



# The GeneROOT Project

Fons Rademakers

CERN openlab Chief Research Officer

# GeneROOT - Using ROOT for Handling Genomics Data

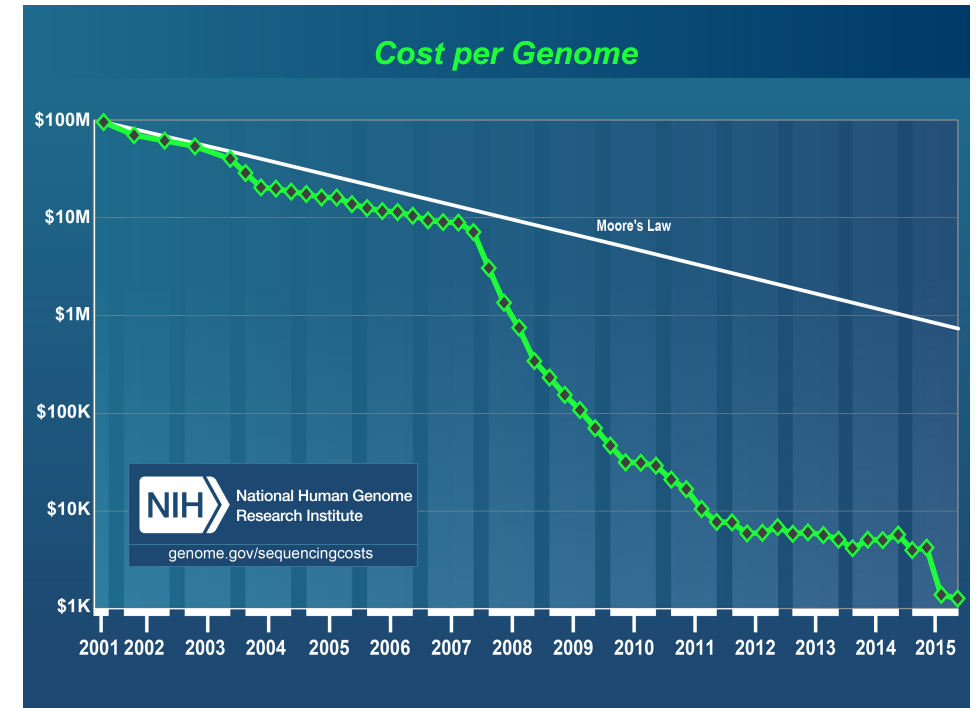
# King College London - TwinsUK Project

- 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

The screenshot shows the homepage of the TwinsUK website. At the top, there is a navigation menu with links for 'ABOUT US', 'TWIN ZONE', 'RESEARCH AREAS', 'PROJECTS', 'DATA ACCESS', 'MEDIA AND ENGAGEMENT', 'SCIENTIFIC PUBLICATIONS', 'TOOLS & INFORMATION', and 'LEARNING & TRAINING'. Below the navigation is a search bar and a 'TwinsUK' logo. The main content area features a purple header with the text: '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.' Below this is a 'Latest News' section with a featured article titled 'World Alzheimer's Day 2017: Q&A with Dr Claire Steves'. To the right, there is a 'SEEKING TWINS' section with the heading 'ARE YOU A TWIN OR DO YOU KNOW A TWIN?' and a list of photos of twins. At the bottom, there are four colored boxes: 'TWIN ZONE' (orange), 'MEDIA & COMMUNICATIONS' (green), 'SCIENTIFIC PUBLICATIONS' (red), and 'DATA ACCESS' (blue). The footer includes logos for 'MAJOR RESEARCH SPONSORS' such as Wellcome Trust, European Commission, MRC, and CDRE, along with social media icons and a disclaimer.

