SIB
Swiss Institute of Bioinformatics

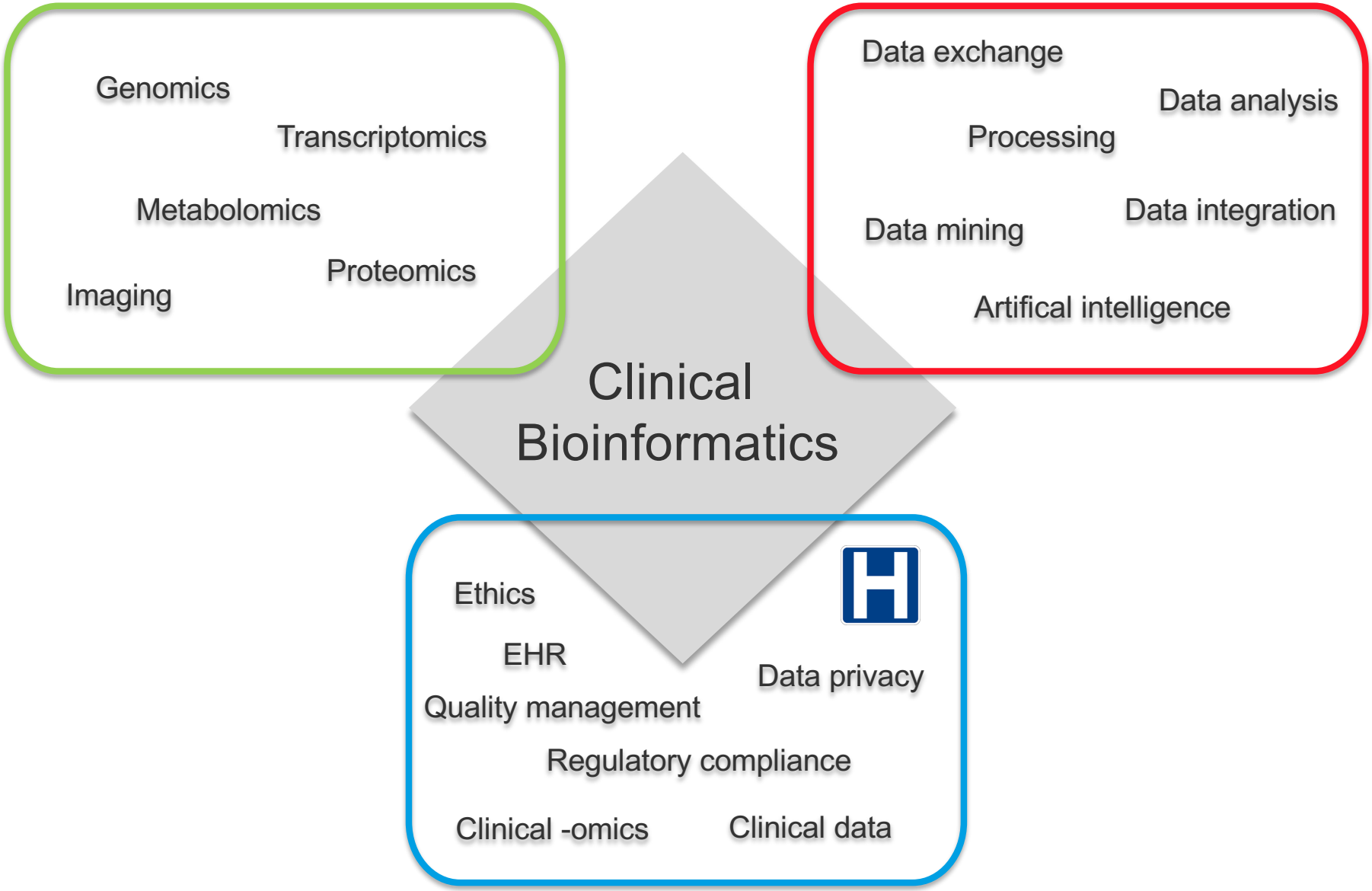
Introduction to bioinformatics: Clinical Bioinformatics

Valérie Barbié, Director SIB Clinical Bioinformatics

Zürich, 24 November 2020

elixir SWITZERLAND

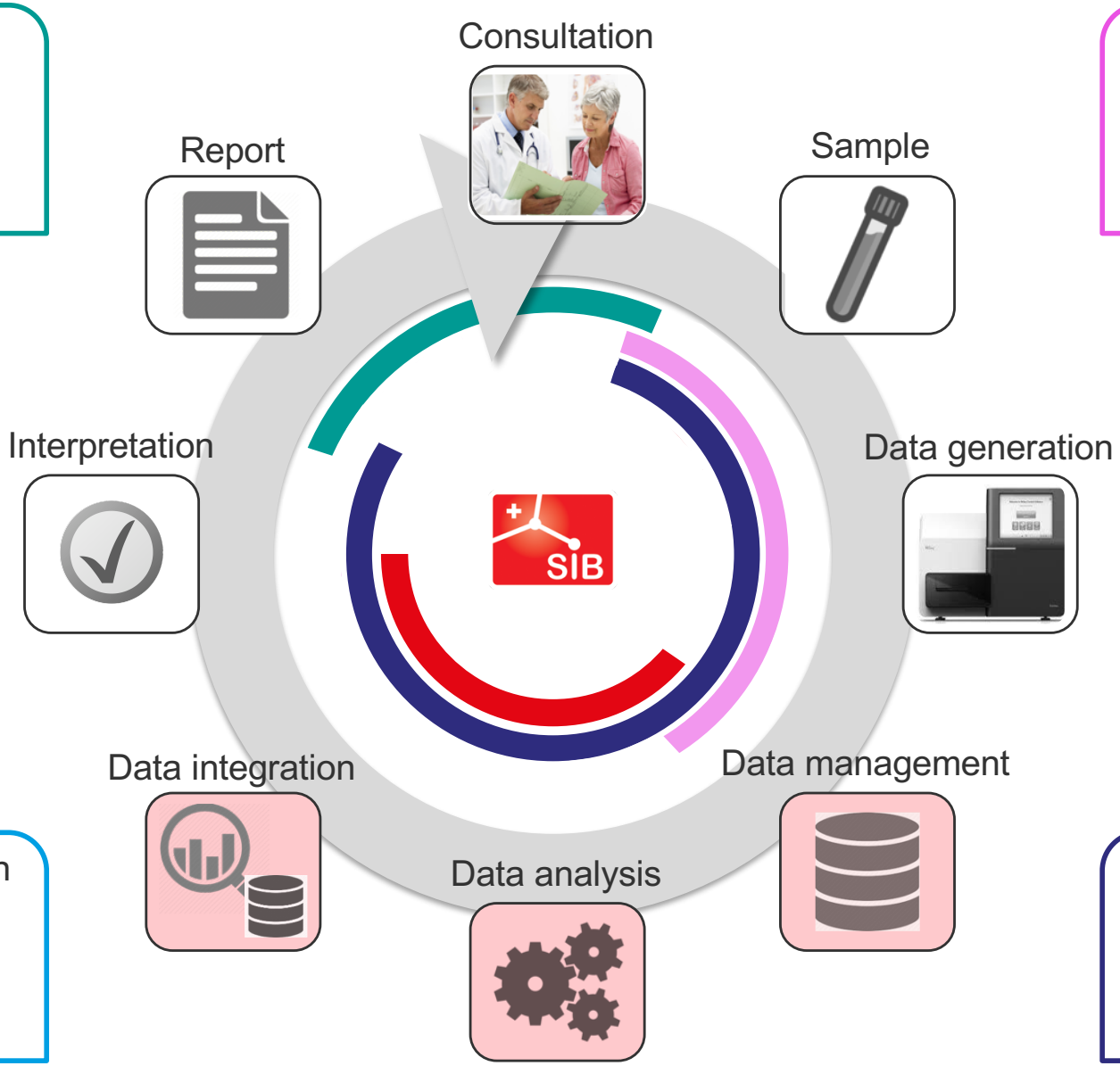
What is clinical bioinformatics?



Multiple expertise

Oncologist
Geneticist
Infectiologist
…

Lab staff
…

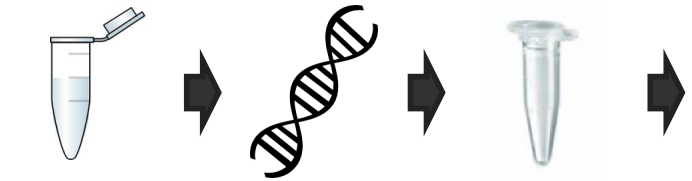


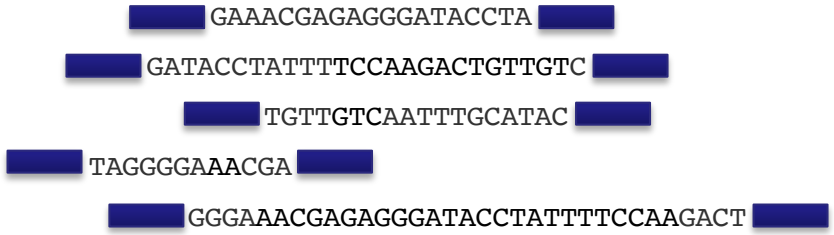
Boinformatician
Data scientist
Developer
IT specialist
…

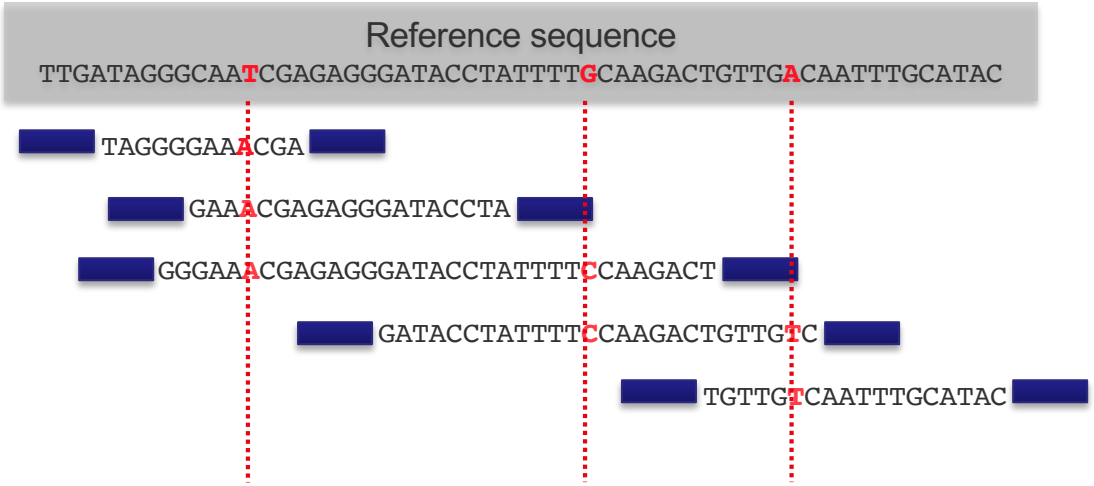
Pathologist
Microbiologist
Pharmacologist
…

