Quest for Standard

Sequence alignment/map format (SAM) and SAMtools

Heng Li

Wellcome Trust Sanger Institute

7 August 2009

Outline

1. Quest for Standards
   - Burst of data volume and software
   - Difficulties in design
2. SAM Alignment Format
   - Overview of SAM/BAM
   - Implementation and Support
   - Technical innovations
3. Displaying Alignments
   - Alignment viewers
   - SAM/BAM and GBrowse

Outline

Difficulties in design

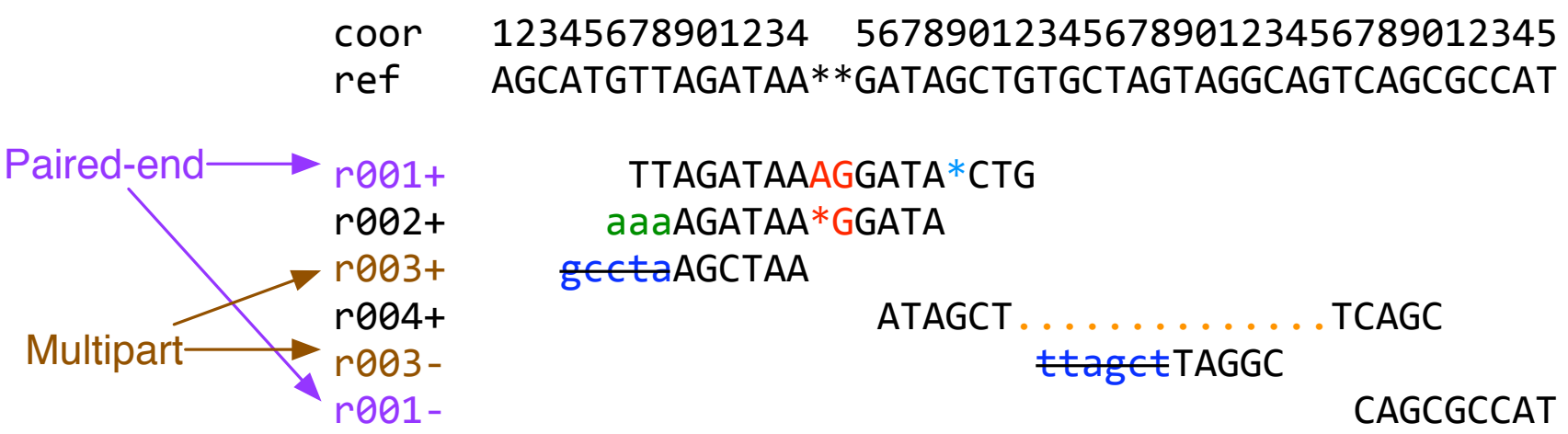
Feature	phrap ACE	GFF	SAM/BAM
Intended use	assembly	genomic features	various aln & assembly
Intended users	developers	more	more
Data volume	medium	small/medium	huge
Compression	no	optional	yes
Streamability	no	yes	yes
Indexing	not builtin	not builtin	yes
Meta data	limited	flexible	flexible

- Collaborative product
- 1000 Genomes Project provides the niche

Outline

SAM: Sequence Alignment/Map (format)

- Motivated by short read alignment but also working with long reads and *de novo* assemblies.
- GFF3-like TAB delimited format
  - 11 mandatory fields for key information
  - variable optional fields
  - predefined tags for non-standard information
  - simple to generate and to parse
- Extended CIGAR string for various types of alignments



Ins & padding
Soft clipping

Splicing
Hard clipping

```
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M * 0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       * 0   0 AGCTAA    *    NM:i:1
r004   0 ref 16 30 6M14N5M    * 0   0 ATAGCTTCAGC        *
r003  16 ref 29 30 6H5M       * 0   0 TAGGC     *    NM:i:0
r001  83 ref 37 30 9M         = 7 -39 CAGCGCCAT          *
```

```
ref  7 T 1 .    | ref 12 T 3 ...      | ref 17 T 3 ...
ref  8 T 1 .    | ref 13 A 3 ...      | ref 18 A 3 .-1G..
ref  9 A 3 ...  | ref 14 A 2 .+2AG.+1G. | ref 19 G 2 *.
ref 10 G 3 ...  | ref 15 G 2 ..       | ref 20 C 2 ..
ref 11 A 3 ..C  | ref 16 A 3 ...      | ...
```

BAM: Binary Alignment/Map (format)

