

# Sequence Read Format (SuRFing the genome)

Asim Siddiqui

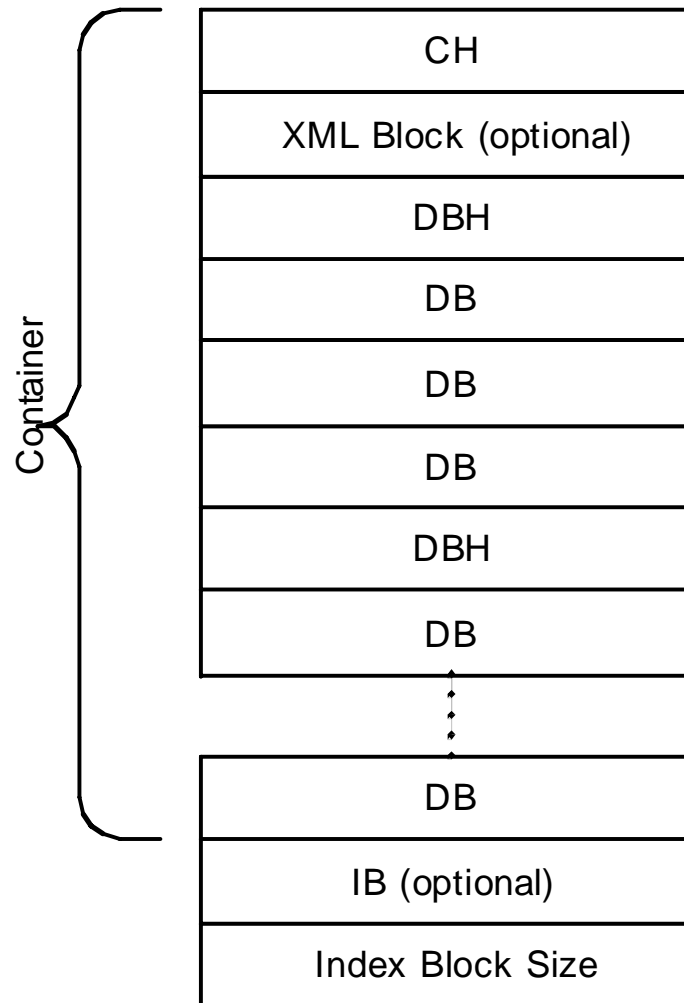
# Acknowledgements

- James Bonfield (Sanger)
- Gabor Marth (Boston College)
- Toby Bloom (Broad)
- Vladimir Alekseyev (NCBI)
- Paul Flicek (EBI)
- Darren Platt (JGI)
- Mike Attali (Helicos)
- James Knight (454/Roche)
- Clive Brown (formely Solexa now Sanger)
- + others

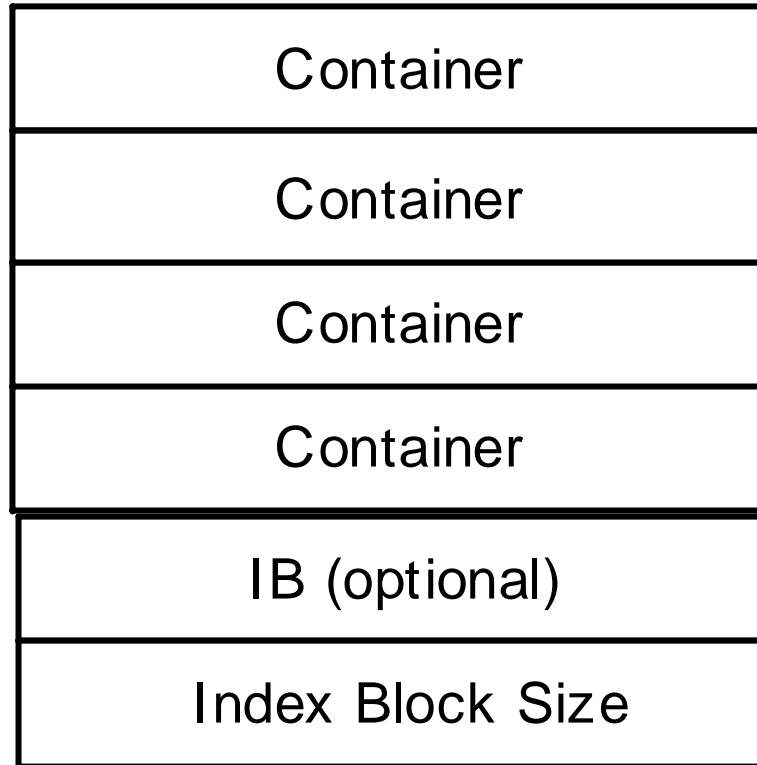
# General Concepts

- Container of block based data
  - Versioned container
- ZTR blocks (future proofed for other block types)
  - Binary format
  - Compression built-in
- Index
- Global and local read ids

# Container Structure



# Container Structure



- Index may be in a separate file

# Format Hierarchy

- Experiment annotation
- Genome annotation
- Summary assembly
- Detailed assembly
- Sequence

# Discussion Points

- Read ids
- Patient data
- Paired reads & other read relationships
- Submission vs. downstream formats
- Metadata
- API
- Timeline

# Read Ids

- Prefix/group id in DBH
- Two types of DBHs
  - Explicit
  - Incremental
- Predefined centre & manufacturer codes
- Global and local in scope
  - Use global ids for data sharing
  - Local ids for internal use



# Patient data

- Global ids may be relabeled by centre to de-identify data.
- Additional issues?

# Related reads

- Handled under a single read id
- ZTR header block in DHR includes read partitioning information and naming of “sub-reads”
- E.g. primer1:T;read1:P;primer2:T;read2:P - a paired end read
- E.g. Region 1;Region 1;Region 2 - for a duplicate read of one portion of the fragment followed by a single read of another

# Submission vs. processed formats

- Can SRF support both?
  - Support built-in
  - uptake will depend on community and enabling features
- Re-distribution of data can utilize global ids
- Creation of the detailed (genome/transcriptome/other-ome) assembly format and supporting functionality will enhance format's value

# Metadata

- XML block
- Purposely undefined
- Reserve an XML record  
“NCBI\_metadata”

# Submission types

- Suggest different records under the NCBI\_metadata record
  - Some common fields
- Does the SRF/ZTR structure provide sufficient flexibility?
- Who is going to define the metadata structure?

# API

- Definition
- Support
- Maintenance

# Timeline