Applications of Next Generation Sequencing (NGS) in medical diagnosis

Next Generation Sequencing principle

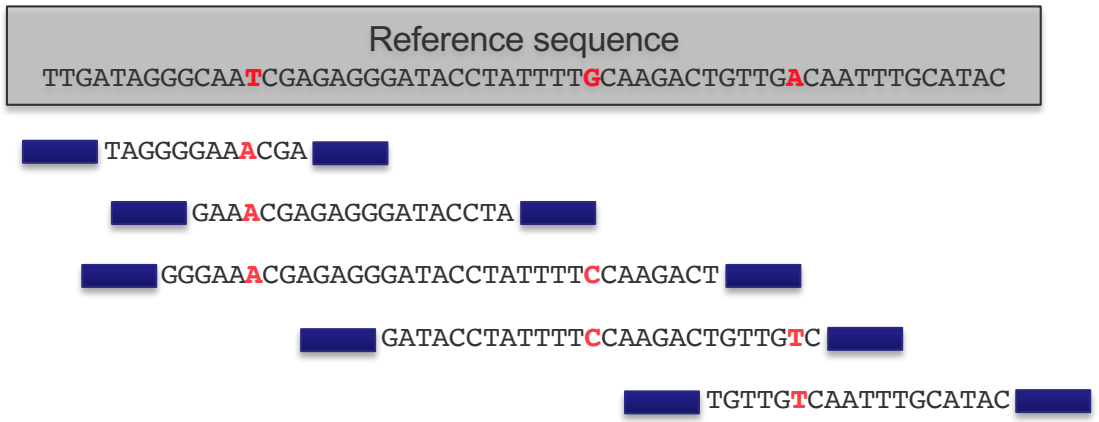




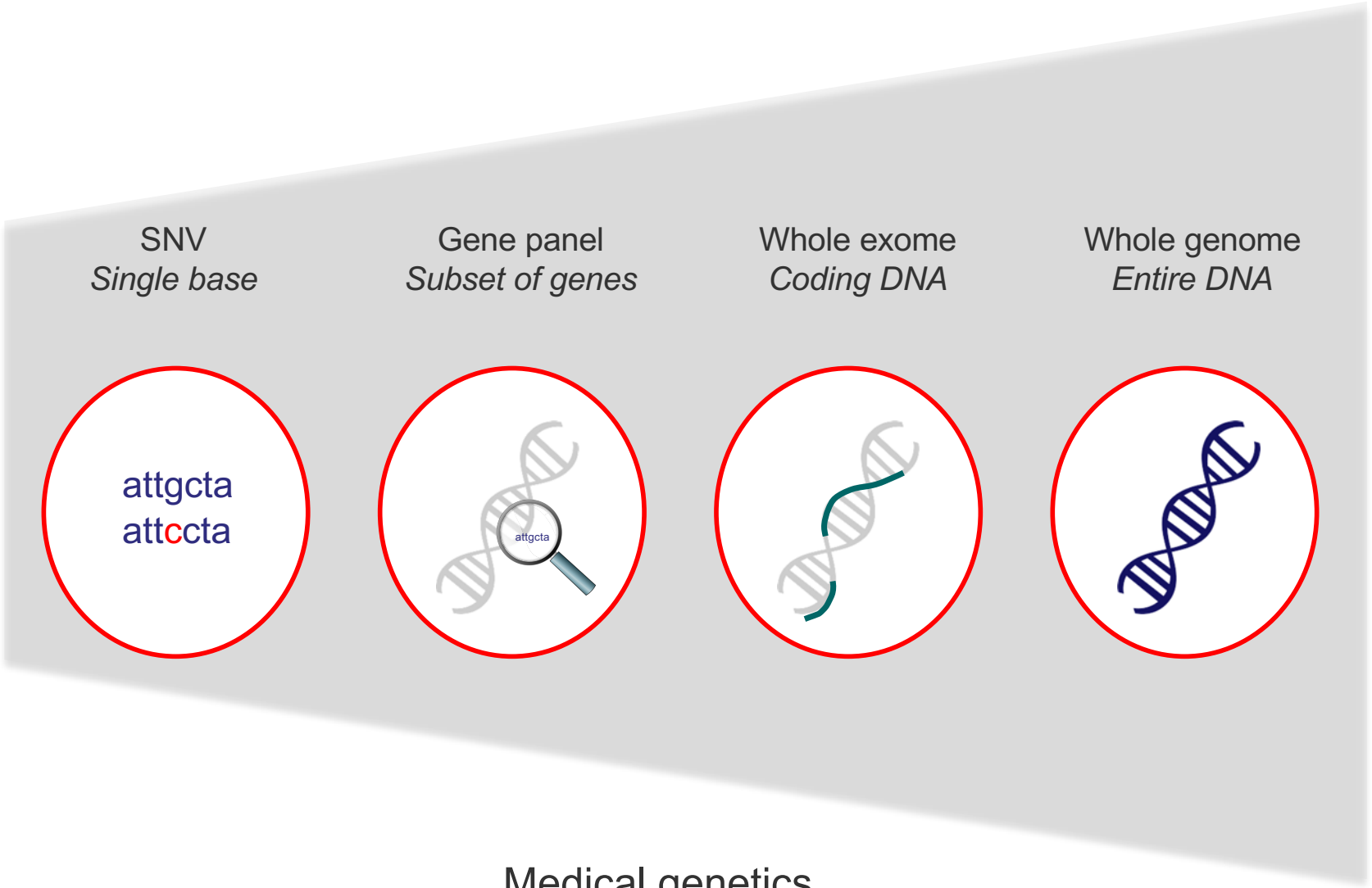


Examples of NGS clinical applications

	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

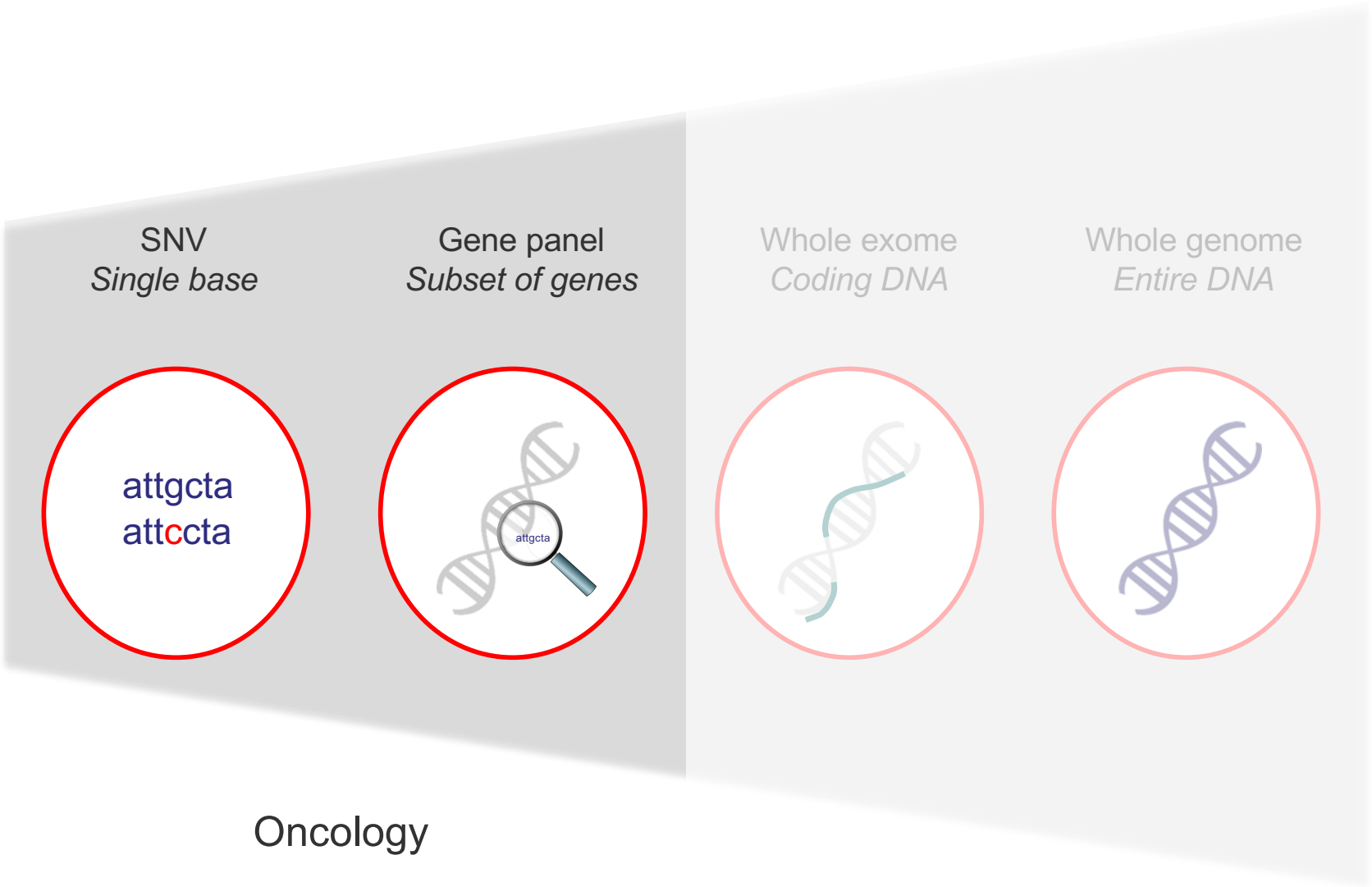


Scale matters...



Medical genetics

Scale matters...



Oncology

Applications of Next Generation Sequencing (NGS) in medical diagnosis

Focus on oncology

PART I

Overview of an NGS bioinformatics pipeline

NGS in cancer diagnosis?

