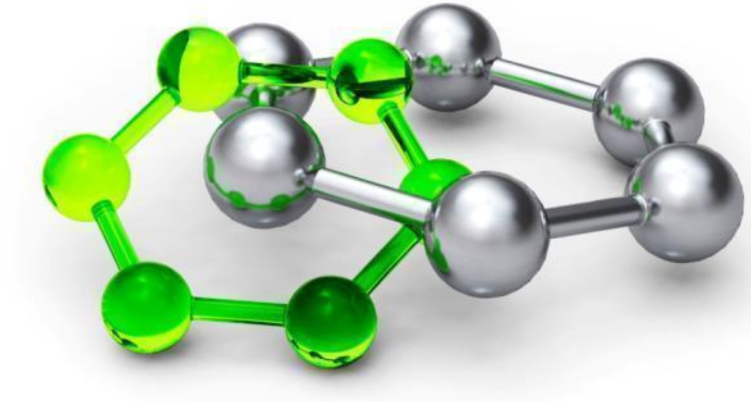
CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

Moderní metody analýzy genomu Bioinformatika I

Mgr. Nikola Tom



Brno,
11.11.20
16

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and
Development for Innovation

Bioinformatics

Bioinformatics is a quite new field… (first NGS in 2005)
How to analyse data defived from NGS = bottleneck of NGS

AIM: clean the data and give them biological sense

Bioinformatics **SOLUTION 1**:

- commercial software and ready to use pipelines

BUT they have usually not-transparent settings and/or not enough of options
(good programs expensive)

Bioinformatics

Bioinformatics **SOLUTION 2**:

- command-line based tools/software

Each tools solves only a part of the analysis

- Need for setup the pipeline & tune programs’ parameters

(challenging & more precise!!!)

ubuntu

gatk

Bioinformatics

Choice of programs & settings heavily depends on type of experiment, library preparation, biological question

Laptop or PC are usually not enough… need for cluster

×

Before we start analysis

We have to know what we are dealing with… and what we want to find out…

Concept of the project
DNA/RNA/methylation/…

DNA

- Targeted sequencing (amplicons, gene panels, exomes)
- Whole genome sequencing
  - Finding differences to known reference genome = re-sequencing

De novo assembly
- Genome construction

Before we start analysis

RNA
- Gene expression, ncRNA, alternative splicing

Metagenomics (bacteria, viruses)
- Composition of organisms in the sample, genetic variants

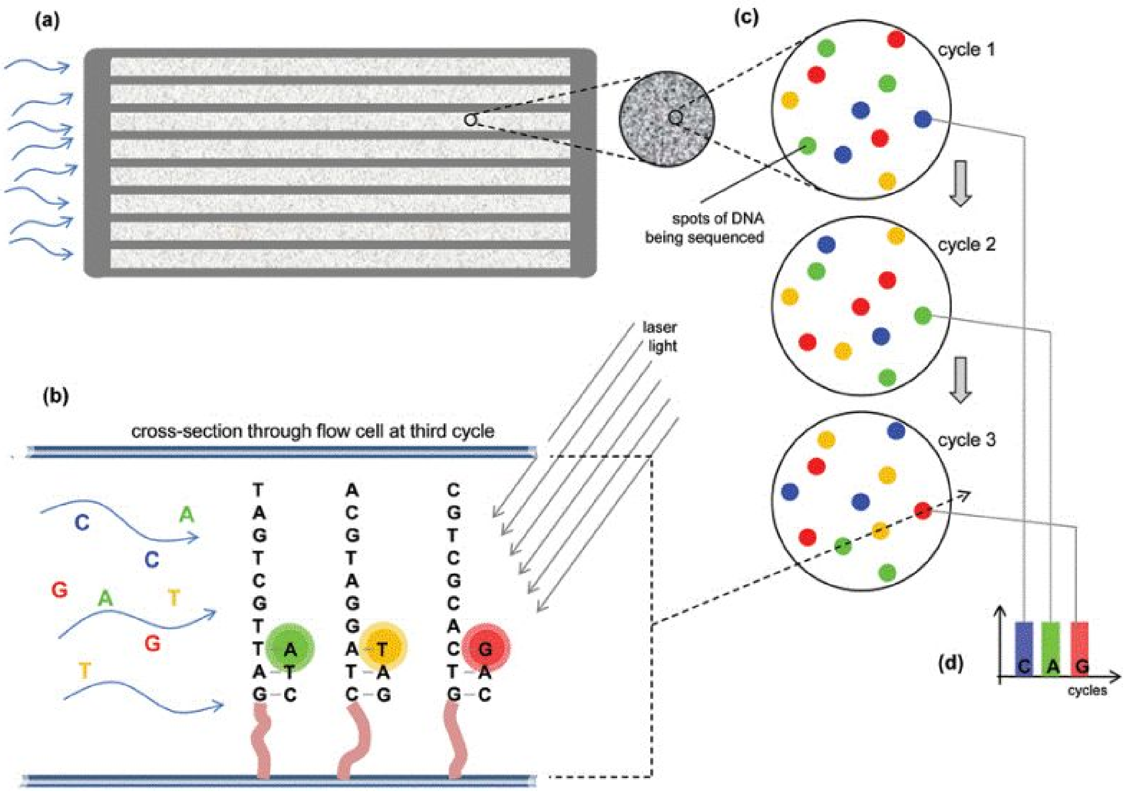
ChIP sequecing (DNA-protein interactions)

Bioinformatics’ starting point

Raw sequencing data - READ

Produced during **base calling**
- signal to sequence conversion and assigning base quality scores

(**fastq** file)



