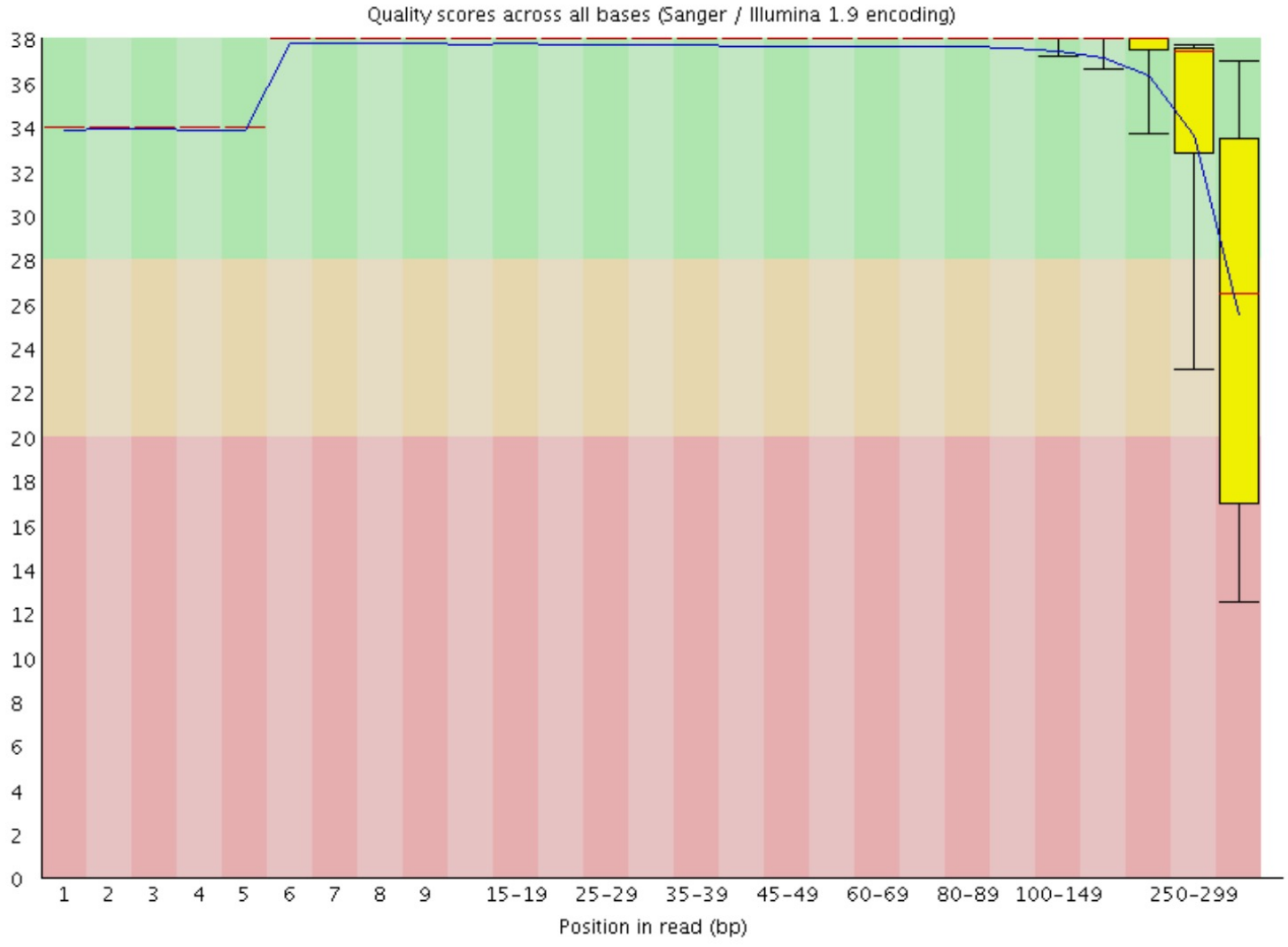
Mapping and Alignment

Philip T.L.C. Clausen

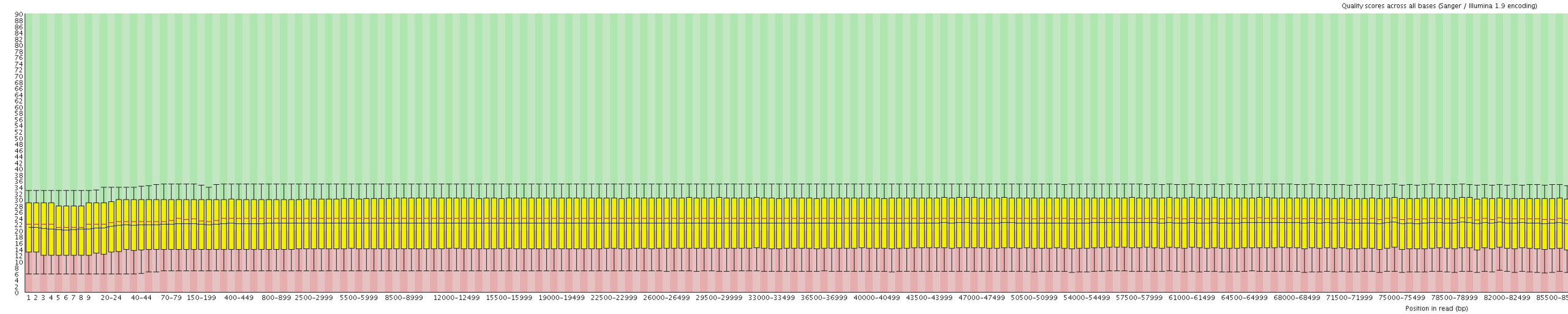
Quick recap

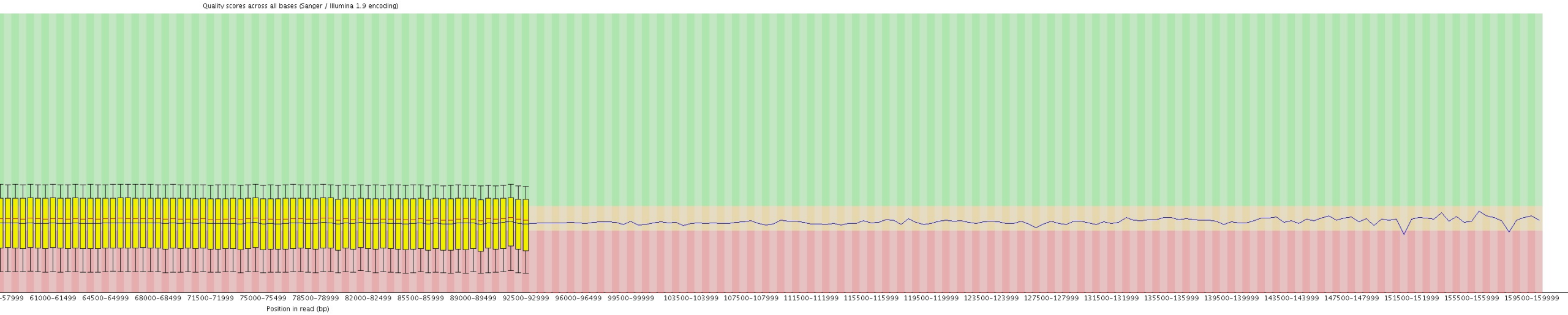
- Trimming of Illumina reads
- Trimming of ONT reads
- De Bruijn Graph assembly

