ゲノムデータからの集団サイズの推定

河合洋介

第一回統合生物考古学セミナー
2023年12月10日　東京大学理学部２号館

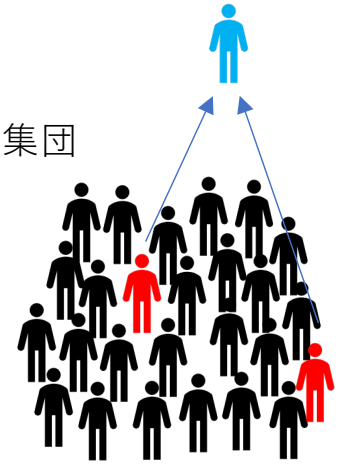
遺伝的多様性と集団史

- 現在の生物集団の遺伝的多様性は過去の集団史(demographic history)を反映している
- Higher diversity, larger population
  - 過去の集団サイズの変動は現在の集団の遺伝的多様性に痕跡を残しているはずである。

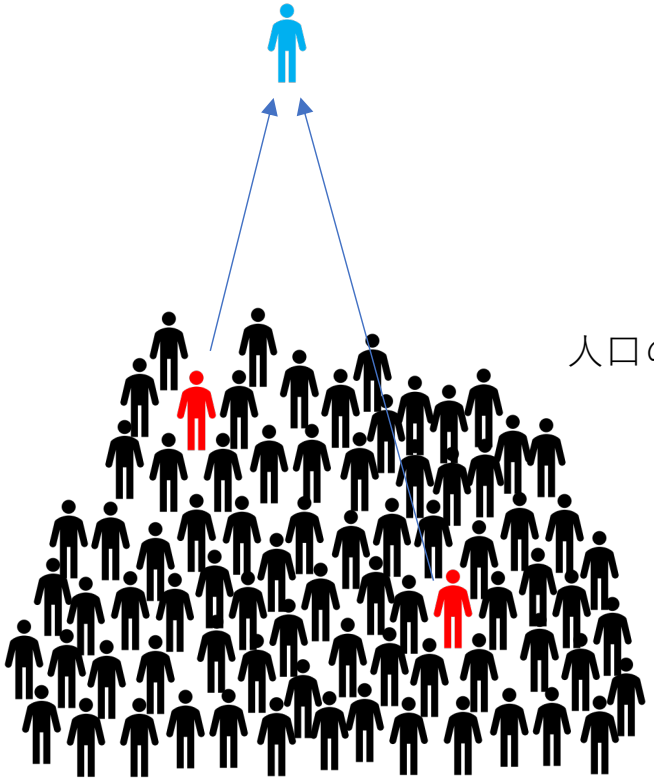
ゲノムデータから過去の集団サイズの変動を推定する

ゲノムDNAから集団史を推定する

人口の少ない集団



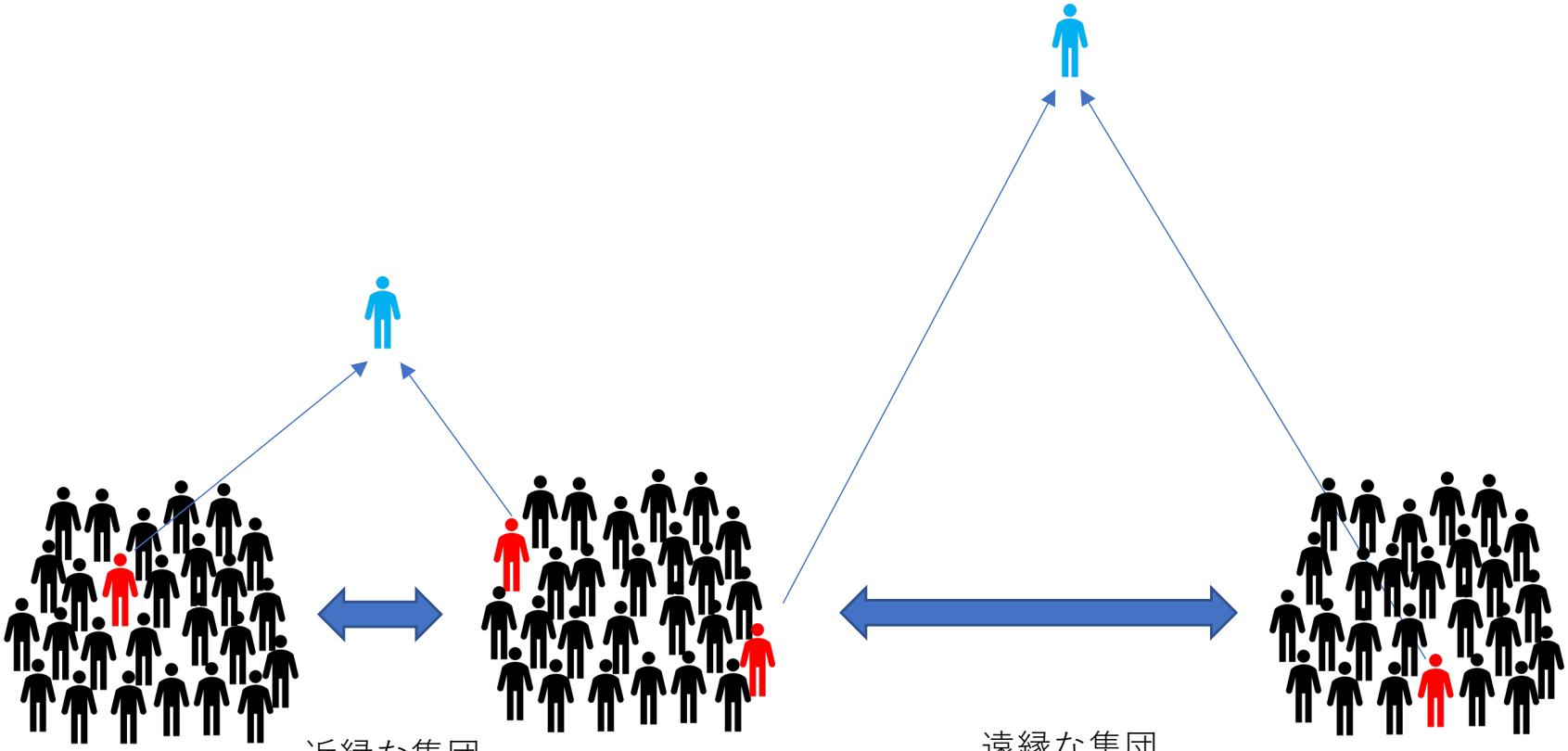
共通祖先までの時間（世代）が短い
=多様性が小さい



人口の多い集団

共通祖先までの時間（世代）が長い
＝多様性が大きい

ゲノムDNAから集団史を推定する

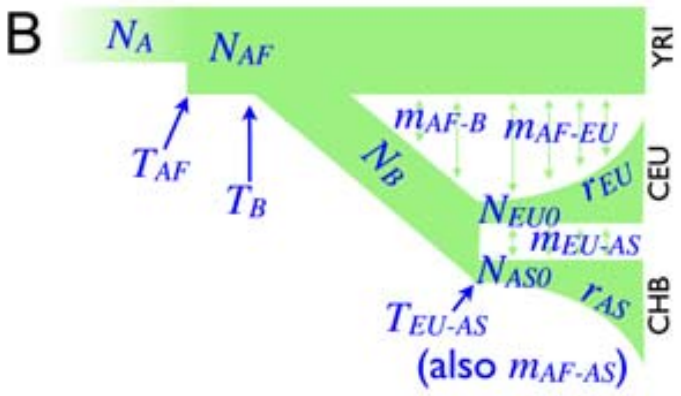


