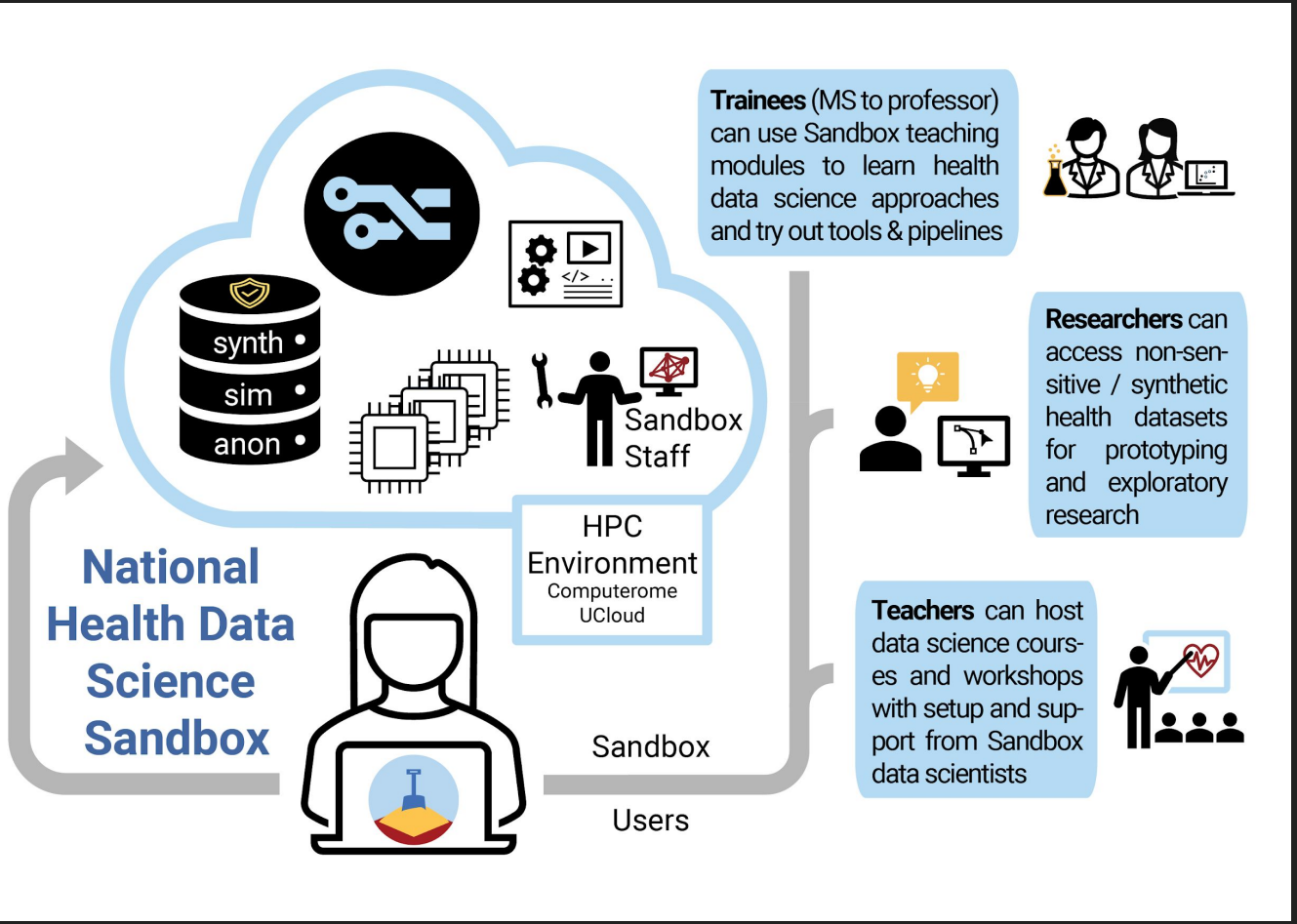
Genomics

OMICS Workshop 29.august.2023

Samuele Soraggi
Health Data Science Sandbox

Health data science sandbox

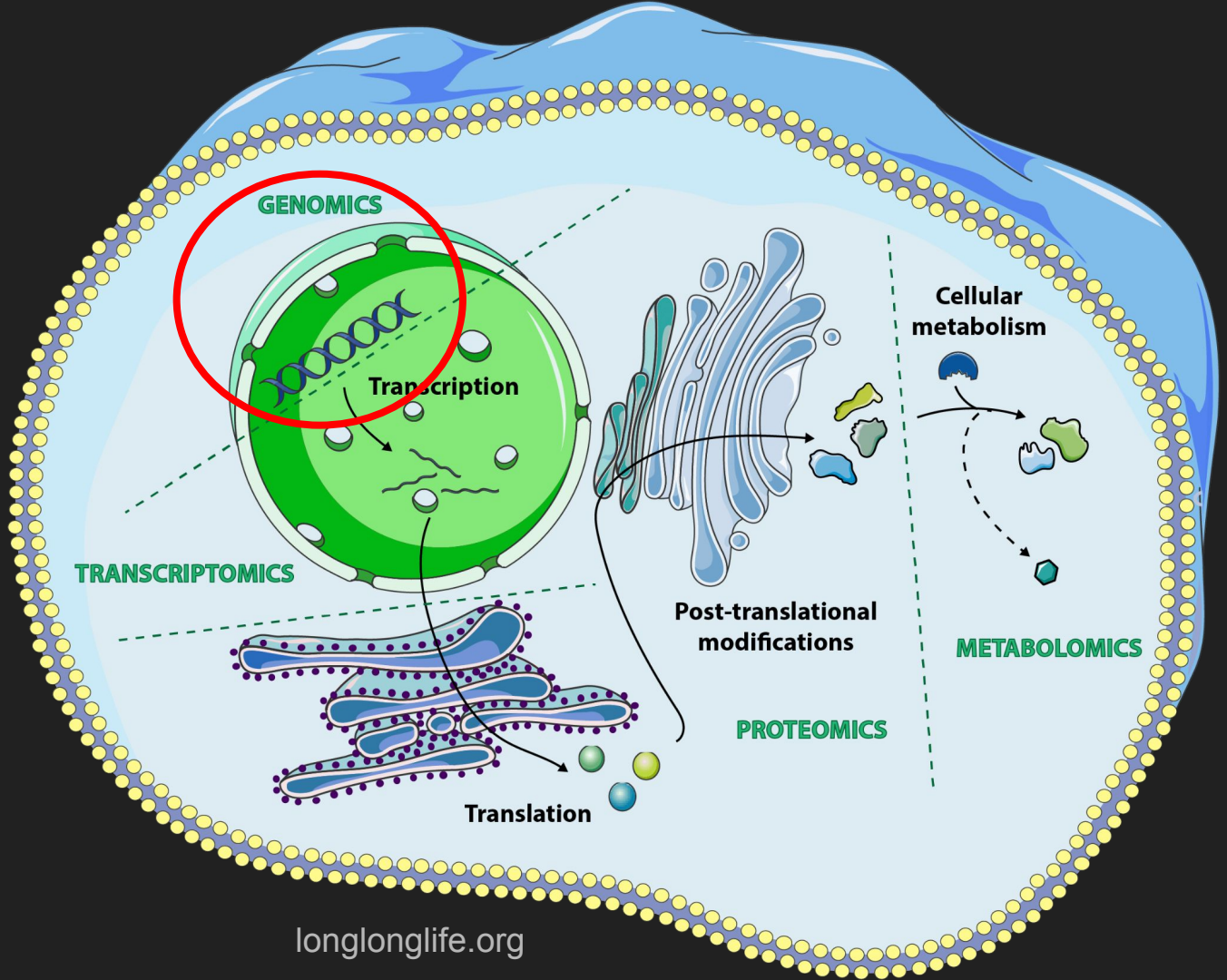


Home Page:

hds-sandbox.github.io

Contact:

samuele@birc.au.dk



Program

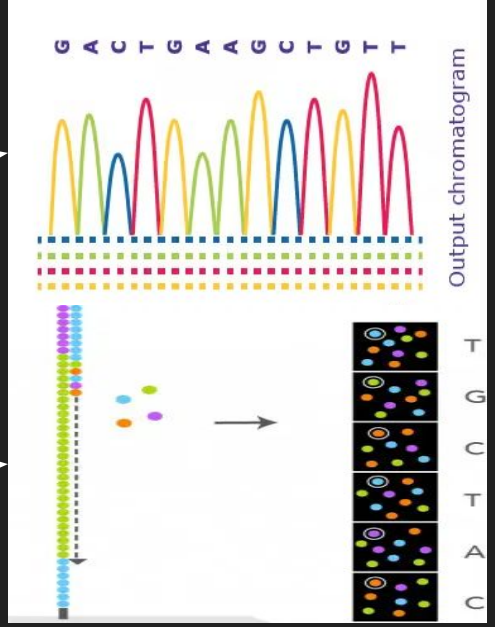
9-9.30	Introduction to genomics and tutorial format
9.30-9.45	Questions/Small break
9.45 - X	Log into uCloud, start the alignment part of the tutorial
X - X+15	Discussion of first part and questions
