# ARTIC Pipeline

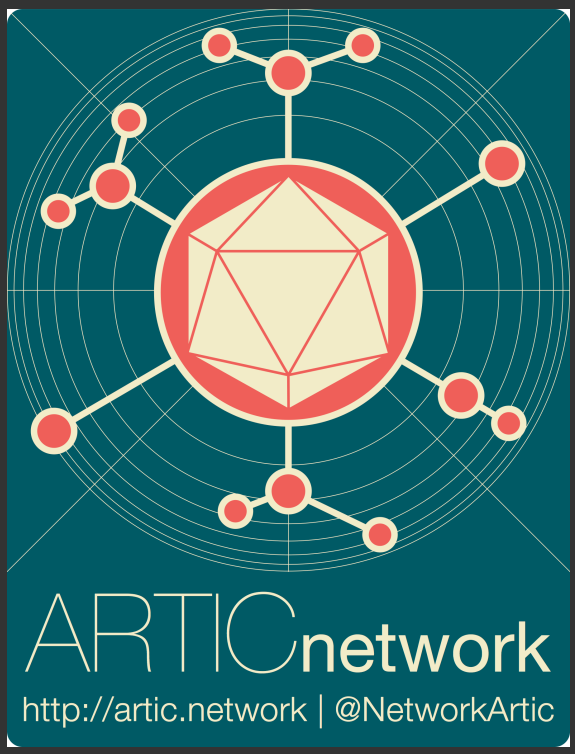
*From Raw ONT Data to Consensus Sequences*

Joanna Malukiewicz, *Ph.D.*

*University of Hamburg / Institute of Tropical Medicine, USP*

Thursday, the 13$^{th}$ of July, 2023

# Introduction




nCoV-2019 novel coronavirus bioinformatics protocol

# ONT Software

- MinKNOW software provides a graphical interface between the minION sequencer and the user
- MinKNOW can be set to be run with or without basecalling
- Guppy is used for calling bases from input FAST5 and outputting basecalls as FASTQ format (more on those formats later)

# ONT Fast5 Format

- The raw “squiggle” signals that come off the minION are stored in the hdf5-based fast5 format
- HDF = Hierarchical Data Format
- Data in this type of file are structured in a nested format, similar as JSON
- more info https://medium.com/@shiansu/a-look-at-the-nanopore-fast5-format-f711999e2ff6

# ARTIC Bioinformatics Steps

https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html

–Basecalling

–Demultiplexing

–Mapping

–Polishing

–Consensus Generation

# Conda Environment

- Manager for programs and environments
- Each environment can have its own versions of Python and/or different versions of a specific program
- We will activate a pre-installed ARTIC environment with Conda
- The following command can be typed into the terminal prompt

```
source activate artic-ncov2019
```

# Guppy Basecalling

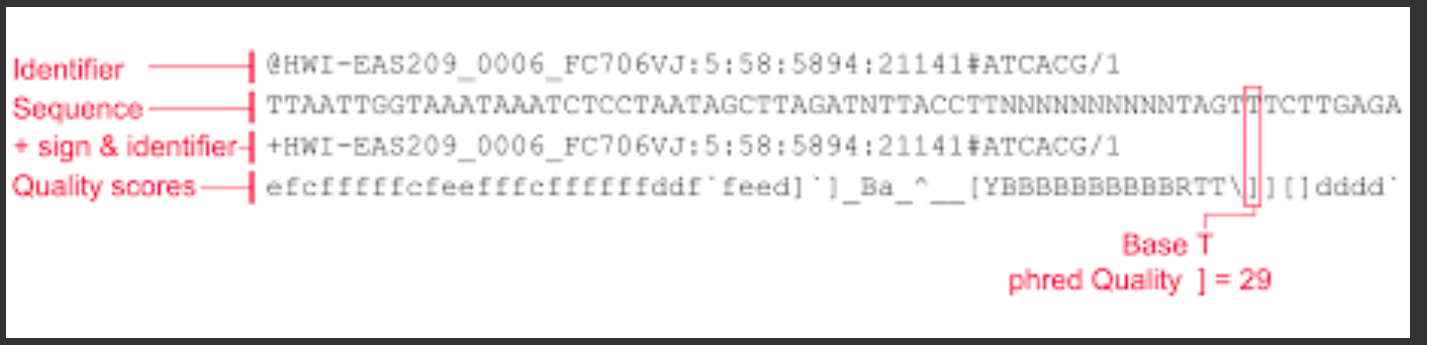
- Barcodes help identify individual samples during sequencing
- Guppy can be used for basecalling (if not previously carried out in minKNOW)
- An example of a basecalling command via guppy would be
- Basecalled data are output in FASTQ format
- FASTQ is similar to FASTA, but it has extra basecall quality information not contained within a FASTA file

```
guppy_basecaller -c dna_r9.4.1_450bps_hac.cfg -i /path/to/reads -s run_name -x
auto -r
```

