CERN openlab

The GeneROOT Project
Status and Plans

Fons Rademakers

CERN openlab Chief Research Officer

GeneROOT - Using ROOT for Handling Genomics Data

King College London - TwinsUK Project

- Collaboration between the KCL and CERN openlab
- Try to optimize genomics data storage and processing using HEP tools
- Working on a local 400TB copy of TwinsUK data
  - 750 Monozygotic twins
  - 900 Dizygotic twins
  - 138 Singletons

TwinsUK

The TwinsUK resource is the biggest UK adult twin registry of 12,000 twins used to study the genetic and environmental aetiology of age related complex traits and diseases.

Latest News

World Alzheimer's Day 2017: Q&A with Dr Claire Steves

[illegible]

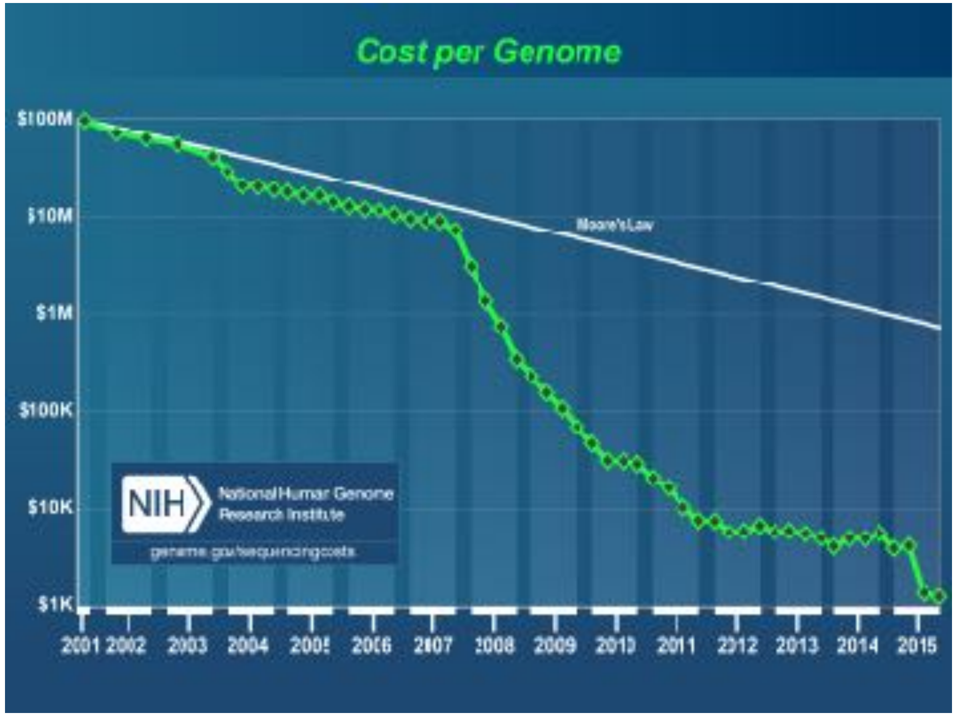
SEEKING TWINS

ARE YOU A TWIN OR DO YOU KNOW A TWIN?

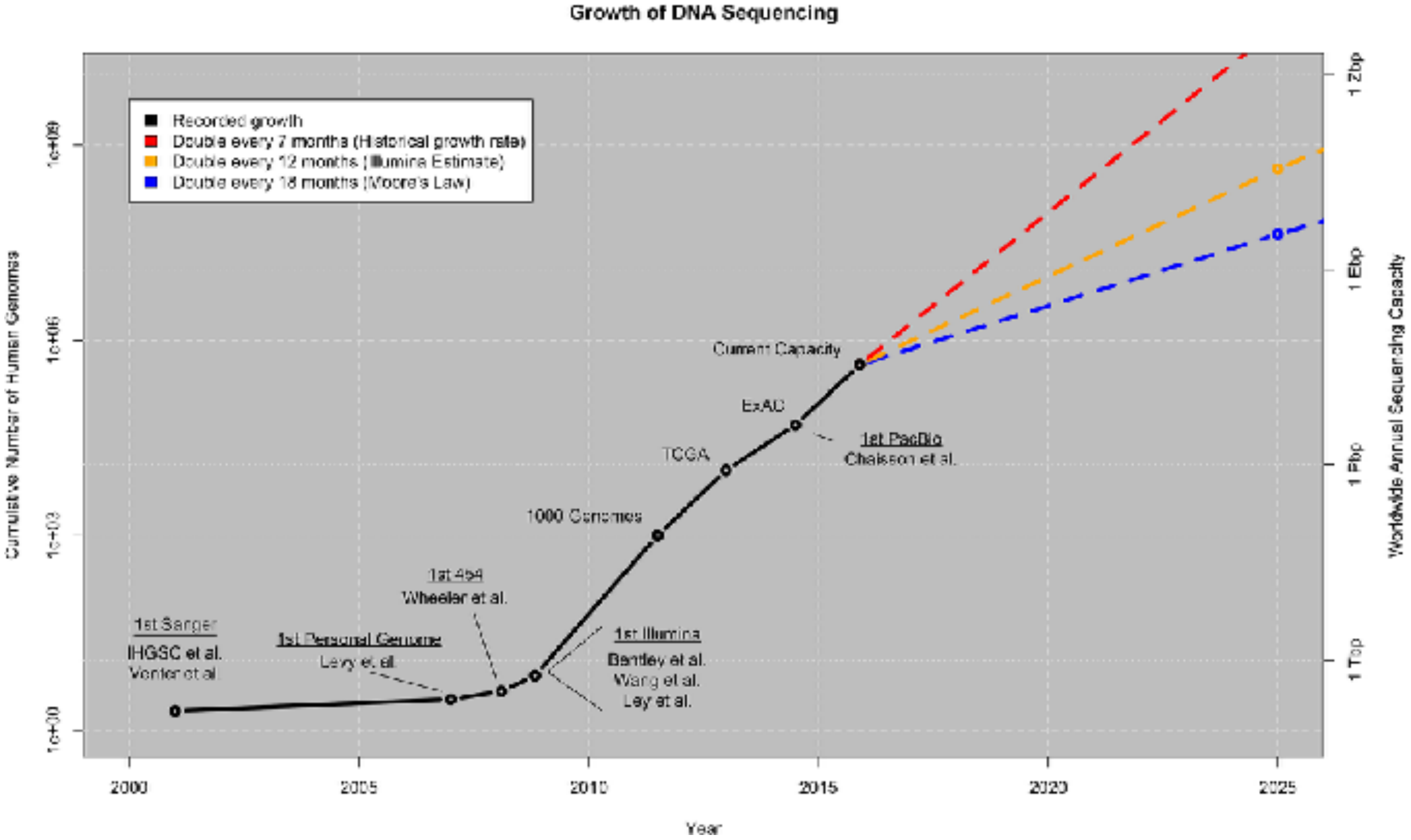
[illegible]

