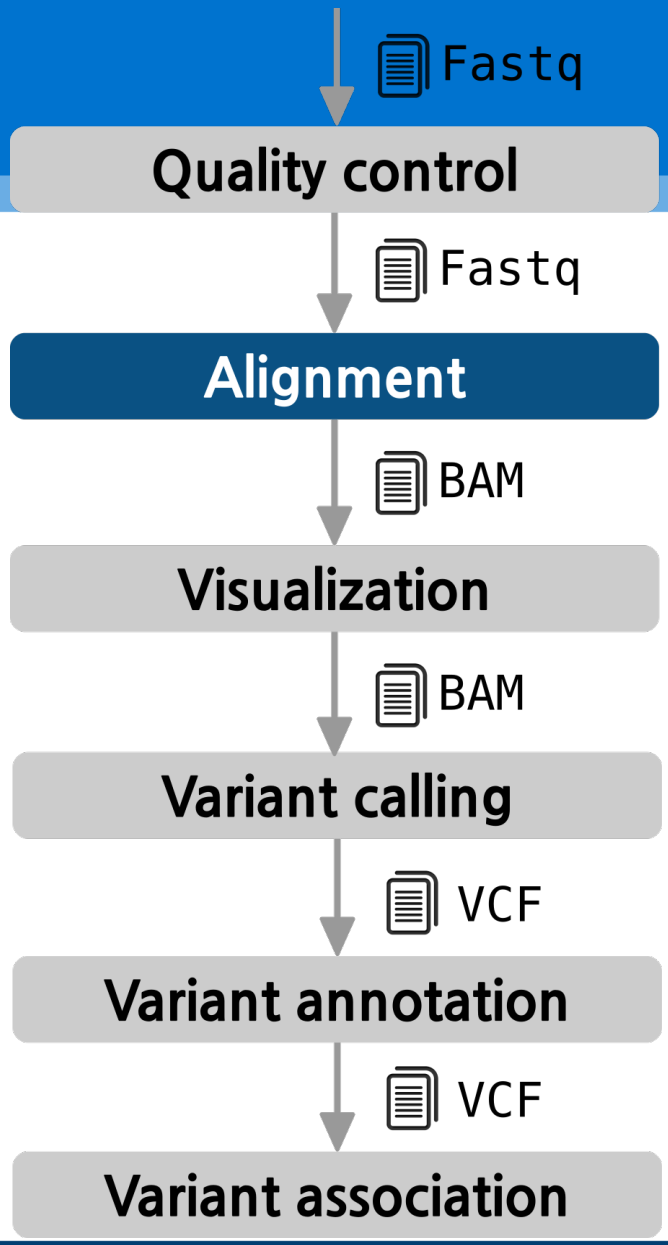
UNIVERSITY OF CAMBRIDGE

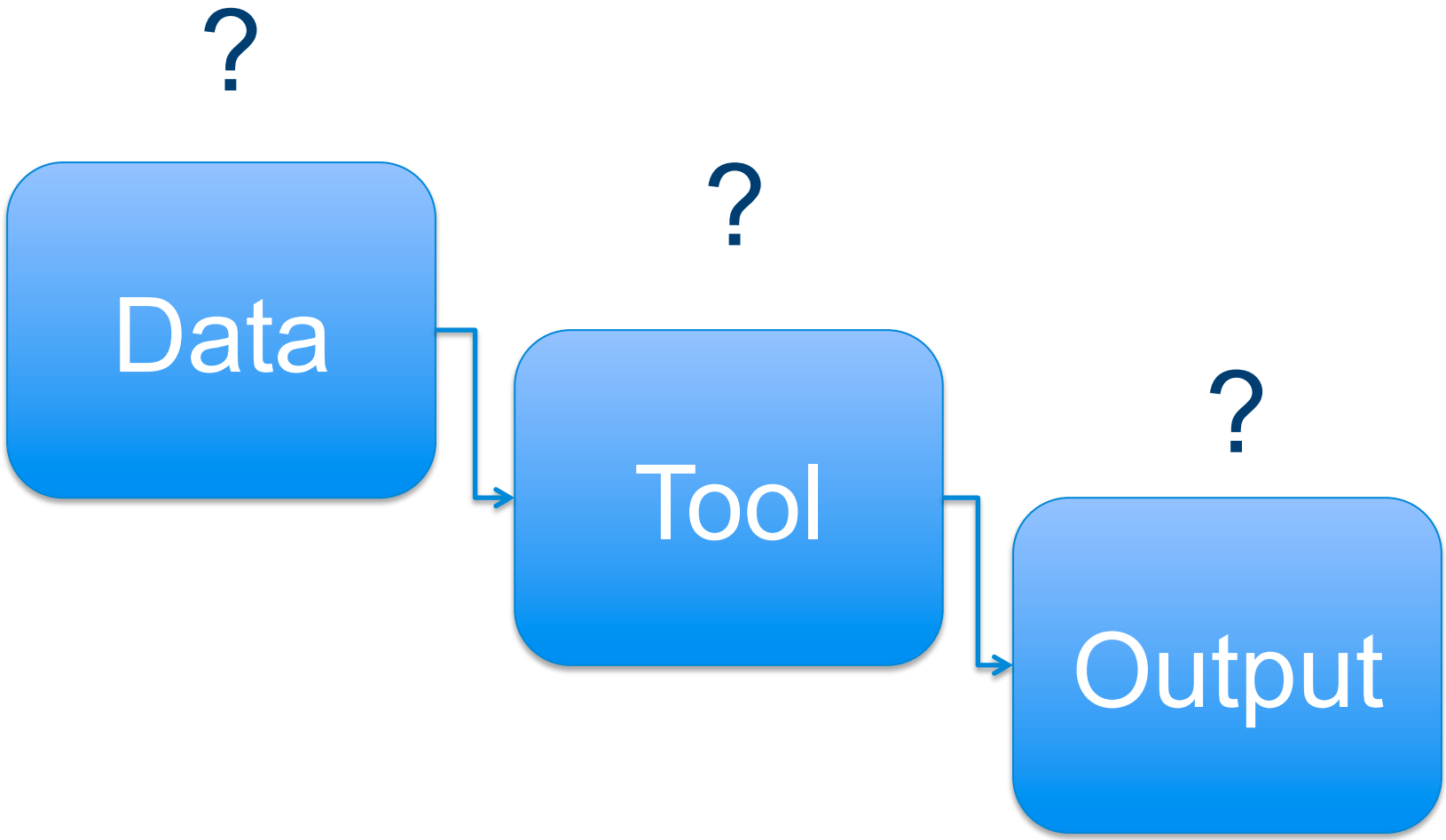
Short Read Alignment (NGS)

