Swiss Institute of Bioinformatics

INTRODUCTION TO BIOINFORMATICS:

Clinical Bioinformatics

V. Barbié, Clinical Bioinformatics

Zürich, 03 December 2024

elixir SWITZERLAND

Outline

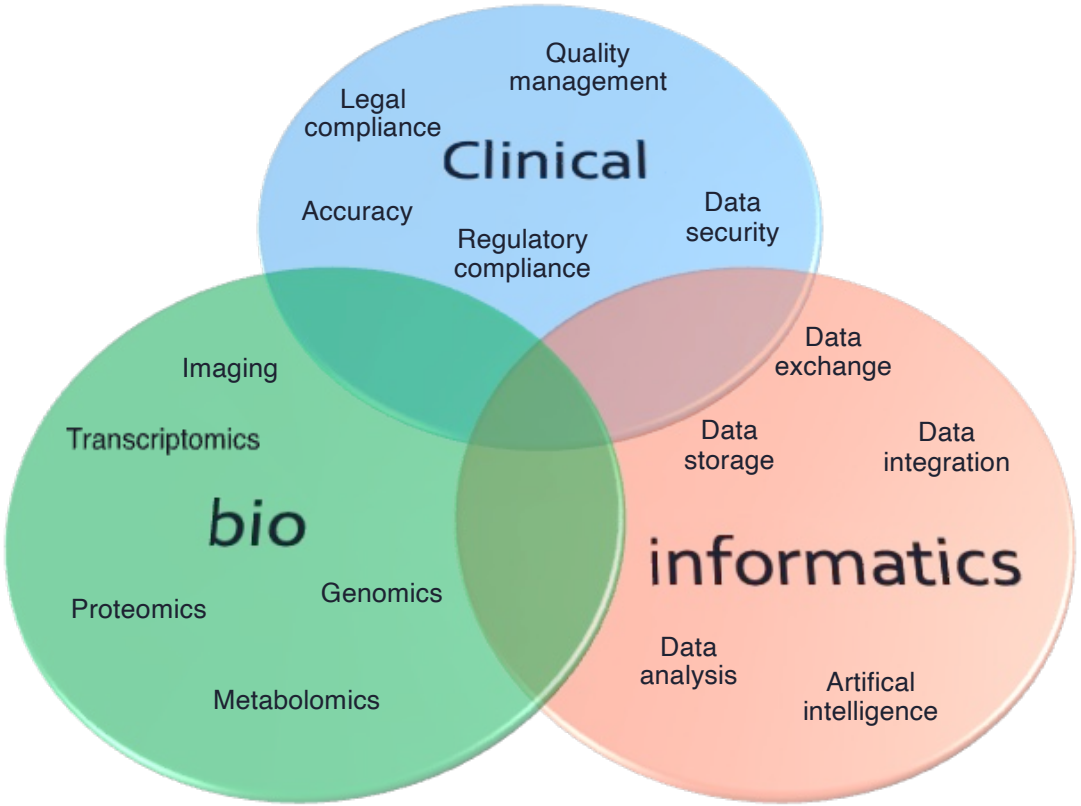
What is
clinical bioinformatics

Why clinical bioinformatics?
Next Generation Sequencing (NGS) in medical diagnosis

Overview of an oncology NGS diagnostic pipeline

Other considerations

What is clinical bioinformatics?



Outline

What is
clinical bioinformatics

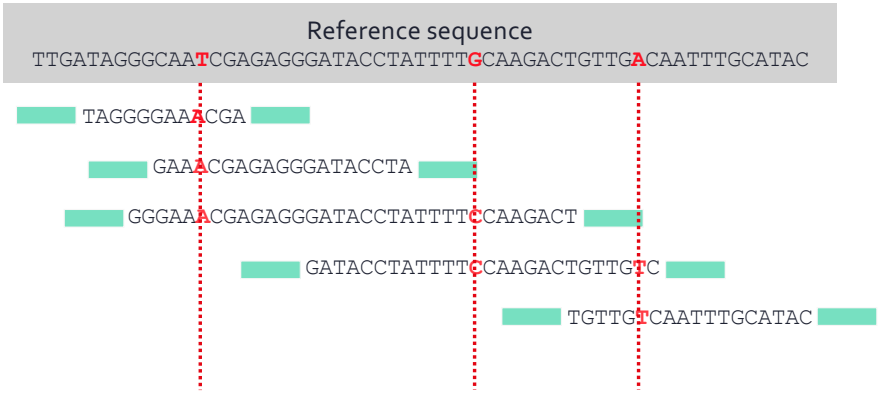
Why clinical bioinformatics?

Next Generation Sequencing (NGS) in medical diagnosis

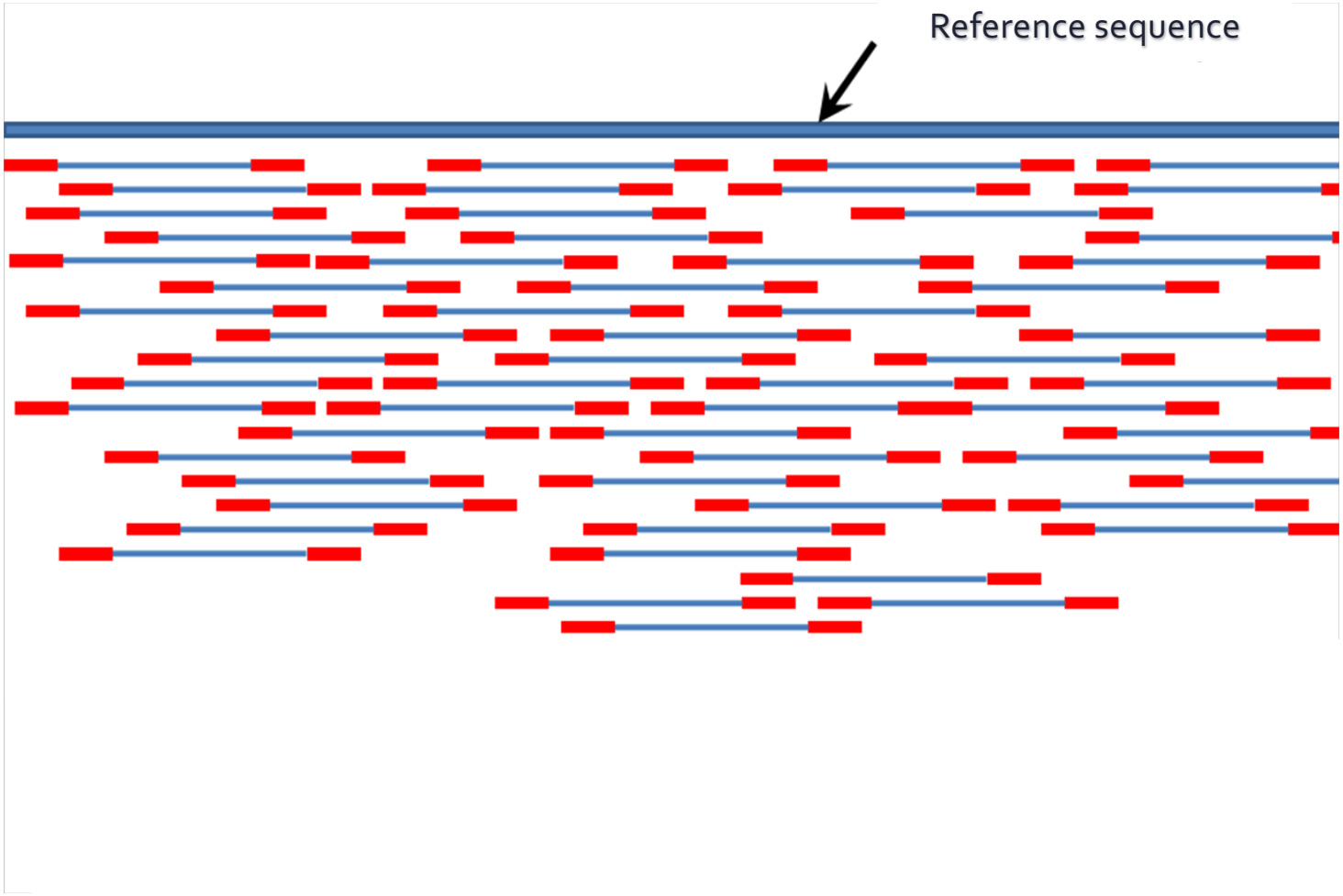
Overview of an oncology NGS diagnostic pipeline

