CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

Moderní metody analýzy genomu
Bioinformatika I

Mgr. Nikola Tom

Brno, 13.11.2017

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)
Intersection of biology, computer science and statistics

AIM: clean the data and give them biological sense

NGS data analysis = bottleneck of NGS

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)

CLC bio
A QIAGEN® Company

SOPHIA GENETICS

Bioinformatics

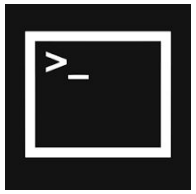
Bioinformatics **SOLUTION 2**:

- command-line based tools/software

Each tools solves only a part of the analysis

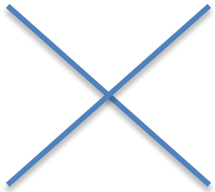
- Need for setup the pipeline & tune programs' parameters

(challenging & more precise!!!)



Bioinformatics

Modern laptop or PC might be enough… but bigger computer = better



Before we start the analysis

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

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

Concept of the project
DNA/RNA/epigenomics/metagenomics…

DNA

- Targeted sequencing - amplicons, gene panels, whole exomes
  (target enrichment methods - PCR, ligation…)
- Whole genome sequencing
  - Finding differences to known reference genome = re-sequencing

- De novo assembly
  - Genome (re)construction

Before we start analysis

RNA
- Gene expression, miRNA, ncRNA, alternative splicing

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

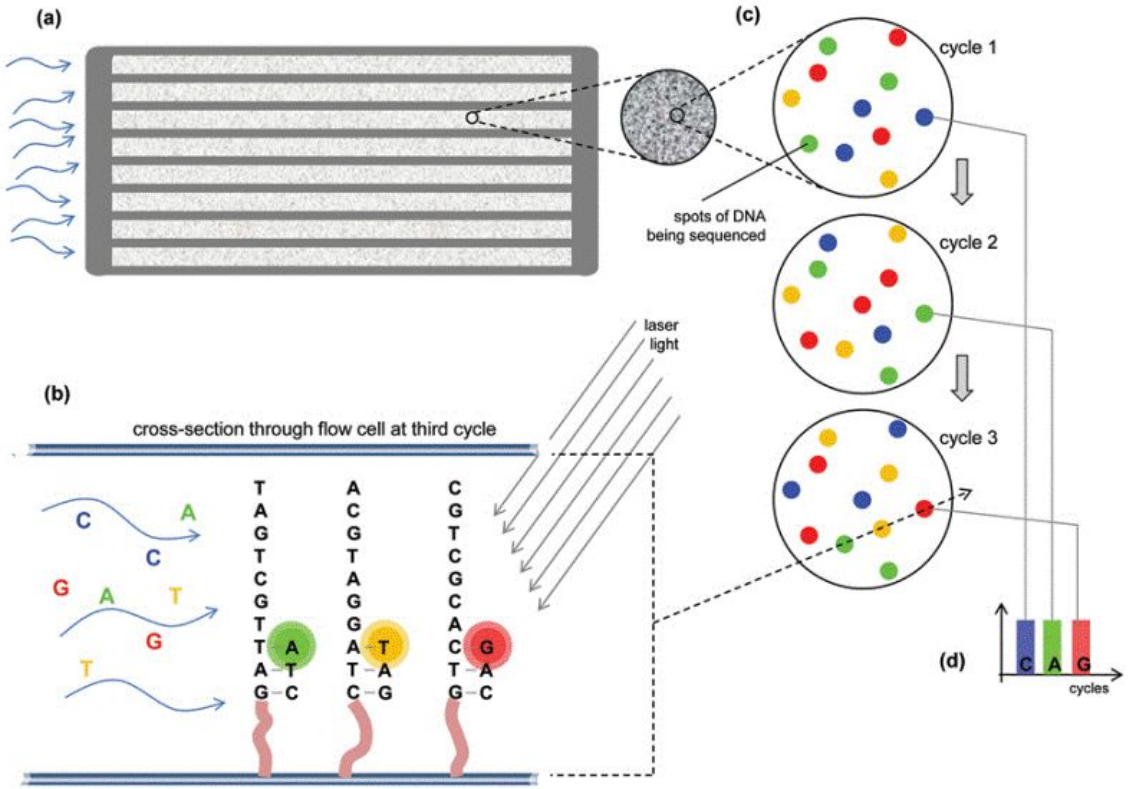
Epigenomics
- DNA-protein interactions, methylations

Bioinformatics' starting point

Raw sequencing data - READ

Produced during **base calling**

- signal (fluorescence, electric current) 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)
- Pair-end sequencing 1 molecule = 2 reads = 2 fastq files (R1, R2)

- Corresponding quality score for each base
- **Phred score** – probability of arising an error (log based)

Q10 = 1 in 10 = 90% base accuracy
Q20 = 1 in 100 = 99% base accuracy
Q30 = 1 in 1 000 = 99.9% base accuracy
Q40 = 1 in 10 000 = 99.99% base accuracy

