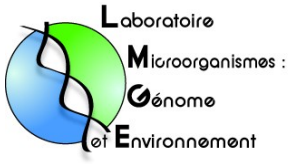
Production et analyse de données au LMGE

Besoins et expériences



CLERMONT UNIVERSITÉ

cnrs

UdA

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

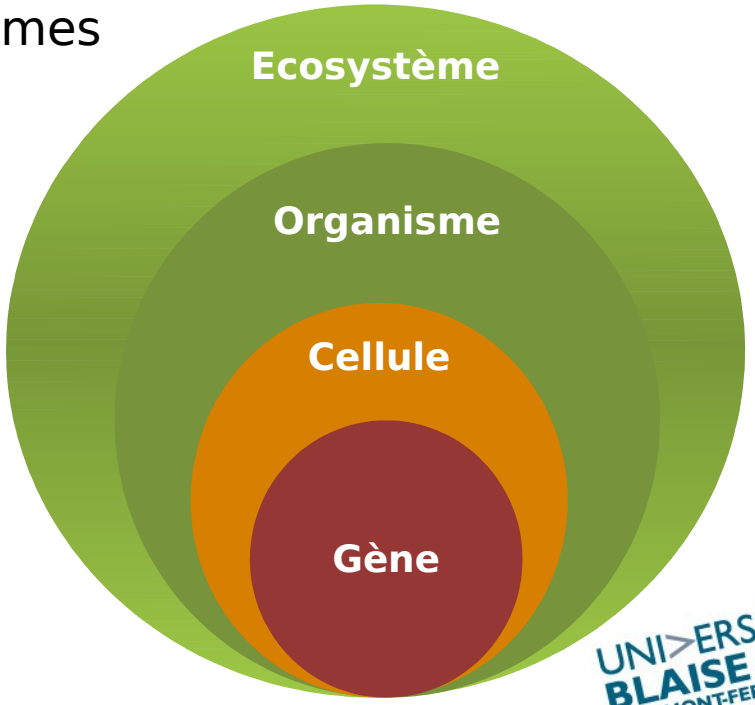
Laboratoire Microorganismes Génome et Environnement

UMR 6023 – CNRS - UBP - UdA

- 120 personnes, 5 équipes de recherche

Objets d'étude : les microorganismes dans leur environnement

- Fonctionnement des écosystèmes
- Biodiversité
- Réseaux trophiques
- Interactions hôtes / parasites
- Ecotoxicologie



- Sol, lac, rivière
- Hôte, parasite
- Milieu hospitalier

LMGE : les données

Production

- Paramètres environnementaux

  Sondes multiparamétriques, cytométrie, chromatographie, etc
- Inventaires taxonomiques

  Microscopie, séquençage
- Post génomique

  NGS, RNASeq, MS-MS, RMN, etc

Format

- Fichiers plats
- Images

Nature

- Alpha-numériques
- Séquences
- Arbres, réseaux

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

LMGE : les données

Production

- Paramètres environnementaux

  Sondes multiparamétriques, cytométrie, chromatographie, etc
- Inventaires taxonomiques

  Microscopie, séquençage
- Post génomique

  NGS, RNASeq, MS-MS, RMN, etc

Observatoire SOERE ?

Format

- Fichiers plats
- Images

Nature

- Alpha-numériques
- Séquences
- Arbres, réseaux

Big data ?

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

LMGE : Exploitation de gros jeux de données

Comparaison de séquences

- Données de référence

NR	17 millions de séquences alignées	20 Go
TREMBL	28 millions de séquences	80 Go

- Données expérimentales

454	1,7 million de séquences	5 Go
Illumina	3 milliards de séquences	300 Go

LMGE : Exploitation de gros jeux de données

Comparaison de séquences

- Données de référence

NR	17 millions de séquences alignées	20 Go
TREMBL	28 millions de séquences	80 Go

- Données expérimentales

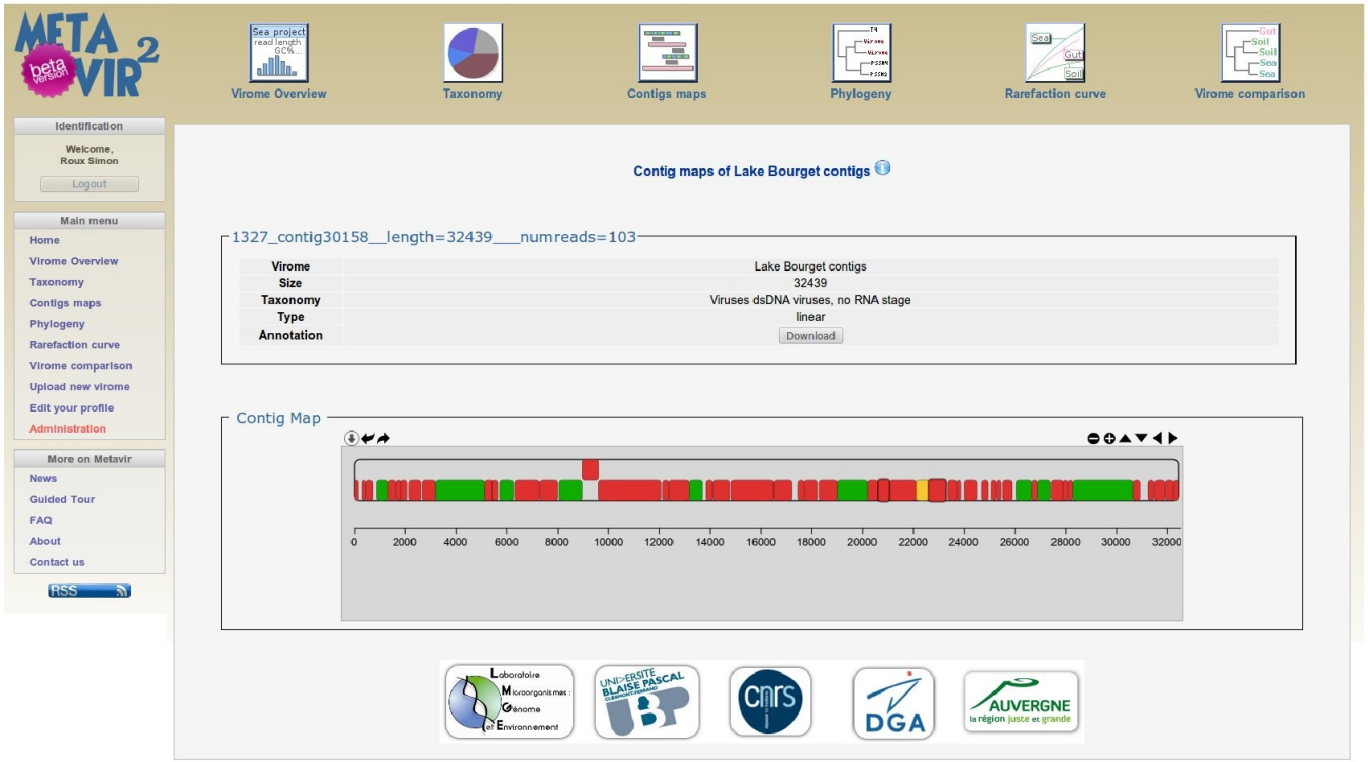
454	1,7 million de séquences	5 Go
Illumina	3 milliards de séquences	300 Go

Expérience

- MetaVir : serveur de métagénomique virale
- PANAM : analyse de données métagénétiques
- Galaxy : plate forme de bioinformatique (Univ Pennsylvanie)

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND UBP

LMGE : Exploitation de gros jeux de données



BIOINFORMATICS **APPLICATIONS NOTE**

Vol. 27 no. 21 2011, pages 3074–3075
doi:10.1093/bioinformatics/btr519

Data and text mining

Advanced Access publication September 11, 2011

Metavir: a web server dedicated to virome analysis

Simon Roux[1,2,*], Michaël Faubladier[1,2], Antoine Mahul[3], Nils Paulhe[1,2], Aurélien Bernard[1,2], Didier Debroas[1,2] and François Enault[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand, [2]CNRS, UMR 6023, LMGE, F-63177 Aubiere and [3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France

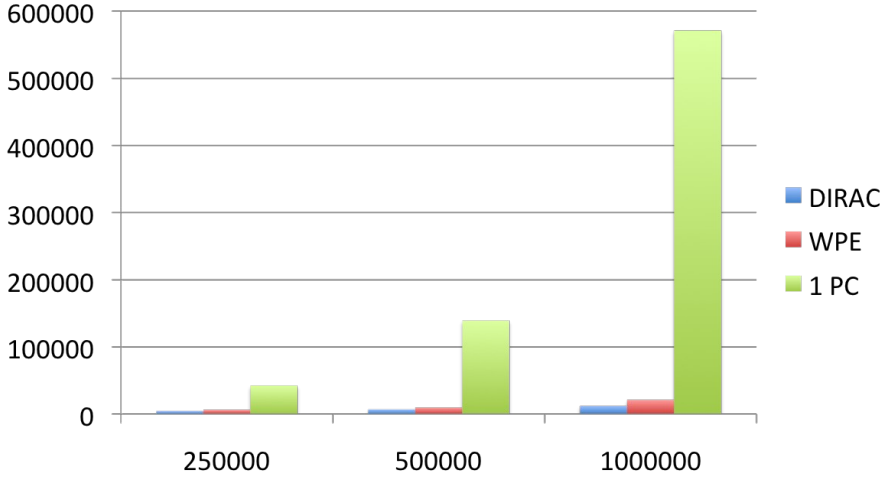
Associate Editor: Alfonso Valencia

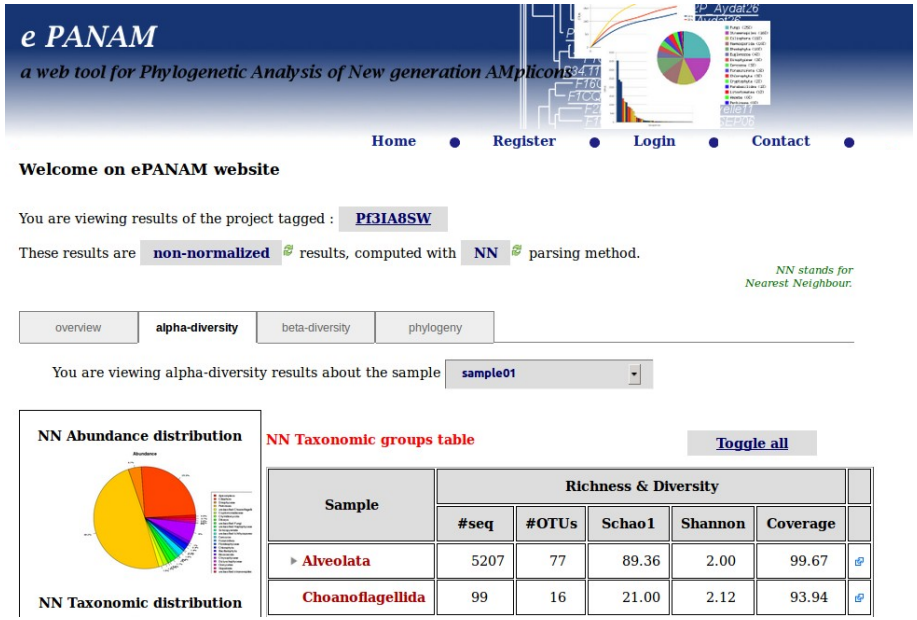
lifegrid

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

LMGE : Exploitation de gros jeux de données

```
Use -gamma for approximate but comparable Gamma(20) log-likelihoods
ML-NNI round 2: LogLk = -81755.250 NNIs 92 max delta 42.55 Time 38.92es (max delta 42.546)
ML-NNI round 3: LogLk = -81176.762 NNIs 63 max delta 70.93 Time 44.12es (max delta 59.209)
ML-NNI round 4: LogLk = -80965.872 NNIs 41 max delta 27.10 Time 47.94es (max delta 27.101)
ML-NNI round 5: LogLk = -80820.670 NNIs 24 max delta 33.02 Time 51.06es (max delta 33.017)
ML-NNI round 6: LogLk = -80737.619 NNIs 12 max delta 18.13 Time 53.32es (max delta 18.127)
ML-NNI round 7: LogLk = -80679.707 NNIs 13 max delta 22.36 Time 54.96
ML-NNI round 8: LogLk = -80494.091 NNIs 10 max delta 51.58 Time 56.34
ML-NNI round 9: LogLk = -80422.407 NNIs 5 max delta 16.75 Time 57.26
ML-NNI round 10: LogLk = -80419.810 NNIs 0 max delta 0.00 Time 58.08
Turning off heuristics for final round of ML NNIs (converged)
ML-NNI round 11: LogLk = -80222.356 NNIs 4 max delta 1.26 Time 65.26 (final) delta 1.256)
Optimize all lengths: LogLk = -80216.857 Time 67.03
Total time: 73.65 seconds Unique: 341/345 Bad splits: 1/338 Worst delta-LogLk 0.459
panam@bioinfo1:~$
```





Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity.

5 **Najwa Taib[1,2], Jean-François Mangot[1,2,3,4], Isabelle Domaizon[3,4], Gisèle Bronner[1,2], and Didier Debroas[1,2,§]**

Accepté dans plosOne

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

LMGE : Exploitation de gros jeux de données

Expérience

- MetaVir : 1 métagénome : 10h de calcul (140 noeuds)
- PANAM (ePANAM) : 1 run 454 : 36h sur un PC
- Galaxy : usage du cluster de l'université de Pennsylvanie

Limitations

- Temps de traitement pour un jeu de données
  - Restriction des offres d'analyse
  - Sous échantillons des jeux de données (50 000 seq)
  - Découpage des données → perte d'intégrité
  - Heuristiques de calcul
- Multiplication des expériences

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND

LMGE : big data ?

Ressources

- Calcul

  PC, cluster (NUMA, CRRI), grille (AUVERGRID)

- Stockage

Comment gérer les données issues des plateformes ?

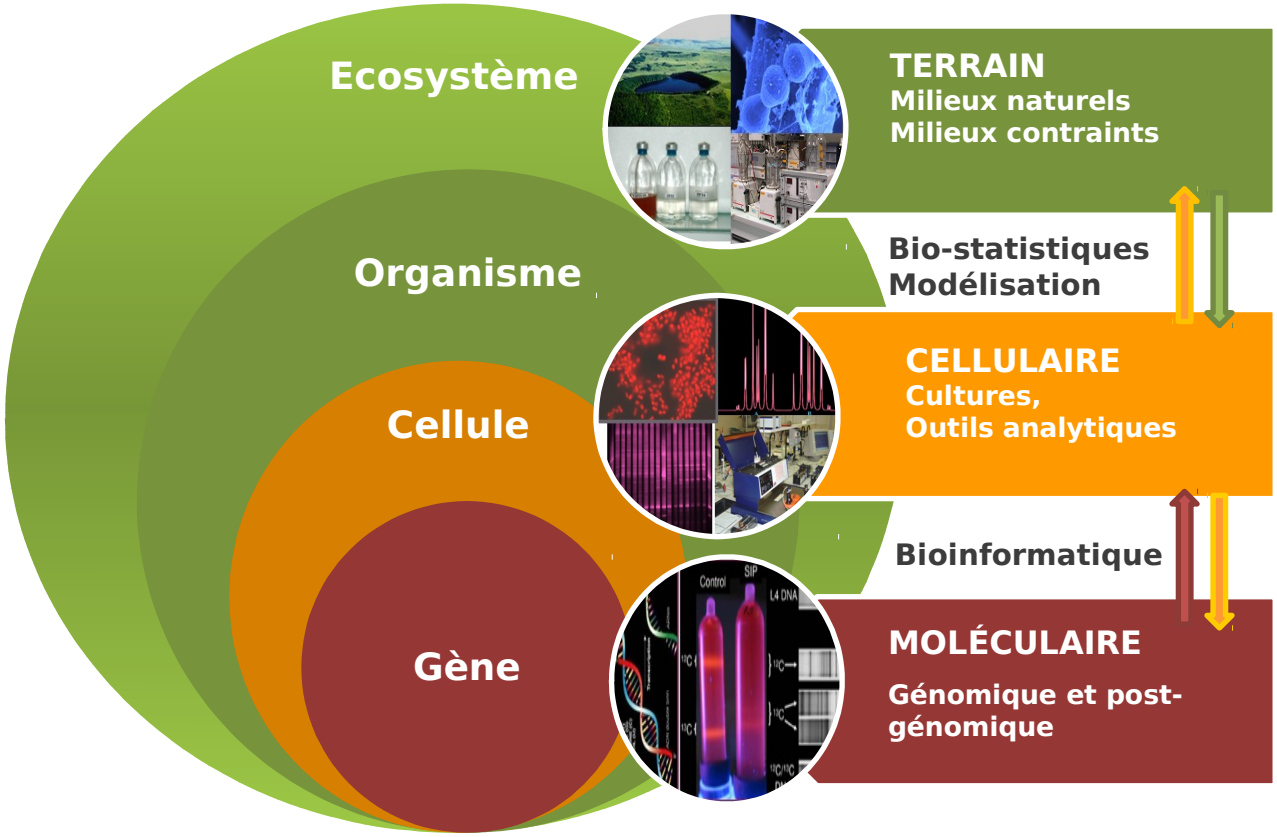
Compétences

- EC Bioinformatique : conception de traitements de données
- IE Informatique : développements HPC (cluster)
- Collaborations PCSV pour développements sur grille


UNIVERSITE
BLAISE PASCAL
CLERMONT-FERRAND
UBP

Laboratoire Microorganismes Génome et Environnement

Etude des microorganismes à différents niveaux d'organisation



Paramètres environnementaux

Inventaires taxonomiques

Post génomique

UNIVERSITE BLAISE PASCAL CLERMONT-FERRAND