Other considerations

Next Generation Sequencing principle



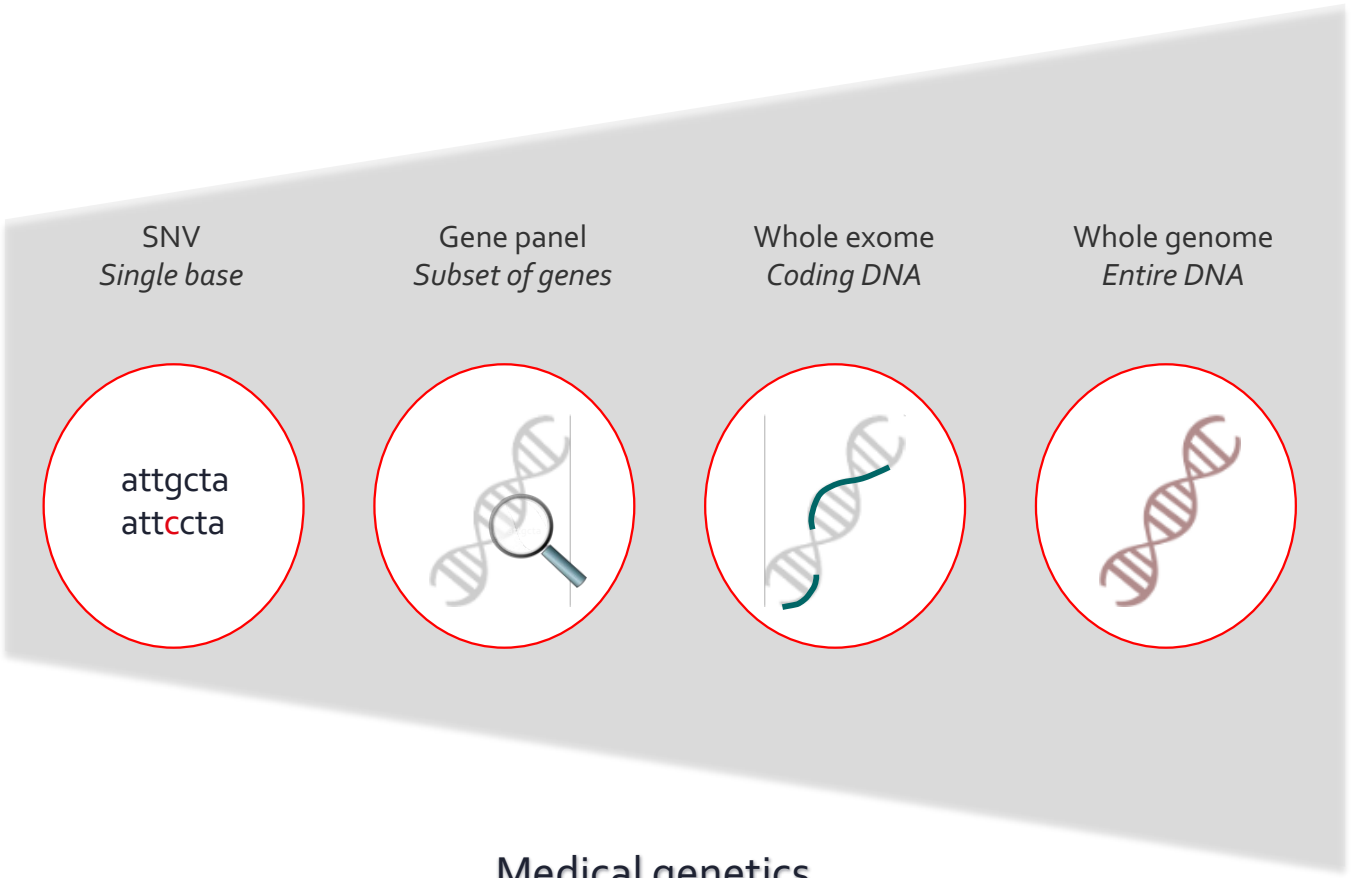
Next Generation Sequencing principle



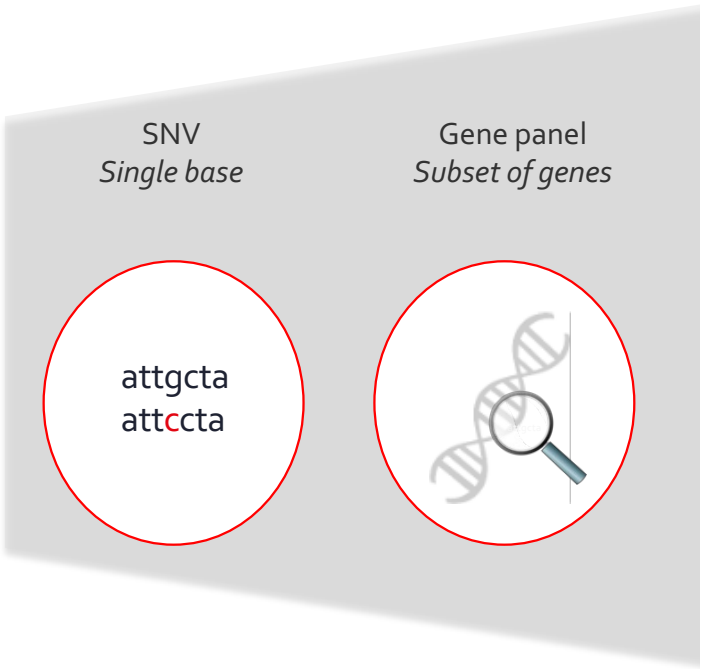
Examples of NGS clinical applications

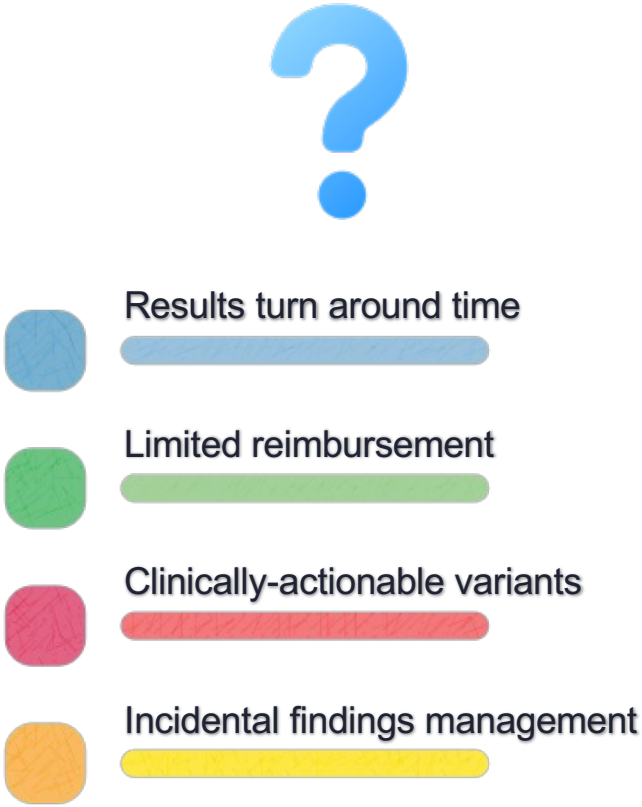
	Source DNA	Reference DNA
Oncology	Patient tumor or blood	Consensus human genome Germline
Microbiology	Patient	Pathogens genomes, resistance genes
Medical genetics	Patient	Family members, known defects
Pharmacogenetics	Patient	Drug-response or -sensitivity mutations

A matter of scale



A matter of scale





Oncology

Outline

What is
clinical bioinformatics

Why clinical bioinformatics?

