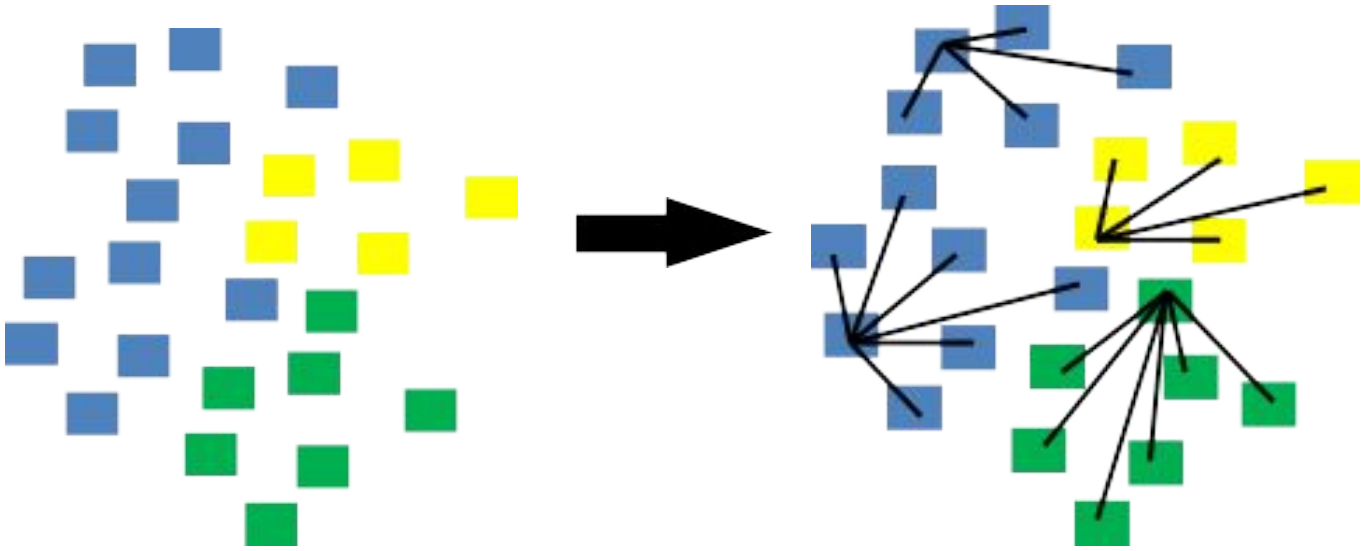
De novo genome assembly



By Dr. Richard Allen White III

Lecture 2 - Sep 3rd, 2019

De novo genome assembly

Concepts:

- What is a genome?
- How do we obtain genomes?
- What is an *de novo* assembly?
- How do we assemble genomes?

Learning Objectives:

- What is various file formats for genomics?
- More UNIX command line

De novo genome assembly

What is a genome?

What is a metagenome?

De novo genome assembly

What is a genome?

- The totality of an organisms complete set of DNA including all the genes encoded present within the DNA of the organism

What is a metagenome?

De novo genome assembly

What is a genome?

- The totality of an organisms complete set of DNA including all the genes encoded present within the DNA of the organism

What is a metagenome?

- Whole community sampling of all the genomes represented as DNA within a microbial community, representing the functional and taxonomic potential of a ecosystem

How to obtain a bacterial genome?

How to obtain a bacterial genome?

Culturing < 1% can by easily cultured

How to obtain a bacterial genome?

Culturing < 1% can by easily cultured

Single cell genomics

Only a few labs in the world, very incomplete genomes (30% avg, 10-90%)

How to obtain a bacterial genome?

Culturing < 1% can by easily cultured

Single cell genomics

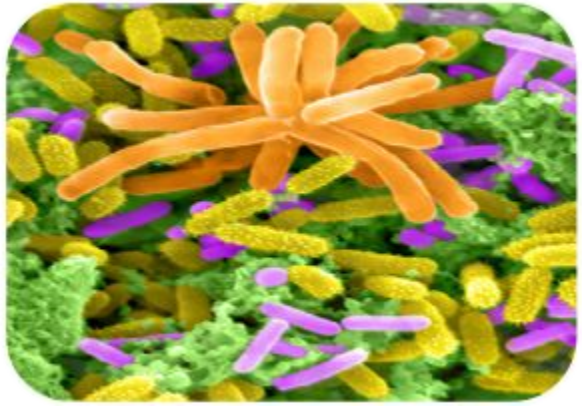
Only a few labs in the world, very incomplete genomes (30% avg, 10-90%)

How to obtain a bacterial genome?

