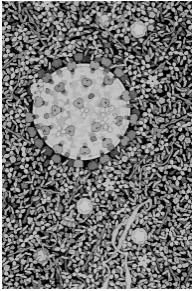
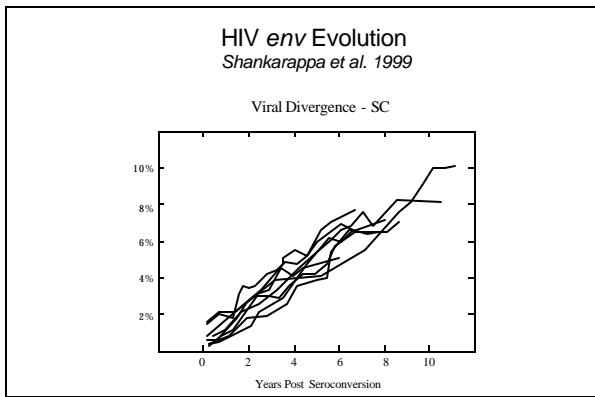
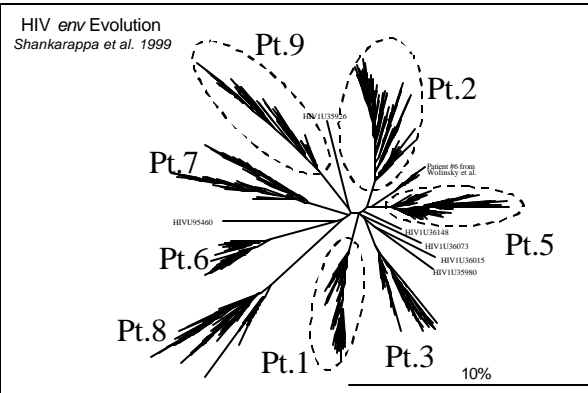


Watercolour by David Goodsell
Scripps Institute



Measurably Evolving Populations

Allen Rodrigo
Bioinformatics Institute, University of Auckland
and
Allan Wilson Centre for Molecular Ecology and Evolution



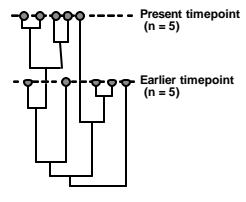
HIV genetic variation

- **Rapid and measurable viral evolution**
 - The HIV envelope gene accumulates substitutions at a rate of approx. 1% per year.
 - Compare this with:
 - ☞ the average prokaryotic or eukaryotic gene - 1% per 4 million years
 - ☞ ribosomal RNA -- 1% per 50 million years

Measurably evolving populations (MEPs)

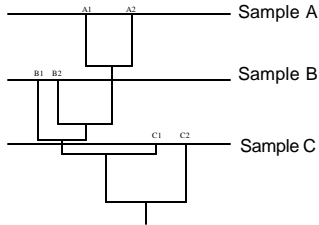
MEP: Any population evolving fast enough so that statistically significant differences between serially sampled sequences can be detected.

- ◆ Rapidly evolving pathogens, e.g., HIV, FIV, Influenza.
- ◆ Ancient DNA: so far mostly mitochondrial, e.g.
 - Adelle penguins
 - Pleistocene bears



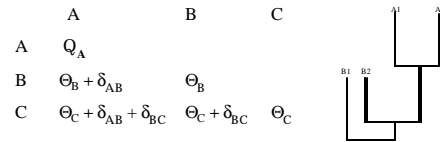
Reconstructing serial genealogies

serial sample Unweighted Pair Group Method of Arithmetic Means (sUPGMA)
(Drummond & Rodrigo, 2000)



A general model

A general model decomposes pairwise distances into within-timepoint diversity (θ) and between time-point divergence (δ)



In this example 5 parameters need to be estimated:
 $\Theta_A, \Theta_B, \Theta_C, \delta_{AB}, \delta_{BC}$

Least squares estimation

Observed	Q_A	Q_B	Q_C	d_{AB}	d_{BC}
d_{A1A2}	1	0	0	0	0
d_{B1B2}	0	1	0	0	0
d_{C1C2}	0	0	1	0	0
d_{A1B1}	0	1	0	1	0
d_{A1B2}	0	1	0	1	0
d_{A2B1}	0	1	0	1	0
d_{A2B2}	0	1	0	1	0
d_{A1C1}	0	0	1	1	1
d_{A1C2}	0	0	1	1	1
d_{A2C1}	0	0	1	1	1
d_{A2C2}	0	0	1	1	1
d_{B1C1}	0	0	1	0	1
d_{B1C2}	0	0	1	0	1
d_{B2C1}	0	0	1	0	1
d_{B2C2}	0	0	1	0	1

