LARGE SCALE BENCHMARKING OF PHYLOGENETIC BOOTSTRAP METHODS

A USE CASE WITH NEXTFLOW

Frédéric Lemoine
Gascuel Lab (Evolutionary Bioinformatics), Institut Pasteur, Paris

2017/09/15

Institut Pasteur

Contents

PART 1

Introduction

1.1 Phylogenetics

What is phylogenetics?

"In biology, **phylogenetics** is the study of the **evolutionary history** and relationships among individuals or groups of organisms (e.g. species, or populations). These relationships are discovered through **phylogenetic inference methods** that evaluate observed heritable traits, such as DNA sequences or morphology under a model of evolution of these traits."

1.1 Phylogenetics

Why we study phylogenetics @Pasteur?

We develop new methods to study:

- Phylogeography/Ancestral state reconstruction:
  - e.g. Tracing the origin and the evolution of virus epidemics;
- Drug resistance:
  - e.g. Modelling the emergence and transmission of HIV drug resistance mutations;
- Phylodynamics/Virulence:
  - e.g. Associating genome evolution with virulence

1.1 Phylogenetics

Why we study phylogenetics @Pasteur?

We develop new methods to study:

- Phylogeography/Ancestral state reconstruction:
  - e.g. Tracing the origin and the evolution of virus epidemics;
- Drug resistance:
  - e.g. Modelling the emergence and transmission of HIV drug resistance mutations;
- Phylodynamics/Virulence:
  - e.g. Associating genome evolution with virulence

Pre-requisite

All these analyses first need one or several phylogenetic tree(s)

1.1 Phylogenetics

Main steps in reconstructing a phylogeny

- Collect sequences (First but not least!);
- Build a multiple alignment (T-Coffee, Clustal, Muscle, MAFFT, etc.);
- Clean/Filter the alignment (Noisy, BMGE, Gblocks);
- Infer a tree (PhyML, RAxML, FastME, etc.);
- Assess the robustness of the inferred tree: HOW?

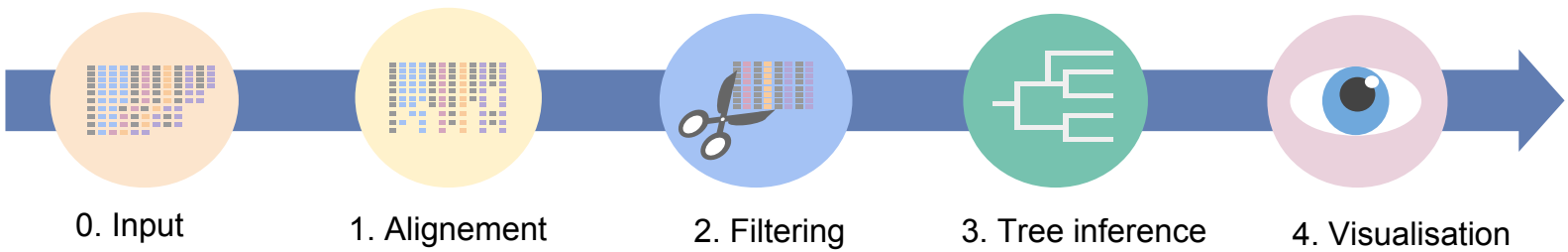


Image from Damien Correia: `http://ngphylogeny.pasteur.fr`

1.2 Bootstrap in phylogenetics

Definition

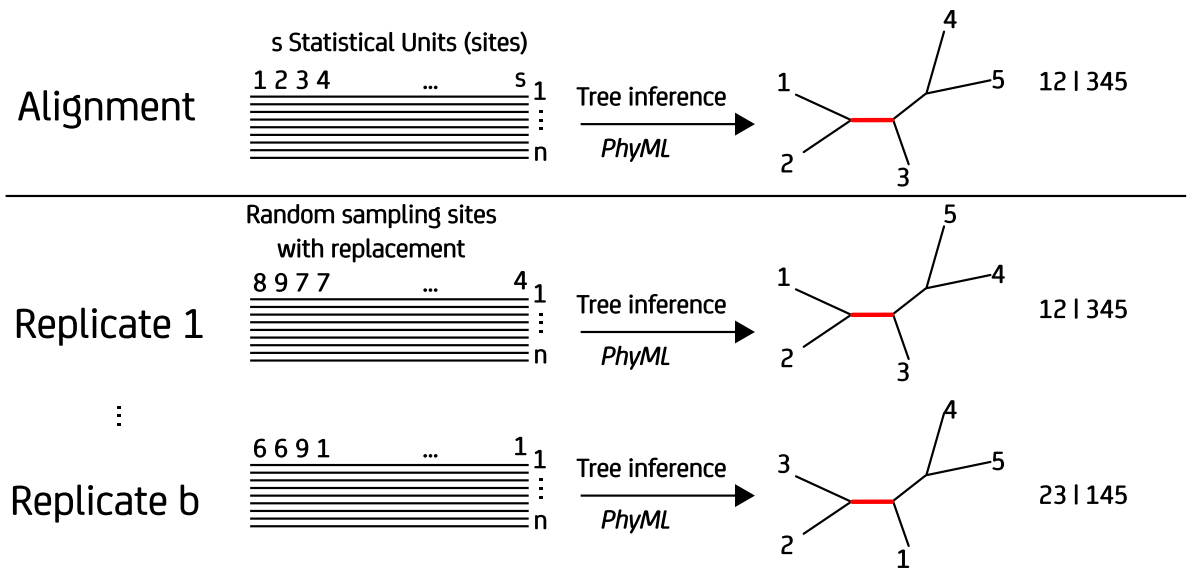
"The bootstrap is a widely used statistical method to study the robustness, bias and variability of numerical estimates (Efron, 1979). It involves **resampling with replacement** from the original dataset to obtain **replications of the original estimate**, and then typically to compute the variance and distribution of this estimate."

```
data=rnorm(500)
boot=unlist(lapply(1:1000,function(b){
          mean(sample(data,size=500,replace=T))
     }))
```

