Introduction to Population Genetics

Selina Carlhoff and Mei-Shin Wu

MPI-SHH
SUMMER SCHOOL
2021

Doorway
to Human History

Introduction to Population Genetics

1. What is population genetics?

2. Genomic variation

3. Applications & Methods

4. Genetics & Culture

5. Multidisciplinary studies

Population genetics

Definition, Principle, Application

SUMMER SCHOOL 2021

Doorway to Human History

Population *Genetics*

Population genetics inspect genetic variation within and among populations and study the **evolutionary factors** that explain this variation.

What is the genetic profile of the population?

Who is/are the ancestor/s of the population?

Which demographic processes affected the studied population?

Application of population genetic studies

Have you heard of 23andMe?

Evolutionary studies : reconstructing human history & migration routes.

Clinical studies : association between human genome and disease

Hardy–Weinberg equilibrium

1. Large population

2. Random mating

3. No gene flow

4. No natural selection

- **A theoretical baseline**

Genomic variation

Genomic variation, Evolution, Genetic drift

SUMMER SCHOOL 2021

Doorway to Human History

Genomic variation

1. Single Nucleotide Polymorphism (SNP)

2. Short insertion or deletion (INDEL)

3. Copy number variation (CNV)

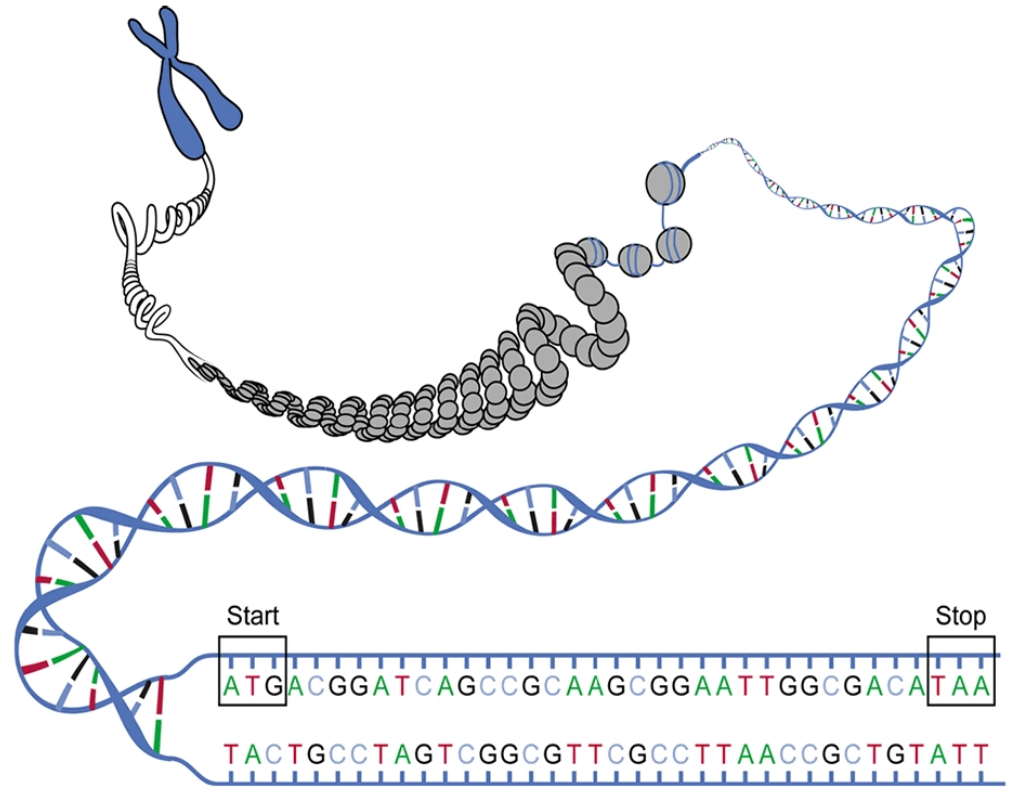


Image source: https://biologywarakwarak.wordpress.com/2012/01/15/the-3-magical-rules-to-determine-the-amino-acid-chain-from-a-dna-piece-without-error/

Single Nucleotide Polymorphisms (SNPs)

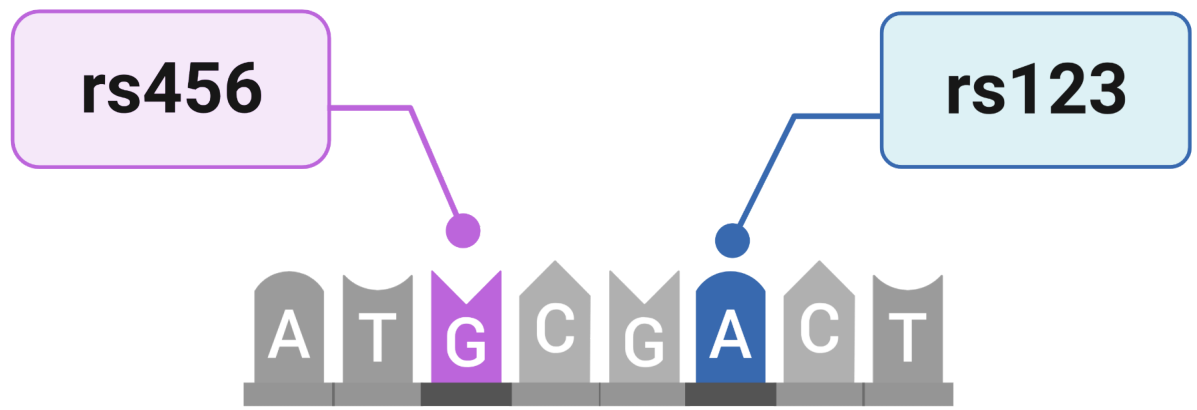
= single nucleotides that exhibit variation in different individuals

most SNP chips developed for genome-wide association studies

fast, easy and inexpensive way to retrieve millions of genotypes

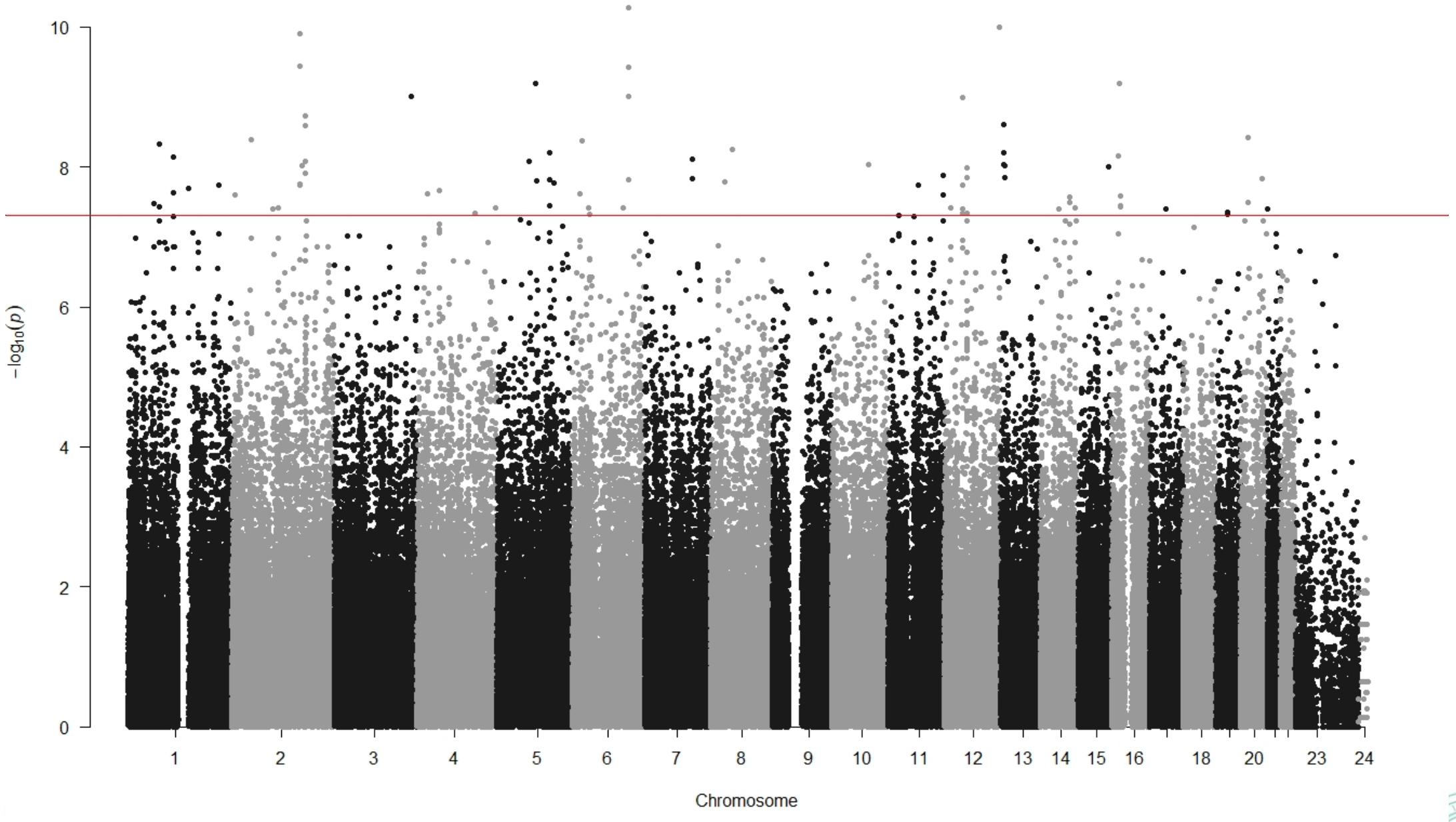
ascertainment bias towards variation in European populations

SNP sets: Human Origins (~500,000 SNPs)
1240K (~1,240,000 SNPs)



created with BioRender

Single Nucleotide Polymorphisms (SNPs)



Single Nucleotide Polymorphisms (SNPs)

a. Autosome

b. Mitochondral DNA

c. Y chromosome

Wild Type		Mutan Type
A T C G A C T G	▶	A T C **T** A C T G
T A G C T G A C		T A G **A** T G A C

Genomic variation

2. Insertion/Deletion polymorphisms (INDELs)

Wild Type		Mutan Type
A T C G A C T G	▶	A T C - A A C T G
T A G C T G A C		T A G - T T G A C

3. Copy number variation (CNV)

   A larger range of insertion/deletion.

