Session 11 -Theory

Genome Assembly with Short Reads

Date: 19/02/2024, 15:00-17:00

Teacher: **Fernando Cruz** (CNAG)

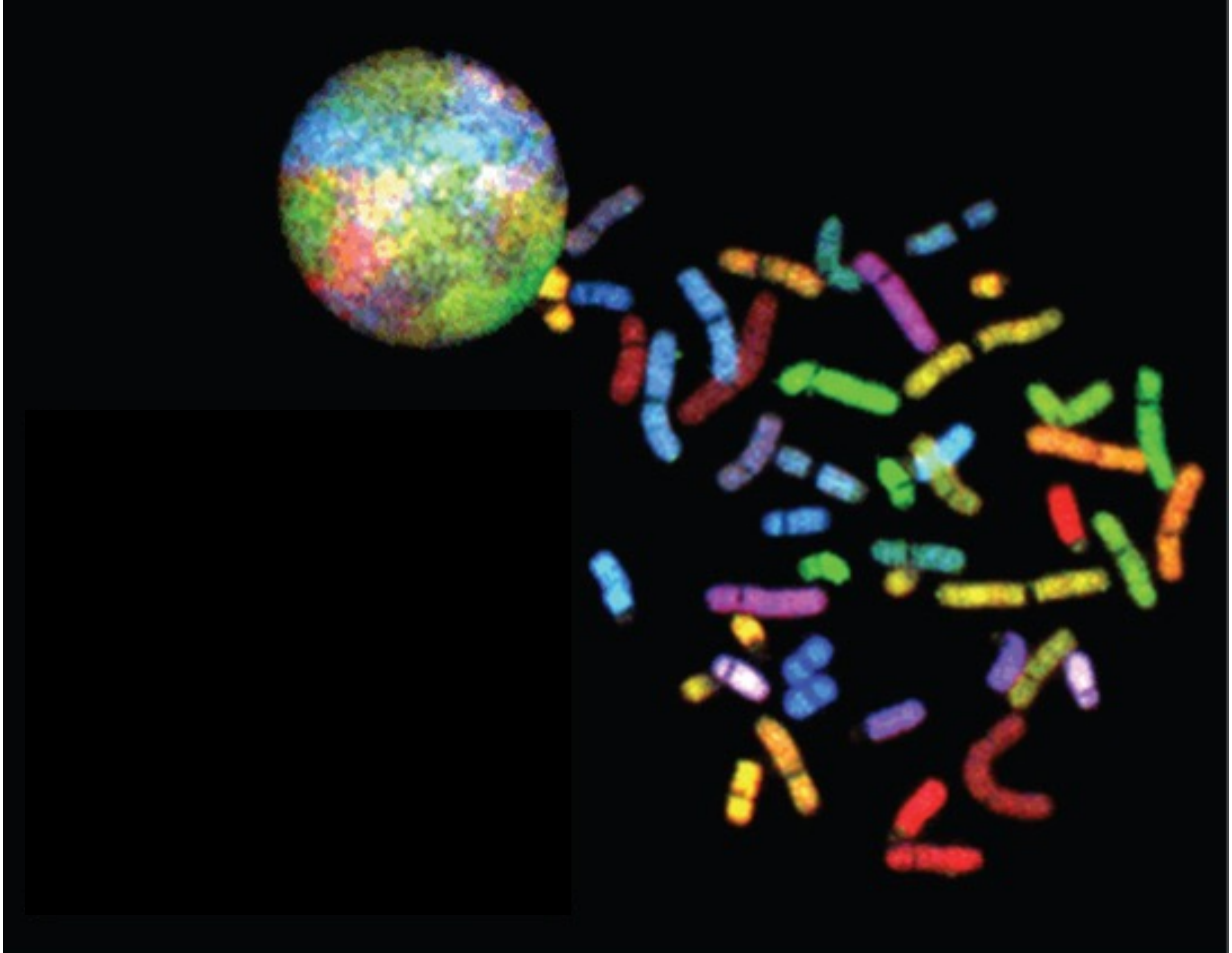
fernando.cruz@prof.esci.upf.edu

Bachelor's Degree in Bioinformatics

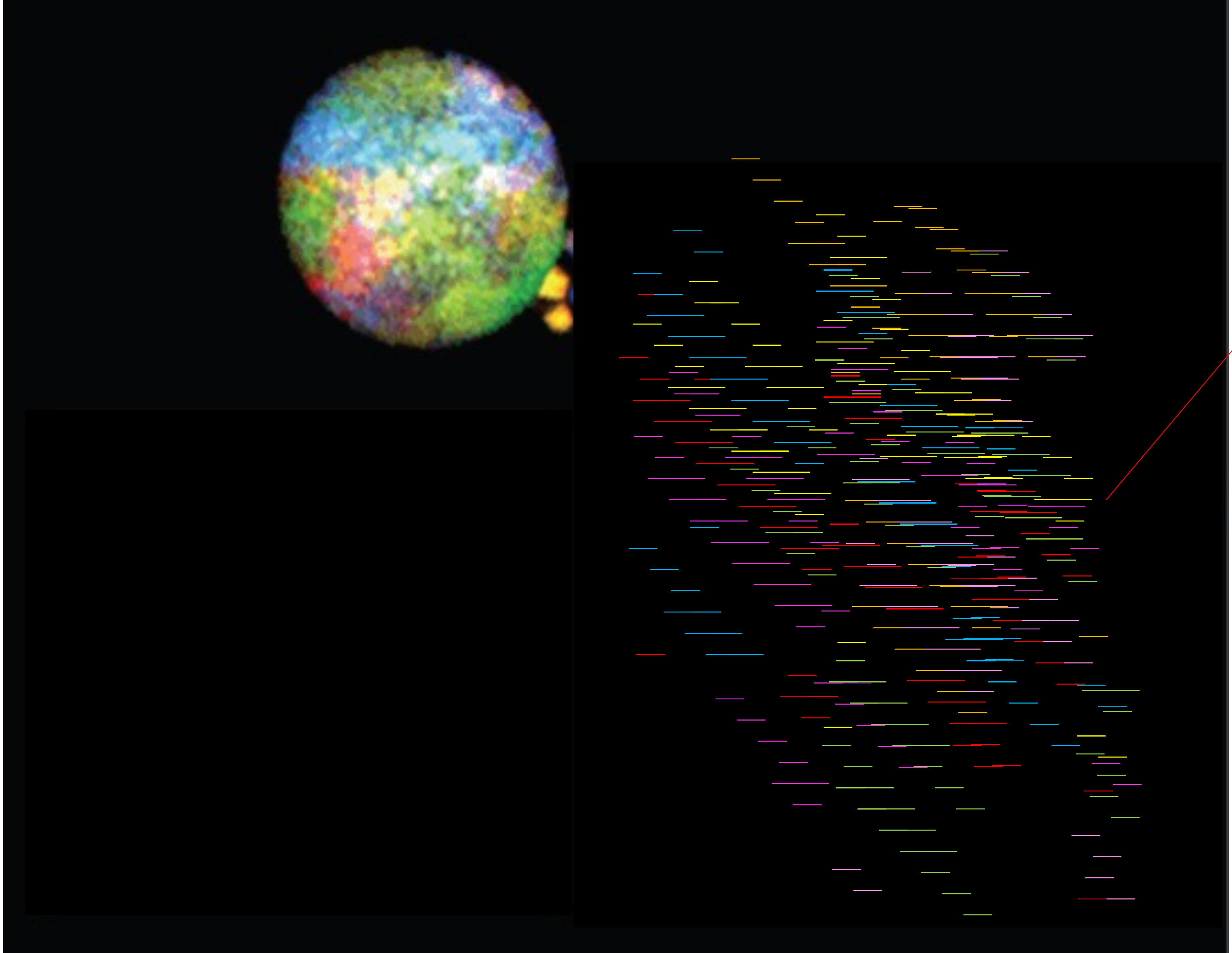
Course 2023-2024

52115 - Algorithms for sequence analysis in Bioinformatics (**ASAB**)

What is ‘*de novo*’ genome assembly?



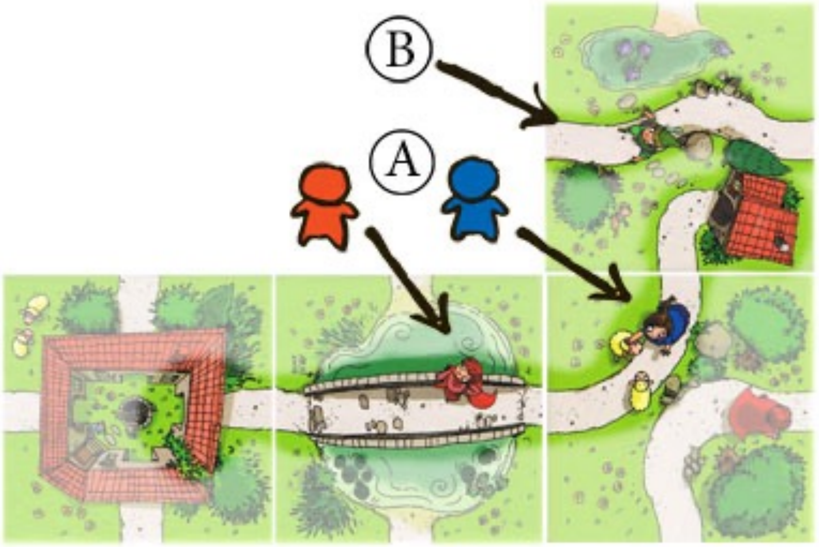
[redacted]



Chrom ???

De novo genome assembly

- Resolving a puzzle
- *The pieces are reads*
  - *Reads* are *"reproductions"* (with varying length and accuracy) of real DNA sequence stretches.
- Closing Paths in *Carcassone*



Short Reads assemble into Contigs

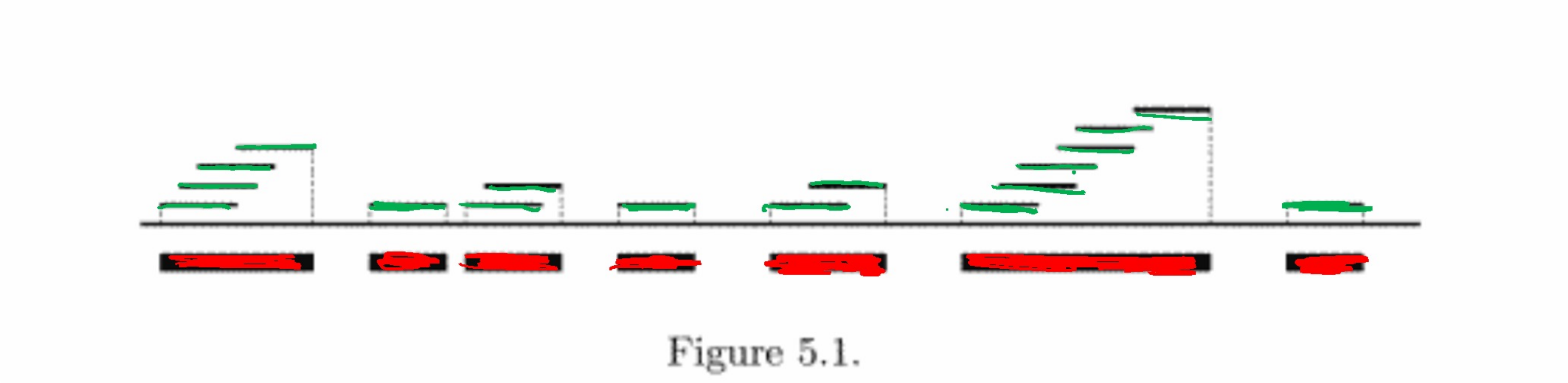
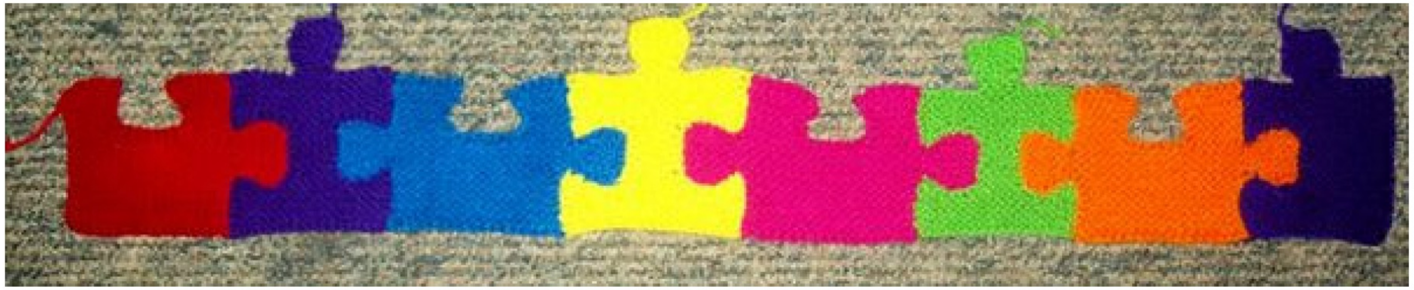
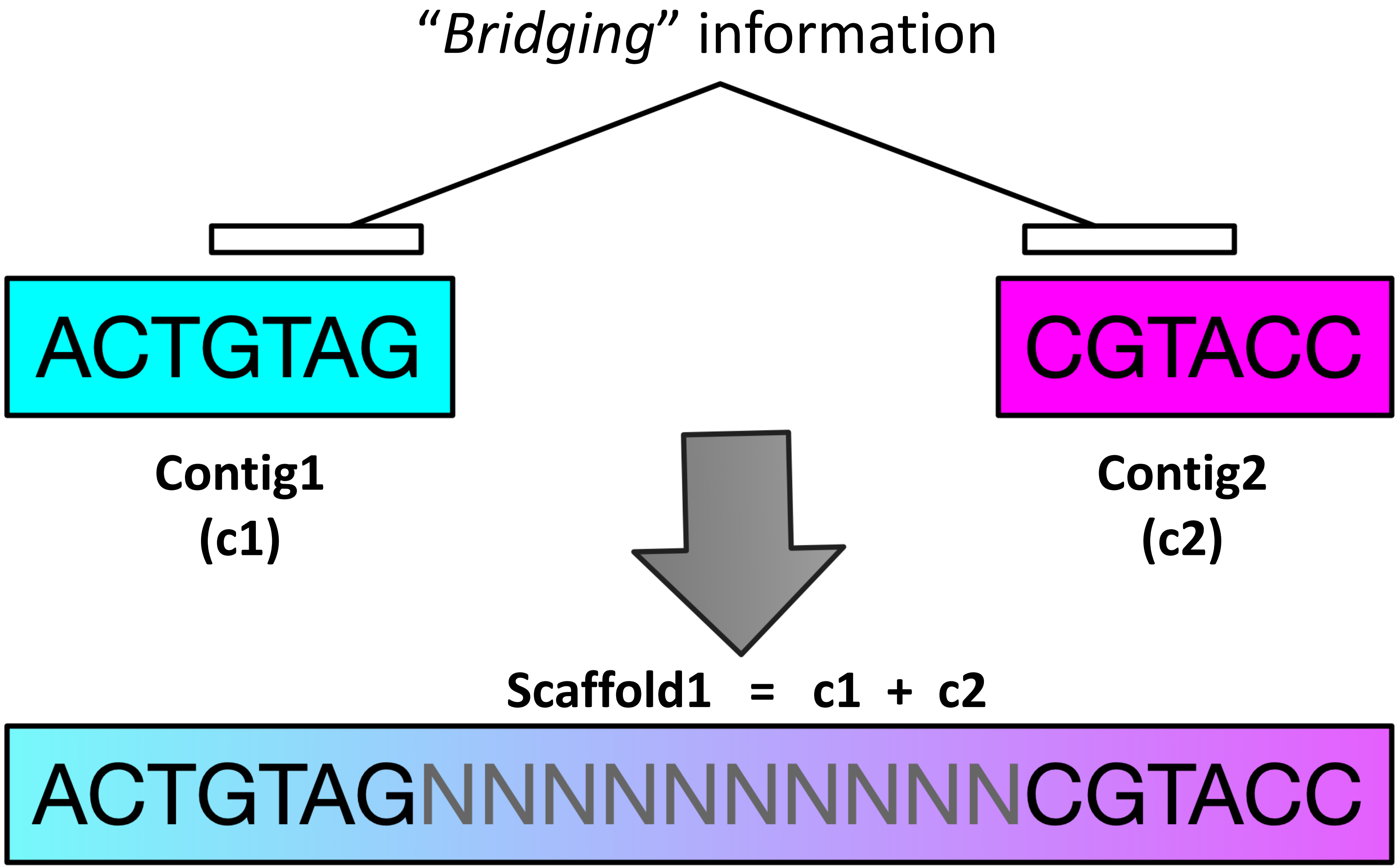


Figure 5.1.