近縁な集団
共通祖先までの時間（世代）が短い
＝多様性が小さい

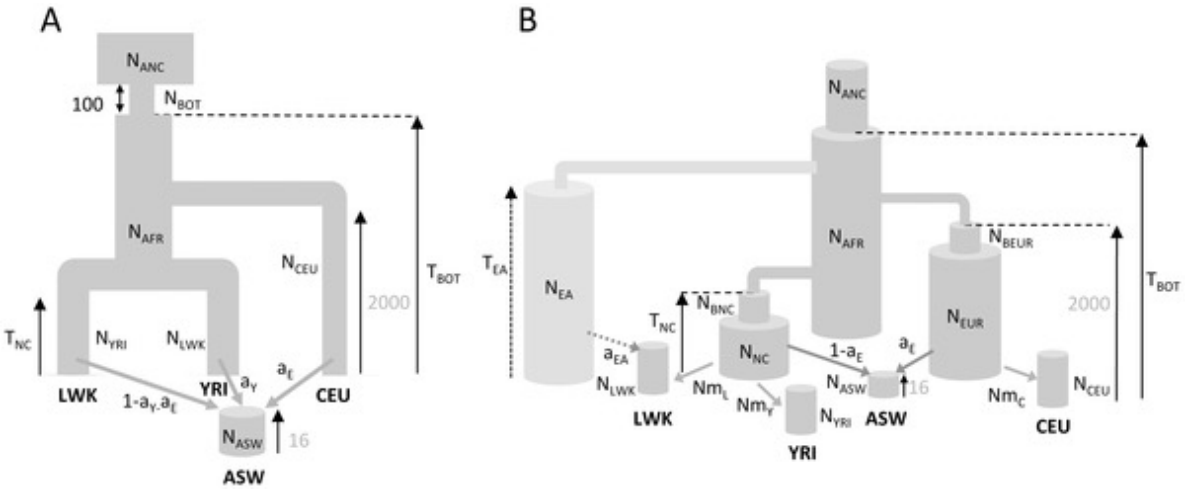
遠縁な集団
共通祖先までの時間（世代）が長い
＝多様性が大きい

人口動態の推定手法

- パラメータを用いる方法
  - 集団サイズ、移住率、増加率などをパラメータとしたモデルを建てて、ゲノムデータから推定する方法



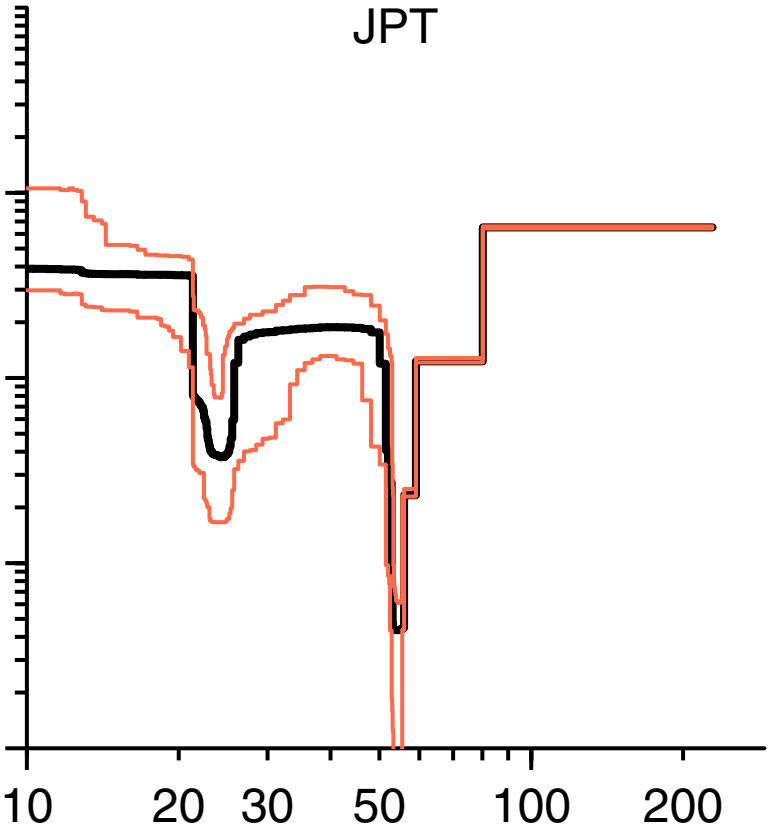
Harris and Nielsen. 2013. PLoS Genet.



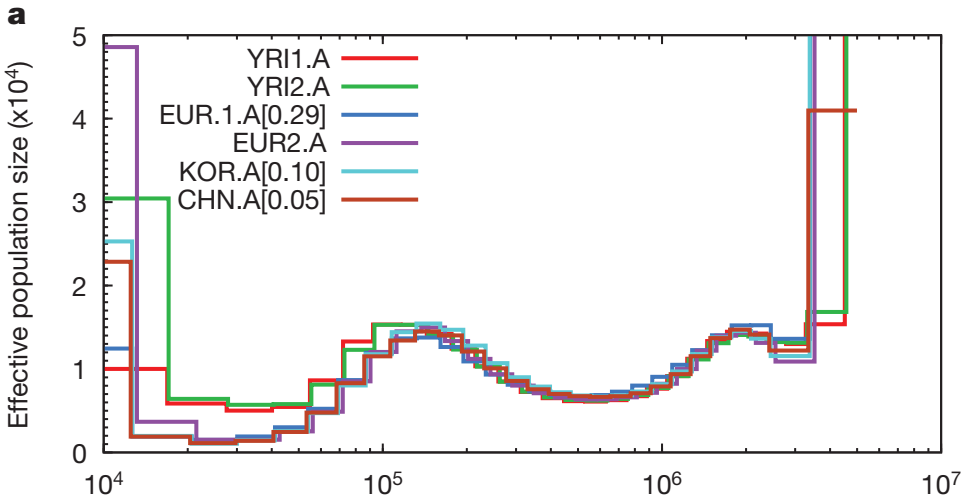
Excoffier et al. 2013. PLoS Genet.

人口動態の推定手法

- パラメータに頼らない方法



Stairway plot(Liu and Fu. 2015. Nat. Genet.)



Estimate from individual genomes
of 1000 Genome Project
Li and Durbin, 2011. Nat. Genet.

