Comment le séquençage haut-débit et les algorithmes bioinformatiques contribuent-ils à répondre aux questions de recherche en Sc. de la vie ?

Sourakhata TIRERA
Doctorant IPG/UG
Bioinformaticien

Cayenne, le 12-06-2019

Institut Pasteur de la Guyane · Université de Guyane · ceba · KWATA Etude-Conservation Faune de GUYANE · l'Europe s'engage avec le FEDER · BioGeCo Biodiversité, gènes & communautés

Sommaire :

- Présentation de comment le séquençage haut-débit et les outils bioinformatiques (algorithmes) permettent-ils de répondre à une question biologique ?

    - Succinctement des généralités et « l'univers du NGS »

    - 2 Cas concrets de questions et la façon dont elles ont été répondues

    - 1 Exemple à voir en pratique : partiellement… du moins

    - Discussions pendant et après…

Généralités :

- **La bioinformatique :**

  Domaine où plusieurs compétences se mettent en œuvre

  Variété de « bioinformatiques »

  Appel à une interdisciplinarité

- **Termes :**

  Biologie « sèche » – Biologie « humide »

  Expérience *in silico* – Expérience *in vitro/in vivo*

Déterminant centrale : la donnée !

- Produire :
    - Travail d'<u>échantillonnage</u>
    - Travail de <u>traitement</u> adapté (ex : biologie moléculare)

- Organiser :
    - Gros volumes : Eviter des copies inutiles
    - Sauvegarde de certains fichiers (longs à produire)
    -etc.
- Traiter :
    - Souvent, se fait en « plusieurs fois »… réaffiner, tester etc.
    - Nécessite au moins : une compréhension de la question bio et les capacités à « conduire » les analyses bioinfo
- Interpréter :
    - Peut faire intervenir les biologistes
    - Nouvelles directions pour les analyses (retour au traitement)

- Publier :
    - Mettre les (nouvelles) connaissances au service du (d'un) public
    - Mettre les données au service de la science !

Ecosystème bioinformatique et cycles de vie

- Serveurs et clusters de calculs : Mésocentres, Instituts de recherche etc.

- Langages de programmation :
    - Gratuits et vivier d'outils d'analyses et de librairies

- Données produites par avancées techniques (séquençage) :
    - Séquençage Sanger
    - 1ere Génération NGS
    - Lecture de Molécule unique

- Design de nouveaux algorithmes :
    - Suite aux nouveaux problèmes et défis

- Données publiées : (Génomes, algorithmes, méthodes, etc.)

- Nouvelles connaissances !

Exemple 1 : On ne comprend plus rien aux souches de plasmodium !

Constat :

Souches de Plasmodium résistantes (re-) devenus sensibles à la Chloroquine.

Pourtant, le génotype « résistant » est resté fixé.

Comment s'explique ce phénotype « discordant » ?

PNAS

Adaptive evolution of malaria parasites in French Guiana: Reversal of chloroquine resistance by acquisition of a mutation in *pfcrt*

Stéphane Pelleau[a], Eli L. Moss[b], Satish K. Dhingra[c], Béatrice Volney[a], Jessica Casteras[a], Stanislaw J. Gabryszewski[c], Sarah K. Volkman[d,e,f], Dyann F. Wirth[d,e], Eric Legrand[a], David A. Fidock[c,g,1], Daniel E. Neafsey[b,1], and Lise Musset[a,2]

[a]Laboratoire de parasitologie, WHO Collaborating Center for Surveillance of Anti-Malarial Drug Resistance, Centre National de Référence du paludisme, Laboratoire associé pour la région Antilles-Guyane, Institut Pasteur de la Guyane, 97300 Cayenne, French Guiana; [b]Genome Sequencing and Analysis Program, Broad Institute, Cambridge, MA 02142; [c]Department of Microbiology and Immunology, Columbia University College of Physicians and Surgeons, New York, NY 10032; [d]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115; [e]Infectious Disease Program, Broad Institute, Cambridge, MA 02142; [f]School of Nursing and Health Sciences, Simmons College, Boston, MA 02115; and [g]Division of Infectious Diseases, Department of Medicine, Columbia University College of Physicians and Surgeons, New York, NY 10032

Significance

This study addresses the evolutionary dynamics of antimalarial drug resistance after changes in drug use. We show that chloroquine resistance in *Plasmodium falciparum* from French Guiana was lost after sustained drug removal, whereas the resistance marker PfCRT K76T remained fixed in the parasite population. This phenotypic reversion was caused by the acquisition of a single additional C350R substitution in PfCRT. This genetic change also impaired susceptibility to piperaquine, suggesting that piperaquine pressure drove the expansion of this allele. These findings have important implications for understanding drug resistance evolution when standard resistance alleles reach fixation and can lose their utility as markers because of adaptive changes at other amino acid positions or loci in the genome.

