmann, and Ch. Brander. [Viral adaptation to immune selection pressure by HLA class I-restricted CTL responses targeting epitopes in HIV frameshift sequences](#). *J. Exp. Med.*, 10.1084/J. Exp. Med.. 20091808, January 2010.
- S. Avila-Rios, C. Ormsby, J. Carlson, H. Valenzuela-Ponce, J. Blanco-Heredia, D. Garrido-Rodriguez, C. Garcia-Morales, D. Heckerman, Z. Brumme, S. Mallal, M. John, E. Espinosa, and G. Reyes-Teran. [Unique features of HLA-mediated HIV evolution in a Mexican cohort: a comparative study](#). *Retrovirology*, 6:72doi:10.1186/1742-4690-6-72, August 2009.
- Z. Brumme, M. John, J. Carlson, C. Brumme, D. Chan, M. Brockman, L. Swenson, I. Tao, S. Szeto, P. Rosato, J. Sela, C. Kadie, N. Frahm, C. Brander, D. Haas, S. Riddler, R. Haubrich, B. Walker, P. Harrigan, D. Heckerman, and S. Mallal. [HLA-associated immune escape pathways in HIV-1 subtype B Gag, Pol and Nef proteins](#). *PLoS ONE*, 4(8):e6687. doi:10.1371/journal.pone.0006687. August 2009.
- Y. Kawashima, K. Pfafferott, J. Frater, P. Matthews, R. Payne, M. Addo, H. Gatanaga, M. Fujiwara, A. Hachiya, H. Koizumi, N. Kuse, S. Oka, A. Duda, A. Prendergast, H. Crawford, A. Leslie, Z. Brumme, C. Brumme, T. Allen, C. Brander, R. Kaslow, J. Tang, E. Hunter, S. Allen, J. Mulenga, S. Branch, T. Roach, M. John, S. Mallal, A. Ogwu, R. Shapiro, J. Prado, S. Fidler, J. Weber, O. Pybus, P. Klennerman, T. Ndung'u, R. Phillips, D. Heckerman, P. Harrigan, B. Walker, M. Takiguchi, and P. Goulder. [Adaptation of HIV-1 to human leukocyte antigen class I](#). *Nature*, February 2009.
- J. Carlson, Z. Brumme, C. Rousseau, C. Brumme, P. Matthews, C. Kadie, J. Mullins, B. Walker, P. Harrigan, P. Goulder, D. Heckerman. [Phylogenetic dependency networks: Inferring patterns of CTL escape and codon covariation in HIV-1 Gag](#). *PLoS Computational Biology*, 5(10):e1000530. doi:10.1371/journal.pcbi.1000530. October 2009.
- Y. Wang, B. Li, J. Carlson, H. Streeck, A. Gladden, R. Goodman, A. Schneidewind, K. Power, I. Toth, N. Frahm, G. Alter, C. Brander, M. Carrington, B. Walker, M. Altfeld, D. Heckerman, and T. Allen. [Protective HLA class I alleles restricting acute-phase CD8⁺ T cell responses are associated with viral escape mutations located in highly conserved regions of HIV-1](#). *Journal of Virology*, November, 2008.
- D. Yerly, D. Heckerman, T. Allen, T. Suscovich, N. Jovic, C. Kadie, W. Pichler, A. Cerny, and C. Brander. [Design, expression, and processing of epitomized hepatitis C virus-encoded CTL epitopes](#). *Journal of Immunology*, 181:6361-6370, November, 2008.
- P. Matthews, A. Prendergast, A. Leslie, H. Crawford, R. Payne, C. Rousseau, I. Honeyborne, J. Carlson, C. Kadie, C. Brander, J. Mullins, H. Coovadia, T. Ndung'u, B. Walker, D. Heckerman, P. Goulder. [Central role of reverting mutations in HLA associations with human immunodeficiency virus set point](#). *Journal of Virology*, 82:8548-59, September, 2008.
- Z. Brumme, C. Brumme, J. Carlson, H. Streeck, M. John, O. Eshlebaum, B. Block, B. Baker, C. Kadie, M. Markowitz, H. Jessen, A. Kelleher, E. Rosenberg, J. Kaldor, Y. Yuki, M. Carrington, T. Allen, S. Mallal, M. Altfeld, D. Heckerman, and B. Walker. [Marked epitope- and allele-specific differences in rates of mutation in human immunodeficiency type 1 \(HIV-1\) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection](#). *Journal of Virology*, 82:9216-9227, September 2008.
- Nature**
J. Timm, A. Berical, N. Lennon, A. Berlin, S. Young, B. Lee, D. Heckerman, J. Carlson, L. Reyor, M. Kleyman, C. McMahon, C. Birch, J. Schulze Zur Wiesch, T. Ledlie, M. Koehrsen, C. Kodira, A. Roberts, G. Lauer, H. Rosen, F. Bihl, A. Cerny, U. Spengler, Z. Liu, A. Kim, Y. Xing, A. Schneidewind, M. Madey, J. Fleckenstein, V. Park, J. Galagan, C. Nusbaum, B. Walker, G. Lake-Bakaar, E. Daar, I. Jacobson, E. Gomperts, B. Edlin, S. Donfield, R. Chung, A. Talal, T. Marion, B. Birren, M. Henn, T. Allen. [Naturally occurring dominant resistance mutations to hepatitis C virus protease and polymerase inhibitors in treatment-naïve patients](#). *Hepatology*, 48:1769-1778, July 2008.