Rapidly Increasing Amount of Genomics Data

- Next generation Sequencing (NGS)
  - Dramatic increase in the amount of data
  - Improved data confidence
- NGS is enabler for more sophisticated research questions in Genomics



Issue: Leaps in sequencing technology have outperformed advances in computing



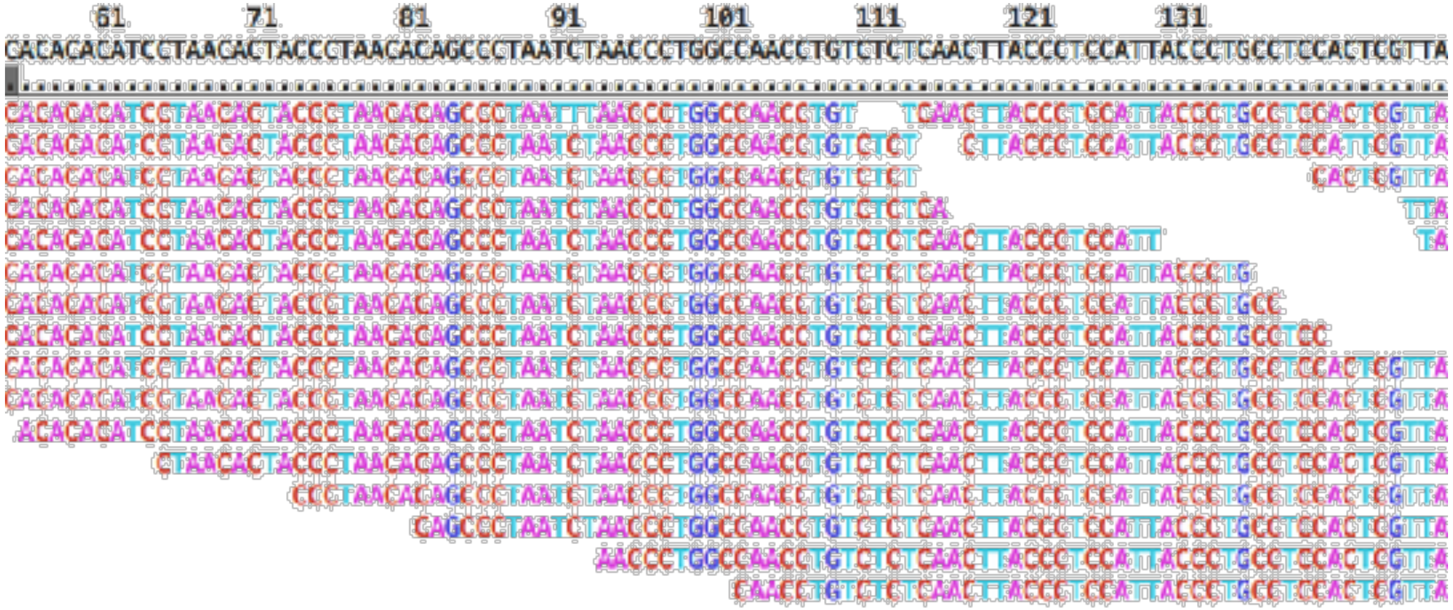
Growth of DNA sequencing data both in terms sequenced human genomes and total sequencing capacity

. Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, “Big data: Astronomical or genomical?”, Plos biology, vol. 13, no. 7, e1002195, 2015.

