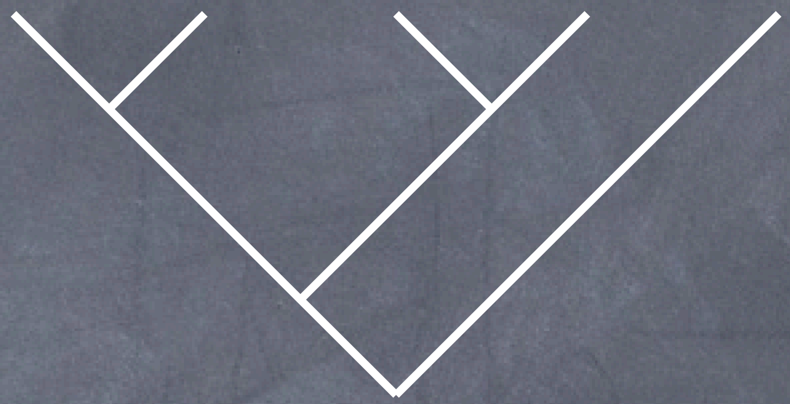
Parts of a Phylogenetic Model

1) Tree (with branch lengths)

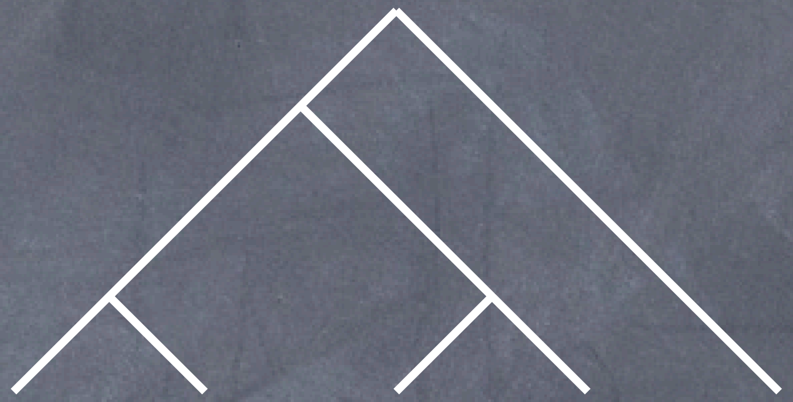


2) Model of character change (continuous-time Markov chain)

$$\begin{pmatrix} - & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & - & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & - & r_{GT}\pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & - \end{pmatrix}$$

Parts of a Phylogenetic Model

1) Tree (with branch lengths)



2) Model of character change (continuous-time Markov chain)

$$\begin{pmatrix} - & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & - & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & - & r_{GT}\pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & - \end{pmatrix}$$

The approach I will describe:

- Expand the pool of candidate models, describing models as partitions
- Calculate the joint posterior probability of the models and the other tree parameters
- Use Bayes factors to evaluate the support for alternative phylogenetic models

For several important problems in phylogenetics, models can be described as partitions.

A partition of a set of distinct objects, S, is a set of disjoint subsets of S whose union is S.

Substitution Models as Partitions

$$\begin{pmatrix} - & r_{AC}\,\pi_C & r_{AG}\,\pi_G & r_{AT}\,\pi_T \\ r_{AC}\,\pi_A & - & r_{CG}\,\pi_G & r_{CT}\,\pi_T \\ r_{AG}\,\pi_A & r_{CG}\,\pi_C & - & r_{GT}\,\pi_T \\ r_{AT}\,\pi_A & r_{CT}\,\pi_C & r_{GT}\,\pi_G & - \end{pmatrix}$$

General Time Reversible Model
(GTR, Tavaré, 1986)

Substitution Models as Partitions

$$\begin{array}{cc|cccc} & & & \text{To} & & \\ & & A & C & G & T \\ \hline & A & - & r_{AC}\,\pi_C & r_{AG}\,\pi_G & r_{AT}\,\pi_T \\ & C & r_{AC}\,\pi_A & - & r_{CG}\,\pi_G & r_{CT}\,\pi_T \\ \text{From} & G & r_{AG}\,\pi_A & r_{CG}\,\pi_C & - & r_{GT}\,\pi_T \\ & T & r_{AT}\,\pi_A & r_{CT}\,\pi_C & r_{GT}\,\pi_G & - \end{array}$$

General Time Reversible Model
(GTR, Tavaré, 1986)

AC	AG	AT	CG	CT	GT
1	1	1	1	1	1

AC	AG	AT	CG	CT	GT
1	1	1	1	1	1

$$\begin{pmatrix} - & 1 & 1 & 1 \\ 1 & - & 1 & 1 \\ 1 & 1 & - & 1 \\ 1 & 1 & 1 & - \end{pmatrix}$$

Jukes & Cantor (1969)

AC	AG	AT	CG	CT	GT
1	2	1	1	2	1

AC	AG	AT	CG	CT	GT
1	2	1	1	2	1

$$\begin{pmatrix} - & 1 & \kappa & 1 \\ 1 & - & 1 & \kappa \\ \kappa & 1 & - & 1 \\ 1 & \kappa & 1 & - \end{pmatrix}$$

Kimura (1980)

AC	AG	AT	CG	CT	GT
1	2	1	1	3	1

AC	AG	AT	CG	CT	GT
1	2	1	1	3	1

$$\begin{pmatrix} - & 1 & \alpha & 1 \\ 1 & - & 1 & \beta \\ \alpha & 1 & - & 1 \\ 1 & \beta & 1 & - \end{pmatrix}$$

Tamura & Nei (1993)

203 time-reversible models

- 1 with one substitution type (111111)
- 31 with two substitution types (e.g., 121111, 112122, 122121, 121121, etc.)
- 90 with three substitution types (e.g., 111213, 123313, 121321, 122133, 121131, etc.)
- 65 with four substitution types (e.g., 121134, 123344, 123134, etc.)
- 15 with five substitution types (e.g., 123452, 123245, 112345, etc.)
- 1 with six substitution types (123456)

The Combinatorics...

The number of ways a set with n objects can be partitioned into disjoint and non-empty sets is described by the Bell numbers (Bell, 1934):

$$B_n = \sum_{k=0}^{n} S(n,k)$$

Bell Numbers

n	B_n
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4140
9	21147
10	115975
11	678570
12	4213597

Bell Numbers

n	B_n
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4140
9	21147
10	115975
11	678570
12	4213597

MCMC for substitution models

- Proposal mechanisms that change the substitution model.
- This often involves a change in the dimension of the problem (e.g., 111111 to 112111)
- Use reversible jump MCMC (Green, 1995)

Split Move

1. Choose one of the substitution groups with at least two members at random.
2. Split the substitutions in this group into two.

Current Model: 111222

Proposed Model: 111233

Merge Move

1. Choose two of the substitution groups at random.
2. Merge the substitutions in these groups into one.

Current Model: 111233

Proposed Model: 111222

Name	Gene(s)	Taxa	Sites
Angiosperms	phyA & phyC	46	1104
Archaea	rRNA	64	1620
Bats	IRBP	13	1255
Butterflies	wingless	106	378
Crocodiles	c-myc	68	818
Gophers	mtDNA	15	379
HIV-1	env	13	273
HIV-1	pol	23	2841
Lice	mtDNA	17	379
Lizards	mtDNA	30	1456
Mammals	mtDNA	23	9741
Parrotfish	mtDNA & timo-4C4	18	1689
Primates	mtDNA	12	898
Vertebrates	β-globin	17	432
Water snakes	mtDNA	34	2866
Whales	mtDNA	31	1140

MCMC

- Chains run for 10,000,000 cycles
- Samples taken during the first 1,000,000 cycles discarded as the "burn in"
- Posterior probability of a model calculated as the fraction of the time the chain visited that model
- Uniform prior assumed on models

Name	AIC	PP	BF
Angiosperms	189	189 (0.57)	266
Archaea	198	168 (0.74)	584
Bats	50	112 (0.34)	103
Butterflies	125	136 (0.27)	74
Crocodiles	134	125 (0.35)	109
Gophers	162	40 (0.46)	175
HIV-1 (env)	25	25 (0.29)	83
HIV-1 (pol)	157	50 (0.62)	322
Lice	15	15 (0.56)	255
Lizards	193	193 (0.68)	435
Mammals	203	193 (0.64)	353
Parrotfish	189	162 (0.56)	258
Primates	112	15 (0.32)	92
Vertebrates	125	125 (0.20)	52
Water snakes	191	166 (0.54)	238
Whales	162	15 (0.60)	111

Name	Size of 95% Credible Set
Angiosperms	4 models
Archaea	3 models
Bats	13 models
Butterflies	12 models
Crocodiles	9 models
Gophers	15 models
HIV-1 (env)	22 models
HIV-1 (pol)	6 models
Lice	11 models
Lizards	5 models
Mammals	2 models
Parrotfish	3 models
Primates	15 models
Vertebrates	17 models
Water snakes	8 models
Whales	12 models

	Number of Substitution Types					
Name	1	2	3	4	5	6
Angiosperms			0.01	0.55	35.16	9.91
Archaea				6.14	3.87	4.36
Bats		0.07	2.68	0.75	0.54	0.47
Butterflies		0.01	0.53	2.81	1.81	0.83
Crocodiles		0.01	0.36	4.64	1.23	0.36
Gophers		0.25	3.05	0.57	0.49	0.20
HIV-1 (env)		2.26	0.95	0.65	0.54	0.70
HIV-1 (pol)			2.21	0.95	0.69	0.38
Lice		7.32	0.66	0.19	0.04	0.02
Lizards			0.01	0.43	42.22	9.81
Mammals					21.91	115.5
Parrotfish				2.75	8.50	6.58
Primates		2.55	1.09	0.53	0.24	0.09
Vertebrates		0.03	0.49	2.59	2.28	1.83
Water snakes			0.11	4.54	3.59	2.53
Whales		8.19	0.59	0.19	0.05	0.02

Number of Substitution Types

Name	1	2	3	4	5	6
Angiosperms			0.01	0.55	35.16	9.91
Archaea				6.14	3.87	4.36
Bats		0.07	2.68	0.75	0.54	0.47
Butterflies		0.01	0.53	2.81	1.81	0.83
Crocodiles		0.01	0.36	4.64	1.23	0.36
Gophers		0.25	3.05	0.57	0.49	0.20
HIV-1 (env)		2.26	0.95	0.65	0.54	0.70
HIV-1 (pol)			2.21	0.95	0.69	0.38
Lice		7.32	0.66	0.19	0.04	0.02
Lizards			0.01	0.43	42.22	9.81
Mammals					21.91	115.5
Parrotfish				2.75	8.50	6.58
Primates		2.55	1.09	0.53	0.24	0.09
Vertebrates		0.03	0.49	2.59	2.28	1.83
Water snakes			0.11	4.54	3.59	2.53
Whales		8.19	0.59	0.19	0.05	0.02

What would LRT say?

- Compare 111111 vs. 121121 (reject 111111 all of the time)
- Compare 121121 vs. 121131 (reject 121121 half of the time)
- Compare 121131 vs. 123456 (reject 121131 13 of 16 times)

Commonly chosen models:

121121, 121131, 123323, 121323, 123343, 121341, 123143, 121343, 123341, 123454, 123324, 123141, 123123, 123345, 123451

Commonly chosen models:

121121, 121131, 123323, 121323, 123343, 121341, 123143, 121343, 123341, 123454, 123324, 123141, 123123, 123345, 123451

A transition rate is not constrained to be the same as a transversion rate for the models with high posterior probability (except for HIV-env and vertebrate β-globin)

```
for (int i=0; i<4; i++)
     {
     double sumL = 0.0, sumR = 0.0;
     for (int j=0; j<4; j++)
          {
          sumL += pL[i][j] * cL[j];
          sumR += pR[i][j] * cR[j];
          }
     cP[i] = sumL * sumR;
     }
```

```
for (int i=0; i<20; i++)
     {
     double sumL = 0.0, sumR = 0.0;
     for (int j=0; j<20; j++)
          {
          sumL += pL[i][j] * cL[j];
          sumR += pR[i][j] * cR[j];
          }
     cP[i] = sumL * sumR;
     }
```

$$\begin{pmatrix} - & r_{AC}\pi_C & r_{AG}\pi_G & r_{AT}\pi_T \\ r_{AC}\pi_A & - & r_{CG}\pi_G & r_{CT}\pi_T \\ r_{AG}\pi_A & r_{CG}\pi_C & - & r_{GT}\pi_T \\ r_{AT}\pi_A & r_{CT}\pi_C & r_{GT}\pi_G & - \end{pmatrix}$$

8 to 11 Free Parameters

$$\mathbf{Q} = \begin{pmatrix} - & \theta_{AR}\pi_R & \theta_{AN}\pi_N & \cdots & \theta_{AW}\pi_W & \theta_{AY}\pi_Y & \theta_{AV}\pi_V \\ \theta_{AR}\pi_A & - & \theta_{RN}\pi_N & \cdots & \theta_{RW}\pi_W & \theta_{RY}\pi_Y & \theta_{RV}\pi_V \\ \theta_{AN}\pi_A & \theta_{RN}\pi_R & - & \cdots & \theta_{NW}\pi_W & \theta_{NY}\pi_Y & \theta_{NV}\pi_V \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \theta_{AW}\pi_A & \theta_{RW}\pi_R & \theta_{NW}\pi_N & \cdots & - & \theta_{WY}\pi_Y & \theta_{WV}\pi_V \\ \theta_{AY}\pi_A & \theta_{RY}\pi_R & \theta_{NY}\pi_N & \cdots & \theta_{YW}\pi_W & - & \theta_{YV}\pi_V \\ \theta_{AV}\pi_A & \theta_{RV}\pi_R & \theta_{NV}\pi_N & \cdots & \theta_{WV}\pi_W & \theta_{YV}\pi_Y & - \end{pmatrix} \mu$$

208 to 379 Free Parameters

Some Fixed Amino Acid Models

- Dayhoff (Dayhoff et al, 1978)
- Jones (Jones et al., 1992)
- MtRev (Adachi & Hasegawa, 1996)
- WAG (Whelan & Goldman, 2001)
- MtMam (Cao et al., 1998; Yang et al., 1998)
- RtRev (Dimmic et al., 2002)
- CpRev (Adachi et al., 2000)
- Blosum (Henikoff & Henikoff, 1992)
- ECM (Kosiol et al., 2007)
- Vt (Muller & Vingron, 2000)
- Poisson (Bishop & Friday, 1987)

‘Centered’ Prior

- Constrain the sum of the 190 substitution rates to be one
- Place a Dirichlet probability distribution prior on these rate proportions
- One can center the prior on fixed amino acid models

Zwickl & Holder (2004)

Prior Probability Distribution on Substitution Rate Parameters

$$f(\boldsymbol{\theta} \,|\, \chi\boldsymbol{\nu}) = \frac{1}{b(\chi\boldsymbol{\nu})} \prod_{i<j\in\mathbf{S}} \theta_{ij}^{\chi\nu_{ij}-1}$$

$$\begin{array}{rcl} \theta_i & : & i\text{-th substitution rate} \\ \nu_i & : & i\text{-th centering parameter} \\ \chi & : & \text{controls the variance} \end{array}$$

$$\mathbf{Q} = \left(\begin{array}{ccc} - & 1.0\,\pi_1 & 1.2\,\pi_2 \\ 1.0\,\pi_0 & - & 1.4\,\pi_2 \\ 1.2\,\pi_0 & 1.4\,\pi_1 & - \end{array} \right) \mu$$

$$\nu_{01} = 1.0/(1.0 + 1.2 + 1.4)$$

$$\nu_{02} = 1.2/(1.0 + 1.2 + 1.4)$$

$$\nu_{12} = 1.4/(1.0 + 1.2 + 1.4)$$

Marginal distribution of the i-th rate is a Beta, with parameters $\nu_{ij}\chi$ and $\chi(1-\nu_{ij})$

Expected value of the i-th rate is $\mathrm{E}(\theta_{ij}) = \nu_{ij}$

Variance of the i-th rate is $\mathrm{Var}(\theta_{ij}) = \dfrac{\nu_{ij}(1-\nu_{ij})}{\chi+1}$

Data Sets

- ADH sequences sampled from 23 Drosophila
- β-globin sequences sampled from 17 vertebrates
- coat protein sequences from 9 bacteriophage
- replicase sequences from 9 bacteriophage
- env sequences from 23 encephalitis virus samples
- pol sequences from 23 HIV samples
- hemagglutin sequences from 28 influenza (type A) samples
- E-glycoprotein sequences from 18 Flavivirus