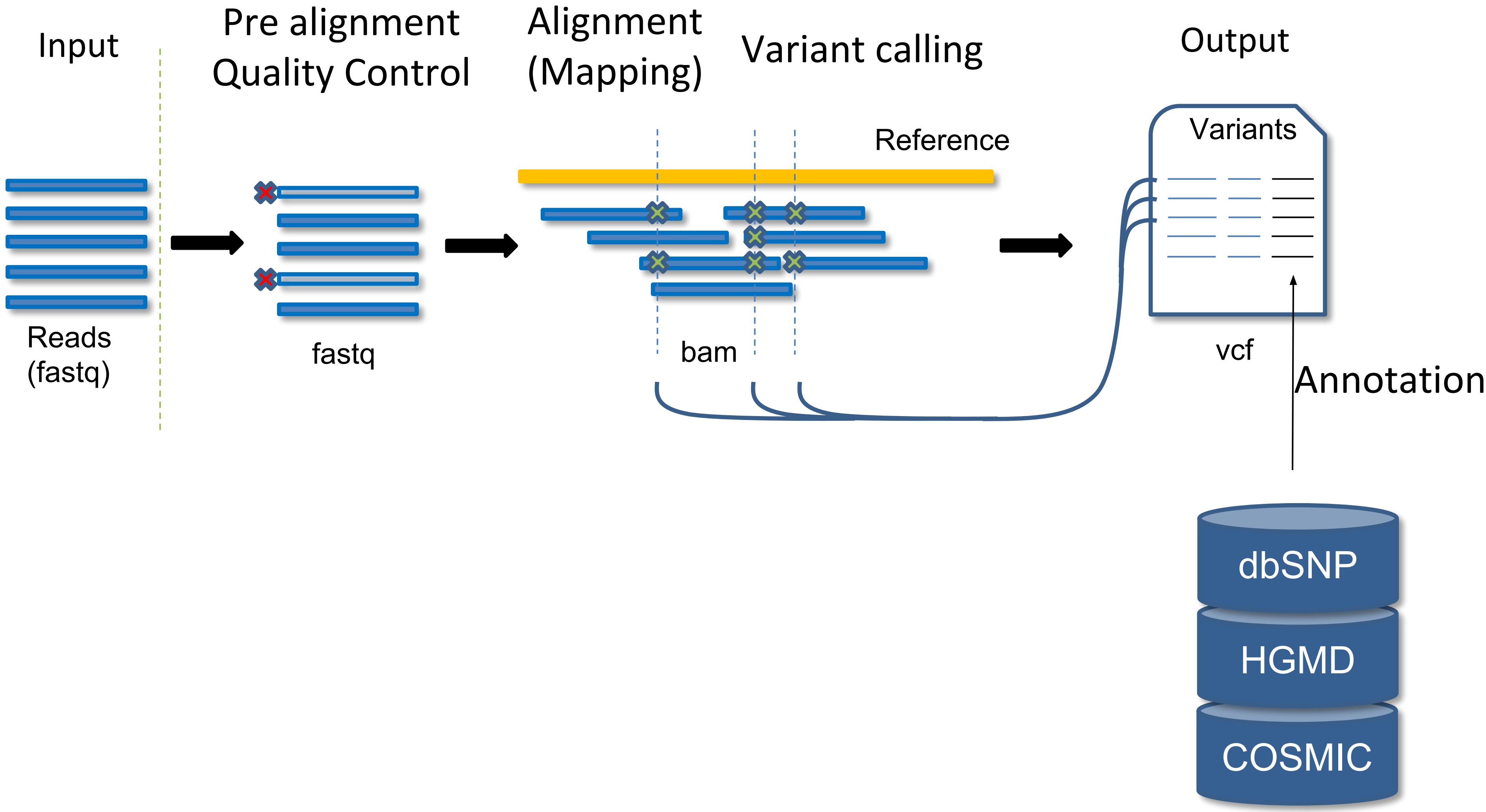
Fastq file

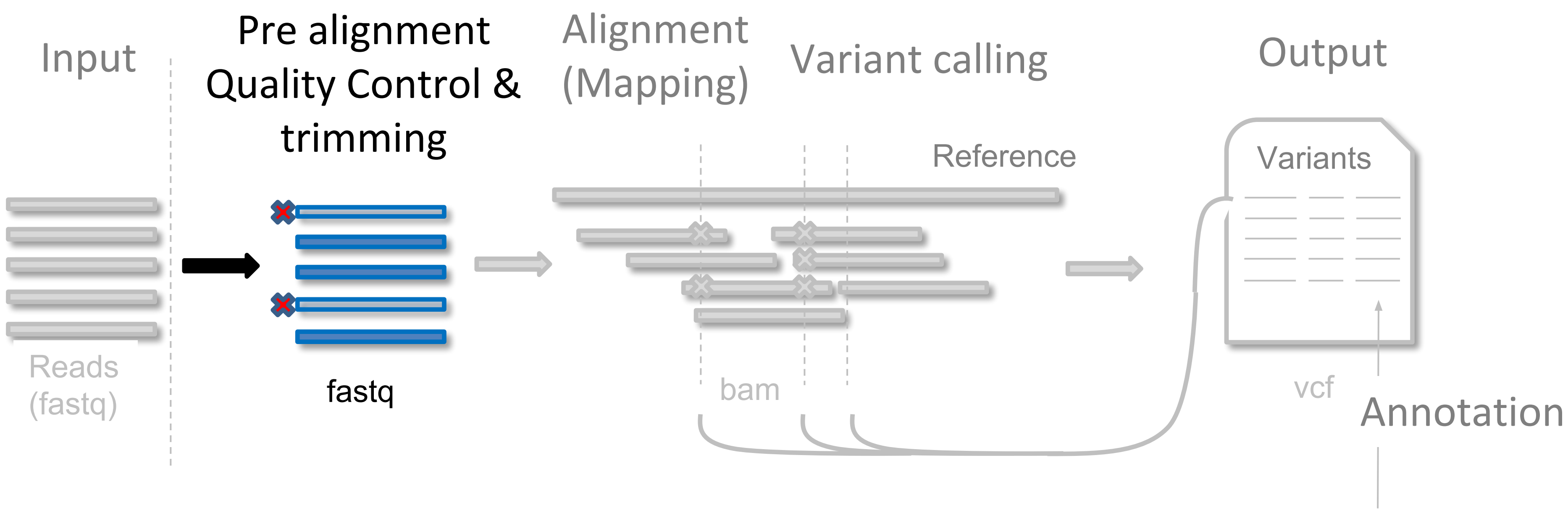
- Consists of reads - biological sequences
  (each read represents 1 input molecule sequenced on flowcell)
- Corresponding quality score for each base
- **Phred score** – probability of arising an error (log based)
- ASCII character
- (fasta+ qual, csfasta + csqual, sff)
- Pair-end sequencing – 2 fastq files

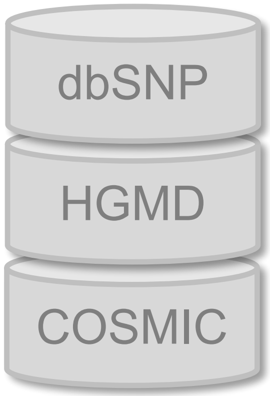
example.fastq

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

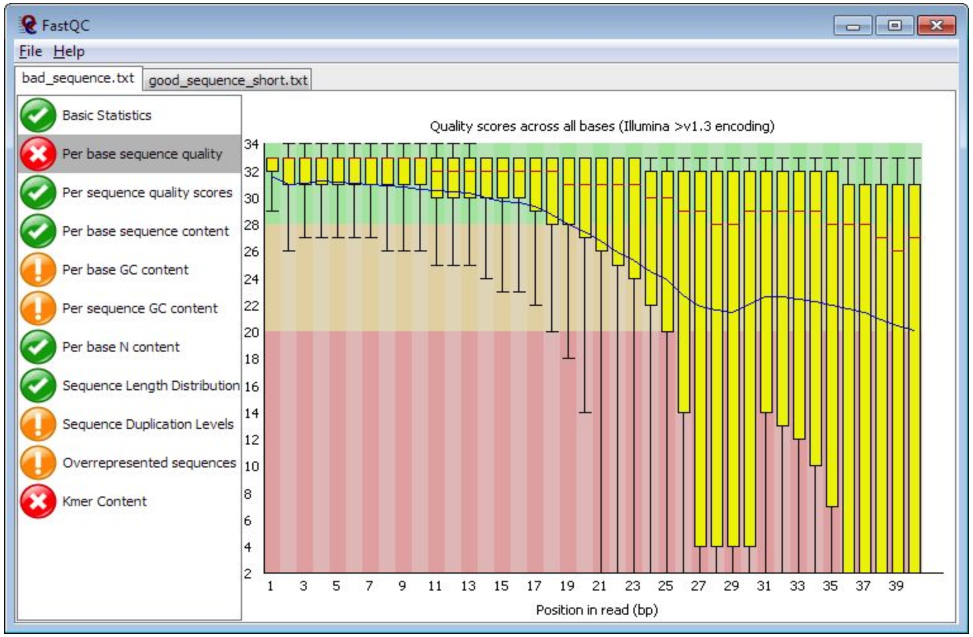
NGS pipeline



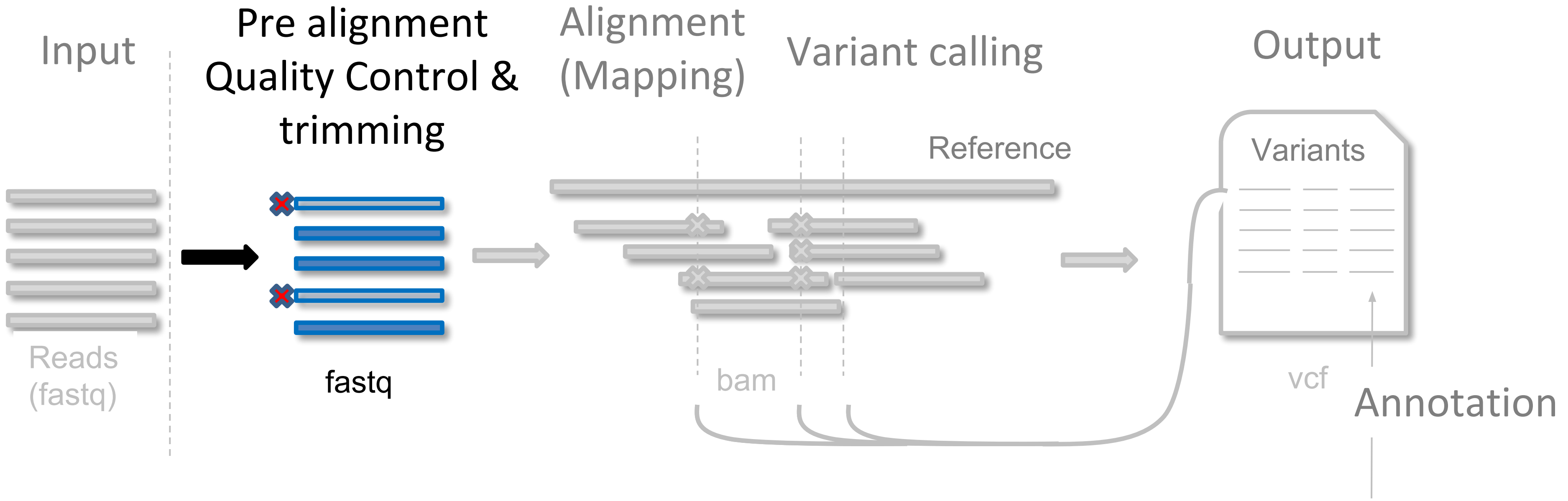


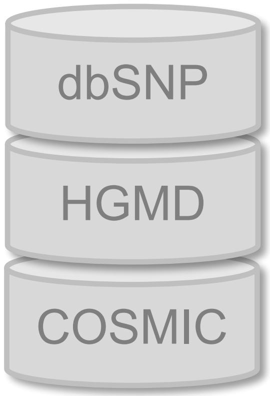


Quality control (FastQC)



- ✓ 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

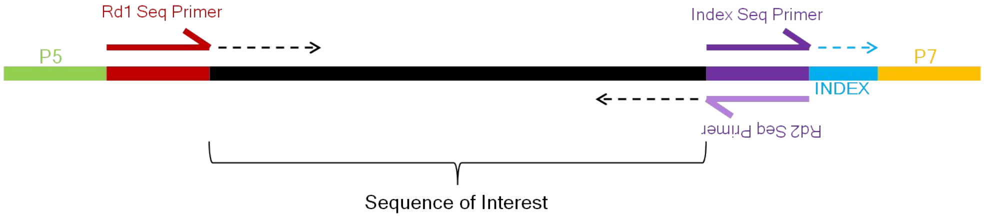


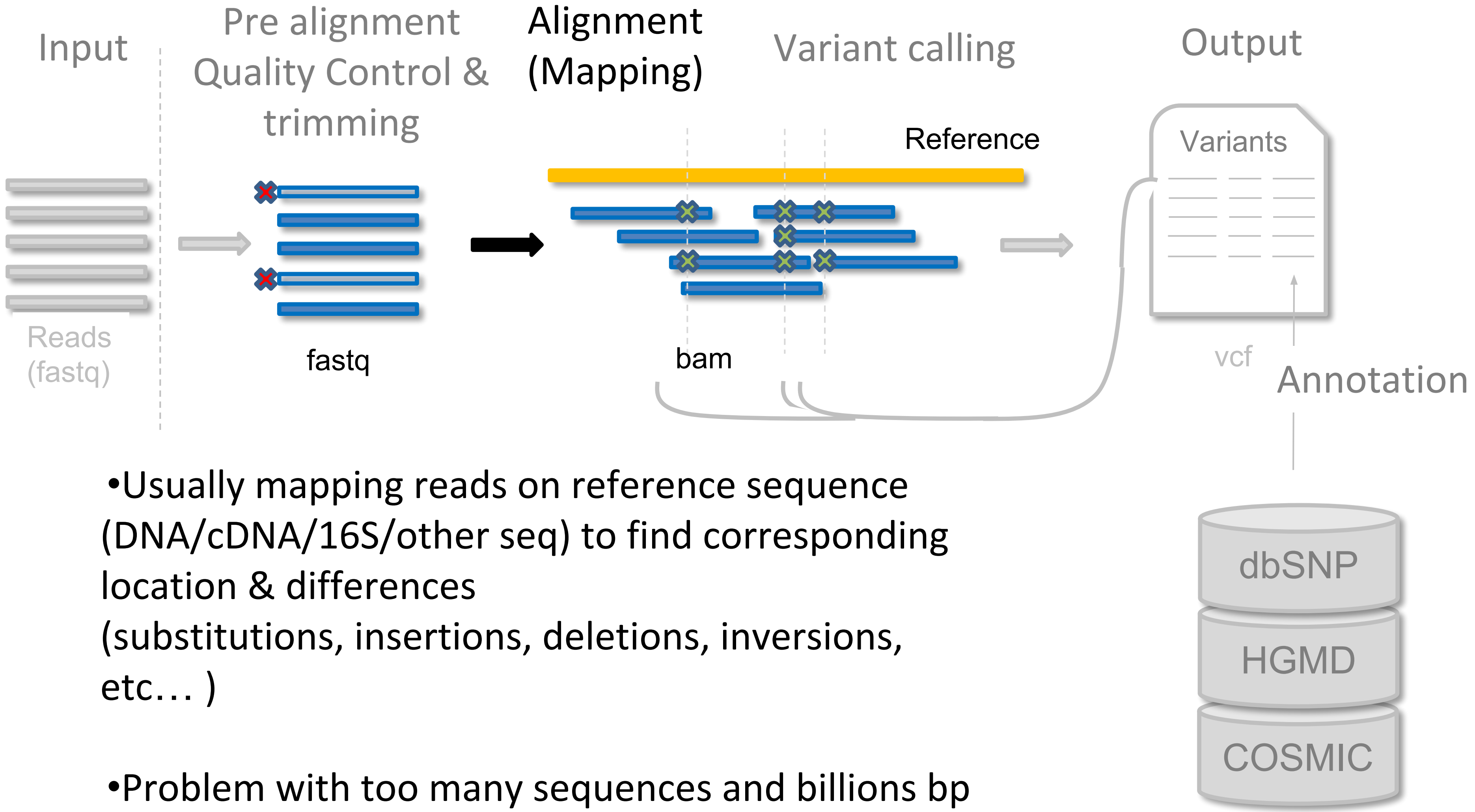


Cleaning reads (Cutadapt)

