FUNDAMENTOS DE SECUENCIACIÓN DE ÚLTIMA GENERACIÓN Y GENÓMICA TRASLACIONAL

Variant Detection

UAM Universidad Autónoma de Madrid

Contact info:

Bioinformatics Unit directed by Dr. Fátima Al-Shahrour
Centro Nacional de Investigaciones Oncológicas

Elena Piñeiro: epineiro@cnio.es

BU CNIO BIOINFORMATICS UNIT

Outline

- Definition, relevance and types of genomic variants
- Clinical implications of genomic variants
- Bioinformatics steps for genomic variant detection
- Algorithms for variant calling

Genomic Variants

Definition, relevance and types

What are genomic variants?

- Genomic variants are **permanent changes** in the **DNA sequence of an organism**.

- They **can emerge by different mechanisms**:
  - Recombination during gametes formation.
  - Errors during the DNA replication.
  - External factors like radiation, viruses, transposons, tobacco, UV light.

Genome is around 99.9 percent identical between humans at the base-pair level.

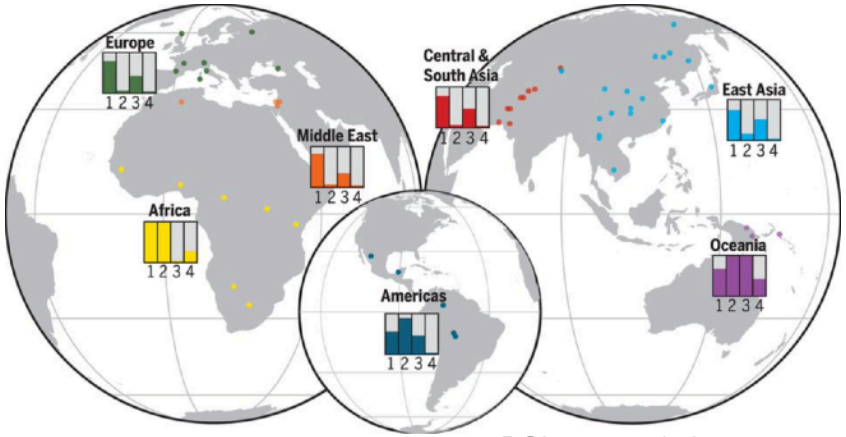
Variation:

- Is a source of variability, laying the foundations for the evolutionary mechanisms.
- Allow phenotypic differences between individuals (hair color, skin color, ...).
- Involved in diseases and drug response.

https://www.genome.gov

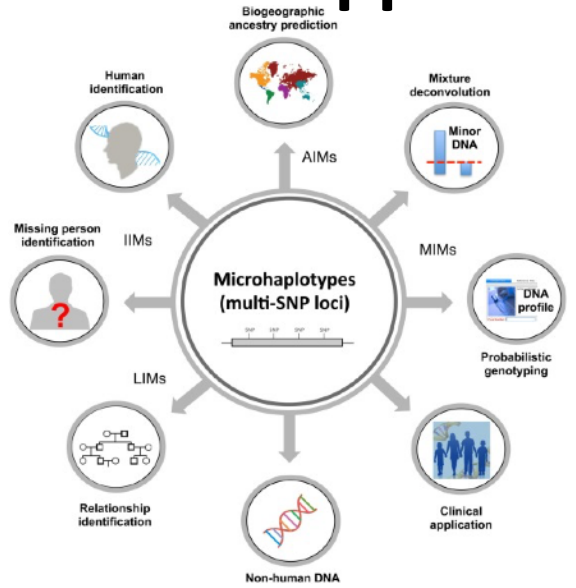
Areas of study of genomic variants

Population and demographical studies



DOI: 10.1126/science.aay5012

Forensic applications

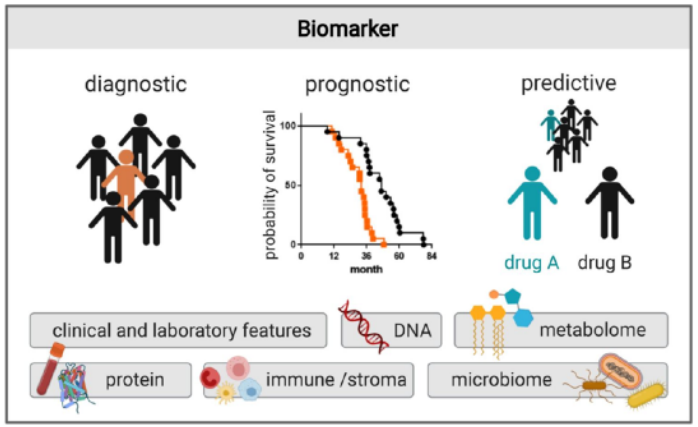


DOI: 10.1016/j.fsigen.2018.09.009

doi: 10.1038/s41598-020-58356-1

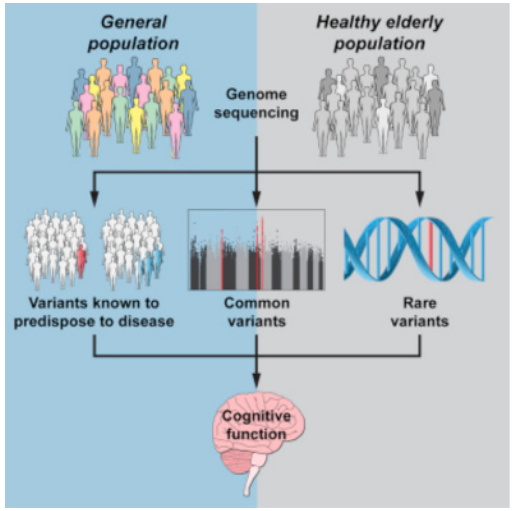
Phylogenetic studies

Clinical applications



DOI: 10.3390/cancers12113234

Epidemiological studies

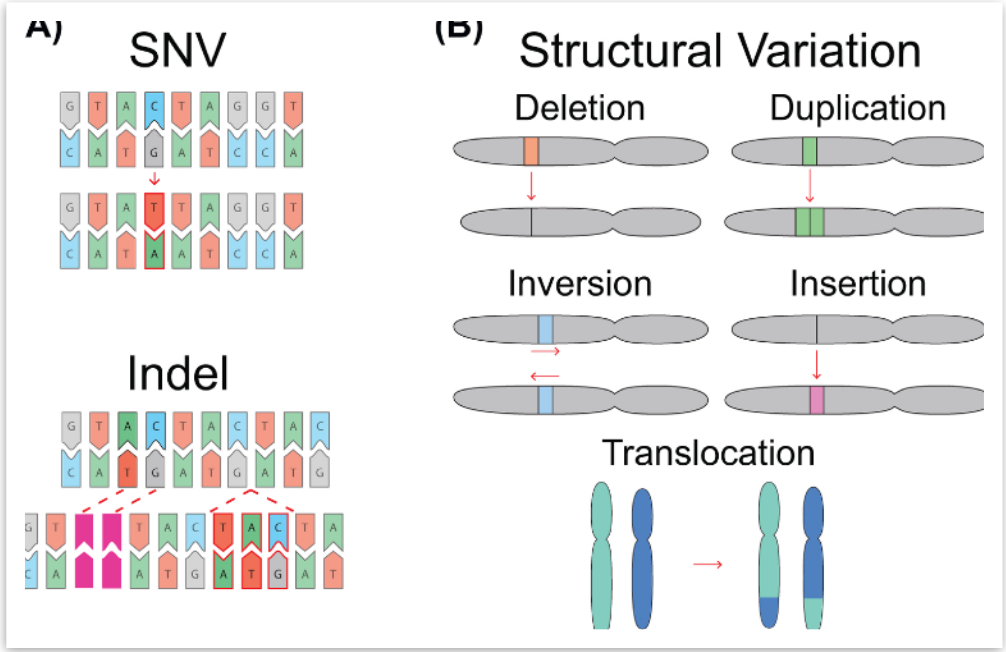


DOI:https://doi.org/10.1016/j.cell.2016.03.022

Types of genomic variants

- Different types of variants according to different criteria:
  - Variant **size**
  - Variant **position** in the DNA sequence
  - **Consequence** of the variant in transcription and translation
  - **Clinical implication**

Variants according to size



doi: 10.1016/j.tig.2020.10.003

Polymorphism

SNV Single Nucleotide Variant - SNP
MNV Multiple Nucleotide Variant
Indel Small Insertion and Deletion

VNTR (Micro, Minisatellites)
Variable Number of Tandem Repetition

CNV Copy Number Variation
Translocation, Inversion

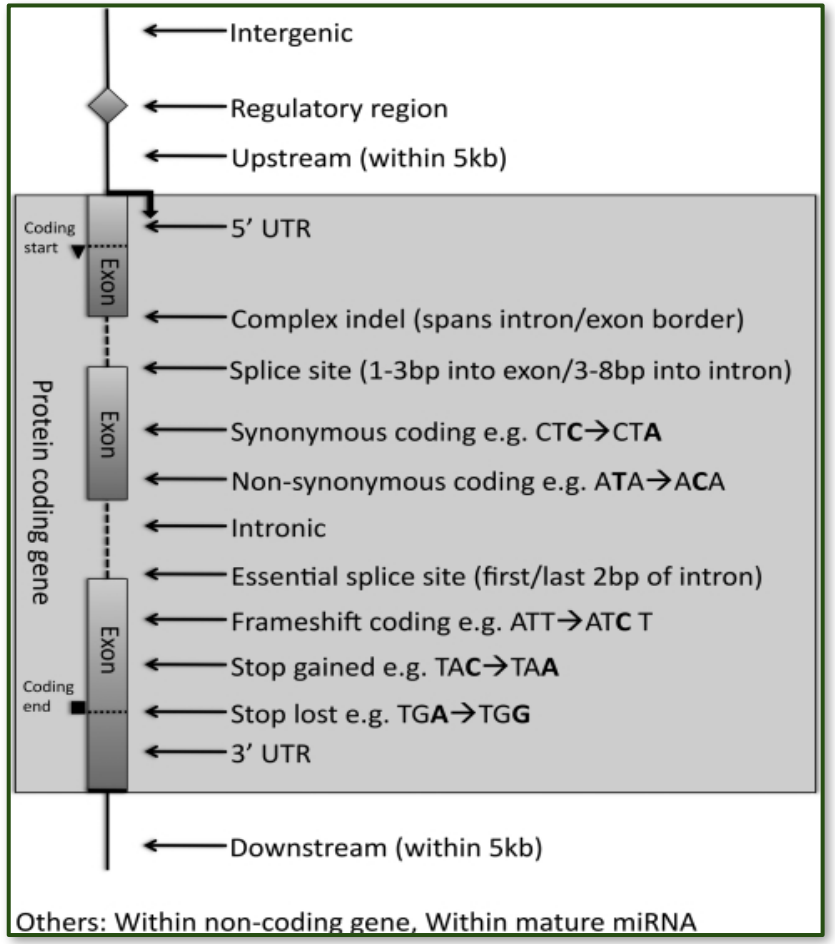
Structural variants

Aneuploidy

Size

Definition, relevance and types of genomic variants

Variants according to the sequence position



McLaren et al. Bioinformatics. 2010.

- Intergenic
  - In regulatory regions
  - Upstream
  - Downstream
- In genes
  - Untranslated regions: 5'UTR y 3'UTR
  - Exons
  - Introns
  - Splicing regions

Clinical implications of genomic variants

Biomarkers

- Associated with a higher/lower **risk** of suffering a pathology (e.g. BRCA variants in some cancer types)
- Allow the establishment of a **diagnosis** (e.g. BRC-ABL1 for CML)
- Give a **prognosis** about the evolution or the appearance of some event
- Allow the **prediction** of treatment response
- Allow **therapeutical adjustments** (doses and secondary effects)

Pharmacogenomic biomarkers

Drug	Biomarker
Abemaciclib (1)	ESR (Hormone Receptor)
Abemaciclib (2)	ERBB2 (HER2)
Ado-Trastuzumab Emtansine	ERBB2 (HER2)
Afatinib	EGFR
Alectinib	ALK
Anastrozole	ESR, PGR (Hormone Receptor)
Arsenic Trioxide	PML-RARA
Atezolizumab	CD274 (PD-L1)
Binimetinib	BRAF
Blinatumomab	BCR-ABL1 (Philadelphia chromosome)
Bosutinib	BCR-ABL1 (Philadelphia chromosome)
Brentuximab Vedotin	TNFRSF8 (CD30)
Brigatinib	ALK
Ceritinib	ALK
Cetuximab (1)	EGFR
Cetuximab (2)	RAS
Cobimetinib	BRAF
Crizotinib (1)	ALK
Crizotinib (2)	ROS1
Dabrafenib	BRAF
Dacomitinib	EGFR
Dasatinib	BCR-ABL1 (Philadelphia chromosome)
Denileukin Diftitox	IL2RA (CD25 antigen)
Enasidenib	IDH2
Encorafenib	BRAF
Erlotinib	EGFR

Drug	Biomarker
Everolimus (1)	ERBB2 (HER2)
Everolimus (2)	ESR (Hormone Receptor)
Exemestane	ESR, PGR (Hormone Receptor)
Fulvestrant (1)	ERBB2 (HER2)
Fulvestrant (2)	ESR, PGR (Hormone Receptor)
Gefitinib	EGFR
Gilteritinib	FLT3
Goserelin	ESR, PGR (Hormone Receptor)
Ibrutinib	Chromosome 17p
Imatinib (1)	KIT
Imatinib (2)	BCR-ABL1 (Philadelphia chromosome)
Imatinib (3)	PDGFRB
Imatinib (4)	FIP1L1-PDGFRA
Ipilimumab	Microsatellite Instability, Mismatch Repair
Ivosidenib	IDH1
Lapatinib (1)	ERBB2 (HER2)
Lapatinib (2)	ESR, PGR (Hormone Receptor)
Larotrectinib	NTRK
Letrozole	ESR, PGR (Hormone Receptor)
Lorlatinib	ALK
Midostaurin	FLT3
Neratinib	ERBB2 (HER2)
Nilotinib (1)	BCR-ABL1 (Philadelphia chromosome)
Nivolumab (1)	BRAF
Nivolumab (3)	Microsatellite Instability, Mismatch Repair
Olaparib (1)	BRCA

Drug	Biomarker
Olaparib (2)	ERBB2 (HER2)
Olaparib (3)	ESR, PGR (Hormone Receptor)
Osimertinib	EGFR
Palbociclib (1)	ESR (Hormone Receptor)
Palbociclib (2)	ERBB2 (HER2)
Panitumumab (2)	RAS
Pembrolizumab (2)	CD274 (PD-L1)
Pembrolizumab (3)	Microsatellite Instability, Mismatch Repair
Pertuzumab (1)	ERBB2 (HER2)
Ponatinib	BCR-ABL1 (Philadelphia chromosome)
Ribociclib (1)	ESR, PGR (Hormone Receptor)
Ribociclib (2)	ERBB2 (HER2)
Rituximab	MS4A1 (CD20 antigen)
Rucaparib (1)	BRCA
Talazoparib (1)	BRCA
Talazoparib (2)	ERBB2 (HER2)
Tamoxifen (1)	ESR, PGR (Hormone Receptor)
Toremifene	ESR (Hormone Receptor)
Trametinib (1)	BRAF
Trastuzumab (1)	ERBB2 (HER2)
Tretinoin	PML-RARA
Vemurafenib (1)	BRAF
Venetoclax (1)	Chromosome 17p
Vincristine	BCR-ABL1 (Philadelphia chromosome)

FDA – Junio 2019

Pharmacogenomic biomarkers

Cetuximab – RAS – Oncology **Resistance**

Indications and Usage:

Erbitux is indicated for the treatment of **K-Ras wild-type**, epidermal growth factor receptor (EGFR)-expressing, metastatic colorectal cancer (mCRC) as determined by FDA-approved test for this use

Drug	Biomarker
Arsenic Trioxide	PML-RARA
Atezolizumab	CD274 (PD-L1)
Binimetinib	BRAF
Blinatumomab	BCR-ABL1 (Philadelphia chromosome)
Bosutinib	BCR-ABL1 (Philadelphia chromosome)
Brentuximab Vedotin	TNFRSF8 (CD30)
Brigatinib	ALK
Ceritinib	ALK
Cetuximab (1)	EGFR
Cetuximab (2)	RAS
Cobimetinib	BRAF
Crizotinib (1)	ALK
Crizotinib (2)	ROS1
Dabrafenib	BRAF
Dacomitinib	EGFR
Dasatinib	BCR-ABL1 (Philadelphia chromosome)
Denileukin Diftitox	IL2RA (CD25 antigen)
Enasidenib	IDH2
Encorafenib	BRAF
Erlotinib	EGFR