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

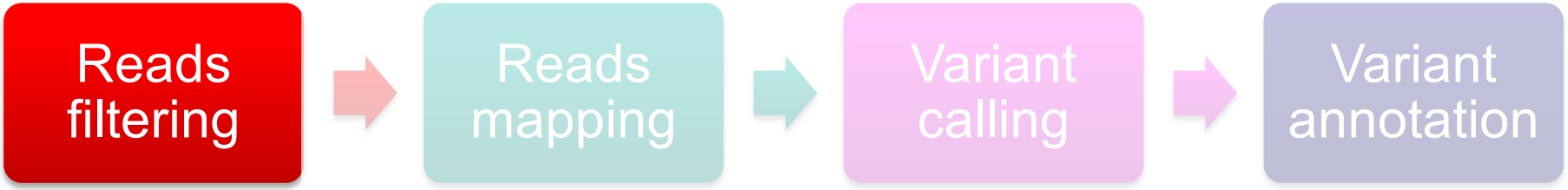
Overview of a NGS bioinformatics pipeline

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

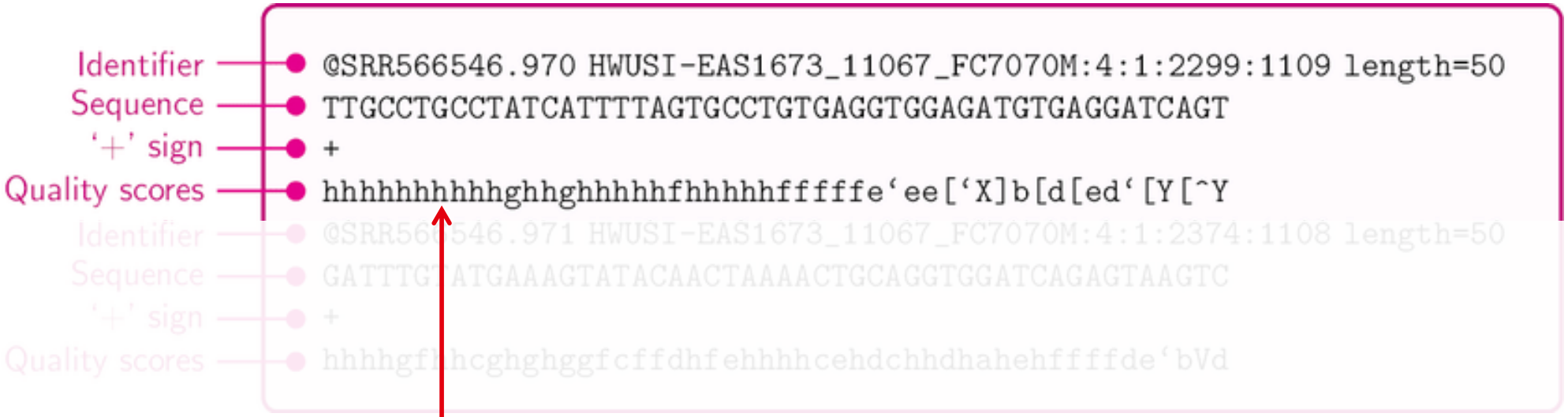
To be discussed during this course

PART II

Quality control



Out of the sequencer: FASTQ file

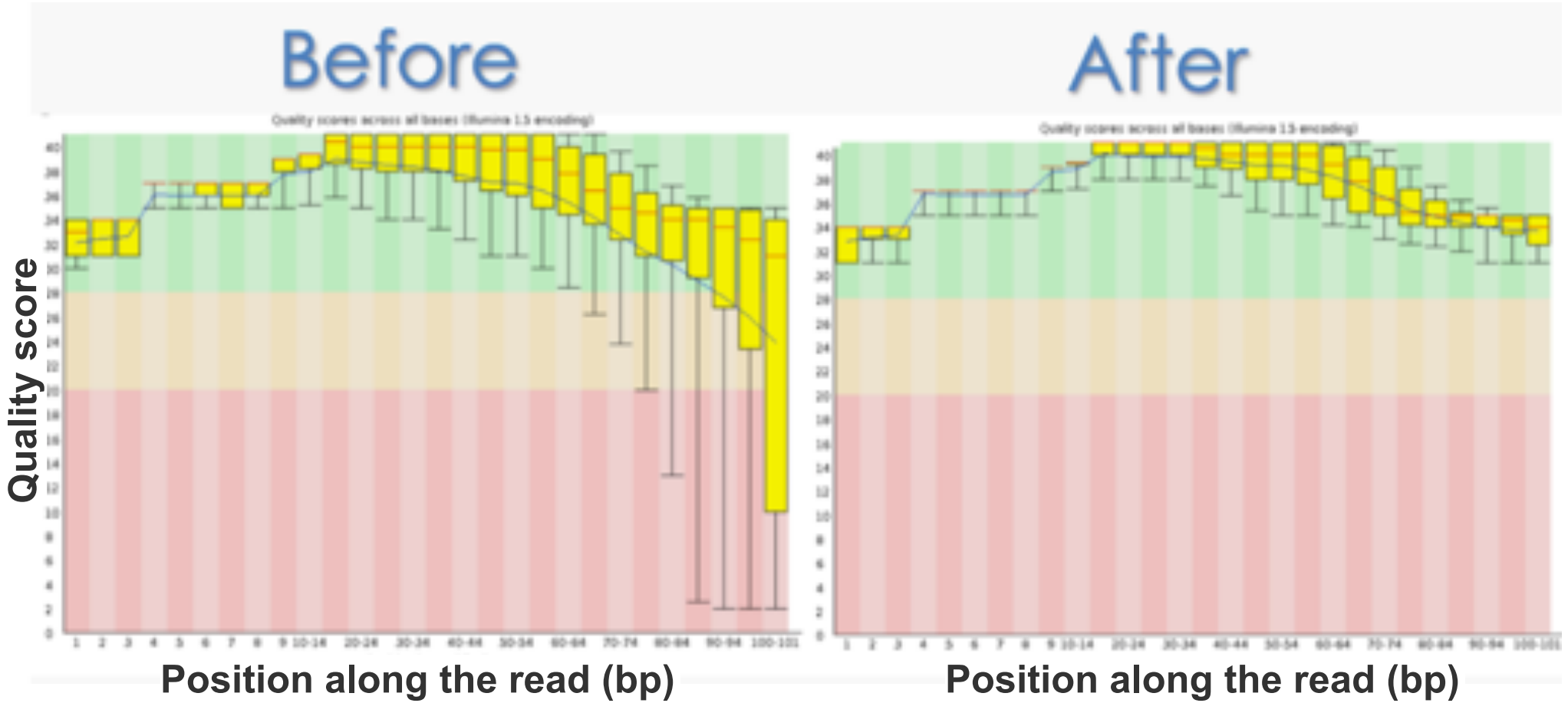


Each nucleotide has a **quality score (Phred score)**
representing the probability that a base was miscalled by the sequencer

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

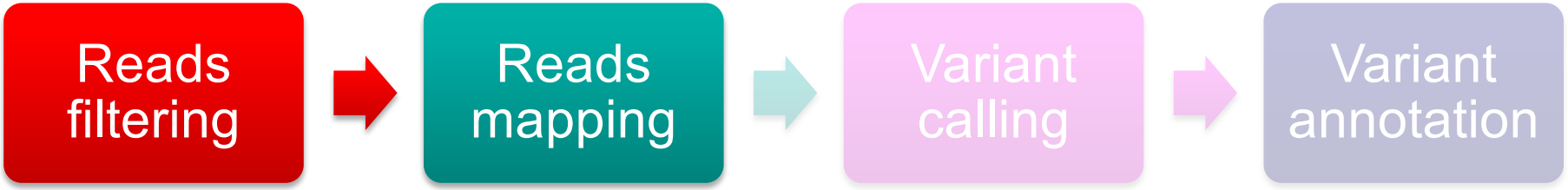
Phred Score	Prob. of incorrect base call	Base call accuracy	Code
10	1 in 10	90%	J
20	1 in 100	99%	T
30	1 in 1'000	99.9%	^
40	1 in 10'000	99.99%	h