$P = (X'X)^{-1}X'd$

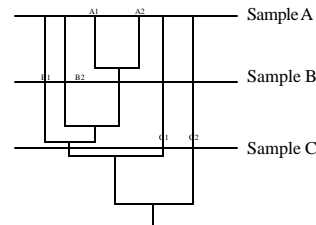
where

$P = (\Theta_A, \Theta_B, \Theta_C, \delta_{AB}, \delta_{BC})$

and d is a vector of pairwise distances.

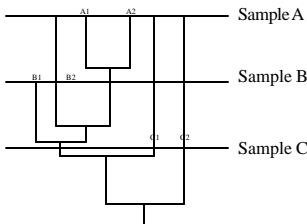
$E[d_{AC}] = \Theta_C + d_{AB} + d_{BC}$

Adjusting branches for tree-building



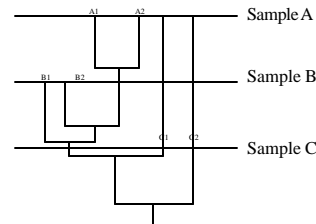
- Observed distance matrix can be adjusted by adding estimated δ to appropriate distances.
- Construct tree using UPGMA on corrected distances.
- Trim off branches by the amounts added.

Adjusting branches for tree-building



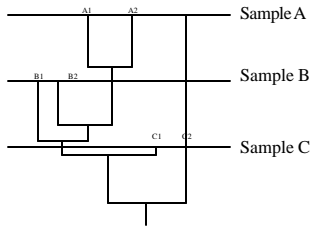
- Observed distance matrix can be adjusted by adding estimated δ to appropriate distances.
- Construct tree using UPGMA on corrected distances.
- Trim off branches by the amounts added.

Adjusting branches for tree-building



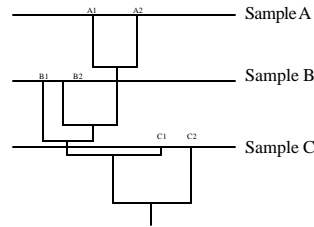
- Observed distance matrix can be adjusted by adding estimated δ to appropriate distances.
- Construct tree using UPGMA on corrected distances.
- Trim off branches by the amounts added.

Adjusting branches for tree-building

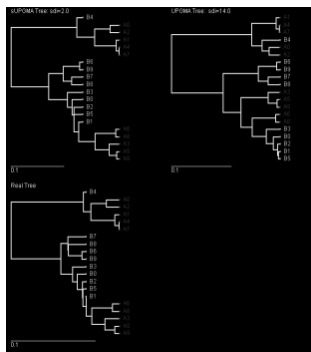


- ❑ Observed distance matrix can be adjusted by adding estimated δ to appropriate distances.
- ❑ Construct tree using UPGMA on corrected distances.
- ❑ Trim off branches by the amounts added.

Adjusting branches for tree-building



- ❑ Observed distance matrix can be adjusted by adding estimated δ to appropriate distances.
- ❑ Construct tree using UPGMA on corrected distances.
- ❑ Trim off branches by the amounts added.

