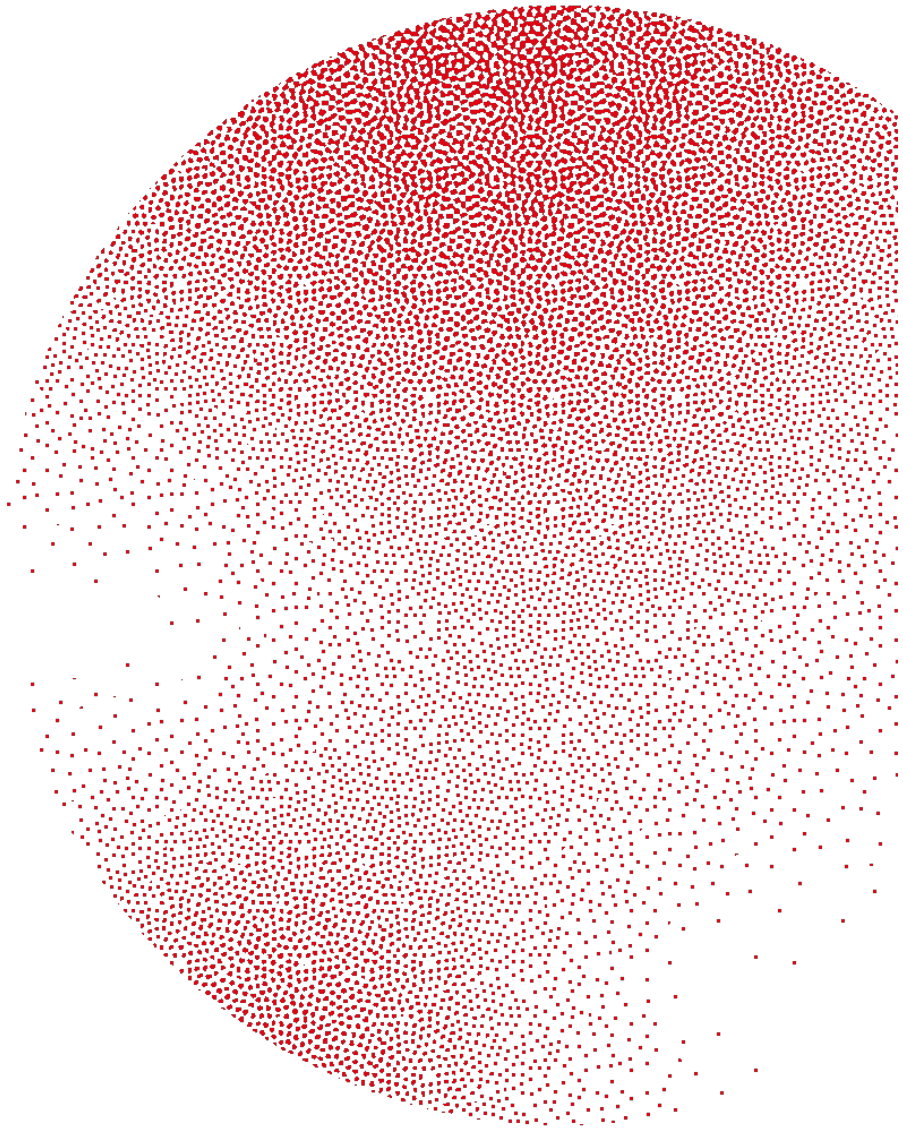
Swiss Institute of Bioinformatics

INTRODUCTION TO BIOINFORMATICS:

Clinical Bioinformatics

V. Barbié, Director Clinical Bioinformatics

Zürich, 13 October 2026



Outline

What is
clinical bioinformatics

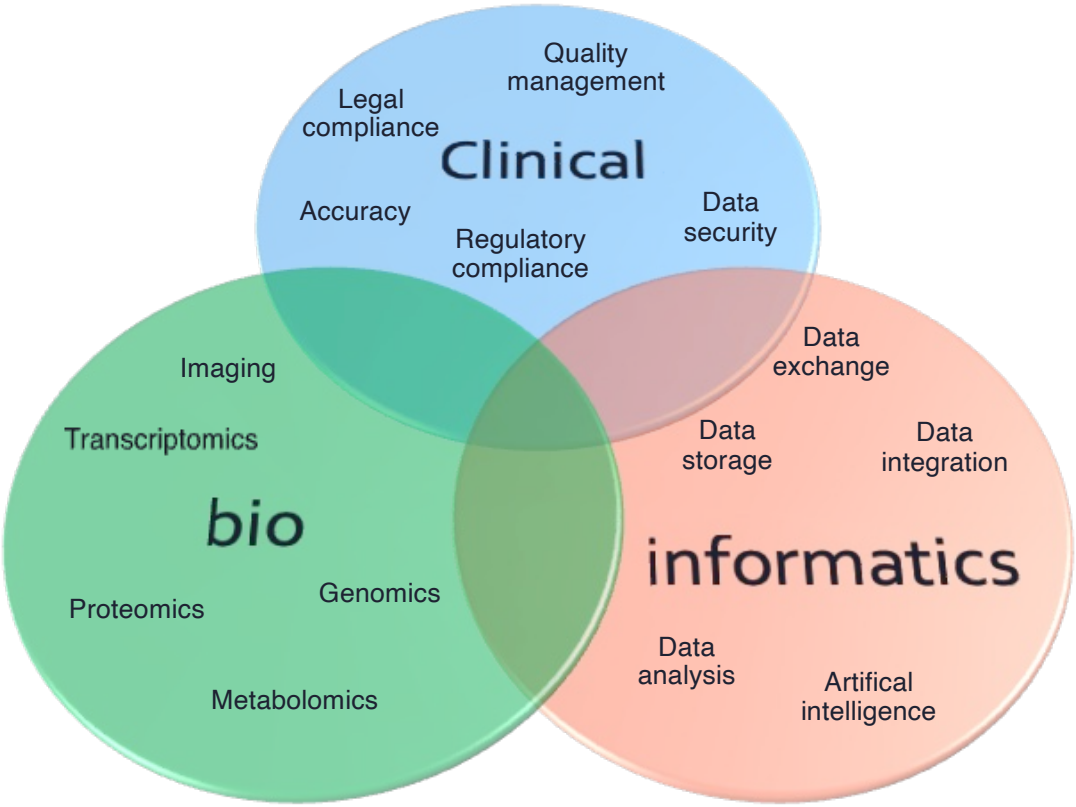
Why clinical bioinformatics?

Next Generation Sequencing (NGS) in medical diagnosis

Overview of an oncology NGS diagnostic pipeline

Other considerations

What is clinical bioinformatics?



