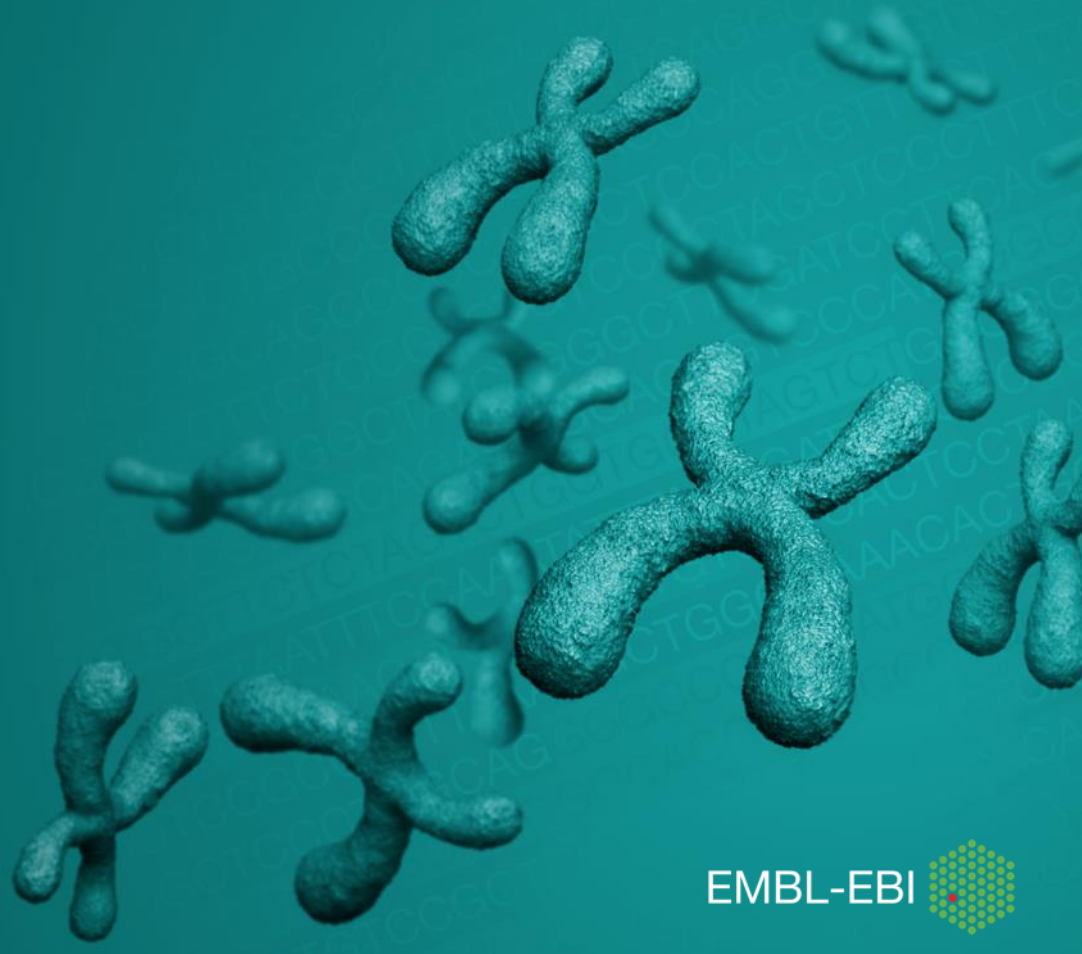
Sequence data compression in the wild

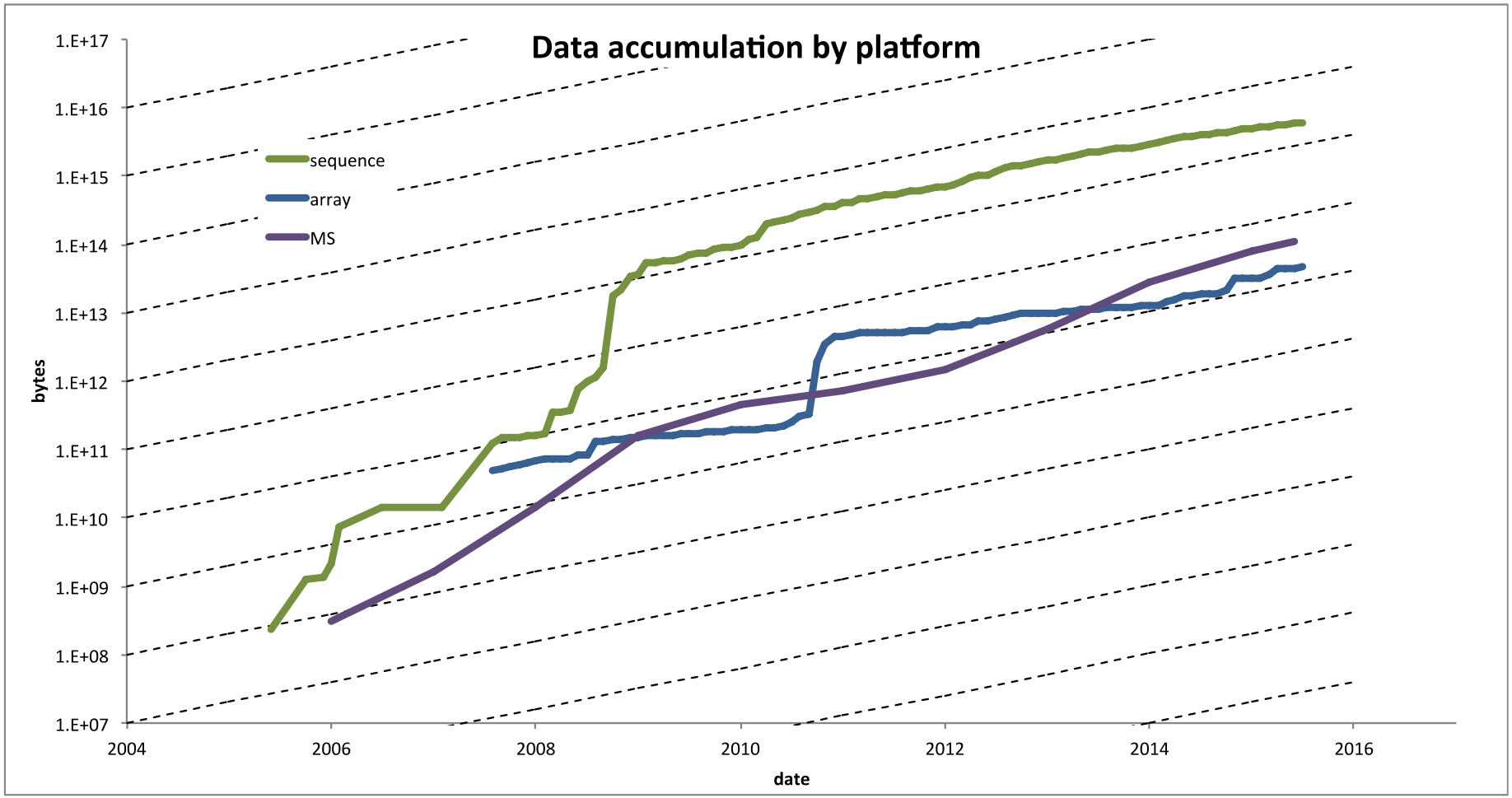
Guy Cochrane

www.ebi.ac.uk

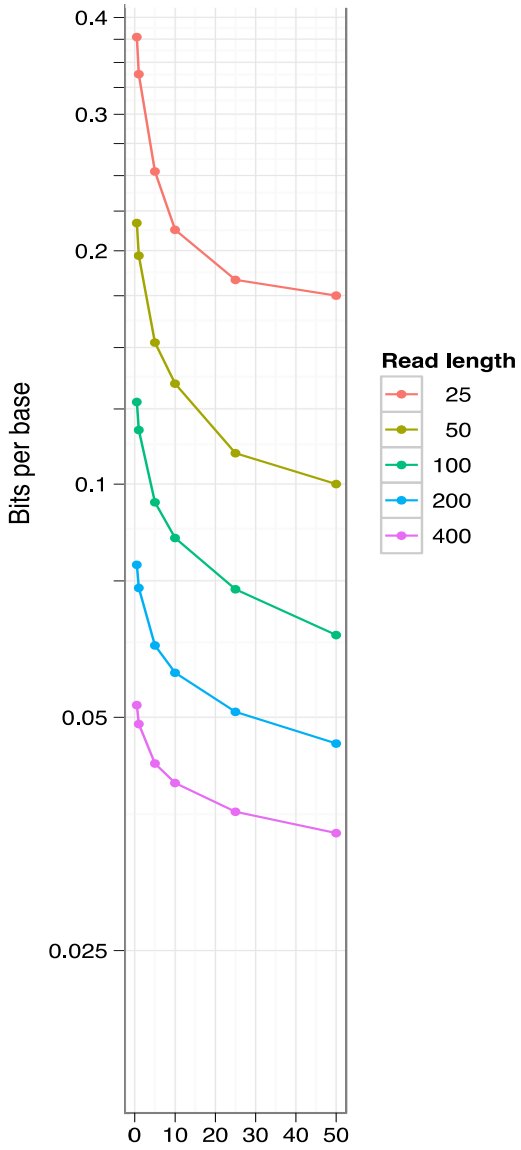


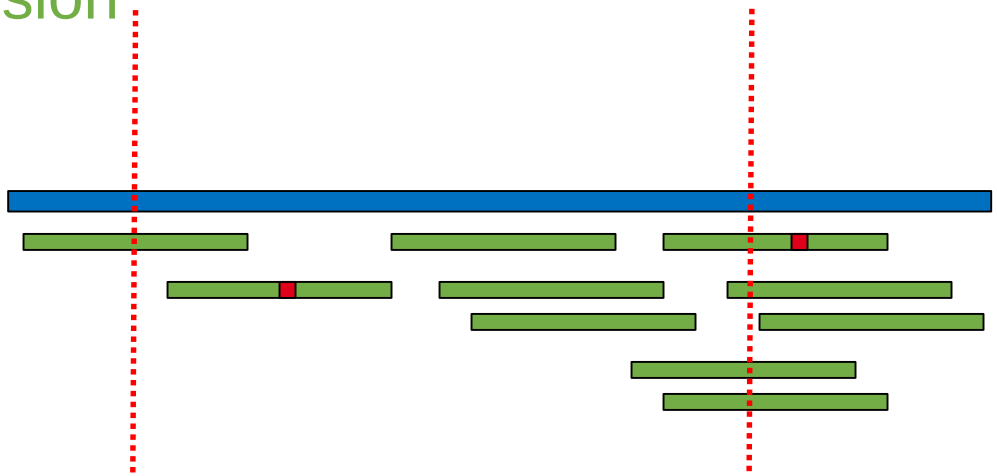
EMBL-EBI

Data growth


Data accumulation by platform
sequence
array
MS
bytes
1.E+17
1.E+16
1.E+15