Poisson Random Field Model

Wright-Fisherモデルの拡散近似

$$\frac{\partial}{\partial t} f(q,t) = \frac{1}{2}\frac{\partial^2}{\partial q^2}\{q(1-q)\rho(t;\lambda) f(q,t)\} - \frac{\partial}{\partial q}\{Sq(1-q)f(q,t)\}$$

集団サイズの時間変化

q:allele frequency
S: selection coefficient
ρ(t;λ): population size

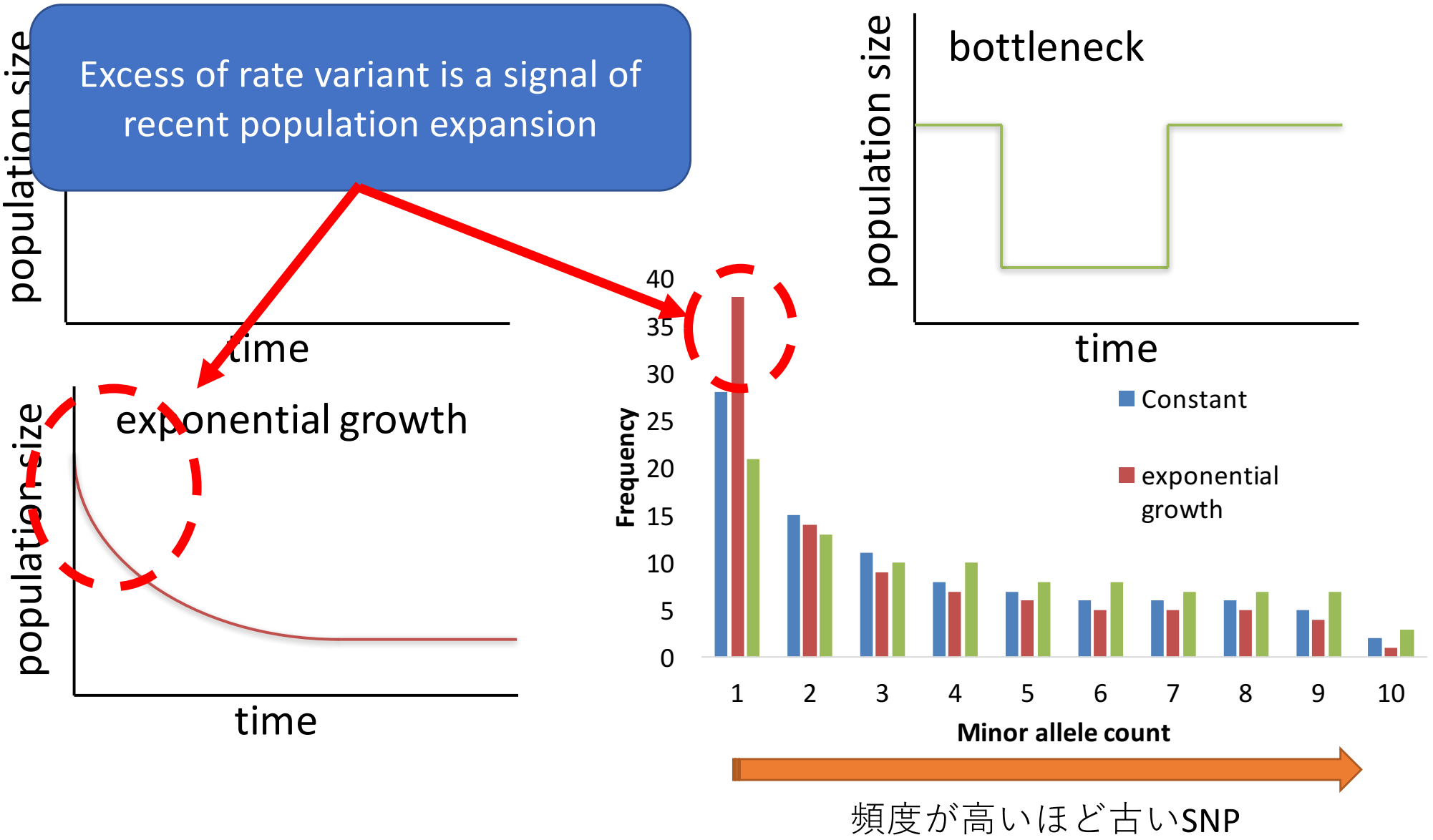
Probability function of allele frequency in sampled data

$$p(n,i;\Theta) = \int_0^1 \binom{n}{i} q^i (1-q)^{n-i} f(q,t;\Theta) dq$$

n: number of chromosomes
i: number of mutants
Θ: parameters (e.g. population sizes)

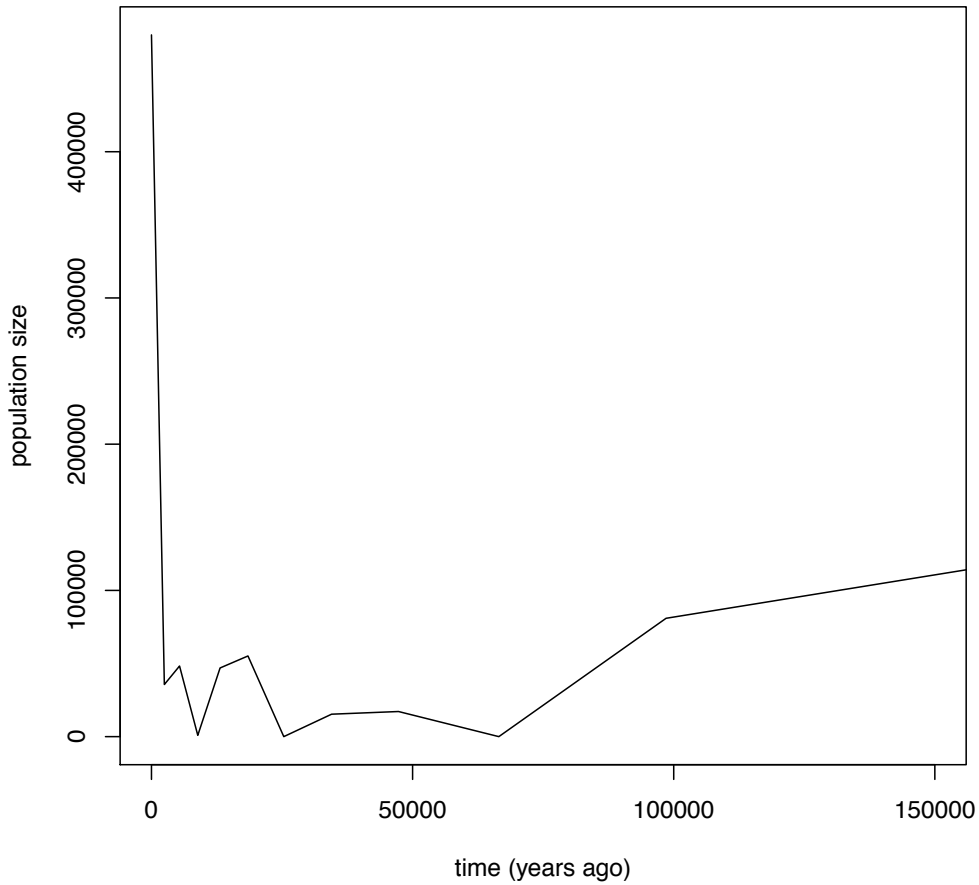
Maximum likelihood Estimation of parameters

