

Introduction to Variant Analysis

Bioinformatics Course UEB-VHIR
November 2020

Ricardo Gonzalo¹, Mireia Ferrer¹, Àlex Sánchez^{1,2}
Berta Miró¹, Angel Blanco^{1,2}

1 Unitat d'Estadística i Bioinformàtica (UEB) VHIR

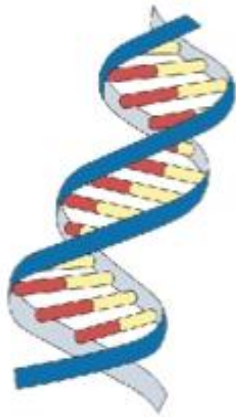
2 Departament de Genètica Microbiologia i Estadística, UB

TABLE OF CONTENTS

1. Introduction to Human genetic variation
2. Application of NGS to the study of genetic variation
3. Steps in NGS data analysis with a focus on variant analysis
4. Challenges in variant analysis
5. Hands on with exome variant analysis → **Session 6**

Introduction to Human Genetic Variation

- All human beings are 99.9% identical in their genetic makeup. Differences in the remaining 0.1% hold important clues about the causes of diseases.
- Gaining a better understanding of the interactions between genes and the environment by means of genomics is helping researchers find better ways to improve health and prevent disease



Genotypes are the genetic make-up of an individual⁵.

Phenotypes are the physical traits and characteristics of an individual and are influenced by their genotype and the environment⁶.

Introduction to Variant Analysis

- Genetic differences (variants) can occur between
 - Individuals of a population
 - Strains of an organism
 - Healthy and diseased tissue
- Why study genetic variation?
 - understand the natural function of affected genes
 - model human migration
 - provide mechanistic insight into disease processes
 - predict disease outcomes, response to treatments
 - forensic applications, paternity tests...

Introduction to Human Genetic Variation

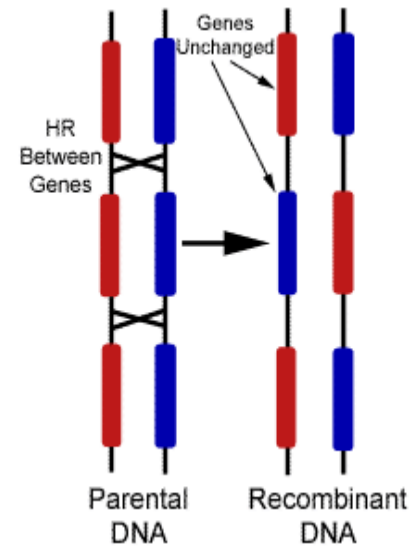
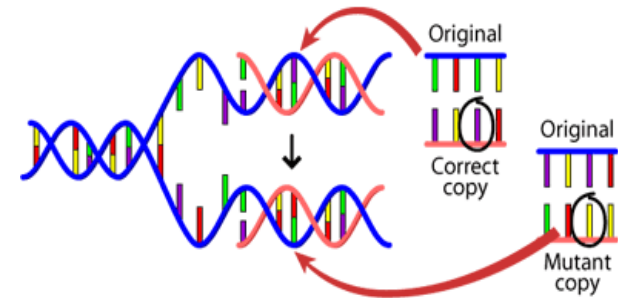
- The term **variant** is used to refer *to a specific region of the genome which differs between two genomes*.

Reference: ...GTGCGTAGACTGCTAGATCGAAGA...
Sample: ...GTGCGTAGACTG**A**TAGATCGAAGA...

- The term **reference allele** refers to the base that is found in the reference genome
 - The reference is just somebody's genome not always the major allele.
 - The **alternative allele** refers to any base, other than the reference
- Alleles at variants close together on the same chromosome tend to occur together more often than is expected by chance. These blocks of alleles are called **haplotypes**.

Introduction to Human Genetic Variation

- Two major sources of genetic variation:
 - **Mutations:** occur when there is an error during DNA replication that is not corrected by DNA repair enzymes
 - **Recombination:** is a process by which pieces of DNA are broken and exchanged to produce new combinations
- They can be hereditary or not, depending if they occur in:
 - **Germ cells** (eg. sperm, egg cells): can be inherited from one individual to another and so affect population dynamics, and ultimately evolution
 - **Somatic cells** (all other): can affect the individual, but they are not passed on to offspring.



Introduction to Human Genetic Variation

- Genetic variation is commonly divided in three types:
 - 1) **Single Nucleotide Variants (among which SNPs)**
 - 2) **Insertion or deletions (“indels”)**
 - 3) **Structural variation**
 - Copy number variation
 - Chromosomal rearrangement events
- All forms of variations are related with disease but we will mostly focus on SNPs

Single Nucleotide Variants (commonly SNPs)

- SNVs result from a substitution of a single base-pair
- They are the most common type of genetic variation among people
- They occur almost once in every 1,000 nucleotides on average, which means there are roughly 4 to 5 million SNVs in a person's genome
- These variations may be unique or occur in many individuals

Introduction to Human Genetic Variation

Single Nucleotide Variants (commonly SNPs)

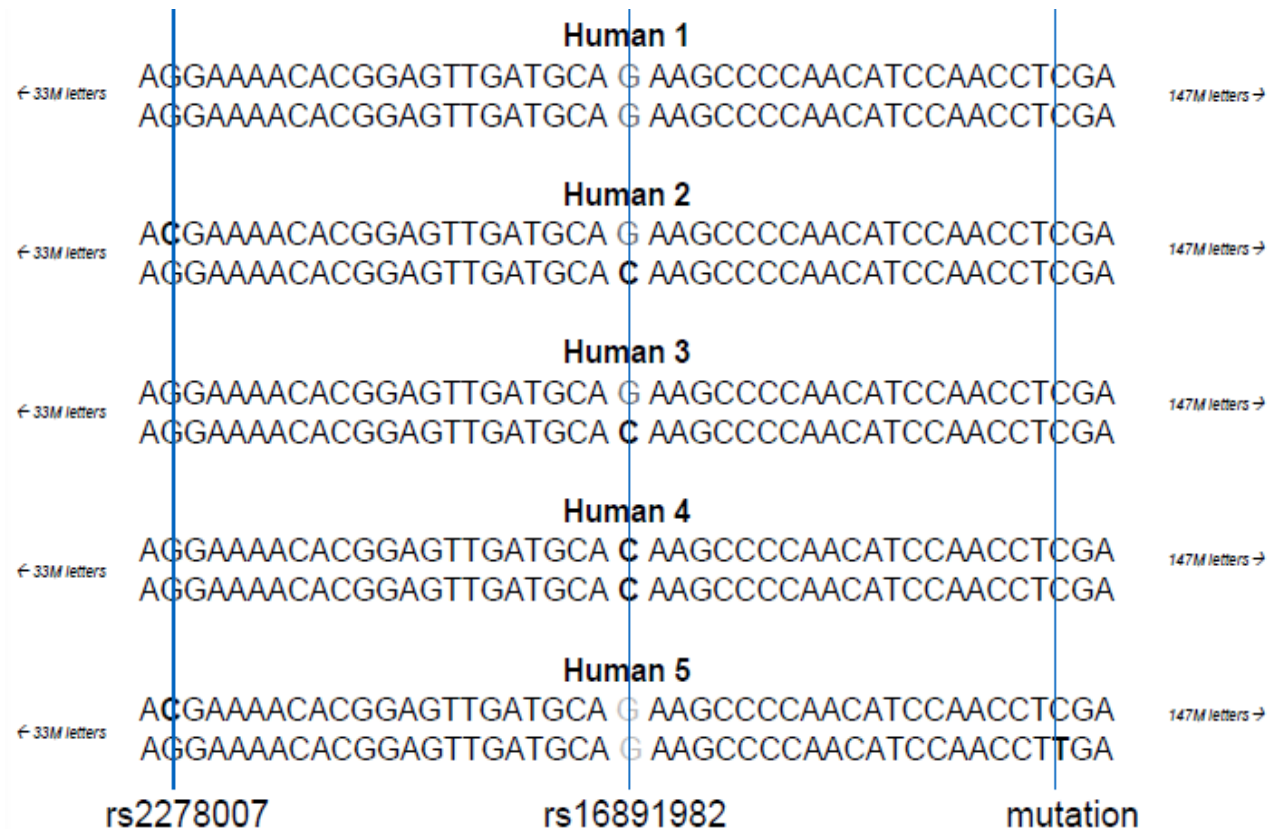
- SNVs result from a substitution of a single base-pair
- They are generally classified as single nucleotide polymorphisms (SNPs) if they are present at a moderately high frequency in the population (>1%)



Introduction to Human Genetic Variation

Single Nucleotide Variants (commonly SNPs)

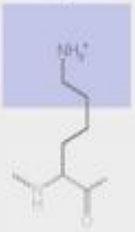
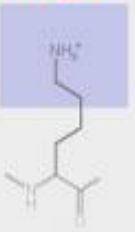
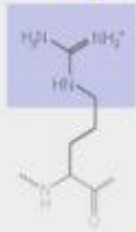
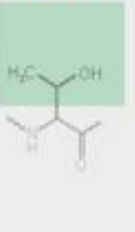
- SNVs result from a substitution of a single base-pair
- They are generally classified as single nucleotide polymorphisms (SNPs) if they are present at a moderately high frequency in the population (>1%)



Introduction to Human Genetic Variation

Single Nucleotide Variants (commonly SNPs)

- Most commonly, these variations are found in the DNA *between* genes.
- When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function.
 - If a variant falls within a coding region, it can be categorised based on how it would affect the codon it falls within

	Point mutations				
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr
					

Introduction to Human Genetic Variation

Insertions/deletions (indels)

- Insertion or deletion of a single stretch of DNA sequence
- Can range from 2-100s of bp

Reference	ACTGACGCATGCATCATGCATGC	
Insertion	ACTGACGCATG GTA CATCATGCATGC	} Indel
Deletion	ACTGACG -- TGCATCATGCATGC	

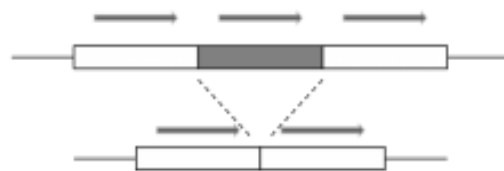
Indels may result in
frameshift mutations

Normal							
mRNA	AUG	GGG	GCC	AAA	AGU	UAG	UUUG...
polypeptide	Met	Gly	Ala	Lys	Ser	Stop	
Insertion					+U		
					↓		
mRNA	AUG	GGC	GCC	AAA	UAG	UUAGUUUG...	
polypeptide	Met	Gly	Ala	Lys	Stop		
Deletion							
			-G				
			↓				
mRNA	AUG	GGC	CCA	AAA	GUU	AGU	UUG
polypeptide	Met	Gly	Pro	Lys	Val	Ser	Leu
			Random				

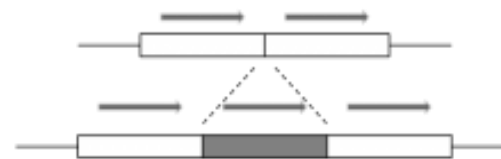
Introduction to Human Genetic Variation

Structural variation

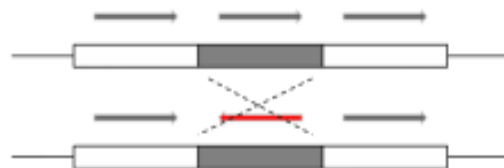
- Genetic variation that occurs over a “larger” DNA sequence.



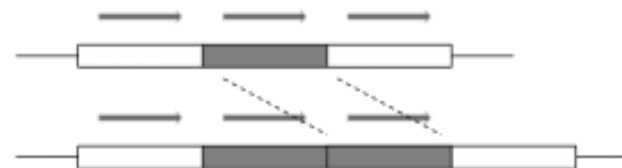
Deletion



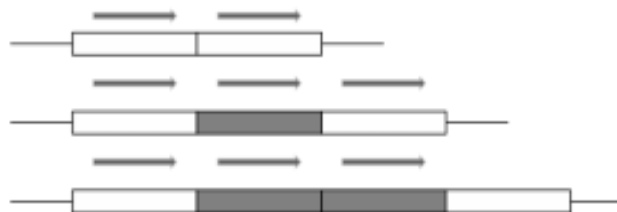
Insertion



Inversion



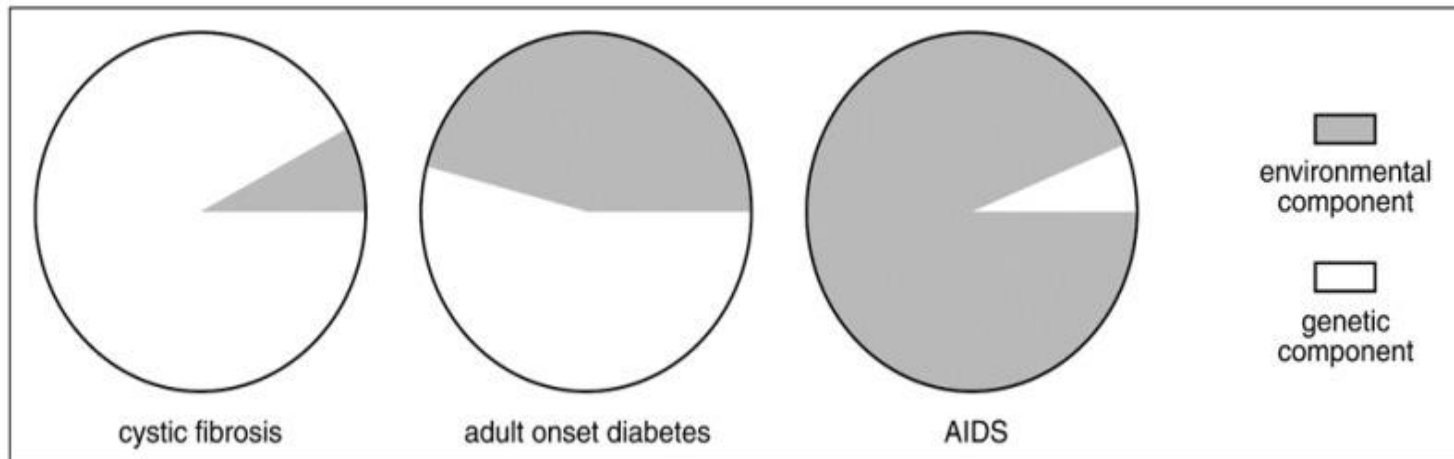
Duplication



Copy Number Variation

Introduction to Human Genetic Variation

- Overall genetic variation may produce distinct effects ranging from innocuous to lethal.
- In many cases the effect will be changes in one or more proteins that will possibly affect the individual's health.



Virtually all human diseases, except perhaps trauma, have a genetic component.

