SevenBridges

Variant Calling

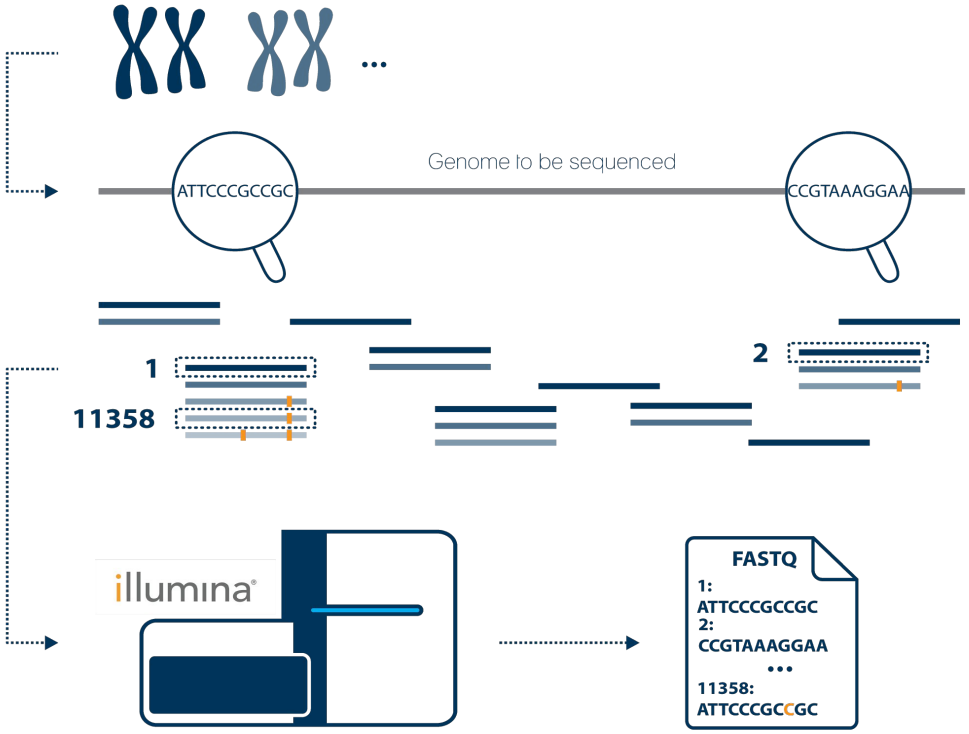
MATF, April 2021.

Ana Đukić

Reminders

Reminder: DNA Sequencing

We got a FASTQ files with the “reads” – little pieces of the genome.

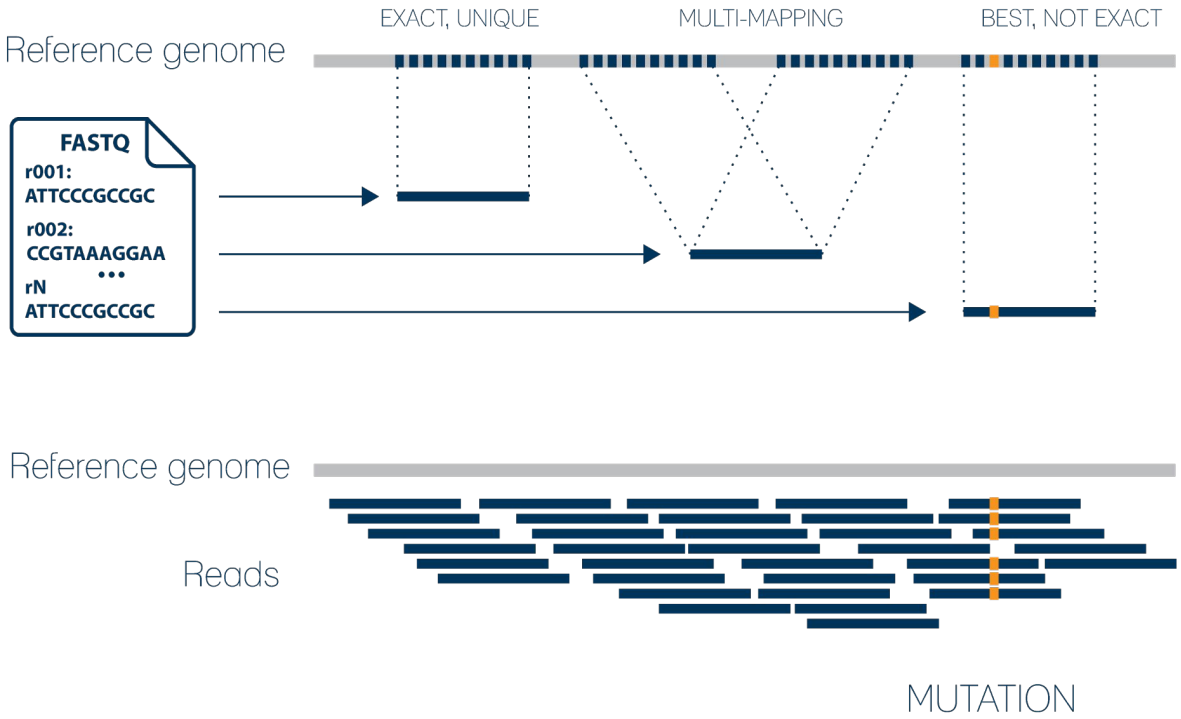


Reminder: DNA Sequencing

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

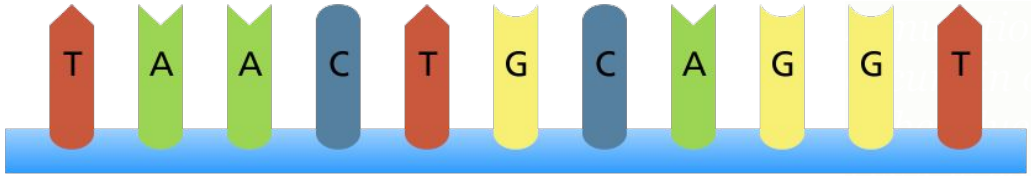
Reminder: Alignment



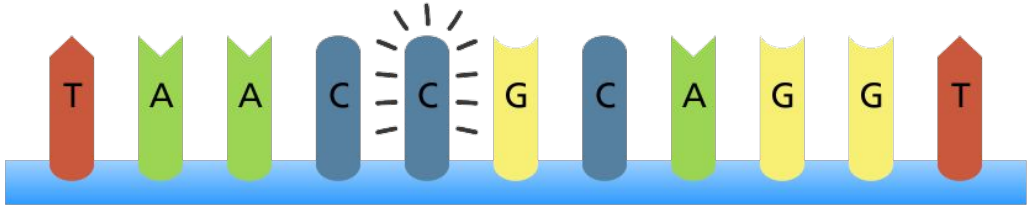
Variants

Variants?

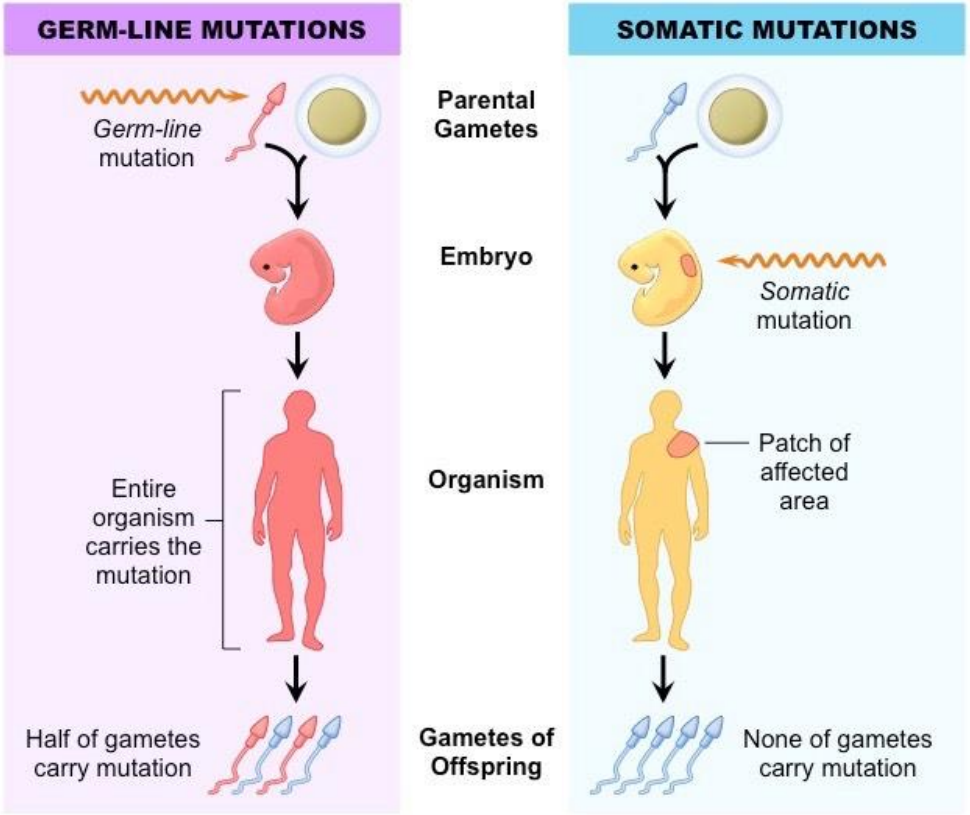
Original sequence



Point mutation



Variants?

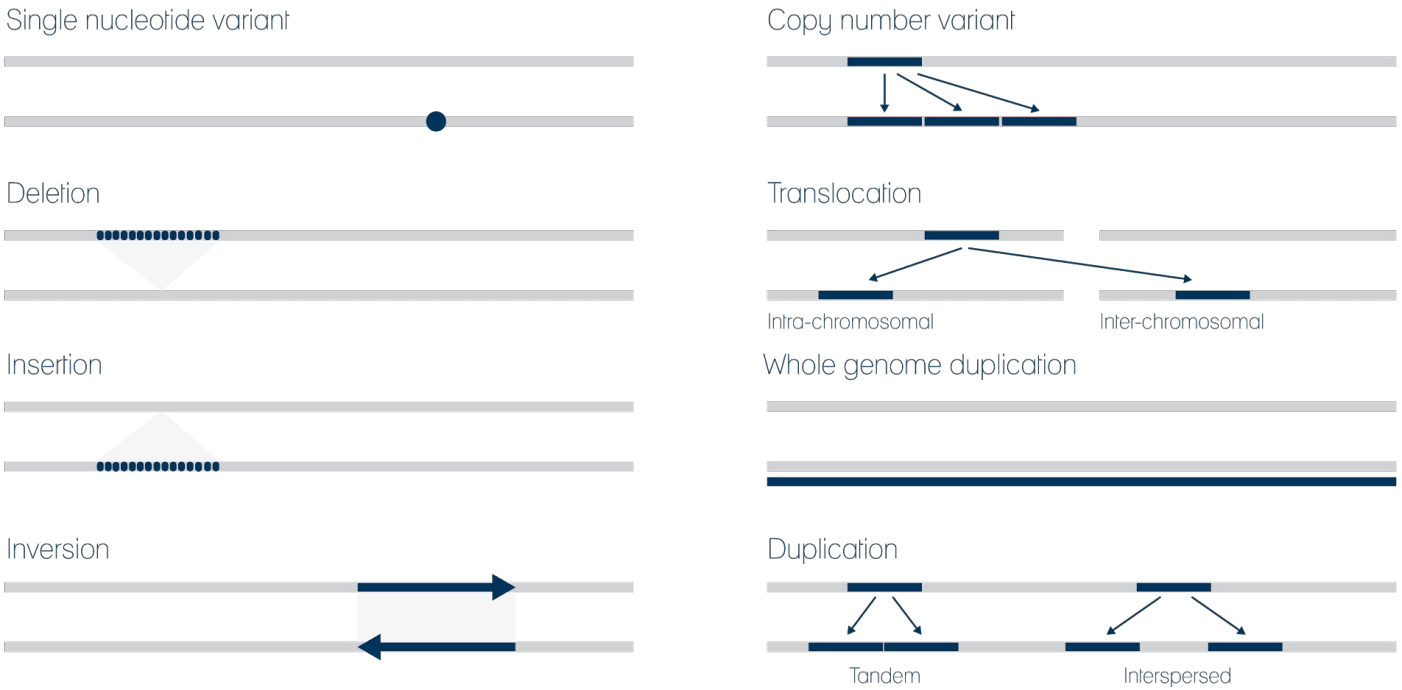

GERM-LINE MUTATIONS
SOMATIC MUTATIONS
Parental Gametes
Embryo
Organism
Gametes of Offspring
Germ-line mutation
Somatic mutation
Entire organism carries the mutation
Patch of affected area
Half of gametes carry mutation
None of gametes carry mutation

