100 To de tuberculose, pour quoi faire ?

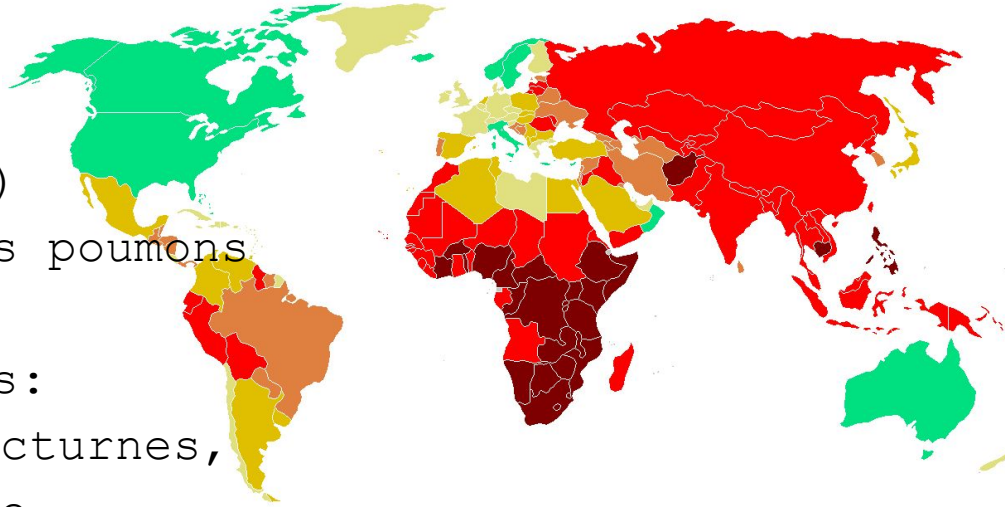
Christophe Guyeux, 13/12/2021
Gaëtan Senelle, Guislaine Refrégier, Christophe Sola

JCAD 2021

INTRODUCTION

La tuberculose

- Maladie infectieuse
- Bactérie (M. tuberculosis)
- Touche le plus souvent les poumons
- Transmission par aérosols
- Signes cliniques variables:
  - toux, fièvre, sueur nocturnes,
  - amaigrissement, fatigue,
  - râles à l'auscultation…
- 1,4 million de morts en 2019



Histoire

- Théophile Laennec conceptualise la maladie (1826)
- J.-A. Villemin : il s'agit d'une maladie infectieuse
- Robert Koch identifie le bacille tuberculeux comme agent étiologique (1882)
- C. von Pirquet développe le test de sensibilité à la tuberculine
- Vaccin par Albert Calmette et Camille Guérin (1921)
- Premier séquençage du génome 1997

Le Monde

G. Christophe

ARCHIVES

Des chercheurs ont séquencé le génome du bacille de Koch

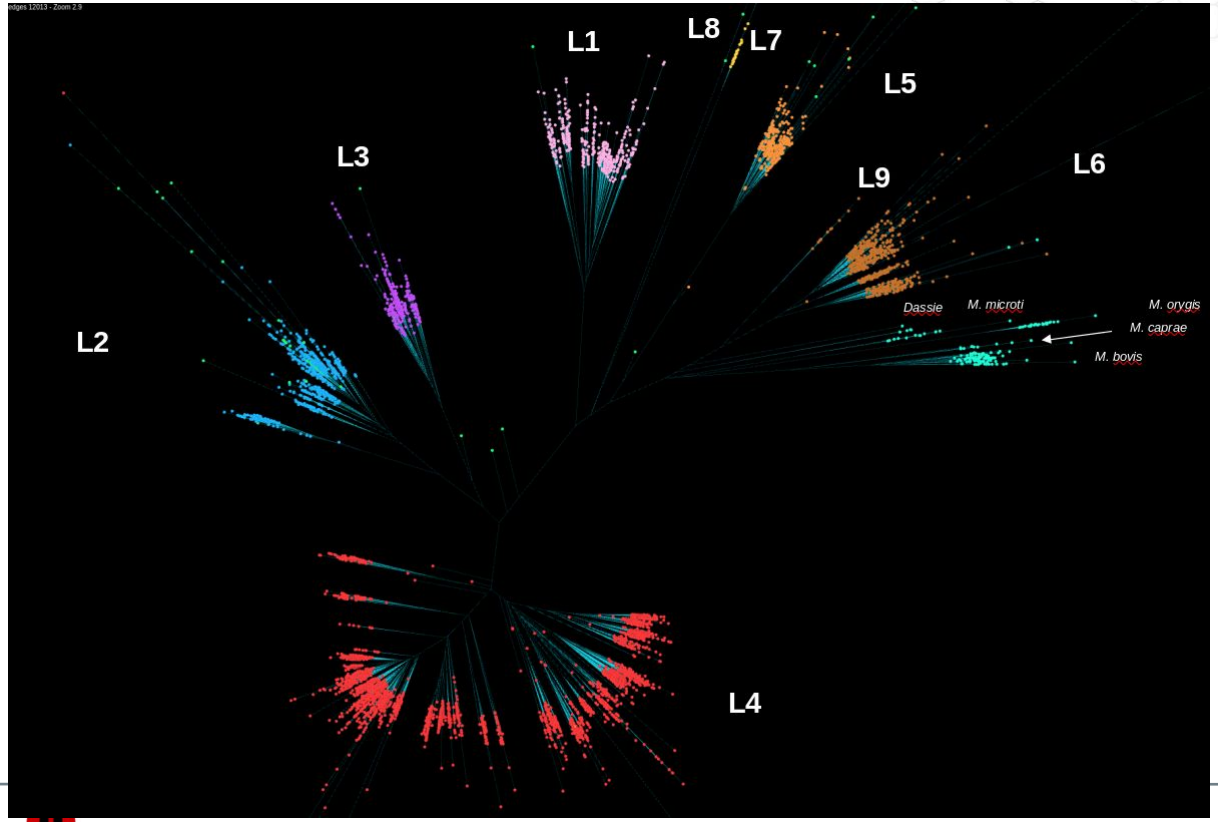
Par CATHERINE VINCENT

Publié le 12 juin 1998 à 00h00 - Mis à jour le 12 juin 1998 à 00h00 · Lecture 2 min.

femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

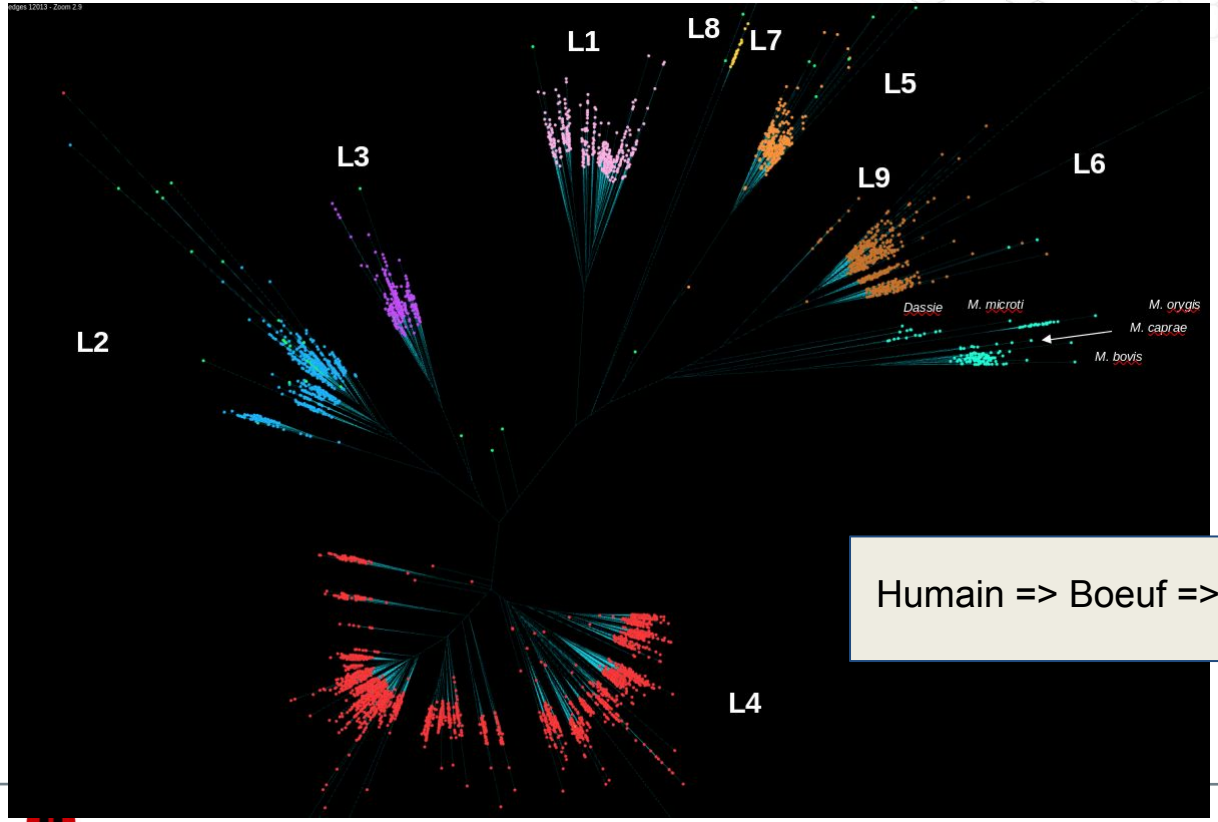
Phylogénie du MTBC



femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

Phylogénie du MTBC

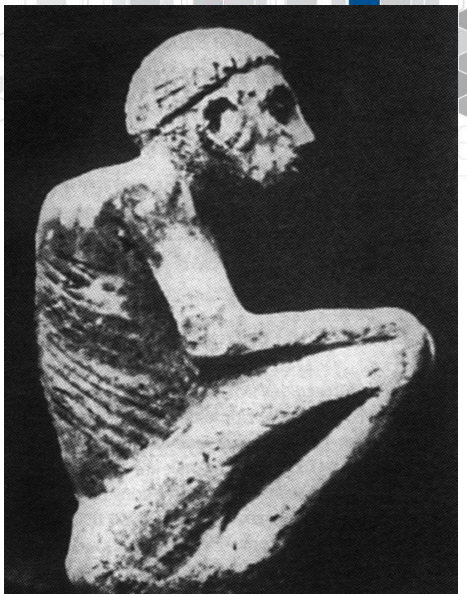


Humain => Boeuf => Humain ?

femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

M.tuberculosis reste méconnue



- Quel âge ?
  - apparue avec l'homme moderne ?
  - paléolithique ? néolithique ?

- Combien de lignées ?
  - L2 (1995), L1 (2001), L3 (2002)
  - L7 Ethiopie (2012), L8 (2020), L9 (2021)

Une fois le génome d'un malade obtenu, quels antibiotiques choisir ?

Quelle émergence par lignée (L1) ?

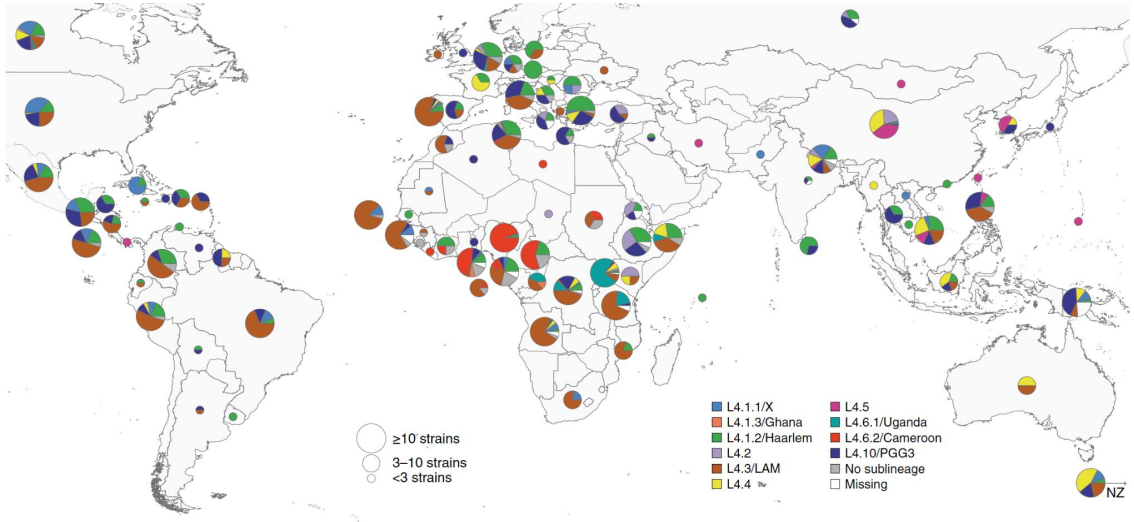
1. African origin, Asian dispersal, Asia to Africa back-migration

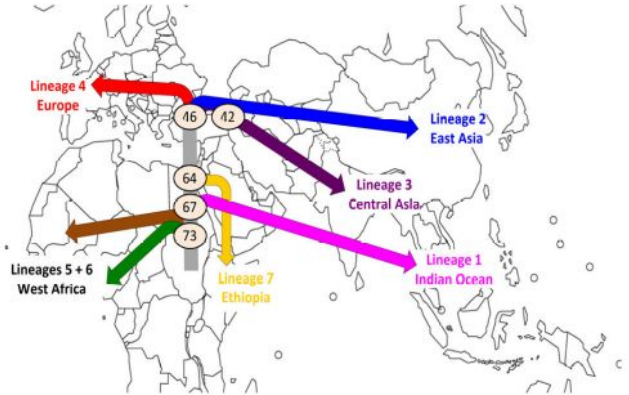
2. Asian origin, West and South Asia dispersal, Arabian Peninsula and East Africa migration

3. Mixed-1 : African origin, African and Asian dispersal, Asia to Africa back-migration

4. Mixed-2 : Asian origin, African and Asian dispersal

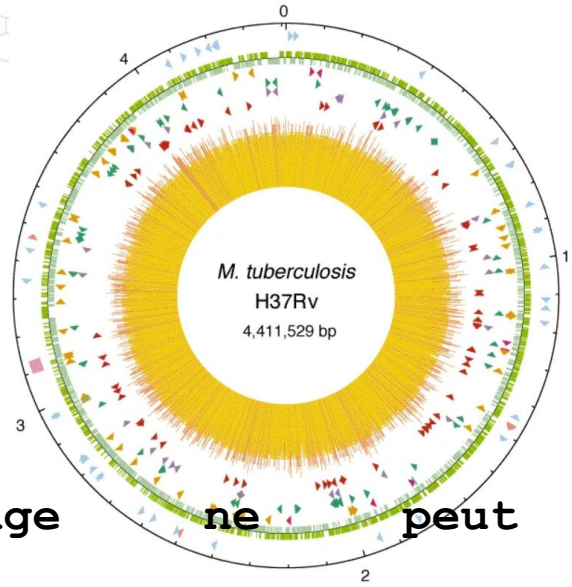
Quelle distribution par sous-lignée (L4) ?





Génomes & Big data

Un génome de *M. tuberculosis* ?