Culturing < 1% can by easily cultured

Single cell genomics

Only a few labs in the world, very incomplete genomes (30% avg, 10-90%)



Metagenomic - population genome binning

De novo assembly terms

De novo assembly terms

ATTACCGG
TTTTCCGG

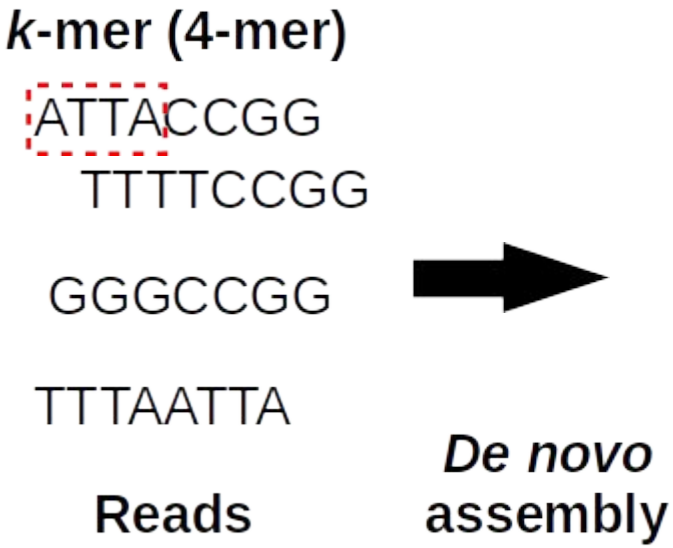
GGGCCGG

TTTAATTA

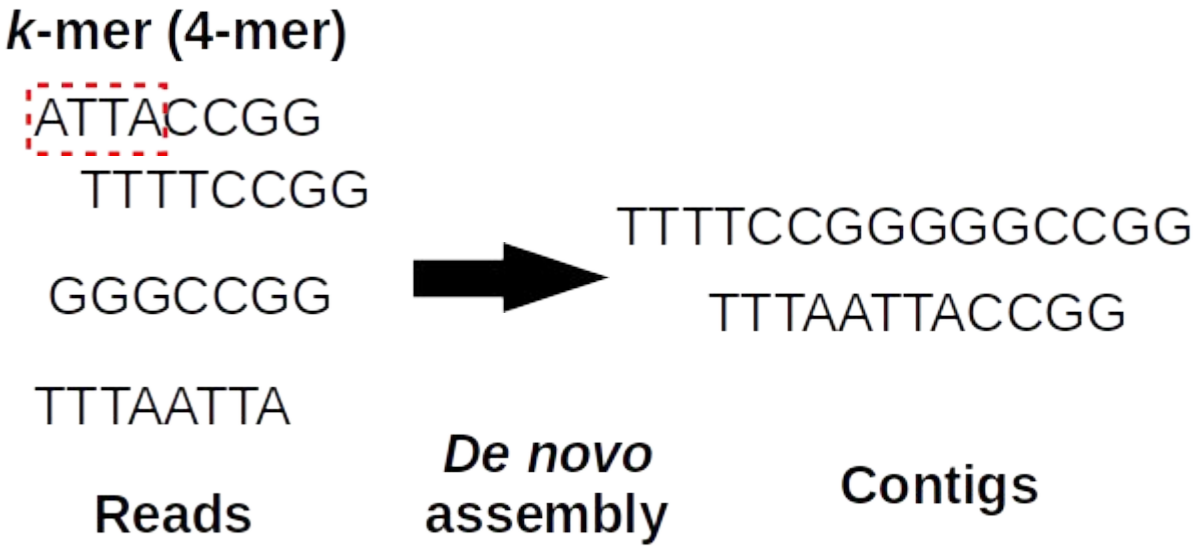
Reads

De novo assembly terms

De novo assembly terms



De novo assembly terms



De novo assembly terms

De novo assembly terms

De novo assembly types

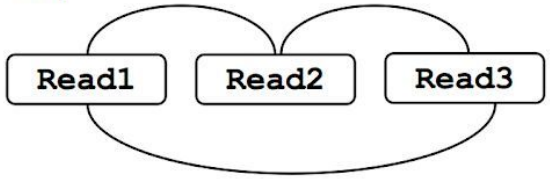
(a) Overlap, Layout, Consensus assembly

(b) De Bruijn graph assembly

De novo assembly types

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps

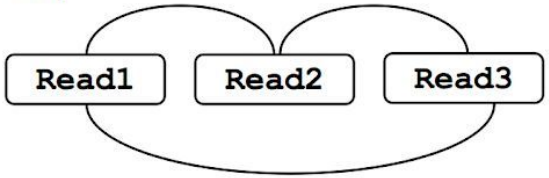


(b) De Bruijn graph assembly

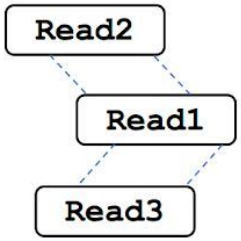
De novo assembly types

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads

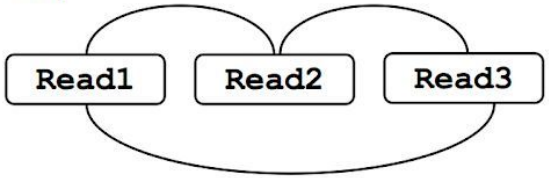


(b) De Bruijn graph assembly

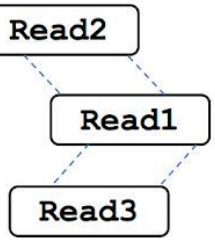
De novo assembly types

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

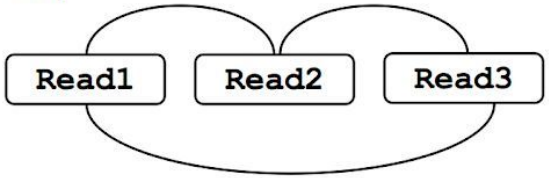
```
CGATTCTA
   TTCTAAGT
 GATTGTAA
--------
CGATTCTAAGT
```

(b) De Bruijn graph assembly

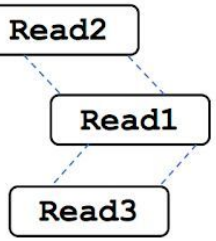
