Human Genomics and Epigenomics

Pratical 1 – 18/01/2021
Pratical 2 – 19/01/2021
Pratical 3 – 25/01/2021
Pratical 4 – 26/01/2021

Prof. Massimo Delledonne
Functional Genomics lab

ALIGNMENT AND VARIANT CALLING

1° Day (3h): Pre-processing of raw reads

- The fastq file
- Quality control of fastq files
- Adapter removing and trimming of fastq files
  - Sickle and scythe
  - Trimmomatic
- Reads alignment:
  - The human reference genome (hg19 and hg38, main differences)
  - The BAM file

2° Day (3h): Alignment

- Alignment of trimmed reads to the reference genome
  - BWA-mem
  - Isaac2 pipeline
- Duplicates removal
- Read Clipping
- Visualization of aligned reads on IGV

ALIGNMENT AND VARIANT CALLING

3° Day (3h): Statistics and Variant Calling

- Statistics on reads alignment: main parameters for the evaluation of NGS data
  - Average coverage
  - Uniformity
  - Fold enrichment (on/near/off target)
  - Genotypability
- Variant calling:
  - The VCF and gVCF files
  - Germline variant calling
    - GATK4 Best practice pipeline

4° Day (3h): Variant Calling

- Germline variant calling
  - GATK4 Best practice pipeline
  - Strelka2
- Visualization of genetic variants on IGV
- Structural detection

STATISTICS ON READS ALIGNMENT: MAIN PARAMETERS FOR THE EVALUATION OF NGS DATA

Average coverage

Coverage (or depth) is the number of unique reads that are aligned in a specific position.

Average coverage for **whole genome**:

$$\frac{N * L}{G}$$

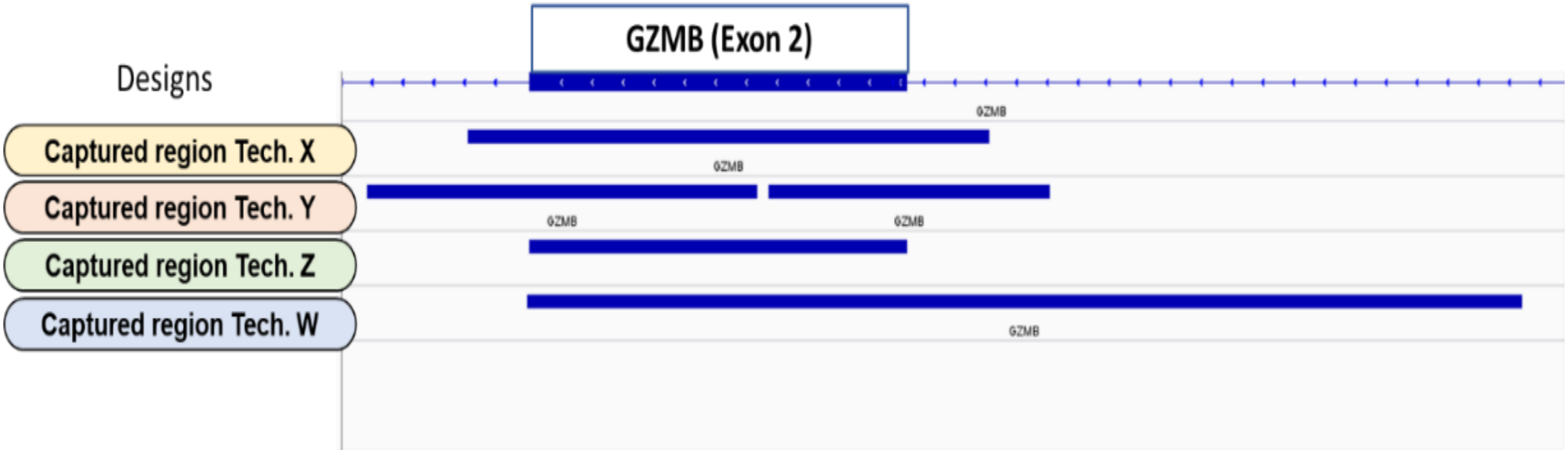
Where:
G = the length of the genome
N = the number of reads
L = average read length

Average coverage for **whole exome**:

$$\frac{N * L}{E}$$

Where:
E = the length of the exome
N = the number of reads
L = average read length

Capture kit (Design) and RefSeq



The captured region is different for each capture kit, depending on how the kit is designed by the company.

Connect to server

1. Enter in the server:
   a. ssh lessons@157.27.80.26
   b. Password: lez2021
2. Enter in the created folder: cd HGE_2021/your_name

Calculate coverage for exome RefSeq

- Calculate coverage for the RefSeq:

```
/opt/bedtools coverage -hist -abam
sample.sorted.dedup.clipped.bwa.bamUtils.bam -b ../ref/refseq.chr6.bed
| gzip > refseq-capture.hist.coverage.gz
```

Calculate coverage for exome RefSeq

- Look at the obtained results:

less refseq-capture.hist.coverage.gz

chr6	17825767	17825934	KIF13A	109	1	167	0.0059880
chr6	17825767	17825934	KIF13A	110	9	167	0.0538922
chr6	17825767	17825934	KIF13A	114	1	167	0.0059880
chr6	17825767	17825934	KIF13A	115	1	167	0.0059880
chr6	17825767	17825934	KIF13A	116	1	167	0.0059880
chr6	17825767	17825934	KIF13A	119	4	167	0.0239521
chr6	17825767	17825934	KIF13A	120	6	167	0.0359281
chr6	17825767	17825934	KIF13A	121	2	167	0.0119760
chr6	17825767	17825934	KIF13A	122	1	167	0.0059880
chr6	17825767	17825934	KIF13A	123	3	167	0.0179641

CHR | START POSITION | END POSITION | GENE | DEPTH | #BASES AT DEPTH | SIZE OF TARGET REFSEQ | %OF TARGET REFSEQ AT DEPTH

Calculate coverage for design kit

- Calculate coverage for the design:

```
/opt/bedtools coverage -hist -abam
sample.sorted.dedup.clipped.bwa.bamUtils.bam -b ../ref/design.chr6.bed
| gzip > design-capture.hist.coverage.gz
```

Calculate coverage for design kit

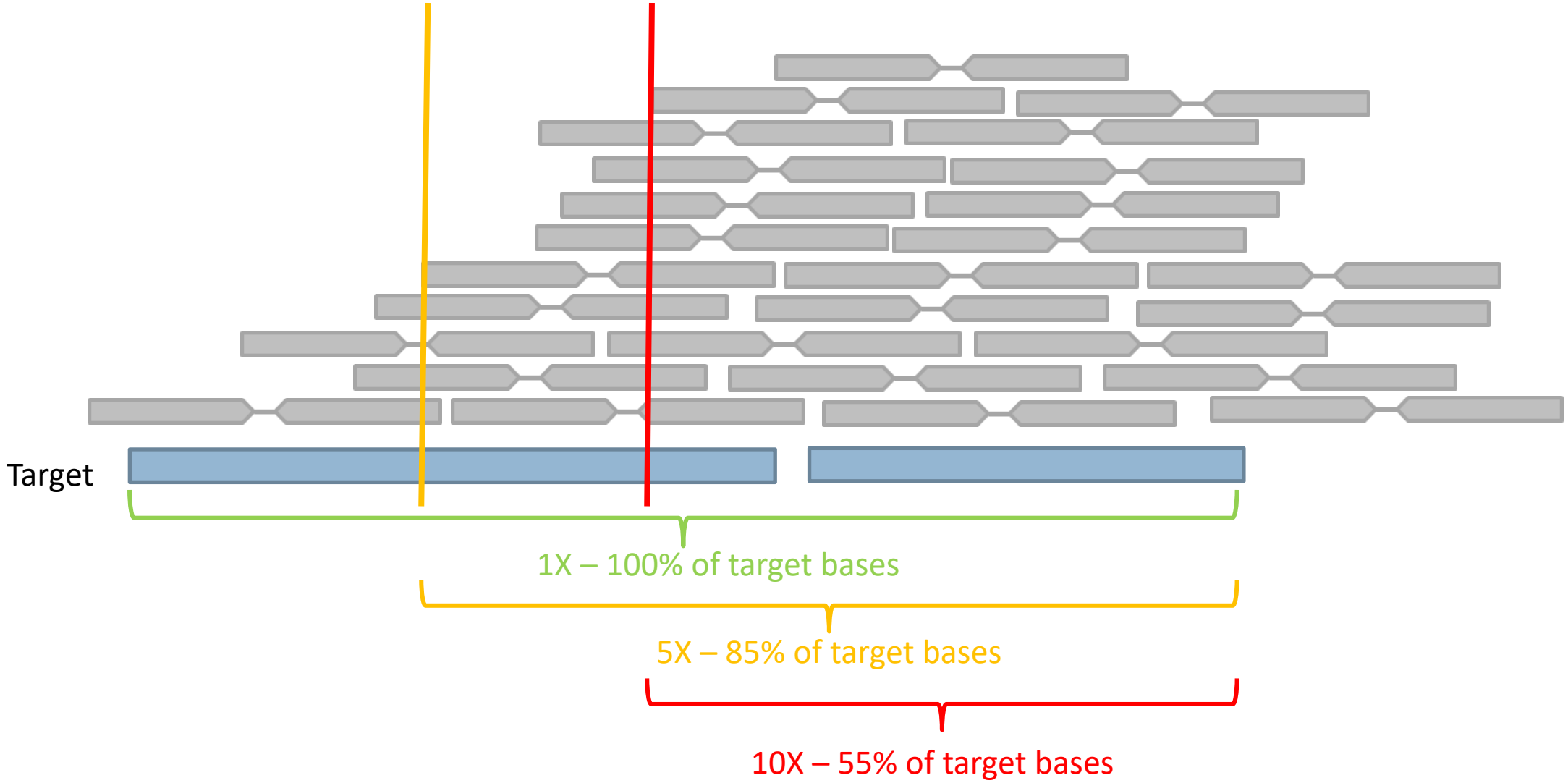
- Look at the obtained results:

less -S design-capture.hist.coverage.gz

CHR	START POSITION	END POSITION	GENE	DEPTH	#BASES AT DEPTH	SIZE OF TARGET REFSEQ	%OF TARGET REFSEQ AT DEPTH
chr6	17825767	17825934	KIF13A	109	1	167	0.0059880
chr6	17825767	17825934	KIF13A	110	9	167	0.0538922
chr6	17825767	17825934	KIF13A	114	1	167	0.0059880
chr6	17825767	17825934	KIF13A	115	1	167	0.0059880
chr6	17825767	17825934	KIF13A	116	1	167	0.0059880
chr6	17825767	17825934	KIF13A	119	4	167	0.0239521
chr6	17825767	17825934	KIF13A	120	6	167	0.0359281
chr6	17825767	17825934	KIF13A	121	2	167	0.0119760
chr6	17825767	17825934	KIF13A	122	1	167	0.0059880
chr6	17825767	17825934	KIF13A	123	3	167	0.0179641
chr6	17825767	17825934	KIF13A	124	1	167	0.0059880
chr6	17825767	17825934	KIF13A	130	1	167	0.0059880

Coverage 1X/5X/10X/20X/30X

Percentage of target bases covered by at least 1/5/10/20/30 reads



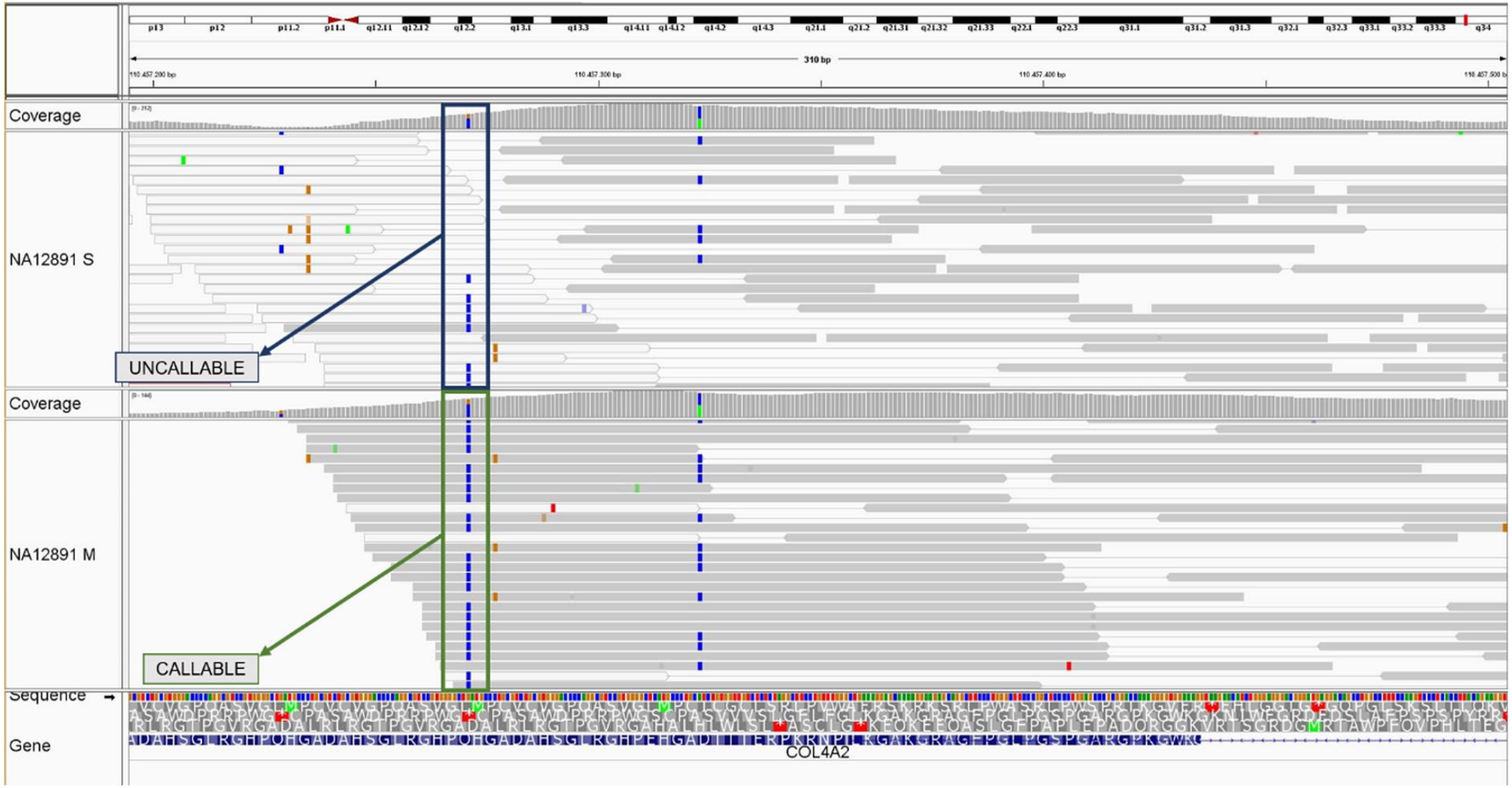
Statistic results example

		DESIGN														
CAPTURE KIT	LIBRARY KIT	Design Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS (RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	36,735,875	72.21	99.36	98.82	98.39	97.52	95.76	95.86	95.34	41.30	38.69	19.40	34.90	1.44	96.30

		REF-SEQ														
CAPTURE KIT	LIBRARY KIT	RefSeq Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS(RD >=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	34,298,612	72.43	99.49	99.08	98.68	97.92	96.39	96.35	95.90	38.97	36.75	23.63	35.17	1.43	96.66

Genotypability

Genotypability is the ability to call the genotype in a specific genomic position.
This parameter is based on both **depth of coverage** and **read mapping quality**



Genotypability

- First we need to add the read group to the sample:

```
java -jar /opt/picard.jar AddOrReplaceReadGroups
I=sample.sorted.dedup.clipped.bwa.bamUtils.bam
O=sample.sorted.dedup.clipped.bwa.bamUtils.rg.bam RGID=sample RGLB=lib1
RGPL=ILLUMINA RGPU=unit1 RGSM=20 VALIDATION_STRINGENCY=SILENT
```

- Create the index:

```
samtools index sample.sorted.dedup.clipped.bwa.bamUtils.rg.bam
```

Genotypability

- Calculate callable regions:

java -jar /opt/gatk-3.8/GenomeAnalysisTK.jar -T CallableLoci -R ../ref/chr6.hg38.fa -I sample.sorted.dedup.clipped.bwa.bamUtils.rg.bam --summary callable.txt -o callable.bed

