Poretools,Porekit, and Jupyter notebook

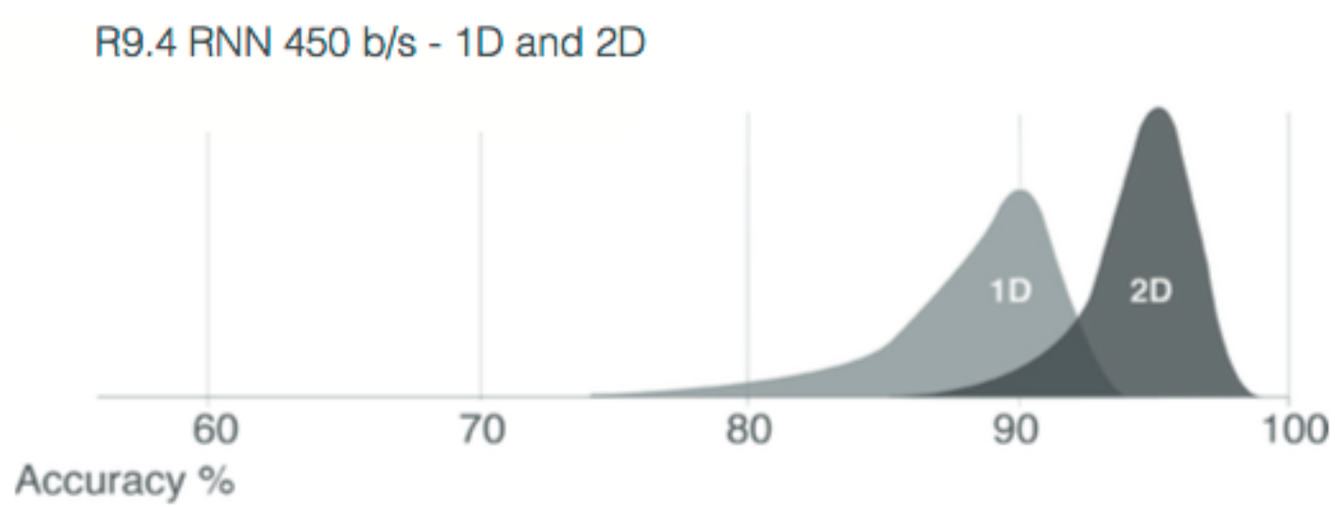
— Exploration tools for Nanopore sequencers —

2016/10/18 Rivas Lab. Meeting

Yosuke Tanigawa
Biomedical Informatics (BMI) program

Oxford Nanopore MinIon

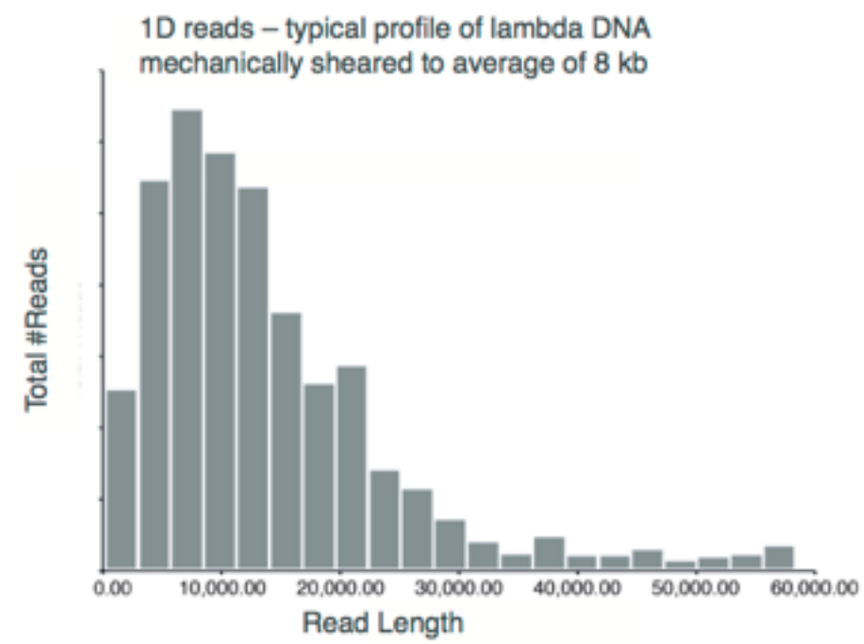
- Handy size DNA sequencer
- Long read sequencer (8kb)
- 3 protocols: 1D, 2D, Rapid
- High error rate
  - 92% (1D), 96% (2D)