SAM - Sequence Alignment/Map

- Text-based format for storing biological sequences aligned to a reference one

- SAM - Plaintext
- BAM - Binary Compressed


61
71
81
91
101
111
121
131
CACACACATCCTAACACTACCCTAACACAGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTA

Challenge: data size is extremely large (about 500GB for a single human genome)

Genomic data size will overtake LHC data in the coming years

SAM Example

```
@SQ     SN:chrM LN:16571
@SQ     SN:chrX LN:155270560
SOLEXA-1GA-2_2  0   chr1  10145  25  36M  *  0  0  AACCCCTAACCCTAACCCTAACCCTA  hhhhHcWhhHTghcKA_ONhAAEEBZ
SOLEXA-1GA-2_2  0   chr1  10148  25  36M  *  0  0  CCCTAACCCTAACCCTAACCCTAACC  hbfhhhXUYhT_ULZdLRTKNIMIKG  NM:i:0
SOLEXA-1GA-2_2  16  chr1  10149  25  36M  *  0  0  CCAAACACTAACCCTAACCCTAACCC  ><>B@>?>?D>>?B?D>DBC?E@BDH  NM:i:1  X1:i:1
SOLEXA-1GA-2_2  0   chr1  10150  25  36M  *  0  0  CTAACCCTAAACCTAACCCTAACCCT  hhW_X]MXNOHQQWMILHGIFMJGJ
```

The file consists of a header section (lines starting with a @)
and an alignment section with 11 mandatory fields + several optional fields

A human genome SAM file consist of about 1 billion lines in the alignement section and is about 500GB large

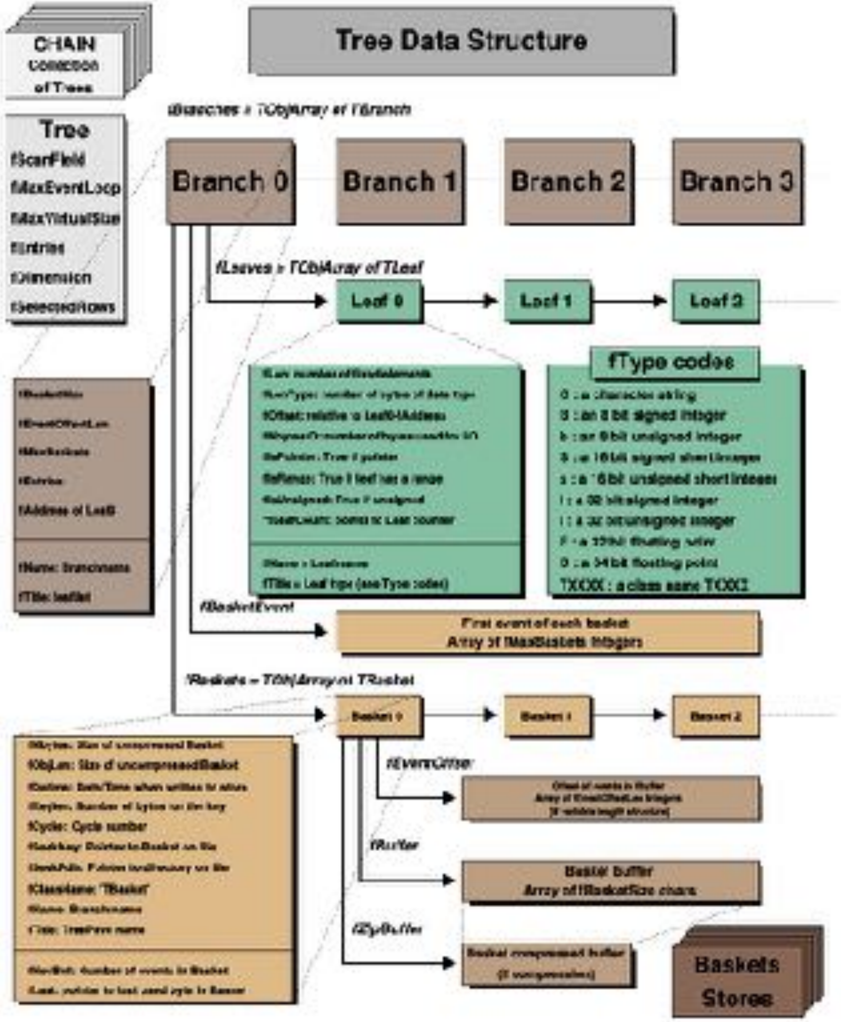
BAM - Binary Alignment/Map

- BGZF is block compression implemented on top of the standard gzip file format
- File is gunzip compatible
- Each 64KB block of data is compressed and added to the file
- Using an .bai index file, random access is supported in the BAM file

ROOT Framework

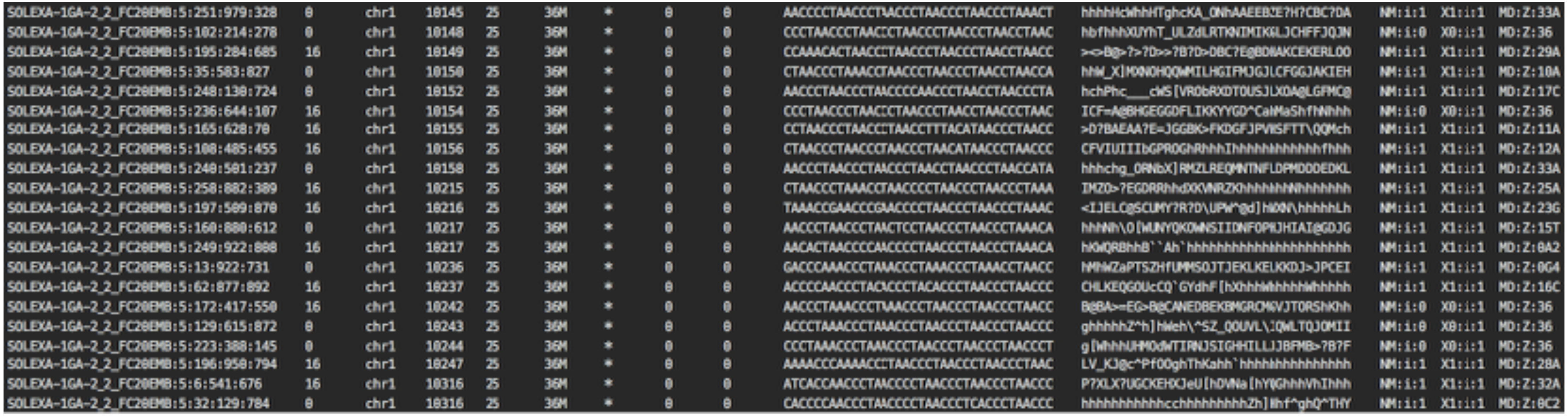
Don’t reinvent the wheel

Most of the challenges with massive data processing are common to HEP. **ROOT** has decades of experience in software design and optimisation