X+15 - 13.00	Continuing with the variants analysis tutorial (If X small enough, we use the final time for discussion again)

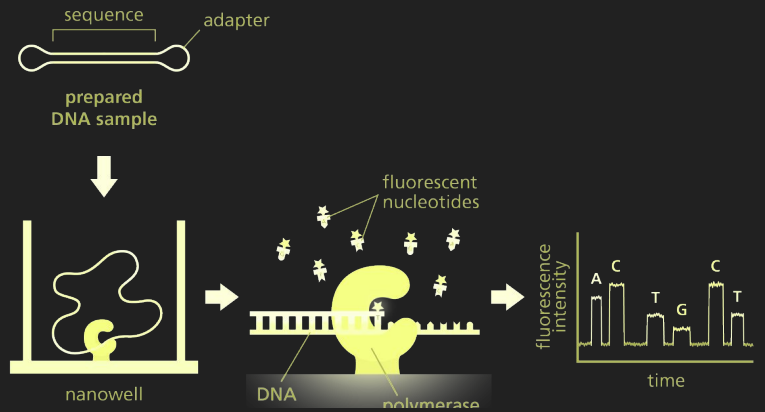
Genomics

Studying the whole DNA set to extract a variety of information from it

Genome data is generated through a variety of technologies (not limited to)

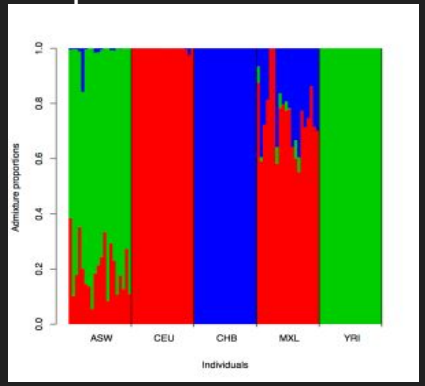
- Sanger sequencing
- Next Generation Sequencing (illumina)
- Third Gen Sequencing (Pacbio, Nanopore)



