

UCLA Institute for Quantitative and Computational Biology

March 2020

QCB Workshop W5a: RNAseq-1 (day 2)

Nicolas Rochette

(EEB/ISG)

Karolina Kaczor-Urbanowicz

(Oral Biology & Medicine)

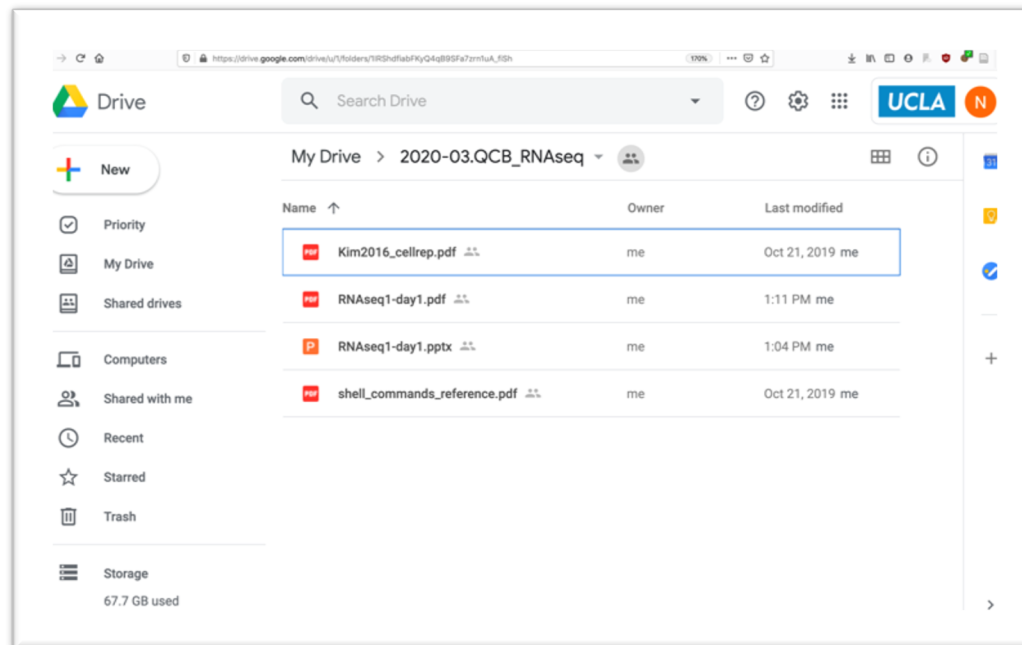
Workshops' google drive folder

Slides & supporting documents:

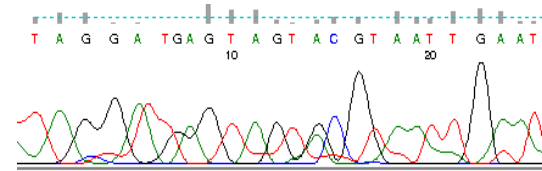
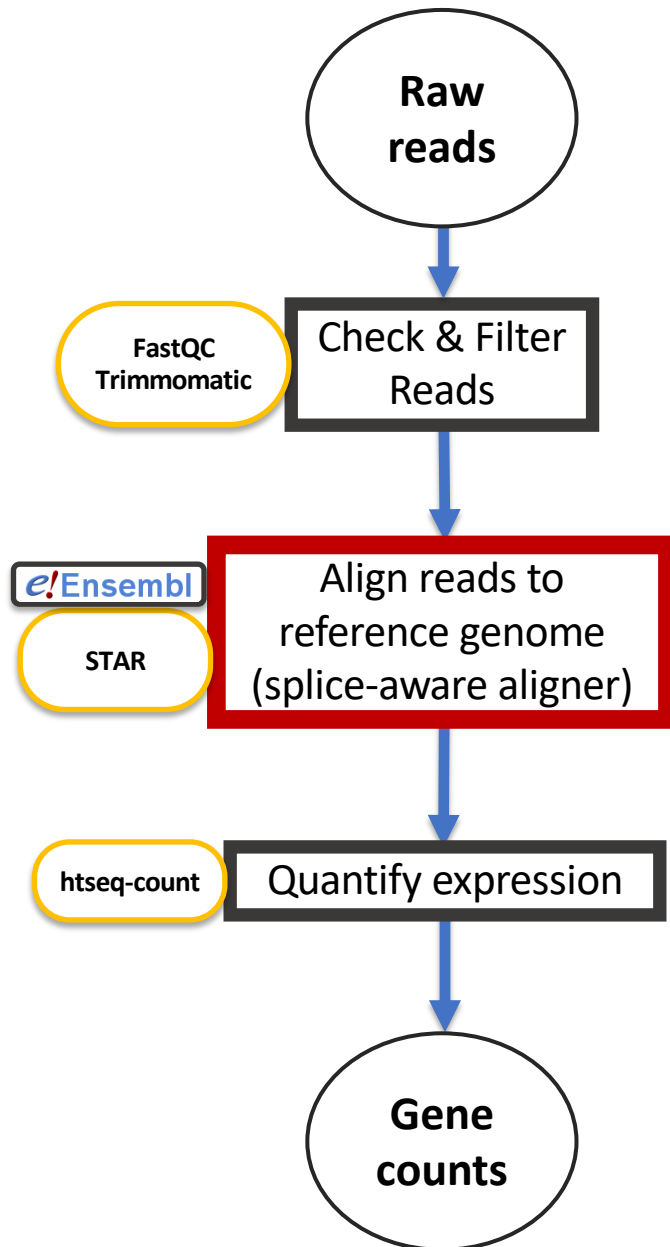
https://drive.google.com/drive/folders/1IRShdfiabFKyQ4qB9SFa7zrn1uA_fiSh