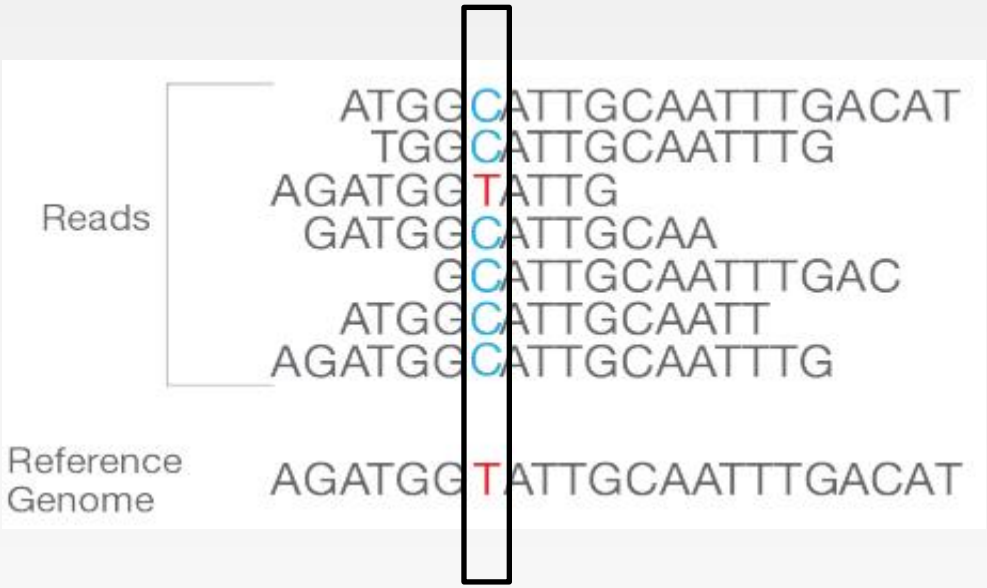
Exemple 1 : On ne comprend plus rien aux souches de plasmodium !

Etude **GWAS** : **G**enome **W**ide **A**ssociation **S**tudies :

- ✓ Séquençage de génomes complets de *P. falciparum*
- ✓ **Identification des mutations**
  - ✓ Algorithmes de mapping :

    millions de reads/génome sur une référence en quelques heures

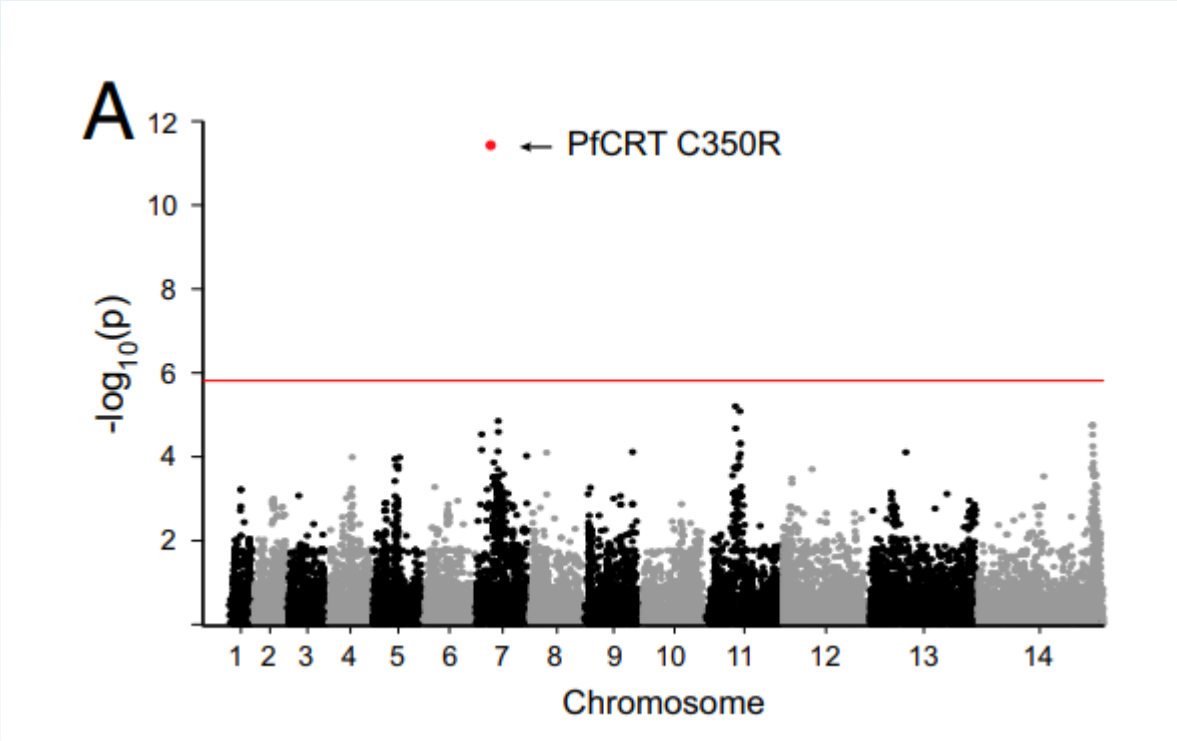
    Inspirés des algorithmes de compression



Exemple 1 : On ne comprend plus rien aux souches de plasmodium !