- Check the output:

less callable.bed

chr6	0	60000	REF_N
chr6	60000	60064	NO_COVERAGE
chr6	60064	60215	LOW_COVERAGE
chr6	60215	60222	NO_COVERAGE
chr6	60222	60373	LOW_COVERAGE
chr6	60373	61797	NO_COVERAGE
chr6	61797	62058	LOW_COVERAGE
chr6	62058	62840	NO_COVERAGE
chr6	62840	62991	LOW_COVERAGE
chr6	62991	63113	NO_COVERAGE
chr6	63113	63264	LOW_COVERAGE
chr6	63264	63465	NO_COVERAGE
chr6	63465	63707	LOW_COVERAGE
chr6	63707	67060	NO_COVERAGE
chr6	67060	67293	LOW_COVERAGE
chr6	67293	67375	NO_COVERAGE
chr6	67375	67526	LOW_COVERAGE
chr6	67526	68211	NO_COVERAGE
chr6	68211	68579	LOW_COVERAGE
chr6	68579	68710	NO_COVERAGE
chr6	68710	68978	LOW_COVERAGE
chr6	68978	69267	NO_COVERAGE
chr6	69267	69553	LOW_COVERAGE
chr6	69553	70653	NO_COVERAGE
chr6	70653	70954	LOW_COVERAGE
chr6	70954	70957	CALLABLE
chr6	70957	71162	LOW_COVERAGE
chr6	71162	71800	NO_COVERAGE

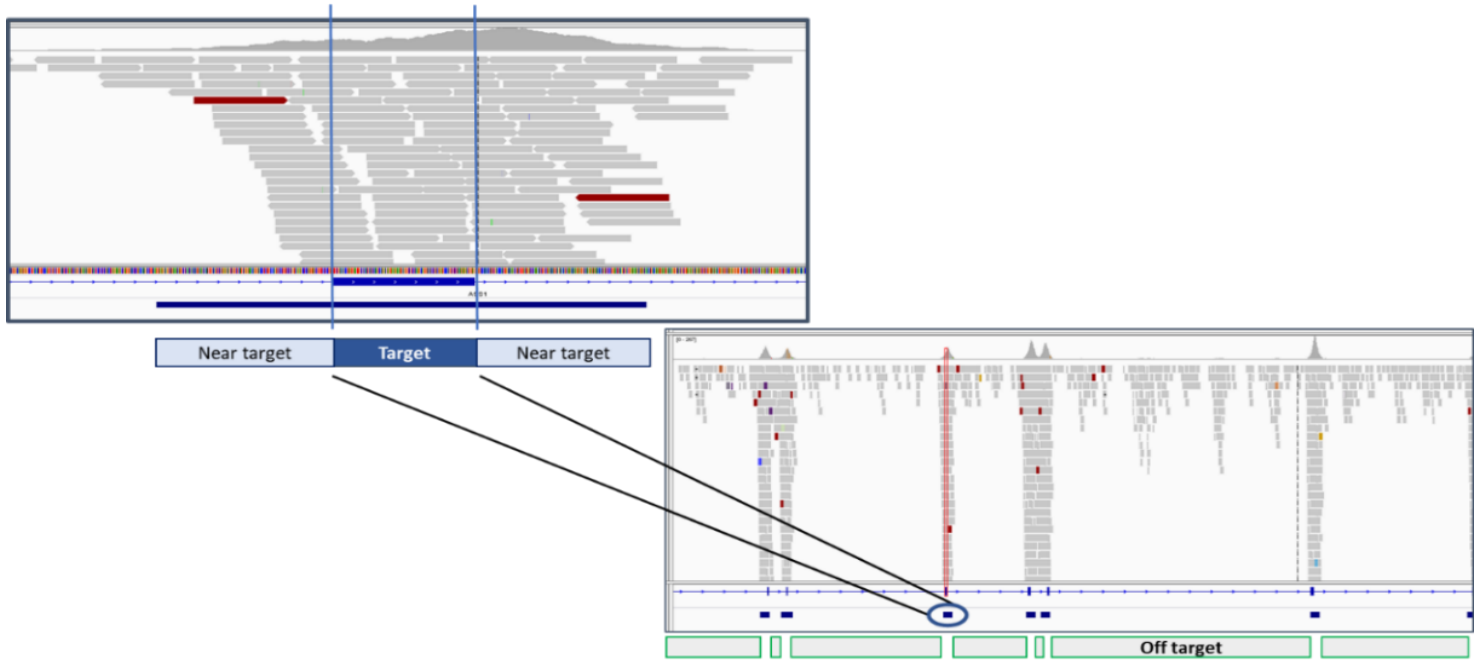
For each region the program say if it is callable or not and why not.

Statistic results example

		DESIGN														
CAPTURE KIT	LIBRARY KIT	Design Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS (RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	36,735,875	72.21	99.36	98.82	98.39	97.52	95.76	95.86	95.34	41.30	38.69	19.40	34.90	1.44	96.30

		REF-SEQ														
CAPTURE KIT	LIBRARY KIT	RefSeq Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS(RD >=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	34,298,612	72.43	99.49	99.08	98.68	97.92	96.39	96.35	95.90	38.97	36.75	23.63	35.17	1.43	96.66

Fold enrichment (on/near/off target)

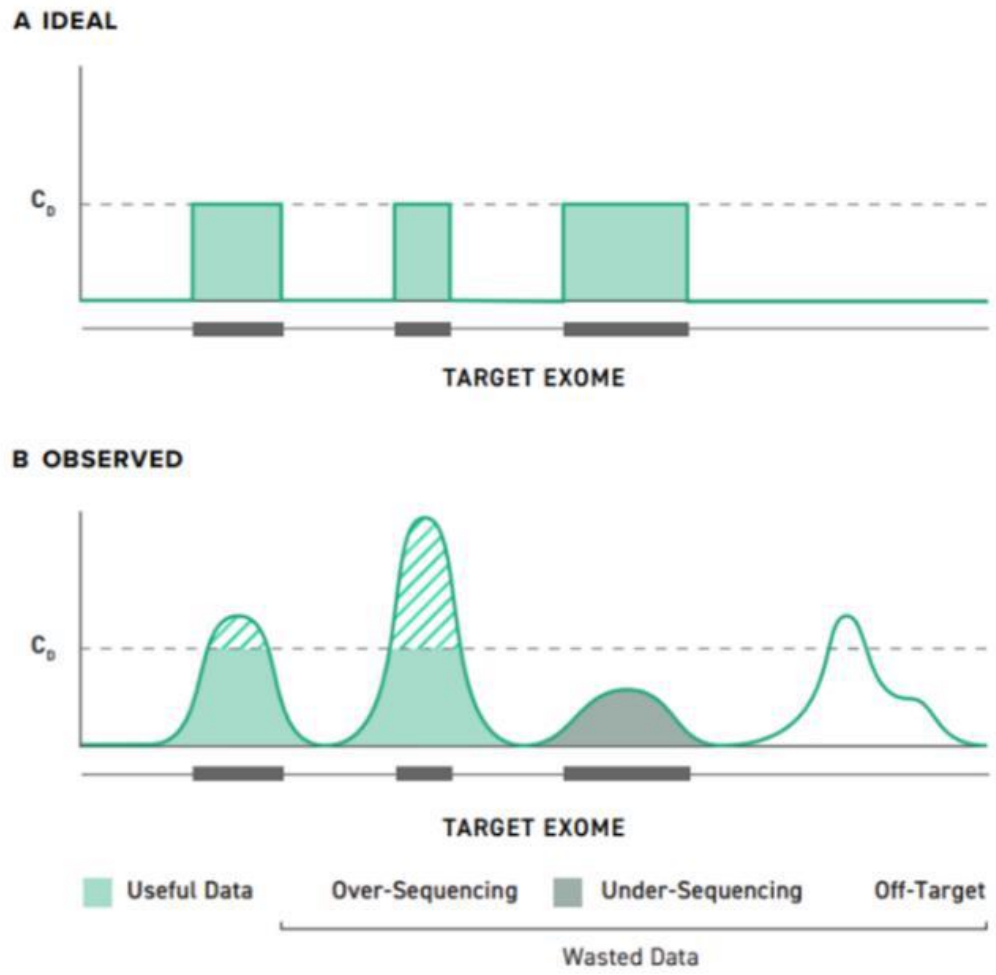


ON TARGET: The number of aligned bases that mapped on the target region of the genome
NEAR TARGET: The number of aligned bases that mapped within a fixed interval close to the target region, but not on the target region (~250 bp)
OFF TARGET: The number of aligned bases that mapped to neither on or near the target region
FOLD ENRICHMENT:

$$\frac{ON\ TARGET / (\text{ON TARGET} + \text{NEAR TARGET} + \text{OFF TARGET})}{TARGET\ LENGTH / \text{GENOME SIZE}}$$

Fold80 base penalty

Fold 80 base penalty is defined as the fold change of **non-zero read coverage** needed to bring 80% of the ROI bases to the observed mean coverage.



