CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

# Moderní metody analýzy genomu - analýza

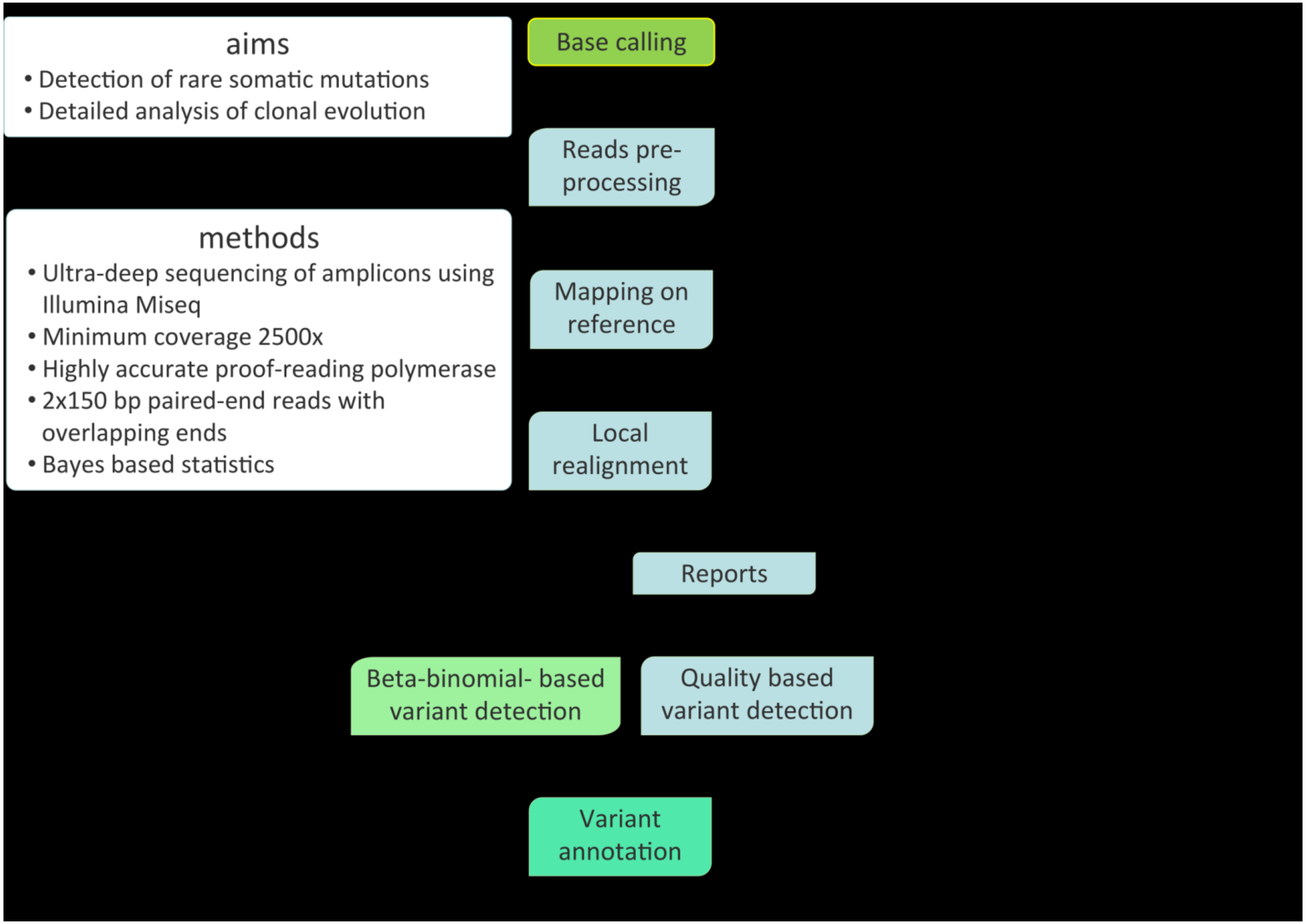
Mgr. Nikola Tom

Brno, 22.4.2015

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and
Development for Innovation

