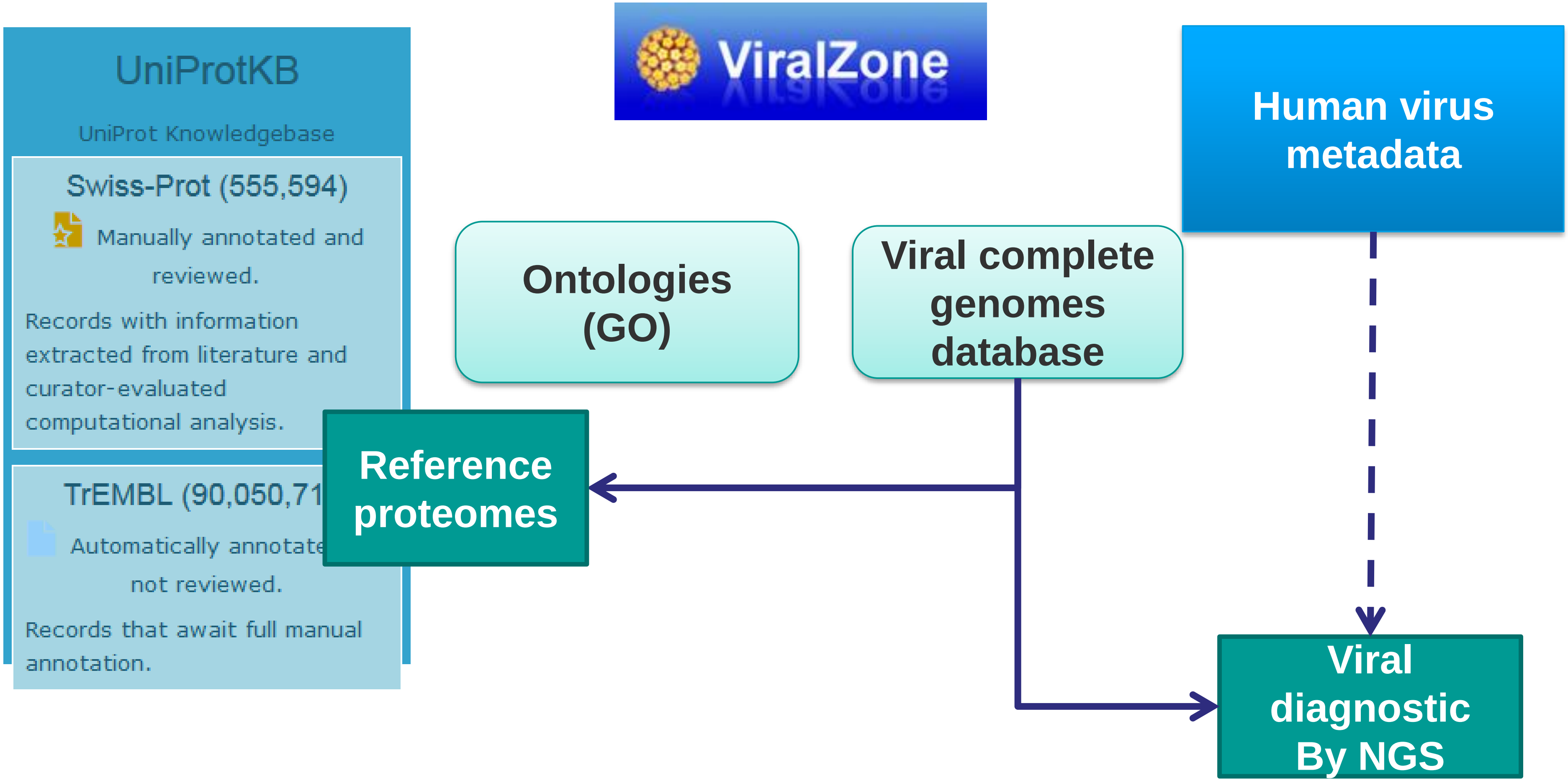
On the importance of curated databases

SIB Swiss Institute of Bioinformatics

Geneva, Switzerland

Digitalisation of virus knowledge

SIB, SwissProt group



UniProtKB

UniProt Knowledgebase

Swiss-Prot (555,594)

Manually annotated and reviewed.

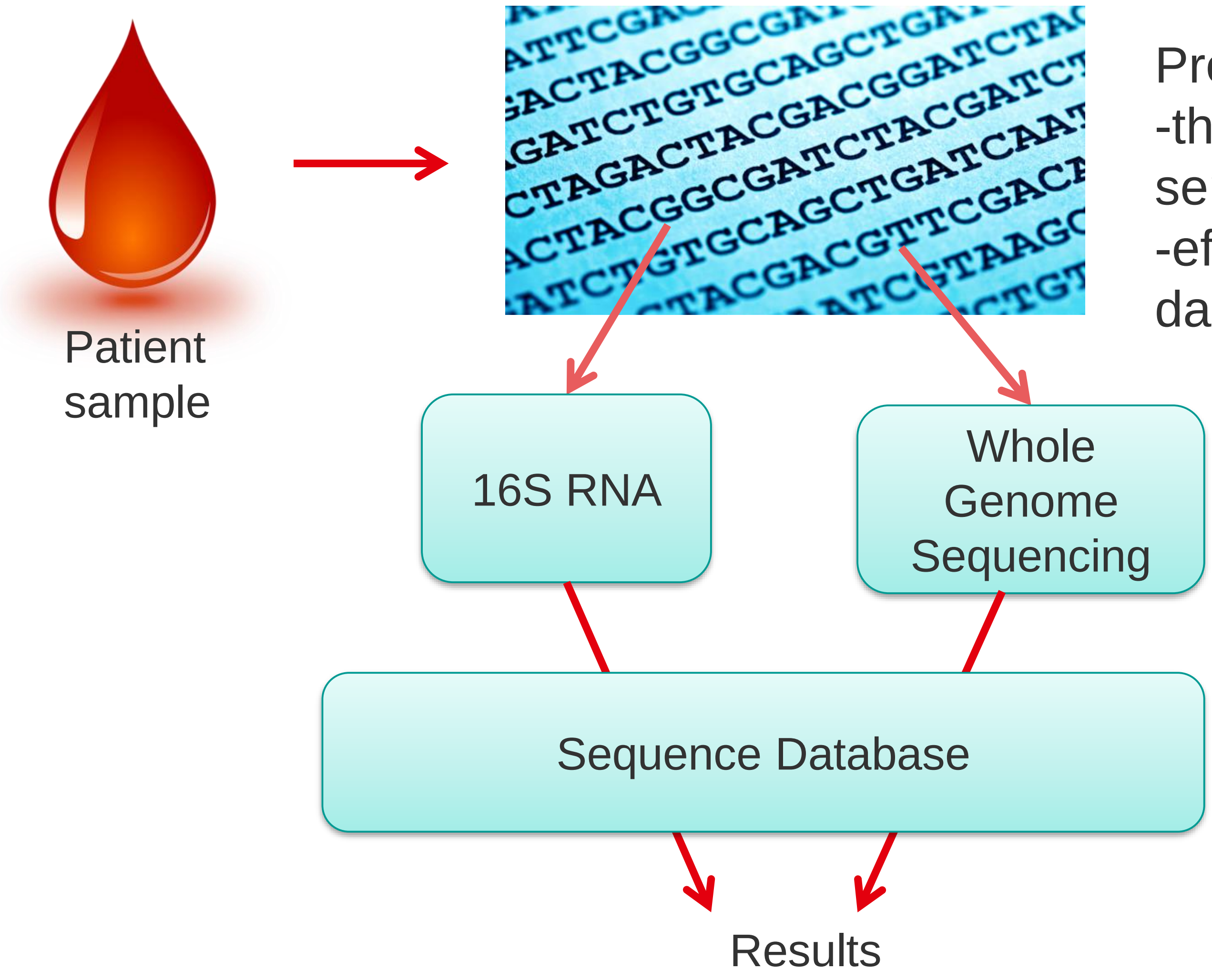
Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (90,050,71

Automatically annotate not reviewed.

Records that await full manual annotation.

Detection by next generation sequencing



Patient sample

16S RNA

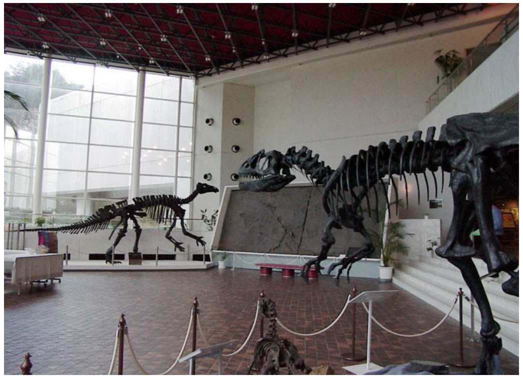
Whole Genome Sequencing

Sequence Database

Results

Present challenge:
-the analysis step (speed, sensivity)
-efficient references databases.