Site Frequency Spectrum(SFS)



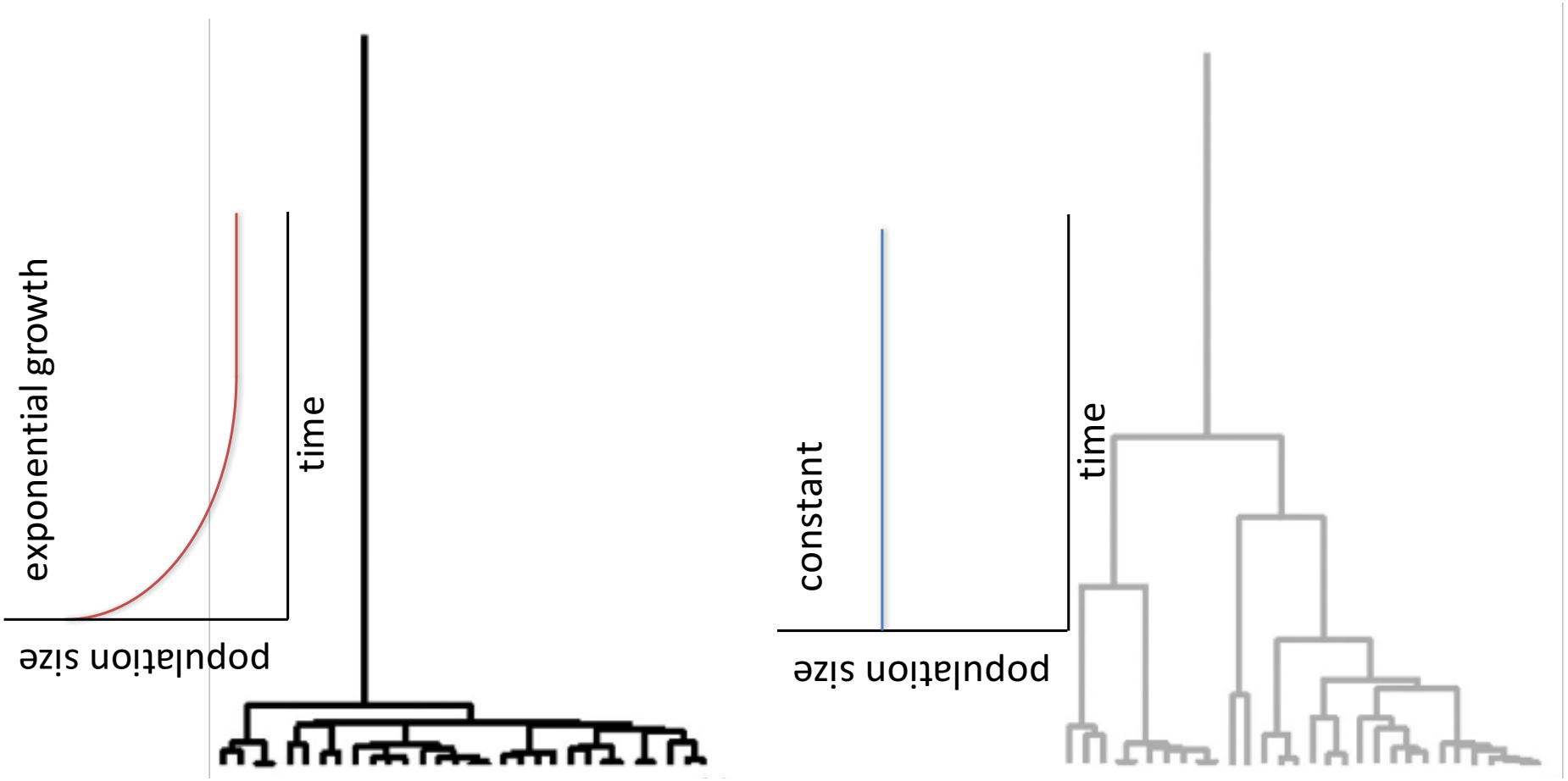
SFSからの集団サイズ変動推定

- 例
  - ∂a∂i (Harris and Nielsen. 2013. PLoS Genet.)
  - Stairway plot (Liu and Fu. 2015 Nat Genet.)
- pros
  - Only necessary for very simple statistics.
- cons
  - Large number of sample is necessary.
  - Inference is affected by SNP ascertainment.
    - SNP array data are unavailable.



Estimate from site frequency spectrum (SFS) of 1,070 individuals
Nagasaki et al. 2015 Nat. Comm.

