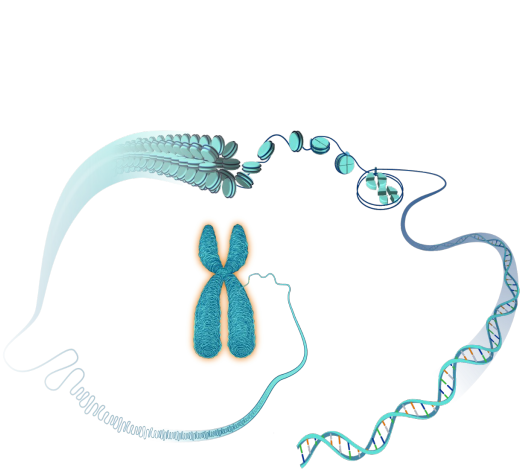
Introduction to Long Read Data Formats

Álvaro Cortés Calabuig

In this presentation

Ideally…



AACCTGGCTG…

However, sequencers:

- Unable to sequence whole chromosomes
- Work with what they are given:
  - Bulk DNA
  - Degraded DNA
  - Contaminated DNA
- Make mistakes: "the longer the read the higher % errors"

From 'raw' to 'raw': Base calling

Illumina

Pacbio

ONT

BCL Files

ZMW "Movie" Files

Fast5/Pod5

Base calling: from raw to text

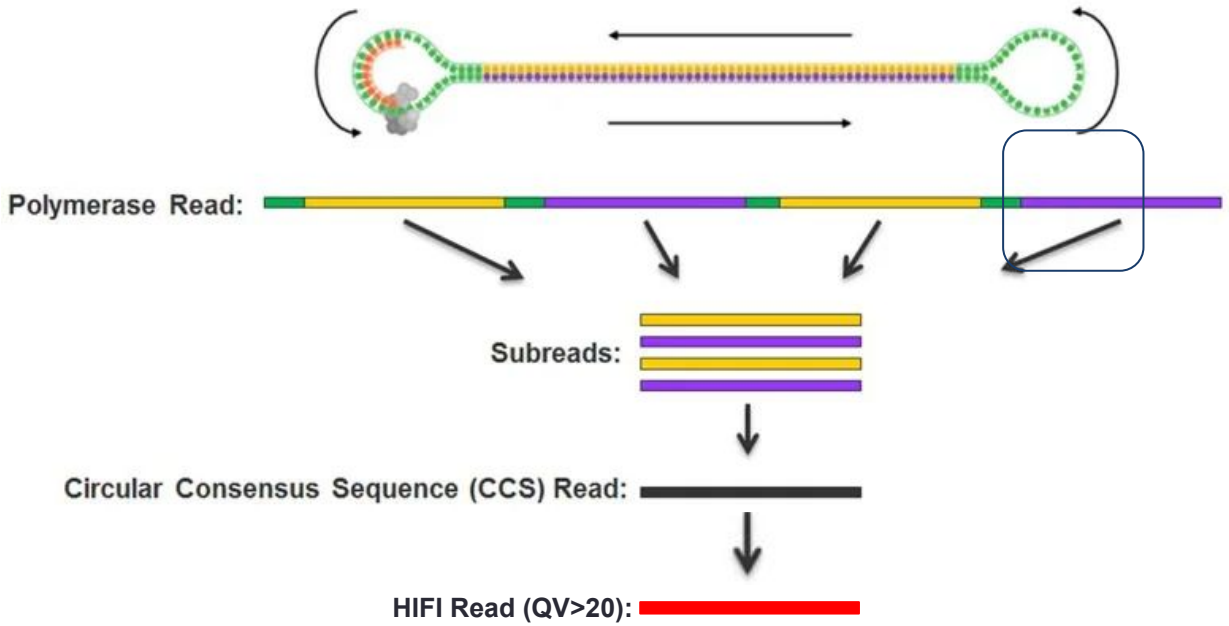
bcl2fastq

Smrt Link

Dorado

Fastq/uBAM

Pacific Biosciences



- `<movie>.`Subreads.bam/fastq: range from 0.5 TB to 1.5 TB
- `<movie>.subreads.`reads.bam/fastq: both HiFi Reads (≥QV 20) and non-HiFi read -> ccs reads
- `<movie>.`Hifi_reads.bam/fastq

