Introduction to Nanopore sequence analysis

Tim Kahlke
tim.kahlke@uts.edu.au
https://github.com/timkahlke
Twitter: @AdvancedTwigTec

Nanopore
Ideology & Resources

How do I get my own MinION ????

Nanopore Store: www.nanoporetech.com

Starter Kit: $1,000US

- MinION
- 2 Flow Cells
- 1 library preparation kit (6 reactions)
- 1 Flow Cell Wash Ki

Don’t forget your import permit!!!

Consumables

- Flow Cell ~$900US
- Library prep kits ~$550US
- Flow Cell Wash Kits ~$190US

Note

The ligation kit needs additional consumables

- Ampure / CleanNGS magentic Beads
- NEB FFPE / U II End repair
- Ligase Master Mix

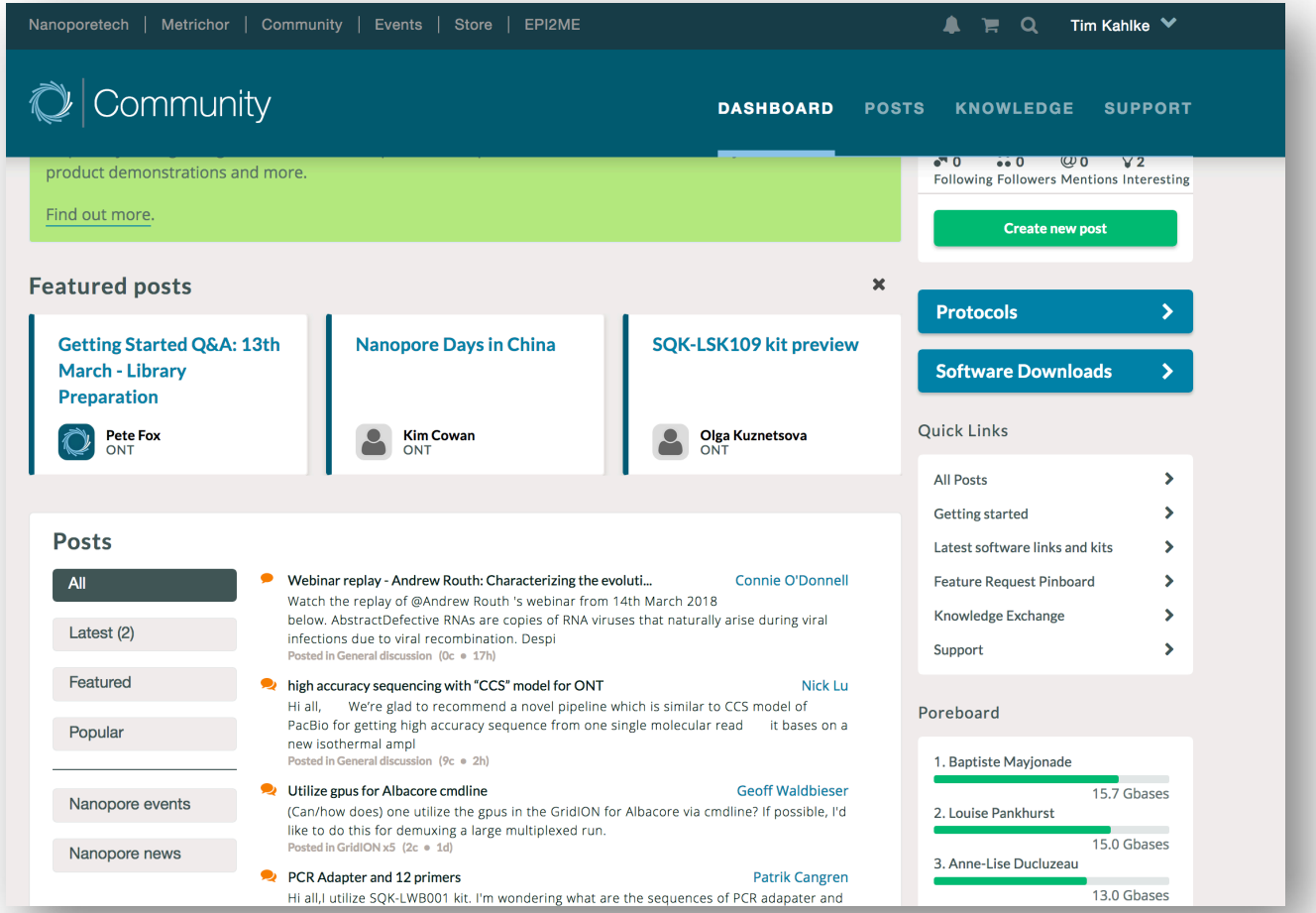
Resources

Nanopore Community

- Customer Support
- Protocols
- Forums

Protocols.io

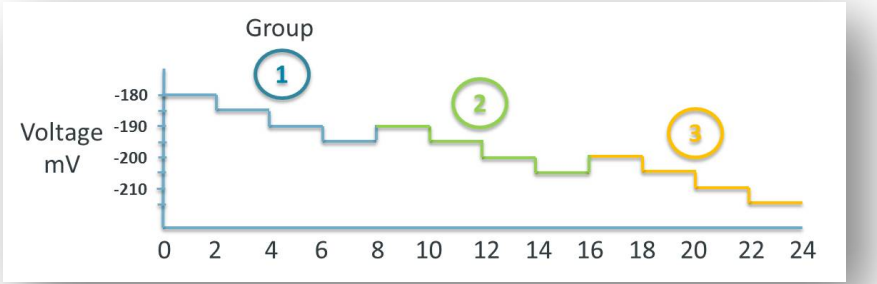
- HMW extraction protocols



MinKNOW – advanced settings

MinKNOW protocol scripts

- Give more flexibility
- Configure:
  - Starting voltage for sequencing run
  - Duration of sequencing run
  - Default working directory
  - Turn raw signal storage on/off
  - Turn event data collection on/off



Source: https://community.nanoporetech.com

MinKNOW protocol scripts

Windows

C:\Program Files\OxfordNanopore\MinKNOW\ont-python\Lib\site packages\bream\core\nc\cli\NC_Sequencing.py

MacOSX

Applications/MinKNOW.app/Contents/Resources/ont-python/lib/python2.7/site-packages/bream/core/nc/cli/NC_Sequencing.py

Linux

/opt/ONT/MinKNOW/ont-python/lib/python2.7/site-packages/bream/core/nc/cli

- Editing requires a command-line text editor
- More details at Nanopore community