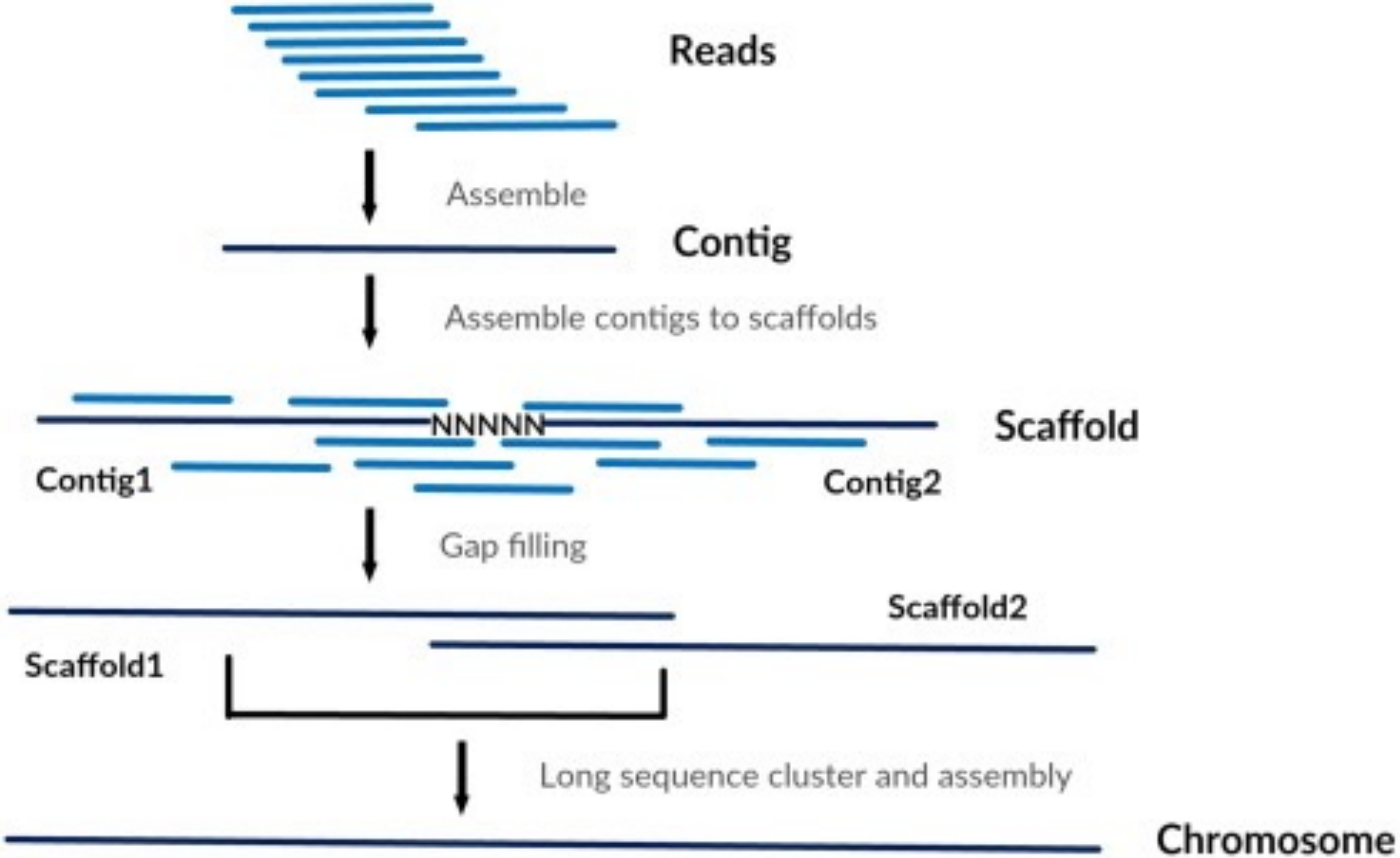
<u>Contigs</u> are blocks of contiguous sequence obtained by assembly of smaller DNA sequences (e.g. reads)