Outline

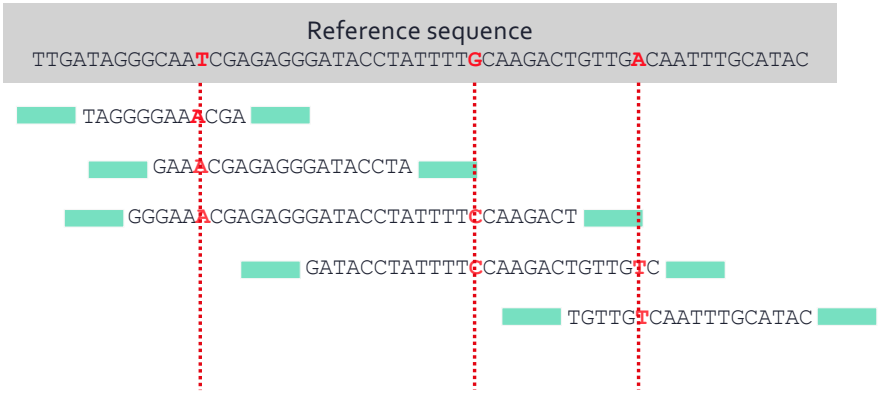
What is
clinical bioinformatics

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

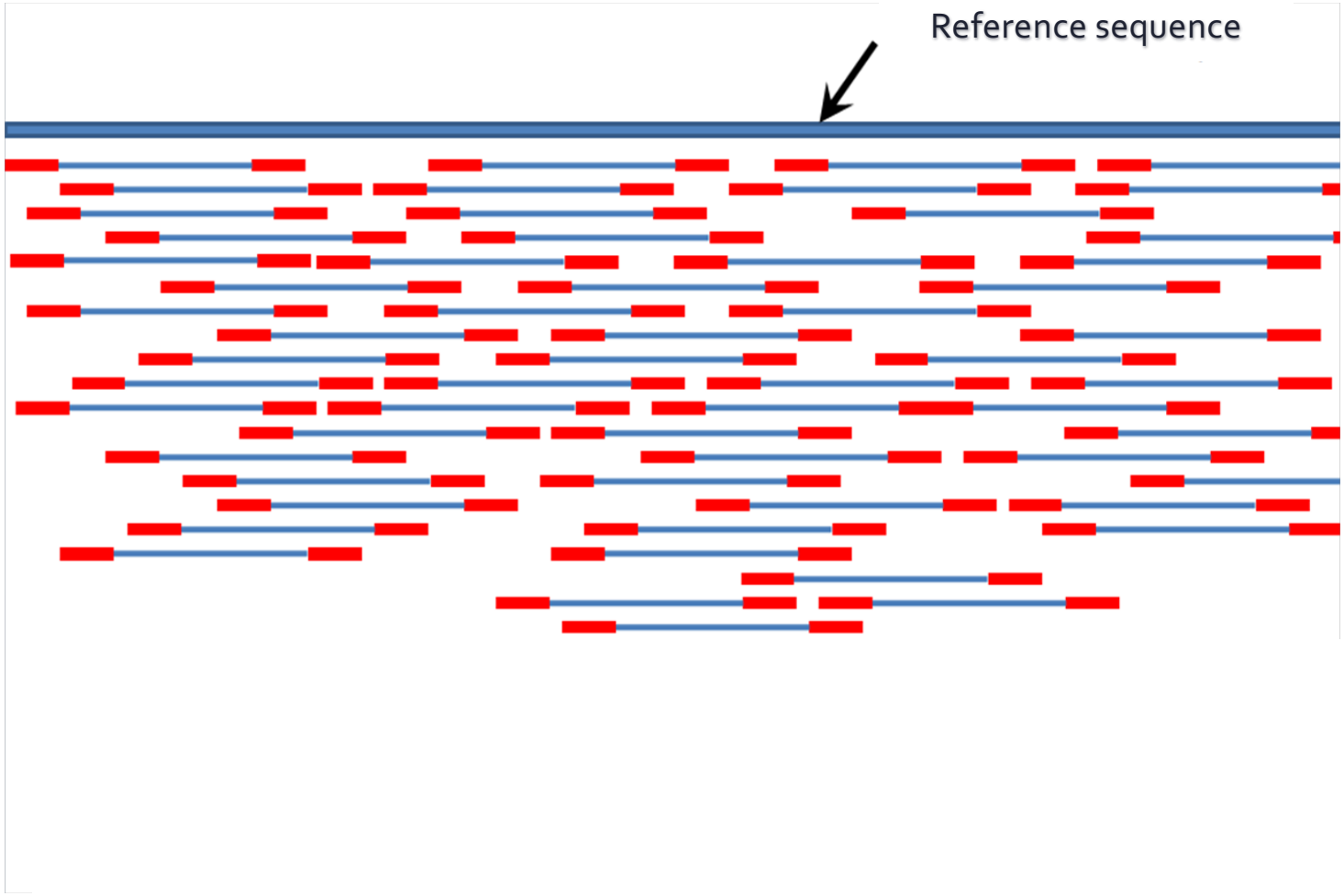
Overview of an oncology NGS diagnostic pipeline

