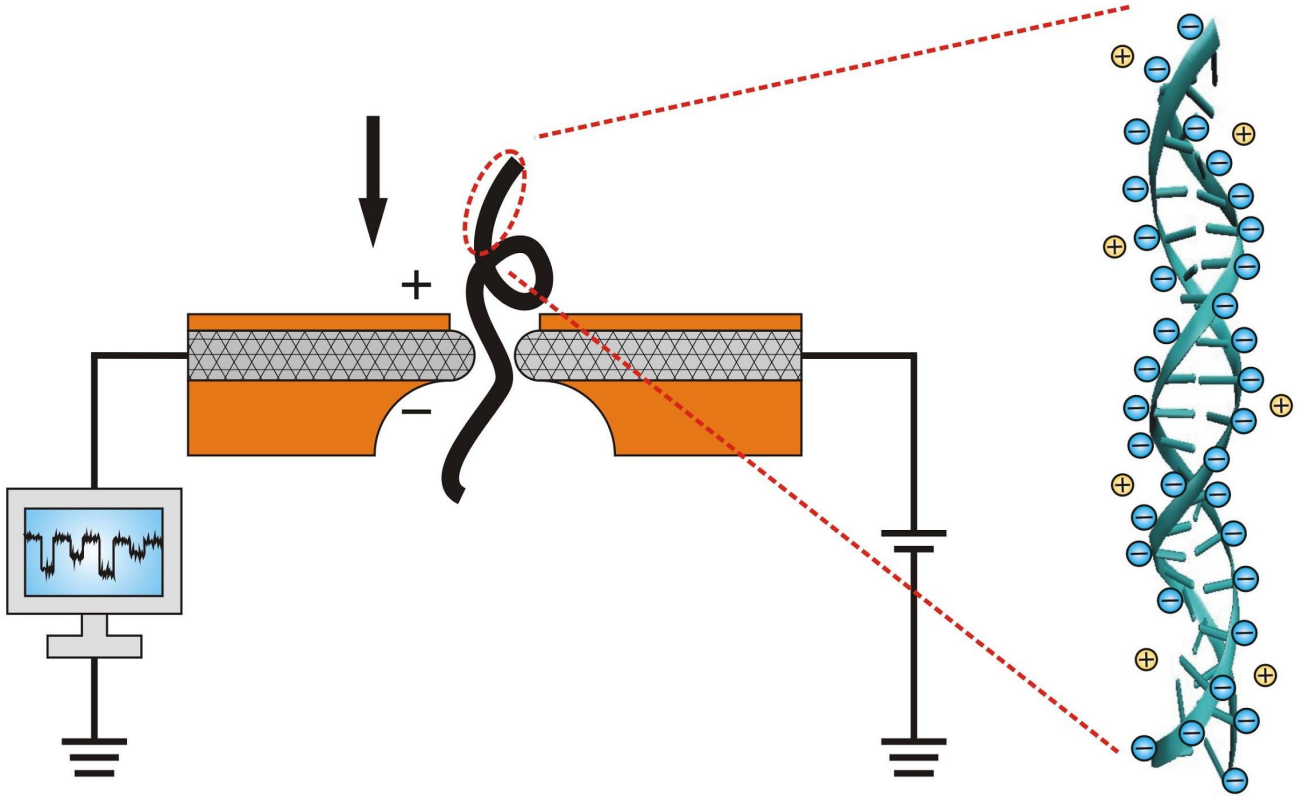
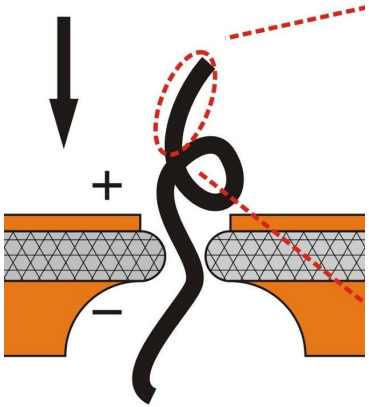
Encoding DNA-Sequenced Nanopore Data

