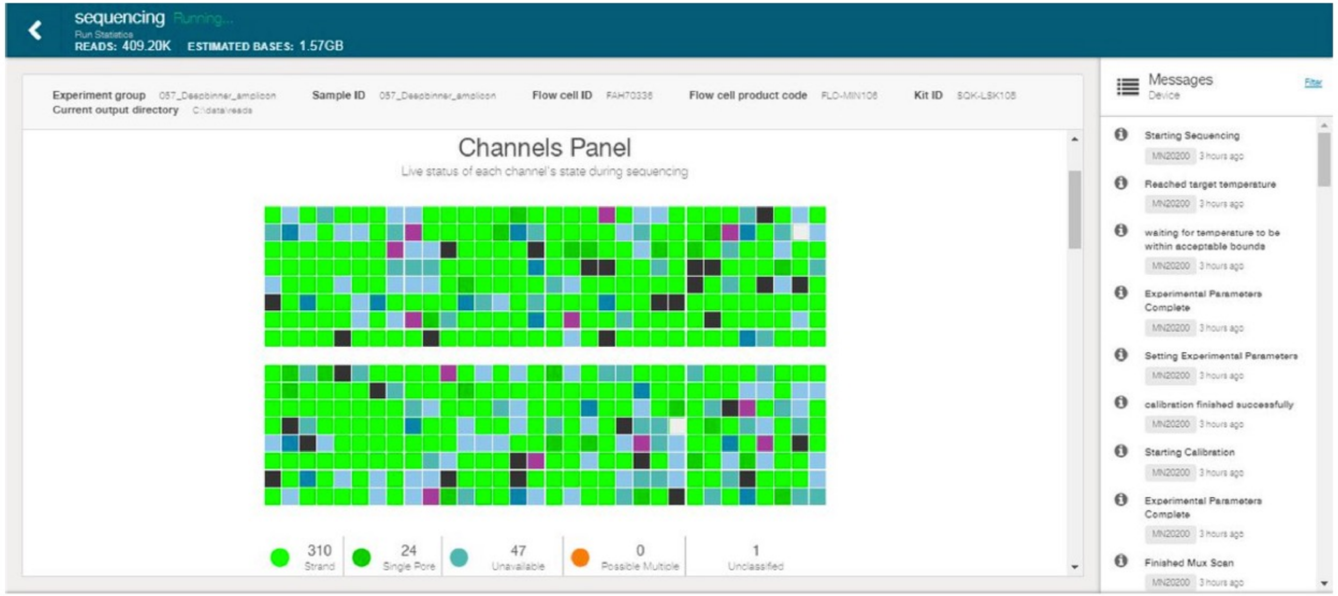
Sequence data

Kirstyn Brunker

RAGE workshop, NCDC, Abuja

12-16th Feb 2024

MinKNOW



sequencing Running...

Run Statistics

READS: 409.20K ESTIMATED BASES: 1.57GB

Experiment group 057_Deepbinner_amplicon
Current output directory C:\data\reads

Sample ID 057_Deepbinner_amplicon

Flow cell ID FAH70336

Flow cell product code FLO-MIN106

Kit ID SQK-LSK108

Channels Panel

Live status of each channel's state during sequencing

● 310 Strand	● 24 Single Pore	● 47 Unavailable	● 0 Possible Multiple	1 Unclassified

