A phylogenetic Hidden Markov Model for Immune Epitope Discovery

Motivation

HIV-1

- Rapidly evolving human retrovirus
- Originated from SIVcpz in West Africa around 1900
- Pandemic, particularly significant in southern Africa
- Medical interventions
  - treatment
  - vaccination
- Viral evolution (associated global HIV diversity) important for HIV disease, treatment and control

Motivation

HIV-1

- Rapidly evolving human retrovirus
- Originated from SIVcpz in West Africa around 1900
- Pandemic, particularly significant in southern Africa
- Medical interventions
  - treatment
  - vaccination
- Viral evolution (associated global HIV diversity) important for HIV disease, treatment and control

Motivation

HIV-1

- Rapidly evolving human retrovirus
- Originated from SIVcpz in West Africa around 1900
- Pandemic, particularly significant in southern Africa
- Medical interventions
  - treatment
  - vaccination
- Viral evolution (associated global HIV diversity) important for HIV disease, treatment and control

Motivation

HIV-1

- Rapidly evolving human retrovirus
- Originated from SIVcpz in West Africa around 1900
- Pandemic, particularly significant in southern Africa
- Medical interventions
  - treatment
  - vaccination
- Viral evolution (associated global HIV diversity) important for HIV disease, treatment and control

Motivation

HIV-1

- Rapidly evolving human retrovirus
- Originated from SIVcpz in West Africa around 1900
- Pandemic, particularly significant in southern Africa
- Medical interventions
  - treatment
  - vaccination
- Viral evolution (associated global HIV diversity) important for HIV disease, treatment and control

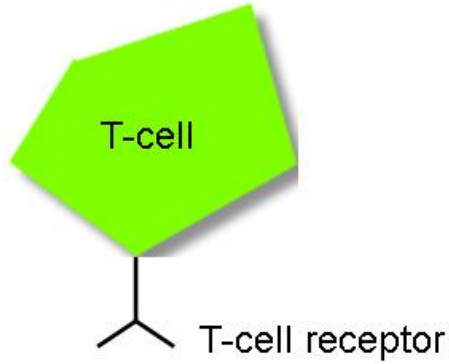
Adaptive immune response

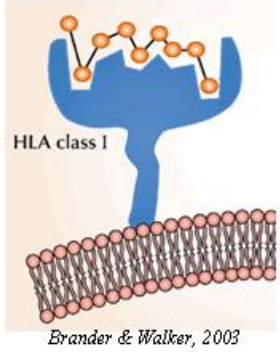
Cellular

(immune cells vs infected cells)

Humoral

(secreted antibodies vs free virus)





Brander & Walker, 2003

Cellular immune response to HIV-1

- Targets short (8 - 11 amino acid) linear peptides (epitopes)
- Key role in controling initial HIV-1 viremia
- A target of several vaccines
  - STEP/Phambili Trials
  - Recent Thai vaccine trial (2009)
- Largest contribution to variation in disease progression
- A major determinant of HIV evolution

Cellular immune response to HIV-1

- Targets short (8 - 11 amino acid) linear peptides (epitopes)
- Key role in controling initial HIV-1 viremia
- A target of several vaccines
  - STEP/Phambili Trials
  - Recent Thai vaccine trial (2009)
- Largest contribution to variation in disease progression
- A major determinant of HIV evolution

Cellular immune response to HIV-1

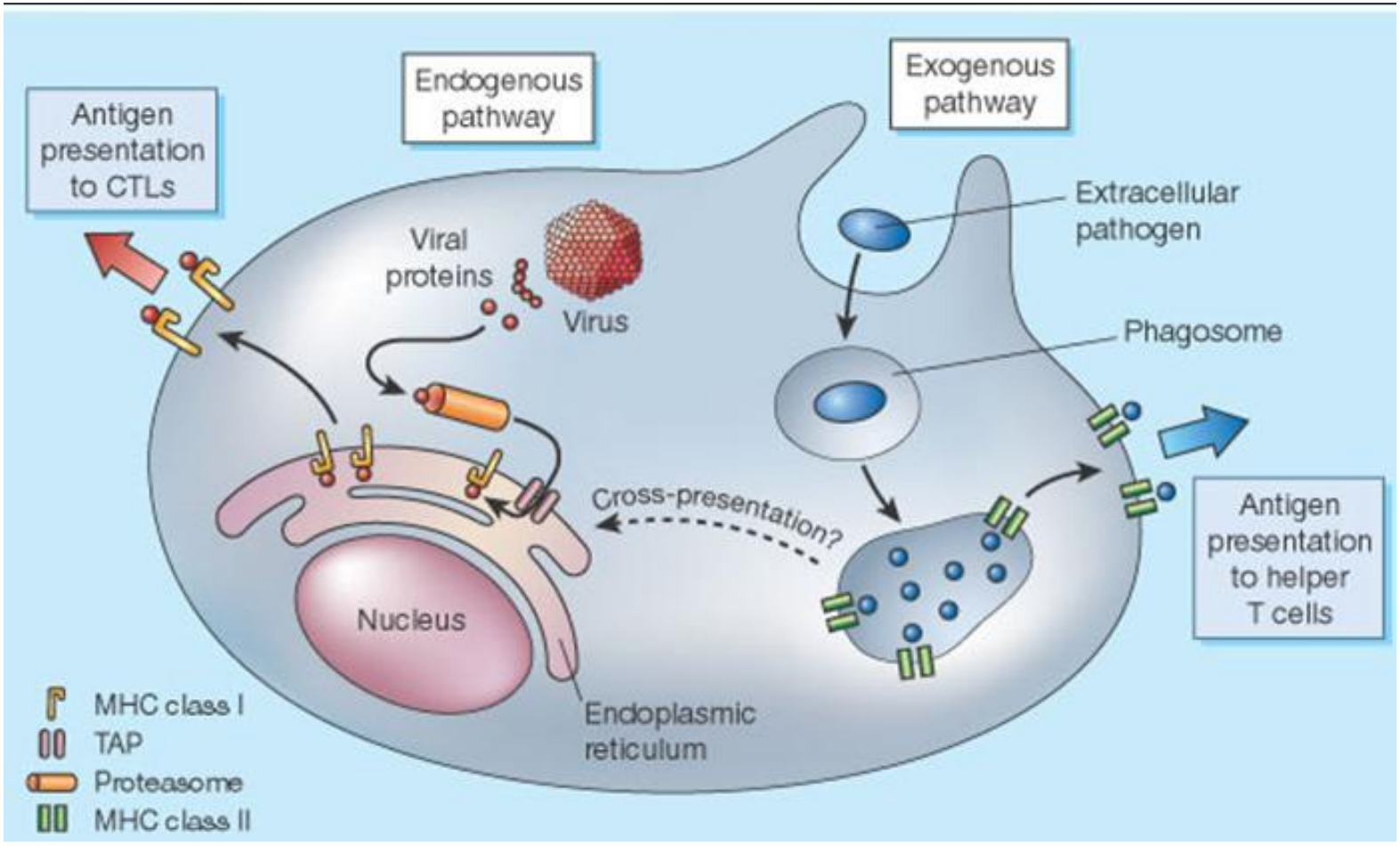
- Targets short (8 - 11 amino acid) linear peptides (epitopes)
- Key role in controling initial HIV-1 viremia
- A target of several vaccines
  - STEP/Phambili Trials
  - Recent Thai vaccine trial (2009)
- Largest contribution to variation in disease progression
- A major determinant of HIV evolution

