GENOME RESEARCH

A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples

Naccache *et al.* 2014

Journal club

Mihkel Vaher

Background

- Nearly all microorganisms can be uniquely identified on the basis of their specific nucleic acid sequence, requirements:
  - Sufficiently long read lengths
  - Multiple hits to the microbial genome
  - Well-annotated reference database

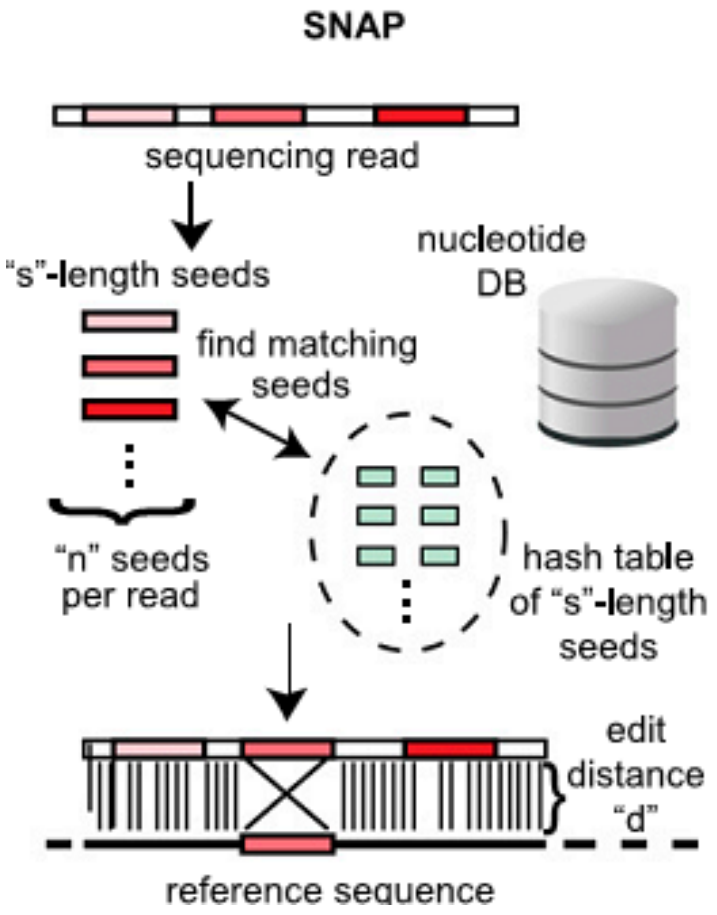
- NGS laboratory workflows incur minimum turnaround of 8 h from clinical sample to sequence
  - Subsequent computational analyses of NGS data should be performed within a timeframe suitable for actionable responses in clinical medicine and public health (i.e., minutes to hours)
  - Pipeline must retain sensitivity, accuracy
- Sparse reads often do not overlap sufficiently to permit *de novo* assembly
  - Individual reads 100–300 nucleotides (nt) in length, must be classified to a high degree of accuracy.

SURPI - *sequence-based ultrarapid pathogen identification*

- *Fast* and *comprehensive* mode
- Generates results in a clinically actionable timeframe of minutes to hours
- Main components: two alignment tools SNAP and RAPSearch
- Pathogen detection using both in silico-generated and clinical data

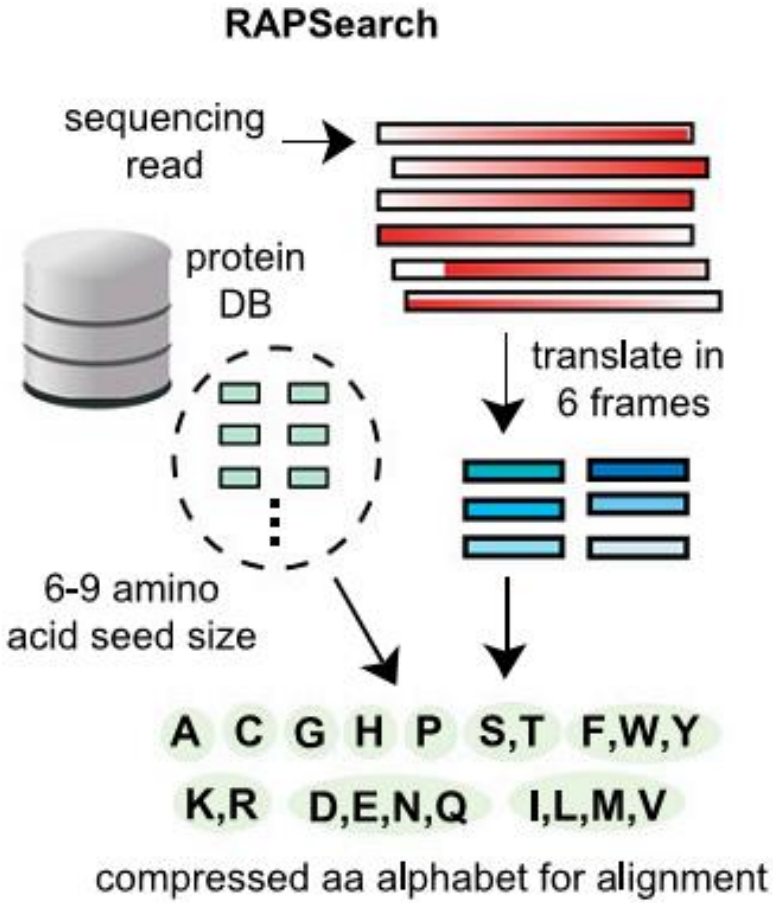
SNAP

- **Read lengths have increased since BLAST**: from 25–30 base pair reads to 100 bp or more.
  - Lets use longer seeds
  - **Fewer "false positive" locations by chance**.
- 10–50× speedup over the textbook O(n^2) edit distance check due to quicker rejection
- Leverages the higher memory capacities on today's servers to index more seeds and perform fewer hash lookups.
- **Align human genome 100 bp read dataset with 30-fold coverage in 20 minutes on a 32-core server**