遺伝子系図と集団サイズの関係



Genealogy estimate results in demographic inference

合祖モデル

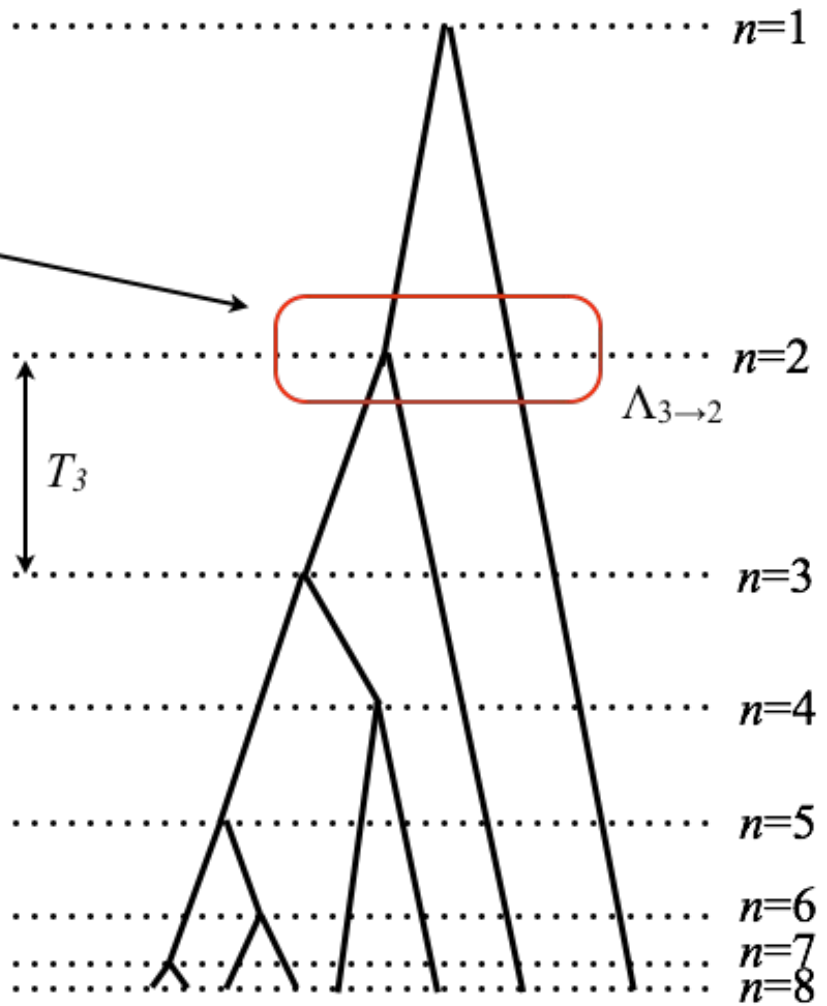
合祖のパターンの確率$P(\Lambda_{n\to n-1})$=

$$\binom{n}{2}^{-1} = \frac{2}{n(n-1)}$$

合祖時間の確率分布$P(T_n)$=

$$\frac{n(n-1)}{4N} exp(-\frac{n(n-1)}{4N} T_n)$$

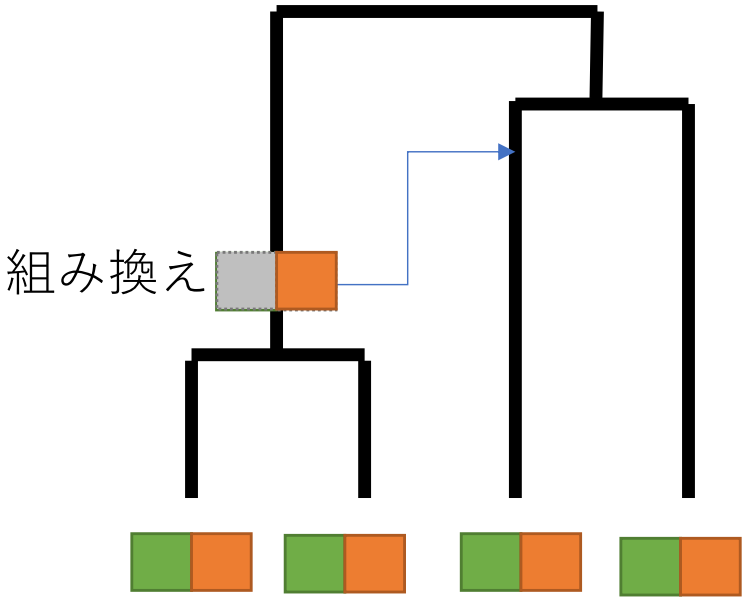
$$P(G;\Theta(t)) = \left(\frac{2}{\Theta(t)}\right)^{n-1} exp\left(\sum_{i=2}^{n} -\frac{i(i-1)}{\Theta(t)} t_i\right)$$



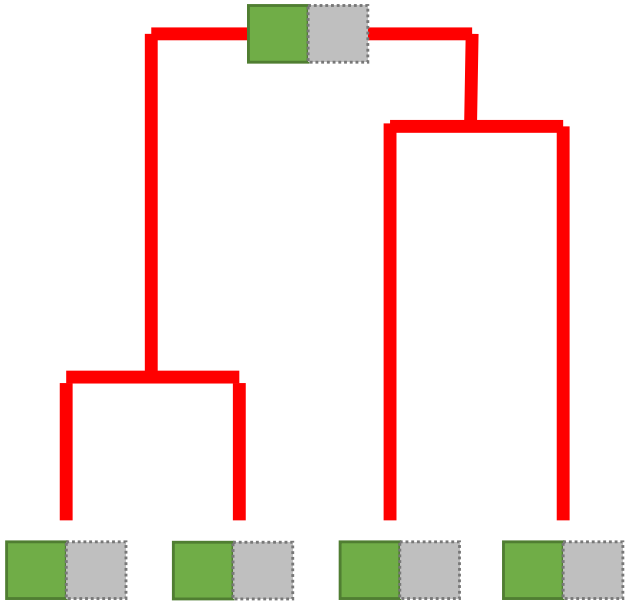
合祖モデルを使った集団サイズ変化の推定方法

- Examples
  - Lamarc (Kuhner et al.2006 Bioinformatics)
  - Bayesian Skyline Plot (Drummond et al. 2005. MBE)
- pros
  - Fully analysis of genetic diveristy
- cons
  - Hard to extend genome-wide analysis