<https://tinyurl.com/sshouvww>

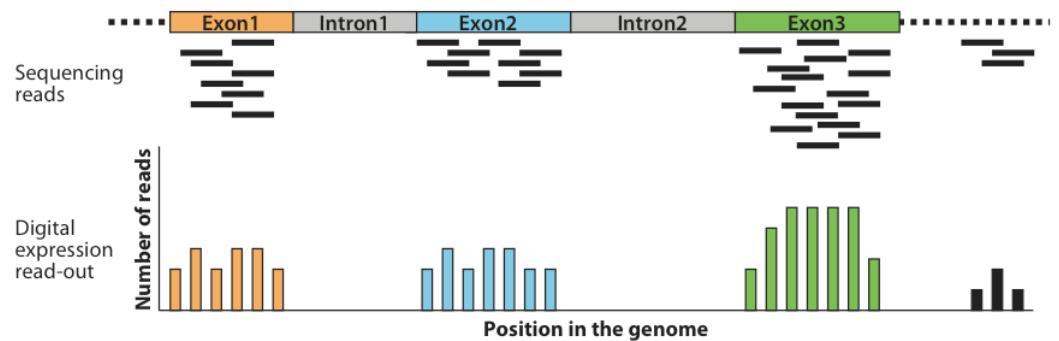
Please bookmark it / write it down!



RNA-seq analysis: Overview



Reference genome sequence



	sample1	sample2	sample3	...
gene1	999	701	616	
gene2	532	520	41	
gene3	14	36	305	
...				

Connecting to the cluster

- Remote shell

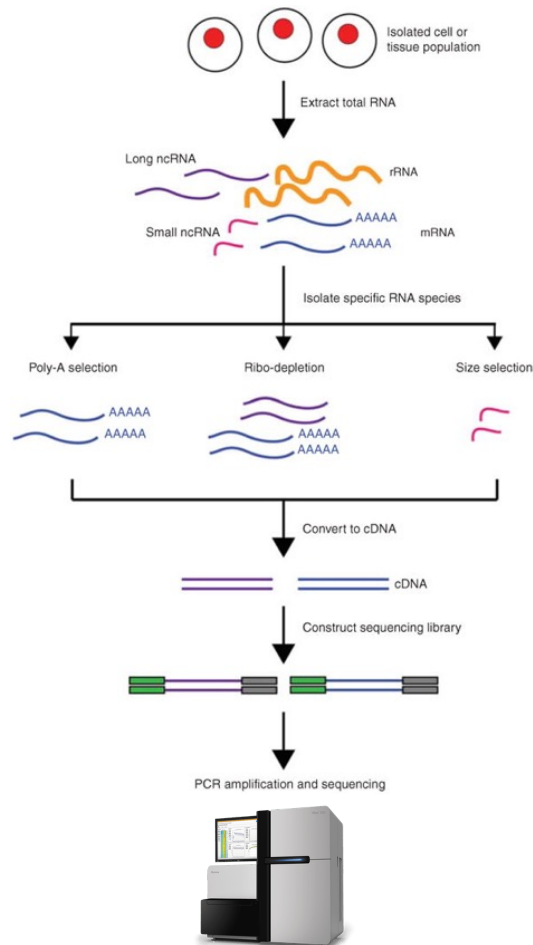
ssh USER@hoffman2.idre.ucla.edu

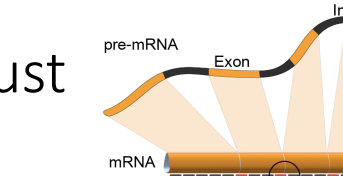
- (through WSL/Putty/Mobaxterm on Windows)
- Once connected, start an 'interactive job' requesting 4 hours and 32GB memory (STAR needs a lot of memory)