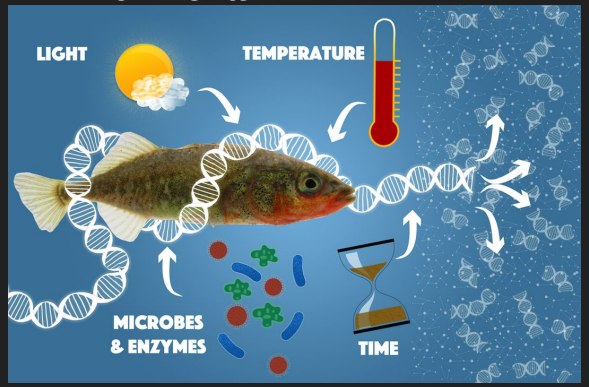


Some applications

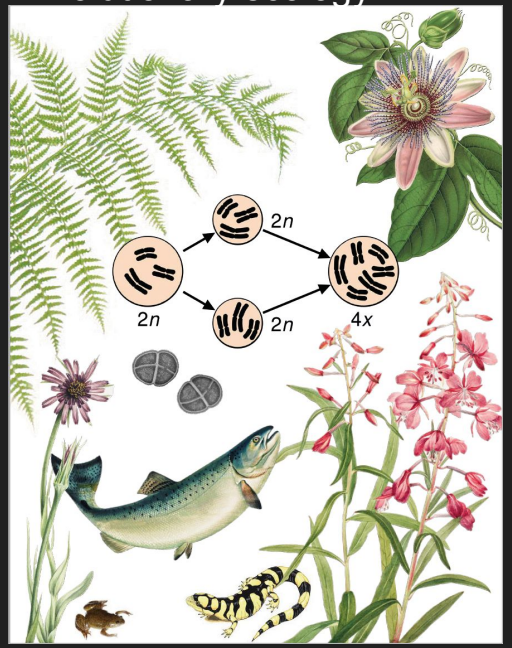
Population Structure



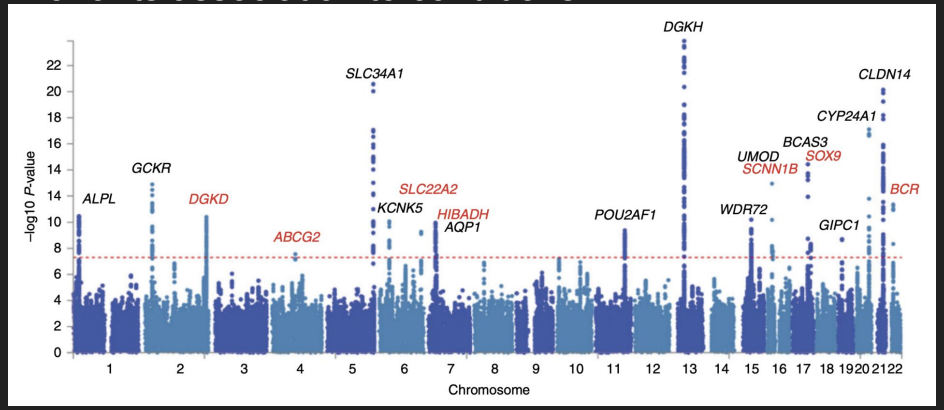
Environmental DNA



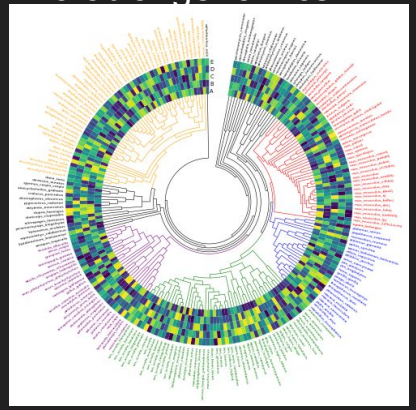
Evolutionary ecology



Variants association to conditions



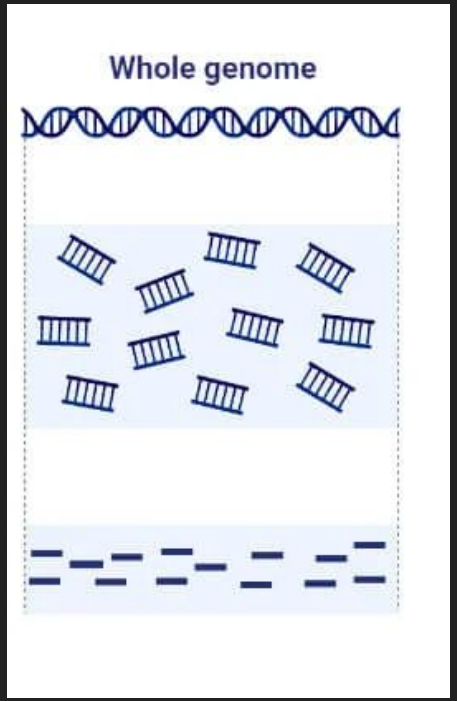
Microbial genomics

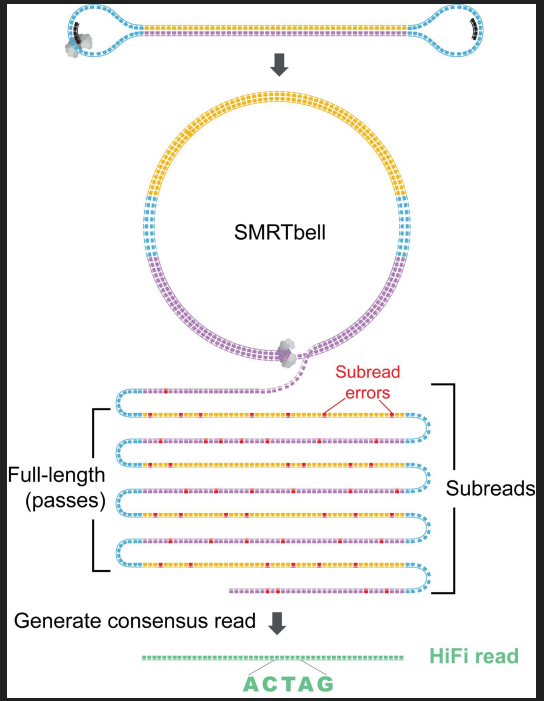


N(3)GS data

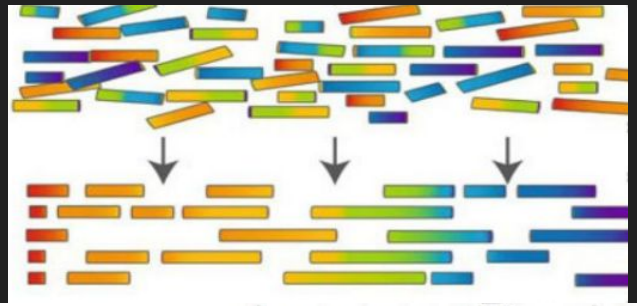
Genome sequencing data created with techniques which are

- Parallel (high throughput)
- Reasonably fast
- Cheap ($500 to $5000)
- Of varying accuracy (.1% to .000000001%)
- High depth (each position in the genome sequenced >30x on average)





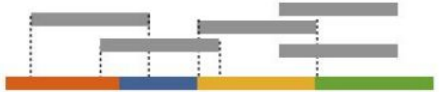
