Next Generation Sequencing & Galaxy

Aldo Jongejan
a.jongejan@amc.uva.nl

Bioinformatics

- A lot of data is produced!!
  - *Raw data, i.e. reads and qualities*

- NGS:
  - *Raw data - TeraBytes*
  - *Text-sequence data - GigaBytes*
  - *Sequence variant data - Mega/KiloBytes*

Current Solution

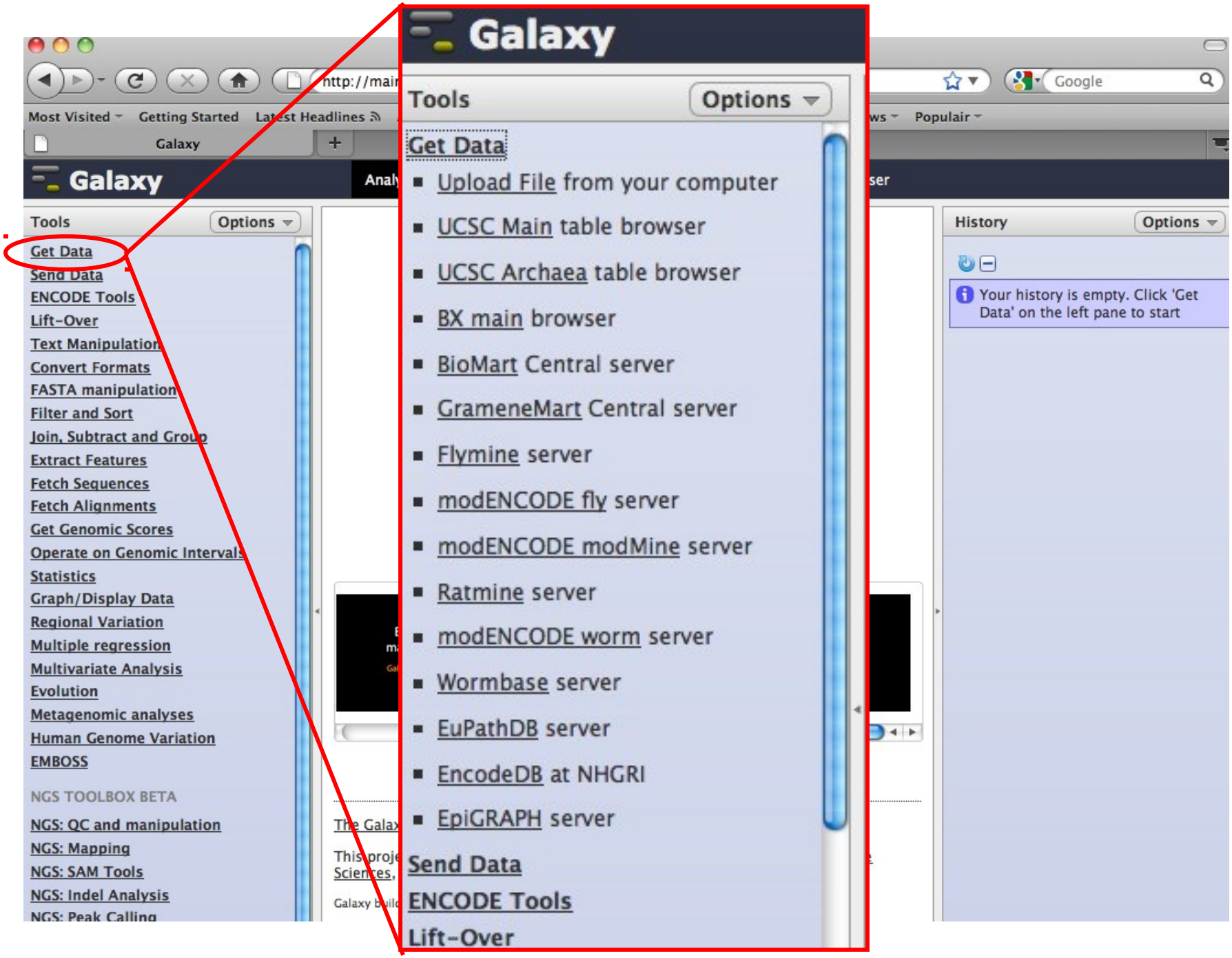
- Indexing
  - *The reads*
  - *The genome*