組み換えと遺伝子系図

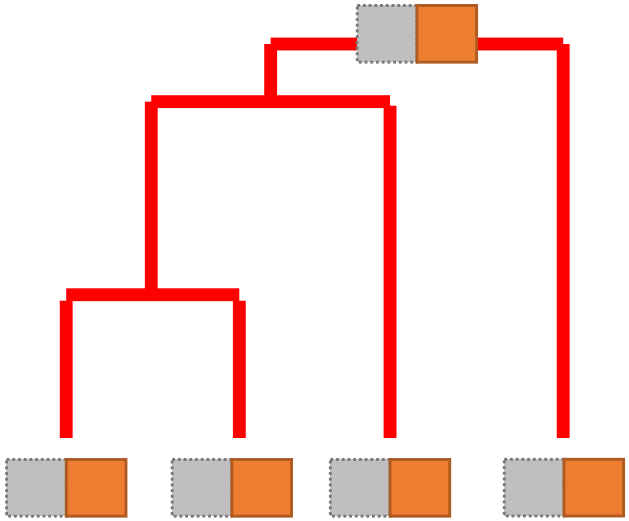


遺伝子座1



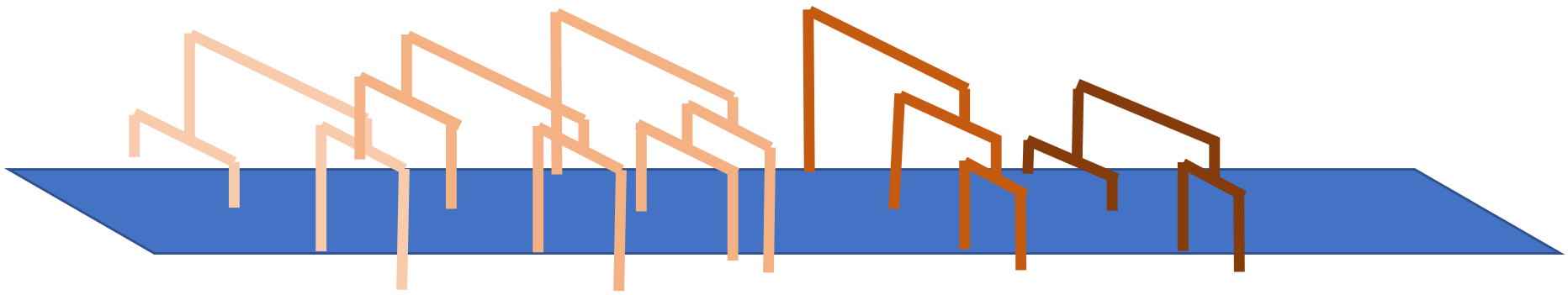
組み換えによって隣り合う遺伝子座間でも異なる遺伝子系図が生じる

遺伝子座 2



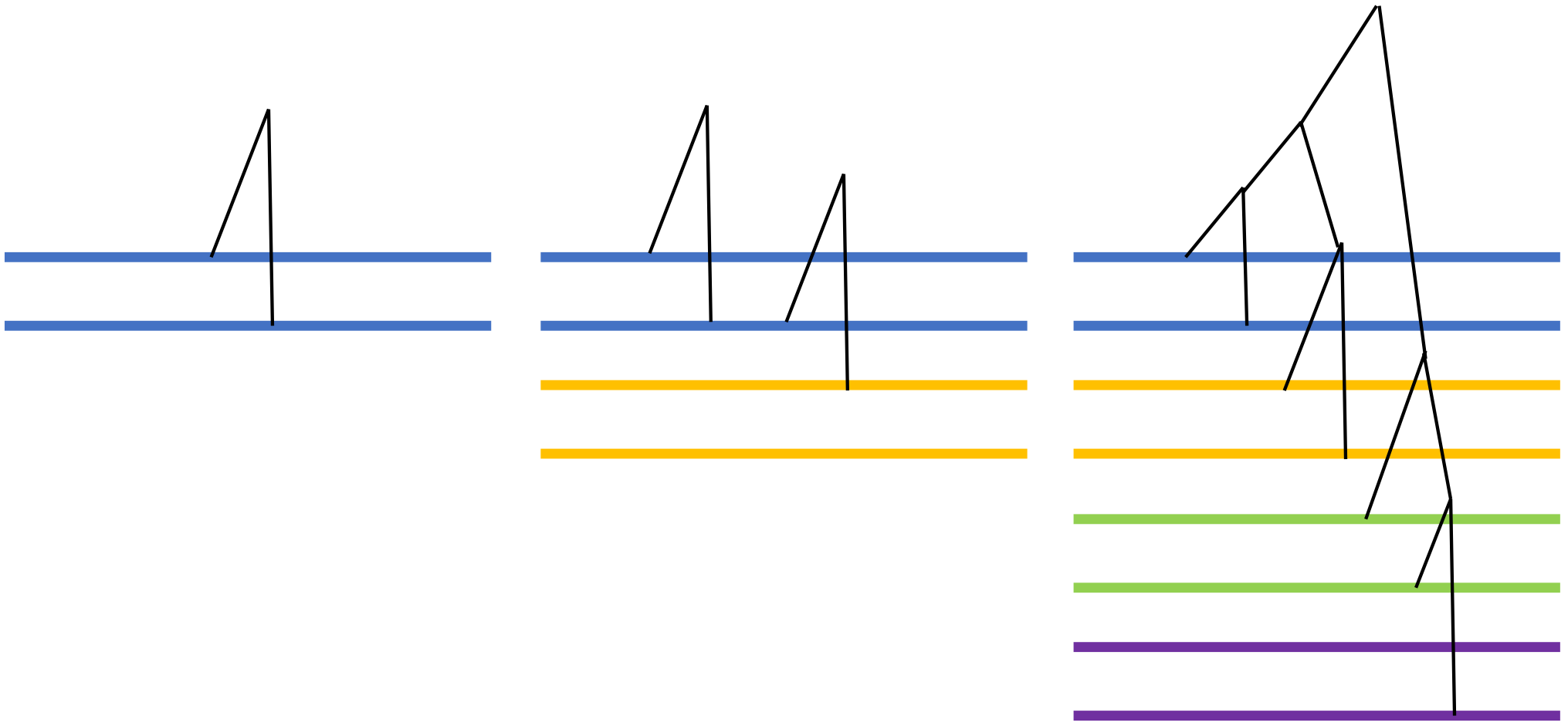
遺伝子座1　遺伝子座 2

- 二倍体生物のゲノムは多数の系統樹の集合によって表される
- 近似法が必要..
  - Sequential Markovian Coalescent (SMC)



SMC methods

- PSMC: Pairwise Sequentially Markovian Coalescent (Li and Durbin, 2011)
- MSMC: The multiple sequentially Markovian coalescent (Schiffels and Durbin, 2014)
- SMC++ (Terhorst et al., 2017)
- Relate (Speidel et al. 2019)



PSMC

集団サイズ

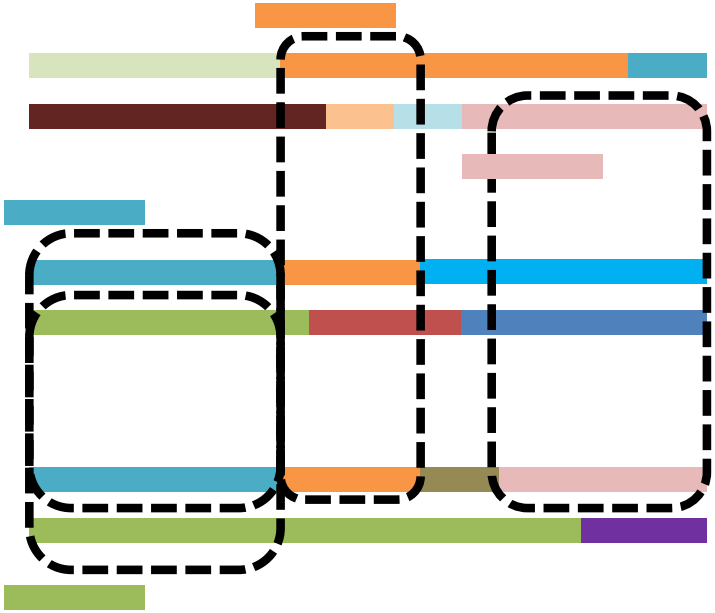
MSMC

集団サイズ
分岐時間

RELATE

いろいろ

IBD(Identical by descent) Sharing



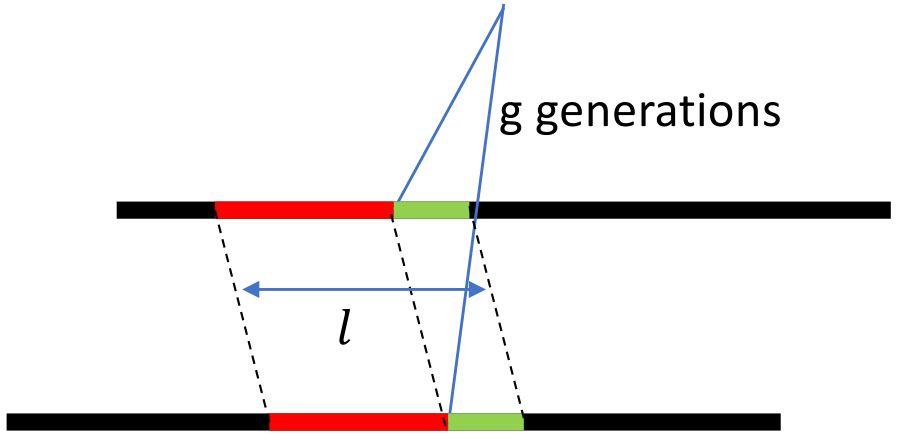
- Pair of identical genomic segment inherited it from a common ancestor without recombination

