DNA Sequencing:

Computational Challenges 101:

The basics

whoami

- 15+ years wetlab biomedical research
- Heart disease
- Lipid metabolism
- Skeletal dysplasia (dwarfism)
- Cancer biology (prostate)
- Infectious disease (HIV)

whoami

- ~5 years computational biology research
- Variant discovery (dwarfism)
- Genome editing (CRISPR)
- Cancer biology (Ovarian)
- Cancer immunology (Brain)

ZYMO

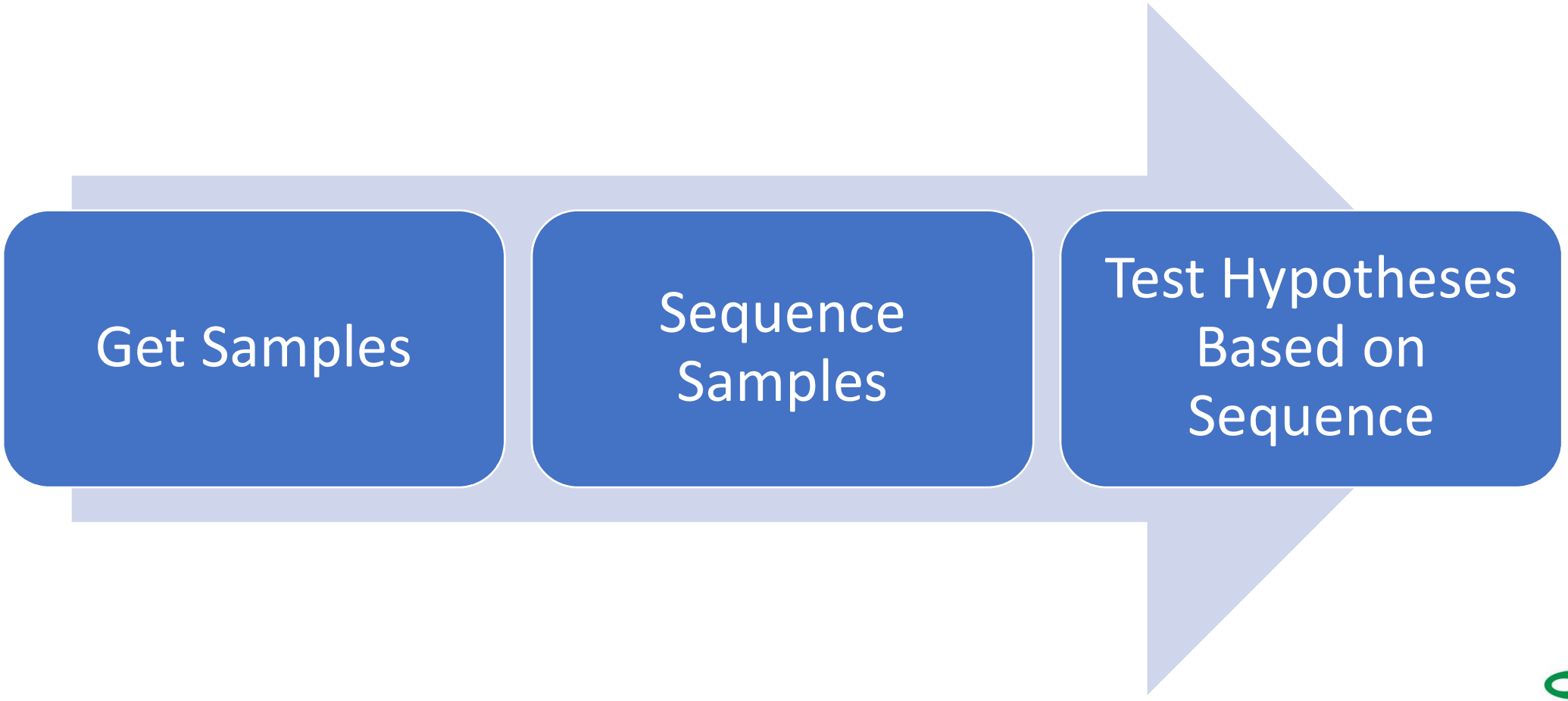
whoami

- Modeling human disease
- Efficient handling of large biological data sets
- Scientific computer program design and design practices
- Bioinformatics standards and practices
- Information security (both offensive and defensive sides)

whoami

- Stuff I like:
  - Data abstraction
  - Especially abstraction of computational concepts to biology
  - Or abstracting biological concepts to computer science analogs
  - Breaking things in a controlled manner

The Challenge:



Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

Overview

The Basic Materials of Life: Biology as information storage and

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

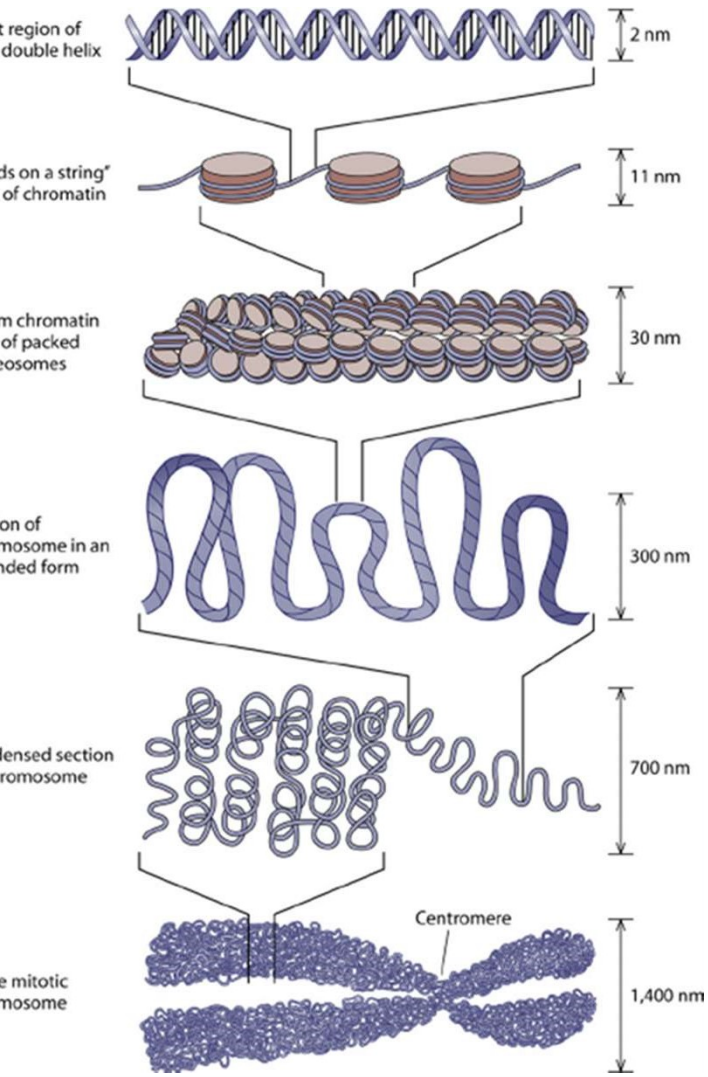
The Basic Materials of Life

*General rules and usage

- DNA: Highly stabile chemical used to encode t instructions for cellular life
  - Unit: Nucleotide = base + backbone
  - Alphabet: Bases = (A, T, G, C) N = unidentified base
  - Operations:
    - Replication to make a copy
    - Transcription to make RNA
- Backbone: Repeating pattern of sugar and phosphat
  - Sugar = deoxyribose
- Double-stranded anti-parallel: A-T and G-C
- Potentially modified by C-methylation and modificati of histones (protein for winding DNA)

ZYMO

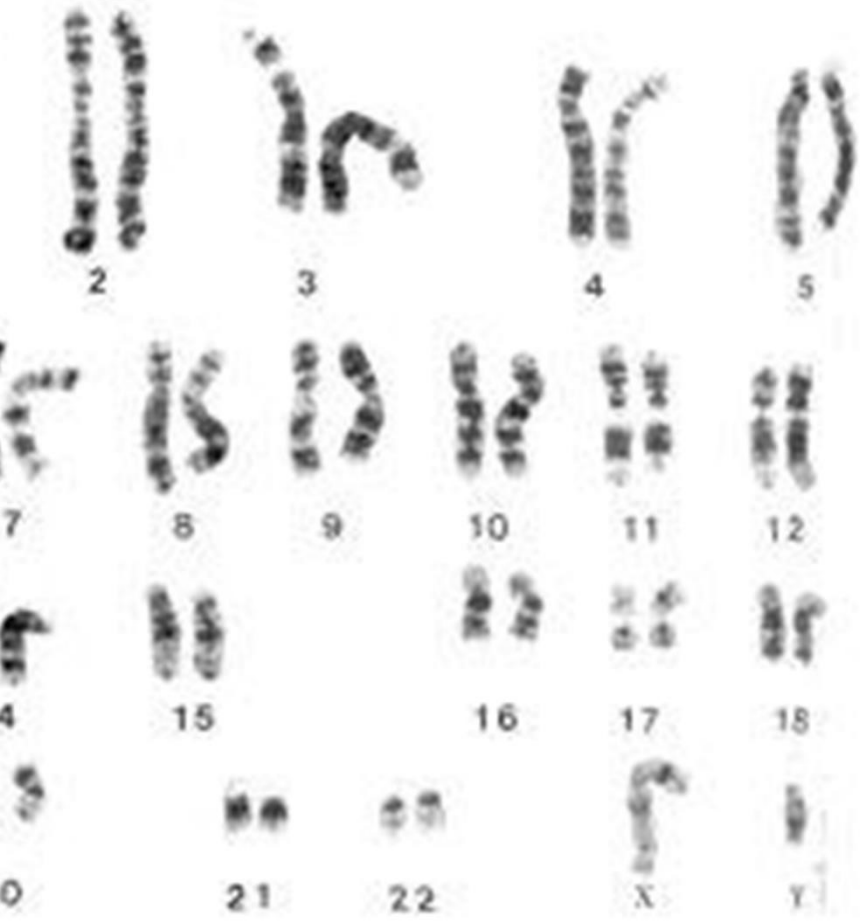
DNA Storage



- Problems with DNA as a physical information storage system
  - Negative charge can make it susceptible to reacting with positively-charged chemicals
  - Physical strings get very long (about 2 meters in a single human cell)
- Solution: coil DNA around positively-charged protein, then super-coil/condense that