Messages
Device

Filter

Starting Sequencing
MN20200 3 hours ago

Reached target temperature
MN20200 3 hours ago

waiting for temperature to be within acceptable bounds
MN20200 3 hours ago

Experimental Parameters Complete
MN20200 3 hours ago

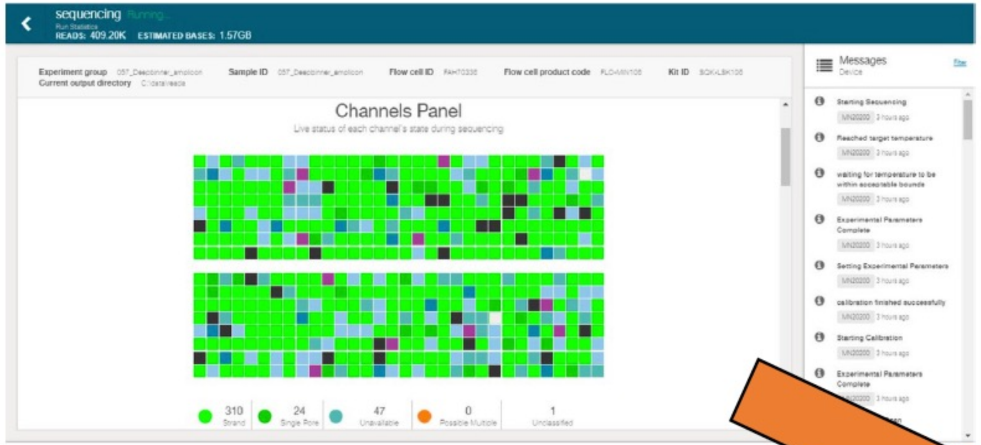
Setting Experimental Parameters
MN20200 3 hours ago

calibration finished successfully
MN20200 3 hours ago

Starting Calibration
MN20200 3 hours ago

Experimental Parameters Complete
MN20200 3 hours ago

Finished Mux Scan
MN20200 3 hours ago



sequencing Running...
Run Statistics
READS: 409.20K ESTIMATED BASES: 1.57GB

Experiment group 057_Deepbinner_amplicon
Current output directory C:\data\reads
Sample ID 057_Deepbinner_amplicon
Flow cell ID FAH70336
Flow cell product code FLO-MIN106
Kit ID SQK-LSK108

Channels Panel
Live status of each channel's state during sequencing

● 310 Strand
● 24 Single Pore
● 47 Unavailable
● 0 Possible Multiple
1 Unclassified

Messages
Device

Starting Sequencing
MN20200 3 hours ago

Reached target temperature
MN20200 3 hours ago

waiting for temperature to be within acceptable bounds
MN20200 3 hours ago

Experimental Parameters Complete
MN20200 3 hours ago

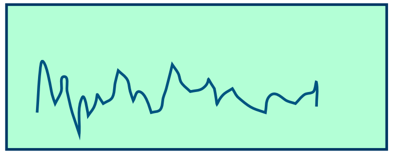
Setting Experimental Parameters
MN20200 3 hours ago

calibration finished successfully
MN20200 3 hours ago

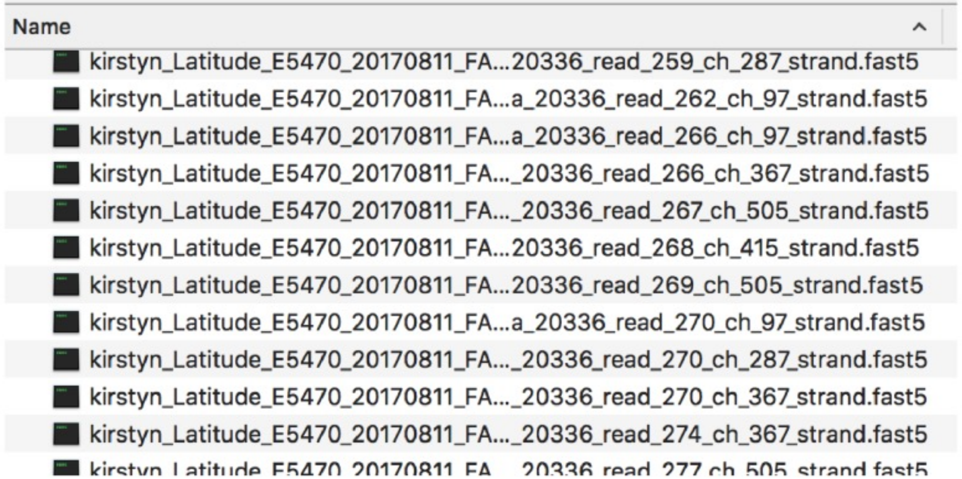
Starting Calibration
MN20200 3 hours ago