CHAIN
Collection of Trees
Tree Data Structure
fBranches : TObjArray of TBranch
Tree
fScanField
fMaxEventLoop
fMaxVirtualSize
fEntries
fDimension
fSelectedRows
Branch 0
Branch 1
Branch 2
Branch 3
fLeaves : TObjArray of TLeaf
Leaf 0
Leaf 1
Leaf 2
fAddress of Leaf0
fName: branchname
fTitle: leaflist
fOffset: relative to Leaf0-fAddress
fIsRange: True if leaf has a range
fName : Leafname
fTitle : Leaf type (see Type codes)
fType codes
C : a character string
B : an 8 bit signed integer
b : an 8 bit unsigned integer
S : a 16 bit signed short integer
s : a 16 bit unsigned short integer
I : a 32 bit signed integer
i : a 32 bit unsigned integer
F : a 32 bit floating point
D : a 64 bit floating point
TXXXX : a class name TXXX
fBasketEvent
First event of each basket
Array of fMaxBaskets integers
fBaskets : TObjArray of TBasket
Basket 0
Basket 1
Basket 2
fObjLen: Size of uncompressed Basket
fCycle: Cycle number
fClassName: "TBasket"
fEventOffset
fBuffer
Basket buffer
Array of fBasketSize chars
fZipBuffer
Basket compressed buffer
Baskets
Stores

