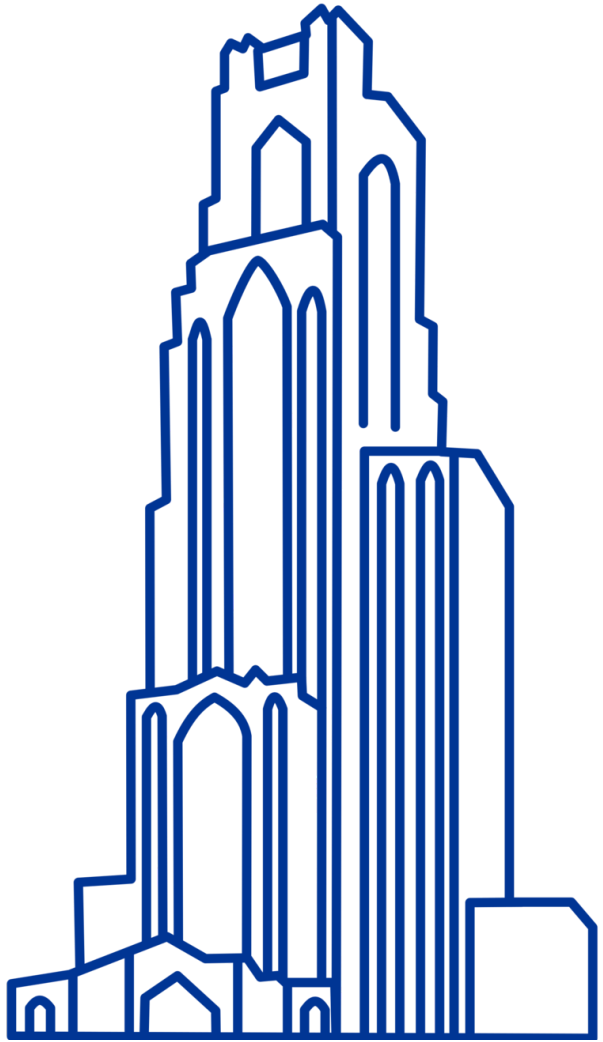
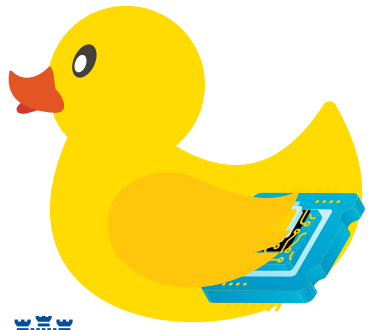


Computational Biology
(BIOSC 1540)

Lecture 03:
Quality control

Sep 3, 2024



University of Pittsburgh

Announcements

- Assignment 01 is due Thursday at 11:59 pm.
- Assignment 02 will be posted on Thursday.
- The last module will be Scientific Python instead of Special Interests
  - Lectures are optional
  - No assigned homework
  - Optional assignments for extra credit?
  - Please complete the Kaggle intro and Python beforehand if you want to participate
- Optional final is Monday, Dec 16 at 10 AM
  - If you do worse, I will not count the final