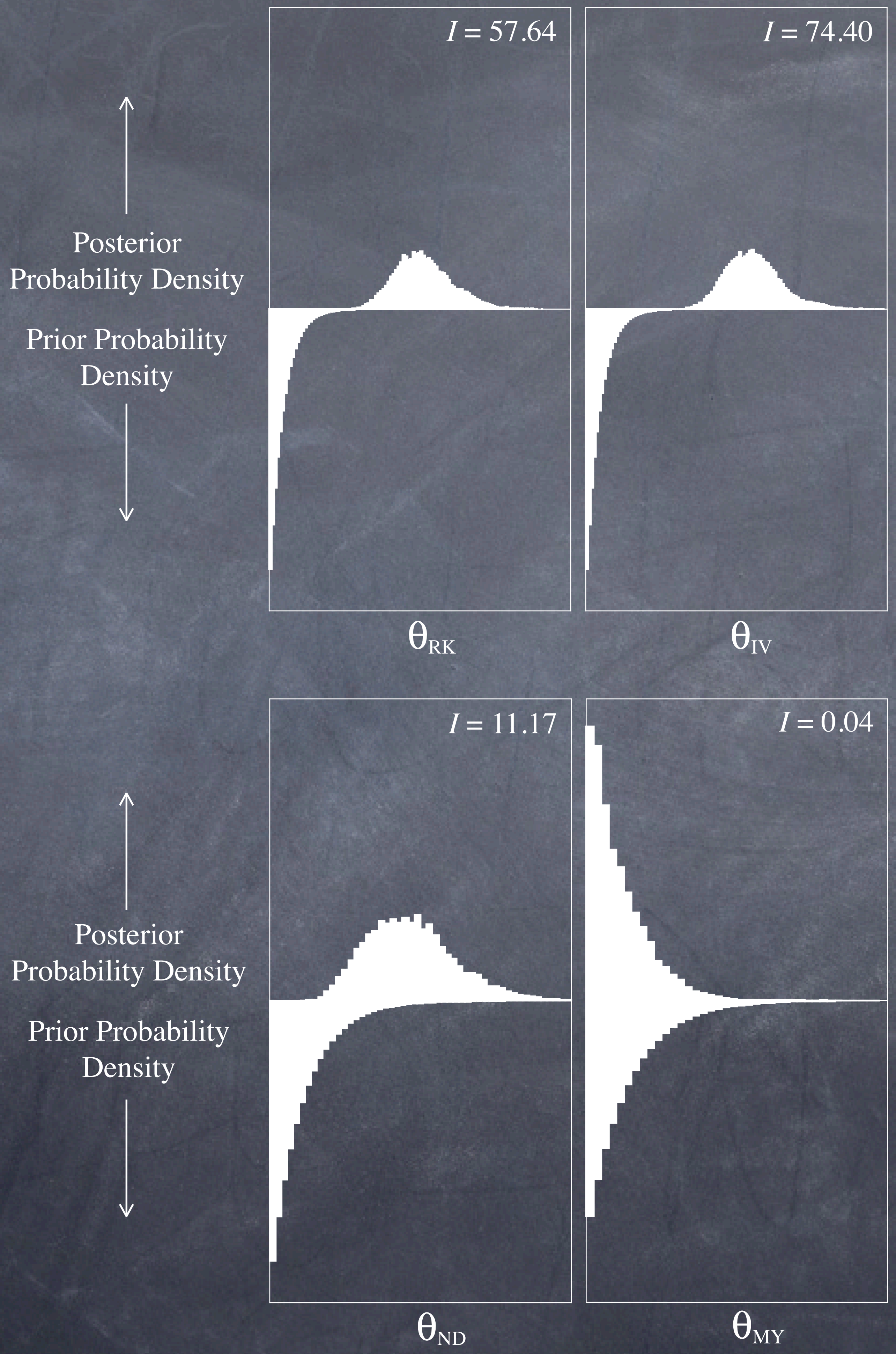
Kullback-Leibler (1951) Divergence

$$I(f,g) = \int f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx$$

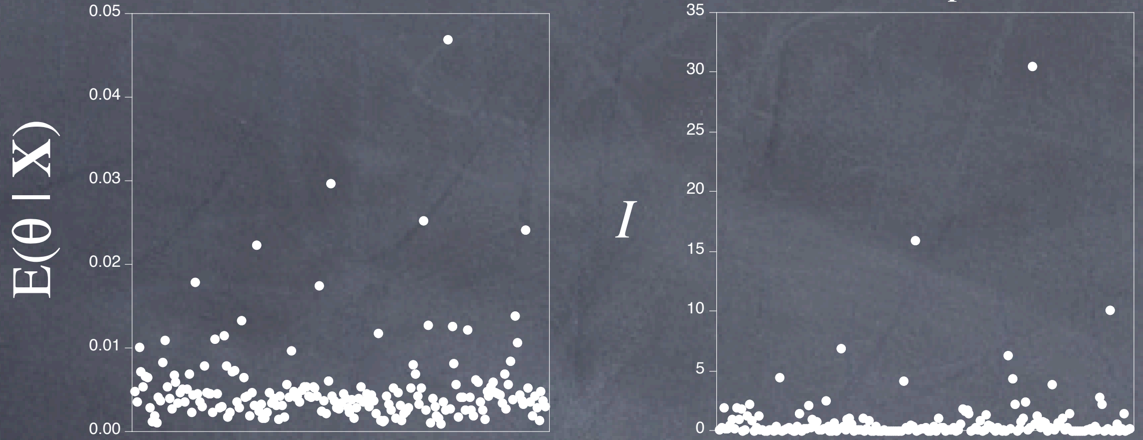
Prior: Beta

Posterior: Beta?

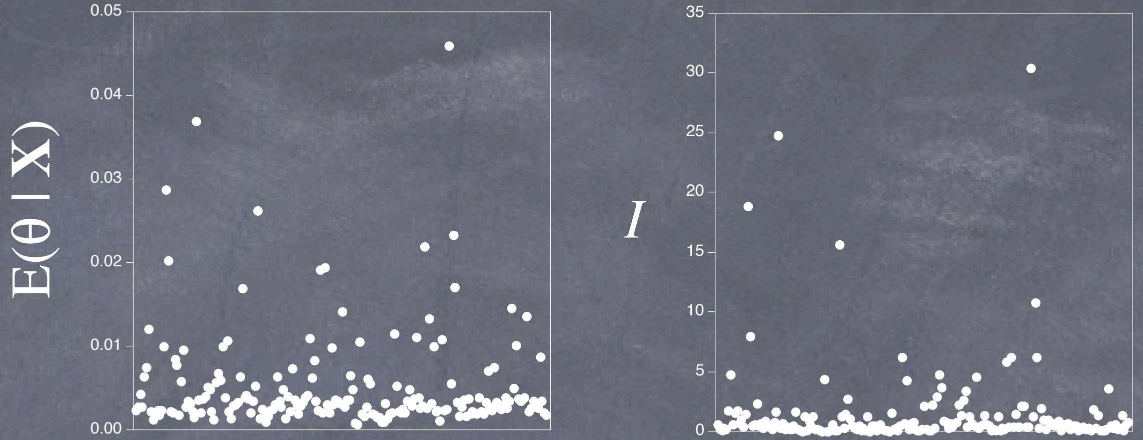
$$I = \ln\frac{b(a_2,b_2)}{b(a_1,b_1)} - (a_2-a_1)\psi(a_1) - (b_2-b_1)\psi(b_1) + (a_2-a_1+b_2-b_1)\psi(a_1+b_1)$$



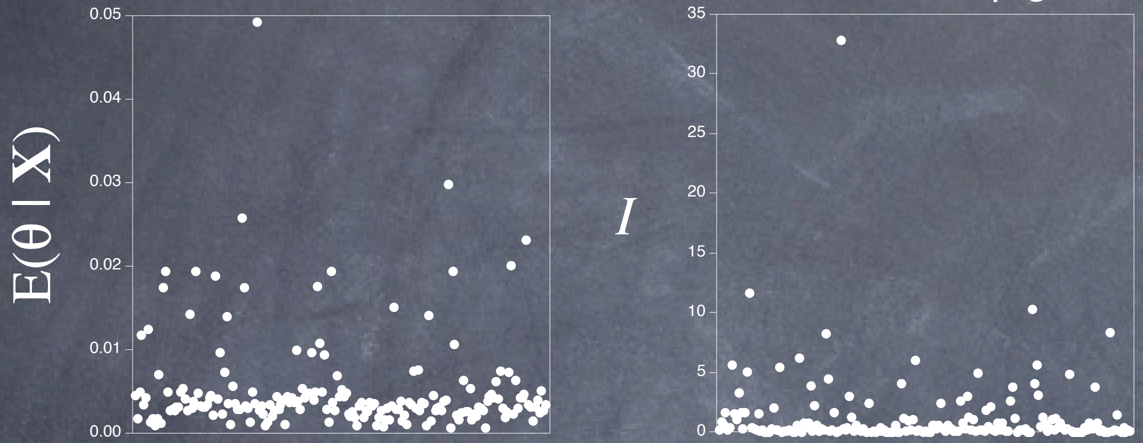
Posterior Probability Density
Prior Probability Density
$I = 57.64$
θ_{RK}
$I = 74.40$
θ_{IV}
Posterior Probability Density
Prior Probability Density
$I = 11.17$
θ_{ND}
$I = 0.04$
θ_{MY}