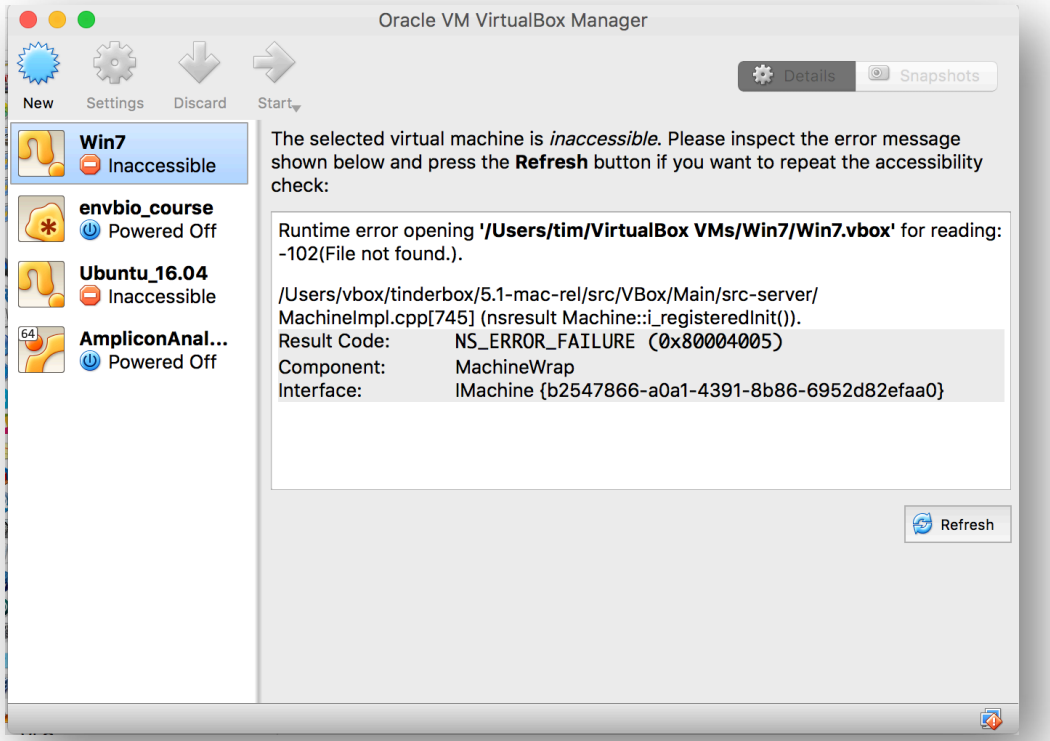
Workshop environment

Google drive

- /Practicals/DayX — Practical data including course virtual machine file
- /Practicals/Day2/Sequencing — Sequencing data from day1
- /Presentations — Presentation .pptx
- /Hand-outs — Handout.pdfs
- /Pre_course — VirtualBox installs etc

