Phylogenetic Code in the Cloud – Can it Meet the Expectations?

Adam Kraut[1], Sébastien Moretti[2,3], Marc Robinson-Rechavi[2],
Heinz Stockinger[3], and Dean Flanders[4]

1) BioTeam Inc., Middleton, MA, USA
2) University of Lausanne, Dept. of Ecology and Evolution, Swiss Institute of Bioinformatics
3) Swiss Institute of Bioinformatics, Vital-IT Group, Lausanne, Switzerland
4) Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

Heinz.Stockinger@isb-sib.ch

HealthGrid 2010, Paris, 28-30 June 2010

BIOTEAM
Enabling Science

SIB
Swiss Institute of
Bioinformatics

UNIL | Université de Lausanne

FMI
Friedrich Miescher Institute
for Biomedical Research

Objectives and outline

- Evaluation of cloud computing for a bioinformatics application in the domain of phylogeny
  - Usability, ease of use, performance, price
- Application benchmark
  - HPC Cluster
  - Amazon’s EC2
- Experimental results
- Lessons learned

Introduction

- Cloud computing has created high expectations
  - Similar like Grid computing about 10 years ago
- In life sciences, several applications are CPU intensive
  - Focus on evolutionary biology problem (phylogeny)
    - Embarrassingly parallel bioinformatics application
- Address question of a typical bioinformatician

  “Where to run a CPU-intensive application?”

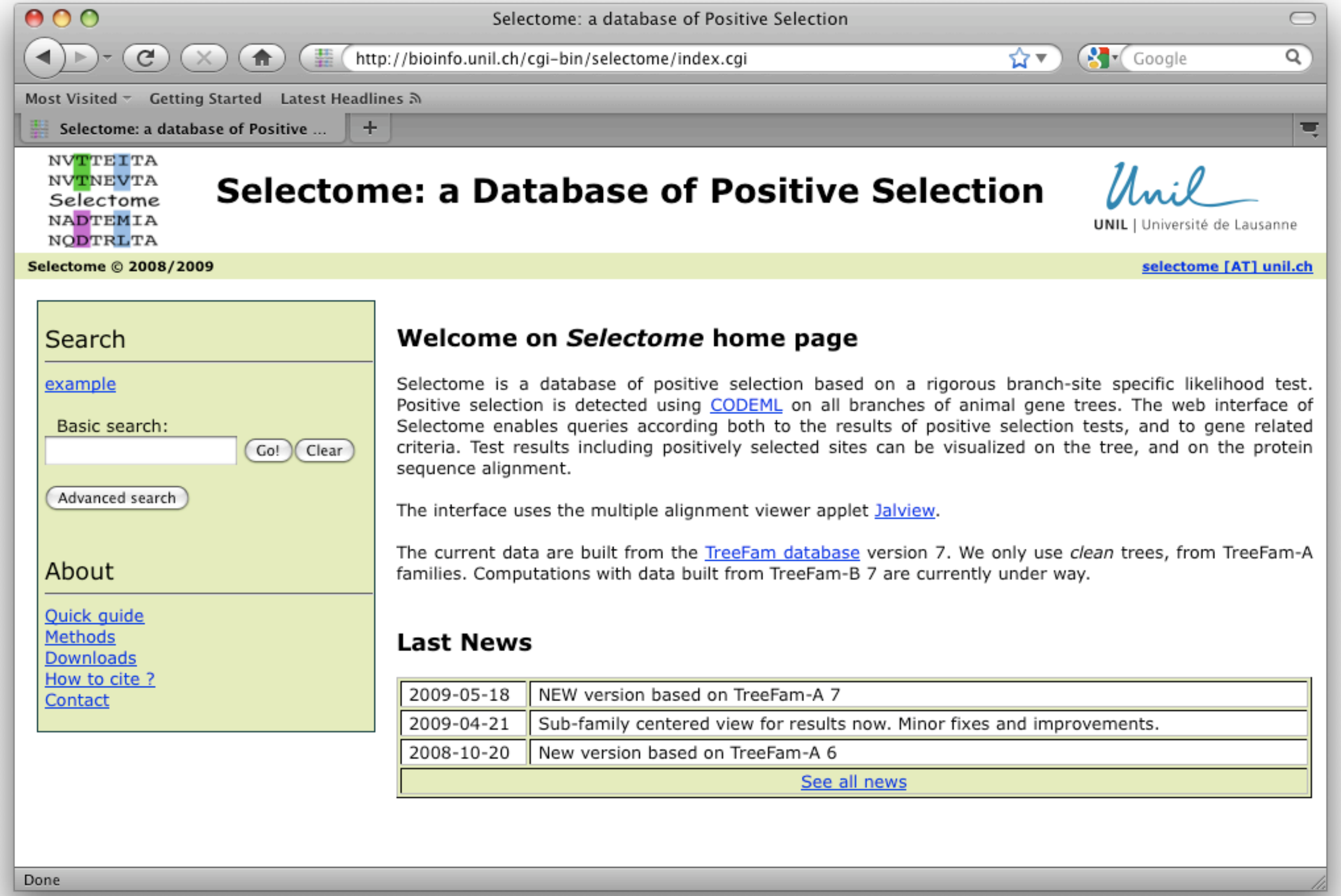
    - Cluster?
    - Cloud?
    - Grid?