Other considerations

Next Generation Sequencing principle

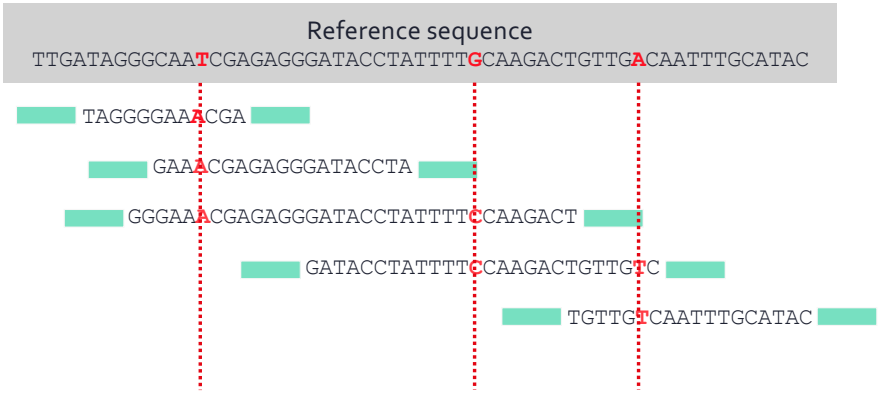


Next Generation Sequencing principle

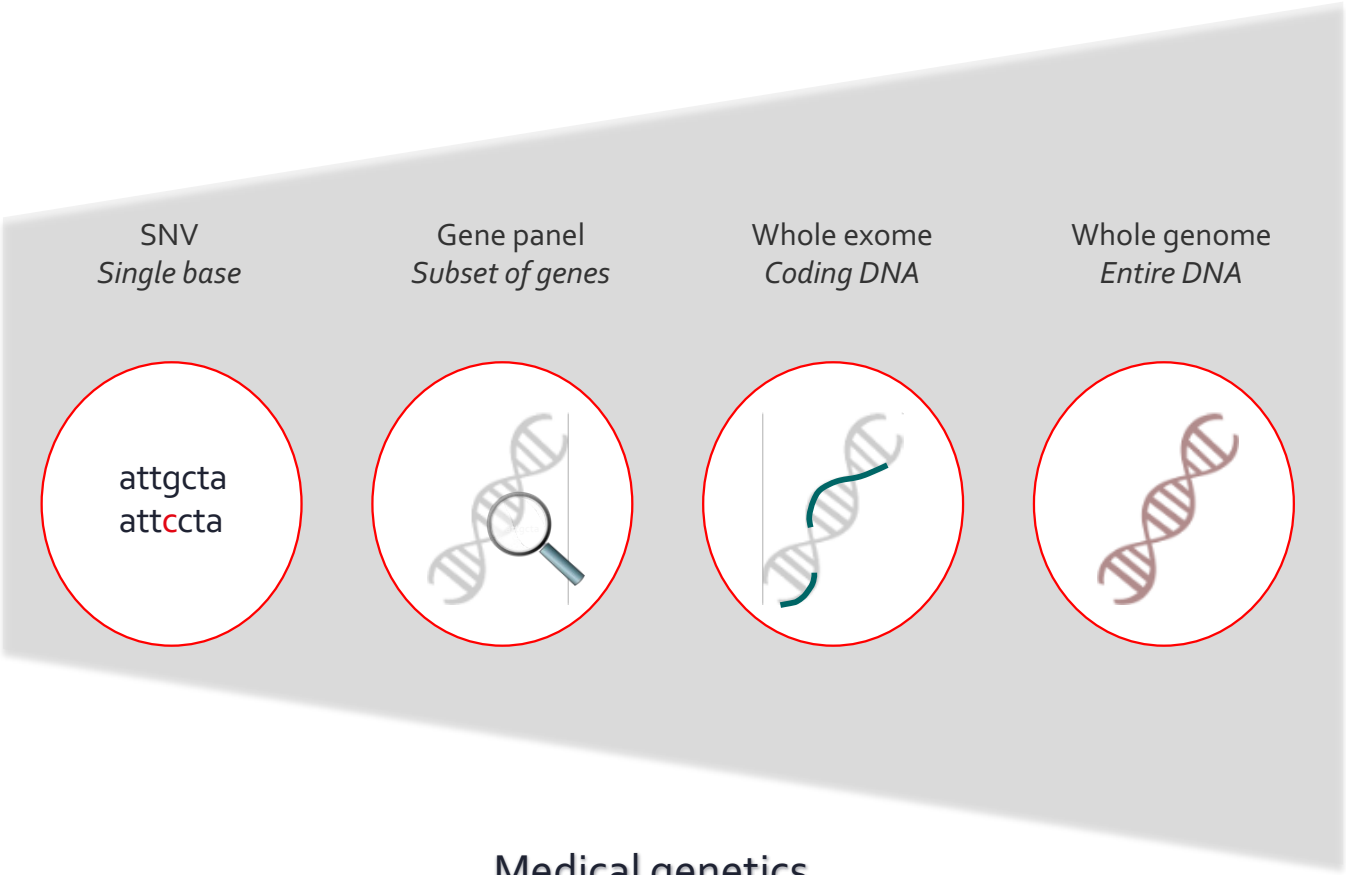


Examples of NGS clinical applications

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

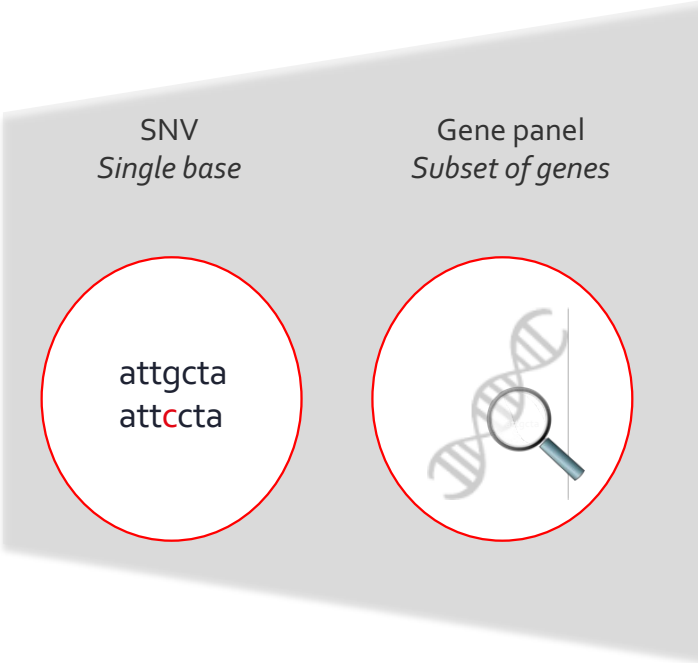


A matter of scale



Medical genetics

A matter of scale



Oncology

- Results turn around time
- Limited reimbursement
- Clinically-actionable variants
- Incidental findings management

Outline

What is
clinical bioinformatics

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

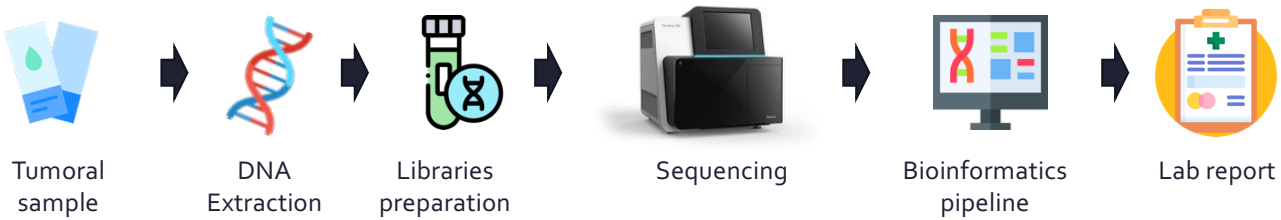
Overview of an oncology NGS diagnostic pipeline

Other considerations

NGS in cancer diagnosis

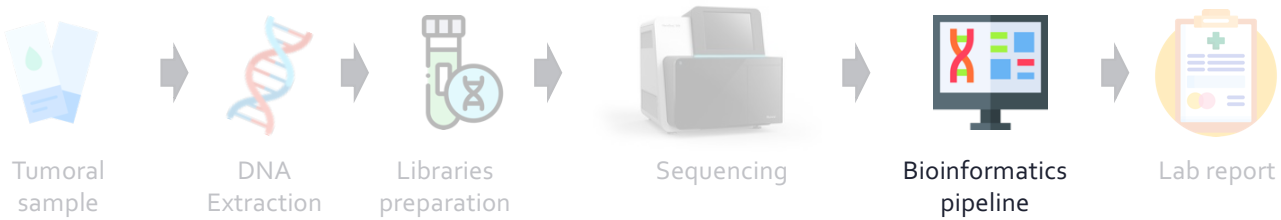
- Identify single nucleotide variants (SNVs), insertions-deletions (indels) to inform clinical management

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

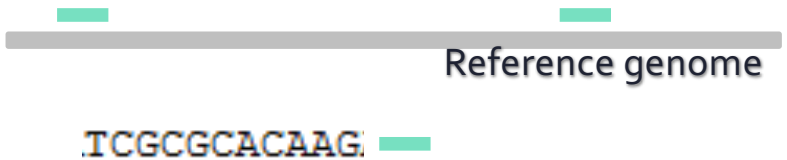
Identifier	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence	TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign	+
Quality scores	hhhhhhhhhhghhghhhhhfhhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence	GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign	+
Quality scores	hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd

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

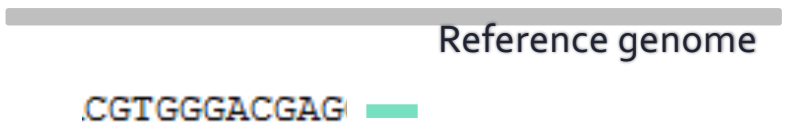
$$Q = -10 \log_{10} P$$

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

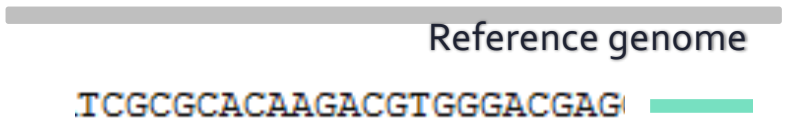
Let's align the reads



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

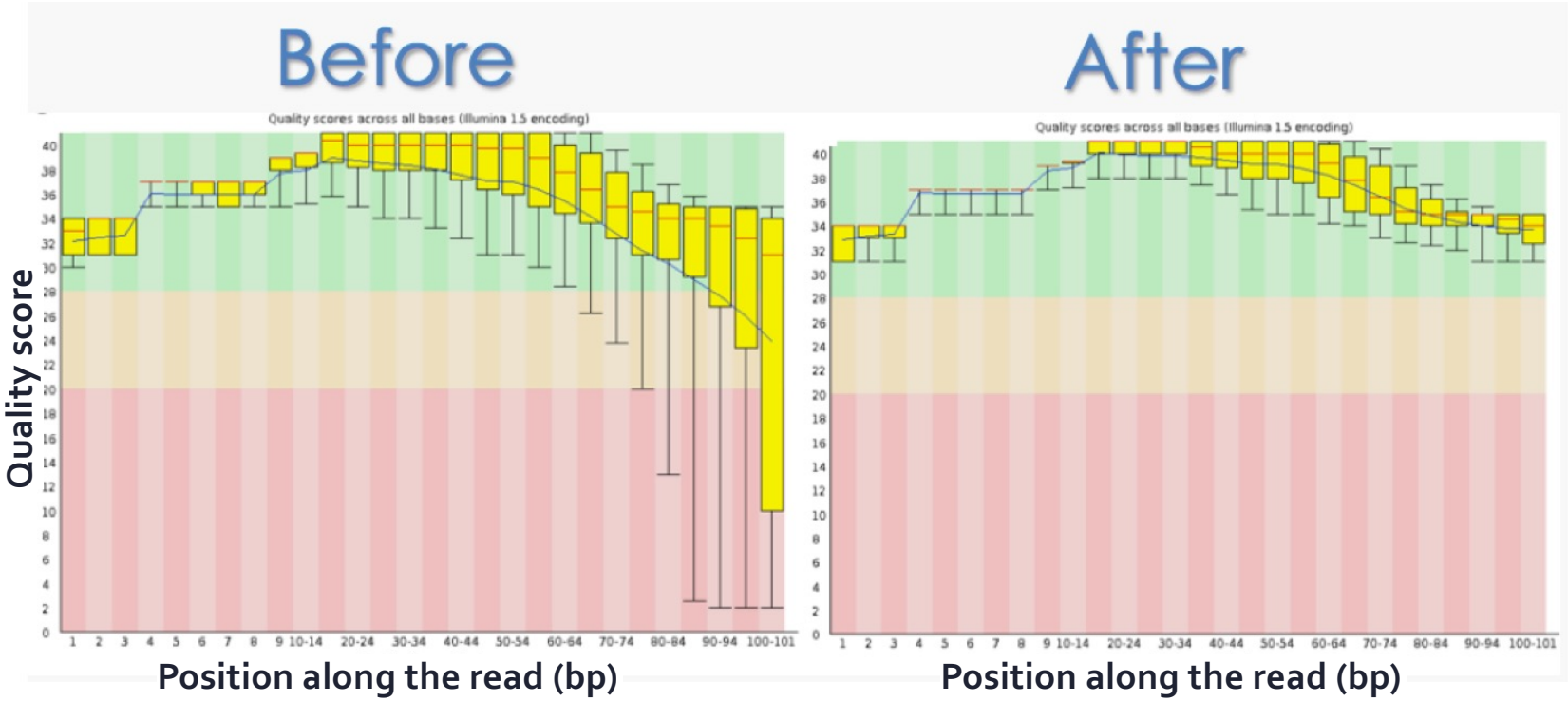


- **Mismatches and gaps allowed**
  - Algorithms use scoring functions

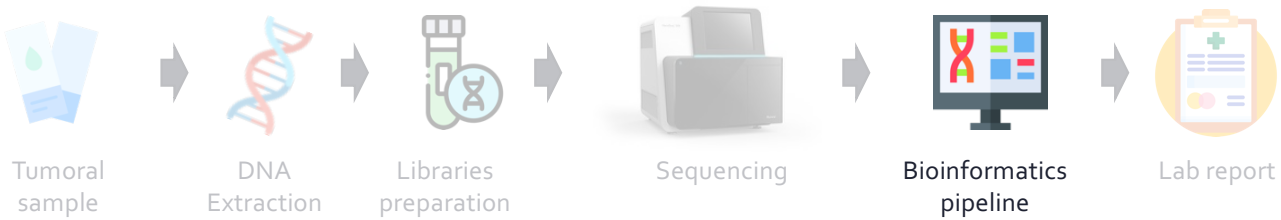


- **Longer reads**
  - Less ambiguous
  - But computationally more expensive

Quality-based reads trimming



Overview of a NGS bioinformatics pipeline

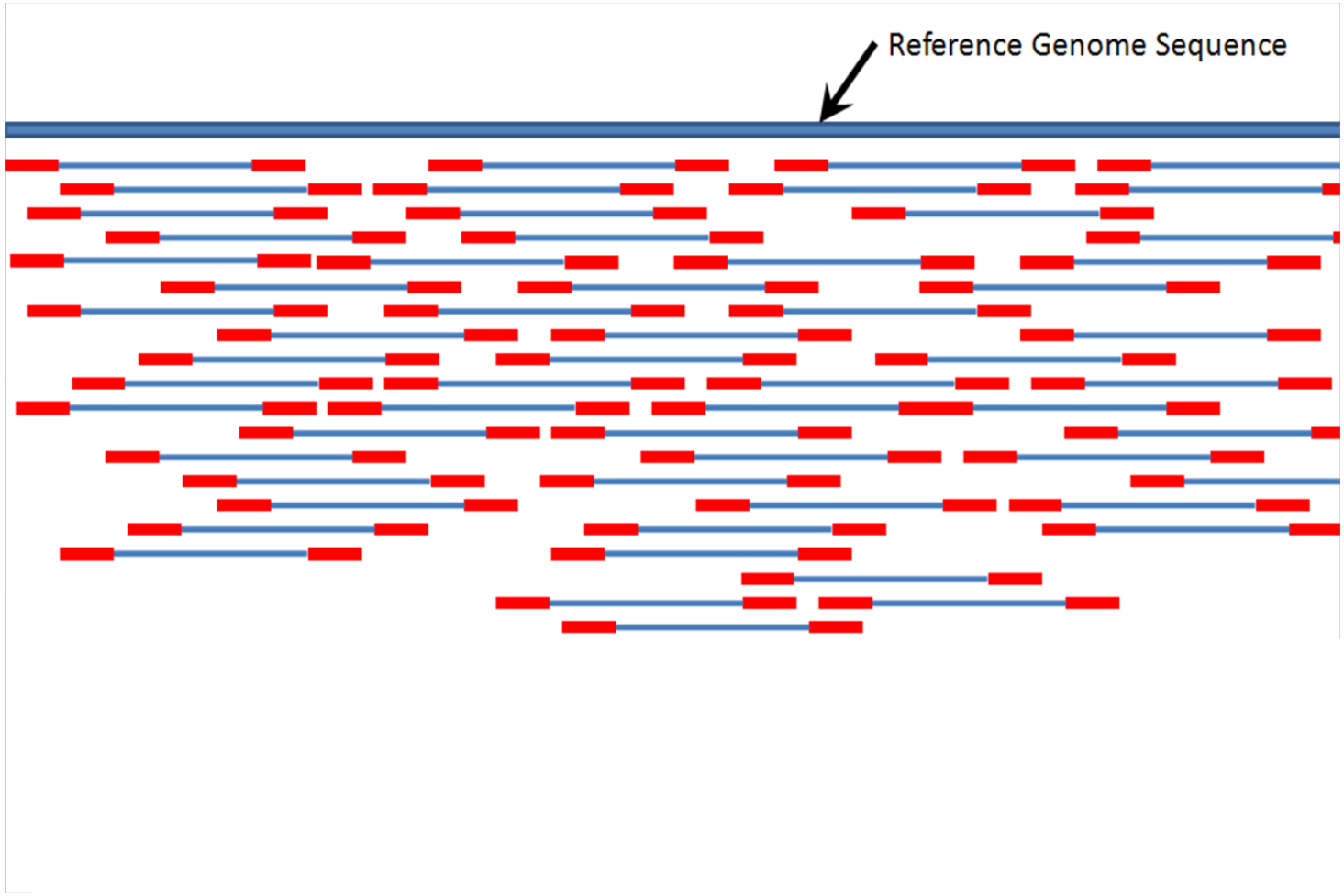


Reads filtering → **Reads mapping**

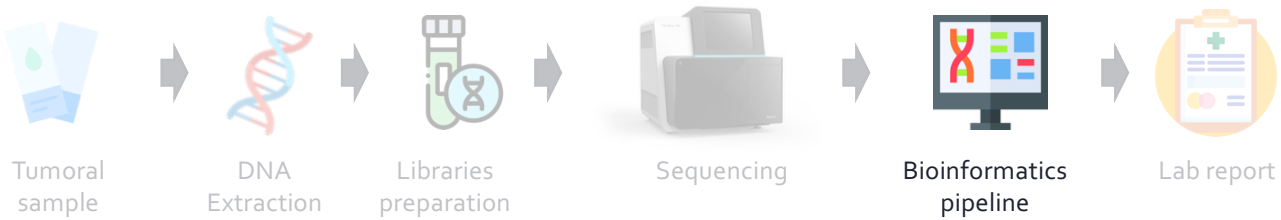
Quality control

Variant identification

Mapping: finding the best position for each read



Overview of a NGS bioinformatics pipeline



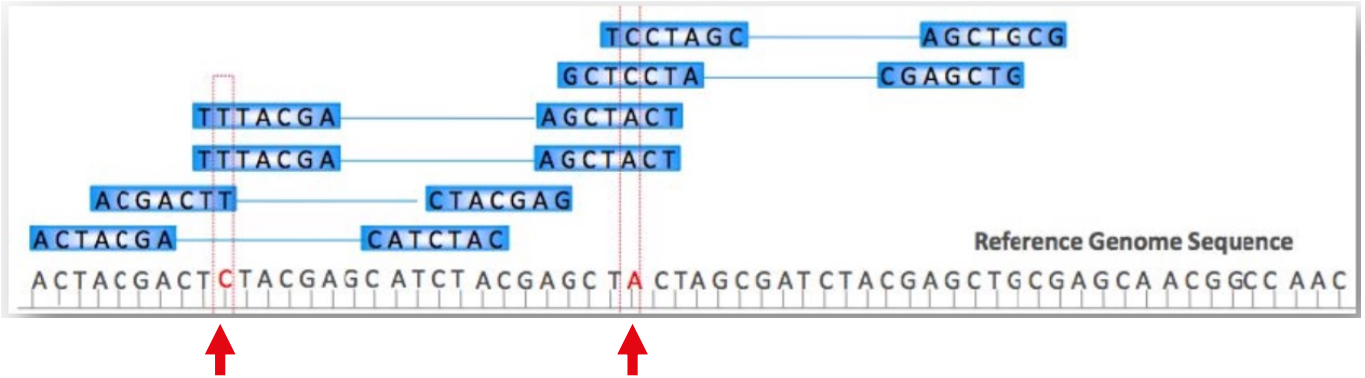
Reads filtering → **Reads mapping** → **Variant calling**

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: Variant Calling Format (VCF)

VCF: Variant Call Format

HEADER

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