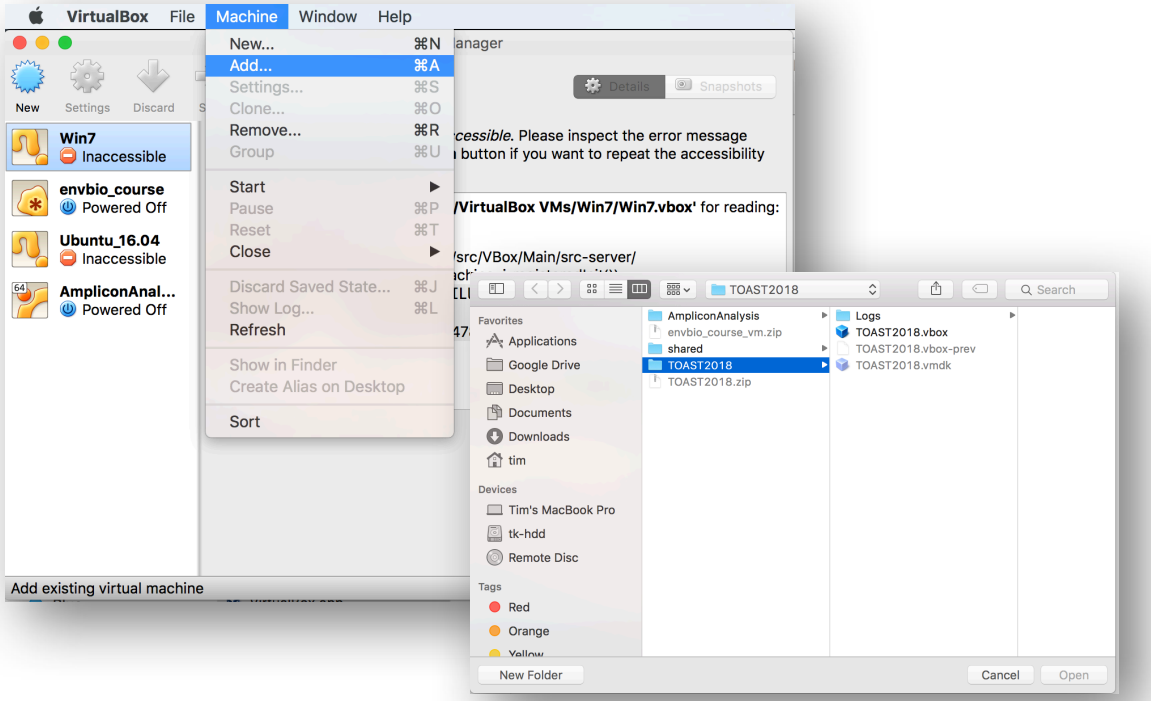
Virtual machine

- ”Virtual computer”
- Enables distribution of pre-configured computer system

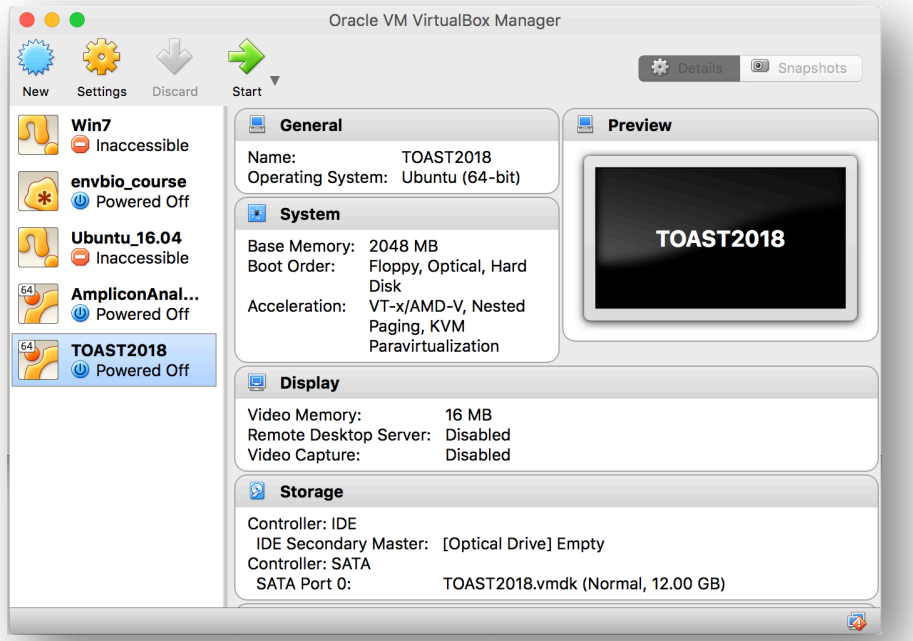


Virtual machine

1. Add TOAST2018 VM

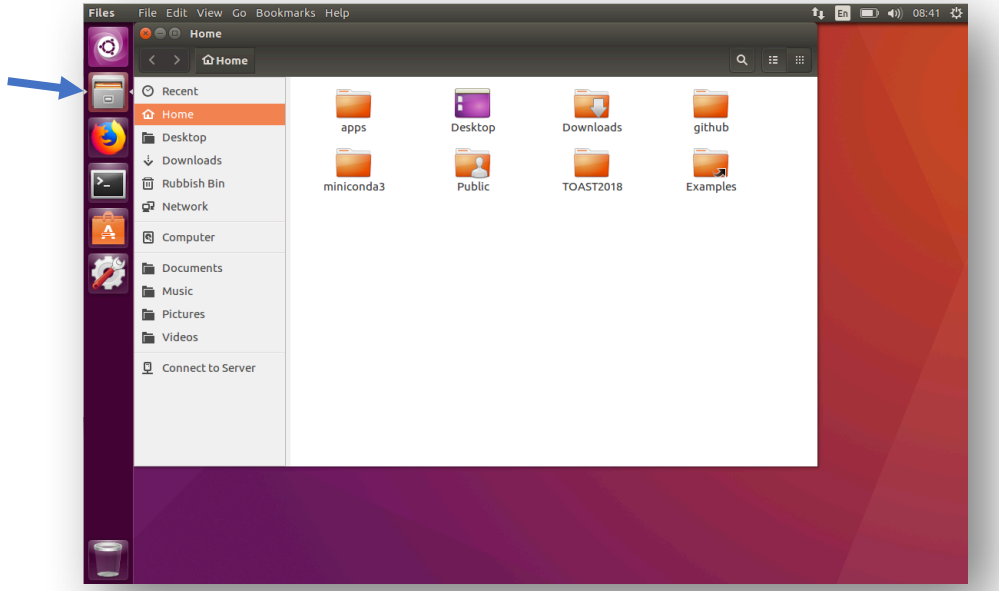


2. Open VM

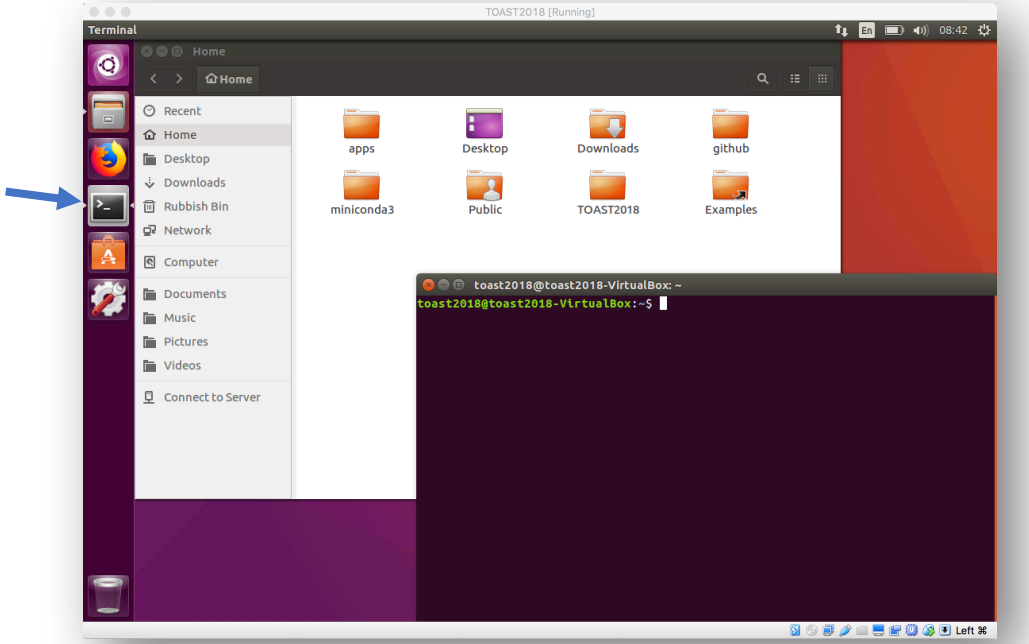


Ubuntu VM

Ubuntu Desktop



Terminal / Command-line



Course directories

/home/toast/2018	Home directory
~/	Short for /home/toast2018
~/TOAST2018	Course data
~/TOAST2018/day1	Download and copy yesterdays sequencing data here
~/TOAST2018/day2	Directory for todays practicals
~/TOAST2018/precomp	Directory with pre-computed data you will need for some tutorials