Drug	Biomarker
Gilteritinib	FLT3
Goserelin	ESR, PGR (Hormone Receptor)
Ibrutinib	Chromosome 17p
Imatinib (1)	KIT
Imatinib (2)	BCR-ABL1 (Philadelphia chromosome)
Imatinib (3)	PDGFRB
Imatinib (4)	FIP1L1-PDGFRA
Ipilimumab	Microsatellite Instability, Mismatch Repair
Ivosidenib	IDH1
Lapatinib (1)	ERBB2 (HER2)
Lapatinib (2)	ESR, PGR (Hormone Receptor)
Larotrectinib	NTRK
Letrozole	ESR, PGR (Hormone Receptor)
Lorlatinib	ALK
Midostaurin	FLT3
Neratinib	ERBB2 (HER2)
Nilotinib (1)	BCR-ABL1 (Philadelphia chromosome)
Nivolumab (1)	BRAF
Nivolumab (3)	Microsatellite Instability, Mismatch Repair
Olaparib (1)	BRCA

Drug	Biomarker
Pembrolizumab (2)	CD274 (PD-L1)
Pembrolizumab (3)	Microsatellite Instability, Mismatch Repair
Pertuzumab (1)	ERBB2 (HER2)
Ponatinib	BCR-ABL1 (Philadelphia chromosome)
Ribociclib (1)	ESR, PGR (Hormone Receptor)
Ribociclib (2)	ERBB2 (HER2)
Rituximab	MS4A1 (CD20 antigen)
Rucaparib (1)	BRCA
Talazoparib (1)	BRCA
Talazoparib (2)	ERBB2 (HER2)
Tamoxifen (1)	ESR, PGR (Hormone Receptor)
Toremifene	ESR (Hormone Receptor)
Trametinib (1)	BRAF
Trastuzumab (1)	ERBB2 (HER2)
Tretinoin	PML-RARA
Vemurafenib (1)	BRAF
Venetoclax (1)	Chromosome 17p
Vincristine	BCR-ABL1 (Philadelphia chromosome)

FDA – Junio 2019

Pharmacogenomic biomarkers

Cetuximab – RAS – Oncology
Indications and Usage:
Erbitux is indicated for the treatment of **K-Ras wild-type**, epidermal growth factor receptor (EGFR)-expressing, metastatic colorectal cancer (mCRC) as determined by FDA-approved test for this use

Alectinib – ALK – Oncology
Indications and Usage:
ALECENCSA is indicated for the treatment of patients with anaplastic lymphoma kinase **(ALK)-positive** metastatic non-small cell lung cancer (NSCLC) as detected by an FDA-approved test

Sensitivity

Drug	Biomarker
Brigatinib	ALK
Ceritinib	ALK
Cetuximab (1)	EGFR
Cetuximab (2)	RAS
Cobimetinib	BRAF
Crizotinib (1)	ALK
Crizotinib (2)	ROS1
Dabrafenib	BRAF
Dacomitinib	EGFR
Dasatinib	BCR-ABL1 (Philadelphia chromosome)
Denileukin Diftitox	IL2RA (CD25 antigen)
Enasidenib	IDH2
Encorafenib	BRAF
Erlotinib	EGFR

Drug	Biomarker
Imatinib (4)	FIP1L1-PDGFRA
Ipilimumab	Microsatellite Instability, Mismatch Repair
Ivosidenib	IDH1
Lapatinib (1)	ERBB2 (HER2)
Lapatinib (2)	ESR, PGR (Hormone Receptor)
Larotrectinib	NTRK
Letrozole	ESR, PGR (Hormone Receptor)
Lorlatinib	ALK
Midostaurin	FLT3
Neratinib	ERBB2 (HER2)
Nilotinib (1)	BCR-ABL1 (Philadelphia chromosome)
Nivolumab (1)	BRAF
Nivolumab (3)	Microsatellite Instability, Mismatch Repair
Olaparib (1)	BRCA

Drug	Biomarker
Rituximab	MS4A1 (CD20 antigen)
Rucaparib (1)	BRCA
Talazoparib (1)	BRCA
Talazoparib (2)	ERBB2 (HER2)
Tamoxifen (1)	ESR, PGR (Hormone Receptor)
Toremifene	ESR (Hormone Receptor)
Trametinib (1)	BRAF
Trastuzumab (1)	ERBB2 (HER2)
Tretinoin	PML-RARA
Vemurafenib (1)	BRAF
Venetoclax (1)	Chromosome 17p
Vincristine	BCR-ABL1 (Philadelphia chromosome)

FDA – Junio 2019

Pharmacogenomic biomarkers

Cetuximab – RAS – Oncology

Indications and Usage:

Erbitux is indicated for the treatment of **K-Ras wild-type**, epidermal growth factor receptor (EGFR)-expressing, metastatic colorectal cancer (mCRC) as determined by FDA-approved test for this use

Alectinib – ALK – Oncology

Indications and Usage:

ALECENCSA is indicated for the treatment of patients with anaplastic lymphoma kinase **(ALK)-positive** metastatic non-small cell lung cancer (NSCLC) as detected by an FDA-approved test

Chloroquine – G6PD – Infectious Diseases

Adverse effects

Adverse Reactions:

Blood and lymphatic system disorders: Pancytopenia, aplastic anemia, reversible agranulocytosis, thrombocytopenia and neutropenia. Hemolytic anemia in G6PD deficient patients

Drug	Biomarker
Dabrafenib	BRAF
Dacomitinib	EGFR
Dasatinib	BCR-ABL1 (Philadelphia chromosome)
Denileukin Diftitox	IL2RA (CD25 antigen)
Enasidenib	IDH2
Encorafenib	BRAF
Erlotinib	EGFR

Drug	Biomarker
Lorlatinib	ALK
Midostaurin	FLT3
Neratinib	ERBB2 (HER2)
Nilotinib (1)	BCR-ABL1 (Philadelphia chromosome)
Nivolumab (1)	BRAF
Nivolumab (3)	Microsatellite Instability, Mismatch Repair
Olaparib (1)	BRCA

Drug	Biomarker
Trastuzumab (1)	ERBB2 (HER2)
Tretinoin	PML-RARA
Vemurafenib (1)	BRAF
Venetoclax (1)	Chromosome 17p
Vincristine	BCR-ABL1 (Philadelphia chromosome)

FDA – Junio 2019

Pharmacogenomic biomarkers

Cetuximab – RAS – Oncology

Indications and Usage:

Erbitux is indicated for the treatment of **K-Ras wild-type**, epidermal growth factor receptor (EGFR)-expressing, metastatic colorectal cancer (mCRC) as determined by FDA-approved test for this use

Alectinib – ALK – Oncology

Indications and Usage:

ALECENCSA is indicated for the treatment of patients with anaplastic lymphoma kinase **(ALK)-positive** metastatic non-small cell lung cancer (NSCLC) as detected by an FDA-approved test

Chloroquine – G6PD – Infectious Diseases

Adverse Reactions:

Blood and lymphatic system disorders: Pancytopenia, aplastic anemia, reversible agranulocytosis, thrombocytopenia and neutropenia. Hemolytic anemia in G6PD deficient patients

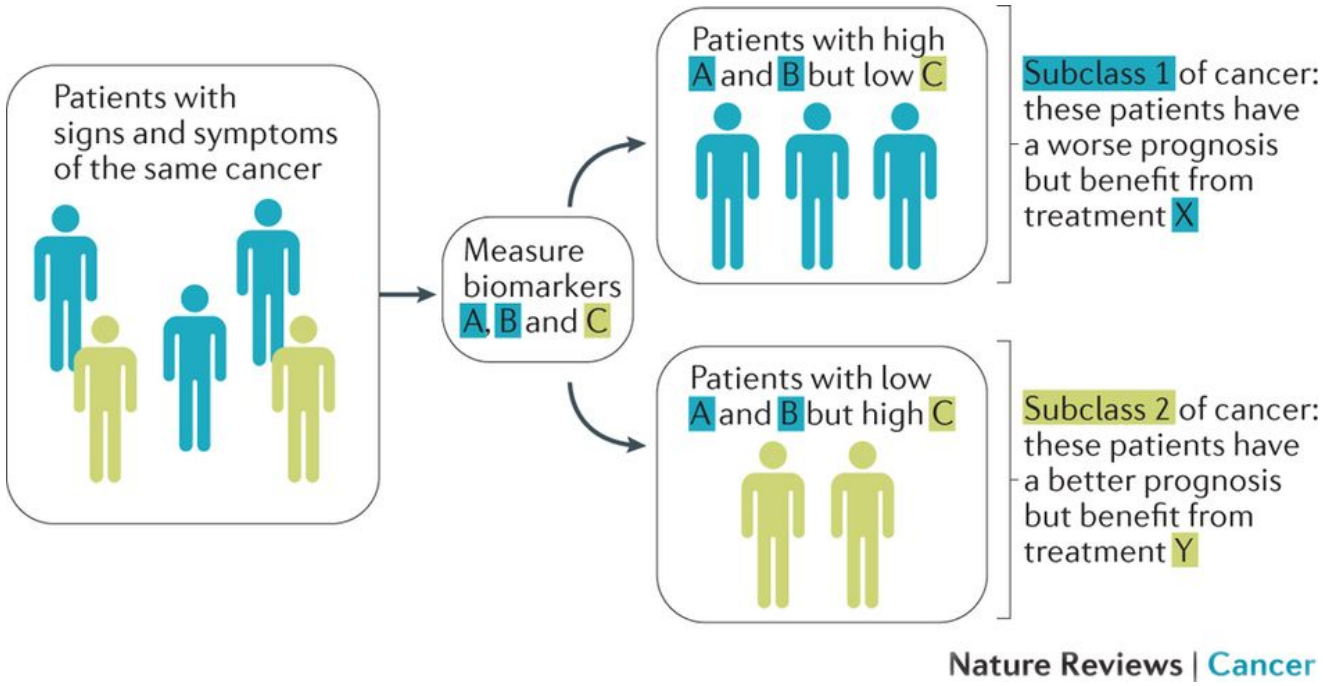
Amifampridine Phosphate – NAT2 – Neurology

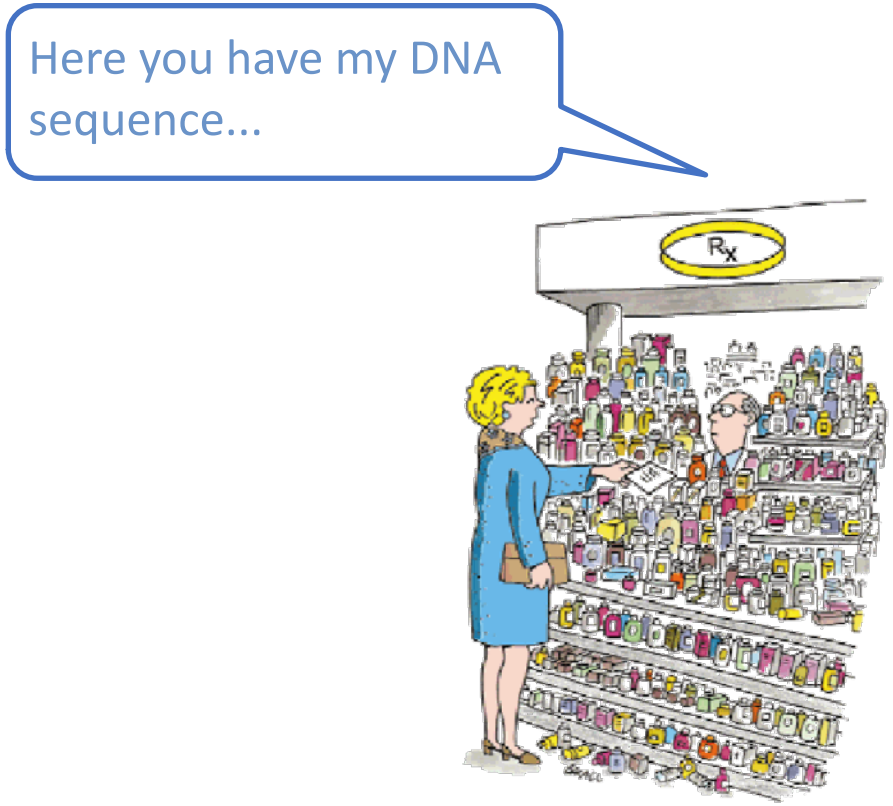
Dose adjustment

Dosage and Administration:

The recommended starting dosage of FIRDAPSE in known N-acetyltransferase 2 **(NAT2) poor metabolizers** is 15 mg daily, taken orally in 3 divided doses

Personalized Medicine

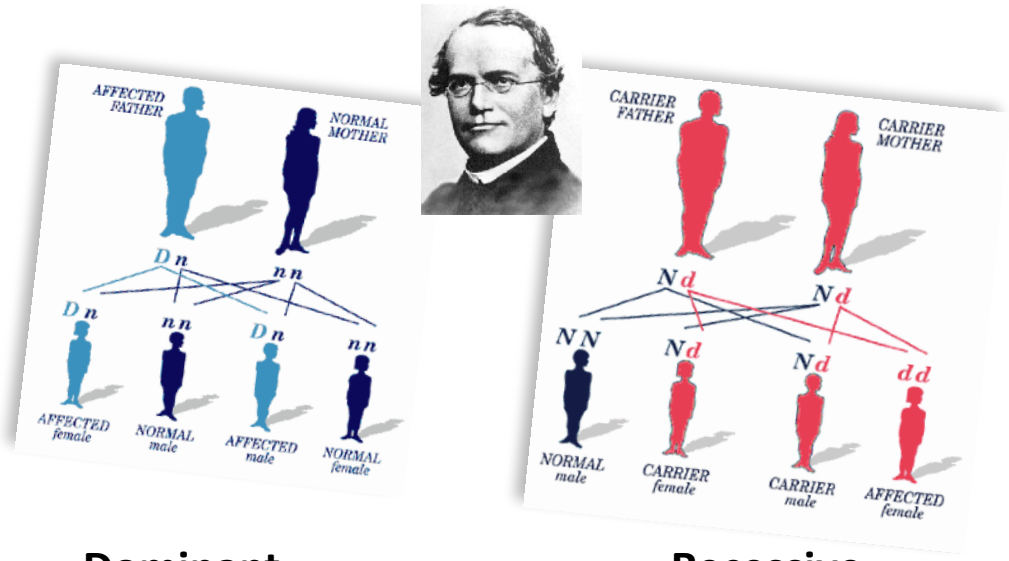




Clinical implications of genomic variants

Monogenic disorders

- Caused by the alteration in one gene
- Follow the **mendelian** inheritance rules



Dominant

e.g.: Huntington disease

(HTT gene)

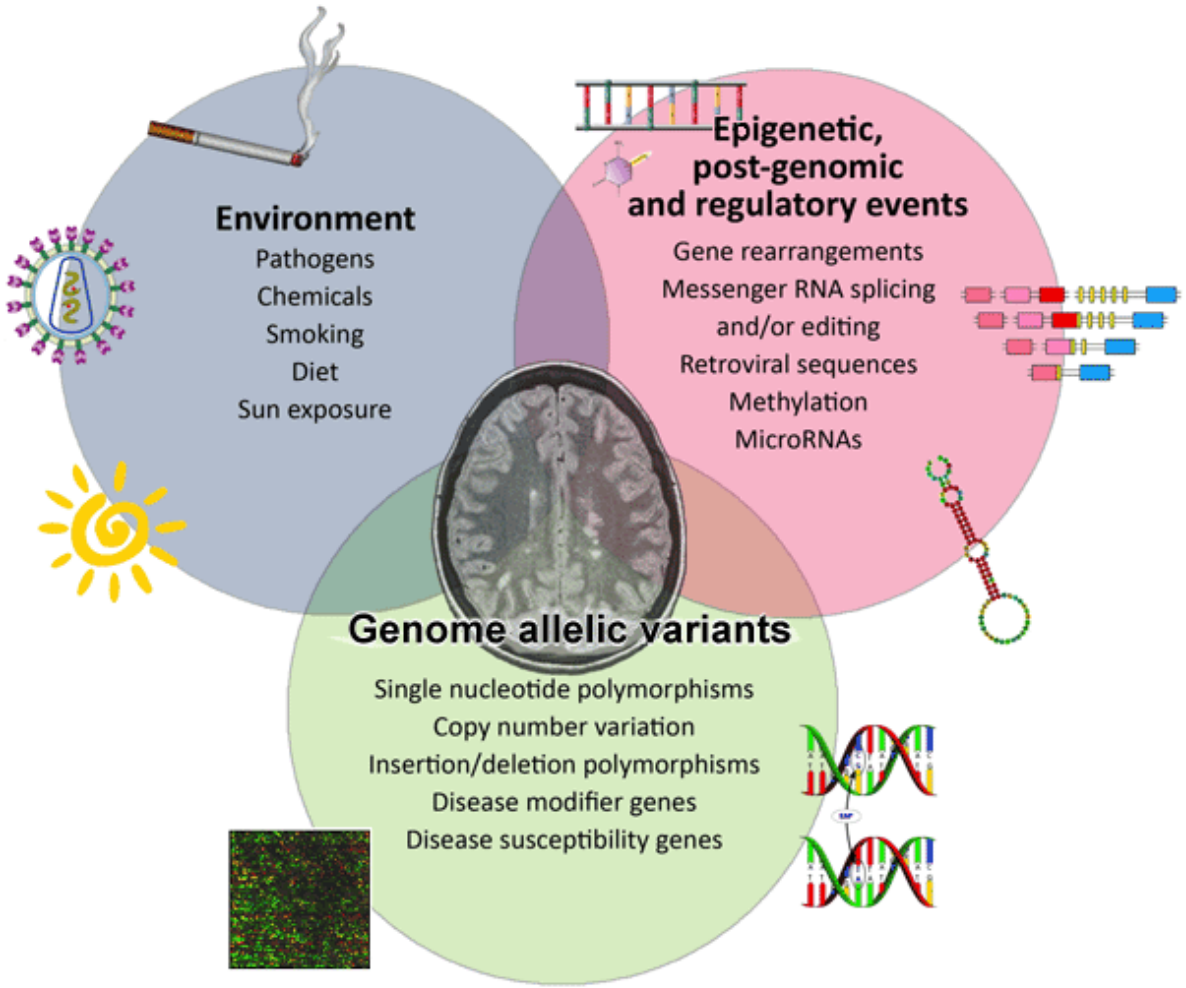
Recessive

e.g.: Cystic fibrosis

(CFTR gene)