Allele frequeny

The occurrence of a genetic variant in a **population.**

For example: The G and T alleles found among populations are 55% and 45% correspondingly.

Wild Type (G)

A T C G A C T G

T A G C T G A C

Mutan Type (T)

A T C T A C T G

T A G A T G A C

Allele frequeny

“Frequeny” means the two alleles should be summed up to 100%

$$G + T = 1$$

Human have two sets of chromosomes:

$$(G + T)^2 = 1^2$$

$$GG + 2GT + TT = 1$$

Example

Geneticists found 200 cats with pink palms among 1000 cats on the Dream Island. Under the assumption of Hardy-Weinberg equilibrium, how many cats with “GG” and “GT” alleles accordingly?

GG + 2GT + TT = 1

0.30 + 0.5 + 0.2 = 1

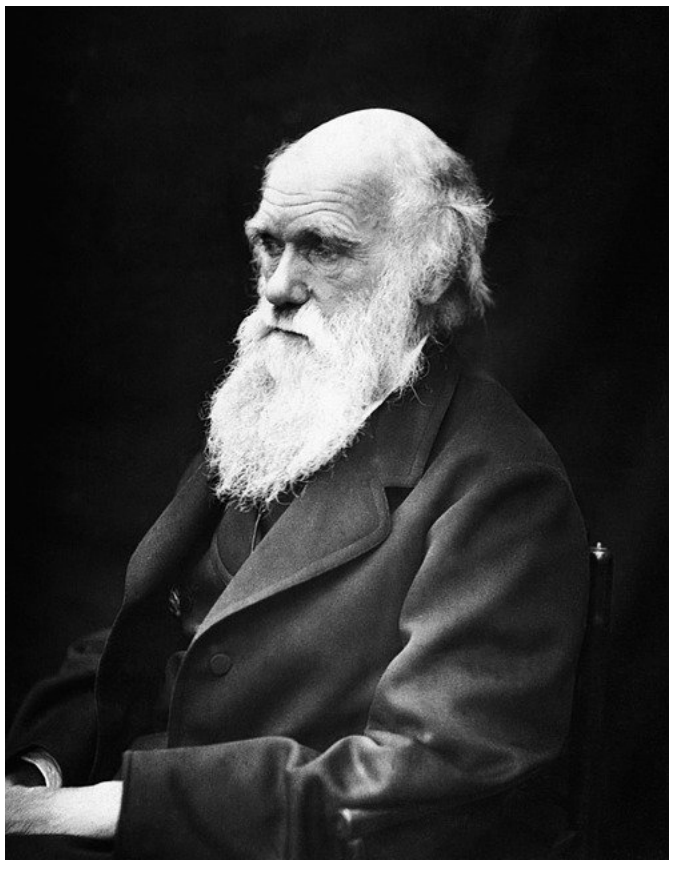
300 + 500 + 200 = 1000 (cats)

If gene flows occurred in the past, the allele frequencies will be significantly different from the assumption.

Evolutionary factors

Natural selection
(Non-random)

Genetic drift
(Random)



Charles Darwin

Sewall Wright

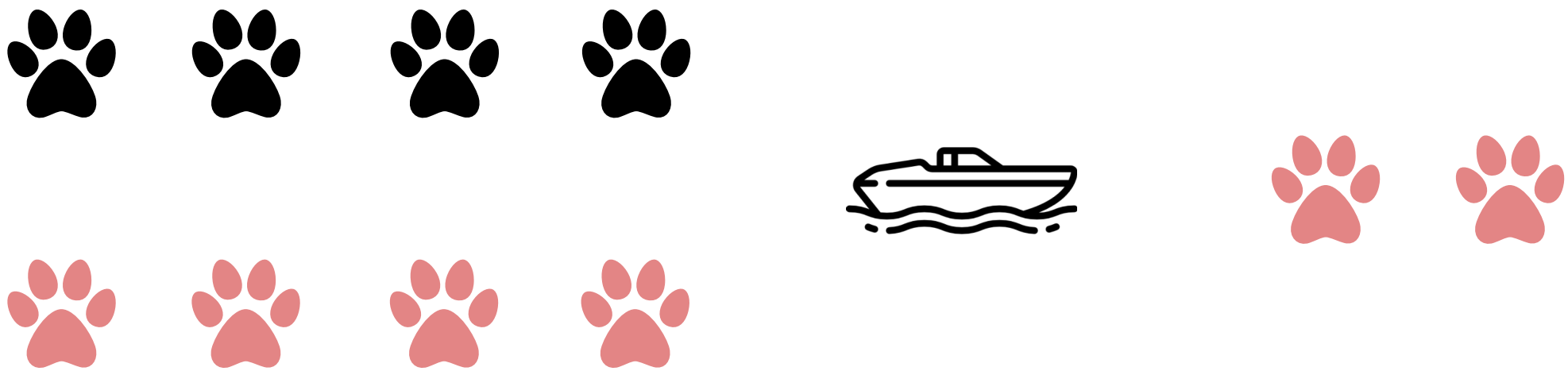
Ronald Fisher

Image by WikiImages from Pixabay and Wikipedia

Random Evolutionary factors

Bottle neck effect (i.e. natural disaster)

Founder effect (i.e. population split)

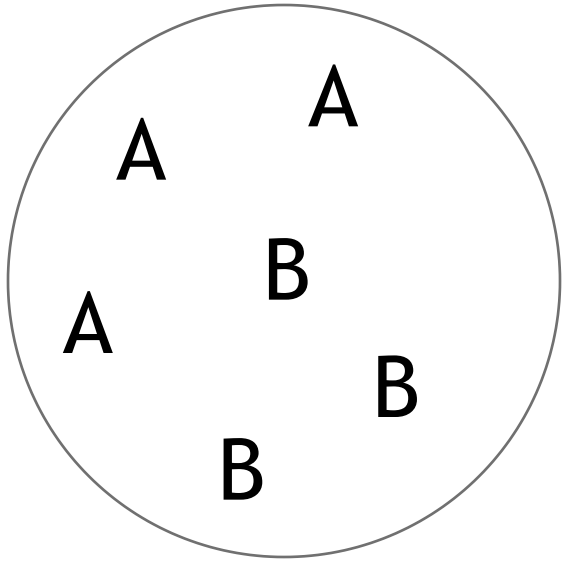


Genetic distance (F_{ST})

Compare total genetic diversity (H_T) to that within group (H_S)

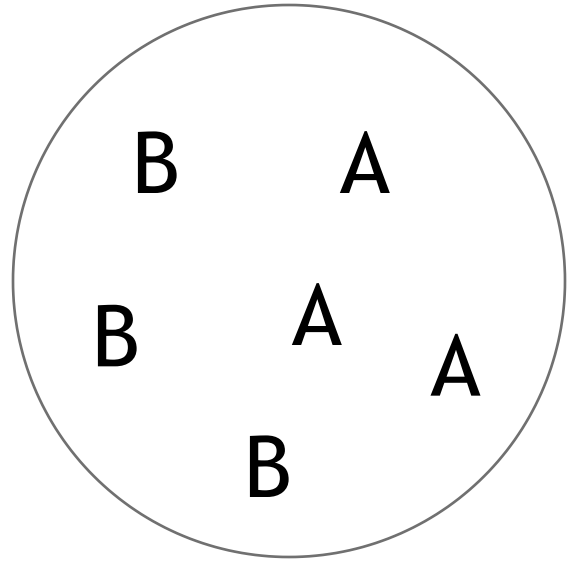
$$F_{ST} = (H_T - H_S) / H_T$$

example: one SNP with two alleles (A or B)



Population 1

$H_1 = 0.5$



Population 2

$H_2 = 0.5$

$H_S = 0.5$

$H_T = 0.5$

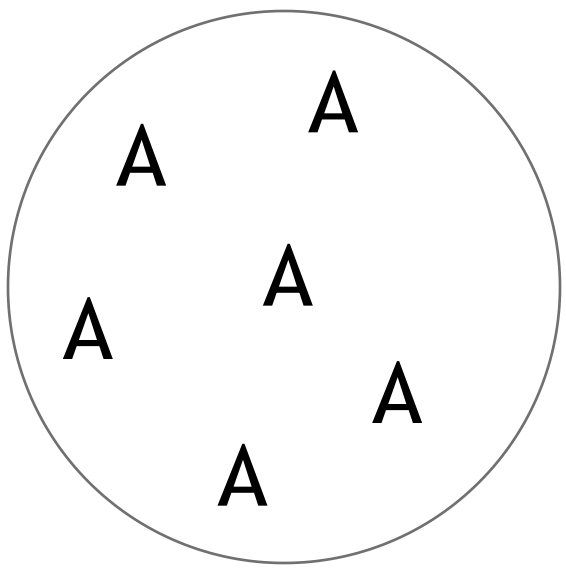
$F_{ST} = (0.5 - 0.5) / 0.5$
$= 0$

—> no genetic difference

Genetic distance (F_{ST})

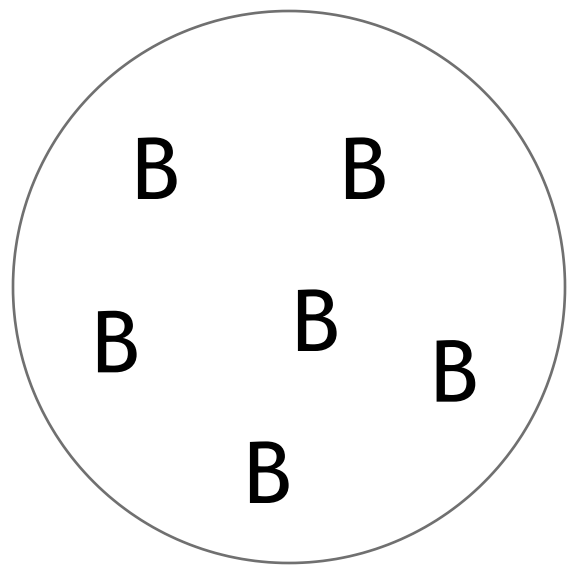
$$F_{ST} = (H_T - H_S) / H_T$$

example: one SNP with two alleles (A or B)