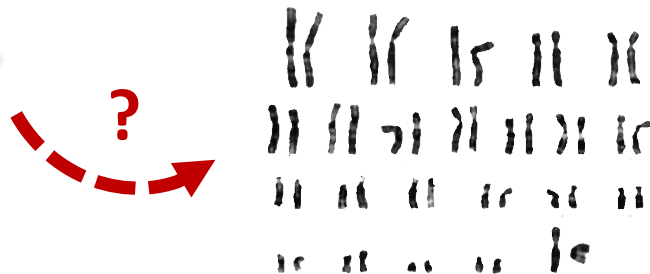
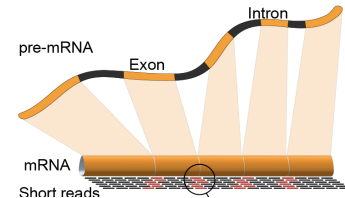
qssh -l h_rt=4:00:00,h_data=32G

Read alignments: principle

Read alignment / “mapping”



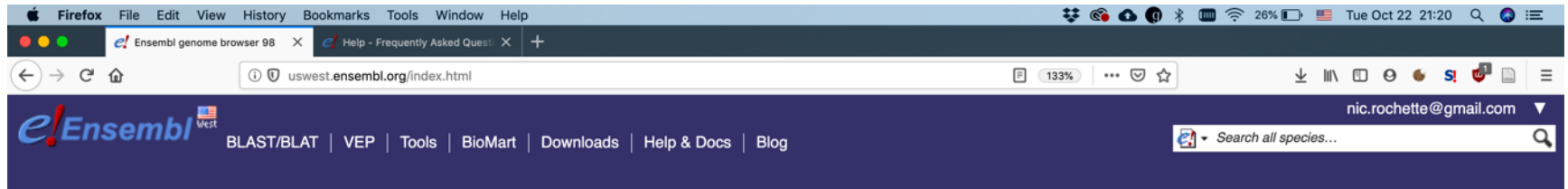
- During library preparation, DNA/RNA is fragmented: **the fragments** from the entire genome/transcriptome **become completely mixed-up**
 - The goal of read alignment/mapping is to put things back into place
i.e. **determine which read goes where in the genome**
 - For RNA data, methods must account for splicing
- 



Reference genomes

Ensembl

<https://www.ensembl.org/>



Tools
[All tools](#)

BioMart >
Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >
Search our genomes for your DNA or protein sequence

Variant Effect Predictor >
Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species for

e.g. **BRCA2** or **rat 5:62797383-63627669** or **rs699** or **coronary heart disease**

All genomes
-- Select a species --

- [View full list of all Ensembl species](#)
- [Edit your favourites](#)

Favourite genomes

-  **Human**
GRCh38.p13
[Still using GRCh37?](#)
-  **Mouse**
GRCm38.p6
-  **Northern American deer mouse**
HU_Pman_2.1

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 98 (September 2019)

- dbSNP152 for human with changes to repeat expansion/retraction representation
- Update to **GENCODE 32** for human
- New beta Post-GWAS analysis pipeline tool online
- New genomes: 11 pig breeds with comparative analysis between them and other agricultural species
- New gene annotation: dog, cat, horse, rabbit, grey short-tailed opossum, marmoset and rhesus monkey
- New genomes: nine fish, one frog, five plants, one worm and one diatom

[More release news](#) on our blog

Other news from our blog

- 03 Oct 2019: [Cool stuff the Ensembl VEP can do: predict pathogenicity of missense variants](#)
- 26 Sep 2019: [Ensembl 98 has been released!](#)
- 19 Sep 2019: [Simplifying our GRCh37 Services](#)

Ensembl (ctd.)

Firefox File Edit View History Bookmarks Tools Window Help

Mus musculus - Ensembl genome browser

uswest.ensembl.org/Mus_musculus/Info/Index

133%

nic.rochette@gmail.com

Search all species...