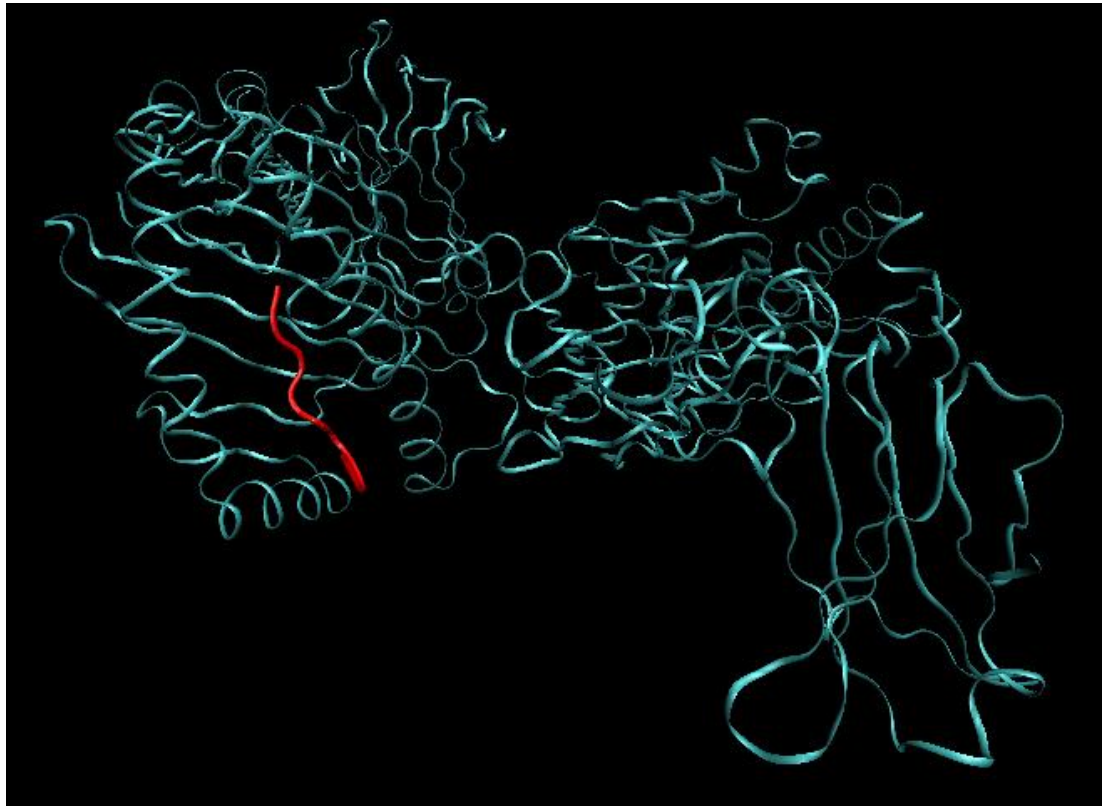
Cellular immune response to HIV-1

- Targets short (8 - 11 amino acid) linear peptides (epitopes)
- Key role in controling initial HIV-1 viremia
- A target of several vaccines
  - STEP/Phambili Trials
  - Recent Thai vaccine trial (2009)
- Largest contribution to variation in disease progression
- A major determinant of HIV evolution

Cellular immune response to HIV-1

- Targets short (8 - 11 amino acid) linear peptides (epitopes)
- Key role in controling initial HIV-1 viremia
- A target of several vaccines
  - STEP/Phambili Trials
  - Recent Thai vaccine trial (2009)
- Largest contribution to variation in disease progression
- A major determinant of HIV evolution


Antigen presentation to CTLs
Endogenous pathway
Exogenous pathway
Extracellular pathogen
Viral proteins
Virus
Phagosome
Cross-presentation?
Antigen presentation to helper T cells
Nucleus
Endoplasmic reticulum
MHC class I
TAP
Proteasome
MHC class II



Human Leukocyte Antigen (HLA) alleles

- Highly polymorphic

Human Leukocyte Antigen (HLA) alleles

- Highly polymorphic
- HLA alleles can bind and present different peptides

Human Leukocyte Antigen (HLA) alleles

- Highly polymorphic
- HLA alleles can bind and present different peptides
- Associated with distinct ‘anchor residue’ motifs

  e.g. A*2602

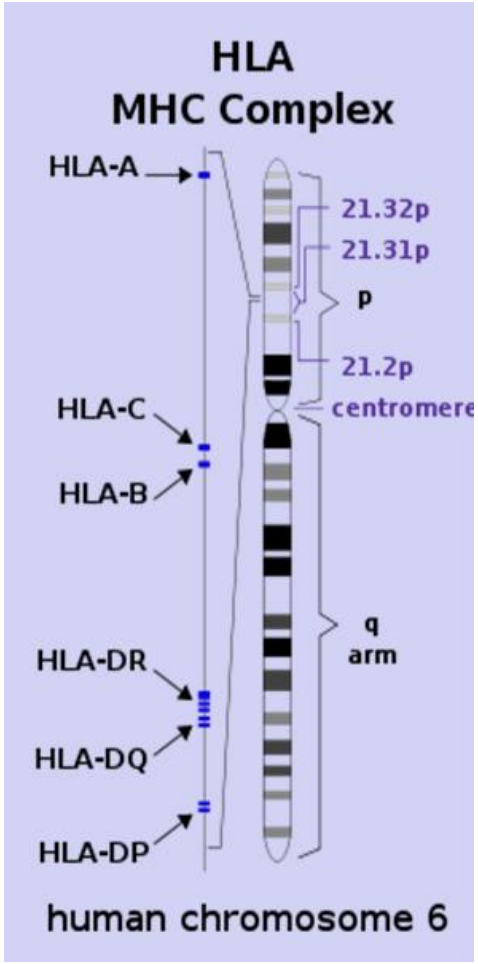
  x-[VTILF]-x-x-x-x-x-[YF(ML)]