  - *Fast sorting/indexing algorithms*
    - BWA (Burrow-Wheeler), SHRiMP, MAQ,Soap, BFAST, Bowtie, ……
    - Indexes become larger and larger…

NGS apps and interfaces

- Availability
- Resources
- User friendliness

- Galaxy
  - *http://main.g2.bx.psu.edu*
  - *local*


Galaxy
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
History
Your history is empty. Click 'Get Data' on the left pane to start
Upload File from your computer
UCSC Main table browser
UCSC Archaea table browser
BX main browser
BioMart Central server
GrameneMart Central server
Flymine server
modENCODE fly server
modENCODE modMine server
Ratmine server
modENCODE worm server
Wormbase server
EuPathDB server
EncodeDB at NHGRI
EpiGRAPH server

Upload your file

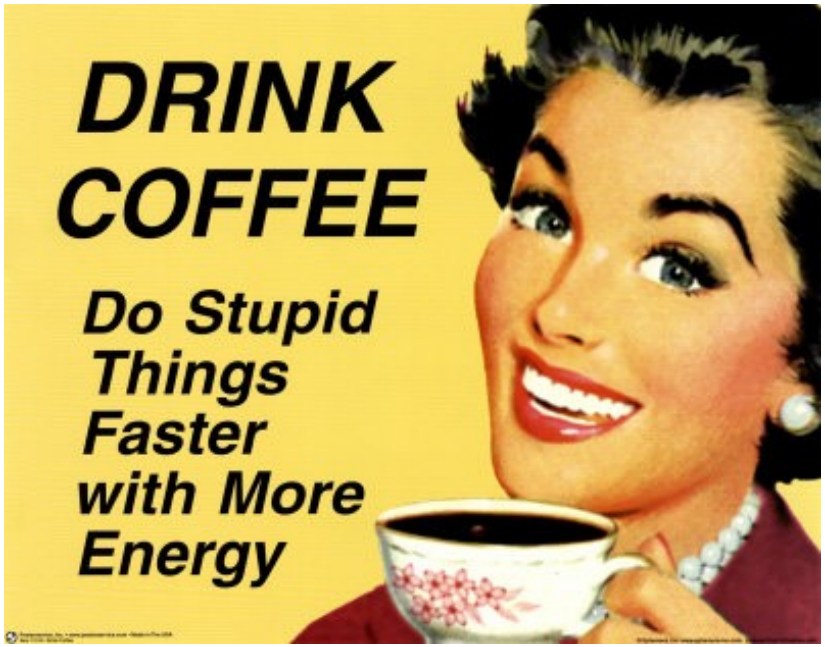
- Many file formats are supported
  - *But be aware of the size of your data!*

- NGS data can be obtained from the 1000 Genome Project:
  - *www.1000genomes.org*

  - *Get some data here!*

Upload your file

- Go back to Galaxy and click “Get Data”
  - *Browse to your file*
  - *Select the genome you want to use as reference (a.k.a Human Genome build GRCh37/hg19)*
  - *And wait….*



History

Options

1: SRR062634.filt.fastq

77.0 Mb

format: fastq, database: hg19

Info: uploaded fastq file

```
@SRR062634.321 HWI-EAS110_103327062:6:1:1446:951/2
TGATCATTTGATTAATACTGACATGTAGACAAGAAGAAAAGTATGTTTCATGCTATTTTGAGTAACTTCCATTTAGAAGCCTACTCCTGAGC
+
B5=BD5DAD?:CBDD-DDDDDCDDB+-B:;?A?CCE?;D3A?B?DB??;DDDEEABD+>DAC?A-CD-=D?C5A@::AC-?AB?=:>CA@##
@SRR062634.488 HWI-EAS110_103327062:6:1:1503:935/2
AATGTTATTAAAAATGGACACCTTTTTCTCACACATTCAGTTTCATTGTCTCGCACCCCATCGTTTTACTTTTCTTCCTTCAGAAAATGATA
```

Galaxy

Tools Options

Evolution

Metagenomic analyses

Human Genome Variation

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

ILLUMINA DATA

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq

FASTQ Groomer

File to groom:

1: SRR062634.filt.fastq

Input FASTQ quality scores type:

Illumina 1.3+

Advanced Options:

Show Advanced Options

Output FASTQ quality scores type:

Sanger (recommended)

Galaxy tools are designed to work with the Sanger Quality score format.