- ASCII character

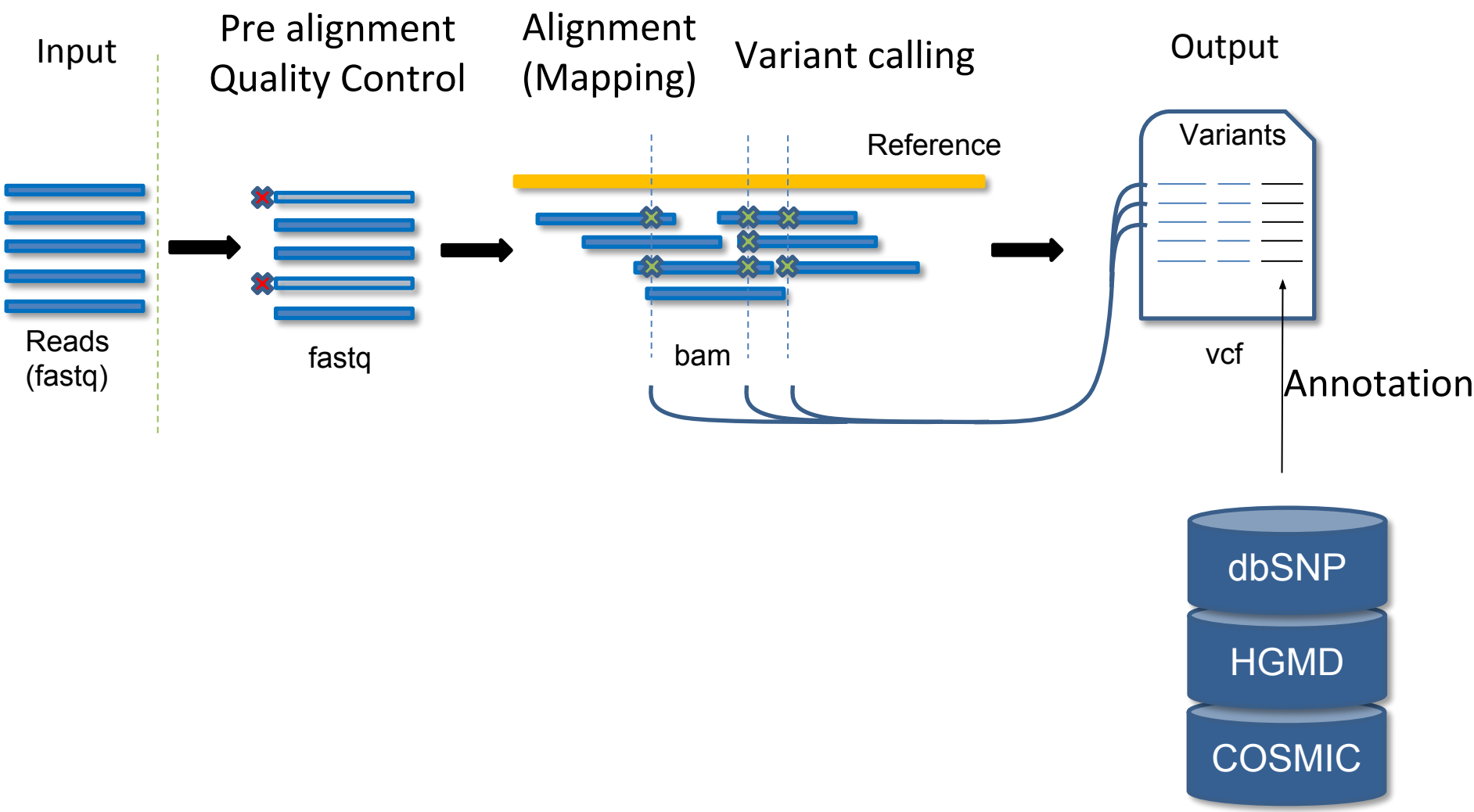
example.fastq

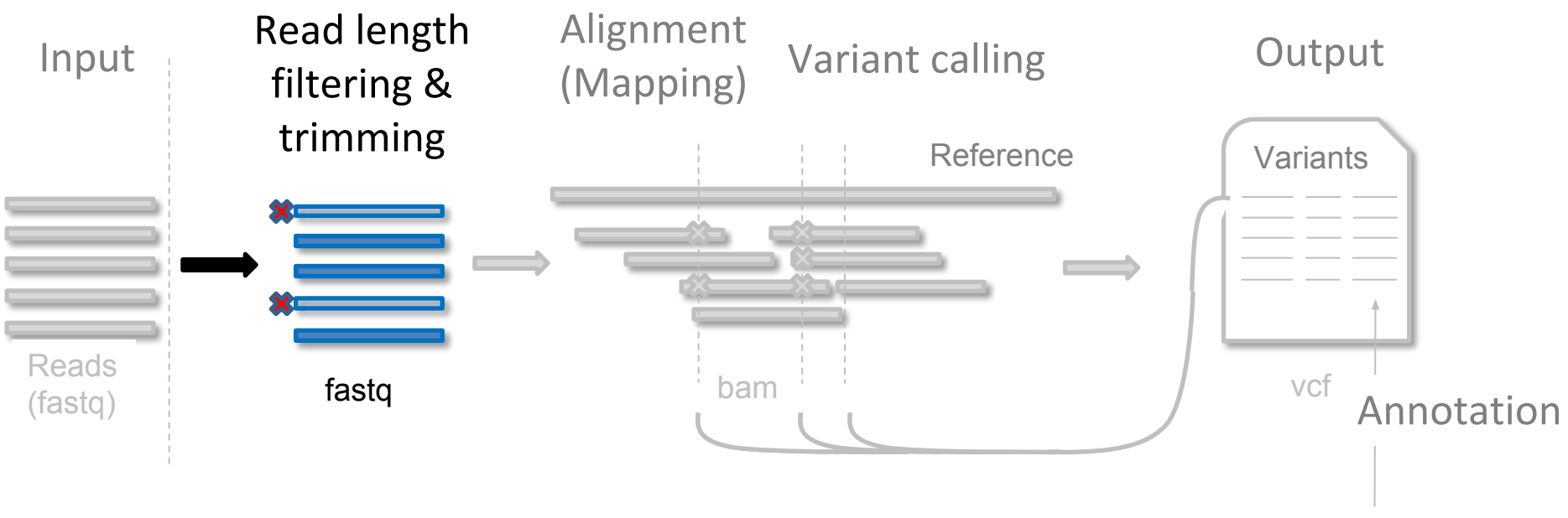
```
@
SEQ_ID GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
 !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

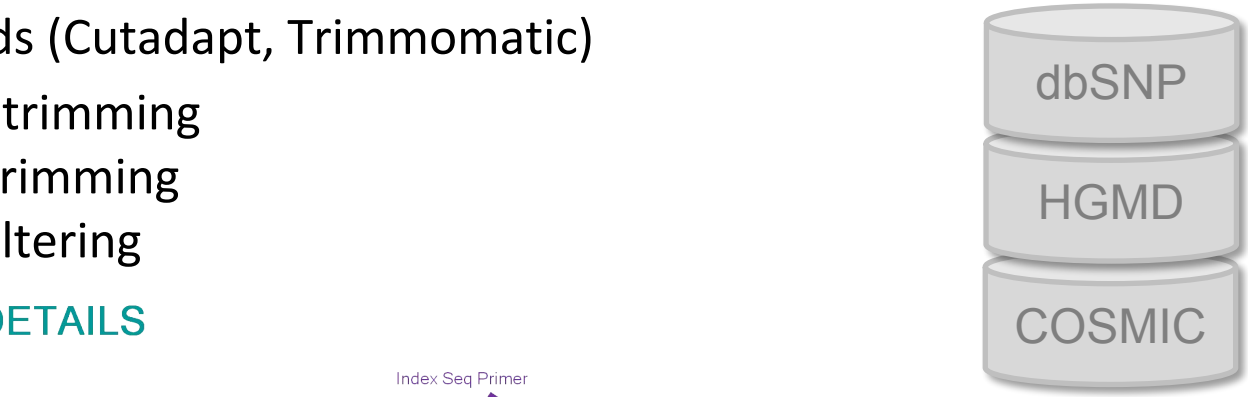
NGS pipeline - DNA re-sequencing



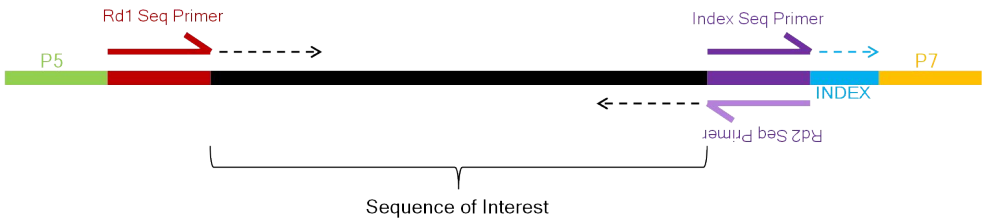


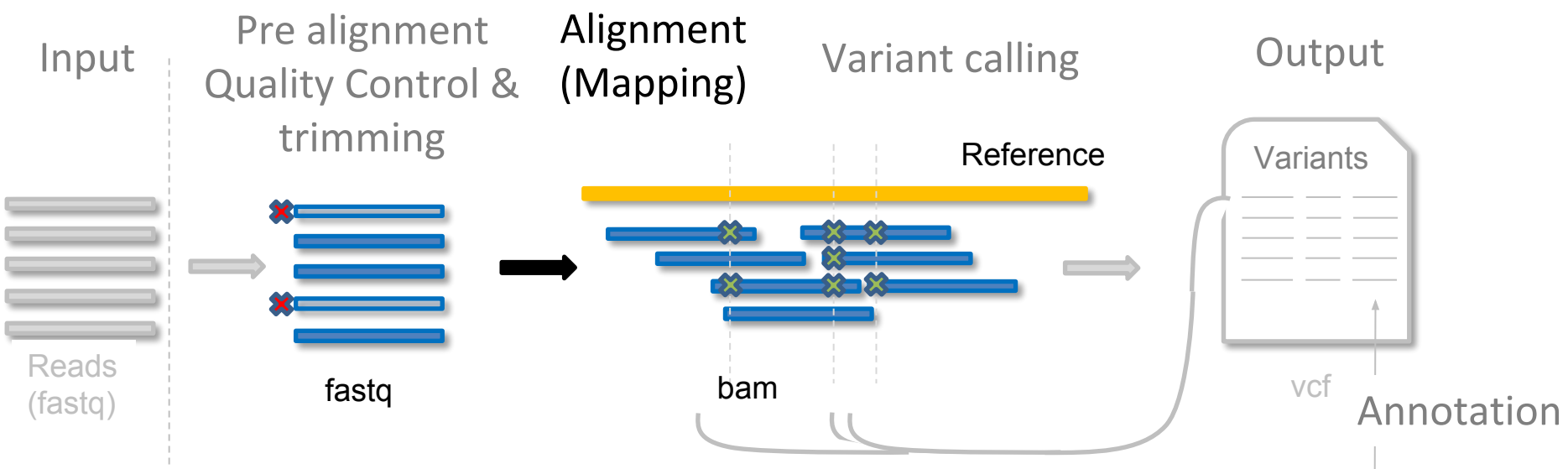
Cleaning reads (Cutadapt, Trimmomatic)

- Adaptor trimming
- Quality trimming
- Length filtering



STRUCTURE DETAILS