Mouse (GRCm38.p6)

Search Mouse (*Mus musculus*)

Search all categories Search Mouse... Go

e.g. [Cntnap1](#) or [4:136366473-136547301](#) or [rs27096498](#) or [adipocyte](#)

Genome assembly: GRCm38.p6 (GCA_000001635.8)

[More information and statistics](#)

[Download DNA sequence \(FASTA\)](#)

[Convert your data to GRCm38 coordinates](#)

[Display your data in Ensembl](#)

Other reference assemblies

- NCBIM37 (Ensembl release 67)

Other strains

This species has data on 15 additional strains. [View list of strains](#)

[View karyotype](#)

[Example region](#)

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

[More about this genebuild](#)

[Download FASTA](#) files for genes, cDNAs, ncRNA, proteins

[Download GTF or GFF3](#) files for genes, cDNAs, ncRNA, proteins

[Update your old Ensembl IDs](#)

[Example gene](#)

[Example transcript](#)

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

[More about comparative analysis](#)

[Download alignments \(EMF\)](#)

[Example gene tree](#)

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

[More about variation in Ensembl](#)

[Download all variants \(GVF\)](#)

[Variant Effect Predictor](#)

[Example variant](#)

Ensembl (ctd.)

The screenshot shows the Ensembl website for Mouse (GRCm38.p6). The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar is present with the text 'Search all species...'. Below the navigation bar, the 'Search Mouse (Mus musculus)' section contains a search input field and a 'Go' button. The main content area is divided into several panels:

- Genome assembly: GRCm38.p6 (GCA_000001635.8)**: This panel includes links for 'More information and statistics', 'Download DNA sequence (FASTA)', 'Convert your data to GRCm38 coordinates', and 'Display your data in Ensembl'. It also lists 'Other reference assemblies' (NCBIM37) and 'Other strains'. A red arrow points to this section.
- Gene annotation**: This panel includes links for 'More about this genebuild', 'Download FASTA files for genes, cDNAs, ncRNA, proteins', 'Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins', and 'Update your old Ensembl IDs'. It also features an 'Example gene' and an 'Example transcript'. A red arrow points to this section.
- Comparative genomics**: This panel includes links for 'More about comparative analysis' and 'Download alignments (EMF)'. It also features an 'Example gene tree'.
- Variation**: This panel includes links for 'More about variation in Ensembl', 'Download all variants (GVF)', and 'Variant Effect Predictor'. It also features an 'Example variant'.

A “reference genome” has two core parts:

1. The **sequence** itself (assembly; FASTA format)
2. The **gene positions** (annotations/genebuild; GTF/GFF3 format)

(NB. on the FTP, look for an assembly named `*.dna_sm.primary_assembly.fa.gz`)

Ensembl (ctd.)

- NR2E3 gene (a key target of NRL)

The screenshot displays the Ensembl genome browser interface for the NR2E3 gene in Mouse (GRCm38.p6). The browser window shows the URL `uswest.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000032292;r=9:59942771-59960659`. The page is titled "Gene: Nr2e3 ENSMUSG00000032292".

Gene-based displays:

- Summary
- Splice variants
- Transcript comparison
- Gene alleles
- Sequence
- Secondary Structure
- Comparative Genomics
- Genomic alignments
- Gene tree
- Gene gain/loss tree
- Orthologues
- Paralogues
- Ensembl protein families
- Strains
- Ontologies
- GO: Molecular function
- GO: Biological process
- GO: Cellular component
- Phenotypes
- Genetic Variation
- Variant table
- Variant image
- Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
- Gene history

Gene: Nr2e3 ENSMUSG00000032292

Description: nuclear receptor subfamily 2, group E, member 3 [Source:MGI Symbol;Acc:MGI:1346317]

Gene Synonyms: Pnr, RNR, photoreceptor-specific nuclear receptor

Location: Chromosome 9: 59,942,771-59,960,659 reverse strand. GRCm38:CM001002.2

About this gene: This gene has 9 transcripts ([splice variants](#)), 236 orthologues, 11 paralogues, is a member of 1 Ensembl protein family and is associated with 10 phenotypes.

Transcripts: [Show transcript table](#)

Summary

Name: [Nr2e3](#) (MGI Symbol)

CCDS: This gene is a member of the Mouse CCDS set: [CCDS23255.1](#)

UniProtKB: This gene has proteins that correspond to the following UniProtKB identifiers: [Q9QXZ7](#)

Ensembl version: ENSMUSG00000032292.8

Gene type: Protein coding

Annotation method: Annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation, see [article](#).