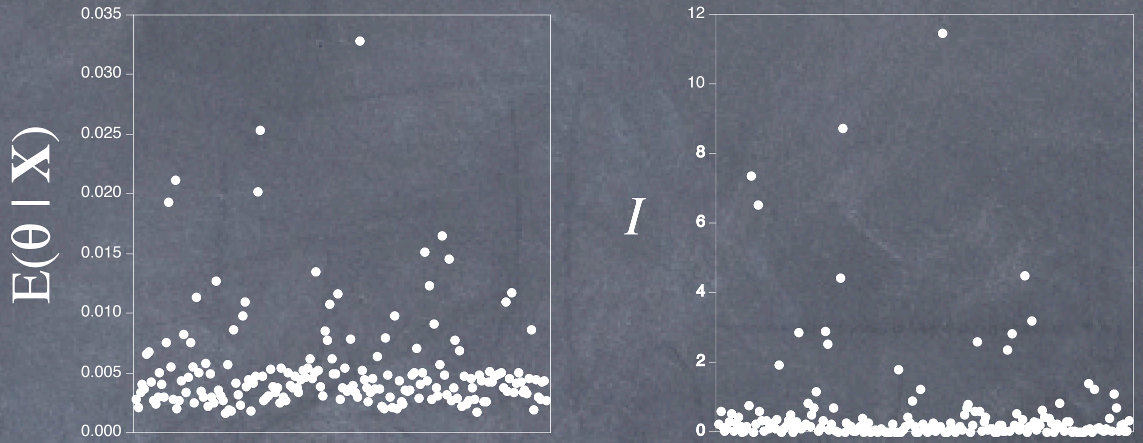
Drosophila ADH



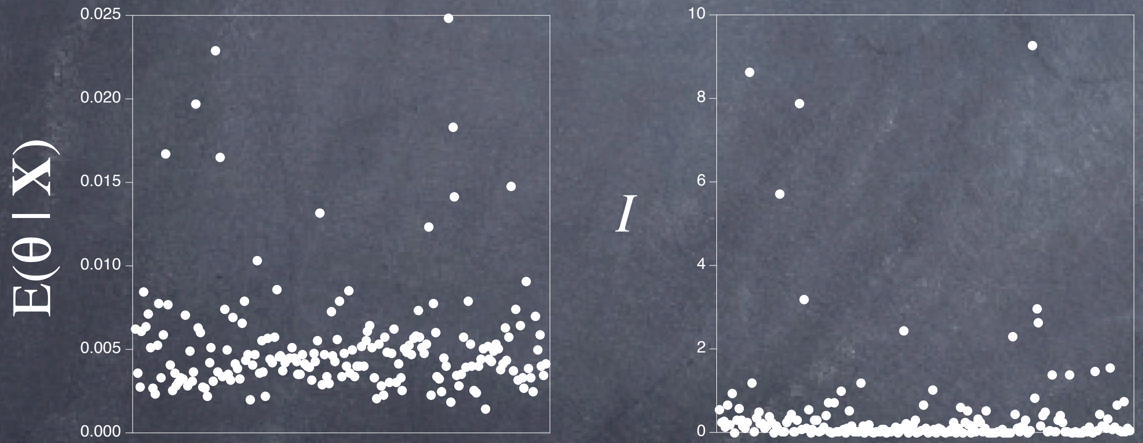
Flavivirus



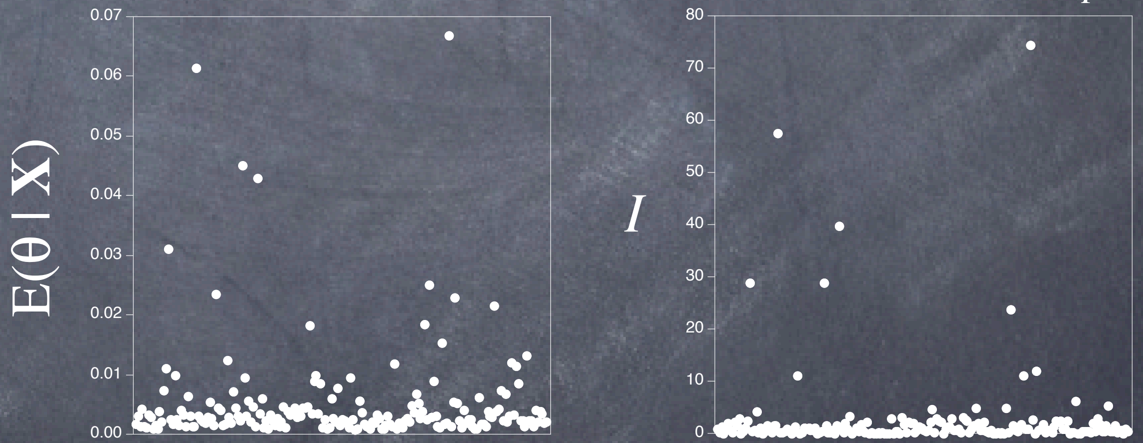
Vertebrate β-globin



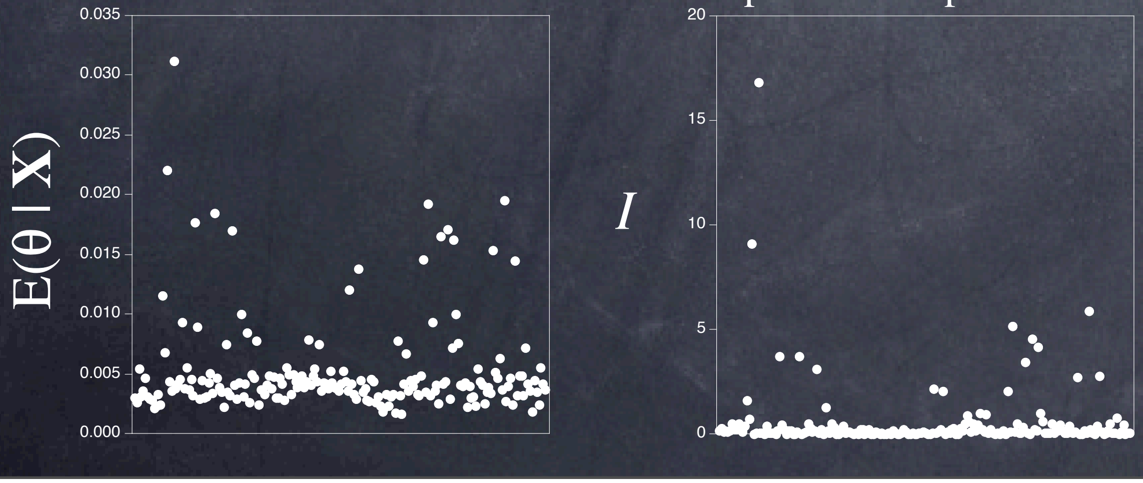
Influenza



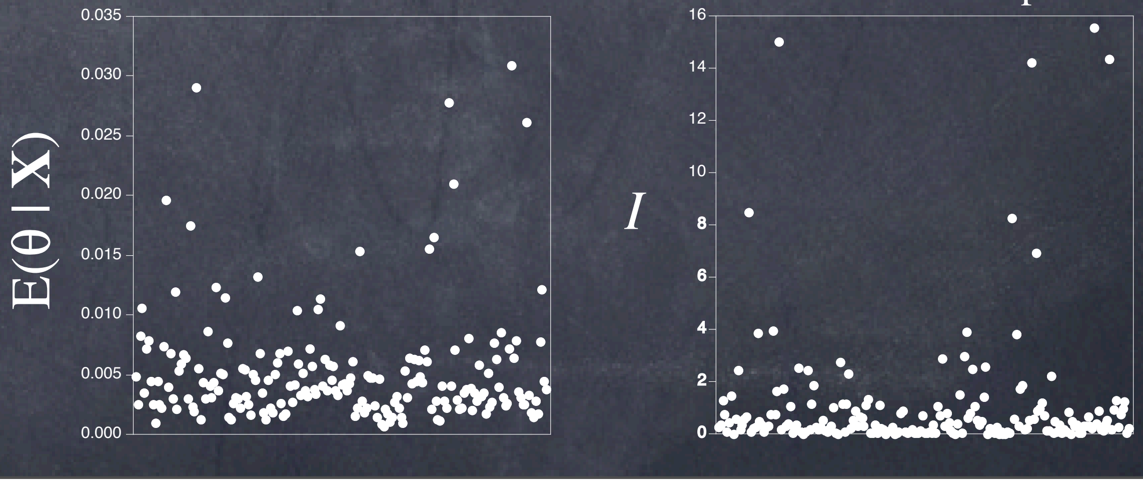
Leviviridae coat



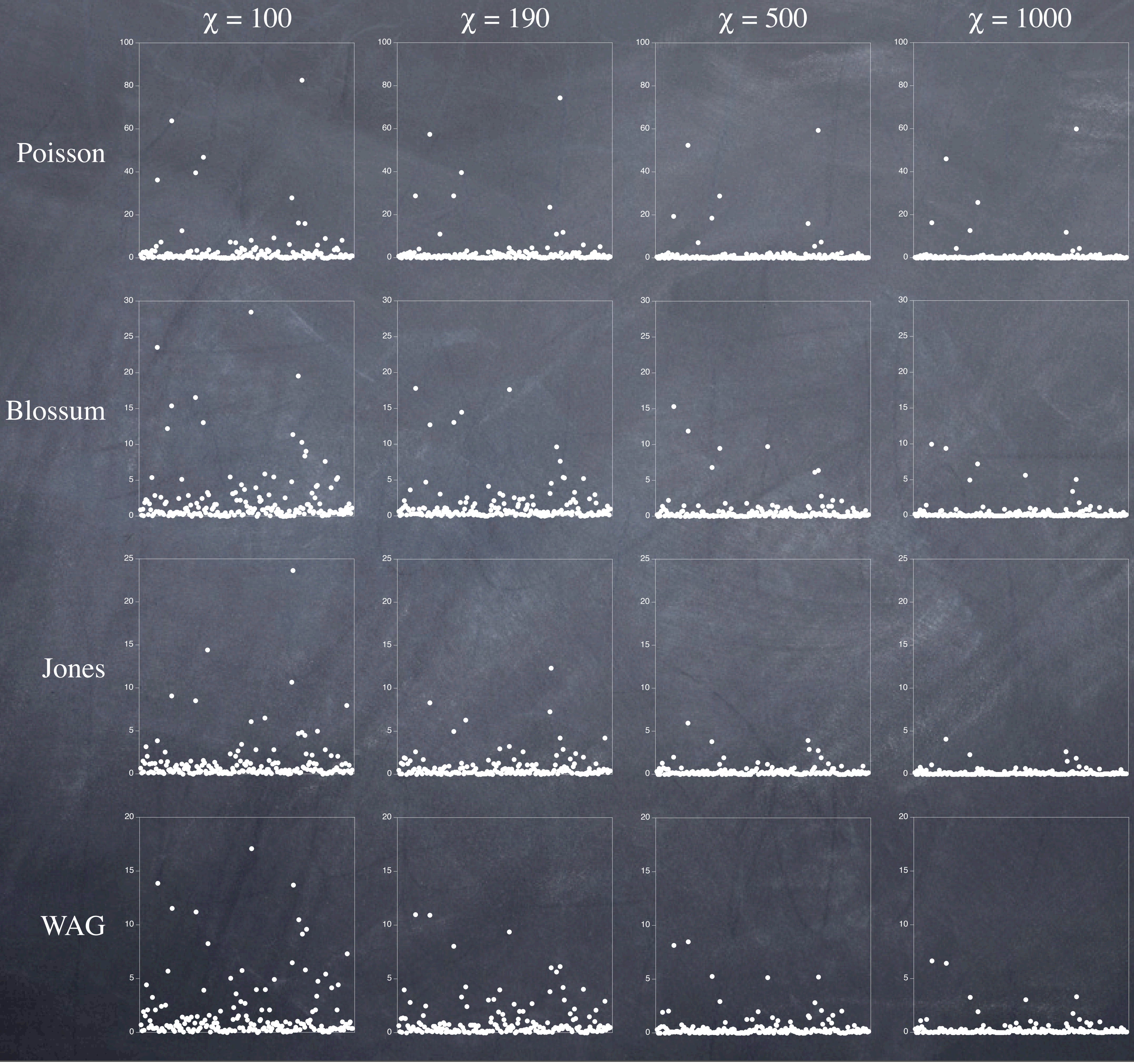
HIV *pol*



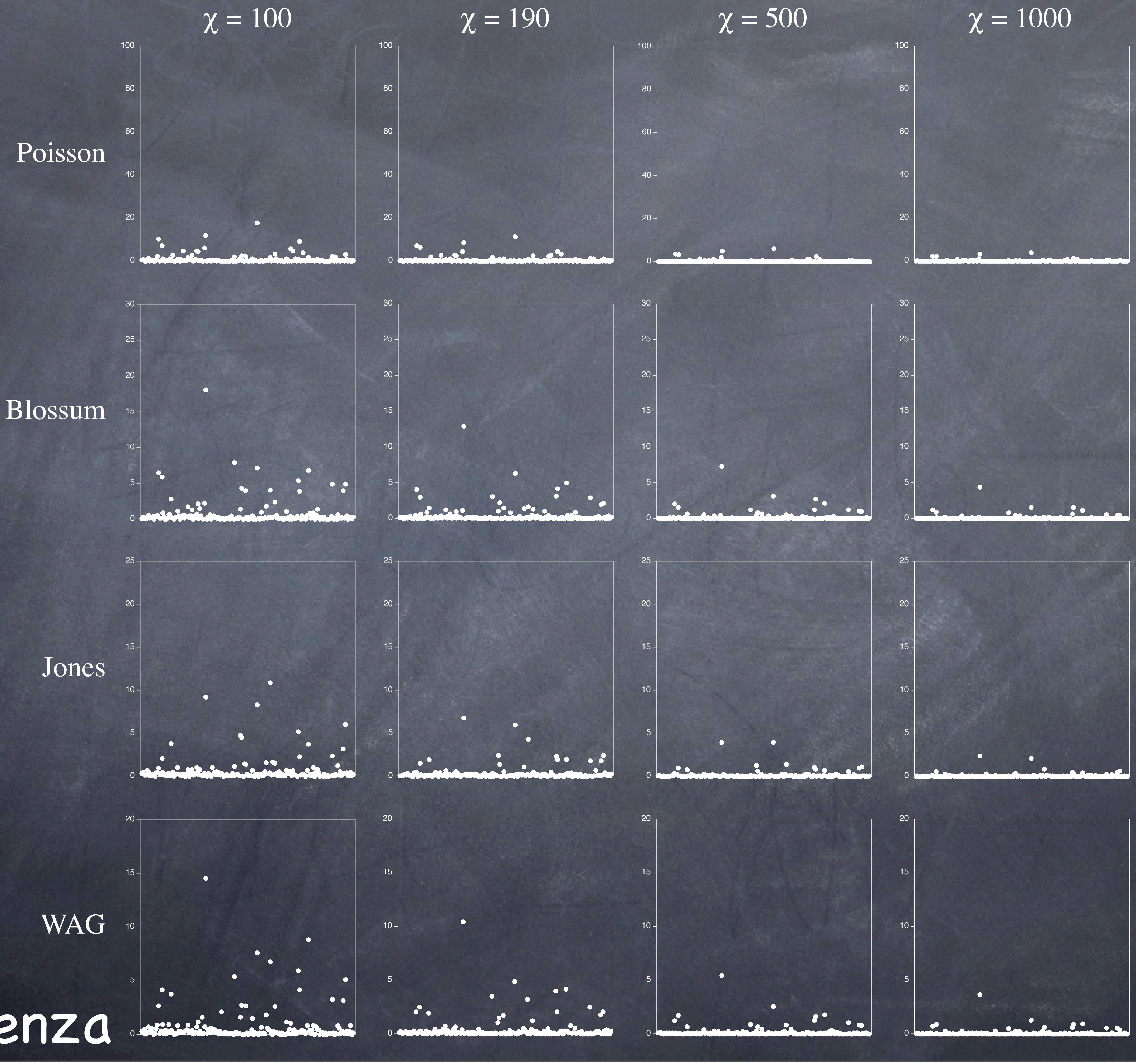
Japanese encephalitis *env*



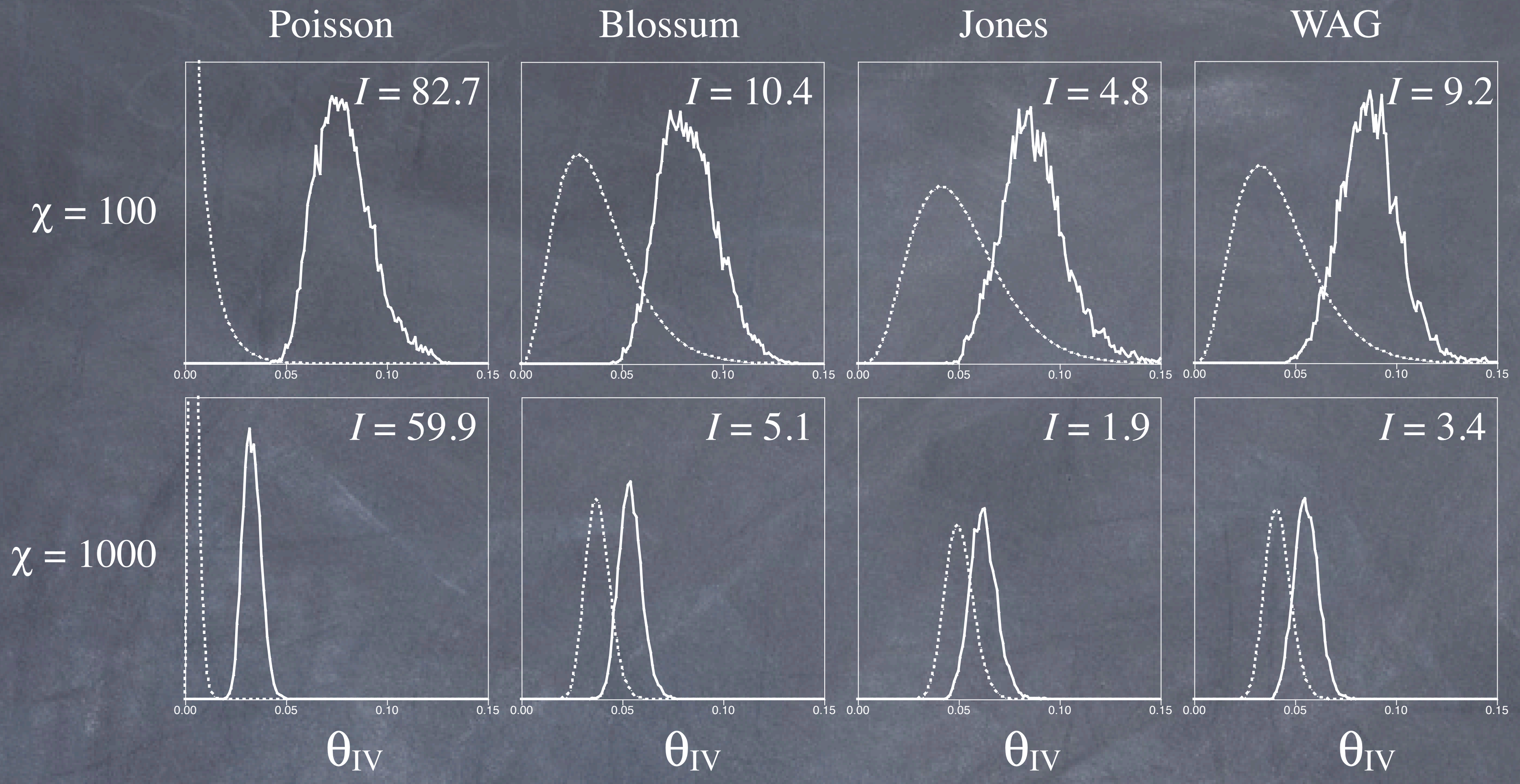
Leviviridae replicase

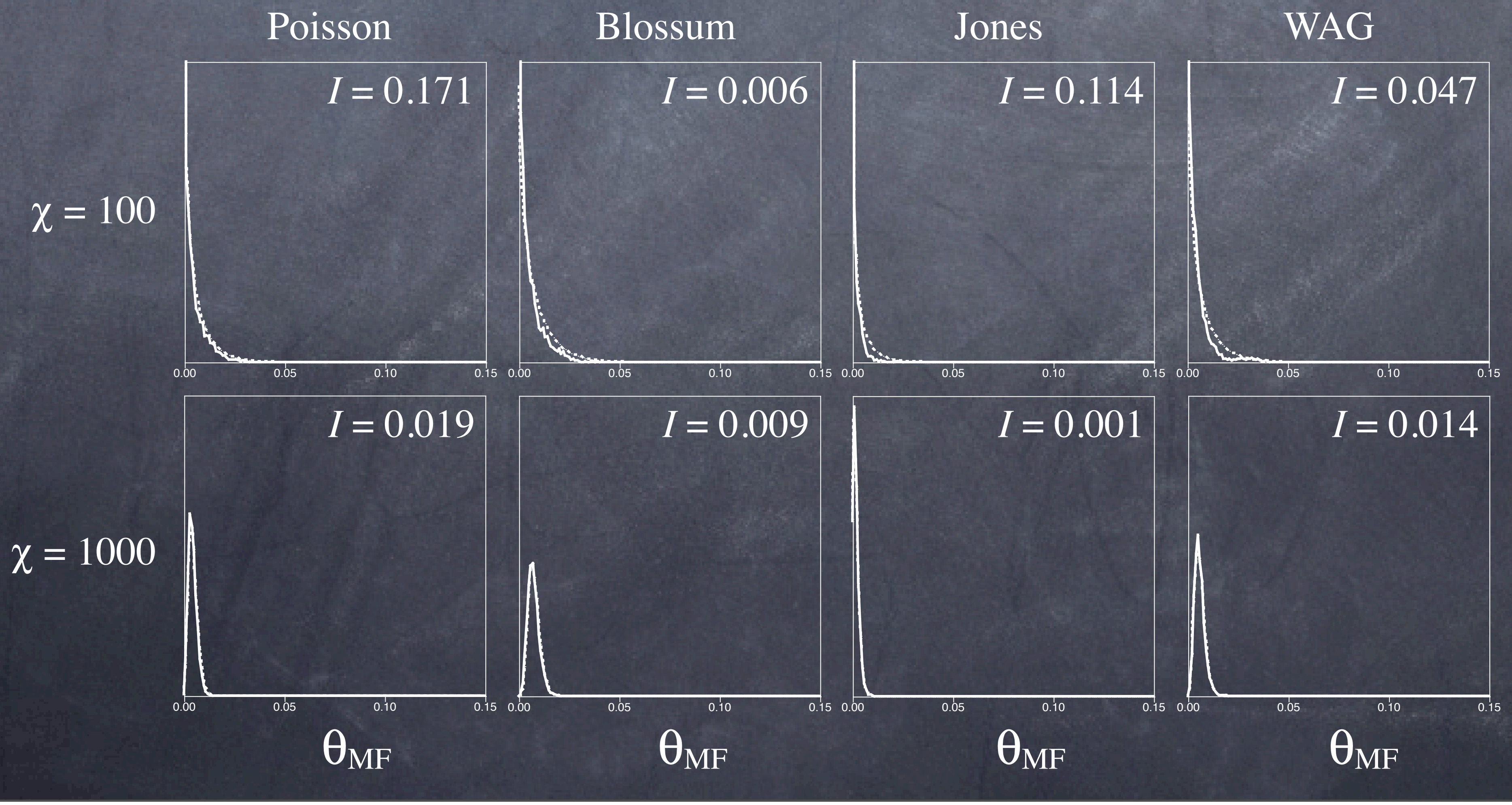

χ = 100
χ = 190
χ = 500
χ = 1000
Poisson
0
20
40
60
80
100
Blossum
0
5
10
15
20
25
30
Jones
0
5
10
15
20
25
WAG
0
5
10
15
20

HIV

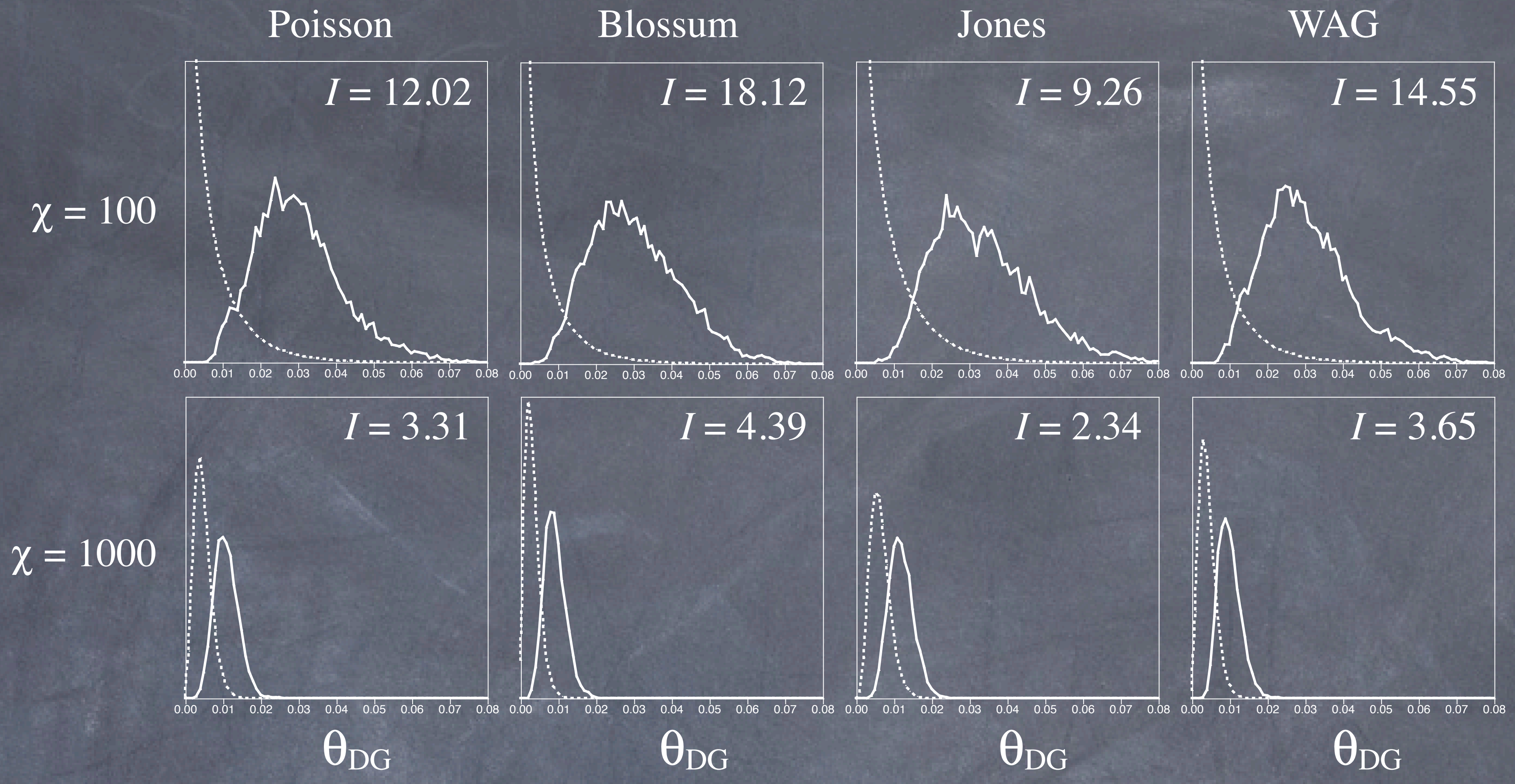

χ = 100
χ = 190
χ = 500
χ = 1000
Poisson
Blossum
Jones
WAG
0
20
40
60
80
100
0
5
10
15
20
25
30
0
5
10
15
20
25
0
5
10
15
20

