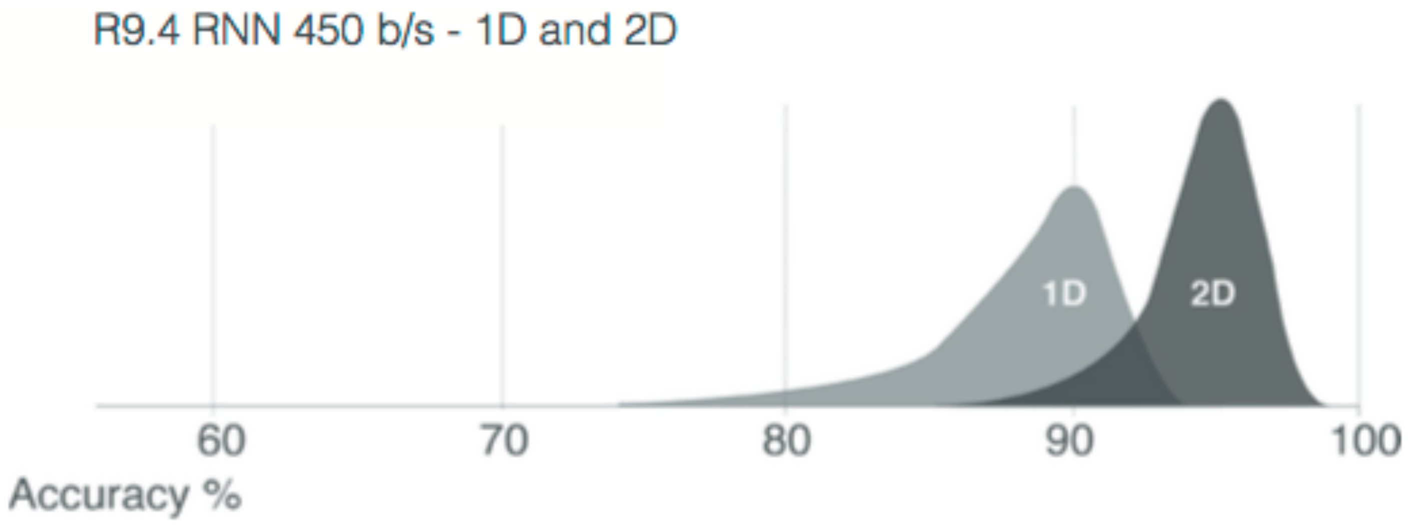
Musing research activities with long read sequence data

Yosuke Tanigawa (ytanigaw@stanford.edu)
M. Rivas Lab (rotation student) | Biomedical Informatics Ph.D. Program

Oxford Nanopore MinIon

- Handy size DNA sequencer
- Long read sequencer (8kb)
- 3 protocols: 1D, 2D, Rapid
- High error rate
  - 92% (1D), 96% (2D)
- R9.4 chemistry



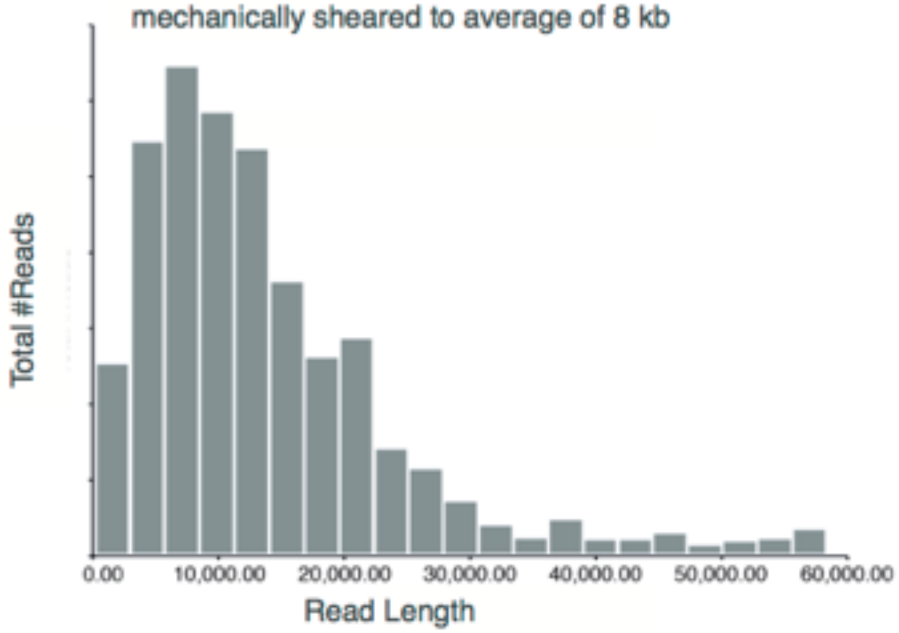
https://nanoporetech.com/products/minion

