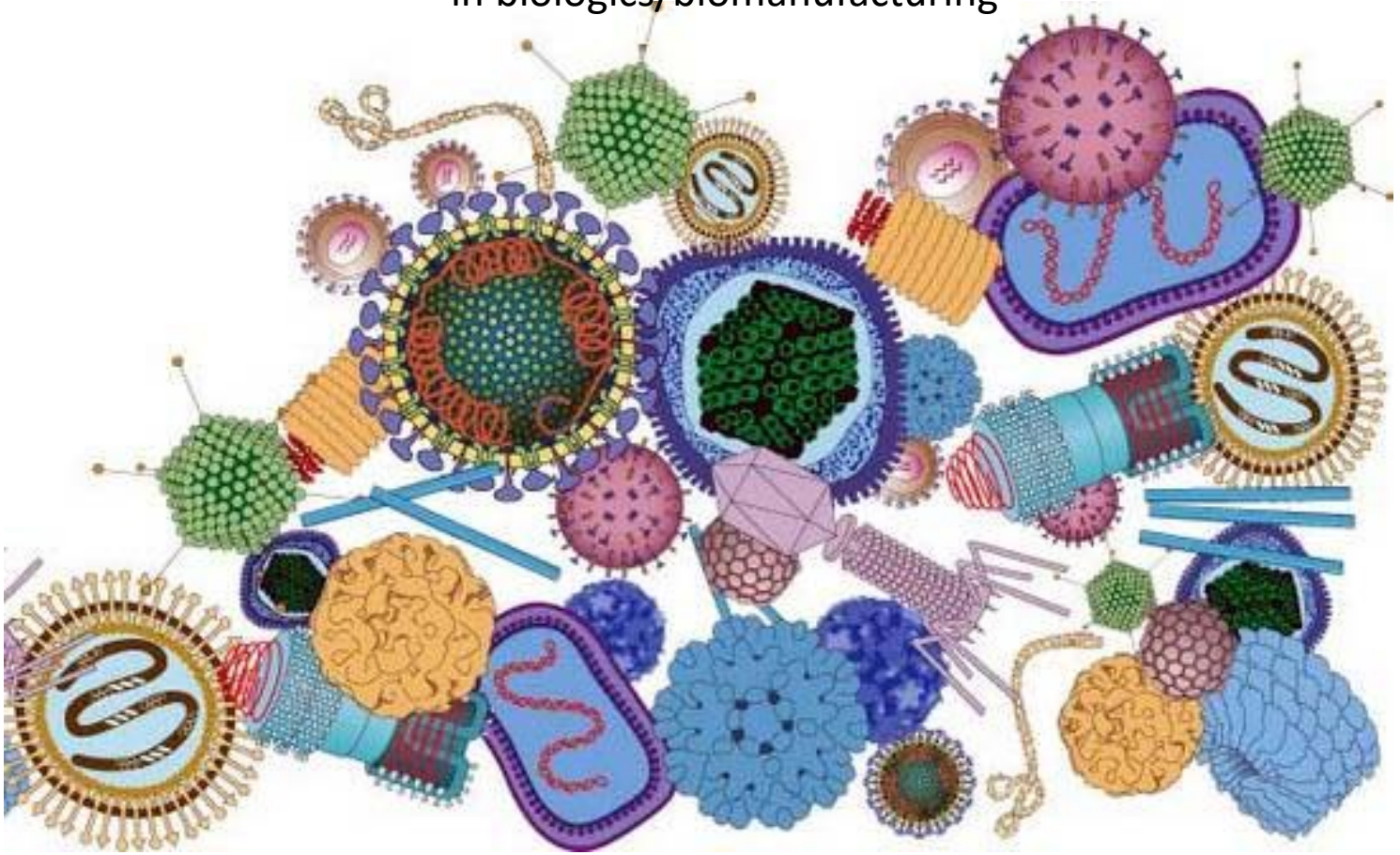
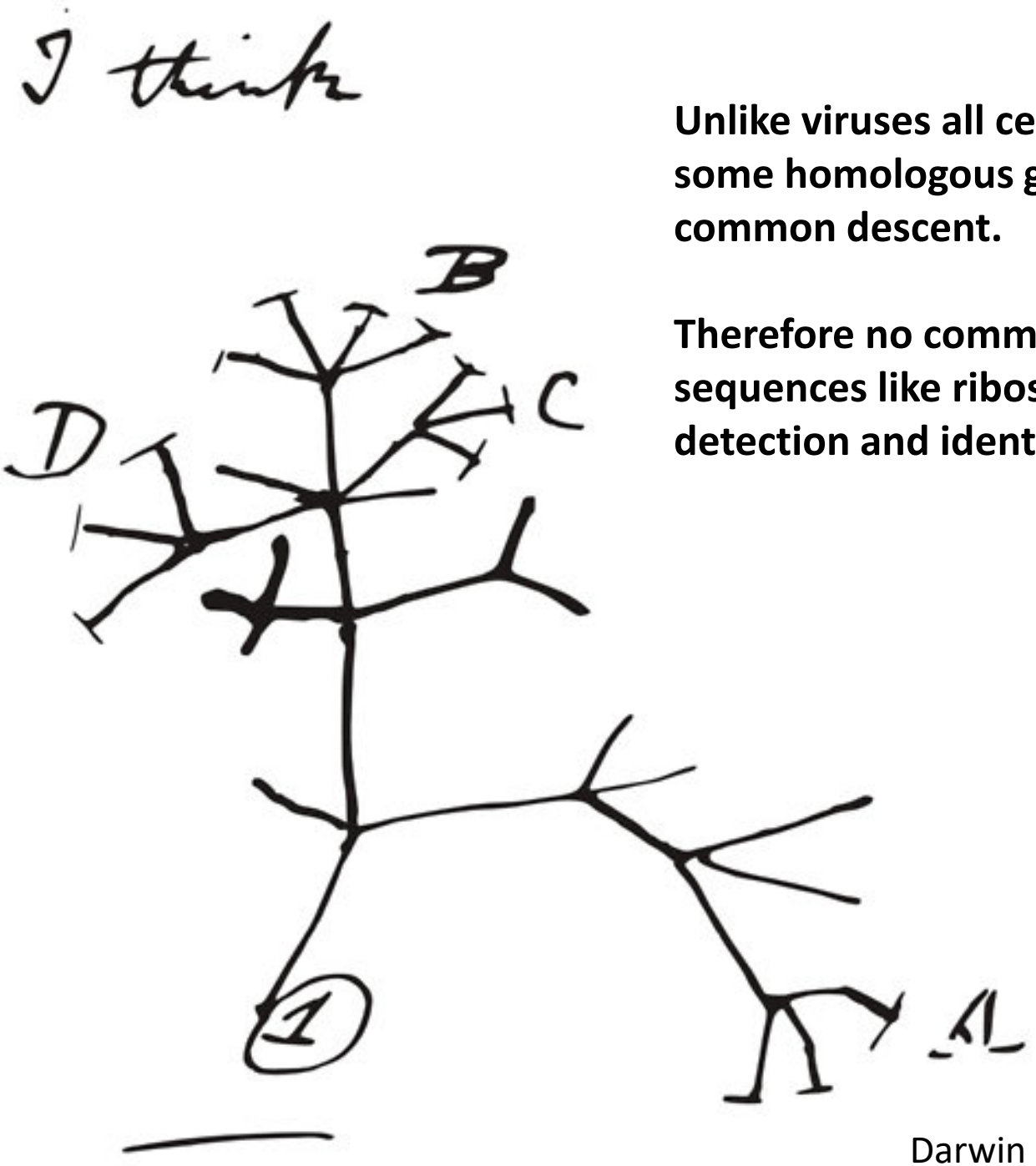
Identification of highly divergent viral sequences.

NIST workshop on standards for NGS detection of viral adventitious agents in biologics/biomanufacturing



Eric Delwart
Vitalant Research Institute & UCSF Lab Medicine.



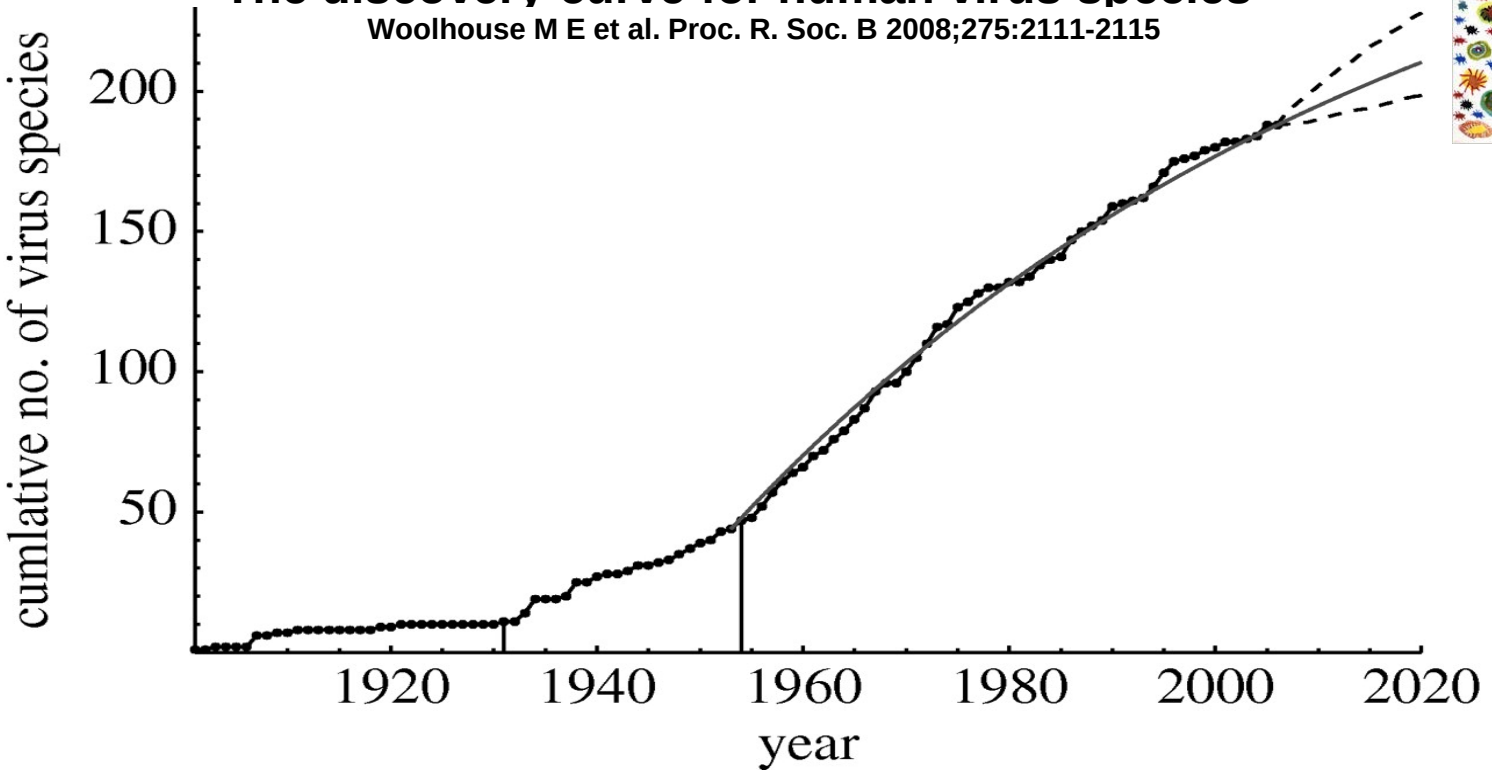
Unlike viruses all cells have some homologous genes from common descent.

