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)

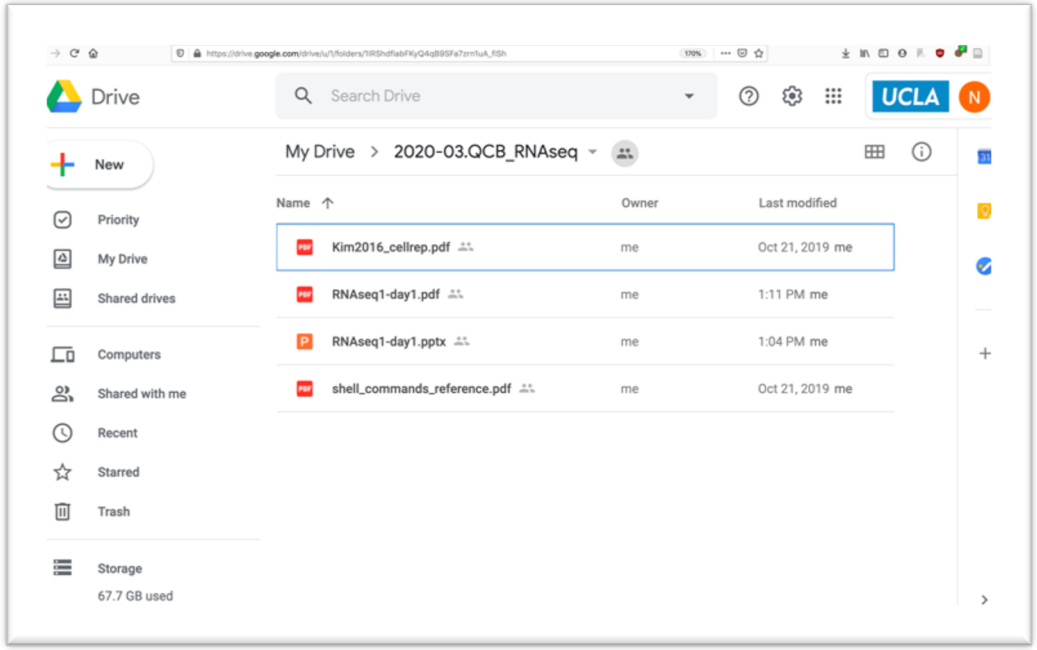
Workshops' google drive folder

Slides & supporting documents:

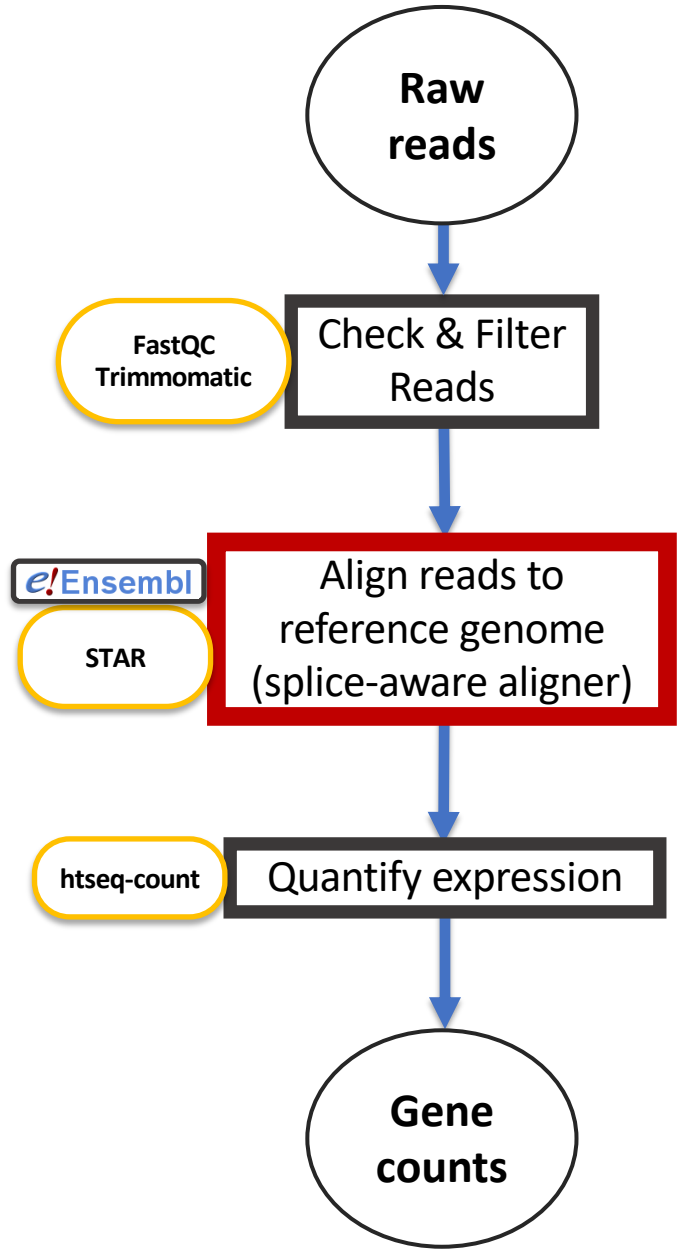
https://drive.google.com/drive/folders/1IRShdfiabFKyQ4qB9SFa7zrn1uA_fiSh