Illumina fastqc report

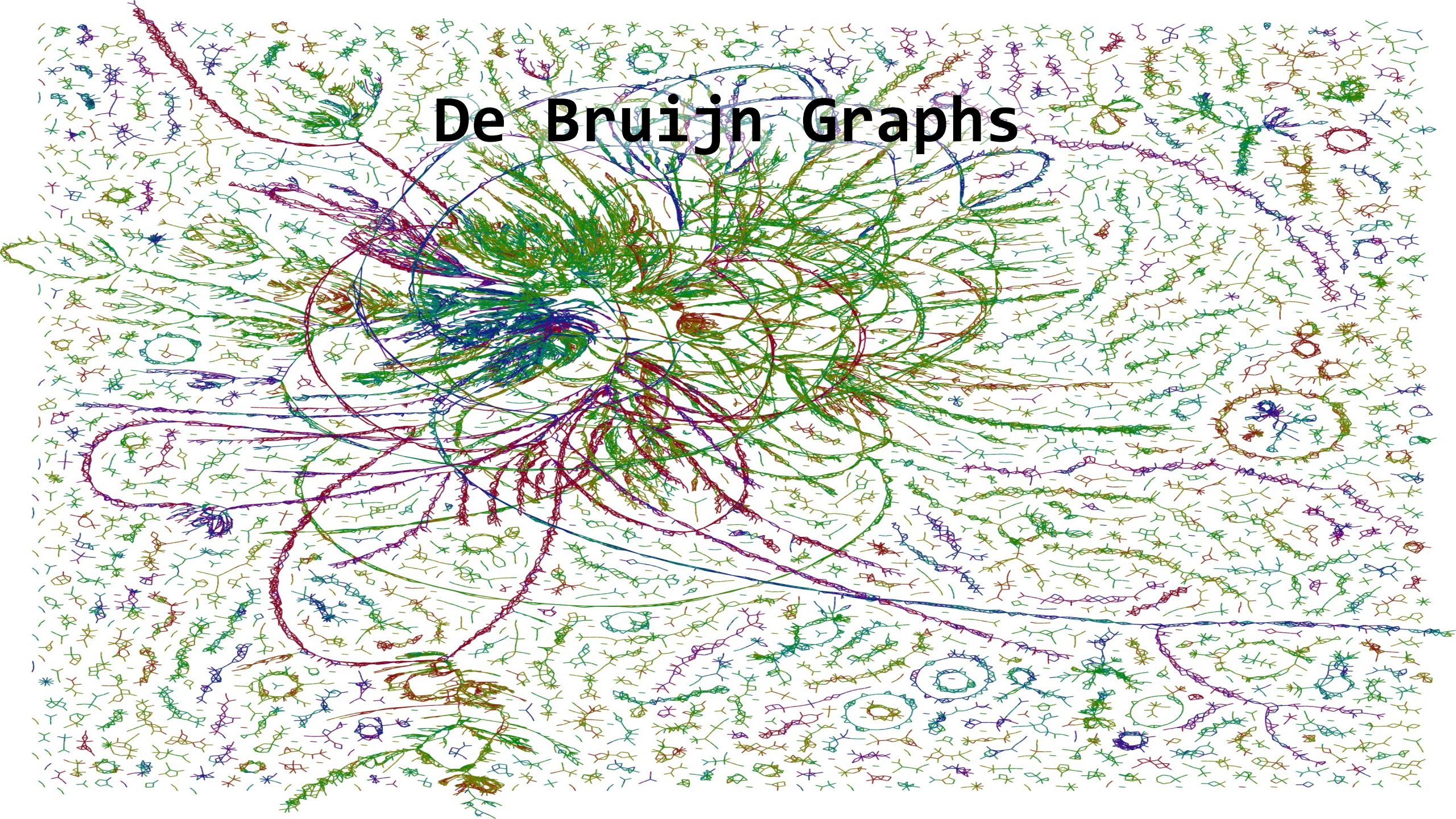


ONT fastqc report (16 threads)

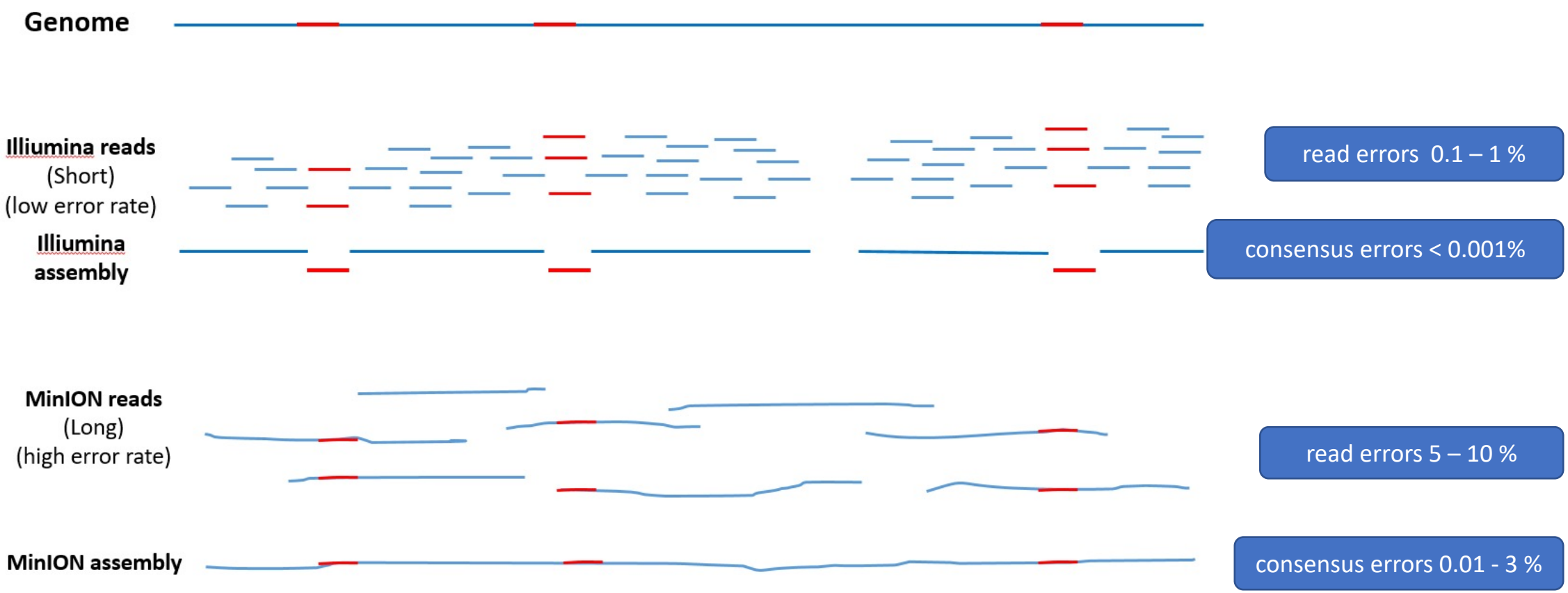




De Bruijn Graphs



Technology impacts



Repeat area (rRNA, IS, homologue genes etc..)

Mapping and Alignment

Philip T.L.C. Clausen

Content:

- Mapping and Chaining
- OLC and Alignment
- Polish and ConClave

Goal:

- Know what the button does
- Know which button to push
- Join the competition

Why:

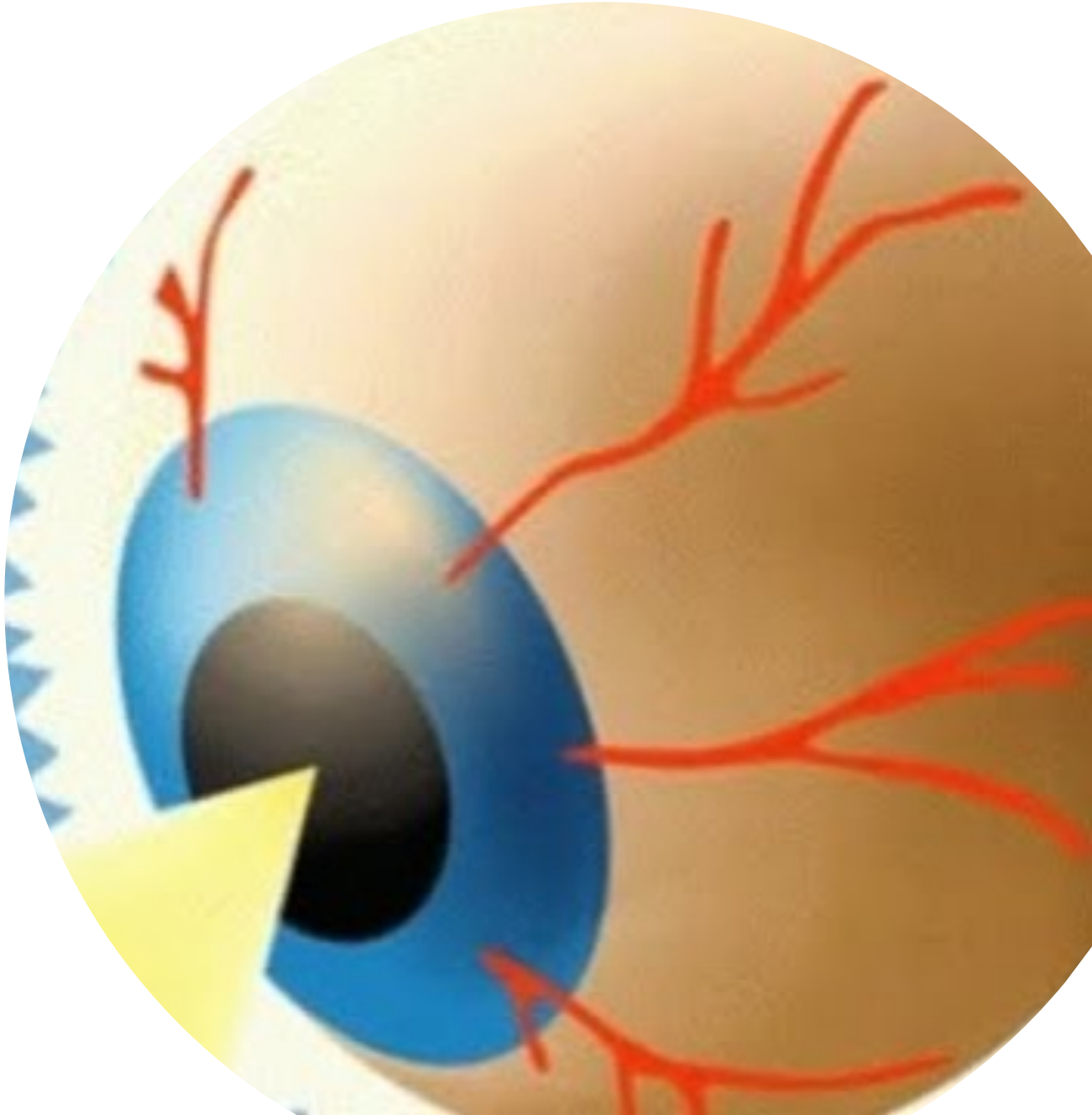
- The computer does only what we commands it to do
- We command the computer through programs
- The programs contains algorithms and methods