Variant calling

Introduction to Variant calling

- Variant calling is the process of finding differences between reference genome and observed sample
- We need aligned reads to the reference genome so we can find – “call” variants
- Different types of genomic variants

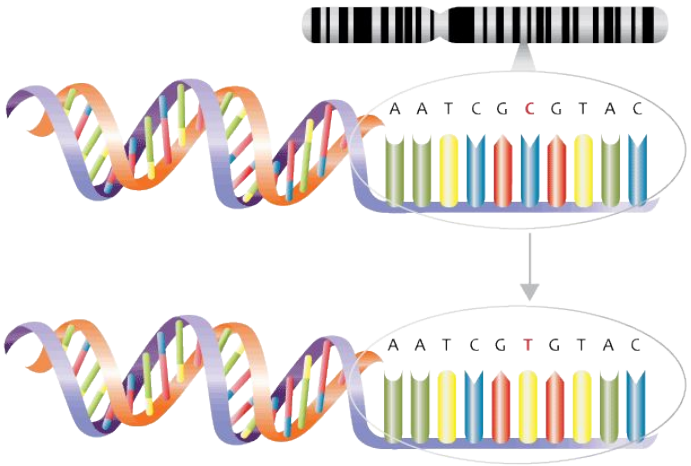
Genomic Variants



Genomic Variants

- SNV (Single Nucleotide Variant)

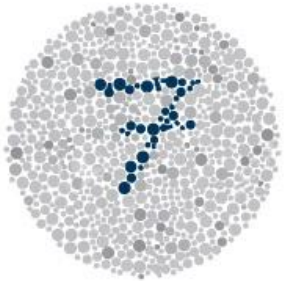
  Simple ones - not a big change on the first look, but...



Genomic Variants

Each of those characteristics causes one SNV

LONGER EYELASHES



DALTONISM

LESS SLEEPING

SUPER STRENGTH

Genomic Variants

Breast Cancer

BRCA2 gene (TS)

SNV id : rs1799954

Chromosome 13
Position 32,340,455

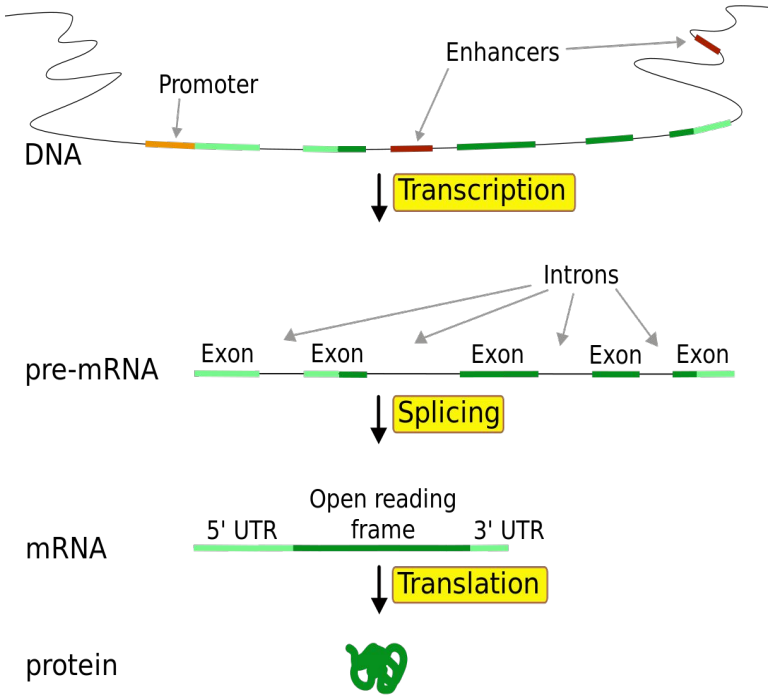
Cancer genotypes: **CC, CT and TT**

http://www.eupedia.com/genetics/cancer_related_snp.shtml
https://www.snpedia.com/index.php/Rs1799954

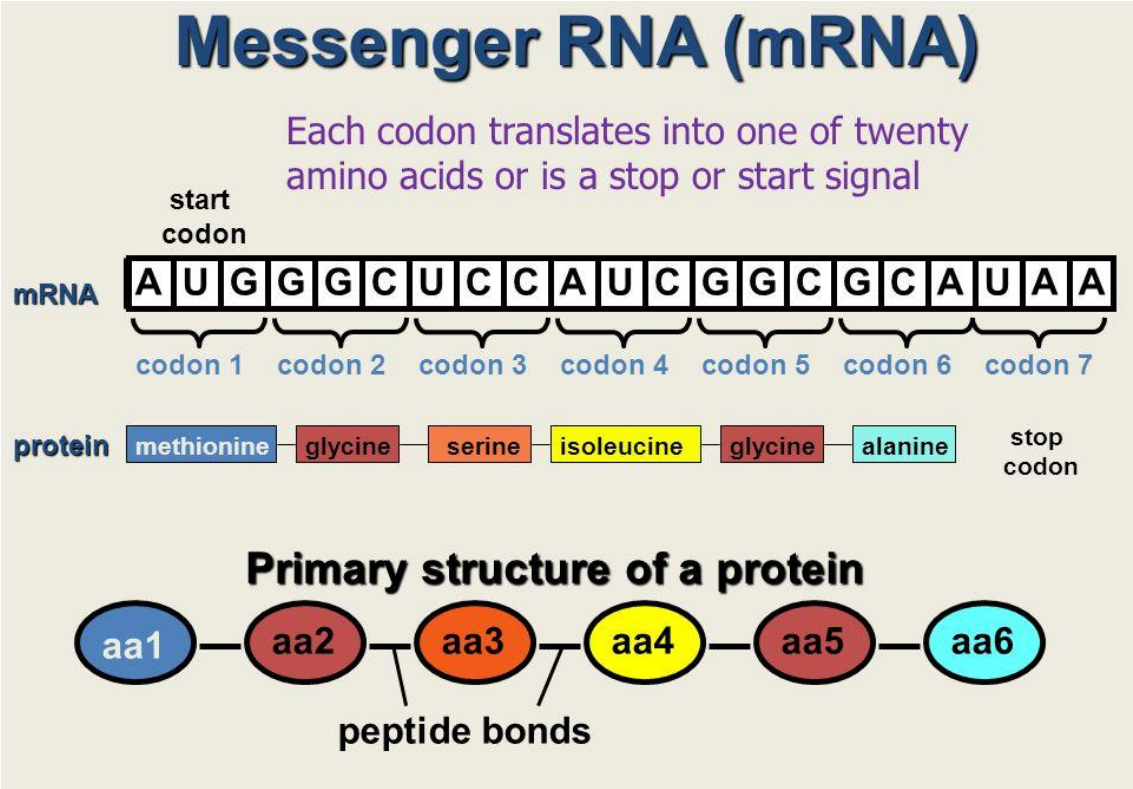
Genomic Variants

Based on the variant location, we can predict if mutation will have impact.

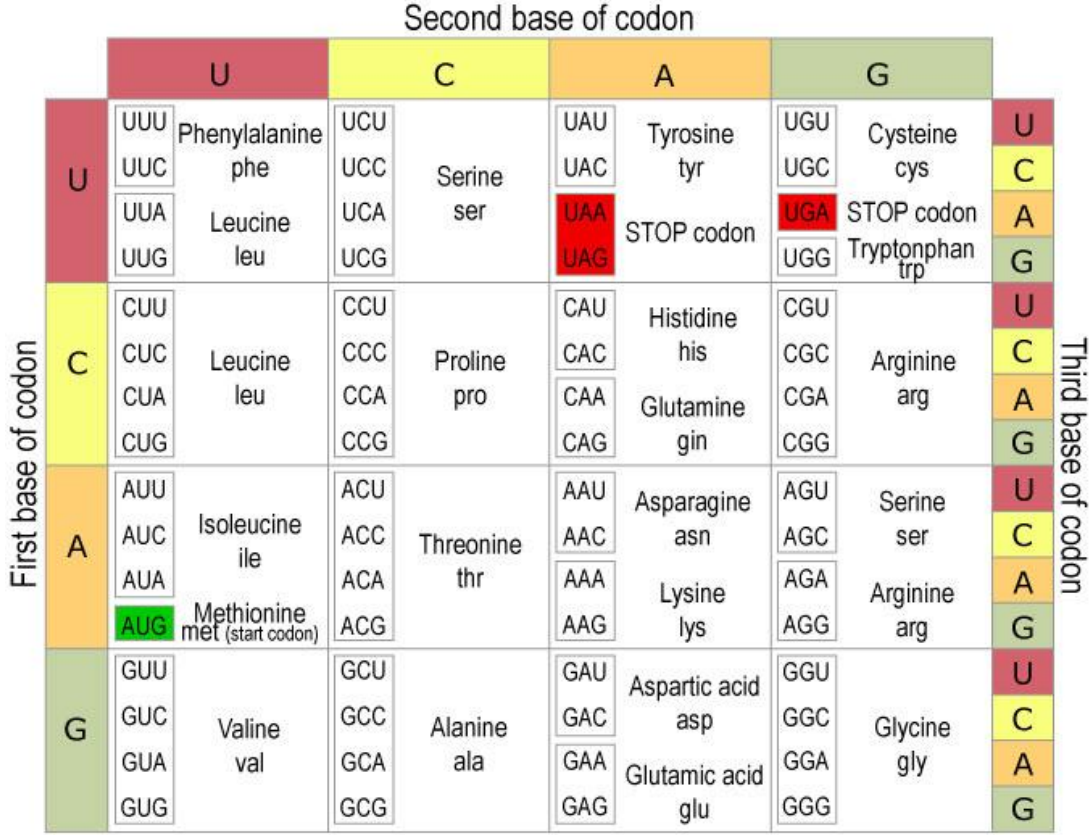
- Central dogma



RNA to Protein



Codon Table


Second base of codon
U
C
A
G
First base of codon
Third base of codon
U
UUU Phenylalanine phe
UUC
UUA Leucine leu
UUG
UCU Serine ser
UCC
UCA
UCG
UAU Tyrosine tyr
UAC
UAA STOP codon
UAG
UGU Cysteine cys
UGC
UGA STOP codon
UGG Tryptonphan trp
C
CUU Leucine leu
CUC
CUA
CUG
CCU Proline pro
CCC
CCA
CCG
CAU Histidine his
CAC
CAA Glutamine gin
CAG
CGU Arginine arg
CGC
CGA
CGG
A
AUU Isoleucine ile
AUC
AUA
AUG Methionine met (start codon)
ACU Threonine thr
ACC
ACA
ACG
AAU Asparagine asn
AAC
AAA Lysine lys
AAG
AGU Serine ser
AGC
AGA Arginine arg
AGG
G
GUU Valine val
GUC
GUA
GUG
GCU Alanine ala
GCC
GCA
GCG
GAU Aspartic acid asp
GAC
GAA Glutamic acid glu
GAG
GGU Glycine gly
GGC
GGA
GGG