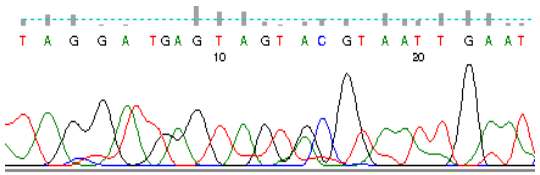
https://tinyurl.com/**sshouvw**

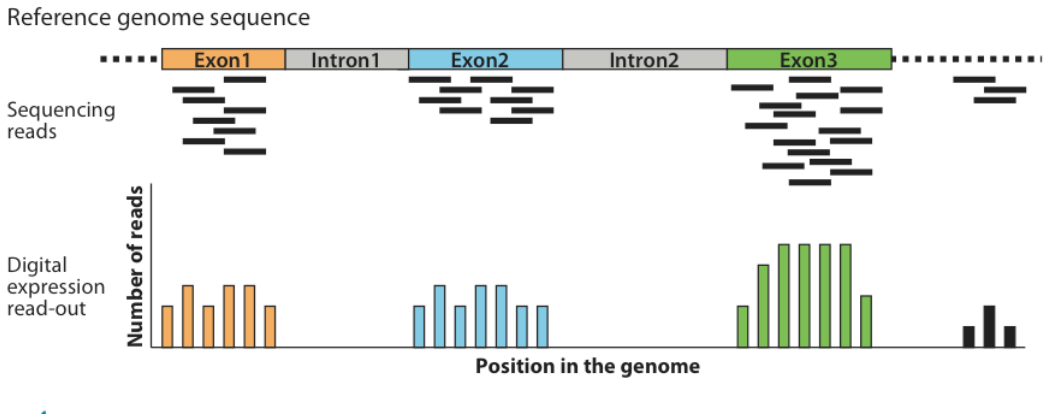
Please bookmark it / write it down!



RNA-seq analysis: Overview







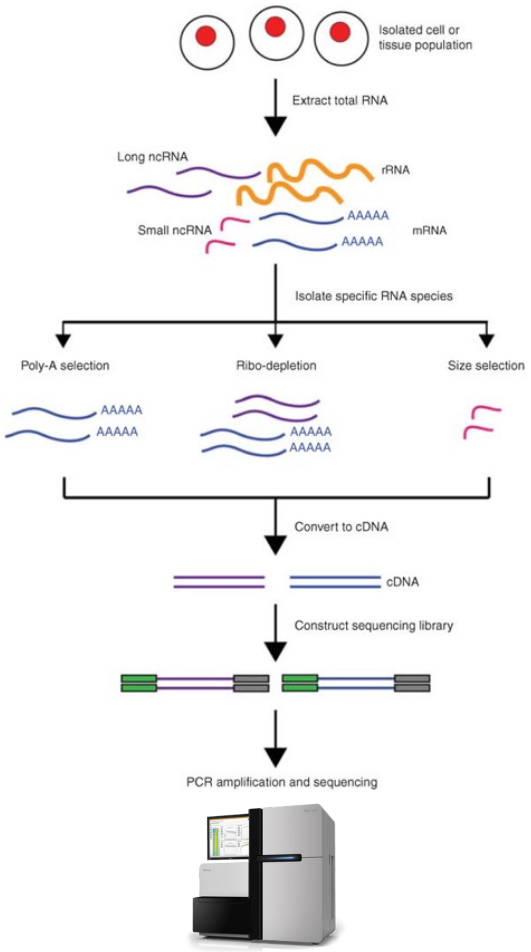
	sample1	sample2	sample3	...
gene1	999	701	616	
gene2	532	520	41	
gene3	14	36	305	
...				

Connecting to the cluster

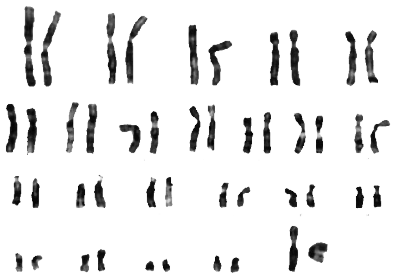
- Remote shell
  ```
  ssh USER@hoffman2.idre.ucla.edu
  ```
  - (through WSL/Putty/Mobaxterm on Windows)
  - Once connected, start an 'interactive job' requesting 4 hours and 32GB memory (STAR needs a lot of memory)
    ```
    qrsh -l h_rt=4:00:00,h_data=32G
    ```

Read alignments: principle

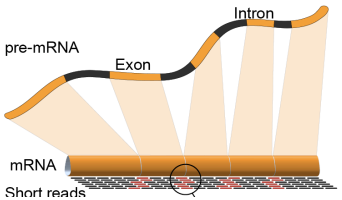
Read alignment / “mapping”



