

Swiss Institute of
Bioinformatics

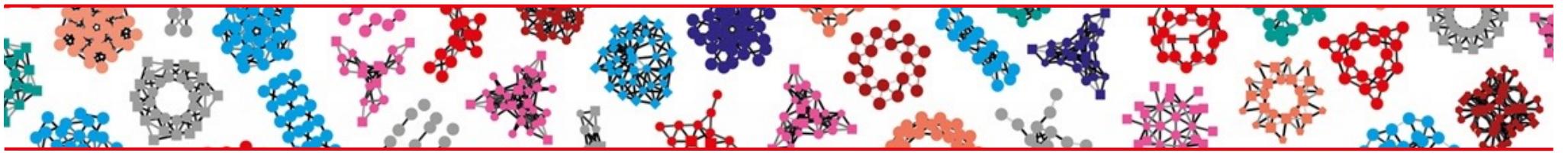
Introduction to bioinformatics: Clinical Bioinformatics

Valérie Barbié, Director SIB Clinical Bioinformatics

Zürich, 07 December 2021

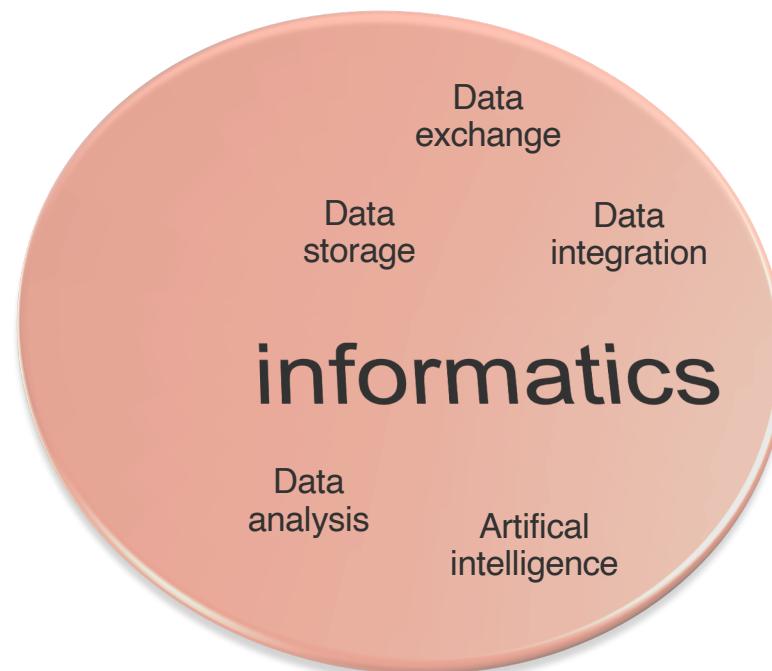


www.sib.swiss

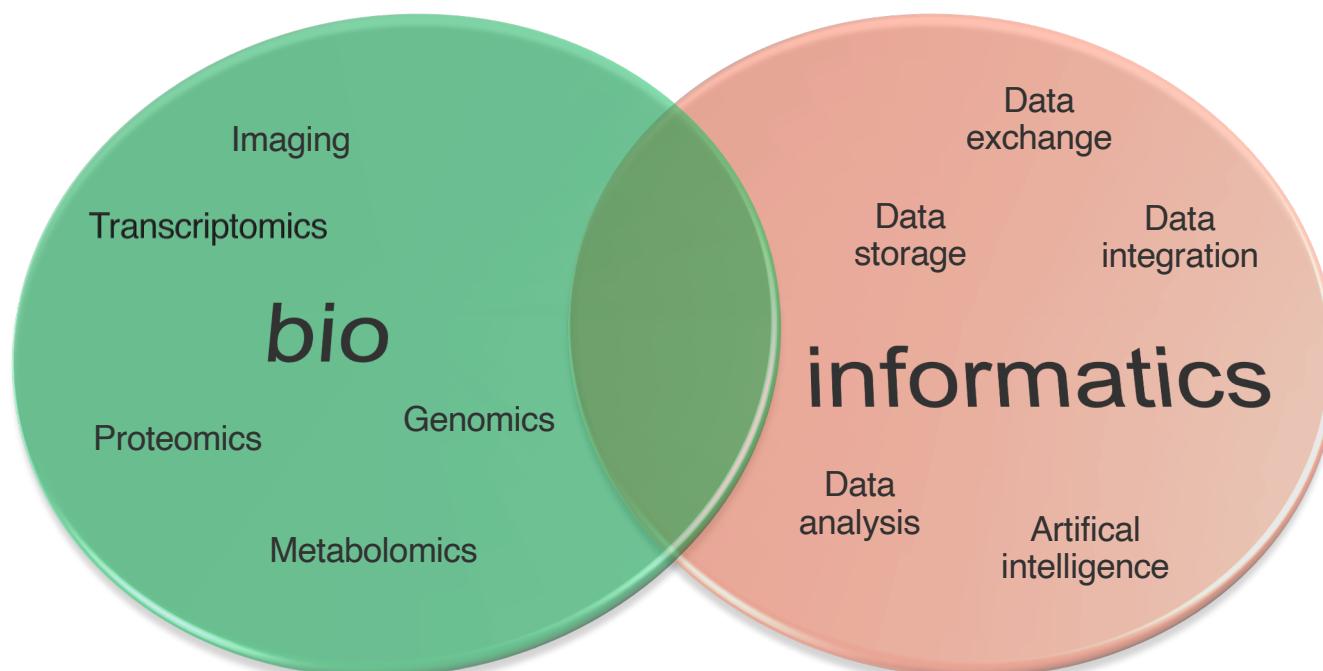


What is clinical bioinformatics?

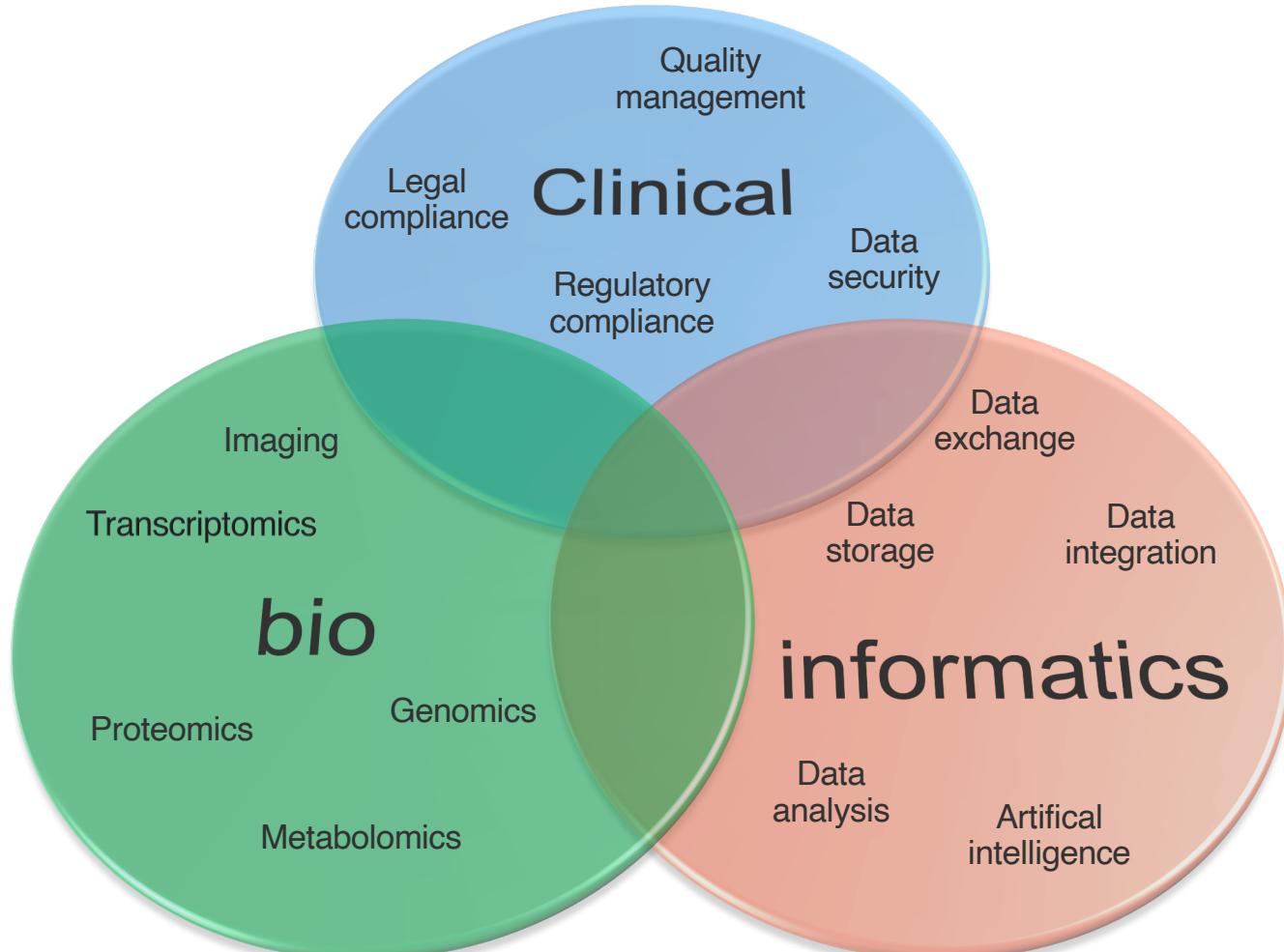
What is clinical bioinformatics?

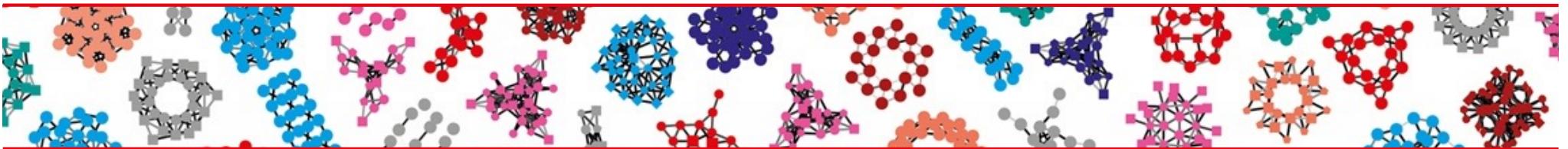


What is clinical bioinformatics?



What is clinical bioinformatics?

