BMI 206

Bayesian Statistics

Ryan D. Hernandez, PhD
Associate Professor
Bioengineering and Therapeutic Sciences
Institute for Human Genetics
Quantitative Biosciences Institute
ryan.hernandez@ucsf.edu

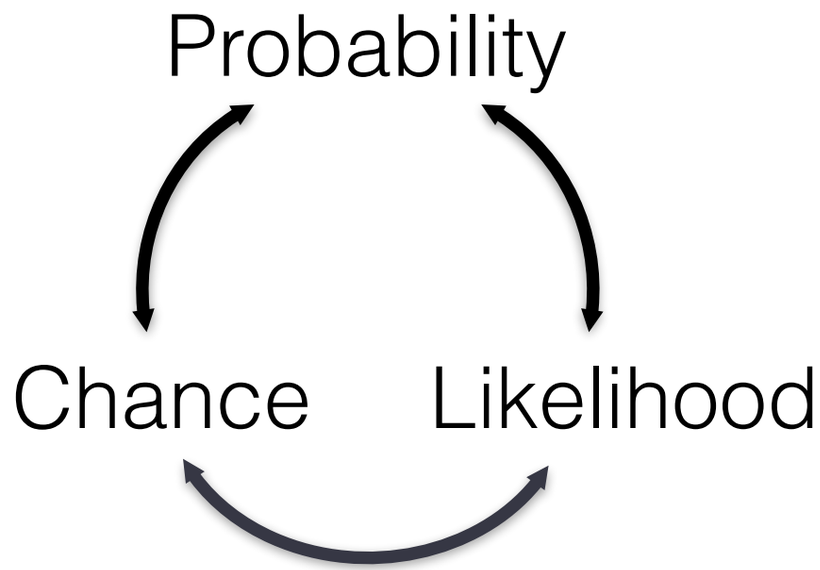
Big Data Era: Drinking from the fire hose

What are statistics?

- The collection, organization, analysis, interpretation, and presentation of data
- Biostatistics represents the application of statistics to biomedical research
- Three main branches of statistics
  - Descriptive statistics
  - Inferential statistics
  - Theoretical statistics

Background

- We are all aware of what the word “probability” means, here are some definitions:



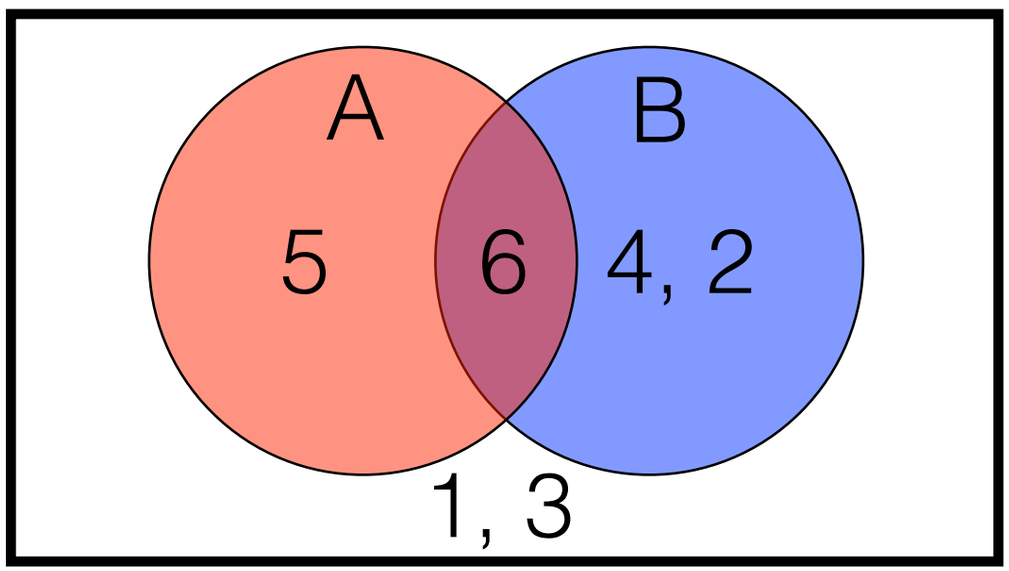
  - a priori
    - The basic notion in our heads: flipping a coin, rolling die
  - frequentist
    - Data-driven, based on observed frequency across experiments
  - subjective
    - Combination of the above

Outline

- There is no way to cover all of Bayesian statistics in a single lecture!!
- Basic probability
    - Addition and multiplication rules
    - Independence
    - Joint and conditional probabilities
    - Bayes' Rule
- Bayesian statistical modeling and inference
- Markov Chain Monte Carlo (MCMC)

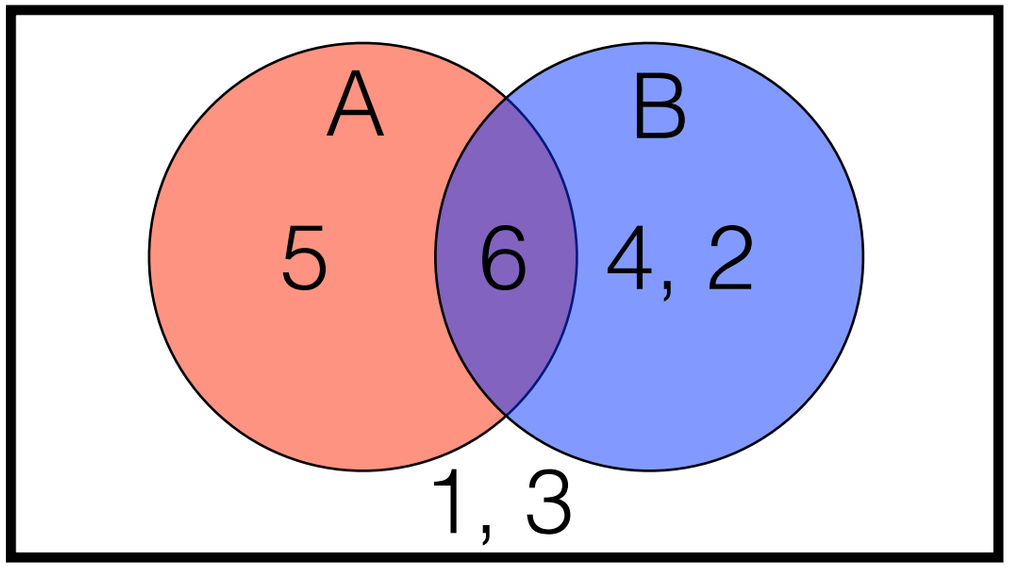
Simple Example

- Let's consider rolling a die.
- We are interested in two **events**:
  - **A**: we roll a number >4.
  - **B**: we roll an even number
- It is easy to calculate the probability of each event:
  - P(A) = 2/6 = 0.333
  - P(B) = 3/6 = 0.5



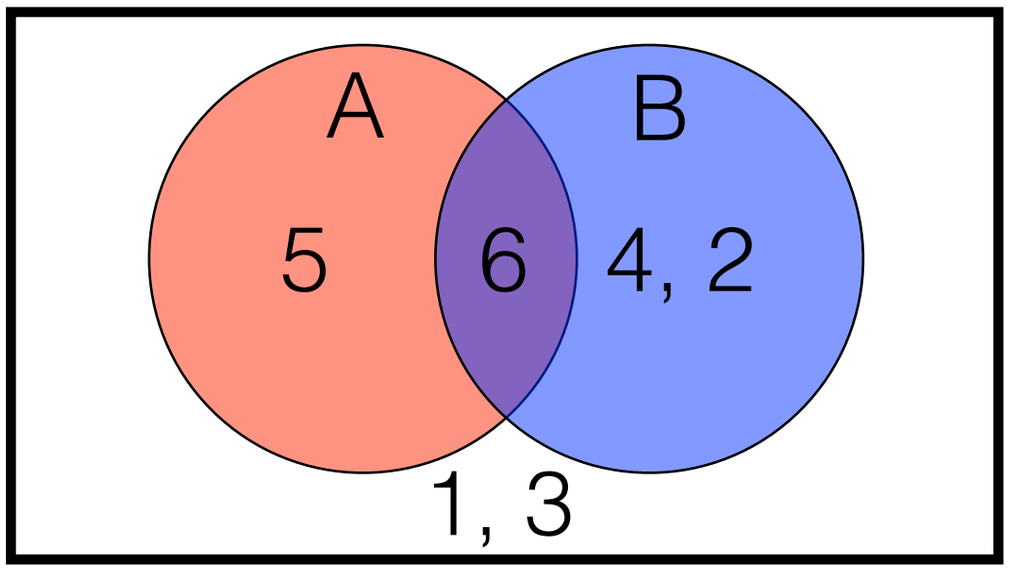
Simple Example

- What is the probability of A or B?
- Addition Rule:
- P(A or B) = P(A) + P(B) - P(A and B)



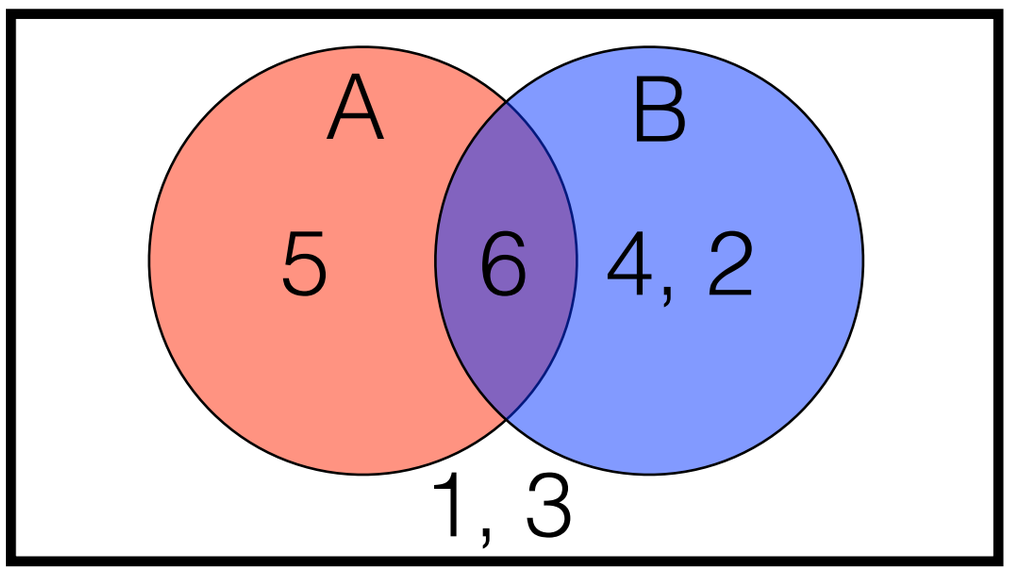
Independence and the Multiplicative Rule

- Let's introduce the idea of conditional probability.
- Consider the effect of one of these events on another: What is the probability that we will see an even number if we already know that we have thrown a number larger than 4?
- This can be written down as: P(B | A).



