

School of Biological Sciences



Bioinformatics tools for profiling viral and bacterial communities

Rhys Parry Asgari Lab 21st Feb 2019





The microbiome

My PhD model; Ae. aegypti the yellow fever mosquito



A glossary of terms

Microbiota

"ecological community of commensal, symbiotic and pathogenic microorganisms" includes bacteria, archaea, protists, fungi and viruses.

Metagenomics

the study of genetic material recovered directly from environmental samples.

De novo assembly

The process of assembling short sequencing reads into contigs or scaffolds

Numerous assemblers exist

Basic Local Alignment Search Tool (BLAST)

Algorithm that searches for similarity between biological sequences (nucleotide or protein)

Algorithm	Query	Database
blastn	Nucleotide	Nucleotide
blastx	Nucleotide	Protein
tblastn	Protein	Nucleotide







Today

Designed to be a primer or overview into metagenomics with an emphasis on freeware tools.

Sections:

- 1. Experimental design
- 2. Viral communities
 - De novo assemblers
 - Building blast databases
 - Mapping and visualising community structure
- 3. Bacterial communities
 - Kraken Unbiased shotgun sequencing
 - Amplicon Tools



http://www.usegalaxy.org.au



A plug for Galaxy

http://www.usegalaxy.org.au

The Galaxy / Australia Analyze Data Workflow Visualize - Shared Data - Help - User -	Using 97%
1 Galaxy Australia FTP service is now at ftp.usegalaxy.org.au.	
Tools	C 🕈 🗆
search tools search datasets	8
FILE AND META TOOLS Get Data Bioinformatics work 12 shown, 18 deleted 5.77 GB	cshop
Send Data Original Send Data Convert Formats O 30: blastx output.fa Collection Operations Vs 'output.fa'	
GENERAL TEXT TOOLS Galaxy Australia is currently running Galaxy version 18.09 Filter and Sort 29: seedump fale	
Join, Subtract and Group 29. Sequence, 14 COMMON GENOMICS TOOLS Galaxy is a web-based platform for data intensive biological research.	
Operate on Genomic Intervals 27: output.fa Extract Features Users without programming experience can specify parameters and run tools and 26: Krann on data 24: k	× × •
Fetch Alignments workflows. Galaxy also automatically captures information so that any user can Zo: Kona on data 24: F ML ML OC and manipulation 25: Kona on data 24: F	
EASTA manipulation Z5: Krona on data 23: F Bisard This service is free to use for any Australian researcher. On-line training material is NL	I () / X
SAM Tools User Data Storage Policy ** JULY 2018 news - note new data limits and retention 24: Kraken-report on d	a 🕐 🖋 🗙
VCF Manipulation 23: Kraken-report on d VED tools times	a () / ×
DeepTools 22: Kraken on data 20: Issification Issification	<u> </u>
Blast + Galaxy Australia Jobs (Last 12 hours) 21: Kraken on data 16:	C • / ×
GENOMICS ANALYSIS 10 20: SRR7975696 (fastq dump) Assembly 8 10	• / ×
Mapping 6 16: SRR8583496 (fastq Variant Calling 4 4 GATK Tools 2 4	• / ×
RNA Analysis 0 14:00 16:00 18:00 20:00 22:00 00:00 Annotation	

No "one size fits all" solution for profiling <u>all</u> microbiota Most studies focus on viral or bacterial communities

	Sequencing	Notes
Viral profiling	Ribosomal RNA depletion (For RNA) "Unbiased" high- throughput shotgun sequencing	Requires high depth of sequencing Typically <i>de novo</i> assembly and then a virus discovery pipeline
Bacterial profiling	16S Ribosomal RNA amplicon Sequencing (Replicates)	Well established workflows Low number of reads required to get a meaningful result. Contamination.
Environmental DNA/RNA	"Unbiased" High-throughput shotgun DNA/RNA sequencing	Requires high depth of sequencing May have a <i>de novo</i> step Accounting for everything is computationally intensive Contamination



F QUEENSLAND STRALIA



"Contaminome"

Contaminate laboratory reagents or appear spuriously in PCR and/or sequencing data

Appropriate negative controls must be added

- "Dry" runs of the kit reagents
- "Dry" runs of the spin columns that are being used
- Familiarise yourself with commonly known bacterial contaminants & exclude them from downstream analysis
- Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC biology. 2014 Dec;12(1):87. doi:10.1186/s12915-014-0087-z
- Laurence M, Hatzis C, Brash DE. Common contaminants in nextgeneration sequencing that hinder discovery of low-abundance microbes. PloS one. 2014 May 16;9(5):e97876





@RhysHParry

The "virome"