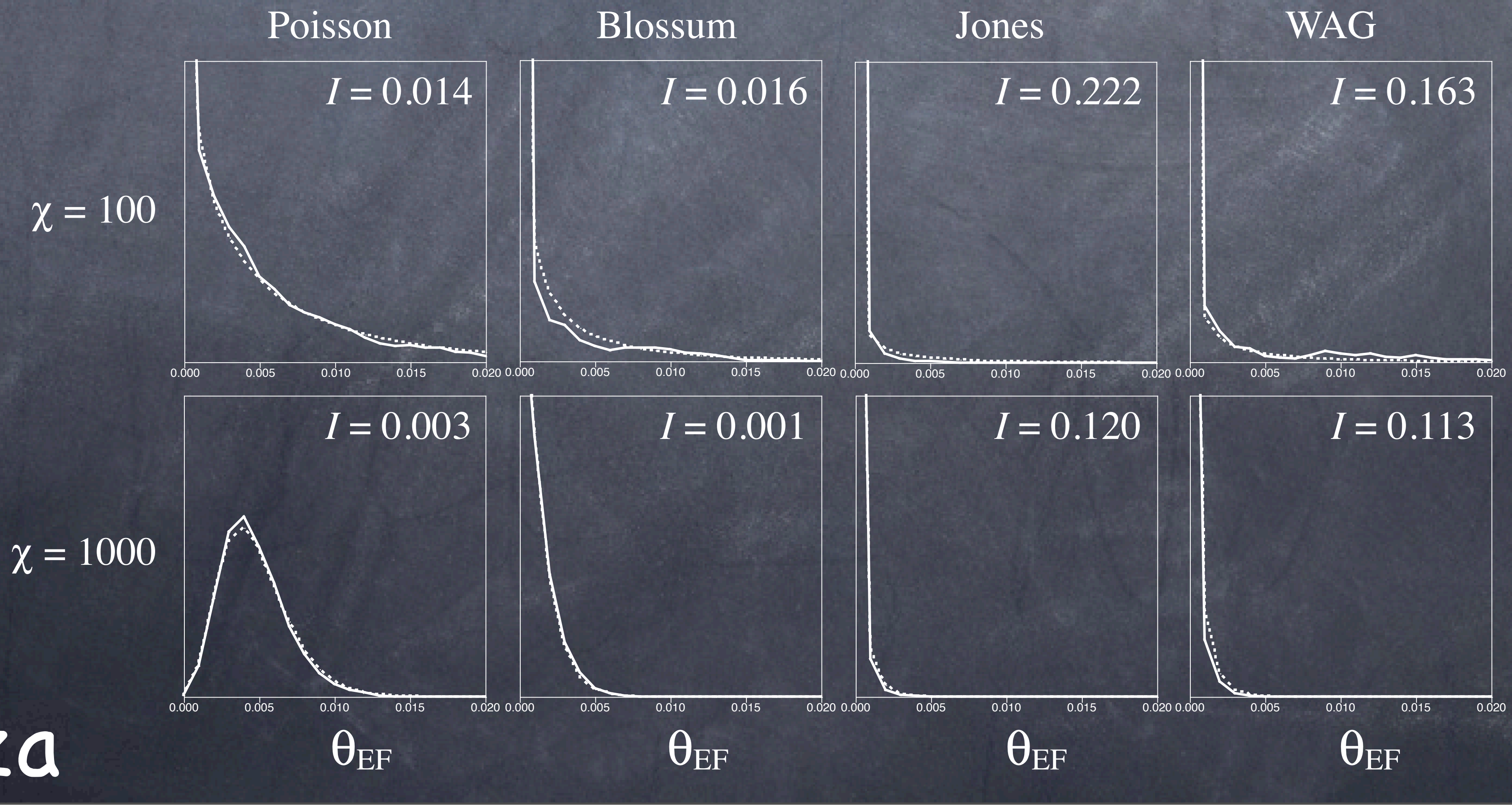
Influenza


Poisson
Blossum
Jones
WAG
χ = 100
I = 82.7
I = 10.4
I = 4.8
I = 9.2
χ = 1000
I = 59.9
I = 5.1
I = 1.9
I = 3.4
0.00
0.05
0.10
0.15
θIV
θIV
θIV
θIV


Poisson
Blossum
Jones
WAG
χ = 100
I = 0.171
I = 0.006
I = 0.114
I = 0.047
χ = 1000
I = 0.019
I = 0.009
I = 0.001
I = 0.014
0.00
0.05
0.10
0.15
θMF
θMF
θMF
θMF

HIV


Poisson
Blossum
Jones
WAG
χ = 100
χ = 1000
$I = 12.02$
$I = 18.12$
$I = 9.26$
$I = 14.55$
$I = 3.31$
$I = 4.39$
$I = 2.34$
$I = 3.65$
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
θ_{DG}


Poisson
Blossum
Jones
WAG
χ = 100
χ = 1000
$I = 0.014$
$I = 0.016$
$I = 0.222$
$I = 0.163$
$I = 0.003$
$I = 0.001$
$I = 0.120$
$I = 0.113$
0.000
0.005
0.010
0.015
0.020
θ_{EF}

Influenza

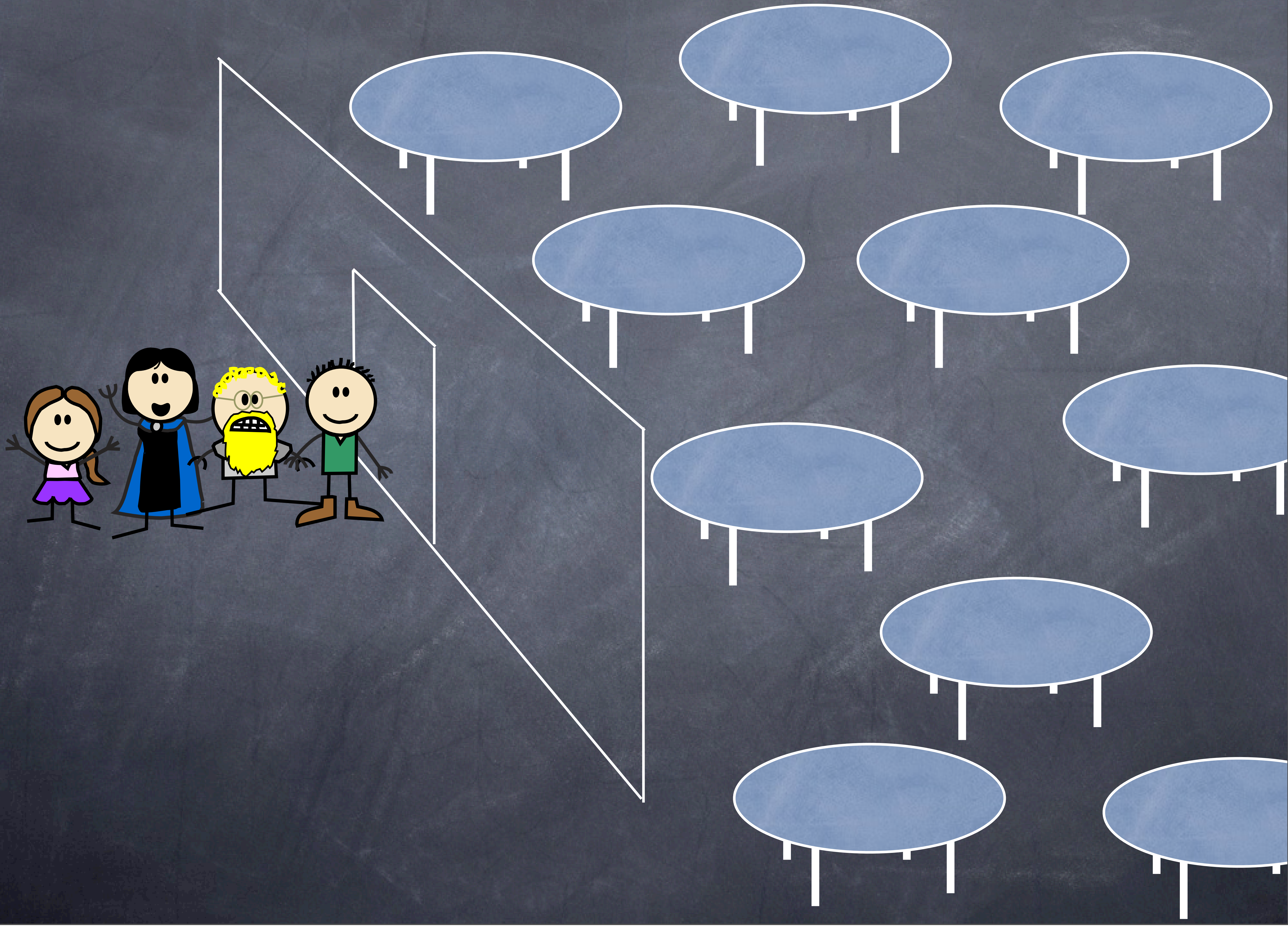
$$\mathbf{Q} = \left(\begin{array}{ccc} - & \theta_{01}\,\pi_1 & \theta_{02}\,\pi_2 \\ \theta_{01}\,\pi_0 & - & \theta_{12}\,\pi_2 \\ \theta_{02}\,\pi_0 & \theta_{12}\,\pi_1 & - \end{array} \right) \mu$$

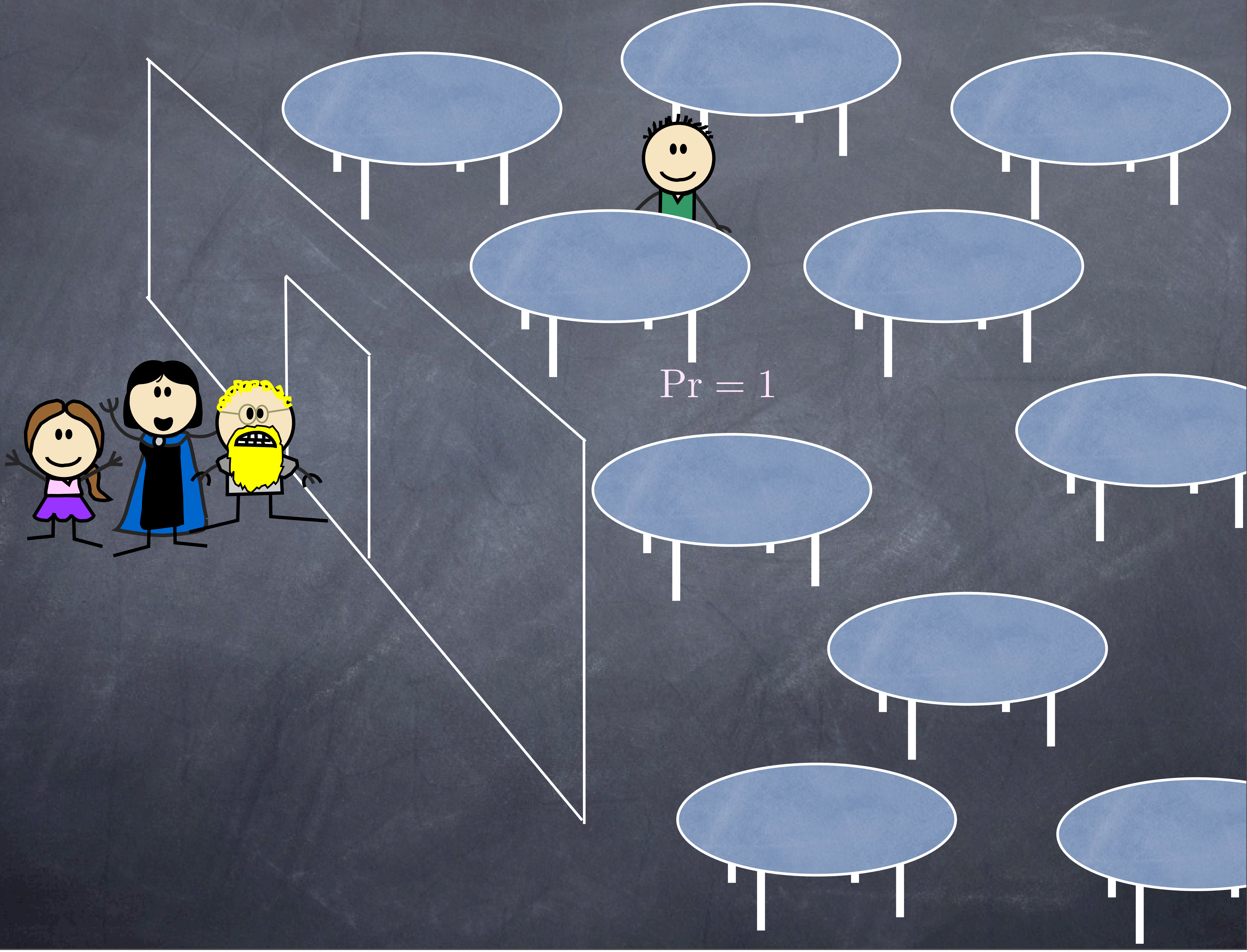
01	02	12
1	1	1
1	1	2
1	2	1
1	2	2
1	2	3

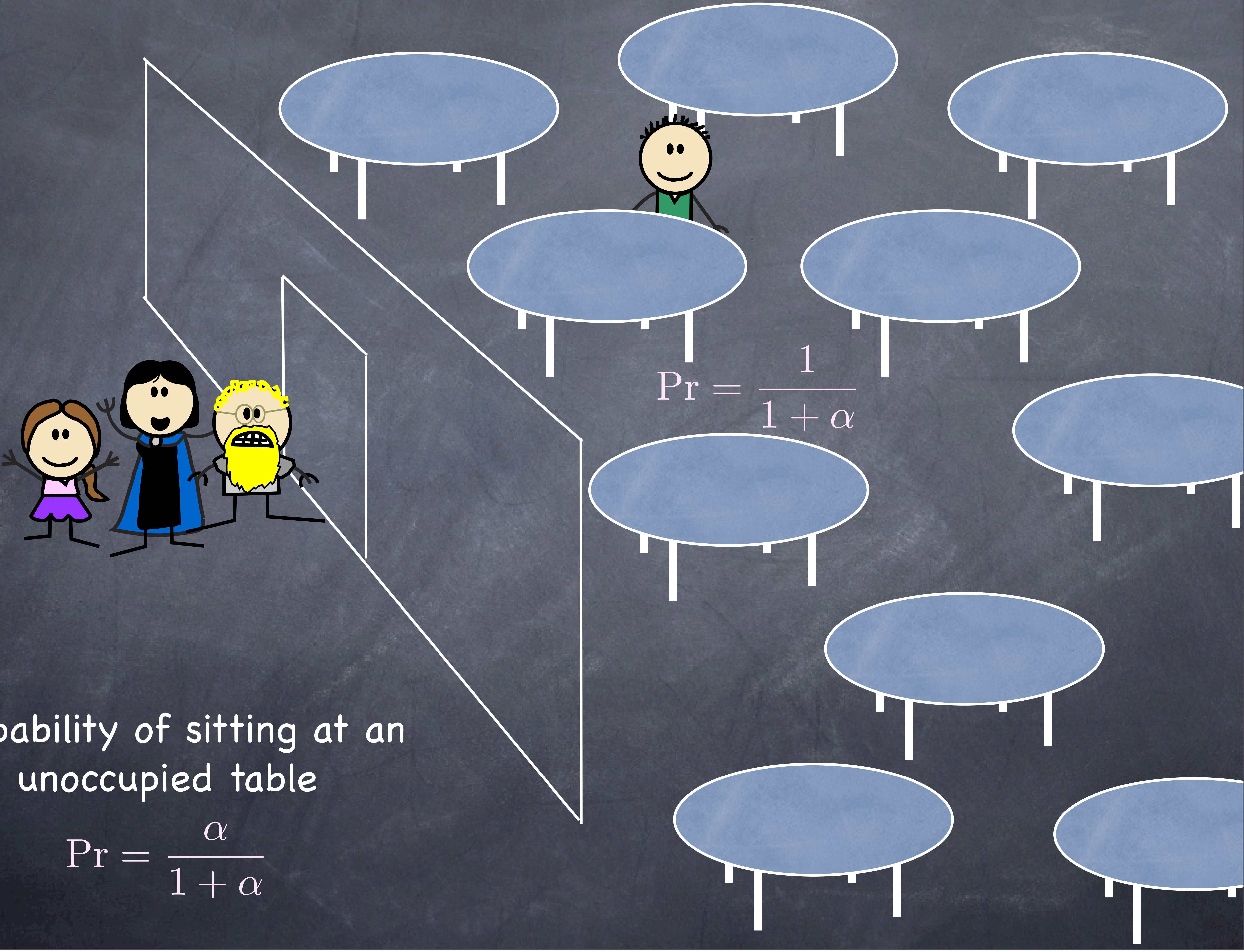
n	$\mathcal{B}(n)$
1	1
2	2
3	5
4	15
5	52
6	203
7	877
8	4140
9	21147
10	115975
100	4.75×10^{115}
190	6.59×10^{258}

Dirichlet Process Prior

- Ferguson (1973), Antoniak (1974)
- Provides a flexible way to model situations in which the data elements are drawn from a mixture of simpler parametric distributions
- Connections to the Ewen's sampling formula
- Was used in a phylogenetic context by Lartillot and Philippe (MBE 2004)
- Often described as the "Chinese Restaurant Table" process