•Mapping reads onto **reference sequence** (organism genome or part of the genome)

- to find corresponding location & differences

(substitutions, insertions, deletions, inversions, etc…)

•Problem with:

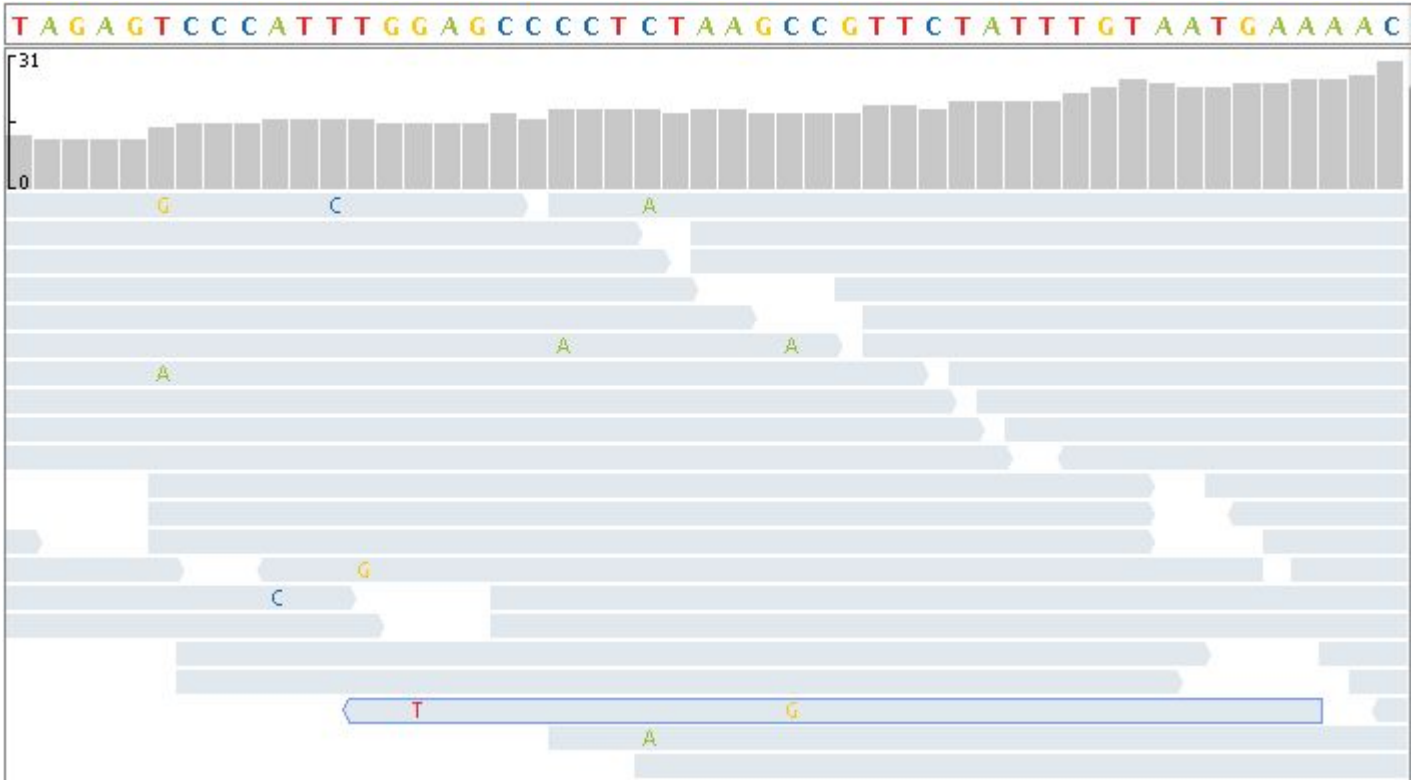
- too many sequences
- billions bp long references
- **non-perfect matches** between reads and reference

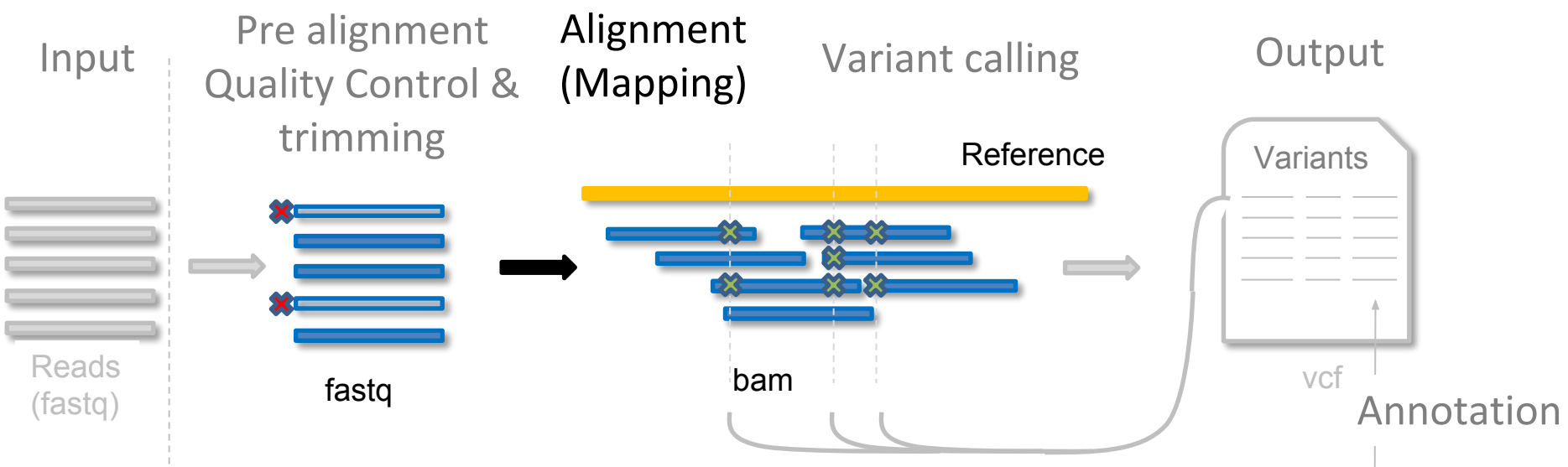
=> need for special algorithms

(Burrows-Wheeler transform, hash table indexing)

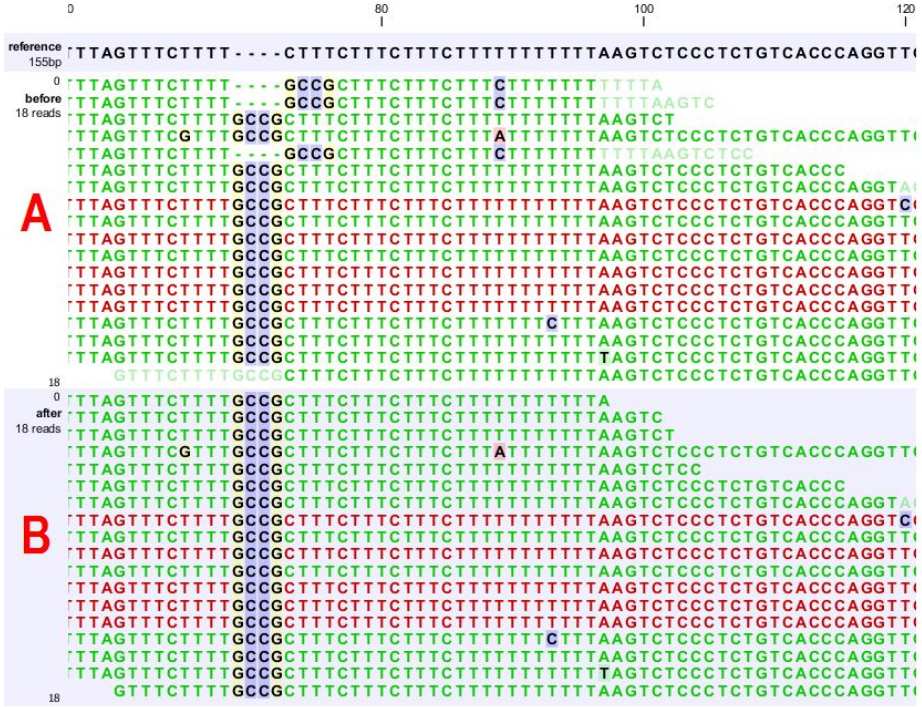
