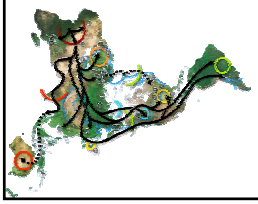
Computational Genomics

10-810/02-710, Spring 2009

Population Stratification from Genetic Data

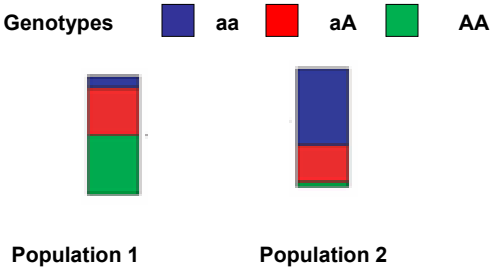
Eric Xing



Lecture 24, April 15, 2009

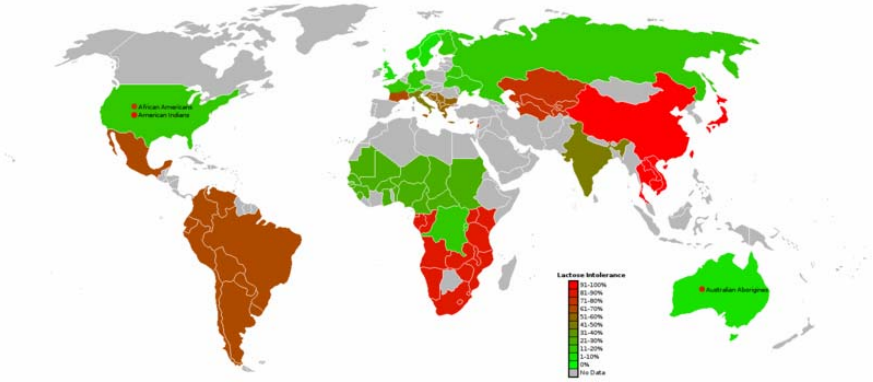
What is population structure?

- Population Structure
  - Among a set of individuals, groups characterized by some measure of genetic distinction
  - A "population" is usually characterized by a distinct distribution over genotypes
  - Example



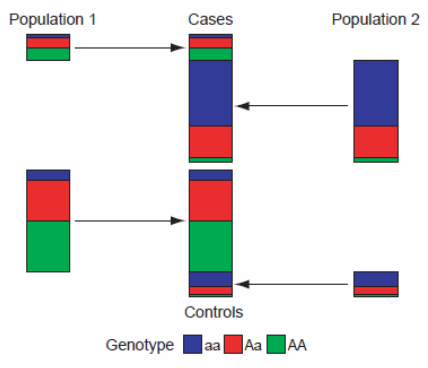
Motivation

- **Reconstructing individual ancestry:** The Genographic Project
  - https://genographic.nationalgeographic.com/genographic/index.html



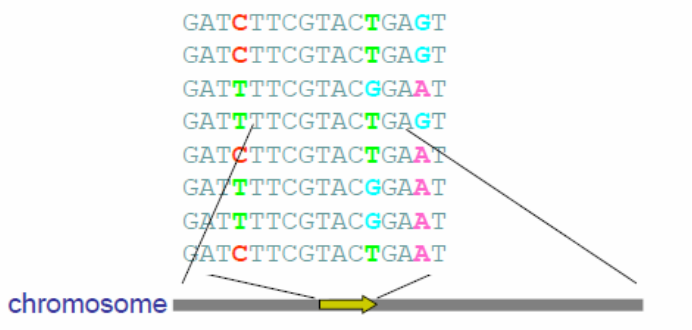
Motivation (continued)

- **Association studies**
  - **Testing genetic basis for diseases.**
  - **Population structure in data causes false positives.**



Genetic Markers

- Single Nucleotide Polymorphism (SNP)
  - Base changes at a single position
  - Each variant called an allele
  - Most common type of polymorphism