Population 1

$H_1 = 0$



Population 2

$H_2 = 0$

$H_S = 0$

$H_T = 0.5$

$F_{ST} = (0.5 - 0) / 0.5$
$= 1$

—> max. genetic difference

Genetic distance (F_{ST})

FST values range from 0 to 1

Can be applied to pairs/many populations and one/many loci

limitations:

F_{ST} = 1 only if alleles fixed

$F_{ST} \neq 0$ even if no alleles in common

For humans F_{ST} = 0.08-0.14

—> 8-14% of genetic variation between populations

—> 86-92% of genetic variation within populations

Methods - 1

mtDNA, Y-chromosome, autosomal DNA

mitochondrial DNA

organelles in the cells involved in energy production

Circular molecule of ~16,500 nucleotides

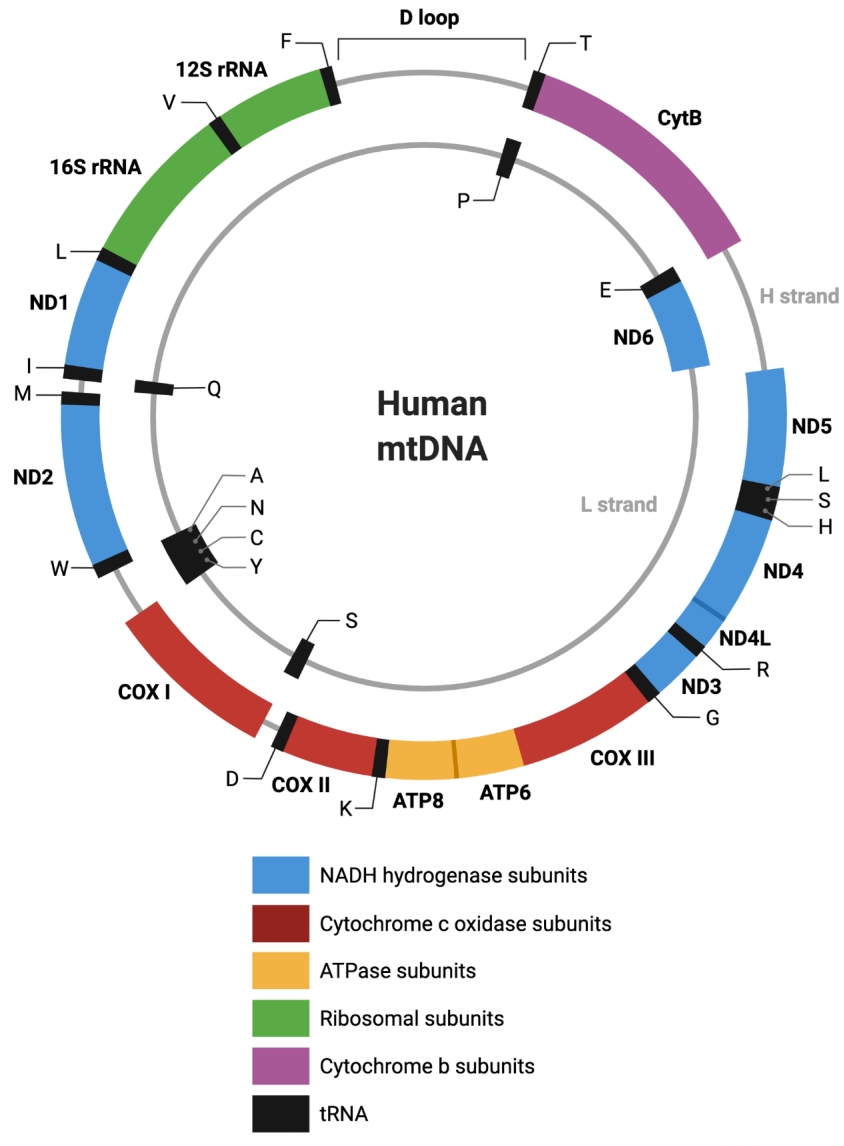
Different genetic coding of amino acids

Evolves 10x faster than nuclear DNA

Several hundred mtDNA per nuclear DNA

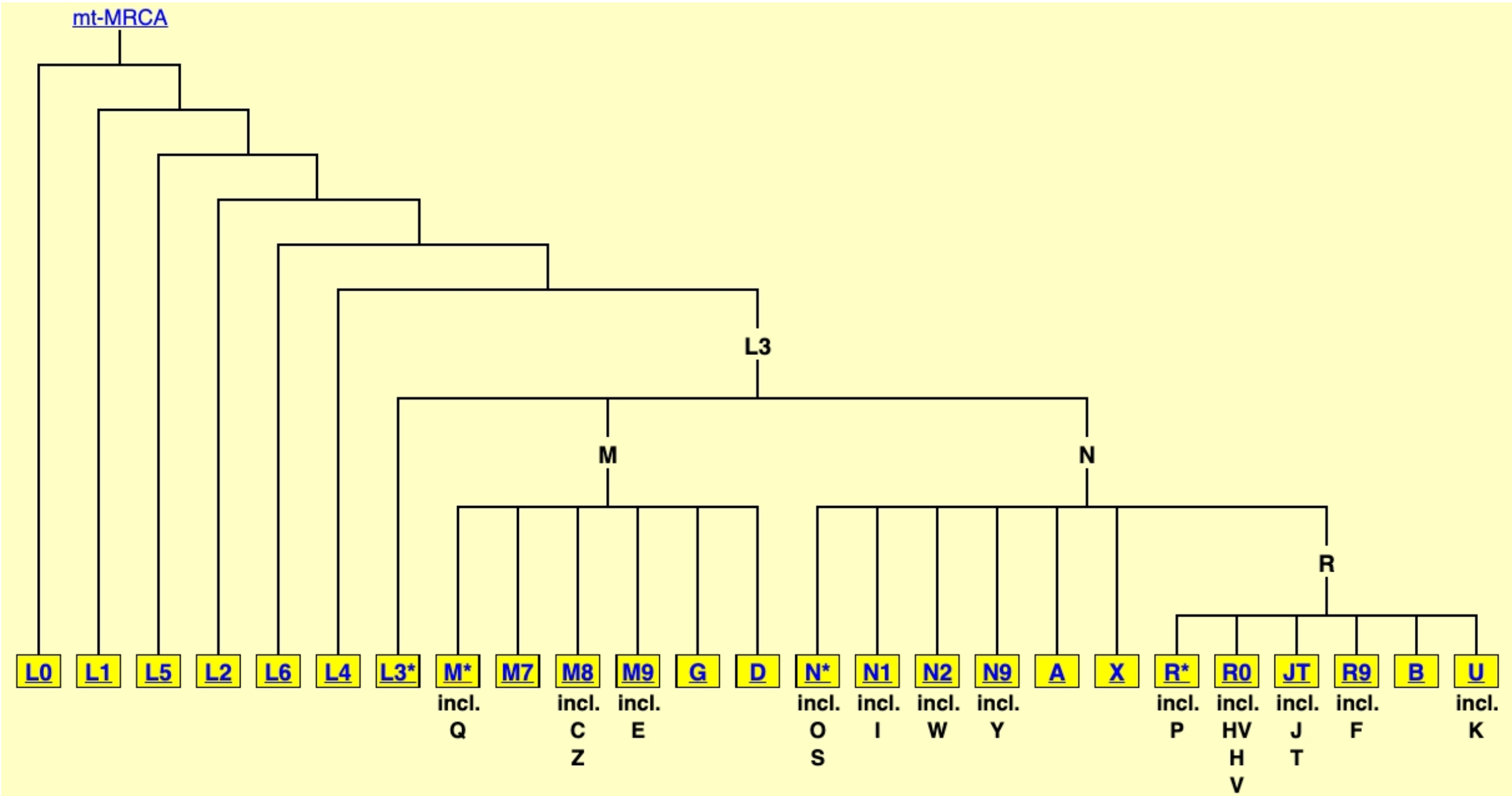
Only inherited through maternal lineage
Without recombination

