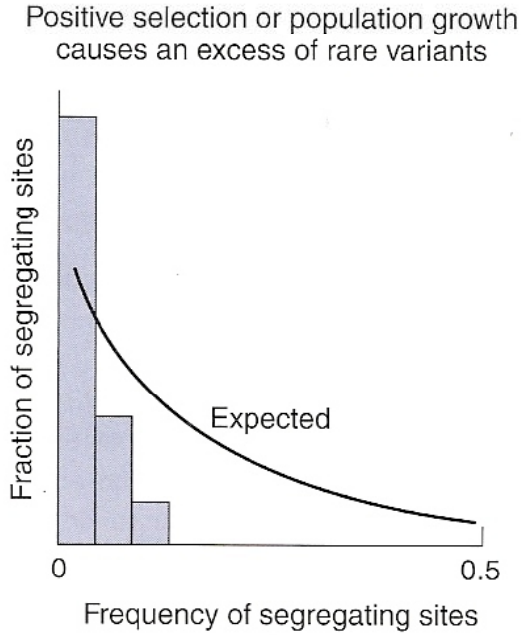
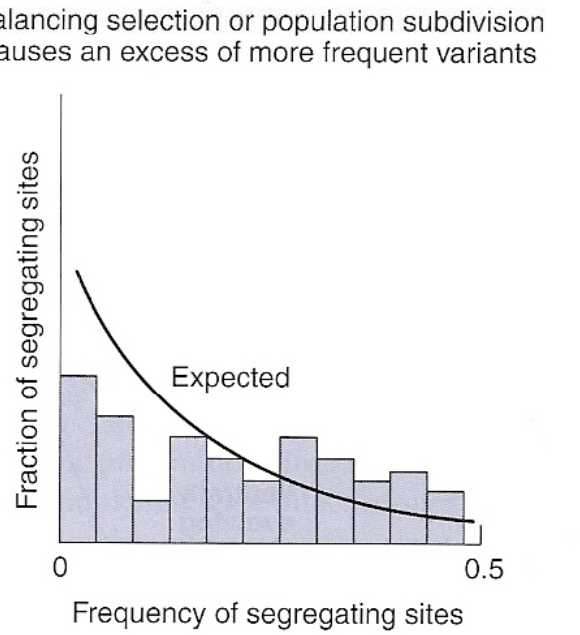
# ESTATÍSTICAS PARA INFERIR PADRÕES DEMOGRÁFICOS E SELECÇÃO

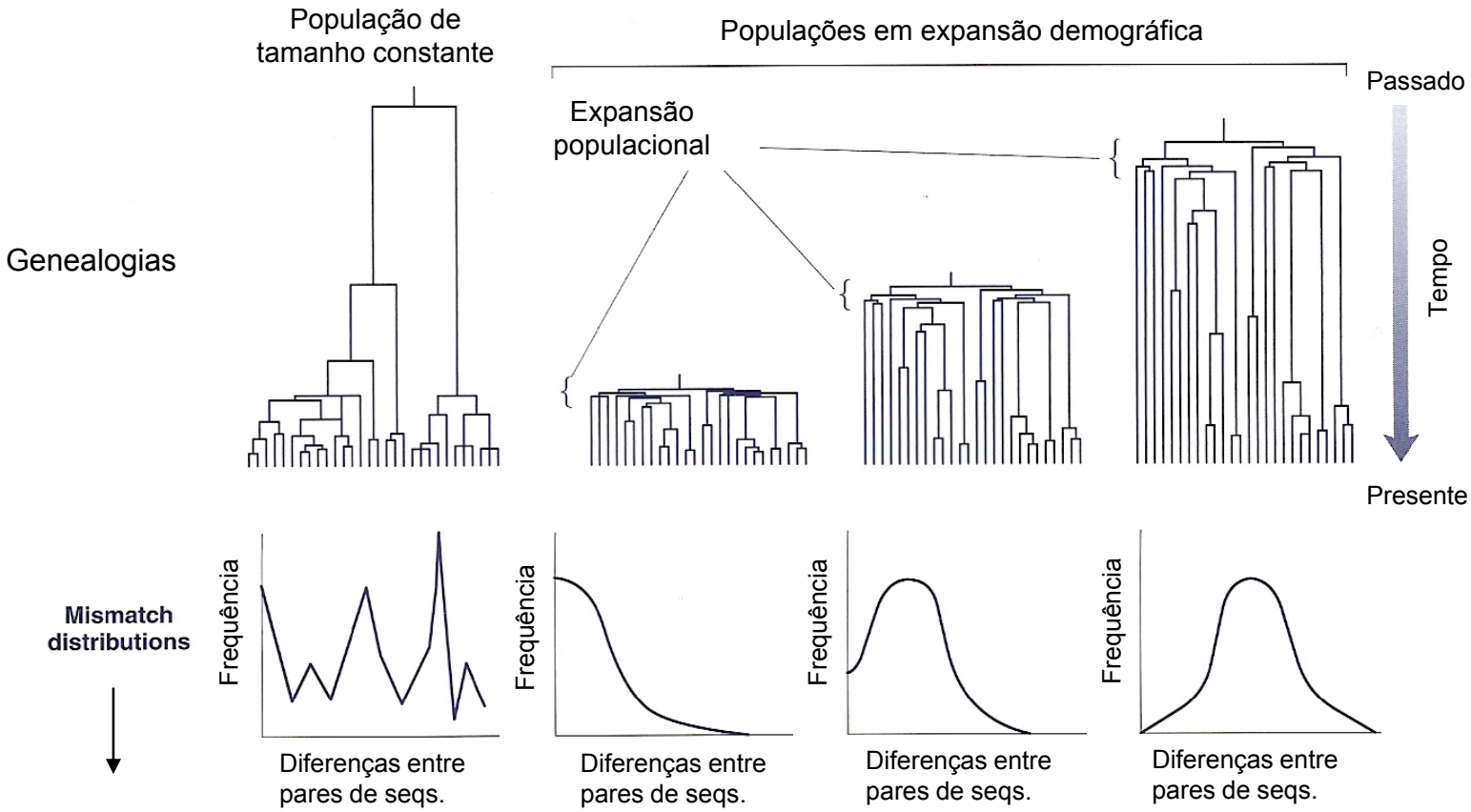
## Inferências demográficas e selectivas