Lectures

Bioinformatics >

Computational structural biology >

Scientific python v

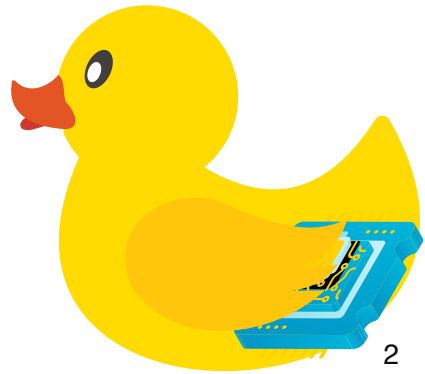
Lecture 19: NumPy

Lecture 20: Polars

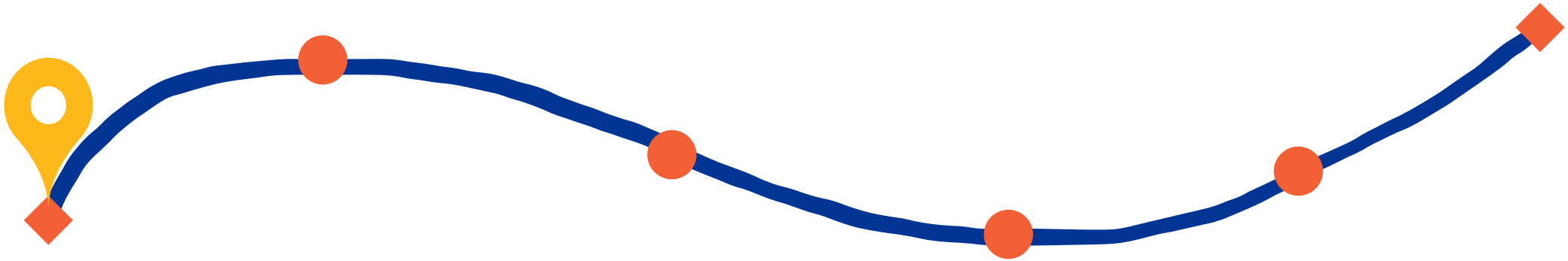
Lecture 21: Matplotlib

Lecture 22: SciPy

Lecture 23: Scikit-learn



After today, you should be able to

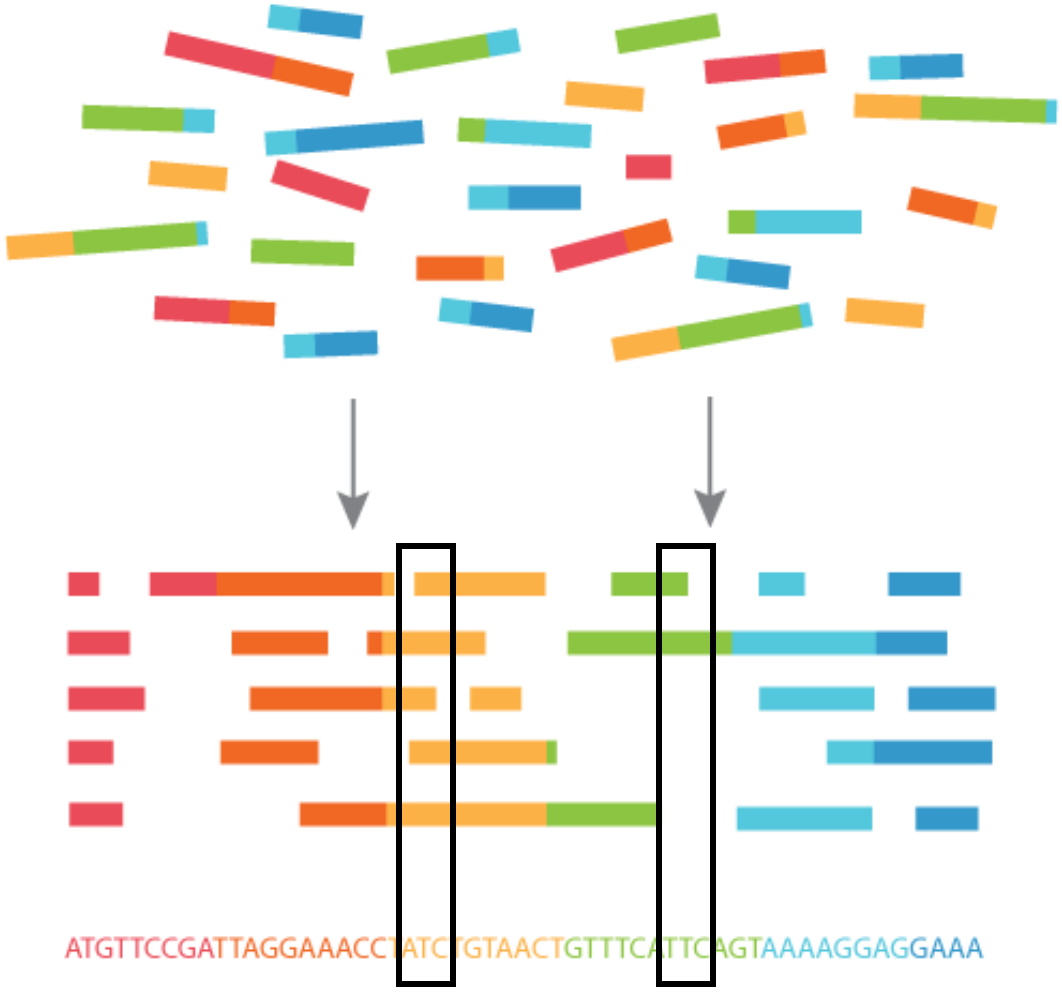


1. **Explain the basic concepts and importance of genome assembly.**
2. Interpret FASTA and FASTQ file formats and their role in storing sequences.
3. Perform and interpret quality control on reads using FastQC.
4. Identify common quality issues in sequencing data and explain their impacts.
5. Describe the process and importance of sequence trimming and filtering.

Sequencing provides short, overlapping reads of DNA

Genome assembly is the process of combining our sequencing reads into a continuous DNA sequence

Having multiple fragments that contain the same portion of the sequence improves our coverage