Length distribution of IBD segment in cM

$$f(l;g) = \int_0^l f_L(x;g) f_R(l-x;g)\,dx$$

$$= \int_0^l \frac{g}{50} e^{\frac{g}{50}x} \frac{g}{50} e^{\frac{g}{50}(l-x)}\,dx$$

$$= \left(\frac{g}{50}\right)^2 l e^{-\frac{g}{50}l}$$



$f_L(x;g) = \frac{g}{50} e^{\frac{g}{50}x}$

$f_R(x;g) = \frac{g}{50} e^{\frac{g}{50}x}$

exponential with mean 50/g

(average distance between recombination per meiosis = 100 cM)

Palamala et al. 2012. AJHG

IBD共有の応用

- 応用例
  - IBDNe (Browning and Browning. 2015. AJHG)
- 利点
  - 近年の集団動態推定の精度が高い
    - 過去に遡るほど精度が下がる
  - SNPアレイデータも利用可能
- 欠点
  - 多数サンプルが必要

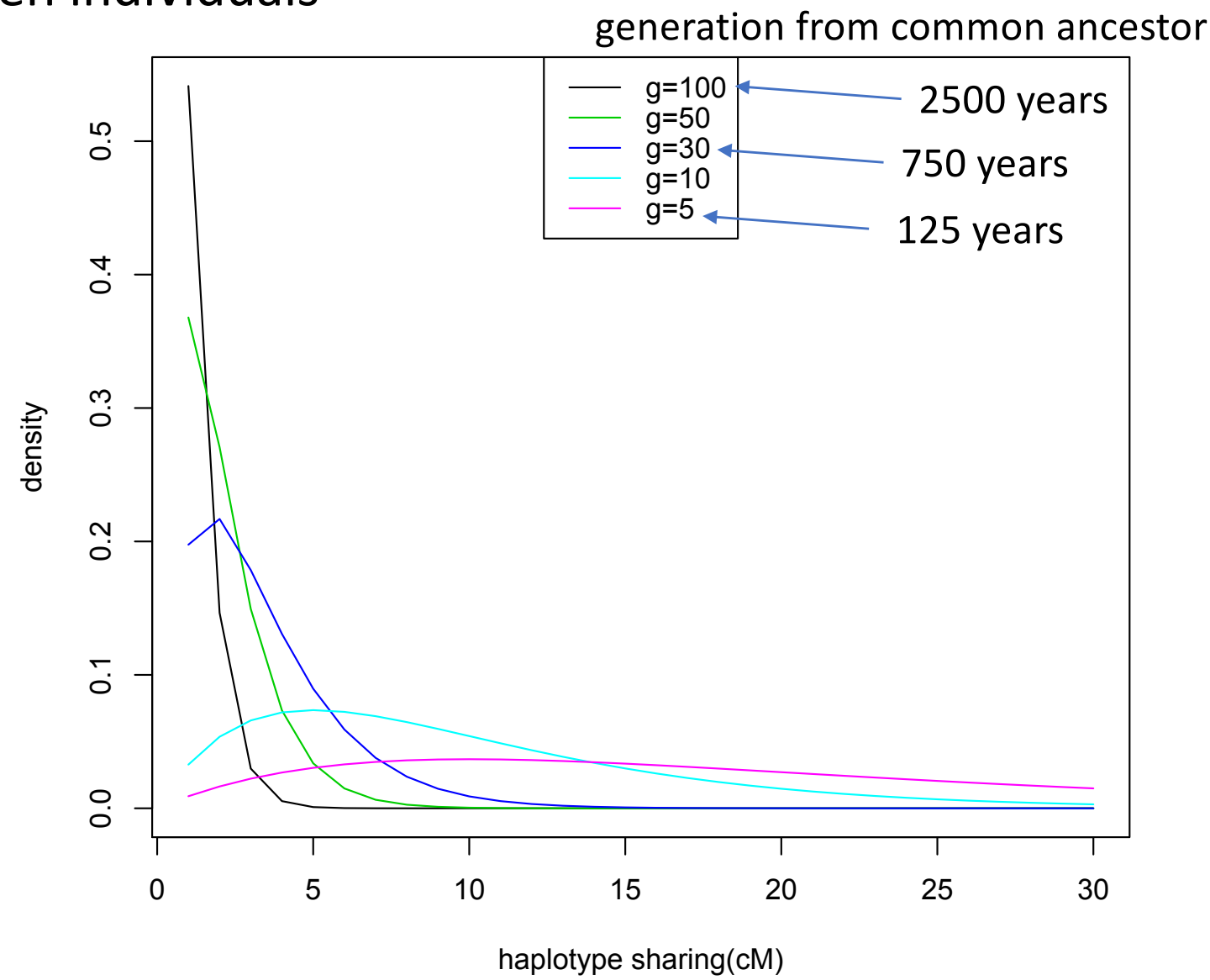
demographicモデルθにおけるIBD共有長の分布

$\int_u^v p(l|\theta)dl = \sum_{g=1}^{\infty} p(g|\theta) \int_u^v f(l;g)dl$　　u ~ v CMのIBDが見つかる確率

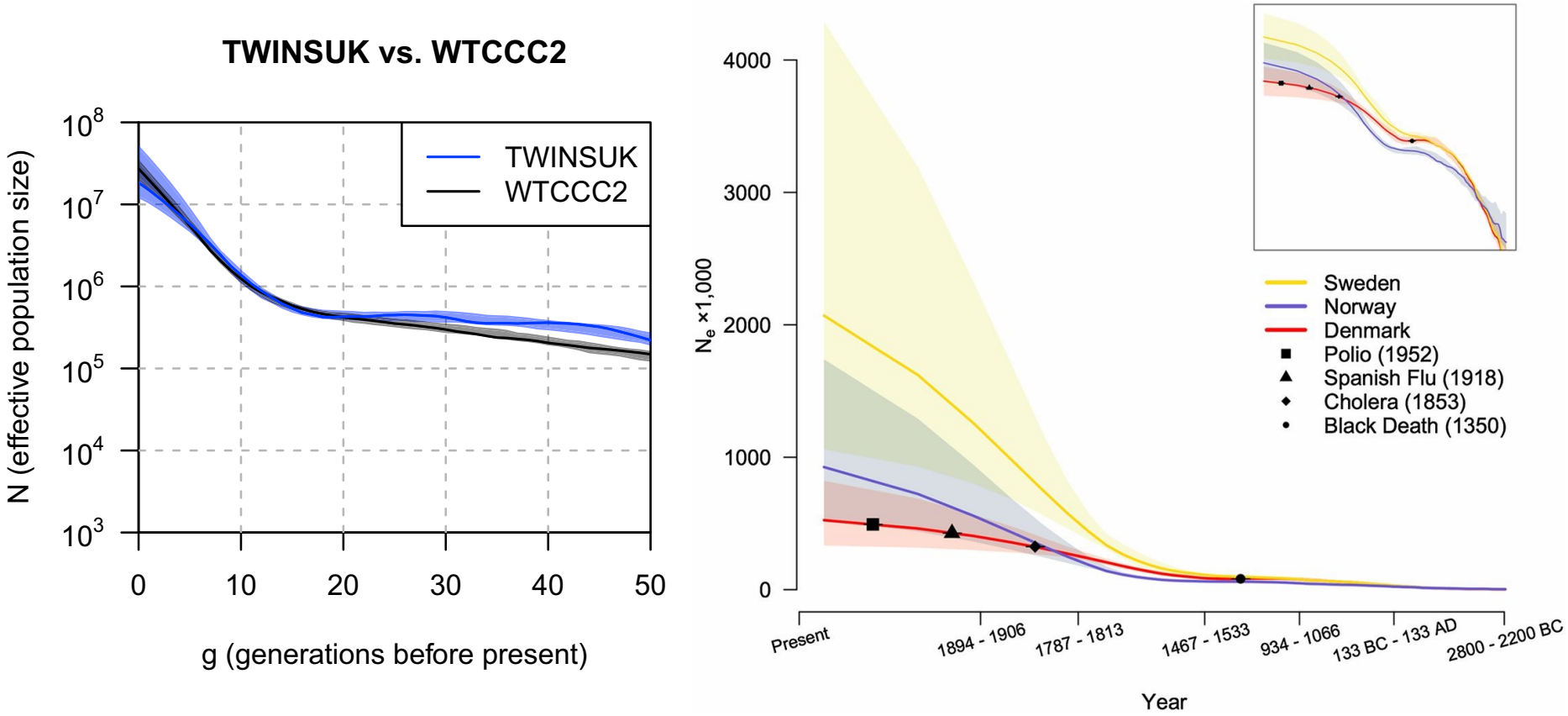
$$\int_u^v f(l;g)dl = \int_u^v (\frac{g}{50})^2 l e^{-\frac{g}{50}l} dl = \left(\frac{tu}{50} + 1\right) e^{-\frac{tu}{50}} - (\frac{tv}{50} + 1) e^{-\frac{tv}{50}}$$

