The University Of Queensland
AUSTRALIA
School of
Biological Sciences

Bioinformatics tools for profiling viral and bacterial communities

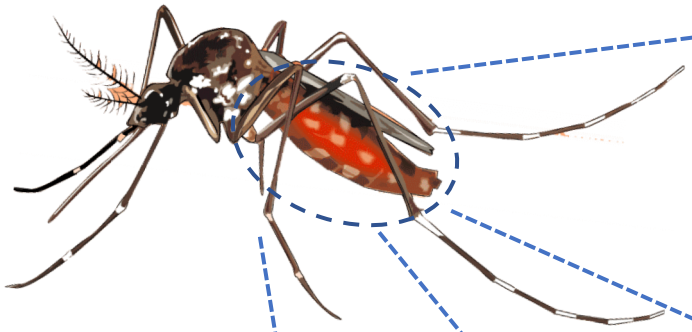
Rhys Parry
Asgari Lab
21st Feb 2019

@RhysHParry

The University Of Queensland
AUSTRALIA

The microbiome

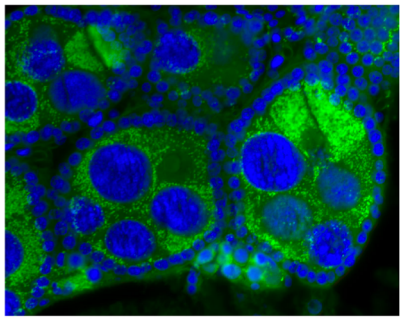
My PhD model; *Ae. aegypti* the yellow fever mosquito



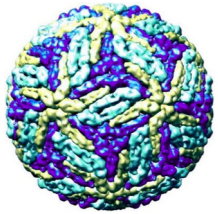
Virome
Commensal viruses

?
?

?
?
?