ZYMO

DNA Storage



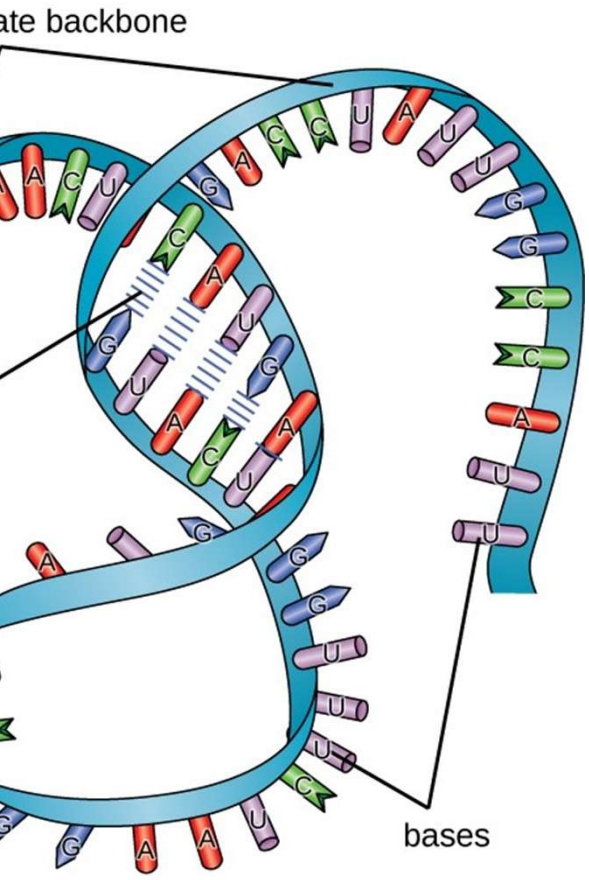
ng patterns = condensation patterns

- Problems with DNA as a physical information storage system
  - Negative charge can make it susceptible to reacting with positively-charged chemicals
  - Physical strings get very long (about 2 meters in a single human cell)
- Solution: coil DNA around positively-charged protein, then super-coil/condense that

ZYMO

The Basic Materials of Life

*General rules and usage

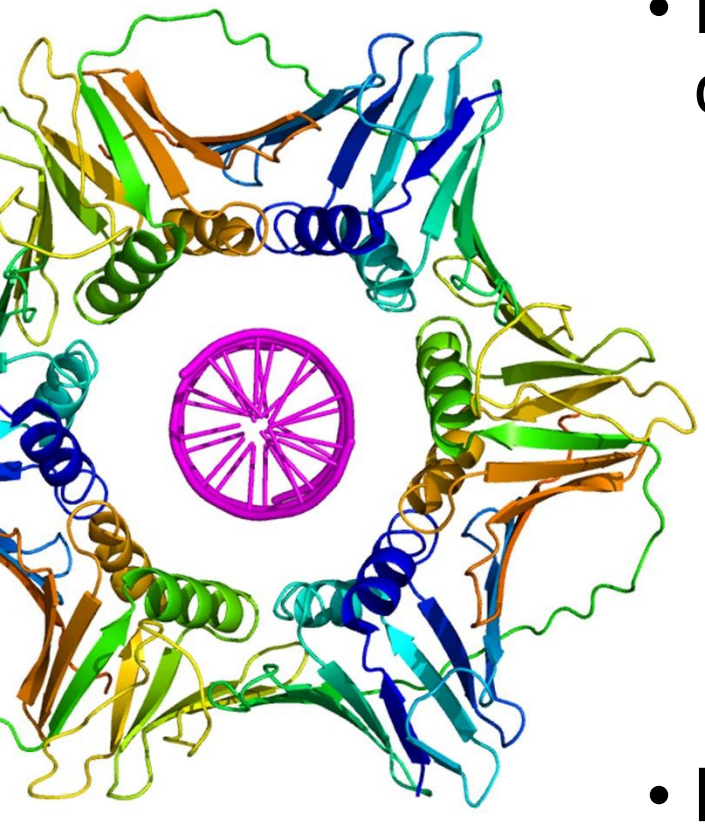


- RNA: Less stabile chemical used to ca protein-making instructions from DNA
  - Unit: Nucleotide = base + backbone
  - Alphabet: Bases = (A, U, G, C)
  - Operations:
    - Transcription to create it from DNA
    - Translation to protein by ribosome
- Backbone: Same as DNA, but with ribo instead of deoxyribose
- Single stranded, can form comp structures. Modifications being studied.

ZYMO

The Basic Materials of Life

*General rules and usage



- Protein: Variable stability molecule w complex structure and interactions
  - Unit: Amino acid
  - Alphabet: Amino acids (20 standard, thousa with modification)
  - Operations:
    - Translation from RNA
    - Catalyzing chemical reactions
    - Creating structures
    - Many others
- Backbone: Amino acid peptide bonds
- Structures of nearly infinite complexity

ZYMO

The Basic Materials of Life

*General rules and usage

ormation flows from DNA to RNA to Protein

s codec was established early when life started

cleic acids (DNA and RNA) are much easier to sequence du
ir lower complexity

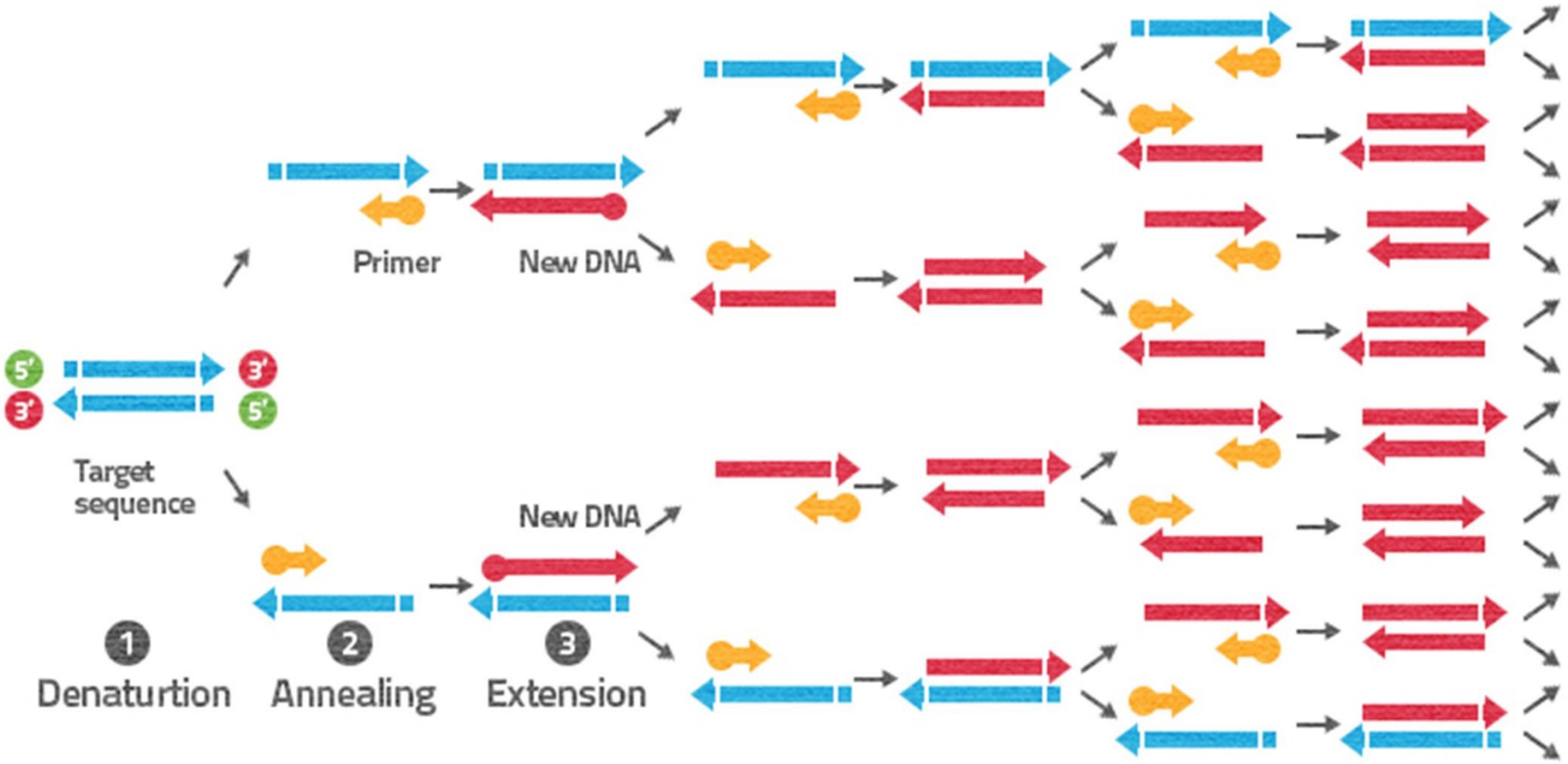
A is easier to sequence than RNA due to its high stability

DNA's double-stranded structure allows for infinite, exponential replication

ZYMO

The Basic Materials of Life

*General rules and usage



ZYMO

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

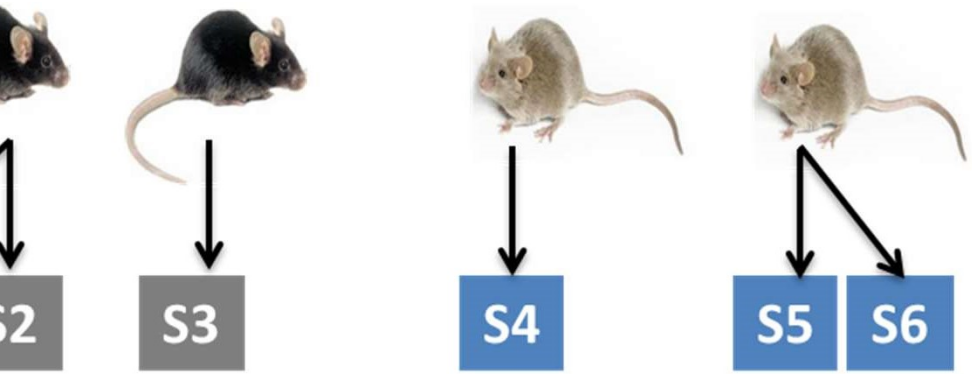
Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

Replicates: Science Must Be Reproducible



- Technical replicates
  - Sequence the same sample twice
  - Identify any noise caused due to processing or handling technique
  - Usually few required
- Biological replicates
  - Different members of the same group
  - Identify normal variance within the group
  - Biology is noisy

General Workflow

Break open cells in sample to get nucleic acids

- Reverse transcribe RNA to DNA if needed
- Cell lysis is easy in some samples (such as tissue)
- Unbiased cell lysis can be hard (mixed microbial populations)

Fragment DNA if needed

- Need consistent length with random start/stop sites

Attach adapters and filter

Quality checks

Sequence

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

d	Read length	Accuracy	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadv
e real-ng	5,500 bp to 8,500 bp avg (10,000 bp N50); maximum read length >30,000 bases	99.999% consensus accuracy; 87% single-read accuracy	50,000 per SMRT cell, or ~400 megabases	30 minutes to 2 hours	$0.33–$1.00	Longest read length. Fast. Detects 4mC, 5mC, 6mA	Moderate Equipmen expensive.
ctor	up to 400 bp	98%	up to 80 million	2h 0m 0s	$1	Less expensive equipment. Fast.	Homopoly
g (454	700 bp	99.90%	1 million	24h 0m 0s	$10	Long read size. Fast.	Runs are e Homopoly
hina)	50 to 300 bp	98%	up to 2 billion	1 to 10 days, depending upon sequencer and specified read length	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipmen expensive. high conce DNA.
ligation cing)	50+35 or 50+50 bp	99.90%	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower tha methods. sequencin palindrom
ion ncing)	400 to 900 bp	99.90%	N/A	20 minutes to 3 hours	$2,400	Long individual reads. Useful for many applications.	More expe impractica sequencin

d	Read length	Accuracy	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadv
e real-ng	5,500 bp to 8,500 bp avg (10,000 bp N50); maximum read length >30,000 bases	99.999% consensus accuracy; 87% single-read accuracy	50,000 per SMRT cell, or ~400 megabases	30 minutes to 2 hours	$0.33–$1.00	Longest read length. Fast. Detects 4mC, 5mC, 6mA	Moderate Equipmen expensive.
ctor	up to 400 bp	98%	up to 80 million	2h 0m 0s	$1	Less expensive equipment. Fast.	Homopoly
	700 bp	99.90%	1 million	24h 0m 0s	$10	Long read size. Fast.	Runs are e Homopoly
hina)	50 to 300 bp	98%	up to 2 billion	1 to 10 days, depending upon sequencer and specified read length	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipmen expensive. high conce DNA.
	50+35 or 50+50 bp	99.90%	1.2 to 1.4 billion	1 to 2 weeks	$0.13	Low cost per base.	Slower tha methods. sequencin palindrom
	400 to 900 bp	99.90%	N/A	20 minutes to 3 hours	$2,400	Long individual reads. Useful for many applications.	More expe impractica sequencin