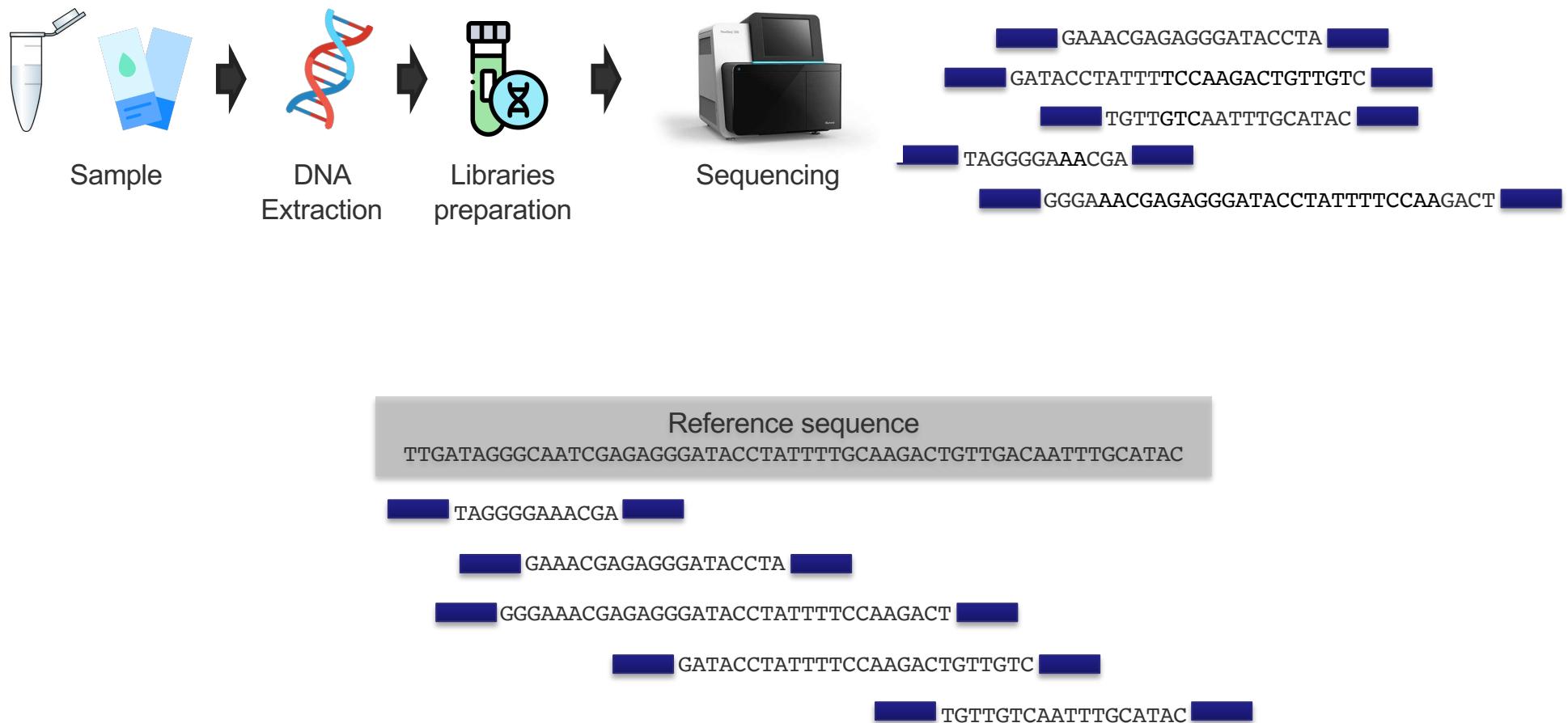




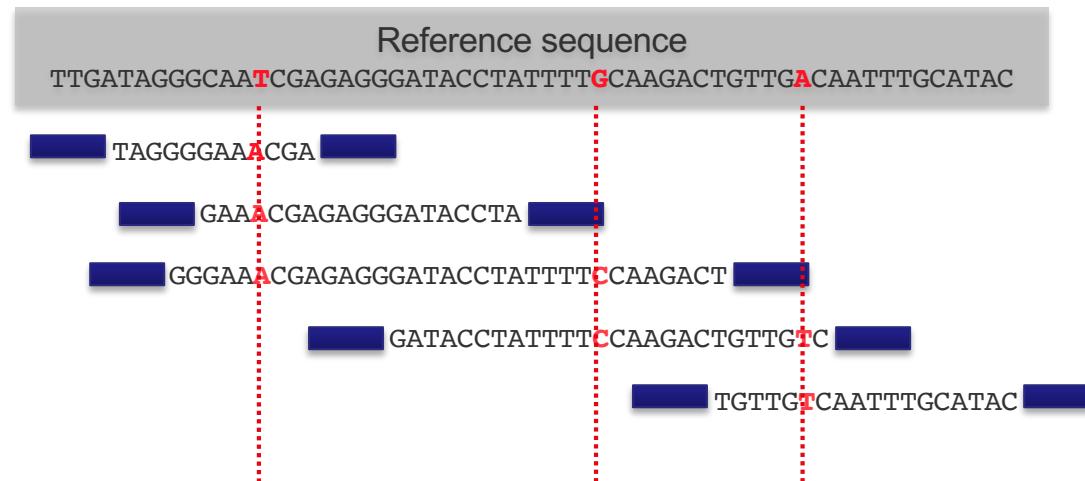
Why clinical bioinformatics?

