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
  $\rightarrow$ Complex to expand and adjust
- Initiated during my stay as a visiting bioinfomartician as NCBI from Oct.2017 to Dec. 2017.
- Ongoing

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

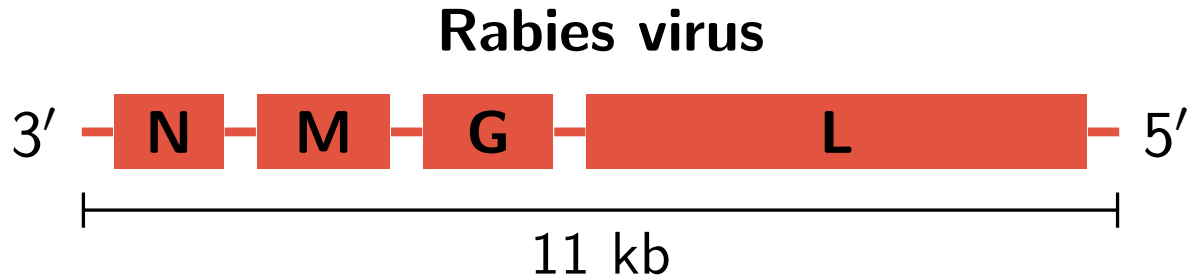
Foreword

Standing on the shoulder of giants

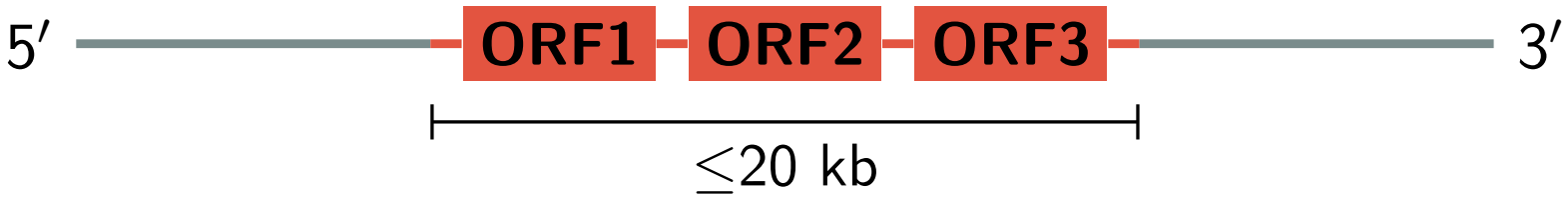
- Karina Zile, karina.zile@gmail.com
- Mitchell A. Ellison II, mae92@pitt.edu
- Jacob Waldman, jaw267@pitt.edu
- Kristyna Kupkova, kupkova@feec.vutbr.cz
- Cody Glickman, cody.glickman@ucdenver.edu
- Andrew Clugston, andrew.clugston@pitt.edu
- Paul G. Cantalupo, pcantalupo@gmail.com
- Vineet Raghu, vineet@cs.pitt.edu

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



Endogenous viral element (EVEs)