Coverage: High Low

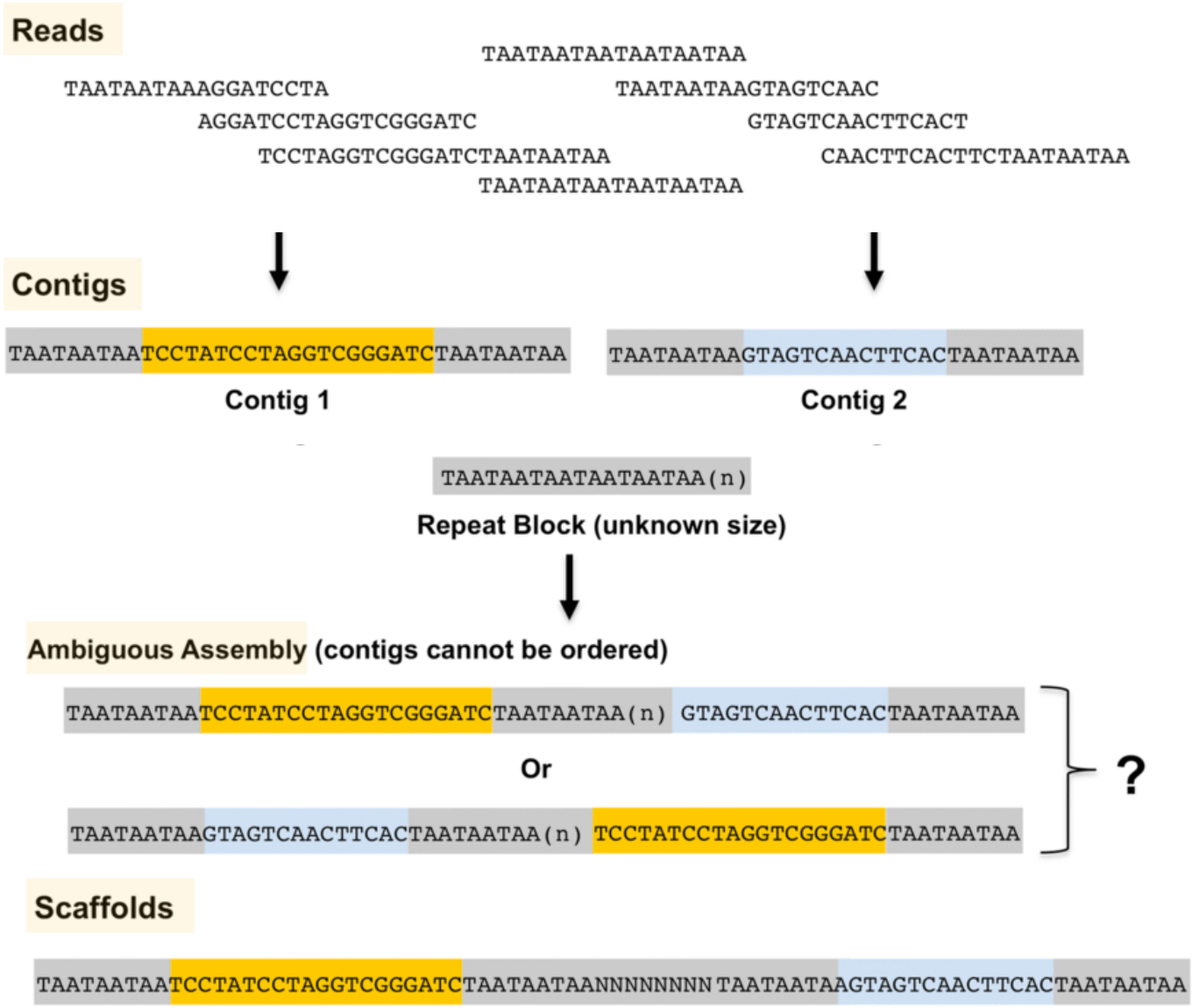
Assembly terminology

Raw sequences coming from our experiments

Continuous stretches of DNA sequence from overlapping sequencing reads

Connecting contigs in an unknown order

Multiple contigs with estimated gaps



Let's build the original sequence from small fragments with copies and errors

Original sequence: **5'- GTACCTAG -3'**

Fragments

1. GTACC
2. TACCT
3. ACCTA
4. CCTAG

Potential copies?

1. GTACC
2. TACCT
3. ACCTA
4. CCTAG

Errors

1. GTACG
2. ACCTT

Hard, right? This is what we ask of computational biologists working in **genome assembly**

Assembly quality metrics

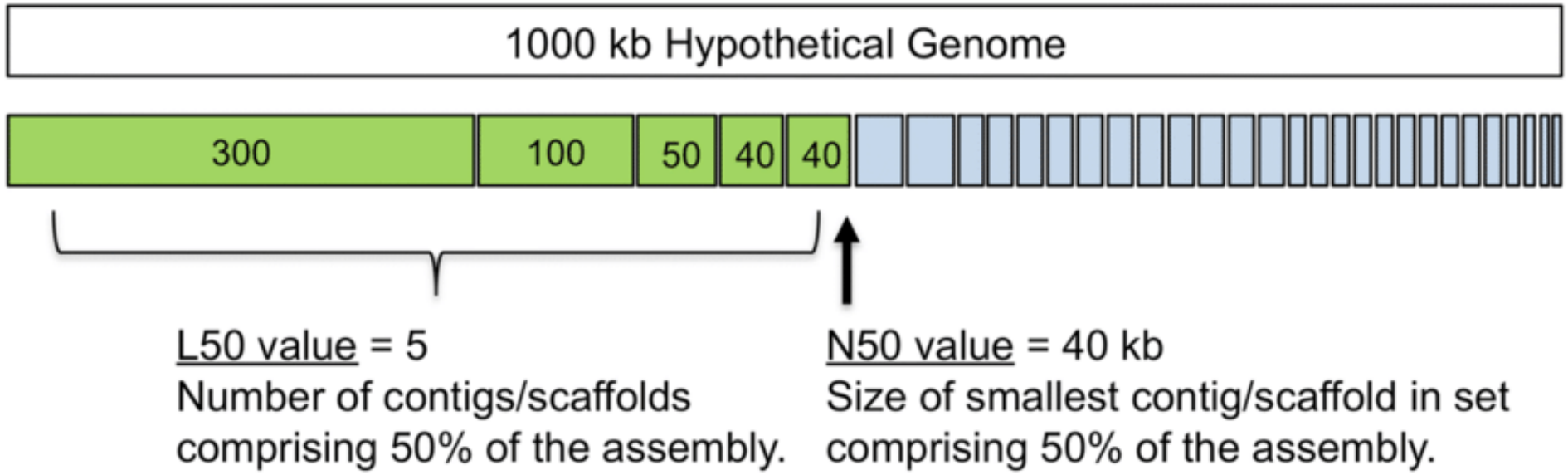
First, sort all contigs from longest to shortest

L50 number of contigs whose combined length is at least 50%

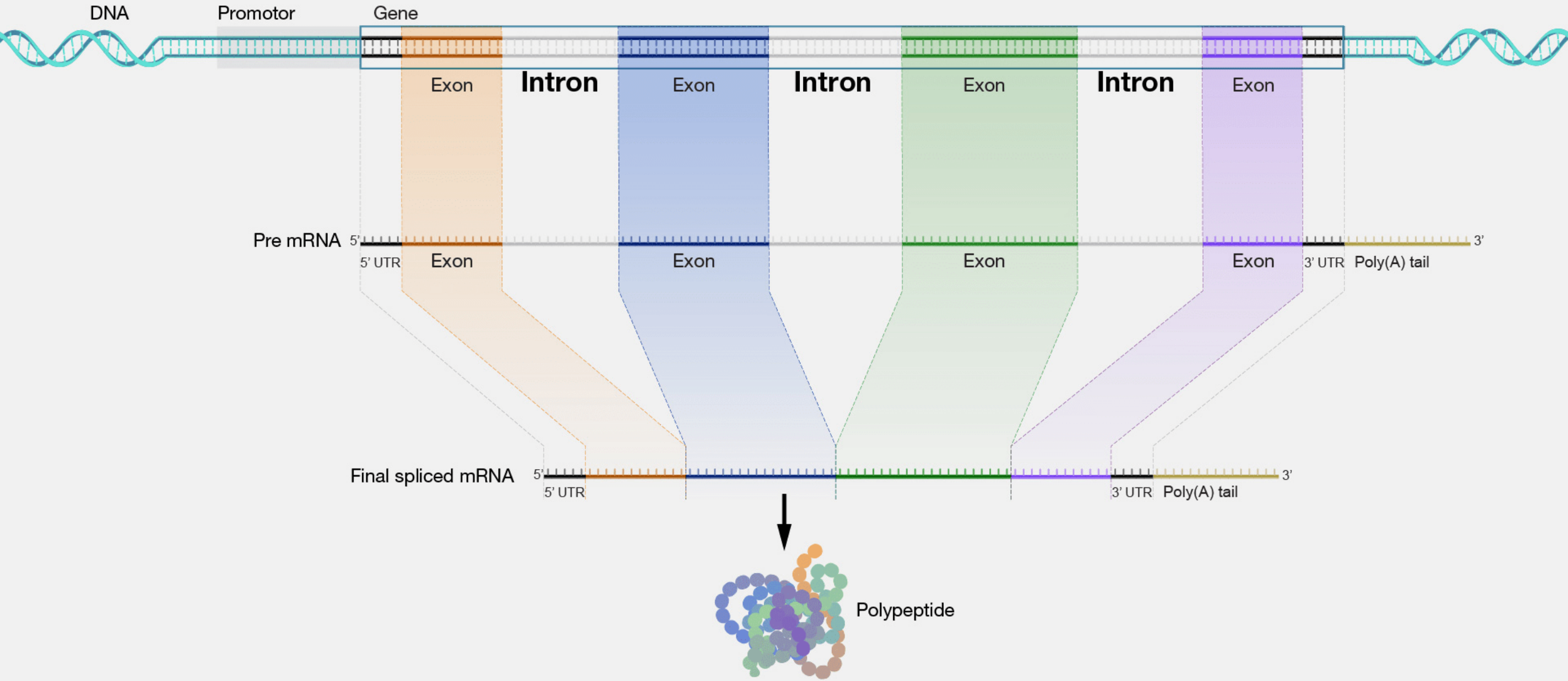
(Lower is better.)