# Deep-seq Workflow/Pipeline



aims

- Detection of rare somatic mutations
- Detailed analysis of clonal evolution

methods

- Ultra-deep sequencing of amplicons using Illumina Miseq
- Minimum coverage 2500x
- Highly accurate proof-reading polymerase
- 2x150 bp paired-end reads with overlapping ends
- Bayes based statistics

Base calling

Reads pre-processing

Mapping on reference

Local realignment

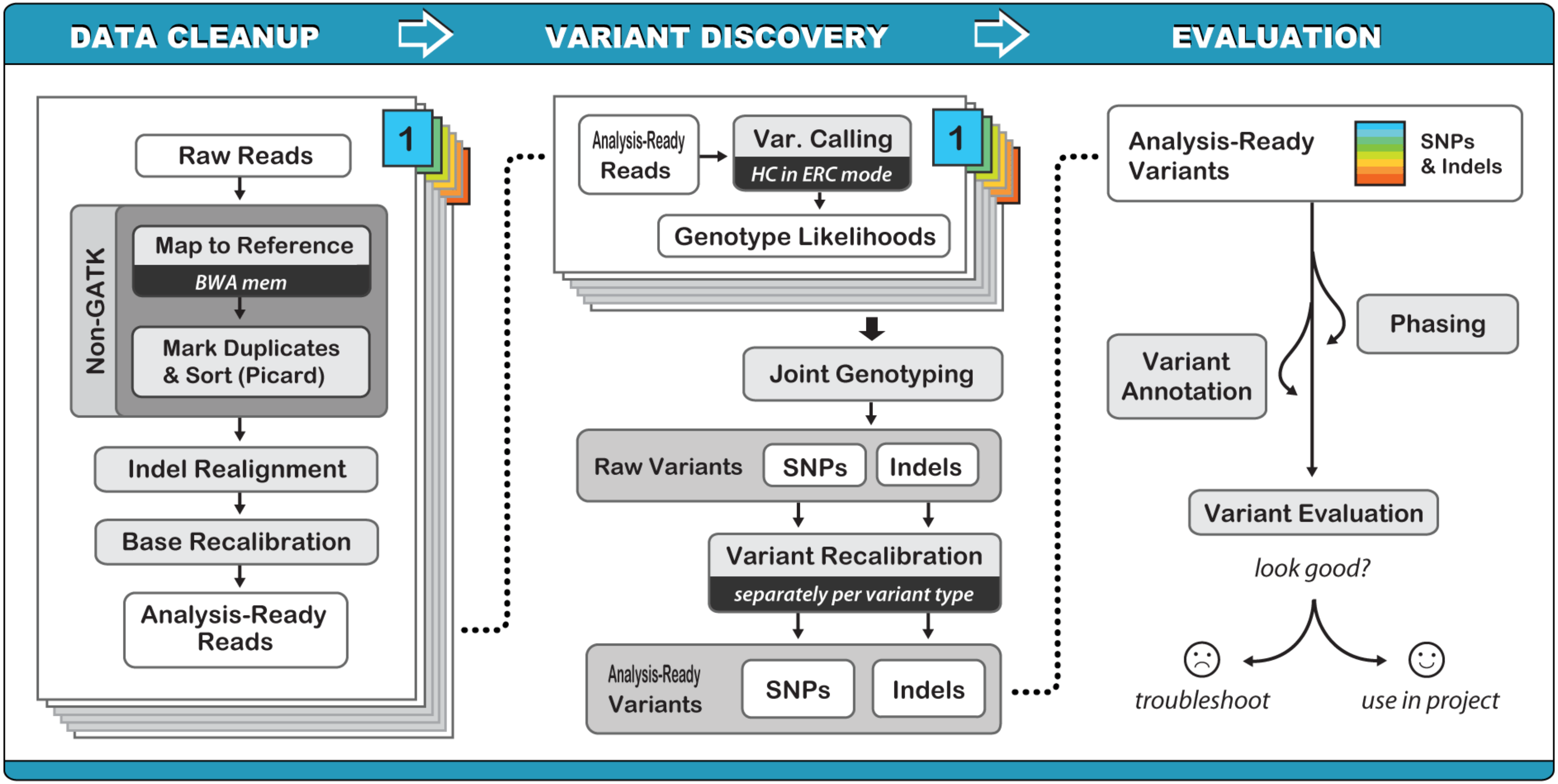
Reports

Beta-binomial- based variant detection

Quality based variant detection

Variant annotation

# GATK Workflow/Pipeline



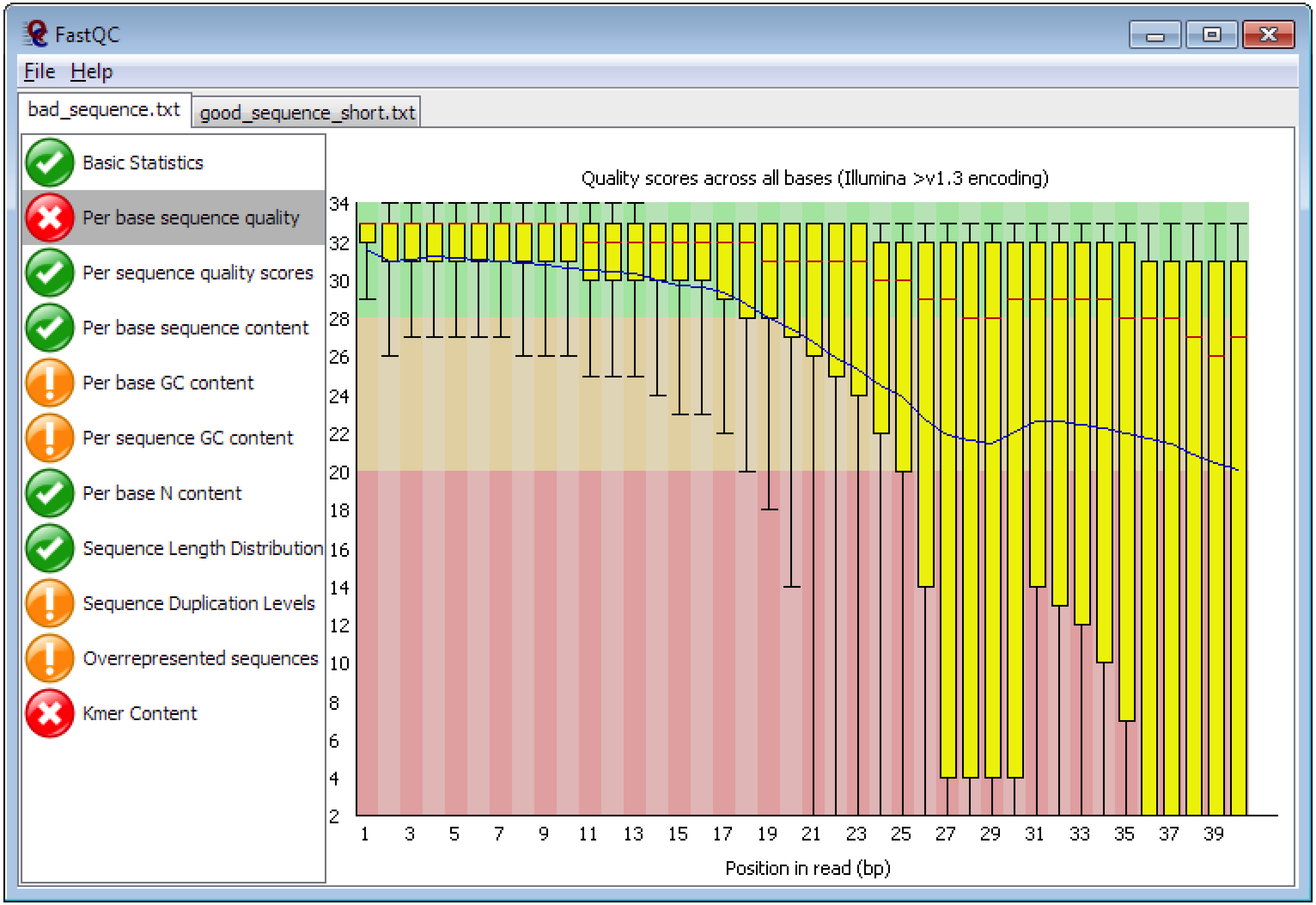
https://www.broadinstitute.org/gatk/guide/best-practices

