

Licence 3 - DLSV316

# Analyse des Séquences Génomiques



# Objectifs du cours

- Rappels programmation R
- Extraction d'information de bases de données de séquences
- Faire et analyser sous R un alignement multiple de séquences ADN/protéines
- Manipuler des protéines en 3D
- Raisonner sur la relation entre conservation de séquence et fonction

# Programme

	21-mars	22-mars	28-mars	29-mars	05-avr	11-avr	12-avr	18-avr	19-avr
	33A	226	313	226	33A	315/336	33A	315	315
8:15-10:15							AL		
10:15-12:30	DG		DG			AL	AL	DG	
13:30-15:30		DG		AL	DG				AL
15:45-17:45					DG				AL
	Prise en main R	exercices R	Fin R . Banques de données génomiques	Alignement multiple	Lecture fasta R, calcul Entropie	Entropie sur ali ADN	Traduction ADN>prot. Entropie sur ali prot	Intro Pymol	Mapping entropie sur 3D. Visu 3D

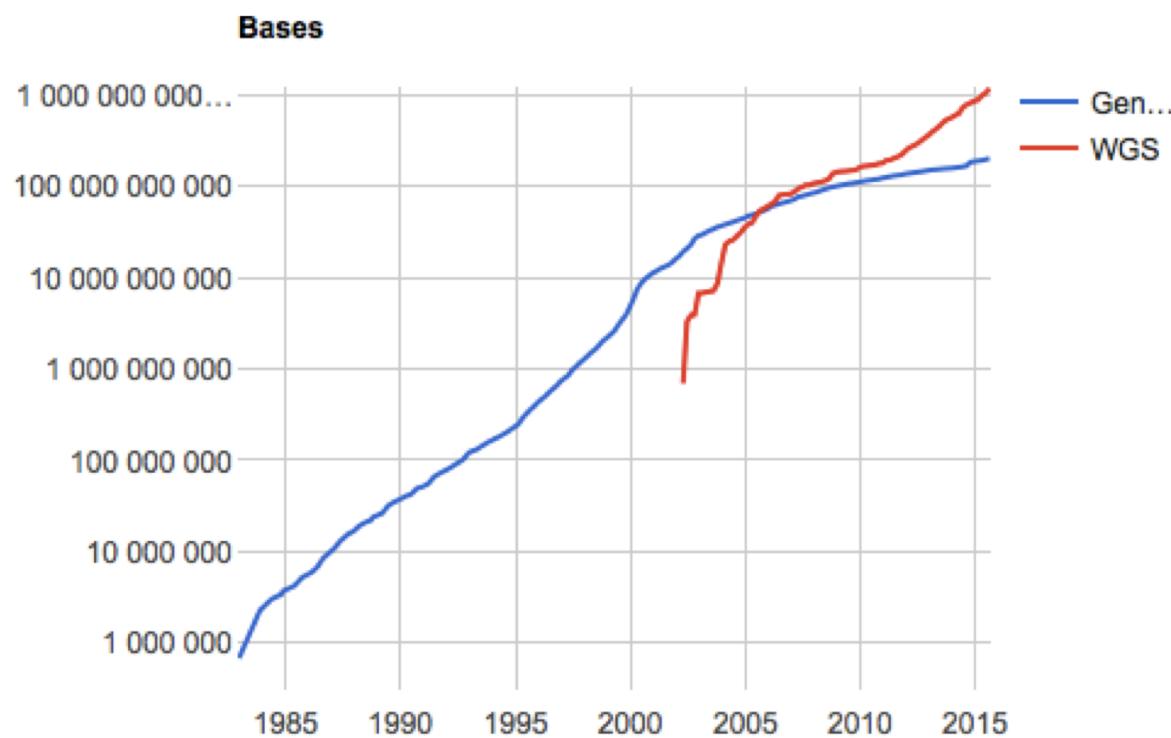
# Les banques de séquences

# Genbank: La banque d'ADN du NIH

- **Etat Genbank 2018**

- 253 Gbases
- 207M séquences
- Genbank double environ tous les 14 mois depuis ses débuts en 1982.
- Nouvelle version tous les 2 mois

(WGS: whole genome shotgun)



# Identifiants Genbank

- Chaque enregistrement se voit attribuer un numéro d'acquisition, stable et unique, et chaque séquence un numéro de version (anciennement numéro GI.)
- Quand un changement est effectué dans un enregistrement Genbank, le num. d'acquisition reste, la version change.

```
LOCUS NC_000913 4641652 bp DNA circular CON 08-AUG-2016
DEFINITION Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION NC_000913
VERSION NC_000913.3
```

Début de la fiche Genbank de E. coli

# Autres banques nucléotidiques

- EMBL: Equivalent européen de Genbank. Format différent, contenu presque identique.
- DDBJ: équivalent au Japon
- Banques spécialisées Certaines collections de séquences, bien que généralement présentes dans Genbank, sont beaucoup plus utiles lorsqu'elles sont rassemblées dans des banques spécialisées, par ex:
  - Récepteurs des lymphocytes T (Réarrangements de l'ADN)
  - Génomes HIV, etc.
- SRA: short read archive...