GenBank is a repository database

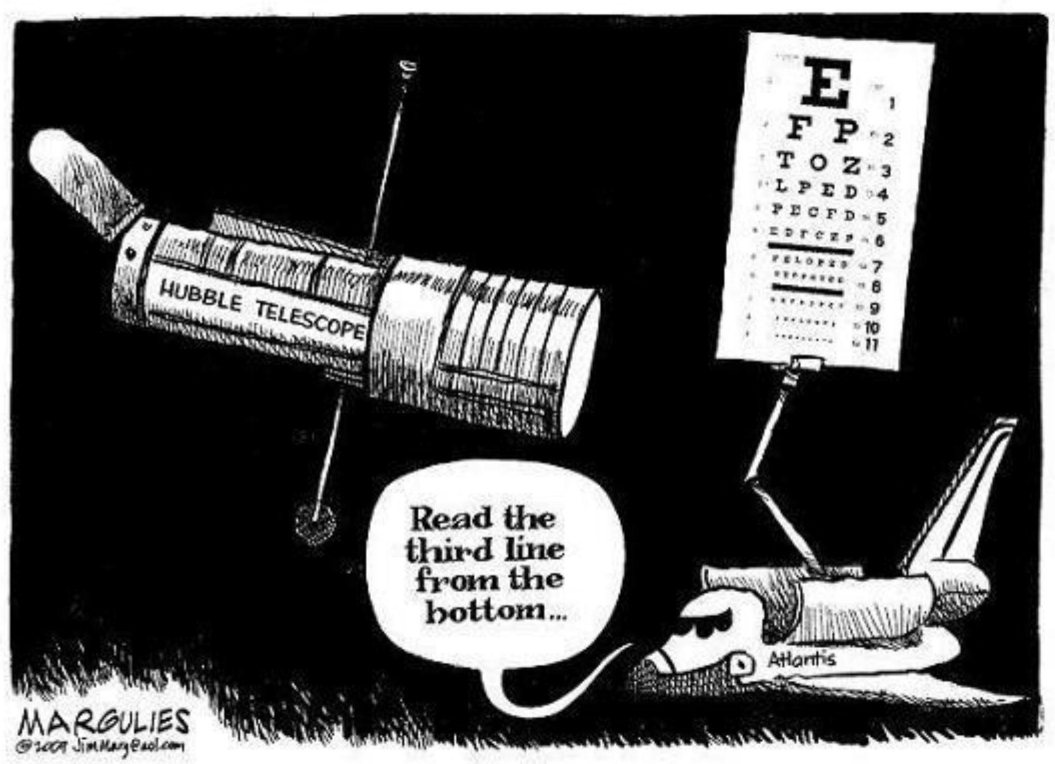


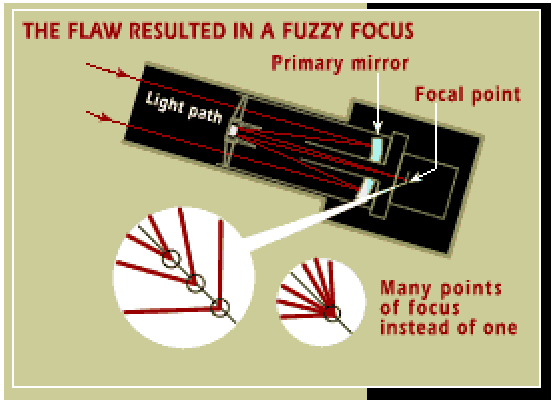
- Unexpected information you can find in *INSDC*=Genbank/EMBL/DDBJ:

```
FT   source        1..124
FT                 /db_xref="taxon:4097"
FT                 /organelle="plastid:chloroplast"
FT                 /organism="Nicotiana tabacum"
FT                 /isolate="Cuban cahibo cigar, gift from
FT                                        President Fidel Castro"
```

```
FT   CDS           complement(45959..47332)
FT                 /db_xref="SPTREMBL:Q9UZ71"
FT                 /note="PAB2386"
FT                 /transl_table=11
FT                 /product="4-AMINOBUTYRATE qui se dilate AMINOTRANSFERASE
FT                 (EC 2.6.1.19)"
FT                 /protein_id="CAB50188.1"
FT                 /translation="MDYPRIVVNPPGPKAKELIEREKRVLSTGIGVKLFPLVPKRGFGP
FT                 FIEDVDGNVFIDFLAGAAAASTGYSHPKLVKAVKEQVELIQHSMIGYTHSERAIRVAEK
FT                 LVKISPIKNSKVLFGLSGSDAVDMAIKVSKFSTRRPWILAFIGAYHGQTLGATSVASFQ
FT                 VSQKRGYSPLMPNVFWVPYPNPYRNPWGINGYEEPQELVNRVVEYLEDYVFSHVVPPDE
FT                 VAAFFAEPIQGDAGIVVPPENFFKELKKLLDEHGILLVMDEVQTGIGRTGKWFASEWFE
FT                 VKPDMIIFGKGVASGMGLSGVIGREDIMDITSGSALLTPAANPVISAAADATLEIIEEE
FT                 NLLKNAIEVGSFIMKRLNELKEQFDIIGDVRGKGLMIGVEIVKENGRPDPEMTGKICWR
FT                 AFELGLILPSYGMFGNVIRITPPLVLTKEVAEKGLEIIEKAIKDAIAGKVERKVVTWH"
```

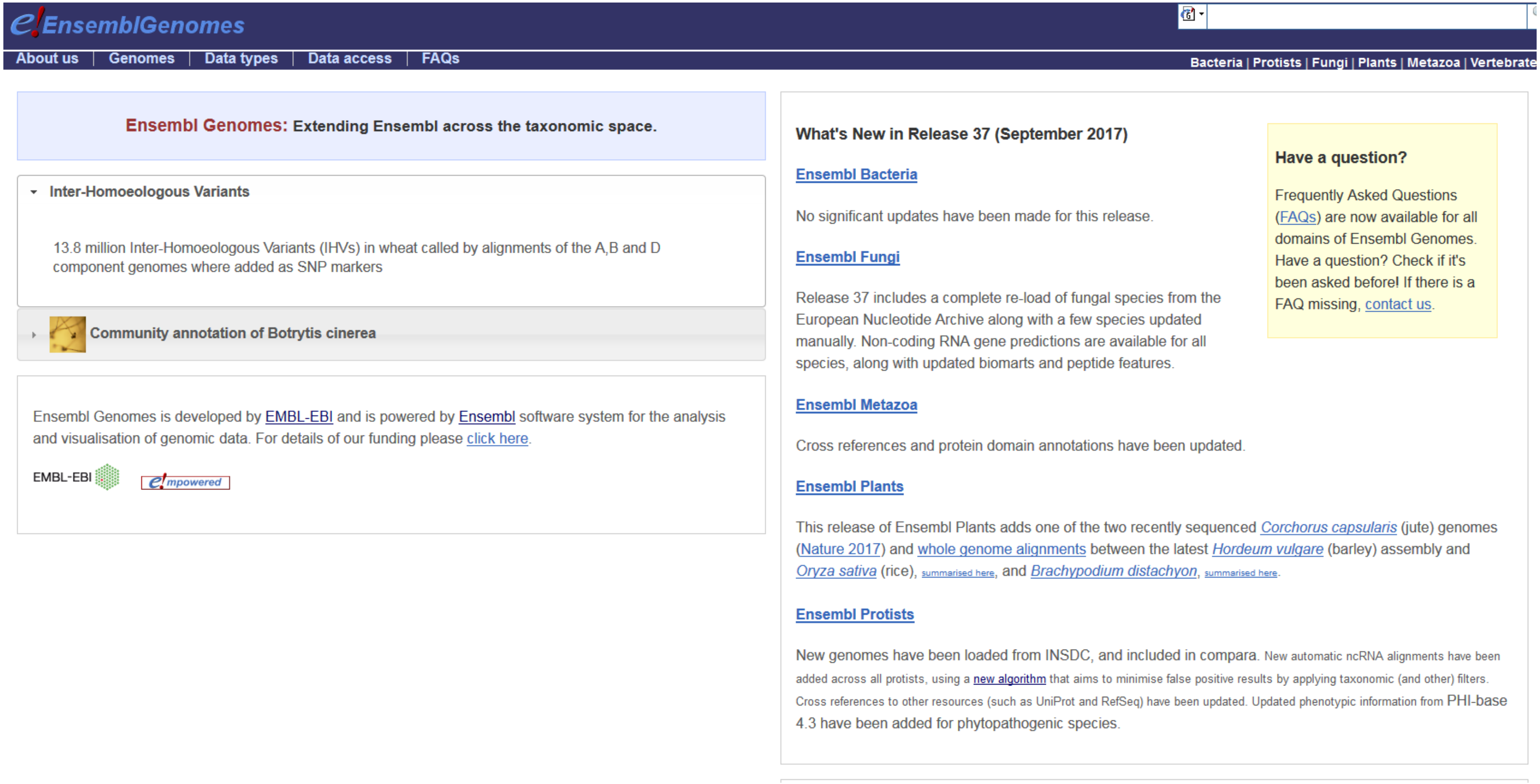
The importance of database




THE FLAW RESULTED IN A FUZZY FOCUS
Primary mirror
Focal point
Light path
Many points of focus instead of one

Whatever the quality of the samples, sequencing and bioinformatics,
a flawed database can blur the results

There was no database for virus complete genomes


EnsemblGenomes
About us | Genomes | Data types | Data access | FAQs
Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrate

Ensembl Genomes: Extending Ensembl across the taxonomic space.

Inter-Homoeologous Variants

13.8 million Inter-Homoeologous Variants (IHVs) in wheat called by alignments of the A,B and D component genomes where added as SNP markers

Community annotation of Botrytis cinerea

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please click here.

EMBL-EBI e!mpowered

What's New in Release 37 (September 2017)

Ensembl Bacteria

No significant updates have been made for this release.

Ensembl Fungi

Release 37 includes a complete re-load of fungal species from the European Nucleotide Archive along with a few species updated manually. Non-coding RNA gene predictions are available for all species, along with updated biomarts and peptide features.

Ensembl Metazoa

Cross references and protein domain annotations have been updated.

Ensembl Plants

This release of Ensembl Plants adds one of the two recently sequenced *Corchorus capsularis* (jute) genomes (Nature 2017) and whole genome alignments between the latest *Hordeum vulgare* (barley) assembly and *Oryza sativa* (rice), summarised here, and *Brachypodium distachyon*, summarised here.