- 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**



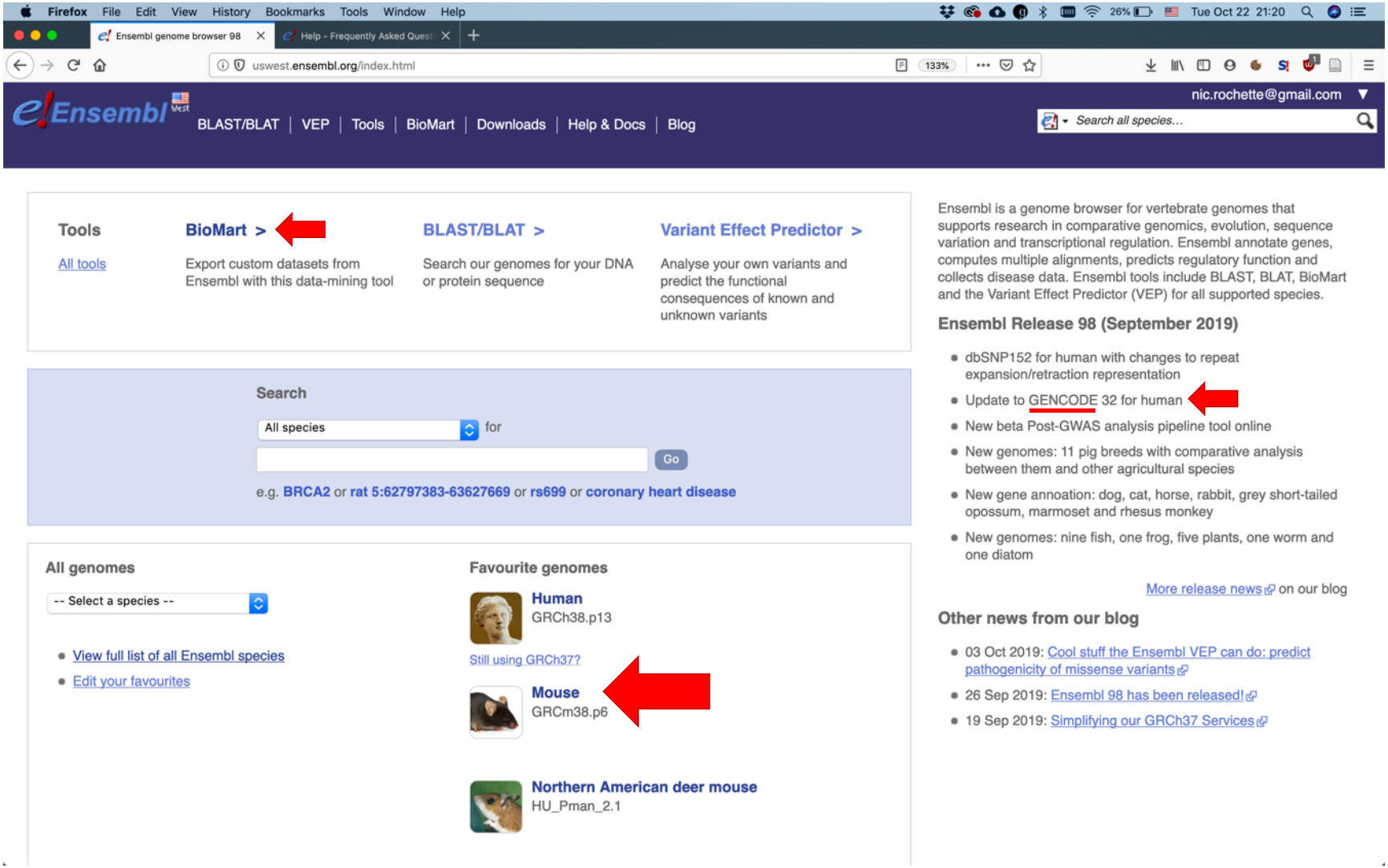
- For RNA data, methods must account for splicing