N50 is the sequence length of the shortest contig at 50% of the total genome length

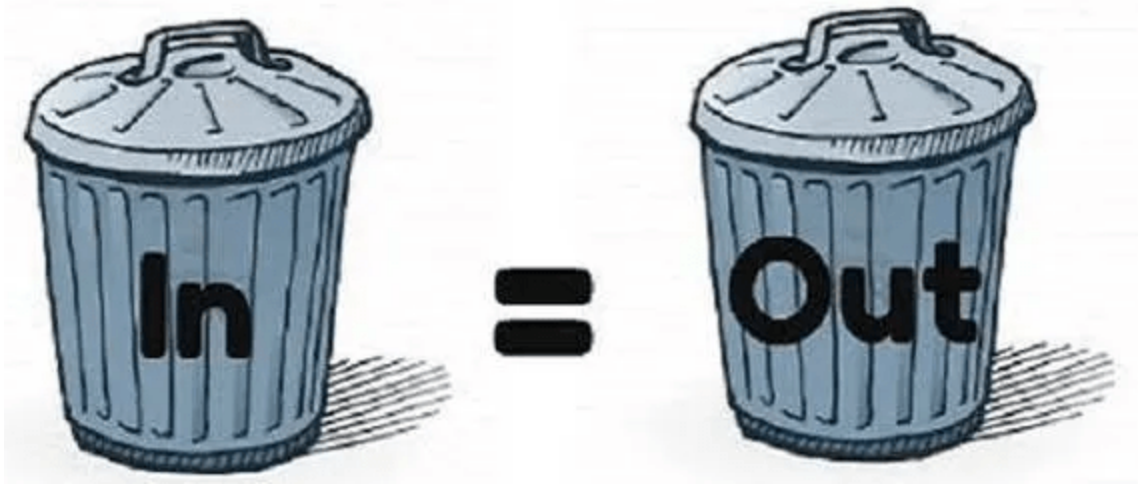
(Higher is better.)



Then we can annotate our genome

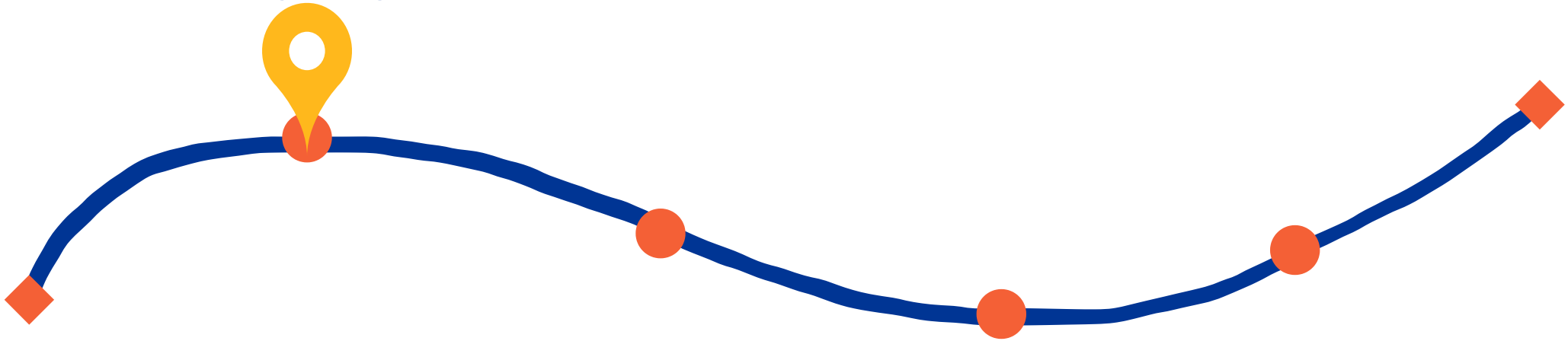


Cleaning our sequencing reads improves our assembly



Garbage in, garbage out

After today, you should be able to



1. Explain the basic concepts and importance of genome assembly.
2. **Interpret FASTA and FASTQ file formats and their role in storing sequences.**
3. Perform and interpret quality control on reads using FastQC.
4. Identify common quality issues in sequencing data and explain their impacts.
5. Describe the process and importance of sequence trimming and filtering.

Sequences are stored in FASTA files

DNA

```
>BTBSCRYR
tgcaccaaacatgtctaaagctggaaccaaaattactttctttgaagacaaaaactttca
aggccgccactatgacagcgattgcgactgtgcagatttccacatgtacctgagccgctg
caactccatcagagtggaaggaggcacctgggctgtgtatgaaaggcccaattttgctgg
gtacatgtacatcctaccccggggcgag
```

Protein

```
>crab_anapl ALPHA CRYSTALLIN B CHAIN (ALPHA(B)-CRYSTALLIN)
MDITIHNPLIRRPLFSWLAPSRIFDQIFGEHLQESELLPASPSLSPFLMR
SPIFRMPSWLETGLSEMRLEKDKFSVNLDVKHFSPEELKVKVLGDMVEIH
GKHEERQDEHGFIAREFNRKYRIPADVDPLTITSSLSLDGVLTVSAPRKQ
SDVPERSIPITREEKPAIAGAQRK
```