- File structure different on Sequel II, Sequel IIe and Revio
- Files to use is analysis dependent: e.g. denovo assembly: subreads

Base Calling Oxford Nanopore

Conductivity (or Resistance) Varies by base

- Direct measurement of the changes in ionic current as a DNA/RNA strand passes through the pore
- Current (pico amperes) is measured: “Squiggles”
- Conductivity is calculated and stored as a 16-bit Integer value
- Base calling transforms measured conductivity into bases

ONT Base Calling

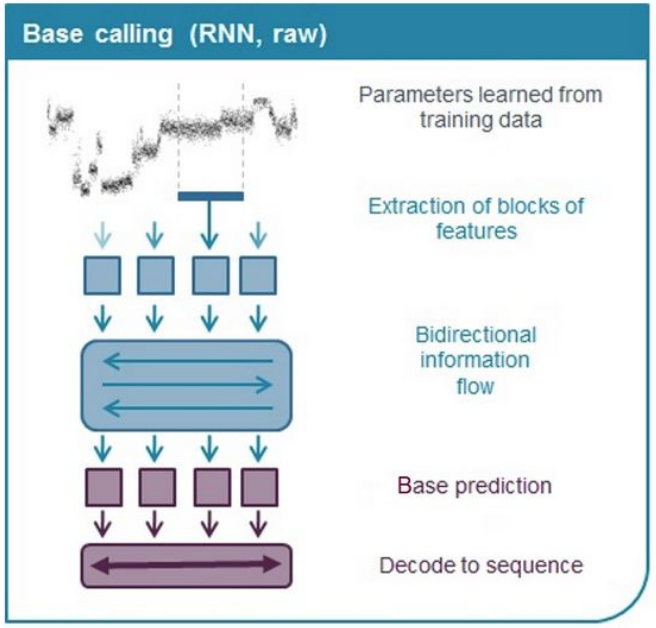
Why is ONT base calling difficult?

- Electrical signals come from single molecules:
  - Noisy
  - Stochastic data
- The electrical resistance of a pore is determined by the bases present within multiple nucleotides that reside in the pore's narrowest point
  - Approximately five nucleotides for pore, yielding a large number of possible states:
    - $4^5 = 1024$ for a standard four-base model

This makes basecalling of ONT device signals a suitable application of machine learning techniques

Base Calling Oxford Nanopore

At the Heart of a Nanopore Sequencer Runs MinKnow Software



- Data acquisition, real-time analysis and feedback
- Local base calling
- MinKnow produces Fast5/pod5 (raw) and Fastq files

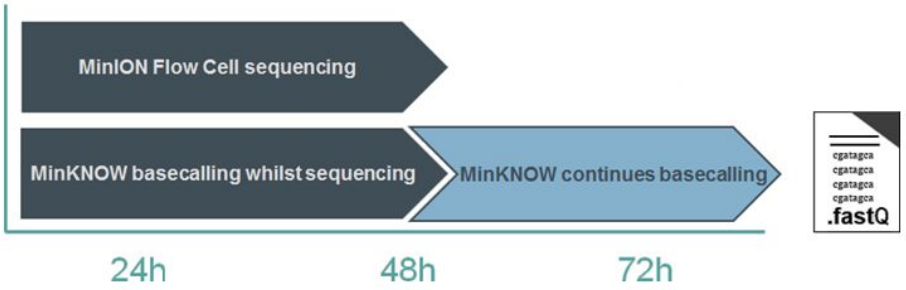
Fast5 Files

Raw Data of Oxford Nanopore: Fast5 or Pod5

- The fast5 format is a specification over a HDF5 file (Data Type and Data Format Specification)
- Each read resulting from sequencing a molecule is stored as its own fast5/pod5 file
- $ h5dump -g "/Raw" read.fast5
- The main data is the “squiggle”

Dorado and MinKNOW

- ONT's Dorado (succesor of Guppy):
  - Latest Nanopore's base calling software
  - Stand alone package as well as incorporated on MinKnow
  - Three running modes:
    - Fast model: Balance between accuracy and speed
    - HAC: higher raw read accuracy than the Fast model and is more computationally-intensive.
    - Super accurate: model has an even higher raw read accuracy, and is even more intensive than the HAC model.