*The example of
Next Generation Sequencing (NGS)
in medical diagnosis*

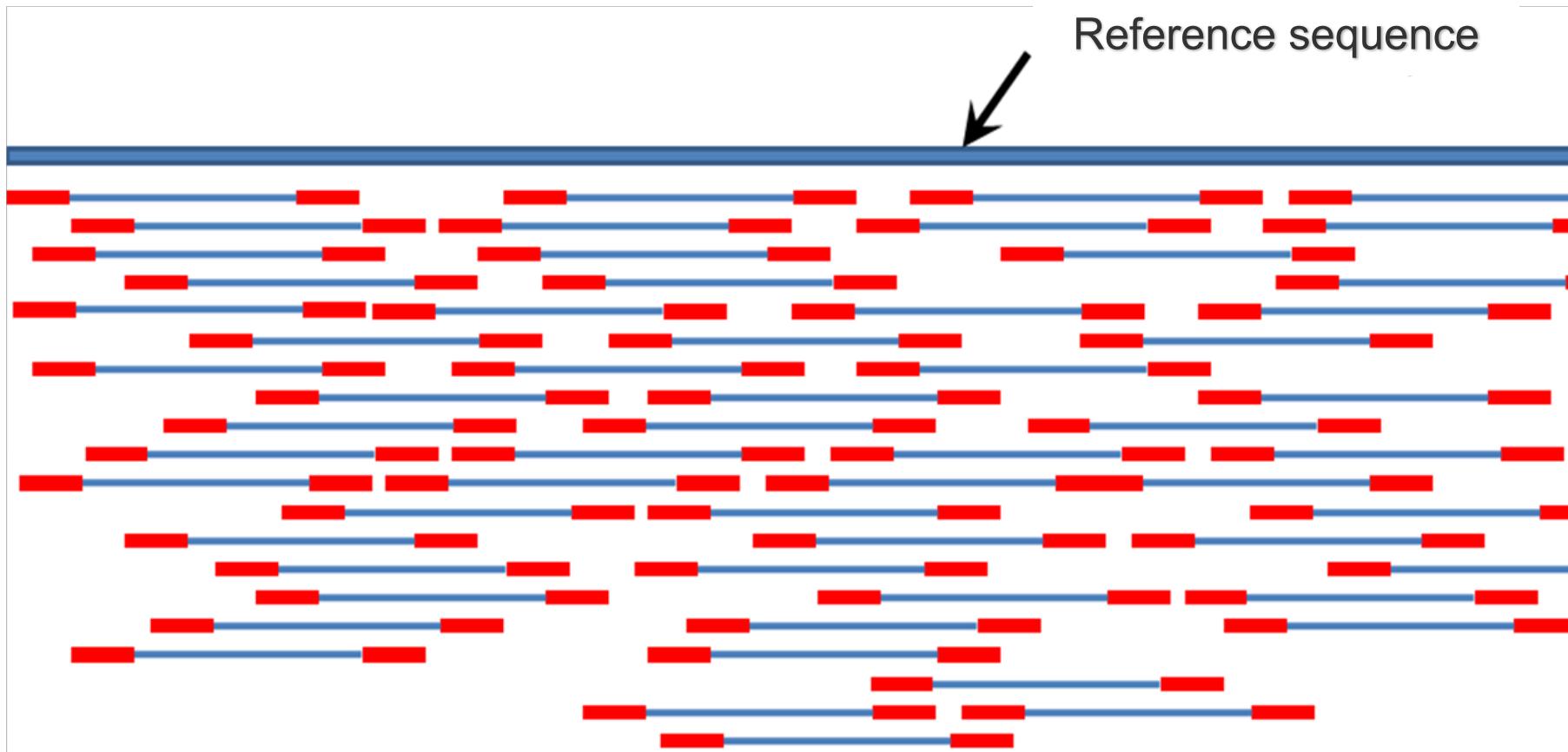
Next Generation Sequencing principle



Next Generation Sequencing principle

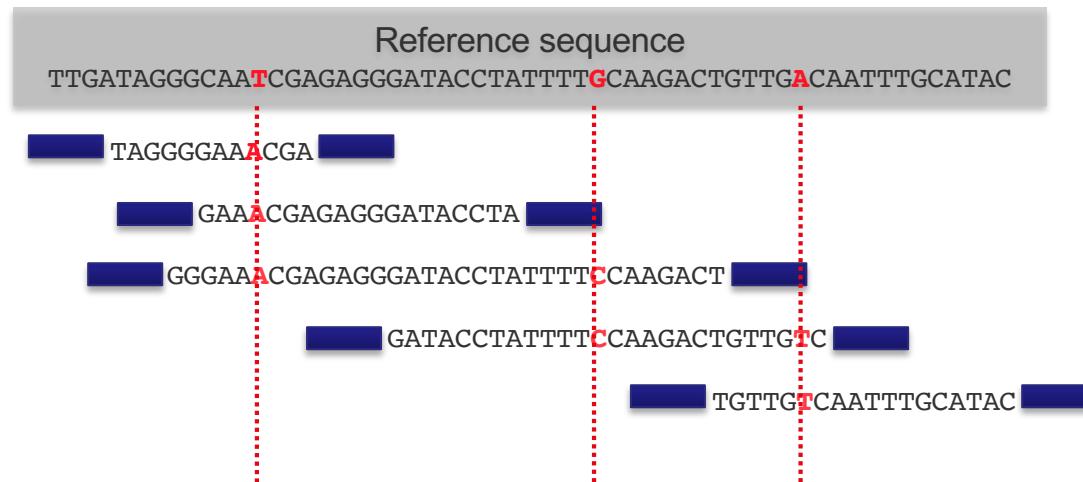


Next Generation Sequencing principle

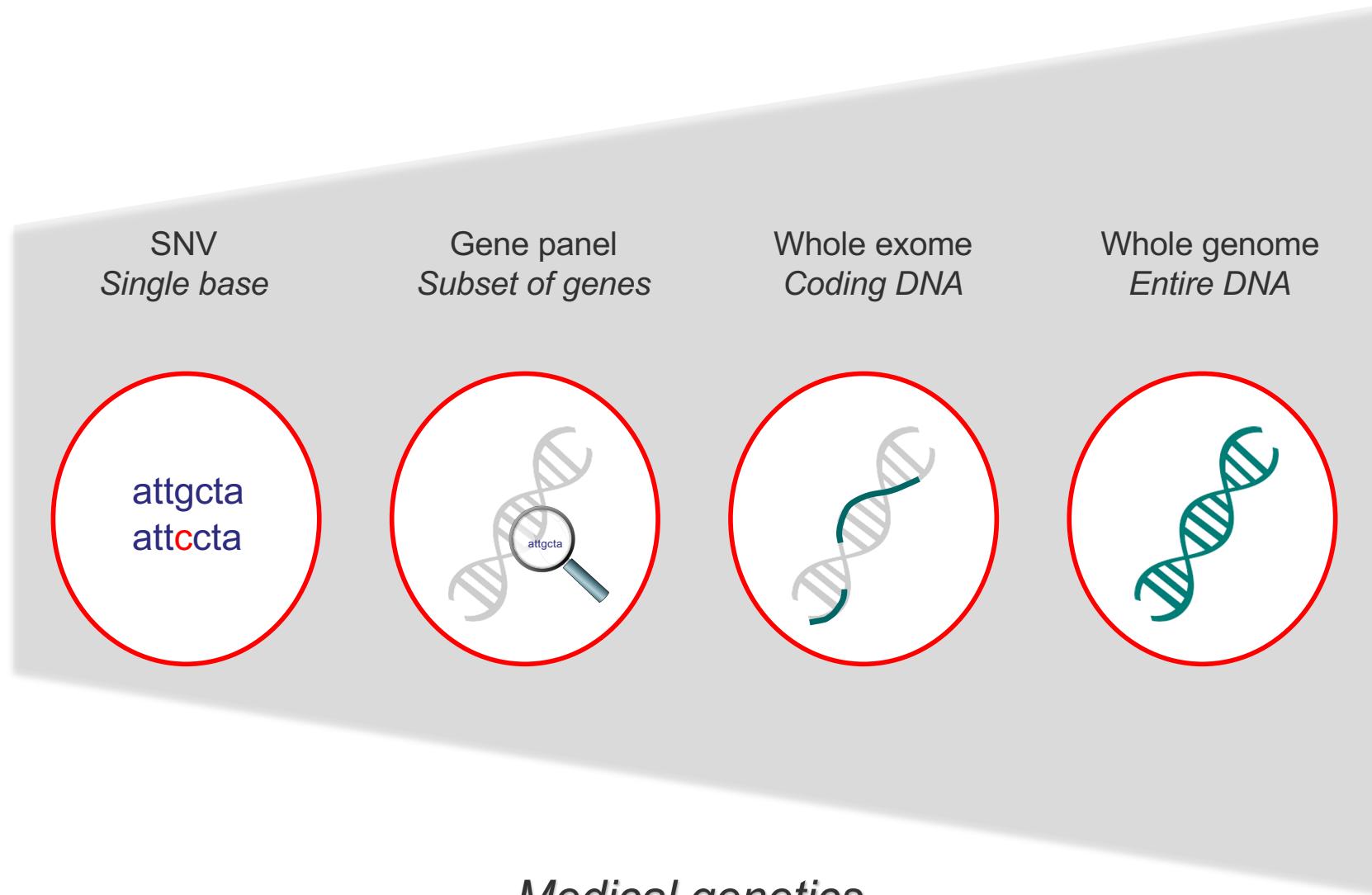


Examples of NGS clinical applications

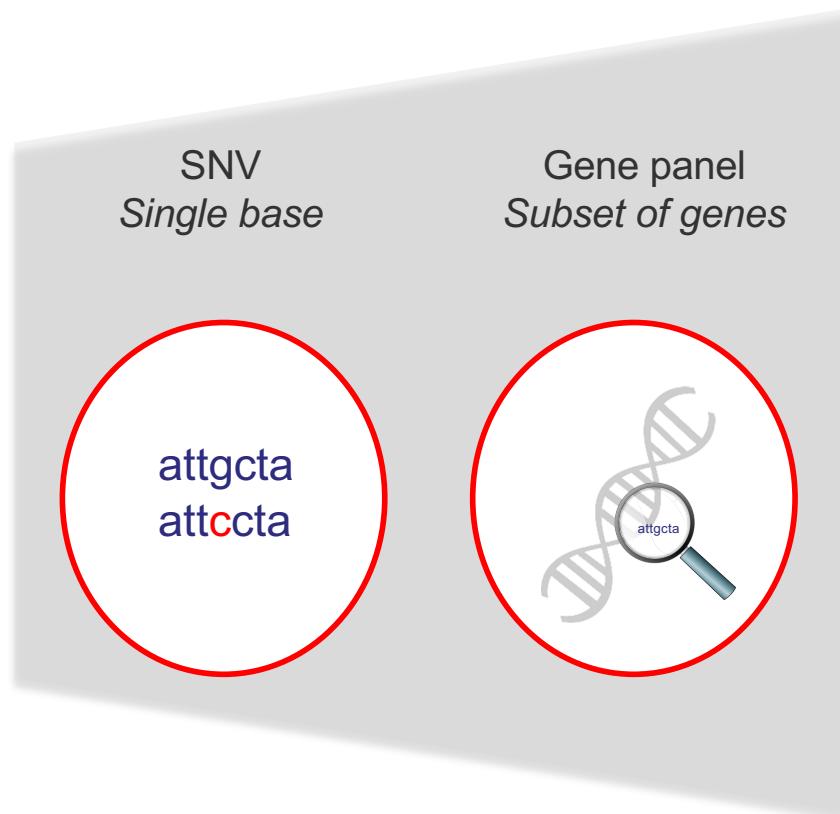
	Source DNA	Reference DNA
Oncology	Patient tumor or blood	Consensus human genome Germline



Scale matters

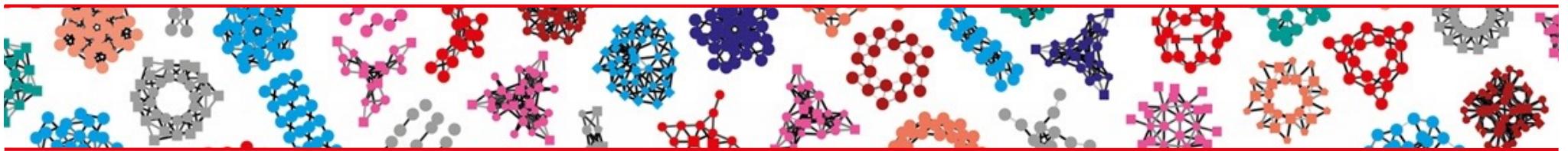


Scale matters



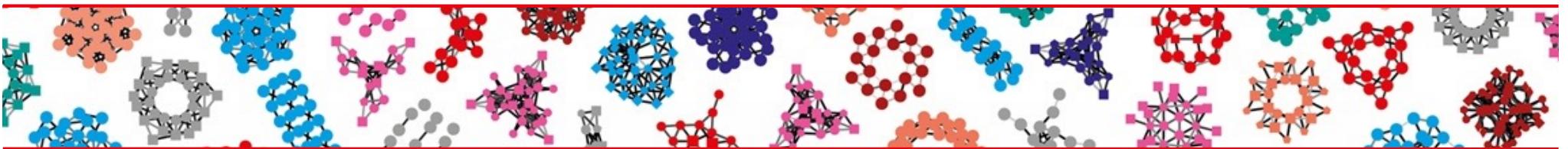
Oncology





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

attcggtcatgccatagggg

NGS in cancer diagnosis?

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

at~~t~~cgggtcatgccatagggg

Single Nucleotide Variant (SNV)

at~~g~~cgggtcatgccatagggg

Insertion

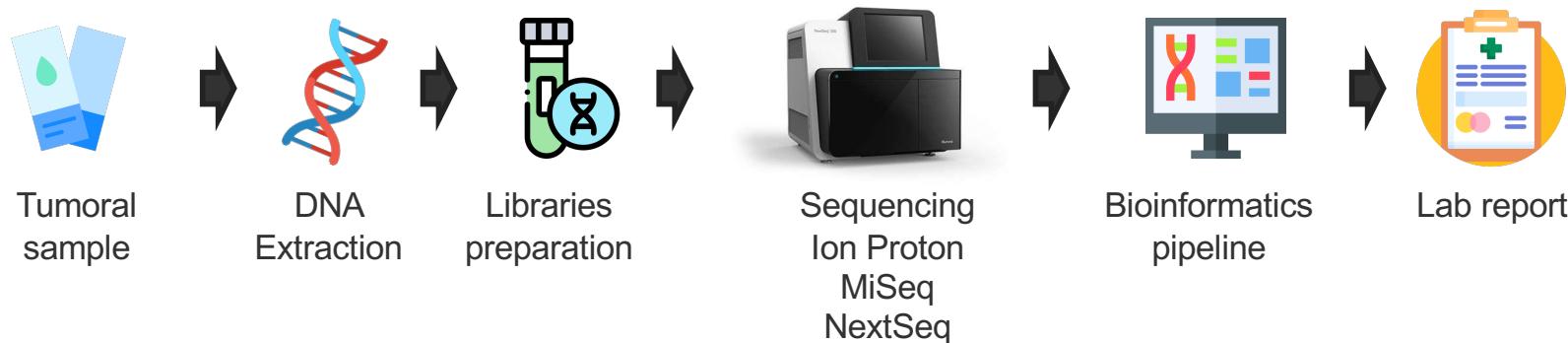
at~~g~~cgggtcatcgtgtccgccatagggg

Deletion

at~~g~~cgggtcatcgtgtccg....tagggg

cccc

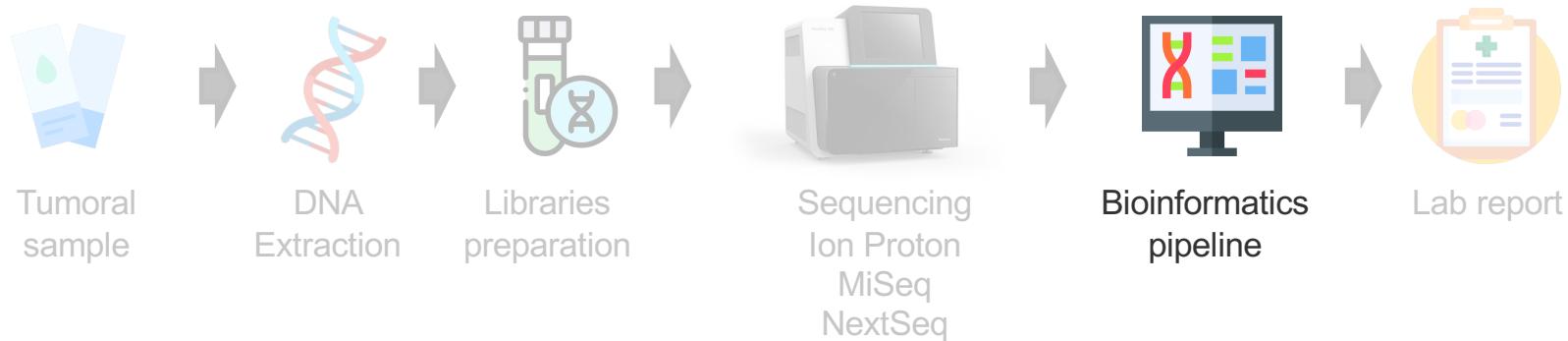
Overview of a NGS bioinformatics pipeline

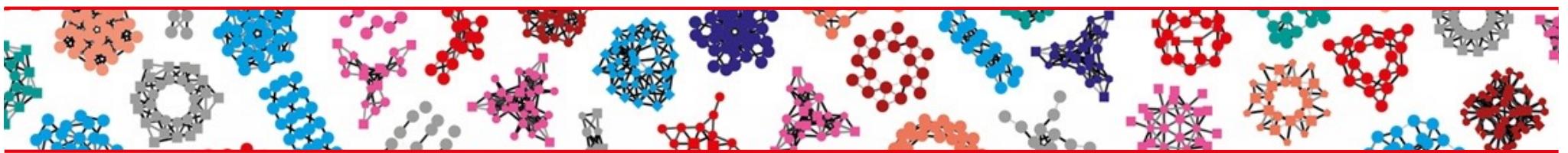


■ Gene panels analysis in clinical routine

- Identify **artifacts**: quality control
- Identify **somatic** vs. germline variants
- Variant **annotation**: does it provide clinically-useful information?

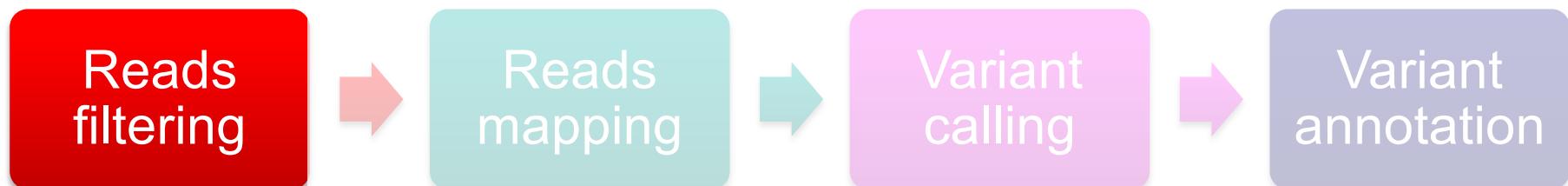
Overview of a NGS bioinformatics pipeline