- T. Miura, M. Brockman, C. Brumme, Z. Brumme, J. Carlson, F. Pereyra, A. Trocha, M. Addo, B. Block, A. Rothchild, B. Baker, T. Flynn, A. Schneidewind, B. Li, Y. Wang, D. Heckerman, T. Allen, and B. Walker, [Genetic Characterization of Human Immunodeficiency Virus type 1 in Elite Controllers: Lack of gross genetic defects or common amino acid changes](#), in *Journal of Virology*, pp. JV1.00535-08, 2008
- Z. Brumme, I. Tao, S. Szeto, C. Brumme, J. Carlson, D. Chan, C. Kadie, N. Frahm, C. Brander, B. Walker, D. Heckerman, and P. Harrigan. [HLA-specific polymorphisms in HIV-1 Gag and their association with viral load in chronic untreated](#). *AIDS*, 22(11):1277-1286, July, 2008.
- P. Goepfert, W. Lumm, P. Farmer, P. Matthews, A. Prendergast, J. Carlson, C. Derdeyn, J. Tang, R. Kaslow, A. Bansal, K. Yusim, D. Heckerman, J. Mulenga, S. Allen, P. Goulder, and E. Hunter. [Transmission of HIV-1 Gag immune escape mutations is associated with reduced viral load in linked recipients](#). *Journal of Experimental Medicine*, 205(5):1009-1017, April, 2008.
- C. Rousseau, M. Daniels, J. Carlson, C. Kadie, H. Crawford, A. Prendergast, P. Matthews, D. Raugi, B. Maust, G. Learn, D. Nickle, N. Frahm, C. Brander, B. Walker, P. Goulder, T. Bhattacharya, D. Heckerman, B. Korber, and J. Mullins. [Class-I driven evolution of human immunodeficiency virus type 1 subtype C proteome: Immune escape and viral load](#). *Journal of Virology*, doi:10.1128/JVI.02455-07, April 2008.
- D. Yerly, D. Heckerman, T. Allen, J. Chisholm, K. Faircloth, C. Linde, N. Frahm, J. Timm, W. Pichler, A. Cerny, and C. Brander. [Increased CTL epitope variant cross-recognition and functional avidity are associated with HCV clearance](#). *J. Virol*, January 2008.
- N. Kholiswa, C. Day, Z. Mncube, K. Nair, D. Ramduth, C. Thobakgale, E. Moodley, S. Reddy, C. de Pierres, N. Mkhwanazi, K. Bishop, M. van der Stok, N. Ismail, I. Honeyborne, H. Crawford, D. Kavanagh, C. Rousseau, D. Nickle, J. Mullins, D. Heckerman, B. Korber, H. Coovadia, P. Goulder, and B. Walker. [Targeting of a CD8 T cell Env epitope presented by HLA-B*5802 is associated with markers of HIV disease progression and lack of selection pressure](#). *AIDS Res Hum Retroviruses*. 2008 Jan;24(1):72-82.
- Z. Brumme, C. Brumme, D. Heckerman, B. Korber, M. Daniels, J. Carlson, C. Kadie, T. Bhattacharya, C. Chui, T. Mo, R. Hogg, J. Montaner, N. Frahm, C. Brander, B. Walker, P. Harrigan. [Evidence of Differential HLA Class I-Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1](#). *PLoS Pathogens*, 3(7): e94, July 2007.
- J. Carlson, C. Kadie, S. Mallal, and D. Heckerman. [Leveraging hierarchical population structure in discrete association studies](#). *PLoS ONE*, 2(7): e1591, July 2007.
- C. Rousseau, G. Learn, T. Bhattacharya, D. Nickle, D. Heckerman, S. Chetty, C. Brander, P. Goulder, B. Walker, P. Kiepiela, B. Korber, and J. Mullins. [Extensive intrasubtype recombination in South African Human Immunodeficiency Virus type I subtype C infections](#). *J Virol*, 81(9): 4492-4500. May, 2007.
- T. Bhattacharya, M. Daniels, D. Heckerman, B. Foley, N. Frahm, C. Kadie, J. Carlson, K. Yusim, B. McMahon, B. Gaschen, S. Mallal, J. Mullins, D. Nickle, J. Herbeck, C. Rousseau, G. Learn, T. Miura, C. Brander, B. Walker, B. Korber. [Founder effects in the assessment of HIV polymorphisms and HLA allele associations](#). *Science*, 315, 1583-1586, March 16 2007.