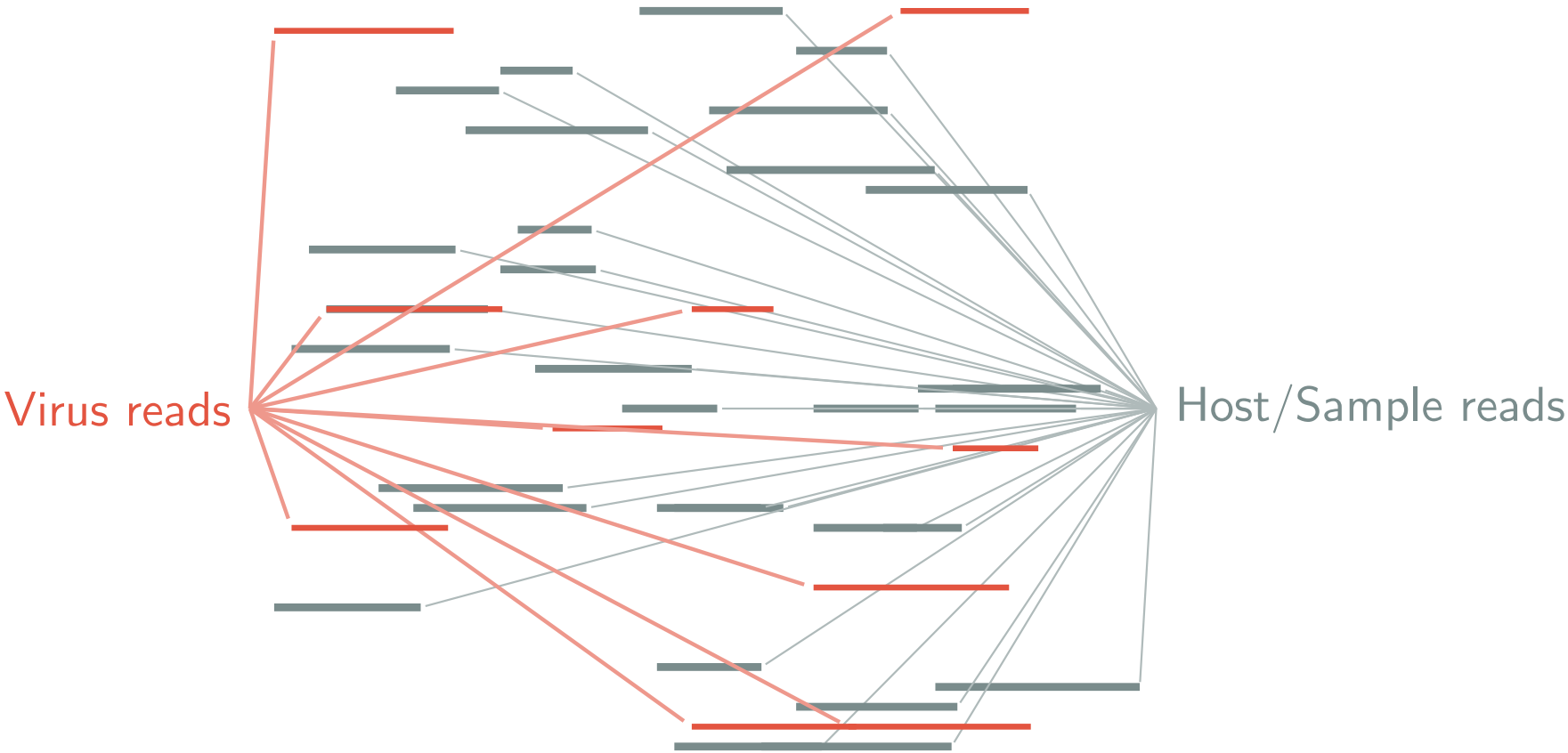


- Integration into host genome mandatory for replication
- Retroviridae, e.g. HIV
- Part of the host genome
- Not widely analyzed: complex analysis and missing data [2]
- Allowing to date virus evolution
- Characterize the mechanisms that underpin the integration of viruses into host genomes