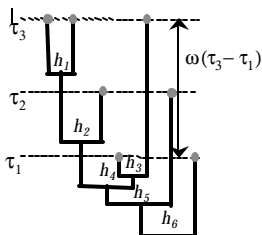


Reconstructing serial genealogies using sUPGMA

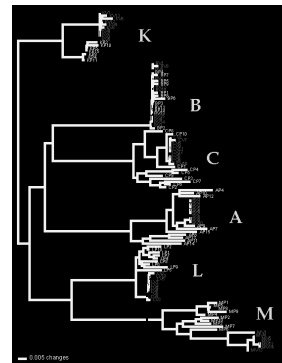
Likelihood with serial-sample trees

Estimating evolutionary rates

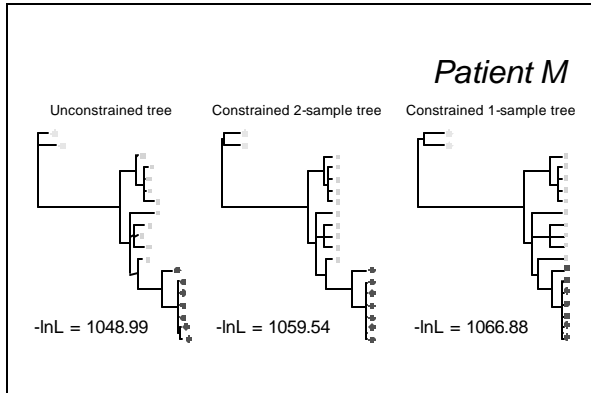
Single Rate with Date Tips (SRDT)



- ◆ Estimates uniform rate (ω) over entire sampling period.
- ◆ Strict molecular clock.
- ◆ Use ML to optimize branch lengths, estimate parameters \mathbf{h}, ω .
- ◆ Maximise $L(\mathbf{h}, \omega) = P(D | T, \mathbf{h}, \omega)$;

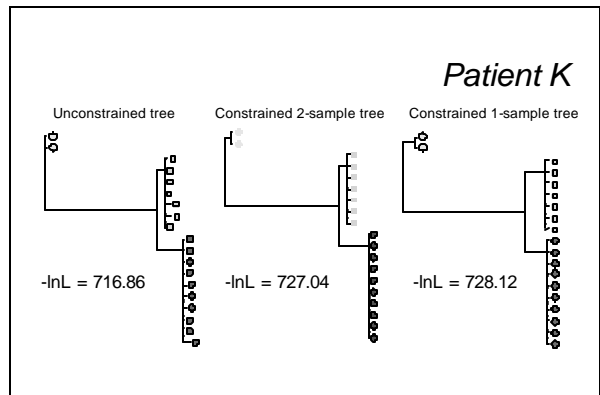
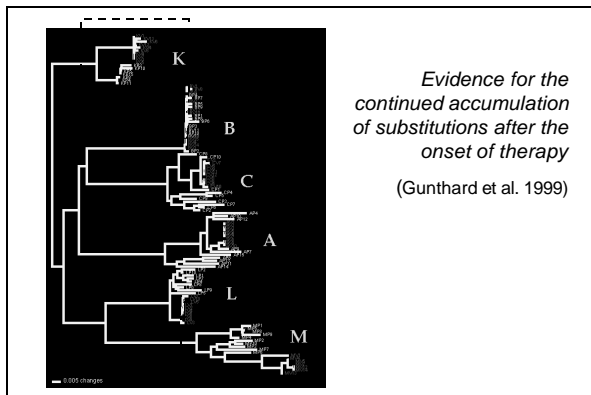


Evidence for the continued accumulation of substitutions after the onset of therapy (Gunthard et al. 1999)



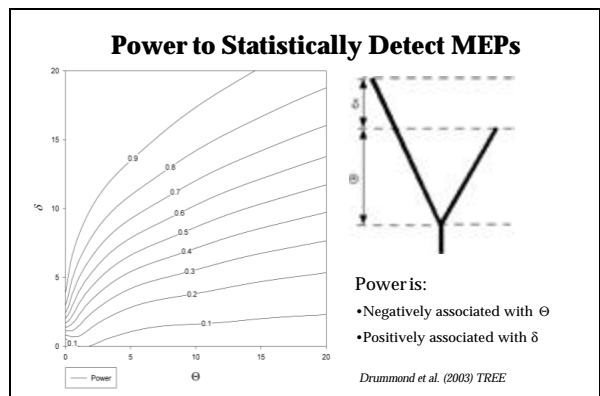
Patient M: hypothesis tests

Comparison	c^2 $2(\ln L_1 - \ln L_2)$	df	<i>p-value</i>
unconstrained vs. 2-sample trees	20.36	15	0.159
2-sample vs. 1-sample tree	7.14	1	0.007



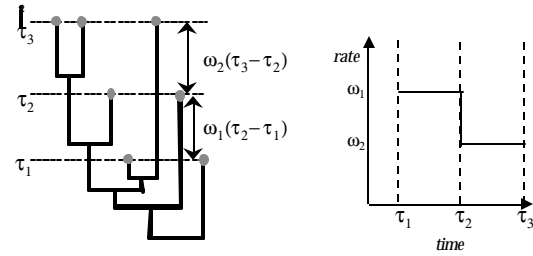
Patient K: hypothesis tests

Comparison	c^2 $2(\ln L_1 - \ln L_2)$	df	<i>p-value</i>
unconstrained vs. 2-sample trees	21.10	15	0.134
2-sample vs. 1-sample tree	2.16	1	0.142