Scaffolds

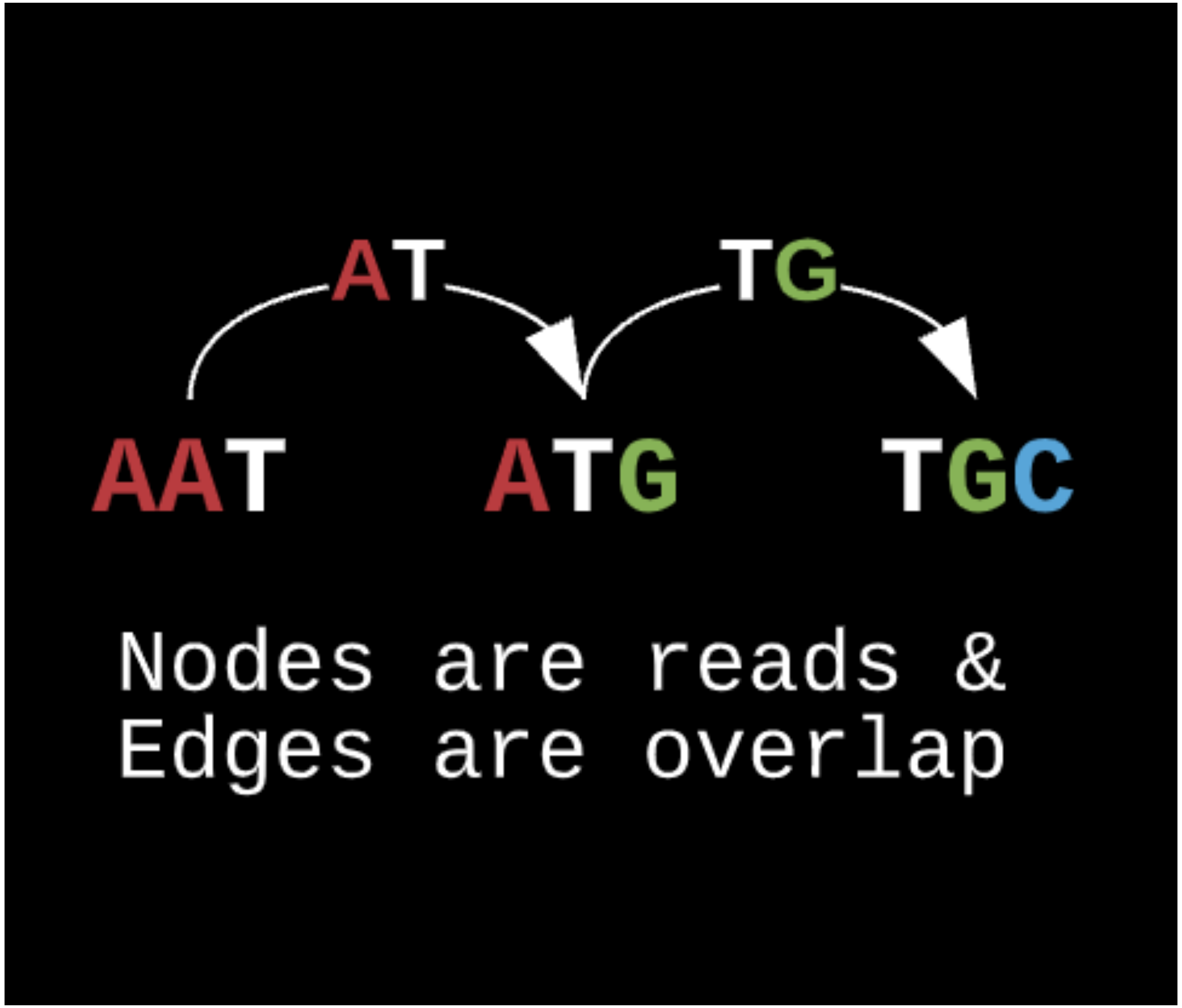


***Scaffolds* are contigs connected by an unknown portion of sequence (gaps)**

Short gaps can be filled



Assembly Graphs

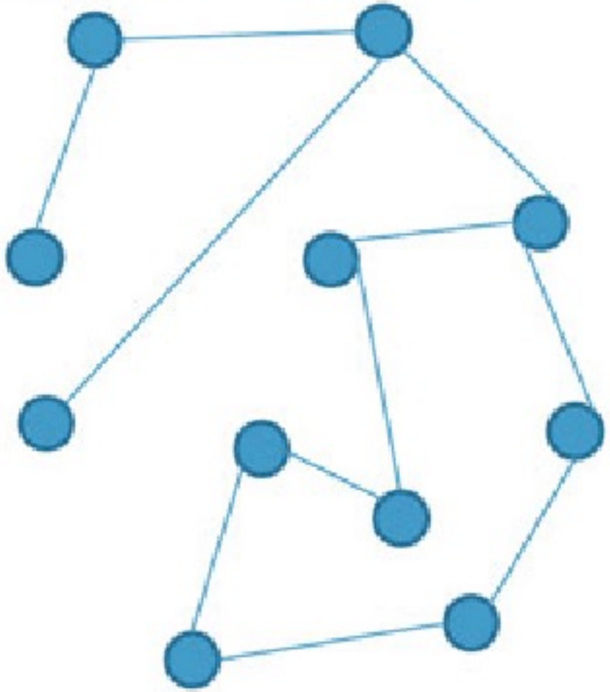


Nodes are reads &
Edges are overlap

Assembly Graphs

Node2 __Edge 2 <-> 3___Node 3

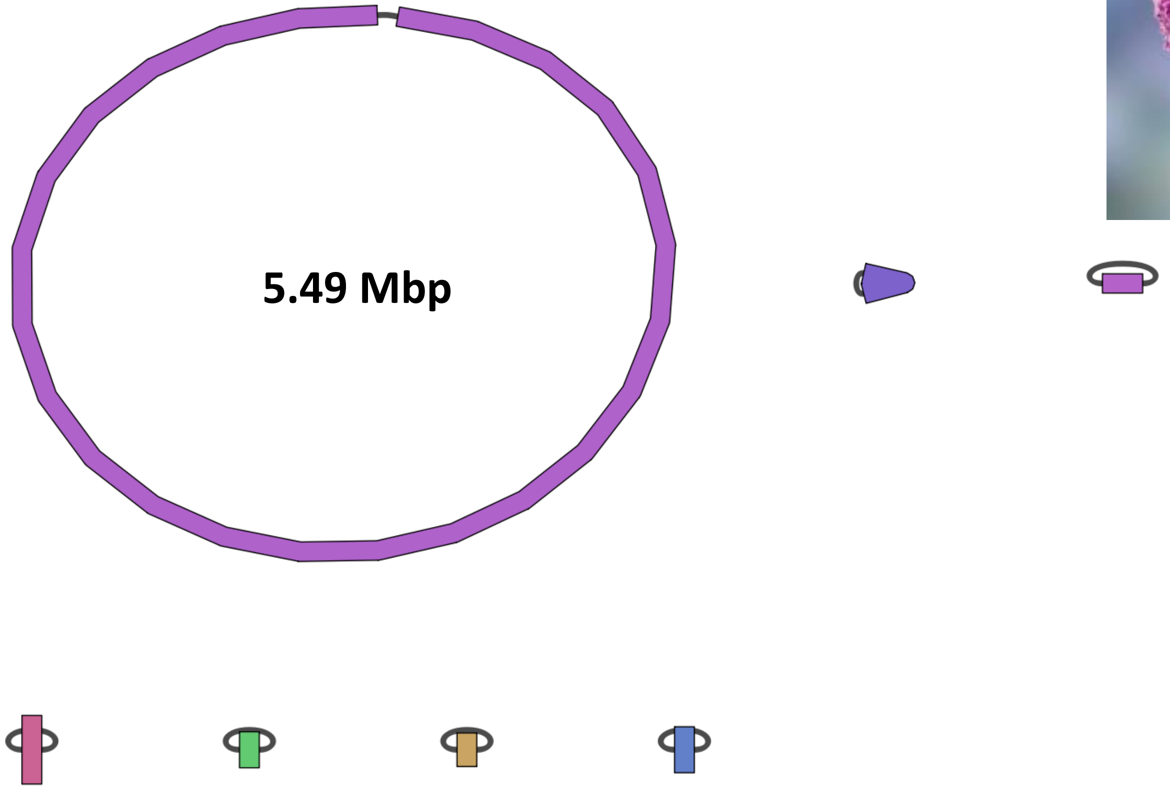
ATTGCCCGGAA CGGAATGTGAT



We would like to achieve high contiguity

Genome size is a limiting factor

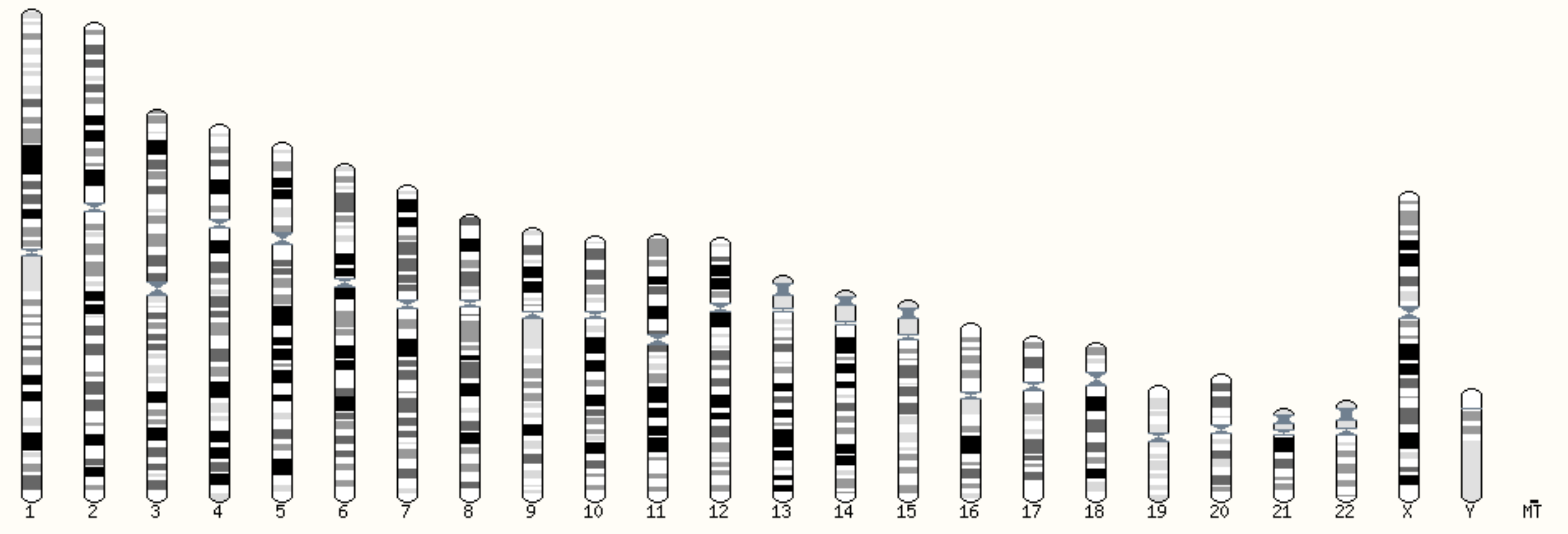
Klebsiella pneumoniae



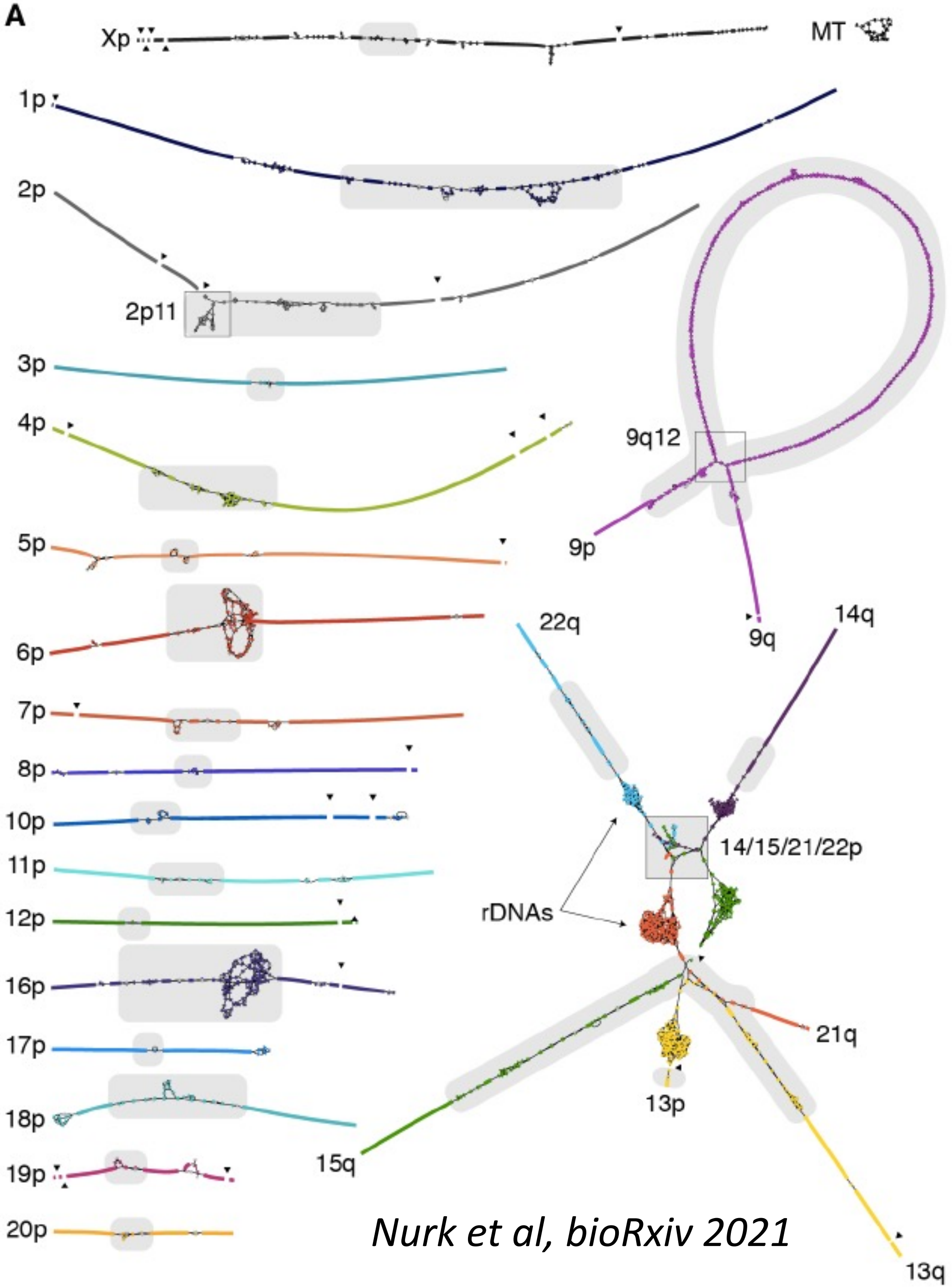
≥30x ONT and ≥60x Illumina
(Unicycler v0.4.6)

Human genome

2n= 46



22 Autosomes + 2 sex chromosomes → Ideally. 24 scaffolds/contigs !



Nurk et al, bioRxiv 2021

TELOMERE-TO-TELOMERE CONSORTIUM

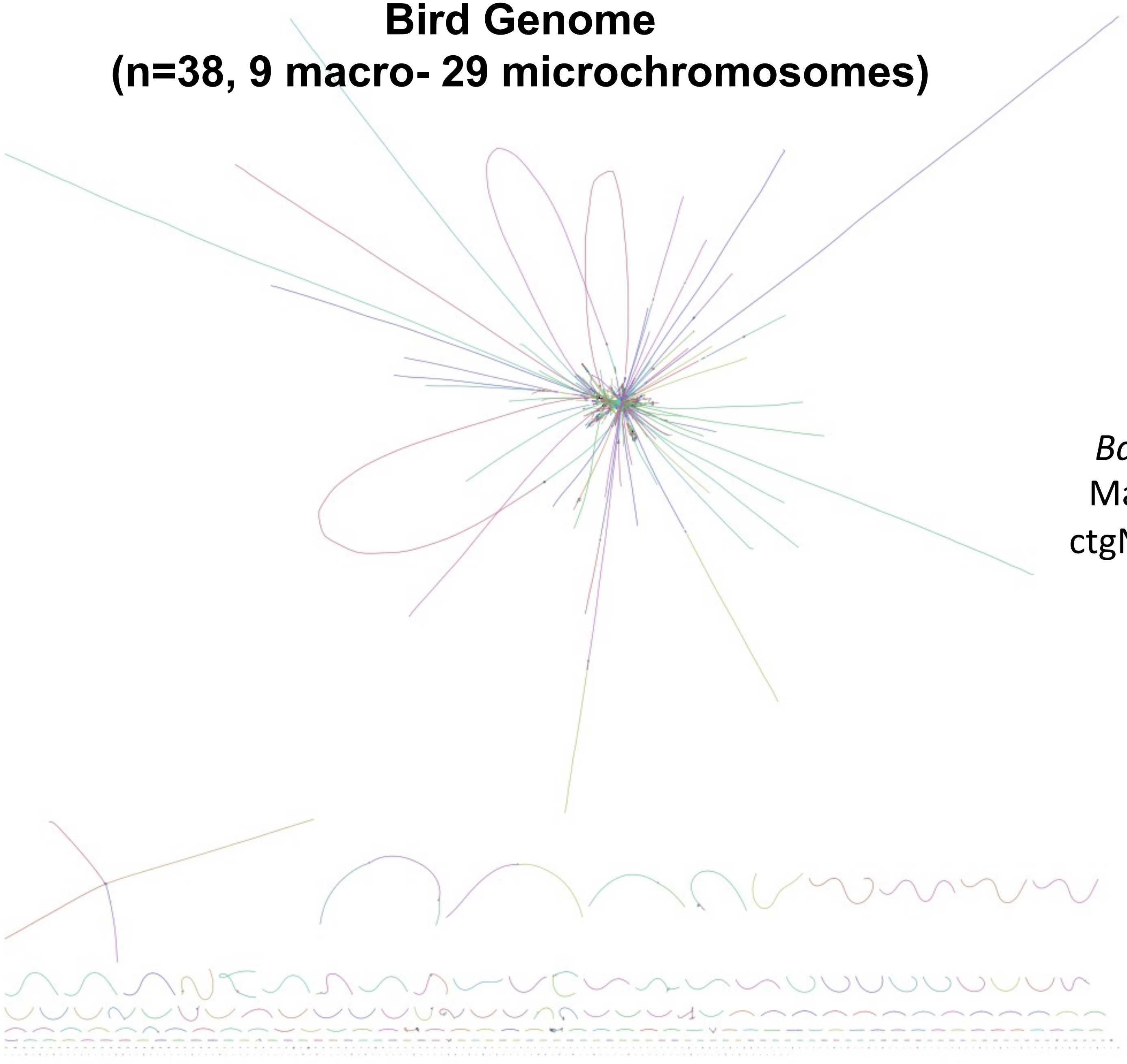
https://sites.google.com/ucsc.edu/t2tworkinggroup/home?authuser=0

- Haploid sample (CHM13 *Hiatidiform*)
- **Terlomere-To-Telomere(T2) assembly**
- Not perfect yet

Remaining Issues:

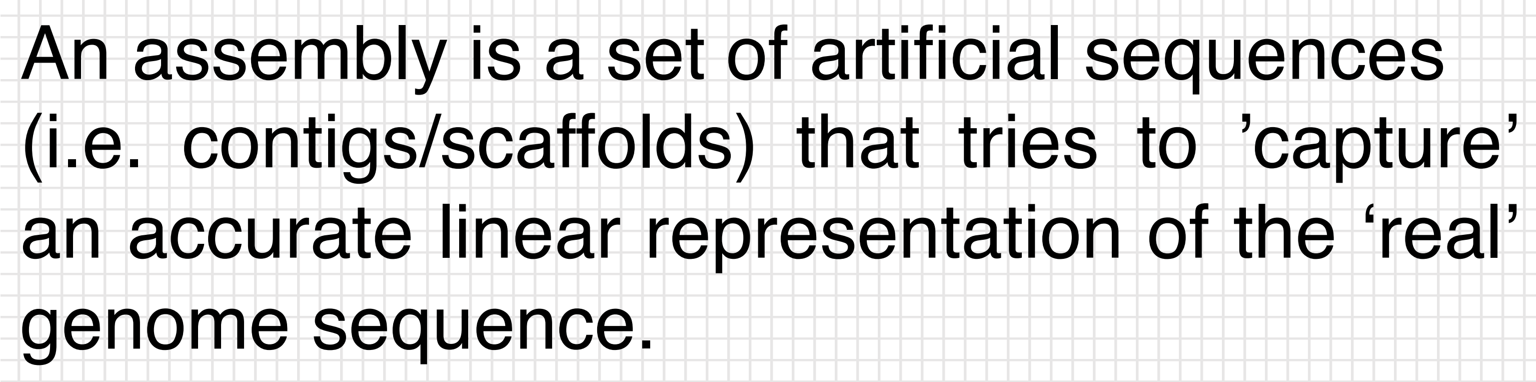
- ***Coverage gaps (GA-rich)***
- ***Centromeric Satellite repeats***
- ***rDNAs array***

Bird Genome
(n=38, 9 macro- 29 microchromosomes)



Base assembly
MaSuRCA+FLYE
ctgN50 = 6.13 Mb

How do we know our assembly is good enough?



An assembly is a set of artificial sequences (i.e. contigs/scaffolds) that tries to ’capture’ an accurate linear representation of the ‘real’ genome sequence.

Assembly Properties

The main properties to evaluate the quality of an assembly are:

- **Contiguity**
- **Gene completeness**
- **Sequence Accuracy**

How do we measure **contiguity**?

Contiguity metrics - **Nseries**

To measure an assembly contiguity we use *Nseries* metrics (Nx)

1. All sequences are **sorted by length**.

2. **Nx**: We determine **the length of the sequence at which the cumulative length is ≥ x%** of the total assembly length

3. **Lx**: We **count the number of sequences** at which the cumulative length is ≥x% (Lx)

Can be applied to contigs or scaffolds!!!

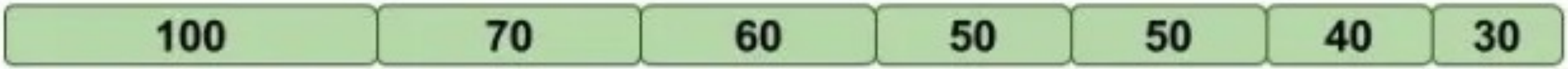
Contiguity metrics – **N50**

- contig ‘N50 length’, defined as the largest length *L* such that 50% of all nucleotides are contained in contigs of size at least *L*.

- scaffold ‘N50 length’, defined as the largest length *L* such that 50% of all nucleotides are contained in scaffolds of size at least *L*.

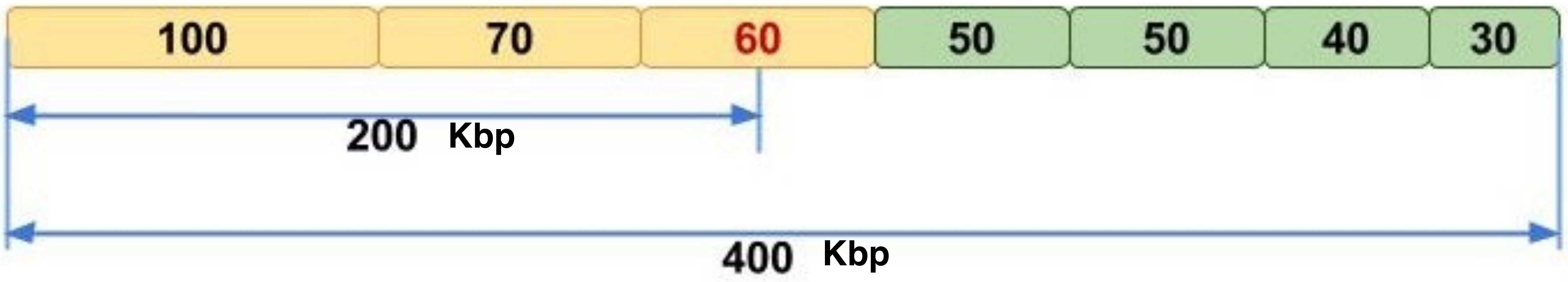
Lander et al. (2001) Initial sequencing and analysis of the human genome. Nature 409.6822: 860-921.

N50 and L50



All contigs are sorted by length

N50 and L50



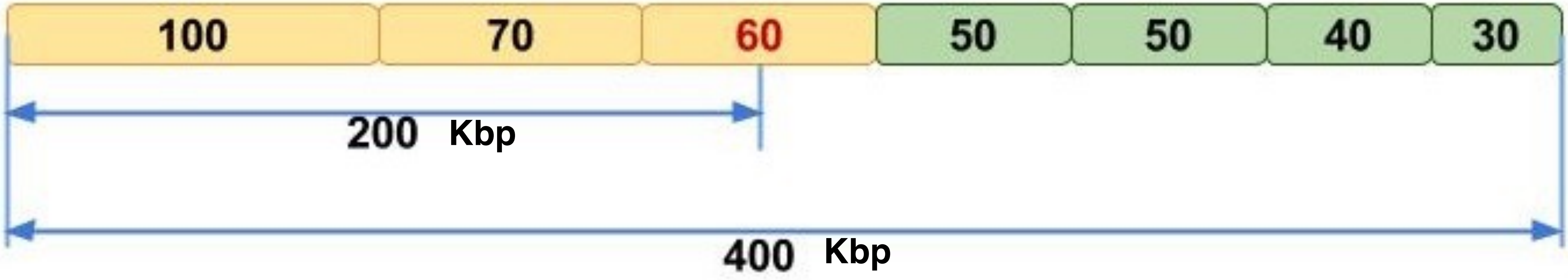
100+70=170 < 200

100+70+60 >= 200

N50 = 60 Kbp

L50 = 3

N100



N100 ?

L100 ?

How do we measure **gene completeness**?

Gene Completeness – **BUSCO**

OrthoDB

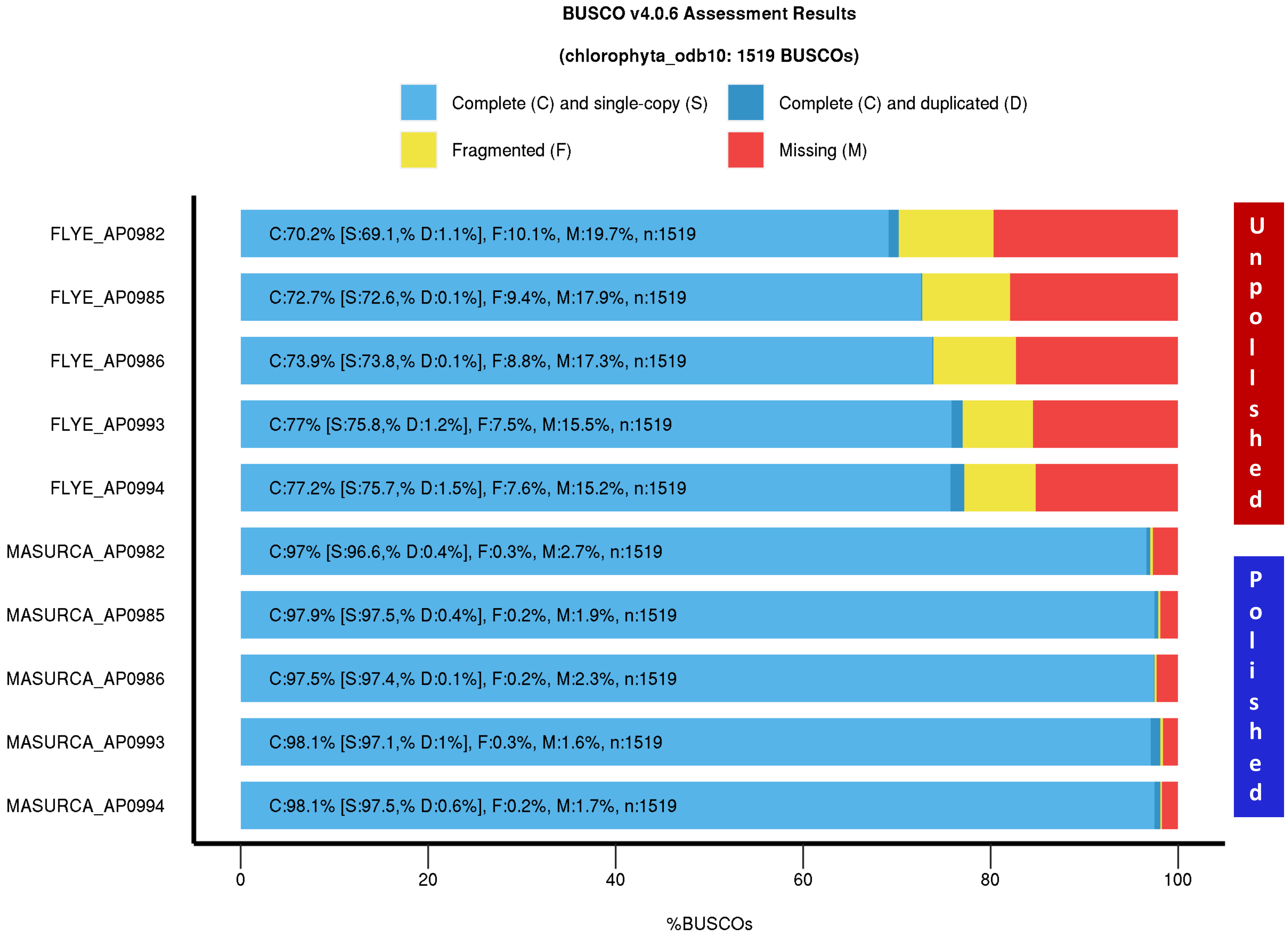
- It uses **orthodb,** a database containing **single copy orthologues** (buscos) on a clade.

