Genome Assembly

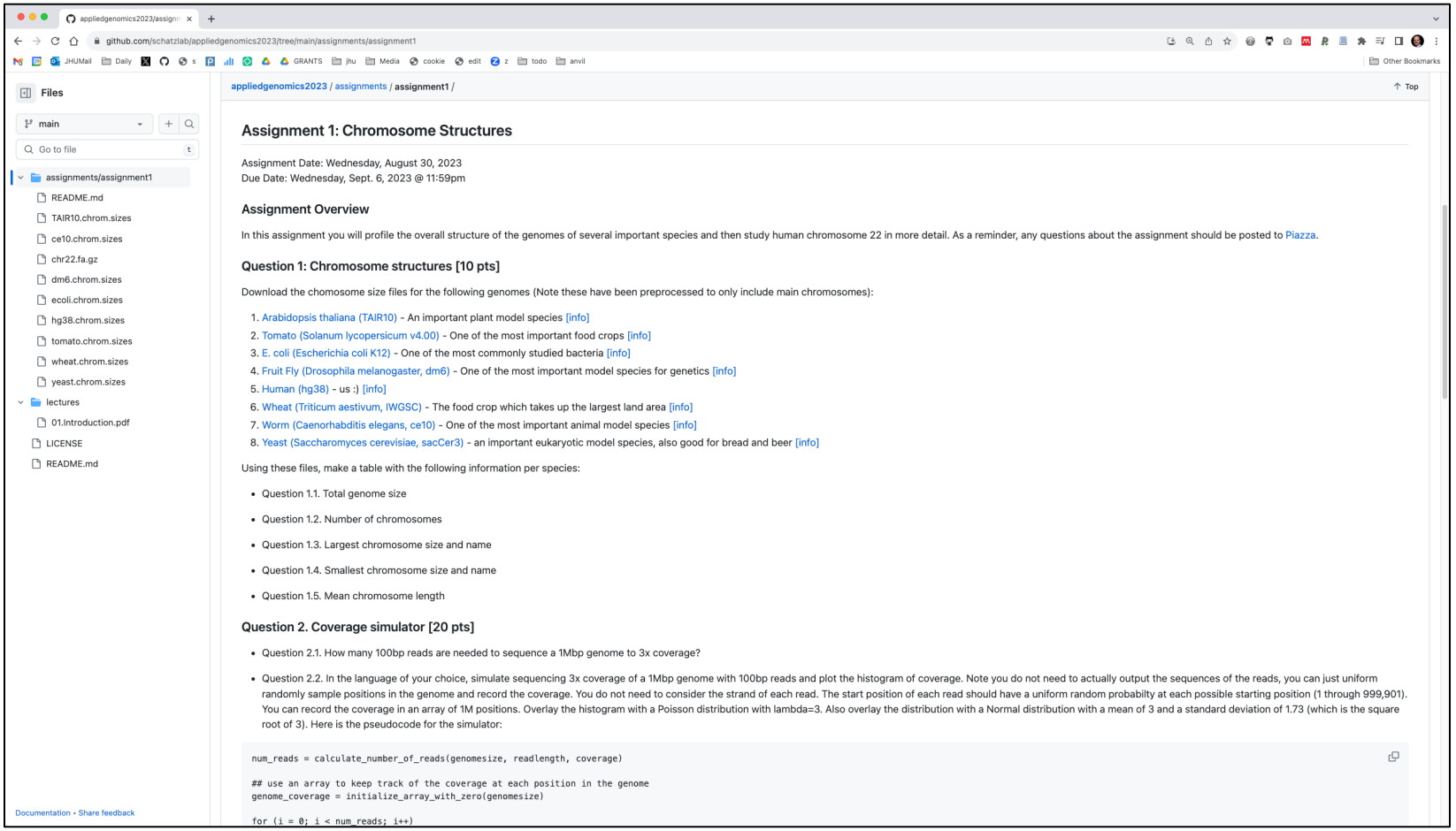
Michael Schatz

Sept 6, 2023

Lecture 3: Applied Comparative Genomics

Assignment 1

Due end of day on Sept 6 (right before midnight)



appliedgenomics2023/assignm

github.com/schatzlab/appliedgenomics2023/tree/main/assignments/assignment1

JHUMail Daily s GRANTS jhu Media cookie edit z todo anvil Other Bookmarks

Files

main

Go to file t

assignments/assignment1

README.md
TAIR10.chrom.sizes
ce10.chrom.sizes
chr22.fa.gz
dm6.chrom.sizes
ecoli.chrom.sizes
hg38.chrom.sizes
tomato.chrom.sizes
wheat.chrom.sizes
yeast.chrom.sizes

lectures

01.Introduction.pdf

LICENSE
README.md

Documentation · Share feedback

appliedgenomics2023 / assignments / assignment1 /

Top

Assignment 1: Chromosome Structures

Assignment Date: Wednesday, August 30, 2023
Due Date: Wednesday, Sept. 6, 2023 @ 11:59pm

Assignment Overview

In this assignment you will profile the overall structure of the genomes of several important species and then study human chromosome 22 in more detail. As a reminder, any questions about the assignment should be posted to Piazza.

Question 1: Chromosome structures [10 pts]

Download the chomosome size files for the following genomes (Note these have been preprocessed to only include main chromosomes):

1. Arabidopsis thaliana (TAIR10) - An important plant model species [info]
2. Tomato (Solanum lycopersicum v4.00) - One of the most important food crops [info]
3. E. coli (Escherichia coli K12) - One of the most commonly studied bacteria [info]
4. Fruit Fly (Drosophila melanogaster, dm6) - One of the most important model species for genetics [info]
5. Human (hg38) - us :) [info]
6. Wheat (Triticum aestivum, IWGSC) - The food crop which takes up the largest land area [info]
7. Worm (Caenorhabditis elegans, ce10) - One of the most important animal model species [info]
8. Yeast (Saccharomyces cerevisiae, sacCer3) - an important eukaryotic model species, also good for bread and beer [info]

Using these files, make a table with the following information per species:

- Question 1.1. Total genome size
- Question 1.2. Number of chromosomes
- Question 1.3. Largest chromosome size and name
- Question 1.4. Smallest chromosome size and name
- Question 1.5. Mean chromosome length

Question 2. Coverage simulator [20 pts]

- Question 2.1. How many 100bp reads are needed to sequence a 1Mbp genome to 3x coverage?
- Question 2.2. In the language of your choice, simulate sequencing 3x coverage of a 1Mbp genome with 100bp reads and plot the histogram of coverage. Note you do not need to actually output the sequences of the reads, you can just uniform randomly sample positions in the genome and record the coverage. You do not need to consider the strand of each read. The start position of each read should have a uniform random probabilty at each possible starting position (1 through 999,901). You can record the coverage in an array of 1M positions. Overlay the histogram with a Poisson distribution with lambda=3. Also overlay the distribution with a Normal distribution with a mean of 3 and a standard deviation of 1.73 (which is the square root of 3). Here is the pseudocode for the simulator:

```
num_reads = calculate_number_of_reads(genomesize, readlength, coverage)

## use an array to keep track of the coverage at each position in the genome
genome_coverage = initialize_array_with_zero(genomesize)

for (i = 0; i < num_reads; i++)
```

https://github.com/schatzlab/appliedgenomics2023/tree/main/assignments/assignment1

On ChatGPT…

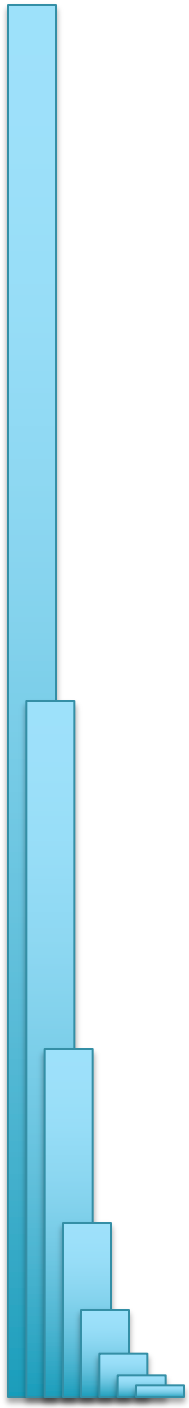
ChatGPT 3.5 output:

… Notably, Professor Michael Schatz, a pioneer in genomics research at Johns Hopkins University, has made significant contributions to the development of high-throughput computational tools and techniques for genome assembly and analysis (Schatz et al., 2019). These advances are crucial for microbiologists working with Acetobacter to harness the wealth of genomic data available, advancing our understanding of their evolution and potential biotechnological applications….

References:

…

4. Schatz, M. C., Maron, L. G., & Stein, J. C. (2019). Assembly and annotation of genomes from long-read sequencing. Current Opinion in Plant Biology, 48, 69-75.

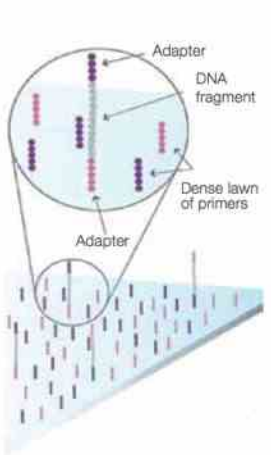


Part 1: Recap and Illumina Sequencing

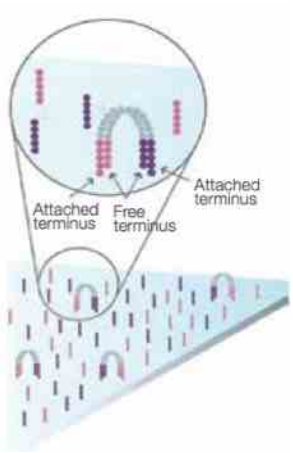
Second Generation Sequencing

Illumina NovaSeq 6000
Sequencing by Synthesis

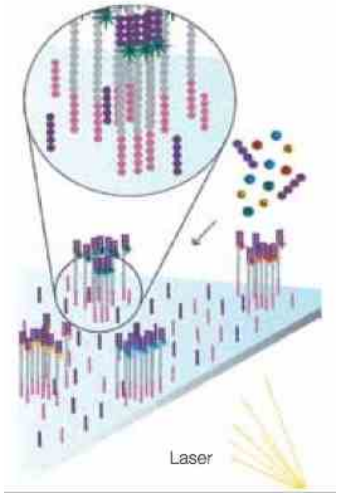
>3Tbp / day
(JHU has 4 of these!)



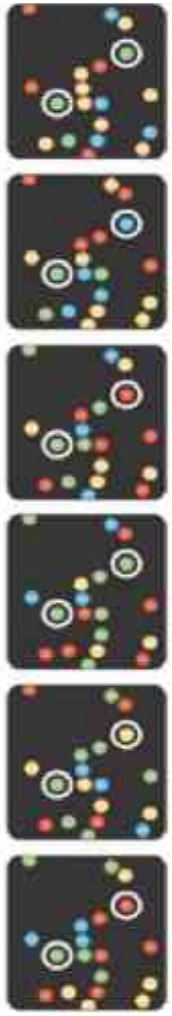
1. Attach



2. Amplify



3. Image

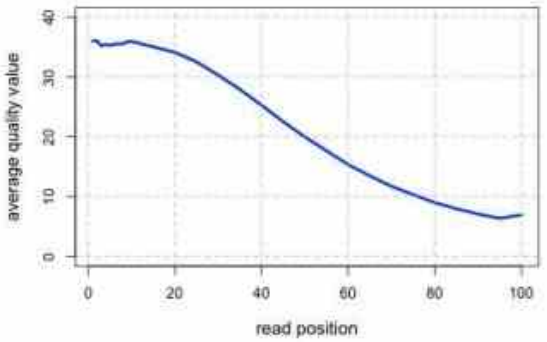


Metzker (2010) Nature Reviews Genetics 11:31-46
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Illumina Quality