# FASTA


FASTA format

# FASTQ




FASTQ format

# Guppy Basecalling

```
guppy_basecaller -c dna_r9.4.1_450bps_hac.cfg -i /path/to/reads -s run_name -x
auto -r
```

- “-c” is the configuration file that gives guppy some information about what library kit and flow cell type you used
- “-i” is the directory where your FAST5 reads are found
- “-s” is the name for your sequencing run
- “-x” determines CPU or GPU mode

# Guppy Demultiplex

- After (or during) basecalling, sequencing reads from specific samples need to be sorted by their given barcode, ie demultiplexed
- Guppy can be
- An example o
would be


```
guppy_barcoder --require_barcodes_both_ends -i run_name -s output_directory --
arrangements_files "barcode_arrs_nb12.cfg barcode_arrs_nb24.cfg"
```

# ARTIC Filtering

- ARTIC filtering step is carried out to remove chimeric reads
- Chimeric reads can be formed by the ligation of two distinct molecules during library prep
- Chimeras also form in silico by the base calling software when two molecules are sequenced in the same pore in short succession (Martin and Legget, 2021)
- In silico chimeras can lead to barcode misidentification

# ARTIC Filtering

- The artic pipeline gets around chimeric reads with two main approaches
- During demultiplex the same barcode is seen at the start and the end of each read
- Reads can also be filtered to ensure they are of the expected size (e.g. amplicon length + adaptor + barcode length, typically around 500bp for our schemes)

# Filtering Command

- Filtering by size occurs via the command line below
- This command will also bring all the fastq files for each barcode into a single *.fastq file

```
artic guppyplex --skip-quality-check --min-length 400 --max-length 700 --
directory output_directory/barcode03 --prefix run_name
```

# MinION Pipeline

- This command has to be carried out individual per barcode
- Bar code and sample name will need to be altered for each command

```
artic minion --normalise 200 --threads 4 --scheme-directory ~/artic-
ncov2019/primer_schemes --read-file run_name_barcode03.fastq --fast5-directory
path_to_fast5 --sequencing-summary path_to_sequencing_summary.txt nCoV-2019/V3
samplename
```

# MinION Pipeline

```
artic minion --normalise 200 --threads 4 --scheme-directory ~/artic-
ncov2019/primer_schemes --read-file run_name_barcode03.fastq --fast5-directory
path_to_fast5 --sequencing-summary path_to_sequencing_summary.txt nCoV-2019/V3
samplename
```

- “threads” is a computing process to support the command
- more threads = more powerful processing
- –scheme-directory is the location of things like primers and reference sequence
- –sequencing-summary file is generated by guppy with information about base calling run such as which reads passed and failed that part of the pipeline

# MinION Pipeline

```
artic minion --normalise 200 --threads 4 --scheme-directory ~/artic-
ncov2019/primer_schemes --read-file run_name_barcode03.fastq --fast5-directory
path_to_fast5 --sequencing-summary path_to_sequencing_summary.txt nCoV-2019/V3
samplename
```

- The command will carry out several step
  - Mapping of sequencing reads to a provided SARS CoV-2 reference genome (located within the scheme directory)
  - Polishing the reads (bioinformatically improving the basecalls)
  - Creating a SARS CoV-2 genome consensus based on the provided sequencing reads

# Output Files

- samplename.rg.primertrimmed.bam -> BAM file for visualisation after primer-binding site trimming
- samplename.trimmed.bam -> BAM file with the primers left on (used in variant calling)
- samplename.merged.vcf -> all detected variants in VCF format
- samplename.pass.vcf -> detected variants in VCF format passing quality filter

# Output Files

- samplename.fail.vcf -> detected variants in VCF format failing quality filter
- samplename.primers.vcf -> detected variants falling in primer-binding regions
- samplename.consensus.fasta -> consensus sequence

# BAM File

- binary alignment map

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

BAM

# VCF File

- variant call format

**VCF**

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

**BCF**

```
2  81170  .  C  T   .  .    AC=9;AN=7424   GT:0/0:0/0:0/1:0/1:1/0:0/0:0/0    DP:4:3:1:9:4:5:4          GQ:12: 9: 3:24:12:15:12
2  81171  .  G  A   .  .    AC=6;AN=7446   GT:0/1:0/0:0/0:0/0:0/1:0/1:0/0    DP:4:3:1:9:4:5:4          GQ:12: 9: 3:24:12:15:12
2  81182  .  A  G   .  .    AC=5;AN=7506   GT:0/0:0/0:0/0:0/0:0/0:0/0:0/0    DP:5:4:5:9:4:4:4          GQ:15:12:15:24:12:12:12
2  81204  .  T  G   .  .    AC=2;AN=7542   GT:1/0:0/0:0/0:0/0:0/0:1/0:0/1    DP:5:9:10:15:9:13:14      GQ:15:27:30:39:27:39:42
```

vcf