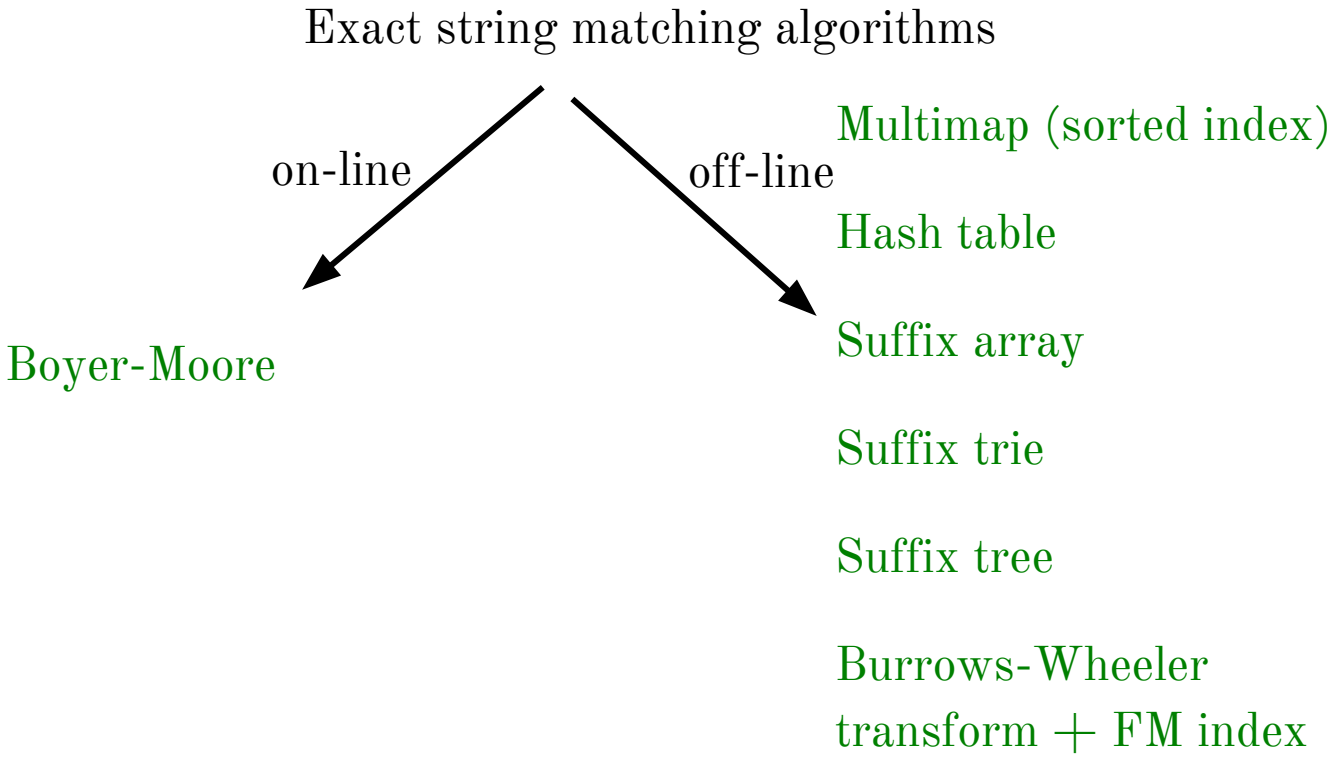
1. Bioinformatic workflows.
2. Variant calling.
3. Cancer analysis.

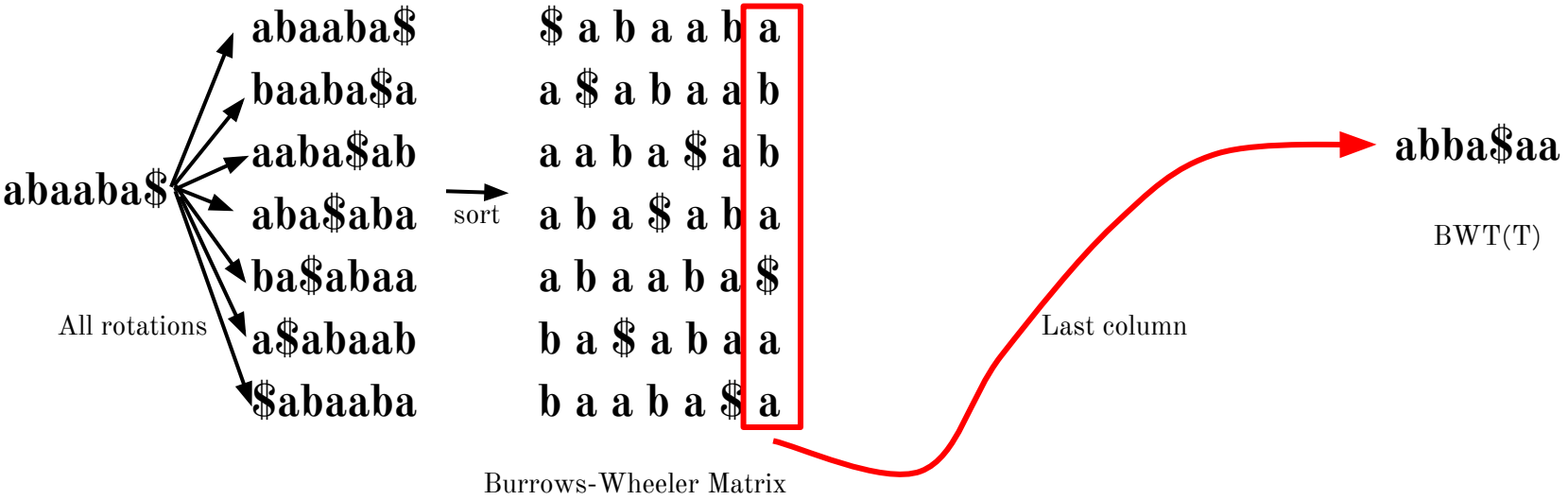
Lesson 05

# Recapitulation



# Burrows-Wheeler Transform

Reversible permutation of the characters of a string, used originally for compression




How is it useful for compression?

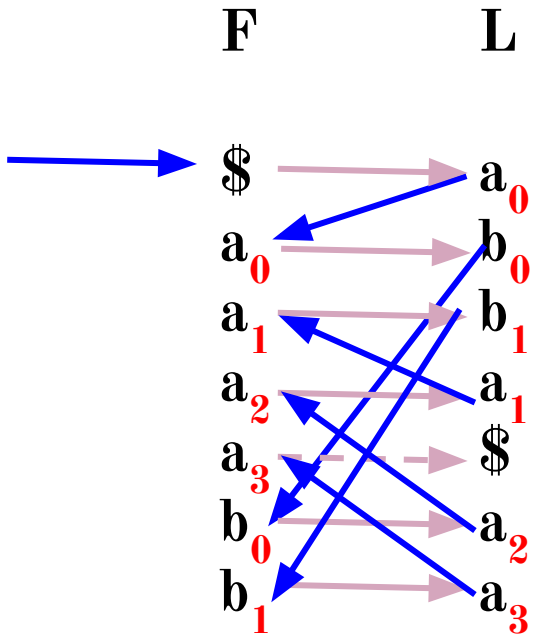
How is it reversible?

How is it an index?

# Burrows-Wheeler Transform: reversing

Reverse BWT(T) starting at right-hand-side of T and moving left

Start in first row. F must have \$.
L contains character just prior to \$: $a_0$
...

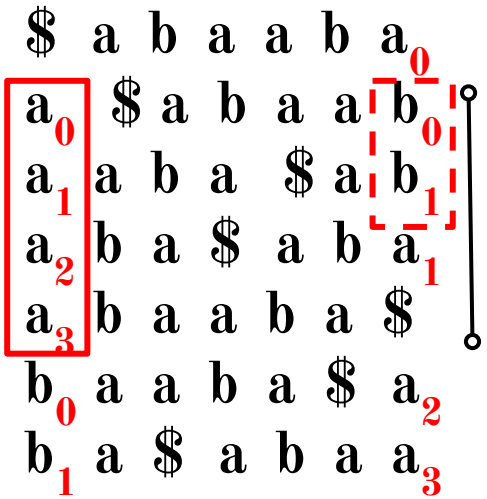



Reverse of chars we visited = $\mathbf{a_3\ b_1\ a_1\ a_2\ b_0\ a_0\ \$}$ = T

# FM Index: querying

Look for range of rows of BWM(T) with P as prefix
Do this for P's shortest suffix, then extend to successively longer suffixes until range becomes empty or we've exhausted P

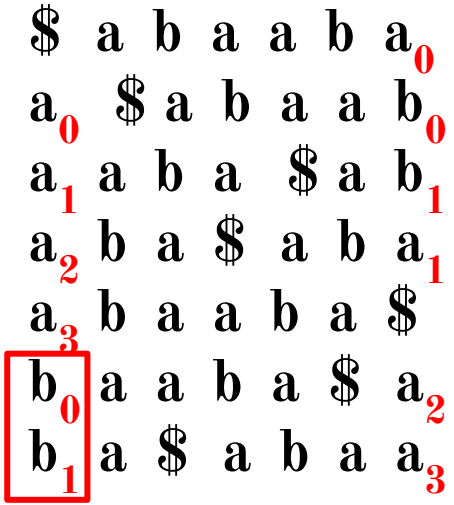
**P = aba**




Look at those rows in L.
$b_0$, $b_1$ are b-s occurring just to left.

Use LF Mapping. Let new range delimit those b-s

**P = aba**



# FM Index: querying

We have rows beginning with **ba**, now we seek rows beginning with **aba**

**P = aba**

$ a b a a b a$$_0$$

a$$_0$$ $ a b a a b$$_0$$

a$$_1$$ a b a $ a b$$_1$$

a$$_2$$ b a $ a b a$$_1$$

a$$_3$$ b a a b a $

b$$_0$$ a a b a $ a$$_2$$

b$$_1$$ a $ a b a a$$_3$$

Occurs just to the left

Use LF mapping

**P = aba**

| F | | | | | | L |
|---|---|---|---|---|---|---|
| $ | a | b | a | a | b | a$$_0$$ |
| a$$_0$$ | $ | a | b | a | a | b$$_0$$ |
| a$$_1$$ | a | b | a | $ | a | b$$_1$$ |
| a$$_2$$ | b | a | $ | a | b | a$$_1$$ |
| a$$_3$$ | b | a | a | b | a | $ |
| b$$_0$$ | a | a | b | a | $ | a$$_2$$ |
| b$$_1$$ | a | $ | a | b | a | a$$_3$$ |

Now we have the rows with prefix **aba**

# FM Index

1. L = BWT(T)
2. First column (number of appearances of each character)
3. Suffix Array (or SA Sample)
4. Tally (rank, occurrences) matrix

# FM Index: Example

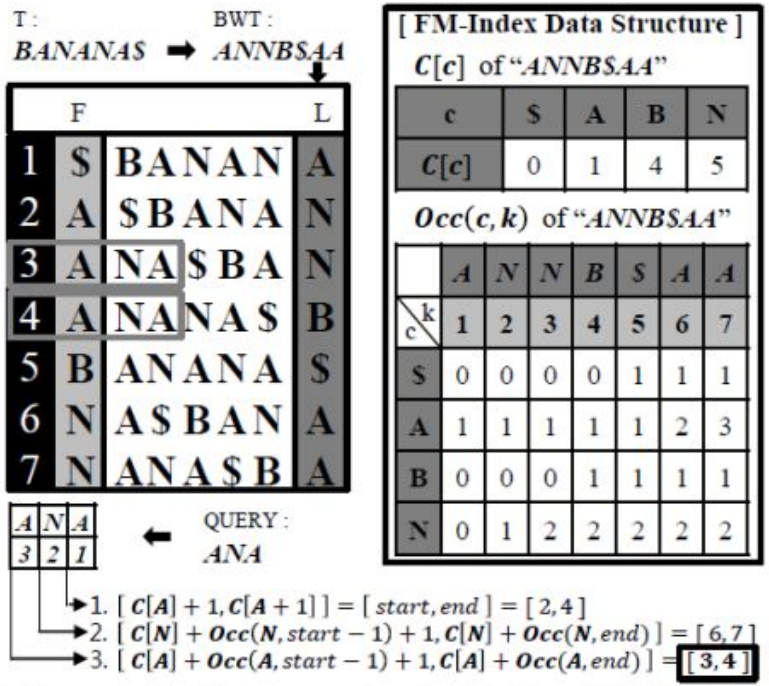


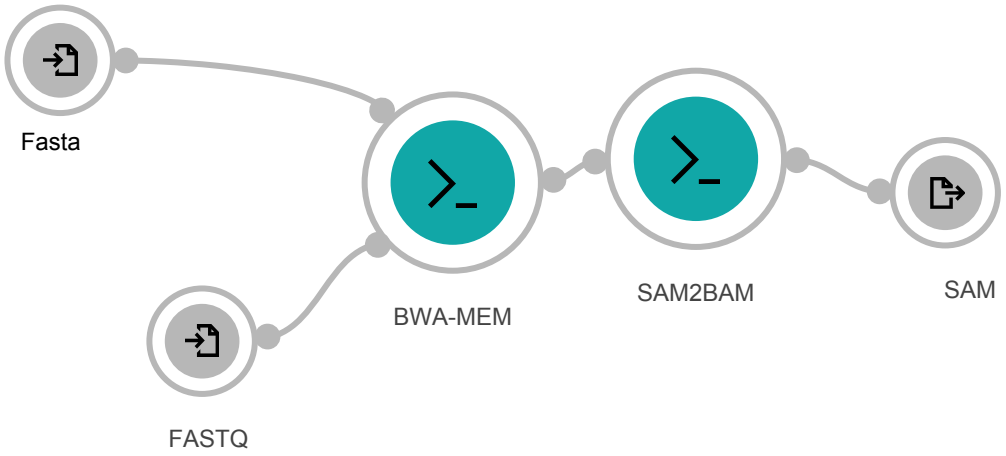

Fig. 3. An example of query search using BWT and FM-index for text T=*BANANA$*.The *$* is 'EOF' character.