- Lack of redundant techniques to characterise viruses
- Typically biased depending on the RNA/DNA sample you want to send away
- Typically two parts
 - A) Virus discovery searching for related viruses among assembled contigs, annotation, phylogenetic analysis
 - B) Community profiling, diversity measures, composition



Virus Discovery



RNA-Seq de novo assembly benchmarks

M.musculus RNA-Seq dataset SRX648736 39179 (22740) genes, 94929 (47618) isoforms

	Trans-ABySS	IDBA-tran	SOAPdenovo Trans	Trinity	SPAdes	rnaSPAdes
Transcripts	61508	38294	47025	51245	48706	34615
Aligned	59666	38201	46891	51121	47979	34381
Unaligned	1842	93	134	124	727	234
Gene database coverage, %	15.2	16.8	13.1	18.2	17.6	16.7
Partially-assembled isoforms (>50%)	5824	6804	4608	7089	6998	6916
Fully-assembled isoforms (>95%)	1552	1599	877	2053	2315	2344
Misassemblies	692	378	21	320	817	527
Avg. mismatches per transcript	0.5	0.9	0.4	1.1	0.9	1.1

Credit: "De novo transcriptome assembly Does anybody even need it?", PG 79 - Andrey Prjibelski

URL: <u>http://bioinformaticsinstitute.ru/sites/default/files/denovo_transcriptome_assembly - prjibelski 2-</u> <u>dec-2015.pdf</u>

Virus Discovery



Creating and curating a BLAST database

- Download the proteins you want to compare against deposited in the nonredundant protein database at NCBI.
- Currently there are 5,228,807 virus protein files
 - ~1.5GB fasta file
- Good luck using all those in a database
- Exclude overrepresented viruses in the ٠ non-redundant database with the search string I have included in the appendix
 - 585,825 virus proteins
 - ~200mb file
- Use CD-HIT to collapse a fasta file into a non-redundant 90% cutoff one.
 - 62,840 representatives
 - 30Mb file

https://www.ncbi.nlm.nih.gov/protein/ ٠

SNCBI Resources 🖸	How To 🖸	Sign in to NCBI
Protein	Protein txid10239[Organism.exp] Create alert Advanced	Search Help
Species Animals (1) Bacteria (2) Viruses (5,228,807)	FASTA + 20 per page + Sort by Default order + Send to: + Choose Destination See the results of this search (2477663 items) in our new Identical Protein Groups	Filters: Manage Filters
Customize Source databases PDB (28,930) RefSeq (325,248) UniProt(15,850) Customize Genetic compartments Blavetid (400)	Items: 1 to 20 of 5228807 Ocelections	[Tree] odeficiency virus 1 (1073475) us (792192) (204851) is (155944) us (125245) 2879820)
Plasmid (186) Sequence length Custom range	>BBI54994.1 nucleoprotein, partial (Measles virus genotype VSSTLASELGITAEDARLVSEILMHITTEDRTSRAUGPRQAQVSFLHCDQSENELPGLGCKEDRRVHRSRG EAMESSGEIGSSRLSDARAAHLPTSTPLDIDTASESGQDPQDSRRSADALLRLQAMAGILEEQGSDTDTP RVYNDRDLLD	Find items
Release date Custom range	nucleoprotein. partial [Measles virus genotype B3] GenBank: BBI54993.1 GenPept Identical Proteins Graphics	Search details
Revision date Custom range Clear all	Previous record Next record >BBI54993.1 nucleoprotein, partial [Measles virus genotype B3] VSSTLASELGITAEDAKLVSEILAMHTEEDRISBANGERQAQVSFLHGDQSSELERLGGKEDRRVKONRG EAGESWEFTERSEANDRAINPEOPCH INTEGSFEREDORGSBERSADLI 21 CAMAGE ISFEREDORGTPTP	
Show additional filters	RVYNORDLLD nucleoprotein.partial [Measles virus genotype B3] GenBank: BBI54992.1 Genbank: Heating Complian	Search See more Recent activity
	Seniment Ioemiscal Provious record Next record	txid10239[Organism:exp] (5228807) Protein
	286134774.1 NUCLEOPTOTEIN, PATLAI [#683.68 VITUS @6665Y96 B3] VISTLASELGTHERALVISELMHTERISANGRAQVSFILGOSESLENIGGKEDREVKQNRG EAGESHEETENSRADURAHPPTGTPLDIDTASEFSQDPQDSRRSADALLRLQMAGISEEQDSDTDTP RVINBROLLD	 txid10239[Organism:exp] NOT "Human metapneumovirus" NOT "Human re Protein Research and laboratory contamination can reliable incode secures a beard mission
	nucleoprotein, partial (Measles virus genotype B3) GenBank: BBIG4991,1 GenPlant Intential Britains Graphics	TSA: Antipaluria urichi breed breeding C275695_a_54_0_I_9946, transcribi Nucleotide
	Soun og svenedi FUIGIIIs SIBJIIIUs	Building a BLAST database with local



