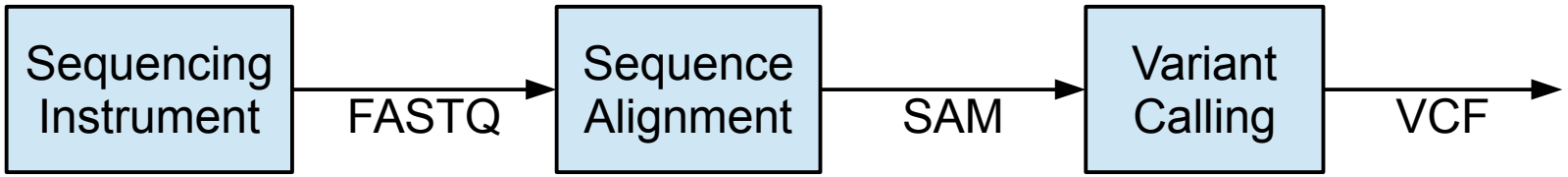
CRAM

Implementation & future directions.

James Bonfield,
Wellcome Trust Sanger Institute

Overview



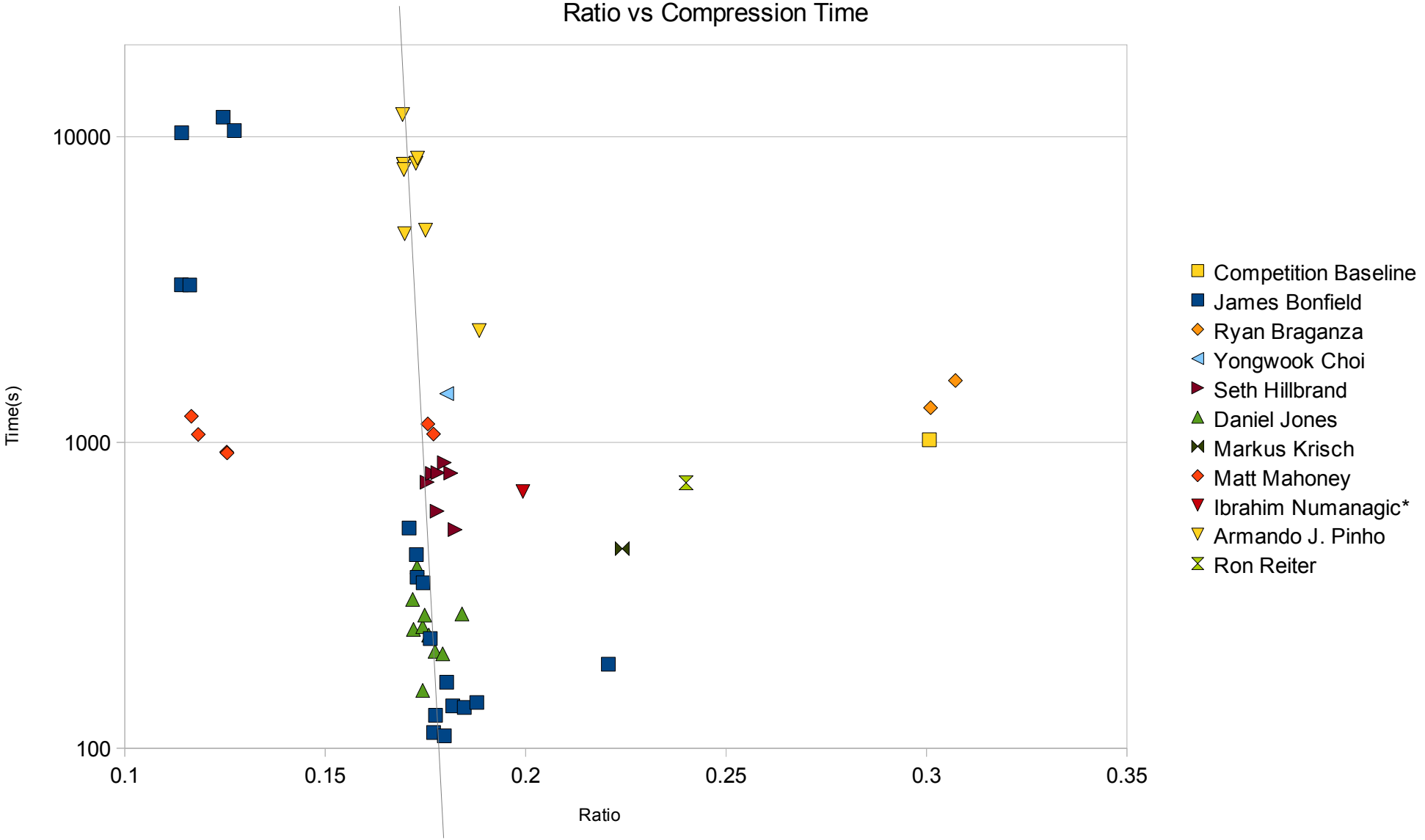
- FASTQ
- SAM / BAM
- CRAM
- Comparisons with other tools
- Additional / experimental codecs
- Lossy compression
- Future directions

FASTQ Format

- @Identifier (name), Sequence, (“+” delimiter), Quality.
- Represents unaligned DNA fragments as they come off instrument.

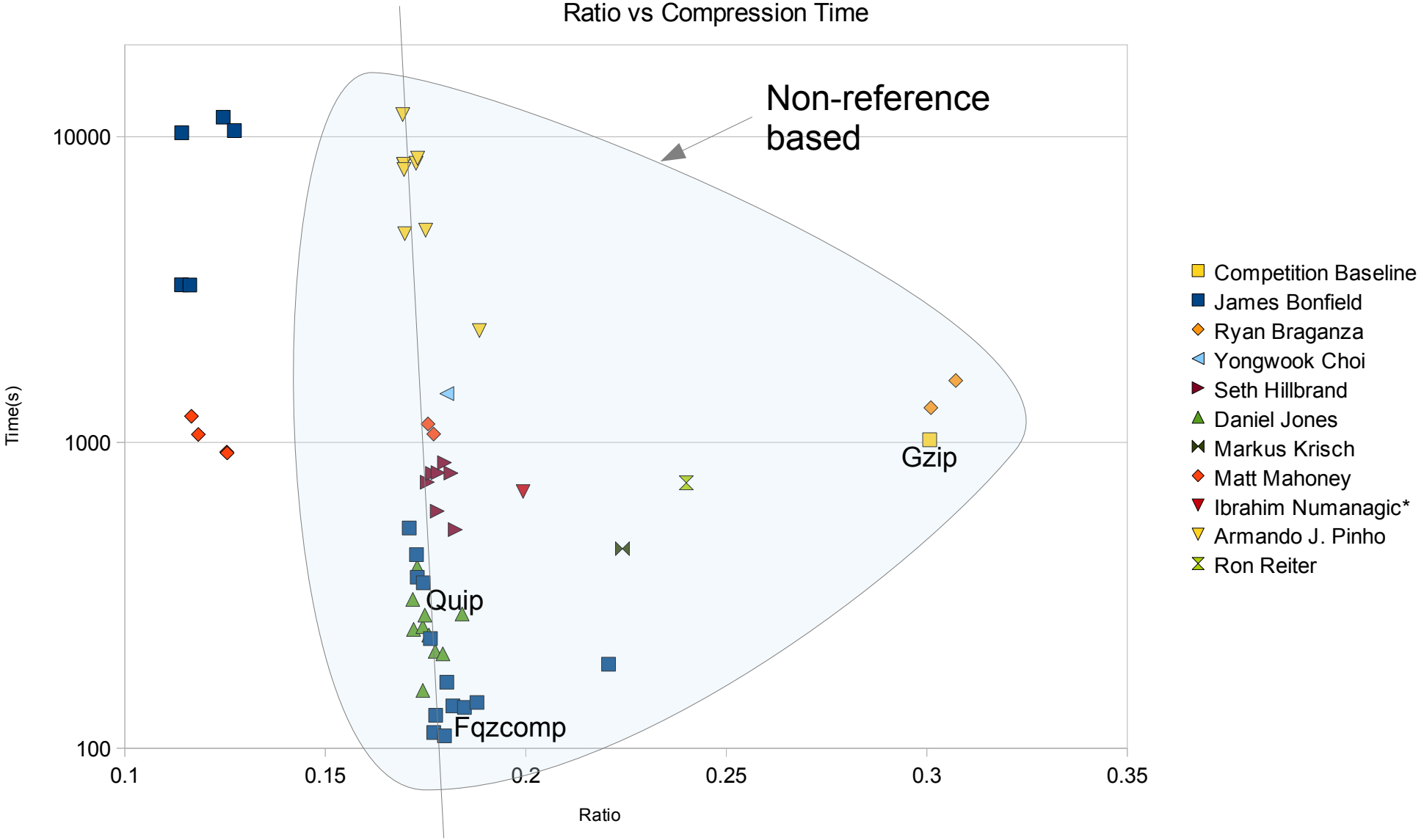
```
@HS25_09827:2:2211:17372:23675#49/2
AAAATTATTAATATTACTGCTGTTAGAGAAATGAATGAGCCTACAGATGATAGGATATTTCATGTGGTGTATGCATCGGGGTAGTCCGAG
+
;CA8DBCCDCF<HEECCDEBD?>78CE@FEGD>@4?D,G5AFEHCCGBF>8DAD>5C@HDDHF?5?HCD=H>H4ED>HB>DF=C;CCA4E
@HS25_09827:2:2109:13544:26093#49/1
TGAAAATTATTAATATTACTGCTGTTAGAGAAATGAATGAGCCTACAGATGATAGGATATTTCATGTGGTGTATGCATCGGGGTAGTCCG
+
CABDFGEFJFGEGJGFGGGF>HIKEJFIFHHHHIIGGHEGGHFHGGHIGEIEFIGIFHFFIGGDFKCGGAHFGGGGHGHHGGIFGGFHFH
@HS25_09827:2:2103:14309:24962#49/2
CGATGCATACACCACATGAAACATCCTATCATCTGTAGGCTCATTCATTTCTCTAACAGCAGTAGTATTAATAATTTTCATGATTTGAGA
+
;CABDDD@FFFGHAGGFFIFFCJFIHEIIEGGFHGCHGCGFEHHFGEEFGHGGGIFEFFFHH?HGJFGFGFHHHEEFIFIHGEHIICEDE
@HS25_09827:2:2211:3294:20164#49/2
CATACACCACATGAAATATCCTATCATCTGTAGGCTCATTCATTTCTCTAACAGCAGTAATATTAATAATTTTCATGATTTGAGAAGCCT
+
<CBCDDGCFEHGAG?EEGDGFGHIHHFEGECGFDGEDGCGHDHH?GGEJ>FDGEEFG@FHGFHFGEFGFDFF6BFABHD>EEFF;IHFHH
@HS25_09827:2:2215:14172:53309#49/2
GCTTCGAAGCGAAGGCTTCTCAAATCATGAAAATTATTAATATTACTGCTGTTAGAGAAATGAATGAGCCTACAGATGATAGGATATTTC
+
;?9C>CC<;H7/HGIGF=8=FHJF;BGIIIGA@@9G:JGGDGHHFGE5FFHHGFGHFBEIDGHH,HFHFIEHHGIH4HGIHEFHFIHCHG
```