Read Length Distribution

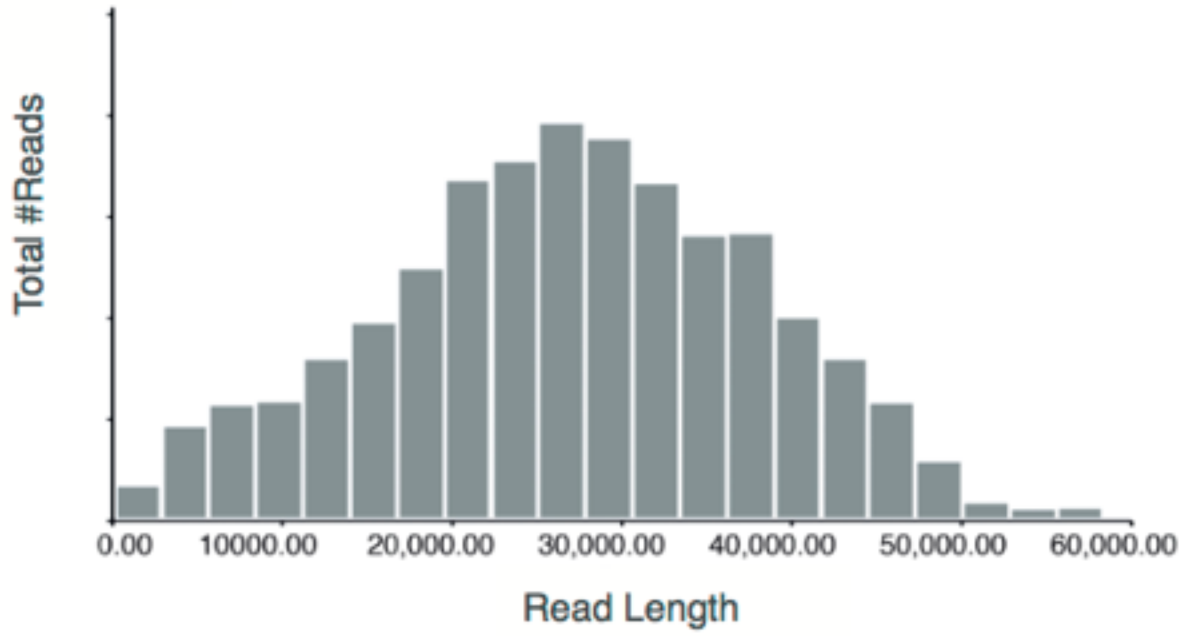
Mode of the distributions

- 1D: 8kb
- 2D: 28kb
- rapid: 3kb

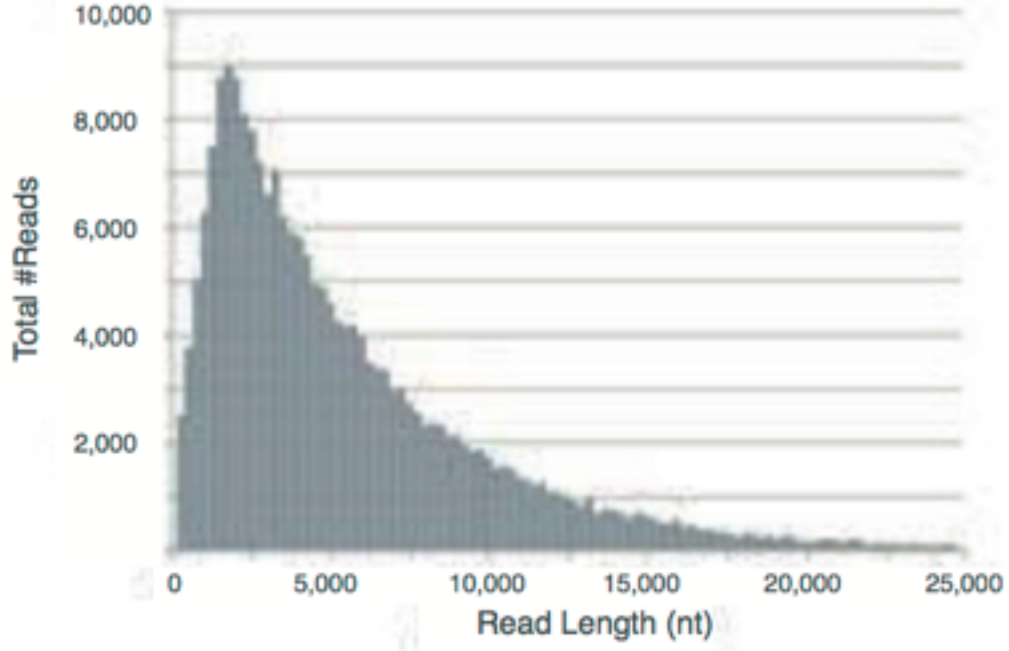
1D reads – typical profile of lambda DNA mechanically sheared to average of 8 kb