```

INFO meta-information

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

FILTER meta-information

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

FORMAT meta-information

```
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

Fixed fields

Optional: FORMAT field specifying data type + Per-sample genotype data

BODY

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

Things to watch out when assessing variant quality

Output of the variant caller: VCF (Variant Call Format)

HEADER

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
```

INFO meta-information

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

FILTER meta-information

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

FORMAT meta-information

```
##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

Fixed fields

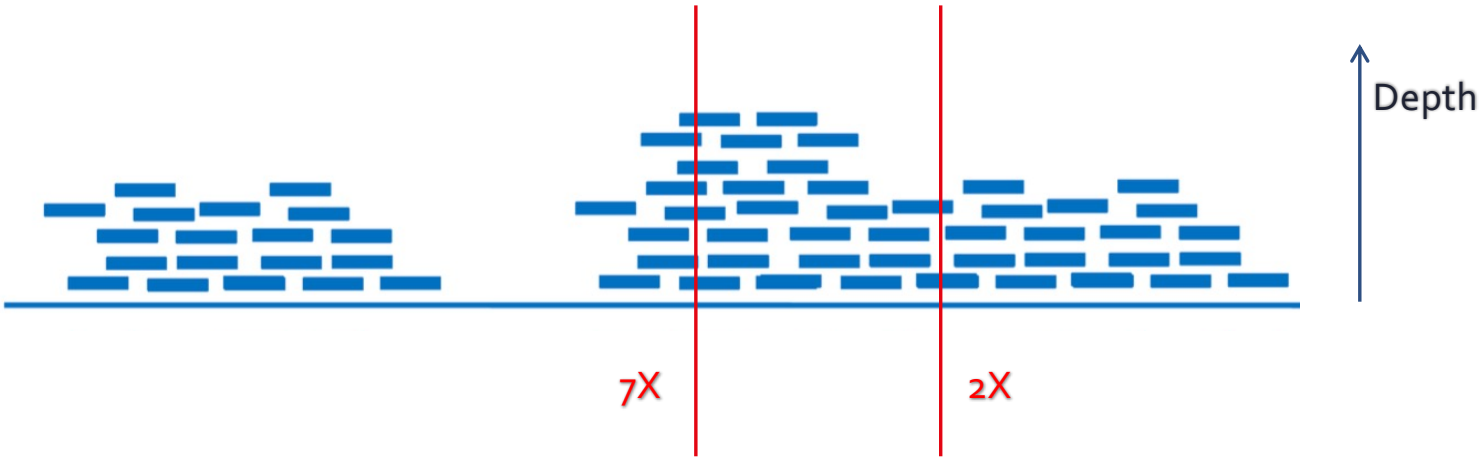
Optional: FORMAT field specifying data type + Per-sample genotype data

BODY

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2

Depth

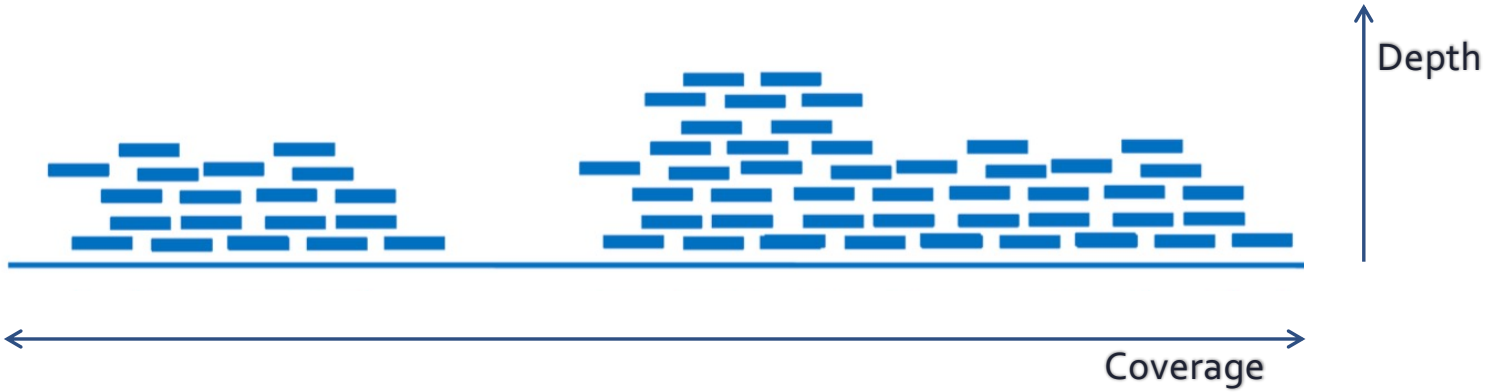
Depth: nb of reads that include a given nucleotide, at a given position