Sequencing data file formats

FastA

- Originally introduced as the main file format for the fasta program suite
- Sequence entry consist of at least 2 lines
    1. First line starting with “>” followed by the sequence identifier and optional additional information, .e.g tags and qualifiers
    2. Following lines contain sequence originally restricted to 80 or 120 characters per line

```
>NC_012064.1 Thalassiosira pseudonana CCMP1335
TCCAAGAGTCGAAAGTAGTAATCTACTAGGACTACGCATAGCTCGCTAGCATCGTCGACTCGATCGACTCGCTCTACGCT
AGAATTGTGTTTCGCATCTTCCATTGTCCTTCAAAAATTTTCCATGTTTCCCCGATTAGCACCGTGGAGAGTTCGACGCA
GAATCCCCGGAGAAGTGCATCACTACCGAGAGACTCTTGTCGCTCGATTGCTCGCAACGTATGTGAGCAGTGTAAAGCAT
ATGGATTCCGAGGGAGATGAACAAGTTTGCAAATCGCGTCA
```

FastQ

- Developed by the *Welcome Sanger Trust Institute* combining sequence and quality scores
- *De facto* standard for Next- or 2nd Generation sequencing technologies
- Every sequence entry consists of 4 lines
    1. Line starting with ”@” followed by the sequence identifier
    2. Characters of the Sequence
    3. A “+”
    4. Quality scores encoded in ASCII characters, one character per corresponding sequence character

```
@NC_012064.1
TCCAAGAGTCGAAAGTAGTAATCTACTAGGACTACGCATAGCTCGCTAGCATCGTCGACTCGATCGACTCGCTCTACGCT
+
SDFL++++++++++++??????????????????????????TI#)%*IY)GEJ)##)J$JDWK#K#DFJVGJ#$T#*@#
```