Human Leukocyte Antigen (HLA) alleles

- Highly polymorphic
- HLA alleles can bind and present different peptides
- Associated with distinct ‘anchor residue’ motifs

  e.g. A*2602

  x-[VTILF]-x-x-x-x-x-x-[YF(ML)]



HIV escape from CTL responses

- Virus adapts to escape from host CTL responses
  - E.g. B57/B5801 and mutation T242N (in HIV-1 Gag)
- Escape can be a result of failure of
  - peptide processing
  - HLA binding to peptide
  - T cell recognition (temporary)
- Often incurs fitness cost
- Reverts to ‘wild-type’ upon transmission to new host
- Adaptation of the virus to human populations recently demonstrated (Kawashima, 2009)

HIV escape from CTL responses

- Virus adapts to escape from host CTL responses
  - E.g. B57/B5801 and mutation T242N (in HIV-1 Gag)
- Escape can be a result of failure of
  - peptide processing
  - HLA binding to peptide
  - T cell recognition (temporary)
- Often incurs fitness cost
- Reverts to ‘wild-type’ upon transmission to new host
- Adaptation of the virus to human populations recently demonstrated (Kawashima, 2009)

HIV escape from CTL responses

- Virus adapts to escape from host CTL responses
  - E.g. B57/B5801 and mutation T242N (in HIV-1 Gag)
- Escape can be a result of failure of
  - peptide processing
  - HLA binding to peptide
  - T cell recognition (temporary)
- Often incurs fitness cost
- Reverts to ‘wild-type’ upon transmission to new host
- Adaptation of the virus to human populations recently demonstrated (Kawashima, 2009)

HIV escape from CTL responses

- Virus adapts to escape from host CTL responses
  - E.g. B57/B5801 and mutation T242N (in HIV-1 Gag)
- Escape can be a result of failure of
  - peptide processing
  - HLA binding to peptide
  - T cell recognition (temporary)
- Often incurs fitness cost
- Reverts to ‘wild-type’ upon transmission to new host
- Adaptation of the virus to human populations recently demonstrated (Kawashima, 2009)

HIV escape from CTL responses

- Virus adapts to escape from host CTL responses
  - E.g. B57/B5801 and mutation T242N (in HIV-1 Gag)
- Escape can be a result of failure of
  - peptide processing
  - HLA binding to peptide
  - T cell recognition (temporary)
- Often incurs fitness cost
- Reverts to ‘wild-type’ upon transmission to new host
- Adaptation of the virus to human populations recently demonstrated (Kawashima, 2009)

Goal: Incorporate sequence evolution into epitope prediction

- Existing methods to predict CTL epitopes based on
  - amino acid sequence motifs
  - structures of HLA alleles
  - machine learning
- Evolutionary information not used

Goal: Incorporate sequence evolution into epitope prediction

- Existing methods to predict CTL epitopes based on
  - amino acid sequence motifs
  - structures of HLA alleles
  - machine learning
- Evolutionary information not used

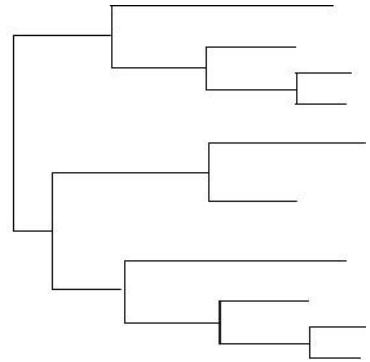
Goal: Incorporate sequence evolution into epitope prediction

- Existing methods to predict CTL epitopes based on
  - amino acid sequence motifs
  - structures of HLA alleles
  - machine learning
- Evolutionary information not used