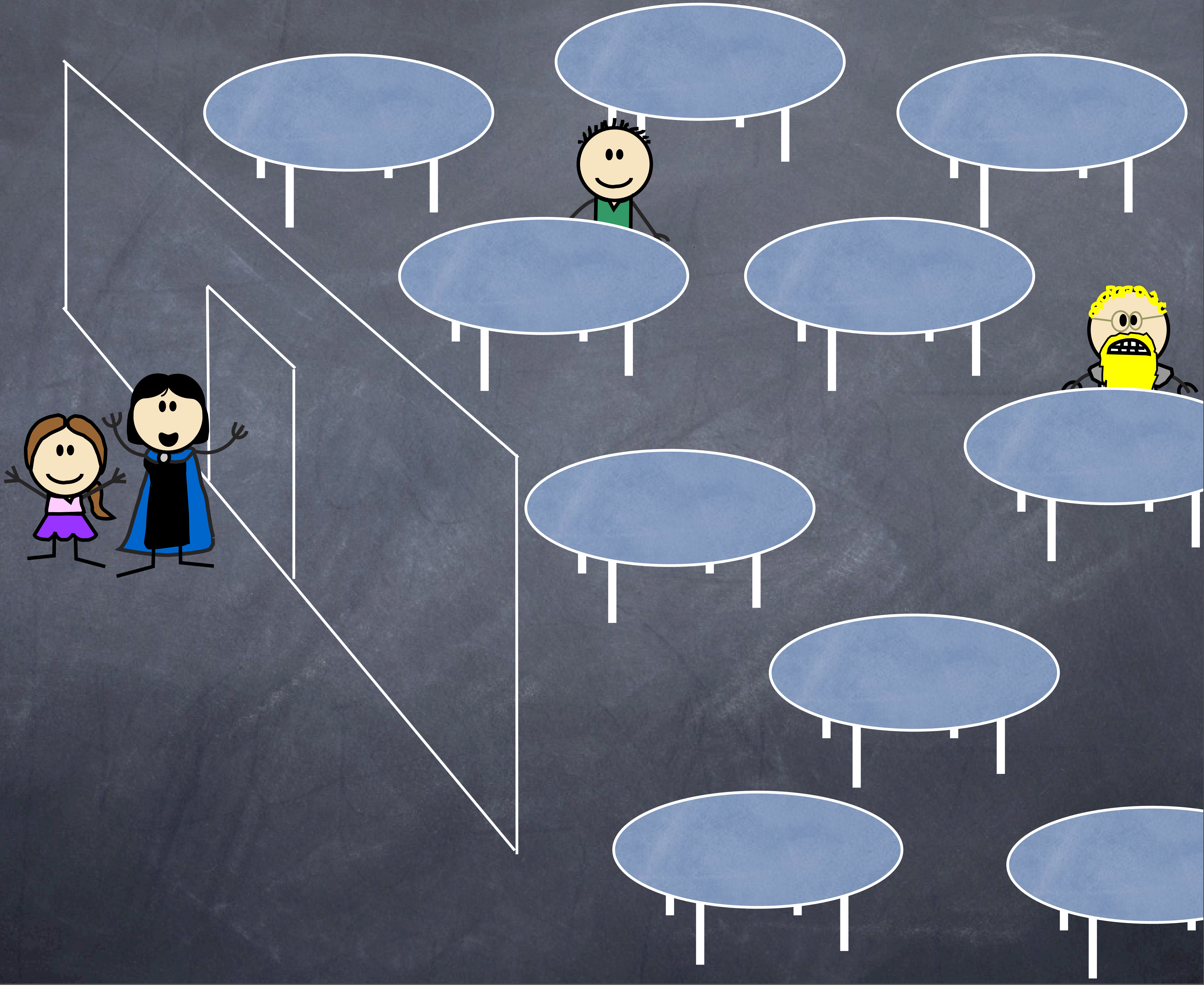

$\Pr = 1$

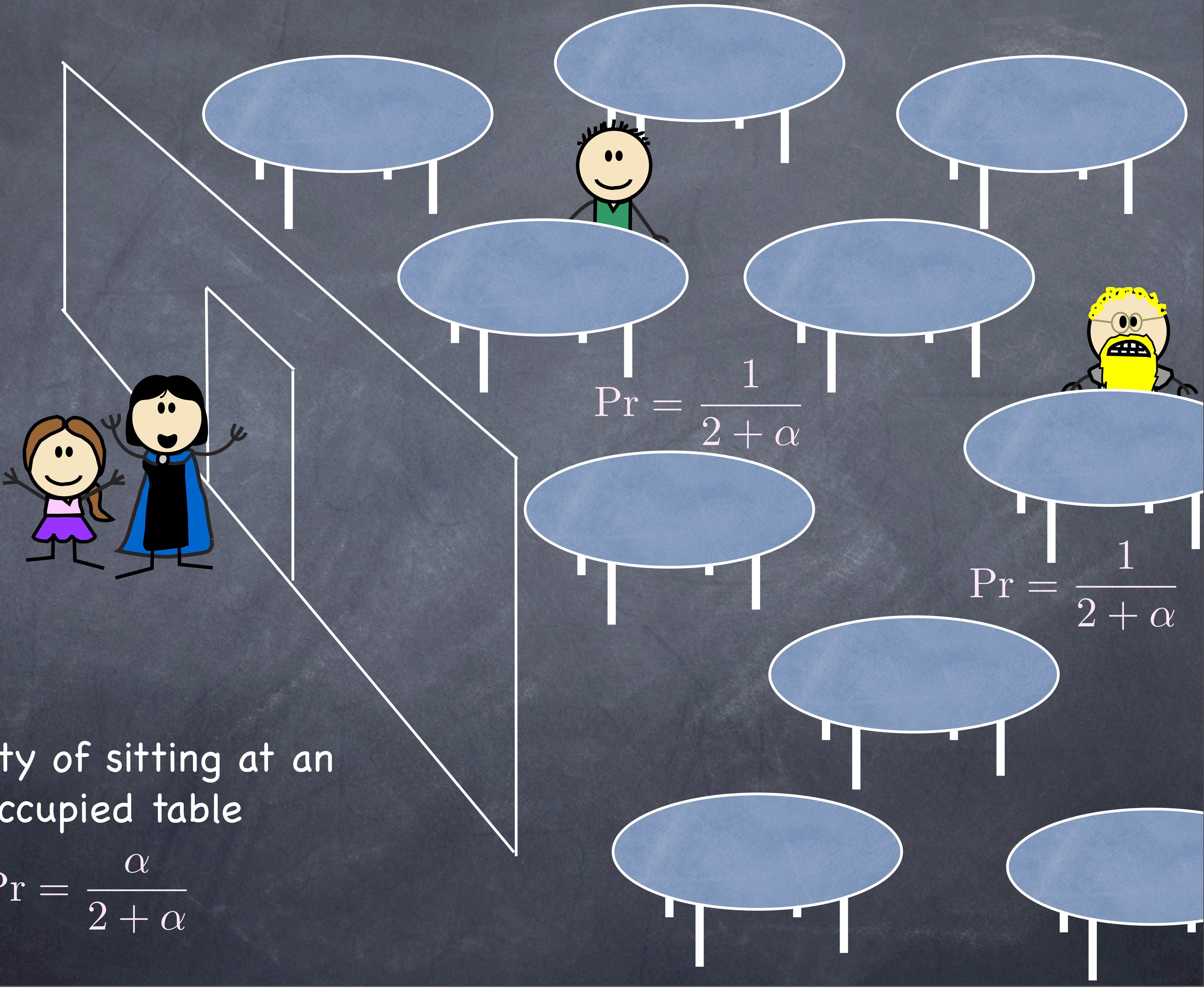


$$\Pr = \frac{1}{1+\alpha}$$

Probability of sitting at an unoccupied table

$$\Pr = \frac{\alpha}{1+\alpha}$$



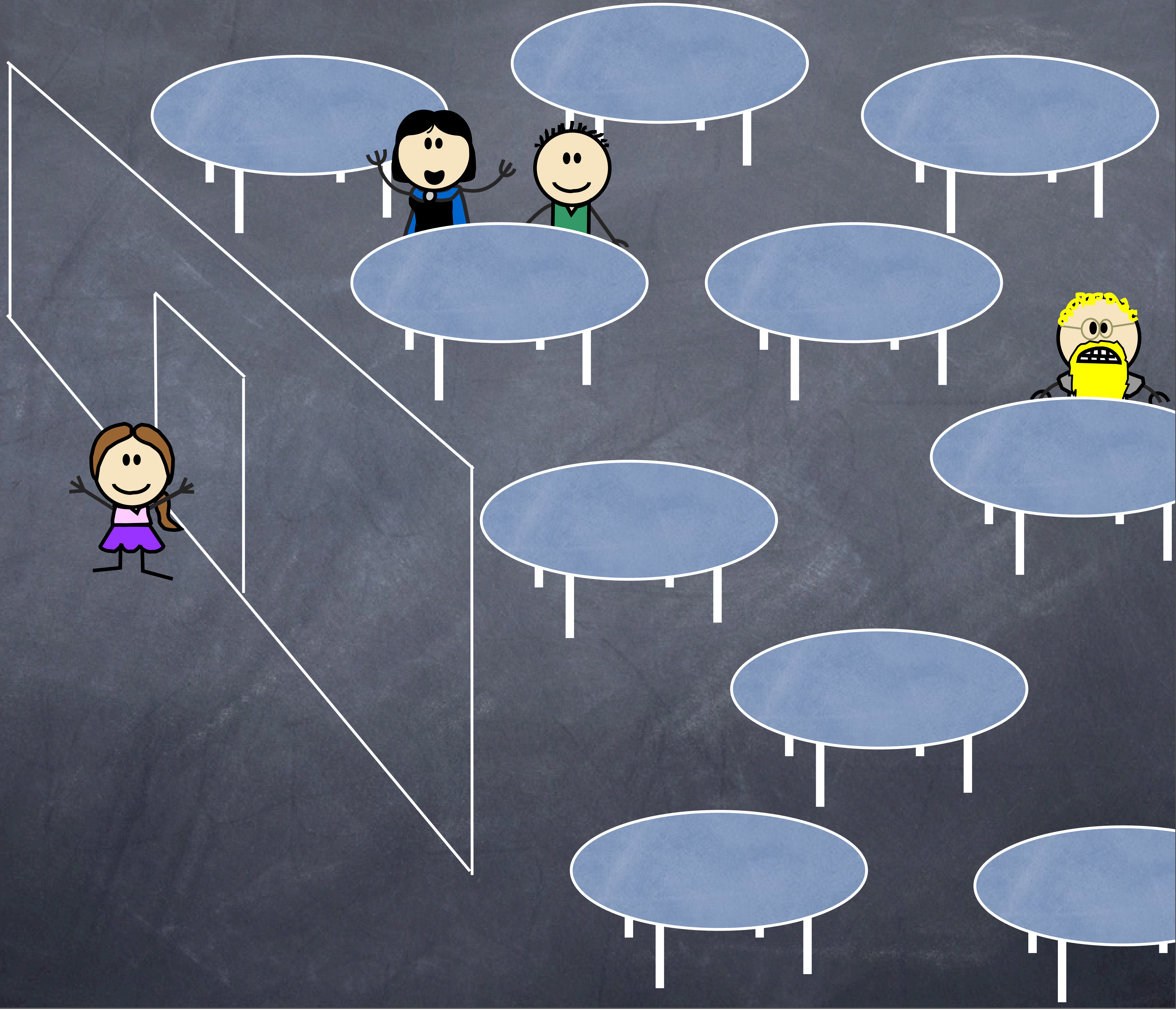


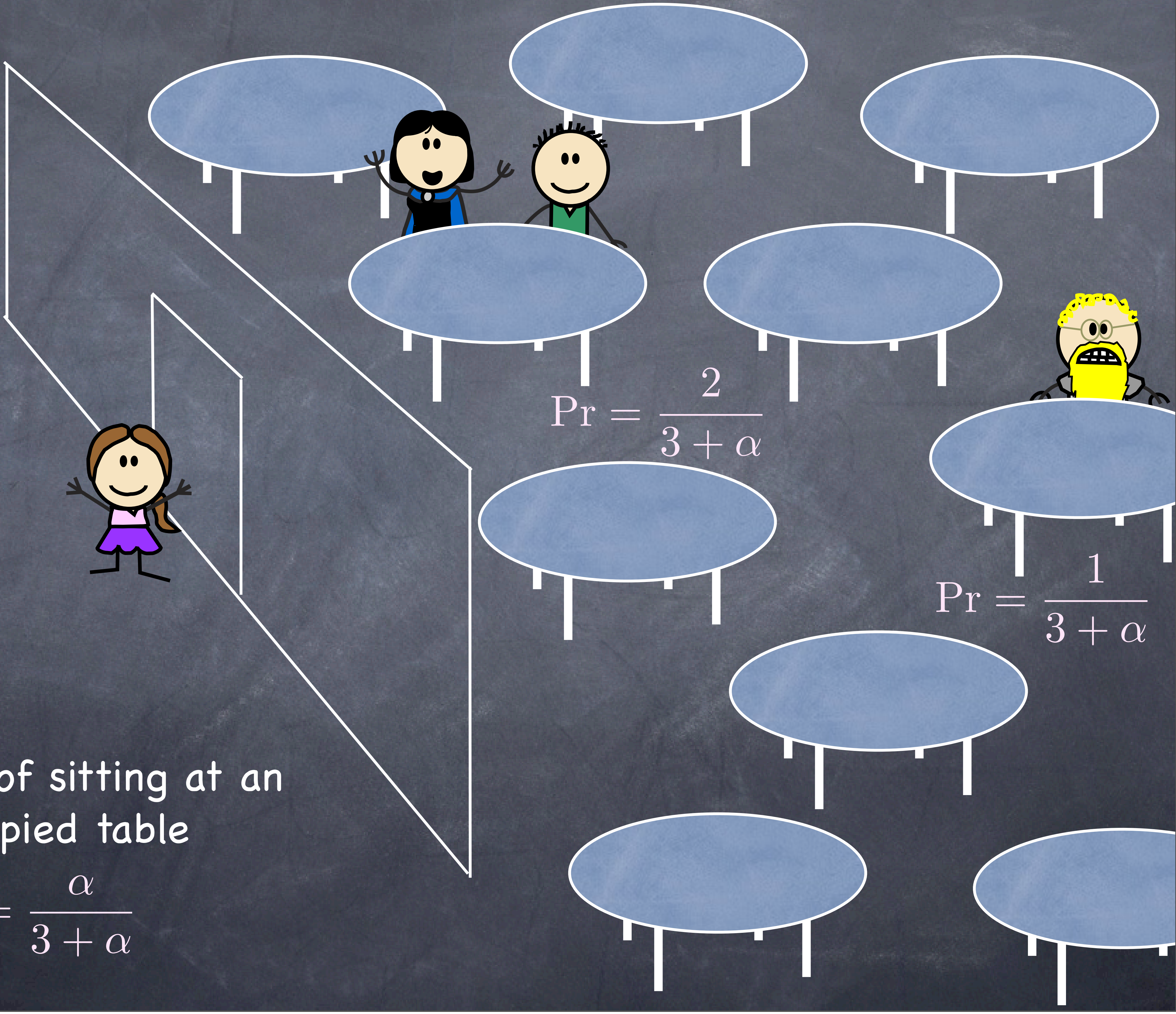
$$\Pr = \frac{1}{2+\alpha}$$

$$\Pr = \frac{1}{2+\alpha}$$

Probability of sitting at an unoccupied table

$$\Pr = \frac{\alpha}{2+\alpha}$$



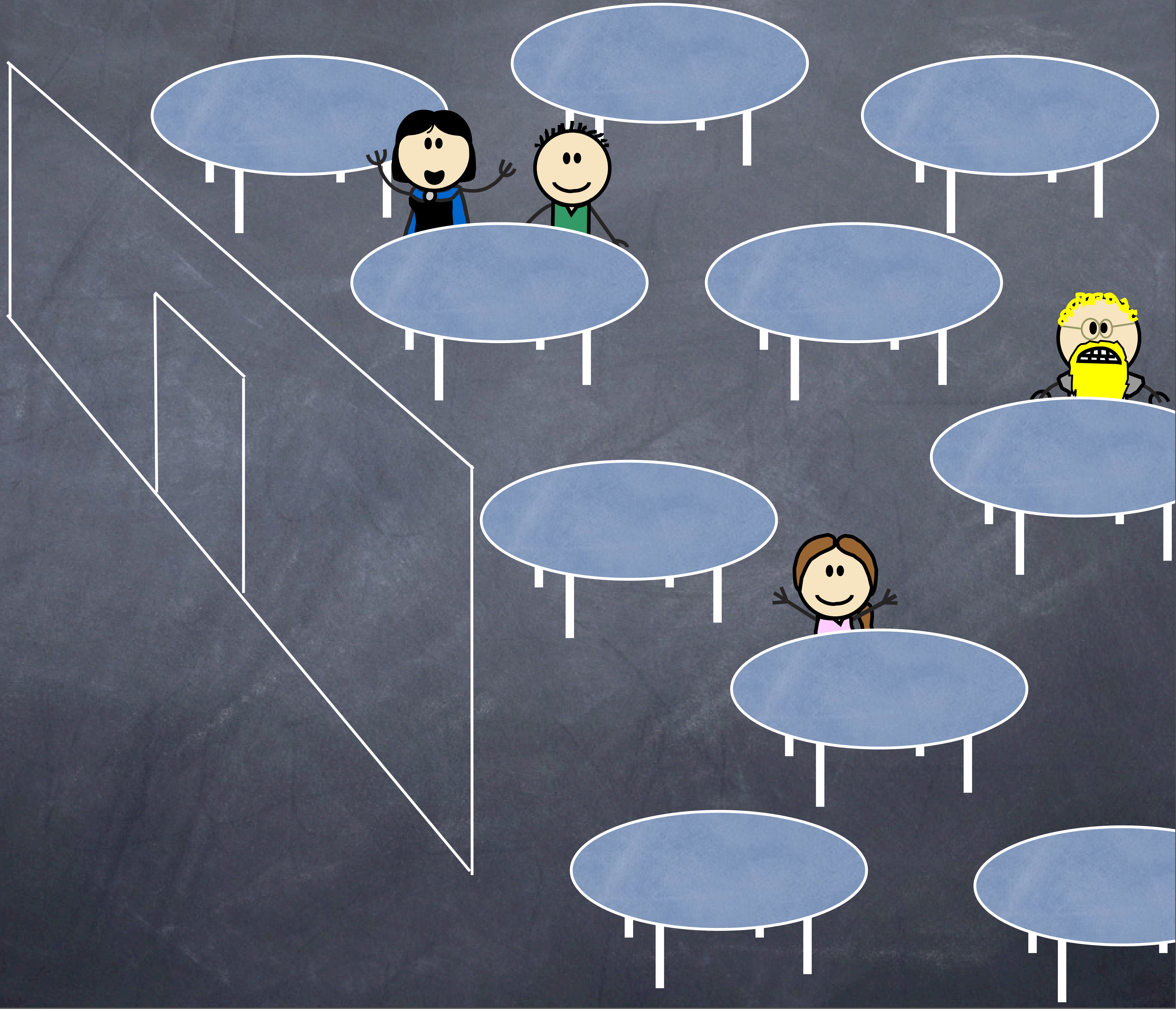