1.E+14
1.E+13
1.E+12
1.E+11
1.E+10
1.E+09
1.E+08
1.E+07
2004
2006
2008
2010
2012
2014
2016
date

Reference-based compression

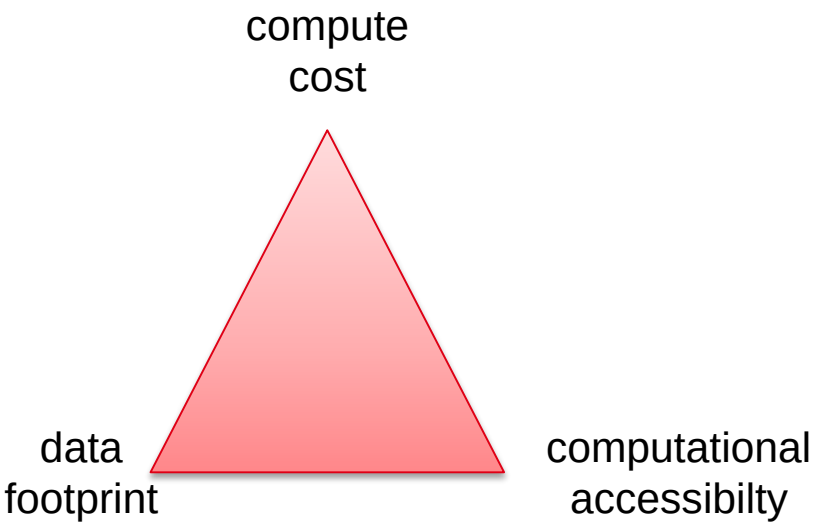




- Encoding of read starts and differences
- 3.5x–100x compression over existing formats
- Scales favourably with increasing read length and density