Frequent Story in Bioinformatics

I had an issue with storing some data

I came up with a new format for storing that data

Now I have two problems

The QSEQ Format

2	2208	33.30	98.40	0	2	CTGGGATTTATTTATTCGGTTTGCAAGCTGTAAGTCTCCCCCACTGCTTC	aaacacYabK[`bdefK[RRa_e^JQQ`dI_ged^Ibfg]U_H
2	2208	38.40	99.10	0	2	CAGGGACACTGACGTAGATCAGCAAGGAGTATTCGACTTTGAGGATGTTG	_Z^ccPca`ceYce`aefhgfb_`bdf^b^^cb]f^Ucecece
2	2208	63.30	98.20	0	2	CCTAGGCAGCACTAGGCAGGCTGGGGGGCCACAAAGCGAAGAAGGCATGG	Z^^ccc`^c^R`e_`d]]eaUaefbH^YEUGQJ[[bREQUUHW
2	2208	60.10	98.60	0	2	AGGAGATGGCCTTTTTGGGCAAGGACAAGCCATCTTCAGAGAATAATGAG	^__`c\`cecgggifhfiihidafecgadfhhfhif`ccafXa
2	2208	66.80	98.80	0	2	GCCATCAATGTCACCAATCAGTGCCTTTGAGGGTTGTCCATCTCCCAAAG	_V^^cdcac``c^bfgiJ`Y^b`gacghfX`ZbbeggfI^[ae
3	**4**	**5**	**6**	**7**	**8**	**9**	**10**

ine name: unique identifier of the sequencer.
umber: unique number to identify the run on the sequencer.
number: positive integer (currently 1-8).
umber: positive integer.
ordinate of the spot. Integer (can be negative).
ordinate of the spot. Integer (can be negative).
: positive integer. No indexing should have a value of 1.
Number: 1 for single reads; 1 or 2 for paired ends.
ence
ty: the calibrated quality string.
Did the read pass Illumina filtering? 0 - No, 1 - Yes

ZYMO

The QSEQ Format

2	2208	33.30	98.40	0	2	CTGGGATTTATTTATTCGGTTTGCAAGCTGTAAGTCTCCCCCACTGCTTC	aaacacYabK[`bdefK[RRa_e^JQQ`dI_ged^Ibfg]U_H
2	2208	38.40	99.10	0	2	CAGGGACACTGACGTAGATCAGCAAGGAGTATTCGACTTTGAGGATGTTG	_Z^ccPca`ceYce`aefhgfb_`bdf^b^^cb]f^Ucecece
2	2208	63.30	98.20	0	2	CCTAGGCAGCACTAGGC[illegible]CAAAGCGAAGAAGGCATGG	Z^^ccc`^c^R`e_`d]]eaUaefbH^YEUGQJ[[bREQUUHW
2	2208	60.10	98.60	0	2	AGGAGATGGCCTTT[illegible]TTCAGAGAATAATGAG	^__`c\`cecgggifhfiihidafecgadfhhfhif`ccafXc
2	2208	66.80	98.80	0	2	GCCATCAATGTC[illegible]CCATCTCCCAAAG	_V^^cdcac``c^bfgiJ`Y^b`gacghfX`ZbbeggfI^[ae
3	4	5	6	7	8		10

ine name: unique identifier of
umber: unique number to ide
number: positive integer (curr
umber: positive integer.
ordinate of the spot. Integer
ordinate of the spot. Integer
: positive integer. No indexing
Number: 1 for single reads; 1
ence
ty: the calibrated quality string.
Did the read pass Illumina filtering? 0 - No, 1 - Yes

ZYMO

The FASTQ Format

Read sequence

unique identifier/name
equivalent to columns
in qseq.txt format

```
@ERR030881.107 HWI-BRUNOP16X_0001:2:1:13663:1096#0/1
ATCTTTTGTGGCTACAGTAAGTTCAATCTGAAGTCAAAACCAACCAATTT
+
5.544,444344555CC?CAEF@EEFFFFFFFFFFFFFFFFEFFFEFFF
@ERR030881.311 HWI-BRUNOP16X_0001:2:1:18330:1130#0/1
TCCATACATAGGCCTCGGGGTGGGGGAGTCAGAAGCCCCCAGACCCTGTG
+
GFFFGFFBFCHHHHHHHHHHIHEEE@@@=GHGHHHHHHHHHHHHHHHHH
@ERR030881.1487 HWI-BRUNOP16X_0001:2:1:4144:1420#0/1
GTATAACGCTAGACACAGCGGAGCTCGGGATTGGCTAAACTCCCATAGTA
+
55*'+&&5'55('''888:8FFFFFFFFF4/1;/4./++FFFFF=5:E#
```

Read

Read

Read

Quality string

a constant field, sometimes used for notes

ZYMO

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

The Challenge

We need a reference genome for any organism we wish to study

There is no existing technology that can read from one end of a chromosome to the other reliably

This reference building is made harder by repetitive sequence and made easier with longer reads

Longer reads with nanopore are starting to help

The Solutions

Cheat/crib the reference assembly process with an already assembled reference

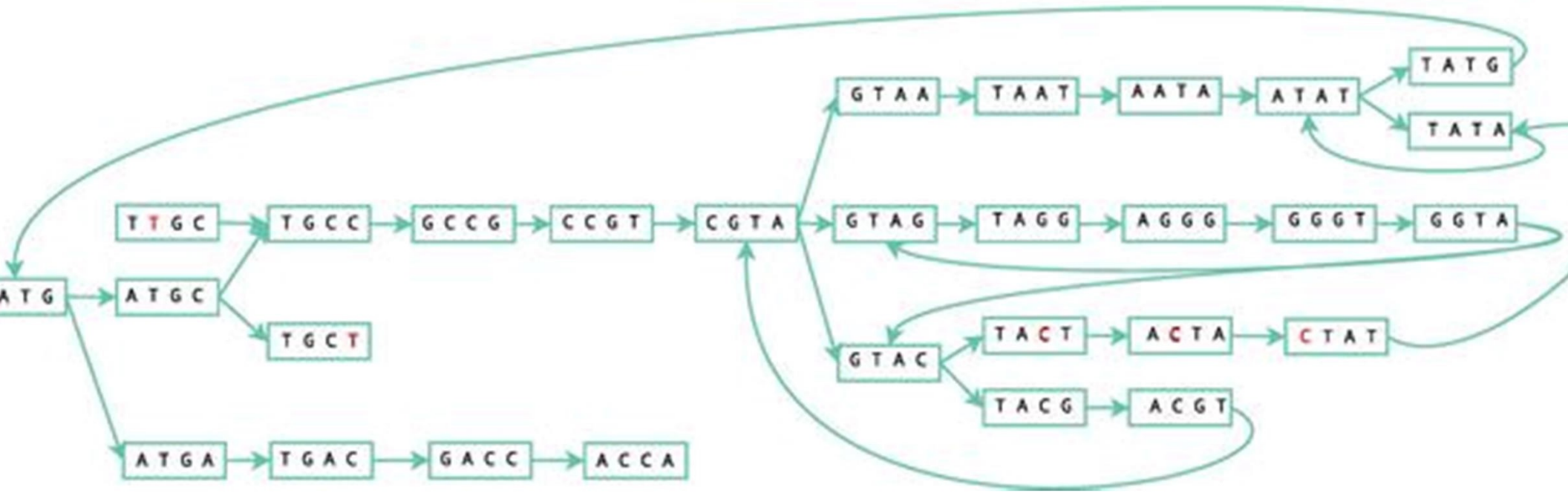
- I want to assemble a reference genome for the Irvine Brown Pigeon
- A colleague has already assembled a genome for the Santa Barbara yellow pigeon
- Jackpot!

Assembling a genome from short reads with some overlap can be reduced to a de Bruijn graph problem

- The solution is the Euler path through the graph

Reads are getting longer due to improved sequencing methods

The de Bruijn Method



The de Bruijn Method

Velvet assembler (serial): needs over 2TB of RAM, takes days

ABySS: Uses MPI for parallelism, 196 Cores for about 96 hours

SOAPdenovo: Uses threading, 40 cores with 150GB of RAM o
more can do the job in about 40 hours

A Weakness in the Assembly Process

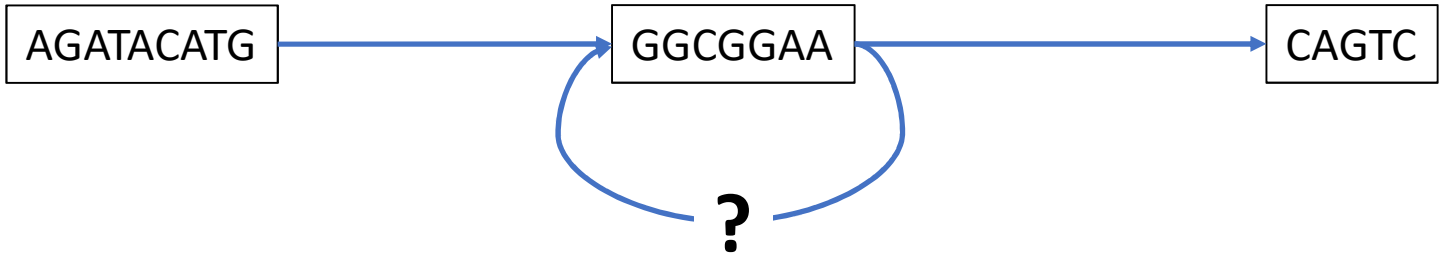
AGATACATGGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAACAG

Low-complexity Sequences: A Weakness in the Assembly Process

AGATACATGGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAACAG

Low-complexity Sequences: A Weakness in the Assembly Process

AGATACATGGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAAGGCGGAACAG

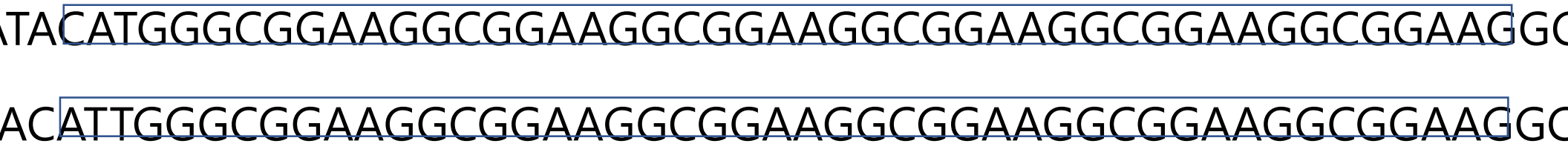


If your read length is less than the length of the low-complexity sequence region, you cannot know how long it is