NGS data: alignment to reference



Commins J. et al, Biol Proced Online 11(1) 2015

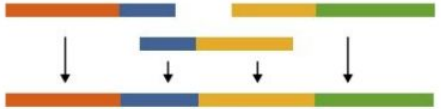
Mapping to reference sequence

Recreate the genome with using prior knowledge as reference



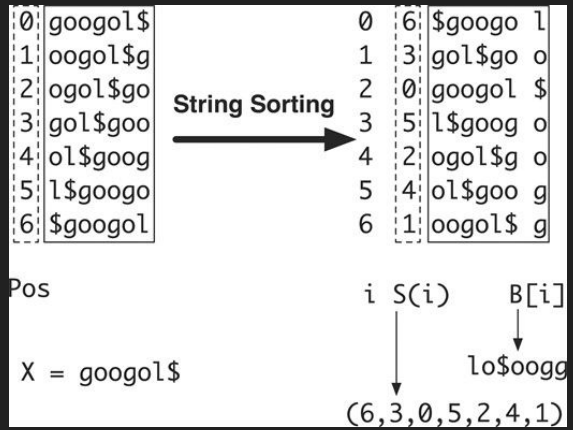
De Novo assembly

Recreate the genome with no prior knowledge



NGS data: alignment to reference - Burrows Wheeler (BW) Aligner

- Builds a BW Transform (BWT) of the reference
- Stores it as indices in fast to access hash tables
- Matches sequenced reads using the indexing doing GLOBAL alignment
- BW Aligner is very popular especially for short sequencing data. Others like Minimap2 can do long reads.
- The output is a .bam file