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

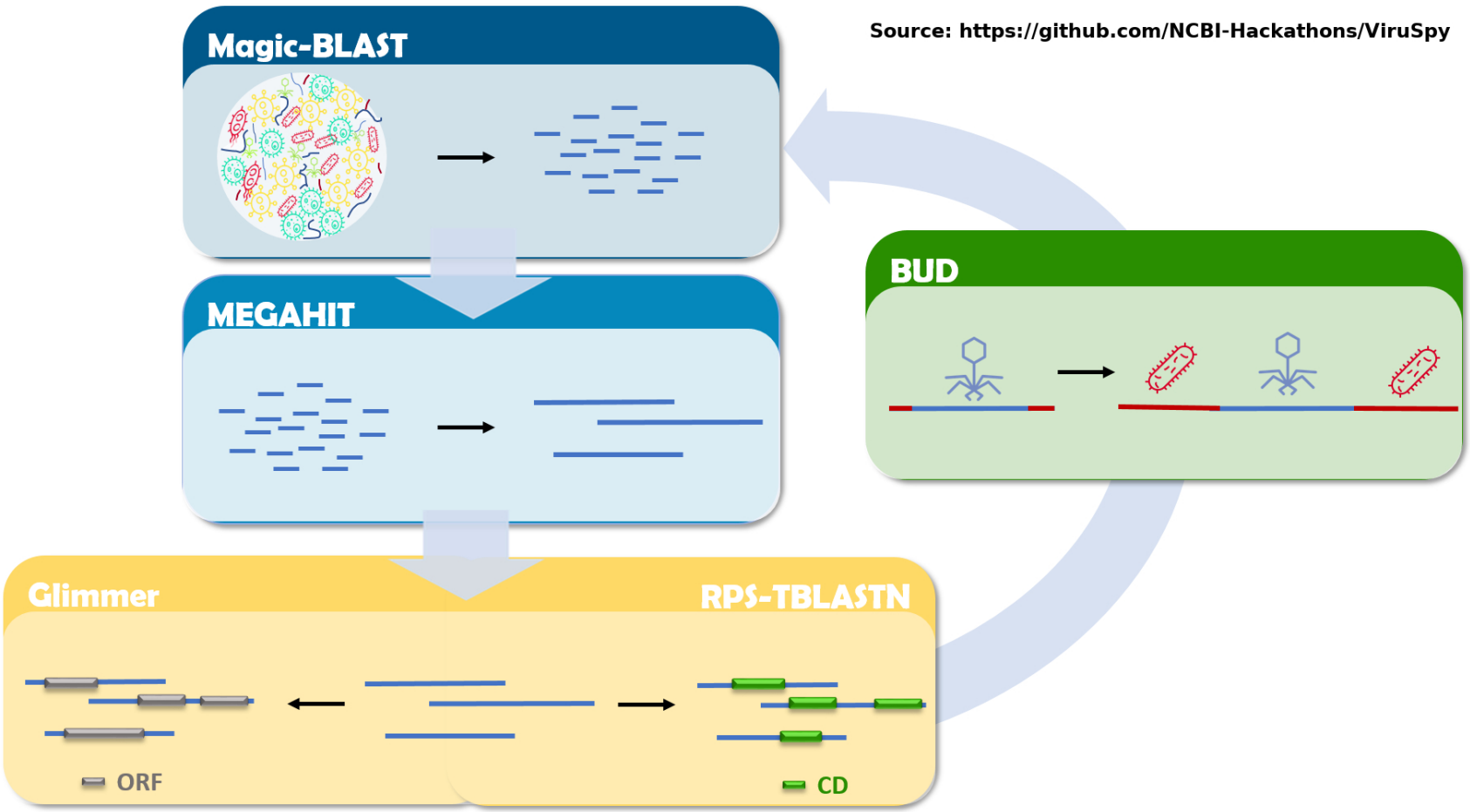
SRA data: a treasure trove for virus genomics

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



BUD algorithm

- "Building Up Domains"