- **Searches these genes** against our assembly.

- Reports how many are **Complete**, how many are **Fragmented** and how many are **Missing**

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31**.

Gene Completeness – **BUSCO**



6 Ostreococcus tauri strains assembled twice

Gene Completeness – **BUSCO**

What are the reasons for **missingness?**

- **Not assembled**
- **Not close-enough database**
- **Not enough sequence quality in assembly**

How do we measure Sequence Accuracy?

Sequence Accuracy– **Consensus Quality (QV)**

- The QV score is expressed logarithmically, and represents the log-scale probability of errors for the consensus basecalls

QV= 30	1 error in 1,000 bp
QV= 40	1 error in 10,000 bp

Rhie A, Walenz BP, Koren S, Phillippy AM: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biology* 2020, **21:**245.

Current goal is to meet EBP standards: 6CQ40

- Main criteria (6CQ40)

- **>1 Mbp contig N50**
- Chromosome-scale scaffolds
- **Error rate <1/10,000bp = QV40**

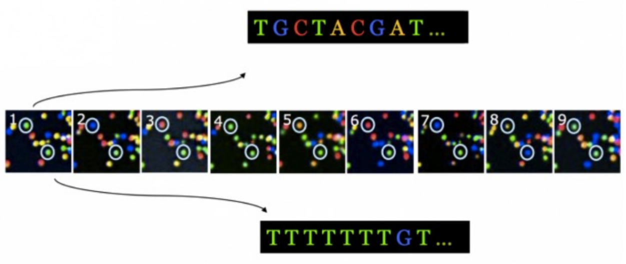
EARTH
BIOGENOME
PROJECT

- Additional requirements

- **>90% single copy complete BUSCOs**
- <5% false duplications
- >90% kmer completeness
- >90% sequence assigned to chr
- >90% transcripts from same organism mappable
- Separate symbionts, organellar genomes, haplotypic alternate seqs

Assembling Short Reads

Illumina



- Sequencing by synthesis
  - reversible terminators
- Ultra-high throughput
  - 100s of millions to billions of reads per run (high coverage)
- Short reads
  - 100-250bp
- Good quality
  - ~1% error (0.1 % after trimming)

What are K-mers?

A **K-mer** is a substring of length K in a string T of DNA with L bases.

AATTGGCCG L=9

2-mers

AATTGGCCG
AA
AT
TT
TG
GG
GC
CC
CG

Total 2-mers: 8

3-mers

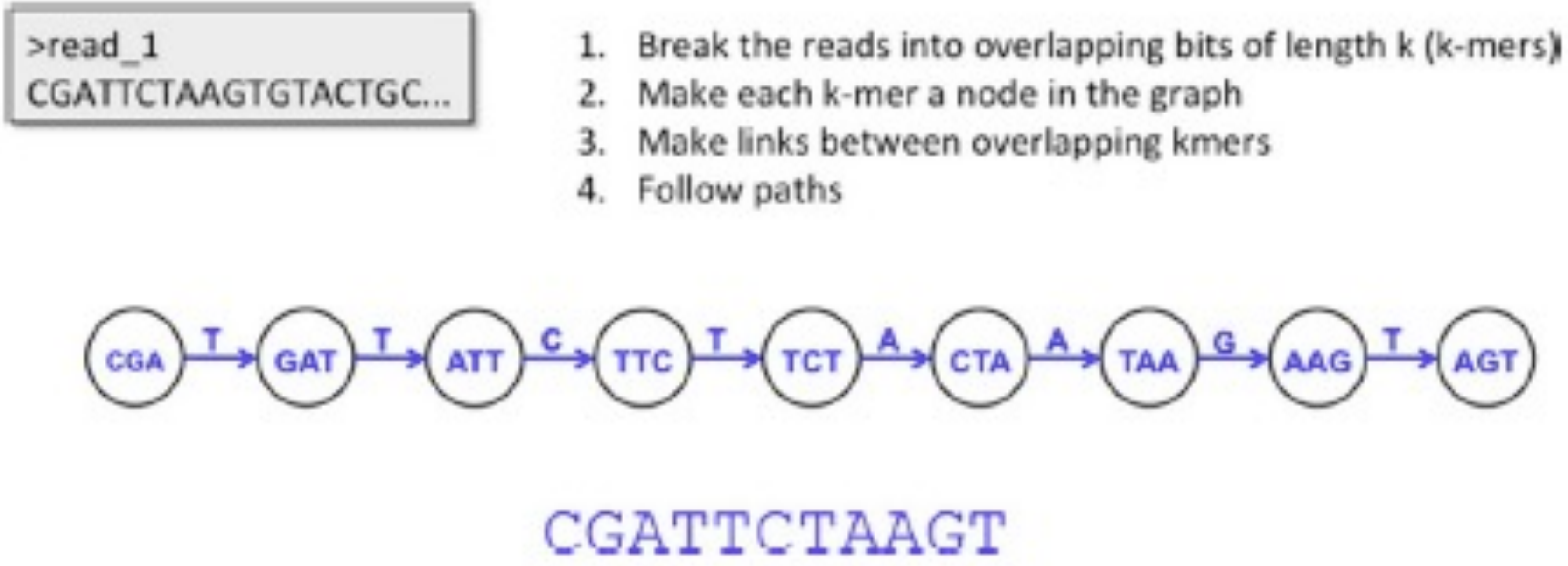
AATTGGCCG
AAT
ATT
TTG
TGG
GGC
GCC
CCG

Total 3-mers: 7

Total k-mers (n)= L – K + 1

All **K-mers** from substring T will **overlap K-1 bases !**

Reads are broken into K-mers



```
>read_1
CGATTCTAAGTGTACTGC...
```

1. Break the reads into overlapping bits of length k (k-mers)
2. Make each k-mer a node in the graph
3. Make links between overlapping kmers
4. Follow paths

CGATTCTAAGT

Anything unusual on the edges?

Leggett RM, MacLean D: **Reference-free SNP detection: dealing with the data deluge.** *BMC Genomics* 2014, **15:**S10

Why eads are broken into K-mers?

- Trap sequencing errors in smaller substrings

- Compute a higher number of overlaps across the genome

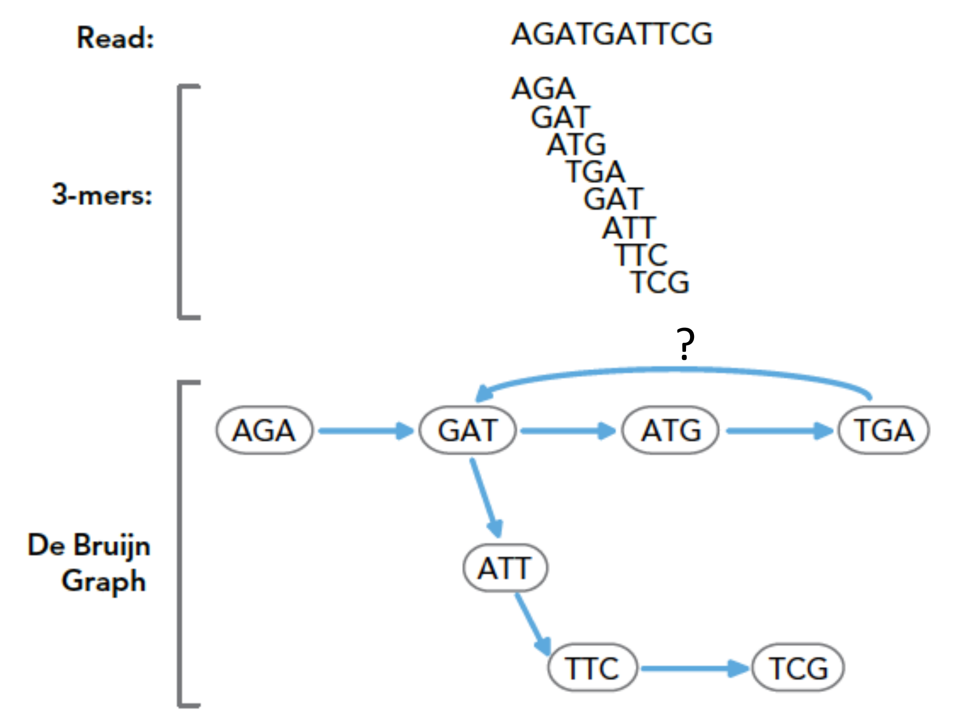
- Overcome coverage ‘holes’

Total Kmer coverage = ((L – K +1)/L) * Read Coverage

De Bruijn Graphs

- More efficient (memory and time) for billions of short reads
- Decompose reads into *k*-mers
- Construct graph where nodes are *k*-mers and edges are *k-1* overlaps

Figure 3: De Bruijn Graph for Read with K=3



The length of overlaps is k-1=2. Gray arrows indicate where all the k-mers derived from the one read are placed in the graph. Blue arrows indicate the order of the k-mers and their overlaps.

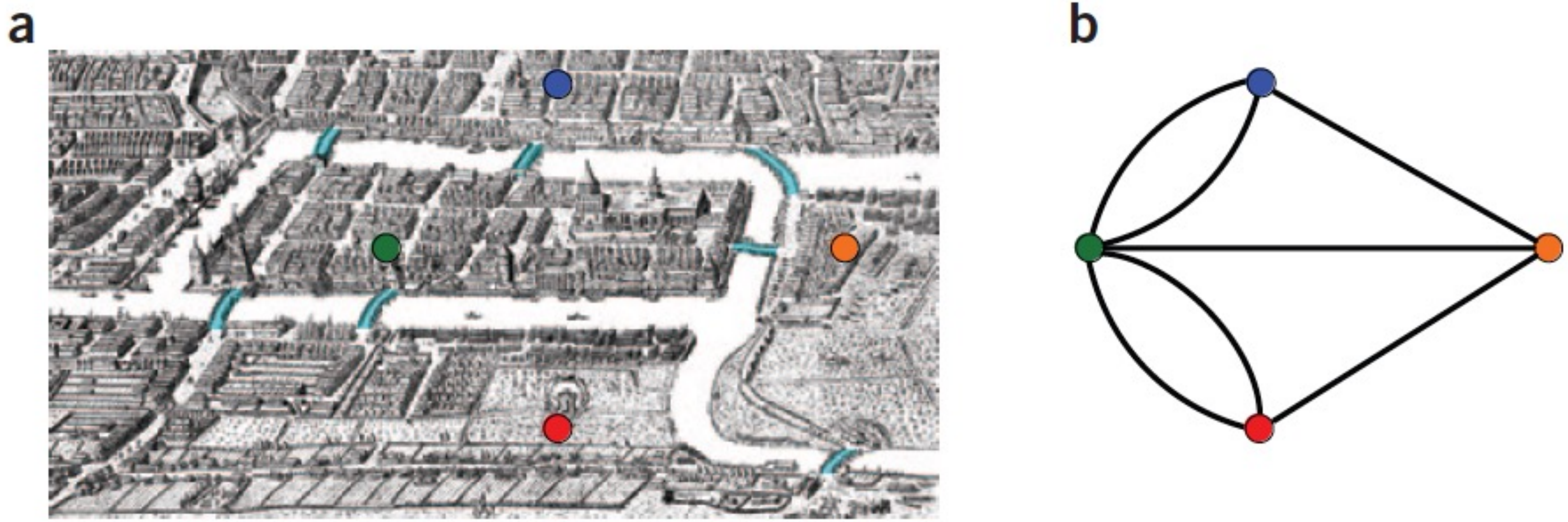
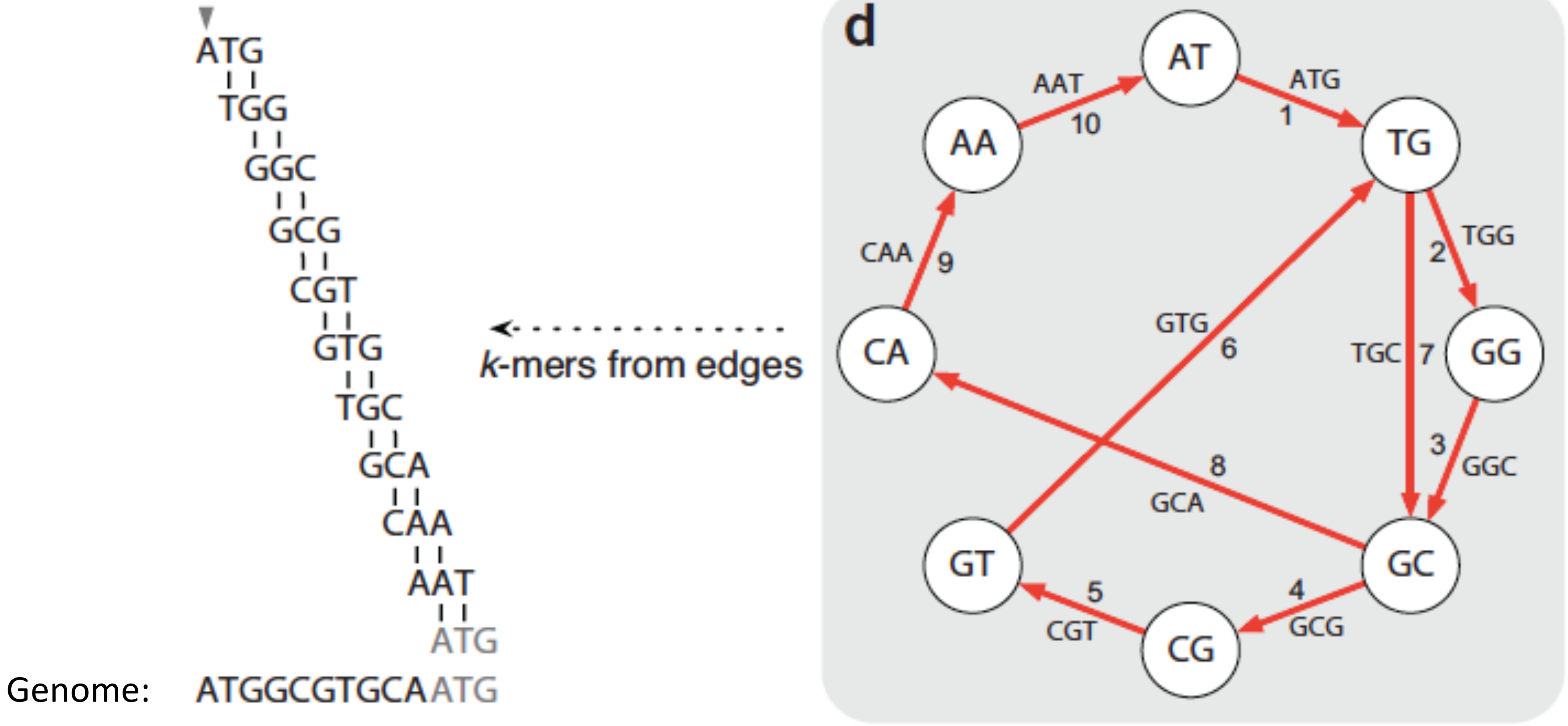


Figure 1 Bridges of Königsberg problem. (**a**) A map of old Königsberg, in which each area of the city is labeled with a different color point. (**b**) The Königsberg Bridge graph, formed by representing each of four land areas as a node and each of the city's seven bridges as an edge.

