# Modelling the co-evolution of genes in bacterial genomes to discover the hidden ecology of accessory genes and mobile genetic elements.

**Florent Lassalle** , Xavier Didelot.
Imperial College London, Deptartment of infectious Disease Epidemiology - MRC Center for Outbreak Analysis and Modelling.
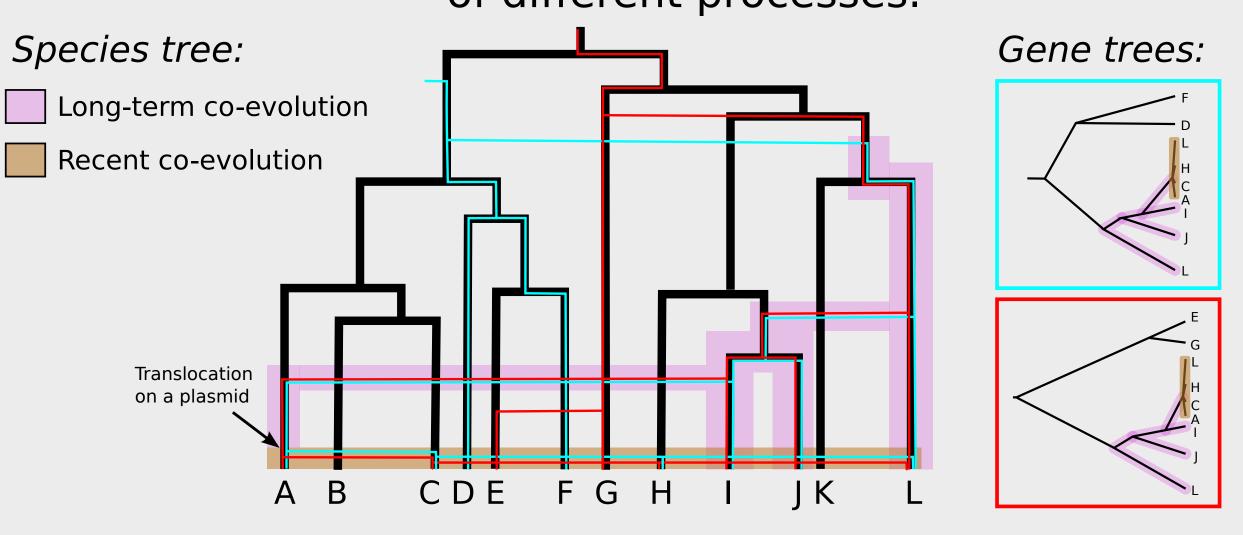
In bacteria, so-called "accessory" genes are present in only a fraction of members of the same species. They can confer important adaptations, such as resistance to an antibiotic, the ability to establish symbiosis with a host plant, etc. These functions are usually only transiently relevant to the ecology of the host bacterium - during antibiotic treatment or in the proximity of a compatible symbiotic partner - yet these accessory genes are observed in many bacteria living away from such selective conditions. Persistence of accessory genes in the "gene pool" of a (community of) species could simply be explained by the fact that they are often carried by mobile genetic elements (MGEs) like plasmids, whose selfish transmission would counter-balance their lack of adaptive value for the host.

However, the conserved association between several accessory genes suggests another hypothesis: the versatile repertoire of biological functions gathered on MGEs can prove adaptive under a variety of environmental conditions, and rare, but periodic, positive selection may drive the survival of the vector and its cargo, and its propagation across various genetic backgrounds. In addition, long-term co-evolution of constituent genes may lead to the establishment of epistatic interactions, which in turn would promote the maintenance of physical linkage between genes.

To study how genes aggregate into evolutionary successful MGEs, I developed a pipeline for the reconstruction of the evolutionary scenarios (with duplication and horizontal transfer) of all genes in a pangenome, and applied it to a dataset of 880 Enterobacteriaceae genomes. Co-evolving gene modules were defined based on the length and significance of shared evolutionary paths. The study of gene module associations and resulting functional combinations, and the evaluation of this pattern under neutral vs. co-selection models, will help unravelling the ecological structure of bacterial pangenomes. This tool will soon be made available as an open-source, platform-independent software.

## PANGEREC: A bioinformatic pipeline for phylogenetic reconciliation of gene histories across bacterial pangenomes.



GenBank **RefSeq** (all available complete or high quality genomes in target clade)

Extract all coding sequences

**MMSeqs[1], ClustalΩ[2]**

Cluster into homologous gene families and align

Concatenate core gene alignments

**RAxML[3]**

Compute reference species tree (core genome tree)

Group closely related species (strains) into populations

(For each gene family:)

**RAxML[3]**

Compute preliminary ML gene trees (all sequences)

Sample gene tree topologies (backbones)

Separate "rake" clades of closely related sequences from the gene tree backbones

**MrBayes[4]**

**ALE[5]**

Reconcile gene tree backbones with species tree = infer HGT → and duplication ■ events & ancestral genome content

Tabulate occurence of species populations in clades of closely related sequences = clusters of recent gene sharing

| | Pop A | Pop B | Pop C | Pop D | Pop E | Pop F | ... |
|---|---|---|---|---|---|---|---|
| fam1-clade1 | 1 | 0 | 2 | 0 | 1 | 0 | |
| fam1-clade2 | 0.25 | 0 | 1 | 0 | 0 | 0 | |
| fam2-clade1 | 1 | 0 | 2 | 0 | 0.5 | 1 | |
| fam2-clade2 | 0 | 0.33 | ... | 0 | 1 | 1 | |
| ... | | | ... | | | | |

### OUTPUT A:

• Compute co-evolution score for all pairs of gene lineages (or just for physically linked genes) based on cumulative gene presence overlap over the species tree.