Complex disorders

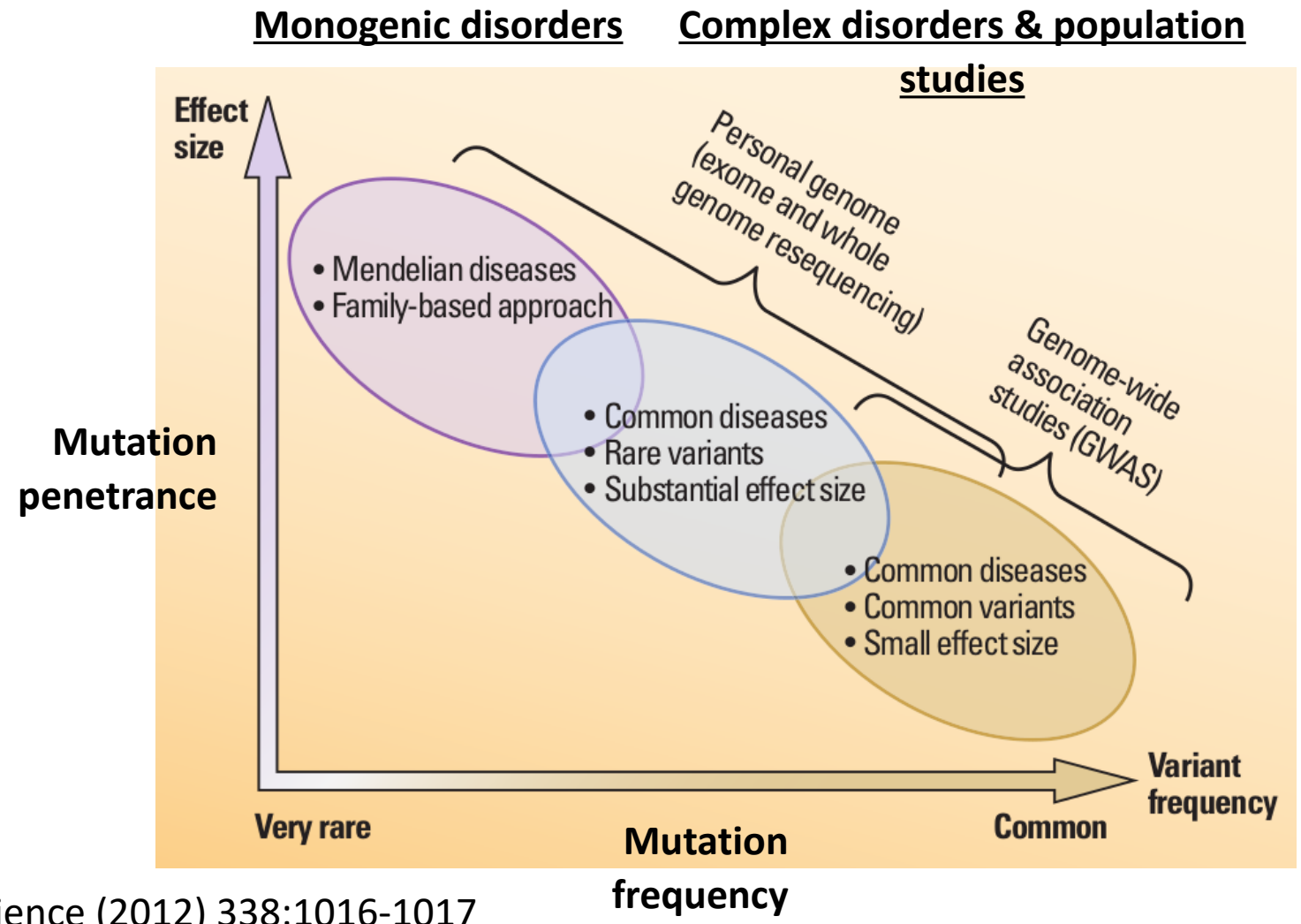
- Caused by the combination of genetic, epigenetic, environmental and lifestyle factors.
- Do not obey the standard mendelian patterns of inheritance



E.g.: **Congenital disorders** (neural tube disorders, harelip, congenital heart diseases) and **Adulthood disorders** (coronary diseases, mellitus diabetes, cancer, etc.)

Clinical implications of genomic variants

Variants in genetics disorders

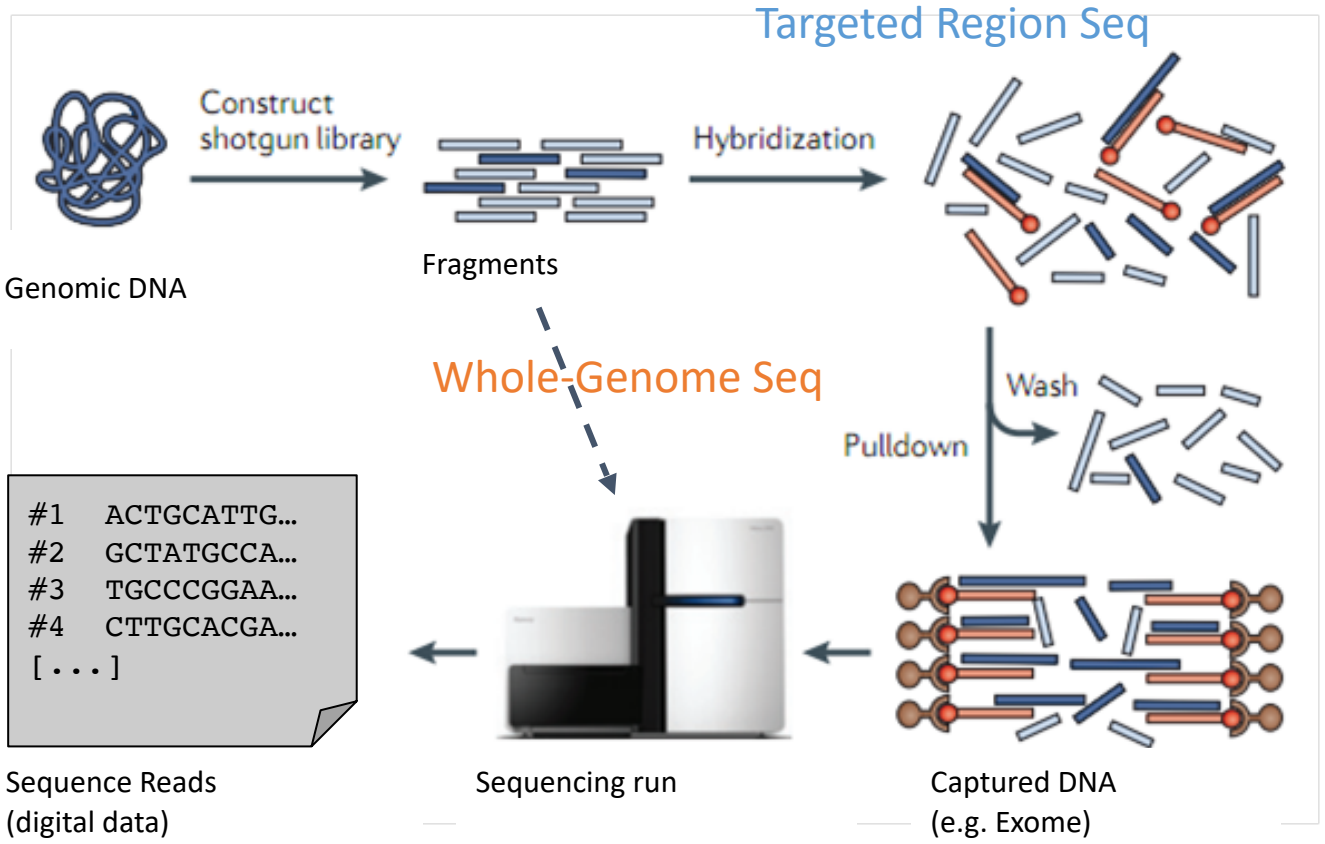


Kaiser J. Science (2012) 338:1016-1017

Variant detection with NGS

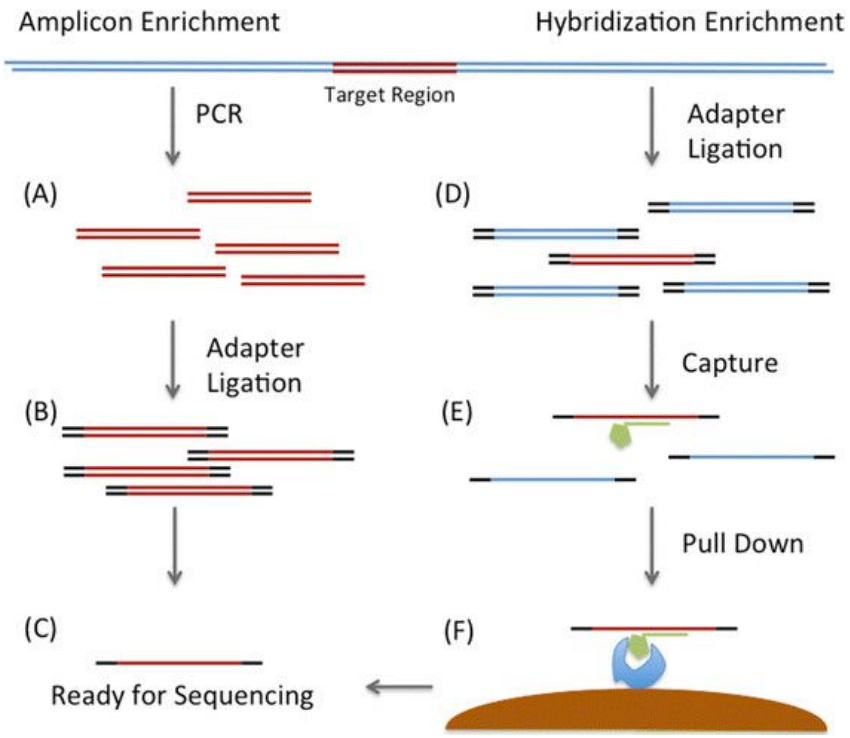
Sequencing and bioinformatics process

DNA Sequencing data generation



The **sequence reads** are the sequence of the ends of the original fragment.

Chemistries employed for targeted DNA-based NGS



	Amplicon	**Hybrid**
Advantages	Low DNA input requirement High sensitivity	Evaluation of structural variants
Disadvantages	More difficult structural variant assessment	Higher input DNA requirements Longer turnaround time

Different NGS technologies have different capabilities

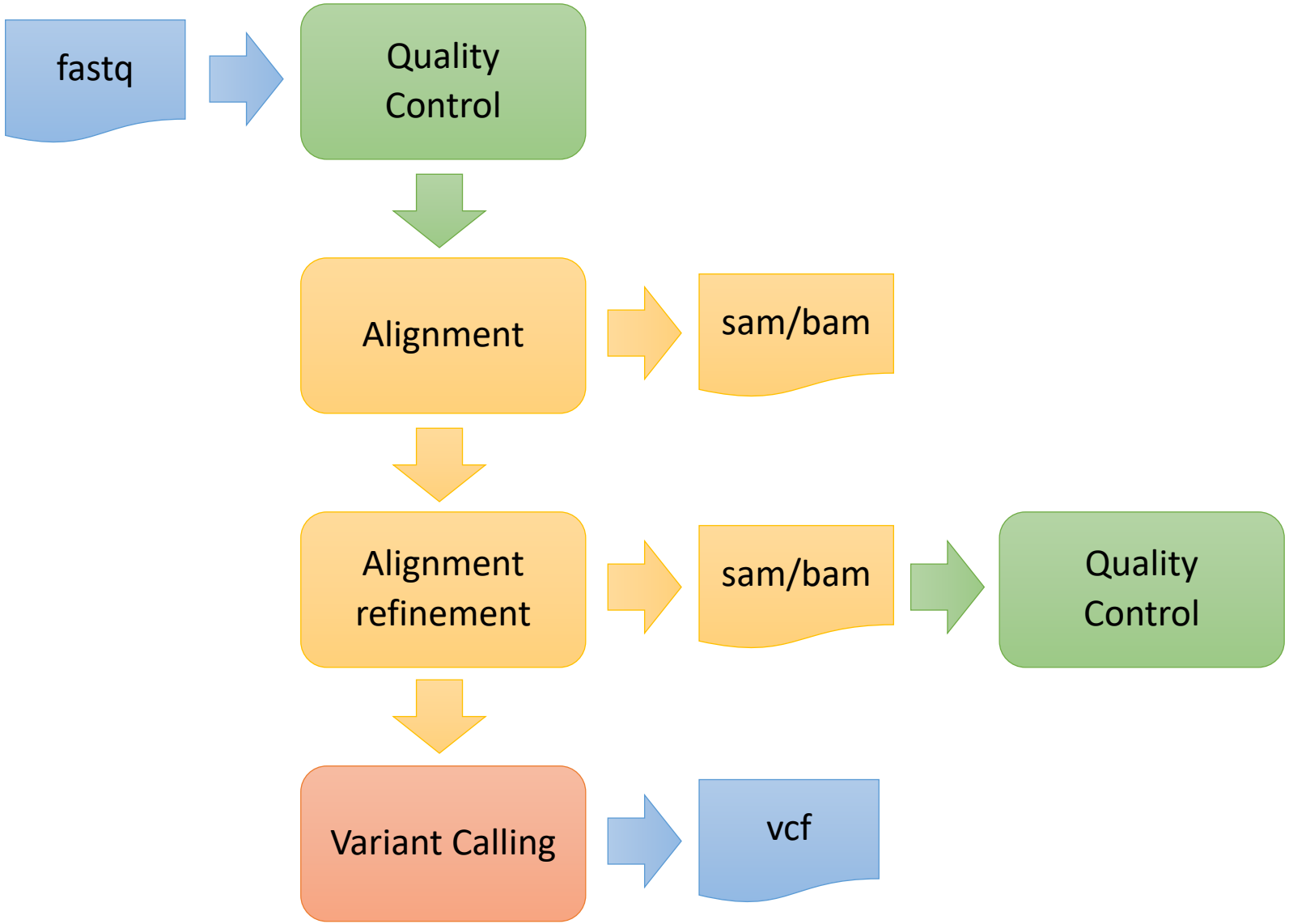
Table 1 | **Main characteristics of current NGS technologies**

Technology	Run type			Maximum read length	Quality scores	Error rates	Refs
	Single end	Paired end	Mate pair				
Illumina	Yes	Yes	Yes	300 bp	>30	0.0034–1%	59
SOLiD	Yes	Yes	Yes	75 bp	>30	0.01–1%	60
IonTorrent	Yes	Yes	No	400 bp	~20	1.78%	22
454	Yes	Yes	No	~700 bp (up to 1 kb)	>20	1.07–1.7%	53,61
Nanopore	Yes	No	No	5.4–10 kb	NA	10–40%	62–66
PacBio	Yes	No	No	~15 kb (up to 40 kb)	<10	5–10%	22,67–69