Matthias Haimel

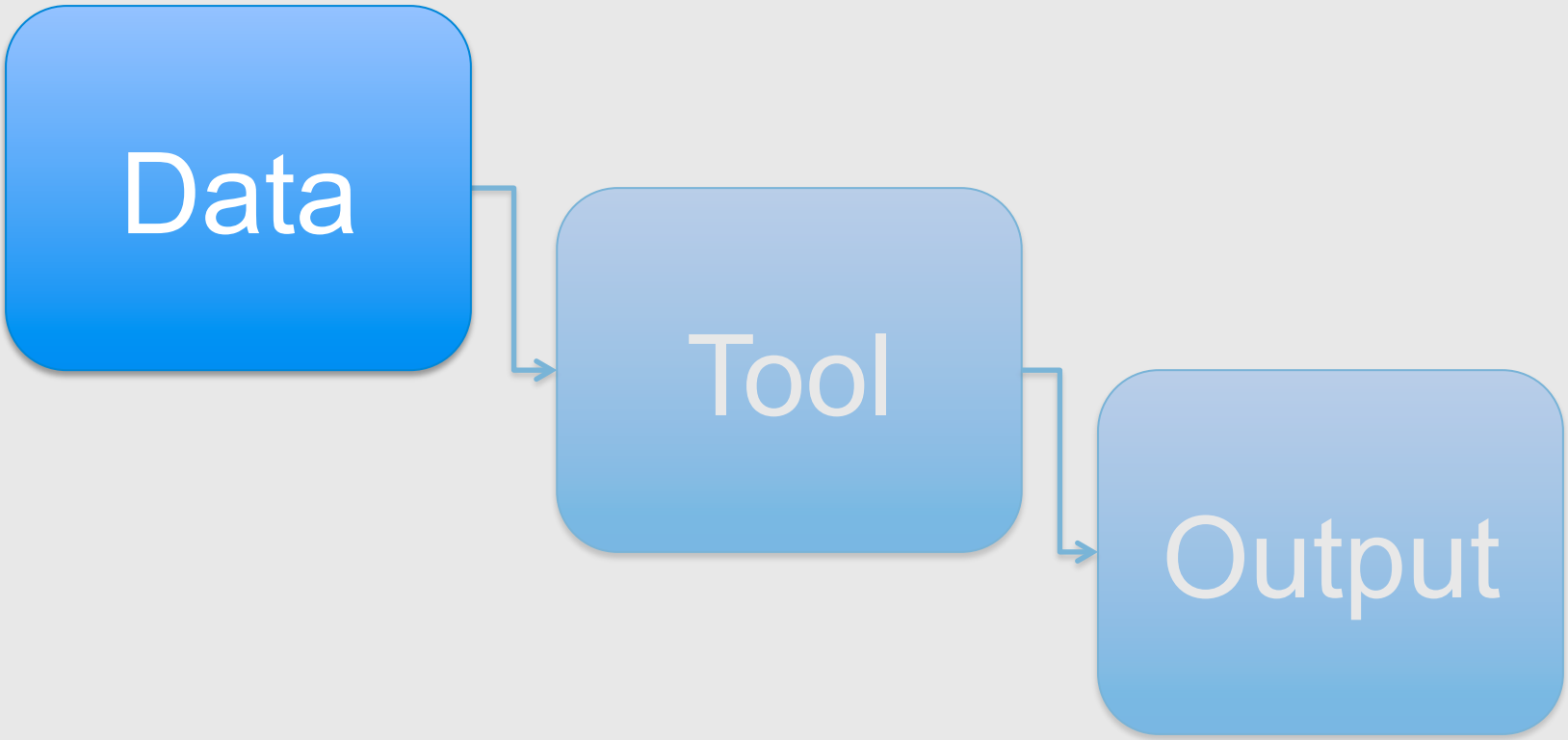
Respiratory Medicine


Fastq
Quality control
Fastq
Alignment
BAM
Visualization
BAM
Variant calling
VCF
Variant annotation
VCF
Variant association

Short Read Alignment (NGS)



Short Read Alignment (NGS)



Short Read Alignment (NGS)

Data

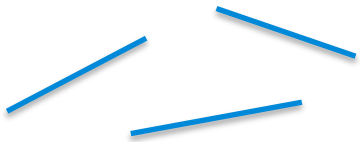
- Reads (FASTQ)

Short Read Alignment (NGS)

Data

- Reads (FASTQ)