Experimental Parameters Complete
MN20200 3 hours ago

Pretty user interface



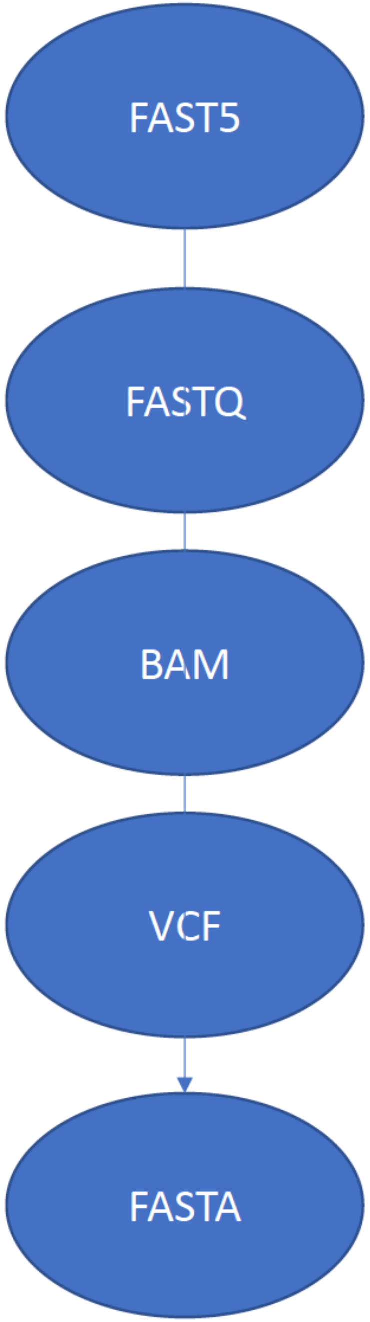
Raw data



Name

```
kirstyn_Latitude_E5470_20170811_FA...20336_read_259_ch_287_strand.fast5
kirstyn_Latitude_E5470_20170811_FA...a_20336_read_262_ch_97_strand.fast5
kirstyn_Latitude_E5470_20170811_FA...a_20336_read_266_ch_97_strand.fast5
kirstyn_Latitude_E5470_20170811_FA..._20336_read_266_ch_367_strand.fast5
kirstyn_Latitude_E5470_20170811_FA..._20336_read_267_ch_505_strand.fast5
kirstyn_Latitude_E5470_20170811_FA...20336_read_268_ch_415_strand.fast5
kirstyn_Latitude_E5470_20170811_FA...20336_read_269_ch_505_strand.fast5
kirstyn_Latitude_E5470_20170811_FA...a_20336_read_270_ch_97_strand.fast5
kirstyn_Latitude_E5470_20170811_FA..._20336_read_270_ch_287_strand.fast5
kirstyn_Latitude_E5470_20170811_FA..._20336_read_270_ch_367_strand.fast5
kirstyn_Latitude_E5470_20170811_FA..._20336_read_274_ch_367_strand.fast5
kirstyn_Latitude_E5470_20170811_FA..._20336_read_277_ch_505_strand.fast5
```

Each fast5 file generated contains 4000 reads



Raw data format

FAST5

Raw signal data from MinION pores

Basecalled reads + quality info (Phred scores)

FASTQ

Text-based format for storing both sequence data and its corresponding quality scores

Binary Alignment/Map files

BAM

A BAM file (.bam) is the binary version of a SAM file, which is a tab-delimited text file that contains sequence alignment data

Variant call format

VCF

Text file used in bioinformatics for storing gene sequence variations

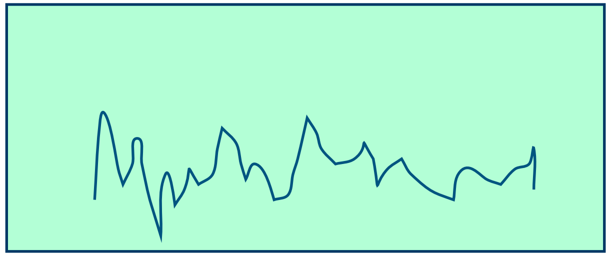
FASTA

Text-based format for storing both sequence data

Fast5

- Raw electical signal data, i.e. **squiggle data**

- HDF5 format: storage and organization of large amounts of heterogeneous data, using a hierarchical structure.



Needs “basecalled”

→

ATTGCCGTAAT….

Fastq

- Common NGS format
- It contains a series of records, where each record represents a single sequence read obtained from the sequencing machine.

