

Endovir

Report NCBI Visiting Bioinf

Jan P. Buchmann

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

2017-12-08

Exogenous / Endogenous viruses

Exogenous virus

- ▶ Integration into host genome not required for replication
- ▶ Can occur occasionally, e.g. Human Hepres virus 6
- ▶ Clearly separated form host genome



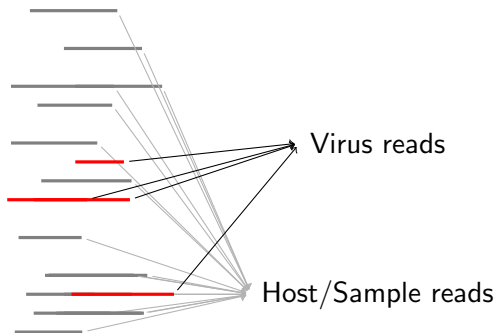
Endogenous virus

- ▶ Integration into host genome mandatory for replication
- ▶ Retroviridae, e.g. HIV
- ▶ Part of the host genome



SRA data

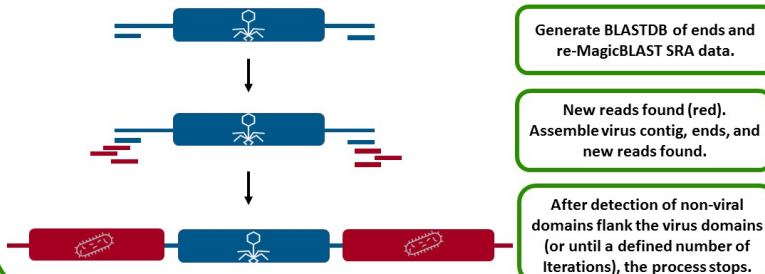
- ▶ SRA data contains host and virus sequences
- ▶ Endogenous and putative novel viruses



Previous work

- ▶ ViruSpy ¹
- ▶ Logic mostly implemented in Bash and Perl
→ Complex to expand and adjust
- ▶ BUD algorithm

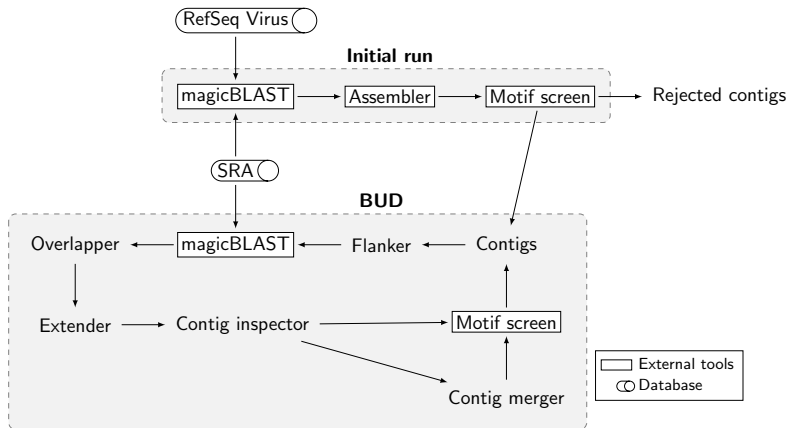
BUD algorithm



Source: <https://github.com/NCBI-Hackathons/VirusSpy>

¹<https://github.com/NCBI-Hackathons/VirusSpy>

Endovir



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

Modularity

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

```
class MagicblastAlignment:

    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

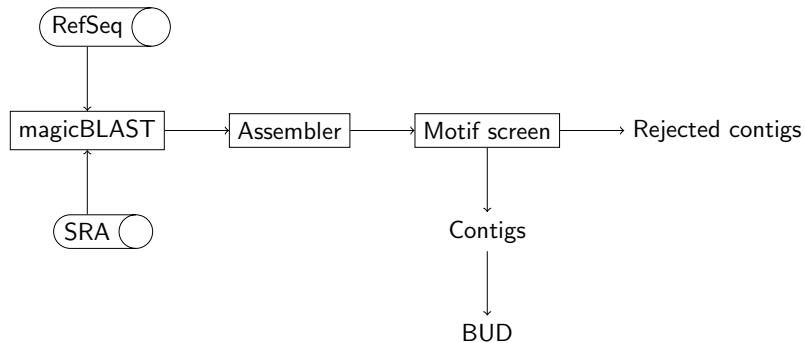
        def get_ordered_coords(self):
            if self.strand == 0:
                return (self.start, self.stop)
            return (self.stop, self.start)

    class Reference:

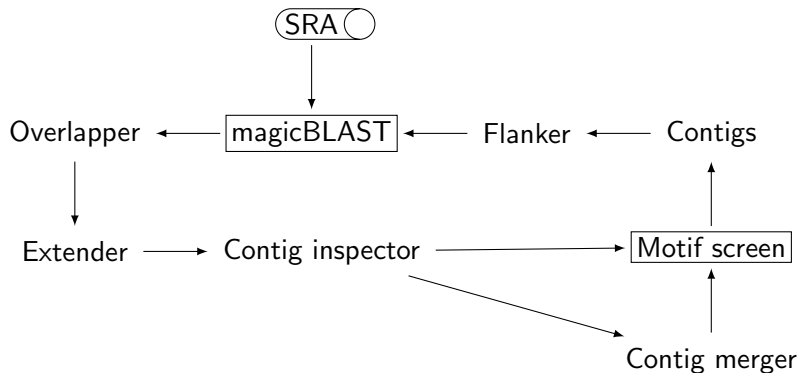
        def __init__(self, name, start, stop, strand):
            self.name = name
            self.start = int(start) - 1
            self.stop = int(stop) - 1
            self.strand = 1 if strand == 'minus' else 0
            self.aln_length = abs(self.stop - self.start) + 1

        def get_ordered_coords(self):
            if self.strand == 0:
```

Init



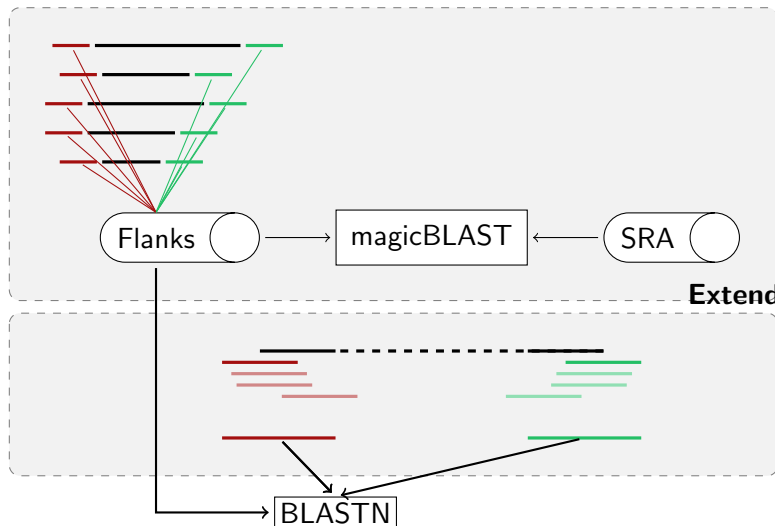
Bud



Flanker

Screen flanks

Extend flanks



Outlook

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