Cyclone UC11

Assembling genomes from sequencing reads

Philippe Veber

November 14th, 2016

Laboratoire de Biométrie et Biologie Évolutive

Context

Univ. Lyon Biology department

- 5 labs involved in microbiology, evolution and ecology
- a bioinformatics platform (PRABI-AMSB)
- a molecular biology platform with sequencing facilities

Context

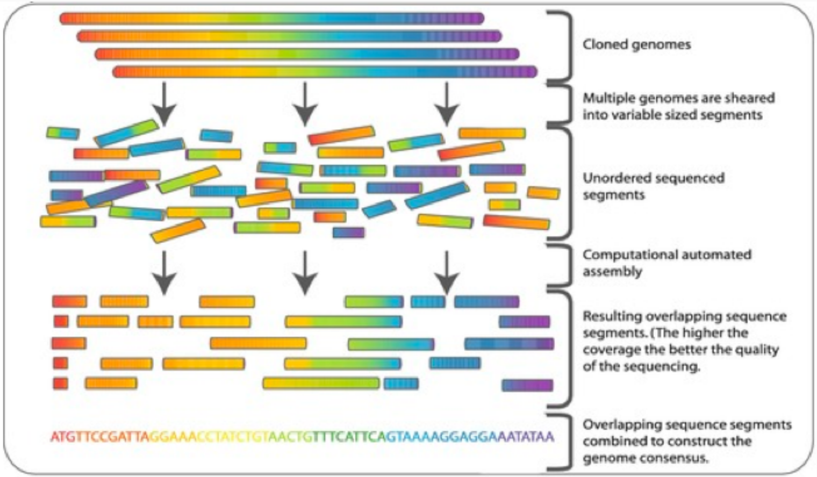
Univ. Lyon Biology department

- 5 labs involved in microbiology, evolution and ecology
- a bioinformatics platform (PRABI-AMSB)
- a molecular biology platform with sequencing facilities

Implementation of a technical partnership between the two platforms on next-generation sequencing

- joint handling of NGS experiments
- from biological samples to bioinformatics analysis

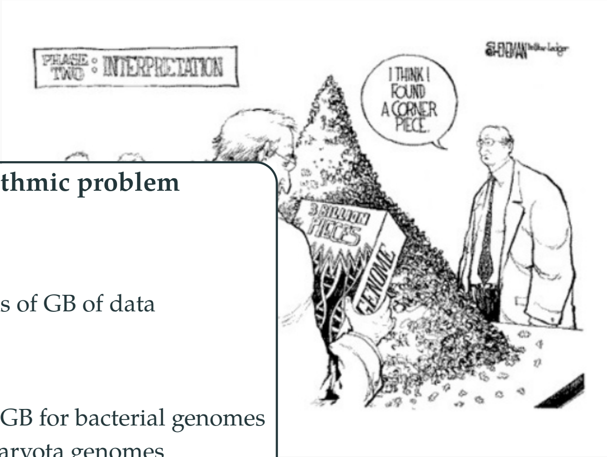
Genome assembly


Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.

from `https://en.wikipedia.org/wiki/Genomic_library`

Computational requirements

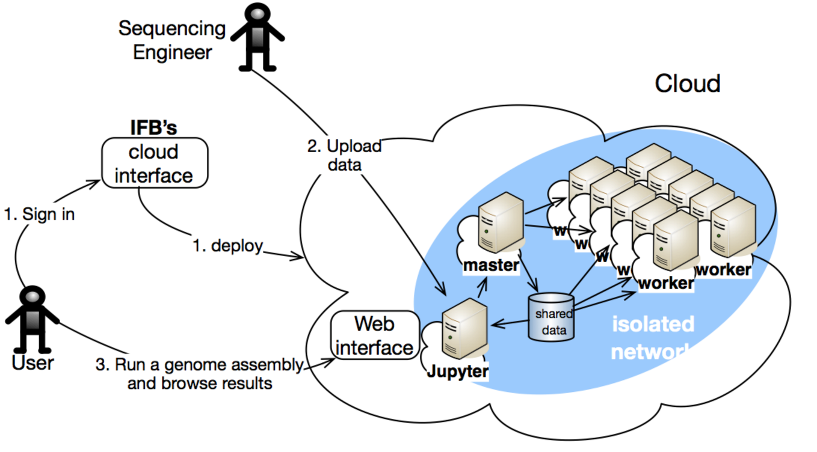
Computational requirements



A tough algorithmic problem

- storage
  - input up to tens of GB of data
  - output ~ 1 GB
- memory
  - from ~ 32 – 64 GB for bacterial genomes
  - to 1 TB for eukaryota genomes
- many cores will help
  - usually shared memory parallelism
- I/O intensive

Expected workflow



Technical issues

- API hooks to deploy a new virtual infrastructure
- and install required configuration/software on it
- authentication for data upload
- storage policy
  - long-run storage?
  - estimate intermediate storage need

Bistro library: describing and deploying complex workflows

An OCaml library to build a workflow as a set of interdependent scripts

Features:

- lightweight encapsulation for scripts/programs
  - with dependency tracking
  - static typing on file formats
- easy composition of large workflows from components
- execution engine
  - distributed
  - resume-on-failure
  - docker friendly
  - HTML execution reports