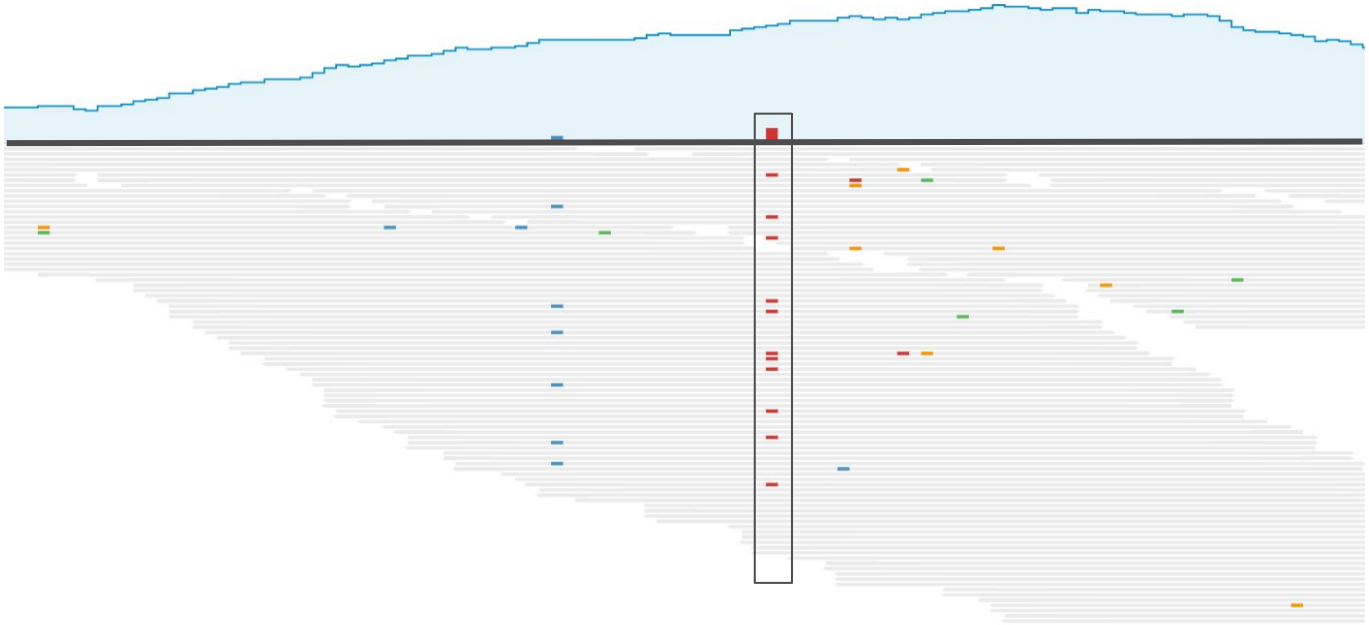
Genomic Variants

Variants can have different impact on human cells and organism

- Single Nucleotide Variants(**SNV**):
  - Harmless
    - **Silent** (Synonymous) – Usually no effect
  - Harmful:
    - **Missense** – Amino acid change
    - **Nonsense**(Start/Stop Gain/Lost) – AUG / UAG, UAA, UGA
  - Depends on the location
    - **Coding** regions
    - **Noncoding** regions
- Insertions/Deletions – **INDELS**
  - **In frame**
  - **Out of frame (Frameshift)**

Basic concepts of variant calling

What is the pileup?

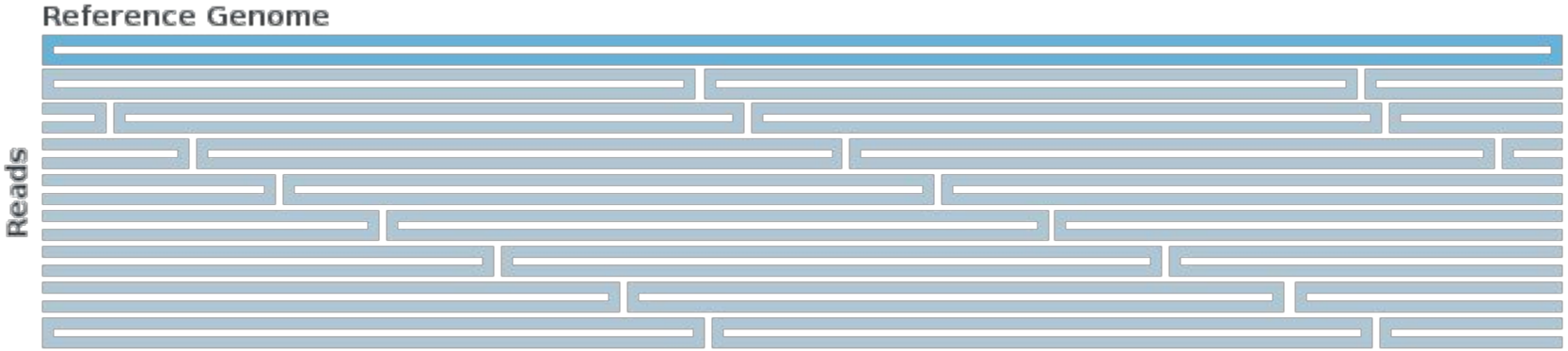


Terminology of Variant Calling

- Checking if all nucleotides in a pileup support the reference genome
- Reference supporting reads – REF
- Variant (Alternative) supporting reads – ALT
- Depth/Coverage = REF + ALT (number of reads covering that position)
- Variant Allele Frequency = ALT / (REF + ALT)
  - Coverage 30 - 20 REF reads, 10 ALT reads
    - VAF = 0,33 or 33%
- Genotypes - human genome is diploid
  - 0/0 = Both alleles match the reference (homozygous)
  - 0/1 = One allele matches reference and one does not (heterozygous)
  - 1/1 = Both alleles do not match reference (homozygous)
  - 1/2 = One allele contains one variant and the other another one (heterozygous)

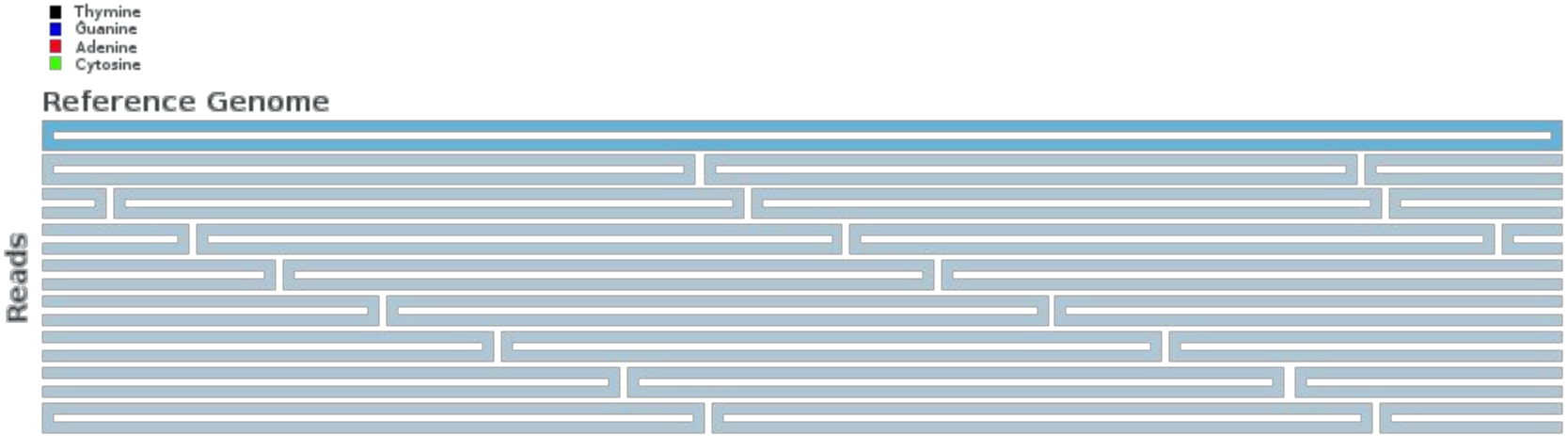
Ideal Variant Calling

Ideal Variant Calling

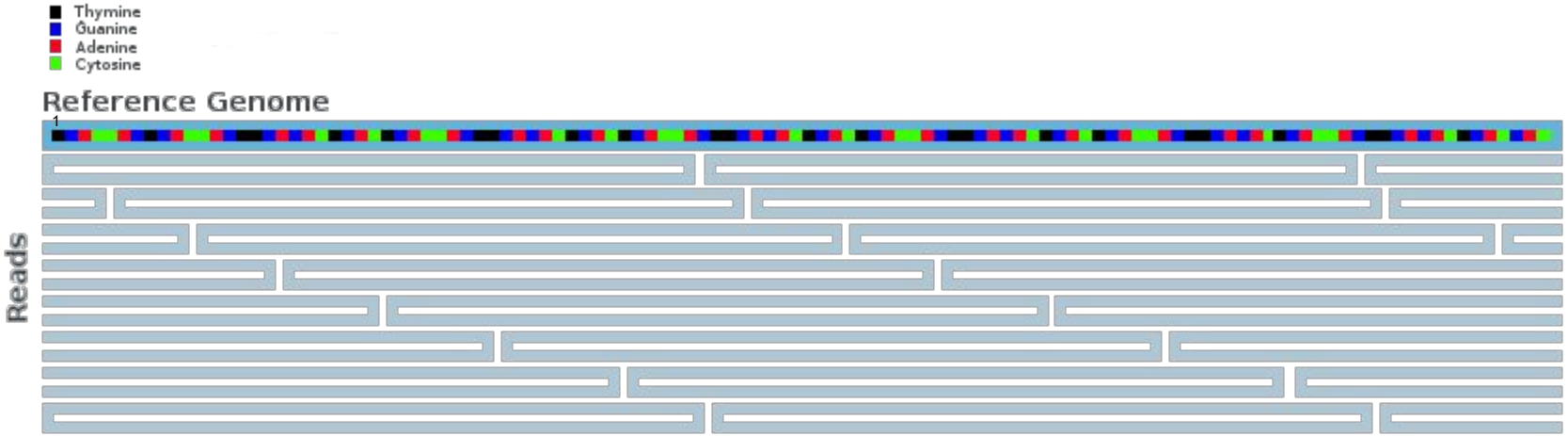


Ideally we will have uniform distribution of reads.

