Characterizing and Correcting Errors in Long Read Sequencing Data

PRESENTER: **Jonn Smith**

INTRO

The PacBio HiFi protocol can generate consensus reads of quality Q20-Q30 by redundant sequencing of circularized fragments. However, this consensus process is only successful if the HiFi adapter is correctly isolated and removed, and if the circular sequencing completes at least two passes of the original input molecule. Reads failing these criteria are not corrected. Empirically this amounts to, on average, 50% the reads being rejected from the HiFi consensus process and left at their original lower quality.

We use the corrected HiFi reads to construct a graph based on their alignment to the reference.
We then align the rejected reads to this graph, favoring edges from the HiFi data and omitting poorly supported edges in the rejected data. This effectively corrects the rejected reads to the same approximate quality as the corrected HiFi reads using the these HiFi reads as a prior.

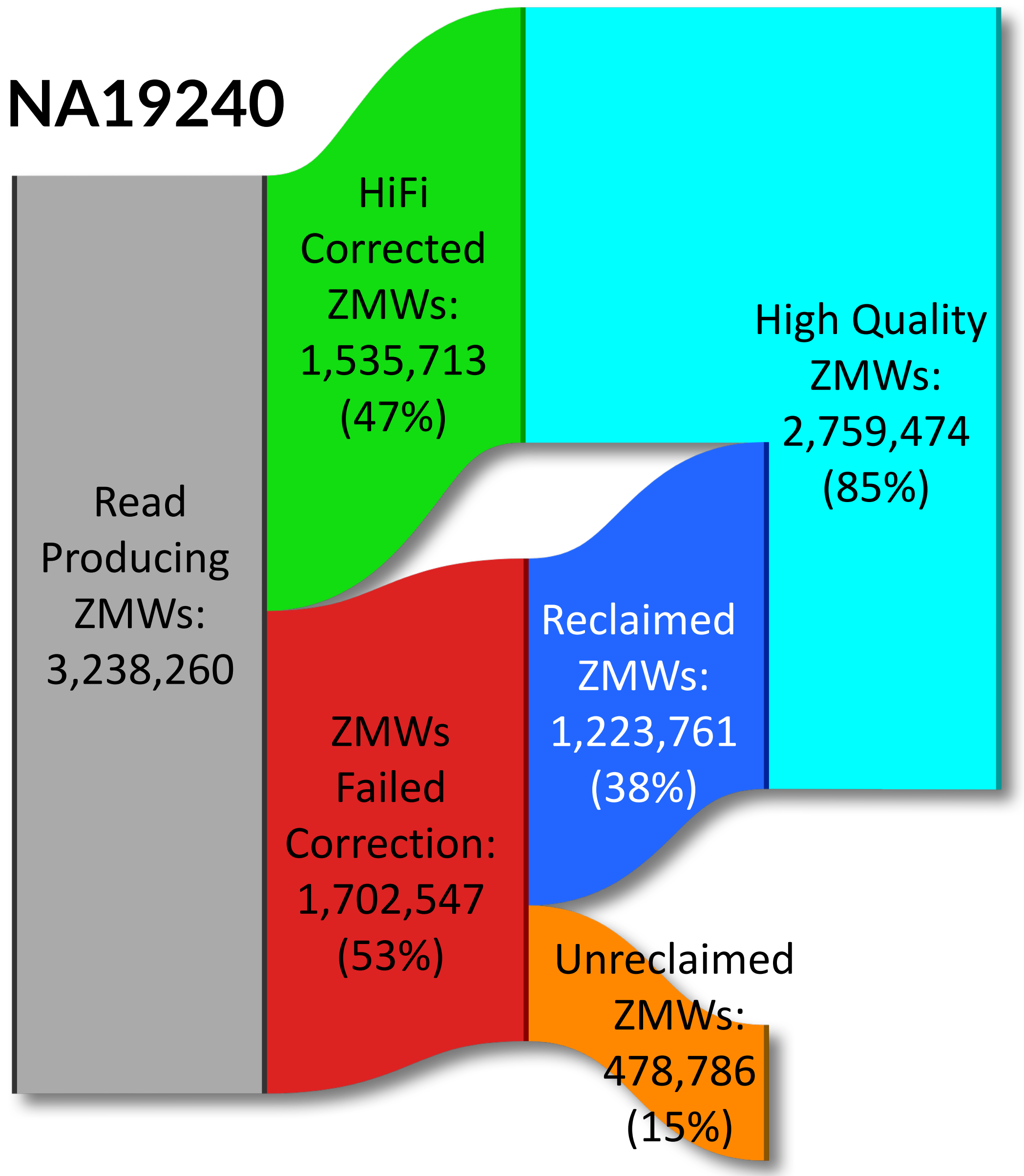
We believe this method is applicable to other consensus-based sequencing techniques as well (such as R2C2 as applied to Oxford Nanopore data).