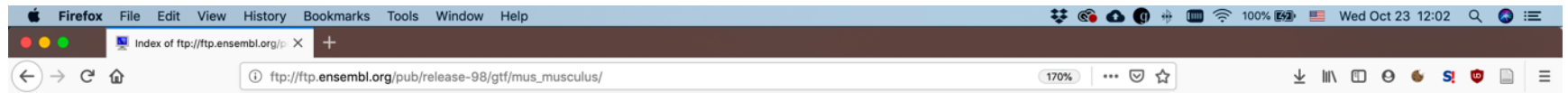
Annotation Attributes: overlapping locus [[Definitions](#)]

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

Regulation Legend:

- CTCF
- Open Chromatin
- Promoter Flank
- Enhancer
- Promoter
- Transcription Factor Binding Site

Ensembl FTP



Index of ftp://ftp.ensembl.org/pub/release-98/gtf/mus_musculus/

 [Up to higher level directory](#)

Name	Size	Last Modified
File: CHECKSUMS	1 KB	9/4/19 8:48:00 PM PDT
File: Mus_musculus.GRCm38.98.abinitio.gtf.gz	3288 KB	8/29/19 8:32:00 PM PDT
File: Mus_musculus.GRCm38.98.chr.gtf.gz	29540 KB	8/29/19 8:24:00 PM PDT
File: Mus_musculus.GRCm38.98.chr_patch_hapl_scaff.gtf.gz	29930 KB	8/29/19 8:27:00 PM PDT
File: <u>Mus_musculus.GRCm38.98.gtf.gz</u>	29548 KB	8/29/19 8:24:00 PM PDT
File: README	10 KB	8/29/19 8:28:00 PM PDT

FASTA files

```
>1 dna_sm:chromosome chromosome:GRCm38:1:1:195471971:1 REF
TTCTGTTTCTATTTTGTGGTTACTTTGAGGAGAGTTGGAATTAGGTCTTCTTTGAAGGTC
TGGTAGAACTCTGCATTAAACCCATCTGGTCCTGGGCTTTTTTTTTTTTTTTTTTTTTT
TTTGGGTGGGAGACTATTGATGACTGCCTCTATTTCTTTAGGGGAAATGGGACTTTTAGT
...

>10 dna_sm:chromosome chromosome:GRCm38:10:1:130694993:1 REF
CCGAATGGCAGAGAAACACCTGAATAAAATGTTCAACATCCTTAATCATCAGGGAAATGC
AAATCAAAACAACACTGAGATTCCACTTCACTCCAGTTAGAATGGCTAAGATCAAAAAC
CAGGTGACAACAGATGTTGGCGAGGATGTGGAGAAAGGGGAACACTCCTCCATTGTTGGT
...

>11 dna_sm:chromosome chromosome:GRCm38:11:1:122082543:1 REF
ACATTTTTCACGTTTTTCACTGTTTCTCGCCATATTCCAAGTCTTACAGTGTGTGTTTCT
TATTTTGTAGGTTTTTTCAGTGTTTGTGGCCATATTCTAGGTCTTAACAGGGGAGTTTTT
AATTTTACTCACTGGTCAGTTTTCTTGCCATAATCCTGGTCCTACCATAGGTGTGTCTCA
...
```

ID & Misc.Infos
Sequence
(ctd.)

ID 2
Sequence 2

ID 3
Sequence 3

Annotations: GTF files (aka. GFF2)

Chrom		Feature type	Start	End	Strand	Metadata
▼		▼	▼	▼	▼	▼-----→
1	ensembl	gene	4430189	4450423	. + .	gene_id "ENSACAG000000011126"; gene_name "TMEM1
1	ensembl	transcript	4430189	4450423	. + .	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	exon	4430189	4430804	. + .	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	CDS	4430503	4430804	. + 0	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	start_codon	4430503	4430505	. + 0	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	exon	4439303	4439440	. + .	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	CDS	4439303	4439440	. + 1	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	exon	4443852	4443930	. + .	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	CDS	4443852	4443930	. + 1	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	exon	4445846	4450423	. + .	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	CDS	4445846	4446022	. + 0	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	stop_codon	4446023	4446025	. + 0	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	five_prime_utr	4430189	4430502	. + .	gene_id "ENSACAG000000011126"; transcript_id "E
1	ensembl	three_prime_utr	4446026	4450423	. + .	gene_id "ENSACAG000000011126"; transcript_id "E

- Tab-delimited text files
- GFF3 files are similar, but the metadata part is more cleanly structured.
- Some programs prefer GTF, others GFF3

Alignments files (BAM & SAMtools)

IGV

<https://software.broadinstitute.org/software/igv/>

The screenshot shows the IGV website homepage. The browser window has a Firefox address bar with the URL <https://software.broadinstitute.org/software/igv/>. The page features a sidebar on the left with a navigation menu including Home, Downloads, Documents, IGV User Guide, Tutorial Videos, File Formats, Hosted Genomes, FAQ, Release Notes, Credits, and Contact. Below the menu is a search bar and copyright information for 2013-2018 Broad Institute and the Regents of the University of California. The main content area has a large banner with the text "Integrative Genomics Viewer" and a visual representation of the software's interface showing genomic tracks. Below the banner are two sections: "Overview" and "Citing IGV". The "Overview" section describes IGV as a high-performance visualization tool for interactive exploration of large, integrated genomic datasets, supporting various data types like array-based and next-generation sequence data. It also lists the available forms: the original IGV (Java desktop application), IGV-Web (web application), and igv.js (JavaScript component for embedding). The "Citing IGV" section provides citation information for publications, including a 2011 Nature Biotechnology article and a 2013 Briefings in Bioinformatics article.