1.2 Bootstrap in phylogenetics

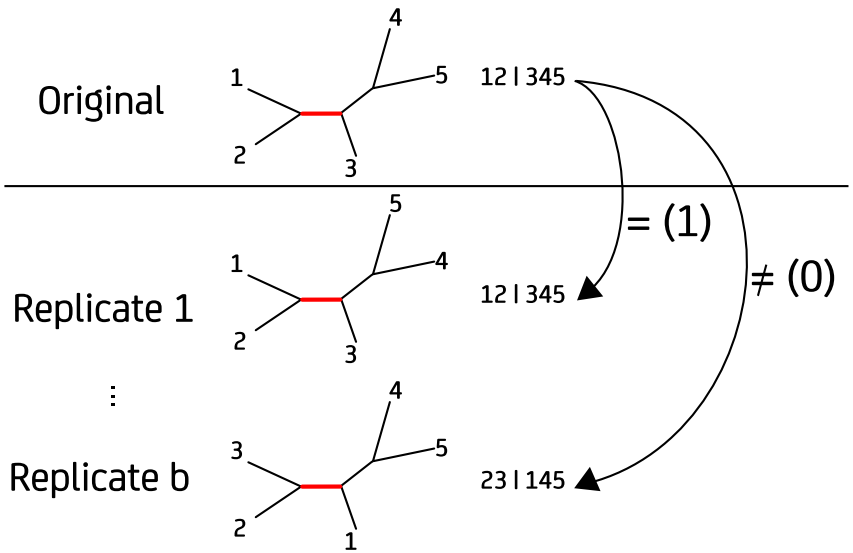
Application to phylogenetics

- Proposed by Felsenstein in 1985 (42nd most cited ever)
- Application of Efron's Bootstrap (1979) in Phylogeny