Conversion - SAM to ROOT

We convert from TAB-delimited SAM format to a ROOT TTree



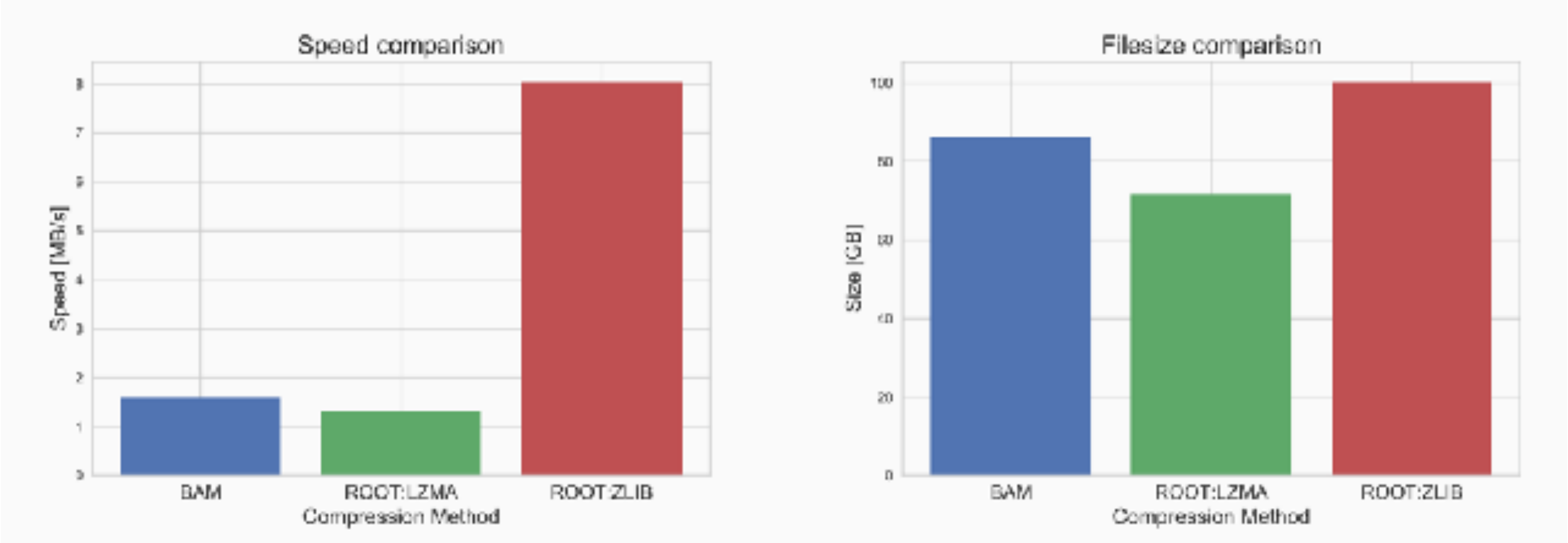
```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328	0	chr1	10145	25	36M	*	0	0	AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT	hhhhHcWhhhTghcKA_ONhAAEEBZE?H?CBC?DA	NM:i:1	X1:i:1	MD:Z:33A
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278	0	chr1	10148	25	36M	*	0	0	CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC	hbfhhhXUYhT_ULZdLRTKNIMIK6LJCHFFJQJN	NM:i:0	X0:i:1	MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685	16	chr1	10149	25	36M	*	0	0	CCAAACACTAACCCTAACCCTAACCCTAACCTAACC	><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO	NM:i:1	X1:i:1	MD:Z:29A
SOLEXA-1GA-2_2_FC20EMB:5:35:583:827	0	chr1	10150	25	36M	*	0	0	CTAACCCTAAACCTAACCCTAACCCTAACCTAACCA	hhW_X]MXNOHQQWMILHGIFMJGJLCFGGJAKIEH	NM:i:1	X1:i:1	MD:Z:10A
SOLEXA-1GA-2_2_FC20EMB:5:248:130:724	0	chr1	10152	25	36M	*	0	0	AACCCTAACCCTAACCCCAACCCTAACCTAACCCTA	hchPhc___cWS[VRObRXDTOUSJLXDA@LGFMC@	NM:i:1	X1:i:1	MD:Z:17C
SOLEXA-1GA-2_2_FC20EMB:5:236:644:107	16	chr1	10154	25	36M	*	0	0	CCCTAACCCTAACCCTAACCCTAACCTAACCCTAAC	ICF=A@BHGEGGDFLIKKYYGD^CaHMaShfhNhhh	NM:i:0	X0:i:1	MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:165:628:70	16	chr1	10155	25	36M	*	0	0	CCTAACCCTAACCCTAACCTTTACATAACCCTAACC	>D?BAEAA?E=JGGBK>FKDGFJPVWSFTT\QQMch	NM:i:1	X1:i:1	MD:Z:11A
SOLEXA-1GA-2_2_FC20EMB:5:108:485:455	16	chr1	10156	25	36M	*	0	0	CTAACCCTAACCCTAACCCTAACATAACCCTAACCC	CFVIUIIIbGPROGhRhhhIhhhhhhhhhhhhfhhh	NM:i:1	X1:i:1	MD:Z:12A
SOLEXA-1GA-2_2_FC20EMB:5:240:501:237	0	chr1	10158	25	36M	*	0	0	AACCCTAACCCTAACCCTAACCTAACCCTAACCATA	hhhchg_ORNbX]RMZLREQMNTNFLDPMDDDEDKL	NM:i:1	X1:i:1	MD:Z:33A
SOLEXA-1GA-2_2_FC20EMB:5:258:882:309	16	chr1	10215	25	36M	*	0	0	CTAACCCTAAACCTAACCCCTAACCCTAACCCTAAA	IMZD>?EGDRRhhdXKVNRZKhhhhhhNhhhhhhh	NM:i:1	X1:i:1	MD:Z:25A
SOLEXA-1GA-2_2_FC20EMB:5:197:509:870	16	chr1	10216	25	36M	*	0	0	TAAACCGAACCCGAACCCCTAACCCTAACCCTAAAC	<IJELC@SCUMY?R?D\UPW^@d]hWXN\hhhhhLh	NM:i:1	X1:i:1	MD:Z:23G
SOLEXA-1GA-2_2_FC20EMB:5:160:880:612	0	chr1	10217	25	36M	*	0	0	AACCCTAACCCTAACTCCTAACCCTAACCCTAAACA	hhhNh\O[WUNYQKOWNSIIDNFOPKJHIAI@GDJG	NM:i:1	X1:i:1	MD:Z:15T
SOLEXA-1GA-2_2_FC20EMB:5:249:922:888	16	chr1	10217	25	36M	*	0	0	AACACTAACCCCAACCCCTAACCCTAACCCTAAACA	hKWQRBhhB``Ah`hhhhhhhhhhhhhhhhhhhhhh	NM:i:1	X1:i:1	MD:Z:0A2
SOLEXA-1GA-2_2_FC20EMB:5:13:922:731	0	chr1	10236	25	36M	*	0	0	GACCCAAACCCTAAACCCTAAACCCTAAACCTAACC	hMhWZaPTSZHfUMMSOJTJEKLKELKKDJ>JPCEI	NM:i:1	X1:i:1	MD:Z:0G4
SOLEXA-1GA-2_2_FC20EMB:5:62:877:892	16	chr1	10237	25	36M	*	0	0	ACCCCAACCCTACACCCTACACCCTAACCCTAACCC	CHLKEQGOUcCQ`GYdhF[hXhhhWhhhhhWhhhhh	NM:i:1	X1:i:1	MD:Z:16C
SOLEXA-1GA-2_2_FC20EMB:5:172:417:550	16	chr1	10242	25	36M	*	0	0	AACCCTAAACCCTAACCCTAACCCTAACCCTAACC	B@BA>=EG>B@CANEDBEKBMGROM6VJTORShKhh	NM:i:0	X0:i:1	MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:129:615:872	0	chr1	10243	25	36M	*	0	0	ACCCTAAACCCTAAACCCTAACCCTAACCCTAACCC	ghhhhhZ^h]hWeh\^SZ_QOUVL\IQWLTQJOMII	NM:i:0	X0:i:1	MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:223:388:145	0	chr1	10244	25	36M	*	0	0	CCCTAAACCCTAAACCCTAACCCTAACCCTAACCCT	g[WhhhUHMOdWTIRNJSIGHHILLJJBFMB>?B?F	NM:i:0	X0:i:1	MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:196:958:794	16	chr1	10247	25	36M	*	0	0	AAAACCCAAAACCCTAACCCTAACCCTAACCCTAAC	LV_KJ@c^PfOOghThKahh`hhhhhhhhhhhhhhh	NM:i:1	X1:i:1	MD:Z:28A
SOLEXA-1GA-2_2_FC20EMB:5:6:541:676	16	chr1	10316	25	36M	*	0	0	ATCACCAACCCTAACCCCTAACCCTAACCCTAACCC	P?XLX?UGCKEHXJeU[hDWNa[hY@GhhhVhIhhh	NM:i:1	X1:i:1	MD:Z:32A
SOLEXA-1GA-2_2_FC20EMB:5:32:129:784	0	chr1	10316	25	36M	*	0	0	CACCCCAACCCTAACCCCTAACCCTCACCCTAACCC	hhhhhhhhhhhcchhhhhhhhhZh]Whf^ghQ^THY	NM:i:1	X1:i:1	MD:Z:0C2
```

ROOT handles compression, memory buffers, datatypes, endian-ness, etc.