Therefore no common sequences like ribosomes for detection and identification

Darwin 1837

The discovery curve for human virus species

Woolhouse M E et al. Proc. R. Soc. B 2008;275:2111-2115



Some "new" human viruses since 2005

- **Parvoviruses:**
  - B19 erythrovirus (3 genotypes)
  - **PARV4** (3 genotypes)
  - **Bocavirus**-1/**2** (2 species)
  - **Bufavirus-1/**2 (2 species)
  - **Tusavirus** (1 species)
  - **Cutavirus** (1 species)
- **Picornaviruses:**
  - Enterovirus (7 species >260 serotypes)
  - Hepatovirus (1 species)
  - Parechovirus (1 species)
  - Kobuvirus (1 species)
  - **Cosaviruses** (5 species >33 genotypes)
  - **Salivirus** (1 species)
  - **Cardioviruses** (1 species >8 genotypes)
- **Astroviruses:**
  - Human mamastrovirus 1 (8 genotypes)
  - Human astroviruses MLB-1/2/3 (3 genotypes)
  - **Human astroviruses HMO-1/2/3** VA4 (4 genotypes)
  - **Human mamastrovirus 20** (1 genotype)
- **Polyomaviruses:**
  - BKV/JC -KIPyV -WUPyV
  - MCPyV -HPyV6 -HPyV7
  - TSPyV -HMWPyV-HPyV9 **-HPyV10** -HPyV11

Sources of viral contaminations in Biologics

- History of contamination starting with live attenuated Yellow Fever Virus vaccine contaminated with HBV from “stabilizer” human serum.
- Frequent use of animal serum in cell cultures.
- Other animal product: porcine trypsin.
- Use of contaminated cell lines.

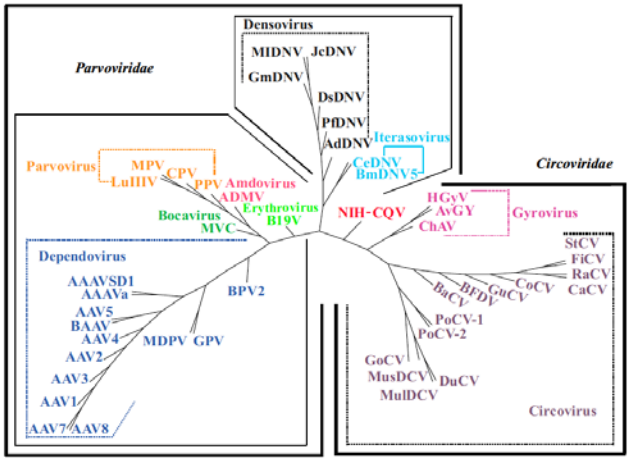
QC of Biologics for viruses

- Avoidance of laboratory contamination requires stringent QC similar to ancient DNA labs (segregation of amplified nucleic acids like PCR amplicons, libraries, and plasmids, unidirectional samples flow).
- NGS data management to detect possible source of contamination (virus present in recent library).
- Awareness of bleed-over between samples sequenced on same Illumina chip.
- Problem bigger for ID diagnostics handling more (and more diverse) samples.

Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing

Baoyan Xu[a,b,1], Ning Zhi[a,1,2], Gangqing Hu[c,1], Zhihong Wan[a], Xiaobin Zheng[d], Xiaohong Liu[a], Susan Wong[a], Sachiko Kajigaya[a], Keji Zhao[c,3], Qing Mao[b,2], and Neal S. Young[a,3]

[a]Hematology Branch and [c]Systems Biology Center, National Heart, Lung, and Blood Institute, Bethesda, MD 20892; [b]Institute of Infectious Disease, Southwest Hospital, Third Military Medical University, Chongqing 400038, China; [d]Department of Embryology, Carnegie Institution for Science, Baltimore, MD 21218

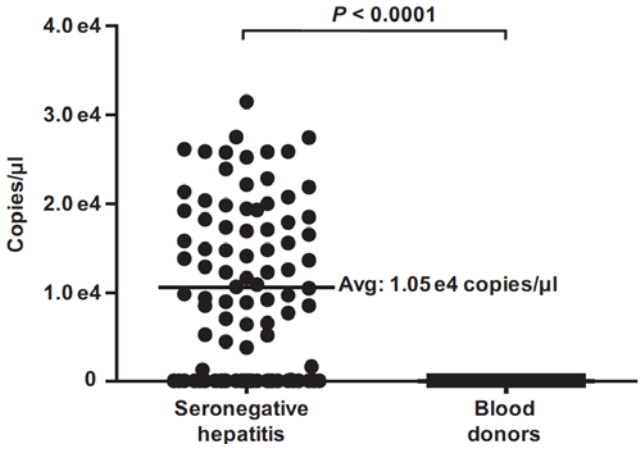


cid extraction kits for parvo·
, full extraction; +, positive l

plicase (247 nt)	Bridge (4
3-1010nt	1554-204

F	C
+	-
+	+
+	-
+	-
+	-
-	-
-	-
-	-
-	-

Water, Molecular Biology Grade NA



PNAS

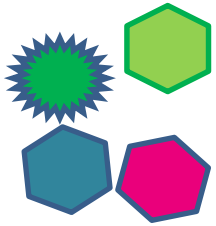
Experimental animals asymptomatically infected

- Mice
- Rat
- Rhesus macaques
- African green monkey
- Zebrafish

Viral metagenomics

virus sized fraction

DNAse and RNAse



Enriched Virus particles

RNA + DNA extraction

Random RT-PCR

5' **GTCCATGCATGACTCGAGTC**NNNNNNNN3'

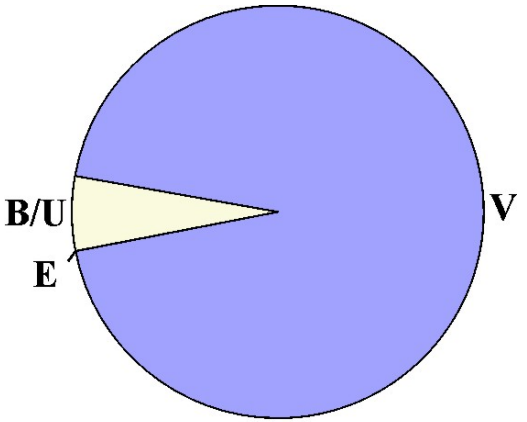
Illumina seq

De novo assembly

BLAST

against all viruses

Oral poliovirus 1,2,3

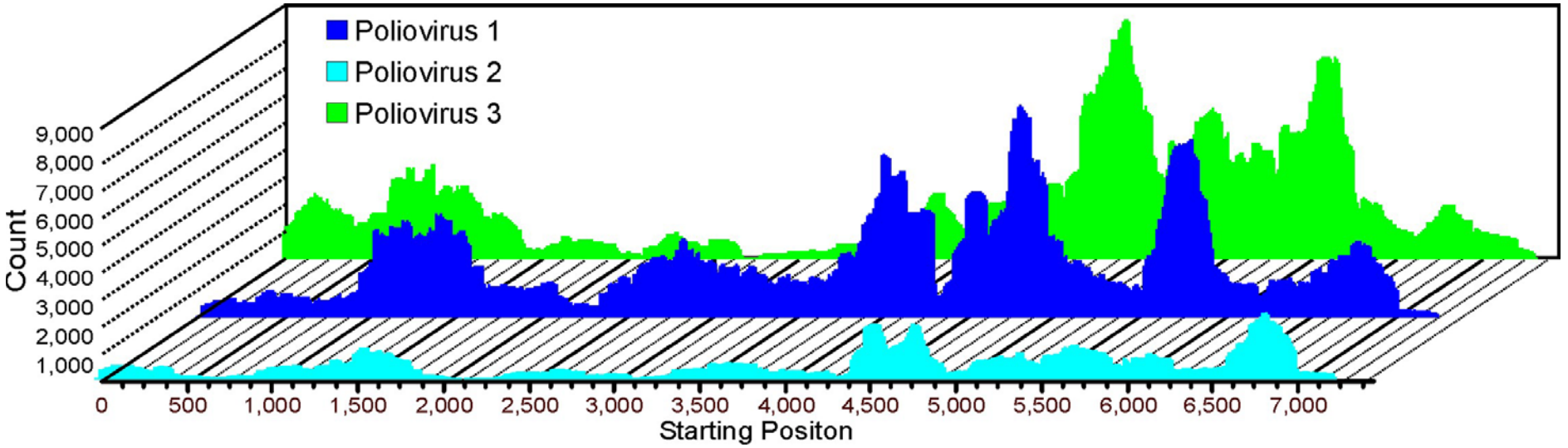


OPV: >400 $TCID_{50}$

Other viruses:

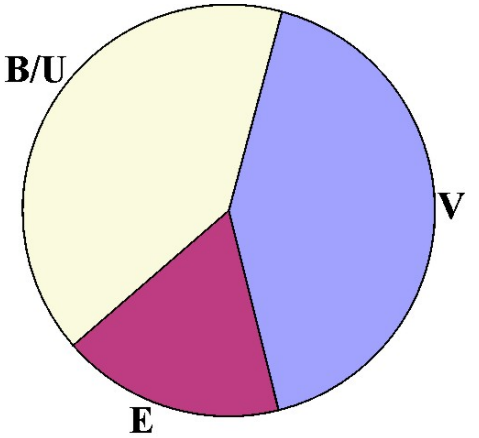
None Detected

60,735 PV1
21,251 PV2
59,943 PV3



Metagenomics is useful tool to detect adventitious viraus contaminations

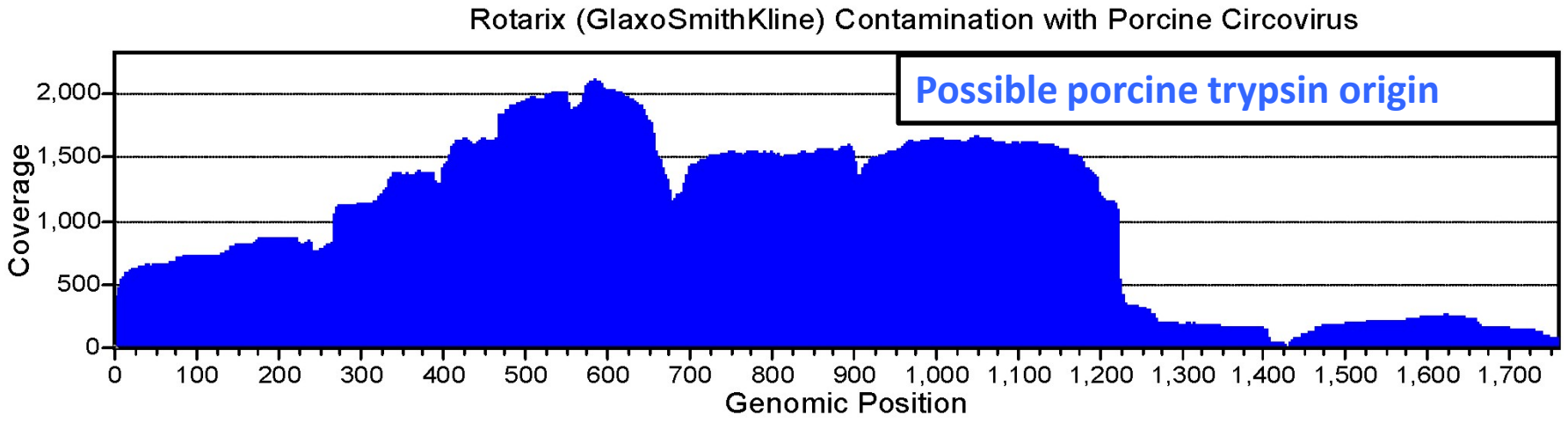
Rotarix



Rotarix: 10^5 $TCID_{50}$

Rotavirus Sequences (458)