1.2 Bootstrap in phylogenetics

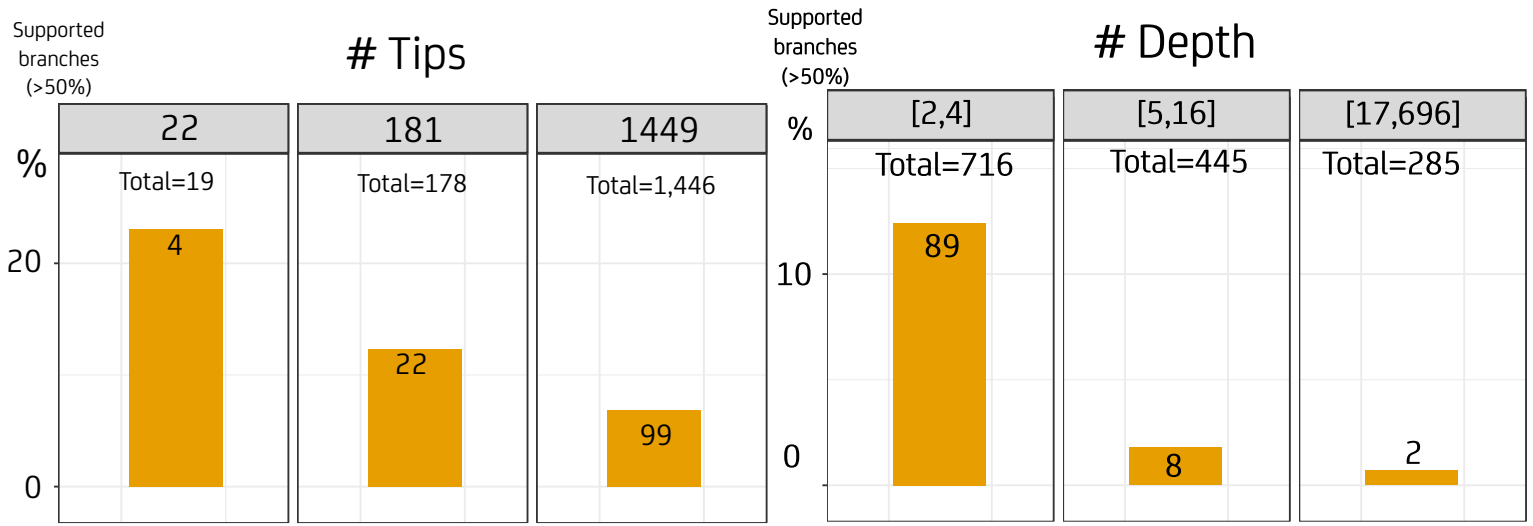
Computation



- Support of a branch is the % of bootstrap trees that contain **the exact same bipartition**: FBP.
- On each bootstrap tree, a branch is either **present (1)** or **absent (0)**

1.2 Bootstrap in phylogenetics

Limitations

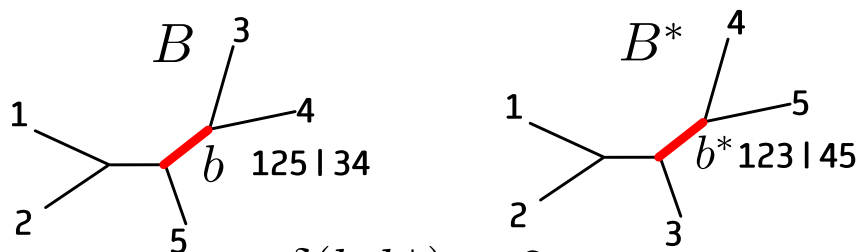


Institut Pasteur

1.3 New measure of branch support

Transfer Bootstrap Expectation (TBE)

- Goal: Assessing the extent to which a reference branch b is present in bootstrap trees;
- $\sim$ average % of stable taxa around b over bootstrap trees;
- Continuous measure in $[0, 1]$ based on the transfer distance;



$$\delta(b, b^*) = 2$$

1.4 Questions

- How does TBE behave?
- Does it overcome limitations of FBP?
- Analysis of 3 datasets
  - Mammalian COI-5p protein
  - HIV pol gene
  - Simulated dataset