The columns are defined using a C++ class.

Performance - SAM to ROOT

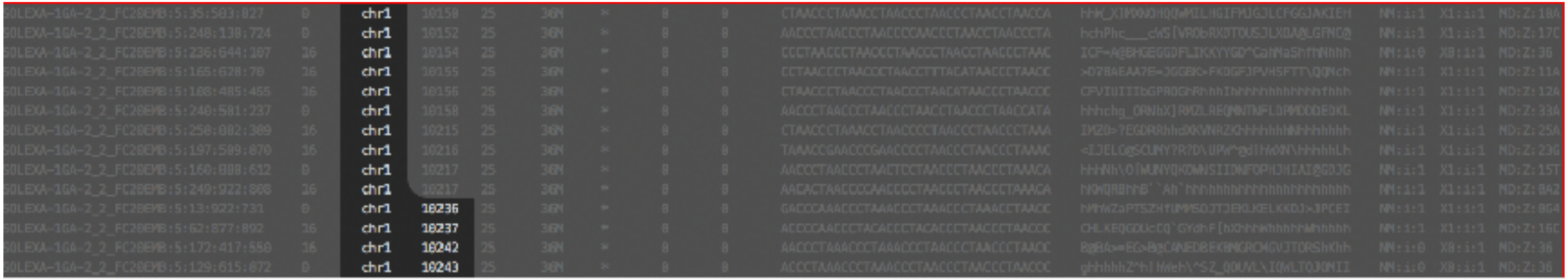
There is a tradeoff between compression and read/write speed for 100GB file.



With ZLIB compression -> 4 times faster
With LZMA compression conversion -> 15% smaller

View - Random Access

- We also need to be able to view the information as fast as possible
- ROOT columnar structure allow us to just look at the chromosome and position columns to optimize performance



```
SOLEXA-1GA-2_2_FC20EMB:5:35:503:027   0   chr1  10150  25  36M  *  0  0  CTAACCCTAAACCTAACCCTAACCCTAACCTAACCA  HHH_XIMXNDHQQWMILHGIFMJGJLCFGGJAKIEH  NM:i:1  X1:i:1  MD:Z:18A
SOLEXA-1GA-2_2_FC20EMB:5:248:138:724  0   chr1  10152  25  36M  *  0  0  AACCCTAACCCTAACCCCAACCCTAACCTAACCCTA  hchPhc___cWS[VRObRXDTOUSJLXDA@LGFND@  NM:i:1  X1:i:1  MD:Z:17C
SOLEXA-1GA-2_2_FC20EMB:5:236:644:107  16  chr1  10154  25  36M  =  0  0  CCCTAACCCTAACCCTAACCCTAACCTAACCCTAAC  ICF=A@BHGEGGDFLIKKYYGD^CahMaShfhNhhh  NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:165:628:70   16  chr1  10155  25  36M  =  0  0  CCTAACCCTAACCCTAACCTTTACATAACCCTAACC  >D7BAEAA7E=JGGBK>FKDGFJPVH5FTT\QQMch  NM:i:1  X1:i:1  MD:Z:11A
SOLEXA-1GA-2_2_FC20EMB:5:188:485:455  16  chr1  10156  25  36M  =  0  0  CTAACCCTAACCCTAACCCTAACATAACCCTAACCC  CFVIUIIIbGPROGhRhhhIhhhhhhhhhhhhfhhh  NM:i:1  X1:i:1  MD:Z:12A
SOLEXA-1GA-2_2_FC20EMB:5:240:581:237  0   chr1  10158  25  36M  *  0  0  AACCCTAACCCTAACCCTAACCTAACCCTAACCATA  hhhchg_ORNbX]RMZLAEQMNTNFLDRMDDDEDKL  NM:i:1  X1:i:1  MD:Z:33A
SOLEXA-1GA-2_2_FC20EMB:5:258:082:389  16  chr1  10215  25  36M  *  0  0  CTAACCCTAAACCTAACCCCTAACCCTAACCCTAAA  IMZ0>?EGDRRhhdXKVNRZKhhhhhhhNhhhhhhh  NM:i:1  X1:i:1  MD:Z:25A
SOLEXA-1GA-2_2_FC20EMB:5:197:589:870  16  chr1  10216  25  36M  *  0  0  TAAACCGAACCCGAACCCCTAACCCTAACCCTAAAC  <IJELC@SCUNY?R?D\UPW^@d|HWXN\hhhhhLh  NM:i:1  X1:i:1  MD:Z:23G
SOLEXA-1GA-2_2_FC20EMB:5:160:888:612  0   chr1  10217  25  36M  *  0  0  AACCCTAACCCTAACTCCTAACCCTAACCCTAAACA  HhHNh\O|WUNYQKOWNSIIDNFOPHJHIAI@GDJG  NM:i:1  X1:i:1  MD:Z:15T
SOLEXA-1GA-2_2_FC20EMB:5:249:922:888  16  chr1  10217  25  36M  =  0  0  AACACTAACCCCAACCCCTAACCCTAACCCTAAACA  hKWQRBhhB``Ah`hhhhhhhhhhhhhhhhhhhhhh  NM:i:1  X1:i:1  MD:Z:8A2
SOLEXA-1GA-2_2_FC20EMB:5:13:922:731   0   chr1  10236  25  36M  =  0  0  GACCCAAACCCTAAACCCTAAACCCTAAACCTAACC  hMhWZaPTSZHfUMMSDJTJEKLKELKKDJ>JPCEI  NM:i:1  X1:i:1  MD:Z:0G4
SOLEXA-1GA-2_2_FC20EMB:5:62:877:892   16  chr1  10237  25  36M  =  0  0  ACCCCAACCCTACACCCTACACCCTAACCCTAACCC  CHLKEQGDUcCQ`GYdhF[hXhhhWhhhhhWhhhhh  NM:i:1  X1:i:1  MD:Z:16C
SOLEXA-1GA-2_2_FC20EMB:5:172:417:550  16  chr1  10242  25  36M  *  0  0  AACCCTAAACCCTAAACCCTAACCCTAACCCTAACC  B@BA==EG>B@CANEDBEKBNGRCHGVJTORShKhh  NM:i:0  X0:i:1  MD:Z:36
SOLEXA-1GA-2_2_FC20EMB:5:129:615:872  0   chr1  10243  25  36M  *  0  0  ACCCTAAACCCTAAACCCTAACCCTAACCCTAACCC  ghhhhhZ^h|HWeh\^SZ_QOUVL\IQWLTOJONII  NM:i:0  X0:i:1  MD:Z:36
```