# Bioinformatic workflows and cloud computing

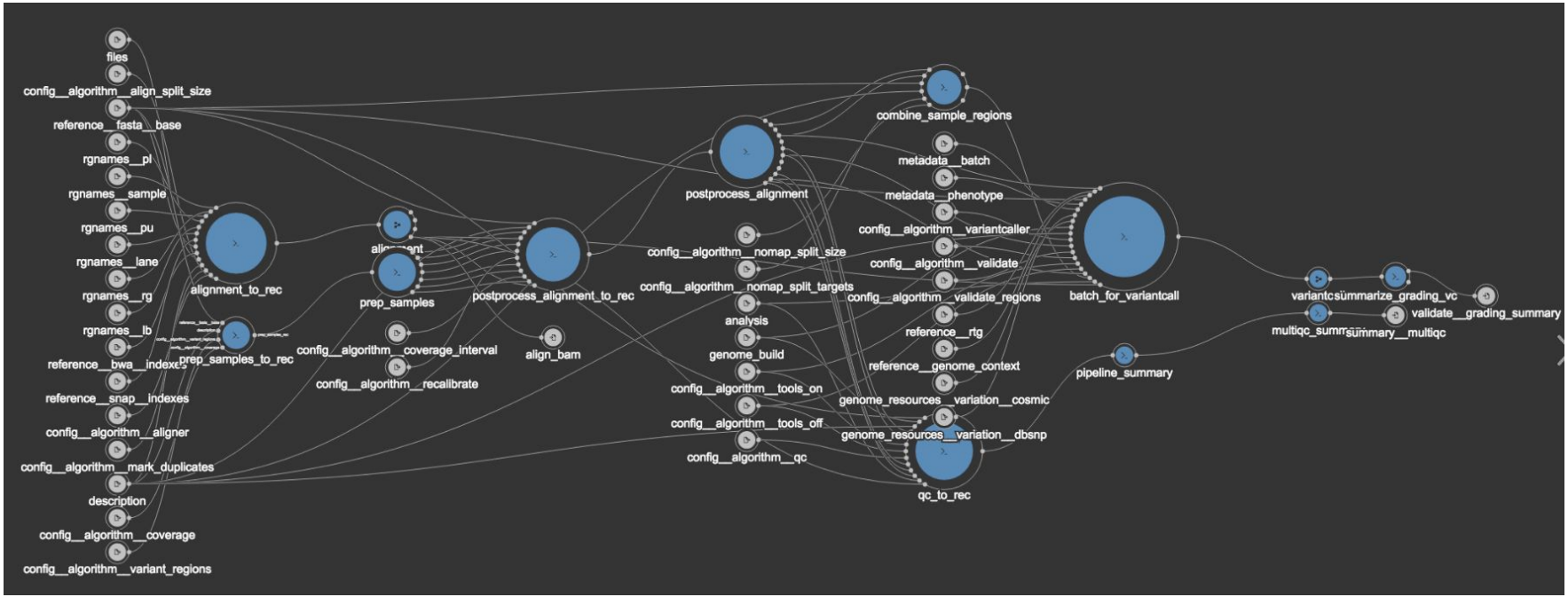
# What is a workflow?

- Acyclic graph of tools connected to perform some analysis
- Workflow's nodes are:
  - Inputs (file or parameter)
  - Tools
  - Outputs
  - Workflow

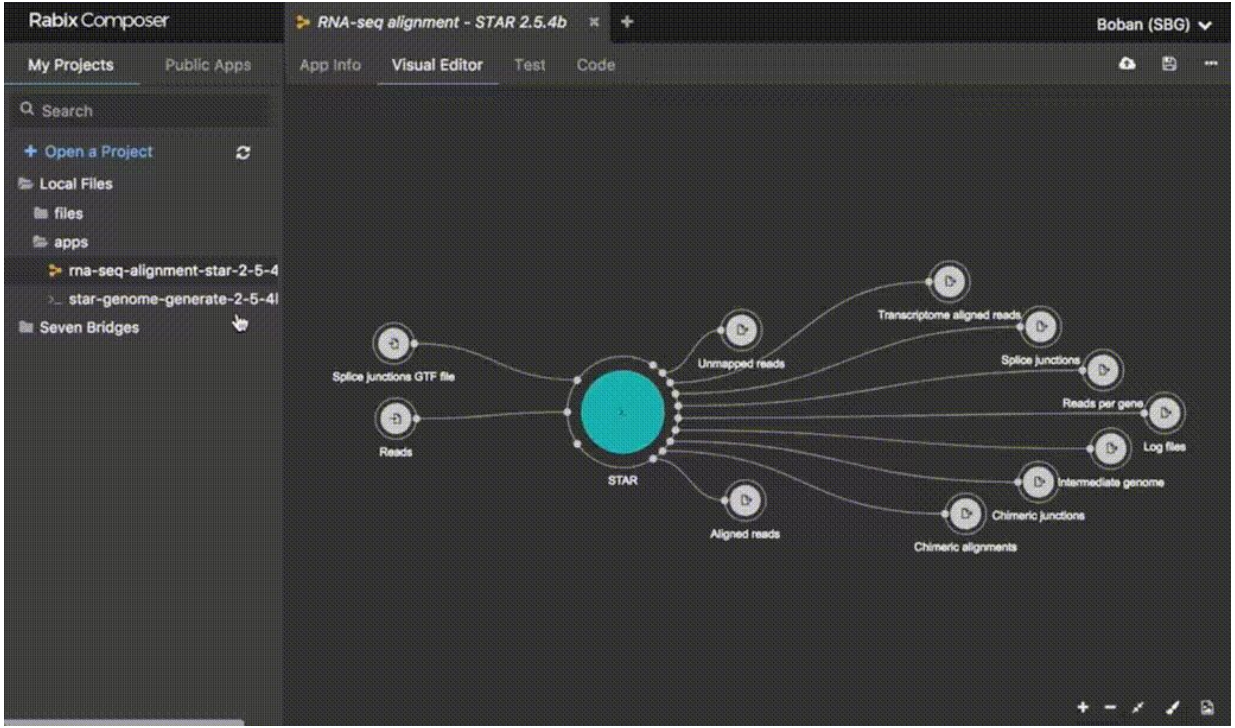
```
bwa mem ref.fa read1.fq read2.fq >
aln.sam
sam2bam aln.sam > aln.bam
```



# Why we need a workflow?



# How to build a workflow?




https://github.com/rabix/composer

# Common Workflow Language

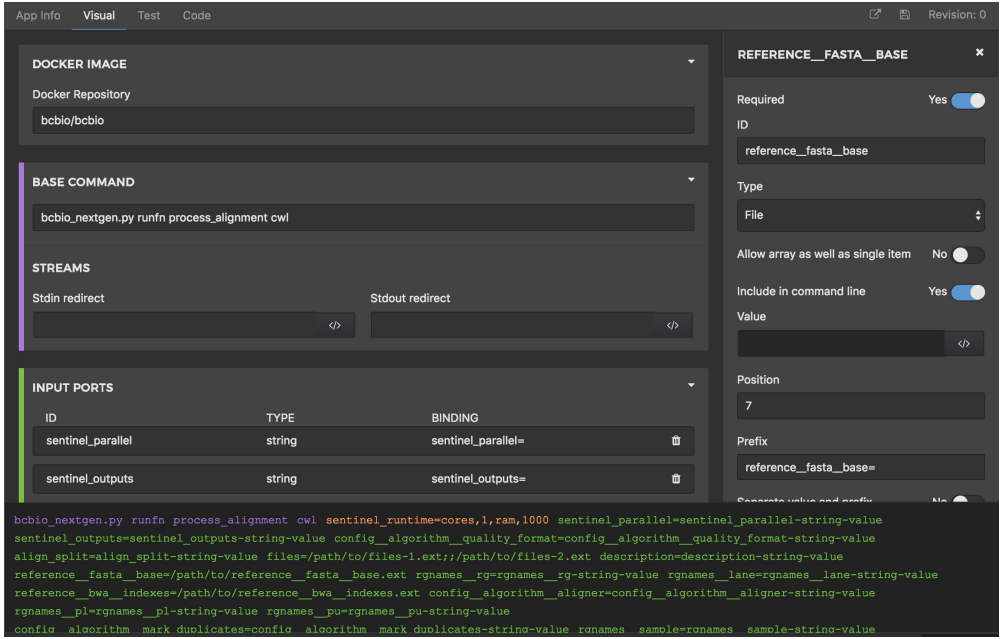
COMMON
WORKFLOW
LANGUAGE

- Define inputs and outputs of a software, runtime and requirements
- Define how to connect software, creating a workflow
- Ensure reproducibility and portability
- Think of CWL as a detailed recipe!

# Common Workflow Language

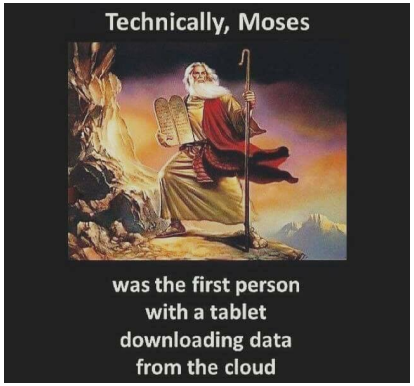
COMMON WORKFLOW LANGUAGE

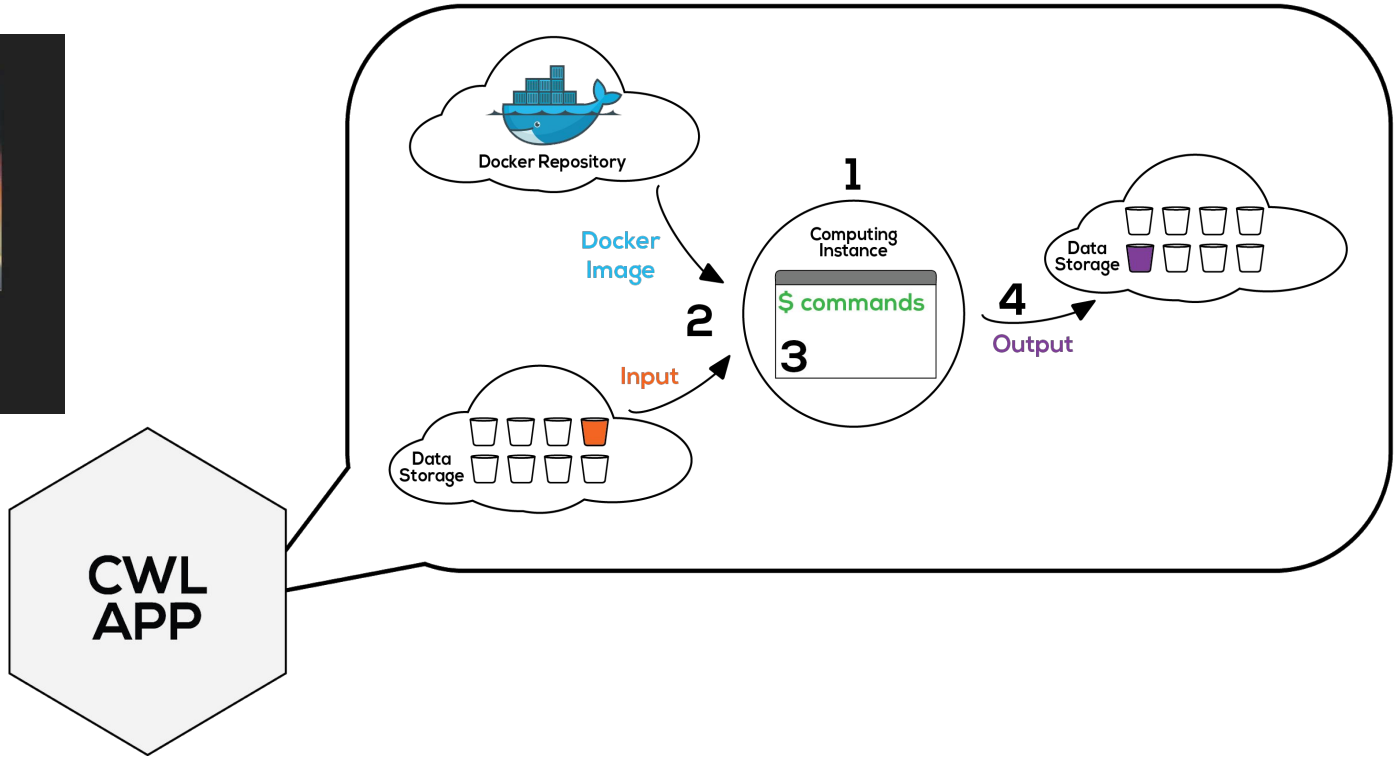
- **Reproducible** analyses (standard)
- Scalable execution
- Metadata & file registry integration
- **Portability** - deployable on multiple platforms
- Revision management and versioning
  User management / permissions