Force Quality Score encoding:

ASCII

Summarize input data:

Summarize Input

Execute

What it does

This tool offers several conversions options relating to the FASTQ format.

NGS TOOLB

NGS: QC and

NGS: Mappin

ILLUMINA

- Map with
- Map with

ROCHE-4

- Lastz map reference
- Megablast against ht databases
- Parse blas

AB-SOLID

- Map with

NGS: SAM To

Map with BWA

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Select a reference genome:

Human (Homo sapiens): hg19 Full

Is this library mate-paired?:

Single-end

FASTQ file:

2: FASTQ Groomer on data 1

Must have Sanger-scaled quality values with ASCII offset 33

BWA settings to use:

Commonly Used

For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

Suppress the header in the output SAM file:

☑

BWA produces SAM with several lines of header information

Execute

What it does

SAMtools

History

3: Map with BWA on data 2: mapped reads

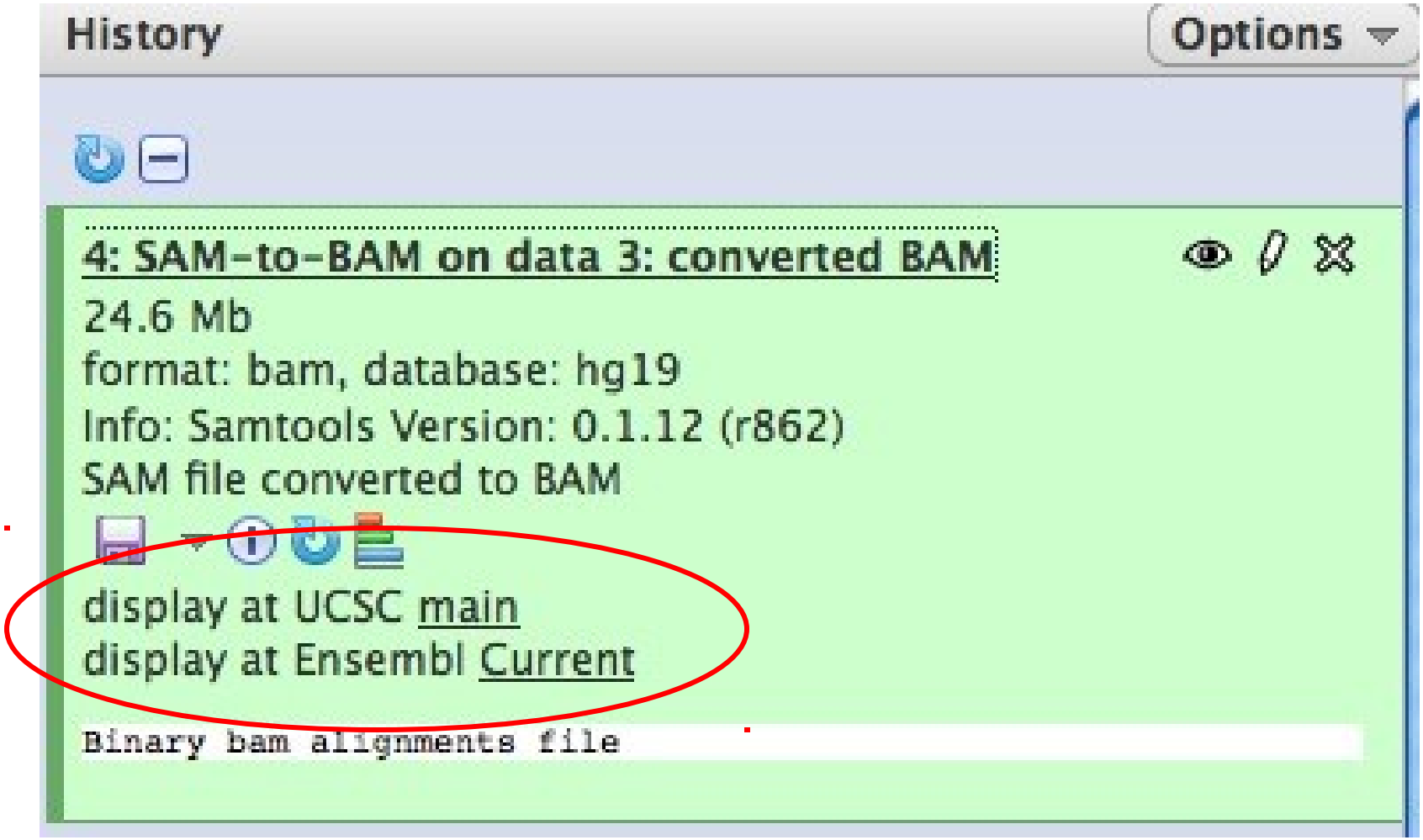
~313,657 lines
format: sam, database: hg19
Info: BWA Version: 0.5.6 (r1303)
BWA run on single-end data

