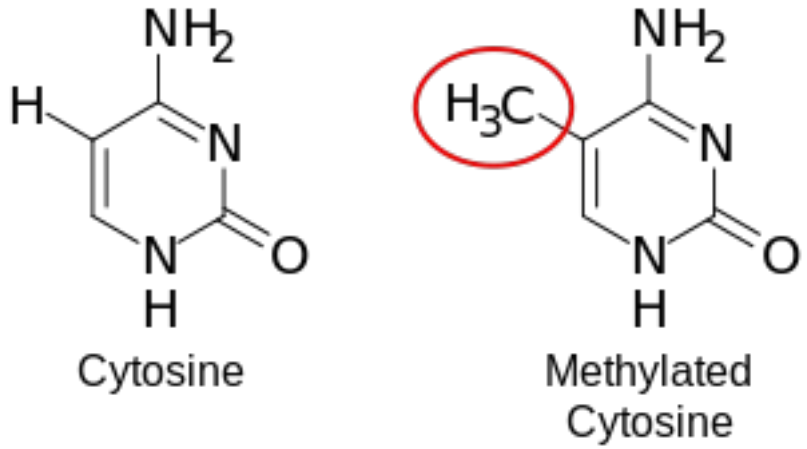
Methylation Analysis Centromere Discovery

BIS180L

Julin Maloof and Matt Davis

Methylation Overview

- Cytosine (C) Residues can be modified by the addition of a methyl group



https://commons.wikimedia.org/wiki/File:DNA_methylation.svg#/media/File:DNA_methylation.svg