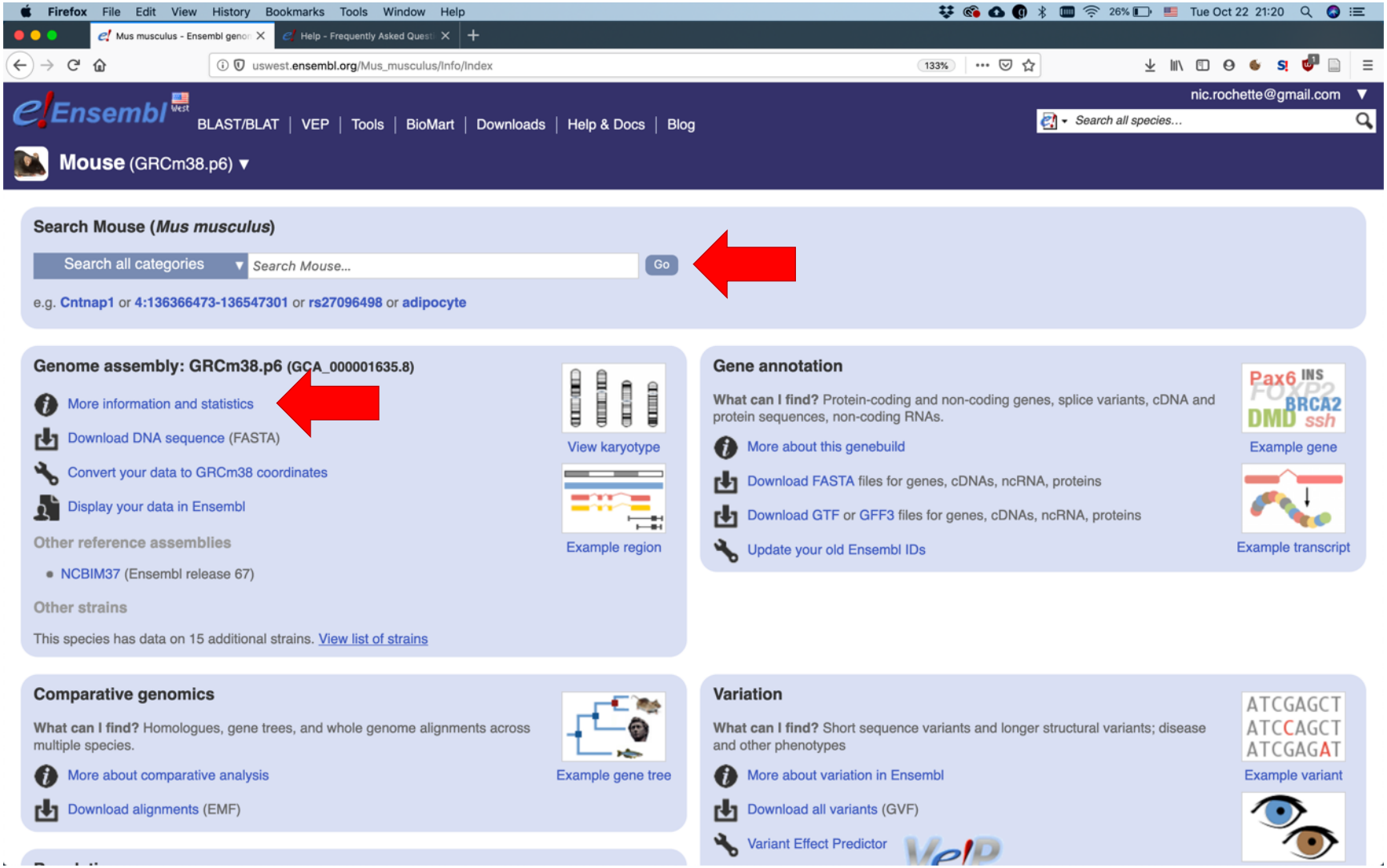
Reference genomes

Ensembl

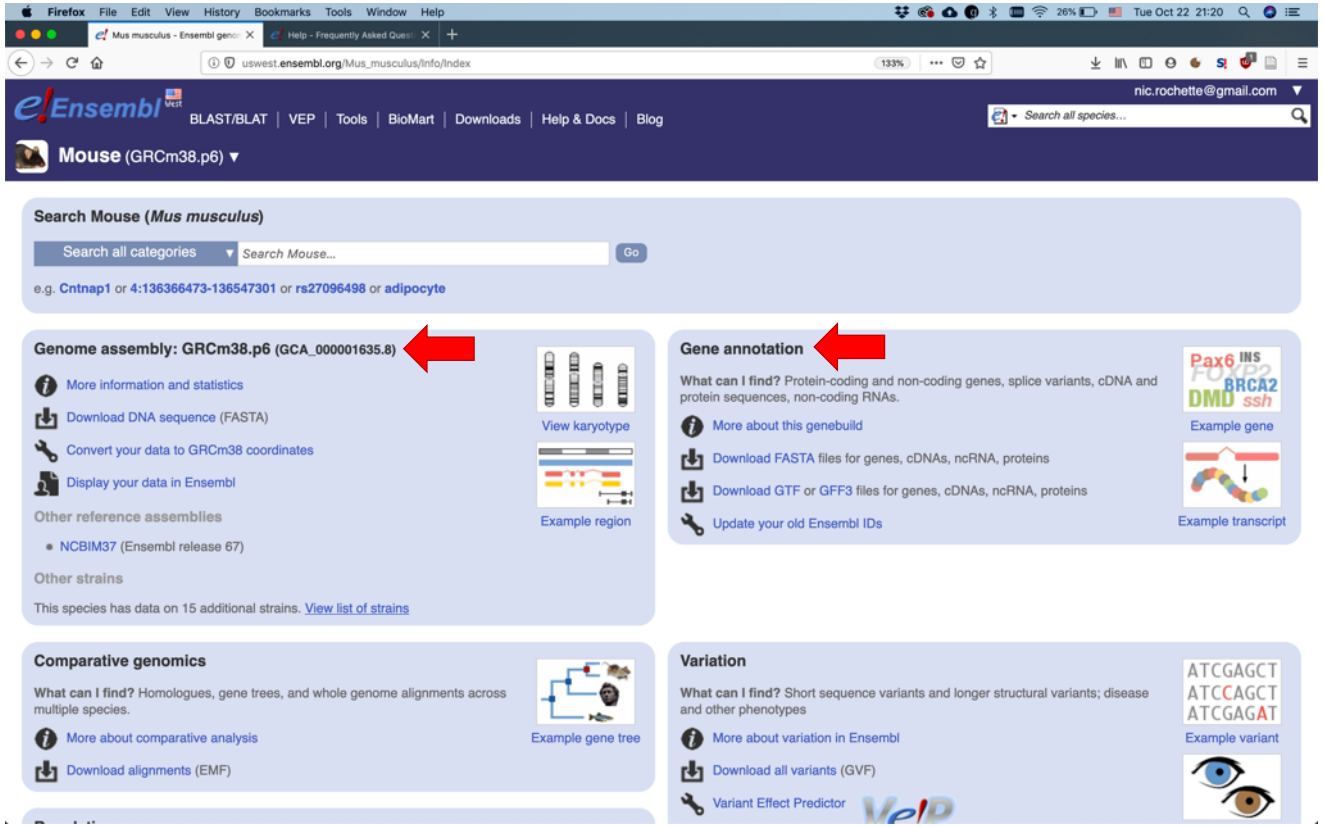
https://www.ensembl.org/



Ensembl (ctd.)


Firefox File Edit View History Bookmarks Tools Window Help
Tue Oct 22 21:20