# Next Generation Sequencing



Nanopore  
Minlon



Lifetech Ion  
torrent PGM



Illumina  
MySeq



Lifetech Ion  
proton



Illumina  
Hi-Seq 2000



Illumina  
NovaSeq

50Mb

400 Mb

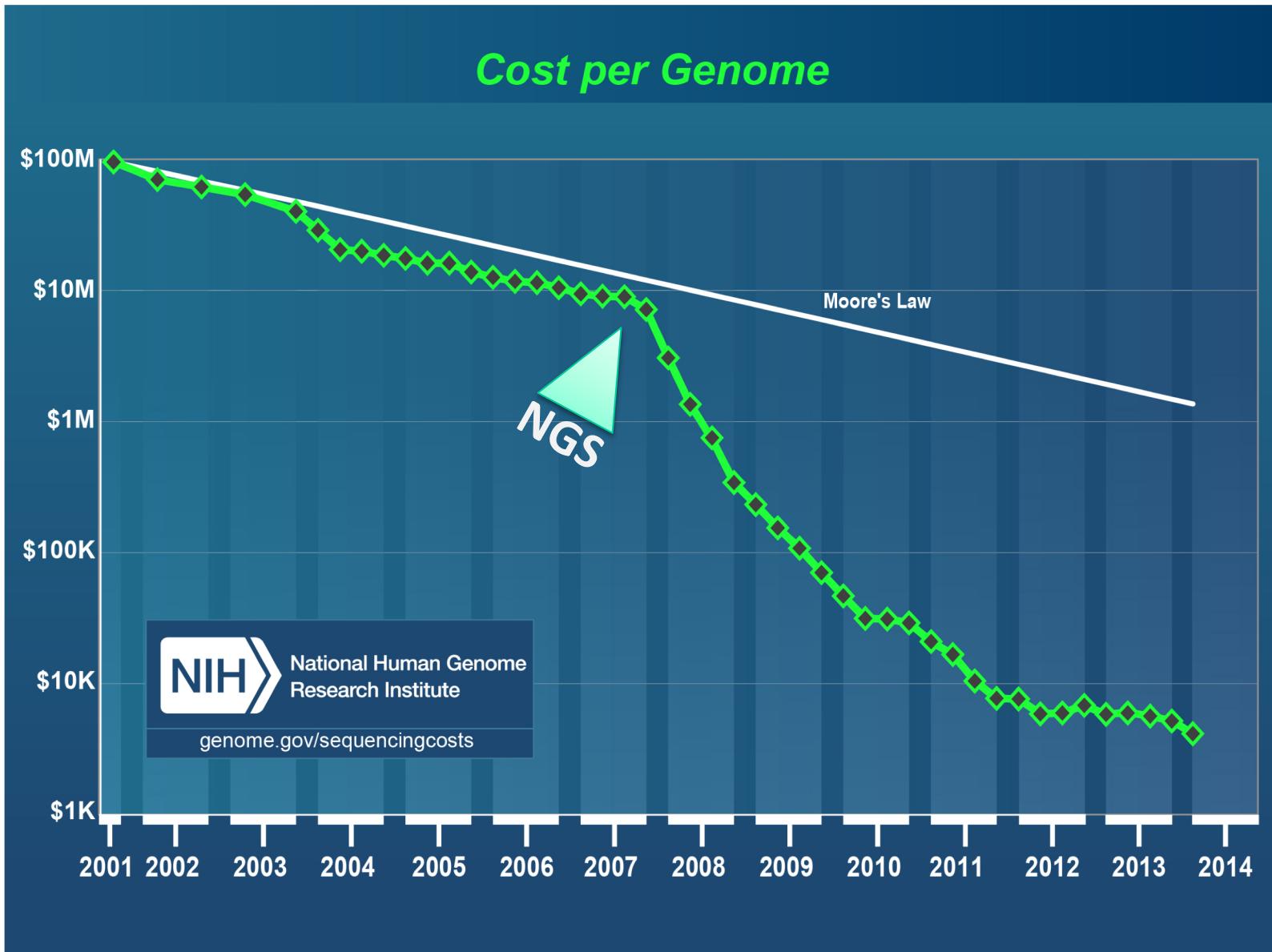
4 Gb

20 Gb

300 Gb

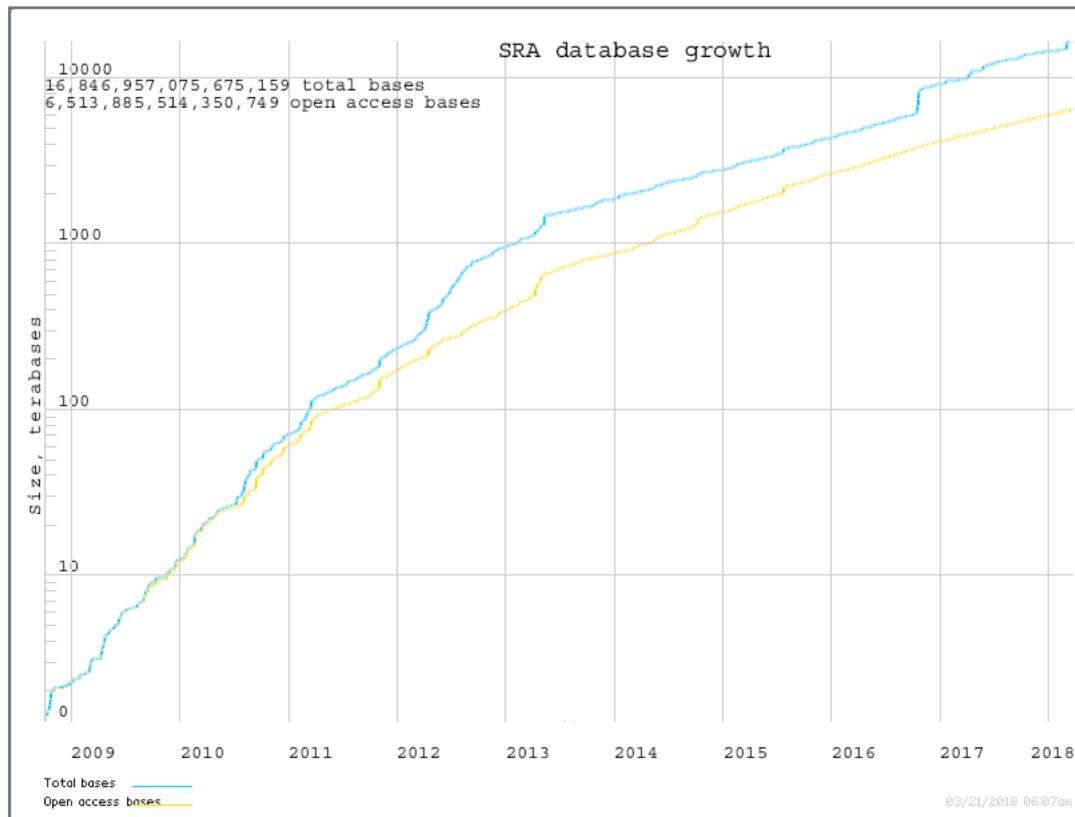
3Tb

# Le bouleversement des NGS



# The SRA (short read archive) database

- 3.200.000 entries (2018)
  - Each entry: ~50M sequences (DNA or cDNA fragments)



# Banques protéiques

- Swissprot (UniProtKB/Swissprot).
  - La mieux annotée des banques protéiques. 2018: 550.000 entrées.
  - Curation par experts seulement (basé sur publis)
  - Attention: toutes les protéines connues n'y sont pas!
- TrEMBL (UniProtKB/TrEMBL):
  - banque protéique produite automatiquement par traduction banque EMBL. 2018: 108.000.000 entrées
- Uniprot=Swissprot+TrEMBL

# Formats de données en bioinformatique

- La majorité des données de bioinfo sont de type texte:
  - FASTA, FASTQ, SAM, VCF, BED, GFF, GTF, TSV, CSV, WML, JSON
- Pour des raisons de performance et d'espace, certains sont en format binaire
  - BAM, VCF.GZ, FASTQ.GZ

# Même donnée, différents formats

Format **JSON**

<https://tools.ietf.org/html/rfc4627>

```
users : {  
first_name: "James", last_name:  
"Watson", birthday: "1928-04-06"  
}
```

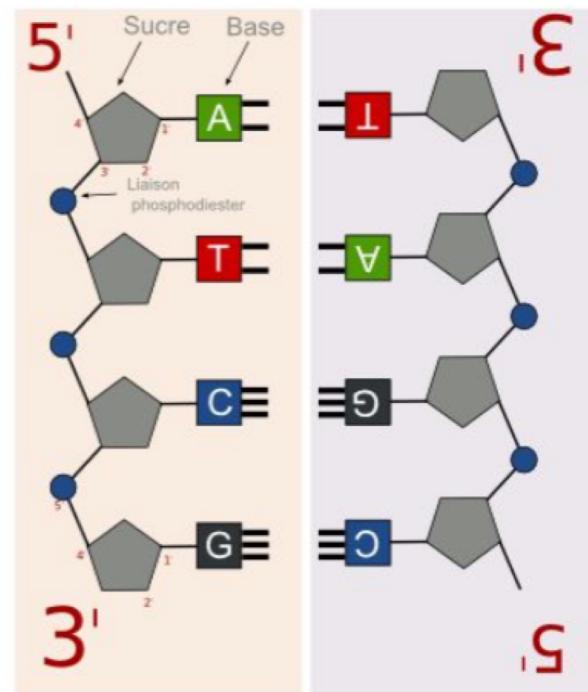
Format **XML**

<https://www.w3.org/TR/REC-xml/>

```
<users>  
<first_name>James</firstname>  
<last_name> Watson</last_name>  
<birthday>19280406</birthday>  
</users>
```

# Formats de séquences d'ADN

- Toujours dans le sens 5'->3'
- Sur quel brin?



# Format fasta

\*.fa , \*.fasta

```
>identifiant1 commentaire libre
CAGCATCGATCGTCGGCGATGCATGCGGATGCTAGCTGATCACGATGC
CGCATGCTAGTCAGGCAGGGAGGGATATTATTAGCGGCGTATCGGATGA
CAGCATTACGGCGGGAGTGCTATTATTATGAGCGGGCGAT
>identifiant2 commentaire libre
CAGGCAGGAGGTTCTTATTATATCGGCAGGGCGGAGGCAGGCGATGCATC
CAGTGCAGTGCAGTAGTCAGCGATGCATTATGACTGACTCAGTTT
CCCGCTAGCTATGCTATGCTATTGATCGATTGTGAGCTGATCTGGC
CAGCTATGCTTAGTA
```

# Protein fasta (from Uniprot)

gene ID	species	full (english) name
>tr Q4A489 Q4A489_BATTH	Actin-1	
CPESLFQPSFLGMESAGIHETTYNSIMKCDVDIRKDLYANTVLSGGTTMFPGIADRMQKE		
ITALAPSTMKIKIIAPPERKYSVWIGGSILA		

# Format Genbank

**LOCUS** L10986 47233 bp DNA linear INV 21-SEP-2004  
**DEFINITION** *Caenorhabditis elegans* cosmid F10E9, complete sequence.  
**ACCESSION** L10986  
**VERSION** L10986.2 GI:38638818  
**KEYWORDS** HTG.  
**SOURCE** *Caenorhabditis elegans*  
**ORGANISM** *Caenorhabditis elegans*  
 Eukaryota; Metazoa; Nematoda; Chromadorea; Rhabditida;  
 Rhabditoidea; Rhabditidae; Peloderinae; *Caenorhabditis*.  
**REFERENCE** 1 (bases 1 to 47233)  
**AUTHORS** .  
**CONSRTM** WormBase Consortium  
**TITLE** Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The *C. elegans* Sequencing Consortium  
**JOURNAL** Science 282 (5396), 2012-2018 (1998)  
**MEDLINE** 99069613  
**PUBMED** 9851916  
**FEATURES**  
 source Location/Qualifiers  
 1..47233  
 /organism="Caenorhabditis elegans"  
 /mol\_type="genomic DNA"  
 /strain="Bristol N2"  
 /db\_xref="taxon:6239"  
 /chromosome="III"  
 /clone="F10E9"  
 gene 265..26728  
 /gene="mig-10"  
 /locus\_tag="F10E9.6"  
 CDS join(265..338,3266..3515,15194..15317,21507..21  
 21727..21887,23171..23335,24302..24472,24524..24608,  
 25012..25827,26284..26430,26478..26728)  
 /gene="mig-10"  
**/translation**=MDSCEECDLEVDSDDEEDQLFGEK CISLLSLLPLSSSTLLSNA  
 INLELDEVERPPPLNVLEEQQFPKV CANIEENELEADTEEDIAETADDEESKDPVE  
 KTNFEPSPVTMDTYDFPDPYPVQIRARPVPPKPPIDTVRYSMNNIKESADWQLDELL  
 EEELEALETQLNSSNGGDQLLLGVSGIPASSSR ENVKSISTLPPP PALS YHQT PQQPQ  
  