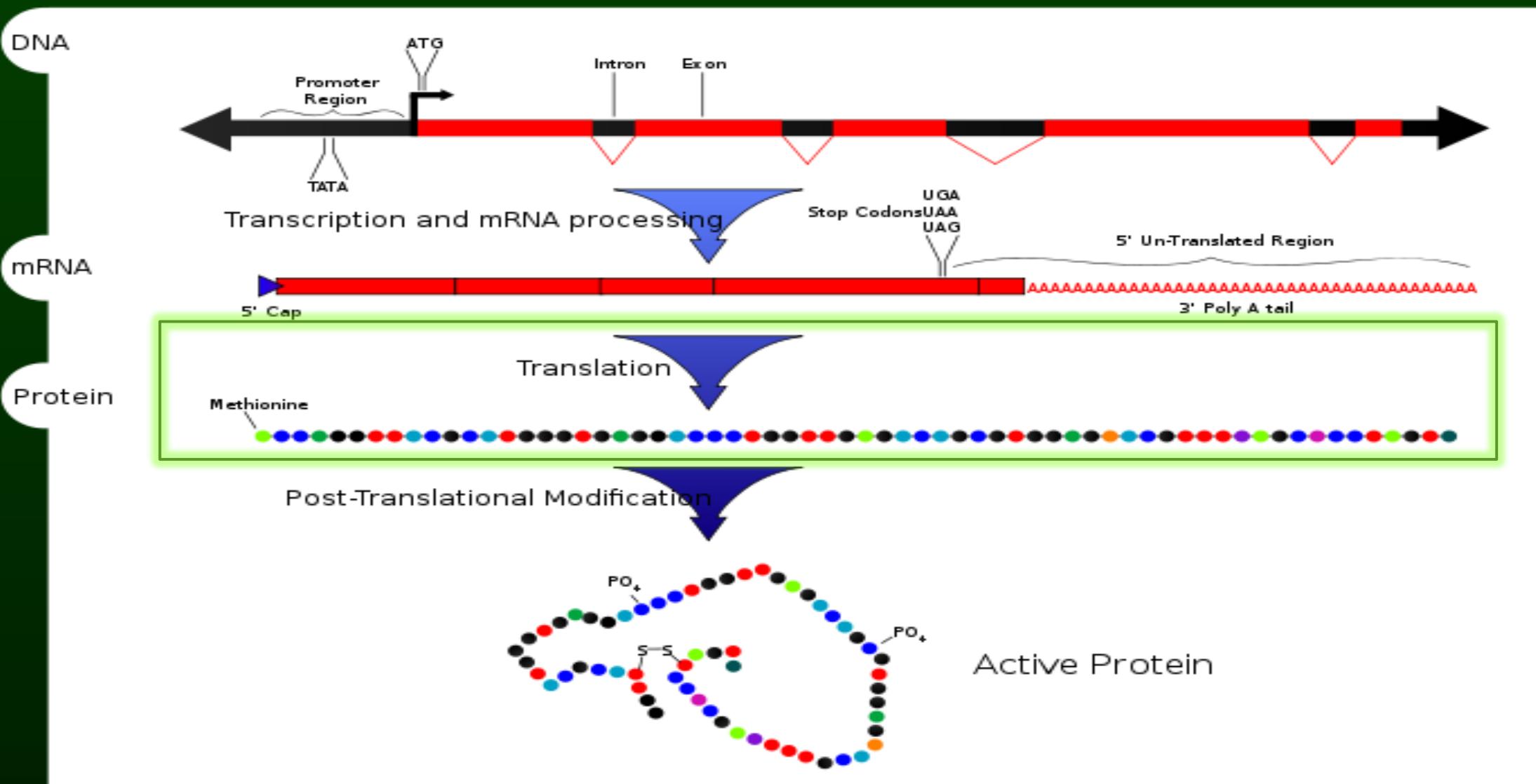
What do they tell us?