Next Generation Sequencing (NGS) in medical diagnosis

Overview of an oncology NGS diagnostic pipeline

Other considerations

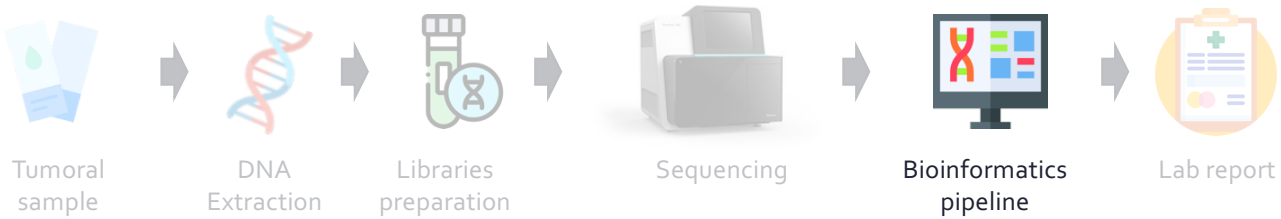
NGS in cancer diagnosis

- Identify single nucleotide variants (SNVs), insertions-deletions (indels) to inform clinical management

Overview of a NGS bioinformatics pipeline

- Gene panels analysis in clinical routine
  - Identify **differences**
  - Identify **artifacts**: quality control
  - Identify **somatic** vs. germline variants
  - Variant **annotation**: does it provide clinically-useful information?

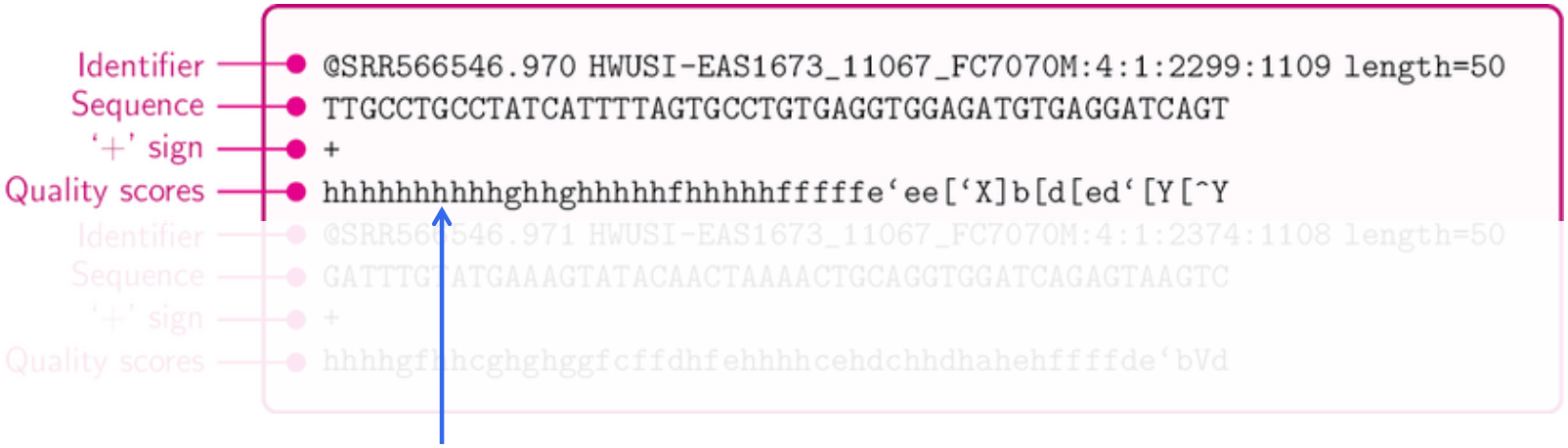
Overview of a NGS bioinformatics pipeline



Reads filtering

Quality control

Out of the sequencer: FASTQ file

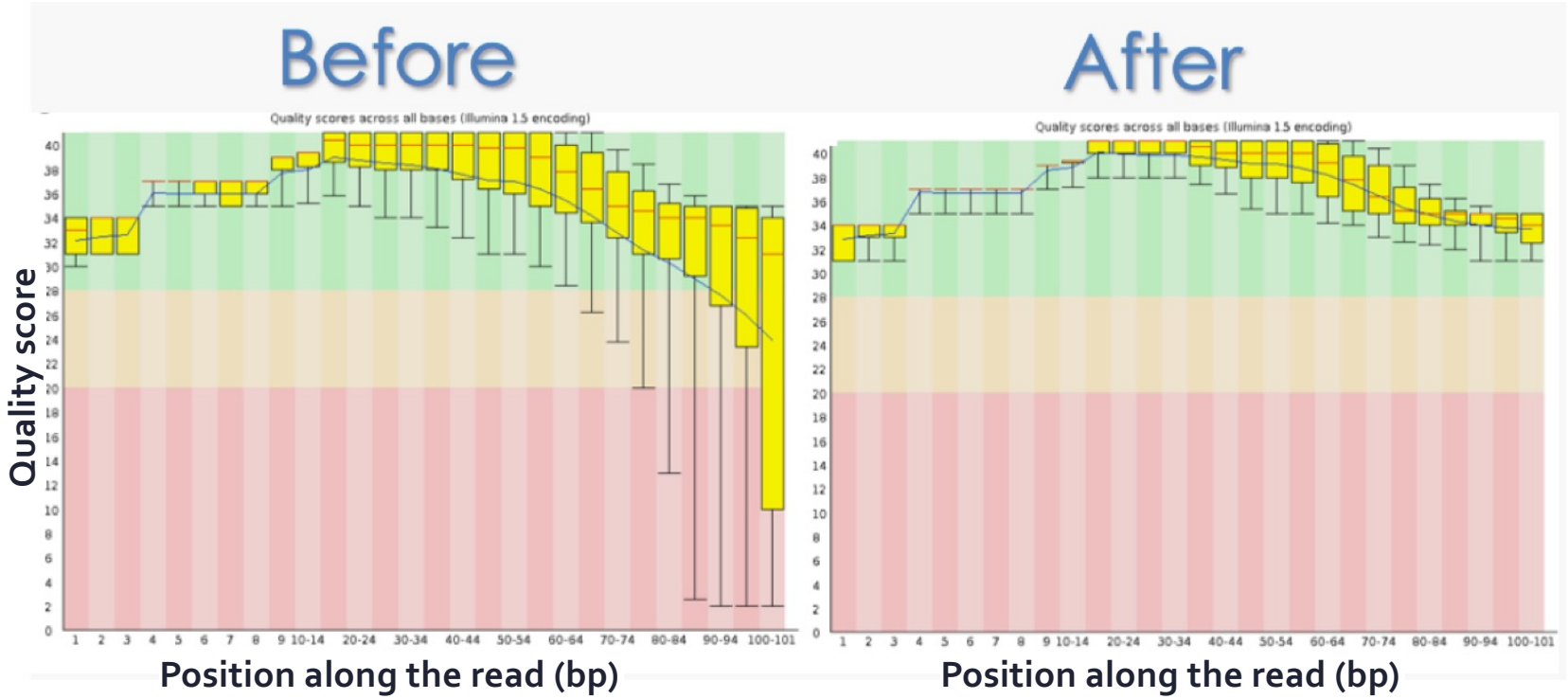


Each nucleotide has a **quality score (Phred score)**
representing the probability that a base was miscalled by the sequencer

$$Q = -10 \log_{10} P$$

Phred Score	Prob. of incorrect base call	Base call accuracy	Code
10	1 in 10	90%	J
20	1 in 100	99%	T
30	1 in 1'000	99.9%	^
40	1 in 10'000	99.99%	h