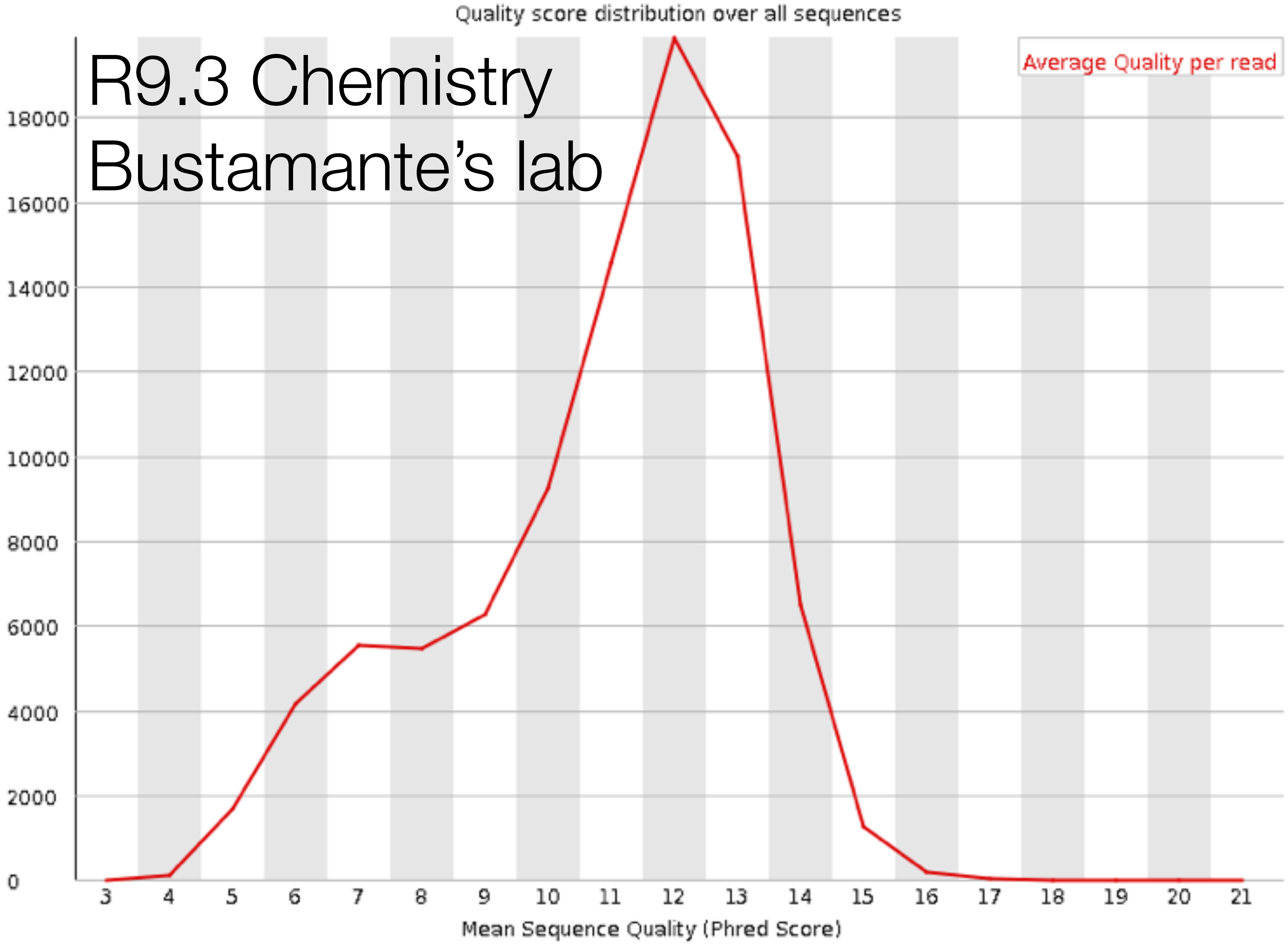
2D reads – typical profile of lambda DNA mechanically sheared to average of 8 kb



1D reads – typical profile of lambda DNA with no shearing



Error rate is around 10%: Quality score freq. dist.



R9.3 Chemistry
Bustamante’s lab

Related research questions for long-read data

- How can we de-noise long-read data
  - Can we learn error profile?
  - Can we fix homopolymers?
    - Improve event calling step in base calling
- Variant calling from collection of mapped reads
  - How can we improve accuracy?

[Resources]
Long read: analysis tools & data sets

- Tools
  - poretools: **sequence extraction** & statistics (partial support for R9.4)
  - porekit: statistics & plots (partial support for R9.4 chem.)
  - minoTour: real-time data analysis (complicated installation)
  - nanopolish: compute consensus by HMM (de-noising)
  - AlignQC: quality assessment of alignment (error model)
  - DeepNano: base caller with Deep RNN (overfitting to a specific dataset?)
- Data sets (nanopore MinION & PacBio)
  - NA12878 on MinION by Oxford Nanopore
  - NA12878 on MinION by Wellcome Trust Centre for Human Genetics
  - NA12878 on PacBio by Genome in a Bottle Consortium (Mt.Sinai)