# Raw sequence = fastq

- Biological sequence
- Corresponding quality scores
- ASCII character
- (fasta+ qual, csfasta + csqual, sff)

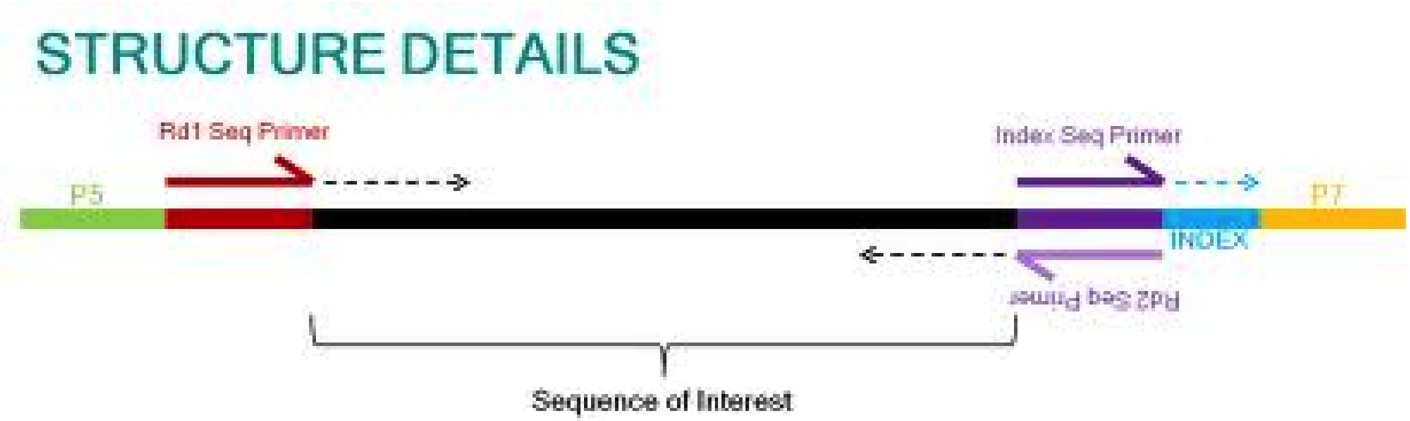
```
@
SEQ_ID GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

# FastQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

# Cutadapt

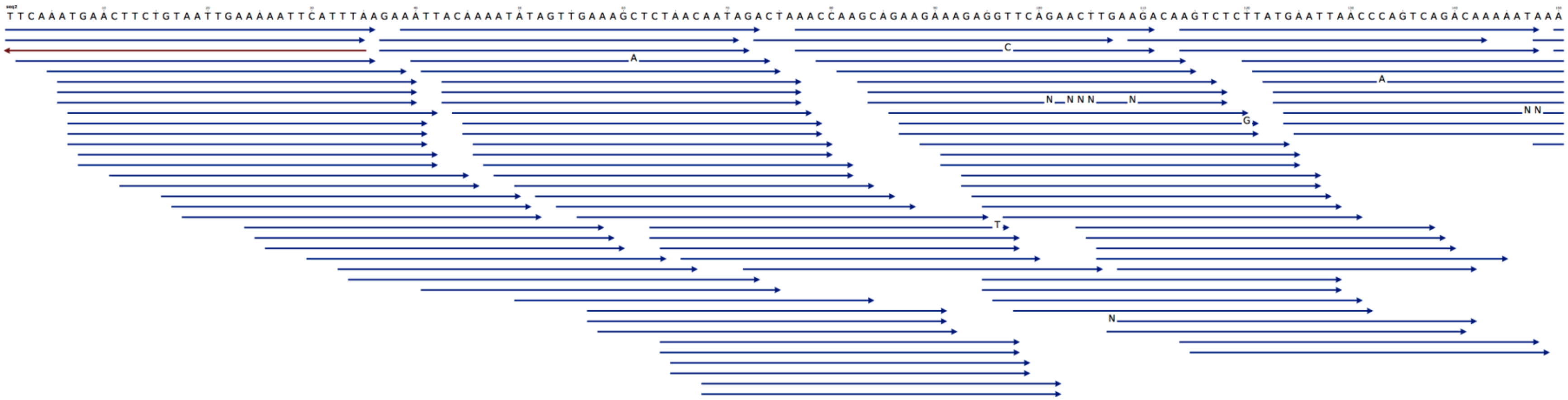
