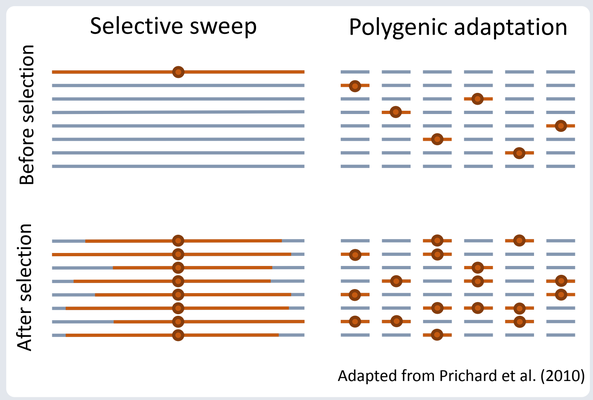
Detecting gene subnetworks under polygenic selection

Alexandre Gouy[1,2], Laurent Excoffier[1,2]
[1]CMPG, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland
[2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

INTRODUCTION

- **Adaptation** is typically viewed as involving selective sweeps at a single locus. But for **quantitative traits**, we expect to observe modest changes in allele frequencies at many loci.
- Tests to detect selection from genomic data based on **single-locus selective sweep** models can be challenged by these **small allele frequency changes at many loci** which may remain below the detection limit of most of these methods.



- This method aims at detecting polygenic selection. The general idea is to **search for subnetworks of genes within biological pathways** that present unusual features.
- This search is a typical **combinatorial optimization** problem that can be solved using **simulated annealing**. The significance test procedure explicitly takes into account this optimization process.
- We searched evidence for **convergent adaptation to altitude in humans**

WORKFLOW

INPUT

Gene scores
Statistic as a **proxy for selection**: F_{ST}, PBS, …

Pathway data
Graphs of genes and their interactions (KEGG, Reactome, NCI)

SEARCH ALGORITHM

Background distribution
Generate the **subnetwork scores distributions for each possible network size** by sampling random subnetworks.

High-scoring subnetworks search
Search in each pathway for a high-scoring subnetwork using **simulated annealing**. The **score *s* is standardized** at each iteration using the background distribution.

Subnetwork score

$$s = \frac{1}{\sqrt{k}} \sum_{i}^{k} g_i$$

$$z = \frac{s - \mu_k}{\sigma_k}$$

$g_i \leftrightarrow$ gene score
$k \leftrightarrow$ subgraph size
$\mu_k \leftrightarrow$ mean of background distribution
$\sigma_k \leftrightarrow$ SD of background distribution

PERMUTATION TEST

For each class of connectivity:
- **Permute** the gene scores;
- **Search** for high-scoring subnetworks in permuted data;
- Repeat *n* times to generate the high-scores **null distribution used to infer p-values**.

High-scoring subnetworks | **Scores** | **Empirical p-values**

STATISTICAL CORRECTIONS

Correction for network overlap
As pathways share some genes, the high-scoring **subnetworks are** also **overlapping**. To correct for this effect, we define **clusters** of subnetworks based on Jaccard's **similarity index**, and only keep the highest scoring subnetwork per cluster.

Correction for multiple tests
We expect to remove autocorrelation among p-values with the overlapping correction, allowing us to perform a multiple testing correction on remaining networks using the **FDR method**.

OUTPUT

For each biological pathway:
- if existing, the **high-scoring subnetwork** identified (list of genes);
- its **score** and associated empirical **p-value**;
- if the network resists to overlap correction, a **q-value**.

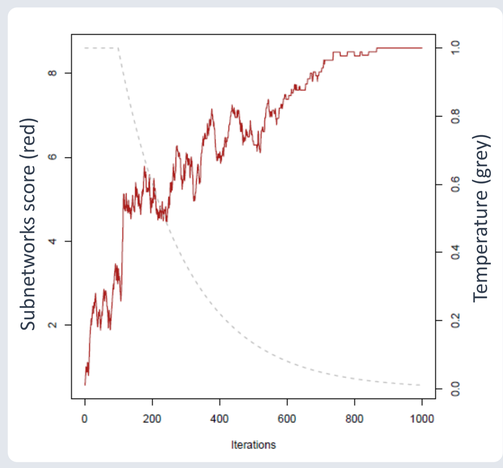
DATA

- 906,600 SNPs for **Tibetans** and **Andeans** populations living at **high altitude**, from Bigham et al. (2012).
- **Probability of convergent adaptation** estimated for each SNP using a hierarchical Bayesian model, from Foll et al. (2014).
- **17,272 genes** in **1,509 pathways** have been **tested**.
- 3 pathways databases: KEGG, Reactome and NCI.
- Pathways with biggest connected components of size < 10 are excluded.