uswest.ensembl.org/Mus_musculus/Info/Index
nic.rochette@gmail.com
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
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.)



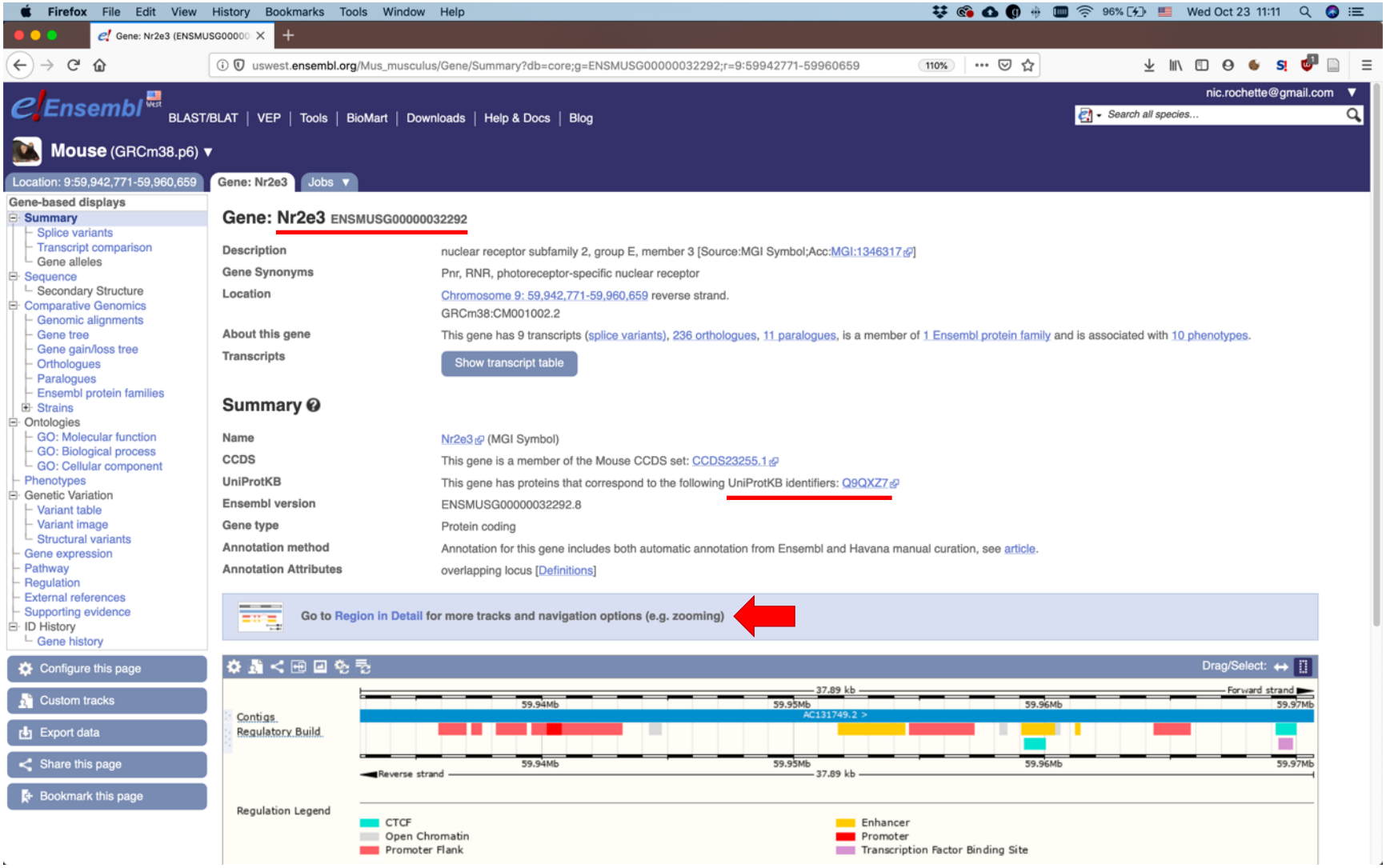
A “reference genome” has two core parts:

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

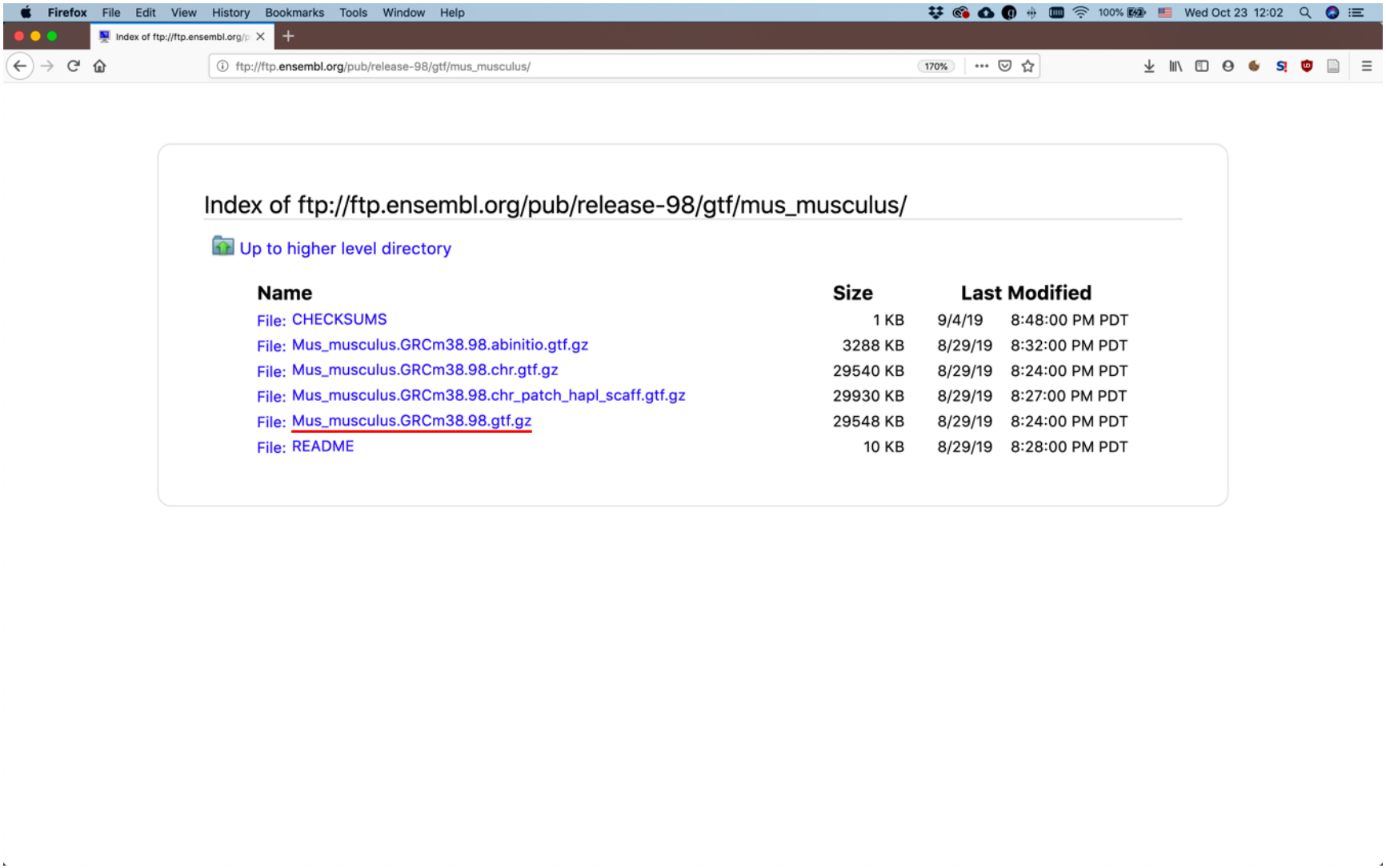
(NB. on the FTP, look for an assembly named `*.dna_sm.primary_assembly.fa.gz`)

Ensembl (ctd.)

- NR2E3 gene (a key target of NRL)



Ensembl FTP


Firefox File Edit View History Bookmarks Tools Window Help
Wed Oct 23 12:02
Index of ftp://ftp.ensembl.org/
ftp://ftp.ensembl.org/pub/release-98/gtf/mus_musculus/
170%
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: Mus_musculus.GRCm38.98.gtf.gz 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
TGGTAGAACTCTGCATTAAACCCATCTGGTCCTGGGCTTTTTTTTTTTTTTTTTTTTTTT
TTTGGGTGGGAGACTATTGATGACTGCCTCTATTTCTTTAGGGGAAATGGGACTTTTAGT
...
>10 dna_sm:chromosome chromosome:GRCm38:10:1:130694993:1 REF
CCGAATGGCAGAGAAACACCTGAATAAAATGTTCAACATCCTTAATCATCAGGGAAATGC
AAATCAAAACAACACTGAGATTCCACTTCACTCCAGTTAGAATGGCTAAGATCAAAAACT
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)