Quality-based reads trimming

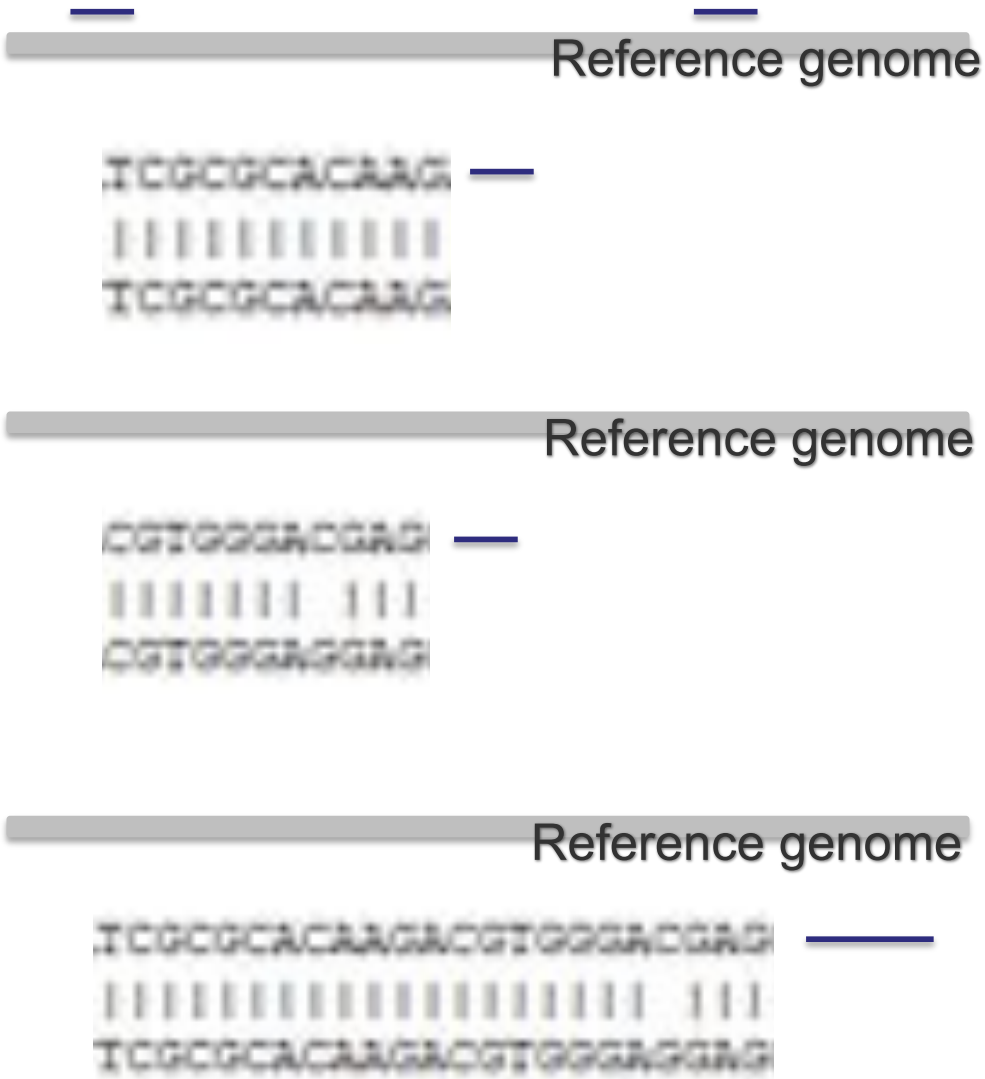


PART II

Quality control



Let's align the reads

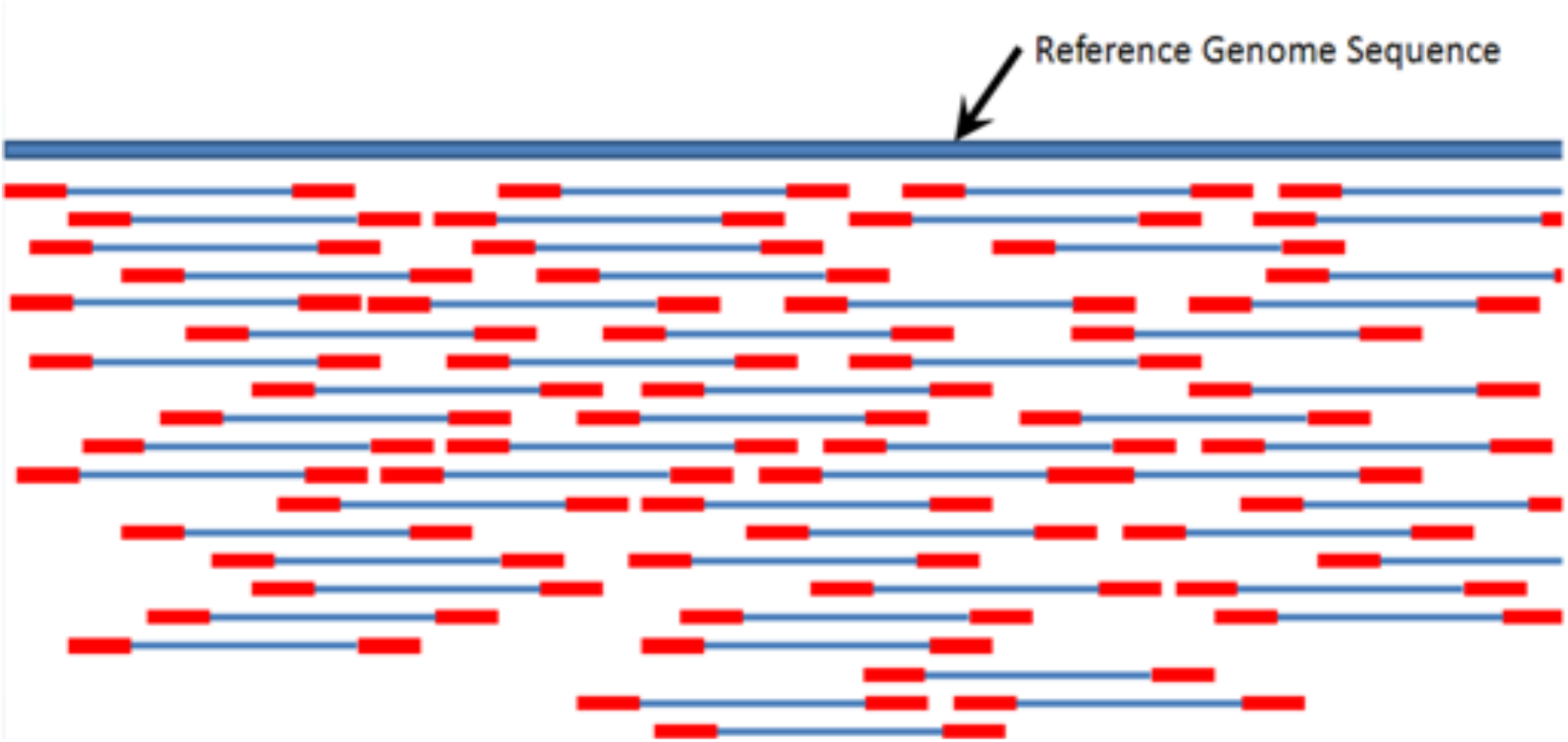


! **Short reads** are likely to map at several positions along the reference genome

! **Mismatches** and **gaps** allowed
→ algorithms have scoring functions

! **Longer reads** are less ambiguous
→ but computationally more expensive

Mapping: finding the best position for each read



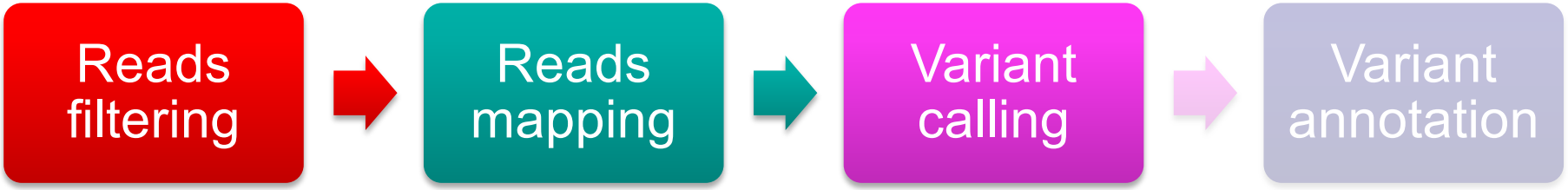
Famous mappers

BWA (Li and Durbin 2009)

Bowtie (Langemead et al. 2009)

PART II

Quality control



Variant calling: putting it all together

- Is it a true variant or a sequencing error?
- Variant callers generally assume that sequencing errors are independent across reads

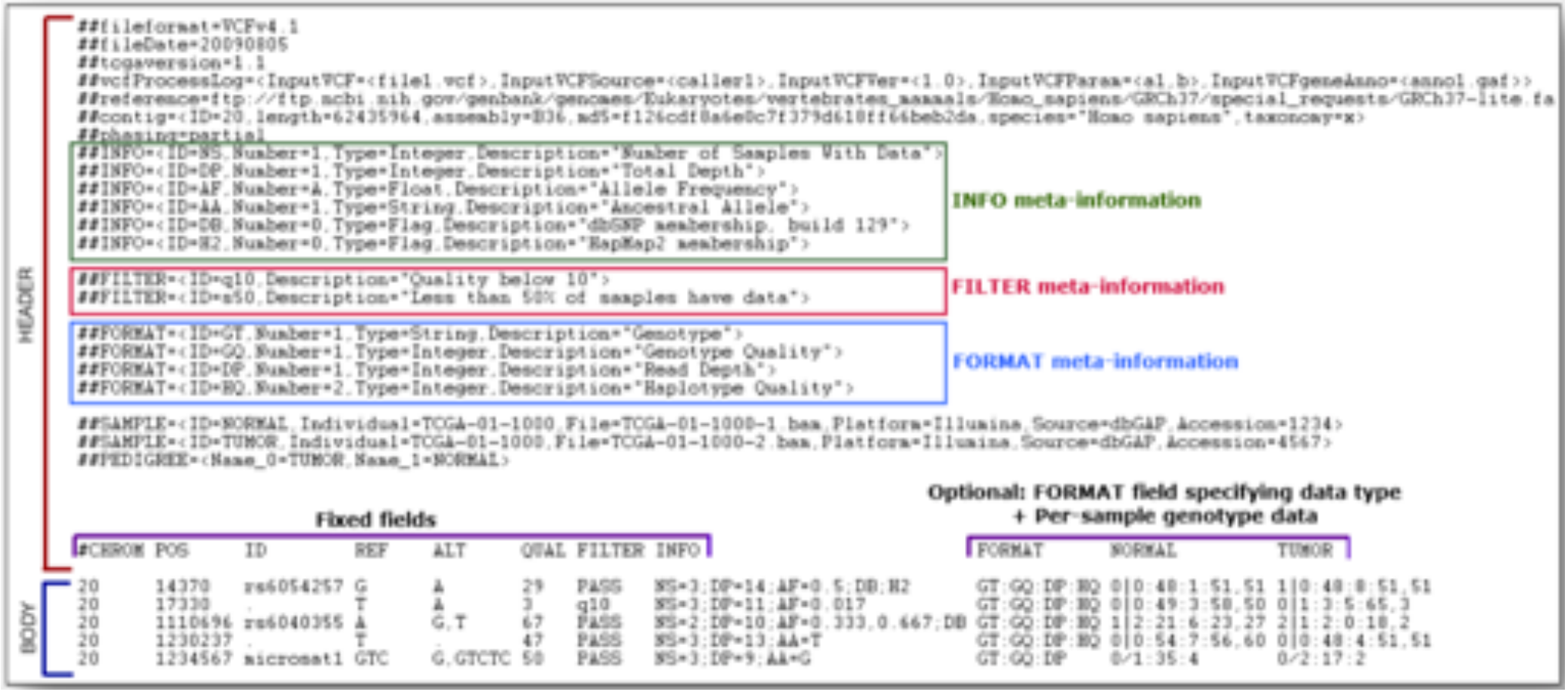
A word of caution: germline vs. somatic caller

- Many popular variant callers are designed for constitutional genome analysis, where variants occur in either 50% (heterozygous) or 100% (homozygous) of the reads.
- Often included in the callers as priors, and thus variants with other allele frequencies (typically for somatic variants) may be filtered out.

➔ Always use a somatic-specific variant caller.

