Performance evaluation of adaptive introgression classification methods

Jules Romieu, Ghislain Camarata, Pierre-André Crochet, Miguel de Navascués, Raphaël Leblois, François Rousset

Performance evaluation of adaptive introgression classification methods

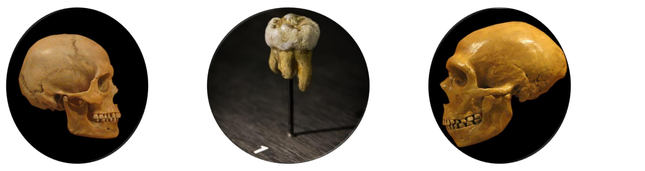
Jules Romieu[1,2], Ghislain Camarata[1,2], Pierre-André Crochet[3], Miguel de Navascués[2], Raphaël Leblois[2], and François Rousset[1]

[1]ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France, [2]CBGP, INRAE, CIRAD, IRD, Institut Agro, Univ Montpellier, Montpellier, France, [3]CEFE, CNRS, Univ Montpellier, EPHE, IRD, Montpellier, France

1 Background and objective

Current adaptive introgression classification methods

Developped and tested → **Human lineages**



Homo sapiens — *Denisovan* — *Homo neanderthalensis*

Recent divergence time

Performance ? → ***Other lineage***

Podarcis genus

Ancient divergence time

Objective: Performance tests of AI methods on simulated datasets under *Podarcis* scenarios

2 Summary statistic and methods tested

Q95 (1%,100%) (Racimo *et al.* 2017)

- **Summary statistic**
- 50kb windows
- **High frequency** of alleles uniquely shared between the donor and the recipient populations
- Uses samples from 3 populations

VolcanoFinder (Setter *et al.* 2020)

- **Composite likelihood-based** genomic scan
- One test site every kb
- **Excess intermediate frequency polymorphism** in the flanking region of the genome
- Uses only the **recipient population** sample

Simulation-based classification methods

Genomatnn (Gower *et al.* 2021)

- **Deep learning** (Convolutional Neural Network)
- 100kb windows
- **Black box** (train on genotype matrix)
- Uses samples from 3 populations

MaLAdapt (Zhang *et al.* 2023)

- **Machine learning** (Extra Trees Classifier)
- 50kb windows
- Train on **summary statistics** (alleles shared between donor and recipient, linkage desequilibrium, genetic diversity, etc)
- Uses samples from 3 populations

3 Simulation test approach

Simulation	Simulated test sets	Classification methods	Classification metrics
Backward + Forward ms prime + SLiM	Simulated genome: Known states: Non-AI window — AI window	• Q95[1] (Summary statistics) • VolcanoFinder[2] (Composite likelihood) • Genomatnn[3] (Deep learning) • MaLAdapt[4] (Machine learning)	Known states Versus Predicted states ROC, TPR, AUC, FPR