Compeau PE, Pevzner PA, Tesler G: **How to apply de Bruijn graphs to genome assembly.** *Nat Biotechnol* 2011, **29:**987-991.

De Bruijn Graph

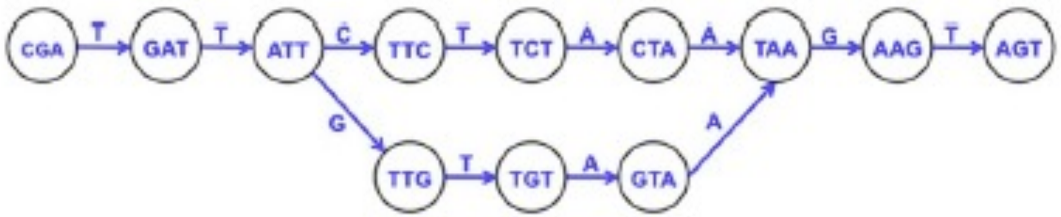


Eulerian cycle
Visit each edge once
(easier to solve)

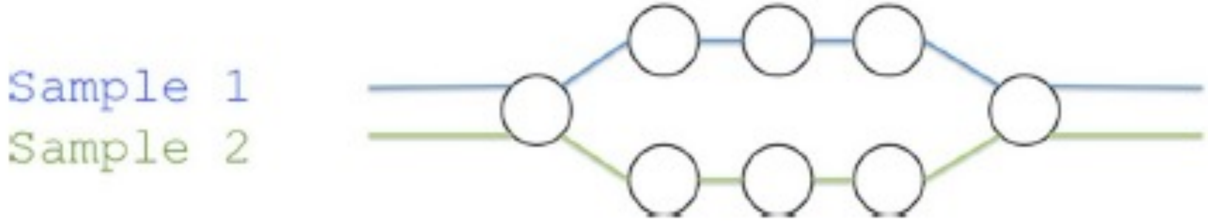
Compeau PE, Pevzner PA, Tesler G: **How to apply de Bruijn graphs to genome assembly.** *Nat Biotechnol* 2011, **29:**987-991.

SNPs create 'Bubbles' in the graph

Unlike errors these branches have similar K-mer coverage



CGATTCTAAGT
CGATTGTAAGT



Leggett RM, MacLean D: **Reference-free SNP detection: dealing with the data deluge.** *BMC Genomics* 2014, **15:**S10

Effect of K-mer Length

K-mer overlap: K determines the length of the overlap between k-mers (K-1)

AATTGGCCG L=9

2-mers

AATTGGCCG
A**A**
AT
TT
TG
GG
GC
CC
CG

Total 2-mers: 8

K-overlap= 1

3-mers

AATTGGCCG
AAT
ATT
TTG
TGG
GGC
GCC
CCG

Total 3-mers: 7

K-overlap= 2

Total k-mers (n)= L – K + 1

Effect of K-mer Length

- **K-mer overlap – increase with** K (+)

- **K-mer coverage – drops with K (-)**

- **Likelihood of error -** increase with K (-)

Effect of K-mer Length

We need to find a balance !!!!!!!!

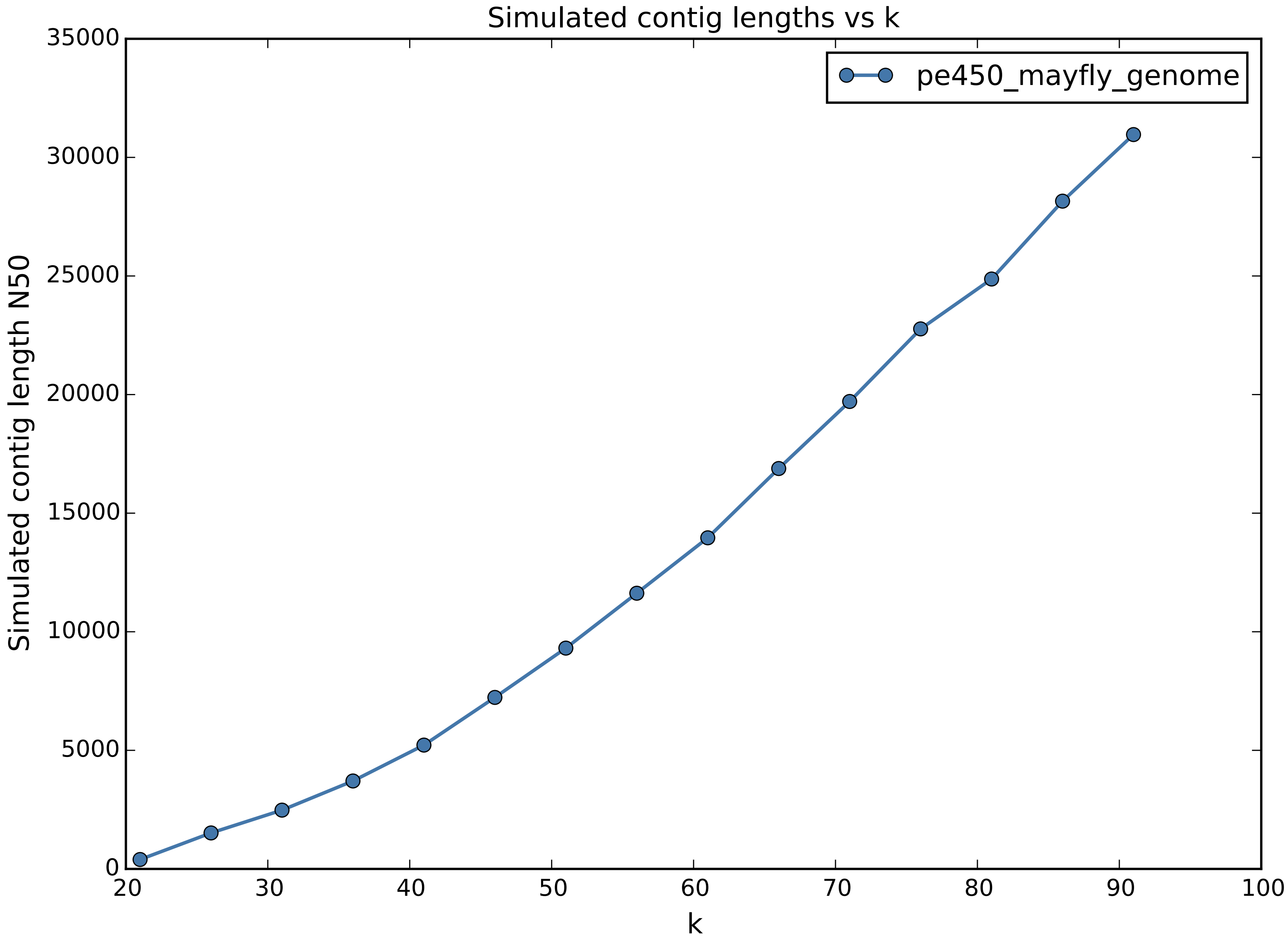
K

Long enough for reliable overlaps,

Short enough to avoid errors

and represent most of the genome (coverage)

Optimal K-mer Length



Optimal K-mer Length

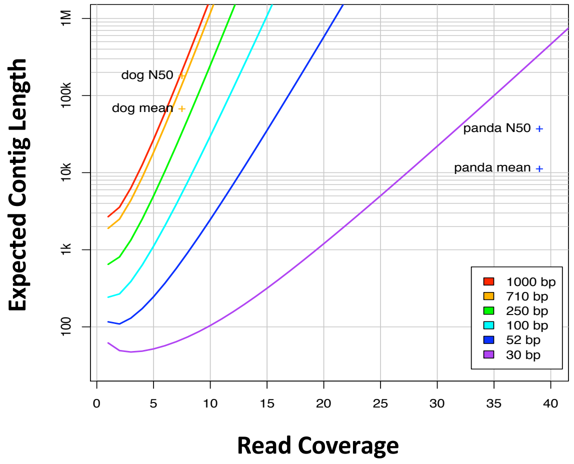
It will depend on:

- **Read Length**
- **Error Rate (reads)**
- **Sequencing Coverage (reads)**
- **Repeats and Heterozygosity Rates (genome)**

Key factors for a good assembly

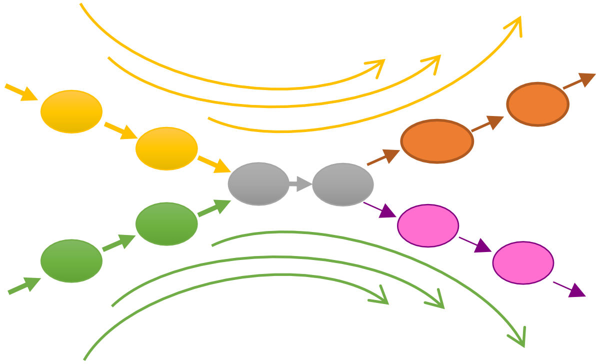
Key factors for a good assembly

Coverage



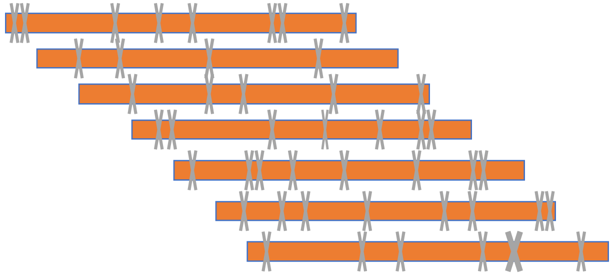
High coverage is required

Read Length



Reads & mates must be longer than the repeats

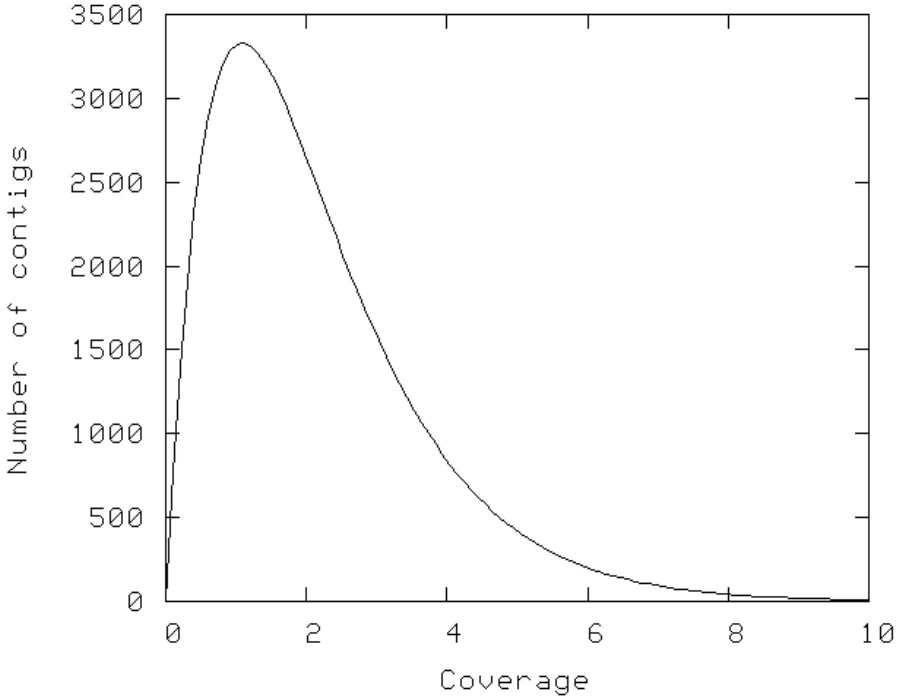
Quality

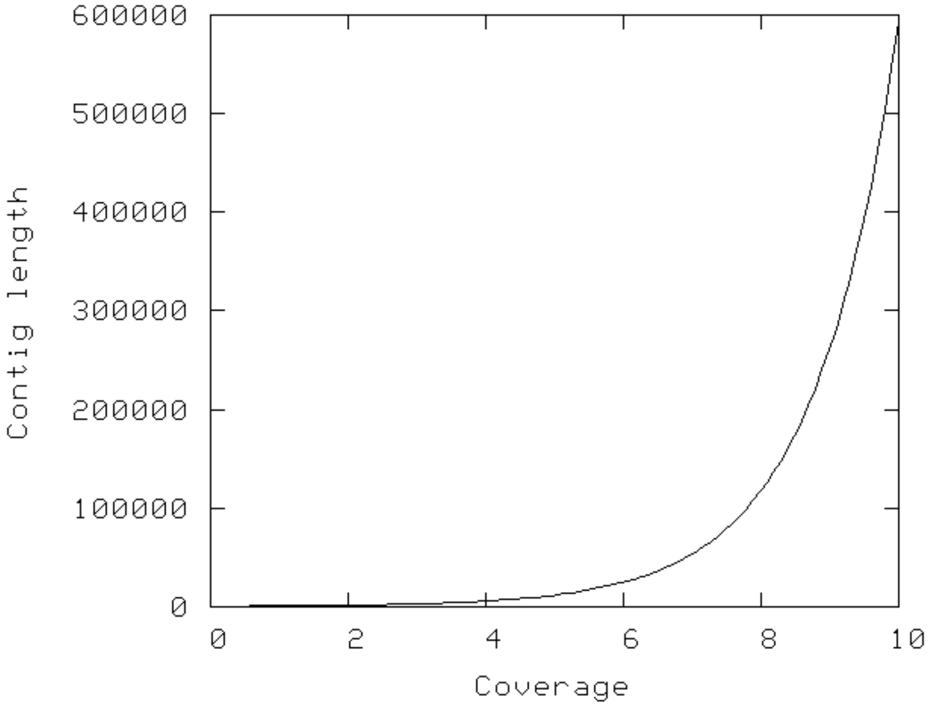


Errors obscure overlaps

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Coverage





Lander-Waterman statistics

E(#islands) = N$e^{-c\sigma}$
E(island size) = L(($e^{c\sigma}$ – 1) / c + 1 – σ)
contig = island with 2 or more reads

L = read length

T = minimum detectable overlap

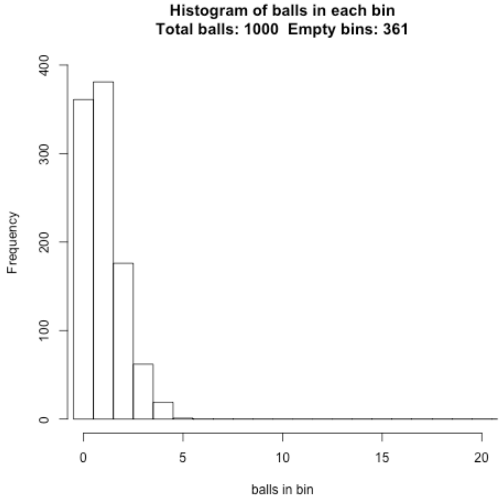
G = genome size

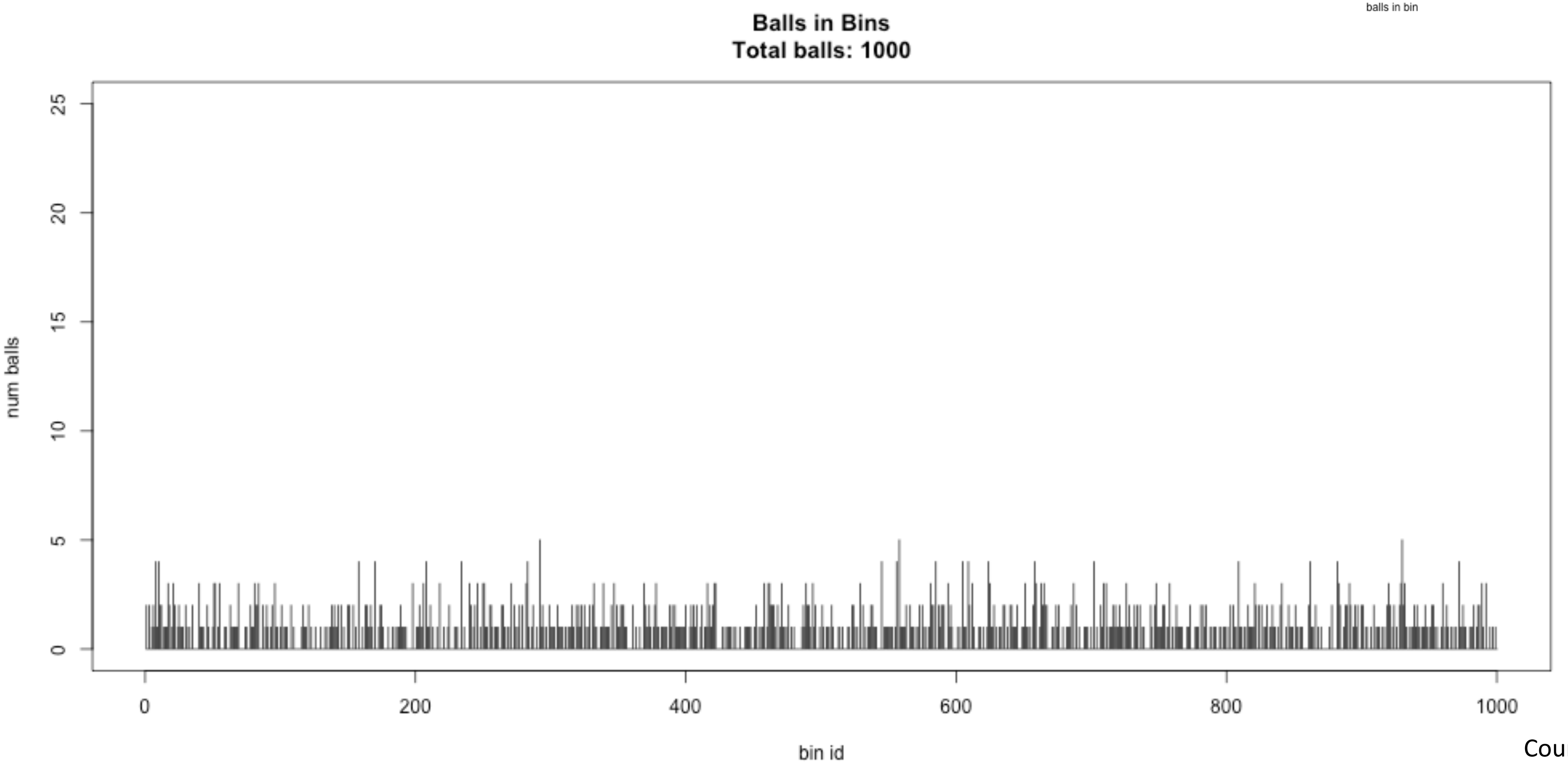
N = number of reads

c = coverage (NL / G)

