Endovir

A pipeline to identify endogenous viruses in SRA datasets

https://github.com/NCBI-Hackathons/EndoVir

Jan P. Buchmann

NCBI, NLM, NIH, Bethesda, MD, USA
The University of Sydney, Sydney, Australia

2018-01-09

Foreword

Based on previous work

- ViruSpy [1]
- Logic mostly implemented in Bash and Perl
  → Complex to expand and adjust
- Initiated during my stay as a visiting bioinformatician as NCBI from Oct.2017 to Dec. 2017.
- Ongoing

---

[1] https://github.com/NCBI-Hackathons/ViruSpy

Foreword

Standing on the shoulder of giants

- Karina Zile, karina.zile@gmail.com
- Mitchell A. Ellison II, mae92@pitt.edu
- Jacob Waldman, jaw267@pitt.edu
- Kristyna Kupkova, kupkova@feec.vutbr.cz
- Cody Glickman, cody.glickman@ucdenver.edu
- Andrew Clugston, andrew.clugston@pitt.edu
- Paul G. Cantalupo, pcantalupo@gmail.com
- Vineet Raghu, vineet@cs.pitt.edu

Exogenous viruses

- Integration into host genome not required for replication
- Hist genome integration occurs occasionally, e.g. Human Herpes virus 6
- Not part of the host genome
- RNA viruses: $\leq \approx$ 40 kb
- DNA viruses: $\leq \approx$ 2 Mbp

Rabies virus

$3'$ — [N] — [M] — [G] — [L] — $5'$

11 kb
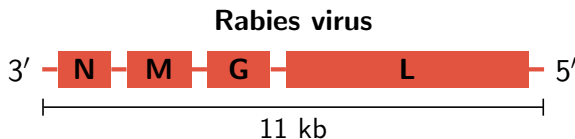
Endogenous viral element (EVEs)
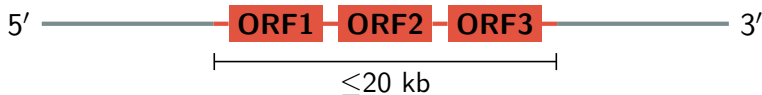
- Integration into host genome mandatory for replication
- Retroviridae, e.g. HIV
- Part of the host genome
- Not widely analyzed: complex analysis and missing data [2]
- Allowing to date virus evolution
- Characterize the mechanisms that underpin the integration of viruses into host genomes

---

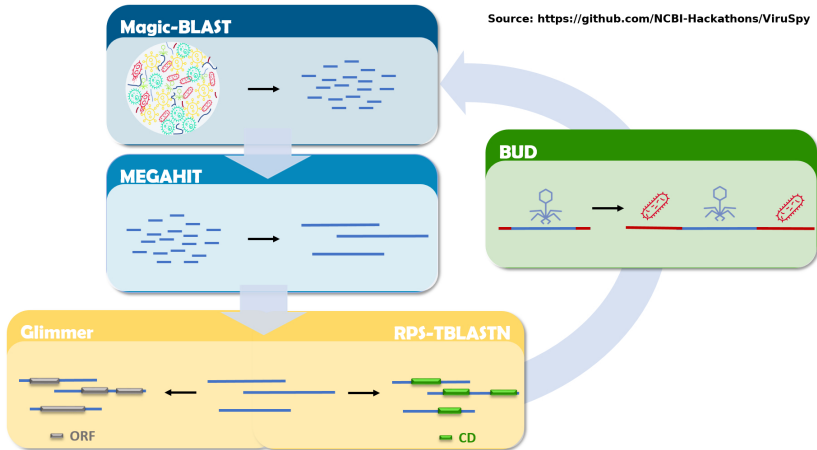[2]Katzourakis, A. and Gifford, R.J. (2010). PLos Genet., 6, e1001191

SRA data: a treasure trove for virus genomics

- ▶ SRA data contains host and virus sequences
- ▶ Exo/Endogenous and putative novel viruses

Virus reads

Host/Sample reads
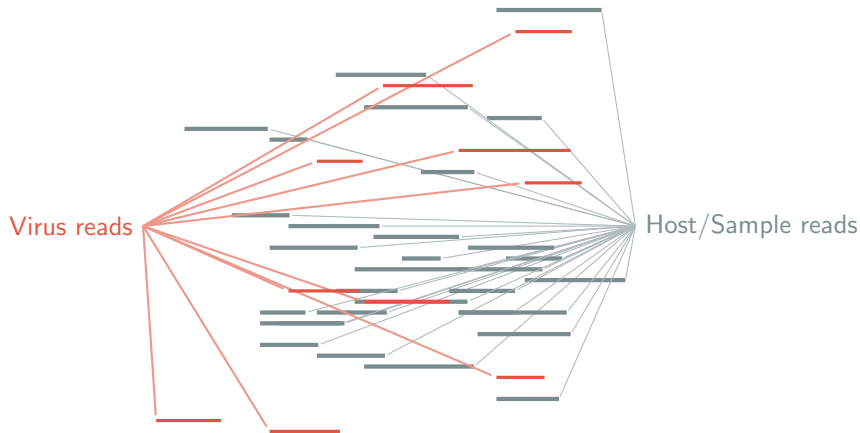
BUD algorithm

▶ "Building Up Domains"

Source: https://github.com/NCBI-Hackathons/ViruSpy

Endovir

Open source	MIT license
Modern	Implemented in Python 3 (3.6)
Independent	Using only Python standard libraries
Limited disk I/O	Using pipes where possible
Modularity	Changing external tools, e.g. assembler, without touching Endovir internals
Status	Alpha, ongoing development

Biological databases

Endovir uses two biological databases:

Sequences Virus genomes sequences from RefSeq

Domains virus-specific Cdd database. We provide a setup
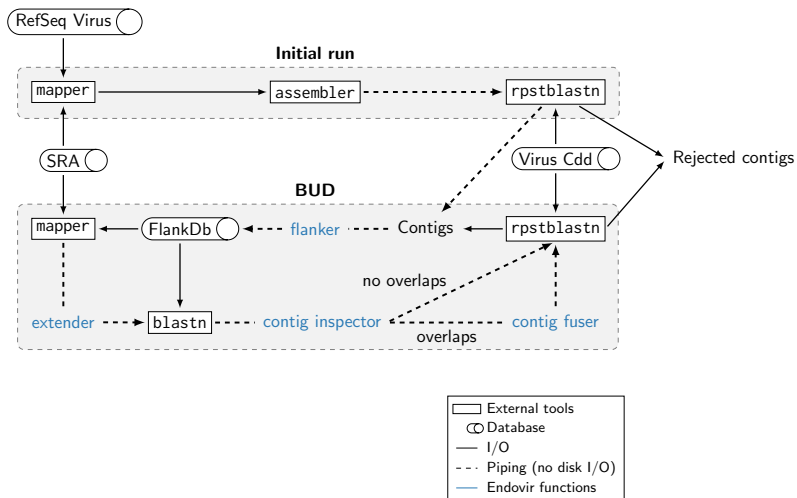script to create the Cdd database