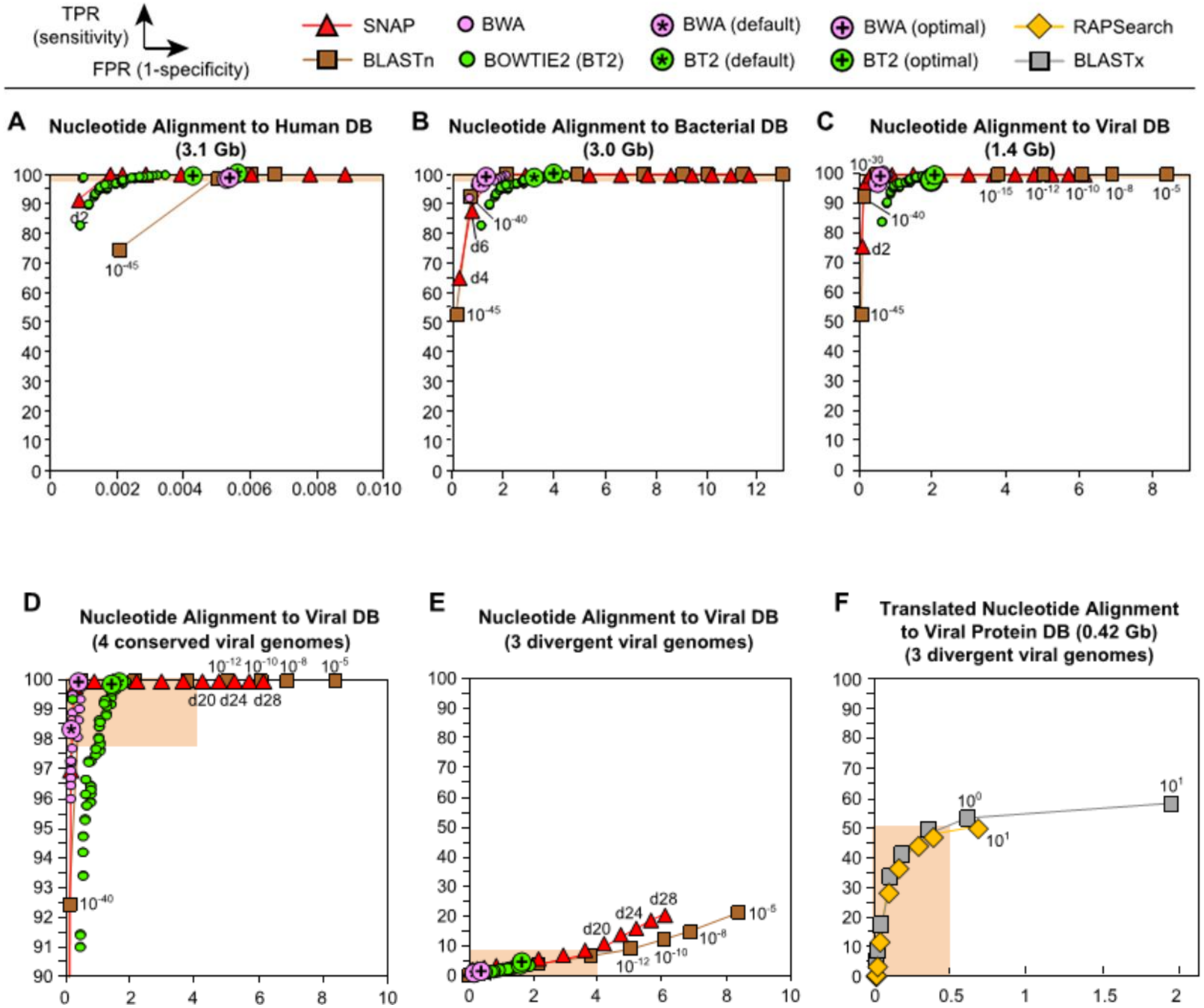
RAPSearch

- ~20–90-fold speedup relative to BLAST with similar of sensitivity
- Output same format as BLAST
- In SURPI:
  - Novel microorganisms with divergent genomes
  - Often viruses can only be identified on the basis of remote amino acid homology