6,344 Porcine Circovirus-1



Method is non-specific

Basic

Local

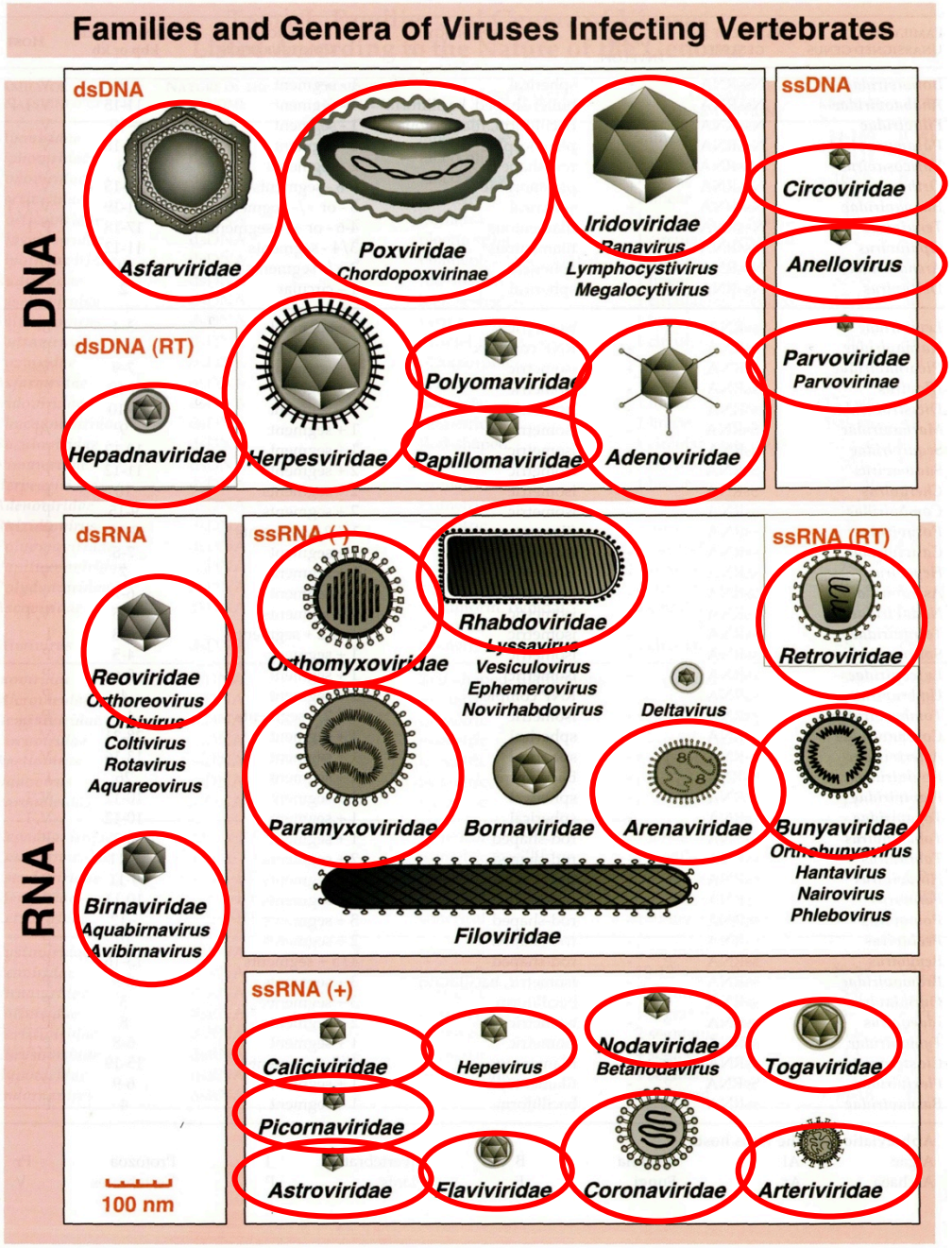
Alignment

Search

Tool

Viruses detected:

BLASTx E score <10^{-5}



Library prep options

- Enrich VLP or extract total NA?
- Total NA versus RNA or DNA?
- Random amplification, primers targeting many viruses, combination?
- Rnadom RT-PCR or Rolling Circle Amplification?
- Libraries finished by ligation or transposon?

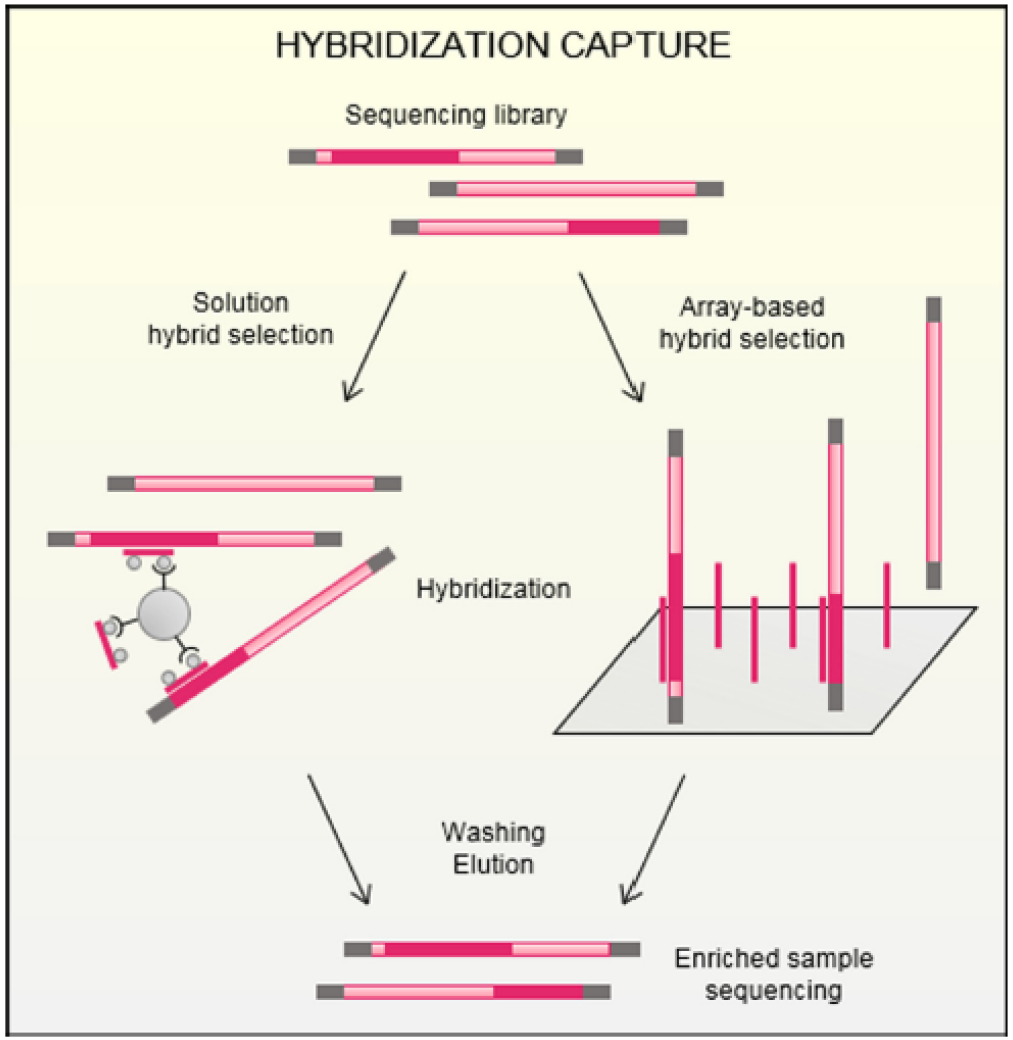
Sensitivity approaching real time PCR using 25 viruses pool (NIBSC)

Virus	Ct	N1	N230	N231	N12	N2	N3	N227	N221	N225	N226	N32	N4	N42	N5	N232	N233	N6
RVA	24.49	5588	3617	6000	1535	4401	7110	5277	5459	2896	1938	1375	9004	1360	18357	24380	24988	1302
HHV5	28.95	156	100	282	62	9	536	457	322	447	298	57	376	40	514	148	112	6
HHV3	29.02	1533	176	43	338	2227	836	826	678	330	219	303	1799	292	1600	350	564	814
HPeV3	29.35	11893	5877	7608	4452	16548	19423	18843	10190	3507	2494	3891	13545	3666	17973	19938	24802	6163
AdV2	29.71	1	0	1	1	249	94	284	301	260	113	18	0	11	16	6	6	0
AstV	30.53	0	0	0	6	69	219	0	0	14	7	1	0	8	24	36	70	0
HHV1	30.59	0	0	0	0	0	0	0	0	11	4	2	0	1	3	0	0	0
CVB4	30.72	1	0	1	6	0	64	12	12	24	4	7	5	9	60	111	189	0
HRV A39	31.16	0	0	0	0	0	0	2	0	6	2	0	0	0	11	32	39	0
HHV4	31.27	0	0	0	3	0	8	0	5	34	31	8	2	11	26	12	6	295
PIV4	31.83	0	0	0	11	1	385	38	53	24	18	18	1	19	18	37	47	0
HMPV A	31.86	0	0	0	0	0	46	0	5	26	35	5	0	0	15	44	40	0
IFV A H1N1	32.02	0	0	0	0	0	7	0	2	2	3	4	0	0	0	0	0	0
HHV2	32.48	0	0	0	0	0	0	0	2	7	9	1	0	0	0	1	0	0
SaV C12	33.37	0	0	0	1	11	0	95	45	14	2	0	0	0	28	84	70	0
PIV2	33.87	3	0	2	35	1	882	102	102	253	164	42	0	30	2	4	4	0
RSV A2	34.33	0	0	0	0	0	0	0	0	4	12	0	0	0	0	5	2	0
PIV1	34.43	4	0	0	57	5	22	32	39	44	26	46	5	59	78	161	207	1480
CoV 229E	36.48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NV GI	ND	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NV GII	ND	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
IFV B	ND	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
IFV A H3N2	ND	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
PIV3	ND	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
AdV41	ND	0	0	0	0	0	0	0	0	4	4	0	0	0	0	0	0	0
Virus detected		8	4	7	12	10	13	11	14	20	20	15	8	12	15	17	15	6

Target capture

uses synthetic RNA baits on magnetic beads to pull down targeted viral sequences increasing sensitivity.

Restricted to already known viruses or close relatives.



Data analysis options

- Use curated viral genome/proteome databases. Update frequently.
- Limit to known human/mammalian/vertebrate/eukaryotic viruses?
- Include virus of unknown tropism (eg many CRESS-DNA viruses, many members of *Picornavirales* order)?
- Nucleotide similarity for fast analyses and close similarities only.
- Protein similarity for more divergent viruses.