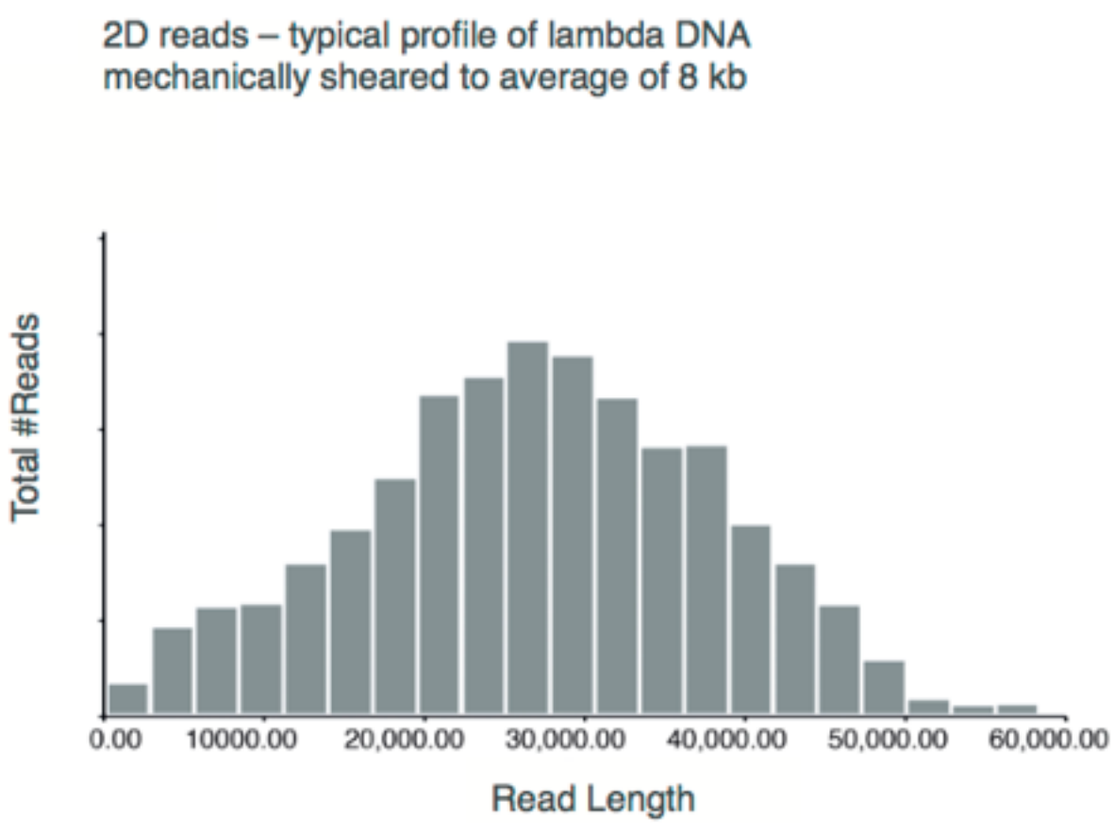


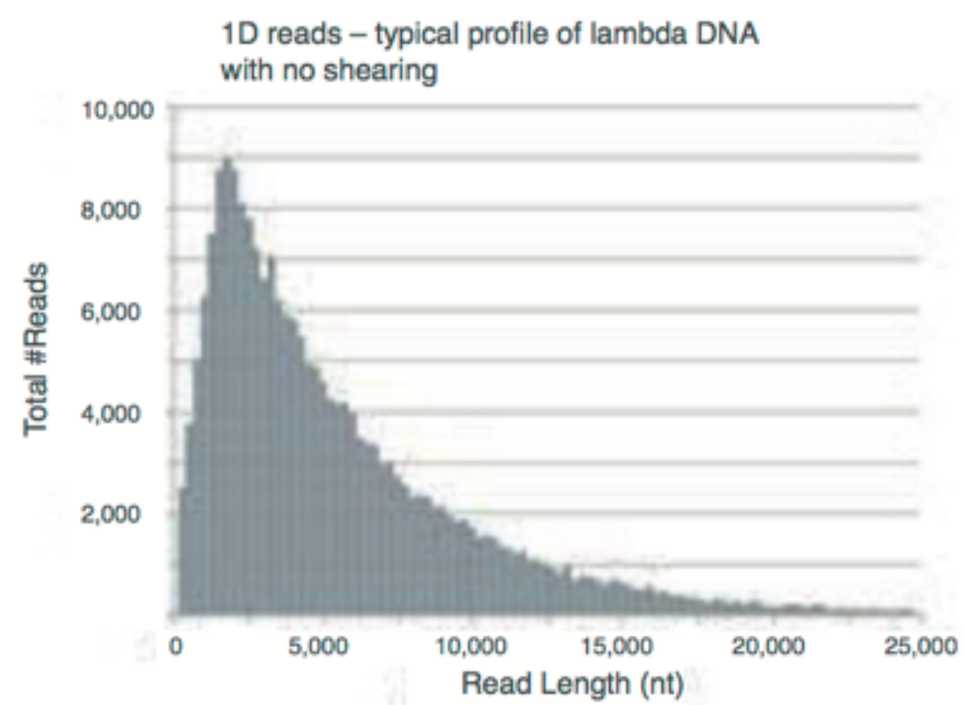
https://nanoporetech.com/products/minion