Os fenómenos demográficos (expansão ou redução do efectivo populacional, subdivisão, migração) e selectivos (selecção positiva, balanceada, etc.) deixam marcas na variabilidade genética das populações.
A dedução da magnitude destes fenómenos faz-se por comparação com um cenário teórico (normalmente uma situação de equilíbrio)



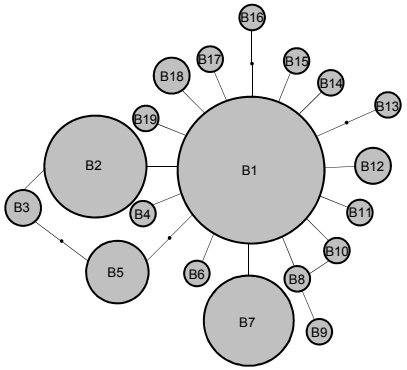



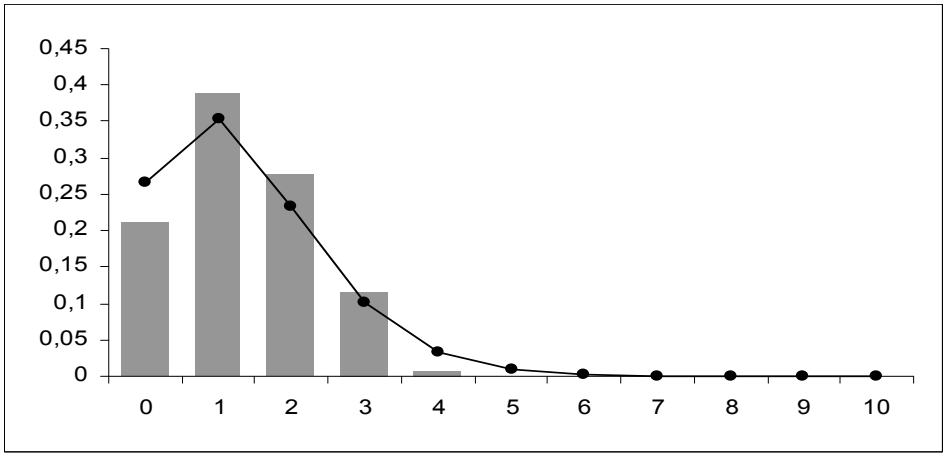
Ex. Expansão demográfica (ou selecção positiva):