PART II

Quality control



Out of the sequencer: FASTQ file

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

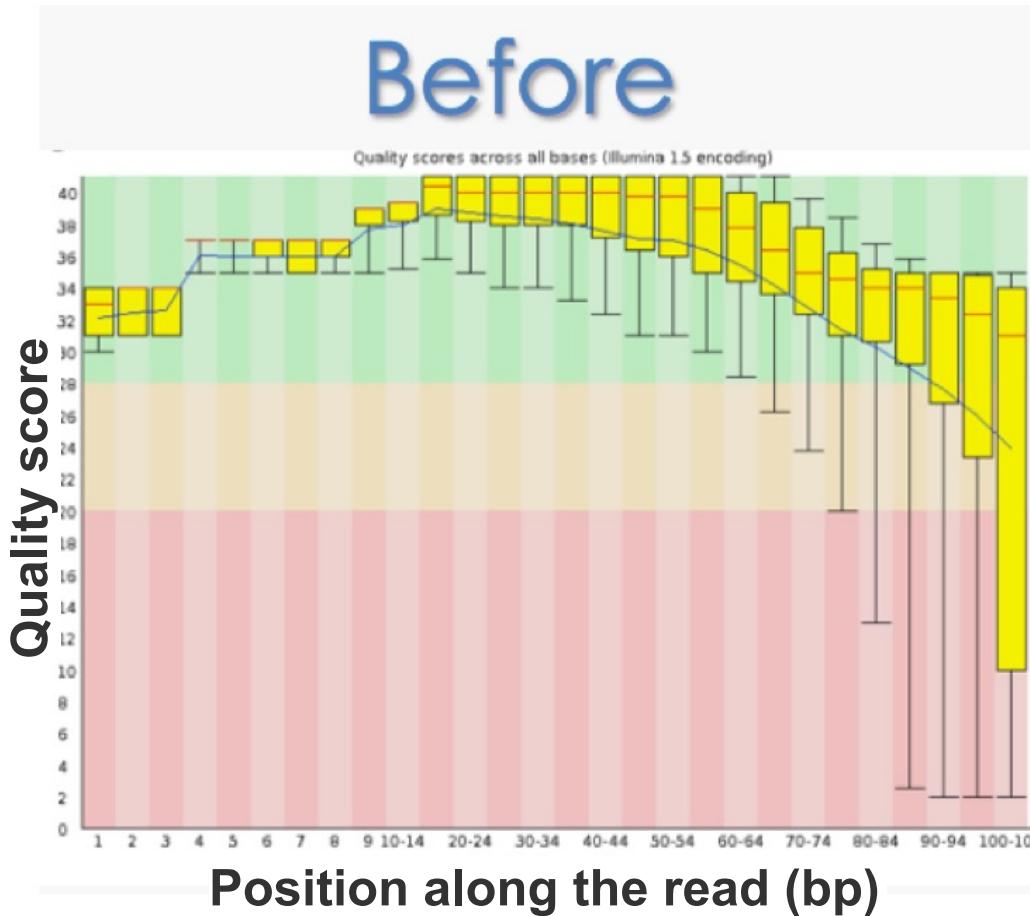


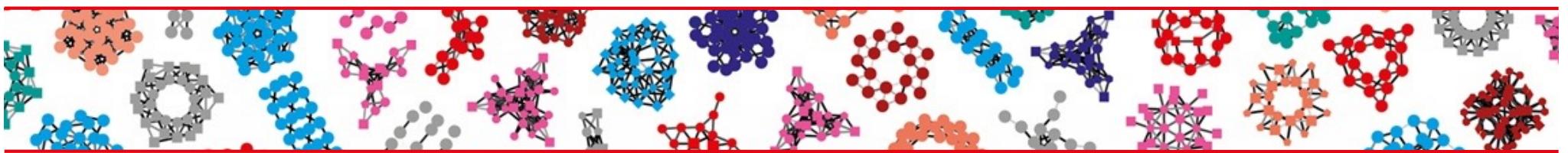
Each nucleotide has a **quality score (Phred score)** representing the probability that a base was miscalled by the sequencer

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

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

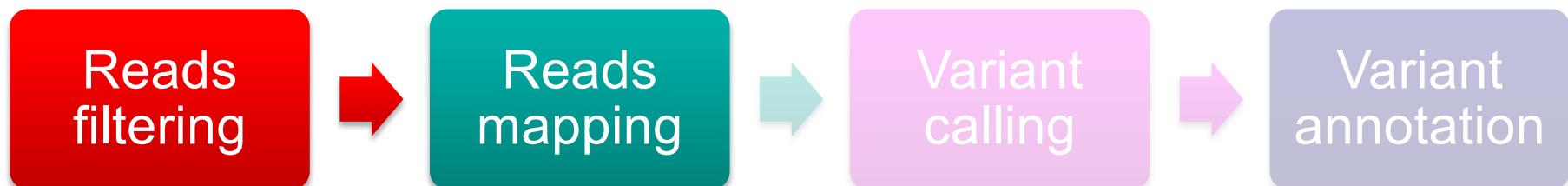
Quality-based reads trimming



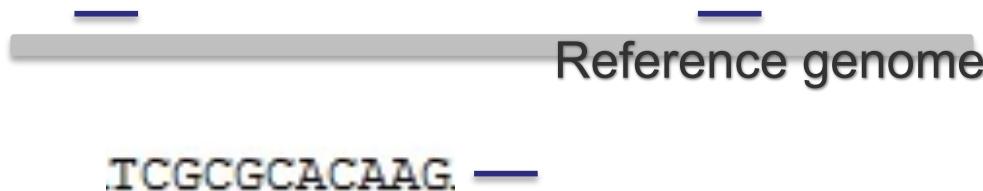


PART II

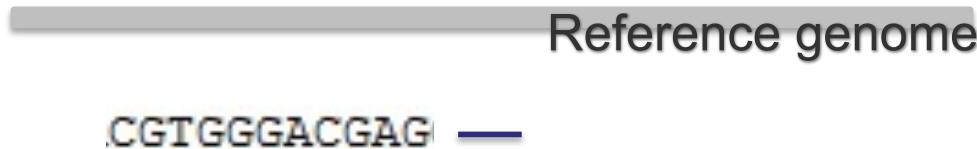
Quality control



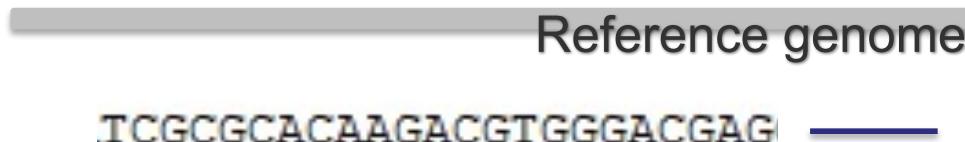
Let's align the reads



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

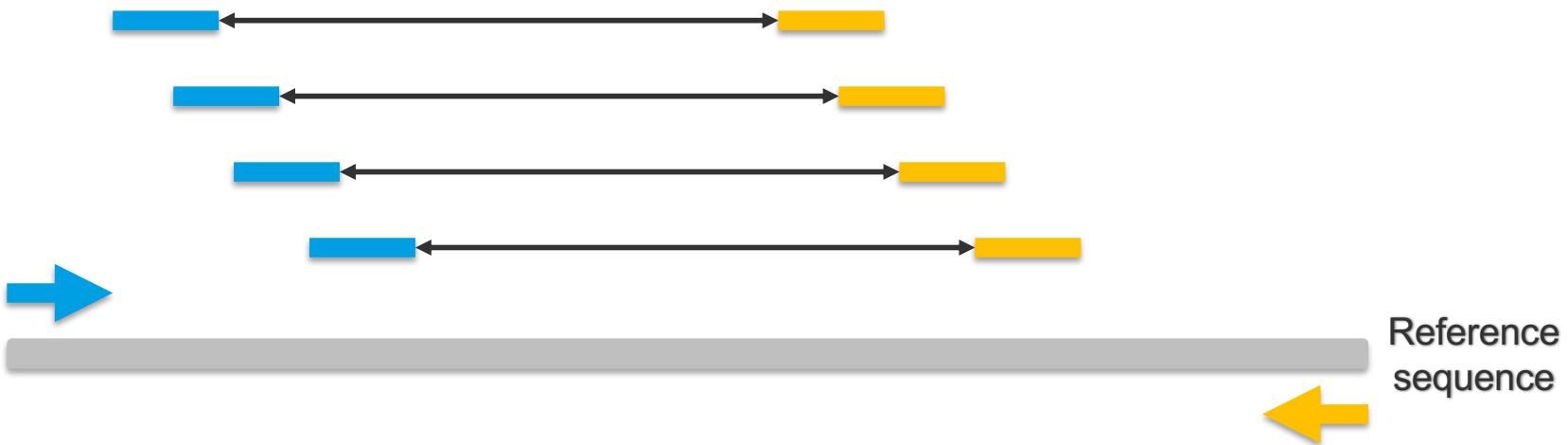


! Mismatches and gaps allowed
→ algorithms have scoring functions



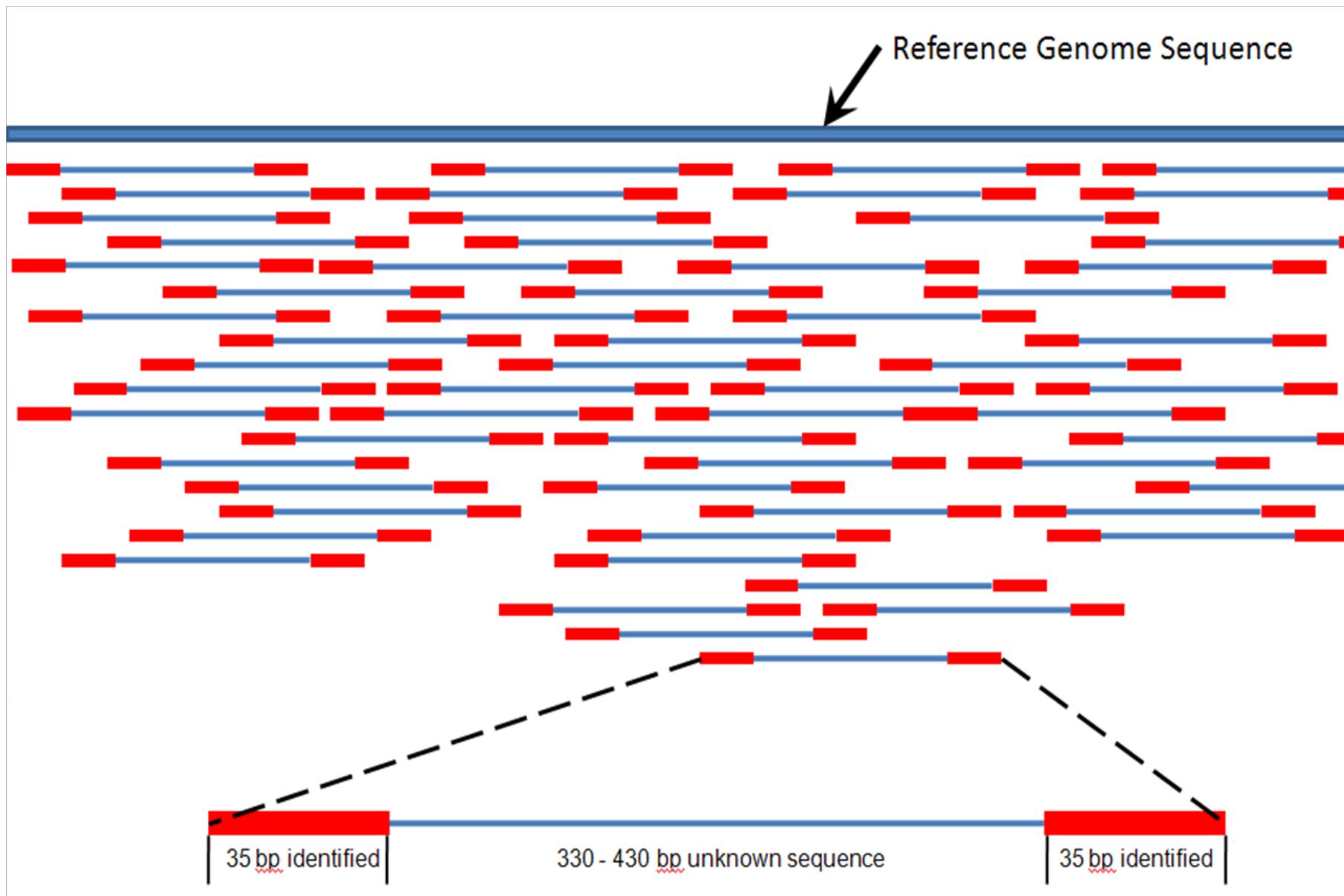
! Longer reads are less ambiguous
→ but computationally more expensive

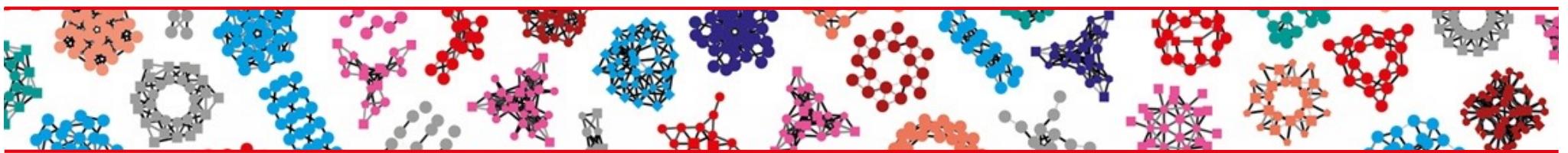
Paired-end sequencing



- Much better alignment on across regions difficult to sequence (e.g. repetitive regions)