Quality-based reads trimming



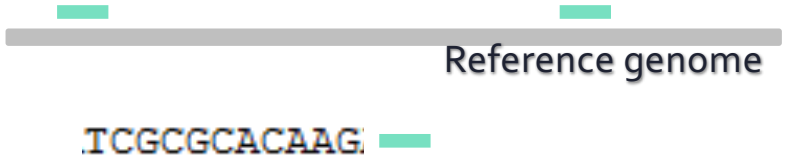
Overview of a NGS bioinformatics pipeline

Reads filtering → **Reads mapping**

Quality control

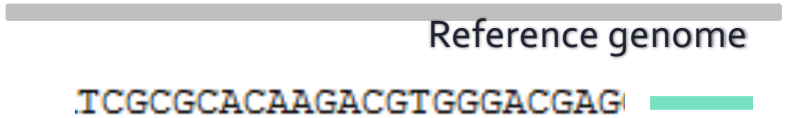
Variant identification

Let’s align the reads



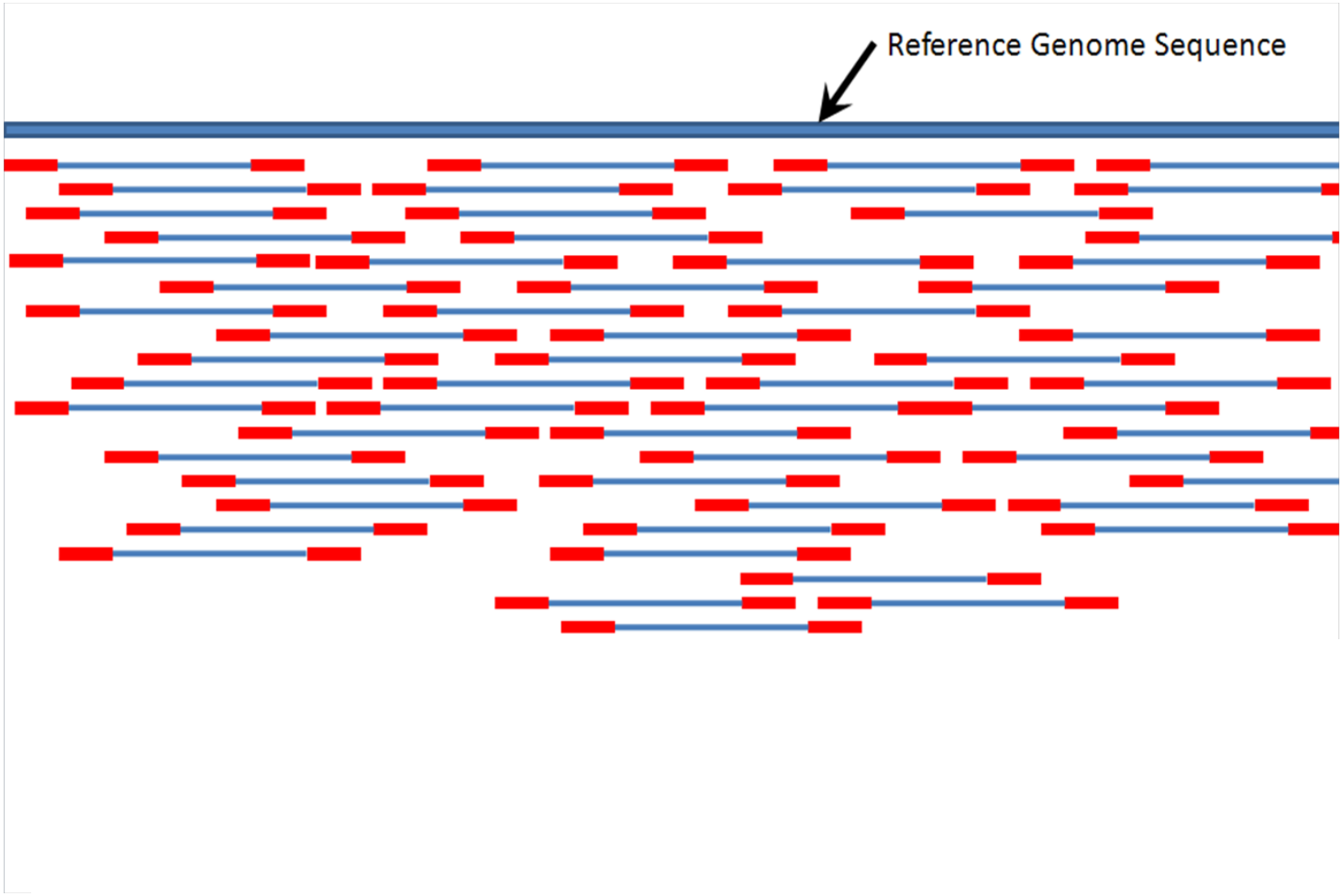
- **Short reads**
  - Likely to map at several positions along the reference genome

- **Mismatches and gaps allowed**
  - Algorithms use scoring functions



- **Longer reads**
  - Less ambiguous
  - But computationally more expensive

Mapping: finding the best position for each read



Overview of a NGS bioinformatics pipeline

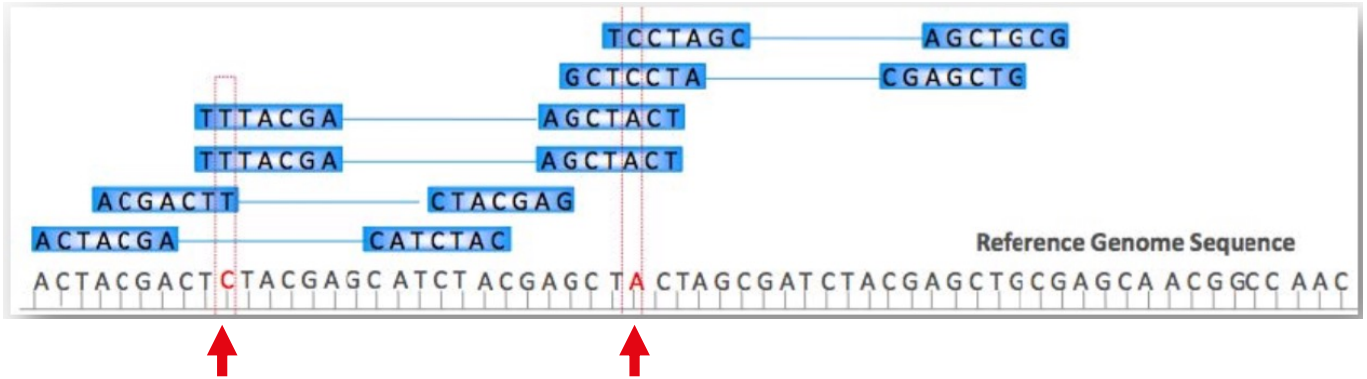
Reads filtering → Reads mapping → Variant calling