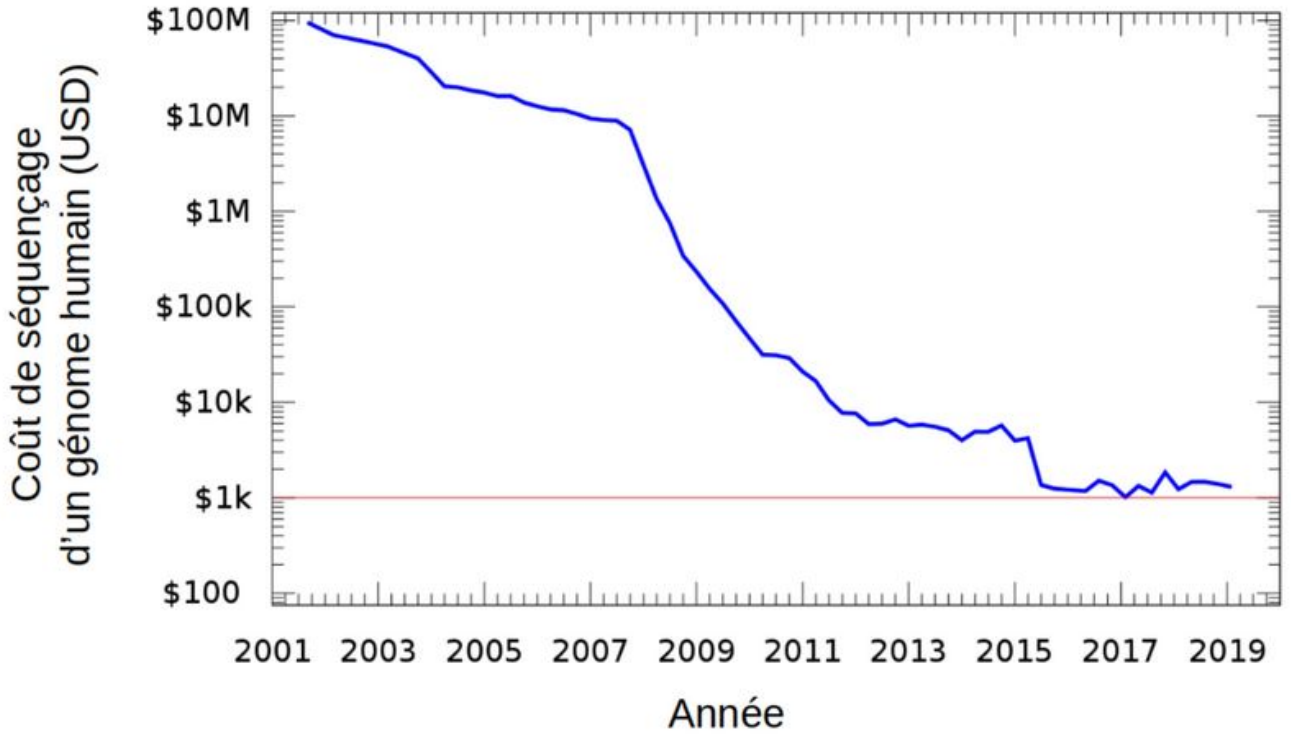


- Une molécule ADN circulaire
  - ~4 millions de nucléotides
  - ~4000 gènes

- **Aucune technique de séquençage ne peut lire une si grosse molécule d'un bloc**
  - séquençage aléatoire de petits reads (~100 nucl.)
  - chaque nucléotide est lu plusieurs centaines de fois

- le coût de séquençage est très abordable
  - mais… le problème d'assemblage est NP complet

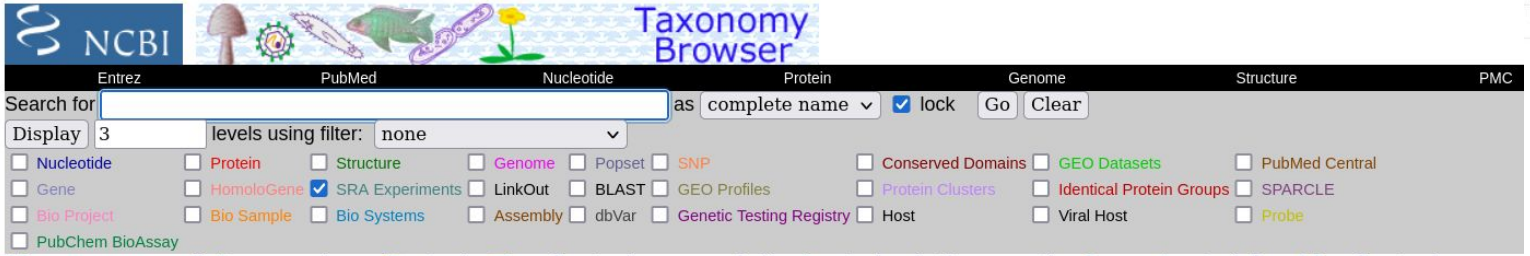
Coût de séquençage : une révolution



Un génome = beaucoup de reads

```
guyeux@bilbo:~$ head Mufindiensis_GCCAAT_L007_R1_001.fastq
@DE18INS60511:181:C8DLNANXX:7:1101:1313:2125 1:N:0:GCCAAT
NAGTAGTGGTTGAAGTAGTGGTTGGCTGATTTGTCCATTTTAGTNNNNCTCCAGTAGTCAAGAATCTTATCACATTGGTTTGATCNNNNNNNNNNNNNNN
+
#<<BBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF####<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFB###############
@DE18INS60511:181:C8DLNANXX:7:1101:1319:2157 1:N:0:GCCAAC
CTTGTTACTTGTATAGAAGACAAATAAGCCAATTCATAATGCTANNAGATTATTGAAATTCATTCTTGATTTACAAGAAACGATGACAAACAAAAACTTG
+
BBB/<<FFF<BFFFF/FFFFFFFFFFF/BFFFBFFFFFF<FFF##<</<FFFBFFFFFFFBBFFF<F<FF/FFFFFFFFB</B<FFFFFBFFF<FFF##
@DE18INS60511:181:C8DLNANXX:7:1101:1451:2173 1:N:0:GCCAAT
AGAATGATAGAGGTGCATATTAAGATACCATGAAGATTCATTTANAAGCTTATGCAATGGTTTGTATCCTGTCCAATTTAGAGATTGCATCCTTCTAATA
guyeux@bilbo:~$ wc -l Mufindiensis_GCCAAT_L007_R1_001.fastq
197124672 Mufindiensis_GCCAAT_L007_R1_001.fastq
guyeux@bilbo:~$ wc -c Mufindiensis_GCCAAT_L007_R1_001.fastq
12985486648 Mufindiensis_GCCAAT_L007_R1_001.fastq
guyeux@bilbo:~$ ls -lh Mufindiensis_GCCAAT_L007_R1_001.fastq
-rwxr-xr-x 1 guyeux and 13G Dec 13 11:01 Mufindiensis_GCCAAT_L007_R1_001.fastq
```

> 100000 génomes disponibles



Lineage (full): cellular organisms; Bacteria; Terrabacteria group; Actinobacteria; Actinomycetia; Corynebacteriales; Mycobacteriaceae

- **Mycobacterium** 146,712 *Click on organism name to get more information.*
  - **Mycobacterium aemonae**
  - **Mycobacterium albicans** 17
  - **Mycobacterium album**
  - **Mycobacterium alsense** 9
  - **Mycobacterium angelicum** 1
  - **Mycobacterium anthracenicum**
  - **Mycobacterium aquaticum**
  - **Mycobacterium aquiterrae**
  - **Mycobacterium asiaticum** 21
    - **Mycobacterium asiaticum DSM 44297**
  - **Mycobacterium attenuatum** 2
  - **Mycobacterium avium complex (MAC)** 5,273
    - **Mycobacterium arosiense** 2
      - **Mycobacterium arosiense ATCC BAA-1401 = DSM 45069**
    - **Mycobacterium avium** 3,488

femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

Analyse génomique : qu'étudier ?

Que faire d'un génome ?

- Existence d'une référence (coûteuse)
  - H37Rv, L4, assemblé et étudié depuis 25 ans

- 99,98% de similarité entre deux souches
  - nb de différences augmente
    => + grand éloignement à la référence (phylogénie)
  - Comparer les différences entre deux génomes à :
    - la virulence, la résistance
    - la lignée
    - intra/extra pulmonaire, humaine/animale
    - …

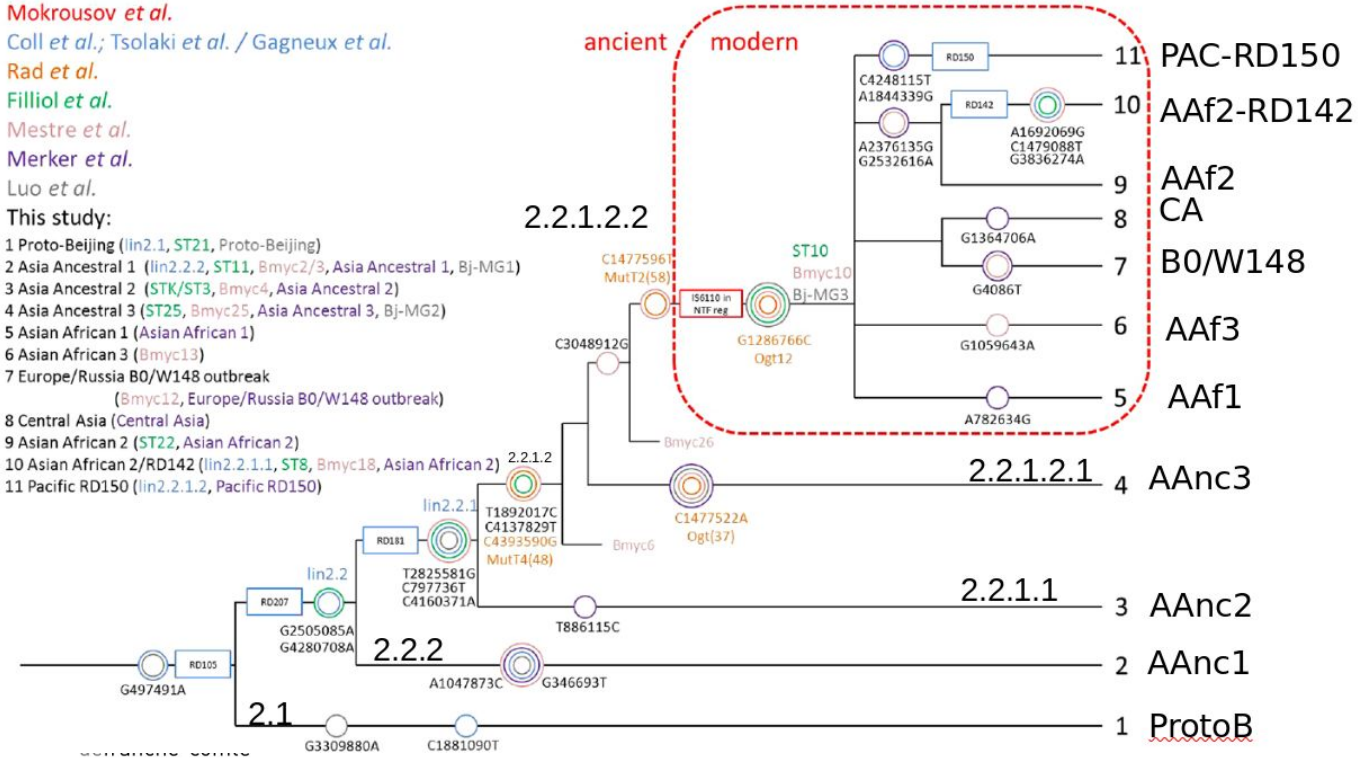
Quelles différences entre génomes ?

- Un nucléotide, localement
  => Classifications à base de SNP

lineage	Position*	Gene coord.	Allele change
lineage1	615938	1104	G/A
lineage1.1	4404247	1056	G/A
lineage1.1.1	3021283	711	G/A
lineage1.1.1.1	3216553	339	G/A
lineage1.1.2	2622402	51	G/A
lineage1.1.3	1491275	1038	G/A
lineage1.2.1	3479545	375	C/A
lineage1.2.2	3470377	303	C/T
lineage2	497491	810	G/A
lineage2.1	1881090	5787	C/T
lineage2.2	2505085	615	G/A
lineage2.2.1	797736	804	C/T
lineage2.2.1.1	4248115	1602	C/T
lineage2.2.1.2	3836274	618	G/A
lineage2.2.2	346693	1059	G/T
lineage3	3273107	894	C/A

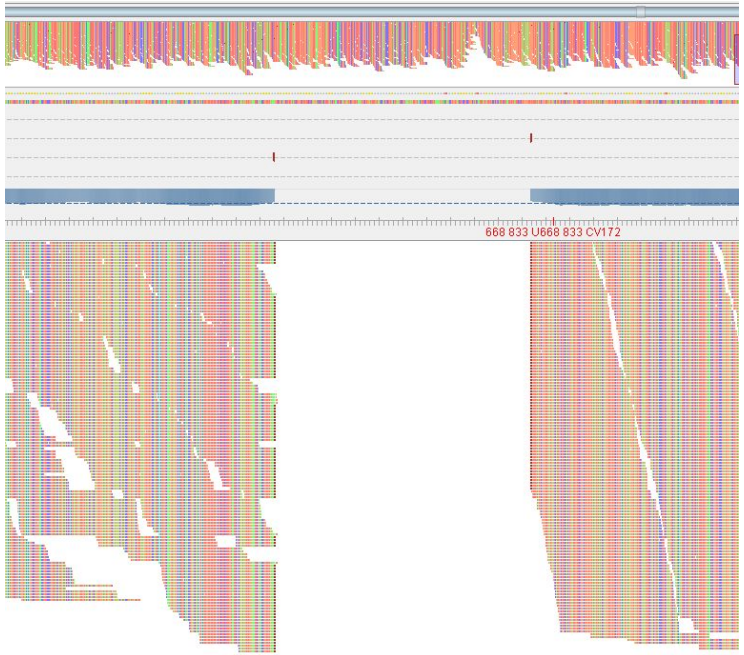
Quelles différences entre génomes ?

- Exemple de la lignée 2 :

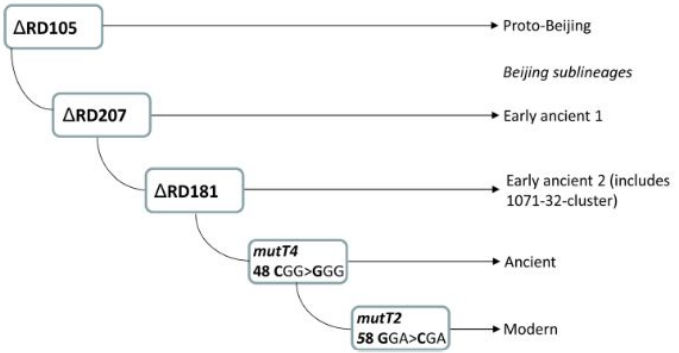


Quelles différences entre génomes ?

- Une région délétée % la référence

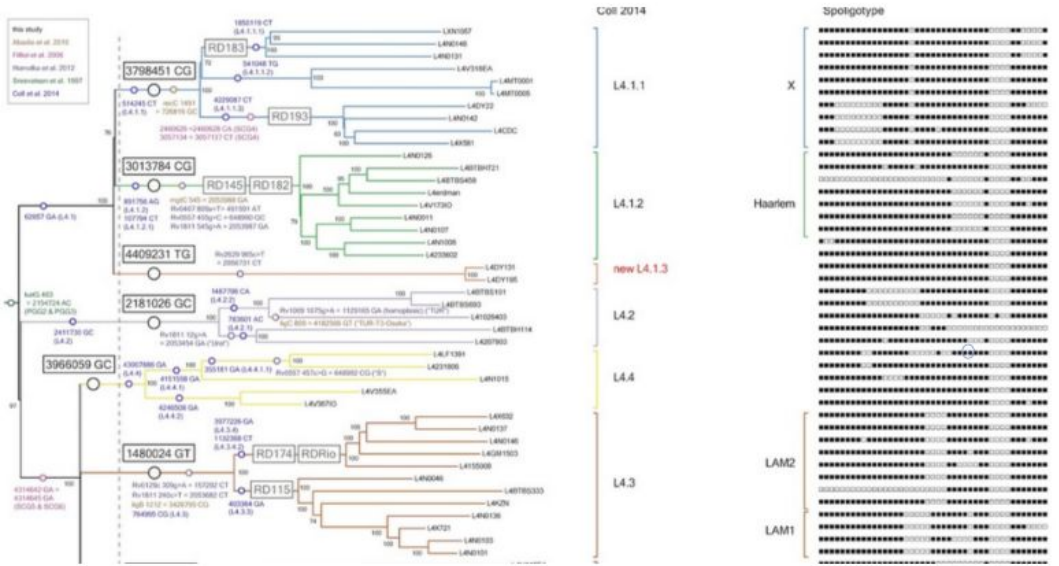


- Duplications possibles
- Quid des régions absentes dans la référence ?



Quelles différences entre génomes ?

- Le locus CRISPR :

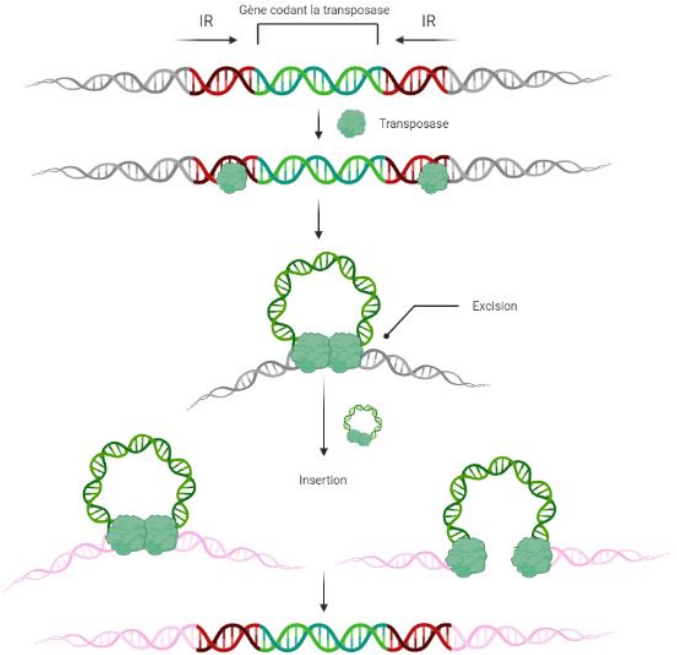


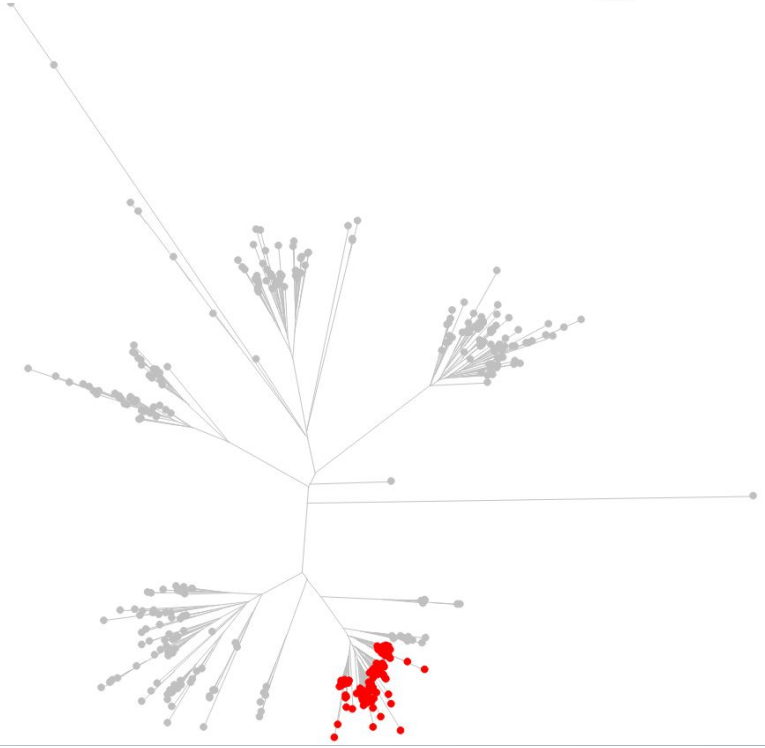
femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

Quelles différences entre génomes ?

- Les séquences d'insertion





Analyse génomique : comment faire ?

Première approche : blast

- Construire une base de données de reads
  - Une par génome (100000 BDD)
- Pour une séquence donnée
  - rechercher les reads similaires dans la bdd
  - alignement local (Smith-Waterman)

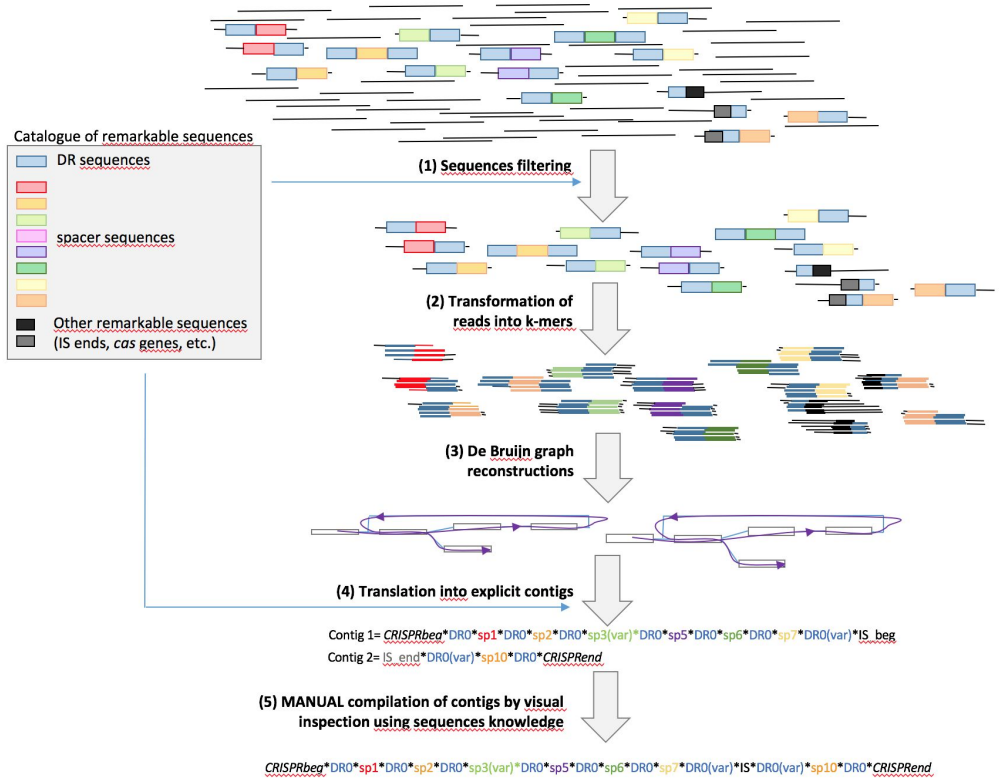
- Exemple : SNP
  - On blaste un bout de séquence autour du SNP
  - On regarde les reads avec / sans le SNP

Première approche : blast

- Exemple : IS
  - On blaste l'en tête de l'IS
  - On regarde les séquences préfixes :
    - AGGAGGAGAGAAGGGA-**ATGCCGTAGGAGGAGTT**
      position 23325 et IS6110
    - CCGTAGCTATTTAGGA-**ATGCCGTAGGAGGAGTT**
      position 1214 et IS6110
  - Les reads sont potentiellement bruités
    => clusteriser les préfixes

Première approche : blast

- Exemple : CRISPR



Première approche : évaluation

- 100 cœurs pour 50 souches / jour
  - 70 000 SNPs
  - 30 RD détectés

- Conclusion :
  - Ne passe pas à l'échelle
    - Lenteur
    - Intervention humaine
  - Basée sur des caractères connus d'avance
    - Par ex., ne trouve pas de nouveaux SNP/RD

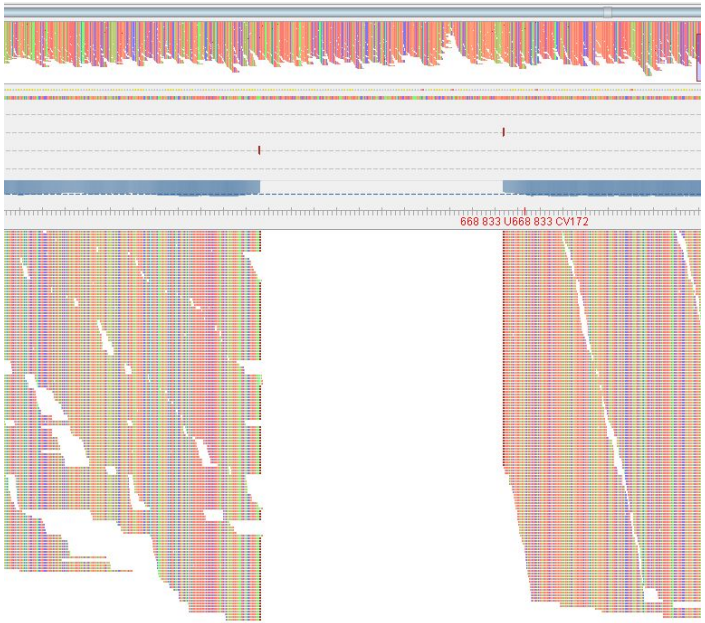
Deuxième approche : idée générale

- On aligne “parfaitement” les reads sur la référence
  - Cela peut se faire rapidement

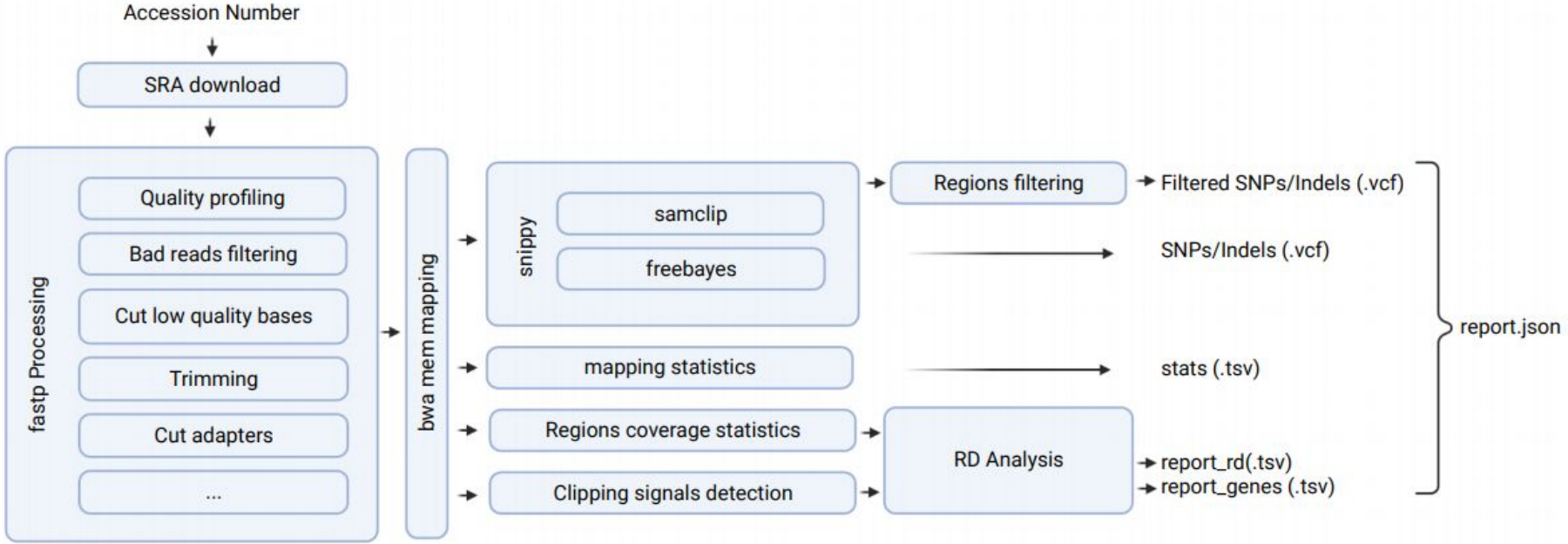
- On ne s’intéresse qu’au “reste” :
  - reads avec erreurs de lecture
  - bouts de séquences distinctes de la ref.
  - on assemble ce qui peut l’être

99,98% de similarités => 1000 fois moins de reads

Deuxième approche : IS & RD



Deuxième approche : pipeline



Deuxième approche : évaluation

- 1000 souches par jour sur machine “musclée”
  - 300 000 SNPs
  - 180 RD détectés
  - Nettoyage et statistiques de qualités
  - Données compressées
  - Introduction d’indices de qualité

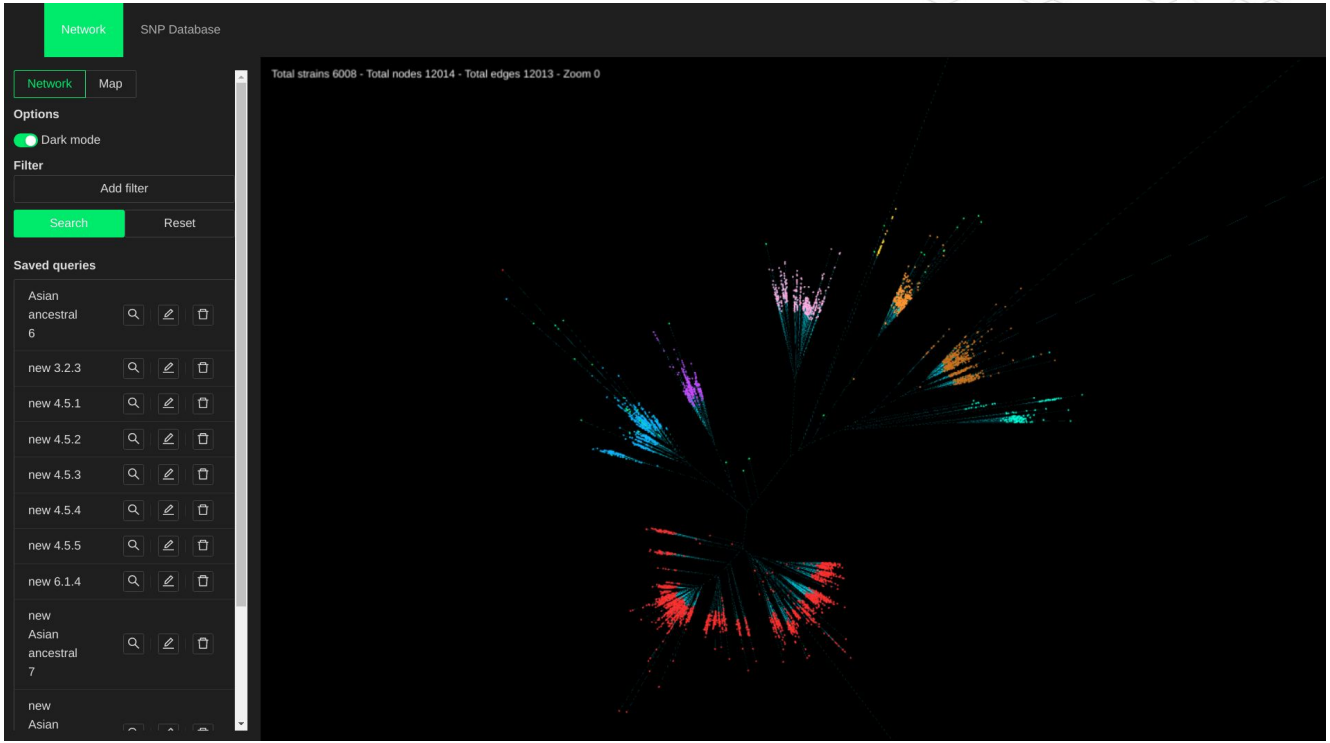
- Conclusion :
  - 100000 génomes atteignables
  - Découverte de nouveaux caractères
  - Des problèmes résistent : CRISPR, arbre phylogénétique

Conclusion

Programmation optimisée

- Proche du matériel
- Index inversé pour la recherche
- WebGL bas niveau pour visualisation
- C# pour le calcul
- GRPC pour la communication

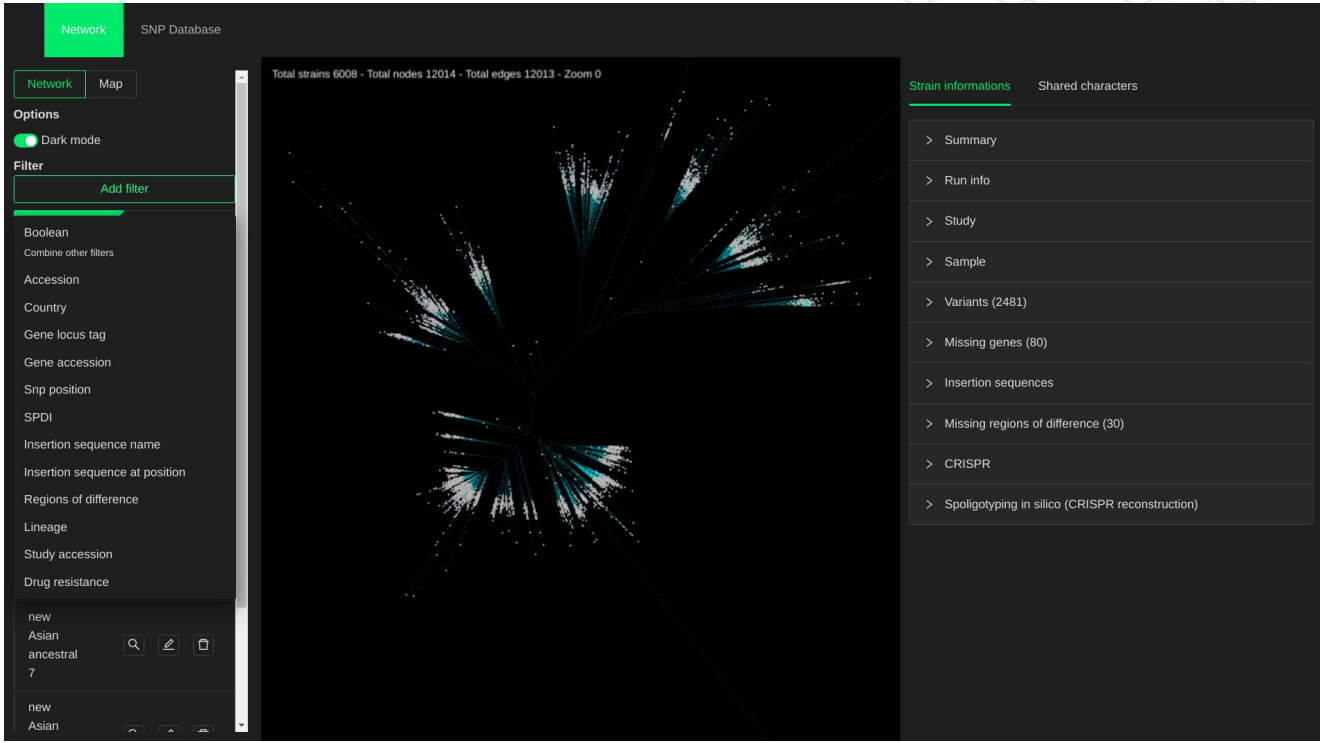
Une interface ergonomique



femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

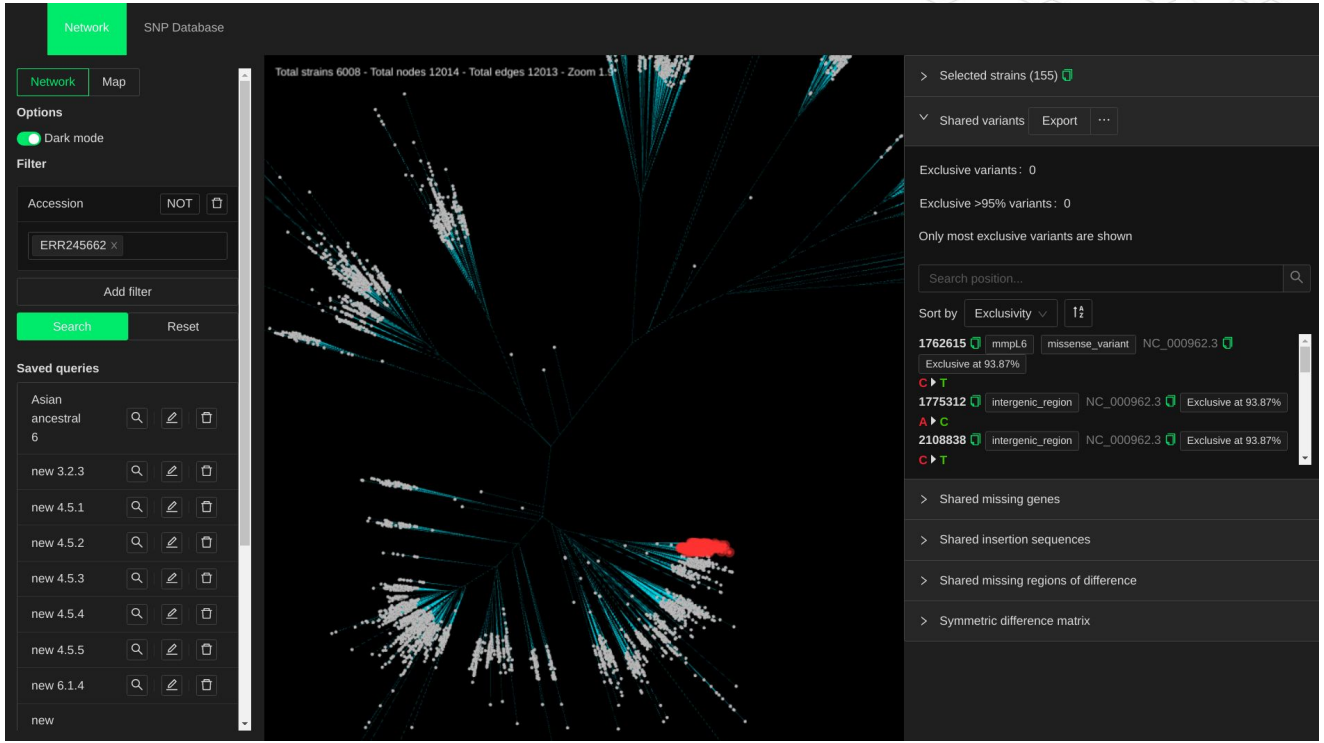
Une interface ergonomique



femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

Une interface ergonomique



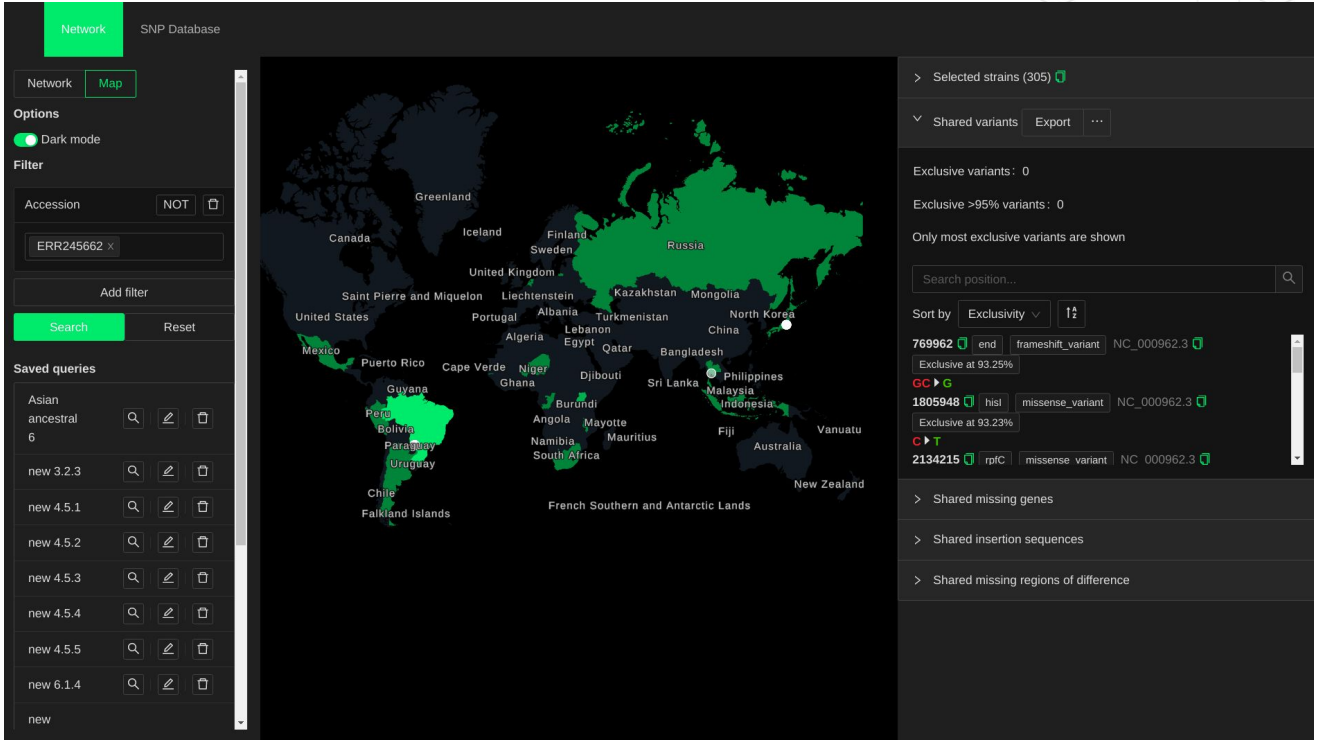
Une interface ergonomique

Supplementary File : Pairwise distance matrix (in SNPs) between all Asian Ancestral 5 strains																				
	DRR034363	DRR034366	DRR034381	DRR034395	DRR034416	DRR034422	DRR034450	DRR034455	DRR034465	DRR034470	DRR034471	DRR034476	DRR034478	DRR034482	DRR034489	DRR034493	DRR130160	DRR130203	DRR157280	DRR157281
DRR034363	0	244	250	223	254	261	266	245	241	231	255	270	258	213	214	233	393	425	272	256
DRR034366	244	0	230	205	234	253	258	227	223	243	255	270	242	227	240	243	377	439	256	234
DRR034381	250	230	0	225	246	263	262	233	259	261	267	278	268	245	232	255	397	417	258	234
DRR034395	223	205	225	0	231	248	203	218	216	232	244	251	227	222	227	222	354	406	251	235
DRR034416	254	234	246	231	0	247	278	247	219	259	273	278	258	209	246	237	389	421	230	228
DRR034422	261	253	263	248	247	0	259	252	262	252	276	265	259	242	231	242	368	412	275	263
DRR034450	266	258	262	203	278	259	0	221	205	191	209	236	202	227	238	249	341	377	280	254
DRR034455	245	227	233	218	247	252	221	0	216	200	202	205	187	208	221	246	332	378	271	243
DRR034465	241	223	259	216	219	262	205	216	0	190	204	233	199	228	211	240	332	382	245	229
DRR034470	231	243	261	232	259	252	191	200	190	0	218	211	205	218	237	250	318	354	279	257
DRR034471	255	255	267	244	273	276	209	202	204	218	0	215	195	246	215	264	320	366	295	265
DRR034476	270	270	278	251	278	265	236	205	233	211	215	0	208	251	240	281	335	379	286	274
DRR034478	258	242	268	227	258	259	202	187	199	205	195	208	0	209	228	259	331	405	266	248
DRR034482	213	227	245	222	209	242	227	208	228	218	246	251	209	0	215	258	364	412	263	251
DRR034489	214	240	232	227	246	231	238	221	211	237	215	240	228	215	0	211	359	415	236	220
DRR034493	233	243	255	222	237	242	249	246	240	250	264	281	259	258	211	0	392	420	271	243
DRR130160	393	377	397	354	389	368	341	332	332	318	320	335	331	364	359	392	0	196	373	379
DRR130203	425	439	417	406	421	412	377	378	382	354	366	379	405	412	415	420	196	0	393	407
DRR157280	272	256	258	251	230	275	280	271	245	279	295	286	266	263	236	271	373	393	0	166
DRR157281	256	234	234	235	228	263	254	243	229	257	265	274	248	251	220	243	379	407	166	0

femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

Une interface ergonomique



femto-st SCIENCES & TECHNOLOGIES

mésocentre de calcul de franche-comté

Présent et futur

- Version à 6000 génomes disponible
  - De nombreuses découvertes rendues possibles
  - Finalisation de la version à 16000 génomes

- À venir (comment ?)
  - 100000 génomes
  - Version publique
    - BDD en lecture-écriture
  - CRISPR, VIRU-VNTR…
  - I.A. et résistance aux antibiotiques

Merci!