- BAM is the exact binary representation of SAM.
- Zlib/gzip compatible compression (decompressed by zlib/gzip).
  - achieving 1 byte per raw base pair, including sequence, quality, read name, position and meta info.
- Streamability: processing alignments without loading the entire alignment into memory
  - BAM is usually sorted by the leftmost chromosomal position.
- Random access (BAM)
  - Quickly retrieving sequences overlapping a specified region.
  - Small index size ($\sim$9MB for deep human resequencing)

Implementations

- SAMtools: command line tools and C APIs
  - Conversion from other formats
  - SAM⇔BAM, indexing, sorting, merging, pileup, SNP/indel calling, alignment viewer ...
  - Native HTTP/FTP support
- Picard: command line tools and pure Java APIs
  - SAM⇔BAM, sorting, merging, ...
  - Better at merging and rmduping
- GATK: pileup, SNP calling and more command line tools in Java
- Bio::DB::Sam: Perl APIs built on top of SAMtools

3rd-party support

Program	r-SAM	w-SAM	r-BAM	w-BAM	comment
ABpipeline		yes			SOLiD pipeline
BLAST		converted			generic alignment
Bowtie		converted			short read aln
BWA		yes			short? read aln
GApipeline		converted			Illumina suite
IGV	yes	no	yes		generic viewer
Karma		yes			short read aln
MAQ		converted			short read aln
NovoAlign		yes			short read aln
PSL format		converted			BLAT aln format
SNP-o-matic		yes			aln&SNP calling
SOAP(2)		converted			short read aln
SSAHA2		yes			read alignment
Stampy		yes			short read aln
TopHat		yes			short RNAseq aln
ZOOM		converted			short read aln

BGZF: generic indexable compression format

- The standard gzip/zlib format is not block-wise. Indexing is intricate and inefficient.
- BGZF is separated into multiple standalone gzip/zlib blocks (64kB each).
- Random access achieved by virtual offset (64 bits): `blockNumber<< 16|inBlockOffset`.
- BGZF can be decompressed by zlib/gzip

BAM indexing

- Difficulty in indexing:
  - B-tree or linear index: inefficient for resolving 'overlap' queries
  - R-tree or binning index: difficult in streaming
- BAM indexing: binning plus linear index for alignments sorted by the leftmost coordinates.
  - For short read alignment, typically one seek function call for the retrieval of reads in a region (more efficient than R-tree).
  - Small index file size ($\sim$9MB for deep human resequencing)