Quality control

Variant identification

Variant identification

Variant calling: putting it all together

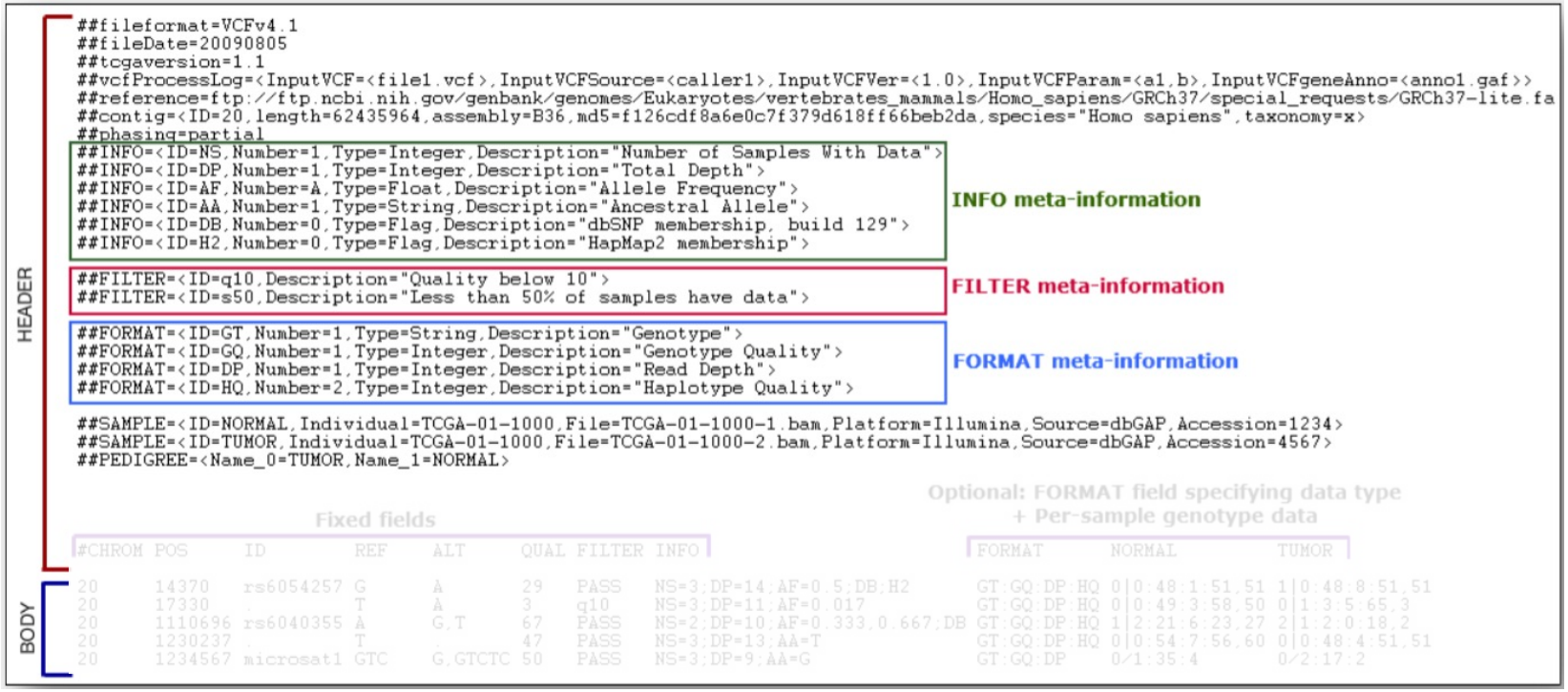


True variant or technical error?

- Performed by the sequencer software or the bioinformatician
- Germline vs somatic calling
  - Germline: constitutional genome analysis, where variants occur in 50% (heterozygous) or 100% (homozygous) of the reads.
  - Somatic: no ploidy assumption, low frequency alleles.

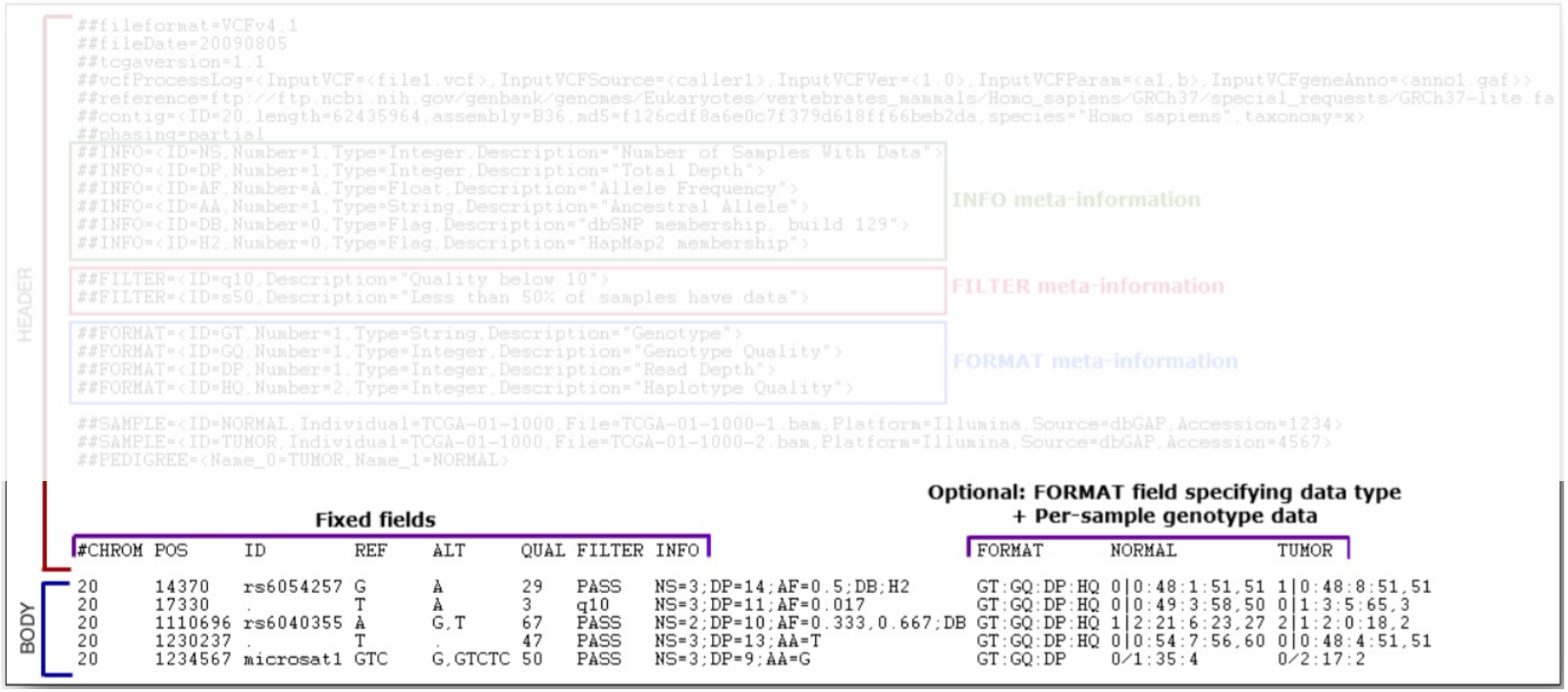
Output of the variant caller: VCF

VCF: Variant Call Format



Output of the variant caller: VCF

VCF: Variant Call Format



Output of the variant caller: VCF

VCF: Variant Call Format

Fixed fields

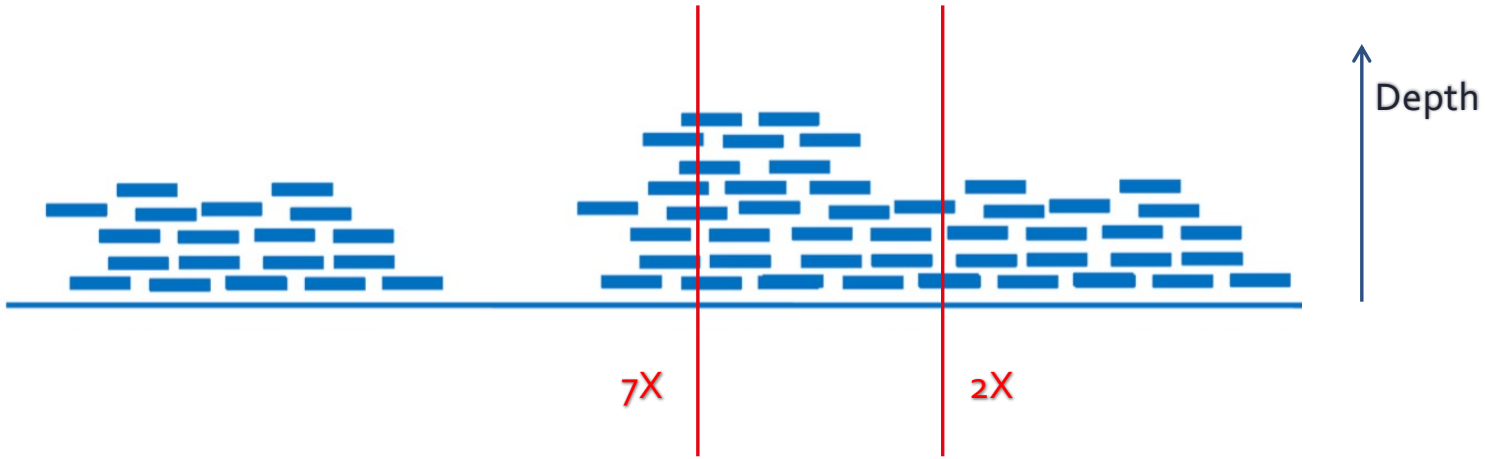
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G