QV	p_{error}
40	1/10000
30	1/1000
20	1/100
10	1/10

$$Q_{\text{sanger}} = -10 \log_{10} p$$

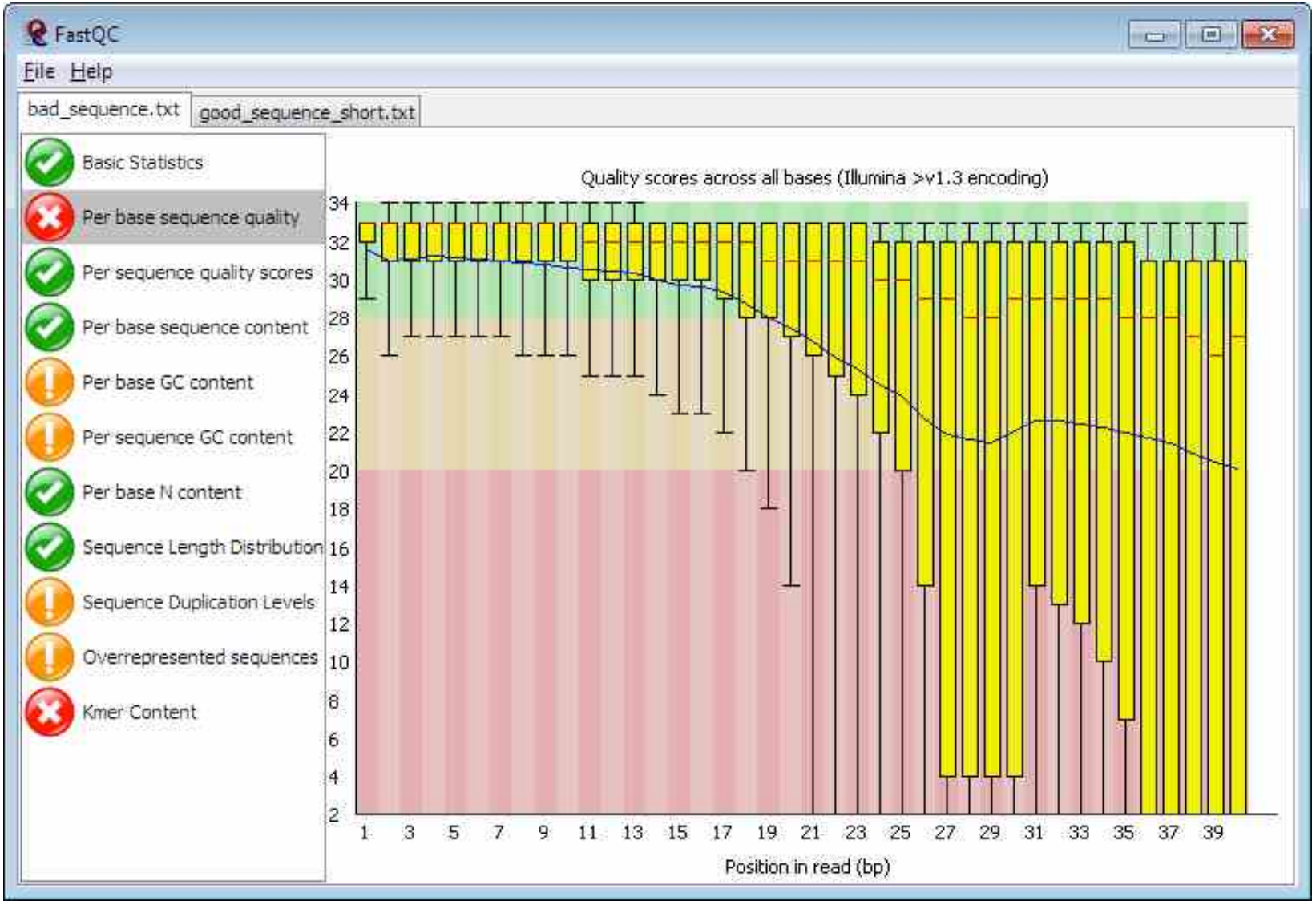



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

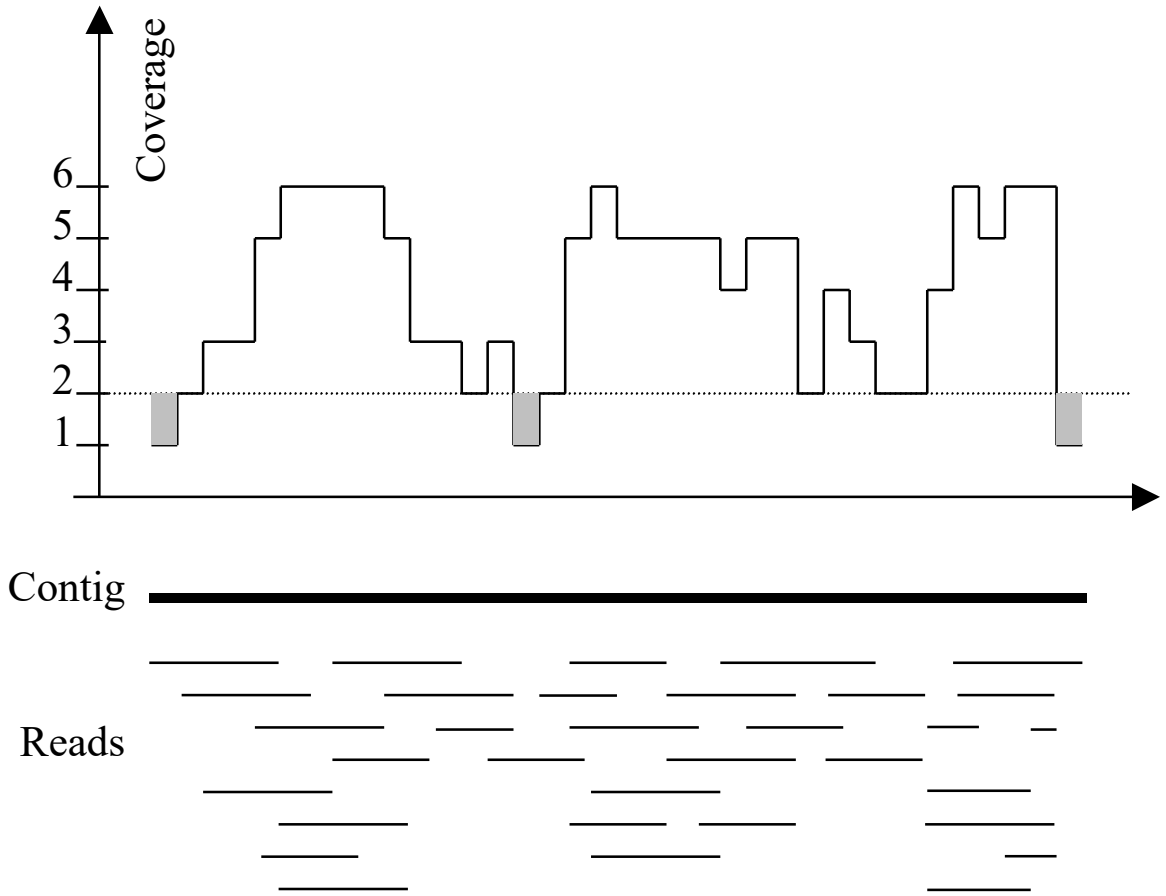
http://en.wikipedia.org/wiki/FASTQ_format

FASTQC: Are my data any good?



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Typical sequencing coverage



Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs $1

If the genome is 10 Mbp, should we sequence 100k 100bp reads?

Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

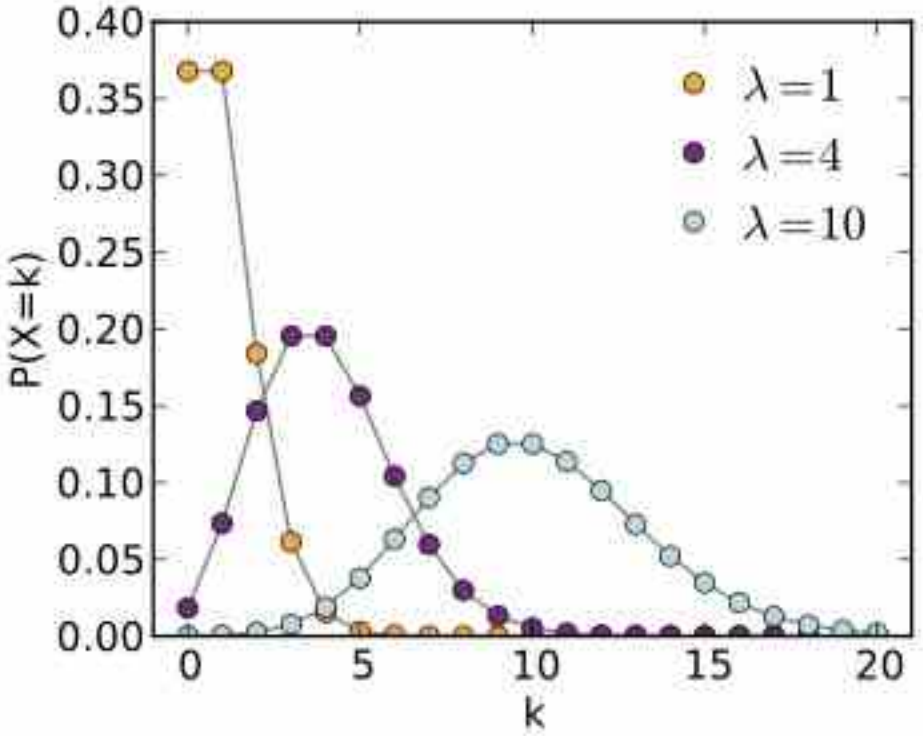
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

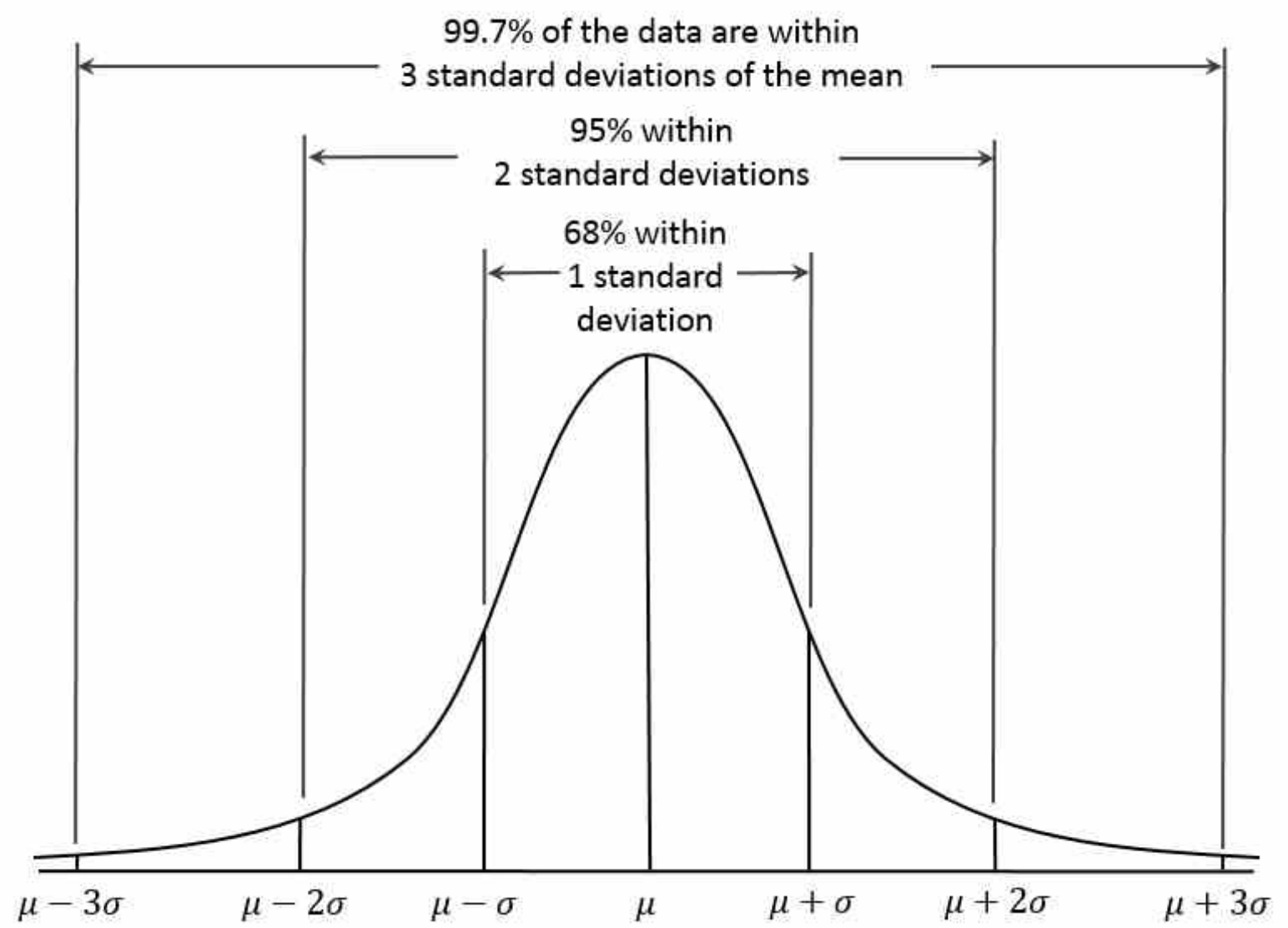
Resembles a normal distribution, but over the positive values, and with only a single parameter.

Key properties:

- ***The standard deviation is the square root of the mean.***
- ***For mean > 5, well approximated by a normal distribution***



Normal Approximation



Can estimate Poisson distribution as a normal distribution when $\lambda > 10$

K-mers and K-mer counting

GATTACATACACATTGGATG

GAT ACA ACA ATT GAT

ATT CAT CAC TTG ATG

TTA ATA ACA TGG

TAC TAC CAT GGA

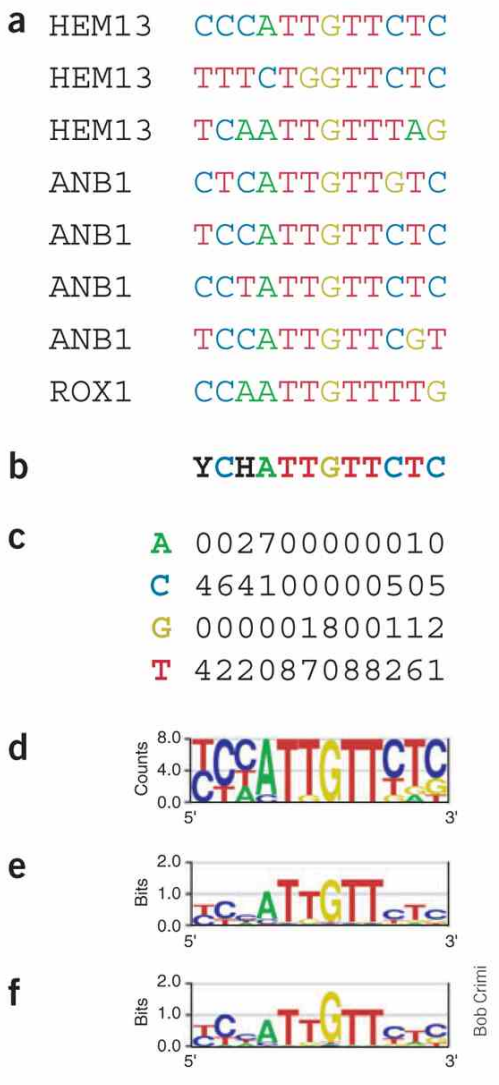
Kmers:
- Divide a string into substrings of length k
- Notice every position is covered k times
- Notice there are G – k + 1 kmers from a string of length G

Computation: Very easy to compute, exact matches, represent 32mers in 64 bits
Biological: The “atomic unit” of a sequence, creates a fingerprint of a genome/read

Transcription Factors

A transcription factor (or sequence-specific DNA-binding factor) is a protein that controls the rate of transcription of genetic information from DNA to messenger RNA, by binding to a specific DNA sequence.