•BWA, Bowtie2, Bfast, SHRiMP (SAM/BAM/CRAM format)

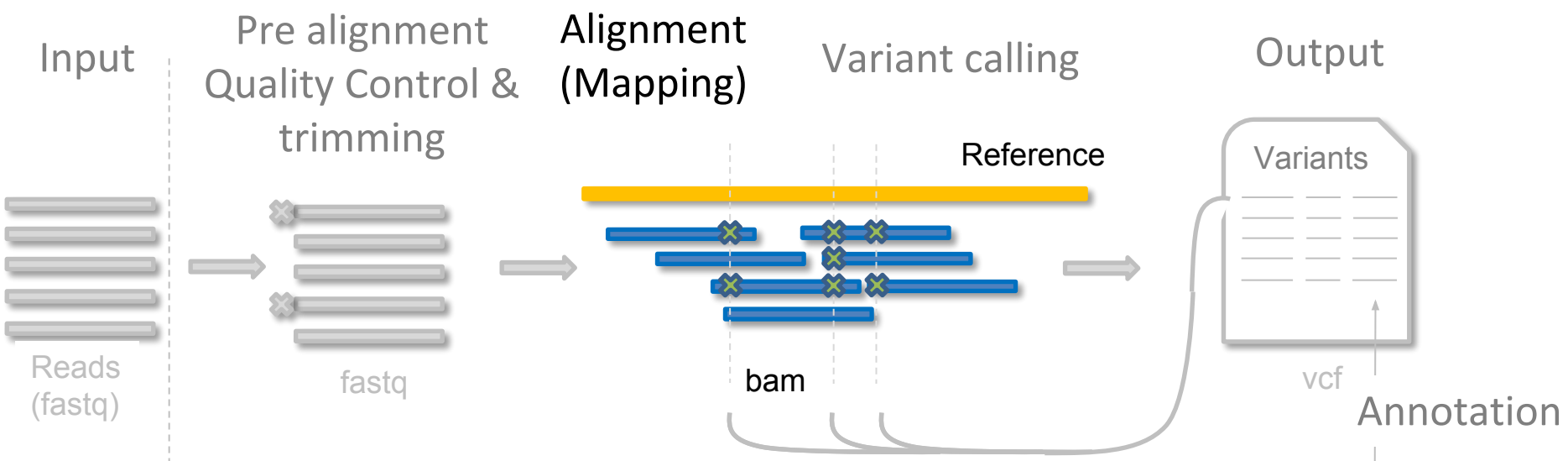
Example of read alignment





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





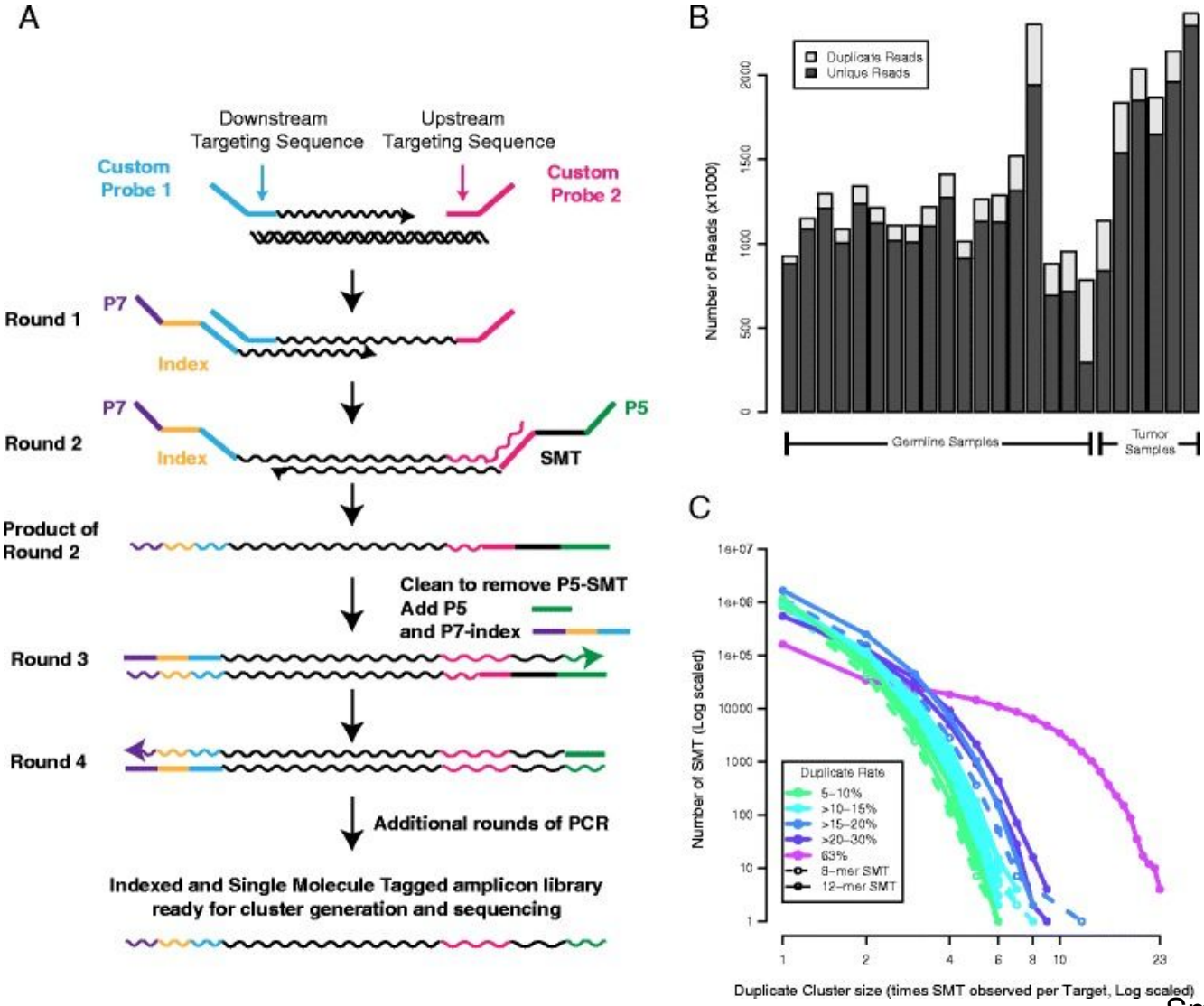
REMOVE PCR DUPLICATES

Each read represents 1 input molecule

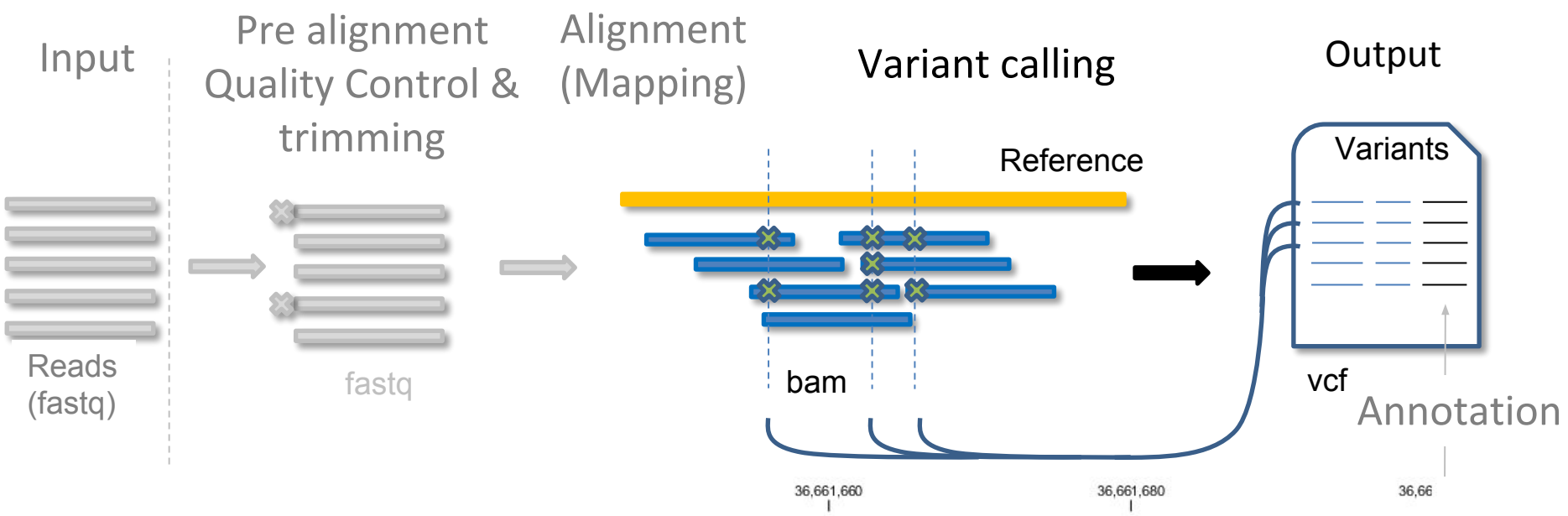
THEORY:

In case of DNA re-sequencing, 1 diploid genome (1 cell)
is represented by 2 reads because of 2 chromosomes
BUT
there is a PCR before the sequencing =>
1 input molecule from 1 cell could be represented