Mapping and alignment

I want to compare two or more sequences,

and so what…

1951



1966

Number of possible alignments

$$\binom{m+n}{n} = \frac{(m+n)!}{m!\,n!}$$


150 X 150

1970

Saul B. Needleman

$$D(i,j) = max \begin{cases} D(i+1,j+1) + S(T_i, Q_j) \\ P(i,j) \\ Q(i,j) \end{cases}$$

$$P(i,j) = max \begin{cases} D(i+1,j) + w_1 \\ P(i+1,j) + U \end{cases}$$

$$Q(i,j) = max \begin{cases} D(i,j+1) + w_1 \\ Q(i,j+1) + U \end{cases}$$

Christian D. Wunsch

Simplified NW

Saul B. Needleman

$$D(i,j) = max \begin{cases} D(i+1, j+1) + S(T_i, Q_j) \\ D(i+1, j) + W \\ D(i, j+1) + W \end{cases}$$

Christian D. Wunsch

What if we made the two sequences form a matrix?

$$D(i,j) = max \begin{cases} D(i+1, j+1) + S(T_i, Q_j) \\ D(i+1, j) + W \\ D(i, j+1) + W \end{cases}$$

$$S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

M = 1 , P = -1 , W = -3

Q\T	A	C	G	T	-
A					
G					
T					
T					
-					0

What if we made the two sequences form a matrix?

$$D(i,j) = max \begin{cases} D(i+1, j+1) + S(T_i, Q_j) \\ D(i+1, j) + W \\ D(i, j+1) + W \end{cases}$$

$$S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

M = 1 , P = -1 , W = -3

Q\T	A	C	G	T	-
A					-12
G					-9
T					-6
T					-3
-	-12	-9	-6	-3	0

What if we made the two sequences form a matrix?

$$D(i,j) = max \begin{cases} D(i+1, j+1) + S(T_i, Q_j) \\ D(i+1, j) + W \\ D(i, j+1) + W \end{cases}$$

$$S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

$\mathbf{M = 1}$, P = -1 , W = -3

$$D(4,4)$$

$$= max \begin{cases} 0 + S(T,T) \\ -3 + (-3) \\ -3 + (-3) \end{cases}$$

$$= max \begin{cases} 0 + 1 \\ -6 \\ -6 \end{cases} = 1$$

Q\T	A	C	G	T	-
A					-12
G					-9
T					-6
T					-3
-	-12	-9	-6	-3	0

What if we made the two sequences form a matrix?

$$D(i,j) = max \begin{cases} D(i+1, j+1) + S(T_i, Q_j) \\ D(i+1, j) + W \\ D(i, j+1) + W \end{cases}$$

$$S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

$\mathbf{M = 1}$, P = -1 , W = -3

$$D(4,4)$$

$$= max \begin{cases} 0 + S(T,T) \\ -3 + (-3) \\ -3 + (-3) \end{cases}$$

$$= max \begin{cases} 0 + 1 \\ -6 \\ -6 \end{cases} = 1$$

Q\T	A	C	G	T	-
A					-12
G					-9
T					-6
T				1	-3
-	-12	-9	-6	-3	0

What if we made the two sequences form a matrix?

$$D(i,j) = max\begin{cases} D(i+1,j+1) + S(T_i, Q_j) \\ D(i+1,j) + W \\ D(i,j+1) + W \end{cases}$$

$$S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

$\mathbf{M = 1}$, P = -1 , W = -3

$$D(4,3)$$

$$= max\begin{cases} -3 + S(T,G) \\ -6 + (-3) \\ 1 + (-3) \end{cases}$$

$$= max\begin{cases} -3 + (-1) \\ -9 \\ -2 \end{cases} = -2$$

Q\T	A	C	G	T	-
A					-12
G					-9
T					-6
T				1	-3
-	-12	-9	-6	-3	0

What if we made the two sequences form a matrix?

$$D(i,j) = max\begin{cases} D(i+1,j+1) + S(T_i, Q_j) \\ D(i+1,j) + W \\ D(i,j+1) + W \end{cases}$$

$$S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

$\mathbf{M = 1}$, P = -1 , W = -3

$$D(4,3)$$

$$= max\begin{cases} -3 + S(T,G) \\ -6 + (-3) \\ 1 + (-3) \end{cases}$$

$$= max\begin{cases} -3 + (-1) \\ -9 \\ -2 \end{cases} = -2$$

Q\T	A	C	G	T	-
A					-12
G					-9
T					-6
T			-2	1	-3
-	-12	-9	-6	-3	0

What if we made the two sequences form a matrix?

$$D(i,j) = max \begin{cases} D(i+1, j+1) + S(T_i, Q_j) \\ D(i+1, j) + W \\ D(i, j+1) + W \end{cases}$$

$$S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

M = 1 , P = -1 , W = -3

```
T:    ACGT
      |__|
Q:    AGTT
```

Q\T	A	C	G	T	-
A	0	-2	-4	-8	-12
G	-4	-1	-1	-5	-9
T	-6	-3	0	-2	-6
T	-8	-5	-2	1	-3
-	-12	-9	-6	-3	0

And now you…

$$D_{i\,j} = max \begin{cases} D_{i+1\,j+1} + S(T_{i'}\,Q_j) \\ D_{i+1\,j} + W \\ D_{i\,j+1} + W \end{cases} , \qquad S(x,y) = \begin{cases} M & x = y \\ P & x \neq y \end{cases}$$

BREAK…

Drastic improvement on computation speed

$$\frac{(m+n)!}{m!\ n!} >> m\ n$$

Sequence lengths	Needleman-Wunsch	Exhaustive search	Number of e^- in the universe
150 X 150	22 500	~ 10^{89}	~ 10^{80}

Is a run time of “m n” good enough?