- Identifying more normal epitopes (11)
- Identifying novel class of epitope targets (2)
- Identifying novel immune responses (1)

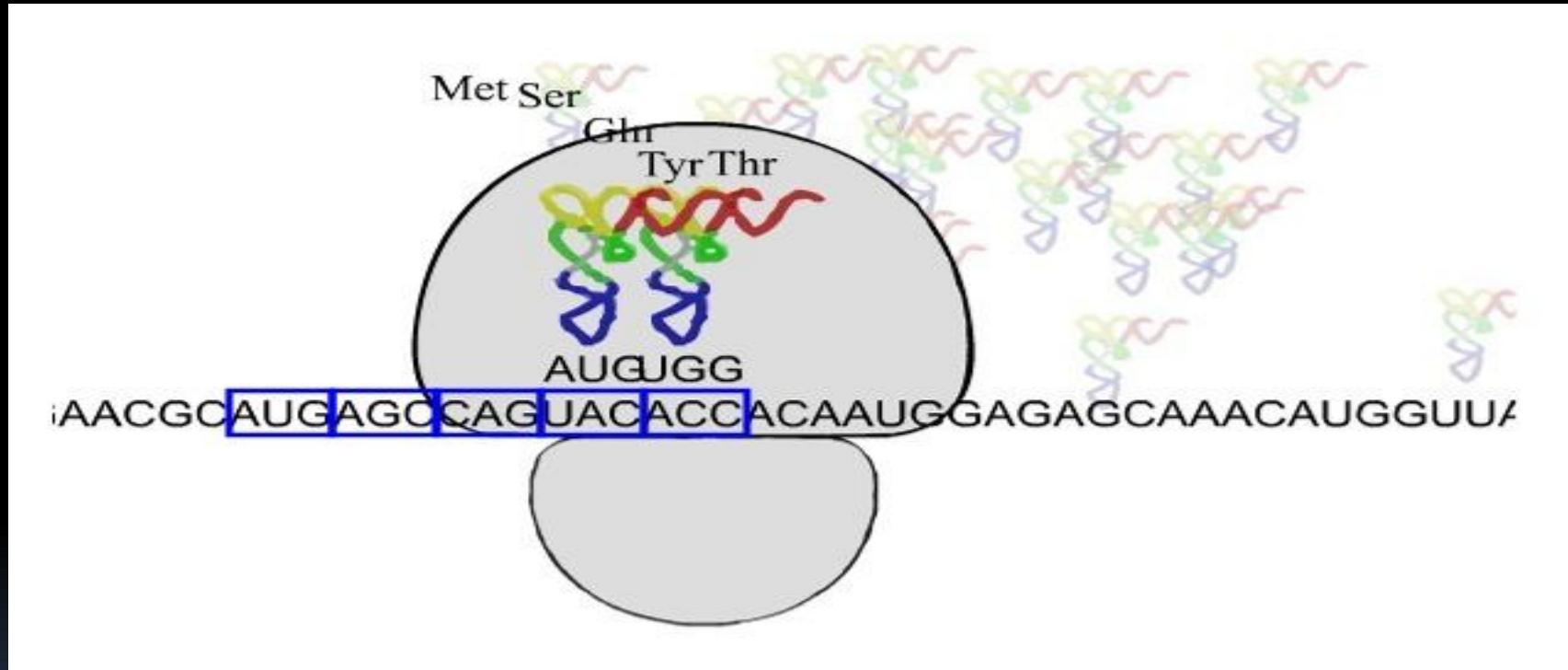
Central Dogma of Molecular Biology : Eukaryotic Model



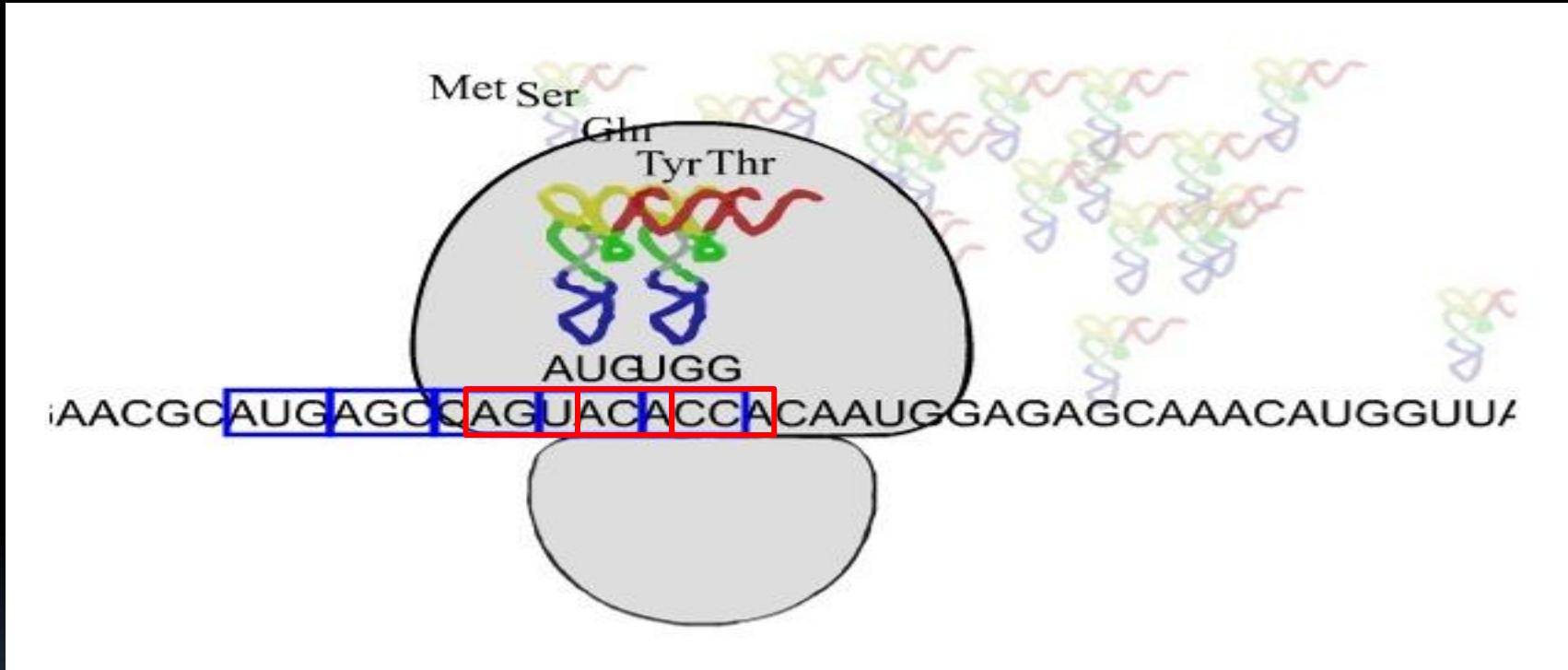
Major errors in translation



Translation



Alternate reading frames lead to gibberish



This gibberish produces a lot of epitopes targeted by the immune system

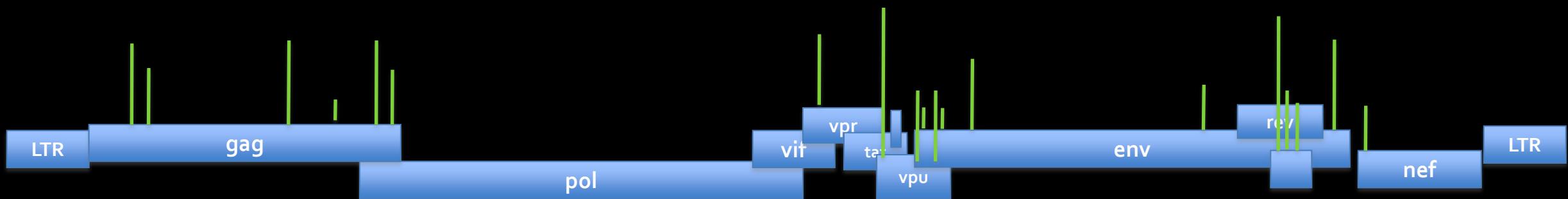
Table 1. Cryptic epitopes predicted based on HLA Class I associated HIV-1 polymorphism