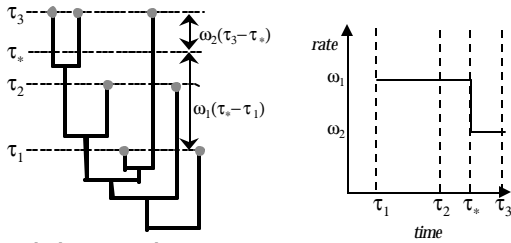
Estimating changes in the rate of evolution

Multiple Rates with Dated Tips (MRDT)



Multiple rates: sampling times coincident with substitution rates.

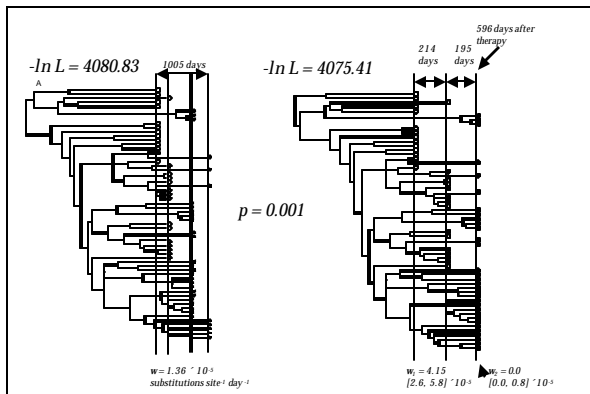
Multiple Rates with Dated Tips (MRDT)



Multiple rates: sampling times **not** coincident with rate changes.

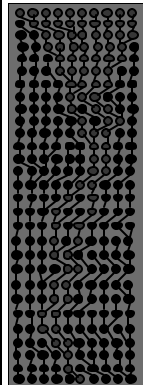
An HIV Example

- ♦ serially sampled partial HIV-1 envelope (*env*) sequences (60 sequences of 660 bases)
- ♦ sampled at days 0, 214, 671, 699 and 1005.
- ♦ Monotherapy with zidovudine was initiated after day 409
- ♦ Therefore the data set contains two samples before and three samples after treatment began.



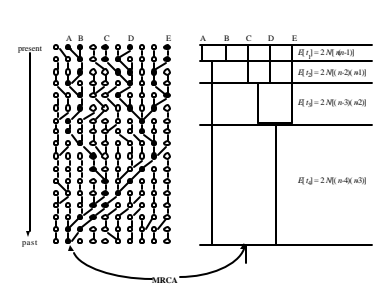
Population Genetics and Gene Genealogies

Wright-Fisher Population Model



- ▶ Constant population size
- ▶ Equal propensity to reproduce
- ▶ Discrete generations

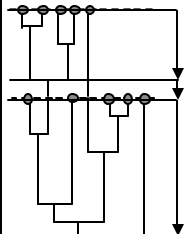
After a finite interval, all individuals will be descended from a common ancestor (MRCA).



Kingman (1982) showed that, for a population of size N , the distribution of each coalescent interval is exponential with parameter $2N/(i-1)$.

Expected time to the MRCA is $2N(n-1)/n$.

The s -coalescent in a likelihood framework (Rodrigo and Felsenstein, 1998)



◆ The likelihood function:

$$P(D|Nt, w) = \sum_G P(D|G, w) P(G|Nt)$$

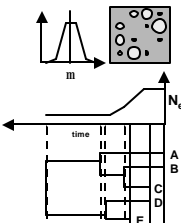
where $w = m/t$

$$P(G|Nt) = \prod_{p=n_1}^{n_2} \frac{2}{2Nt} \exp\left(-\frac{p(p-1)}{2Nt} n_p\right) \times \exp\left(-\frac{(n_2-c)(n_2-c-1)}{2Nt} (v_2 - n_1)\right) \times \prod_{q=2}^{n_1+n_2-c} \frac{2}{2Nt} \exp\left(-\frac{q(q-1)}{2Nt} n_q\right)$$