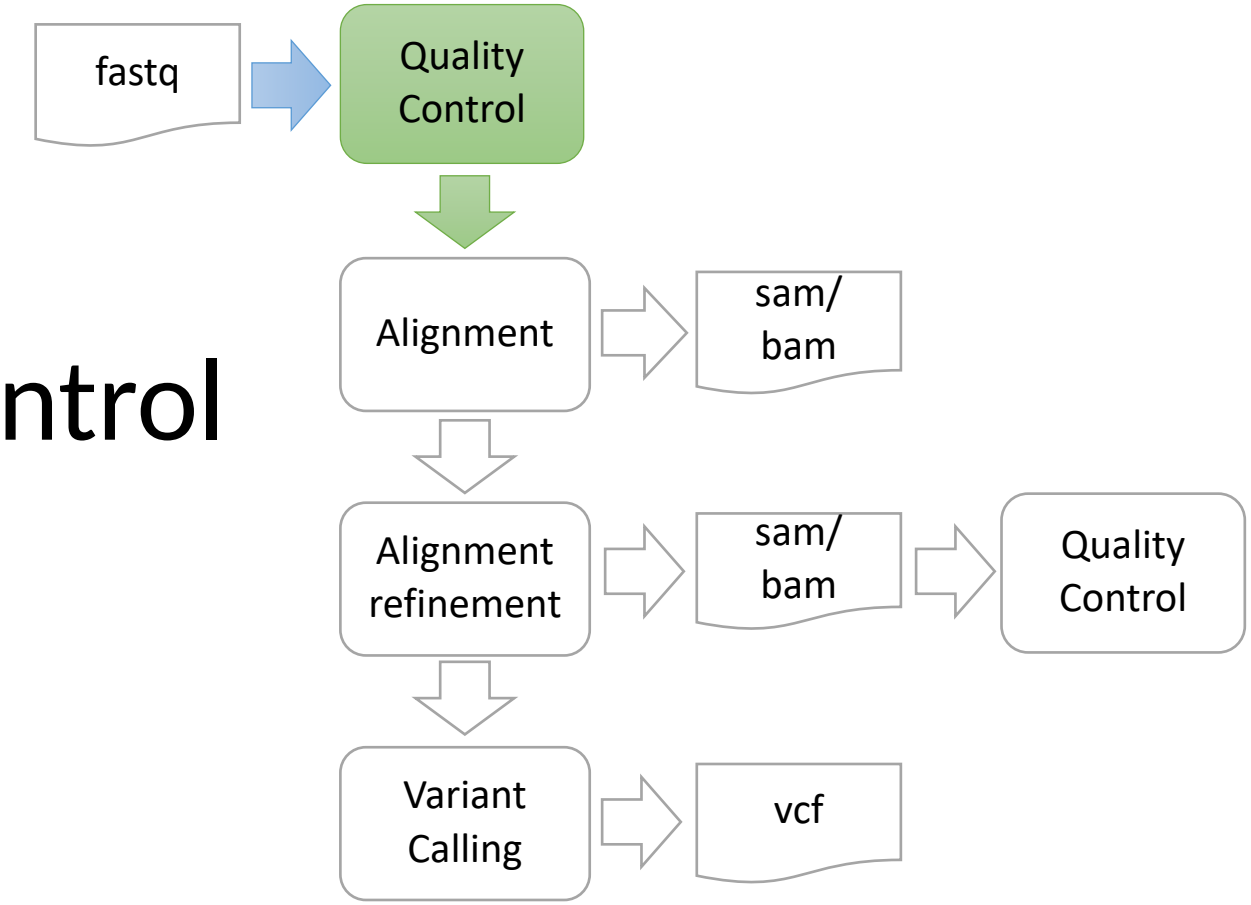
454, 454 pyrosequencing (Roche); NA, not applicable; Nanopore, Oxford Nanopore Technologies; NGS, next-generation sequencing; PacBio, Pacific Biosciences; SOLiD, sequencing by oligonucleotide ligation and detection (Thermo Fisher).

Nature Reviews Genetics 17,459–469(2016) doi:10.1038/nrg.2016.57

General steps



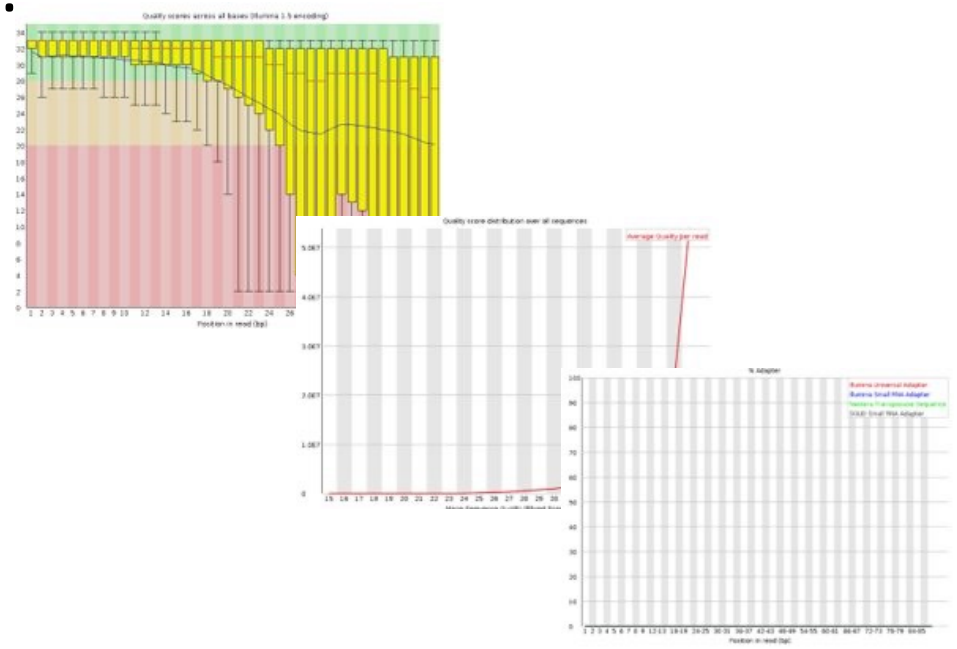
Raw data Quality Control



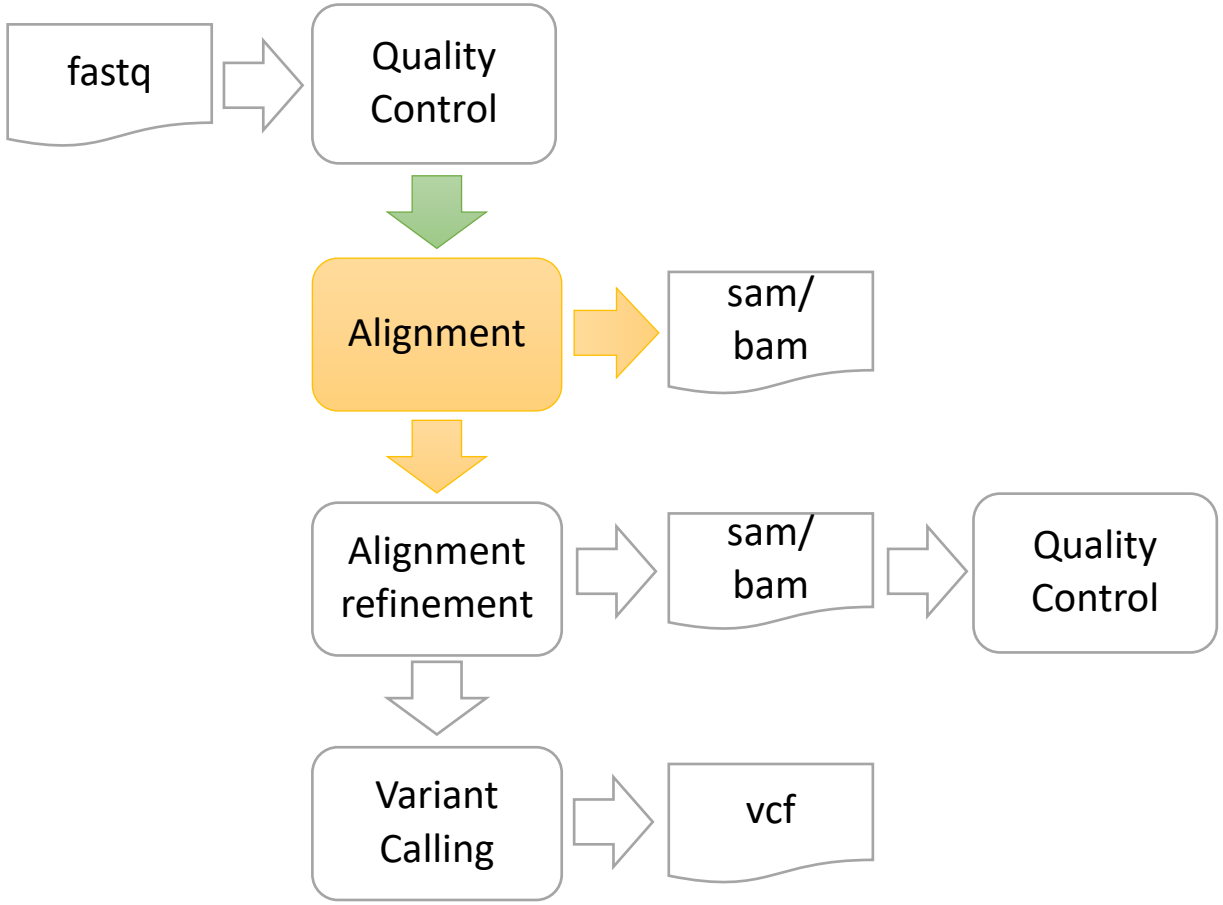
Quality control

- Is the read number in the expected range?
- Have bases a good quality?
- Is there any contamination in the samples?

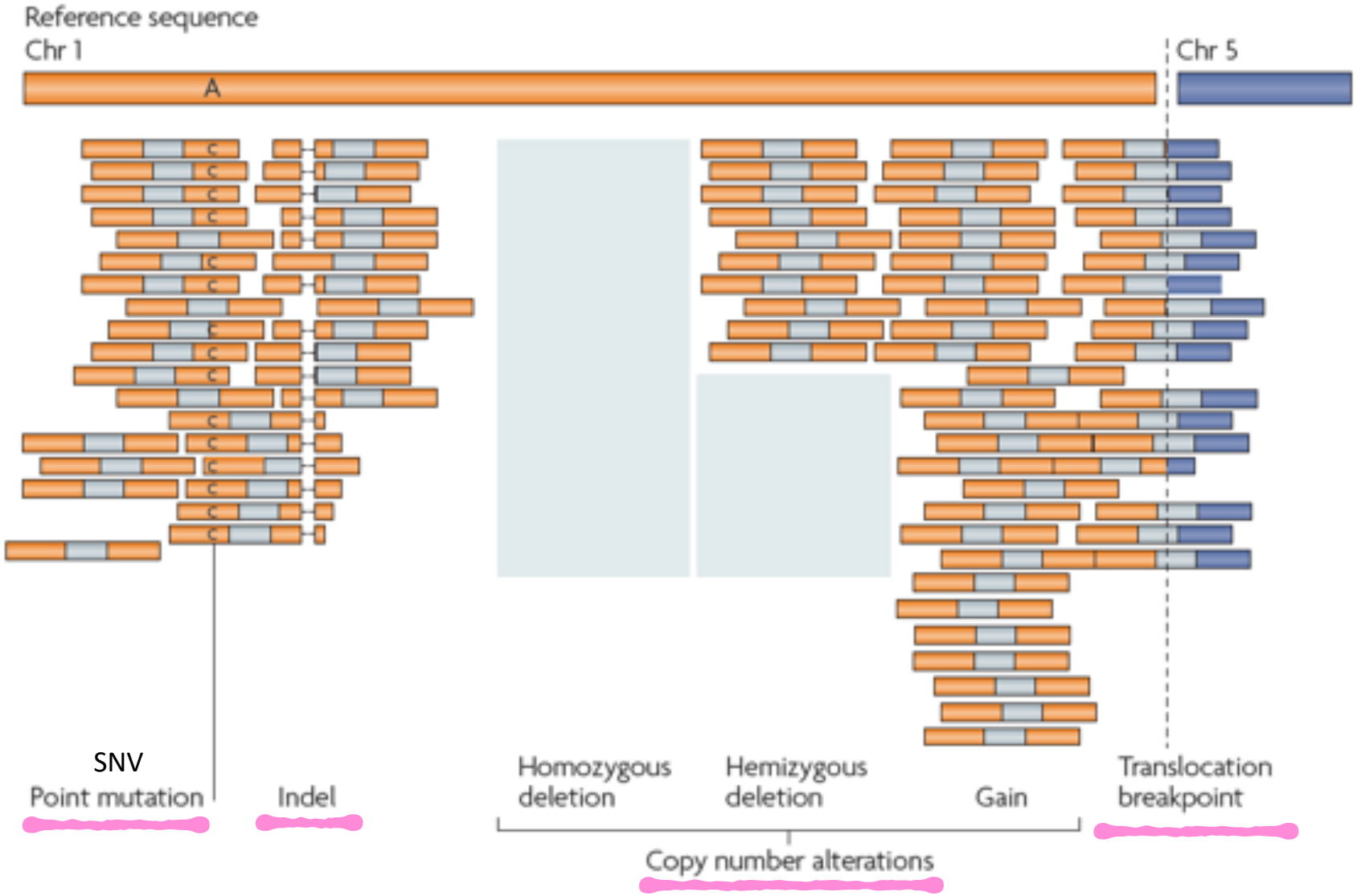
- Software:
  - FastQC
  - FastQ Screen



Alignment


fastq
Quality Control
Alignment
sam/ bam
Alignment refinement
sam/ bam
Quality Control
Variant Calling
vcf

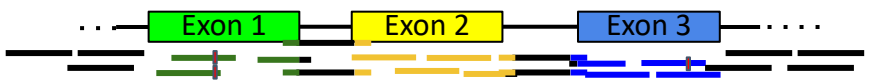
Alignment of reads uncover potential variant sites



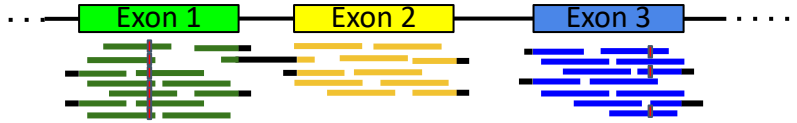
Variant detection with NGS

DNA-seq strategies – Sequencing coverage

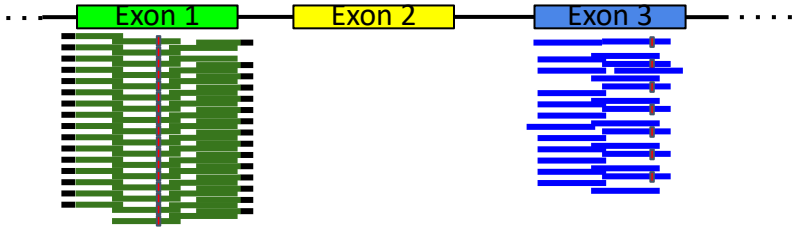
Whole-Genome Sequencing

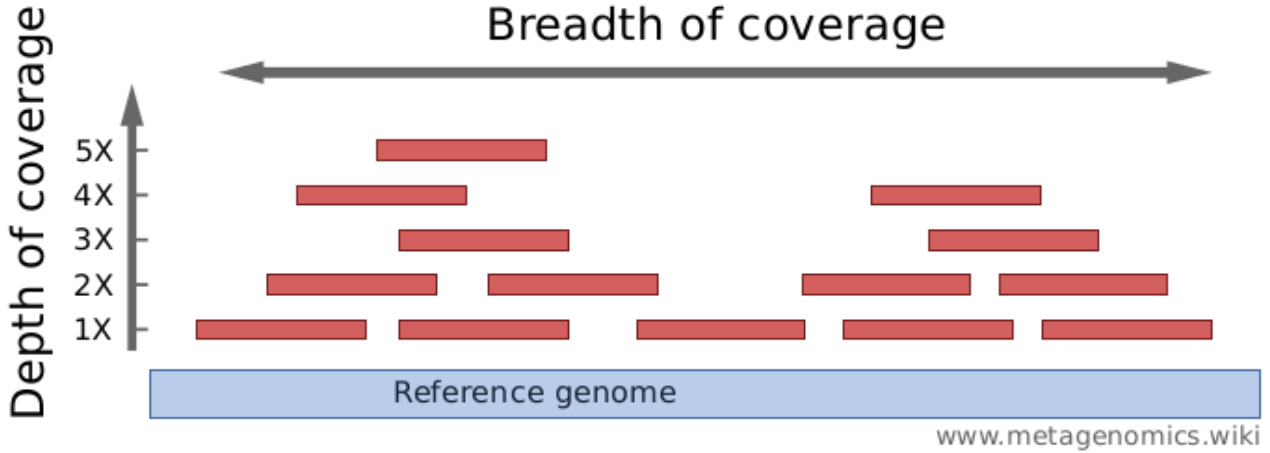


Whole-Exome Sequencing



Panel of genes by amplicon/PCR approach





Depth: times that a base is sequenced

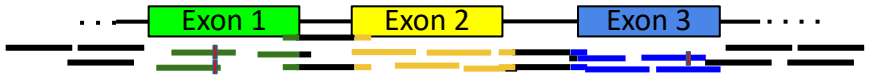
On average:

$$\frac{\sum \textit{Number of times a sequenced base is covered by reads}}{\textit{Length of the sequenced genome}}$$