1983

David J. Lipman

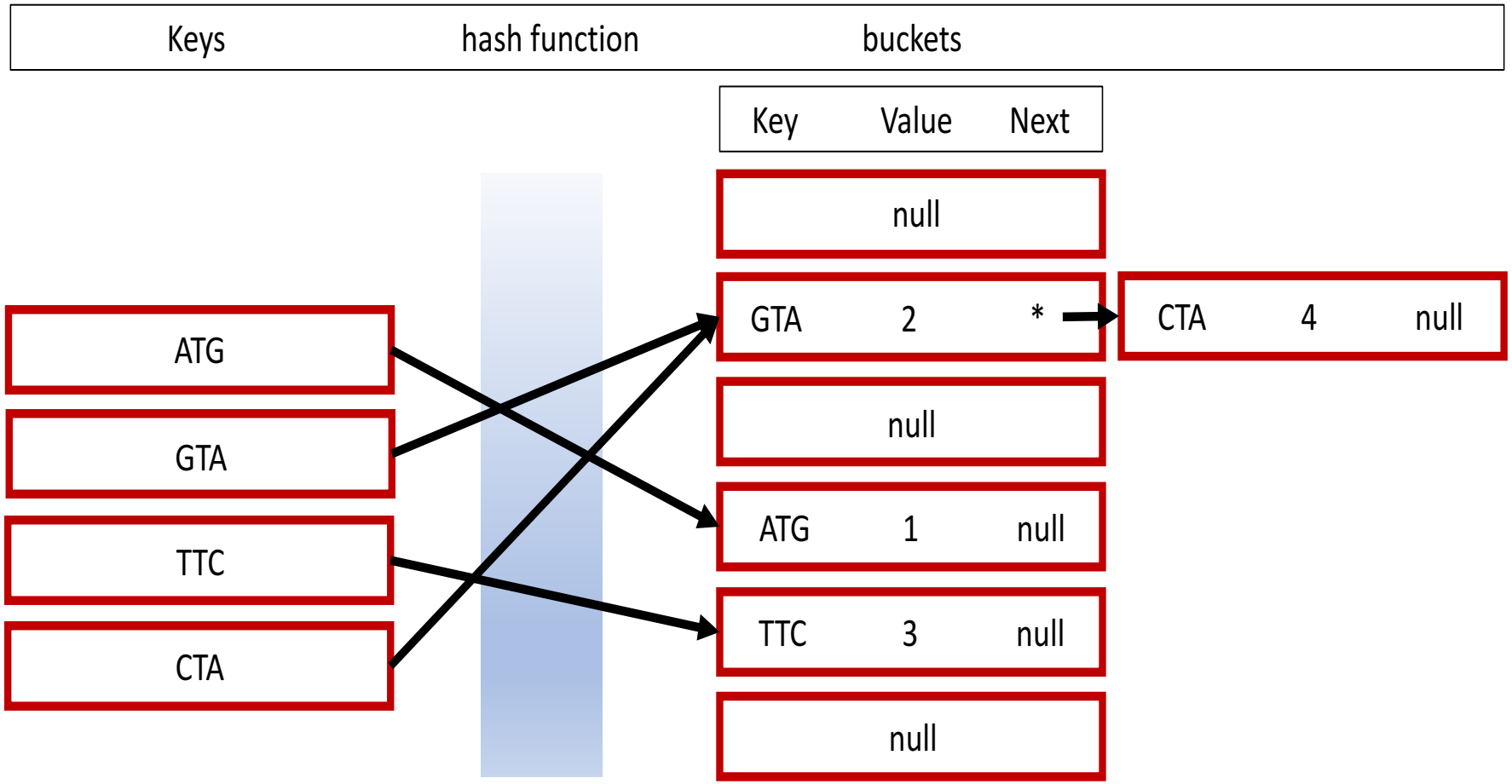
Exact matching is fast

- Feasible for entire sequences?
- What if we divide the sequences into overlapping subsequences (*k*-mers).
- And make an index revealing where each subsequence belong.

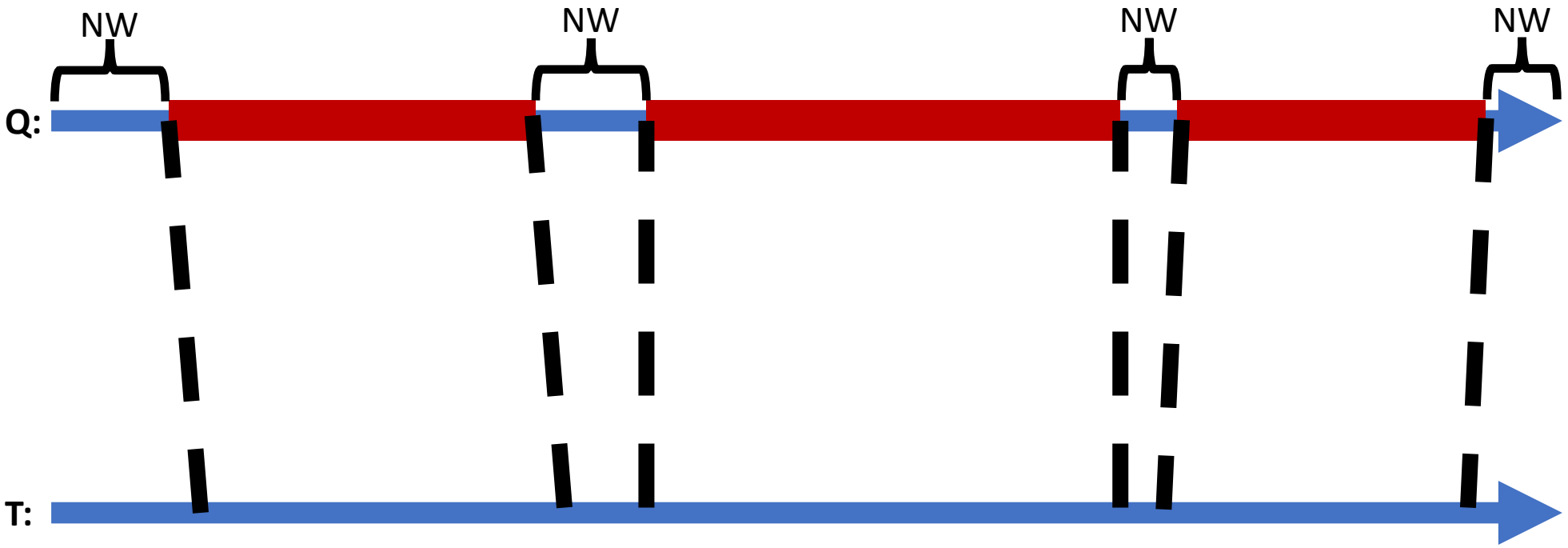
K-mer, where $k \in \mathbb{N}$, here eight

```
  CGATTTAATATTATGAAATATATTAAAGGGGAACGTATTACTTATGAGCAGT
1 CGATTTAA
2  GATTTAAT
3   ATTTAATA
4    TTTAATAT
5     TTAATATT
        … … … … … … … …
```

Stor *k*-mers in hashmaps



Fasta, 1985



Seed and extend

Simple example, k = 5:

Q: AAGACTCCGACTGGGACTTTGATGTTCGAAAGA

T: GACTGGGACTTTGATG

Seed and extend

Simple example, k = 5:

Q: AAGACTCCGACTGGGACTTTGATGTTCGAAAGA

T: GACTGGGACTTTGATG

Seed and extend

Simple example, k = 5:

Q: AAGACTCCGACTGGGACTTTGATGTTCGAAAGA

T: GACTGGGACTTTGATG

Seed and extend

Simple example, k = 5:

T: GACTGGGACTTTGATG

Extended *k*-mer matches are also noted as Anchors, Maximum Exact Matches (MEMs), or Maximum Unique Matches (MUMs) if they are uniquely occurring in the reference.

Seed and extend, impact on computation

$$m\,n >> m + n$$

And now you…

3-piece handout:
Reads, Reference and Index

Index is alphabetically sorted *k*-mers of all 10-mers in the reference prefixed with ‘T’.