Fold80 and Fold Enrichment

- Calculate on/near/off target, fold enrichment and fold80 for the RefSeq:

```
java -jar /opt/picard.jar CollectHsMetrics
I=sample.sorted.dedup.clipped.bwa.bamUtils.bam O=refseq.HsMetrics.txt
R=../ref/chr6.hg38.fa BAIT_INTERVALS=../ref/refseq.chr6.intervals
TARGET_INTERVALS=../ref/refseq.chr6.intervals
PER_TARGET_COVERAGE=refseq.PER_TARGET_COVERAGE.txt
PER_BASE_COVERAGE=refseq.PER_BASE_COVERAGE.txt
VALIDATION_STRINGENCY=SILENT NEAR_DISTANCE=261
```

You can do the same for the design…

Fold80 and Fold Enrichment

```
## htsjdk.samtools.metrics.StringHeader
# CollectHsMetrics BAIT_INTERVALS=[../ref/refseq.chr6.intervals] TARGET_INTERVALS=[../ref/refseq.chr6.intervals] INPUT=sample.sorted.dedup.clipped.bwa.bamU
tils.bam OUTPUT=refseq.HsMetrics.txt PER_TARGET_COVERAGE=refseq.PER_TARGET_COVERAGE.txt PER_BASE_COVERAGE=refseq.PER_BASE_COVERAGE.txt NEAR_DISTANCE=261 VA
LIDATION_STRINGENCY=SILENT REFERENCE_SEQUENCE=../ref/chr6.hg38.fa    METRIC_ACCUMULATION_LEVEL=[ALL_READS] MINIMUM_MAPPING_QUALITY=20 MINIMUM_BASE_QUALITY=
20 CLIP_OVERLAPPING_READS=true INCLUDE_INDELS=false COVERAGE_CAP=200 SAMPLE_SIZE=10000 ALLELE_FRACTION=[0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5]
 VERBOSITY=INFO QUIET=false COMPRESSION_LEVEL=5 MAX_RECORDS_IN_RAM=500000 CREATE_INDEX=false CREATE_MD5_FILE=false GA4GH_CLIENT_SECRETS=client_secrets.json
 USE_JDK_DEFLATER=false USE_JDK_INFLATER=false
## htsjdk.samtools.metrics.StringHeader
# Started on: Thu Jan 14 10:10:53 CET 2021

## METRICS CLASS        picard.analysis.directed.HsMetrics
BAIT_SET        BAIT_TERRITORY  BAIT_DESIGN_EFFICIENCY  ON_BAIT_BASES   NEAR_BAIT_BASES OFF_BAIT_BASES  PCT_SELECTED_BASES      PCT_OFF_BAIT    ON_BAIT_VS_
SELECTED     MEAN_BAIT_COVERAGE      PCT_USABLE_BASES_ON_BAIT        PCT_USABLE_BASES_ON_TARGET      FOLD_ENRICHMENT HS_LIBRARY_SIZE HS_PENALTY_10X  HS_PEN
ALTY_20X  HS_PENALTY_30X  HS_PENALTY_40X  HS_PENALTY_50X  HS_PENALTY_100X TARGET_TERRITORY        GENOME_SIZE     TOTAL_READS     PF_READS        PF_BASES
        PF_UNIQUE_READS PF_UQ_READS_ALIGNED     PF_BASES_ALIGNED        PF_UQ_BASES_ALIGNED     ON_TARGET_BASES PCT_PF_READS    PCT_PF_UQ_READS PCT_PF_UQ_R
EADS_ALIGNED MEAN_TARGET_COVERAGE    MEDIAN_TARGET_COVERAGE  MAX_TARGET_COVERAGE     MIN_TARGET_COVERAGE     ZERO_CVG_TARGETS_PCT    PCT_EXC_DUPE    PCT_EX
C_ADAPTER PCT_EXC_MAPQ    PCT_EXC_BASEQ   PCT_EXC_OVERLAP PCT_EXC_OFF_TARGET      FOLD_80_BASE_PENALTY    PCT_TARGET_BASES_1X     PCT_TARGET_BASES_2X     P
CT_TARGET_BASES_10X    PCT_TARGET_BASES_20X    PCT_TARGET_BASES_30X    PCT_TARGET_BASES_40X    PCT_TARGET_BASES_50X    PCT_TARGET_BASES_100X   PCT_TARGET_B
ASES_250X   PCT_TARGET_BASES_500X   PCT_TARGET_BASES_1000X  PCT_TARGET_BASES_2500X  PCT_TARGET_BASES_5000X  PCT_TARGET_BASES_10000X PCT_TARGET_BASES_25000X
 PCT_TARGET_BASES_50000X PCT_TARGET_BASES_100000X        AT_DROPOUT      GC_DROPOUT      HET_SNP_SENSITIVITY     HET_SNP_Q       SAMPLE  LIBRARY READ_GROUP
refseq  1725835 1       219577840       188673872       140140715       0,744452        0,255548        0,537849        127,229915      0,335649        0,2
77313        39,627882       7104606 3,664435        3,710367        3,761402        3,809886        3,866535        4,178872        1725835 170805979
      4443492 4443492 654188889       4019859 4018504 548392427       486450279       181415393       1       0,904662        0,999663        105,117461
      101     412     0       0,003778        0,112952        0       0,013314        0,011246        0,017987        0,513689        1,34766 0,995415
        0,994263        0,991928        0,990428        0,988696        0,985885        0,976554        0,52246 0,00149 0       0       0       0       0
       0       0       0       1,335066        0,872188        0,994381        23
```

All information are reported.

Exctract the information we are interested on:

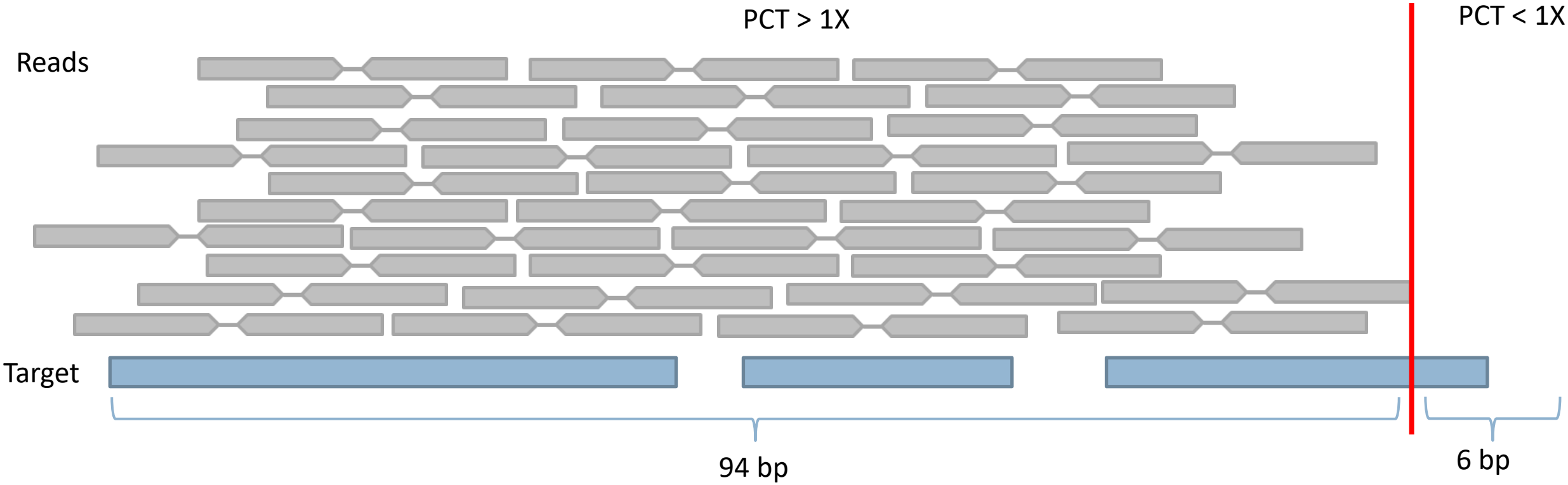
../script/metricsParser_mod.pl refseq.HsMetrics.txt

```
[lessons@localhost denise]$ ../script/metricsParser_mod.pl refseq.HsMetrics.txt
Argument "39,627882" isn't numeric in sprintf at ../script/metricsParser_mod.pl line 227, <FILE> line 8.
Argument "1,34766" isn't numeric in sprintf at ../script/metricsParser_mod.pl line 227, <FILE> line 8.
.       40.04   34.40   25.55   39.00   1.00
```

%ON TARGET BASES | %NEAR TARGET BASES | %OFF TARGET BASES | FOLD ENRICHMENT | FOLD80

Uniformity of coverage (Pct > 0.2*mean)

Uniformity of coverage (Pct > 0.2*mean) is defined as the percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.