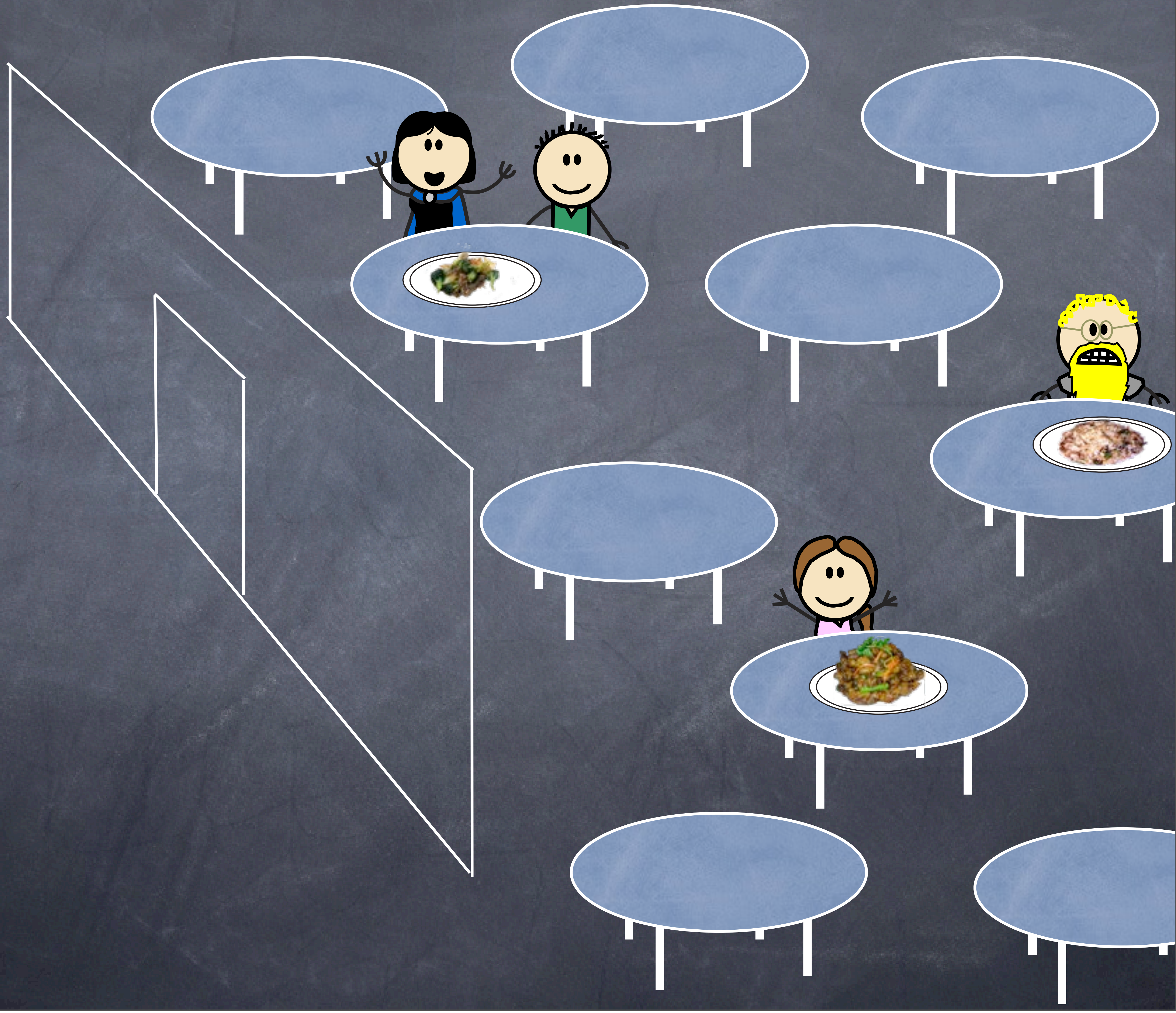
$$\Pr = \frac{2}{3+\alpha}$$

$$\Pr = \frac{1}{3+\alpha}$$

Probability of sitting at an unoccupied table

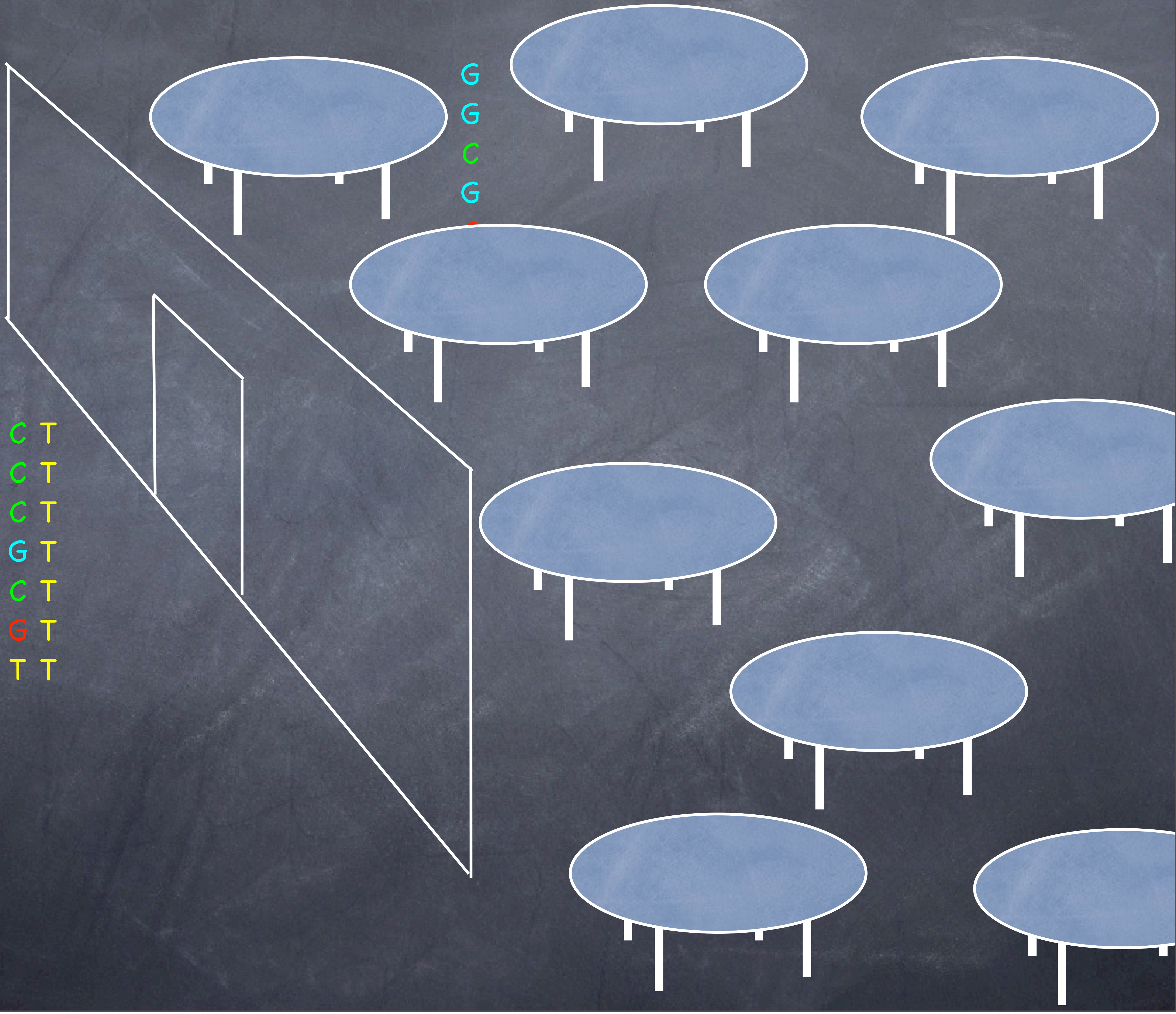
$$\Pr = \frac{\alpha}{3+\alpha}$$

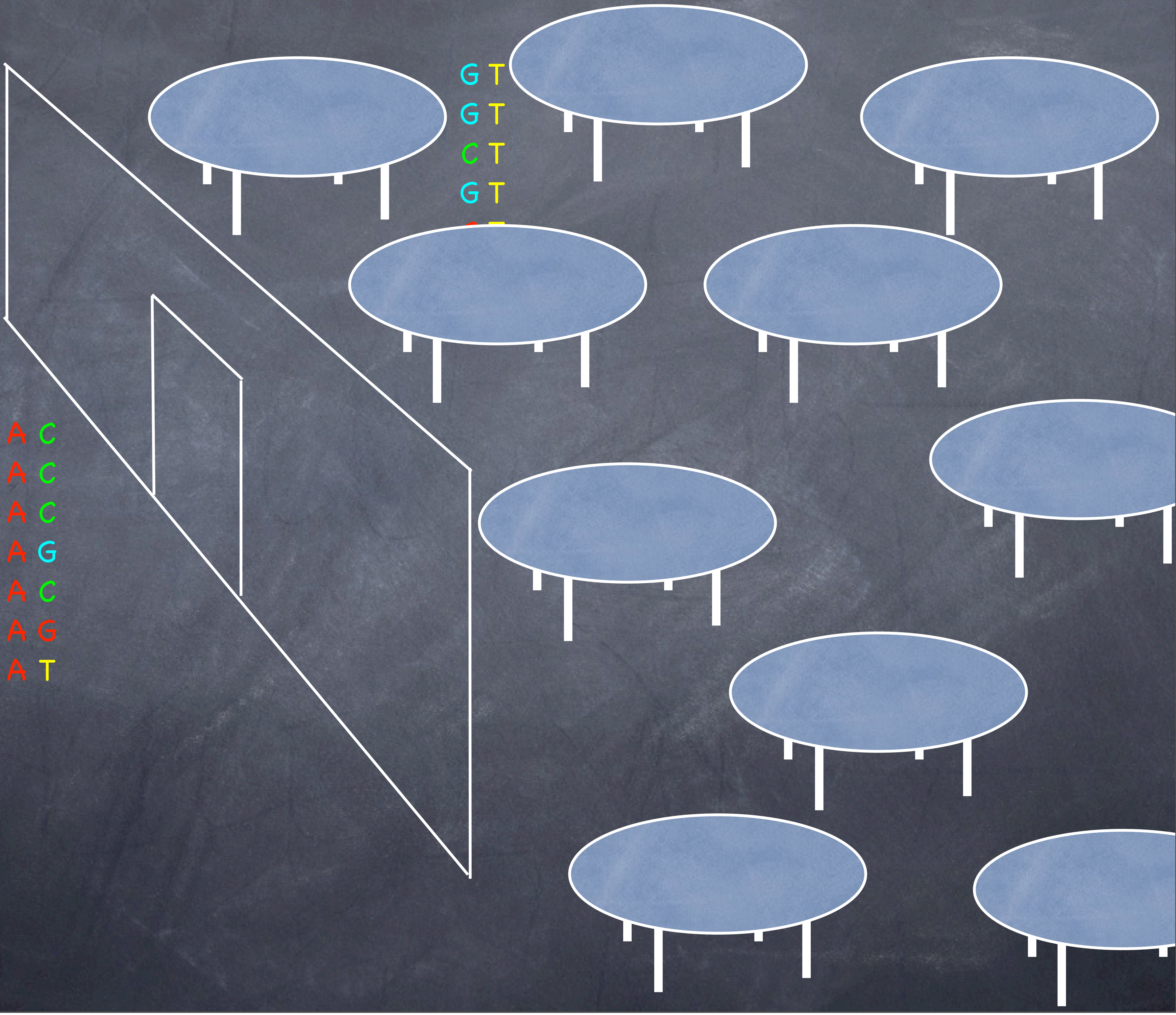




Human	A	C	T	G
Mouse	A	C	T	G
Rat	A	C	T	C
Elephant	A	G	T	G
Gopher	A	C	T	C
Chicken	A	G	T	A
Shark	A	T	T	T

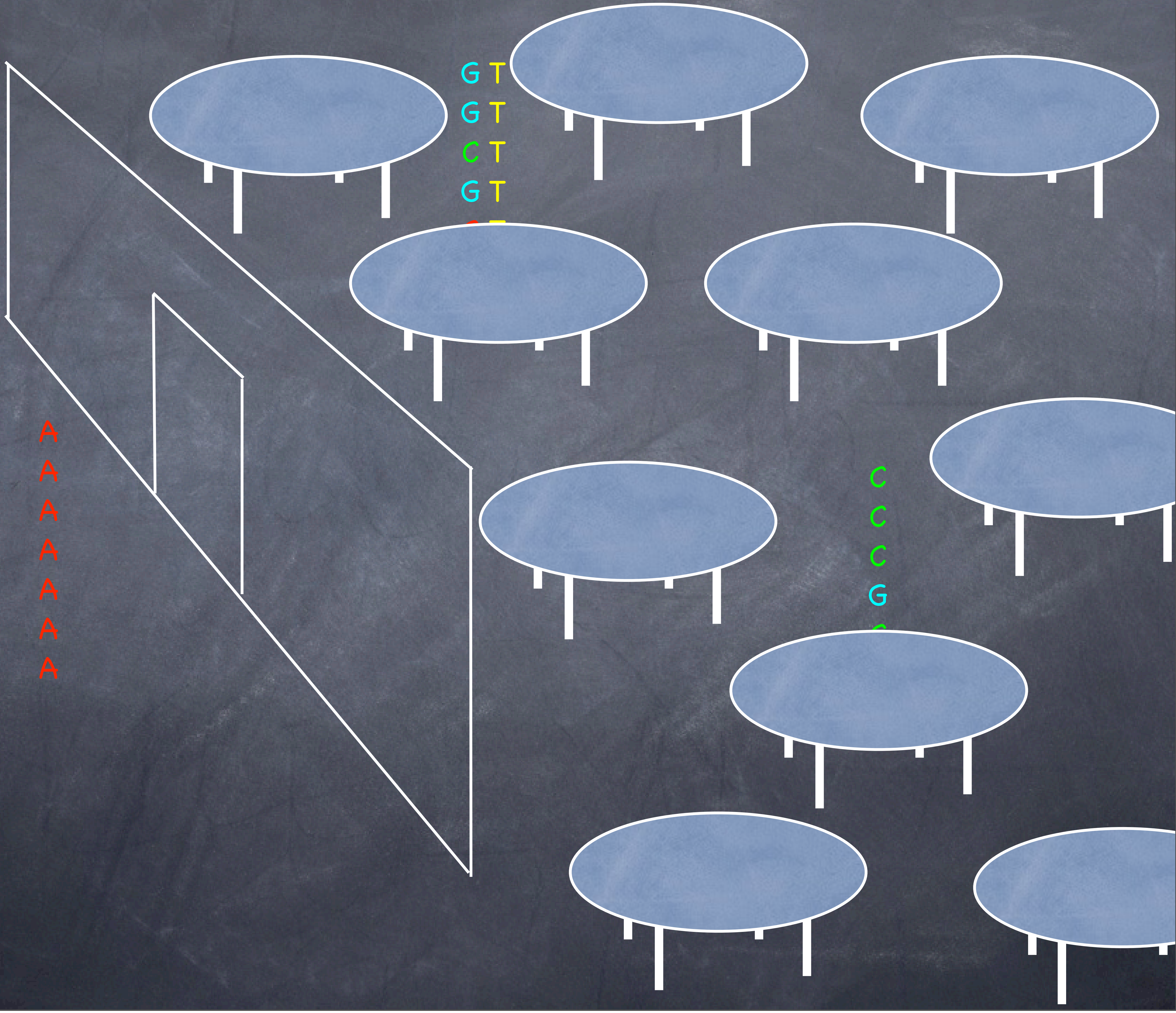
Human	A C T
Mouse	A C T
Rat	A C T
Elephant	A G T
Gopher	A C T
Chicken	A G T
Shark	A T T