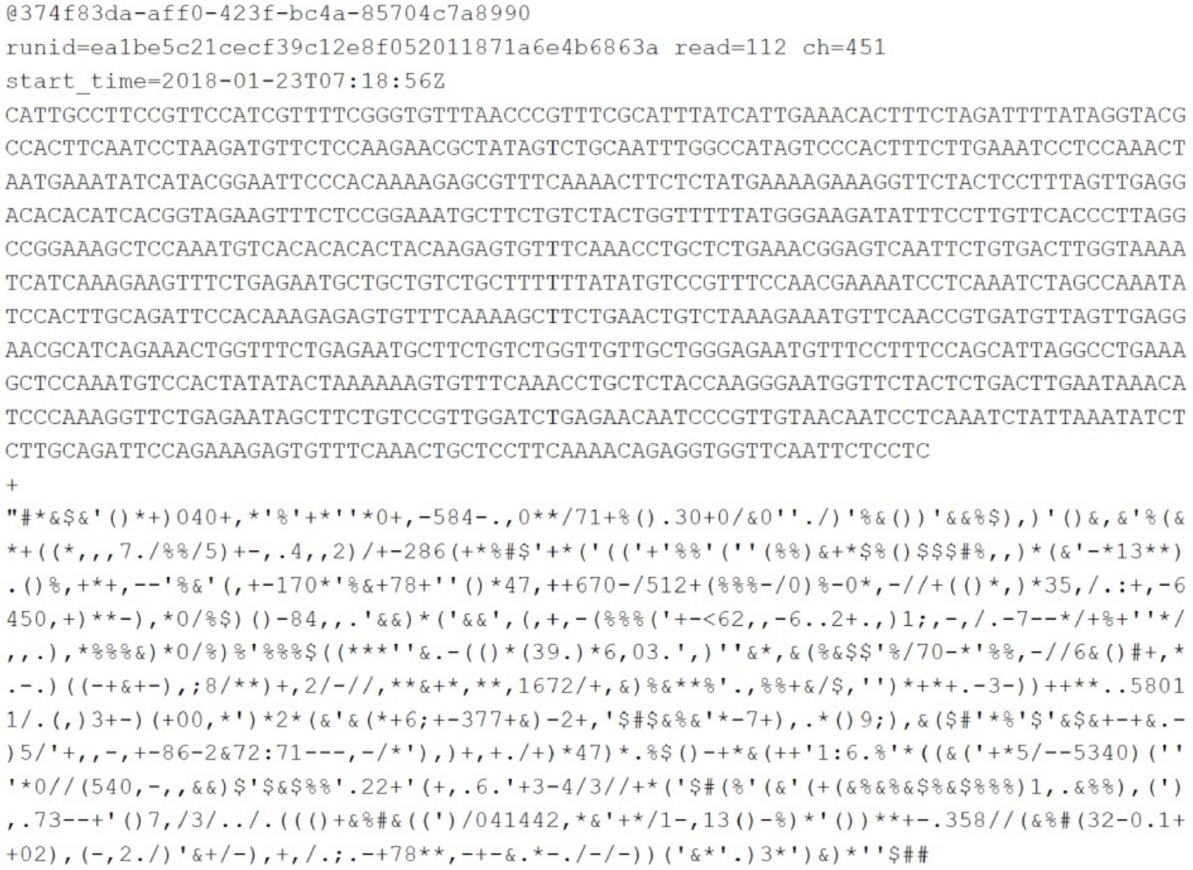
- Each record in the FASTQ file consists of four lines:
  1. sequence identifier
  2. raw DNA sequence
  3. a separator line
  4. quality scores corresponding to each base in the DNA sequence, representing the confidence or accuracy of the base call.

- Each record in the FASTQ file consists of four lines:
  1. sequence identifier
  2. raw DNA sequence
  3. a separator line
  4. quality scores corresponding to each base in the DNA sequence, representing the confidence or accuracy of the base call.

  IMPORTANT: Lines 2 and line 4 must have the same length or the sequence record is not valid.