Reads

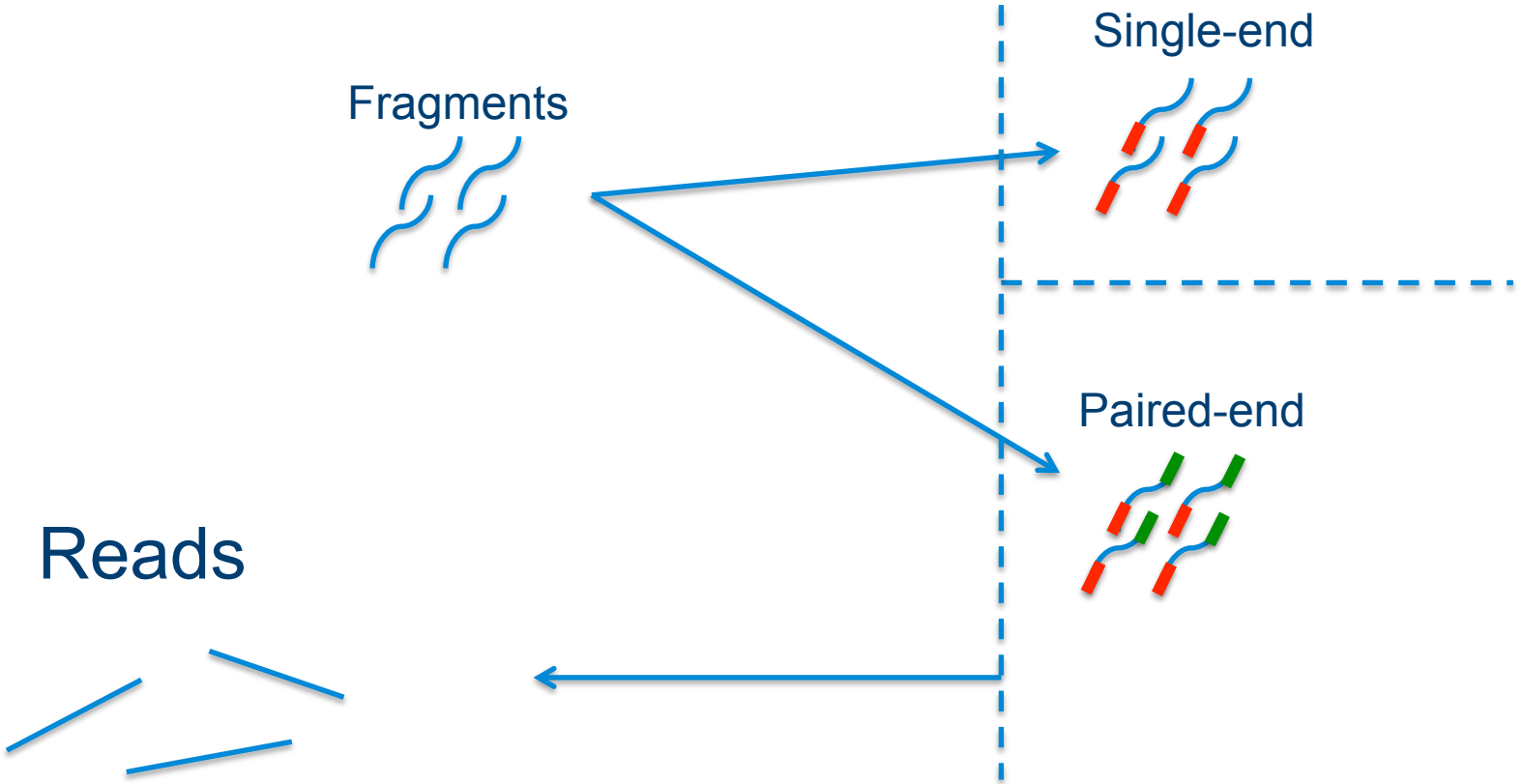


```
@ERR001268.10000142/1
CTATCCTAATCCCCAGAACACATGAAGATGTGACCT
+
IIIIIICIIIIDIB@5=30;?0?F/37257029.*F
@ERR001268.10000296/1
TCGGAGAATTCATACTGGAGAGAAACCTTACAAATG
+
8F@BII.IICII:4<?EB44670:9)1712.5*000
@ERR001268.10000504/1
CCTTGCTTTCAGTTGTCCCACCTTTCCAGACCAAAC
+
=IIIIIIIII:IIIIF?73809?946*/+03,+/6
```

Short Read Alignment (NGS)

Data

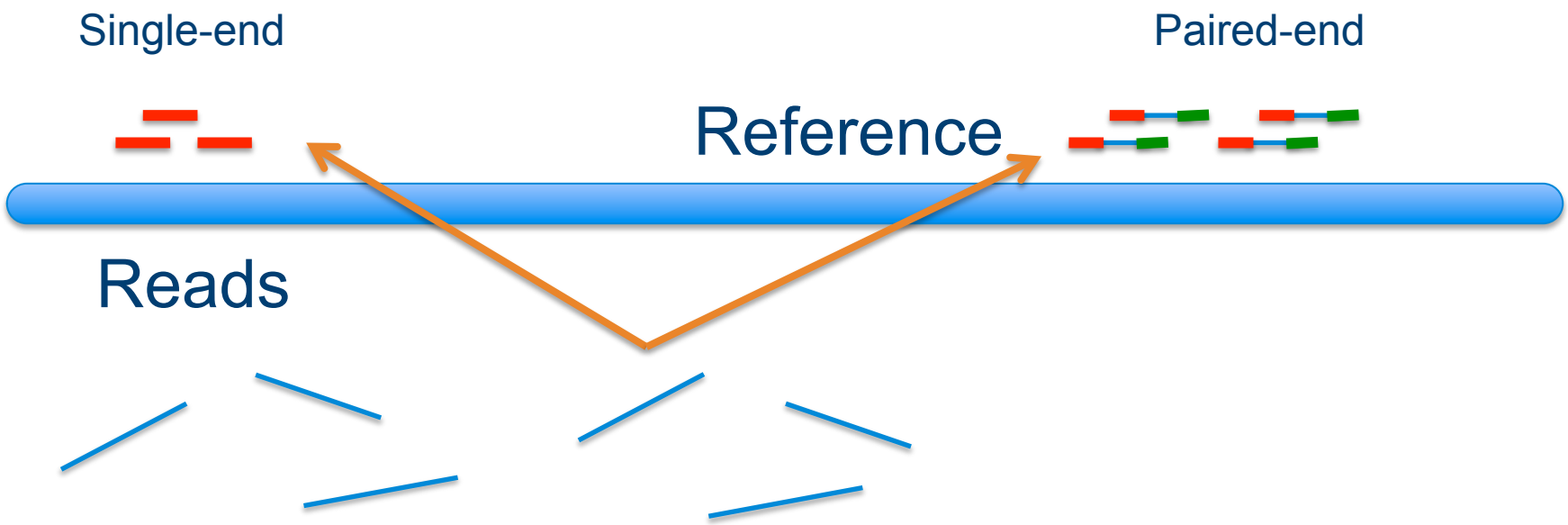
- Reads (FASTQ)