SequenceSqueeze FASTQ Leaderboard



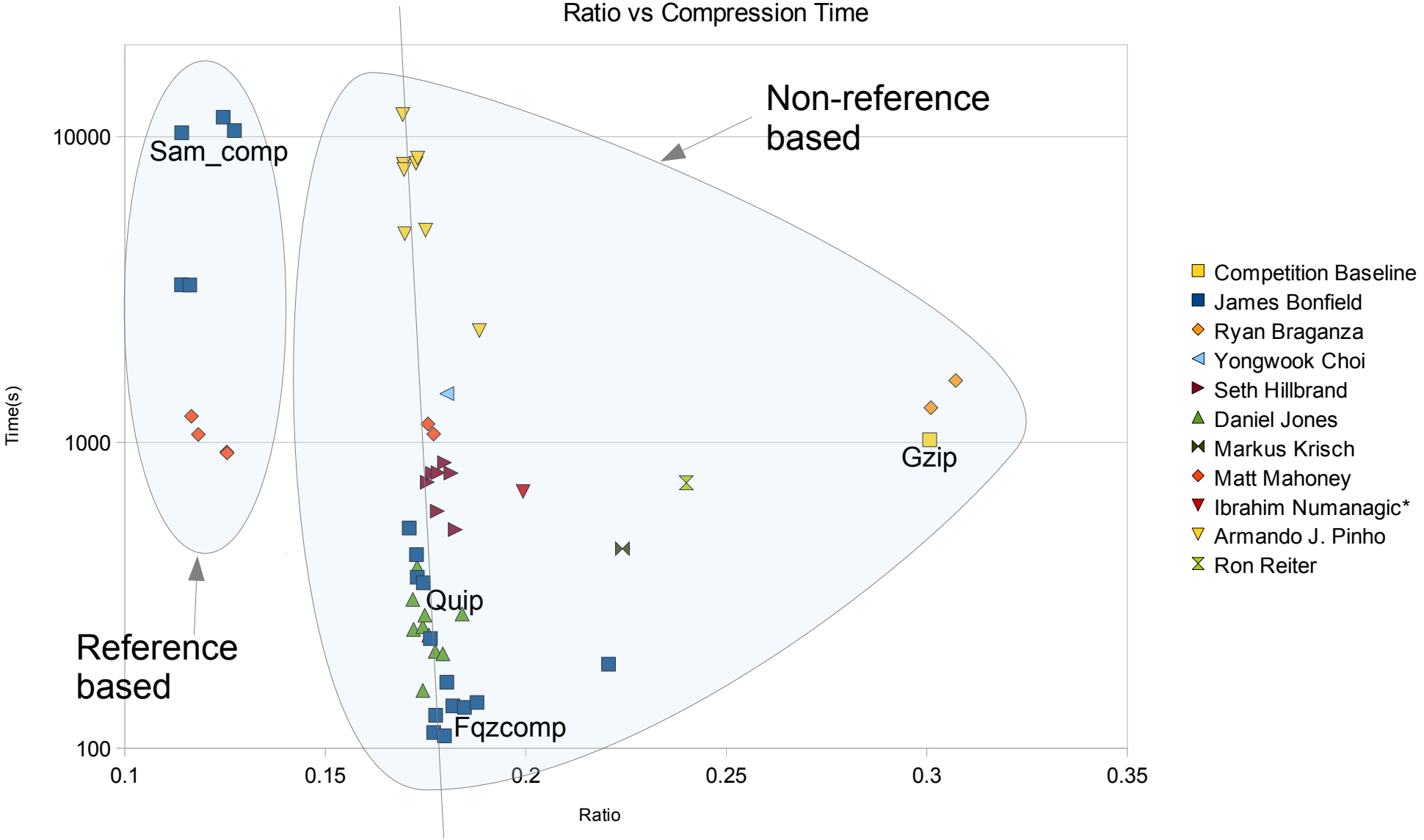
wellcome trust sanger institute

SequenceSqueeze FASTQ Leaderboard



wellcome trust sanger institute

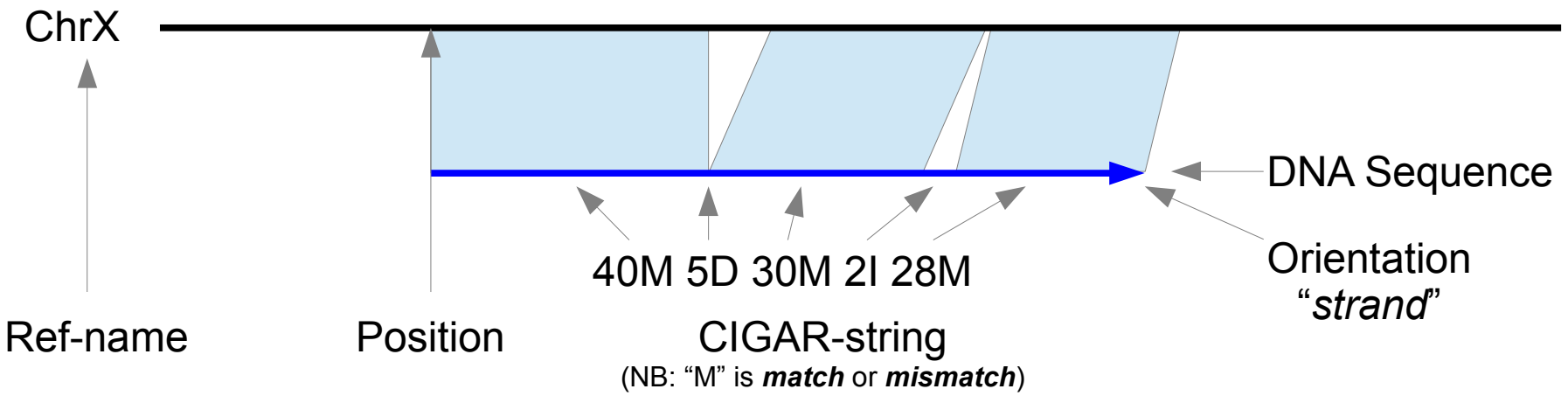
SequenceSqueeze FASTQ Leaderboard



wellcome trust sanger institute

SAM format

- One record (a single DNA fragment alignment) per line describing alignment between fragment and reference.
- 11 fixed columns + optional key:type:value tuples.