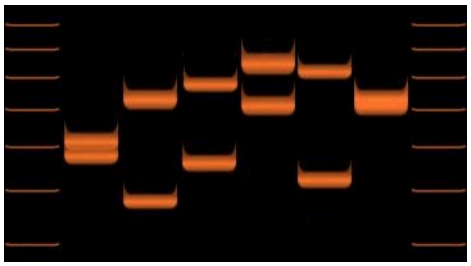


More about SNPs

- SNPs account for around 90% of human genomic variation
- About 10 million SNPs exist in human populations
- Most SNPs are outside protein coding regions, i.e., in exons
- 1 SNP every 100-300 base pairs

Markers (contd)

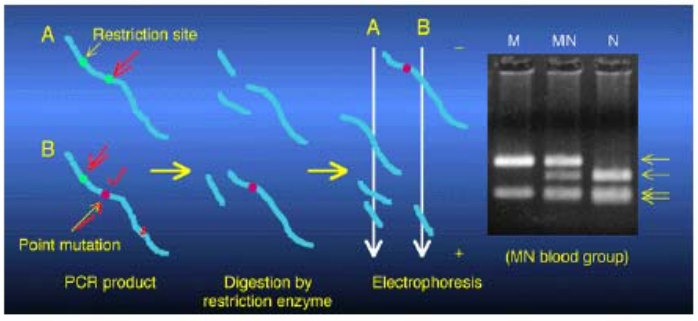
- Variable Number Tandem Repeats (VNTRs)
  - Short nucleotide sequence repeating in the genome.
  - Often show length variation between individuals
    - Microsatellites (4-5 base repeating units)
    - Minisatellites (longer repeating units)
  - Represented as number of counts



VNTR (DS180) alleles in 6 individuals.

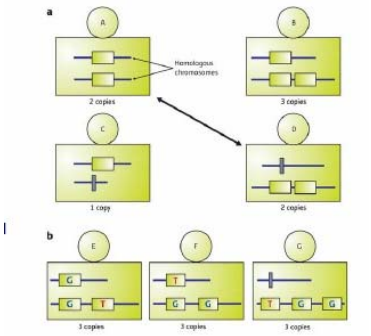
Restriction fragment length polymorphisms

- DNA variations that can be detected by breaking DNA using restriction enzymes, followed by gel electorphoresis.
- Method used before modern DNA sequencing techniques.



Other genetic markers

- Copy Number Variation
  - DNA segment whose numbers differ in different genomes
    - Kilobases to megabases in size
  - Usually two copies of all autosomal regions, one per chromosome
  - Variation due to deletion or duplication



Copy-number variation (CNV) can occur in ambiguous patterns. (a) Individuals in a population may have different copy numbers on homologous chromosomes at CNV loci. For example, here individual A and D have two copies, although the patterns are different: A has one copy on each chromosome, whereas D has two on one chromosome and zero on the other. (b) Individuals may also have CNVs that contain SNPs. For example, individuals E, F, and G each have three copies, but the patterns can be distinguished by the numbers of copies on each chromosome and variations defined by SNPs.

Non-autosomal markers

- Y-chromosome
  - Inherited paternally
  - Finding Y-chromosomal Adam
    - Patrilineal most recent common ancestor of all modern Y chromosomes
- Mitochondrial DNA
  - Inherited maternally
  - Finding Mitochondrial Eve
    - Matrilineal most recent common ancestor of all mitochondrial DNA

Methods based on analysis of simple markers

- Low-dimensional projection
  - PCA-based methods
    - Cavalli-Sforza et al (1978)
    - Patterson et al (2006)
- Clustering
  - Distance-based
    - Bowcock et al (1994)
  - Model-based
    - STRUCTURE- Pritchard et al (2000)
    - mStruct – Shringarpure and Xing (2008)