Ok, let’s go for it

Main questions related to cloud usage

- How easy is it to port an existing application to Amazon’s EC2?
- Given prior experience with a cluster job submission system, can a similar interface be used?
- What is the performance of the cloud with respect to a compute cluster?
- What is the actual price?

Evolutionary biology: Selectome



Selectome: a database of Positive Selection
http://bioinfo.unil.ch/cgi-bin/selectome/index.cgi
Google
Most Visited
Getting Started
Latest Headlines
Selectome: a database of Positive ...
NVTTEITA
NVTNEVTA
Selectome
NADTEMIA
NODTRLTA
UNIL | Université de Lausanne
Done

Selectome: a Database of Positive Selection

Selectome © 2008/2009

selectome [AT] unil.ch

Search

example

Basic search:

Go! Clear

Advanced search

About

Quick guide
Methods
Downloads
How to cite ?
Contact

Welcome on *Selectome* home page

Selectome is a database of positive selection based on a rigorous branch-site specific likelihood test. Positive selection is detected using CODEML on all branches of animal gene trees. The web interface of Selectome enables queries according both to the results of positive selection tests, and to gene related criteria. Test results including positively selected sites can be visualized on the tree, and on the protein sequence alignment.

The interface uses the multiple alignment viewer applet Jalview.

The current data are built from the TreeFam database version 7. We only use *clean* trees, from TreeFam-A families. Computations with data built from TreeFam-B 7 are currently under way.