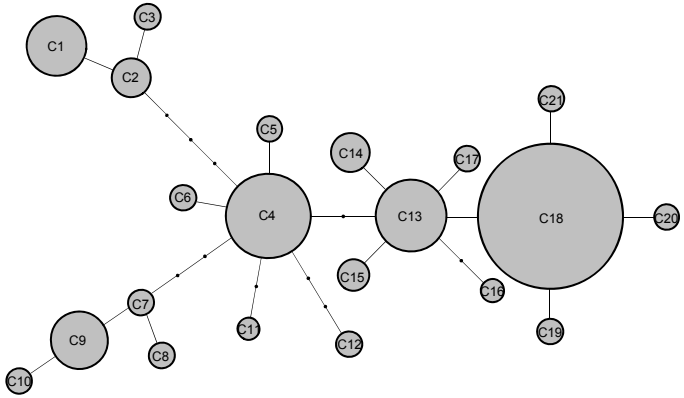



Distribuição do número de diferenças entre pares de sequências

# Mismatch distribution



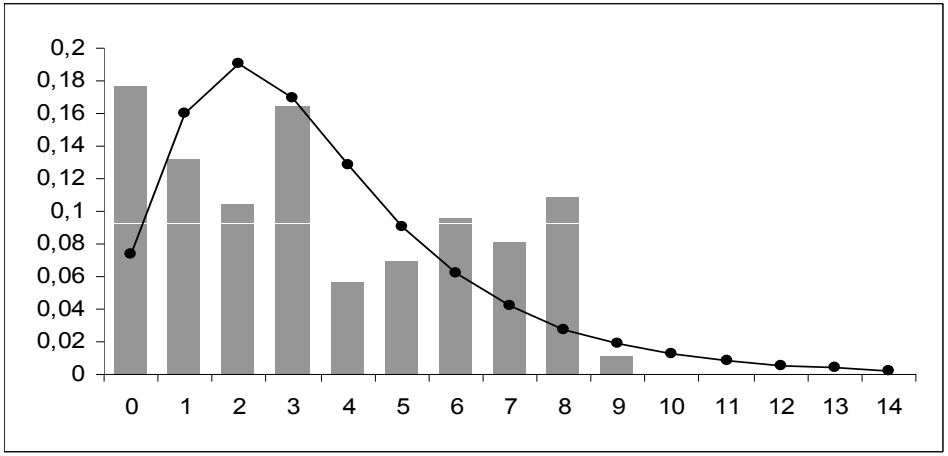



**TABLE 1** Statistical tests of selection[a]

| Test | Type | Designed to detect | Best use | Caveats | Reference(s) |
|---|---|---|---|---|---|
| HKA | Within vs between spp. (two loci) | Differences in variation levels not accountable by constraints | Balancing selection; recent selective sweeps or other variation-reducing forces | High recombination rates may reduce effectiveness of test | 49 |
| McDonald (run test) | Within- vs between-spp. (contiguous region) | Regions with non-neutral patterns of poly. and div. | Equilibrium balancing selection | Has some advantages over the HKA test | 71, 72 |
| McDonald Kreitman G | Within- vs between-spp. (syn. vs nonsynon.) | Adaptive evolution | Adaptive protein evolution: mutation/selection | Selection on codon usage can seriously jeopardize test | 73 |
| Tajima's D | Within sp. | Skew in frequency spectrum | General purpose test of frequency spectrum skew | See reference 27 for situations in which the test performs poorly | 96 |
| Fu & Li's D | Within sp. | Recent vs ancient mutations | General purpose test of frequency spectrum skew | Fu's more recent tests may be more powerful | 29 |
| Fu W | Within sp. | Departures in frequency spectrum | Population subdivision | Hudson's Gst test is more powerful for detecting subdivision | 27 |
| Fu $G\eta$ | Within sp. | Departures in frequency spectrum | Population subdivision, shrinkage, and overdominance selection | Little power against excess number of rare alleles 28 | 27 |
| Fu $G\xi$ | Within sp. | Departures in frequency spectrum | Population subdivision, shrinkage, and overdominance selection | Little power against excess number of rare alleles | 27 |
| Fu $F_s$ | Within sp. | Excess or rare alleles (one sided) | Population growth, genetic hitchhiking, and background selection | May be best overall test for detecting genetic hitchhiking and population growth | 28 |
| Hudson | Within sp. and allele | Unexpectedly low variation within an allele class | Directional selection | A good test for young alleles driven to high frequency | 45 |
| Wall $B$ ans $Q$ | Within sp. | Linkage disequil. between adjacent segregating sites | Population subdivision, balancing selection | $Q$ is more powerful when there is substantial recombination | 100 |
| Andolfatto's $S_k$ | Within sp (sliding window) | Non-neutral haplotype structure | Balancing and directional selection; pop. subdivision | Interpretation may be difficult | 2 |

Alguns dos testes mais utilizados para averiguar se houve expansão demográfica:

**D de Tajima (1989)**
Compara duas estimativas do θ (uma baseada no valor de π e a outra baseada no S). Quando há um excesso de mutações recentes, o S aumenta mas o π mantém-se relativamente baixo (os haplótipos são todos muito próximos), por isso o θ(S) é muito maior do que o θ(π) -> valores negativos da estatística D.

**Fs de Fu (1997)**
Baseia-se na probabilidade de numa população em equilíbrio observarmos mais haplótipos do que os que observamos na nossa população (com base no valor estimado de θ(π)). Valores negativos indicam expansão.

**R2 de Ramos-Onsins e Rozas (2002)**
Baseia-se na diferença entre o número de "singletons" (que numa população em crescimento estão em excesso) e o número médio de diferenças entre sequências. Valores muito baixos indicam crescimento.

**Etc… etc… etc…**

**Todas estas estatísticas são calculadas pelo DnaSP**

Alguns dos testes mais utilizados para averiguar se houve expansão demográfica:

**D de Tajima (1989)**
Compara duas estimativas do θ (uma baseada no valor de π e a outra baseada no S). Quando há um excesso de mutações recentes, o S aumenta mas o π mantém-se relativamente baixo (os haplótipos são todos muito próximos), por isso o θ(S) é muito maior do que o θ(π) -> valores negativos da estatística D.

**Fs de Fu (1997)**
Baseia-se na probabilidade de numa população em equilíbrio observarmos mais haplótipos do que os que observamos na nossa população (com base no valor estimado de θ(π)). Valores negativos indicam expansão.

**R2 de Ramos-Onsins e Rozas (2002)**
Baseia-se na diferença entre o número de "singletons" (que numa população em crescimento estão em excesso) e o número médio de diferenças entre sequências. Valores muito baixos indicam crescimento.

**Etc… etc… etc…**

**Todas estas estatísticas são calculadas pelo DNAsp**

**Teoria da coalescência**

Permite simular genealogias andando para trás no tempo até ao ancestral comum mais recente

Permite gerar dados simulados a partir dos nossos, obtendo-se uma distribuição empírica de determinadas estatísticas que serve como hipótese nula para avaliar o seu grau de significância nos dados reais