Haplogroups based on shared mutations



created with BioRender

mtDNA haplogroups



SUMMER SCHOOL 2021 Doorway to Human History

https://www.phylotree.org/

Y-chromosome

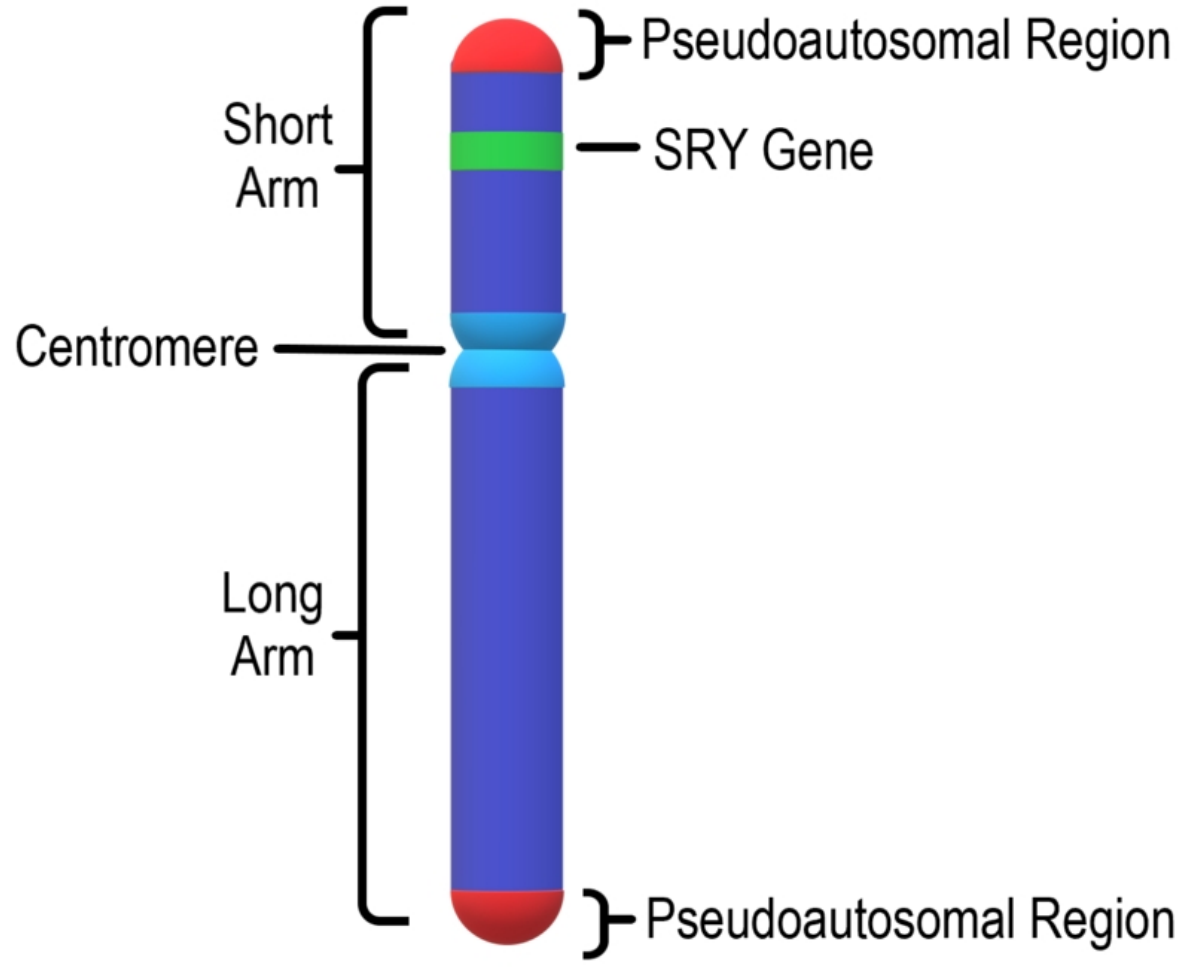
Only inherited through paternal lineage

Pseudoautosomal - large part non-recombining (NRY)

Very short, very few genes

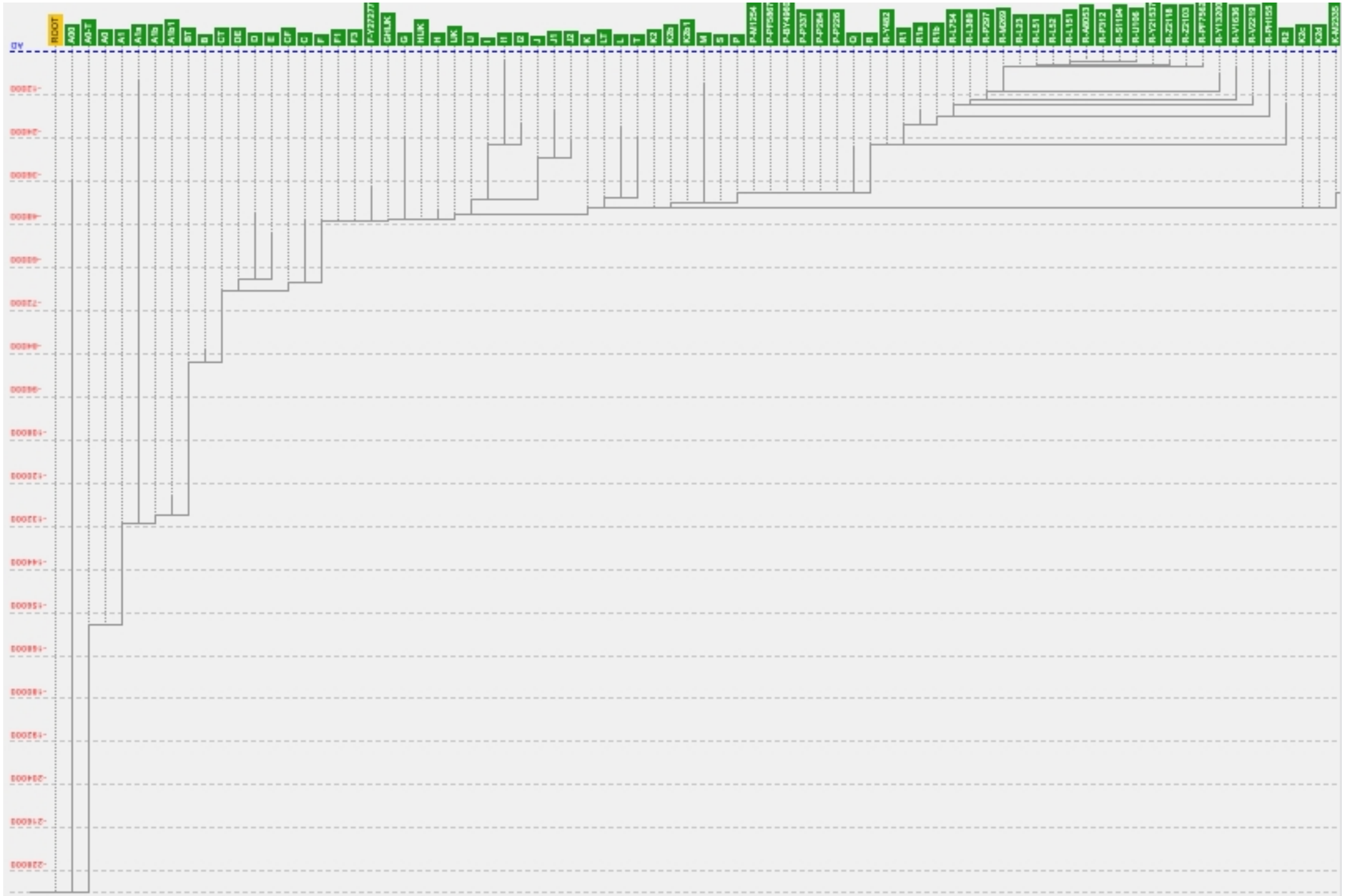
Single, haploid locus

Phylogeny split into haplogroups



(c) Christinelmiller, Wikimedia Commons

Y-haplogroups

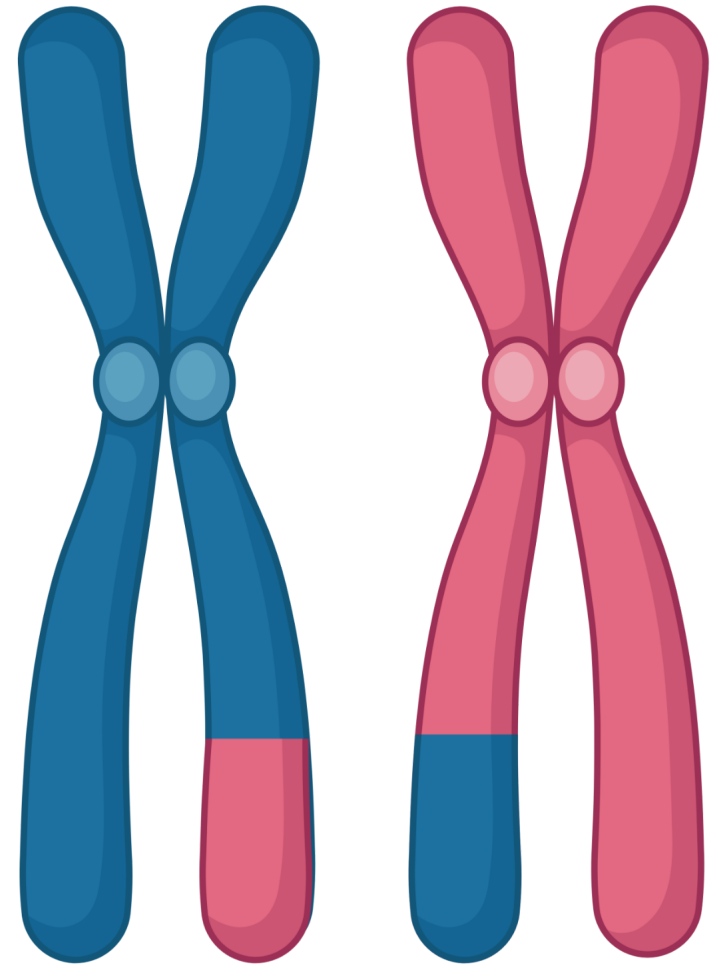


Autosomal DNA

Vastly greater amount of genetic information

Recombination mixes parental chromosomes each generation

Averaging across many positions reduces effects of genetic drift



created with BioRender

Genetic linkage

The tendency that physically closely positioned loci are inherited together

Linkage disequilibrium (LD) = nonrandom association between genotypes of linked SNPs