Arboviruses
Dengue-2



Bacteriome
Endosymbionts
Gut bacteria

The University Of Queensland
AUSTRALIA

@RhysHParry

A glossary of terms

Microbiota

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

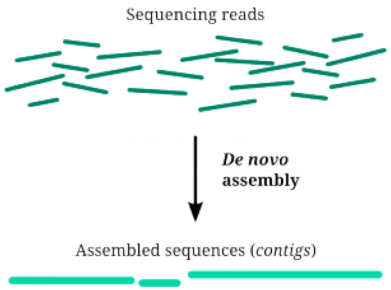
Metagenomics

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

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

@RhysHParry

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

Galaxy Australia

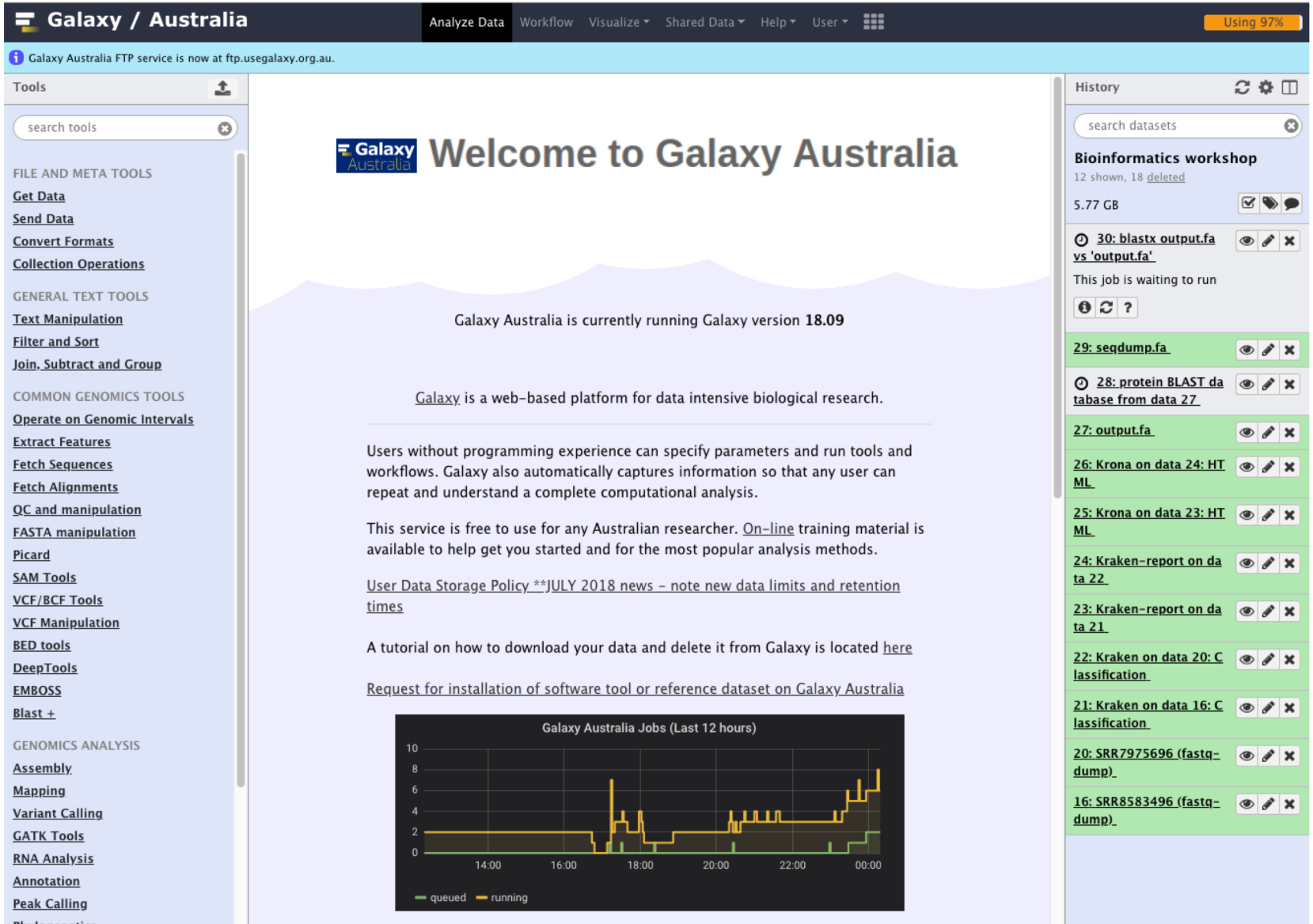
http://www.usegalaxy.org.au

The University Of Queensland AUSTRALIA

@RhysHParry

A plug for Galaxy

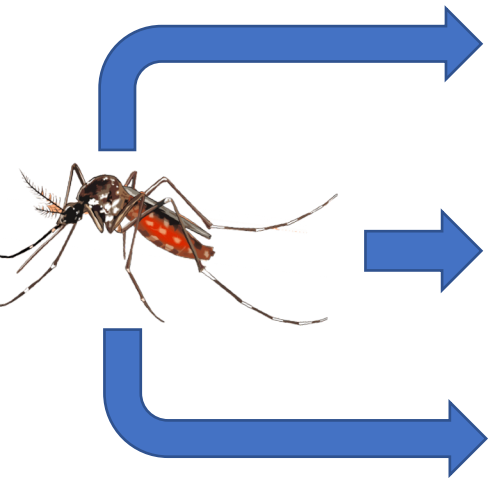
http://www.usegalaxy.org.au



Sequencing for your research question

No “one size fits all” solution for profiling **all** 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 *de novo* 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 *de novo* step Accounting for everything is computationally intensive **Contamination**

The University Of Queensland AUSTRALIA

@RhysHParry

A note on contamination

"Contaminome"

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

Appropriate negative controls must be added

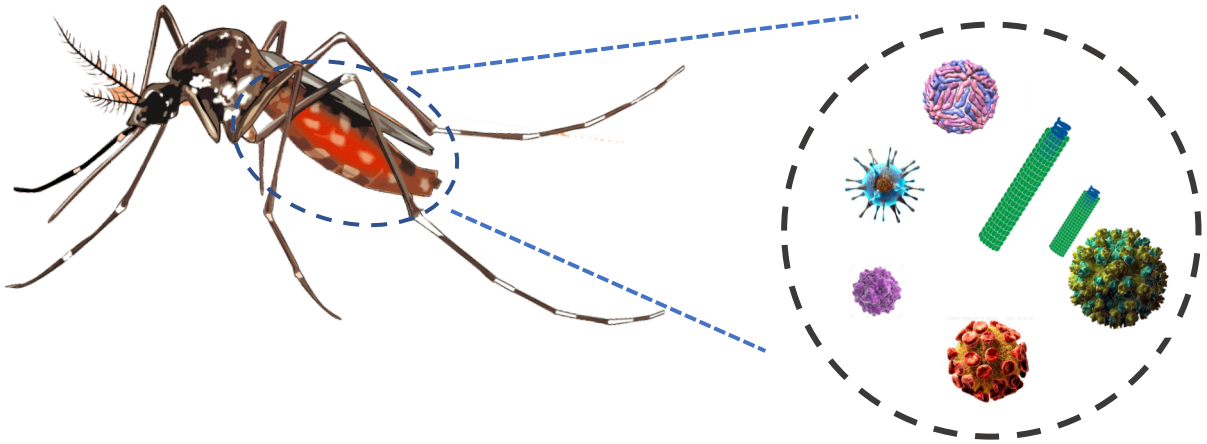
- "Dry" runs of the kit reagents
- "Dry" runs of the spin columns that are being used
- Familiarise yourself with commonly known bacterial contaminants & exclude them from downstream analysis

- 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 next-generation sequencing that hinder discovery of low-abundance microbes. PloS one. 2014 May 16;9(5):e97876

The University Of Queensland AUSTRALIA

@RhysHParry

The “virome”