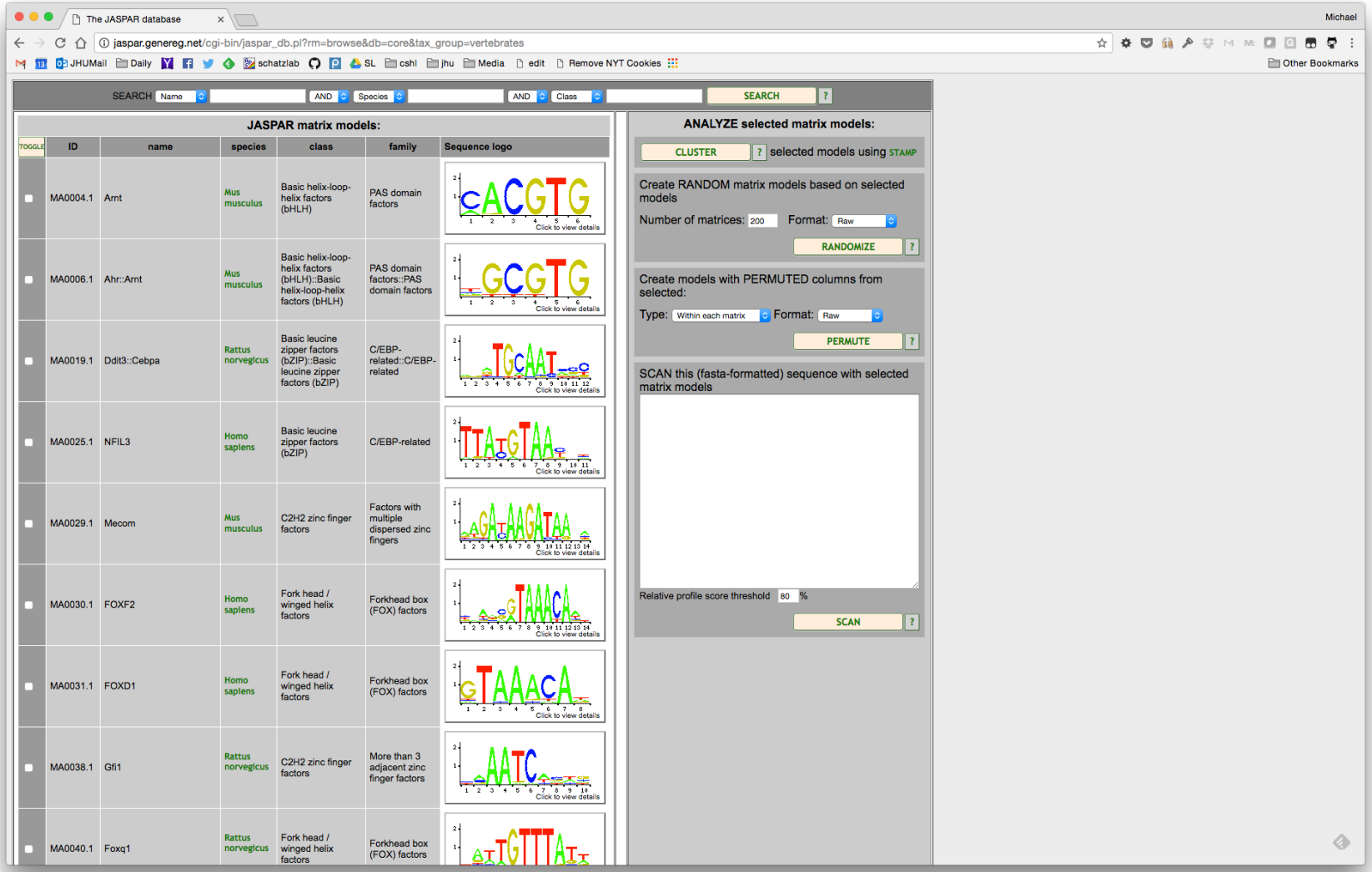
- Transcription factors work alone or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase to specific genes.
- A defining feature of transcription factors is that they contain at least one DNA-binding domain (DBD)
- Figure (a) Eight known genomic binding sites in three S. cerevisiae genes. (b) Degenerate consensus sequence. (c,d) Frequencies of nucleotides at each position. (e) Sequence logo (f) Energy normalized logo using relative entropy to adjust for low GC content in S. cerevisiae.



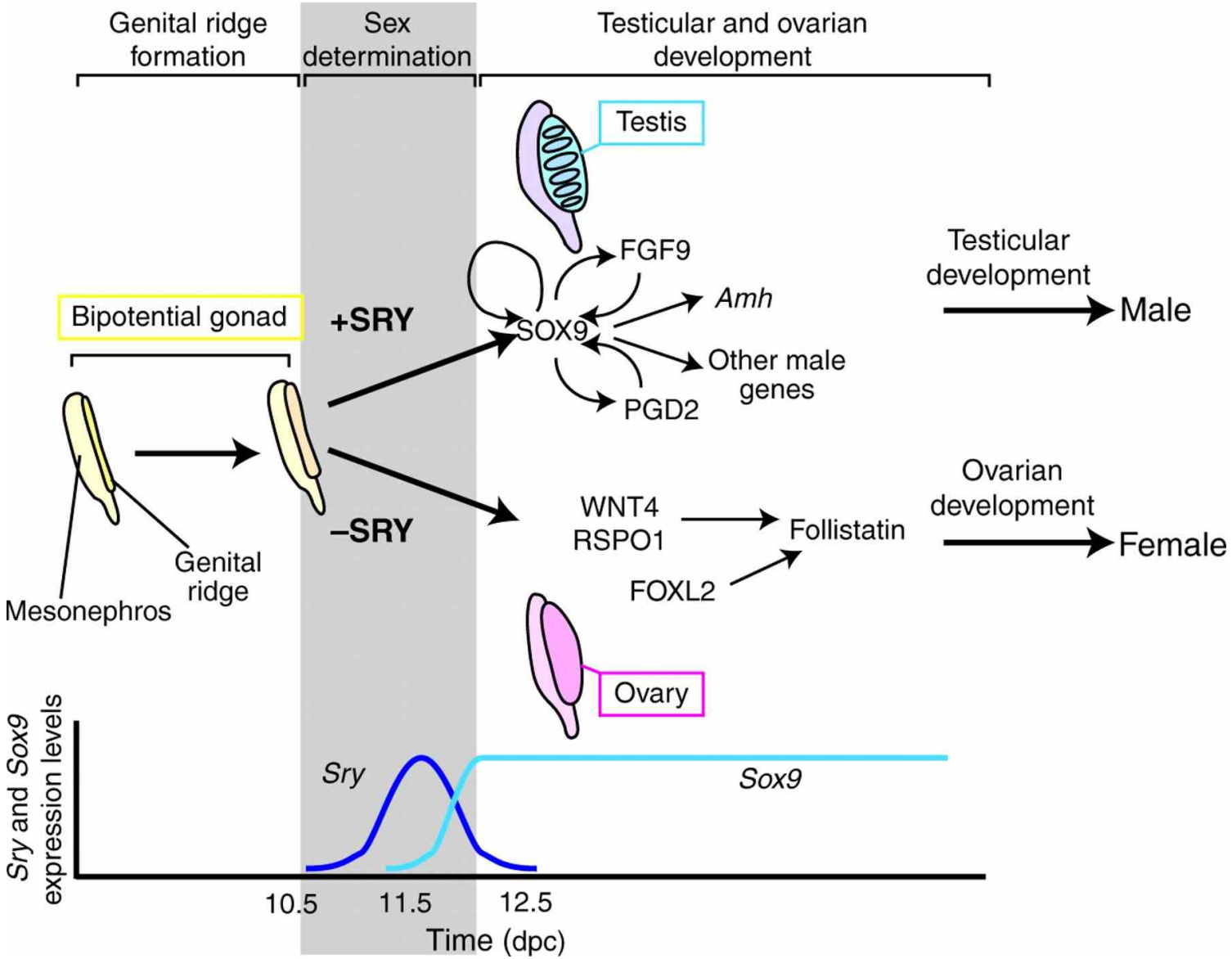
What are DNA sequence motifs?

D'haeseleer (2006) Nature Biotechnology 24, 423 – 425 doi:10.1038/nbt0406-423

Transcription Factors Database

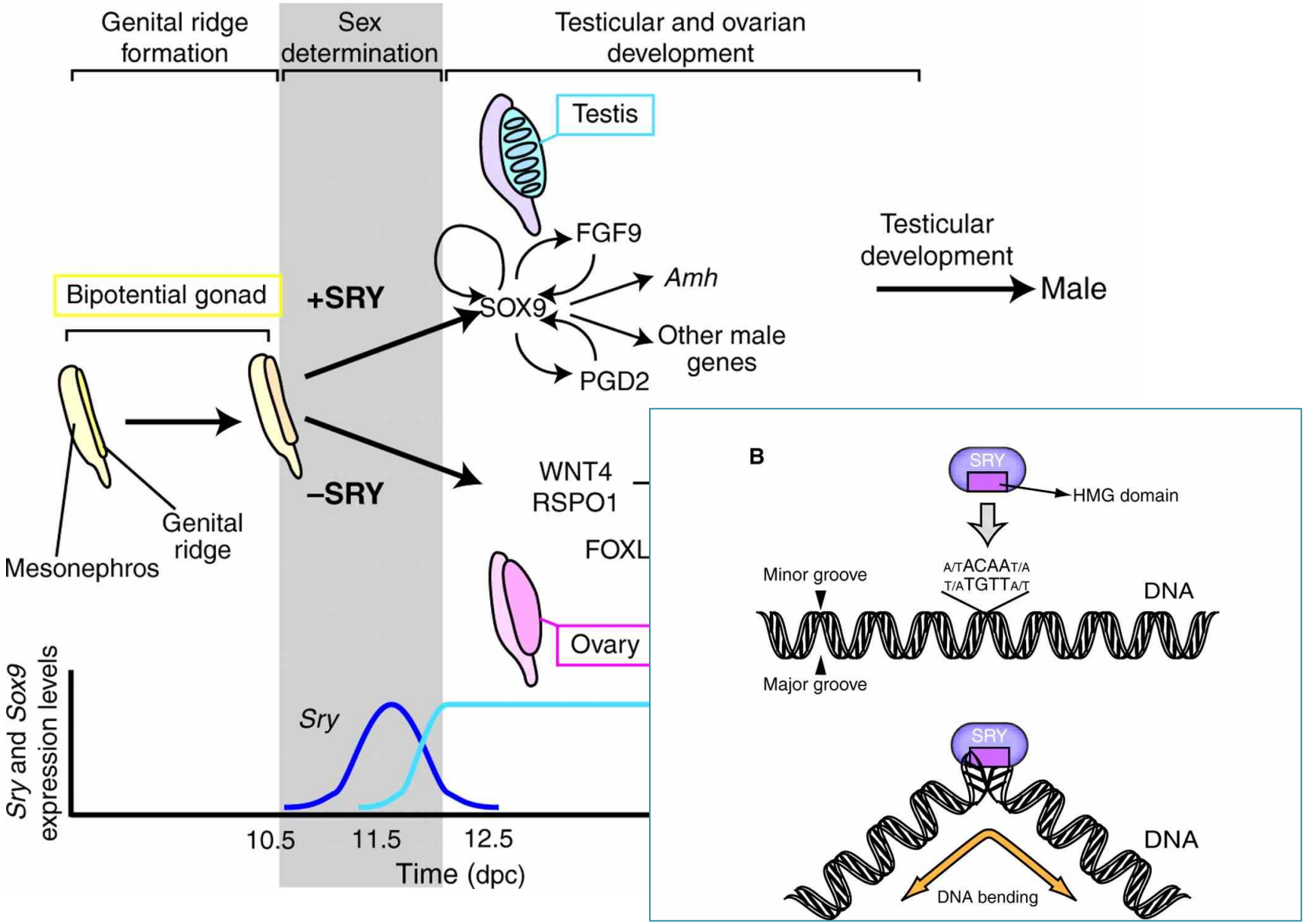


JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles
Anthony Mathelier (2014) Nucleic Acids Res. 42 (D1): D142-D147. DOI: https://doi.org/10.1093/nar/gkt997