Short Read Alignment (NGS)

Data

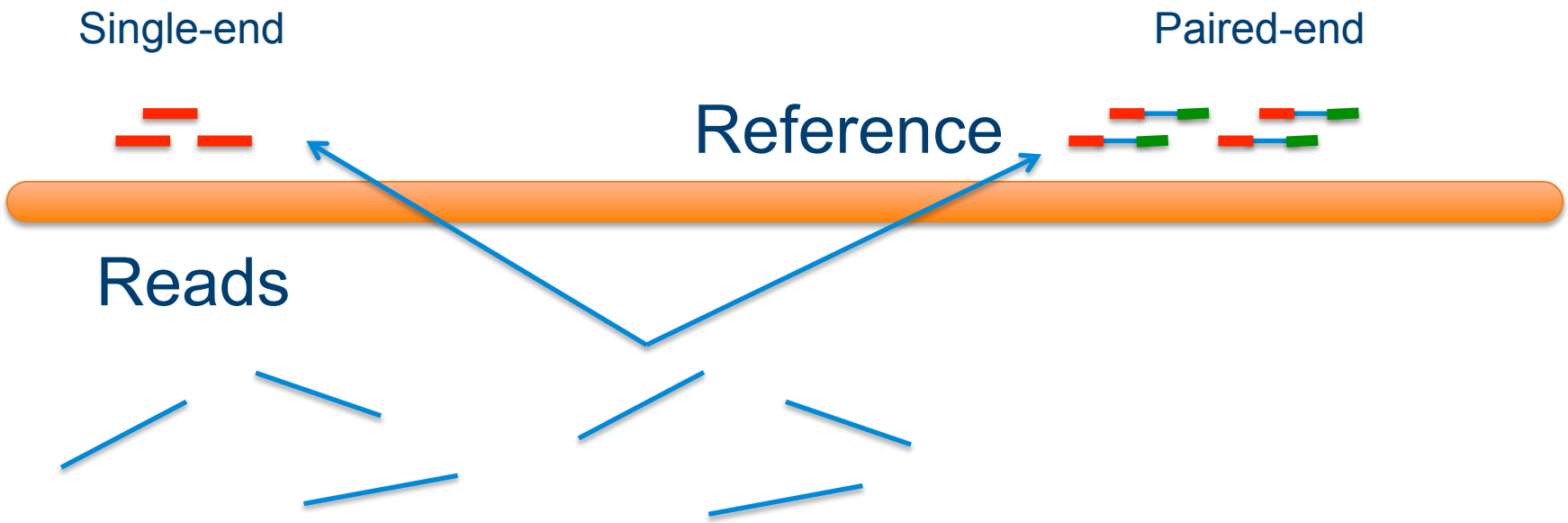
- Reference



Short Read Alignment (NGS)