Home

Integrative Genomics Viewer

Overview

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

IGV is available in multiple forms, including:

- the original **IGV** - a Java desktop application,
- IGV-Web** - a web application,
- igv.js** - a JavaScript component that can be embedded in web pages (*for developers*)

Citing IGV

To cite your use of IGV in your publication, please reference one or more of:

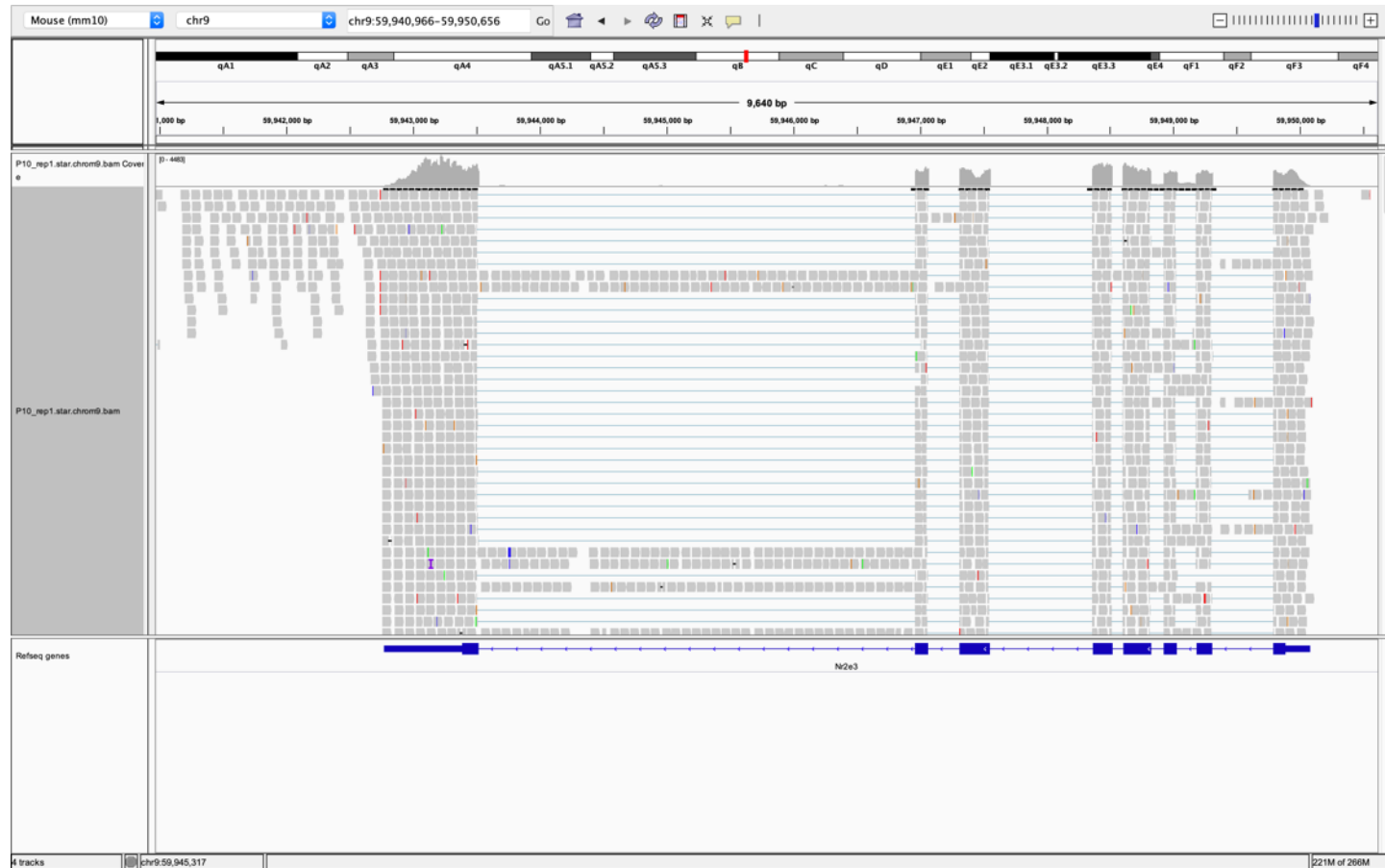
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 \(2011\)](#). (Free PMC article [here](#)).