• Identify modules of genes sharing high co-evolution scores and test significance of association using **EpiCs[6]**.

### OUTPUT B:

• Compute co-occurence score for all pairs of clades and for all pairs of populations.

• Identify modules of co-segregating genes sharing high co-occurrence scores.

• Test for preferential gene sharing amongst species populations.

## Long-term vs. recent co-evolution are signatures of different processes.



*Species tree:*
▨ Long-term co-evolution
▨ Recent co-evolution

*Gene trees:*

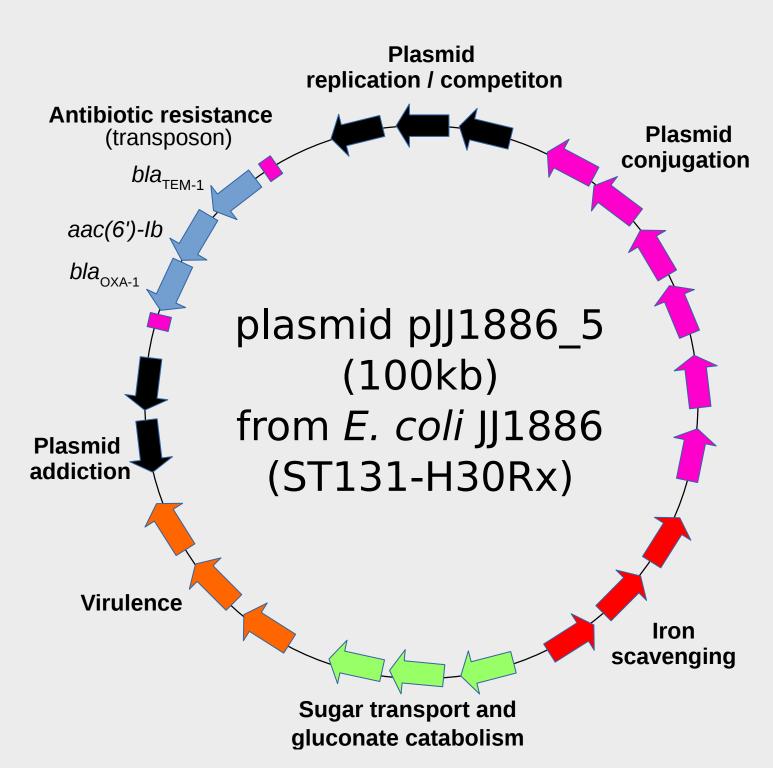Translocation on a plasmid

A B   C D E   F G H   I   J K   L

Conserved association of gene over long times may reflect selective constraints (complementary ecology, epistasis) between co-evolving genes, whereas the joint dissemination of genes over recent times often reflects their selfish spread (via a MGE), and is most likely neutral or only transiently adaptive to the host.

## Case study application: correlated histories of genes on resistance- and virulence-conferring plasmid of *E. coli*

ST131 is the major epidemic lineage of *E. coli* in the USA since 2007, causing severe (sometimes lethal) urinary tract infections. H30-Rx is a hyper-virulent, multi-drug resistant sub-lineage which phenotype is linked to the acquisition of an IncF plasmid.

In order to track the history of aggregation of this super-pathogenic construct, we reconstructed the pattern of co-evolution of gene homologs in the pangenome, then focusing on a set of 23 related plasmids.

In the long term, the plasmid replication, addiction and conjugation genes (backbone) were tightly linked together, and to a lesser extent with catabolic and virulence genes, whereas the antibiotic resistance (AbR) and iron scavenging (Iron) genes are recent additions to the backbone. Catabolism, virulence and Iron cassettes occur in a greater diversity of contemporary plasmid backbones than the AbR cassette, suggesting these function shape the ecological niche of this plasmid lineage.



Plasmid replication / competiton
Plasmid conjugation
Iron scavenging
Sugar transport and gluconate catabolism
Virulence
Plasmid addiction
Antibiotic resistance (transposon)
$bla_{TEM-1}$
$aac(6')-Ib$
$bla_{OXA-1}$

plasmid pJJ1886_5 (100kb) from *E. coli* JJ1886 (ST131-H30Rx)

**A**



(Long-term) co-evolution score: mean number of shared events across sampled scenarios

**B**

(Recent) co-occurence score (Jaccard index)

## References:

1. Hauser et al. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics. 32, 1323–1330.
2. Sievers et al (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539.
3. Stamatakis (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, btu033.
4. Ronquist & Huelsenbeck (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinforma. Oxf. Engl. 19, 1572–1574.
5. Szöllősi et al. (2013). Lateral Gene Transfer from the Dead. Syst. Biol. 62, 386–397.
6. Behdenna et al. (2016). Testing for Independence between Evolutionary Processes. Syst. Biol. 65, 812–823.