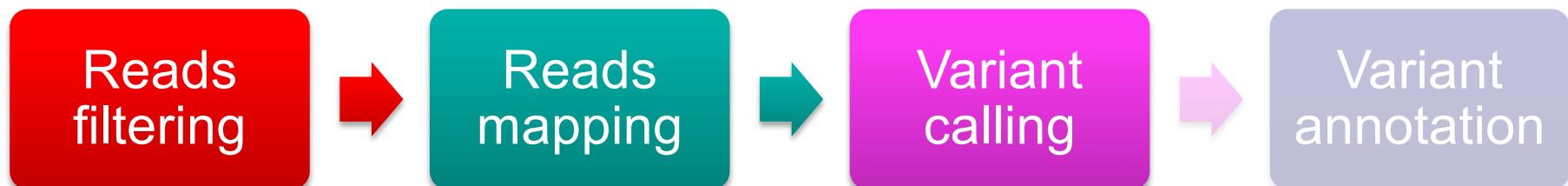
Mapping: finding the best position for each read



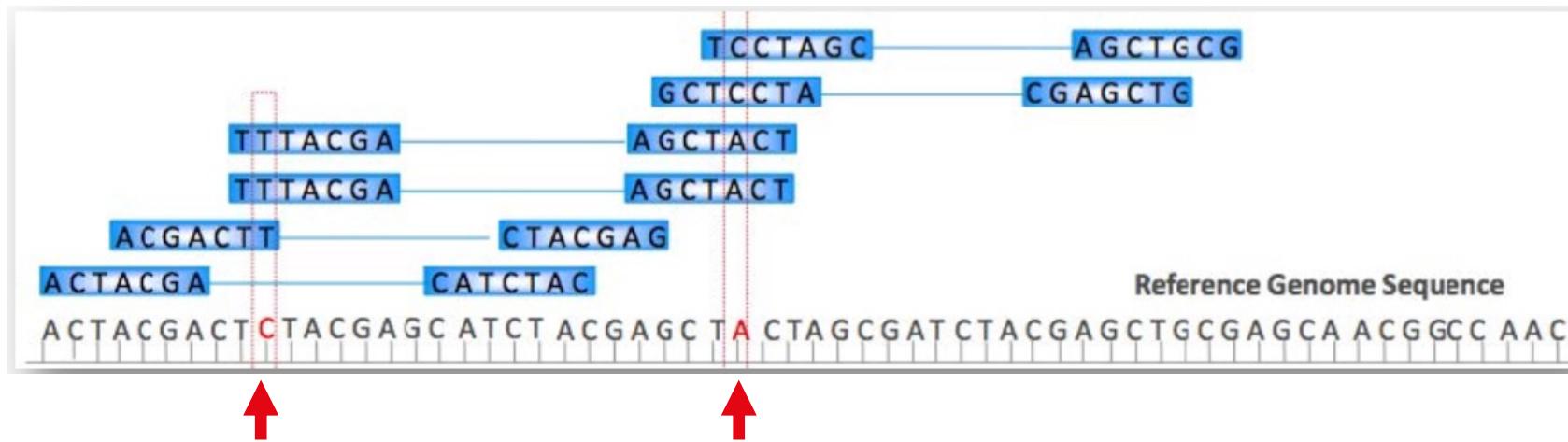


PART II

Quality control



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

HEADER	<pre>##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</pre>	
	<pre>##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"></pre>	INFO meta-information
	<pre>##FILTER=<ID=q10,Description="Quality below 10"> ##FILTER=<ID=s50,Description="Less than 50% of samples have data"></pre>	FILTER meta-information
	<pre>##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"></pre>	FORMAT meta-information
BODY	<pre>##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></pre>	Optional: FORMAT field specifying data type + Per-sample genotype data
	Fixed fields	
	<pre>#CHROM POS ID REF ALT QUAL FILTER INFO</pre>	<pre>FORMAT NORMAL TUMOR</pre>
	<pre>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</pre>	<pre>GT:GQ:DP:HQ 0 0:48:1:51,51 1 0:48:8:51,51 GT:GQ:DP:HQ 0 0:49:3:58,50 0 1:3:5:65,3 GT:GQ:DP:HQ 1 2:21:6:23,27 2 1:2:0:18,2 GT:GQ:DP:HQ 0 0:54:7:56,60 0 0:48:4:51,51 GT:GQ:DP 0/1:35:4 0/2:17:2</pre>

Output of the variant caller: VCF

VCF: Variant Call Format

HEADER	<pre>##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"></pre>	INFO meta-information						
	<pre>##FILTER=<ID=q10,Description="Quality below 10"> ##FILTER=<ID=s50,Description="Less than 50% of samples have data"></pre>	FILTER meta-information						
BODY	<pre>##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"></pre>	FORMAT meta-information						
	<pre>##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></pre>	Optional: FORMAT field specifying data type + Per-sample genotype data						
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	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

Output of the variant caller: VCF

VCF: Variant Call Format

HEADER	<pre>##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"></pre>	INFO meta-information																																																																	
	<pre>##FILTER=<ID=q10,Description="Quality below 10"> ##FILTER=<ID=s50,Description="Less than 50% of samples have data"></pre>	FILTER meta-information																																																																	
BODY	<pre>##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"></pre>	FORMAT meta-information																																																																	
	<pre>##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></pre>	Optional: FORMAT field specifying data type + Per-sample genotype data																																																																	
BODY	Fixed fields	FORMAT NORMAL TUMOR																																																																	
	<table border="1"><thead><tr><th>#CHROM</th><th>POS</th><th>ID</th><th>REF</th><th>ALT</th><th>QUAL</th><th>FILTER</th><th>INFO</th><th>FORMAT</th><th>NORMAL</th><th>TUMOR</th></tr></thead><tbody><tr><td>20</td><td>14370</td><td>rs6054257</td><td>G</td><td>A</td><td>29</td><td>PASS</td><td>NS=3;DP=14;AF=0.5;DB:H2</td><td>GT:GQ:DP:HQ</td><td>0 0:48:1:51,51</td><td>1 0:48:8:51,51</td></tr><tr><td>20</td><td>17330</td><td>.</td><td>T</td><td>A</td><td>3</td><td>q10</td><td>NS=3;DP=11;AF=0.017</td><td>GT:GQ:DP:HQ</td><td>0 0:49:3:58,50</td><td>0 1:3:5:65,3</td></tr><tr><td>20</td><td>1110696</td><td>rs6040355</td><td>A</td><td>G,T</td><td>67</td><td>PASS</td><td>NS=2;DP=10;AF=0.333;0.667;DB</td><td>GT:GQ:DP:HQ</td><td>1 2:21:6:23,27</td><td>2 1:2:0:18,2</td></tr><tr><td>20</td><td>1230237</td><td>.</td><td>T</td><td>.</td><td>47</td><td>PASS</td><td>NS=3;DP=13;AA=T</td><td>GT:GQ:DP:HQ</td><td>0 0:54:7:56,60</td><td>0 0:48:4:51,51</td></tr><tr><td>20</td><td>1234567</td><td>microsat1</td><td>GTC</td><td>G,GTCTC</td><td>50</td><td>PASS</td><td>NS=3;DP=9;AA=G</td><td>GT:GQ:DP</td><td>0/1:35:4</td><td>0/2:17:2</td></tr></tbody></table>	#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
#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																																																									

Output of the variant caller: VCF

VCF: Variant Call Format

HEADER	<pre>##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"></pre>	INFO meta-information																																																
	<pre>##FILTER=<ID=q10,Description="Quality below 10"> ##FILTER=<ID=s50,Description="Less than 50% of samples have data"></pre>	FILTER meta-information																																																
BODY	<pre>##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"></pre>	FORMAT meta-information																																																
	<pre>##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></pre>	Optional: FORMAT field specifying data type + Per-sample genotype data																																																
Fixed fields																																																		
<table border="1"><thead><tr><th>#CHROM</th><th>POS</th><th>ID</th><th>REF</th><th>ALT</th><th>QUAL</th><th>FILTER</th><th>INFO</th></tr></thead><tbody><tr><td>20</td><td>14370</td><td>rs6054257</td><td>G</td><td>A</td><td>29</td><td>PASS</td><td>NS=3;DP=14;AF=0.5;DB;H2</td></tr><tr><td>20</td><td>17330</td><td>.</td><td>T</td><td>A</td><td>3</td><td>q10</td><td>NS=3;DP=11;AF=0.017</td></tr><tr><td>20</td><td>1110696</td><td>rs6040355</td><td>A</td><td>G,T</td><td>67</td><td>PASS</td><td>NS=2;DP=10;AF=0.333,0.667;DB</td></tr><tr><td>20</td><td>1230237</td><td>.</td><td>T</td><td>.</td><td>47</td><td>PASS</td><td>NS=3;DP=13;AA=T</td></tr><tr><td>20</td><td>1234567</td><td>microsat1</td><td>G,GTCTC</td><td>50</td><td>PASS</td><td>NS=3;DP=9;AA=G</td><td></td></tr></tbody></table>		#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	G,GTCTC	50	PASS	NS=3;DP=9;AA=G		
#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	G,GTCTC	50	PASS	NS=3;DP=9;AA=G																																												
		<table border="1"><thead><tr><th>FORMAT</th><th>NORMAL</th><th>TUMOR</th></tr></thead><tbody><tr><td>GT:GQ:DP:HQ</td><td>0 0:48:1:51,51</td><td>1 0:48:8:51,51</td></tr><tr><td>GT:GQ:DP:HQ</td><td>0 0:49:3:58,50</td><td>0 1:3:5:65,3</td></tr><tr><td>GT:GQ:DP:HQ</td><td>1 2:21:6:23,27</td><td>2 1:2:0:18,2</td></tr><tr><td>GT:GQ:DP:HQ</td><td>0 0:54:7:56,60</td><td>0 0:48:4:51,51</td></tr><tr><td>GT:GQ:DP</td><td>0/1:35:4</td><td>0/2:17:2</td></tr></tbody></table>	FORMAT	NORMAL	TUMOR	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	GT:GQ:DP	0/1:35:4	0/2:17:2																														
FORMAT	NORMAL	TUMOR																																																
GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51																																																
GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3																																																
GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2																																																
GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51																																																
GT:GQ:DP	0/1:35:4	0/2:17:2																																																

Output of the variant caller: VCF

VCF: Variant Call Format

HEADER	<pre>##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=f126cdf8a6e0c7f1379d618ff66beb2da,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"></pre>	INFO meta-information																																															
	<pre>##FILTER=<ID=q10,Description="Quality below 10"> ##FILTER=<ID=s50,Description="Less than 50% of samples have data"></pre>	FILTER meta-information																																															
BODY	<pre>##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"></pre>	FORMAT meta-information																																															
	<pre>##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></pre>	Optional: FORMAT field specifying data type + Per-sample genotype data																																															
BODY	Fixed fields	<table border="1"><thead><tr><th>FORMAT</th><th>NORMAL</th><th>TUMOR</th></tr></thead><tbody><tr><td>GT:GQ:DP:HQ</td><td>0 0:48:1:51,51</td><td>1 0:48:8:51,51</td></tr><tr><td>GT:GQ:DP:HQ</td><td>0 0:49:3:58,50</td><td>0 1:3:5:65,3</td></tr><tr><td>GT:GQ:DP:HQ</td><td>1 2:21:6:23,27</td><td>2 1:2:0:18,2</td></tr><tr><td>GT:GQ:DP:HQ</td><td>0 0:54:7:56,60</td><td>0 0:48:4:51,51</td></tr><tr><td>GT:GQ:DP</td><td>0/1:35:4</td><td>0/2:17:2</td></tr></tbody></table>	FORMAT	NORMAL	TUMOR	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	GT:GQ:DP	0/1:35:4	0/2:17:2																													
FORMAT	NORMAL	TUMOR																																															
GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51																																															
GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3																																															
GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2																																															
GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51																																															
GT:GQ:DP	0/1:35:4	0/2:17:2																																															
<table border="1"><thead><tr><th>#CHROM</th><th>POS</th><th>ID</th><th>REF</th><th>ALT</th><th>QUAL</th><th>FILTER</th><th>INFO</th></tr></thead><tbody><tr><td>20</td><td>14370</td><td>rs6054257</td><td>G</td><td>A</td><td>29</td><td>PASS</td><td>NS=3;DP=14;AF=0.5;DB;H2</td></tr><tr><td>20</td><td>17330</td><td>.</td><td>T</td><td>A</td><td>3</td><td>q10</td><td>NS=3;DP=11;AF=0.017</td></tr><tr><td>20</td><td>1110696</td><td>rs6040355</td><td>A</td><td>G,T</td><td>67</td><td>PASS</td><td>NS=2;DP=10;AF=0.333,0.667;DB</td></tr><tr><td>20</td><td>1230237</td><td>.</td><td>T</td><td>.</td><td>47</td><td>PASS</td><td>NS=3;DP=13;AA=T</td></tr><tr><td>20</td><td>1234567</td><td>microsat1</td><td>G,GTCTC</td><td>50</td><td>PASS</td><td>NS=3;DP=9;AA=G</td><td></td></tr></tbody></table>	#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	G,GTCTC	50	PASS	NS=3;DP=9;AA=G		
#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	G,GTCTC	50	PASS	NS=3;DP=9;AA=G																																											