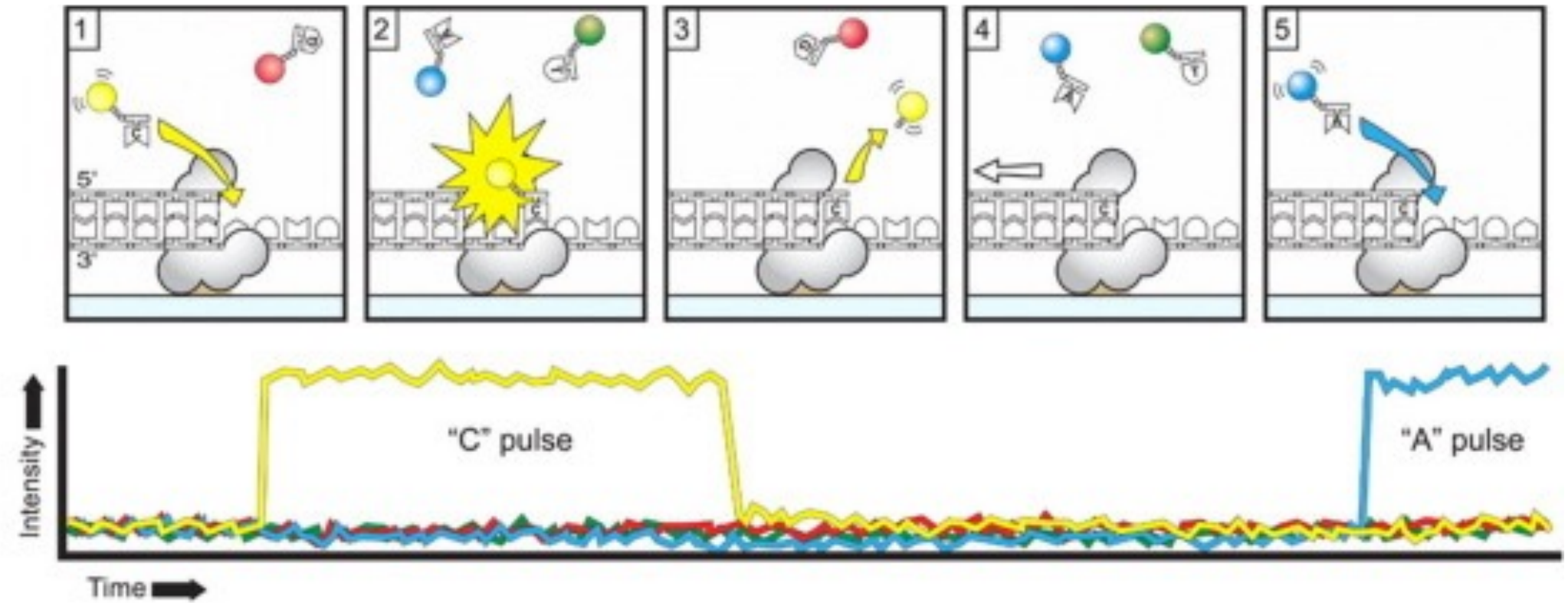
Methylation Context and Roles

- Cytosine methylation can occur in different sequence contexts
  - Animals: “CG” aka ”CpG”
  - Plants: “CG”, “CHH”, “CHG”. Each controlled by different enzymes and with likely different roles
- Cytosine methylation
  - Silencing transposable elements
  - Regulation of gene expression
    - Can be dynamic
    - Role in development
    - Role in response to environment
    - Role in imprinting of parental genome

https://commons.wikimedia.org/wiki/File:DNA_methylation.svg#/media/File:DNA_methylation.svg

Cytosine methylation can be measured by PacBio

- PacBio
  - delayed kinetics of base addition.
  - Complicated signal, requires deep-learning to decode
  - Accurate
  - CpG only



Cytosine methylation can be measured by ONT

- ONT
  - C vs 5mC cause different current signal across membrane
  - Accurate
  - All Contexts. Probably better for plants
- Unfortunately, we only have PacBio

Goals / Questions

- Extract the `CG` methylation data for the *S. diversifolius* genome.
- Determine if the `CG` methylation marks follow expected patterns in genes.
- Calculate methylation proportion for each gene and ask if that corelates with gene expression levels
- Display methylation levels across the chromosome and ask if that pattern is non-random at the chromosome level.

Some file types that you will meet

- ".gff" – General Feature Format
  - Standard for describing the location of features (genes, introns, exons, repeats, etc) in a genome.

```
(base) exouser@julin-2:/revio-data/methylation$ head HiFiasm_S.div.small.gff3
##gff-version 3
ptg000341l      maker   gene    12350   13413   .       -       .       ID=Sdiv_ptg000341l_0001;Name=Sdiv_ptg000341l_0001;Alias=maker-p
A:1.20.85.10,InterPro:IPR000484,InterPro:IPR036854,PANTHER:PTHR33149,PFAM:PF00124,PRINTS:PR00256,PROSITE:PS00244,SUPERFAMILY:SSF81483;N
n D1 (Lepidium virginicum);Ontology_term=GO:0009772,GO:0019684,GO:0045156;Note=Similar to psbA: Photosystem II protein D1 (Lepidium vir
terPro:IPR000484,InterPro:IPR036854,PANTHER:PTHR33149,PFAM:PF00124,PRINTS:PR00256,PROSITE:PS00244,SUPERFAMILY:SSF81483;Ontology_term=GO
ptg000341l      maker   mRNA    12350   13413   .       -       .       ID=Sdiv_ptg000341l_0001-R;Parent=Sdiv_ptg000341l_0001;Name=Sdiv
snap-gene-0.6-mRNA-1;Dbxref=Gene3D:G3DSA:1.20.85.10,InterPro:IPR000484,InterPro:IPR036854,PANTHER:PTHR33149,PFAM:PF00124,PRINTS:PR00256
=Similar to psbA: Photosystem II protein D1 (Lepidium virginicum);Ontology_term=GO:0009772,GO:0019684,GO:0045156;_AED=0.48;_QI=0|0|0|1|
A: Photosystem II protein D1 (Lepidium virginicum);Dbxref=Gene3D:G3DSA:1.20.85.10,InterPro:IPR000484,InterPro:IPR036854,PANTHER:PTHR331
0244,SUPERFAMILY:SSF81483;Ontology_term=GO:0009772,GO:0019684,GO:0045156;
ptg000341l      maker   exon    13216   13413   .       -       .       ID=Sdiv_ptg000341l_0001-R:5;Parent=Sdiv_ptg000341l_0001-R;
ptg000341l      maker   exon    13086   13175   .       -       .       ID=Sdiv_ptg000341l_0001-R:4;Parent=Sdiv_ptg000341l_0001-R;
ptg000341l      maker   exon    12656   13043   .       -       .       ID=Sdiv_ptg000341l_0001-R:3;Parent=Sdiv_ptg000341l_0001-R;
ptg000341l      maker   exon    12568   12585   .       -       .       ID=Sdiv_ptg000341l_0001-R:2;Parent=Sdiv_ptg000341l_0001-R;
ptg000341l      maker   exon    12350   12546   .       -       .       ID=Sdiv_ptg000341l_0001-R:1;Parent=Sdiv_ptg000341l_0001-R;
ptg000341l      maker   CDS     13216   13413   .       -       0       ID=Sdiv_ptg000341l_0001-R:cds;Parent=Sdiv_ptg000341l_0001-R;
ptg000341l      maker   CDS     13086   13175   .       -       0       ID=Sdiv_ptg000341l_0001-R:cds;Parent=Sdiv_ptg000341l_0001-R;
(base) exouser@julin-2:/revio-data/methylation$
```