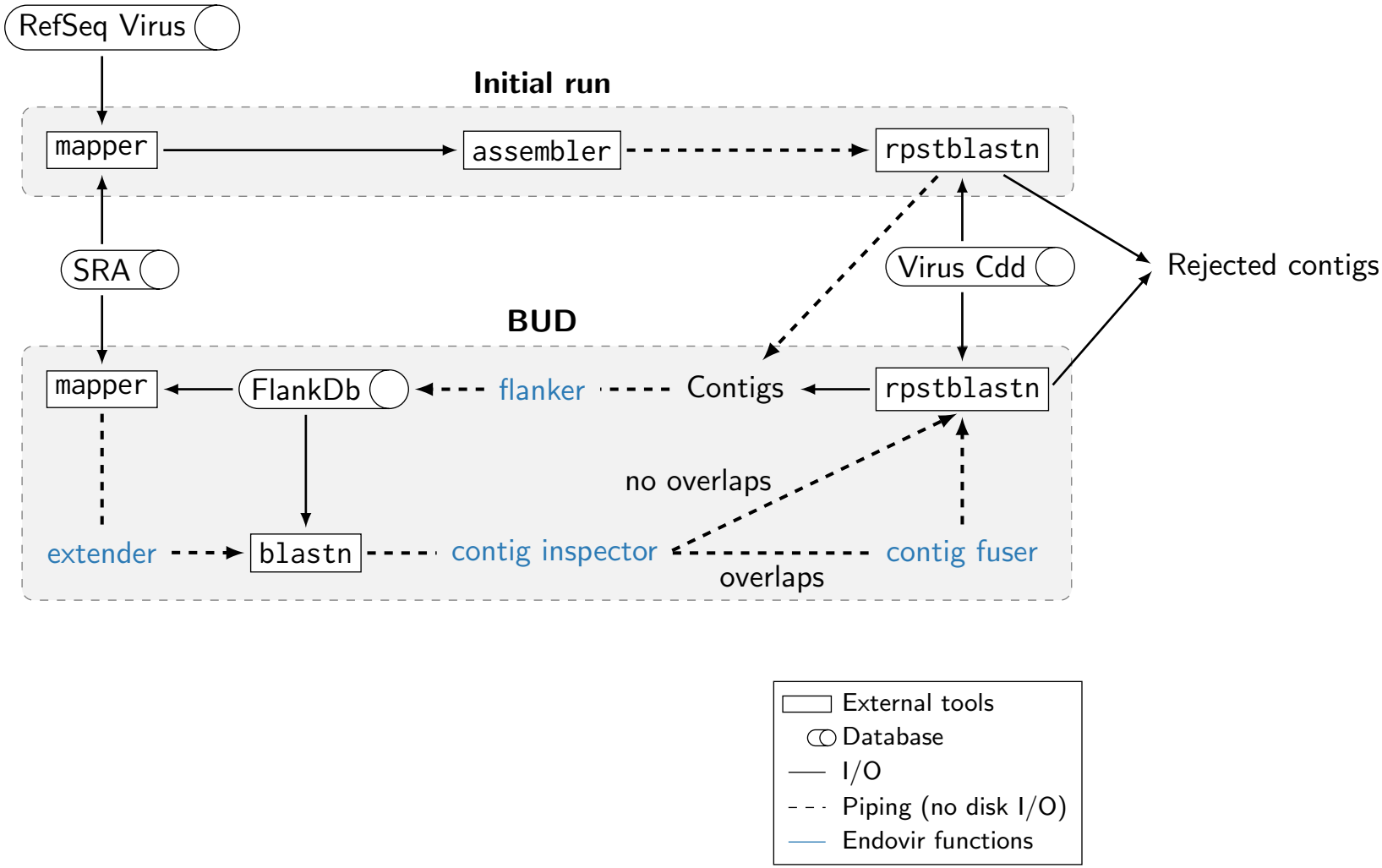
Endovir

Open source	MIT licence
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