by more reads - PCR duplicates => **Biased variant allele frequency**
(**EXAMPLE...**)
How to solve it?
1) Molecular barcodes (very new method)
2) Identity of start-end positions of read pair (not suitable for amplicons)

Introduction of Molecular barcodes during library preparation



Smith et al. 2014



Mutation types:

Germinal mutations

Somatic mutations

Substitutions

Insertions

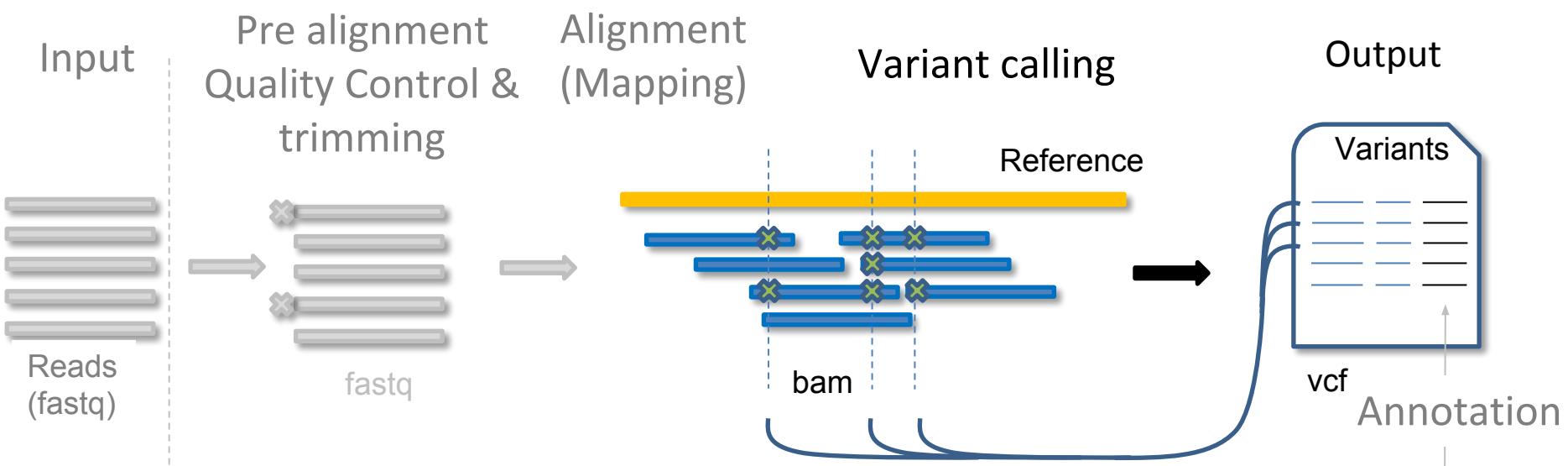
Deletions

Complex variants

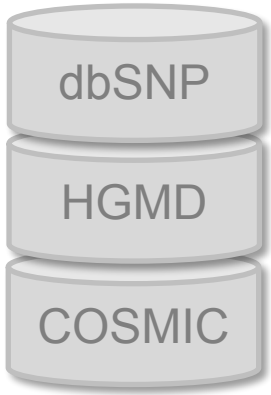