LD “decays” over time due to recombination & mutations

LD can be used to infer time of admixture & selection

Methods - 2

Unsupervised analyses & admixture modelling

Example

RESEARCH

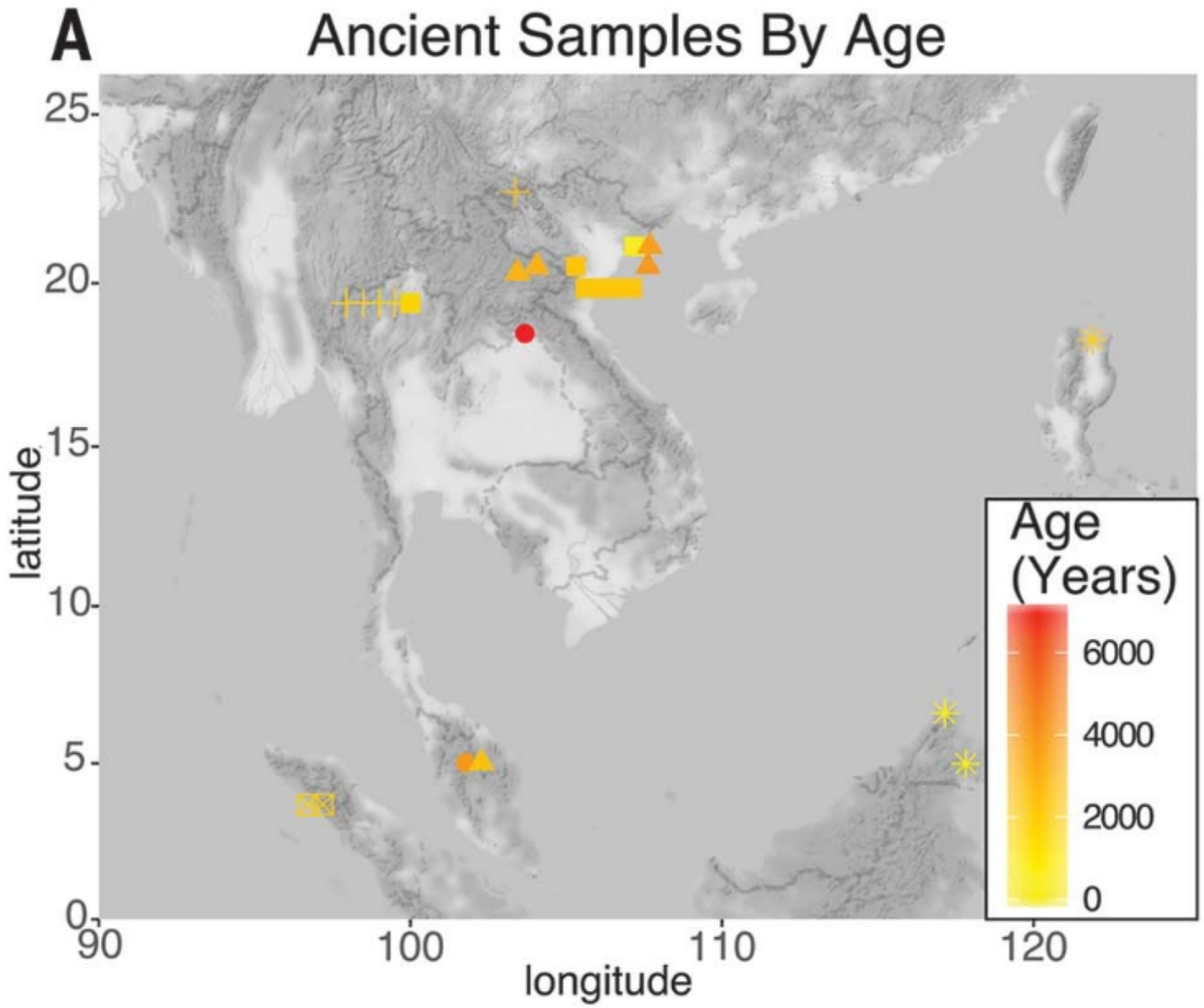
HUMAN GENOMICS

The prehistoric peopling of Southeast Asia

Hugh McColl[1]*, Fernando Racimo[1]*, Lasse Vinner[1]*, Fabrice Demeter[1,2]*, Takashi Gakuhari[3,4], J. Víctor Moreno-Mayar[1], George van Driem[5,6], Uffe Gram Wilken[1], Andaine Seguin-Orlando[1,7], Constanza de la Fuente Castro[1], Sally Wasef[8], Rasmi Shoocongdej[9], Viengkeo Souksavatdy[10], Thongsa Sayavongkhamdy[10], Mohd Mokhtar Saidin[11], Morten E. Allentoft[1], Takehiro Sato[12], Anna-Sapfo Malaspinas[13], Farhang A. Aghakhanian[14], Thorfinn Korneliussen[1], Ana Prohaska[15], Ashot Margaryan[1,16], Peter de Barros Damgaard[1], Supannee Kaewsutthi[17], Patcharee Lertrit[17], Thi Mai Huong Nguyen[18], Hsiao-chun Hung[19], Thi Minh Tran[18], Huu Nghia Truong[18], Giang Hai Nguyen[18], Shaiful Shahidan[11], Ketut Wiradnyana[20], Hiromi Matsumae[4], Nobuo Shigehara[21], Minoru Yoneda[22], Hajime Ishida[23], Tadayuki Masuyama[24], Yasuhiro Yamada[25], Atsushi Tajima[12], Hiroki Shibata[26], Atsushi Toyoda[27], Tsunehiko Hanihara[4], Shigeki Nakagome[28], Thibaut Deviese[29], Anne-Marie Bacon[30], Philippe Duringer[31,32], Jean-Luc Ponche[33], Laura Shackelford[34], Elise Patole-Edoumba[35], Anh Tuan Nguyen[18], Bérénice Bellina-Pryce[36], Jean-Christophe Galipaud[37], Rebecca Kinaston[38,39], Hallie Buckley[38], Christophe Pottier[40], Simon Rasmussen[41], Tom Higham[29], Robert A. Foley[42], Marta Mirazón Lahr[42], Ludovic Orlando[1,7], Martin Sikora[1], Maude E. Phipps[14], Hiroki Oota[4], Charles Higham[43,44], David M. Lambert[8], Eske Willerslev[1,15,45]†

(McColl et al. 2018)

Research question



(McColl 2018)

Principle Component Analysis

Unsupervised analysis = doesn’t require population affiliation

Reducing dimensionality while retaining max. information

Developed to deal with large datasets

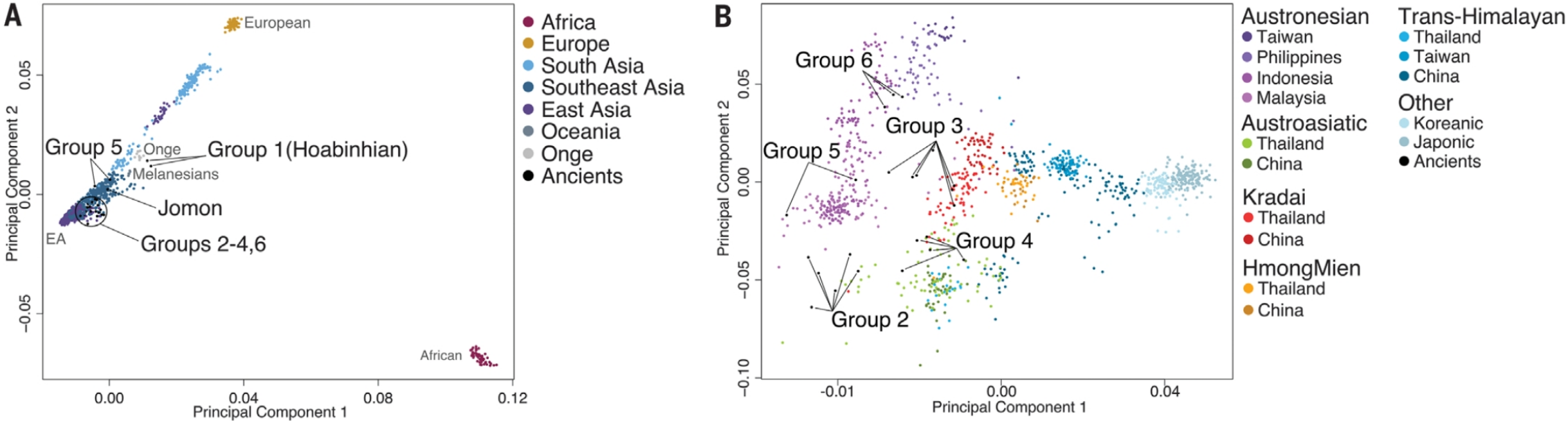
Transforms huge amount of SNPs’ allele frequencies into pairwise distance matrix

Extracts eigenvalues and independent principal components (PCs; most to least representative of variability)

Principle Component Analysis

Two components are treated as “coordinates” (PC1 vs PC2)

Each individual is plotted on the graph according to their coordinates.



Structure analysis

Unsupervised analysis

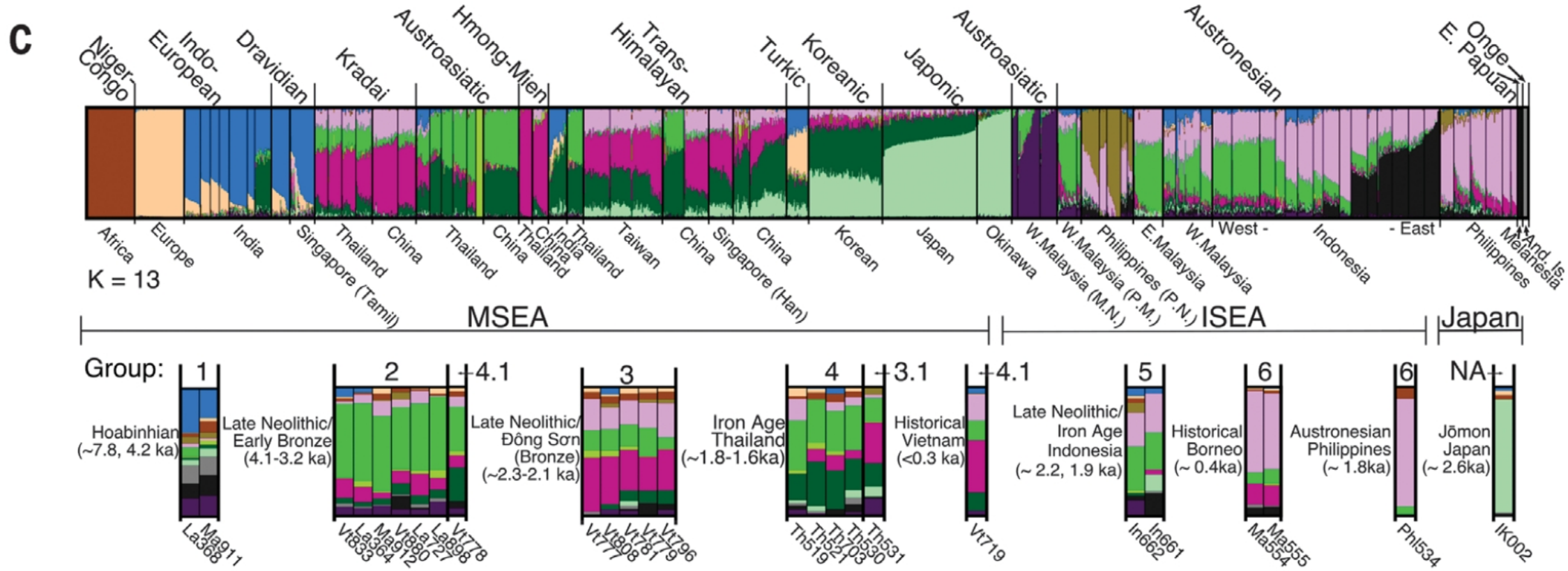
Assume that individuals are descended from multiple ancestral “populations” with different allele frequencies