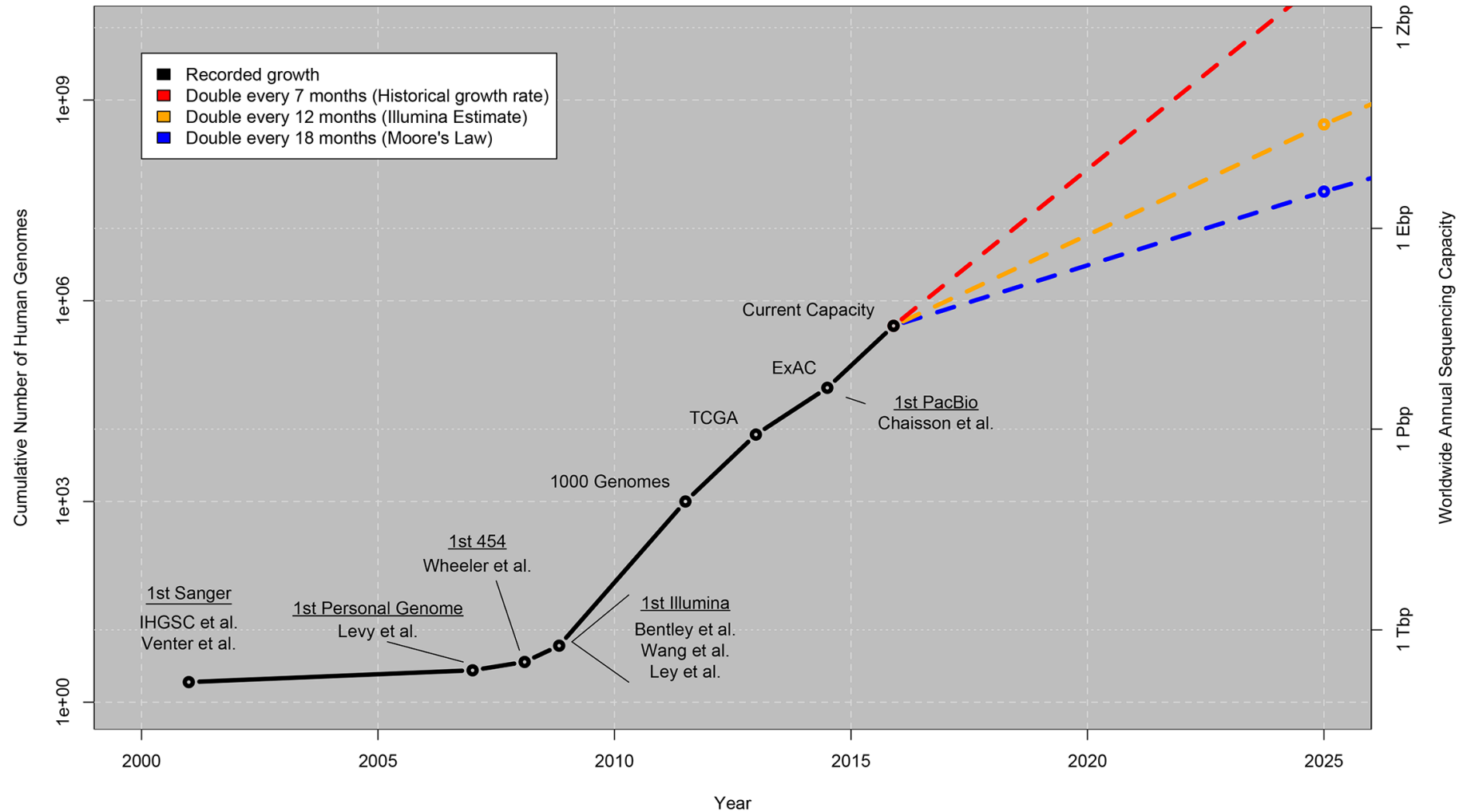
# Rapidly Increasing Amount of Genomics Data