@RhysHParry

Annotation + Composition



Virus annotation:

ORFfinder <u>https://www.ncbi.nlm.nih.gov/orffinder/</u>

Protein domain prediction:

https://www.genome.jp/tools/motif/ Aggregates:

> Pfam protein database NCBI-CDD database PROSITE Pattern

Annotation + Composition

A)

Composition:

- Re-map the sequencing data to the viruses
- Visualise using heat maps



B)





log10 Reads per million



Bacteriome Endosymbionts Gut bacteria

- Total shotgun metagenomics
 - Kraken : Uses k-mer matching against a database
 - Krona : Visualises the output from Kraken

Amplicon sequencing

MAPseq

- QIIME2 <u>https://giime2.org/</u>
- Mothur <u>https://www.mothur.org/</u>
 - https://github.com/jfmrod/mapseq
- MG-Rast <u>https://www.mg-rast.org/</u>
- MASQUE <u>https://github.com/aghozlane/masque</u>
- Total shotgun metagenomics
 - Kraken
 - QIIME2
- Bacterial databases
 - No need to build your own
 - Many groups have built curated chimera-free databases







- Total shotgun metagenomics
 - Kraken : Uses k-mer matching against a database
 - Krona : Visualises the output from Kraken

Amplicon sequencing

MAPseq

.

.

- QIIME2 https://giime2.org/
- Mothur https://www.mothur.org/
 - https://github.com/jfmrod/mapseq
- MG-Rast https://www.mg-rast.org/
- MASQUE <u>https://github.com/aghozlane/masque</u>
- Total shotgun metagenomics
 - Kraken
 - QIIME2
- Bacterial databases
 - No need to build your own
 - Many groups have built curated chimera-free databases



- Uses k-mer matching against a database
- Has a good tradeoff between sensitivity and time

	Genus	Genus	Speed
Classifier	precision	sensitivity	(reads/min)
Naïve Bayes Classifier	97.64	97.64	7
PhymmBL	96.11	96.11	76
PhymmBL (conf. > 0.65)	99.08	95.45	76
Megablast w/ best hit	96.93	93.67	4511
Kraken	99.90	91.25	1307161
Kraken (quick operation)	99.92	89.54	4101162
MiniKraken 2014 (Kraken w/ 4GB DB)	99.95	65.87	1441476
MiniKraken 2014 (quick operation)	99.98	65.31	2693119
MetaPhIAn	n/a	n/a	370770

Cons (Colossal computational requirements)

- "Construction of Kraken's standard database will require at least 500 GB of disk space as of Oct. 2017.
- After construction, the minimum required database files require approximately 200 GB of disk space."

Memory

- To run efficiently, Kraken requires enough free memory to hold the database in RAM.
- The default database size is 174 GB (as of Oct. 2017), and so you will need at least that much RAM if you want to build or run with the default database.



THE UNIVERSITY

	Genus	Genus	Speed
Classifier	precision	sensitivity	(reads/min)
Naïve Bayes Classifier	97.64	97.64	7
PhymmBL	96.11	96.11	76
PhymmBL (conf. > 0.65)	99.08	95.45	76
Megablast w/ best hit	96.93	93.67	4511
Kraken	99.90	91.25	1307161
Kraken (quick operation)	99.92	89.54	4101162
MiniKraken 2014 (Kraken w/ 4GB DB)	99.95	65.87	1441476
MiniKraken 2014 (quick operation)	99.98	65.31	2693119
MetaPhIAn	n/a	n/a	370770

Pros

You do not need to install and compile the standard database

Two smaller, more workable databases exist:

MiniKraken DB_4GB (2.9 GB) (On galaxy Australia)

MiniKraken DB_8GB (6.0 GB)