Output of the variant caller: VCF

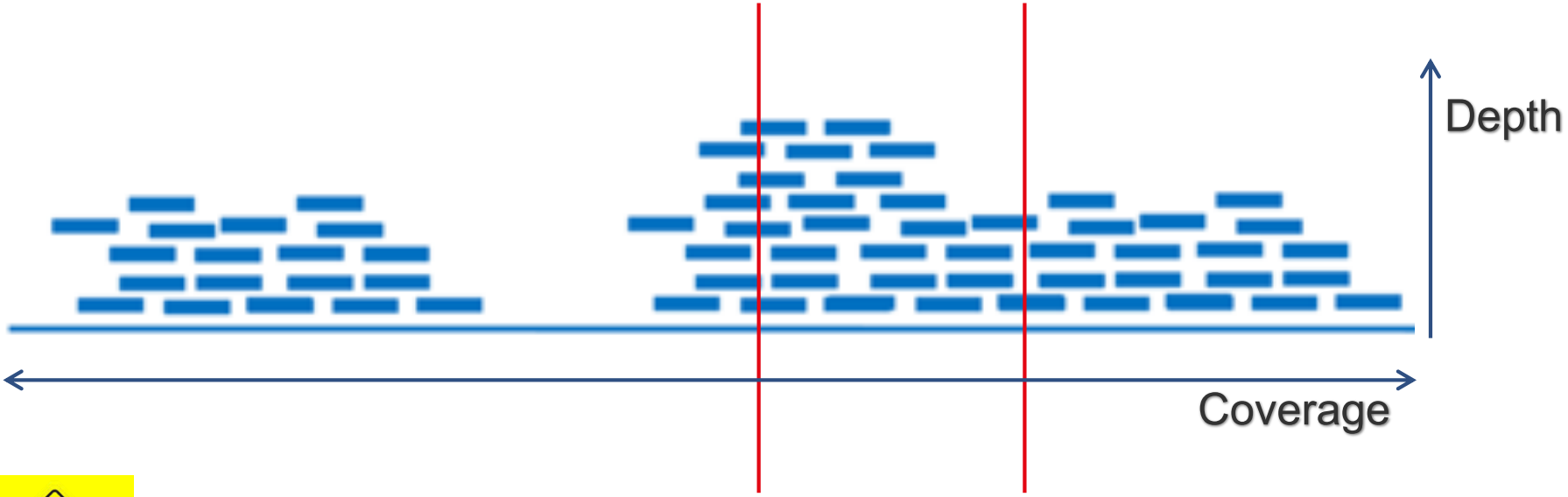
VCF: Variant Call Format



Things to watch out for when assessing variant quality

Depth and coverage

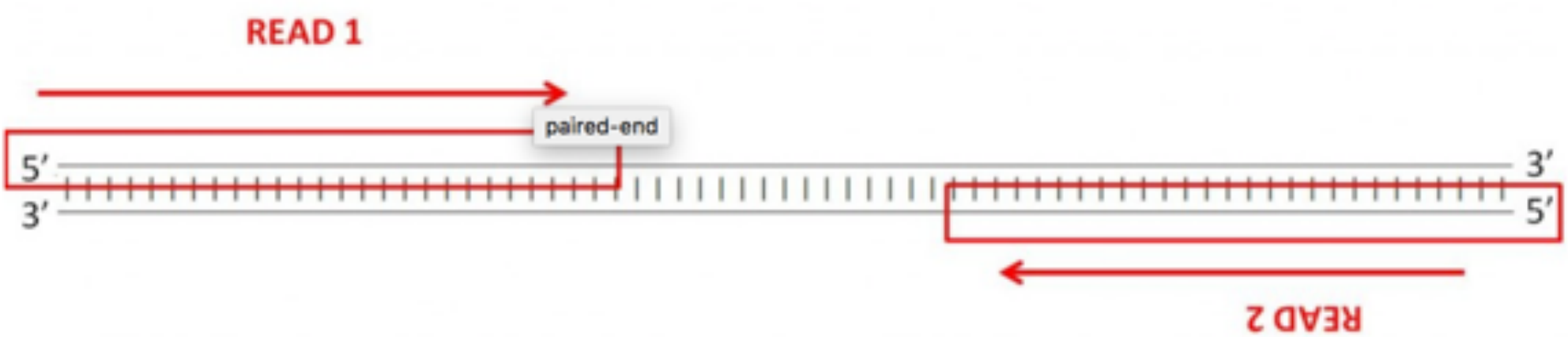
- **Depth:** nb of reads that include a given nucleotide, at a given position (e.g. 1000X)
- **Coverage:** percentage or nb of bases of a reference genome covered with a certain depth, e.g. 90% at 5X



Many people use "coverage" for "depth".
Watch out if % or X

Strand bias

- **Both DNA strands are sequenced**
- Bias occurs when one strand is favored over the other
- Normal mutations should occur on both + and – strands with equal frequencies

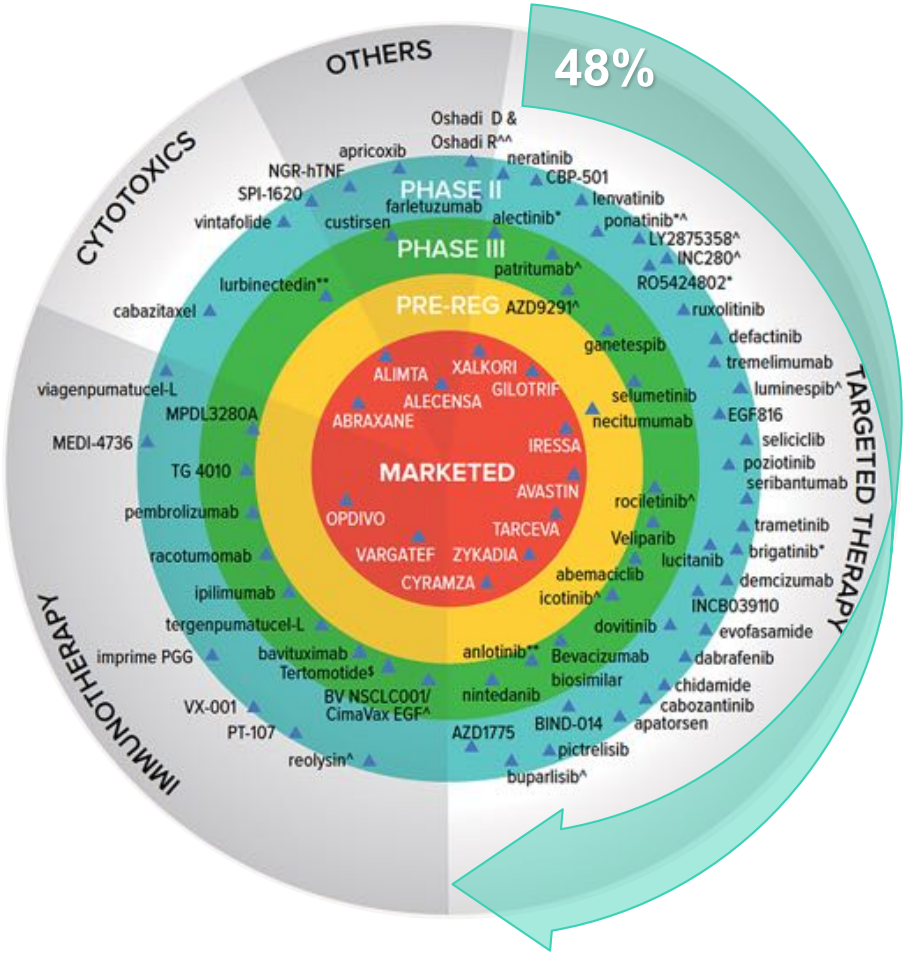


http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/

PART III

Variant annotation and interpretation

Predictive: personalized molecular oncology



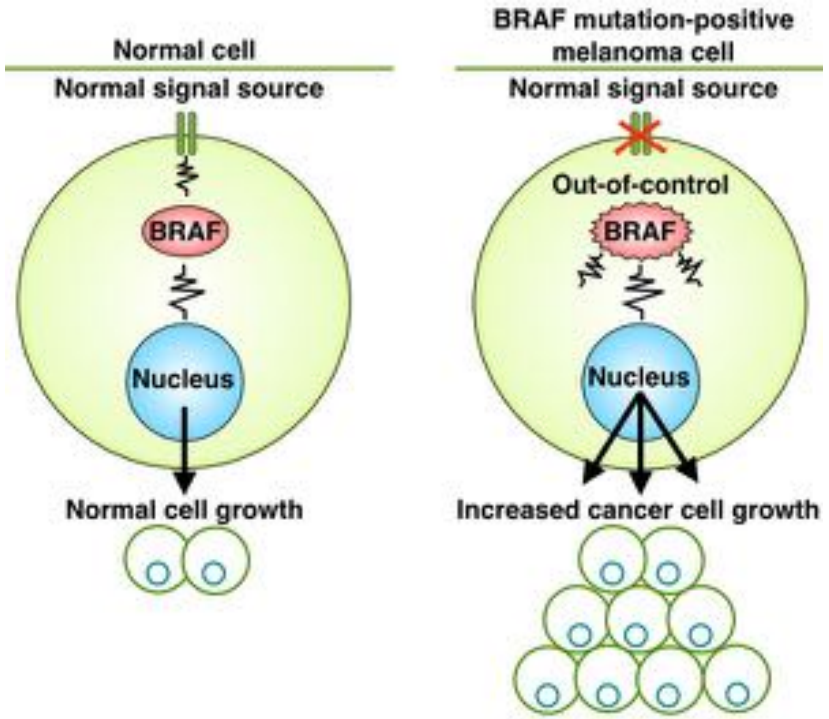
Lung cancer: about 100 drugs at different stages of development process

Source: IMS Health R&D Focus™, clinicaltrials.gov, company websites Dec. 2014

Medical genetics: focus on pathogenicity

- **Pathogenic variant**: genetic alteration that increases an individual's **susceptibility or predisposition** to a certain disease.

- **5 levels (guidelines)**
  - Pathogenic
  - Likely pathogenic
  - Variant of unknown significance
  - Likely benign
  - Benign



www.cancer.gov
www.sinhhocphantu.net

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

CrossMark

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

Finding actionable variants

"Unlike interpretation of germline sequence variation, which focuses on pathogenicity (…), ***interpretation of somatic variants should be focused on their impact on clinical care****".*

Finding actionable variants

Filtered list
of variants

Predisposition	*Indicates risk to develop a disease*
Diagnostic	*Supports disease characterization*
Prognostic	*Indicates disease evolution*
Predictive	*Supports treatment decisions*

Other important questions

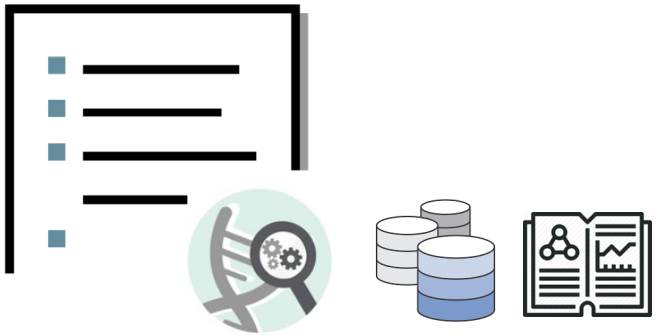
- Is it prevalent in the cancer subtype of interest?
- Is it known in other cancer subtypes or diseases?
- Is it present in the general population?
- Are there other known variants in the same gene?
- Is it related to an ongoing clinical trial?
- What is the evidence level? Observed vs. predicted

PART III

Variant annotation and interpretation

… bioinformatics at the rescue

Bioinformatics to the rescue… for annotation



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

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

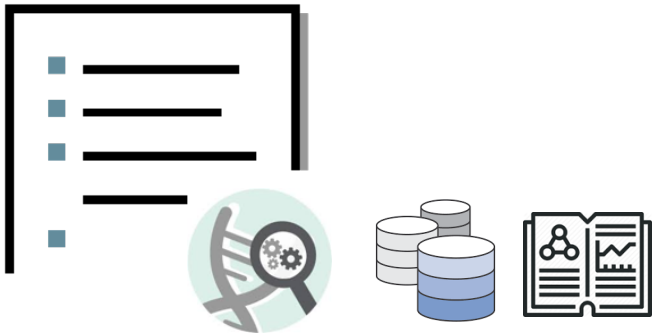
Locating variants

- Convert **genomic coordinates** (chromosome, position) to the corresponding **cDNA/amino-acid coordinate** system

- **HGVS nomenclature** (*http://varnomen.hgvs.org*)
  - Substitution c.76A>T
  - Deletion c.76delA
  - Insertion c.76_77insG
  - Protein sequence p.Lys76Asn
  - Genomic sequence g.476A>T

- **Important to store for tracking**
  - Version of the human genome assembly
  - Accession and version of the mRNA transcripts