Fastq and (u)BAM/Sam

- After base calling, ONT and Pacbio sequencers produce standard Fastq and (u)BAM/SAM files

- ZMW Files

Smrt Link

- Subreads
- CCS
- Hifi

Fastq (u)BAM

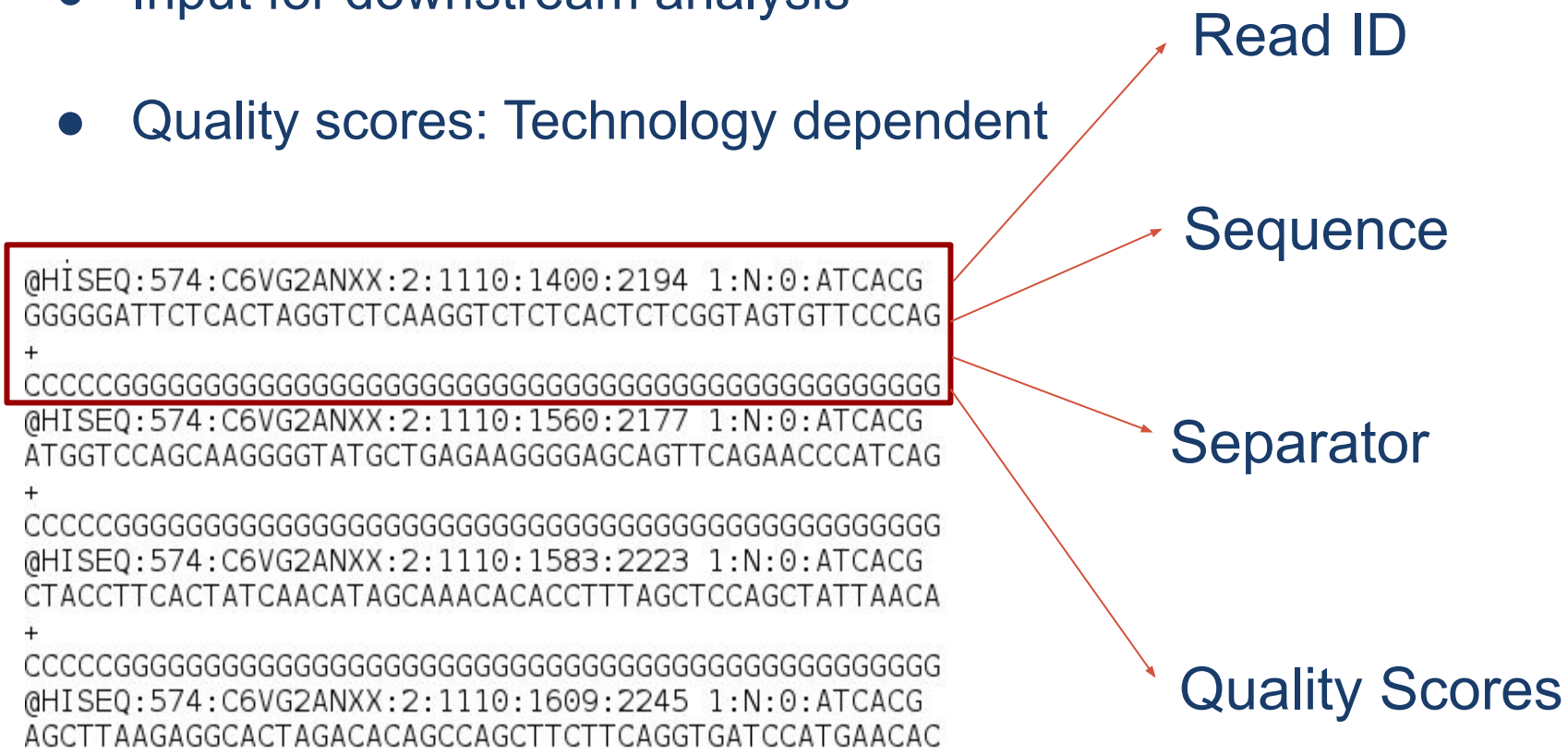
- Fast5
- Pod5

Dorado

- Fastq
- (u)BAM

Fastq Format

- Human readable (often .fastq.gz format)
- Input for downstream analysis
- Quality scores: Technology dependent