- 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**

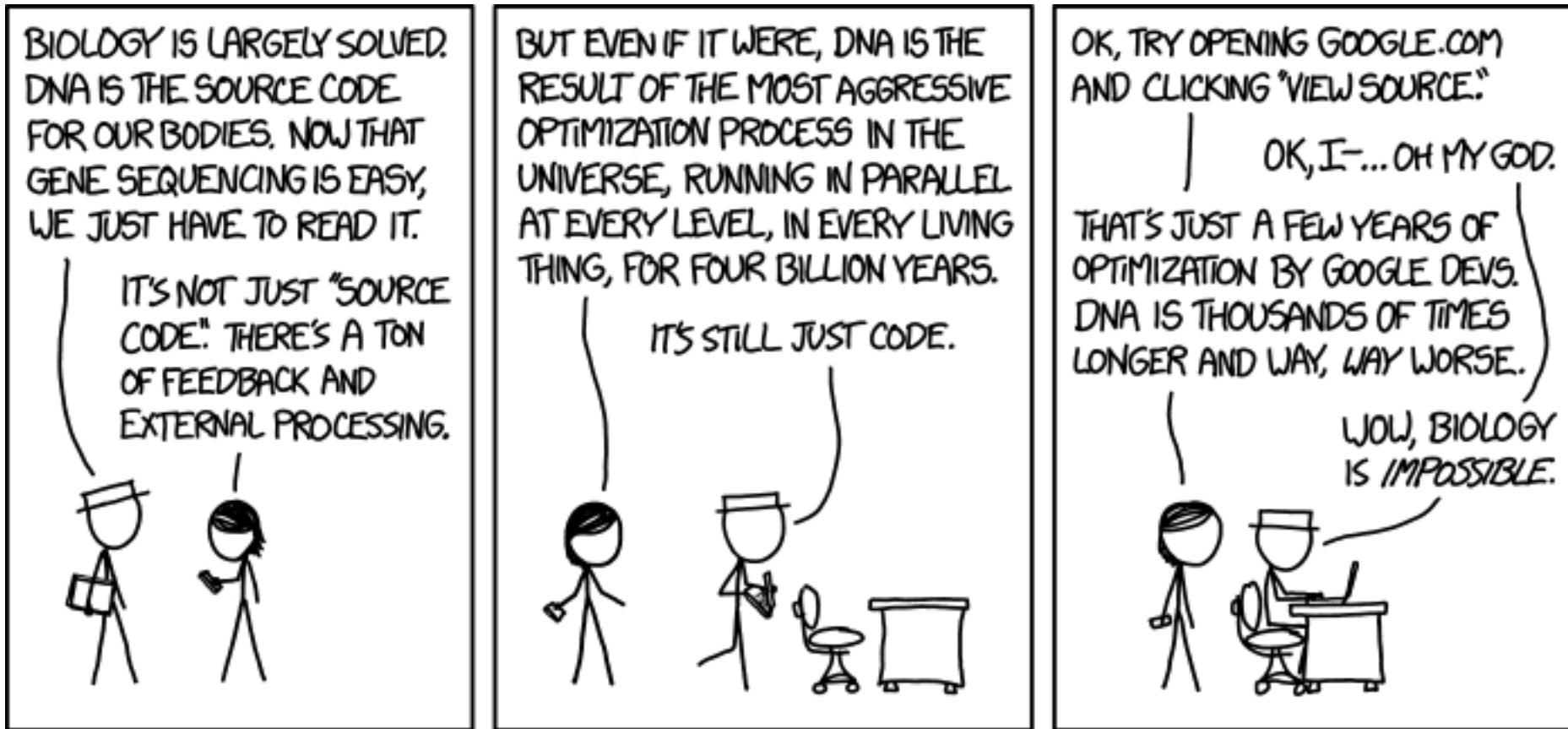
## Growth of DNA Sequencing



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

# Genomics Data

Understanding genomics data is not trivial at all.