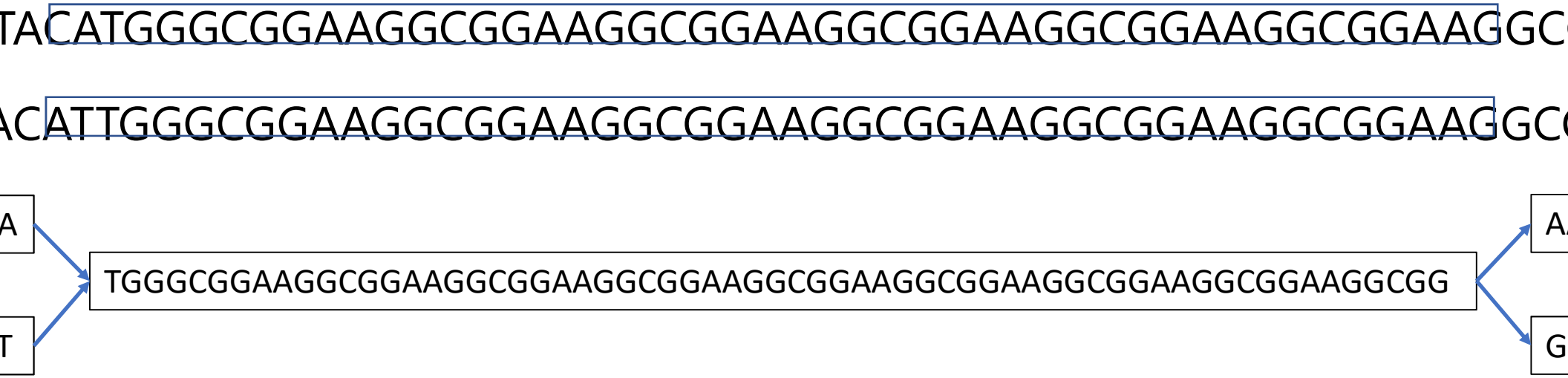
Newer methods (nanopore specifically) have been generating significantly longer reads to resolve some of these

ZYMO

A Weakness in the Assembly Process



Repeated Sequences: A Weakness in the Assembly Process



If your read length is less than the length of the repeated sequence, you may be unable to place it appropriately

Longer reads help with this problem as well. Some sequences end up being unplacable in the genome.

ZYMO

Storing Reference Sequence

uence Name

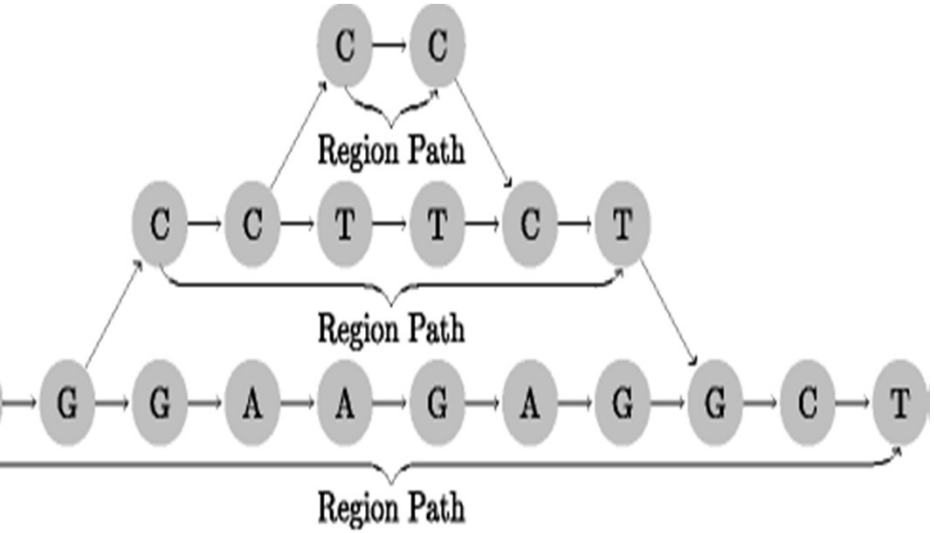
d

uence

```
>yneN
TTAATGCCTCTTCTCATTTCTTCTGCTGTCATCCGCACAGCAGAAGAATTCCTCATTGAC
TATTATTTCGCAATTTGCTCACATGGATTAAATTAAACTACATACTATAAGATATAAACT
TCTGCCTACAGCTGTAAGAAACTCCGCTCAGTACTGAAGCACCAGTCCTATTTCCTCTTT
TCTCCAGCCTGTTATATTAAGCATACTGATTAACGATTTTTAACGTTATCCGCTAAATAA
ACATATTTGAAATGCATGCGACCACAGTGAAAAACAAAATCACGCAAAGAGACAACTATA
A
>yegR
ACTAACGGCTGCCACCGATAAATTTCAAAAAAGAGCATATACCTAATATTCAACTAAACA
GTGGCATCTTCAATATAATATATTAAAGCCCCCATGGAGTTACCCTGAAGGGCCTCAATG
TCCGTAATTCCTACTTATGTAGGAAATGTTGTACAGAACATTTATTATAATCCTATTCAA
TTATAATAATCATGCCATTATTATATTTAAACACTAGAGAGTGTCGTTGGTATTTAATGG
GGGAAGGTGAGATGAAAAAGATAGCTGCTATATCATTAATTAGTATTTTTATTATGTCTG
G
>emrK
AAATCAGGGATTGTACCGATGATTTATAGTTTCAAGTTGGCACTATAAGTCTTCTTACTA
ATCCTACAGGCGTAAGAATTGTATTGCAAAAGCCACGGTTTAGTCCTCTGTTGTTTTTTT
TGCACCTCATTTAAATTAGGCCTCCAACGTTCCTGGGATAATGTGCAACACATGCACTGT
GTTTGATATGAAGAATGAATGCTCTTTTCATTCAATTCATAAATTTCATCTATGAGAAAT
GAGAGATAATAGTGGAACAGATTAATTCAAATAAAAAACATTCTAACAGAAGAAAATACT
T
>evgA
AATACAATTCTTACGCCTGTAGGATTAGTAAGAAGACTTATAGTGCCAACTTGAAACTAT
AAATCATCGGTACAATCCCTGATTTTATTGTTGACATTTCATTTATGCCGACTATTTATA
TGGTATACTTGTCGAATTATCTTAAAGGAAGCTCAGATTTTCTTATTTTTATTGAGAAAA
TGAGATGACGCCTTATGTCTGTATTACTACAGGGAGAAGGGAGATGCTTCATTGCAAAGG
GAATAATCTATGAACGCAATAATTATTGATGACCATCCTCTTGCTATCGCAGCAATTCGT
>yfdX
TGGCTGTATTTACATTTAATTAATCAGTATTTACATCGATATAATAAATGACATCTCTTT
GTGGTATATAAGAATAGTTCTCTGCGACAGGAAGCATATTCCTACAATTGTAAGACTAAA
ATACTTCTTGCGATAATAACTACAACTGTAAGATAACCCTTTCAAAATGACCGTTGCTCT
CTGATTTCTCATTTCATGCTCACCCAATATGATGGCGGCGTTTTCTAAAACTGTTAAAGA
ATGAGGTAAGTATGAAACGTTTAATTATGGCCACGATGGTCACAGCAATTCTGGCATCTT
C
```

ZYMO

A Better Reference Genome?



et al., 2017

- The human reference geno matches nobody.
- Some positions are invarian
  - Base changes are incompati with life
- Some positions are highly variable
  - Contribute to observed huma diversity
- A graph-based reference ha been proposed

ZYMO

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

The Challenge

We have a big collection of reads from an organism with an assembled reference genome

We want to figure out where in the genome each read came from without having to reassemble a new genome for the sample.

Brute-force Alignment

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGGGACTCGGGATCTCAGT

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGGGACTCGGGATCTCAGT

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGGGACTCGGGATCTCAGT

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGGGACTCGGGATCTCAGT

• • •

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGGGACTCGGGATCTCAGT

Brute-force Alignment

- Assumptions
  - 50 bp reads
  - 3.3 x 10^9 bases in the genome
  - 5 x 10^7 reads from an experiment
- 8.25 x 10^{18} (8,250,000,000,000,000,000) ops
- Computation time
  - 4.125 billion seconds on a computer that can do about 2 billion operations / second
  - About 130 years on a single computer…
  - ...and we haven't even worried about indels yet

2 Big Questions

1. How do we deal with insertions and deletions?

2. How do we complete our alignments sometime before our great-grandchildren forget why we were doing the alignment in the first place?