Fast5

- Oxford Nanopore file format for storing raw MinION sequencing output and base call information
- Based on HDF5, a standard to store large complex data types

```
/{attributes: file_version}
|-UniqueGlobalKey/
| |-tracking_id/{attributes: standard tracking fields}
| |-channel_id/{attributes: channel_number, digitisation, offset, range, sampling_rate}
| |-context_tags/{attributes: set when the experiment is configured} |-Raw/
| |-Reads/
| |-Read_42/{attributes: start_time, duration, read_number, start_mux, read_id}
| |-Signal{samples} |-Analyses/
| |-Segmentation_000{attributes: name, version}
| | |-Configuration/
| | | |-stall_removal/{attributes: parameters for stall removal}
| | | |-split_hairpin/{attributes: parameters for hairpin splitting}
| | |-Summary/{attributes: return_status}
| | | |-segmentation/{attributes: has_template, has_complement, first_sample_template, duration_template, first_sample_complement, duration_complement,
                                num_events_template, num_events_complement}
| |-Basecall_1D_000/{attributes: name, version}
| | |-Configuration/
| | | |-basecall_1d/{attributes: parameters for basecalling}
| | |-BaseCalled_template/
| | | |-Events{mean, stdv, start, length, model_state, move, weights, p_model_state, mp_state, p_mp_state, p_A, p_C, p_G, p_T}
| | | |-Fastq{string}
| | |-Summary/{attributes: return_status}
| | | |-basecall_1d_template/{attributes: num_events, called_events, mean_qscore, strand_score, sequence_length, stay_prob, step_prob, skip_prob}
|
```

Phred Quality scores

Phred33

- Encodes the error probability of the corresponding sequence character
- Given quality score *Q* the error probability *Pe* is calculated with

$$Pe = 10^{\frac{-Q}{10}}$$

Phred score	Error probability
10	0.1 (10%)
20	0.01 (1%)
30	0.001 (0.1%)
40	0.0001 (0.01%)
50	0.00001 (0.001%)

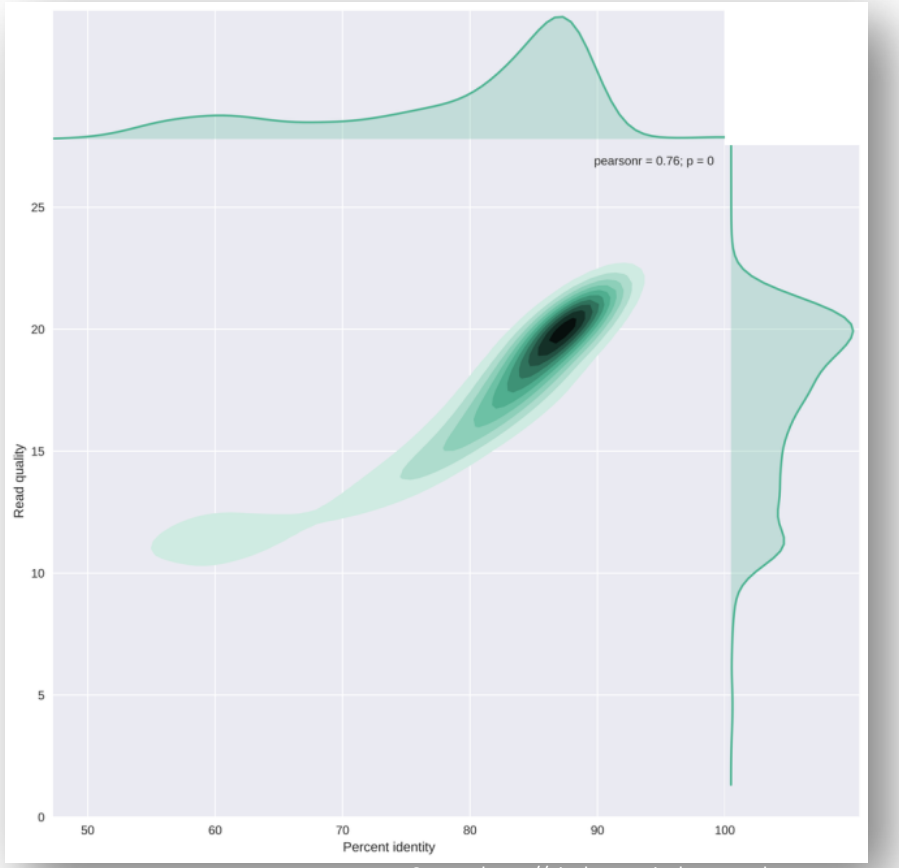
- In FastQ files *Q* for each base is given in a single ASCII character ranging from 33-126