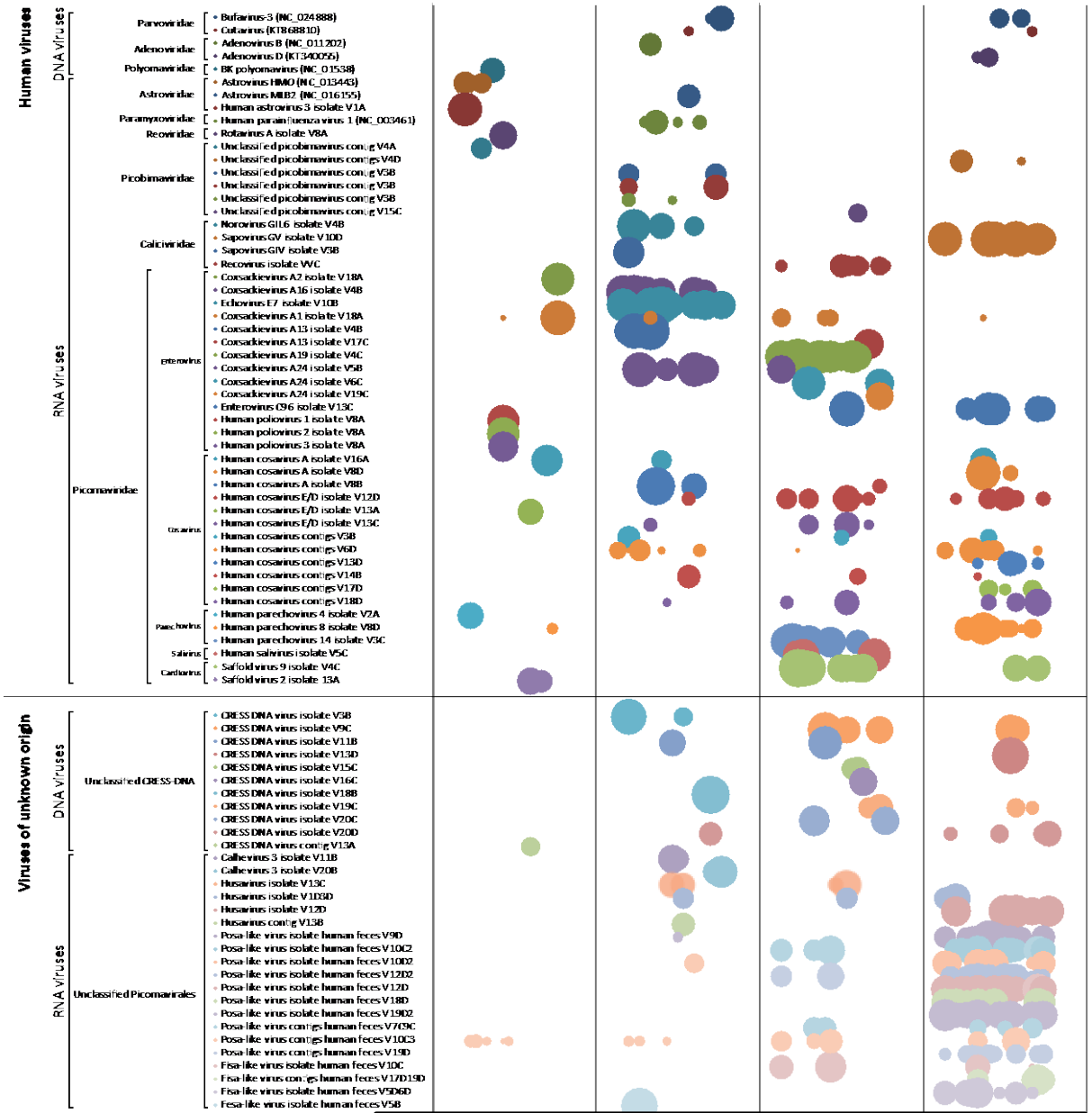
Distribution of eukaryotic viruses in feces from 80 healthy children