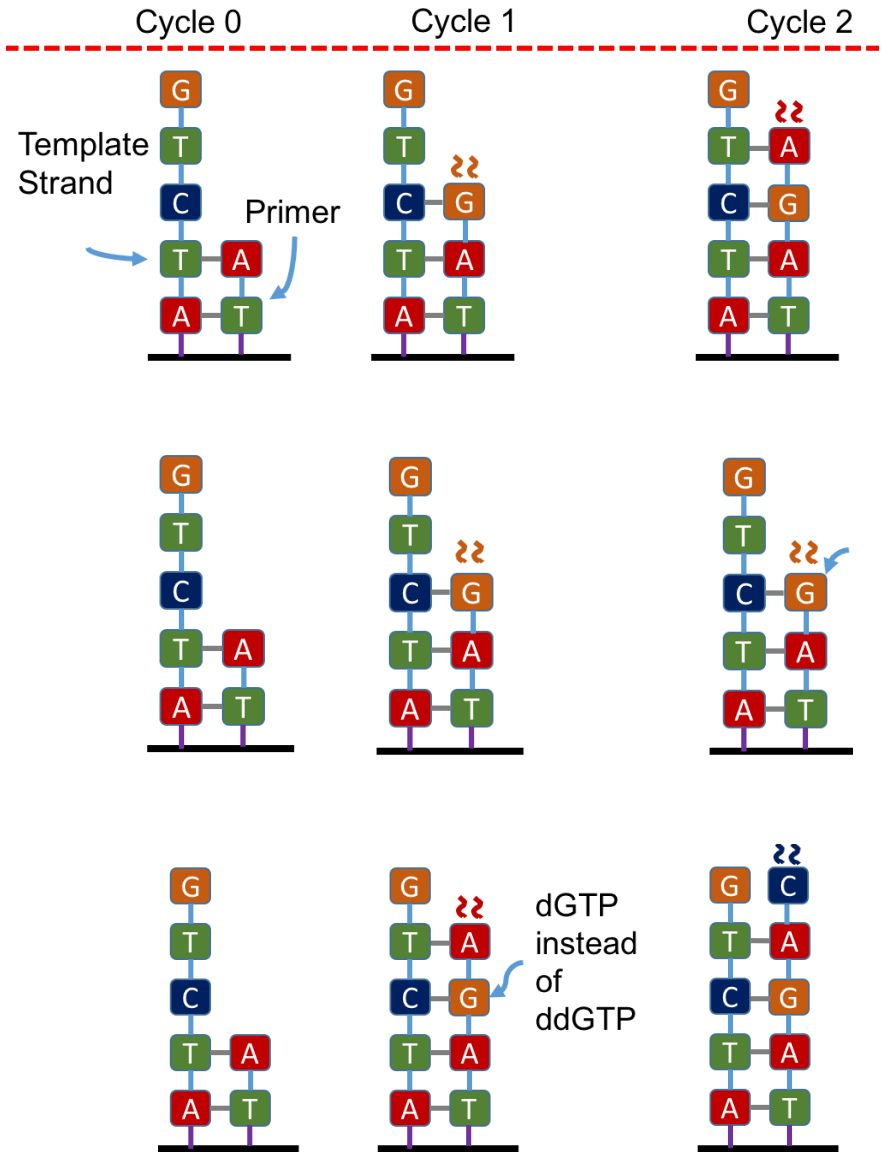
- One line starts with a ">" and a sequence identification code.
  - It is optionally followed by a description of the sequence.
- One or more lines containing the sequence itself.

However, base calling is not perfect

Normal sequencing by synthesis

Lagging synthesis by failure to remove blocking fluorophore

Leading synthesis by addition of dNTP instead of ddNTP



Signal cross-talk degrades quality

Clean

Noisy

ML models and algorithms compute the probability of error (i.e., quality)

FASTQ files store sequence and quality

Quality scores measure the probability that a base is called incorrectly

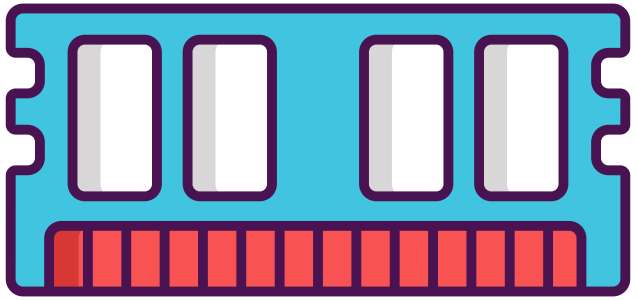
@Identifier
Sequence
+
Per-nucleotide quality

```
1 @HWI-M01876:76:000000000-AF16W:1:1101:10853:1000 1:N:0:CGTGACAGAT
2 NTGTACTTCATCCGAAACTCGTGCTCATCTCTGCTCAGATCGGAAGAGCACACGTCTGAACTCCAG
3 +
4 #8ABCFGGGFCEDCFGGGGGGGFFCGEFGGGGGGFGGGGGGGGDEFGGGGGGGGGGGGGGGGFFF
5 @HWI-M01876:76:000000000-AF16W:1:1101:16471:1000 1:N:0:CGTGAACTTG
6 NTTCCAGATATTCGATGCATGTGCCGCTCCTGTCGGAGATCGGAAGAGCACACGTCTGAACTCCAG
7 +
8 #8BCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGGFGGGGGGGGGGGGGGG
```

What does "G" or "8" quality mean?