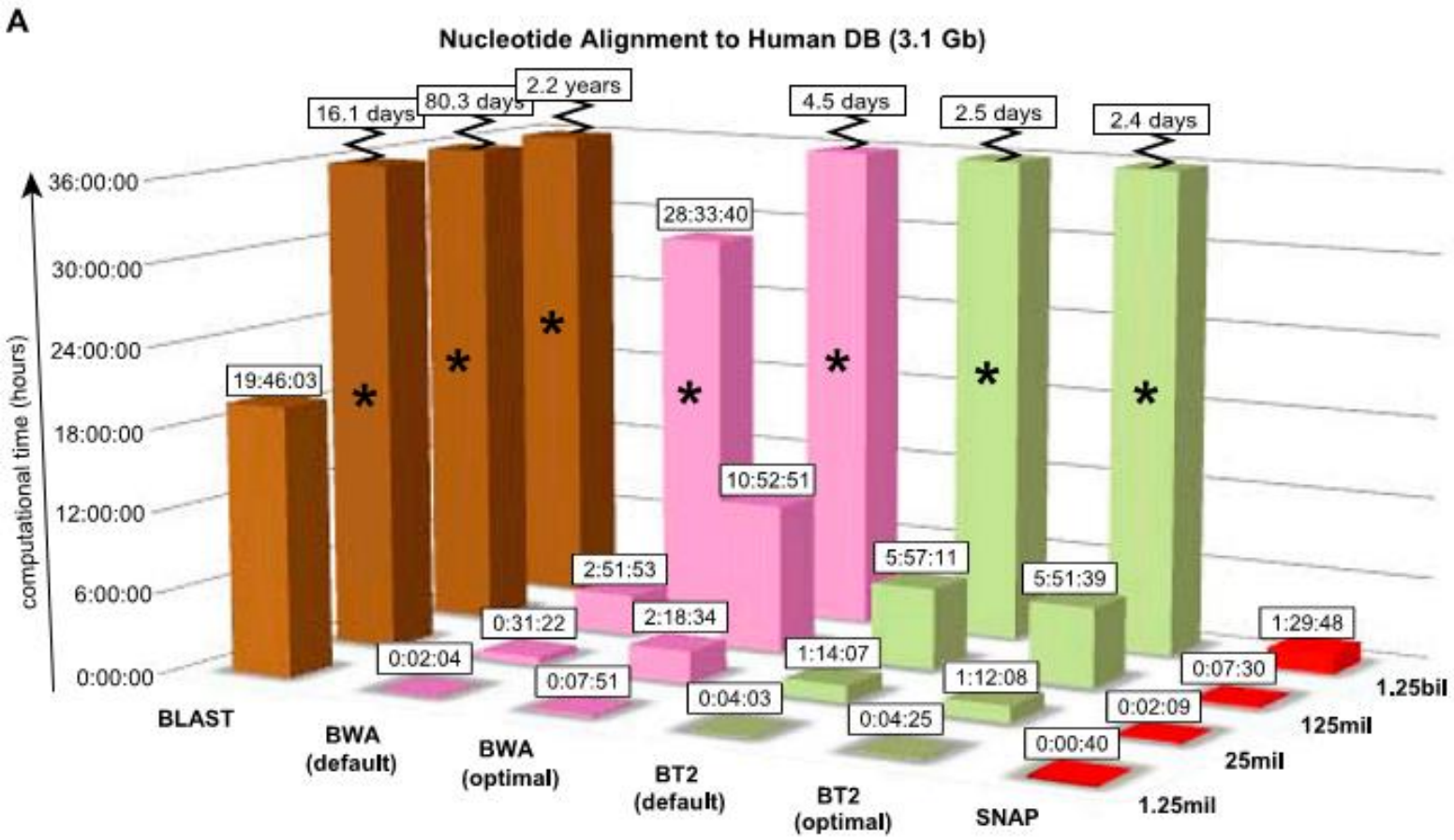


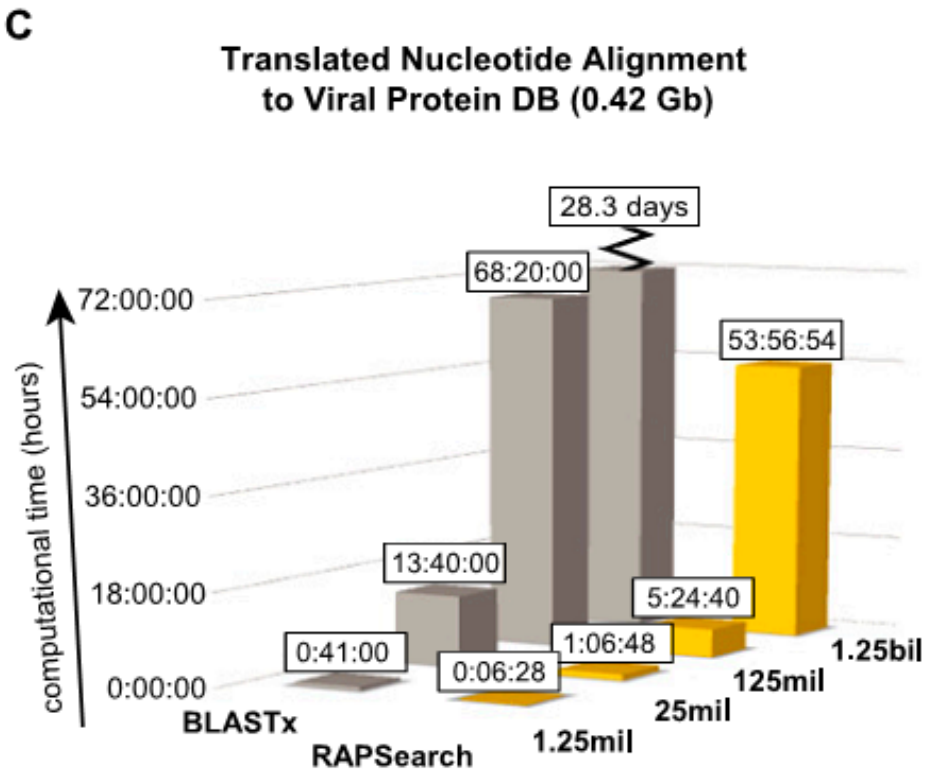
In silico-generated data aligner accuracy