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

STRUCTURE DETAILS



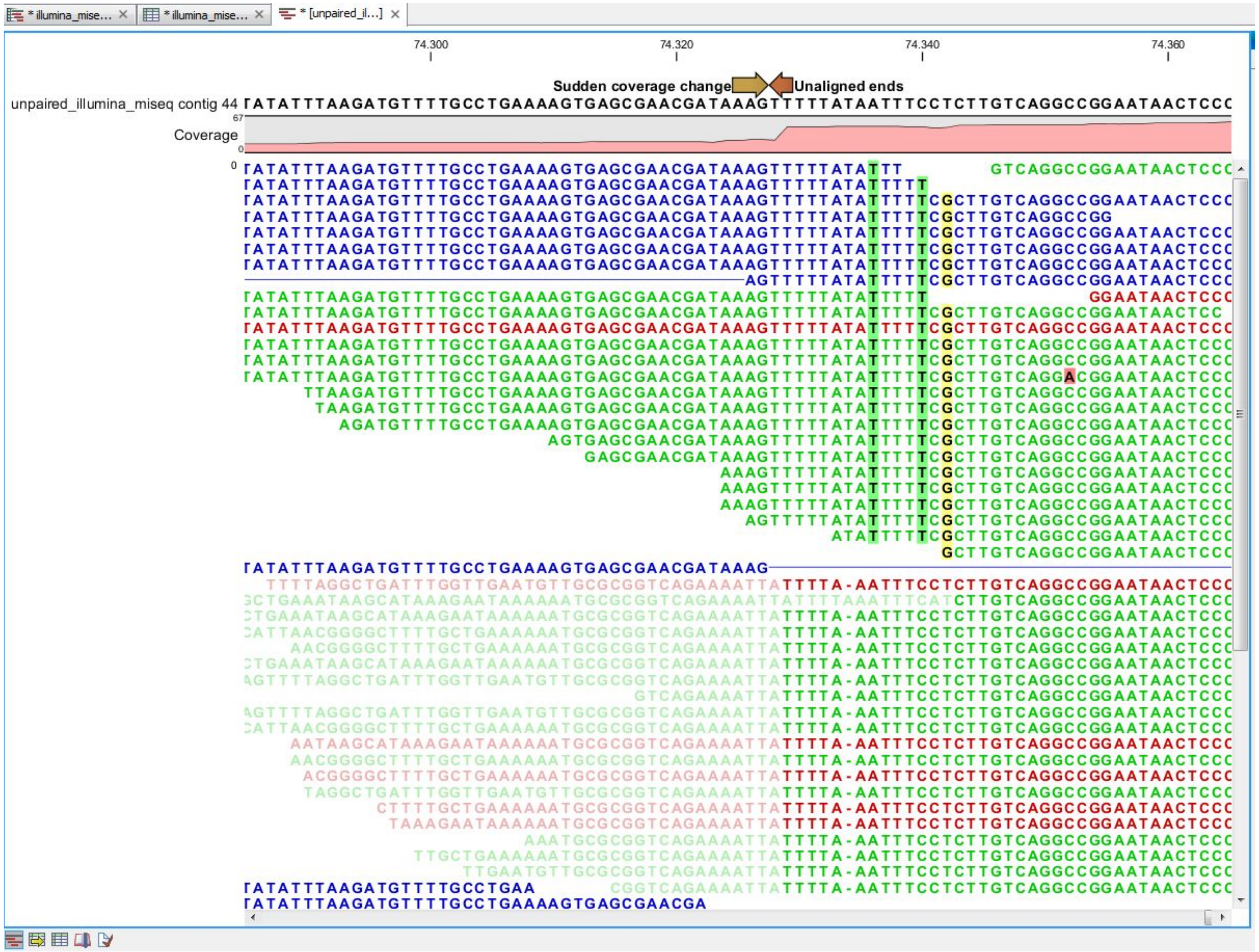


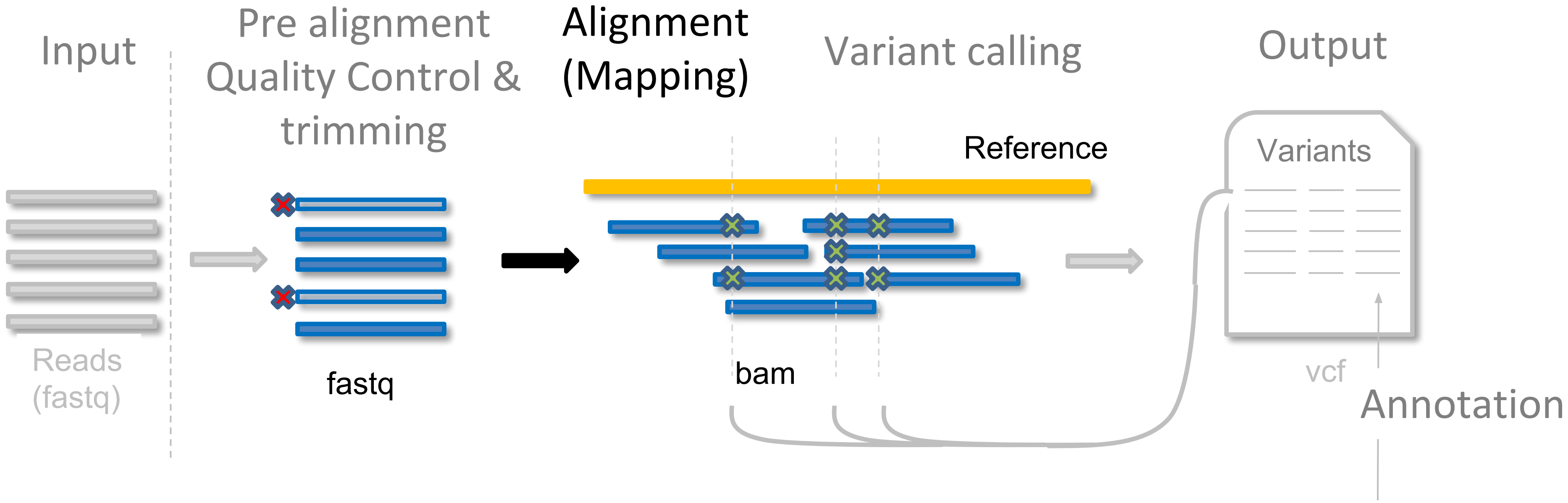
- Usually mapping reads on reference sequence (DNA/cDNA/16S/other seq) to find corresponding location & differences (substitutions, insertions, deletions, inversions, etc…)

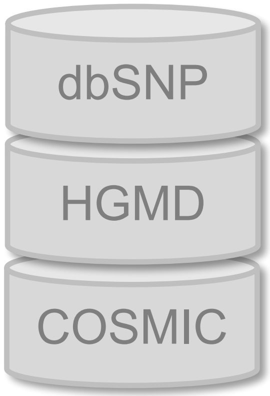
- Problem with too many sequences and billions bp long references – need for special algorithms (Burrows-Wheeler transform, hash table indexing)

- BWA, Bowtie, Bfast, SHRiMP (BAM format)

