Measuring genetic differentiation from pooled population samples

Valentin Hivert, Raphaël Leblois, Eric Petit, Mathieu Gautier, Renaud Vitalis

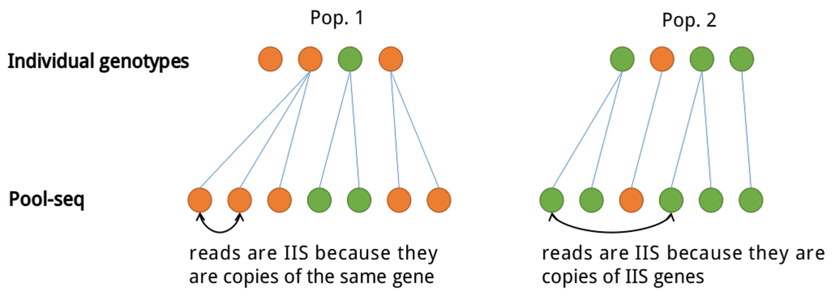
Measuring genetic differentiation from pooled population samples

Valentin Hivert[1,†,*], Raphaël Leblois[1,†], Eric J. Petit[2], Mathieu Gautier[1,†,§], and Renaud Vitalis[1,†,§]

[1]CBGP, INRA, CIRAD, IRD, Montpellier SupAgro, Univ Montpellier, Montpellier, France
[2]ESE, Ecology and Ecosystem Health, INRA, Agrocampus Ouest, Rennes, France
[†]Institut de Biologie Computationnelle, Univ Montpellier, Montpellier, France
[§]These authors are joint senior authors on this work
[*]**valentin.hivert@inra.fr**

Background

- To characterize genetic diversity at a population level, sequencing pools of individual DNAs (Pool-seq) was recently proposed as a valuable and cost-effective alternative to individual genotyping.
- F_{ST}, which measures the extent of differentiation between populations, **is best defined as the intraclass correlation for gene frequencies.** Intraclass correlations may be estimated following an analysis-of-variance framework [5] or, equivalently, by measuring the probability of identity between pairs of genes within and between demes.
- In Pool-seq experiments, because individual genotypes are not observed, distinct reads may be identical because they were sequenced from the same gene, or because they were sequenced from distinct, yet IIS genes.



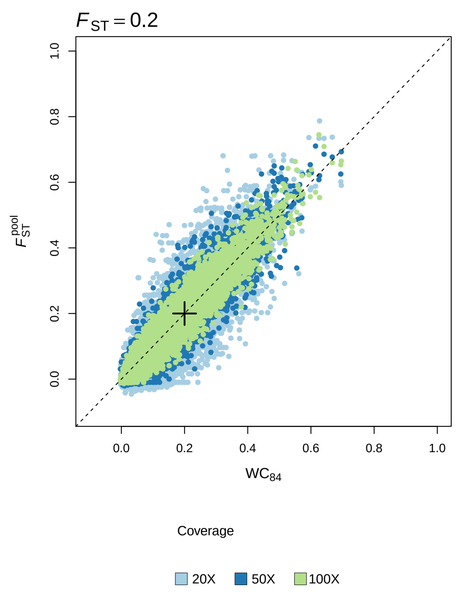
A new estimator of F_{ST} for Pool-seq data

- **Appropriate estimators of differentiation parameters must account for both the sampling of individual genes from the pool and the sampling of reads from these genes.**
- We have developed $\hat{F}_{ST}^{pool}$, a new estimator of F_{ST} for Pool-seq data, in an analysis-of-variance framework [2] (a QR code that gives access to the published article is provided in the poster title area).
- We show that, in the limit case where all pools have the same size n:

$$\hat{F}_{ST}^{pool} = 1 - \left(\frac{1-\hat{Q}_1^r}{1-\hat{Q}_2^r}\right)\left(\frac{n}{n-1}\right)$$

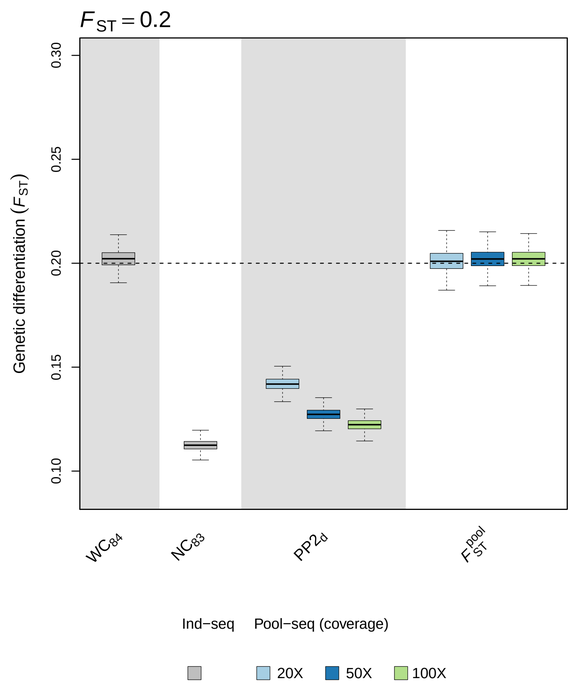
where $\hat{Q}_1^r$ and $\hat{Q}_2^r$ are the frequencies of identical pairs of reads within and between pools, respectively, computed by simple counting of IIS pairs.