Bioinformatics to the rescue… for annotation



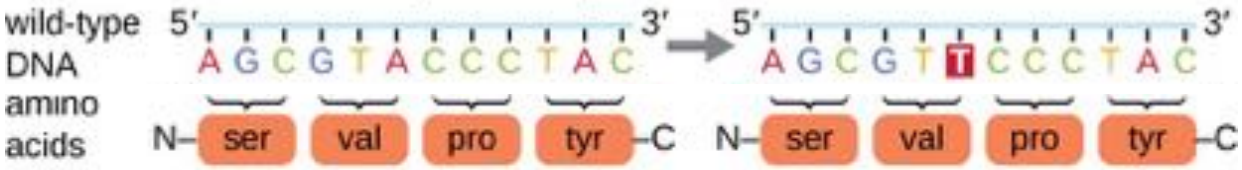
- **Genes** and **transcripts** affected by the variant
- **Location** of the variants (e.g. coding, noncoding region…)
- Predict **variant effect** (e.g. stop gained, missense…)
- Predict **variant impact** on protein function
- Retrieve annotations from **public databases**

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

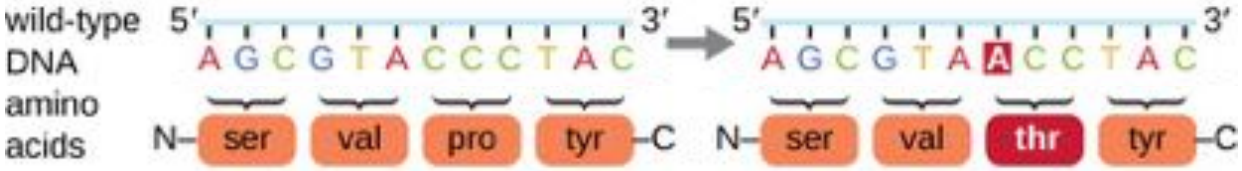
Variant effects on the protein

point mutation: substitution of a single base

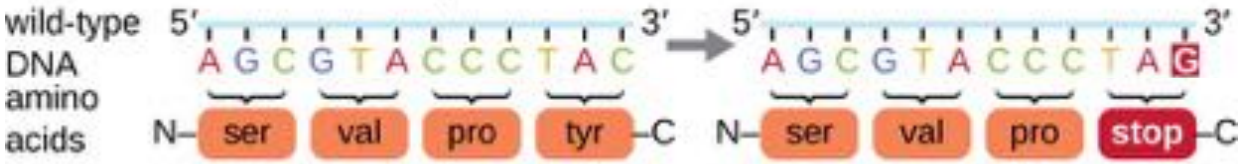
silent: has no effect on the protein sequence



missense: results in an amino acid substitution



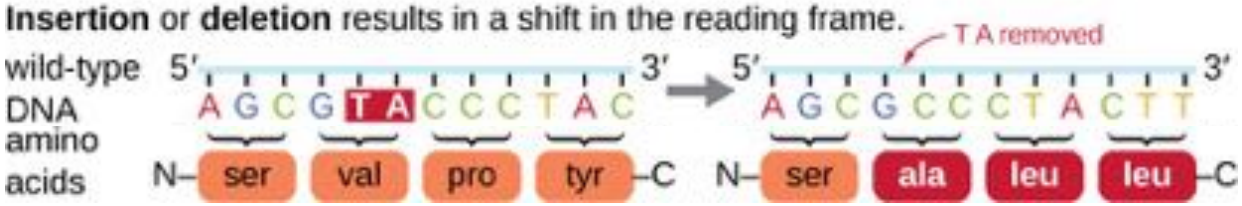
nonsense: substitutes a stop codon for an amino acid



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

Variant effects on the protein

frameshift mutation: insertion or deletion of one or more bases

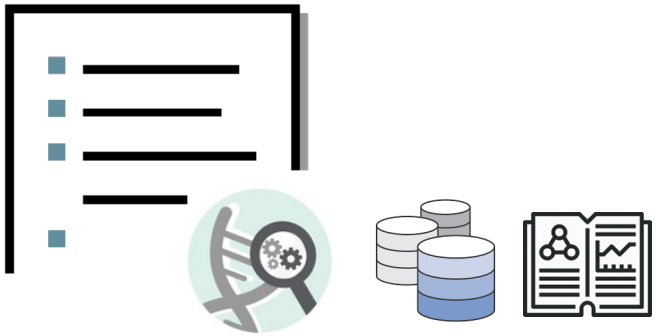


There is an ontology to describe variant effects: sequence ontology

http://www.sequenceontology.org/browser/obob.cgi

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

Bioinformatics to the rescue… for annotation



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

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

What is the impact of non-silent mutations?

- Is the mutation in a **gene**, **exon, regulatory** region…?
- Is the mutation in an **evolutionarily conserved** region accross species?

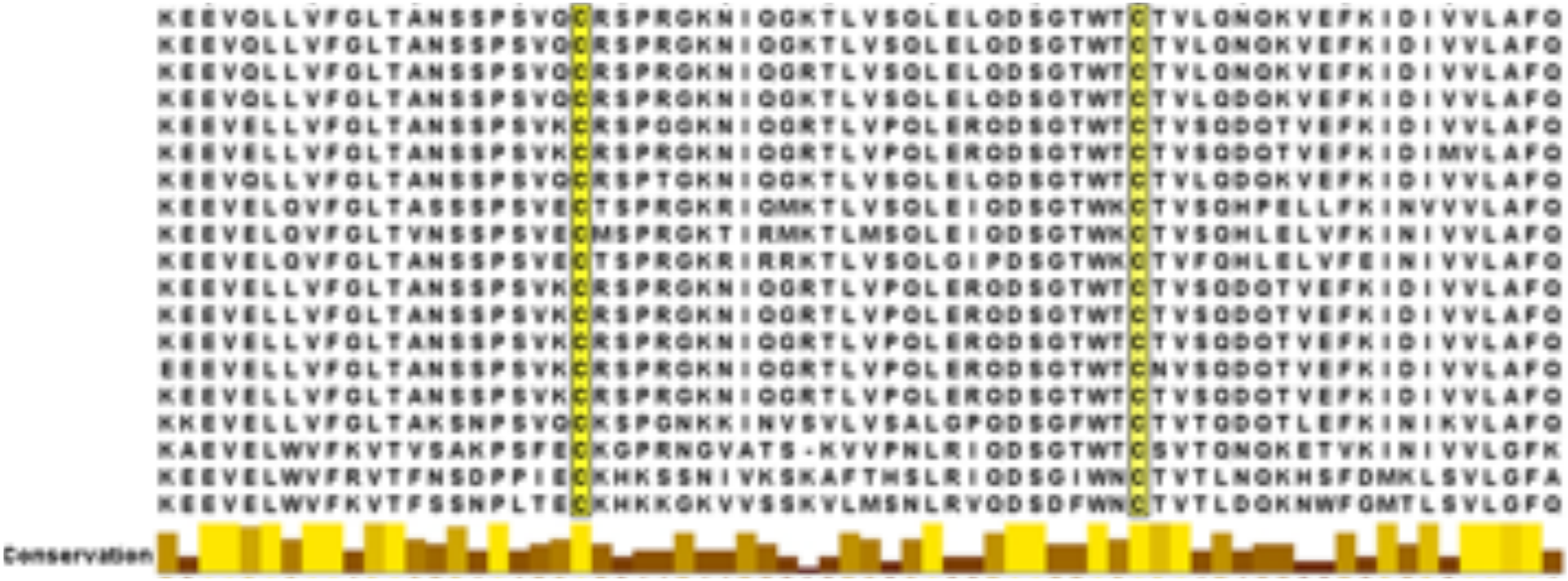


Image from Front Pharmacol. 2015 Mar 10;6:1. doi: 10.3389/fphar.2015.00001

Predicting the impact: examples of tools

not exhaustive

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		aa conservation in related seq.	aa conservation + structural feat.	aa conservation + protein tolerance to mutations

SnpEff impact rules

Putative Impact	Sequence Ontology term
HIGH	start_lost
HIGH	stop_gained
HIGH	stop_lost

http://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf

What if different tools predict different things?

ACMG/AMP Guidelines

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. Genet Med. 2015;17:405–24.

- Use a combination of tools
- Only keep variants with consensus predictions

But which combination of tools to use?

- A recent study proposes combinations of tools with increased concordance for clinically relevant variants

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1353-5

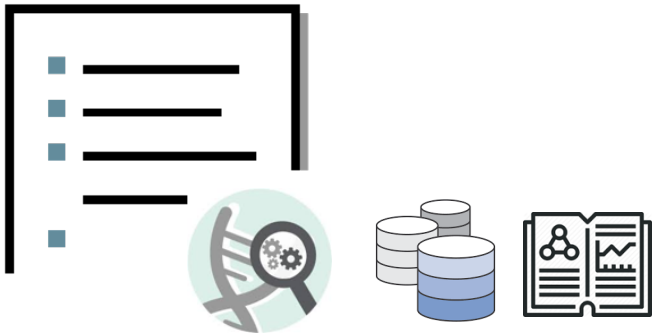
I found a damaging mutation: is it always bad?