Presented by Sasha Jenner

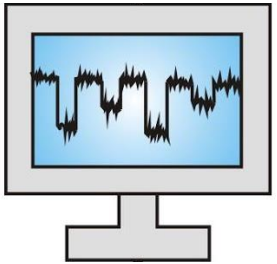
Background



Nanopore sequencing of DNA

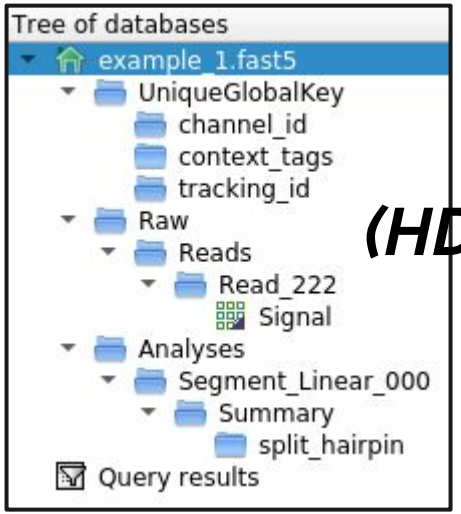


Read

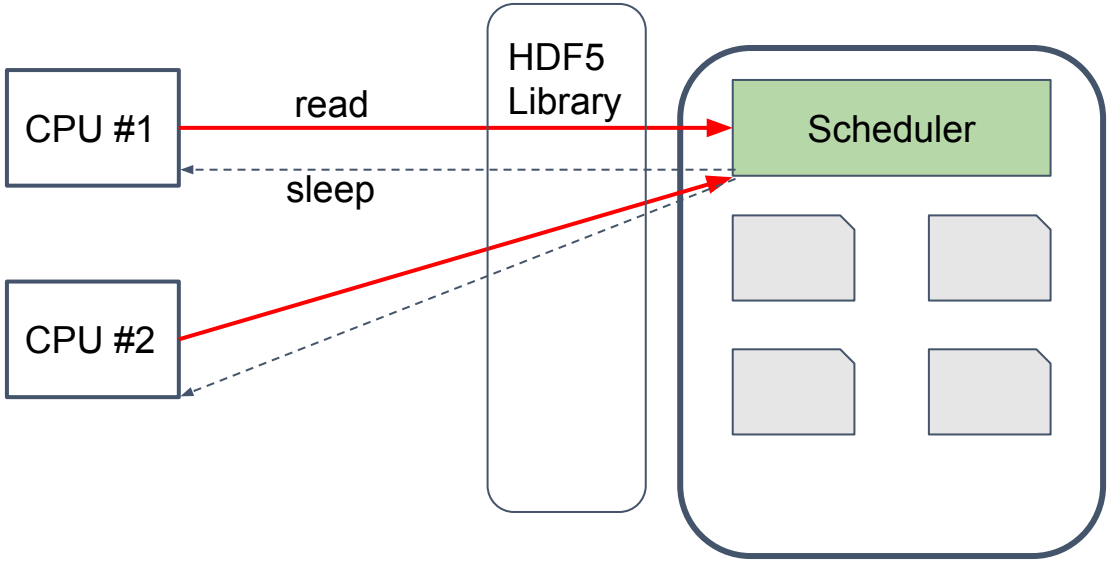


Signal

FAST5



(HDF5)


CPU #1
read
sleep
CPU #2
HDF5
Library
Scheduler

SLOW5 TSV File

#fileformat: slow5v1.0
#exp_start_time: 2020-01-01T00:00:00Z
#run_id: 855cdb4b26948
#flow_cell_id: FAH00000

Nanopore Metadata

#read_id	n_samples	digitisation	offset	range	sampling_rate	raw_signal
read-0	123456	8192	6	1467.6	100000	498,492,501,508,503,505,509,...
read-1	2000	8192	5	1467.6	4000	400,401,500,403,407,478,510,...
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
read-N	10000	8192	3	1467.6	4000	559,545,560,551,550,565,701,...

