Pràctica amb ordinador 1. Coalescència

Principals Transicions Evolutives

2023-04-18

Preparació de l'ordinador

En aquesta pràctica utilitzarem el llenguatge de programació R, opcionalment el programa RStudio i el paquet `learnPopGen` (Revell 2019) per comprovar alguns dels principis de la teoria de la coalescència. Si no t'has familiaritzat encara amb R, conulta el document **Introducció a R**.

Primer, cal instal · lar `learnPopGen`, si no s'ha instal · lat abans. Inicia una sessió d'R o d'RStudio i tecleja (o copia i pega) el següent en la consola:

```
install.package('learnPopGen')
```

Una vegada instal · lat, cal carregar el paquet `learnPopGen` en la sessió de treball. Per fer-ho, tecleja el següent en la consola:

```
library('learnPopGen')
```

Per últim, cal importar la funció `genealogia()`, que ens ajudarà a visualitzar els llinatges d'individus concrets. Està guardada a l'arxiu `genealogia.R`, en la mateixa carpeta on has trobat este guió. Pots obrir l'arxiu, copiar la definició de la funció i pegar-la en la consola (i donar-li al botó `enter`). O bé pots utilitzar la funció `source()` per llegir i executar l'arxiu des de la consola. Per exemple, si la carpeta de treball de la sessió de R és la mateixa on està l'arxiu `genealogia.R`, pots utilitzar el comandament següent:

```
source('genealogia.R')
```

Si l'arxiu està en una carpeta diferent a la carpeta des de la qual estàs treballant en la sessió de R, pots indicar l'adreça completa de l'arxiu. Per exemple:

```
source('C:/Users/pep/Downloads/Coalescencia/genealogia.R')
```

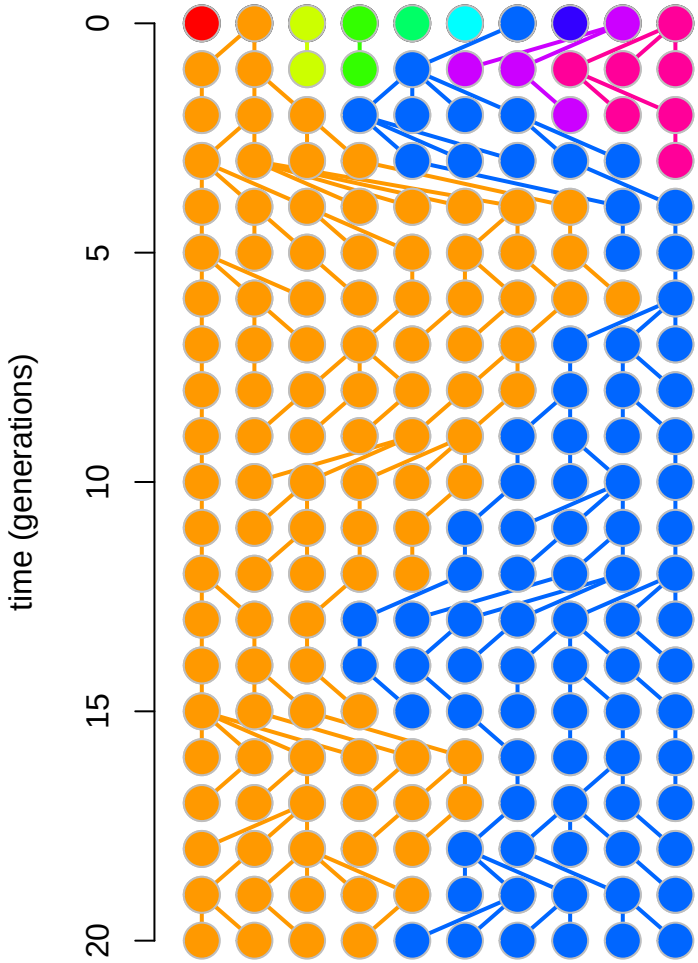
Per saber quina és la teua carpeta de treball actual en R, utilitza la funció `getwd()`. I per canviar-la, `setwd()`. Per exemple:

```
setwd('C:/Users/pep/PTE/Practica/Coalescencia')
```

Simulació

La funció `coalescent.plot()` permet simular i representar gràficament una població de Fisher-Wright: un nombre constant d'individus haploides (o de còpies d'un gen), amb generacions separades, on tots els individus (o còpies d'un gen) tenen la mateixa probabilitat de deixar descendència (*eficàcia biològica*), i per tant les variacions de les freqüències de cada llinatge només depenen de la **deriva genètica**: són aleatòries, no hi ha selecció natural.

```
sim01 <- coalescent.plot(n = 10, ngen = 20, sleep = 0)
```



La funció `coalescent.plot()` necessita que especifiquem el nombre d'individus (paràmetre `n`) i el nombre de generacions (`ngen`). A més, indiquem `sleep = 0` perquè represente la gràfica més ràpidament. L'objecte `sim01` guarda el resultat de la simulació. Així, podràs tornar a representar gràficament el resultat cada vegada que tecleges `plot(sim01)`.

Nombre de descendents

El primer exercici consisteix en averiguar quina és l'esperança (la mitjana) del nombre de descendents d'un individu qualsevol que quedaran **vius al cap de 40 generacions**. Utilitza una població de 20 individus per calcular-ho. Després d'averiguar-ho, contesta les preguntes següents:

- El resultat depén del nombre de generacions?
- Depén de la mida de la població?
- És un valor representatiu?

Llegeix el text següent i pensa quina opinió et mereix.

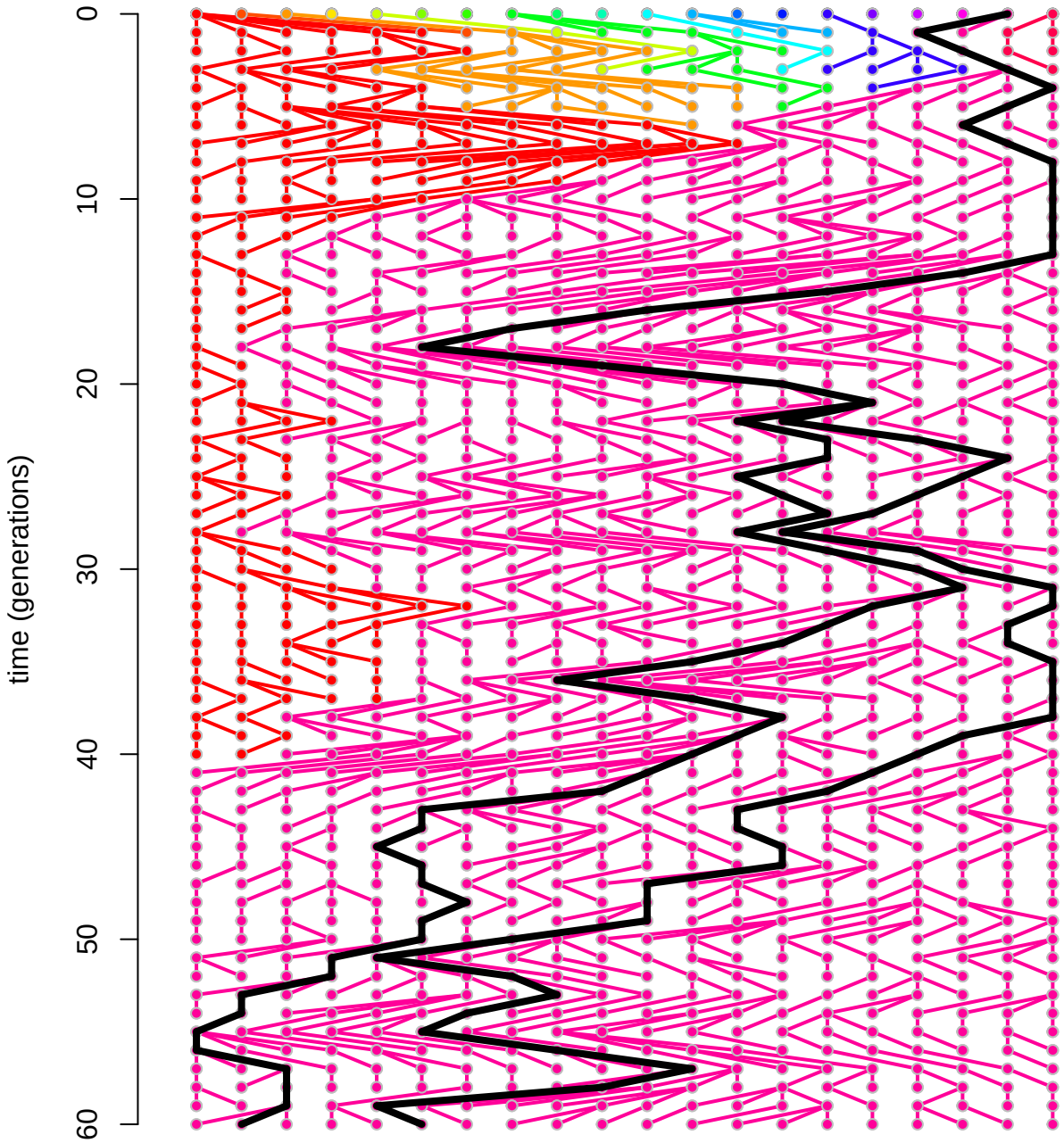
> "Suposa que la població de Dinamarca compleix les suposicions: 6 milions de persones, un nombre indefinit de generacions en el passat, sense estructura social i amb un temps de generació de 25 anys. Per estar pràcticament segurs que un individu d'aleshores o bé no té cap descendent en el present o bé és l'ancestre de tots en el present, ens hauríem de remuntar 966 anys. Com que el rey Gorm (primer rei danés, mort l'any 958) té descendents en la població actual, ha de ser necessàriament ancestre de tot el regne."
>
> — Hein, J., Schierup, M.H. i Wiuf, C. 2005, p. 251

Temps de coalescència de 2 individus

L'objectiu és estimar la mitjana i la variància del nombre de generacions que ens hem de remuntar en el passat per trobar l'ancestre comú *més recent* de **dos** individus actuals **triats a l'atzar**. Per obtenir uns valors representatius cal utilitzar almenys 10 simulacions diferents. Simularem poblacions de 20 individus durant 60 generacions. Per triar 2 individus aleatoris entre 20, utilitza la funció `sample(1:20, 2)`. Per facilitar la identificació de l'ancestre comú més recent, utilitza la funció `genealogia()`, que necessita que especifiquem dos paràmetres: el nom de la simulació realitzada i els individus triats aleatòriament, dels quals volem conéixer la genealogia:

```
sim02  <- coalescent.plot(n = 20, ngen = 60, sleep = 0)
```

```
mostra <- sample(1:20, 2)
genealogia(sim02, mostra)
```



Per exemple, en la gràfica anterior s'observa que l'ancestre comú més recent dels individus 2 i 6 va viure fa 39 generacions. Si fa 60 generacions encara existien 2 ancestres diferents per als dos individus seleccionats aleatòriament, caldria afegir una segona simulació per identificar l'ancestre comú d'aquells dos ancestres de fa 60 generacions.

Guarda els resultats en un vector en l'espai de treball. Per exemple, si aquests fóren els teus 10 resultats faries:

```
Exercici2 <- c(58, 20, 56, 51, 52, 35, 33, 15, 31, 59)
```

Així podràs calcular la mitjana i la variància fàcilment amb les funcions `mean(Exercici2)` i `var(Exercici2)`. Després, contesta les preguntes següents:

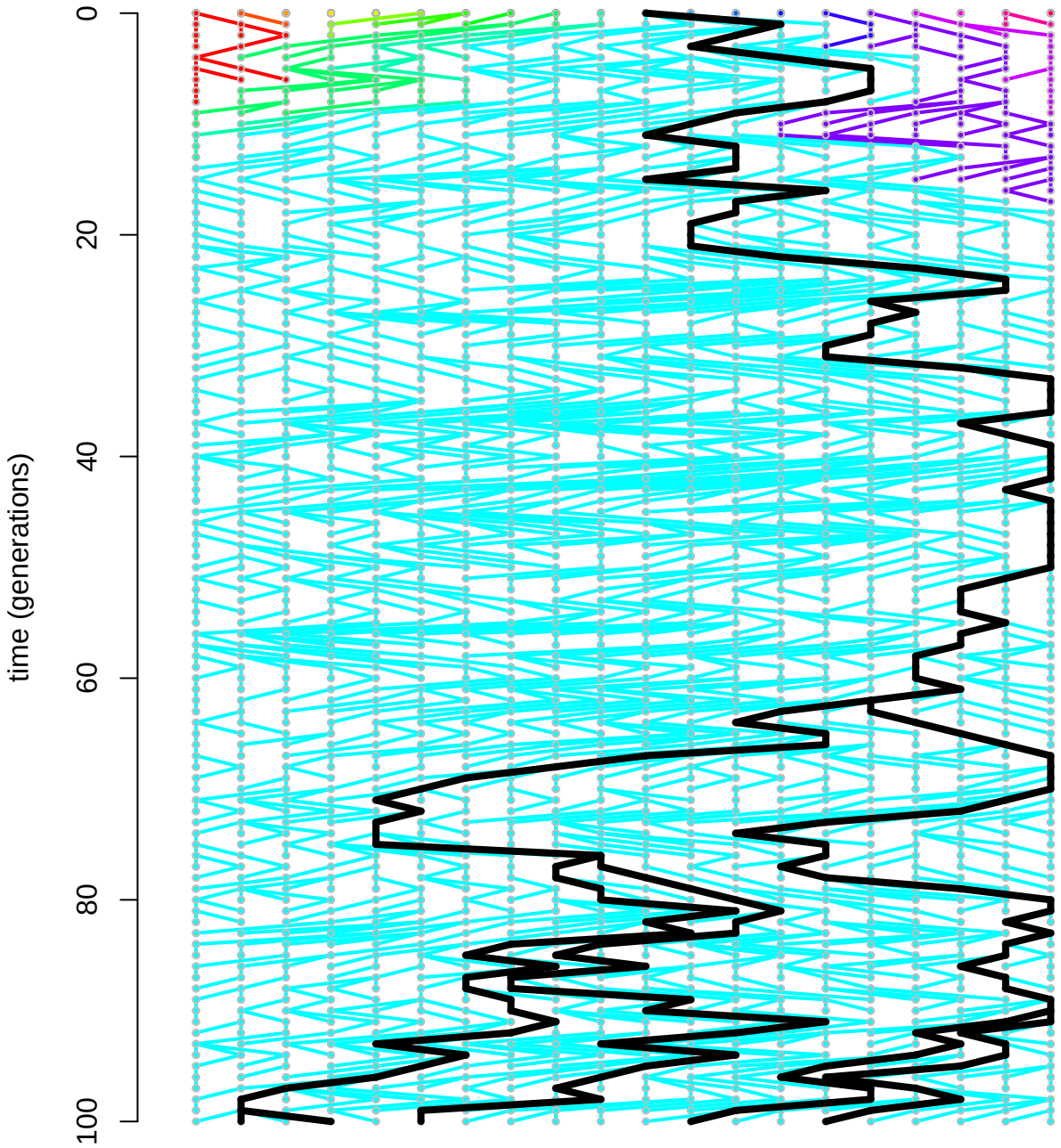
- Com es comparen els valors estimats amb els valors esperats d'acord amb la teoria?
- Si la mida poblacional efectiva de la població humana fóra 10000 individus, quantes generacions ens hauríem de remuntar en el passat per trobar l'ancestre comú més recent de l'ADN mitocondrial de dos individus triats a l'atzar?
- I per trobar l'ancestre comú més recent de dues còpies d'un gen nuclear?

Temps de coalescència de 5 individus

Cal repetir l'exercici anterior, però ara buscant l'ancestre comú més recent de **5** individus *triats a l'atzar* en una població de 20 individus. Pots fer el mostreig amb la funció `sample(1:20, 5)`. Aquesta vegada, per ajudar-te a comptar les generacions, pots utilitzar la funció `mrca()`, que necessita dos arguments: la simulació obtinguda amb `coalescent.plot()` i la mostra dels 5 individus triats a l'atzar. Aprofitant que no cal identificar l'ancestre comú en el gràfic, pots simular 100 (o més) generacions, per assegurar-te que trobaràs l'ancestre comú dels 5 individus en una sola simulació. Per exemple:

```
sim03 <- coalescent.plot(n = 20, ngen = 100, sleep = 0)
```

```
mostra <- sample(1:20, 5)
genealogia(sim03, mostra)
```




```
mrca(sim03, mostra)
```

```
## [1] 38
```

La funció `mrca(sim03, mostra)` ens diu que l'ancestre comú més recent dels individus 2, 4, 6, 12 i 15 va viure fa 38 generacions, com pots comprovar a la gràfica. Com abans, guarda els resultats en un vector i determina la mitjana i la variància. Aleshores, contesta les preguntes següents:

- Com es compara el resultat empíric amb l'esperança teòrica del nombre de generacions que cal retrocedir en el passat per trobar l'ancestre comú més recent de 5 individus?
- Quina era la fòrmula de l'esperança teòrica del temps de coalescència de tota una espècie?
- Quin resultat donaria per a la coalescència global del cromosoma mitocondrial humà? Utilitza un temps de generació de 20 anys per traduir generacions a anys.
- I quin és el temps de coalescència global esperat per un gen nuclear humà?

- Compara els temps de coalescència de 2 individus ($T_{MRCA}(2)$) i de tota l'espècie ($T_{MRCA}(N_e)$): et sembla que estan ben proporcionats?

Representació dinàmica de la coalescència

Si queda temps, visita l'enllaç http://bedford.io/projects/coaltrace i observa la representació dinàmica i en temps continu del procés de coalescència. Les boles de colors representen individus i les línies que deixen enrere, relacions d'ancestralitat. Els llinatges extingits desapareixen. En tot moment veiem un arbre amb la forma típica del procés de coalescència, amb aproximadament la meitat de tota la profunditat de l'arbre ocupada per només dos llinatges. Si fas *click* sobre el gràfic, activaràs els comandaments. Aleshores, prem la tecla H per veure les opcions i prova l'efecte de modificar els paràmetres següents sobre la forma de l'arbre:

- El temps de generació.
- La mida poblacional.
- Les mutacions.
- La migració.

Apèndix 1. Resum de la teoria

En una població de Fisher-Wright de mida N, on N és el nombre de *gàmetes* que passen d'una generació a la següent, el temps mig fins la primera coalescència entre n gens mostrejats a l'atzar és:

$$E(T_n) = \frac{N}{\binom{n}{2}} = \frac{2N}{n(n-1)}$$

I la variància seria:

$$\sigma^2(T_n) = \frac{4N^2}{(n(n-1))^2}$$

Per tant, el temps esperat per trobar l'ancestre comú de dos gens en una població de N gàmetes és $T_2 = N$ generacions, amb una variància $\sigma^2(T_2) = N^2$.

El temps mig fins la coalescència completa d'una mostra de n gens i la seua variància són:

$$E(T_{MRCA}(n)) = \sum_{j=2}^{n} E(T_j) = \sum_{j=2}^{n} \frac{2N}{j(j-1)}$$

$$= 2N\left(1 - \frac{1}{n}\right)$$

$$\sigma^2(T_{MRCA}(n)) = 4N \sum_{j=2}^{n} \frac{1}{j^2(j-1)^2}$$

$$\sim 1.16N$$

El temps mig fins la coalescència de tots els N gens de la població, en principi és: $E(T_{MRCA}(N)) = 2N - 2$, d'acord amb la fòrmula anterior. A pesar que les fòrmules estan pensades per a mostres molt menors que la mida real de la població, aquest resultat és adequat, perquè el nombre de llinatges es redueix ràpidament durant les primeres generacions.

Havent definit N com el nombre de gàmetes, les fòrmules anteriors són vàlides tant per a poblacions haploides com diploides. Però en poblacions diploides és més comú referir-se al nombre d'individus. Per tant, es pot

substituir $N = 2N_e$, on N_e és el nombre efectiu d'individus. En qualsevol cas, es pot generalitzar a unitats de coalescència, definides com el nombre de generacions esperades fins la coalescència de 2 gens. És a dir, si $T_2 = 1$ unitats de coalescència, aleshores, $E(T_n) = \frac{2}{n(n-1)}$, $E(T_{MRCA}(n)) = 2(1 - 1/n)$ i $E(T_{MRCA}(N)) \sim 2$. Així les fòrmules s'apliquen a poblacions de qualsevol mida. La mida poblacional només afecta l'escala temporal. En totes les poblacions de Fisher-Wright s'espera que l'ancestre comú a tots els individus es remonta només al doble del nombre de generacions necessàries per trobar l'ancestre comú de dos individus triats a l'atzar.

Bibliography

Hein, J., M. H. Schierup, and C. Wiuf. 2005. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.

Revell, Liam J. 2019. "learnPopGen: An r Package for Population Genetic Simulation and Numerical Analysis." *Ecology and Evolution* 9 (14): 7896–7902. https://doi.org/https://doi.org/10.1002/ece3.5412.