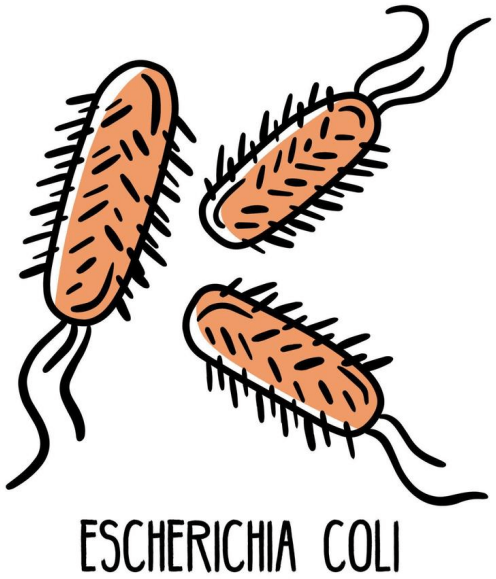
ASCII-encoded probabilities

We need to store millions upon millions of floats (e.g., 0.92829) per nucleotide



One million float32 values are about 3.8 MB

Seems small, but one *E. coli* genome is ~5 million base pairs and we have multiple copies



ASCII characters require ~1/4 the memory, and we already have to store nucleotides

Hexadecimal characters have an associated int

Phred quality (Q) is the integer associated with the ASCII symbol

$$P(Q) = 10^{-Q/10}$$

Probability that an error occured

The smallest value 33, because lower hexadecimal cannot be rendered on screen

Dec	Char	Dec	Char	Dec	Char	Dec	Char
0	NUL (null)	32	SPACE	64	@	96	`
1	SOH (start of heading)	33	!	65	A	97	a
2	STX (start of text)	34	"	66	B	98	b
3	ETX (end of text)	35	#	67	C	99	c
4	EOT (end of transmission)	36	$	68	D	100	d
5	ENQ (enquiry)	37	%	69	E	101	e
6	ACK (acknowledge)	38	&	70	F	102	f
7	BEL (bell)	39	'	71	G	103	g
8	BS (backspace)	40	(	72	H	104	h
9	TAB (horizontal tab)	41	)	73	I	105	i
10	LF (NL line feed, new line)	42	*	74	J	106	j
11	VT (vertical tab)	43	+	75	K	107	k
12	FF (NP form feed, new page)	44	,	76	L	108	l
13	CR (carriage return)	45	-	77	M	109	m
14	SO (shift out)	46	.	78	N	110	n
15	SI (shift in)	47	/	79	O	111	o
16	DLE (data link escape)	48	0	80	P	112	p
17	DC1 (device control 1)	49	1	81	Q	113	q
18	DC2 (device control 2)	50	2	82	R	114	r
19	DC3 (device control 3)	51	3	83	S	115	s
20	DC4 (device control 4)	52	4	84	T	116	t
21	NAK (negative acknowledge)	53	5	85	U	117	u
22	SYN (synchronous idle)	54	6	86	V	118	v
23	ETB (end of trans. block)	55	7	87	W	119	w
24	CAN (cancel)	56	8	88	X	120	x
25	EM (end of medium)	57	9	89	Y	121	y
26	SUB (substitute)	58	:	90	Z	122	z
27	ESC (escape)	59	;	91	[	123	{
28	FS (file separator)	60	<	92	\	124	|
29	GS (group separator)	61	=	93	]	125	}
30	RS (record separator)	62	>	94	^	126	~
31	US (unit separator)	63	?	95	_	127	DEL

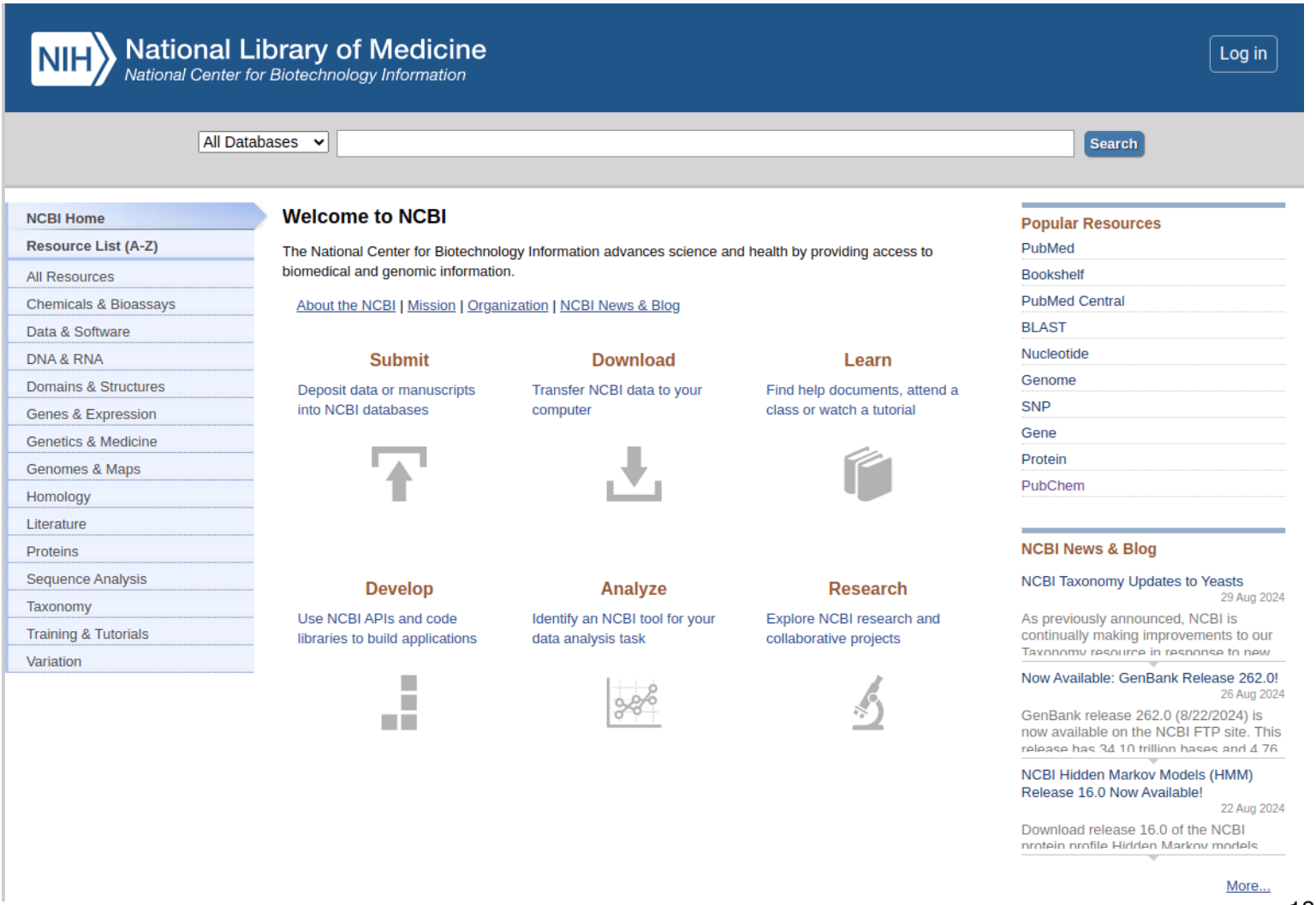
$$P(!) = 10^{-(33-33)/10} = 1.0$$

$$P(\#) = 10^{-(35-33)/10} \approx 0.63$$

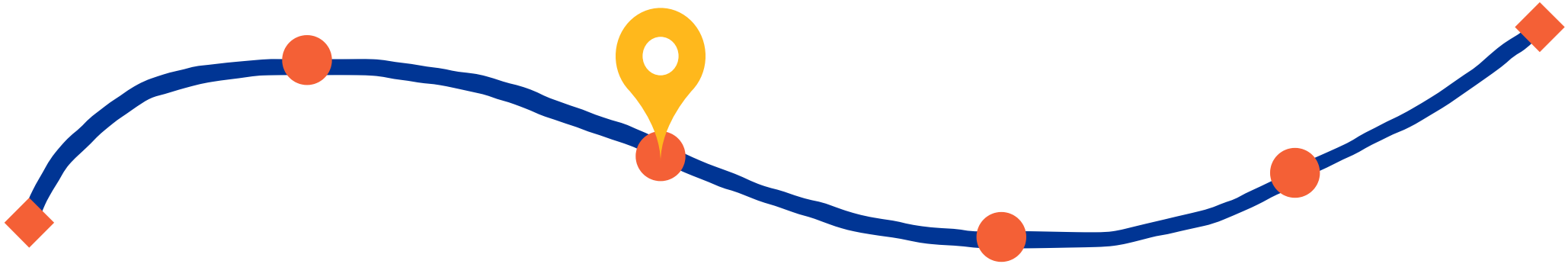
Sequencing runs store millions of FASTQ entries

Scientists will deposit FASTQ files into NIH databases

- **GeneBank** for genomic sequences
- **Sequence read archive (SRA)** for sequencing data
- **RefSeq** for reference genomes
- **BioProject** for curated resources for a specific project
- Many more



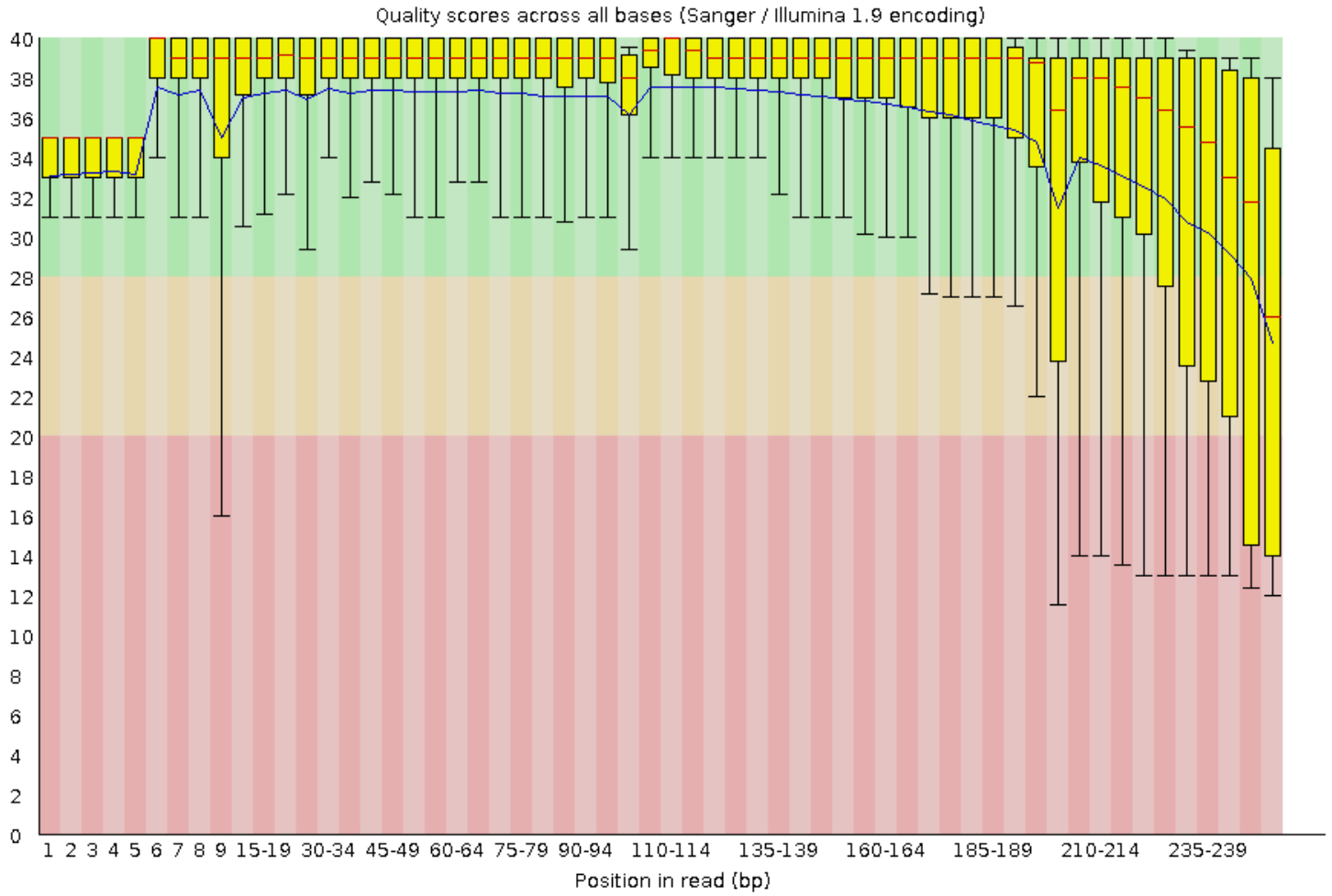
After today, you should be able to



1. Explain the basic concepts and importance of genome assembly.
2. Interpret FASTA and FASTQ file formats and their role in storing sequences.
3. **Perform and interpret quality control on reads using FastQC.**
4. **Identify common quality issues in sequencing data and explain their impacts.**
5. Describe the process and importance of sequence trimming and filtering.