4 Demographic model

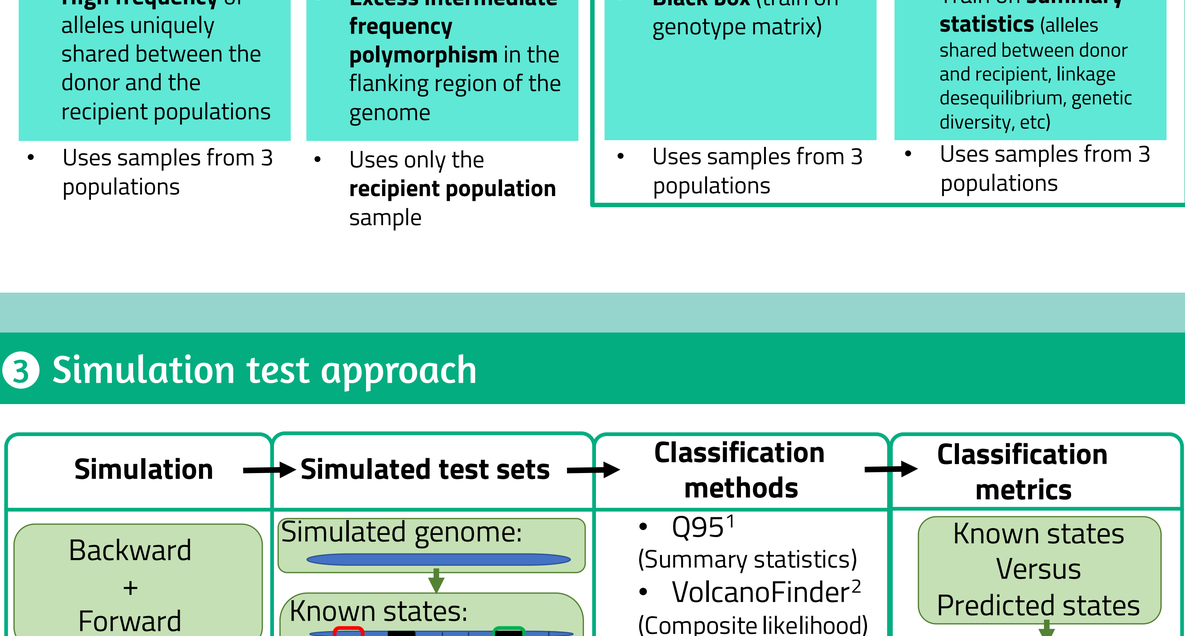
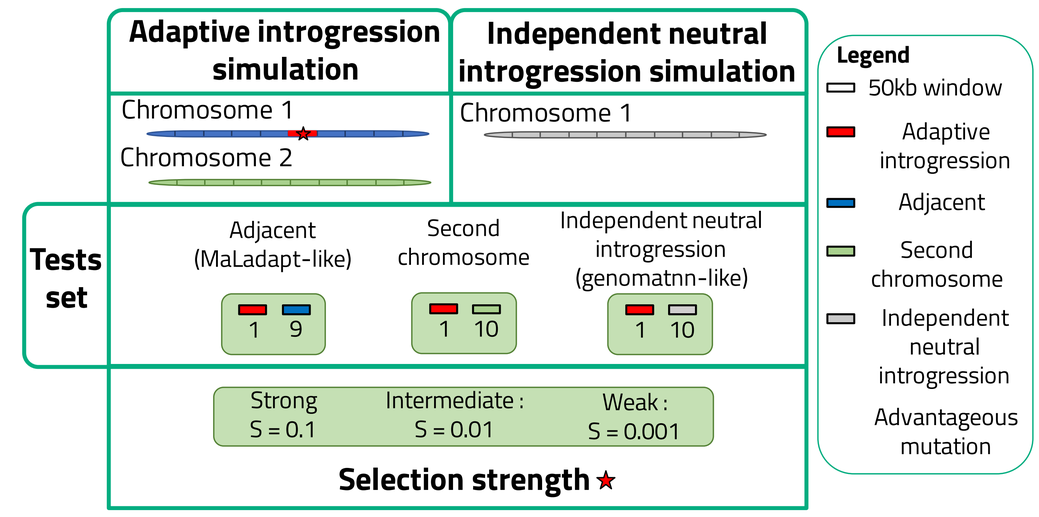


Figure 1 – Demographic model used to generate simulated dataset. Anc_O_I = Outgroup and Ingroup ancestral population, O = Outgroup population, I = Ingroup population, Anc_D_DS = Donor and non-introgressing ancestral population, Anc_R_RS = Recipient and non-introgressed ancestral population. DS = sister to the donor population (non-introgressing population), D = Donor population (introgressing population), R = Recipient population (introgressed population), RS = Recipient sister population (non-recipient population), Population size (N) = 10,000.

5 Importance of genetic architecture

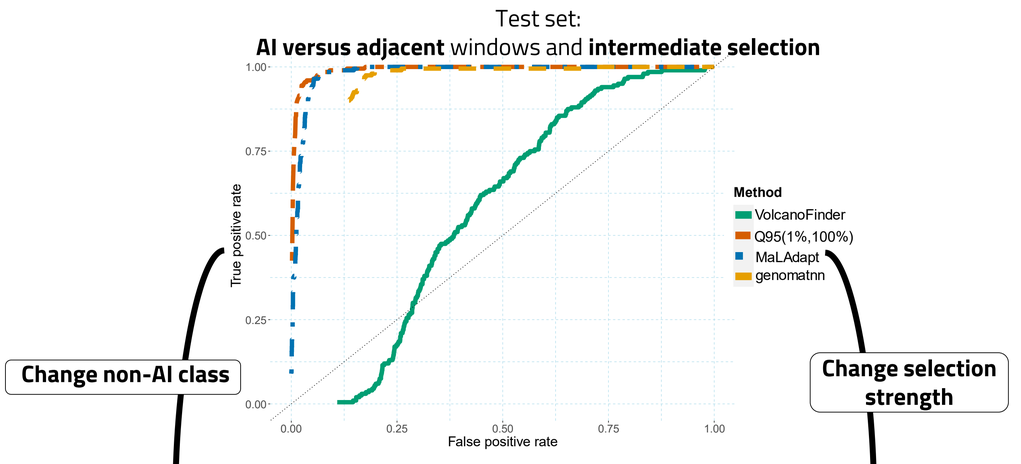
Hitchhiking effect on method's performance?