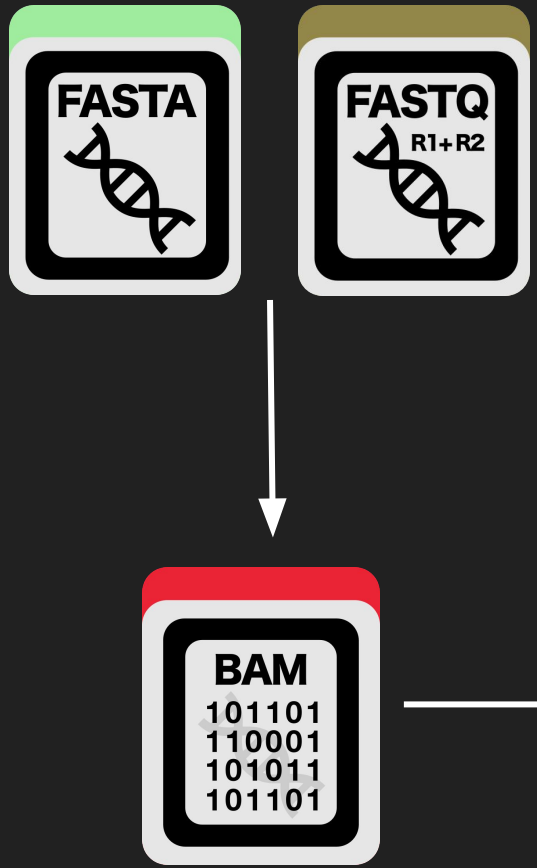

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714        16      chr20   190930  3       100M    *       0       0
      CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961        16      chr20   193953  50      100M    *       0       0
      TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030        16      chr20   270877  50      100M    *       0       0
      GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699        0       chr20   271218  50      50M4700N50M     *       0
      0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
```

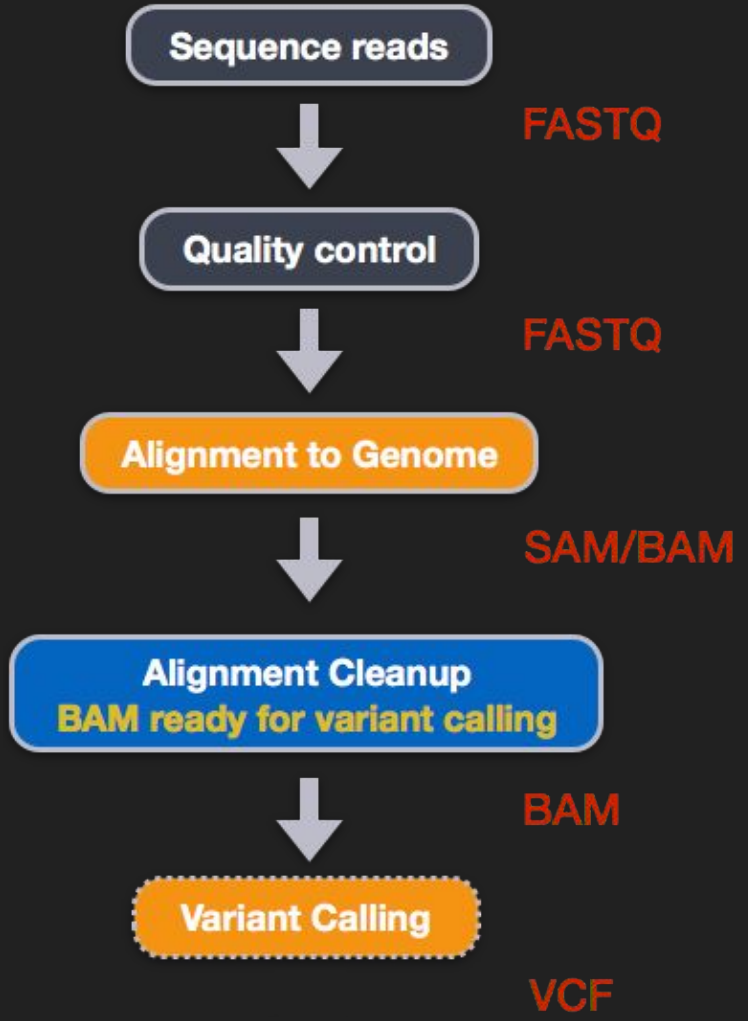