Symbol	ASCII value	Q	Pe
+	43	10	0.1
?	63	30	0.001

Nanopore Q-score

- Plotting of Albacore quality scores (Nov. 2016)
- Human Genome (Nanopore) vs GRCh37

Phred score	Error probability
10	0.1 (10%)
20	0.01 (1%)
30	0.001 (0.1%)
40	0.0001 (0.01%)
50	0.00001 (0.001%)



Source: https://gigabaseorgigabyte.wordpress.com

MinKNOW – Data storage

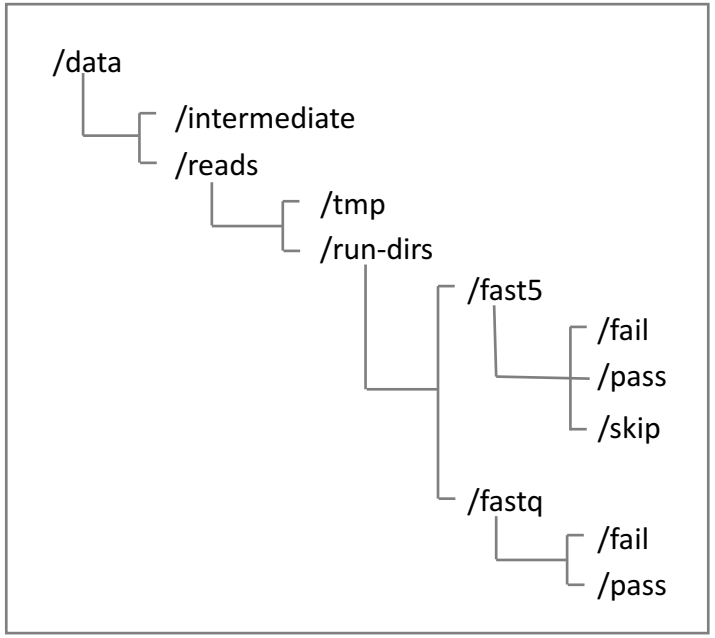
MinKNOW working directory

- Can be set via MinKNOW script (advanced option)
- Default
  - Windows: C:\data
  - MacOSX: /Library/MinKNOW/data
  - Unix: /var/lib/MinKNOW/data
- Two main sub-directories:
  - /intermediate | Directory for unprocessed raw signal data.

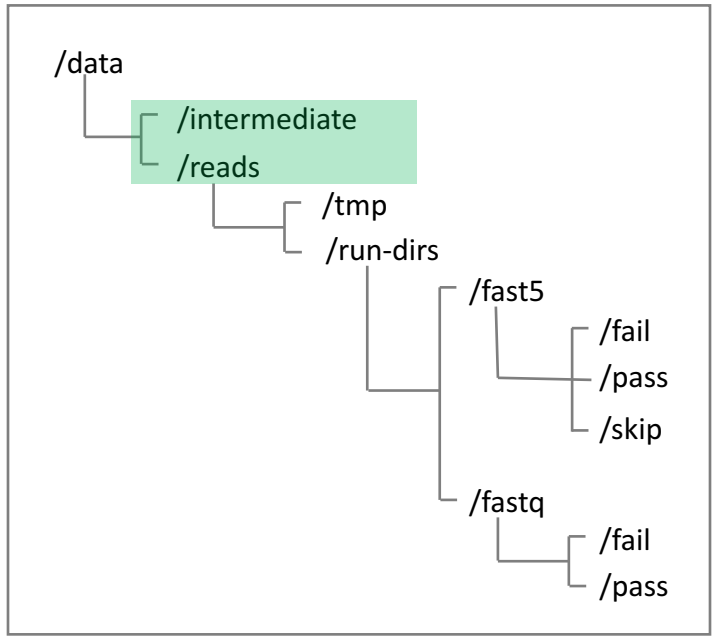
    These Files can not be recovered if MinKNOW fails of exits prematurely!
  - /reads | Directory for raw and base called read files