Inversions

Large structural variations (translocations, indels)

Copy number variations



AIM of variant calling - to identify the variant and distinguish from an error
(library preparation, sequencing, alignment)

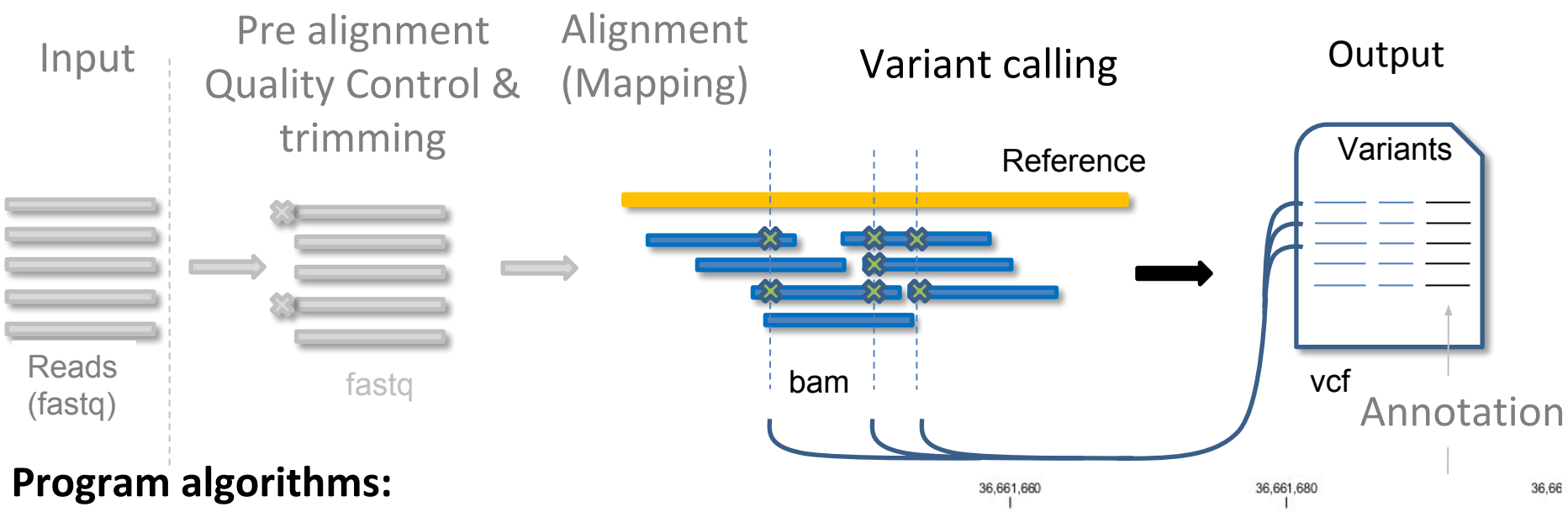


Experimental designs (also depends on types of samples available):

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

Tumor + related normal control
Tumor collected in time

Family (rare diseases, genotyping)



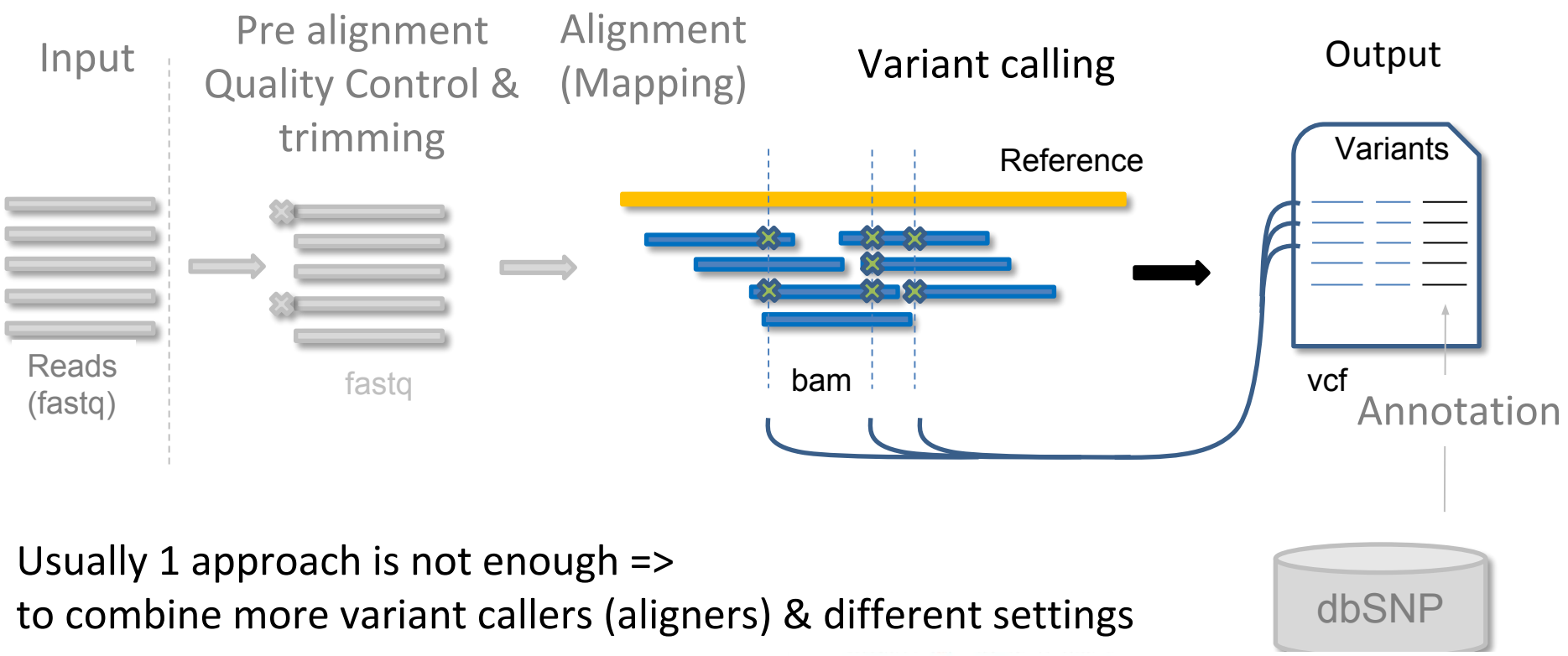