De novo assembly types

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

```
CGATTCTA
   TTCTAAGT
 GATTGTAA
--------
CGATTCTAAGT
```

(b) De Bruijn graph assembly

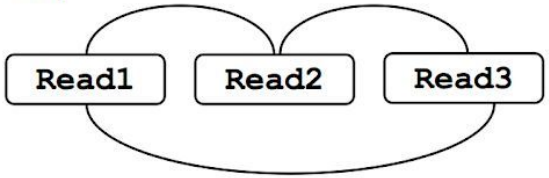
(i) Make kmers

```
Read1: TTCTAAGT    Read2: CGATTCTA    Read3: GATTGTAA
Kmers: TTC         Kmers: CGA         Kmers: GAT
        TCT                GAT                ATT
         CTA                ATT                TTG
          TAA                TTC                TGT
           AAG                TCT                GTA
            AGT                CTA                TAA
```

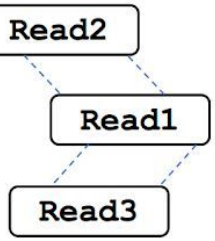
De novo assembly types

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

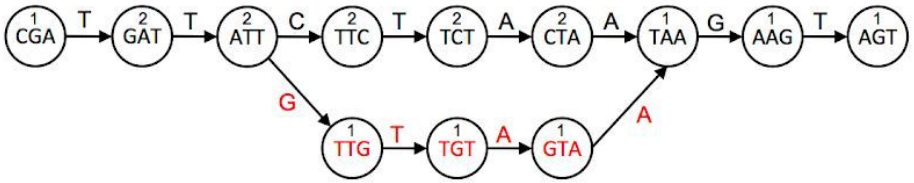
```
CGATTCTA
   TTCTAAGT
GATTGTAA
--------
CGATTCTAAGT
```

(b) De Bruijn graph assembly

(i) Make kmers

```
Read1: TTCTAAGT     Read2: CGATTCTA     Read3: GATTGTAA
Kmers: TTC          Kmers: CGA          Kmers: GAT
        TCT                 GAT                 ATT
         CTA                 ATT                 TTG
          TAA                 TTC                 TGT
           AAG                 TCT                 GTA
            AGT                 CTA                 TAA
```