Mean Coverage: 10X → Percentage of target bases (PCT) > 0.2 * 10 → Percentage of target bases (PCT) > 1X → Uniformity of coverage (Pct > 0.2*mean): 94%

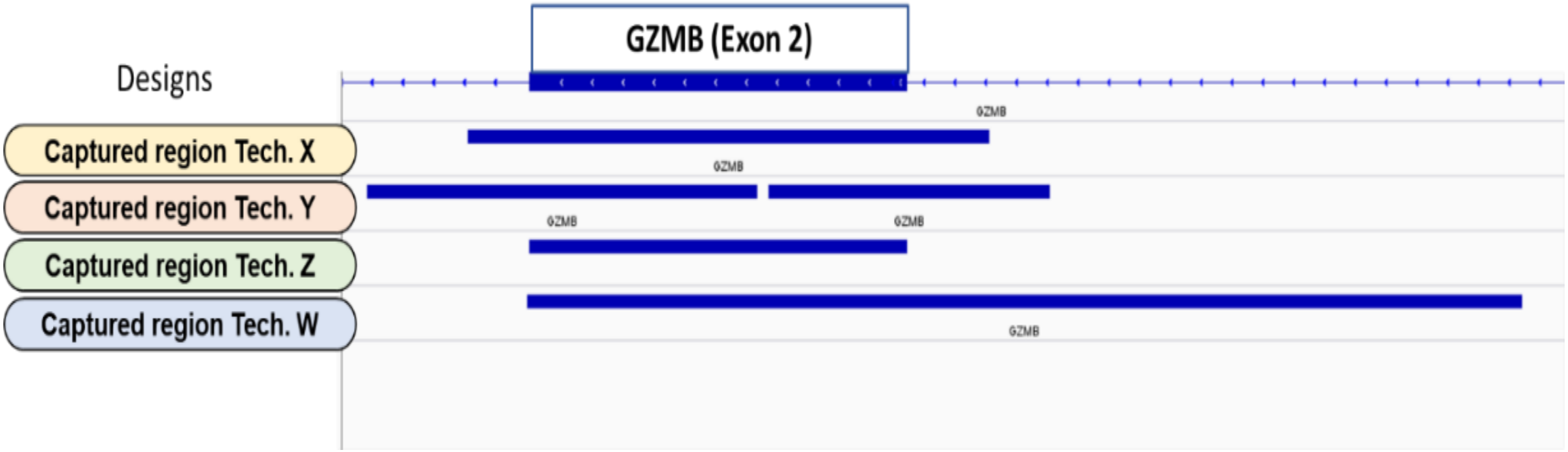
Statistic results example

		DESIGN														
CAPTURE KIT	LIBRARY KIT	Design Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS (RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	36,735,875	72.21	99.36	98.82	98.39	97.52	95.76	95.86	95.34	41.30	38.69	19.40	34.90	1.44	96.30

		REF-SEQ														
CAPTURE KIT	LIBRARY KIT	RefSeq Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS(RD >=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	34,298,612	72.43	99.49	99.08	98.68	97.92	96.39	96.35	95.90	38.97	36.75	23.63	35.17	1.43	96.66

The design is longer than the RefSeq, so the percentage of ON/NEAR target is higher.

Capture kit (Design) and RefSeq



The captured region is different for each capture kit, depending on how the kit is designed by the company.

Statistic results example

							DESIGN														
CAPTURE KIT	LIBRARY KIT	#fragments	%GC	Insert size	#map dedup	%dupl	Design Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS (RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	28,607,138	48.00	287.96	25,216,350	11.51	36,735,875	72.21	99.36	98.82	98.39	97.52	95.76	95.86	95.34	41.30	38.69	19.40	34.90	1.44	96.30
Capture X	KIT 1 Meccanic	26,559,760	48.00	309.54	23,160,195	12.45	36,735,875	66.37	99.52	98.83	98.32	96.89	94.53	95.96	95.33	39.65	39.76	20.56	33.51	1.46	96.29
Capture Y	KIT 2 Enzimatic	48,276,252	49.00	312.91	42,372,357	11.84	35,826,357	137.17	99.55	99.40	99.31	99.10	98.88	96.43	96.38	43.55	45.87	10.71	37.73	1.48	97.43
Capture Y	KIT 2 Meccanic	50,692,820	48.00	311.62	42,977,163	14.37	35,826,357	137.03	99.54	99.34	99.16	98.74	98.20	96.47	96.31	42.42	46.73	10.92	36.75	1.45	96.90

		REF-SEQ														
CAPTURE KIT	LIBRARY KIT	RefSeq Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS(RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	34,298,612	72.43	99.49	99.08	98.68	97.92	96.39	96.35	95.90	38.97	36.75	23.63	35.17	1.43	96.66
Capture X	KIT 1 Meccanic	34,298,612	66.60	99.63	99.06	98.62	97.38	95.29	96.45	95.89	37.43	37.83	24.74	33.78	1.45	96.67
Capture Y	KIT 2 Enzimatic	34,298,612	136.46	99.60	99.49	99.42	99.25	99.04	96.75	96.71	41.72	43.14	15.28	37.65	1.48	97.56
Capture Y	KIT 2 Meccanic	34,298,612	136.47	99.59	99.42	99.26	98.88	98.36	96.78	96.62	40.62	43.89	15.57	36.66	1.45	97.03

Statistic results example

							DESIGN														
CAPTURE KIT	LIBRARY KIT	#fragments	%GC	Insert size	#map dedup	%dupl	Design Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS (RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	28,607,138	48.00	287.96	25,216,350	11.51	36,735,875	72.21	99.36	98.82	98.39	97.52	95.76	95.86	95.34	41.30	38.69	19.40	34.90	1.44	96.30
Capture X	KIT 1 Meccanic	26,559,760	48.00	309.54	23,160,195	12.45	36,735,875	66.37	99.52	98.83	98.32	96.89	94.53	95.96	95.33	39.65	39.76	20.56	33.51	1.46	96.29
Capture Y	KIT 2 Enzimatic	48,276,252	49.00	312.91	42,372,357	11.84	35,826,357	137.17	99.55	99.40	99.31	99.10	98.88	96.43	96.38	43.55	45.87	10.71	37.73	1.48	97.43
Capture Y	KIT 2 Meccanic	50,692,820	48.00	311.62	42,977,163	14.37	35,826,357	137.03	99.54	99.34	99.16	98.74	98.20	96.47	96.31	42.42	46.73	10.92	36.75	1.45	96.90

		REF-SEQ														
CAPTURE KIT	LIBRARY KIT	RefSeq Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS(RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	34,298,612	72.43	99.49	99.08	98.68	97.92	96.39	96.35	95.90	38.97	36.75	23.63	35.17	1.43	96.66
Capture X	KIT 1 Meccanic	34,298,612	66.60	99.63	99.06	98.62	97.38	95.29	96.45	95.89	37.43	37.83	24.74	33.78	1.45	96.67
Capture Y	KIT 2 Enzimatic	34,298,612	136.46	99.60	99.49	99.42	99.25	99.04	96.75	96.71	41.72	43.14	15.28	37.65	1.48	97.56
Capture Y	KIT 2 Meccanic	34,298,612	136.47	99.59	99.42	99.26	98.88	98.36	96.78	96.62	40.62	43.89	15.57	36.66	1.45	97.03

Statistic results example: statistics on downsampled data on 60X mapped coverage

					DESIGN														
CAPTURE KIT	LIBRARY KIT	Avg.Size	Insert size	#map_dedup	Design Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS (RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	455	287.96	20,516,147	36,735,875	59.86	99.33	98.76	98.22	96.92	93.00	95.79	95.20	41.30	38.69	19.40	34.90	1.44	96.27
Capture X	KIT 1 Meccanic	487	309.54	20,939,234	36,735,875	59.83	99.52	98.77	98.21	96.48	92.21	95.92	95.21	39.65	39.76	20.56	33.51	1.47	96.29
Capture Y	KIT 2 Enzimatic	503	312.90	18,721,238	35,826,357	60.32	99.49	99.28	99.02	97.37	91.61	96.38	96.15	43.55	45.87	10.71	37.73	1.51	97.30
Capture Y	KIT 2 Meccanic	495	311.64	18,758,166	35,826,357	60.30	99.46	99.11	98.62	96.79	91.71	96.33	95.86	42.42	46.73	10.91	36.75	1.48	96.80

					REF-SEQ														
CAPTURE KIT	LIBRARY KIT	Avg.Size	Insert size	#map_dedup	RefSeq Length	MEAN	%1X	%5X	%10X	%20X	%30X	%PASS	%PASS (RD>=10)	%ON TARGET BASES	%NEAR TARGET BASES	%OFF TARGET BASES	Fold enrich	fold80	Uniformity of coverage (Pct > 0.2*mean)
Capture X	KIT 1 Enzimatic	455	287.96	20,516,147	34,298,612	60.00	99.46	99.02	98.55	97.43	93.89	96.30	95.75	38.97	36.75	23.63	35.17	1.45	96.63
Capture X	KIT 1 Meccanic	487	309.54	20,939,234	34,298,612	59.99	99.63	99.01	98.52	97.04	93.16	96.41	95.78	37.43	37.83	24.74	33.78	1.45	96.65
Capture Y	KIT 2 Enzimatic	503	312.90	18,721,238	34,298,612	60.00	99.56	99.40	99.18	97.53	91.73	96.71	96.49	41.71	43.14	15.28	37.65	1.51	97.43
Capture Y	KIT 2 Meccanic	495	311.64	18,758,166	34,298,612	60.00	99.52	99.22	98.76	96.95	91.88	96.65	96.19	40.62	43.89	15.57	36.66	1.48	96.93

Capture Y:

More bases covered 10X.