Data

- Human Reference
  - GRCh37 <-> GRCh38

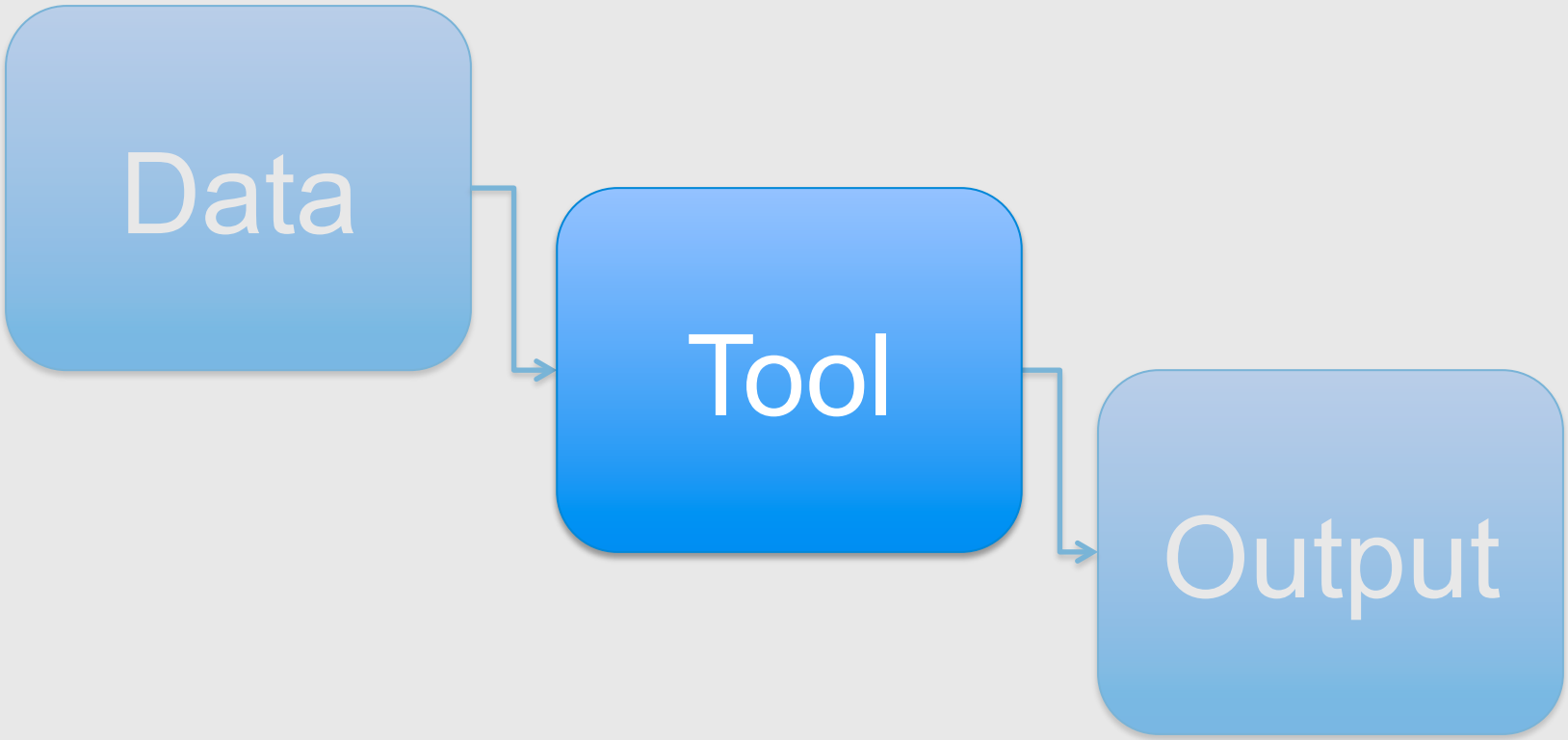


Reference Genome

Data

- **GRCh38** (submitted December 24, 2013 http://www.ncbi.nlm.nih.gov/news/12-23-2013-grch38-released/)
  - New coordinate system
  - Alternate sequence representation
  - Centromeres and heterochromatin models added
  - > 8000 sequencing errors corrected
- **Primary assembly**
  - Chromosomes + unlocalised & unplaced contigs + MT
- **Full assembly**
  - Primary assembly + alternate loci
- **Decoy contigs**
  - 2,385 contigs identified by Heng Li

Short Read Alignment (NGS)



Short Read Alignment (NGS)

Tools

- Global alignment
  - Attempt to align every base
  - Good for equal size sequence

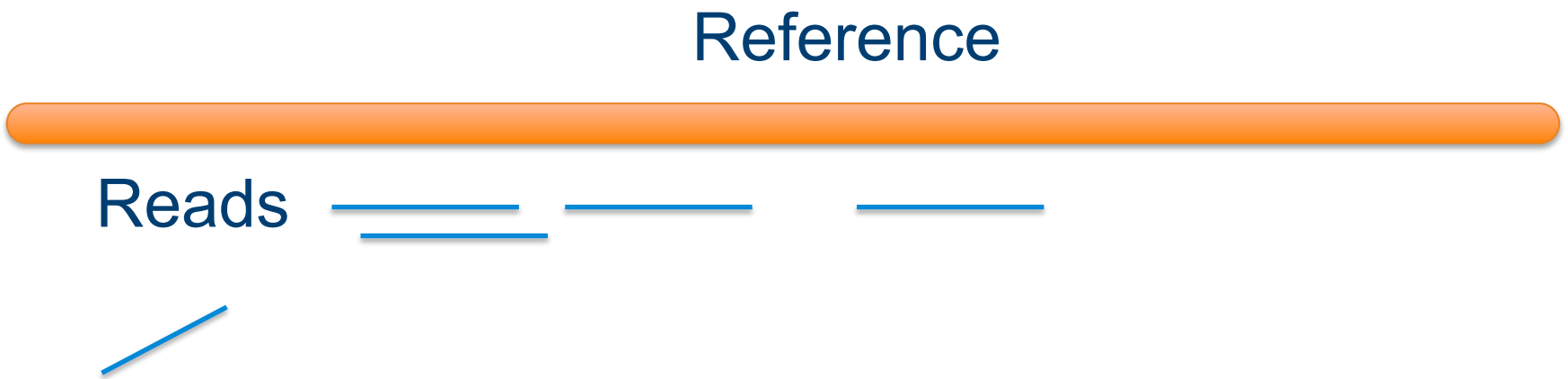
```
Global    CT----AT-TTACT----AT
          CTGGCTATGTTACTATGCAT
 Local    ----CTAT-TTACTAT----
```

- Local alignment
  - Contain a region of target sequence
  - Good for unequal sequence length

Short Read Alignment (NGS)

Tools

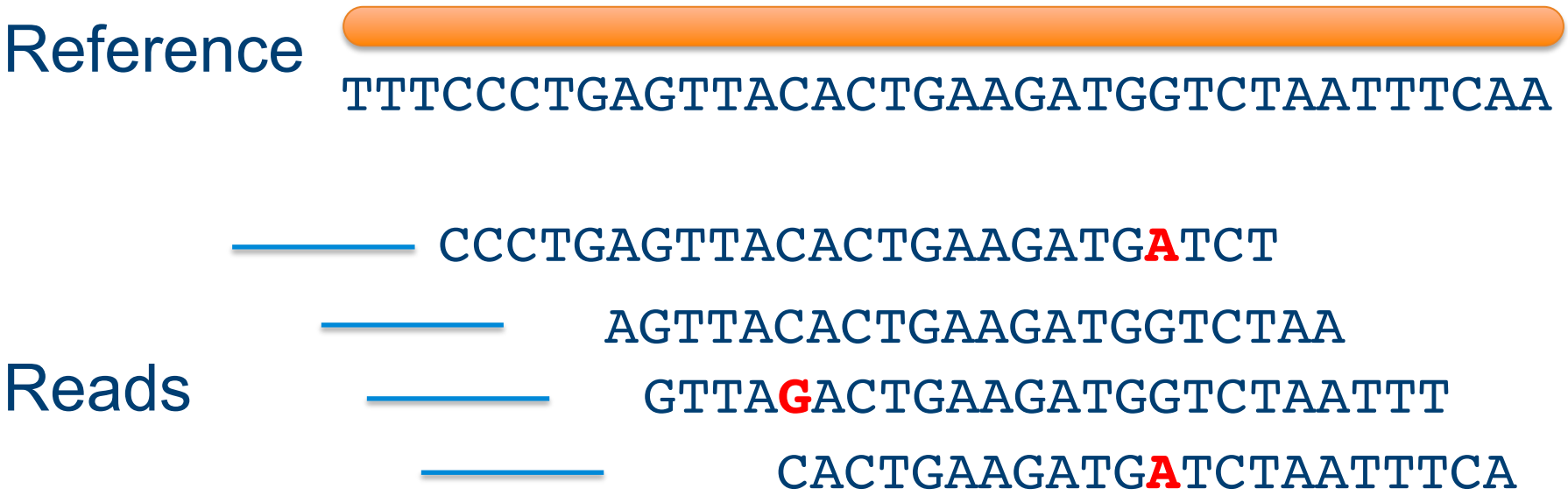
- Local alignment
- Fast
- Sensitive
- Specific



Short Read Alignment (NGS)

Tools

- Align each read to the reference sequence
- Reads are not perfect!!!