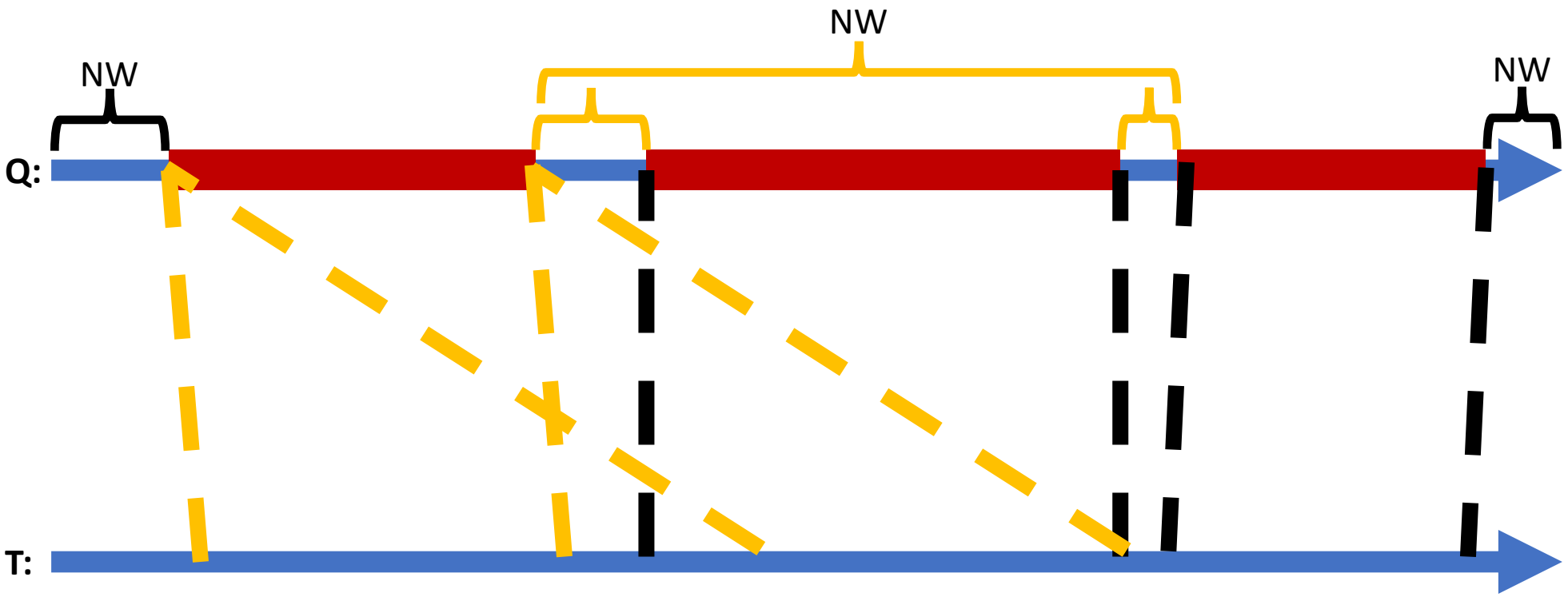
Find identify the *k*-mer matches between reads and reference using the index,
and extend on both sides of the matches.
When the *k*-mer cannot be extended further, find the next *k*-mer match.

K = 10 , prefix : T

TAAAAGATGC	91	TCTAGCTTCC	593	TGTTTTTGCT	59
TAAAGTTCTG	209	TCTCAACAGC	143	TTAAAGTTCT	208
TAAAGTTGCA	635	TCTCAGAATG	285	TTAACTGGCG	573
TAAATCTGGA	692	TCTCGCGGTA	717	TTAAGCATTG	847
TAACACTGCG	386	TCTGACAACG	407	TTAATAGACT	612
TAACCATGAG	373	TCTGCGCTCG	653	TTACATCGAA	128
TAACCGCTTT	436	TCTGCTATGT	215	TTACGGATGG	331
TAACTCGCCT	469	TCTGGAGCCG	696	TTACTCTAGC	589
TAACTGGCGA	574	TCTTACGGAT	329	TTACTTCTGA	402
TAAGAGAATT	349	TGAAAGTAAA	85	TTATCCCGTG	234
TAAGATCCTT	155	TGAACGAAAT	806	TTATCTACAC	772
TAAGCATTGG	848	TGAAGATCAG	101	TTATGCAGTG	357
TAAGCCCTCC	752	TGAAGCCATA	503	TTATTCCCTT	28
TAATAGACTG	613	TGAATGAAGC	499	TTATTGCTGA	682
TACACGACGG	777	TGACAACGAT	409	TTCAACATTT	7
TACACTATTC	277	TGACACCACG	527	TTCCAATGAT	190
TACATCGAAC	129	TGACAGTAAG	343	TTCCCGGCAA	599
TACCAAACGA	511	TGACGCCGGG	245	TTCCCTTTTT	31
TACGGATGGC	332	TGACTTGGTT	293	TTCCGGCTGG	667
TACTCACCAG	306	TGAGAGTTTT	164	TTCCGTGTCG	15
TACTCTAGCT	590	TGAGATAGGT	827	TTCCTGTTTT	55
TACTTACTCT	586	TGAGCACTTT	199	TTCGCCCCGA	172
TACTTCTGAC	403	TGAGCGTGGG	707	TTCTCAGAAT	284
TAGACAGATC	815	TGAGTACTCA	302	TTCTGACAAC	406
TAGACTGGAT	616	TGAGTATTCA	1	TTCTGCGCTC	652
TAGCTTCCCG	595	TGAGTGATAA	379	TTCTGCTATG	214
TAGGTGCCTC	832	TGATAAATCT	689	TTGACGCCGG	244
TAGTTATCTA	769	TGATAACACT	383	TTGAGAGTTT	163
TATCATTGCA	725	TGATCGTTGG	479	TTGAGTACTC	301
TATCCCGTGT	235	TGATGAGCAC	196	TTGATCGTTG	478
TATCGTAGTT	764	TGATTAAGCA	844	TTGCACAACA	447
TATCTACACG	773	TGCACAACAT	448	TTGCAGCACT	730
TATGCAGTGC	358	TGCACGAGTG	116	TTGCAGGACC	640
TATGGATGAA	800	TGCAGCAATG	542	TTGCCTTCCT	50
TATGTGGCGC	220	TGCAGCACTG	731	TTGCGCAAAC	561
TATTAACTGG	571	TGCAGGACCA	641	TTGCGGCATT	40
TATTATCCCG	232	TGCAGTGCTG	360	TTGCTCACCC	64
TATTCAACAT	5	TGCCATAACC	368	TTGCTGATAA	685
TATTCCCTTT	29	TGCCTCACTG	836	TTGGGAACCG	485
TATTCTCAGA	282	TGCCTGCAGC	538	TTGGGTGCAC	111
TATTGCTGAT	683	TGCCTTCCTG	51	TTGGTTGAGT	297
TCAACAGCGG	145	TGCGCAAACT	562	TTTAAAGTTC	207
TCAACATTTC	8	TGCGCTCGGC	655	TTTATTGCTG	681
TCACAGAAAA	316	TGCGGCATTT	41	TTTCCAATGA	189
TCACCAGTCA	309	TGCGGCCAAC	392	TTTCCGTGTC	14
TCACCCAGAA	68	TGCTATGTGG	217	TTTCGCCCCG	171
TCACTGATTA	840	TGCTCACCCA	65	TTTGCACAAC	446
TCAGAATGAC	287	TGCTGAAGAT	98	TTTGCCTTCC	49

BREAK…

There is some more to it…



Seed-Extend issues

- MEMs might be placed several places in the reference.
- Some MEMs might be erroneous mappings (frequent with smaller *k*-mers).
- Depending on the mapping method, some MEMs might have been missed (i.e. too large *k*-mer).

Chaining, 90’s

$$f(i) = \max\Big\{ \max_{i>j\geq 1} \{f(j) + \alpha(j,i) - \beta(j,i)\}, w_i \Big\}$$