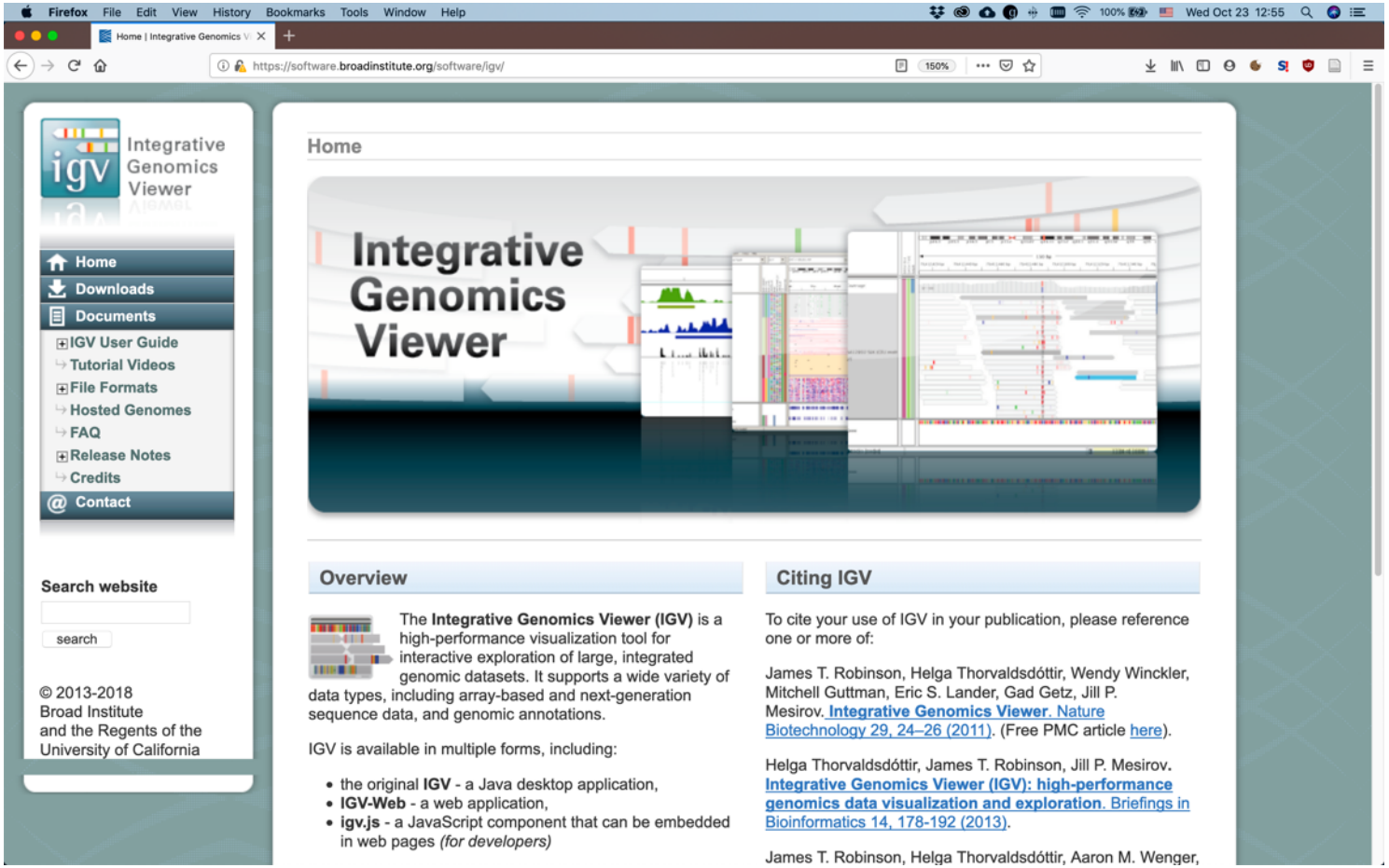
Chrom		Feature type	Start	End	Strand		Metadata
1	ensembl	gene	4430189	4450423	.	+	.	gene_id "ENSACAG00000011126"; gene_name "TMEM1
1	ensembl	transcript	4430189	4450423	.	+	.	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	exon	4430189	4430804	.	+	.	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	CDS	4430503	4430804	.	+	0	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	start_codon	4430503	4430505	.	+	0	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	exon	4439303	4439440	.	+	.	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	CDS	4439303	4439440	.	+	1	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	exon	4443852	4443930	.	+	.	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	CDS	4443852	4443930	.	+	1	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	exon	4445846	4450423	.	+	.	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	CDS	4445846	4446022	.	+	0	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	stop_codon	4446023	4446025	.	+	0	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	five_prime_utr	4430189	4430502	.	+	.	gene_id "ENSACAG00000011126"; transcript_id "E
1	ensembl	three_prime_utr	4446026	4450423	.	+	.	gene_id "ENSACAG00000011126"; 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)