Introduction to Human Genetic Variation

- Given the importance and potential effects of variants a relevant aspect in biological studies becomes the **identification** and **analysis** of *variants associated with a specific trait* of a population.
- Bioinformatics is key to each stage of this process and is essential for **handling genome-scale data**. It also provides us with a **standardised framework to describe variants**.

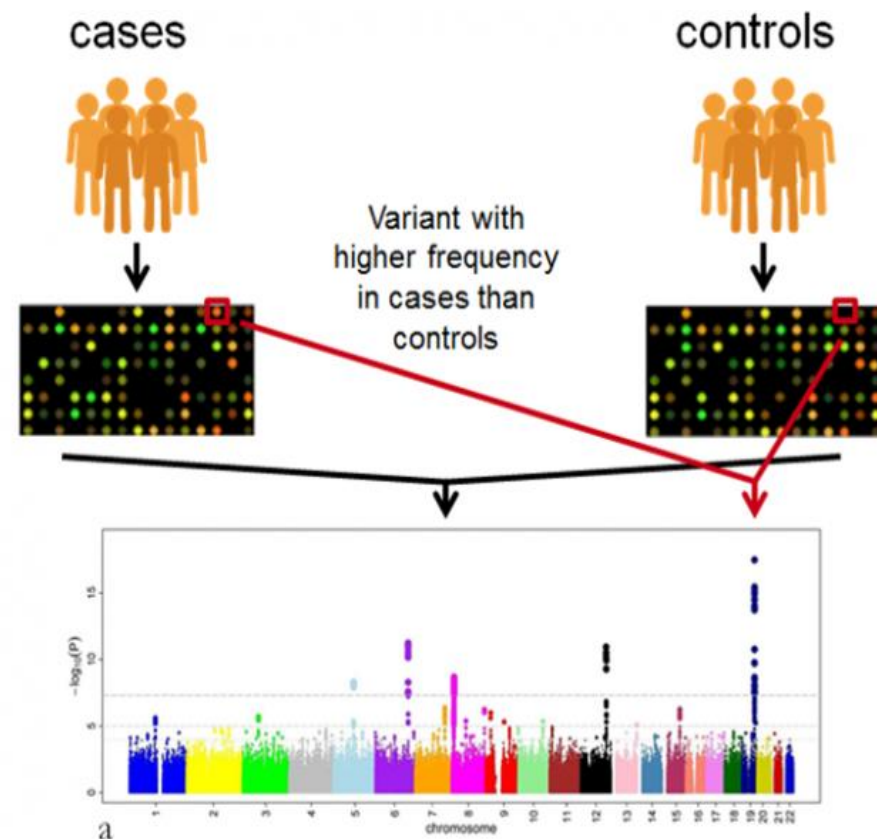
Introduction to Human Genetic Variation

Types of studies on genetic variation

• There are many ways in which you can study genetic variation however most studies can be loosely classified as:

– Genome wide association studies (GWAS):

- associate variants with a phenotype, trait or disease based on the fact that a variant leading to a phenotype is found at a higher frequency in cases than control



Types of studies on genetic variation

- There are many ways in which you can study genetic variation however most studies can be loosely classified as:
 - **Genome wide association studies (GWAS)**
 - **Studies on Functional consequences of variants:** aim to understand the molecular mechanisms and pathways that link genotype to phenotype
 - **Population genetics:** study of variation within populations of individuals, and the forces which shape it. Involves the examination and modelling of changes in the frequencies of genes and alleles in populations over space and time.

Bioinformatics resources

- Bioinformatics resources for studying variants:

- **Databases of variant annotations**

dbSNP: contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

1000 Genomes Project: the largest public catalogue of human variation and genotype data. The goal was to find most genetic variants with frequencies of at least 1% in the populations studied.

dbGAP: database of Genotypes and Phenotypes

ClinVar: archive of reports of relationships among medically important variants and phenotypes

OMIM: a database of known mendelian disorders

COSMIC: Catalogue of somatic mutations in cancer

Introduction to Human Genetic Variation

Bioinformatics resources

Variants may have identifiers from multiple databases.

Identifier type	Example	Description
ssID	ss335	Submitted SNP ID assigned by dbSNP or EVA
rsID	rs334	Reference SNP ID assigned by dbSNP or EVA. ssIDs of the same variant type that colocalise are combined to give an rsID for that locus.
HGVS*	ENST00000366667.4:c.803T>C	Expresses the location of the variant in terms of a transcript or protein.
COSMIC ID	COSM1290	ID assigned by COSMIC for somatic variants.
HGMD	CD830010	ID assigned by HGMD to variants known to be associated with human inherited diseases.
ClinVar	RCV000016573	ID assigned to dbSNP or dbVar/DGVa annotated variants, linking them to human health.
UniProt	VAR_010085	ID assigned by UniProt for reviewed human

Bioinformatics resources

- Bioinformatics resources for studying variants:
 - Databases of variant annotations
 - **Software that evaluate variant consequence**
 - Algorithms such as [SIFT](#) and [PolyPhen](#) or [Ensembl's Variant Effect Predictor \(VEP\)](#) estimate how likely this amino acid change is to affect protein function.
 - These estimates are based on how well conserved the protein is, the chemical difference between the amino acids, and the 3D structure of the protein (PolyPhen only).
 - Both provide a score out of one (0 is the most severe for SIFT, whereas 1 is the most severe for PolyPhen) along with a qualitative prediction.
 - These are predictions only, not experimental validations of the effect. Known or predicted functional consequences for variants of a specific protein are summarised in [UniProt](#).

Introduction to Human Genetic Variation

Example

- Searching for the *rs334* variant in [Ensembl](#)

Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

Human (GRCh38.p13) ▾

Location: 11:5,226,502-5,227,502 Variant: rs334

Variant displays

- Explore this variant
 - Genomic context
 - Genes and regulation
 - Flanking sequence
 - Population genetics
 - Phenotype data
 - Sample genotypes
 - Linkage disequilibrium
 - Phylogenetic context
 - Citations
 - 3D Protein model
- Configure this page
- Custom tracks
- Export data
- Share this page
- Bookmark this page

rs334 SNP

Most severe consequence: missense variant | [See all predicted consequences](#)

Alleles: **T/A/C/G** | Ancestral: T | MAF: 0.03 (A) | Highest population MAF: 0.14

Change tolerance: CADD: A:15.65, C:15.88, G:9.350 | GERP: -1.22

Location: [Chromosome 11:5227002](#) (forward strand) | VCF: 11 5227002 rs334 T A,C,G

Co-located variants: HGMD-PUBLIC CD830010, CM097155, CM880038; dbSNP rs63749819 (T/-)

Evidence status:

Clinical significance:

HGVS names: This variant has 42 HGVS names - [Show](#)

Synonyms: This variant has 22 synonyms - [Show](#)

Genotyping chips: This variant has assays on: Illumina_HumanOmni5, Illumina_ExomeChip

Original source: Variants (including SNPs and indels) imported from dbSNP (release 153) | [View in dbSNP](#)

About this variant: This variant overlaps [5 transcripts](#), [1 regulatory feature](#), has [2504 sample genotypes](#), is associated with [sickle cell anemia](#); one report concludes that everyone with this mutation

Description from SNPedia

Explore this variant

- Genomic context
- Genes and regulation
- Flanking sequence
- Population genetics
- Phenotype data
- Sample genotypes
- Linkage disequilibrium
- Phylogenetic context
- Citations
- 3D Protein model

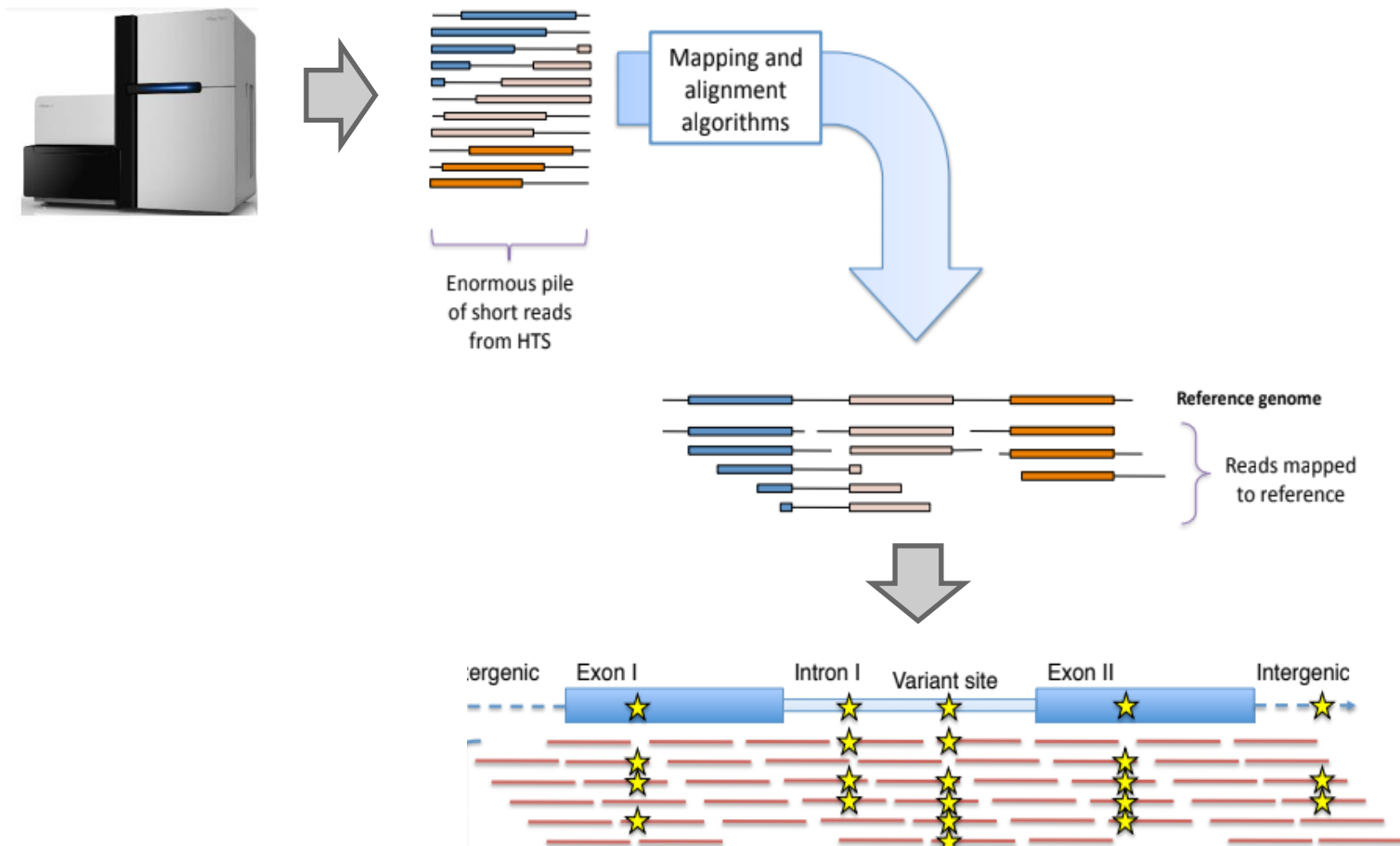
Example

- Searching for the *rs334* variant in [Ensembl](#)
 - Functional consequence: *rs334* is a missense variant in *HBB*, a haemoglobin subunit.
 - Phenotype association studies: It is associated with sickle cell anaemia and malaria resistance
 - Population genetics: the phenotype-associated A allele is mostly found in African populations

Application of NGS to the study of genetic variation

NGS for studying genetic variation

- Next generation sequencing (NGS) technology has had a transformatory effect upon population-level studies linking genetic variation to gene function.
- NGS allows for nucleotide variation profiling and large-scale discovery of genetic markers



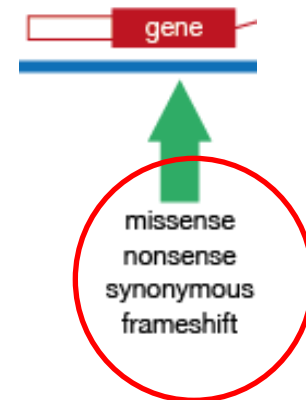
NGS for studying genetic variation

- Can be performed at different levels
 - **Whole Genome Sequencing (WGS)**
 - **Whole Exome Sequencing (WES)**
 - **Targeted subgenomic sequencing**
 - identify trait loci by re-sequencing candidate genes in a large number of patients and controls
 - likely to be supplanted by whole-exome sequencing

NGS for studying genetic variation

Exome Sequencing (WES)

- Exome sequencing or Whole Exome Sequencing (WES) is **the sequencing of all the expressed protein-coding genes in a genome** (known as the exome)
 - <2% of the human genome
- The goal of this approach is to identify genetic variants that alter protein sequences

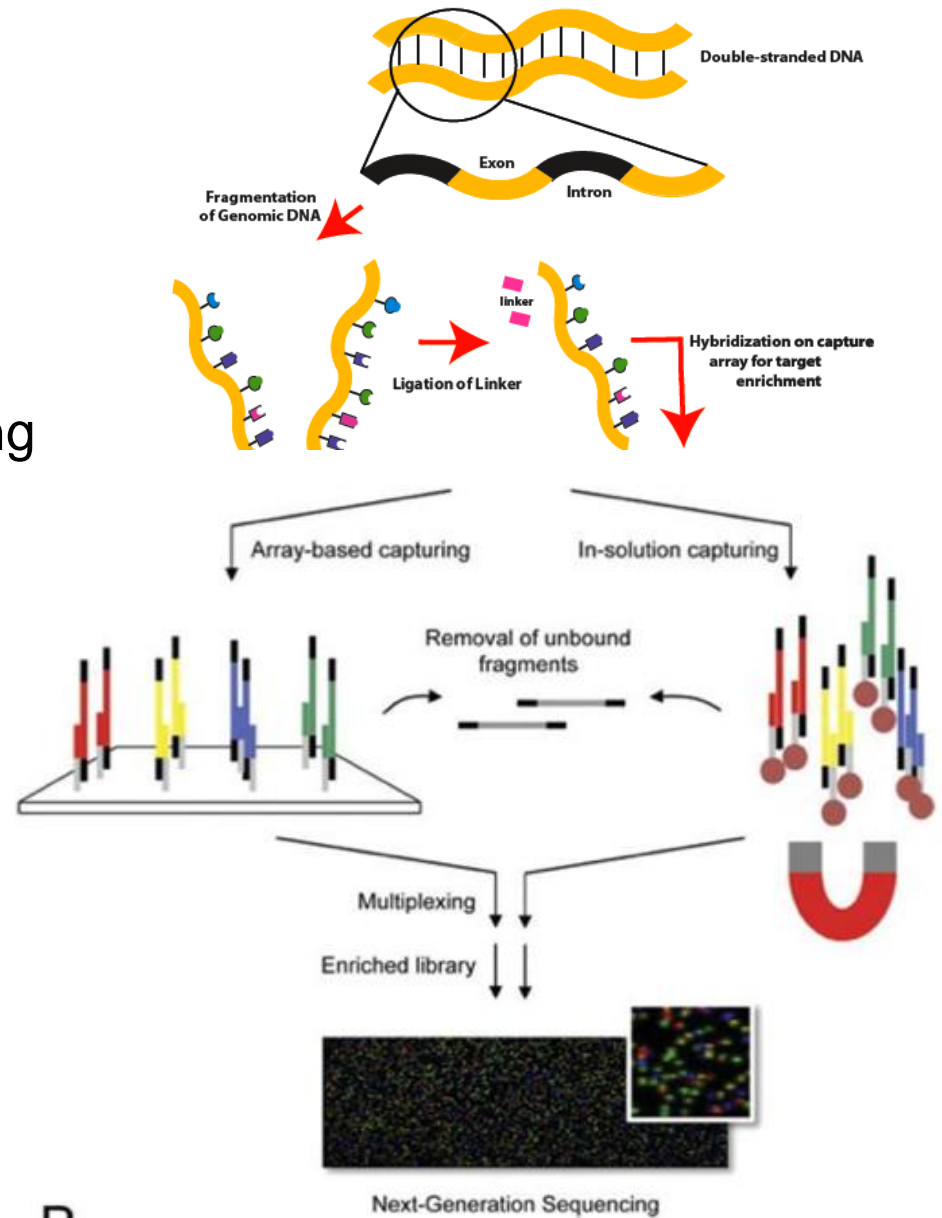


NGS for studying genetic variation

Exome Sequencing (WES)