- Randomly generated 100 base pair (bp) reads
  - 1 million human reads
  - 250,000 bacterial reads
  - 25,000 viral reads
  - 1000 reads each from four known viruses three divergent ‘‘novel’’ viruses (genomes removed a priori from the reference database)
- All aligners performed well at detecting known reads but
- Poorly in detecting divergent viral reads
  - The need for translated nucleotide alignment algorithms such as RAPSearch and BLASTx


TPR
(sensitivity)
FPR (1-specificity)
SNAP
BLASTn
BWA
BOWTIE2 (BT2)
BWA (default)
BT2 (default)
BWA (optimal)
BT2 (optimal)
RAPSearch
BLASTx
A
Nucleotide Alignment to Human DB
(3.1 Gb)
B
Nucleotide Alignment to Bacterial DB
(3.0 Gb)
C
Nucleotide Alignment to Viral DB
(1.4 Gb)
D
Nucleotide Alignment to Viral DB
(4 conserved viral genomes)
E
Nucleotide Alignment to Viral DB
(3 divergent viral genomes)
F
Translated Nucleotide Alignment
to Viral Protein DB (0.42 Gb)
(3 divergent viral genomes)

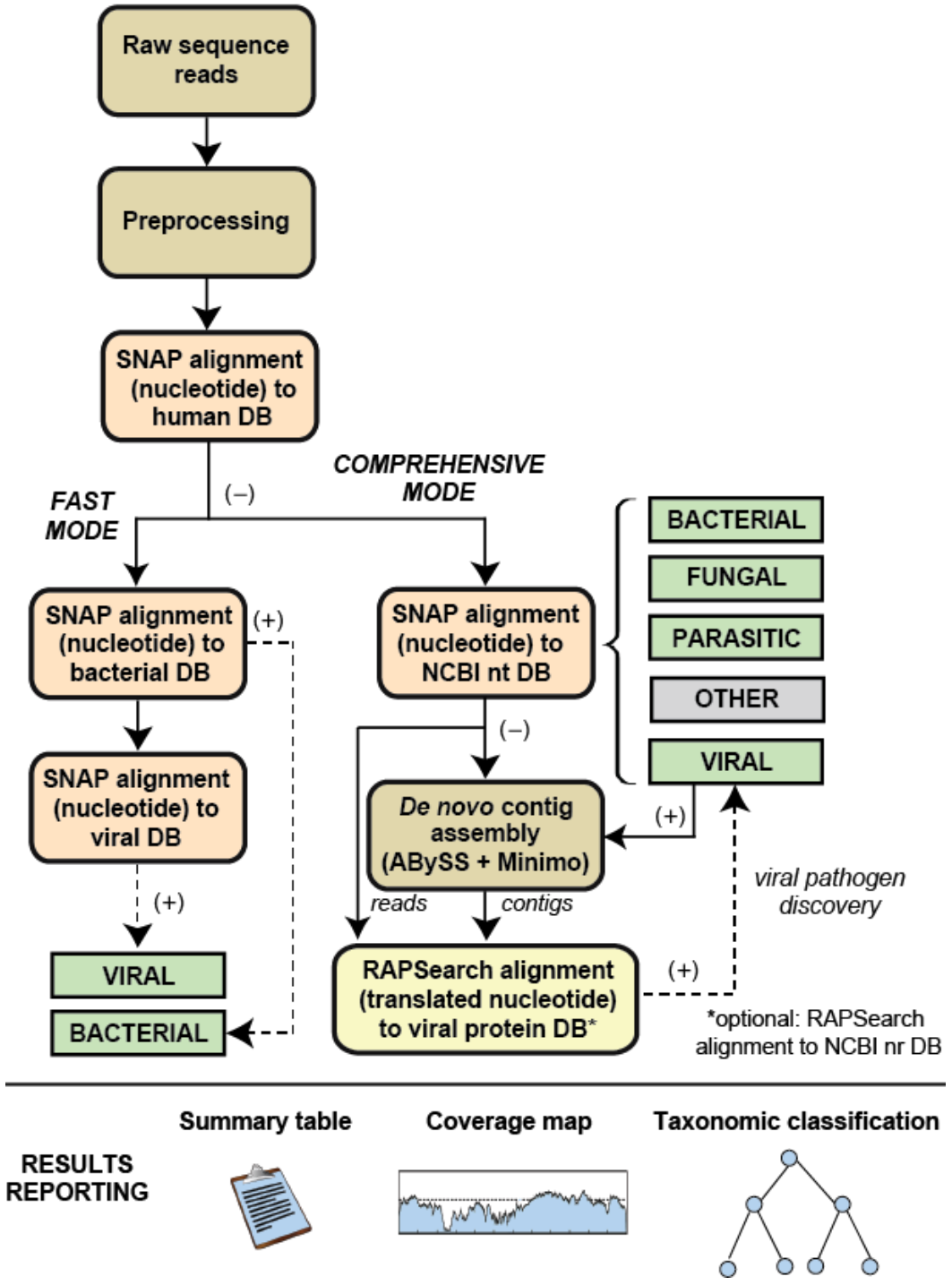
Speed of SNAP and RAPSearch





SURPI pipeline

- FASTQ files as input, recognizes the presence of multiple barcodes used for indexing.
- Shell, Python and Perl scripts
- Fixed external software and database dependencies
- Open-source tools including
  - SNAP and
  - RAPSearch aligners