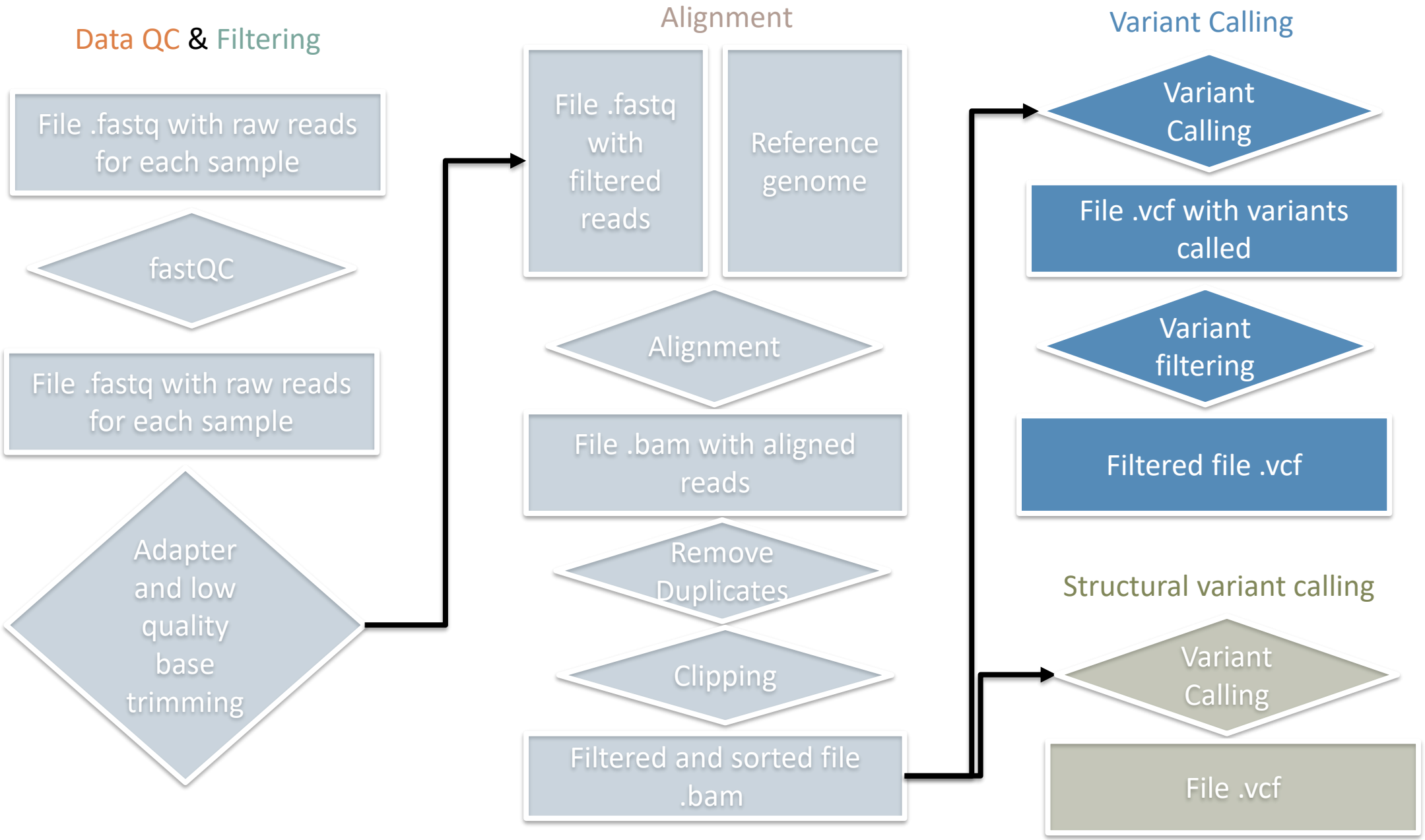
More genotypable regions.

Higher fold80.

Increased uniformity of coverage.

VARIANT CALLING

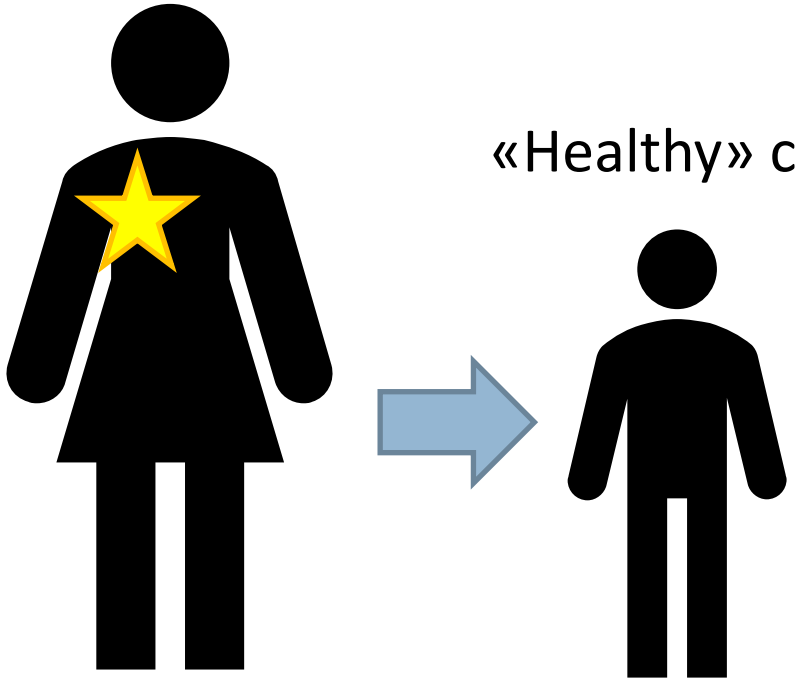
Pipeline


Data QC & Filtering
File .fastq with raw reads for each sample
fastQC
File .fastq with raw reads for each sample
Adapter and low quality base trimming
Alignment
File .fastq with filtered reads
Reference genome
Alignment
File .bam with aligned reads
Remove Duplicates
Clipping
Filtered and sorted file .bam
Variant Calling
Variant Calling
File .vcf with variants called
Variant filtering
Filtered file .vcf
Structural variant calling
Variant Calling
File .vcf