• It consists of two steps:

1. Target enrichment: select only exons
2. Sequence the exonic DNA using any high-throughput DNA sequencing technology



NGS for studying genetic variation

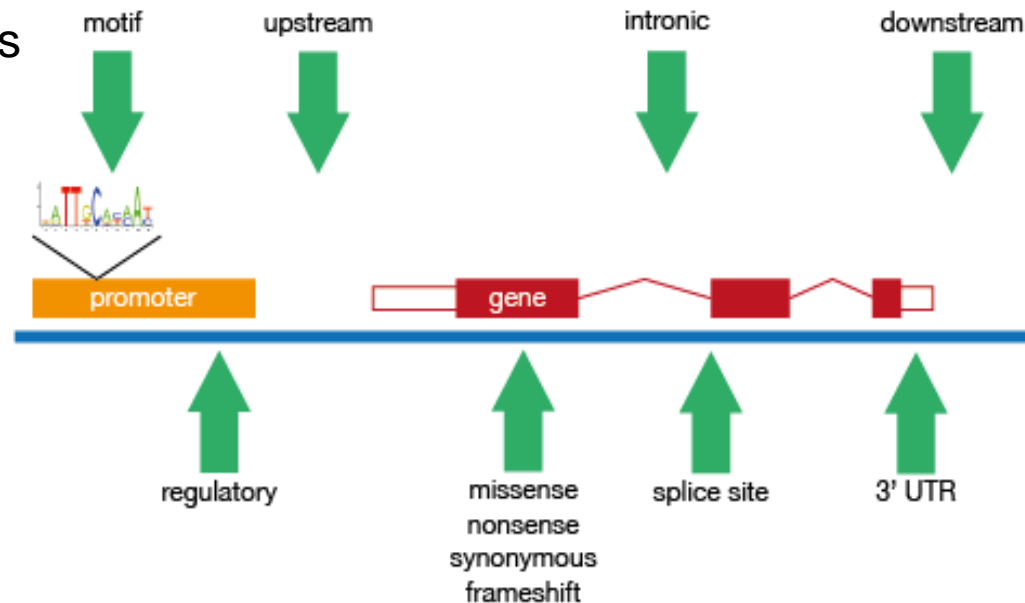
WES vs WGS

- Why WES instead of WGS (Whole Genome Sequencing)?
 - less size
 - Cheaper
 - Allows higher depth coverage for more accurate variant calling
 - **Most known variants related to disease are in coding regions**

NGS for studying genetic variation

WES vs WGS

- Limitations of WES:
 - may not target 100% of the genes in the human genome (approximately 97% of exons are targeted)
 - limited in detecting the following types of mutations
 - Structural variants
 - Triplet repeat disorders
 - Other copy number variants
 - Epigenetic factors
 - Introns
 - Regulatory sequences
 - Gene-gene interactions

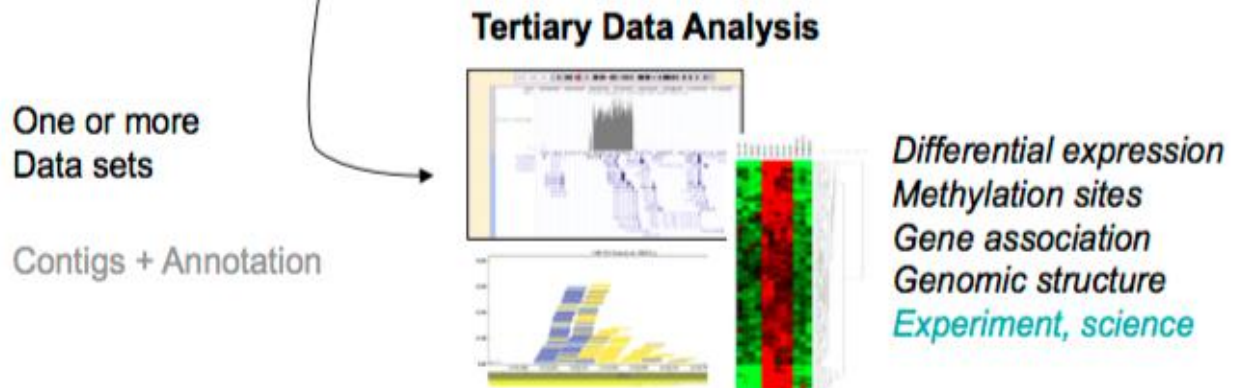
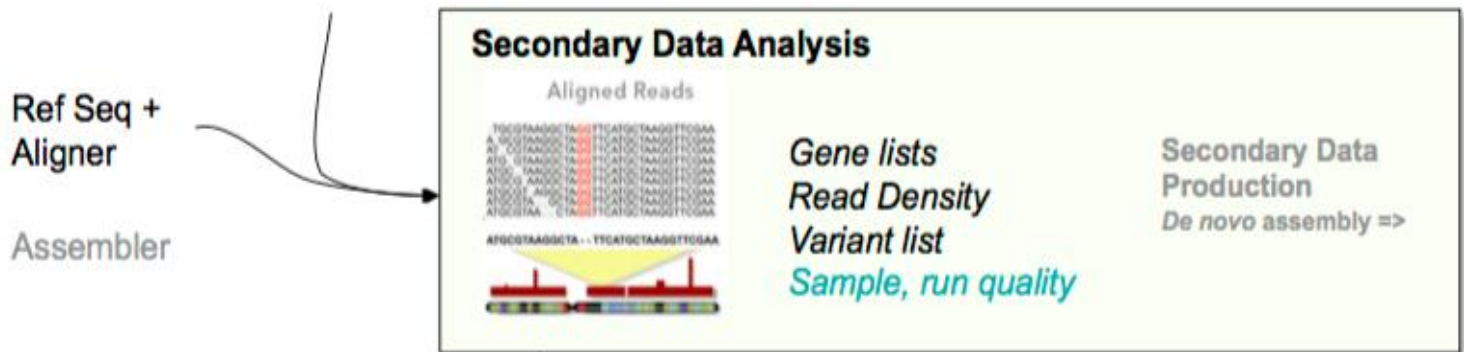


Steps in NGS analysis with a focus on variant analysis

Steps in NGS analysis

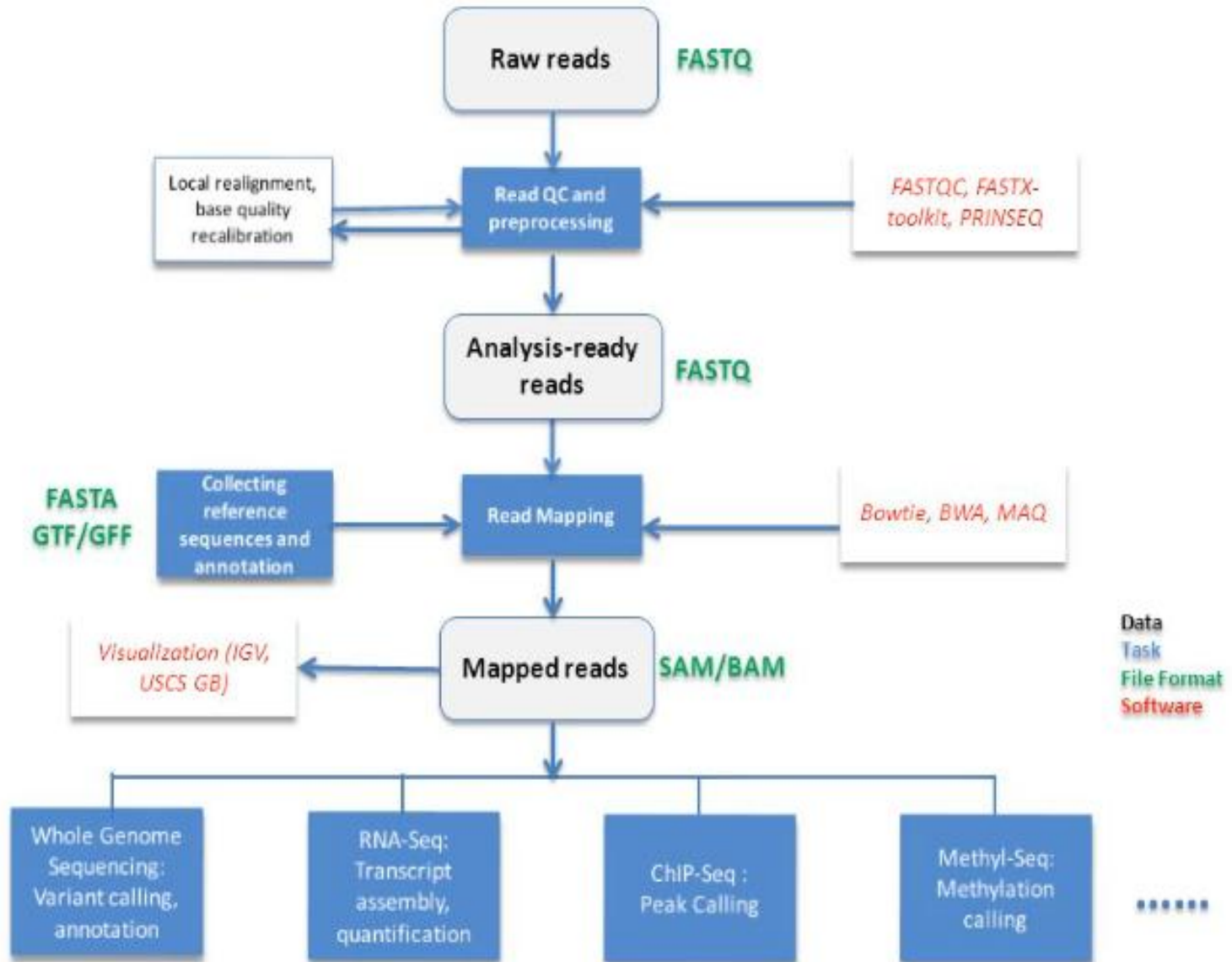
- NGS data is analyzed in three stages

Primary Data Analysis - Images to bases



Steps in NGS analysis

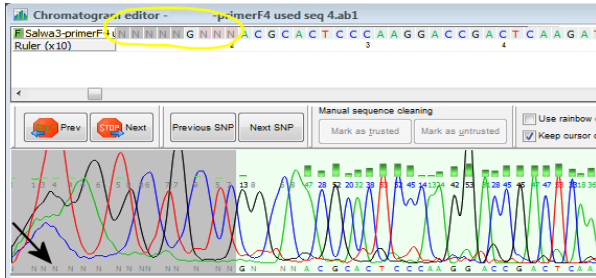
- We will have different data (file) formats and tools for each step



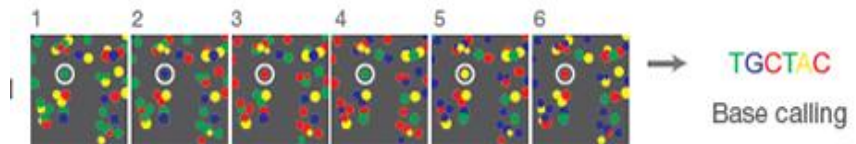
Steps in NGS analysis

- **Base calling: obtaining the raw read sequences (FASTQ files)**

Sanger



Illumina (NGS)



- Base calling accuracy often measured by the Phred Quality Score (Q score) which assesses the accuracy of a sequencing platform.
- It indicates the probability that a given base is called incorrectly by the sequencer.