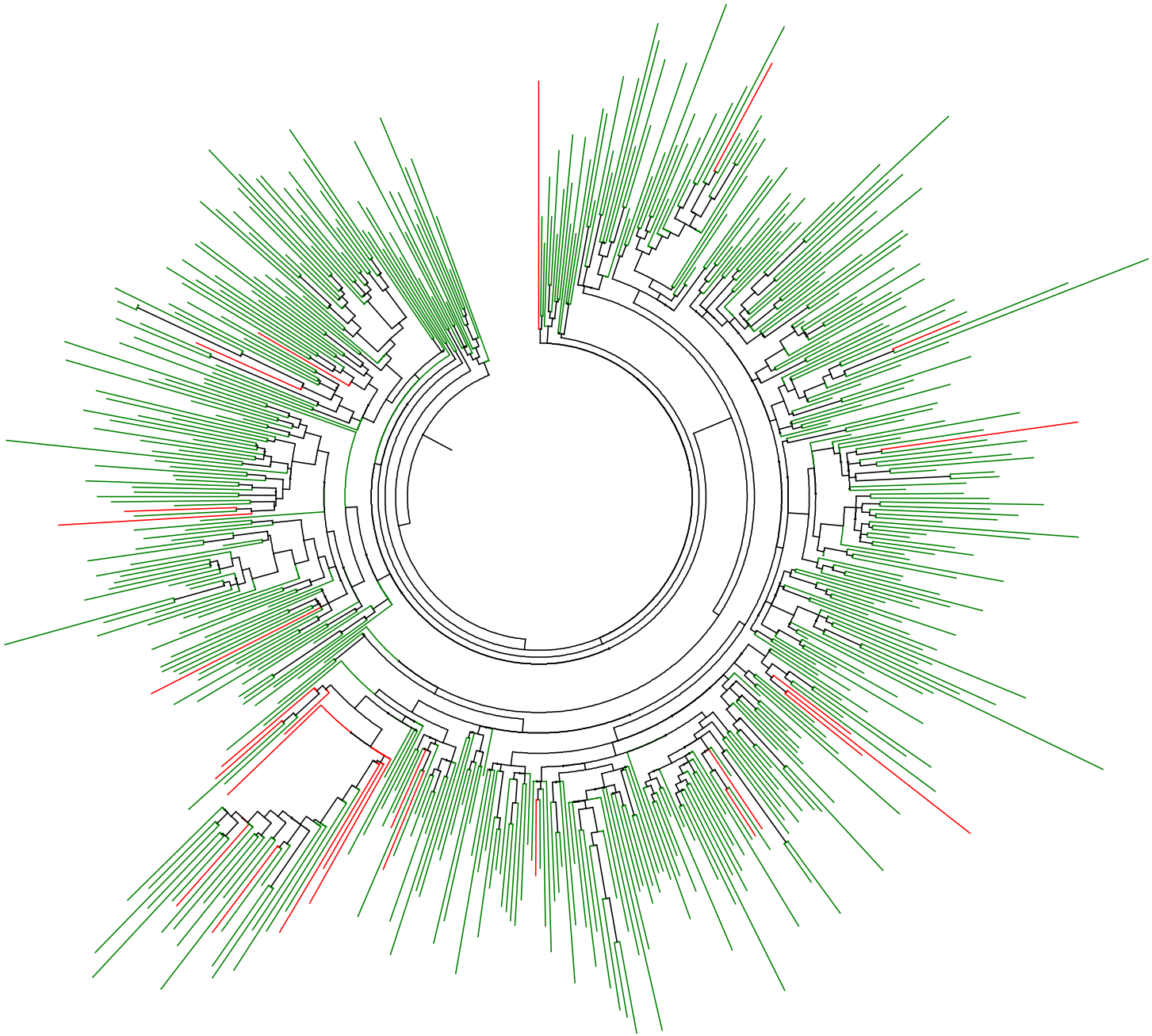
Data

- Gag protein-coding sequences from 506 patients in Durban (Rousseau et al. 2008)

```
'TTATGCTACAGGAGACATAATAGGAGATATAA
'TTATGCAACAGGAGAAATAATAGGAAATATAA
'TTAT---ACAGGAGAAATAGTAGGAGATATAA
'TTAT---ACAGGAGAAATAATAGGAGATATAA
'TTAT---ACAGGAGAAATAATAGGAGATATAA
'TTAT---ACAGGAGAAATAGTAGGAGATATAA
'TTATGCA---GGAGAAATAATAGGAGATATAA
'TTATGCA---GGAGAAATAATAGGAGATATAA
'TTAT---ACAGGAGAAATAATAGGAGATATAA
'TTAT---ACAGGGGAAATAATAGGAGATATAA
'TTATACAACAGGAGACATAATAGGAGATATAA
'TTATGCAACAGGAGAAATAATGGGAGATATAA
'TTATACAACAGGAGAAATAATAGGAGATATAA
'TTATGCAACAGGAAAAATAATAGGAGATATAA
```



- HLA genotypes for all patients



Q under Halpern & Bruno - like model:

$$g_{ij} = \frac{\mu_{ij} ln\left(\frac{\pi_j \mu_{ji}}{\pi_i \mu_{ij}}\right)}{1 - \frac{\pi_i \mu_{ij}}{\pi_j \mu_{ji}}}$$

This has a mutation component, μ_{ij} and a selection component (probability of fixation of the mutation)

Introduce a model of epitope evolution

- Distinct model for each human HLA allele
- Epitope model: rate of evolution from codon *I* to codon *J* may depend on presence/absence of the allele

Define wild-type amino acid at a site: the most fit amino acid in the absence of an immune response targeting the site.

Introduce parameters, γ_0 and γ_1, describing the equilibrium frequency (and hence fitness) of the wild-type when HLA allele absent, present

Introduce a model of epitope evolution

- Distinct model for each human HLA allele
- Epitope model: rate of evolution from codon *I* to codon *J* may depend on presence/absence of the allele

Define wild-type amino acid at a site: the most fit amino acid in the absence of an immune response targeting the site.

Introduce parameters, γ_0 and γ_1, describing the equilibrium frequency (and hence fitness) of the wild-type when HLA allele absent, present

Introduce a model of epitope evolution

- Distinct model for each human HLA allele
- Epitope model: rate of evolution from codon *I* to codon *J* may depend on presence/absence of the allele

Define wild-type amino acid at a site: the most fit amino acid in the absence of an immune response targeting the site.

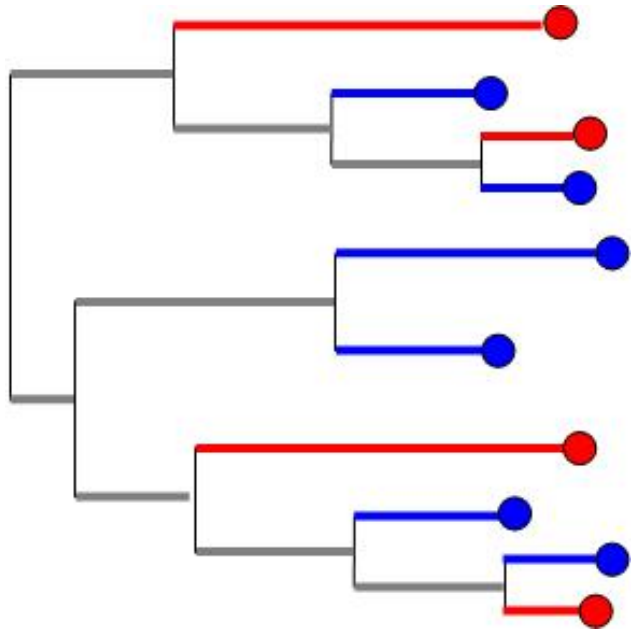
Introduce parameters, γ_0 and γ_1, describing the equilibrium frequency (and hence fitness) of the wild-type when HLA allele absent, present

Introduce a model of epitope evolution

- Distinct model for each human HLA allele
- Epitope model: rate of evolution from codon *I* to codon *J* may depend on presence/absence of the allele

Define wild-type amino acid at a site: the most fit amino acid in the absence of an immune response targeting the site.