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

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

Keep the gene function in mind when interpreting its deleteriousness

Bioinformatics to the rescue… for annotation



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

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

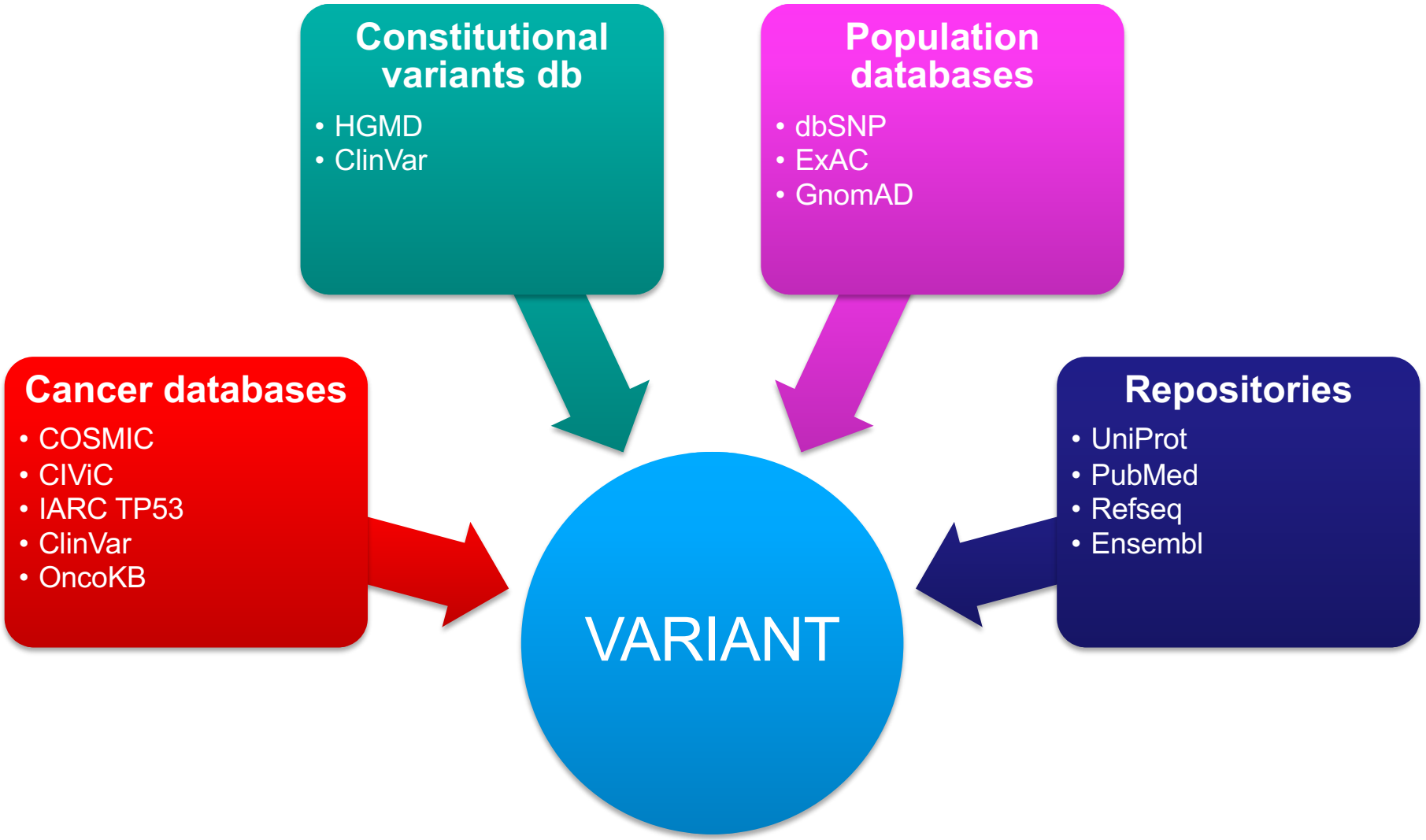
PART III

Variant annotation and interpretation

… with knowledge-bases

Annotating a variant: knowledgebases

not exhaustive

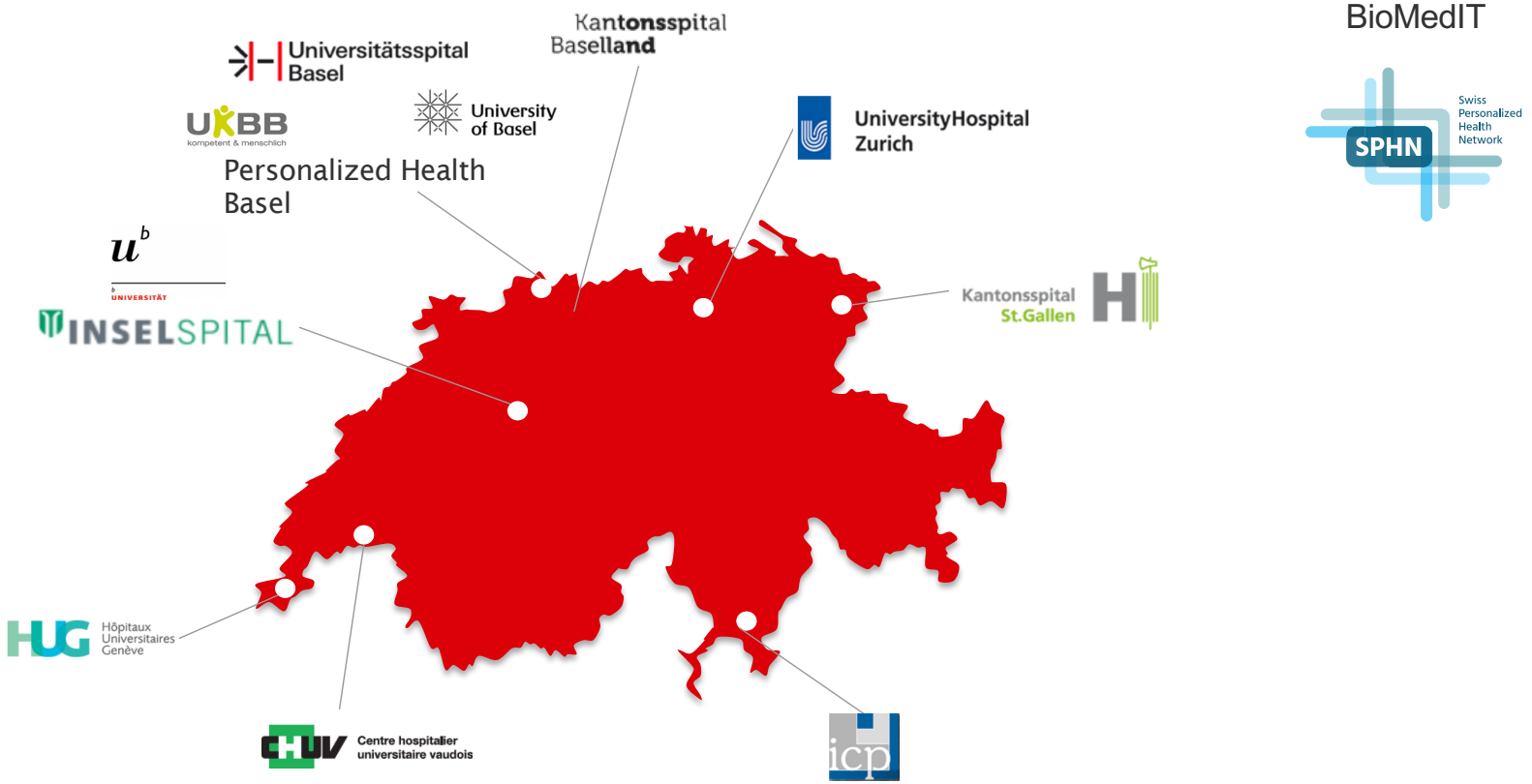


Real-life implementations and constraints

Swiss Variant Interpretation Platform

SVIP

BioMedIT

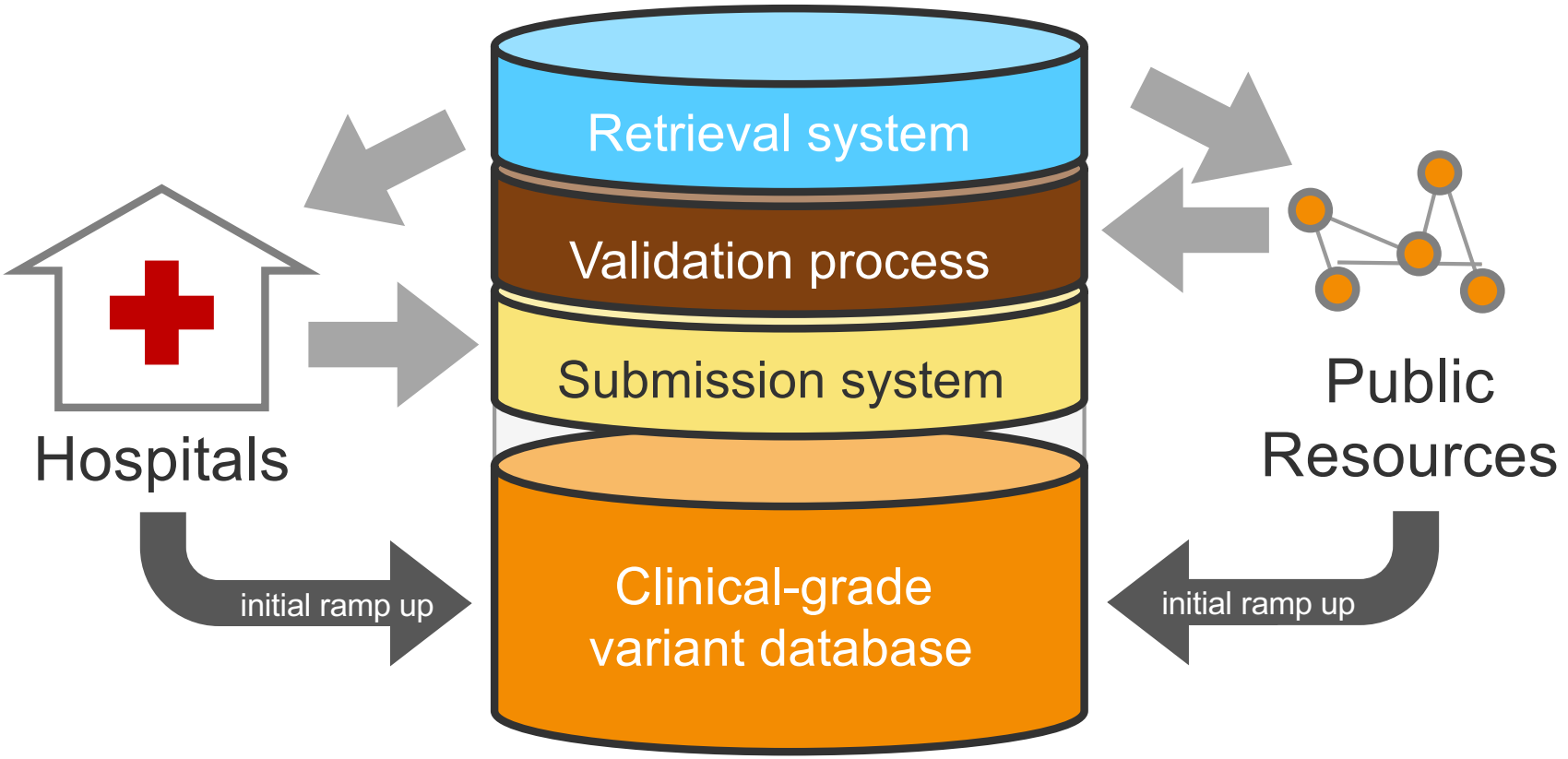


- Centralize cancer variants identified in Swiss patients in one single place, agreeing on their clinical interpretation
  - *www.svip.ch*