His-Yang Fritz, M.H., Leinonen, R., Cochrane, G., and Birney, E. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734-40.

Needs



Present

Data reduction mark I

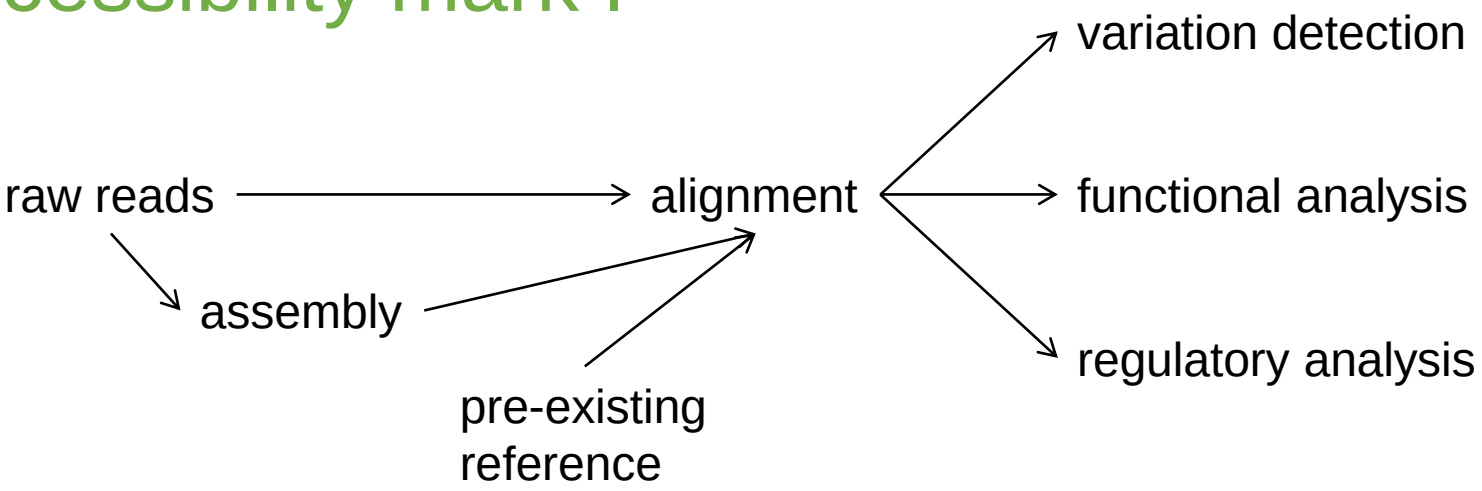
- ‘Horizontal’ model

```
atgctagatgcatgactagcatcgactaactatacggacatacgactacga
93847361297346139237427362354923746927346923397647 8
```

