# Two M2 internships on Deep Learning for population genetics

**Supervisors: Flora Jay, Burak Yelmen, Guillaume Charpiat, Cyril Furtlehner**
**Contact: flora.jay@lri.fr** ; web: https://flora-jay.blogspot.com/

**Location**: LISN (Paris-Saclay University), machine learning and bioinformatics groups

**When/Duration**: 2022, preferably 5-6months for M2 students. Outstanding M1 applicants will be considered as well.

We are looking for two highly motivated interns to work on the follow-up research of two of the lab papers:
- "Creating artificial human genomes using generative neural networks", Yelmen et al 2021
- “Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation”, Sanchez et al 2020

These projects could lead to a PhD thesis afterwards.

**Requirements**
The ideal candidate should at least be good at python scripting and basic machine learning/statistics concepts. Experience with deep learning is a plus. Familiarity with genomics, population genetics, generative models, bash scripting and high-performance computing is not mandatory but a plus.

Salary: the regular stipend for internships in academia is ~550eur/month.

* **Subject 1: Creating artificial human genomes using generative neural networks**

*Context : Using machine learning, we could generate synthetic genomes that successfully mimic the real ones but are not identical to any of them. We relied on two type of neural network architectures that we trained on human genetic databases: (1) Generative Adversarial Networks, that were a breakthrough in the domain of computer vision, allowing the generation of extremely realistic images; (2) Restricted Boltzmann Machines, another family of generative models capable of learning complex data distributions. We measured the quality of the generated genomes in terms of data hidden structure, population structure, linkage disequilibrium, haplotype diversity, etc. and demonstrated that they provided an accurate representation of the real ones. Without duplicating any of the individuals, most key characteristics of the data were conserved. Additionally, we showed that releasing these genomes would lead to a privacy gain. A direct implication is the increase in richness of public datasets, e.g. with populations still under-represented in genetic studies.*

**Internship tasks**

The multidisciplinary project combines genomics, population genetics and machine learning with two major focal points, depending on the candidate's background:

1) Improving the proposed generative models and implementing new variants for genomic data. In particular, the original GAN architecture will be modified to scale to very large genomic data. The intern will implement one or multiple models adapted from high-resolution image or text generation, and will compare them to the baseline.

2) Extending the applicability of generative models for various other genomic tasks such as imputation and Genome Wide Association Studies (GWAS).

* **Subject 2: Interpreting neural networks for population genetic inference**

*Context. Our lab and others have recently developed inference methods based on deep learning to infer evolutionary models directly from genetic data (e.g. for reconstructing demographic history, Sanchez et al 2020). We have also developed a generic software, dnadna, for implementing and applying neural networks in population genetics (Sanchez, Bray et al). Despite their good performances, neural networks are often (understandably) criticized for their lack of interpretability. Interpretability is not only crucial for explaining a prediction, but also for avoiding artifact and biases. As for neural networks, theoretical and applied research in this direction is moving fast. Many methods address the question of what information the network uses globally to construct a model, or what information contributed to the output for a particular example. Yet very few have been applied to DNN based on genetic data for population genetic inference. On the other hand, interpretability has been investigated in a more traditional setup, where SNP data are reduced into handcrafted expert statistics that do not contribute equally to the prediction.*

**Internship tasks**
Depending on the candidate's background:

- Reviewing interpretability methods (e.g. Shrikumar et al 2017, ...) that are relevant to population genetics tasks
- Testing the direct application of state-of-art approaches in computer vision and genomics to population genetics.
- Developing and testing a novel approach based on dimensionality reduction for identifying the link between handcrafted features and neural networks (through linear correlation analysis, for instance). This information will be cross-checked with prediction quality and neuron importance.

*** References**

B Yelmen, A Decelle, L Ongaro, D Marnetto, C Tallec, F Montinaro, C Furtlehner, L Pagani, F Jay. Creating Artificial Human Genomes Using Generative Models. PLoS genetics, 17(2), e1009303. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009303

T Sanchez, J Cury, G Charpiat, F Jay (2020). Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. Molecular Ecology Ressources DOI:10.1111/1755-0998.13224
Link:
https://www.researchgate.net/profile/Flora-Jay/publication/342822922_Deep_learning_for_population_size_history_inference_Design_comparison_and_combination_with_Approximate_Bayesian_Computation/links/5f353ad192851cd302f1829c/Deep-learning-for-population-size-history-inference-Design-comparison-and-combination-with-Approximate-Bayesian-Computation.pdf

*T Sanchez*, EM Bray*, P Jobic, J Guez, G Charpiat, J Cury°, F Jay° (2021) Dnadna: Deep Neural Architecture for DNA - A deep learning framework for population genetic inference*
*https://hal.archives-ouvertes.fr/hal-03352910*

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." International Conference on Machine Learning. PMLR, 2017.