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

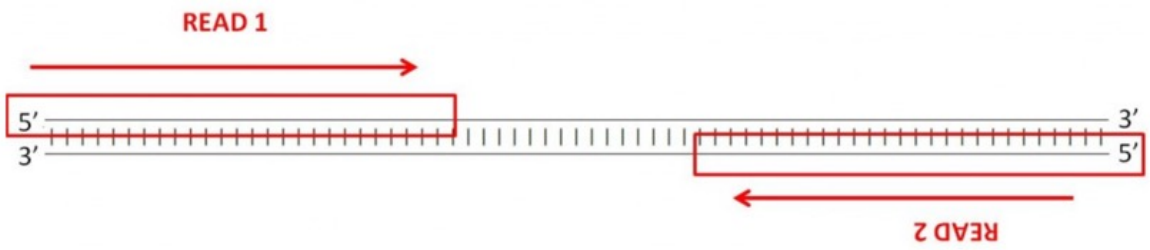
Coverage

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



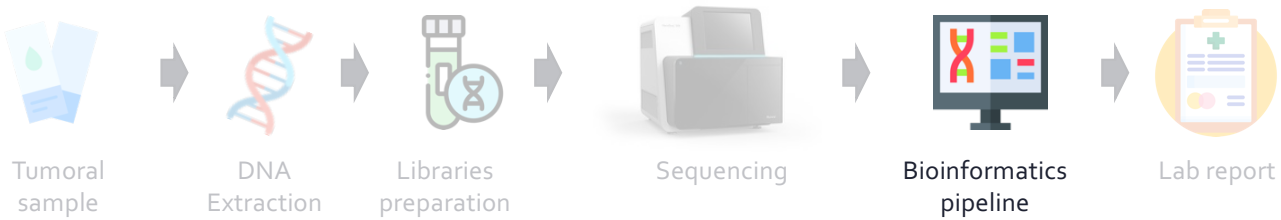
Strand bias in paired-end sequencing

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



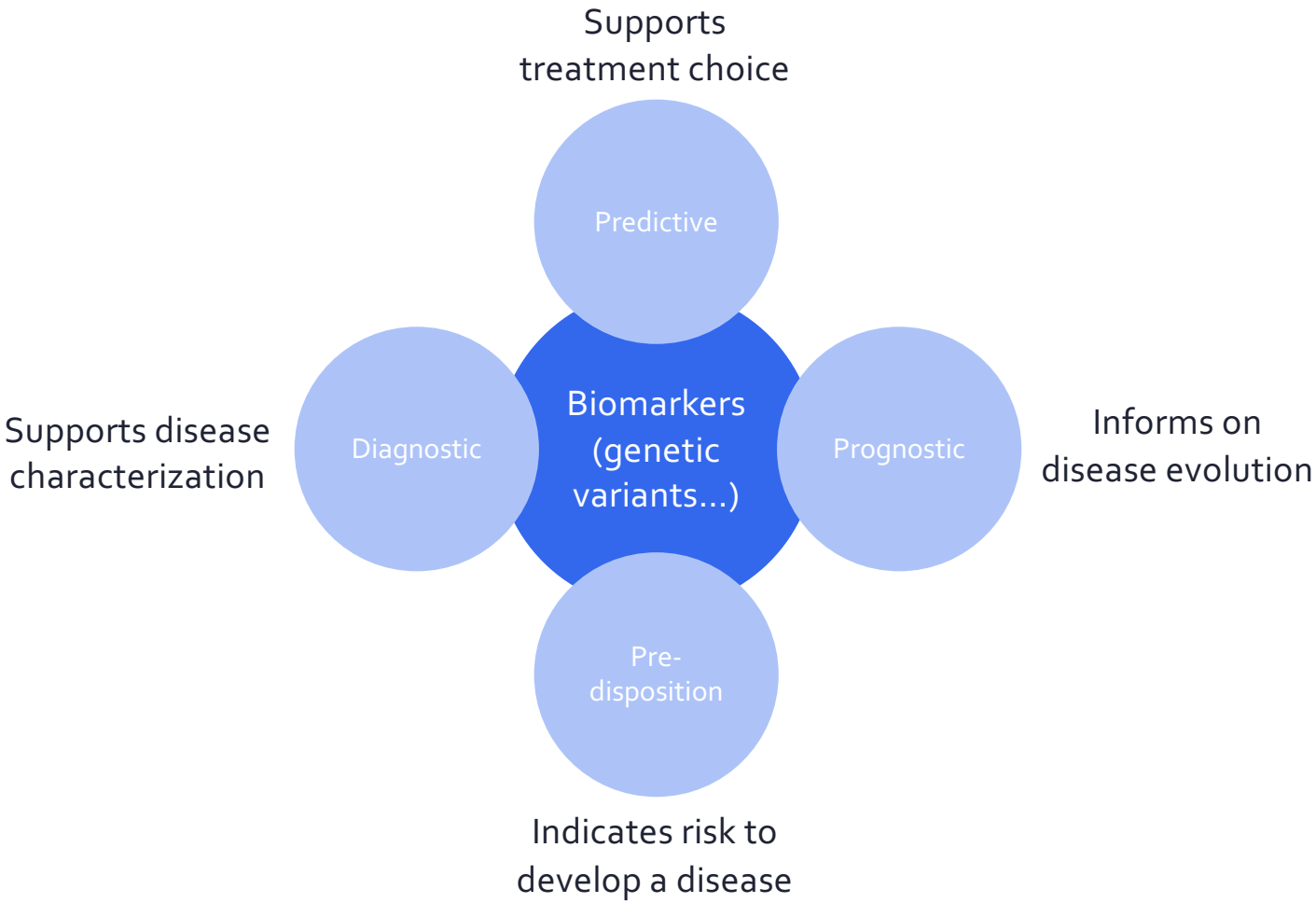
http://www.cureffi.org/2012/12/19/forward-and-reverse-reads-in-paired-end-sequencing/