Some file types that you will meet

- “bed” – browser extensible data
  - Simple format to describe an attribute (e.g. CpG) at locations in a genome
  - First 3 columns are always sequenceID, start, end.
  - Remaining columns are more flexible

```
(base) exouser@julin-2:/revio-data/methylation$ head S_div_cpg.combined.bed
ptg000001l      4568    4569    93.7    Total   35      33      2       94.3
ptg000001l      4635    4636    94.8    Total   35      34      1       97.1
ptg000001l      4641    4642    94.3    Total   35      34      1       97.1
ptg000001l      4695    4696    73.2    Total   35      26      9       74.3
ptg000001l      4720    4721    95.1    Total   35      34      1       97.1
ptg000001l      4733    4734    95.5    Total   35      34      1       97.1
ptg000001l      4747    4748    94.1    Total   35      33      2       94.3
ptg000001l      4793    4794    95.3    Total   37      36      1       97.3
ptg000001l      4796    4797    95.7    Total   37      36      1       97.3
ptg000001l      4812    4813    96.5    Total   37      36      1       97.3
```

An important R object class: GRanges

- The “GRanges” class is used for storing genomic data
  - Both .bed and .gff files can be represented as Granges
  - Excellent set operations, overlaps, etc.
  - Powerful but can be painful

```
> gff
GRanges object with 33433 ranges and 3 metadata columns:
            seqnames            ranges strand |     type                  ID                Name
               <Rle>         <IRanges>  <Rle> | <factor>         <character>         <character>
      [1] ptg000001l         5286-9436      + |     mRNA Sdiv_ptg000001l_0001-R Sdiv_ptg000001l_0001-R
      [2] ptg000001l       11141-11320      + |     mRNA Sdiv_ptg000001l_0003-R Sdiv_ptg000001l_0003-R
      [3] ptg000001l       18862-20472      + |     mRNA Sdiv_ptg000001l_0004-R Sdiv_ptg000001l_0004-R
      [4] ptg000001l       32918-33824      + |     mRNA Sdiv_ptg000001l_0007-R Sdiv_ptg000001l_0007-R
      [5] ptg000001l       34123-34566      + |     mRNA Sdiv_ptg000001l_0008-R Sdiv_ptg000001l_0008-R
```

An important R object class: GRanges

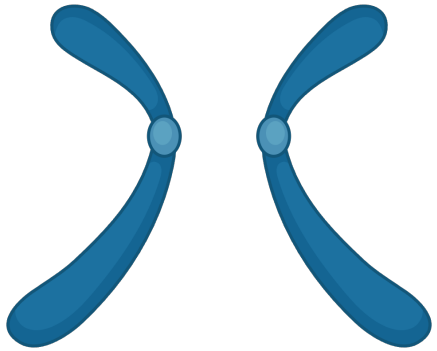
- The “GRanges” class is used for storing genomic data
  - Both .bed and .gff files can be represented as Granges
  - Excellent set operations, overlaps, etc.
  - Powerful but can be painful

```
> cpg
GRanges object with 7218137 ranges and 2 metadata columns:
             seqnames              ranges strand |     score  coverage
                <Rle>           <IRanges>  <Rle> | <numeric> <integer>
        [1] ptg000001l         4568-4569      * |      93.7         35
        [2] ptg000001l         4635-4636      * |      94.8         35
        [3] ptg000001l         4641-4642      * |      94.3         35
        [4] ptg000001l         4695-4696      * |      73.2         35
        [5] ptg000001l         4720-4721      * |      95.1         35
```

