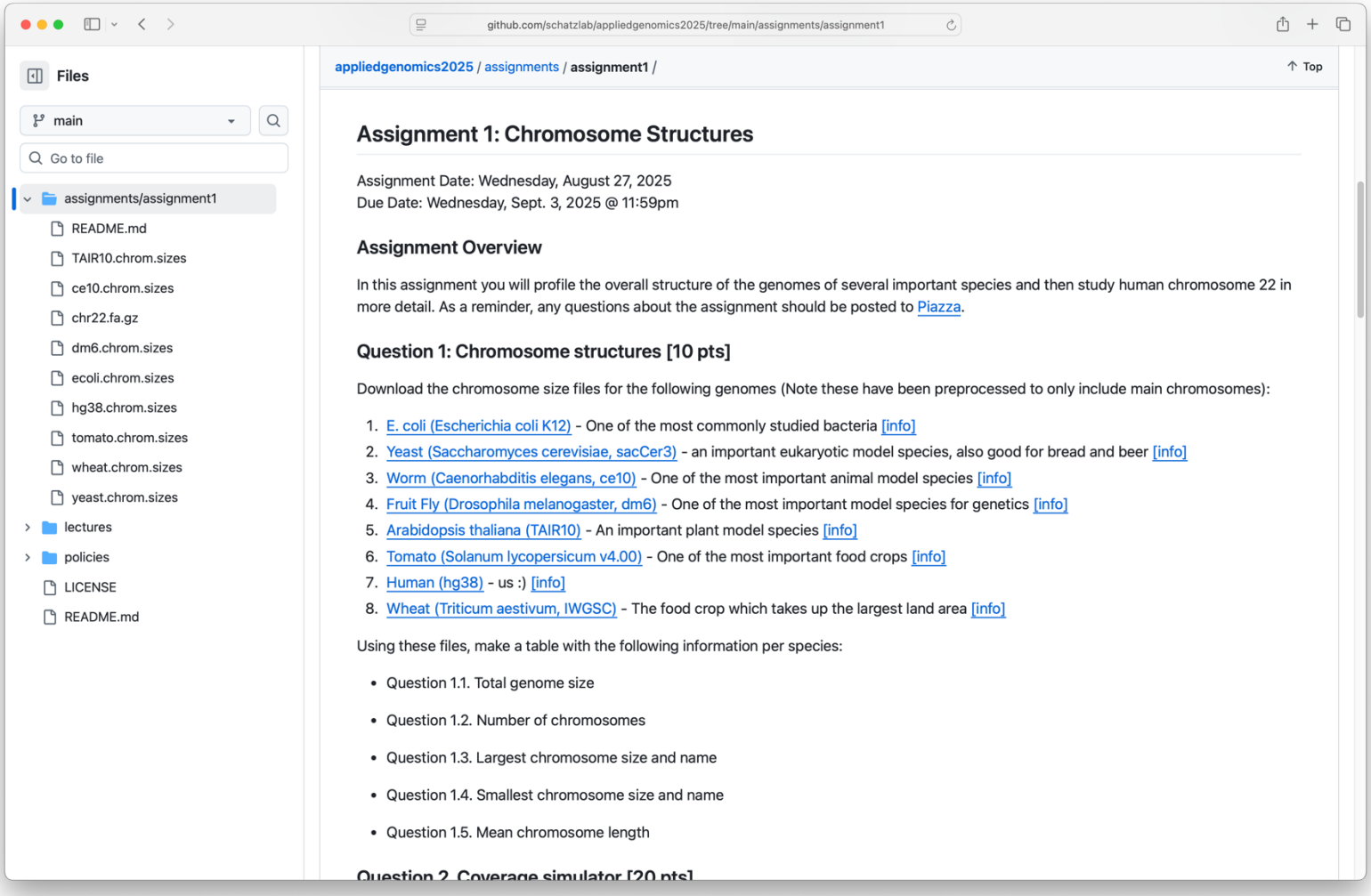
Practical Assembly

Michael Schatz

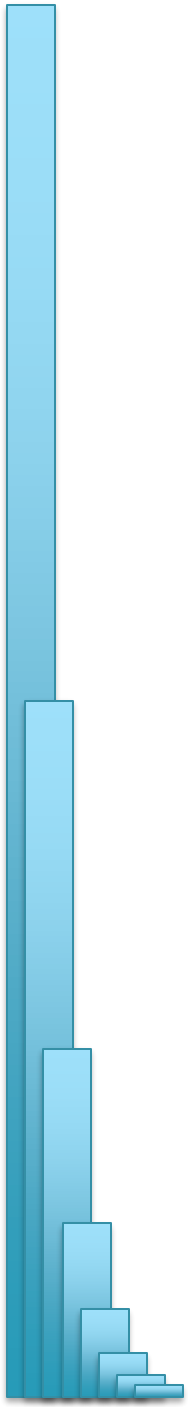
Sept 8, 2025

Lecture 4: Applied Comparative Genomics

Assignment 1



https://github.com/schatzlab/appliedgenomics2025/tree/main/assignments/assignment1
Due end of day on Wednesday Sept 3 (right before midnight)



Part 0: Recap

K-mers and K-mer counting

GATTACATACACATTGGATG

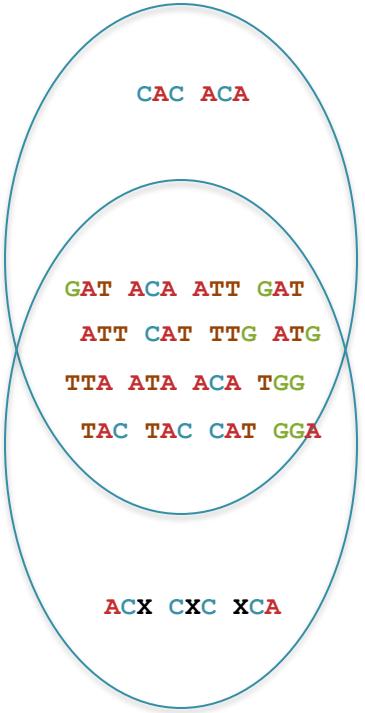
GAT:2	**CAT:2**	**ATG:1**	**TGG:1**
ACA:3	**CAC:1**	**TTA:1**	**TAC:2**
ATT:2	**TTG:1**	**ATA:1**	**GGA:1**

1: 7 (ATG, TGG, …)
2: 4 (GAT, CAT, ATT, TAC)
3: 1 (ACA)

See HW1

GATTACATACACATTGGATG
GAT ACA ACA ATT GAT
ATT CAT CAC TTG ATG
TTA ATA ACA TGG
TAC TAC CAT GGA

GATTACATACXCATTGGATG
GAT ACA ACX ATT GAT
ATT CAT CXC TTG ATG
TTA ATA XCA TGG
TAC TAC CAT GGA



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{16}{21} = 76\%$$

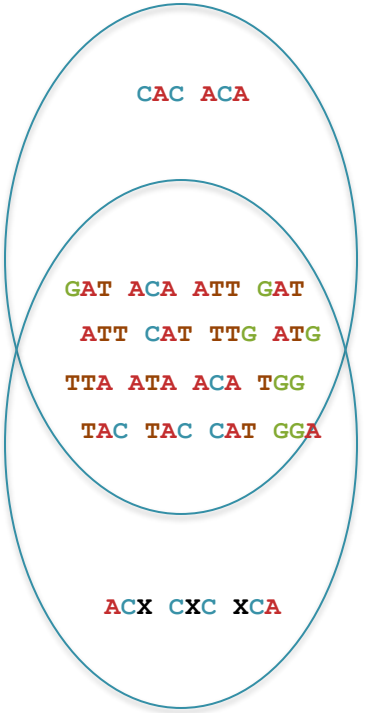
$$\text{ANI} \approx 1 + \frac{1}{k}\ln\left(\frac{2J}{1+J}\right)$$

ANI = 95.1%

GATTACATACACATTGGATG
GAT ACA ACA ATT GAT
ATT CAT CAC TTG ATG
TTA ATA ACA TGG
TAC TAC CAT GGA

GATTACATACXCATTGGATG
GAT ACA ACX ATT GAT
ATT CAT CXC TTG ATG
TTA ATA XCA TGG
TAC TAC CAT GGA

ERRORS!
Thank you to Piazza!



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{16}{21} = 76\%$$

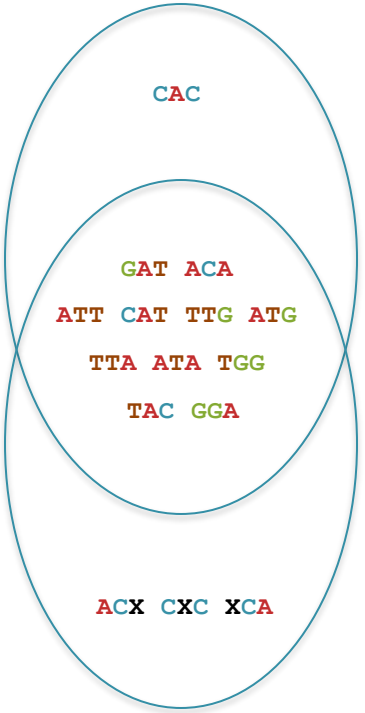
$$\text{ANI} \approx 1 + \frac{1}{k}\ln\left(\frac{2J}{1+J}\right)$$

ANI = 95.1%

```
GATTACATACACATTGGATG
GAT ACA ACA ATT GAT
 ATT CAT CAC TTG ATG
  TTA ATA ACA TGG
   TAC TAC CAT GGA
```

```
GATTACATACXCATTGGATG
GAT ACA ACX ATT GAT
 ATT CAT CXC TTG ATG
  TTA ATA XCA TGG
   TAC TAC CAT GGA
```

Corrected analysis



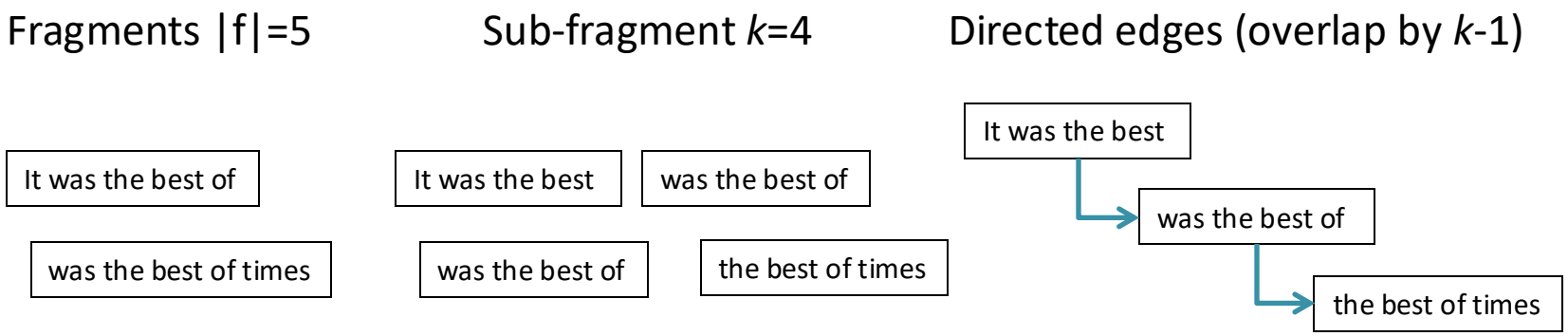
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{11}{15} = 73.3\%$$

$$\mathrm{ANI} \approx 1 + \frac{1}{k}\ln\left(\frac{2J}{1+J}\right)$$

ANI = 94.4%

de Bruijn Graph Construction

- G_k = (V,E)
  - V = Length-*k* sub-fragments
  - E = Directed edges between consecutive sub-fragments
    - Sub-fragments overlap by *k*-1 words