BAM/SAM Format

- SAM: Sequence Alignment Map file
- BAM files are machine-readable versions of SAM files

```
HISEQ:574:C6VG2ANXX:3:1307:21149:38188  0       1       3216913 50      51M     *       00
CTGGTAGGAGGCTAGGGCCCAAGCCAAGGACACAAGGGAGGCTGCTGCTGT     BBBCBGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGFGGGGGGGGGGGG       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:-  NH:i:1  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:2208:7076:98530   16      1       3532557 3       51M     *       00
TGTTCGGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTAT     GGBGGGGGGGGGGGGGGBGGGGGGGGG
GFCGGGGGGGGGGGGGGGGGBCBBC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736175  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:2110:2743:48952   272     1       3532562 3       51M     *       00
GGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTATTTAAG     GGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGCCCCC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736170  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:1115:10182:55408  16      1       3532562 3       51M     *       00
GGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTATTTAAG     GGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGCCCCC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736170  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:1107:5697:49698   16      1       3532562 3       51M     *       00
GGACACTTCAAGTTCTGCATCACTCTCTGCGGAGGATACATTCTATTTAAG     GGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGCCCCC       AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:51 YT
:Z:UU   XS:A:+  NH:i:2  CC:Z:10 CP:i:117736170  HI:i:0  RG:Z:GC024982
HISEQ:574:C6VG2ANXX:3:1315:2671:58947   256     1       3592903 1       51M     *       00
TTAAGACTGAATTCTGACATAGCTAAAGCCTTCCGCCAGTGTTCCAACAGT     BBCCCGGGGGGGGGGGGGGGGGGFGG
GGGGGGGGGGGGGGGGGGGGGGGGG       AS:i:-10        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD
:Z:6T1A42       YT:Z:UU XS:A:-  NH:i:3  CC:Z:15 CP:i:10497071   HI:i:0  RG:Z:GC024982
```

(u)BAM files are used by Minknow and Smrt Link to store methylation calls from unaligned reads

●

This information is not stored on fastq files!

Modifications are Stored in SAM/BAM Files

- Each modified base prediction listed has a quality value associated with it.
-
- Base modification quality should be interpreted as the likelihood of this modification being correct given an assumption the original call is correct.

- Base modification:

  MM:Z:([ACGTUN][-+]([a-z]+|[0-9]+)[.?]?(,[0-9]+)*;)*

- Modification quality:

  ML:B:C,scaled-probabilities

Base Modifications

Unmodified base	Code	Abbreviation	Name	ChEBI
C	m	5mC	5-Methylcytosine	27551
C	h	5hmC	5-Hydroxymethylcytosine	76792
C	f	5fC	5-Formylcytosine	76794
C	c	5caC	5-Carboxylcytosine	76793
C	C		Ambiguity code; any C mod	
T	g	5hmU	5-Hydroxymethyluracil	16964
T	e	5fU	5-Formyluracil	80961
T	b	5caU	5-Carboxyluracil	17477
T	T		Ambiguity code; any T mod	
U	U		Ambiguity code; any U mod	
A	a	6mA	6-Methyladenine	28871
A	A		Ambiguity code; any A mod	
G	o	8oxoG	8-Oxoguanine	44605
G	G		Ambiguity code; any G mod	
N	n	Xao	Xanthosine	18107
N	N		Ambiguity code; any mod	

Example MM:Z: field

C+m,5,12,0;

- There are three potential 5-Methylcytosine bases on the top strand of SEQ.
-
- The first 5 ‘C’ bases are unmodified and the 6th, 19th and 20th have modification status indicated by the corresponding probabilities in the ML tag
- The 12 cytosines between the 6th and 19th cytosine are unmodified
- Modification probabilities for the 17 skipped cytosines are not provided.

Thank you!

GENOMICS CORELEUVEN