Software assigns ancestry of each individual to one or more of these “populations” based on observed genotypes

Increasing number of ancestry components (K)

software: ADMIXTURE, DYSTRUCT, fastNGSadmix

Admixture - Structure



Austroasiatic

Hmong-Mien

Austronesian

broad East Asian

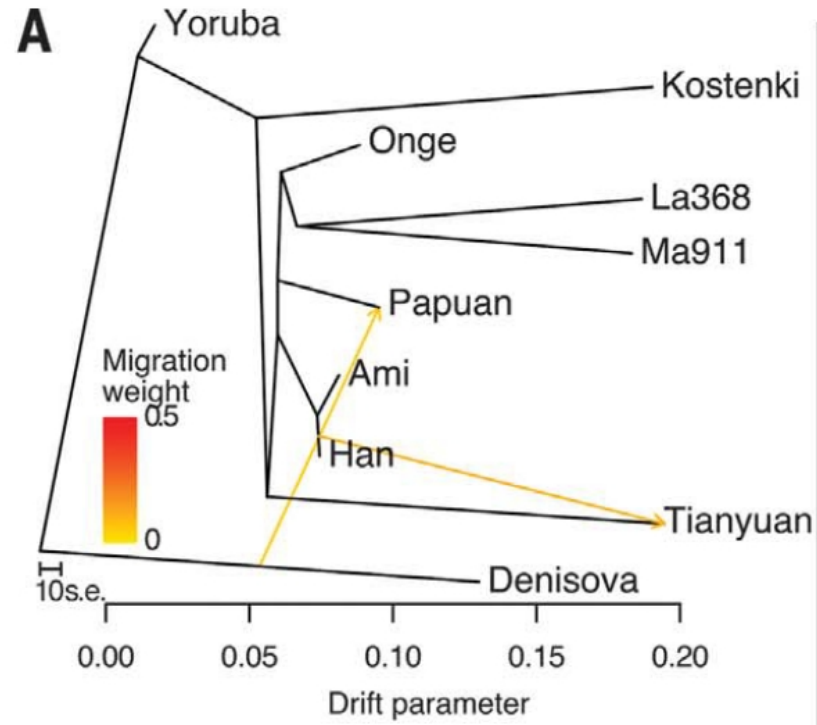
Admixture - TreeMix

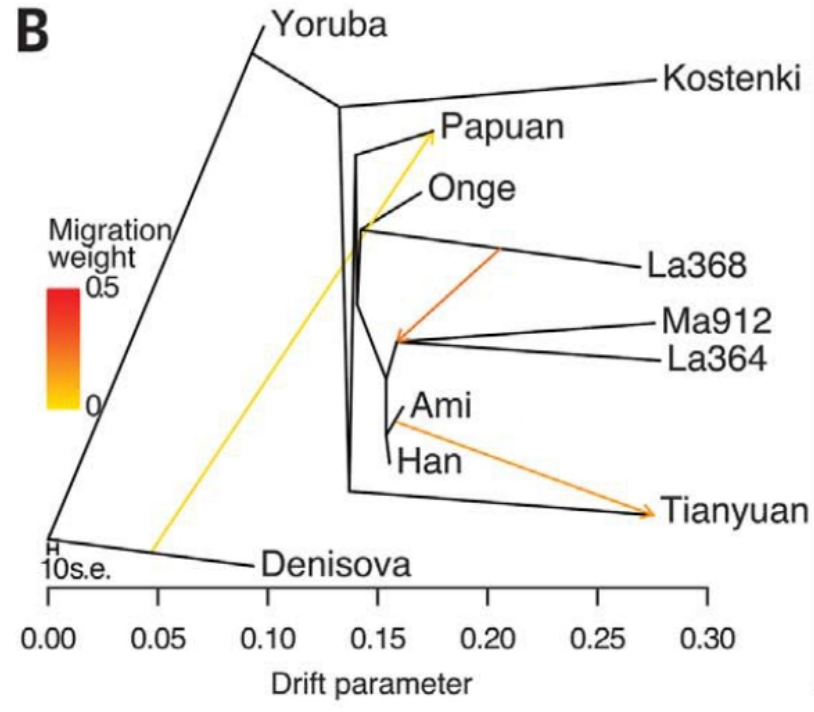
Estimates the historical relationships among populations using a graphical representation that allows both population splits and migration events.

Based on a similarity/covariance matrix

Maximum likelihood tree

Algorithm introduces “migration edges” to reduce residual variation





Pickrell & Pritchard (2012)

f-statistics (Admixtools)

F_{ST} = genetic diversity = $(H_T - H_S) / H_T = \{0;1\}$

f-statistics = shared genetic drift between 2/3/4 populations

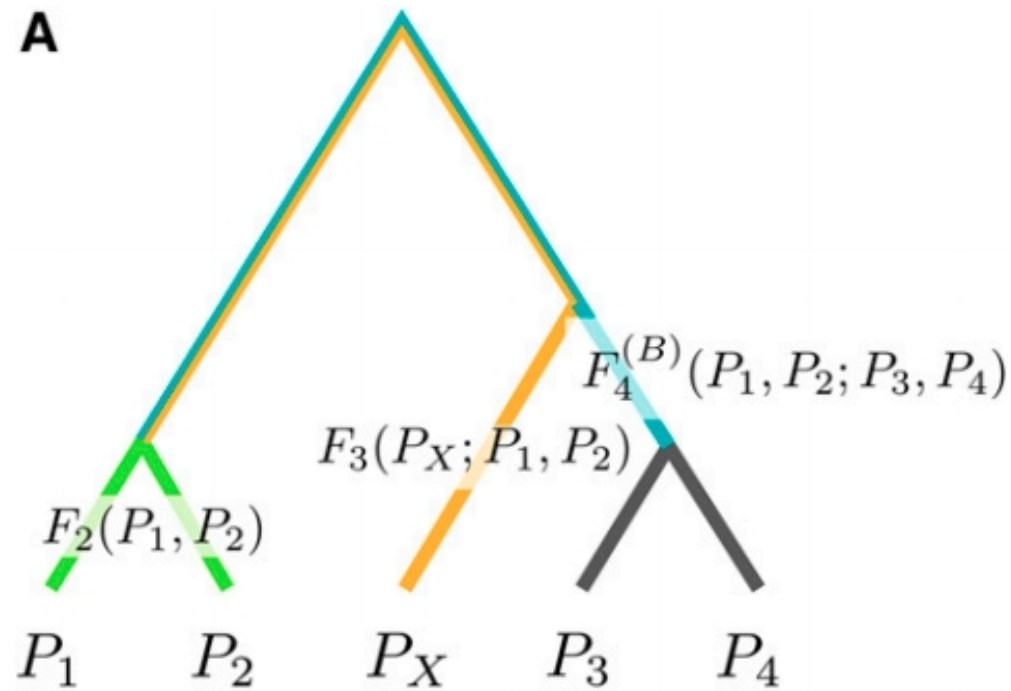
$f_2(P_1, P_2)$ = path from P_1 to P_2

$f_3(P_X; P_1, P_2)$ = path from P_X to node connecting all 3 populations

$f_4(P_1, P_2; P_3, P_4)$ = path from node (P_1, P_2) to node (P_3, P_4)

$f_4 > 0$ —> P_2 more similar to P_4

$f_4 < 0$ —> P_2 more similar to P_3



Reich et al. (2009)
Patterson et al. (2012)
Peter (2016)