Fastq files

Header contains:
@<READ_ID> runid=<RUNID> read=<read_number>
ch=<channel_number>
start_time=<start_time>



```
@374f83da-aff0-423f-bc4a-85704c7a8990
runid=ea1be5c21cecf39c12e8f052011871a6e4b6863a read=112 ch=451
start_time=2018-01-23T07:18:56Z
CATTGCCTTCCGTTCCATCGTTTTCGGGTGTTTAACCCGTTTCGCATTTATCATTGAAACACTTTCTAGATTTTATAGGTACG
CCACTTCAATCCTAAGATGTTCTCCAAGAACGCTATAGTCTGCAATTTGGCCATAGTCCCACTTTCTTGAAATCCTCCAAACT
AATGAAATATCATACGGAATTCCCACAAAAGAGCGTTTCAAAACTTCTCTATGAAAAGAAAGGTTCTACTCCTTTAGTTGAGG
ACACACATCACGGTAGAAGTTTCTCCGGAAATGCTTCTGTCTACTGGTTTTTATGGGAAGATATTTCCTTGTTCACCCTTAGG
CCGGAAAGCTCCAAATGTCACACACACTACAAGAGTGTTTCAAACCTGCTCTGAAACGGAGTCAATTCTGTGACTTGGTAAAA
TCATCAAAGAAGTTTCTGAGAATGCTGCTGTCTGCTTTTTTATATGTCCGTTTCCAACGAAAATCCTCAAATCTAGCCAAATA
TCCACTTGCAGATTCCACAAAGAGAGTGTTTCAAAAGCTTCTGAACTGTCTAAAGAAATGTTCAACCGTGATGTTAGTTGAGG
AACGCATCAGAAACTGGTTTCTGAGAATGCTTCTGTCTGGTTGTTGCTGGGAGAATGTTTCCTTTCCAGCATTAGGCCTGAAA
GCTCCAAATGTCCACTATATACTAAAAAAGTGTTTCAAACCTGCTCTACCAAGGGAATGGTTCTACTCTGACTTGAATAAACA
TCCCAAAGGTTCTGAGAATAGCTTCTGTCCGTTGGATCTGAGAACAATCCCGTTGTAACAATCCTCAAATCTATTAAATATCT
CTTGCAGATTCCAGAAAGAGTGTTTCAAACTGCTCCTTCAAAACAGAGGTGGTTCAATTCTCCTC
+
"#*&$&'()*+)040+,*'%'+*''*0+,-584-.,0**/71+%().30+0/&0''./)'%&())'&&%$),)'()&,&'%(&
*+((*,,,7./%%/5)+-,.4,,2)/+-286(+*%#$'+*('(('+'%%'(''(%%)&+*$%()$$$#%,,)*(&'-*13**)
.()%,+*+,--'%&'(,+-170*'%&+78+''()*47,++670-/512+(%%%-/0)%-0*,-//+(()*,)*35,/.:+,-6
450,+)**-),*0/%$)()-84,,.'&&)*('&&',(,+,-(%%%('+-<62,,-6..2+.,)1;,-,/.-7--*/+%+''*/
,,.),*%%%&)*0/%)%'%%%$((***''&.-(()*(39.)*6,03.',)''&*,&(%&$$'%/70-*'%%,-//6&()#+,*
.-.)((-+&+-),;8/**)+,2/-//,**&+*,**,1672/+,&)%&**%'.,%%+&/$,'')*+*+.-3-))++**..5801
1/.(,)3+-)(+00,*')*2*(&'&(*+6;+-377+&)-2+,'$#$&%&'*-7+),.*()9;),&($#'*%'$'&$&+-+&.-
)5/'+,,-,+-86-2&72:71---,-/*'),)+,+./+)*47)*.%$()-+*&(++'1:6.%'*((&('+*5/--5340)(''
'*0//(540,-,,&&)$'$&$%%'.22+'(+,.6.'+3-4/3//+*('$#(%'(&'(+(&%&%&$%&$%%%)1,.&%%),(')
,.73--+'()7,/3/.././((()+&%#&((')/041442,*&'+*/1-,13()-%)*'())**+-.358//(&%#(32-0.1+
+02),(-,2./)'&+/-),+,/.;.-+78**,-+-&.*-./-/-))('&*'.)3*')&)*''$##
```