Independence & Conditional Probability

- The probability of event B given event A := Pr(B|A)
- It is a conditional probability since it depends on A having occurred.
  - If A occurred, then we must have thrown either a “5” or a “6”
  - The probability of an even number given that you have thrown a number larger than 4, is ½.
  - This is the conditional probability of B given A = Pr(B|A)
- The unconditional probability of B is Pr(B) = 3/6 = ½

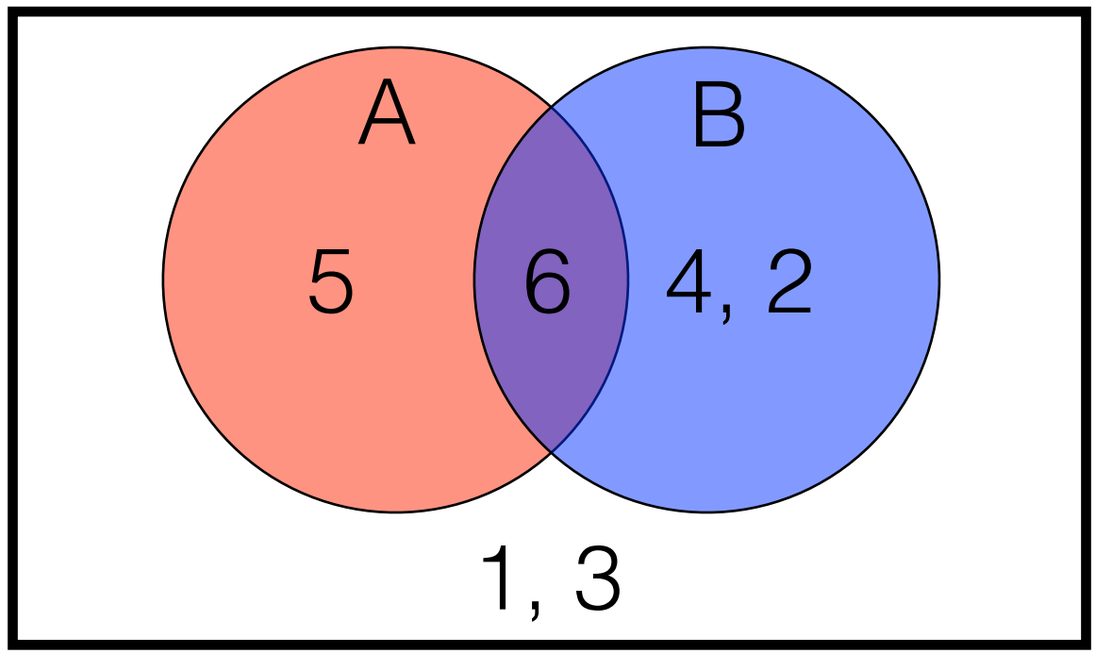


Independence (cont'd)

- The multiplication rule for probabilities
  - **Pr(A and B)** = Pr(A)×Pr(B|A) = Pr(B)×Pr(A|B)
  - IF two events A and B are **independent**, then:
    - Pr(A|B) = Pr(A) and Pr(B|A) = Pr(B)
    - Therefore: **Pr(A and B)** = Pr(A)×Pr(B)

Law of Total Probability

- What if we want to know the overall probability of an event?
  - What is the Pr(B)?
- **The Law of Total Probability:**
  - $\Pr(B) = \Pr(B|A)\times\Pr(A)+\Pr(B|A^c)\times\Pr(A^c)$
    $= 1/2\times1/3 + 1/2\times2/3$
    $= 1/6+2/6 = 1/2$
- Implication: Using just a little bit of algebra, we can now come up with explicit forms for conditional probability!



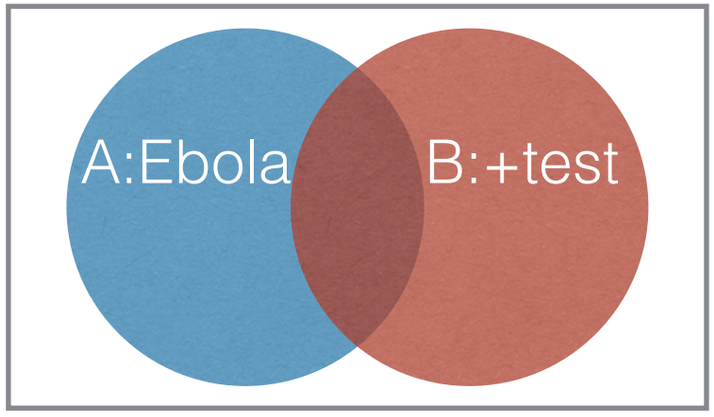
Hypothetical Example

- Suppose TSA imposes mandatory Ebola testing of all travelers on domestic flights in the USA.
- You go on a flight, and are tested for Ebola.
- Your test comes back positive...

Hypothetical Example

- What is the probability that you are actually infected with Ebola?
- Suppose the **sensitivity** of the test is high:
  - 99.9% of people infected with Ebola test positive.
- Suppose the **specificity** of the test is also high:
  - 99.9% of people not infected with Ebola test negative.
- Given your positive test and this information, should you be quarantined?!

Hypothetical Example



- Bayes' Rule comes to the rescue!!
- Let A be the event “Have Ebola”
- Let B be the event “Test Positive for Ebola”
- We want P(A | B) in terms we can easily quantify.
- Recall: P(A and B) = P(A|B)×P(B) = P(B|A)×P(A)

“Probability of having Ebola given positive test”

“Probability of having positive test given you are infected with Ebola”: Sensitivity

“Probability of being infected with Ebola”

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

“Probability of testing positive for Ebola”

$$P(B) = P(B|A)P(A) + P(B|A^C)P(A^C)$$

Hypothetical Example

- A = “Have Ebola infection”; B = “Test Positive for Ebola”
- In the USA, P(A) = Pr(have Ebola) ≈ 4/316,100,000=1.3e-8.
- Sensitivity: P(B|A) = 0.999; Specificity: $P(B^C|A^C)$=0.999

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

$$= \frac{0.999 \times 1.3e-8}{0.999 \times 1e-8 + 0.001 \times (1-1.3e-8)}$$

$$= 1.26e-5$$

- Thus, there is only a small chance you are actually infected, despite the high sensitivity and specificity!!

Hypothetical Example

- A = “Have HIV infection”; B = “Test Positive for HIV”
- In **Liberia**, P(A) = Pr(have Ebola) ≈ 4665/4,294,000=0.0011.
- Sensitivity: P(B|A) = 0.999; Specificity: P(BC|AC)=0.999

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

$$= \frac{0.999 \times 0.0011}{0.999 \times 0.0011 + 0.001 \times (1 - 0.0011)}$$

$$= 0.5207$$

- Thus, there is 52.07% chance you are actually infected, which is a much better test.
- The important difference is the **prior probability** of Ebola!

Bayesian Statistics

- In this class, you previously talked about **Hypothesis Testing** and **Parameter estimation**.
- These were largely discussed from the *frequentist* perspective (i.e., **Maximum Likelihood**)
- In that case, you wanted to calculate the probability of the observed data under a model:
  - P(Data | H_0)
- For parameter estimation, the goal was to find the parameter values that maximize this probability (i.e., **maximum likelihood estimate**):
  - $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ P(\mathrm{Data}|\theta)$
- In Bayesian Statistics, we turn this around!

Bayesian Statistics

- Main reasons to use Bayesian Statistics:
  - to account for previous knowledge about a parameter
  - logically update our knowledge about a parameter after we observe data
  - make formal probability statements about the parameter
  - to specify model assumptions and check model quality/sensitivity to these assumptions in a straightforward way.

Bayesian Statistics

- Bayesians treat unobserved data and unknown parameters in similar ways:
  - Each has a probability distribution!
- In a Bayesian model, we will need two things:
  - A **likelihood function** describing the probability of the data given the parameter values
  - A **prior distribution**, which describes the behavior of the parameter(s) **unconditional** on the data.
- The prior could reflect:
  - Uncertainty about a parameter that is actually fixed
  - The variety of values that a truly stochastic parameter could take.

Hemophilia Example

- In humans, males have an X and a Y chromosome, while females have two X chromosomes.
- Hemophilia is a genetic disease caused by a recessive X-linked mutation.
  - Much more common in males! (though still rare)
- Consider a woman with an affected brother.
- What is the probability she is a **carrier**?

Hemophilia Example

- We are told her father is not affected, so her mother must have been a carrier.
- The **prior** probability of being a carrier for this woman is 50%:
  - $P(\theta=1) = P(\theta=0) = 0.5$

θ: Carrier status (0=no, 1=yes)

Hemophilia Example

- The **prior** probability of being a carrier for this woman is 50%:
  - P(θ=1) = P(θ=0) = 0.5
- Suppose the woman has a son that is unaffected.
- Let y_1=1 and y_1=0 denote the case that the son is affected or unaffected.
- We can then write down two probabilities for the son being unaffected:
  - P(y_1=0 | θ=1) = 0.5
  - P(y_1=0 | θ=0) = 1
- We can now use Bayes' rule to combine the data with the prior probability to produce the **posterior probability**:

$$P(\theta=1|y_1) = \frac{P(y_1|\theta=1)P(\theta=1)}{P(y_1|\theta=1)P(\theta=1)+P(y_1|\theta=0)P(\theta=0)} = 0.5$$

Hemophilia Example

- What if the woman has another unaffected son?
- Let $y_2=0$ denote the case that the 2nd son is unaffected
- We can then write down two probabilities for both sons being unaffected:
  - $P(y_1=0, y_2=0 \mid \theta=1) = 0.5\times0.5 = 0.25$
  - $P(y_1=0, y_2=0 \mid \theta=0) = 1\times1 = 1$

This is not exactly what we want…

- Let $y=(y_1,y_2)$, then Bayes' rule gives use the **posterior probability**:

$$P(\theta=1|y) = \frac{P(y|\theta=1)P(\theta=1)}{P(y|\theta=1)P(\theta=1)+P(y|\theta=0)P(\theta=0)}$$

$$= \frac{0.25\times0.5}{0.s5\times0.5+1\times0.5} = 0.2$$

Hemophilia Example

- Intuitively, the more unaffected children the woman has, the less probable it is that she is a carrier.
- Bayes rule provides a formal mechanism for determining the extent of the correction!
- A key aspect of Bayesian analysis is the ease with which sequential analyses can be performed.
- Suppose the woman has a 3rd son, who is also unaffected.
- The entire calculation does not need to be redone:
  - Use the previous posterior probability as the new prior!

$$P(\theta = 1|y_1, y_2, y_3) = \frac{0.5 \times 0.2}{0.5 \times 0.2 + 1 \times 0.8} = 0.111$$

Setting up a Bayesian Model

- The key to Bayesian Inference is that the unknown parameter(s) θ are treated as random variables with prior distribution $f(\theta)$.
  - Sometimes in Bayesian world the prior is denoted $\pi(\theta)$.
- The prior distribution represents what we think we know about the parameters **before we observe any data**.
  - This is different from likelihood theory, where θ is treated as an unknown constant!
- Given some observed data X=x, we are interested in:

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}$$

Setting up a Bayesian Model: Asthma mortality

- Let's develop a Bayesian model for asthma mortality rates per year for a city with 200,000 people.
- It is found that 3 people died from asthma.
- This gives a crude estimate of 1.5 deaths per 100k people

Setting up a Bayesian Model: Asthma mortality

- We can do better!
- Let y be the number of deaths, and θ death rate per 100k
- We can model the likelihood: $\mathrm{P}(y|\theta) = \mathrm{Poisson}(2\theta)$.
- What about the prior?!
- In Western countries, typical asthma mortality rates are around 0.6 per 100k.
  - We can use this information!