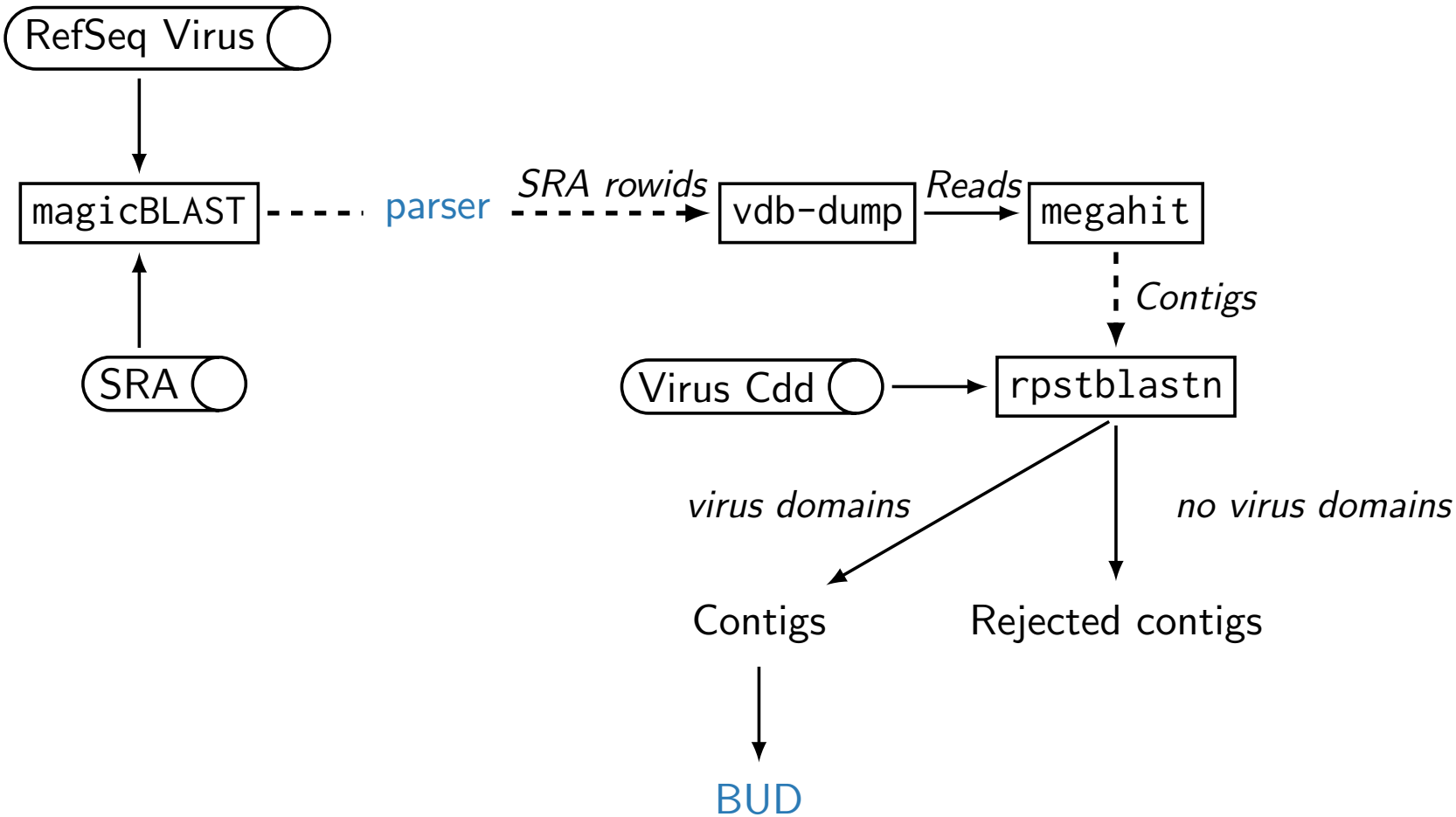
Usage

```
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
```

