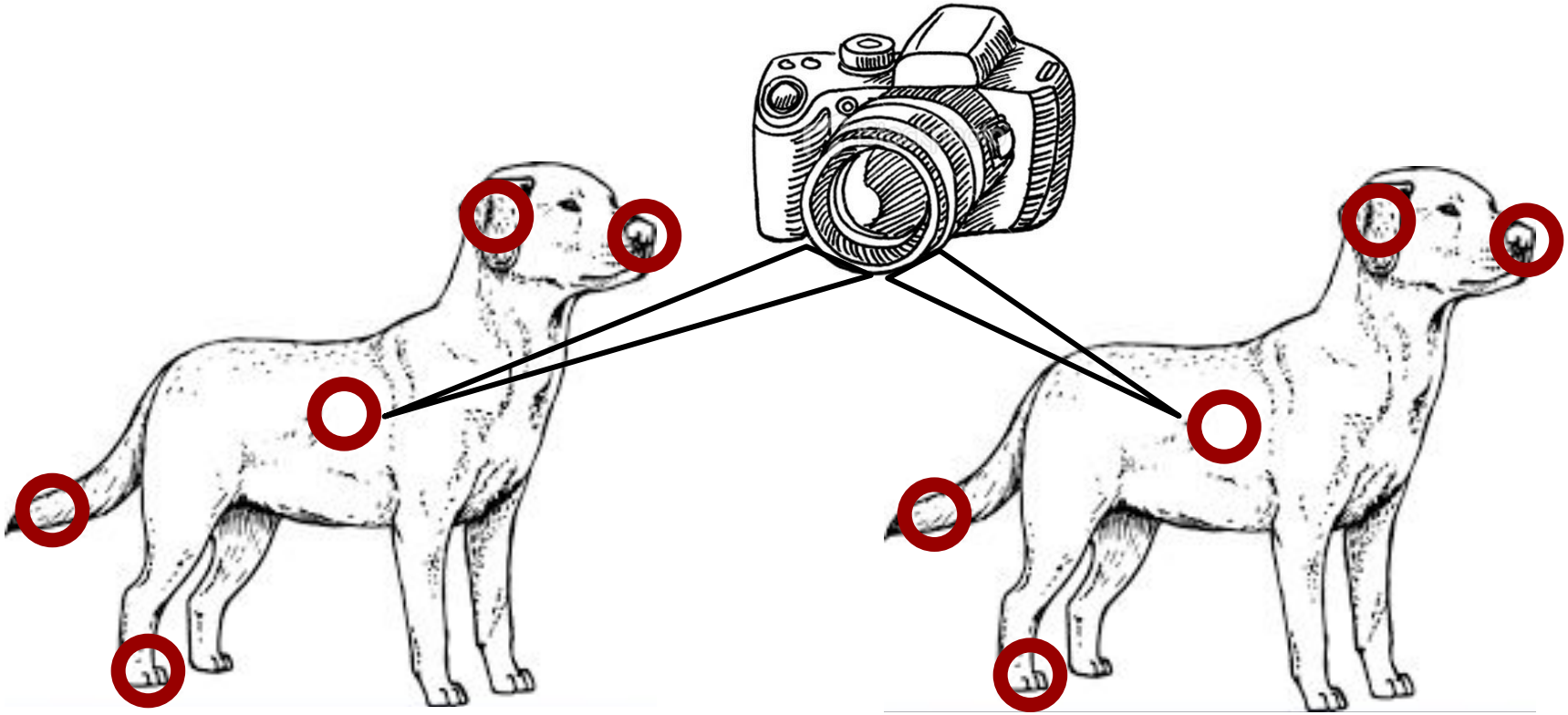
What's in a Mutt? An Intro to Dog DNA Analysis

Lecture 5
Jan 16th, 2019

So far, all our data has been SNP genotypes



Single Nucleotide Polymorphisms (SNPs)

Tasha ...ATCGGAATAGCGAGTA...

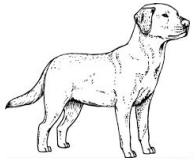
dog of interest (two copies)

G T
G G

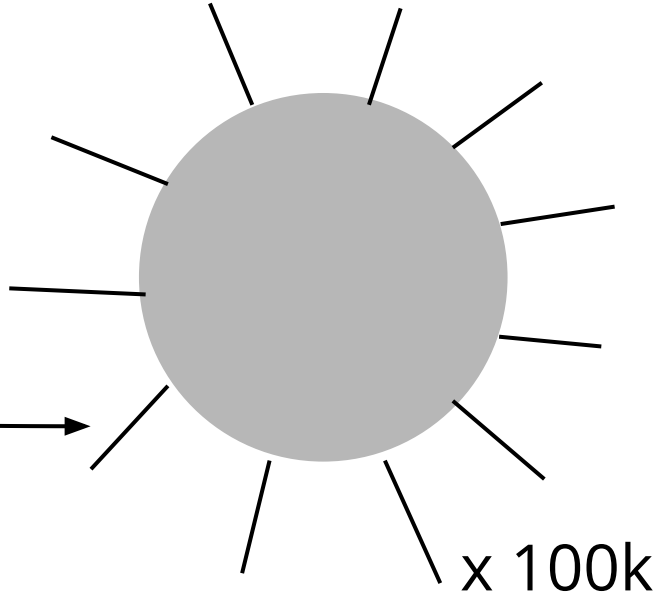
We get these SNP chips aka genotyping arrays. These ***only*** look at predetermined sites.

Illumina Bead SNP Array

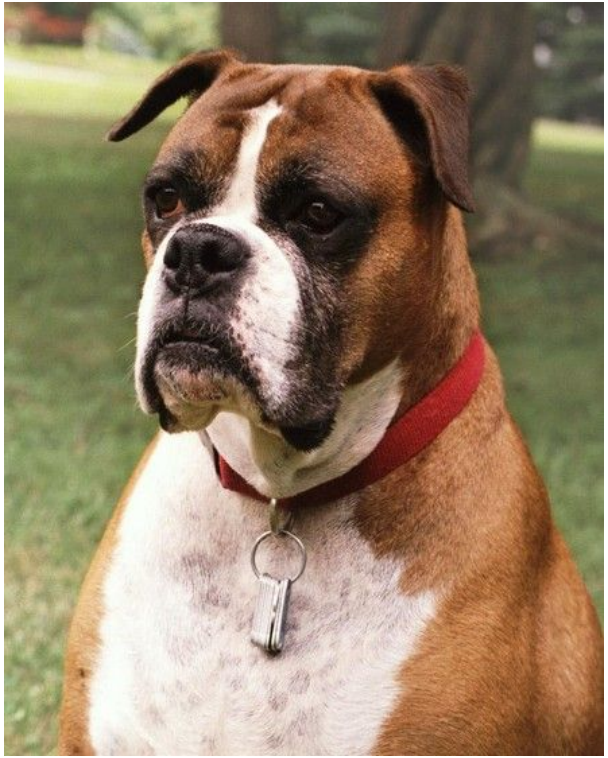
SNP genotype: ??