Human DNA viruses

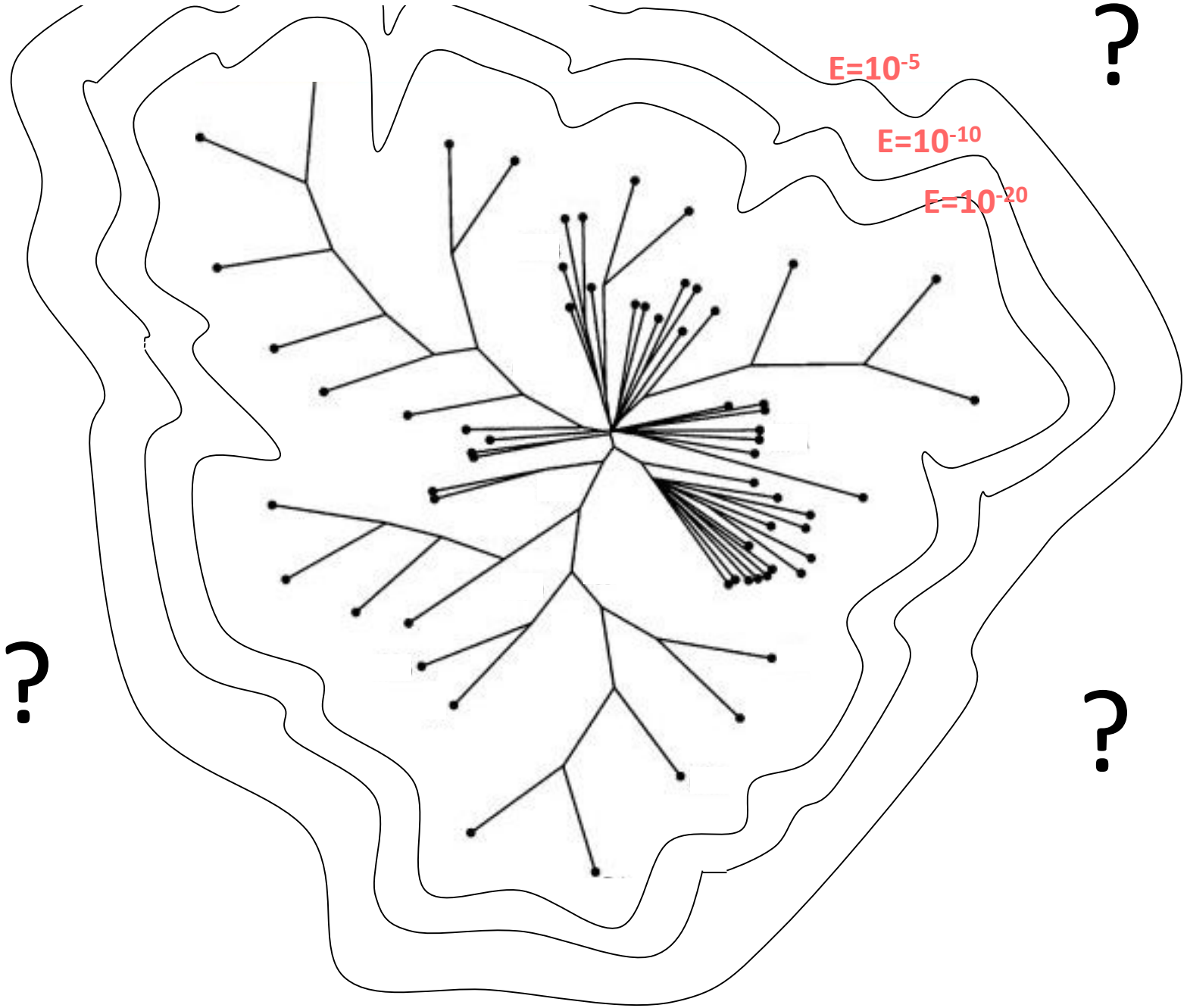
Human RNA viruses

DNA viruses unknown tropism

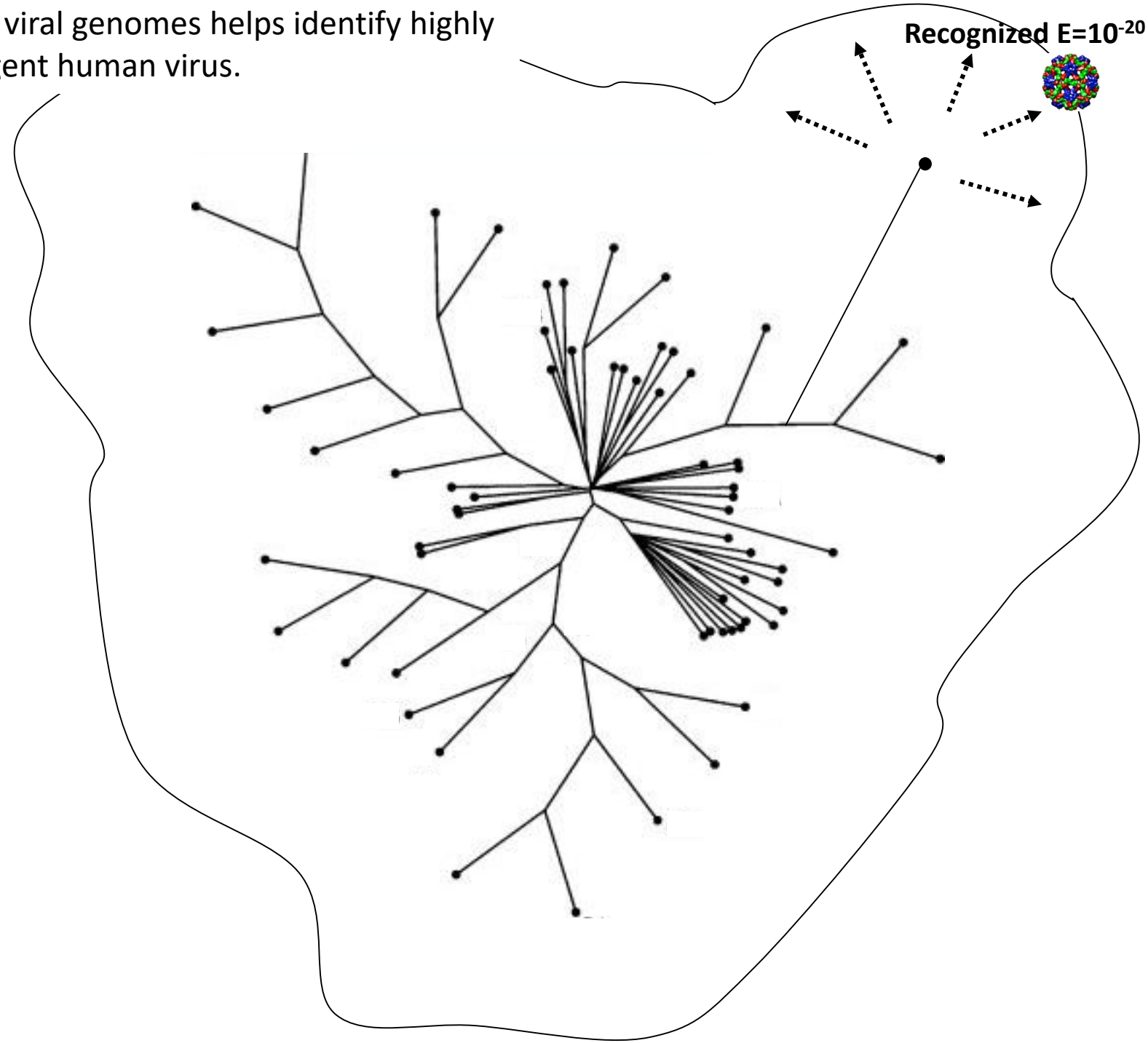
RNA viruses unknown tropism



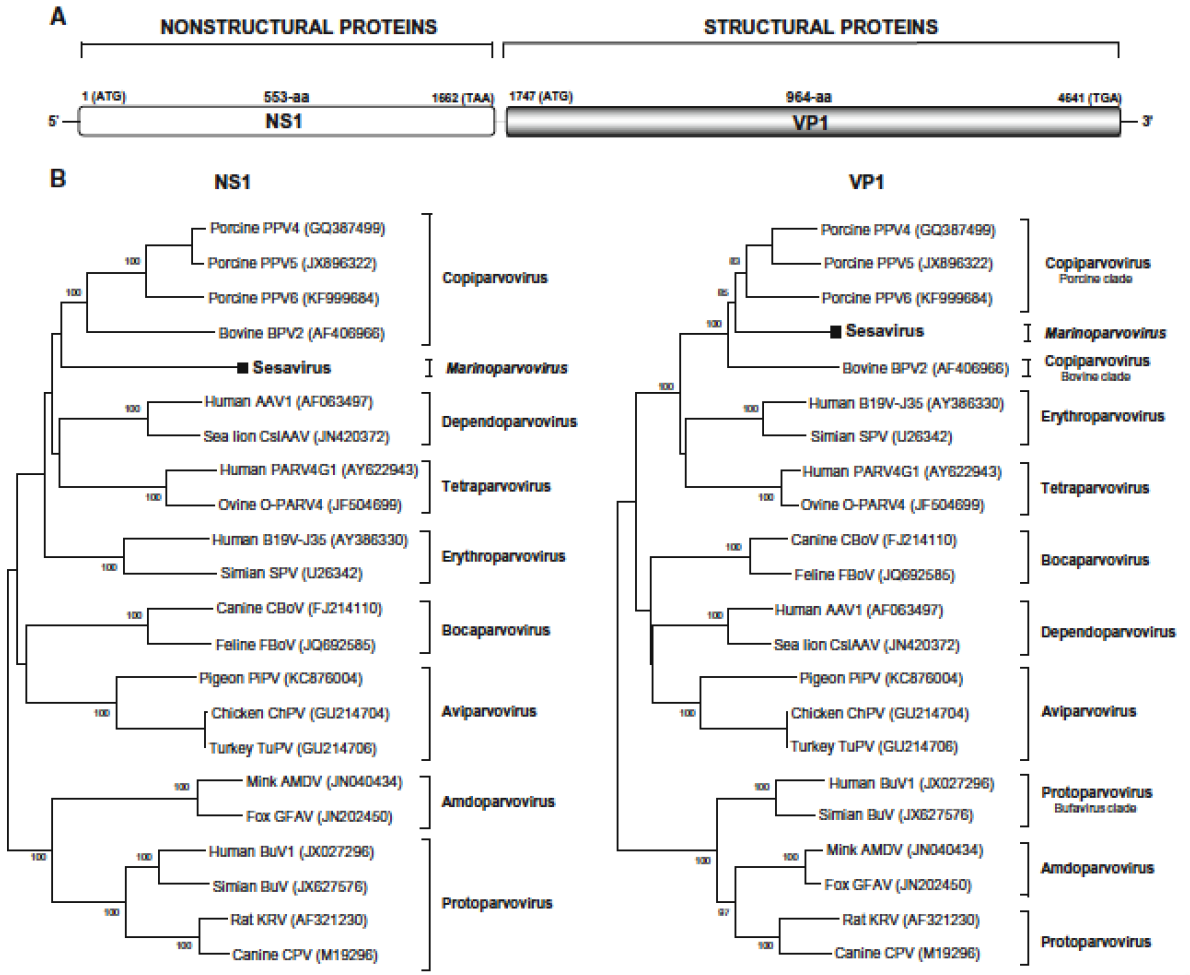
Highly divergent viruses are not recognized by similarity search.



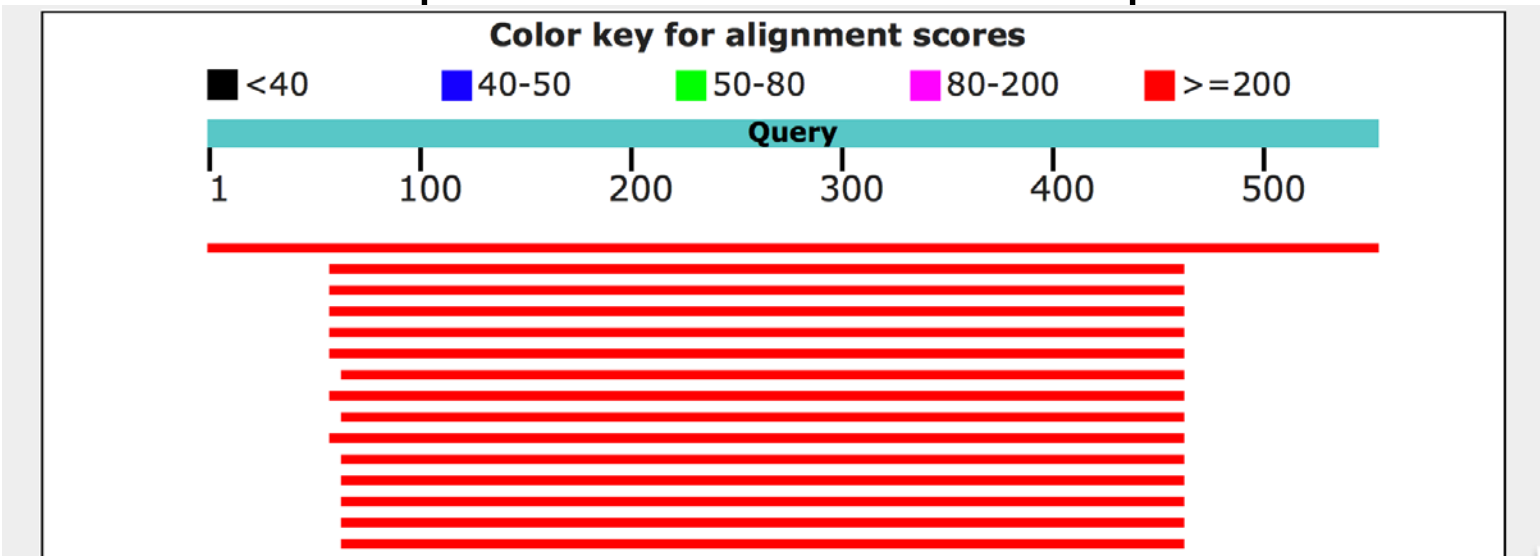
Novel viral genomes helps identify highly divergent human virus.



Identifying a new parvovirus genus

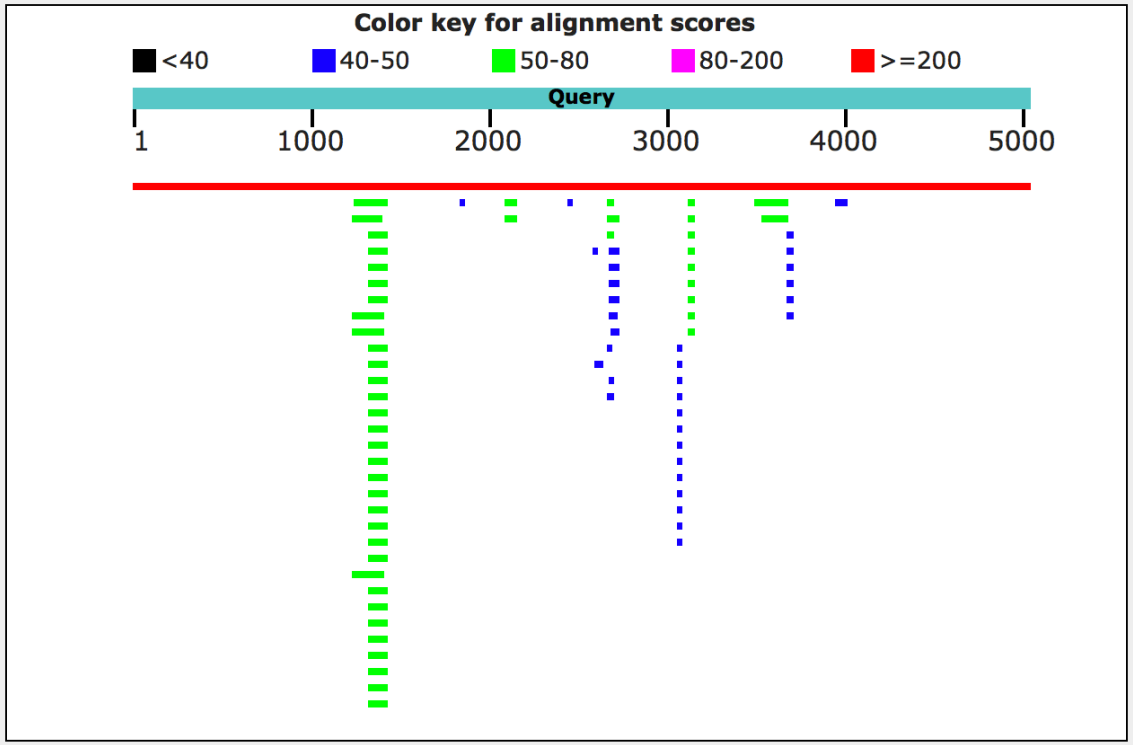


Protein BLASTx of novel NS



	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
NS1 [Sesavirus CSL10538]	1162	1162	100%	0.0	100.00%	YP_009116876.1
replicase [Porcine parvovirus 5]	241	241	72%	3e-68	33.02%	AFU93187.1
nonstructural protein 1 [Porcine parvovirus 5]	239	239	72%	2e-67	32.79%	ANM72084.1
replicase [Porcine parvovirus 5]	239	239	72%	2e-67	32.95%	YP_008888533.1
replicase [Porcine parvovirus 5]	237	237	72%	2e-67	32.56%	QBA84740.1
replicase [Porcine parvovirus 5]	239	239	72%	2e-67	32.79%	QBA84795.1
replicase [Porcine parvovirus 5]	236	236	71%	3e-67	33.18%	QBA84791.1

DNA BLASTn of novel genome



	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Sesavirus CSL10538 NS1 and VP1 genes, complete cds	9106	9106	100%	0.0	100.00%	KM035804.1
Goat protoparvovirus strain C4 nonstructural protein 1 gene, partial cds	70.7	70.7	3%	1e-06	68.72%	MH835426.1
Bosavirus sp. strain 738 putative structural protein gene, partial cds	69.8	69.8	3%	1e-06	68.53%	MH248790.1
Goat protoparvovirus strain C3 nonstructural protein 1 gene, partial cds	66.2	66.2	3%	1e-05	68.18%	MH835425.1
Chicken parvovirus strain GX-CH-PV-12, complete genome	59.9	59.9	2%	0.002	72.38%	KX133419.1
Galliform aveparvovirus 1 isolate IPV, complete genome	59.9	59.9	2%	0.002	72.38%	KU569162.1
Chicken parvovirus isolate ADL120686, complete genome	59.9	59.9	2%	0.002	72.38%	KJ486491.1

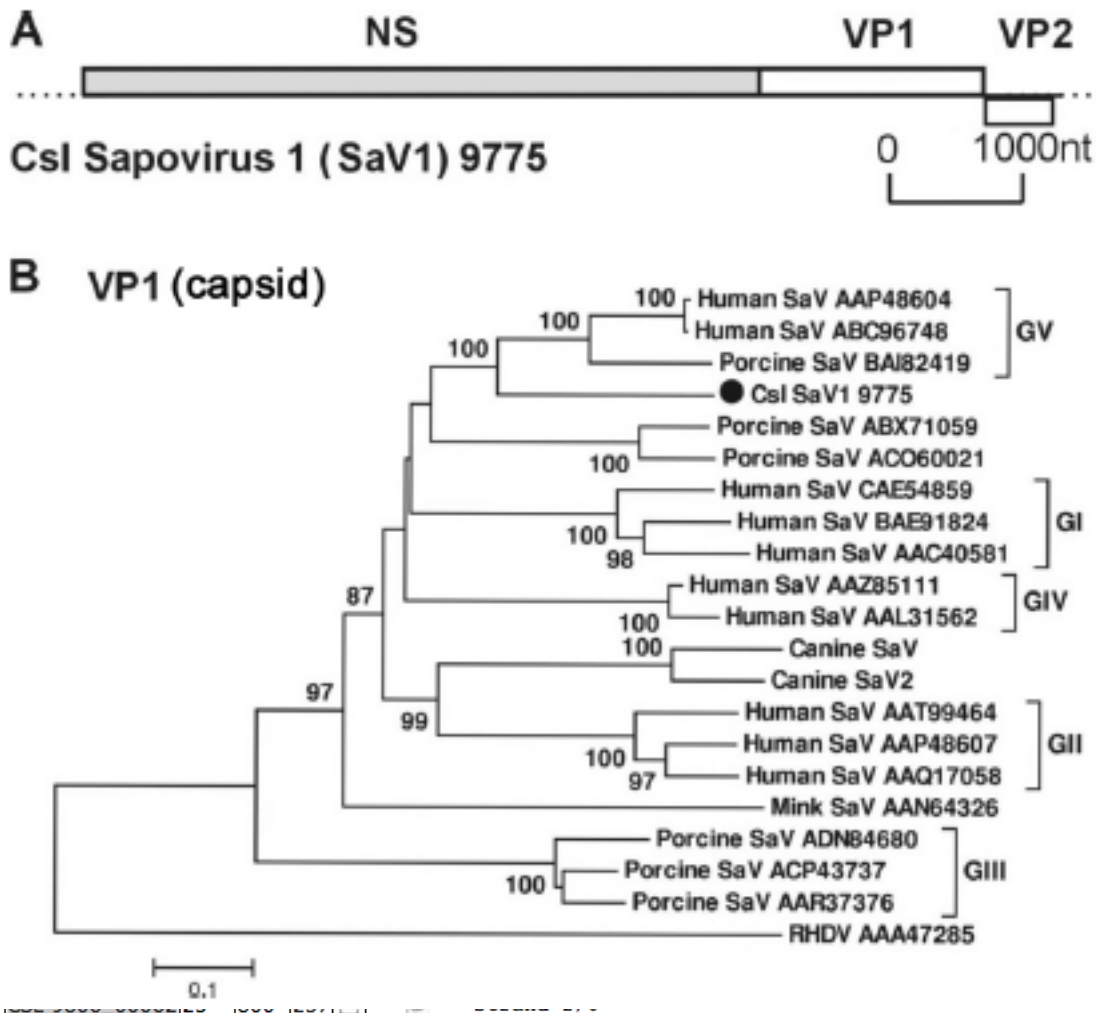
Deposited after reference

A divergent calicivirus with borderline E score $3x10^{-4}$ (23% protein identity)

all tag blast logout

- Plasma pools 5
- Plasma pools 6
- Cancer Tissue
- Rat Stool
- Danish respiratory 1
- Danish respiratory 2
- Pig
- RatVole
- Sea lions

vhit	
1125	1927
1136	956
1137	673
1140	890
1141	574
1148	393
1153	1034
1157	1518
1162	1056
1166	767
1169	865
1170	873
1174	1081
1181	822
1182	1316
1185	1234
1187	1309
1194	1185
1199	960