# CWL @ Cloud

COMMON WORKFLOW LANGUAGE

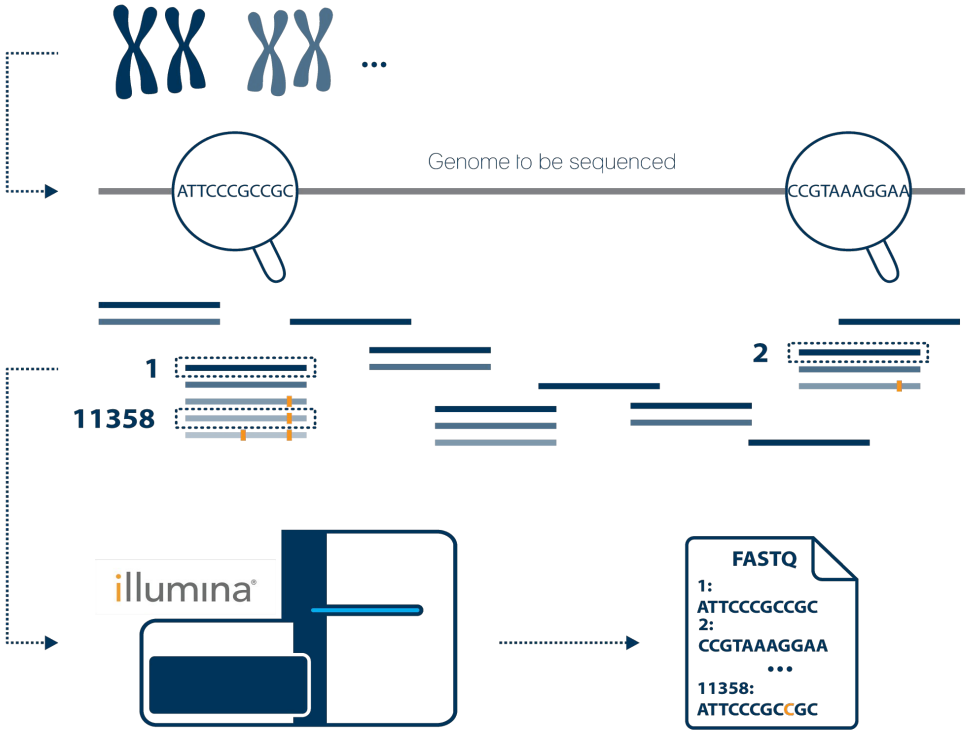


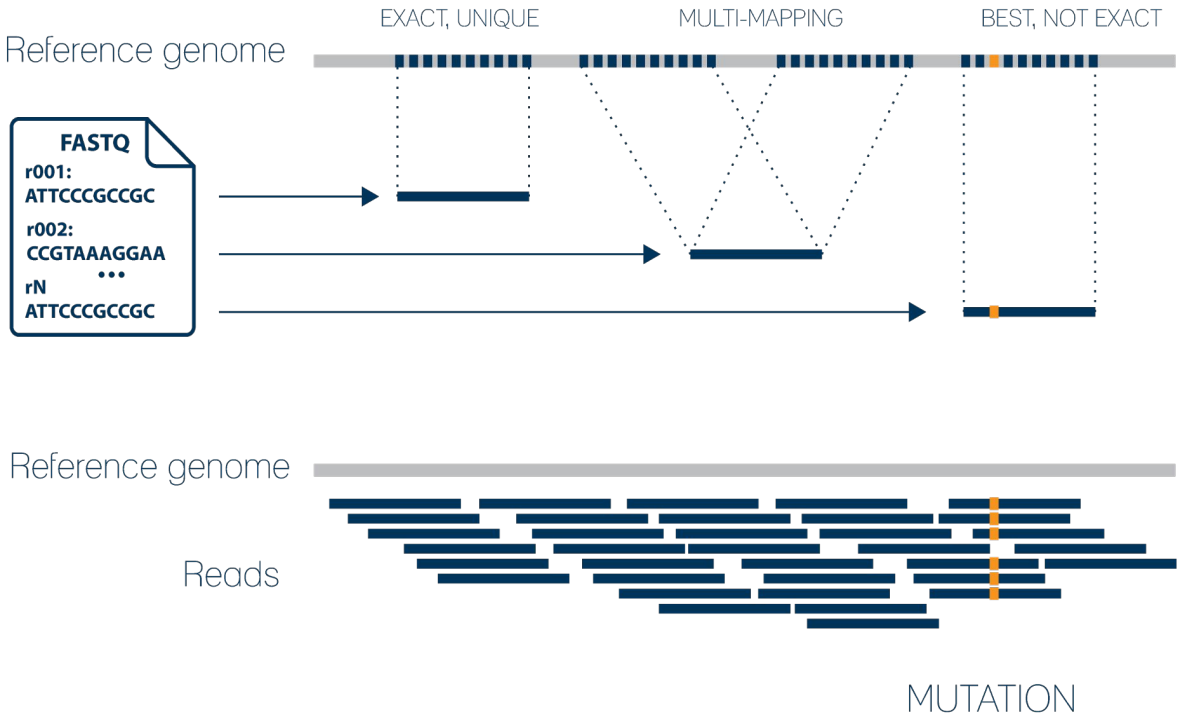
# Variant Calling

# Reminder: DNA Sequencing

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



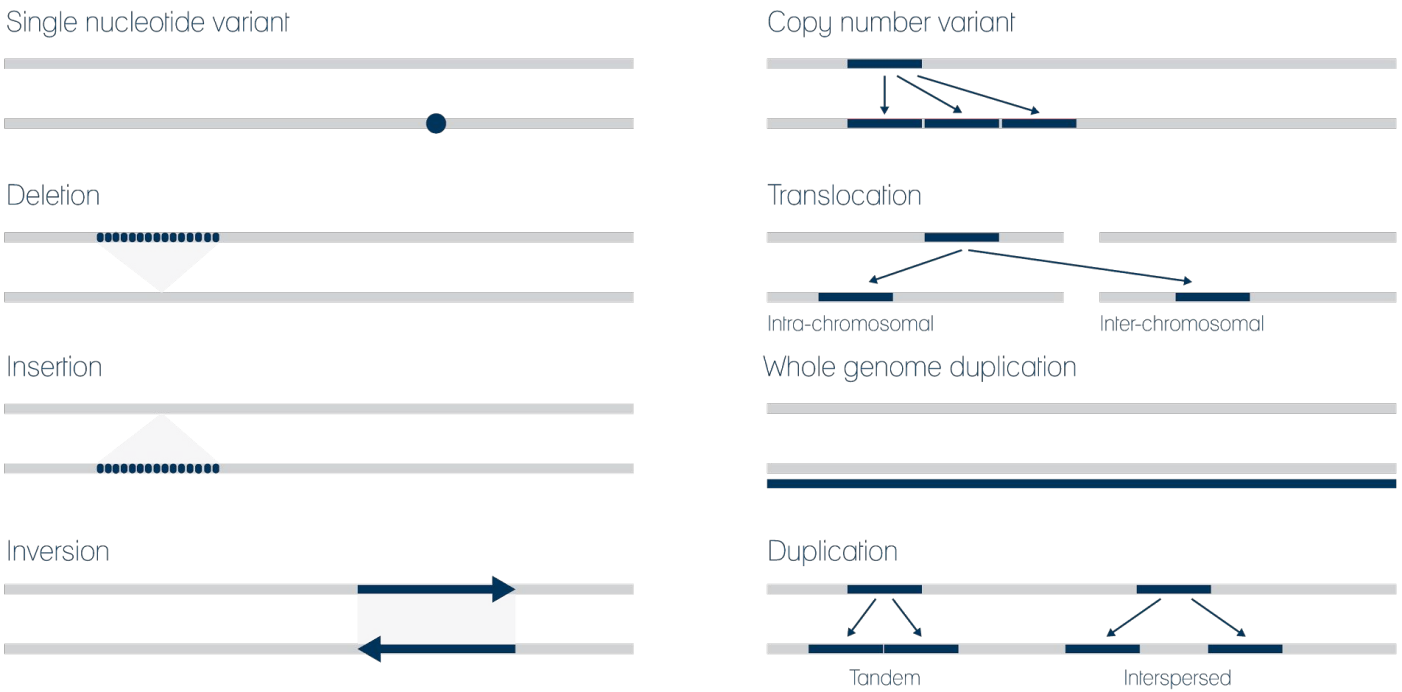
# Reminder: Alignment



# 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

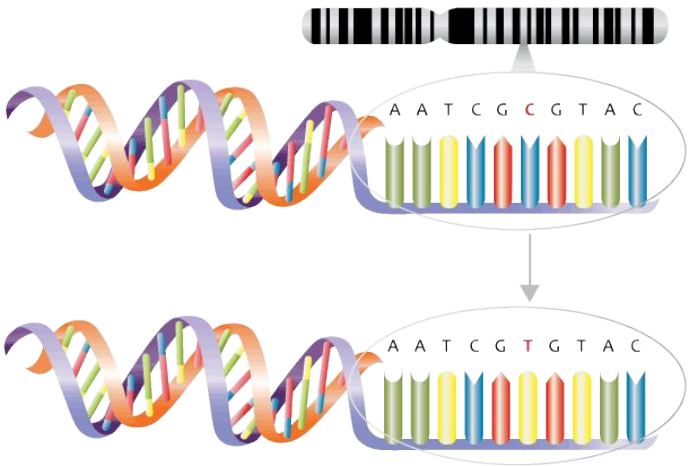
# Genomic Variants



# Genomic Variants

- SNV ( Single Nucleotide Variant)

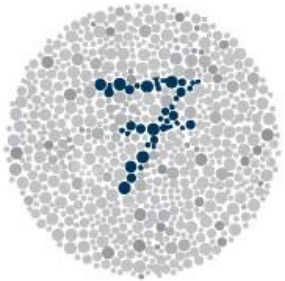
  Simple ones - not a big change on the first look, but...



# 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**

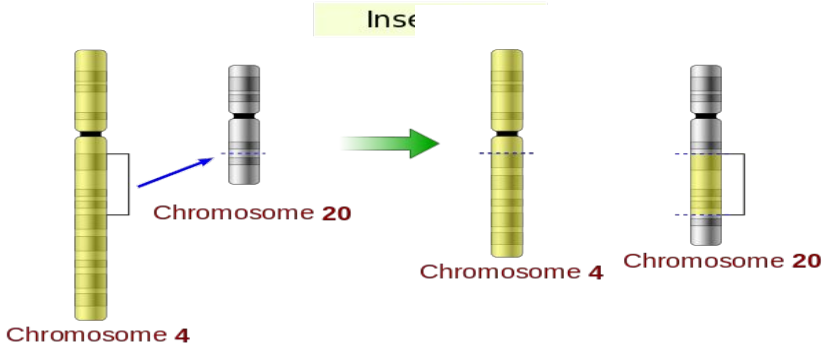
Chromosome 13
Position 32,340,455

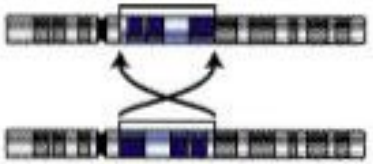
Cancer genotypes: **CC, CT and TT**

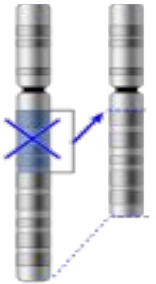
http://www.eupedia.com/genetics/cancer_related_snp.shtml
https://www.snpedia.com/index.php/Rs1799954

# Genomic Variants

Deletions, Insertions, Translocations, Inversions, and some others...