To answer each question, we implemented a pipeline in Nextflow

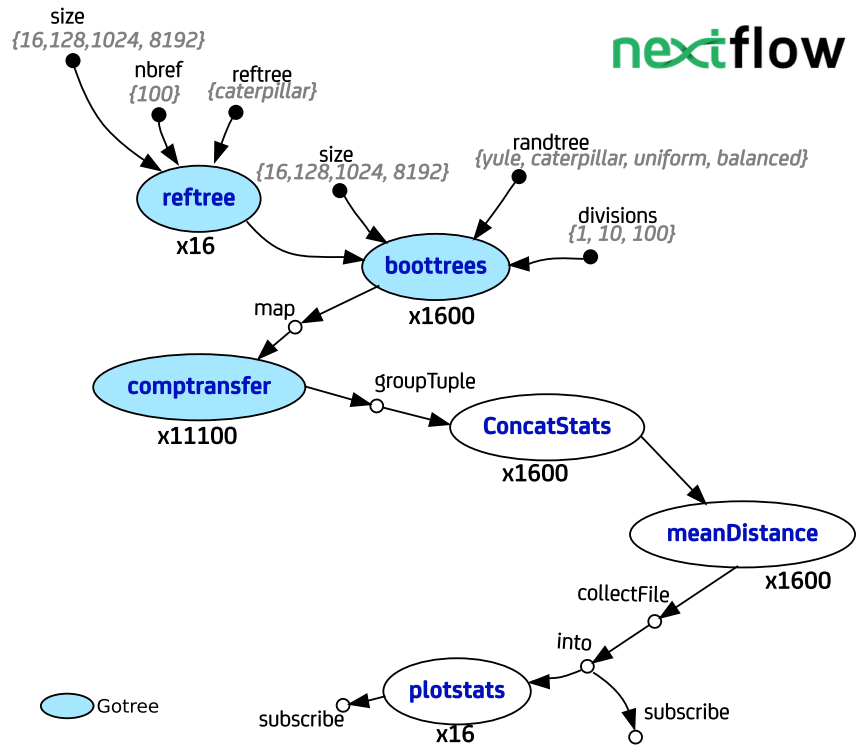
PART 2

Designing workflows for bootstrap analysis

2.1 Theoretical distribution of TBE

Workflow

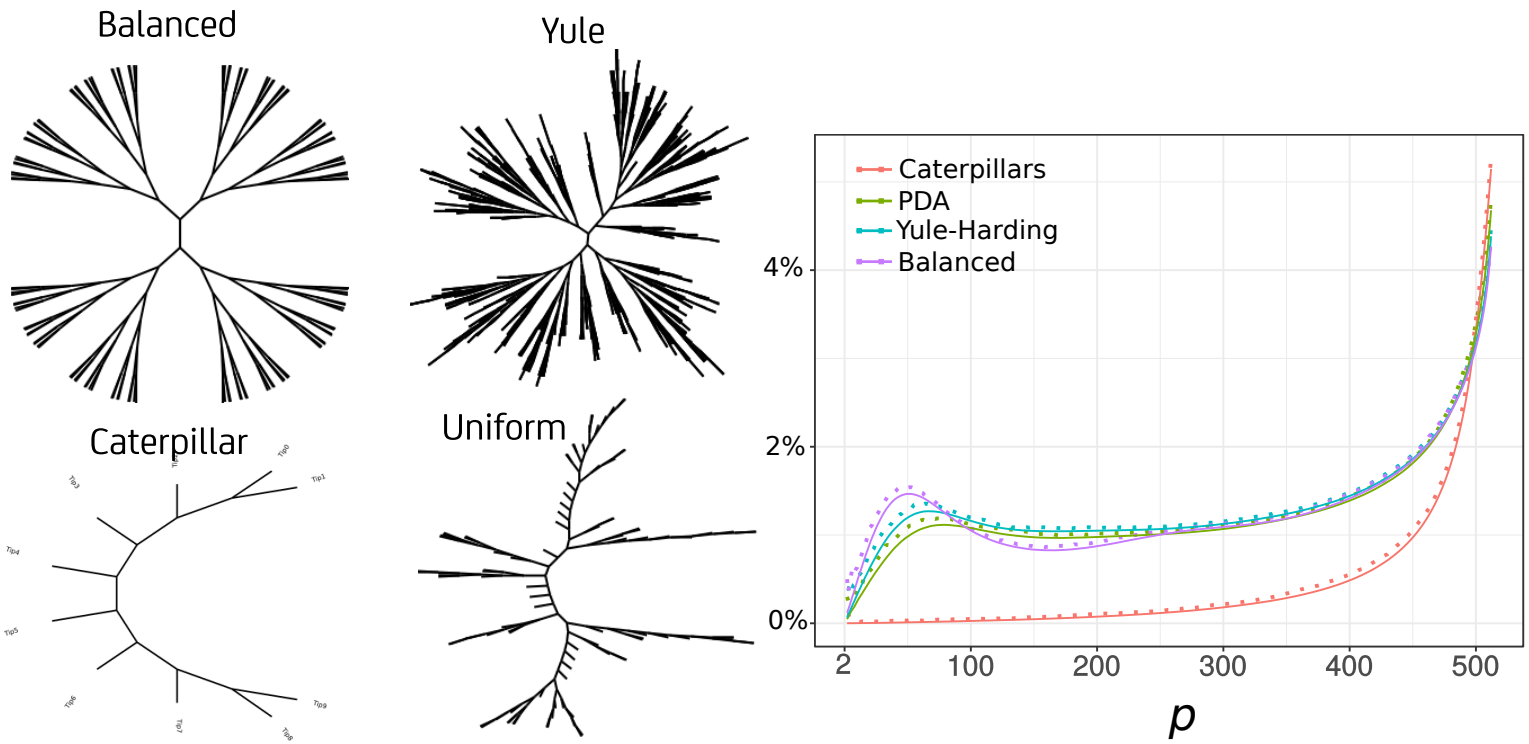
- How does TBE capture phylogenetic signal?
- If we take random trees, what support should we expect?