σ = 1 – T/L

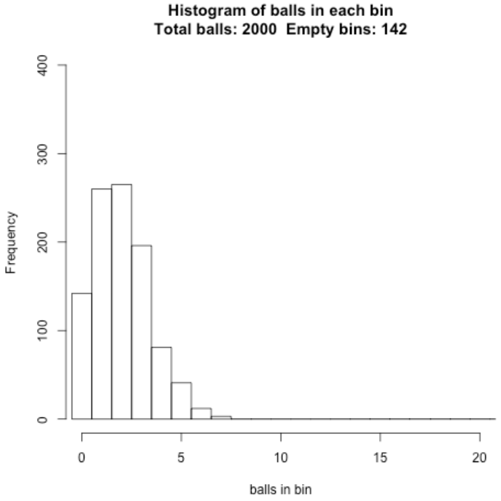
Balls in Bins 1x

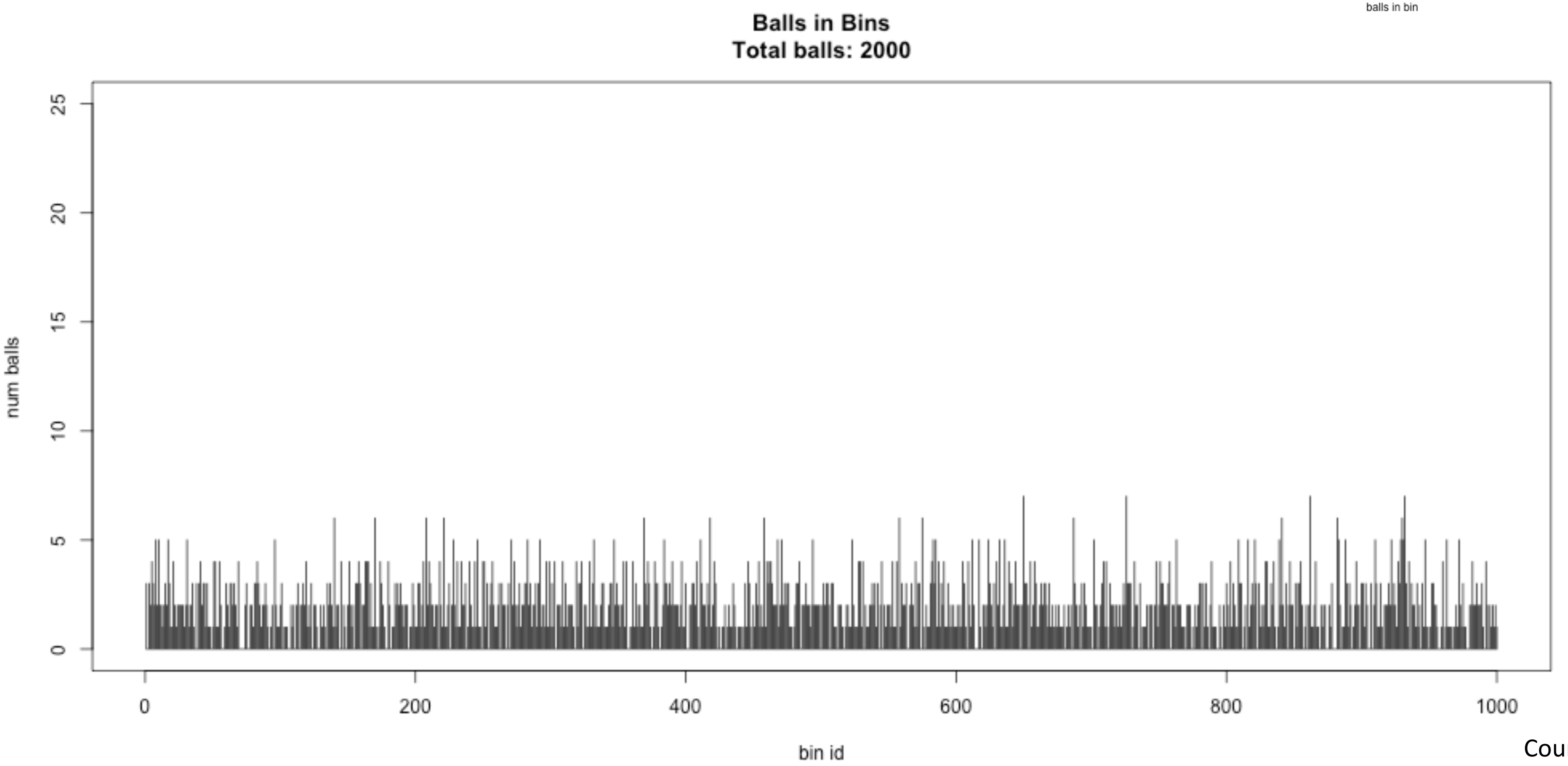




Courtesy of T. Alioto

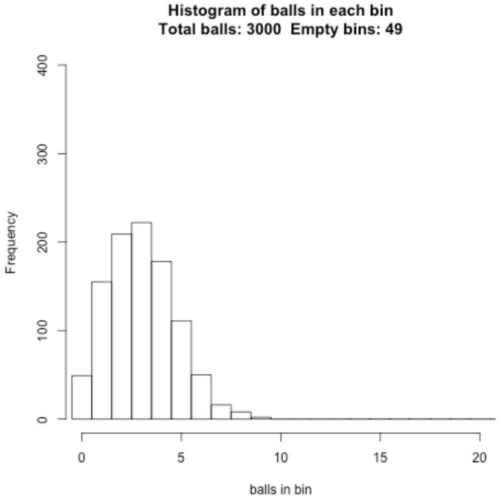
Balls in Bins 2x

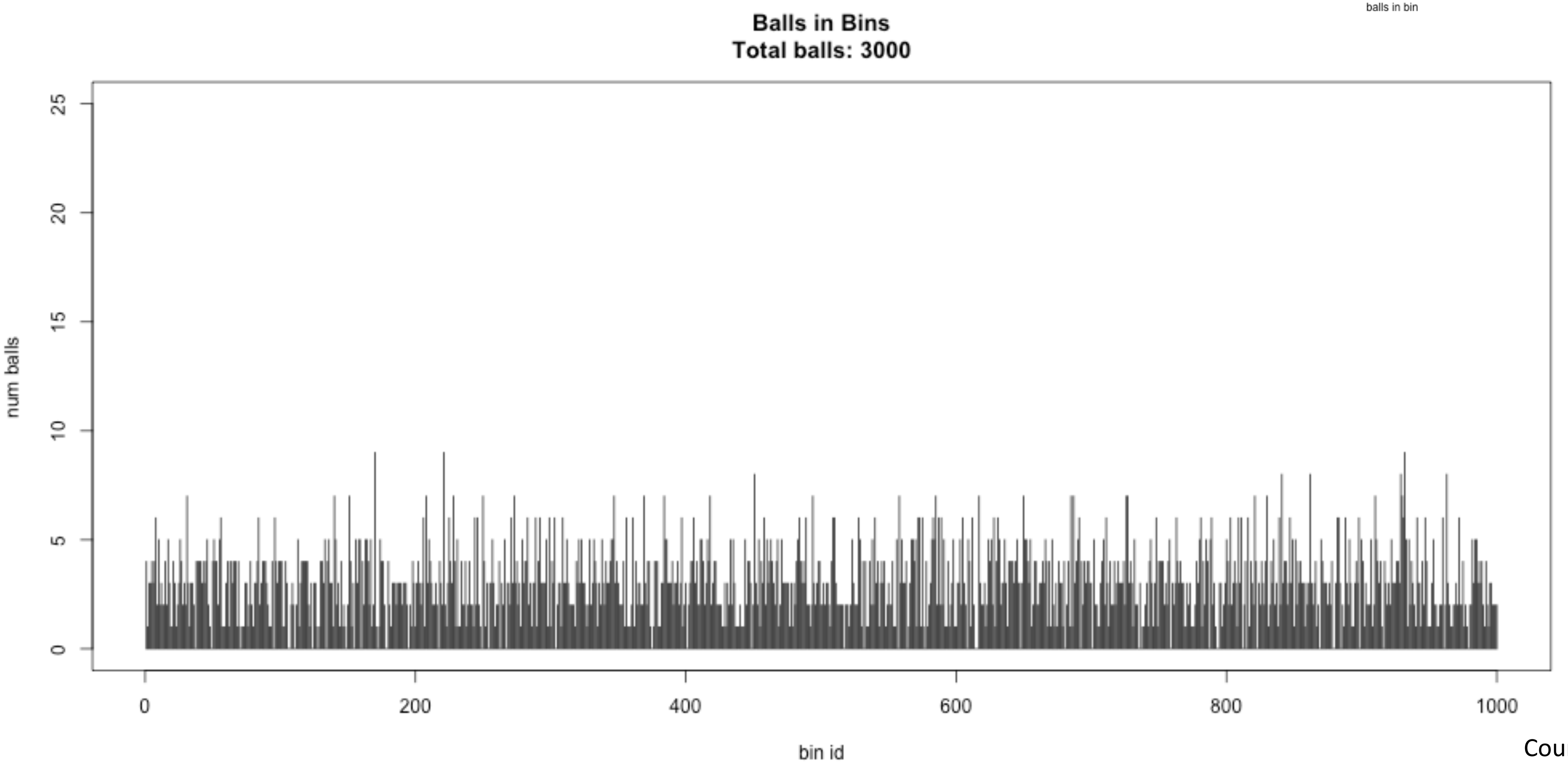




Courtesy of T. Alioto

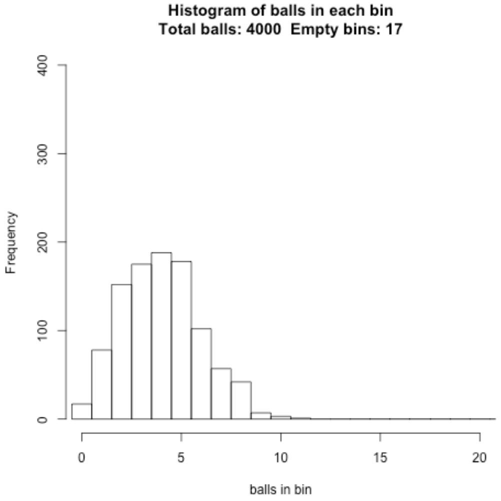
Balls in Bins 3x

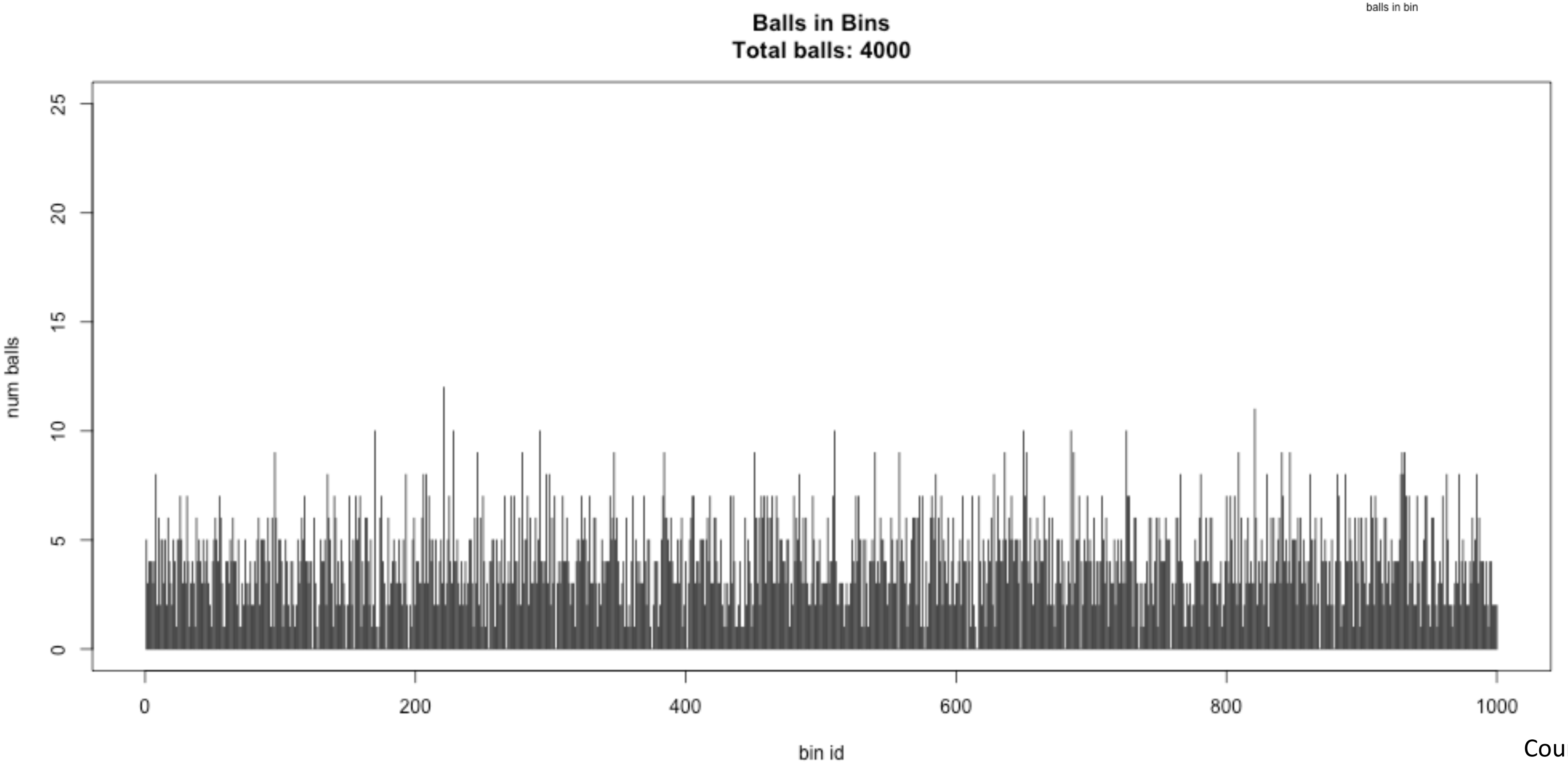




Courtesy of T. Alioto

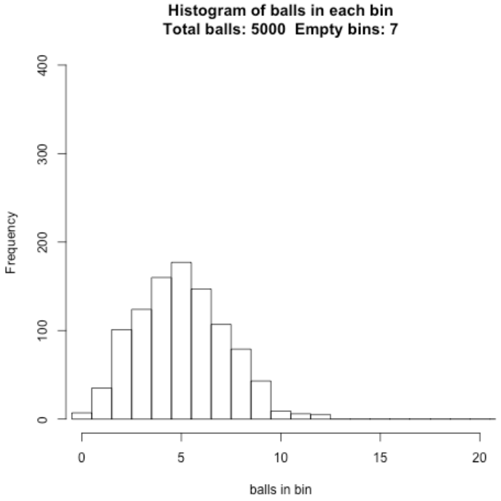
Balls in Bins 4x