# SAM Example

```
@SQ      SN:chrM LN:16571
@SQ      SN:chrX LN:155270560
SOLEXA-1GA-2_2  0   chr1  10145  25  36M  *  0  0  AACCCCTAACCCCTAACCCCTAACCCCTA  hhhhHcWhhHTghcKA_ONhAAEEBZ
SOLEXA-1GA-2_2  0   chr1  10148  25  36M  *  0  0  CCCTAACCCCTAACCCCTAACCCCTAACC  hbfhhhXUYhT_ULZdLRTKNIMIKG  NM:i:0
SOLEXA-1GA-2_2  16  chr1  10149  25  36M  *  0  0  CCAAACACTAACCCCTAACCCCTAACCC  ><>B@>?>?D>>?B?D>DBC?E@BDH  NM:i:1  X1:i:1
SOLEXA-1GA-2_2  0   chr1  10150  25  36M  *  0  0  CTAACCCTAACCCCTAACCCCTAACCCCT  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 alignment section and is about 500GB large

# SAM Data Structure

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1-based leftmost mapping POSition
5	MAPQ	Int	MAPing Quality
6	CIGAR	String	CIGAR string
7	RNEXT	String	Ref. name of the mate/NEXT read
8	PNEXT	Int	Position of the mate/NEXT read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

# BAM - Binary Alignment/Map

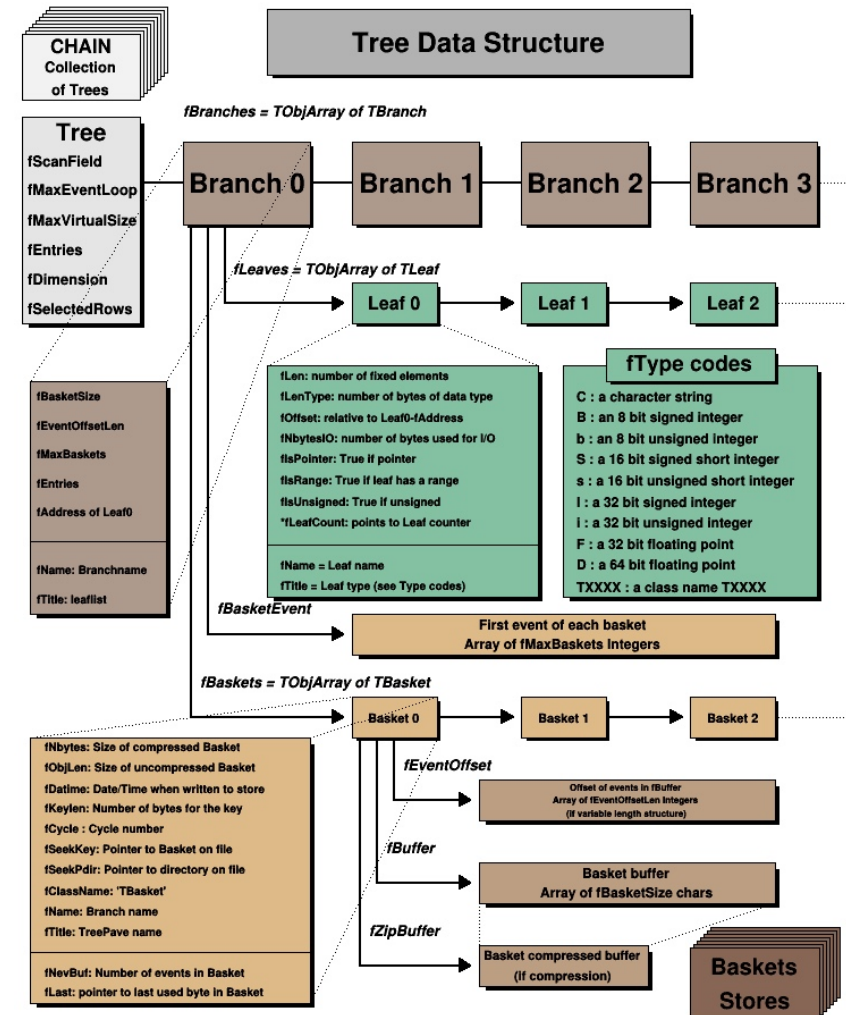
- BGZF is block compression implemented on top of the standard gzip file format
- File is gunzip compatible
- Each 64KB block of data is compressed and added to the file
- Using an .bai index file, random access is supported in the BAM file

# 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



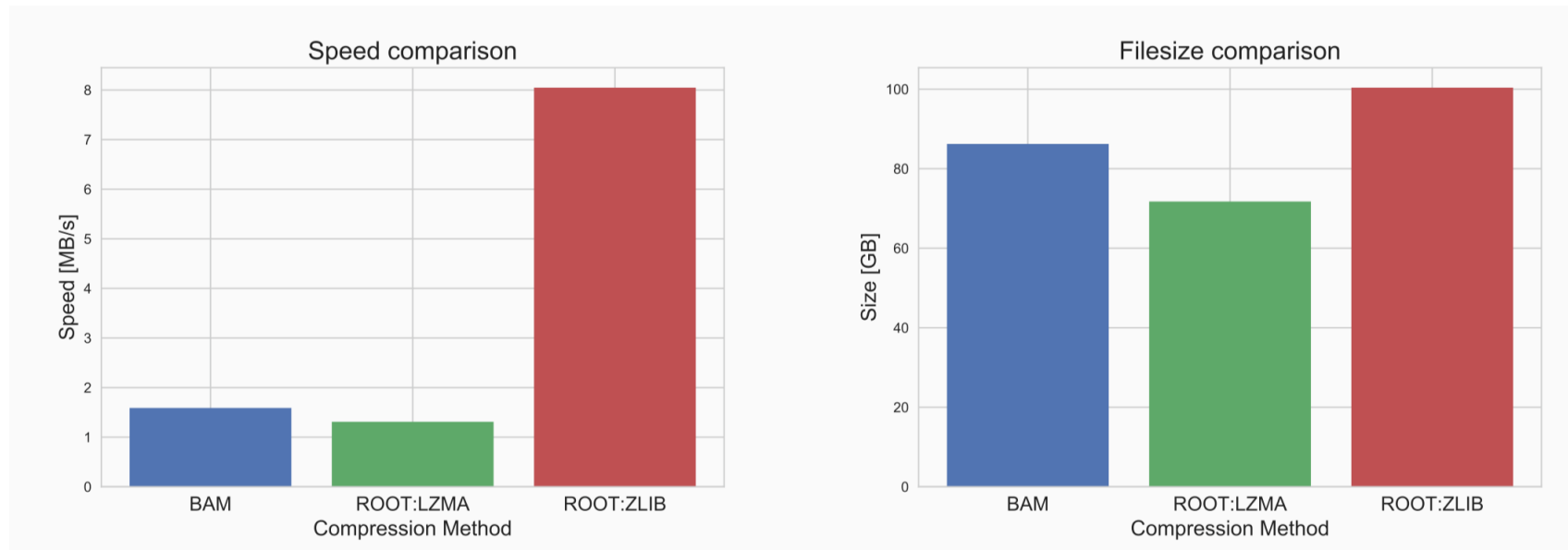


# RAMRecord Class