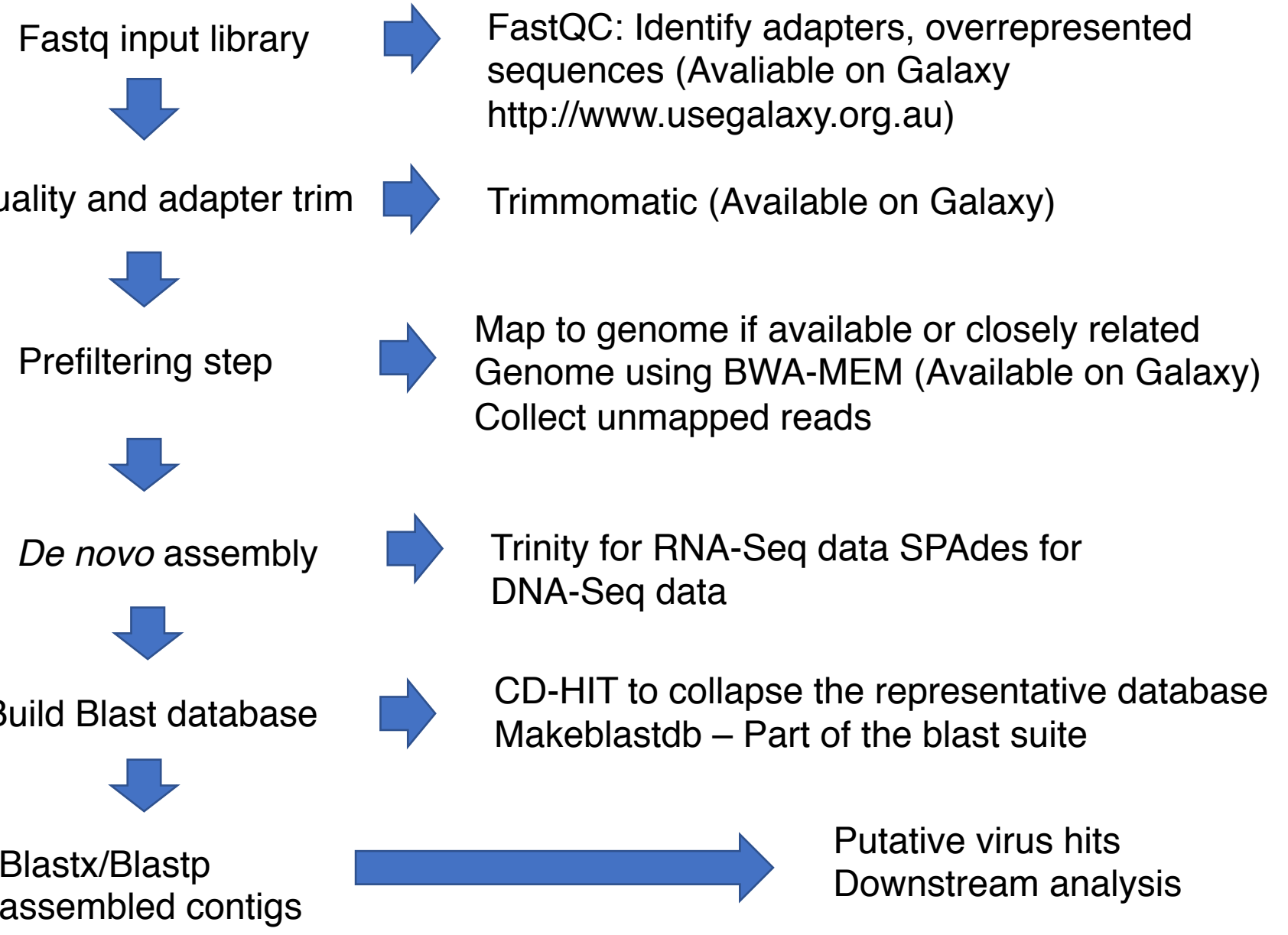


- Lack of redundant techniques to characterise viruses
- Typically biased depending on the RNA/DNA sample you want to send away
- Typically two parts
  - A) Virus discovery – searching for related viruses among assembled contigs, annotation, phylogenetic analysis
  - B) Community profiling, diversity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

The University Of Queensland
AUSTRALIA

@RhysHParry

Virus Discovery



RNA-Seq *de novo* assembly benchmarks

M.musculus RNA-Seq dataset SRX648736

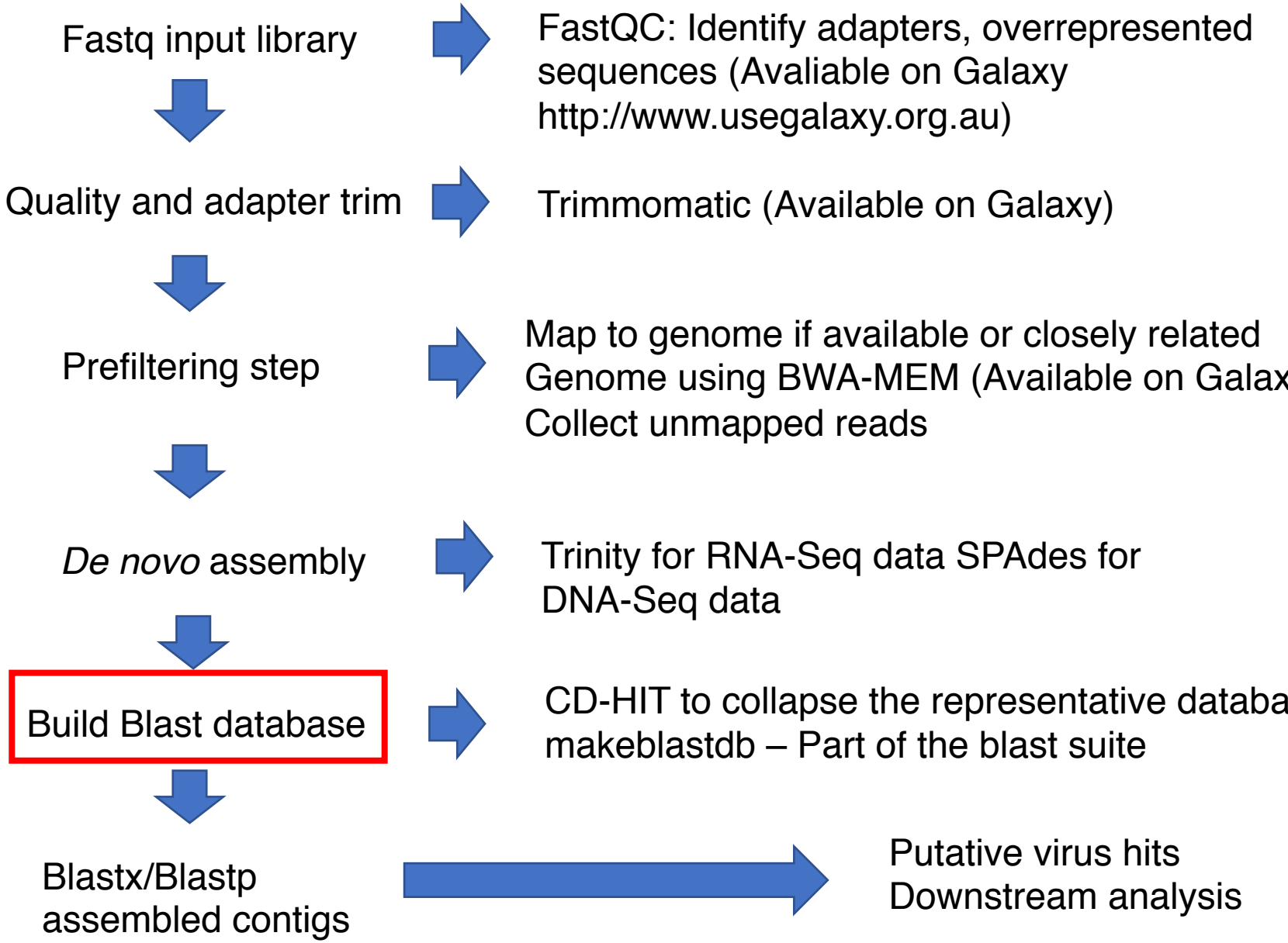
39179 (22740) genes, 94929 (47618) isoforms