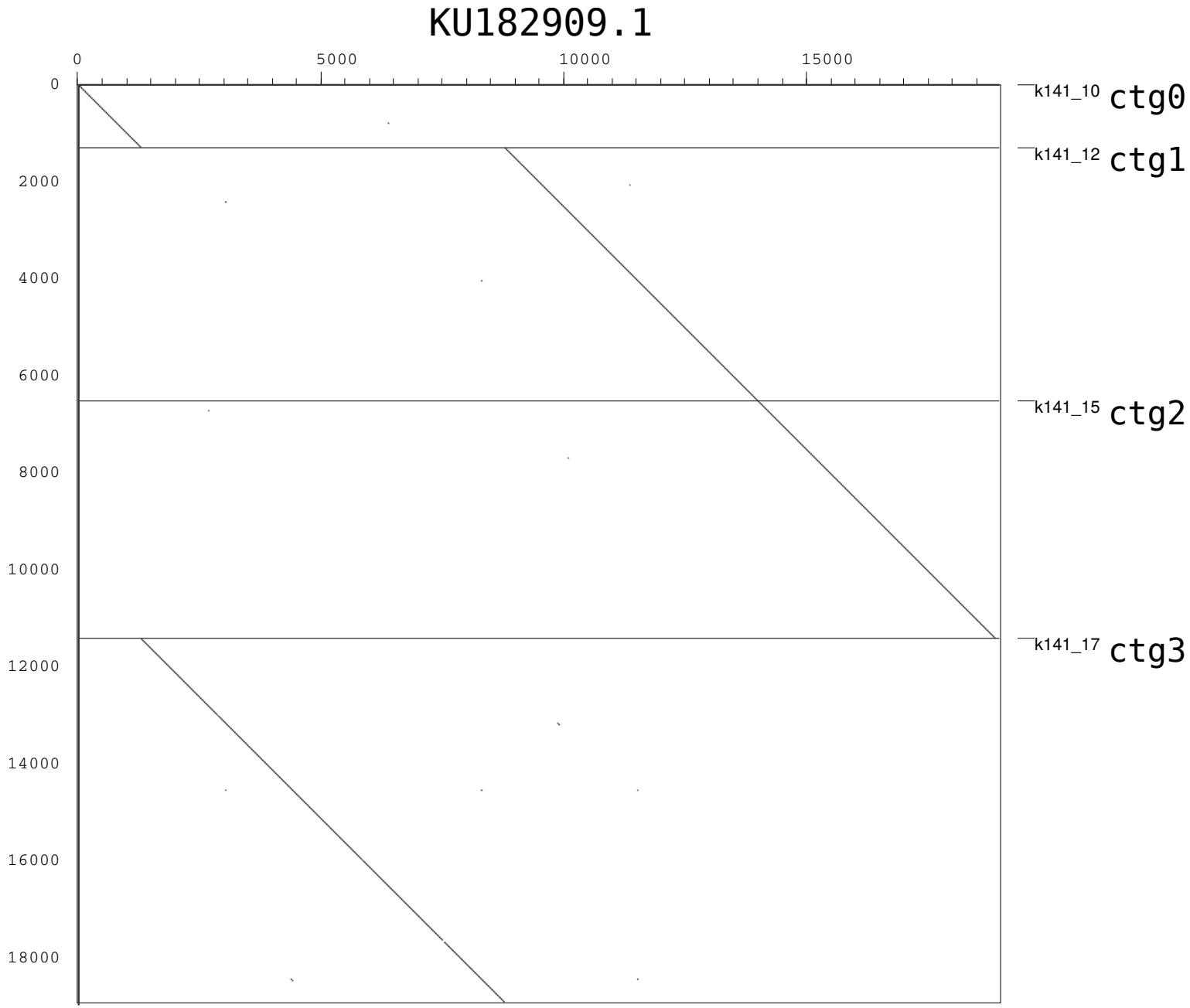
Workflow


RefSeq Virus
Initial run
mapper
assembler
rpstblastn
SRA
Virus Cdd
Rejected contigs
BUD
mapper
FlankDb
flanker
Contigs
rpstblastn
no overlaps
extender
blastn
contig inspector
contig fuser
overlaps
External tools
Database
I/O
Piping (no disk I/O)
Endovir functions

