A PIPELINE TO DETECT GENOMIC REGIONS UNDER NATURAL SELECTION IN 1000GENOMES DATA

Marc Pybus+ [1], Pierre Luisi+ [1], Giovanni M Dall'Olio+ [1], Manu Uzkudun+ [1], Pavlos Pavlidis[2], Ferran Casals[1], Hafid Laayouni[1], Jaume Bertranpetit* [1], Johannes Engelken* [1,3]

[1]Institute of Evolutionary Biology, CSIC-UPF, Barcelona, 08003, Spain,
[2] Max Planck Institute for Evolutionary Anthropology, Department of Evolutionary Genetics, Leipzig, 04103, Germany
+ Equal contribution
* Equal contribution

A major challenge in population genetics is the inference of naturally selected genetic variants. Recent progress in the development of new statistics as well as more accurately inferred demographic histories and unprecedented amounts of data hold great promise for the future. Specifically, the human 1000genomes project aims to make available the majority of genetic variants with an allele frequency of >1% from numerous worldwide populations (www.1000genomes.org).

We have implemented a large number of tests for natural selection in a bioinformatic pipeline, including XP-CLR, ω, CLR, Tajima´s D, Fay & Wu´s H, Fu & Li´sD, iHS, δiHH, XPEHH, Fst, δDAF etc. These statistics are mainly based on population differentiation, long range haplotype and allele frequency spectrum models. We used extensive coalescent simulations of neutral and selected genomic regions in order to evaluate each statistic for (i) their sensitivity to detect selection and for (ii) their power to localize the center of selection as well as for (iii) their robustness in diverse demographic scenarios.

Further, by combining the statistics in a single composite score through machine learning, we both improve the power and facilitate the interpretation of the diverse tests for selection. Specifically, we have obtained and tested classifiers that are optimized for detecting regions with different selective scenarios, including complete and partial sweeps as well as differentiating ancient and recent selection events. We have applied our methods successfully to experimental data from the 1000genomes project and have found interesting new patterns of selection. Results are visualized on a local version of the UCSC genome browser. We will discuss the latest results that we obtain from this ongoing project, showing that different layers of selection can be observed in the human genome.