 QVYTGIGWEKKYKSPTPWCISIKLTALQMRSQFIKYICAEDEMTFKKWLVALRIAKN  
 GAELLEN YERACQIRRET LGPASSMSAASSSTAISEVPHSLSHHQRTPSVASSIQLSS  
 HMMNNNPTHPLSVNVRNQSPASFSVNCSQSHPSRTSAKLEI QYDEQPTGTIKRAPLDV  
 LRRVSRASTSSPTIPQEESDSDEEF PAPPVAVSVMRMPVVTPPKCPTLTSK KAPPP  
 PPKRSDTTKLQSASPMAKNDLEAALARRREKMATMEC"

BASE	COUNT	2598	a	2024	c	1888	g	2449	t
ORIGIN									
1	ttctaaaagt	cgaaaaacga	gcaattttg	atgtctagatt	ttttgatttg	acgaattttt			
61	tcatgttttt	ttcttaaaa	aaggttttt	accctttaaa	gttttcctt	cccttccaaat			
121	tttttccttc	tttcttatac	gacttctcaa	gttcaactc	taaacaaag	ctacatgtac			
181	atttccggta	aaccttgtt	ctcagaagat	ccatttctt	tttggatcat	ttattcaaga			
241	ttgaattcca	aaatttcagc	caatatggac	agttgcgaag	aggaatgcga	tctggaaagg			
301	gacagtgcg	aagaagatca	actttttgtt	gaaaagtggt	gagttcttat	tgttgtaaacc			
361	aaagaaatgt	cagtggccg	taaacacttg	actcccaaat	ggttctcg	aattaccta			
421	tgcacactt	tcaagtgtt	gccgtttagt	cttagccaa	ttgaacgtt	tagatgttaa			
481	atggaaaatg	ggtaaagttt	tttattttat	agaaaaaaagg	tttggaaaaa	aatcgagtca			
541	ctgaatagtt	tgaagaacgg	aaaaataaaa	ctttccaaaa	atcataaaa	atttatgttt			
601	tcgaaaaattt	tagtggat	tttggatgtt	tgtttgaca	aaagctaaac	catctttat			
661	gtatgtttgt	aaaatgttca	caaagatgcg	tttttttttc	aaatttgca	ggctatcttt			
721	acattcacat	ttggataattt	caaatttttc	ttatcgctaa	caaattttcc	tattttccaa			
781	attattcgtt	tttataaaggc	tttggatgtt	tgttgcgtt	atctttatgt	gtcatcgat			
		..	..	..	..	..			



# Format fastq



Sequenceur NGS

Descriptif du read (position sur la piste de séquençage, taille,...)

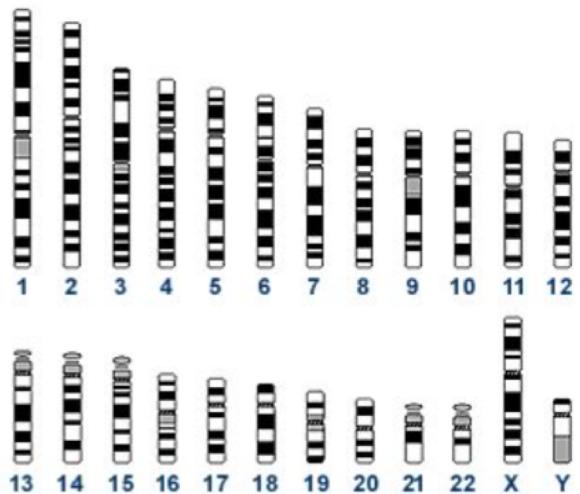
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Qualité (probabilité que la base soit correcte) encodé par code ASCII

# Les régions

Les coordonnées génomiques permettent de définir une région exacte du génome

<chromosome>:<start>-<end>  
chr7:117465784-117715971



# Accéder directement à une région

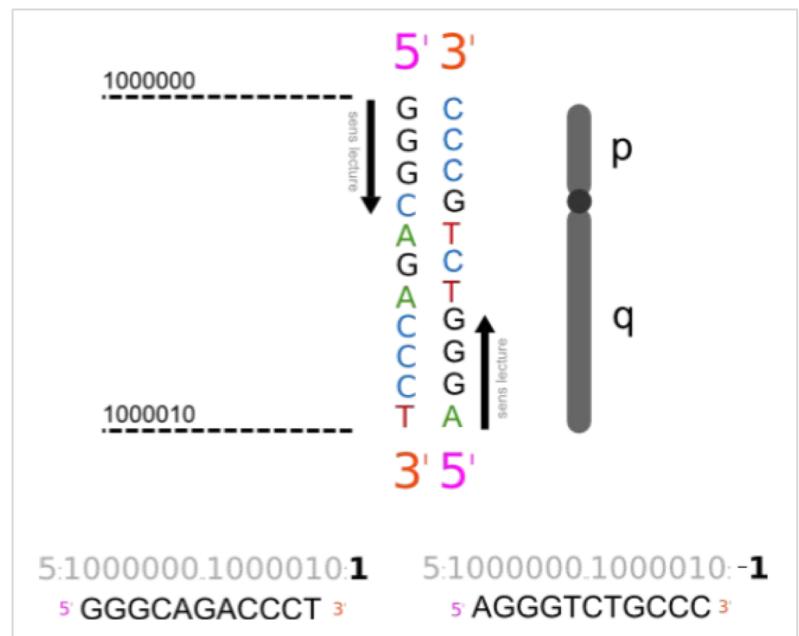
Via le browser Ensembl: <https://www.ensembl.org>

Puis choisir region du génome humain: 7:117465784..117715971

Via une URL:

<http://rest.ensembl.org/sequence/region/human/7:117465784..117715971:-1>

Attention aux versions  
d'assemblage du  
génome (Hg19, Hg38..)



# Exercice Genbank

Récupérez sur le site du NCBI *dans la section nucleotides* l'accession NC\_002549.1, au format gb

- De quelle séquence s'agit-il?
- Visualisez gènes, protéines, séquences régulatrices
- 10 premières bases du gène 1?

# Exercices

Projet

Bioinformatics

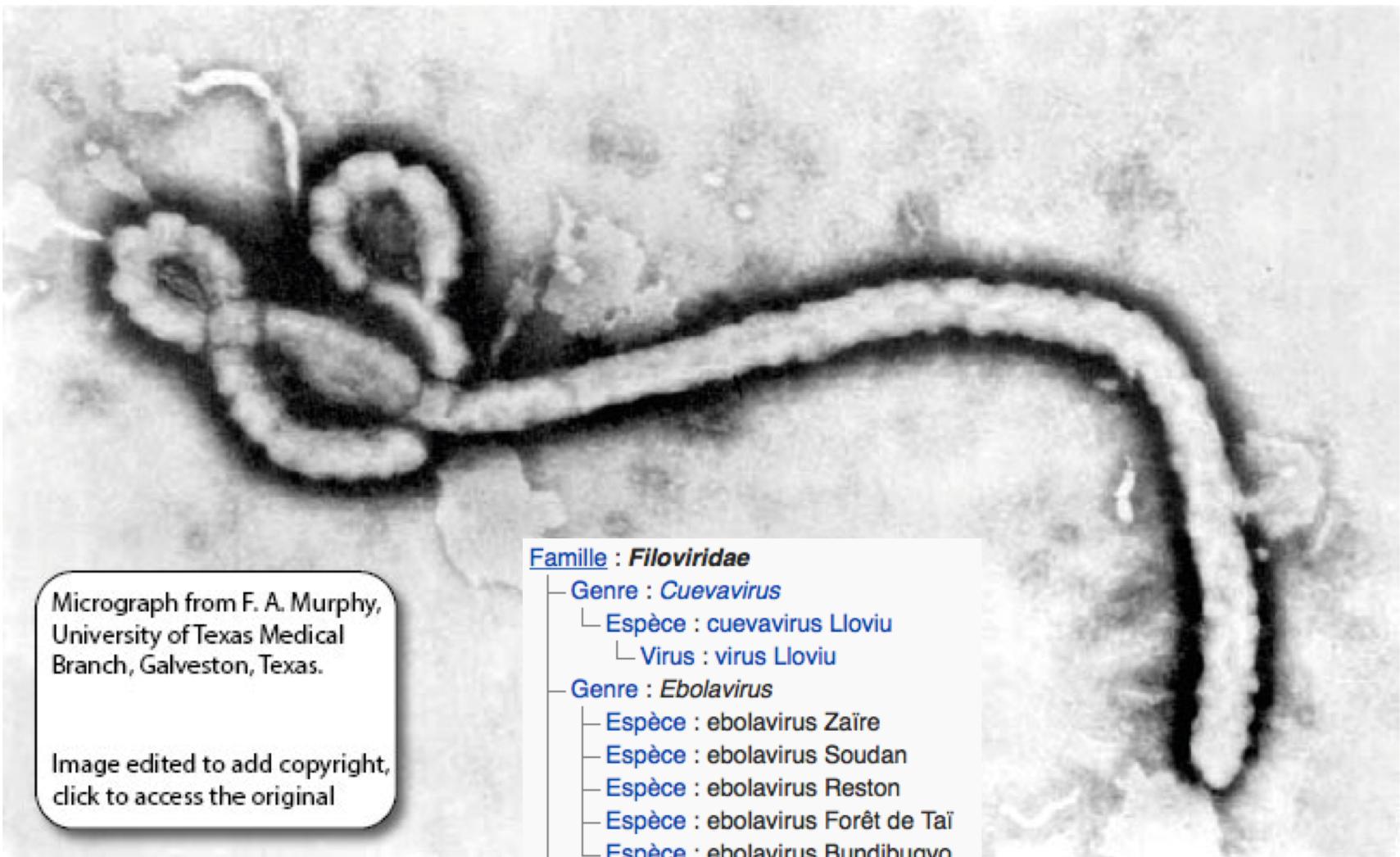
VS.

EbolaVirus

# La maladie à virus Ebola

- Virus transmis à l'homme à partir d'animaux sauvages (chauve-souris, primates), puis entre humains
- Fièvre hémorragique
- Transmission: par contact fluides/muqueuses
- Taux de léthalité moyen de 50%
- Aucun vaccin ni traitement

# EbolaVirus (EBOV): famille des filovirus



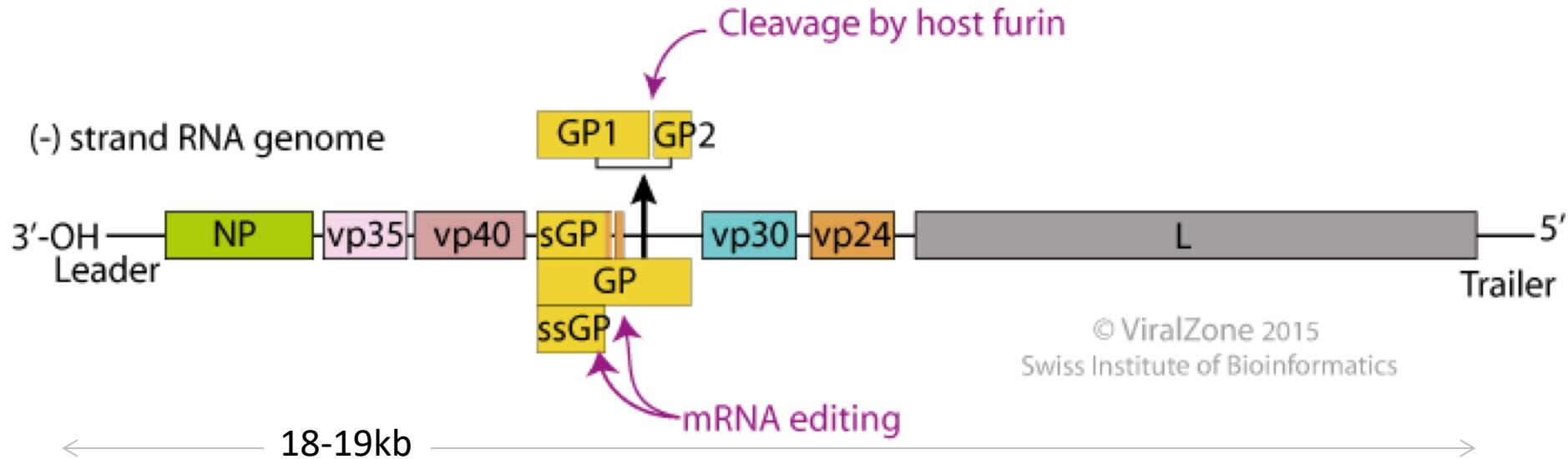
Micrograph from F. A. Murphy,  
University of Texas Medical  
Branch, Galveston, Texas.

Image edited to add copyright,  
click to access the original

## Famille : Filoviridae

- Genre : *Cuevavirus*
  - └ Espèce : cuevavirus Lloviu
    - └ Virus : virus Lloviu
- Genre : *Ebolavirus*
  - └ Espèce : ebolavirus Zaïre
  - └ Espèce : ebolavirus Soudan
  - └ Espèce : ebolavirus Reston
  - └ Espèce : ebolavirus Forêt de Taï
  - └ Espèce : ebolavirus Bundibugyo
- Genre : *Marburgvirus*
  - └ Espèce : marburgvirus Marburg
    - └ Virus : virus Marburg
    - └ Virus : virus Ravn

# EBOV: un virus à ARN brin (-)



NP	nucléoprotéine
VP35	co-facteur RNA-pol
VP40	protéine de matrice
xGPx	glycoprotéines
VP30	activateur transcription
VP24	protéine de matrice
L	RNApol RNA-dépendante + coiffe et polyA

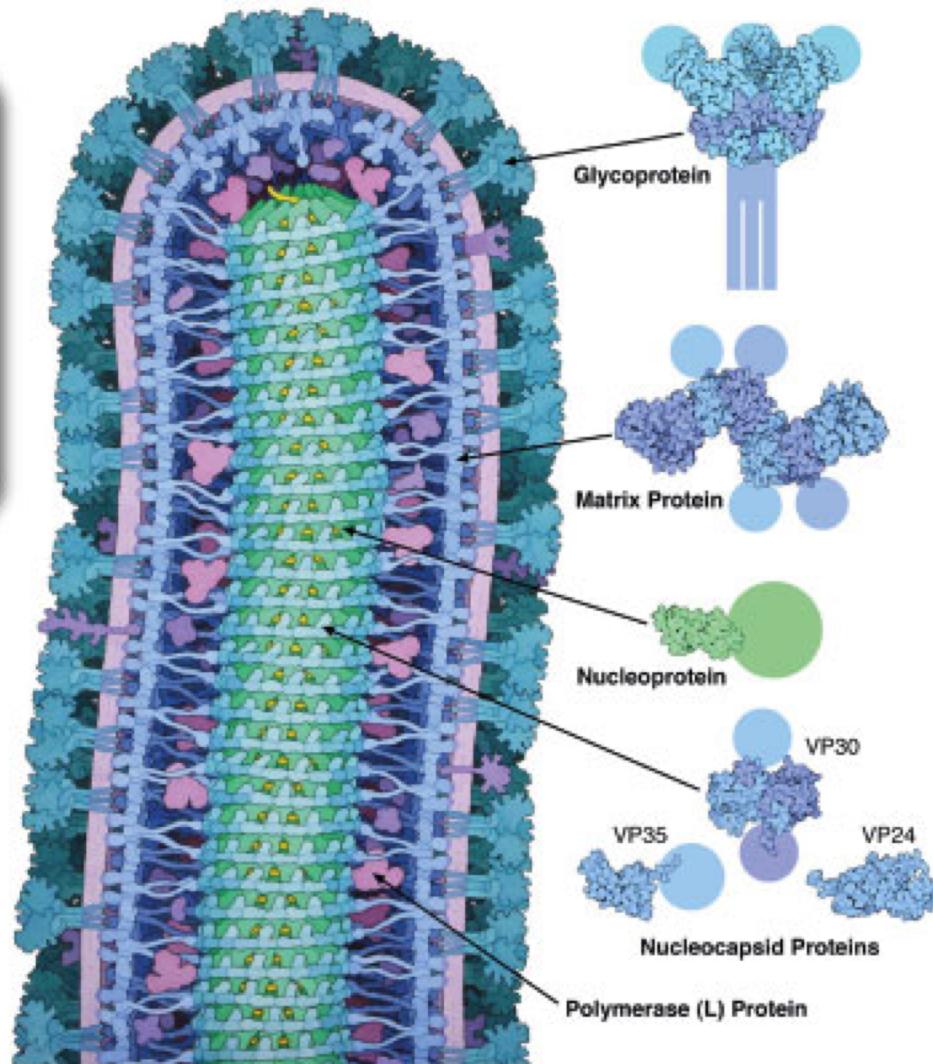
# Composition d'EBOV

## Ebolavirus Proteins

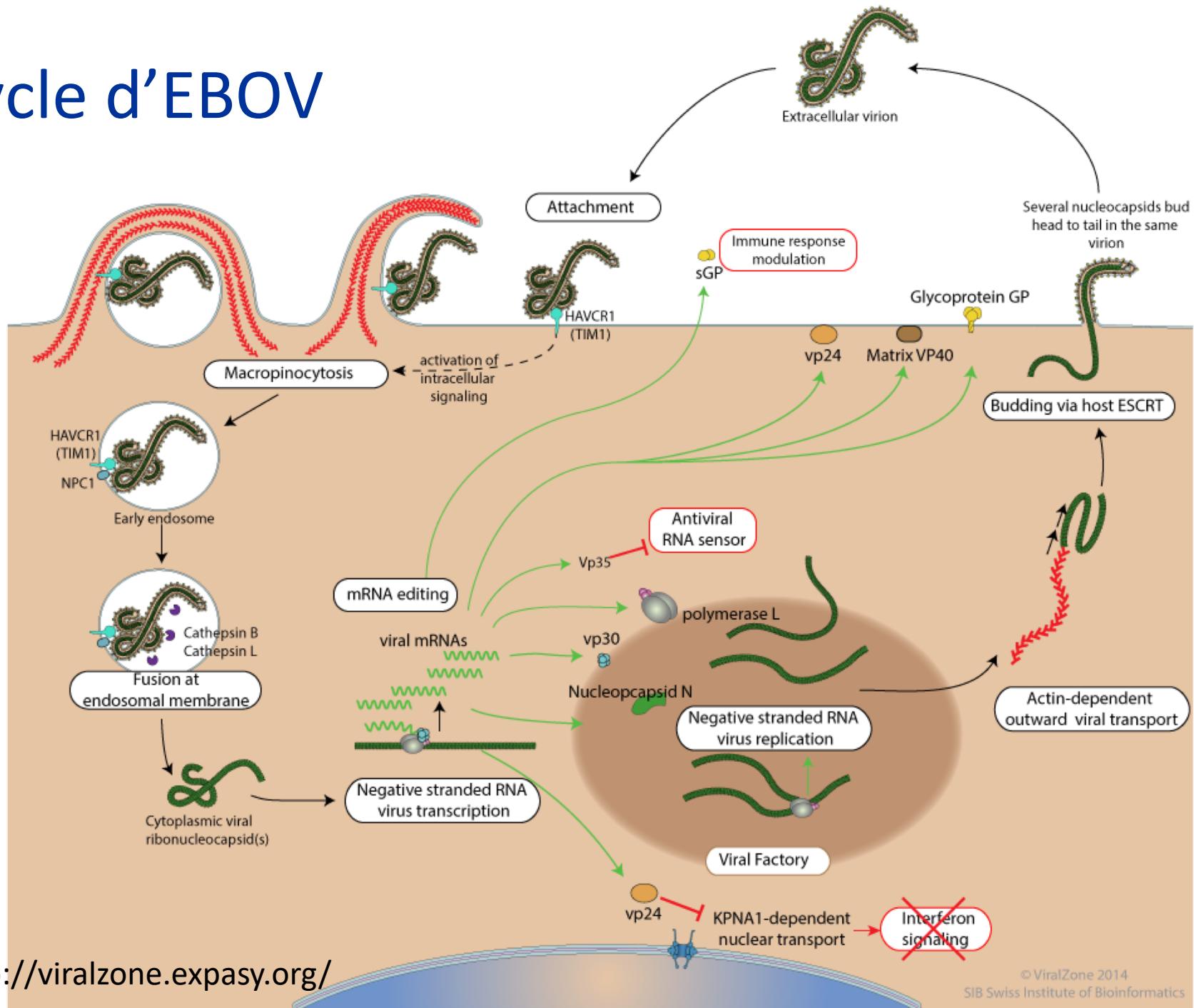
RCSB PDB-101

Credits: David S. Goodsell  
and the RCSB PDB

Image edited to add copyright,  
click to access the original

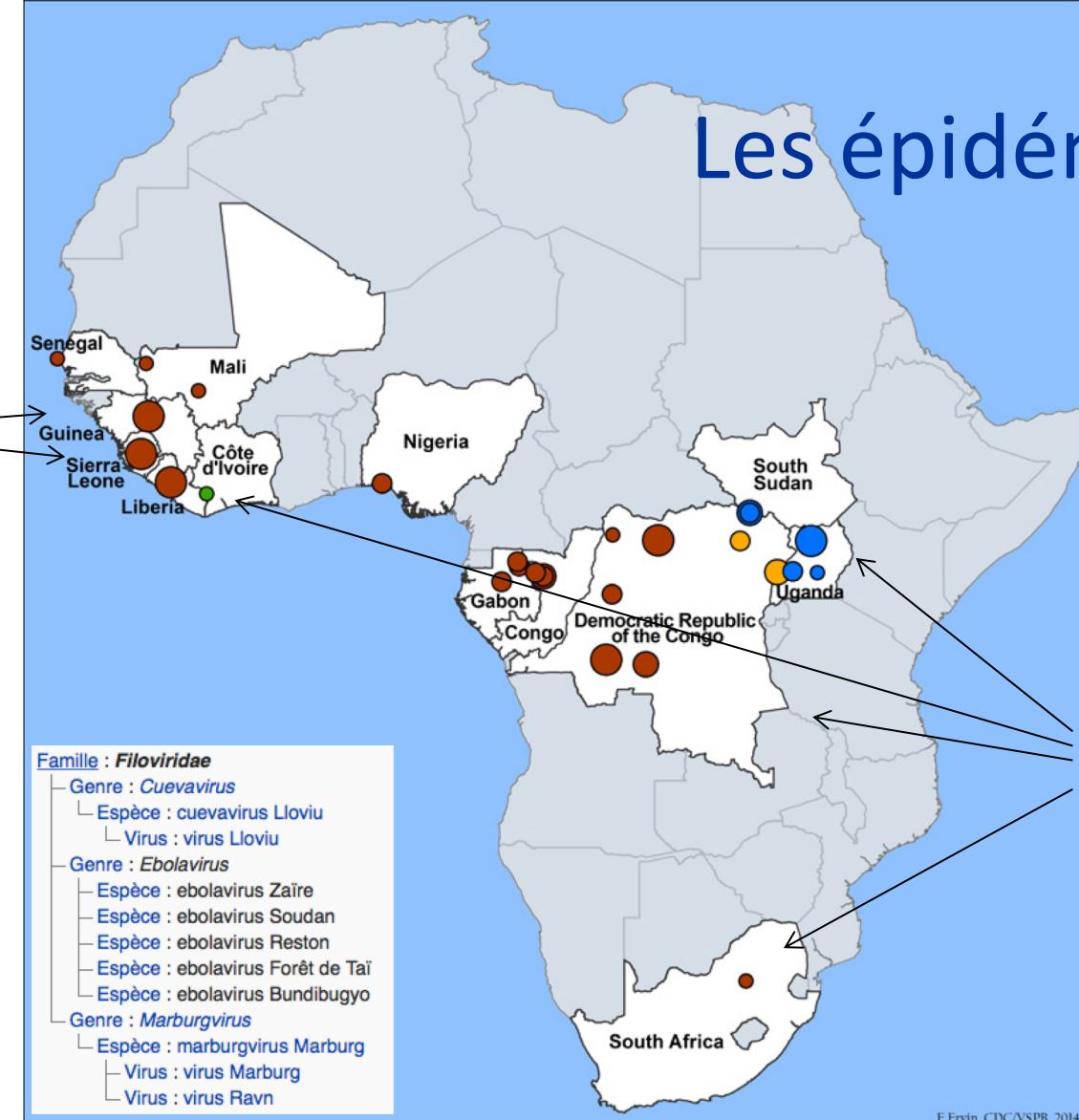


# Cycle d'EBOV

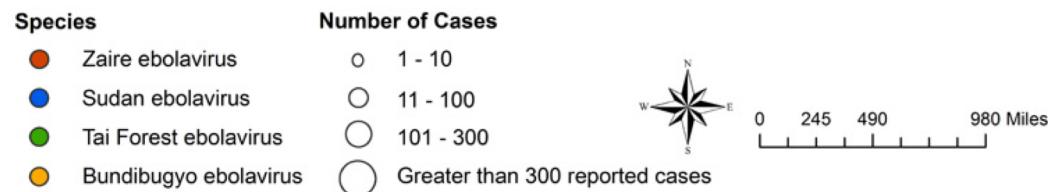


# Les épidémies

Epidémie 2014  
 (Sierra Leone, Libéria,  
 Guinée):  
 28000 cas, 11000 décès



EBOLAVIRUS OUTBREAKS BY SPECIES AND SIZE, 1976 - 2014



# Mini-Plateforme de séquençage EBOV

Déployée en 2015 en Guinée  
Séquenceur MinION



*Sequencing Ebolavirus in Guinea. A researcher prepares samples on the right, while MinIONs plugged into laptops are visible on the left.*

IMAGE COURTESY OF EUROPEAN MOBILE LAB. PHOTOGRAPH BY TOMMY TRENCHARD

Un Séquenceur MinION



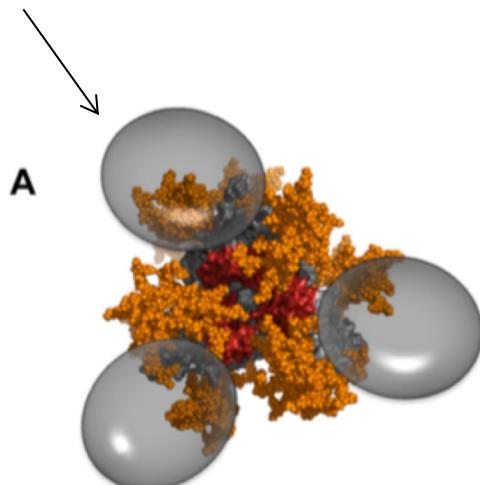
2016: 1693 génomes complets ou partiels séquencés

# Les glycoprotéines GP

- Protéines d'enveloppe
- Permettent l'attachement du virus aux récepteurs de l'hôte
- Sont exprimées en surface de la cellule-hôte
- Induisent une réponse immunitaire
  - Développement de vaccin
  - Comprendre l'évolution/adaptation du virus

# Structure de la GP

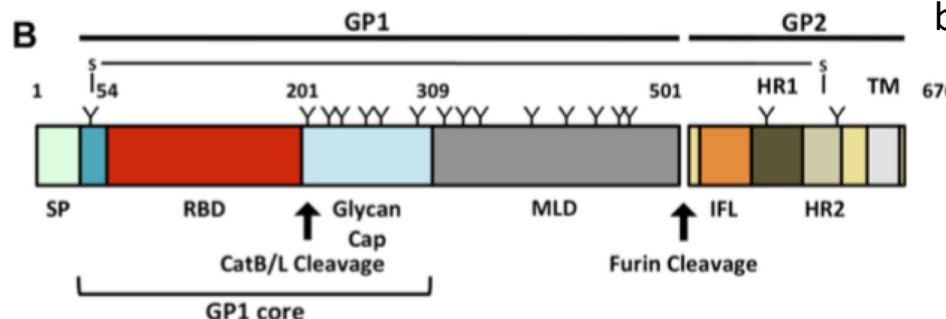
Trimère de GP1/GP2 en surface du virion



receptor-binding domain (RBD)  
très conservé

Glycan cap & mucin-like domain (MLD): très variables

GP2: dom. transmembranaire



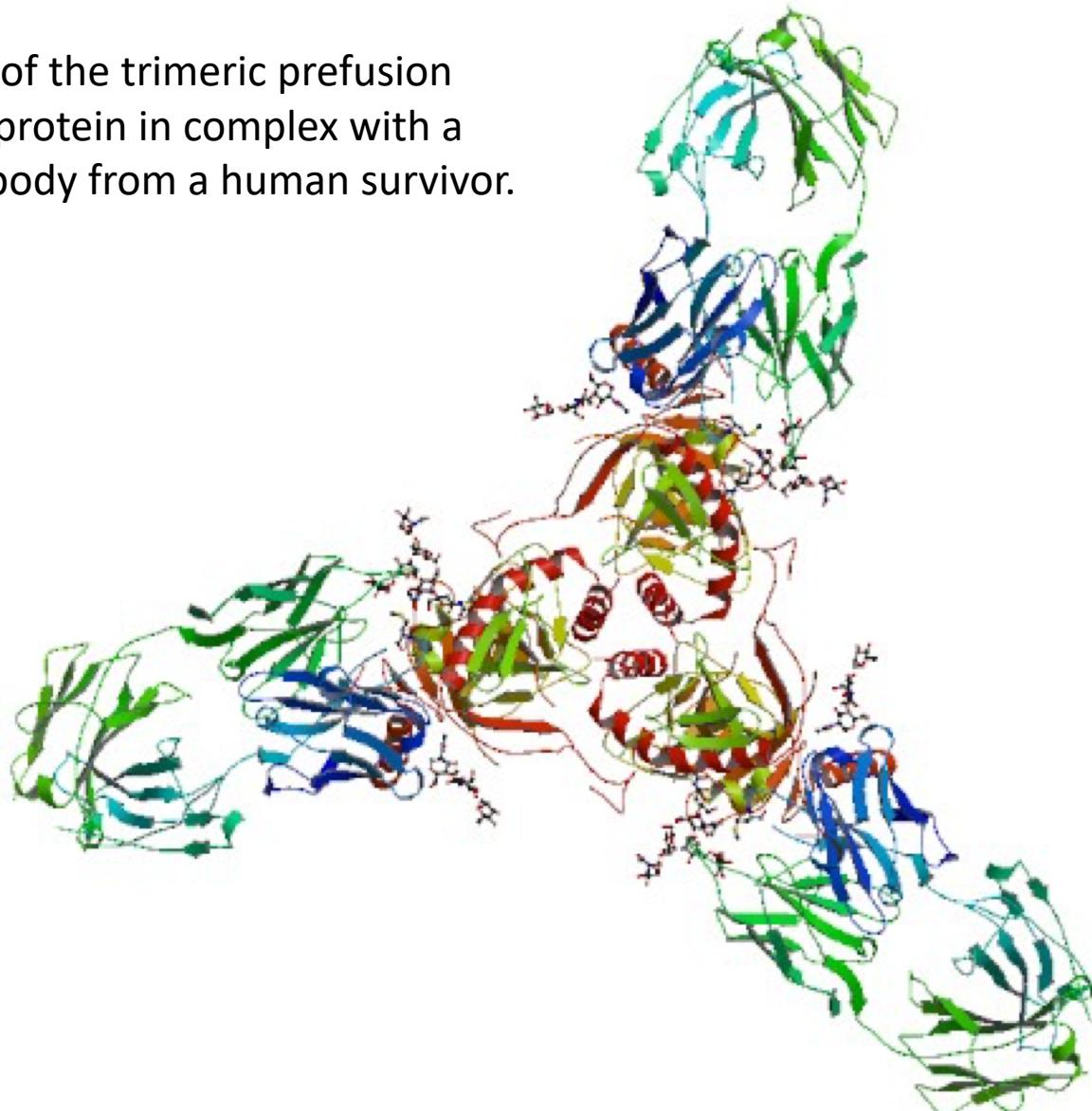
	N40	N204	N228	N238	N257	N268	N296
Ebola	IHNSTLQ	PVNATEDPSS	GTNETETLFEVDNLTYV	QLNETIYTSGKRSNTTGK	KKNLTRK		
Sudan	VTNSTLLE	AVNYTENTSS	GAQHSTTLFKINNNNTFV	QLNDTIHLHQQLSNTTGK	KKNLSEQ		
Tai Forest	VHNNTLQ	PANMTTDPS	GTNTTEFLFQVDHLYV	LLNETIYSDNRRSNTTGK	KKNFTKT		
Bundibugyo	VHNNTLQ	PANMTTDPS	GTNMTNFLFQVDHLYV	QLNETIYTNGRRSNTTGT	KKNFTKT		
Reston	VTNSTLK	PVNTTDDSTS	GGNESNTLFKVSNHTYV	QLNETLRRNNRLNSTGR	KKNFSQQ		

Sites de fixation glycanes

FIG 1 Schematic diagrams of Ebola virus GP. (A) A molecular model of EBOV GP1/2 shown in a top-down view. Complex N-glycans are shown in orange, GP is shown in light gray, RBD is shown in red, and MLD structure that has not been solved is represented as a gray sphere. PDB ID [3CSY](#). (B) Linear model of EBOV GP. The disulfide bond between GP1 and GP2 is indicated, as well as the locations of N-linked glycans (marked with "Ys") in the GP1 and -2 domains, and the known protease cleavage sites are noted. SP, signal peptide; RBD, receptor-binding domain; MLD, mucin-like domain; IFL, internal fusion loop; HR1 and -2, heptad repeats 1 and 2; TM, transmembrane domain. (C) Alignment of predicted N-linked glycan sites within the GP1 core of the five Ebola virus species. N-X-S/T sequons are highlighted with a black background.

3CSY

Crystal structure of the trimeric prefusion  
Ebola virus glycoprotein in complex with a  
neutralizing antibody from a human survivor.  
Lee et al. 2008



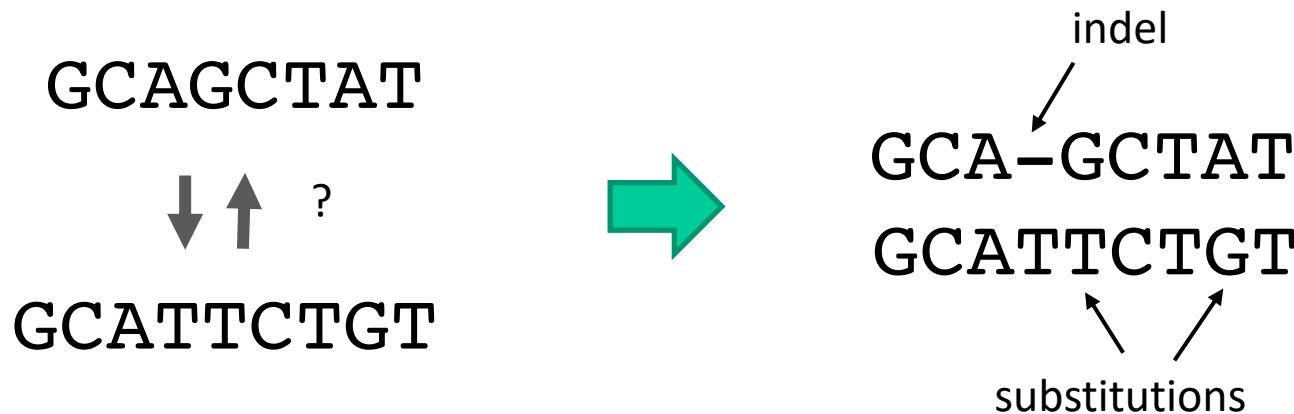
# Questions

- Variation des GP au cours de l'évolution de EBOV
- Positions les plus variables dans la GP?
- Variations ADN vs. variations protéine
- A quoi correspondent ces positions sur la structure tridimensionnelle?

# Notions de comparaison de séquences

# Comparaison de séquences

Quel est le taux de ressemblance entre 2 séquences?



Comparer = trouver l'alignement optimal

# Un alignement entre 2 séquences

	M	N	A	L	S	Q	L	N
N		•						•
A			•					
L				•			•	
M	•							
S					•			
Q						•		
N		•						•
H								

	M	N	A	L	S	Q	L	N
N		•						•
A			•					
L				•			•	
M	•							
S					•			
Q						•		
N		•						•
H								



Trouver le chemin optimal maximisant les identités

# Alignment global

- Algorithme de Needleman & Wunsh
  - Programmation dynamique
  - Trouve l'alignment de score optimal
- Score pour
  - Résidus identiques 
  - Substitution 
  - Indel 

# Matrices de Substitution

- Matrice 4X4 (nt) ou 20x20 (aa) décrivant la distance ou la similitude entre résidus.
- Estiment le coût ou le taux de remplacement d'un résidu par un autre (distance).
- Le choix d'une matrice affecte fortement le résultat de l'analyse. Chaque matrice de score représente implicitement une théorie évolutive donnée

## Matrices DNA

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

Matrice identité

	A	C	G	T
A	3	0	1	0
C	0	3	0	1
G	1	0	3	0
T	0	1	0	3

Matrice transition/transversion

# Protéines: matrice de Dayoff (1979)

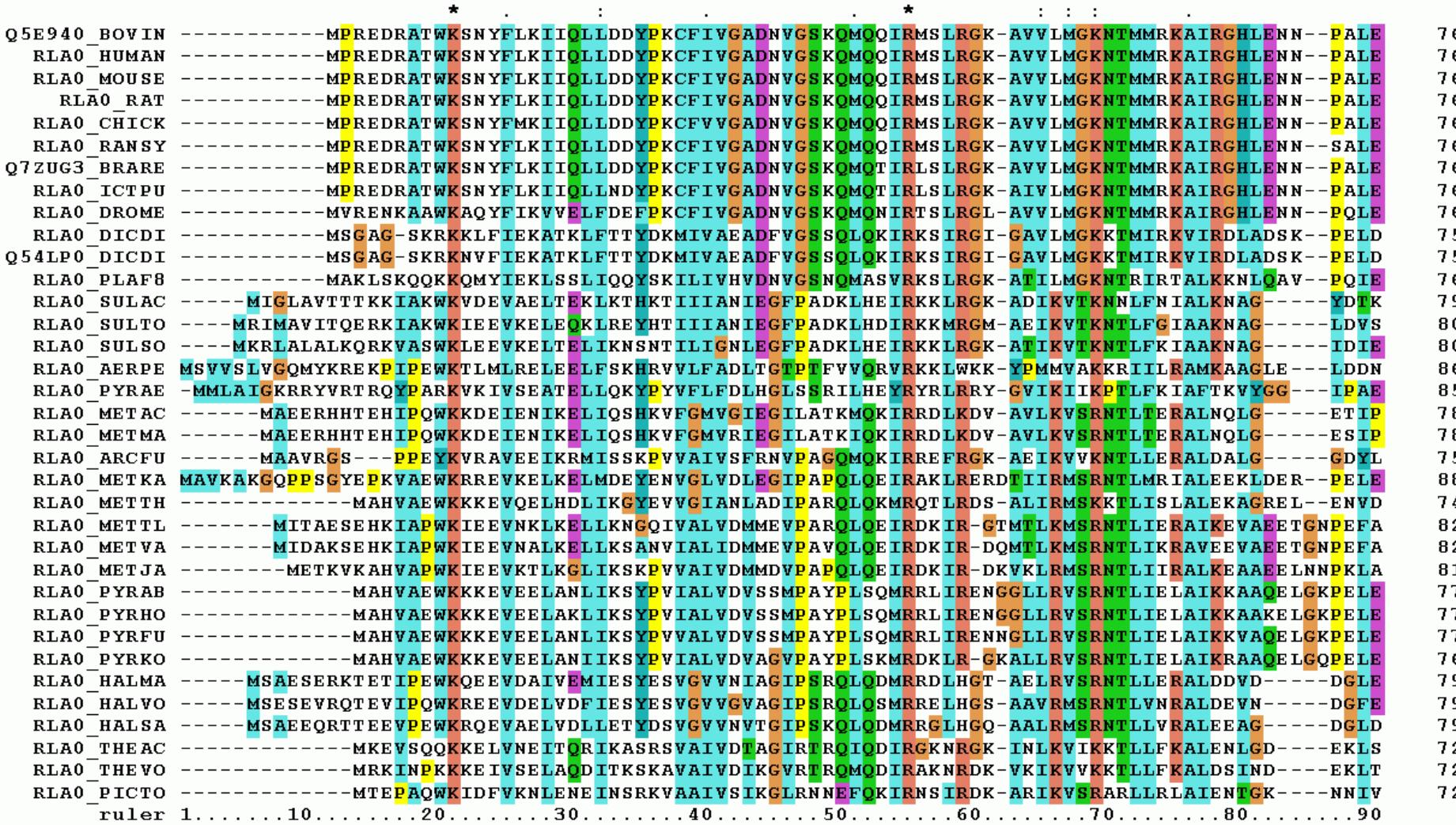
A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
0.4	0.0	-0.4	0.0	0.0	-0.8	0.2	-0.2	-0.2	-0.4	-0.2	0.0	0.2	0.0	-0.4	0.2	0.2	0.0	-1.2	-0.6	0.0	A
0.5	0.9	0.6	0.4	-1.0	0.1	0.3	-0.4	0.1	-0.7	-0.5	0.4	-0.2	0.3	-0.1	0.1	0.0	-0.4	-1.1	-0.6	0.4	B
	2.4	-1.0	-1.0	-0.8	-0.6	-0.6	-0.4	-1.0	-1.2	-1.0	-0.8	-0.6	-1.0	-0.8	0.0	-0.4	-0.4	-1.6	0.0	-1.0	C
	0.8	0.6	-1.2	0.2	0.2	-0.4	0.0	-0.8	-0.6	0.4	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.1	-0.8	0.5	D	
	0.8	-1.0	0.0	0.2	-0.4	0.0	-0.6	-0.4	0.2	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.6	E		
	1.8	-1.0	-0.4	0.2	-1.0	0.4	0.0	-0.8	-1.0	-1.0	-0.8	-0.6	-0.6	-0.2	0.0	1.4	-1.0	F			
	1.0	-0.4	-0.6	-0.4	-0.8	-0.6	0.0	-0.2	-0.2	-0.6	0.2	0.0	-0.2	-1.4	-1.0	-0.1	G				
	1.2	-0.4	0.0	-0.4	-0.4	0.4	0.0	0.6	0.4	-0.2	-0.2	-0.4	-0.6	0.0	-0.4	H					
	1.0	-0.4	0.4	0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.2	0.0	0.8	-1.0	-0.2	-0.4	I					
	1.0	-0.6	0.0	0.2	-0.2	0.2	0.6	0.0	0.0	-0.4	-0.6	-0.6	-0.8	0.1	K						
	1.2	0.8	-0.6	-0.6	-0.4	-0.6	-0.6	-0.4	0.4	-0.4	-0.4	-0.4	-0.2	-0.5	L						
	1.2	-0.4	-0.4	-0.2	0.0	-0.4	-0.2	0.4	-0.4	-0.8	-0.4	-0.4	-0.2	-0.3	M						
	0.4	-0.2	0.2	0.0	0.2	0.0	-0.4	-0.8	-0.4	-0.4	-0.4	-0.8	-0.4	0.2	N						
	1.2	0.0	0.0	0.2	0.0	-0.2	-1.2	-1.0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	P						
	0.8	0.2	-0.2	-0.2	-0.4	-1.0	-0.8	-0.6	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	Q						
	1.2	0.0	-0.2	-0.4	0.4	-0.8	-0.6	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	0.6	R						
	0.4	0.2	-0.2	-0.4	-0.6	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	S						
	0.6	0.0	-1.0	-0.6	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	T						
	0.8	-1.2	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	V						
	3.4	0.0	-1.2	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	-0.8	W						
	2.0	-0.8	Y													Z					
																	0.6	Z			

Matrice dérivée des substitutions observées dans les régions bien conservées des protéines

Chaque case représente la probabilité de voir ces deux résidus remplacés l'un par l'autre dans un alignement. (matrice Iod-score, de "log-odds" ou "log des chances").

$$S = \log(F_{ij} / (F_i \times F_j))$$

# Alignement multiple



# Les usages des alignements multiples

- Phylogénie moléculaire
  - via calcul de distance
- Prédiction 3D
  - Par homologie avec autres protéines de repliement 3D connu
- Inférence fonctionnelle
  - Positions conservées=fonction conservée