Conjugate Priors

- When choosing a prior distribution there are 2 approaches:
  - Choose a distribution matching what you know
  - Choose a distribution that is convenient.
- Certain probability distributions have a very nice property:
  - When you multiply them together, terms combine to give you a nice functional form.
- This is a really nice property in Bayesian statistics, because we are always multiplying the likelihood function and a prior distribution.
- These are called **conjugate priors**.
- A few examples:

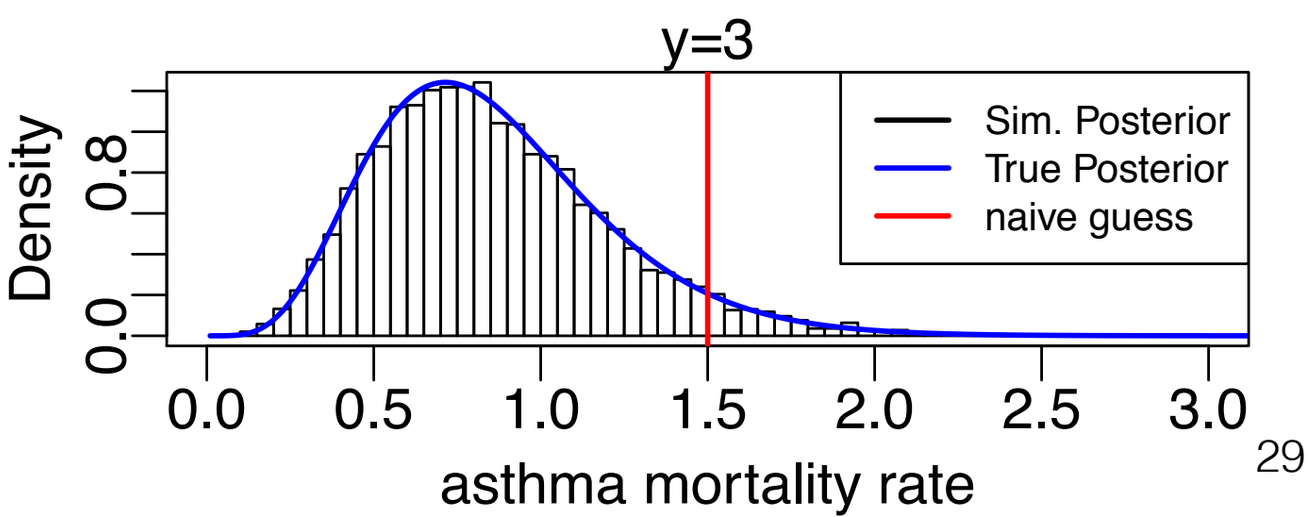
Likelihood	Prior
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Multinomial	Dirichlet
Exponential	Gamma
Normal	Normal

Setting up a Bayesian Model: Asthma mortality

- What about the prior?!
- In Western countries, typical asthma mortality rates are around 0.6 per 100k.
  - We can use this information! Pick a **conjugate prior**:
  - $\theta \sim \text{Gamma}(3,5)$

The parameters of the prior distribution are referred to as **hyper parameters**.

$$y \sim Poisson(2\theta)$$
$$\theta \sim Gamma(\alpha, \beta)$$
$$\theta|y \sim Gamma(\alpha + y, \beta + 2)$$



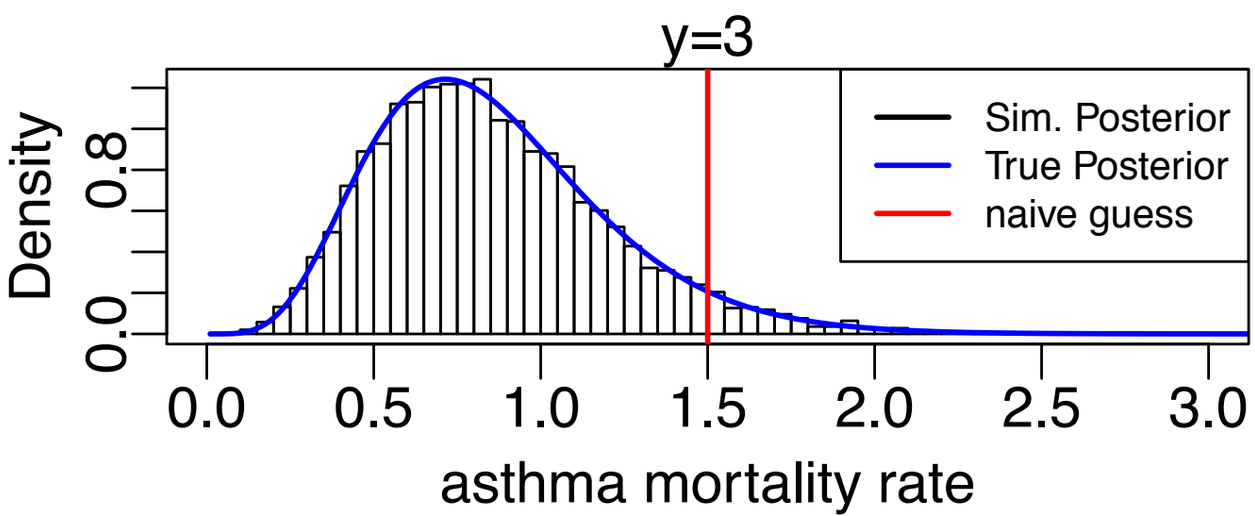
Setting up a Bayesian Model: Asthma mortality

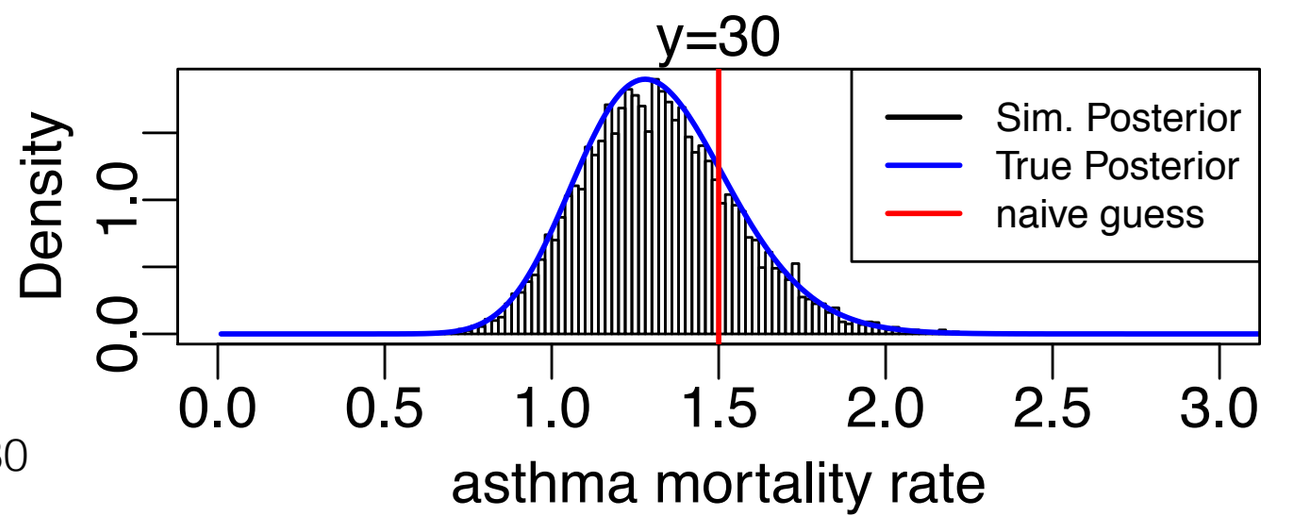
- What if we get more data?!
- Suppose we follow the same city for 10 years, and see a total of 30 deaths due to asthma.

$$y_i \sim Poisson(2\theta)$$

$$\theta \sim Gamma(\alpha, \beta)$$

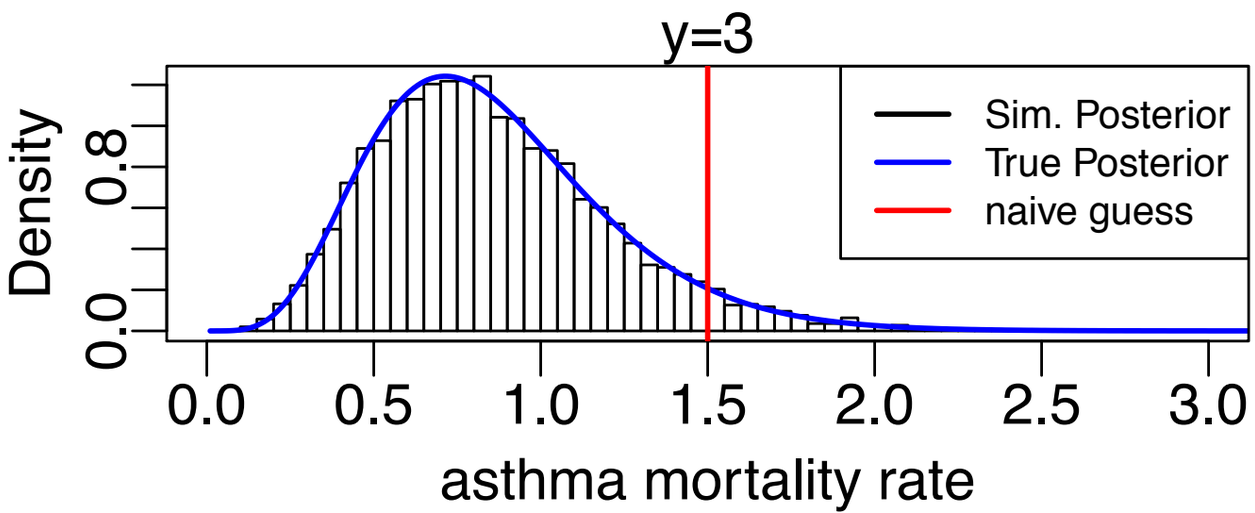
$$\theta|y \sim Gamma\left(\alpha + \sum_{i=1}^{10} y_i, \beta + 2n\right)$$