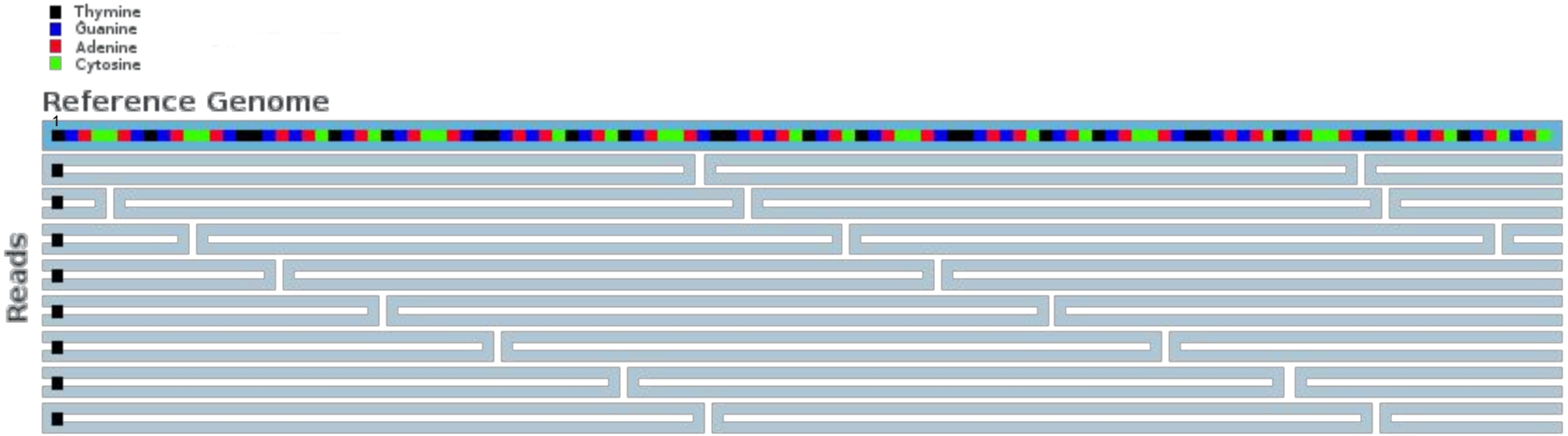
Ideal Variant Calling



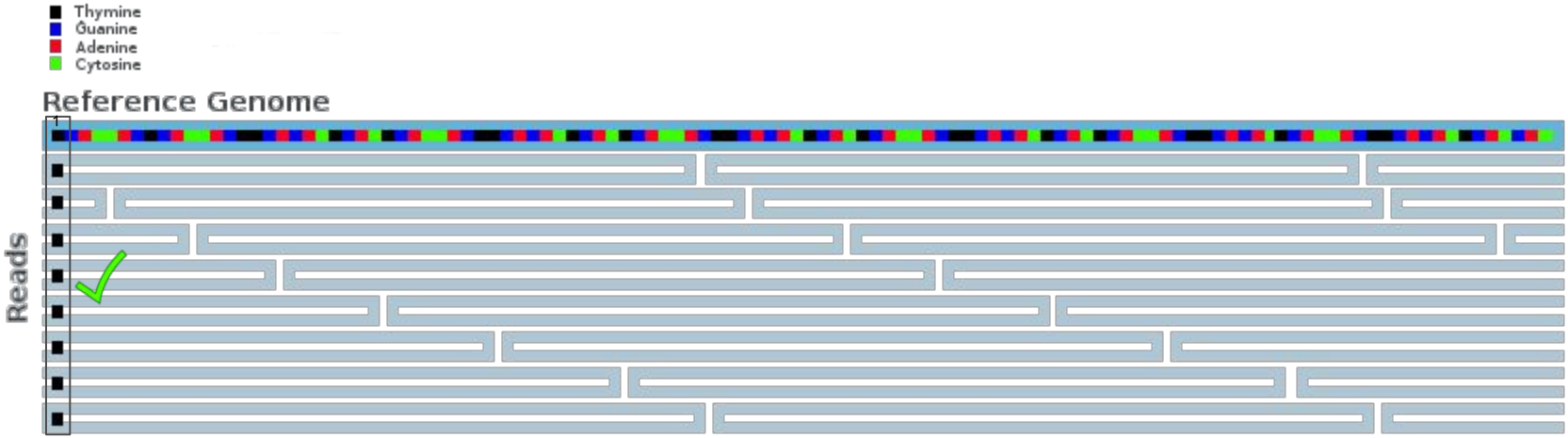
Ideal Variant Calling

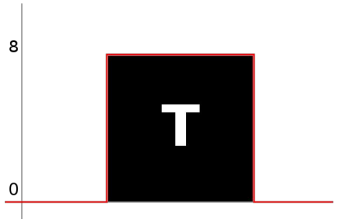


Ideal Variant Calling



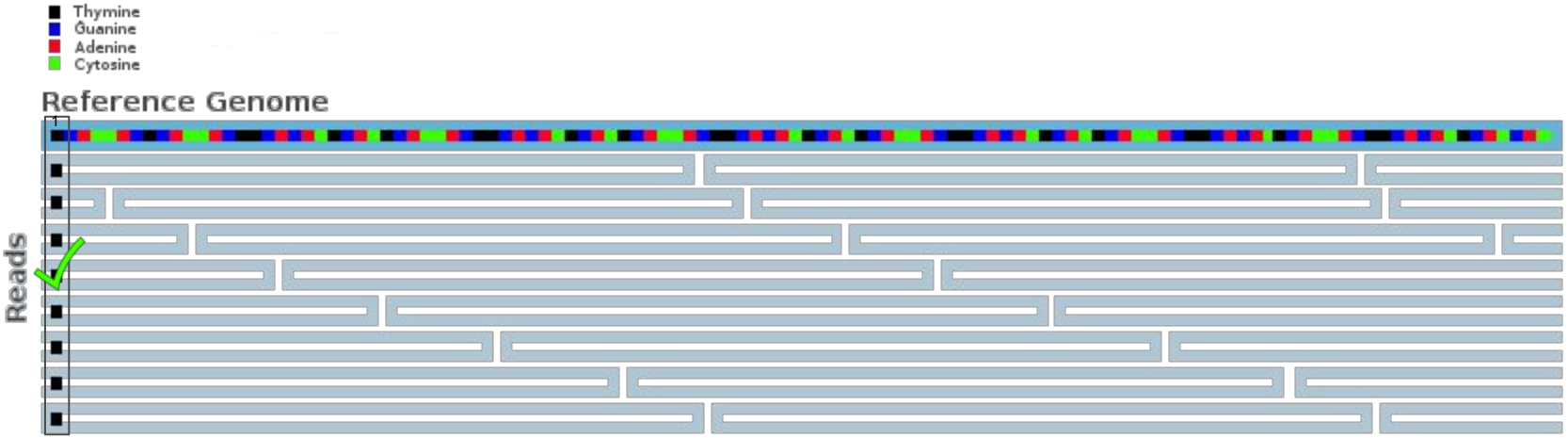
Ideal Variant Calling

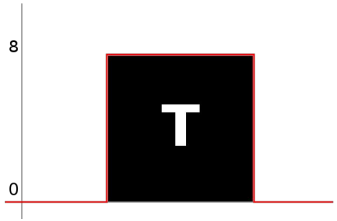




- We have “T” in the all reads covering that position

Ideal Variant Calling

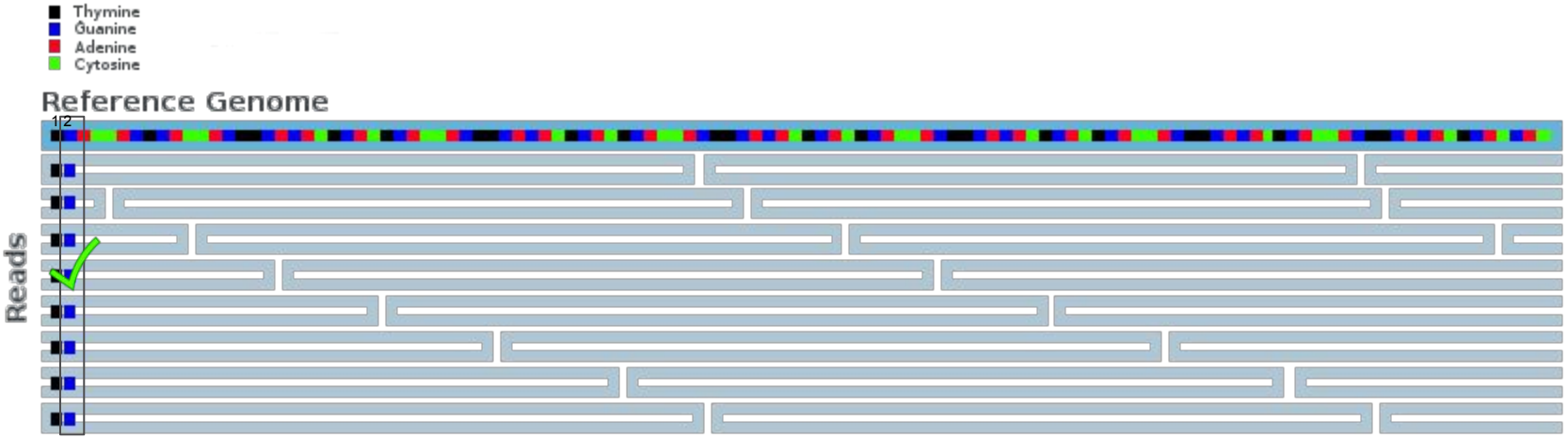


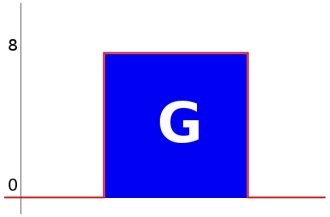


How can we represent what we have observed?