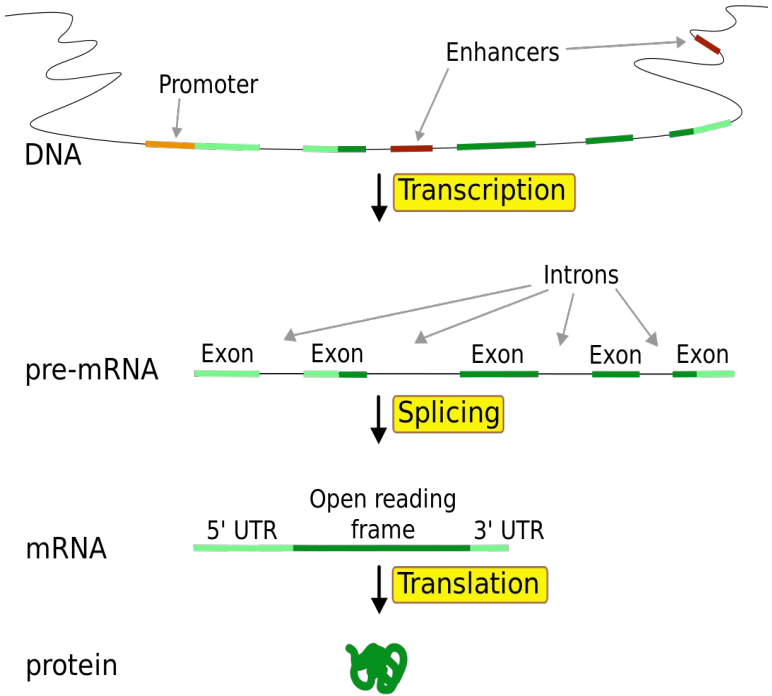





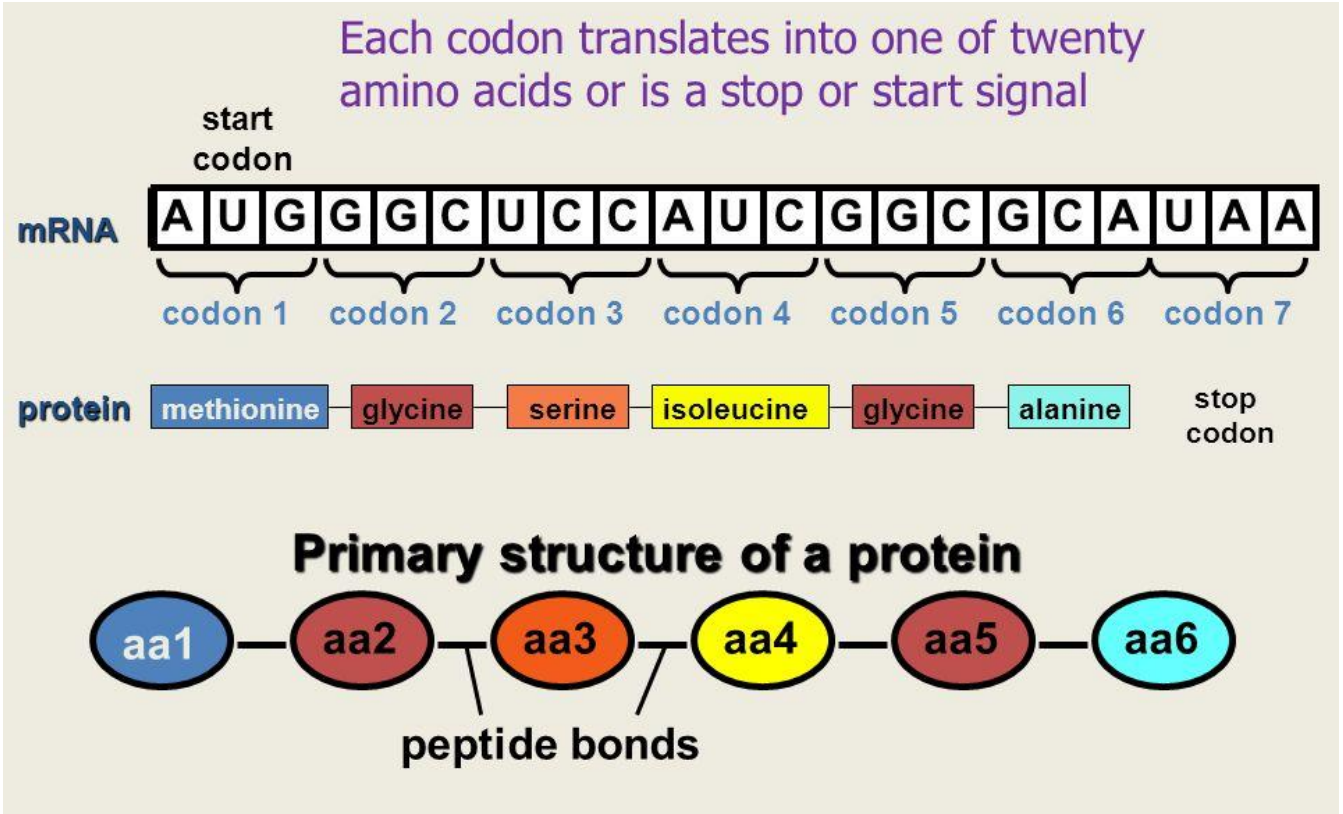
# Genomic Variants

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


- Central dogma



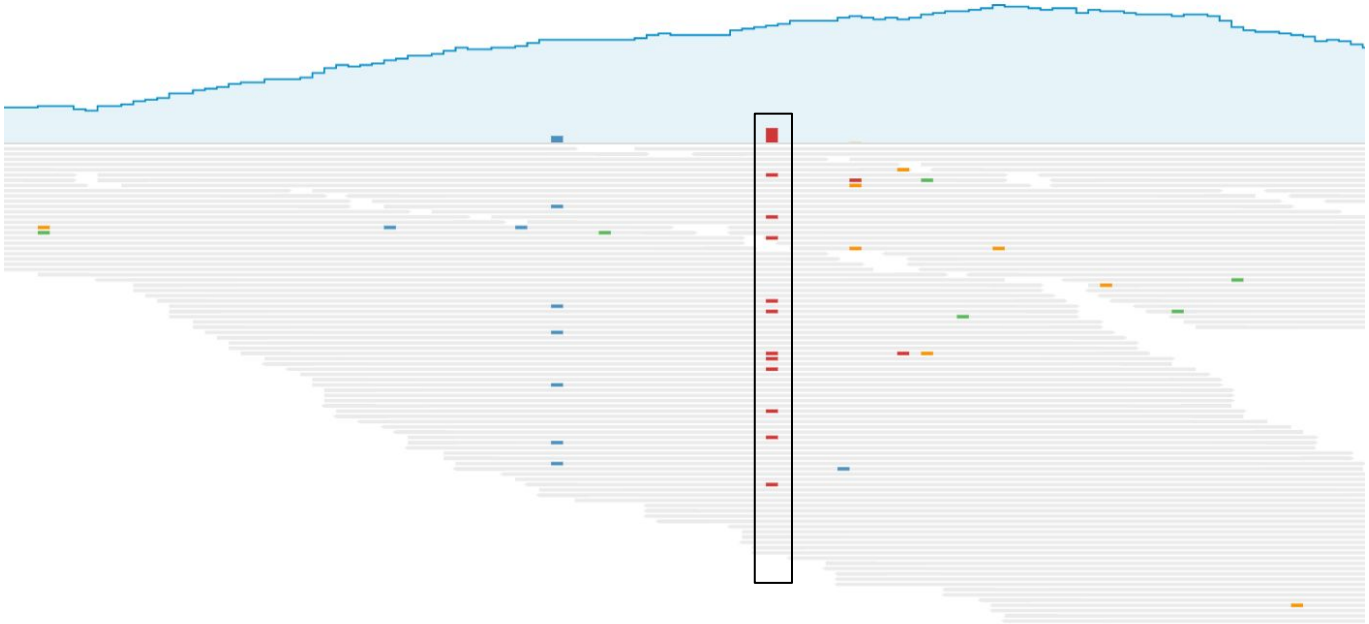
# RNA to Protein



# Genomic Variants

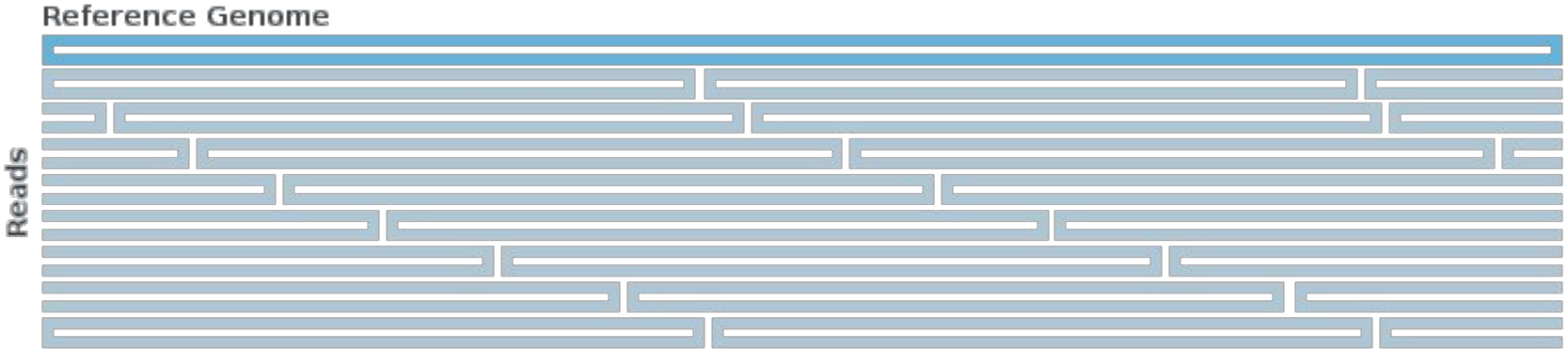
- Variants can have different impact on human cells and organism
- Single Nucleotide Variants(**SNV**):
  - Harmless
    - **Silent** – Usually no effect
  - Harmful:
    - **Missense** – Amino acid change
    - **Nonsense**(Start/Stop Gain/Lost) – AUG / UAG, UAA, UGA
  - Depends on the location
    - **Noncoding regions** ( Promoter, Enhancer, lncRNA, miRNA...)
- Insertions/Deletions – **INDELS**
  - **In frame**
  - **Out of frame (Frameshift)**

# What is the pileup?



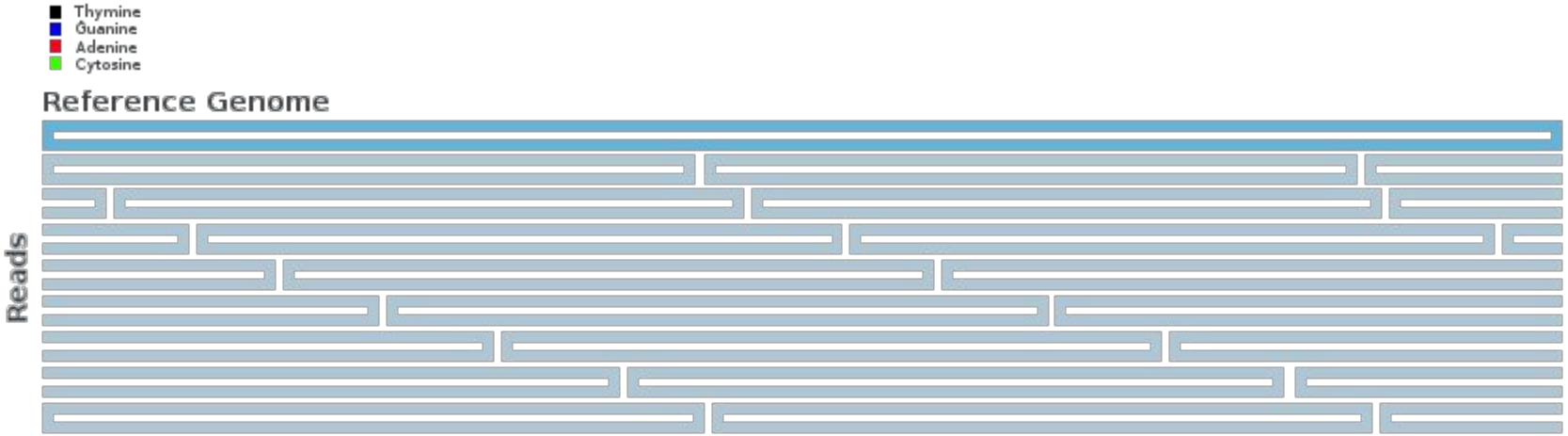
# Ideal Variant Calling

# Ideal Variant Calling

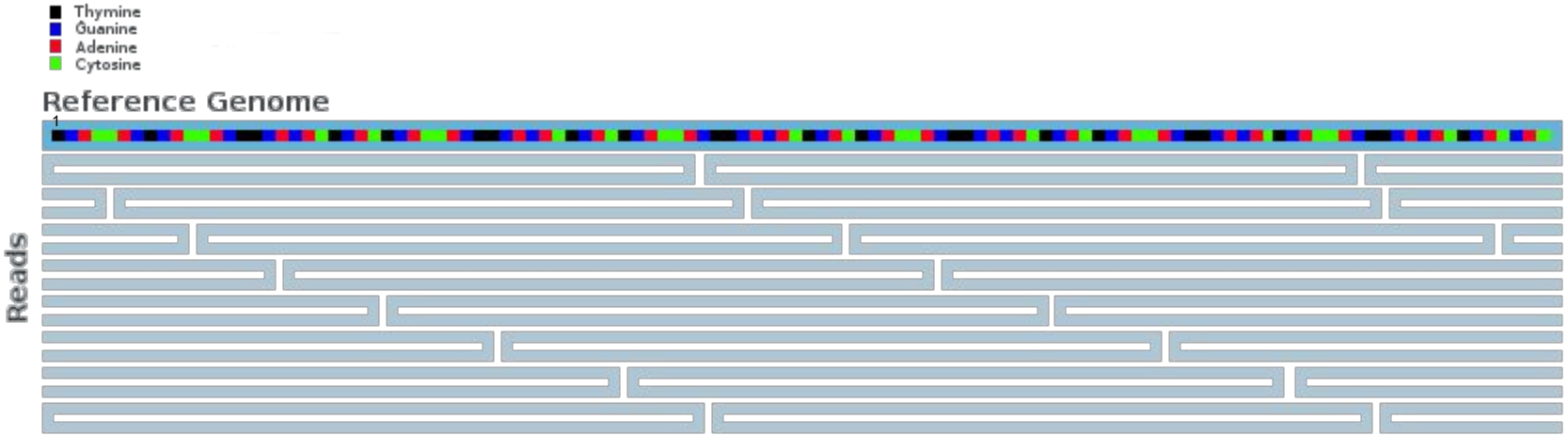



Ideally we will have uniform distribution of reads.