$$\alpha(j,i) \;=\; \min\Big\{ \min\{y_i - y_j, x_i - x_j\}, w_i \Big\}$$

$$\beta(j,i) = \gamma_c\big((y_i - y_j) - (x_i - x_j)\big)$$

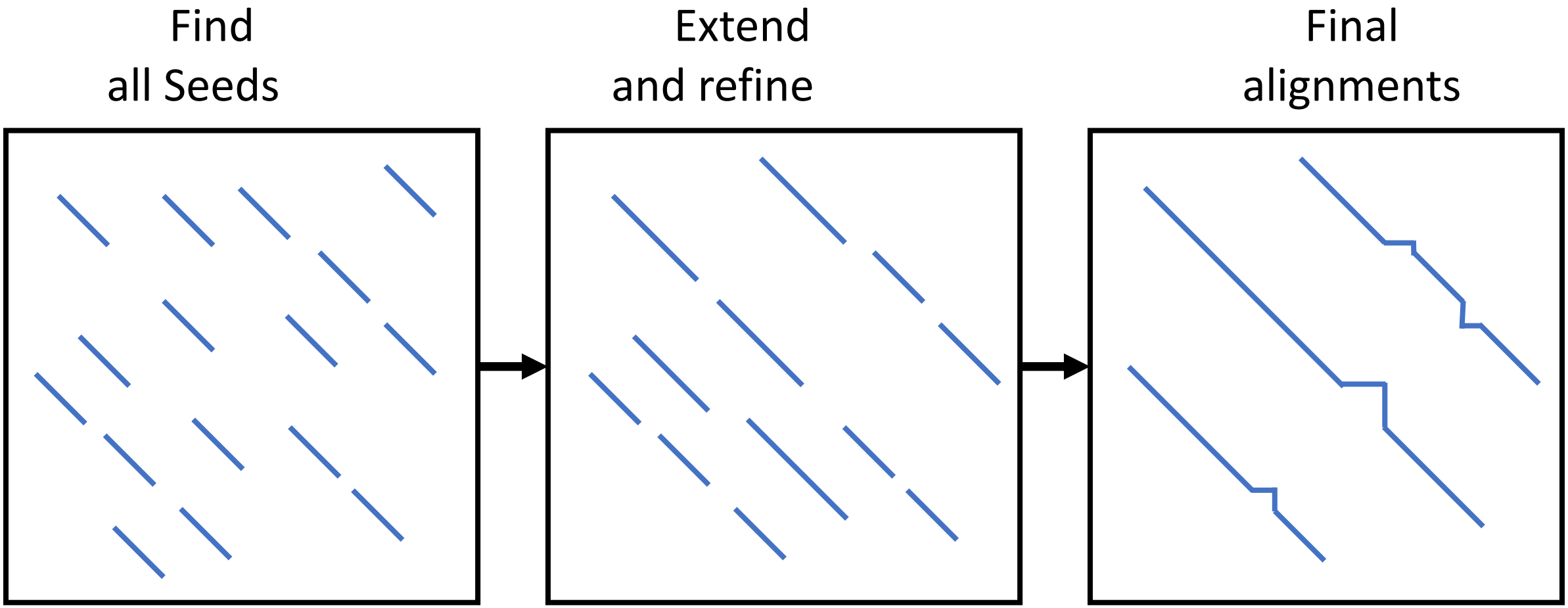
Li H. 2018

The Idea of Chaining

- Co-linear anchors (MEMs) are likely to belong together and produce high quality alignments, and can be chained together.

- Distantly placed (horizontally or vertically) placed anchors are likely to be erroneous mappings, and will be unlikely to produce high quality overlaps if chained together.

Mapping and Alignment



Methods these algorithms cover

- BLAST
  k-mer mapping, chaining, alignment
- BWA-MEM / Bowtie2
  FM-mapping, chaining, alignment
- Minimap2
  minimizer mapping, chaining, alignment
- GraphMap
  k-mer mapping, chaining, alignment

Methods these algorithms cover

- **BLAST**
  _k_-mer mapping, chaining, alignment
- BWA-MEM / Bowtie2
  FM-mapping, chaining, alignment
- **Minimap2**
  minimizer mapping, chaining, alignment
- GraphMap
  _k_-mer mapping, chaining, alignment

Why does Minimap(2) sort anchors w.r.t. to template positions prior to chaining?

Does BLAST use the same strategy…

-max_target_seqs

Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows

Nidhi Shah, Michael G Nute, Tandy Warnow, Mihai Pop

OLC, Combined with last week

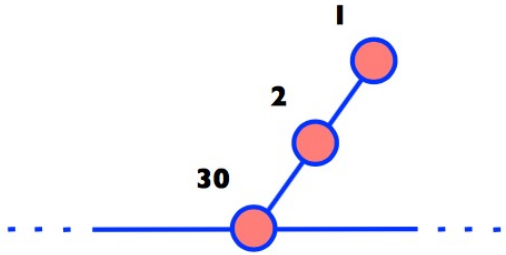
- Overlap Layout Consensus (OLC) assemblies
- Identify overlaps between reads with Minimap2 (mapping and chaining).
- Form unitigs with Miniasm.

Miniasm

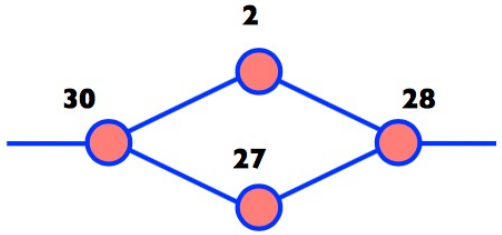
- Trim low depth regions (and tips).
- Drop contained reads.
- Remove transitive edges.
- Pop short bubbles, favor largest overlap when forming the consensus.

Basic Graph Trimming

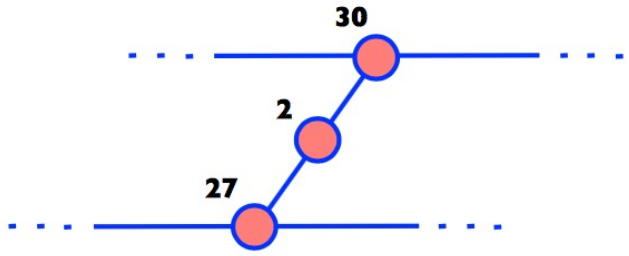
Clip tips
(seq err, end)



Pinch bubbles
(seq err, middle, SNP)



Remove low cov. links



Miniasm, pitfalls

- Quick and dirty solution:
  Gives the overall synteny of the assembled sequences.

  Usually followed by realigning the reads to assembly iteratively.
  E.g. Minimap2 or KMA

  “Polishers” exist that can correct for some of the common ONT errors.
  E.g. Medaka or Pilon

BREAK…

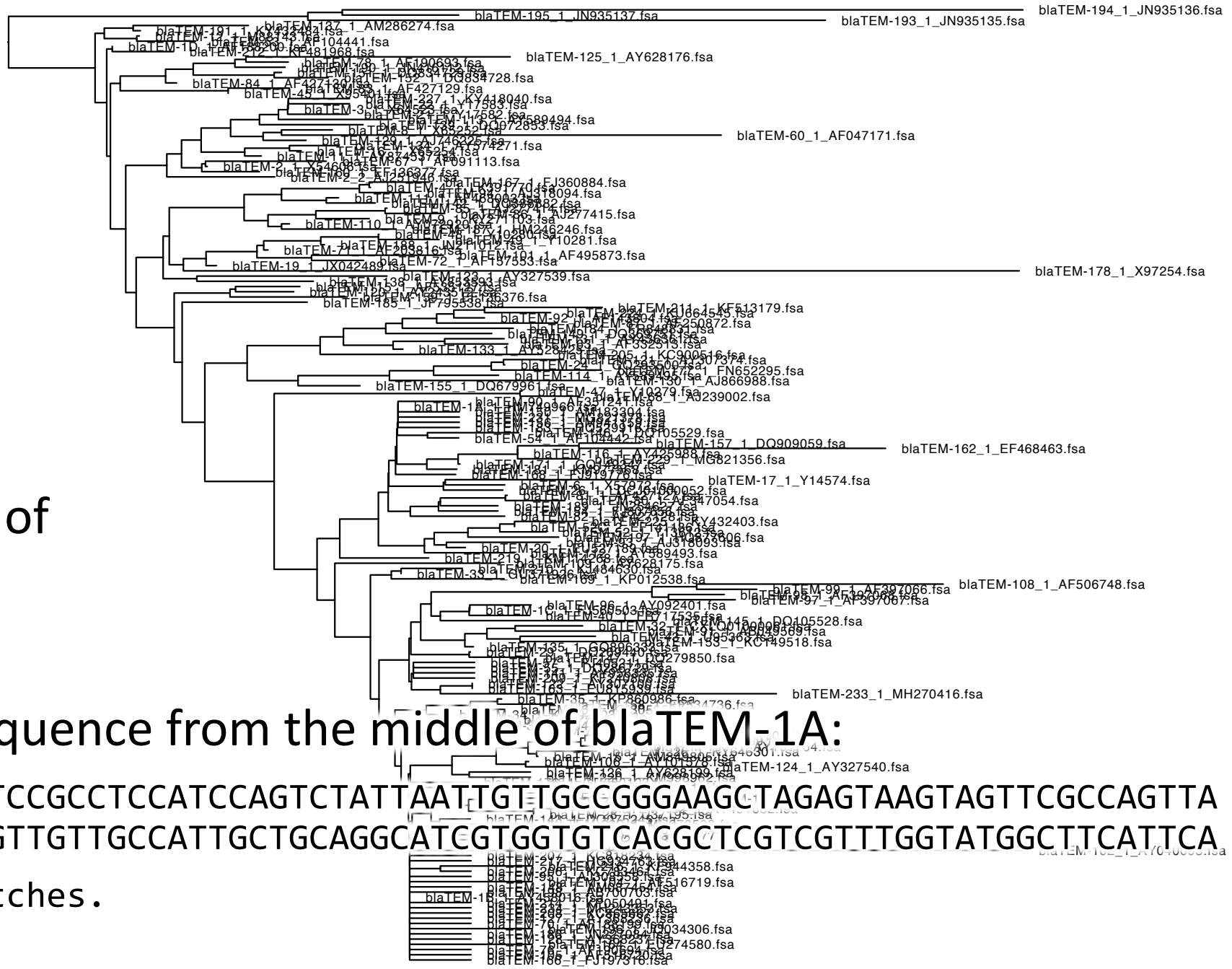
The redundant sequences

- Use the mapping and alignment methods just introduced.
- Something else?