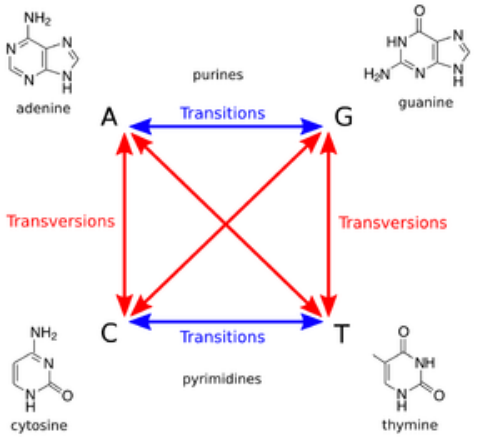
WGS example calculation

- Human genome:
  - $3.2 * 10^9$
- Phred Quality Score:
  - 40 (1 in 10,000 probably incorrect)
- Probably incorrect bases:
  - 3.2*10^5 = 320,000
- 40x sequence coverage (average):
  - 12,800,000 probably incorrect bases

Short Read Alignment (NGS)

Tools

- Reference missing?
  - *de novo* assembly
- Mismatches
  - Species polymorphism
  - Sequencing error



Ti/Tv ratio
- random: 0.5
- whole genome: ~2
- whole exome: ~3

Yan Guo et al (2012).
Exome sequencing generates high quality data in non-target regions. *BMC Genomics*, *13*, 194.

Reference

TTTCCCTGAGTTACACTGAAGATGGTCTAATTTCAA

Reads

CCCTGAGTTACACTGAAGATG**A**TCT

AGTTACACTGAAGATGGTCTAA

GTTA**G**ACTGAAGATGGTCTAATTT

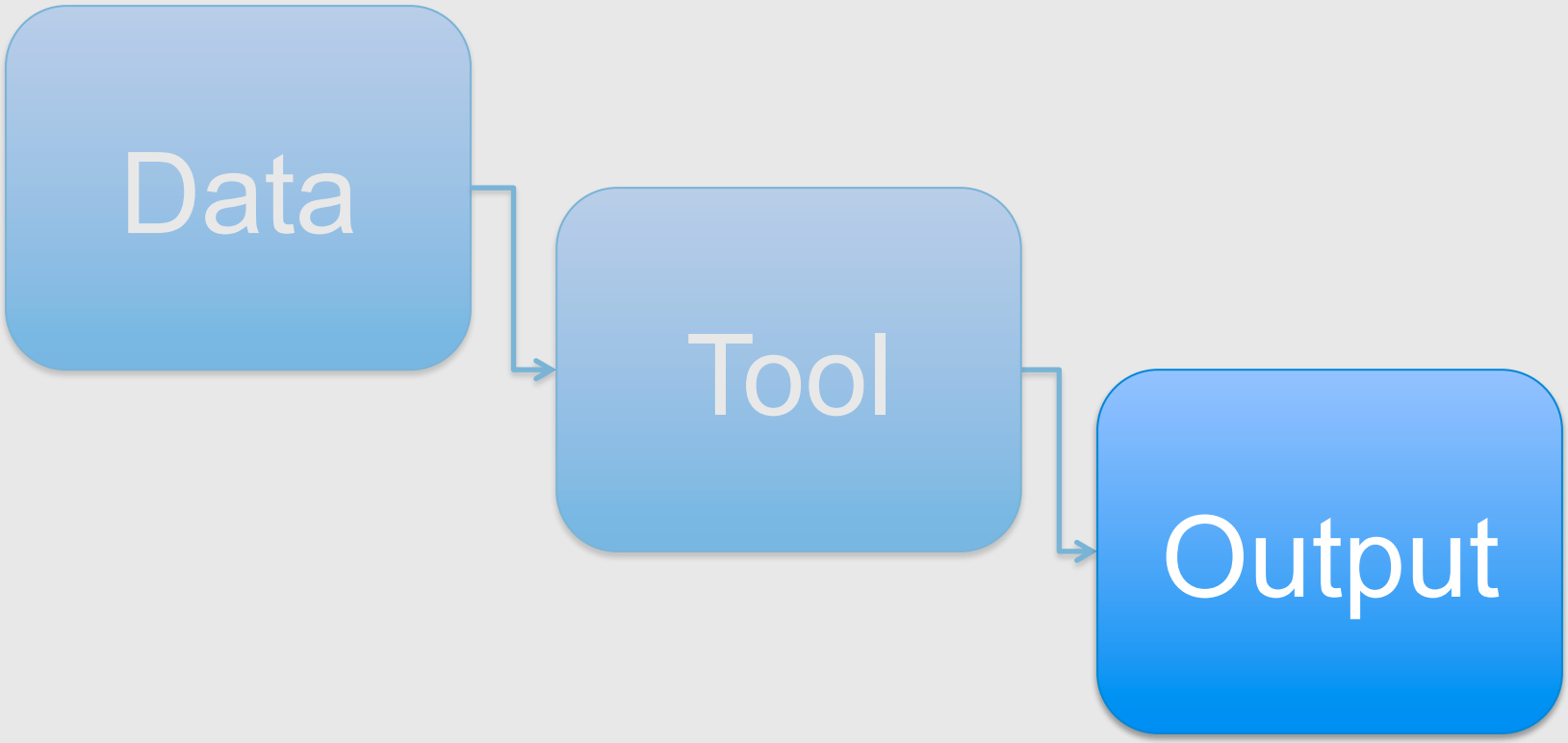
TG**A**TCTAATTTCAATG

Short Read Alignment (NGS)