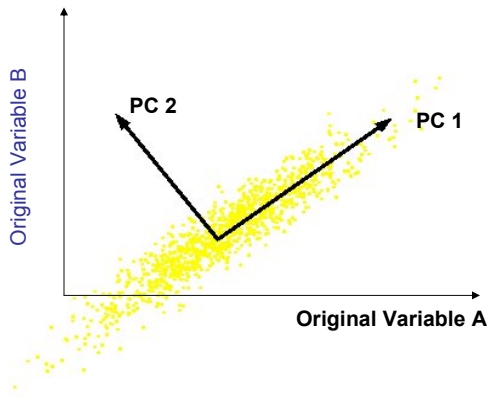
Low-dimensional projections

- Genetic data is very large
  - Number of markers may range from a few hundreds to hundreds of thousands
  - Thus each individual is described by a high-dimensional vector of marker configurations
  - A low-dimensional projection allows easy visualization
- Technique used
  - Factor analysis
  - Many statistical methods exist – ICA, PCA, NMF etc.
  - Principal Components Analysis (next slide)
- Allows projection of individuals into a low dimensional space
- Usually projected to 2 dimensions to allow visualization

Principal Component Analysis

- Most common form of factor analysis
- The new variables/dimensions ...
  - Are linear combinations of the original ones
  - Are uncorrelated with one another
    - Orthogonal in original dimension space
  - Capture as much of the original variance in the data as possible
  - Are called Principal Components
- Demo at http://www.cs.mcgill.ca/~sqrt/dimr/dimreduction.html

What are the new axes?



- **Orthogonal directions of greatest variance in data**
- **Projections along PC1 discriminate the data most along any one axis**

Principal Components

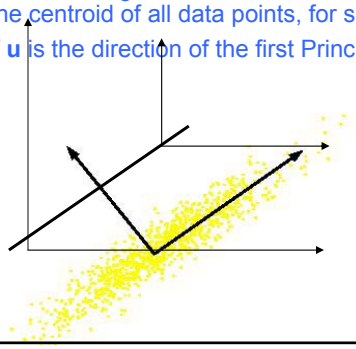
- First principal component is the direction of greatest variability (covariance) in the data
- Second is the next orthogonal (uncorrelated) direction of greatest variability
  - So first remove all the variability along the first component, and then find the next direction of greatest variability
- And so on …

Principal Components Analysis (PCA)

- Principle
  - Linear projection method to reduce the number of parameters
  - Transfer a set of correlated variables into a new set of uncorrelated variables
  - Map the data into a space of lower dimensionality
  - A form of unsupervised learning
- Properties
  - It can be viewed as a rotation of the existing axes to new positions in the space defined by original variables
  - New axes are orthogonal and represent the directions with maximum variability

Computing the Components

- Data points are vectors in a multidimensional space
- Projection of vector **x** onto an axis (dimension) **u** is **u.x**
- Direction of greatest variability is that in which the average square of the projection is greatest
  - I.e. **u** such that $E((\mathbf{u.x})^2)$ over all **x** is maximized
  - (we subtract the mean along each dimension, and center the original axis system at the centroid of all data points, for simplicity)
  - This direction of **u** is the direction of the first Principal Component



Computing the Components

- $E((\mathbf{u.x})^2) = E((\mathbf{u.x})(\mathbf{u.x})^T) = E(\mathbf{u.x.x}^T.\mathbf{u}^T)$
- The matrix $\mathbf{C} = \mathbf{x.x}^T$ contains the correlations (similarities) of the original axes based on how the data values project onto them
- So we are looking for w that maximizes $\mathbf{uCu}^T$, subject to **u** being unit-length
- It is maximized when w is the principal eigenvector of the matrix **C**, in which case
  - $\mathbf{uCu}^T = \mathbf{u}\lambda\mathbf{u}^T = \lambda$ if **u** is unit-length, where λ is the principal eigenvalue of the correlation matrix C
  - The eigenvalue denotes the amount of variability captured along that dimension

Why the Eigenvectors?