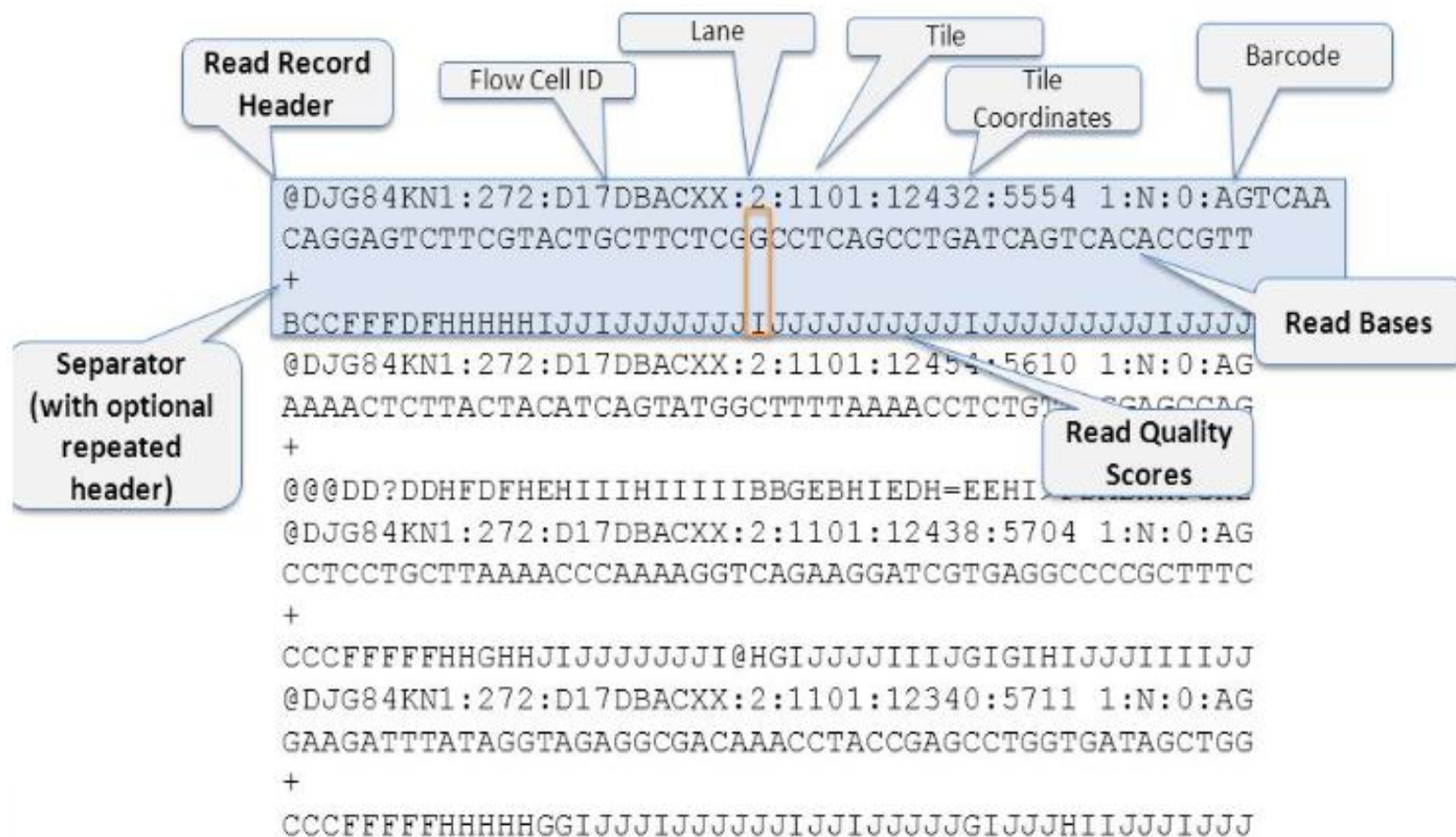
$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

Phred quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%

- Ambiguous positions with Phred scores ≤ 20 are labeled with N.
- To assign each base a unique score identifier (instead of numbers of varying character length), Phred scores are typically represented as ASCII characters.

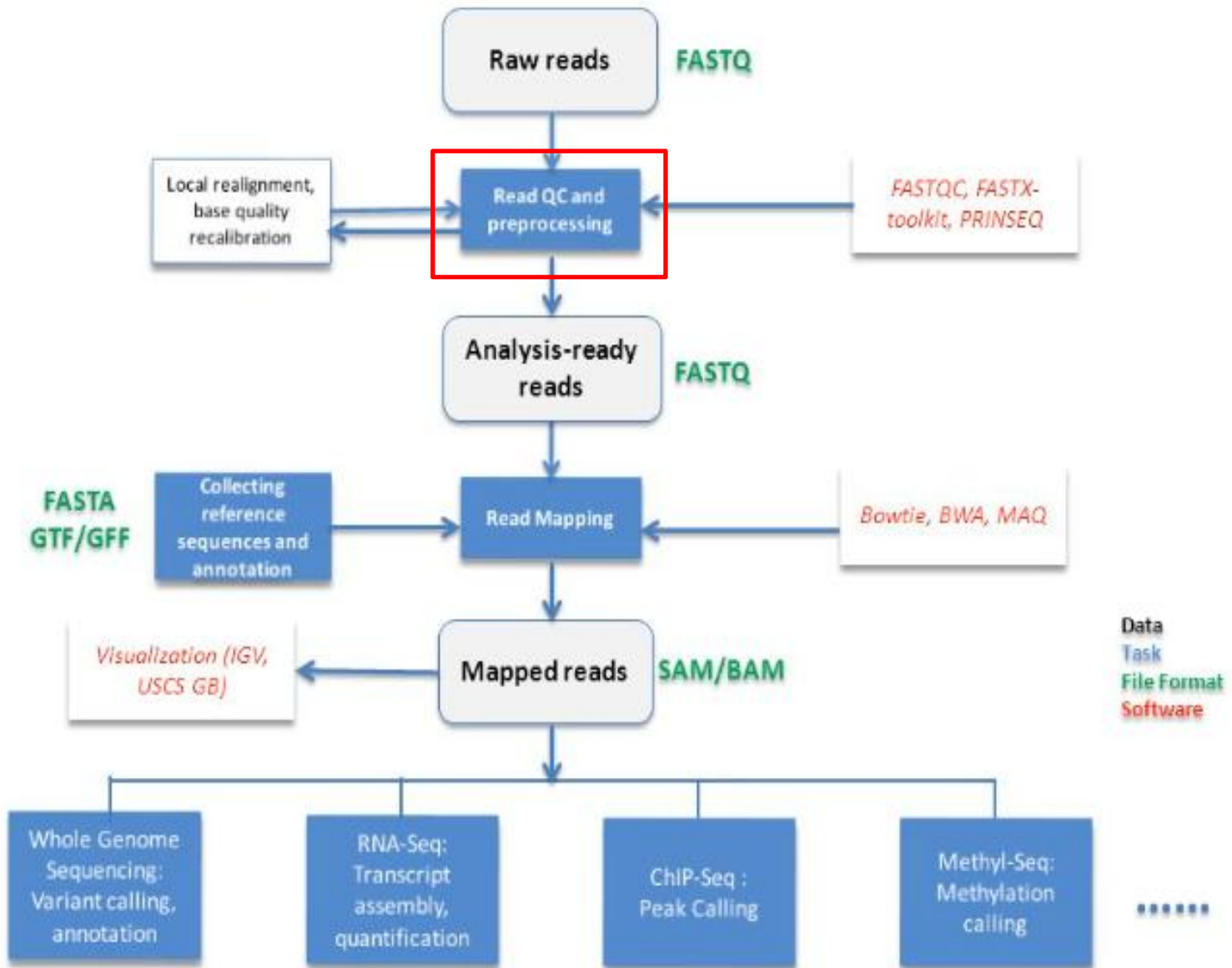
Steps in NGS analysis

FASTQ format = DNA sequence data + Phred quality scores of each base



NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads.

Steps in NGS analysis

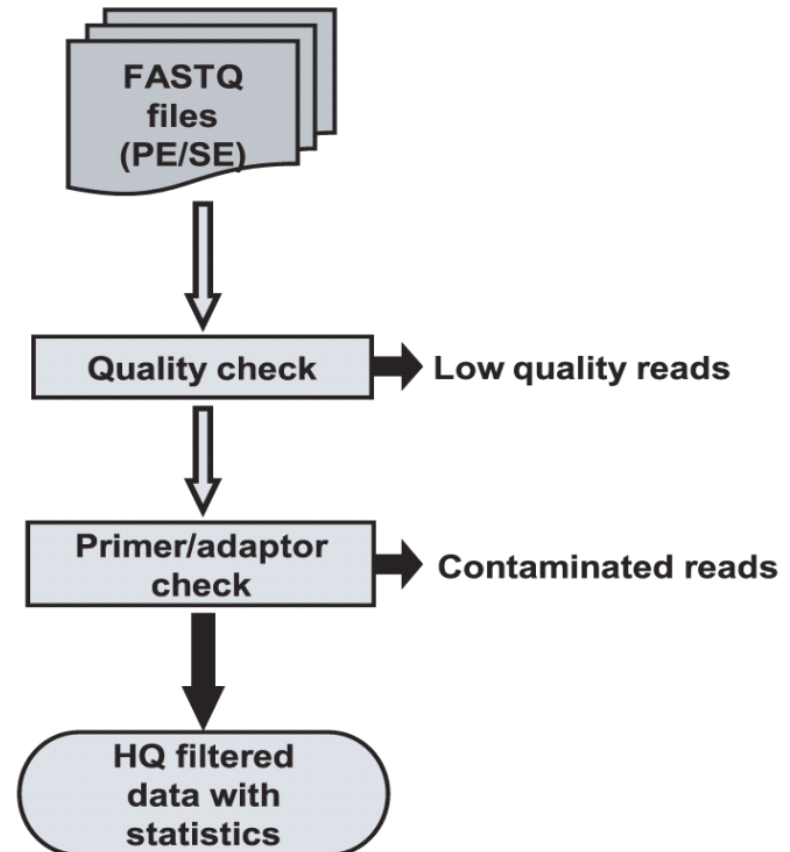


Steps in NGS analysis

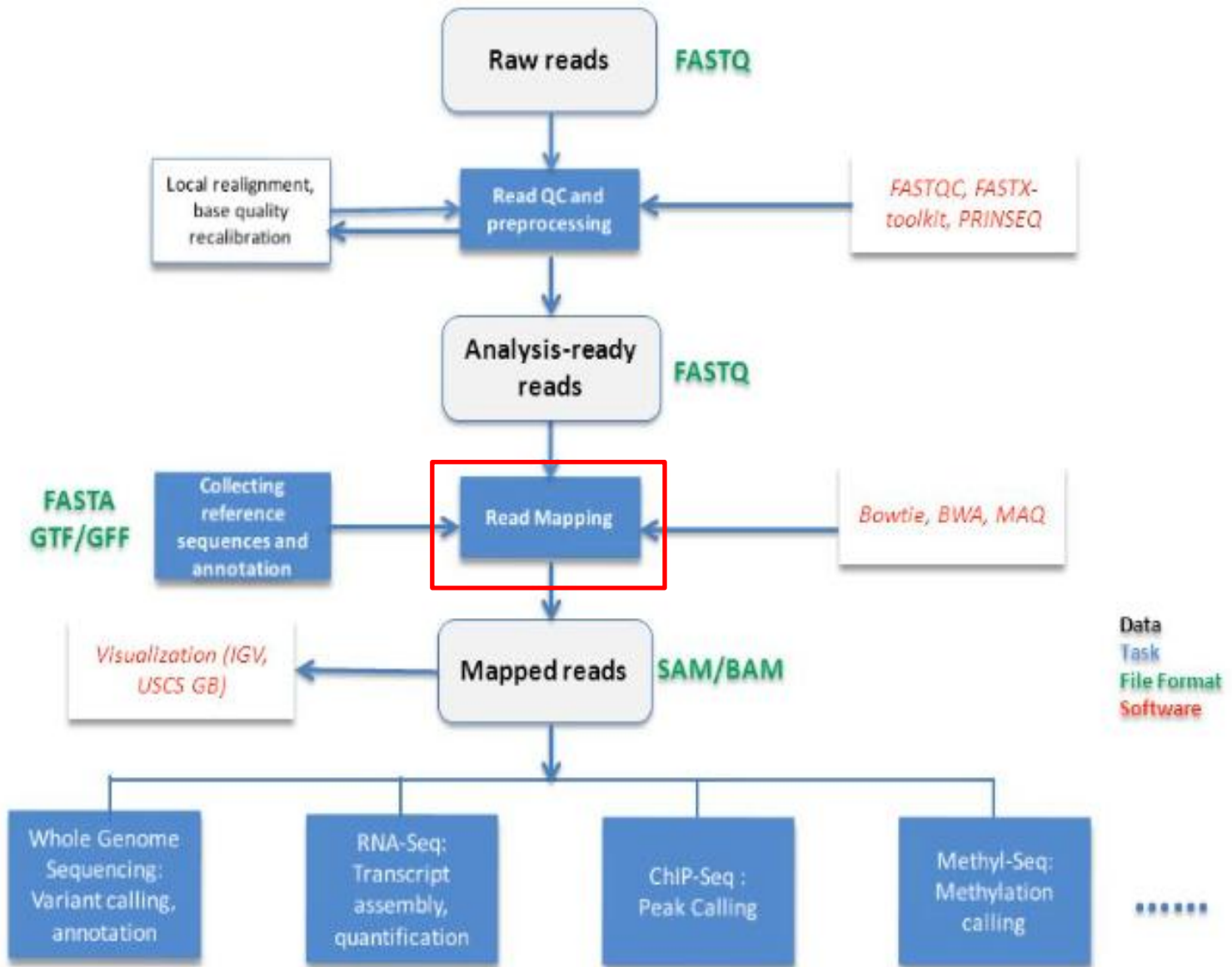
Quality Control

- Quality Control analysis of sequence data is extremely important for meaningful downstream analysis

- To analyze problems in quality scores/ statistics of sequencing data
- To check whether further analysis with sequence is possible
- To remove redundancy (filtering)
- To remove low quality reads from analysis
- To remove adapter contamination

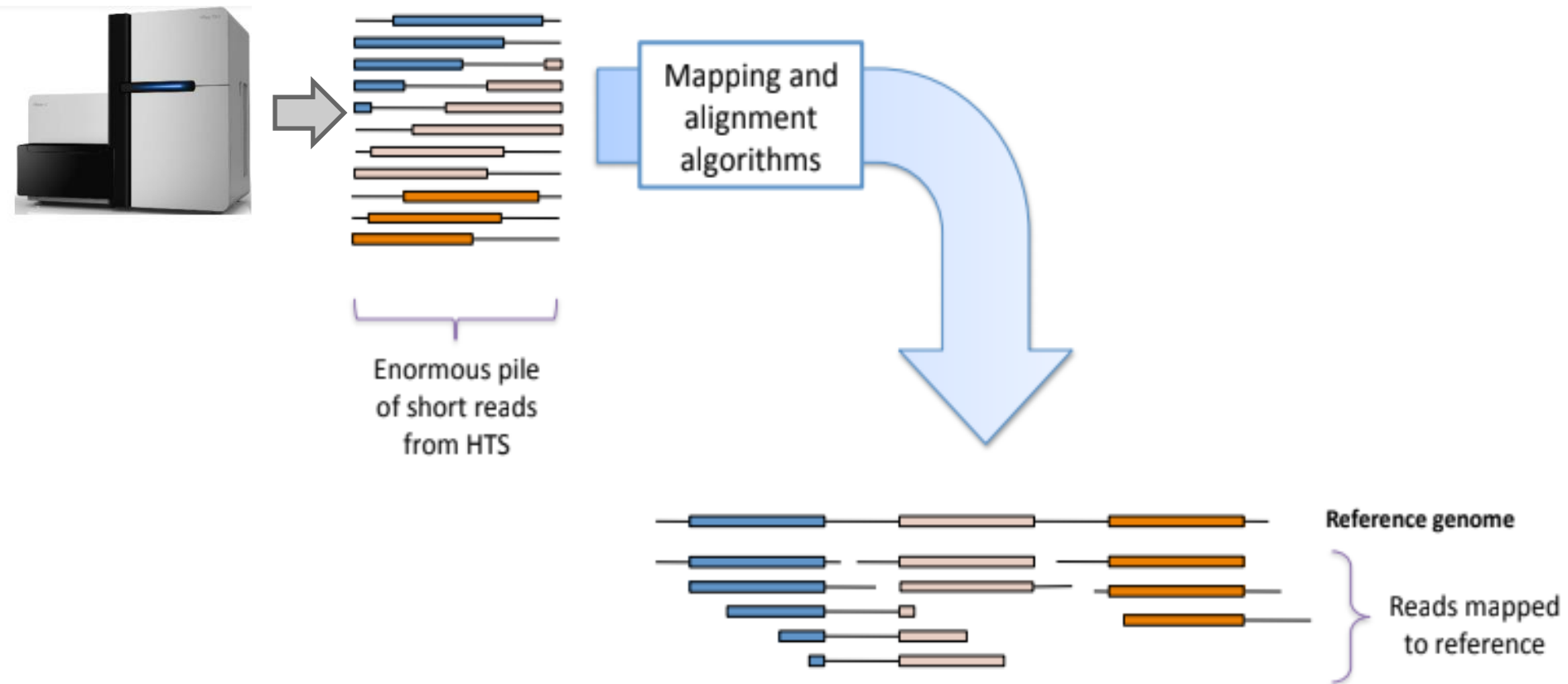


Steps in NGS analysis



Steps in NGS analysis

Mapping reads to the genome



Steps in NGS analysis

Mapping reads to the genome

- Determine position of short read on the reference genome

Reference:	. . . A A - C G C C T T . . .	= match
-	: - :	: = mismatch
Read:	A G G G G C C T T	- = gap

Steps in NGS analysis

Mapping reads to the genome

Challenging!