NGS data: SNP calling

- SNP: location where the aligned nucleotide is substituted (wrt reference genome)
- Function of the frequency, but also of the mapping and base quality
- SNP calling softwares such as bcftools and GATK

- Genetic profiling (personalized medicine), population genetics (common intolerances, characteristic variants, ancestries and breeding), Association studies (GWAS), Clinical Genetics

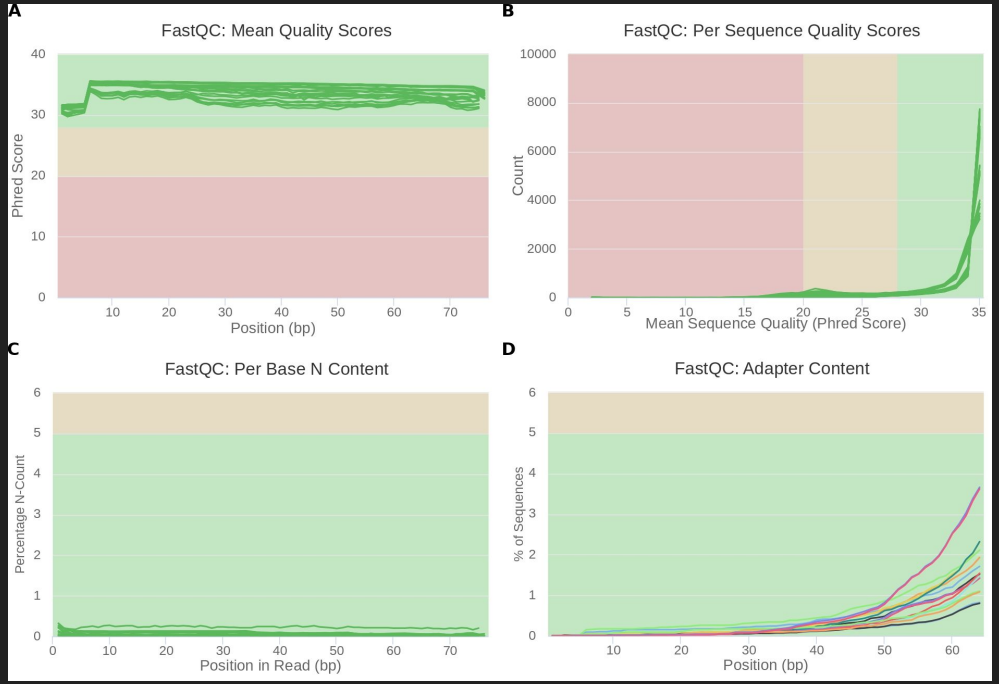
NGS data: Alignment and SNP calling file set and pipeline to create them

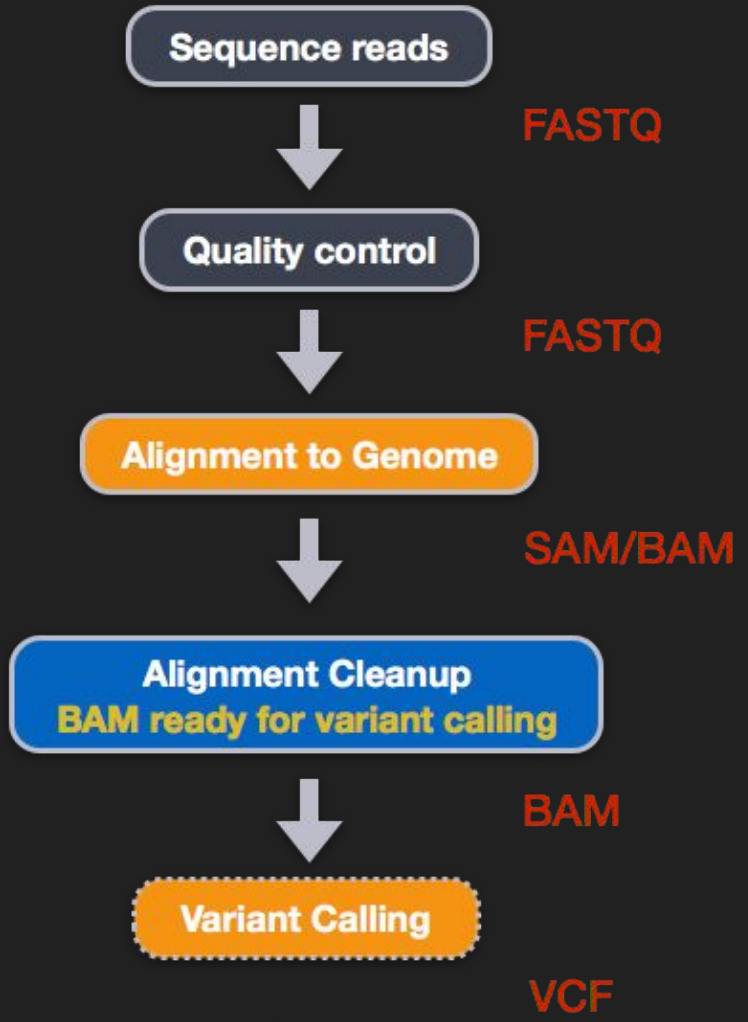




NGS data: Quality control for fastq files

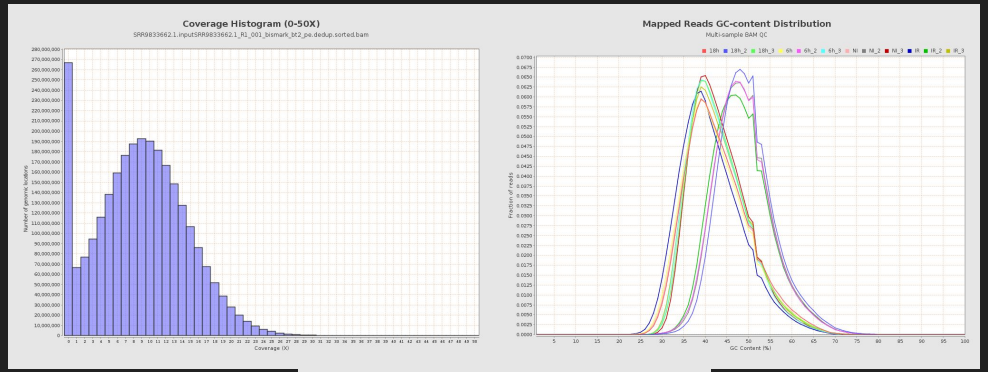