Per base sequence quality

Box and whisker plot of base-call accuracy

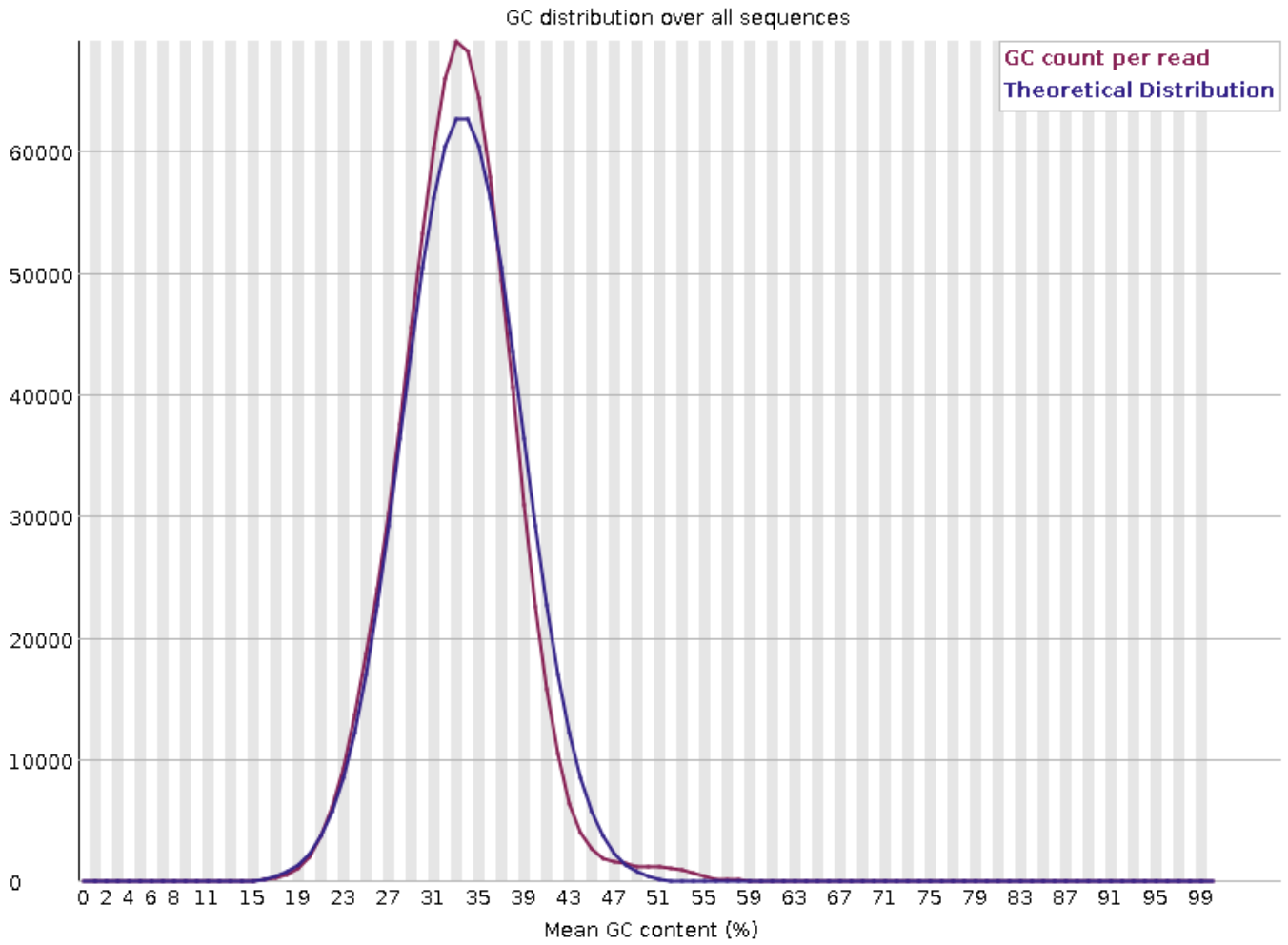


Excellent

Good

Poor

Per sequence GC content

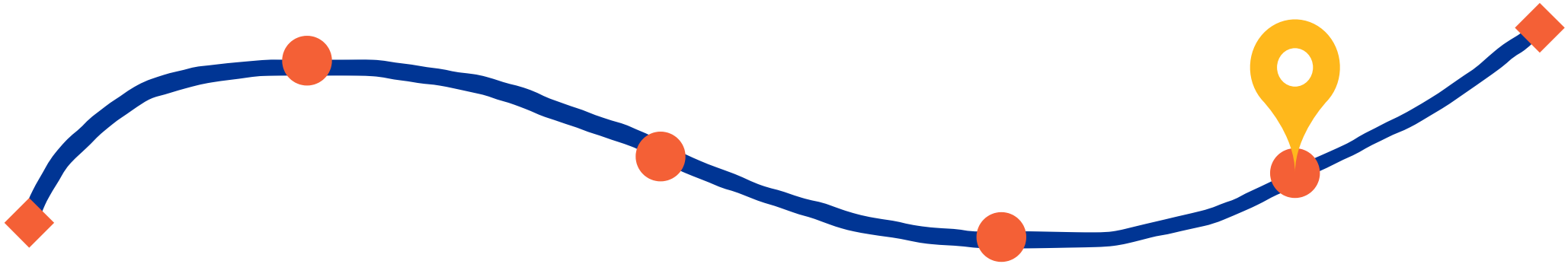


Strong deviations from normal distribution could indicate contamination

Activity: Quality control

TopHat Questions

After today, you should be able to



1. Explain the basic concepts and importance of genome assembly.
2. Interpret FASTA and FASTQ file formats and their role in storing sequences.
3. Perform and interpret quality control on reads using FastQC.
4. Identify common quality issues in sequencing data and explain their impacts.
5. **Describe the process and importance of sequence trimming and filtering.**

Activity: Read trimming

With cleaned data, we can now assemble our reads into contigs/scaffolds

Before the next class, you should

Lecture 03:
Quality control

Lecture 04:
De novo assembly

Today

Thursday

- Finish Assignment 01, which is due Thursday at 11:59 pm.