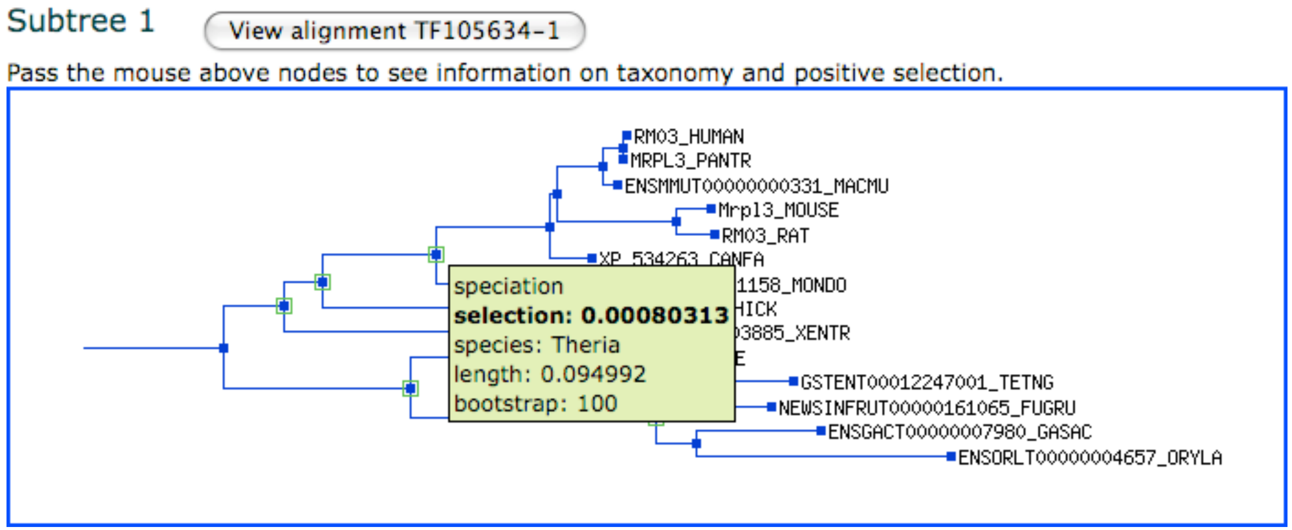
Last News

2009-05-18	NEW version based on TreeFam-A 7
2009-04-21	Sub-family centered view for results now. Minor fixes and improvements.
2008-10-20	New version based on TreeFam-A 6
See all news	

http://selectome.unil.ch

Selectome: database of positive selection

- Selectome uses the PAML[1] package to study the selective pressure
- Codeml (one software tool in PAML) is used with branch-site model to detect positive selection



Subtree 1

View alignment TF105634-1

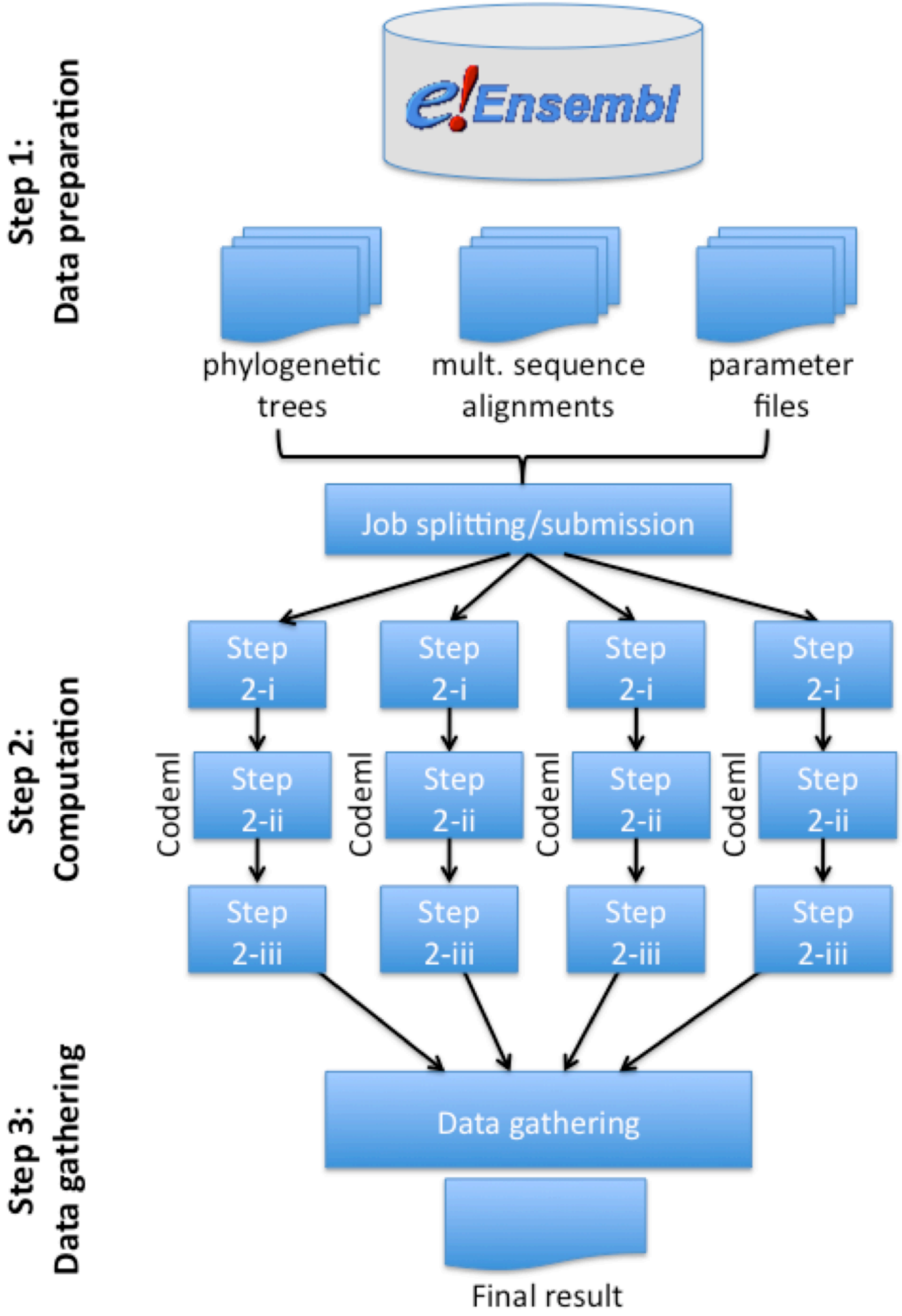
Pass the mouse above nodes to see information on taxonomy and positive selection.