SAM format

- One record (a single DNA fragment alignment) per line describing alignment between fragment and reference.
- 11 fixed columns + optional key:type:value tuples.

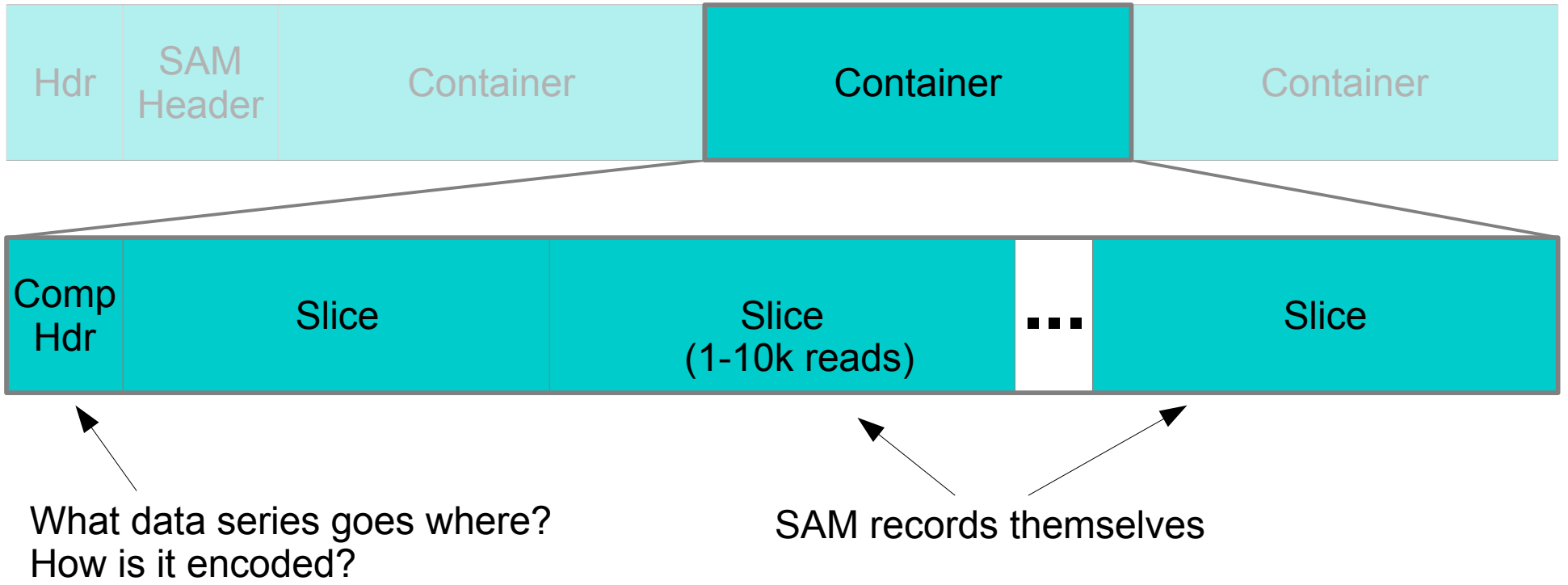
```
HS25_09827:2:2206:15555:91393#49        147     1       172818  0       100M    =       172413  -505
GAGGCATGGAGAGTTCACAGCTACCAAGTGTAGGAGTCTGGATTCAAAGCCAACGGCGTGACTCCAAAGTCCCTGCCCTAGCCCCTGGACCACCCTTGCA
GEHHGF?GGLHGG4IIDGGFJFHGEEFHKGDFEHIGFFEFFJHFEEJGHHIJIHGEHHJHG=6HEEHEGGIIFGEBCFEFGIGHGCGHGFGEBDEDDBC<
X0:i:5  X1:i:6  MD:Z:100        RG:Z:1#49       XG:i:0  AM:i:0  NM:i:0  SM:i:0    XM:i:0  XO:i:0
XT:A:R
HS25_09827:2:1206:20031:27537#49        83      1       173002  0       100M    =       172621  -481
ACCTACCCCTGGTGCCCCGCCTCTCACCACCCTTCTTCCTGCTTTTACCTCAACCCCTACACAAAGCCTGGGCCACTTAATGTGGCATCAAACAGACGCC
BIHBC+HGG<HF>>GJHIGJH?GFGFIGDHFGGGFEHGGIFIBDCLHHIIHF?HGIIJDD@HHGBIIFHHHGIFDHGIHGHFGGHGJGGGFIGECFFBAC
X0:i:1  X1:i:4  BC:Z:NGTCTATC   MD:Z:5A39A54    RG:Z:1#49       XG:i:0  AM:i:0    NM:i:2  SM:i:0
XM:i:2  XO:i:0  QT:Z:!4:BDDD=   XT:A:U
```

- 1 byte quality per DNA base (likelihood of error)
- BAM is just serialised, binary encoded and gzipped version.