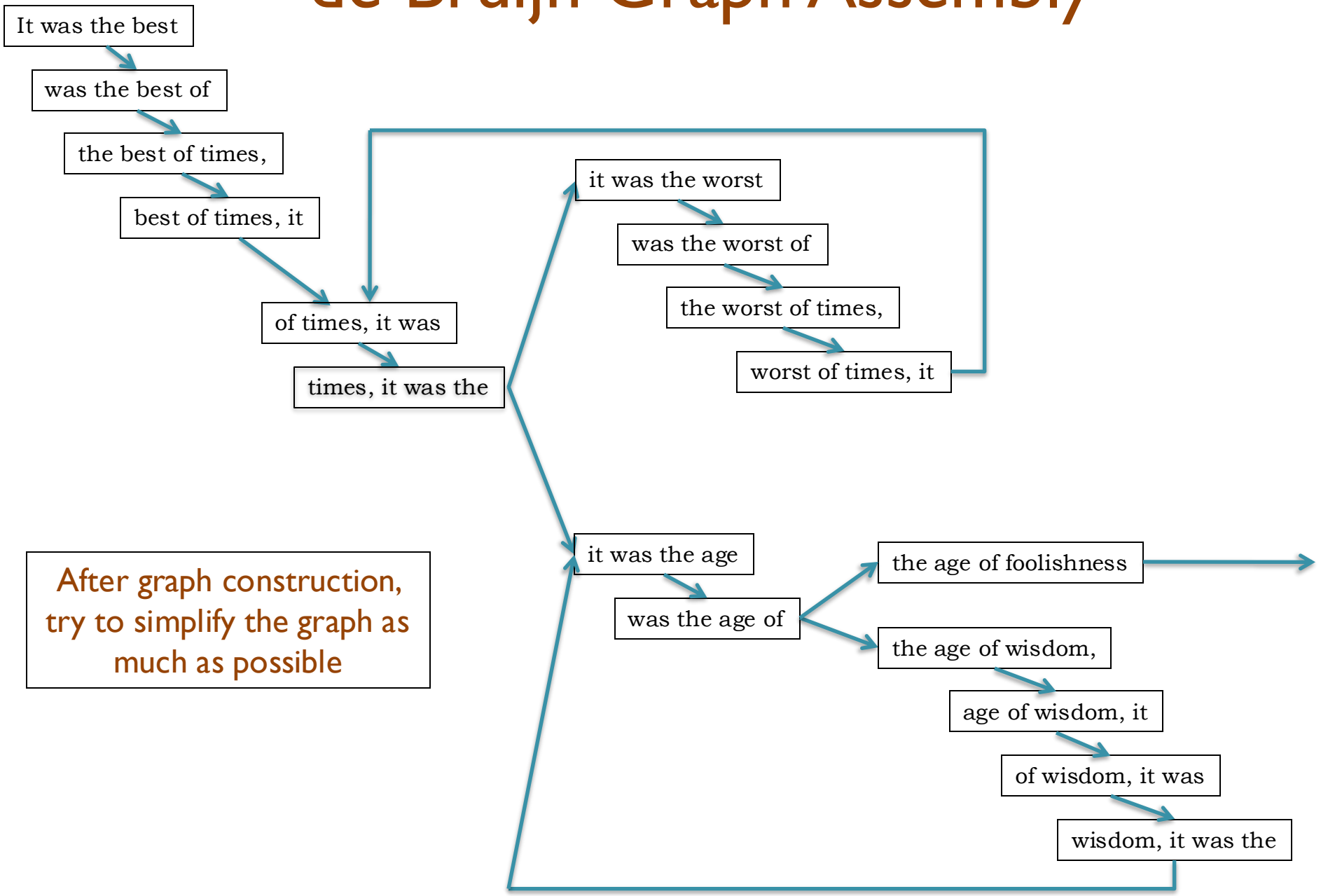


– Overlaps between fragments are implicitly computed

How to pronounce:
https://forvo.com/word/de_bruijn/

de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~

~~ACGT~~

~~ATAC~~

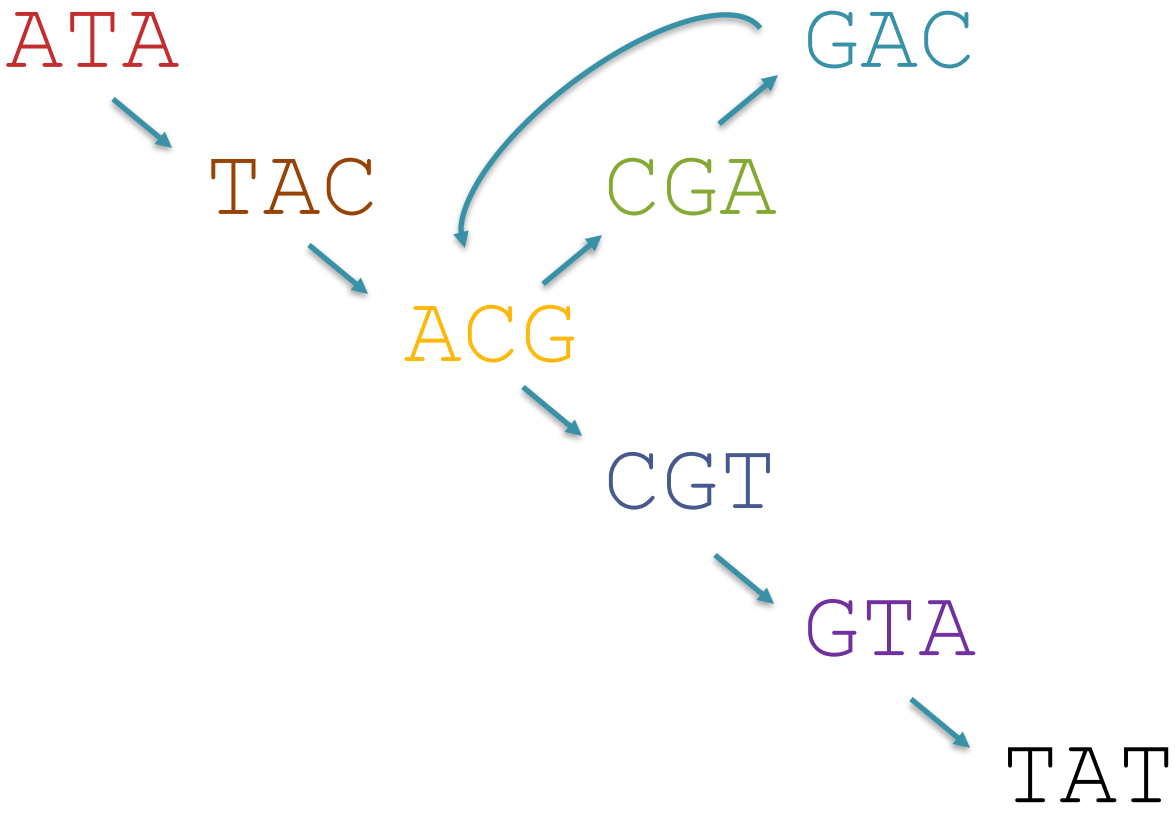
~~CGAC~~

~~CGTA~~

~~GACG~~

~~GTAT~~

~~TACG~~



Whats another possible genome?

ATACGACGTAT

Pop Quiz 2

Assemble these reads using a de Bruijn graph approach (k=3):

~~ACGA~~

~~ACGT~~

~~ATAC~~

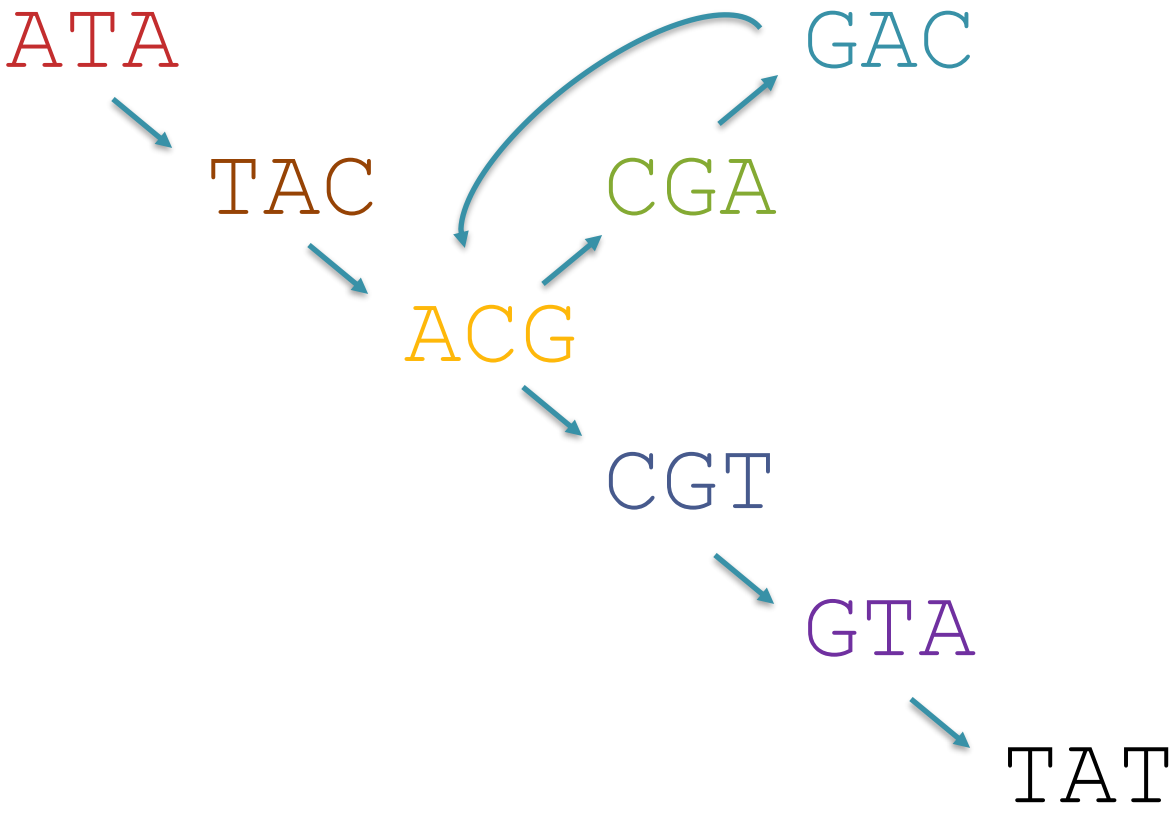
~~CGAC~~

~~CGTA~~

~~GACG~~

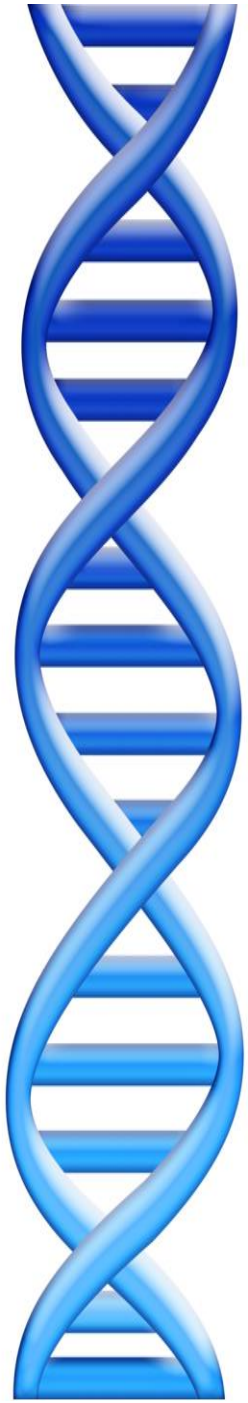
~~GTAT~~

~~TACG~~



Should we add the edge TAT -> ATA?