Program algorithms:

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

Giving p-value based on different features…

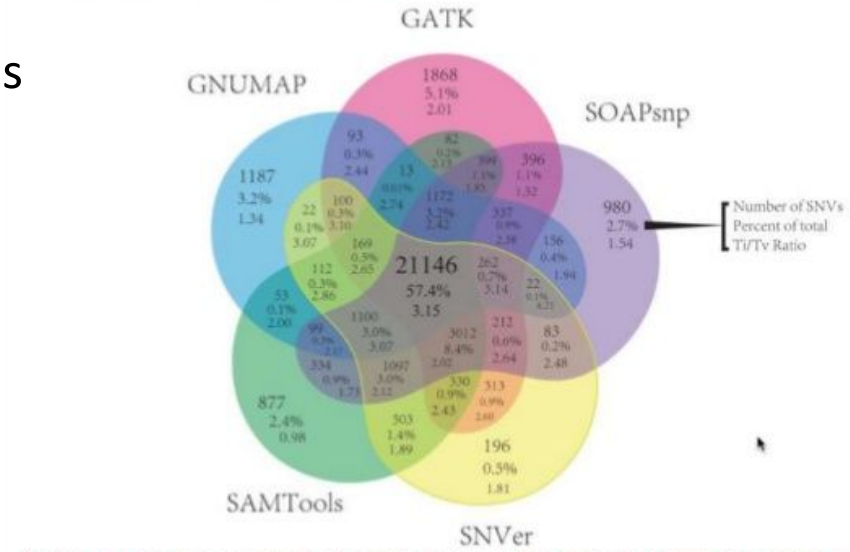
Options for many parameters & filters:

- 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)

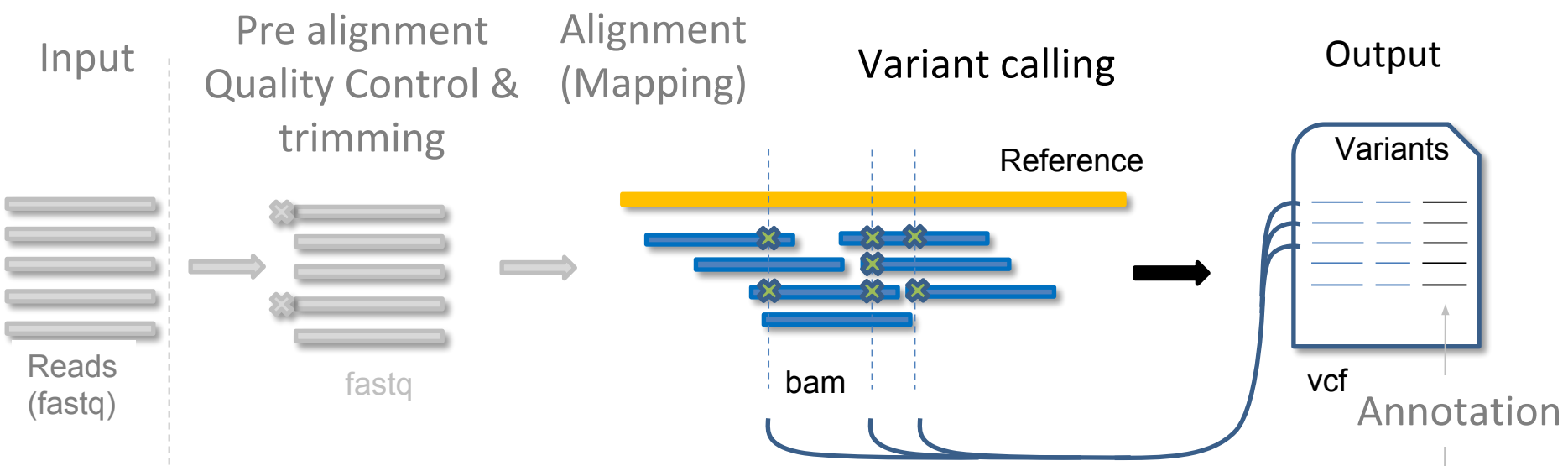


Usually 1 approach is not enough =>
to combine more variant callers (aligners) & different settings

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)

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