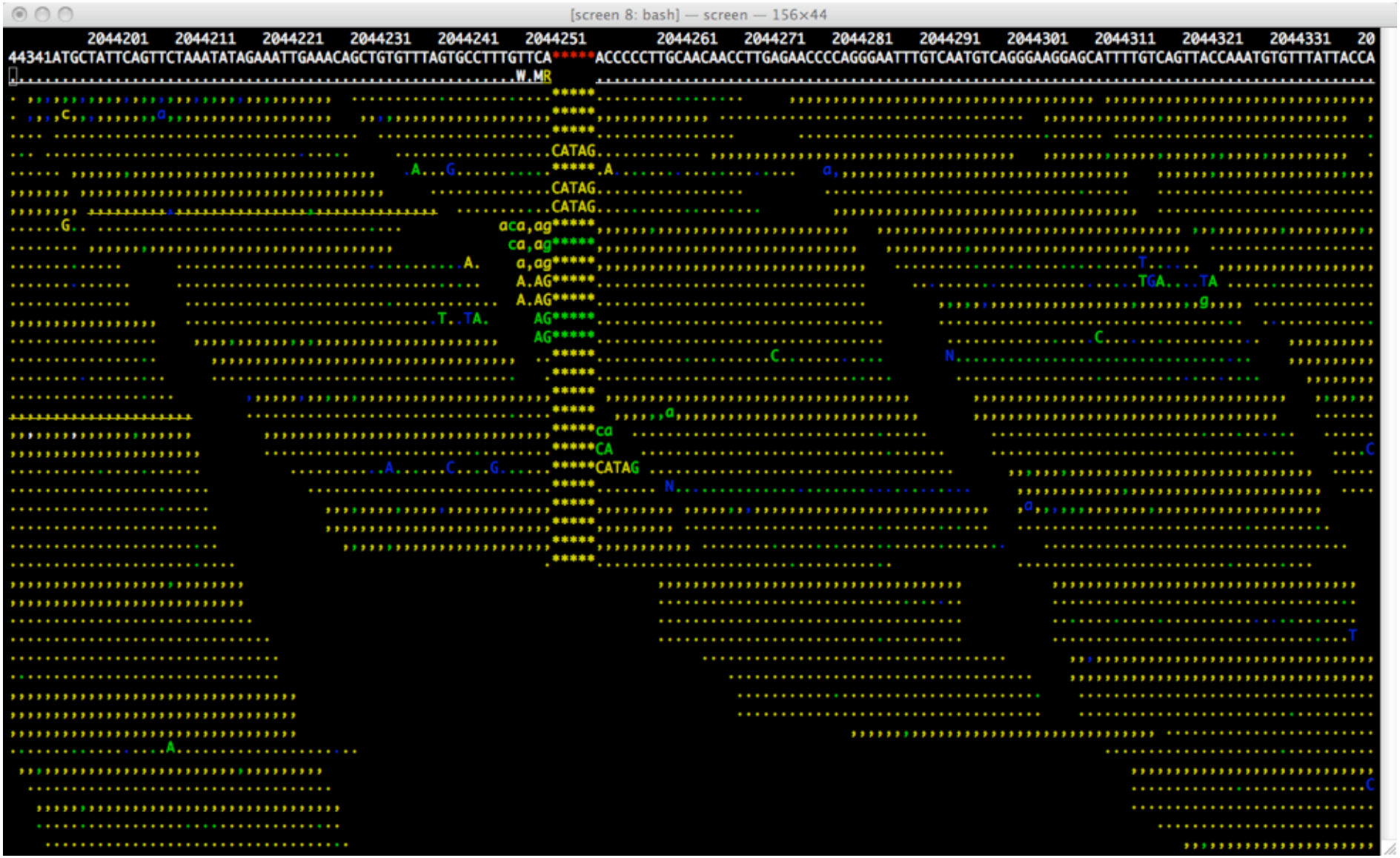
Native HTTP/FTP support

- Retrieve alignments overlapping a specified region from a remote file on http/ftp.
- Usage: simply replace the input BAM file name as a URL (http/ftp only).
- Applications:
  - Downloading part of alignments directly from ftp/http.
  - Viewing alignments without downloading huge alignment files.
  - Genome browser: custom tracks without uploading entire alignment.

Outline

What is viewer used for?

- A great help for method development:
  - Visually understand the alignment: the error rate, the depth, etc.
  - Validate aligner results: even read depth? right coordinates? right gaps?
  - Validate SNP/indel calls: human eyes are always better.
  - Validate structural variations: pair-end information
- Who will look at alignments from the 1000 Genomes Project? (an open question)