- There is ambiguity mapping a read with a mismatch versus a deletion

location 1 (mismatch)

... TTT **AGAATGAGCCGAG** TTCGCGCGCGGGT **AGAAT-AGCCGAG** TT ...

||||| |||||
AGAATTAGCCGAG

13 bp read

location 2 (deletion)

genomic DNA

||||| |||||
AGAATTAGCCGAG

13 bp read

Steps in NGS analysis

Mapping reads to the genome

Challenging!

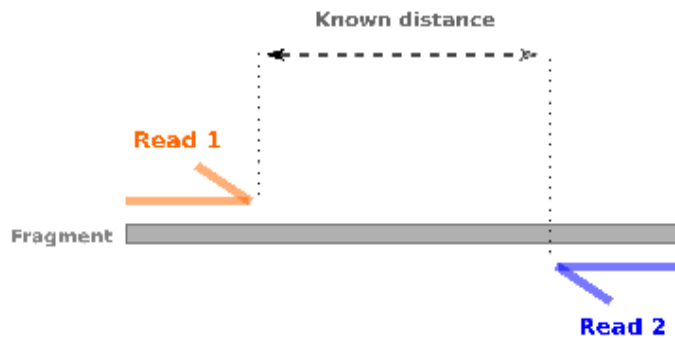
- There is ambiguity mapping a read with a mismatch versus a deletion
- A read could align to multiple places (repeats)



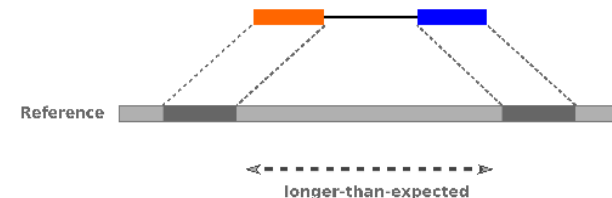
Steps in NGS analysis

Mapping reads to the genome

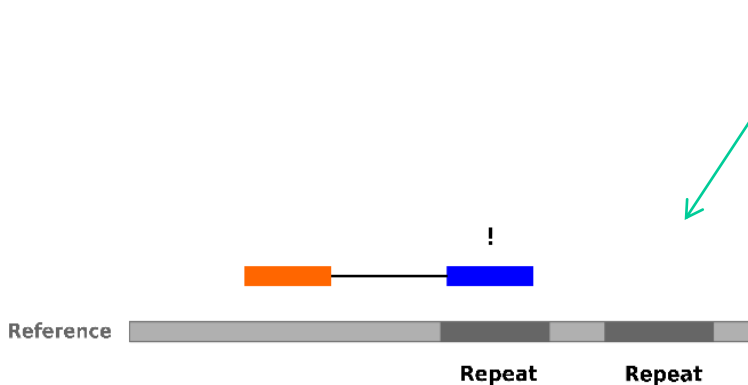
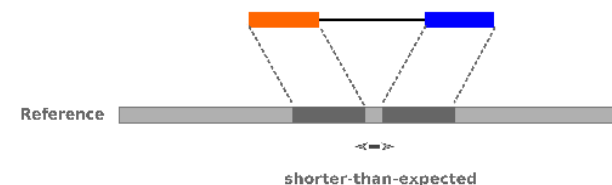
- Paired-end sequencing improves accuracy of mapping
- Sequencing:** Cut longer fragments of DNA, sequence only the ends



- Deletions:** Longer mapping distance than expected



- Insertions:** Shorter mapping distance than expected



Steps in NGS analysis

Mapping reads to the genome

- Quality scores to assess mapping accuracy
 - quantify the probability that a read is misplaced.
 - Function of factors such as:
 - uniqueness (ie not a multi-mapper)
 - number of mismatches in read
 - number of insertions/deletions in read
 - quality of bases in read

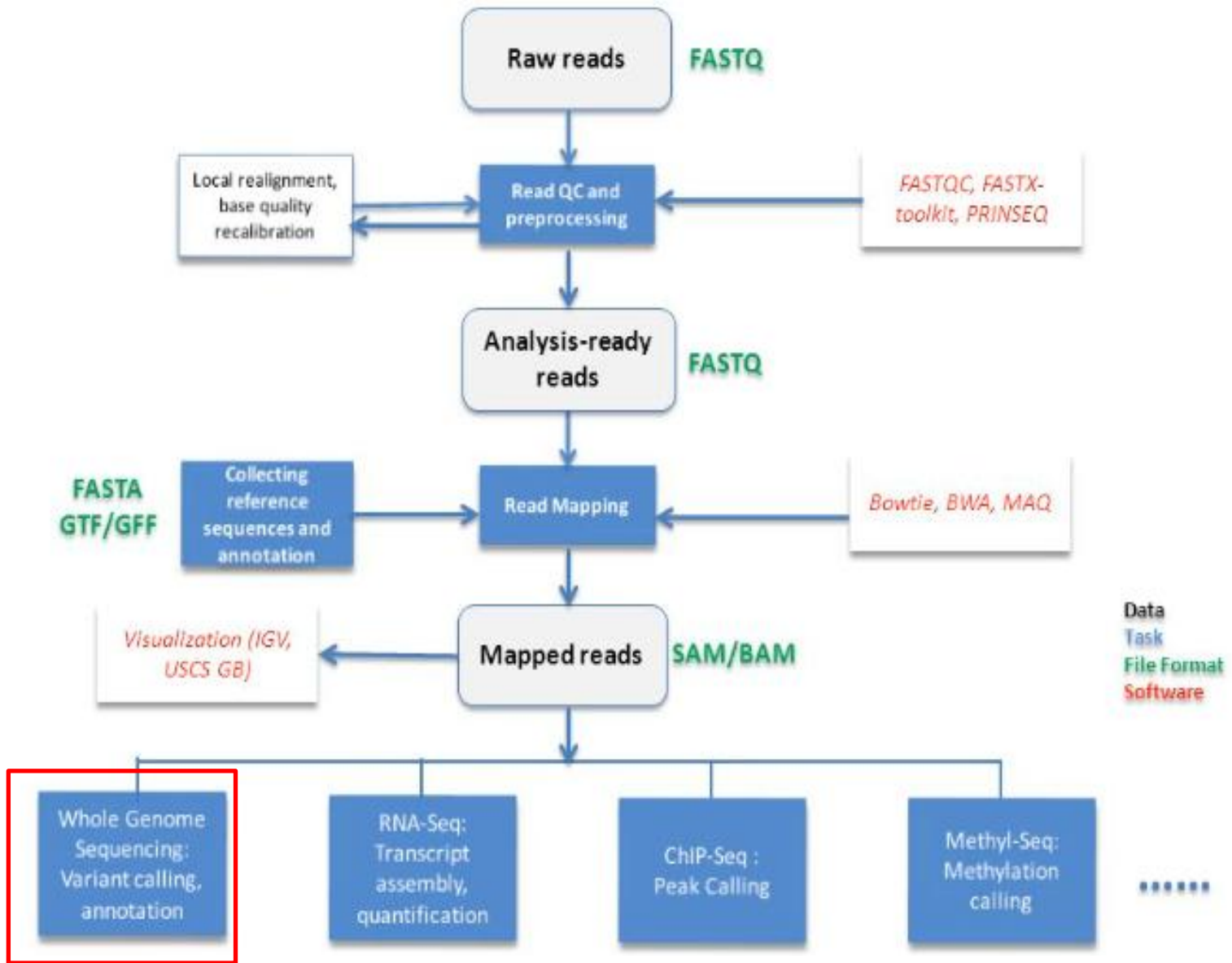
Sequence One : GGCTGG

Sequence Two : GAGG

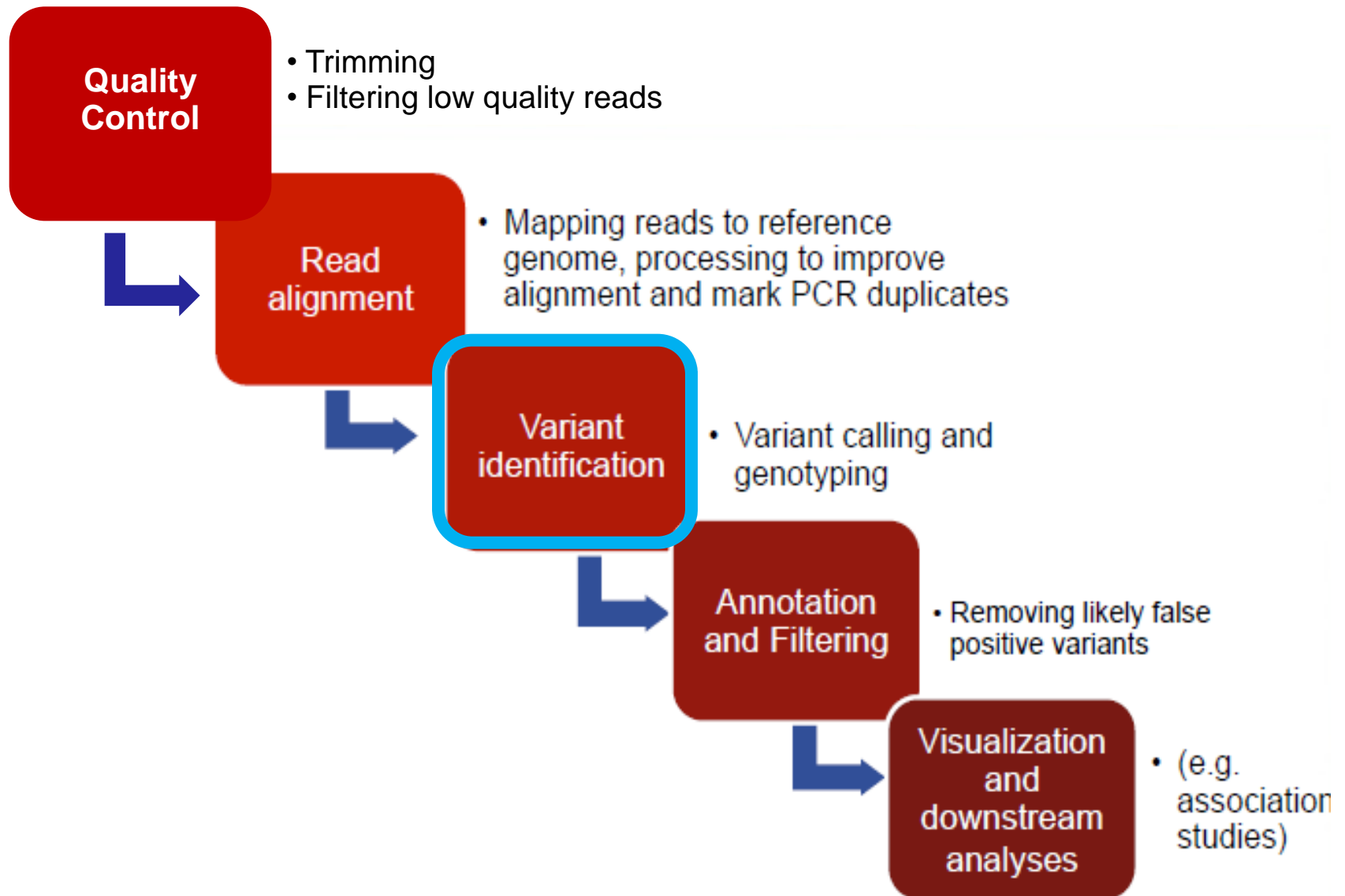
	G	G	C	T	G	G
	G	A	-	-	G	G
	10	-5	-5	-1	10	10
	10	5	0	-1	9	19

Your cumulative score

Steps in NGS analysis



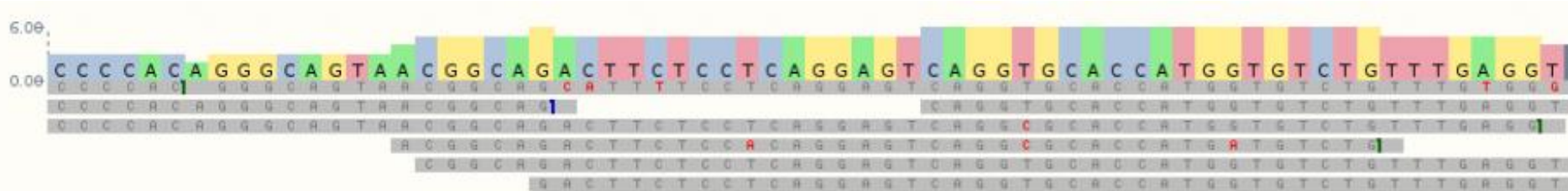
Steps in variant analysis



Steps in variant analysis

Variant Calling

- Variant calling involves comparing a sample sequence, which may be a single gene sequence, a whole exome or a whole genome, with a *reference* sequence.
- Goal: Identify variant bases, genotype likelihood and allele frequency while avoiding instrument noise

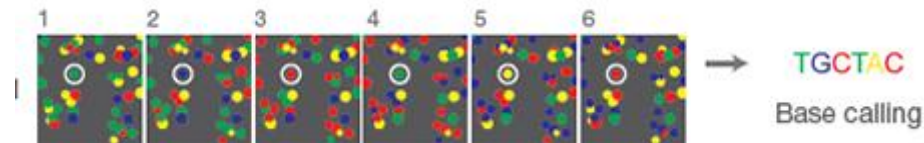


- A variant call can be:
 - a true variant
 - an experimental artifact, e.g. a library preparation error
 - a base calling error
 - an analysis error, e.g. a misalignment