1.QNAME	2.FLAG					
SRR062634.321	16	chr13	32427664			37
M:i:2	X0:i:1	X1:i:0	XM:i:2	XO:i:0	XG:i:0	MD:
SRR062634.488	0	chr7	52876683			37
i:1	X0:i:1	X1:i:0	XM:i:1	XO:i:0	XG:i:0	MD:
SRR062634.849	0	chr4	134221215			37
:i:2	X0:i:1	X1:i:0	XM:i:2	XO:i:0	XG:i:0	MD:

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

Visualization

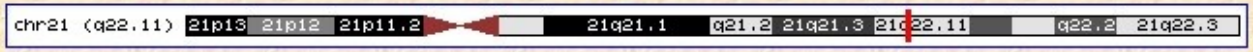


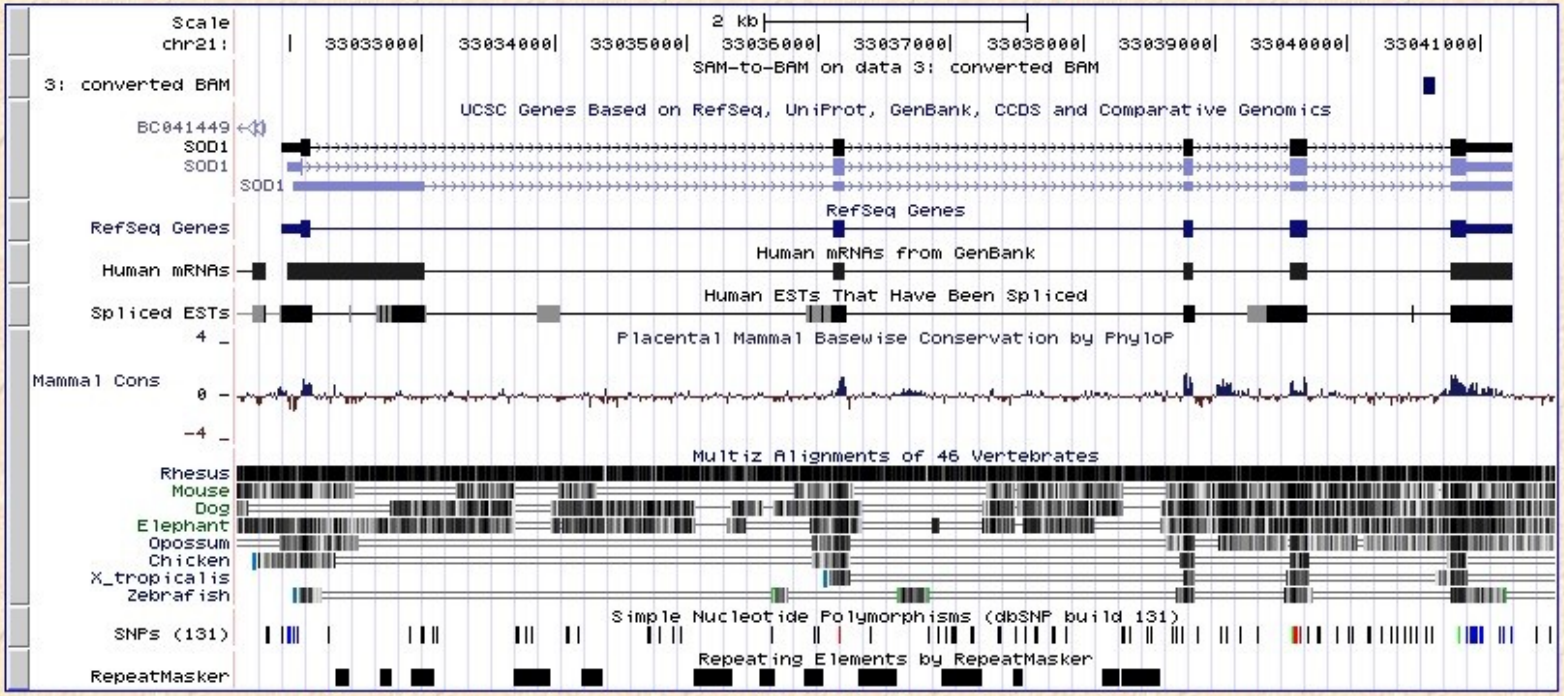
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr21:33,031,597-33,041,570 gene jump clear size 9,974 bp. configure





move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks.