Introduce parameters, γ_0 and γ_1, describing the equilibrium frequency (and hence fitness) of the wild-type when HLA allele absent, present



At internal branches

$$\gamma = p_0\gamma_0 + p_1\gamma_1$$

where p_1, empirical frequency of HLA allele in population ($p_0 = 1 - p_1$)

Probability under epitope model

$$p_E(\mathbf{X}_\ell|\psi') = \int_{\gamma_1}\int_{\gamma_0} p(\mathbf{X}_\ell|\psi', \gamma_1, \gamma_0)p(\gamma_0|\gamma_1, \xi_0)p(\gamma_1|\xi)\ d\gamma_0 d\gamma_1,$$

$p(\gamma_0|\gamma, \xi_0)$ and $p(\gamma_1|\xi)$, prior distributions with hyperparameters ξ_0,ξ_1

ψ', all parameters of the model, other than γ 's

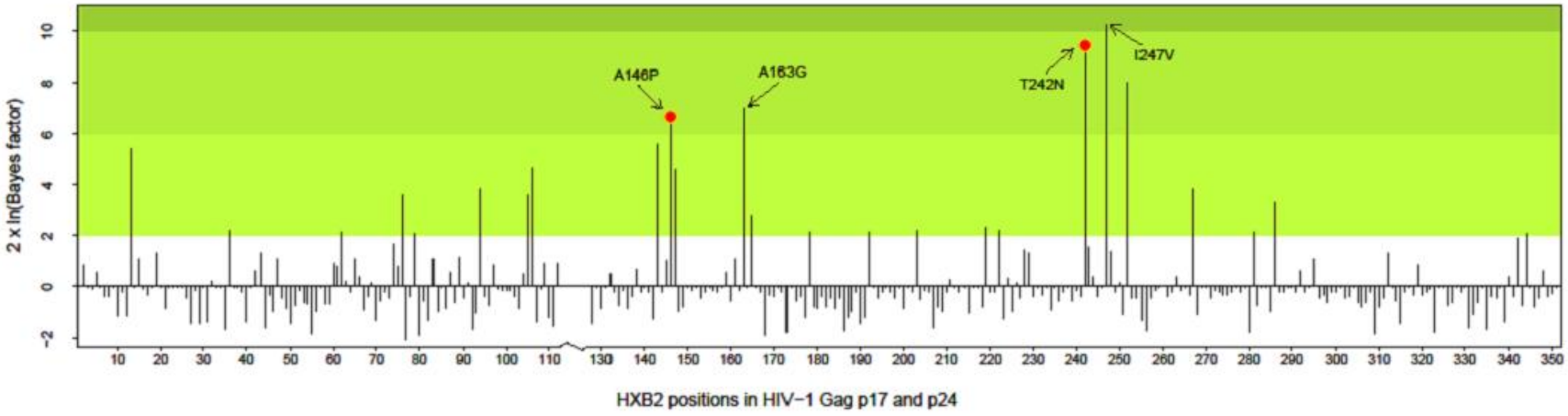
ξ_0, ξ_1 and transition probabilities estimated using Baum-Welch algorithm

Probability under non-epitope model

$\gamma_0 = \gamma_1 \Rightarrow$

$$p_N(\mathbf{X}_\ell|\psi') = \int_\gamma p(\mathbf{X}_\ell|\psi', \gamma)p(\gamma|\xi)d\gamma.$$

Bayes Factor along Gag



Phylogenetic Hidden Markov Models (PhyloHMMs)

- Hidden states emit columns of the sequence alignment
- Emission probabilities: probability of the alignment column
- To date a small number of practical applications
- Generally computationally intensive

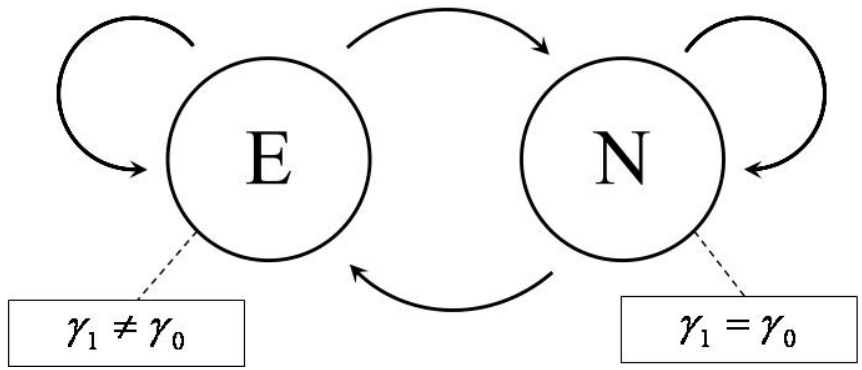
PhyloHMM for epitope discovery

- Hidden Markov model with
  - Epitope or not-epitope as hidden states
  - Emission probabilities expressed as $P_E(X)$ and $P_N(X)$
- Standard algorithms can be used to
  - Find most probable path through the hidden states (Viterbi)
  - Work out the posterior probability of each site being within an epitope (Backward-Forward algorithm)

PhyloHMM for epitope discovery

- Hidden Markov model with
  - Epitope or not-epitope as hidden states
  - Emission probabilities expressed as $P_E(X)$ and $P_N(X)$
- Standard algorithms can be used to
  - Find most probable path through the hidden states (Viterbi)
  - Work out the posterior probability of each site being within an epitope (Backward-Forward algorithm)

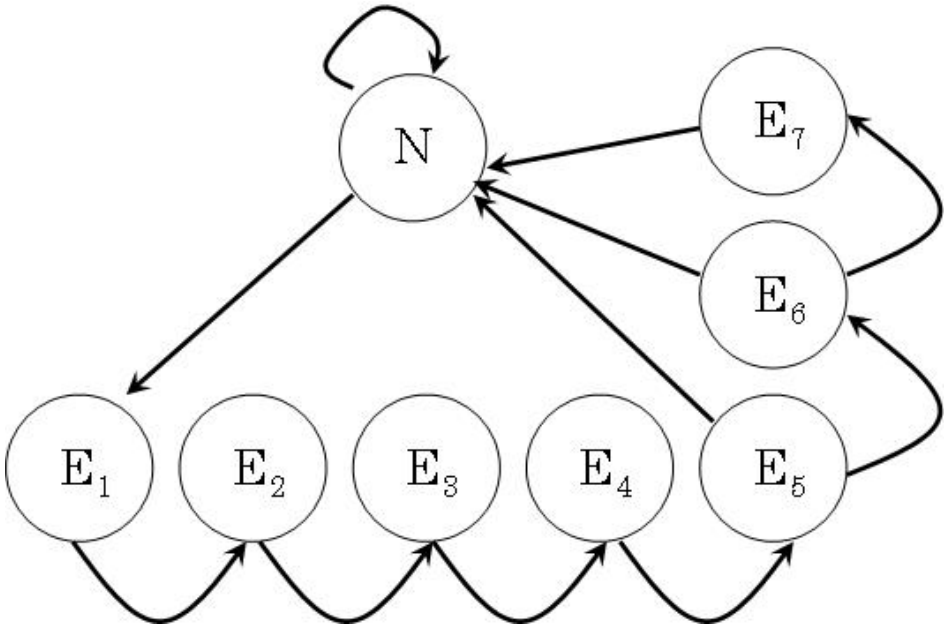
Simple two-state epitope model



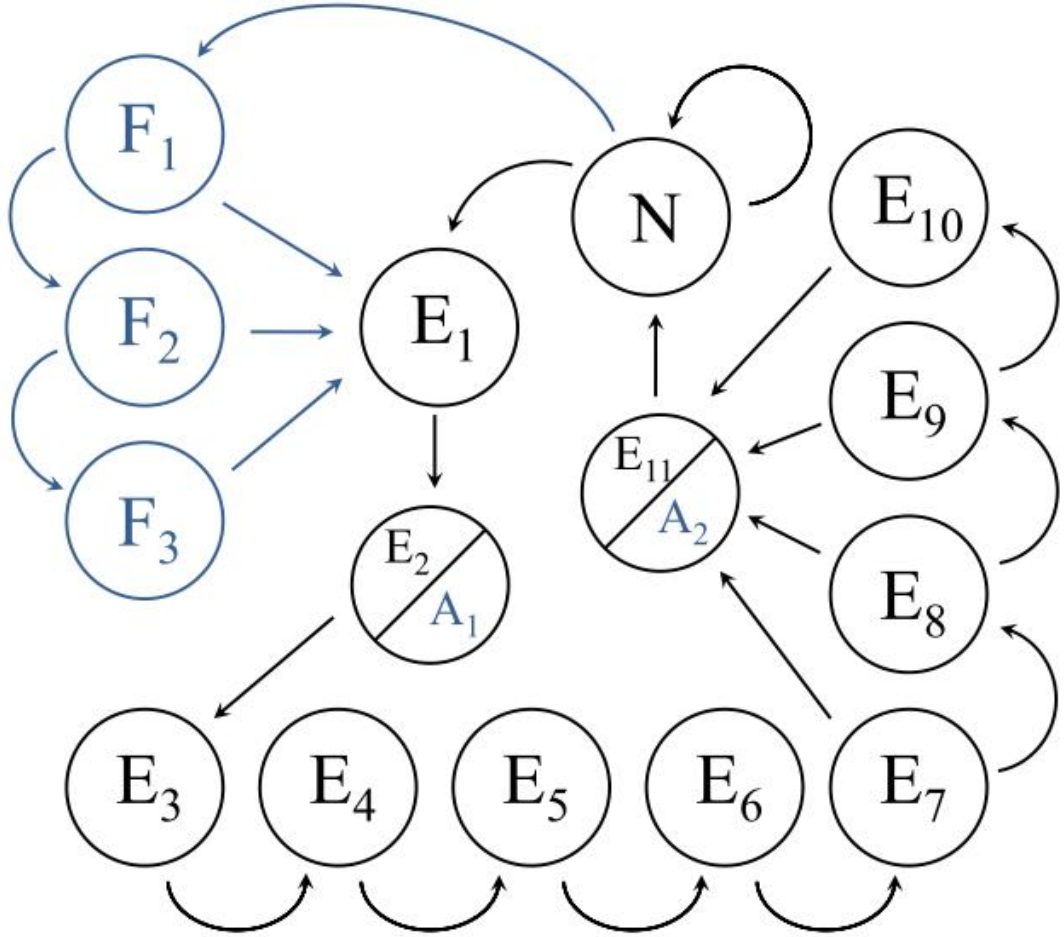
‘Structured’ epitope model

- Take account of expected epitope length and structure
- Can be used to incorporate sequence motifs required for HLA alleles to bind

Structured epitope models



Structured epitope models



Incorporating anchor residue motifs into the model

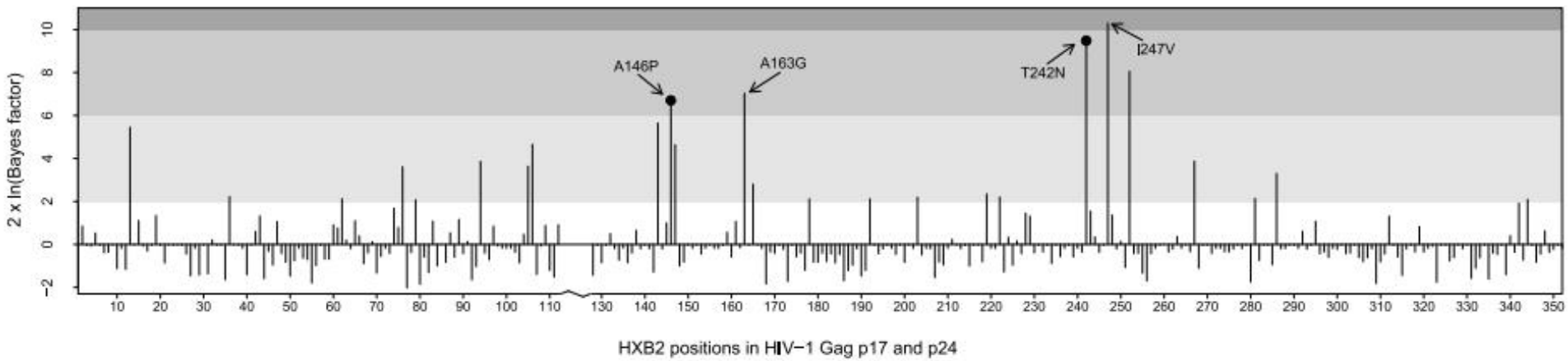
Indicator variable $\mathbb{1}_{\ell i}$