Preprocessing

- Trimming low-quality and adapter sequences using cutadapt, retaining reads of trimmed length >50 bp
- Removing low-complexity sequences using the DUST algorithm in PRINSEQ
- Normalizing read lengths for SNAP alignment by cropping reads of length >75 to 75 bp
- Remove human reads (SNAP)

Fast mode (nucleotide seqs)

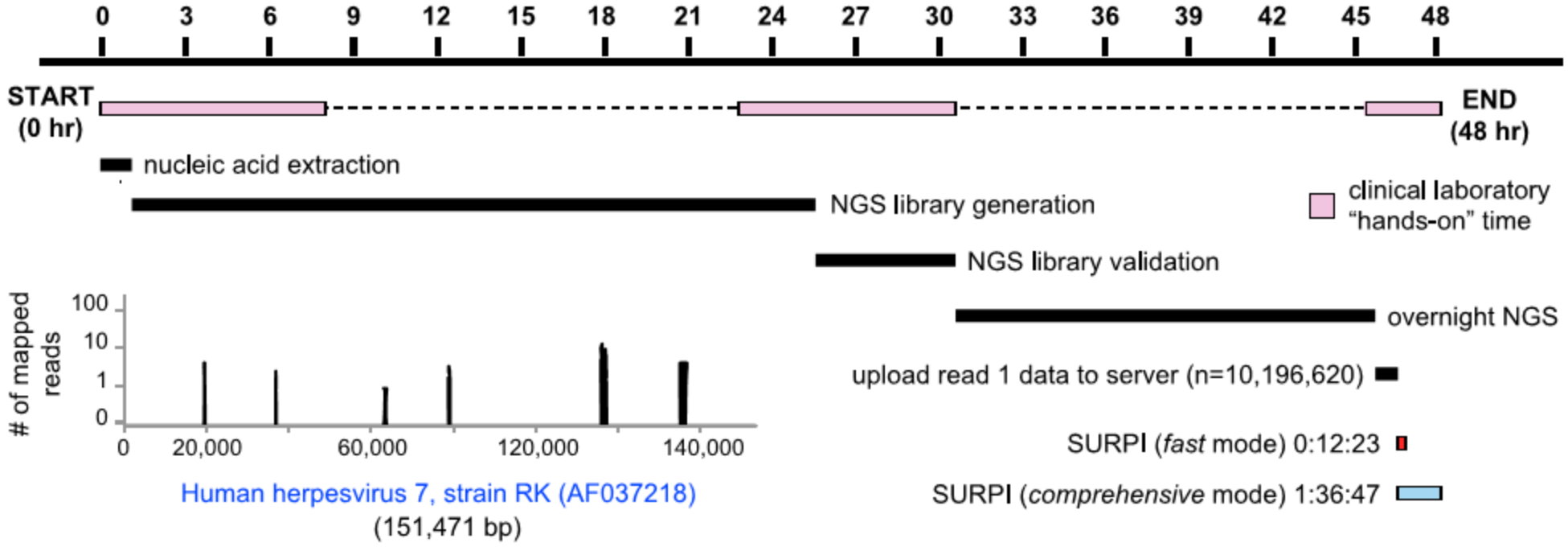
- Viruses and bacteria are identified by SNAP alignment to viral and bacterial nucleotide databases.

Comprehensive mode (nucleotide + protein seqs)

- Reads are aligned using SNAP to all nucleotide sequences in the NCBI nt collection
- Unclassified reads and contigs generated from *de novo* assembly
- Alignment to a viral protein database using RAPSearch for pathogen discovery of divergent viruses
- For each barcode, the best coverage map for each viral genus identified in the dataset is generated.

Speed of SURPI and feasibility for real-time clinical analysis

- 20-yr-old female patient, 3 d of fever to ~~101.5 °C~~ (38.6 °C), myalgias and headache from hiking in a region of Australia endemic for the mosquito-borne Ross River and Barmah Forest alphaviruses
- Within a 48-h sample-to-answer turnaround time and 13 min SURPI analysis time, sequences spanning the genome of human herpesvirus 7 (HHV-7) were detected.
- Subsequent PCR supported a diagnosis of primary HHV-7 infection



Output

- List of all classified reads with taxonomic assignments
- Table of read counts
- Both viral and bacterial genomic coverage maps

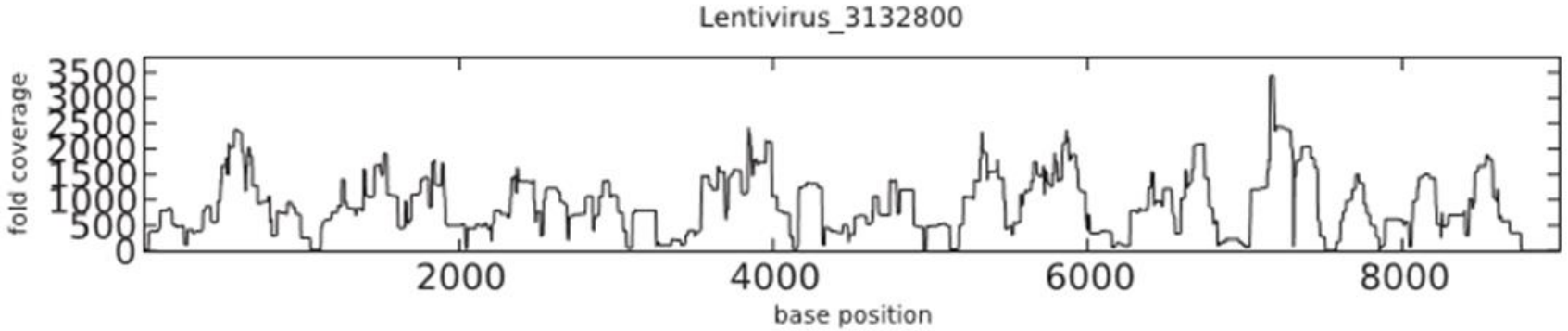
By nucleotides (containing HPV-18)