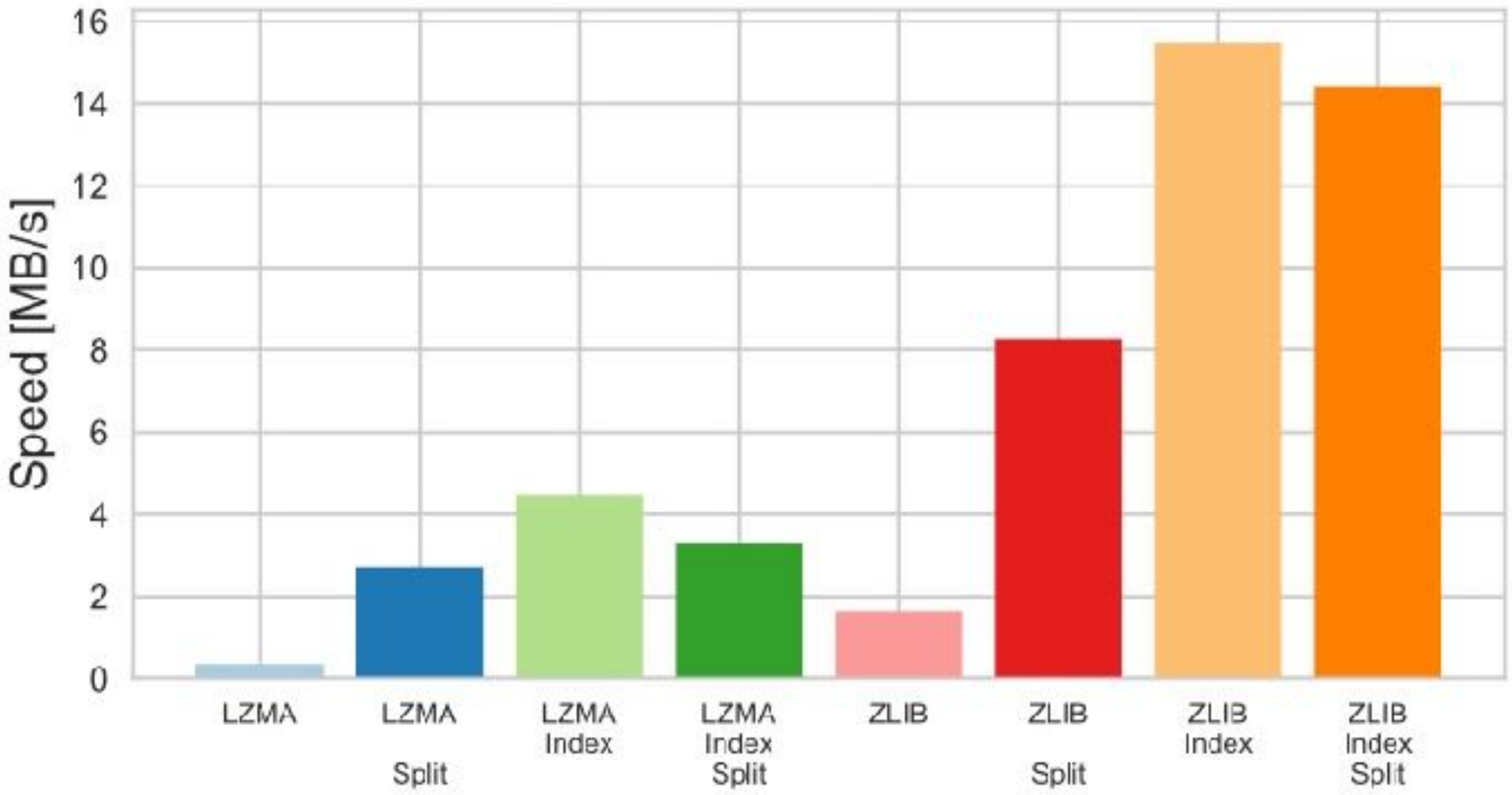
Indexing For Fast Access

- Improve read speed by using an index
  - For random access BAM format needs a BAI index file
  - For ROOT implemented a RAMIndex, sparse index in combination with fast columnar ROOT file scanning
  - 16 bytes per entry, compressed

```
class RAMIndex {
private:
   typedef std::pair<int,int> Key_t;            // refid (of rname) and pos
   typedef std::map<Key_t,Long64_t> Index_t;   // map of Key_t and TTree entry number

   Index_t fIndex;
   …
};
```