Read Length Distribution

Mode of the distributions
- 1D: 8kb
- 2D: 28kb
- rapid: 3kb







Poretools

- Provides basic summary statistics from data
  - Developed by Nick Loman and Aaron Quinlan
  - http://doi.org/10.1093/bioinformatics/btu555
  - https://github.com/arq5x/poretools
- Does NOT fully support data from R9 run (the most recent file format)
- https://github.com/rivas-lab/nanopore/blob/master/notes/20161011_poretools.ipynb

Summary Statistics

- Summary stats from two data sets
  - cDNA: 20161006_minion_human_cDNA
  - WGS: 20161008_wgs_caucasian_48hr
- Reads are shorter than expected

	cDNA	WGS
total reads	26,854	29,964
total base	46,314,462	44,839,915
mean	1724.68	1496.46
median	1094	925
min	58	35
max	108262	94024
N25	5201	4547
N50	2529	2227
N75	1327	1140

Porekit

- Metadata extraction and make some nice plots
  - https://github.com/akloster/porekit-python
- https://github.com/rivas-lab/nanopore/blob/master/notes/20161017_porekit.ipynb
- This does not support the most recent file format of Nanopore data

Jupyter Notebook

- Run & Plot python code on your web browser
- You can use it on cluster machine (sherlock, etc..)
- 3 steps
  - start Jupyter notebook on cluster
  - connect to cluster
  - open on your browser

(Step 1) Start Jupiter

yosuke — ytanigaw@sh-5-36:~ — ssh — 80×24

```
 For support:
   * email: research-computing-support@stanford.edu
   * office hours: Tuesdays 10-11am, Thursdays 3-4pm, room 261 @ Polya Hall
-------------------------------------------------------------------------------

[ytanigaw@sherlock-ln03 login_node ~]$ sdev
[ytanigaw@sh-5-36 ~]$ ml load anaconda/anaconda3
[ytanigaw@sh-5-36 ~]$ jupyter-notebook --no-browser
[W 15:30:07.921 NotebookApp] Unrecognized JSON config file version, assuming ver
sion 1
[I 15:30:09.212 NotebookApp] [nb_conda_kernels] enabled, 2 kernels found
[I 15:30:09.831 NotebookApp] ✓ nbpresent HTML export ENABLED
[W 15:30:09.831 NotebookApp] ✗ nbpresent PDF export DISABLED: No module named 'n
bbrowserpdf'
[I 15:30:09.839 NotebookApp] [nb_conda] enabled
[I 15:30:09.927 NotebookApp] [nb_anacondacloud] enabled
[I 15:30:09.933 NotebookApp] Serving notebooks from local directory: /home/ytani
gaw
[I 15:30:09.933 NotebookApp] 0 active kernels
[I 15:30:09.933 NotebookApp] The Jupyter Notebook is running at: http://localhos
t:8888/
[I 15:30:09.933 NotebookApp] Use Control-C to stop this server and shut down all
 kernels (twice to skip confirmation).
```

(Step 2) Connect

```
yosuke — ytanigaw@sh-5-36:~ — ssh — 80×24
ytanigaw@sh-5-36:~        ytanigaw@sh-5-36:~
DN52ec7g:~ yosuke$ PORT=$((30000+RANDOM%29999))
DN52ec7g:~ yosuke$ ssh -t -L 8888:localhost:$PORT sherlock ssh -L $PORT:localhos
t:8888 sh-5-36
Last login: Wed Oct 12 16:26:42 2016 from sherlock-ln01.local
Rocks Compute Node
Rocks 6.1 (Emerald Boa)
Profile built 17:35 28-Jul-2016

Kickstarted 17:41 28-Jul-2016
[ytanigaw@sh-5-36 ~]$
```

Port Forwarding (forward HTTP)