ARF ^a	Protein	HLA	Best Epitope ^b	Peptide	PP ^c	PHI ^d	CHI ^e
2	pol	B*3910	---ELKTFGREF---	EF8	30%	1 (0)	0 (0)
3	gag	A*01	---SSHANVKRY---	SY9*	54%	7 (0)	7 (0)
3	pol	B*15	---YRYSISRTY---	YY9	40%	1 (0)	3 (0)
3	pol	C*07	---YRFSISRAY---	YY9*	20%	13 (1)	17 (2)
3	pol	C*07	---NLWKKGYRE---	NF9	30%	13 (0)	17 (2)
4	pol	A*3001	---FCFPPWYYL---	FL9	51%	2 (0)	9 (0)
4	pol	A*34	---NIPCFSYF---	NF8*	26%	2 (0)	4 (0)
4	pol	B*15	---LCFYVATGY---	LY9	31%	1 (0)	3 (2)
4	pol	B*36	---SPAILFWQL---	SL9	30%	4 (0)	7 (0)
4	pol	B*42	---LPKSDLREW---	LV9	51%	1 (0)	2 (0)
5	pol	A*0205	---SVNCFTSLV---	SV9*	35%	14 (0)	20 (0)
5	gag	A*3001	---CLQPSDVSK---	CK9*	29%	2 (0)	9 (1)
5	pol	A*3002	---AYFPVFRFL---	AL9*	28%	2 (0)	9 (1)
5	pol	A*33	---TGHLPANE---	TF8	26%	1 (0)	4 (0)
5	nef	A*6801	---SLTAGHPTM---	SM9	30%	1 (0)	8 (1)
5	gag	B*08	---FPHFQQPF---	FF8*	36%	5 (1)	9 (0)
5	pol	B*36	---IPNAYCESV---	IV9	42%	4 (0)	7 (0)
5	pol	B*5802	---ASFIWPPTF---	AF9	40%	4 (1)	5 (2)
5	gag	C*0801	---NVAPGPNAL---	NL9*	58%	1 (0)	2 (0)
5	pol	C*0804	---FPTINFCISL---	FL9	27%	1 (0)	2 (0)
5	pol	C*18	---DPTYKSSI---	DI8*	26%	0 (0)	1 (0)
6	pol	A*0205	---SLLVHVWLPL---	SL10	29%	14 (1)	20 (1)
6	pol	A*29	---NMHPPHPVPL---	NL9	69%	4 (3)	1(1)
6	pol	B*5802	---LPSPPFLHKL---	LL9	28%	4 (1)	5 (2)

Bansal, *J. Exp. Med.*, 2010

Berger, *J. Exp. Med.*, 2010

First evidence that innate arm of immune system drives HIV evolution



Points of attack by natural killer cells

Summary and next steps

- HIV is not invulnerable
- We can use machine learning and HPC to find HIV's Achilles' heel(s)
- Test the vaccine (with Jim Mullins)

PhyloD.net is part of Microsoft Biology Foundation

Microsoft Research

Search Microsoft Research

Home Our Research Collaboration Careers

About External Research | Tools and Services | Opportunities

[Home](#) > [Collaboration](#) > [Tools and Services](#) > The Microsoft Biology Foundation

The Microsoft Biology Foundation



The bioinformatics community has developed a strong tradition of open development, code sharing, and cross-platform support, and a number of language-specific bioinformatics toolkits are now available. These toolkits serve as valuable nucleation points for the community, promoting the sharing of code and establishing *de facto* standards.

The Microsoft Biology Foundation (MBF) is a language-neutral bioinformatics toolkit built as an extension to the Microsoft .NET Framework. Currently it implements a range of parsers for common bioinformatics file formats; a range of algorithms for manipulating DNA, RNA, and protein sequences; and a set of connectors to biological Web services such as NCBI BLAST. MBF is available under an open source license, and executables, source code, demo applications, and documentation are freely downloadable.

People involved



Jonathan Carlson
Nico Pfeifer
Jennifer Listgarten
Carl Kadie
Nebojsa Jojic
MSR-SCR Team



Philip Goulder
Philippa Matthews



Mary Carrington



Zabrina Brumme
Mark Brockman



P. Richard Harrigan
Chanson Brumme



Bruce Walker
Florencia Pereyra
Todd Allen
Marcus Altfeld



Thunbi Ndung'u



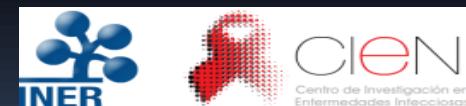
Richard Haubrich
Susan Riddler
ACTG 5142



Bette Korber
T. Bhattacharya



Paul Goepfert
Anju Bansal



Santiago Avila-Rios



Jim Mullins



Simon Mallal
Mina John



Toshiyuki Miura



Eric Hunter



Martin Markowitz