SRY: The master switch in mammalian sex determination

Kashimada and Koopman (2010) Development 137: 3921-3930; doi: 10.1242/dev.048983



SRY: The master switch in mammalian sex determination

Kashimada and Koopman (2010) Development 137: 3921-3930; doi: 10.1242/dev.048983

K-mers and K-mer counting

GATTACATACACATTGGATG

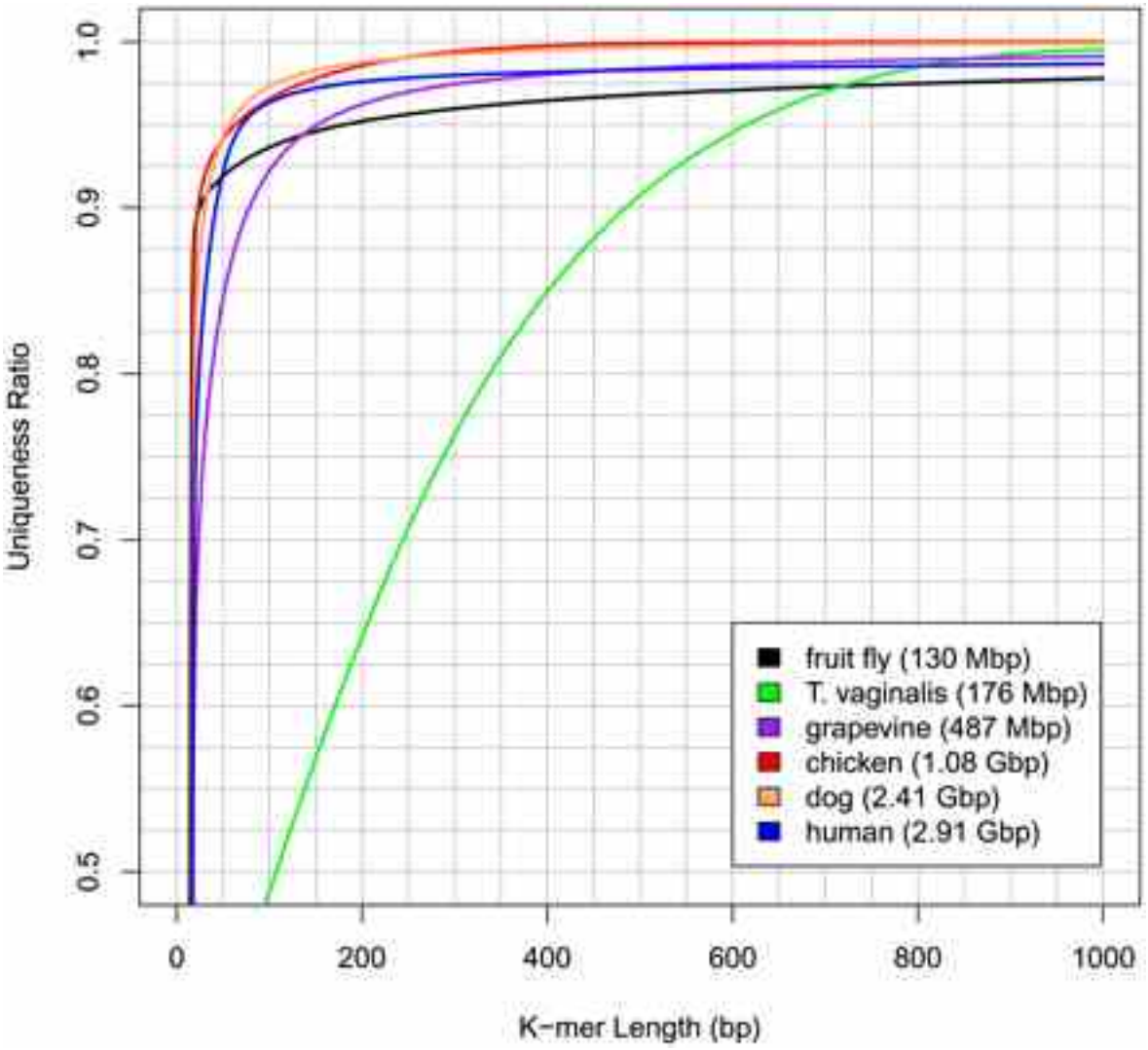
```
1:  7  (ATG, TGG, …)
2:  4  (GAT, CAT, ATT, TAC)
3:  1  (ACA)
```

How long should k be?

K=1 : Too short, every base is present
K=2 : Too short, every pair of bases will be present

Pick k so that G/(4^k) << 1
k = log_4 (G)
At least 15 for human, often a bit longer
But not too long or could loose resolution

K-mer Uniqueness



Assembly of large genomes using second-generation sequencing
Schatz et al. (2010) Genome Research. doi: 10.1101/gr.101360.109

Question?

We would love to generate
longer and longer reads with this technology

What can we do?

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#: ALWZ01000000000 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 b (Simpson *et al.*, 2009) using human and was later used to assemble the SOAPdenovo tool (Li *et al.*, 2010). high quality results, as demonstrate *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 uni generation end, those challenges inclu whole-genome shotgun sequencing da ing reduced representation resources tude of the problem. On the bioin massive sequencing datasets is extrem ing cycles, memory usage, storage req programming implementations on co

We addressed the data representati and sequencing multiple whole-genon 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 sequencin

In this work, we demonstrate that t 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 (Roche 454 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 circularized 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 Inc., Hayward, 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 454 Titanium with 6 kb paired-end libraries at the PlateForm 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 BAC DNA was fragmented using a Hydroshear with a standard assembly tube (Digilab) at speed code 18, 6–10 kb fragments were extracted from the gel, and the 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 the BAC sequencing data as previously described (Hamberger and Bohlmann, 2006; 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 reagent cartridge together (Supplementary Figs S1B and S2), giving access to the reagent tubes, yet allowing the cartridge to be put back together 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

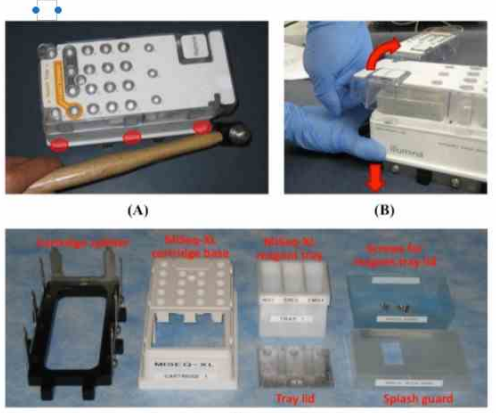


Figure S1. Modification of the MiSeq cartridge. MiSeq reagent cartridge was modified to allow for longer read lengths. **(A, B)** Opening of the clamshell style cartridge. **(C)** Contents of the modified cartridge. This was initially used to combine two PE150 kits for PE300 runs. When Illumina introduced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

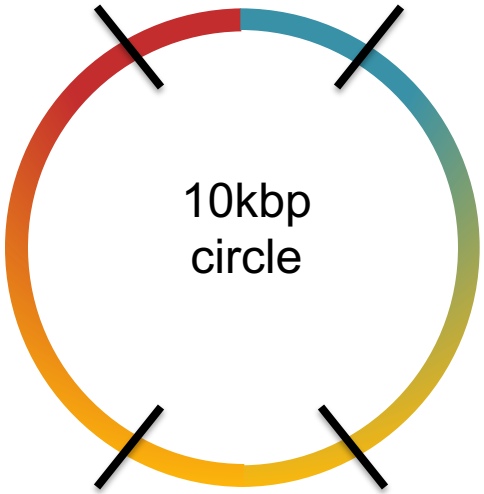
Paired-end and Mate-pairs

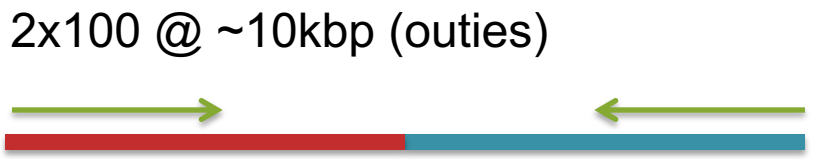
Paired-end sequencing

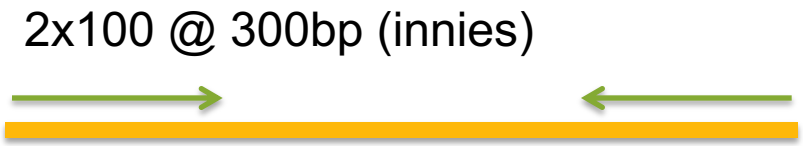
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