Etude **GWAS** : **G**enome **W**ide **A**ssociation **S**tudies :

- ✓ Séquençage de génomes complets de *P. falciparum*
- ✓ **Identification des mutations**
- ✓ Statistiques d'association (génotype/phénotype)



<u>Manhattant plot</u> : points de mutation et probabilité qu'ils expliquent le phénotype

Exemple 1 : On ne comprend plus rien aux souches de plasmodium !

PNAS

Adaptive evolution of malaria parasites in French Guiana: Reversal of chloroquine resistance by acquisition of a mutation in *pfcrt*

Stéphane Pelleau[a], Eli L. Moss[b], Satish K. Dhingra[c], Béatrice Volney[a], Jessica Casteras[a], Stanislaw J. Gabryszewski[c], Sarah K. Volkman[d,e,f], Dyann F. Wirth[d,e], Eric Legrand[a], David A. Fidock[c,g,1], Daniel E. Neafsey[b,1], and Lise Musset[a,2]

[a]Laboratoire de parasitologie, WHO Collaborating Center for Surveillance of Anti-Malarial Drug Resistance, Centre National de Référence du paludisme, Laboratoire associé pour la région Antilles-Guyane, Institut Pasteur de la Guyane, 97300 Cayenne, French Guiana; [b]Genome Sequencing and Analysis Program, Broad Institute, Cambridge, MA 02142; [c]Department of Microbiology and Immunology, Columbia University College of Physicians and Surgeons, New York, NY 10032; [d]Department of Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115; [e]Infectious Disease Program, Broad Institute, Cambridge, MA 02142; [f]School of Nursing and Health Sciences, Simmons College, Boston, MA 02115; and [g]Division of Infectious Diseases, Department of Medicine, Columbia University College of Physicians and Surgeons, New York, NY 10032

Une fois la pression médicamenteuse « chloroquine » retirée les souches de *P. falciparum* ont acquis une nouvelle mutation les rendant sensible.

Hypothèse : Apparition puis expansion de l'allèle du à la soumission à la « piperaquine ».

Validation Biologique faite !

Significance

This study addresses the evolutionary dynamics of antimalarial drug resistance after changes in drug use. We show that chloroquine resistance in *Plasmodium falciparum* from French Guiana was lost after sustained drug removal, whereas the resistance marker PfCRT K76T remained fixed in the parasite population. This phenotypic reversion was caused by the acquisition of a single additional C350R substitution in PfCRT. This genetic change also impaired susceptibility to piperaquine, suggesting that piperaquine pressure drove the expansion of this allele. These findings have important implications for understanding drug resistance evolution when standard resistance alleles reach fixation and can lose their utility as markers because of adaptive changes at other amino acid positions or loci in the genome.

Exemple 2 : Leishmania RNA Virus 1, Projet GENEVIL

PLOS | NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

Unraveling the genetic diversity and phylogeny of *Leishmania* RNA virus 1 strains of infected *Leishmania* isolates circulating in French Guiana

Sourakhata Tirera[1☯], Marine Ginouves[2,3☯], Damien Donato[1☯], Ignacio S. Caballero[1,4], Christiane Bouchier[5], Anne Lavergne[1], Eliane Bourreau[6], Emilie Mosnier[2,7,8], Vincent Vantilcke[9], Pierre Couppié[2,3,10], Ghislaine Prevot[2‡], Vincent Lacoste[1,11‡*]

Leishmania guyanensis
Leishmania braziliensis

Quand le virus infecte les leishmanies ….

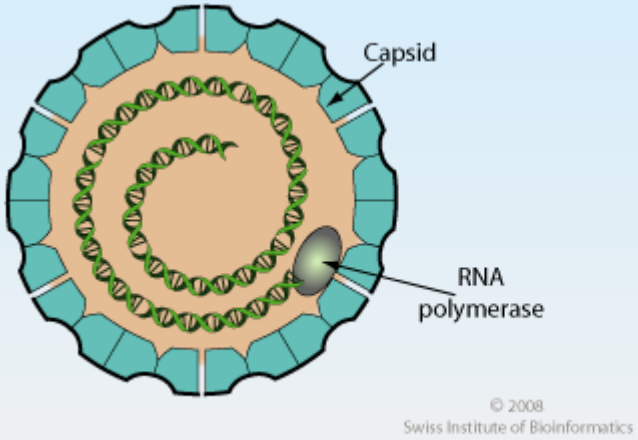
- Echecs thérapeutique
- Leishmaniose plus sévère

