Notebook_200605_rce_diamond

Summary

Throughput and sensitivity of `diamond` vs. `bowtie2` was tested on two datasets, one with Coronavirus (Cov+) and one without (Cov-). Sensitivity to novel genera was tested by hold-out validation: the Cov+ dataset has SARS-Cov-2 which is a Betacoronavirus, the mapping references for hold-out tests had Alphacoronavirus only.

Results

Query	Aligner	Reference	Time	Mem	Alignments
Cov+	`bowtie2`	Pan-Cov	18:23	79 Mb	954 k
		Pan-Alpha	2:45	35 Mb	11 k
	`diamond`	Cov pol	15:00	387 Mb	271 k
		Cov genes	28:09	459 Mb	980 k
		Alpha genes	7:19	311 Mb	334 k
Cov-	`bowtie2`	Pan-Cov	3:39	65 Mb	2 k
		Pan-Alpha	2:48	27 Mb	0
	`diamond`	Cov pol	3:34	367 Mb	0
		Cov genes	3:51	450 Mb	0
		Alpha genes	3:52	453 Mb	0

Note that `bowtie2` finds 11/954 = 1% of the novel genus hits, compared to 334/980 = 34% for `diamond`. Thus `diamond` is **30× more sensitve** than `bowtie2` to novel genus alignments. On the Cov- test, the elapsed times were very similar. Memory use of `diamond` was well under 1Gb.

Methods

Datasets

Cov +	`SRR11454614`	Human hCov-19 infected patients bronchoalveolar lavage
Cov-	`ERR3568641`	Sheep thyroid deficiency before birth

Hardware

Linux server with Intel i7-7820X CPU @ 3.60GHz, SSHD.

Diamond command line and parameters

With default options, diamond uses a lot of memory and many threads. I used this command-line:

```
cat FASTQFILE \
  | $res/diamond/diamond blastx \
      -d $res/diamond/MAPPING_INDEX \
      -k 1 \
      -p 1 \
      -b 0.1 \
      -q /dev/stdin \
      -t /tmp \
      -o TSVFILE
```

`-k 1`	Maximum hits per query sequence. Reduce output file size for Cov+.
`-p 1`	Single-threaded for `t2.micro` or `t2.nano`.
`-b 0.1`	Limits memory, RAM used is roughly $<6\times b$ in Gb, so here is <600Mb.
`-t /tmp`	Temp directory. Required to avoid bug with input from /dev/stdin.
`-o TSVFILE`	Output file, or /dev/stdout for post-processing e.g. summarizer.

Mapping references

`bowtie2`	Pan-Cov	Coronavirus pan-genome `covref3`
	Pan-Alpha	Coronavirus pan-genome `covref3`, Alpha only.
`diamond`	Cov pol	Coronavirus, pol gene only.
	Cov genes	Coronavirus, all genes.
	Alpha genes	Coronavirus, all genes, Alpha only.