- E.G. quantize from 10- to 3-point scale

```
atgctagatgcatgactagcatcgactaactatacggacatacgactacga
31323121133122113113213121122311322313122311133223 3
```

Accessibility mark I



- Principles
  - Favour a biologically-relevant reference and index consistently with (broadly understood) coordinate system for reference
  - Allow in-file computational access with minimal decompression
  - Harmonise indexing scheme with that of other tools (e.g. SAM/BAM) and integrate into tool chain
  - Provide acceptable performance at the cost of maximum compression
- Layered services
  - Sequence similarity search
  - look up by annotated gene or other feature

Reference-based compression in use

- CRAM
  - WTSI (CRAM as production format)
  - EMBL-EBI (15% sequenced bases, 9% of bytes now submitted and presented in CRAM)
  - 1,000 Genomes
  - Global Alliance for Genomes and Health (GA4GH)
  - Others
- cSRA

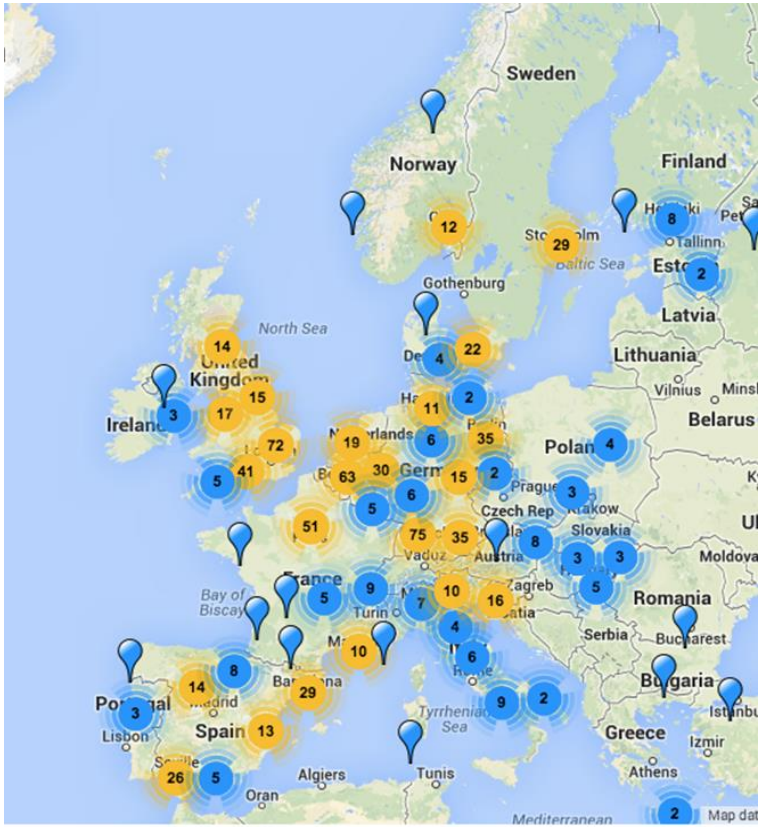
Challenges

Technological advance

- Aggressive growth within platforms
  - E.g. Illumina read lengths, coverage, cost, library methods (e.g. moleculo)
- Step changes as new platforms emerge
  - E.g. 454 vs. Sanger, Illumina vs. 454 and Sanger, ONT

Stakeholder community

- Diverse and dispersed
- Organic rather than top-down in its organisation
- Standards and practices ‘become’, as much as are driven
- Rapid advance of technology leads to rapidly evolving informatics
- Trust is key



What are raw data?