- Spot reads of bad quality
- Find out if there are missing bases in some reads
- Remaining adapters from sequencing usually need trimming

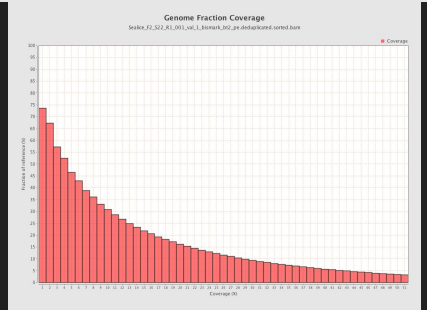


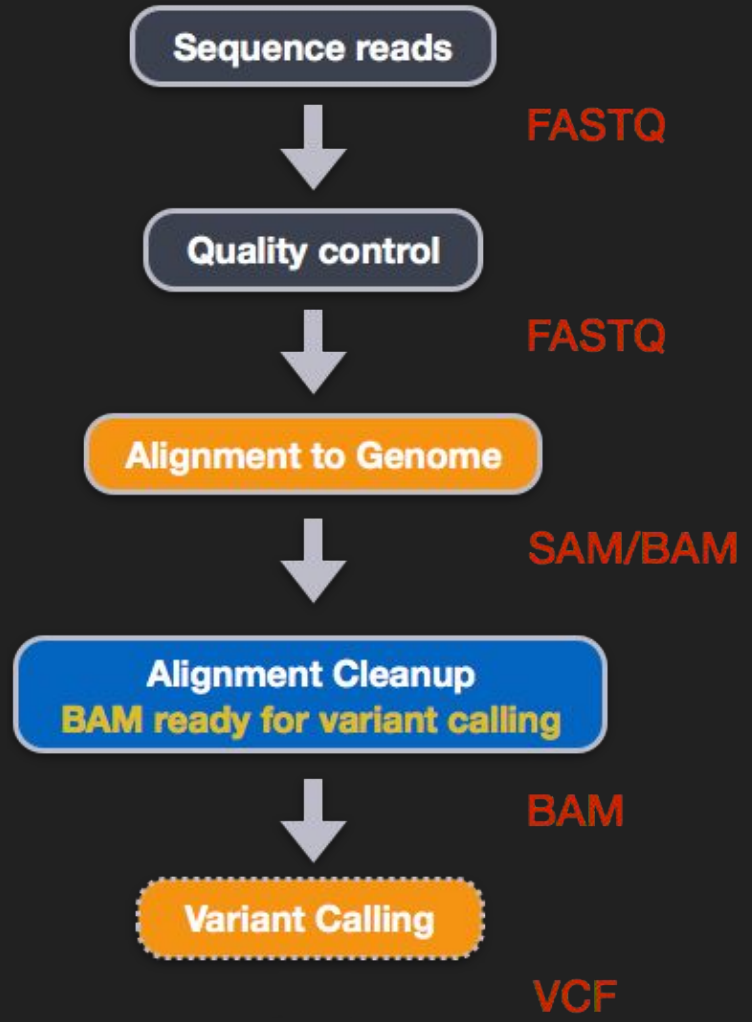


NGS data: Quality control for bam files

- Spot low coverage alignments
- Check GC content distribution
- Control how much of the genome has been covered at varying depths

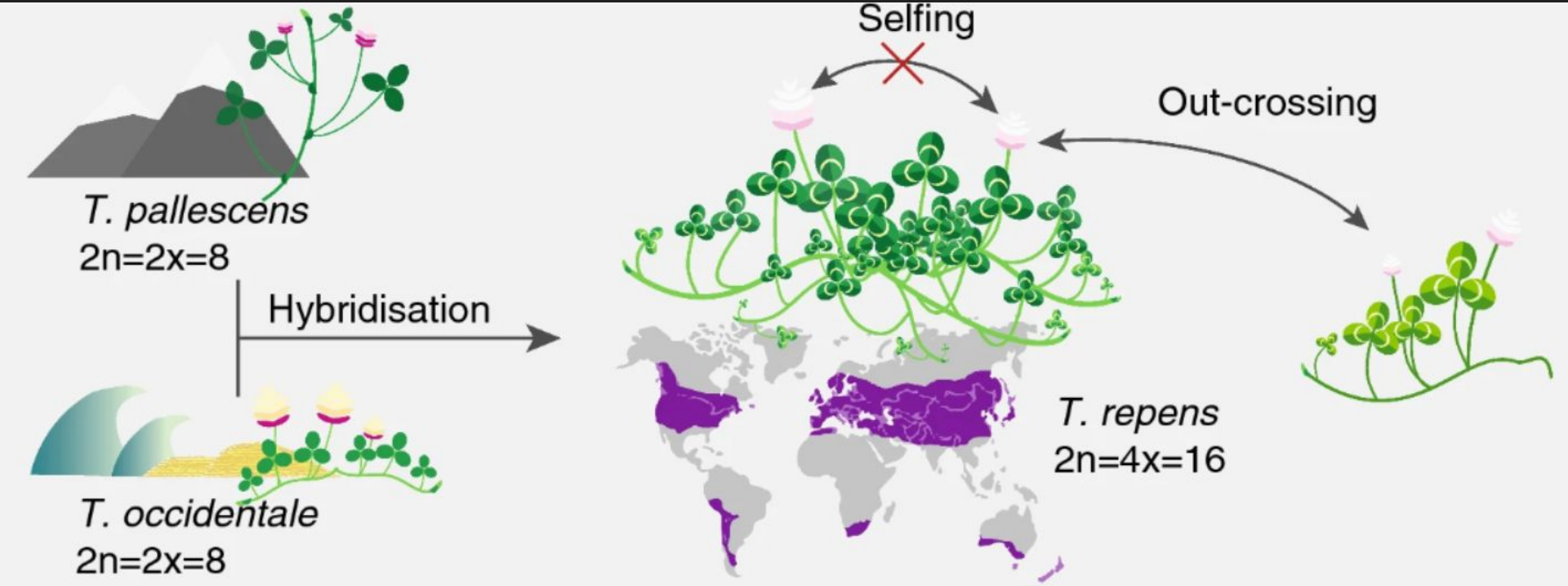




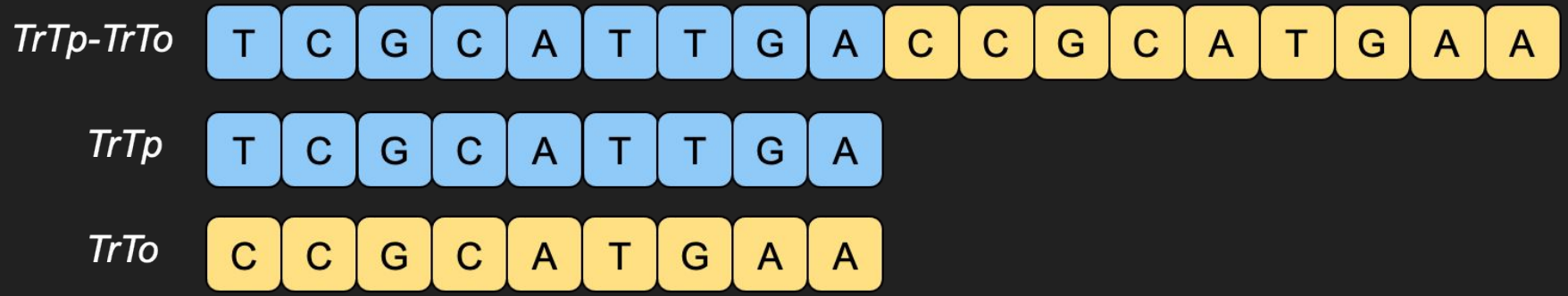


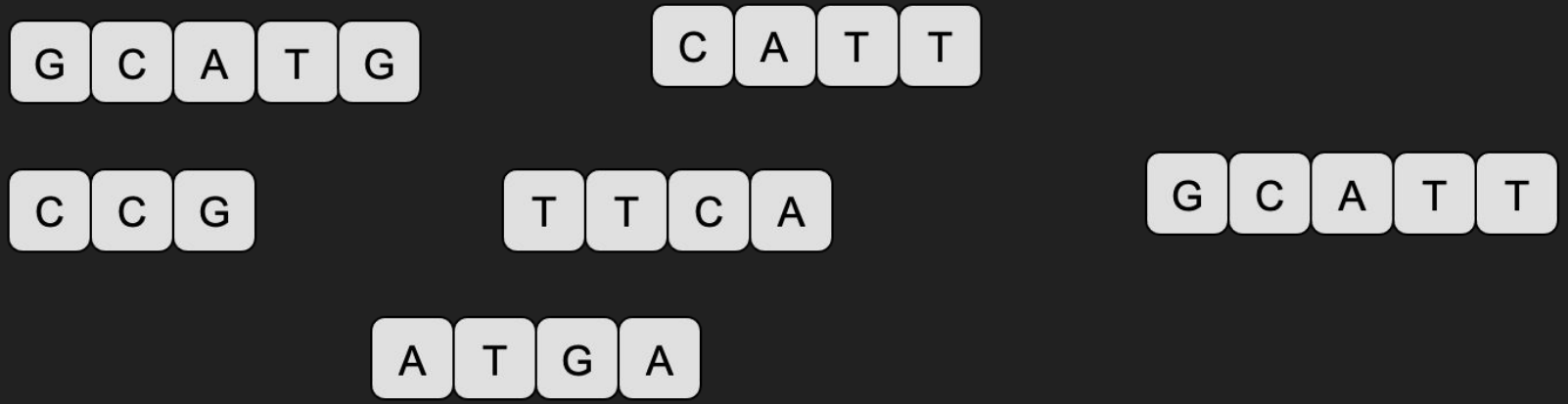
Tutorial

Raw data from white clover

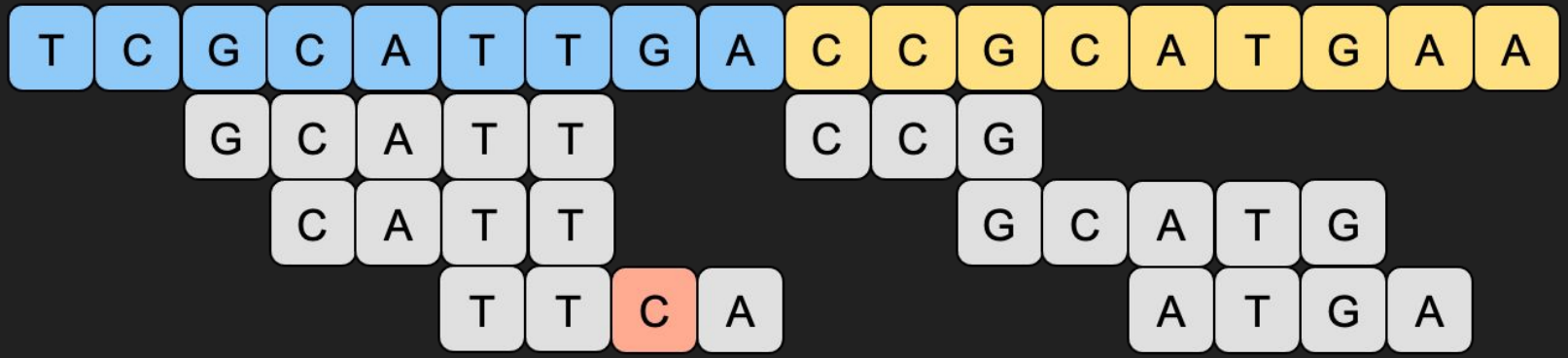


Hybridization of two species


TrTp-TrTo
T C G C A T T G A C C G C A T G A A