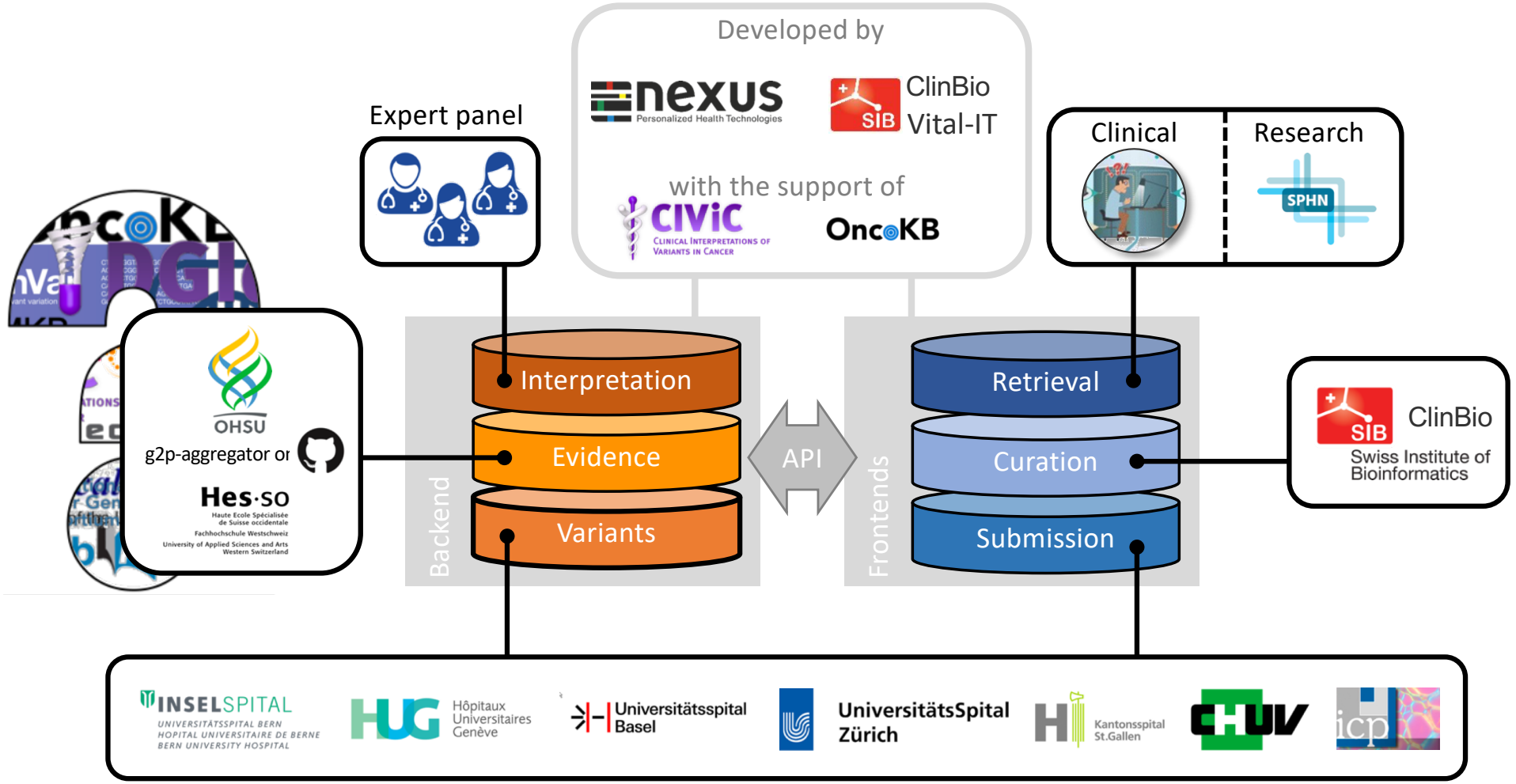
SGPath SSPath
Société Suisse de Pathologie
Schweizerische Gesellschaft für Pathologie
Swiss Society of Pathology

SWISS SOCIETY OF MOLECULAR PATHOLOGY

Platform workflow…



…but there's more to it



Specific requirements

- **Contain some personal identifiable data**

- **Two access layers**
  - Clinical layer, accessible to partner clinicians only
  - Research layer, public

- **Need for a specific IT and system architecture**

- **Need for a specific legal framework**
  - Consortium Agreement
  - Data Transfer and Use Agreement
  - Compliance with GDPR (catalog of processing…)
  - …….

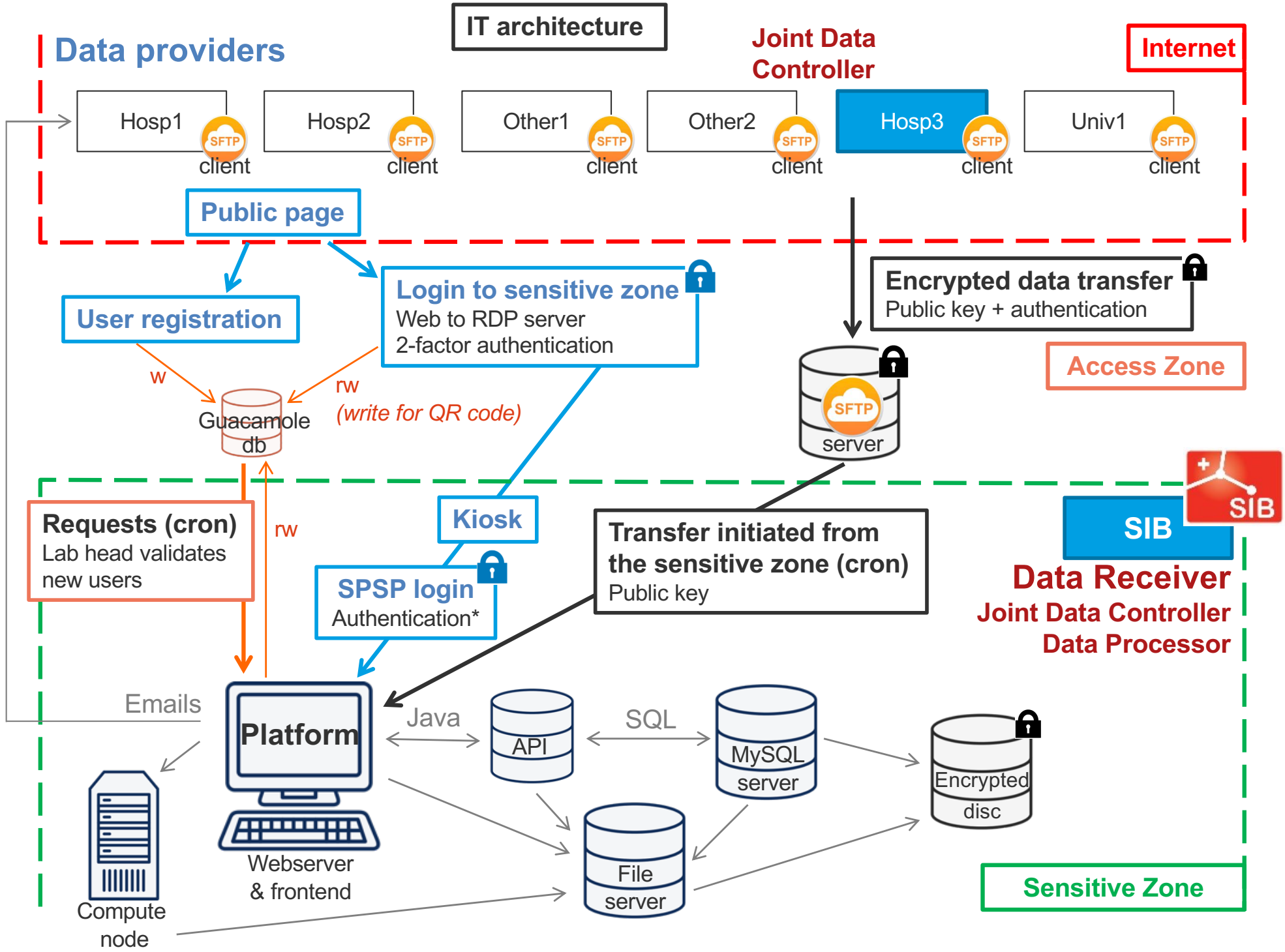
Swiss Pathogen Surveillance Platform

FNSNF
Swiss National Science Foundation

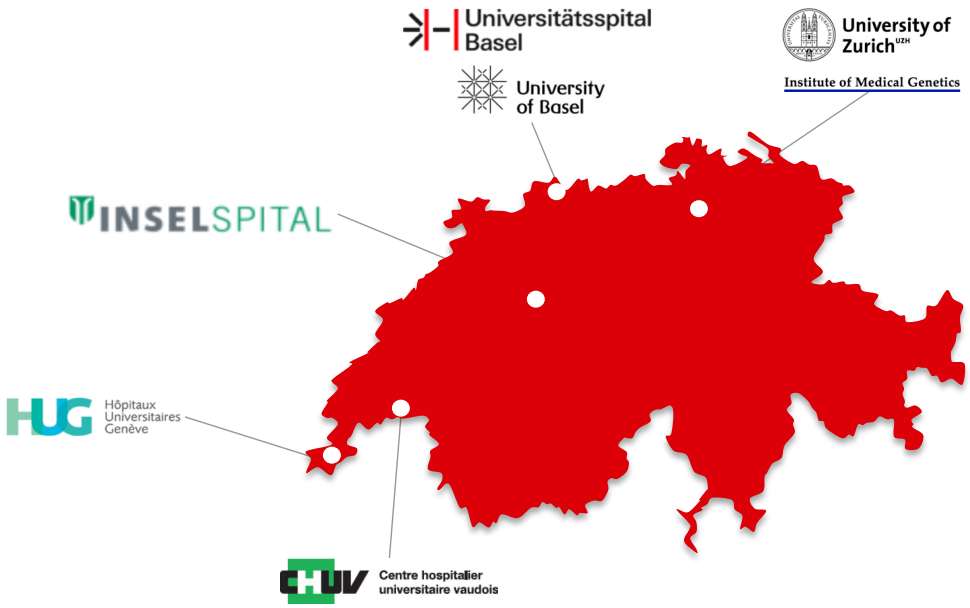
72 NRP Antimicrobial Resistance
National Research Programme

Swiss national project
led by Dr. Egli

- Common platform for microbial WGS analyses
- Identify risks of multi-drug resistant bacterial pathogens
- Explore risks by predicting dynamics of spread
- Inter-cantonal, standardized data, harmonized methods


IT architecture
Data providers
Joint Data Controller
Internet
Hosp1
Hosp2
Other1
Other2
Hosp3
Univ1
SFTP client
Public page
User registration
Login to sensitive zone
Web to RDP server
2-factor authentication
Encrypted data transfer
Public key + authentication
Access Zone
w
rw
(write for QR code)
Guacamole db
SFTP server
Requests (cron)
Lab head validates new users
rw
Kiosk
SPSP login
Authentication*
Transfer initiated from the sensitive zone (cron)
Public key
SIB
Data Receiver
Joint Data Controller
Data Processor
Emails
Platform
Java
API
SQL
MySQL server
Encrypted disc
Webserver & frontend
File server
Compute node
Sensitive Zone

SwissGenVar



Swiss Personalized Health Network

- Collection of genetic variants identified in patients by Swiss clinical genetic laboratories
- With patients clinical phenotype
- Even more sensitive data!

Specific training needs

Training & outreach

- Certificate of Advanced Studies in Personalized Molecular Oncology (pmo.unibas.ch)
- NGS QC and annotation for cancer diagnosis
- MOOC on Precision Medicine
- ESCMID Workshop on bioinformatics for bacterial genomics
- Outreach events (health exhibitions, scientific cafés, schools kids…)

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

SIB Clinical Bioinformatics

Ivo de Carvalho

Valérie Barbié

Miriam Tesfai

Florent Tassy

Yann Christinat
(HUG, Molecular Pathology)

Abdullah Kahraman
(USZ, Molecular Pathology)

Dillenn Terumalai

Aitana Lebrand

Steffen Pade

Blanca Cabrera Gil