# Un arbre phylogénétique réalisé à partir d'un alignement de séquences de gènes GP35 de Ebola & Marburg V.

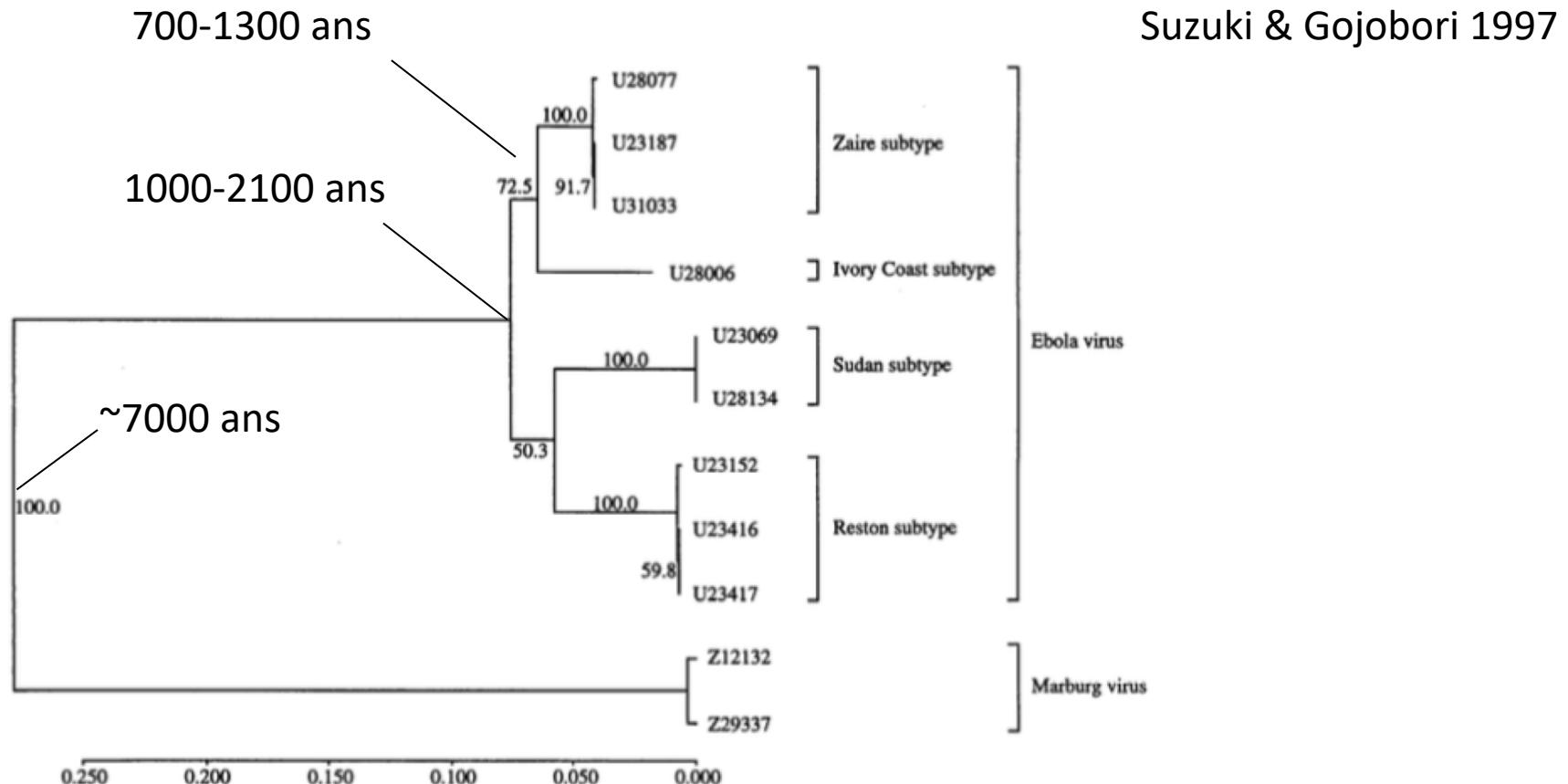


FIG. 1.—Phylogenetic tree constructed for the GP gene of Ebola and Marburg viruses by the neighbor-joining method (Saitou and Nei 1987), with distances for nonsynonymous sites estimated by the method of Nei and Gojobori (1986). The bootstrap probability for each node is also indicated (Felsenstein 1985). When we estimated the substitution rate of Ebola virus, we excluded the sequences of Marburg virus and constructed another phylogenetic tree (data not shown), in which the topology among Ebola virus strains was identical with that of the former one.

# Entropie

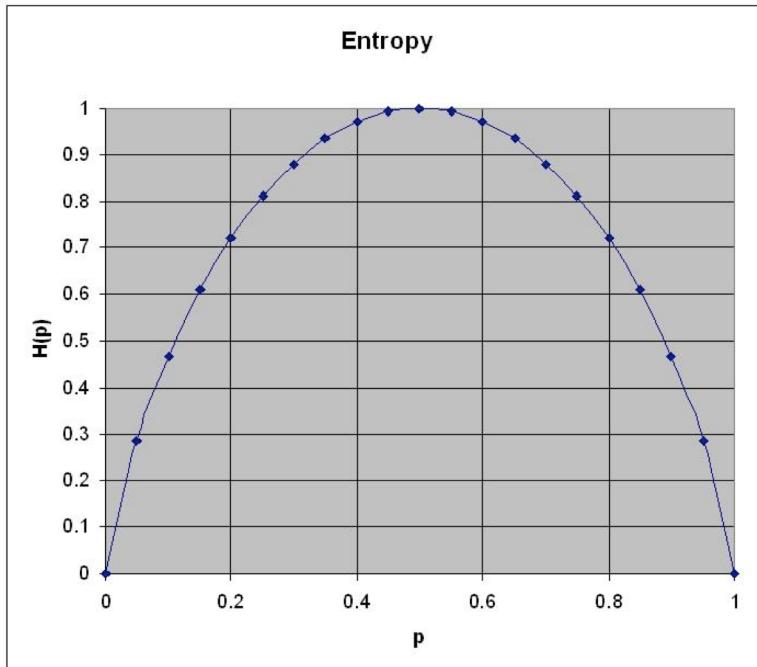
# Entropie de Shannon

- Entropie de Shannon à la position  $i$ :

$$H_i = - \sum_{a=A,T,G,C} f_{a,i} \log_2(f_{a,i})$$

$f_{a,i}$  : fréquence lettre  $a$  à la position  $i$ .

$f \cdot \log(f)$  tend vers 0 quand  $f$  tend vers 0 ou 1

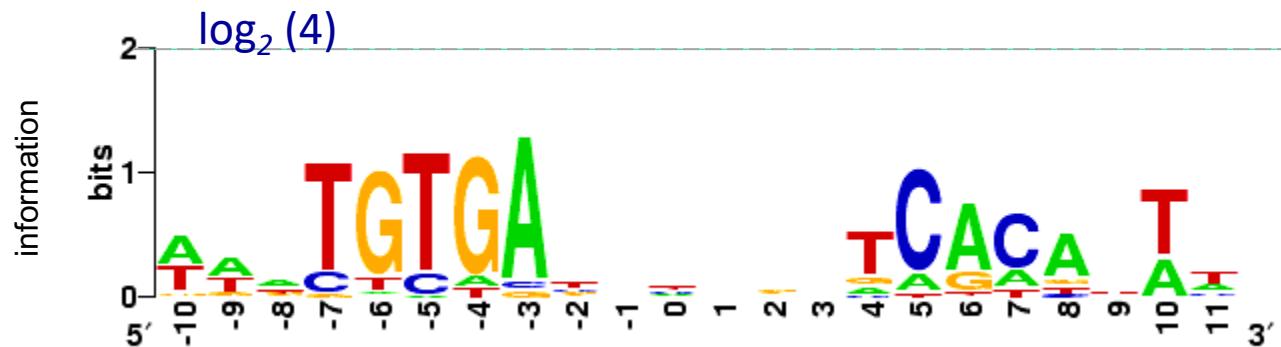


Variation de l'entropie en fonction de  $p$  pour un système à deux états

# Entropie et contenu en information

- Le contenu en information est proportionnel à:

$$\log_2 (4) - H_i \quad = \text{information (pour ADN: } n=4)$$



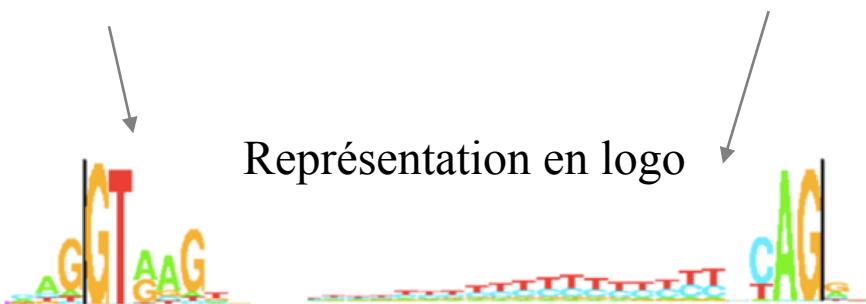
Hauteur des lettres =  $f * \text{information}$

# Sequence logos

(Schneider TD, Stephens RM. NAR. 1990)

Sites d'épissage

Représentation en fréquence:



*Beaucoup mieux  
qu'une fréquence!  
Fait ressortir  
régions  
conservées/  
variables*

The DNA-binding helix-turn-helix motif of the CAP family