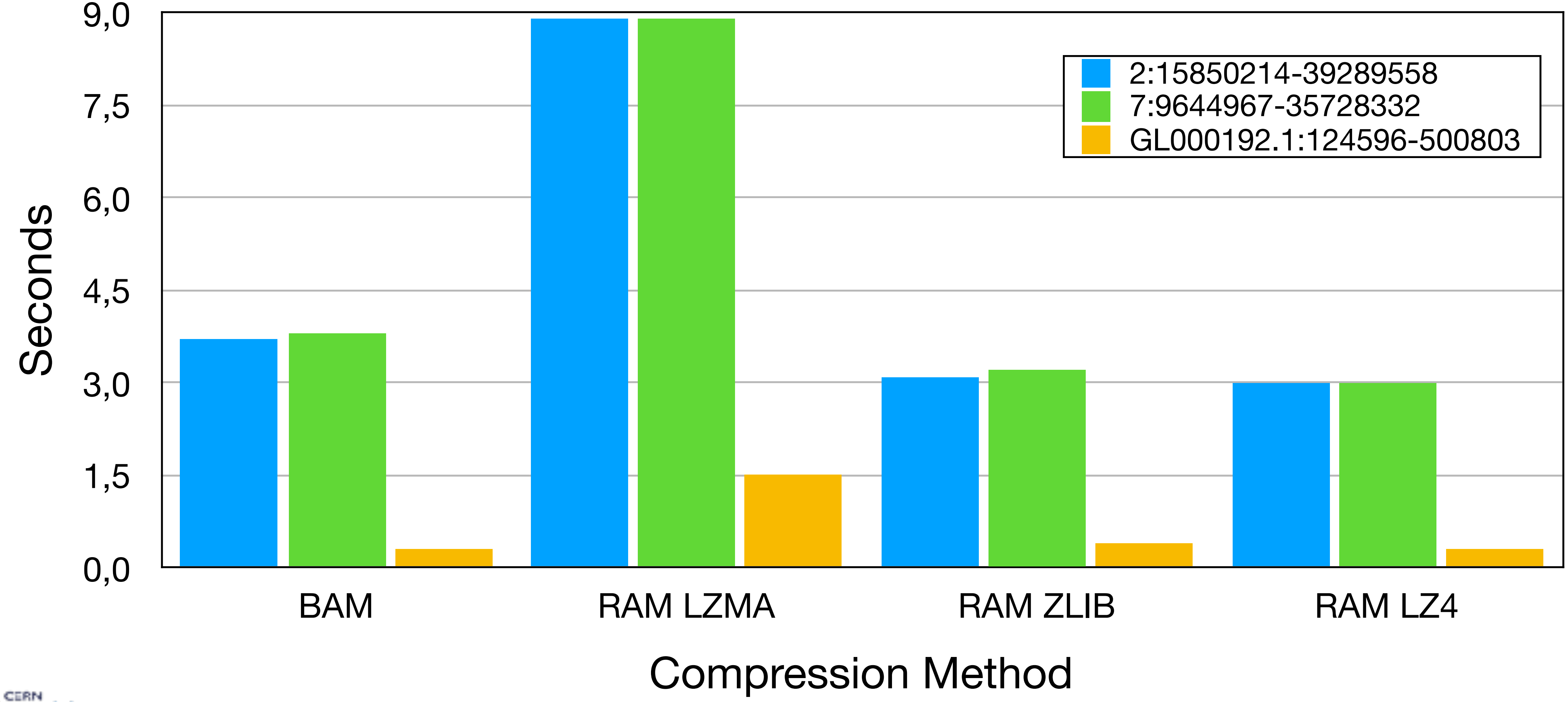
View - Performance



Median read speed for ramtools across different parameter settings

View - Performance

Viewing 3 different regions from a 9-18 GB file



Future Work

- Increase file sample - The study was done on only one base SAM file (subsampled to 10% and 1%). Conversion, indexing and views should be run on different SAM files so sample bias is reduced
- Extended format comparison - The current work compared only to the well established BAM format. However, comparisons to formats such as CRAM should be made to see the relative performances in compression rate and speed and read speed, respectively
- Support for additional operations - Add to ramtools support for sort, merge, split or stats.
- FASTQ/FASTA conversion - The raw sequencer data format. Simpler but closely resembling SAM. Most of the TTree advantages apply also to this type of data. In fact, recent research in formats like LFQC has many analogies to how ROOT branches are used to compress data fields separately, optimizing compression and read speed.

CERN openlab GeneROOT Technology Transfer Benefits

- Additional use case for CERN’s ROOT technology
- Return flow of know-how benefiting the ROOT User community
- CERN openlab provides a doctoral student who gains valuable experience
- Entire Omics community would benefit from improved analysis tools to handle rapidly growing amounts of data
- Project is joint effort between CERN openlab, CERN medical applications group, King’s College London

- Big thank you to my 2017 openlab summer student
Jose Javier Gonzalez Ortiz