- Scheduler: Slurm
- Tools : Module
- Jobs: $\sim$ 16000
- Time: $>$ 5 days

2.1 Theoretical distribution of TBE

Results



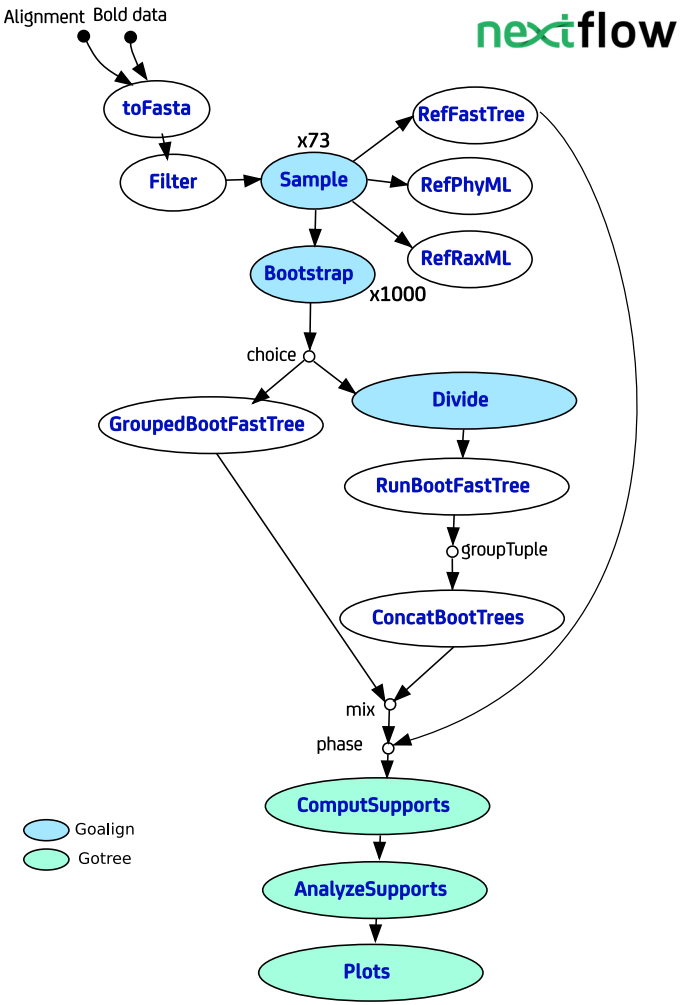
Institut Pasteur

2.2 Mammal data workflow

- How does TBE compare to FBP?
- Does it detect more signal, without false positives?
- General workflow
  1. Build reference and bootstrap trees
  2. Compute TBE and FBP supports
  3. Compare TBE and FBP supports

2.2 Mammal data workflow

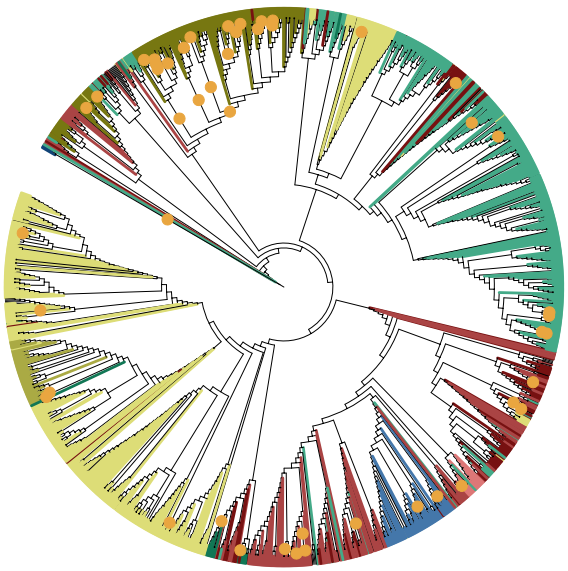
Workflow



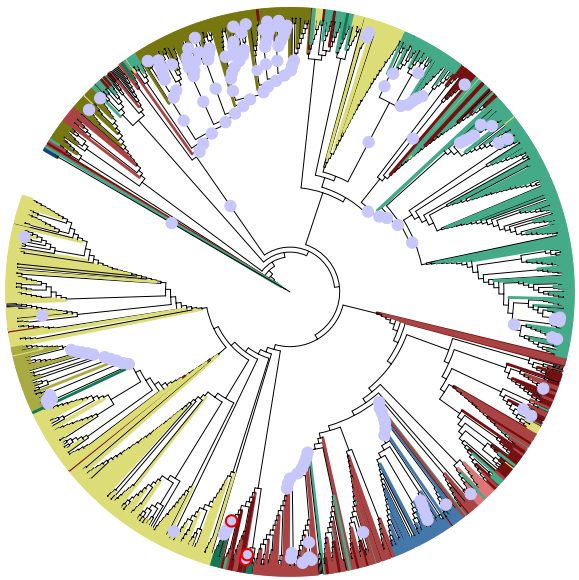
- Scheduler: Slurm
- Tools : Module
- Jobs: $\sim$ 1000
- Time: $\sim$ 1 Day

2.2 Mammal data workflow

Results



FBP



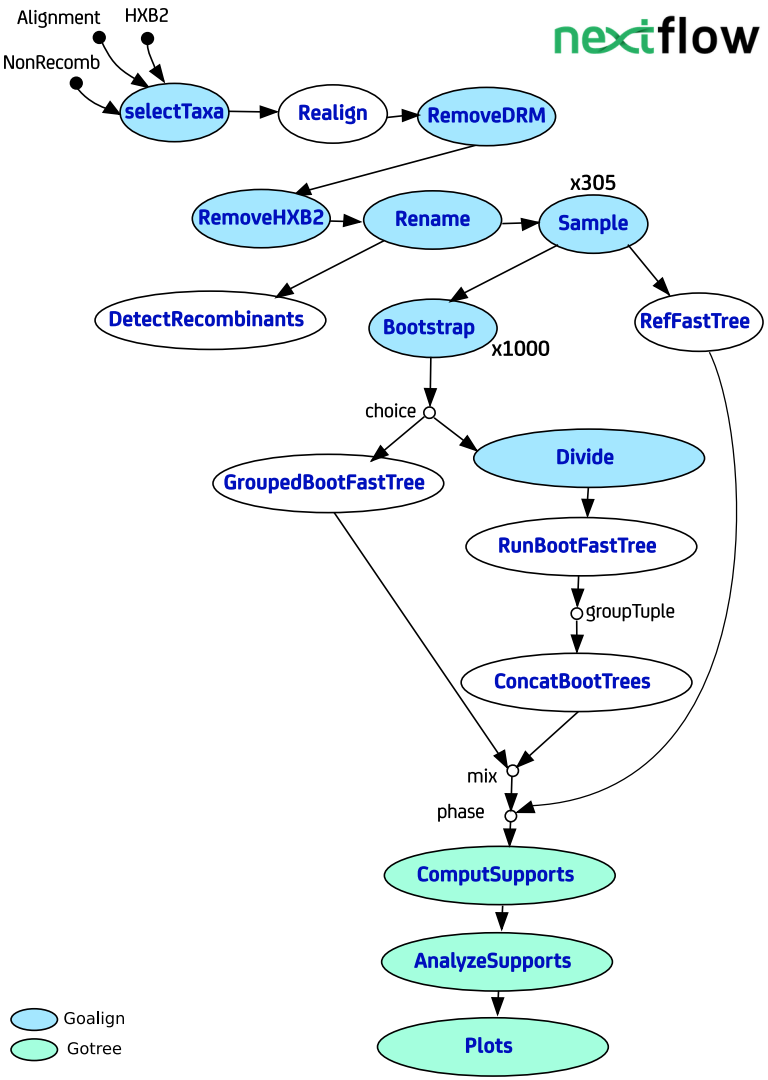
TBE

2.3 HIV data workflow

- How does TBE compare to FBP with HIV large dataset?
- Does it detect more signal, without false positives?
- General workflow
  1. Build reference and bootstrap trees
  2. Compute TBE and FBP supports
  3. Compare TBE and FBP supports

2.3 HIV data workflow

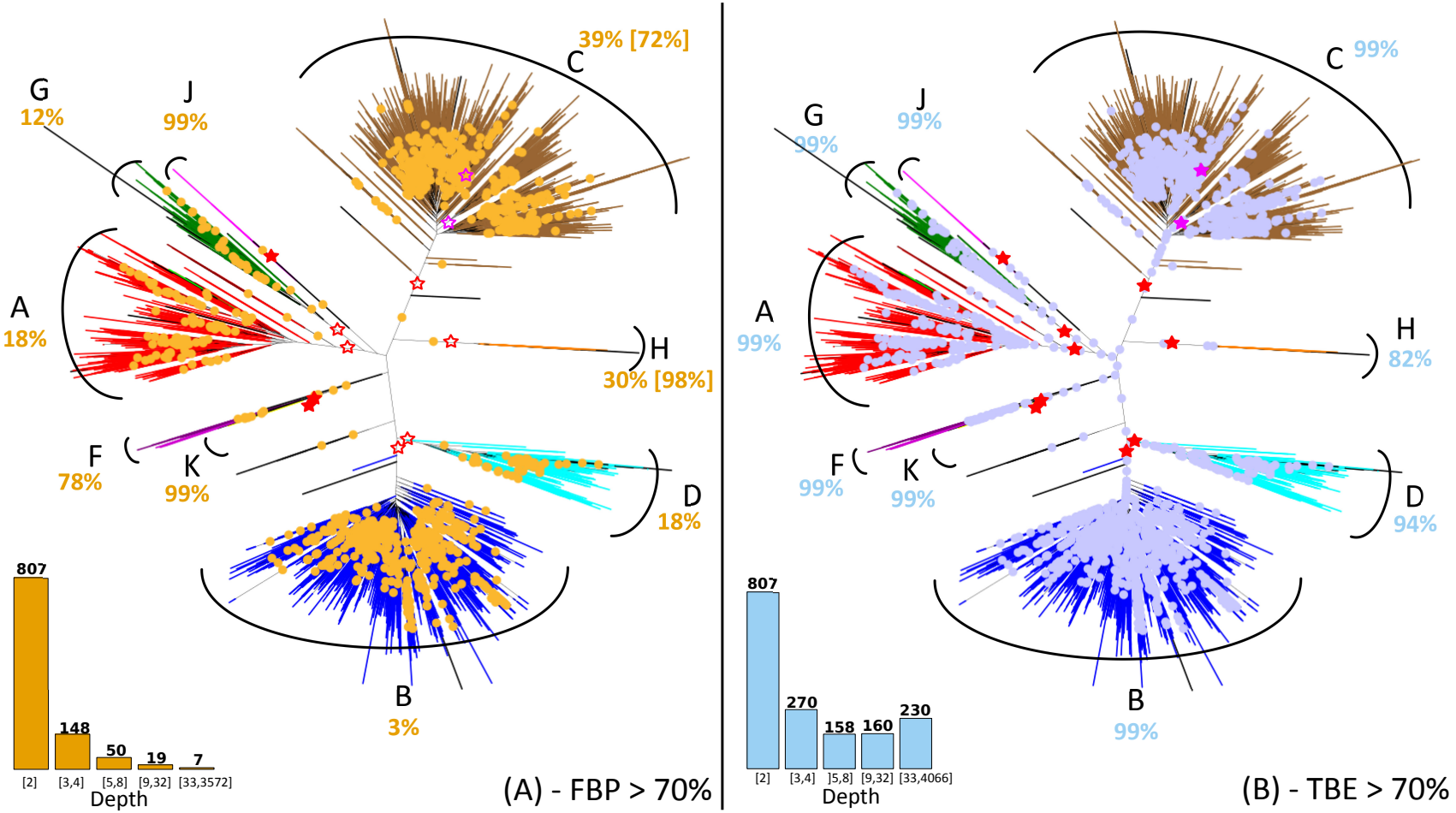
workflow



- Scheduler: Slurm
- Tools : Module
- Jobs: $\sim 50,000$
- Time: $\sim$ 2 days

2.3 HIV data workflow

Results: Supports



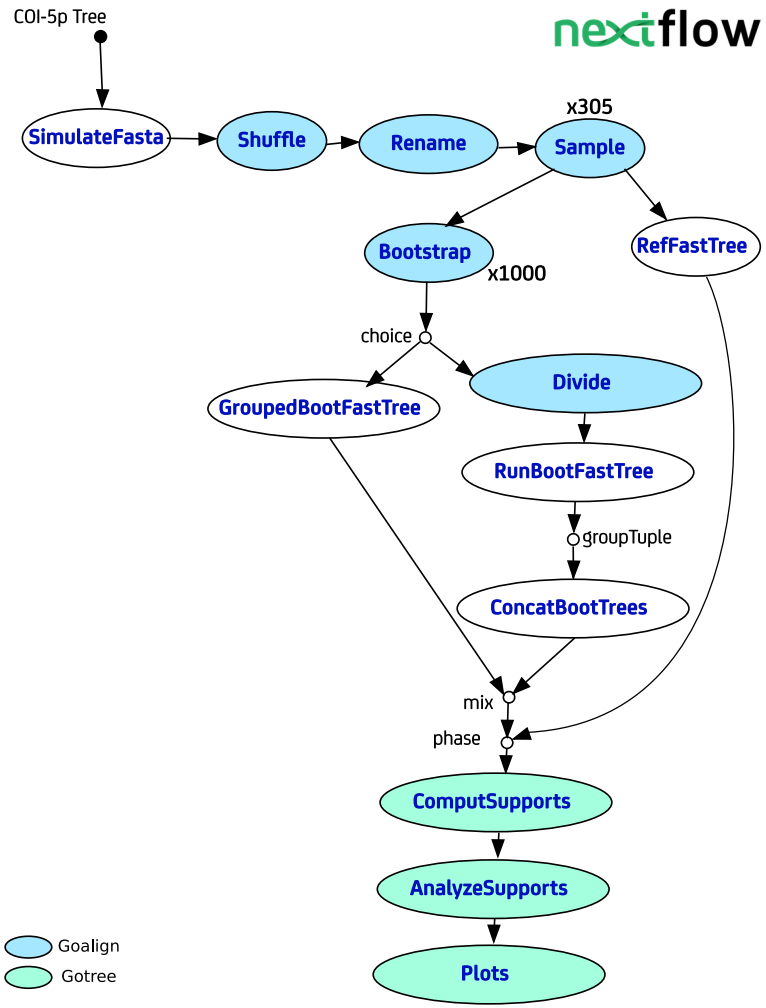
Institut Pasteur

2.4 Simulated data workflow

- How does TBE compare to FBP with simulated data?
- Does it detect more signal, without false positives?
- General workflow
  1. Simulate an alignment from mammal COI-5p tree
  2. Add potential noise
  3. Build reference and bootstrap trees
  4. Compute TBE and FBP supports
  5. Compare TBE and FBP supports

2.4 Simulated data workflow

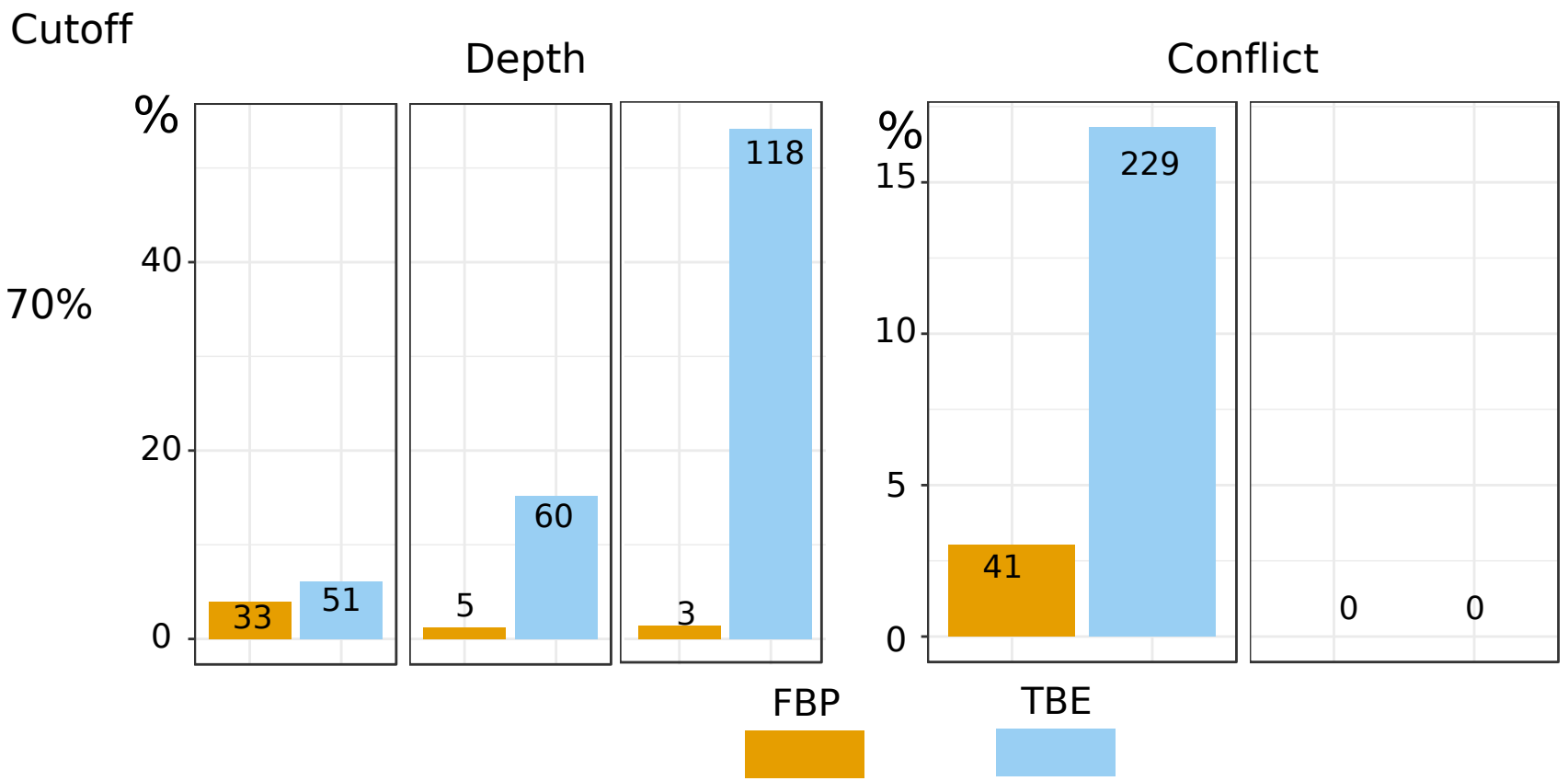
workflow



- Scheduler: Slurm
- Dependency
- Tools : Module
- Jobs: $\sim 1,000$
- Time: $\sim$ 7 hours

2.4 Simulated data workflow

Results: Histograms



2.5 Gotree/Goalign

Phylogenetic command line tools

`github.com/fredericlemoine/{goalign,gotree}`

- We developped Gotree/Goalign toolkit to ease reproducibility of phylogenetic workflows.
- It implements major phylogenetics commands (reformating, rerooting, consensus, bootstrap, etc.)
- It is implemented in Go:
  1. Static binaries $\rightarrow$ Easily distributable/installable without dependencies
  2. Easy parallelization to multiple cores (go routines/channels)
  3. Nice way of giving access to a public API

PART 3

Conclusion

3.2 Nextflow

Nextflow helped us **A LOT** to :

- Keep trace of analyses;
- Cope with several HPC environments (slurm, sge, etc.);
- Relaunch parts of analyses;
- Keep focused on the essential : more on what we do than how to do it.

3.3 Tools

- Preprint : `www.biorxiv.org/content/early/2017/06/23/154542`
- Web interface (Go) : `booster.c3bi.pasteur.fr/` / `github.com/fredericlemoine/booster-web`
- Workflows (Nextflow) : `github.com/evolbioinfo/booster-workflows`
- Booster computation (C) : `github.com/evolbioinfo/booster`
- Goalign/Gotree toolkit (Go) : `github.com/fredericlemoine/{goalign,gotree}`

3.4 Difficulties / Perspectives

- Cluster environment does not yet include a ready to use singularity installation: In progress;
- Several sub-workflows are defined multiple times. Reuse of sub-workflows “may be” useful in some cases (not always);

Thanks for your attention!

Evolutionary Bioinformatics @Pasteur

https://research.pasteur.fr/en/team/evolutionary-bioinformatics/

Institut Pasteur
25-28, rue du Docteur Roux
75724 Paris Cedex 15 - France
www.pasteur.fr