CONTIG	POS	REF	ALT	GT
X	1	T	-	0/0

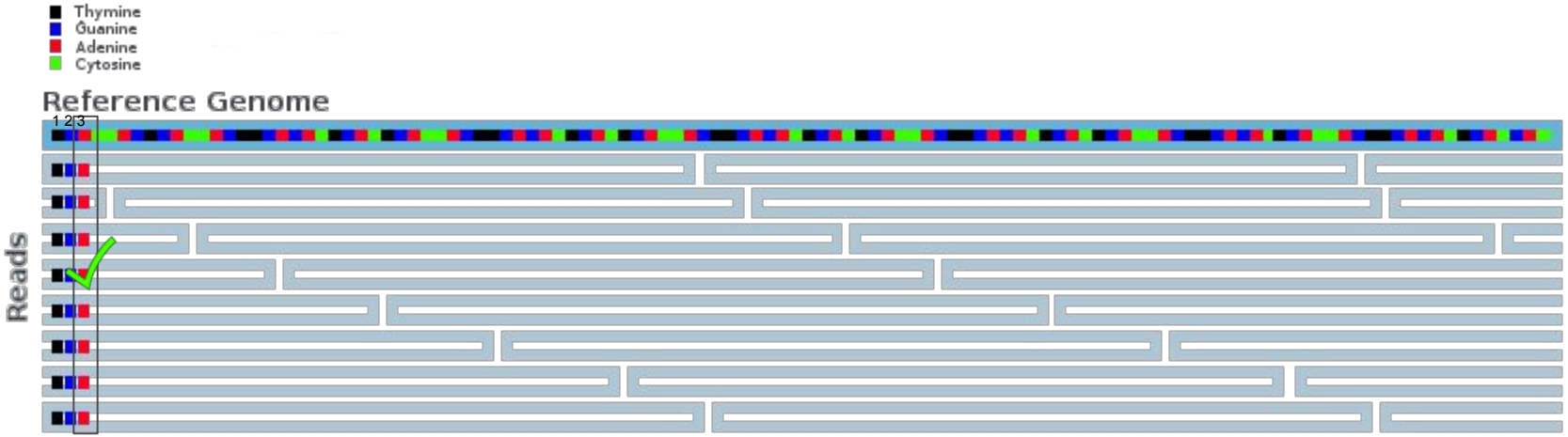
Ideal Variant Calling

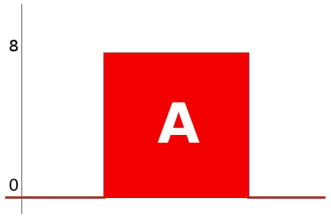




CONTIG	POS	REF	ALT	GT
X	1	T	-	0/0
X	2	G	-	0/0

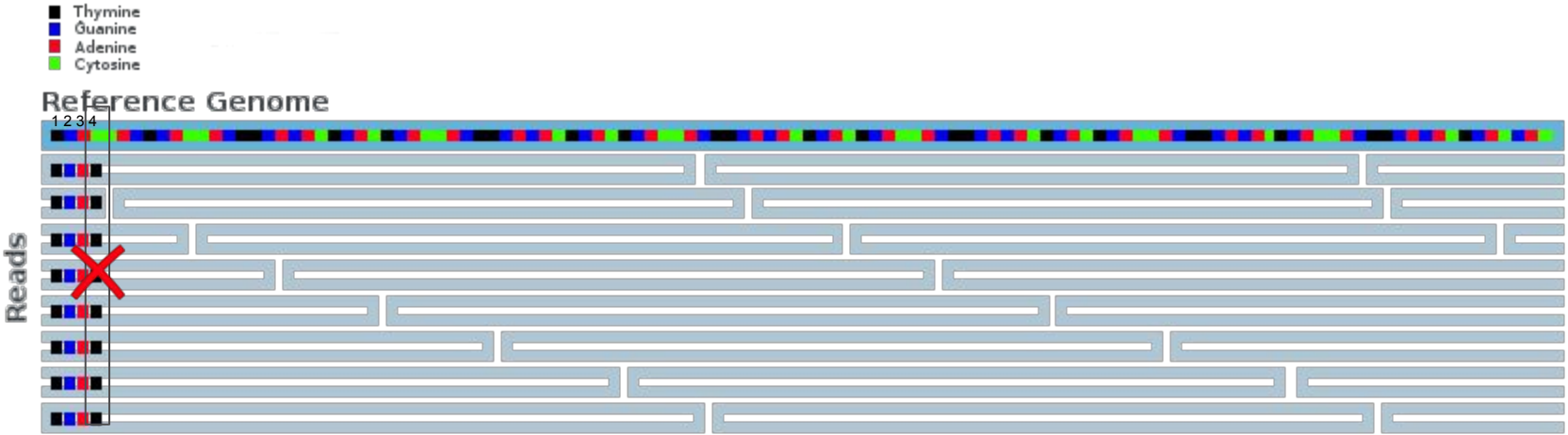
Ideal Variant Calling

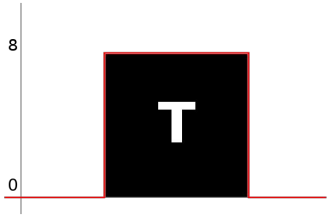




CONTIG	POS	REF	ALT	GT
X	1	T	-	0/0
X	2	G	-	0/0
X	3	A	-	0/0

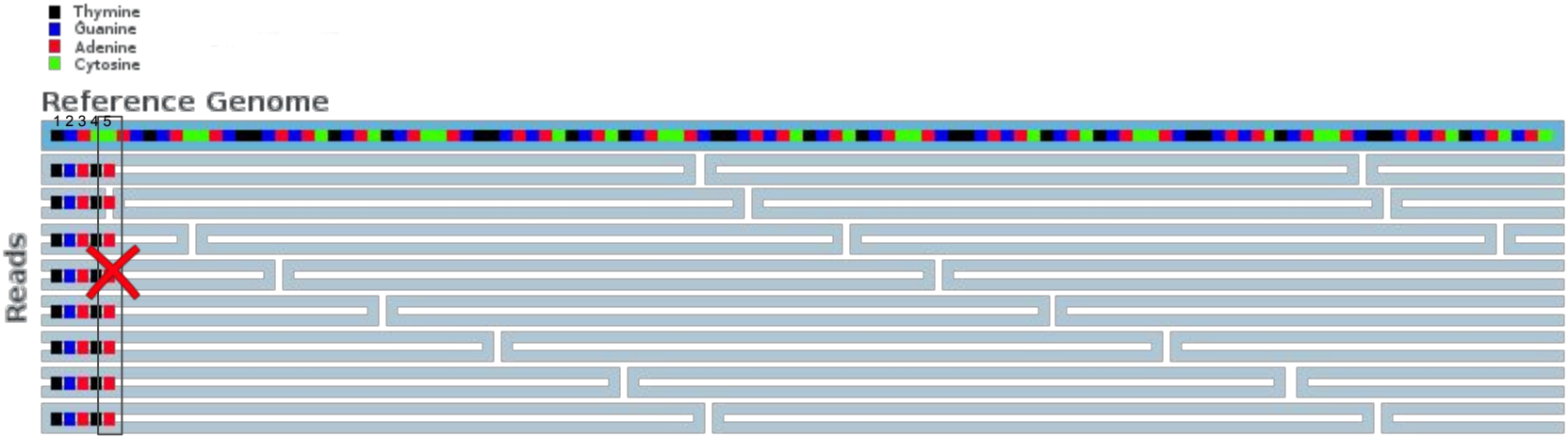
Ideal Variant Calling

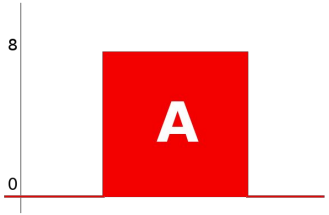




CONTIG	POS	REF	ALT	GT
X	2	G	-	0/0
X	3	A	-	0/0
X	4	**C**	**T**	**1/1**

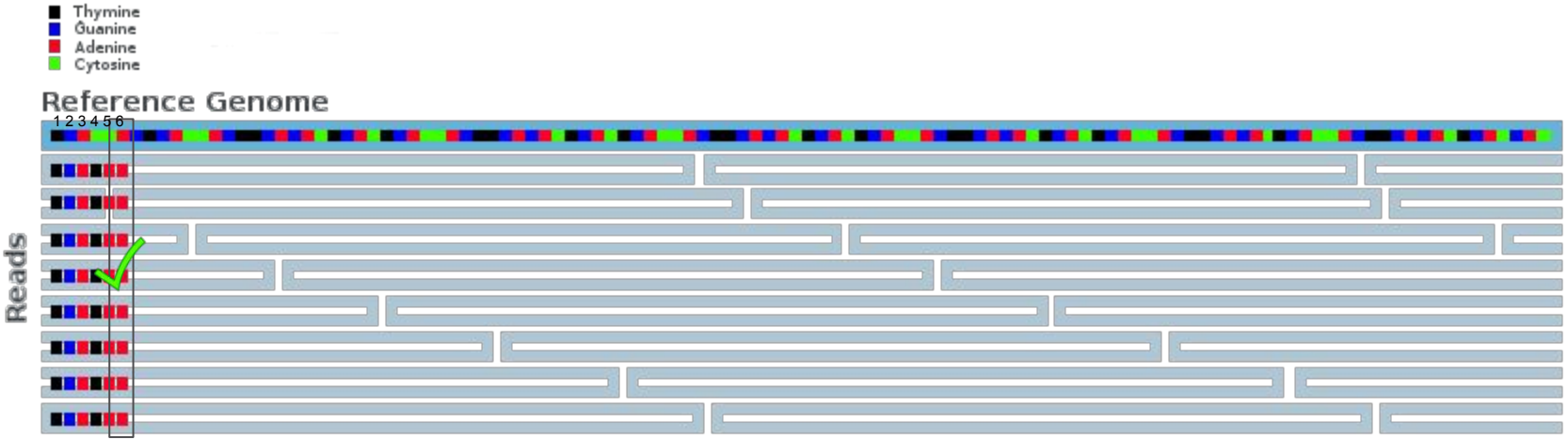
Ideal Variant Calling

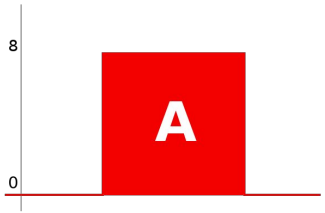




CONTIG	POS	REF	ALT	GT
X	3	A	-	0/0
X	4	**C**	**T**	**1/1**
X	5	**C**	**A**	**1/1**

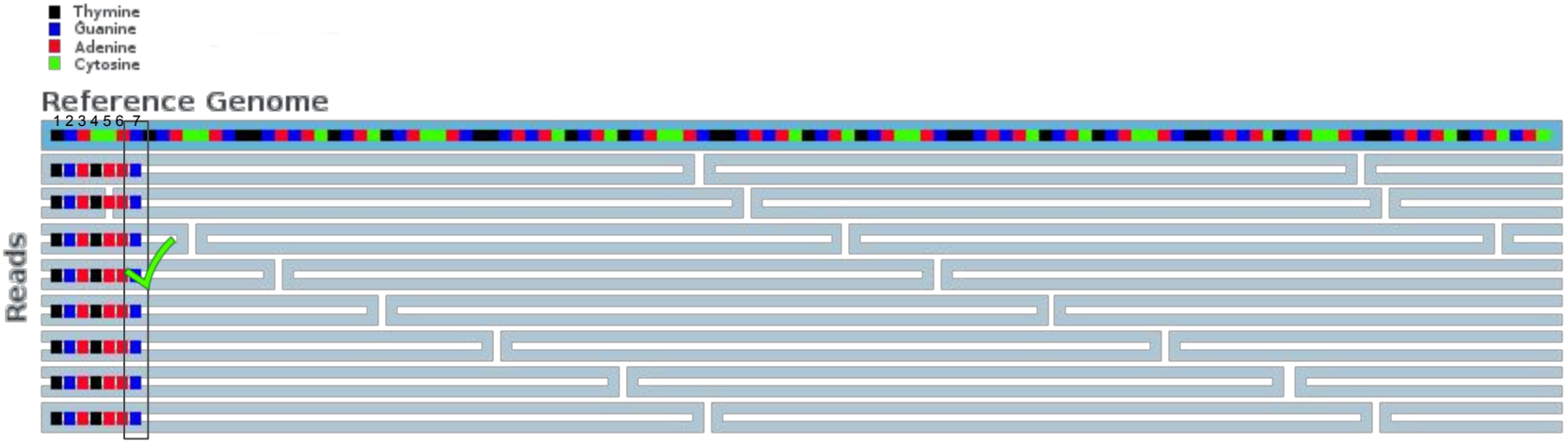
Ideal Variant Calling

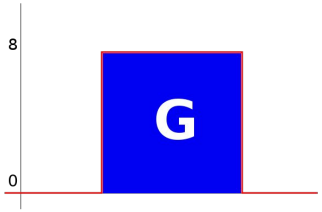




CONTIG	POS	REF	ALT	GT
X	4	**C**	**T**	**1/1**
X	5	**C**	**A**	**1/1**
X	6	A	-	0/0

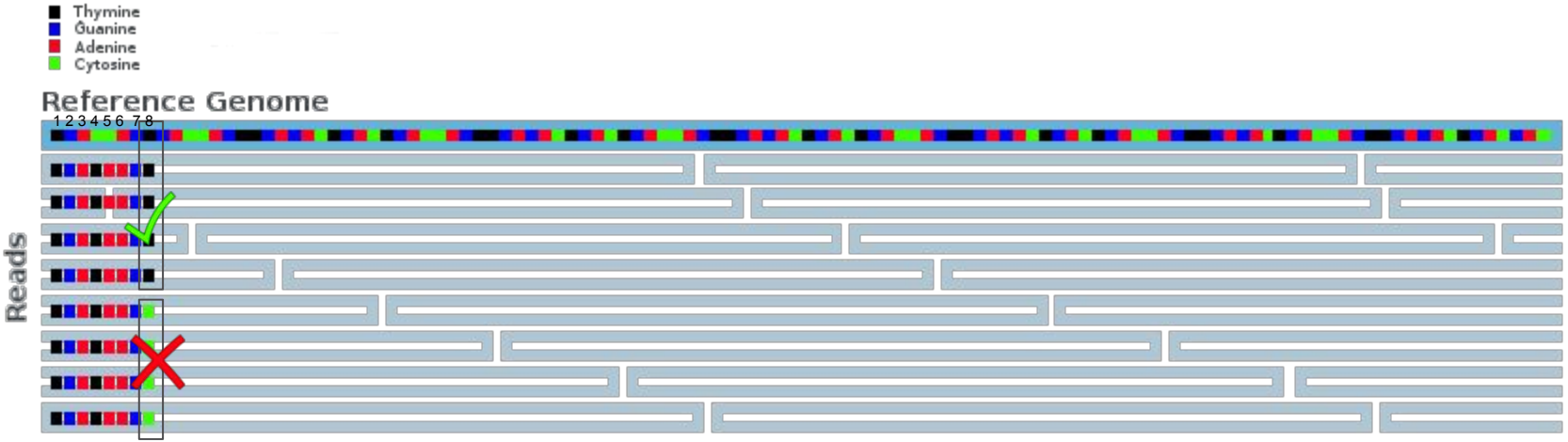
Ideal Variant Calling

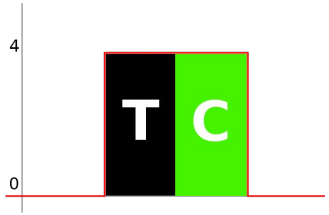




CONTIG	POS	REF	ALT	GT
X	5	**C**	**A**	**1/1**
X	6	A	-	0/0
X	7	G	-	0/0

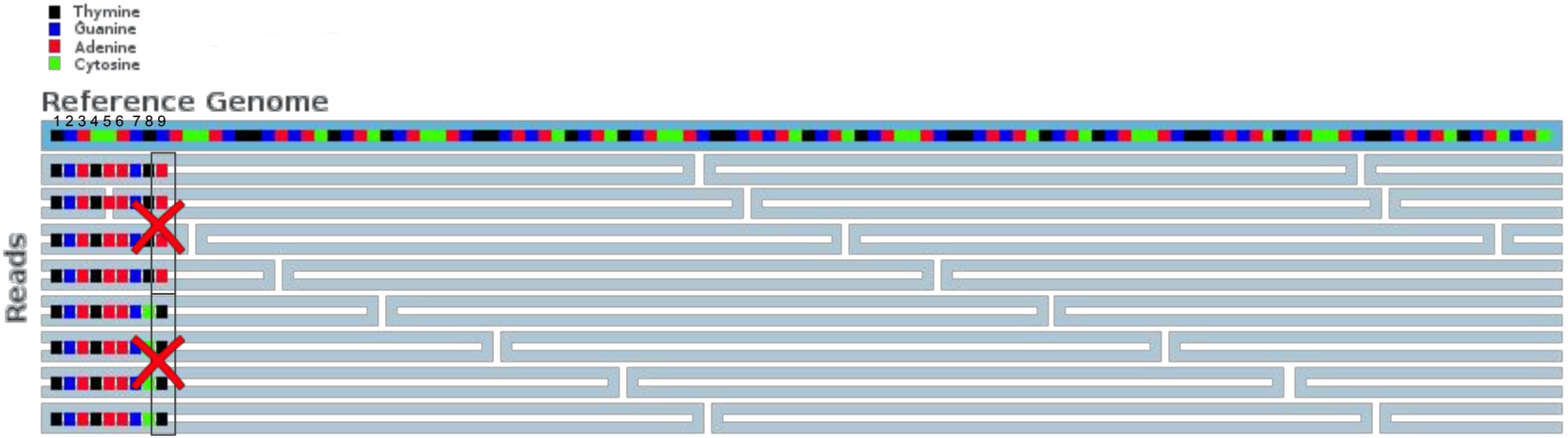
Ideal Variant Calling

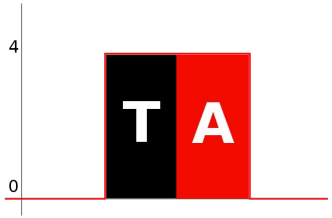




CONTIG	POS	REF	ALT	GT
X	6	A	-	0/0
X	7	G	-	0/0
X	8	**T**	**C**	**0/1**

Ideal Variant Calling



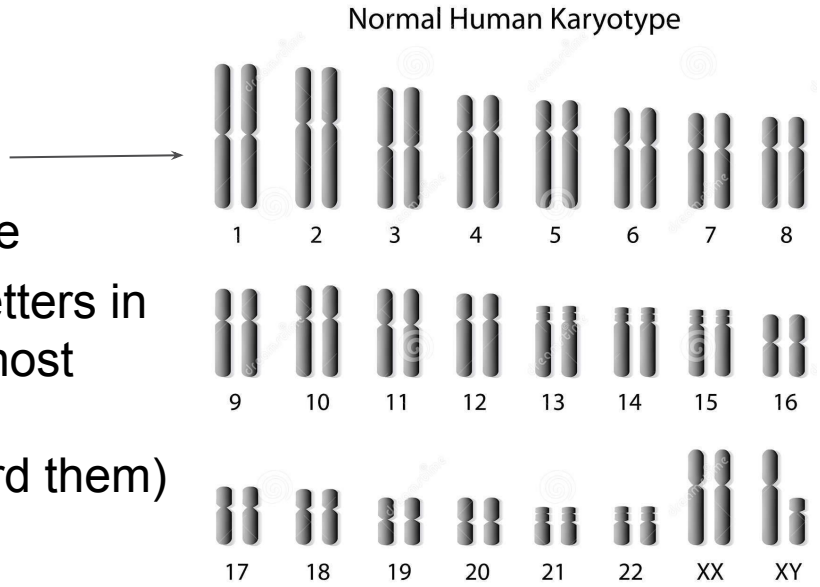


CONTIG	POS	REF	ALT	GT
X	7	G	-	0/0
X	8	**T**	**C**	**0/1**
X	9	**G**	**A,T**	**1/2**

Variant Calling