$\mathbb{1}_{\ell i} = 1$ if anchor residue present at site ℓ in some predefined proportion of sequences ($i \in \{1, 2\}$)

$$p_{A_i}(\mathbf{X}_\ell|\psi') = \begin{cases} p_E(\mathbf{X}_\ell|\psi')/p_E(\mathbb{1}_{\ell i} = 1|\psi') & \text{if } \mathbb{1}_{\ell i} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

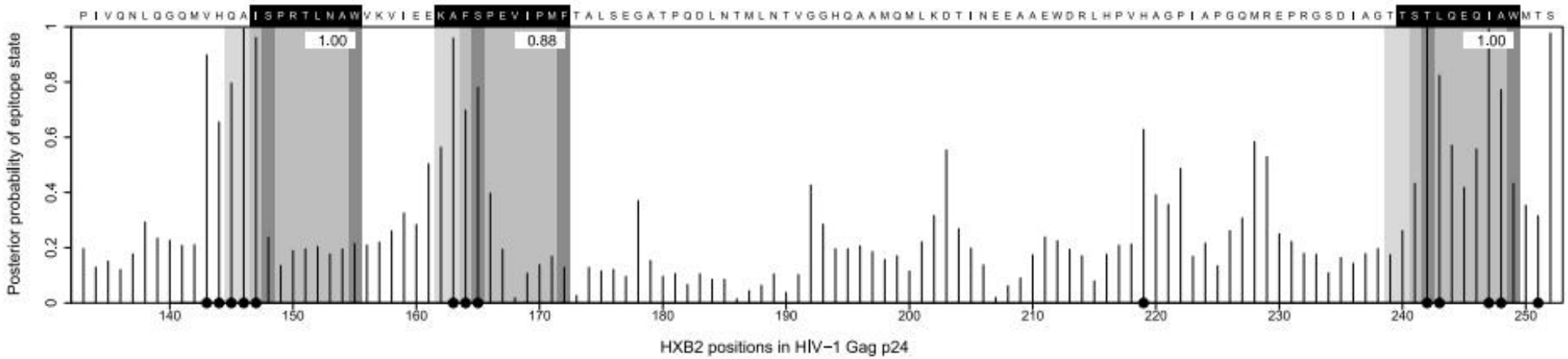
Results of application to real data

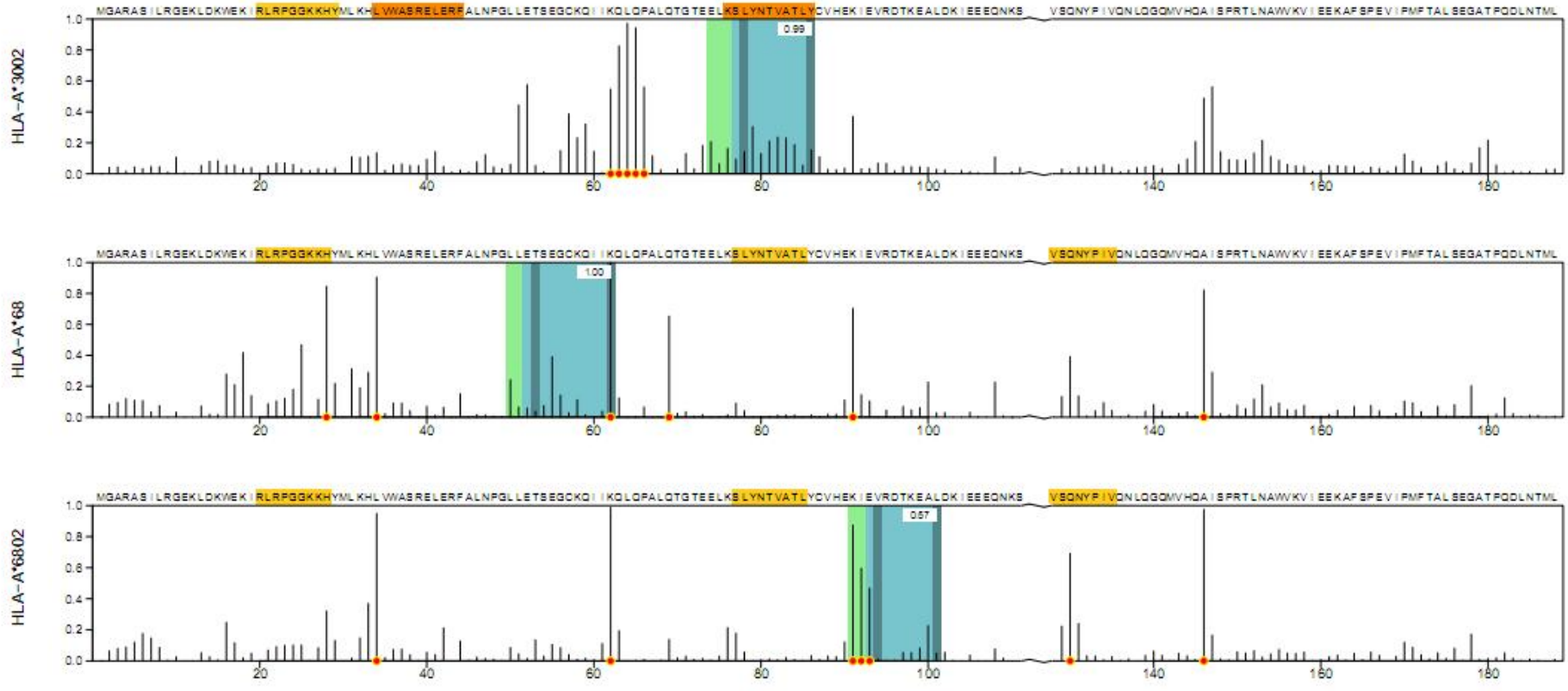
Comparison of fit of epitope and non-epitope models at sites across Gag