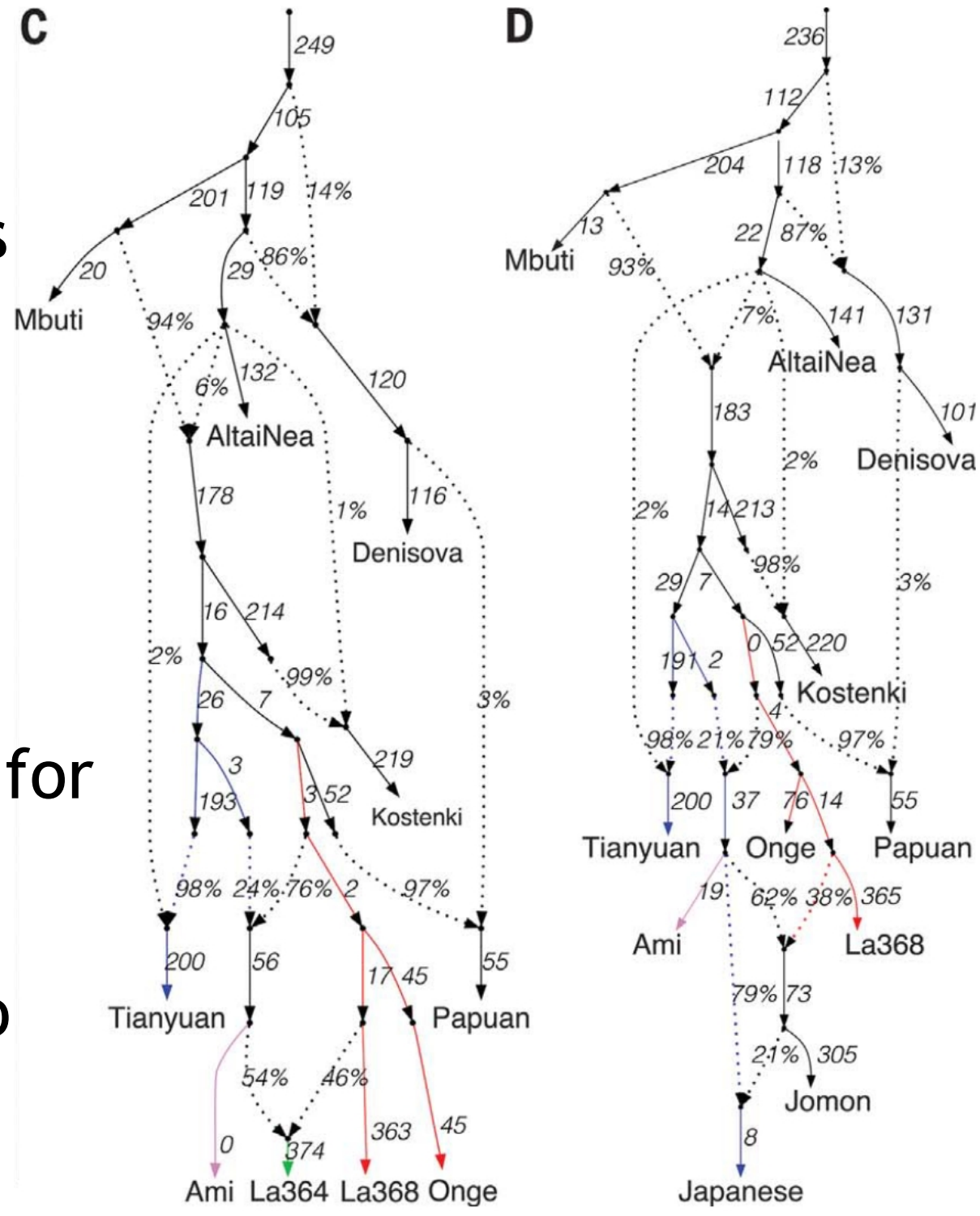
Admixture - qpGraph

Reconstruct genetic relationships between different groups allowing for the addition of admixture events

Based on a given topology

Estimates f2, f3 and f4-statistics for all pairs, triplet, and quadruples

Compares the expected values to the tested topology



Patterson et al. (2012)

Genetics & Culture

Ethnic group and linguistic group

SUMMER SCHOOL 2021
Doorway to Human History

Population Genetics

Population means "a group of people". The definition of "group" is dynamically depending on the research fields and questions.

Biological sense: abstract model in which individuals in one isolated group mate freely with others from the same group

Medicinal research: compare patients with healthy individuals.

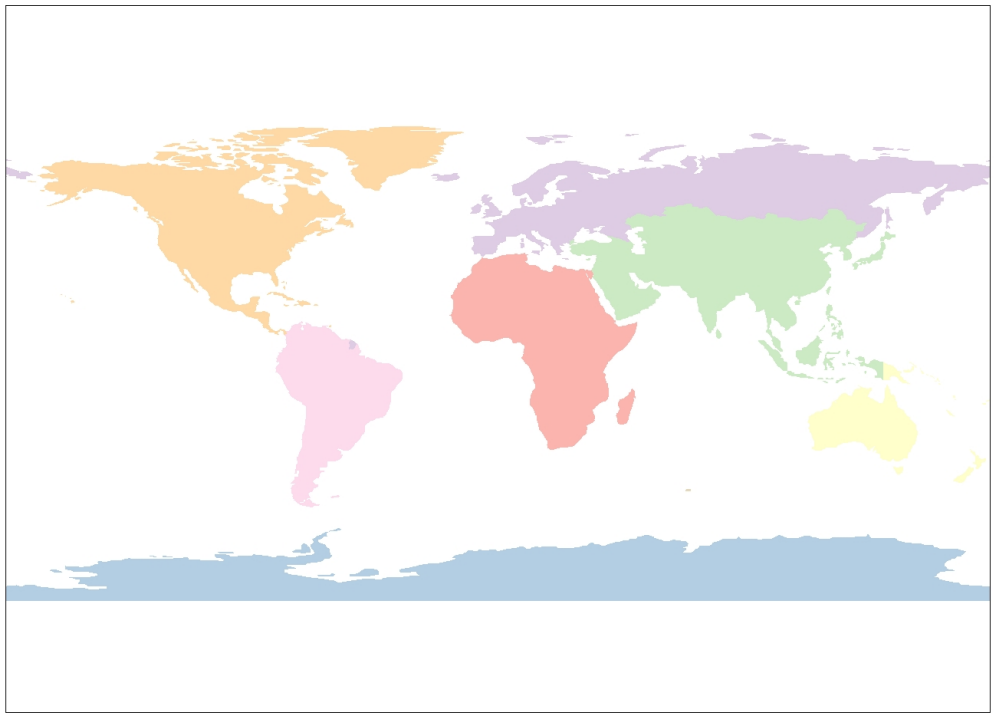
Animal study: compare between different phenotypes

Anthropological: group that shares a culture (shared activities, Behaviours, goods which are transmitted between individuals)

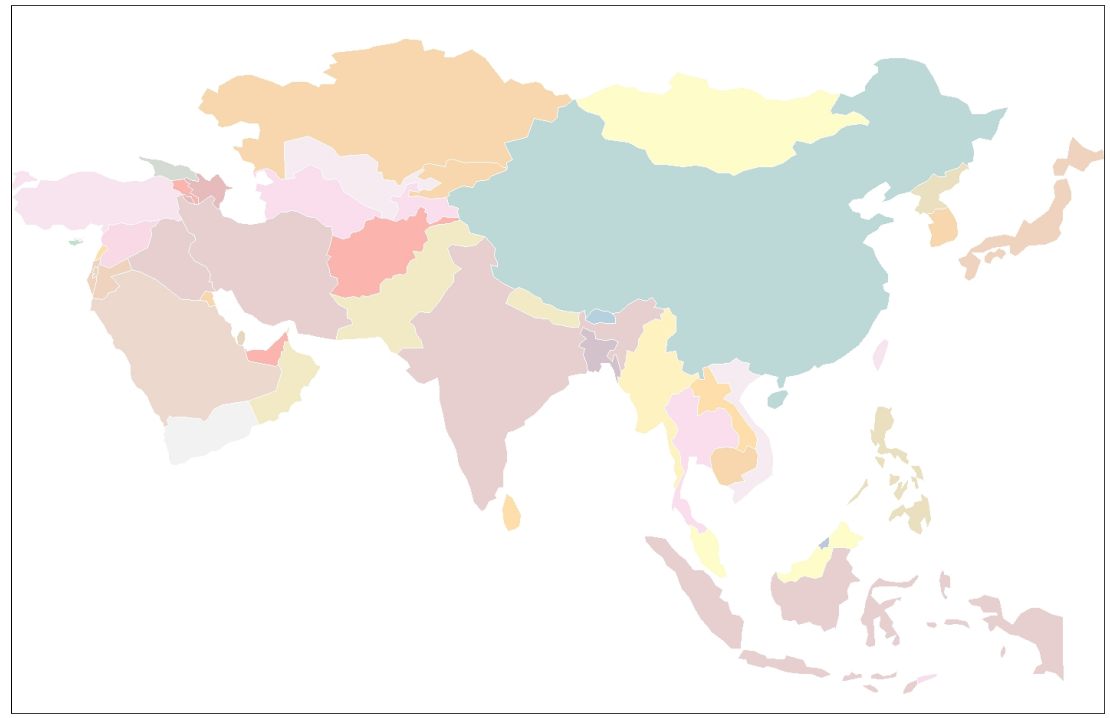
Genetic variation & linguistics

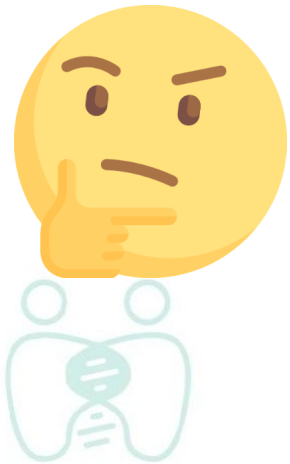
How to define a "group"?

Geography



Nationality





AND??

Genetic variation & linguistics

How to define ethnicity is a discussion started in 1960s and continuously discussed till today. Ethnicity commonly realised as a form of group identity based on shared cultural traits.

Primordial: ethnic identity is fixed and permanent.

Three important factors: biology, language, and culture

Situational: ethnicity is fluid and dynamic.

Three different theories: instrumentalist, materialist, constructionist

Genetic variation & linguistics

The genetic profile of each individual is fixed.

Each individual can acquire multiple languages, switch to a different religion, move to a new residence area, or have multiple nationalities.

Cultural practises can change without substantial migration of people (cultural diffusion, language replacement, elite

Naming genetic groups

Vocabulary differences between fields

Archaeological culture = genetic cluster = historical people?

Correlations exist and can be meaningful, but different underlying evidence & approaches

Group names need to be:
short, coherent, accessible, flexible, stable

Geographic-temporal system
eg. CentralGermany_BA or Stonehenge_2800_2200BP

Eisenmann et al. (2018)

Genetic variation & linguistics

Cultural label and language label aren’t the same!