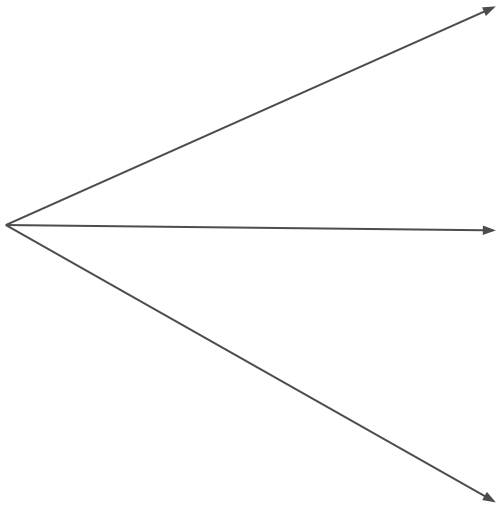
- Two possible cases:
  1. All of the bases in pileup are the same nucleotide [A,T,C,G]
  2. Different nucleotides exist in the pileup

- In the simplest case, assume diploidy
  - There can be only two alleles at a site
  - If there are more than two different letters in the pileup we will only consider the most common two
    (assume others are errors and discard them)



Genomic Variants

DNA Sequencing Data



Whole Genome Data

- ~30x coverage
- Novel variants

Whole Exome Data

- ~100x coverage
- Coding variants

Targeted Sequencing Data

- ~1000x coverage
- Gene variants

Variant Calling

- When all of the bases in pileup are the same nucleotide [A,T,C,G]
  - All bases are the same and **match the reference**
    - Consider the site to be homozygous reference
  - All bases are the same and **do not match the reference**
    - Consider the site to be homozygous variant
    - But what if the pileup contains only one or two bases?
    - Probably an error, but still make the call and leave it to filtering
  - Making the call looks fairly simple

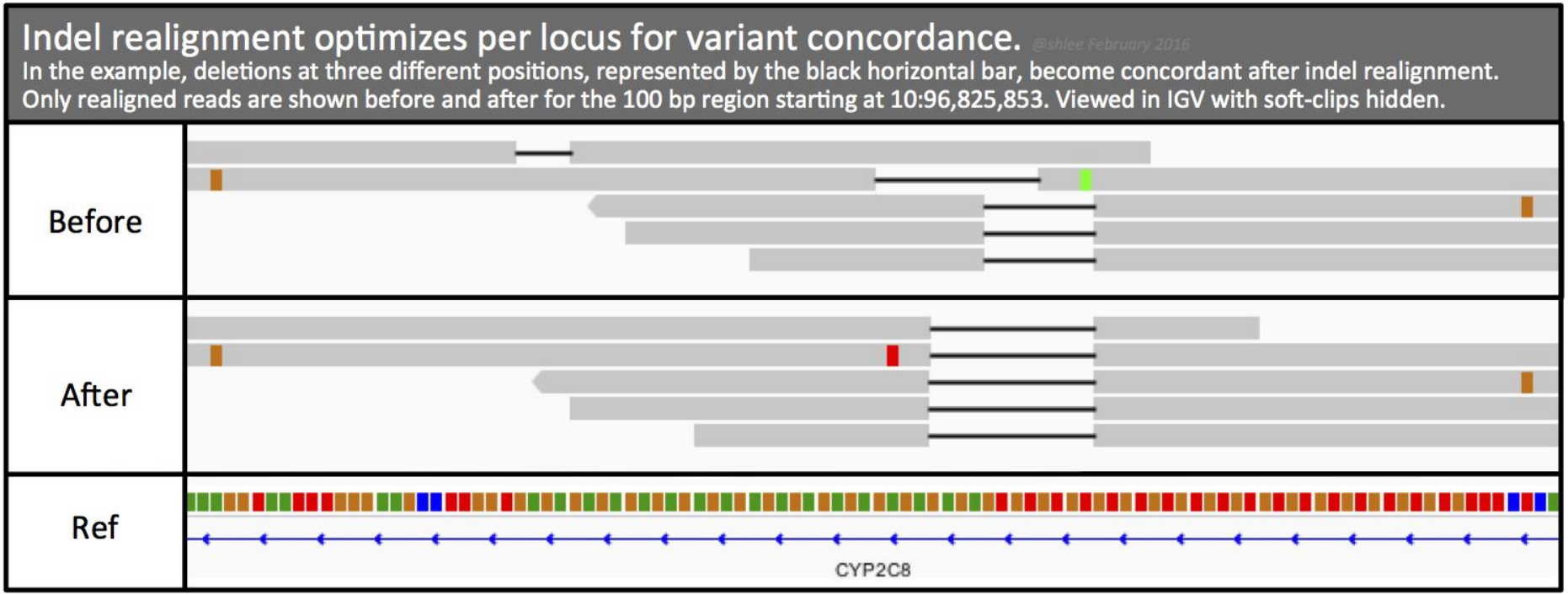
Variant Calling

- When different nucleotides exist in the pileup
  - If we have 15 “A” and 15 “T”, it’s a heterozygote!
  - If we have 29 “A” and 1 “T”, the “T” is probably an error!
  - What about 5 “T”? Or 7?
    - Where is the threshold?
    - What happens with more or less than 30 bases?