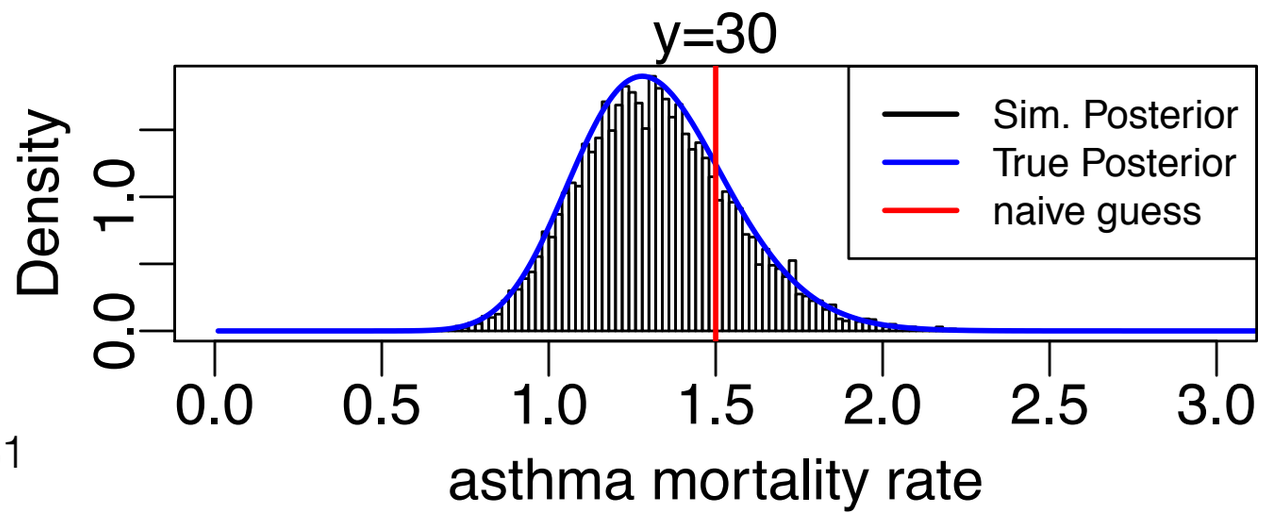




Setting up a Bayesian Model: Asthma mortality

- We are often interested in these summaries of the posterior distribution:
  - **Posterior mean** (“average value”)
  - **Posterior mode** (“most probable value”)
  - **High posterior density interval** (analog of confidence interval)





Two More Examples

- The examples we've examined so far had nice closed form solutions (thanks conjugate priors!).
- This isn't always the case!!
  - Bayesian clustering in population genetics
  - Bayesian protein structure prediction

- But first a bit more background.

Bayesian Inference Example

- Data: $Y_1, Y_2, ..., Y_n \sim N(\mu, \sigma^2)$, and a **non-informative prior**: $\pi_0(\mu, \sigma^2) \propto 1/\sigma^2$
- Joint posterior:

$$\pi(\mu, \sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{n/2+1} \times \exp\left\{-\frac{\sum(y_i - \mu)^2}{2\sigma^2}\right\}$$

- This is not the form of any standard probability distribution…
- Suppose we just wanted the posterior mean

Monte Carlo Integration

- The definition of the mean of a distribution is:

$$E(X) = \int x\pi(x)dx$$

- This can be generalized to any function h:

$$E(h(X)) = \int h(x)\pi(x)dx$$

- But this can sometimes be hard to evaluate!

Monte Carlo Integration

- Let's take a random sample $X^{(1)}, X^{(2)}, ..., X^{(N)} \sim \pi(x)$.
- If these are independent samples, then by the *law of large numbers*:

$$\bar{h}_N = \frac{1}{N} \sum_{i=1}^{N} h\left(X^{(i)}\right) \underset{\lim N \to \infty}{\longrightarrow} E(h(X))$$

- This is *Monte Carlo (MC) integration*.
- This holds for (almost) **any** proposal distribution $\pi(x)$.
- Of course, some proposal distributions are better than others...

Markov Chain Monte Carlo

- Back to our goal: determine the posterior distribution $\pi(\theta \mid y)$!
- Let's simplify to start out, and just write our **target distribution** as $\pi(x)$.
- Metropolis et al (1953) solved the problem for a symmetric proposal distribution, and Hastings (1970) generalized the solution to all distributions.
- This solution (the **Metropolis-Hastings algorithm**) is what was originally referred to as MCMC.

Metropolis Algorithm

- At each iteration t:

1. Sample $y \sim q(y|x^{(t)})$.

"Candidate" point

Symmetric "Proposal" distribution: $q(y|x) = q(x|y)$

2. Calculate acceptance ratio: $r = \frac{\pi(y)}{\pi\left(x^{(t)}\right)}$

3. Set $x^{(t+1)} = \begin{cases} y & \text{with probability } \min(1, r) \\ x^{(t)} & \text{else} \end{cases}$

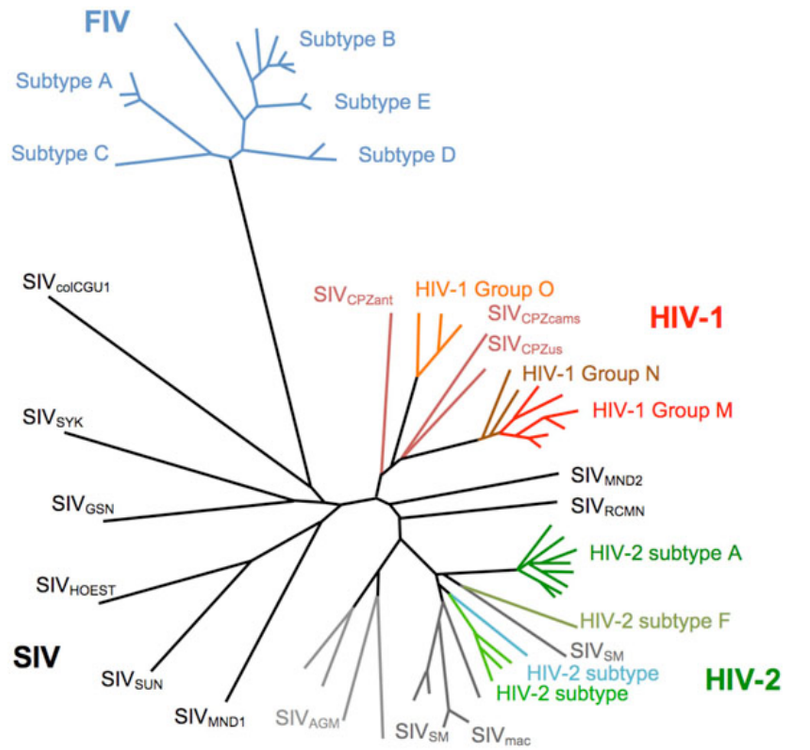
4. Repeat m times.

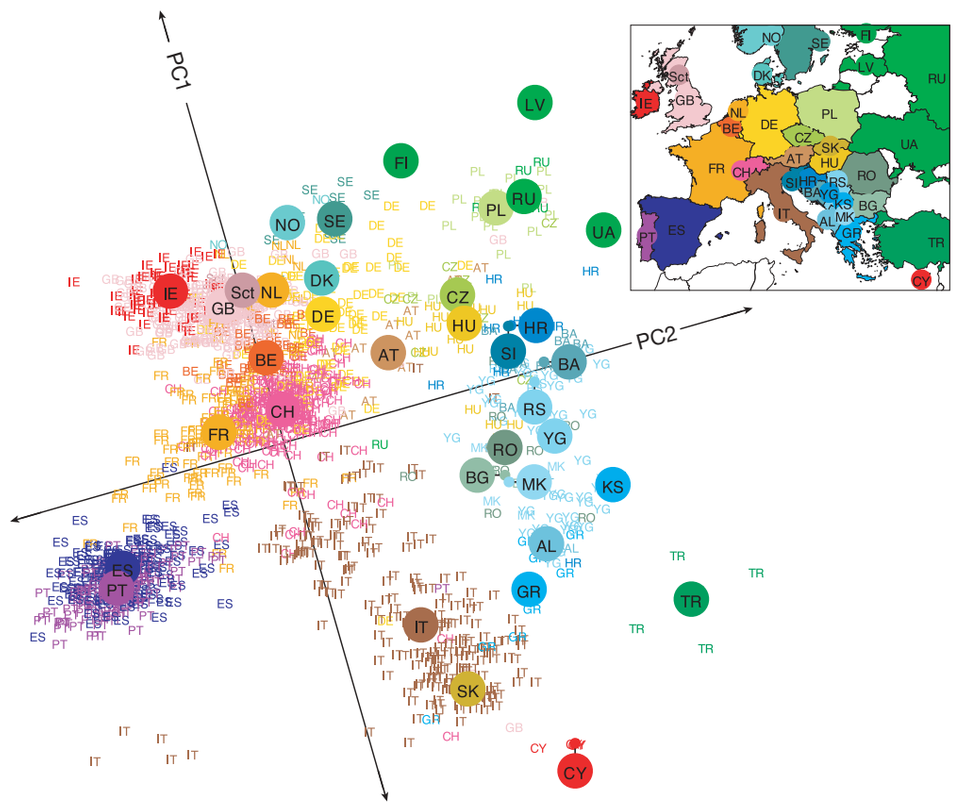
STRUCTURE Setup

- You start with a sample of individuals with genotypes
    - Are they from a single homogeneously mixing population?
    - If there is substructure in your data, how many populations contributed to your sample?
- We refer to this as the “K problem”...

Distance-Based methods

- There are many non-parametric models for identifying population structure in a dataset.
  - Neighbor Joining
    - hierarchical clustering
  - PCA
- What are some of the problems with distance-based methods?