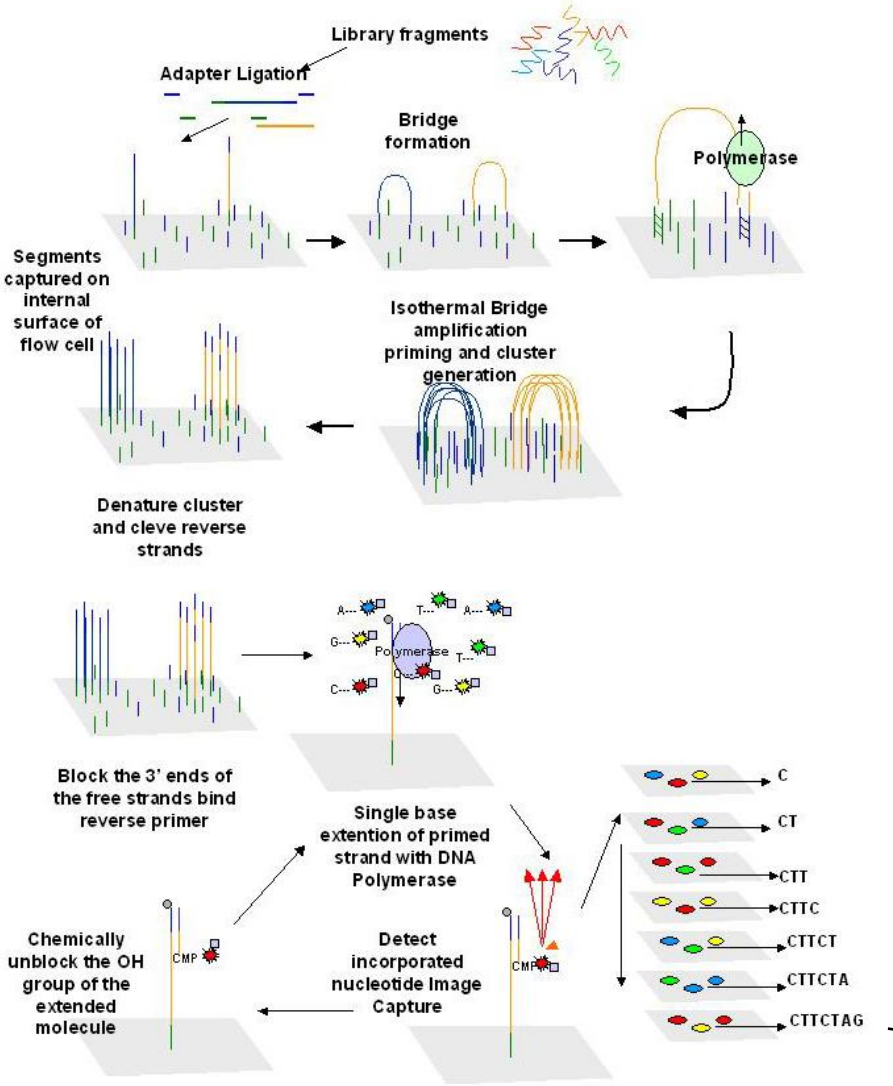


image → region

etc.