↓

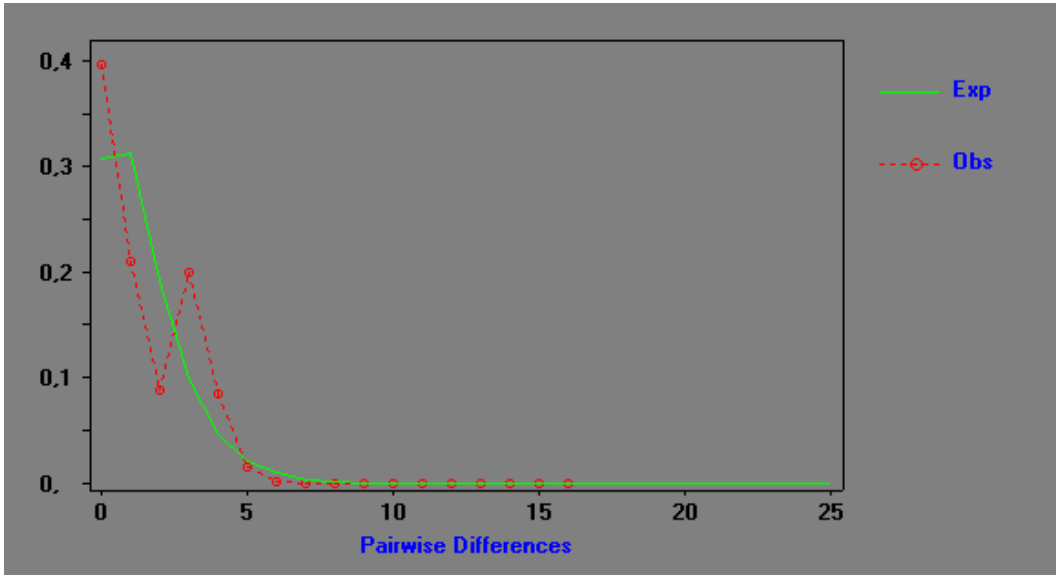
DnaSP

Permite estudar genealogias de acordo com modelos mais ou menos complexos, estimando os valores dos parâmetros do modelo que mais se adequam aos nossos dados

↓

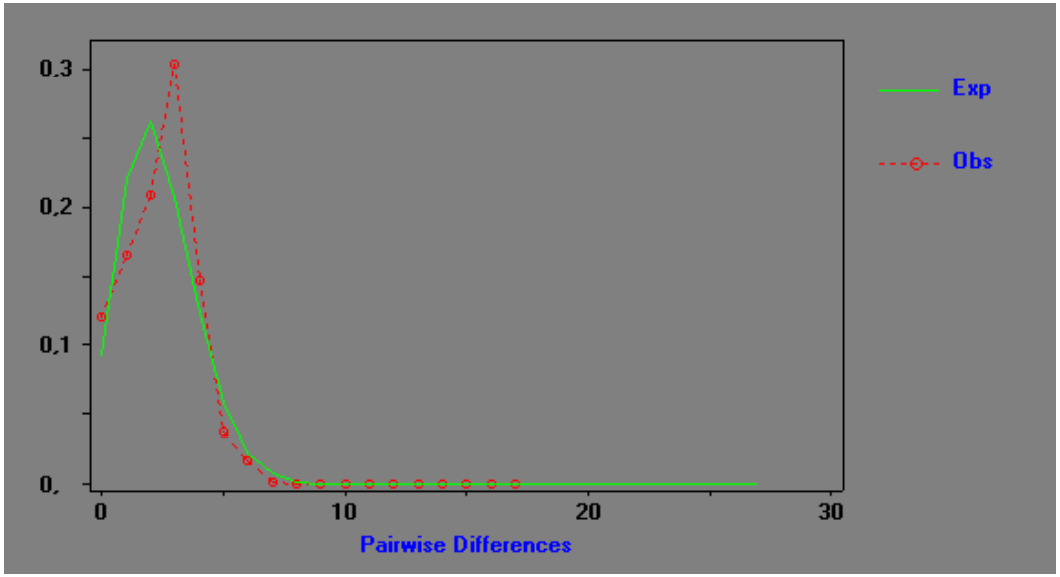
Genetree
Fluctuate
Migrate, Lamarc
IM, Mdiv
BEAST