Different type of variants

Somatic Variant

Parent

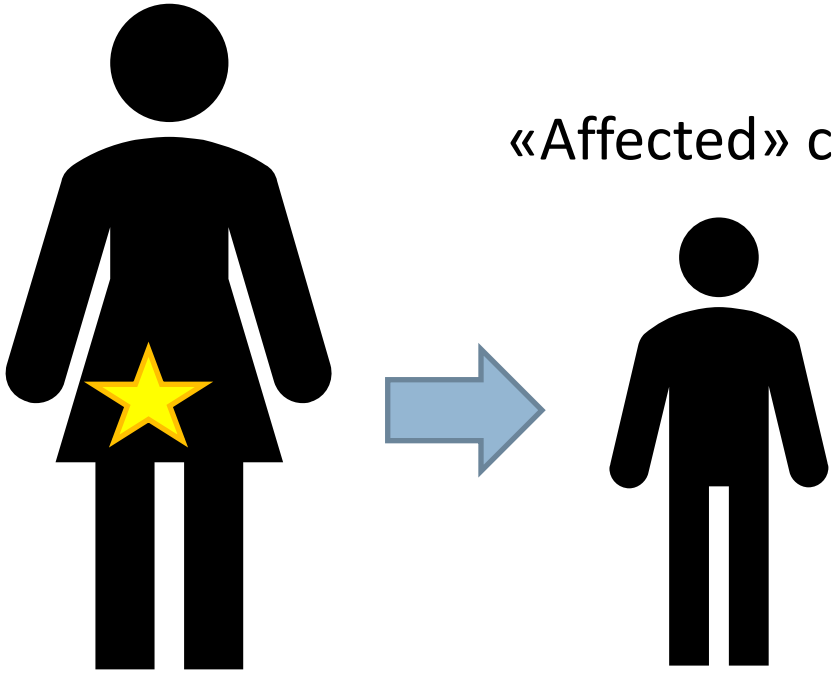


«Healthy» child

- Non germline tissue
- Not trasmitted to child

Germline Variant

Parent



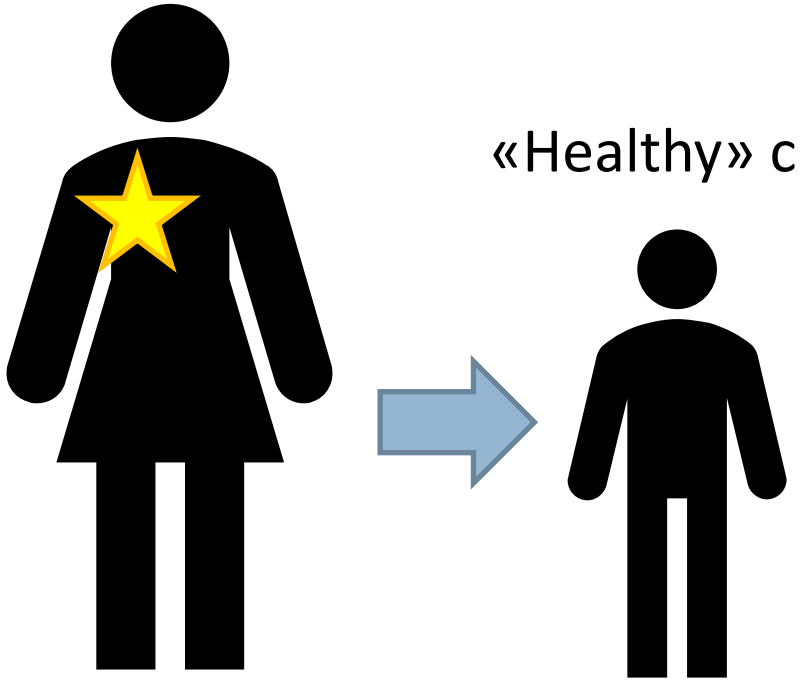
«Affected» child

- Mutation in egg or sperm
- Trasmitted to child

Different type of variants

Somatic Variant

Parent

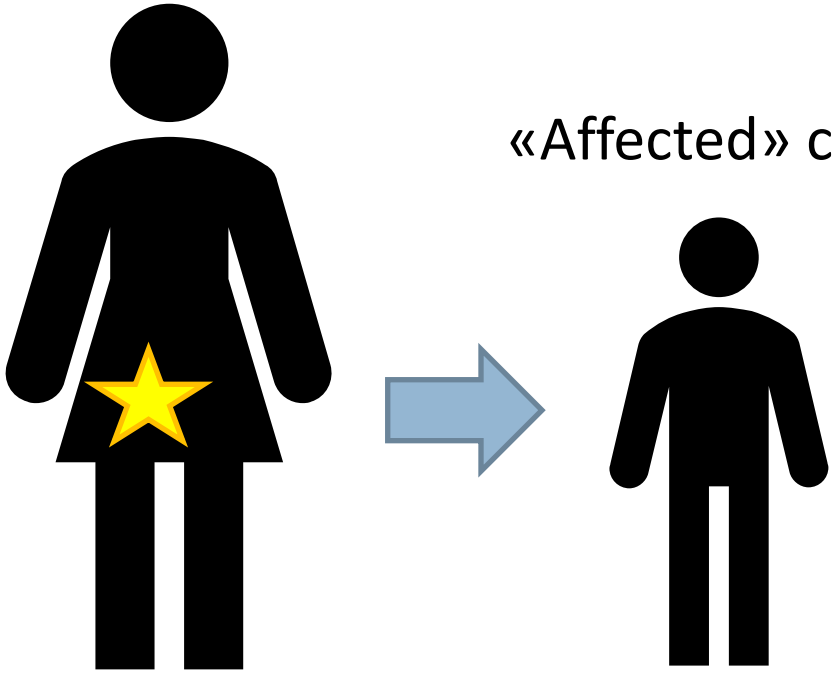


«Healthy» child

- Non germline tissue
- Not trasmitted to child

Germline Variant

Parent



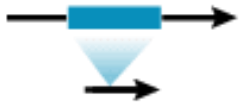
«Affected» child

- Mutation in egg or sperm
- Trasmitted to child

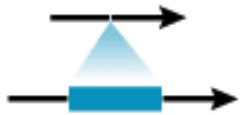
Different type of variants

Single Nucleotide Variant

Deletion



Insertion



Tandem Duplication

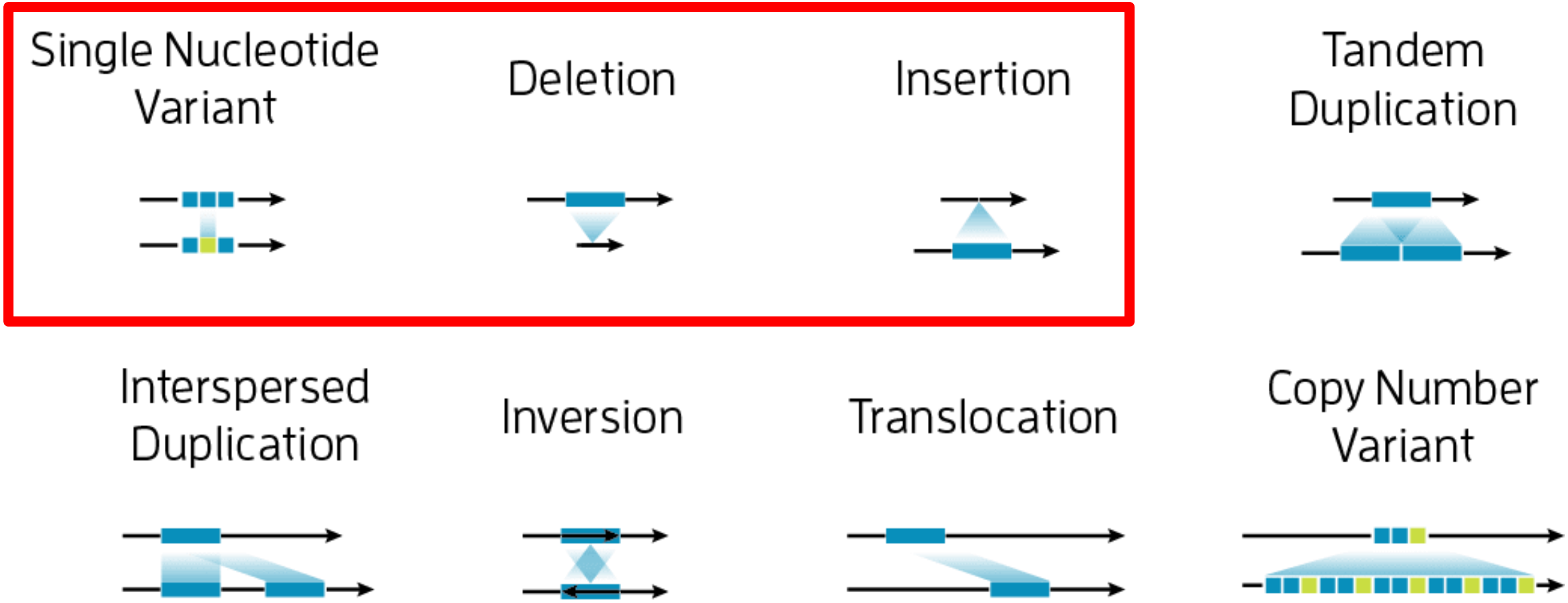
Interspersed Duplication

Inversion

Translocation

Copy Number Variant

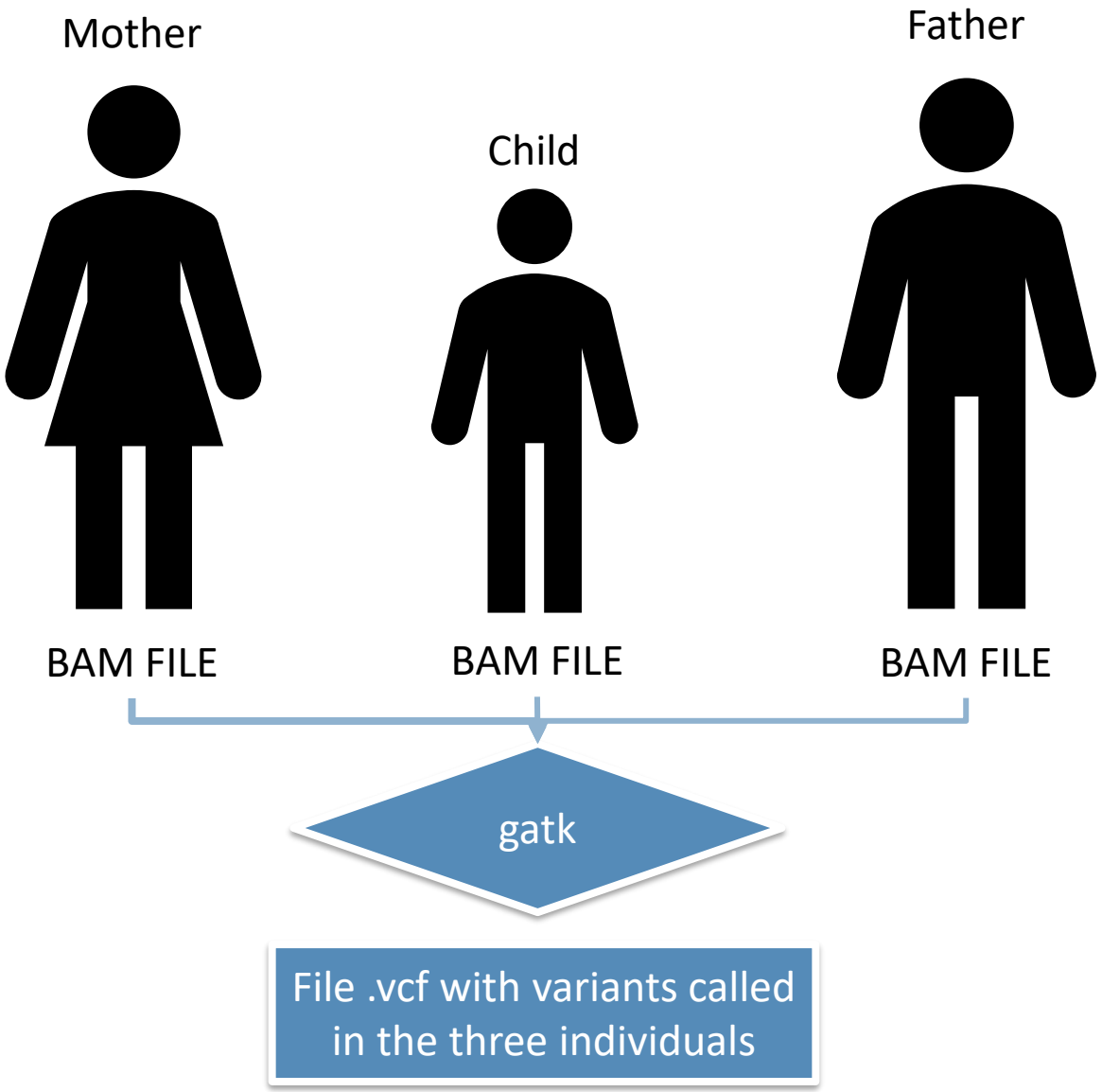
Different type of variants



Variant Calling

- Starts from alignment data to find differences on the genome
- Decision to call a variant depends on many aspects:
  - Alignment quality
  - Read quality
  - Base coverage
  - …
- Many software are available:
  - SOAP2
  - SamTools
  - **GATK**
  - Strelka2
  - …

Family Analysis



VCF

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

```
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1   SAMPLE2
1       1  .      ACG  A,AT    .   PASS    .                    GT:DP      1/2:13    0/0:29
1       2  rs1    C    T,CT    .   PASS    H2;AA=T              GT:GQ      0|1:100   2/2:70
1       5  .      A    G       .   PASS    .                    GT:GQ      1|0:77    1/1:95
1     100  .      T    <DEL>   .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3  0/0:20
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

Large SV

Insertion

Other event

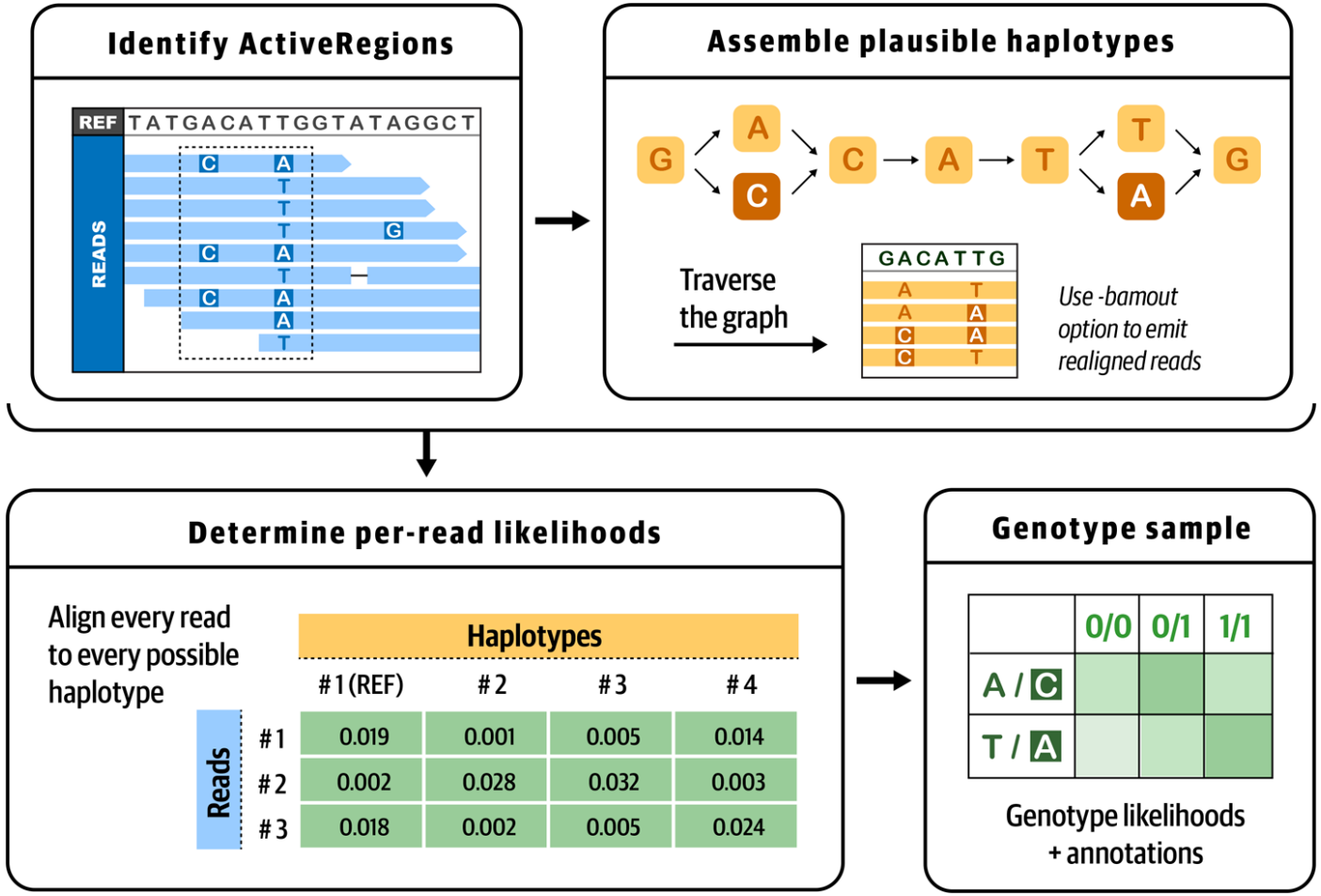
gVCF

Single-sample GVCF containing non-variant records and <NON_REF> symbolic allele

```
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
20 10000204 . A <NON_REF> . . END=10000210 GT:DP:GQ:MIN_DP:PL 0/0:33:84:31:0,84,1260
20 10000211 . C T,<NON_REF> 326.77 . BaseQRankSum=2.340;ClippingRankSum=-1.162;DP=35;MLEAC=1,0;
MLEAF=0.500,0.00;MQ=60.00;MQRankSum=0.623;ReadPosRankSum=0.152
GT:AD:DP:GQ:PL:SB0/1:21,14,0:35:99:355,0,526,418,568,986:12,9,7,7
20 10000212 . A <NON_REF> . . END=10000216 GT:DP:GQ:MIN_DP:PL 0/0:35:90:33:0,90,1350
```

gVCF files contains also reference positions, saved as blocks

Germline variant calling – GATK4



1. **Identify ActiveRegions**
   Identify regions were variants are present.
2. **Assemble plausible haplotypes**
   For each region, creates a DeBruijn graph and identifies the possible variation present in the data
3. **Determine per-read likelihoods**
   Each read is aligned to every possible identified haplotype and a score is given.
4. **Genotype sample**
   The likelihood for each genotype is calculated and the most likely genotype is given.