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-gpu_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` in `artic-smk.py` 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

The following command should work as-is using the test data. It will process the given fast5 directory according to `sample_sheet.tsv`. 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 The following printout may be out of date:

```
Run artic pipeline

optional arguments:
  -h, --help                        show this help message and exit
  --sample-sheet SAMPLE_SHEET, -s SAMPLE_SHEET
                                    Tab-separated file of samples, barcodes, and
                                    other sample-specific options. See online
                                    doumentation for details [required]
  --output OUTPUT, -o OUTPUT        Output directory [artic-out]
  --medaka-scheme-directory MEDAKA_SCHEME_DIRECTORY, -sd MEDAKA_SCHEME_DIRECTORY
                                    Path to scheme directory [primer-schemes]
  --fast5-dir FAST5_DIR, -f5 FAST5_DIR
                                    Directory of fast5 file. Typically the
                                    Nanopore run directory
  --fastq-dir FASTQ_DIR, -fq FASTQ_DIR
                                    Input alternative to fastq5-dir: Directory
                                    of demultiplexed fastq files. fastq-dir
                                    contains subdirectories named after the
                                    sample barcodes and containing the
                                    respective fastq files
  --genome-name GENOME_NAME, -g GENOME_NAME
                                    Name for consensus genome [genome]
  --guppy-config GUPPY_CONFIG       For fast5 input: Configuration for
                                    guppy_basecaller
                                    [dna_r9.4.1_450bps_fast.cfg]
  --guppy-barcode-kit GUPPY_BARCODE_KIT
                                    For fast5 input: Barcode kit passed to
                                    guppy_barcoder [EXP-NBD104]
  --guppy-basecaller-opts GUPPY_BASECALLER_OPTS
                                    Additional options passed to
                                    guppy_basecaller as a string with leading
                                    space e.g. " --num_caller 10" []
  --guppy-path GUPPY_PATH           Full path to guppy bin directory. Leave
                                    empty if guppy is already on your search
                                    PATH []
  --min-length MIN_LENGTH, -L MIN_LENGTH
                                    Ignore reads less than min-length [350]
  --medaka-model MEDAKA_MODEL       Model to use for medaka [r941_min_fast_g303]
  --jobs JOBS, -j JOBS              Number of jobs to run in parallel [1]
  --dry-run, -n                     Run pipeline dry-run mode
  --snakemake-opts SNAKEMAKE_OPTS, -smk SNAKEMAKE_OPTS
                                    Additional options to snakemake as a string
                                    with leading space e.g. " --rerun-incomplete
                                    -k" []
  --version, -v                     show program's version number and exit
```

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

- **Options 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.
- **Options 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 subdirecties 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
```

Devel

To run test suite:

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

Compile this markdown to pdf:

```
pandoc -V colorlinks=true -V geometry:margin=1in README.md -o README.pdf
```