```
class RAMRecord : public TObject {
public:
    enum EQualCompressionBits {
        kPhred33          = BIT(14),    // Default Phred+33 quality score
        kIlluminaBinning  = BIT(15),    // Illumina 8 bin compression
        kDrop              = BIT(16),    // Drop quality score
    };
private:
    TString          v_qname;           // Query template NAME
    UShort_t        v_flag;            // Bitwise FLAG
    Int_t           v_refid;           // Reference sequence ID (maps to reference seq name)
    Int_t           v_pos;             // 0-based left most mapping POSition
    UChar_t         v_mapq;            // MAPing Quality
    Int_t           v_ncigar_op;       // Number of CIGAR operands
    UInt_t          *v_cigar;          //[v_ncigar_op] (op_len<<4|op. "MIDNSHP=X" -> "012345678")
    Int_t           v_refnext;         // Reference ID of the mate/next read
    Int_t           v_pnext;           // 0-based position of the mate/next read
    Int_t           v_tlen;            // Observed Template LENgth
    Int_t           v_lseq;            // Length of segment SEQUENCE
    Int_t           v_lseq2;           // (Length+1)/2 of segment SEQUENCE
    UChar_t         *v_seq;            //[v_lseq2] segment SEQUENCE
    UChar_t         *v_qual;           //[v_lseq] ASCII of Phred-scaled base QUALity+33
    Int_t           v_nopt;            // Number of optional fields
    TString         *v_opt;            //[v_nopt] Optional fields
};
```

# 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:5:35:583:827	0	chr1	10150	25	36M	*	0	0	CTAACCCCTAACCCCTAACCCCTAACCCCTAACCA	hhw_X]MxNQHQQWMLHGIFMJGJLCFGGJAKIEH	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	AACCCTAACCCCTAACCCCTAACCCCTAACCCCTA	hchPhc___cWS [VR0bRXDIOUSJLX0A@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	CCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC	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	CCTAACCCCTAACCCCTAACCTTACATAACCCTAAC	>D7BAEAA?E=JGGBK>FKDGFJPVHSFTT\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	CTAACCCCTAACCCCTAACCCCTAACCCCTAACCC	CFVIUIIibGPR0GhRhhhIhhhhhhhhhhhhfhhh	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	AACCCTAACCCCTAACCCCTAACCCCTAACCCATA	hhhchg_ORNbX]RMZLREQMNTNFLDPMDDDEDKL	NM:i:1	X1:i:1	MD:Z:33A
SOLEXA-1GA-2_2_FC20EMB:5:258:882:389	16	chr1	10215	25	36M	*	0	0	CTAACCCCTAACCCCTAACCCCTAACCCCTAAAA	IMZ0>?EGDRRhdXKVNRZKhhhhhhhNhhhhhh	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	TAAACCGAACCCGAACCCCTAACCCCTAACCCTAAC	<IJELC@SCUMY?R?D\UPW^@]hW0N\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	AACCCTAACCCCTAACCCCTAACCCCTAACCCATA	hhhNh\0 [WUNYQKOWNSIIDNFOPHJHIAI@gDJG	NM:i:1	X1:i:1	MD:Z:15T
SOLEXA-1GA-2_2_FC20EMB:5:249:922:808	16	chr1	10217	25	36M	*	0	0	AACACTAACCCCAACCCCTAACCCCTAACCCATA	hKwQRBhb`^`Ah`hhhhhhhhhhhhhhhhhhhh	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	GACCCAAACCCCTAACCCCTAACCCCTAACCCTAAC	hMhwZaPTSZHfUMMSOJTJEKLEKLDJ>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	ACCCCAACCCCTAACCCCTAACCCCTAACCCTAACCC	CHLKEQ0UcCQ`GYdhF[hXhhhhhhhhhhhhhh	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	AACCCTAACCCCTAACCCCTAACCCCTAACCCTAAC	B@BA>=EG>B@CANEDBEKBMGRMGVJTORShKhh	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	ACCCCTAACCCCTAACCCCTAACCCCTAACCCTAACCC	ghhhhhZ^h]hWeh^*SZ_Q0UJVL\IQWLTQJOMII	NM:i:0	X0:i:1	MD:Z:36

# Indexing For Fast Access

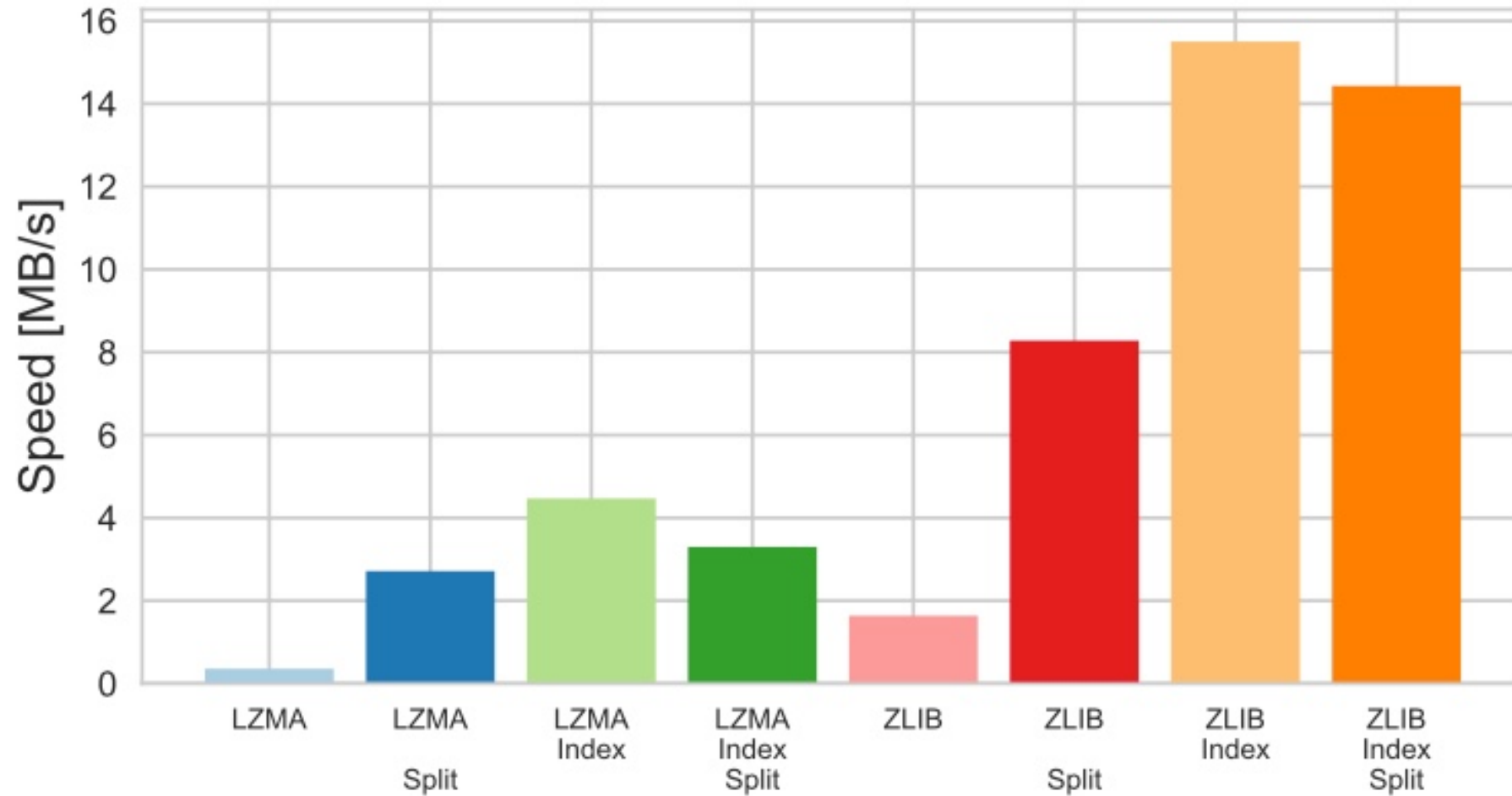
- Improve read speed by using an index
  - For random access BAM format needs a BAI index file
  - For ROOT implemented a RAMIndex, sparse index in combination with fast columnar ROOT file scanning
  - 16 bytes per entry, compressed