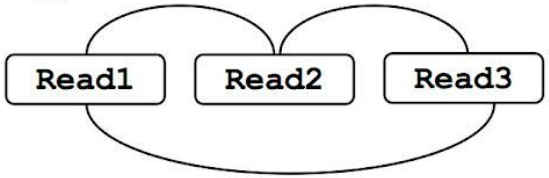
(ii) Build graph



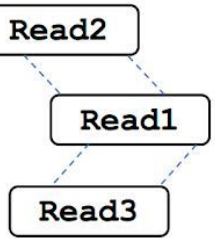
De novo assembly types

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

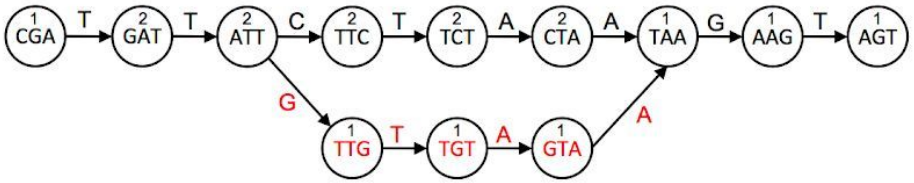
```
CGATTCTA
   TTCTAAGT
 GATTGTAA
----------
CGATTCTAAGT
```

(b) De Bruijn graph assembly

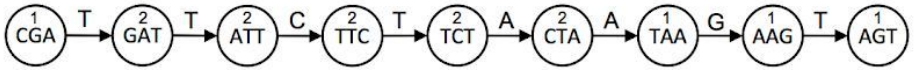
(i) Make kmers

```
Read1: TTCTAAGT     Read2: CGATTCTA     Read3: GATTGTAA
Kmers: TTC          Kmers: CGA          Kmers: GAT
        TCT                 GAT                 ATT
         CTA                 ATT                 TTG
          TAA                 TTC                 TGT
           AAG                 TCT                 GTA
            AGT                 CTA                 TAA
```

(ii) Build graph



(iii) Walk graph and output contigs



CGATTCTAAGT

What is a draft vs complete genome?

- Draft genome contains many contigs (>10)
- Complete genome in bacteria, is usually (<5 contigs) often containing a circular chromosome but can contain a linear chromosome (e.g., *Rhodobacter*), with plasmids.

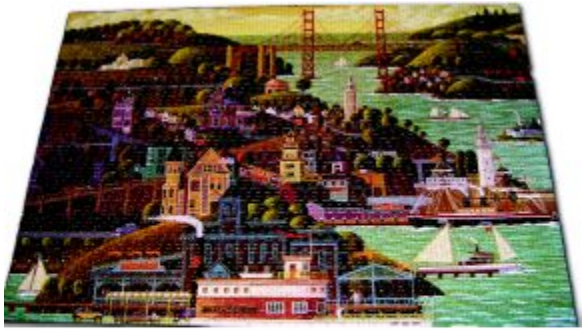
Why do we need complete genomes?

- Functional studies demand an error-free genome sequence as a starting point[1]
- Availability of data on genome organization provides biological insights[1]
- A complete genome sequence is a permanent, valuable scientific resource[1]

[1]Fraser et al., 2002 JBac
https://jb.asm.org/content/184/23/6403

De novo assembly genome vs. community

De novo assembly genome vs. community



Genome

De novo assembly genome vs. community

Genome

Reads

De novo assembly genome vs. community

Genome

Reads

Contigs

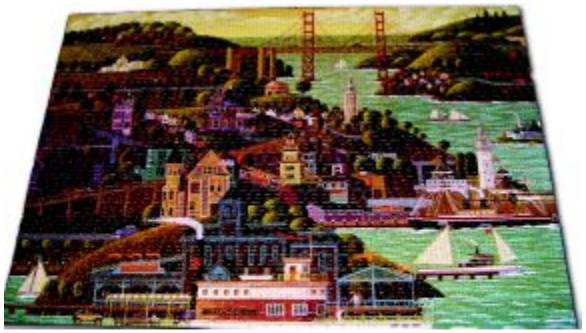
De novo assembly genome vs. community

Genome

Reads

Contigs

De novo assembly genome vs. community



Genome

Reads

Contigs

Metagenome

De novo assembly genome vs. community

Genome

Reads

Contigs

Metagenome

Reads

De novo assembly genome vs. community

Genome

Reads

Contigs

Metagenome

Reads

Contigs

De novo assembly genome vs. community

Genome

Reads

Contigs

Metagenome

Reads

Contigs

Metagenomic assembly (Soil)

[1]Howe *et al.*, 2014 PNAS.
[2]Li *et al.*, 2015 Bioinformatics.