move end < 2.0 >

track search default tracks default order hide all manage custom tracks configure reverse refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

Custom Tracks

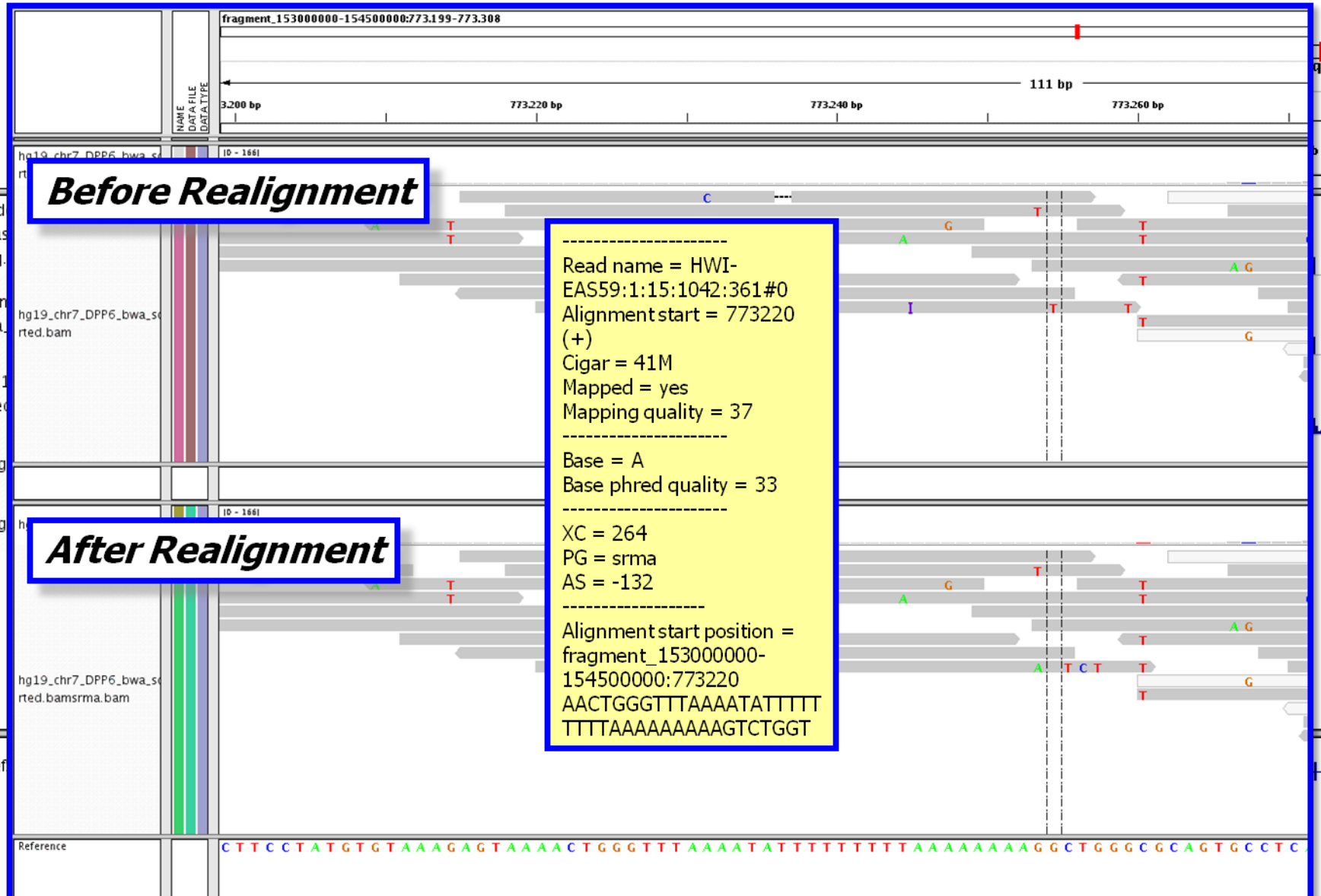
refresh

SAM-to-BAM on data 3: converted BAM