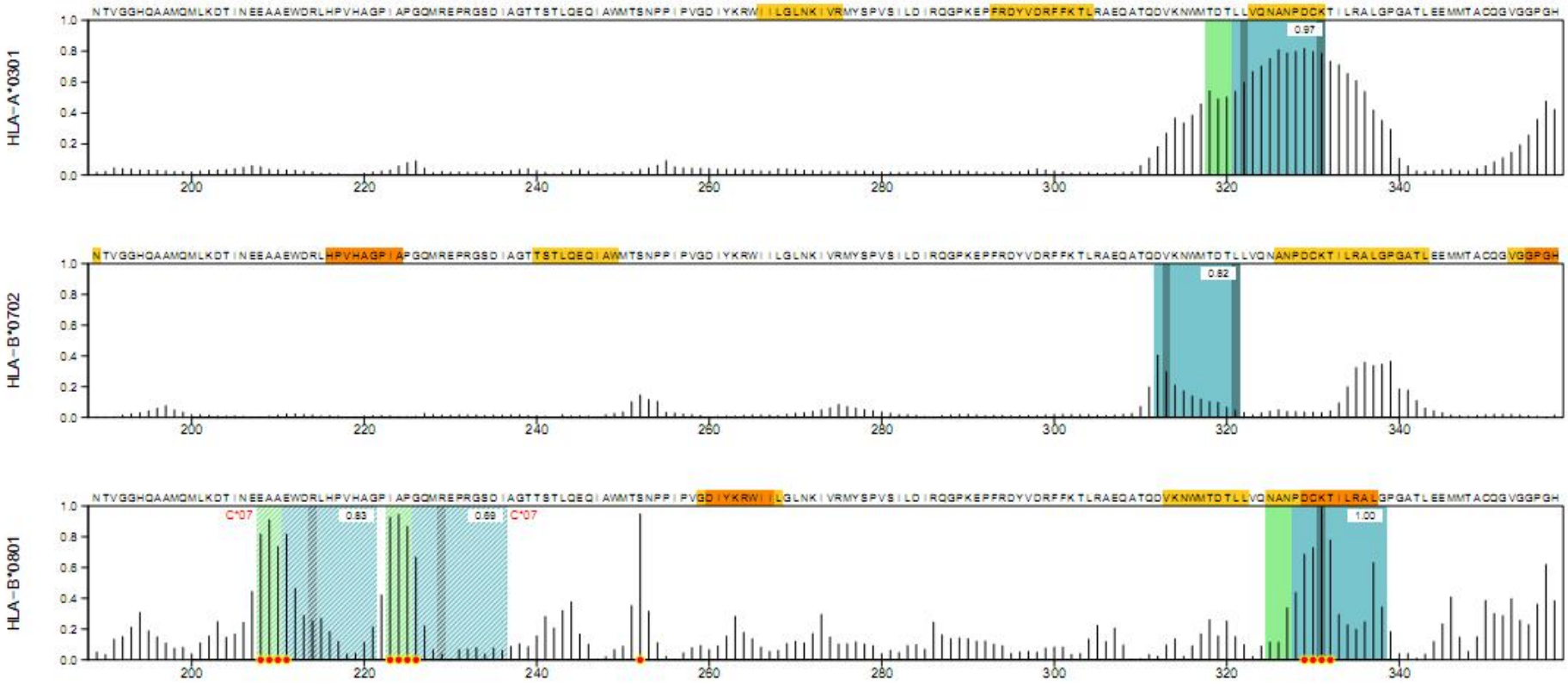


Results of application to real data

Posterior probabilities of epitope model




HLA-A*3002
HLA-A*68
HLA-A*6802
0.99
1.00
0.67


HLA-A*0301
HLA-B*0702
HLA-B*0801
0.97
0.82
C*07
0.83
0.69
C*07
1.00

	Anchors Only		Phylo-HMMwD		Phylo-HMMwD+M	
HLA allele	Known	Total	Known	Total	Known	Total
A*02	11	12	0	0	2	2
A*23	1	6	1	4	1	4
A*29	1	10	0	3	1	3
A*30	3	12	2	6	2	5
A*3001	3	12	1	3	1	3
A*3002	2	4	1	5	1	3
A*68	3	12	1	8	0	4
A*6802	3	10	1	8	0	3
A*0301	5	11	2	2	2	3
B*0702	5	12	0	0	1	3
B*0801	6	11	3	5	2	4
B*15	5	12	1	5	1	3
B*1503	2	3	0	1	0	1
B*1510	2	7	1	6	2	3
B*42	3	5	1	2	3	4
B*4201	3	5	1	2	3	4
B*44	3	8	2	3	1	2
B*4403	3	8	0	1	1	2
B*57	7	7	3	6	3	3
B*5801	6	7	1	1	3	3
B*5802	1	1	2	7	1	1
B*81	1	5	1	4	1	1
C*02	0	9	0	0	0	1
C*0202	0	9	0	0	0	2
C*04	1	13	2	10	1	5
C*06	1	3	0	5	0	6
C*0602	1	3	0	4	0	5
C*07	2	15	0	7	1	10
C*08	3	12	1	6	0	2
C*17	0	9	0	2	0	0
C*1701	0	9	0	4	0	0
C*18	2	13	1	5	0	6
Total[a]	63	191	21	81	24	65

Conclusions

- Mutation-selection model of immune escape
- An evolutionary model that includes environmental variables
- Provides a probabilistic framework for epitope prediction, incorporating evolutionary information

Issues outstanding and further work

- Not all T cell responses have the same effect on sequence evolution
- Linkage disequilibrium between HLA alleles
  - to-date performed post-hoc analysis
- Compensatory mutations
- Integration with other epitope-prediction methods

Issues outstanding and further work

- Not all T cell responses have the same effect on sequence evolution
- Linkage disequilibrium between HLA alleles
  - to-date performed post-hoc analysis
- Compensatory mutations
- Integration with other epitope-prediction methods

Issues outstanding and further work

- Not all T cell responses have the same effect on sequence evolution
- Linkage disequilibrium between HLA alleles
  - to-date performed post-hoc analysis
- Compensatory mutations
- Integration with other epitope-prediction methods

Issues outstanding and further work

- Not all T cell responses have the same effect on sequence evolution
- Linkage disequilibrium between HLA alleles
  - to-date performed post-hoc analysis
- Compensatory mutations
- Integration with other epitope-prediction methods

Acknowledgements

- Miguel Lacerda
- Konrad Scheffler
- Carolyn Williamson (UCT)
- Sergei Kosakovsky-Pond (UCSD)


TITLE
TITLE


TITLE
TITLE


$$\nu_J = \begin{cases} \frac{\gamma \pi_J}{\sum_{s \in W} \pi_s} & \text{if } J \in W \\ \frac{(1-\gamma)\pi_J}{1-\sum_{s \in W} \pi_s} & \text{if } J \notin W \end{cases}$$