Sequence

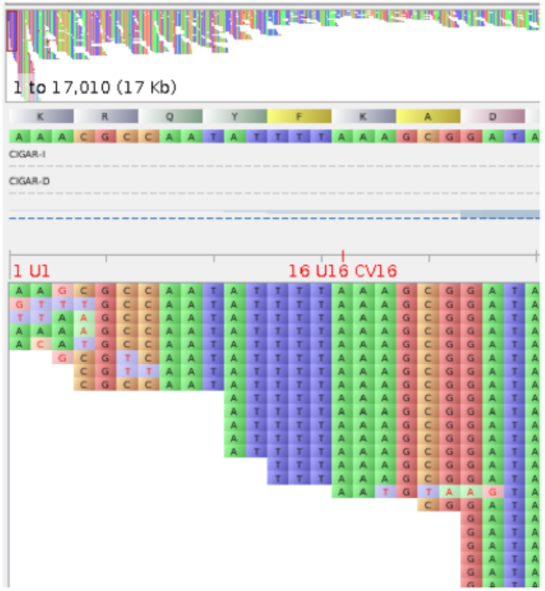
Separator

Quality data

BAM files

- Binary file format commonly used in genomics to store DNA sequence alignment data obtained from NGS
- It contains aligned reads, which are short DNA sequences from the original sample, along with information about their mapping locations to a reference genome
- The files are compressed and structured- so efficient for storage and analysis of large-scale sequencing data
- Allow various bioinformatics tools and algorithms to work with the aligned data efficiently

BAM



Have to be sorted and indexed for visualization in different programs

e.g. Tablet,

BAM files contain a header section and an alignment section:

Header—Contains information about the entire file, such as sample name, sample length, and alignment method. Alignments in the alignments section are associated with specific information in the header section.

Alignments—Contains read name, read sequence, read quality, alignment information, and custom tags. The read name includes the chromosome, start coordinate, alignment quality, and the match descriptor string.

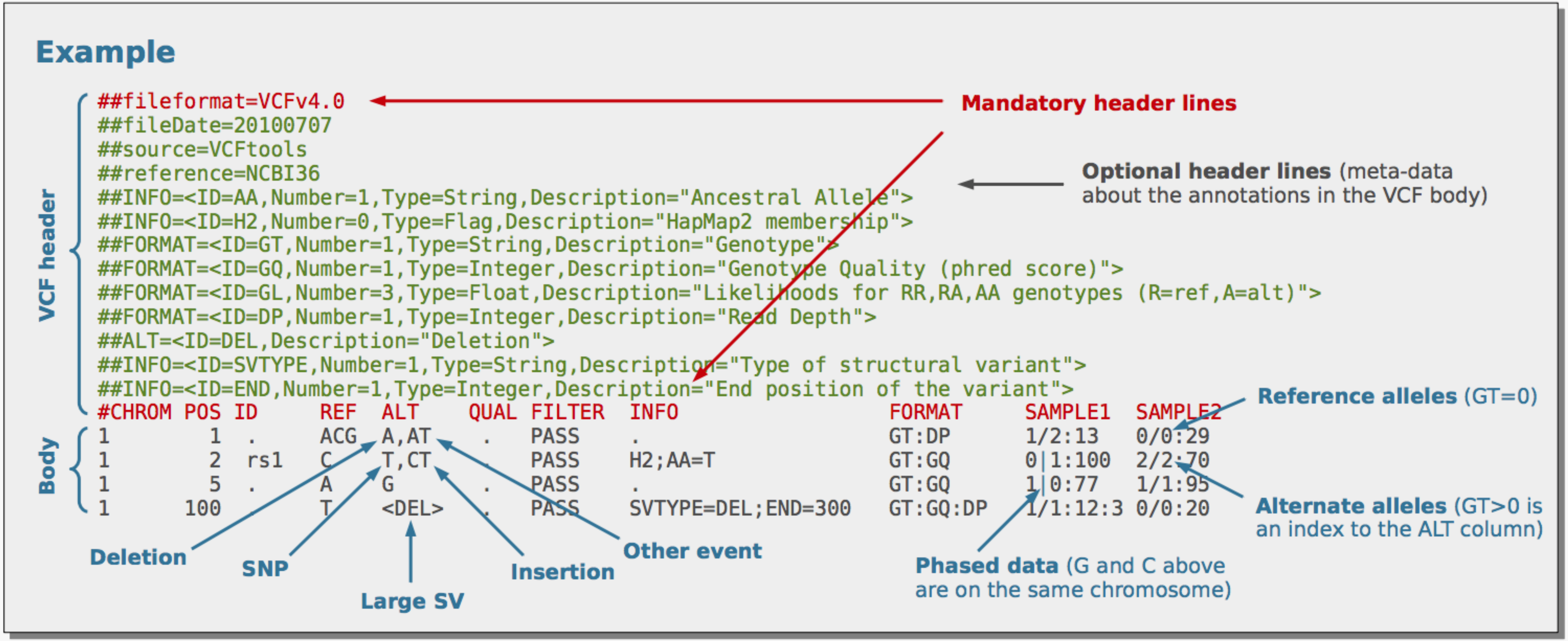
VCF (Variant Call Format)