BODY

Things to watch out when assessing variant quality

Depth

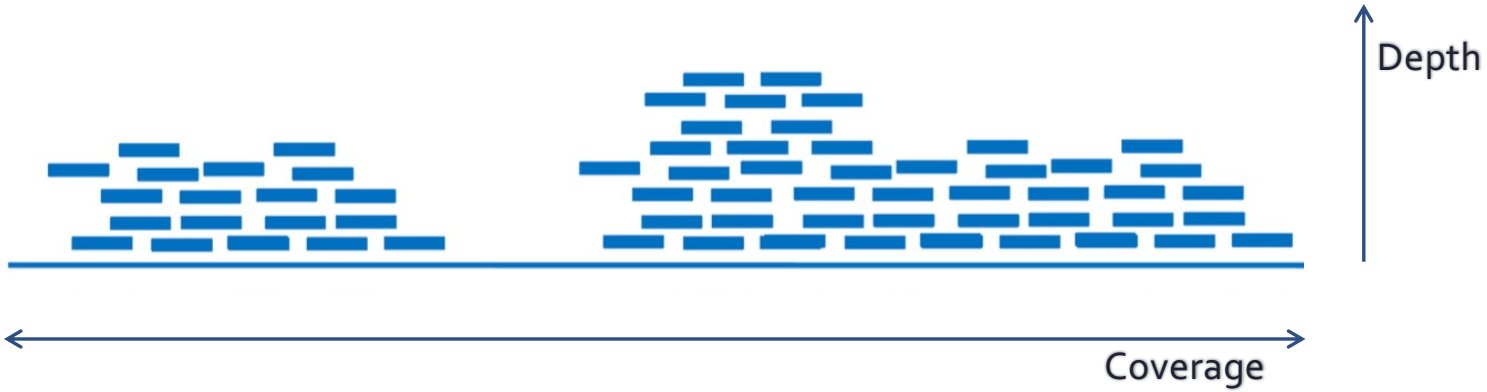
Depth: nb of reads that include a given nucleotide, at a given position



- Diagnosis: gene panel at 1500X, whole exome at 100X
- In oncology, impossible to detect low frequency clones with exome analyses

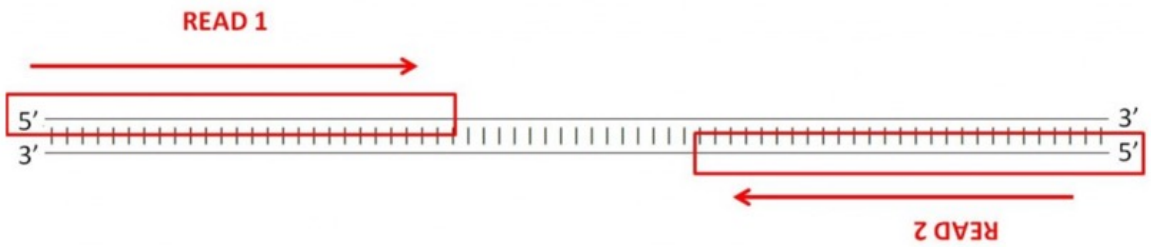
Coverage

Coverage: % or nb of bases of a reference genome that are covered with a certain depth, e.g. 90% at 5X



Strand bias in paired-end sequencing

- Both DNA strands are sequenced
- Normal mutations should occur on both with equal frequencies



http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/

Overview of a NGS bioinformatics pipeline

Reads filtering	Reads mapping	Variant calling	Variant annotation
Quality control	Variant identification	Variant identification	Variant annotation and interpretation

Medical genetics: focus on pathogenicity

© American College of Medical Genetics and Genomics ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

GENETICS in MEDICINE | Volume 17 | Number 5 | May 2015

Find pathogenic variants

i.e. genetic alterations increasing an individual's susceptibility or predisposition to a certain disorder