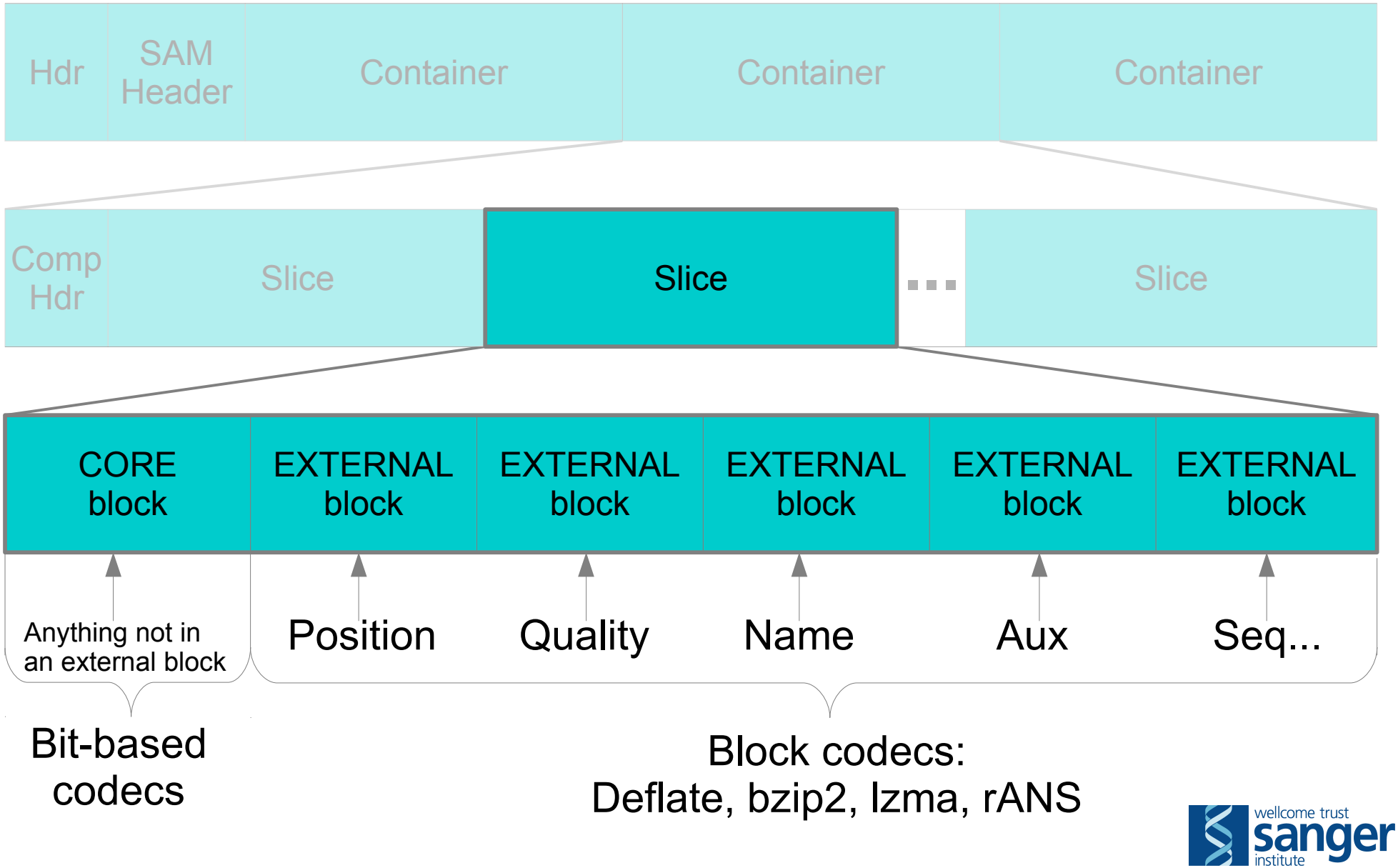
wellcome trust sanger institute

CRAM Internals

CRAM Internals



CRAM Internals



wellcome trust sanger institute

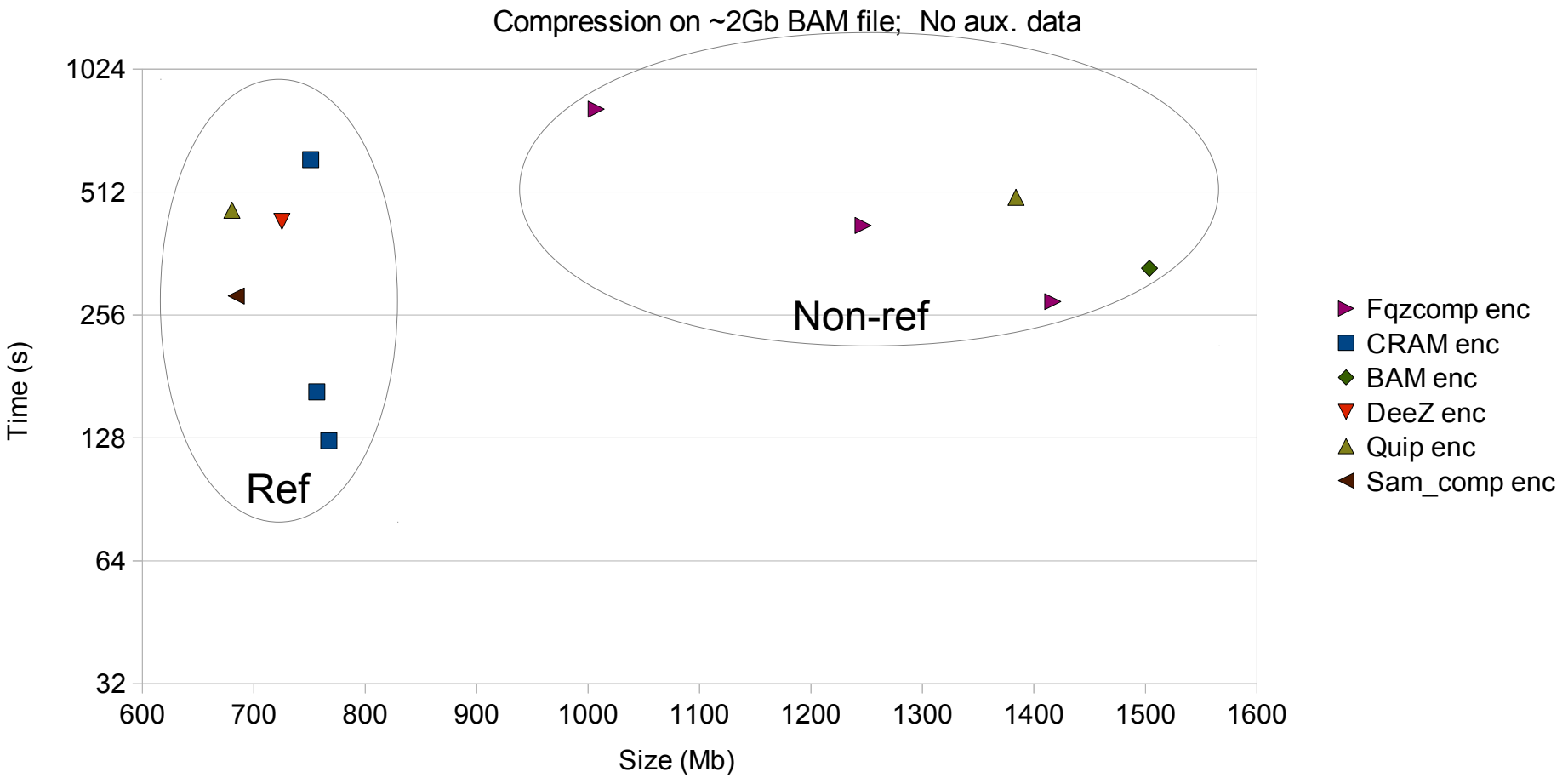
cram_dump: Aggregated Data Series

```
Block CORE          , total size      78256
Block content_id  11, total size   96112606 x      RN Identifiers (custom)
Block content_id  12, total size  520881483 x      QS Quality Scores (custom)
Block content_id  13, total size     187797 gb r   IN Inserted bases (mostly gzip)
Block content_id  14, total size   11436419   l R  SC
Block content_id  15, total size   10747303  b  R  BF Flags (mostly rANS order 1)
Block content_id  16, total size    4269965    rR  CF
Block content_id  17, total size   11618289    r   AP Position (rANS order 0)
Block content_id  19, total size    1992014 g lr   MQ
Block content_id  20, total size     829444 g      NS
Block content_id  21, total size     322980    r   MF
Block content_id  22, total size    1394973 gbl    TS
Block content_id  23, total size    5228135  bl    NP
Block content_id  24, total size   11643750 g l    NF
Block content_id  25, total size    1579439  b  R  RL
Block content_id  26, total size    5767216 g  r   FN Feature; no. of
Block content_id  27, total size    2344908    rR  FC Feature code
Block content_id  28, total size   17366343 g l    FP Feature position
Block content_id  29, total size      28994 g  r   DL
Block content_id  30, total size     722651 g      BA
Block content_id  31, total size    3731687     R  BS
Block content_id  32, total size    1921980 g lrR  TL
Block content_id  36, total size    1711720 gb     HC

Block content_id  47, total size    8485619 g lr   ASC
Block content_id  48, total size   11875678   lr  XSC
Block content_id  49, total size    1222057 g  r   MQC
Block content_id  50, total size   35519403  bl   MSS
Block content_id  51, total size   39629656   l   MCi Aux. MC:i:<int> tag (lzma)
Block content_id  52, total size   29451404  b    SAZ Aux. SA:Z:<str> tag (bzip2)
Block content_id  53, total size   15520701  bl   XAZ
Block content_id  54, total size      48737 gbl   MCS
Block content_id  55, total size     128305 g     asC
```