	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: http://bioinformaticsinstitute.ru/sites/default/files/denovo_transcriptome_assembly_-_prjibelski_2-dec-2015.pdf

Virus Discovery



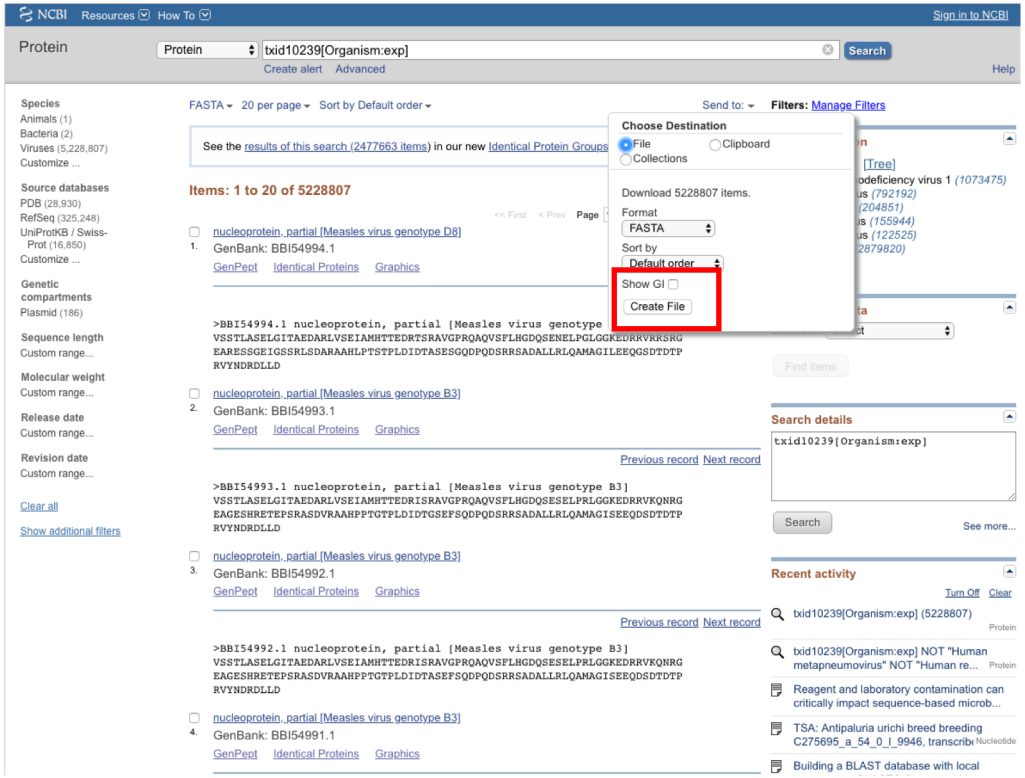
Creating and curating a BLAST database

- Download the proteins you want to compare against deposited in the non-redundant 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 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

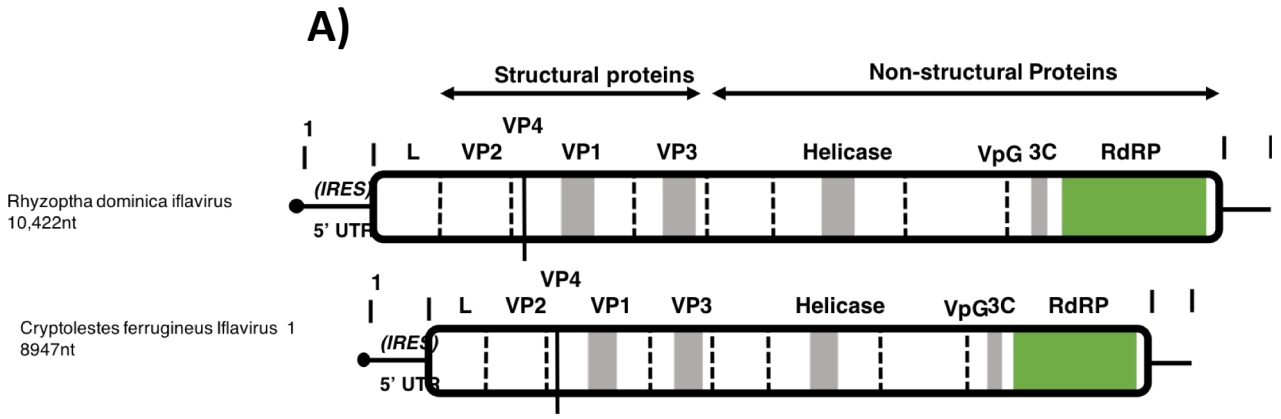
- https://www.ncbi.nlm.nih.gov/protein/



The University Of Queensland
AUSTRALIA

@RhysHParry

Annotation + Composition



Virus annotation:

ORFfinder https://www.ncbi.nlm.nih.gov/orffinder/

Protein domain prediction:

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

Aggregates:

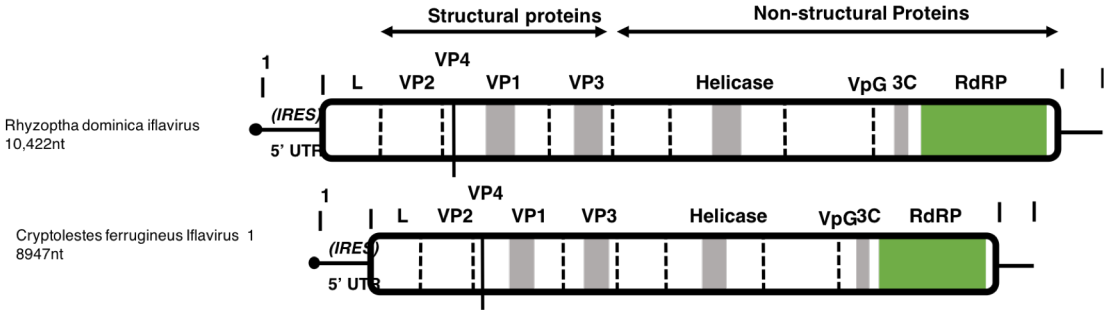
- Pfam protein database
- NCBI-CDD database
- PROSITE Pattern

Annotation + Composition

Composition:

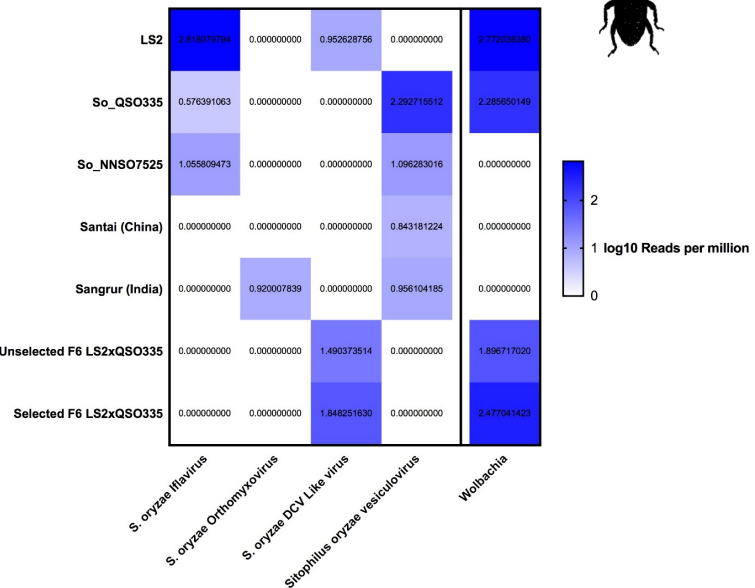
- Re-map the sequencing data to the viruses
- Visualise using heat maps

A)



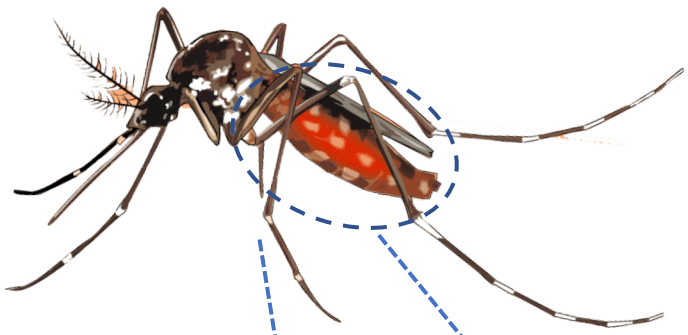
B)