Species	Genus	Family	Index #0 (prostate cancer tissue)
Human herpesvirus 5	Cytomegalovirus	Herpesviridae	1
Murid herpesvirus 1	Muromegalovirus	Herpesviridae	1
Suid herpesvirus 1	Varicellovirus	Herpesviridae	1
Human herpesvirus 1	Simplexvirus	Herpesviridae	2
Alphapapillomavirus 7	Alphapapillomavirus	Papillomaviridae	11955
Human papillomavirus		Papillomaviridae	30
Murine xenotropic virus NZB		Retroviridae	1
Human immunodeficiency virus 1	Lentivirus	Retroviridae	3
Murine leukemia virus	Gammaretrovirus	Retroviridae	31
Murine leukemia-related retroviruses	Gammaretrovirus	Retroviridae	5
Enterobacteria phage phi80	Lambdalikevirus	Siphoviridae	1

By translated nucleotides (containing HPV-18)

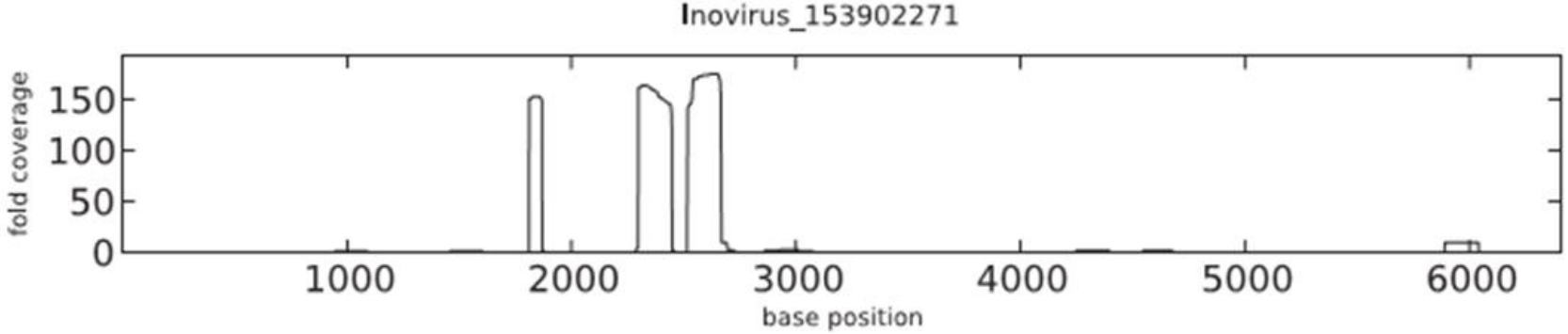
Species	Genus	Family	Reads #80
Alphapapillomavirus 7	Alphapapillomavirus	Papillomaviridae	2