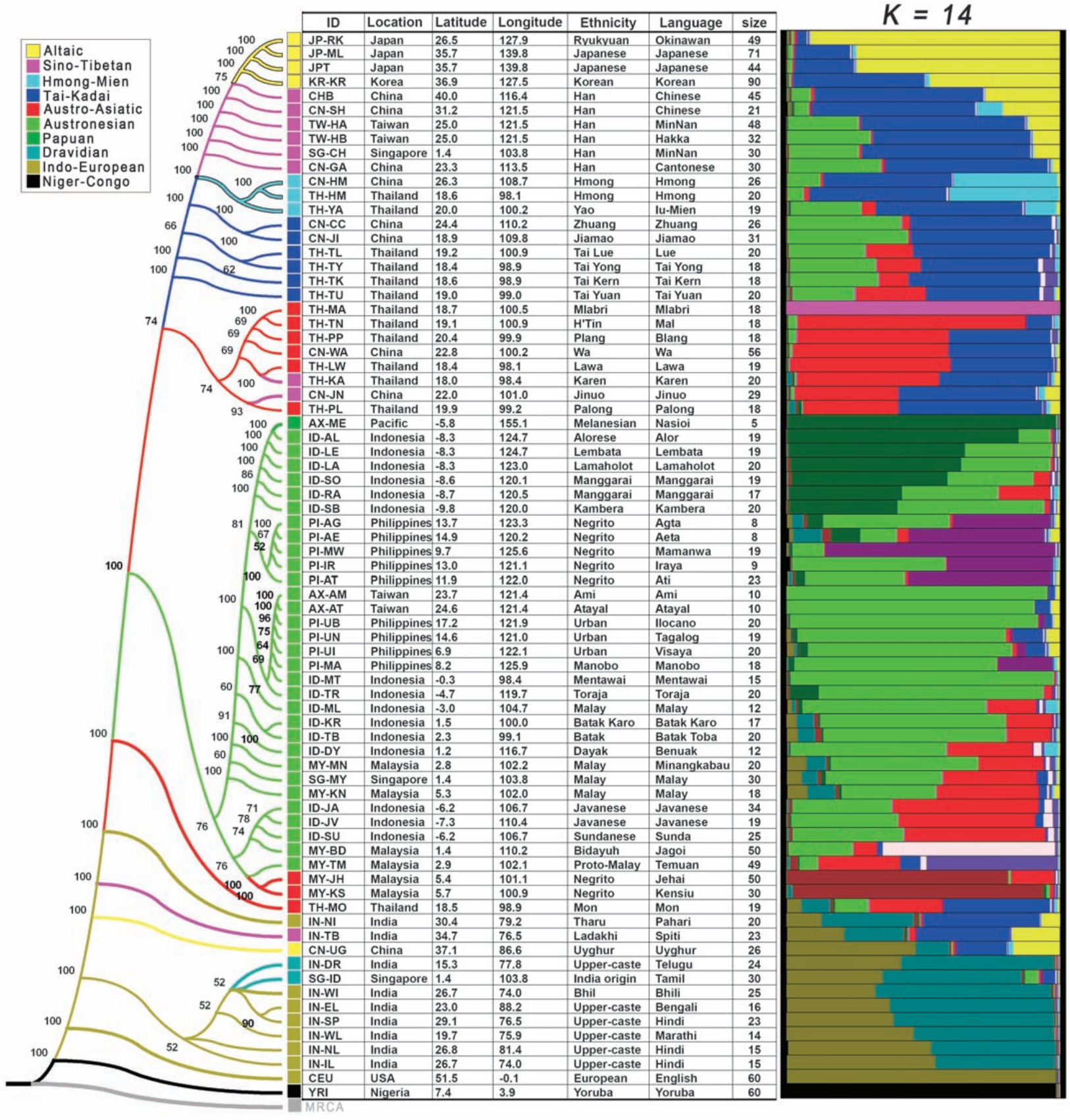




Model-Based Methods

- Assumes that individuals are random draws from some parametric model.
- Inference for the parameters corresponding to each cluster is then done jointly with inference for the cluster membership of each individual.

Population Structure of Asia



ID	Location	Latitude	Longitude	Ethnicity	Language	size
JP-RK	Japan	26.5	127.9	Ryukyuan	Okinawan	49
JP-ML	Japan	35.7	139.8	Japanese	Japanese	71
JPT	Japan	35.7	139.8	Japanese	Japanese	44
KR-KR	Korea	36.9	127.5	Korean	Korean	90
CHB	China	40.0	116.4	Han	Chinese	45
CN-SH	China	31.2	121.5	Han	Chinese	21
TW-HA	Taiwan	25.0	121.5	Han	MinNan	48
TW-HB	Taiwan	25.0	121.5	Han	Hakka	32
SG-CH	Singapore	1.4	103.8	Han	MinNan	30
CN-GA	China	23.3	113.5	Han	Cantonese	30
CN-HM	China	26.3	108.7	Hmong	Hmong	26
TH-HM	Thailand	18.6	98.1	Hmong	Hmong	20
TH-YA	Thailand	20.0	100.2	Yao	Iu-Mien	19
CN-CC	China	24.4	110.2	Zhuang	Zhuang	26
CN-JI	China	18.9	109.8	Jiamao	Jiamao	31
TH-TL	Thailand	19.2	100.9	Tai Lue	Lue	20
TH-TY	Thailand	18.4	98.9	Tai Yong	Tai Yong	18
TH-TK	Thailand	18.6	98.9	Tai Kern	Tai Kern	18
TH-TU	Thailand	19.0	99.0	Tai Yuan	Tai Yuan	20
TH-MA	Thailand	18.7	100.5	Mlabri	Mlabri	18
TH-TN	Thailand	19.1	100.9	H'Tin	Mal	18
TH-PP	Thailand	20.4	99.9	Plang	Blang	18
CN-WA	China	22.8	100.2	Wa	Wa	56
TH-LW	Thailand	18.4	98.1	Lawa	Lawa	19
TH-KA	Thailand	18.0	98.4	Karen	Karen	20
CN-JN	China	22.0	101.0	Jinuo	Jinuo	29
TH-PL	Thailand	19.9	99.2	Palong	Palong	18
AX-ME	Pacific	-5.8	155.1	Melanesian	Nasioi	5
ID-AL	Indonesia	-8.3	124.7	Alorese	Alor	19
ID-LE	Indonesia	-8.3	124.7	Lembata	Lembata	19
ID-LA	Indonesia	-8.3	123.0	Lamaholot	Lamaholot	20
ID-SO	Indonesia	-8.6	120.1	Manggarai	Manggarai	19
ID-RA	Indonesia	-8.7	120.5	Manggarai	Manggarai	17
ID-SB	Indonesia	-9.8	120.0	Kambera	Kambera	20
PI-AG	Philippines	13.7	123.3	Negrito	Agta	8
PI-AE	Philippines	14.9	120.2	Negrito	Aeta	8
PI-MW	Philippines	9.7	125.6	Negrito	Mamanwa	19
PI-IR	Philippines	13.0	121.1	Negrito	Iraya	9
PI-AT	Philippines	11.9	122.0	Negrito	Ati	23
AX-AM	Taiwan	23.7	121.4	Ami	Ami	10
AX-AT	Taiwan	24.6	121.4	Atayal	Atayal	10
PI-UB	Philippines	17.2	121.9	Urban	Ilocano	20
PI-UN	Philippines	14.6	121.0	Urban	Tagalog	19
PI-UI	Philippines	6.9	122.1	Urban	Visaya	20
PI-MA	Philippines	8.2	125.9	Manobo	Manobo	18
ID-MT	Indonesia	-0.3	98.4	Mentawai	Mentawai	15
ID-TR	Indonesia	-4.7	119.7	Toraja	Toraja	20
ID-ML	Indonesia	-3.0	104.7	Malay	Malay	12
ID-KR	Indonesia	1.5	100.0	Batak Karo	Batak Karo	17
ID-TB	Indonesia	2.3	99.1	Batak	Batak Toba	20
ID-DY	Indonesia	1.2	116.7	Dayak	Benuak	12
MY-MN	Malaysia	2.8	102.2	Malay	Minangkabau	20
SG-MY	Singapore	1.4	103.8	Malay	Malay	30
MY-KN	Malaysia	5.3	102.0	Malay	Malay	18
ID-JA	Indonesia	-6.2	106.7	Javanese	Javanese	34
ID-JV	Indonesia	-7.3	110.4	Javanese	Javanese	19
ID-SU	Indonesia	-6.2	106.7	Sundanese	Sunda	25
MY-BD	Malaysia	1.4	110.2	Bidayuh	Jagoi	50
MY-TM	Malaysia	2.9	102.1	Proto-Malay	Temuan	49
MY-JH	Malaysia	5.4	101.1	Negrito	Jehai	50
MY-KS	Malaysia	5.7	100.9	Negrito	Kensiu	30
TH-MO	Thailand	18.5	98.9	Mon	Mon	19
IN-NI	India	30.4	79.2	Tharu	Pahari	20
IN-TB	India	34.7	76.5	Ladakhi	Spiti	23
CN-UG	China	37.1	86.6	Uyghur	Uyghur	26
IN-DR	India	15.3	77.8	Upper-caste	Telugu	24
SG-ID	Singapore	1.4	103.8	India origin	Tamil	30
IN-WI	India	26.7	74.0	Bhil	Bhili	25
IN-EL	India	23.0	88.2	Upper-caste	Bengali	16
IN-SP	India	29.1	76.5	Upper-caste	Hindi	23
IN-WL	India	19.7	75.9	Upper-caste	Marathi	14
IN-NL	India	26.8	81.4	Upper-caste	Hindi	15
IN-IL	India	26.7	74.0	Upper-caste	Hindi	15
CEU	USA	51.5	-0.1	European	English	60
YRI	Nigeria	7.4	3.9	Yoruba	Yoruba	60

HUGO Pan-Asian SNP Consortium. Science (2009).

STRUCTURE Setup

- Detecting population structure in a sample is really a missing data problem!
- If we knew:
    - The frequency of every allele in each ancestral population
    - The ancestral population that each person derives from
- Then we could write down a simple likelihood function for our data
    - Assuming sites are independent, we could just multiply the frequencies of all alleles in the ancestral population.
- If we didn't know the ancestral population, we could iterate through all populations, and choose the population with maximum likelihood!

STRUCTURE Setup

- Of course, we don't know either of those key elements!
- In fact, because of genetic drift, knowing the ancestral population may not even be that helpful.
- MCMC to the rescue!

STRUCTURE Setup

- X = genotypes of the sampled individuals
- Z = the (unknown) populations of origin of individuals
- P = the (unknown) allele frequencies in all populations

- Assumptions:
  - Hardy-Weinberg Equilibrium within populations (but not necessarily between populations)
  - Complete linkage equilibrium between loci within populations (i.e., independence).