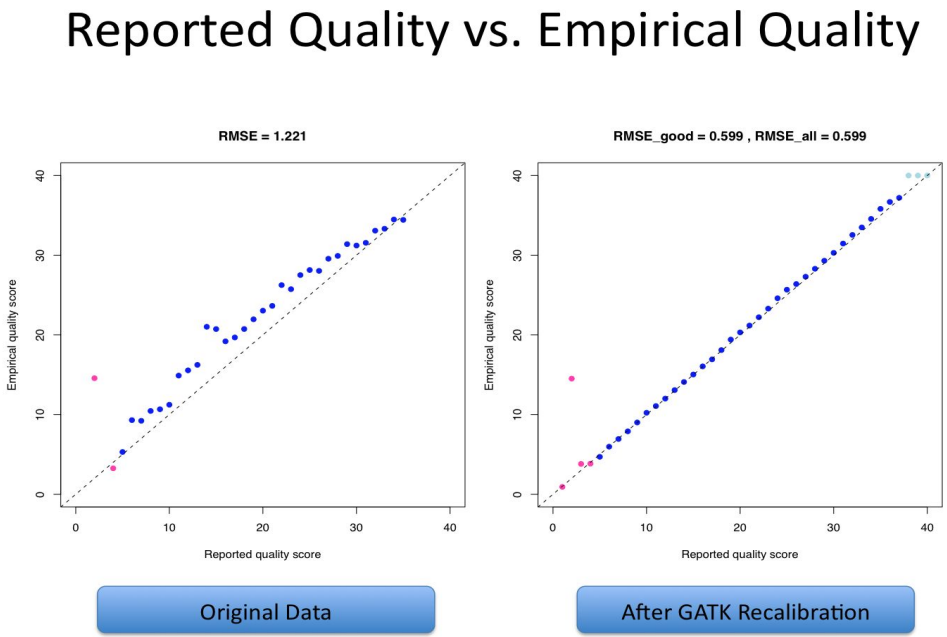
Variant Calling

- Improve calling with pre-processing steps:
    - Mark duplicate reads which came from same DNA fragment
  - Indel Realignment
    - Realign near insertions and deletions
  - Base quality Recalibration
    - Recalibrate quality of bases which sequencer outputs

Variant Calling



Variant Calling



Let's say the machine reads an A nucleotide, and assigns a quality score of Q20 - in Phred-scale, that means it's 99% sure it identified the base correctly.

- One wrong base out of 100
- If we sequence 90 billion bases we get 900 million wrong called bases!

Variant Calling

- So, when we have two letters in the pileup, what should we call?
  - Let's call the two "letters" **b** and **b'** *(b, b' ∈ [A, C, T, G])*
  - Let **n** be the total number of bases, and **k** number of b' bases
  - Three possible explanations for the pileup:
    - Genotype is bb; k bases are errors, n-k are correct
    - Genotype is b'b'; n-k bases are errors, k are correct
    - Genotype is bb'; all n bases are correct
  - Now we need to find the probabilities of these three cases
    - Will pick the most probable one!

Variant Calling – advance

- We assumed a flat error rate
  - But we have Base qualities from the sequencer
  - Machine-specific error profiles
- We can look at mapping qualities
  - Mapping errors are a big source of errors
- We can look at haplotypes
  - Errors don’t segregate nicely

Variant Calling Results

Variant calling results

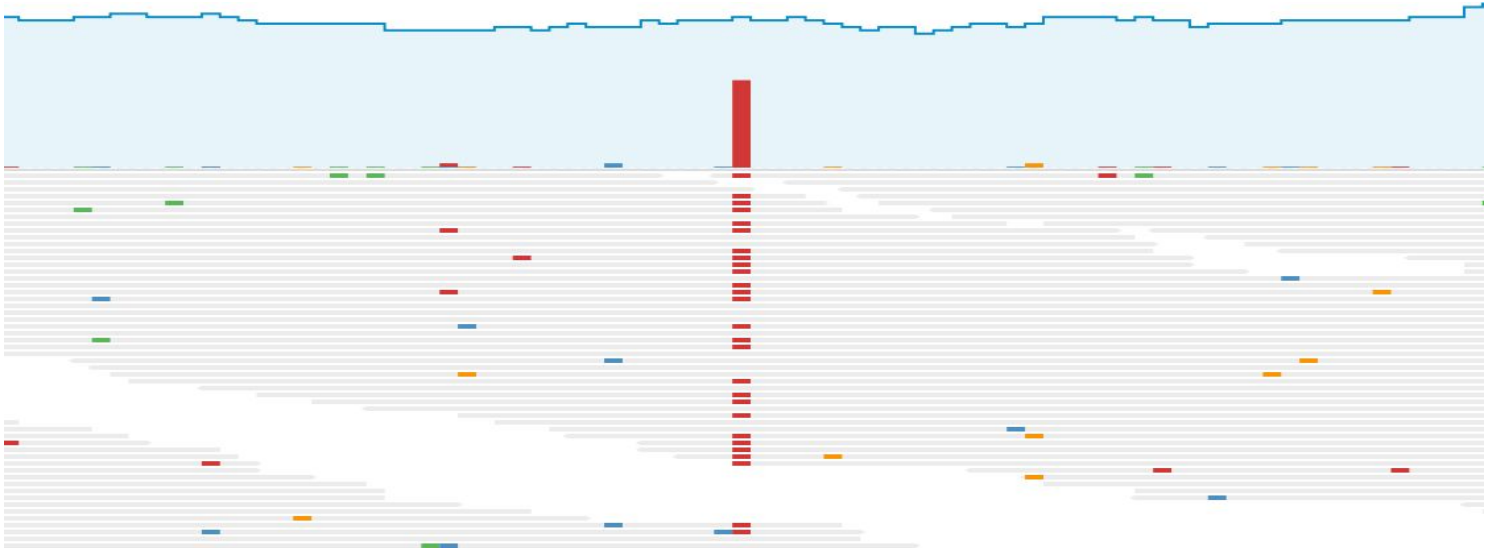
- The result of Variant Calling is a file in VCF format, which contains mutations
- A plain text file format for storing variant data
- A number of line starting with ## -the header
- Main header line:
  #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1
- This is followed by the actual variant data, one entry per line
  22 10001 . A C 40 PASS DP=14 GT 0/1
- More than one sample can be in one line
- For details: VCF specification

Variant calling results

- Example of VCF format
- Each row represents one mutation

CHR	POS	REF	ALT	FORMAT	NA12878
1	14300	A	G	GT, VAF	0/1, 0.4
2	15367	A	C	GT, VAF	1/1, 0.9
3	25612	C	G,A	GT, VAF	1/2, ?
5	5632	TA	T	GT, VAF	0/1, 0.5
7	7824	T	TA	GT, VAF	1/1, 0.8

Variant calling results – check out BAM file

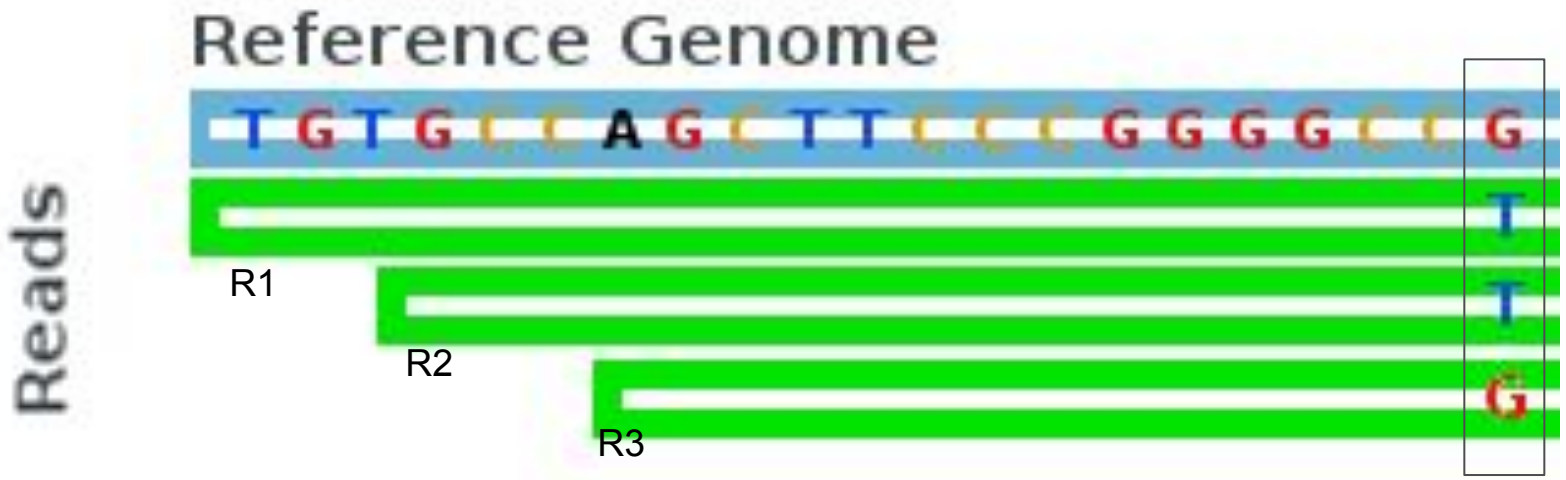


CHR	POS	REF	ALT	FORMAT	NA12877
1	14125	T	C	GT, VAF	0/1, 0.6

ana.djukic@sevenbridges.com
milan.kovacevic@sevenbridges.com

Example how to determine genotype

STEP 1



STEP 2

- Calculate the probability of every possible base b∈ {A, C, G, T} given the observed allele A.
- Observed alleles in our example at given locus are A = {T, G}.