Dealing With Indels Using Seeds

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGGGACTCGGGATCTCAGT

Dealing With Indels Using Seeds

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

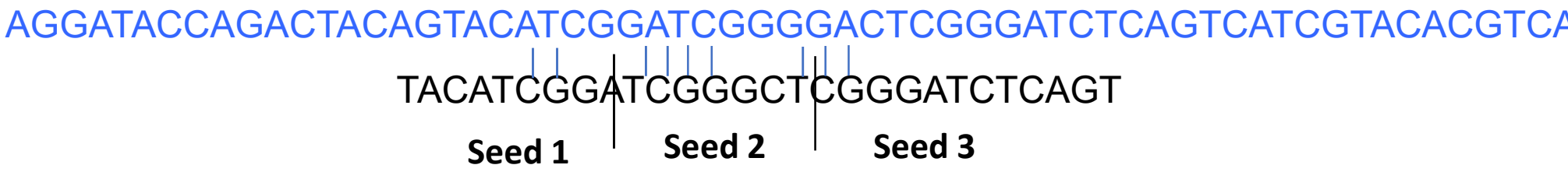
TACATCGGATCGGG--CTCGGGATCTCAGT

Dealing With Indels Using Seeds

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGGCTCGGGATCTCAGT

Dealing With Indels Using Seeds



Dealing With Indels Using Seeds

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGA
TCGGGCTCG
GGATCTCAGT

Dealing With Indels Using Seeds

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

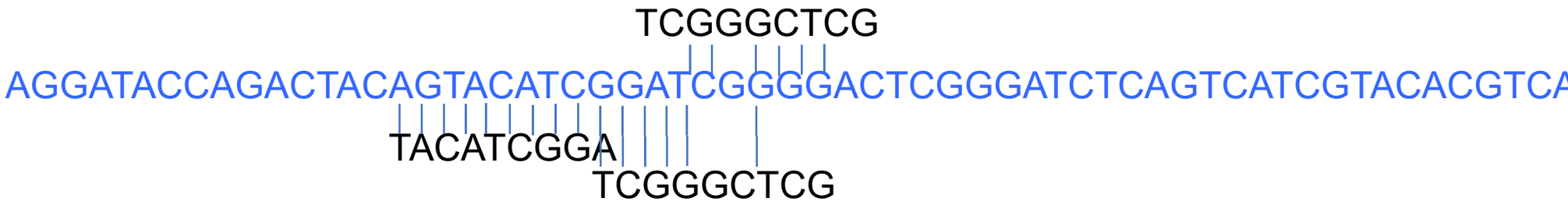
TACATCGGA

TCGGGCTCG
GGATCTCAGT

Dealing With Indels Using Seeds

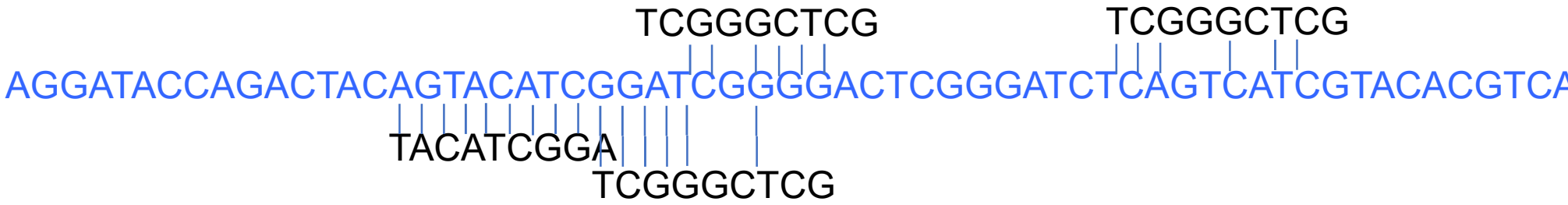
GGATCTCAGT

Dealing With Indels Using Seeds



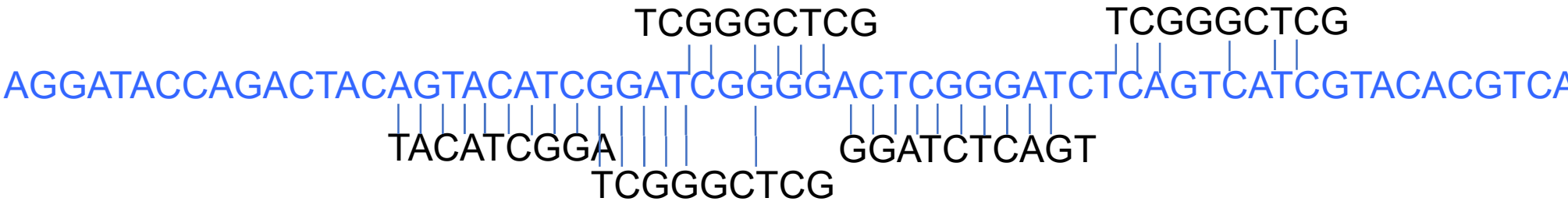
GGATCTCAGT

Dealing With Indels Using Seeds

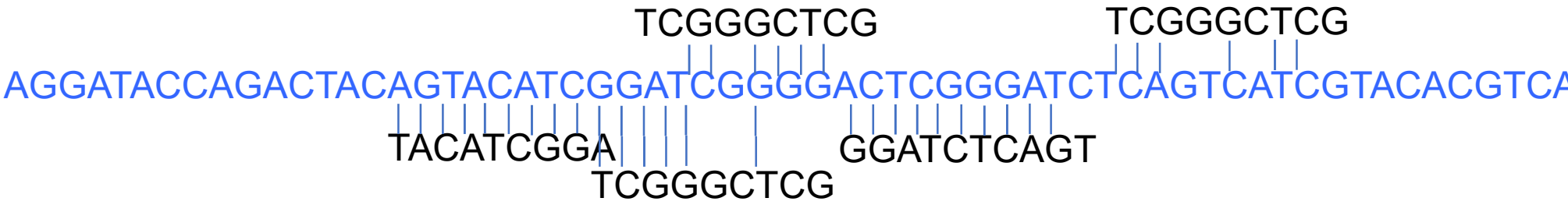


GGATCTCAGT

Dealing With Indels Using Seeds



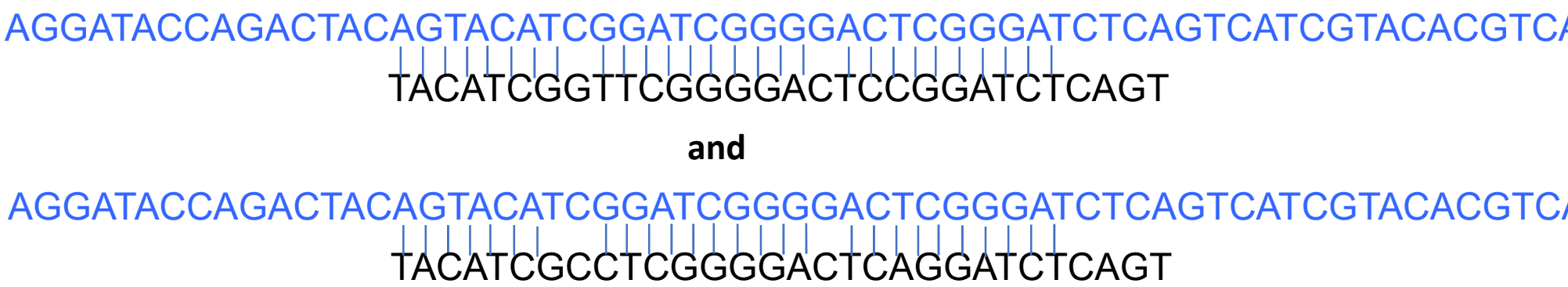
Dealing With Indels Using Seeds



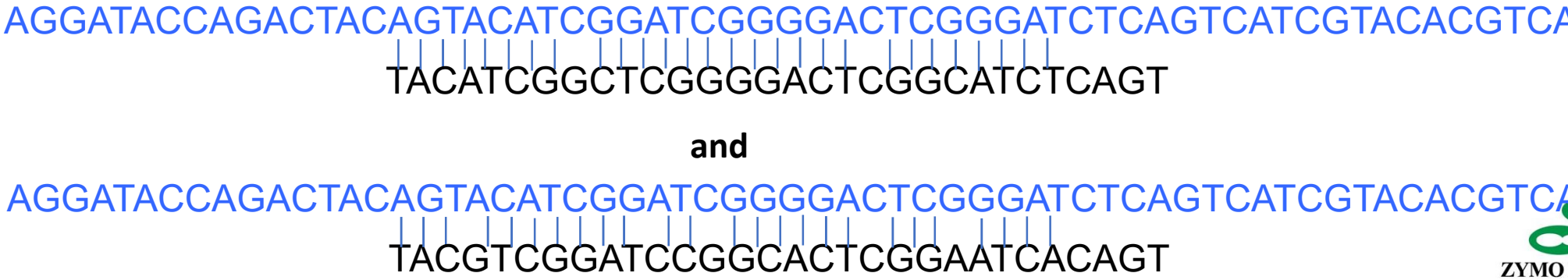
Could this be modified to become the alignment method?

Scoring an Alignment

Scoring an Alignment



vs.



Scoring an Alignment

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGGATCGGG--CTCGGGATCTCAGT

vs.

AGGATACCAGACTACAGTACATCGGATCGGGGACTCGGGATCTCAGTCATCGTACACGTCA

TACATCGCATCGGGGCCTCGGGATTTCAGT

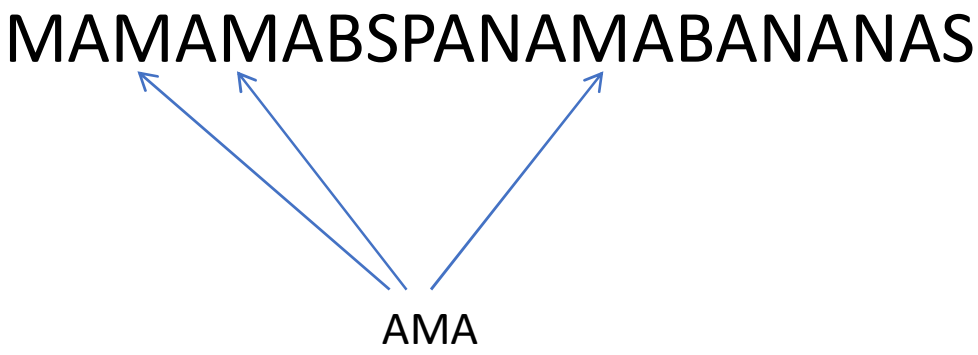
Mama Mab's Panama Bananas

ZYMO

Mama Mab's Panama Bananas

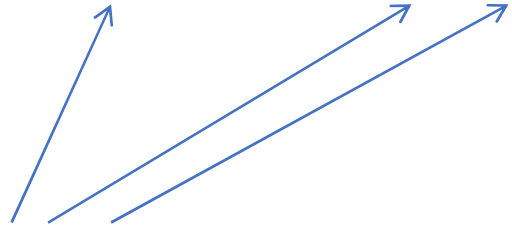
MAMAMABSPANAMABANANAS

Mama Mab's Panama Bananas



Mama Mab's Panama Bananas

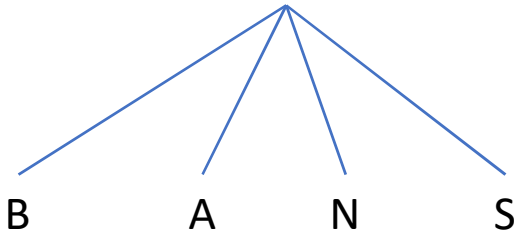
MAMAMABSPANAMABANANAS



ANA

Mama Mab's Panama Bananas

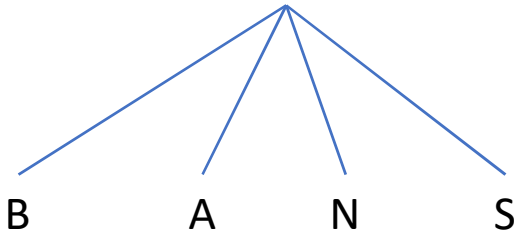
0 BANANAS
1 ANANAS
2 NANAS
3 ANAS
4NAS
5 AS
6 S