speciation
selection: 0.00080313
species: Theria
length: 0.094992
bootstrap: 100

[1] PAML = Phylogenetic Analysis by Maximum Likelihood

Selectome/Codeml

Embarrassingly parallel computations using Codeml

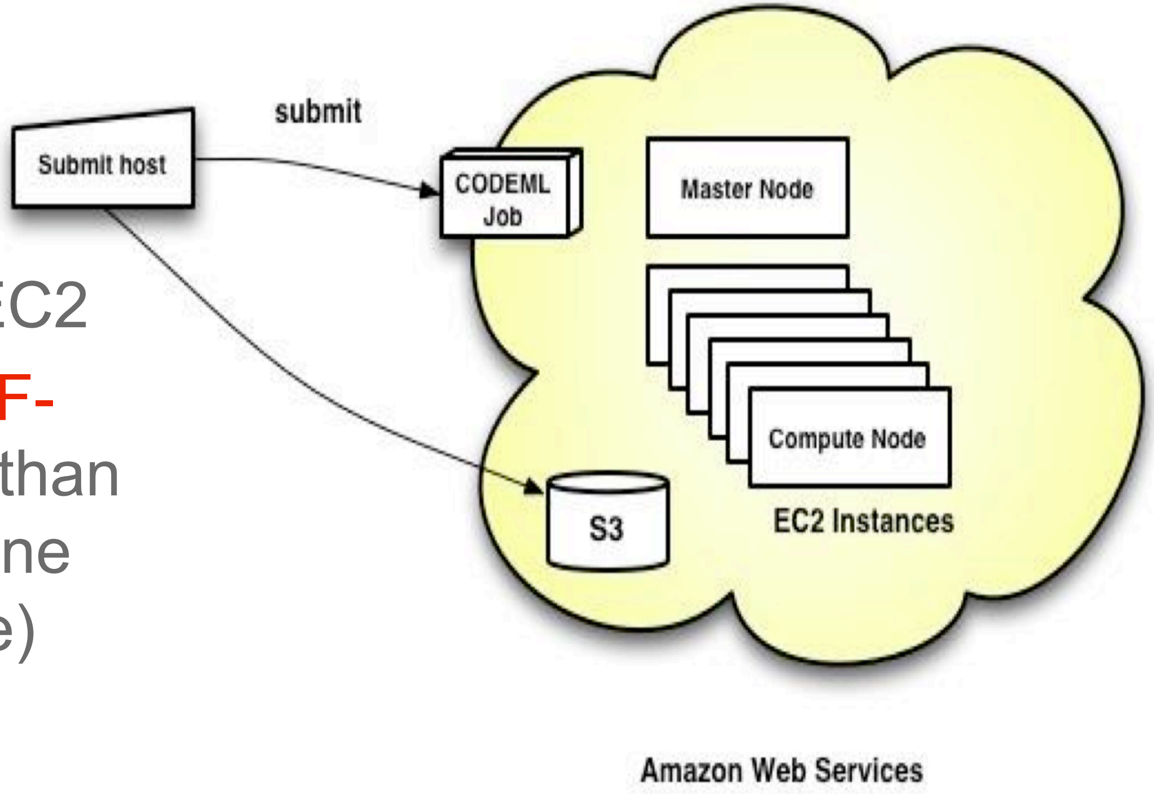


The computational challenge

- For current vertebrate phylogenetic families, a Selectome release needs **1’400’000** Codeml jobs
- A single-threaded Codeml job takes about 20 min. on SIB/Vital-IT[1] cluster
  - Script-based job submission based on LSF is available
- Want to provide a new Selectome release every two months
  - Be “aligned” with Ensembl releases

[1] Vital-IT is the SIB’s HPC centre. It operates an HPC cluster of more than 1000 cores.

Cloud approach: run Codeml on Amazon EC2

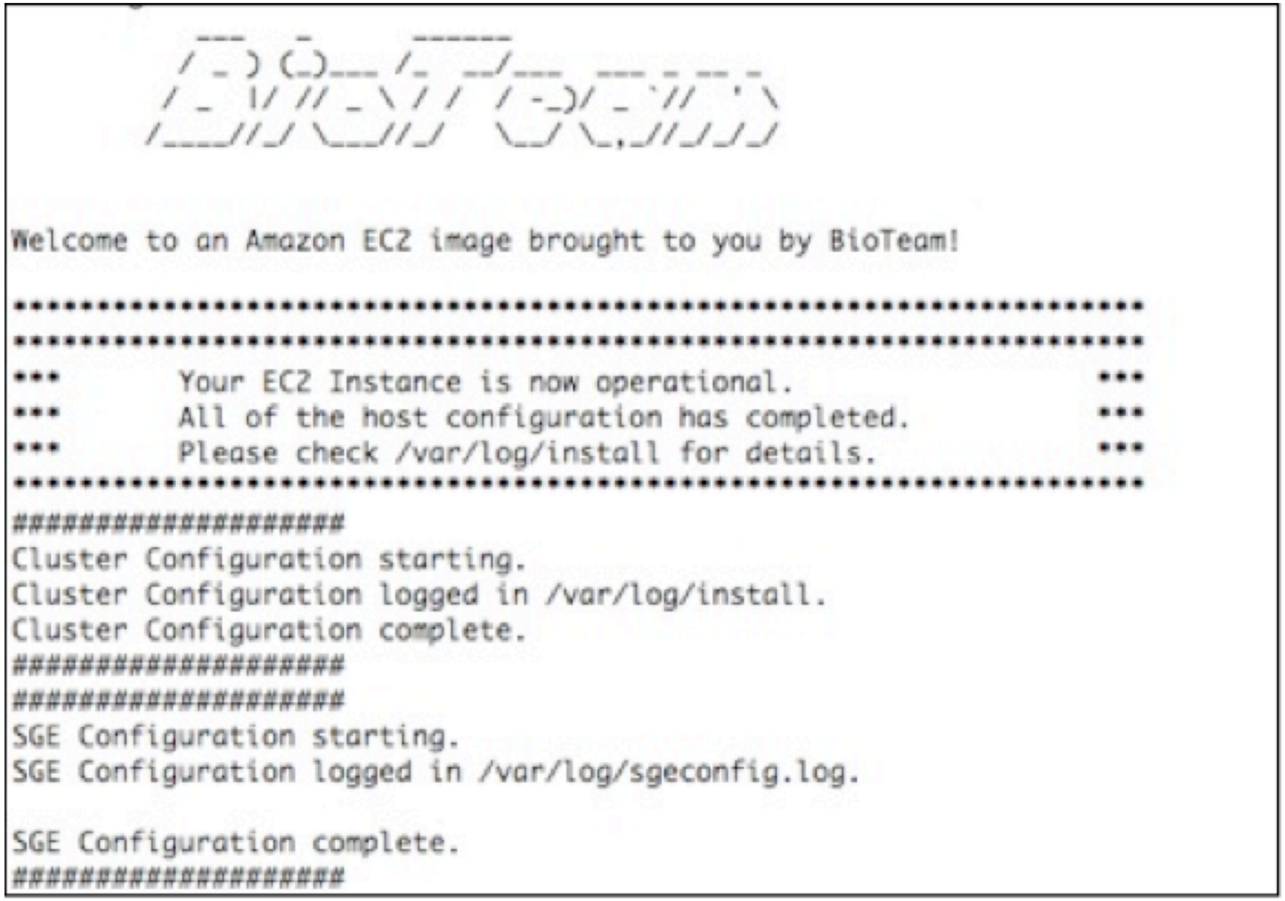


- Port application to EC2
- Want to have an LSF-like interface rather than pure Amazon Machine Images (ease of use)