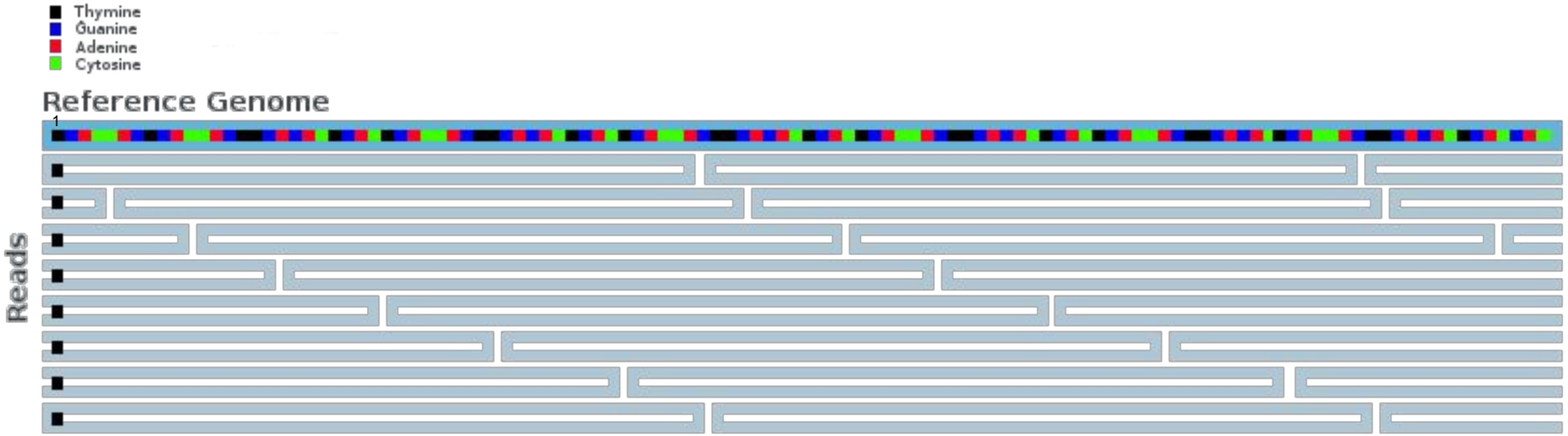
# Ideal Variant Calling



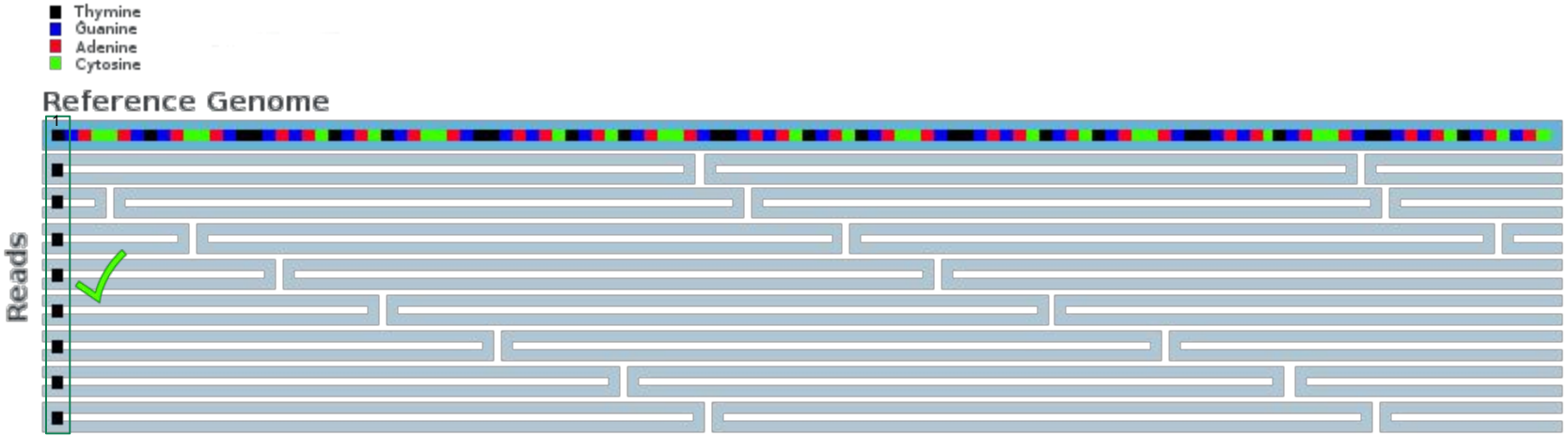
# Ideal Variant Calling

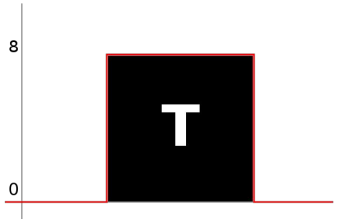


# Ideal Variant Calling



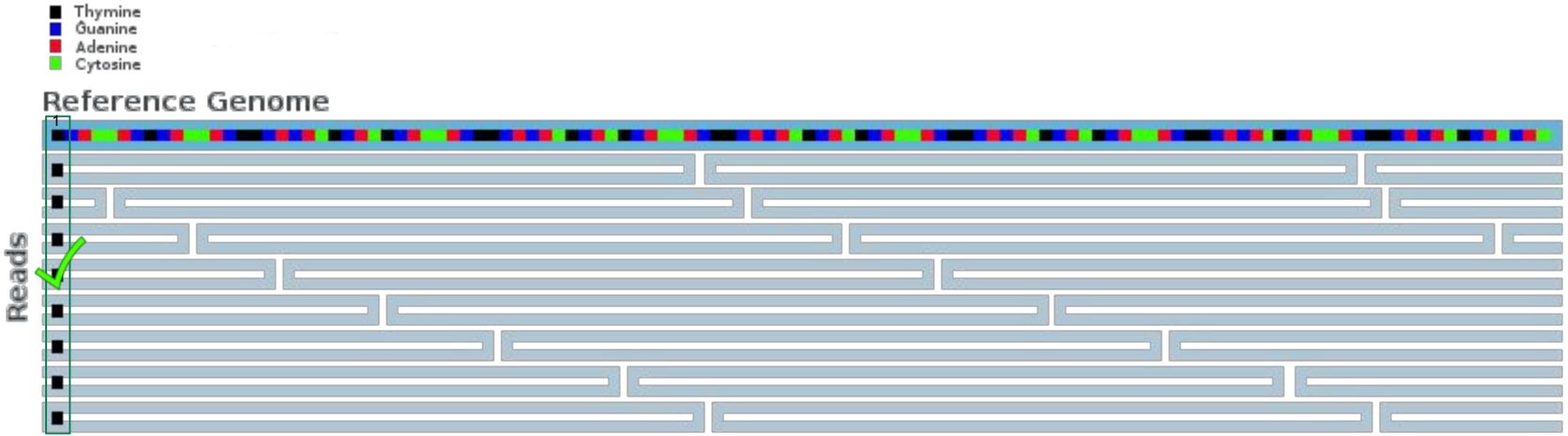
# Ideal Variant Calling

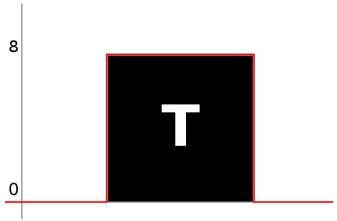



- We have “T” in the all reads covering that position

# Ideal Variant Calling

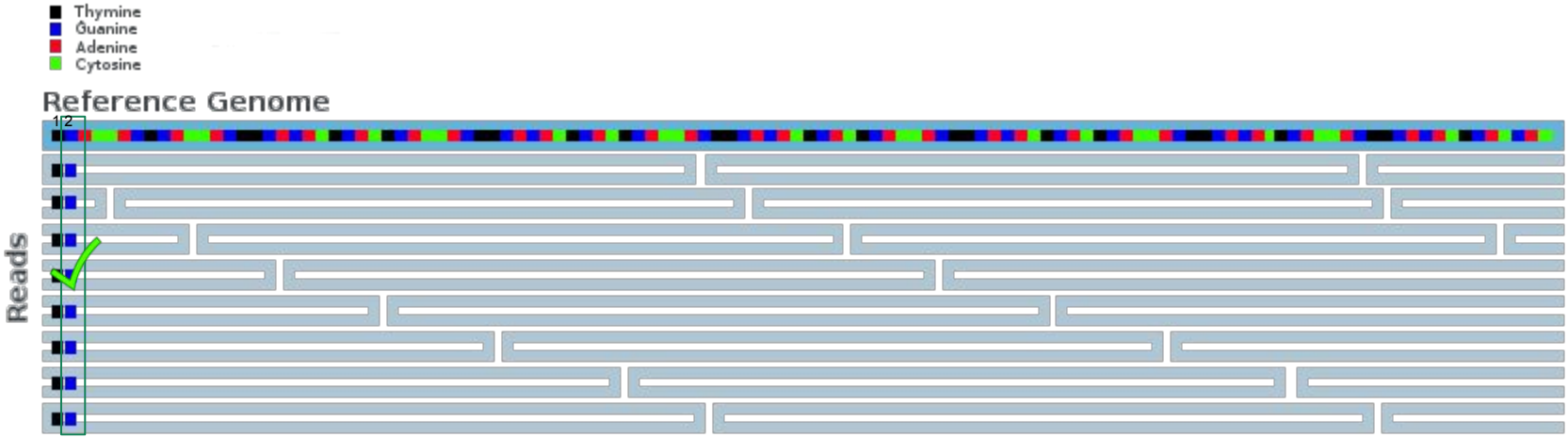


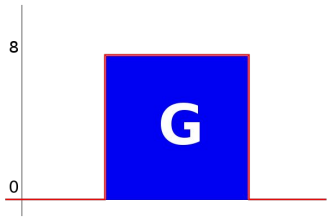




## How can we represent what we have observed?

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

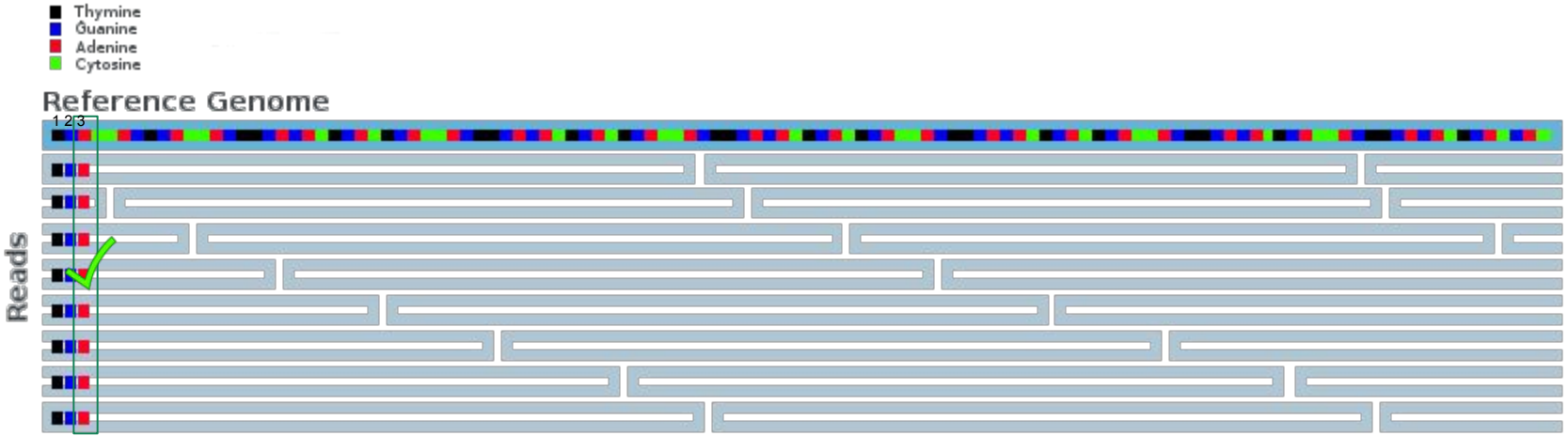
# Ideal Variant Calling

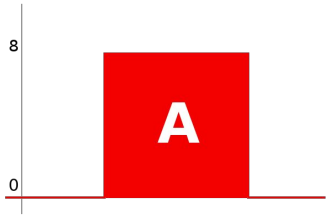



| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 1 | T | - | 0/0 |
| X | 2 | G | - | 0/0 |

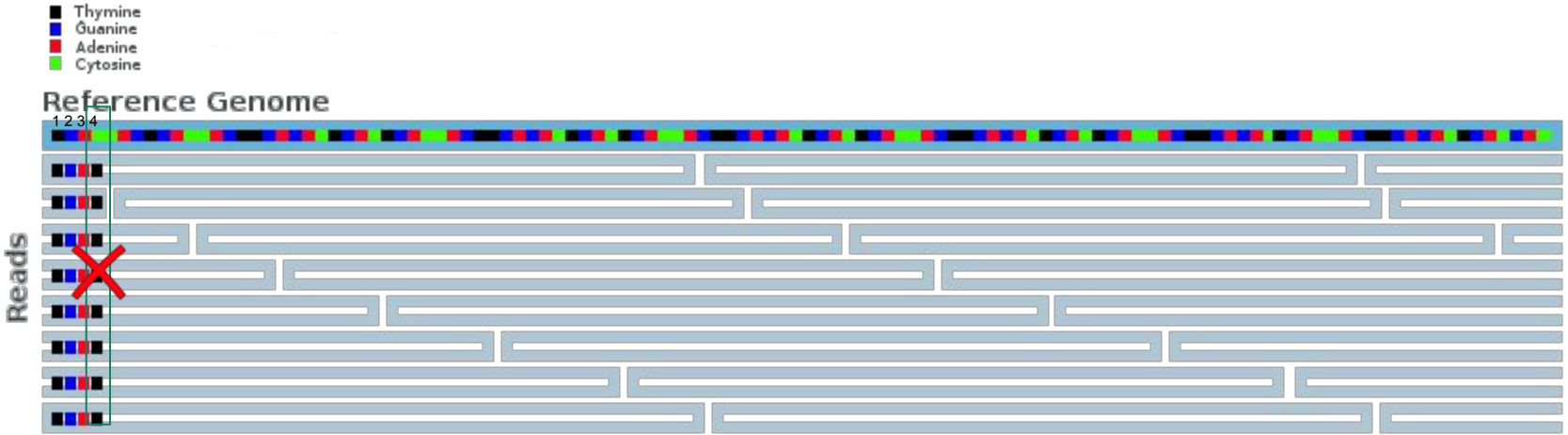
# Ideal Variant Calling

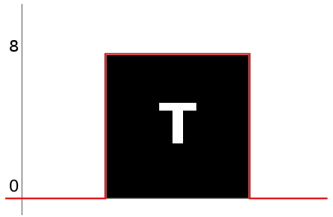



| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 1 | T | - | 0/0 |
| X | 2 | G | - | 0/0 |
| X | 3 | A | - | 0/0 |

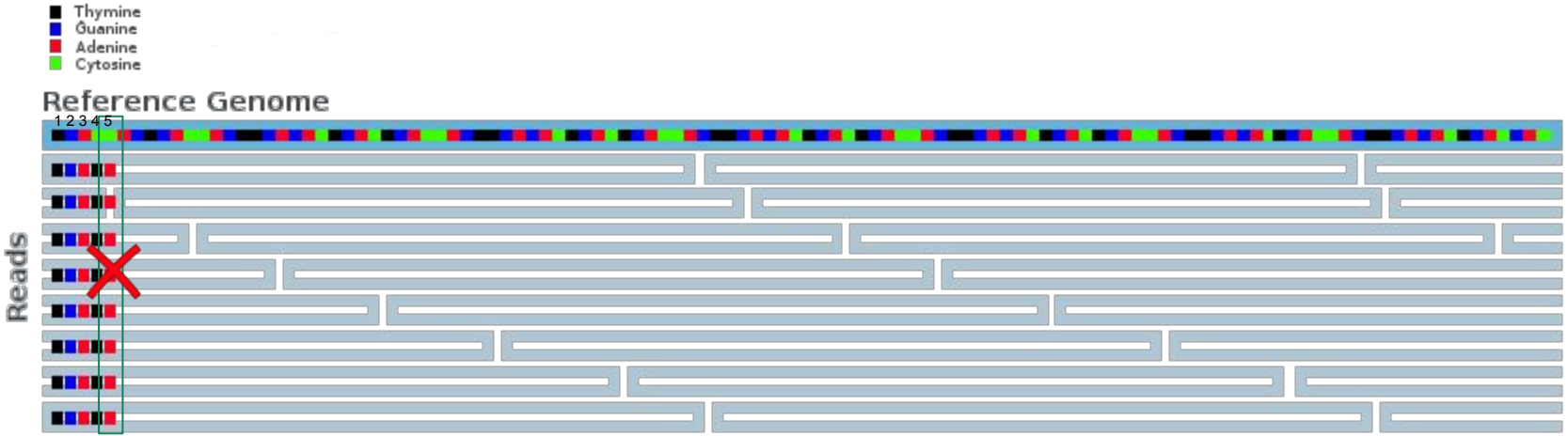
# Ideal Variant Calling

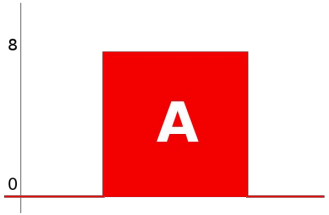



| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 2 | G | - | 0/0 |
| X | 3 | A | - | 0/0 |
| X | 4 | **C** | **T** | **1/1** |

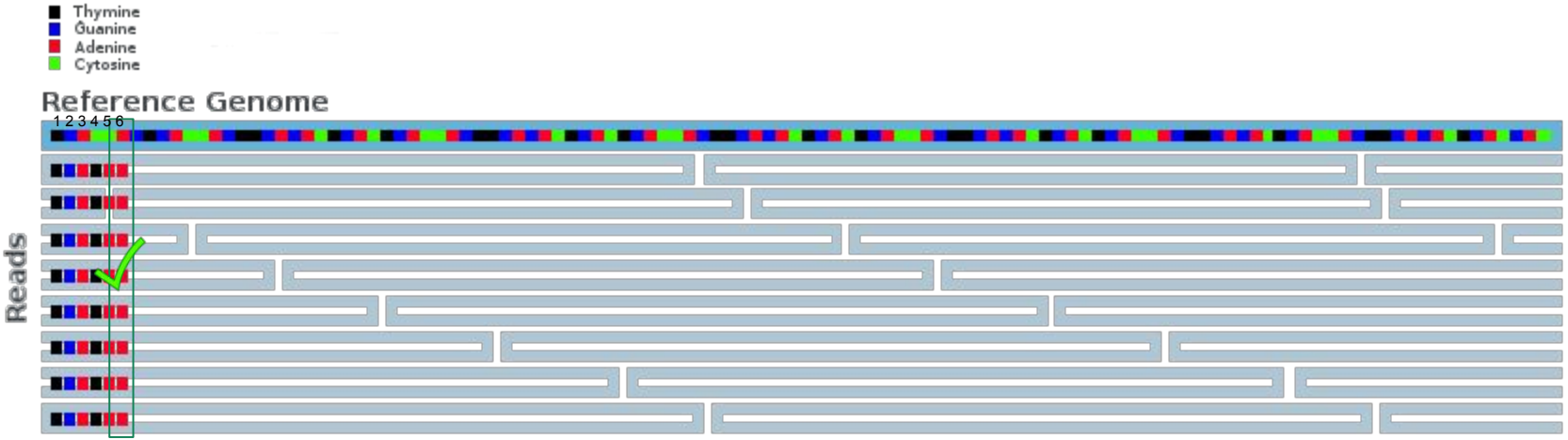
# Ideal Variant Calling

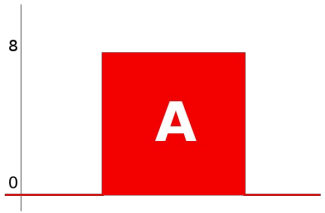



| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 3 | A | - | 0/0 |
| X | 4 | **C** | **T** | **1/1** |
| X | 5 | **C** | **A** | **1/1** |

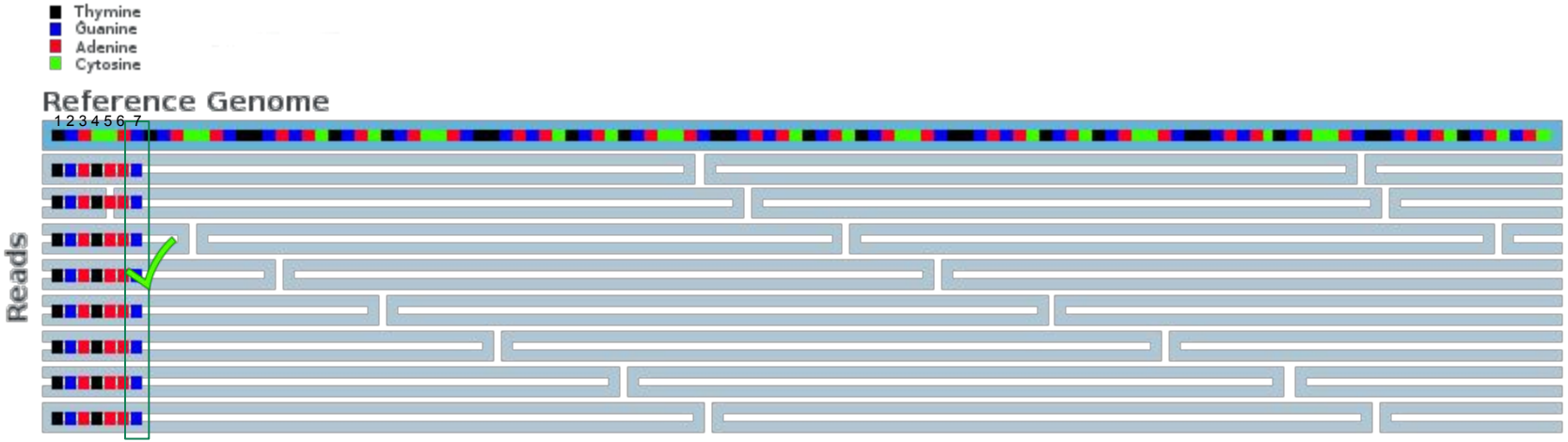
# Ideal Variant Calling

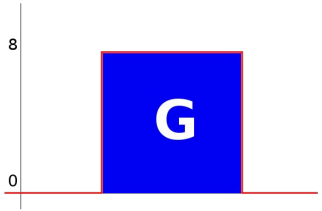



| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 4 | **C** | **T** | **1/1** |
| X | 5 | **C** | **A** | **1/1** |
| X | 6 | A | - | 0/0 |

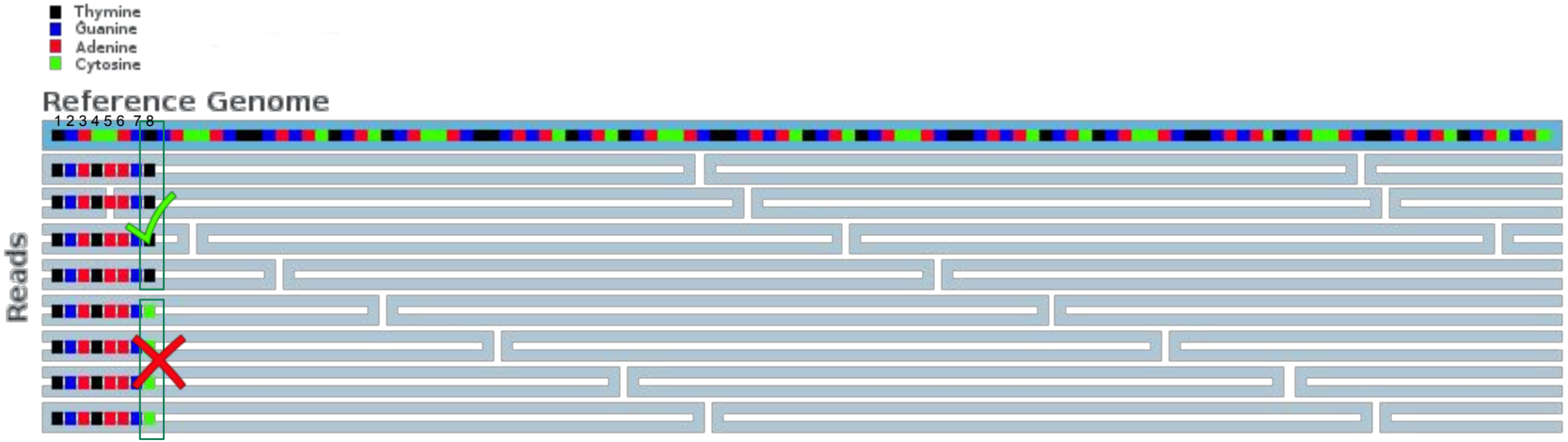
# Ideal Variant Calling

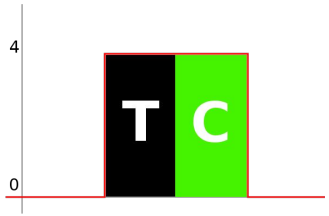



| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 5 | **C** | **A** | **1/1** |
| X | 6 | A | - | 0/0 |
| X | 7 | G | - | 0/0 |

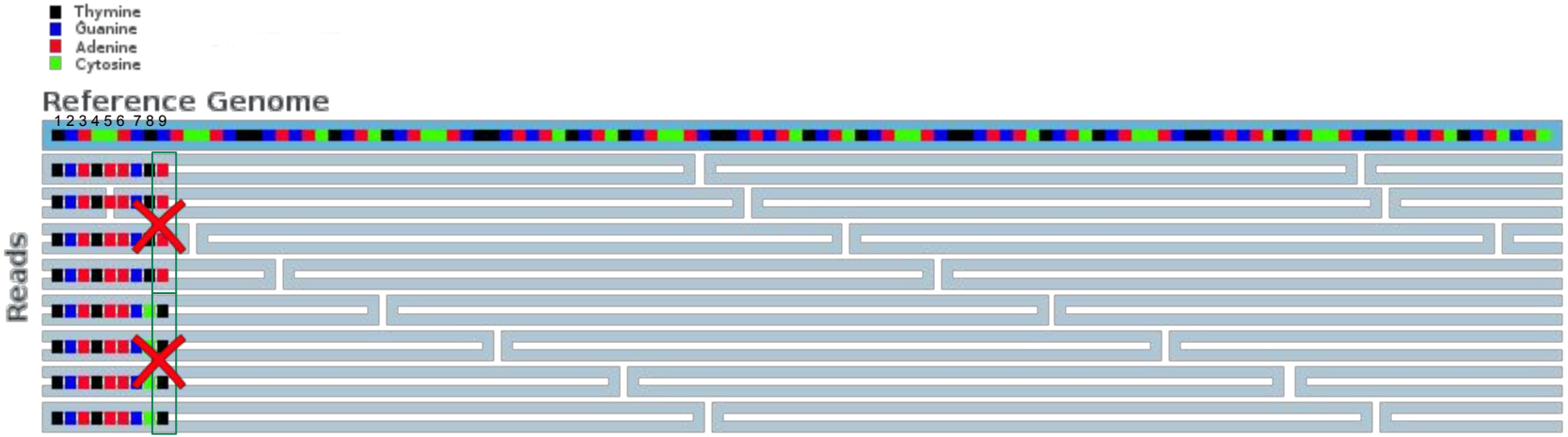
# Ideal Variant Calling

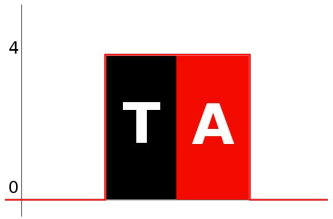



| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 6 | A | - | 0/0 |
| X | 7 | G | - | 0/0 |
| X | 8 | **T** | **C** | **0/1** |

# Ideal Variant Calling



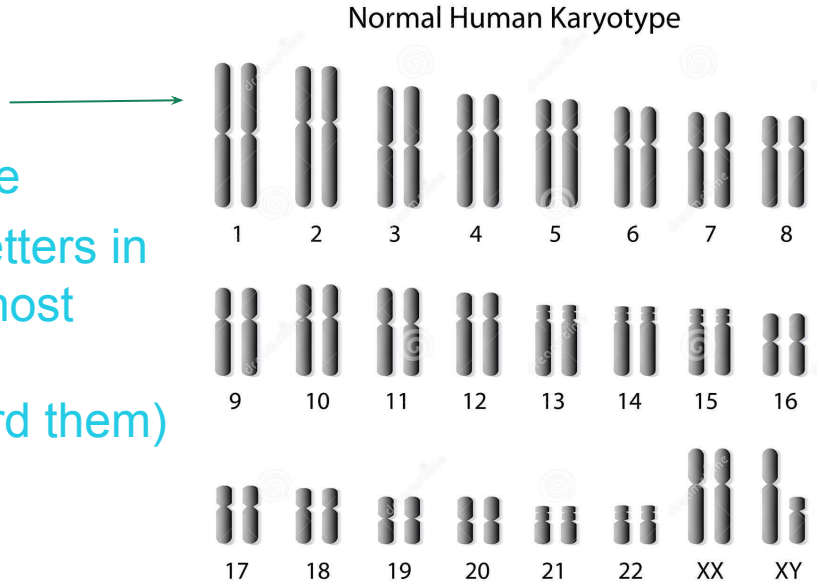




| CONTIG | POS | REF | ALT | GT |
|---|---|---|---|---|
| X | 7 | G | - | 0/0 |
| X | 8 | **T** | **C** | **0/1** |
| X | 9 | **G** | **A,T** | **1/2** |

# 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)

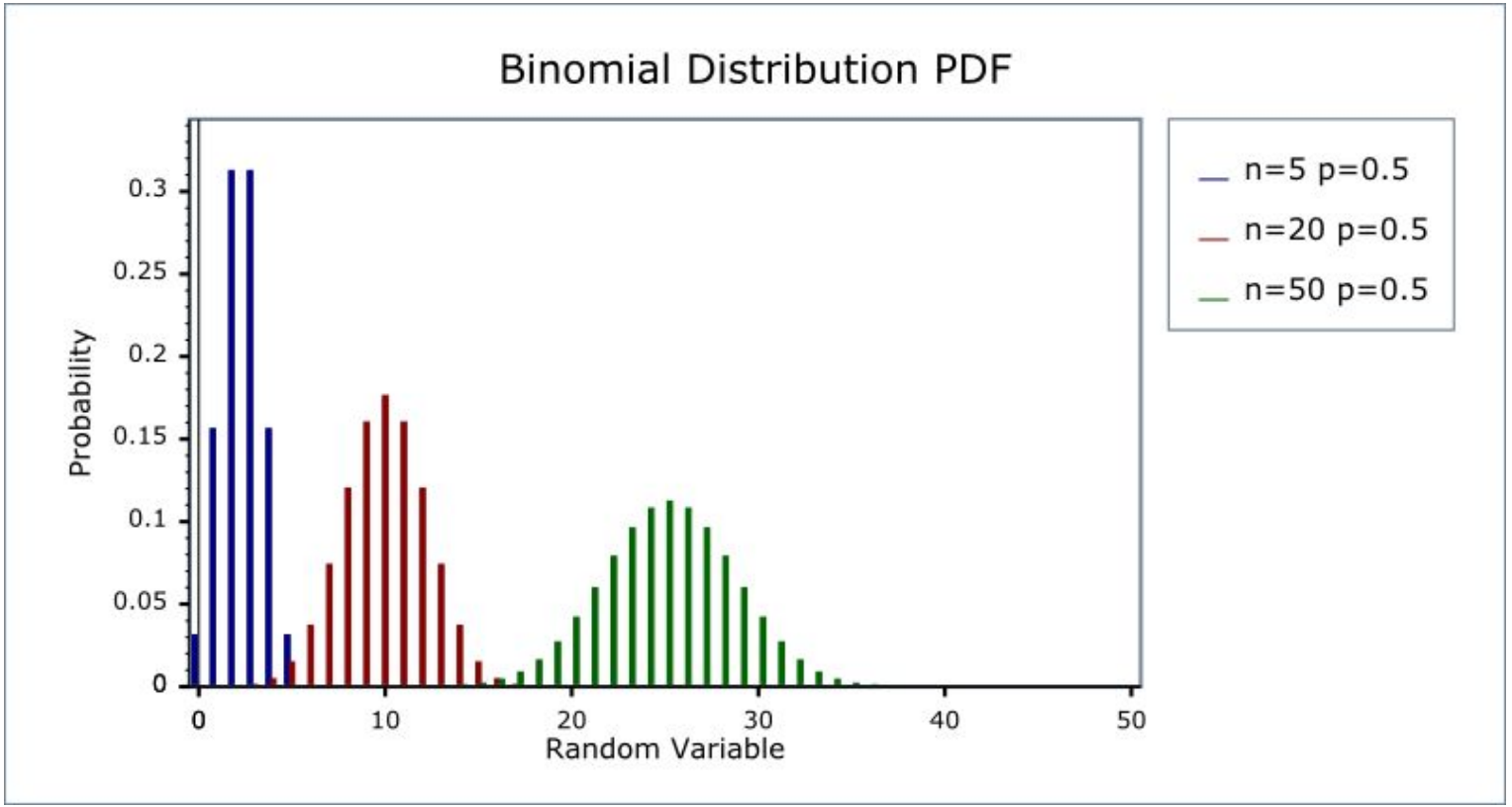


# Binomial distribution

- Models the number of successes in a sequence of yes/no experiments
- Parameters:
  - n - number of trials
  - p - probability of a success in a single trial
  - Probability that K out of n trials will be success

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

# Binomial distribution



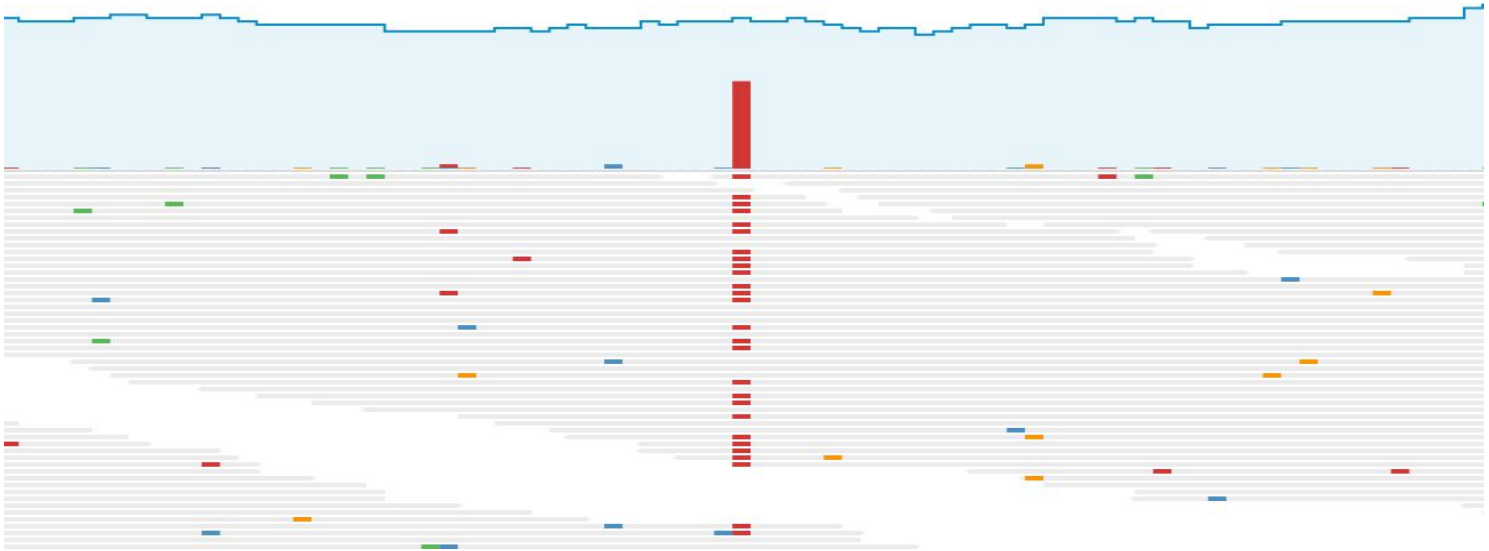
# Variant Calling

- 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
- Population-based methods
  - Separate variant calling from genotyping

