# Exercise 7

Jindřich Matuška, 525183

As part of this exercise

- create two WDLs and run them in terra.
- Discuss the results, runtime, cost, and your experience.
- Share which jobs specifically you consider as your final submission.

1) Create a WDL that processes an assembly as an input and outputs the lengths of all gaps
   - (sum of all unknown sequences or Ns).
   - Use grep for counting and allow preemptible.
2) Create a similar WDL but parallelize it to process each sequence within the assembly individually;
   - again report the total number of unknown sequences in the whole assembly.

## Creating WDLs

My nonparallel workflow consists of a single task in which I unzip the file using `gzip`, search with `grep` for all occurences of `N` (unknown nucleotides in sequence) and count the number of occurences with `wc`. Output of each command is piped directly as input of the next command for a speedup by not saving into intermediate file. This makes the workflow very fast (approx. 1 minute).

The parallel workflow first splits the assembly file into multiple ones so that each sequence is in its own file. This is done using `seqkit split` program and takes most of the time of the computation. This step can not be sped up by parallelization as `seqkit` seem to only profit from parallel reading/writing operations. Each of the resulting files is then processed in its own task using the same command as in the non-parallel workflow. Finally, the results of all these tasks are summed and written back into the table.

Counting in both workflows assume `N` as the only character for unknown nucleotide. This does not seem to be a problem as all four assemblies seem to conform to this.

## Running workflows in Terra

Both workflows were tested and ran in Terra.bio. Both of these workflows do give the same results. The submissions are under following submission IDs.

- Linear workflow: `86c0e06c-2a81-4fe9-b200-8867a27d104f`
- Parallel workflow: `8f338901-384c-4caa-af93-6ac81987242f`

The workflows are in GitHub repository

- `https://github.com/Ardnij123/pv269_7_slow`

**Runtime, cost**

The linear workflow was many times faster than the parallel as was argued in the previous section. The run cost of the linear was also many times lower. In this case, the parallelization is outpowered by the reading/writing operations.

**Experience**

There were a few large steps needed when creating the workflow. First problem came up when I was first creating the workflow. It took some time to get used to the WDL. The second problem was in working out problems with seqkit. Terra.bio does seem to assign 4 threads to a task by default. The seqkit also uses 4 threads by default. And whenever a task uses all assigned threads, Terra.bio kills the task. And prints out exception that says it is probably due to memory.

Apart from that, my experience with using Terra was mostly positive. The user interface takes a little bit of time to get used to, but otherwise it is quite friendly. WDL seems to me like a very useful language and tool.