Week 3

Agenda

1. Team Updates
2. Ciera Updates
3. Intro to conservation, phylogenetics and alignments

Ciera Updates

- Use Issue trackers
- You all should write permissions
  - you can make new repos!
- data_management.md
  - Pull requests
- readme.md

Conservation

Similar or identical nucleotide sequences between species (**orthologous**) *or within a genome (paralogous)*.

All organisms accumulate **mutations**, those nucleotides that persist in time are conserved sequences. The conserved sequences are maintained through by **natural selection**.

Conservation

You can observe conservation between species from sequence **alignments**.

```
Species 1   ACTTCTGAAGATGACACT---------AC
Species 2   ACTTCTGAAGATGACACT---------AC
Species 3   ACTTCTGAAGATGACACTTCCGCTAACC
Species 4   ACGGCTG--GATGACCCT---------AC
Species 5   ACGGCTG--GATGACCCT---------CC
```

Conservation

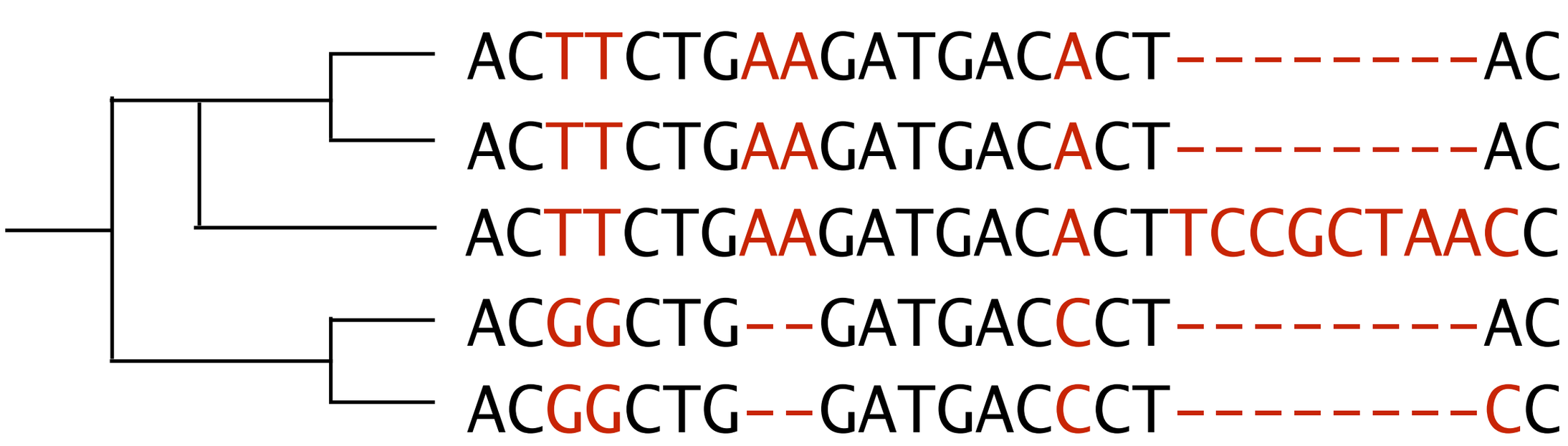
Patterns of conservation allow you to establish **phylogenetic relatedness** between species.

```
Species 1   ACTTCTGAAGATGACACT---------AC
Species 2   ACTTCTGAAGATGACACT---------AC
Species 3   ACTTCTGAAGATGACACTTCCGCTAACC
Species 4   ACGGCTG--GATGACCCT---------AC
Species 5   ACGGCTG--GATGACCCT---------CC
```

Conservation

Patterns of conservation allow you to establish **phylogenetic relatedness** between species.

Which is represented by **phylogenetic trees**.



Conservation

Conservation can occur in coding and non-coding nucleic acid sequences.

Highly conserved DNA sequences are thought to have functional value, although the role for conservation in non-coding DNA sequences is poorly understood.

Non-coding regions are a different beast.