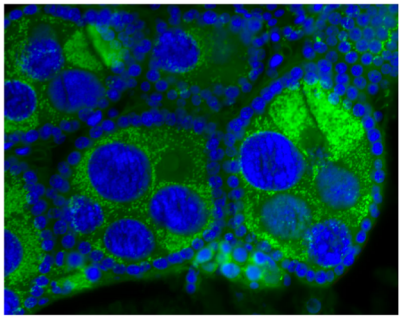
Sitophilus oryzae



Rhyzopertha dominica

Bacterial communities





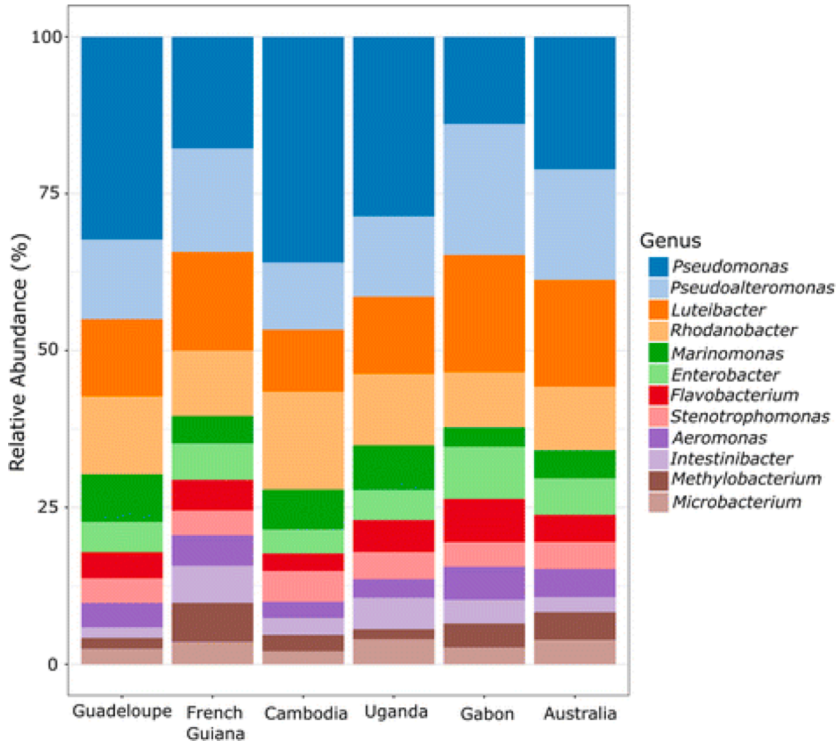
Bacteriome
Endosymbionts
Gut bacteria

- Total shotgun metagenomics
  - Kraken : Uses k-mer matching against a database
  - Krona : Visualises the output from Kraken
- Amplicon sequencing
  - QIIME2 https://qiime2.org/
  - Mothur https://www.mothur.org/
  - MAPseq https://github.com/jfmrod/mapseq
  - MG-Rast https://www.mg-rast.org/
  - MASQUE https://github.com/aghozlane/masque
- Total shotgun metagenomics
  - Kraken
  - QIIME2
- Bacterial databases
  - No need to build your own
  - Many groups have built curated chimera-free databases

The University Of Queensland AUSTRALIA

@RhysHParry

Bacterial communities



- Total shotgun metagenomics
  - Kraken : Uses k-mer matching against a database
  - Krona : Visualises the output from Kraken
- Amplicon sequencing
  - QIIME2 https://qiime2.org/
  - Mothur https://www.mothur.org/
  - MAPseq https://github.com/jfmrod/mapseq
  - MG-Rast https://www.mg-rast.org/
  - MASQUE https://github.com/aghozlane/masque
- Total shotgun metagenomics
  - Kraken
  - QIIME2
- Bacterial databases
  - No need to build your own
  - Many groups have built curated chimera-free databases

The University Of Queensland
AUSTRALIA

@RhysHParry

Bacterial communities

Classifier	Genus precision	Genus sensitivity	Speed (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
MetaPhlAn	n/a	n/a	370770

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

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 OF QUEENSLAND
AUSTRALIA

@RhysHParry

Bacterial communities

Classifier	Genus precision	Genus sensitivity	Speed (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
MetaPhlAn	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

The University Of Queensland AUSTRALIA

@RhysHParry

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	P	1224	Proteobacteria
74.94	22158	934	C	1236	Gammaproteobacteria
38.97	11521	1727	O	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