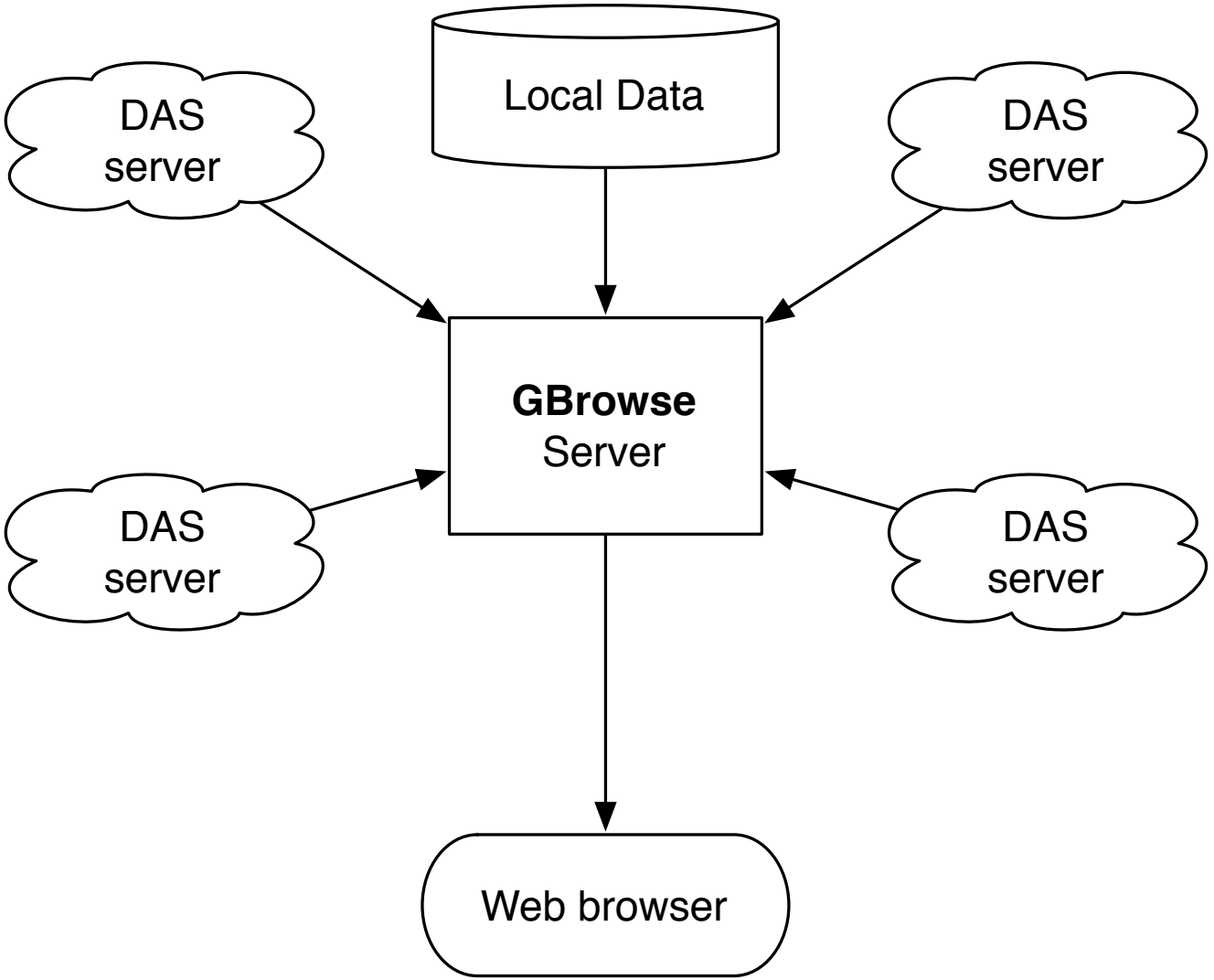

[screen 8: bash] — screen — 156×44
2044201 2044211 2044221 2044231 2044241 2044251 2044261 2044271 2044281 2044291 2044301 2044311 2044321 2044331
44341ATGCTATTCAGTTCTAAATATAGAAATTGAAACAGCTGTGTTTAGTGCCTTTGTTCA*****ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAATTTGTCAATGTCAGGGAAGGAGCATTTTGTCAGTTACCAAATGTGTTTATTACCA
CATAG
CATAG
CATAG
aca,ag*****
ca,ag*****
a,ag*****
A.AG*****
A.AG*****
AG*****
AG*****
*****ca
*****CA
*****CATAG