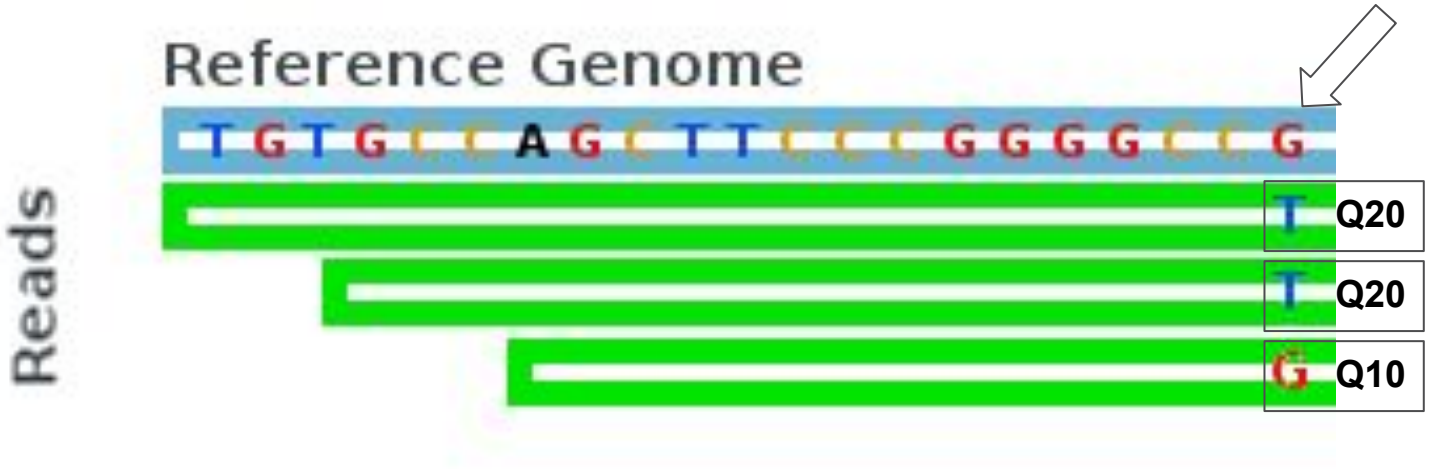
Verovatnoca da dodje do greske prilikom ocitavanja baze tokom sekvenciranja. Primer: ako je Q20 to znaci da je -10*log10(e)=20, iz cega sledi da je e=0.01

base

Allele

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases},$$

STEP 2 (Example)



STEP 2 (Example)

A	T	p(b=A| A=T)=
C	T	p(b=C|A=T)=
G	T	p(b=G|A=T)=
T	T	**p(b=T|A=T)= 0.99**

$$\sum p(b|A) = 1$$

p(A|T) + p(C|T) + p(G|T) + p(T|T) = 1

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases}$$

Since base quality is Q20, e = 0.01

p(b=T|A=T) = **1 - 0.01**
= **0.99**

STEP 2 (Example)

A	T	p(b=A| A=T)=
C	T	p(b=C|A=T)=
G	T	p(b=G|A=T)=
T	T	**p(b=T|A=T)= 0.99**

$$\sum p(b|A) = 1$$

p(A|T) + p(C|T) + p(G|T) + p(T|T) = 1

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases},$$

T for R1 ----- **Q20**

Q20 = 0.01

p(b=A|A=T) = **(0.01)/3**
= **0.003**

p(b=C|A=T) = **(0.01)/3**
= **0.003**

p(b=G|A=T) = **(0.01)/3**
= **0.003**

STEP 2 (Example)

A	T	p(b=A| A=T)= 0.003
C	T	p(b=C|A=T)= 0.003
G	T	p(b=G|A=T)= 0.003
T	T	**p(b=T|A=T)= 0.99**

$$\sum p(b|A) = 1$$

p(A|T) + p(C|T) + p(G|T) + p(T|T) = 1

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases},$$

T for R1 ----- **Q20**

Q20 = 0.01

p(b=A|A=T) = **(0.01)/3**
= **0.003**

p(b=C|A=T) = **(0.01)/3**
= **0.003**

p(b=G|A=T) = **(0.01)/3**
= **0.003**

STEP 2 (Example)

A	G	p(b=A| A=G)=
C	G	p(b=C|A=G)=
G	**G**	**p(b=G|A=G)= 0.9**
T	G	p(b=T|A=G)=

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases},$$

Since base quality is Q10, e = 0.1

p(b=G|A=G) = **(1-0.1)**
= 0.9

STEP 2 (Example)

A	G	p(b=A| A=G)= 0.03
C	G	p(b=C|A=G)= 0.03
G	**G**	**p(b=G|A=G)= 0.9**
T	G	p(b=T|A=G)= 0.03

$$p(b|A) = \begin{cases} \frac{e}{3} : b \neq A \\ 1 - e : b = A \end{cases},$$

T for R3 ----- **Q10**

Q10 = 0.1

p(b=A|A=G) = **(0.1)/3**
= 0.03

p(b=C|A=G) = **(0.1)/3**
= 0.03

p(b=T|A=G) = **(0.1)/3**
= 0.03

STEP 2 (Example)

A	T	p(b=A| A=T)= 0.003
C	T	p(b=C|A=T)= 0.003
G	T	p(b=G|A=T)= 0.003
T	**T**	**p(b=T|A=T)= 0.99**

A	T	p(b=A| A=T)= 0.003
C	T	p(b=C|A=T)= 0.003
G	T	p(b=G|A=T)= 0.003
T	**T**	**p(b=T|A=T)= 0.99**

A	G	p(b=A| A=G)= 0.03
C	G	p(b=C|A=G)= 0.03
G	**G**	**p(b=G|A=G)= 0.9**
T	G	p(b=T|A=G)= 0.03

STEP 3

3. Calculate the probability

$$p(b|G) = p(b|\{A_1, A_2\}) = \frac{1}{2}p(b|A_1) + \frac{1}{2}p(b|A_2),$$

of each base b given a genotype from the set of plausible genotypes G ∈ {TT, TG, GG} (which are based on the observed bases at this position within the pileup T and G)

STEP 3 (Example)

$$p(b|G) = p(b|\{A_1, A_2\}) = \frac{1}{2}p(b|A_1) + \frac{1}{2}p(b|A_2),$$

T	TT	p(b=T| G=TT) = 0.99/2 + 0.99/2 = 0.99
T	TG	p(b=T|G=TG) = 0.99/2 + 0.003/2 = 0.49
T	GG	p(b=T|G=GG) = 0.003/2 + 0.003/2= 0.003

STEP 3 (Example)

$$p(b|G) = p(b|\{A_1, A_2\}) = \frac{1}{2}p(b|A_1) + \frac{1}{2}p(b|A_2),$$

G	TT	p(b=G| G=TT) = 0.03/2 + 0.03/2 = 0.03
G	TG	p(b=G|G=TG) = 0.03/2 + 0.9/2 = 0.47
G	GG	p(b=G|G=GG) = 0.9/2 + 0.9/2= 0.9

STEP 4

4. Sum over all bases to calculate the probabilities

$$p(D|G) = \prod_{b \in \text{pileup}} p(b|G)$$

the probability of the observed data given all plausible genotypes **G ∈ {TT, TG, GG}**.

STEP 4 (Example)

T	TT	p(b=T| G=TT) = 0.99
T	TG	p(b=T|G=TG) = 0.49
T	GG	p(b=T|G=GG) = 0.003

G	TT	p(b=G| G=TT) = 0.03
G	TG	p(b=G|G=TG) = 0.47
G	GG	p(b=G|G=GG)= 0.9

TT	TT	p(D=TT| G=TT) = 0.99 * 0.99 * 0.03 = **00.29**
TG	TG	p(D=TG|G=TG) = 0.49 * 0.49 * 0.47 = **0.11**
GG	GG	p(D=GG|G=GG) = 0.003 * 0.003 * 0.9 = **8x10^-6**

STEP 5

5. Given P(G) = 1 and P(D) is the sum of all P(D|G) probabilities for each plausible genotype, we can calculate the posterior probability

$$p(G|D) = \frac{p(G)p(D|G)}{p(D)}$$

of each plausible genotype by dividing each P(D|G) by P(D).

STEP 5 (Example)

P(G) = 1

P(D) = 0.029 + 0.11 + 0.0000081 = 0.139

TT	TT	p(D=TT| G=TT) = 0.29/0.139 = 0.21
TG	TG	p(D=TG|G=TG) = 0.11/0.139 = 0.79
GG	GG	p(D=GG|G=GG) = 8x10^-6/0.139 = 5.8 x 10^-5

STEP 5 (Example)

P(G) = 1

P(D) = 0.029 + 0.11 + 0.0000081 = 0.139

TT	TT	p(D=TT| G=TT) = 0.29/0.139 = 0.21
TG	TG	p(D=TG|G=TG) = 0.11/0.139 = 0.79
GG	GG	p(D=GG|G=GG) = 8x10^-6/0.139 = 5.8 x 10^-5

max(TT: 0.21, **TG: 0.79**, GG: 5.8x10^-5**) = 0.79** **TG**

Notebook exercise

PRACTICE - CGC interactive analysis:

CODE MINI VARIANT CALLER

- Login into CGC Platform
- Go to Public Apps, Search for SAMTools MPileup tool, copy the tool to your project
- Go to Data > Public Test Files, Search for files: merged-normal.bam, copy files to your project
- Check metadata of merged-normal.bam file, find the reference genome. Go to Public Reference Files and find the reference FASTA and its FAI, and copy files to your project
- Create and run a task for SAMTools MPileup tool with input BAM and reference files
- Explore app settings of the tool
- Setup Data Cruncher Interactive Analysis in your project
- Test the Notebook with output PILEUP file that was created by previously executed task

The Notebook:

- Call Variants
- Genotype
- Use Qualities
- Outputs VCF

Pileup File Format

- Facilitates SNP/indel calling
- You can call variants “manually”
- Each line consists of 6 tab-separated columns

1. Sequence identifier
2. Position in sequence
3. Reference base at that pos.
4. Depth of coverage
5. Bases at that position
6. [Quality of bases]

```
seq1 272 T 24  ,.$....,,.,.,...,,,.,..^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,.,.,...,,,.,..A <<<;<<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$....,,.,.,...,,,.,...   7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$....,,.,.,...,,,.,...^l.  <+;9*<<<<<<<<<=<<:;<<<<
seq1 276 G 22  ...T,,.,.,...,,,.,....  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ....,,.,.,.C.,,,.,..G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  ....,,.,.,...,,,.,....^k.  %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,.,.,...,,,.,..... ;75&<<<<<<<<<<=<<<9<<:<<<
```

Pileup File Format

Column 5: The bases string [edit]

- . (dot) means a base that matched the reference on the forward strand
- , (comma) means a base that matched the reference on the reverse strand
- </> (less-/greater-than sign) denotes a reference skip. This occurs, for example, if a base in the reference genome is intronic and a read maps to two flanking exons. If quality scores are given in a sixth column, they refer to the quality of the read and not the specific base.
- AGTCN denotes a base that did not match the reference on the forward strand
- agtcn denotes a base that did not match the reference on the reverse strand
- A sequence matching the regular expression \+[0-9]+[ACGTNacgtn]+ denotes an insertion of one or more bases starting from the next position
- A sequence matching the regular expression -[0-9]+[ACGTNacgtn]+ denotes a deletion of one or more bases starting from the next position
- ^ (caret) marks the start of a read segment and the ASCII of the character following `^' minus 33 gives the mapping quality
- $ (dollar) marks the end of a read segment
- * (asterisk) is a placeholder for a deleted base in a multiple basepair deletion that was mentioned in a previous line by the -[0-9]+[ACGTNacgtn]+ notation

Column 6: The base quality string [edit]

This is an optional column. If present, the ASCII value of the character minus 33 gives the mapping Phred quality of each of the bases in the previous column 5. This is similar to quality encoding in the FASTQ format.