Contain between 2.7-5% of the k-mers of the standard library

Workflow

Quality trim your data

Use CD-HIT-DUP to identify duplicates from single or paired Illumina reads

Use Kraken and a representative database

Visualise your report using Krona plots



Example output OTU Table

1	2	3	4	5	6
9.26	2738	2738	U	0	unclassified
90.74	26829	2	-	1	root
90.72	26824	0	-	131567	cellular organisms
90.72	26824	71	D	2	Bacteria
87.74	25941	620	Ρ	1224	Proteobacteria
74.94	22158	934	С	1236	Gammaproteobacteria
38.97	11521	1727	0	135622	Alteromonadales
21.58	6381	0	F	267888	Pseudoalteromonadaceae
21.58	6381	6381	G	53246	Pseudoalteromonas
8.85	2618	111	F	72275	Alteromonadaceae
8.38	2478	0	G	226	Alteromonas
7.77	2297	0	S	314275	Alteromonas mediterranea
7.77	2297	2297	-	1300257	Alteromonas mediterranea U8
0.61	181	181	S	715451	Alteromonas naphthalenivorans
0.05	16	0	G	1621534	Paraglaciecola
0.05	16	0	S	326544	Paraglaciecola psychrophila
0.05	16	16	-	1129794	Paraglaciecola psychrophila 170
0.04	11	0	G	89404	Glaciecola
0.03	10	0	S	300231	Glaciecola nitratireducens
0.03	10	10	-	1085623	Glaciecola nitratireducens FR1064
0.00	1	1	S	983545	Glaciecola sp. 4H-3-7+YE-5
0.01	2	0	G	2742	Marinobacter
0.00	1	0	S	2743	Marinobacter hydrocarbonoclasticus
0.00	1	1	-	351348	Marinobacter hydrocarbonoclasticus VT8
0.00	1	0	S	1033846	Marinobacter adhaerens
0.00	1	1	-	225937	Marinobacter adhaerens HP15
2.65	783	0	F	267890	Shewanellaceae
2.65	783	729	G	22	Shewanella
0.11	33	0	S	192073	Shewanella denitrificans
0.11	33	33	-	318161	Shewanella denitrificans OS217
0.02	7	0	S	56812	Shewanella frigidimarina
0.02	7	7	-	318167	Shewanella frigidimarina NCIMB 400

9.26% Unclassified





Visualising an OUT table using a KRONA plot



https://usegalaxy.org.au/datasets/306e74f5448b28fa/display/?preview=True&dataset=0&node=0 &collapse=true&color=false&depth=5&font=11&key=true



 Most 16S Amplicon sequencing workflows have similar steps

Figure adapted from:

Plummer E, Twin J, Bulach DM, Garland SM, Tabrizi SN. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. Journal of Proteomics & Bioinformatics. 2015 Jan 1;8(12):283.



Comparisons

	QIIME	mothur	MG-RAST
Wahaita			http://metagenomics.anl.g
website	nttp://qlime.org/	nttp://www.motnur.org/	<u>ov</u>
	TES (http://www.h3phele.com/)		
Web-based interface	Not supported/maintained by the	NO	YES (at website
			above)
Primary usage	Command line	Command line	GUI (at website above)
Amplicon analysis	YES	YES	YES
Whole metagenome shotgun	-	-	
analysis	YES*	NO	YES
Sequencing technology	Illumina, 454, Sanger, Ion	Illumina, 454, Sanger, Ion	Illumina, 454, Sanger,
compatibility	Torrent, PacBio	Torrent, PacBio	Ion Torrent, PacBio
16S rRNA gene Databases	RDP, SILVA, Greengenes and	RDP, SILVA, Greengenes	M5RNA, RDP, SILVA
searched	custom databases	and custom databases	and Greengenes
		Needleman-Wunsch, blastn,	
Alignment Method	PyNAST, MUSCLE, INFERNAL	gotoh	BLAT
Taxonomic			
analysis/assignment	UCLUST, RDP, BLAST, mothur	Wang/RDP approach	BLAI
Clustering algorithm	UCLUST, CD-HIT, mothur,	mothur, adapts DOTUR and	UCLUST
Clustering algorithm	BLAST	CD-HIT	UCLUSI alaba
Phylogenetic Tree		Clearcut algorithm	aipila VES
	lasinee		115
Chimera detection	UCHIME chimera slaver BLAST	more	No
		libic	
		Dondrograms, host mans	PCA plots, heat maps,
	PCA plots. OTU networks. bar	Venn diagrams, bar plots.	Krona and Circos for
Visualisation	plots, heat maps	PCA plots	visualisation





A note on databases

- You are only as good as your database
- Green genes <u>http://greengenes.secondgenome.com/</u>
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol.. 2006 Jul 1;72(7):5069-72.
- RDP (Ribosomal Database Project) <u>https://rdp.cme.msu.edu/</u>
- Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic acids research. 2013 Nov 27;42(D1):D633-42.
- SILVA <u>https://www.arb-silva.de/</u>
- SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).
- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research. 2012 Nov 27;41(D1):D590-6.

What is the best?