Meat loaf from the FLI canteen

Cannibalism…

Höper D. et al. 2018



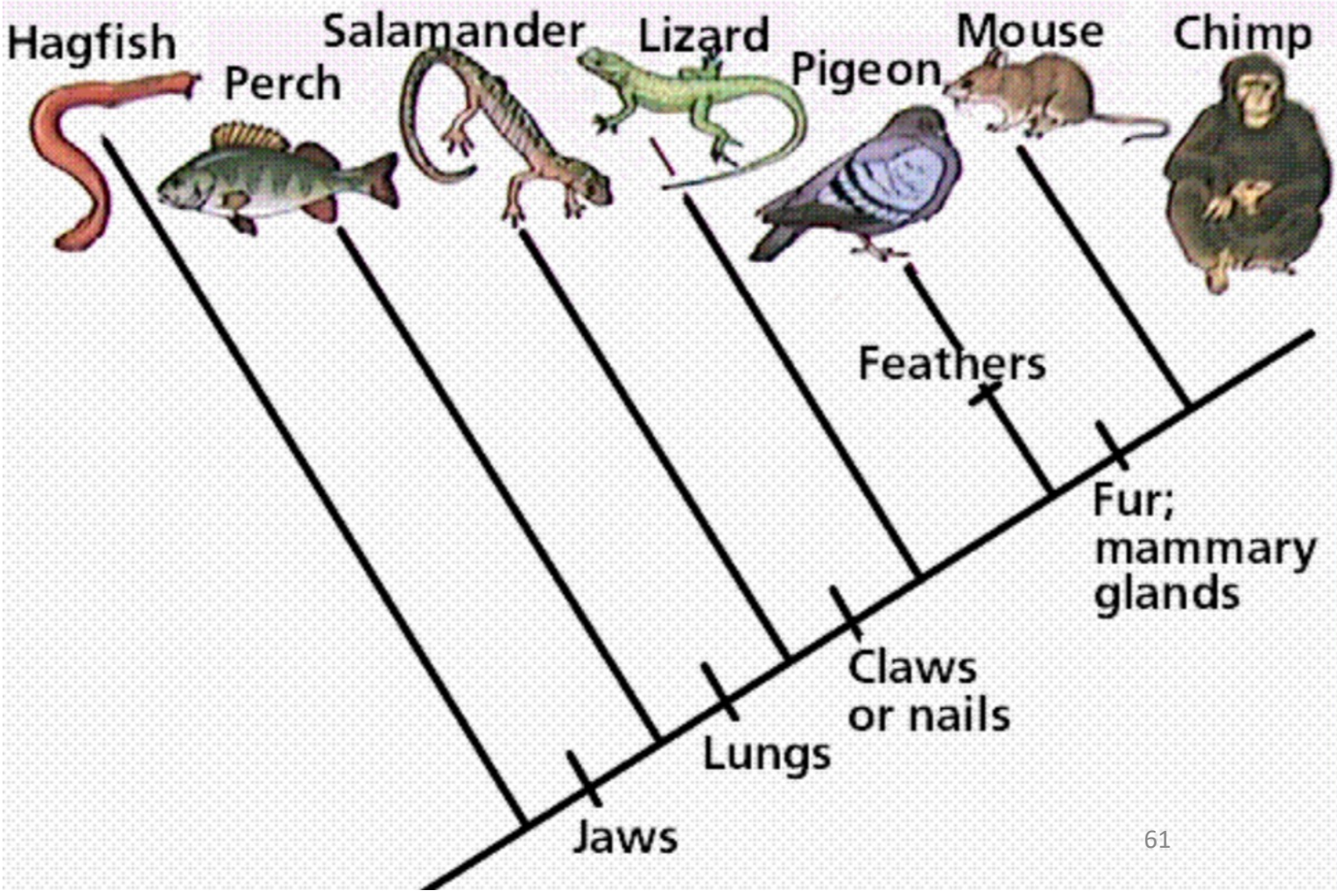
- 183 blaTEM variants
- Max distance of 38 SNPs
- 150 bp subsequence from the middle of blaTEM-1A:
- GTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTA ATAGTTTGCGCAACGTTGTTGCCATTGCTGCAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCA
- 131 perfect matches.

ConClave, 2018

Find best candidates

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \}$$

- Feathers
  - $T_m \in$ {Pigeon}
- Jaws
  - $T_m \in$ {Perch, Salamander, Lizard, Pigeon, Mouse, Chimp}
- Lungs
  - $T_m \in$ {Salamander, Lizard, Pigeon, Mouse, Chimp}



Calculate the total score of all reads against the best candidates

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases}$$

- Feathers
  - $T_m \in$ {Pigeon}
- Jaws
  - $T_m \in$ {Perch, Salamander, Lizard, Pigeon, Mouse, Chimp}
- Lungs
  - $T_m \in$ {Salamander, Lizard, Pigeon, Mouse, Chimp}

t	C(t)
Hagfish	0
Perch	1
Salamander	2
Lizard	2
Pigeon	3
Mouse	2
Chimp	2

Calculate the total score of all reads against the best candidates

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \}$$

- Feathers
  - $T_m \in$ {Pigeon}
  - $S_k \in$ {Pigeon}
- Jaws
  - $T_m \in$ {Perch, Salamander, Lizard, Pigeon, Mouse, Chimp}
  - $S_k \in$ {Pigeon}
- Lungs
  - $T_m \in$ {Salamander, Lizard, Pigeon, Mouse, Chimp}
  - $S_k \in$ {Pigeon}

t	C(t)
Hagfish	0
Perch	1
Salamander	2
Lizard	2
Pigeon	3
Mouse	2
Chimp	2

And now you…

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \tag{1}$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \tag{2}$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \tag{3}$$

Query sequence	Score	Templates
GCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGTAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG	64	2,12
GAATGACTTGCTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA	63	1,2,3,7,8,9
TAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTG	72	1,2,3,4,5,6,8,9,11,12
CCACTTCTGCGCTCGGCCCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCG	66	2,5
TCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA	74	2,5,12
GGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTGGGAATCATG	63	1,2,3,4,5,6,8,9,11,12
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATTAAGTTGCAGGACCACTTCTGCGCTCGGC	39	1,2,3,4,6,7,8,9,10,11
CCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC	71	2,5,12

#	Templates:	ConClave Scores:
01	blaTEM-1A_4_HM749966	_____
02	blaTEM-2_1_X54606	_____
03	blaTEM-2_3_AJ251946	_____
04	blaTEM-3_1_X64523	_____
05	blaTEM-6_1_X57972	_____
06	blaTEM-8_1_X65252	_____
07	blaTEM-10_1_AF093512	_____
08	blaTEM-11_1_AY874537	_____
09	blaTEM-12_1_M88143	_____
10	blaTEM-15_1_AM849805	_____
11	blaTEM-16_1_X65254	_____
12	blaTEM-17_1_Y14574	_____

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \tag{1}$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \tag{2}$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \tag{3}$$

Predicted Template: _____

Query sequence	Score	Templates
GCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGTAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG	64	2,12
GAATGACTTGCTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA	63	1,2,3,7,8,9
TAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTG	72	1,2,3,4,5,6,8,9,11,12
CCACTTCTGCGCTCGGCCCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCG	66	2,5
TCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA	74	2,5,12
GGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTGGGAATCATG	63	1,2,3,4,5,6,8,9,11,12
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATTAAGTTGCAGGACCACTTCTGCGCTCGGC	39	1,2,3,4,6,7,8,9,10,11
CCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC	71	2,5,12

#	Templates:	ConClave Scores:
01	blaTEM-1A_4_HM749966	_____
02	blaTEM-2_1_X54606	64
03	blaTEM-2_3_AJ251946	_____
04	blaTEM-3_1_X64523	_____
05	blaTEM-6_1_X57972	_____
06	blaTEM-8_1_X65252	_____
07	blaTEM-10_1_AF093512	_____
08	blaTEM-11_1_AY874537	_____
09	blaTEM-12_1_M88143	_____
10	blaTEM-15_1_AM849805	_____
11	blaTEM-16_1_X65254	_____
12	blaTEM-17_1_Y14574	64