Usage

```
usage: endovir.py [-h] [-srr SRR] [--wd WD] [--max_cpu MAX_CPU]

Endovir

optional arguments:
  -h, --help            show this help message and exit
  -srr SRR              SRR number, e.g. SRR5150787
  --wd WD               Working directory for analysis
  --max_cpu MAX_CPU, -p MAX_CPU
                        Max number of cores to use. NOT YET IMPLEMENTED
```
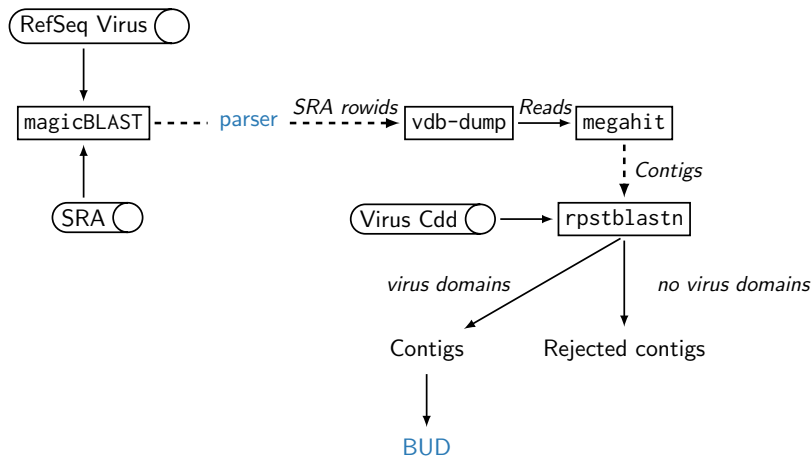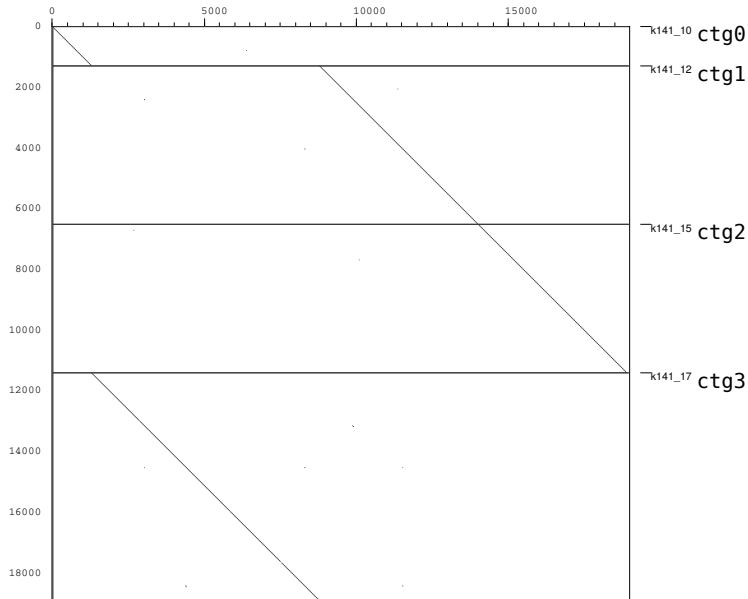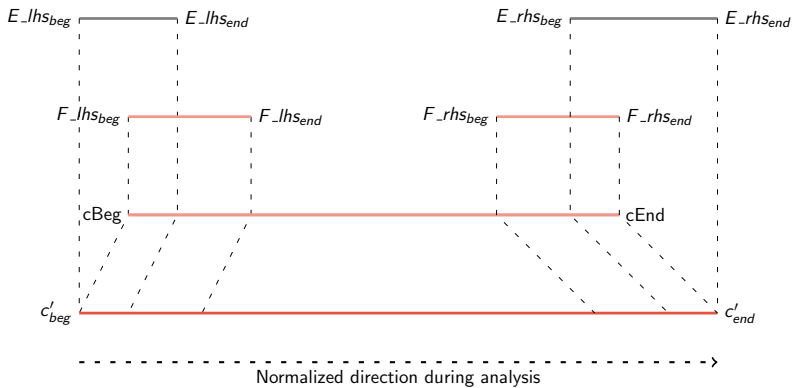
Workflow