-VVPLKSDPDETRWHARTHNYTF 323
V K+ PD +W ++T
TTVTAKTQPD-LKWKGLDADHTA 117

-------SEVFFQAGQPNAVSRQ 503
++V P A++ +
PLPDVIATQVRPYQDPPVAINGE 123

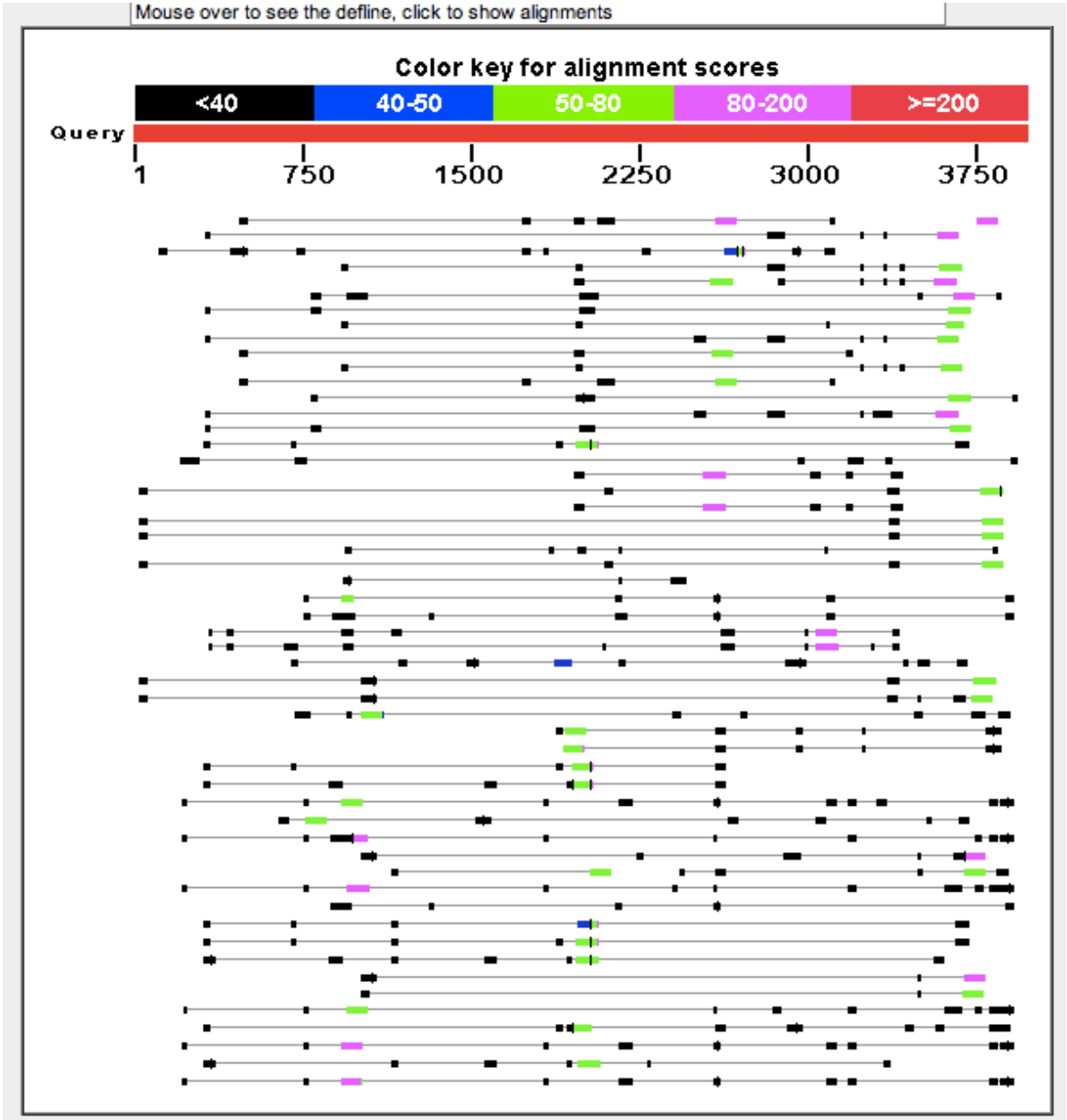
TAAGPSYGVVKQSVFNPDGAPAK 683
++ GP + VK+ DG P
SSNGPWFPGVKKDYTTEDGGPNP 129

AGLTRPIFCFD 760
G R ++ D
DGKRRLLWGCD 1336

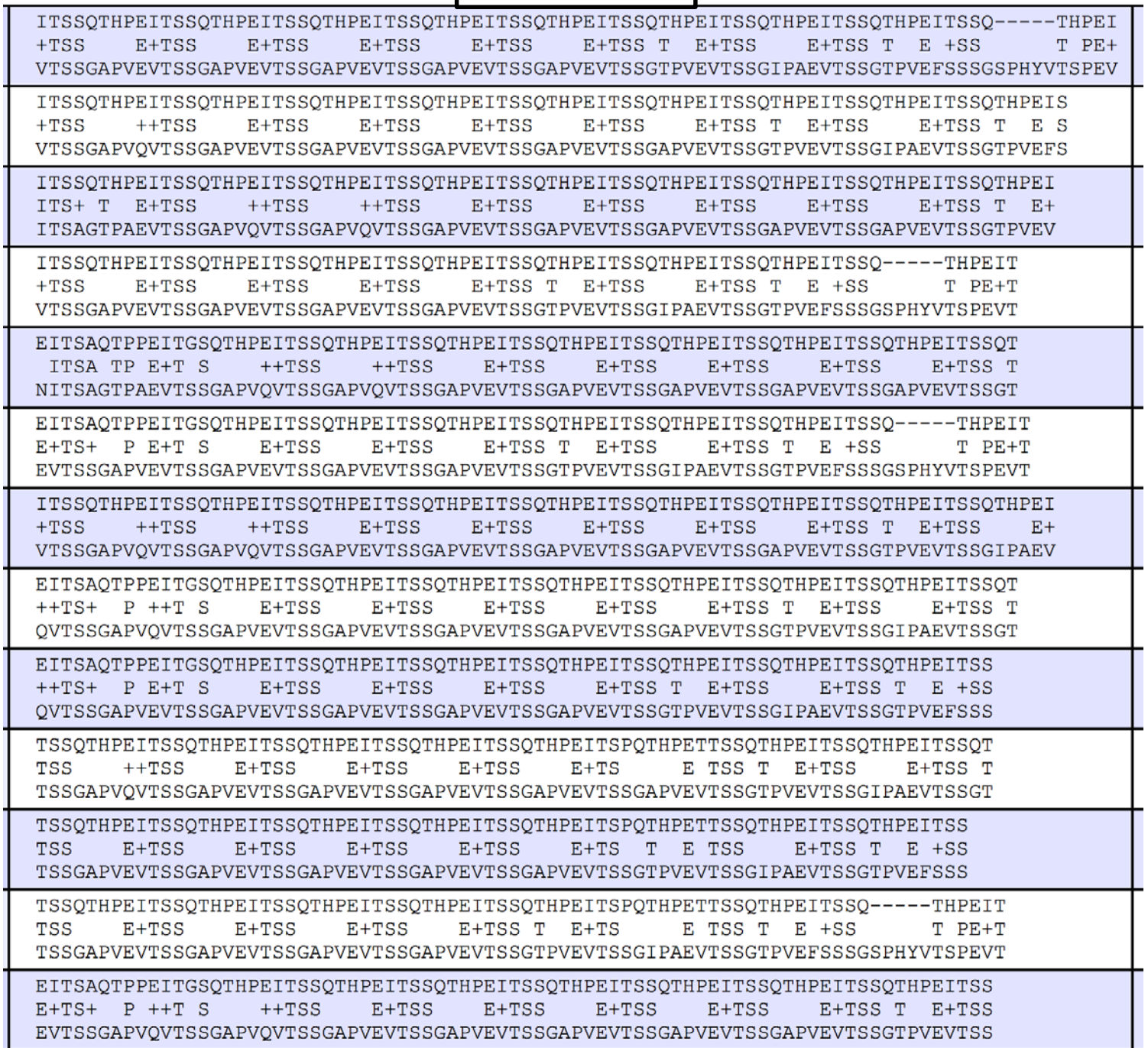
licivirus]