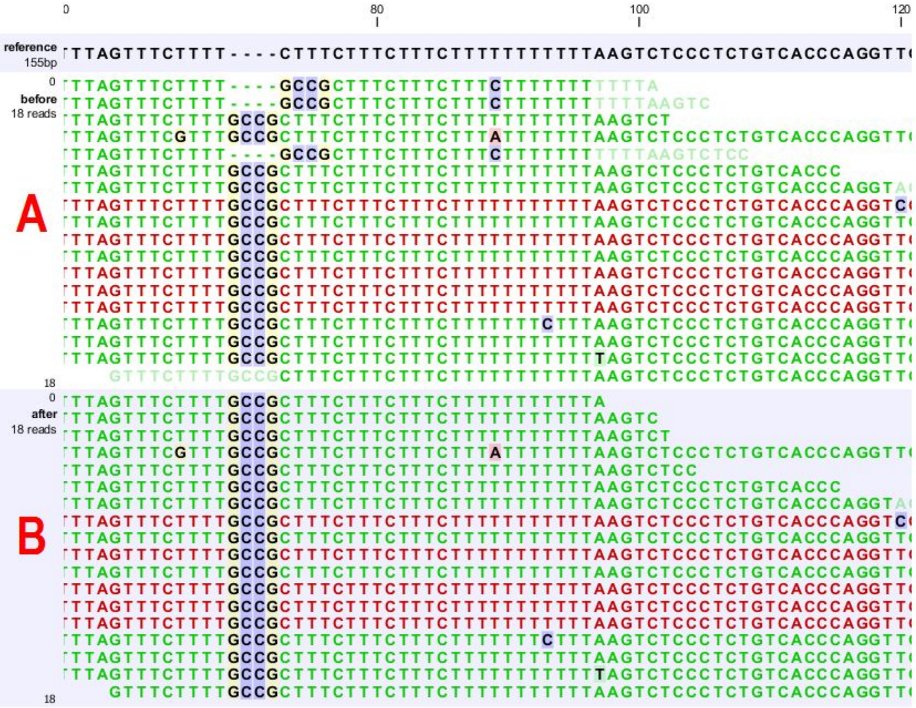
Example of read mapping



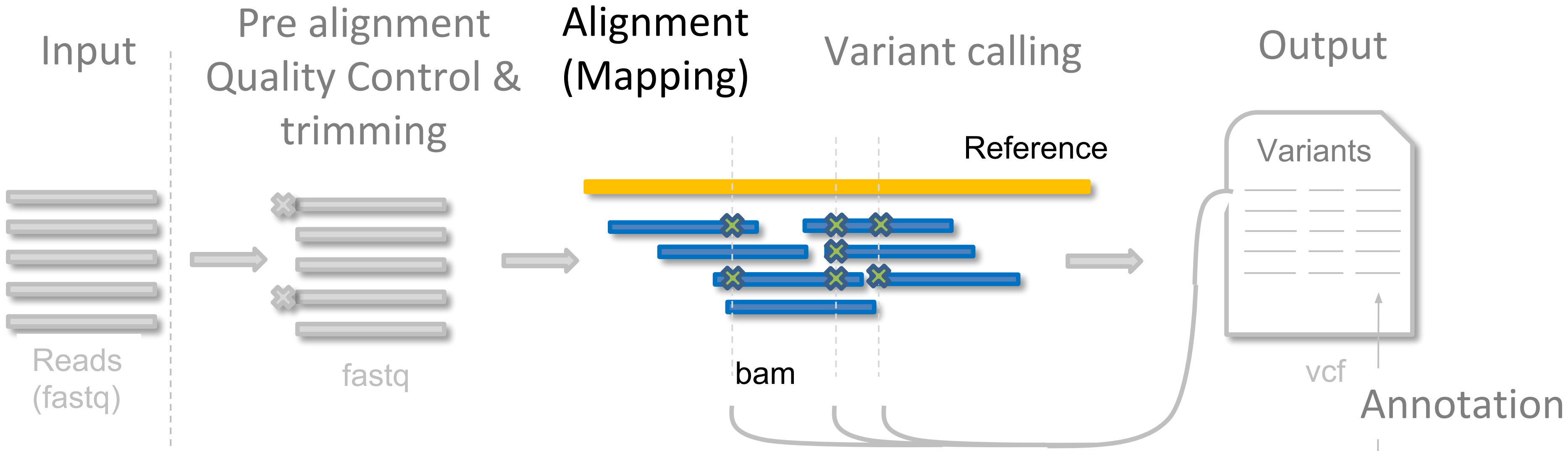


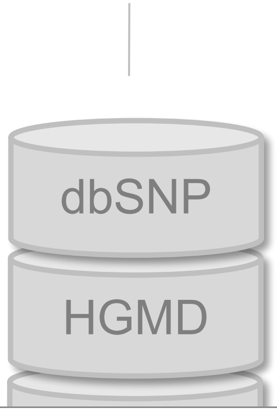


Usually alignment is not perfect – false positive indels & Substitutions => Need for local indel realignment



```
reference  TTAGTTTCTTTT----CTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
155bp
0
before     TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTA
18 reads   TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCT
           TTAGTTTCGTTTGCCGCTTTCTTTCTTTCTTTATTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTT----GCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTA
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTCTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTTAGTCTCCCTCTGTCACCCAGGTT
              GTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
18
0
after      TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTA
18 reads   TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCT
           TTAGTTTCGTTTGCCGCTTTCTTTCTTTCTTTATTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTA
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTC
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTCTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
           TTAGTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTTAGTCTCCCTCTGTCACCCAGGTT
              GTTTCTTTTGCCGCTTTCTTTCTTTCTTTTTTTTTTTAAGTCTCCCTCTGTCACCCAGGTT
18
```



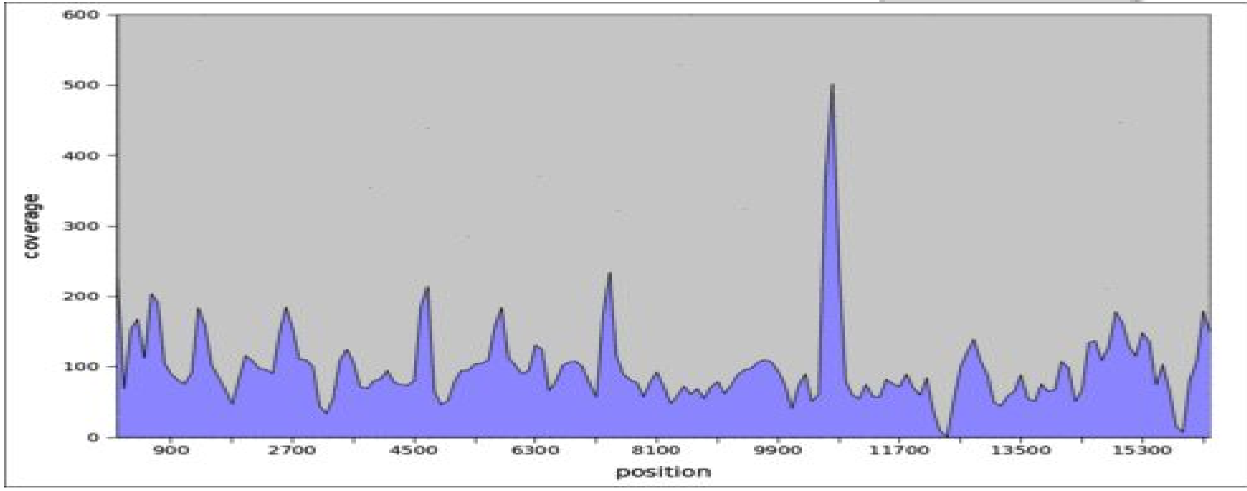


Mapping, Coverage reports

- Repeat alignment/other steps with different criteria?
- Important checkout for lab protocol
- Specificity of PCR
- Settings of variant calling threshold, CNV
- Target bed file (Browser Extensible Data)

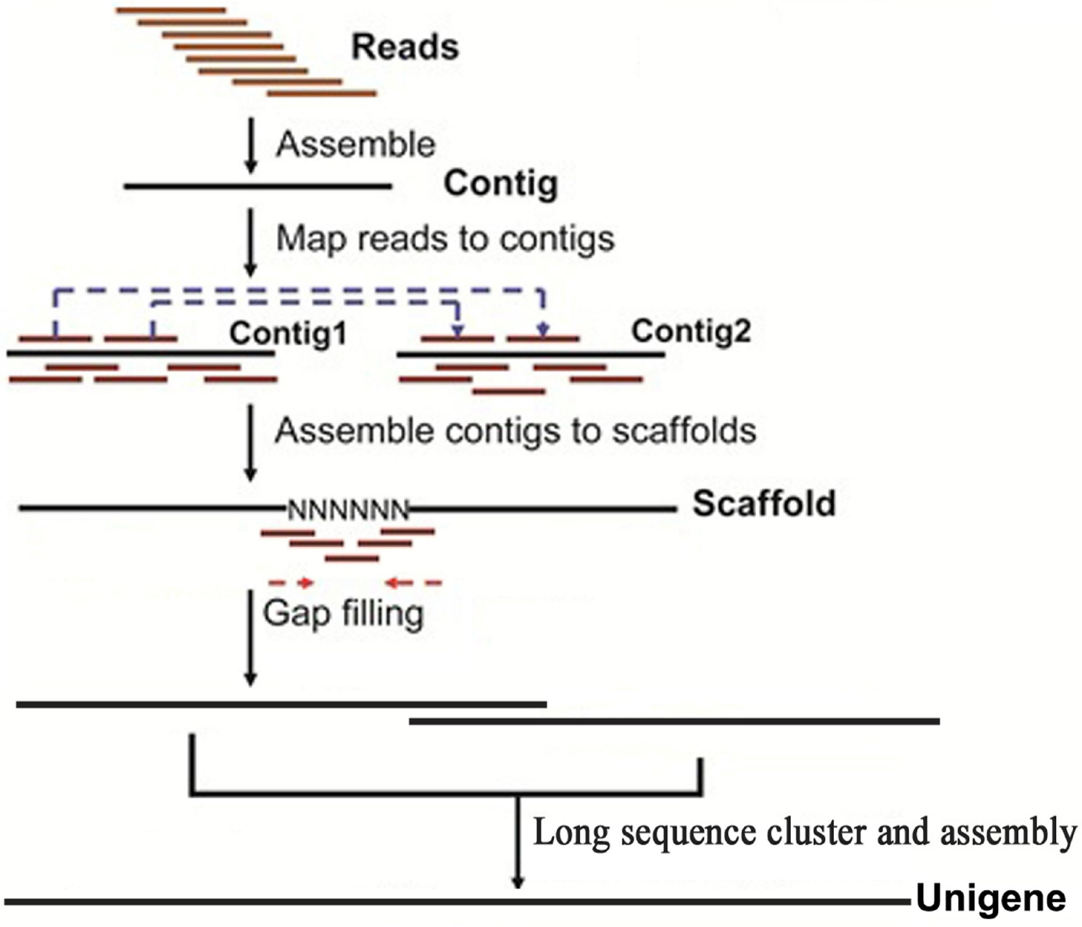
chr7	127471196	127472363
chr7	127472363	127473530
chr7	127473530	127474697
chr7	127474697	127475864
chr7	127475864	127477031

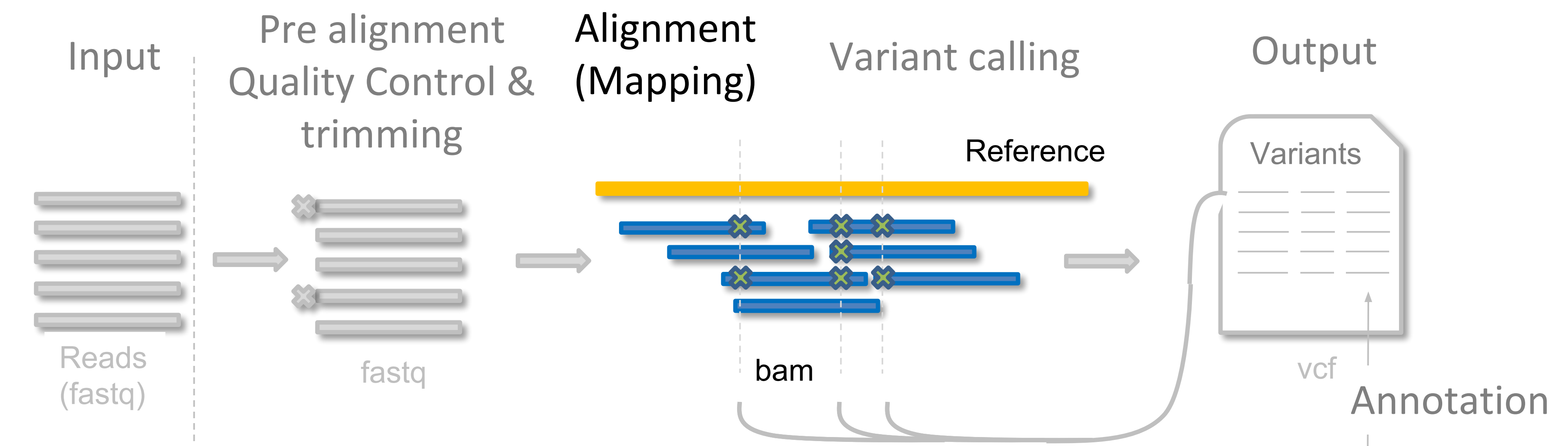
(bed format)