# Tau = 2ut



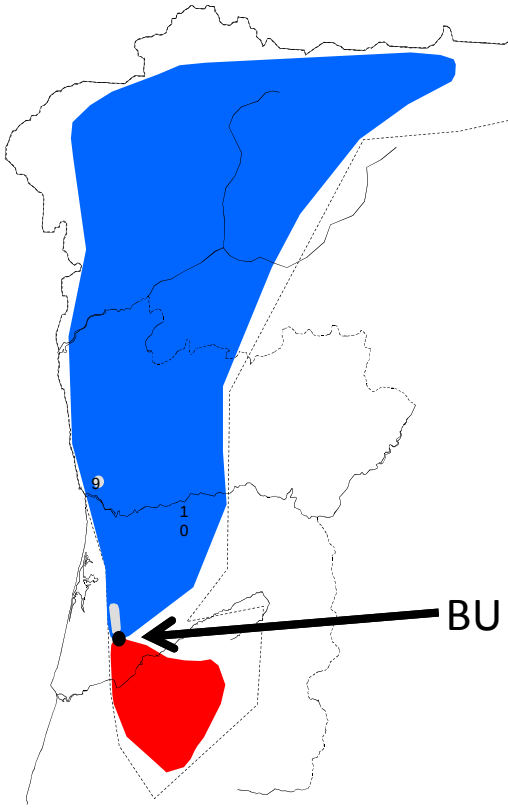

**NORTE**:

Tajima's D = -1.53754
Fu's Fs = -8.322*
R2 = 0.0463*

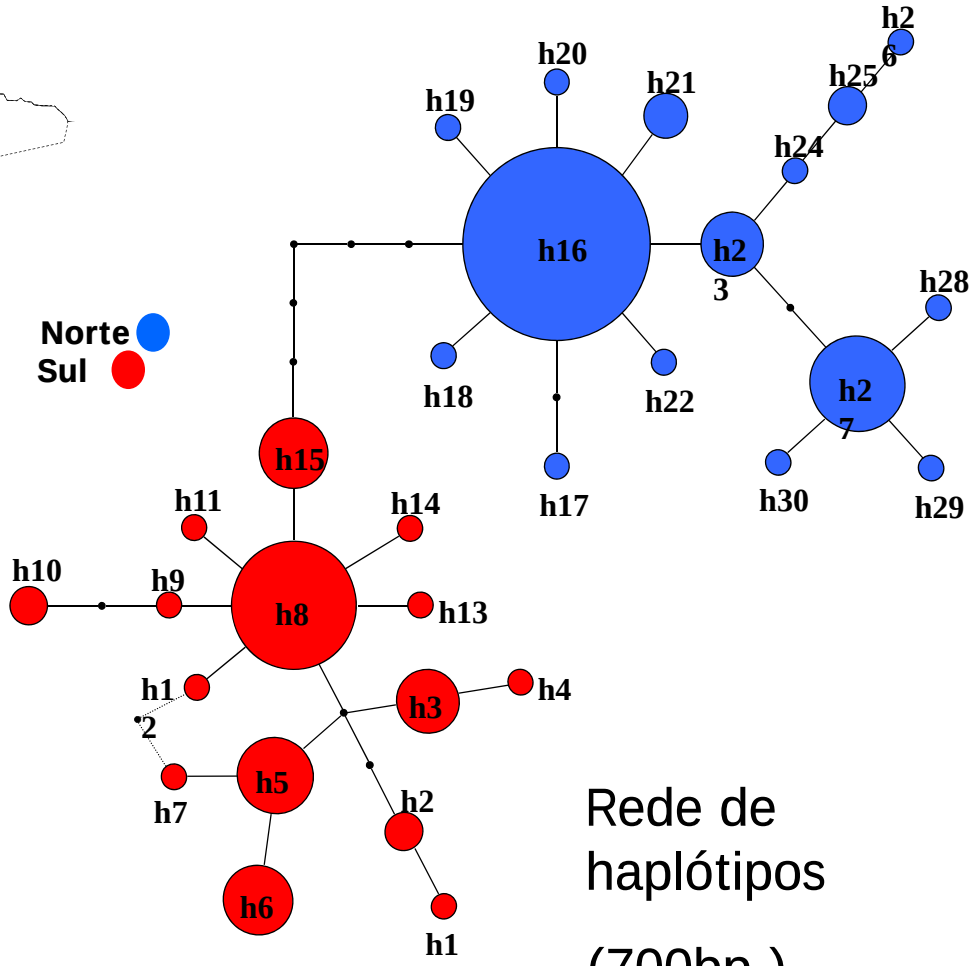



**SUL**:

Tajima's D = -1.05829
Fu's Fs = -5.912*
R2 = 0.0705




Distribuição geográfica dos grupos




Rede de haplótipos

(700bp )

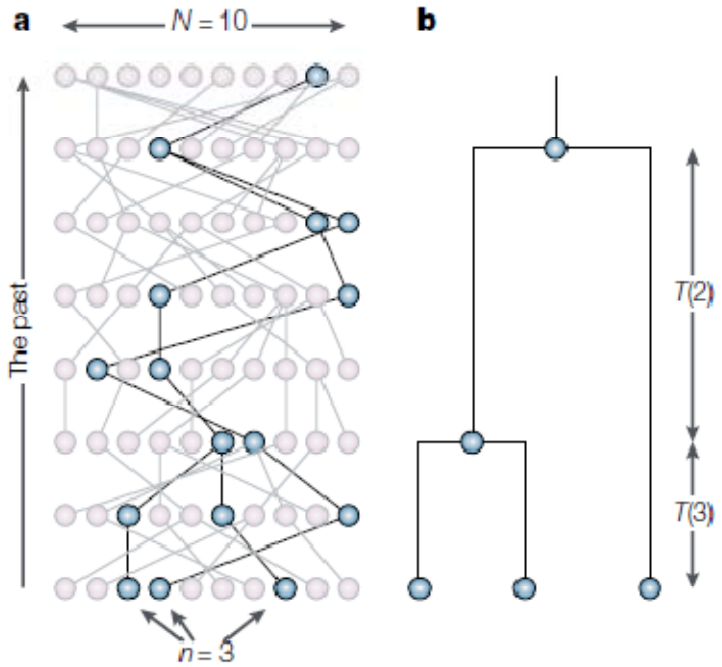
# Inferências demográficas com base na teoria da coalescência e métodos MCMC

Estes métodos que acabámos de utilizar resumem-se ao cálculo de estatísticas sumárias a partir dos nossos dados. São importantes para termos uma ideia geral dos processos demográficos.

No entanto, recentemente foram desenvolvidos outros métodos muito mais robustos para fazer este tipo de estimativa, que se baseiam na teoria da coalescência.

Já utilizámos a teoria da coalescência para estudar se um determinado valor é ou não é significativo. Contudo, podemos utilizá-la para estimar directamente esses parâmetros.

A teoria da coalescência baseia-se no “pensar para trás”: começamos por uma amostra no presente e tentamos investigar o seu passado.