STRUCTURE Setup

- Goal: $\Pr(Z, P \mid X)$

  - $\propto \Pr(\mathrm{X} \mid \mathrm{Z}, \mathrm{P}) \Pr(Z, P)$

  - $= \Pr(\mathrm{X} \mid \mathrm{Z}, \mathrm{P}) \Pr(Z) \Pr(P)$

- $\Pr(Z, P \mid X) \propto \Pr(X \mid Z, P) \Pr(Z) \Pr(P)$

Joint probability of pop membership & their freqs given obs. genotypes

STRUCTURE Setup

- Our goal is to construct a Markov chain $\theta^{(0)}, \theta^{(1)}, \ldots$ with stationary distribution $\pi(\theta) = \Pr(Z, P, Q \mid X)$.
- This means that for m very large, $\theta^{(m)} \sim \pi(\theta)$
- And for c very large, $\theta^{(m)}, \theta^{(m+c)}, \theta^{(m+2c)}, \ldots$ are independent draws from $\pi(\theta)$.
- m is referred to as the ***burn-in***
- c is referred to as the ***thinning interval.***

STRUCTURE Setup

- Suppose we have N diploid individuals genotyped at L loci.
- Each individual comes from one of K populations.

$(x_l^{(i,1)}, x_l^{(i,2)})$ = genotype of the ith individual at the lth locus, where $i = 1, 2, \ldots, N$ and $l = 1, 2, \ldots, L$;
$z^{(i)}$ = population from which individual i originated;
p_{klj} = frequency of allele j at locus l in population k, where $k = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, J_l$,

- If we knew which population each individual came from, then we could write:

$$\Pr(x_l^{(i,a)} = j | Z, P) = p_{z(i)lj}$$

  - But we don't...

STRUCTURE Setup

- $\Pr(Z, P \mid X) \propto \Pr(X \mid Z, P) \Pr(Z) \Pr(P)$
- We don't know anything about the population of origin.
    - What's a good prior distribution to use?

$$\Pr(z^{(i)} = k) = 1/K,$$

- We don't know anything about the population allele frequencies.
    - What's a good prior distribution to use?

$$p_{kl\cdot} \sim \mathcal{D}(\lambda_1, \lambda_2, \ldots, \lambda_{J_l}),$$

    - Dirichlet distribution! (i.e., generalized Beta)

STRUCTURE Setup

- **Step 1**: Pretend we know population membership of each individual, and sample population frequencies.
- **Step 2**: Pretend we know population allele frequencies, and sample population membership.

- This is a special type of MCMC called a **Gibbs sampler.**

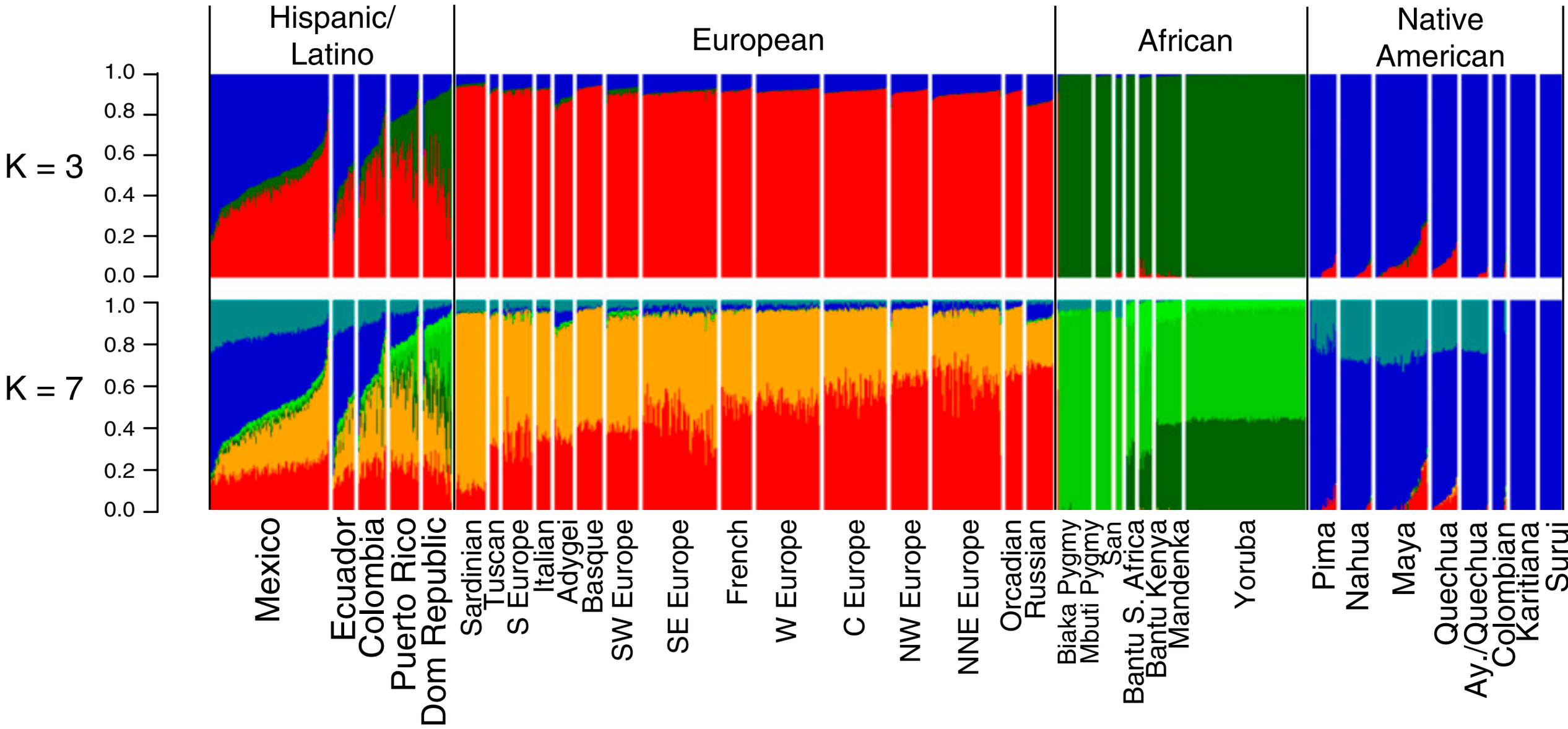
STRUCTURE Setup

- **Step 1**: Sample $P^{(m)}$ from $\Pr(P \mid X, Z^{(m-1)})$
  - If $\Pr(P) \sim \text{Dir}(\lambda_1, ..., \lambda_j)$ and $\Pr(X \mid Z, P)$ = allele frequencies, then
    - $\Pr(P \mid X, Z^{(m-1)}) \sim \text{Dir}(\lambda_1 + n_{kl1}, ..., \lambda_j + n_{klJ})$

- **Step 2**: Sample $Z^{(m)}$ from $\Pr(Z \mid X, P^{(m)})$
  - Key insight is that

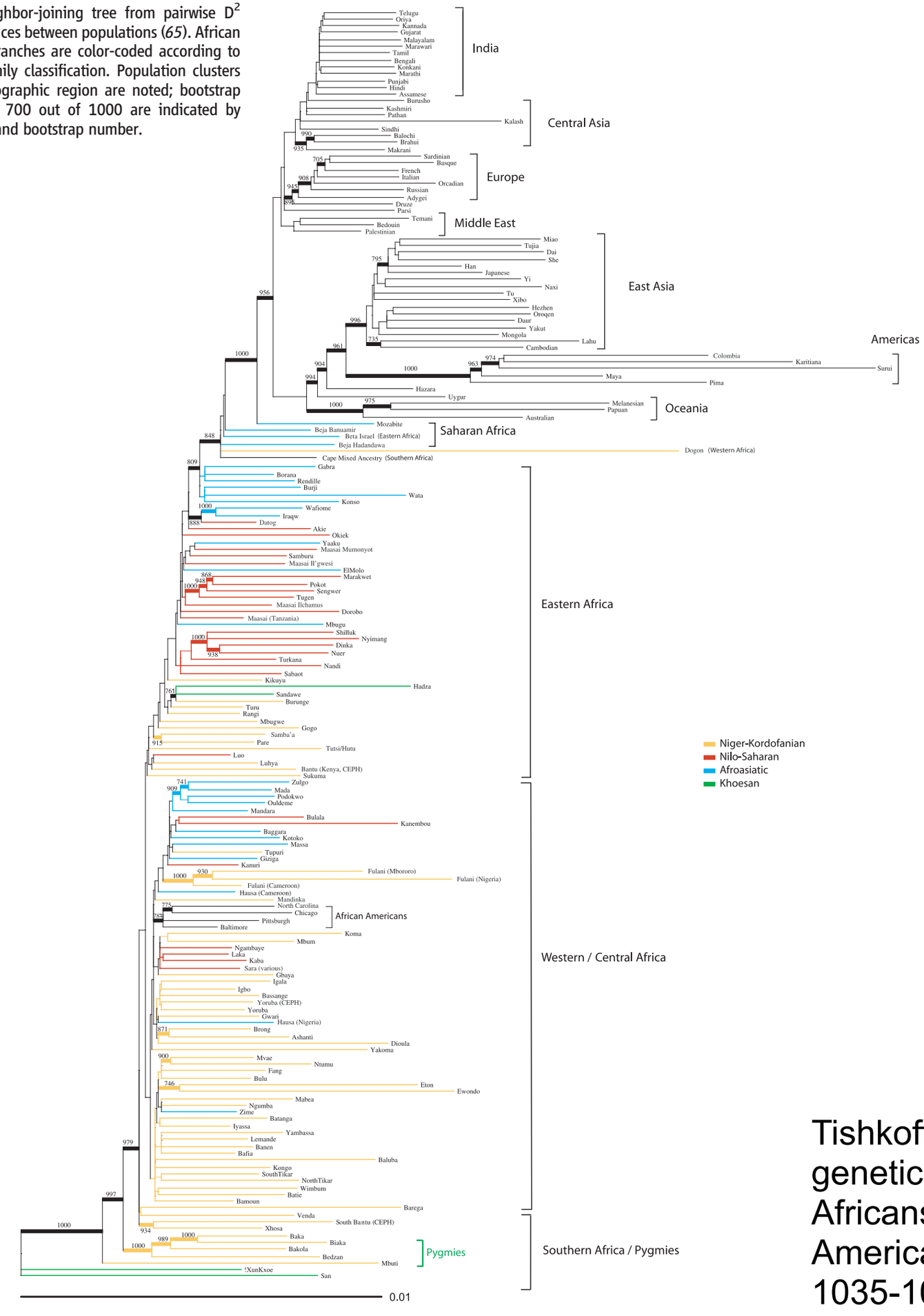
$$\Pr(z^{(i)} = k | X, P) = \frac{\Pr(x^{(i)} | P, z^{(i)} = k)}{\sum_{k'=1}^{K} \Pr(x^{(i)} | P, z^{(i)} = k')}$$

Applications



Bryc, K. et al. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. PNAS (2010).