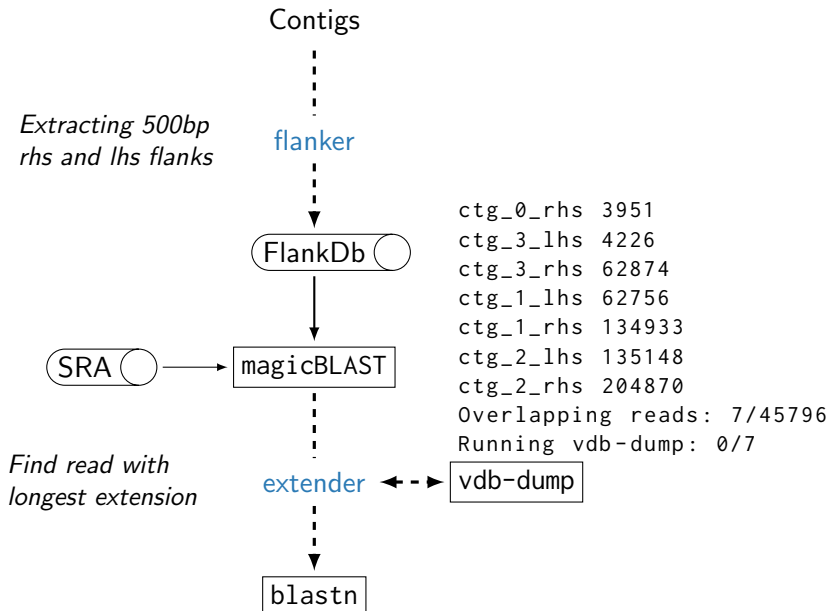# Init: current version

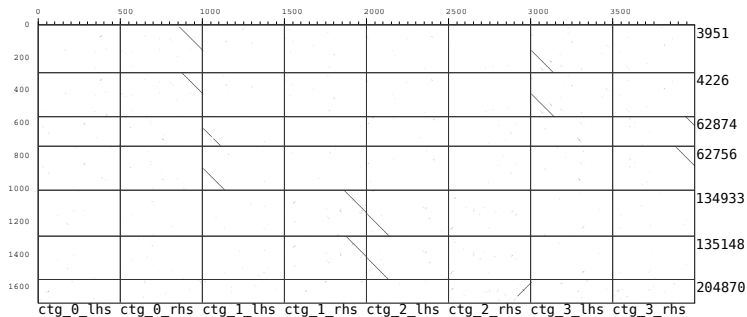# Init: SRR5150787 example (Ebola)

KU182909.1

Endovir's contig model

BUD: Extending

Extracting 500bp rhs and lhs flanks

Find read with longest extension

```
ctg_0_rhs 3951
ctg_3_lhs 4226
ctg_3_rhs 62874
ctg_1_lhs 62756
ctg_1_rhs 134933
ctg_2_lhs 135148
ctg_2_rhs 204870
Overlapping reads: 7/45796
Running vdb-dump: 0/7
```