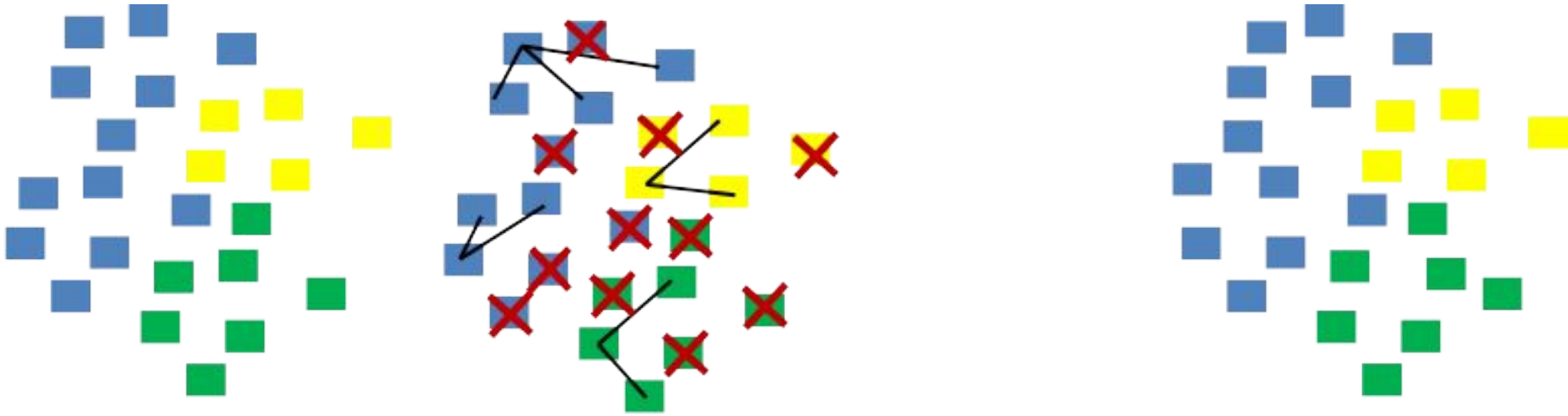
Metagenomic assembly (Soil)

[1]Howe *et al.*, 2014 PNAS.
[2]Li *et al.*, 2015 Bioinformatics.

Metagenomic assembly (Soil)

[1]Howe *et al.*, 2014 PNAS.
[2]Li *et al.*, 2015 Bioinformatics.

Metagenomic assembly (Soil)