Here we present our prototype of this reclamation process.

METHODS

1. Create *Partial Order Alignment Graph* from HiFi reads aligned to the reference
2. Align rejected reads to HiFi read graph
3. Correct and extract rejected read sequences based on alignment with graph
4. Select "best" reclaimed read from each ZMW

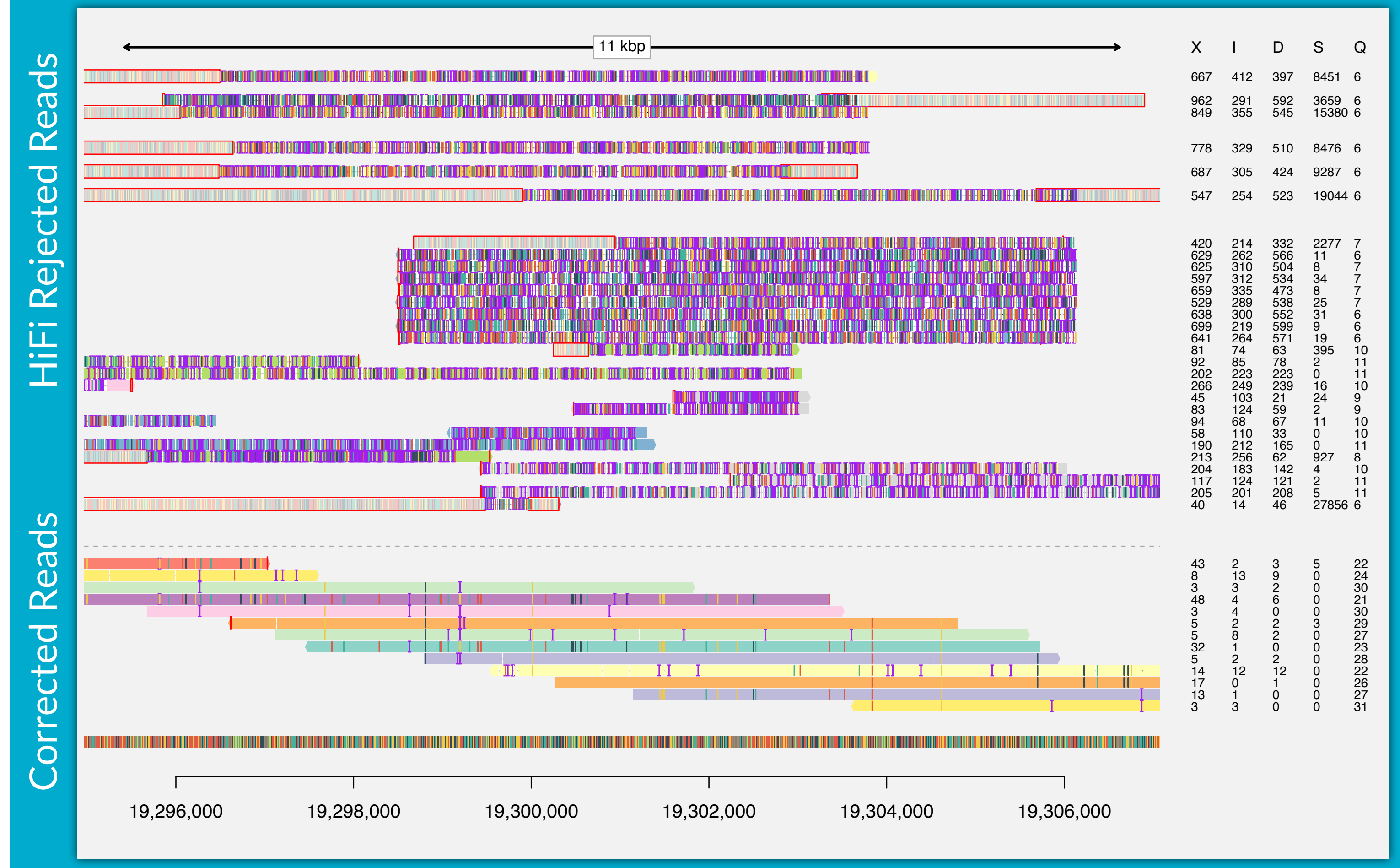
RESULTS



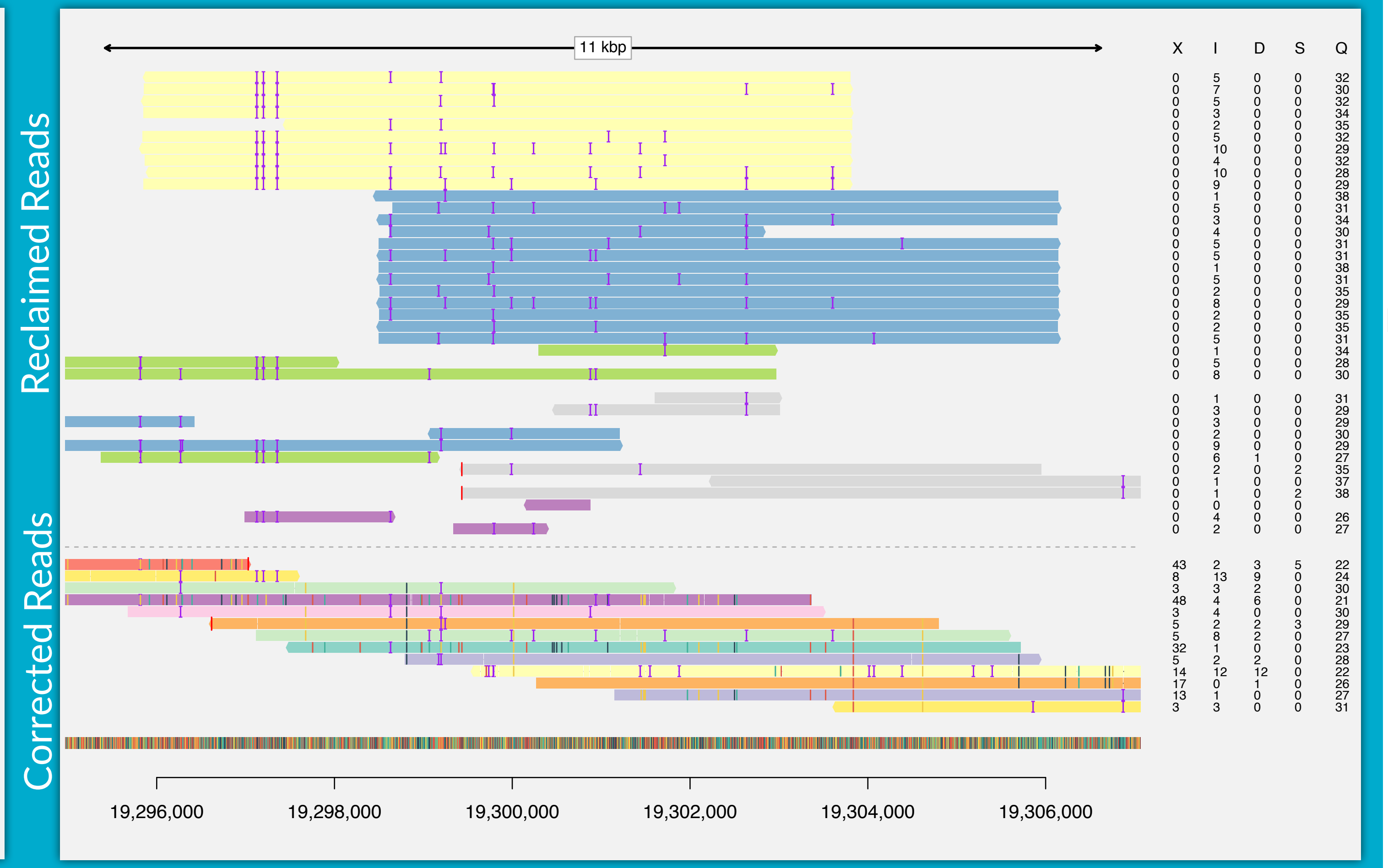
NA19240
Read Producing ZMWs: 3,238,260
HiFi Corrected ZMWs: 1,535,713 (47%)
ZMWs Failed Correction: 1,702,547 (53%)
High Quality ZMWs: 2,759,474 (85%)
Reclaimed ZMWs: 1,223,761 (38%)
Unreclaimed ZMWs: 478,786 (15%)