wellcome trust sanger institute

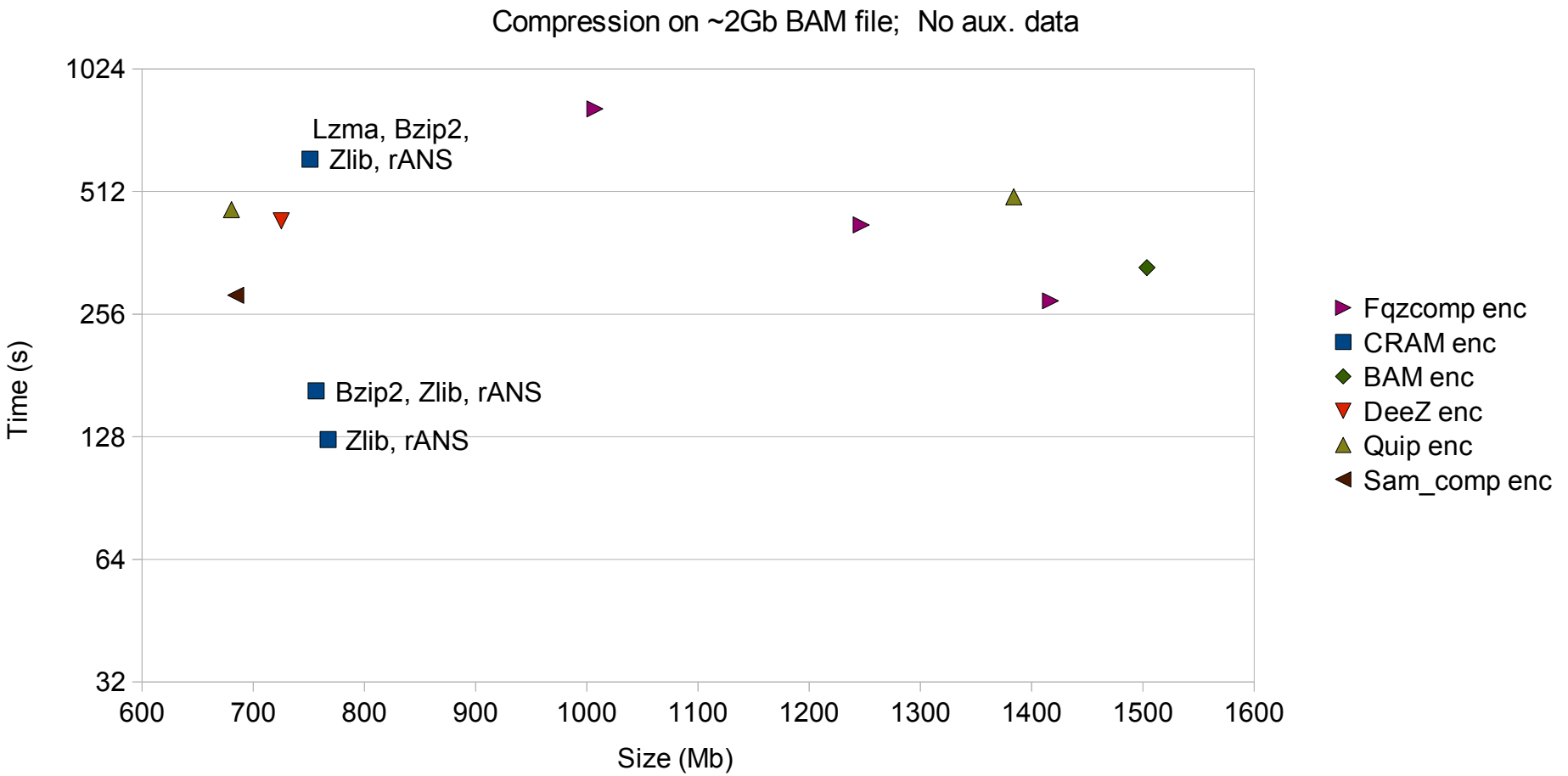
Results: No Aux. Data



Random access: BAM ~200/block. CRAM 10,000/block. DeeZ 1,000,000/block
Quip & Sam_comp no random access.

wellcome trust sanger institute

Results: No Aux. Data

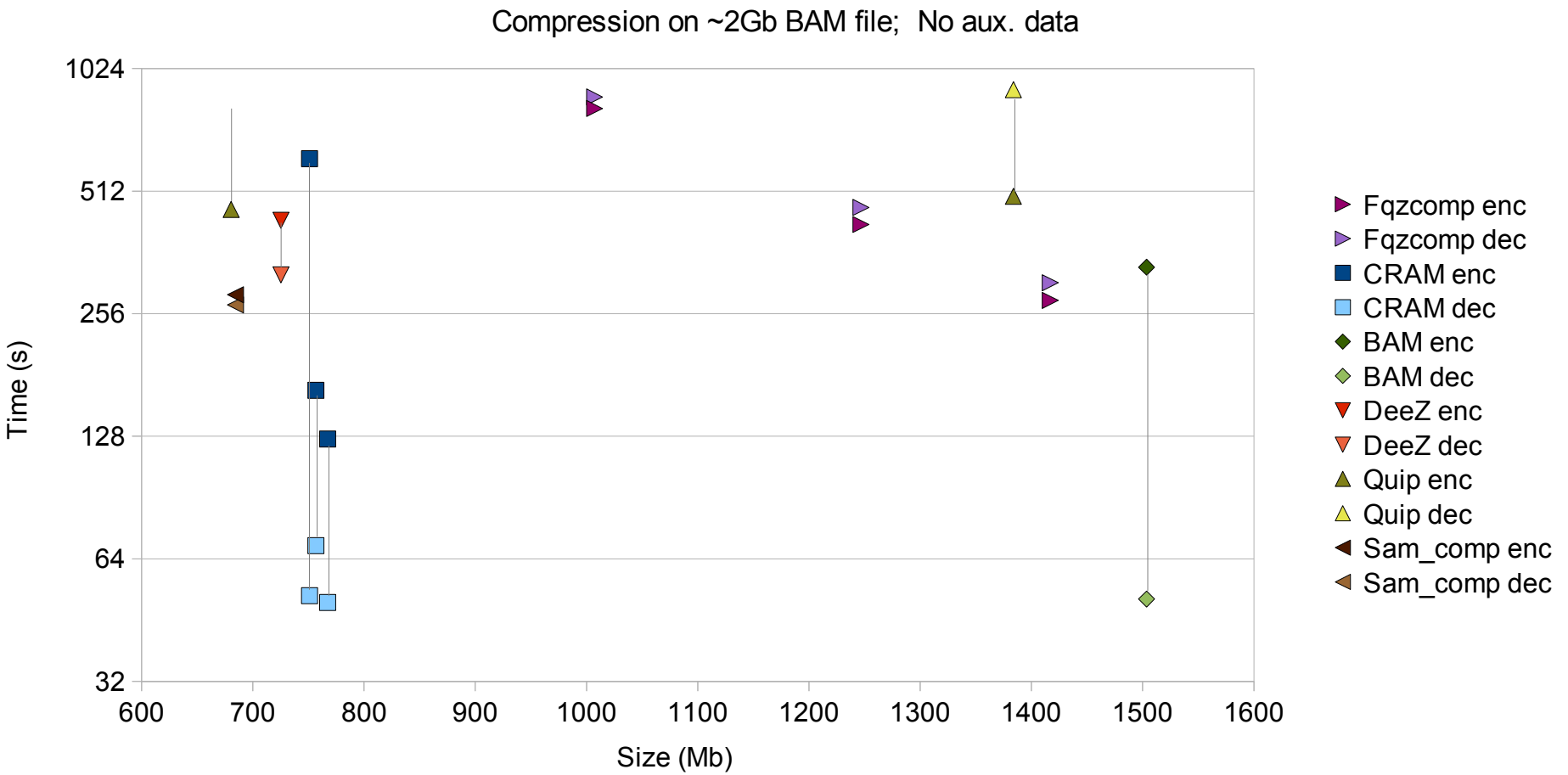


CRAM has built-in support for Deflate (Zlib), Bzip2, LZMA and rANS (Asymmetric Numeral Systems; J.Duda)

Quip smallest; ~10-13% smaller than CRAM, but no random access.