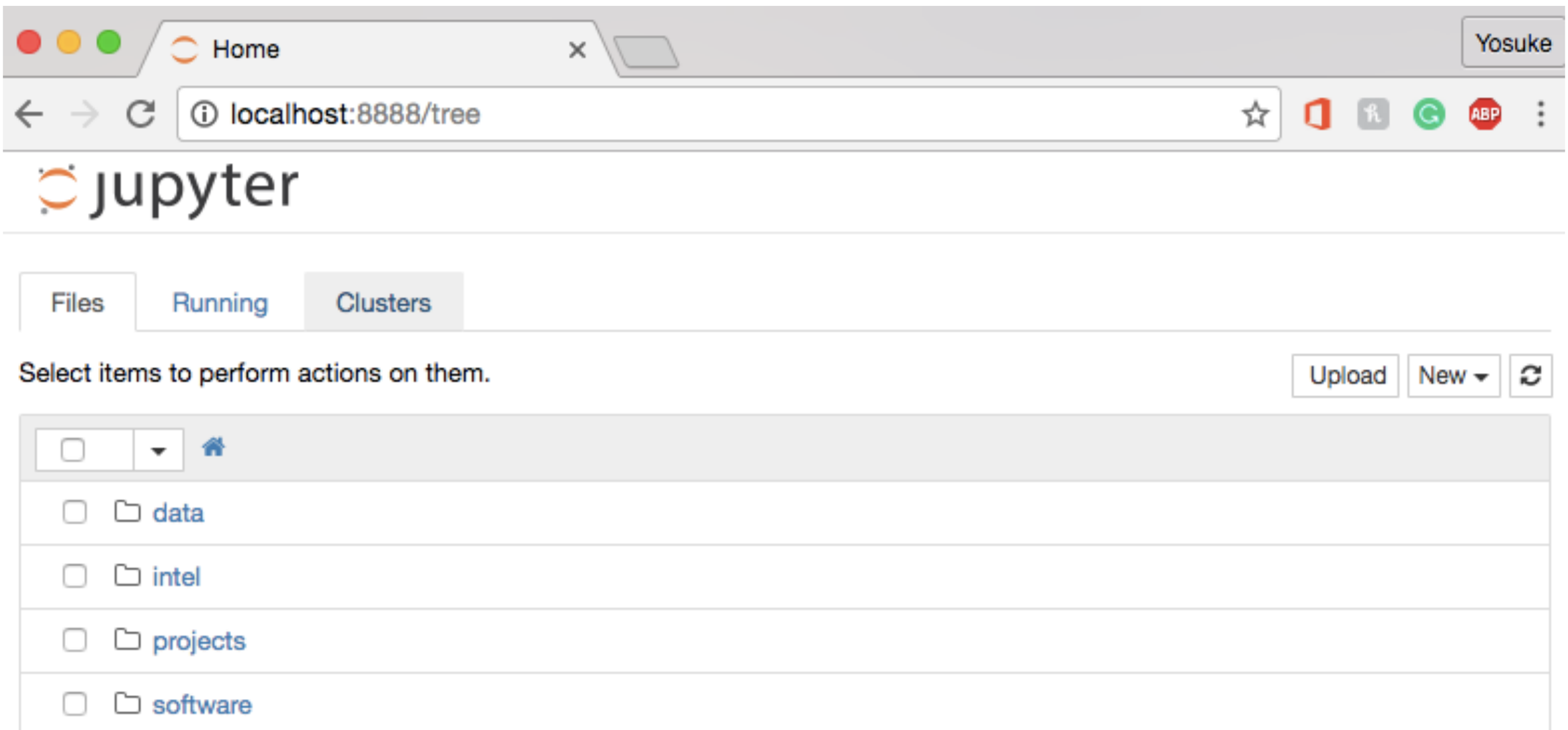
My Laptop <—> sherlock-ln03 <—> sh-5-36

http://sherlock.stanford.edu/mediawiki/index.php/Ipython_Notebooks

We should use anaconda module instead of python/2.7.5

(Step3) Open in browser

Open http://localhost:8888/



Example usage

20161011_poretools

localhost:8888/notebooks/projects/nanopore/notes/20161011_poretools.ipynb

jupyter 20161011_poretools Last Checkpoint: Last Wednesday at 5:12 PM (autosaved)

File Edit View Insert Cell Kernel Help Python 2

Markdown CellToolbar

```
Submitted batch job 10161819
```

Summary stats of '20161006_minion_human_cDNA'

- the batch job described above provided the summary statistics

In [3]:
```
!cat /home/ytanigaw/projects/nanopore/scripts/20161011/poretools-q-20161006_minion_human_cDNA.er
```

In [4]:
```
!cat /home/ytanigaw/projects/nanopore/scripts/20161011/poretools-q-20161006_minion_human_cDNA.ou
```

```
total reads     26854
total base pairs        46314462
mean    1724.68
median  1094
min     58
max     108262
N25     5201
N50     2529
N75     1327
```

(2) 20161008_wgs_caucasian_48hr

- similary, I wrote a script `/home/ytanigaw/scripts/20161011/poretools-q-20161008_wgs_caucasian_48hr.sbatch` and obtained the summary stats. for this sample

In [1]:
```
!cat /home/ytanigaw/projects/nanopore/scripts/20161011/poretools-q-20161008_wgs_caucasian_48hr.e
```