Comparing $\hat{F}_{ST}^{pool}$ with inferences based on genotype data



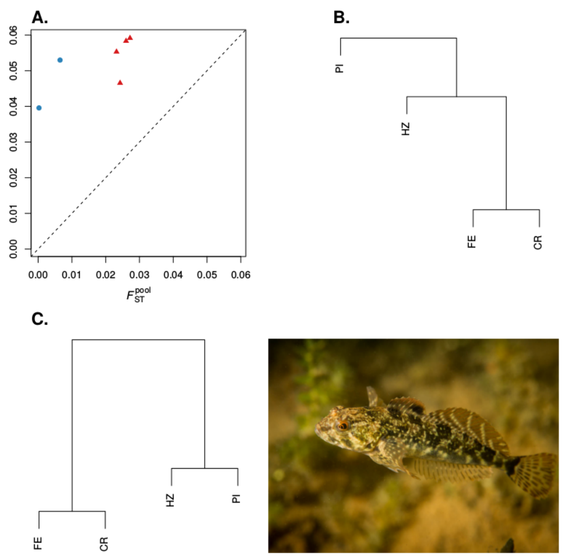
- We compared single-locus estimates of $\hat{F}_{ST}^{pool}$ from pooled data to single locus estimates of F_{ST} based on individual genotypes, using the Weir and Cockerham (1984) [5] estimates (WC_{84}).
- We found that single-locus estimates $\hat{F}_{ST}^{pool}$ are highly correlated with WC_{84} computed on individual data.
- Lastly, the variance of $\hat{F}_{ST}^{pool}$ decreases as the coverage increases.

Comparing $\hat{F}_{ST}^{pool}$ with alternative estimators



- From a simulation study, we found that the accuracy of multilocus $\hat{F}_{ST}^{pool}$ estimators is barely distinguishable from that of multi-locus WC_{84} estimates computed on individual data [5]; furthermore, the accuracy does not depend on the coverage.
- Contrastingly, our analyses showed that the default estimator ($PP2_d$) implemented in `Popoolation2` [3] is biased, and that the extent of the bias depends on the coverage. It converges to the Nei and Chesser's estimator (NC_{83}) [4] as the coverage increases

Application example



- We reanalysed the Pool-seq data published by Dennenmoser et al. [1], who investigated the adaptive genomic divergence between freshwater and brackish-water ecotypes of the prickly sculpin (*Cottus asper*) in Northwestern North-America.
- Comparing pairwise estimates $PP2_d$ [3] and $\hat{F}_{ST}^{pool}$ [2], we found that $\hat{F}_{ST}^{pool}$ (but not $PP2_d$) revealed lower differentiation within (blue dots) than between (red triangles) ecotypes.
- We further found a clear-cut clustering of the estuarine (CR and FE) and freshwater (PI and HZ) samples using the estimator $\hat{F}_{ST}^{pool}$ (C) as opposed to the analyses based on $PP2_d$ (B)
- Our result is in agreement with previous microsatellite-based studies that showed higher genetic differentiation between ecotypes rather than within ecotypes.

Take home message

We developed an unbiased estimator of F_{ST} for Pool-seq data, in an analysis-of-variance framework.

- **The accuracy is barely distinguishable from the analysis-of-variance estimator for individual data [5].**
- **The accuracy does not depend on the coverage or on the pool size.**
- **Although our estimator is sensitive to uneven contributions of individual DNAs in each pool, we found that it was robust to unequal sample sizes and variable coverages.**

Package poolfstat: Computing F-Statistics from Pool-Seq Data

The R package poolfstat includes functions for the computation of F-statistics from Pool-Seq data in population genomics studies.
It is available at the Comprehensive R Archive Network (CRAN) :

https://cran.r-project.org/web/packages/poolfstat/index.html

Acknowledgements

Valentin Hivert's PhD is funded by the ERA-Net BiodivERsA project EXOTIC and the French National Institute for Agricultural Research ("Plant Health and Environment" division). Part of this work was supported by the ANR project SWING (ANR-16-CE02-0015) of the French National Research Agency, and by the CORBAM project of the French region Hauts-de-France.

References

[1] S. Dennenmoser et al. "Adaptive genomic divergence under high gene flow between freshwater and brackish-water ecotypes of prickly sculpin (*Cottus asper*) revealed by Pool-Seq". In: *Mol. Ecol.* 26 (2017), pp. 25–42.
[2] V. Hivert et al. "Measuring Genetic Differentiation from Pool-seq Data". In: *Genetics* (2018). in press. doi: 10.1534/genetics.118.300900.
[3] R. Kofler, R. V. Pandey, and C. Schlötterer. "PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)". In: *Bioinformatics* 27 (2011), pp. 3435–3436.
[4] M. Nei and R. K. Chesser. "Estimation of fixation indices and gene diversities". In: *Ann. Hum. Genet.* 47 (1938), pp. 253–259.
[5] B. S. Weir and C. C. Cockerham. "Estimating F-statistics for the analysis of population structure". In: *Evolution* 38 (1984), pp. 1358–1370.