Oncology: focus on clinical significance

The Journal of Molecular Diagnostics, Vol. 19, No. 1, January 2017

ELSEVIER

the Journal of Molecular Diagnostics
jmd.amjpathol.org

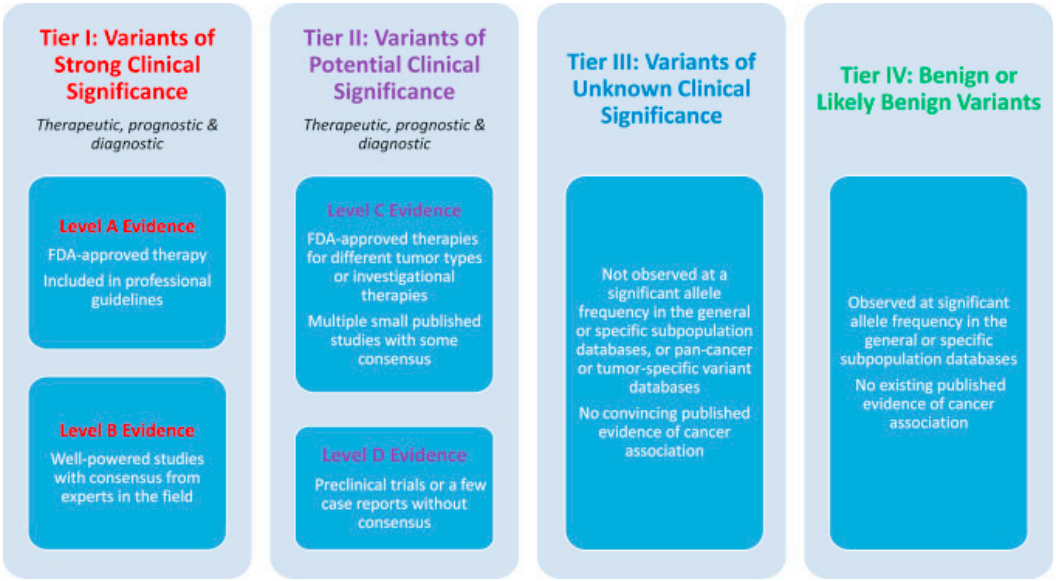
SPECIAL ARTICLE

Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer

A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists

Find actionable variants

i.e. genetic alterations possibly having an impact on clinical care



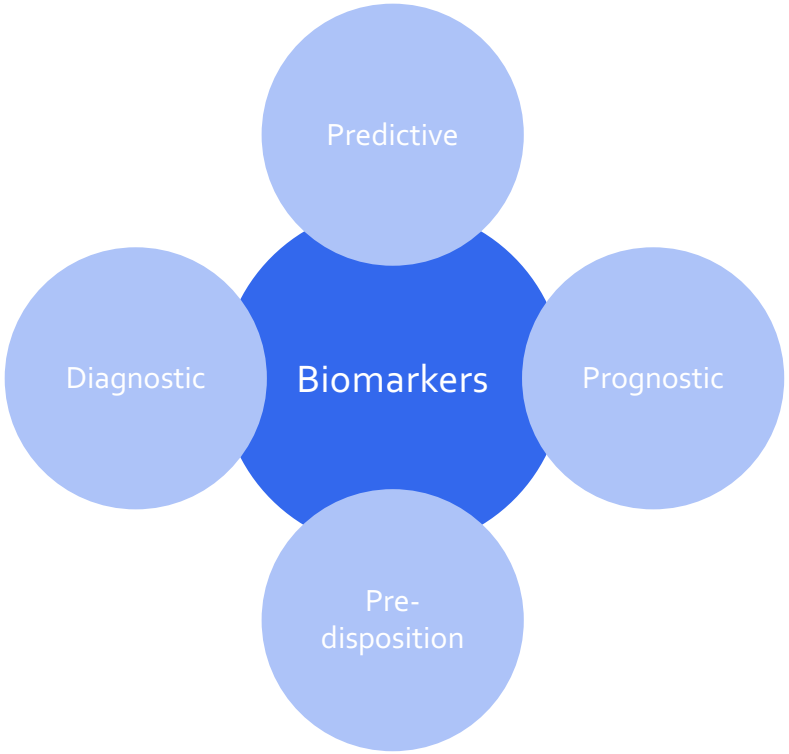
Categories of markers

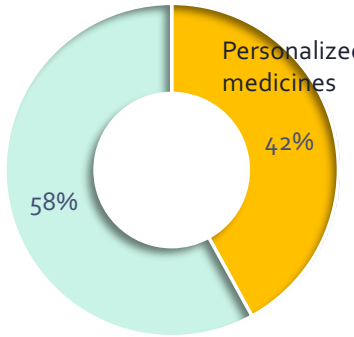
Supports treatment decisions

Supports disease characterization

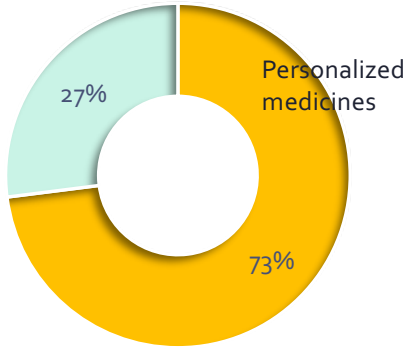
Informs on disease evolution

Indicates risk to develop a disease





All drugs in development



Oncology drugs in development

Bioinformatics to the rescue…

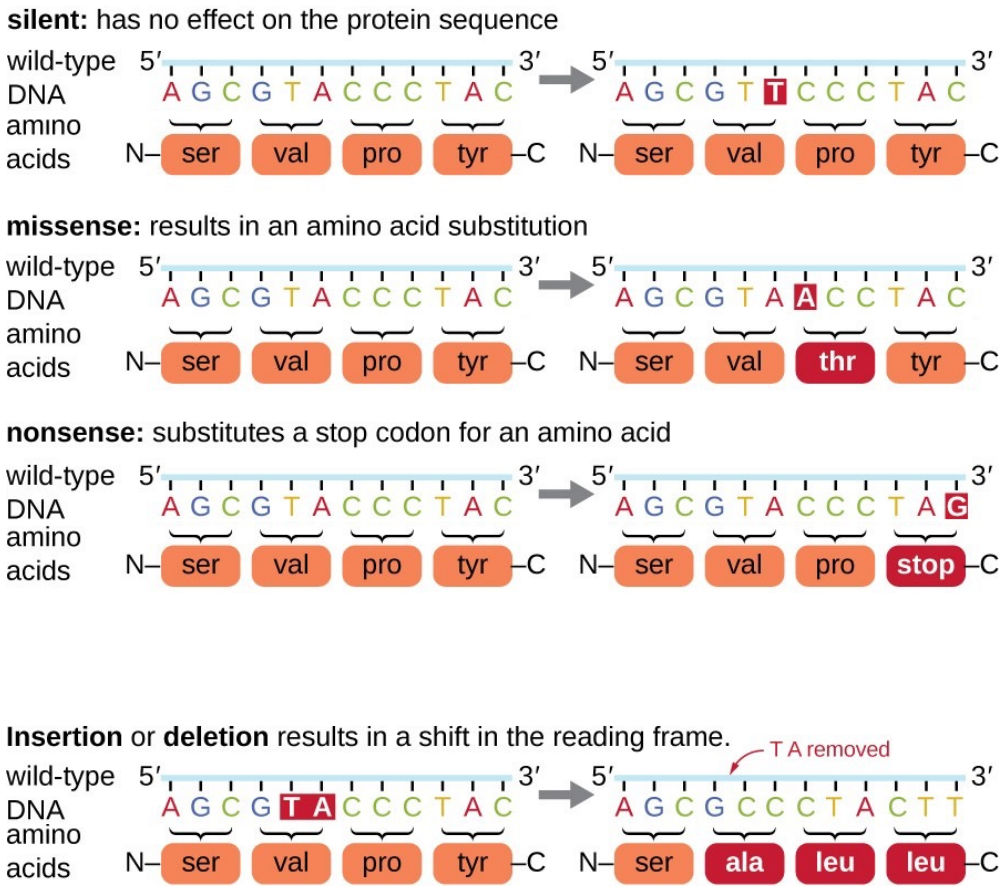
- Location of the variant (e.g. intron, exon, regulatory region…)
- Genes and transcripts affected by the variant
- Predict variant effect (e.g. stop gained, missense…)

http://www.ensembl.org/info/docs/tools/vep/index.html

Locating variants

- Convert genomic coordinates (chromosome, position) to the corresponding cDNA/amino-acid coordinates
- HGVS nomenclature (http://varnomen.hgvs.org)
  - Substitution c.76A>T
  - Deletion c.76delA
  - Insertion c.76_77insG
  - Genomic sequence g.476A>T
  - Protein sequence p.Lys76Asn
- Important to store for tracking
  - Version of the human genome assembly
  - Accession and version of the mRNA transcripts

Predicting variants effect on the protein



Point mutations (single base substitution)

Frameshift mutations (insertion or deletion of one or several bases)

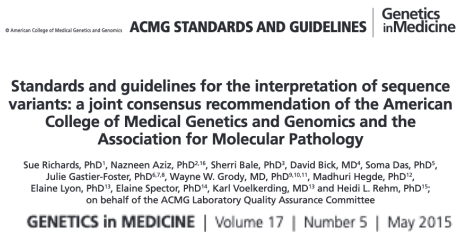
https://courses.lumenlearning.com/microbiology/chapter/mutations/

Bioinformatics to the rescue…

- Location of the variant (e.g. intron, exon, regulatory region…)
- Genes and transcripts affected by the variant
- Predict variant effect (e.g. stop gained, missense…)
- Predict variant impact on protein function, splicing

Predicting variants impact: examples of tools

TOOLS	SnpEff (ClinEff)	VEP	SIFT	PolyPhen-2	FATHMM
Variant effect and location (sequence ontology)	✓	✓			
Prediction of impact (score or category)	✓	←	✓	✓	✓
Features used for impact prediction	Rules based on variant effect (stop gained, lost…)		AA conservation in related seq.	AA conservation and structural features	AA conservation and protein tolerance to mutations



Use a combination of tools and keep variants with consensus prediction.

(not exhaustive)

Bioinformatics to the rescue…

- Location of the variant (e.g. intron, exon, regulatory region…)
- Genes and transcripts affected by the variant
- Predict variant effect (e.g. stop gained, missense…)
- Predict variant impact on protein function, splicing
- Retrieve annotations from public databases

Knowledge bases

Genetic disorders

OMIM®

DECIPHER GRCh38 Celebrating 20 years

orphanet

Cancer

OncoKB

COSMIC Catalogue Of Somatic Mutations In Cancer

CIViC Clinical Interpretation of Variants in Cancer

ClinVar

IARC TP53 Database

Constitutional variants

dbSNP Short Genetic Variations

HGMD®

ClinVar

Repositories

UniProt

RefSeq

Ensembl

PubMed

Population

gnomAD Genome Aggregation Database

Non exhaustive

SIB

Important questions

- Is it prevalent in the cancer subtype of interest?
- Is it known in other cancer subtypes or diseases?
- Is it present in the general population?
- Is it related to an ongoing clinical trial?
- What is the evidence level? Observed vs. predicted
- Are there other known variants in the same gene?