CSL-9806_01301	1	170	56	☐

MINING EXISTING DATABASE:
Re-tBLASTx with “new” viral genomes



Herpesviridae



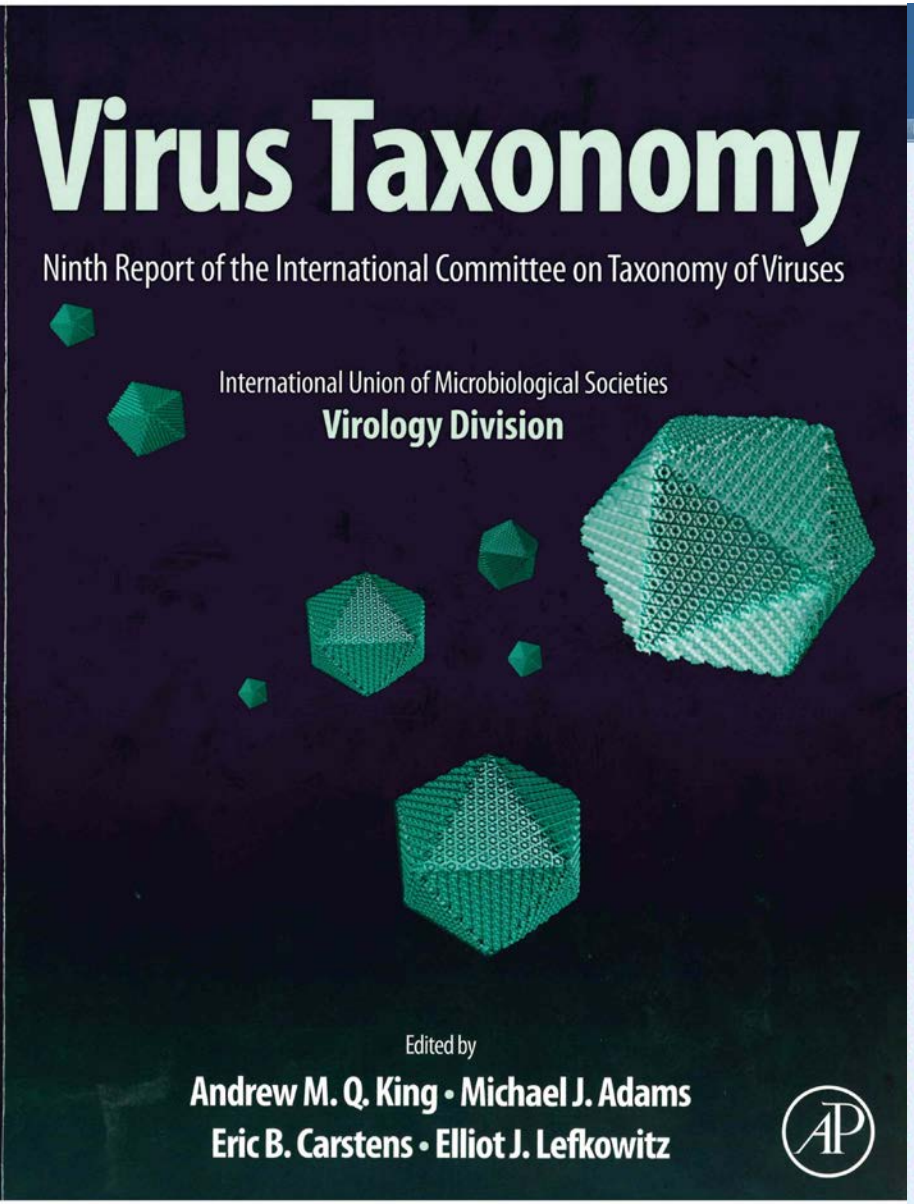
Asfaviridae

```
TCGNTCEDTCGNTYEDTYEHTCG------------NTCWNICEHT----CGNTCENTCENTCGNTCRDTCGNTCGNTCWNSCAQTCEDTCGNTCED
TC +TC  TC +T  DT   TC             NTC ++C  T    C +TC +TC +TC +TC  TC +TC +TC ++CA TC  TC +TC +
TCASTCASTCADTNVDTCVSTCADTNVDTCADTNVNTCASMCADTNVDTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCAN

CEHTCENTCGNTCDDTCENTCGNICWNICENTCGNTCE----DTCKHTCG------------NTCEDTCGNTYDDTYEHTCGDTCWNICEHTCGNTC
C  TC +TC +TC  TC +TC + C + C +TC +TC     DTC  TC             NTC  TC +T  DT   TC  TC + C  TC +TC
CASTCASTCASTCASTCASTCASTCASTCASTCASTCADTNVDTCVSTCADTNVDTCADTNVNTCASTCADTNVDTCASTCASTCASTCASTCASTC

TCGNTCEDTCGNTYEDTYEHTCGNTCW------------NTCEHTCGNTCEDTCENTCGNTYTETCENTCGNICWNTWEHTCAHTCANNCDYT
TC +TC  TC +T  DT   TC +TC             NTC  TC +T  DTC +TC +T   TC +TC + C +T   TCA TCA+ C  T
TCASTCASTCASTCADTNVDTCVSTCADTNVDTCADTNVNTCASTCADTNVDTCASTCASTCASTCASTCASTCASTCASTCASTCAST

TCGNTCEDTCGNTYEDTYEHTCGNTCWNICEHTCGNTCENTCENTCGNTCRDTCGNTCGNTCWNSCAQTCEDTCGNTCED
TC +TC  TC +T   T   TC +TC + C  TC +TC +TC +TC +TC  TC +TC +TC ++CA TC  TC +TC D
TCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCAD

TCGNTCEDTCGNTYEDTYEHTCGNTCWNICEHTCGNTCENTCENTCGNTCRDTCGNTCGNTCWNSCAQTCEDTCGNTCED
TC +TC  TC +T   T   TC +TC + C  TC +TC +TC +TC +TC  TC +TC +TC ++CA T  DTC +TC D
TCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCADTNVDTCVSTCAD

TCGNTCEDTCGNTYEDTYEHTCGNTCWNTCEHTCGNTCEDT------------------------CENTCGNTYTETCENTCGNICWNTWEHTCAHTCANNCDYT
TC +TC  TC +T   T   TC +TC +TC  TC +TC DT                        C +TC +T  +TC +TC + C +T   TCA TCA+ C  T
TCASTCASTCASTCASTCASTCASTCASTCASTCASTCADTNVDTCVSTCADTNVDTCADTNVNTCASTCADTNVDTCASTCASTCASTCASTCASTCASTCAST

TCGNTCEDTCGNTYEDTYEHTCG------------NTCWNTCEHT----CGNTCEDTCENTCGNTYTETCENTCGNICWNTWEHTCAHTCANNC
TC +TC  TC +T  DT   TC             NTC +TC  T    C +TC  TC +TC +T   TC +TC + C +T    TCA TCA+ C
TCASTCASTCADTNVDTCVSTCADTNVDTCADTNVNTCASTCADTNVDTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTC

ICEHTCEDTCENACGNICWNICENTCGNTCKDICGNTCEHTC------------------------VNTCAKSCGNTYADTCENTYEHTCGNTCANICENTCGNTC
 C  TC  TC + C + C + C +TC +TC   C +TC  TC                        VNTCA  C +T  DTC +T   TC +TCA+ C +TC +TC
TCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCADTNVDTCVSTCADTNVDTCADTNVNTCASMCADTNVDTCASTCASTCASTCASTCASTCASTC

RELTCGNTCEDTCGNTYEDTYEHTCGNTCWNICEHTCGNTCENTCENTCGNTCRDTCGNTCGNTCWNSCAQTCEDTCGNTCED
   TC +TC  TC +T   T   TC +TC + C  TC +TC +TC +TC +TC  TC +TC +TC ++CA TC  TC +TC
SAYTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCAS

TCGNTCEDTCGNTYEDTYEHTCGNTCWNICEHTCGNTCENTCENTCGNTCRDTCGNTCGNTCWNSCAQTCEDTCGNTCED
TC +TC  TC +T   T   TC +TC + C  TC +TC +TC +TC +TC  TC +TC +TC ++CA TC  TC +TC
TCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCAS

TCGNTCEDTCGNTYEDTYEHTCGNTCWNICEHTCGNTCENTCENTCGNTCRDTCGNTCGNTCWNSCAQTCEDTCGNTCED
TC +TC  TC +T   T   TC +TC + C  TC +TC +TC +TC +TC  TC +TC +TC ++CA TC  TC +TC
TCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCAS

TCGNTCEDTCGNTYEDTYEHTCGNTCWNICEHTCGNTCENTCENTCGNTCRDTCGNTCGNTCWNSCAQTCEDTCGNTCED
TC +TC  TC +T   T   TC +TC + C  TC +TC +TC +TC +TC  TC +TC +TC ++CA TC  TC +TC
TCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCASTCAS
```

Large fraction of reads are unidentifiable by BLAST

- This **“dark matter”** may contain “new” viral families not yet found in GenBank’s Virus RefSeq
- Weak protein similarity searches using sequence profiles (HHPRED)
- Most new viral families are phages.

Known viral families



NCBI

HOME SEARCH SITE MAP

Entrez Genome

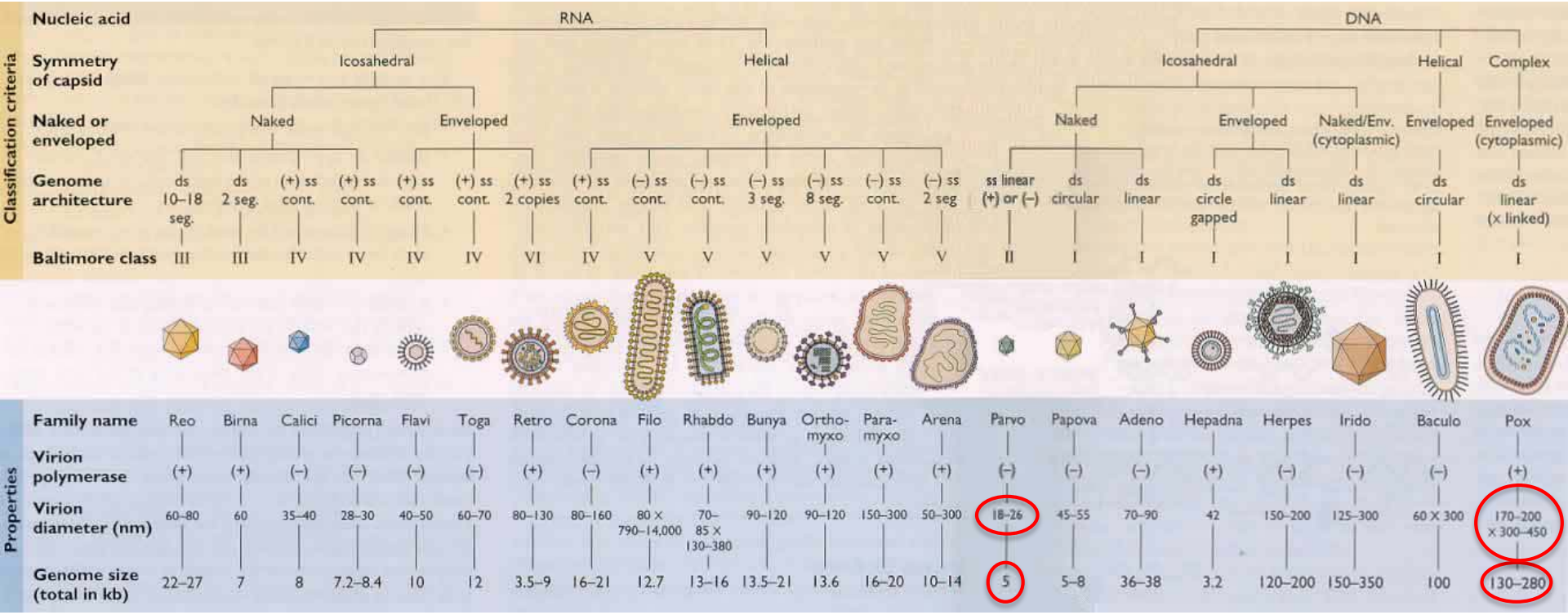
Viral Genomes Home | Taxonomy gr

Viruses - taxonomy groups

- Adenoviridae
- Alphatetraviridae
- Ampullaviridae
- Arteriviridae
- Astroviridae
- Benyviridae
- Bicaudaviridae
- Bornaviridae
- Caliciviridae
- Caulimoviridae
- Circoviridae
- Corticoviridae
- Dicistroviridae
- Endornaviridae
- Fusariviridae
- Geminiviridae
- Hepeviridae
- Hypoviridae
- Iflaviridae
- Leviviridae
- Luteoviridae
- Marseilleviridae
- Microviridae
- Myoviridae
- Nidovirales
- Nudiviridae
- Orthomyxoviridae
- Paramyxoviridae
- Permutotetraviridae
- Picornavirales
- Plasmaviridae
- Polydnaviridae
- Poxviridae
- Retro-transcribing viruses
- Rhizidiovirus
- Salterprovirus
- Siphoviridae
- Tectiviridae
- Tombusviridae
- Tymovirales
- Alloherpesviridae
- Alvernaviridae
- Anelloviridae
- Ascoviridae
- Baculoviridae
- Benyvirus
- Bidnaviridae
- Bromoviridae
- Carmotetraviridae
- Chrysoviridae
- Closteroviridae
- Cystoviridae
- Dinodnavirus
- Filoviridae
- Fuselloviridae
- Globuloviridae
- Herpesvirales
- Hytrosaviridae
- Inoviridae
- Ligamenvirales
- Malacoherpesviridae
- Megabirnaviridae
- Mimiviridae
- Nanoviridae
- Nimaviridae
- Nyamiviridae
- Ourmiavirus
- Partitiviridae
- Phycodnaviridae
- Picornavirales environmental samples
- Podoviridae
- Polyomaviridae
- Quadriviridae
- Retroviridae
- Roniviridae
- Satellites
- Sobemovirus
- Tenuivirus
- Totiviridae
- Tymoviridae
- Alphaflexiviridae
- Amalgaviridae
- Arenaviridae
- Asfarviridae
- Barnaviridae
- Betaflexiviridae
- Birnaviridae
- Bunyaviridae
- Caudovirales
- Cilevirus
- Coronaviridae
- Deltavirus
- Emaravirus
- Flaviviridae
- Gammaflexiviridae
- Hepadnaviridae
- Herpesviridae
- Idaeovirus
- Iridoviridae
- Lipothrixviridae
- Marnaviridae
- Mesoniviridae
- Mononegavirales
- Narnaviridae
- Nodaviridae
- Ophioviridae
- Papillomaviridae
- Parvoviridae
- Picobirnaviridae
- Picornaviridae
- Polemovirus
- Potyviridae
- Reoviridae
- Rhabdoviridae
- Rudiviridae
- Secoviridae
- Sphaerolipoviridae
- Togaviridae
- Turriviridae
- Umbravirus