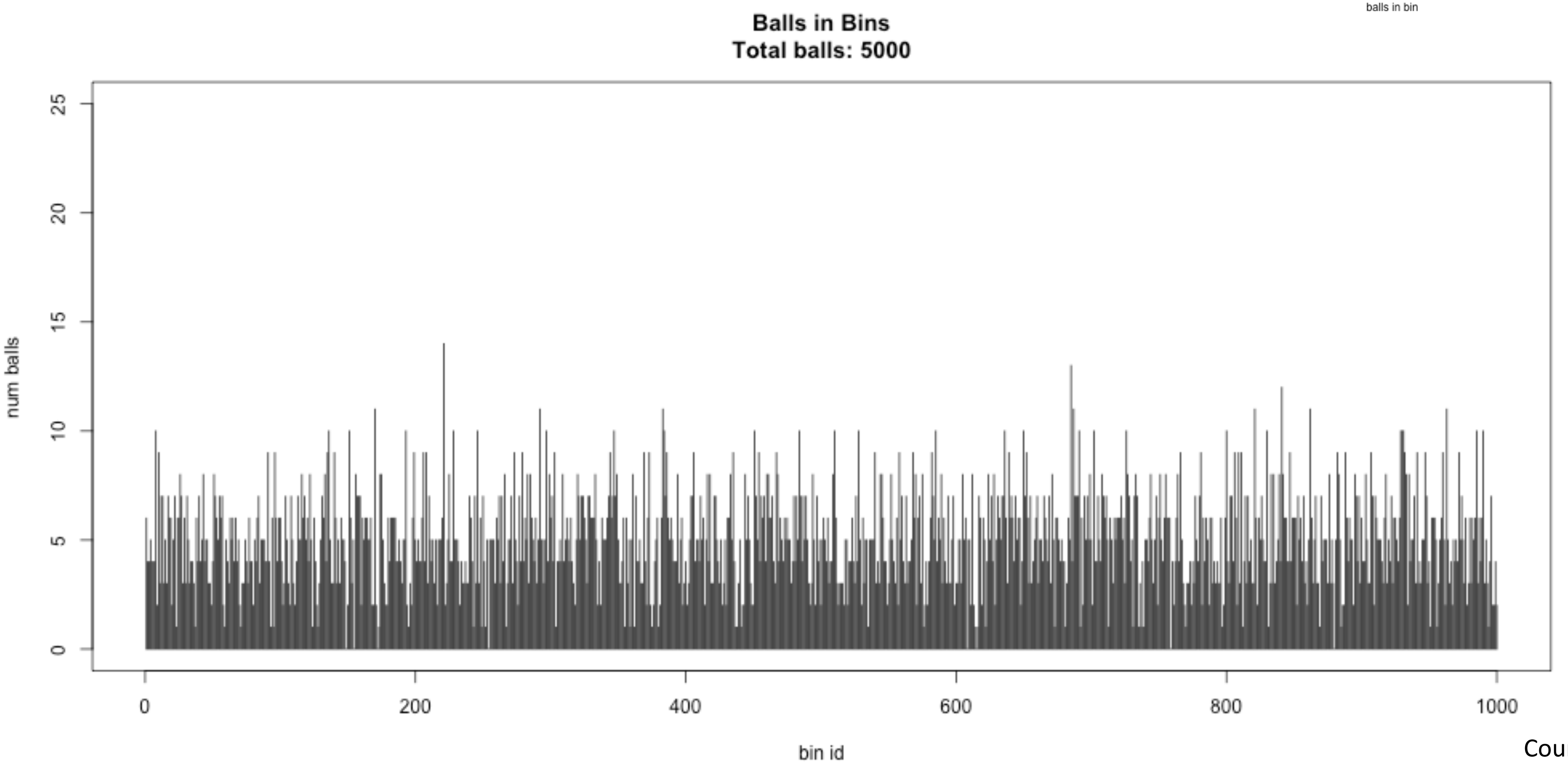




Courtesy of T. Alioto

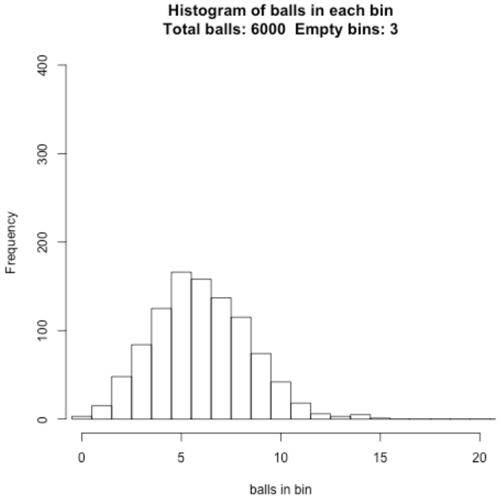
Balls in Bins 5x

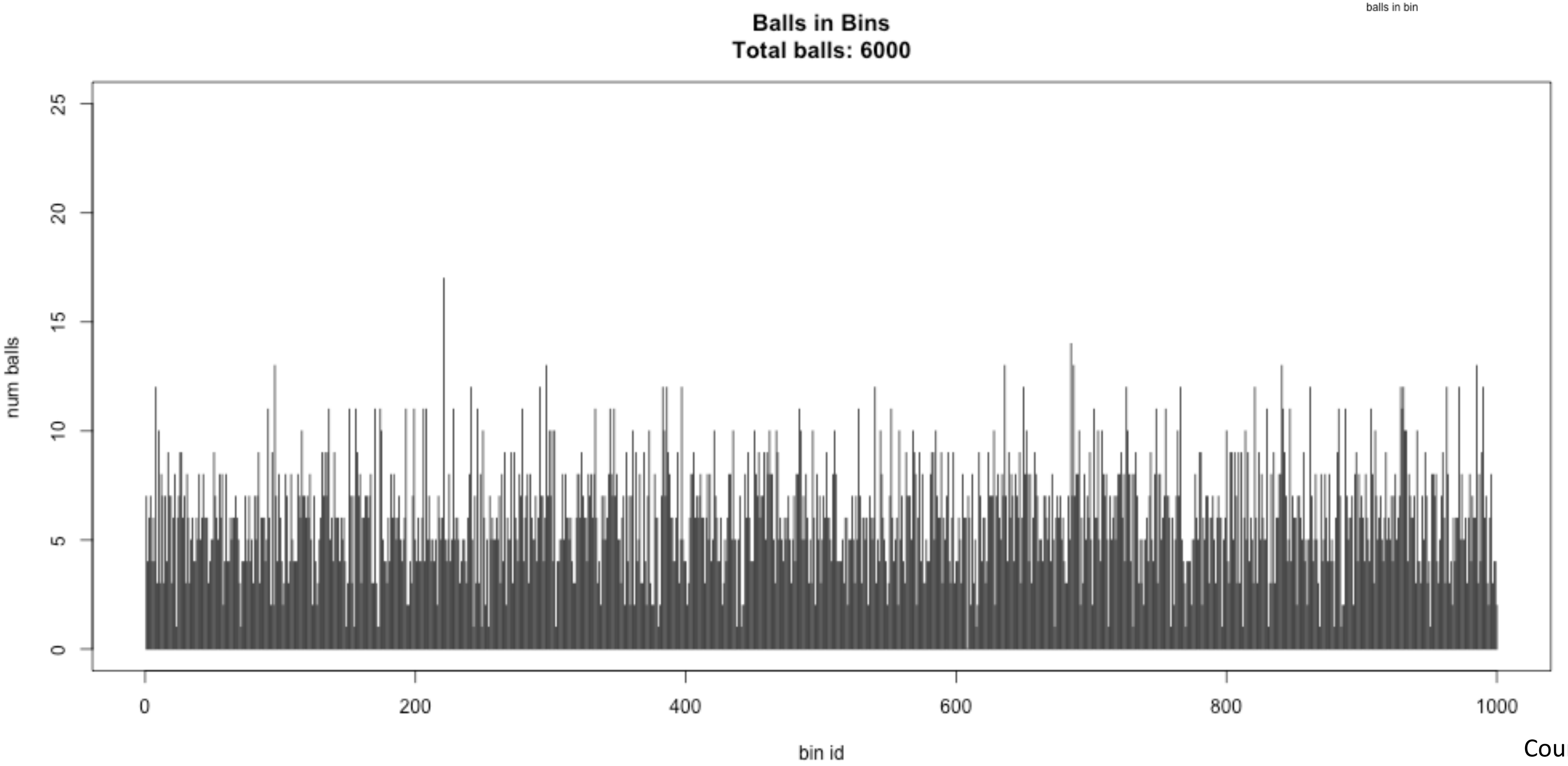




Courtesy of T. Alioto

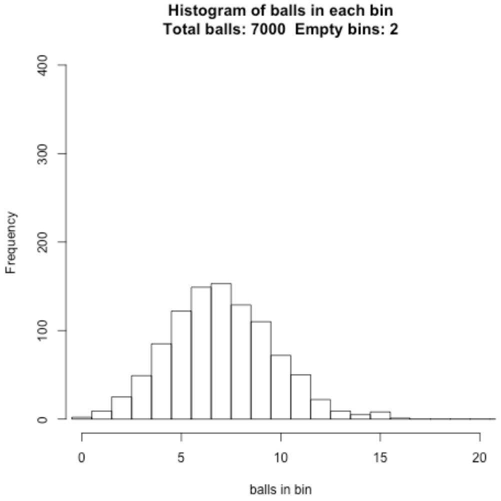
Balls in Bins 6x

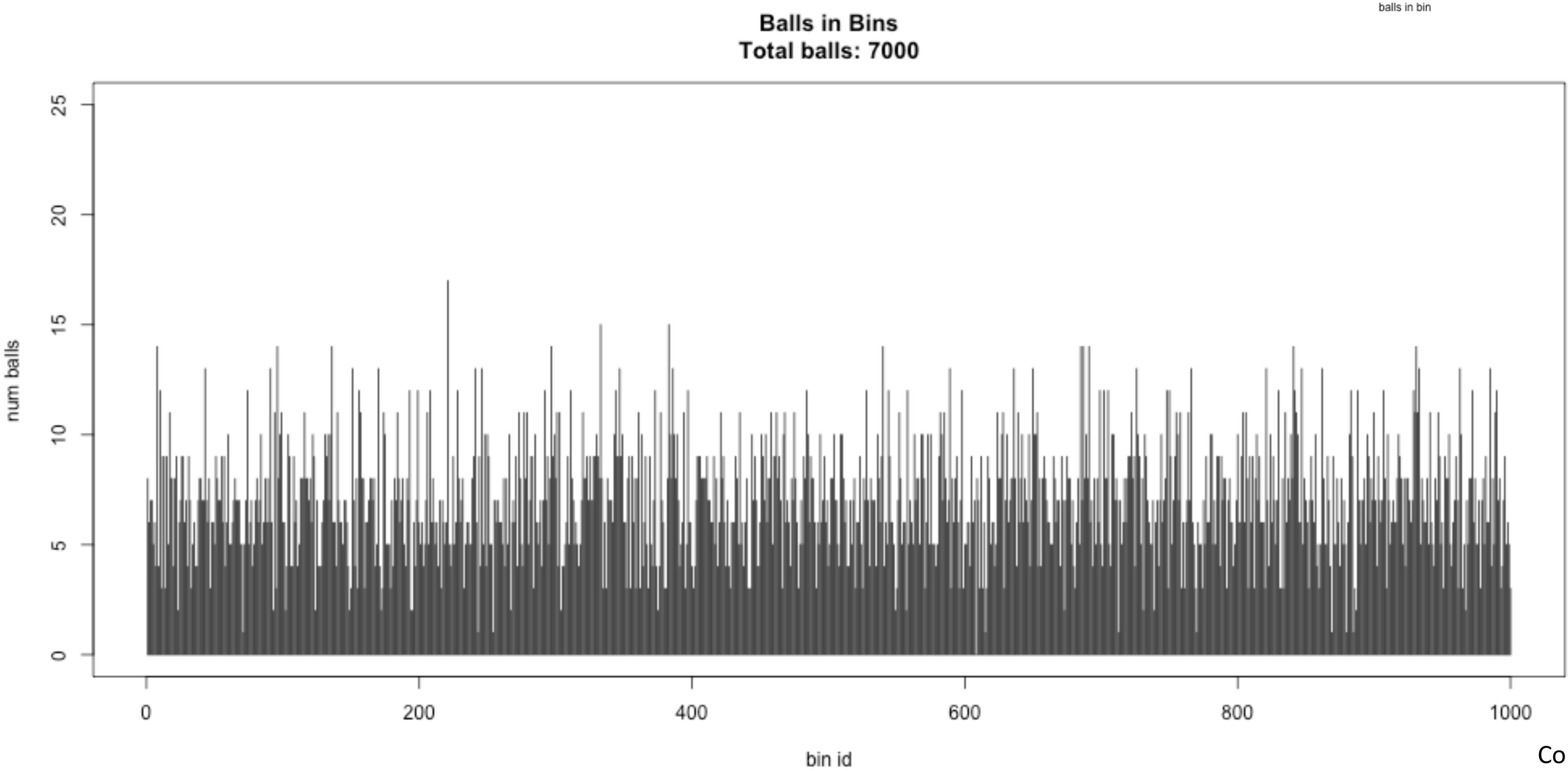




Courtesy of T. Alioto

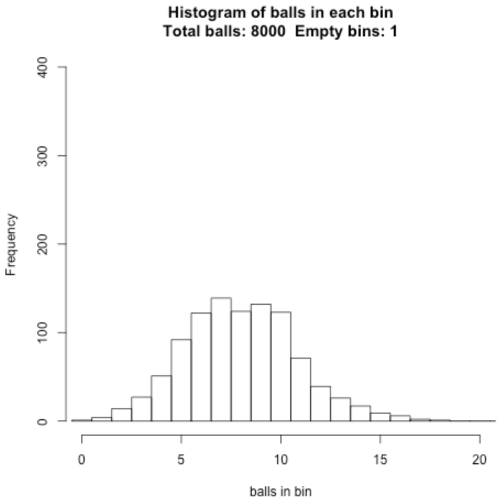
Balls in Bins 7x

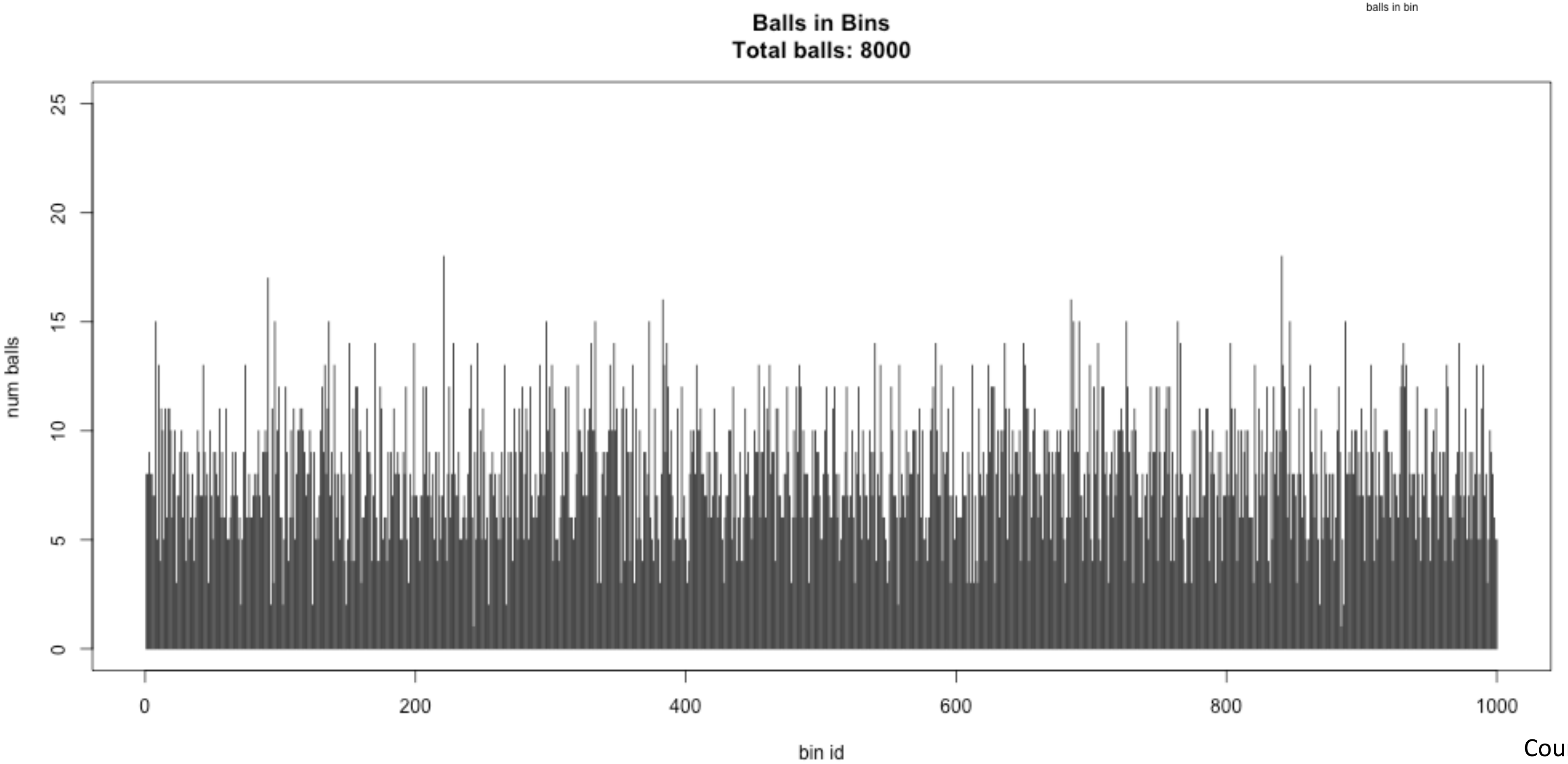




Courtesy of T. Alioto

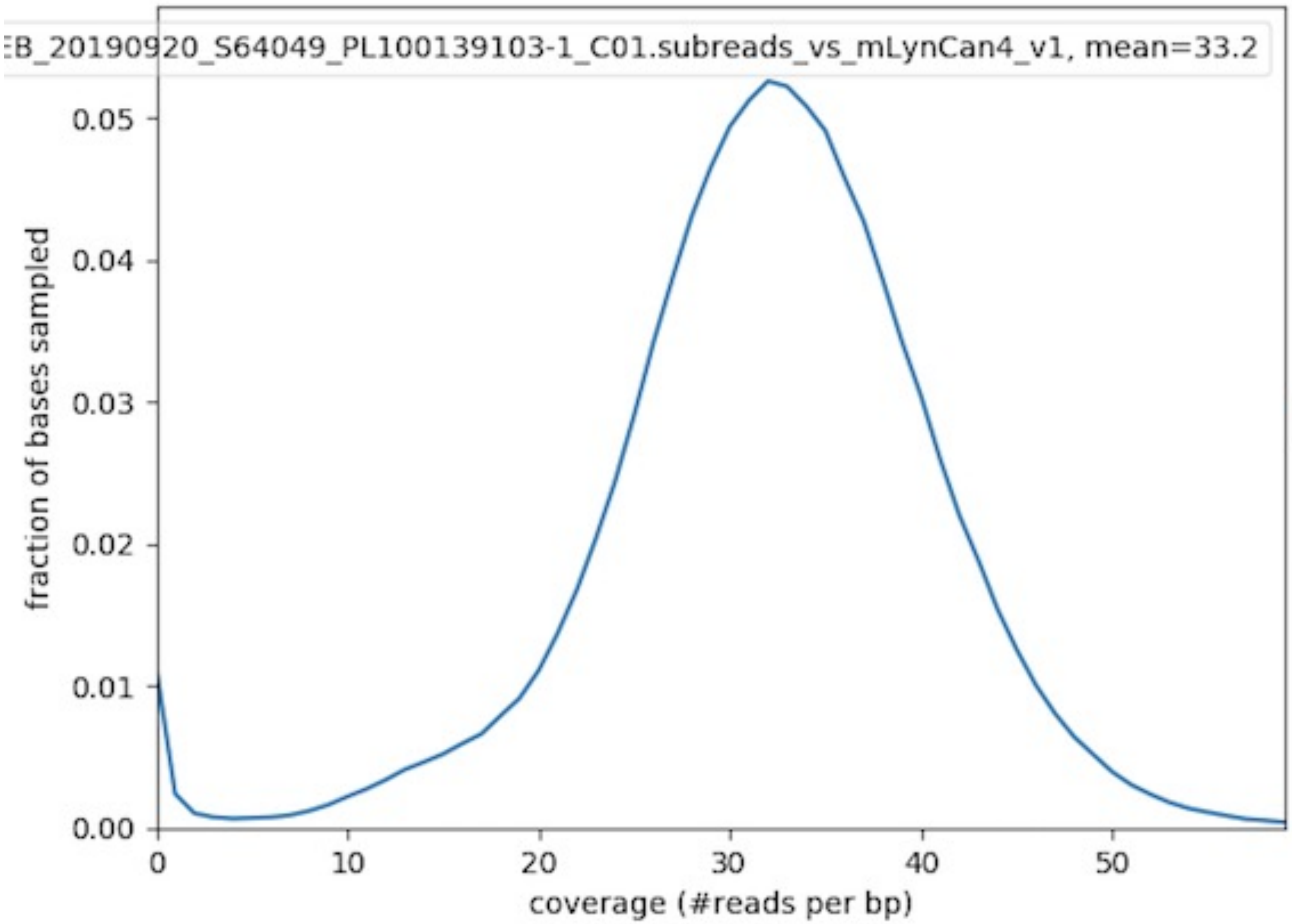