- Adaptor trimming (miRNA)
- Quality filtering
- Length filtering

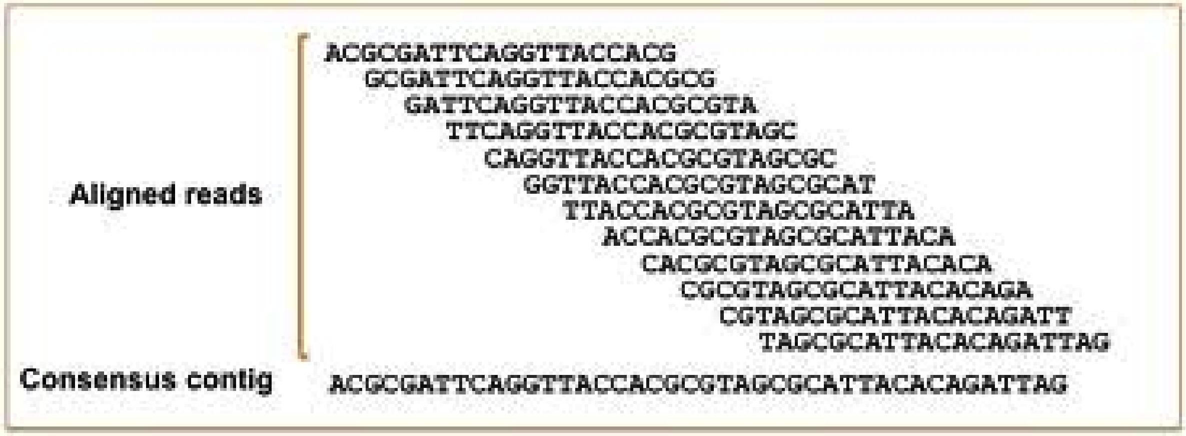


# Read mapping => SAM, BAM

- Usually mapping reads on reference
- miRNA - special case
  - Grouping and annotate against mirBase
- **DNA**
  - BWA, Bowtie, Bfast, SHRiMP, CLC
- **RNA**
  - TopHat (*de novo* splice aligner)