TACCTTATCG**A**TCAT
ATGGAATAGCTAGTA

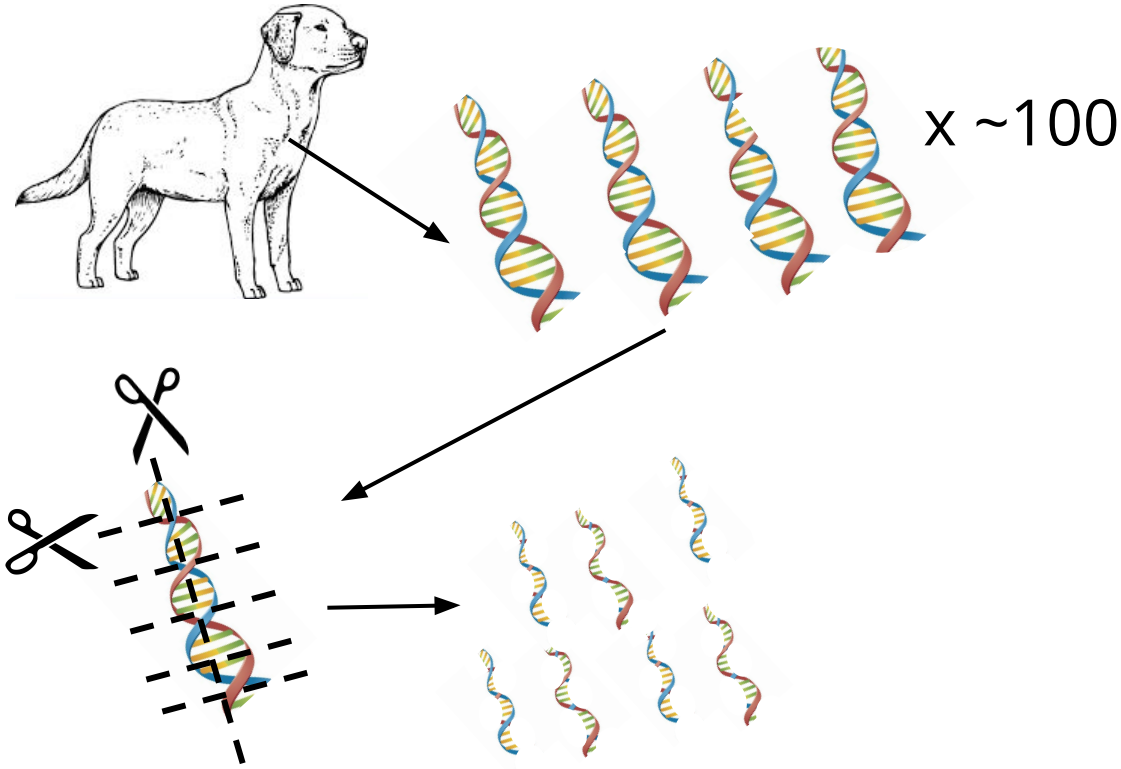


How did we get Tasha's genome?

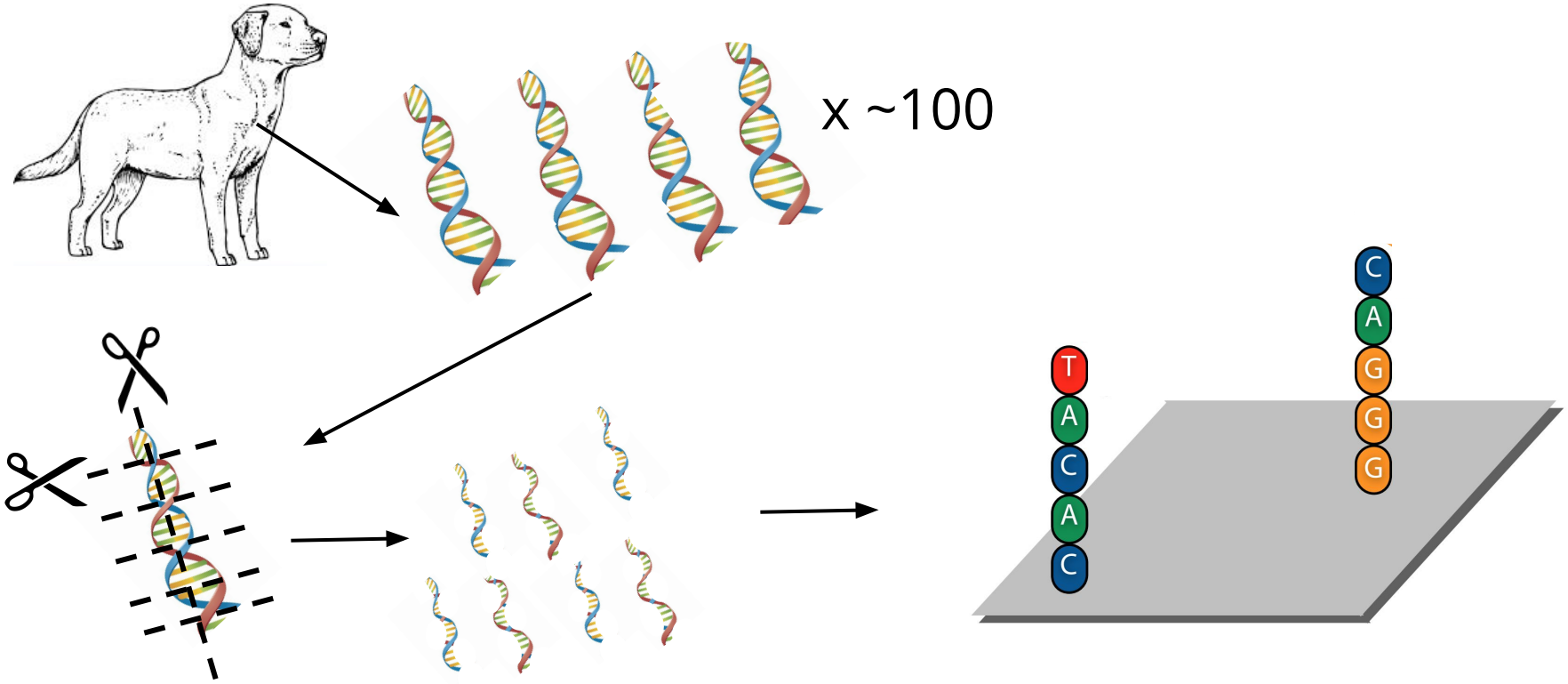


... and how do we pick sites to look for SNPs?

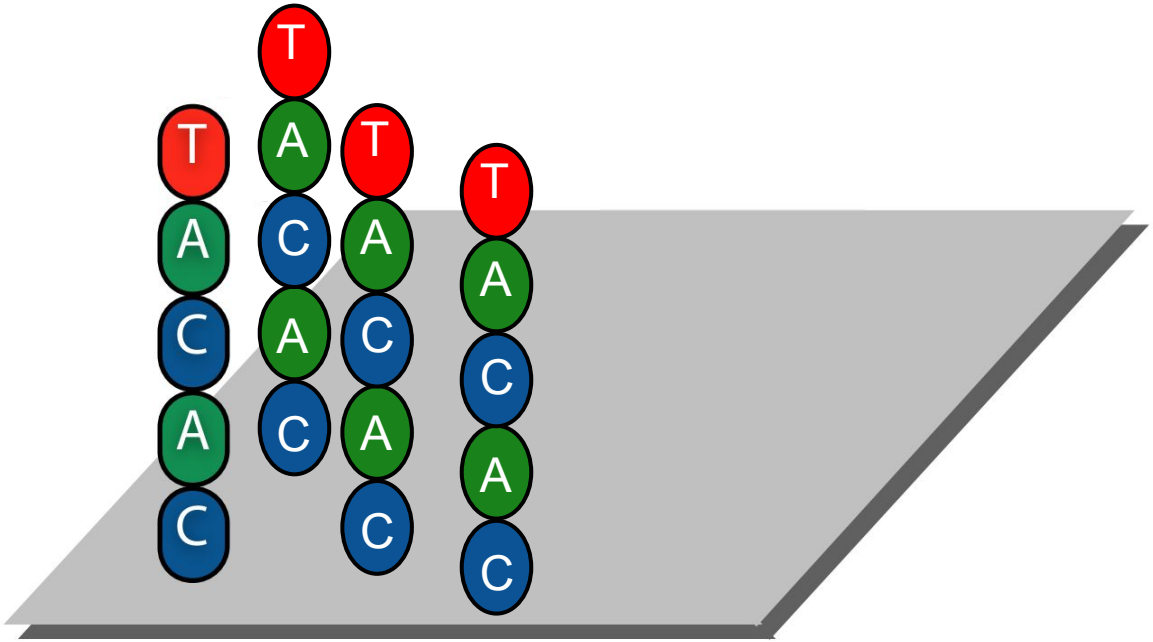
Whole genome sequencing



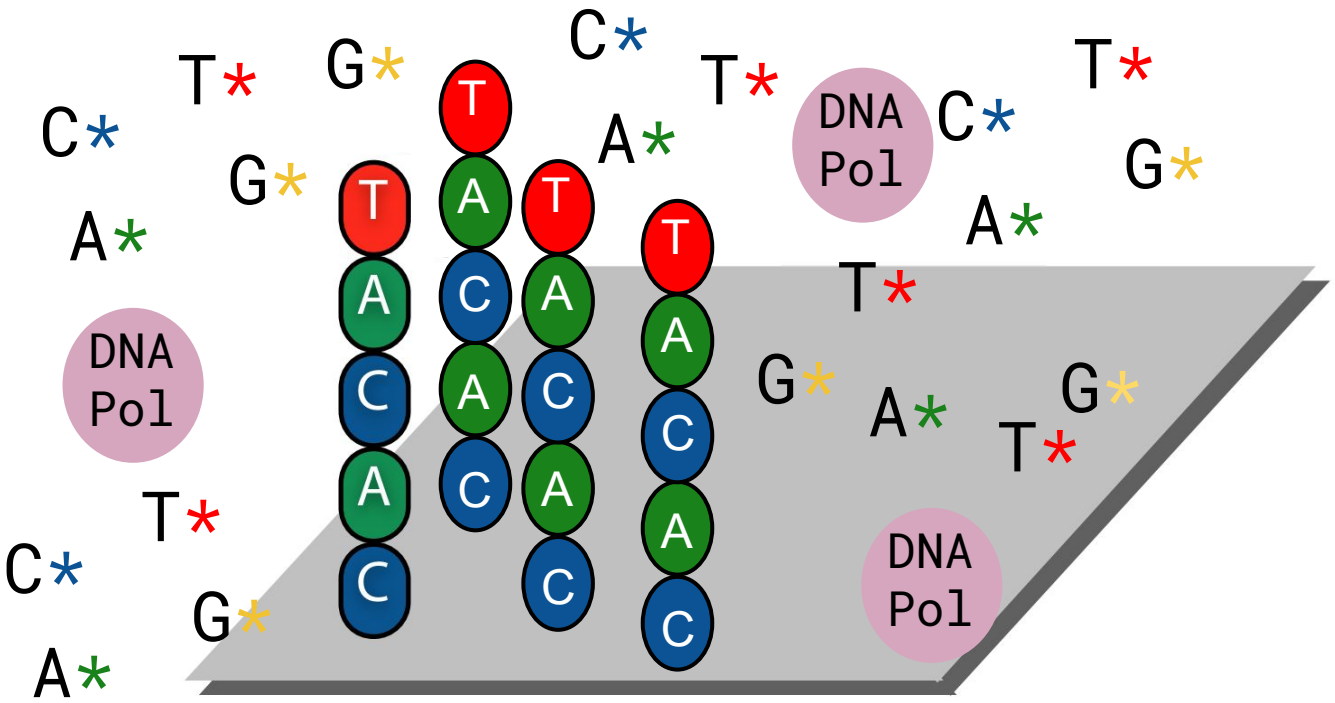
Whole genome sequencing



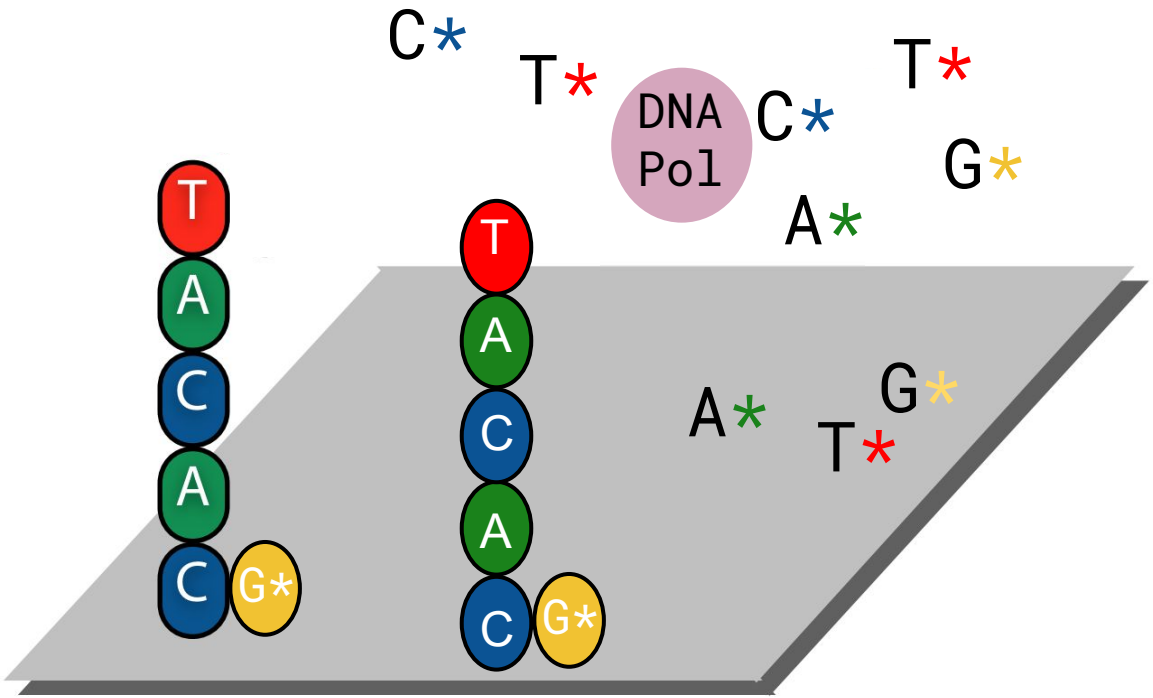
Whole genome sequencing



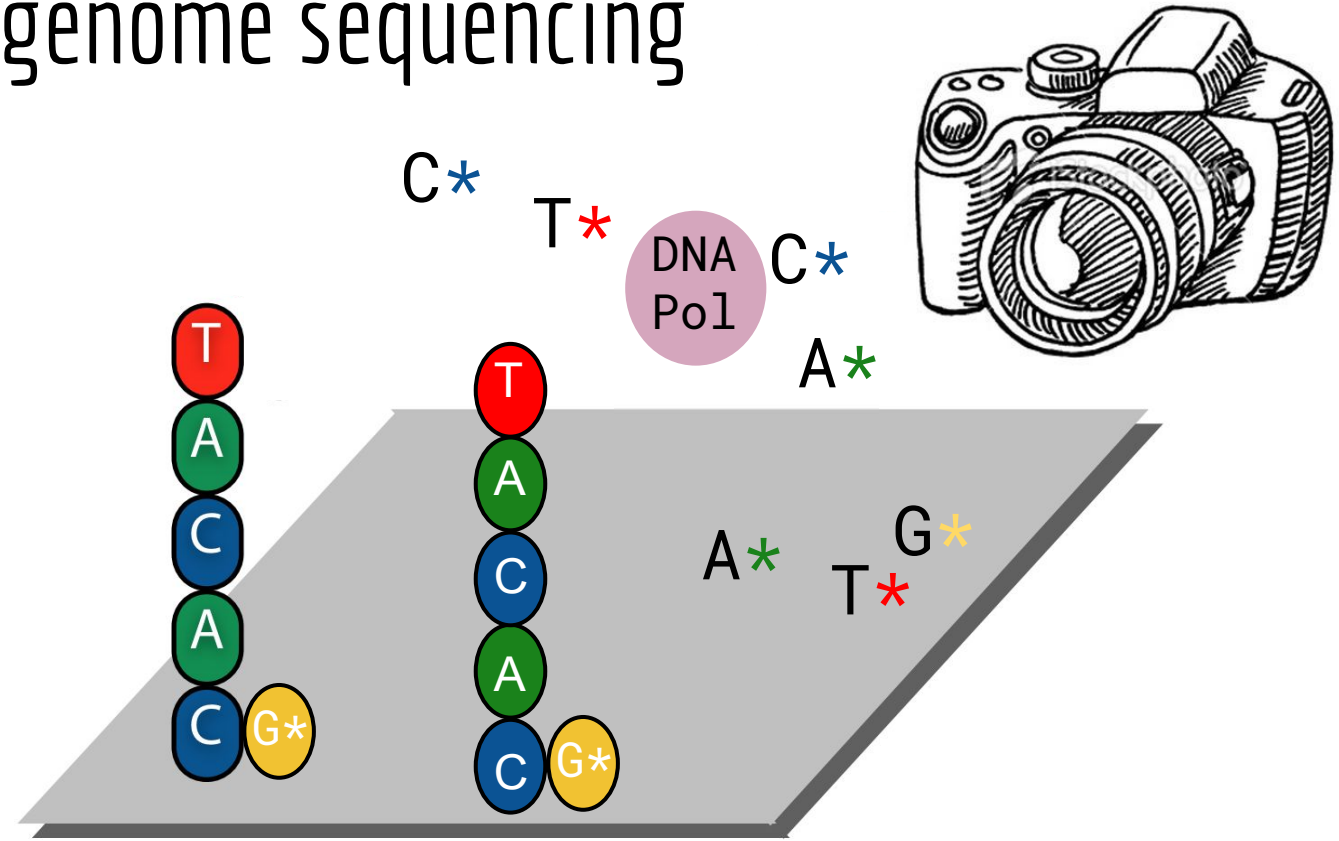
Whole genome sequencing



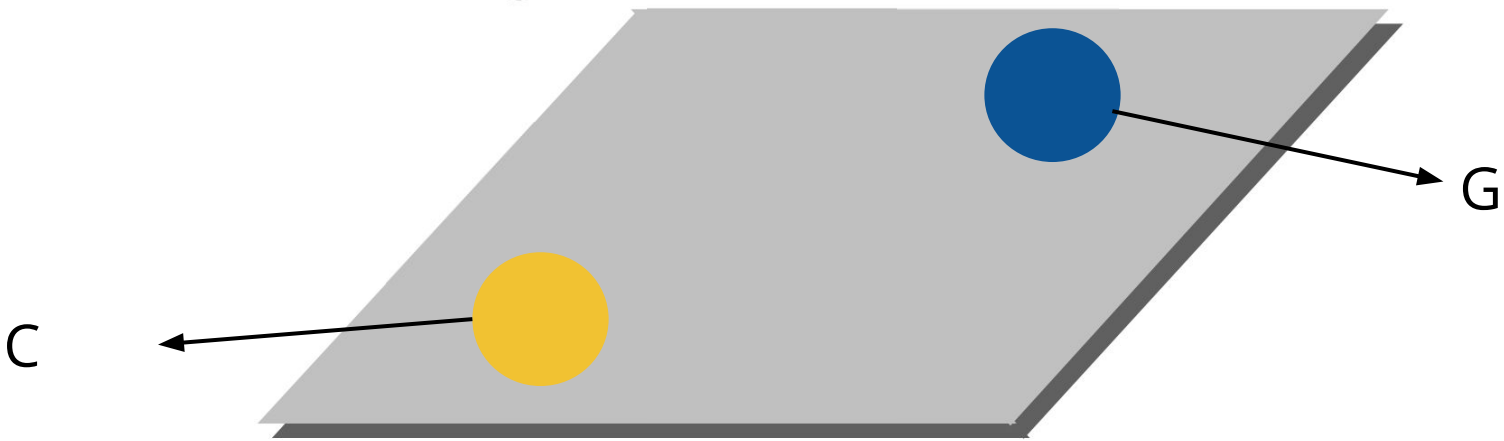
Whole genome sequencing



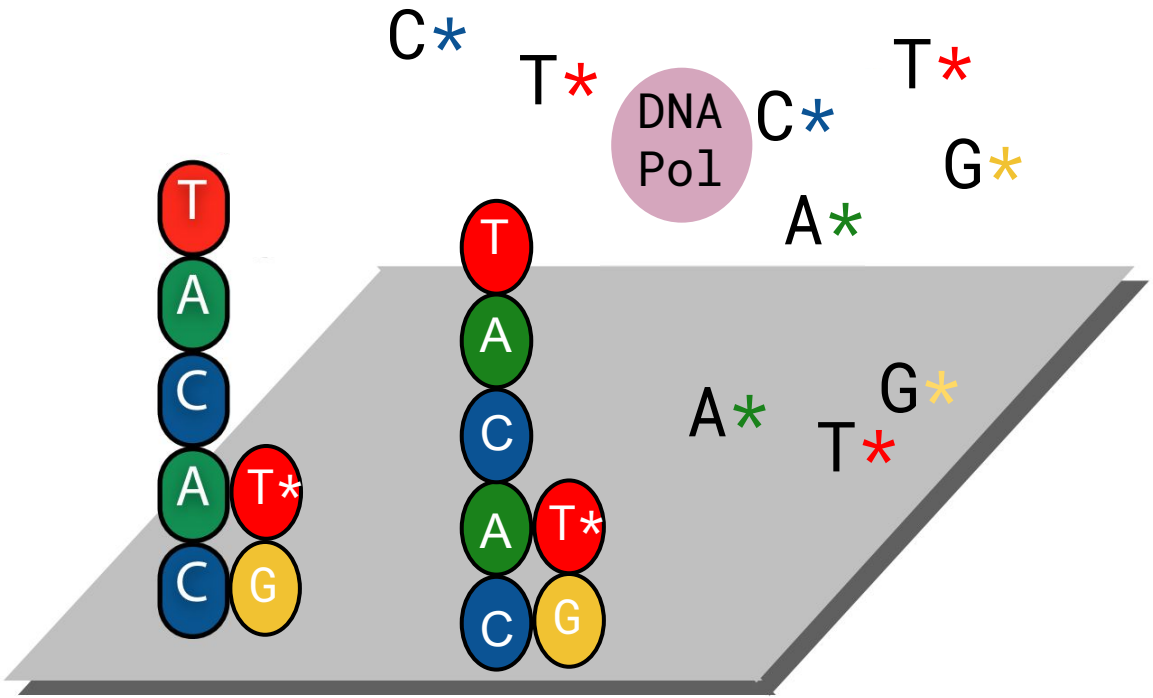
Whole genome sequencing



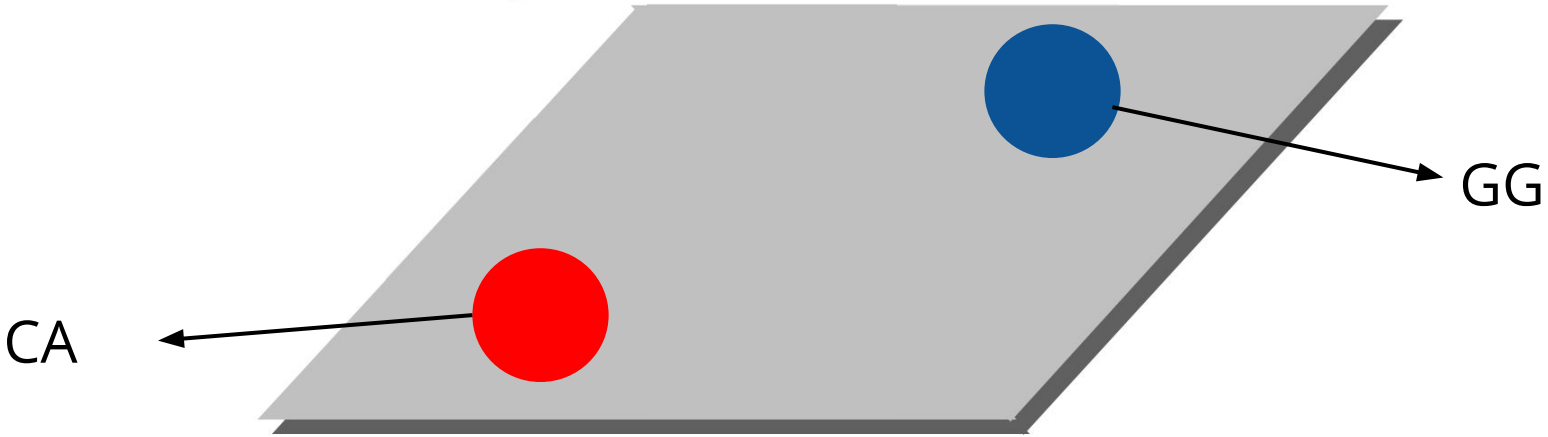
Whole genome sequencing



Whole genome sequencing



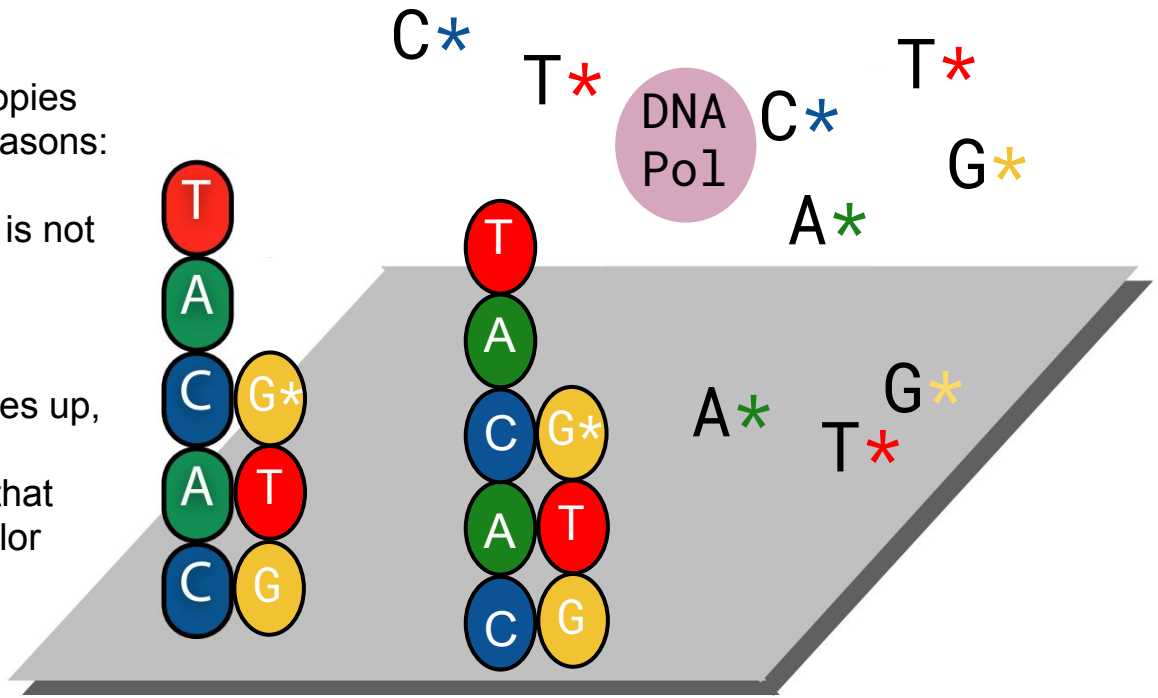
Whole genome sequencing



Whole genome sequencing

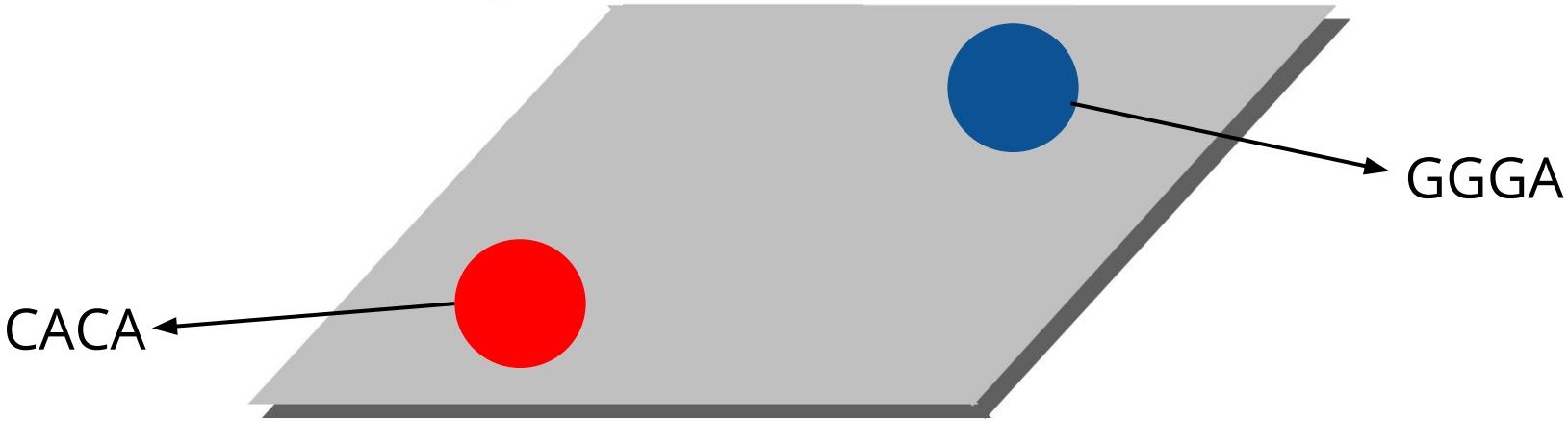
We have many copies of each for two reasons:

1. Light signal is not detectable otherwise
2. If one messes up, the others overpower that incorrect color

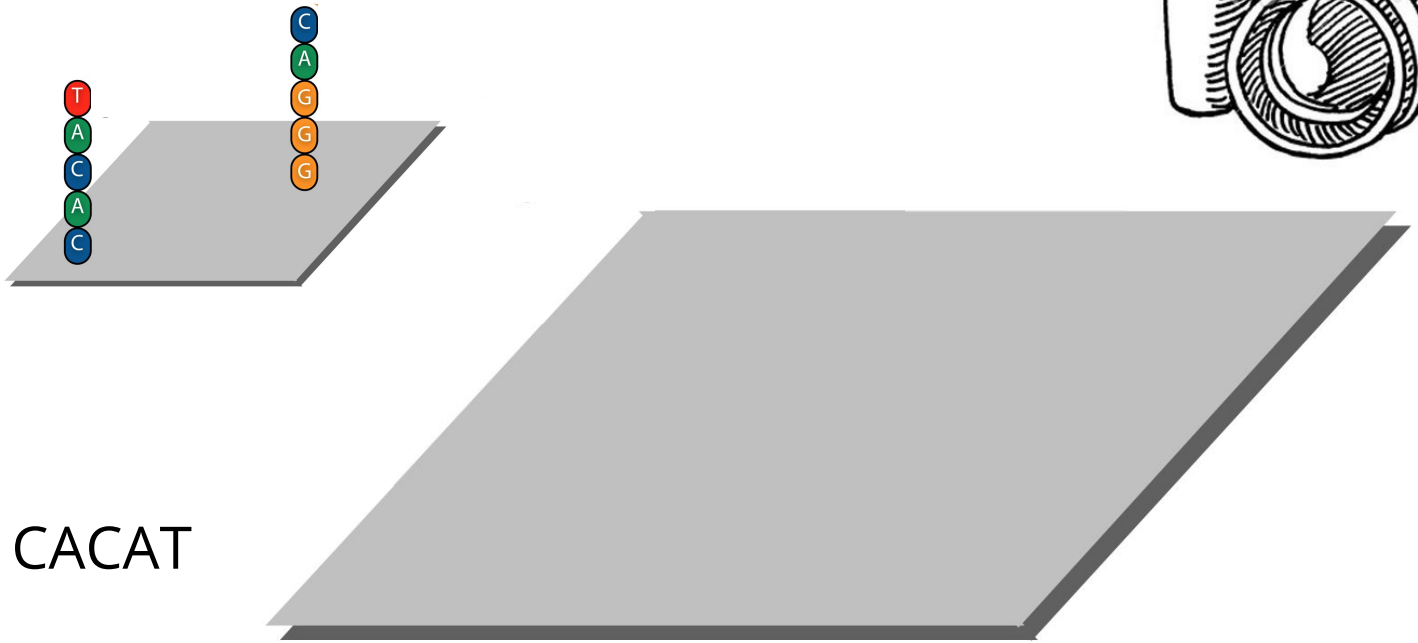


Whole genome sequencing

Whole genome sequencing



Whole genome sequencing

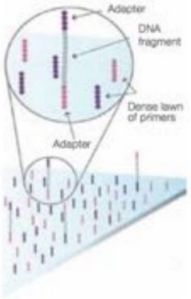


GGGAC

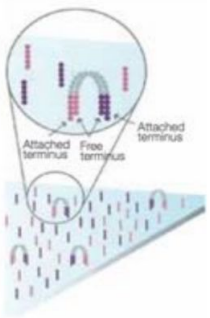
CACAT

Second generation sequencing

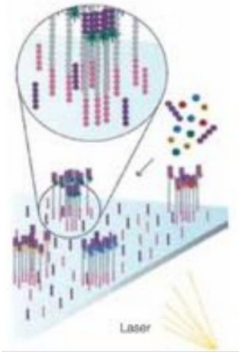
Illumina HiSeq 2000
Sequencing by Synthesis



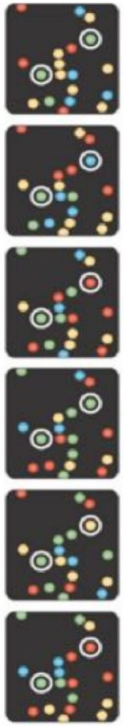
1. Attach



2. Amplify



3. Image



Metzker (2010) Nature Reviews Genetics 11:31-46

Whole genome sequencing data

GGGAC...

CATTT...

CACAT...

ATAAG...

CATGT...

CAAAG...

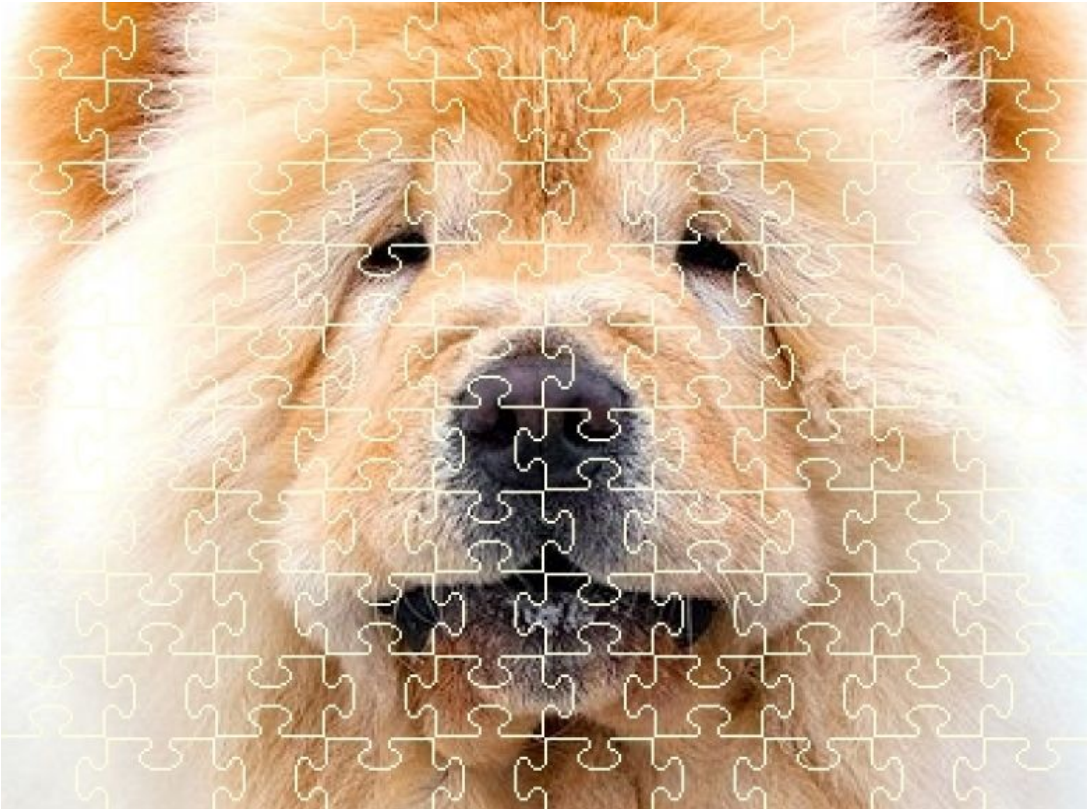
ATGAT...