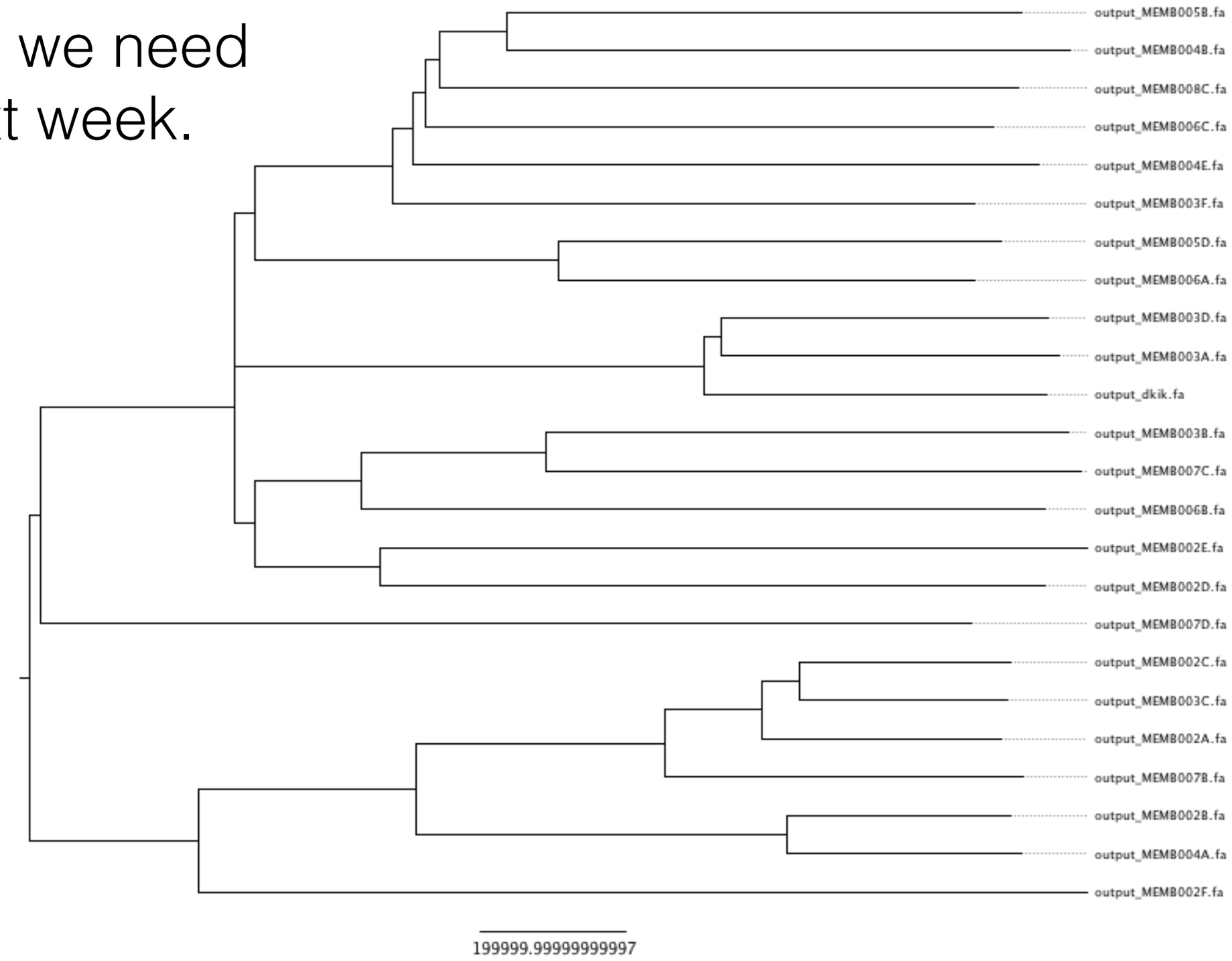
Conservation

Since all regions of a genome do not accumulate mutations in the same way, to find accurate phylogenetic relatedness between species, you need to apply special mutation rate models and do very careful analysis.

We left that another lab and we have a nice tree!

Conservation

- More on why we need this tree next week.



Conservation

Anyway, linear sequence conservation of enhancer architecture may not be as important as it is in coding regions.

```
Species 1   ACTTCTGAAGATGACACT---------AC
Species 2   ACTTCTGAAGATGACACT---------AC
Species 3   ACTTCTGAAGATGACACTTCCGCTAACC
Species 4   ACGGCTG--GATGACCCT---------AC
Species 5   ACGGCTG--GATGACCCT---------CC
```

Conservation

Linear sequence conservation of enhancer architecture may not be as important…

One reason:

Transcription Factor Binding Site (TFBS) Turnover

TFBS Motif = **ACTTCTG**

Conservation

We cannot just look at linear sequence conservation of enhancer architecture…

One reason:

Transcription Factor Binding Site (TFBS) Turnover

TFBS Motif = **ACTTCTG**

Main Question: Are there patterns of constraint on mutation in enhancers regions?