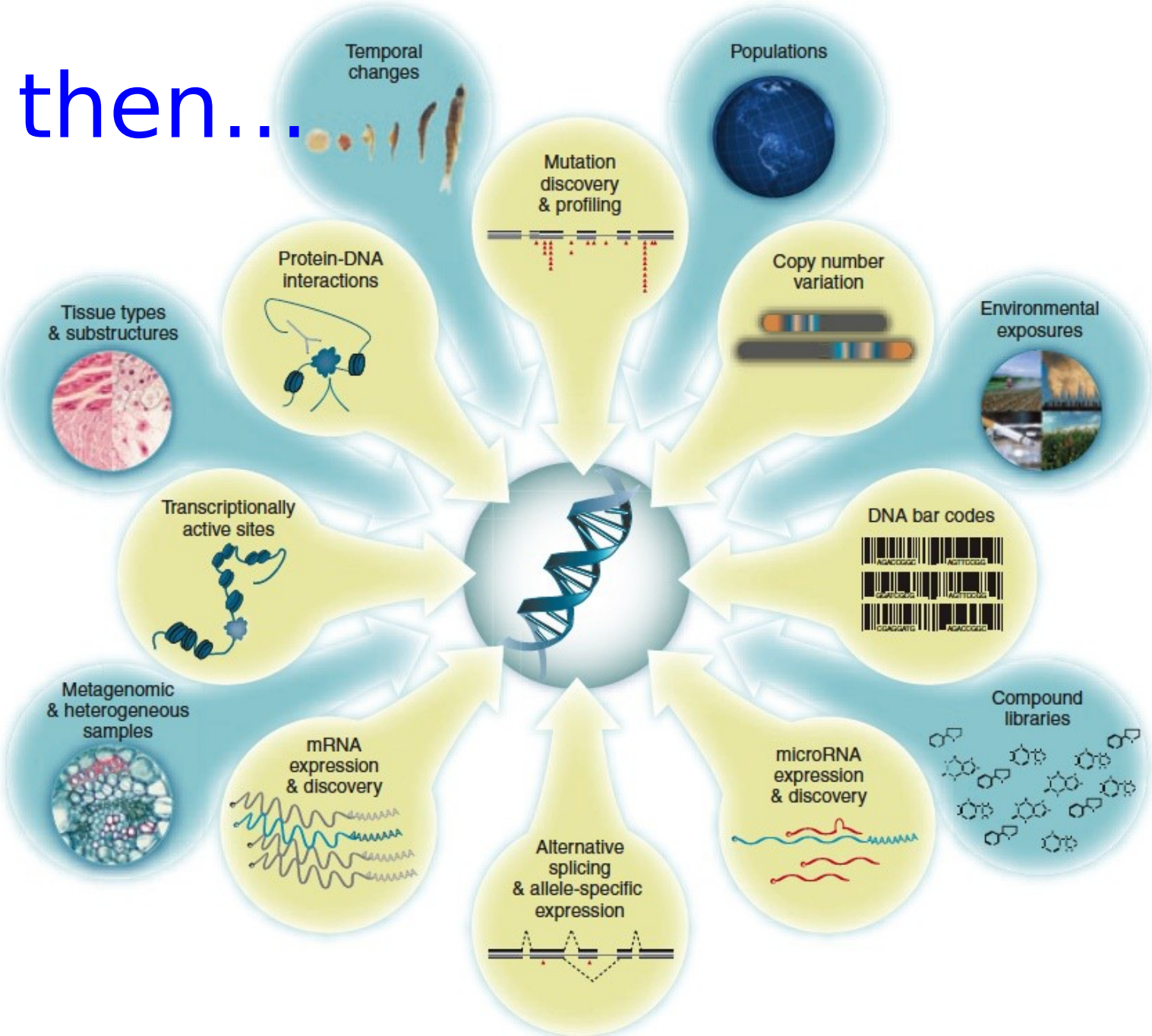
Visualization

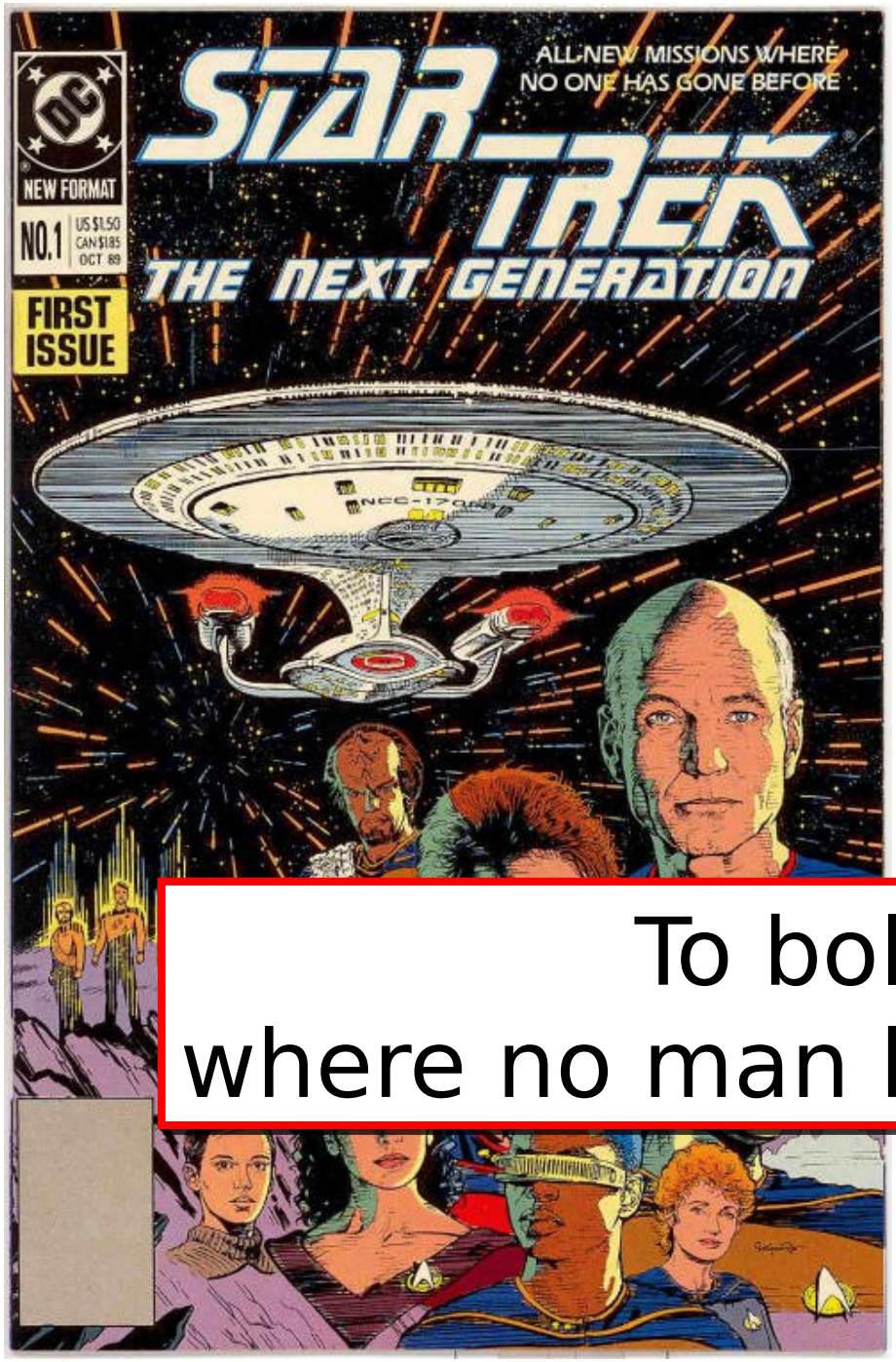
- Other Genome Browsers
  - *Ensembl, NCBI*
  - *IGV (can be installed locally)*
  - *Savant (can be installed locally)*
  - *…*

Visualization



And then…


Temporal changes
Populations
Mutation discovery & profiling
Protein-DNA interactions
Copy number variation
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
Compound libraries
mRNA expression & discovery
microRNA expression & discovery
Alternative splicing & allele-specific expression



To boldly go
where no man has gone before!