$$p(g|\theta) = \frac{1}{N(g,\theta)} \prod_{j=1}^{g} (1 - \frac{1}{N(g,\theta)})$$

Distribution of shared length (cM) of each IBD segment between individuals

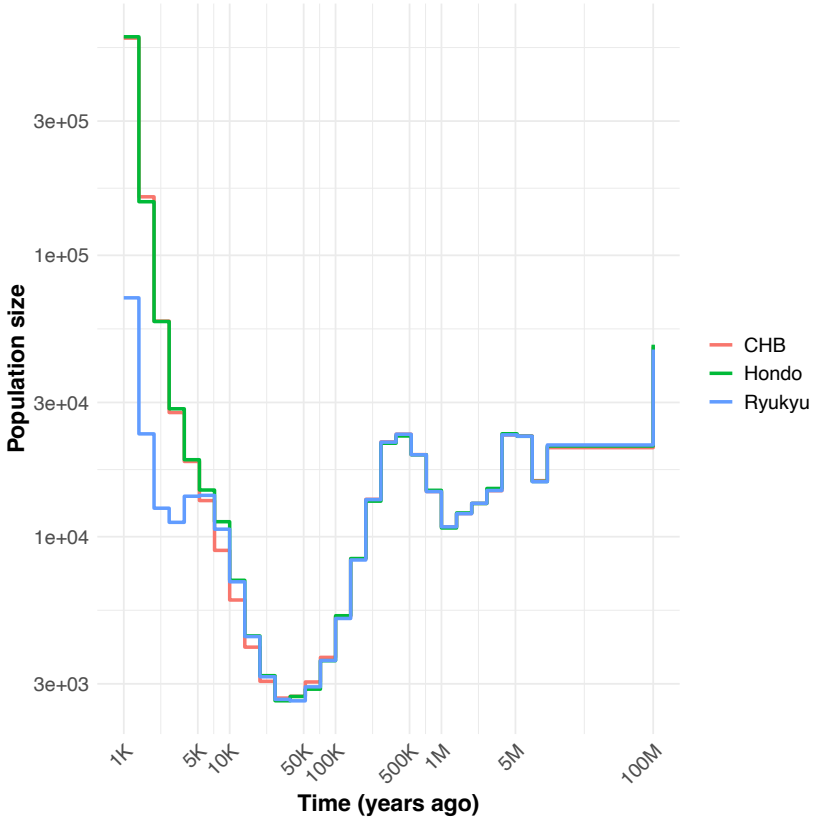


Length of IBD sharing is sensitive to time to common ancestor in terms of generation.

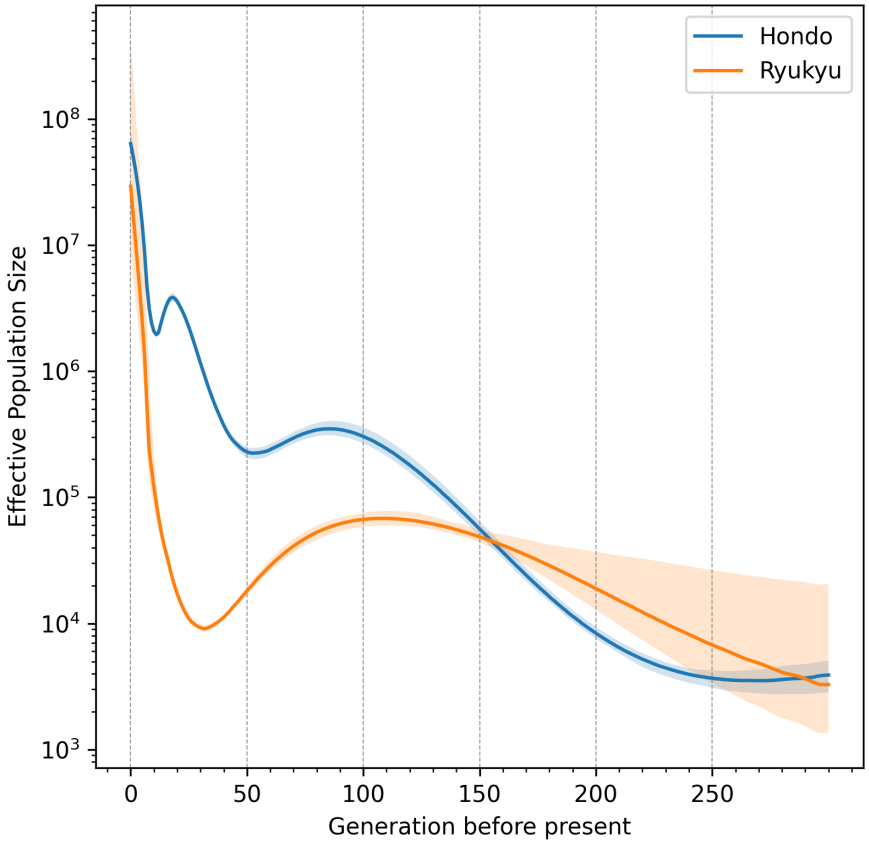


IBDNe によるイギリス人（左; Browning and Browning. 2015. AJHG ）、
北欧（右; Athanasiadis et al. 2016. Genetics）の集団動態推定

日本人の集団サイズ推定



RELATEによる推定



IBDNeによる推定

Kawai et al. 2023. PLOS Genetics