Tools

- Reference missing?
  - *de novo* assembly
- Mismatches
  - Species polymorphism
  - Sequencing error
- Type of data
  - Whole Genome / Exome
  - RNA-Seq
  - ChIP-Seq
  - …

Short Read Alignment (NGS)



Short Read Alignment (NGS)

Output

- **SAM** (**S**equence **A**lignment / **M**ap format)
- **BAM** (**B**inary version of SAM)
- Well defined specification
  - http://samtools.sourceforge.net/SAMv1.pdf
- Large tool collection for reading / writing
  - samtools: basic utility tool
  - Picard: more comprehensive utility tool

Sequence Alignment / Map format

Output

SAM format

- Header section
- Alignment section

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   163 ref   7 30 8M2I4M1D3M = 37   39 TTAGATAAAGGATACTG *
r002     0 ref   9 30 3S6M1P1I4M *  0    0 AAAAGATAAGGATA     *
r003     0 ref   9 30 5S6M       *  0    0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M     *  0    0 ATAGCTTCAGC        *
r003  2064 ref 29 17 6H5M       *  0    0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001    83 ref 37 30 9M         =  7  -39 CAGCGGCAT          * NM:i:1
```

Sequence Alignment / Map format

Output

- Header stores extra information about the alignment
  - E.g. Reference, Program, Read groups

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
```