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







FASTQ Files

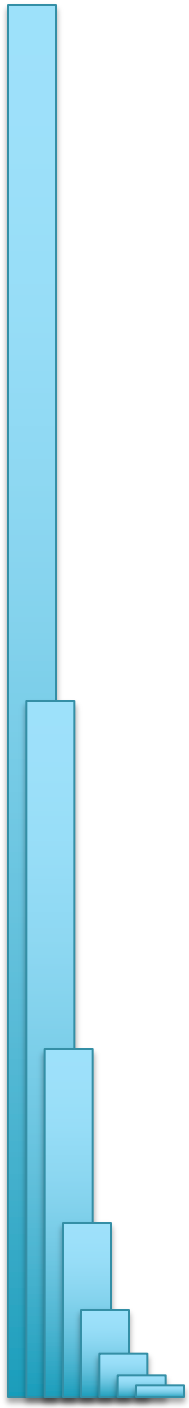
reads.1.fq

reads.2.fq

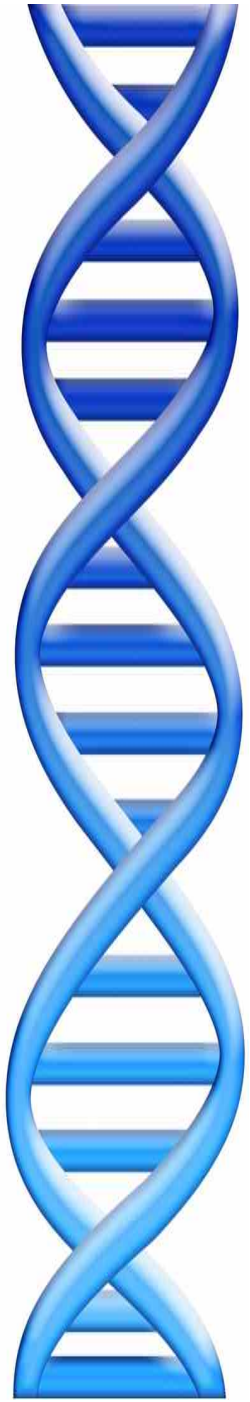
```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

@Identifier
Sequence
+Separator
Quality Values
…

http://en.wikipedia.org/wiki/FASTQ_format



Part 2: De novo genome assembly



Outline

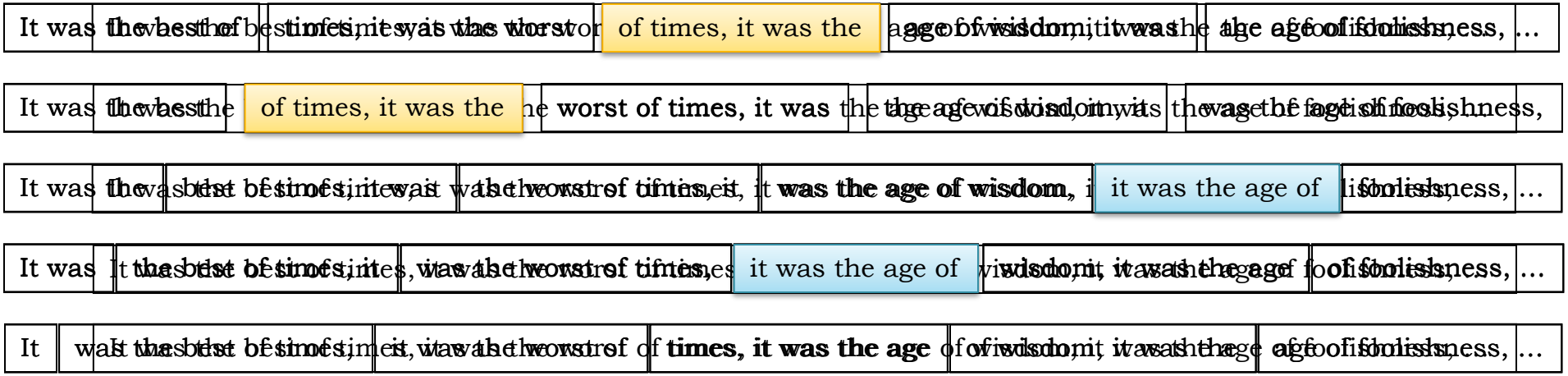
1. ***Assembly theory***
   - Assembly by analogy

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

3. ***Whole Genome Alignment***
   - MUMmer recommended

Shredded Book Reconstruction

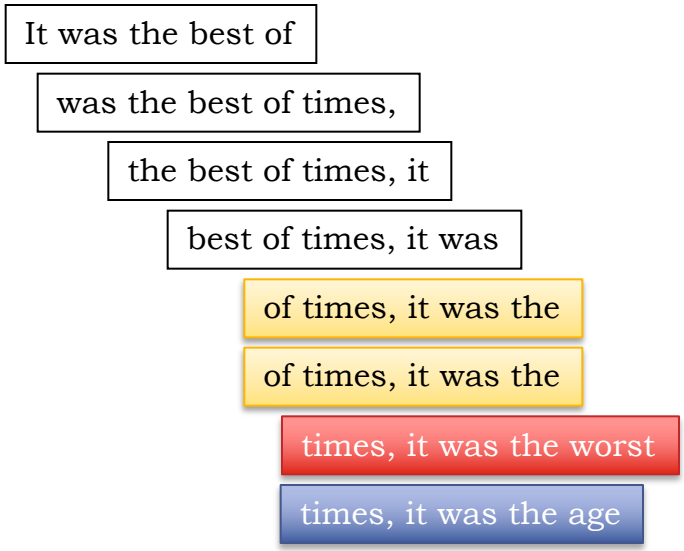
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was



The repeated sequence make the correct reconstruction ambiguous

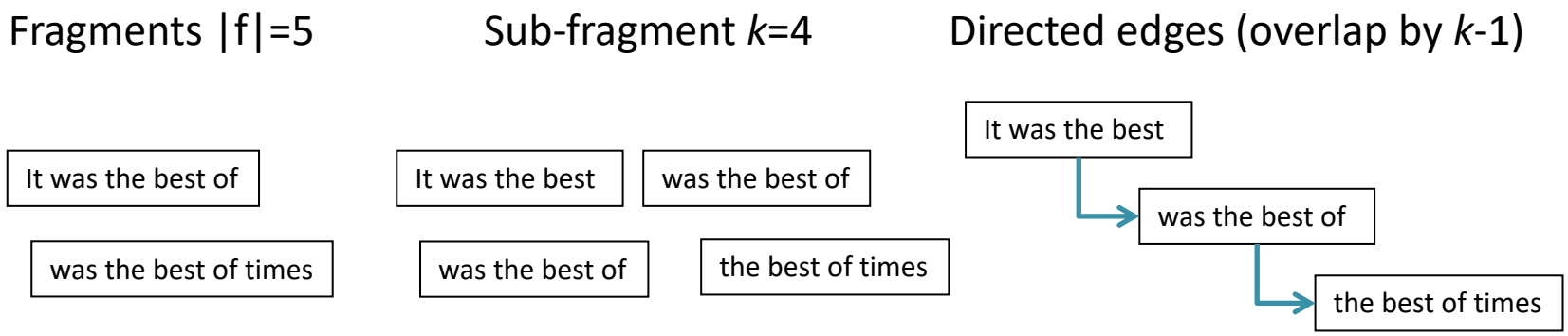
- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

How long will it take to compute the overlaps?

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

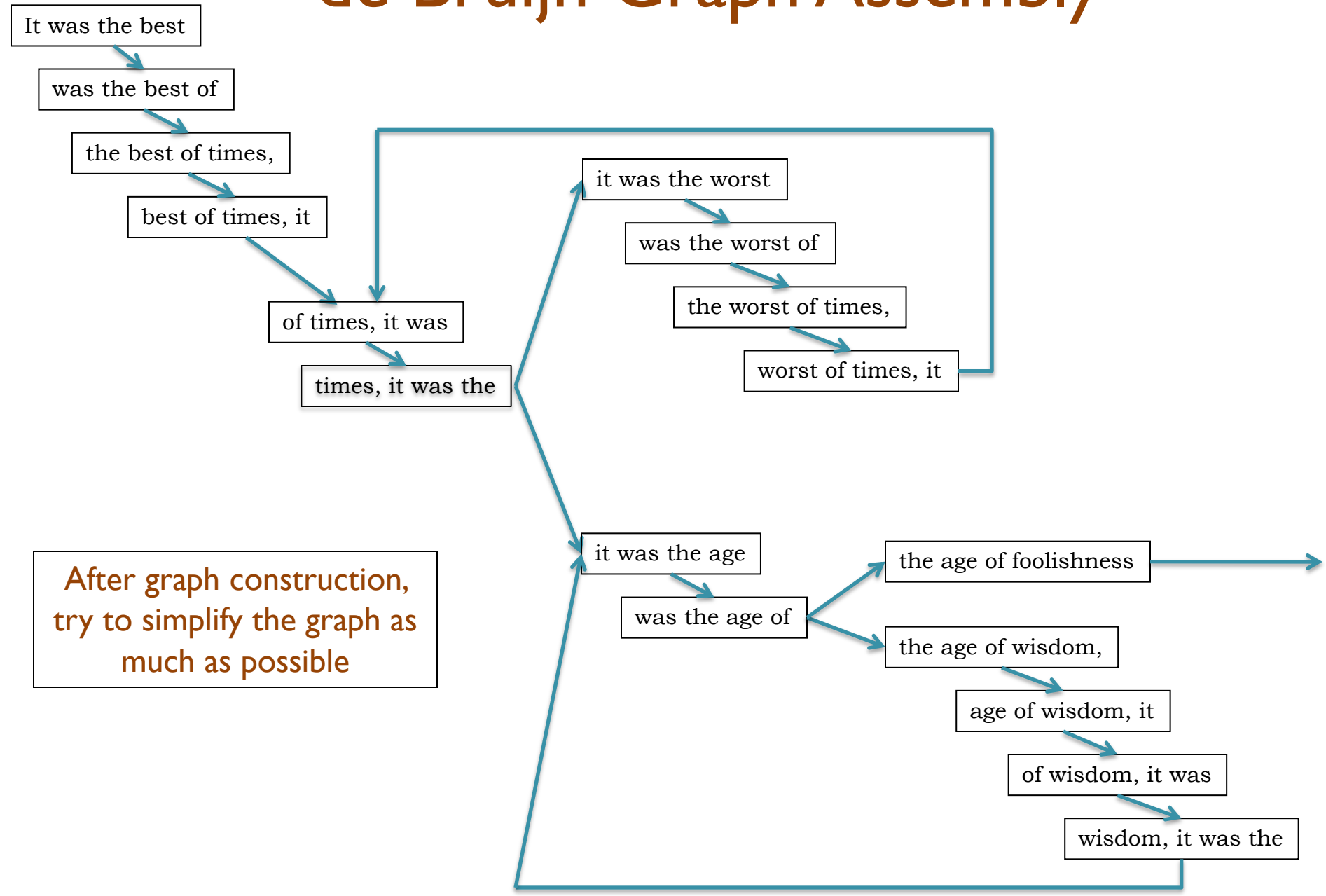


– Overlaps between fragments are implicitly computed

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

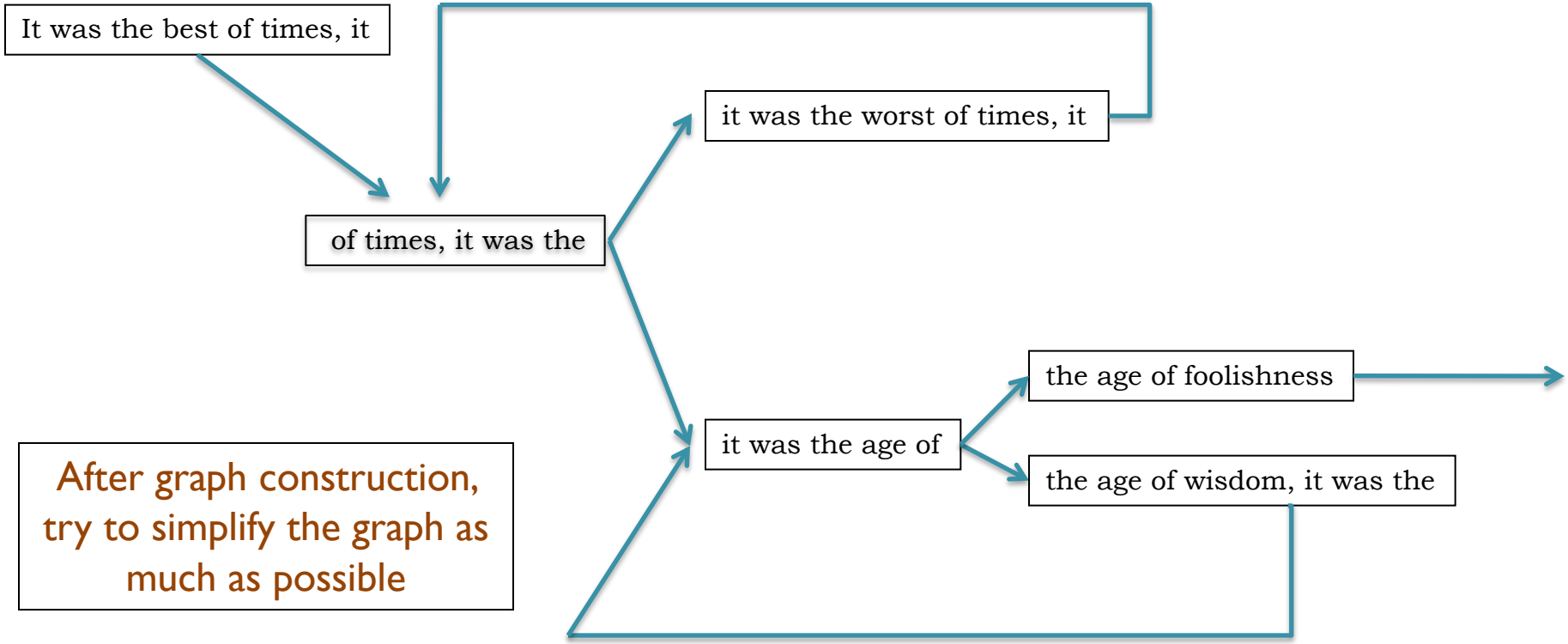
de Bruijn, 1946
Idury et al., 1995
Pevzner et al., 2001

de Bruijn Graph Assembly



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

de Bruijn Graph Assembly



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

The full tale

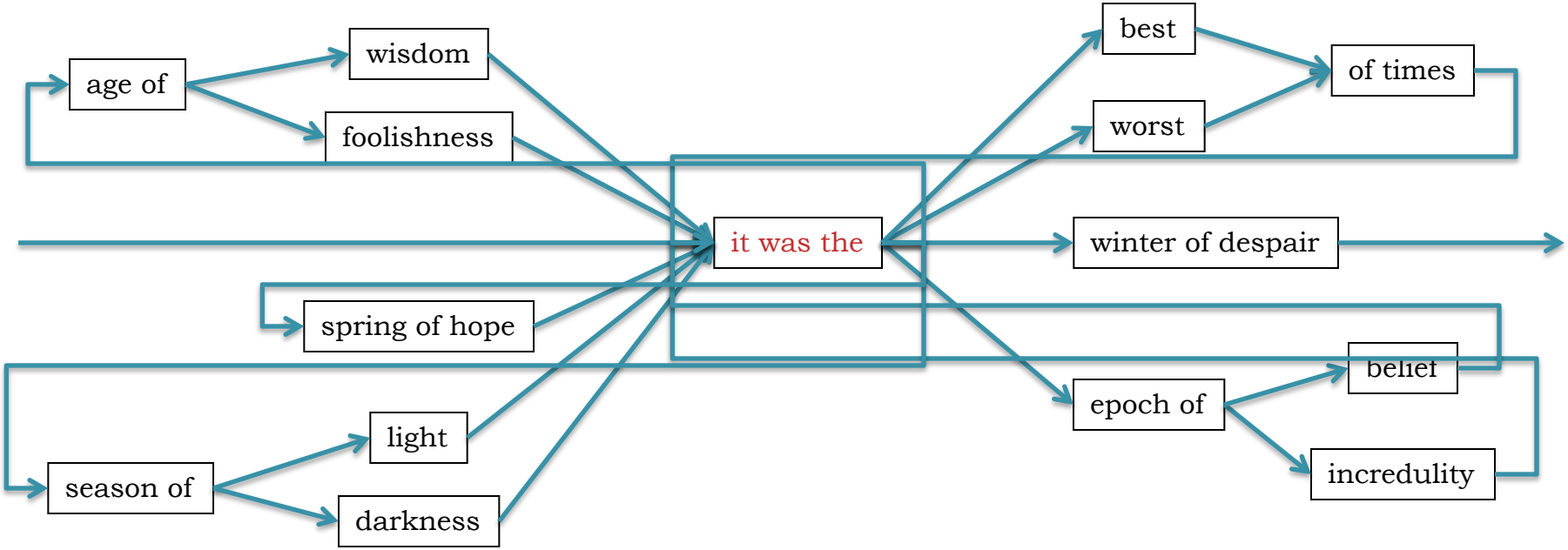
… it was the best of times it was the worst of times …

… it was the age of wisdom it was the age of foolishness …

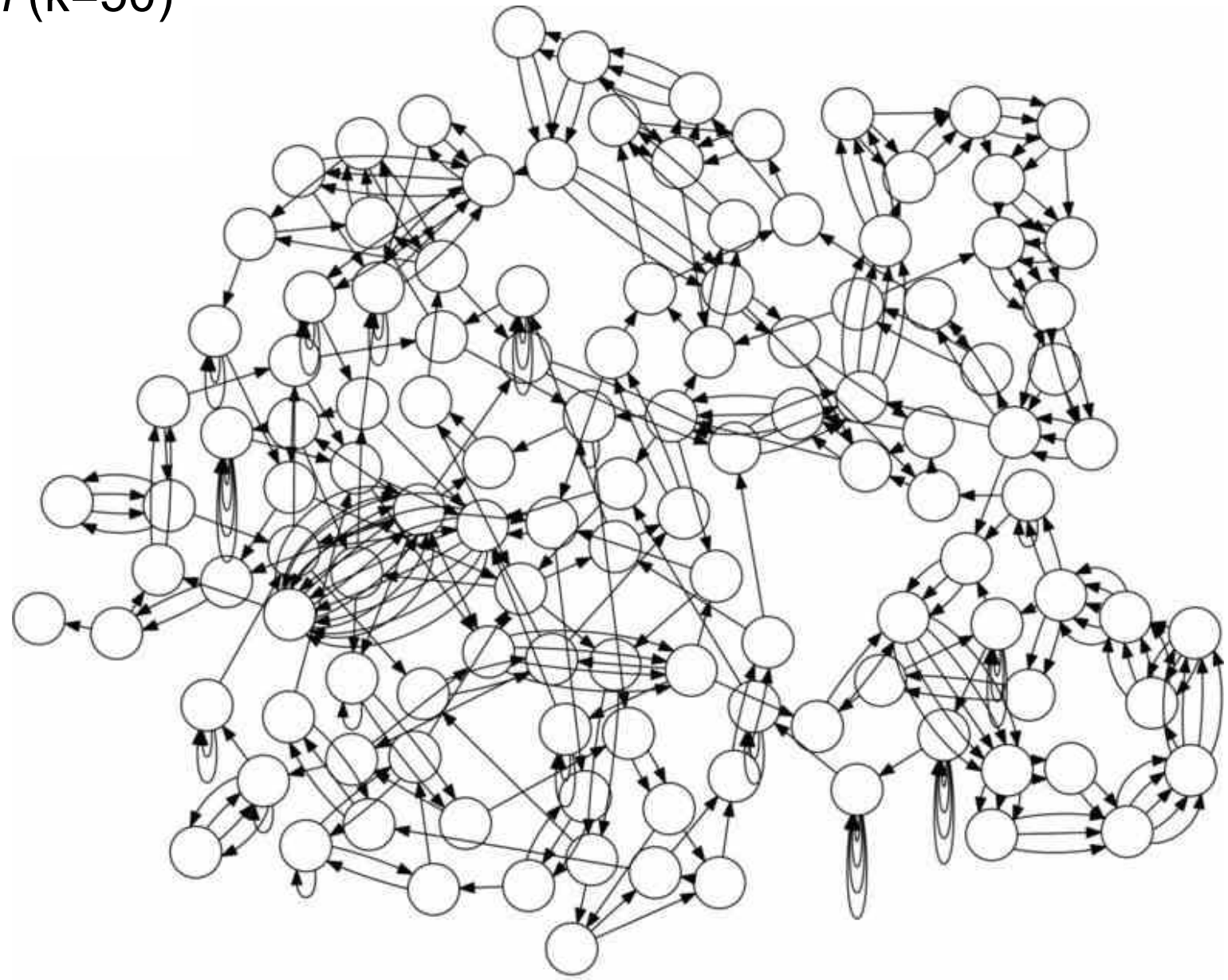
… it was the epoch of belief it was the epoch of incredulity …

… it was the season of light it was the season of darkness …

… it was the spring of hope it was the winder of despair …

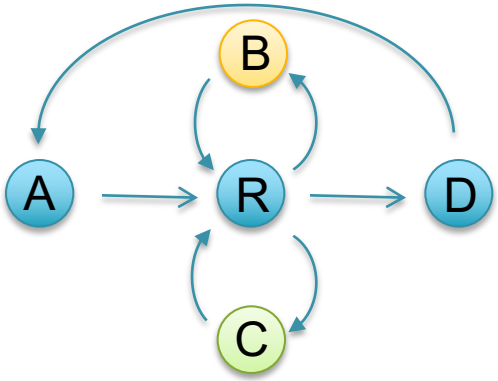


E. coli (k=50)



Reducing assembly complexity of microbial genomes with single-molecule sequencing
Koren et al (2013) Genome Biology. **14**:R101 https://doi.org/10.1186/gb-2013-14-9-r101

Counting Eulerian Cycles



ARBRCRD
or
ARCRBRD

Generally an exponential number of compatible sequences

– Value computed by application of the BEST theorem (Hutchinson, 1975)

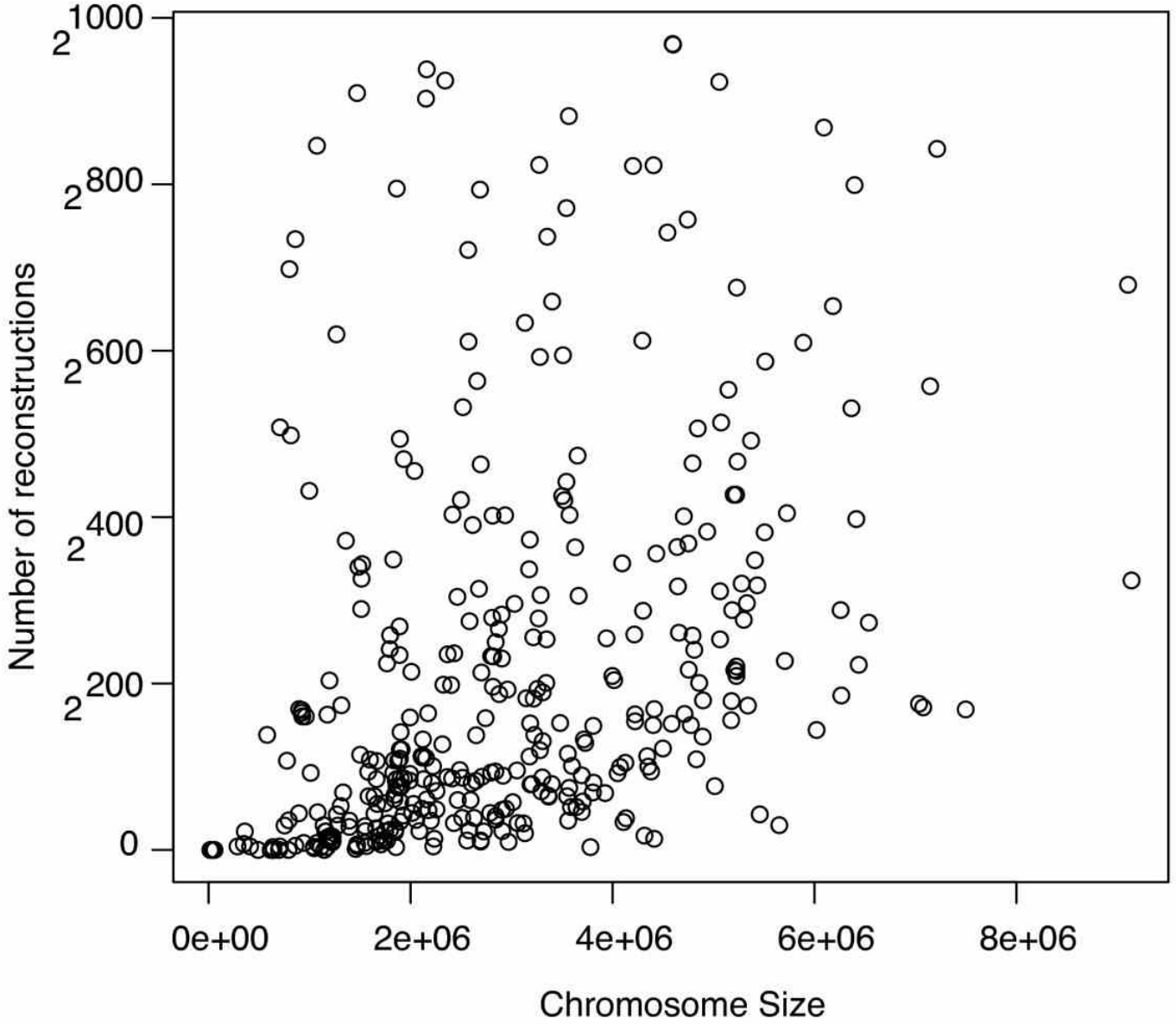
$$\mathcal{W}(G,t) = (\det L)\left\{\prod_{u\in V}(r_u-1)!\right\}\left\{\prod_{(u,v)\in E} a_{uv}!\right\}^{-1}$$

L = *n x n* matrix with r_u-a_{uu} along the diagonal and -a_{uv} in entry uv

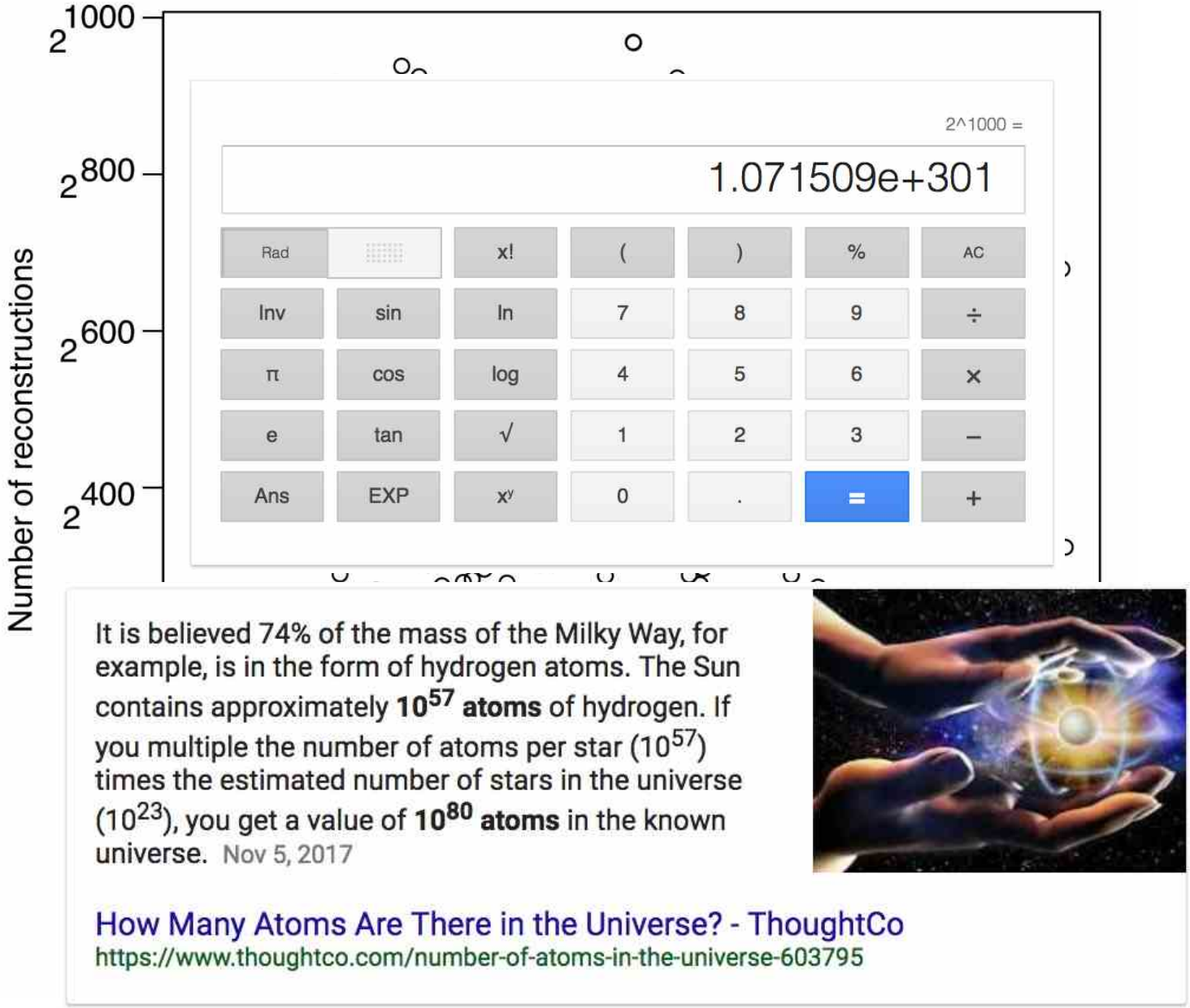
$r_u = d^+(u)+1$ if $u=t$, or $d^+(u)$ otherwise

a_{uv} = multiplicity of edge from *u* to *v*

Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.



Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.



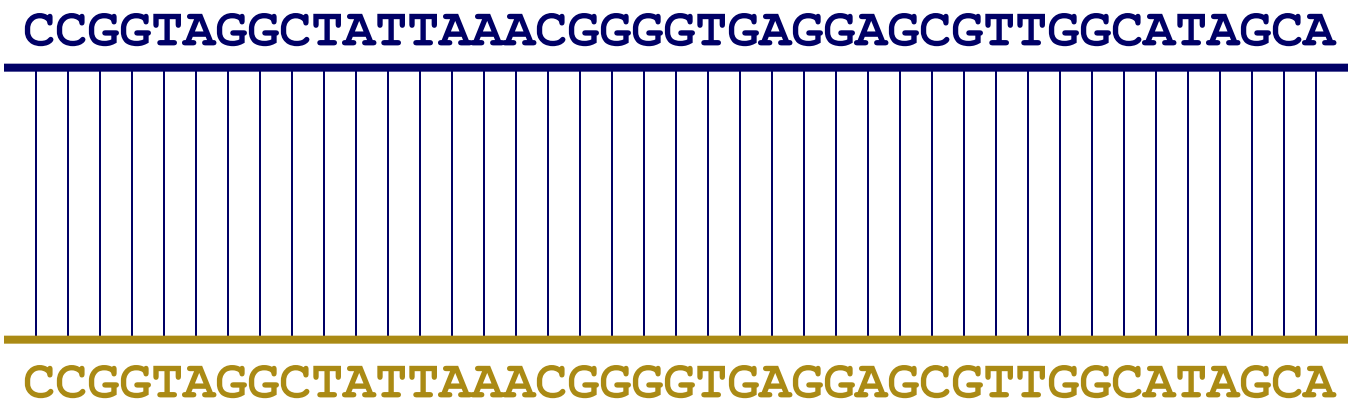
Assembly Complexity of Prokaryotic Genomes using Short Reads.
Kingsford C, Schatz MC, Pop M (2010) *BMC Bioinformatics*.

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
NHGRI

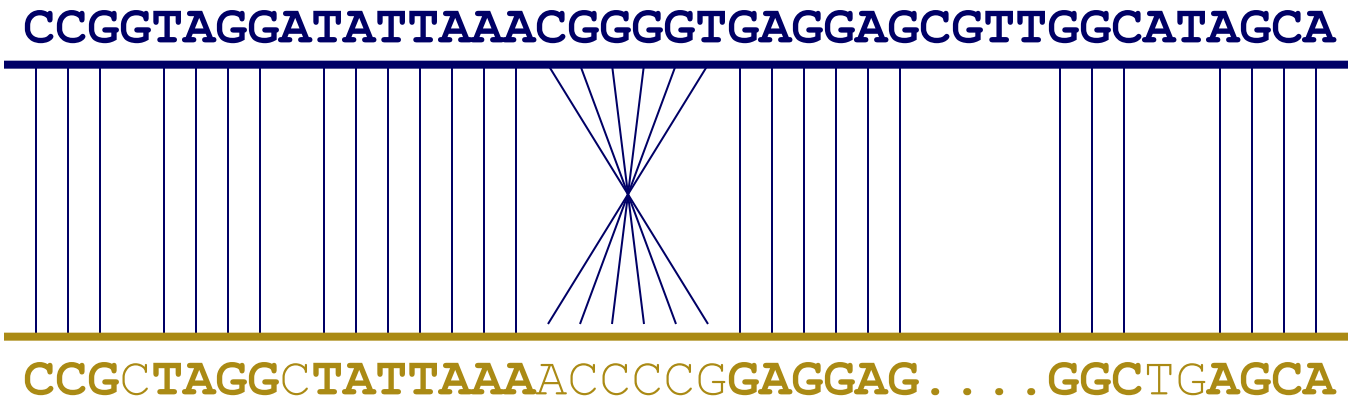
Goal of WGA

- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



Not so fast...

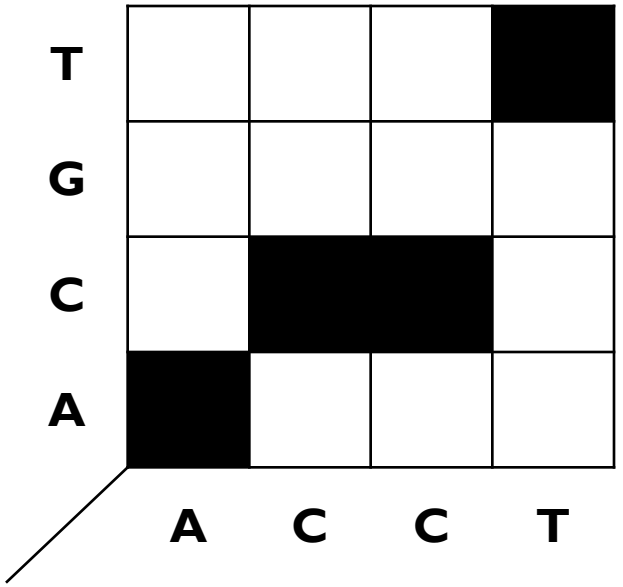
- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)



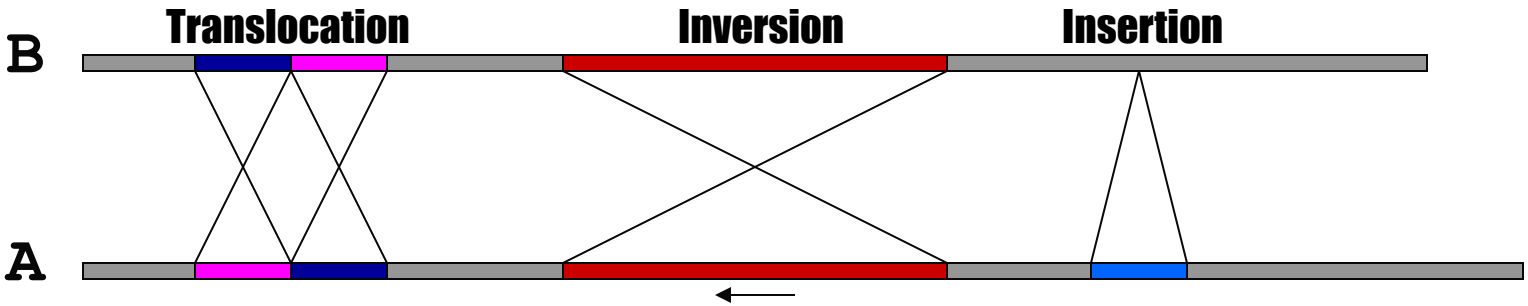
WGA visualization

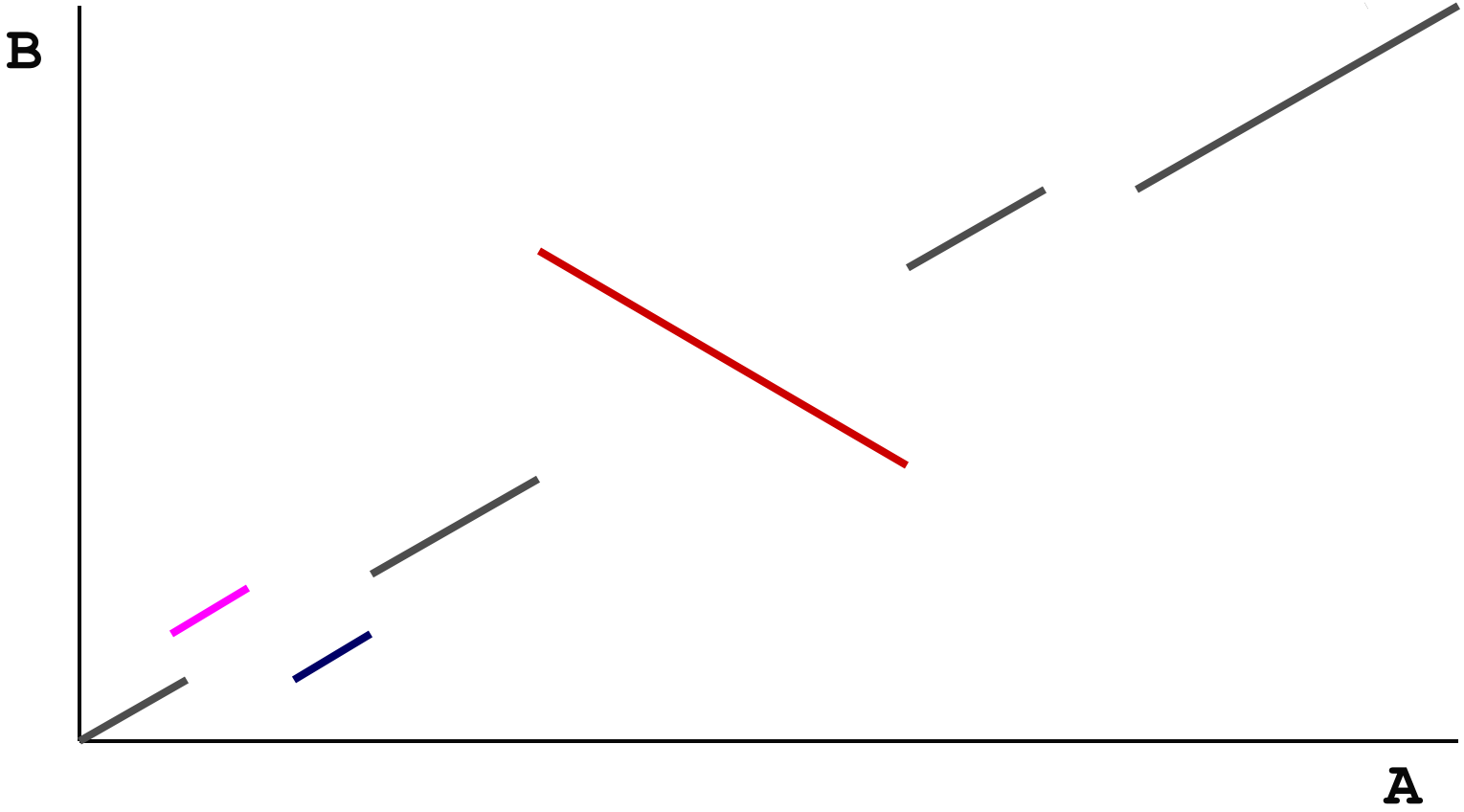
- How can we visualize *whole* genome alignments?

- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j

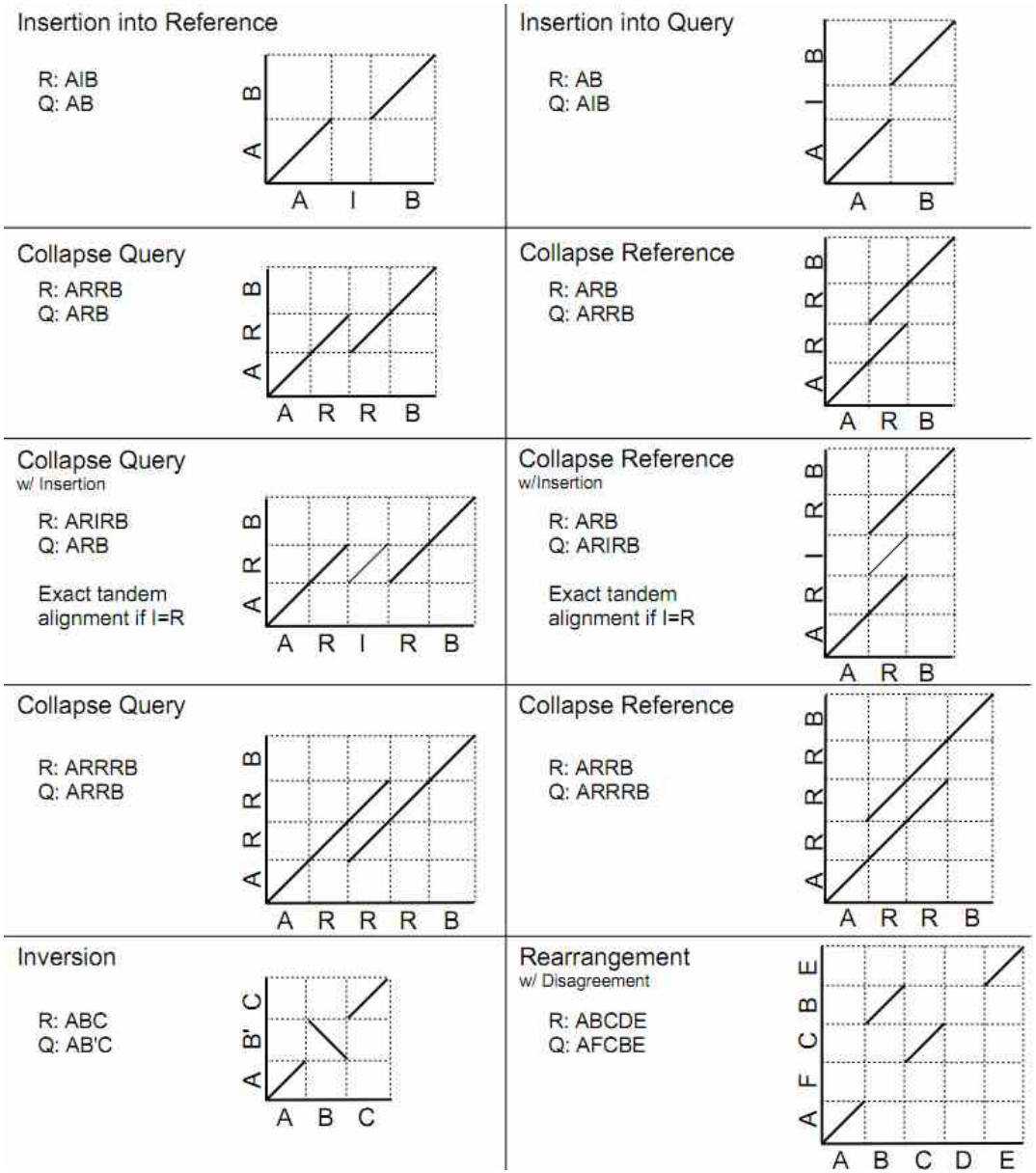


  - A perfect alignment between *A* and *B* would completely fill the positive diagonal


Translocation
Inversion
Insertion
B
A

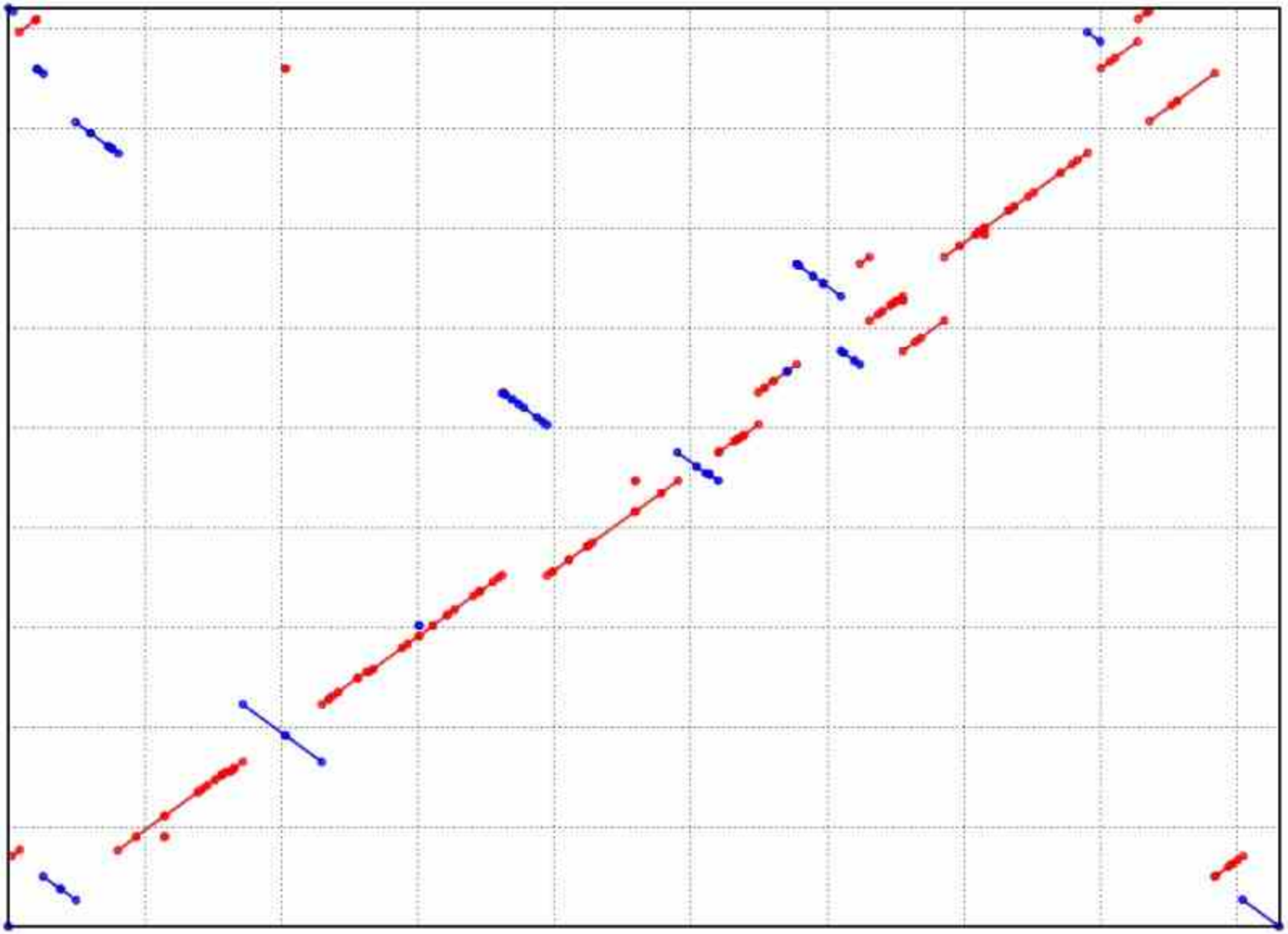

B
A

SV Types



- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

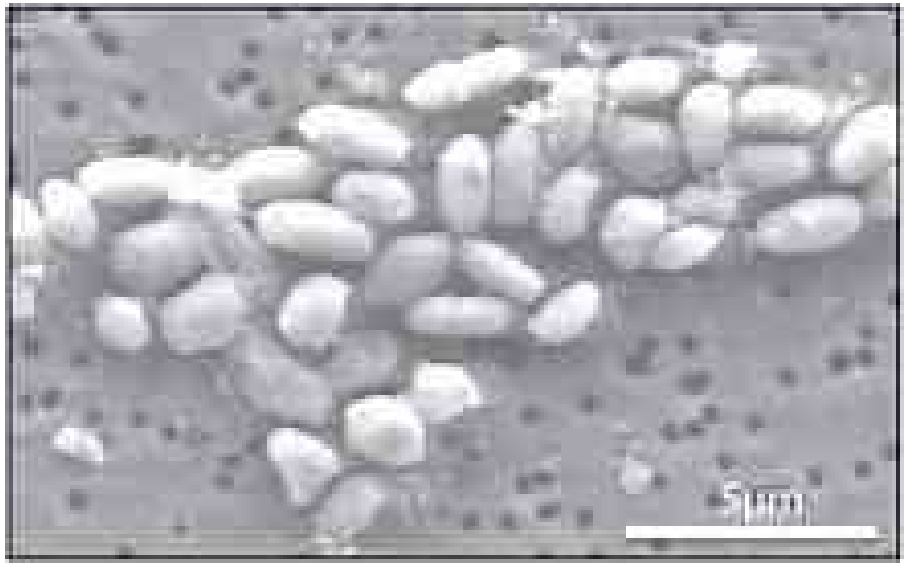
http://mummer.sf.net/manual/AlignmentTypes.pdf



Alignment of 2 strains of *Y. pestis*

http://mummer.sourceforge.net/manual/

Halomonas sp. GFAJ-1



Library 1: Fragment

Avg Read length: 100bp

Insert length: 180bp

Library 2: Short jump

Avg Read length: 50bp

Insert length: 2000bp

A Bacterium That Can Grow by Using Arsenic Instead of Phosphorus
Wolfe-Simon et al (2010) *Science. 332(6034)1163-1166.*

Digital Information Storage

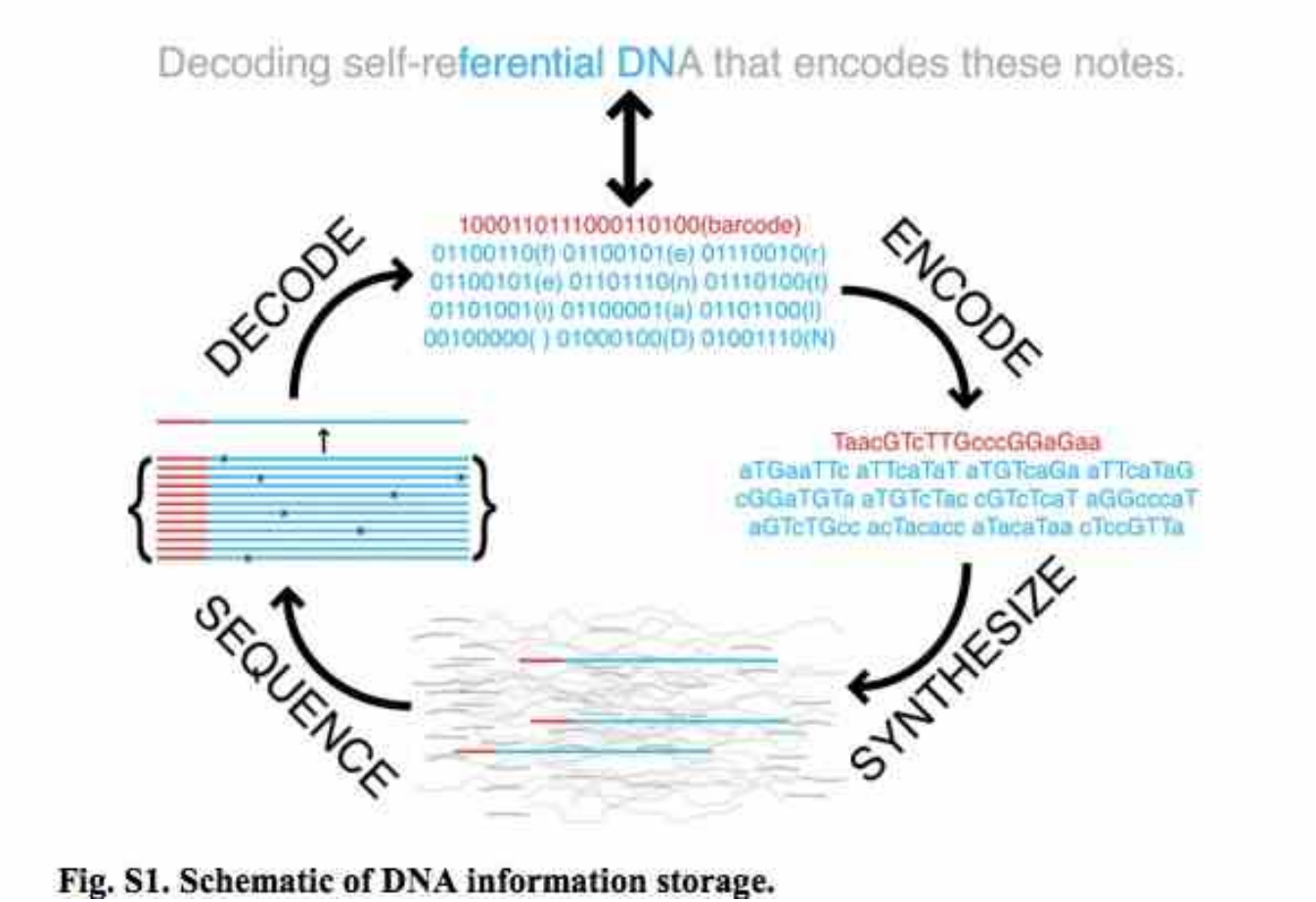


Fig. S1. Schematic of DNA information storage.

Encoding/decoding algorithm implemented in dna-encode.pl from David Dooling.

Next-generation Digital Information Storage in DNA
Church et al (2010) *Science*. 337(6102)1628

Assignment 2: Genome Assembly
Due Wednesday Sept 13 @ 11:59pm

1. ***Setup Conda/Docker/Ubuntu***
2. ***Initialize Tools***
3. ***Download Reference Genome & Reads***
4. ***Decode the secret message***
    1. *Estimate coverage, check read quality*
    2. *Check kmer distribution*
    3. *Assemble the reads with spades*
    4. *Align to reference with MUMmer*
    5. *Extract foreign sequence*
    6. *dna-encode.pl -d*

https://github.com/schatzlab/appliedgenomics2023/blob/master/assignments/assignment2/README.md

Find and decode

nucmer --maxmatch ref.fasta contigs.fasta