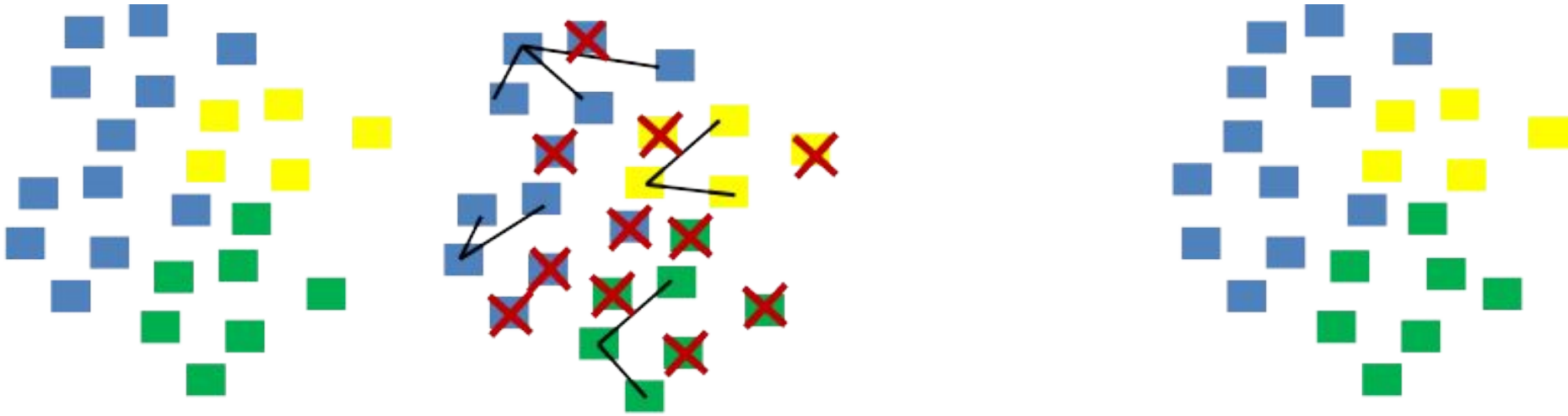


Divide and conquer

- 9 contigs
- >10 kbp
- 10% mapping

[1]Howe *et al.*, 2014 PNAS.
[2]Li *et al.*, 2015 Bioinformatics.

Metagenomic assembly (Soil)

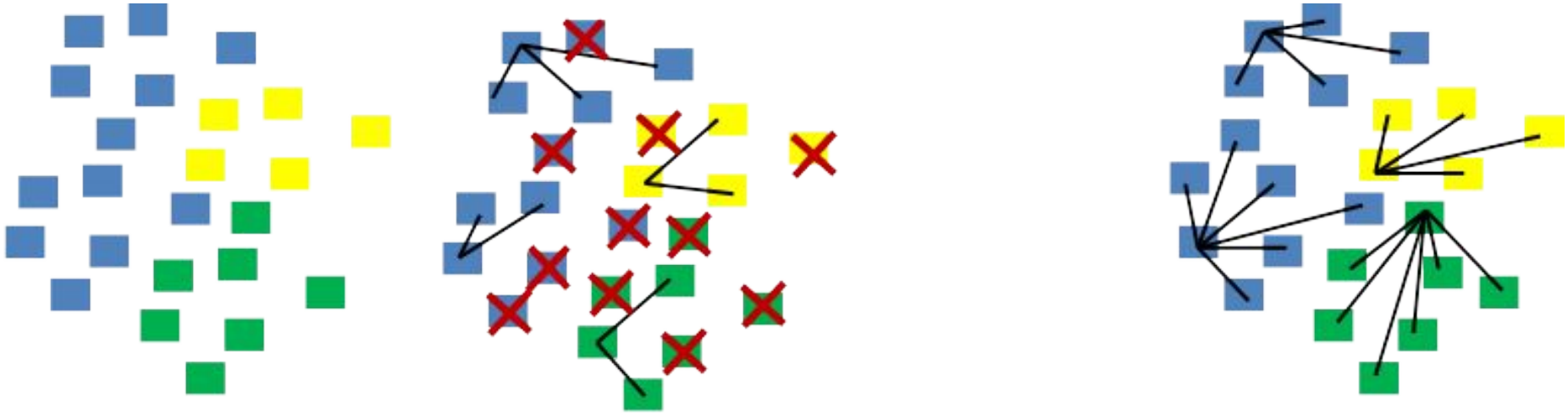


Divide and conquer

- 9 contigs
- >10 kbp
- 10% mapping

[1]Howe *et al.*, 2014 PNAS.
[2]Li *et al.*, 2015 Bioinformatics.

Metagenomic assembly (Soil)