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \quad (1)$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \quad (2)$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \quad (3)$$

Predicted Template: _____

Query sequence	Score	Templates
GCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGTAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG	64	2,12
GAATGACTTGCTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA	63	1,2,3,7,8,9
TAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTG	72	1,2,3,4,5,6,8,9,11,12
CCACTTCTGCGCTCGGCCCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCG	66	2,5
TCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA	74	2,5,12
GGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTGGGAATCATG	63	1,2,3,4,5,6,8,9,11,12
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATTAAGTTGCAGGACCACTTCTGCGCTCGGC	39	1,2,3,4,6,7,8,9,10,11
CCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC	71	2,5,12

#	Templates:	ConClave Scores:
01	blaTEM-1A_4_HM749966	63
02	blaTEM-2_1_X54606	64+63
03	blaTEM-2_3_AJ251946	63
04	blaTEM-3_1_X64523	
05	blaTEM-6_1_X57972	
06	blaTEM-8_1_X65252	
07	blaTEM-10_1_AF093512	63
08	blaTEM-11_1_AY874537	63
09	blaTEM-12_1_M88143	63
10	blaTEM-15_1_AM849805	
11	blaTEM-16_1_X65254	
12	blaTEM-17_1_Y14574	64

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \quad (1)$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \quad (2)$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \quad (3)$$

Predicted Template: _____

Query sequence	Score	Templates
GCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGTAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG	64	2,12
GAATGACTTGCTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA	63	1,2,3,7,8,9
TAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTG	72	1,2,3,4,5,6,8,9,11,12
CCACTTCTGCGCTCGGCCCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCG	66	2,5
TCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA	74	2,5,12
GGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTGGGAATCATG	63	1,2,3,4,5,6,8,9,11,12
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATTAAGTTGCAGGACCACTTCTGCGCTCGGC	39	1,2,3,4,6,7,8,9,10,11
CCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC	71	2,5,12

#	Templates:	ConClave Scores:
01	blaTEM-1A_4_HM749966	63+72
02	blaTEM-2_1_X54606	64+63+72
03	blaTEM-2_3_AJ251946	63+72
04	blaTEM-3_1_X64523	72
05	blaTEM-6_1_X57972	72
06	blaTEM-8_1_X65252	72
07	blaTEM-10_1_AF093512	63
08	blaTEM-11_1_AY874537	63+72
09	blaTEM-12_1_M88143	63+72
10	blaTEM-15_1_AM849805	
11	blaTEM-16_1_X65254	72
12	blaTEM-17_1_Y14574	64+72

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \quad (1)$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \quad (2)$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \quad (3)$$

Predicted Template: _____

Query sequence	Score	Templates
GCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGTAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG	64	2,12
GAATGACTTGCTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA	63	1,2,3,7,8,9
TAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTG	72	1,2,3,4,5,6,8,9,11,12
CCACTTCTGCGCTCGGCCCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCG	66	2,5
TCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA	74	2,5,12
GGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTGGGAATCATG	63	1,2,3,4,5,6,8,9,11,12
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATTAAGTTGCAGGACCACTTCTGCGCTCGGC	39	1,2,3,4,6,7,8,9,10,11
CCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC	71	2,5,12

#	Templates:	ConClave Scores:
01	blaTEM-1A_4_HM749966	63+72
02	blaTEM-2_1_X54606	64+63+72+66
03	blaTEM-2_3_AJ251946	63+72
04	blaTEM-3_1_X64523	72
05	blaTEM-6_1_X57972	72+66
06	blaTEM-8_1_X65252	72
07	blaTEM-10_1_AF093512	63
08	blaTEM-11_1_AY874537	63+72
09	blaTEM-12_1_M88143	63+72
10	blaTEM-15_1_AM849805	
11	blaTEM-16_1_X65254	72
12	blaTEM-17_1_Y14574	64+72

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \quad (1)$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \quad (2)$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \quad (3)$$

Predicted Template: _____

Query sequence	Score	Templates
GCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGTAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG	64	2,12
GAATGACTTGCTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA	63	1,2,3,7,8,9
TAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTG	72	1,2,3,4,5,6,8,9,11,12
CCACTTCTGCGCTCGGCCCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCG	66	2,5
TCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA	74	2,5,12
GGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTGGGAATCATG	63	1,2,3,4,5,6,8,9,11,12
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATTAAGTTGCAGGACCACTTCTGCGCTCGGC	39	1,2,3,4,6,7,8,9,10,11
CCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC	71	2,5,12

#	Templates:	ConClave Scores:
01	blaTEM-1A_4_HM749966	63+72+63+39
02	blaTEM-2_1_X54606	64+63+72+66+74+63+39+71
03	blaTEM-2_3_AJ251946	63+72+63+39
04	blaTEM-3_1_X64523	72+63+39
05	blaTEM-6_1_X57972	72+66+74+63+71
06	blaTEM-8_1_X65252	72+63+39
07	blaTEM-10_1_AF093512	63+39
08	blaTEM-11_1_AY874537	63+72+63+39
09	blaTEM-12_1_M88143	63+72+63+39
10	blaTEM-15_1_AM849805	39
11	blaTEM-16_1_X65254	72+63+39
12	blaTEM-17_1_Y14574	64+72+74+63+71

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \quad (1)$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \quad (2)$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \quad (3)$$

Predicted Template: _____

Query sequence	Score	Templates
GCGCGGTATTATCCCGTGTTGACGCCGGGCAAGAGTAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTG	64	2,12
GAATGACTTGCTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCA	63	1,2,3,7,8,9
TAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTG	72	1,2,3,4,5,6,8,9,11,12
CCACTTCTGCGCTCGGCCCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCG	66	2,5
TCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTA	74	2,5,12
GGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATTGGGAATCATG	63	1,2,3,4,5,6,8,9,11,12
ACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATTAAGTTGCAGGACCACTTCTGCGCTCGGC	39	1,2,3,4,6,7,8,9,10,11
CCATCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCAC	71	2,5,12

#	Templates:	ConClave Scores:
01	blaTEM-1A_4_HM749966	237
02	blaTEM-2_1_X54606	512
03	blaTEM-2_3_AJ251946	237
04	blaTEM-3_1_X64523	174
05	blaTEM-6_1_X57972	346
06	blaTEM-8_1_X65252	174
07	blaTEM-10_1_AF093512	102
08	blaTEM-11_1_AY874537	237
09	blaTEM-12_1_M88143	237
10	blaTEM-15_1_AM849805	39
11	blaTEM-16_1_X65254	174
12	blaTEM-17_1_Y14574	344

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \quad (1)$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \quad (2)$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \quad (3)$$

Predicted Template: blaTEM-2_1_X54606

But is it Relevant for ONT…

Final definitions

- Global Alignment
- Local Alignment
- Minimizer

And KMA

$$T_m \in \underset{i \in T}{\operatorname{argmax}}\{ f(t_i) \} \tag{1}$$

$$C(t) = \sum_{k \in K} \max \begin{cases} f(t_k) & \tau \leq f(t_k) \\ 0 & else \end{cases} \tag{2}$$

$$S_k \in \underset{i \in T_m}{\operatorname{argmax}}\{ C(t_i) \} \tag{3}$$

KMA nano-branch

$$T_m(q) \in \underset{t \in T}{\operatorname{argmax}}\{ f(q,t) \} \tag{1}$$

$$C(t) \in \sum_{q \in Q} \begin{cases} f(q,t) & \tau \leq f(t_k) \wedge t \in T_m(q) \\ 0 & else \end{cases}. \tag{2}$$

$$S_q \in \underset{t \in T_m(q)}{\operatorname{argmax}}\{ C(t) \} \tag{3}$$

$$T_m(q) \in \underset{t \in T}{\operatorname{argmax}} \left\{ \min \left\{ \begin{matrix} \max\limits_{r \in T}\{f(q,r)\} \\ \dfrac{f(q,t)}{\epsilon} \end{matrix} \right\} \right. \quad \varepsilon \in [0;1] \tag{4}$$

ONT Assembly Challenge

- Built the best ONT assembly pipeline
- The same criteria as last week, but ONT instead of Illumina.
- Same deadline and prices too.