```
   --maxmatch    Find maximal exact matches (MEMs) without repeat filtering
   -p refctg     Set the output prefix for delta file
```

mummerplot --layout --png out.delta

```
   --layout      Sort the alignments along the diagonal
   --png         Create a png of the results
```

show-coords -rclo out.delta

```
   -r            Sort alignments by reference position
   -c            Show percent coverage
   -l            Show sequence lengths
   -o            Annotate each alignment with BEGIN/END/CONTAINS
```

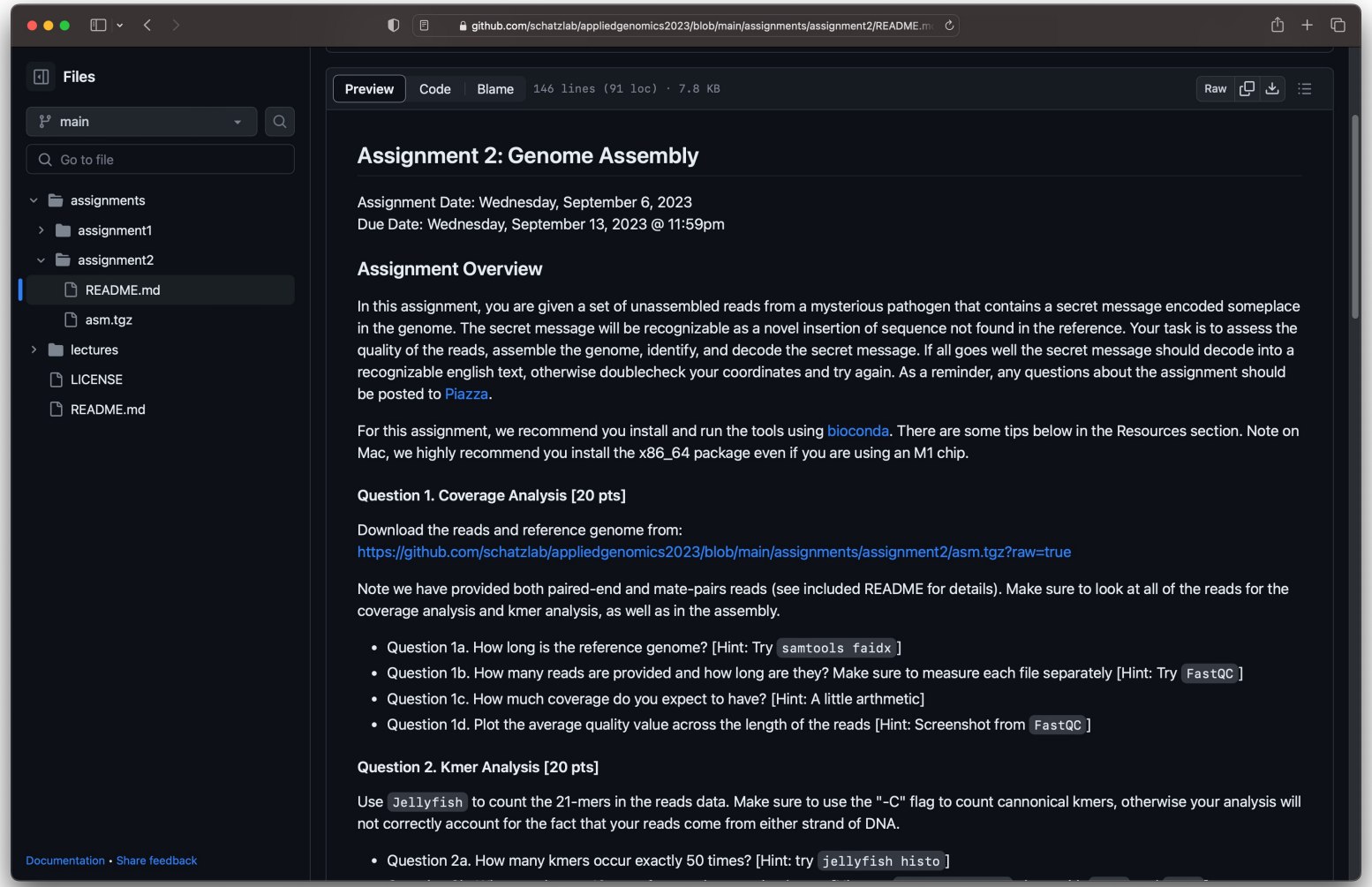
samtools faidx contigs.fasta

```
   Index the fasta file
```

samtools faidx contigs.fasta contig_XXX:YYY-ZZZ > msg.fa
dna-decode.py -d -input msg.fa

See manual at http://mummer.sourceforge.net/manual

Assignment 2: Genome Assembly
Due Wednesday Sept 13 by 11:59pm



https://github.com/schatzlab/appliedgenomics2023/tree/main/assignments/assignment2

Check Piazza for questions!