Bayesian Inference of Evolutionary Parameters

Methodological Challenges for MEPs

- ◆ GOALS: Given serially sampled sequences, D , can we estimate values of:
 - mutation parameters (m and Q)
 - population history (N_e , growth rate)
 - Serial phylogenies, T



Bayesian Inference of Evolutionary Parameters

◆ In a Bayesian framework, uncertainty can be incorporated by specifying prior probabilities of the parameters of interest.

$$P(T, m, N_e, Q | D) = \frac{1}{Z} P(D | T, m, Q) f_G(T | N_e) f_m(m) f_{N_e}(N_e) f_Q(Q)$$

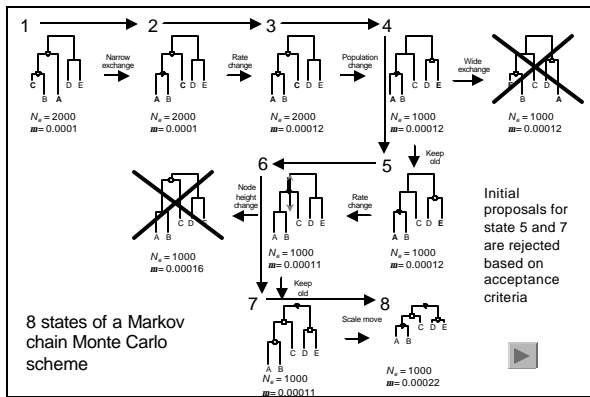
Labels in the equation: Posterior probability, Likelihood function, other priors, Unknown normalizing constant, coalescent prior.

Markov chain Monte Carlo (MCMC)

- ◆ It is not possible to evaluate the posterior distribution $P(T, m, N_e, Q | D)$ analytically.
- ◆ In a Bayesian analysis, MCMC integration can be used to draw correlated samples from a proposal distribution of $\mathbf{X} = \{T, m, N_e, Q\}$ to recover the target distribution P .
- ◆ This is done by accepting or rejecting \mathbf{X} on the basis of the ratio of $P(\mathbf{X}_{i+1} | D)$ to $P(\mathbf{X}_i | D)$, and a Hastings ratio to correct for the density of the proposal distribution.

Markov chain Monte Carlo (MCMC)

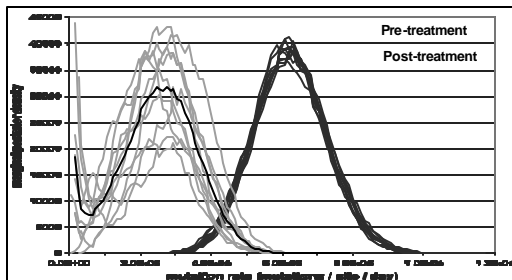
- ◆ The proposal distribution is typically generated by making small (and random changes) to T , m , N_e , or Q or their component parts.
- ◆ The sampled values of \mathbf{X} can be used to build up marginal posterior distributions of T , m , N_e and Q .



An HIV-1 env example

- HIV-1 envelope (*env*) sequences (60 sequences of 660 bases) from infected patient.
- sampled at days 0, 214, 671, 699 and 1005.
- Monotherapy with zidovudine was initiated after day 409.
- Split into **pre-treatment** (n=28) and **post-treatment** (n=32) data sets and analyse separately using MCMC.

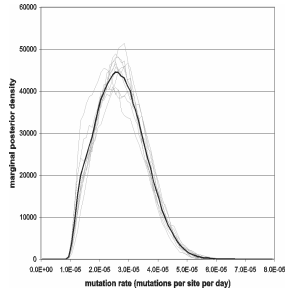
Summary of ten independent MCMC analyses: mutation rate



Can we 'fix' post-treatment? t_{MRCA} prior

- ◆ Sometimes there is prior knowledge about the age of the root (t_{MRCA}).
- ◆ It is widely believed that in HIV-1 infection, the virus population is the outgrowth of a single virus particle from early infection.
- ◆ Therefore, t_{MRCA} will be less than the age of the infection.
- ◆ We suggest an upper limit (t^*) on t_{MRCA} representing this belief. For this example we chose t^* to be 3650 days (10 years).

Post-treatment with informative prior



- Ten independent analyses.
- 'Spurious' mode at zero is removed.
- Variation between runs is reduced.
- Prior allows us reduce earlier ambiguity.

Drummond et al. 2002 Genetics