Maximise $\mathbf{u}^T\mathbf{x}\mathbf{x}^T\mathbf{u}$ s.t $\mathbf{u}^T\mathbf{u} = 1$

Construct Langrangian $\mathbf{u}^T\mathbf{x}\mathbf{x}^T\mathbf{u} - \lambda\mathbf{u}^T\mathbf{u}$

Vector of partial derivatives set to zero

$$\mathbf{x}\mathbf{x}^T\mathbf{u} - \lambda\mathbf{u} = (\mathbf{x}\mathbf{x}^T - \lambda\mathbf{I})\,\mathbf{u} = 0$$

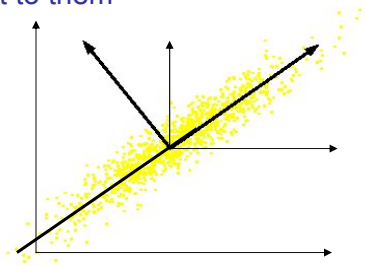
As $\mathbf{u} \neq \mathbf{0}$ then $\mathbf{u}$ must be an eigenvector of $\mathbf{x}\mathbf{x}^T$ with eigenvalue λ

Singular Value Decomposition

- The first root is called the prinicipal eigenvalue which has an associated orthonormal ($\mathbf{u}^T\mathbf{u} = 1$) *eigenvector* $\mathbf{u}$
- Subsequent roots are ordered such that $\lambda_1 > \lambda_2 > \ldots > \lambda_M$ with rank($\mathbf{D}$) non-zero values.
- Eigenvectors form an orthonormal basis i.e. $\mathbf{u}_i^T\mathbf{u}_j = \delta_{ij}$
- The eigenvalue decomposition of $\mathbf{x}\mathbf{x}^T = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$, where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M]$ and $\boldsymbol{\Sigma} = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_M]$
- Similarly the eigenvalue decomposition of $\mathbf{x}^T\mathbf{x} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^T$
- The SVD is closely related to the above $\mathbf{x} = \mathbf{U}\,\boldsymbol{\Sigma}^{1/2}\,\mathbf{V}^T$
- The left eigenvectors **U,** right eigenvectors **V,**
- singular values = square root of eigenvalues**.**

Computing the Components

- Similarly for the next axis, etc.
- So, the new axes are the eigenvectors of the matrix of correlations of the original variables, which captures the similarities of the original variables based on how data samples project to them



- **Geometrically: centering followed by rotation**
  - **Linear transformation**

PCs, Variance and Least-Squares

- The first PC retains the greatest amount of variation in the sample
- The k^{th} PC retains the k^{th} greatest fraction of the variation in the sample
- The k^{th} largest eigenvalue of the correlation matrix C is the variance in the sample along the k^{th} PC

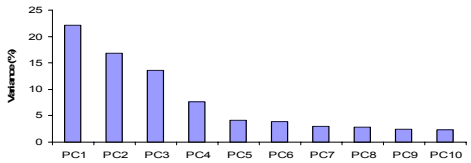
- The least-squares view: PCs are a series of linear least squares fits to a sample, each orthogonal to all previous ones

How Many PCs?

- For n original dimensions, correlation matrix is nxn, and has up to n eigenvectors. So n PCs.
- Where does dimensionality reduction come from?

Dimensionality Reduction

Can *ignore* the components of lesser significance.

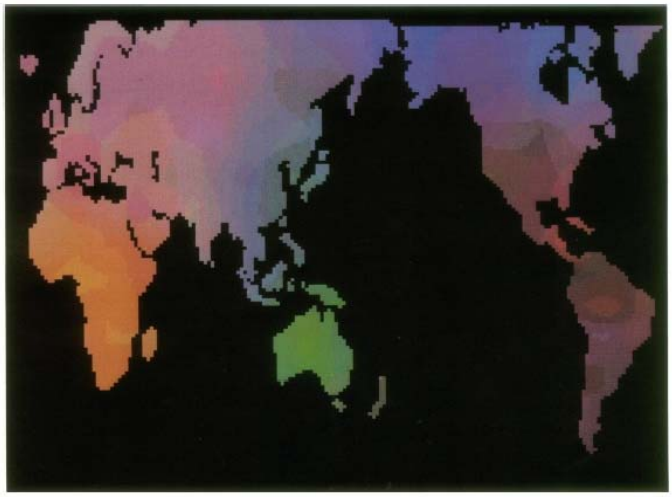


You do lose some information, but if the eigenvalues are small, you don't lose much