ATACGACGTAT



Outline

1. ***Assembly theory***
   - Assembly by analogy

2. ***Practical Issues***
   - Coverage, read length, errors, and repeats

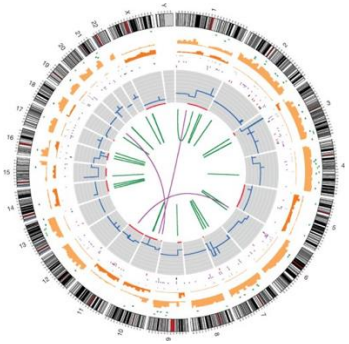
3. Whole Genome Alignment
   - MUMmer recommended

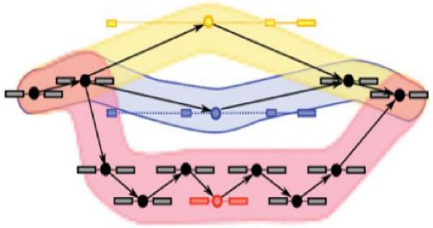
Assembly Applications

- Novel genomes

- Metagenomes

- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …





Why are genomes hard to assemble?

1. ***Biological***:
   - (Very) High ploidy, heterozygosity, repeat content

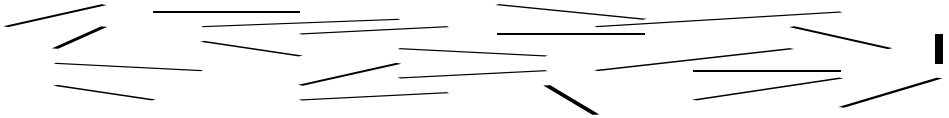
2. ***Sequencing***:
   - (Very) large genomes, imperfect sequencing

3. ***Computational***:
   - (Very) Large genomes, complex structure

4. ***Accuracy***:
   - (Very) Hard to assess correctness

Assembling a Genome

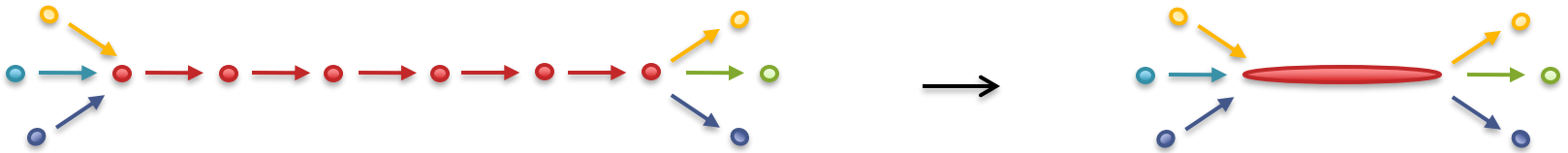
1. Shear & Sequence DNA



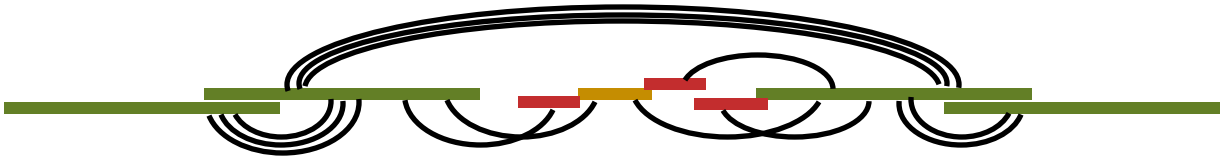
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

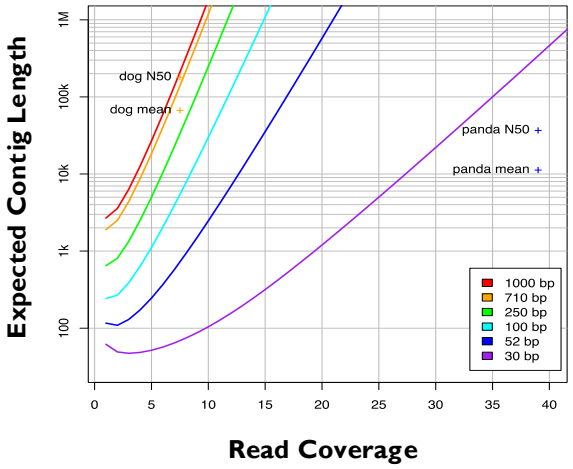


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

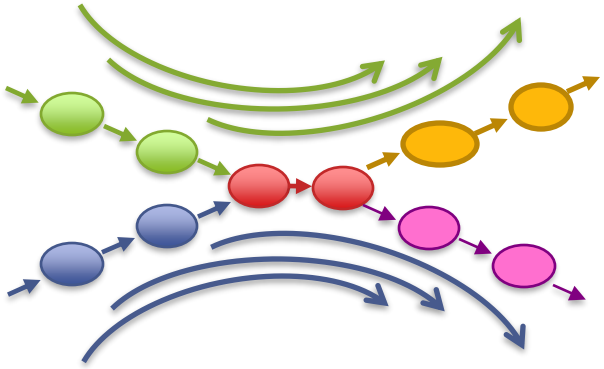
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

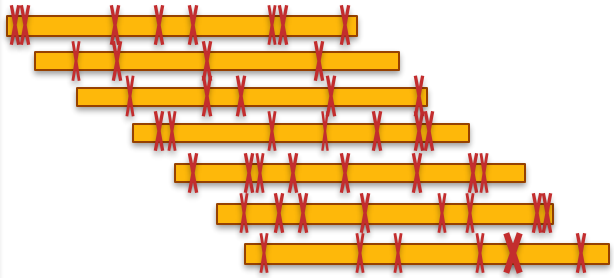
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

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

Coverage Statistics

$$\text{sequencing\_coverage} = \frac{\text{total\_bases\_sequenced}}{\text{genome\_size}}$$

$$\text{genome\_size} = \frac{\text{total\_bases\_sequenced}}{\text{sequencing\_coverage}}$$

$$\text{genome\_size} = \frac{100\text{Gb}}{50\text{x}} = 2\text{Gb}$$

But how can you figure out
the coverage without a genome?

K-mer counting

Kmer-ize

```
Read 1: GATTACA  => GAT,ATT,TTA,TAC,ACA
Read 2: TACAGAG  => TAC,ACA,CAG,AGA,GAG
Read 3: TTACAGA  => TTA,TAC,ACA,CAG,AGA
```

list

```
GAT   ACA   ACA:3
ATT   ACA
TTA   ACA
TAC   AGA   AGA:2
ACA   AGA
TAC   ATT   ATT:1
ACA   CAG   CAG:2
CAG   CAG
AGA   GAG   GAG:1
GAG   GAT   GAT:1
TTA   TAC   TAC:3
TAC   TAC
ACA   TAC
CAG   TTA   TTA:2
AGA   TTA
```

sort **count**

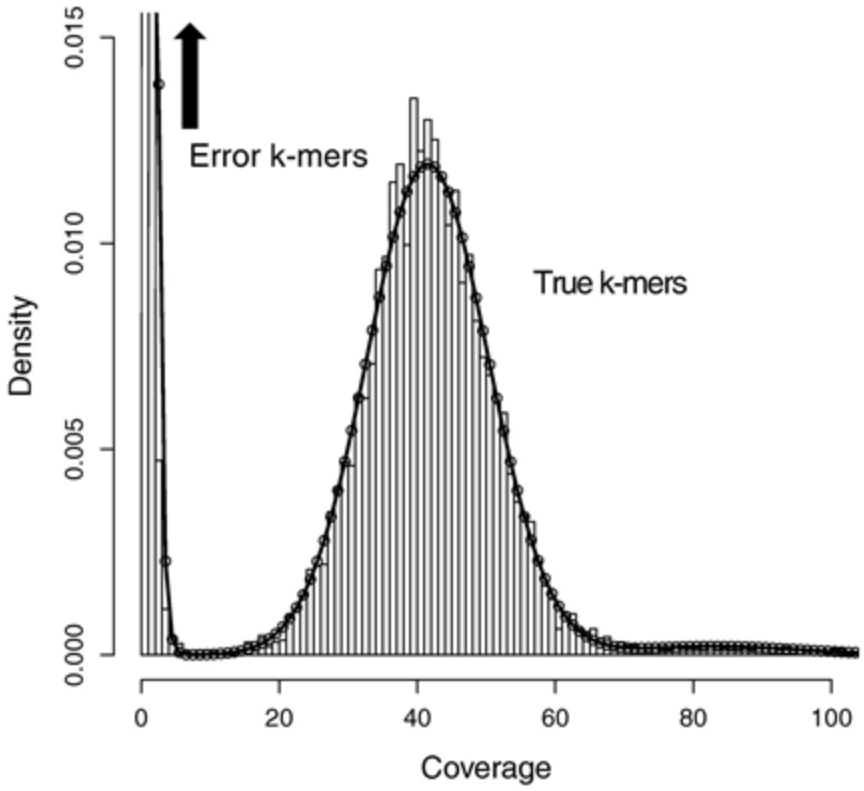
tally

```
3 kmers occur 1x
3 kmers occur 2x
2 kmers occur 3x
```

From read k-mers alone, can learn something about how frequently different sequences occur (aka coverage)