Fig. 1. Neighbor-joining tree from pairwise D^2 genetic distances between populations (*65*). African population branches are color-coded according to language family classification. Population clusters by major geographic region are noted; bootstrap values above 700 out of 1000 are indicated by thicker lines and bootstrap number.



Tishkoff, S. A. et al. The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044 (2009).

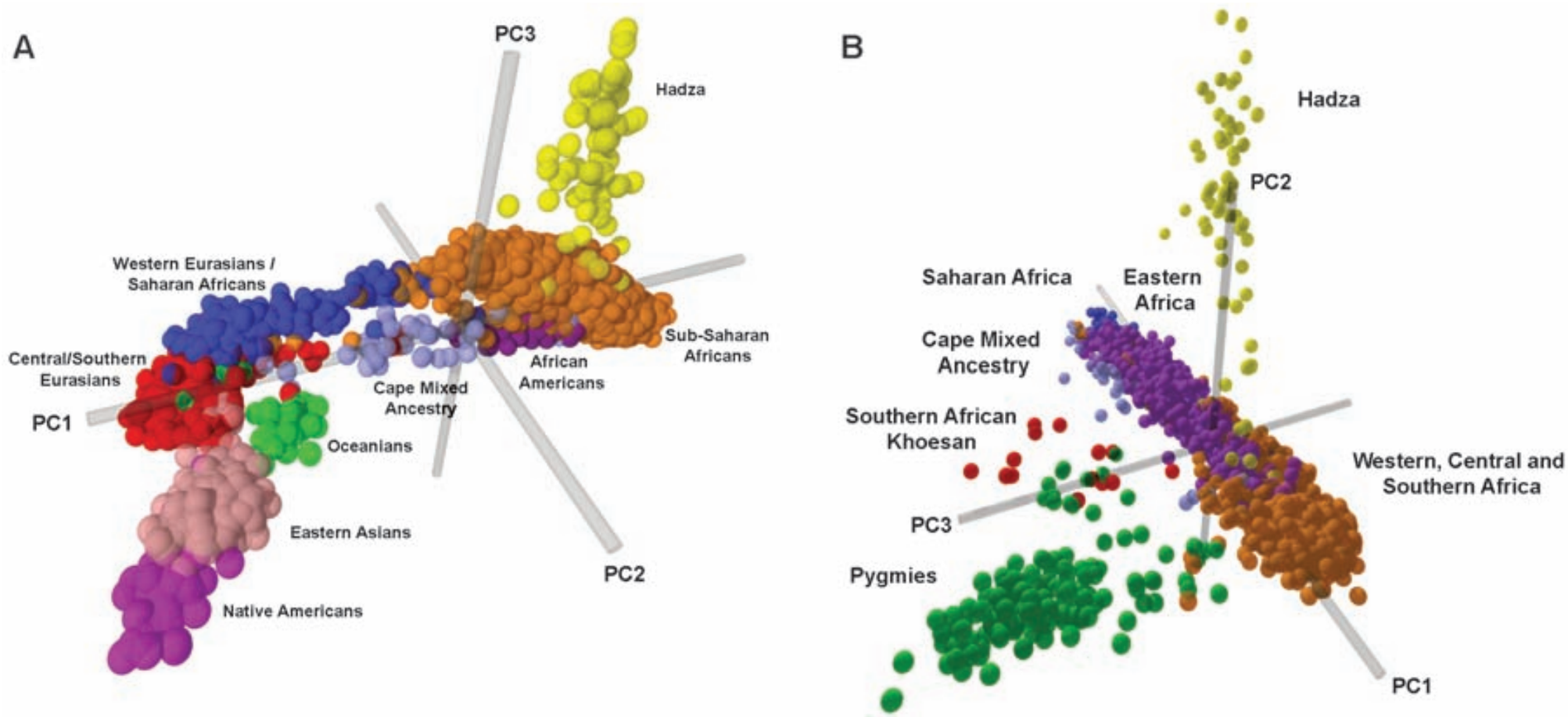
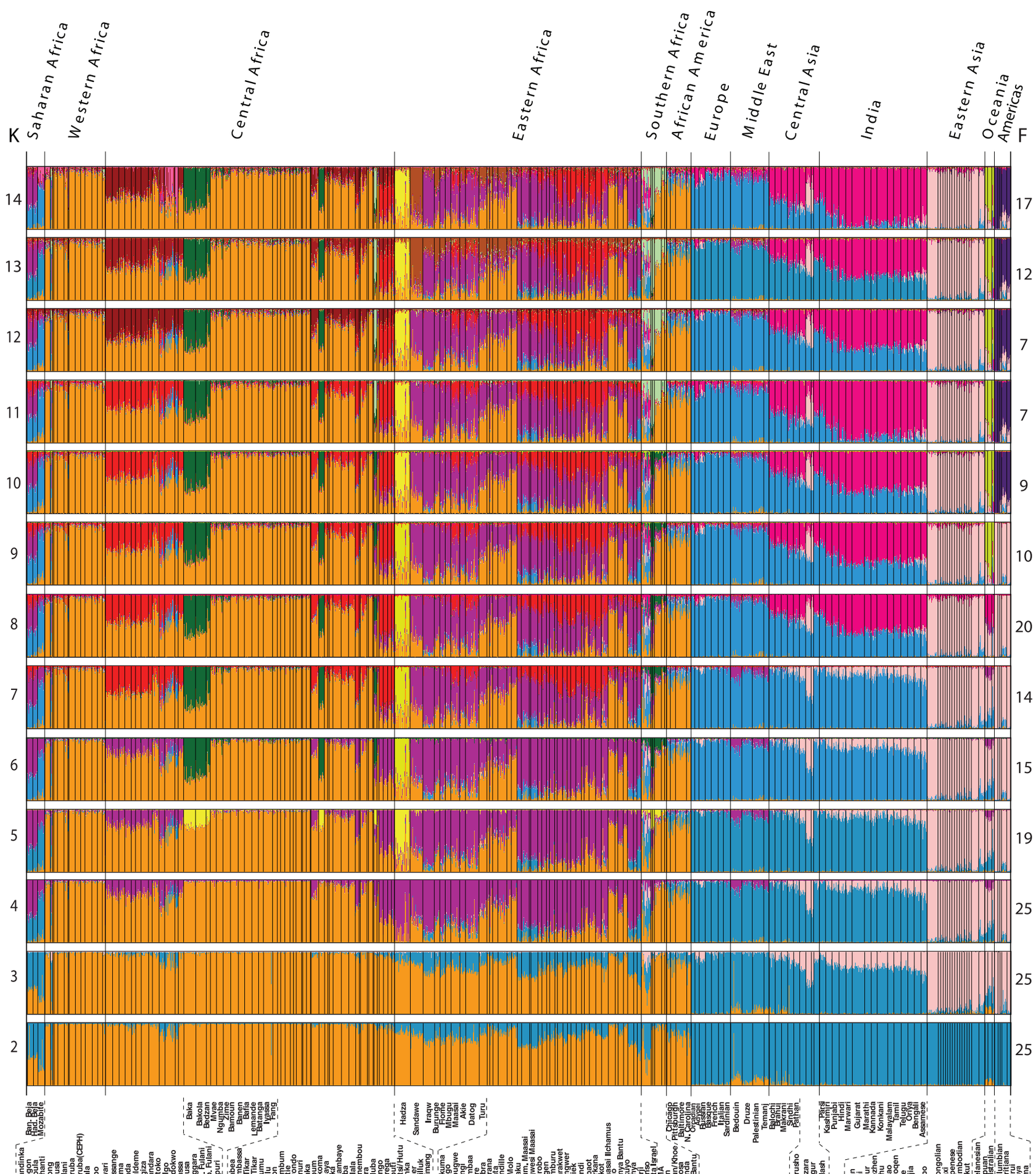


Fig. 2. Principal components analysis (*22*) created on the basis of individual genotypes. (**A**) Global data set and (**B**) African data set.

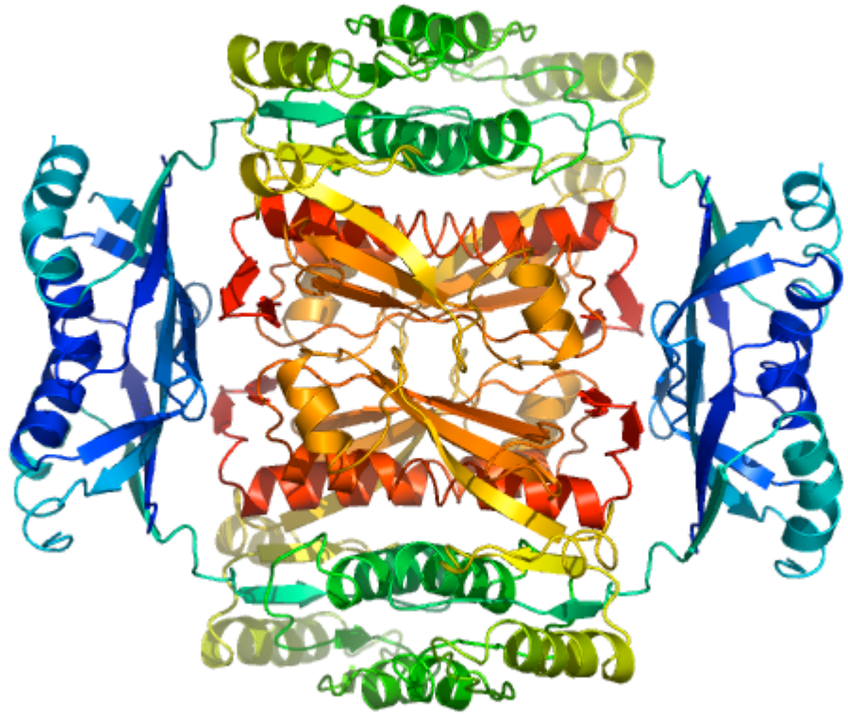
Tishkoff, S. A. et al. The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044 (2009).



Tishkoff, S. A. et al. The genetic structure and history of Africans and African Americans. *Science* **324**, 1035-1044 (2009).