↑ confirmed variant

↑ candidate variant

↑ reference aligned read + mapping quality

↑ demultiplexed read

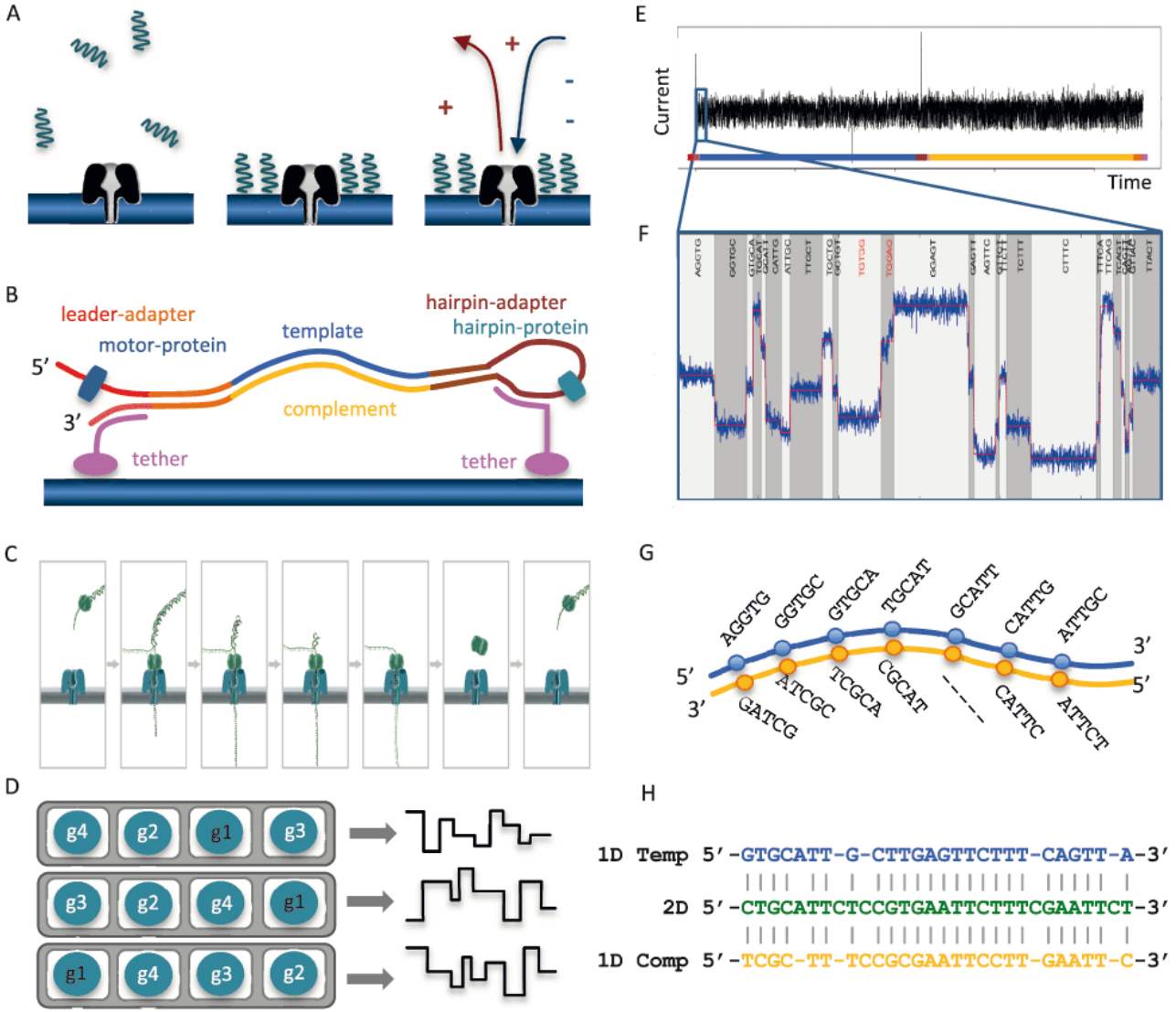
↑ adaptor-stripped read

↑ base call + per-position quality

↑ point intensity + noise

Images: Ken Laing IPC, Centre for Infection and Immunity Division of Clinical Sciences, St George's University of London, http://www.ipc.nxgenomics.org/newsletter/images/BridgeAmp.JPG and http://www.ipc.nxgenomics.org/newsletter/images/Extension.JPG

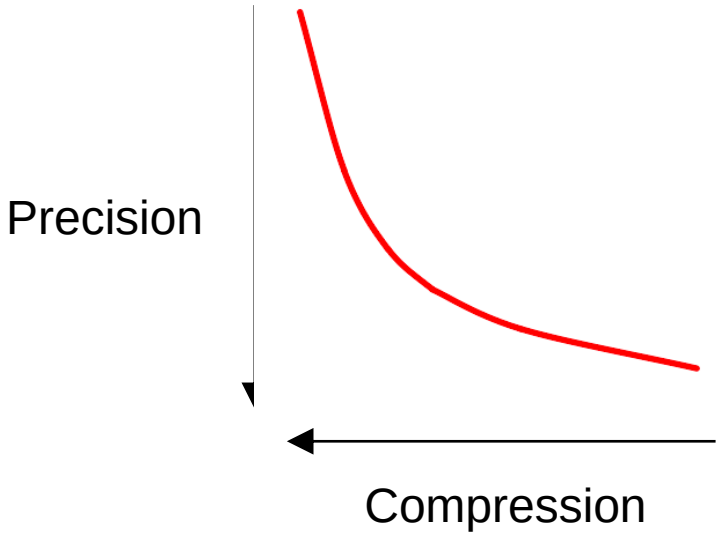
Figure 1. The Oxford Nanopore sequencing process.