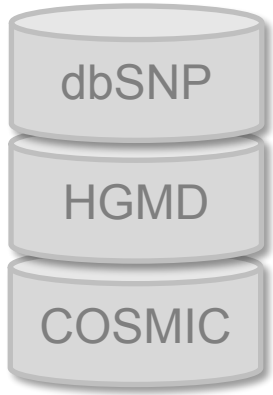
Deletion

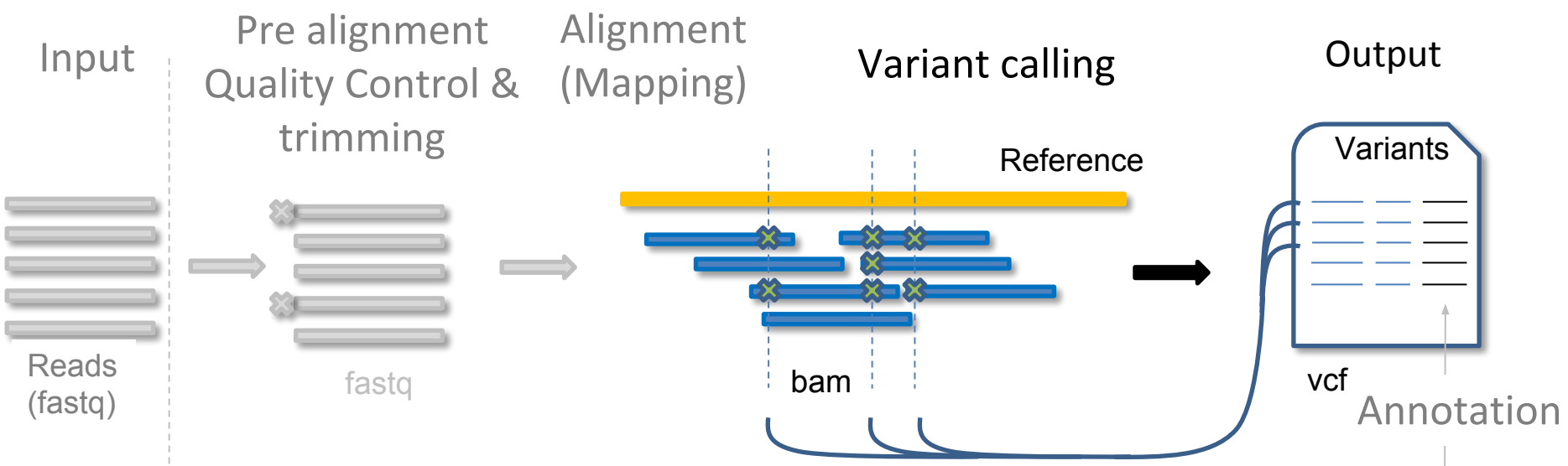
SNP

Large SV

Insertion

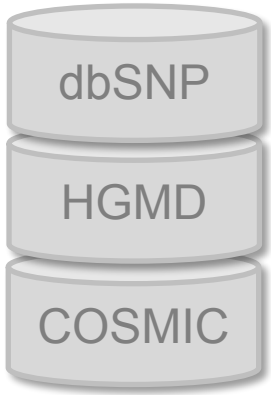
Other event

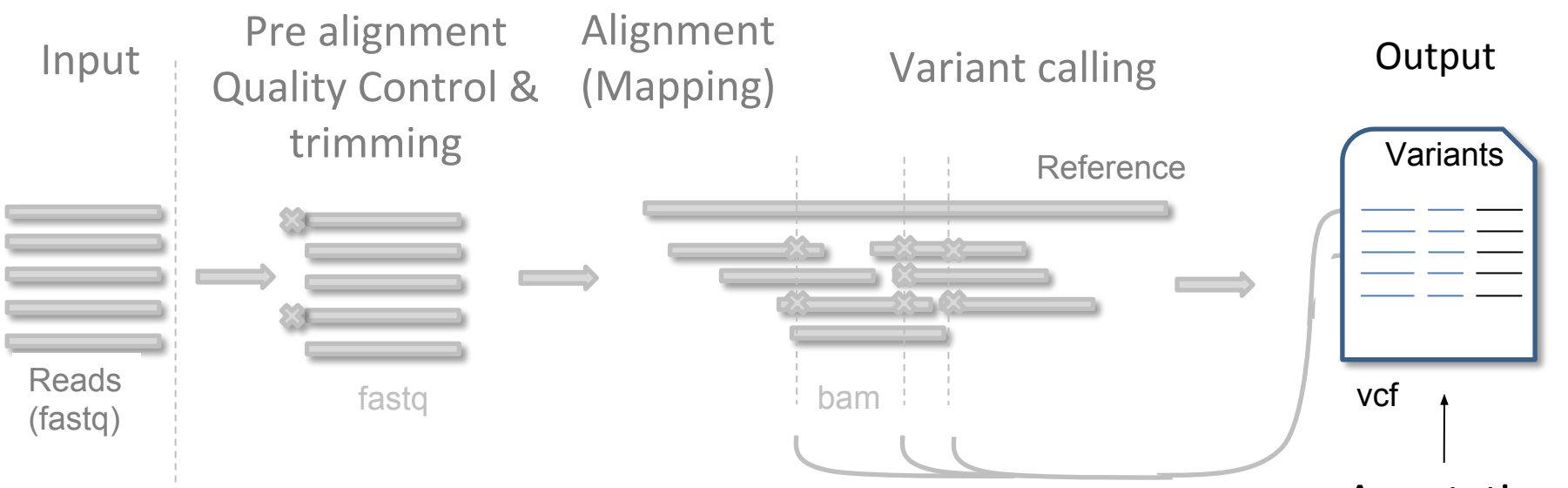




Visualization of variants and alignments (IGV)

Input vcf





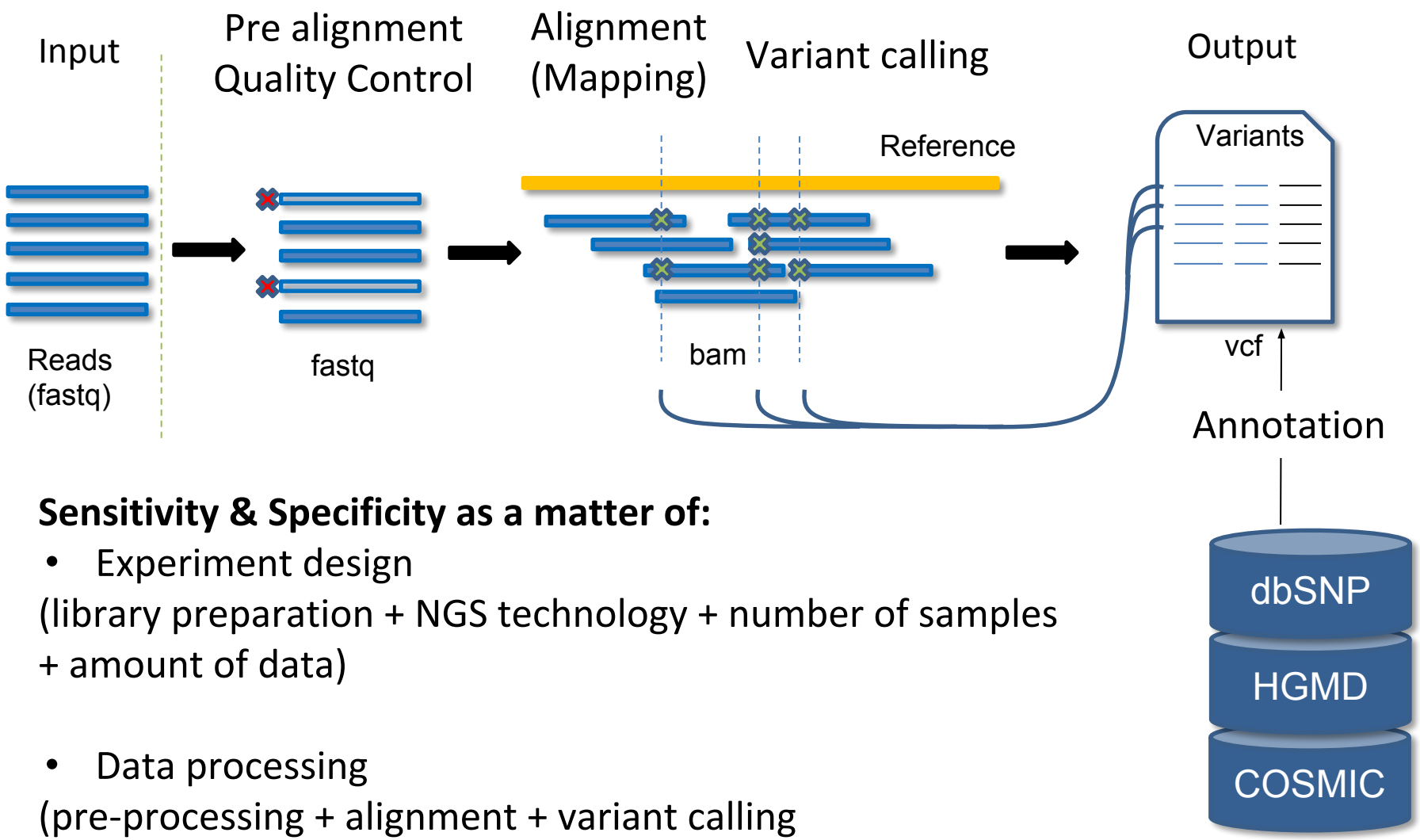
Annotation

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
- 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://www.embo.org/events/events-calendar

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