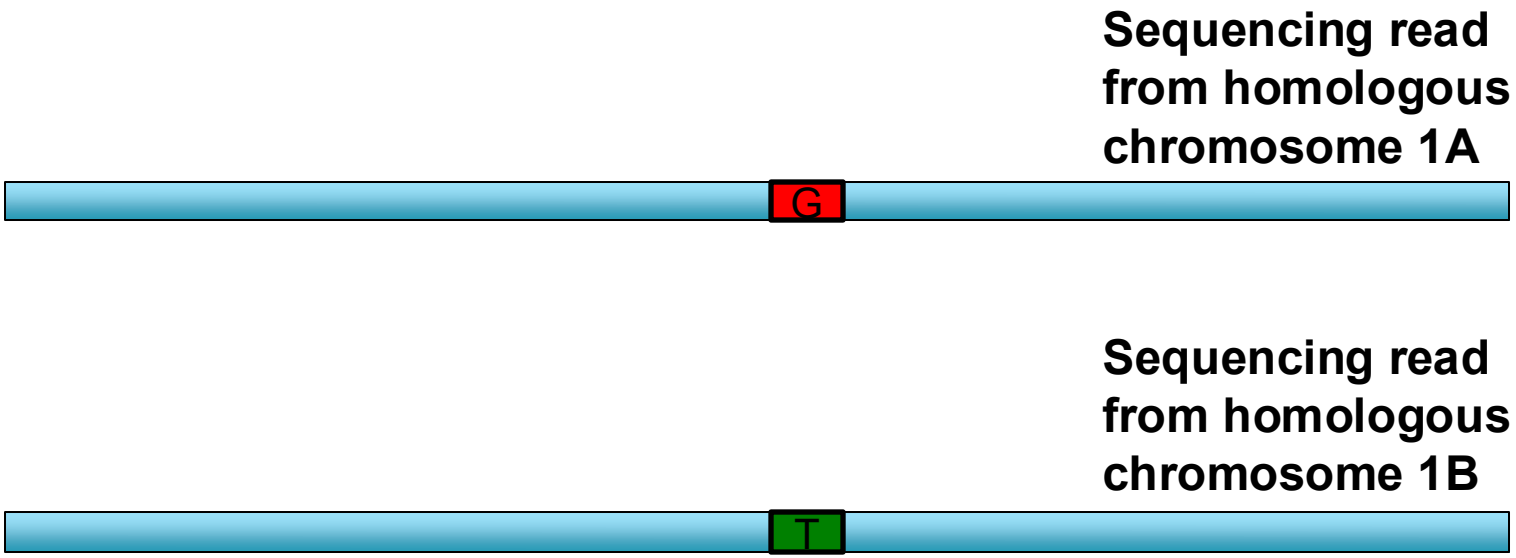
Fast to compute even over huge datasets

K-mer counting in real genomes

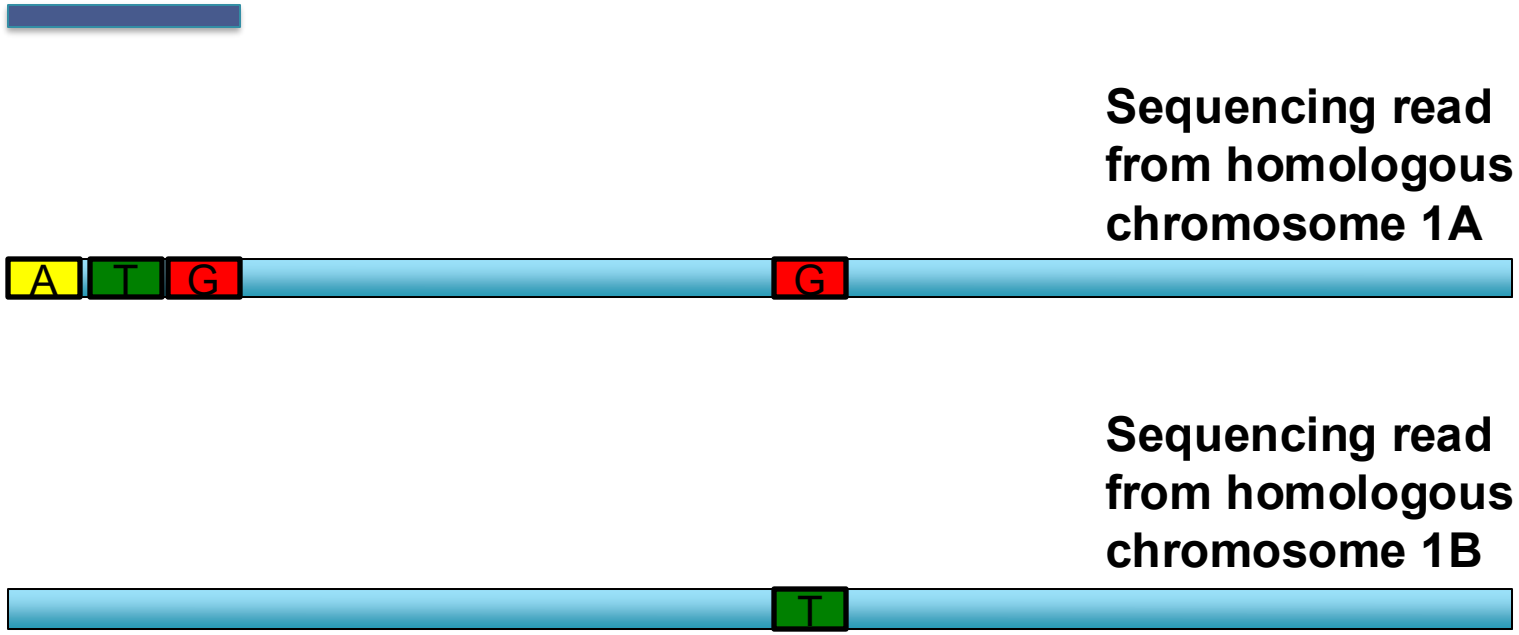


- The tally of k-mer counts in real genomes reveals the coverage distribution.
- Here we sequenced 120Gb of reads from a female human (haploid human genome size is 3Gb), and indeed we see a clear peak centered at 40x coverage
- There are also many kmers that only occur <5 times. These are from errors in the reads
- There are also kmers that occur many times (>>70 times). These are repeats in the genome