Steps in variant analysis

Challenges in variant analysis

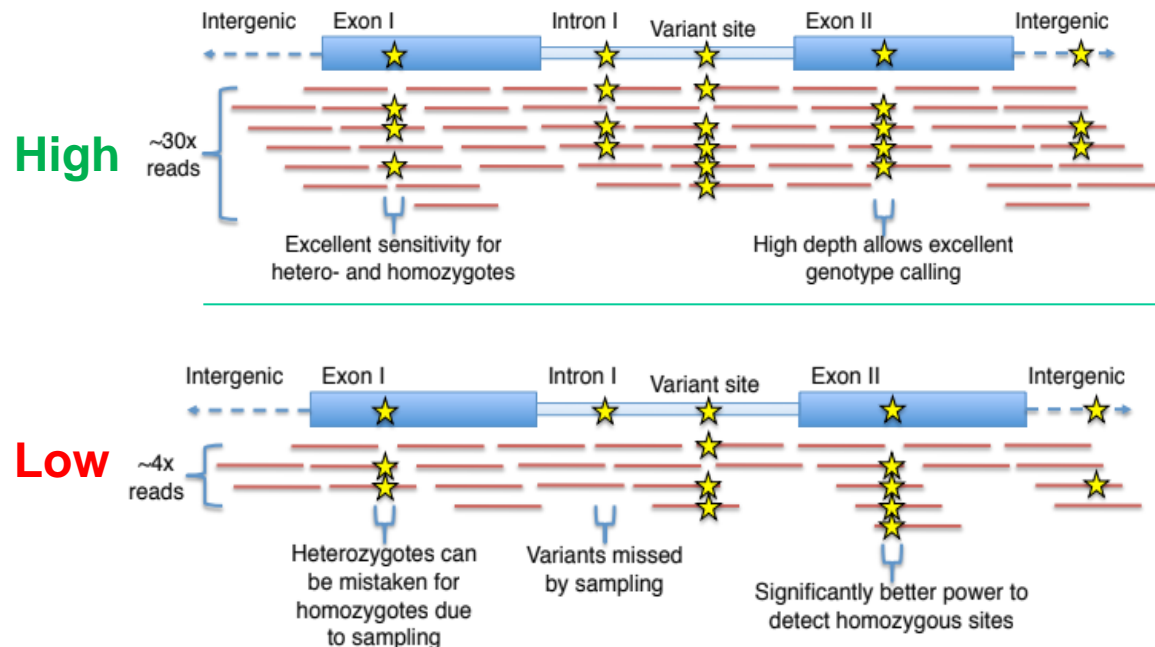
- Base calling errors
 - Different types of errors that vary by technology, sequence cycle and sequence context



Steps in variant analysis

Challenges in variant analysis

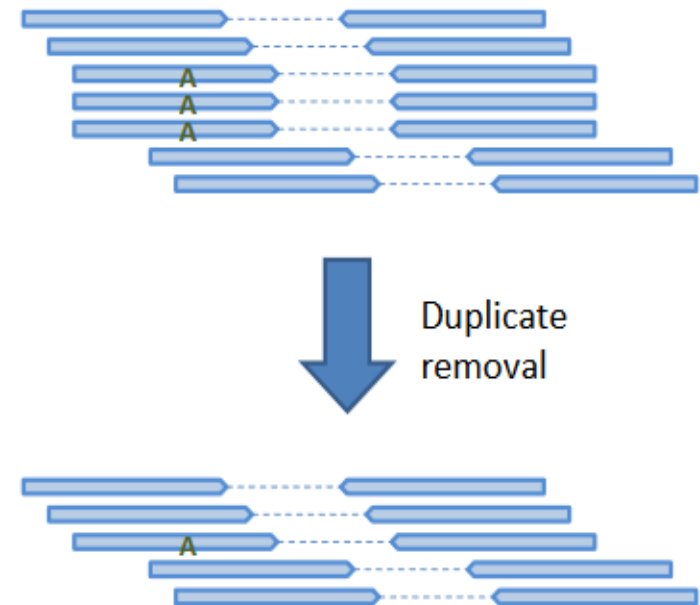
- Base calling errors
 - Different types of errors that vary by technology, sequence cycle and sequence context
- Low coverage sequencing
 - Lack of sequence from two chromosomes of a diploid individual at a site



Steps in variant analysis

Challenges in variant analysis

- Base calling errors
 - Different types of errors that vary by technology, sequence cycle and sequence context
- Low coverage sequencing
 - Lack of sequence from two chromosomes of a diploid individual at a site
- PCR duplicates during library preparation
 - A library that is composed mainly of PCR duplicates could produce inaccurate variant calling (altered coverage depth and variant frequency)
 - It is not required to remove duplicate reads prior to mapping but instead it is recommended to mark duplicates after the alignment.



Steps in variant analysis

Challenges in variant analysis

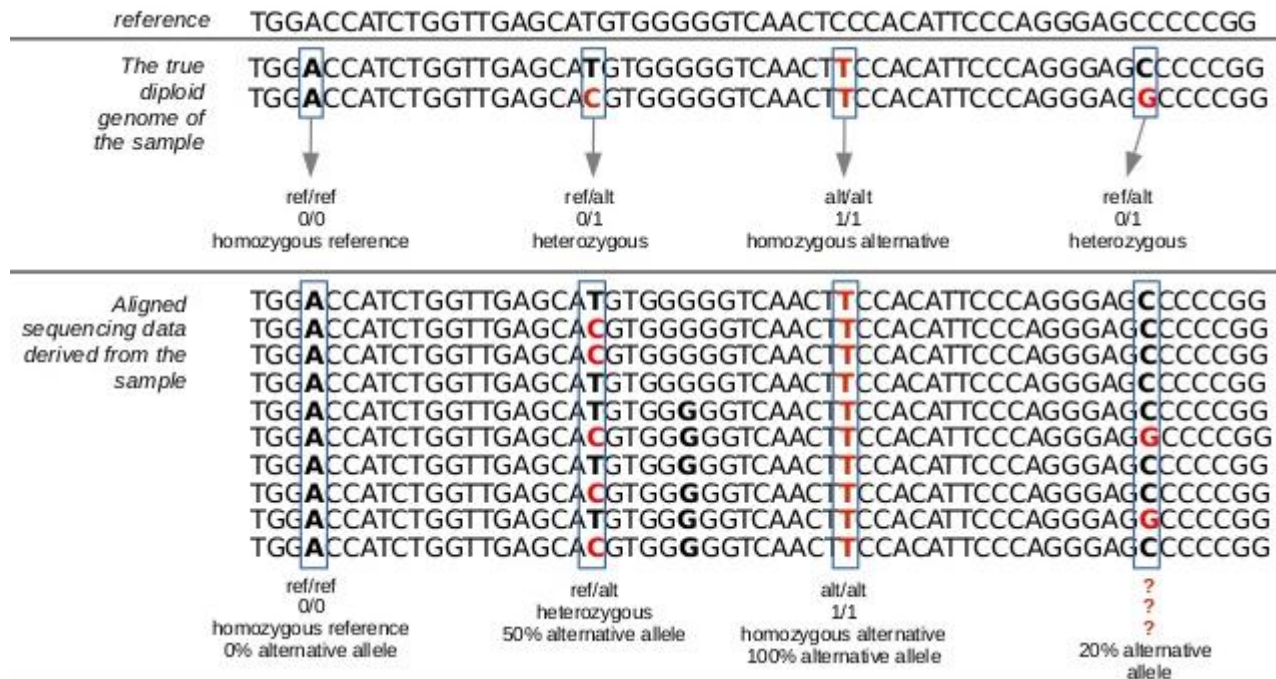
- Base calling errors
 - Different types of errors that vary by technology, sequence cycle and sequence context
- Low coverage sequencing
 - Lack of sequence from two chromosomes of a diploid individual at a site
- PCR duplicates during library preparation
- Inaccurate mapping
 - Aligned reads should be reported with mapping quality score



Steps in variant analysis

Variant Calling

- SNP calling software (SNP callers) can be used to look for SNPs
- Early SNP callers and some commercial packages use a simple method of counting reads for each allele that have passed a mapping quality threshold.
 - This is not good enough, in particular when coverage is low



Steps in variant analysis

Variant Calling

- Advanced SNP callers add more statistics for more accurate variant calling (eg. [FreeBayes](#), [GATK](#))

- Haplotype-based variant calling
 - Looking at a haplotype window instead of individual positions makes misalignments tolerable.

	Ref	Variant Region		Variant Region							
Reads	TACCGAT	CATTGGATCA	CGATTCC...GCATTGC	AAAAAAA-	GACCGCA						
	TACCGAT	CATTGGATCA	CGATTCC...GCATTGC	-AAAAAA-	GACCGCA						
	ACCGAT	TATTGCATCG	CGATTCC...GCATTGC	-AAAAAA-	GACCGCA						
	ACCGAT	CATTGGATCA	CGATTCC...GCATTGC	AAAAAA-A	GACCGCA						
	ACCGAT	TATTGGATCG	CGATTCC...GCATTGC	-AAAAAAA	GACCGCA						
	CCGAT	C-TTGGATCA	CGATTCC...GCATTGC	AAAAAAA-	GACCGCA						
	CCGAT	CAT G GGATCA	CGATTCC...GCATTGC	AAAAAAA A	GACCGCA						

Haplotypes	<div>CATTGGATCA</div>		x8	<div>(A)₇</div>		x10					
	<div>TATTGGATCG</div>		x9	<div>(A)₆</div>		x7					

- Bayesian reasoning
 - Joint genotyping
 - Prior on genotype distributions

Bayesian model

$$\Pr\{G|D\} = \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}} \overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\sum_i \Pr\{G_i\} \Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

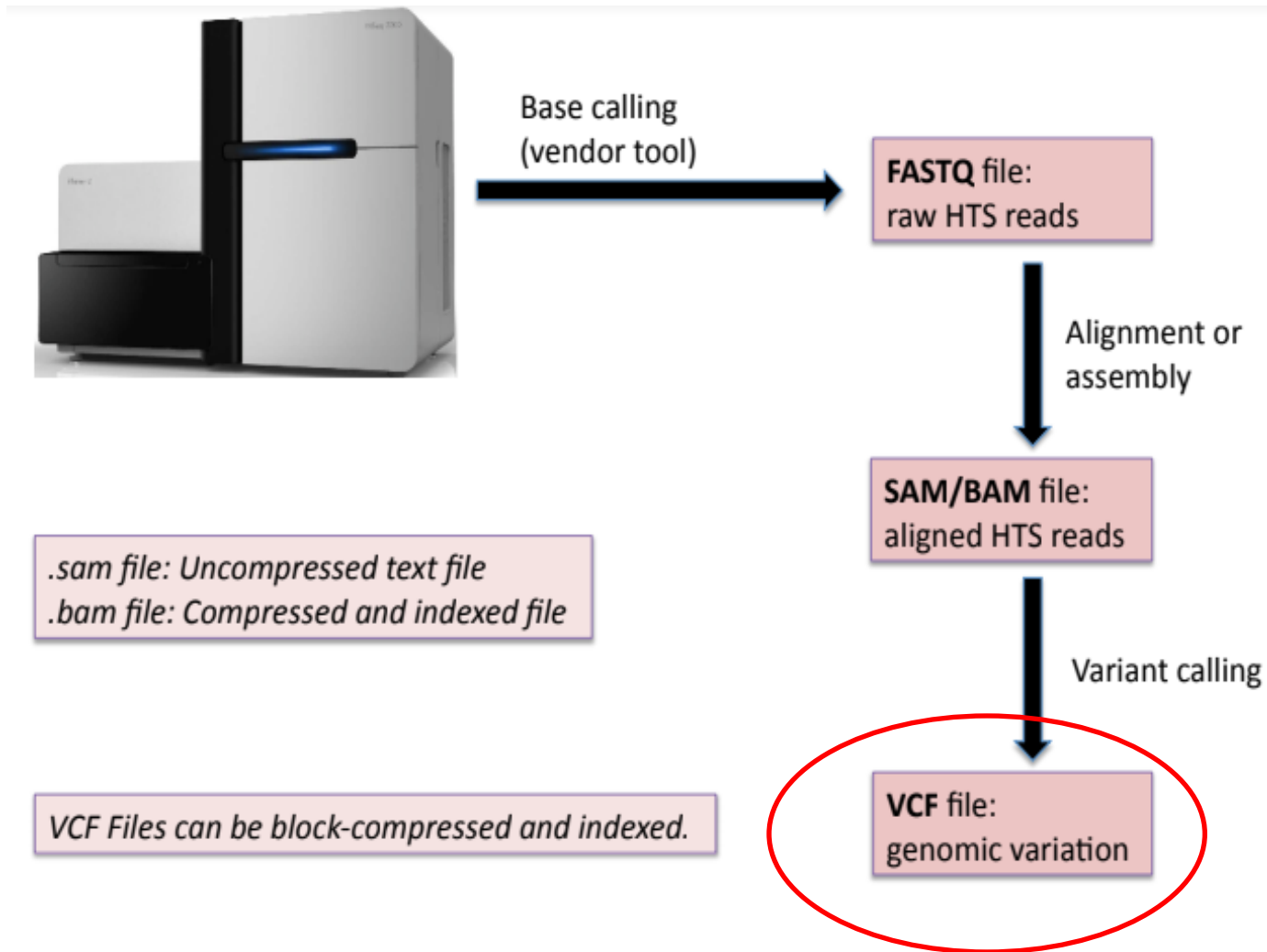
$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } G = H_1 H_2 \text{ (Diploid assumption)}$$

$\Pr\{D|H\}$ is the haploid likelihood function

Steps in variant analysis

Variant Calling

- The output is written into a Variant Call Format (VCF) file



Steps in variant analysis

Variant Calling

- A VCF file includes the following information:

Column	Mandatory	Description
CHROM	Yes	Chromosome
POS	Yes	1-based position of the start of the variant
ID	Yes	Unique identifier of the variant; the dbSNP entry rs1413368 is given in our example
REF	Yes	Reference allele
ALT	Yes	A comma-separated list of alternate nonreference alleles
QUAL	Yes	Phred-scaled quality score
FILTER	Yes	Site filtering information; in our example it is PASS
INFO	Yes	A semicolon-separated list of additional information. These fields include the gene identifier GI (here the gene is NEGR1); the transcript identifier TI (here NM_173808); and the functional consequence FC (here a synonymous change, T296T).
FORMAT	No	Defines information in subsequent genotype columns; colon separated. For example, GT:AD:DP:GQ:PL:VF:GX in our example refers to genotype (GT), allelic depths for the ref and alt alleles in the order listed (AD), approximate read depth (reads with MQ=255 or with bad mates are filtered) (DP), genotype quality (GQ), normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification (PL), variant frequency, the ratio of the sum of the called variant depth to the total depth (VF), and minimum of {genotype quality assuming variant position, genotype quality assuming nonvariant position} (GX).
Sample	No	Sample identifiers define the samples included in the VCF file

