Artic pipeline

Description

...

Setup

Set up Guppy

Download the appropriate version of guppy from Oxford Nanopore (requires registration, which is free), e.g. `ont-guppy_6.4.2_linux64.tar.gz` (GPU) or `ont-guppy-cpu_6.4.2_linux64.tar.gz` (CPU).

Or get it from

```
# CPU
wget https://mirror.oxfordnanoportal.com/software/analysis/ont-guppy-cpu_6.4.2_linux64.tar.gz

# GPU
wget https://mirror.oxfordnanoportal.com/software/analysis/ont-guppy_6.4.2_linux64.tar.gz
```

Extract files:

```
tar zxvf ont-guppy_6.4.2_linux64.tar.gz
```

Then add the `bin` directory to your `PATH` variable:

```
export PATH=/full/path/to/ont-guppy_6.4.2_linux64/bin:$PATH
```

To permanently have guppy available on your PATH, add the command above to the file `~/.bashrc`.

If you don't or you can't edit your PATH, use option `--guppy-path` to point to the guppy bin directory. E.g. `--guppy-path /path/to/ont-guppy_6.4.2_linux64/bin`

Set up conda environment

- Install conda, mamba, and configure for bioconda.
- Create a dedicated environment for this pipeline

```
conda create --yes -n artic-smk
conda activate artic-smk
mamba install --yes --file requirements.txt -n artic-smk
```

Usage

A basic example

The following command should work *as is* using the test data. It will process the given `fast5` directory according to `sample_sheet.tsv`. It assumes guppy is on the search `PATH`. Since option `--dry-run` is set it will only print what

would be executed, remove it for the real processing.

```
./artic-smk.py --sample-sheet test/data/sample_sheet.tsv \
    --fast5-dir test/data/fast5 \
    --genome-name my-genome \
    --output test_out \
    --dry-run
```

Run `./artic.smk.py -h` to see the list of available options (the following printout may be out of date):

```
positional arguments:
  targets                           Target file(s) to create or rule(s) to execute [all]

optional arguments:
  -h, --help                        show this help message and exit
  --version, -v                     show program's version number and exit

Main input/output options:
  --sample-sheet FILE, -s FILE      Tabular file of samples and barcodes. See online docs for
                                    details [required]
  --fast5-dir DIR, -f5 DIR          Directory of fast5 files
  --fastq-dir DIR, -fq DIR          Directory of demultiplexed fastq files. fast5-dir OR fastq-dir
                                    is required
  --output DIR, -o DIR              Output directory [artic-out]

Options for guppy (for fast5 input only):
  --guppy-config STR                Configuration for guppy_basecaller [dna_r9.4.1_450bps_fast.cfg]
  --guppy-barcode-kit STR           Barcode kit [EXP-NBD104]
  --guppy-path DIR                  Full path to guppy bin directory. Leave empty if guppy is on
                                    your search PATH []
  --guppy-basecaller-opts STR       Additional options passed to guppy_basecaller as a string with
                                    leading space e.g. " --num_callers 10" []

Options for artic minion/medaka:
  --medaka-model STR                Model for medaka [r941_min_fast_g303]
  --medaka-scheme-directory DIR, -sd DIR
                                    Path to scheme directory [primer-schemes]
  --medaka-scheme DIR               Scheme for medaka [rabv_ea/V1]
  --normalise N                     Normalise down to moderate coverage to save runtime [200]

Miscellanea:
  --genome-name STR, -g STR         Name for consensus genome [genome]
  --min-length N, -L N              Ignore reads less than min-length [350]

Workflow management options passed to snakemake:
  --jobs N, -j N                    Number of jobs to run in parallel [1]
  --dry-run, -n                     Only show what would be executed
  --snakefile FILE                  Snakefile of the pipeline. The directory "lib" is expected to be
                                    in the same directory as this file [Snakefile]
  --snakemake-opts STR, -smk STR    Additional options to snakemake as a string with leading space
                                    e.g. " --rerun-incomplete -k" [--rerun-incomplete]
```

Input

Sample sheet

This is a tabular file tab or comma separated with first non-skipped line as header. Lines starting with '#' are skipped. Columns are:

Column	Description
sample	Sample name. Avoid names with spaces or special characters (dots, underscores, hyphens are ok)
barcode	Sample barcode

Additional columns are ignored

Input reads

- **Option 1** A directory of **fast5** files that will be passed to `guppy_basecaller` and `guppy_barcoder`. Typically this is the output of the Nanopore run. Use `--fast5-dir/-f5` option to start from here.
- **Option 2** A directory of **fastq** files already demultiplexed and ready for further processing. Use `--fastq/fq` option to start from here, guppy installation is not required. Fastq-dir contains subdirectories named after the sample barcodes. They don't need to be real barcode names as long as they match the sample sheet column `barcode`. Each subdirectory can contain multiple fastq files, possibly gzip'd. This is the test data example:

```
test/data/fastq/
  barcode01
    tvla1_run2.fastq.gz
  barcode02
    tvla1_run1.fastq.gz
  barcode04
    dummy04.fastq.gz
  barcode06
    dummy06.fastq.gz
```

Testing & Development

To run the test suite:

```
./test/test.py
```

Compile this markdown to pdf:

```
pandoc -V colorlinks=true -V geometry:margin=0.8in README.md -o README.pdf
```