De novo assembly – alternative for mapping on reference sequence

- To uncover unknown genomes/transcriptomes
- To detect large structural variants





Input

Pre alignment Quality Control & trimming

Alignment (Mapping)

Variant calling

Output

Reference

Variants

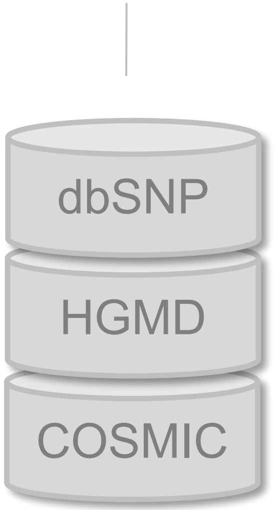
Reads (fastq)

fastq

bam

vcf

Annotation



dbSNP

HGMD

COSMIC

REMOVE PCR DUPLICATES

Each read represents 1 input molecule

THEORY:

E.g. in case of DNA re-sequencing, 1 diploid cell is represented by 2 reads because of 2 chromosomes

BUT

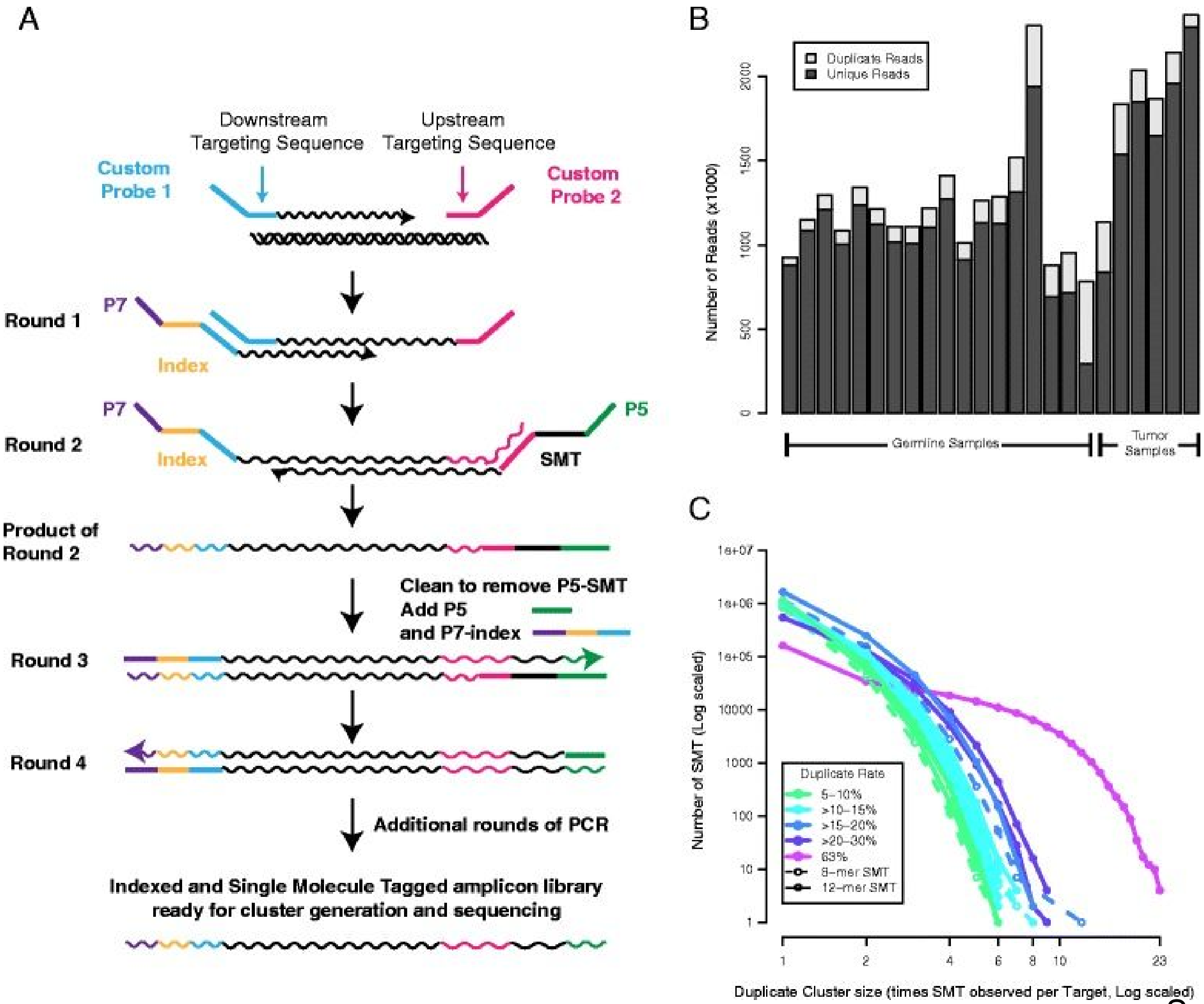
there is a PCR to amplify genetic material to be analyzable => 1 input molecule from 1 cell could be after PCR represented by more reads => Biased variant allele frequency

How to solve it?

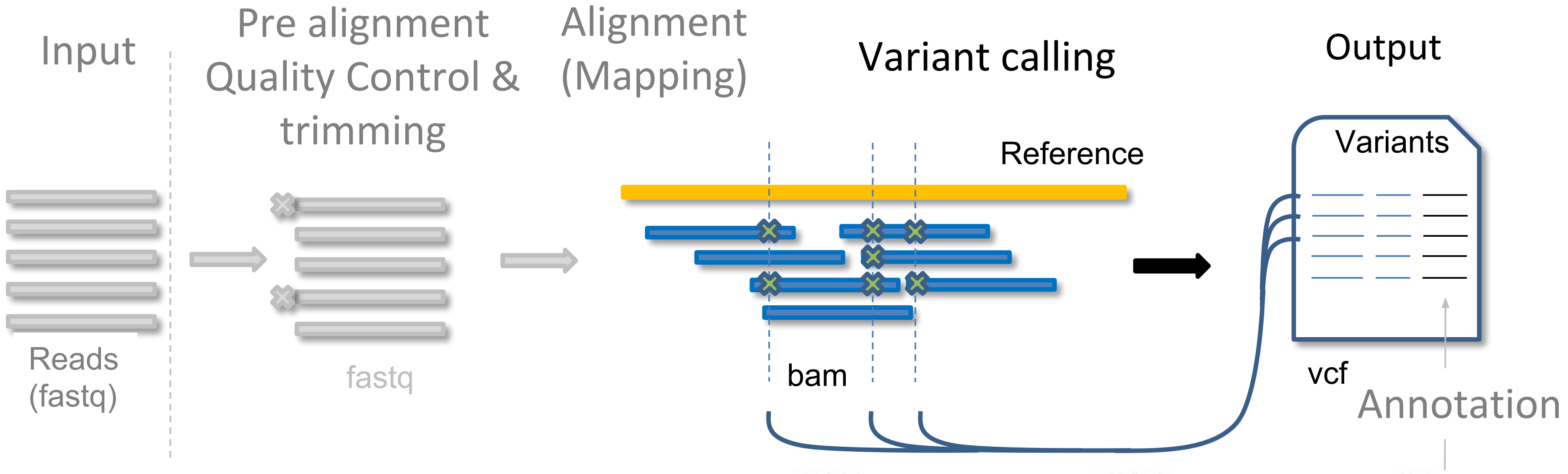
1) Molecular barcodes (very new method)

2) Identity of start-end positions of read pair

Introduction of Molecular barcodes during library preparation



Smith et al. 2014



Input

Pre alignment Quality Control & trimming

Alignment (Mapping)

Variant calling

Output

Mutation types:

Germinal mutations

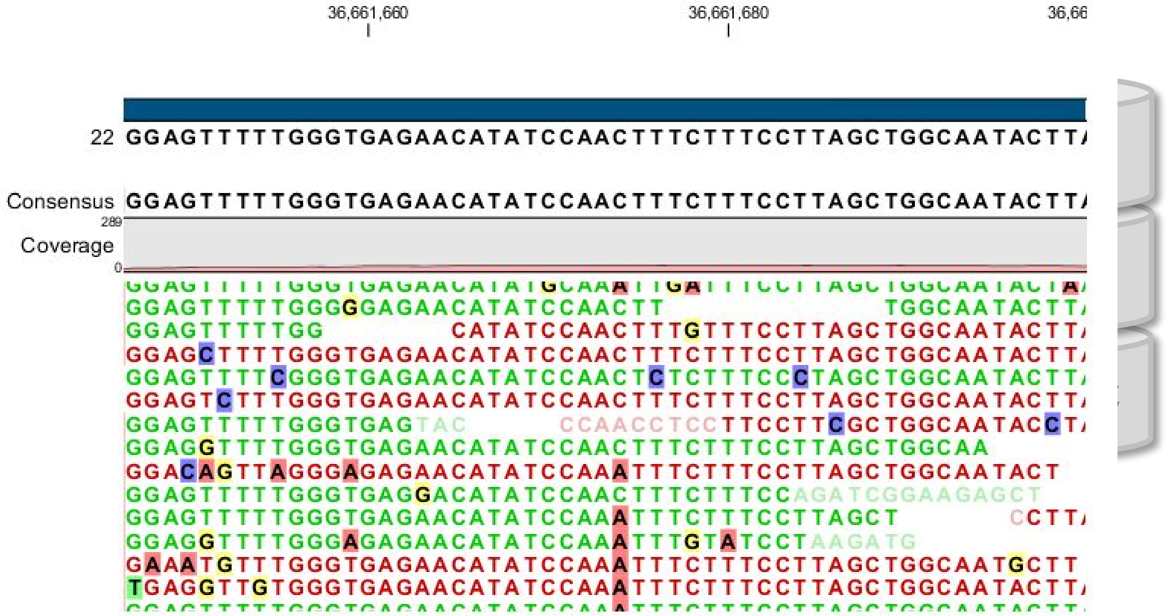
Somatic mutations

Substitutions

Insertions

Deletions

Complex variants

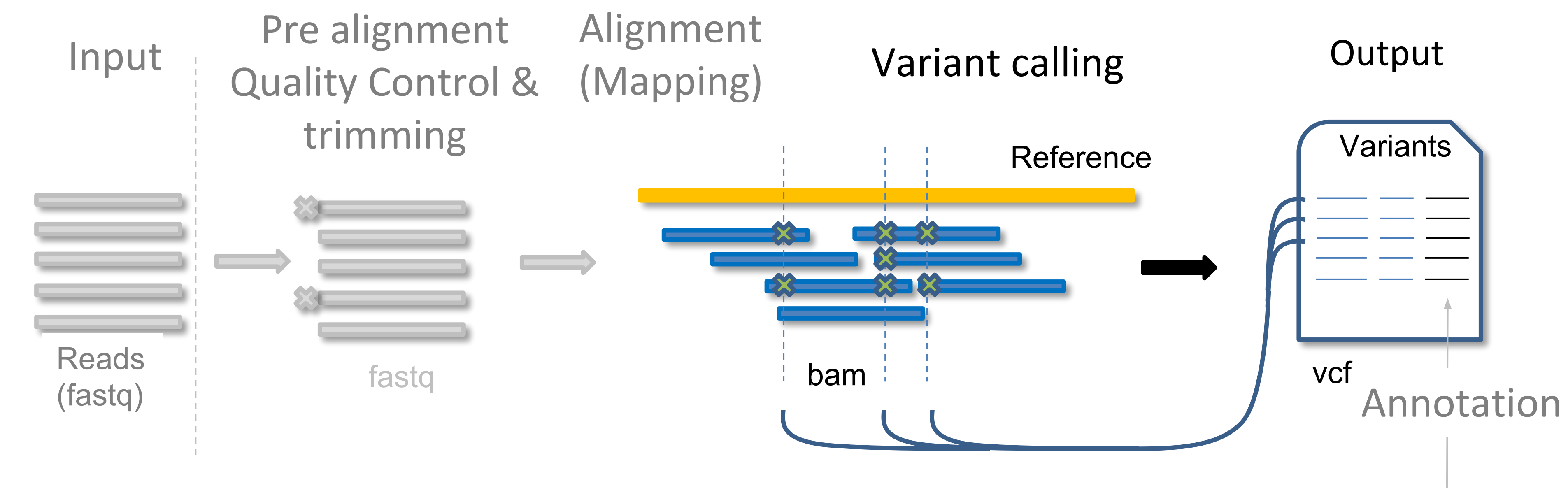


36,661,660
36,661,680
36,66
22 GGAGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTTA
Consensus GGAGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTTA
289
Coverage
0

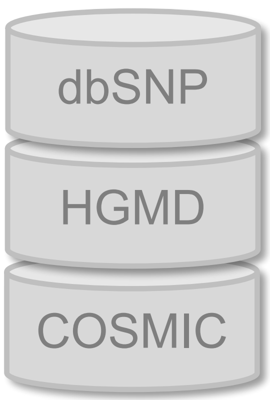
Inversions

Large structural variations (translocations, indels)

Copy number variations



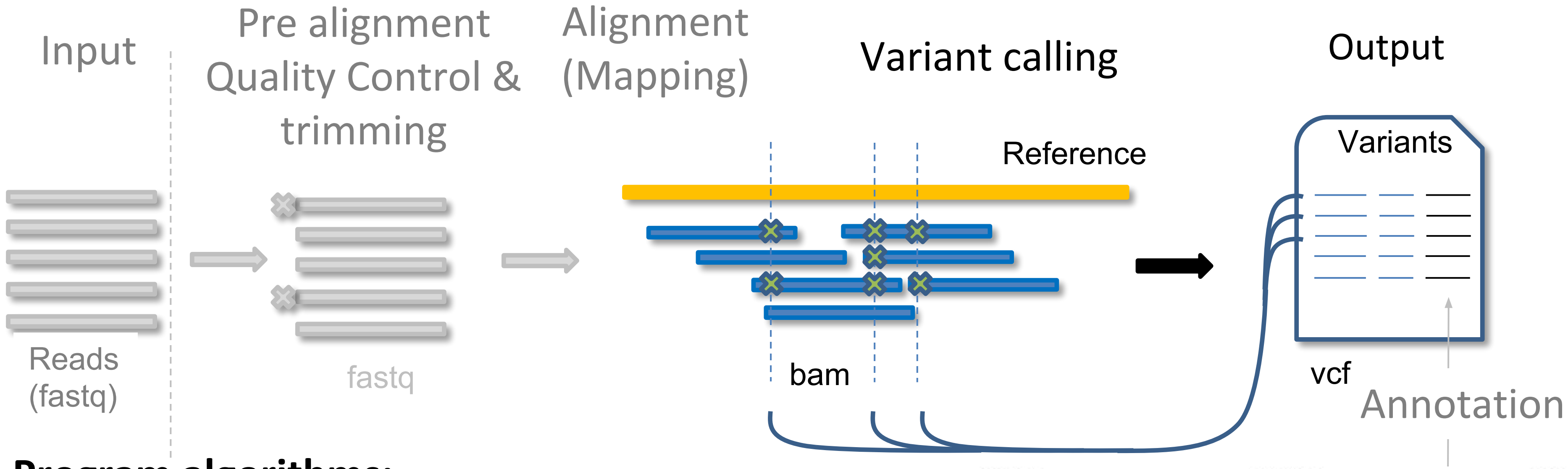
Experimental designs (also depends on types of samples available):

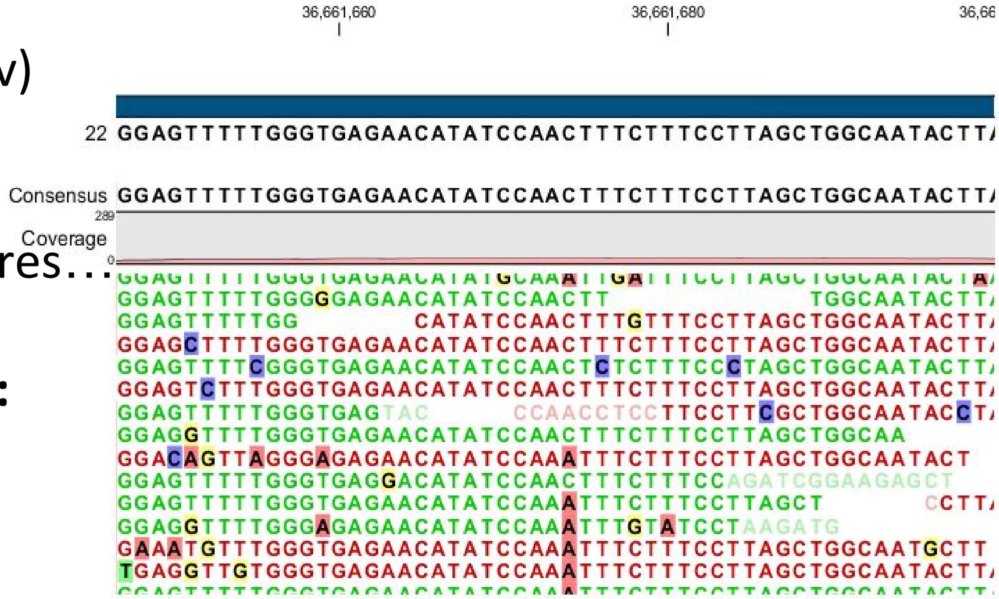


Normal only (genotyping)
Tumor only (genotyping, somatic mutations)

Tumor + related normal control
Tumor + unrelated normal controls
Tumor in time

Family (rare diseases, genotyping)





```
22        GGAGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
Consensus GGAGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
          GGAGTTTTGGGTGAGAACATATGCAAATTGATTTCCTTAGCTGGCAATACTA
          GGAGTTTTGGGGGAGAACATATCCAACTT            TGGCAATACTT
          GGAGTTTTGG       CATATCCAACTTTGTTTCCTTAGCTGGCAATACTT
          GGAGCTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
          GGAGTTTCGGGTGAGAACATATCCAACTCTCTTTCCCTAGCTGGCAATACTT
          GGAGTCTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAATACTT
          GGAGTTTTGGGTGAGTAC     CCAACCTCCTTCCTTCGCTGGCAATACCT
          GGAGGTTTTGGGTGAGAACATATCCAACTTTCTTTCCTTAGCTGGCAA
          GGACAGTTAGGGAGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATACT
          GGAGTTTTGGGTGAGGACATATCCAACTTTCTTTCCAGATCGGAAGAGCT
          GGAGTTTTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCT      CCTT
          GGAGGTTTTGGGAGAGAACATATCCAAATTTGTATCCTAAGATG
          GAAATGTTTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATGCTT
          TGAGGTTGTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATACTT
          GGAGTTTTGGGTGAGAACATATCCAAATTTCTTTCCTTAGCTGGCAATACTT
```

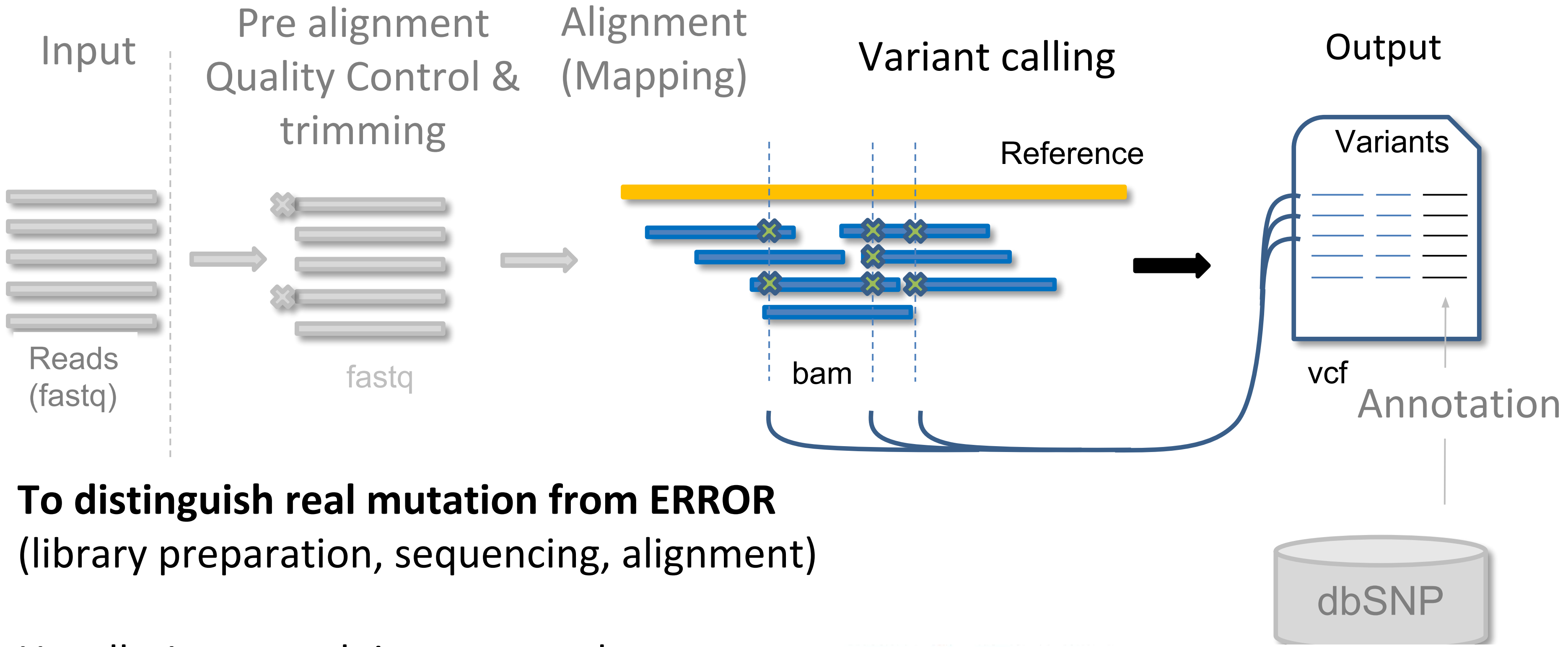
Program algorithms:

- Bayesian statistics (Mutect, DeepSnv)
- Fisher exact test (Varscan, Vardict)
- …

Giving p-value based on different features…

Options for many parameters & filters:

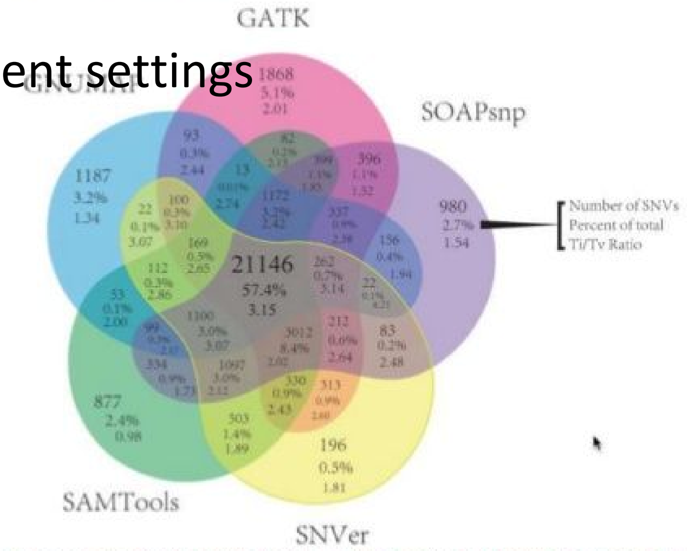
- Minimum coverage
- Variant allele frequency
- Base quality
- Genomic context (homopolymers)
- Position in read (errors at the reads end)
- Mapping quality
- Presence in both forward and reverse reads (strand bias)



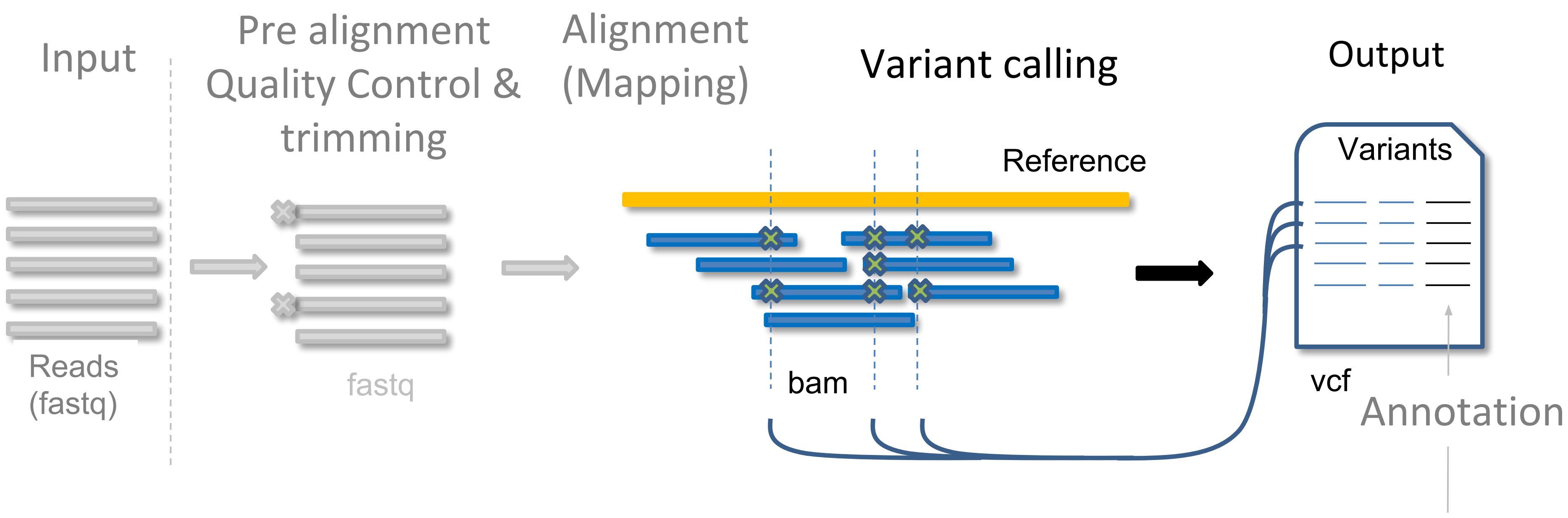
To distinguish real mutation from ERROR
(library preparation, sequencing, alignment)

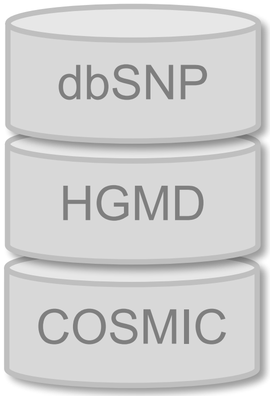
Usually 1 approach is not enough =>
to combine more variant callers (aligners) & different settings

Specific pipeline for each type of mutations
(SNV, INDELS, CNV…)

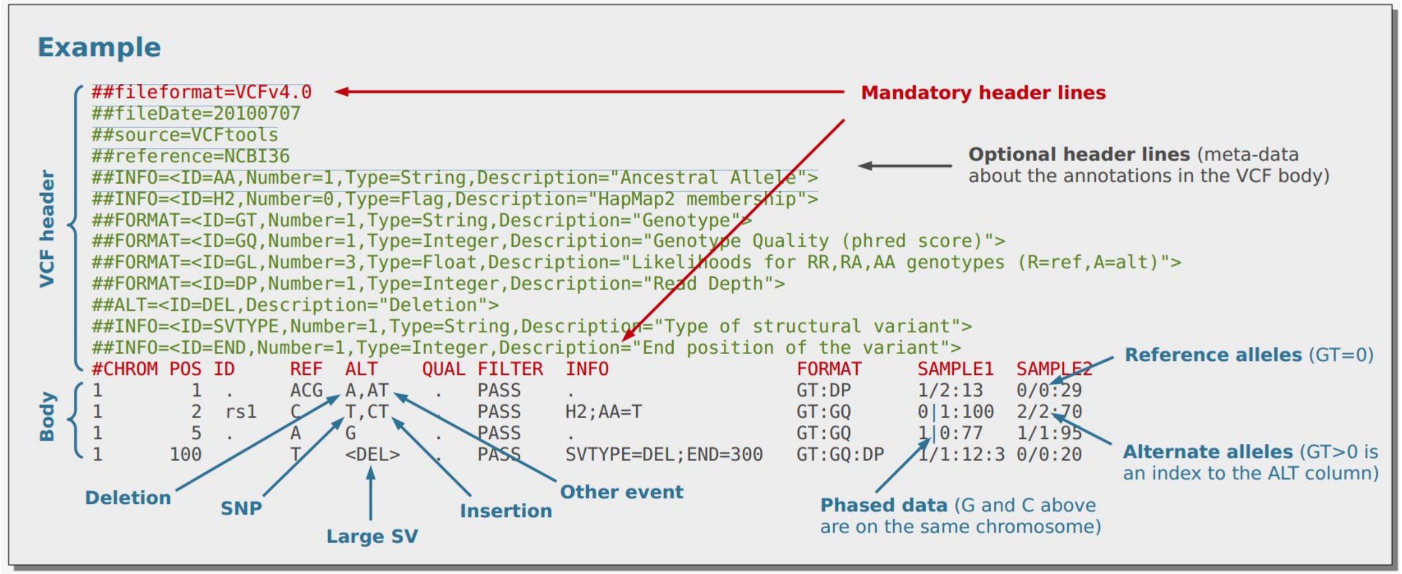


O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* **5**, 28 (2013).