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. [Integrative Genomics Viewer \(IGV\): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 \(2013\)](#).

James T. Robinson, Helga Thorvaldsdóttir, Aaron M. Wenger,

- Lets you run a genome browser on your laptop

IVG (ctd.)



(Read alignments to the NR2E3 region, for sample P10_rep1)

Read alignments files: the BAM/SAM format

- **BAM** is the Binary equivalent of **SAM** (which is plain text)
In practice: we essentially use BAM files
- ‘Sequence Alignment/Map’ format
<http://samtools.sourceforge.net/SAMv1.pdf>
- Supported by all major alignment packages.
- SAM/BAM files can be manipulated with **SAMtools** (and others)
- **Header** section:
 - Metadata about the genome, the samples, the pipeline
 - Header lines start with **@**
- **Alignments** (or ‘records’) section:

```
6_1303_10584_85775 99    groupVIII    311  3    63M3I34M    =    780  572    GGGTATTGGGC @CFFFFFFHFFH
6_1111_20943_90813 163  groupVIII    315  40    100M        =    809  594    TAATGAAGCCAT @BDDFDDA+<A<
6_2111_2016_88235  355  groupVIII    315  3    100M        =    856  573    TAATGAAGCCAT @?DADDBD>D>B
6_1104_8139_99999  163  groupVIII    316  14    100M        =    818  602    AATGAAGCCATT @@FFFFFFGHGHH
6_1304_4167_91751  163  groupVIII    322  5    52M3I29M    =    812  573    GCCATTTTAC  <<BDBDEHHDF
6_2301_14383_16382 163  groupVIII    323  40    51M3I46M    =    809  589    CCATTTTACT  CCFFFFFFHHHH
```

SAM alignment section

cf. FASTQ format

Read Name	FLAG	Chrom	AlnStart	CIGAR				Sequence	BaseQuals
6_1303_10584_85775	99	groupVIII	311	3	63M3I34M	=	780	572	GGGTATTGGGC @CFFFFFFH
6_1111_20943_90813	163	groupVIII	315	40	100M	=	809	594	TAATGAAGCCAT @BDDFDDA+<A<
6_2111_2016_88235	355	groupVIII	315	3	100M	=	856	573	TAATGAAGCCAT @?DADDBD>D>B
6_1104_8139_99999	163	groupVIII	316	14	100M	=	818	602	AATGAAGCCATT @@FFFFFFGHGHH
6_1304_4167_91751	163	groupVIII	322	5	52M3I29M	=	812	573	GCCATTTTAC <<BDBDEHHDF
6_2301_14383_16382	163	groupVIII	323	40	51M3I46M	=	809	589	CCATTTTACT CCFFFFFFHHH

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2 ¹⁶ -1]	bitwise FLAG
3	RNAME	String	* [!-()+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 ³¹ -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 ⁸ -1]	MAPping Quality
6	CIGAR	String	* ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	* = [!-()+-<>-~] [!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 ²⁹ -1]	Position of the mate/next read
9	TLEN	Int	[-2 ²⁹ +1,2 ²⁹ -1]	observed Template LENgth
10	SEQ	String	* [A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

SAM flags

Description	Base 16	Base 10		Base 2	
template having multiple segments in sequencing	0x1	1	2 ⁰	00000000	00000001
each segment properly aligned according to the aligner	0x2	2	2 ¹	00000000	00000010
segment unmapped	0x4	4	2 ²	00000000	00000100
next segment in the template unmapped	0x8	8	2 ³	00000000	00001000
SEQ being reverse complemented	0x10	16	2 ⁴	00000000	00010000
SEQ of the next segment in the template being reversed	0x20	32	2 ⁵	00000000	00100000
the first segment in the template	0x40	64	2 ⁶	00000000	01000000
the last segment in the template	0x80	128	2 ⁷	00000000	10000000
secondary alignment	0x100	256	2 ⁸	00000001	00000000
not passing quality controls	0x200	512	2 ⁹	00000010	00000000
PCR or optical duplicate	0x400	1024	2 ¹⁰	00000100	00000000
supplementary alignment	0x800	2048	2 ¹¹	00001000	00000000

- SAM flags are displayed in base 10 (decimal) but only have a meaning when converted to binary numbers
- You can decode them with **samtools flags**:
samtools flags 163
0xA3 163 PAIRED,PROPER_PAIR,MREVERSE,READ2

‘CIGAR strings’ (Alignments)

