NextFlow Introduction

Maxime Borry, Quentin Letourneur

07/09/2018

Introduction

The goal of this exercise is to recreate the pipeline showned in Figure 1 represented by its DAG. First clone the GitHub repository and `cd` into it. Name your pipeline `coverage.nf` and save it at the root of the cloned repository. You can then execute it with the command `nextflow run coverage.nf -with-dag workflow.nf`

Channel factory

You may want to use the have a look at the `fromFilePairs` methods to create the channel when working with paired end data

Mapping

You can use the `conda` directive to add bowtie2 to your pipeline: `conda "bioconda::bowtie2"`

The mapping command to use is the following:

```
bowtie2 -q -1 reads_1.fastq -2 reads_2.fastq -x index_prefix -S output.sam -p nb_cpus
```

Samtools view

You can use the `conda` directive to add samtools to your pipeline: `conda "bioconda::samtools"`

The goal is to convert the `sam` file into a compressed binary `bam` file using `samtools view`

The samtools command to use is the following:

```
samtools view -S -@ nb_cpus -b -o output.bam input.sam
```

Samtools sort

You can use the `conda` directive to add samtools to your pipeline: `conda "bioconda::samtools"`

The goal is to sort the reads mapped on the reference genome by position

The samtools command to use us the following:

```
samtools sort -@ nb_cpus -o sorted_output.bam input.bam
```

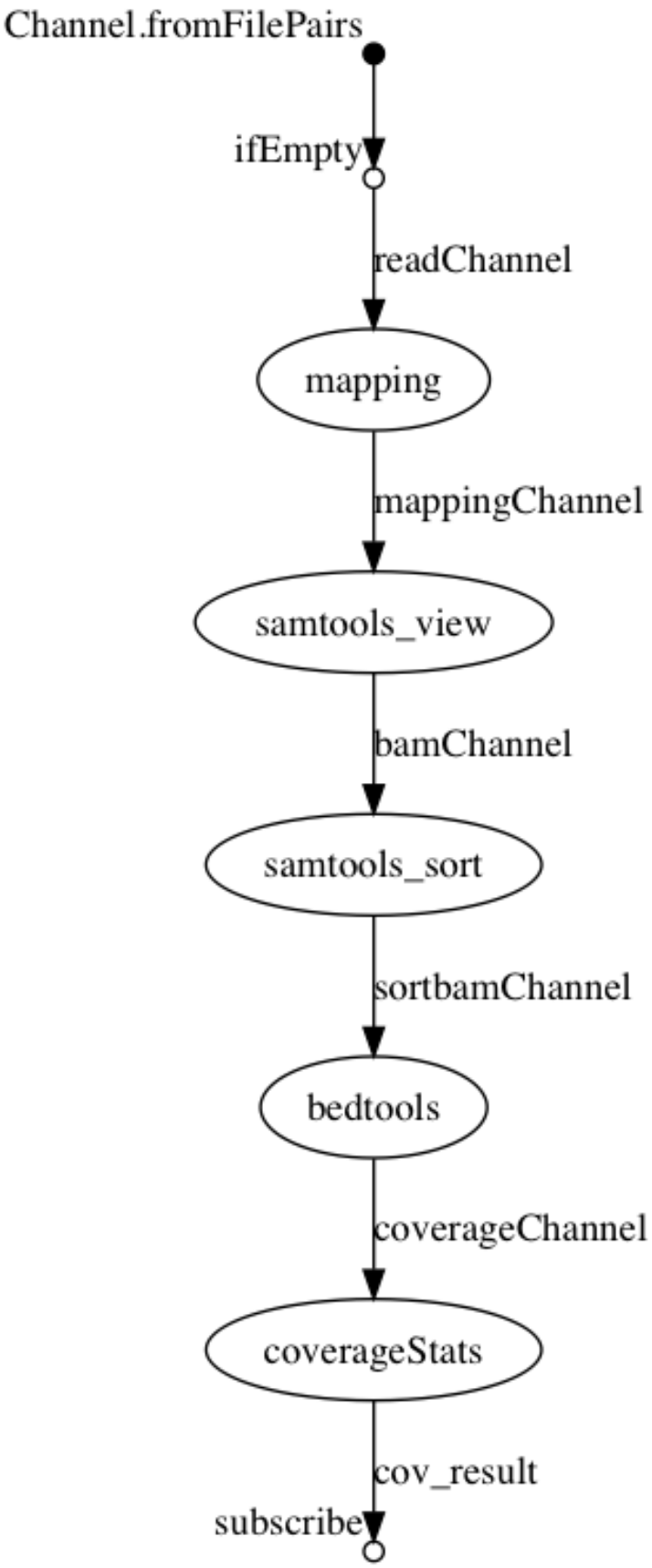

Figure 1: Directed Acyclic Graph (DAG) of the desired pipeline. Ellipses represent processes, arrows represents channels

Bedtools

You can use the `conda` directive to add bedtools to your pipeline: `conda "bioconda::bedtools"`

The goal is to compute the a position specific coverage of the reference genome provided with the aligned reads

The bedtools command to use is the following:

```
bedtools genomecov -ibam sorted_input.bam -d > output.gcbout
```

CoverageStats

You can use the `conda` directive to add Python 3.6 and Numpy to your pipeline: `conda "python=3.6 numpy"`

The goal is to compute the coverage (mean and median) statistics from the position specific coverage file

The script to use is located in the `bin` directory, and therefore is detected by Nextflow automatically.

The command to run this script is the following:

```
bed2coverage input.gcbout
```