BUD: Extending


```
SRR5150787
Flank: ctg_1_lhs  0 500 500 0 0 135
Read:  62756  129 264 265 0
Ext:   0   264 267
Flank: ctg_1_rhs  4843  5343  500 0 359 499
Read:  134933 0 140 280 0
Ext:   5202 5342  282
ctg_1 5482
  lhs ctg_1_lhs : 0 500 500 129
  62756 : lhs 0 264 136 131 267
  rhs ctg_1_rhs : 4982  5482  500 139
  134933 :  rhs 5202  5342  141 141 282
```
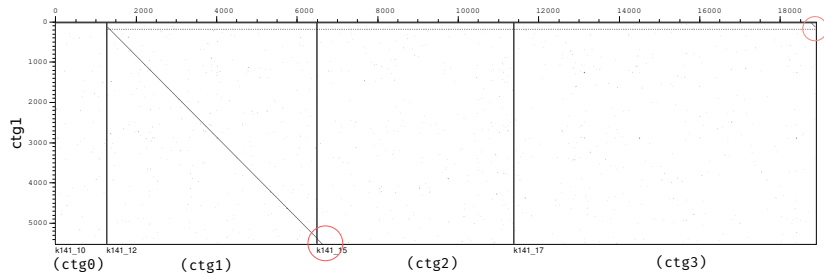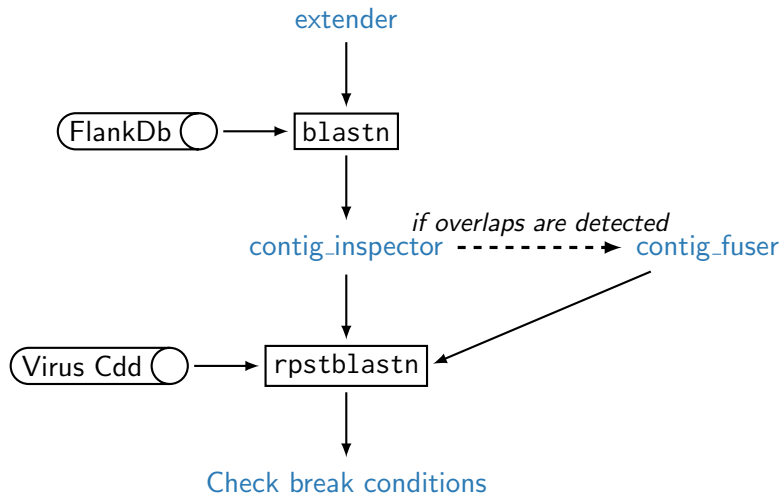
BUD: Extending

BUD: Overlap check

BUD: Overlap check

```
Checking flanks for overlaps
Blast results: ctg_0_lhs vs. ctg_0_lhs
Not an overlap
-----------
Blast results: ctg_0_lhs vs. ctg_2_rhs
Checking ctg_0_lhs and ctg_2_rhs
ctg_2:rhs + ctg_0:lhs
LHS:   130
RHS 117
oLHS: 129
oRHS  139
Src: ctg_2  7724
 RHS_flank: 7224  7724  500
   Extension: 7544 7606  182
Dst: ctg_0  5482  0 126 0
Brk: ctg_2  7597  ctg_0 129
------------------------------------
Blast results: ctg_0_rhs vs. ctg_1_lhs
Checking ctg_0_rhs and ctg_1_lhs
   Update: ctg_0 is now ctg_2
ctg_2 ctg_1
ctg_2:rhs + ctg_1:lhs
LHS:   130
RHS 0
oLHS: 134
oRHS  12
Src: ctg_2  7724
 RHS_flank: 7224  7724  500
   Extension: 10428  10568 0
Dst: ctg_1  5042  0 134 0
Brk: ctg_2  7589  ctg_1 134
```

BUD: Break conditions

Extending into host genome Empty Cdd database screen

No grow Contig was not extended

No contigs No contigs were identified

Modularity

- ▶ Facilitating extending and adding other tools using classes which act as messenger between tool and Endovir.

```python
class MappingAlignment:

  class Read:
    def __init__(self, name, start, stop, strand, qlen):

  class Flank:
    def __init__(self, name, start, stop, strand):

  def __init__(self, cols):
    self.read = self.Read(cols[0], cols[6], cols[7], cols[13], cols[15])
    self.flank = self.Flank(cols[1], cols[8], cols[9], cols[14])
    self.pident = float(cols[2])

class MagicblastParser:

  def __init__(self):
    self.alignments = []

  def parse(self, src):
    self.alignments = []
    for i in src:
      if i[0] != '#':
        self.alignments.append(MagicblastAlignment(i.strip().split('\t')))
```

Outlook

- Adjust tool parameters on the fly
- Replace SRA toolkit with NCBI's ngs-python
- Test on more diverse datasets
- Docker

Acknowledgments

- Ben Busby
- Niki Athanasiadou
- Yadi Zhou
- Richard Copin
- Jeff Robinson
- Greg Boratyn
- Virus Genome Refseq Group at NCBI
- Edward C. Holmes