Ensembl Protists

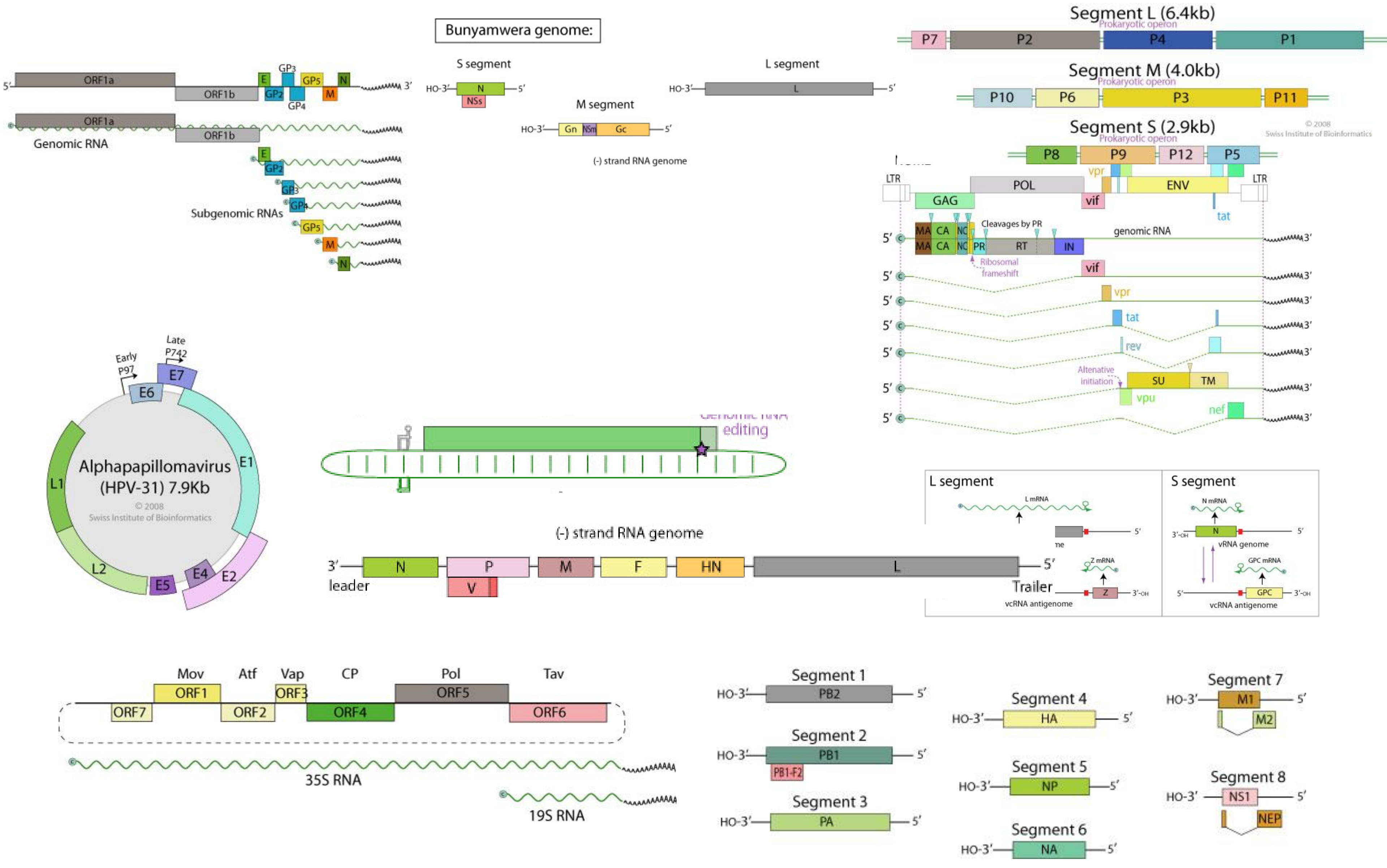
New genomes have been loaded from INSDC, and included in compara. New automatic ncRNA alignments have been added across all protists, using a new algorithm that aims to minimise false positive results by applying taxonomic (and other) filters. Cross references to other resources (such as UniProt and RefSeq) have been updated. Updated phenotypic information from PHI-base 4.3 have been added for phytopathogenic species.

Have a question?

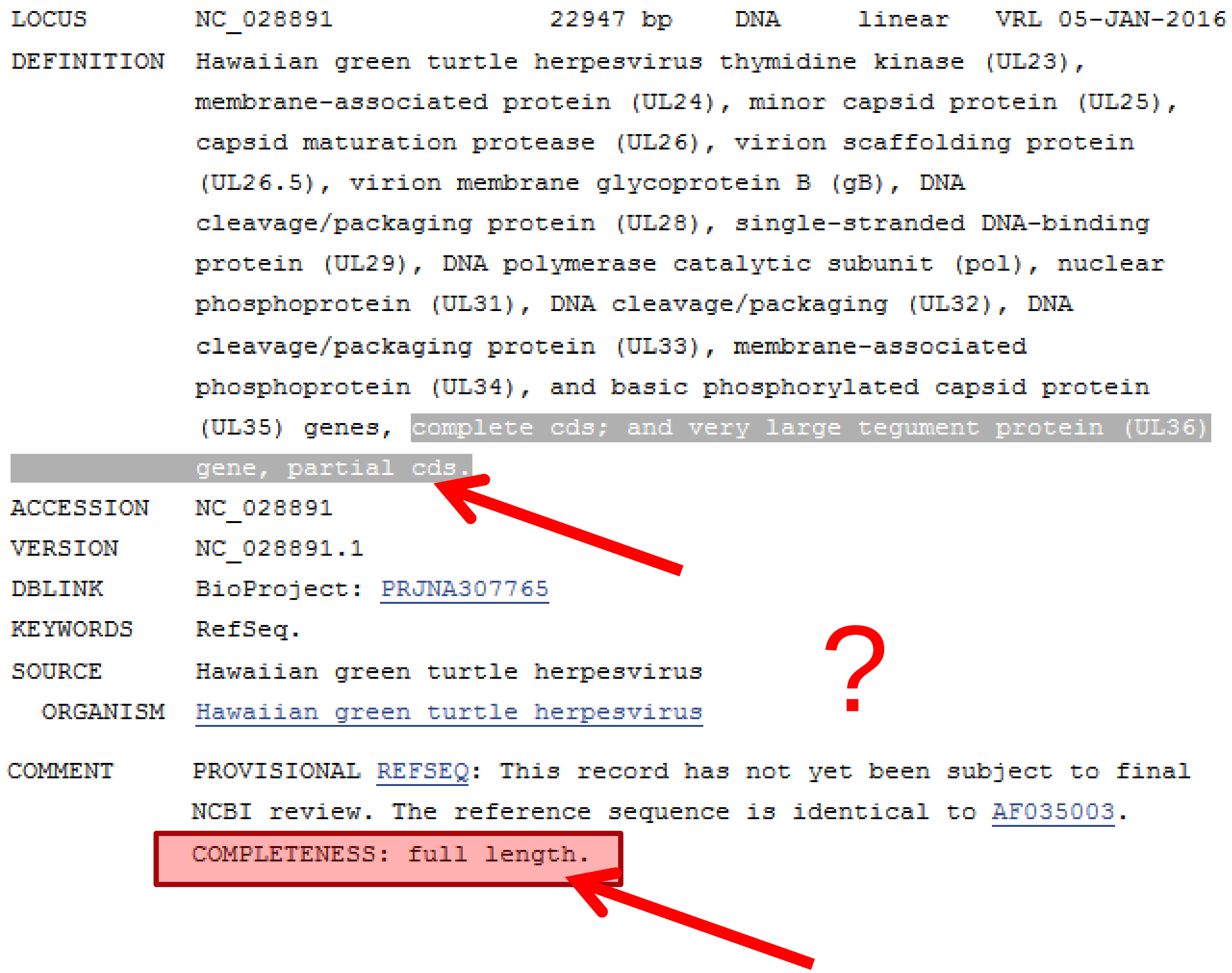
Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing, contact us.

Virus??

Viral genomes



Avoid RefSeq to assess full length criteria

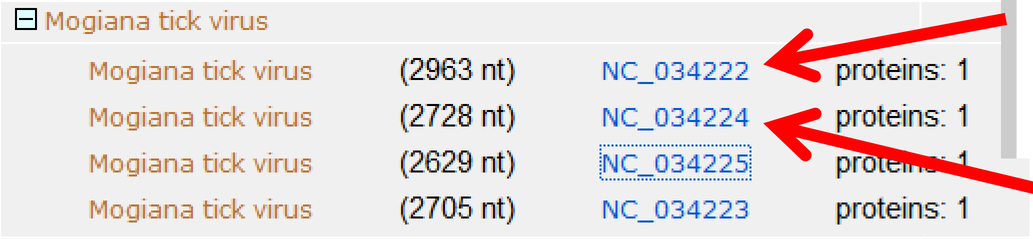


```
LOCUS       NC_028891              22947 bp    DNA     linear   VRL 05-JAN-2016
DEFINITION  Hawaiian green turtle herpesvirus thymidine kinase (UL23),
            membrane-associated protein (UL24), minor capsid protein (UL25),
            capsid maturation protease (UL26), virion scaffolding protein
            (UL26.5), virion membrane glycoprotein B (gB), DNA
            cleavage/packaging protein (UL28), single-stranded DNA-binding
            protein (UL29), DNA polymerase catalytic subunit (pol), nuclear
            phosphoprotein (UL31), DNA cleavage/packaging (UL32), DNA
            cleavage/packaging protein (UL33), membrane-associated
            phosphoprotein (UL34), and basic phosphorylated capsid protein
            (UL35) genes, complete cds; and very large tegument protein (UL36)
            gene, partial cds.
ACCESSION   NC_028891
VERSION     NC_028891.1
DBLINK      BioProject: PRJNA307765
KEYWORDS    RefSeq.
SOURCE      Hawaiian green turtle herpesvirus
  ORGANISM  Hawaiian green turtle herpesvirus

COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence is identical to AF035003.
            COMPLETENESS: full length.
```

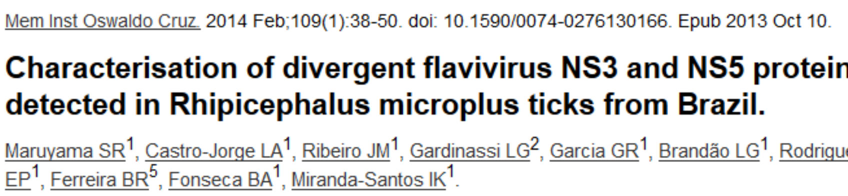
?

Refseq provisional



⊟ Mogiana tick virus

Mogiana tick virus	(2963 nt)	NC_034222	proteins: 1
Mogiana tick virus	(2728 nt)	NC_034224	proteins: 1
Mogiana tick virus	(2629 nt)	NC_034225	proteins: 1
Mogiana tick virus	(2705 nt)	NC_034223	proteins: 1

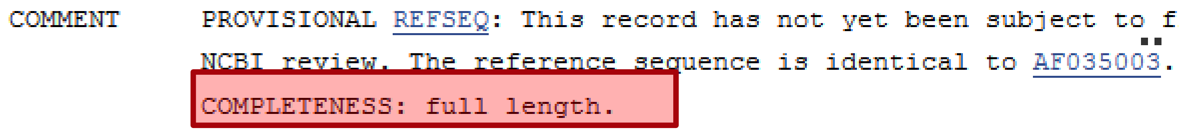


Mem Inst Oswaldo Cruz. 2014 Feb;109(1):38-50. doi: 10.1590/0074-0276130166. Epub 2013 Oct 10.

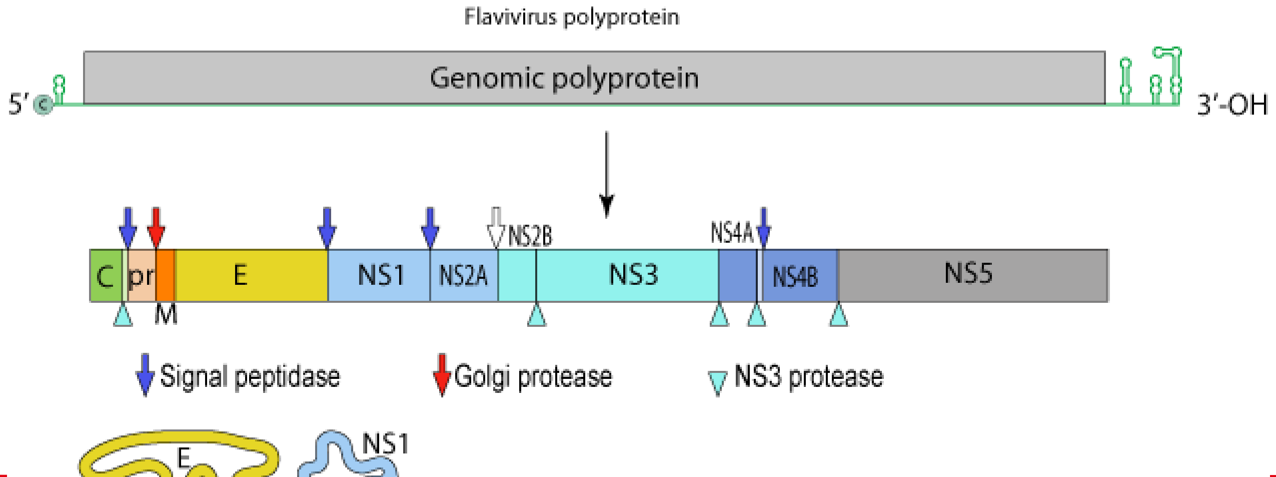
Characterisation of divergent flavivirus NS3 and NS5 protein detected in Rhipicephalus microplus ticks from Brazil.

Maruyama SR[1], Castro-Jorge LA[1], Ribeiro JM[1], Gardinassi LG[2], Garcia GR[1], Brandão LG[1], Rodrigu EP[1], Ferreira BR[5], Fonseca BA[1], Miranda-Santos IK[1].

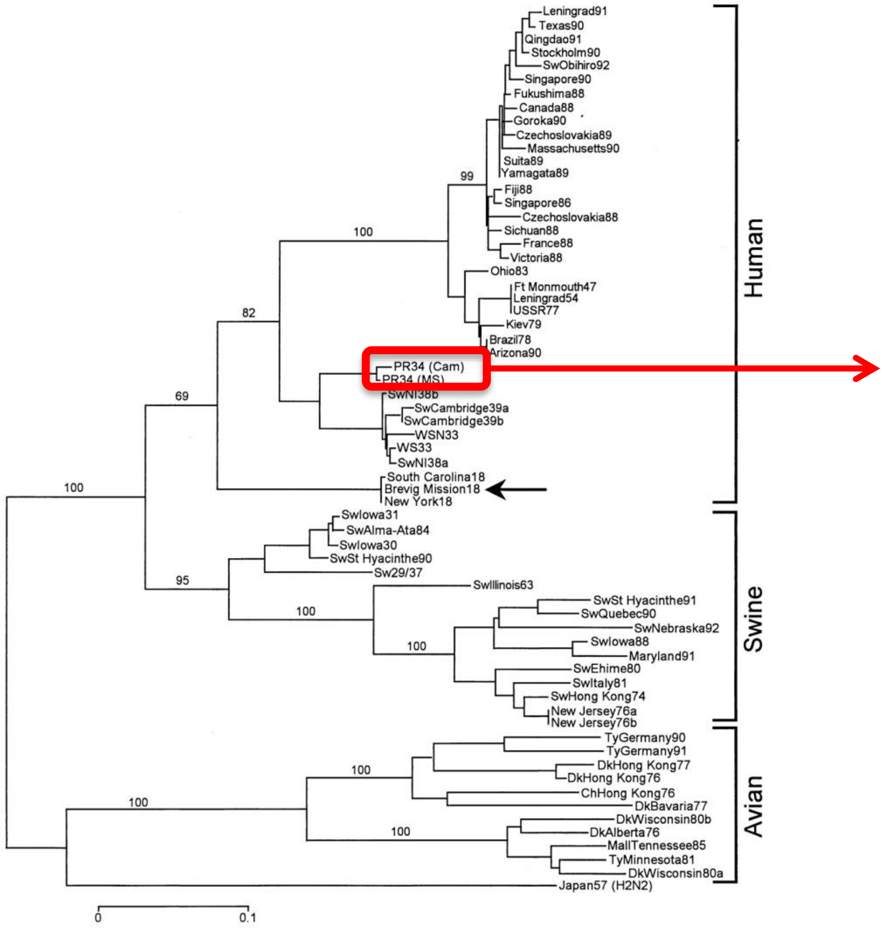
VP2-3 gene, complete cds,



```
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence is identical to AF035003.
            COMPLETENESS: full length.
```

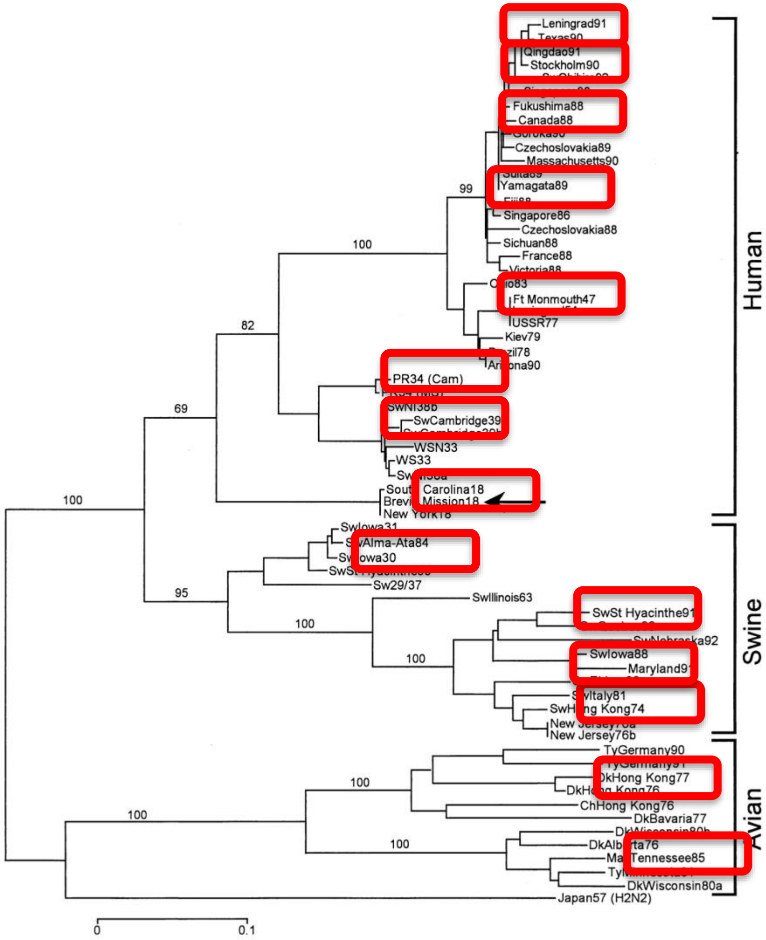


Refseq is for annotation



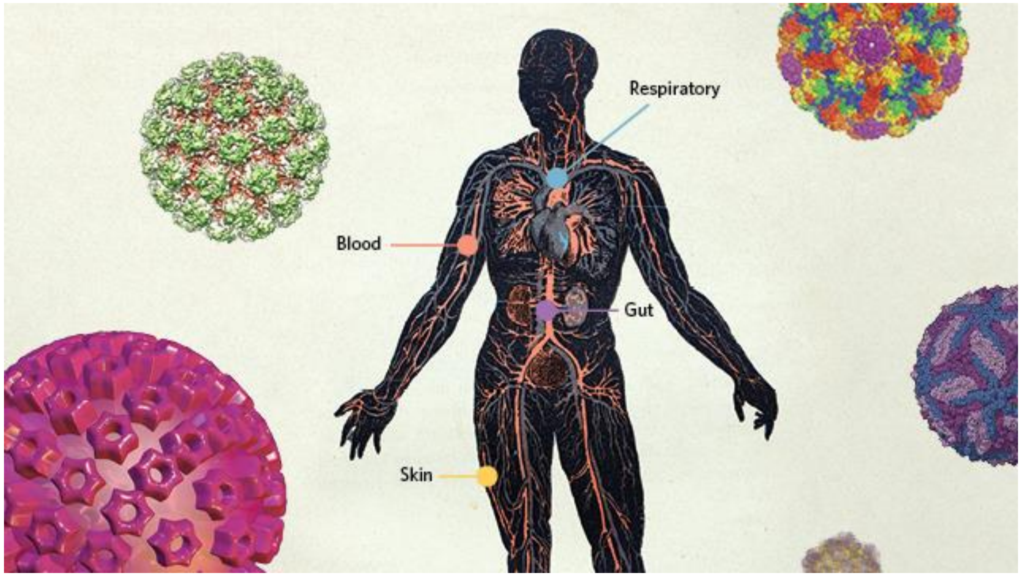
Reference
Puerto rico 1934

Representing the diversity calls for many references

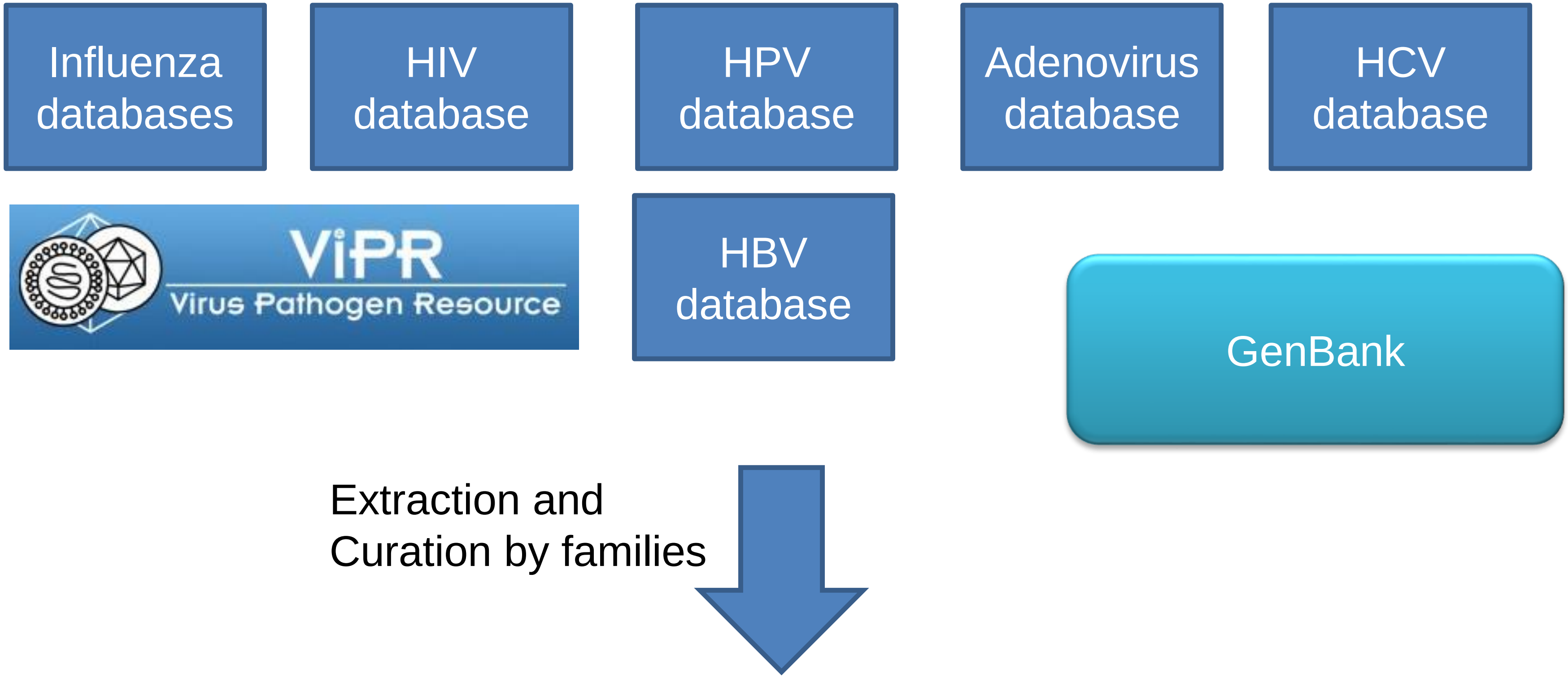


Toward detecting all viruses

- Virology have focused on pathogens
- Recent genomic explorations of human samples have revealed dozens of previously unrecognized viruses.



ViralZone complete genome dataset



Influenza databases

HIV database

HPV database

Adenovirus database

HCV database

ViPR Virus Pathogen Resource

HBV database

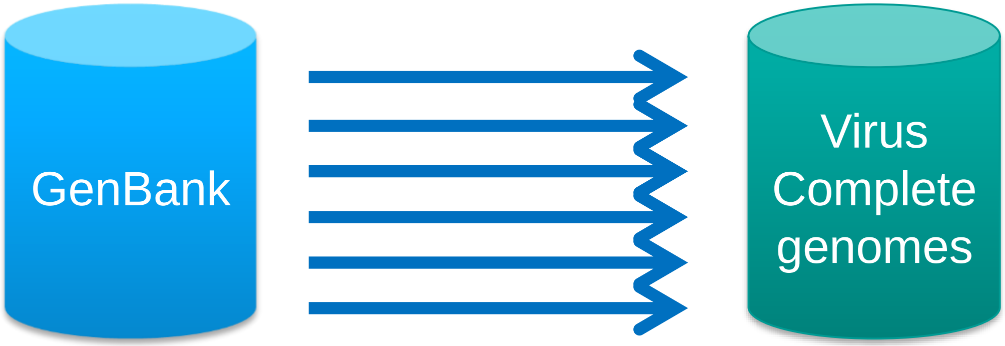
GenBank

Extraction and
Curation by families

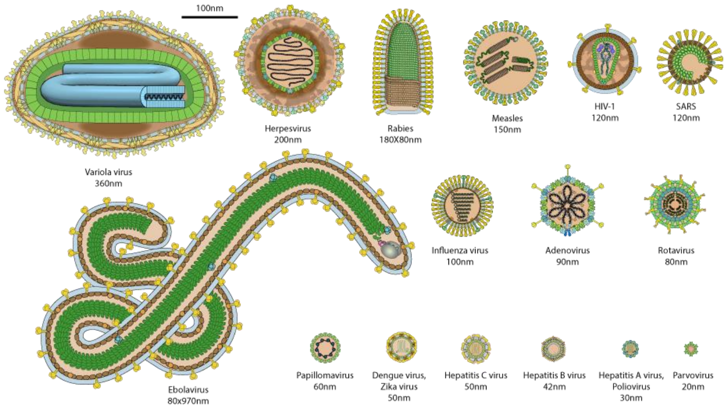
For Eukaryotic viruses: 70,352 complete genomes
317,979 sequences

Toward an automatic detection of virus complete genomes

Indidual criteria
for each 121 virus families

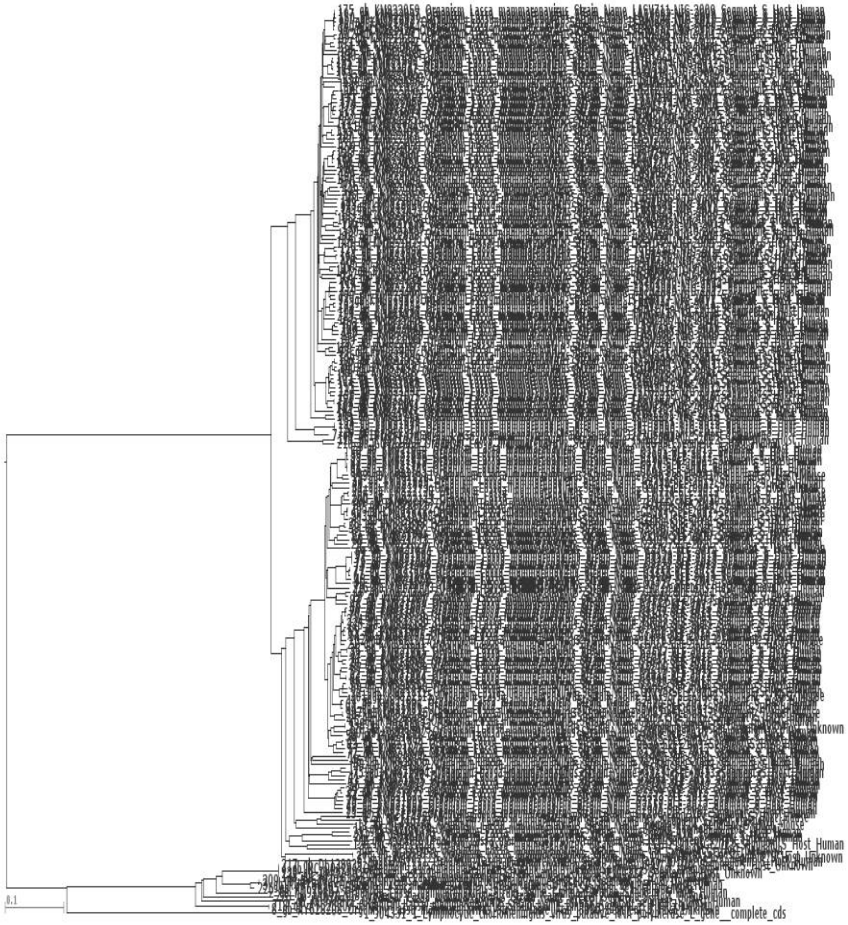


Testing negative selection by families;
-genome length
-CDS all complete
-Number of CDS

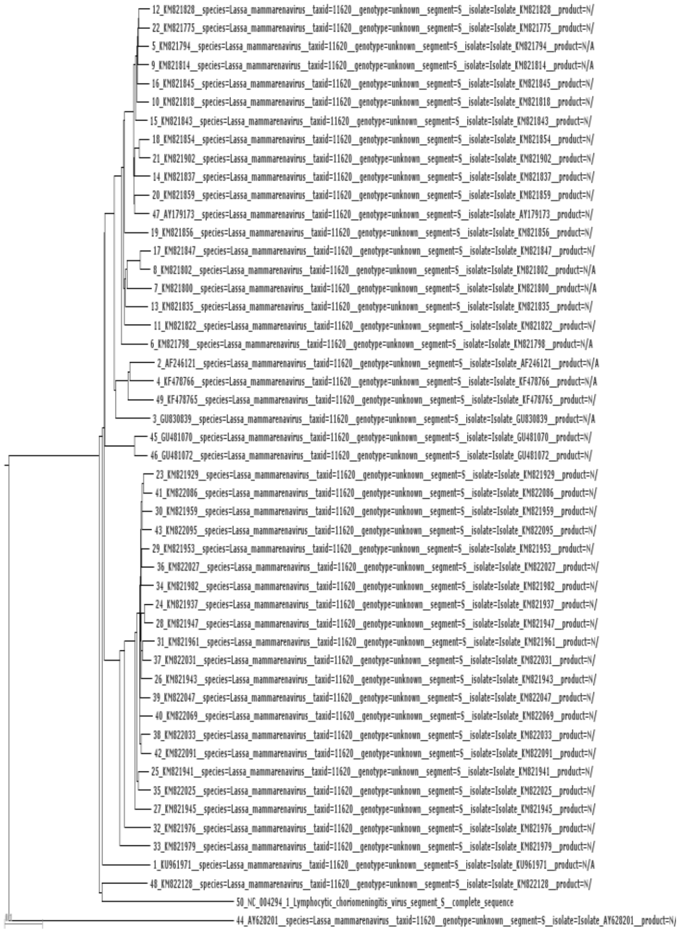


Trimming the branches

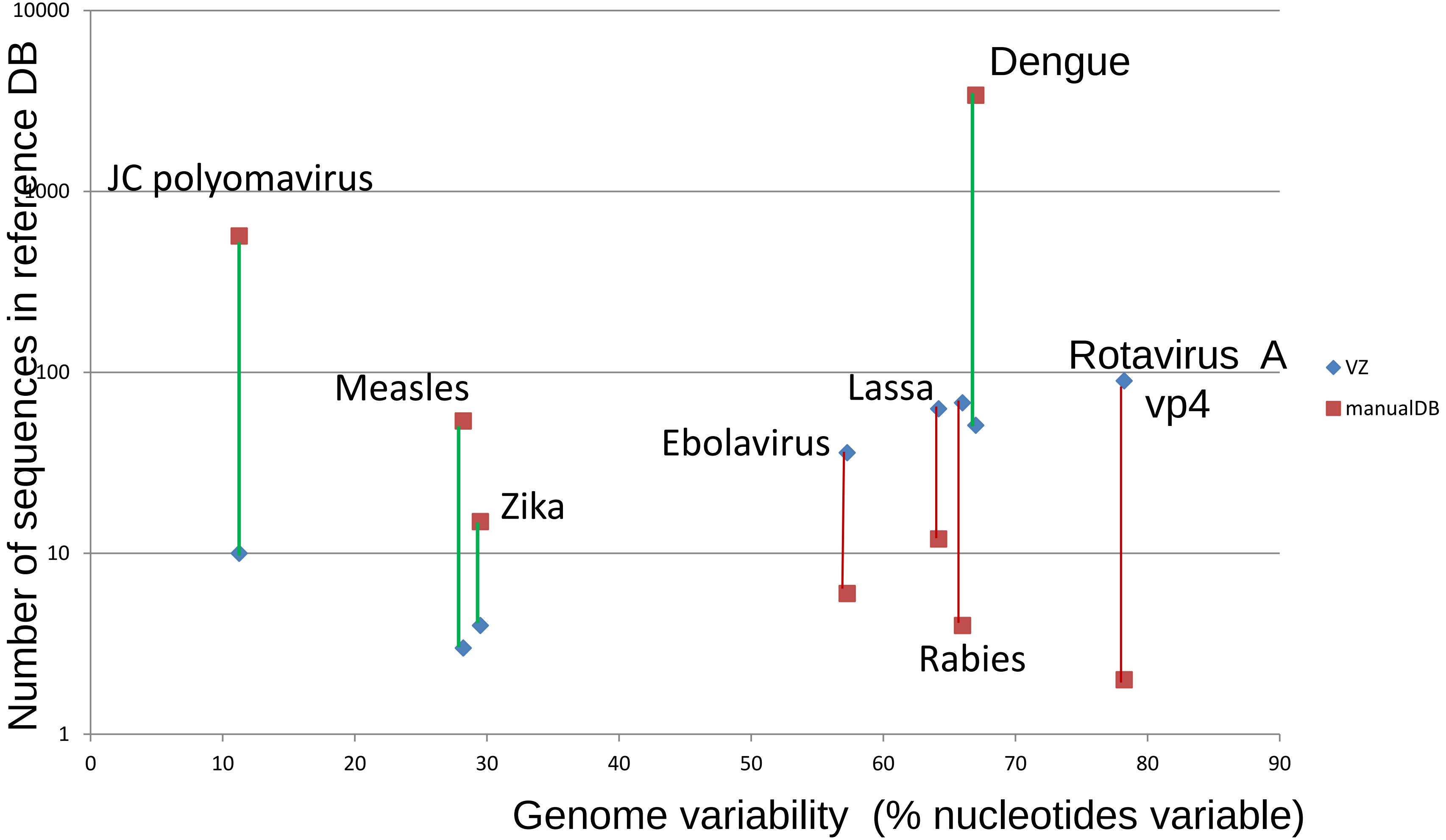
223 complete sequences


175_gb_KU933059_Organism_Lassa_mammarenavirus_Strain_Name_LASV711-NIG-2009_Segment_S_Host_Human
0.1

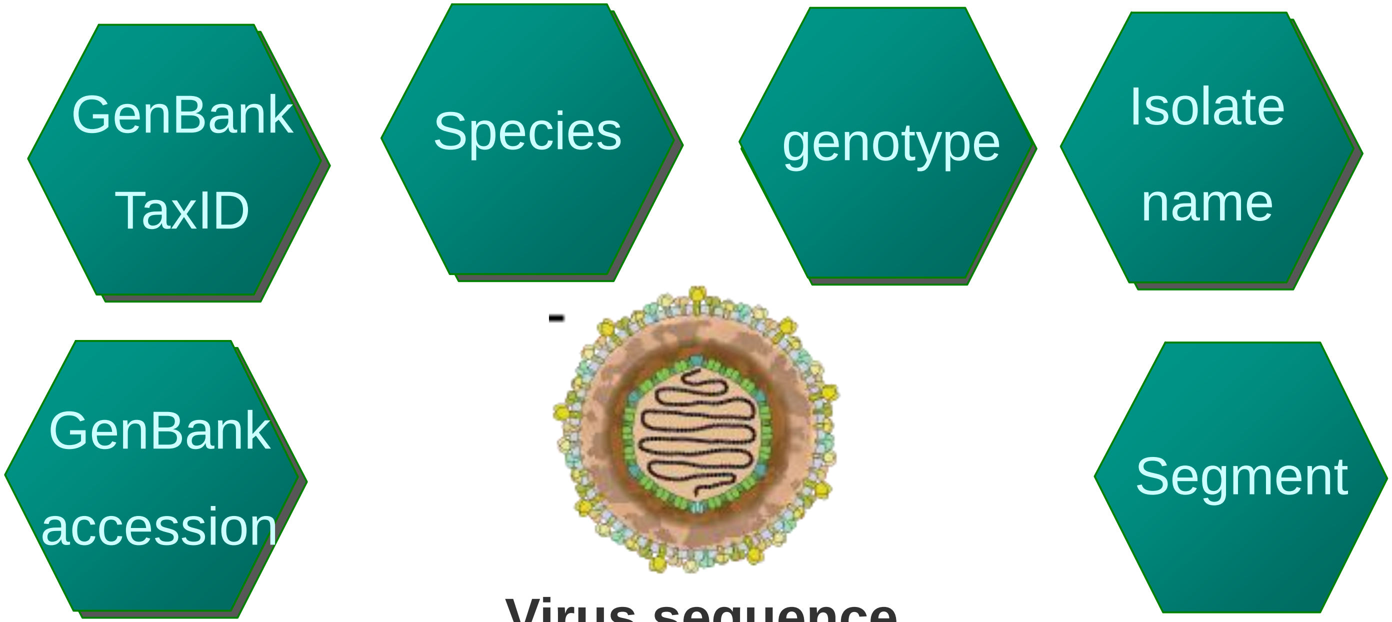
54 clusters at 95% identity


12_KM821828__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821828__product=N/
22_KM821775__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821775__product=N/
5_KM821794__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821794__product=N/A
9_KM821814__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821814__product=N/A
16_KM821845__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821845__product=N/
10_KM821818__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821818__product=N/
15_KM821843__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821843__product=N/
18_KM821854__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821854__product=N/
21_KM821902__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821902__product=N/
14_KM821837__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821837__product=N/
20_KM821859__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821859__product=N/
47_AY179173__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_AY179173__product=N/
19_KM821856__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821856__product=N/
17_KM821847__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821847__product=N/
8_KM821802__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821802__product=N/A
7_KM821800__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821800__product=N/A
13_KM821835__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821835__product=N/
11_KM821822__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821822__product=N/
6_KM821798__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821798__product=N/A
2_AF246121__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_AF246121__product=N/A
4_KF478766__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KF478766__product=N/A
49_KF478765__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KF478765__product=N/
3_GU830839__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_GU830839__product=N/A
45_GU481070__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_GU481070__product=N/
46_GU481072__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_GU481072__product=N/
23_KM821929__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821929__product=N/
41_KM822086__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822086__product=N/
30_KM821959__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821959__product=N/
43_KM822095__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822095__product=N/
29_KM821953__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821953__product=N/
36_KM822027__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822027__product=N/
34_KM821982__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821982__product=N/
24_KM821937__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821937__product=N/
28_KM821947__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821947__product=N/
31_KM821961__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821961__product=N/
37_KM822031__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822031__product=N/
26_KM821943__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821943__product=N/
39_KM822047__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822047__product=N/
40_KM822069__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822069__product=N/
38_KM822033__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822033__product=N/
42_KM822091__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822091__product=N/
25_KM821941__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821941__product=N/
35_KM822025__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822025__product=N/
27_KM821945__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821945__product=N/
32_KM821976__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821976__product=N/
33_KM821979__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM821979__product=N/
1_KU961971__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KU961971__product=N/A
48_KM822128__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_KM822128__product=N/
50_NC_004294_1_Lymphocytic_choriomeningitis_virus_segment_S__complete_sequence
44_AY628201__species=Lassa_mammarenavirus__taxid=11620__genotype=unknown__segment=S__isolate=Isolate_AY628201__product=N/

Manual vs clustering regarding virus variability



Adding annotation to reference sequences



GenBank TaxID

Species

genotype

Isolate name

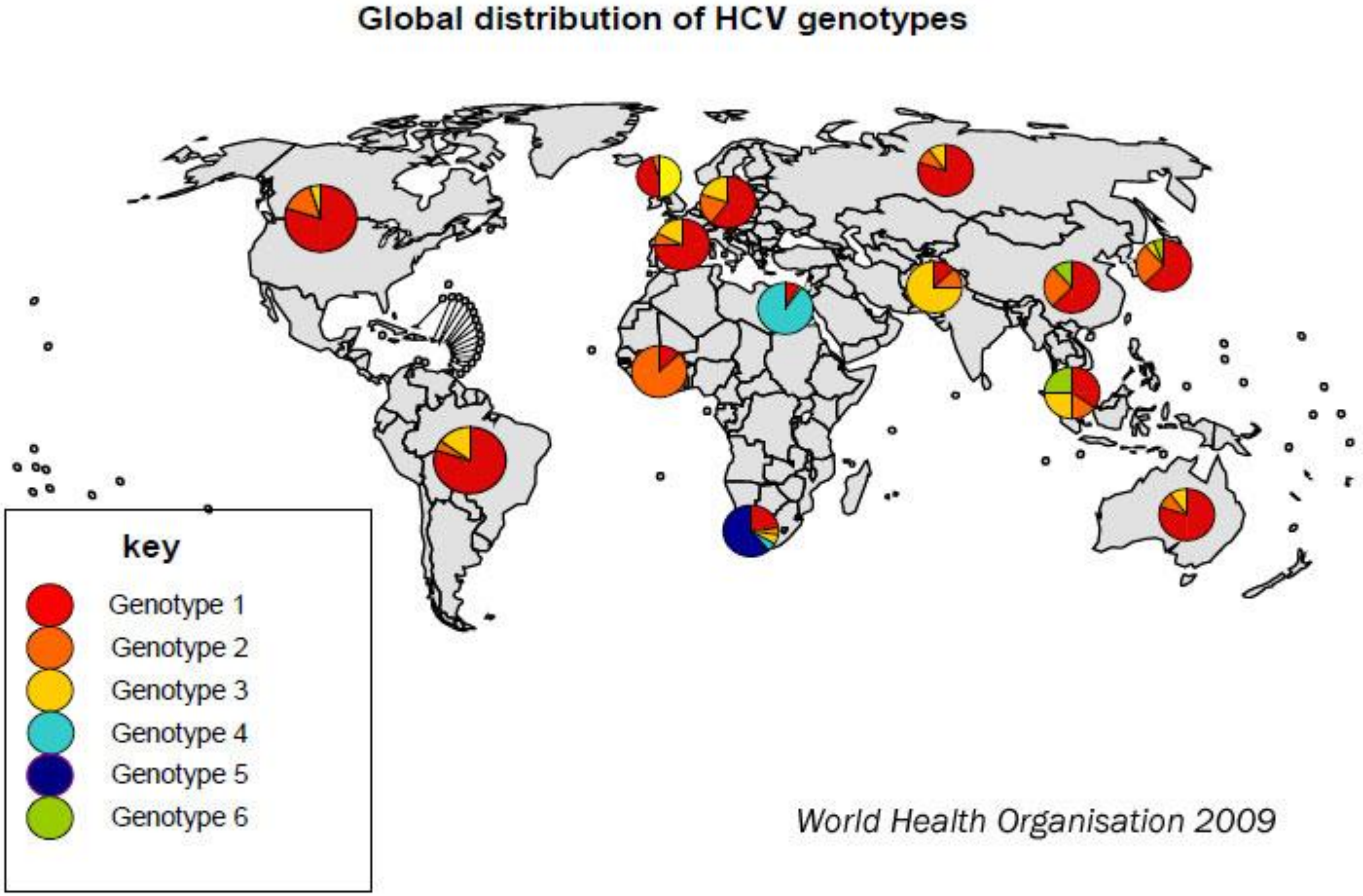
GenBank accession

Segment

-

Virus sequence

Genotyping within a virus species



World Health Organisation 2009

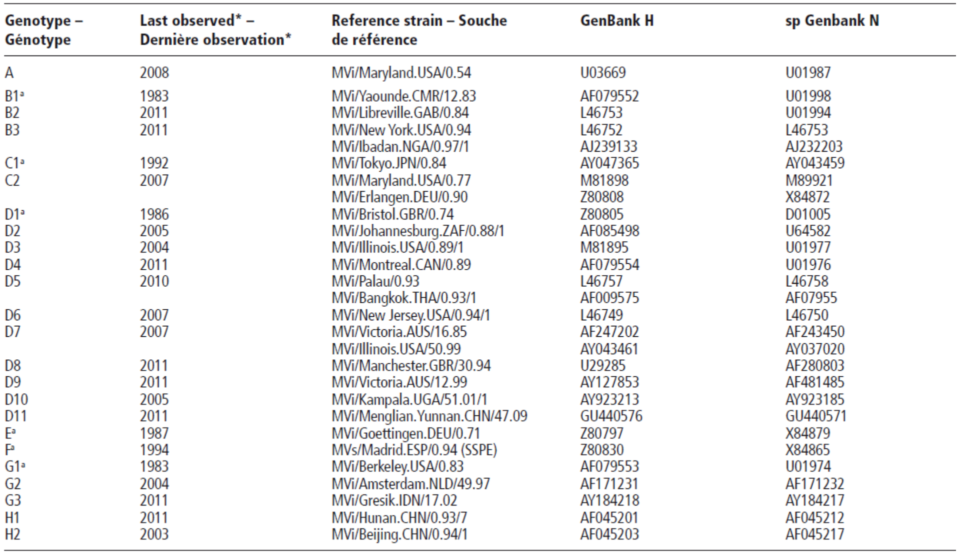
Genotype references data

We are gathering data from various sources: WHO, CDC, publications, Book, virus databases…

Status: in construction

Already done: HAV, HBV, HCV, HSV-1, HEV, HRV, HPV, measles virus, pegivirus, rotavirus, rubella virus, TTV, VZV, West Nile virus, Yellow fever virus, Zika virus

Example: WHO measles



Genotype – Génotype	Last observed* – Dernière observation*	Reference strain – Souche de référence	GenBank H	sp Genbank N
A	2008	MVi/Maryland.USA/0.54	U03669	U01987
B1[a]	1983	MVi/Yaounde.CMR/12.83	AF079552	U01998
B2	2011	MVi/Libreville.GAB/0.84	L46753	U01994
B3	2011	MVi/New York.USA/0.94	L46752	L46753
		MVi/Ibadan.NGA/0.97/1	AJ239133	AJ232203
C1[a]	1992	MVi/Tokyo.JPN/0.84	AY047365	AY043459
C2	2007	MVi/Maryland.USA/0.77	M81898	M89921
		MVi/Erlangen.DEU/0.90	Z80808	X84872
D1[a]	1986	MVi/Bristol.GBR/0.74	Z80805	D01005
D2	2005	MVi/Johannesburg.ZAF/0.88/1	AF085498	U64582
D3	2004	MVi/Illinois.USA/0.89/1	M81895	U01977
D4	2011	MVi/Montreal.CAN/0.89	AF079554	U01976
D5	2010	MVi/Palau/0.93	L46757	L46758
		MVi/Bangkok.THA/0.93/1	AF009575	AF07955
D6	2007	MVi/New Jersey.USA/0.94/1	L46749	L46750
D7	2007	MVi/Victoria.AUS/16.85	AF247202	AF243450
		MVi/Illinois.USA/50.99	AY043461	AY037020
D8	2011	MVi/Manchester.GBR/30.94	U29285	AF280803
D9	2011	MVi/Victoria.AUS/12.99	AY127853	AF481485
D10	2005	MVi/Kampala.UGA/51.01/1	AY923213	AY923185
D11	2011	MVi/Menglian.Yunnan.CHN/47.09	GU440576	GU440571
E[a]	1987	MVi/Goettingen.DEU/0.71	Z80797	X84879
F[a]	1994	MVs/Madrid.ESP/0.94 (SSPE)	Z80830	X84865
G1[a]	1983	MVi/Berkeley.USA/0.83	AF079553	U01974
G2	2004	MVi/Amsterdam.NLD/49.97	AF171231	AF171232
G3	2011	MVi/Gresik.IDN/17.02	AY184218	AY184217
H1	2011	MVi/Hunan.CHN/0.93/7	AF045201	AF045212
H2	2003	MVi/Beijing.CHN/0.94/1	AF045203	AF045217

Adding annotation to reference sequences

Virus sequence

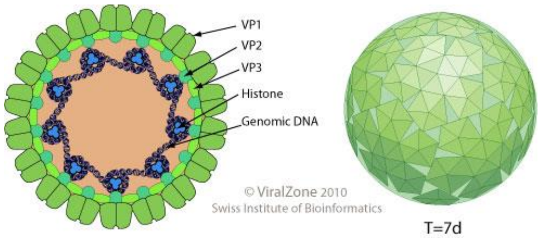
```
>LC190490; species=Rotavirus A; taxid=28875; genotype=unknown; segment=1; isolate=Isolate LC190490;

 AGTTGTTGATCTGTGTGAATCAGACTGCGACAGTTCGAGTTTGAAGCGAAAGCTAGCAACAGTATCAACA
GGTTTTATTTTGGATTTGGAAACGAGAGTTTCTGGTCATGAAAAACCCAAAAAAGAAATCCGGAGGATTC
CGGATTGTCAATATGCTAAAACGCGGAGTAGCCCGTGTGAGCCCCTTTGGGGGCTTGAAGAGGCTGCCAG
CCGGACTTCTGCTGGGTCATGGGCCCATCAGGATGGTCTTGGCAATTCTAGCCTTTTTGAGATTCACGGC
AATCAAGCCATCACTGGGTCTCATCAATAGATGGGGTTCAGTGGGGAAAAAAGAGGCTATGGAAATAATA
AAGAAGTTCAAGAAAGATCTGGCTGCCATGCTGAGAATAATCAATGCTAGGAAGGAGAAGAAGAGACGAG
GCGCAGATACTAGTGTCGGAATTGTTGGCCTCCTGCTGACCACAGCTATGGCAGCGGAGGTCACTAGACG
TGGGAGTGCATACTATATGTACTTGGACAGAAACGATGCTGGGGAGGCCATATCTTTTCCAACCACATTG
GGGATGAATAAGTGTTATA
```