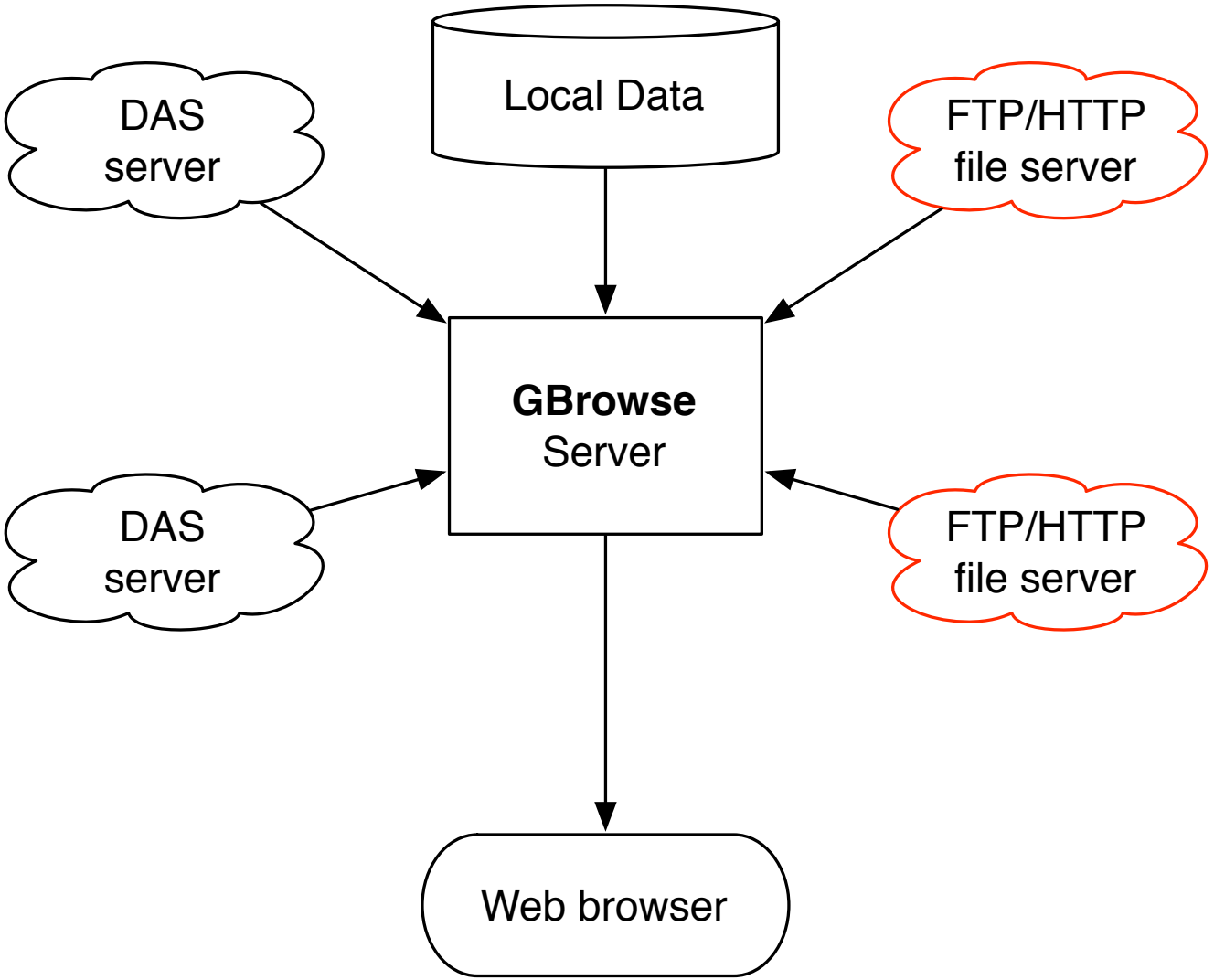
Alignment viewers

- Samtools tview (text viewer):
  - Based on ncurses.
  - Viewing alignments on FTP/HTTP.
  - Simple functionality (no annotation, paired-end, multiple tracks...)
- Broad's IGV (Integrative Genomics Viewer):
  - Server-client: IGV protocol?
  - High-quality Java application
  - Annotation, multiple tracks and more.

SAM/BAM and GBrowse

- Premise: SAM/BAM parser in Perl: Bio::DB::Sam (based on SAMtools C APIs)
- For SAM/BAM, GBrowse is a versatile shared alignment viewer:
  - mutliple tracks and gene annotations
  - thin client (web browser)
- For GBrowse, SAM/BAM can provide an efficient way to
  - access large-scale new sequencing data
  - store various types of alignment (EST, mRNA, etc.) as an alternative to SQL database.
  - realize distributed alignment resources


Local Data
DAS server
DAS server
GBrowse Server
DAS server
DAS server
Web browser


DAS
server
Local Data
FTP/HTTP
file server
GBrowse
Server
DAS
server
FTP/HTTP
file server
Web browser

Distributed alignments

- Feasibility:
  - Native HTTP/FTP support in SAMtools and Bio::DB::Sam (yet?)
  - Compressed alignment files.
  - For short reads, one seek call (establishing network connection) is required to get alignments in a region.
- Advantages:
  - Little configuration at the server hosting alignments.
  - Compressed data transfer between file servers and the GBrowse server.
- Major obstacles:
  - Index files ($\sim$9MB) have to sit on local disks at the GBrowse server.
  - Matching the reference sequences may be an issue.
  - Bandwidth and caching.

Summary

- SAM/BAM is a generic nucleotide alignment format that is
  - is simple to understand, easy to generate and easy to parse
  - is compact in file size
  - is streamable
  - supports fast random access

Acknowledgements

- Bob Handsaker
- Richard Durbin
- Goncalo Abecasis
- Fiona Hyland
- Richa Agarwala
- Gabor Marth
- Tim Fennel
- Alec Wysoker
- Jue Ruan
- Nils Homer
- James Bonfield
- John Marshall