Han population

▼

Northern Han

Central Han

Southern Han

▼

26 Sinitic languages (Glottolog)

AJHG
Volume 85, Issue 6, 11 December 2009, Pages 762-774

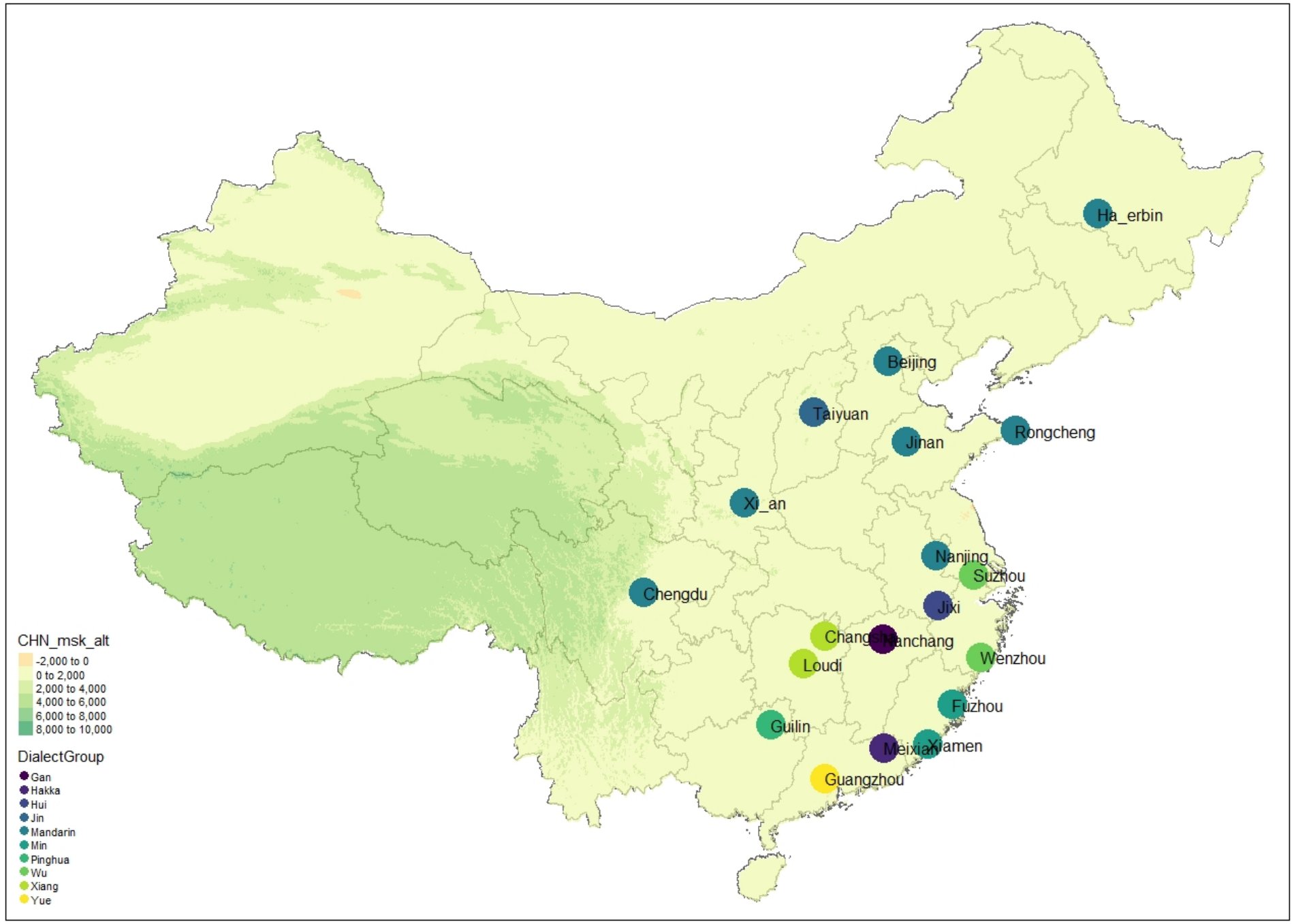
Article

Genomic Dissection of Population Substructure of Han Chinese and Its Implication in Association Studies

Shuhua Xu [1, 2, 7], Xianyong Yin [4, 7], Shilin Li [3], Wenfei Jin [1, 2], Haiyi Lou [1, 2], Ling Yang [1, 2], Xiaohong Gong [3], Hongyan Wang [3], Yiping Shen [3, 5], Xuedong Pan [3], Yungang He [1, 2], Yajun Yang [3], Yi Wang [3], Wenqing Fu [3], Yu An [3], Jiucun Wang [3], Jingze Tan [3], Ji Qian [3] ... Li Jin [1, 2, 3, 6]

(Xu 2009)

Genetic variation & linguistics



Líu, L.; Wáng, H.; Bǎi, Y. (2007): Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí 现代汉语方言核心词·特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Nánjīng: Fènghuáng.

Naming linguistic groups

Choose the population labels wisely.

- Use labels that are based on the same definition.

- Think about the potential of the dataset.

Give detail descriptions about the participants and the area.

- How many languages are spoken by the selected population?

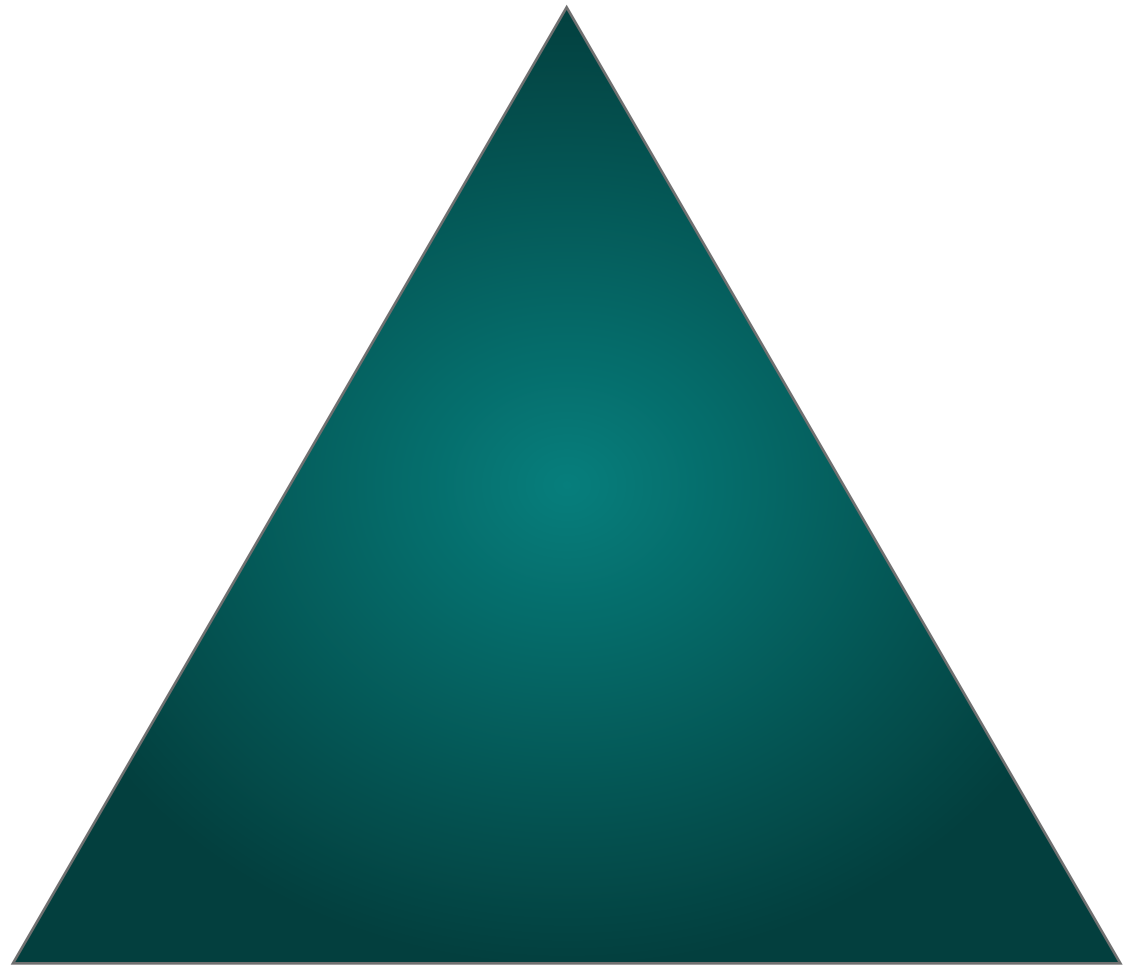
- Longitude and Latitude.

Multidisciplinary studies

Ethnic group and linguistic group

Evolution and co-evolution

Human genome



Culture

Language

Genetic variation & culture

Direct influence
- lactase persistence (strong selection in populations that drink milk)
- higher number of amylase genes (starch digestion)

Indirect influence:
- patrilocality (geographically separated populations show bigger NRY difference than mtDNA & autosome
- social systems eg. caste mobility

Genetic variation & Language

vs.

Human genome

Language family

Genetic variation & linguistics

SNPs vs. Lexicon

40 - 250 core vocabulary

Decide words' etymology (whether they are coming from the same proto-word or not).

Calculate the "distance" between language subgroups according to the amount of words that have shared ancestors.

Compare the two distance matrices.

Compare the two trees.

Obstacles

1. Labelling issues

2. Various types of contact-induced changes trigger language diversification.

3. The linguistic analysis does not yet incorporate the admixture modelings into the framework.

Anything else?

Thank you!

TreeMix workshop: next Tuesday @ 11 am

Case Study Southeast Asia: next Thursday @ 11 am

Mei-Shin Wu

✉ mei_shin_wu@eva.mpg.de

@meishinwu

Selina Carlhoff

✉ selina_carlhoff@eva.mpg.de

@selina_carlhoff

SUMMER SCHOOL 2021 Doorway to Human History

References

[1] Jolliffe Ian T. and Cadima Jorge (2016). Principal component analysis: a review and recent developments *Phil. Trans. R. Soc.* A.37420150202. http://doi.org/10.1098/rsta.2015.0202

[2] Stoneking Mark (2016). *An Introduction to Molecular Anthropology*. John Wiley & Sons, Inc., Hoboken, New Jersey.

[3] Baumann Timothy (2004). Defining ethnicity. The SAA archaeological record 4.4 (2004): 12-14.

[4] Xu, S., Yin, X., et al. (2009). Genomic Dissection of Population Substructure of Han Chinese and Its Implication in Association Studies. *The American Journal of Human Genetics*, *85(6)*, 762–774. https://doi.org/10.1016/j.ajhg.2009.10.015

References

[5] Líu, L.; Wáng, H.; Bǎi, Y. (2007): Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí 现代汉语方言核心词·特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Nánjīng: Fènghuáng.