```
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 - Understanding ROOT I/O Parameters

- ROOT offer many parameters we need to check and profile:
  - Compression Algorithm: ZLIB vs LZMA vs LZ4
  - Split: Columns vs Row
  - Cache: Enabled vs Disabled
  - Storage: Local vs Remote
- For 20 views, the number of tests per file gives a Combinatorial Explosion
  - $3 \times 2 \times 2 \times 2 \times 20 = 480$  views per file
- Would be good to have a machine learning module in ROOT to optimize these parameters depending on different usage patterns

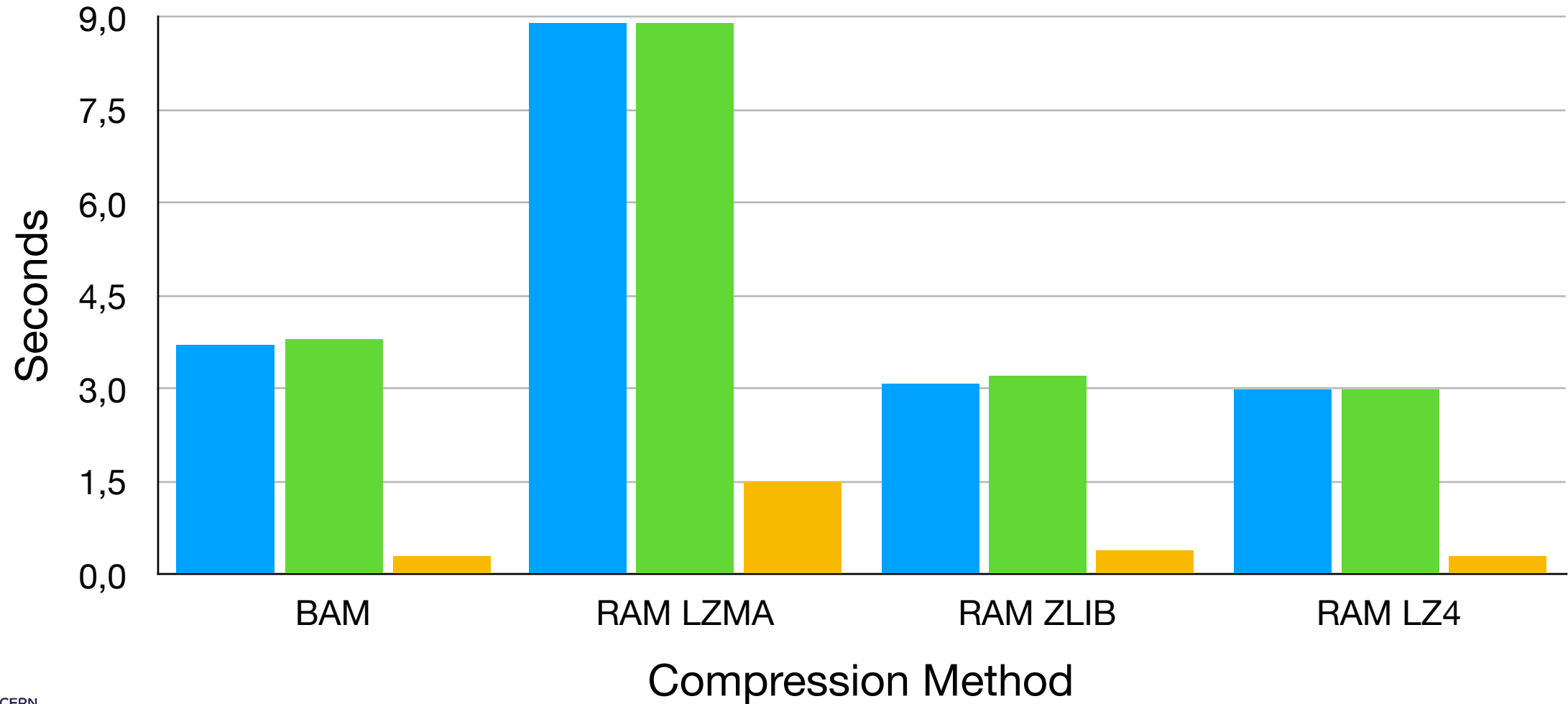
# View - Performance



Median read speed for ramtools across different parameter settings

# View - Performance

Viewing 3 different regions from a 9-18 GB file



# Further Work

- Increase file sample - The study was done on only one base SAM file (subsamped 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