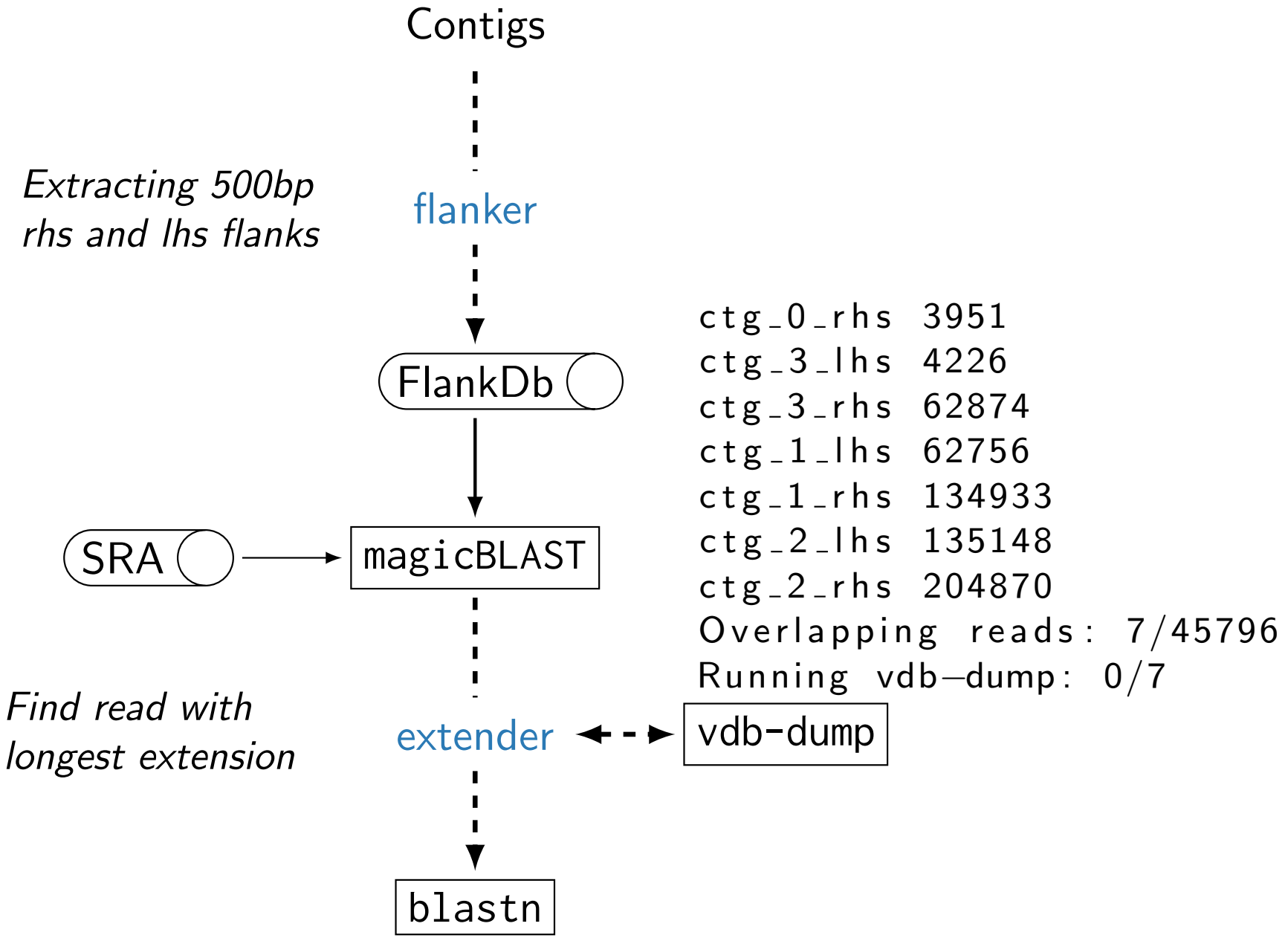
Init: current version



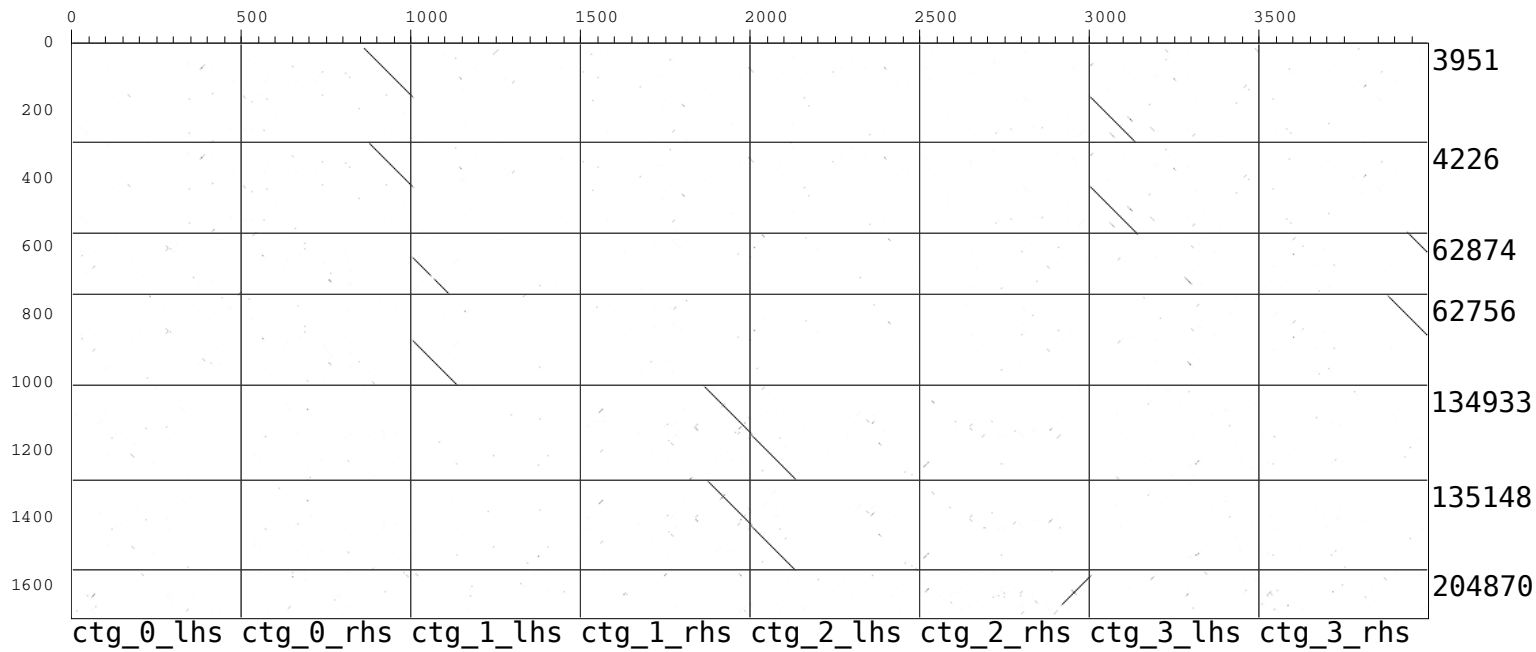
Init: SRR5150787 example (Ebola)



BUD: Extending



BUD: Extending



Modularity

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

```
class MagicblastAlignment(mapping_alignment.MappingAlignment):

  def __init__(self, cols):
    super().__init__(cols)
```

Modularity

```
class MappingAlignment:

  class Query:

    def __init__(self, name, start, stop, strand, qlen):
      self.name = name
      self.length = int(qlen)
      self.sra_rowid = name.split('.')[1]
      self.start = int(start) - 1
      self.stop = int(stop) - 1
      self.strand = 1 if strand == 'minus' else 0
      self.aln_length = abs(self.stop - self.start) + 1

    def get_ordered_coords(self):
      if self.strand == 0:
        return (self.start, self.stop)
      return (self.stop, self.start)

  class Reference:

    def __init__(self, name, start, stop, strand):
      self.name = name
      self.start = int(start) - 1
      self.stop = int(stop) - 1
      self.strand = 1 if strand == 'minus' else 0
      self.aln_length = abs(self.stop - self.start) + 1

    def get_ordered_coords(self):
      if self.strand == 0:
        return (self.start, self.stop)
      return (self.stop, self.start)

  def __init__(self, cols):
    self.qry = self.Query(cols[0], cols[6], cols[7], cols[13], cols[15])
    self.ref = self.Reference(cols[1], cols[8], cols[9], cols[14])
```

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 Yhou
- Richard Copin
- Jeff Robinson
- Edward C. Holmes
- Virus Genome Refseq Group at NCBI