Coverage maps



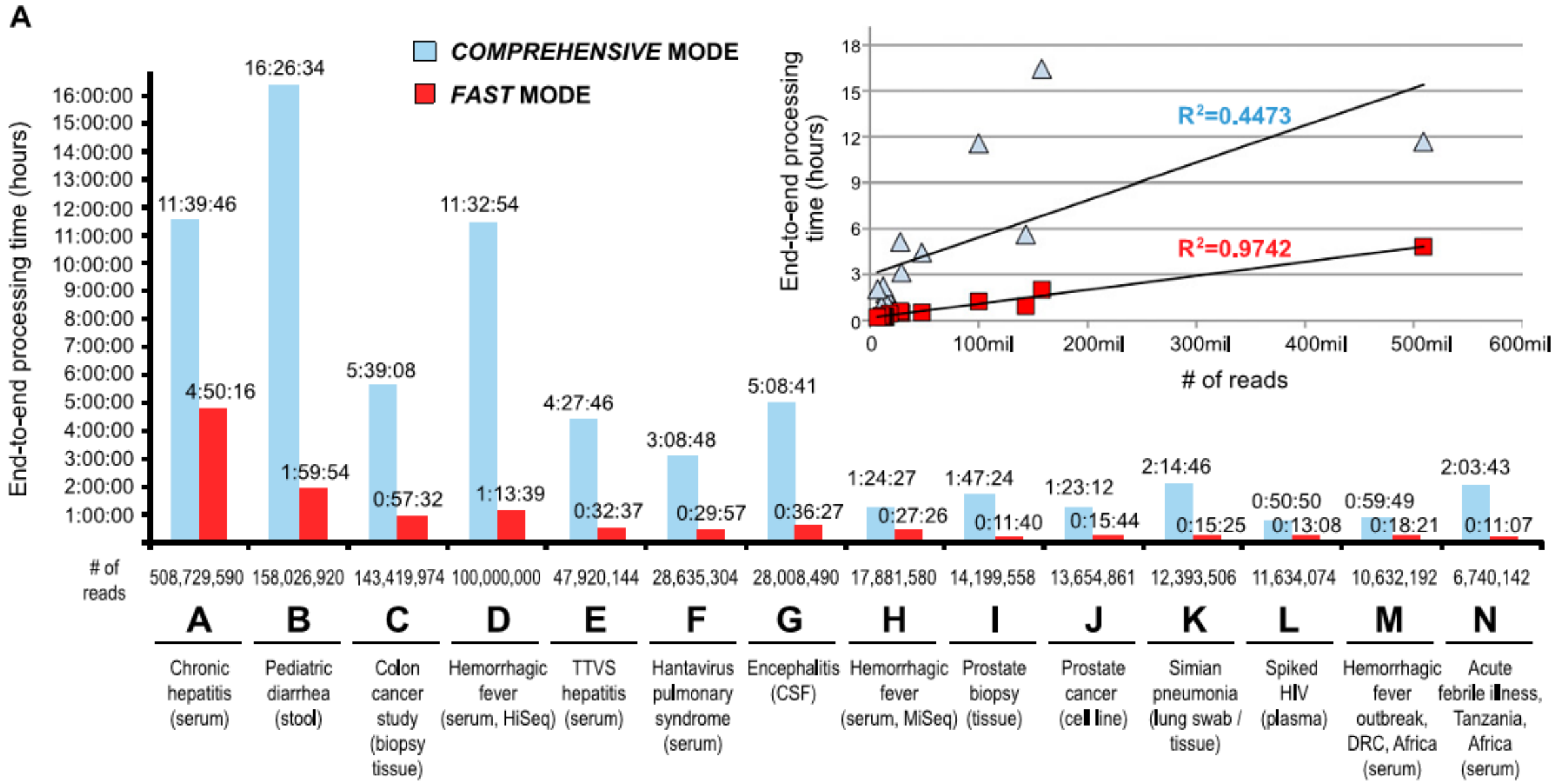
mapping Lentivirus.bar.#GCCAAT@_.HIV.NT.snap.matched.Viruses
against Lentivirus.3132800.fasta with gi definition
gi|3132800|gb|AF063223.1| HIV-1 isolate DJ263 from Djibouti, complete genome

Reference sequence length = 9002 bp
Coverage in bp = 8970
%Coverage = 99.644523
Average depth of coverage = 892.990335
Number of reads contributing to assembly = 54537



mapping Inovirus.bar.#GCCAAT@_.HIV.NT.snap.matched.Viruses
against Inovirus.153902271.fasta with gi definition
gi|153902271|dbj|AB334721.1| Enterobacteria phage f1 DNA, complete genome, isolate: NBRC 20015

Reference sequence length = 6407 bp
Coverage in bp = 1422
%Coverage = 22.194474
Average depth of coverage = 9.526611
Number of reads contributing to assembly = 358


A
COMPREHENSIVE MODE
FAST MODE
End-to-end processing time (hours)
16:00:00
15:00:00
14:00:00
13:00:00
12:00:00
11:00:00
10:00:00
9:00:00
8:00:00
7:00:00
6:00:00
5:00:00
4:00:00
3:00:00
2:00:00
1:00:00
11:39:46
4:50:16
16:26:34
1:59:54
5:39:08
0:57:32
11:32:54
1:13:39
4:27:46
0:32:37
3:08:48
0:29:57
5:08:41
0:36:27
1:24:27
0:27:26
1:47:24
0:11:40
1:23:12
0:15:44
2:14:46
0:15:25
0:50:50
0:13:08
0:59:49
0:18:21
2:03:43
0:11:07
# of reads
508,729,590
158,026,920
143,419,974
100,000,000
47,920,144
28,635,304
28,008,490
17,881,580
14,199,558
13,654,861
12,393,506
11,634,074
10,632,192
6,740,142
A
Chronic hepatitis (serum)
B
Pediatric diarrhea (stool)
C
Colon cancer study (biopsy tissue)
D
Hemorrhagic fever (serum, HiSeq)
E
TTVS hepatitis (serum)
F
Hantavirus pulmonary syndrome (serum)
G
Encephalitis (CSF)
H
Hemorrhagic fever (serum, MiSeq)
I
Prostate biopsy (tissue)
J
Prostate cancer (cell line)
K
Simian pneumonia (lung swab / tissue)
L
Spiked HIV (plasma)
M
Hemorrhagic fever outbreak, DRC, Africa (serum)
N
Acute febrile illness, Tanzania, Africa (serum)
End-to-end processing time (hours)
18
15
12
9
6
3
0
R²=0.4473
R²=0.9742
0
100mil
200mil
300mil
400mil
500mil
600mil
# of reads

Work progress

- Monitoring the log file
- Send notifications via Twitter at various stages within the pipeline

Processing time and cost of SURPI (SNAP, RAPSearch) vs PathSeq (blastn, blastx) on cloud server

Dataset	Number of reads	SURPI (fast mode)		SURPI (comprehensive mode)		PathSeq	
		Time	Cost	Time	Cost	Time	Cost
Prostate cancer (cell line)	13,654,861	0:19:47	$3.10	1:03:22	$6.20	18:50:00	$91.60
TTVS hepatitis (serum)	47,920,144	0:57:09	$3.10	4:28:55	$15.50	27:25:04	$134.24
Pediatric diarrhea (stool)	158,026,920	2:50:17	$9.30	14:43:12	$43.40	~9.2 days	$1,082.42