Overview of a NGS bioinformatics pipeline



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

Personalized medicine



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

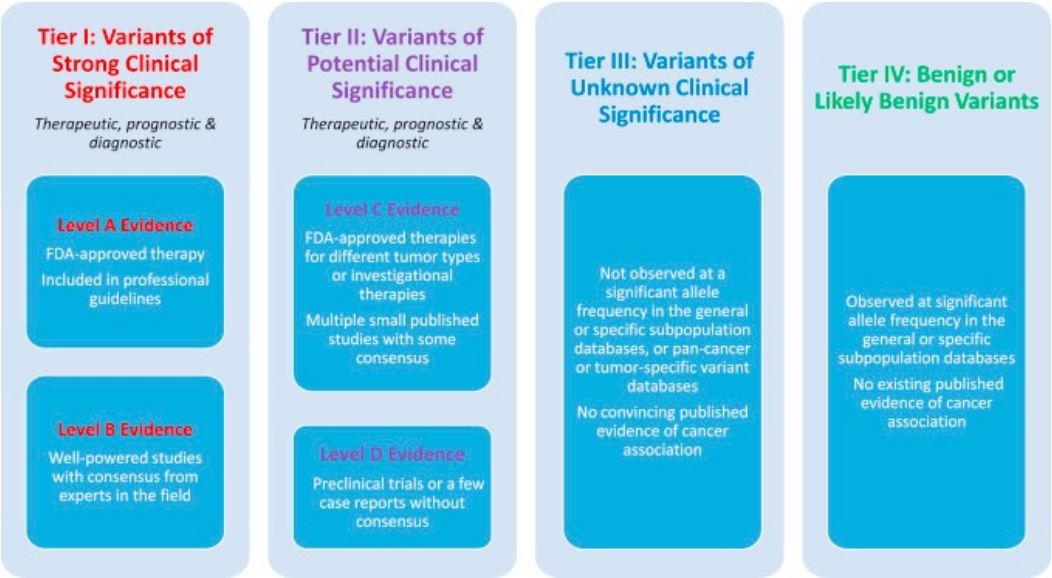
Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer

CrossMark

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

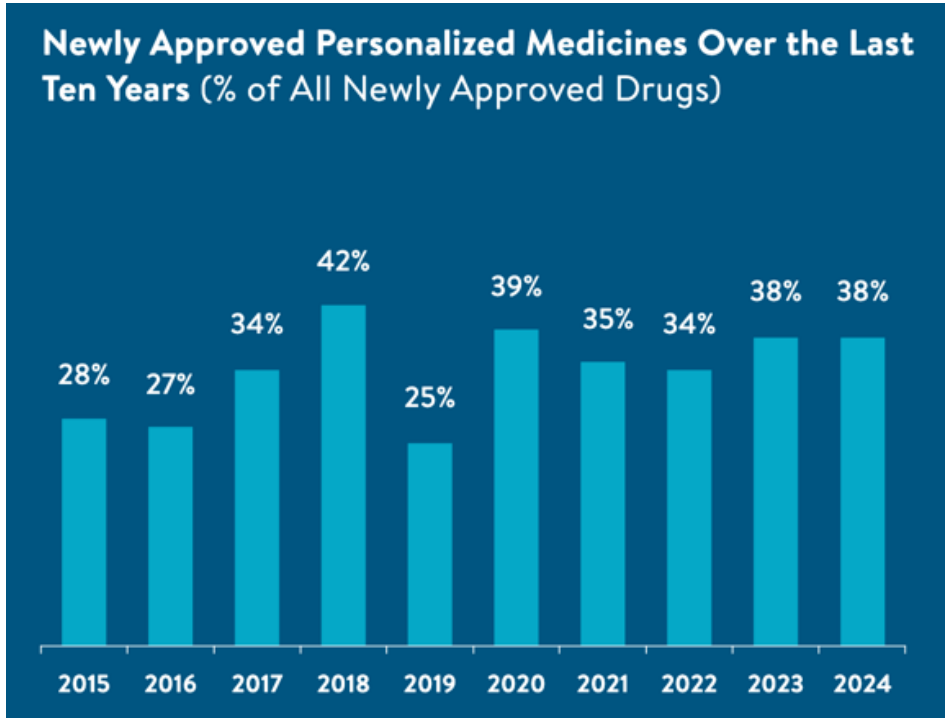
Find actionable variants

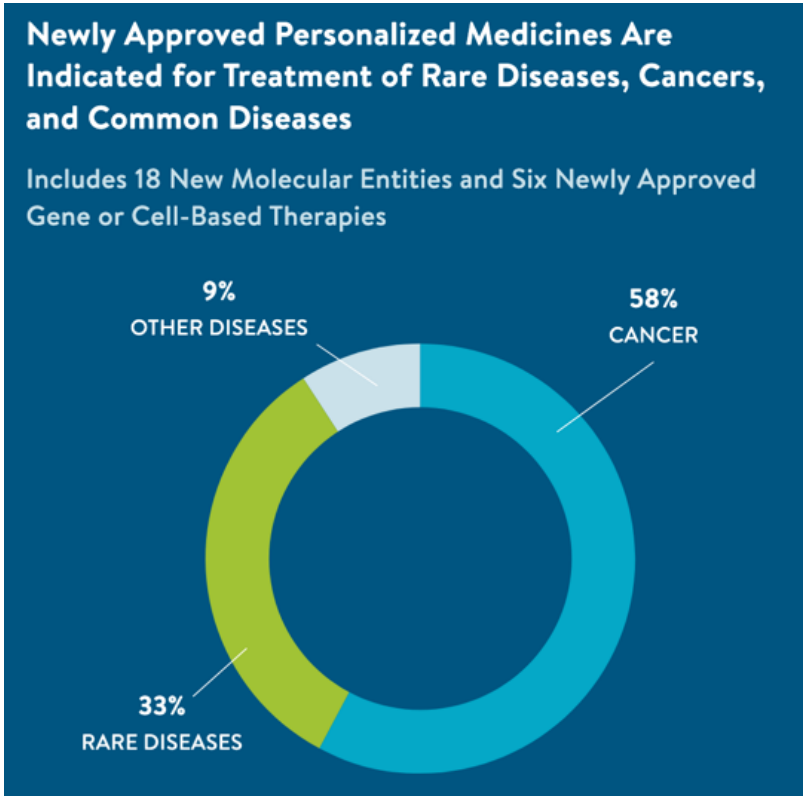
i.e. genetic alterations possibly having an impact on clinical care



Personalized medicine and drug development

Newly FDA-approved drugs





Personalized Medicine Coalition (PMC) : report 2024 on PERSONALIZED MEDICINE AT FDA

How does bioinformatics support variant interpretation?

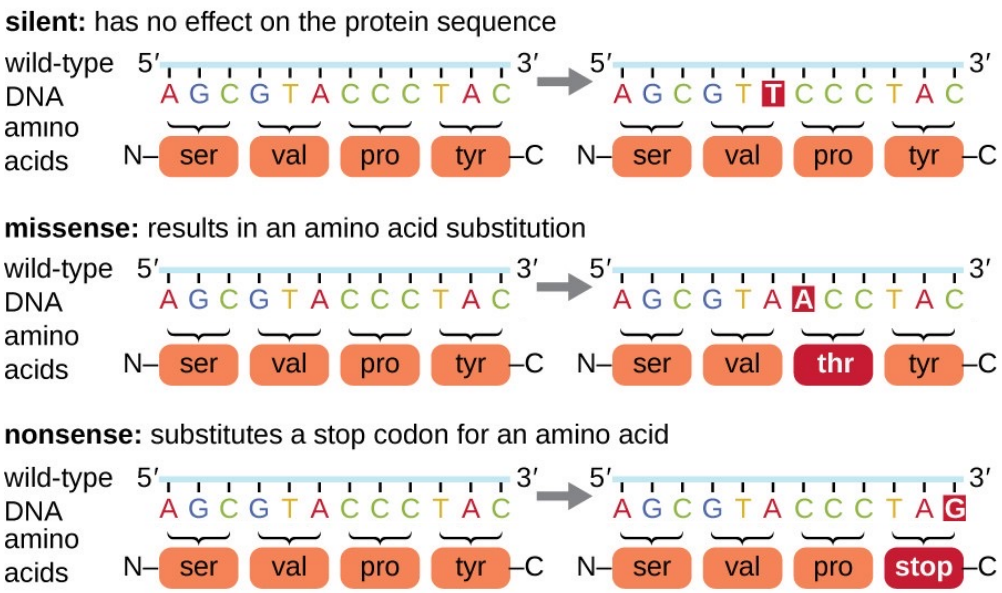
Bioinformatics to guide the interpretation