→ **non-AI window types : adjacent, second chromosome and independent neutral introgression**



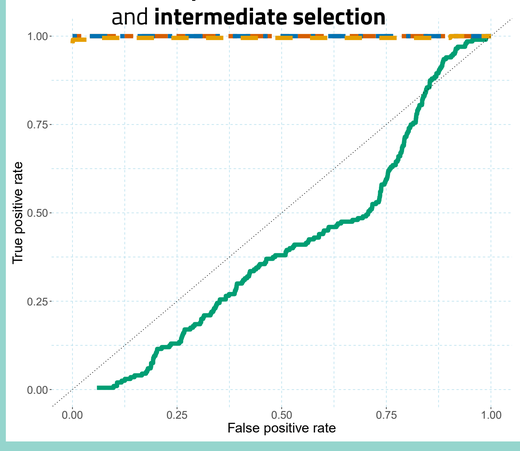
6 Results: ROC curves

Test set:
AI versus adjacent windows and **intermediate selection**



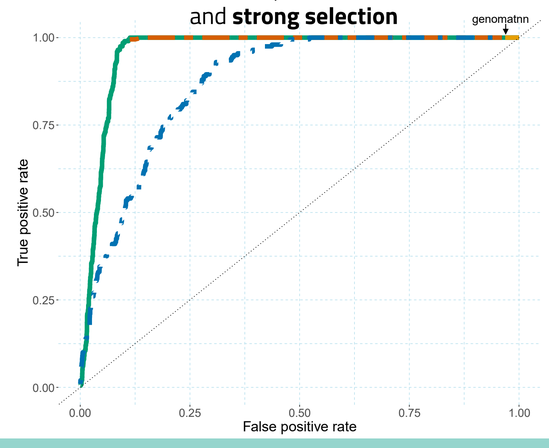
Change non-AI class

Test set:
AI versus independent neutral windows and **intermediate selection**



Change selection strength

Test set:
AI versus adjacent windows and **strong selection**



7 Discussion and conclusion

- Best method in our tests: Q95 summary statistic
- Better performance with strong selection but leads to an increase in hitchhiking effect

  → Take adjacent windows into account

- Simulation-based inference methods: Factors that have negative impacts on performance:
  - Test sets with non-AI windows types non-used in train sets (ex :genomatnn)
  - Misspecification of demographic model used in train sets

  → Importance to take into account different types of neutral windows in trained sets

- Classification methods problems:
  - Define a threshold to discriminate AI/non-AI windows
  - Necessity to use FDR control methods

  → Solution: Develop methods to estimate genomic-level introgression

Bibliography

[1]Racimo F, Marnetto D, Huerta-Sánchez E (2017). Signatures of Archaic Adaptive Introgression in Present-Day Human Populations. Molecular Biology and Evolution 34, 296–317.

[2]Setter D, Mousset S, Cheng X, Nielsen R, DeGiorgio M, Hermisson J (2020). VolcanoFinder: Genomic scans for adaptive introgression. PLOS Genetics 16. e1008867.

[3]Gower G, Picazo PI, Fumagalli M, Racimo F (2021). Detecting adaptive introgression in human evolution using convolutional neural networks. eLife 10. e64669.

[4]Zhang X, Kim B, Singh A, Sankararaman S, Durvasula A, Lohmueller KE (2023). MaLAdapt Reveals Novel Targets of Adaptive Introgression From Neanderthals and Denisovans in Worldwide Human Populations. Molecular Biology and Evolution 40, msad001.