ZYMO

Mama Mab's Panama Bananas

0 BANANAS
1 ANANAS
2 NANAS
3 ANAS
4 NAS
5 AS
6 S

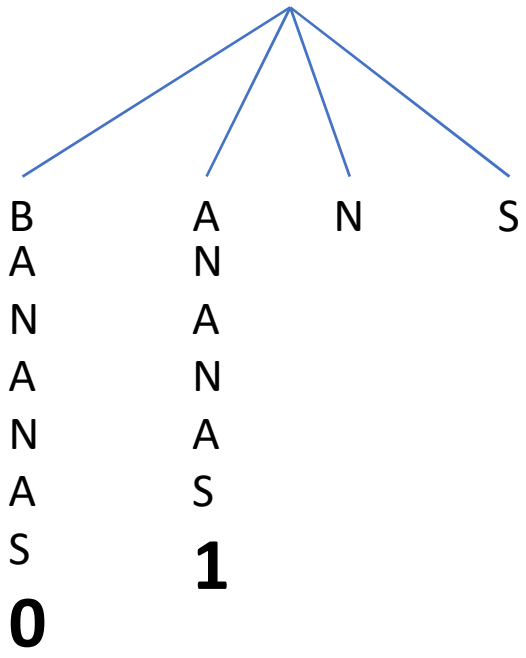


0

Mama Mab's Panama Bananas

0 BANANAS

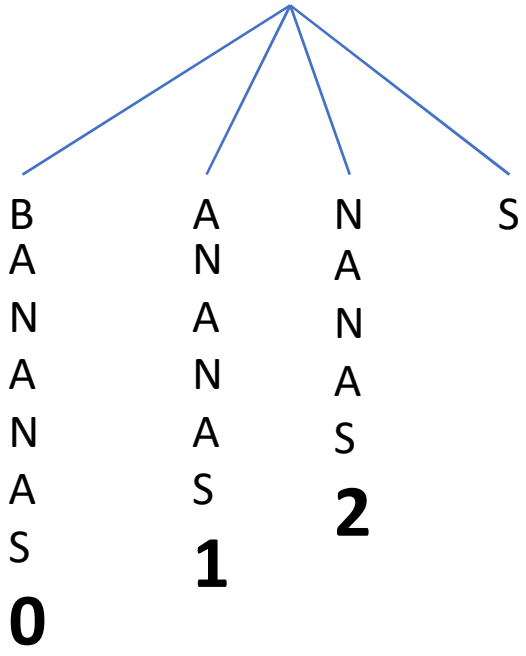
2 NANAS
3 ANAS
4 NAS
5 AS
6 S



Mama Mab's Panama Bananas

0 BANANAS

3 ANAS
4 NAS
5 AS
6 S



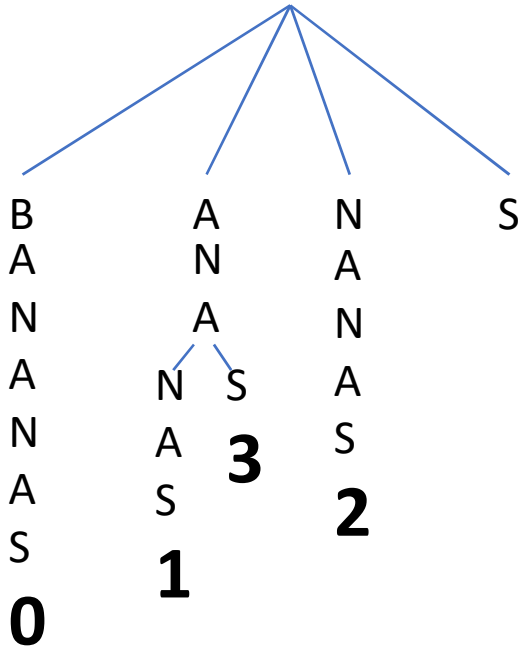
Mama Mab's Panama Bananas

0 BANANAS

4 NAS

5 AS

6 S

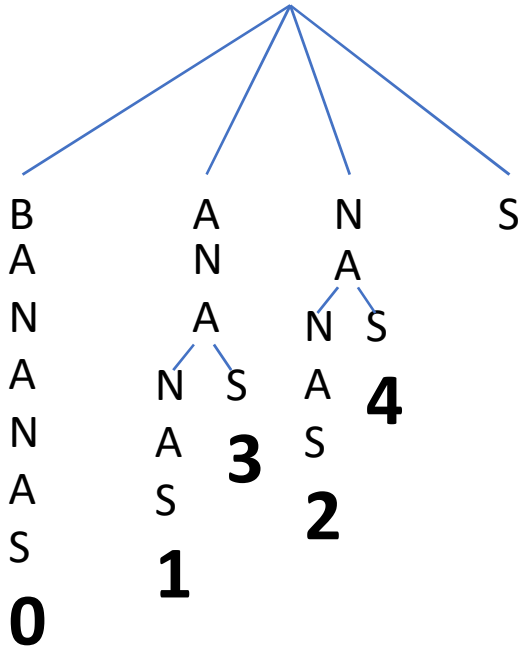


Mama Mab's Panama Bananas

0 BANANAS

5 AS

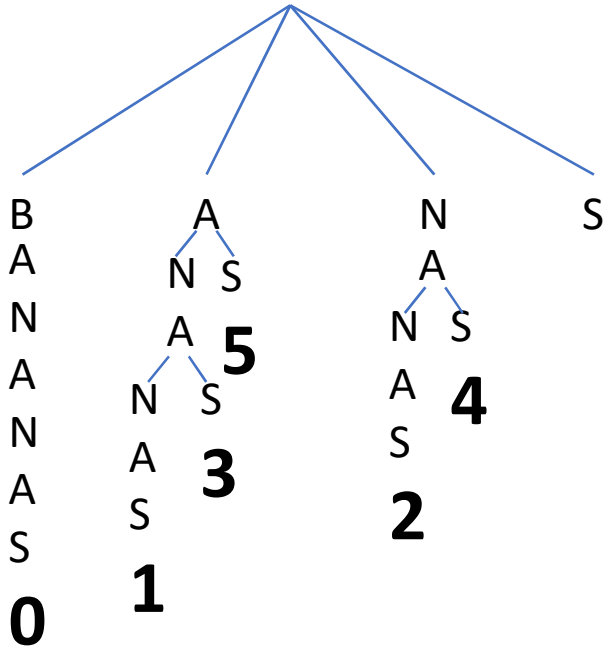
6 S



ZYMO

Mama Mab's Panama Bananas

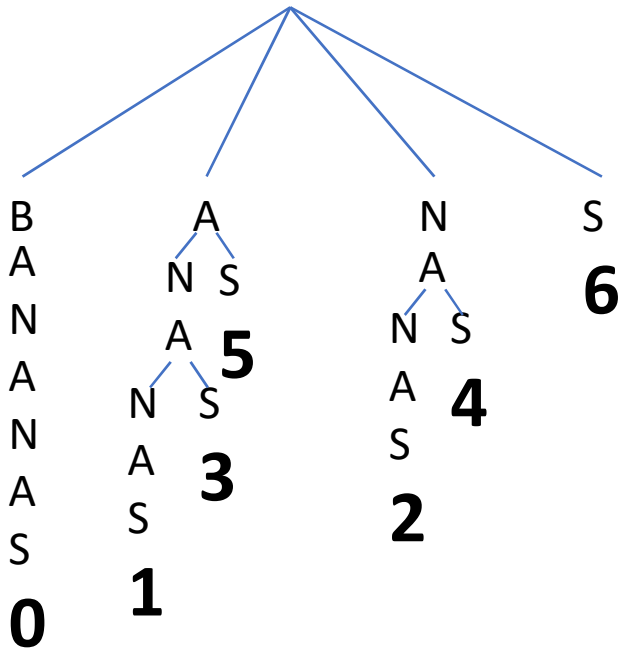
0 BANANAS



6 S

Mama Mab's Panama Bananas

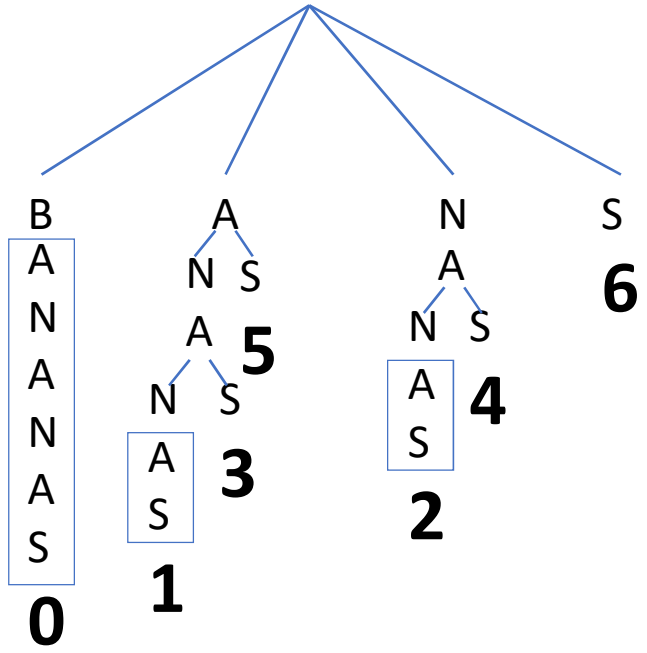
0 BANANAS



Suffix Trie

Mama Mab's Panama Bananas

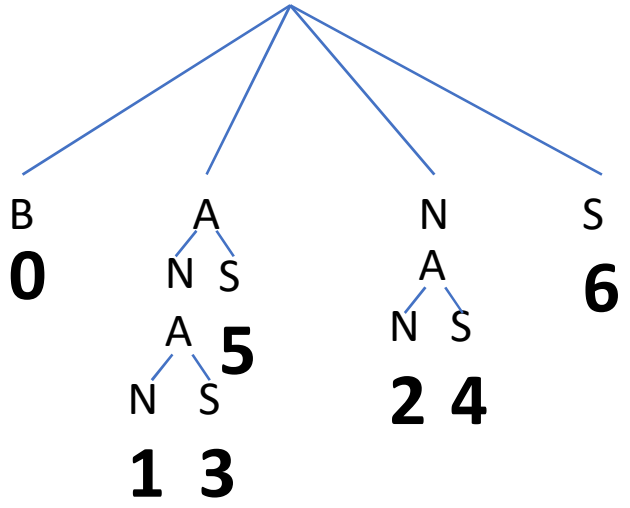
0 BANANAS



Suffix Trie

Mama Mab's Panama Bananas

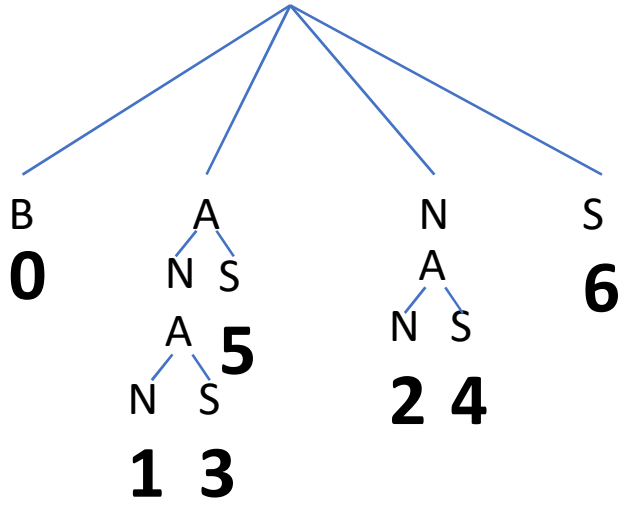
0 BANANAS



Suffix Tree

Mama Mab's Panama Bananas

0 BANANAS



Suffix Tree

Good, but can become very complex/large to store in memory

ZYMO

Suffix Trees

Pro: Alignment operations come down to one operation per letter of read

Con: On the scale of the human genome, they are painful to store in memory

Hack: Burrows-Wheeler transformation

- Effectively sorts every letter by its surrounding letters
- Still effectively a suffix tree, but much more complex to implement and much more efficient to use in terms of resources
- Sequences with limited alphabets and repeating patterns are extremely compressable while still being searchable/alignable
- One of the two main methods being used for alignments, along with seed-based
- Too complicated to walk through during this talk

The SAM Format Standard

```
rsden@login2 workshop2_data]$ head 4_bwa_human001_pe12.sam
    SN:chr14        LN:107349540
    ID:bwa  PN:bwa  VN:0.7.7-r441   CL:/u/local/apps/bwa/current/bwa sampe /u/home/c/cmarsden/workshop2_data/refe
4.fa 3_bwa_human001_pe1.sai 3_bwa_human001_pe2.sai 2_sic_human001_pe1.fastq 2_sic_human001_pe2.fastq
67577.3006  99      chr14   57829687        60      101M    =       57829772        170     CAATCTATTTAAAGTAATCCC
TATGCTCCCCCACAGCCCTTATAATATTTTAAGAGCATGTCTTTTTGTTTACATTTTTCCCATTAAATTG      IIIHIHIIIIIIIIIHIHIHIIHIIIHGGIIIIIGHHI
BIEDHFFFGGGEHD0BB>E@BCEDEGDECACF9CCBBEB>B@>??<A;??;A##      XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i
    XG:i:0  MD:Z:101
67577.3006  147     chr14   57829772        60      85M     =       57829687        -170    TTTTCCCATTAAATTGTGAGC
ACCCTTTATCCAGTTAATCTTTTTGTCCCTCCAAAGCCAGAATGTACAGCTCTG      @5@7@7DBDC>>?::@@E@BBB,EBACEBDEABEEDGGEGHGDHHHFCFIHHC
IIIHIFIIIIGIIIIIHIIIII      XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:85
```

The SAM Format Standard

```
rsden@login2 workshop2_data]$ head 4_bwa_human001_pe12.sam
    SN:chr14        LN:107349540
    ID:bwa  PN:bwa  VN:0.7.7-r441   CL:/u/local/apps/bwa/current/bwa sampe /u/home/c/cmarsden/workshop2_data/refe
4.fa 3_bwa_human001_pe1.sai 3_bwa_human001_pe2.sai 2_sic_human001_pe1.fastq 2_sic_human001_pe2.fastq
67577.3006  99      chr14   57829687        60      101M    =       57829772        170     CAATCTATTTAAAGTAATCCC
TATGCTCCCCCACAGCCCTTATAATATTTTAAGAGCATGTCTTTTTGTTTACATTTTTCCCATTAAATTG      IIIHIHIIIIIIIIIHIHIHIIHIIIHGGIIIIIGHHI
BIEDHFFFGGGEHD0BB>E@BCEDEGDECACF9CCBBEB>B@>??<A;??;A##      XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i
    XG:i:0  MD:Z:101
67577.3006  147     chr14   57829772        60      85M     =       57829687        -170    TTTTCCCATTAAATTGTGAGC
ACCCTTTATCCAGTTAATCTTTTTGTCCCTCCAAAGCCAGAATGTACAGCTCTG      @5@7@7DBDC>>?::@@E@BBB,EBACEBDEABEEDGGEGHGDHHHFCFIHHC
IIIHIFIIIIGIIIIIHIIIII      XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:85
```

ZYMO

The SAM Format Standard

```
rsden@login2 workshop2_data]$ head 4_bwa_human001_pe12.sam
    SN:chr14        LN:107349540
    ID:bwa  PN:bwa  VN:0.7.7-r441   CL:/u/local/apps/bwa/current/bwa sampe /u/home/c/cmarsden/workshop2_data/refe
4.fa 3_bwa_human001_pe1.sai 3_bwa_human001_pe2.sai 2_sic_human001_pe1.fastq 2_sic_human001_pe2.fastq
67577.3006  99      chr14   57829687        60      101M    =       57829772        170     CAATCTATTTAAAGTAATCCC
TATGCTCCCCCACAGCCCTTATAATATTTTAAGAGCATGTCTTTTTGTTTACATTTTTCCCATTAAATTG      IIIHIHIIIIIIIIIHIHIHIIHIIIHGGIIIIIGHHI
BIEDHFFFGGGEHD0BB>E@BCEDEGDECACF9CCBBEB>B@>??<A;??;A##      XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i
    XG:i:0  MD:Z:101
67577.3006  147     chr14   57829772        60      85M     =       57829687        -170    TTTTCCCATTAAATTGTGAGC
ACCCTTTATCCAGTTAATCTTTTTGTCCCTCCAAAGCCAGAATGTACAGCTCTG      @5@7@7DBDC>>?::@@E@BBB,EBACEBDEABEEDGGEGHGDHHHFCFIHHC
IIIHIFIIIIGIIIIIHIIIII      XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:85
```

The BAM Format

...sorta

ZYMO

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

The Challenge

Identify variants with potential clinical significance based on the genomic read data

Sequencers all have non-zero error rates

Normal human variation is common at many sites

The human genome is about 3.3 billion bases

- Very small needle
- Very large haystack
- Several pieces of wire in there as well

Our Tools

Reference genome

Variation data from many ethnically diverse individuals

The ability to sequence family members and know if they are affected

Error modeling for our sequencing platform of choice

Coverage

- We don’t just sequence a single position once, we sequence it several times per run

Reads at a Site

- Reading each base multiple times increases the chance of reading from both copies of the chromosome
- High depth of coverage makes error modeling simple
  - Particularly for Illumina data
- Large numbers of reads can be used to ensure that the variant is not artifact

ZYMO

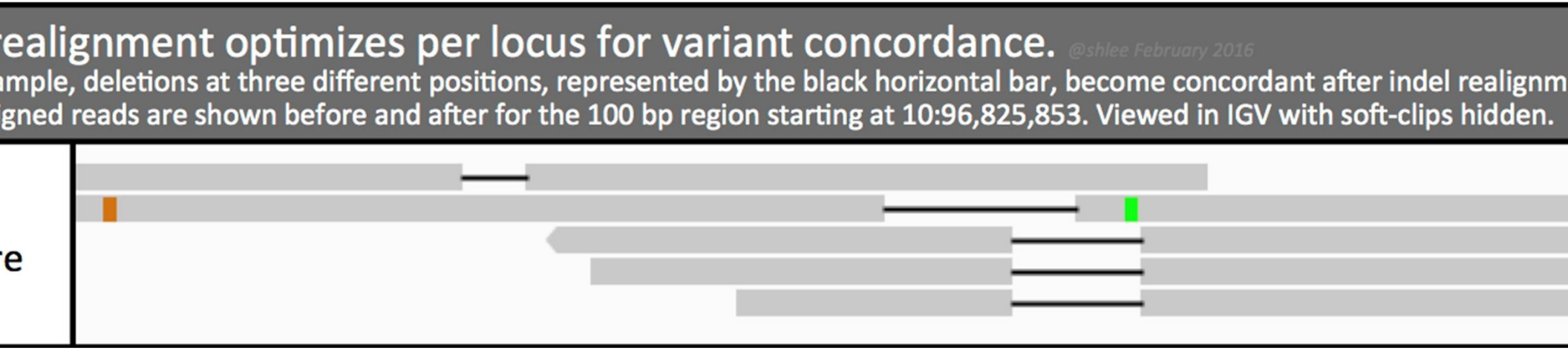
Reads at a Site

- Reading each base multiple times increases the chance of reading from both copies of the chromosome
- High depth of coverage makes error modeling simple
  - Particularly for Illumina data
- Large numbers of reads can be used to ensure that the variant is not artifact

Reads at a Site

- Reading each base multiple times increases the chance of reading from both copies of the chromosome
- High depth of coverage makes error modeling simple
  - Particularly for Illumina data
- Large numbers of reads can be used to ensure that the variant is not artifact

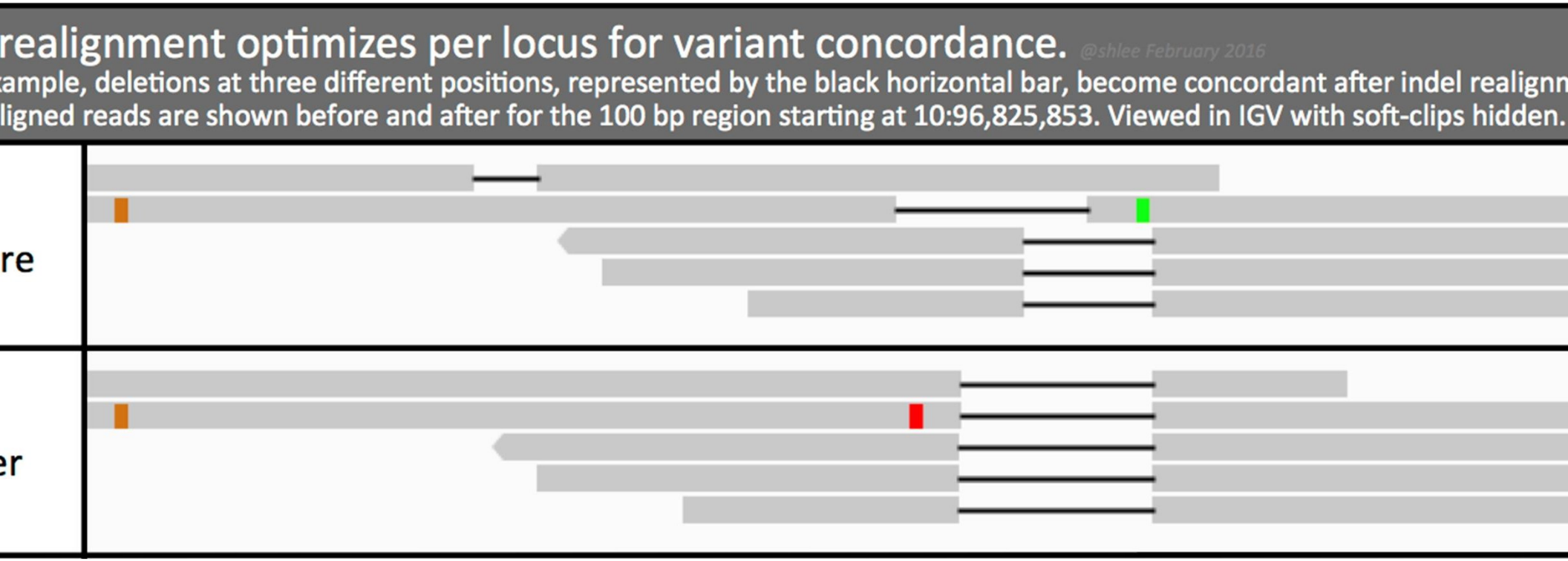
Dealing with Indels



- Aligning individual reads is an iterative process
- Information gained during previous/subsequent read alignments is not available during current alignment
- Determining where an insertion or deletion starts and ends can be difficult with only a single read
  - Especially difficult if that read begins or ends near the site

ZYMO

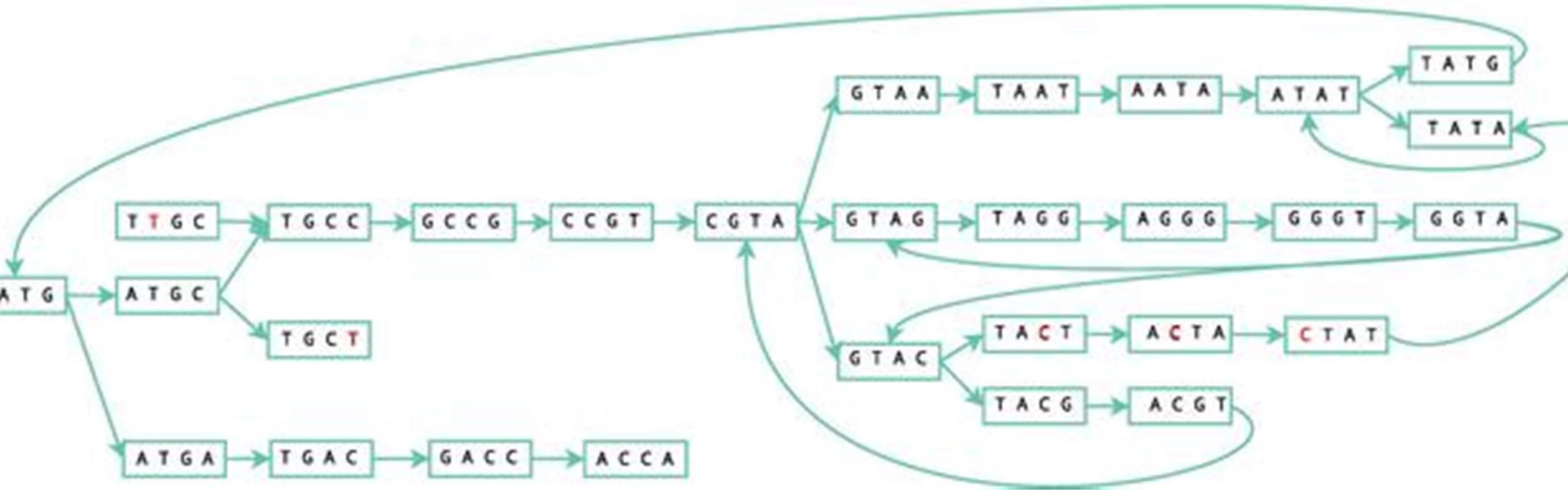
Dealing with Indels



- Use the now-aligned reads in the region to inform realignment around the insertion/deletion site
- Still carries some burden of reference bias…

ZYMO

Local Reassembly (de Bruijn)



- Usually conducted on only a few hundred base segment of the genome
- Very fast/efficient operation free of reference bias

ZYMO

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis *Becomes Dynamic*: Identifying *Tumor* Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

The Challenge

Identify variants between healthy tissue and tumor in a patient

Sequencers still have non-zero error rates

Tumors are not homogeneous and can have high mutation rates

Tumors may rearrange their genomes significantly

- Lose copies of genes
- Gain additional copies of other genes

Tumors have selective pressures from within and without

Great progress, but still very much an open challenge

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

The Challenge

Cells use regulation of RNA production as a quick control/response for many different stimuli

Splicing can cause the same gene to have multiple transcripts

A few RNA molecules are present in very high numbers

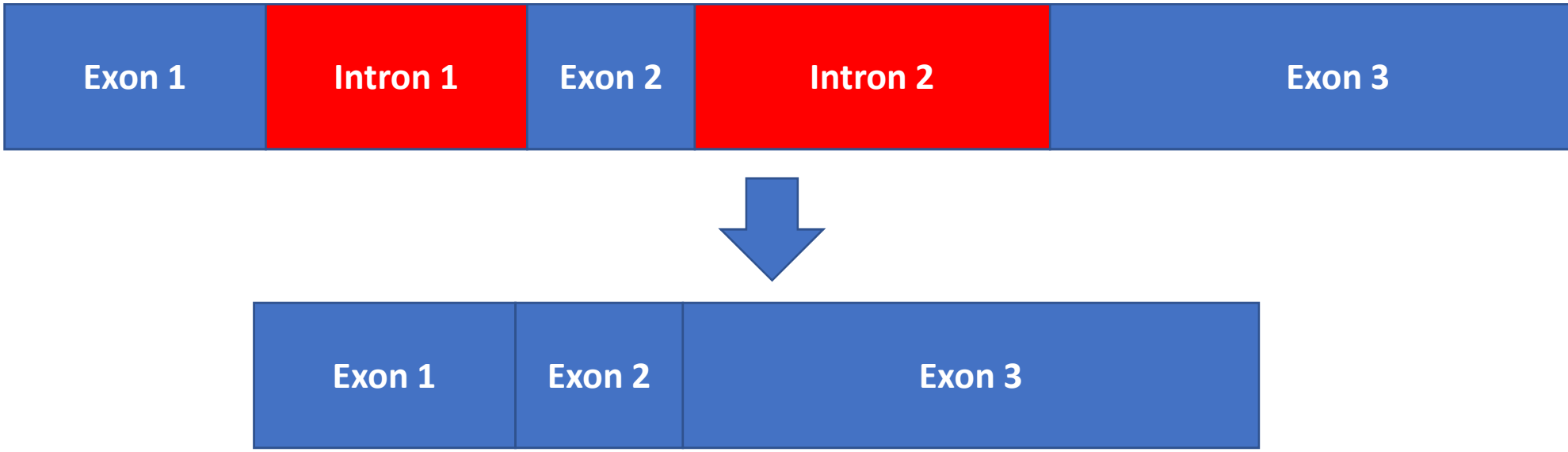
Some RNA molecules can be very important, but present in low amounts

- Great example: Transcription factors… the proteins that control how much of each RNA molecule is made

Our budgets will only allow so many reads per sample per study

- Sequencing becomes a sampling problem

RNA Splicing



The Solutions

Seed based alignment can solve the issue of splicing

- Splicing turns out to create a deletion problem, something we solved previously

A few RNA molecules are present in very high numbers

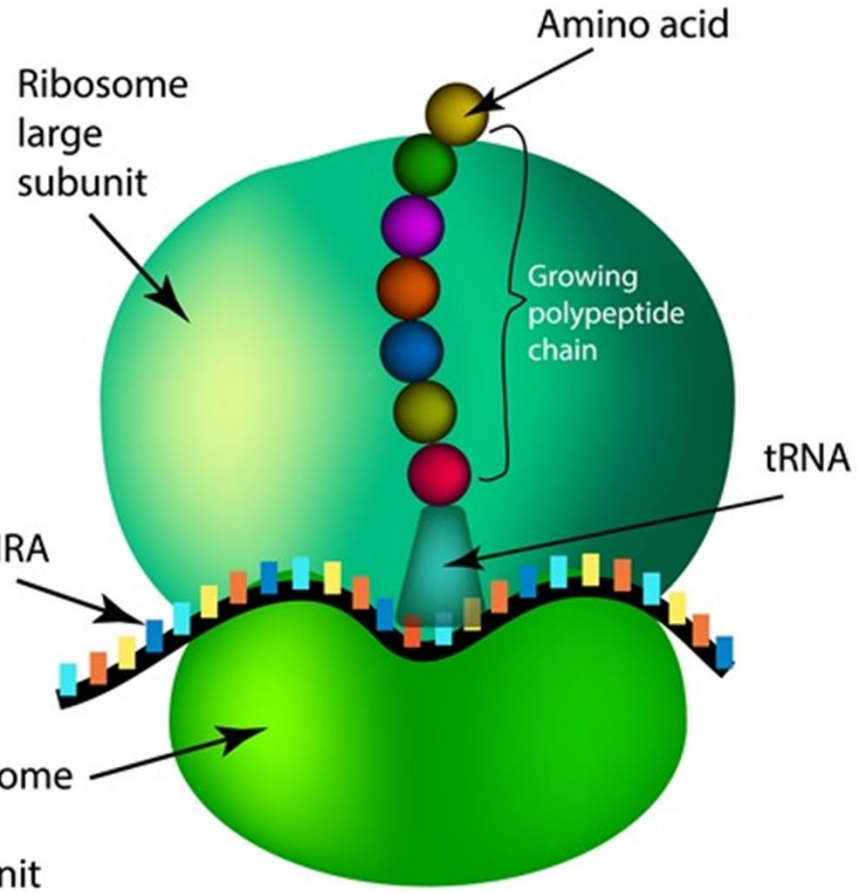
Some RNA molecules can be very important, but present in low amounts

- Great example: Transcription factors… the proteins that control how much of each RNA molecule is made

Our budgets will only allow so many reads per sample per study

- Sequencing becomes a sampling problem

Some RNA Molecules are Present in High Amounts



- Ribosomal RNA is present in every cell
  - Often about 70% of the cell's total RNA
  - Highly stabile
  - Present in bacteria and nucleated cells
- Other cells may have other over-represented RNA molecules

The Solutions

Seed based alignment can solve the issue of splicing

- Splicing turns out to create a deletion problem, something we solved previously

Use molecular techniques to remove unwanted molecules before sequencing

Some RNA molecules can be very important, but present in low amounts

- Great example: Transcription factors… the proteins that control how much of each RNA molecule is made

Our budgets will only allow so many reads per sample per study

- Sequencing becomes a sampling problem

The Solutions

Seed based alignment can solve the issue of splicing

- Splicing turns out to create a deletion problem, something we solved previously

Use molecular techniques to remove unwanted molecules before sequencing

Still an open question, requires better statistical methods or deeper sequencing

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

- Bacteria and mammalian cells have ribosomal RNA that is very different between groups and similar within groups
  - Combination of variable and constant regions
- Microbes are influenced by environment
- Microbes influence environmen
- Microbes transfer between hosts

ZYMO

The Challenge

Identify the microbes present in a sample with resolution down to genus or species

Quantify functional groups of microbes present in a sample

- Such as microbes able to oxidize a specific molecule

Quantify the ratios of different microbe families present in a sample

The Solutions

If just genus and/or species needs to be identified, we can use the ribosomal RNA as a “barcode” for different microbes

Improved microbiological techniques can be used to avoid lysis bias

De Bruijn graph-based methods can be used to reconstruct all possible ribosomal RNA sequences in a sample

Overview

The Basic Materials of Life

Obtaining Material for Sequencing

Reading that Sequence

Drafting a Genome

Aligning to a Genome

Static Analysis: Identifying Variants

Dynamic Analysis: Looking at RNA Expression

Diversity Analysis: The Microbiome

Future and Conclusions

ZYMO

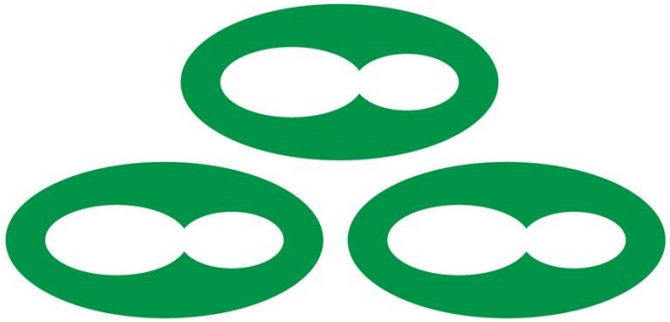
Future Applications

Increase speed of reading to allow for rapid diagnostics and rapid identification of contaminants

Increase sensitivity for non-invasive prenatal testing or early cancer detection using only small samples of blood

Understanding the epigenome

Questions?



ZYMO RESEARCH

The Beauty of Science is to Make Things Simple®

 info@zymoresearch.com

 www.zymoresearch.com

 Toll Free: (888) 882-9682