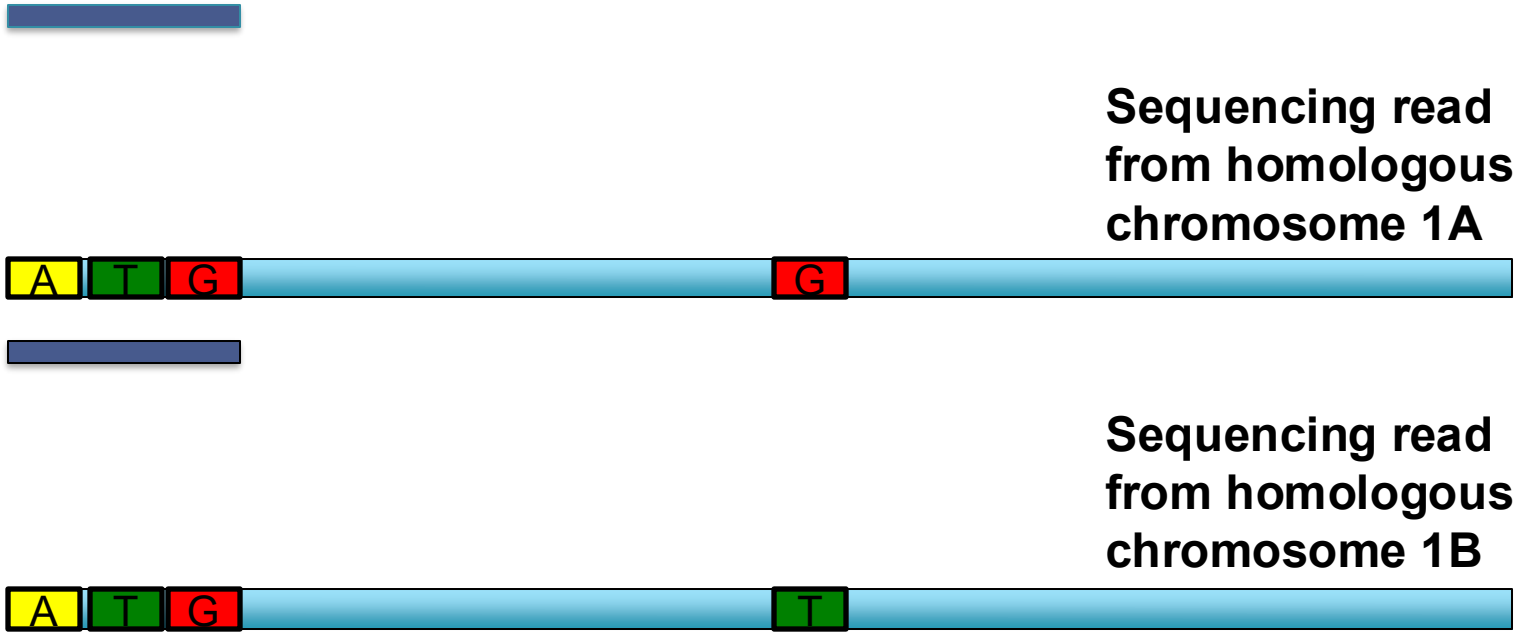
K-mer counting in heterozygous genomes



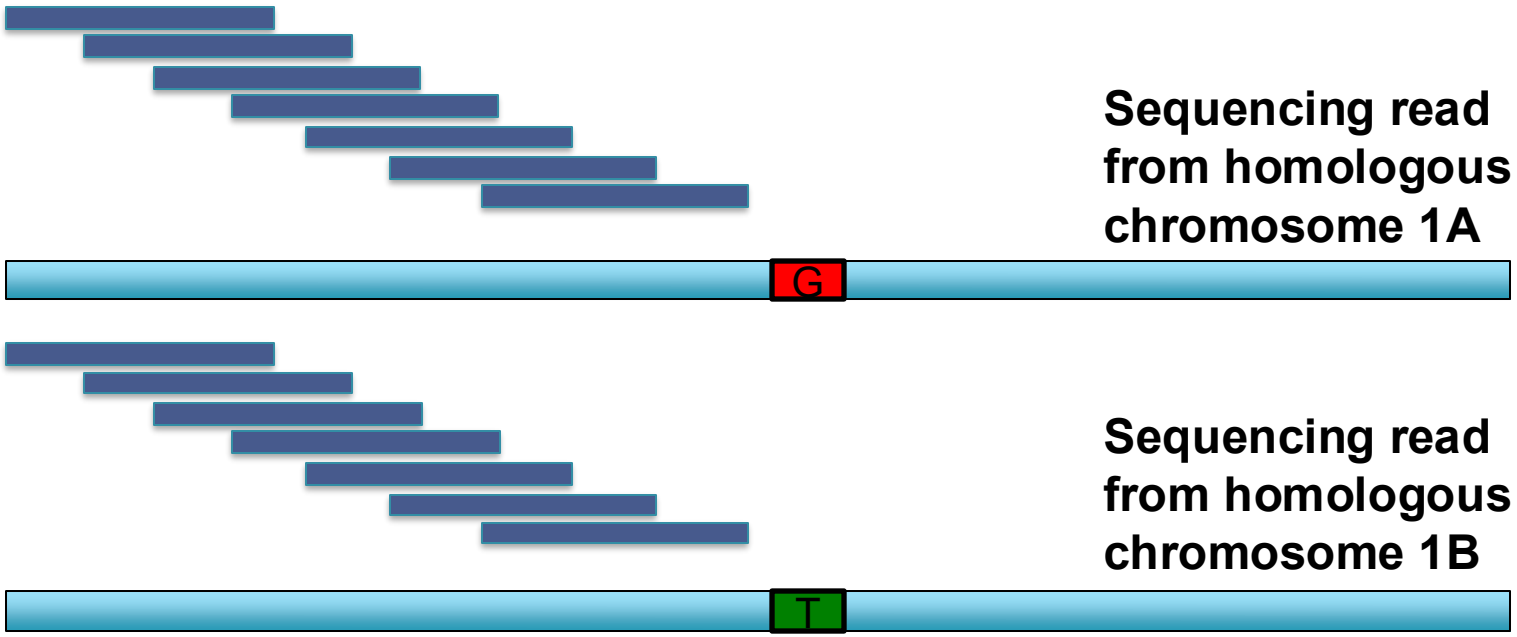
K-mer counting in heterozygous genomes



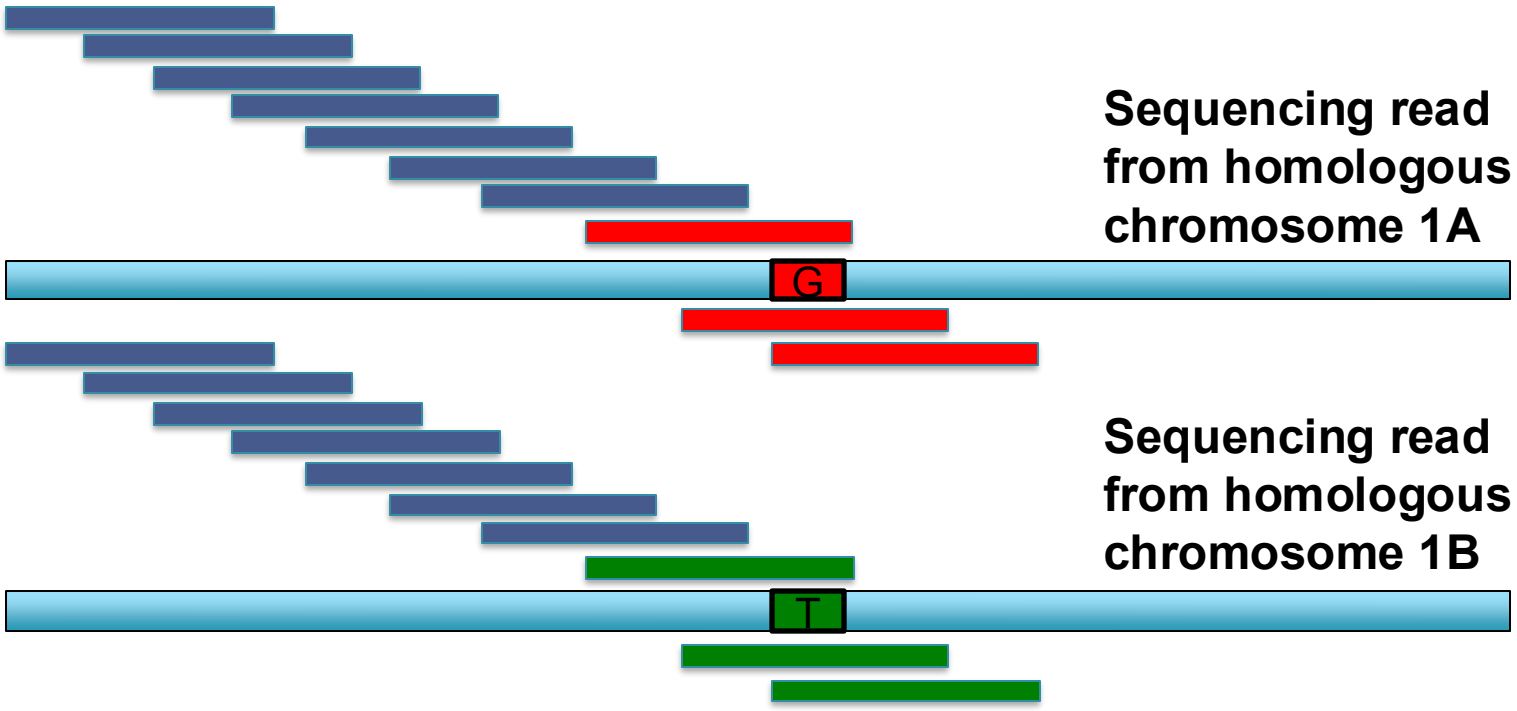
K-mer counting in heterozygous genomes



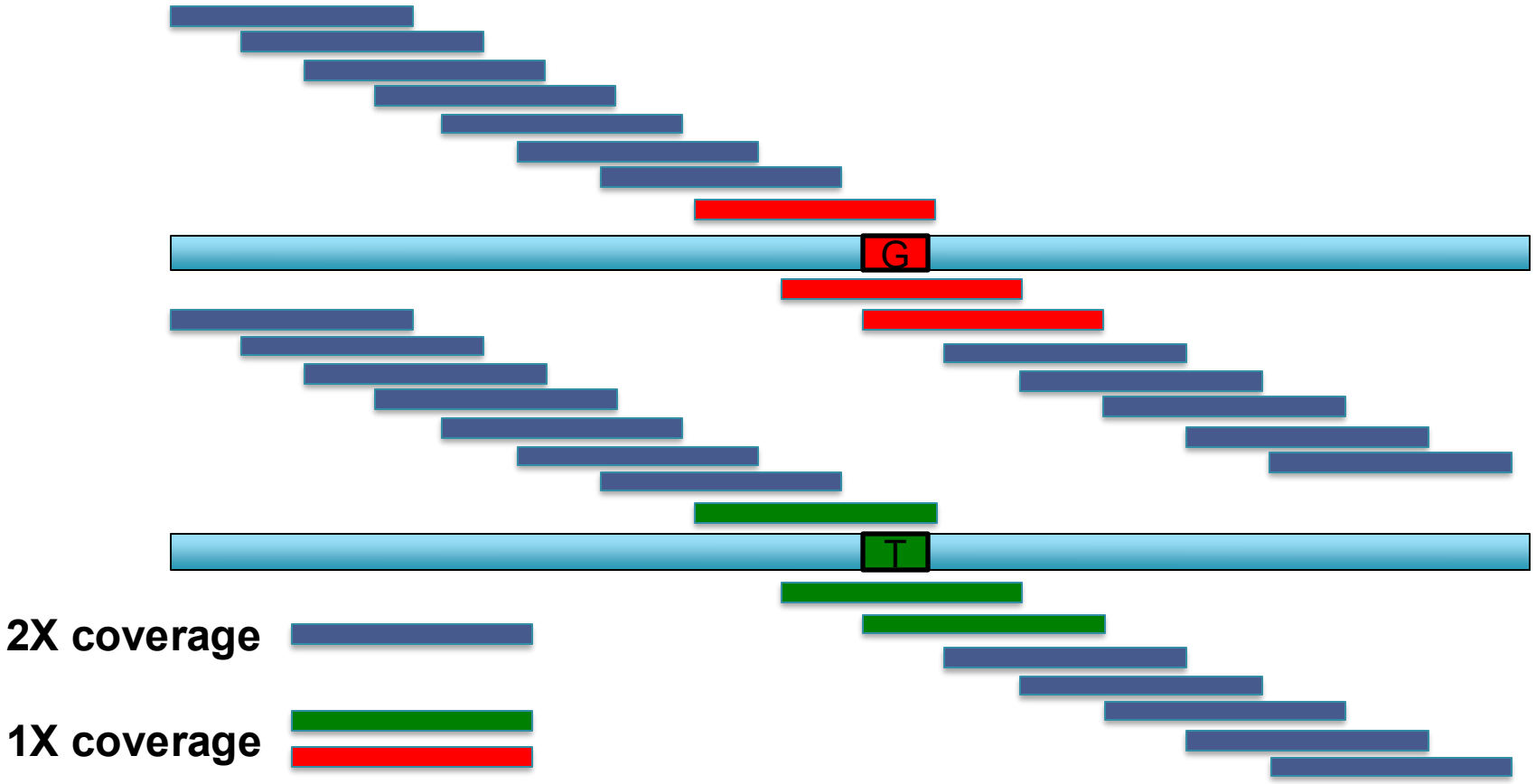
K-mer counting in heterozygous genomes



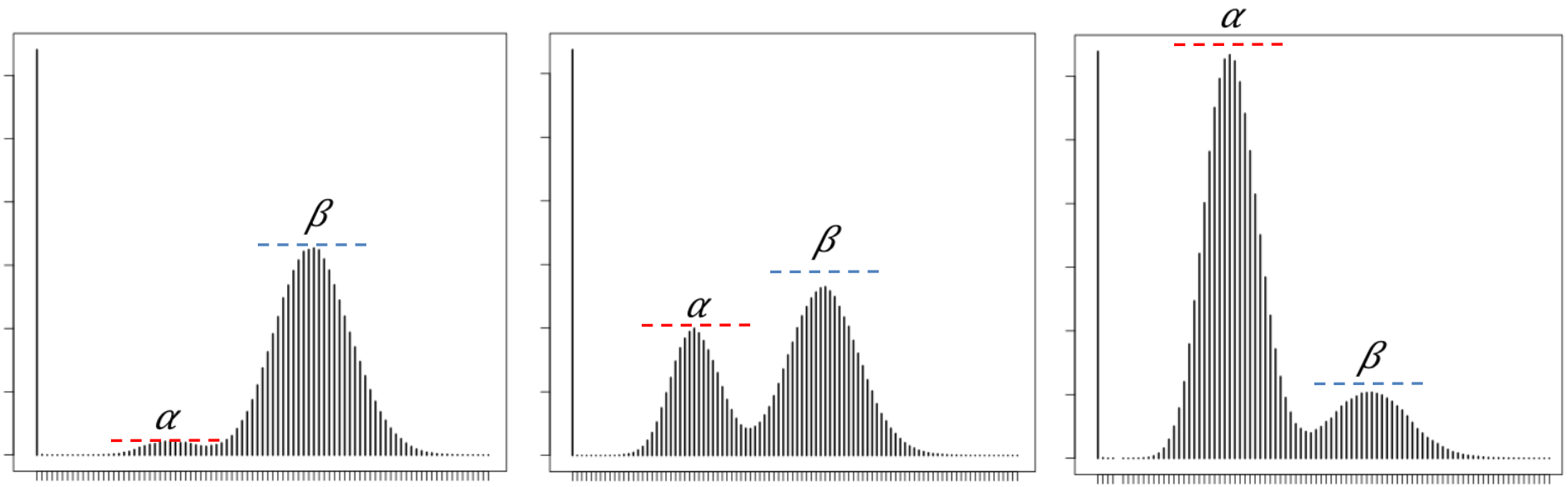
K-mer counting in heterozygous genomes



K-mer counting in heterozygous genomes



Heterozygous Kmer Profiles

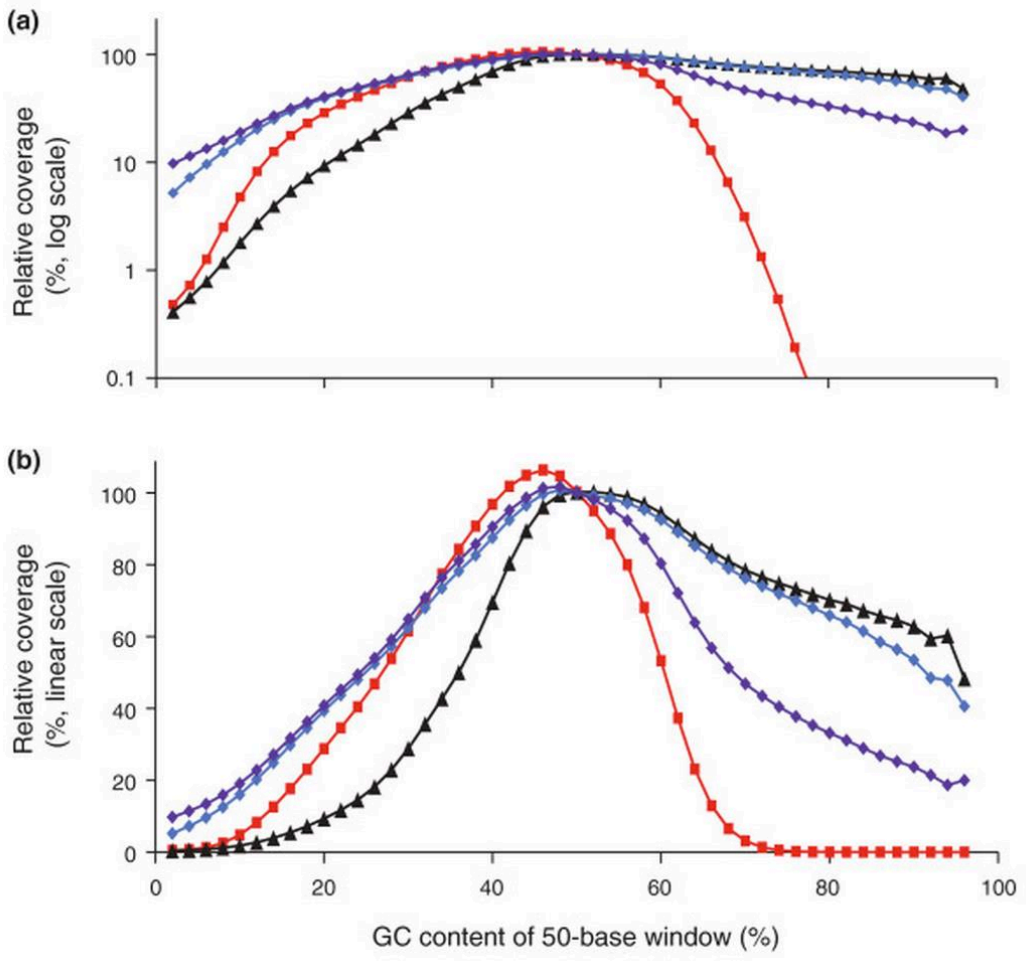


0.1% heterozygosity **1% heterozygosity** **5% heterozygosity**

- ***Heterozygosity creates a characteristic "double-peak" in the Kmer profile***
  - Second peak at twice k-mer coverage as the first: heterozygous kmers average 50x coverage, homozygous kmers average 100x coverage

- ***Relative heights of the peaks is directly proportional to the heterozygosity rate***
  - The peaks are balanced at around 1.25% because each heterozygous SNP creates 2*k heterozygous kmers (typically k = 21)

Beware of GC Biases



Illumina sequencing does not produce uniform coverage over the genome

- Coverage of extremely high or extremely low GC content will have reduced coverage in Illumina sequencing
- Biases primarily introduced during PCR; lower temperatures, slower heating, and fewer rounds minimize biases
- This makes it very difficult to identify variants (SNPs, CNVs, etc) in certain regions of the genome

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology.* 12:R18.

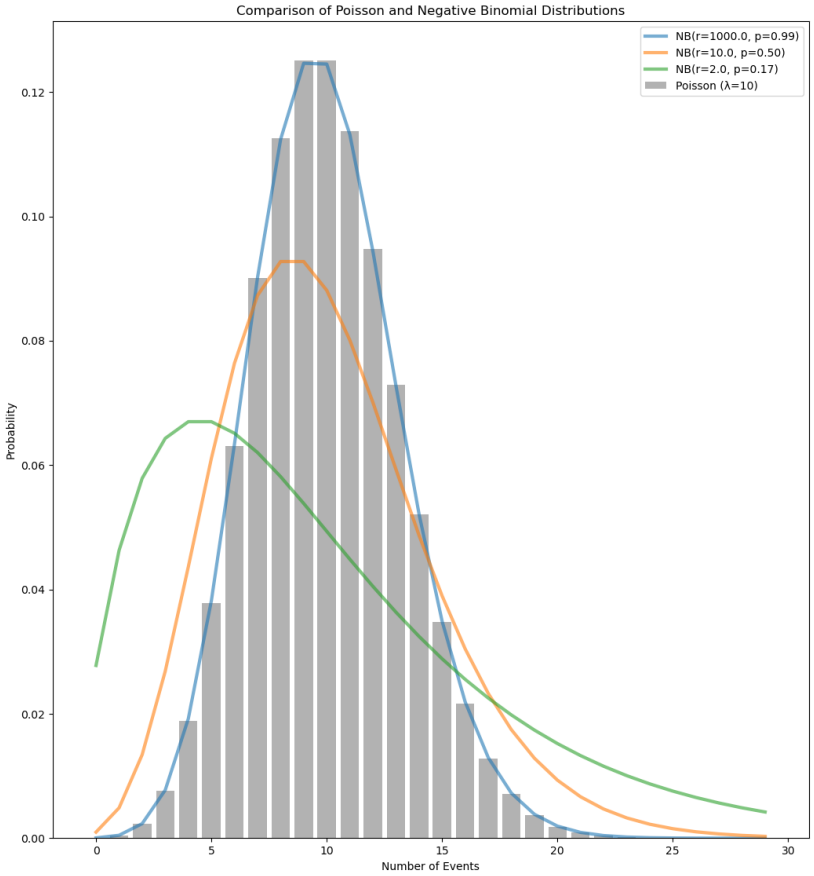
Negative Binomial Distribution

Models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified number of successes (r) occurs

- Commonly used to model over-dispersed count data

Also arises as a continuous mixture of Poisson distributions where the mixing distribution of the Poisson rate is a gamma distribution.

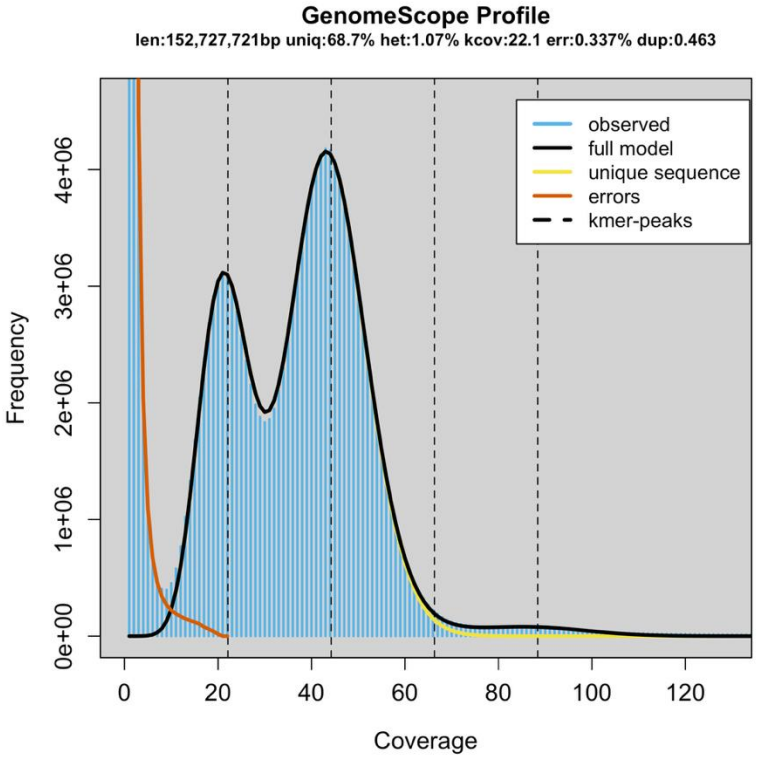
- Sequence coverage distribution where there is a non-uniform probability of a read starting at each position

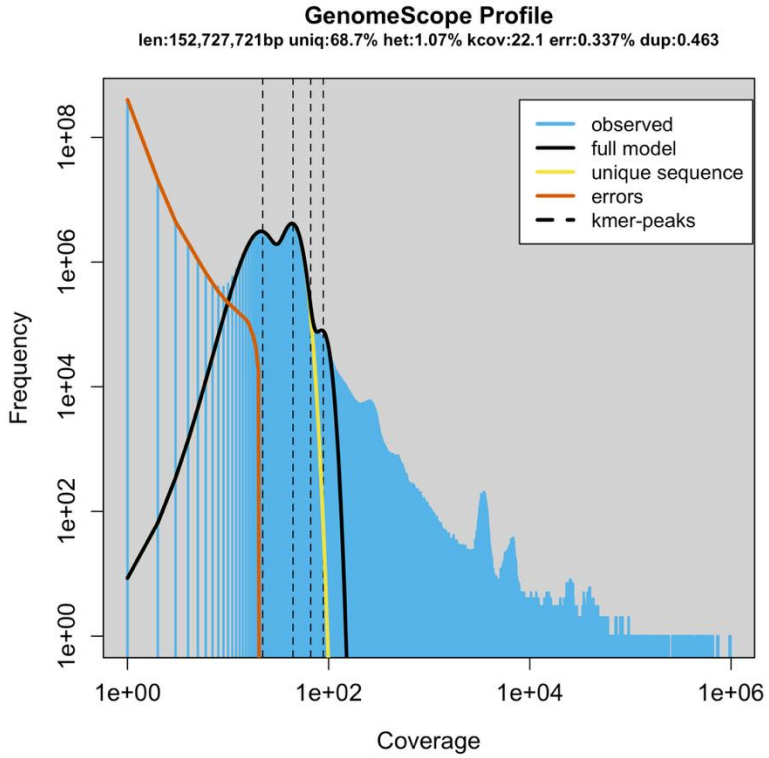


$$f(k;r,p) \equiv \Pr(X = k) = \binom{k+r-1}{k}(1-p)^k p^r$$

GenomeScope: Fast genome analysis from short reads

http://genomescope.org





$$f(k) = \alpha \cdot \mathrm{NB}(k; r_1, p_1) + \beta \cdot \mathrm{NB}(k; r_2, p_2)$$

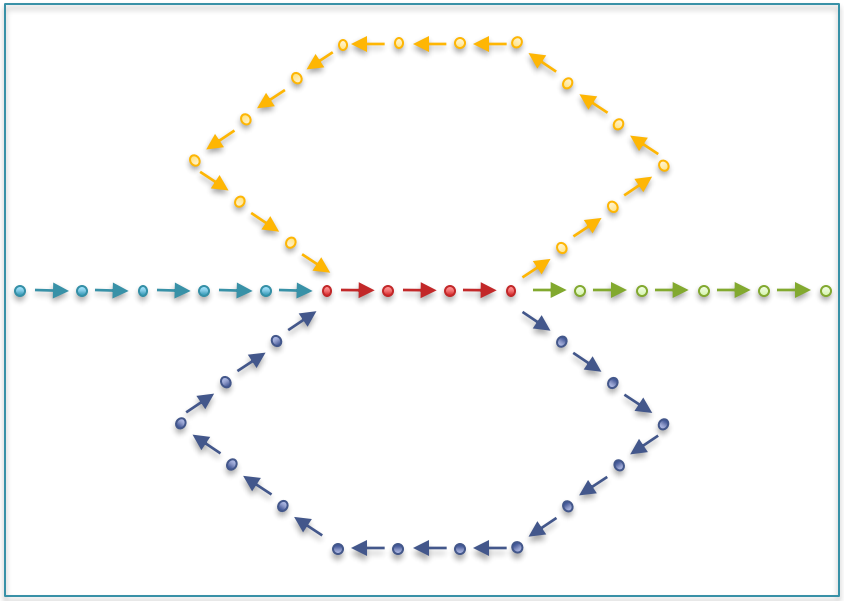
- Theoretical model agrees well with published results:
  - Quickly estimate genome size, rate of heterozygosity, and other genome properties
  - Generalized to higher ploidies by introducing additional terms
  - “Reference-free analysis” does not require the use of an assembled genome

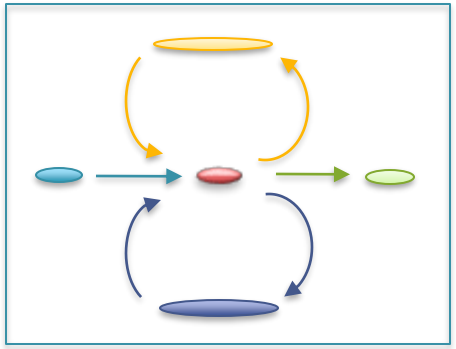
Vurture, GW*, Sedlazeck FJ*, et al. (2017) *Bioinformatics*
Ranallo-Benavidez, TR. *et al.* (2020) *Nature Communication*

Errors

Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"





Why do contigs end?

(1) End of chromosome! ☺, (2) lack of coverage, (3) errors,
(4) heterozygosity and (5) repeats

Errors in the graph

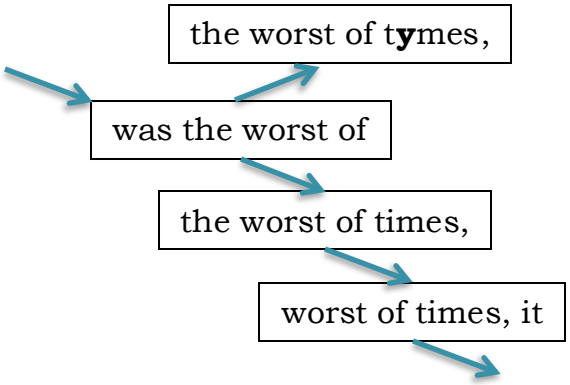
(Chaisson, 2009)

Clip Tips

was the worst of times,

was the worst of t**y**mes,

the worst of times, it



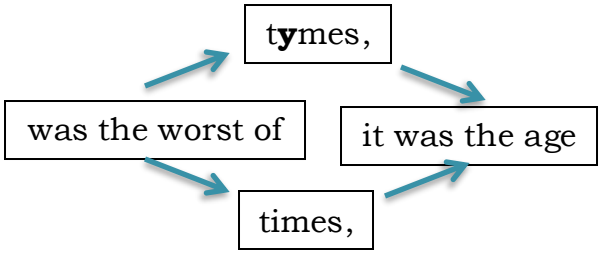
Pop Bubbles

was the worst of times,

was the worst of t**y**mes,

times, it was the age

t**y**mes, it was the age

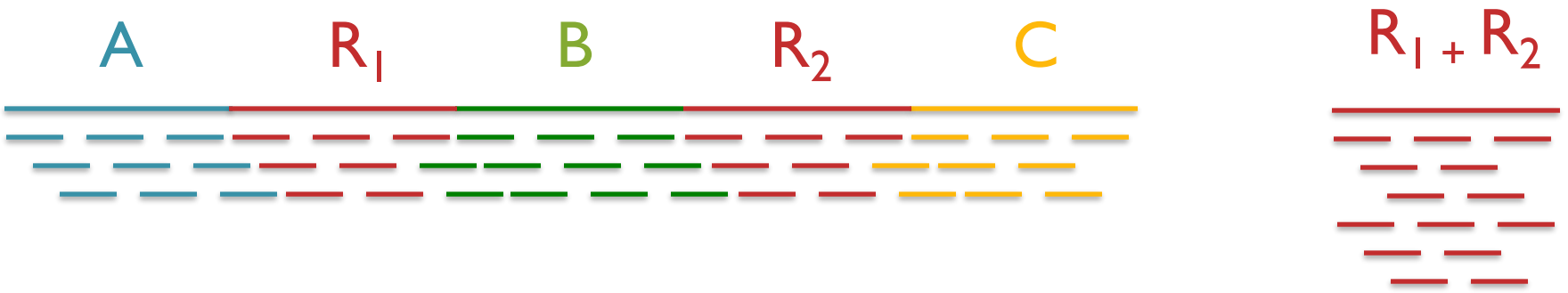


Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ where $1 \leq k \leq 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	*Alu* sequence (~280 bp) Mariner elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

- Over 50% of mammalian genomes are repetitive
  - Large plant genomes tend to be even worse
  - Wheat: 16 Gbp; Pine: 24 Gbp

Repeats and Coverage Statistics



- If *n* reads are a uniform random sample of the genome of length *G*, we expect $k=n\Delta/G$ reads to start in a region of length Δ.
  - If we see many more reads than k (if the arrival rate is > A) , it is likely to be a collapsed repeat

$$\Pr(X-copy)=\binom{n}{k}\left(\frac{X\Delta}{G}\right)^{k}\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

$$A(\Delta,k)=\ln\left(\frac{\Pr(1-copy)}{\Pr(2-copy)}\right)=\ln\left(\frac{\frac{(\Delta n/G)^{k}}{k!}e^{\frac{-\Delta n}{G}}}{\frac{(2\Delta n/G)^{k}}{k!}e^{\frac{-2\Delta n}{G}}}\right)=\frac{n\Delta}{G}-k\ln 2$$

The fragment assembly string graph
Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

Illumina Hacking

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 12 2013, pages 1492–1497
doi:10.1093/bioinformatics/btt178

Genome analysis

Advance Access publication May 22, 2013

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D. Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A. Taylor[1], Macaire Man Saint Yuen[4], Christopher I. Keeling[4], Dana Brand[1], Benjamin P. Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A. Moore[1], Yongjun Zhao[1], Andrew J. Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Carol Ritland[4,6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6], Steven J.M. Jones[1,2,9]

[1]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada, [2]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada, [3]School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada, [4]Michael Smith Laboratories, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [5]British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2, Canada, [6]Department of Forest Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, [7]Institute for Systems and Integrative Biology, Université Laval, Québec, QC G1V 0A6, Canada, [8]Department of Wood and Forest Sciences, Université Laval, Québec, QC G1V 0A6, Canada and [9]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Associate Editor: Michael Brudno

ABSTRACT

White spruce (*Picea glauca*) is a dominant conifer of the boreal forests of North America, and providing genomics resources for this commercially valuable tree will help improve forest management and conservation efforts. Sequencing and assembling the large and highly repetitive spruce genome though pushes the boundaries of the current technology. Here, we describe a whole-genome shotgun sequencing strategy using two Illumina sequencing platforms and an assembly approach using the ABySS software. We report a 20.8 giga base pairs draft genome in 4.9 million scaffolds, with a scaffold N50 of 20 356 bp. We demonstrate how recent improvements in the sequencing technology, especially increasing read lengths and paired end reads from longer fragments have a major impact on the assembly contiguity. We also note that scalable bioinformatics tools are instrumental in providing rapid draft assemblies.

Availability: The *Picea glauca* genome sequencing and assembly data are available through NCBI (Accession#: ALWZ0100000000 PID: PRJNA83435). http://www.ncbi.nlm.nih.gov/bioproject/83435.

Contact: ibirol@bcgsc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 20, 2013; revised on April 10, 2013; accepted on April 11, 2013

1 INTRODUCTION

The assembly of short reads to develop genomic resources for non-model species remains an active area of development (Schatz *et al.*, 2012). The feasibility of the approach and its scalability to large genomes was demonstrated by (Simpson *et al.*, 2009) using human and was later used to assemble the SOAPdenovo tool (Li *et al.*, 2010). high quality results, as demonstrated *et al.*, 2012; Ladner *et al.*, 2013; Rib been successfully applied numerous ti omes (Chan *et al.*, 2011; Chu *et al.*, 2 2011; Godel *et al.*, 2012; Swart *et al.*,

Estimated at 20 giga base pairs (Gb cing and assembly of the genome of t the pine (*Pinaceae*) family present unic generation end, those challenges inclu whole-genome shotgun sequencing dat ing reduced representation resources tude of the problem. On the bioinf massive sequencing datasets is extreme ing cycles, memory usage, storage req programming implementations on co

We addressed the data representatio and sequencing multiple whole-genom HiSeq 2000 and MiSeq sequencers f CA, USA). Compared with localized as building and sequencing fosmid approach of isolating ~10 kb DNA s sequencing fragments in high through CA, USA), a shotgun only sequencing sequence data effectively covering th that can be an order of magnitude les especially substantial when sequencing

In this work, we demonstrate that s at this scale remains viable and pro

*To whom correspondence should be addressed.

assemble the spruce genome, we used the ABySS algorithm (Simpson *et al.*, 2009), which captures a representation of read-to-read overlaps by a distributed de Bruijn graph and uses parallel computations to build the target genome. The modular nature of the tool allowed us to execute a large number of tests to tune the message passing interface for a successful execution, train the assembly parameters for an optimal assembly and quantify the utility of long reads for large genome assemblies. To the best of our knowledge, the ABySS algorithm is unique in its ability to enable genome assemblies of this scale using whole-genome shotgun sequencing data.

2 METHODS

2.1 Sample collection

Apical shoot tissues were collected in April 2006 from a single white spruce (*Picea glauca*, genotype PG29) tree at the Kalamalka Research Station of the British Columbia Ministry of Forests and Ranges, Vernon, British Columbia, Canada. Genomic DNA was extracted from 60 gm tissue by BioS&T (http://www.biost.com/, Montreal, QC, Canada) using an organelle exclusion method yielding 300 µg of high quality purified nuclear DNA.

2.2 Library preparation and sequencing

DNA quality was assessed by spectrophotometry and gel electrophoresis before library construction. DNA was sheared for 45 s using an E210 sonicator (Covaris) and then analysed on 8% PAGE gels. The 200–300 bp (for libraries with 250 bp insert size) or 450–550 bp (for libraries with 500 bp insert size) DNA size fractions were excised and eluted from the gel slices overnight at 4°C in 300 µl of elution buffer {5:1 [vol/vol] LoTE buffer {3 mM Tris–HCl (pH 7.5), 0.2 mM EDTA}/7.5 M ammonium acetate} and was purified using a Spin-X Filter Tube (Fisher Scientific) and ethanol precipitation. Genome libraries were prepared using a modified paired-end tag (PET) protocol supplied by Illumina Inc. This involved DNA end repair and formation of 3′ adenosine overhangs using the Klenow fragment of DNA polymerase I (3′–5′ exonuclease minus) and ligation to Illumina PE adapters (with 5′ overhangs). Adapter-ligated products were purified on QIAquick spin columns (Qiagen) and amplified using Phusion DNA polymerase (NEB) and 10 PCR cycles with the PE primer 1.0 and 2.0 (Illumina). PCR products of the desired size range were purified from adapter ligation artifacts using 8% PAGE gels. DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay (Agilent) and Nanodrop 7500 spectrophotometer (Nanodrop). DNA was subsequently diluted to 8 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen).

The mate pair (MPET, a.k.a. jumping) libraries were constructed using 4 µg of genomic DNA with the Illumina Nextera Mate-Pair library construction protocol and reagent (FC-132-1001). The genomic DNA sample was simultaneously fragmented and tagged with a biotin containing mate pair junction adapter, which left a short sequence gap in the tagmented DNA. The gap was filled by a strand displacement reaction using a polymerase to ensure that all fragments were flush and ready for circularization. After an AMPure Bead cleanup, size selection was done on a 0.6% agarose gel to excise 6–9 kb and 9–13 kb fractions, which were purified using a Zymoclean Large Fragment DNA Recovery Kit. The fragments were circularized by ligation, followed by a digestion to remove any linear molecules and left circularized DNA for shearing. The sheared DNA fragments that contain the biotinylated junction adapter (mate pair fragments) were purified by means of binding to streptavidin magnetic beads, and the unwanted unbiotinylated molecules were washed away. The DNA fragments were then end repaired and A-tailed following the protocol and ligated to indexed TruSeq adapters. The final library was enriched by a 10-cycle PCR and purified by AMPure bead clean-up. Library quality and size were assessed by Agilent DNA 1000 series II assay and KAPA Library Quantification protocol. The two fractions were pooled for sequencing paired end 100 bp using Illumina HiSeq2000.

The construction of the 12 kb mate pair libraries was achieved by a hybrid 454/Illumina procedure. Briefly, 50 µg of genomic DNA was fragmented for 20 cycles at speed code 12 using a Hydroshear (Digilab, Marlborough, MA) equipped with a large assembly module. The fragmented DNA was loaded on a 1% agarose gel, and fragments of 18 kb were extracted. Biotinylated circularization adapters (454 Life Sciences/Roche Titanium Paired-end Adaptor set (454 Life Sciences/Roche, Branford, CT) were added to ends of the gel-extracted fragments. After recombination of the ends was performed with Cre recombinase (New England Biolabs, Ipswich, MA), and linear molecules remaining after digestion were removed with Plasmid Safe (Epicentre, Madison, WI). The circular molecules were fragmented using GS Rapid Library Nebulization System (454 Life Sciences/Roche, Branford, CT), and fragment end-repair followed by A-tailing was performed with the GS Rapid Library preparation kit (454 Life Sciences/Roche, Branford, CT). TruSeq Adaptors (Illumina, San Diego, CA) were ligated to the repaired/A-tailed ends. Biotinylated fragments were enriched using Streptavidin-coupled Dynabeads (Life Technologies, Grand Island, NY) and amplified by PCR using Illumina TruSeq primers.

Random bacterial artificial chromosome (BAC) sequencing was performed using DNA from the same genotype on a 454 Titanium with 6 kb paired-end libraries at the PlateForme Génomiques of the Institute for Systems and Integrative Biology (Université Laval, Quebec City, QC). A single paired-end library was prepared on a pool of 15 BACs (equimolar concentrations) as described earlier in the text with the following modifications: 15 µg of DNA was fragmented using a Hydroshear with a standard assembly program at speed code 18, 6–10 kb fragments were extracted from a 0.5% GS-FLX library adaptors were ligated to the repaired/A-tailed fragments. GS-FLX sequencing using the titanium chemistry was performed according to manufacturer's instructions (454 Life Sciences/Roche, Branford, CT). Sanger sequencing method was used to obtain BAC sequencing data as previously described (Hamberger *et al.*, 2011; Keeling *et al.*, 2010).

2.3 MiSeq modification

In sequencing the spruce genome, we generated longer read data by modifying the MiSeq platform. The MiSeq uses a clamshell style cartridge (Supplementary Fig. S1A) to hold reagent tubes in an array accessed by the MiSeq's sippers. Most of the reagents are required for length independent steps such as denaturation and cluster generation. The three reagents, the Scan, Cleavage and Incorporation mix, are consumed at each cycle. Although the MiSeq allows any read length specified in the control software, the reagent cartridge cannot be changed during the run without stopping it. Increasing the read length, however, requires increasing the quantity of the length-dependent reagents in the cartridge. This led to the solution of combining the length-dependent reagents of two kits into one.

A tool was designed that opens the snap-hook latches of the MiSeq cartridge together (Supplementary Figs S1B and S2), giving access to the reagent tubes, yet allowing the cartridge to be put back on the sipper without damage to its components (Supplementary Fig. S1C). The 40 ml, the stock length-dependent reagent containers allow a maximum of ~650 cycles in total. To maximize the potential of the MiSeq kit approach, a new reagent tray with 70 ml wells was designed and placed in a modified clamshell base.

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A Taylor[1], Macaire Man Saint Yuen[4], Christopher I Keeling[4], Dana Brand[1], Benjamin P Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A Moore[1], Yongjun Zhao[1], Andrew J Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Carol Ritland[4,6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6], Steven JM Jones[1,2,9]

[1] British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, BC V5Z 4S6
[2] University of British Columbia, Department of Medical Genetics, Vancouver, BC V6H 3N1
[3] Simon Fraser University, School of Computing Science, Burnaby, BC V5A 1S6
[4] University of British Columbia, Michael Smith Laboratories, Vancouver, BC V6T 1Z4
[5] British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2
[6] University of British Columbia, Department of Forest Sciences, Vancouver, BC V6T 1Z4
[7] Université Laval, Institute for Systems and Integrative Biology, Québec, QC G1V 0A6
[8] Université Laval, Department of Wood and Forest Sciences, Québec, QC G1V 0A6
[9] Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC V5A 1S6

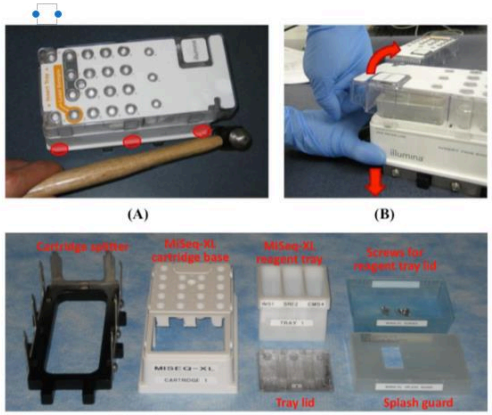


Figure S1. Modification of the MiSeq cartridge. MiSeq reagent cartridge was modified to allow for longer read lengths. **(A, B)** Opening of the clamshell style cartridge. **(C)** Contents of the modified cartridge. This was initially used to combine two PE150 kits for PE300 runs. When Illumina introduced the P250 kit, the same apparatus was used to enable PE500 runs.

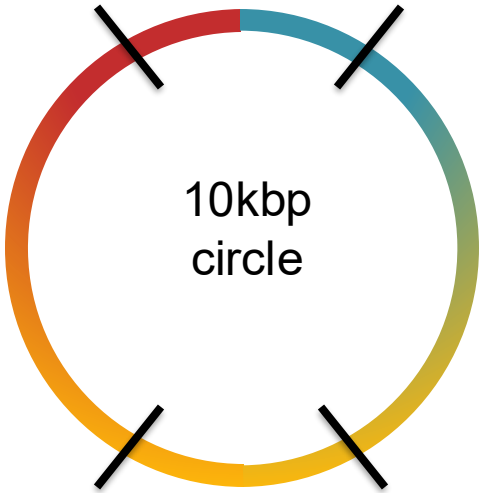
Paired-end and Mate-pairs

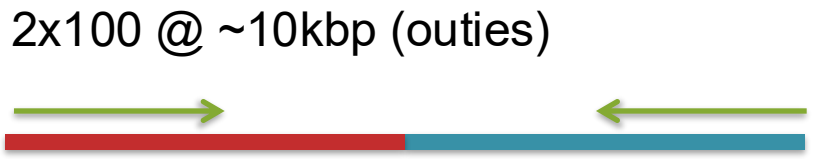
Paired-end sequencing

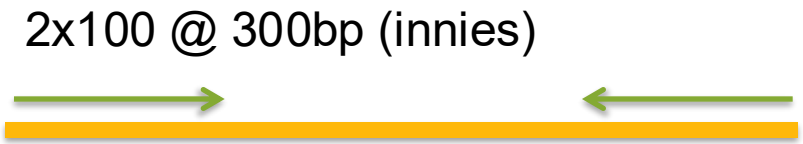
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads

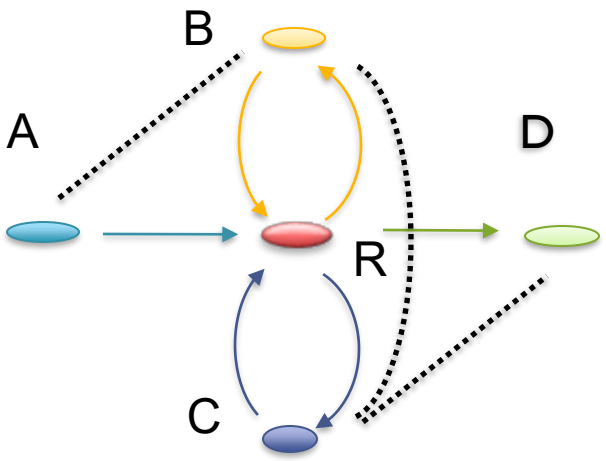




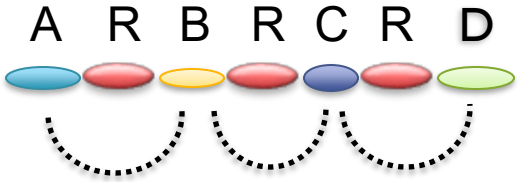


Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries



- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled



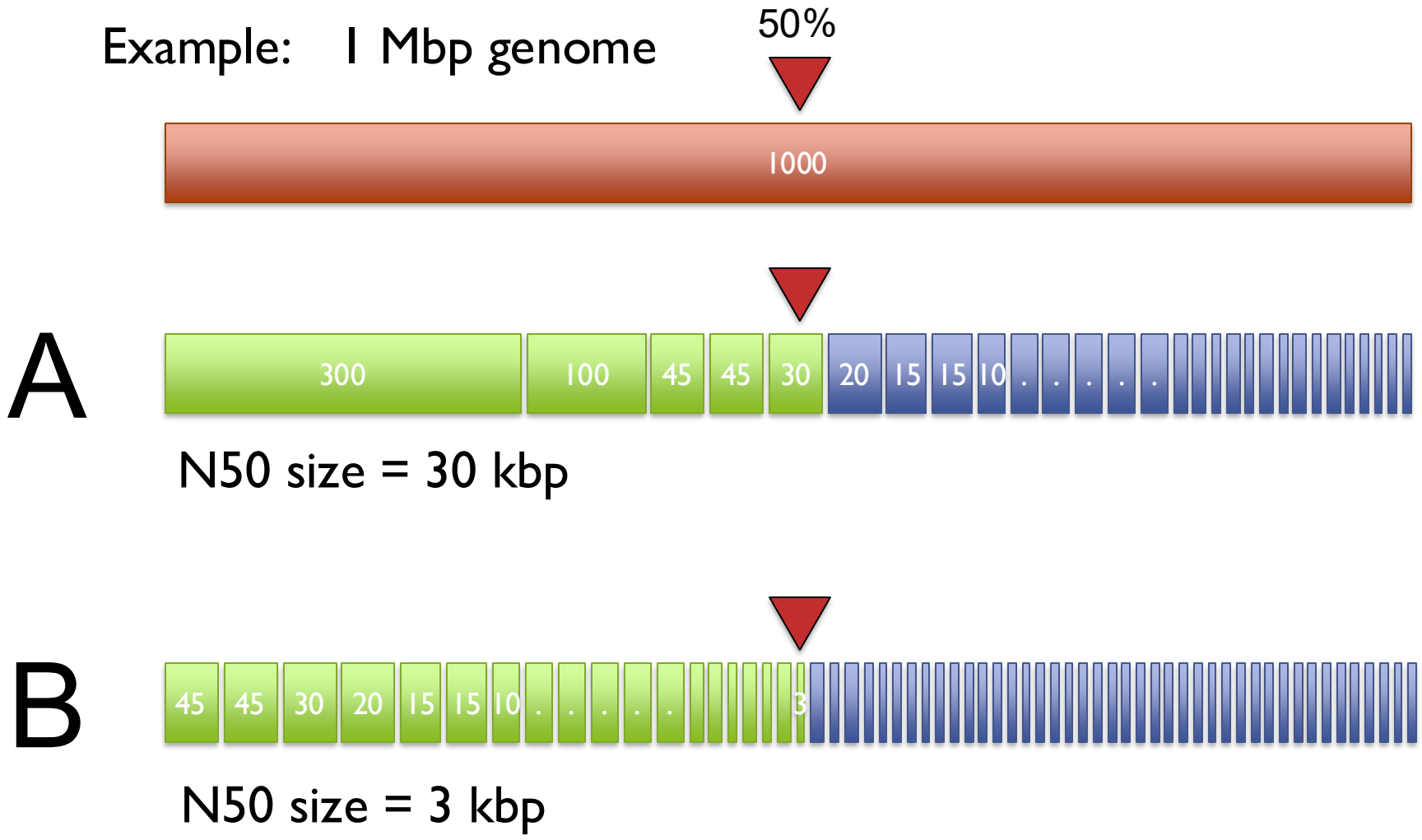
- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

Why do scaffolds end?

Contig/Scaffold N50

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



Contig/Scaffold N50

Def: 50% of the genome is in contigs as large as the N50 value

Better N50s improves the analysis in every dimension

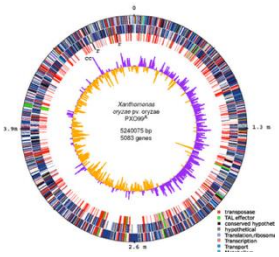
- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

Just be careful of N50 inflation!

- *A very very very bad assembler in 1 line of bash:*
- *cat *.reads.fa > genome.fa*

N50 size = 3 kbp

Assembly Summary



Assembly quality depends on

1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set

- Recommend spades for short read assembl
  - Integrates error correction and scaffolding