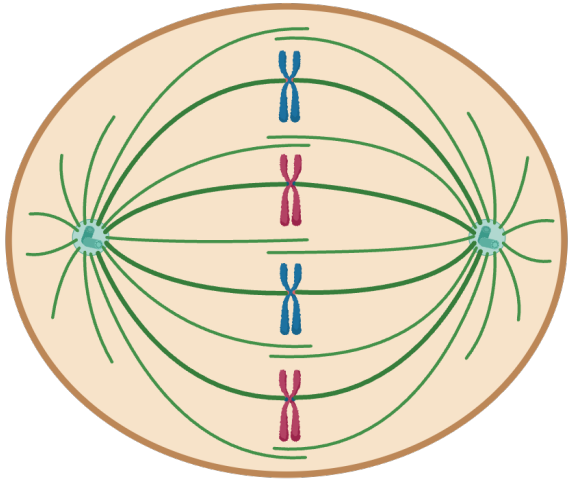
Basic Outline

1. Map Hifi Reads back to the genome assembly (you have already done this)
2. Get methylation percentage for each `CG` site in the genome. (you or I have already done this)
3. Load the data into R.
4. Use R to summarize and analyze the methylation data.

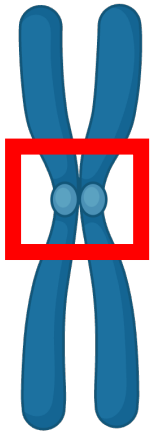
What are centromeres?



They link pairs of sister chromosomes



They are where the spindle fibers attach during mitosis and meiosis

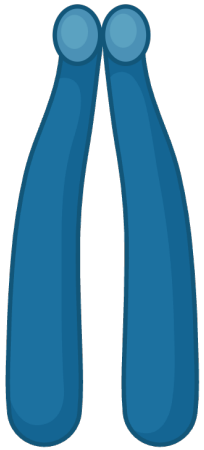


They are what give chromosomes their characteristic X shape

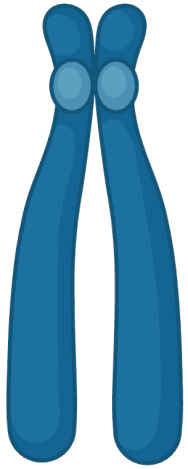
AATTGGTTAATTGGTTAATTGGTTAATTGGTTAATTGGTT