- **Commercial**
  - CLC Genomics Workbench
- ***De novo* assembly** – unknown genomes

# Alignment


seq2
TTCAAATGAACTTCTGTAATTGAAAAATTCATTTAAGAAATTACAAAATATAGTTGAAAGCTCTAACAATAGACTAAACCAAGCAGAAGAAAGAGGTTCAGAACTTGAAGACAAGTCTCTTATGAATTAACCCAGTCAGACAAAAATAAA
C
A
A
N N N N N
N N
G
G
T
N



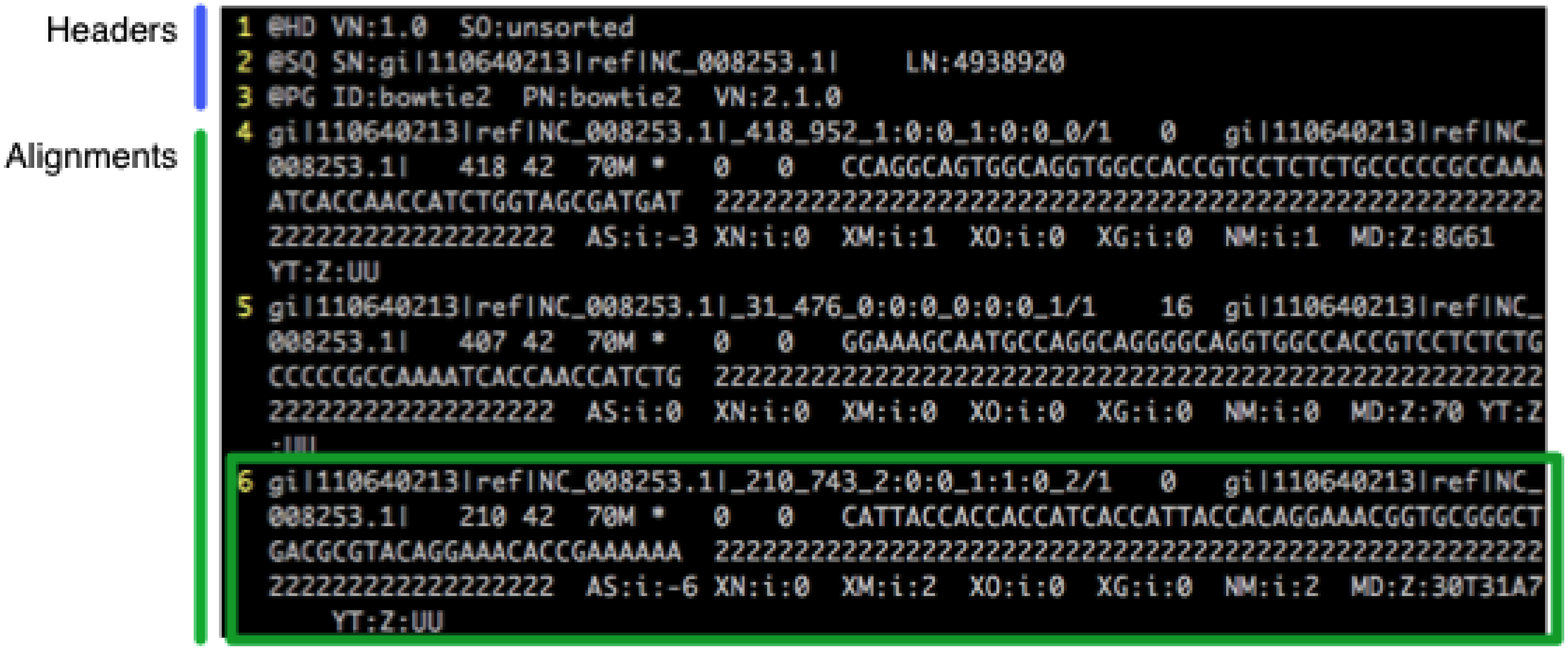
Aligned reads