Computational Biology and Data Science

Graph-based read correction **reclaims 40-90% of reads originally** rejected from PacBio HiFi data **without resequencing.**



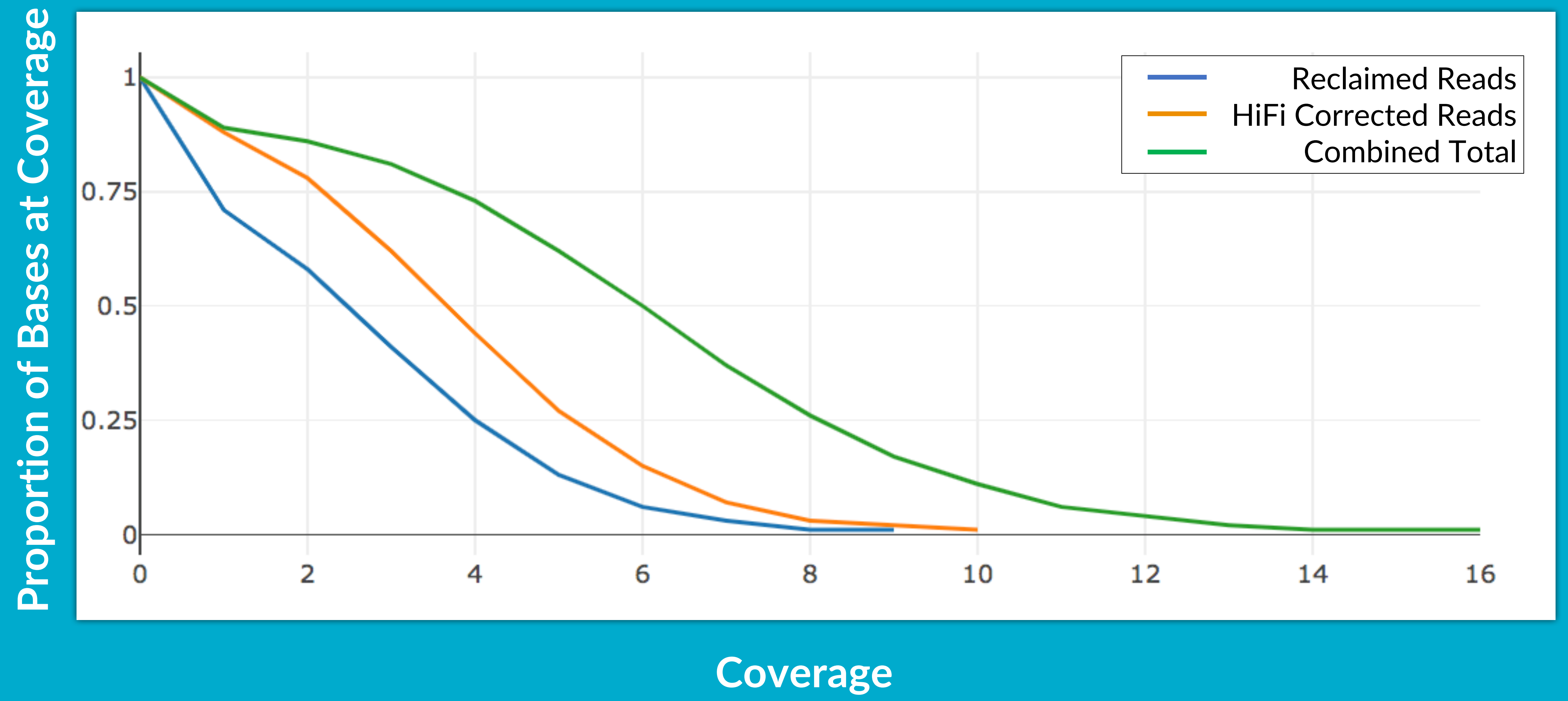
X	I	D	S	Q
667	412	397	8451	6
962	291	592	3659	6
849	355	545	15380	6
778	329	510	8476	6
687	305	424	9287	6
547	254	523	19044	6
420	214	332	2277	7
629	262	566	11	6
625	310	504	8	7
597	312	534	34	7
659	335	473	8	7
529	289	538	25	7
638	300	552	31	6
699	219	599	9	6
641	264	571	19	6
81	74	63	395	10
92	85	78	2	11
202	223	223	0	11
266	249	239	16	10
45	103	21	24	9
83	124	59	2	9
94	68	67	11	10
58	110	33	0	10
190	212	165	0	11
213	256	62	927	8
204	183	142	4	10
117	124	121	2	11
205	201	208	5	11
40	14	46	27856	6
43	2	3	5	22
8	13	9	0	24
3	3	2	0	30
48	4	6	0	21
3	4	0	0	30
5	2	2	3	29
5	8	2	0	27
32	1	0	0	23
5	2	2	0	28
14	12	12	0	22
17	0	1	0	26
13	1	0	0	27
3	3	0	0	31



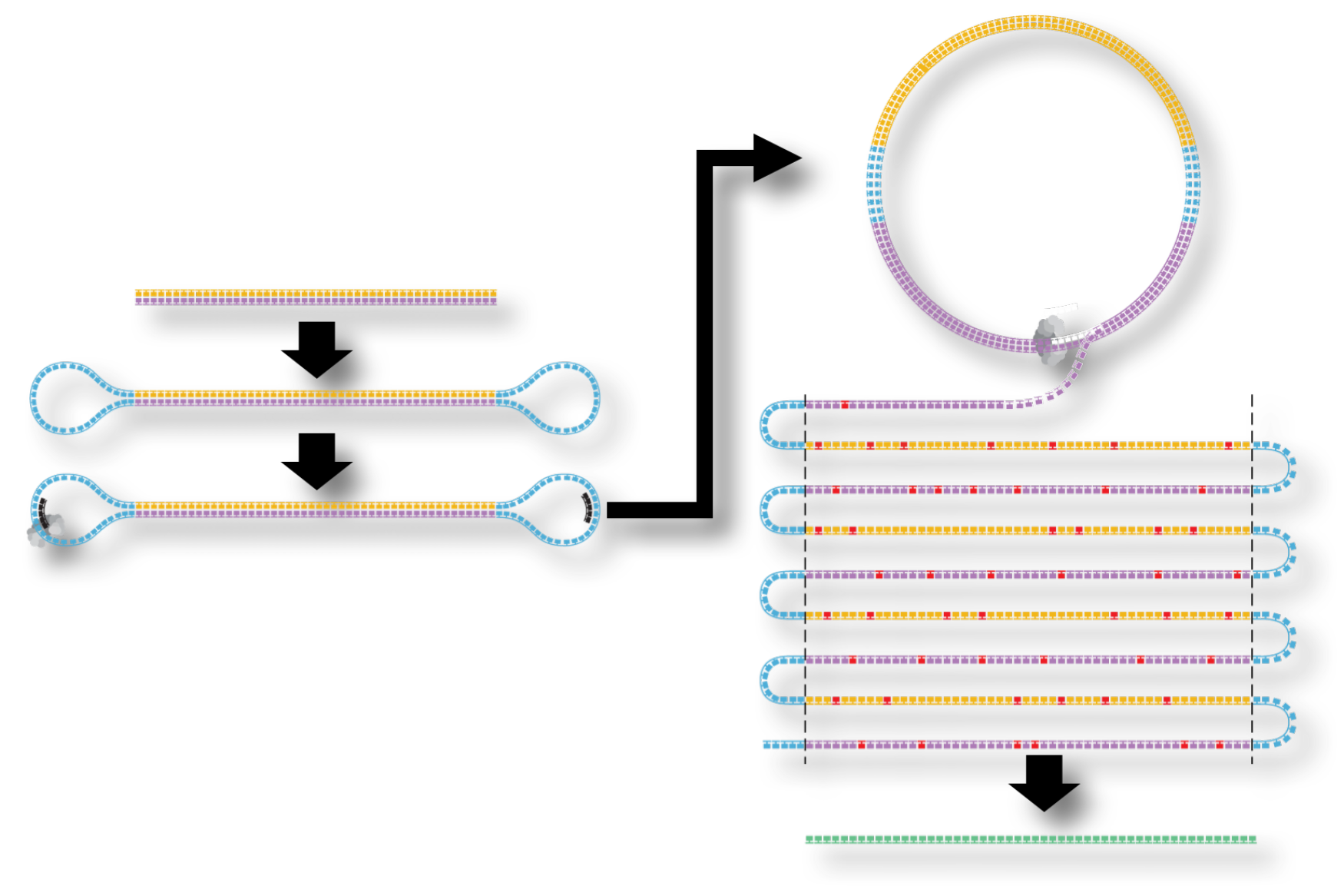
X	I	D	S	Q
0	5	0	0	32
0	7	0	0	30
0	5	0	0	32
0	3	0	0	34
0	2	0	0	35
0	5	0	0	32
0	10	0	0	29
0	4	0	0	32
0	10	0	0	28
0	9	0	0	29
0	1	0	0	38
0	5	0	0	31
0	3	0	0	34
0	4	0	0	30
0	5	0	0	31
0	5	0	0	31
0	1	0	0	38
0	5	0	0	31
0	2	0	0	35
0	8	0	0	29
0	2	0	0	35
0	2	0	0	35
0	5	0	0	31
0	1	0	0	34
0	5	0	0	28
0	8	0	0	30
0	1	0	0	31
0	3	0	0	29
0	3	0	0	29
0	2	0	0	30
0	9	0	0	29
0	6	1	0	27
0	2	0	2	35
0	1	0	0	37
0	1	0	2	38
0	0	0	0	
0	4	0	0	26
0	2	0	0	27
43	2	3	5	22
8	13	9	0	24
3	3	2	0	30
48	4	6	0	21
3	4	0	0	30
5	2	2	3	29
5	8	2	0	27
32	1	0	0	23
5	2	2	0	28
14	12	12	0	22
17	0	1	0	26
13	1	0	0	27
3	3	0	0	31

X – Mismatches, I – Insertions, D- Deletions S – Soft Clips, Q – Quality

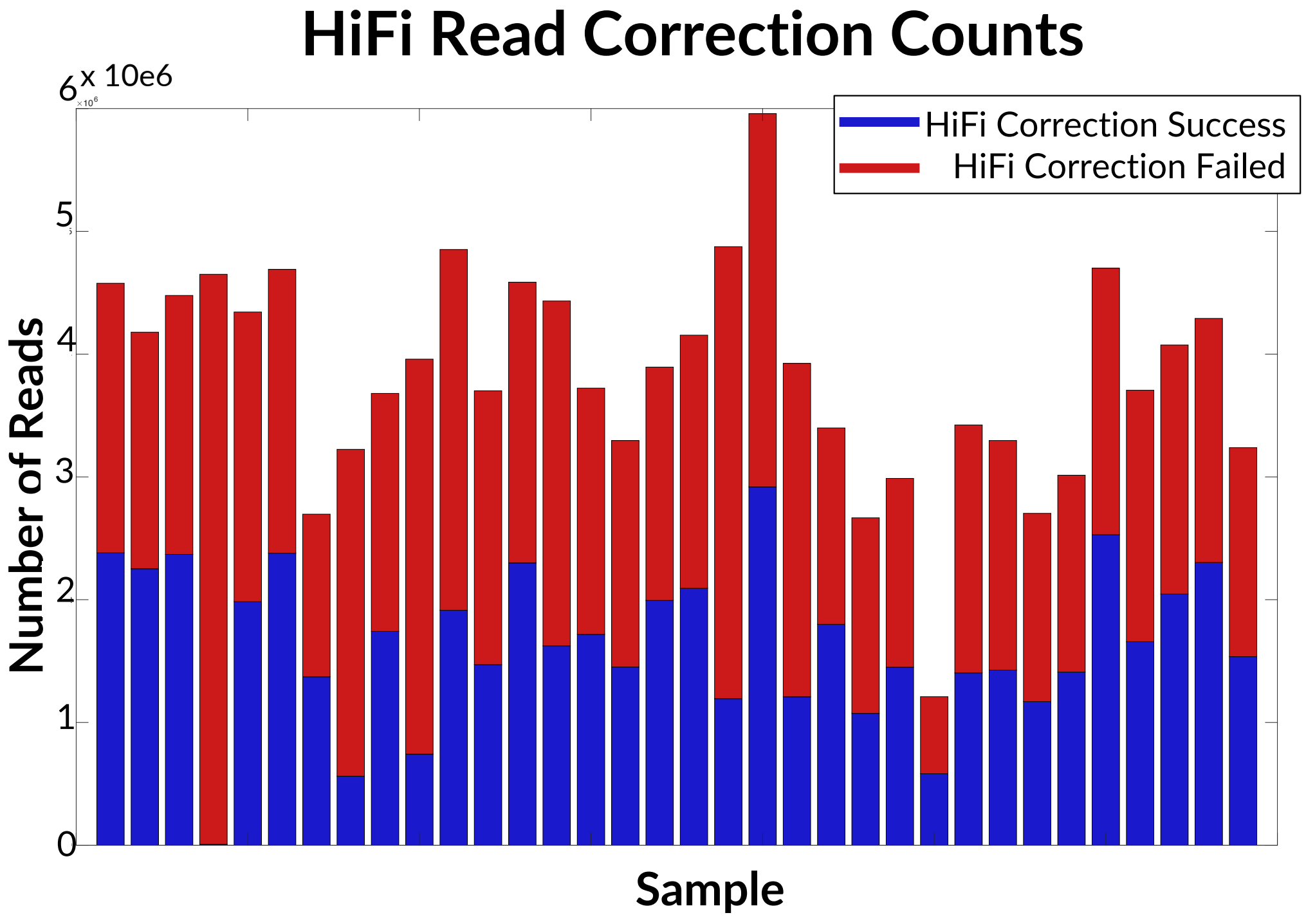
NA19240 Coverage



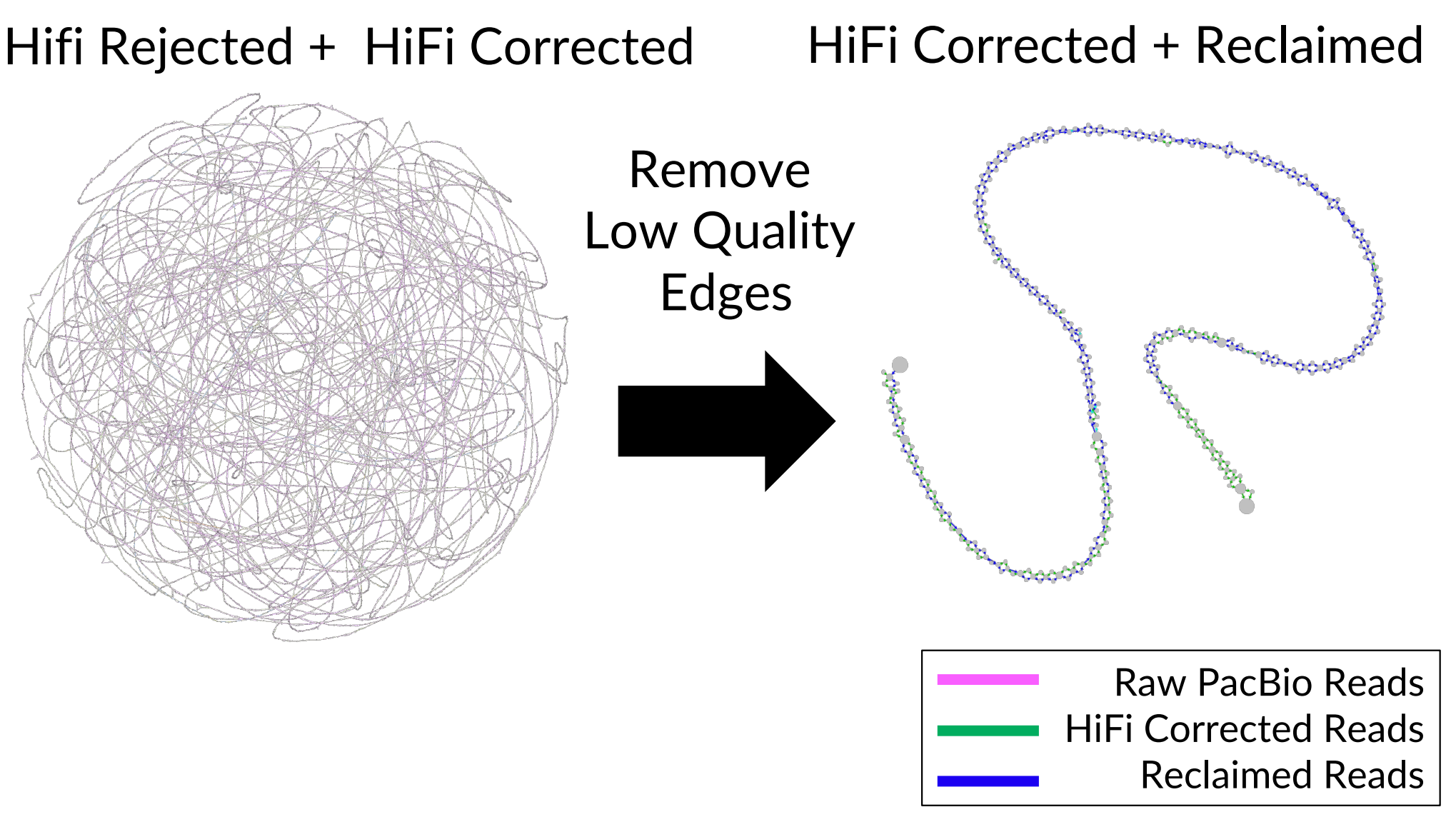
PacBio HiFi Sequencing Process



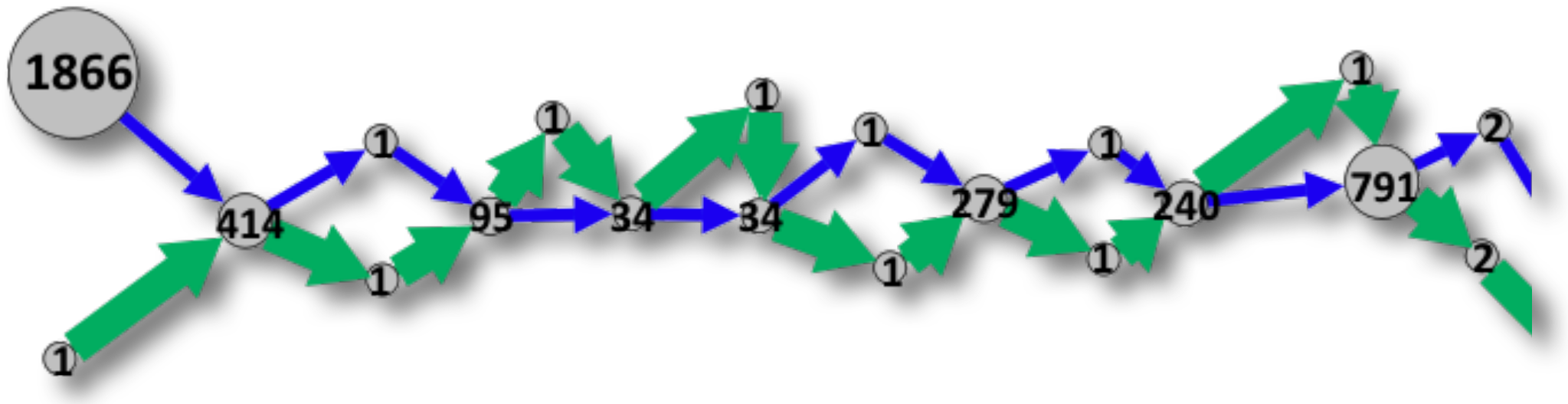
HiFi Read Correction Counts



Partial Order Alignment Graphs of Reads



HiFi Corrected + Reclaimed
(Graph Start)



Jonn Smith, Michael Gatzen, Steve Huang, Maura Costello, Tera Bowers, Kiran Garimella

BROAD INSTITUTE

References:

- Lee, C. (2003). Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, *19*(8), 999–1008. https://doi.org/10.1093/bioinformatics/btg109
- Rautiainen, M., & Marschall, T. (2019). Graphaligner: Rapid and versatile sequence-to-graph alignment. *BioRxiv*, 810812. https://doi.org/10.1101/810812
- Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., … Hunkapiller, M. W. (2019). Highly-accurate long-read sequencing improves variant detection and assembly of a human genome. *BioRxiv*, 519025. https://doi.org/10.1101/519025