# Variant calling results – check out BAM file



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

# Variant calling results

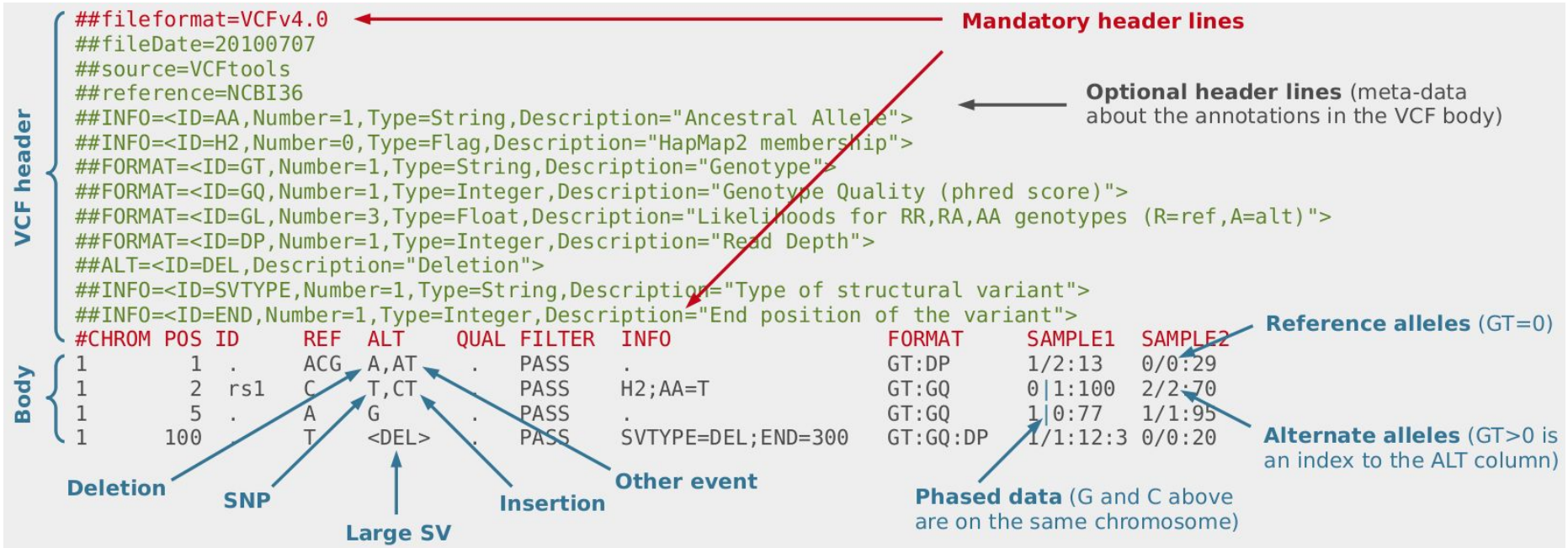
- The result of Variant Calling is a file in VCF format, which contains mutations
- A plain text file format for storing variant data
- A number of line starting with ## -the header
- Main header line:
  #CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1
- This is followed by the actual variant data, 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: http://samtools.github.io/hts-specs/VCFv4.2.pdf

# Variant calling results

- Example of VCF format
- Each row represents one mutation

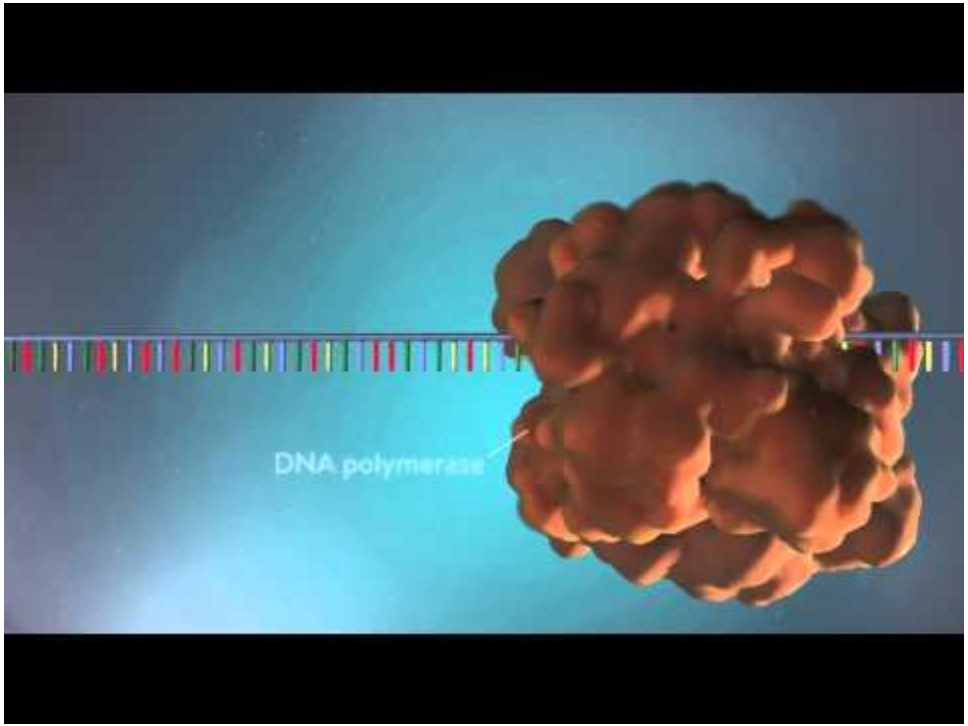
| CHR | POS | REF | ALT | FORMAT | NA12878 |
|---|---|---|---|---|---|
| 1 | 14300 | A | G | GT, VAF | 0/1, 0.4 |
| 2 | 15367 | A | C | GT, VAF | 1/1, 0.9 |
| 3 | 25612 | C | G,A | GT, VAF | 1/2, ? |
| 5 | 5632 | TA | T | GT, VAF | 0/1, 0.5 |
| 7 | 7824 | T | TA | GT, VAF | 1/1, 0.8 |

# Variant Calling Format File



# Computational Cancer Analysis

# DNA replication



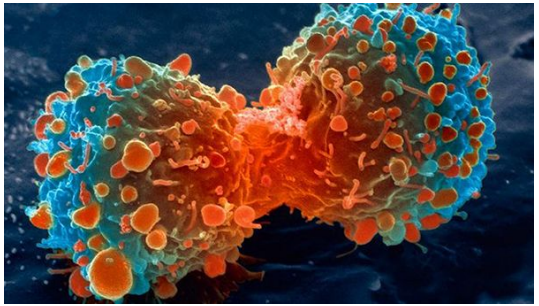
# What is cancer?

Mutation during DNA replication can fall to:

1. Intron (no change)
2. Important gene (cell dies, organism lives)
3. Gene that stops cell division (cell lives, organism...)

What causes cancer (increases probability of mutation)?

1. EM radiation
2. Chemical agents
3. Free radicals
4. Genetic factors
5. Infections (viruses)



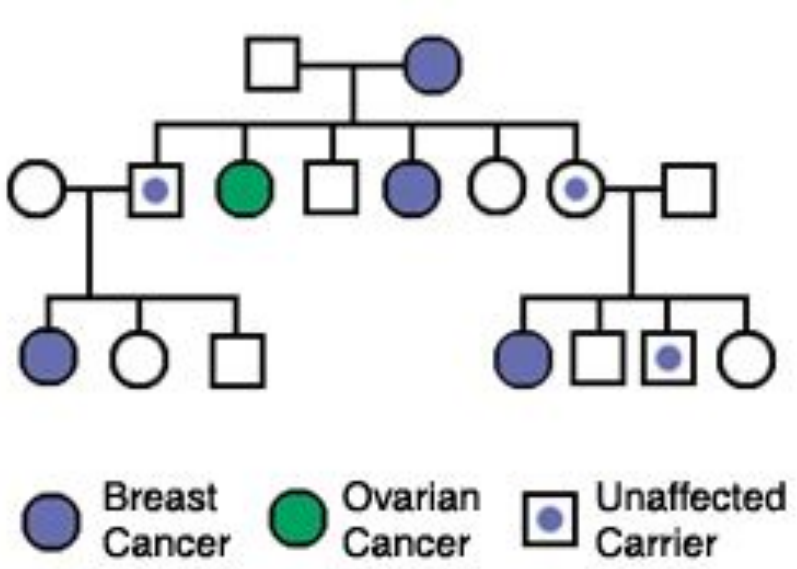
A dividing lung cancer cell.
Credit: National Institutes of Health

# Genetic factors

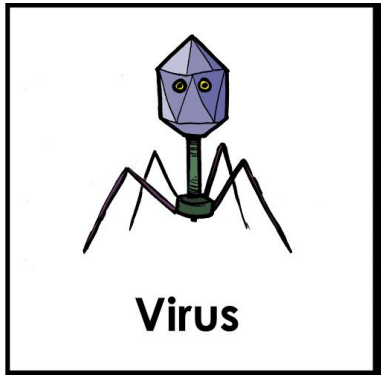
A pedigree from a family with a mutation in the BRCA1 (tumor suppressor) gene

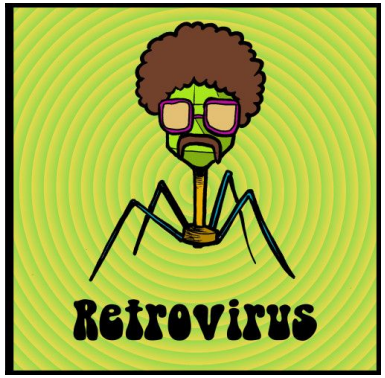
Fathers can be carriers and pass the mutation onto offspring

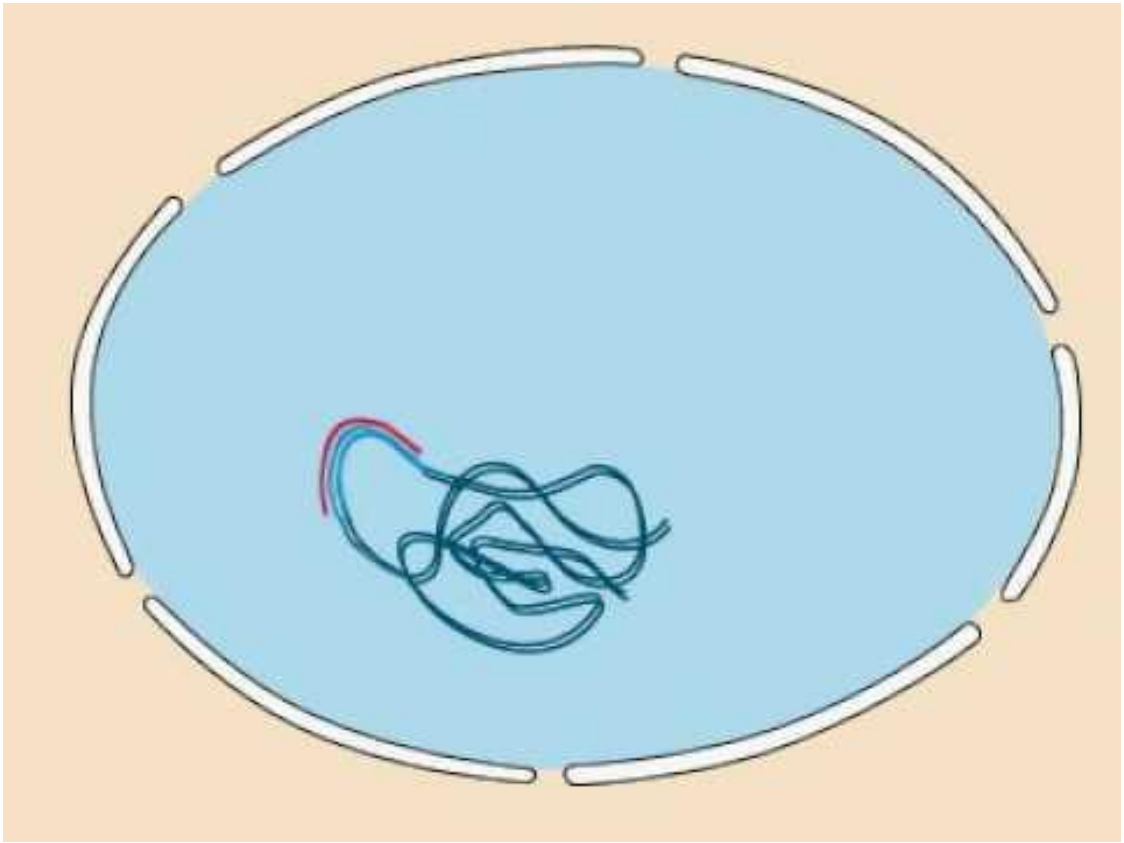
Not all people who inherit the mutation develop the disease, thus patterns of transmission are not always obvious



# Viruses and retroviruses

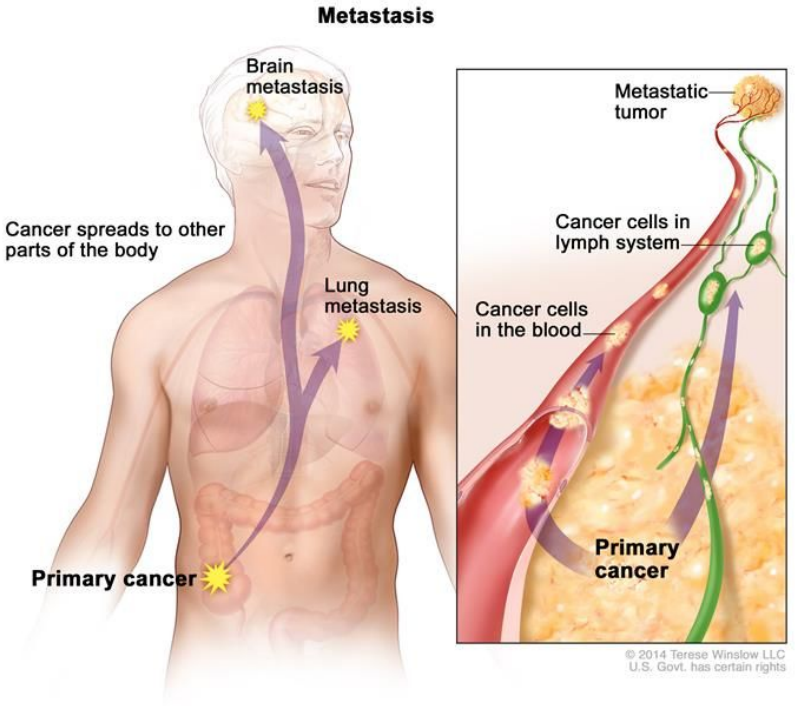


# What is metastasis?

Body's cells begin to divide without stopping and spread into surrounding tissues
Cancer cells - ignore signals that normally tell cells to stop dividing or that begin a process known as programmed cell death, or **apoptosis**, which the body uses to get rid of unneeded cells



# "Drivers" of Cancer

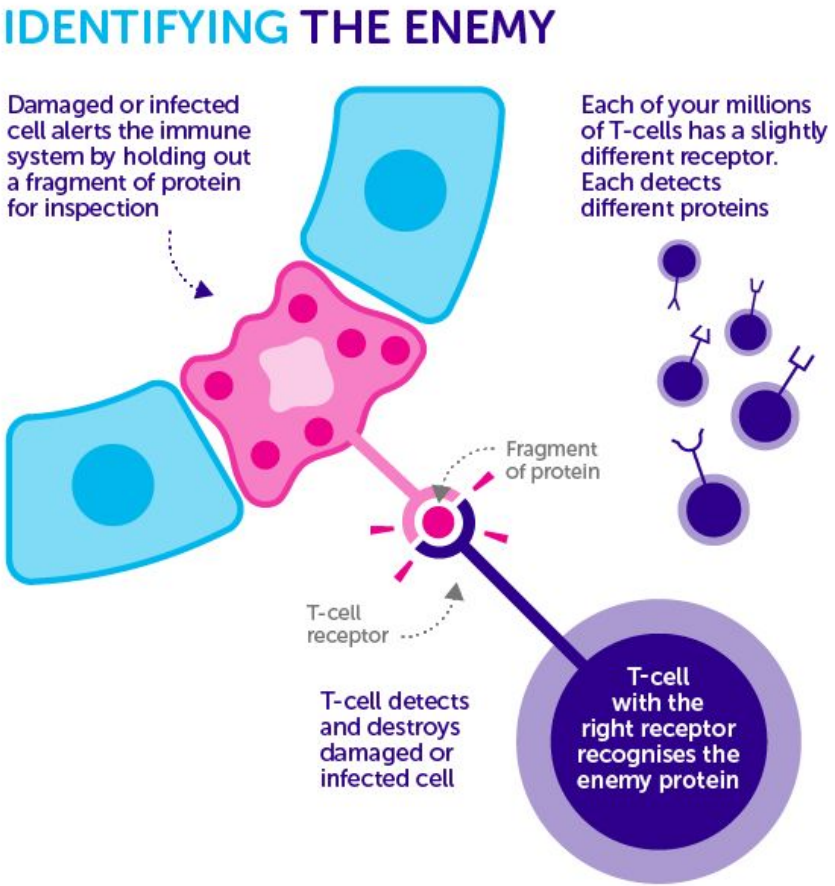
Cancer is a genetic disease that is caused by changes to genes which control the way our cells function, especially how they grow and divide:

1. **Abnormal growth (proto-oncogenes)** Cellular growth mechanism is damaged and cell starts to multiply uncontrollably
2. **Damaged control mechanism** (tumor suppressors) - Cells with certain alterations in tumor suppressor genes may divide in an uncontrolled manner (TP53 - Apoptosis)
3. **Damaged DNA repair mechanism** (Accumulated errors in this group of genes can lead to uncontrollable proliferation)

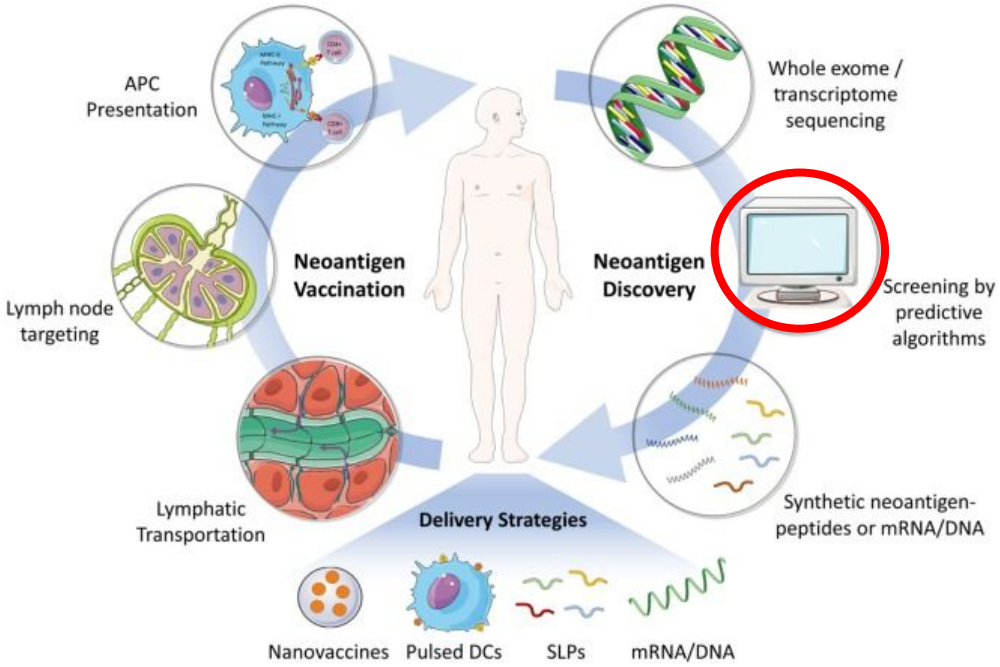
# Cancer cells

**Our body develops thousands cancer cells every day.**

**OMG! OMG! OMG!**



# Neoantigens

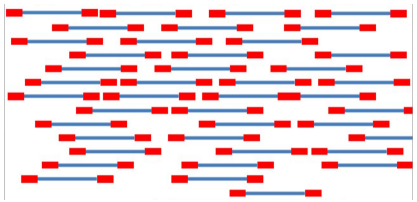


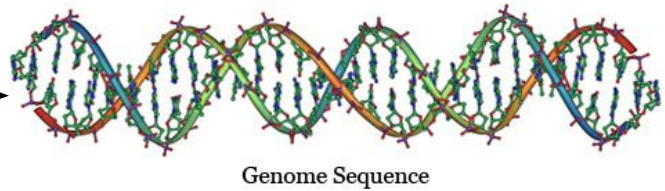

- Neoantigens - proteins presented only by cancer cells
- When neoantigen is known -> immune T-cells can be “programmed” to destroy cancer cells
- These unique cancer markers could be a key to developing a new generation of personalized, targeted cancer immunotherapies

Yugang Guo et all. Neoantigen Vaccine Delivery for Personalized Anticancer Immunotherapy

# How can we discover neoantigens?

1. Reconstruct DNA of tumor and normal tissue

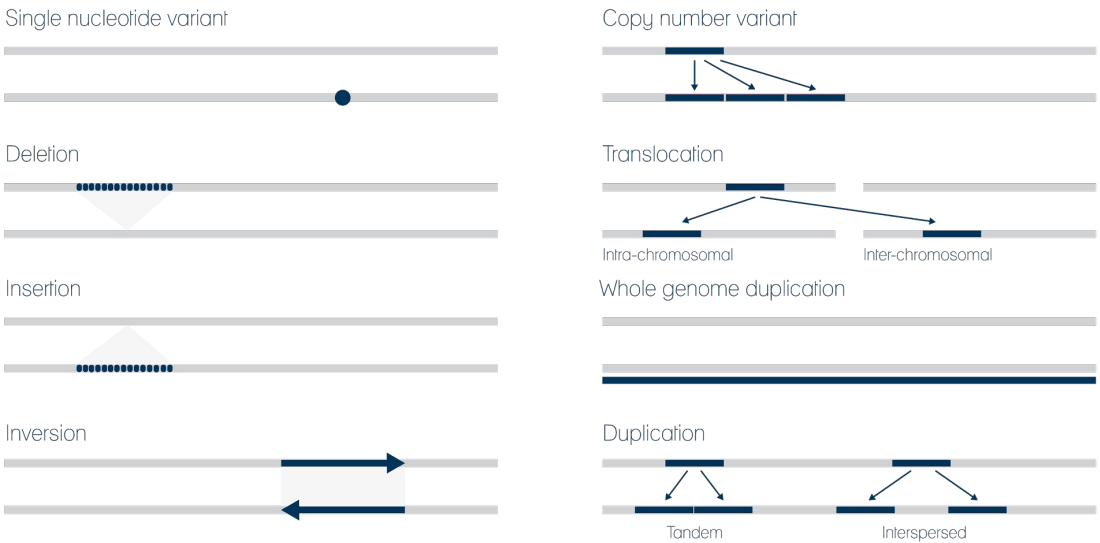




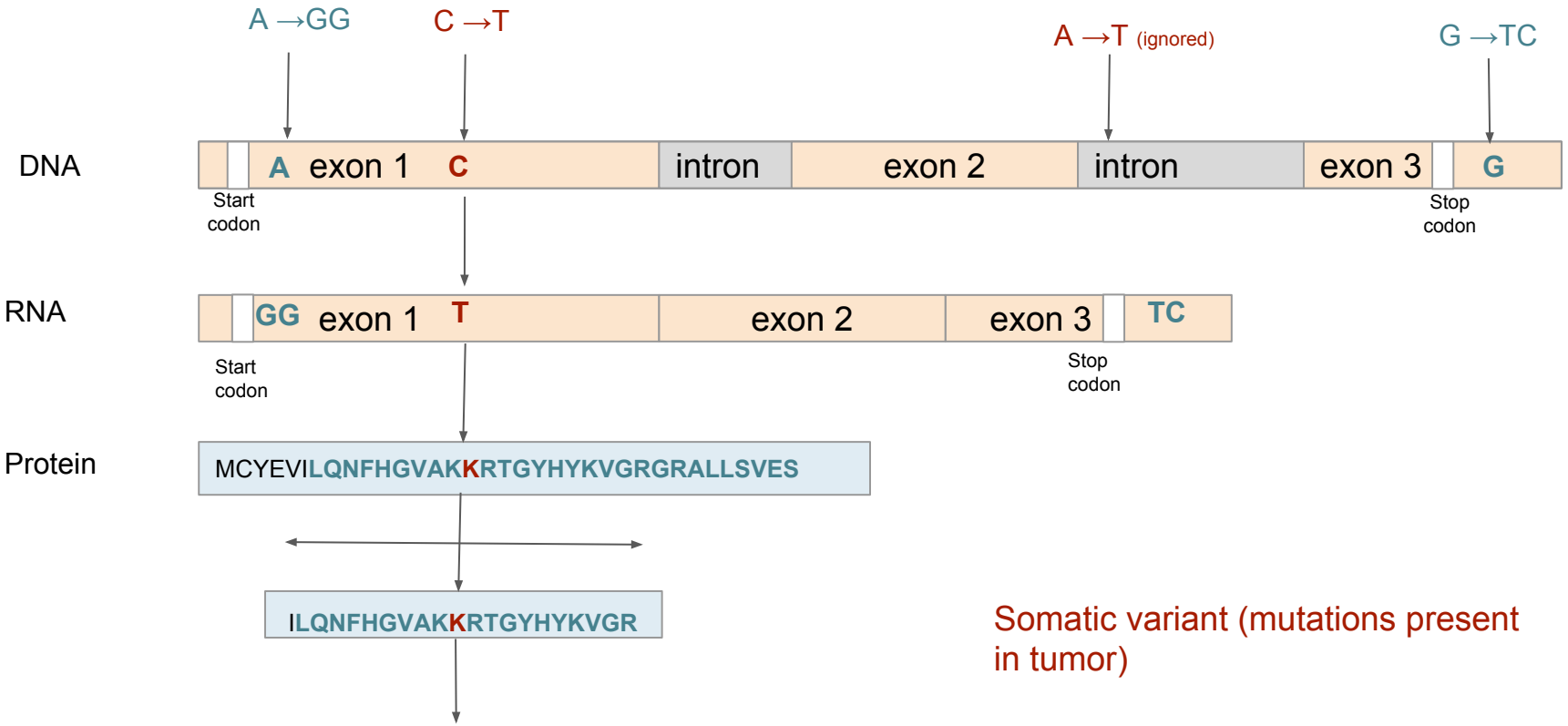
Genome Sequence

# How can we discover neoantigens?

2. Compare DNA from Tumor and Normal tissue



# How can we discover neoantigens?




Somatic variant (mutations present in tumor)

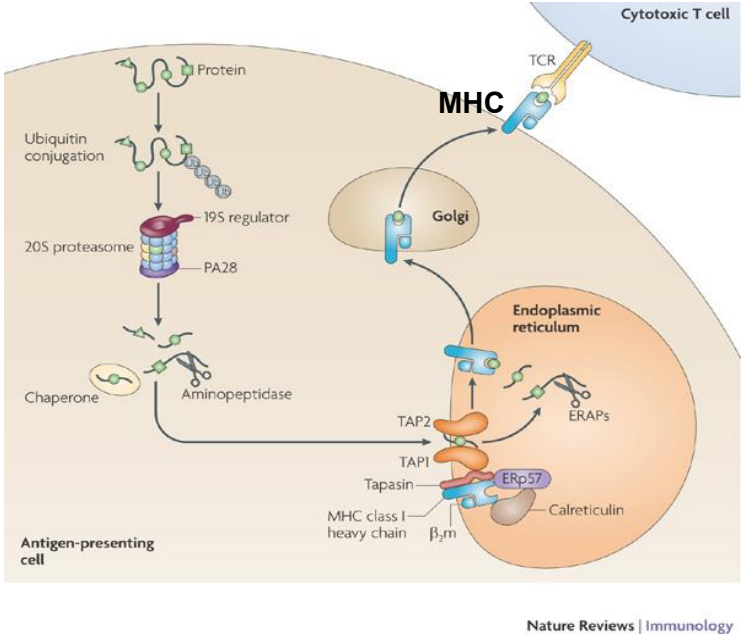
3. Protein extraction

# How can we discover neoantigens?

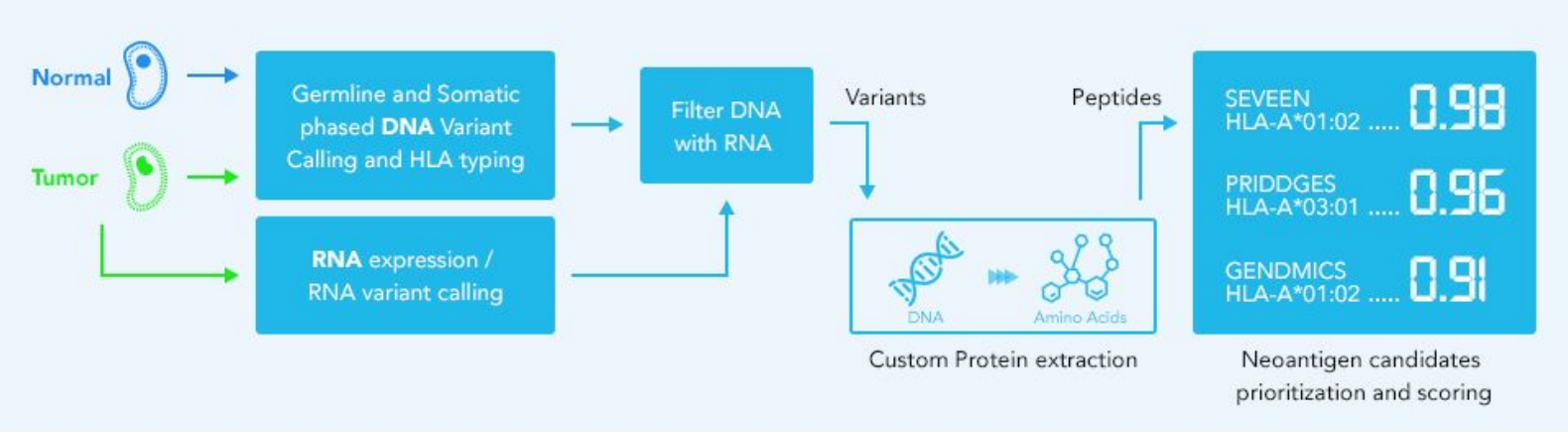
4. Discover HLA type from genome (translates to MHC molecule)

5. Perform scoring of protein-HLA sets (IEDB processing tools)

| Mutation | HLA type | peptide | NetMHC score | Pickpocket score | NetCTLPan score | RNA expression |
|---|---|---|---|---|---|---|
| 1_111957245_C_A | HLA-A*02:01 | MMLSSSPV | 0.881 | 0.633 | 1.09815 | 11.5 |
| 8_144392368_T_C | HLA-A*02:01 | WLLEKLEQL | 0.828 | 1.097 | 1.06133 | 12.5 |
| 17_28537638_C_T | HLA-A*02:01 | VLDEFPHV | 0.836 | 0.374 | 1.06015 | 23 |



# Neoantigen workflow



# References

**How the Immune System Works - Lauren Sompayrac**

JOHNS HOPKINS
WHITING SCHOOL
*of* ENGINEERING

Seven Bridges