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

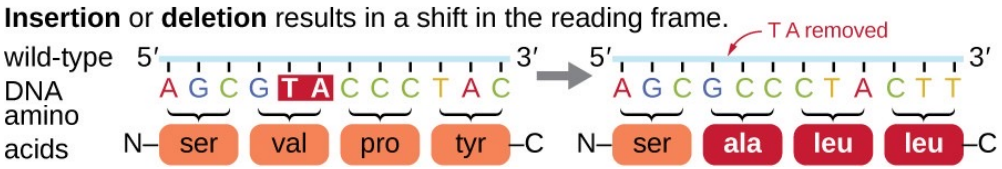
http://www.ensembl.org/info/docs/tools/vep/index.html

Predicting variants effect on the protein sequence

Point mutations (single base substitution)



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



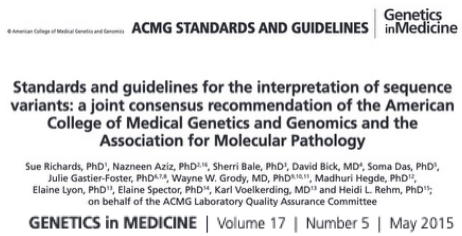
https://courses.lumenlearning.com/microbiology/chapter/mutations/

Bioinformatics to guide the interpretation

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

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


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[15] and Heidi L. Rehm, PhD[16]; on behalf of the ACMG Laboratory Quality Assurance Committee

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

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

(not exhaustive)

Bioinformatics to guide the interpretation

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

Knowledge bases

Cancer

OncoKB
COSMIC
ClinVar
CIVIC
IARC TP53 Database

Genetic disorders

OMIM®
DECIPHER
orphanet

Constitutional variants

dbSNP Short Genetic Variations
ClinVar
HGMD

Population

gnomAD

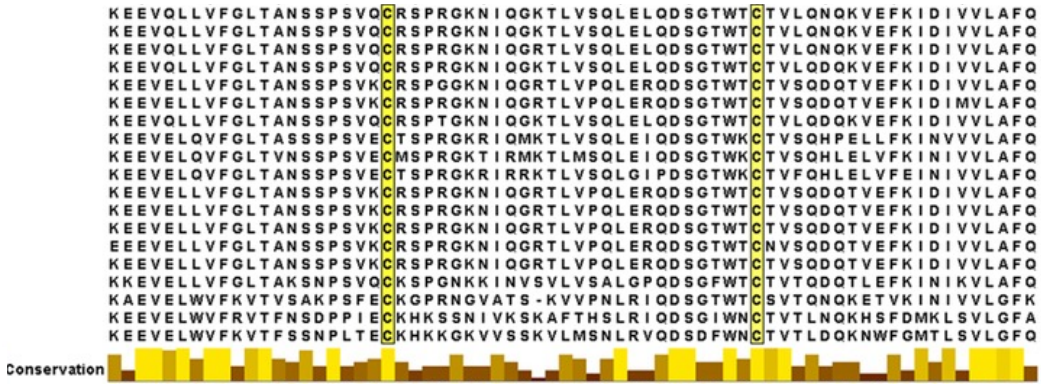
Repositories

UniProt
RefSeq
Ensembl
PubMed

Non exhaustive

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



Identifying 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

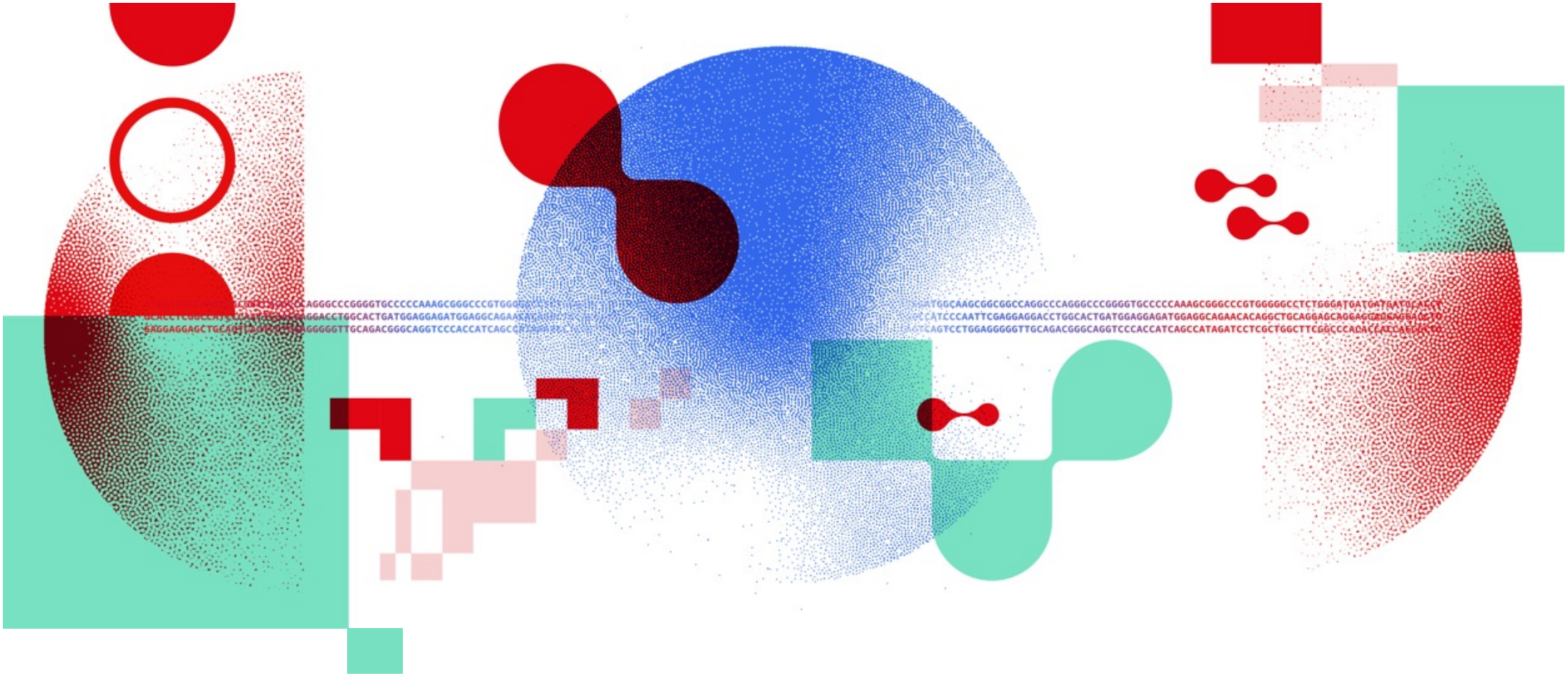
What is
clinical bioinformatics

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

Overview of an oncology NGS diagnostic pipeline

Other considerations in clinical settings

Real-life constraints in the clinics



Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss