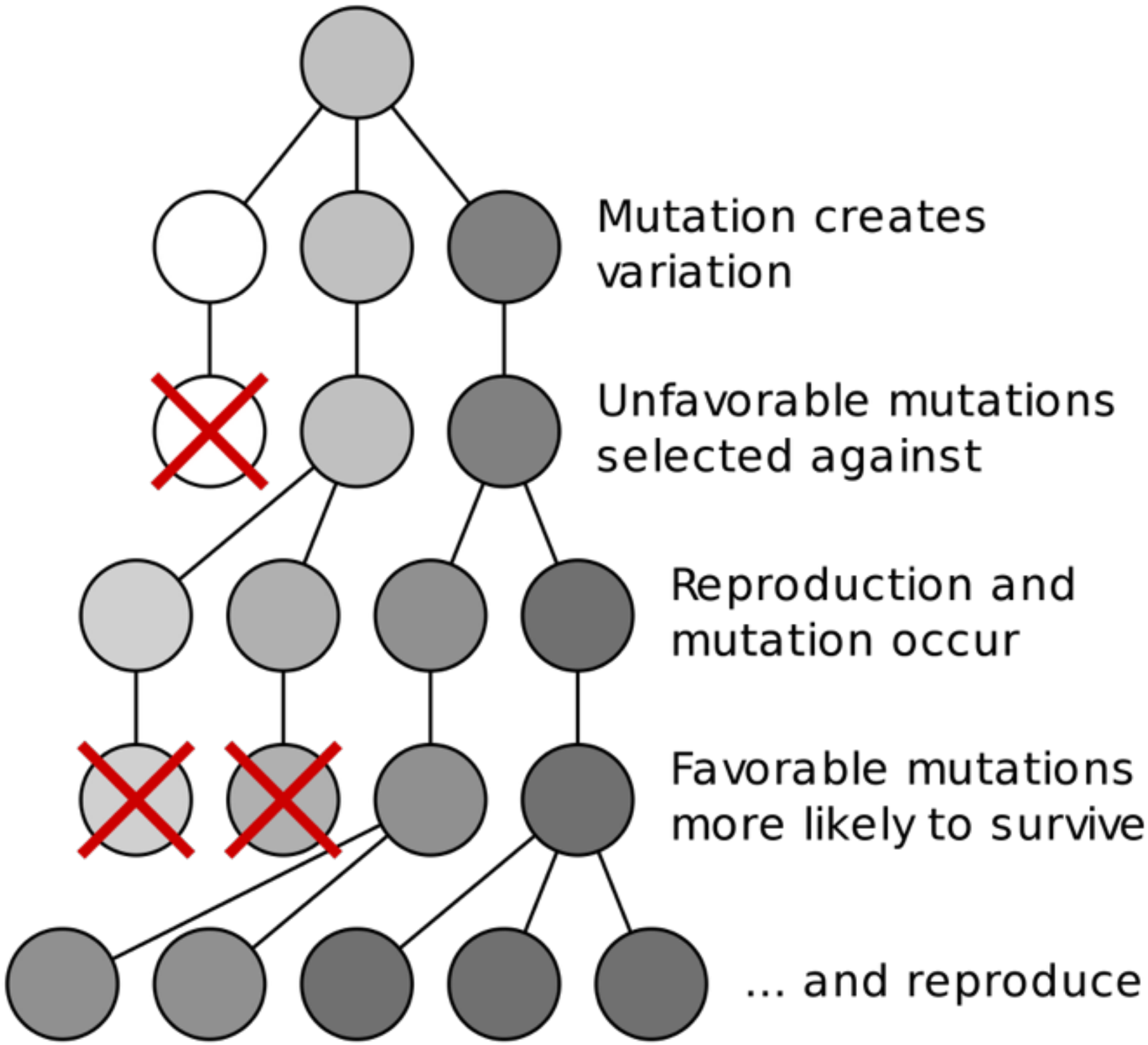
Week 1

Welcome

All organisms accumulate mutations in time

Example: mutation from A to C

ACTTCTGGATGACACTAC
ACTTCTGGATGACCCTAC


Mutation creates variation
Unfavorable mutations selected against
Reproduction and mutation occur
Favorable mutations more likely to survive
... and reproduce

Which fuels the process of

Evolution

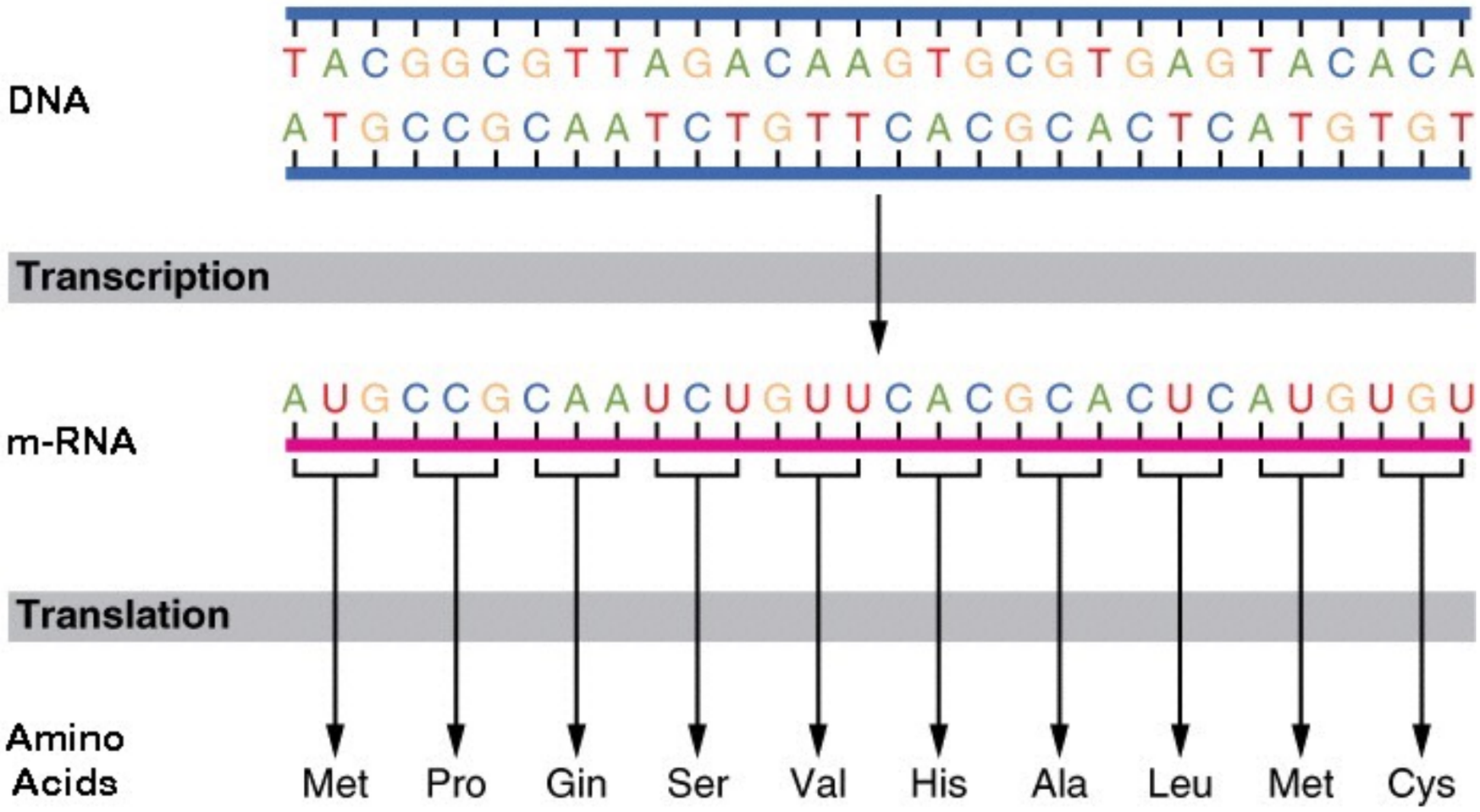
me

What can a string of DNA letters tell you?

A LOT if you are looking at coding regions

coding regions = genes

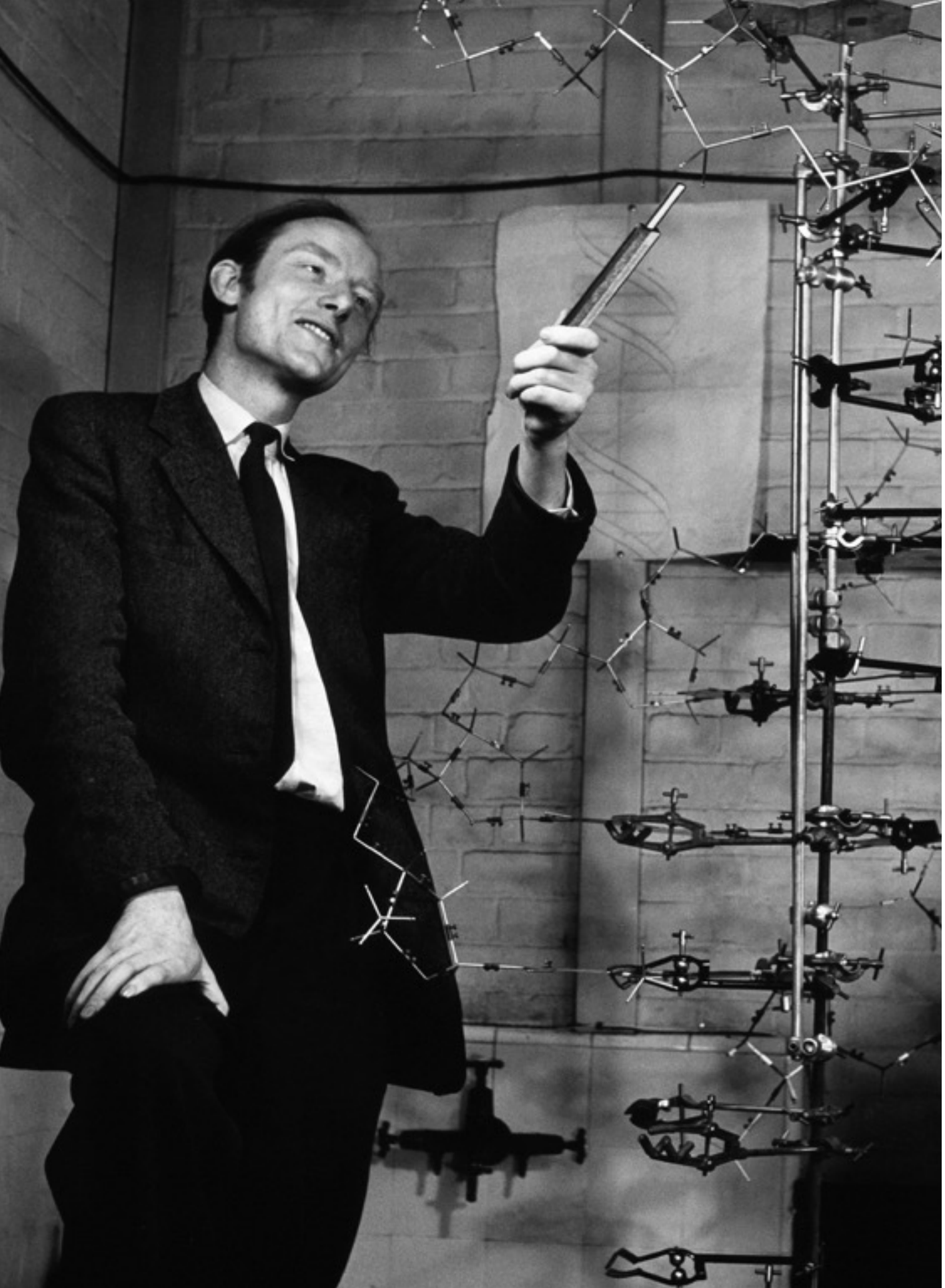
DNA → **RNA** → **Protein**



Humans

97%

Non-coding



“little better than junk.”

- Francis Crick

Short Answer: Not Junk

Non-coding regions which direct gene expression:

Non-coding RNA

Promoters

Insulators

Silencers

Enhancers

Short Answer: Not Junk

Non-coding regions which direct gene expression:

Non-coding RNA

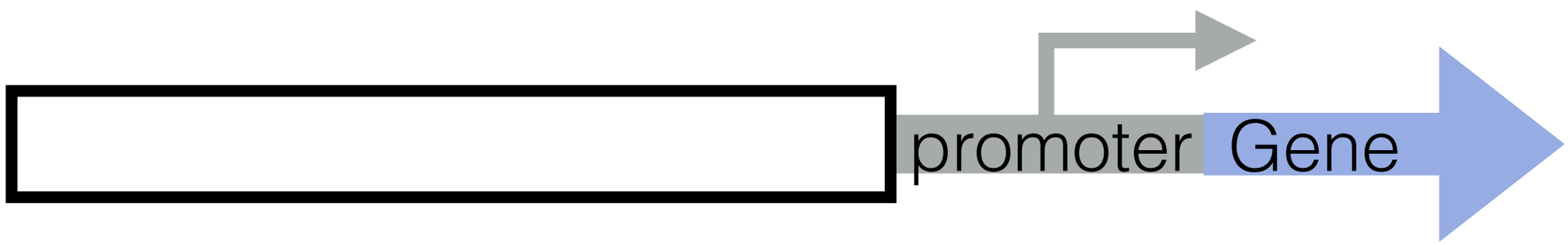
Promoters

Insulators

Silencers

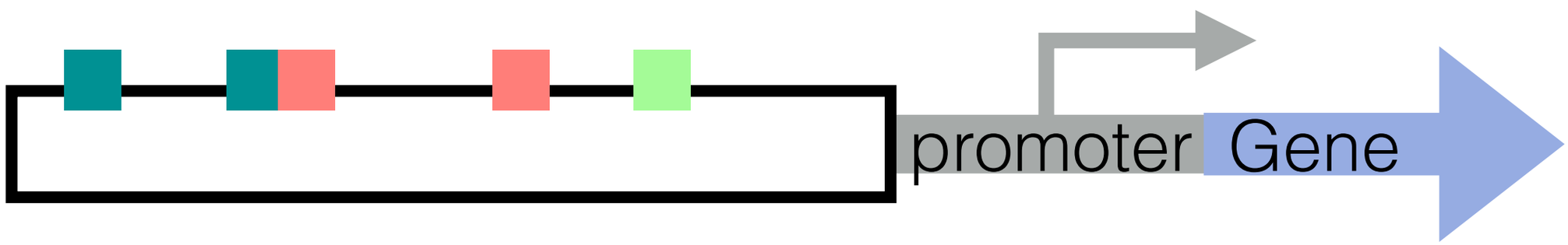
Enhancers

We *think* enhancers direct gene expression like this:



We *think* enhancers direct gene expression like this:

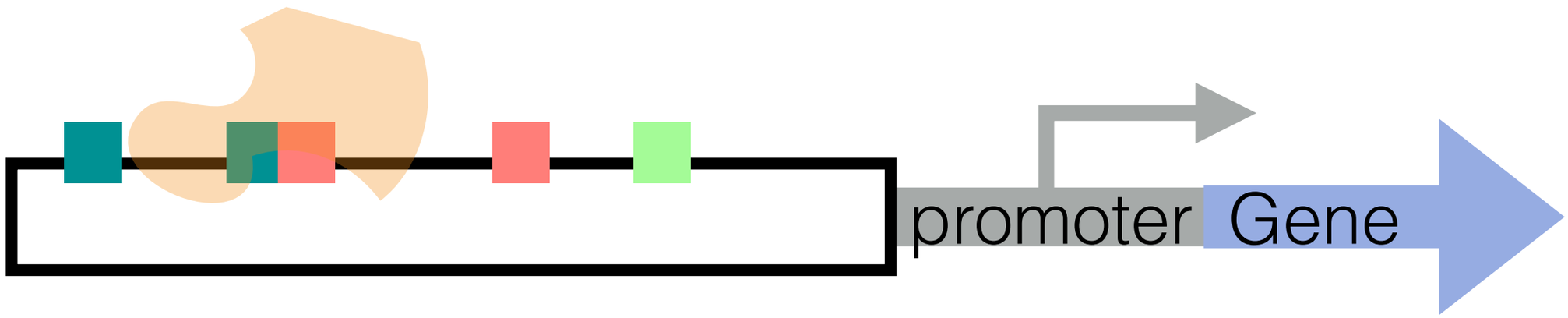
-They have Transcription Factor Binding Sites



We *think* enhancers direct gene expression like this:

-They have Transcription Factor Binding Sites

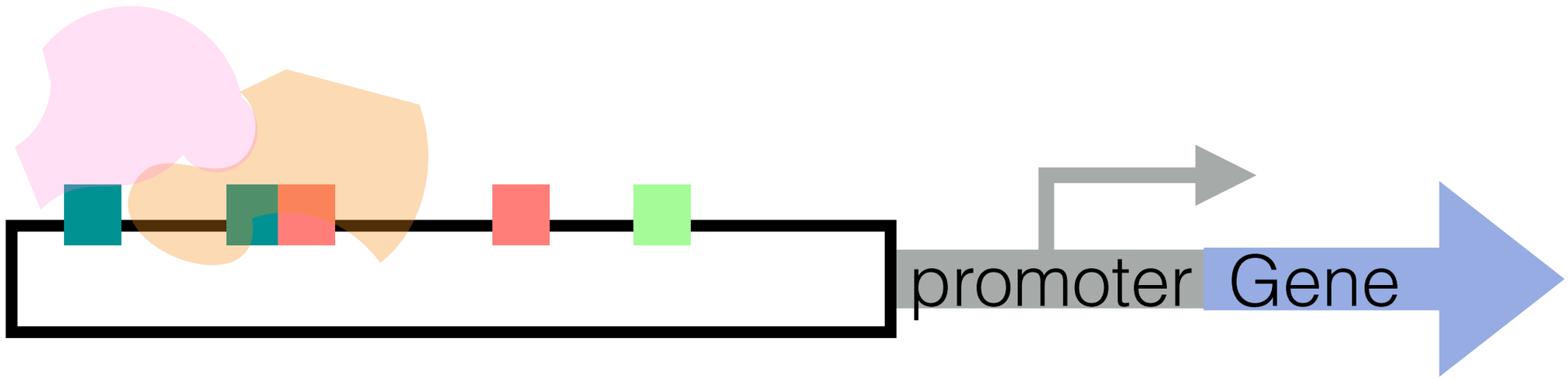
- That bind Transcription Factors



We *think* enhancers direct gene expression like this:

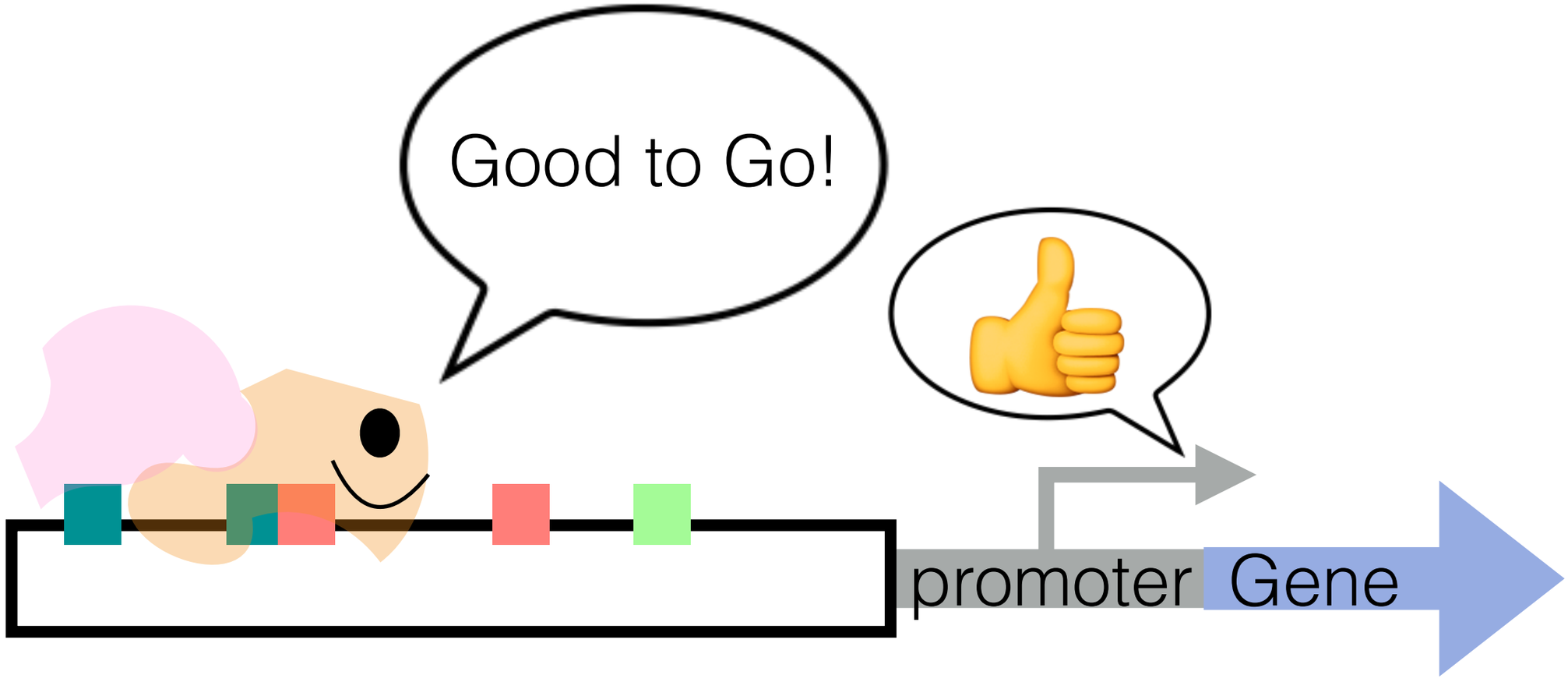
-They have Transcription Factor Binding Sites

- That bind Transcription Factors



We *think* enhancers direct gene expression like this:

-They have Transcription Factor Binding Sites
- That bind Transcription Factors
- Which somehow tells the promoter to turn on or off a gene



?????

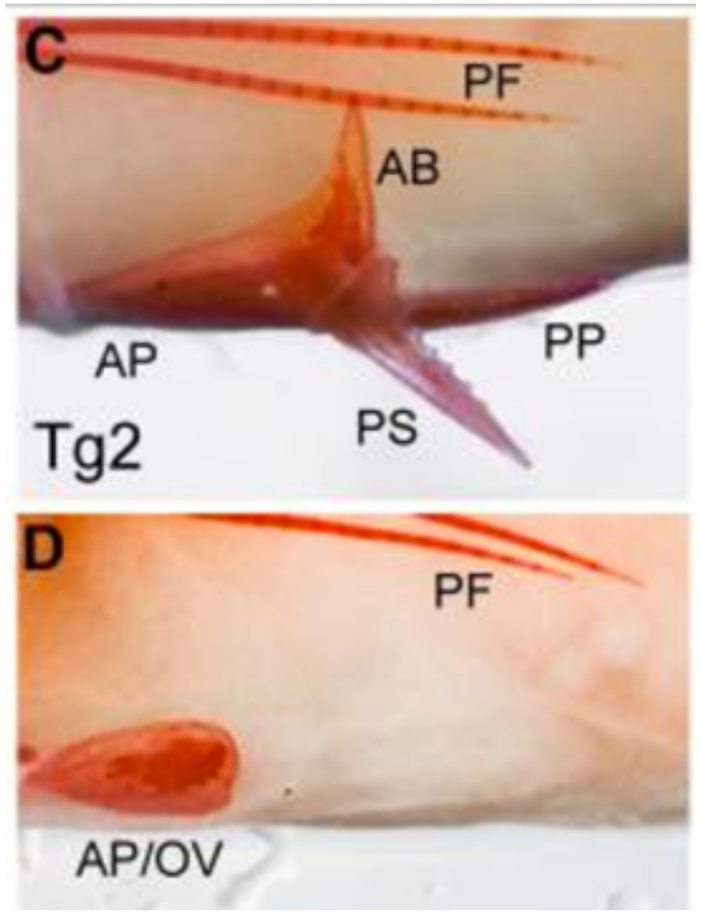

Good to Go!

We do know:

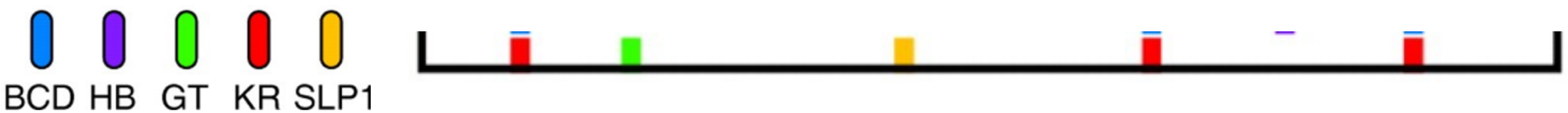
Enhancers can have a profound effect on gene expression

- Mutations in enhancers have been found to cause human disease

- Loss of enhancer function can fuel adaptive evolution by directing entire loss of appendages



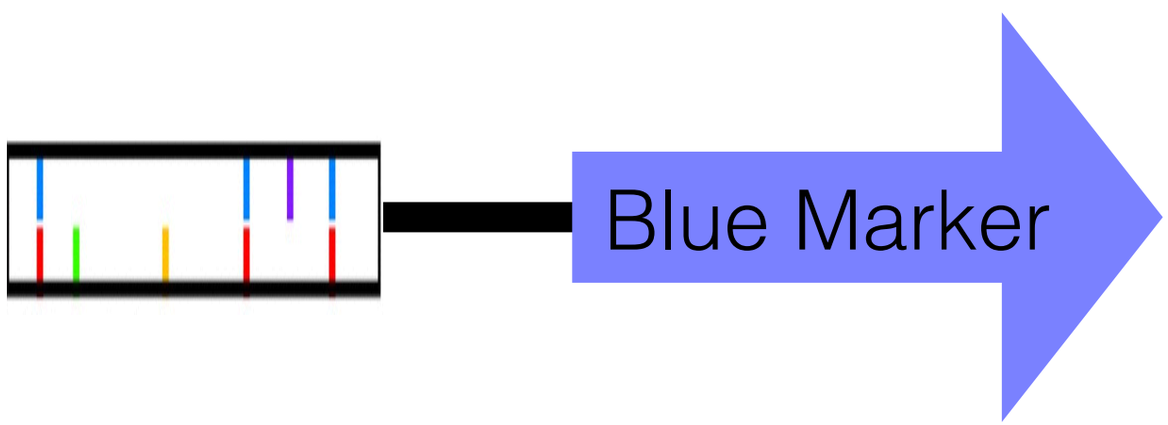
Schematic of DNA sequence with Transcription Factor Binding Sites

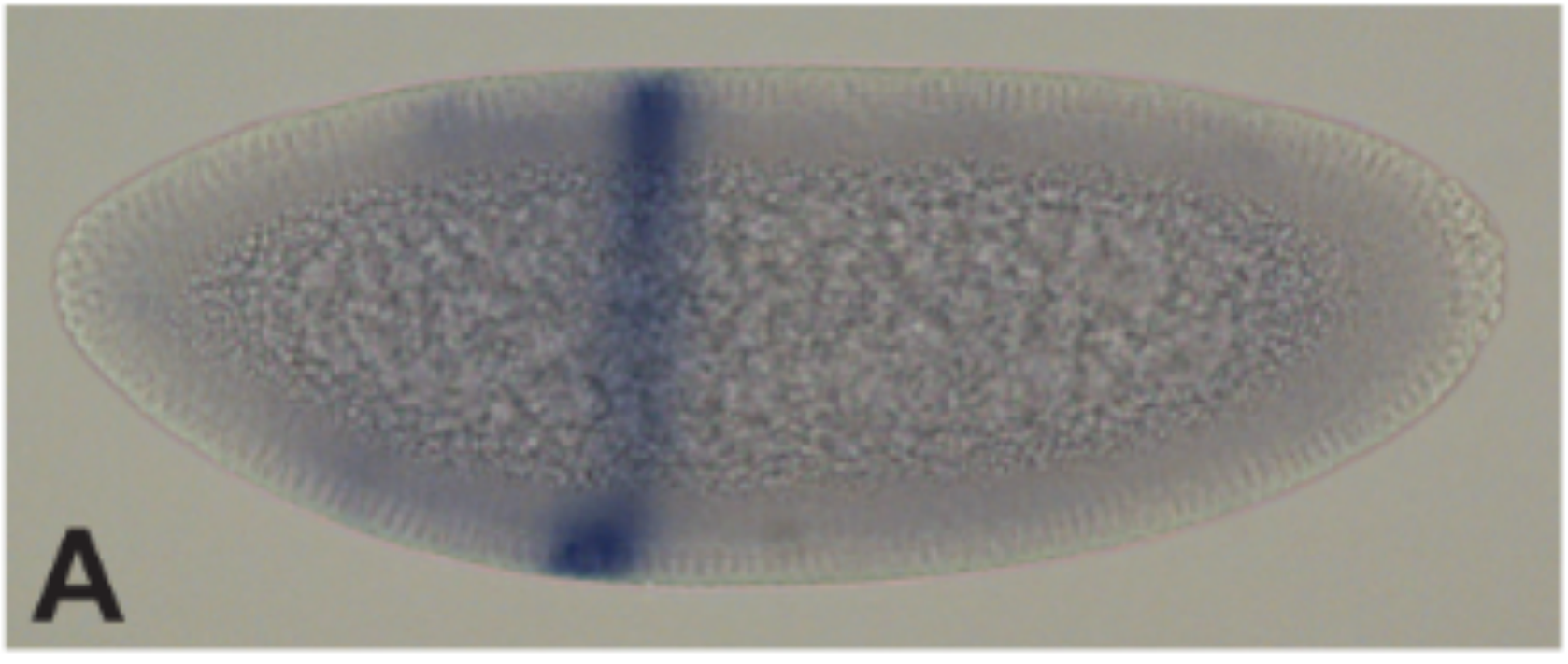


DNA can be interpreted in two directions (more later)

eve stripe 2 enhancer binding sites

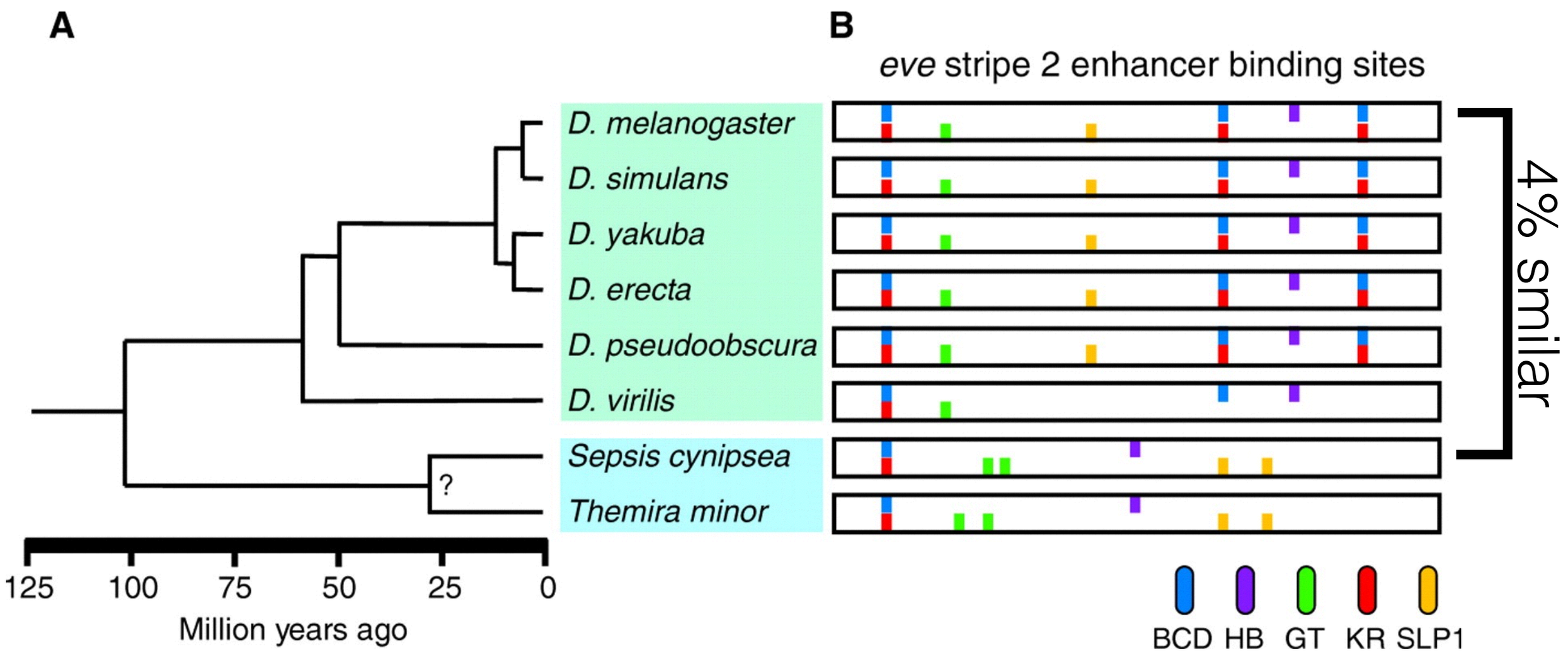
D. melanogaster





Borock et al., 2010, Hare et al., 2008

Mutations at high rates in enhancer regions, resulting in high frequencies of nucleotide changes between species

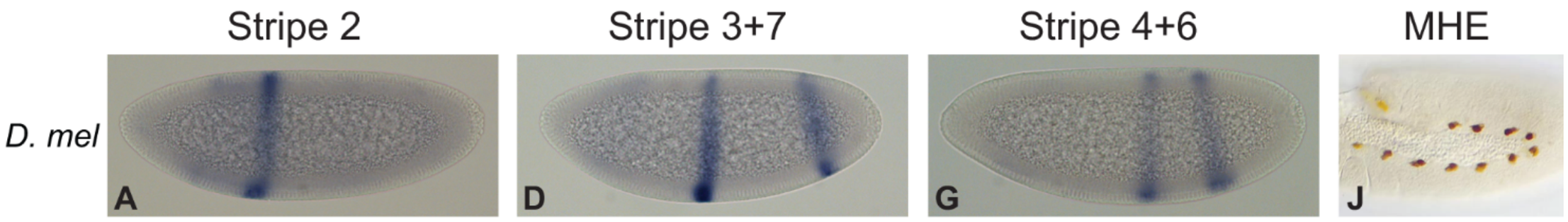


Borock et al., 2010, Hare et al., 2008

So what happens to the enhancer function when only 4% of the sequence remains the same?

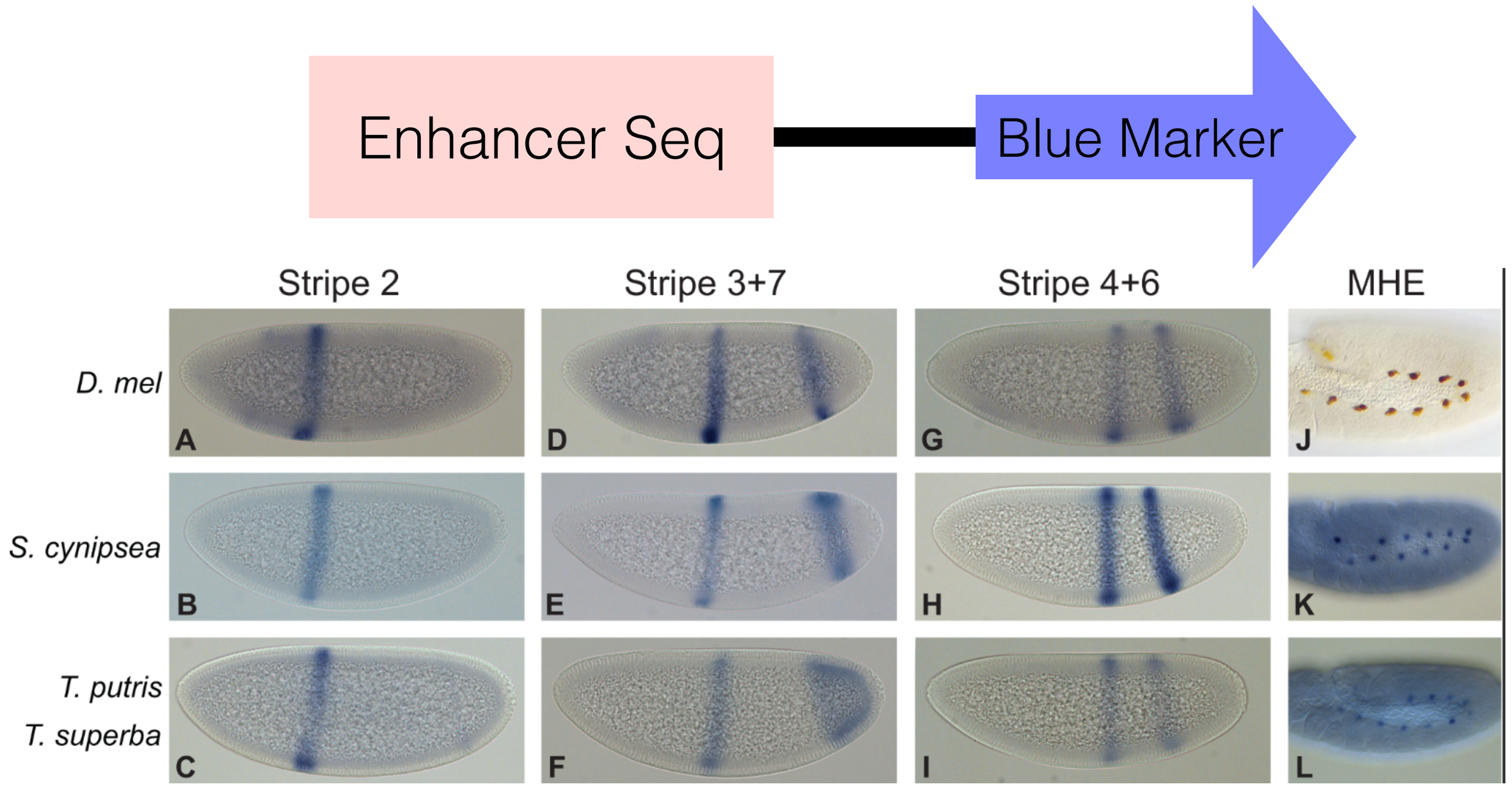
D. mel
Enhancer Seq.

Blue Marker



Hare et al., 2008

Substantial alterations in enhancer sequence structure has little effect on regulatory function!!



Hare et al., 2008

1. Observation / Question

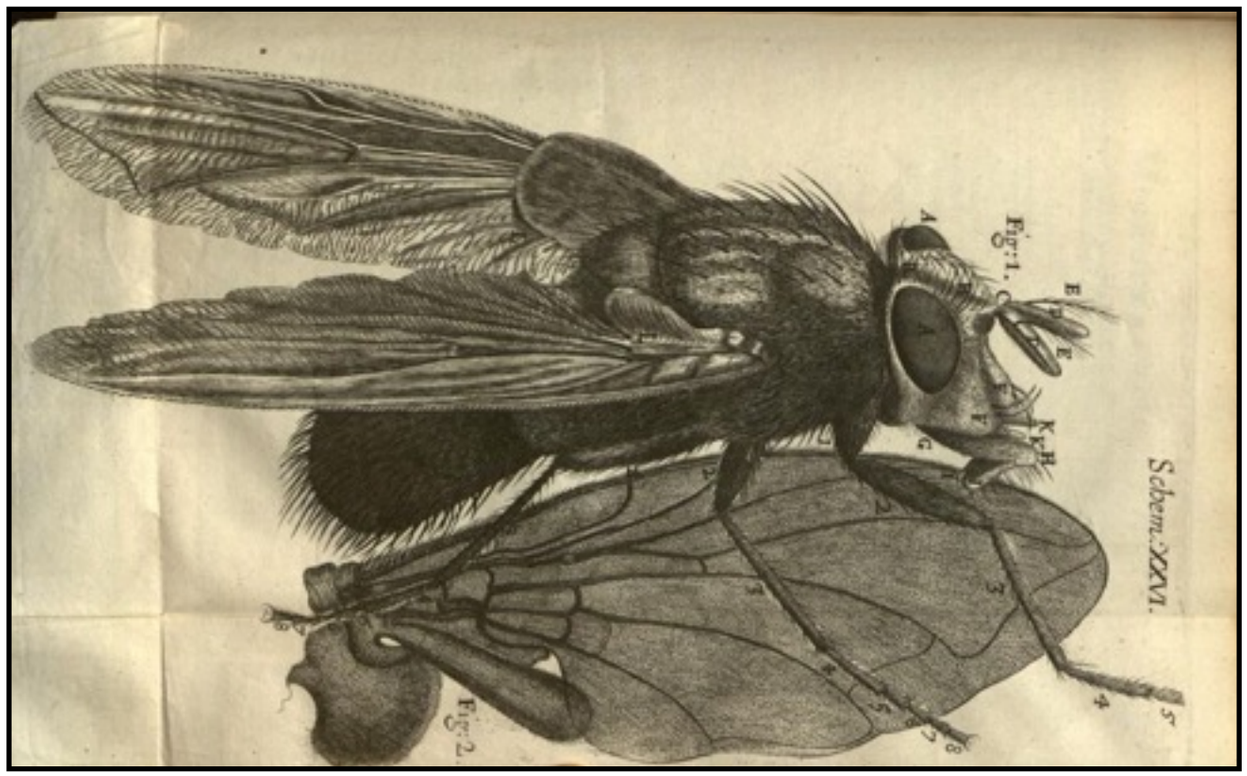
1. How is function maintained despite divergence in enhancer sequence architecture?

2. Is there an hidden enhancer syntax?

3. How do we define enhancers in vast non-coding sequence space?

2. Evolutionary Scale / Species to study

And now study **Fruit Fly (Drosophila)**



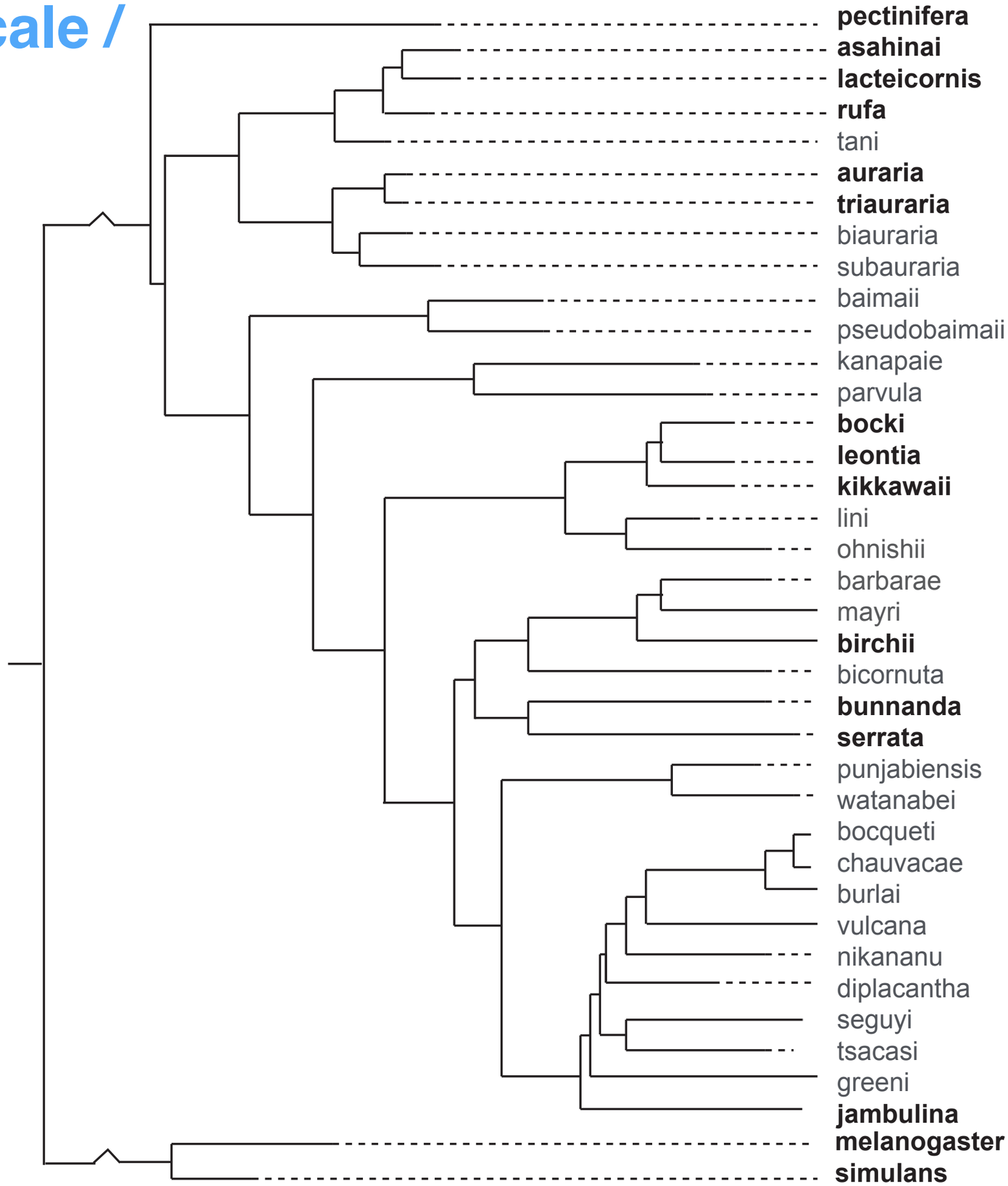
- Small, well established genome (smaller and cleaner dataset)
- Development well studied
- Fast and ample molecular tools

2. Evolutionary Scale

Level	Between Distant Species	Within species (Population Level)
Seq Similarity	too low	too high
Evolutionary Scale	too Big	too Small

2. Evolutionary Scale / Species to study

Data set 1:
Montium
Genomes



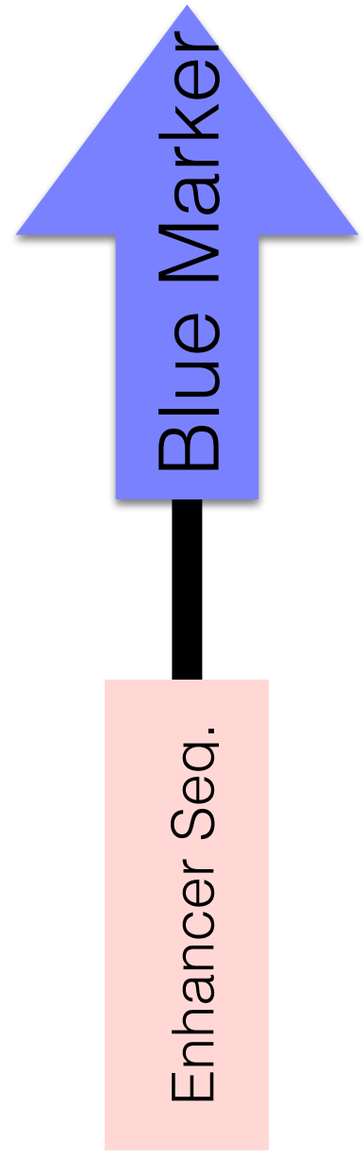
w/ Michael Bronki

Data Set 2: Kvon et al., 2014.

Classified non-coding regions that have verifiable expression patterning during embryogenesis.

- 13% of the non-coding genome
- 3557 / 7705 are active

ID	***Coordinates***	***Expression (4–6, 7–8, 9–10, 11–12, 13–14, 15–16)***
VT0267	chr2L 580471–582588 (2117 bp)	 active
VT0268	chr2L 581470–583741 (2271 bp)	 active
VT0269	chr2L 584274–586488 (2214 bp)	 weak



Approach

Data sets

- 24 Montium species genomes
- classified as active or non-active in 6 developmental stages
- 7705 regions non-coding seq regions

Machine Learning Classification

1. On and off function
2. Stage specific function

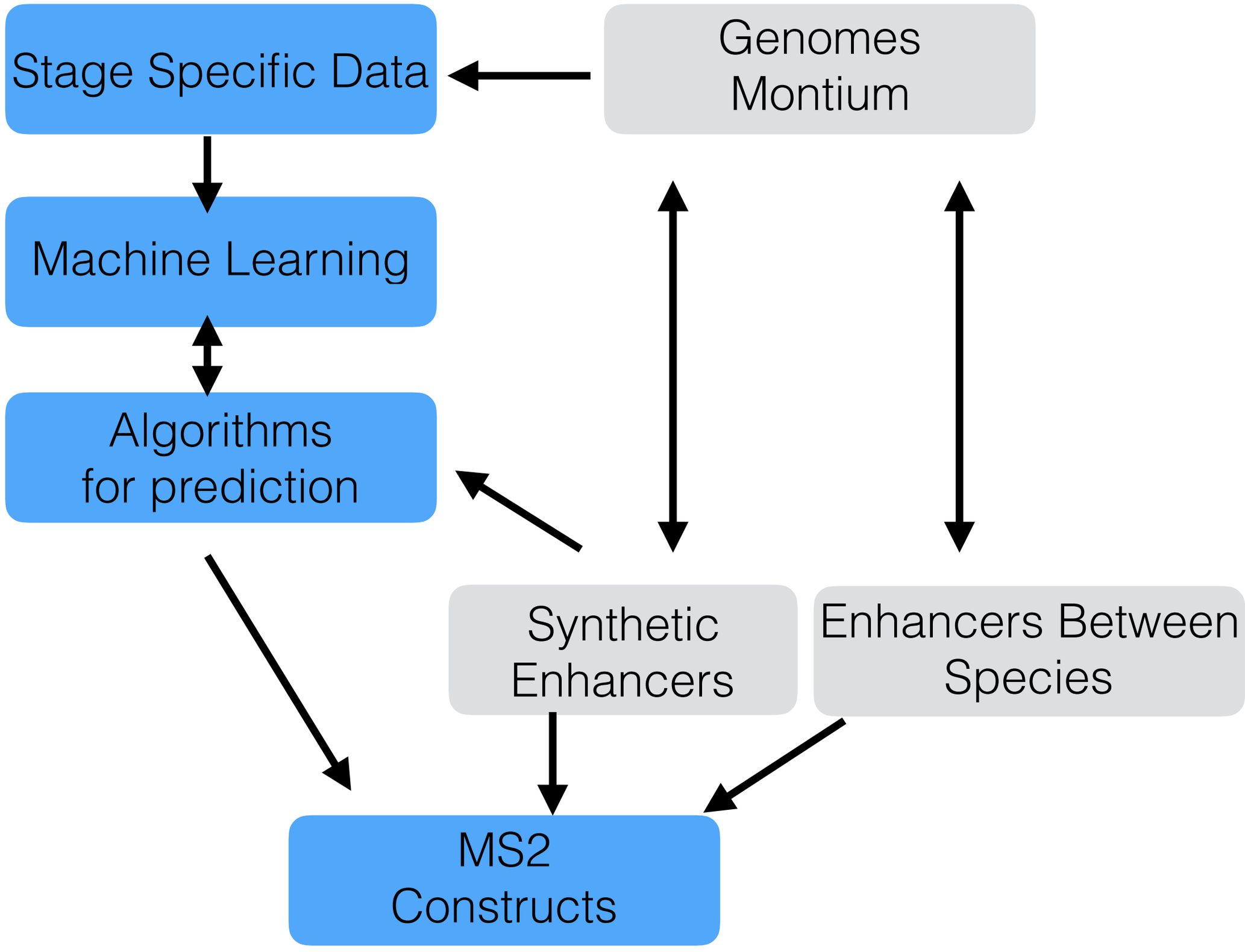
Research Plan

Input

Training

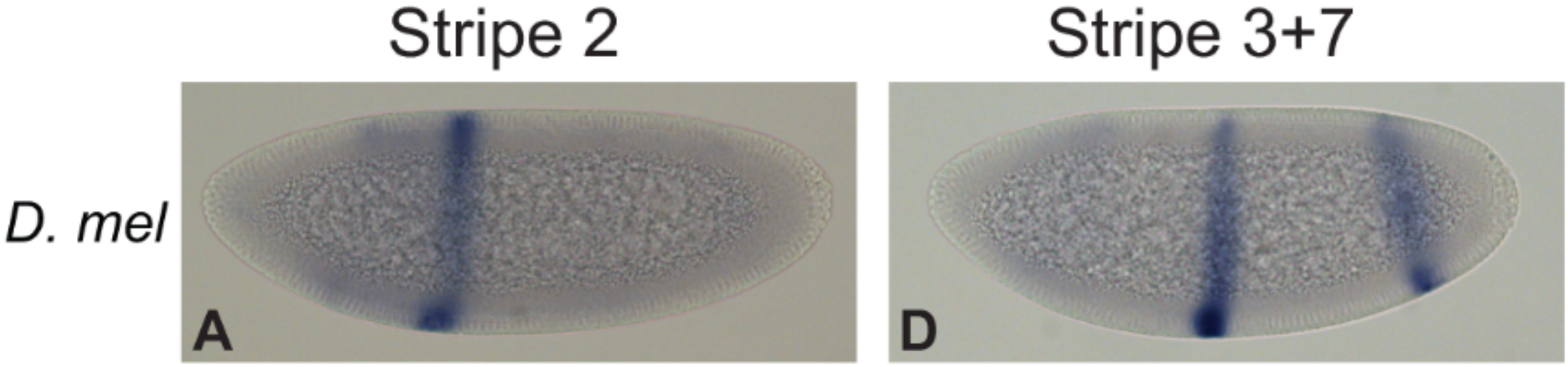
Simulation

Testing MS2

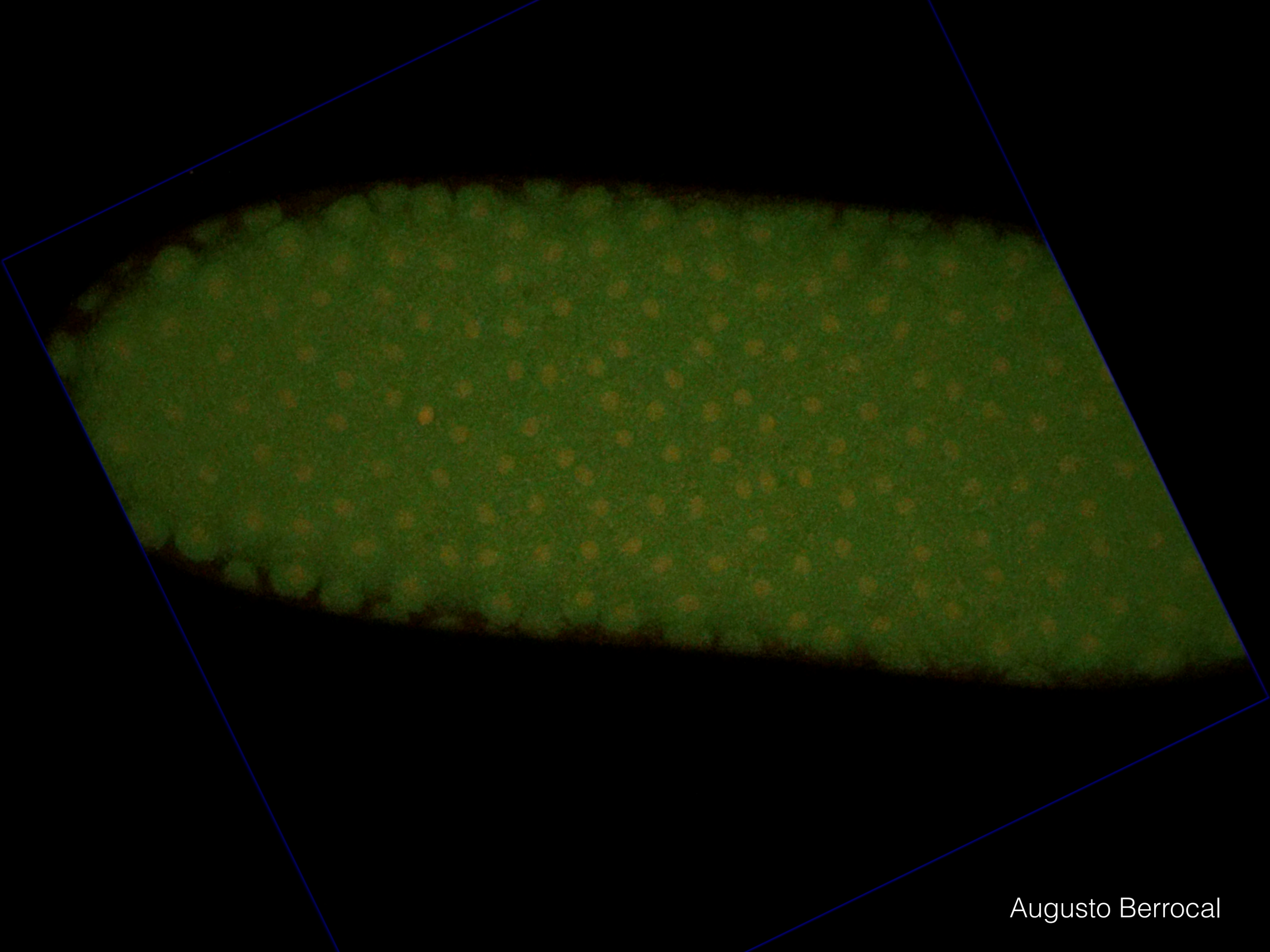


How to test Enhancer Function?

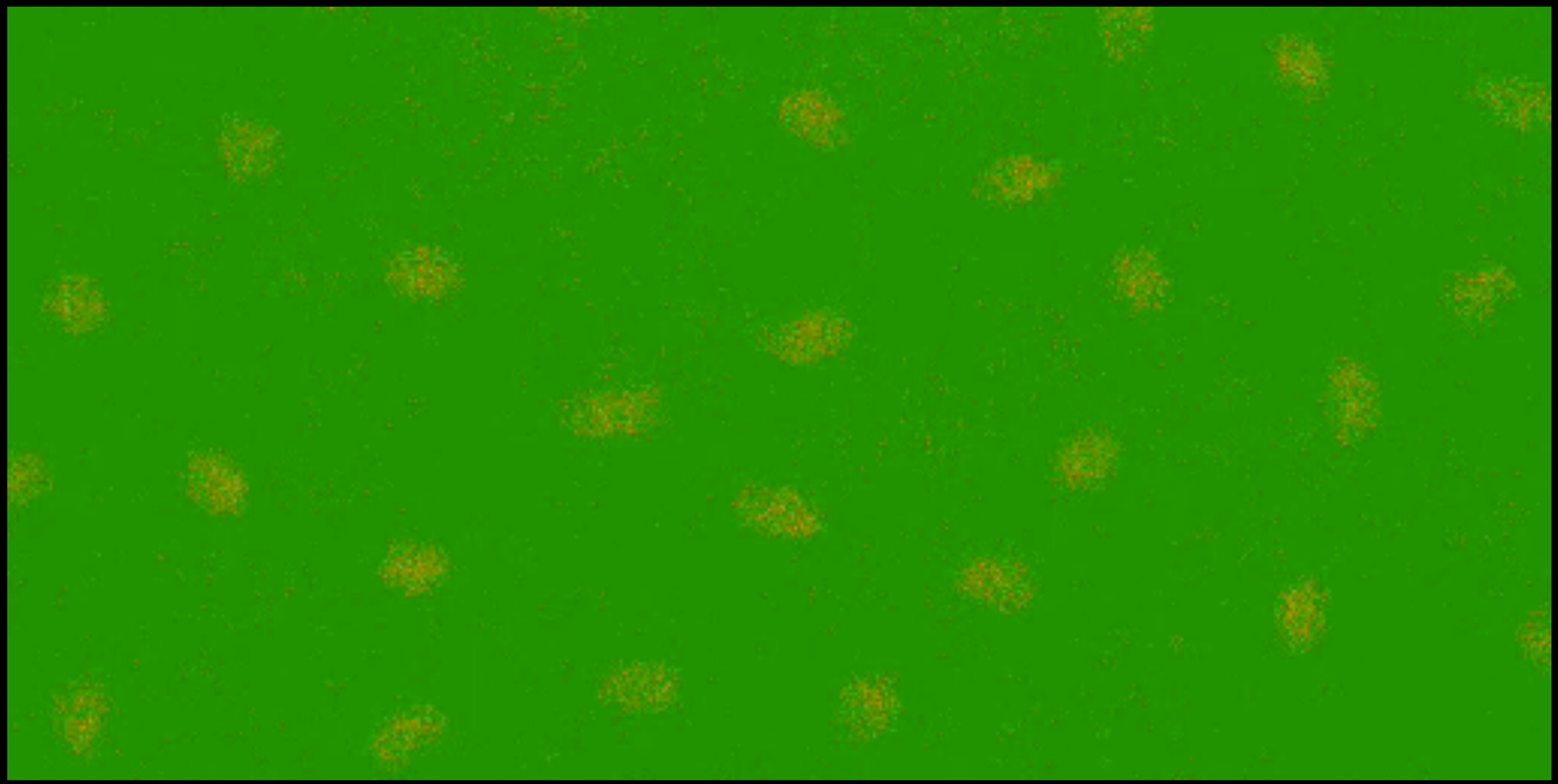
Previously:



In-situs: Static View in Dead Embryos



Augusto Berrocal



Yuya Karita

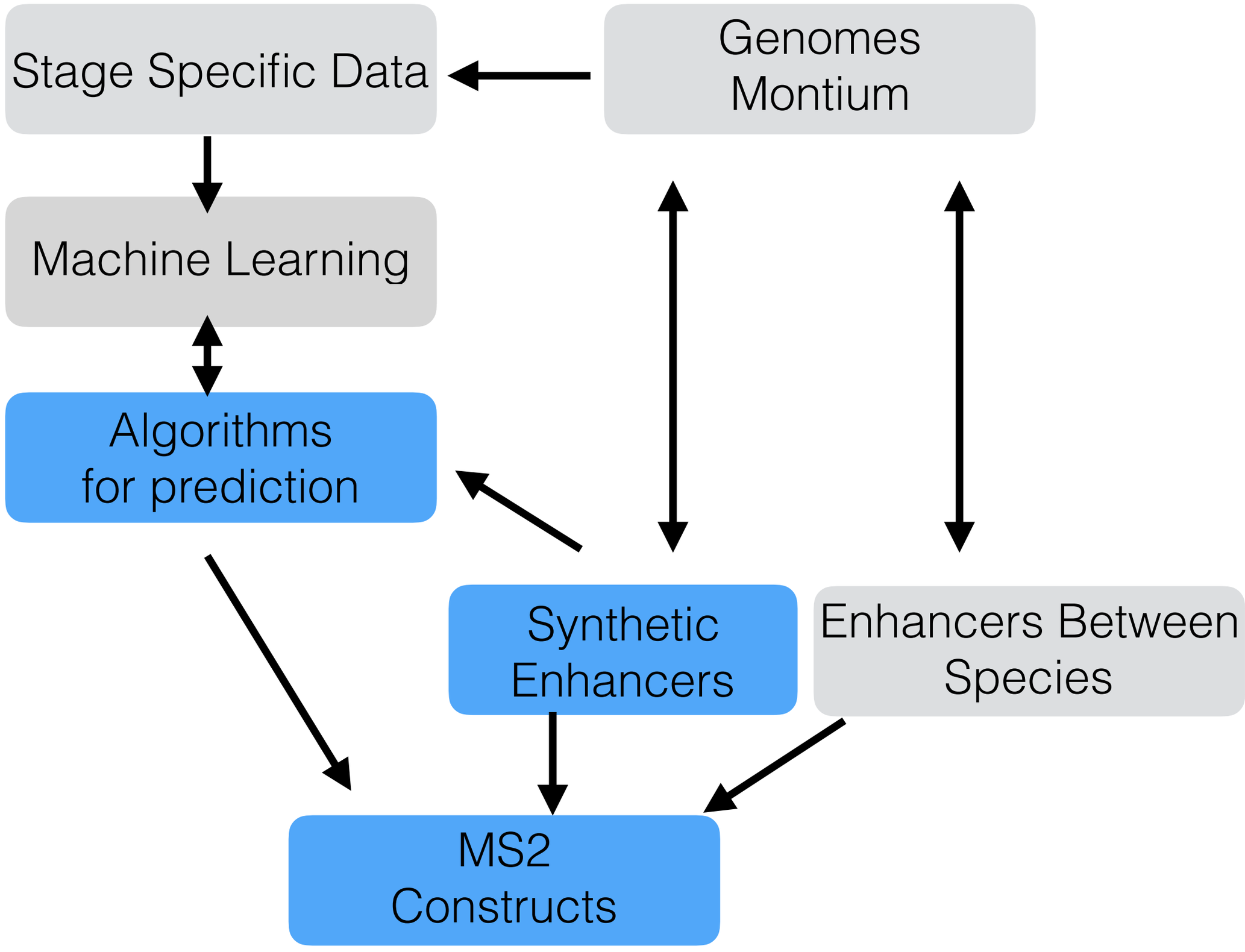
Research Plan

Input

Training

Simulation

Testing MS2



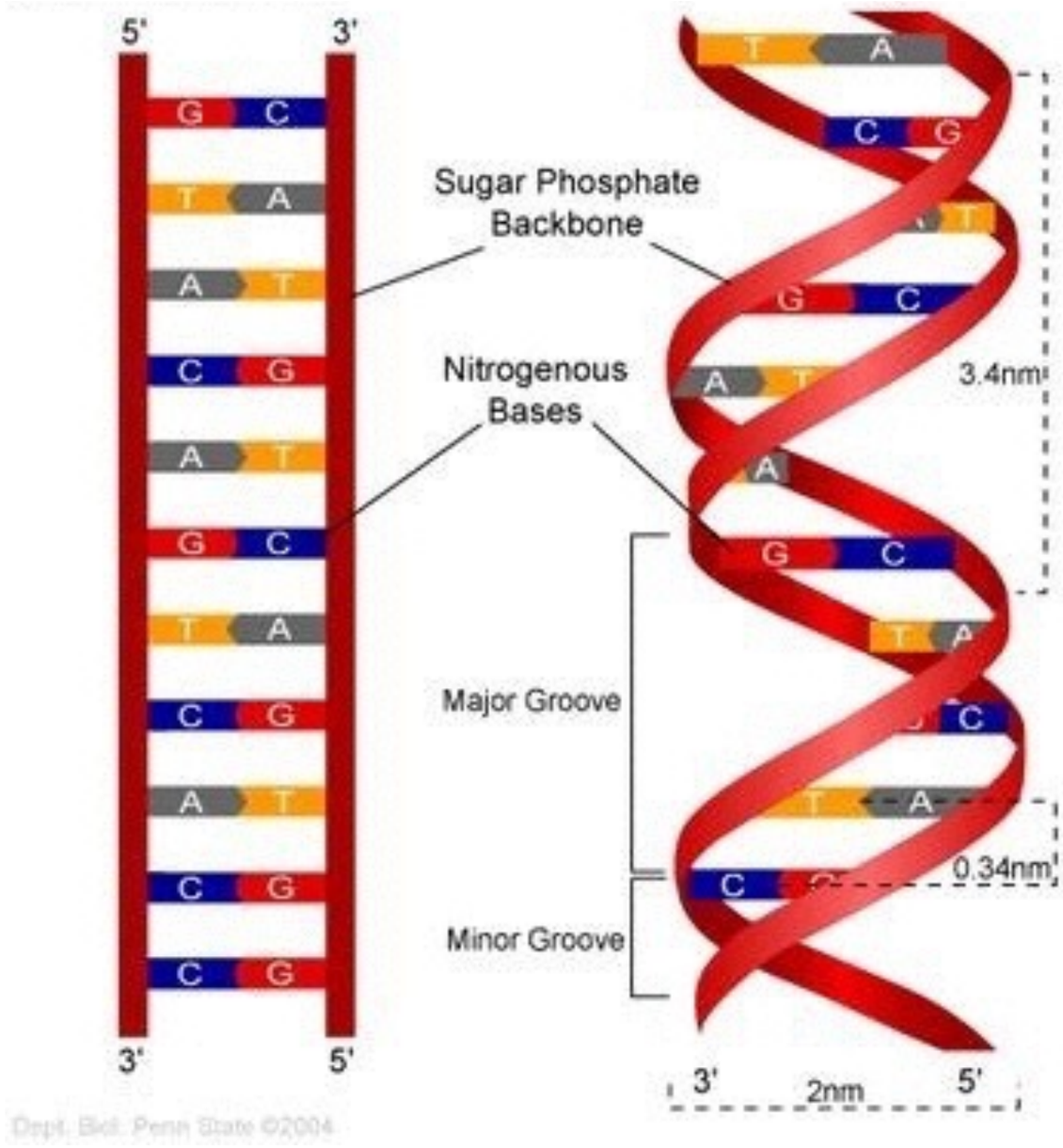
Holy Grail

We create an algorithm that can create synthetic enhancers that verify direct gene expression.

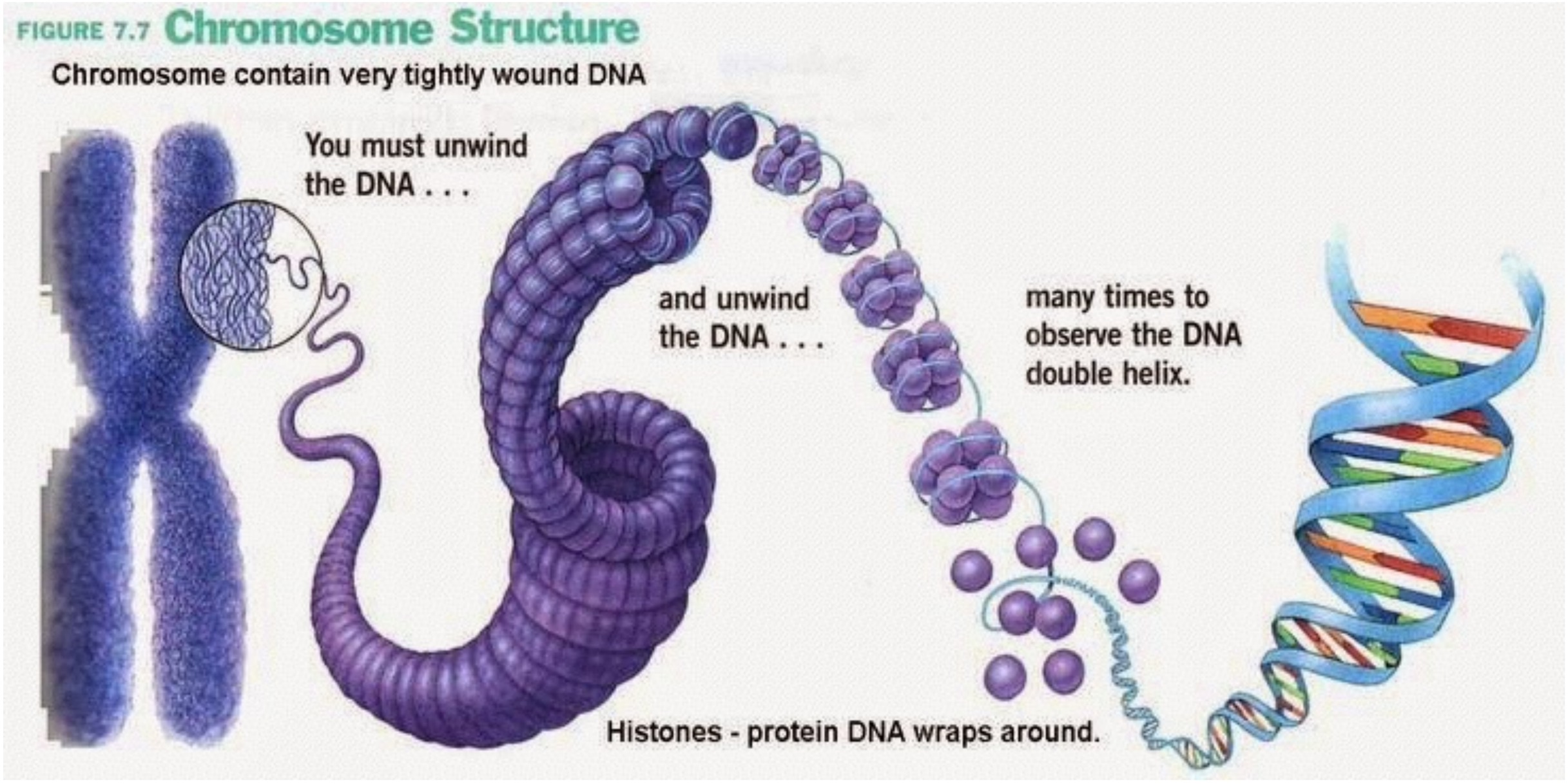
DNA

ACTTCTGGATGACACTACTACAGTGCA

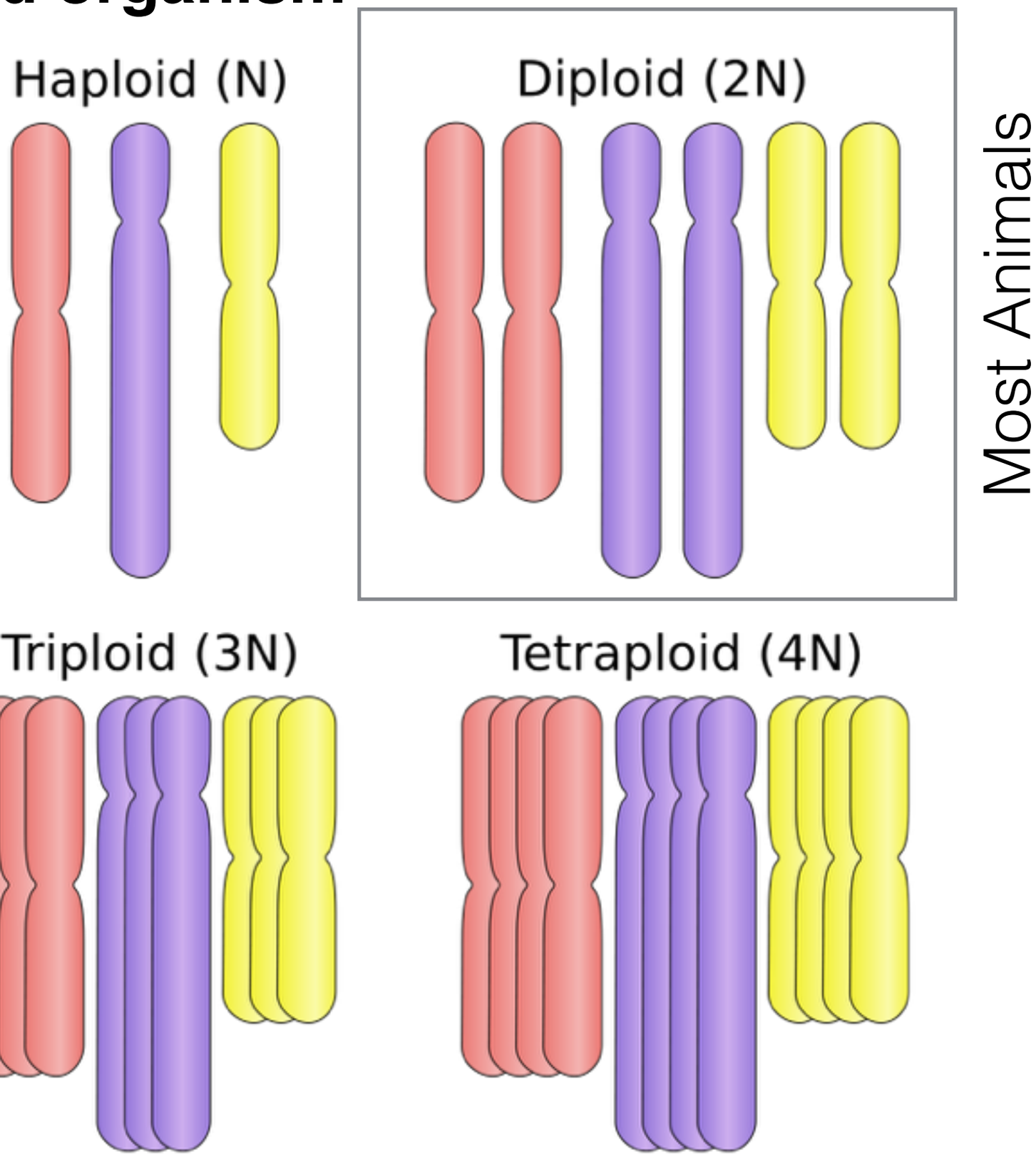
TGAAGACCTACTGTGATGGATGTCACT



DNA is packaged into chromosomes



DNA in a diploid organism



But…lucky you:

You will be getting a string of letters

Teams

Neural Network Implementation

Quality Control Pipeline

TFBS mapping

Team Neural Network Implementation

Data Input: So many options!

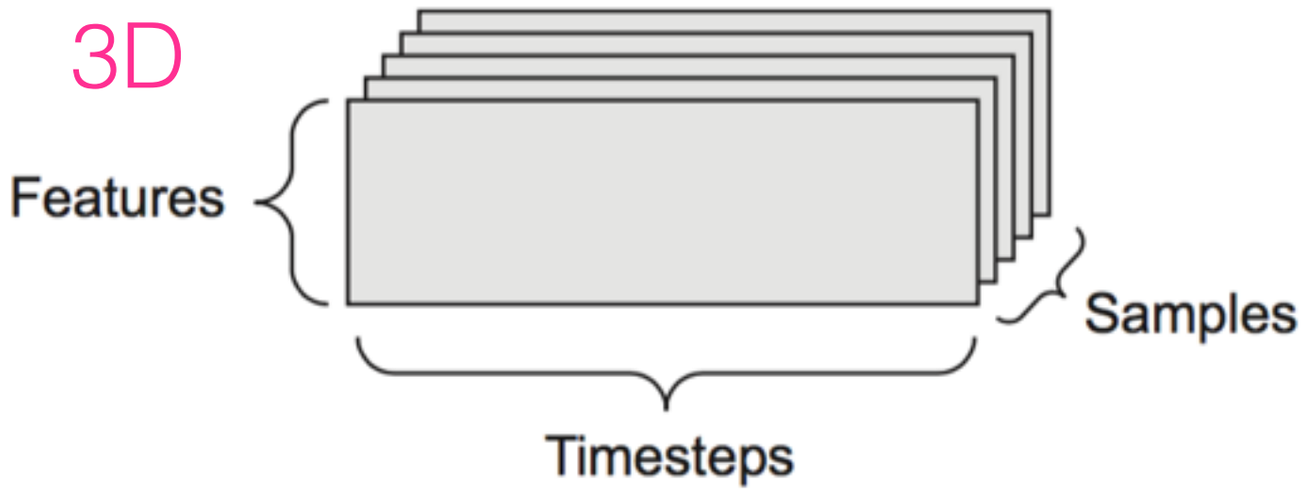
1D

```
> str(x)
 num [1:5] 12 3 6 14 10
```

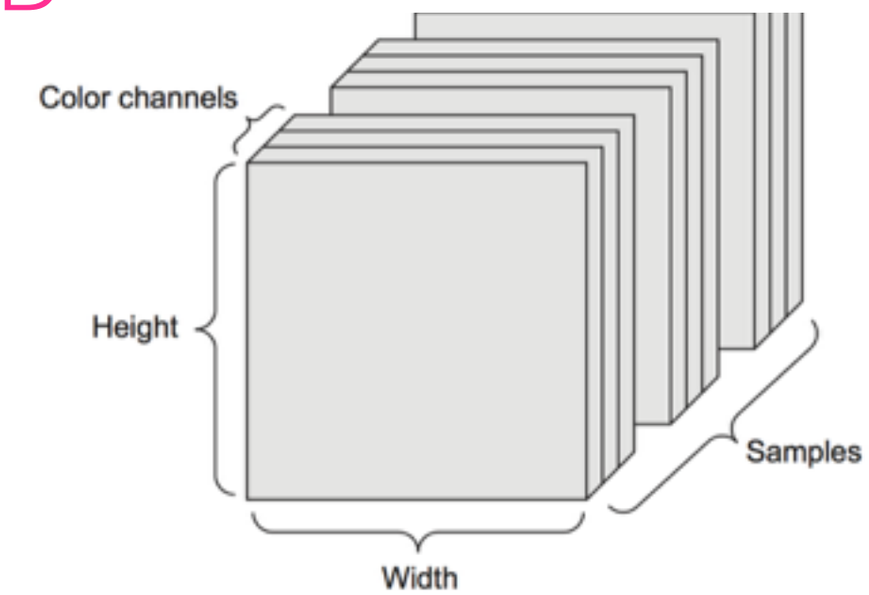
2D

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    0    0    0    0
[2,]    0    0    0    0    0
[3,]    0    0    0    0    0
```

3D



4D



Neural Network Implementation

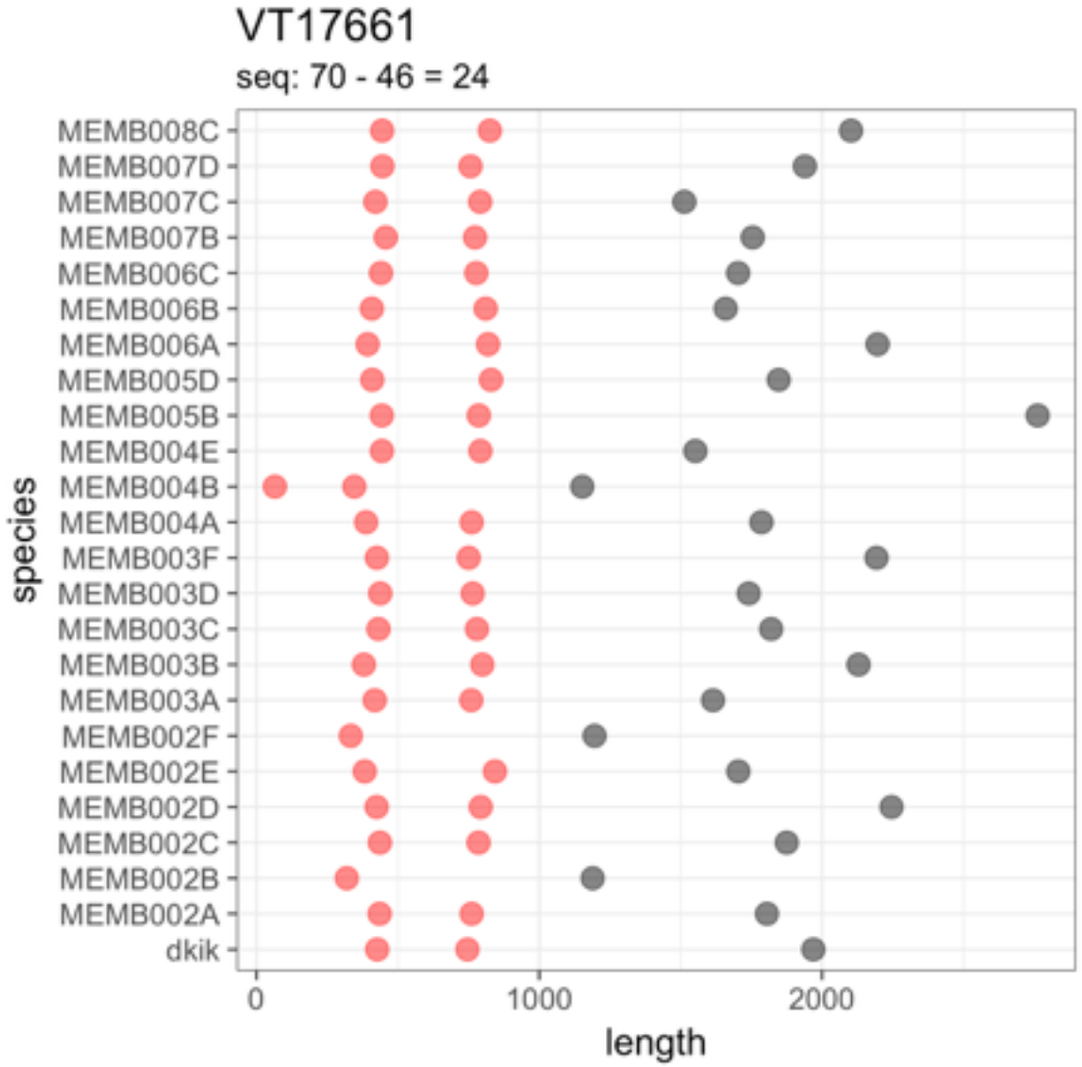
To start:

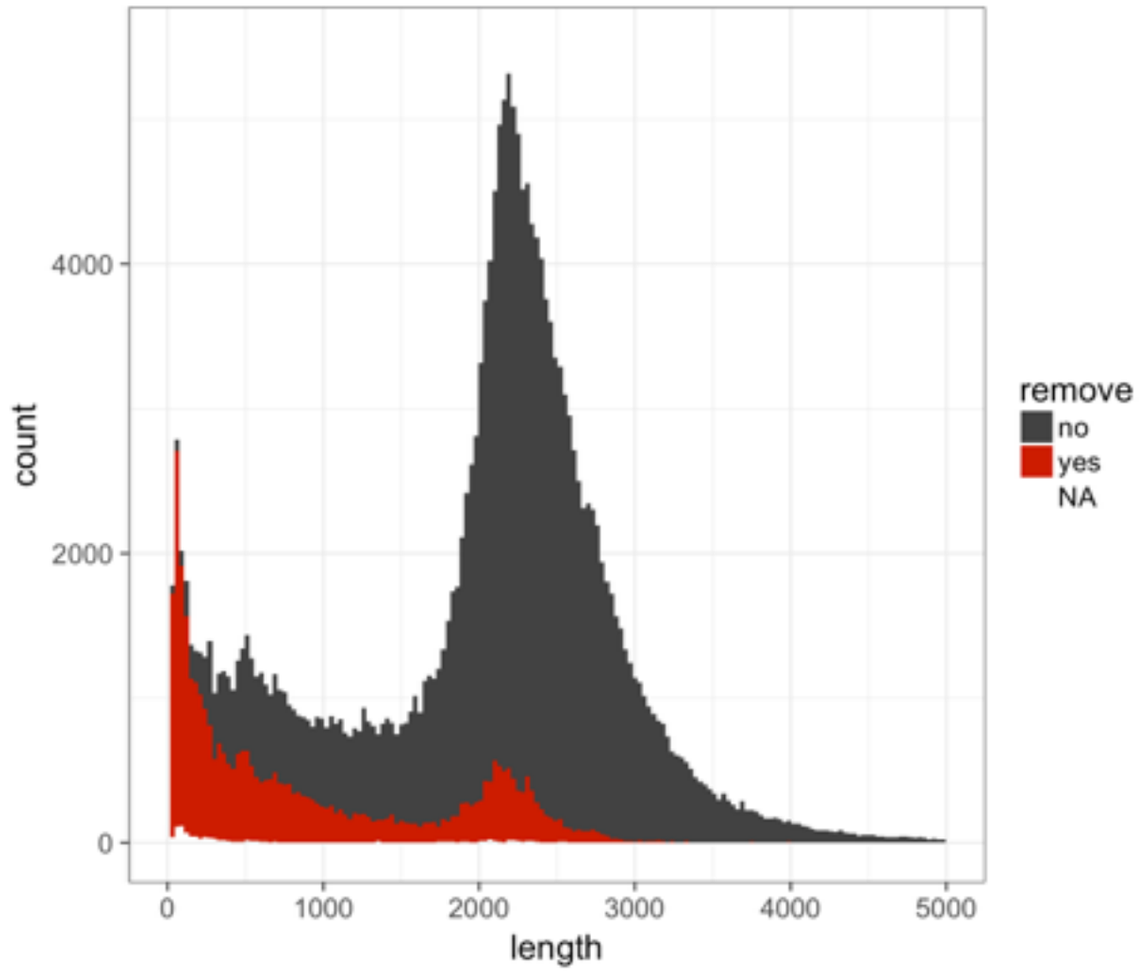
- Raw DNA to Tensors, What are our options?
  - One Hot encoding (2D)
- What layers make sense to include?
- Incorporate more data (mapped TFBS, DNAseI, species relatedness, ect)

Team Quality Control Pipeline

To start:

- Quantify and analyze the input data
- Are these sequences valid?
- Which sequences do we remove?
- Visualize





Team TFBS mapping

- Map all TFBS onto sequence
- Visualize TFBS and conservation
- Make tools for community

A

ID:VT58706
GAL4 expression: yes
DNAse Peaks: 0

score 4 6 8 10

motif_file bcd_FlyReg cad_FlyReg gt_nar2008 kr_FlyReg zelda_

species

D. serrata
D. rufa
D. pectinifera
D. lacteicornis
D. jambulina
D. bunnanda
D. bocki
D. birch
D. asahinai

negative strand

positive strand

B

post.prob

1.00
0.75
0.50
0.25
0.00

0
1000
2000
3000
4000

align_position

To do this week:

- Email your team preference
- give me your github username
- Make and upload a program that reads in fasta sequence and performs one or all of these tasks:
  - aligns sequence
  - turns sequences into a basic python data structure
  - measures GC content per sequence

Questions?