- Represent genetic variations identified in DNA sequencing data.

- Information about genetic variants such as single nucleotide polymorphisms (SNPs), insertions, deletions (INDELS)

- Used for variant analysis, genotyping, and variant calling

Example



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER INFO                 FORMAT     SAMPLE1  SAMPLE2
1        1  .     ACG  A,AT    .   PASS    .                    GT:DP      1/2:13   0/0:29
1        2  rs1   C    T,CT    .   PASS    H2;AA=T              GT:GQ      0|1:100  2/2:70
1        5  .     A    G       .   PASS    .                    GT:GQ      1|0:77   1/1:95
1      100  .     T    <DEL>   .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3 0/0:20
```

VCF header

Body

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

Large SV

Insertion

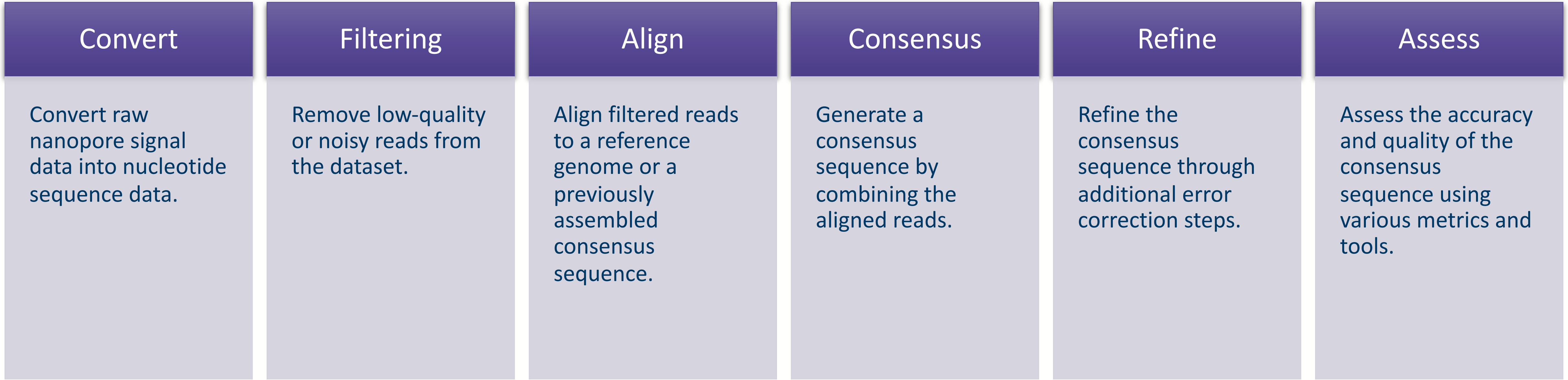
Other event

Getting to a consensus

- Think about what you did to prepare your DNA for the MinION

- Added barcodes and adaptors

Getting to a consensus



Convert

Convert raw nanopore signal data into nucleotide sequence data.

Filtering

Remove low-quality or noisy reads from the dataset.

Align

Align filtered reads to a reference genome or a previously assembled consensus sequence.

Consensus

Generate a consensus sequence by combining the aligned reads.

Refine

Refine the consensus sequence through additional error correction steps.

Assess

Assess the accuracy and quality of the consensus sequence using various metrics and tools.

Advice

- Get an understanding of how the data needs to be processed
- Learn the basics in command line
- Use validated tools

- But unless you want to be a bioinformatician
  - Use a validated and trusted pipeline
  - Get help from a bioinformatican (easier said than done!)
  - Hand over to a bioinformatician for this part