```
ACGCGATTCAGGTTACCACG
  GCGATTCAGGTTACCACGCG
    GATTCAGGTTACCACGCGTA
      TTCAGGTTACCACGCGTAGC
        CAGGTTACCACGCGTAGCGC
          GGTTACCACGCGTAGCGCAT
            TTACCACGCGTAGCGCATTA
              ACCACGCGTAGCGCATTACA
                CACGCGTAGCGCATTACACA
                  CGCGTAGCGCATTACACAGA
                    CGTAGCGCATTACACAGATT
                      TAGCGCATTACACAGATTAG
```

Consensus contig

```
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

# SAM

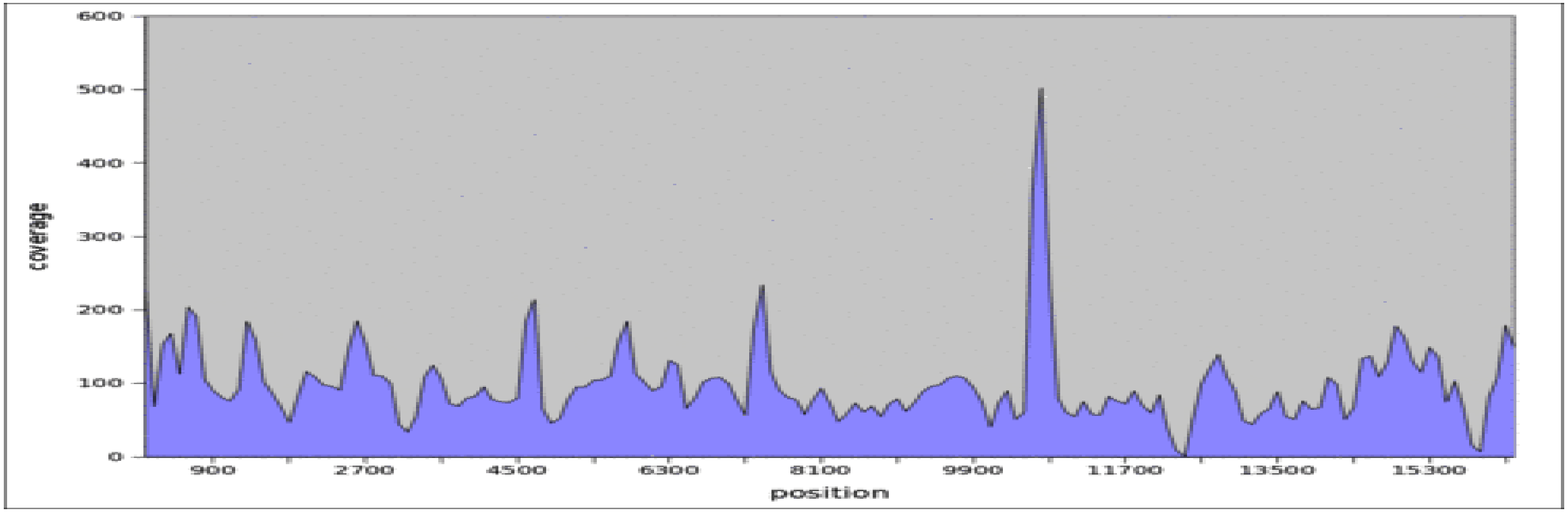


```
@HD VN:1.0  SO:unsorted
@SQ SN:gi|110640213|ref|NC_008253.1|    LN:4938920
@PG ID:bowtie2  PN:bowtie2  VN:2.1.0
gi|110640213|ref|NC_008253.1|_418_952_1:0:0_1:0:0_0/1   0   gi|110640213|ref|NC_
008253.1|   418 42  70M *   0   0   CCAGGCAGTGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAA
ATCACCAACCATCTGGTAGCGATGAT  22222222222222222222222222222222222222222222222222222
22222222222222222  AS:i:-3 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:8G61
YT:Z:UU
gi|110640213|ref|NC_008253.1|_31_476_0:0:0_0:0:0_1/1    16  gi|110640213|ref|NC_
008253.1|   407 42  70M *   0   0   GGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTG
CCCCCGCCAAAATCACCAACCATCTG  22222222222222222222222222222222222222222222222222222
22222222222222222  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:70 YT:Z
:UU
gi|110640213|ref|NC_008253.1|_210_743_2:0:0_1:1:0_2/1   0   gi|110640213|ref|NC_
008253.1|   210 42  70M *   0   0   CATTACCACCACCATCACCATTACCACAGGAAACGGTGCGGGCT
GACGCGTACAGGAAACACCGAAAAAA  22222222222222222222222222222222222222222222222222222
22222222222222222  AS:i:-6 XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:30T31A7
    YT:Z:UU
```

Each row describes a single alignment of a raw read against the reference genome.
Each alignment has 11 mandatory fields, followed by any number of optional fields.