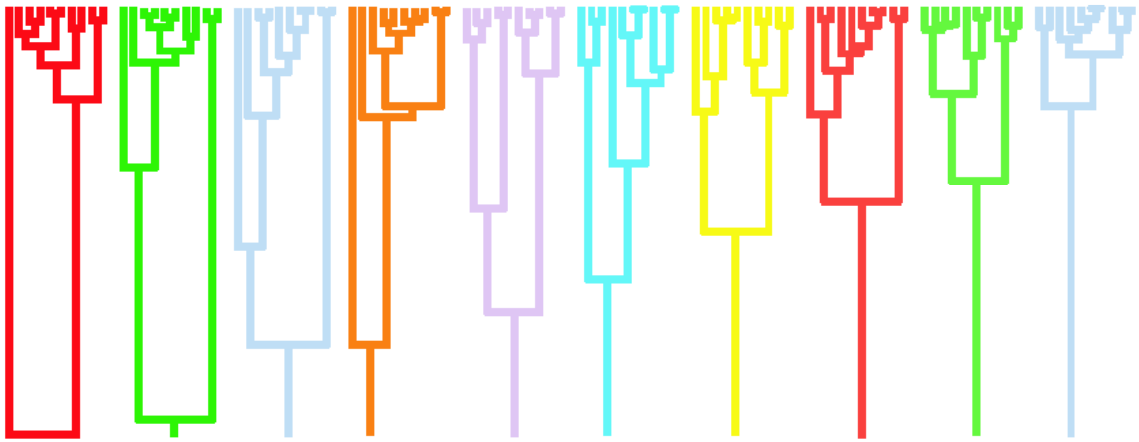

A teoria da coalescência fornece bases matemáticas para avaliar a probabilidade de diferentes cenários, estabelecendo sempre uma relação entre o n (tamanho da amostra) e o N (efectivo populacional).

A TC consiste num modelo matemático que nos permite avaliar a probabilidade de uma genealogia à luz de parâmetros diferentes (por exemplo, o Ne, a migração entre grupos, a taxa de expansão populacional, o tempo de divergência, etc. etc.), e assim saber quais os valores desses parâmetros que melhor explicam essa genealogia.

Mas não é simples fazer isto de uma forma adequada. Há sempre duas grandes fontes de incerteza:

**1.De certeza que estamos a estimar a árvore certa?**

**2. Mesmo que sim, de certeza que a estamos a interpretar correctamente?**
Isto porque a evolução é estocástica:  podemos ter genealogias sugestivas de um fenómeno simplesmente por força do acaso, mesmo sem este ter ocorrido.



Os métodos mais recentes de análise em genética populacional têm em conta estas duas possíveis fontes de erro. Como?

1.Consideram não uma, mas **todas** as genealogias possíveis para os nossos dados (quer ao nível da topologia quer ao nível do tamanho de cada "ramo"). Algumas são mais prováveis que outras considerando os nossos dados, e isto é tido em conta.

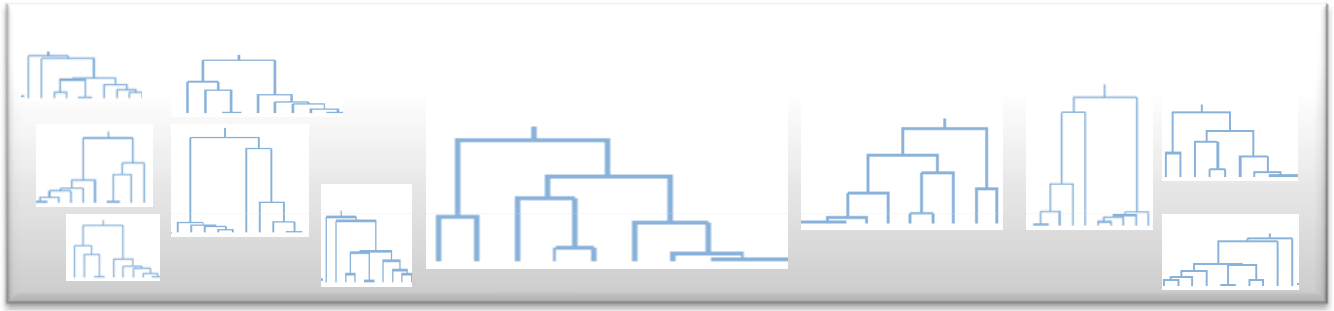
2. Através da teoria da coalescência, estimam a probabilidade de se obter cada uma das genealogias possíveis dados diferentes valores de cada parâmetro que se pretende estimar.

Em teoria, podemos calcular a probabilidade de um parâmetro tomar determinado valor (ex. θ ser igual a 10) somando (ou integrando) a probabilidade de cada genealogia assumindo esse valor.

Mas na prática isto é completamente IMPOSSÍVEL de se calcular analiticamente (o espaço de resultados possíveis é quase infinito). Tem de ser por "tentativa e erro"!

Os métodos mais recentes usam um algoritmo que se chama **Markov Chain Monte Carlo (MCMC).**

PASSO #1



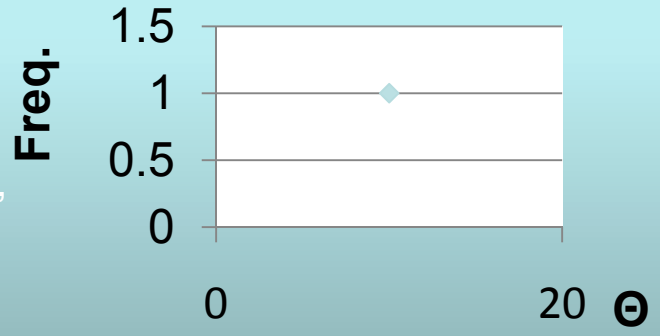
Genealogias amostradas de acordo com a probabilidade dos nossos dados dada a genealogia