62 59 59 59 61 Genera 66 55 64 61 61 100 75 Relative abundance (%) 50 25 0 Expected MAPseq (GG) MAPseq (NCBI MAPseq (SILVA) QIIME 2 (SILVA) mothur (SILVA) QIIME 2 (GG) QIIME (GG) QIIME (SILVA) mothur (RDP) Genus Alcanivorax Halomonas Parvibaculum Spirulina Hyphomicrobium Phaeobacter Staphylococcus Alteromonas Arcobacter Lewinella Sulfurimonas Photobacterium Bacillus Loktanella Planctomyces Synechococcus Brevibacillus Maribacter Polaribacter Tenacibaculum Brumimicrobium Marinimicrobium Prochlorococcus Tepidibacter Burkholderia Marinobacter Pseudoalteromonas Thalassobacter Citromicrobium Marinomonas Pseudomonas Thalassomonas

Ocean

• QIIME2

- Undisputedly the best (most sensitive) 16S
 workflow
- Not on Galaxy but is installed on all HPC platforms at UQ
- But as you see: the composition is really not that much different

Figure adapted from:

Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience. 2018 May 11;7(5):giy054.



Computational requirements



Figure adapted from:

Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience. 2018 May 11;7(5):giy054.



Tools I actually find helpful

BBTools

https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/

BBTools is a suite of tools developed by the Joint Genome Institute. The suite includes a pretty reasonable mapper (BBMap) but the most valuable is the **reformat.sh** and **repair.sh** tools for repairing libraries and changing little quirks of fastq data. Re-interlacing, removing adapters from read header files.

Galaxy workshops and tutorials:

https://galaxy-au-training.github.io/tutorials/ https://galaxy-au-training.github.io/tutorials/modules/metagenomics/

Overview for the BIOM format which is used in microbial profiling http://biom-format.org/



References

FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Lightweight java application with a nice GUI that allows you to see potential issues with your sequencing libraries. FastQC is bundled into most Galaxy webservers.

CD-HIT: http://weizhongli-lab.org/cd-hit/

W. Li, and A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22 (2006) 1658-9 An easy to use command line tool to collapse similar fasta or fastq files (Although a webserver is avaliable for datasets up to 100mb: <u>http://weizhongli-lab.org/cdhit_suite/cgi-bin/index.cgi</u>)

Trimmomatic: (A Fastq/fasta trimming tool, bundled into Galaxy)

A.M. Bolger, M. Lohse, and B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (2014) 2114-2120.

Trinity (A *de novo* assembler, in my experience this is the best RNA-Seq *de novo* assembler) M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q.D. Zeng, Z.H. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotech 29 (2011) 644-U130.

BLAST + (<u>https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi</u> Also bundled into Galaxy) C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T.L. Madden, BLAST plus : architecture and applications. Bmc Bioinformatics 10 (2009)

References

BWA-MEM

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997. 2013 Mar 16.

Kraken

Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology. 2014 Mar;15(3):R46.

Search string to exclude commonly overrepresented viruses in the non-redundant database:

txid10239[Organism:exp] NOT "Human metapneumovirus" NOT "Human respiratory syncytial virus A" NOT "Hepatovirus A" NOT "Porcine circovirus 2" NOT "Mumps rubulavirus" NOT "Foot-and-mouth disease virus" NOT txid156614[Organism:exp] NOT txid12333[Organism:exp] NOT "Norovirus GII "NOT "Enterovirus C" NOT txid451344[Organism:exp] NOT txid35237[Organism:exp] NOT "Human alphaherpesvirus 3" NOT "Acanthamoeba polyphaga mimivirus" NOT "Avian avulavirus 1" NOT "Avian coronavirus" NOT "Vaccinia virus" NOT "Hepacivirus C" NOT Influenza NOT Baculoviridae NOT Dengue NOT phage NOT HIV NOT "Hepatitis C Virus" NOT baculovirus NOT "Hepatitis B virus" NOT "Rotavirus A" NOT "Human betaherpesvirus 5" NOT "Norwalk virus" NOT "Simian immunodeficiency virus" NOT "Human orthopneumovirus" NOT "Human gammaherpesvirus 4" NOT "Porcine reproductive and respiratory syndrome virus" NOT "poxvirus" NOT "Alphapapillomavirus 9" NOT "Enterovirus A" NOT "Rabies lyssavirus" NOT "Human alphaherpesvirus 1" NOT "Human papillomavirus type 16" NOT "Zaire ebolavirus" NOT Nudivirus NOT megavirus NOT "Enterovirus B" NOT "Measles morbillivirus" NOT "Cowpox virus" NOT "Human betaherpesvirus 6" NOT "Orthohepevirus A"