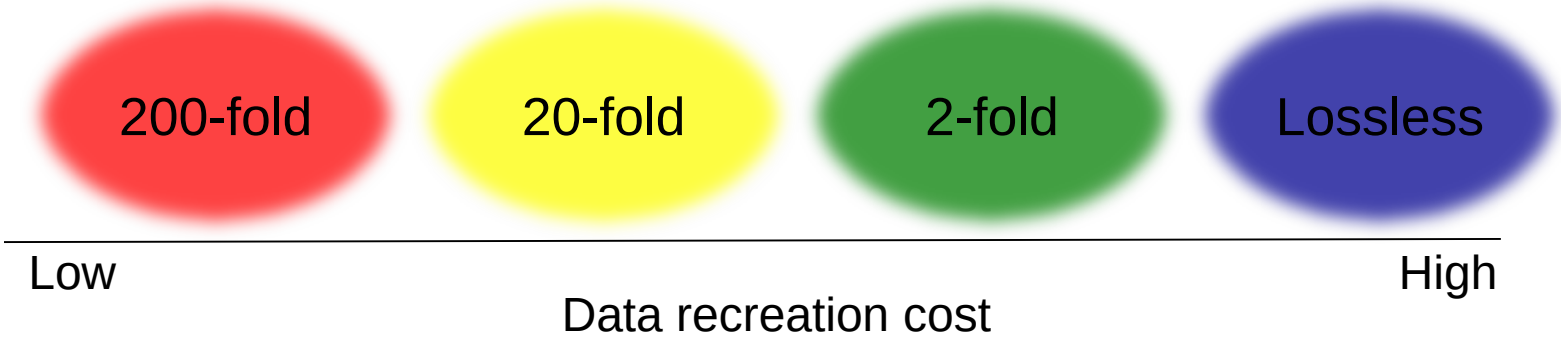


Ip CLC, Loose M, Tyson JR et al. 2015 [version 1; referees: awaiting peer review] F1000Research 2015, 4:1075 (doi: 10.12688/f1000research.7201.1)

F1000Research

Data reduction mark II: finely controlled loss of precision





Cochrane G., Cook C.E. and Birney E. (2012) The future of DNA sequence archiving. *GigaScience* 2012, **1**:2

Future

Future

- More of the same – GA4GH, Global Microbial Identifier, International Cancer Genome Consortium, 100K Genomes
  - Cells, conditions, developmental stages, organs, individuals, microbiomes, populations, environments, etc.
- New platforms
- New applications
- New paradigms for data use

Pathogen surveillance applications: turnaround

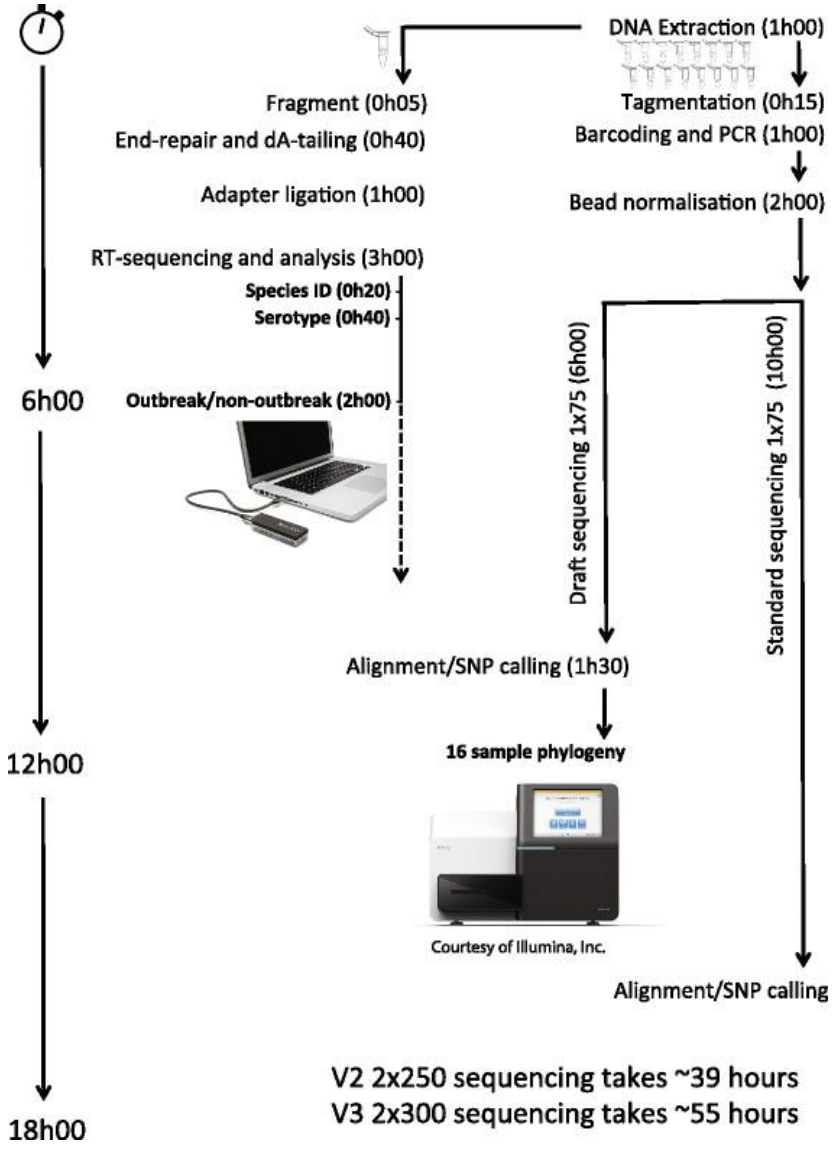
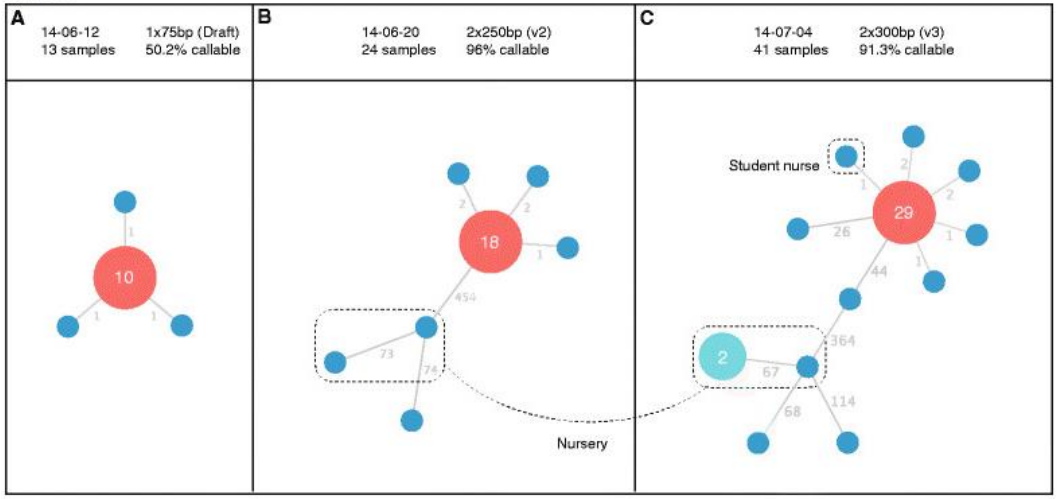