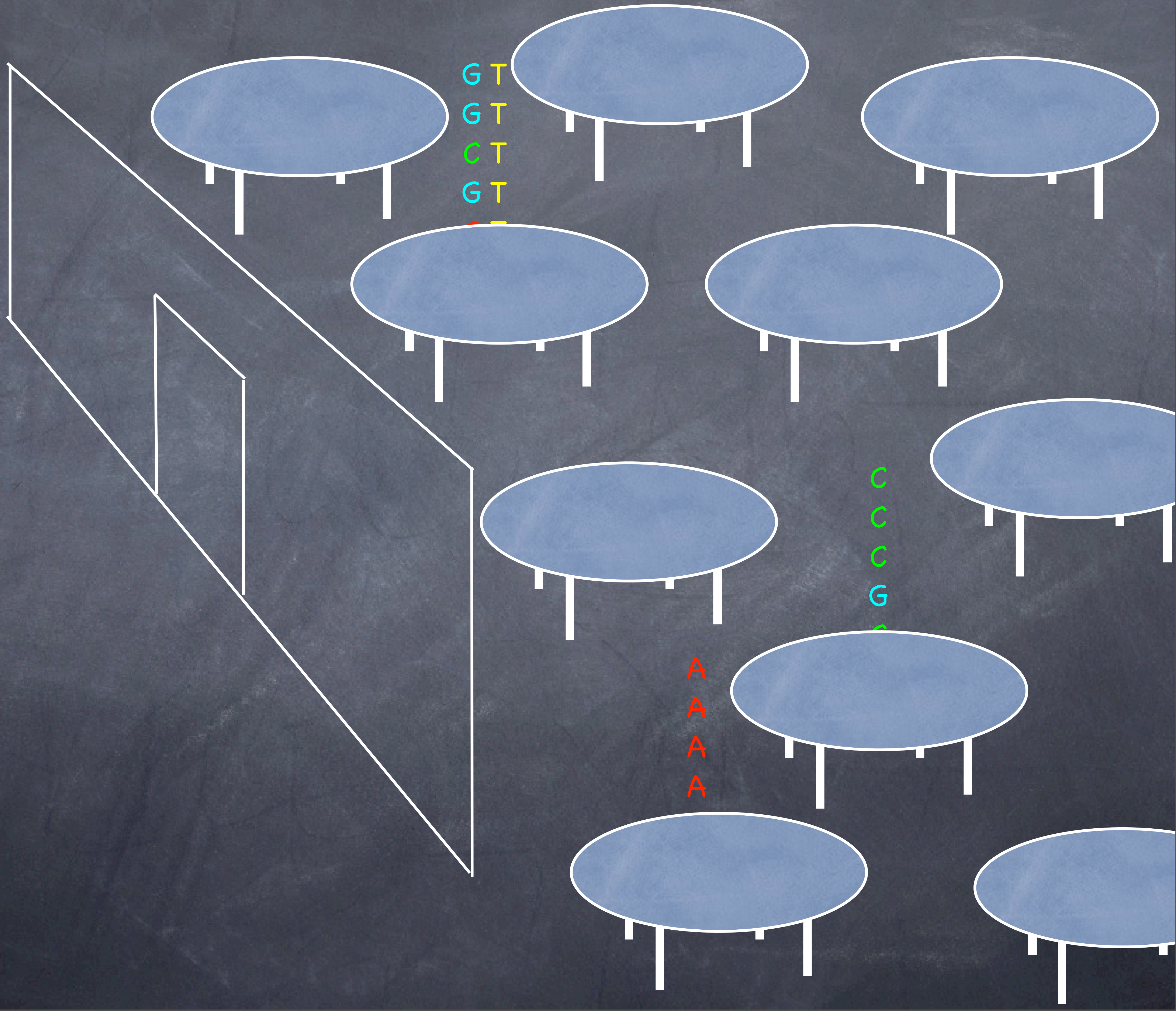


Human	A	C
Mouse	A	C
Rat	A	C
Elephant	A	G
Gopher	A	C
Chicken	A	G
Shark	A	T

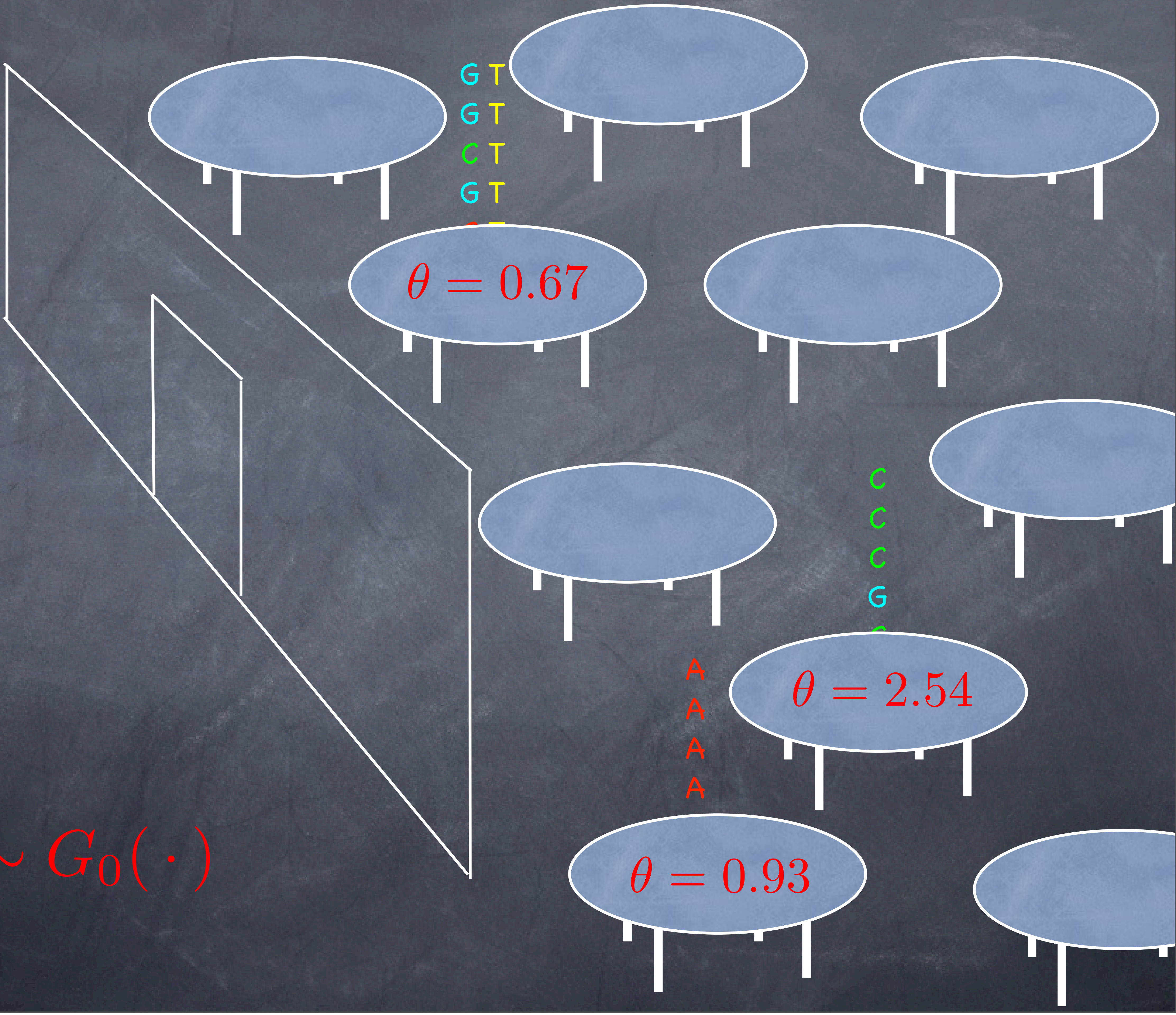
Human	A
Mouse	A
Rat	A
Elephant	A
Gopher	A
Chicken	A
Shark	A



Human
Mouse
Rat
Elephant
Gopher
Chicken
Shark



Human
Mouse
Rat
Elephant
Gopher
Chicken
Shark



$$\theta \sim G_0(\cdot)$$

Probability of the number of classes and the number of elements in each class

$$f(\mathbf{z}, k|\alpha, n) = \alpha^k \frac{\prod_{i=1}^{k}(\eta_i - 1)!}{\prod_{i=1}^{n}(\alpha + i - 1)}$$

Probability of the number of classes and the number of elements in each class

$$f(\mathbf{z}, k|\alpha, n) = \alpha^k \frac{\prod_{i=1}^{k} (\eta_i - 1)!}{\prod_{i=1}^{n} (\alpha + i - 1)}$$

Probability of the number of classes

$$f(k|\alpha, n) = \frac{{}_n a_k \alpha^k}{\prod_{i=1}^{n} (\alpha + i - 1)}$$

Probability of the number of classes and the number of elements in each class

$$f(\mathbf{z}, k|\alpha, n) = \alpha^k \frac{\prod_{i=1}^{k}(\eta_i - 1)!}{\prod_{i=1}^{n}(\alpha + i - 1)}$$

Probability of the number of classes

$$f(k|\alpha, n) = \frac{{}_n a_k \alpha^k}{\prod_{i=1}^{n}(\alpha + i - 1)}$$

Expected number of classes

$$E(k|\alpha, n) = \sum_{i=1}^{n} i\, f(k = i|\alpha, n) \approx \alpha \ln\left(1 + \frac{n}{\alpha}\right)$$

Probability of the number of classes and the number of elements in each class

$$f(\mathbf{z}, k|\alpha, n) = \alpha^k \frac{\prod_{i=1}^{k}(\eta_i - 1)!}{\prod_{i=1}^{n}(\alpha + i - 1)}$$

Probability of the number of classes

$$f(k|\alpha, n) = \frac{{}_n a_k \alpha^k}{\prod_{i=1}^{n}(\alpha + i - 1)}$$

Expected number of classes

$$E(k|\alpha, n) = \sum_{i=1}^{n} i\, f(k = i|\alpha, n) \approx \alpha \ln\left(1 + \frac{n}{\alpha}\right)$$

Probability that two data elements are grouped together in the same class

$$f(z_i = z_j|\alpha, n) = \frac{1}{1 + \alpha}$$

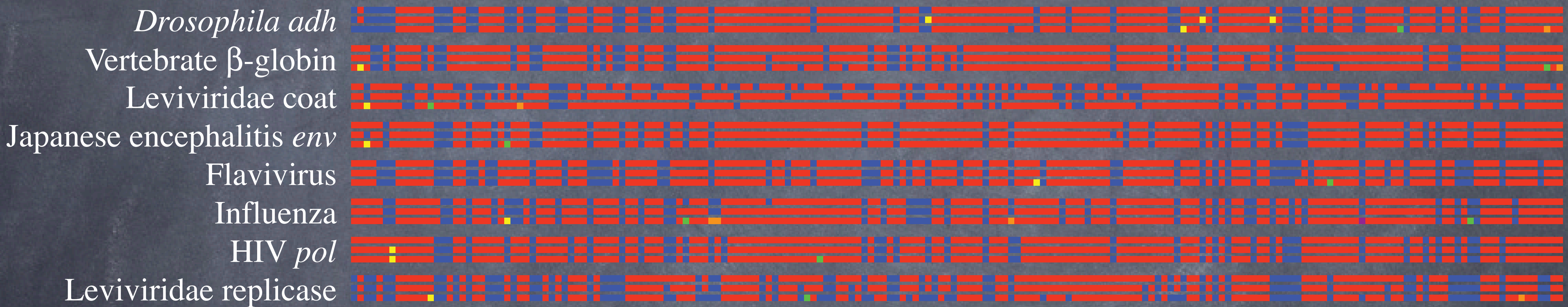
```
(1,2,1,3,3,1,3,1,1,1,1,3,2,3,3,2,4,1,1,1,1,3,3,2,1, ..... ,1,3,1,1,1,3,3,2,1,4,2,1,1,3,1,1,1,1,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,3,4,1,1, ..... ,2,2,1,1,1,2,2,1,1,1,1,1,1,2,1,3,1,1,1,3,2,1,1)
(1,2,2,3,3,2,3,2,2,2,2,3,4,3,3,2,2,4,2,4,4,4,2,4,4, ..... ,2,3,2,4,2,3,3,2,4,2,4,2,2,3,2,2,4,2,2,2,4,4,4)
(1,2,1,3,3,1,3,1,2,2,2,3,1,3,3,2,2,2,2,2,1,3,1,2,1, ..... ,2,3,2,1,2,3,3,2,1,2,1,1,2,3,1,2,1,2,2,2,3,1,2)
(1,1,1,2,2,3,2,1,1,3,1,2,1,2,2,1,1,1,1,3,3,1,2,3,1, ..... ,3,2,3,1,1,2,2,1,2,3,3,1,3,2,3,1,3,3,1,3,2,3,3)
(1,2,2,3,3,2,3,2,1,2,1,3,2,3,3,2,1,2,1,2,1,2,3,1,3, ..... ,2,3,1,2,2,3,3,2,2,2,1,1,2,3,1,1,1,1,1,1,2,2,2)
(1,1,2,3,3,1,3,1,1,2,2,2,2,3,3,1,1,1,1,2,2,1,2,1,2, ..... ,1,3,2,1,1,3,3,2,2,1,1,1,1,3,2,1,2,1,1,2,1,1,1)
(1,1,1,2,2,1,2,3,1,3,1,2,1,2,2,3,1,3,1,1,3,1,1,3,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,3,1,3,1,1,1,1,1)
(1,1,1,2,2,1,2,3,1,3,1,2,1,2,2,3,1,3,1,1,3,1,1,3,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,3,1,3,1,1,1,1,1)
(1,1,2,3,3,1,3,4,1,4,2,3,2,3,3,2,1,1,1,2,1,1,1,2,1, ..... ,1,3,2,1,1,3,3,2,1,2,2,1,1,3,5,2,1,1,2,1,2,2,1)
(1,2,1,3,3,1,3,1,2,2,2,3,2,3,3,2,2,2,1,1,2,2,1,2,2, ..... ,2,3,4,2,2,3,3,1,1,2,1,2,2,3,2,2,2,2,1,2,2,1,2)
(1,2,1,2,3,1,2,3,3,1,3,2,3,2,2,3,3,3,3,3,1,3,3,1,3, ..... ,3,2,1,3,3,2,2,3,3,3,3,3,3,2,1,3,1,3,3,1,1,3,1)
(1,2,1,2,3,1,2,3,3,3,1,2,3,2,2,2,1,3,1,3,3,3,3,1,1, ..... ,3,2,3,1,1,2,2,3,1,3,1,1,3,2,3,1,3,1,1,1,1,1,1)
(1,2,1,2,2,1,2,1,1,1,1,2,3,2,2,1,1,1,1,1,1,1,1,1,1, ..... ,1,2,1,1,3,2,2,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,3)
(1,2,1,2,2,1,1,1,1,1,1,2,1,2,2,3,1,1,1,1,1,3,2,3,1, ..... ,2,2,1,3,1,2,2,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,1)
(1,2,1,2,2,1,1,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,3,1,2,1,1,1,1,1,1,1,1,1)
(1,2,1,2,2,1,1,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,3,1,2,1,1,1,1,1,1,1,1,1)
(1,2,1,2,2,1,3,1,1,1,1,2,1,2,2,1,1,3,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,1,1,1,1,1,2,1,1)
(1,2,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,1)
(1,2,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,3,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,3,1,1,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,3,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,3,1,1,1,1,1,1,1)
(1,2,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,1,1,1,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,1,1,1,1,1,2,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,1,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,2,1,2,1,1,1,1,1,1,1,1,1)
(1,1,1,2,1,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,2,1,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,1,3,1,1,1,1,2,1,1)
(1,1,2,3,1,1,3,2,1,1,1,3,2,3,3,2,1,1,1,1,1,3,3,1,1, ..... ,1,3,1,1,2,3,3,1,1,1,1,2,2,3,1,1,1,1,1,1,3,1,1)
(1,1,2,3,3,1,3,1,2,1,1,3,1,3,3,2,1,1,1,1,1,2,1,1,3, ..... ,1,3,2,1,1,3,3,1,1,1,1,1,1,3,1,2,1,2,1,2,1,1,1)
(1,1,2,3,3,1,3,1,2,1,1,3,1,3,3,2,1,1,1,1,1,2,1,1,3, ..... ,1,3,2,1,1,3,3,1,1,1,1,1,1,3,1,2,1,2,1,2,1,1,1)
(1,1,1,2,2,1,2,3,1,3,1,2,1,2,2,1,1,1,3,3,1,3,2,1,1, ..... ,1,2,1,3,1,2,2,1,1,1,1,3,1,2,3,1,3,1,1,1,1,1,1)
(1,2,2,3,3,2,3,1,2,2,2,3,1,3,3,2,2,2,1,2,2,2,2,2,2, ..... ,2,3,2,2,2,3,3,2,3,2,2,2,2,3,2,2,2,2,2,2,2,2,2)
(1,1,1,2,2,3,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,1,1,1, ..... ,1,2,1,1,1,2,2,1,1,1,1,1,1,2,4,1,1,1,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,2,2,1,1,4,2,2,1,1,1,1,1,1,2,1,3,1,1,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,1,3,1,1,1,1,1,1,1, ..... ,1,2,1,1,1,2,2,1,2,1,1,1,1,2,1,1,1,1,1,1,1,1,1)
(1,1,2,3,3,1,3,4,1,1,1,3,1,3,3,4,1,1,1,4,1,1,3,1,1, ..... ,1,3,1,1,1,3,3,1,1,1,1,1,1,3,1,1,1,1,1,1,1,1,1)
(1,1,2,3,3,1,3,1,1,1,1,3,1,3,3,1,1,1,1,1,2,2,2,1,1, ..... ,1,3,1,2,1,3,3,1,1,1,2,1,1,3,2,1,1,2,1,1,1,1,1)
(1,1,2,3,3,1,3,1,1,1,1,3,1,3,3,1,1,1,1,1,2,2,2,1,1, ..... ,1,3,1,2,1,3,3,1,1,1,2,1,1,3,2,1,1,2,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,3,1,2,3,2,2,1,1,1,1,1,1,1,2,4,1, ..... ,1,2,1,3,3,2,2,1,1,1,1,3,1,2,1,1,1,1,1,1,1,1,1)
(1,1,1,2,2,1,2,1,1,3,1,2,3,2,2,1,1,1,1,1,1,1,2,4,1, ..... ,1,2,1,3,3,2,2,1,1,1,1,3,1,2,1,1,1,1,1,1,1,1,1)
(1,2,2,1,1,2,1,2,2,2,2,1,2,1,1,2,2,3,2,2,2,3,3,1,2, ..... ,2,1,2,2,2,2,1,2,2,2,2,2,3,1,2,3,2,2,2,2,1,3,2)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,3,1,1,1,3,1,2,1,1, ..... ,2,2,1,1,3,1,2,1,1,1,1,1,1,2,1,1,3,1,1,1,2,1,1)
(1,1,1,2,2,1,2,1,1,1,1,2,1,2,2,1,3,1,1,1,3,1,2,1,1, ..... ,2,2,1,1,3,1,2,1,1,1,1,1,1,2,1,1,3,1,1,1,2,1,1)
(1,1,2,3,3,1,3,1,1,1,2,3,2,3,3,1,1,1,2,3,2,1,3,1,1, ..... ,1,3,1,1,1,3,3,1,1,1,4,1,1,3,1,1,1,1,2,1,3,1,1)
(1,1,1,2,2,2,2,1,1,1,1,2,1,2,2,1,1,1,1,1,1,1,2,1,1, ..... ,1,2,1,1,1,2,2,1,2,1,3,1,2,2,3,3,1,1,1,1,2,1,1)
(1,2,2,1,2,1,1,2,2,2,2,1,2,1,1,1,2,2,2,2,2,2,2,2,2, ..... ,2,1,2,2,2,1,1,2,1,2,2,2,1,1,2,2,2,2,2,2,1,2,2)
```

N	1	2	3	4	5	6	7	8
1	🟥	🟦	🟦	🟥	🟦	🟩	🟩	🟩
2	🟥	🟦	🟥	🟥	🟦	🟩	🟩	🟩
3	🟥	🟥	🟥	🟥	🟦	🟩	🟨	🟩
4	🟥	🟥	🟥	🟥	🟦	🟩	🟩	🟩
5	🟥	🟦	🟦	🟥	🟥	🟩	🟨	🟩
6	🟥	🟦	🟥	🟥	🟦	🟦	🟩	🟩
7	🟥	🟥	🟥	🟥	🟦	🟩	🟨	🟨
8	🟥	🟦	🟥	🟥	🟦	🟩	🟩	🟩
9	🟥	🟦	🟥	🟥	🟦	🟩	🟩	🟩
10	🟥	🟦	🟥	🟥	🟦	🟦	🟩	🟩
$\bar{\mathbf{z}}$	🟥	🟦	🟥	🟥	🟦	🟩	🟩	🟩

$$\left.\begin{array}{c} 1 \\ \vdots \\ 10 \end{array}\right\}$$ MCMC Samples

$\bar{\mathbf{z}}$ Mean Partition

Gusfield, D. 2002. Partition-distance: a problem and class of perfect graphs arising in clustering. Information Processing Letters 82:159-164.


Drosophila adh
Vertebrate β-globin
Leviviridae coat
Japanese encephalitis *env*
Flavivirus
Influenza
HIV *pol*
Leviviridae replicase

Part 2 Summary

- The various models of DNA sequence evolution can be considered as partitions, resulting in a total of 203 time-reversible models
- Do the same thing with amino acid models
- Place a Dirichlet process prior probability distribution on the substitution rates

- One can use information from other data bases, in the form of a fixed amino acid model, but temper ones assumptions about rates for any particular data set.