...

We end up with billions of short sequences, called ***reads***, which are each ~250 bases long.

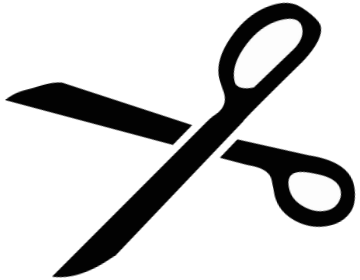
We don't know where each came from in the genome, and some of them came from roughly the same place since we had ~100 copies of the genome to start with.

Assembling a genome



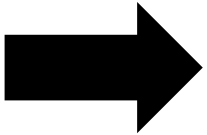
Assembling a genome

GATTACCGCA
GATTACCGCA
GATTACCGCA
GATTACCGCA



Assembling a genome

GATTACCGCA
GATTACCGCA
GATTACCGCA
GATTACCGCA



GATT ACCG
TTAC
CGCA
CCGC
ATTA
GATT
CGCA

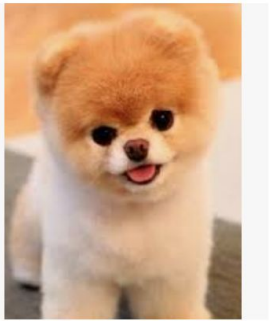
Assembling a genome

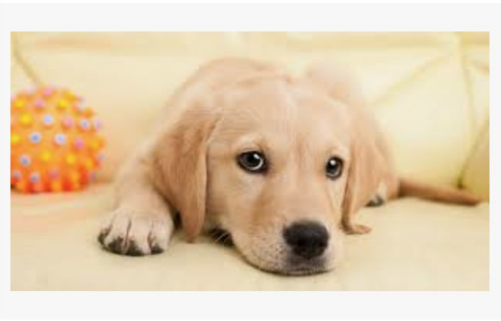
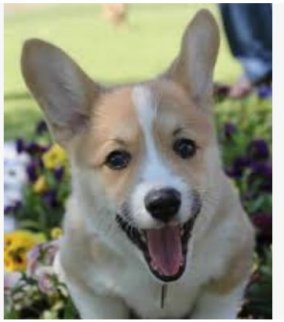
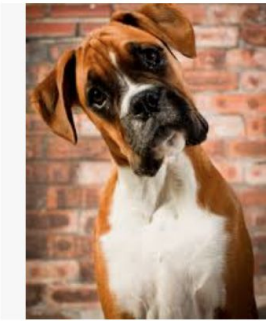
```
GATT    ACCG
                TTAC
 CGCA
        CCGC
                ATTA
GATT
                CGCA
```

```
GATT
 ATTA
  TTAC
    ACCG
     CCGC
      CGCA
----------------------
GATTACCGCA
```

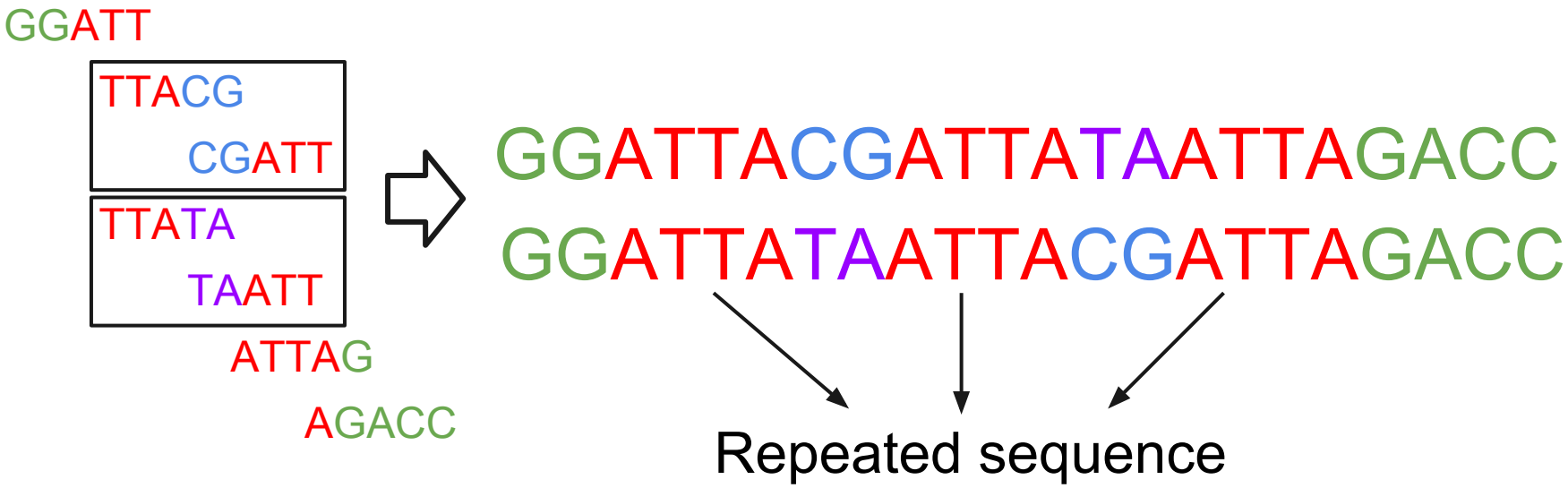
Assembling a genome

Let’s assemble a genome from reads!



Repetitive sequences + short reads = ambiguity



Assembly challenges

So why not do this for all dogs?

- $$$
- Assembly is hard
  - Repeats mean assemblies will be fragmented
  - Data takes up a lot of disk space (can’t fit on your laptop)
  - Algorithms use a lot of compute time/resources
  - We typically don’t consider heterozygous bases when we assemble

More info than SNP chips, less work than assembly

Since we have an assembly for Tasha already, there’s an easier way to find where the reads go!

We’ll talk more about this on Friday.

Logistics

Today's assignment is the last formal assignment!

Part 1 - Wednesday, Jan 16.
Part 2 - Friday, Jan 18.
Part 3 - Wednesday, Jan 23.



Clarence
Reilly
Finch

If you need extra time beyond these dates, that's fine, but if you do please shoot me an email letting me know where you and your partner are on the assignment.

Please turn in your code and question answers to rsherman@jhu.edu and include EN.601.147 in the subject line. Make sure both your names are on your writeups!