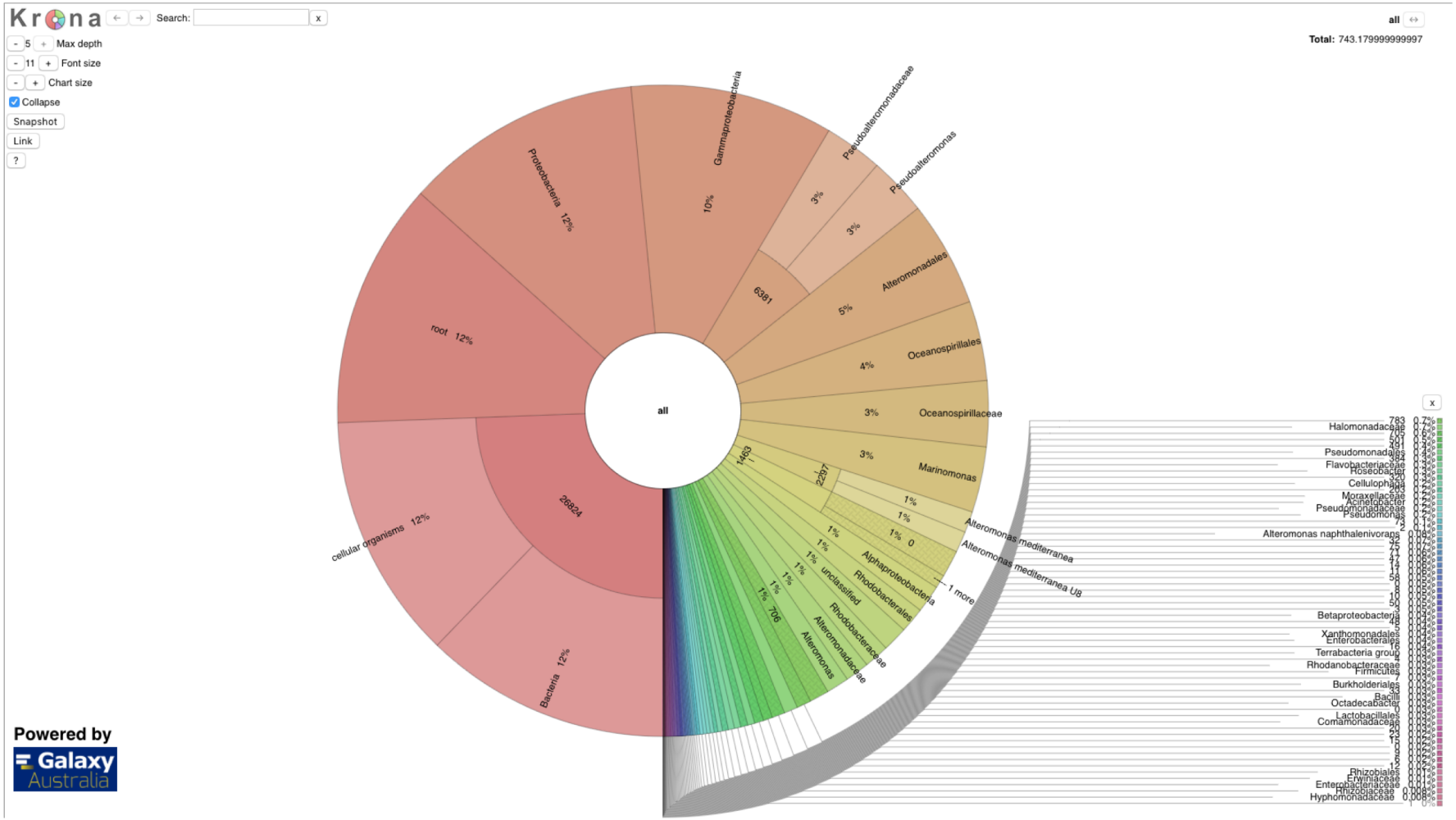
9.26% Unclassified

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

@RhysHParry

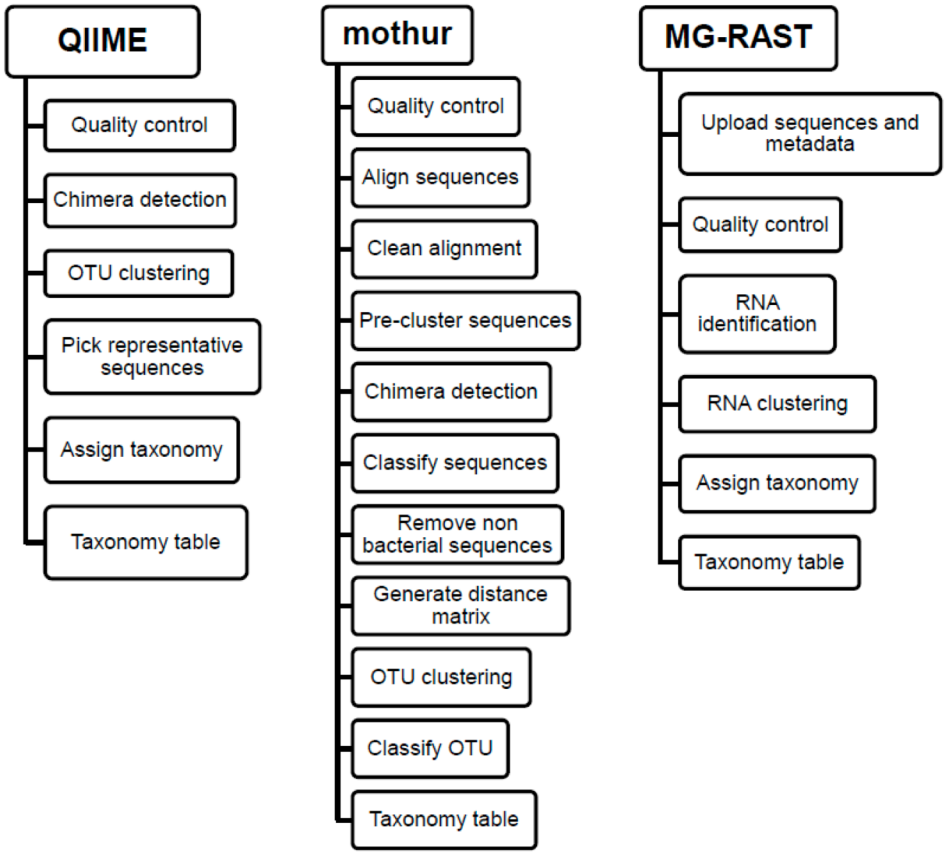
Bacterial communities

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

Bacterial communities



- 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.

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
@RhysHParry

Comparisons

	QIIME	mothur	MG-RAST
Website	http://qiime.org/	http://www.mothur.org/	http://metagenomics.anl.gov
Web-based interface	YES (http://www.n3phele.com/) Not supported/maintained by the QIIME team	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 compatibility	Illumina, 454, Sanger, Ion Torrent, PacBio	Illumina, 454, Sanger, Ion Torrent, PacBio	Illumina, 454, Sanger, Ion Torrent, PacBio
16S rRNA gene Databases searched	RDP, SILVA, Greengenes and custom databases	RDP, SILVA, Greengenes and custom databases	M5RNA, RDP, SILVA and Greengenes
Alignment Method	PyNAST, MUSCLE, INFERNAL	Needleman-Wunsch, blastn, gotoh	BLAT
Taxonomic analysis/assignment	UCLUST, RDP, BLAST, mothur	Wang/RDP approach	BLAT
Clustering algorithm	UCLUST, CD-HIT, mothur, BLAST	mothur, adapts DOTUR and CD-HIT	UCLUST
Diversity analysis	alpha and beta	alpha and beta	alpha
Phylogenetic Tree	FastTree	Clearcut algorithm	YES
Chimera detection	UCHIME, chimera slayer, BLAST	UCHIME, chimera slayer, and more	No
Visualisation	PCA plots, OTU networks, bar plots, heat maps	Dendrograms, heat maps, Venn diagrams, bar plots, PCA plots	PCA plots, heat maps, pie charts, bar plots, Krona and Circos for visualisation