- The variance parameter, χ, has a strong affect on inferences of substitution rates; there is not a lot of information about the 190 exchangeability parameters in the data sets we examined.

- It may be useful to summarize data bases of protein alignments as distributions on rates, instead of as fixed parameter estimates.

Detecting the Footprint of Natural Selection

Nonsynonymous Substitution: A nucleotide substitution that causes a change in the amino acid sequence

Synonymous Substitution: A nucleotide substitution that, because of the redundancy of the genetic code, does not cause a change in the amino acid sequence

$$\omega = d_N/d_S$$

$\omega < 1$	Purifying Selection
$\omega = 1$	Neutral(ish)
$\omega > 1$	Positive Selection

Model of Nielsen & Yang (1998)

$$q_{ij} = \begin{cases} \kappa\omega\pi_j & \text{:nonsynonymous transition} \\ \omega\pi_j & \text{:nonsynonymous transversion} \\ \kappa\pi_j & \text{:synonymous transition} \\ \pi_j & \text{:synonymous transversion} \\ 0 & :i \text{ and } j \text{ differ at two or more positions} \end{cases}$$

κ = transition/transversion rate ratio

ω = nonsynonymous/synonymous rate ratio

π_j = frequency of codon j

M3 Model of Yang et al. (2000)

Three omega classes with $\omega_1 < \omega_2 < \omega_3$

Class	d_N/d_S	Likelihood	Prior
1	ω_1	$\Pr[X \mid \omega_1]$	π_1
2	ω_2	$\Pr[X \mid \omega_2]$	π_2
3	ω_3	$\Pr[X \mid \omega_3]$	π_3

$\Pr[X] = \Pr[X \mid \omega_1]\,\pi_1 + \Pr[X \mid \omega_2]\,\pi_2 + \Pr[X \mid \omega_3]\,\pi_3$

Posterior probability of a site being in selection class 3 is:

$$\Pr[\omega_3 \mid X] = \frac{\Pr[X \mid \omega_3]\, \pi_3}{\Pr[X]}$$

$$\Pr[X] = \Pr[X \mid \omega_1]\, \pi_1 + \Pr[X \mid \omega_2]\, \pi_2 + \Pr[X \mid \omega_3]\, \pi_3$$

Empirical Bayes Approach

- Substitute maximum likelihood estimates for model parameters.
- Assuming the parameters take their maximum likelihood values, calculate the posterior probability of each site being under positive selection.
- PAML

Fully Bayesian Approach

- Specify priors on parameters of model
- Calculate joint posterior probability of all parameters
- Use MCMC to approximate posterior distribution
- MrBayes 3.0

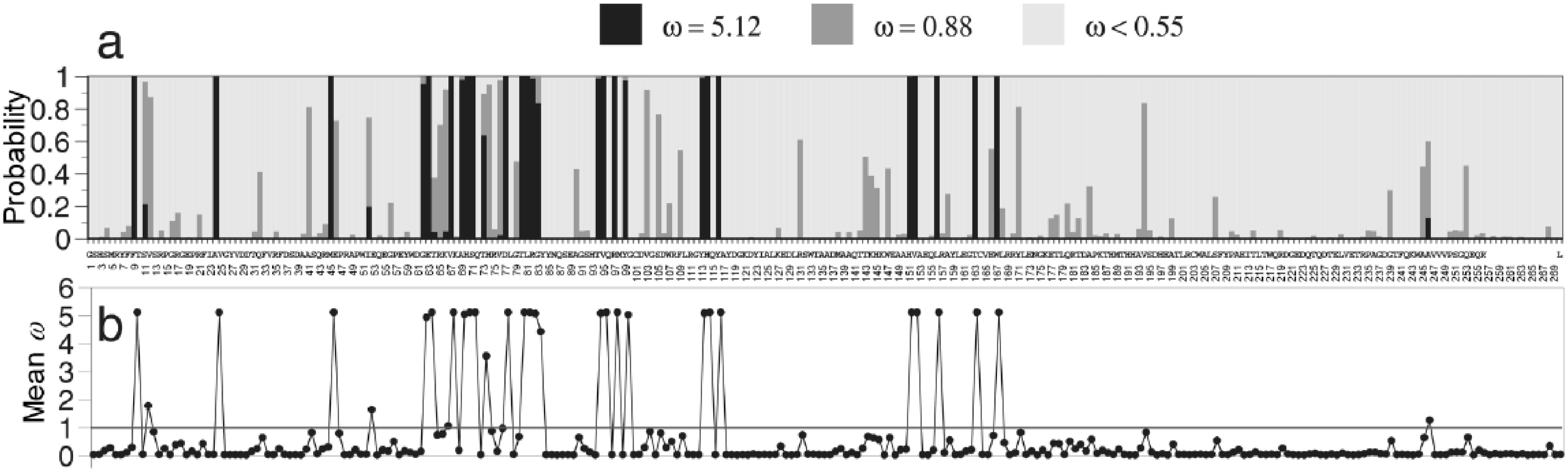


FIG. 1.—*a,* Posterior probabilities of site classes for sites along the MHC class I gene under the random-sites model M8 (beta & ω). Ten equal-probability categories are used to approximate the beta distribution (Yang et al. 2000), so that the model has 11 categories. The ω ratios are 0.00000, 0.00002, 0.00045, 0.00333, 0.01480, 0.04835, 0.12776, 0.28569, 0.54798, 0.88078, and 5.12163. Each of the first 10 categories has proportion 0.08998, where the last category has proportion 0.10019 (table 2). The first nine categories are collapsed into one category represented by $\omega < 0.55$. *b,* Posterior means of ω, calculated as the average of ω over the 11 site classes, weighted by the posterior probabilities. The amino acid sequence is from the structure file (Protein Data Bank file 1AKJ chain A; see fig. 2).

What is the appropriate distribution for d_N/d_S across sites?

Models, Models, Models...

Model	Description
M_0	Common ratio across sites
M_1	$\omega_1 = 0$, $\omega_2 = 1$
M_2	$\omega_1 = 0$, $\omega_2 = 1$, $0 < \omega_3 < \infty$
M_3	$0 < \omega_1 < \omega 2 < \ldots < \omega_k < \infty$
M_4	$\omega_1 = 0$, $\omega_2 = 1/3$, $\omega_3 = 2/3$, $\omega_4 = 1$, $\omega_5 = 3$
M_5	$\omega \sim \mathrm{Gamma}(\alpha, \beta)$
M_6	$\omega_1 \sim \mathrm{Beta}(\alpha_1, \beta_1)$, $\omega_2 \sim \mathrm{Gamma}(\alpha_2, \beta_2)$
M_7	$\omega \sim \mathrm{Beta}(\alpha, \beta)$
M_8	$\omega_1 \sim \mathrm{Beta}(\alpha, \beta)$, $0 < \omega_2 < \infty$
M_9	$\omega_1 \sim \mathrm{Beta}(\alpha_1, \beta_1)$, $\omega_2 \sim \mathrm{Gamma}(\alpha_2, \beta_2)$
M_{10}	$\omega_1 \sim \mathrm{Beta}(\alpha_1, \beta_1)$, $\omega_2 \sim \mathrm{Gamma}(\alpha_2, \beta_2) + 1$
M_{11}	$\omega_1 \sim \mathrm{Beta}(\alpha, \beta)$, $\omega_2 \sim \mathrm{Normal}(\mu, \sigma^2)_{\omega>1}$
M_{12}	$\omega_1 = 0$, $\omega_2 \sim \mathrm{Normal}(1, \sigma_1^2)_{\omega>1}$, $\omega_3 \sim \mathrm{Normal}(\mu, \sigma_2^2)_{\omega>1}$
M_{13}	$\omega_1 \sim \mathrm{Normal}(0, \sigma_1^2)_{\omega>1}$, $\omega_2 \sim \mathrm{Normal}(1, \sigma_2^2)_{\omega>1}$, $\omega_3 \sim \mathrm{Normal}(\mu, \sigma_2^2)_{\omega>1}$

Yang, Z., R. Nielsen, N. Goldman, and A. Pedersen. 2000. Codon substitution model for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Approach

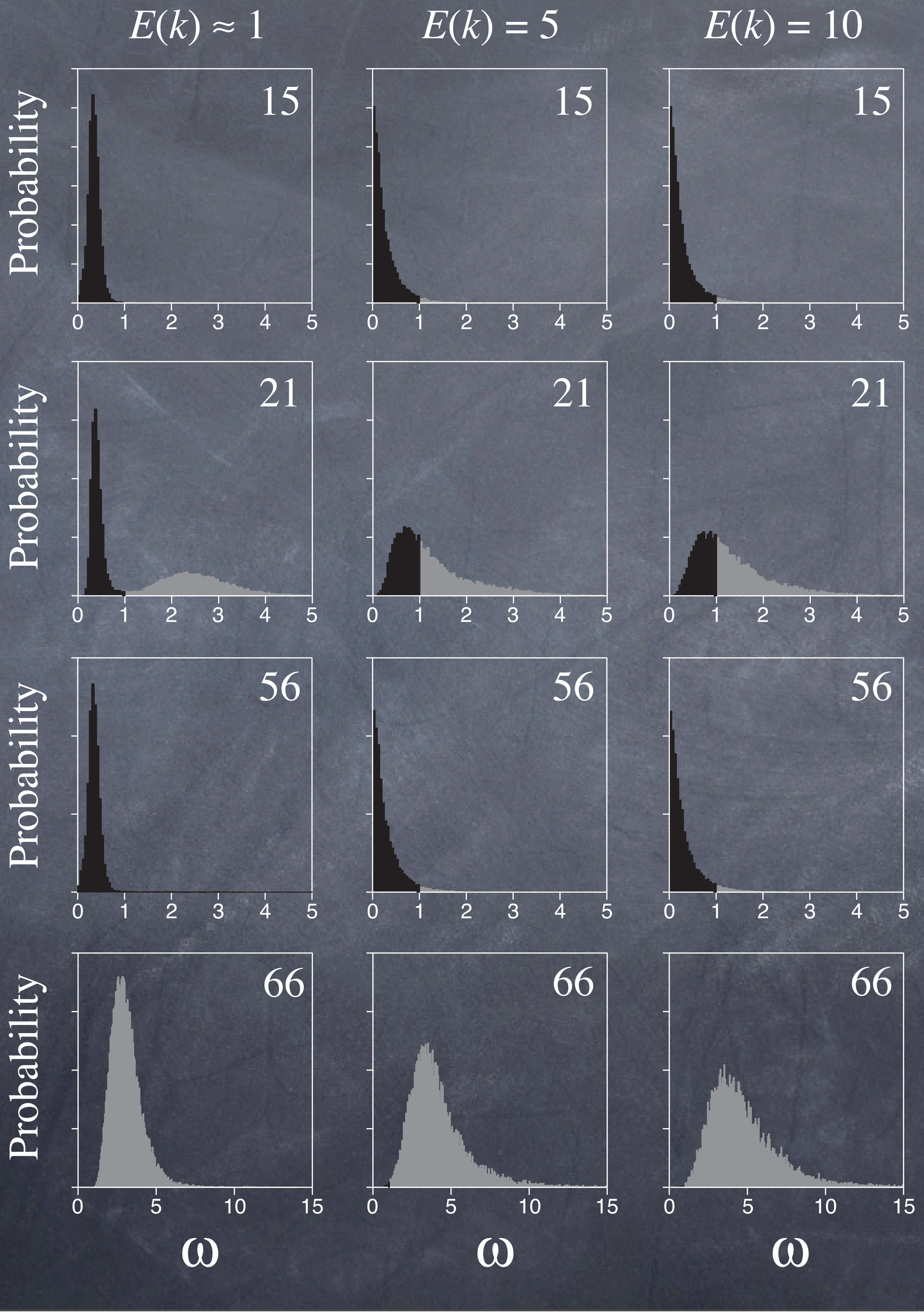
- Treat all parameters of the model as random variables with a prior probability distribution
- Inferences of positive selection are based upon the marginal posterior probability distribution for d_N/d_S at each site
- I use a Dirichlet process prior to describe how d_N/d_S varies across the sequence
- I use Markov chain Monte Carlo to approximate posterior probability distributions of parameters

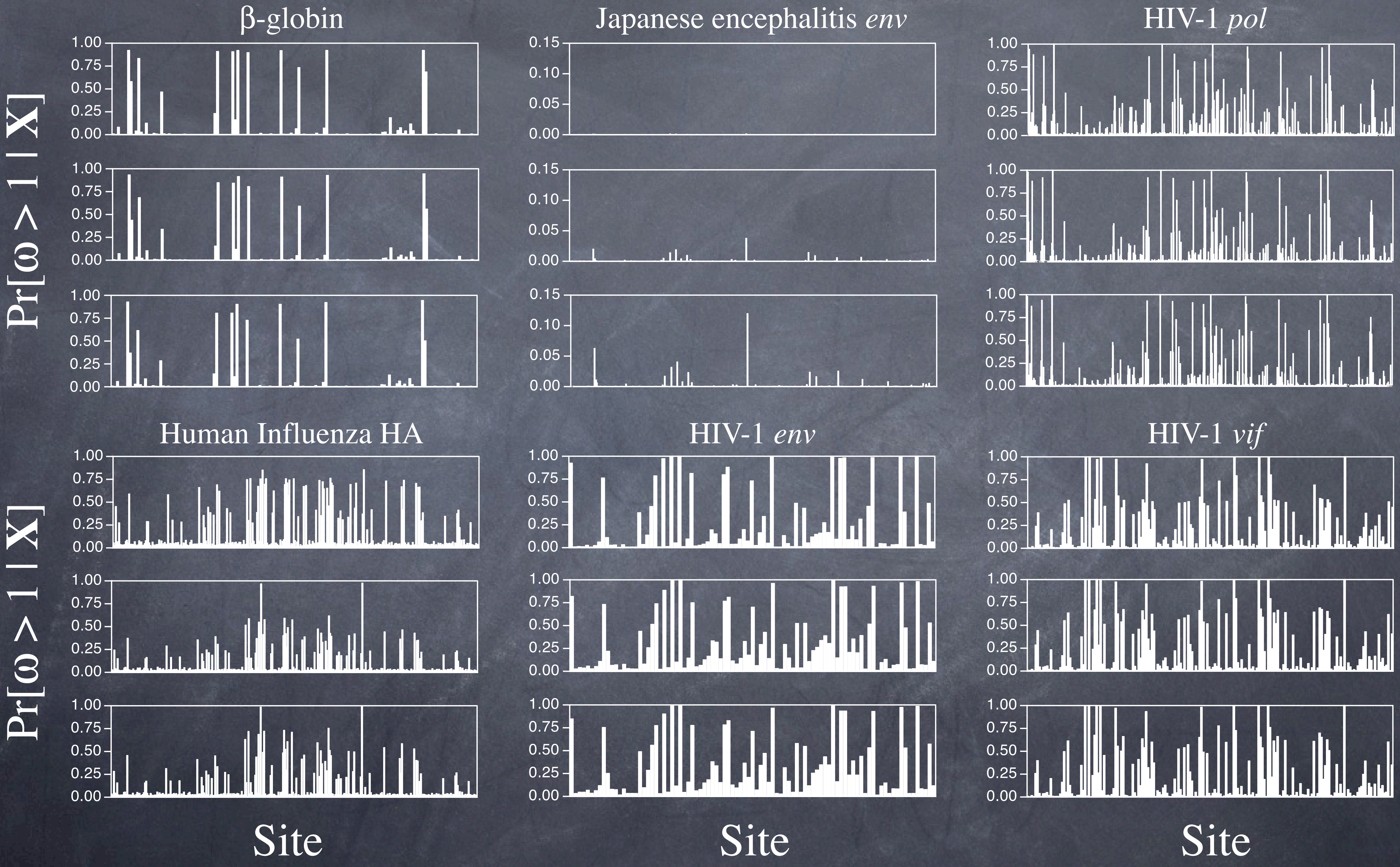
Parameter	Prior
Tree	All trees have equal prior probability
Branch Lengths	Branch lengths are exponential(10) r.v.s
Codon Frequencies	Flat Dirichlet distribution
Transition/Transversion Rate Ratio	Ratio of two identical exponential distributions
d_N/d_S Rate Ratio	Ratio of two identical exponential distributions
Category Information	Dirichlet process
Dirichlet process parameter	Fixed such that E(k) is small

Group	Gene	No. Taxa	No. Sites
Vertebrates	β-globin	17	144
Japanese Encephalitis	env	23	500
Human Influenza	HA1 domain of hemagglutinin	28	329
HIV-1	env	13	91
HIV-1	pol	23	947
HIV-1	vif	29	192

MCMC

- Chains run for 2,000,000 cycles
- Chains thinned, with samples taken every 100 update cycles
- Samples taken during the first 100,000 cycles discarded as the burn-in phase
- All analyses repeated
- Convergence assessed using the program Tracer


$E(k) \approx 1$
$E(k) = 5$
$E(k) = 10$
Probability
15
21
56
66
0
1
2
3
4
5
10
15
ω


β-globin
Japanese encephalitis *env*
HIV-1 *pol*
Human Influenza HA
HIV-1 *env*
HIV-1 *vif*
Pr[ω > 1 | X]
Pr[ω > 1 | X]
1.00
0.75
0.50
0.25
0.00
0.15
0.10
0.05
0.00
Site
Site
Site

Group	E(k)	Sites with probability greater than 0.95 of being under positive selection
Vertebrates	1	
	5	
	10	
Encephalitis	1	
	5	
	10	
Influenza	1	
	5	226, 135
	10	226, 135

Group	E(k)	Sites with probability greater than 0.95 of being under positive selection
HIV-1 env	1	28, 66, 26, 87, 51, 83, 76, 69, 68, 24
	5	28, 66, 26, 87, 83, 51
	10	28, 66, 26, 87, 83, 51
HIV-1 pol	1	67, 347, 478, 779, 568, 761
	5	67, 347, 779, 478, 3, 568
	10	67, 347, 779, 478, 3, 568
HIV-1 vif	1	33, 167, 33, 127, 39, 109, 122, 47, 92, 37
	5	33, 167, 127, 31, 37, 109, 39, 122, 92, 47, 63
	10	33, 127, 167, 31, 37, 109, 122, 39, 92, 47