- Start with '@' followed by a two-letter code

Sequence Alignment / Map format

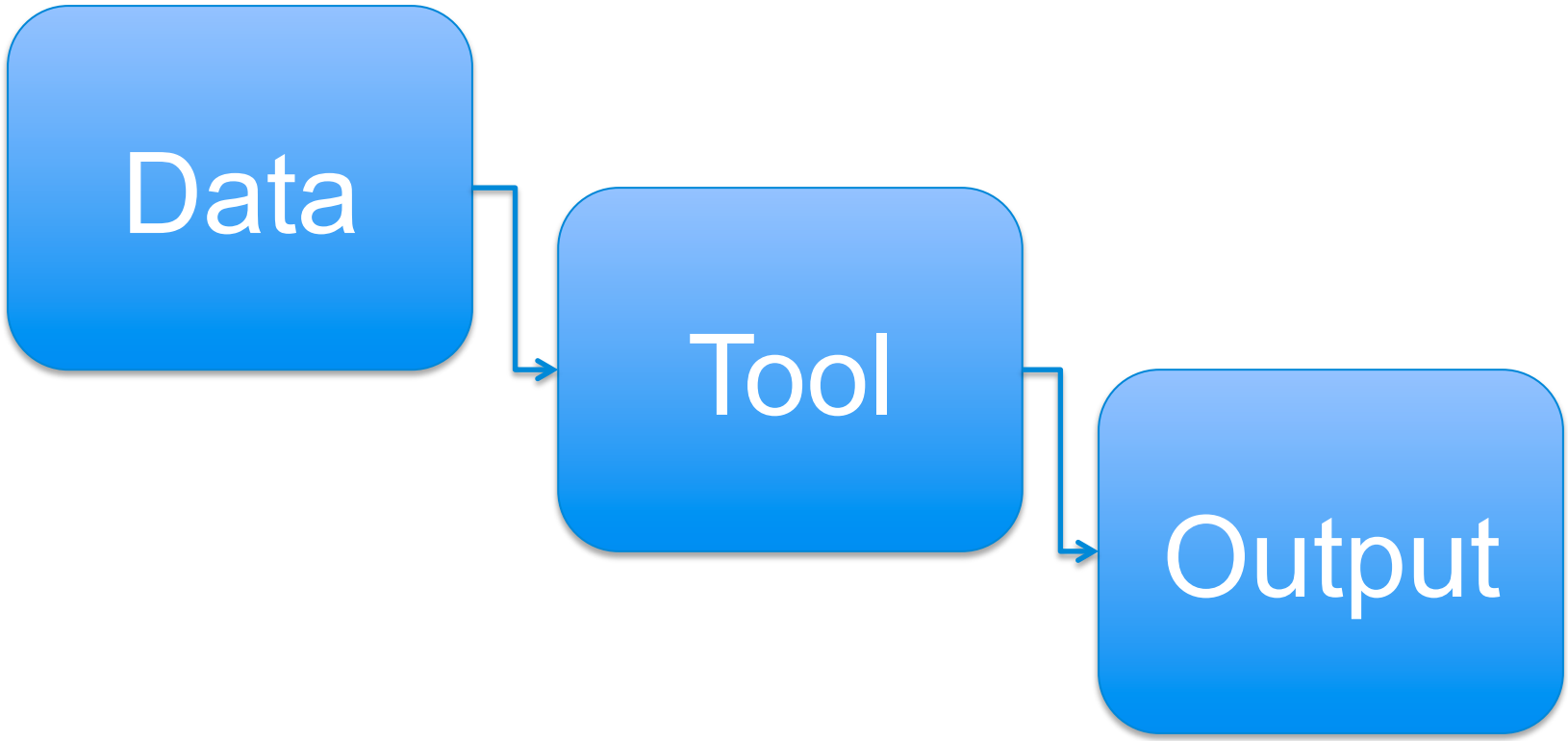
Output

- Alignment

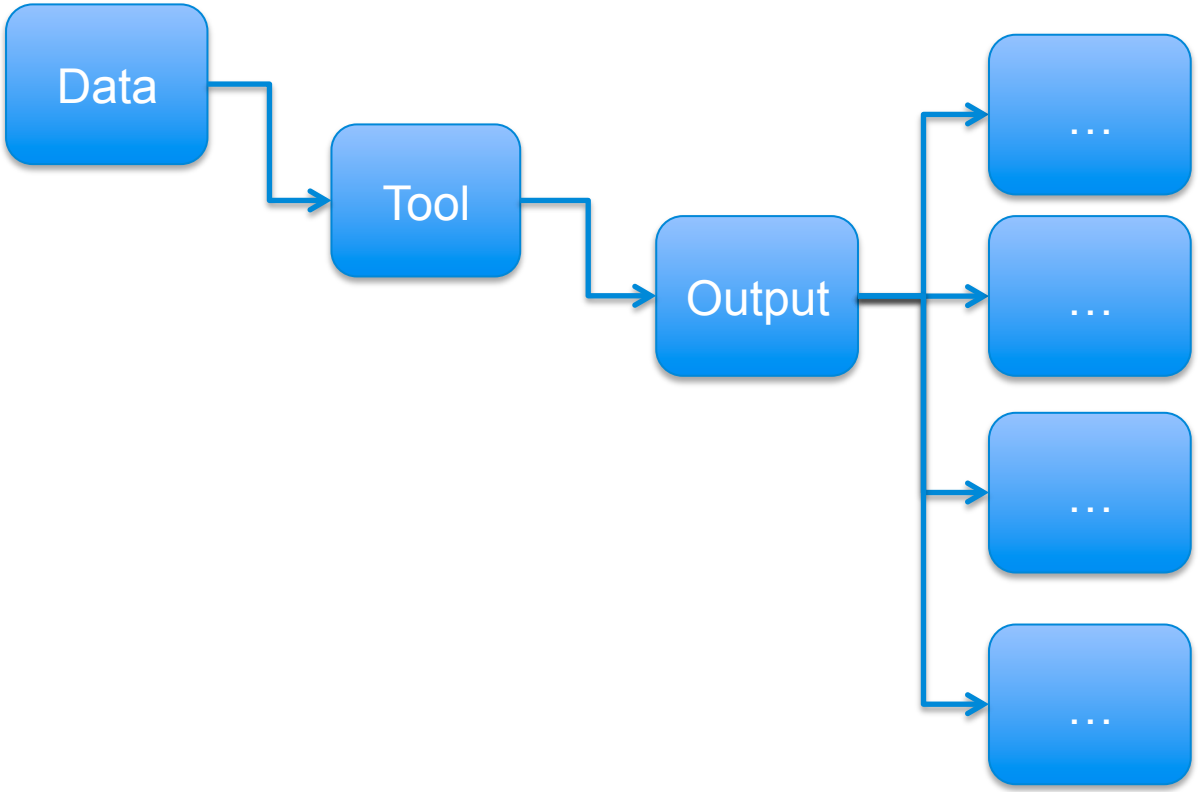
```
r001   83 ref 37 30 9M          =  7 -39 CAGCGGCAT          * NM:i:1
```

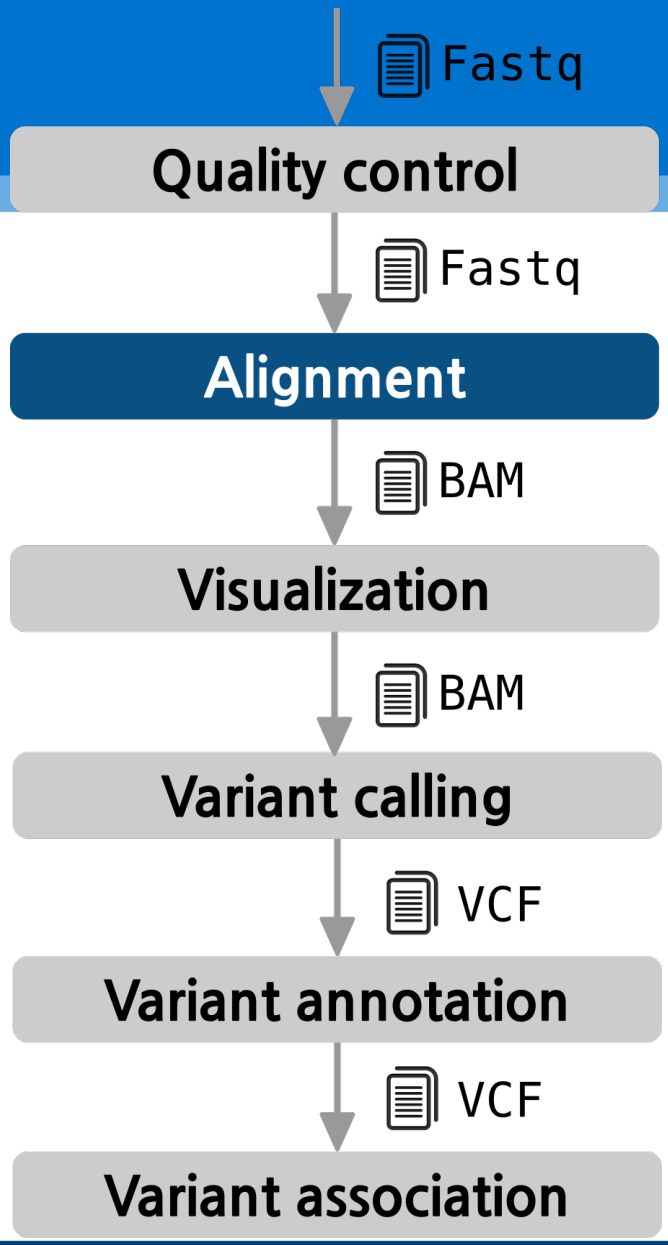
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Short Read Alignment (NGS)



Short Read Alignment (NGS)




Fastq
Quality control
Fastq
Alignment
BAM
Visualization
BAM
Variant calling
VCF
Variant annotation
VCF
Variant association

UNIVERSITY OF CAMBRIDGE

Hands-on