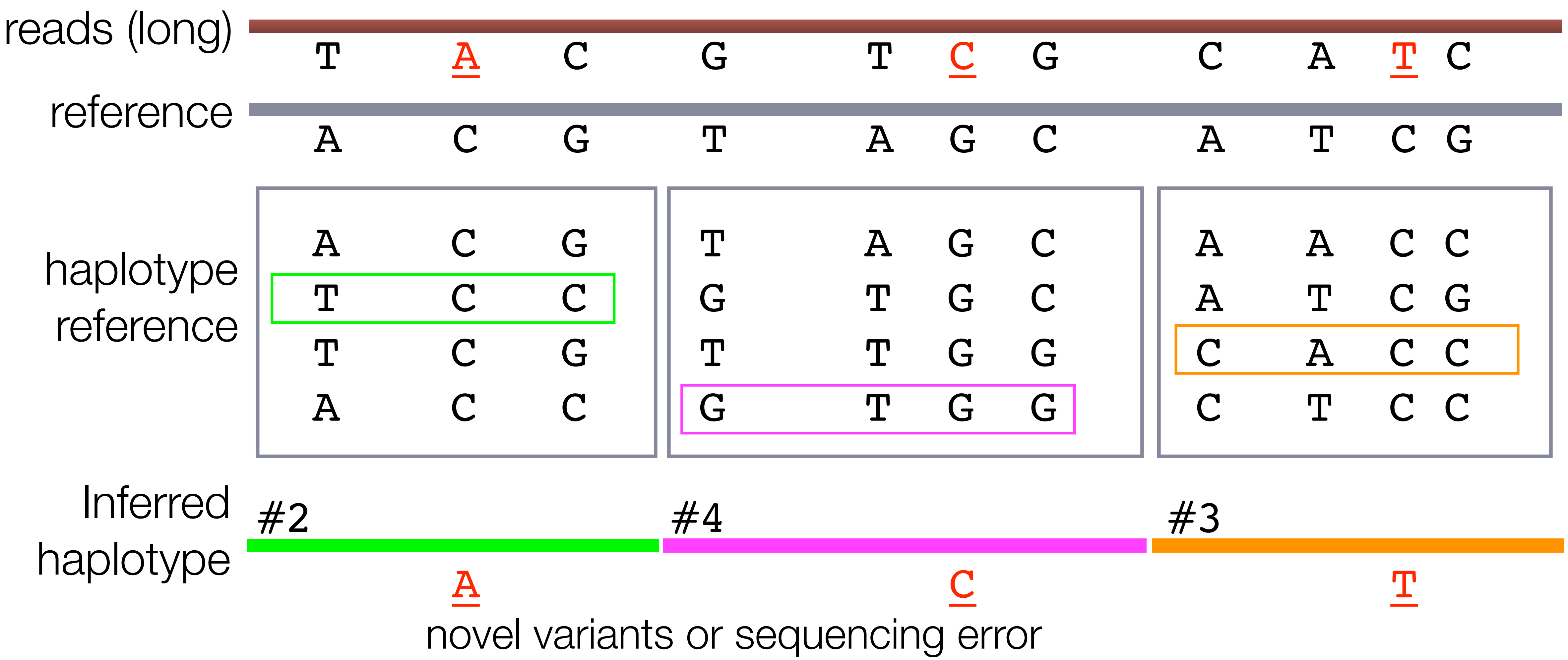
My project
Long read sequencer & Haplotype reference panel

- Long read sequencers (Oxford Nanopore, PacBio)
  - Read length >= 8 kb
  - Review paper of Nanopore

- Haplotype reference
  - UK BioBank (152,729 imputed haplotypes)
  - Haplotype reference consortium (not available yet?)
- Compressed data representation of haplotypes
  - PLINK2

[Research Question]
Compressed representation of personal genomes

- How can we compress personal human genome ?



reads (long)

T A C G T C G C A T C

reference

A C G T A G C A T C G

haplotype reference

A	C	G	T	A	G	C	A	A	C	C
T	C	C	G	T	G	C	A	T	C	G
T	C	G	T	T	G	G	C	A	C	C
A	C	C	G	T	G	G	C	T	C	C

Inferred haplotype

#2 #4 #3

A C T

novel variants or sequencing error

- Indices for haplotypes + novel variants

Compressed representation of personal genomes

- Validating the method with NA12878
- How can we discriminate variants from sequence errors?
- Is it possible to infer haplotypes on the fly?
- Is it scalable to a large population (say 100M)?
- Examples of applications:
    - Haplotyping for forensic science
    - Haplotyping on HLA region
    - Clinical applications