## Espaço de “todos” os valores de um parâmetro

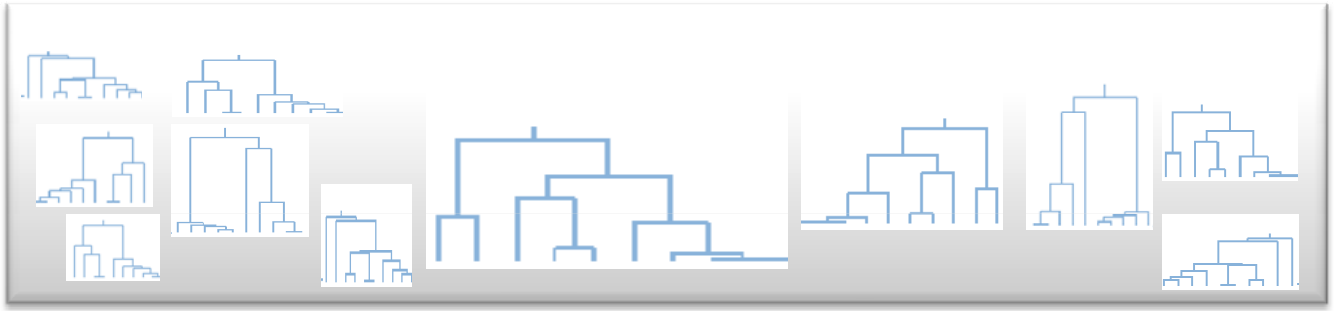
Vamos experimentar, ao acaso, **Θ = 10**.
Vamos calcular a probabilidade da genealogia com este valor (P).

## Resultados

Vamos registar este valor (10), que é agora o nosso valor de referência.



PASSO #2



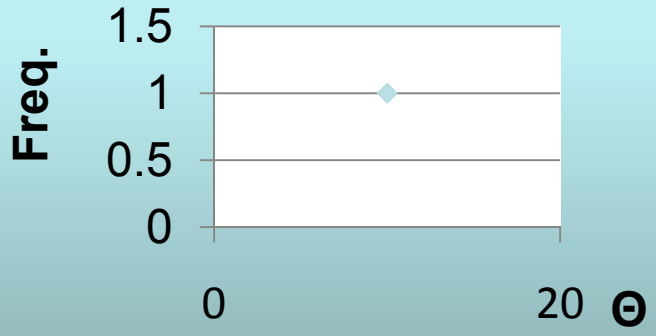
Genealogias amostradas de acordo com probabilidade dos nossos dados dada a genealogia.

## Espaço de “todos” os valores de um parâmetro

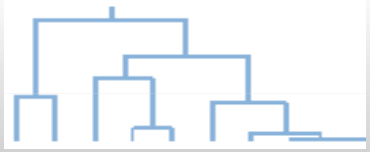
Nova proposta:
Vamos experimentar, ao acaso, **Θ = 15**.
Vamos calcular a probabilidade da genealogia com este valor (P*).

## Resultados

Se P*>P, registamos 15.
Se P> P*, registamos 10 outra vez (ou aceitamos 15) dado um critério probabilístico.
Ex. Vamos assumir que P*>P:



PASSO #2



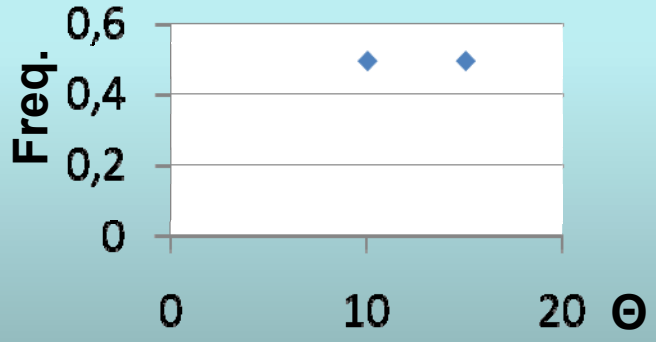
Genealogias amostradas de acordo com probabilidade dos nossos dados dada a genealogia.

## Espaço de “todos” os valores de um parâmetro

Nova proposta:
Vamos experimentar, ao acaso, **Θ = 15**.
Vamos calcular a probabilidade da genealogia com este valor (P*).