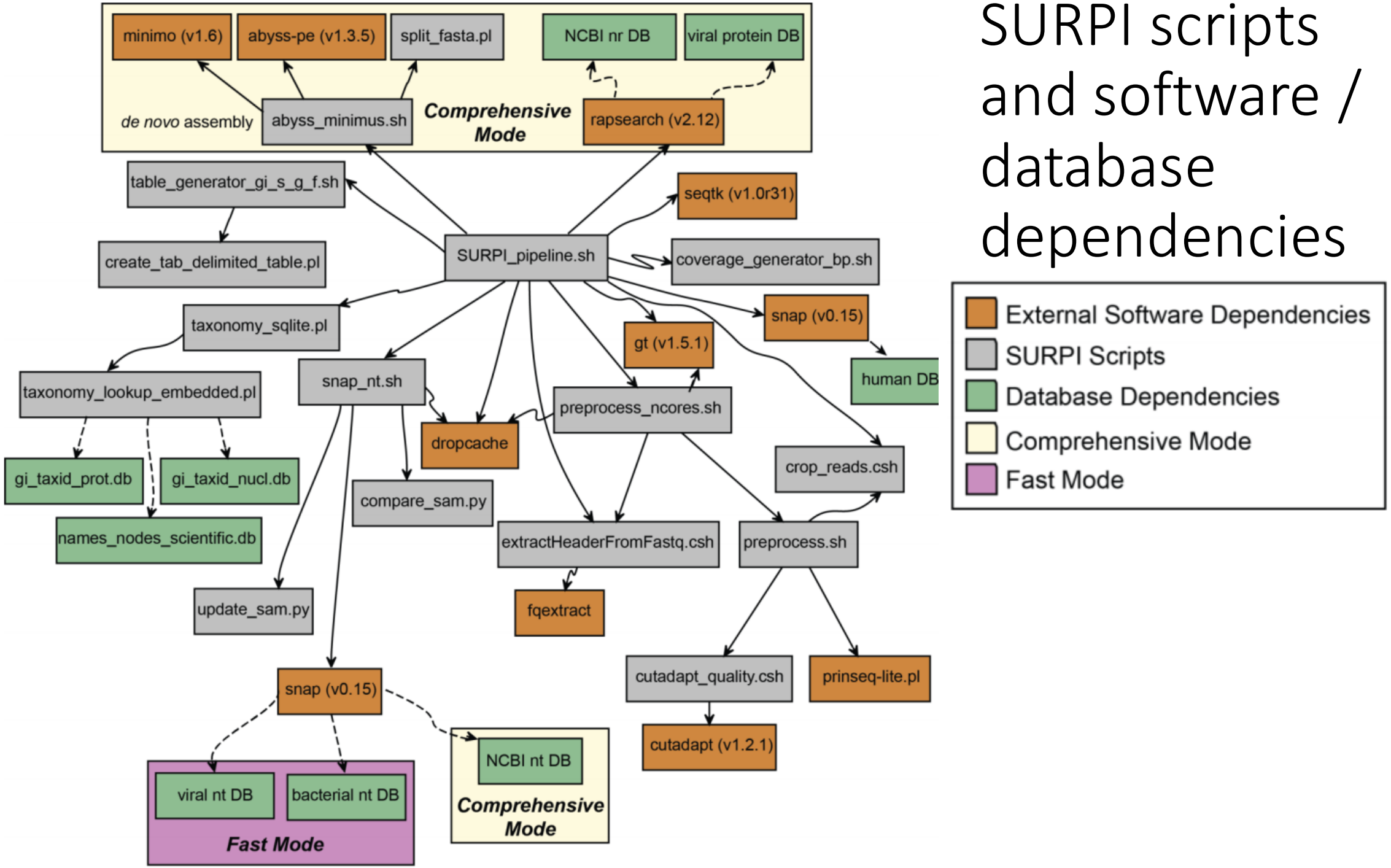
Amazon Web Services - Pay only for what you use. There is no minimum fee.

Troubleshooting

1.The create_snap_to_nt.sh program uses -Ofactor as 1000, on line 29, which may not work for your machine. You need to figure out the correct value and make necessary changes. Read snap aligner document for details.
2.The abyss instalation requires mmap. Make sure you have installed it before compiling abyss. http://hackage.haskell.org/package/mmap-0.5.9/mmap-0.5.9.tar.gz
3.Make sure formatdb is there in your path. It can be downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/
4.The taxonomy_lookup.pl program, at line 84 has sort --parallel=$cores, where you may need to remove --parallel=$cores option, if the sort utility on you machine does not support --parallel option.
5.The abyss_minimus.sh program tries to use mpirun to make it parallel. If the mpirun is not configured properly, you need to remove the option 'np=$cores' in line 86, so that it will not be run parallelly.
6.The ribo_snap_bac_euk.sh program is hardcoded to use the 10,75 as arguments to crop_reads.csh, which you may need to change in line 43.
7.The coveragePlot.py program uses mlab.load() at line 47, which is deprecated in latest version of matplotlib. Hence, you may need to change it to np.loadtxt()

https://www.biostars.org/p/118719/

SURPI scripts and software / database dependencies



Conclusion

Thanks!