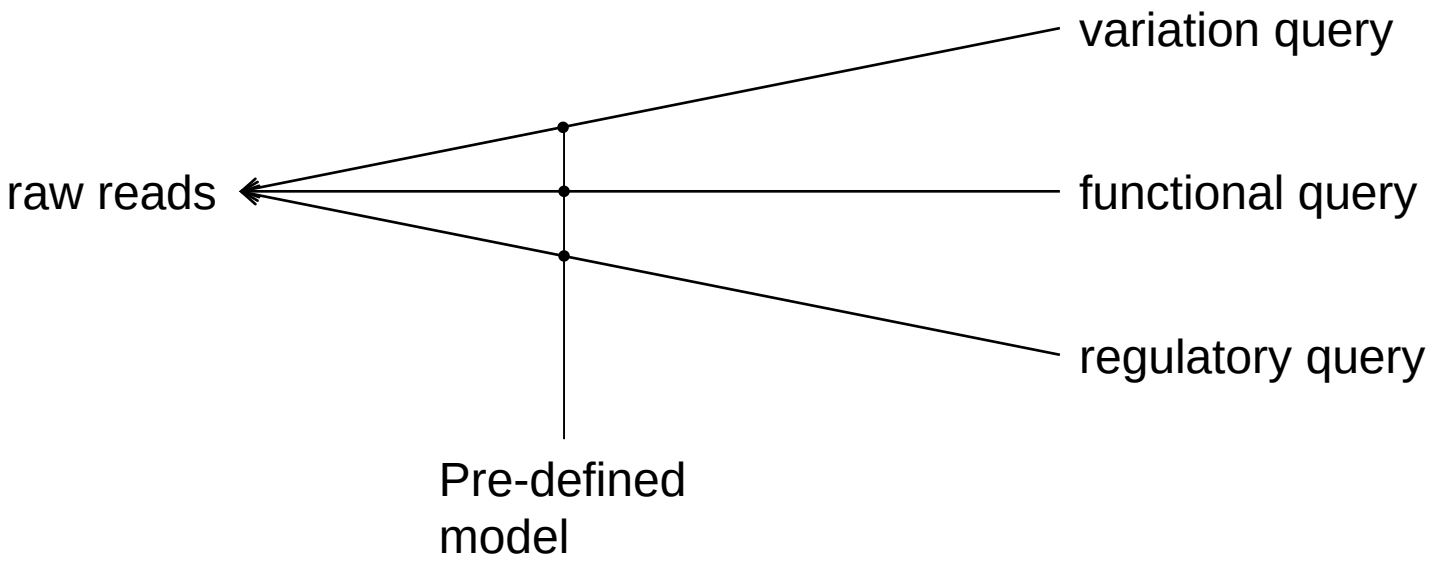
Fig. 1.



Quick, J., Ashton, P., *et al*. (2015) Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella. Genome Biology* 16:114

Pathogen surveillance applications: analysis paradigm

- Reference-based compression is expected to work well in terms of performance and footprint reduction
- Accessibility mark I -> accessiblity mark II



Acknowledgements

- EMBL European Bioinformatics Institute
  - Vadim Zalunin, Markus Hsi-Yang Fritz, Ewan Birney, Dmitriy Smirnov, Rasko Leinonen
- Wellcome Trust Sanger Institute
  - James Bonfield