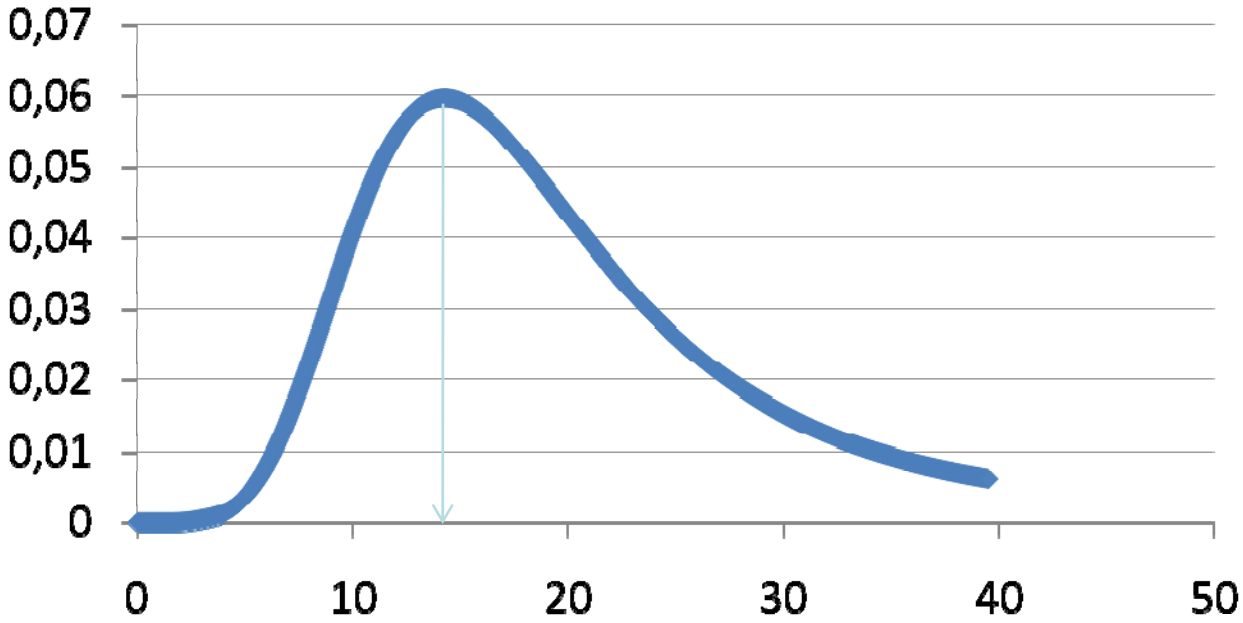
## Resultados

Se P*>P, registamos 15.
Se P> P*, registamos 10 outra vez (ou aceitamos 15) dado um critério probabilístico.
Ex. Vamos assumir que P*>P:
Chama-se a isto um “update”



AO FIM DE MUITOS PASSOS….

Percorremos grande parte do espaço das genealogias (amostrando mais vezes as mais prováveis).
Experimentámos milhares de valores diferentes para cada parâmetro e só aceitámos os que melhoram a nossa estimativa.




A distribuição obtida converge para a distribuição da probabilidade do parâmetro dado os nossos dados.
Chama-se a isto a **distribuição estacionária**.

É A ESTE ALGORITMO DE PROCURA QUE SE CHAMA **MARKOV CHAIN MONTE CARLO**

Cálculo da probabilidade dos nossos dados para uma dada genealogia (atendendo a um **modelo de mutação**, por exemplo).

= **Análise filogenética convencional** (procura a melhor árvore)

↓

Espaço de "todos" os valores de um parâmetro

Cálculo da probabilidade de cada genealogia dados os valores do parâmetro a estimar (atendendo a um **modelo de genética populacional**).

É aqui que a **teoria da coalescência** entra (permite calcular estas probabilidades "facilmente")

↓

Resultados

Ao fim de um número suficiente de passos obtemos a distribuição estacionária da probabilidade de um parâmetro para os nossos dados.
A genealogia é apenas mais um parâmetro (mas do qual não queremos muitas vezes saber).

Algumas notas:

-O método de procura e o critério para fazer “updates” varia conforme o método; isto é bem mais complexo do que os esquemas anteriores fazem crer!

-Por vezes, para facilitar o processo de “procura”, usa-se mais do que uma cadeia em simultâneo (Metropolis coupling; – MCMCMC).

- Estes métodos podem ser implementados de duas maneiras: no âmbito de critérios de verosimilhança (likelihood) ou bayesianos (probabilidade posterior)

Apesar da sua complexidade, estes métodos têm muitas vantagens:

-São muito mais robustos porque não se baseiam apenas na estimativa de uma árvore, que pode sempre estar sujeita a erro.

-Percorrem todo o espaço possível dos parâmetros, por isso se tudo correr bem temos a certeza de que nada ficou por testar.

-Assim, dão-nos uma estimativa da “certeza” (ou incerteza) com que nós estamos a estimar esses parâmetros.

-Do ponto de vista de quem os implementa, é relativamente fácil adicionar parâmetros ao modelo – basta que encontrem uma relação probabilística entre esses parâmetros e a forma da genealogia.

- Por isso permitem testar modelos evolutivos (incluindo muitos parâmetros ao mesmo tempo) mais complexos do que o normal.

Ex. Programa Lamarc

Permite estimar, ao mesmo tempo:

- o θ para n populações (θ1, θ2, θ3, …, θn)

-o parâmetro de crescimento exponencial (g) para cada uma dessas populações (g1, g2, g3,…, gn)

$$\theta_t = \theta_{actual} \times e^{-gt}$$

- a taxa de migração entre todas as populações (m12, m21, m13, m31, …)

- a taxa de recombinação para cada locus

Tudo ao mesmo tempo!

Uma desvantagem destes métodos é que exigem computadores bons e mesmo assim demoram muito tempo.

Por isso não vamos usar o Lamarc, mas o seu parente mais pobre, o Fluctuate, que demora muito menos tempo a correr.

Parâmetros que vamos estimar usando o Fluctuate:

-θ
-parâmetro *g*

# BEAST
# (Bayesian Evolutionary Analysis by Sampling Trees)

O BEAST é um programa “Bayesiano” que permite analisar sequências de DNA ou proteínas.

Pode ser utilizado como método de reconstrução de filogenias mas permite também, aplicando a teoria da coalescência, testar diversas hipóteses evolutivas sem se estar condicionado a uma única topologia.

Usa o MCMC para percorrer o espaço das árvores e parâmetros de forma a que o peso de cada uma é proporcional à sua probabilidade posterior.

O BEAST PERMITE:

* Assumir um relógio molecular constante ou aplicar um modelo de relógio molecular de taxa variável (relaxado).

* Assumir um modelo de mutação heterogéneo (e vários modelos de mutação).

* Aplicar modelos demográficos de acordo com a teoria da coalescência.

* Uma escolha flexível dos *priors* e do modo de busca.

* Utilizar sequências não contemporâneas.

* Estimar datas de divergência (TMRCA).

O BEAST PERMITE:

* Assumir um relógio molecular constante ou aplicar um modelo de relógio molecular de taxa variável (relaxado).

* Assumir um modelo de mutação heterogéneo (e vários modelos de mutação).

* Aplicar modelos demográficos de acordo com a teoria da coalescência.

* Uma escolha flexível dos *priors* e do modo de busca.

* Utilizar sequências não contemporâneas.

* Estimar datas de divergência (TMRCA).