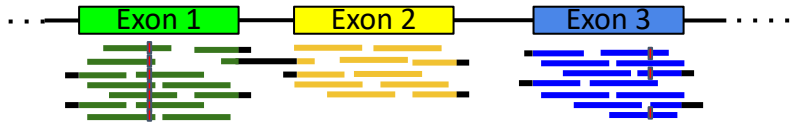
Breadth: percentage of the sequenced genome covered by the reads (at a certain depth)

DNA-seq strategies – Type of variants

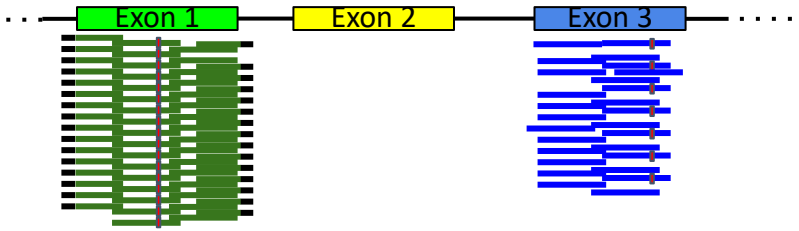
Whole-Genome Sequencing



Whole-Exome Sequencing



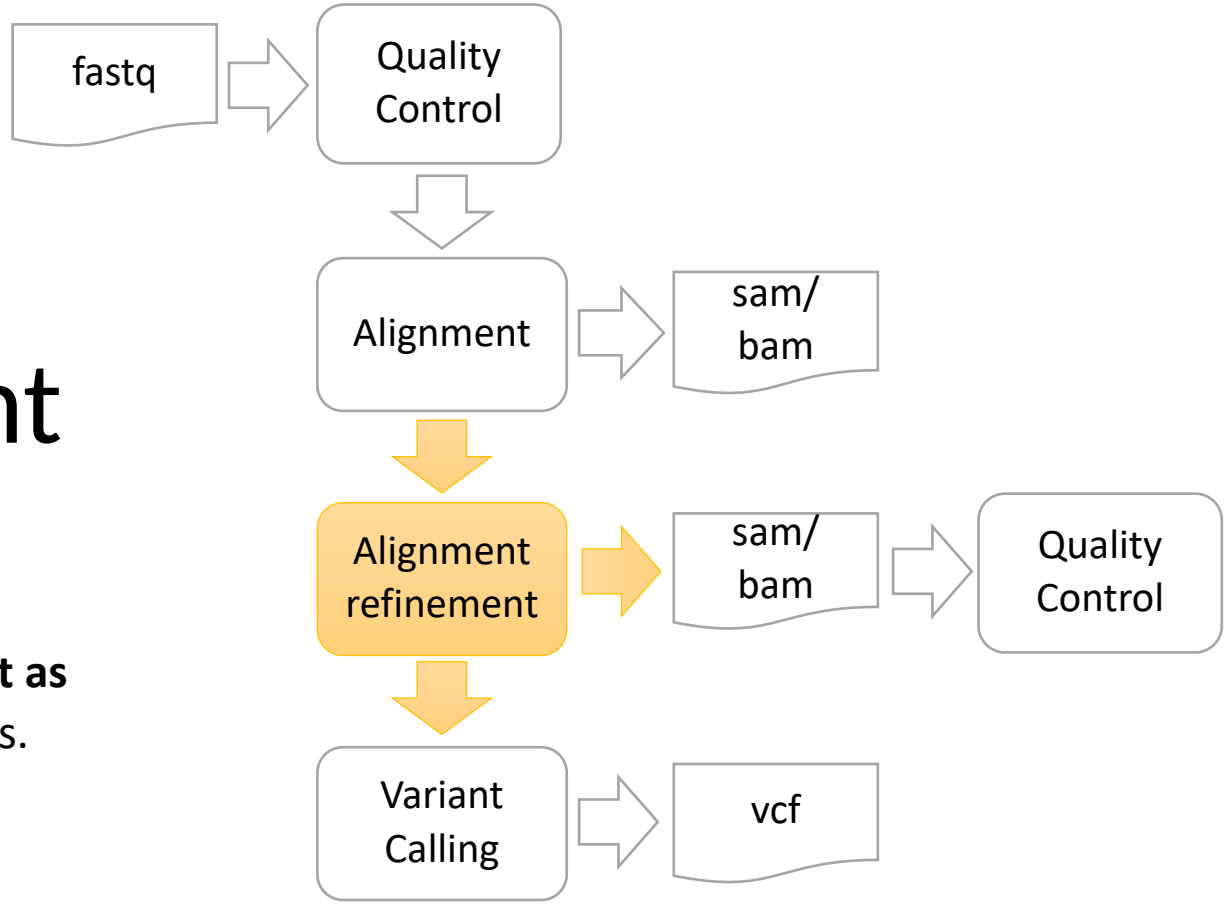
Panel of genes by amplicon/PCR approach



Target	Type of variants discovered	Avg depth per pos	Cost
Entire genome	Coding variants, intronic and regulatory sites. Structural variants #Variants= 3M - 4M.	> 30x Most uniform	High
2% of the genome	Coding variants Some intronic and regulatory sites. Issues in the detection of structural variants **#**Variants= 20k - 60k.	> 50x - 100x	Low
Variable	Depends on the design (customizable) Challenging detection of structural variants # variants = *ND*	> 500x	Lower

Alignment refinement

Variant calling requires the most perfect alignment as possible to avoid False Positives and False Negatives.



Mark/remove duplicates

- Duplicates derive from **PCR amplification** (library preparation): one fragment is sequenced multiple times.
- An error at the beginning of the PCR (first steps) is propagated.

- Duplicates in hybrid-seq are **worthless** for the subsequent analysis:
  Duplicates are source of False Positives calls while only provide redundancy.

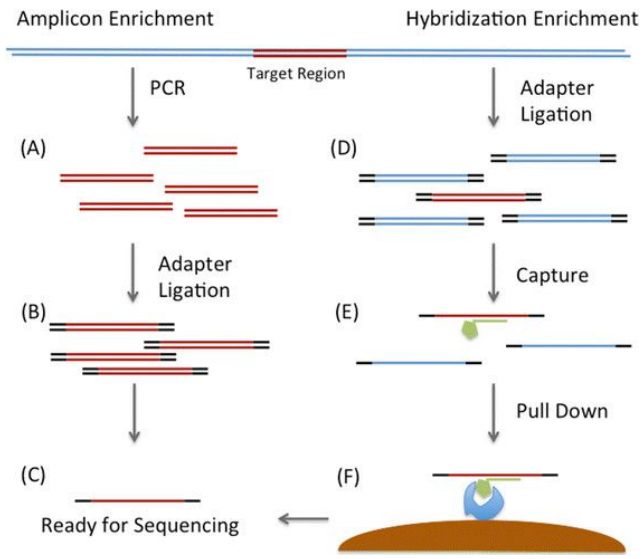
Solution: retrieve the best one, discard the duplicates:

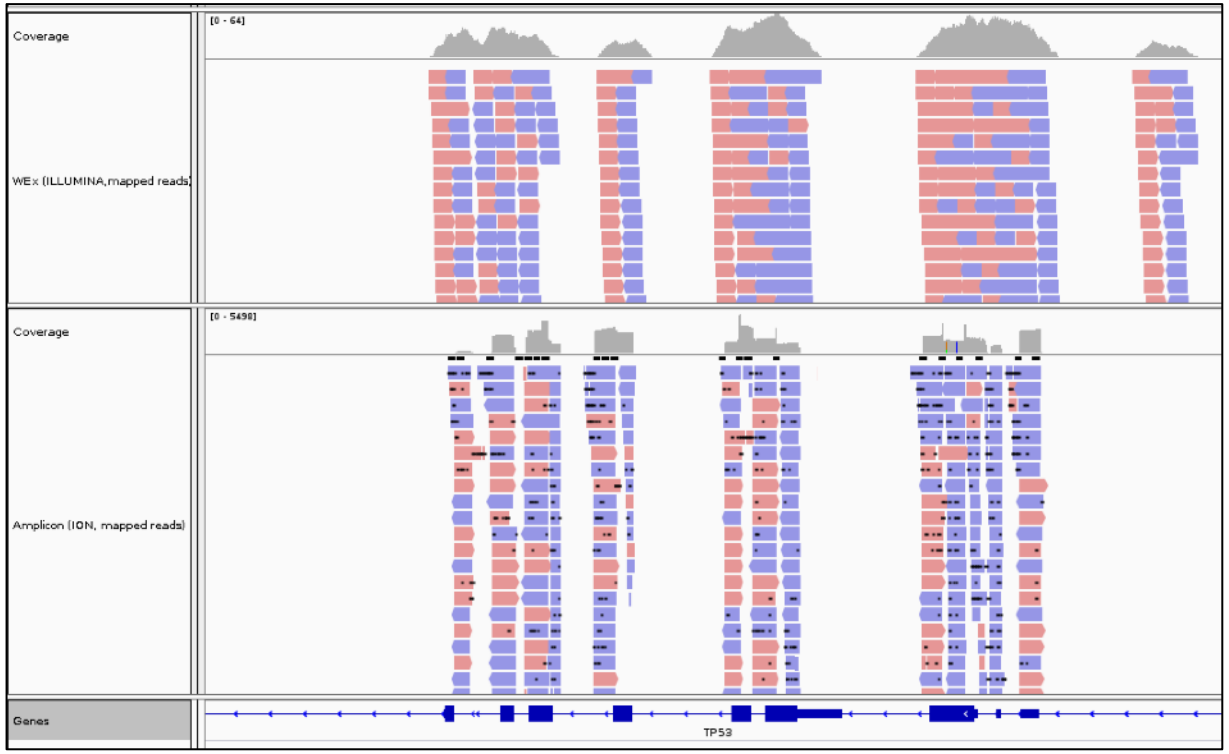
METHOD: by **Picard-tools**
http://broadinstitute.github.io/picard/
(alternatives : **samtools**)

Duplicates share the **same alignment properties** : sequence, start and end positions

Adapted from GATK

Mark/remove duplicates: Amplicon seq





WARNING: **Do NOT remove duplicates in data derived from** amplicon techniques (**Ion Torrent**).
More info.: https://github.com/broadgsa/gatk/blob/master/doc_archive/tutorials/(How_to)_Mark_duplicates_with_MarkDuplicates_or_MarkDuplicatesWithMateCigar.md

Indel realignment

- Algorithms align reads very fast with high accuracy, but not perfectly.

  During alignment, ***penalties on mismatches are much cheaper than gaps*** *(indels).*

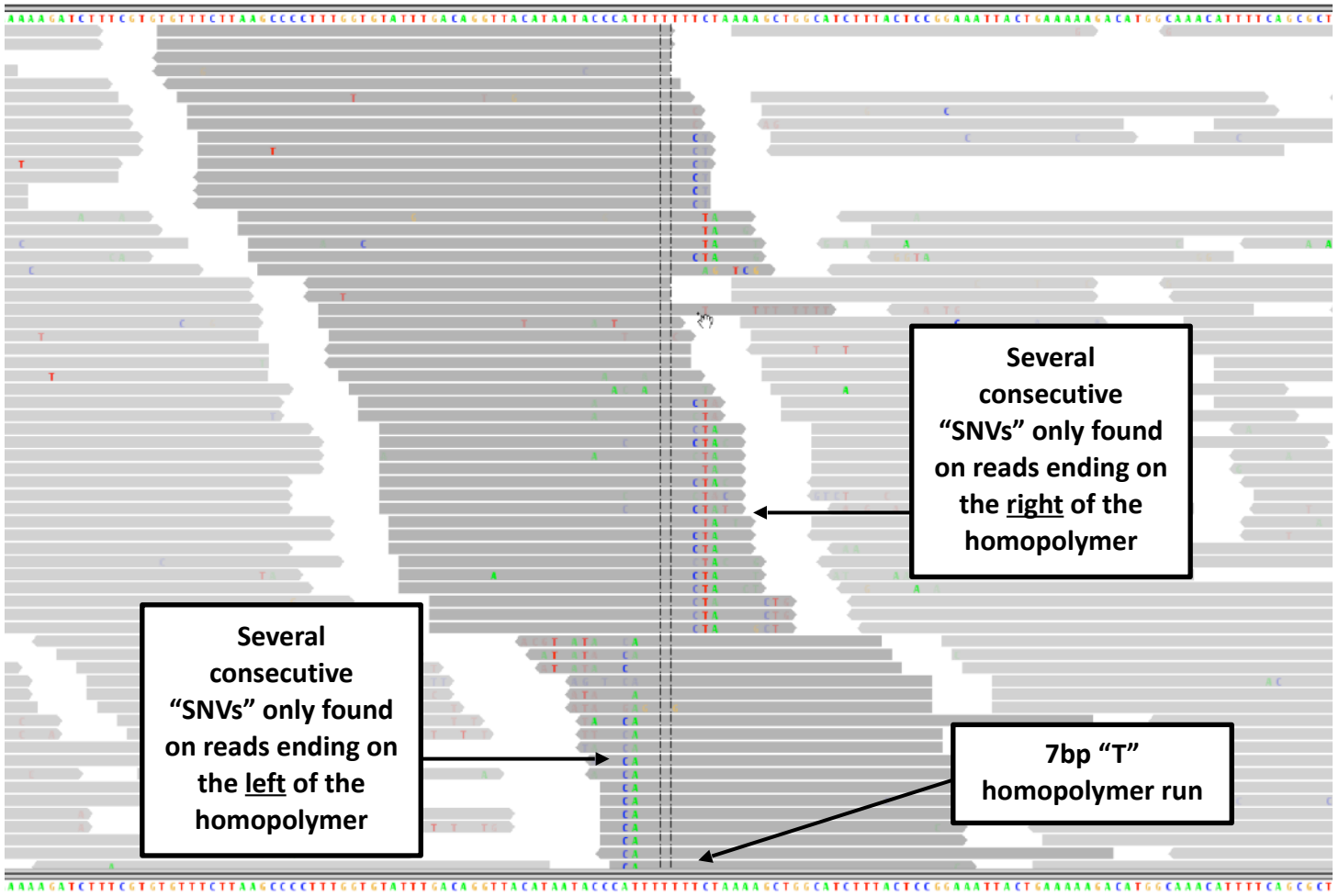
- Also, there are sometimes multiple solutions (alignments) for a given read. **Aligners can choose one randomly**.

- Reads are aligned separately (one by one).

- **Indels can be no properly identified** in the alignment of the read.

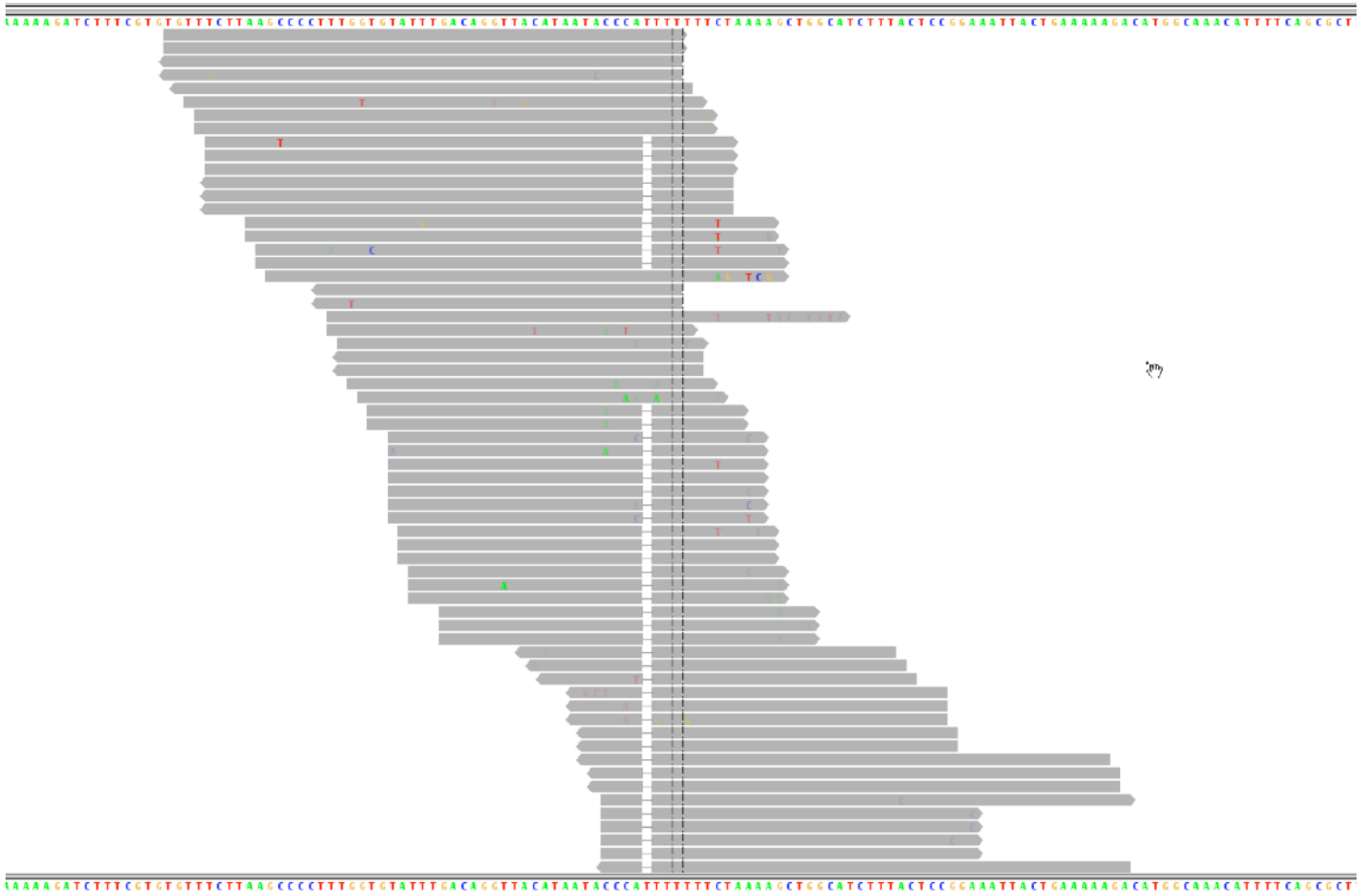
METHOD: by **GATK**
https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-tutorials/(howto)_Perform_local_realignment_around_indels.md

Indel realignment



Taken from GATK team

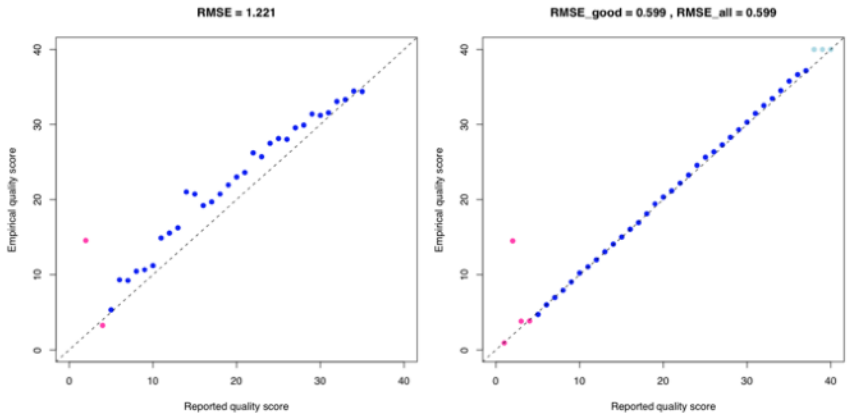
Indel realignment



Taken from GATK team

Base Quality Score Recalibration

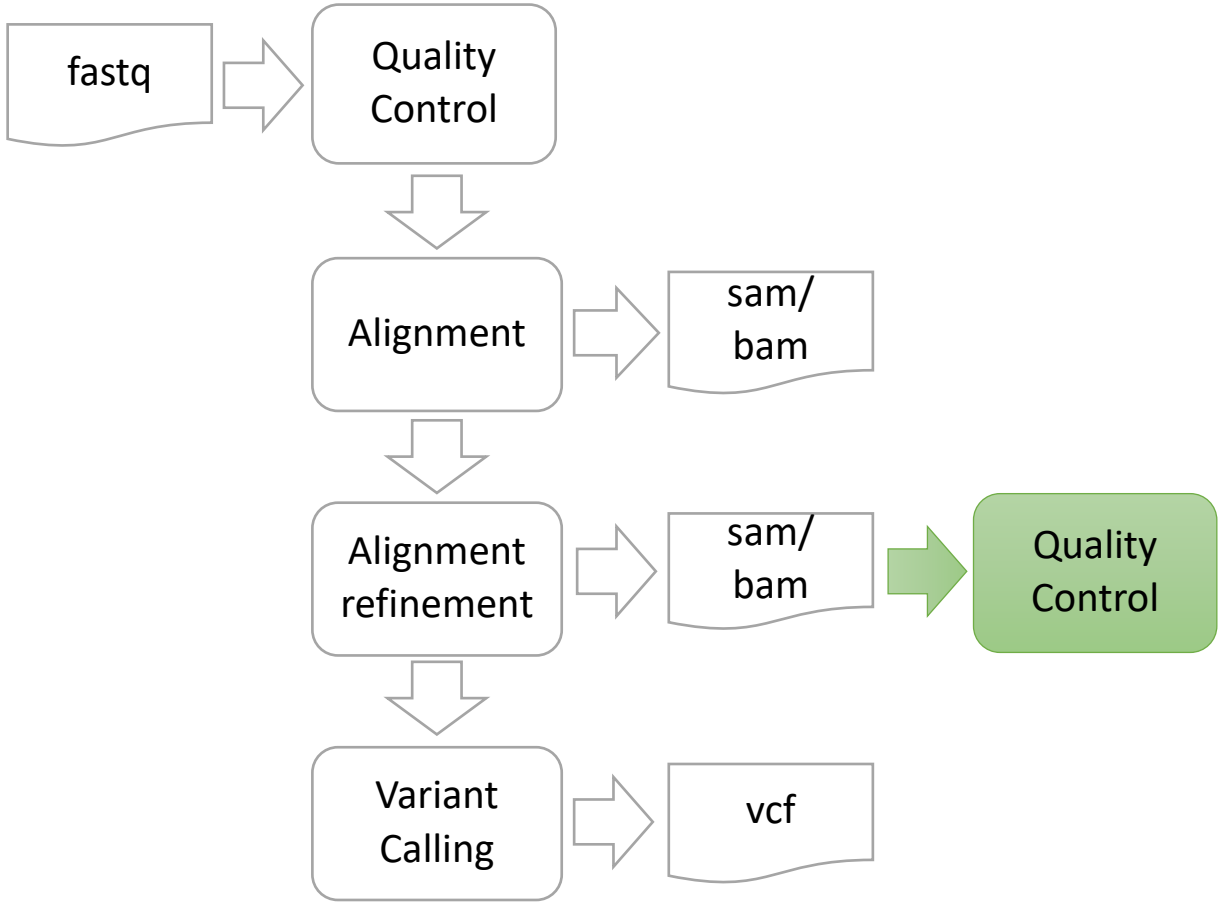
- **Phred Quality score**: each position of the sequence has its particular **base Quality score**.
- The individual quality measures **are crucial during Variant calling**.
- Different NGS technologies have their particular **bias in Quality Score** depending on the context. Recalibration **correct empirically** these biases.



Original After BQSR recalibration

METHOD: by **GATK**
https://gatk.broadinstitute.org/hc/en-us/articles/360035890531-Base-Quality-Score-Recalibration-BQSR-

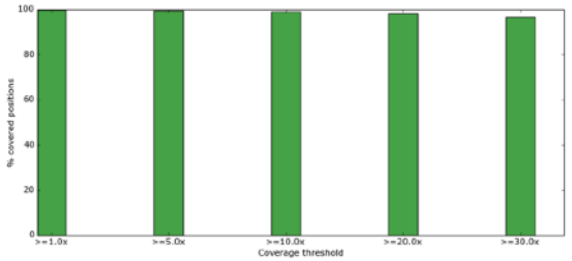
Alignment Quality Control

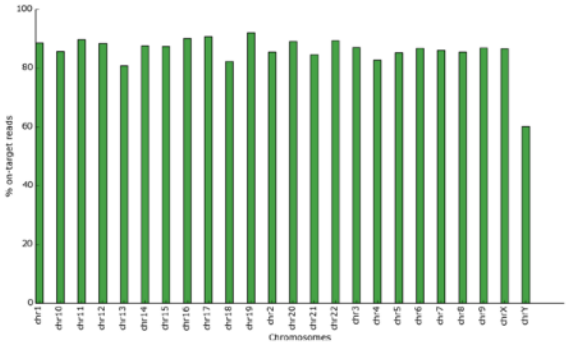


Alignment Quality Control

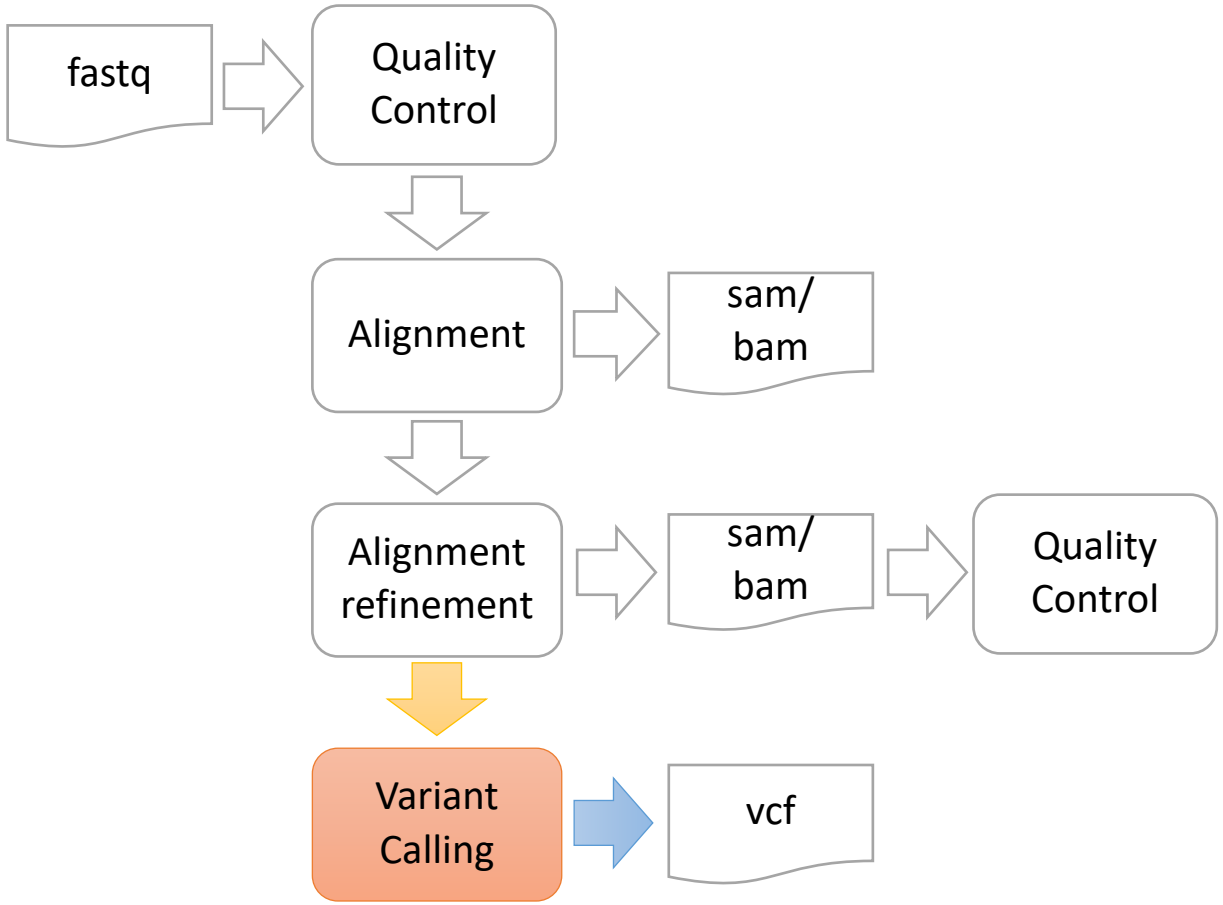
- Mean sequencing depth
- Is there enough coverage in regions of interest?
- Are the reads on-target?

- Software:
  - ngsCAT
  - QualiMap

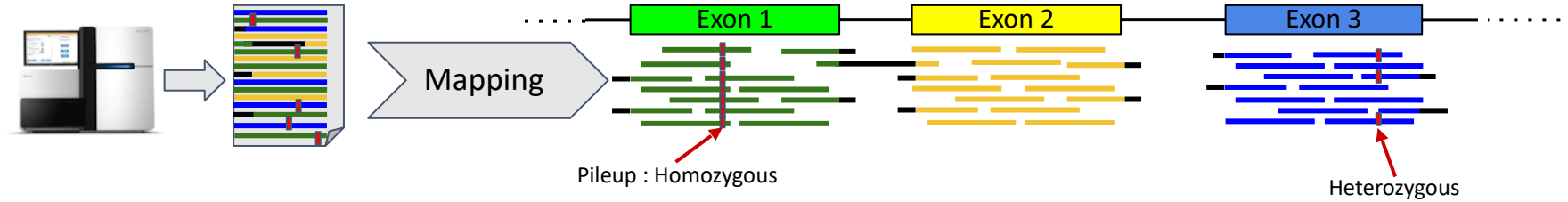




Variant calling



Fundamentals of Variant Calling



1 Identify the most likely **genotype for each genomic position** using statistical methods.

2 **Identify the differences** by comparing with the reference genome.

What is Crucial in Variant calling

- For clinical practices, the use of **gold standard methods** and **reproducible analysis** are mandatory.

- The analysis is based on the comparison against the reference genome:

  A single consensus sequence for the whole genome. It was built up from a high quality set of representative samples of the species (from different populations).
  It is the first-line comparison during analysis.

  By **Genome Reference Consortium (GRC)** (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/)

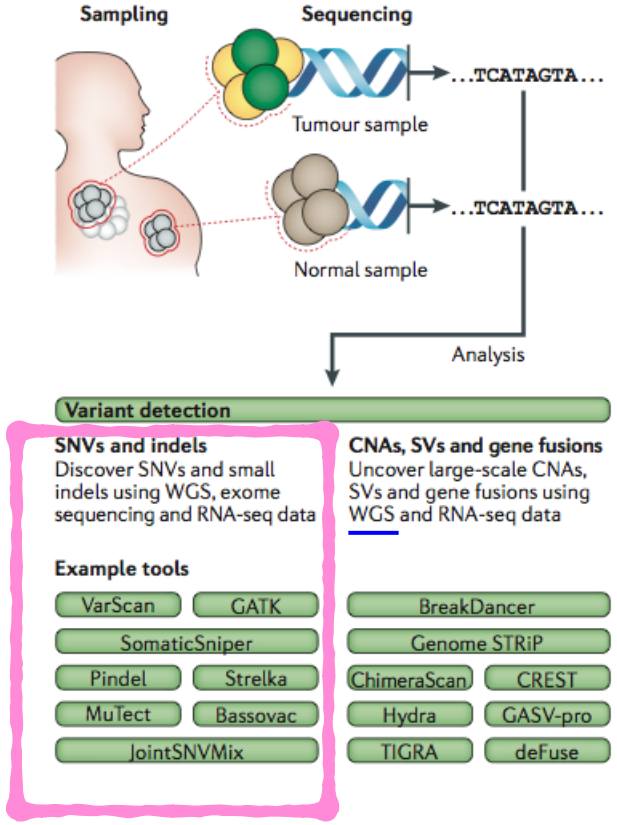
  - **Human assemblies (Versions):**
    + **GRCh37/hg19** : former version. Released in 2012. It is still used for analysis.
    + **CRCh38/hg38** : current version (Sep. 2017). Released in 2014. More accurate, comprehensive (includes Haplotypes) and sophisticated.

We must **keep consistency in the Genome Reference Version** through the variant analysis.

- We must know what **regions along the genome were sequenced** in the experiment, that is, the sequencing library.

Algorithms for Variant Calling

SNVs and Indels



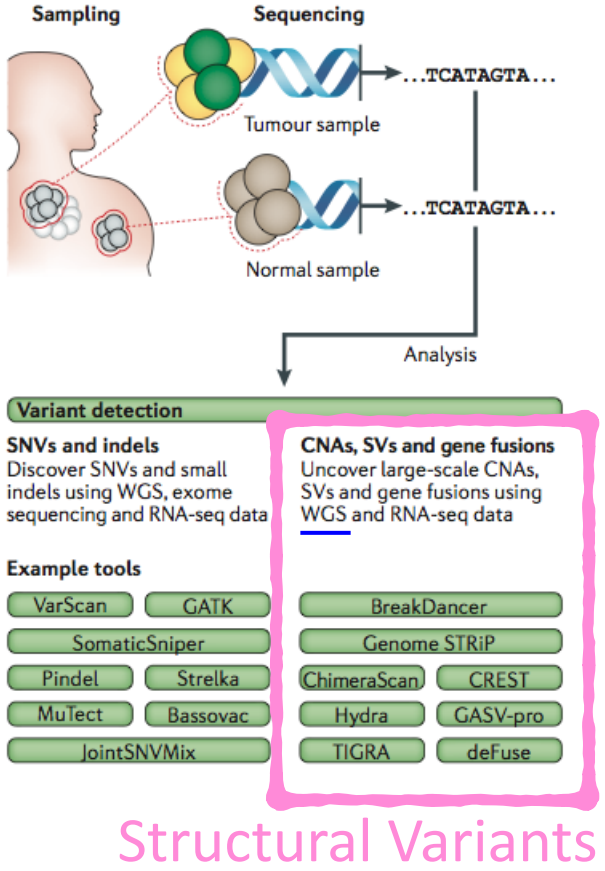
Several Methods have been published.

Further reading:

Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Nat Rev Genet – (2014). doi:10.1038/nrg3767

Nam J.N. et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics (2015)

Algorithms for Variant Calling



Structural Variants

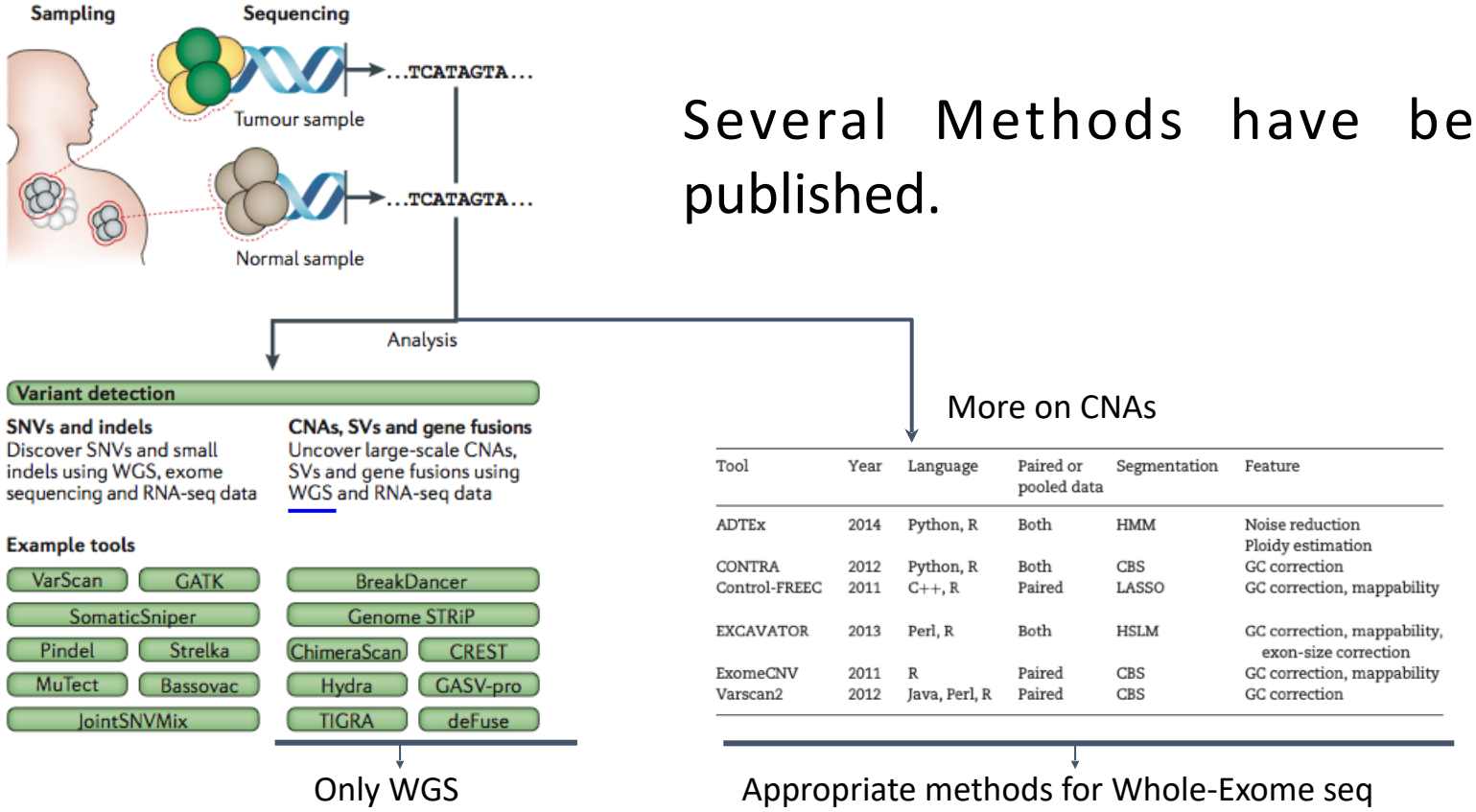
Several Methods have been published.

Further reading:

Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Nat Rev Genet – (2014). doi:10.1038/nrg3767

Nam J.N. et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics (2015)

Algorithms for Variant Calling



Several Methods have been published.

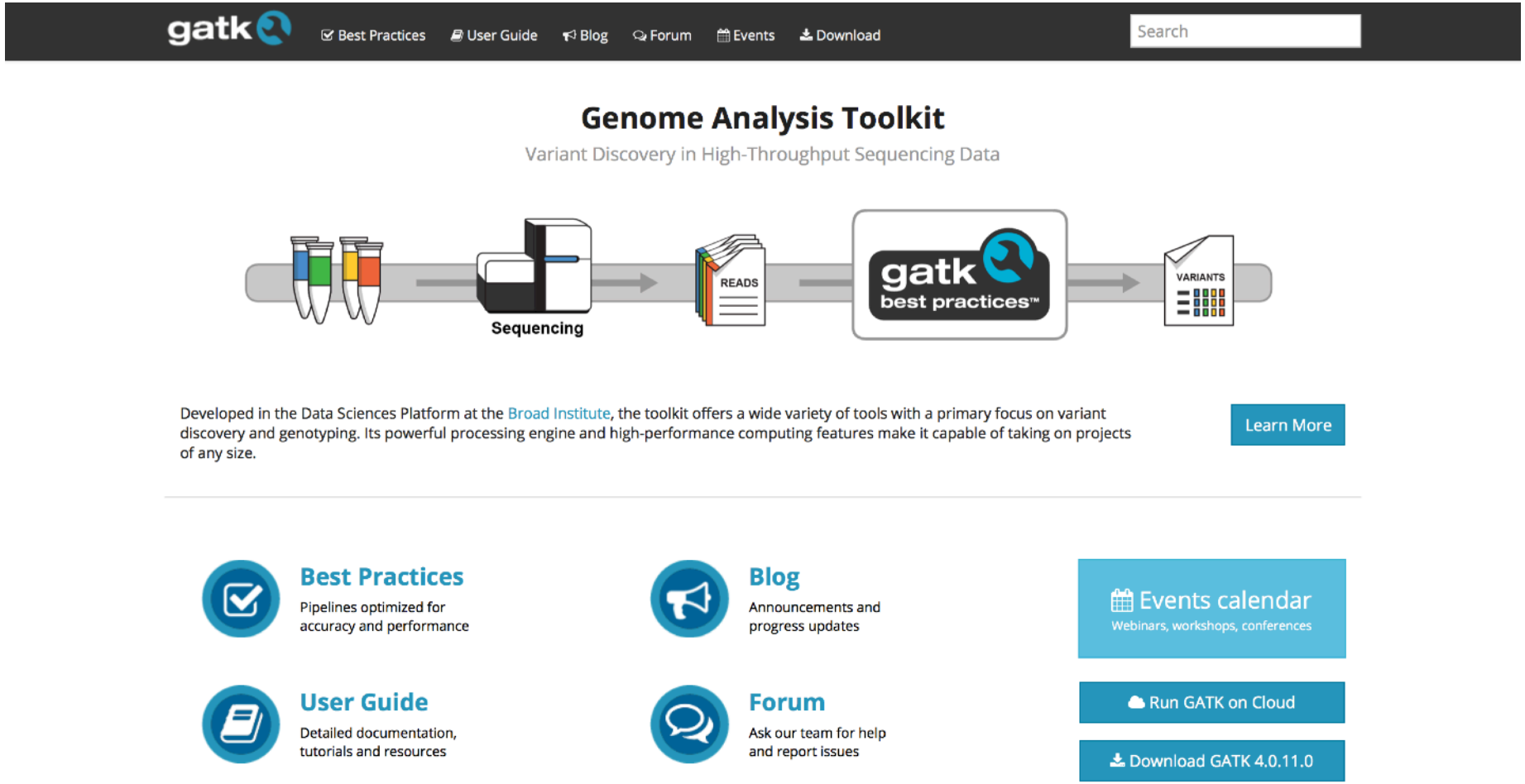
Tool	Year	Language	Paired or pooled data	Segmentation	Feature
ADTEx	2014	Python, R	Both	HMM	Noise reduction Ploidy estimation
CONTRA	2012	Python, R	Both	CBS	GC correction
Control-FREEC	2011	C++, R	Paired	LASSO	GC correction, mappability
EXCAVATOR	2013	Perl, R	Both	HSLM	GC correction, mappability, exon-size correction
ExomeCNV	2011	R	Paired	CBS	GC correction, mappability
Varscan2	2012	Java, Perl, R	Paired	CBS	GC correction

Further reading:

Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Nat Rev Genet – (2014). doi:10.1038/nrg3767

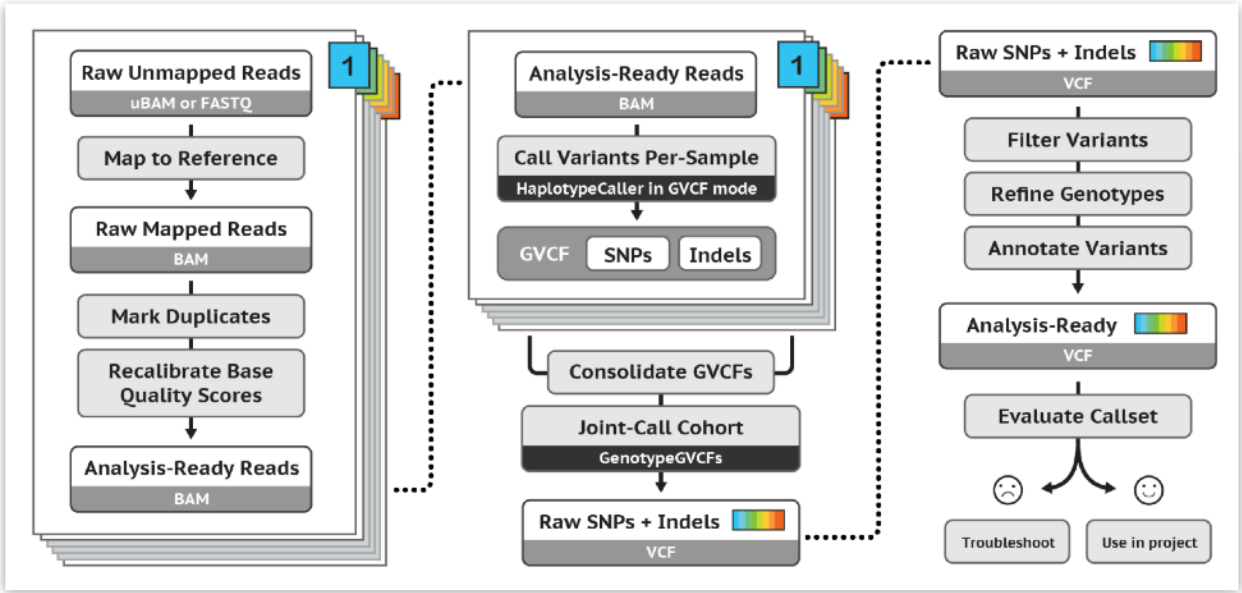
Nam J.N. et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics (2015)