Output of the variant caller: VCF

VCF: Variant Call Format

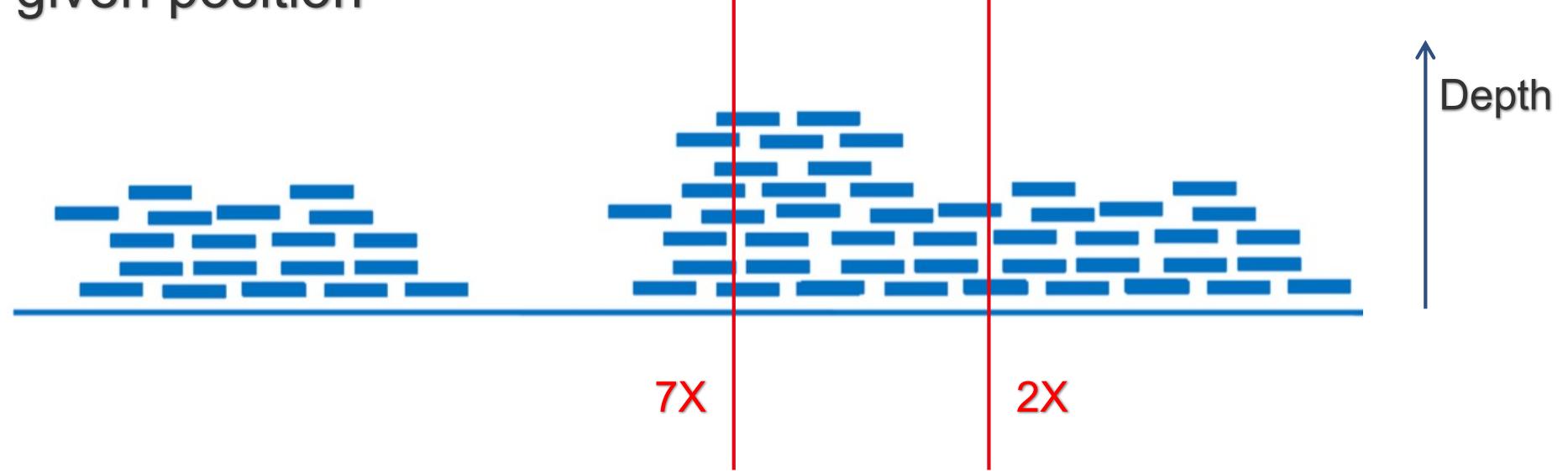
Fixed fields								
	#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
BODY	20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2
	20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017
	20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB
	20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T
	20	1234567	microsat1	GTC	G,GTCTC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



Things to watch out when assessing variant quality

Depth

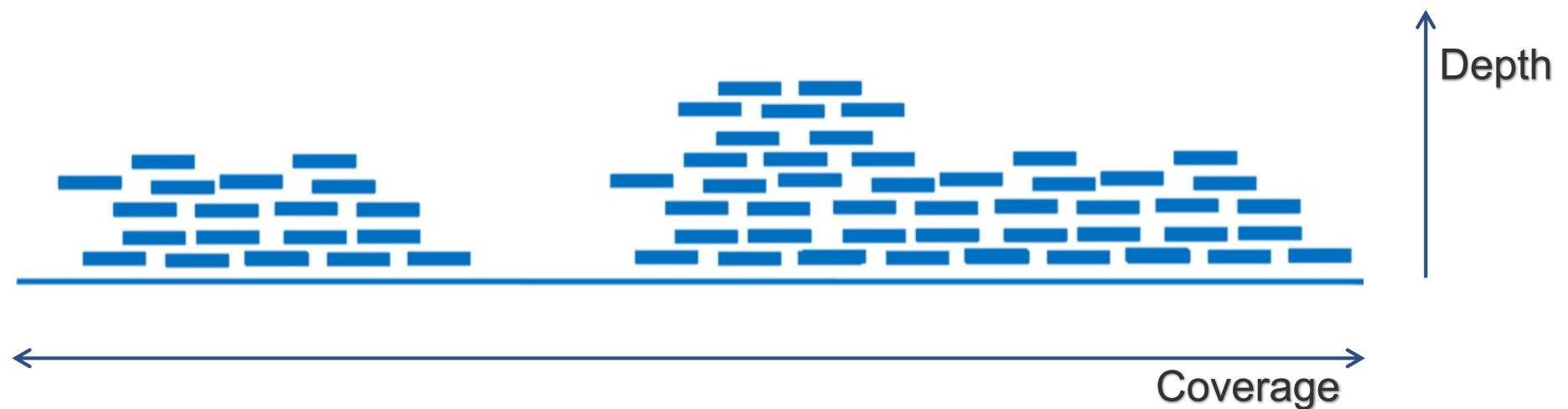
- **Depth:** nb of reads that include a given nucleotide, at a given position



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

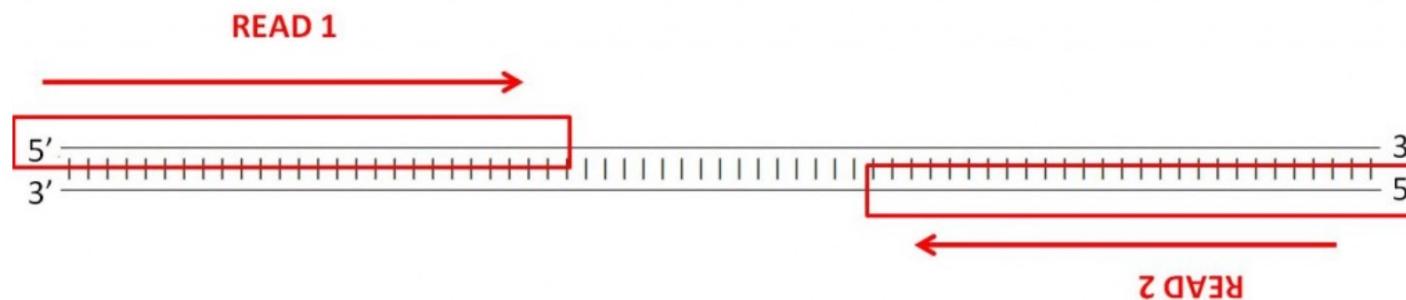
Coverage

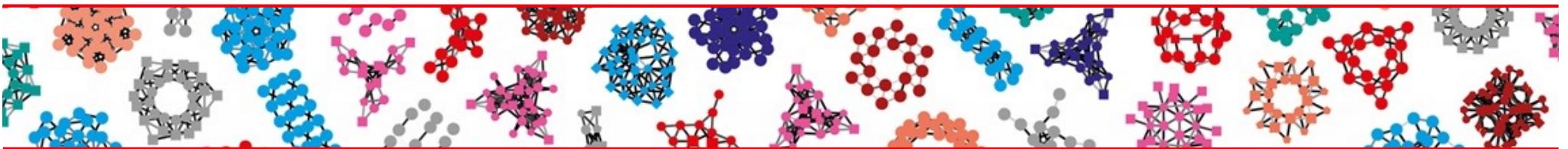
- **Coverage:** % or nb of bases of a reference genome that are covered with a certain depth, e.g. 90% at 5X



Strand bias in paired-end sequencing

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





PART III

Variant annotation
and interpretation

Medical genetics: focus on pathogenicity

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

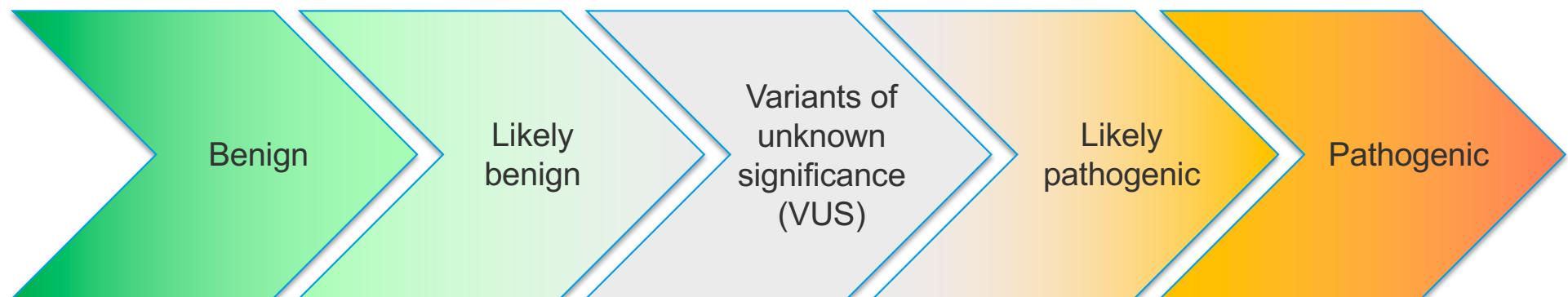
Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD¹, Nazneen Aziz, PhD^{2,16}, Sherri Bale, PhD³, David Bick, MD⁴, Soma Das, PhD⁵, Julie Gastier-Foster, PhD^{6,7,8}, Wayne W. Grody, MD, PhD^{9,10,11}, Madhuri Hegde, PhD¹², Elaine Lyon, PhD¹³, Elaine Spector, PhD¹⁴, Karl Voelkerding, MD¹³ and Heidi L. Rehm, PhD¹⁵; 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



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

Tier I: Variants of Strong Clinical Significance

Therapeutic, prognostic & diagnostic

Level A Evidence

FDA-approved therapy
Included in professional guidelines

Level B Evidence

Well-powered studies with consensus from experts in the field

Tier II: Variants of Potential Clinical Significance

Therapeutic, prognostic & diagnostic

Level C Evidence

FDA-approved therapies for different tumor types or investigational therapies
Multiple small published studies with some consensus

Level D Evidence

Preclinical trials or a few case reports without consensus

Tier III: Variants of Unknown Clinical Significance

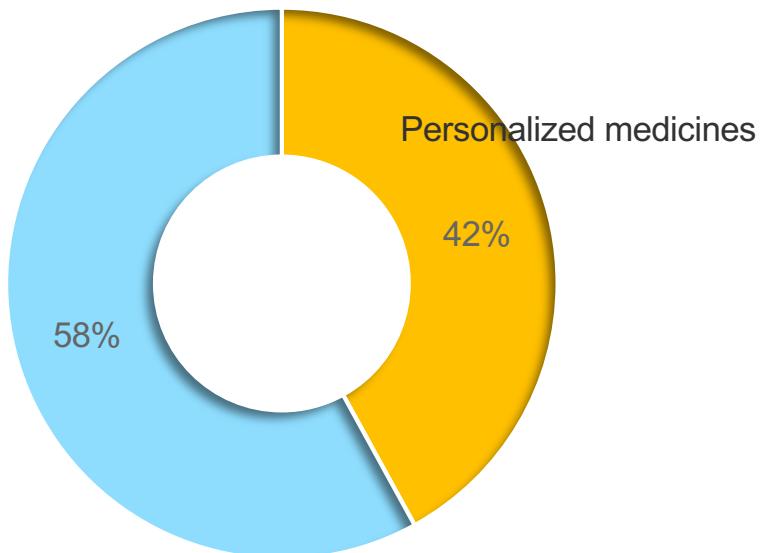
Not observed at a significant allele frequency in the general or specific subpopulation databases, or pan-cancer or tumor-specific variant databases
No convincing published evidence of cancer association

