Genetic relatedness through the lens of tree sequences

Genetic relatedness through the lens of tree sequences

P29

Brieuc Lehmann[1], Gregor Gorjanc[2], Jerome Kelleher[3], Peter Ralph[4], Georgia Tsambos[5]

1. University College London, 2. University of Edinburgh, 3. University of Oxford, 4. University of Oregon, 5. University of Melbourne

ProbGen 2022

The many facets of genetic relatedness

Relatedness has been characterised in many different ways since the seminal paper by Sewall Wright on 'Coefficients of inbreeding and relationship' 100 years ago. In general, genetic relatedness refers to some notion of similarity between individuals, where similarity can be defined according to pedigree, genotype, or genealogy. See [1] for a great review.

Pedigree-based
Coancestry: $\theta(B,C) =$ P(allele drawn at random is IBD in B and C)

$$\theta(B,C) = \textstyle\sum_A \frac{1+f_A}{2^{g_A+1}}, \qquad f_A\text{: coancestry of A's parents} \qquad g_A\text{: length of lineage path } B \to A \to C$$

N.B. in expectation only & requires knowledge of the pedigree!

Genotype-based
Similarity ≈ P(allele drawn at random matches in B and C). Different versions with centring/standardising. Let $X \in \{0,1,2\}^{n\times p}$ be the genotype matrix. E.g.

$$K_{as}(B,C) = \frac{1}{2} + \frac{1}{2m}(X_B - 1)^T(X_C - 1) \qquad \text{‘allele−sharing coefficient’}$$

$$K_{c\alpha}(B,C) = \frac{1}{m}(X_B - 2p_j)^T D_\alpha (X_C - 2p_j), \qquad diag(D_\alpha)_j = \left(2p_j(1-p_j)\right)^\alpha$$

N.B. sensitive to SNPs selected and choice of reference population

Genealogy-based
Similarity = f(Time to Most Recent Common Ancestor across genome)

N.B. based on the Ancestral Recombination Graph… enter tree sequences!

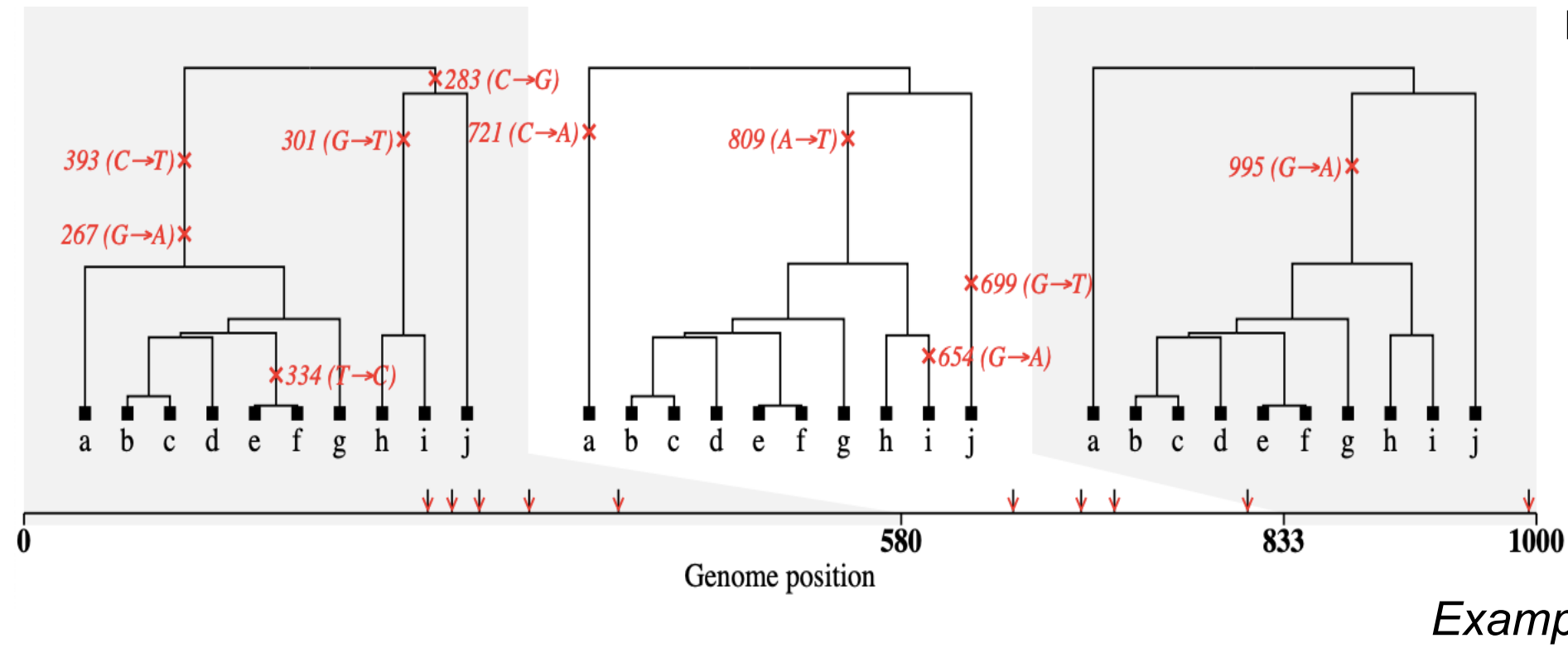
What has genetic relatedness ever done for us?

Directly or indirectly, genetic relatedness plays a key role in a number of common population genetics analyses. Let Σ be the genetic relatedness matrix (GRM).

- **Principal components analysis (PCA)**
  Based on an eigendecomposition of Σ. Used to identify structure in the distribution of genetic variation – see example over on the right →.
- **Phenotype prediction**
  Linear mixed model approaches (e.g., BLUP, "Bayesian alphabet"). For example, $Y \sim N(Z\beta_0,\ \Sigma\sigma_g^2 + \sigma_e^2)$.
- **Heritability estimation**
  As above, but interest is in estimating $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_e^2)$.
- **Controlling for population structure in GWAS**
  Common methods include regressing out the first n principal components, or using linear mixed models (e.g., BOLT-LMM).
- **Any more for anymore?**
  We're very interested to hear any other examples of genetic relatedness being used in the wild – please get in touch!

What is a tree sequence?

A *succinct tree sequence* represents the evolutionary relationships between a set of DNA sequences. Tree sequences are essentially an encoding of Ancestral Recombination Graphs; they can be created by simulation or by inferring relationships from empirical DNA data.



```
Pos: 267 283 301 334 393 654 699 721 809 995
  a:  A   C   G   T   T   G   G   A   A   G
  b:  A   C   G   T   T   G   G   C   T   A
  c:  A   C   G   T   T   G   G   C   T   A
  d:  A   C   G   T   T   G   G   C   T   A
  e:  A   C   G   C   T   G   G   C   T   A
  f:  A   C   G   C   T   G   G   C   T   A
  g:  A   C   G   T   T   G   G   C   T   A
  h:  G   G   T   T   C   G   G   C   T   A
  i:  G   G   T   T   C   A   G   C   T   A
  j:  G   G   G   T   C   G   T   C   A   G
```

Example taken from https://tskit.dev/tutorials/what_is.html

Relatedness as covariance

We can use the tree sequence structure to calculate the genetic relatedness matrix Σ for a group of individuals. We can do so just as easily for branch-based or site-based genetic relatedness.

We have haploid* individuals $I_{1:N}$ with traits $Y_{1:N} = g_{1:N} + e_{1:N}$ where g_i is the genetic value, with population mean $\bar{g} = \frac{1}{N}\Sigma g_i$. Let $g_i^* := g_i - \bar{g}$.

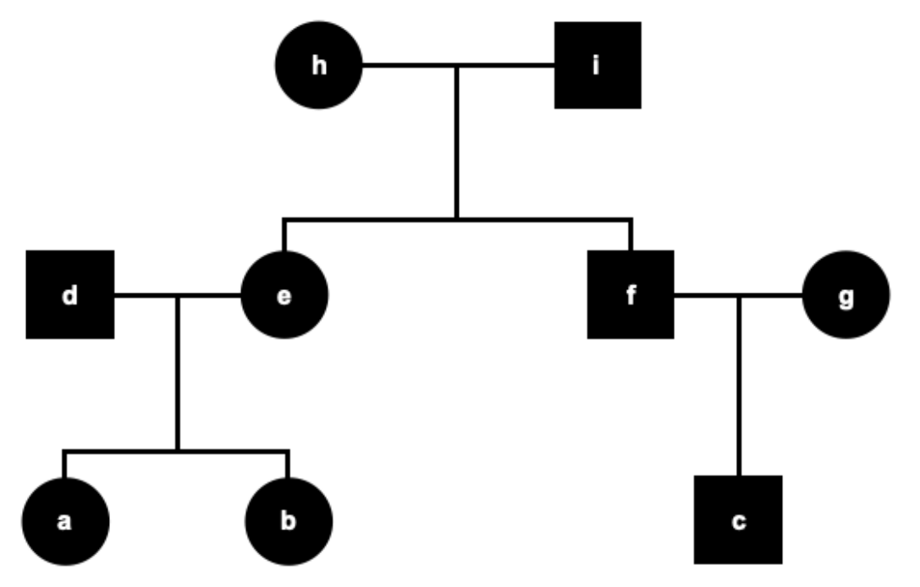
Site-based	Branch-based
Each (biallelic) site j is associated with an effect Z_j. Let $E[Z_j] = 0$ and $Var(Z_j) = 1$.	Each branch b is associated with an effect Z_b. Let $E[Z_b] = 0$ and $Var(Z_b) = a_b$ (‘area’ of branch b)
Let $X_{ij} = 1$ iff individual i has a mutation at site j. $X_{ij} = 0$ otherwise.	Let $T_{ib} = 1$ iff branch b is ancestral to individual i. $T_{ib} = 0$ otherwise.
The genetic value of a trait is the sum of the site effects carried by an individual: $g_i = \sum_{j:G_{ij}=1} Z_j$	The genetic value of a trait is the sum of the branch effects carried by an individual: $\mathrm{g_i} = \sum_{b:T_{ib}=1} Z_b$
Genetic relatedness between individuals i and j is the covariance between the centered genetic values g_i^* and g_j^*. $\Sigma_{ij} = Cov(g_i^*, g_j^*)$	
Let m(B,C) be the no. of pairwise site matches between B and C.	Let A(B,C) be the total area of branches ancestral to both B and C.
Let U and V be individuals selected uniformly at random from $I_1, \dots, I_N$.	
Then, $\Sigma_{BC} = E[m(B,C) - m(B,V) - m(C,U) + m(U,V)]$	Then, $\Sigma_{ij} = E[A(B,C) - A(B,V) - A(C,U) + A(U,V)]$
‘Number of pairwise allelic matches relative to the rest of the sample.’	‘Total area of branches ancestral to pair relative to the rest of sample.’

This is implemented in `tskit` through `ts.genetic_relatedness()`. See also [2] for some closely related (pun intended) work.

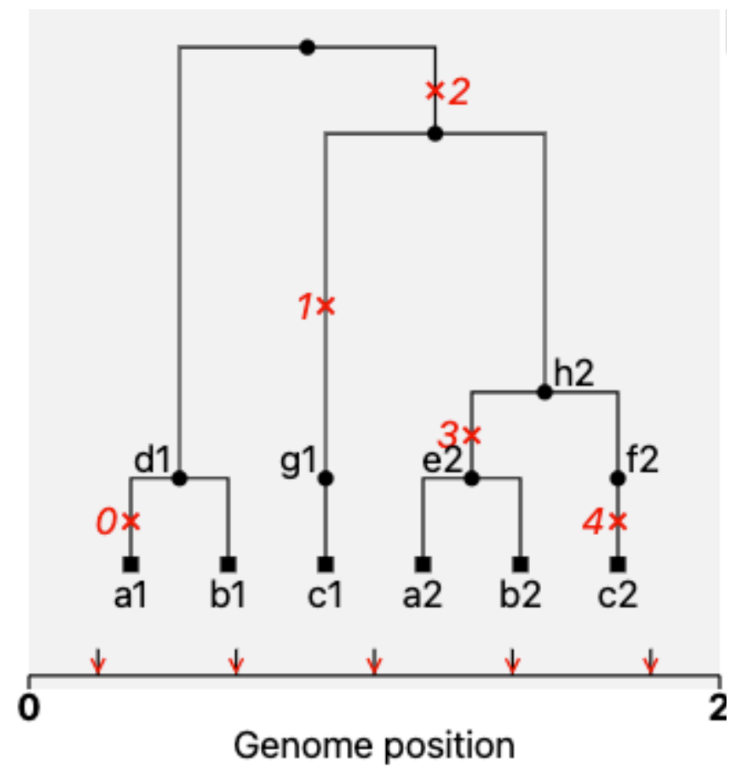
*Extends straightforwardly to other ploidy levels!

An illustrative example

Pedigree



Tree sequence



Genotype matrix

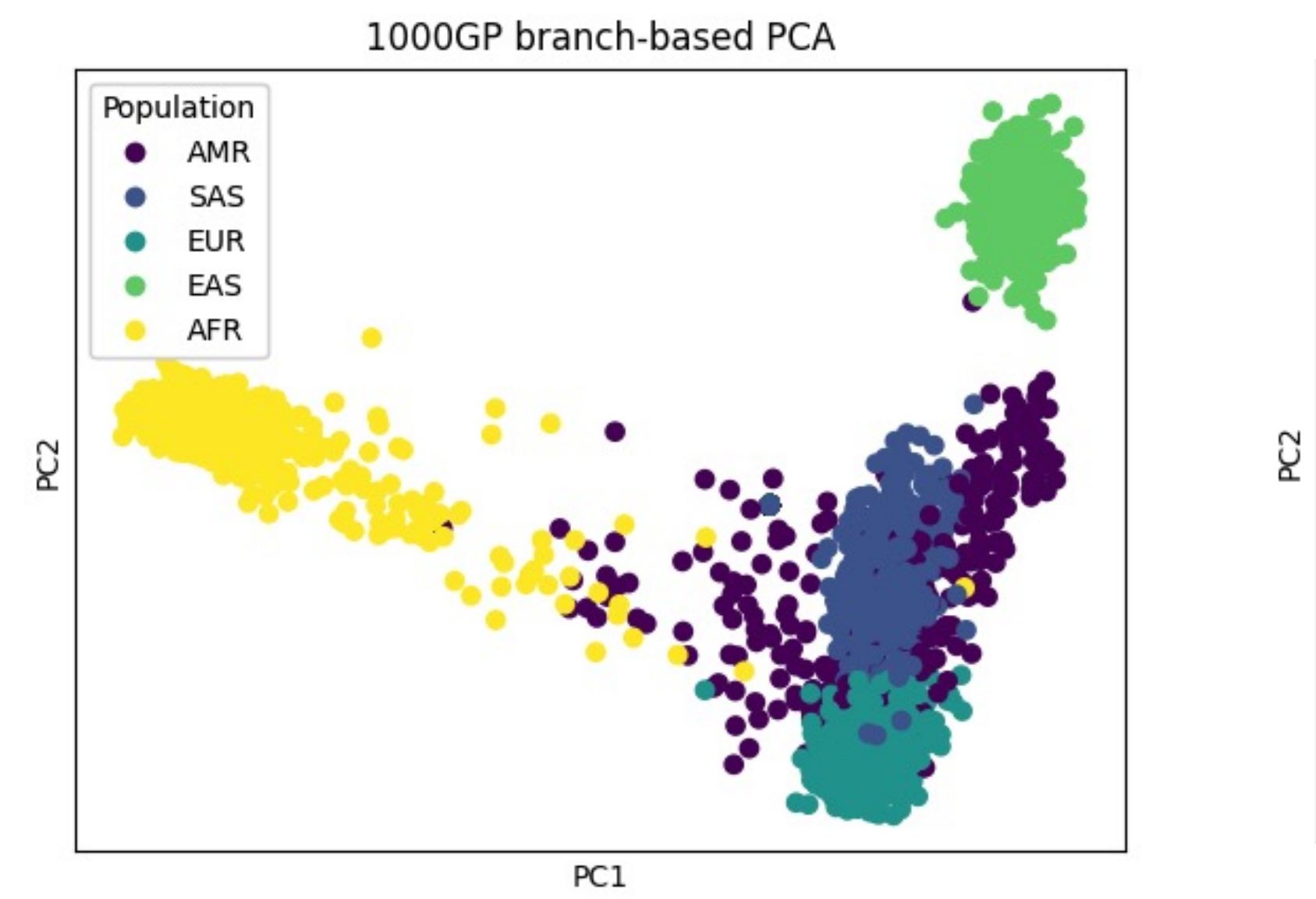
```
Pos:  0   1   2   3   4
  a:  1   0   1   1   0
  b:  0   0   1   1   0
  c:  0   1   2   0   1
```

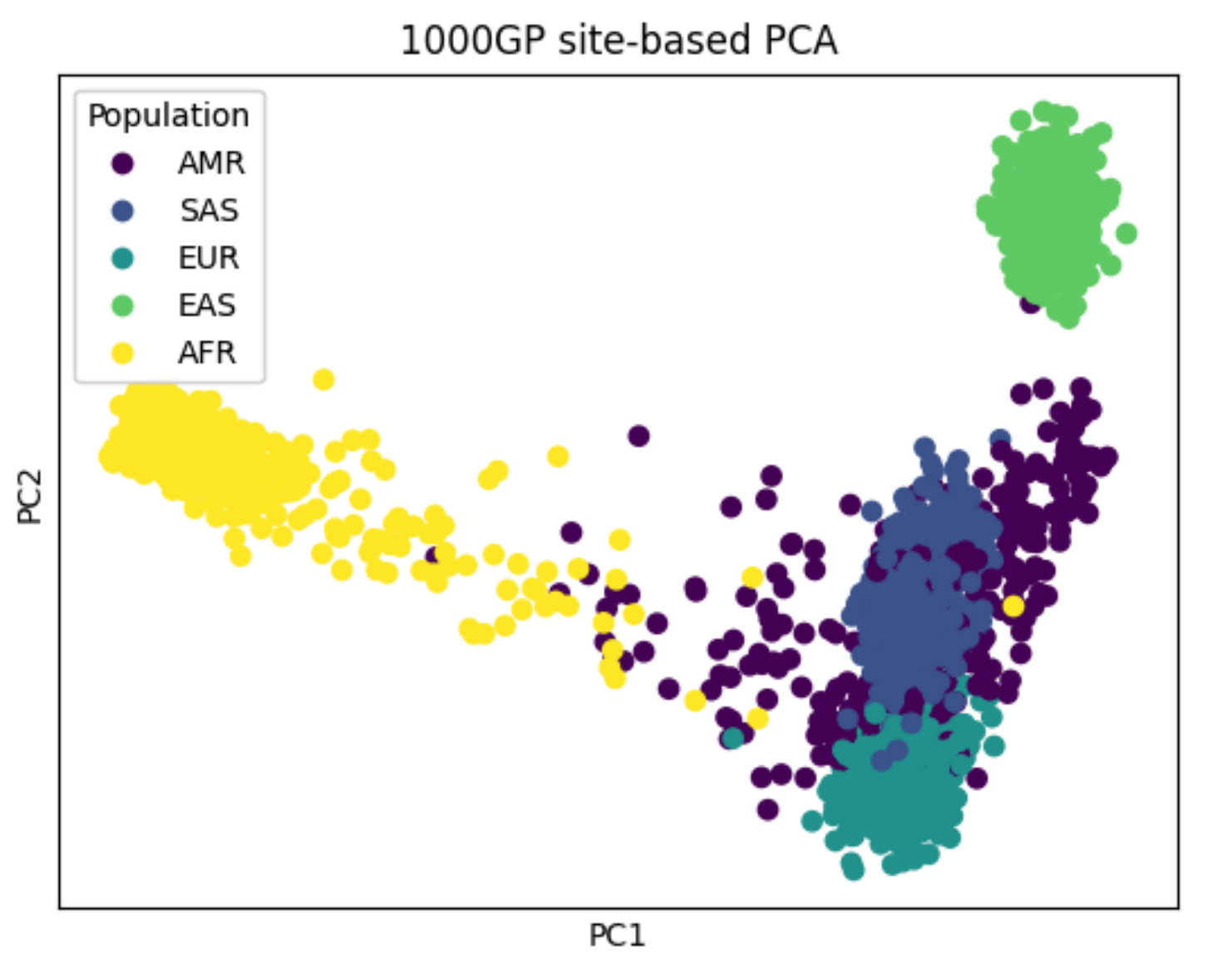
$\Sigma_{ab} =$ `ts.genetic_relatedness([(a1,a2),(b1,b2)], (0,1), ‘site’)`

$$\Sigma_{site} = \begin{pmatrix} 0.89 & 0.22 & -1.11 \\ 0.22 & 0.56 & -0.78 \\ -1.11 & -0.78 & 1.89 \end{pmatrix} \qquad \Sigma_{branch} = \begin{pmatrix} 5.33 & 1.33 & -6.67 \\ 1.33 & 5.33 & -6.67 \\ -6.67 & -6.67 & 13.33 \end{pmatrix}$$

PCA on tree sequences

We study the inferred tree sequence based on 3,601 modern and 8 ancient human genomes [3]. In python, we can define a linear operator based on `ts.genetic_relatedness()`, and use common linear algebra libraries (`scipy`) to perform site- and branch-based principal components analyses.


1000GP branch-based PCA


1000GP site-based PCA

We're actively working on efficient genetic relatedness computations in tskit – stay tuned!

References

[1] Speed, D., Balding, D. Relatedness in the post-genomic era: is it still useful?. *Nat Rev Genet* **16,** 33–44 (2015). doi.org/10.1038/nrg3821

[2] Fan et al. A genealogical estimate of genetic relationships. *bioRxiv (2021)* doi.org/10.1101/2021.08.18.456747

[3] A.W. Wohns et al. A unified genealogy of modern and ancient genomes. *Science* **375,** 6583 (2022). doi.org/10.1126/science.abi8264

Brieuc Lehmann
@brieuclehmann
b.lehmann@ucl.ac.uk
brieuclehmann.github.io