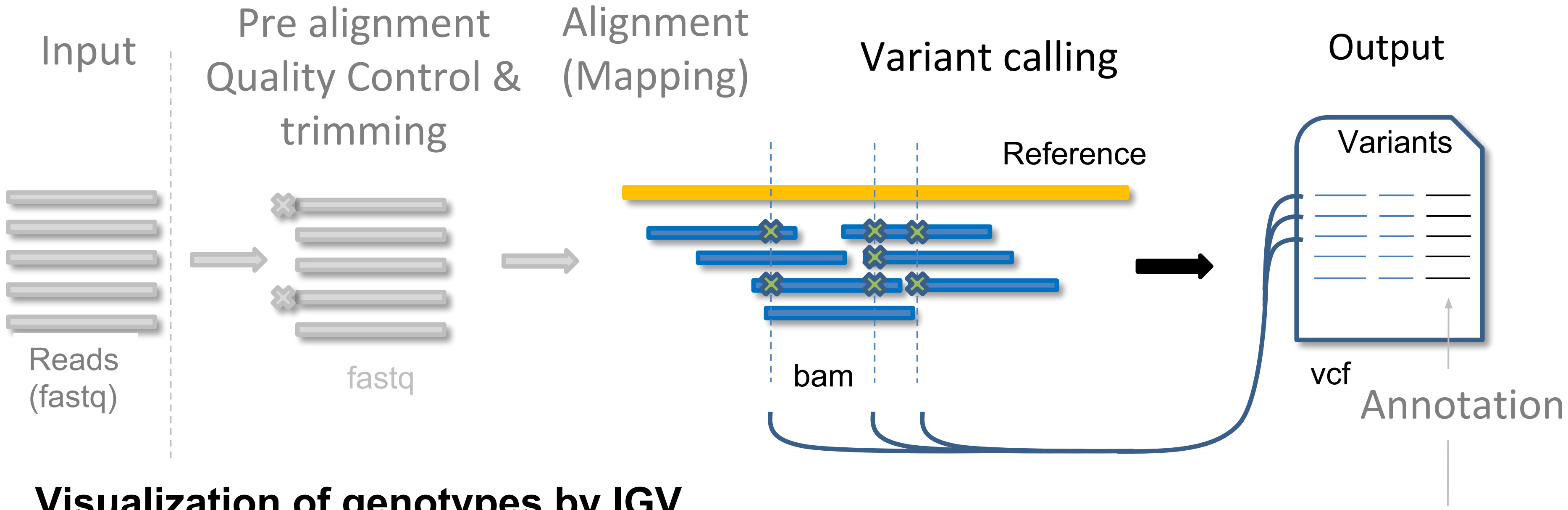


VCF file


Example

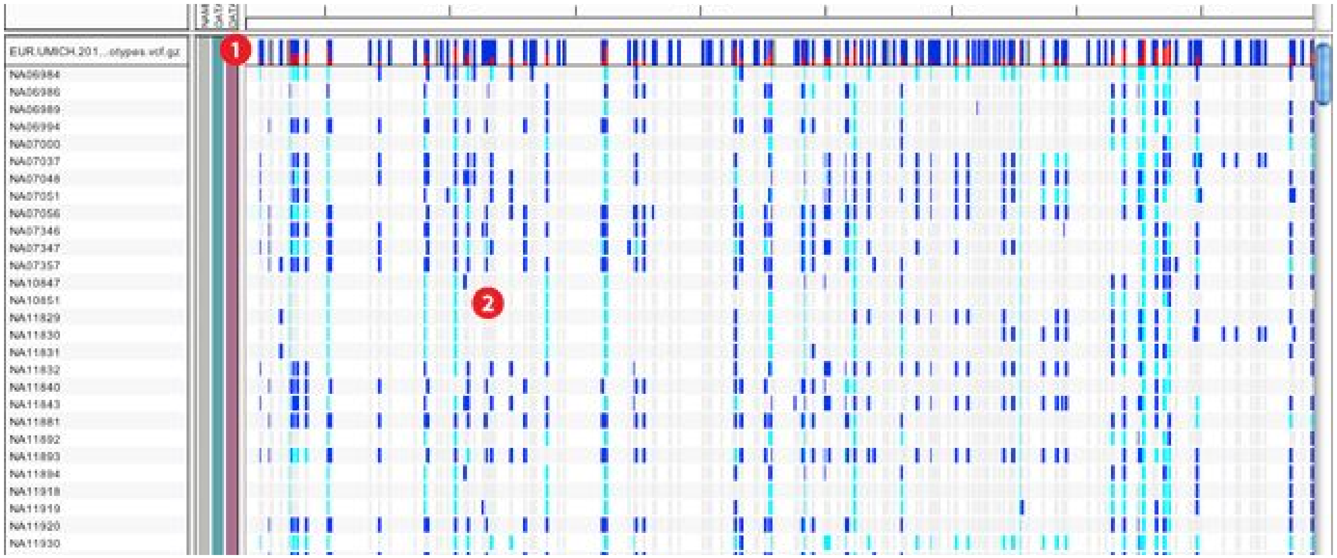
```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype 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
```

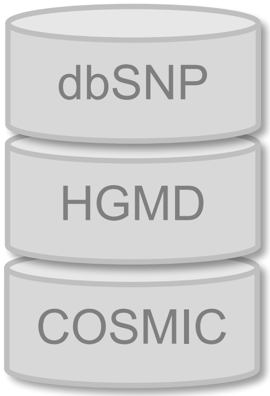
VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)



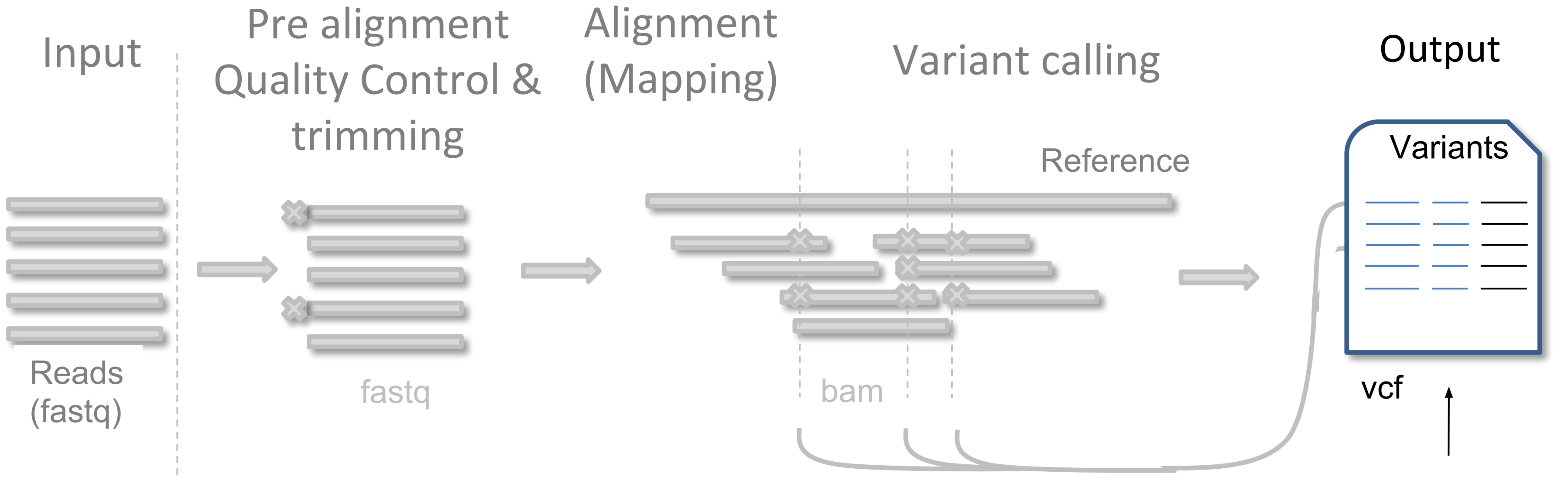
Visualization of genotypes by IGV

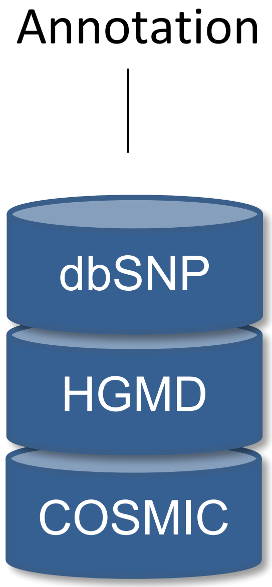
Input vcf





1) Each bar across the top of the plot shows the allele fraction for a single locus.
2) The genotypes for each locus in each sample. Dark blue = heterozygous, Cyan = homozygous variant, Grey = reference.





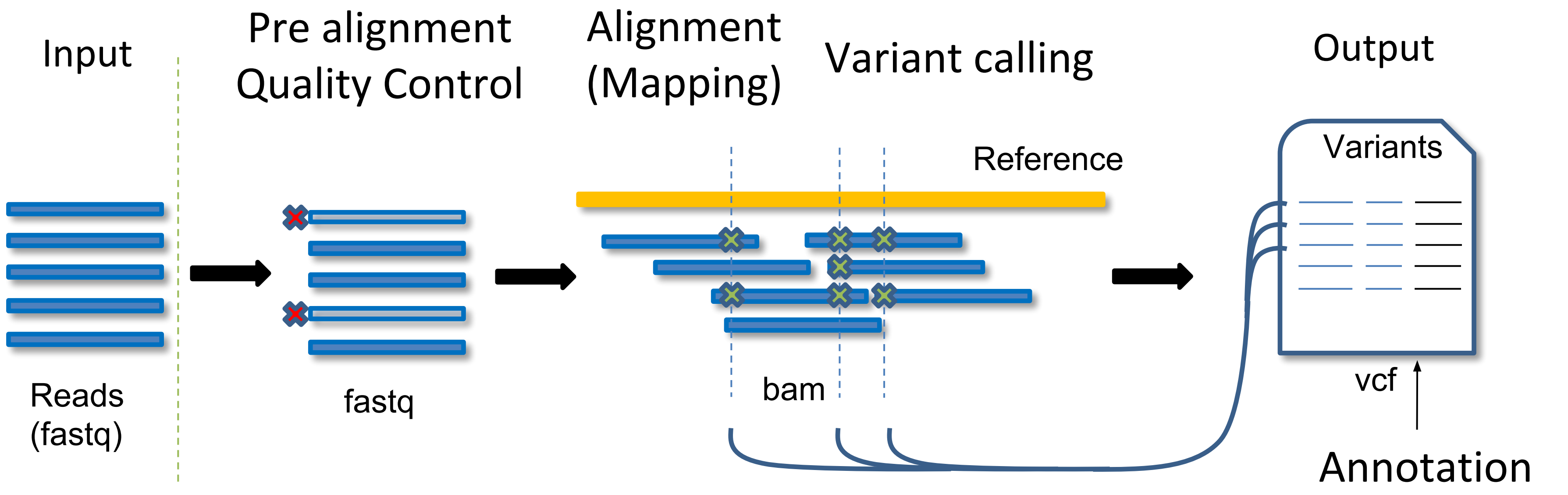
Annotation

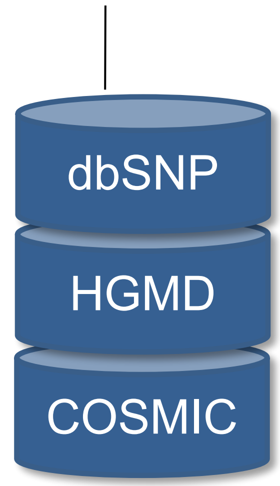
From genomic coordinate to biological meaning

Provide links to various databases (RefSeq, dbSNP, etc.)

To distinguish significant variant from non-significant (synonymous vs. non-synonymous, gene, exon, intron, cDNA, codon, transcript, freq in population, presence in other diseases…)

- RefSeq
- dbSNP
- Regulation
- Comparative genomics
- Repeats
- Functional
- Gene ontology
- Etc.





Sensitivity & Specificity as a matter of:

- Experiment design

(library preparation + NGS technology + number of samples + amount of data)

- Data processing

(pre-processing + alignment + variant calling + annotations + filtering)

Courses

http://meetings.embo.org/event/17-genome

http://www.embo.org/events/practical-courses