wellcome trust sanger institute

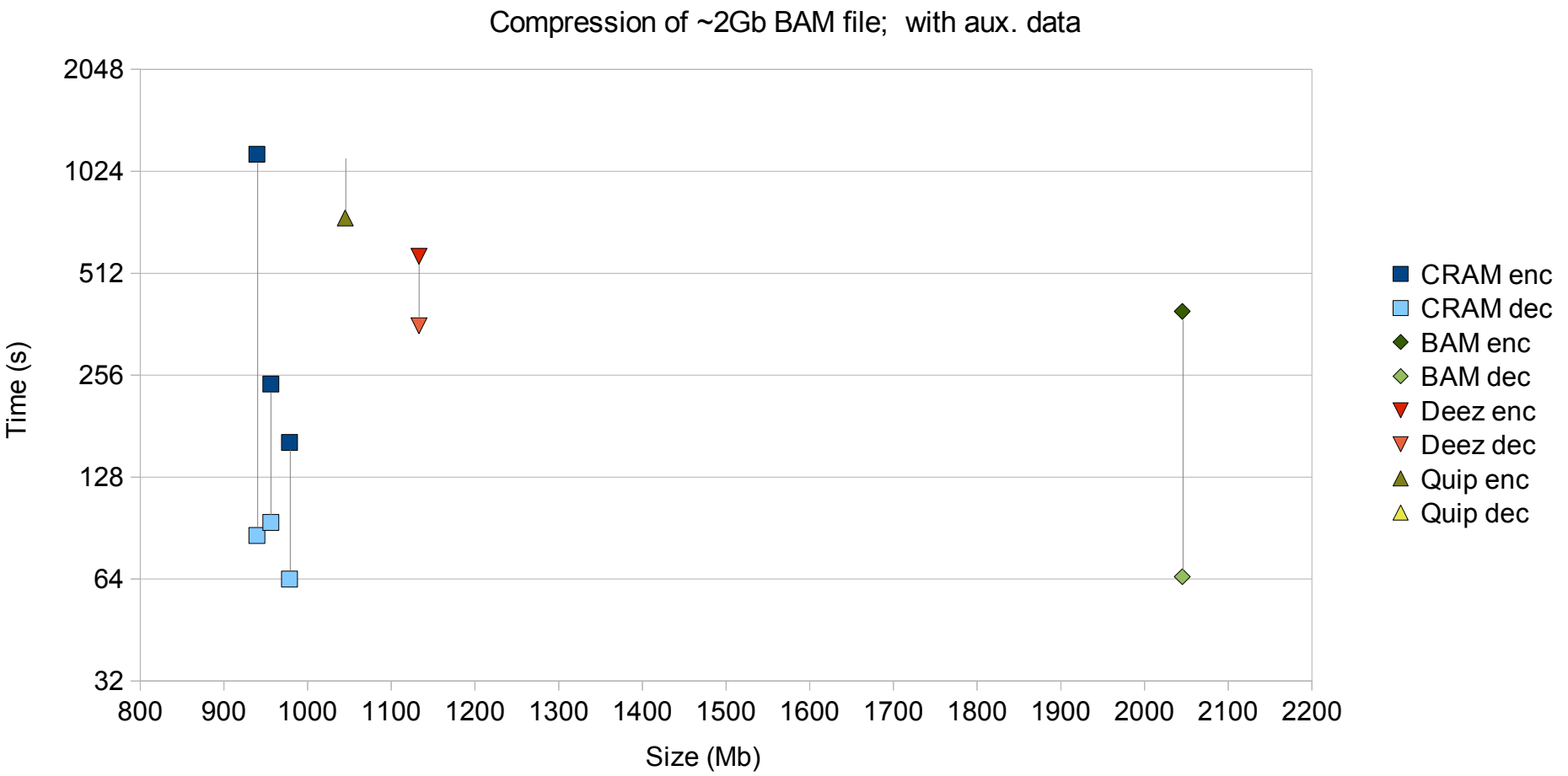
Results: No Aux. Data



Decoding: strongly asymmetric times from LZ based methods vs statistical modelling.

wellcome trust sanger institute

Results: With Auxiliary Fields

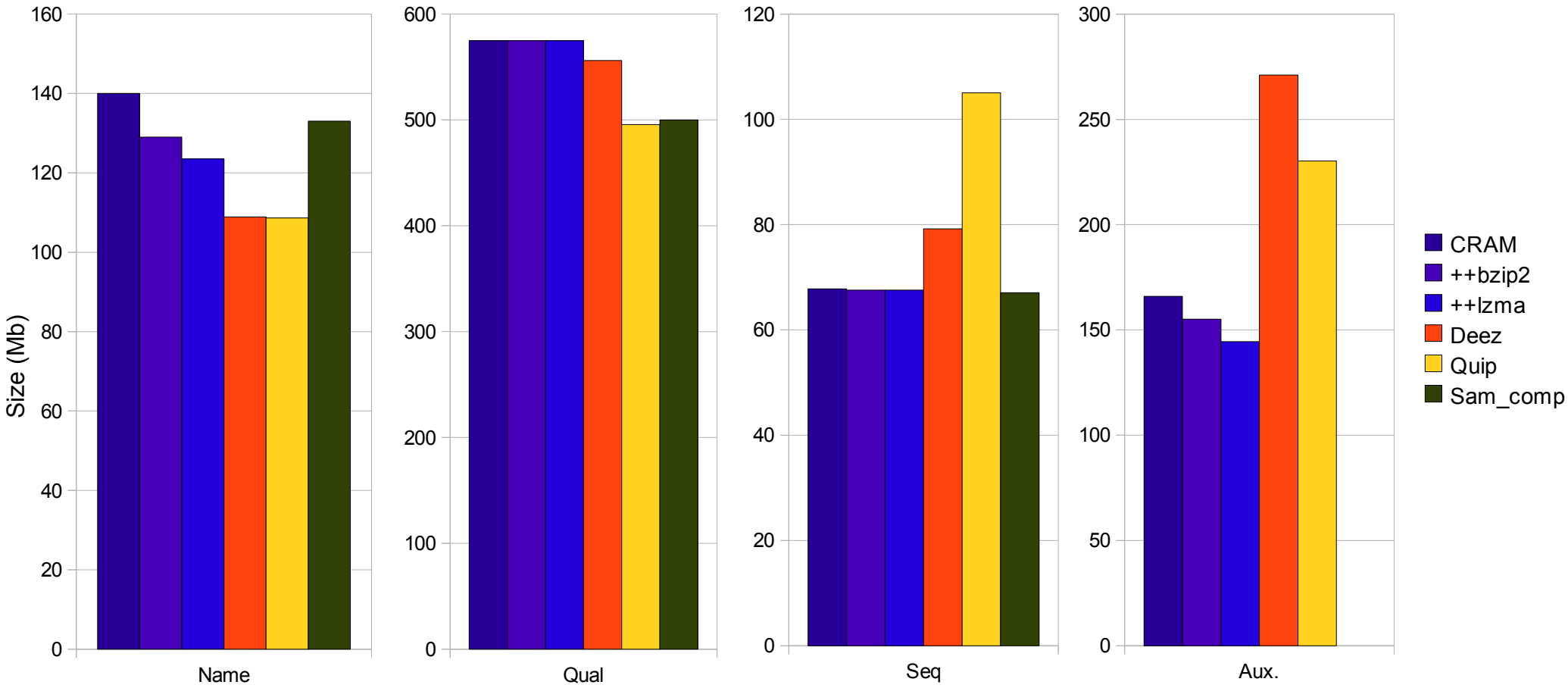


=> CRAM is strong with auxiliary fields, less so with fastq fields

wellcome trust sanger institute

Sequence, Name, Quality, Aux...

- State of the art would be a mixture of several methods. (NB: not tested on wide variety of data.)

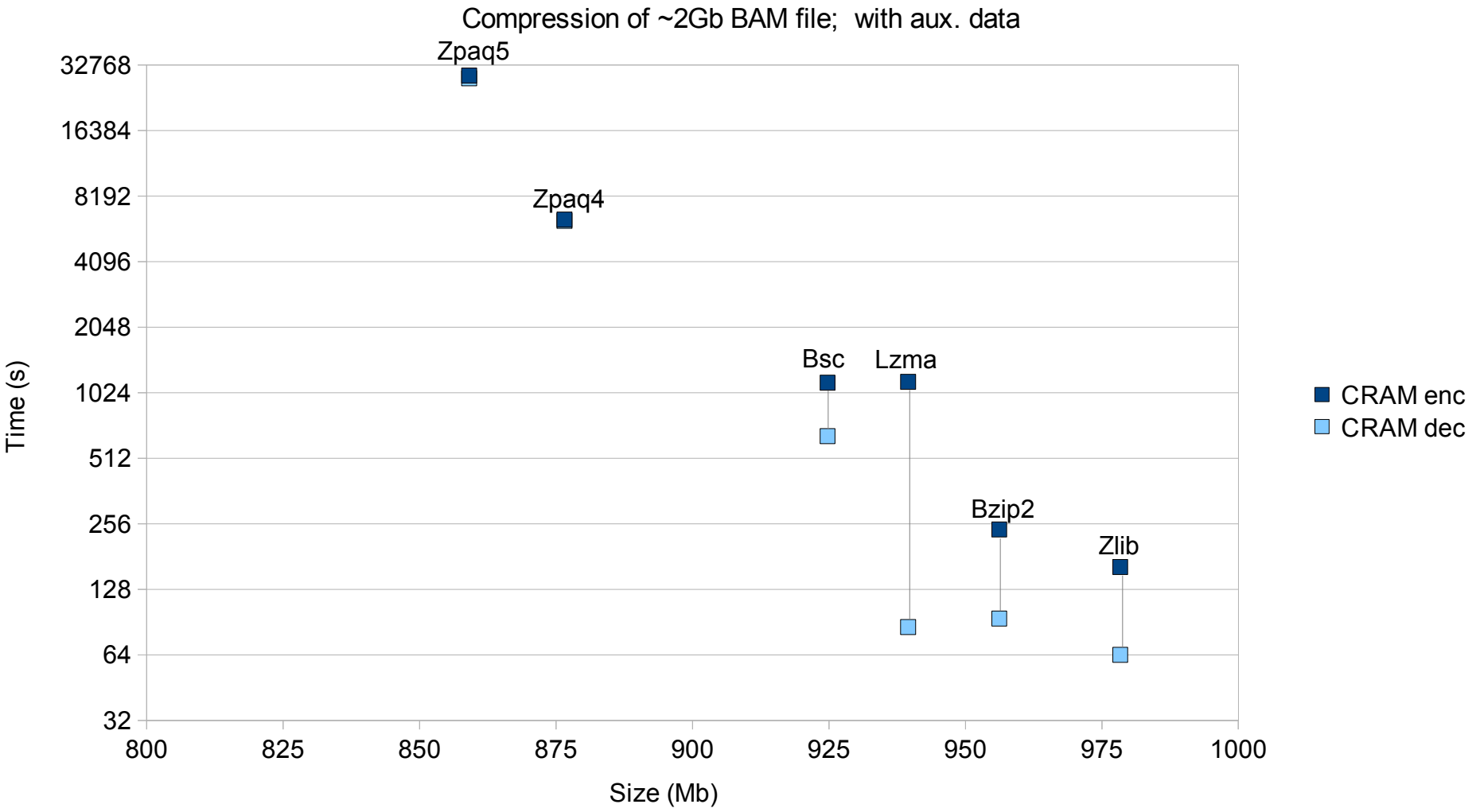


wellcome trust sanger institute

Codec Plugins

- Experimental branch of *Scramble* with plugins to allow use of arbitrary block compression methods.
  - Useful to use Squash: https://github.com/quixdb/squash/
  - Can experiment with extreme cpu/size trade-off to discover data correlations. (ZPAQ)
  - Allows for dedicated name and quality codecs from fqzcomp.
- Codec selection: try all and learn which works best.
  - Avoids over-tuning of e.g. Illumina read names vs 454.
  - Some data series for unknown data types (auxiliary tags).
- Do we want plugins long-term?
  - Proliferation of codecs in the wild; maybe not desirable.

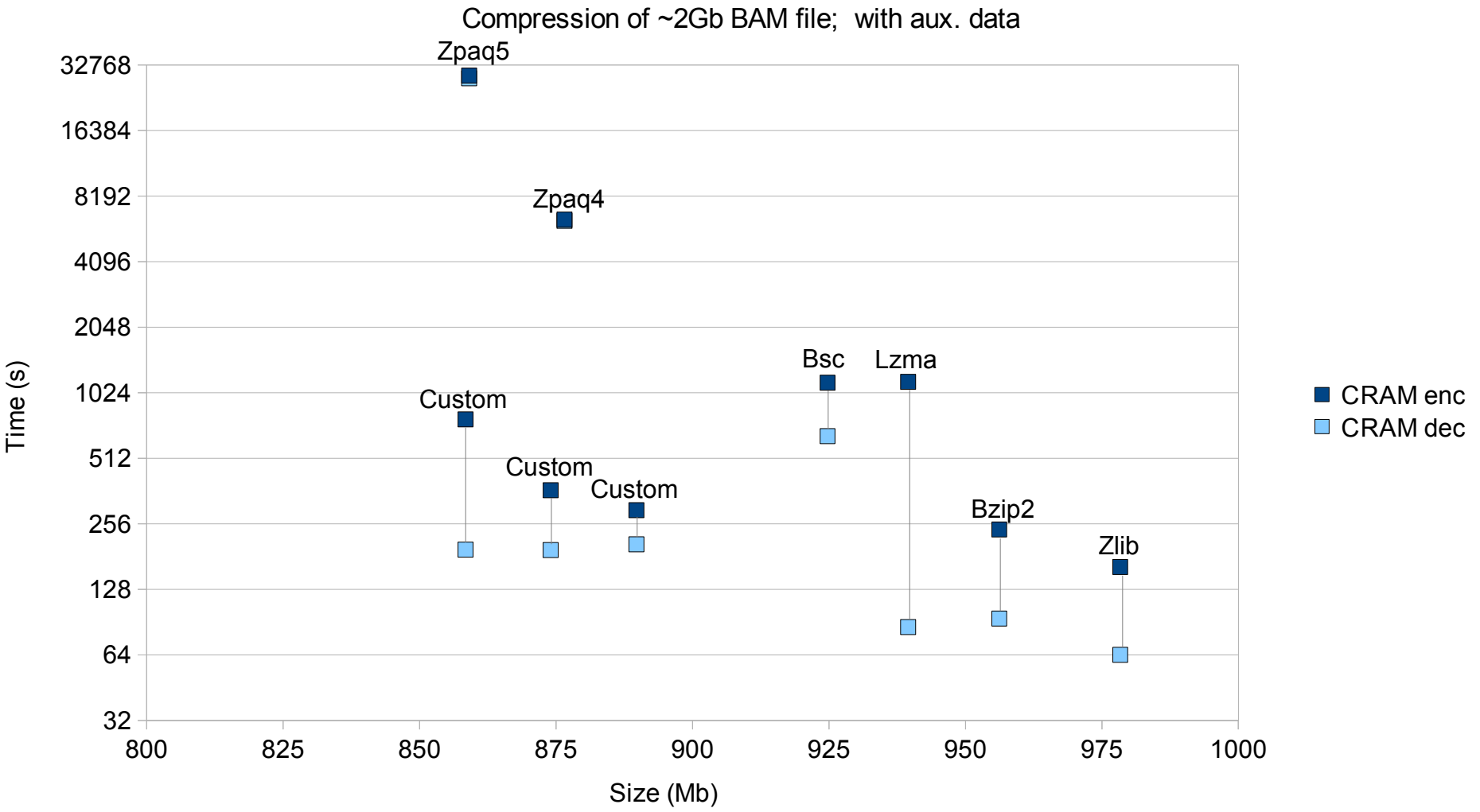
Results: With Plugins



Generic compression algorithms via Squash: Bsc, ZPAQ

wellcome trust sanger institute

Results: With Plugins

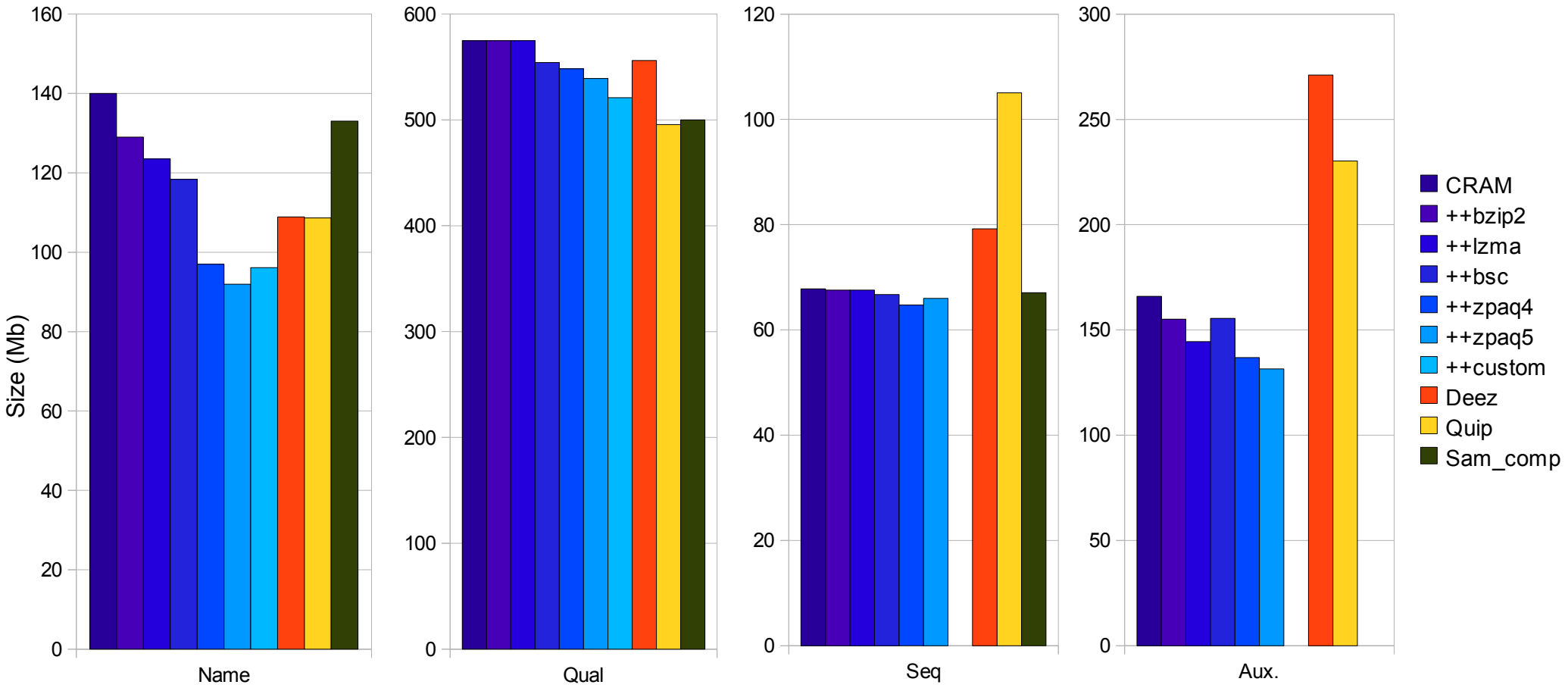


Custom fqzcomp derived codecs.

wellcome trust sanger institute

Sequence, Name, Quality, Aux...