TrTp
T C G C A T T G A
TrTo
C C G C A T G A A


G C A T G
C A T T
C C G
T T C A
G C A T T
A T G A

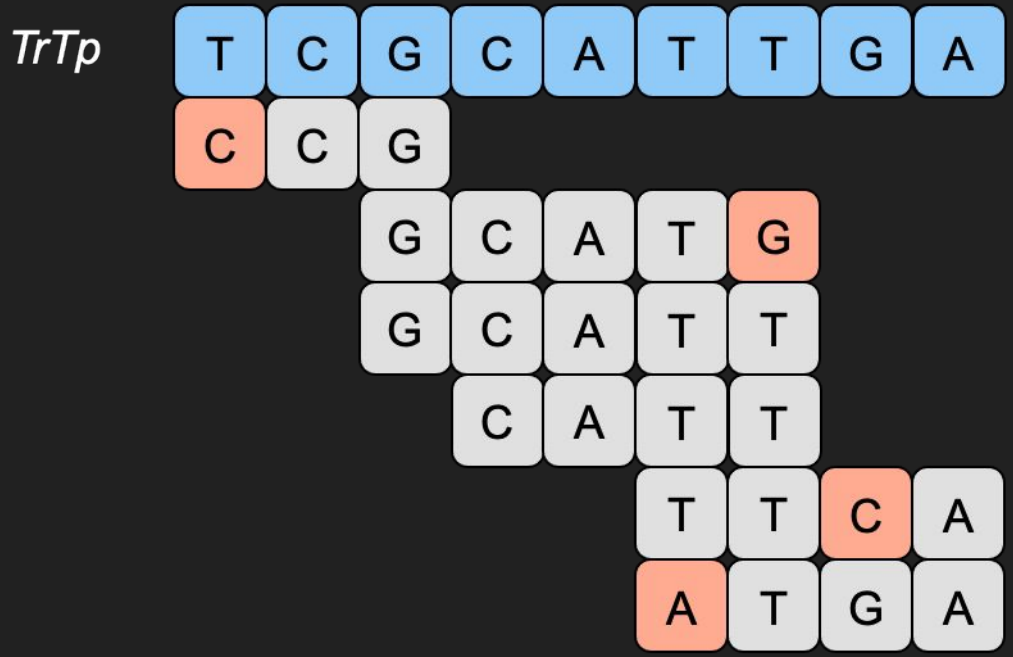
TrTp-TrTo



Nucleotides aligned: 24
Mismatches: 1

Ratio:

1 / 24 = 0.0042



Nucleotides aligned: 21
Mismatches: 4

Ratio:

4 / 21 = 0.19

- Go to https://hds-sandbox.github.io/OMICS-workshop/
- Follow the "**uCloud access**" instructions in the menu if it is your first access on uCloud
- Go on "**Day 1 - Genomics**" to follow the tutorial instructions

More material:

- We have the whole NGS summer school, a population genomics and an introductory GWAS tutorial on the **Genomics Sandbox App on uCloud**