- **n dimensions in original data**
- **calculate n eigenvectors and eigenvalues**
- **choose only the first p eigenvectors, based on their eigenvalues**
- **final data set has only p dimensions**

PCA analysis (Cavalli-sforza,1978)

- Plot o
  - Firs
  - Sec
  - Thir
  - Inte



Comments

Advantages

- Statistical tests
  - For significance of results (Patterson et al. 2006)
- Easy visualization

Disadvantages

- Only mathematical analysis
- No intuition about underlying processes

Distance-based Clustering

Idea

- Compute genetic distance between individuals
  - Many distance measures possible.
  - Nei's genetic distance (Nei, 1972)
  - Cavalli-Sforza chord measure (Cavalli-Sforza and Edwards, 1967)
  - Reynolds, Weir, and Cockerham's genetic distance (1983)
- Construct a pairwise distance matrix for given set of individuals
- Visualize using some representation
  - Tree-based (neighbor-joining tree)
  - Multi-dimensional scaling

Comments

Advantages

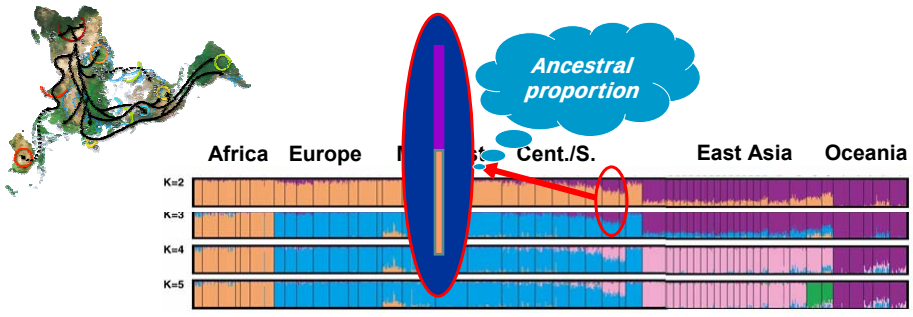
- Simple to compute
- Easy to visualize

Disadvantages

- Clustering depends on distance measure chosen
- Difficult to determine confidence in clustering

Model-based clustering: Structure

- How to display population structure?
  - *Structure (Pritchard et al, 2000)*



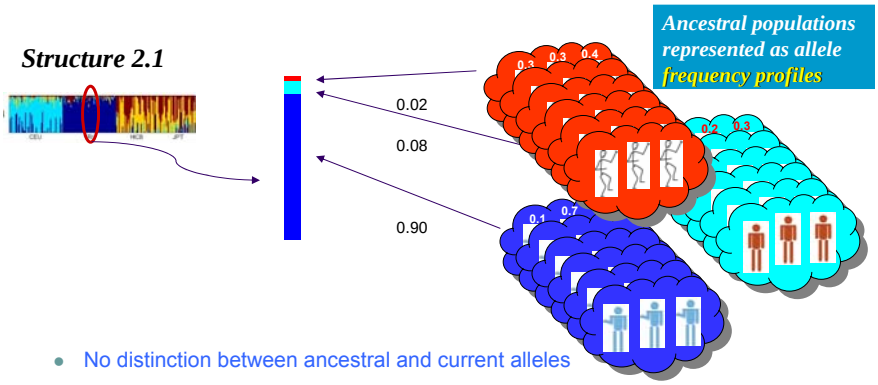
Genetic structure of Human Populations (Rosenberg et al. 2002)

Structure model

- Hypothesis: Modern populations are created by an intermixing of ancestral populations.
- An individual's genome contains contributions from one or more ancestral populations.
- The contributions of populations can be different for different individuals.
- Other assumptions
  - Hardy-weinberg equilbrium
  - No linkage disequilbrium
  - Markers are i.i.d (independent and identically distributed)