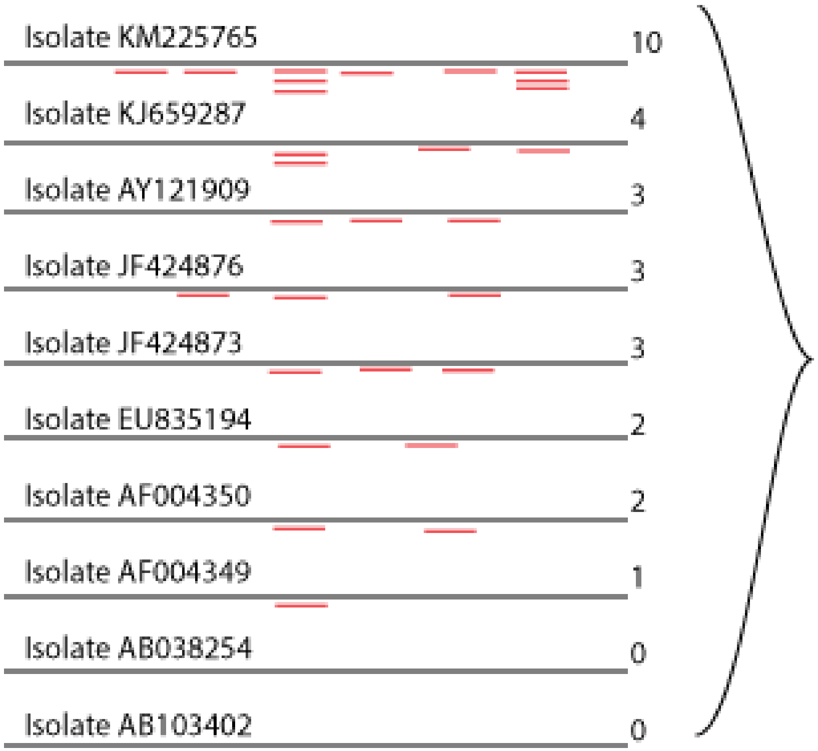
Reads can be assigned to a species level

Isolates

Species



© ViralZone 2010
Swiss Institute of Bioinformatics



Isolate KM225765 10

Isolate KJ659287 4

Isolate AY121909 3

Isolate JF424876 3

Isolate JF424873 3

Isolate EU835194 2

Isolate AF004350 2

Isolate AF004349 1

Isolate AB038254 0

Isolate AB103402 0

Human Polyomavirus 2 (polyomavirus JC)

In search of the best database

Classical approach

Manual Db
(11,256)

Complete
genomes
Database
(16,980)

Source: ViralZone
(247,326)

Can be used in existing pipelines

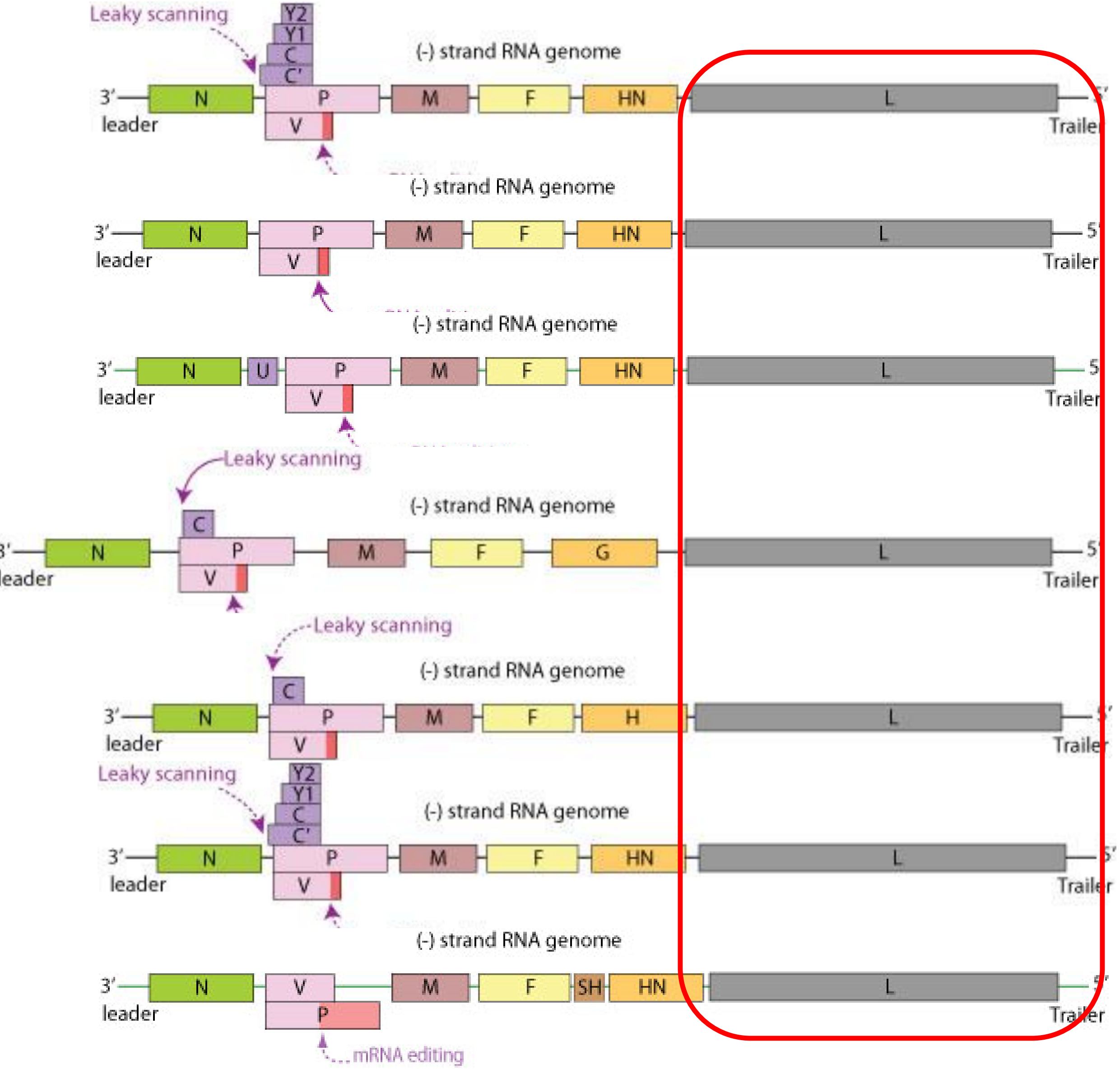
Research approach

ORF
database
(102,280)

Source: GenBank
(1,423,048)

Better representation of virus variability
Haplotypes

Getting a «16S RNA» for viruses



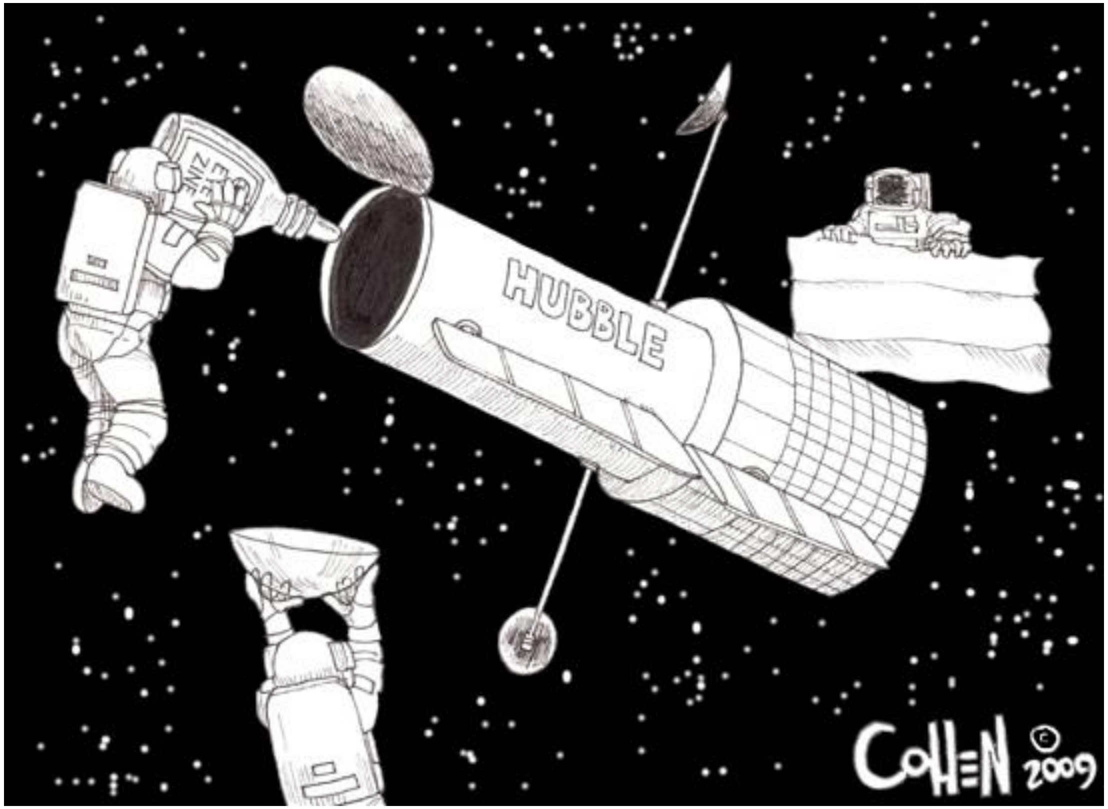
All paramyxoviridae can be identified by their L polymerase polymorphism

Conclusion

- Databases are a **key elemen**t to identify/characterize microbial organisms

- Clinical metagenomics needs **dedicated databases**

- **Sequence curation** with clinical focus facilitates the intrepretation of results



Thank you for your attention

Acknowledgements

Swiss-Prot and Vital-IT
Ioannis Xenarios

Swiss-Prot Group
-Alan Bridge
-Patrick Masson
-Chantal Hulo
-Edouard de Castro
-Andrea Auchinchloss

SIB
Jacques Fellay
Valérie Barbié
Aitana Lebrand

Vital-IT
-Anne Gleizes
-Nicolas Guex
-Christian Iseli
-Thomas Junier

Collaborations
Laurent Kaiser
Samuel Cordey
Florian Laubscher