A lot like telomeres, centromeres are composed of repetitive sequences

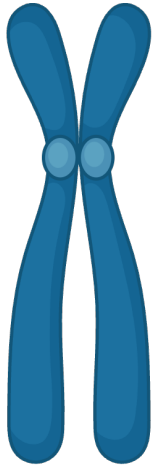
Types of centromeres



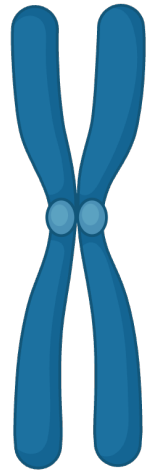
Telocentric



Acrocentric

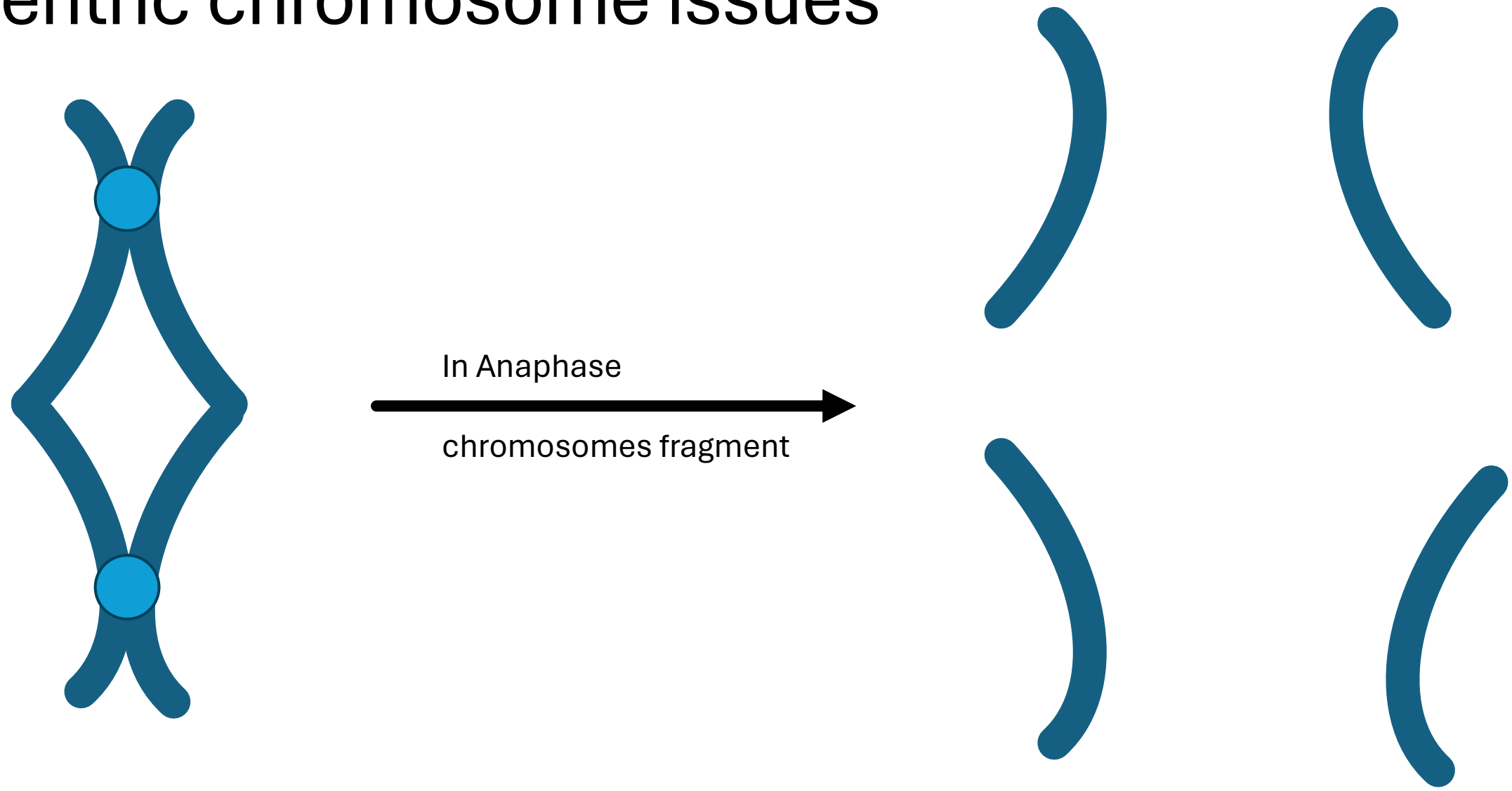


Sub-Metacentric



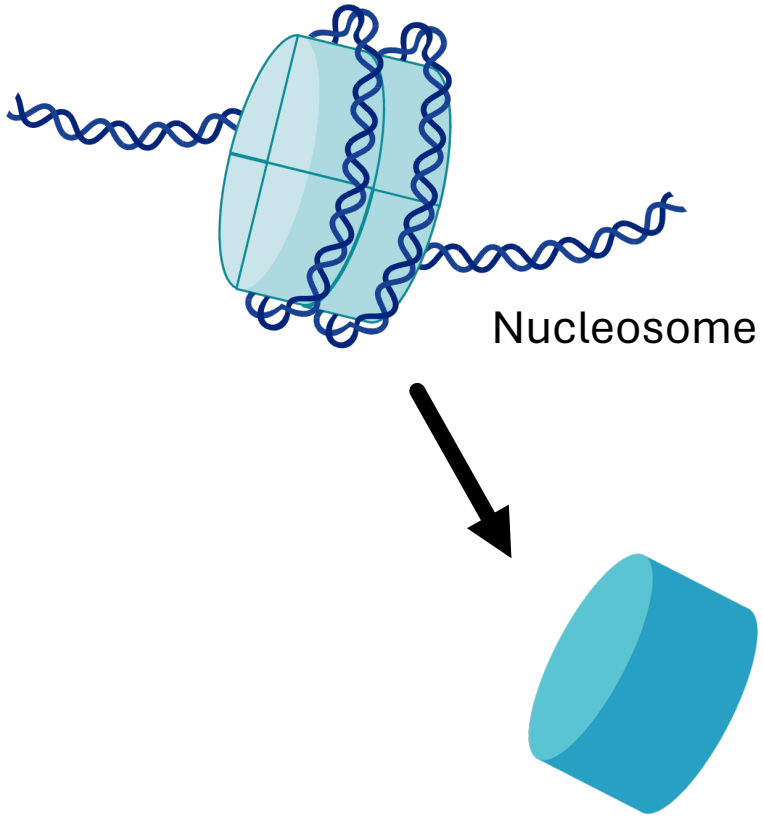
Metacentric

Dicentric chromosome issues



How have we detected centromeres?

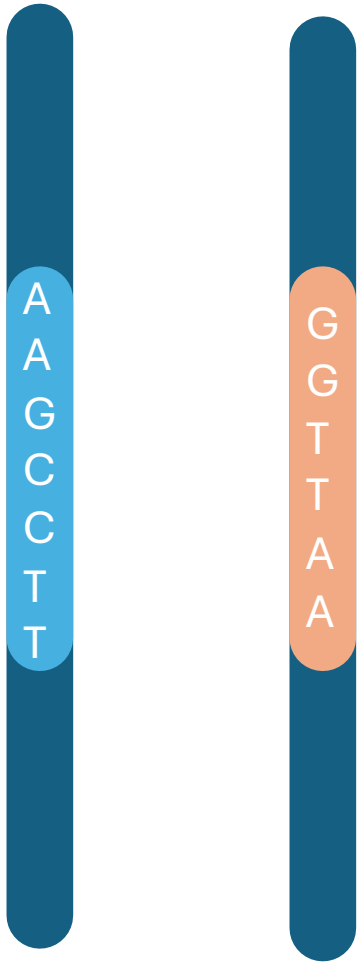
Telomere sequence is very conserved



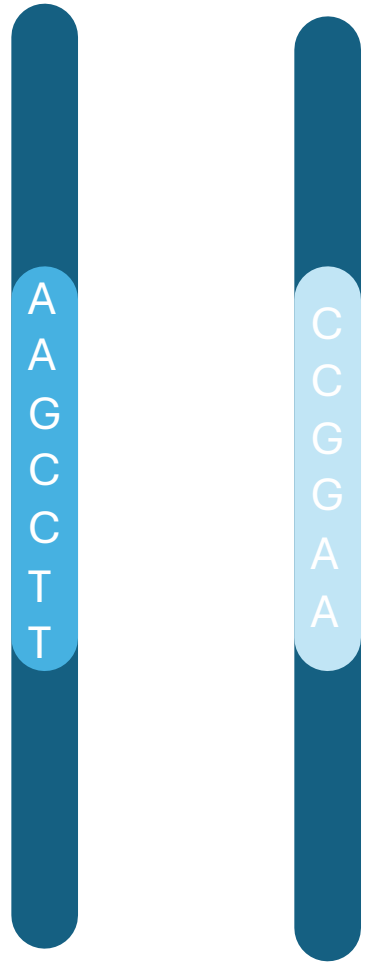
Centromeres are epigenetically conserved

Why is it so challenging to detect them bioinformatically?

Centromere repeats are not conserved and much longer



The centromeric sequence can vary within a genus



The centromeric sequence can vary between chromosomes

How are we going to approach centromere detection today?

Looking for regions of low sequence diversity. AKA regions with a lot of repeats

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

RepeatOBserver: tandem repeat visualization and centromere detection

Cassandra Elphinstone, Rob Elphinstone, Marco Todesco, Loren Rieseberg

doi: https://doi.org/10.1101/2023.12.30.573697

This article is a preprint and has not been certified by peer review [what does this mean?].

Unfortunately this takes a long time. We have run this for you and you will access the data