GATK for variant calling analysis

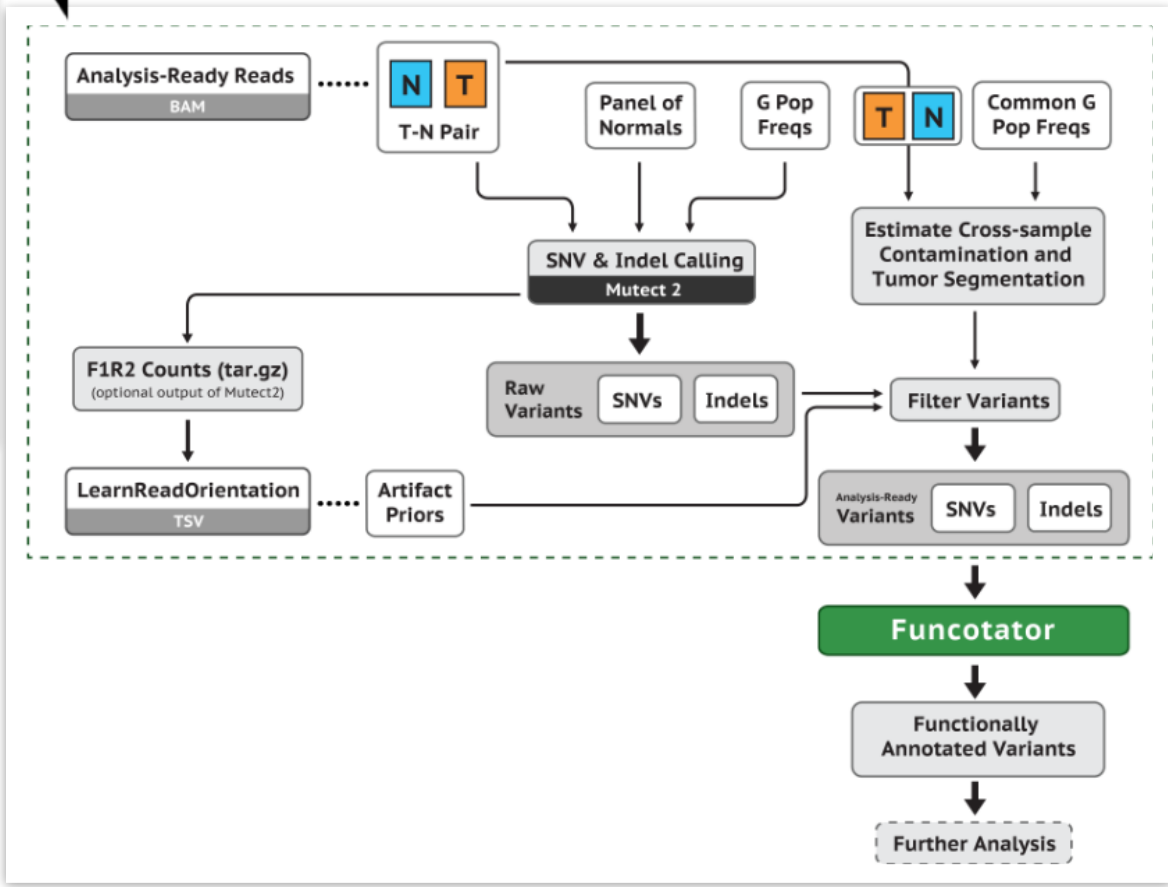


Algorithms for variant calling

GATK Best Practices



Germline small-scale variants



Somatic small-scale variants

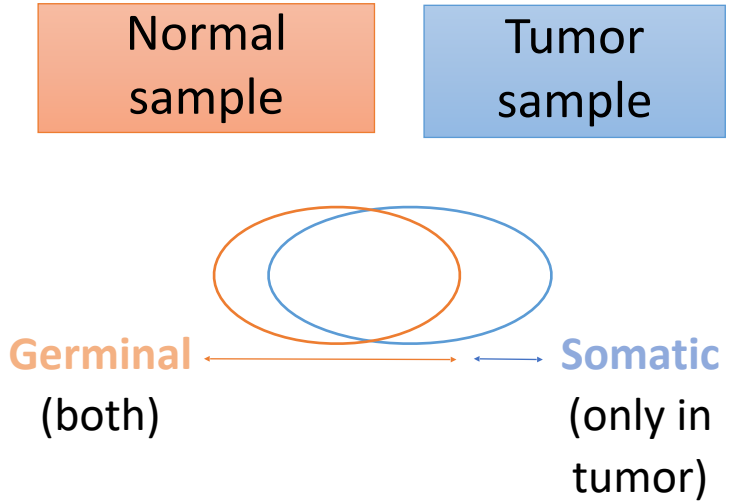
Algorithms for variant calling

Somatic vs germline variants

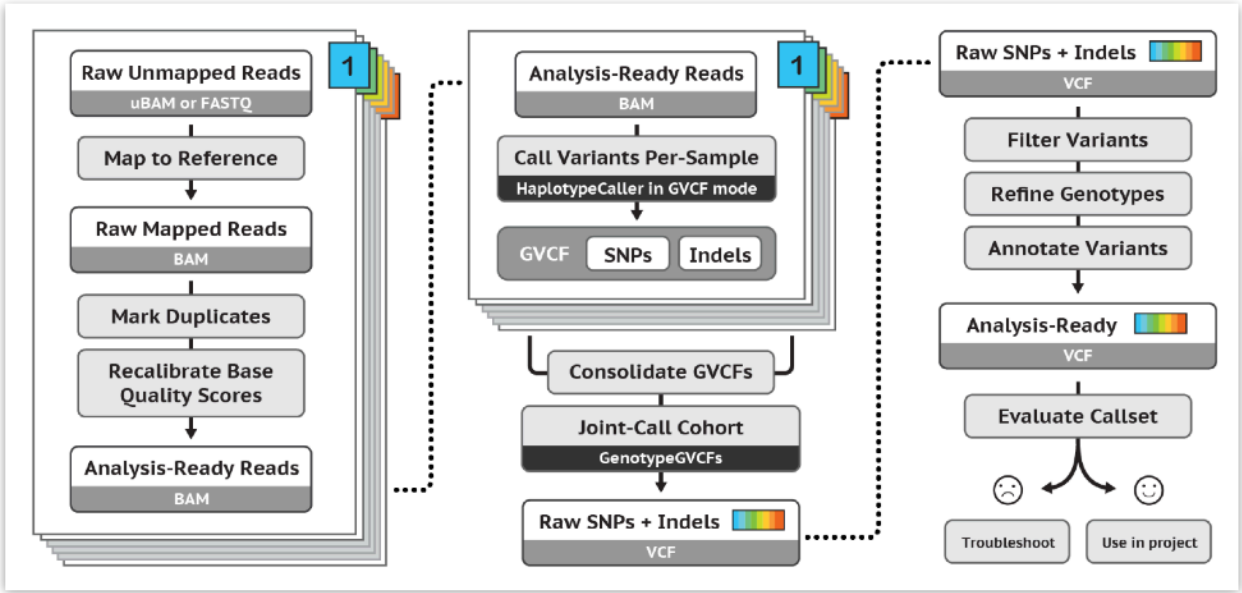
Germline: appear in gametes
Inheritable
Affect to future generations
e.g.: variants involved in rare diseases

Somatic: appear in different from germline cells
Acquired
Only affect to the own affected cell linage
e.g.: variants causing cancer

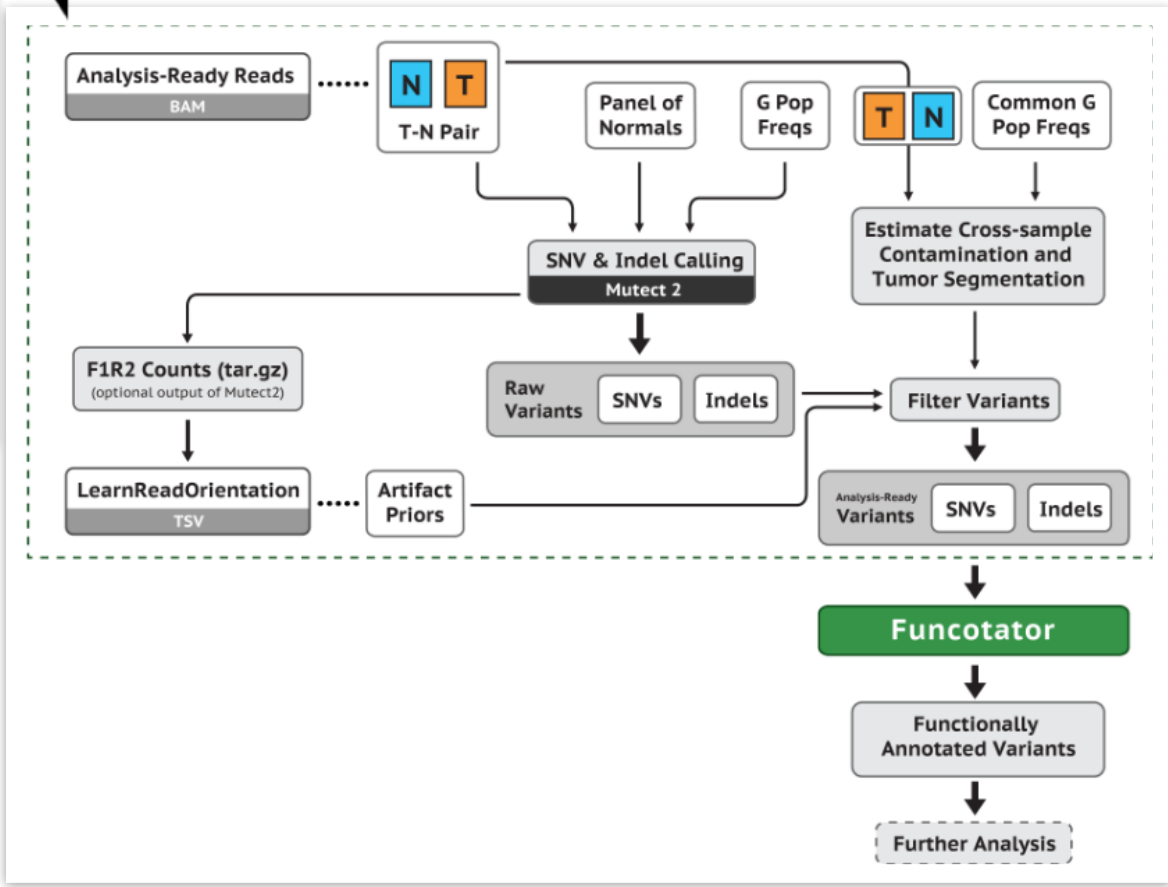
Identified by intersection:



GATK Best Practices



Germline small-scale variants

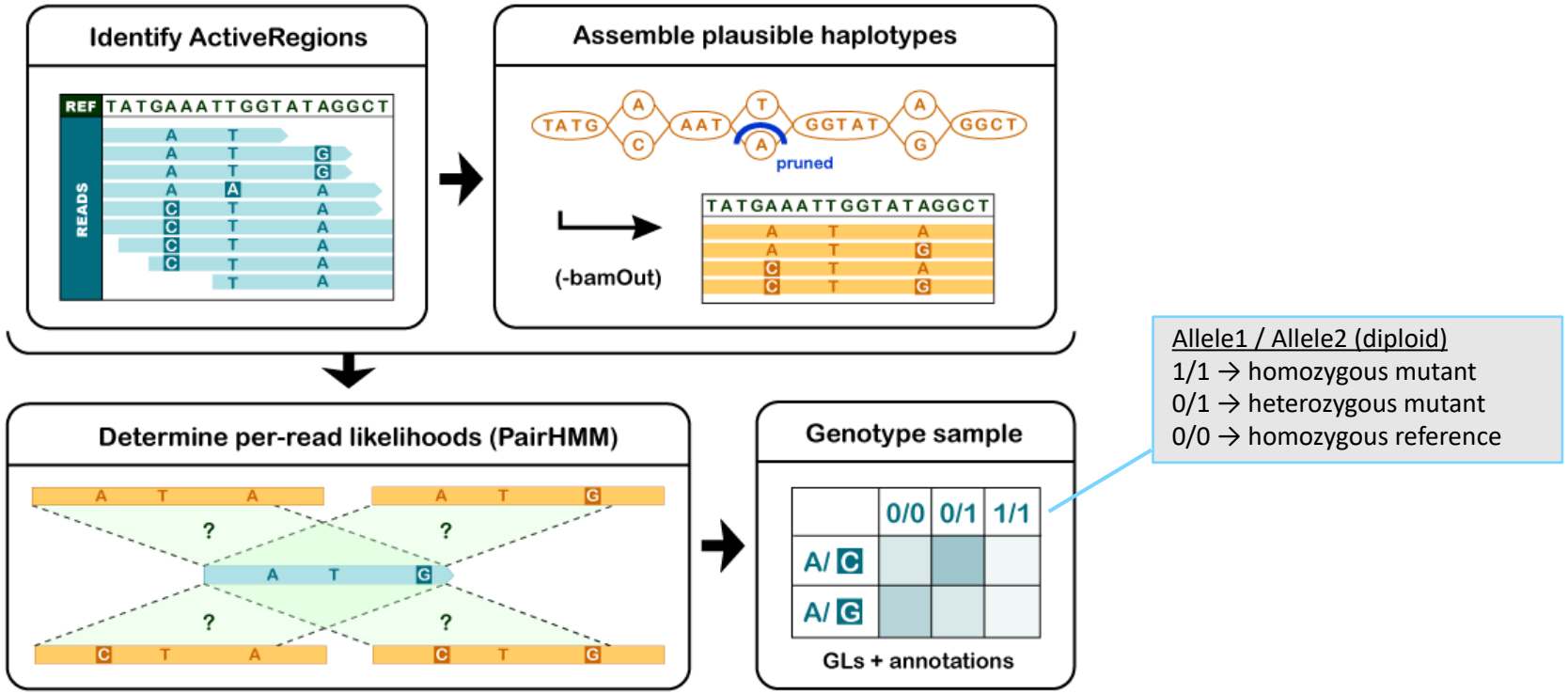


Somatic small-scale variants

Algorithms for variant calling

Variant Calling for SNVs and Indels

Haplotype Caller : Variant calling based on the calculation of genotype likelihoods:



Allele1 / Allele2 (diploid)
1/1 → homozygous mutant
0/1 → heterozygous mutant
0/0 → homozygous reference

Assumptions: It bases the calling in the indicated ploidy (e.g. 2n)

Limited detection of low allele frequencies.

Further reading:

https://github.com/broadgsa/gatk/blob/master/doc_archive/methods/HC_overview:_How_the_HaplotypeCaller_works.md

https://gatk.broadinstitute.org/hc/en-us/sections/360007226771?name=methods

Intra-tumoral heterogeneity effect in VAF

Variant Allele Frequency

Proportion of DNA molecules in the sample carrying the variant

$$VAF = \frac{sequence\ reads\ with\ a\ DNA\ variant}{overall\ coverage\ at\ that\ locus}$$

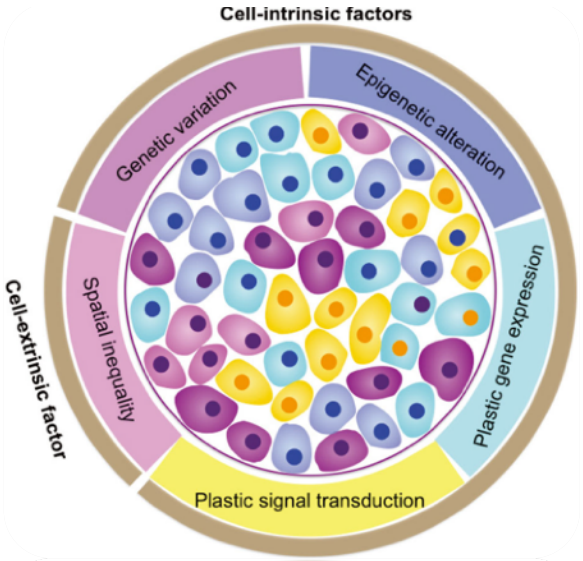
For a diploid organism:

- **heterozygous loci** should be near 0.5 VAF
- **homozygous loci** should be near 1 VAF
- **reference loci** should be near 0 VAF

doi: 10.28092/j.issn.2095-3941.2016.0004

Clonal composition in cancer

changes 0.5/1.0 diploid variant allele frequencies



doi: 10.1038/aps.2015.92

Variant Calling for somatic variants: MuTect2

SNV and Indel caller**.**

Similar logic to Haplotype Caller but:

- It **allows variable allele frequencies**.
- It **includes logic to avoid germline variants**.

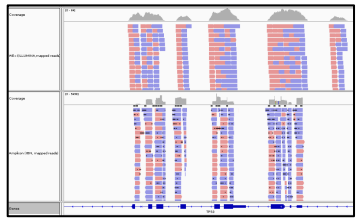
Cibulskis, K. et al.
Nat Biotechnology (2013).doi:10.1038/nbt.2514

https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2

bcftools

http://samtools.github.io/bcftools/

Set of tools to call variants and manage VCF files.

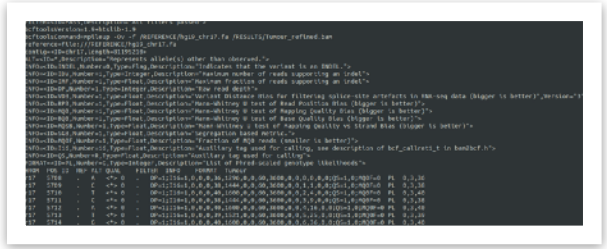


mpileup

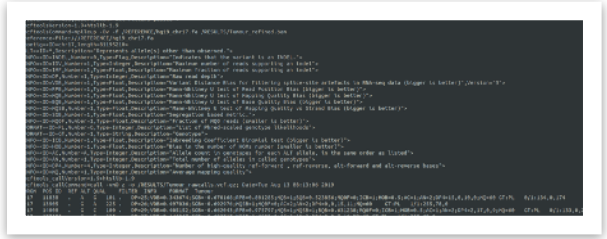
call

Variant Detection

Genotype Calculation

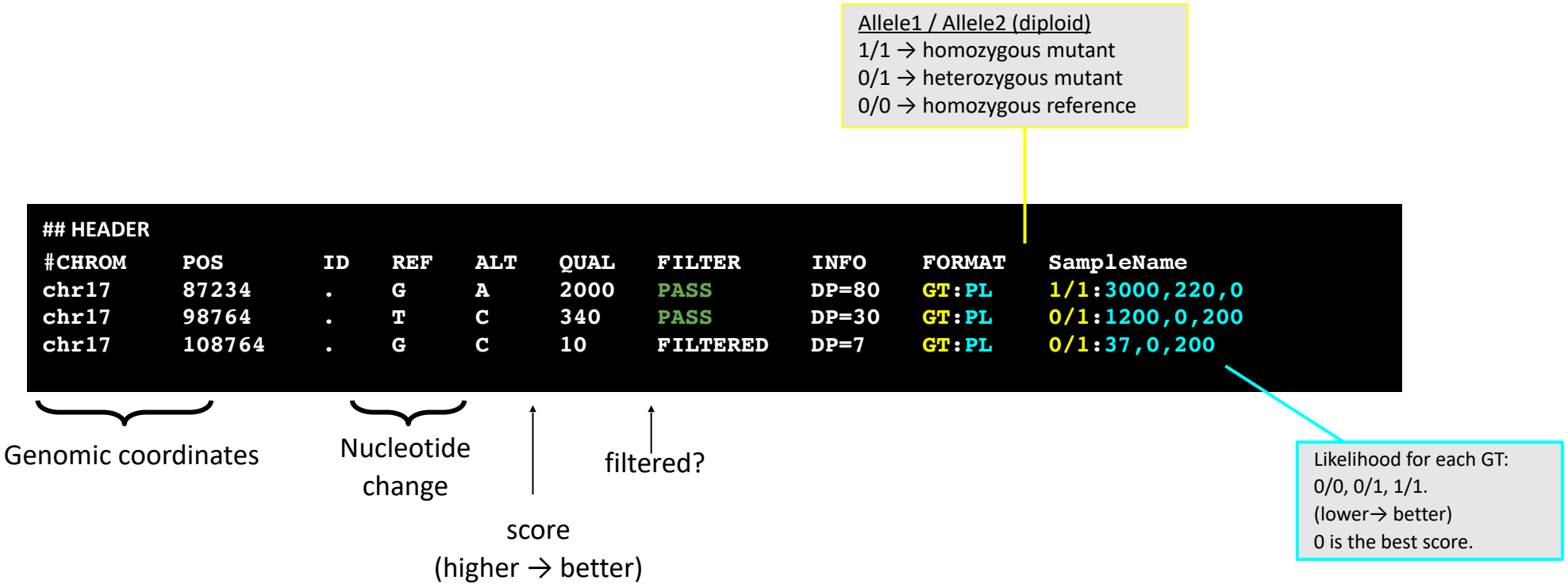


Variant Calling



Algorithms for variant calling

VCF file



More info.:

https://samtools.github.io/hts-specs/VCFv4.2.pdf

https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format

Data for the next class

Data:

Raw Data

Reference Genome

Software:

```
conda install -c bioconda fastqc
conda install -c bioconda bwa
conda install -c bioconda samtools
conda install -c bioconda bcftools
```