Shared file system across compute nodes

Sun Grid Engine (SGE) on EC2

- Execute Codeml on EC2 in the same way as on a cluster with LSF
- Users get impression to use a local cluster
- `scp` or `sftp` to upload data to cloud



```
     ___   _      ______
    / _ ) (_)___ /_  __/___  ___ _ __ _
   / _  |/ // _ \/ /  / -_)/ _ `//  ' \
  /____//_/ \___//_/   \__/ \_,_//_/_/_/

Welcome to an Amazon EC2 image brought to you by BioTeam!

*****************************************************************
*****************************************************************
***         Your EC2 Instance is now operational.             ***
***         All of the host configuration has completed.      ***
***         Please check /var/log/install for details.        ***
*****************************************************************
####################
Cluster Configuration starting.
Cluster Configuration logged in /var/log/install.
Cluster Configuration complete.
####################
####################
SGE Configuration starting.
SGE Configuration logged in /var/log/sgeconfig.log.

SGE Configuration complete.
####################
```

Comparison Cloud vs. Cluster

- Cloud performance analysis
  - How long does it take to run a certain experiment in the cloud?
- Performance comparison with local cluster
  - Compare the performance to a production cluster.
- Price
  - How much does it cost to run a full biological experiment?

Experimental results with benchmark data[1]

Cloud: EC2/SGE

- 8 EC2 Instances of type 'c1.xlarge'
  - 64 Intel Xeon E5410
  - 2.33 GHz, 2 GB RAM

Cluster: LSF

- Sub-set of Vital-IT cluster (52 cores)
  - Intel Xeon 5160
  - 3 GHz, 2 GB RAM

52 Codeml jobs	Cloud	Cluster
Shortest job	10 seconds	10 seconds
Longest job	15.48 min	13.88 min
Avg. processing time	8.65 min	6.41 min
Cost	USD 5.44 (8*0.68)	N/A

[1] http://bioinfo.unil.ch/selectome/download/TEST_CASE/codeml_testCase_data.tar.gz
benchmark on single Xeon CPU: ~6 hours

Discussion of results

- BioTeam's EC2/SGE solution is easy to use for users with LSF (or equivalent) experience
- Amazon charges full hours
  - Our experiment only lasted about 15 min
- Price: how much does it cost to use 64 cores for a year?
  - 8-core CPU is USD 1'820; 64 cores: USD 14'560
- Solution is limited by scalability of NFS used between SGE worker nodes
  - Might use Amazon S3 directly instead of NFS
  - Hadoop (MapReduce) as an alternative?

Conclusion

- Successfully ported a bioinformatics application to the cloud
  - However, using AMI is not straight forward for conventional users (hidden by our approach)
  - By default, EC2 does not provide a high-level job submission interface
- Our approach might be interesting for small bioinformatics groups to cover peak performance requirements
- Complete Selectome release on EC2: expensive
  - 344 cores (43 EC2 compute units): USD 42’105.6

- Future work
  - Run Selectome/Codeml in a Grid environment
  - Codeml on Swiss supercomputers
    - Algorithmic improvements
    - Code optimisation
    - Emerging architectures
- Contact

selectome@unil.ch

Heinz.Stockinger@isb-sib.ch

This work was in part funded via the SystemsX.ch initiative in Switzerland.

The cluster computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the Swiss Institute of Bioinformatics

BIOTEAM
Enabling Science

SIB
Swiss Institute of Bioinformatics

Unil
UNIL | Université de Lausanne

FMI
Friedrich Miescher Institute for Biomedical Research