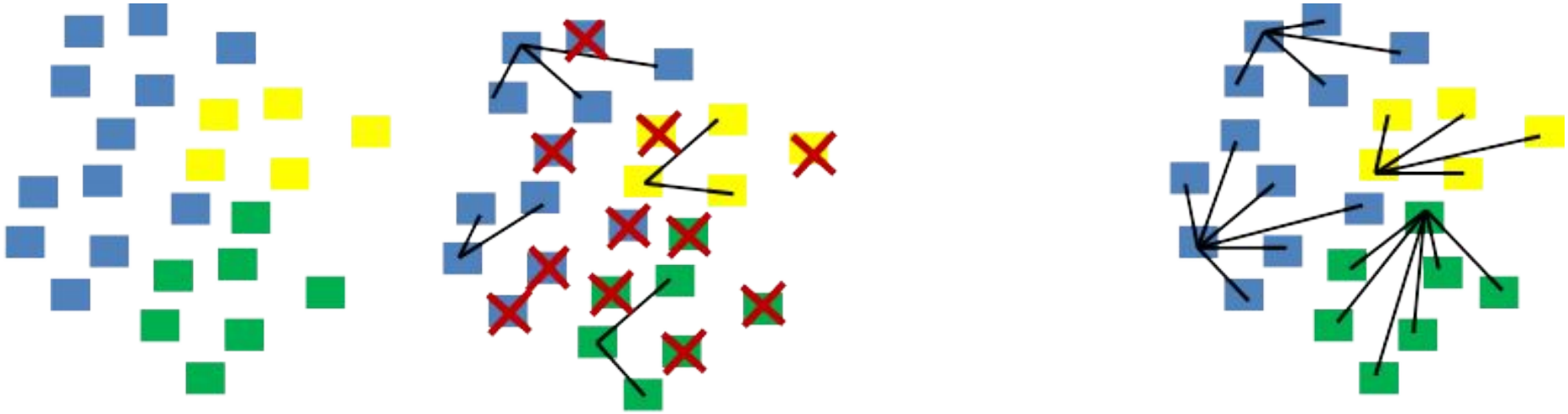


Divide and conquer

- 9 contigs
- >10 kbp
- 10% mapping

[1]Howe *et al.*, 2014 PNAS.
[2]Li *et al.*, 2015 Bioinformatics.

Metagenomic assembly (Soil)



[1]Divide and conquer
- 9 contigs
- >10 kbp
- 10% mapping

[2]Store succinct and go
- 100 contigs
- >10 kbp
- 30% mapping

[1]Howe *et al.*, 2014 PNAS.
[2]Li *et al.*, 2015 Bioinformatics.

De novo metagenomic assembly of genomes

De novo metagenomic assembly of genomes

Soil Metagenome

De novo metagenomic assembly of genomes

Soil Metagenome

Binning

De novo metagenomic assembly of genomes

Soil Metagenome

Binning

HOW?

-GC content
-Abundance/Coverage
-Tetranucleotide frequency
-Composition
-Statistical method

De novo metagenomic assembly of genomes

Soil Metagenome

Binning

HOW?

-GC content
-Abundance/Coverage
-Tetranucleotide frequency
-Composition
-Statistical method

MAGs are?

A single genome isn't a community metagenome?

- High number of types and sizes of genomes present with very sparse sampling leading to low coverage

- Lots of data required usually short reads are used which requires high RAM (working memory) required to assemble

- De novo assembly of soil metagenomes (most complex) yield few contigs and few of the contigs have representative coverage (few mapped reads)

- Few or any computational tools that scale to the large amounts of data required for analysis of data

File formats in genomics

*.fna, .fasta, .fa	?
*.faa	?
*.gff	?
*.gbk	?
*.fq or fastq	?
*.gbk	?

Read vs. sequence?

Contig?

Scaffold?

kmer?

File formats in genomics (fasta)

>sequence1
ATACTCTACTCGTCTCATATCAT
>sequence2
GCGCGCGNCAGCGATCTCTCA
>sequence3
TTTCGCGNNCAGCGATCTCTC

Nucleotide fasta extensions - .fa, .fna, .ffa, .fasta, .contig, .scaf
Best .fna (fasta nucleotide)

File formats in genomics (.faa)

>sequence1
MLQLQPKKRSNLRIWAG
>sequence2
MSTQLKQEPNHQPSGLL
>sequence3
MTAWLKRIVYTALAAYLLSFW

Protein fasta formats - .faa (fasta amino acid)
IT SHOULD NEVER BE .FASTA or OTHER!

File formats in genomics (fastq)

Example 1

@SIM:1:FCX:1:15:6329:1045 1:N:0:2

TCGCACTCAACGCCCTGCATATGACAAGACAGAATC

+

<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=

Example 2

@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36

GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC

+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36

IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC

File formats in genomics (gff/gtf)

GFF

X Ensembl Repeat2419108 2419128 42 . . hid=trf; hstart=1; hend=21

GTF

1 transcribed_unprocessed_pseudogene gene 11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";

Pull files from NCBI

- For *Rhodobacter sphaeroides* 2.4.1 (pubmed.com)
  - Get contig fasta
  - Gbk file
  - Gff or gtf file
  - Protein fasta

UNIX Review

cd - ?
pwd - ?
mkdir - ?
rm - ?
head - ?
tail - ?
more - ?
| - ?
mv - ?
touch - ?