The University Of Queensland AUSTRALIA

@RhysHParry

A note on databases

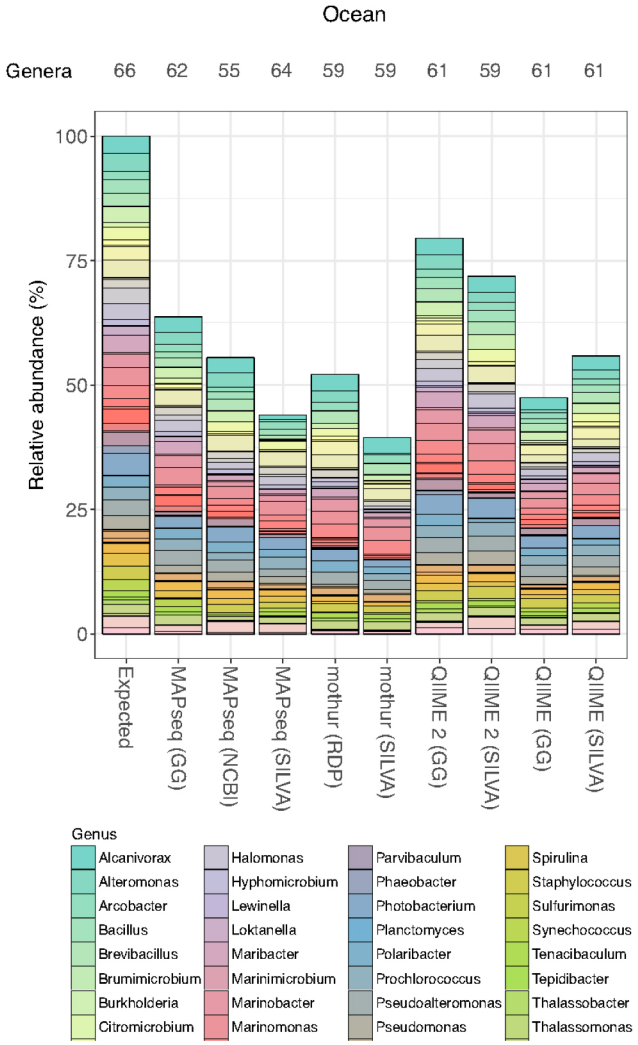
- You are only as good as your database

- Green genes http://greengenes.secondgenome.com/
- 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) https://rdp.cme.msu.edu/
- 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 https://www.arb-silva.de/
- 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?



- **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.

The University Of Queensland AUSTRALIA

@RhysHParry

Computational requirements

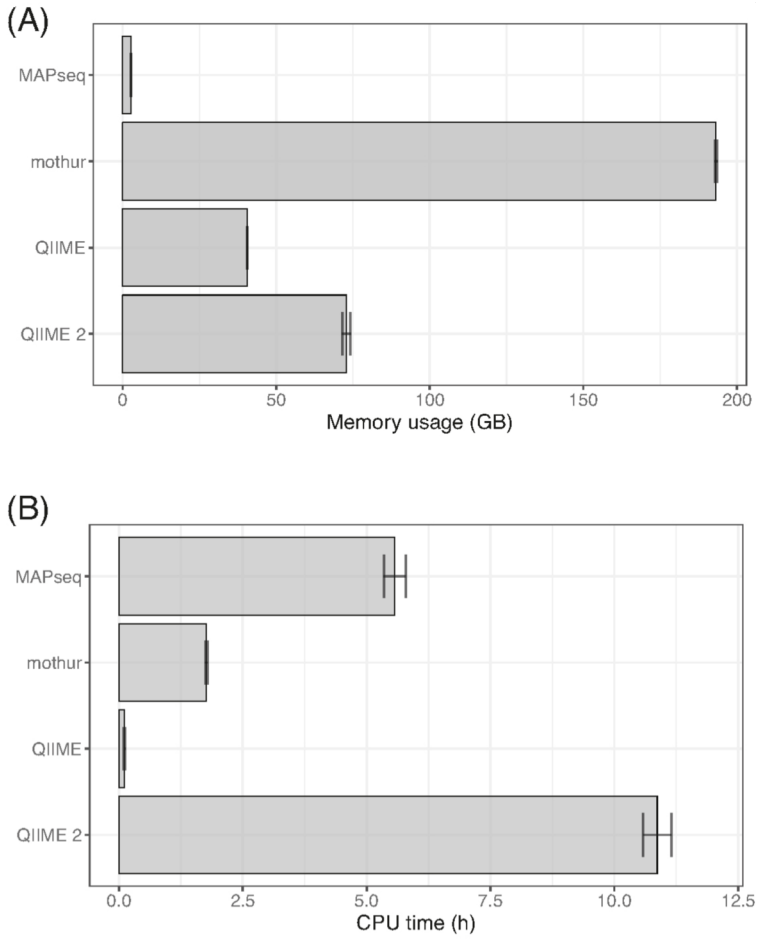


Figure adapted from:
Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. GigaScience. 2018 May 11;7(5):giy054.

The University Of Queensland
AUSTRALIA

@RhysHParry

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/

The University Of Queensland
AUSTRALIA

@MrRhysParry

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: http://weizhongli-lab.org/cdhit_suite/cgi-bin/index.cgi)

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 + (https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi 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 build workable protein database

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"