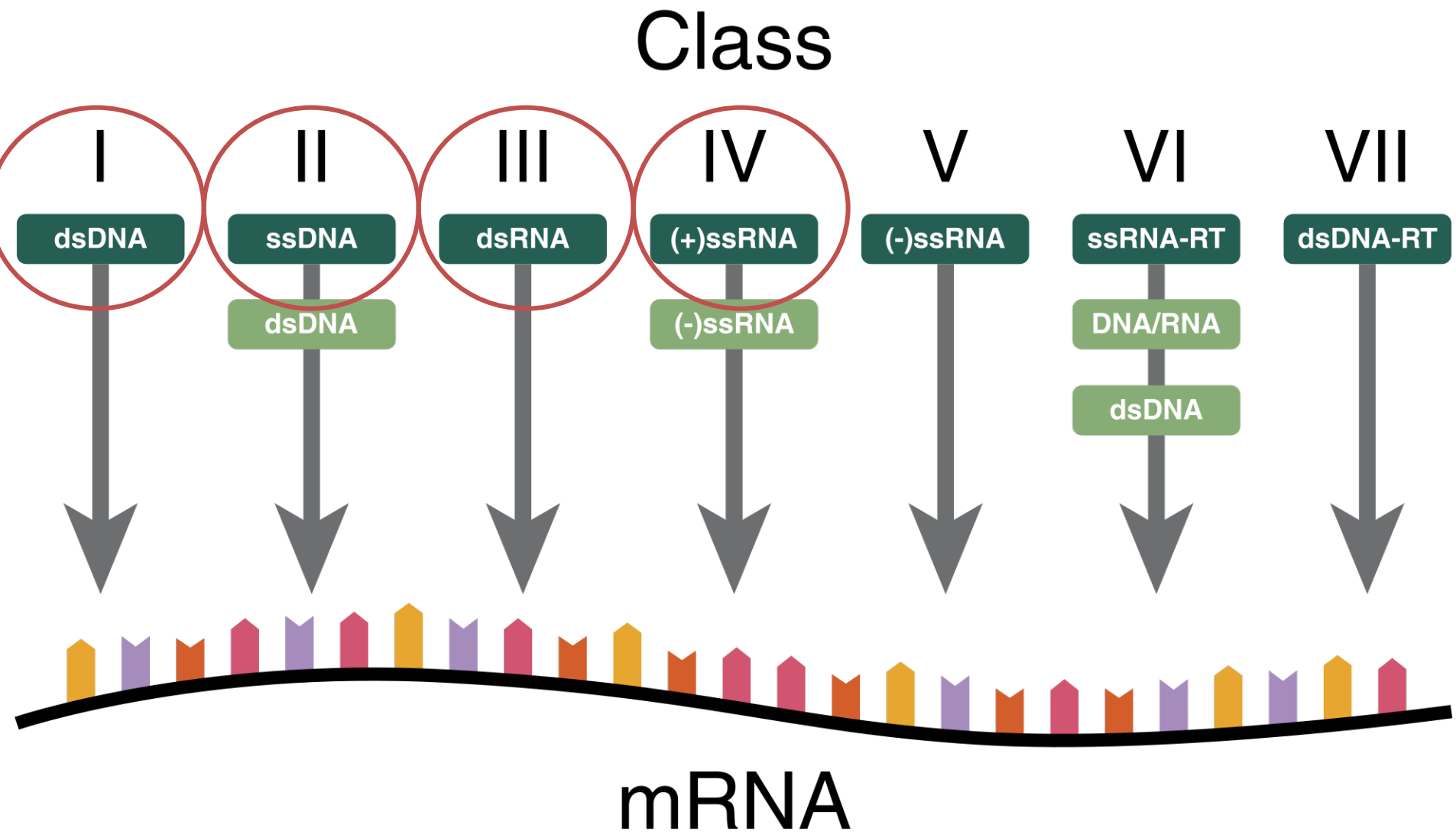
"I'm searching for my keys."

A wide diversity of virion and genome size



Viruses can vary in genome size >1000 fold and in volume >10000 fold

VIRUSES CAN HAVE DIFFERENT TYPES GENOME
(Baltimore classification)



Classic NGS QC standard

For viral particle purification, nucleic acid amplification, and sequencing.

- Include small and large virions +/- lipid envelope. Include ds and ss DNA and RNA viruses.
- Quantify each virus by digital PCR.
- Pool should have approx. same genome equivalent of different viruses.
- Compare efficiency different NGS methods using reads per million (RPM) and % genome coverage.
- Set minimum standard LOD expected for each of the different spiked virus.
- Low concentration unlikely to reduce yield of possible contaminating virus

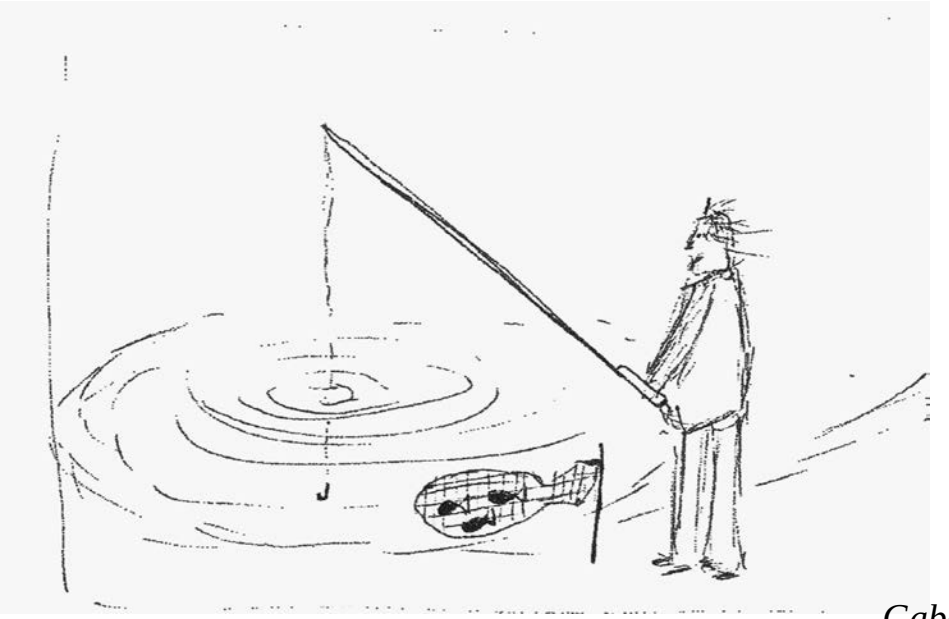
Synthetic NGS QC standard

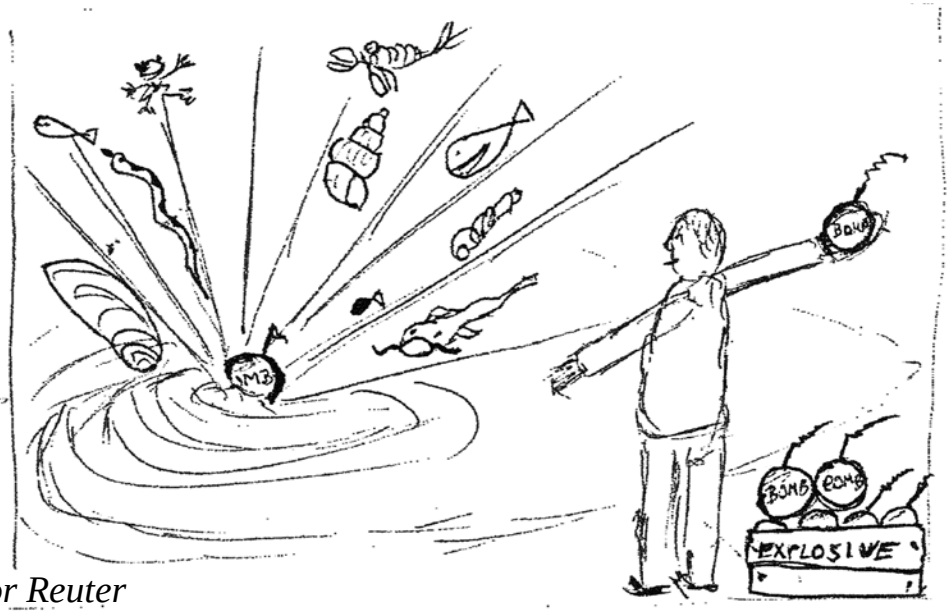
For nucleic acid amplification, and sequencing.

- Synthetize kb long nucleic acids of all known viral genome type: DNA and RNA, single and double stranded.
- Add to extracted nucleic acid to measure efficiency of library construction and sequencing of different type molecules.
- Spike biological sample to measure stability or incorporate into lipid particles of different size to mimic virions.
- Use any sequences. Keep typical viral GC/TA ratio.
- For the same nucleic acid structures (eg ds RNA) use different sequences at different concentrations to measure limit of sensitivity.
- Safer/cheaper/reproducible.

“conventional PCR”

“metagenomics”

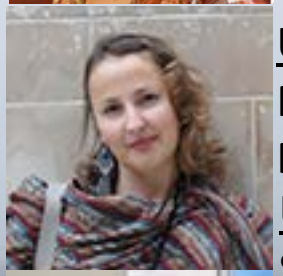


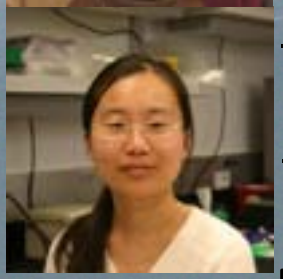


Gabor Reuter

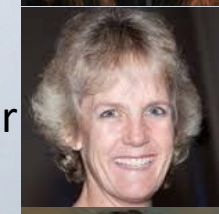
Vitalant Research Institute (aka Blood Systems) & UCSF

Xutao Deng-Bioinformatics
Eda Altan
Liz Fahsbender
Beatrix Kapusinszky (CA Dept Public Health)
Tung Phan (Indiana University, Indianapolis)
Juliana Siquera (Rio de Janeiro,Cancer institute)
Linlin Li (CA Dept Public Health)
Terry Ng (CDC)
Nikola Kondov (10X Genomics)
Lark Coffey (UC Davis)
Wen Zhang (Jiangsu University)
Tongling Shan (Shanghai Vet. Res. Institute)
Amit Kapoor (Ohio State U, Columbus)
Joe Victoria (Boehringer-Ingelheim)
Morris Saffold Jones (Mariposa Dept of Health)



UC Davis
Pat Pesavento
Alison Van Eenennaam
Ardesh Ardeshir
Denis Hartigan-O'Connor

UK NIBSC
Phillip Minor
Ed Mee

UCSF
Samia Naccache
Charles Chiu

Hungarian Natl Ref Lab
Gabor Reuter
Akos Boros

UCLA
Nelson Friemer

IDEXX
Christian Leutenegger

WHO polio eradication group in Pakistan
Sohail Zaidi

Oxford University
Peter Simmonds

Many other collaborators for human, animal, and environmental samples

Funding:
NHLBI, NIAID
Vitalant Research Institute