6_1303_10584_85775 99	groupVIII	311	3	63M3I34M	=	780	572	GGGTATTGGGC @CFFFFFFHFH
6_1111_20943_90813 163	groupVIII	315	40	100M	=	809	594	TAATGAAGCCAT @BDDFDDA+<A<
6_2111_2016_88235 355	groupVIII	315	3	100M	=	856	573	TAATGAAGCCAT @?DADDBD>D>B
6_1104_8139_99999 163	groupVIII	316	14	100M	=	818	602	AATGAAGCCATT @@FFFFFFGHGHH
6_1304_4167_91751 163	groupVIII	322	5	52M3I29M	=	812	573	GCCATTTTTTAC <<BDBDEHHDF
6_2301_14383_16382 163	groupVIII	323	40	51M3I46M	=	809	589	CCATTTTTTACT CCFFFFFFHHHH

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

100M — 100 matching nucleotides (i.e. no gaps)

63M-3I-34M — 63 matching nucleotides
 3 nucleotides not in the reference (3bp insertion)
 34 matching nucleotides

Aligned Read

TGCAGGATGGATGTGTTCTCCTCAGCTGCTTATTTTAACTCCAC**TGC**ACAACATGTTTGTGTTATATTCTTTTCGCTGTGTAGTCTGTAAGC

TGCAGGGACTGCAGGATGGATGTGTTCTCCTCAGCTGCTTATTTTAACTCCAC---ACAACATGTTTGTGTTATATTCTTTTCGCTGTGTAGTCTGTAAGCAGAGTATGATACTG

Reference

The SAMtools program

samtools <subprogram> [options]

Major SAMtools subprograms and options:

- **view** [options] <file>
 - Default: converts to 'beheaded' SAM;
 - **-H**: output the SAM header, **-h**: convert to full SAM, **-b**: leave in (or convert to) BAM format
 - **-f/-F** <int>: filter for/against flags matching the integer.
e.g. **-F 0x904** removes all records flagged 0x4 (unmapped), 0x100 (secondary) or 0x800 (supplementary).
 - If file is - (hyphen) STDIN is used
- **sort** <file> — Sorts the alignments within a BAM file (ie. by chromosome and coordinate). If file is - (hyphen) STDIN is used.
- **index** <file> — Creates an index file (**.bai**) for the given (sorted) BAM file
- **faidx** <file> — Creates an index file (**.fai**) for the given FASTA file
- **flags** <integer> — Print the meaning(s) of the given flag.
- **flagstats** <file> — Print statistics on the alignments in the BAM file.

Aligning reads to a genome (using STAR)

- Make sure the commands work. (hypens; updated paths, matching sample names...)
- Provide any new files for the day (eg. trimmed files...)!
- Check file permissions!

Using STAR to align reads

- *Assuming a previously built STAR genome index!*

STAR genome index (database)

- Created from the FASTA & GTF files downloaded from Ensembl
- Commands:

```
module load STAR
```

```
fasta_file=./Mus_musculus.GRCm38.dna_sm.primary_assembly.fa  
gtf_file=./Mus_musculus.GRCm38.98.gtf  
mkdir star_db
```

```
STAR \  
  --runMode genomeGenerate \  
  --genomeFastaFiles $fasta_file \  
  --sjdbGTFfile $gtf_file \  
  --genomeDir ./star_db/
```

- Submitted as a grid job with 92G memory
- ***Already done!!***

The STAR genome database is in:

```
/u/project/scampbel/nrochett/GENOMES/m38p6.M22/e98/star_db
```

Using STAR to align reads

- *The (previously built) STAR genome index is in:*
/u/project/scampbel/nrochett/GENOMES/m38p6.M22/e98/star_db

- On the cluster:

```
module load STAR
module load samtools
# STAR -h | less
```

```
sample=P10_rep1
```

```
star_db_dir=/u/project/scampbel/nrochett/GENOMES/m38p6.M22/e98/star_db
```

```
STAR \
  --genomeDir $star_db_dir \
  --readFilesIn $sample.five_percent.trimmo.fastq.gz \
  --readFilesCommand zcat \
  --outFileNamePrefix $sample. \
  --outSAMtype BAM SortedByCoordinate \
  --outFilterMismatchNmax 5 \
  --outFilterMultimapNmax 1
```

Assessing read alignments !

- **Percentage of reads that successfully aligned:**
typically >80% for human/mouse references
- **Outcome differences between samples** (quality outliers and/or batch effects)
- *If paired-end sequencing:* Percentage of pairs that co-align ('properly paired' statistics)
- Dedicated quality-control tools exist:
 - RSeQC (Wang et al. 2012)
 - QoRTs (Hartley&Mullikin2015)