Steps in variant analysis

Variant Calling

- A VCF file includes the following information:

Example

VCF header

```

##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">

```

Body

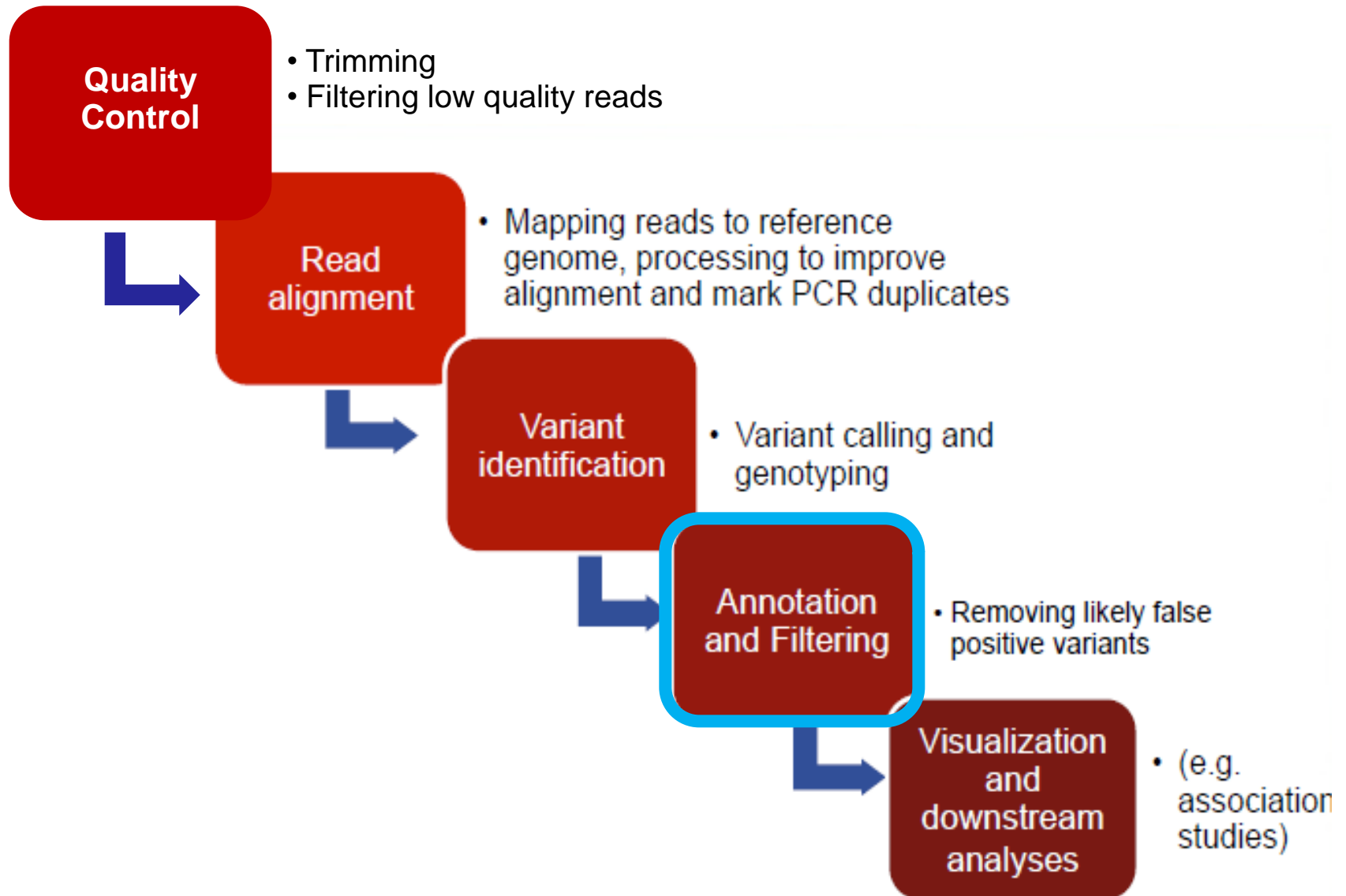
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0/1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1/0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Annotations:

- Mandatory header lines:** ##fileformat=VCFv4.0
- Optional header lines (meta-data about the annotations in the VCF body):** ##INFO, ##FORMAT, ##ALT
- Reference alleles (GT=0):** A, T, G, C
- Alternate alleles (GT>0 is an index to the ALT column):** AT, CT, G,
- Phased data (G and C above are on the same chromosome):** 1/1:12:3
- Deletion:**
- SNP:** rs1
- Large SV:** SVTYPE=DEL;END=300
- Insertion:** T, CT
- Other event:** H2;AA=T

A typical VCF file from a human whole exome sequence experiment may contain ~80,000 rows. A typical human whole genome sequence experiment produces a VCF with ~4 million rows.

Steps in variant analysis



Steps in variant analysis

Annotation and filtering

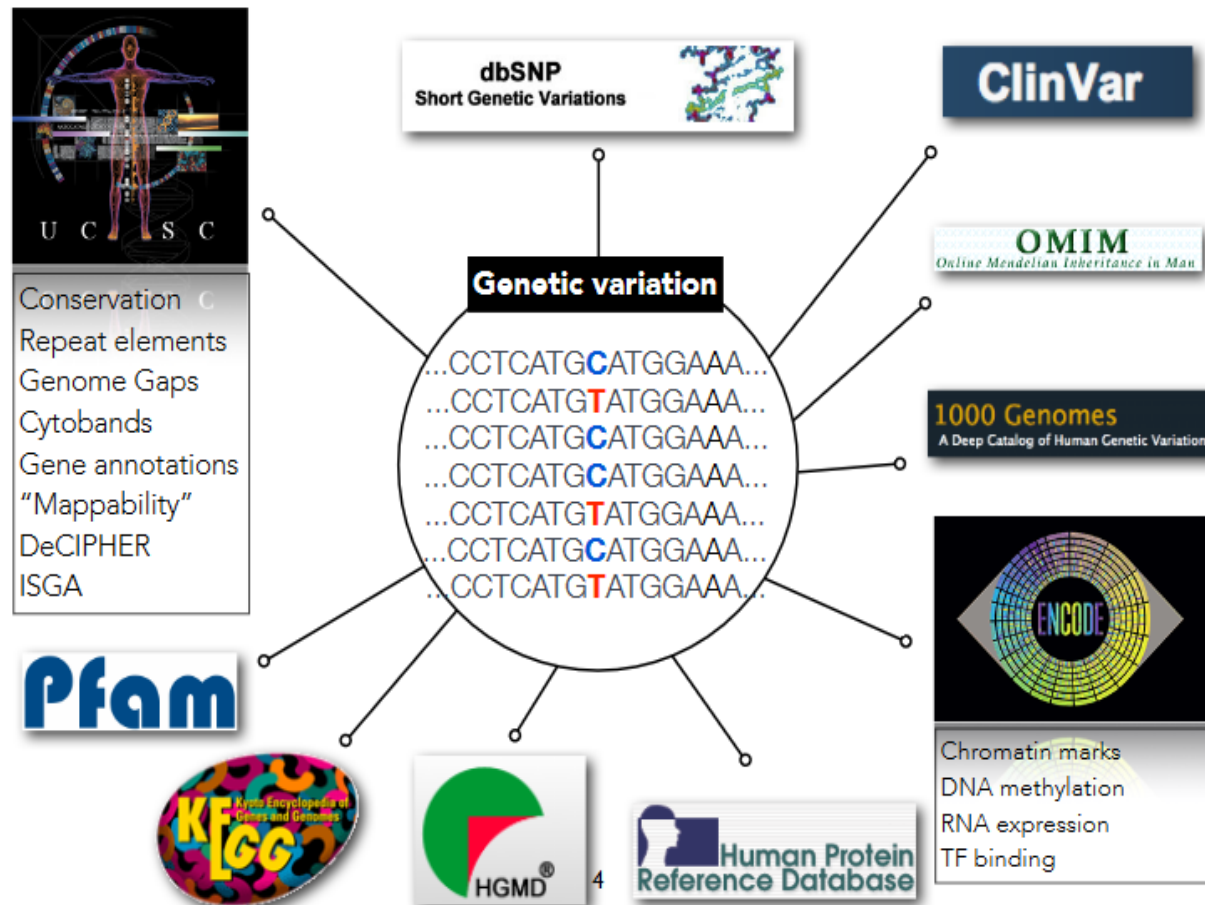
From the output of thousands of variants, which ones should we consider?

- Various important considerations
 - Is the variant call of good quality?
 - Was this variant previously found in the population? At what frequency?
 - If you expect a rare mutation, is the variant commonly found in the general population?
 - Is this variant associated with some disease / trait?
 - Which gene is affected by the variant?
- What is the predicted effect of the variant?
 - Non-synonymous
 - Detrimental for function

Steps in variant analysis

Annotation and filtering

- Annotations provide context
- Prioritizing variants and assessing functional significance





Steps in variant analysis

Annotation and filtering

- Evaluating functional consequences

Missense variants 

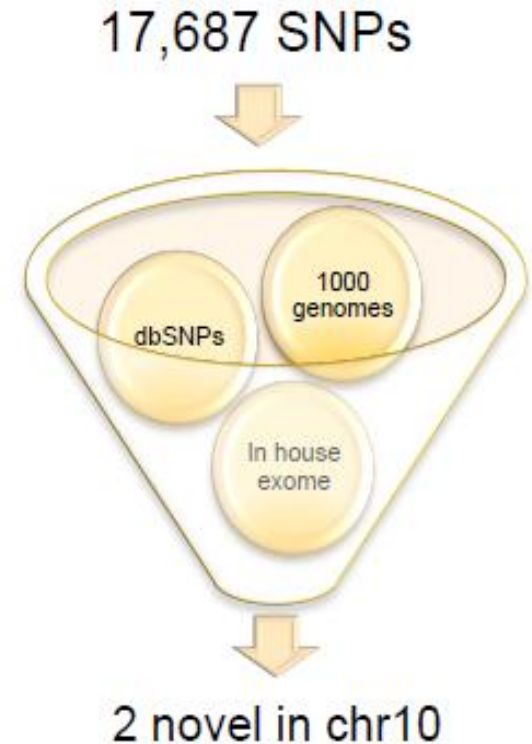
[\[back to top\]](#)

Show All  entries Show/hide columns Filter 									
ID	Chr: bp	Alleles	Class	Source	Type	AA	AA coord	SIFT	Poly Phen
rs121909815	11:5248247	A/G	SNP	dbSNP	Missense variant	V/A	2	0.01	0.119
rs121909830	11:5248247	A/C	SNP	dbSNP	Missense variant	V/G	2	0.07	0.007
rs121909815	11:5248247	A/G	SNP	dbSNP	Missense variant	V/A	2	0.01	0.119
rs121909830	11:5248247	A/C	SNP	dbSNP	Missense variant	V/G	2	0.01	0.007
rs33958358	11:5248248	C/T/A	SNP	dbSNP	Missense variant	V/L	2	0.01	0.001
rs33958358	11:5248248	C/T/A	SNP	dbSNP	Missense variant	V/M	2	0	0.271
rs33958358	11:5248248	C/T/A	SNP	dbSNP	Missense variant	V/L	2	0.02	0.001
rs33958358	11:5248248	C/T/A	SNP	dbSNP	Missense variant	V/M	2	0	0.271
rs35906307	11:5248245	G/A	SNP	dbSNP	Missense variant	H/Y	3	0.02	0.135

Steps in variant analysis

Annotation and filtering

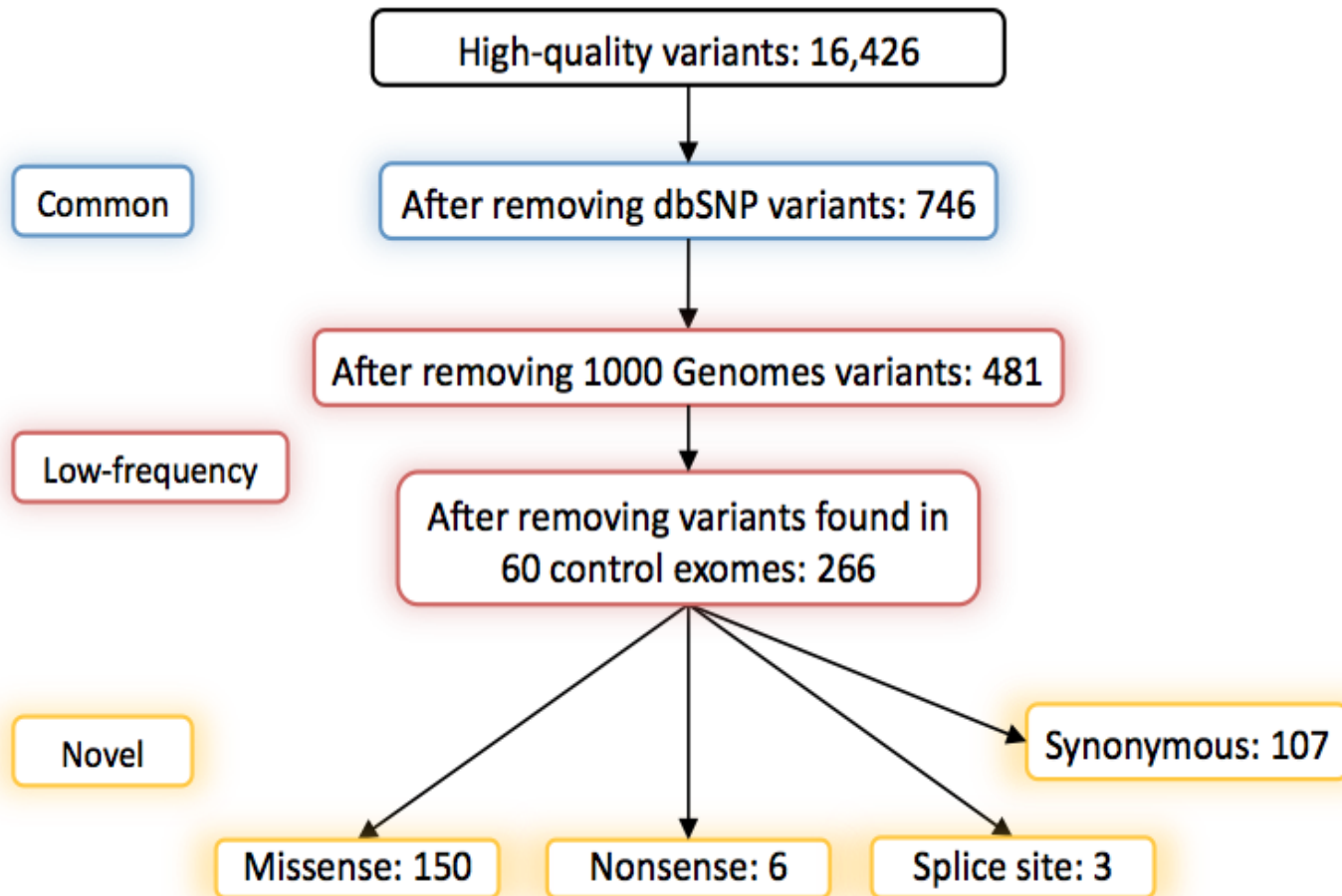
- The initial set of variants is usually filtered extensively in hopes of removing false positives.
- More filtering can include public data or custom filters based on in-house data.
- Most exome studies will then filter common variants ($>1\%$) such as those in dbSNP database
- Filtering by variant consequence to get the deleterious ones



Steps in variant analysis

Annotation and filtering

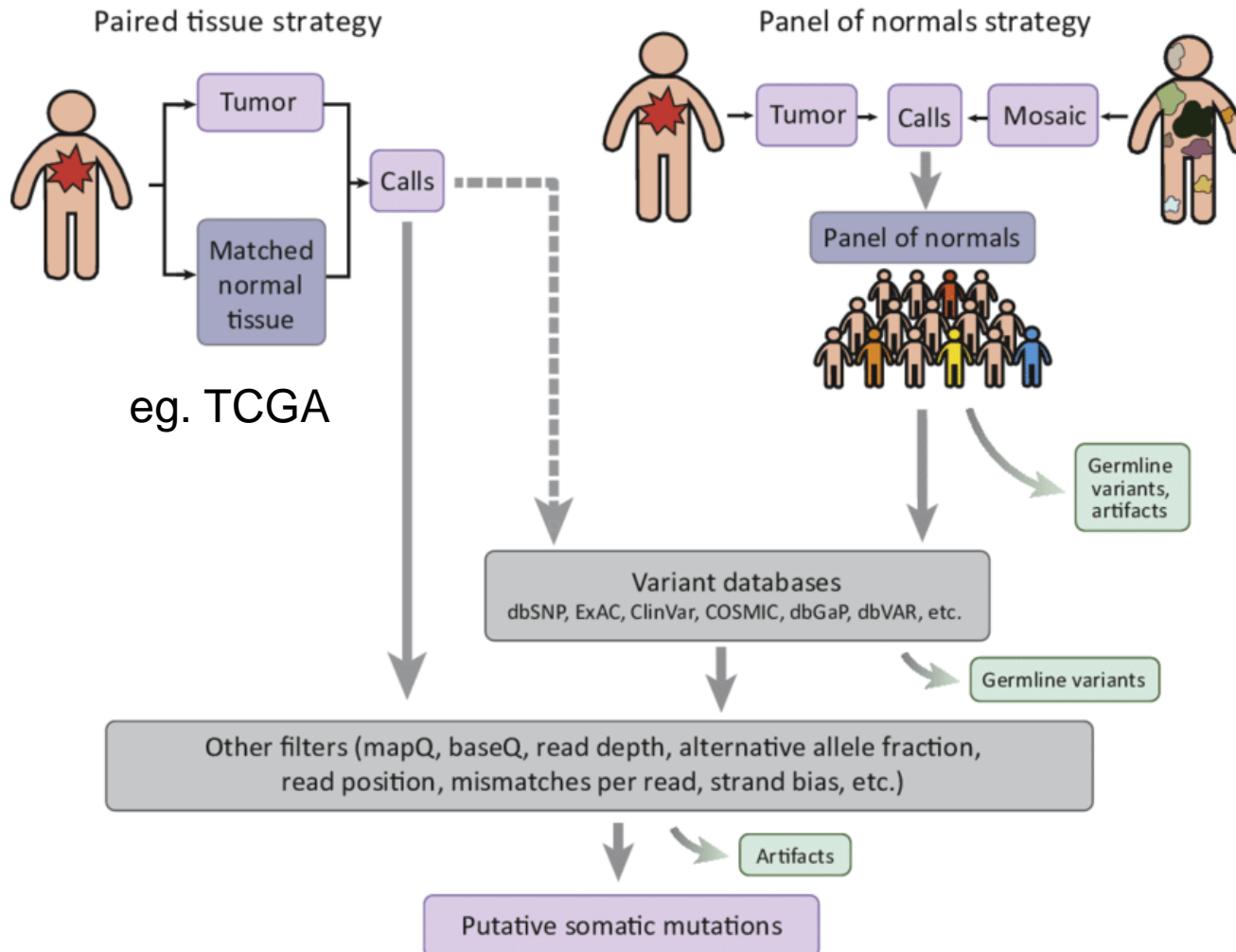
- Example: finding novel variants in Individual X



Steps in variant analysis

Annotation and filtering

- A different strategy may be used for somatic mutations (eg. cancer)



Steps in variant analysis

Annotation and filtering

- A different strategy may be used for Mendelian disorders (trio analysis)

vcf file ("good quality variants - all genotypes")



Apply model of inheritance: e.g. autosomal recessive

Select variants with genotype 0/1 in the parents and 1/1 in the daughter

filtered vcf file ("recessively inherited variants")

CHR	POS	REF	ALT	GT_daughter	GT_father	GT_mother
1	200827638	A	G	1/1	0/1	0/1
1	22158157	A	G	1/1	0/1	0/1
2	171256597	A	C	1/1	0/1	0/1
2	208976955	A	C	1/1	0/1	0/1
4	48496368	A	G	1/1	0/1	0/1
7	14017007	C	T	1/1	0/1	0/1
8	143310815	G	A	1/1	0/1	0/1
8	41517860	G	A	1/1	0/1	0/1
11	47437403	C	T	1/1	0/1	0/1
11	59837097	C	T	1/1	0/1	0/1
12	9833628	C	T	1/1	0/1	0/1
13	36699762	G	A	1/1	0/1	0/1
13	52523808	C	T	1/1	0/1	0/1
14	38256944	T	C	1/1	0/1	0/1
15	79026001	C	A	1/1	0/1	0/1
16	10788129	G	T	1/1	0/1	0/1
16	1498197	A	G	1/1	0/1	0/1
19	49640002	G	T	1/1	0/1	0/1
20	10026357	T	C	1/1	0/1	0/1
22	23657980	G	A	1/1	0/1	0/1

Steps in variant analysis

Annotation and filtering



Steps in variant analysis

Annotation and filtering

- Tools for annotating/filtering variants:

- **ANNOVAR** – <http://annovar.openbioinformatics.org>

- **SnEff** – <http://snpeff.sourceforge.net>

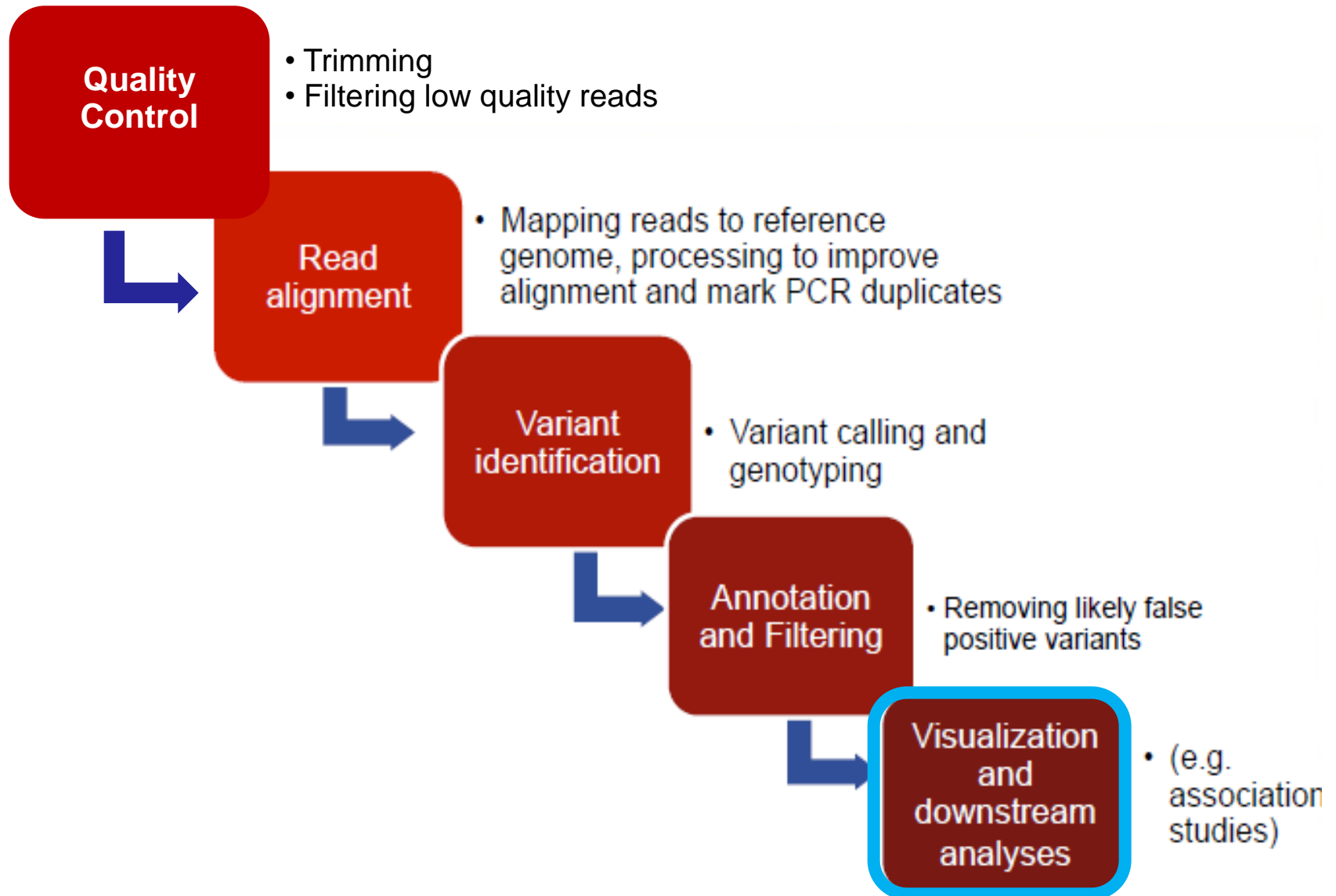
- **Ensembl Variant Effect Predictor** -

<https://www.ensembl.org/info/docs/tools/vep/index.html>

- **PheGenI** - <https://www.ncbi.nlm.nih.gov/gap/phegeni>

- **GEMINI** - <https://gemini.readthedocs.io/en/latest/>

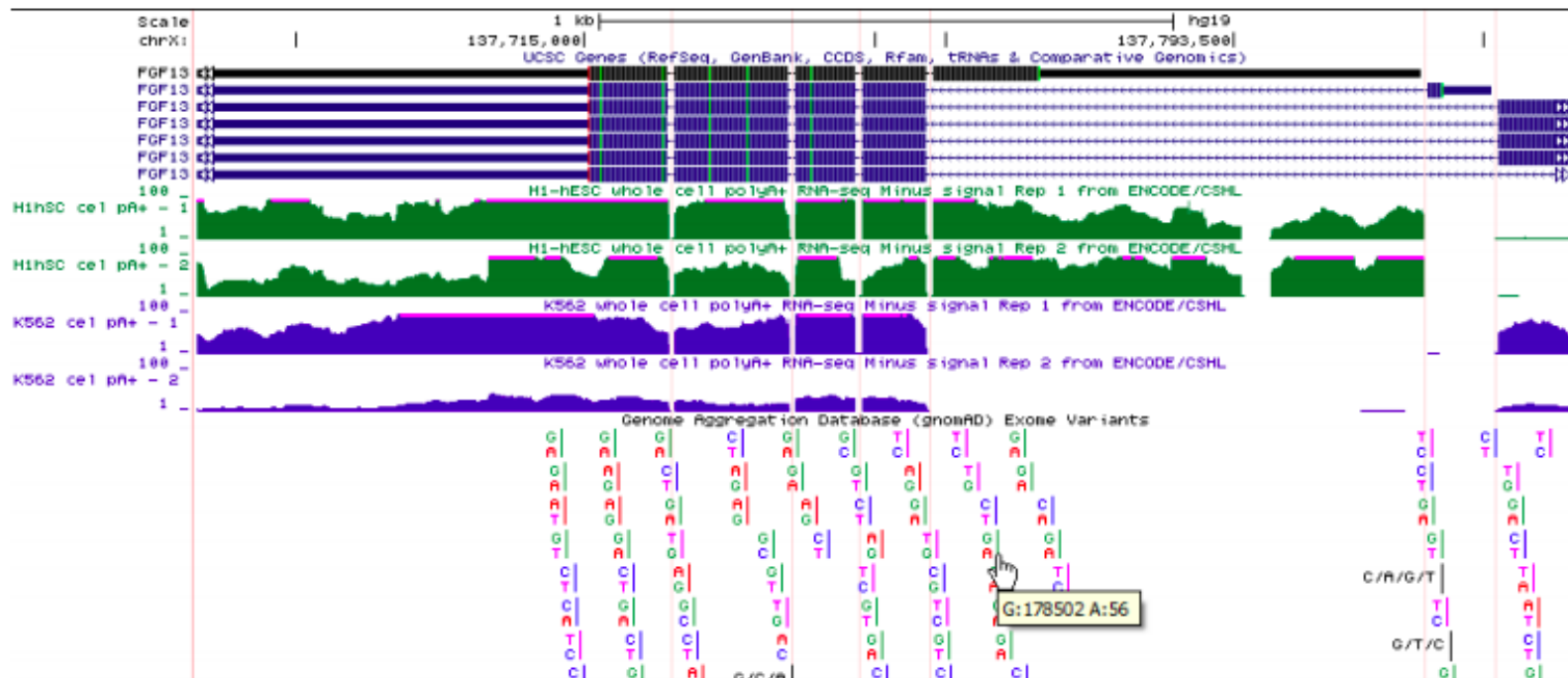
Steps in variant analysis



Steps in variant analysis

Visualization and downstream analysis

- Variants can be visualized in context of the genome using any of the available Genome Browsers (UCSC, Ensembl, IGV, ...)



Steps in variant analysis

Visualization and downstream analysis

- Validation of identified variants (Sanger sequencing)
- Statistics, association studies, integrative analysis...

References and resources

Bibliography

- Pabinger et al. *A survey of tools for variant analysis of next-generation genome sequencing data*. Briefings in Bioinf. 2012

Links and resources

Galaxy tutorials

<https://galaxyproject.github.io/training-material/topics/variant-analysis/>

EMBL-EBI materials

<https://www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction-2019/summary>

<https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data>

This lecture is based on many presentation freely available on the web.

We wish to acknowledge the authors for their efforts and for making their work available.