Tier IV: Benign or Likely Benign Variants

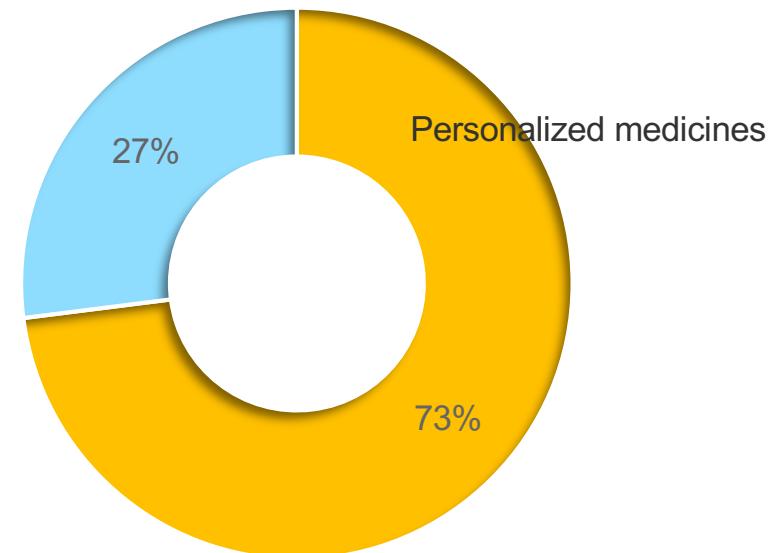
Observed at significant allele frequency in the general or specific subpopulation databases
No existing published evidence of cancer association

Personalized molecular oncology

All drugs
in development



Oncology drugs
in development



Finding actionable variants



Predisposition

*Indicates risk
to develop a disease*

Diagnostic

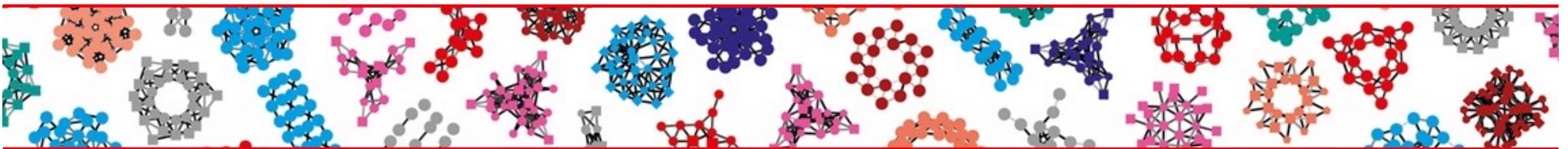
*Supports disease
characterization*

Prognostic

*Indicates disease
evolution*

Predictive

*Supports treatment
decisions*

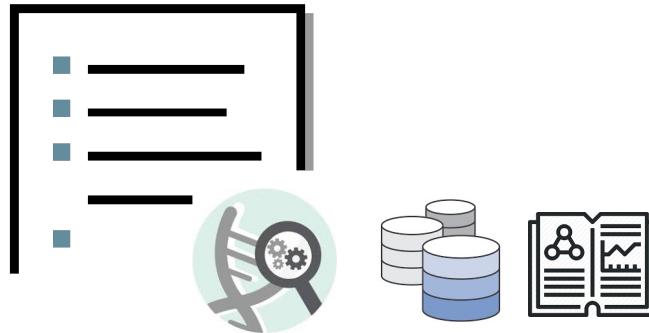


PART III

Variant annotation and interpretation

... bioinformatics at the rescue

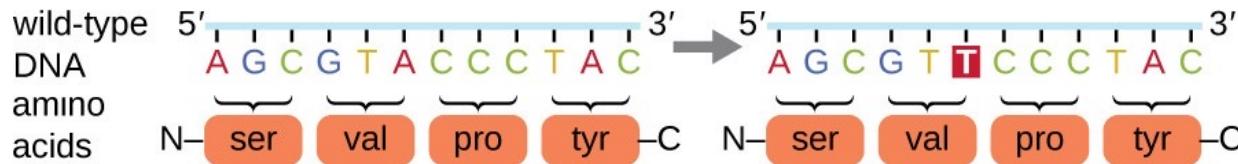
Bioinformatics to the rescue... for annotation



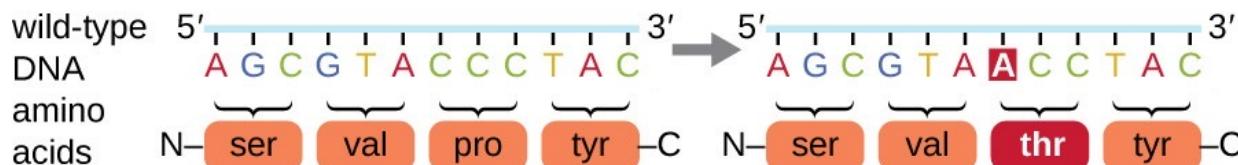
- **Genes and transcripts affected by the variant**
- **Location of the variants (e.g. intron, exon, regulatory region... ...)**
- **Predict variant effect (e.g. stop gained, missense...)**

Variant effects on the protein

silent: has no effect on the protein sequence

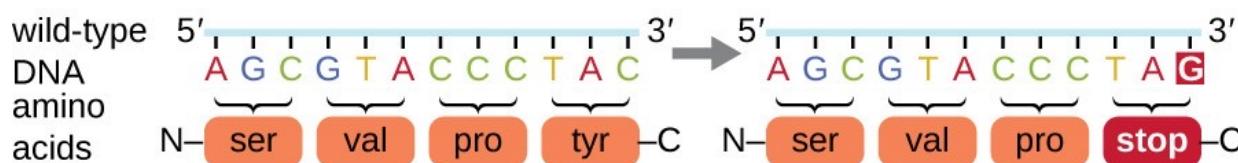


missense: results in an amino acid substitution

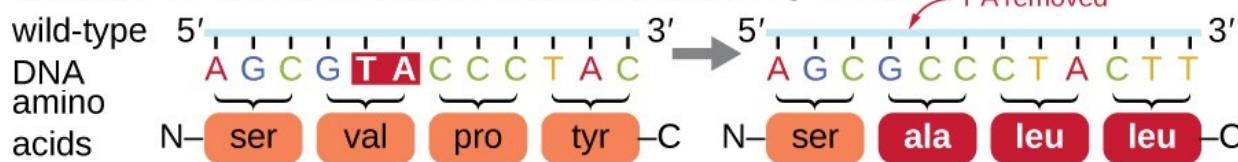


Point mutations
(single base
substitution)

nonsense: substitutes a stop codon for an amino acid

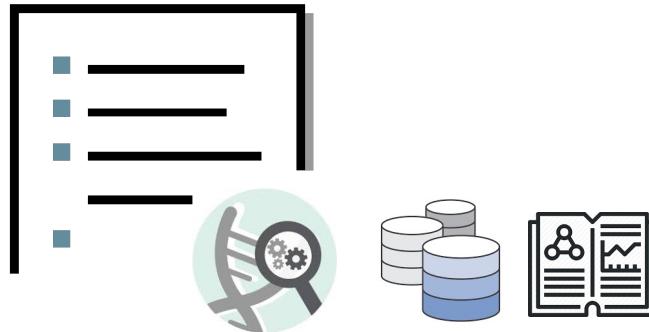


Insertion or deletion results in a shift in the reading frame.



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

Bioinformatics to the rescue... for annotation



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

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 (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.

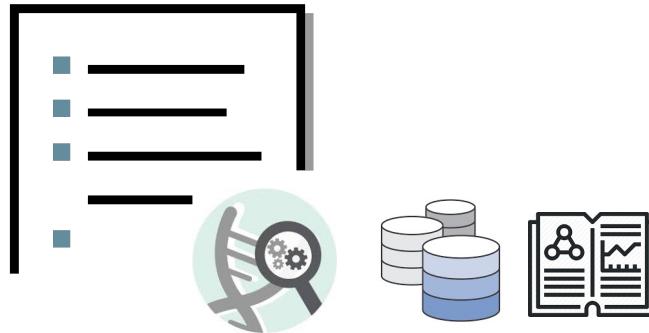
© American College of Medical Genetics and Genomics | ACMG STANDARDS AND GUIDELINES | Genetics in Medicine

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

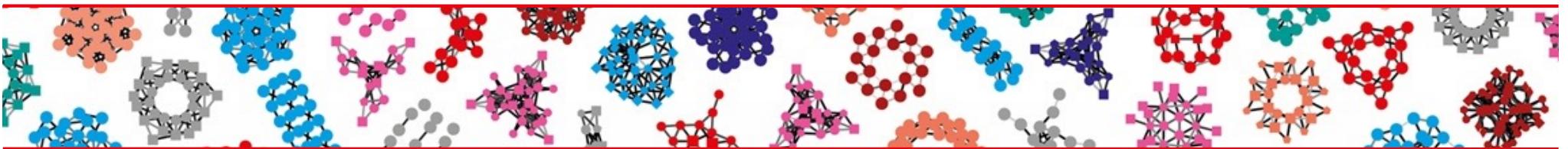
Sue Richards, PhD¹, Nazneen Aziz, PhD^{1,6}, Sherri Bale, PhD², David Bick, MD⁴, Soma Das, PhD⁵, Julie Gastier-Foster, PhD^{6,7,8}, Wayne W. Grody, MD, PhD^{1,10,11}, Madhuri Hegde, PhD¹², Elaine Lyon, PhD⁹, Elaine Spector, PhD¹⁴, Karl Voelkerding, MD¹³ and Heidi L. Rehm, PhD¹⁵, on behalf of the ACMG Laboratory Quality Assurance Committee