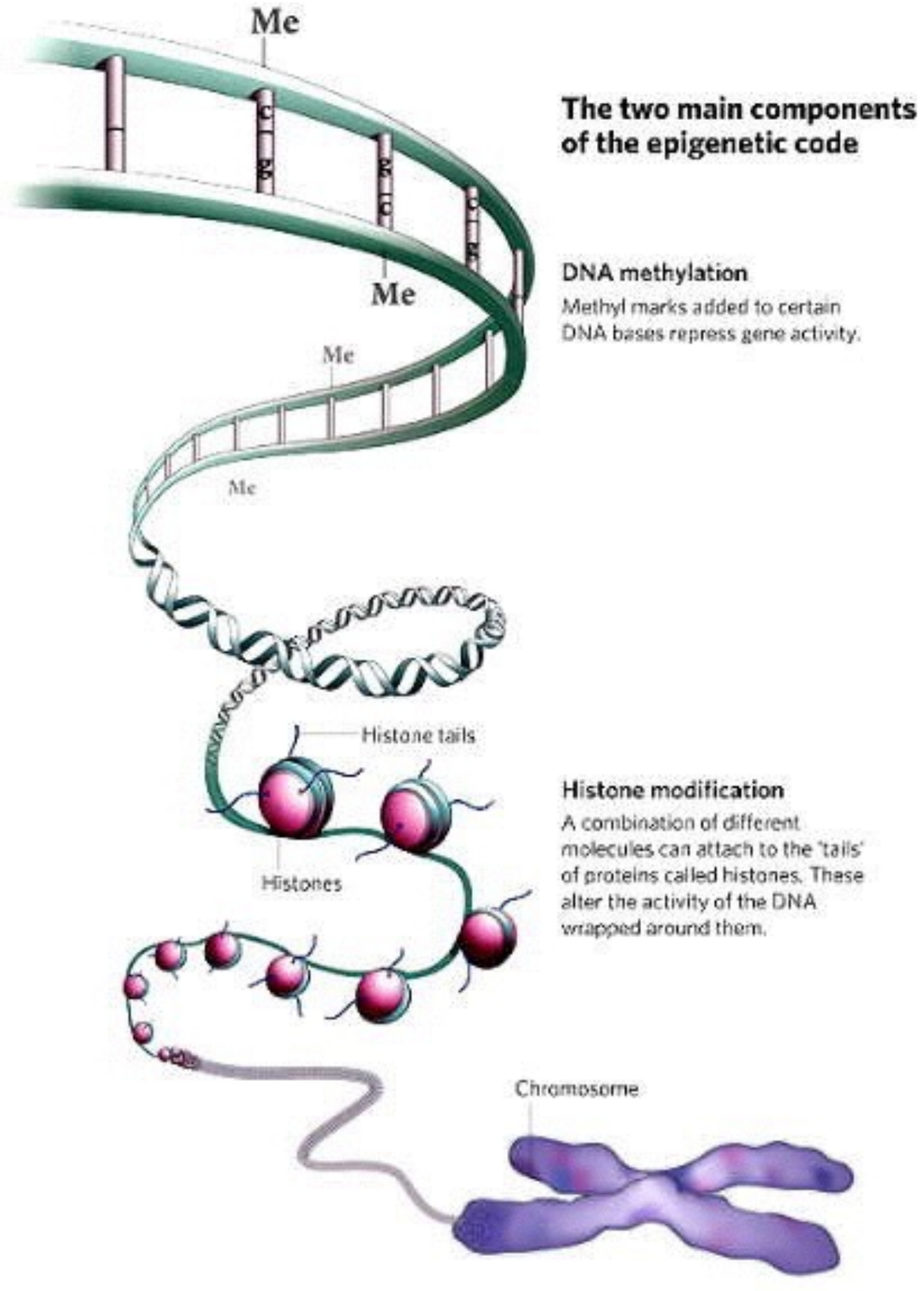
Goal: Capture enhancer architectural changes (like TFBS turnover) while they are occurring. Record what is possible and what is not possible.

- **Team TFBS:** Map TFBS
- **Data Exploration:** Analyze, ex. when Event 1 happens what is the likelihood of Event 2

DNA Accessibility

We cannot just look at linear sequence conservation of enhancer architecture…

because DNA is not linear, and it's structural changes dictate DNA activity through **DNA accessibility.**



DNA Accessibility (Epigenetics)

- DNA that is packaged and folded likely has no activity (function)
- DNA accessibility is a way to control gene expression
- DNA accessibility changes during developmental time
- Patterns of DNA packaging is heritable between generations

DNA Accessibility

How do we know DNA accessibility in the sequences we have? - yes (and no)

- DNase Seq : DNase I hypersensitity sequencing.
  - denoting DNA accessibility in non-coding sites. Which changes in developmental time.

In summary :

- We have a lot of data
- That has never before been combine
- We have a lot of questions to answer

Motif/Transcription Factor Binding Site

* Higher order binding. TFBS aggregation patterning. Cooperative binding.
* Number of TFs within each enhancer (quantified by how many times each unique TF is present)
* Binding Site Turnover - presence in same orthologous position vs presence overall
* Length between same and different types of TFs (homo-dimeric or hetero-dimeric factor, and overlapping motifs = competitive binding by steric hindrance)
* Biased arrangements of TFBS