Read ID

Read Metadata

Signal

SLOW5 Index

Location

Read Size

#read_id	file_offset	rec_length
read-0	67	500000
read-1	500001	1000000
.	.	.
.	.	.
.	.	.
read-N	364459005610	1580072

Aim

Smaller SLOW5 encodings

Methods

SLOW5 Encodings

ASCII → 1. Binary → 2. Compressed

In 1 Byte?

ASCII		Binary
	VS	
0-9		0-255

Binary

VS

ASCII

```
##file_format=blow5v0.1
#read_id    n_samples   digitisation    offset  range   sample_rate raw_signal  num_bases   sequence    fast5_pa
$^@^@^@^@^@^@^@^@11b6cd19-3958-4264-a6f0-04aef956ebbd<8c>^W^@^@^@^@^@^@^@^@^@F^@^@@@<85>s·D^@^@zE]^EÈ^Bâ^BË^BÌ^BÍ^
^B^^^B^\^B^L^B^_^B^X^B<9c>^BÒ^BÙ^Bè^BÛ^B<9f>^Bn^B<89>^Bö^Bã^BÚ^B¸^BÇ^BÏ^Bª^B®^B¨^B^^Bk^Ba^Bg^Bh^B<87>^B«^B<9e>^B
^C^T^Cö^Bø^B^V^C^O^C^U^C^K^C^[^CÝ^BË^B×^Bê^BQ^B:^B:^Bs^B¦^B¨^B<9f>^B ^B¥^B<89>^B<8c>^B<8b>^B<95>^B²^B<91>^B<8e>^
^Bø^A^A^B^G^Bÿ^Aþ^A^E^B÷^A^K^B^E^Bþ^A
^Bû^A÷^Aÿ^Aü^Aø^Aú^A^Q^B<90>^BÁ^BÂ^B·^B°^B®^Bª^B¦^B<85>^B ^B<93>^B<93>^B<94>^B<9f>^B<8c>^B<94>^B<88>^B<88>^B<91>
^B^L^B^[^B#^B^Q^B^T^B^R^B^T^B^W^BË^BÒ^BÃ^BÂ^BÂ^B¿^B<83>^B<81>^B<8b>^B<96>^BÃ^B±^B»^B·^Bi^B?^B9^Bù^Aå^AÝ^Aß^AÞ^AÕ
^B^E^B^A^B:^BÊ^B®^B¸^B¹^B¼^B´^B«^B¶^B»^BÈ^BÂ^Bº^BÅ^B½^BÒ^BÐ^BÞ^BÍ^BÙ^BÓ^BÜ^BÈ^BÕ^BÂ^B×^Bê^BÛ^BÙ^BÚ^BØ^BÖ^BÁ^BÍ^B
^B^M^Bó^Aÿ^A^D^B^B^Bù^A?^Bm^Bo^B_^Bs^B<86>^B<83>^BÌ^Bà^BÕ^Bå^BÞ^BÇ^Bá^Br^B|^B<80>^Ba^B^Y^B^R^Bö^Aø^A^D^Bó^A^E^B^
^B^P^Bÿ^Aý^A^A^B^H^B^C^B^B^B^@^B^E^B^D^B^E^Bþ^Aþ^A^G^Bþ^Aû^A^E^B^L^Bÿ^A^B^Bú^A^B^B^M^Bÿ^A^O^B^G^Bú^A^@^B
^B
^B^B^B^E^B^M^B^D^Bÿ^A^@^B^A^Bù^A^N^Bý^A^A^Bõ^Aý^A   ^B^K^B^C^Bú^A^K^B^K^Bþ^A^K^B^A^B^@^B^@^B^N^Bû^A^D^Bý^Aþ^A^A^
^B^D^B^E^Bþ^A^D^Bî^A^H^B^E^B^A^B^@^Bø^Aú^Aú^A^A^Bö^A
^B  ^B^K^B^C^B^H^Bú^A^C^Bü^Aÿ^A^G^Bü^Aø^A
example.blow5                                                                                   9059,1,1        Top
##file_format=slow5v0.1
#read_id    n_samples   digitisation    offset  range   sample_rate raw_signal  num_bases   sequence    fast5_pa
11b6cd19-3958-4264-a6f0-04aef956ebbd    6028    8192.0  3.0 1467.6  4000.0  1373,712,738,715,716,719,728,724,727
a649a4ae-c43d-492a-b6a1-a5b8b8076be4    59676   8192.0  23.0    1467.6  4000.0  1039,588,588,593,586,574,570,585
c3491225-815c-408b-abc6-ed864f545f4b    37454   8192.0  14.0    1467.6  4000.0  1099,665,655,638,622,620,627,642
75d7303c-726a-407f-8df6-59e98ef86e34    15665   8192.0  18.0    1467.6  4000.0  1235,612,611,573,581,593,572,577
9dc4d6c4-1dc0-49d0-aaa2-078408a749cf    52190   8192.0  6.0 1467.6  4000.0  1023,606,624,546,600,581,569,564,580
52b95332-1cf5-4a6f-8bc4-88fbb1cb0c2c    44141   8192.0  3.0 1467.6  4000.0  1377,956,721,682,675,695,707,700,698
8c395415-c8d4-4476-b77c-30c878bd8a1d    57421   8192.0  3.0 1467.6  4000.0  1173,717,722,709,714,721,712,711,718
3fdd0b4a-2183-45ed-a817-c96e0b692df5    36568   8192.0  3.0 1467.6  4000.0  1370,765,680,684,696,696,684,667,612
ca0779cd-f7a9-4784-bd69-d50d61ce1c72    13002   8192.0  6.0 1467.6  4000.0  1257,658,527,534,533,524,531,520,527
627a9fdf-1655-4b39-a413-c8f0dfb73dc6    45690   8192.0  5.0 1467.6  4000.0  1255,773,617,574,568,555,557,554,551
~
~
example.slow5                                                                                   111,12,1        All
```

Compressed Binary SLOW5

```
^_<8b>^H^@^@^@^@^@^@^@^C%Ë1^NÂ0^LFáÙ½Fv^DCGÎbý%N°<94>8%véõ©ÄúÔ½<94><8a>6á2fG<·6Îõ{¿=<96>4^E<99>5<93>±£ïM<9c>²V^Mu<
84>^N£Q<8a>KÐ<84>U¡¿à<89><90>«<9c>ìZ^M<8d>ìè¼Á¯Õâs<88>½<84>
<VÞ^Qïå^GÒ<91>w z^@^@^@^_<8b>^H^@^@^@^@^@^@^C5<9a>^G|<8f>×÷ÇóÜ«ª<8b>Ö(-<9d>´ZU<9b>
~5j<8f><96><96>j<89>^R$f<83>$ö¨^]A^DA<8c><90><84><90>h<8c>Ø3F<90>È^R
^R<91><90>aïMKñ^?<9f>ÃÿÜ¼<9e>çûÌ{Ïø<9c>qÏ½<95>\^þjÖi]¯Øß<9a>®ßÖv-[ýÛ:
µêÕùÖ½<9e>g<8d>okÔq÷ðt-[Ï£wï¾³Ëºüÿ¯¥<8b>K<93>&Ó|v4wq^YÓ¢ÇkGM¡I6)&Í<9c>6'M<96><89>3ûÌ^F³Å^\3Ç¹*
`Ï5WÙÓÌ^Y<93>mr¸{Ò\`KãY¦É`;Î<97>    &<9d>ûcM<90>    0^KÌL3Ëø<9b>@³<98>}^QwFro²<99>Ç}?
3<9d>{<93>ÍBý^_ËÙ_¼9<8d>}^F÷æðv<88>Ycfs<9c>k<86><9b>¾ÆË^L<80><9a>lú;n<8a>Û ^[jCìZ;Û.³^A<9c>¯²^[ì"^[n<83>í,
î¯±avµ<9d>Ê<93>ùv<81>]b×Ùiv)oz[?ÛËv³^SØ^?·^Wà7Í¼g3ày¿Ùk^NÑr<90>ùÎT4®æ^SóµùÖTá¼^FûWÜ-c:
<9b>U&^Zþv<99>ÃP²ÛlGB^[¡v¼q7¿<99>^& y-<82>âIðãËq^D\^M2-LkÓÎ´dkc~5½M?øèo^F#  /3Ðø <9f><93>ô<9b>È¶ÄtãN_ãÆó    f*
\ÿdî:Ö\pî:÷<9d>Êôï^G^Eë<90>^4rH7×Ì^C<93>o¶AÉ>^S^E5;M,
ÚÚeö<98>xÞû<85>¯<9b><9a>º¦<99>io^Z<9b>F¦<8b>ùÑt4_<9a>§Î^Kç©s<8b>ý^^ÇÛÎEçuzÜÄv<90>ï^O©,6<9b>^CHf?
ÇÝpµ<9b>³Ý´¹<97>ë-ìþèg6z^Lf^OC^V^[ùn<89><89>D<8f>íMsÓ^V:Û±u¦7^_´u
ZOòu<8a><99>b~@ªî`,<91>~^Rh9<8d>ûòDÚ><8a>L^FòU^?<95>e7Ó^]i^MP<,0ËÍ»ö=û<99>=O/
Q&^Tú^F<98><8f>Ìûhæ#SÙ<94>2^_<98><8f>Í<87>¦<8c>ùÆ41<9f>s·ª©Âyeó<99>)ÇY^Uó<85>)Ë^[_p§^\÷>D^KóÍzt½^L^^öCÿ!
(JE^BqP<92>^C%§Ñ£+ßÕ<86><9f>ÚHÑ^[Ä^_<81><8b>Sì©Èi<^TplzÀë×´ú^E}µ<87>j7ó;ôwáØ^[¼|g¾GúÕ9¶1<9d>L+ôP^OÚÊ@
í^Sç]<8e>oBQ^Eó6H+^GÆ3<90>Ò^Uú^N^E^Uþæ<91>9O<8f>;<91>Q<9a>ù^[¬E<99>?@<92>;Øé^O^F¢Í
¶h<95>ÿ^F4=^OmÌÃVæÁÑ^R^ScÖ^Z^W{ÇÜ^C ^AjQÞ`l2xôB<8a>ïØâö]¶÷í^G¶<88>}ÝÞ¥×xz<8f>ÃÊÓy§#ÔwR¬^HfêÀÿwðÿÐük<8a>Ú^Mô?
^PTö<81><96>iHà^O4~^LÜE<81>U/ôÕ^W-^NE^V^Máó^[8¯f*ñäoè<89>@z·Ì]sß\§¥^\Þì<89>¶ÝÐ²^[_^N<83>«p¤~^P-$<80>´]èã$W©ØÅZøù
í^UEbÅ9¶B.ñÈj^[ÈÚ<8e>g^XÉÓ<9f>¾ÝÔnº^Y^Oãiº²{A«^[(^Q^?^S<82>·    Ab^?<99>¥à'<86>-
<9a>}½9<81>ðd<81>Â½H-^Q_5<86>ò^?<9d><93>N¦sÜ©Æ^WËAÁ"¾X^GMÑ c^C^ZÙH¯ëñf<9e>pÐ^Miõ^@ûGÑÒ68õE7^S<90>ô4ð1^Nì<8f>¢G<9
1><96>`¡&úo<86>D¾B2Uð)<95>¡å+[×Ö³õíwl5ms<8e>_Ø<8a>ì<95>m<´<89>UÄÑg,
Ö½^Kêg"ç<83>øv2Ç^\sÃ^T³7<91>éq0y^^I^UÐç^Thí<83><94>J<83>©<92>ìµñ<^M8þLÏ<8d>9w<85><8e>Fü·^@¥-Ðj^SÎ^[b<93>M@
ïB³<92>m'<9c>^Å«<9c>^A^Q©ô<95>
gëàg,¼<8c>ÇÒKØ<8b>ê-C<90>X^W<90>ý?Zn<87>N<9a>&lt;82&gt;ñÖÈÁÅ^5OAÔßÐ»^P^[^K<85>K?d¸éU,
^YBo-y÷^Gèj Ï·ypy<96>Þn£Ñëô|<86>÷î<99><8b>Hx:øø^@.>ÄBÊÁ<93><8b>¹í<p^^ã±^L^ZÝ£þj^K¨<8a>Dò<9b>ÑæF<90>²^T[Þ<8f>töÑJ,
ïÄâ«â *<81>7ö<81>¨XøIFRùD«ÍÈÅ3(^Y<88>¦^R¸w^XJv±ïyå^E^O«íÉ×±<9c>ïGÊÿAc!
2êÇÖ<97>ïú<83>Û<89><Iå<8d>^L<90>{L<91>»^^4íÔ¶â±;oìa^DÛP°ò+¶T    ½?
pN8^G<9d>t'ÇÉuòÀÚ^M'^[<8f>ì<82>^E^DÁM$ý<9e>&lt;80&gt;J<89>¹;<91>ü^Dº^XÉY^D^\<8a>5oS;Ù<86>^F$º^^E¶3xk^T¨^[gF<9b>?
ÁÛ^D<90>ç^K^V¼¹öçß^Wk^]ÆýÀ`Ã<97>/$Ò^\U~wÃ_
example.clow5                                                    9305,1,1        Top
```

gzip format

Algorithm

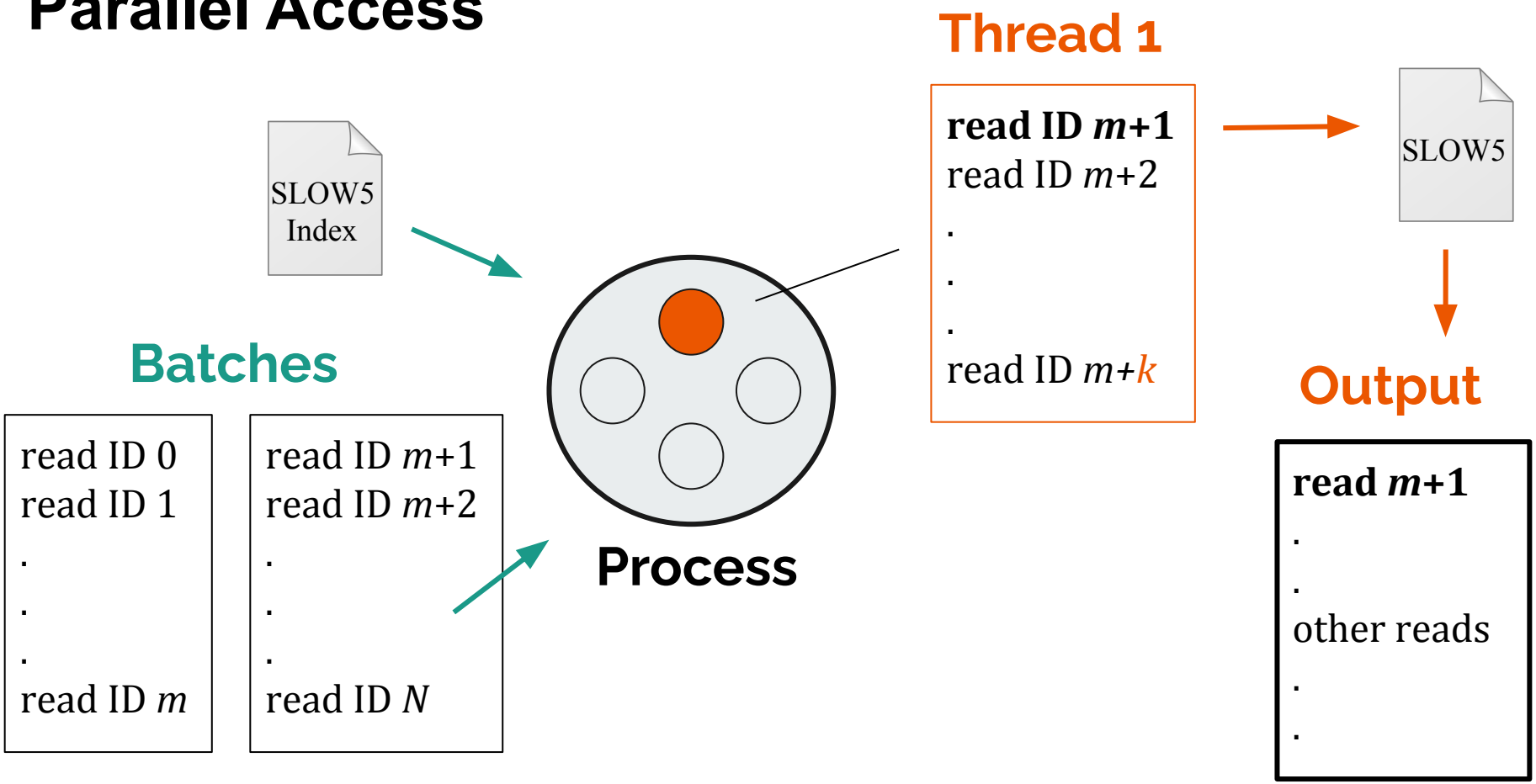
Input

read ID 0
read ID 1
.
.
.
read ID *N*

Output

read 0
read 1
.
.
.
read *N*

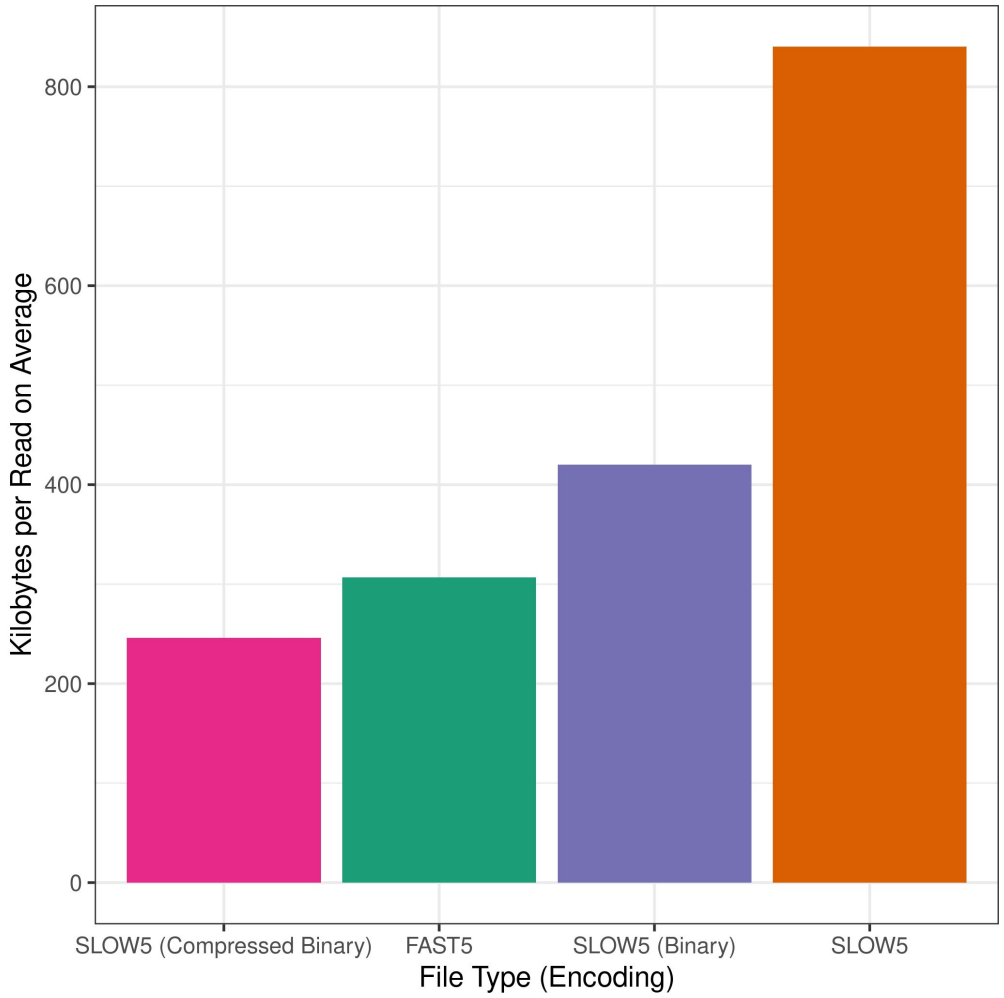
Parallel Access



Results & Discussion

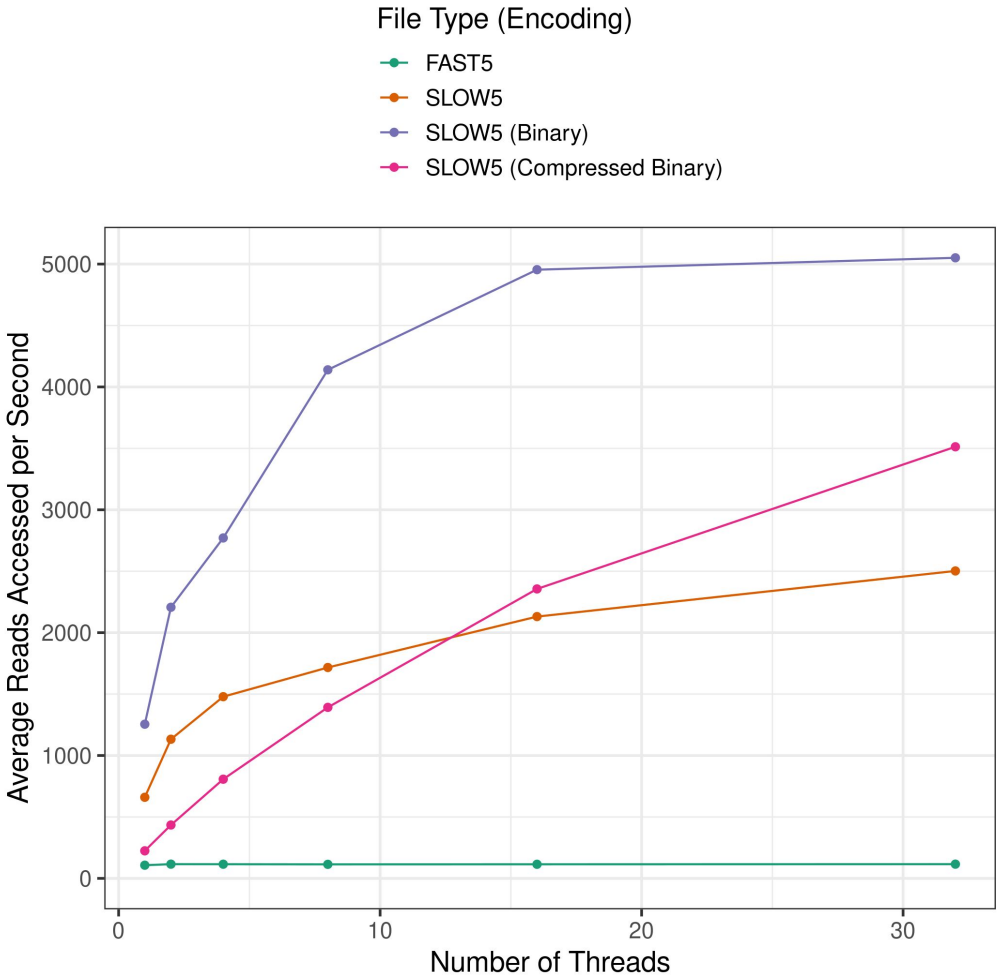
Benchmarking

1. File size
2. Access time



Compressed binary:

20% reduction in FAST5 size



E.g. 500 000 reads
FAST5: > 1h
SLOW5: < 5 min

Conclusion

- Compressed binary - most useful
- Binary encoding - fastest
- Both?