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

Bioinformatics to the rescue... for annotation



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

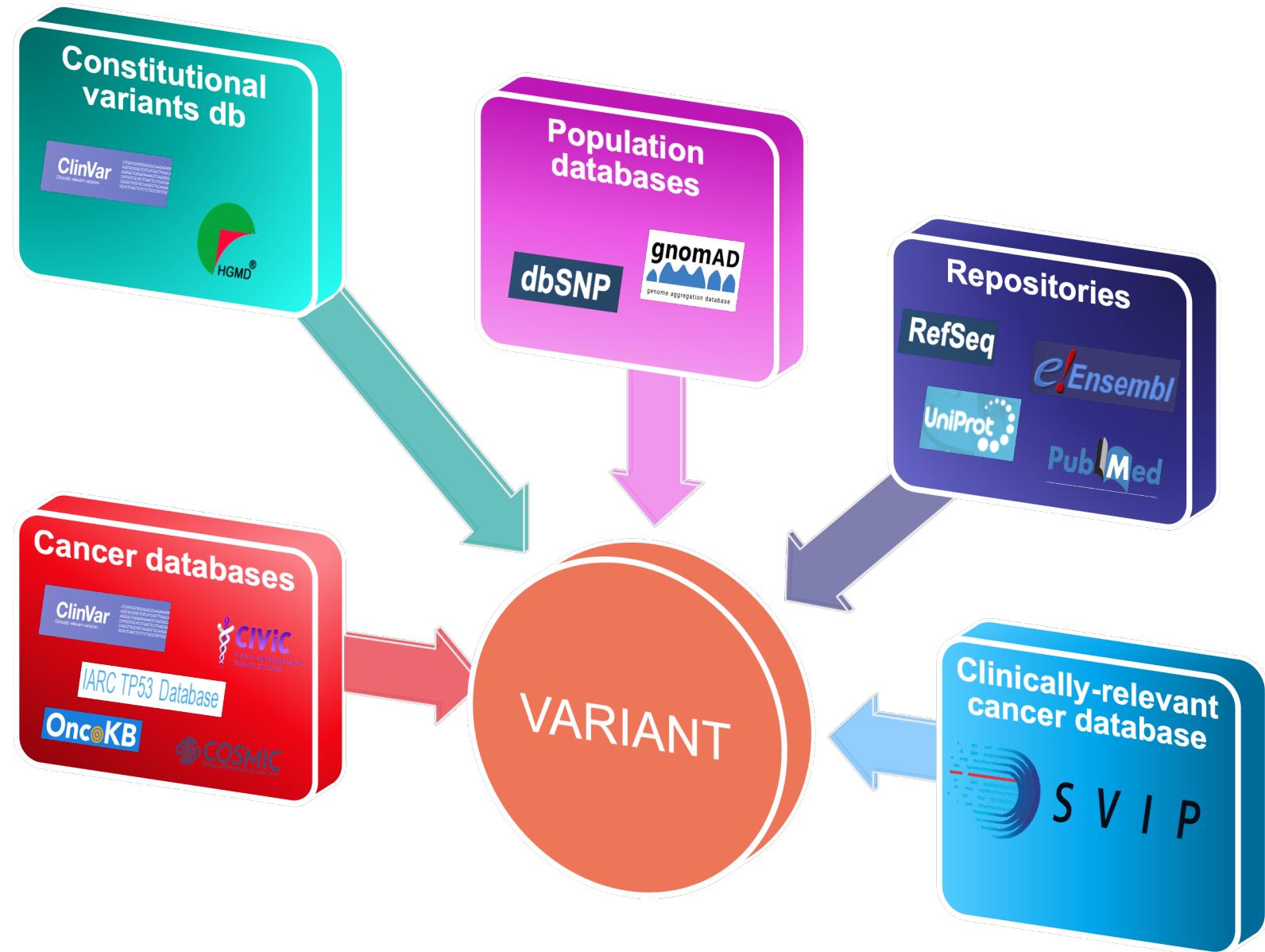


PART III

Variant annotation
and interpretation

... with knowledge-bases

Knowledge bases



Non exhaustive

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

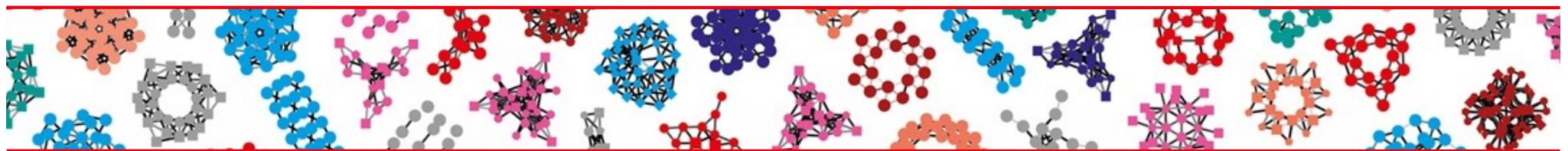
- Is the mutation in an **evolutionarily conserved** region accross species?



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**



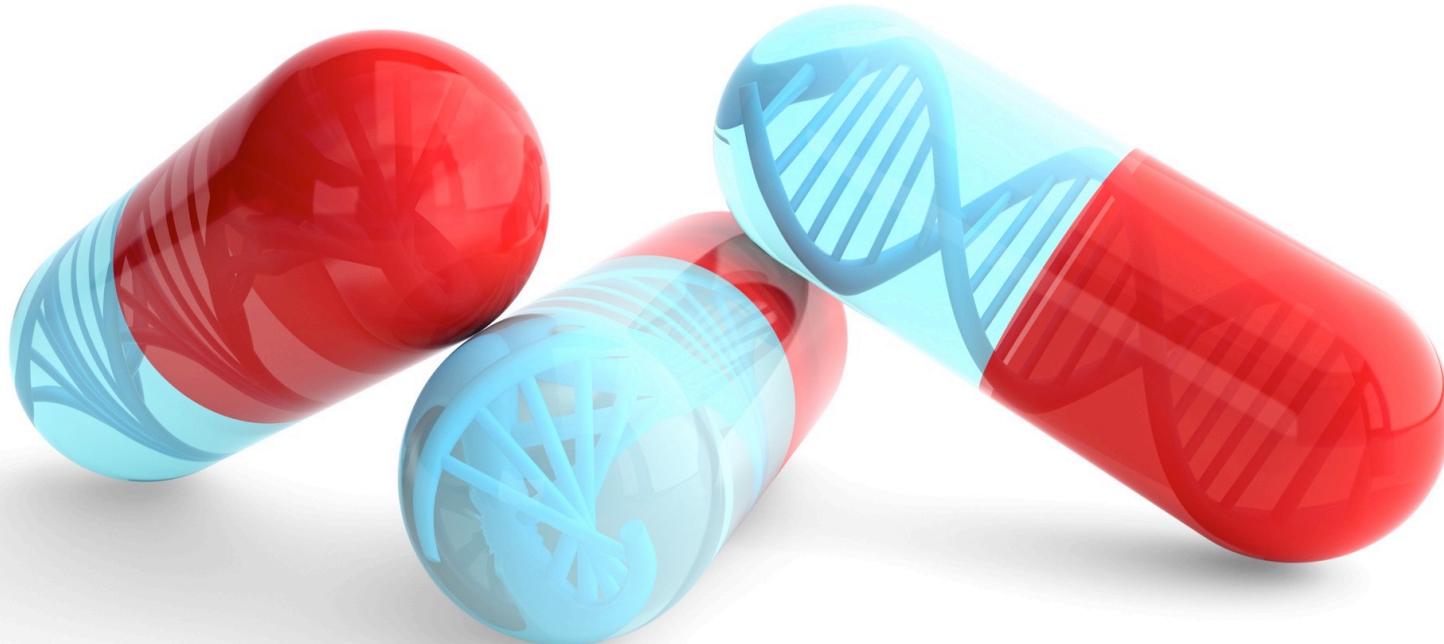
Other considerations...

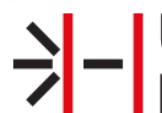
Real-life constraints in the clinics



Certificate of Advanced Studies (CAS) in Personalized molecular oncology

pmo.unibas.ch



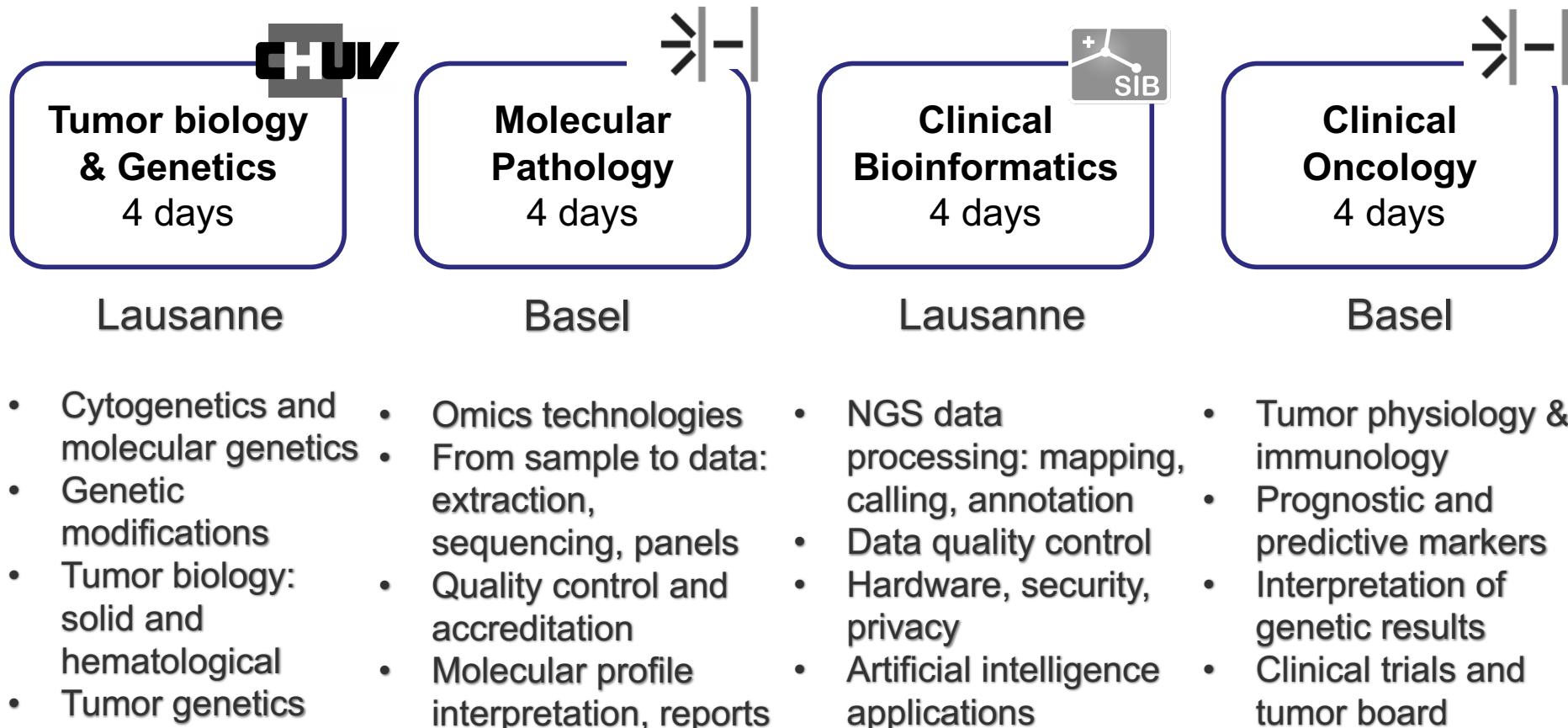
 Universitätsspital
Basel

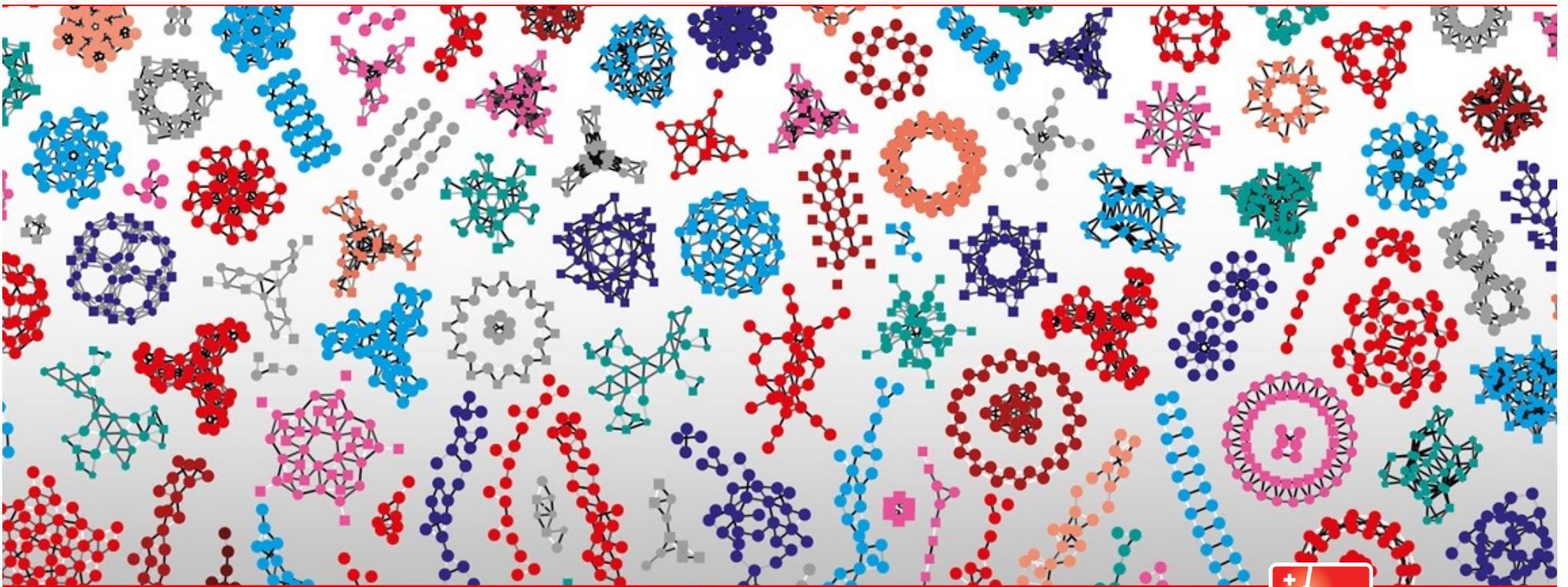


Swiss Institute of
Bioinformatics



CAS PMO: 4 modules and a mini-thesis





SIB
Swiss Institute of
Bioinformatics

Thank You