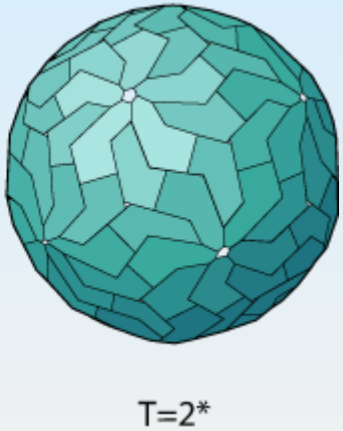
Manque d'information sur ce virus en Guyane…

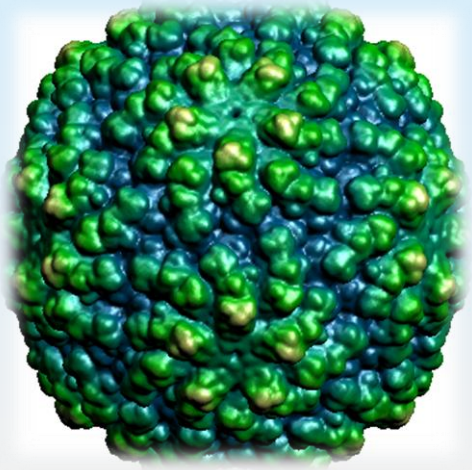
But GENEVIL

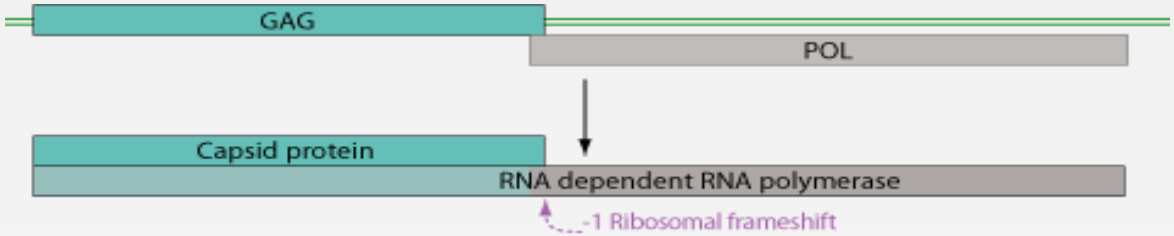
- Caractérisation génomique de LRV1
- Patients infectés en Guyane
  - *L. guyanensis*
  - *L. braziliensis*
- Très peu de génomes disponibles (en 2016)
- Possibilité de distance

Leishmania RNA Virus 1 : Génome









Virus à ARN double brin
Famille des Totiviridae

Génome 5,3 kb

Chaîne de traitement *in vitro*

1. Prélèvement humain
2. Identification des Leishmanies
3. Culture des leishmanies
4. Extraction des acides nucléiques
- Séquençage haut-débit illumina®

```
@D00395:508:HN5F2BCX2:1:1101:1626:98193/1
GCCCCAACACATGGCTCACCACACCATGGTTGGAACGGAAAAGACTCCCAGGAGGAAAAGCCCCCTGGGAAG
+
GGGGGIIIIIIIIIGGIGIIIIIIIIIGGIIIIIIIGIIIIIIIIIIIIIIIIIGGIIIGIIIIIIIIII
```

1 Read illumina au format fastq

Chaîne de traitement *in silico*

1. Contrôle Qualité
2. Assemblage
3. Recherche d'homologie
4. Mapping - Validation (contigs)

Unipro UGENE

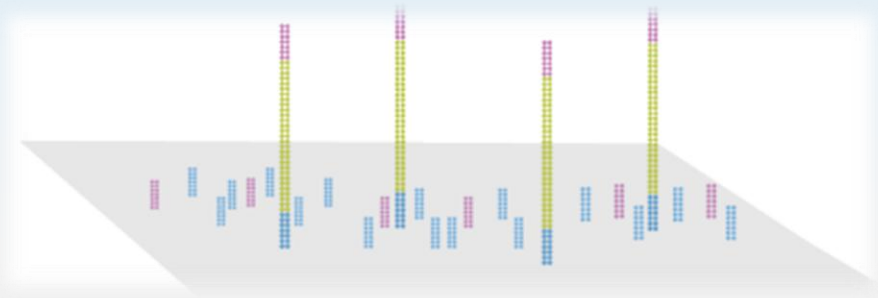
Galaxy

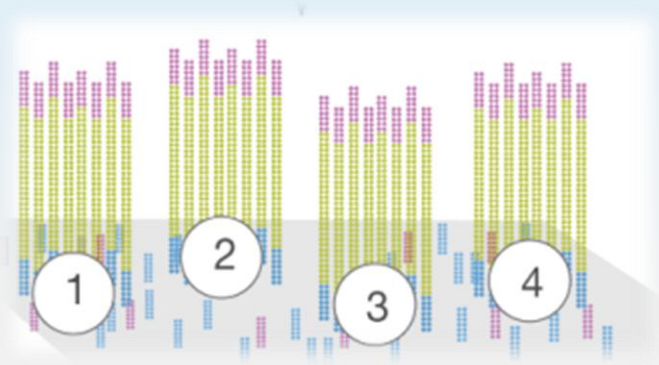
GNU/Linux

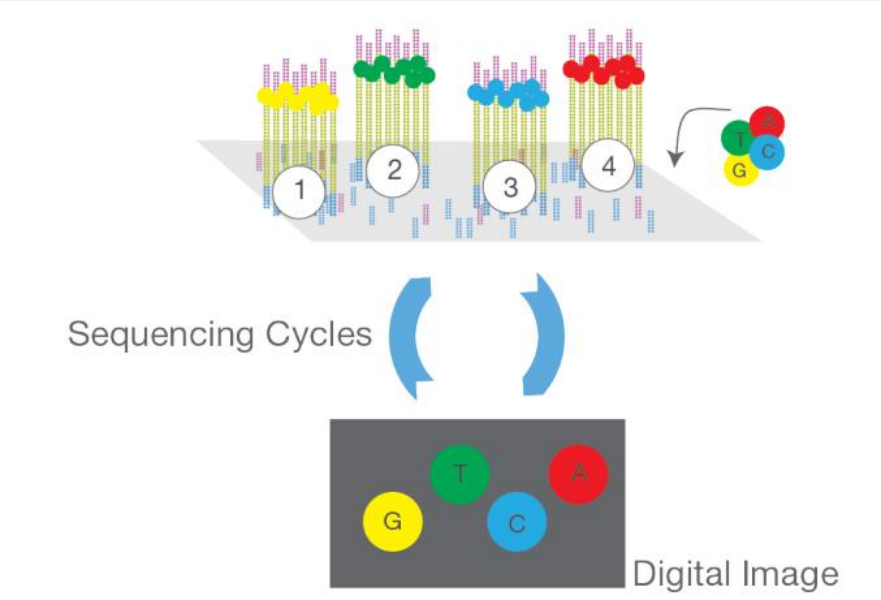
Séquençage haut-débit : Principe

Besoin d'amplifier un fragment connu : Séquençage classique (Sanger)

Amplification de fragments sans à priori : NGS







Séquençage haut-débit : illumina

https://www.youtube.com/watch?v=fCd6B5HRaZ8

- Fragmentation des acides nucléiques
- Sélection de la taille de fragments
- Ligation d'adaptateurs
- Dépôt sur plaque (Flowcell) (contient le complément des adaptateurs)
- Amplification
- Lecture par synthèse (Sequencing by synthesis) à l'aide de fluorochromes
- Génération des fichiers fastq sur disque dur

Entête - identifiant

Séquence d'ADN

```
@D00395:508:HN5F2BCX2:1:1101:1626:98193/1
GCCCCAACACATGGCTCACCACACCATGGTTGGAACGGAAAAGACTCCCAGGAGGAAAAGCCCCCTGGGAAG
+
GGGGGIIIIIIIIIGGIGIIIIIIIIIGGIIIIIIIGIIIIIIIIIIIIIIIIIIIGGIIIGIIIIIIIIII
```

Qualité

1 Read illumina au format fastq

1

Contrôle Qualité

Séquences adaptateurs	⟹	N'appartiennent pas à l'échantillon
Séquences dupliquées	⟹	Inutiles (ne rajoutent pas d'information)
Bases de très faible qualité	⟹	Eviter des variants faux positifs
Motifs surreprésentés	⟹	Peuvent être signe d'anomalie (régions répétées, biais d'amplification)

Etape permettant de corriger les données avant tout Traitement

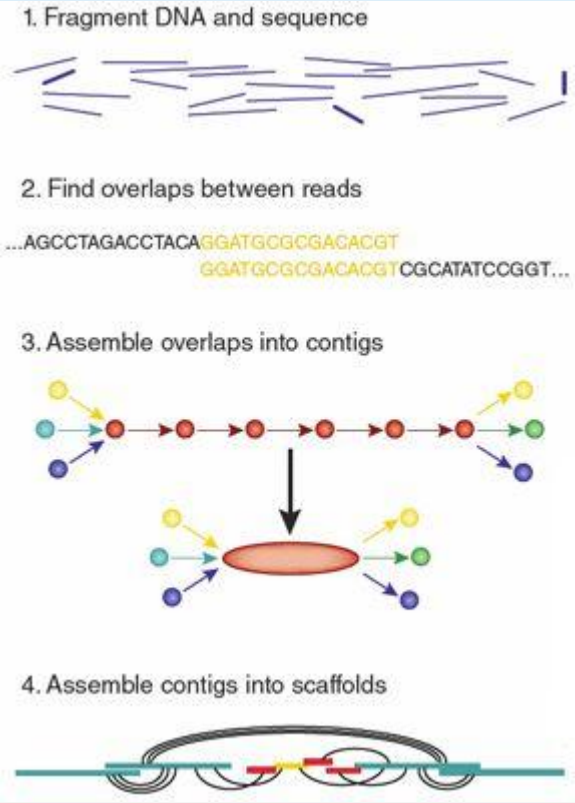
2

Assemblage denovo

Problème : **Reconstruire le puzzle !**

En entrée :

- Séquences courtes générées après fragmentation aléatoire
- Elles ont donc des parties chevauchantes



Solution :

Reconstruire les génomes/gènes/chromosomes
à partir des chevauchements !

Approche 1 : OLC (Overlap Layout Consensus)
Approche 2 : Graphe de de Bruijn

Reads ⟹ **Contigs**
Fastq **Fasta**

2

Assemblage denovo

Graphe de Debruijn

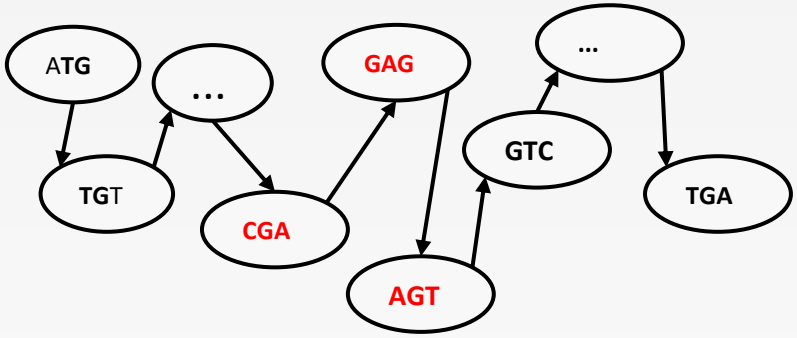
Graphe = structure de données adaptée aux chevauchements

Read A : ATGTCGTCAATCGAGT (l=16)
Read B : CGAGTCTATTCGATGA (l=16)

K=3 -> k-1 identiques :

ATG	TGT	GTC	TCA	CAA	ATC	TCG	CGA	GAG	AGT

CGA	GAG	AGT	GTC	TCT	CTA	TAT	...	ATG	TGA



ATGTCGTCAATCGAGTCTATTCGATGA

Contig_AB (l=27)

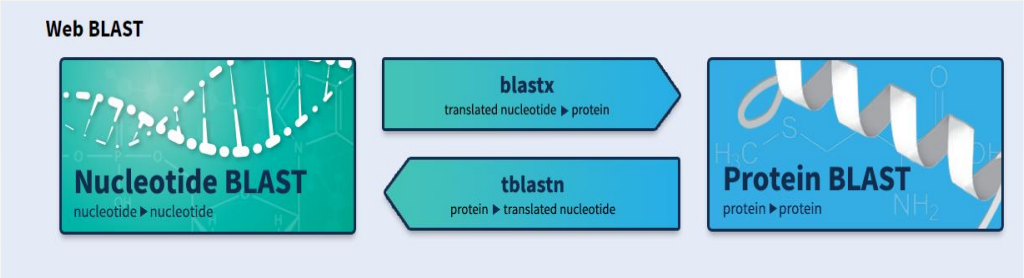
3

Recherche d'homologies

Identifier une séquence inconnue au départ

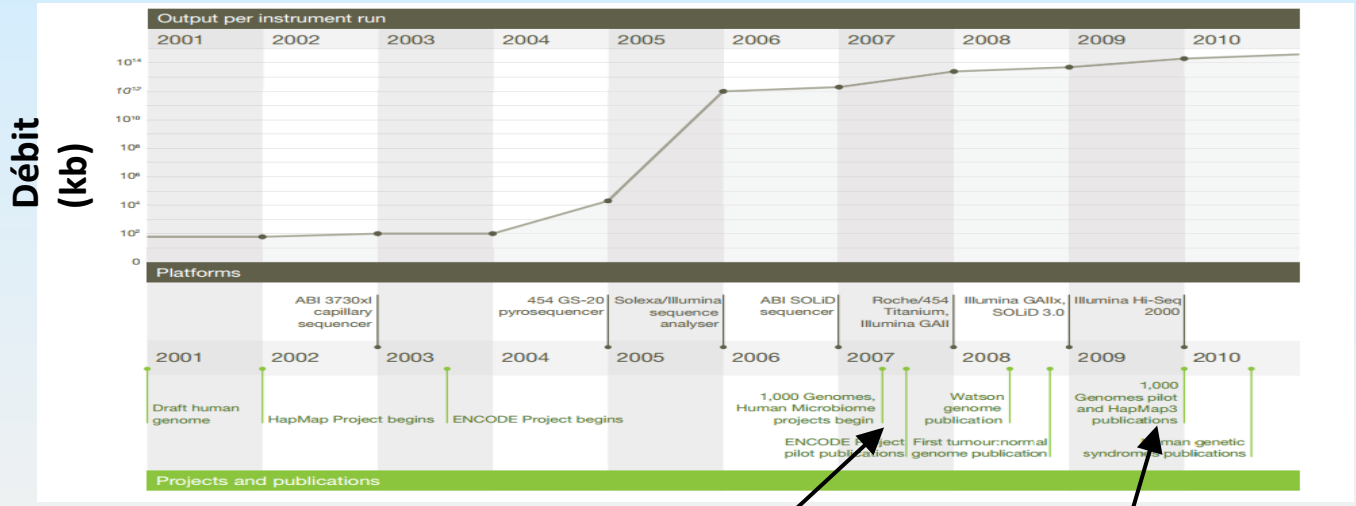
Séquence inconnue → **Séquence + homologues avec % d'identité**

Bases de données génomiques publiques



Croissance exponentielle comme le débit du séquençage et chute des prix

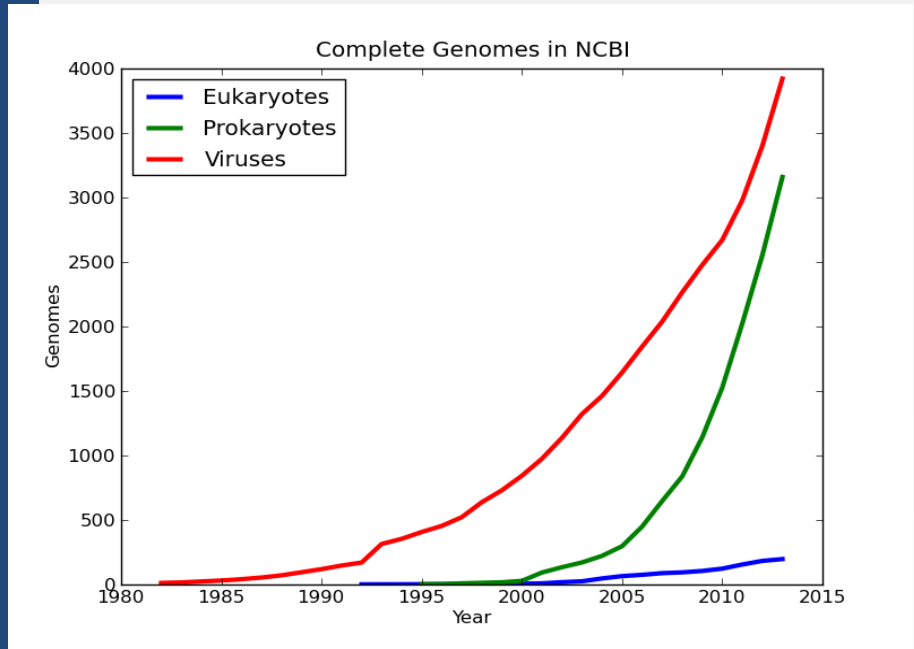
Evolution des dimensions

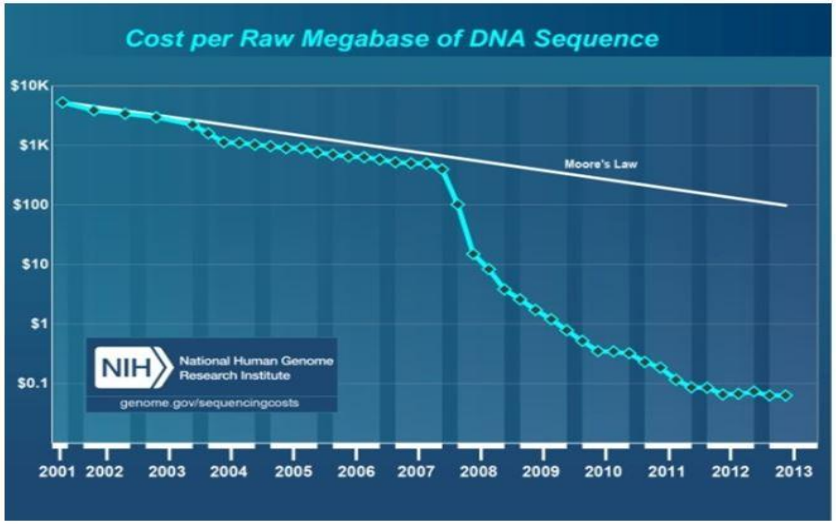


(Mardis et al, 2013)