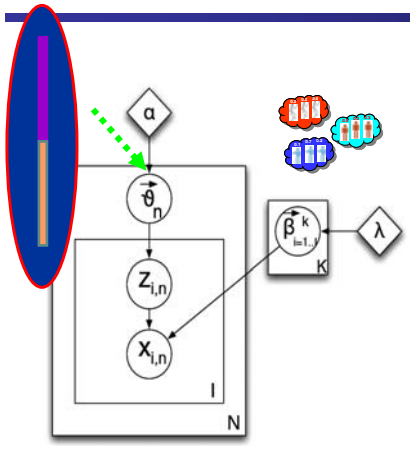
The Admixture Model

- Admixture of "ancestral frequency profiles (AP)"



- No distinction between ancestral and current alleles
- Does not model mutation and chromosomal recombination

The Admixture Model



- β = Distribution over alleles
  - One per population –locus pair
- To generate an individual's genome
- Sample θ from Dirichlet(α)
- For each locus
  - Sample z from Multinomial(θ)
  - Sample x from β corresponding to the population chosen by z

Comments

Advantages

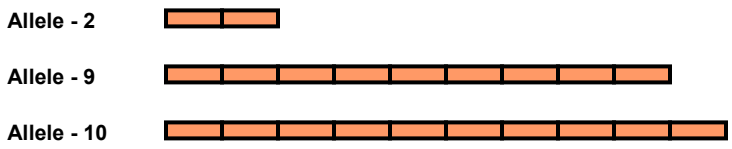
- Generative process
  - Explicit model of admixture
- Meaningful results
- Clustering is probabilistic
  - Can interpret confidence level of clusters

Disadvantages

- Alleles are same in ancestral and modern populations
- No models of mutation, recombination
  - Note: Recombination added in extension by Falush et al.

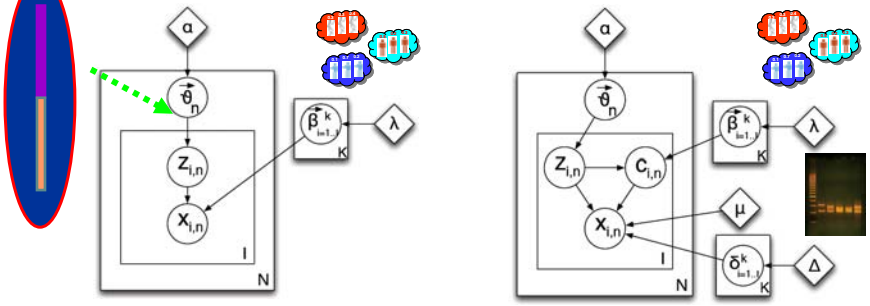
Allele mutations- modeling allele similarity

- Microsatellite units 



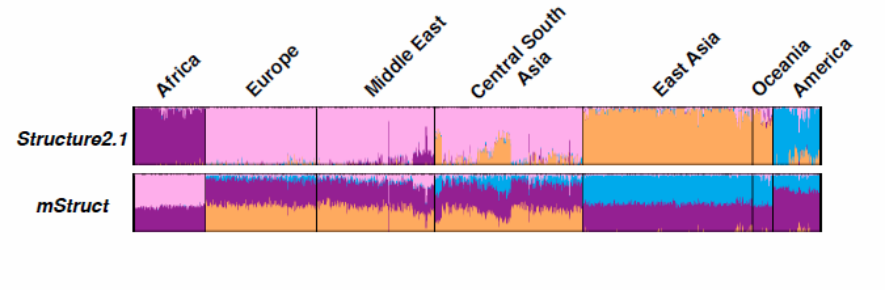
- Allele 9 is more similar to allele 10 than allele 2
- Allele 10 might be a mutation of allele 9
- Structure considers all alleles to be unique
- What if we model allele similarity?
- mStruct – Structure under mutations

From *Structure* to *mStruct*



- From admixture of APs to admixture of MIMs
  - MiM: population-specific Mixture of Inheritance Models
- The inheritance model:
  - Microsatellite: $P(b|a) = \frac{1-\delta}{1-\delta^a+\delta}\delta^{|b-a|}$. SNPs: $P(b|a) = \delta^{\mathcal{I}[b=a]} \times (1-\delta)^{\mathcal{I}[b\neq a]}; \quad a,b \in \{0,1\}.$