Inferential Structure Determination

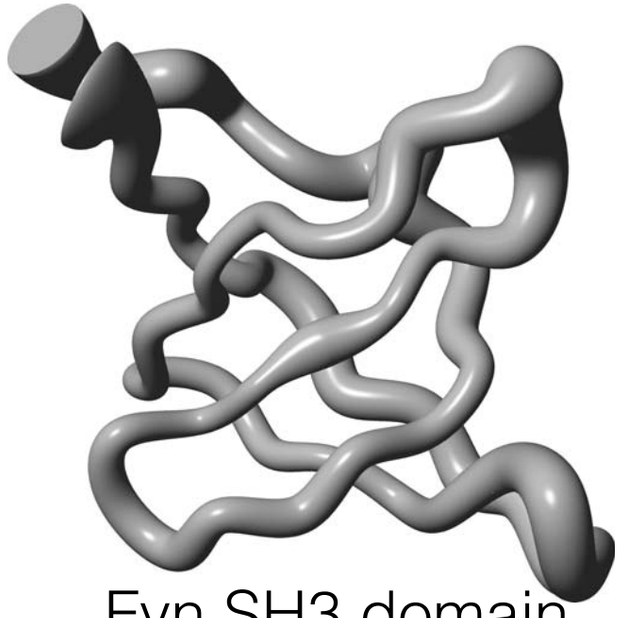
- Major challenge in the determination of three-dimensional macromolecular structures:
  - Experimental data are indirect.
- We observe physical effects that depend on the atomic geometry and use a forward model to relate the observed data to the atomic coordinates.
- Challenges:
  - inherently degenerate
  - data often incomplete
  - data/model rife with uncertainties
  - ill-posed problem: no single structure!!
- Bayesian model can help overcome these!



Rieping, W., Habeck, M., & Nilges, M. (2005). Science 309(5732), 303–306

Inferential Structure Determination

- Rather than try to obtain a single best protein structure, Bayesian statistics can be used to obtain an ensemble.
- In this “sausage plot”, the thickness of the sausage is proportional to the atom-wise error bars from the model.

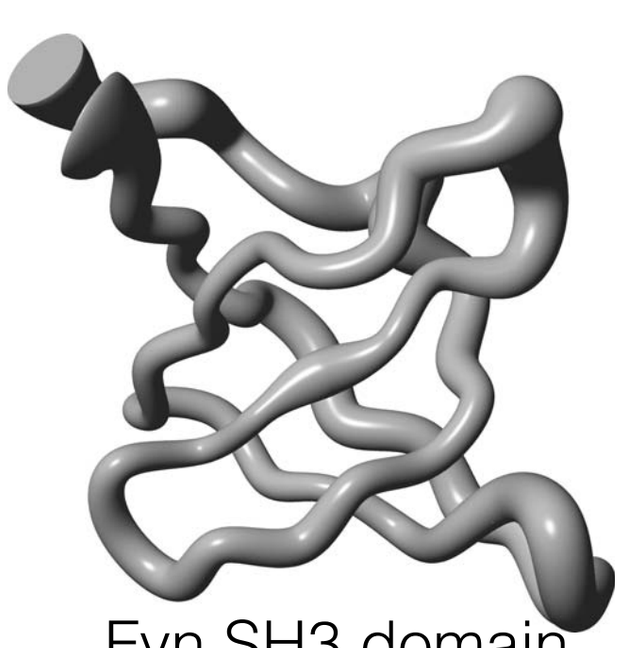


Fyn SH3 domain

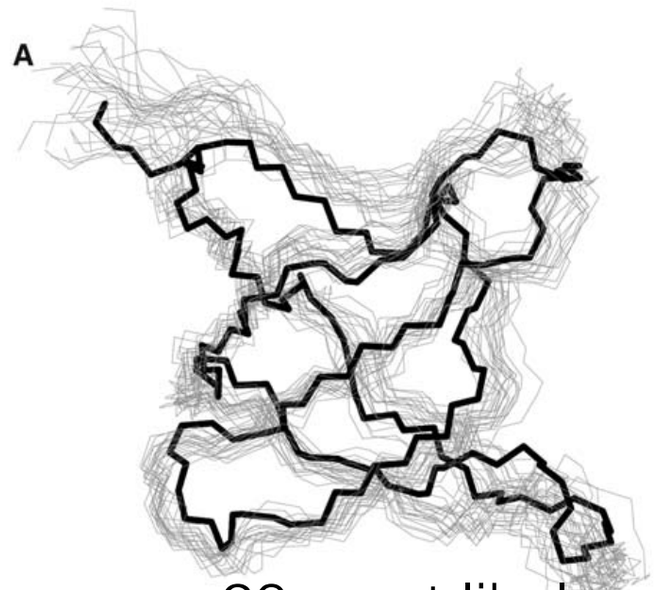
Rieping, W., Habeck, M., & Nilges, M. (2005). Science 309(5732), 303–306

Inferential Structure Determination

- Goal:
  - Get a score (P_i) for every possible conformation (X_i)
  - Rank scores, and keep the best ones



Fyn SH3 domain



20 most likely conformations

Rieping, W., Habeck, M., & Nilges, M. (2005). Science 309(5732), 303–306

Inferential Structure Determination

- Goal:
  - Get a score (P_i) for every possible conformation (X_i)
  - Rank scores, and keep the best ones
- In this case:
  - $P_i = P(X_i \mid D, I)$: Probability of a conformation given the data (D) and prior information (I).
- As usual, apply Bayes' Rule!
  - $P(X|D,I) \propto P(D|X,I)P(X|I)$

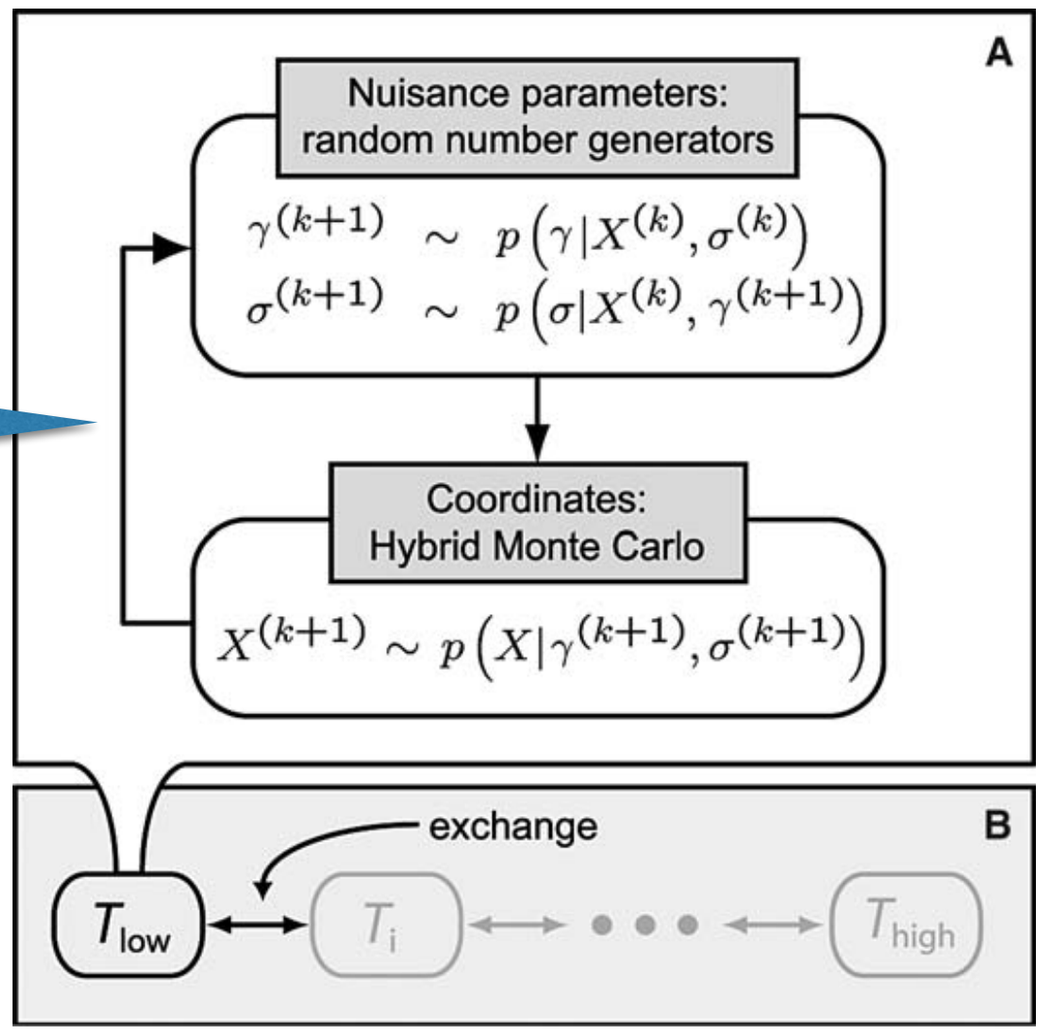
- The likelihood $P(D|X,I)$ combines a forward model that relates observed data to atomic coordinates and an error distribution.
- The prior distribution $P(X|I)$ uses prior information about bimolecular structures, determined by physical energy and temperature of the system.

Rieping, W., Habeck, M., & Nilges, M. (2005). Science 309(5732), 303–306

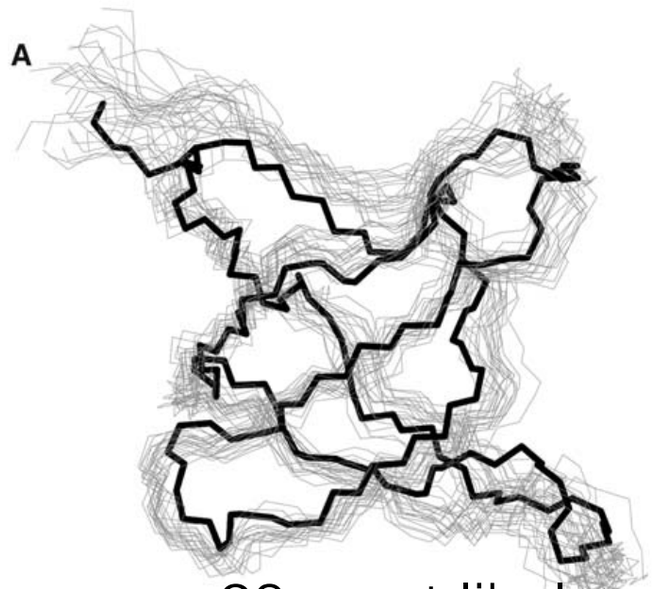
Inferential Structure Determination

$$P(X, \xi | D, I) \propto P(D | X, \xi, I) P(X | I) p(\xi | I)$$

- The full model evaluated incorporates nuisance parameters ($\xi = \{\gamma, \sigma\}$).
- Inference is then performed using MCMC.



Calculate likelihood: $P(D|X,\xi,I)$



20 most likely conformations

Rieping, W., Habeck, M., & Nilges, M. (2005). Science 309(5732), 303–306