MinKNOW directory structure

- Default
  - Windows: C:\data
  - MacOSX: /Library/MinKNOW/data
  - Unix: /var/lib/MinKNOW/data



MinKNOW directory structure



- **/intermediate**

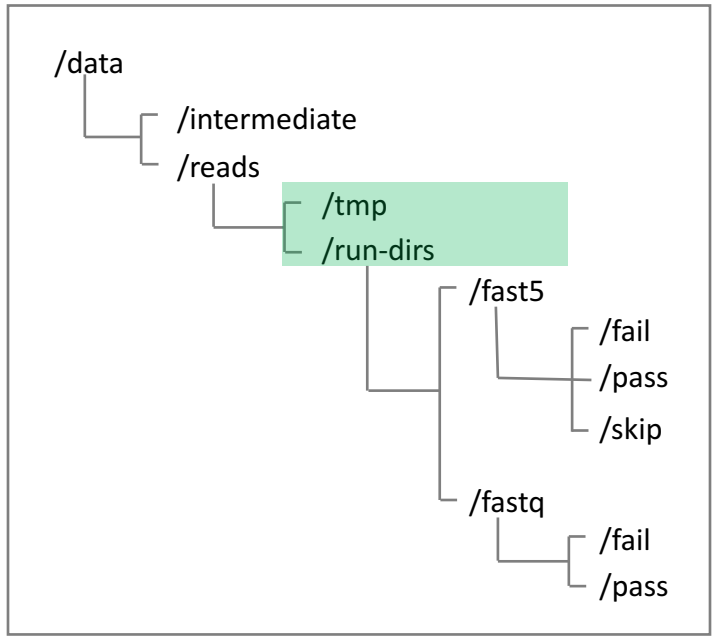
  Raw unprocessed signal files

  These can not be recovered if MinKNOW fails or exits unexpectedly

- **/reads**

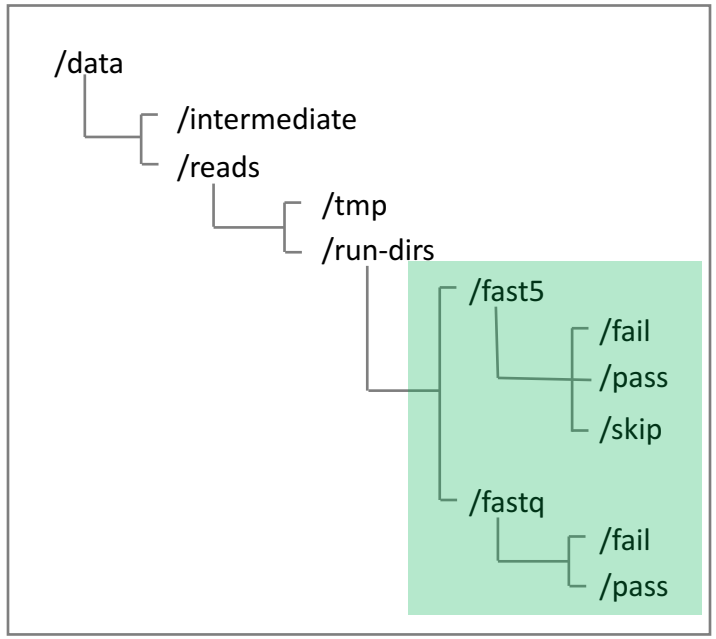
  Directory for processed signal files, individual sequencing runs etc

MinKNOW directory structure



- **/tmp**
  - Directory for not yet base called fast5 files.
  - After base calling files will be re-named from *fast5.tmp* to *fast5* and copied to *sequence-run/fast5* folder
  - **Files can be used for local base calling if MinKNOW fails or exits unexpectedly**
- **/run-dirs**
  - Directories of individual sequencing runs usually named *yyyymmdd_sampleID_id*
  - Containing base called *fast5* and *fastq* files

MinKNOW directory structure



- **/fast5**

  Containing base called reads in *fast5* format, one read per file.

- ***/fastq***

  Containing base called reads in fastq format, 4,000 reads per file

- **/fail**

  Directory for reads that failed QC

- **/pass**

  Directory of reads that passed QC

- **/skip (only in */fast5/* directory)**

  Directory of reads that MinKNOW could not base call

Questions?