SIMULATED ANNEALING

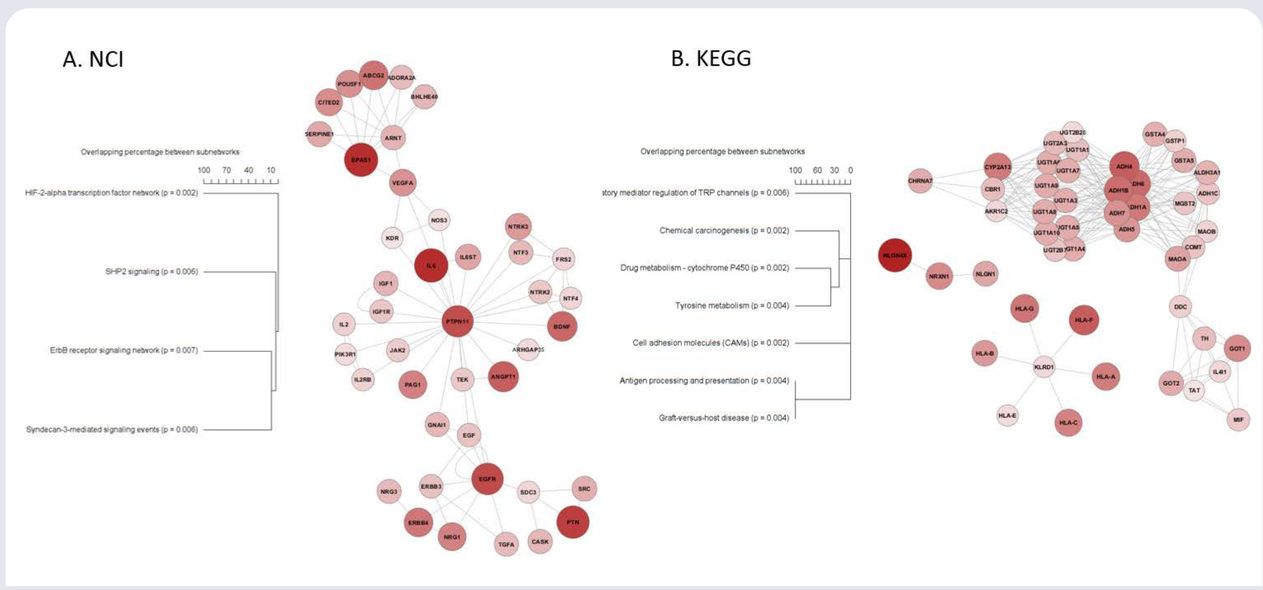
For a graph $G(V, E)$ with V nodes and E edges, N iterations and a temperature function T_i , which decreases geometrically:

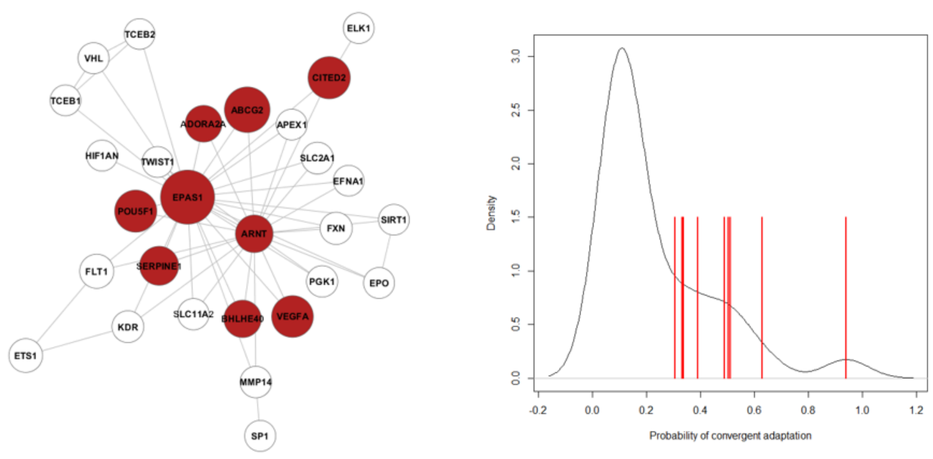
1. Select a **random active subgraph** of $k_{\min}$ nodes $v \in V$
2. Randomly **pick a node** $v \in V$ from the boundary and bordering genes and **update its state** (active ↔ inactive)
3. Compute the subgraph normalized **score** z_{k_i}
4. **Keep new v with a probability** $P = \min(1, \exp\left(\frac{z_{k_i} - z_{k_{i-1}}}{T_i}\right))$
5. If $i < N$, go back to 2.
6. Return the final subnetwork



MAIN RESULTS

- **Convergent evolution** of networks of genes involved in **response to hypoxia** in Tibetans and Andeans
- Different types of adaptive responses: **vascular** (angiogenesis, …), **neural** (response to glutamate toxicity, neurogenesis), **metabolic** (ADH cluster)





References

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J. M., Mei, R., ... & Brutsaert, T. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet*, *6*(9), e1001116.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., & Excoffier, L. (2014). Widespread signals of convergent adaptation to high altitude in Asia and America. The American Journal of Human Genetics, 95(4), 394-407.

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current biology, 20(4), R208-R215.

Contact
alexandre.gouy@iee.unibe.ch

Acknowledgments
We would like to thank J. Daub and I. Dupanloup.
This project is partially funded by the Swiss NSF.