# Mapping, Coverage reports

- Important checkout for lab protocol
- Specificity of PCR
- Normalization
- Settings of variant calling threshold, CNV



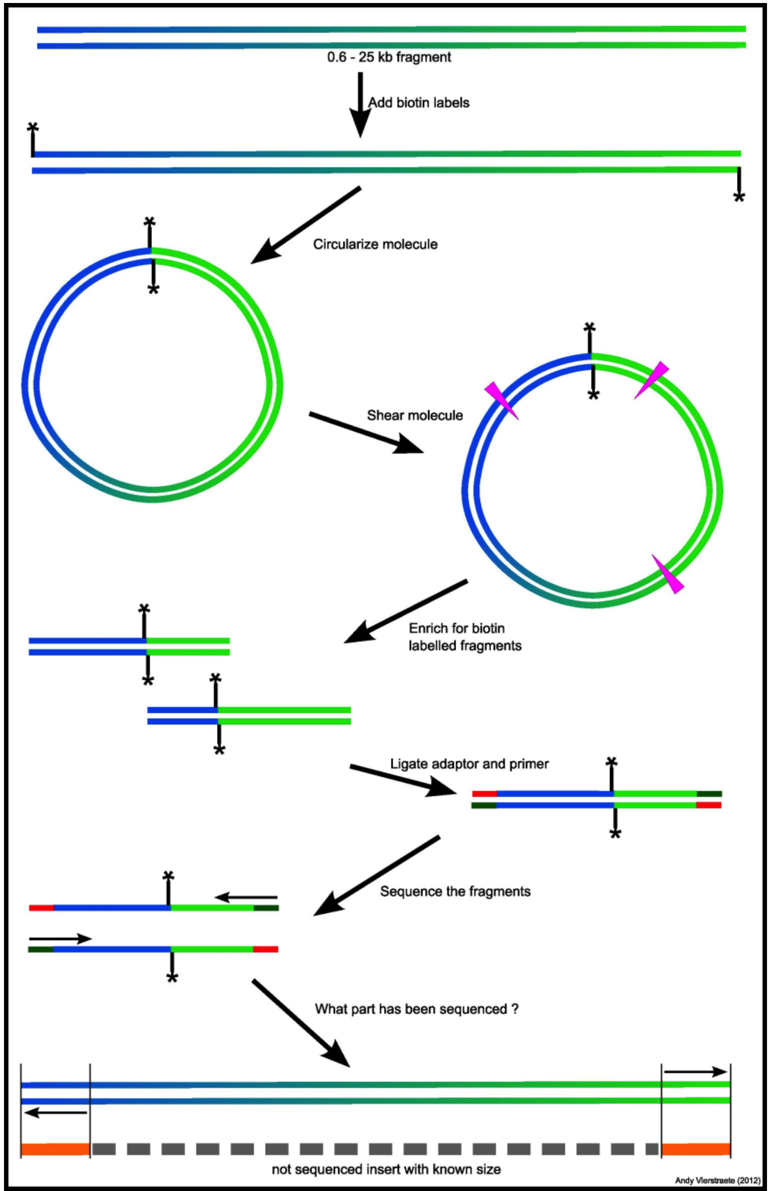
# SNV and small InDel Calling

- Coverage
- Frequency
- Base quality

- !!!
- Genomic context (homopolymers)
- Nucleotide type
- Position in read (errors at the read end)
- Alignment errors

# Structural variations

- Mate-pair library
- Long InDel
- Translocation


0.6 - 25 kb fragment
Add biotin labels
Circularize molecule
Shear molecule
Enrich for biotin labelled fragments
Ligate adaptor and primer
Sequence the fragments
What part has been sequenced ?
not sequenced insert with known size

Andy Vierstraete (2012)

# Annotating and filtering

- Gene
- Transcript
- dbSNP
- Regulation
- Comparative genomics
- Repeats
- Functional
- Gene ontology
- miRNA targets
- Etc.