Comparing population structure maps

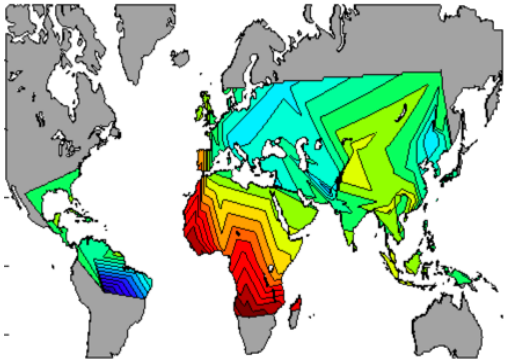


Ancestry structure maps inferred from microsatellite portion of the HGDP dataset, using *mStruct* and *Structure* with 4 ancestral population. The colors represent different ancestral populations.

- **A common ancestral population is now seen across all continents!**
- **Clusters remain unchanged**

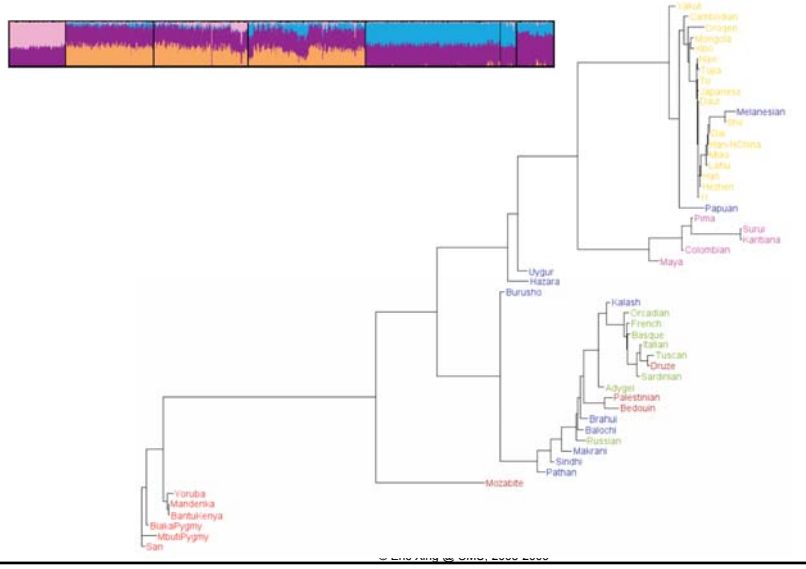
Analyzing mutation empirically

- Contours of empirical accumulated mutation over the world map.
  - Red – high, Blue –low

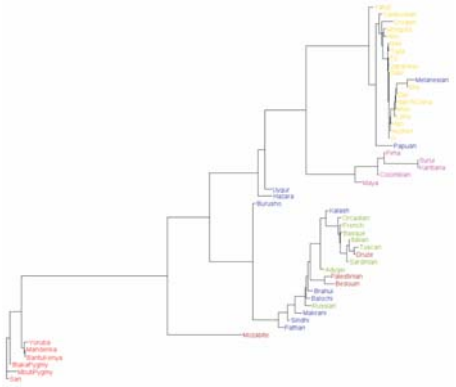


A gradient outward from Africa!

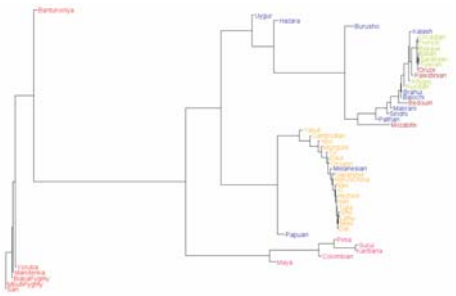
Phylogenetic tree from mStruct Structural map



Neighbour-joining Phylogenetic Trees from the Structural Maps



mStruct



Structure

Novel approaches

- Using indirect approaches to study population history
- Do not use human genetic data, but study human population evolution
- Using language phylogenies
  - Gray et al. (2009)
- Using gut bacteria from human populations
  - Falush et al. (2003)