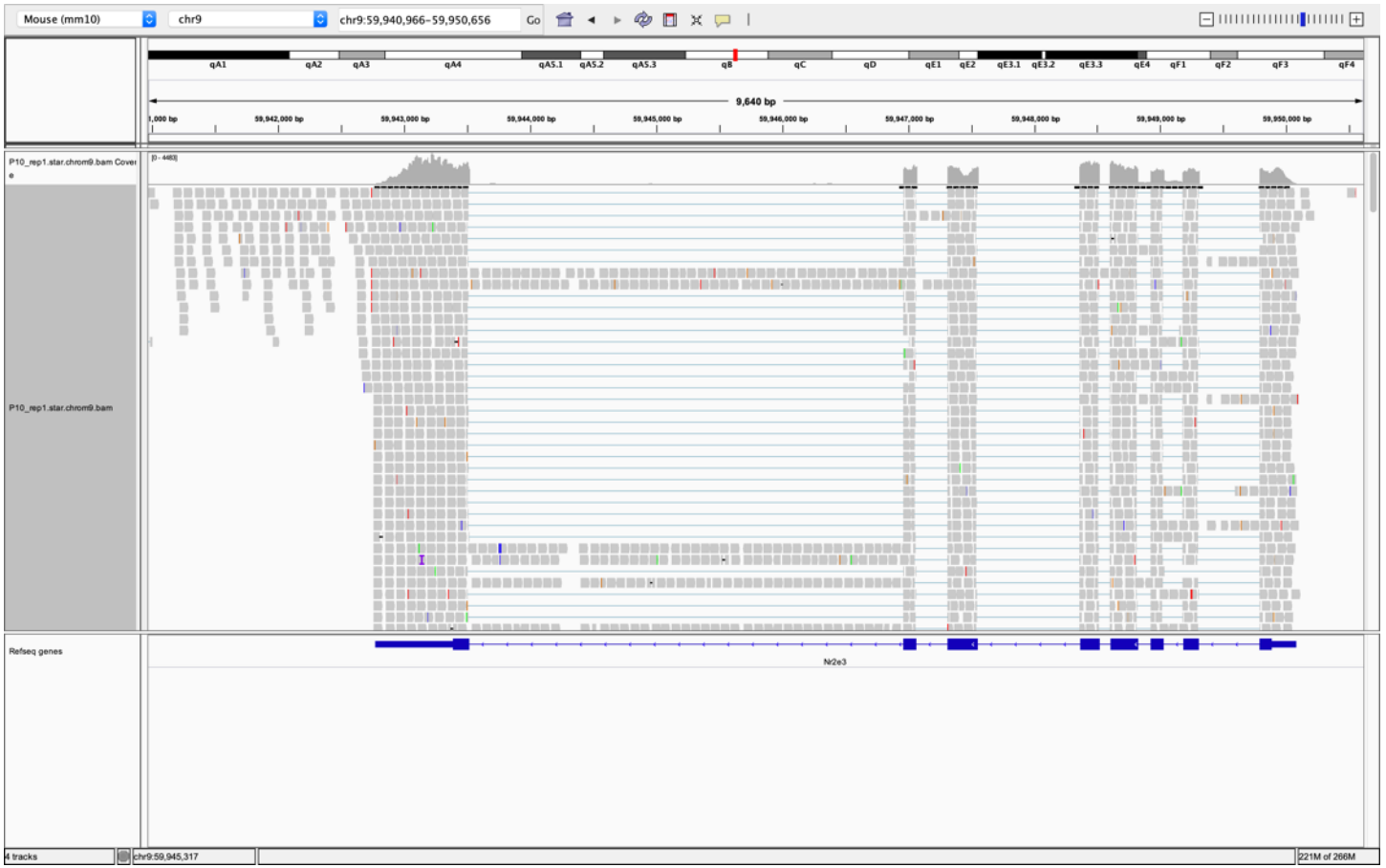
IGV

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



- 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 **B**inary 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 an 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  @CFFFFFHFH
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 @@FFFFFGHGHH
6_1304_4167_91751  163   groupVIII   322  5    52M3I29M   =   812   573   GCCATTTTTAC  <<BDBDEHHDF
6_2301_14383_16382 163   groupVIII   323  40   51M3I46M   =   809   589   CCATTTTTACT  CCFFFFFHHHH
```

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  @CFFFFFHFH
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  @@FFFFFGHGHH
6_1304_4167_91751  163   groupVIII   322  5   52M3I29M  =  812  573  GCCATTTTTAC   <<BDBDEHHDF
6_2301_14383_16382 163   groupVIII   323  40  51M3I46M  =  809  589  CCATTTTTACT   CCFFFFFHHHH
```

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  @CFFFFFHFH
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  @@FFFFFGHGHH
6_1304_4167_91751  163   groupVIII  322  5   52M3I29M  =  812  573  GCCATTTTTAC  <<BDBDEHHDF
6_2301_14383_16382 163   groupVIII  323  40  51M3I46M  =  809  589  CCATTTTTACT  CCFFFFFHHHH
```

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

TGCAGGATGGATGTGTTCCTCCTCAGCTGCTTATTTTAACTCCAC**TGC**ACAACATGTTTTGTGTTATATTCTTTCGCTGTGTAGTCTGTAAGC

TGCAGGGACTGCAGGATGGATGTGTTCCTCCTCAGCTGCTTATTTTAACTCCAC---ACAACATGTTTTGTGTTATATTCTTTCGCTGTGTAGTCTGTAAGCAGAGTATGATACTG

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)