Human microbiome Project

1000 Génomes, HapMap publications



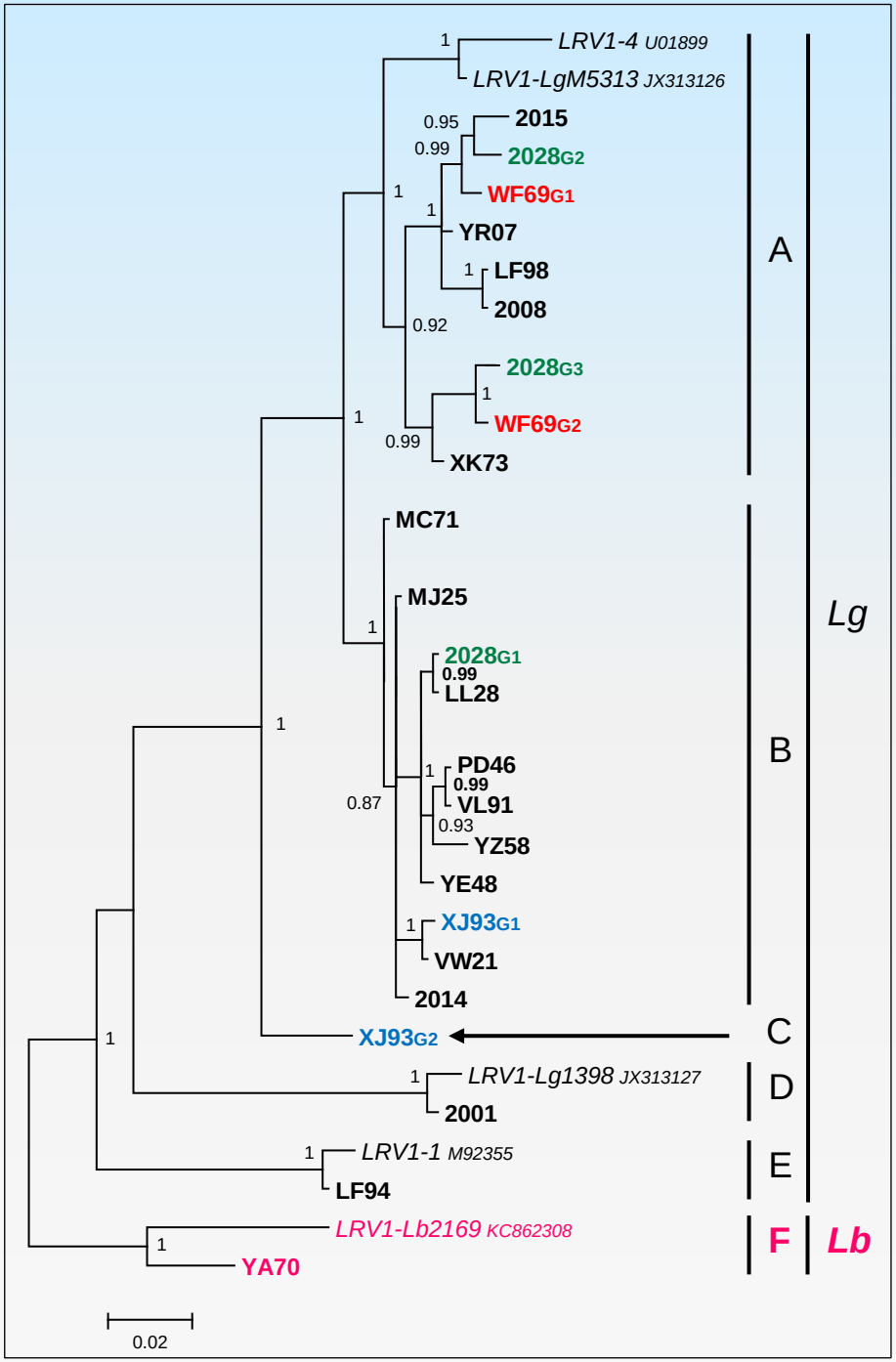


En conclusion de l'étude

Séparation en fonction des hôtes :
L. guyanensis / L. braziliensis

Première identification de multiples infections de LRV

24 nouveaux génomes :
Ressources pour études du virus…
Y compris pour le diagnostic



Avant les classes…

- Il est possible de télécharger des données publiques
- Accéder des logiciels gratuits pour travailler
- Reproduire des exemples concrets… prenez ce qui vous parle !
- Se passer des « complications » de l'informatique

- Sans se passer de :
    - l'informatique : un accompagnement et une organisation centrale ?
    - Ressources de calcul : Pour être à l'aise, un serveur

Exemple : Une interface Galaxy centralisé sur un serveur au rectorat ?

https://usegalaxy.org/

Merci de votre attention !

Pratique avec

Unipro UGENE

Institut Pasteur de la Guyane

Université de Guyane

ceba

KWATA Etude - Conservation Faune de GUYANE

l'Europe s'engage avec le FEDER

BioGeCo Biodiversité, gènes & communautés