Balls in Bins 8x





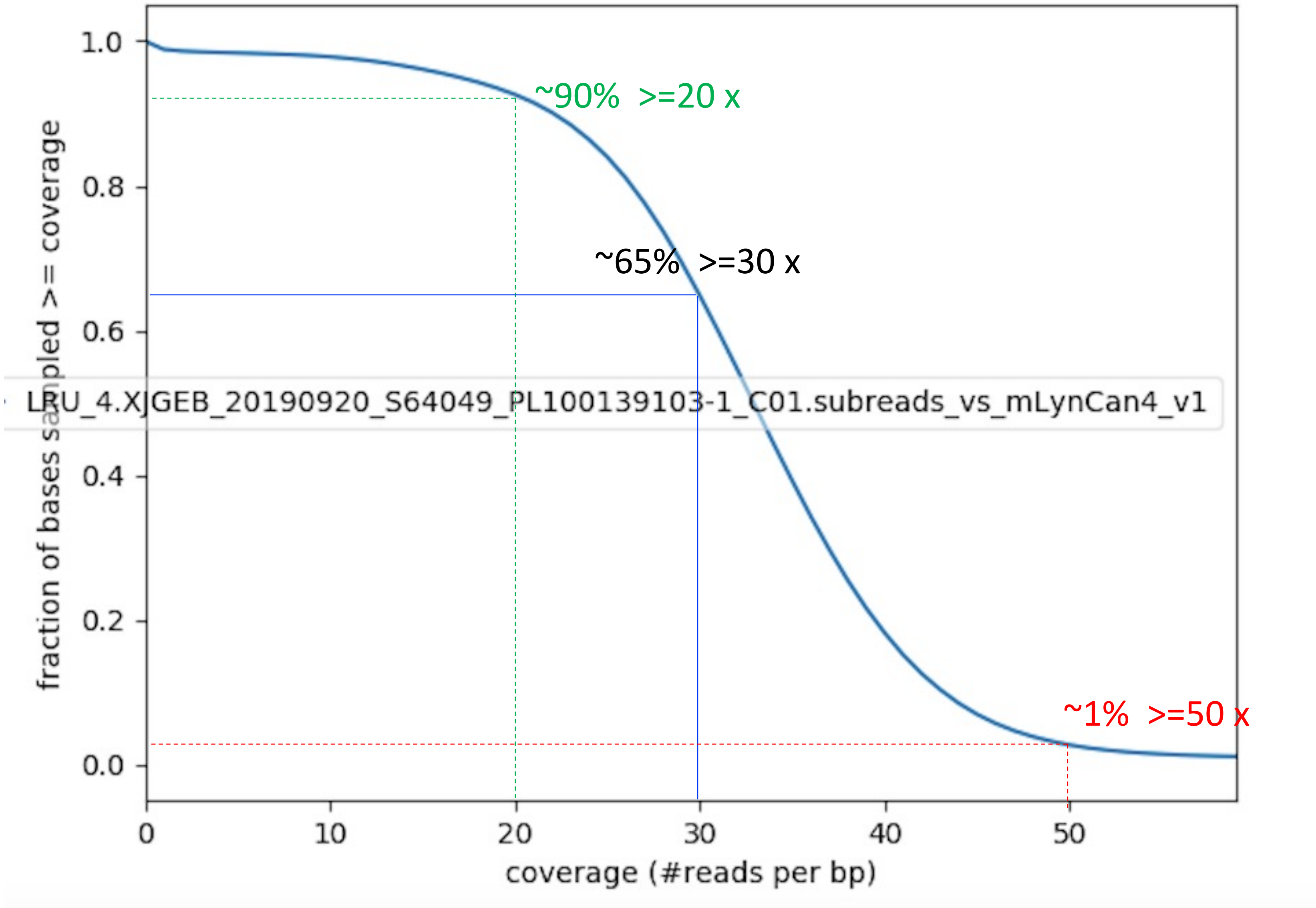
Courtesy of T. Alioto

Coverage Distribution: Normal



Normal genome coverage with mean 33.2x

Coverage Distribution: Even



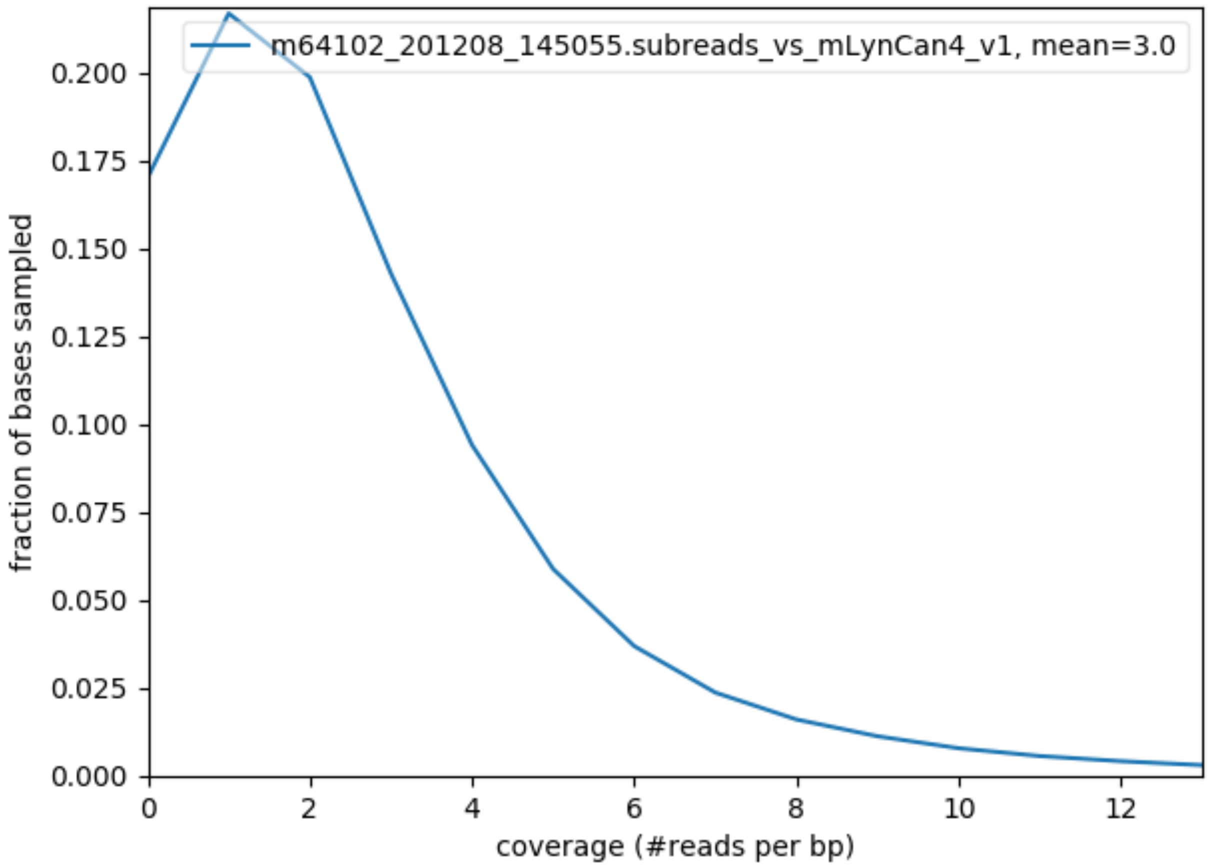
~90% >=20 x

~65% >=30 x

~1% >=50 x

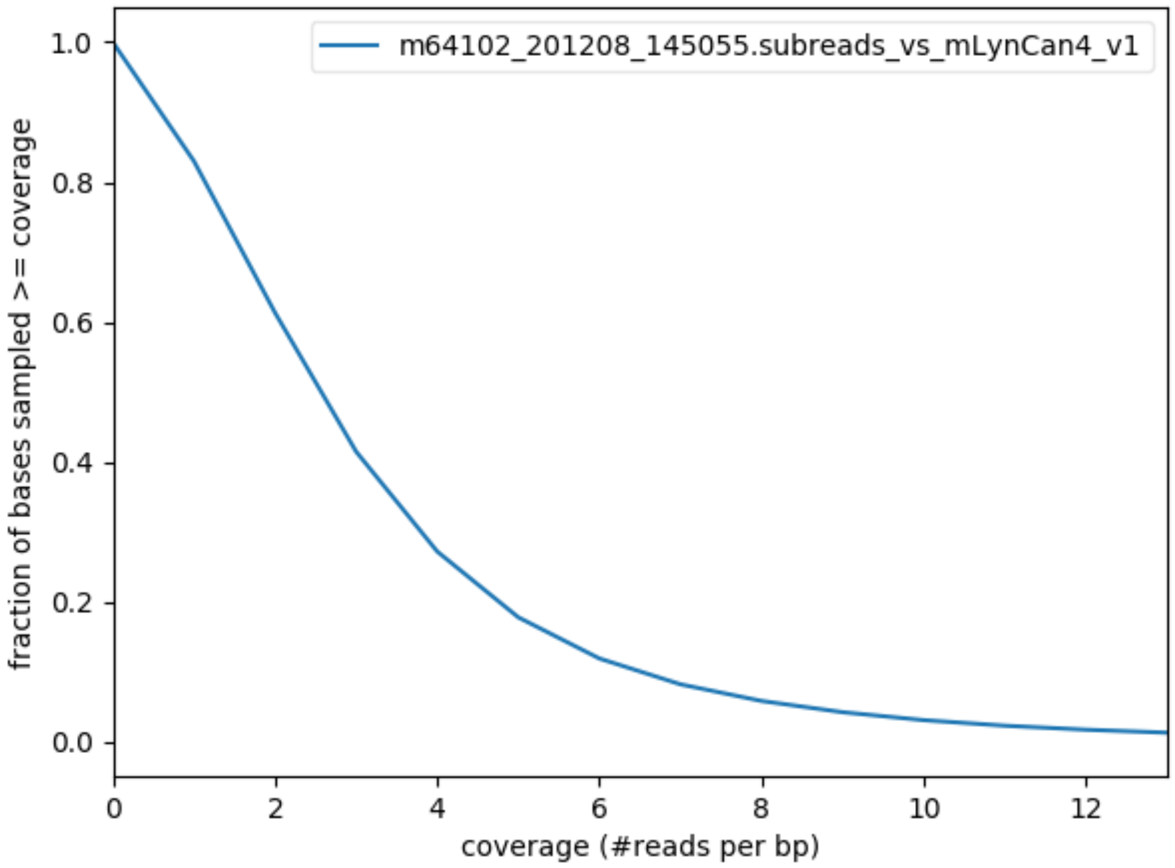
Avoid Abnormal Coverage !

Truncated Distribution



coverage, mean 3x

Skewed towards low coverage



80% of the genome at ≥ **2x**