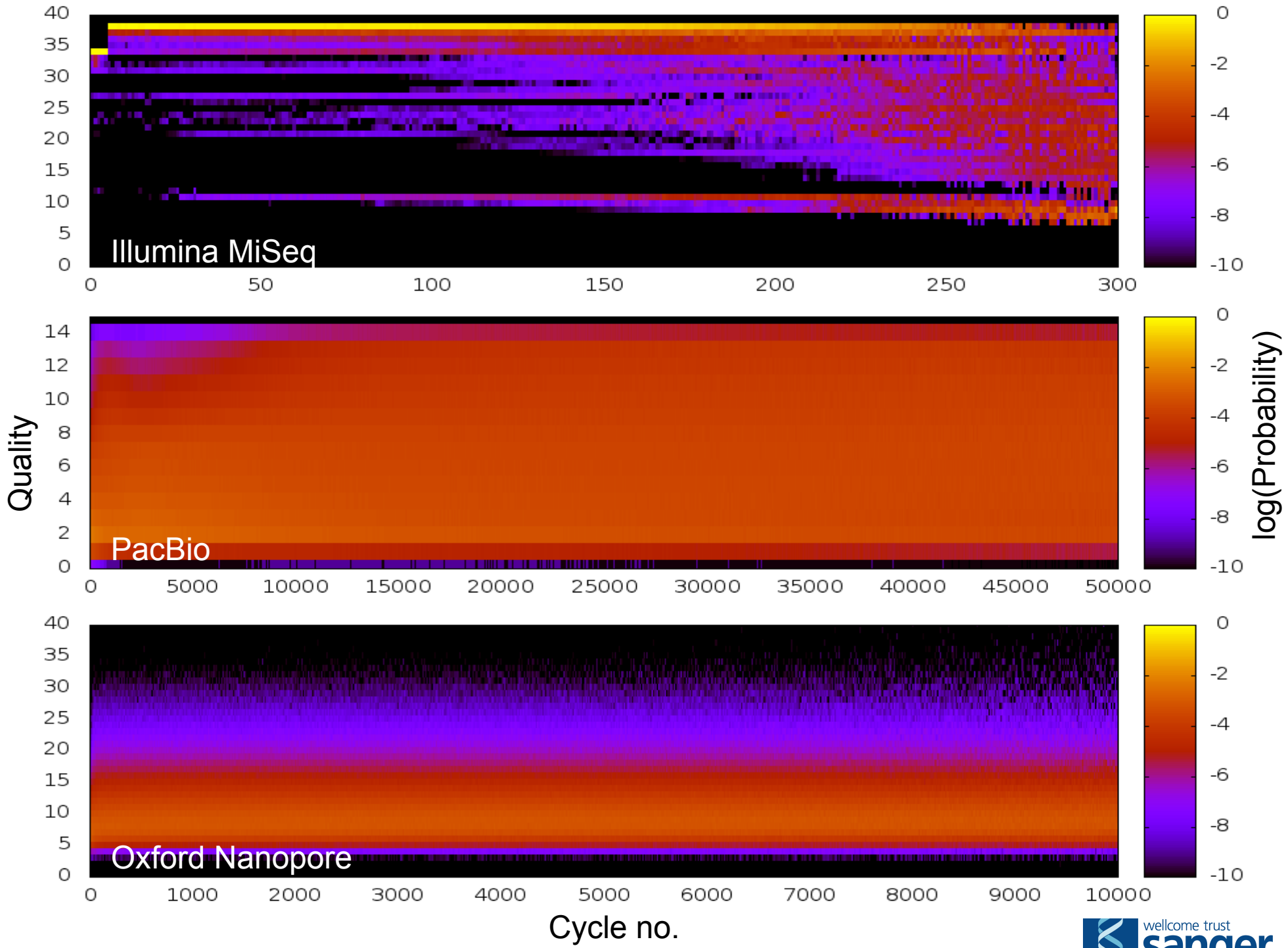
- State of the art would be a mixture of several methods. (NB: not tested on wide variety of data.)



wellcome trust sanger institute

Other Instrument Types

- Illumina: Short, but very high quality.
  - Current bulk of sequencing data.
- PacBio, Oxford Nanopore Technology: Very long, lower quality.
  - Assembly + consensus caller (quiver) for high quality.
  - ONT: true single molecule sequencing.
- Epigenetics / mthylation
  - Modified base calls, not just ACGT.



wellcome trust sanger institute

Lossy Compression

- Quality values are the largest block of data.
  - Originally ~40 discrete values, now quantised to 8.
- Better alternatives to quantisation.
  - Per record: smoothing + RLE, rate distortion.
  - Whole dataset: decisions on which to keep / throw away, *k*mer analysis, BWT (BEETL), De Bruijn graph. **Not every quality value is useful.**
  - Quartz, QVZ, QualComp, LFQC, Leon, …
- Evidence shows some lossy quality compression ***improves*** the downstream analysis tools. Denoising.
  - Evaluate based on downstream analysis methods.

LEON paper: Table 2

SNP calling precision/recall test on data from human chromosome 20, compared to a gold standard coming from the "1000 genomes project"

Prog	Precision	Recall	Ratio
lossless	85.02	67.02	2.95
SCALCE	85.15	66.13	4.1
FASTQZ	85.46	66.63	5.4
LIBCSAM	84.85	67.09	8.4
FQZCOMP	85.09	66.61	8.9
LEON	**85.63**	**67.17**	11.4
RQS	85.59	67.15	12.4
no quality	57.73	68.66	0

No quality means all qualities were discarded and replaced by 'H'. The ratio is given by the original quality size divided by the compressed size. For the lossless line, the best compression ratio obtained by lossless compression tools is given (obtained here with FQZCOMP).

Benoit et al. BMC Bioinformatics 2015 16:288 doi:10.1186/s12859-015-0709-7

Reference-free compression of high throughput sequencing data with a probabilistic de Bruijn graph

What's Next?

- Simplify – CRAM is 23K lines of code; BAM ~4K.
- New DNA library preparation methods to produce consensus sequences. Consensus enough?
  - PacBio quiver
  - 10X Genomics
  - Moleculo
- Compressed sequence representation for analysis
  - BWT for string searching & variant detection
  - Align directly from BWT of fastq.
- In 10 years, will we be wanting raw sequence at all?
  - Raw data (Fastq) → aligned data (SAM) → variant calls (VCF)
  - Little work so far on BCF.

Acknowledgements

- Sanger colleagues
  - Rob Davies, David Jackson.
- CRAM authors
  - Vadim Zalunin
  - Markus Fritz (doi: 10.1101/gr.114819.110)
- Compression tools
  - **DeeZ** (Faraz Hach) doi: 10.1038/nmeth.3133)
  - **Quip** (Daniel Jones) doi: 10.1093/nar/gks754)
  - **Samtools** (Heng Li) doi: 10.1093/bioinformatics/btp352)
  - **Scramble** (doi: 10.1093/bioinformatics/btu390)
  - **ANS** (Jarek Duda) http://arxiv.org/abs/1311.2540)
  - **Quartz** (Y. Yu) doi:10.1038/nbt.3170)
  - **QVZ** (G. Malysa) doi: 10.1093/bioinformatics/btv330)
  - **QualComp** (Idoia Ochoa) doi:10.1186/1471-2105-14-187)
  - **LFQC** (Sudipta Pathak) doi: 10.1093/bioinformatics/btu701)
  - **Leon** (Gaëtan Benoit) doi:10.1186/s12859-015-0709-7)
  - **ZPAQ** (Matt Mahoney)
  - **Zlib** (Jean-Loup. Gailly, Mark Adler),
  - **Bzip2** (Julian Seward; Burrows & Wheeler)
  - **Lzma** (Igor Pavlov)
  - **Bsc** (Ilya Grebnov)

wellcome trust sanger institute