Important questions

- Is the mutation in an evolutionarily conserved region across species?

Front Pharmacol. 2015 Mar 10;6:1. doi: 10.3389/fphar.2015.00001

I found a damaging mutation: is it always bad?

- Keep the mutation in context: what is the gene function?
  - **Tumor suppressor gene**
    Damaging mutations are pathogenic.

  - **Oncogene**
    Activating mutations are pathogenic.
    (beware: damaging mutation can be activating!)

Keep the gene function in mind
when interpreting its deleteriousness

Outline

What is
clinical bioinformatics

Why clinical bioinformatics?
Next Generation Sequencing (NGS) in medical diagnosis

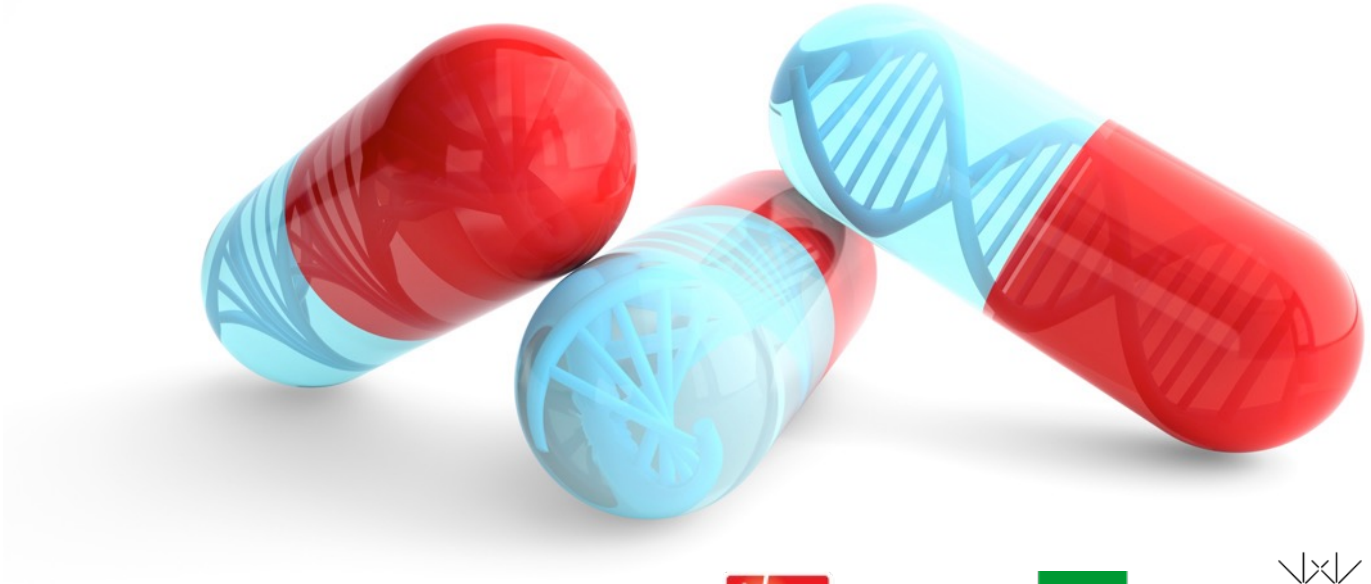
Overview of an oncology NGS diagnostic pipeline

Other considerations

Real-life constraints in the clinics

Certificate of Advanced Studies (CAS) in Personalized molecular oncology

pmo.unibas.ch



Universitätsspital Basel

SIB Swiss Institute of Bioinformatics

CHUV

UNI BASEL

CAS PMO: 4 modules and a mini-thesis

Tumor biology & Genetics
4 days

Lausanne

- Cytogenetics and molecular genetics
- Genetic modifications
- Tumor biology: solid and hematological
- Tumor genetics

Molecular Pathology
4 days

Basel

- Omics technologies
- From sample to data: extraction, sequencing, panels
- Quality control and accreditation
- Molecular profile interpretation, reports

Clinical Bioinformatics
4 days

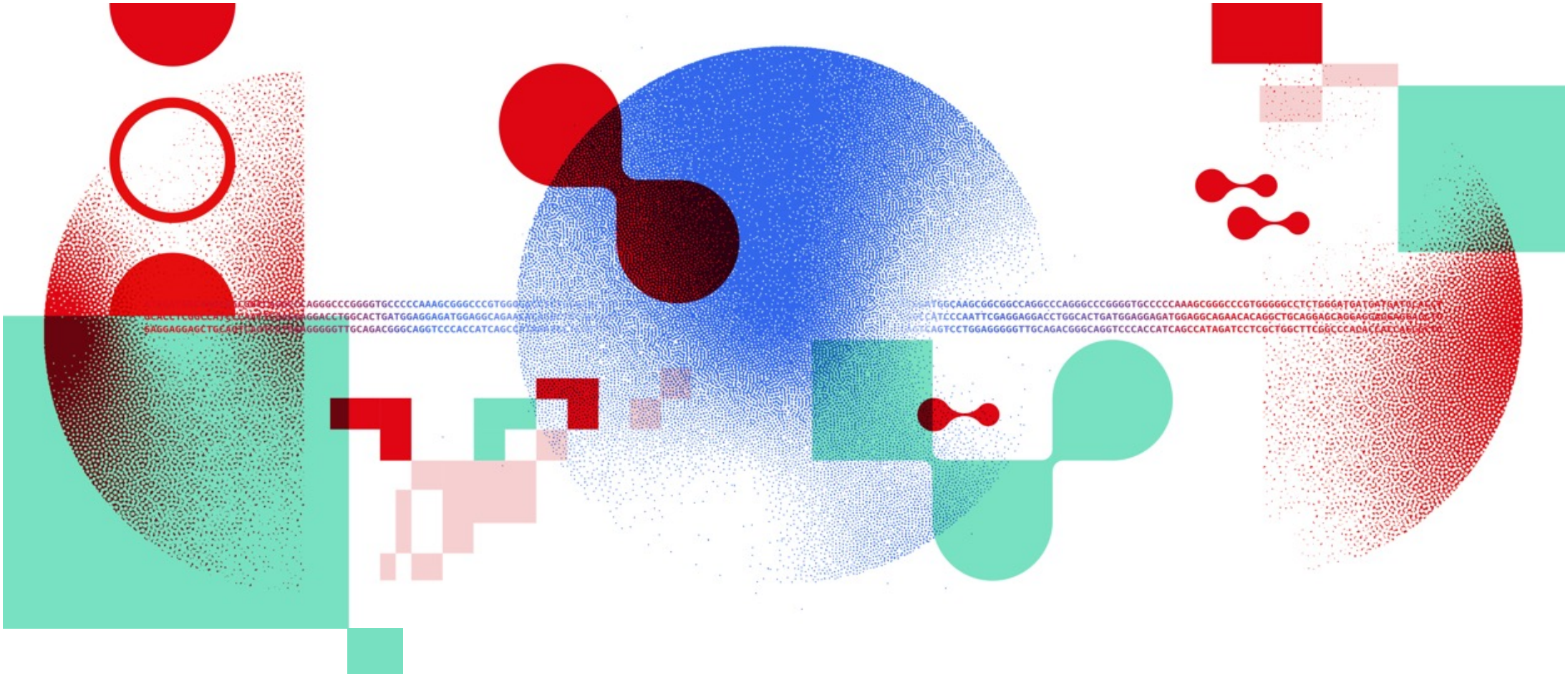
Lausanne

- NGS data processing: mapping, calling, annotation
- Data quality control
- Hardware, security, privacy
- Artificial intelligence applications

Clinical Oncology
4 days

Basel

- Tumor physiology & immunology
- Prognostic and predictive markers
- Interpretation of genetic results
- Clinical trials and tumor board



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss