BCB 716 - Sequence Analysis

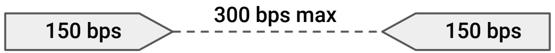
©GABRIEL UTASI 06-12-07

- Problem Set #1 should go out tonight
- Your course logins should work now
- Password is your PID

A look at DNA Alignments

From last time

- **Aligners generate SAM files**
  - An attempt is made to find the closest match for a given read, or read-pair to a reference
  - Alignments are performed independently and in parallel
  - SAM files include
    - the original sequence and quality string from the FASTQ file
    - Initially read pairs are considered together
      - Alignment tolerances
      - Opposite strands
      - Must satisfy a maximum gap distance
    - A placement of the first base that is "normalized" to reference orientation
    - An alignment represented as a CIGAR string
    - Various alignment scores (edit distances, etc.)



- **SAM files are a lot to interpret**
  - Statistic provide a rough idea
  - Localized analysis provides more insights

SAM to BAM

- SAM files tend to be large and difficult to index and manipulate
- Converted into Binary Alignment Maps (BAM files)
- This is done using a toolset called SAMtools
- First to convert a SAM file to a BAM file

```
$ samtools view -S -b CC053.sam -o CC053.bam
$ ls -l CC053.*
-rw-rw-r-- 1 mcmillan its_faculty_psx 5.1G Nov  8 14:57 CC053.bam
-rw-rw-r-- 1 mcmillan its_faculty_psx  24G Nov  8 14:22 CC053.sam
```

- BAM files are smaller, and not simply text, making them easier to search

```
$ samtools view CC053.bam | head -1
A00434:231:H2K7FDSX2:1:1101:10529:1157  99      14      55067154        42
100M    =       55067503        449
GGCTGGAGATGGGGCTGGAGAAGGCGGCTGATCAGGGCTTTCTGAGGGCTCCCTGGAGCCCTCGACTGGCGCCAGGGAAGG
CTCAAGAGGAGGATCTGGG
FFFFFFF:FFFFFFF:FFFFFFFFFF:FFFFFFFFFFFF:FFFFFF:FFFF:FFFFFFFFFFFFFFFFFFFFFFFFFF:FF
FFFFFFFFFFFFFFFFFFF AS:i:-5 XN:i:0  XM:i:1  XO:i:0       XG:i:0  NM:i:1
MD:Z:77G22      YS:i:-4 YT:Z:CP
```

Sorted and Indexed BAMs

- The reads in a BAM file are roughly in the order they can out of the sequencer
- SAM tools provides a tool to sort the reads genomically

```
$ samtools sort CC053.bam -o CC053.sorted.bam
$ ls -l CC053.*
-rw-rw-r-- 1 mcmillan its_faculty_psx 5.1G Nov  8 14:57 CC053.bam
-rw-rw-r-- 1 mcmillan its_faculty_psx  24G Nov  8 14:22 CC053.sam
-rw-rw-r-- 1 mcmillan its_faculty_psx 3.0G Nov  8 15:11 CC053.sorted.bam
```

- BAM files are even smaller, nearby sequences overlap and compress better
- Last of all we build an index so that the BAM file is easier to search/load

```
$ samtools index CC053.sorted.bam
$ ls -l CC053.*
-rw-rw-r-- 1 mcmillan its_faculty_psx 5.1G Nov  8 14:57 CC053.bam
-rw-rw-r-- 1 mcmillan its_faculty_psx  24G Nov  8 14:22 CC053.sam
-rw-rw-r-- 1 mcmillan its_faculty_psx 3.0G Nov  8 15:11 CC053.sorted.bam
-rw-rw-r-- 1 mcmillan its_faculty_psx 3.0M Nov  8 15:18 CC053.sorted.bam.bai
```

Exercise

Go to the following website:

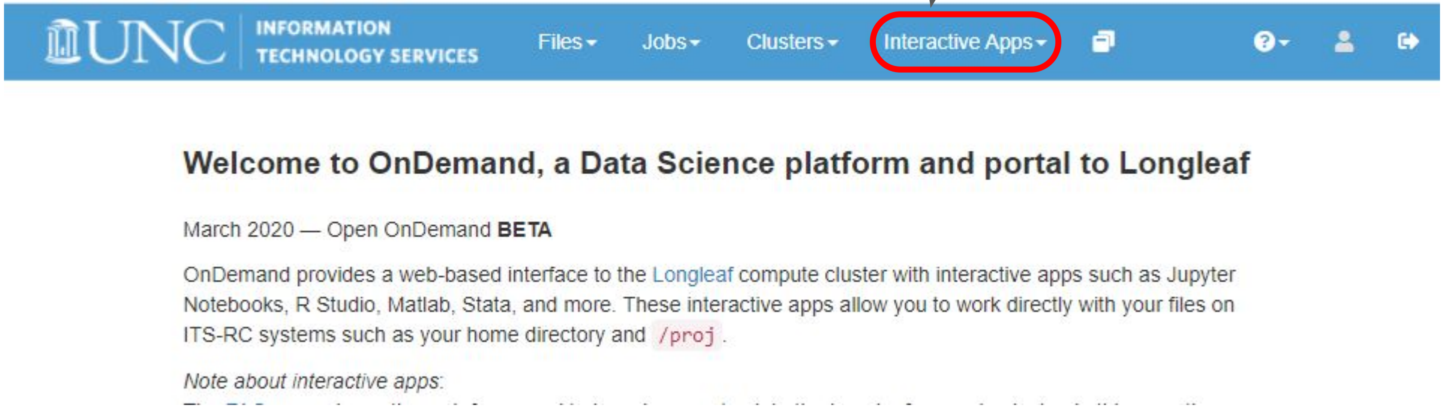
https://ondemand.rc.unc.edu

You will need to authenticate with your ONYEN

Eventually you will get here:

Click here and pick:

🖵 longleaf Desktop



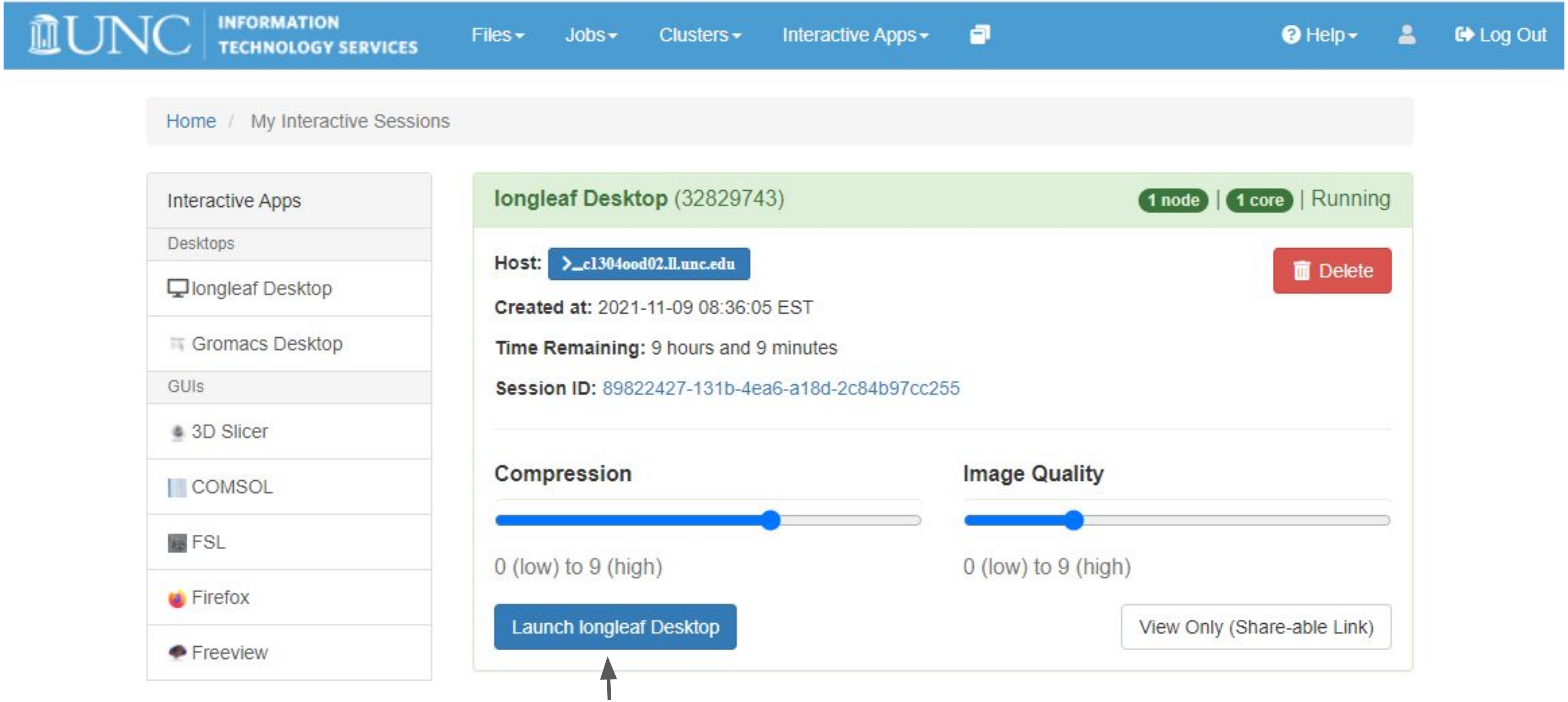
Welcome to OnDemand, a Data Science platform and portal to Longleaf

March 2020 — Open OnDemand **BETA**

OnDemand provides a web-based interface to the Longleaf compute cluster with interactive apps such as Jupyter Notebooks, R Studio, Matlab, Stata, and more. These interactive apps allow you to work directly with your files on ITS-RC systems such as your home directory and `/proj` .

Note about interactive apps:

Wait here for a few seconds



longleaf Desktop (32829743) | Running

Host: >_c1304ood02.ll.unc.edu

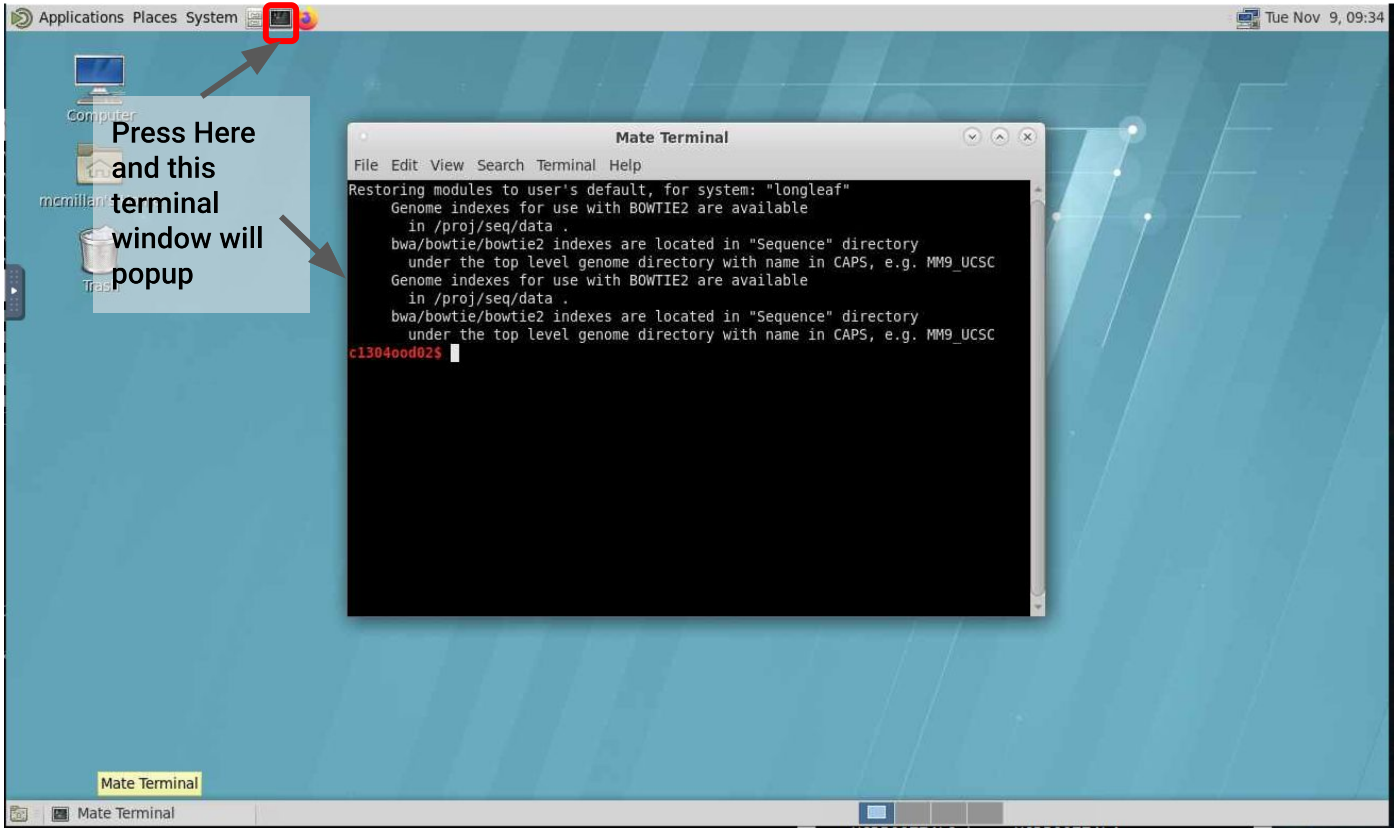
Created at: 2021-11-09 08:36:05 EST

Time Remaining: 9 hours and 9 minutes

Session ID: 89822427-131b-4ea6-a18d-2c84b97cc255

Wait for this button to appear.
Then press it

Eventually you'll get here



Press Here and this terminal window will popup

```
Restoring modules to user's default, for system: "longleaf"
     Genome indexes for use with BOWTIE2 are available
       in /proj/seq/data .
     bwa/bowtie/bowtie2 indexes are located in "Sequence" directory
       under the top level genome directory with name in CAPS, e.g. MM9_UCSC
     Genome indexes for use with BOWTIE2 are available
       in /proj/seq/data .
     bwa/bowtie/bowtie2 indexes are located in "Sequence" directory
       under the top level genome directory with name in CAPS, e.g. MM9_UCSC
c1304ood02$ 
```

Now type a few commands at the command line

- Install an initial set of bioinformatic modules:

```
$ cp /proj/mcmillanlab/BCB716F21/loadModules .
$ cat loadModules
$ module list


Currently Loaded Modules:
1) samtools/1.9    3) bowtie2/2.4.1    5) minimap2/2.17
2) bwa-mem2/2.2.1  4) igv/2.8.7
```

- Today we'll discuss IGV

Integrative Genomics Viewer (IGV)

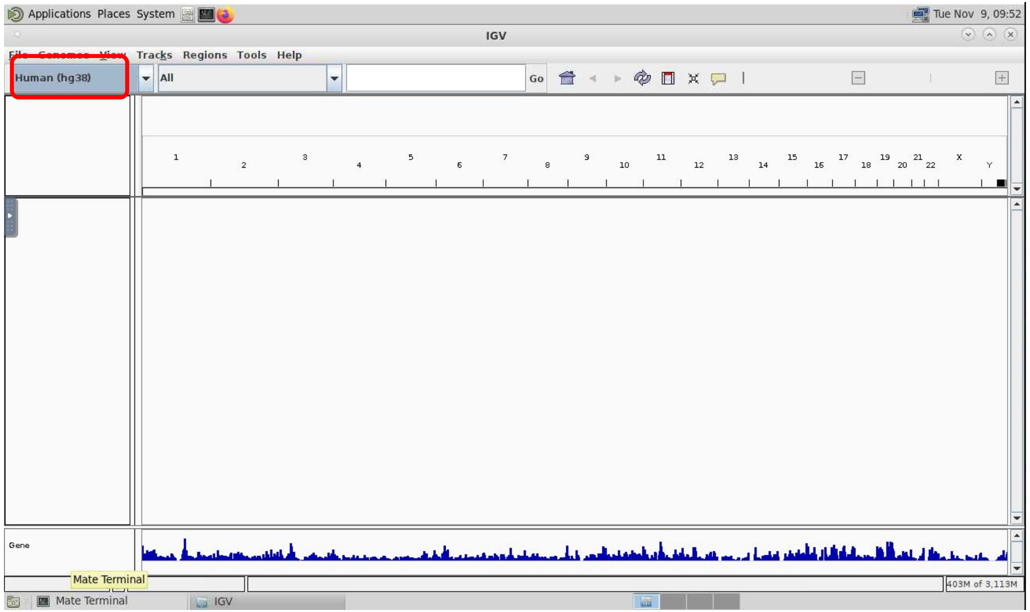
- **I typed:**

```
$ igv &      # starts the viewer as a background process
```

- **After some machinations, and maximizing**

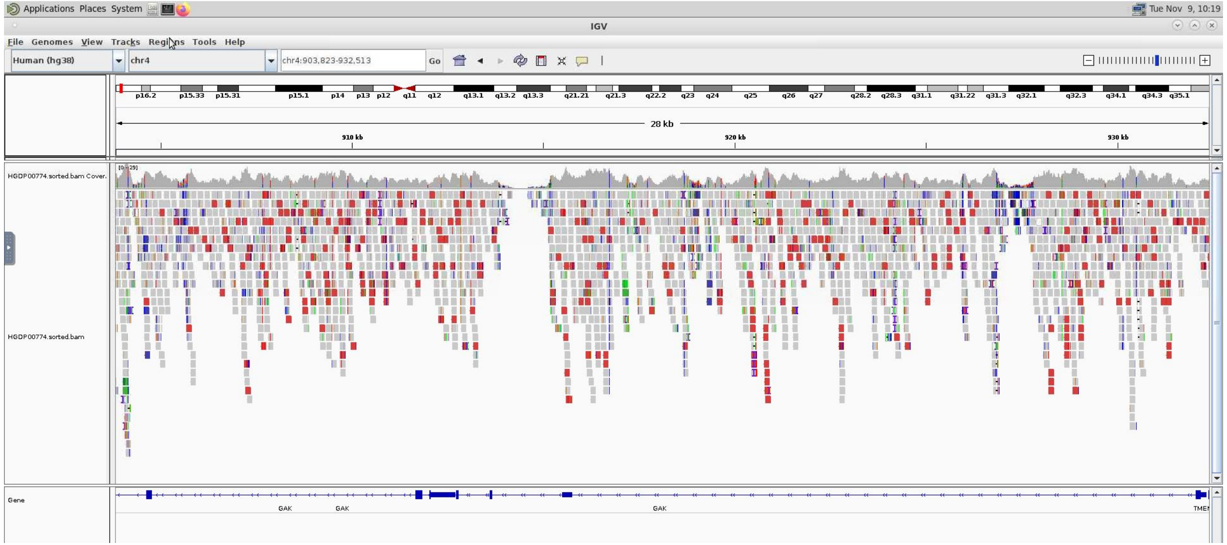
First, you'll need to make sure you are using the correct genome.

I'll use Human (hg38)


Applications Places System
Tue Nov 9, 09:52
IGV
File Genomes View Tracks Regions Tools Help
Human (hg38)
All
Go
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Gene
Mate Terminal
403M of 3,113M
Mate Terminal
IGV

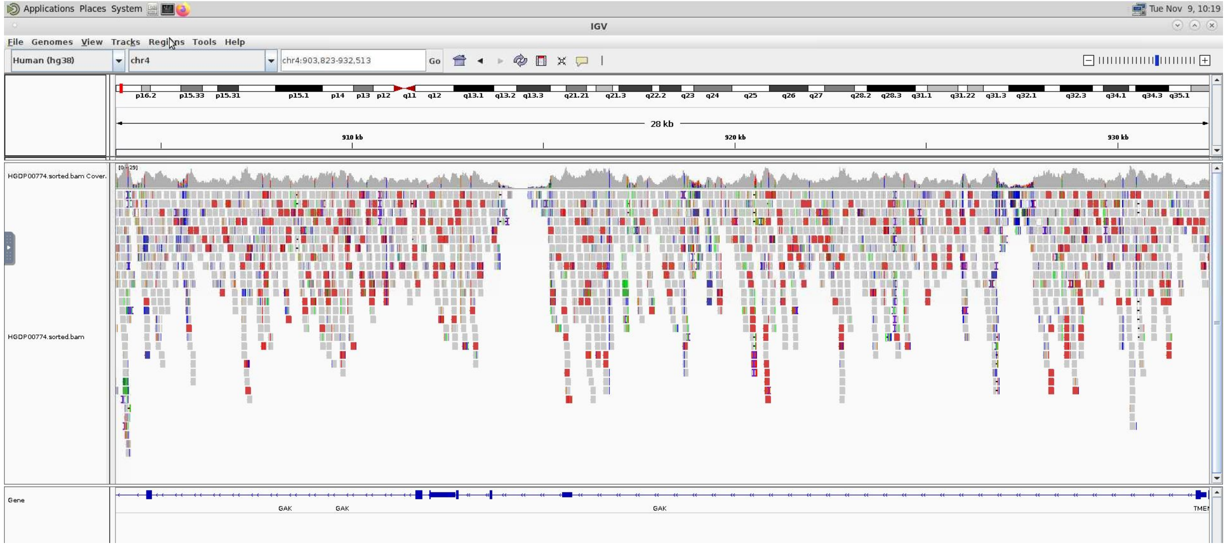
Visualizing BAM files

- The Interactive Genome Viewer (IGV) is a standard tool for visualizing sorted BAM files with index files
- You won't see any reads until you get to a window smaller than 30 kb (configurable, but)
- Coverage above
- Alignments below

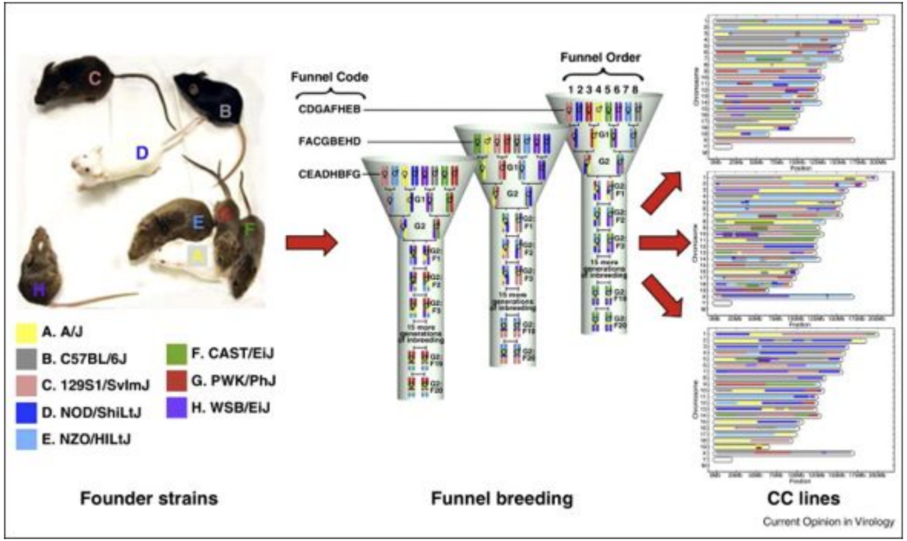

Applications Places System
Tue Nov 9, 10:19
IGV
File Genomes View Tracks Regions Tools Help
Human (hg38)
chr4
chr4:903,823-932,513
Go
p16.2 p15.33 p15.31 p15.1 p14 p13 p12 q11 q12 q13.1 q13.2 q13.3 q21.21 q21.3 q22.2 q23 q24 q25 q26 q27 q28.2 q28.3 q31.1 q31.22 q31.3 q32.1 q32.3 q34.1 q34.3 q35.1
28 kb
910 kb
920 kb
930 kb
[0 - 29]
HGDP00774.sorted.bam Cover.
HGDP00774.sorted.bam
Gene
GAK
GAK
GAK
TMEM

Visualizing BAM files

- The reads are labelled with variants and INDELS that differ from the reference
- Red reads are separated from mates by a larger gap than expected

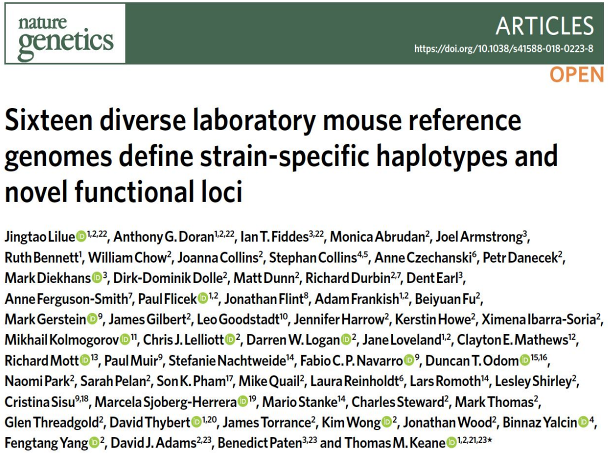

Applications Places System
Tue Nov 9, 10:19
IGV
File Genomes View Tracks Regions Tools Help
Human (hg38)
chr4
chr4:903,823-932,513
Go
p16.2 p15.33 p15.31 p15.1 p14 p13 p12 q11 q12 q13.1 q13.2 q13.3 q21.21 q21.3 q22.2 q23 q24 q25 q26 q27 q28.2 q28.3 q31.1 q31.22 q31.3 q32.1 q32.3 q34.1 q34.3 q35.1
28 kb
910 kb
920 kb
930 kb
[0 - 29]
HGDP00774.sorted.bam Cover.
HGDP00774.sorted.bam
Gene
GAK
GAK
GAK
TMEI

A Genome Model for a Population



Srivastava, Anuj, et al. "Genomes of the mouse collaborative cross." *Genetics* 206.2 (2017): 537-556.

Lilue, Jingtao, et al. "Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci." *Nature genetics* 50.11 (2018): 1574.



https://doi.org/10.1038/s41588-018-0223-8

OPEN

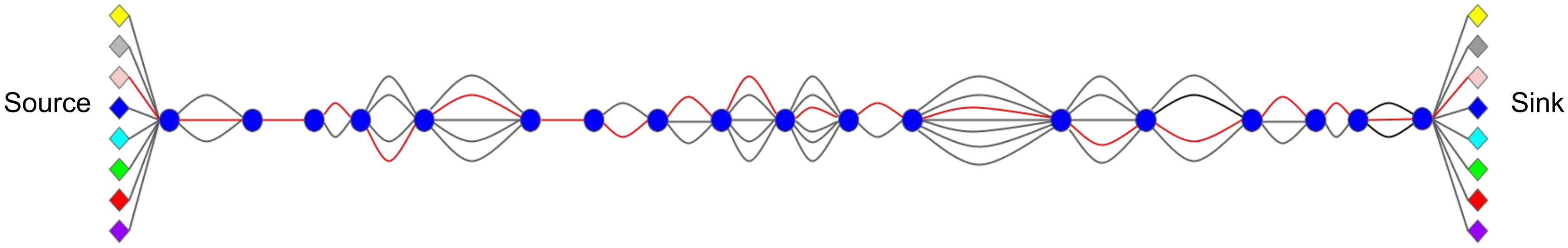
Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci

Jingtao Lilue[1,2,22], Anthony G. Doran[1,2,22], Ian T. Fiddes[3,22], Monica Abrudan[2], Joel Armstrong[3], Ruth Bennett[1], William Chow[2], Joanna Collins[2], Stephan Collins[4,5], Anne Czechanski[6], Petr Danecek[2], Mark Diekhans[3], Dirk-Dominik Dolle[2], Matt Dunn[2], Richard Durbin[2,7], Dent Earl[3], Anne Ferguson-Smith[7], Paul Flicek[1,2], Jonathan Flint[8], Adam Frankish[1,2], Beiyuan Fu[2], Mark Gerstein[9], James Gilbert[2], Leo Goodstadt[10], Jennifer Harrow[2], Kerstin Howe[2], Ximena Ibarra-Soria[2], Mikhail Kolmogorov[11], Chris J. Lelliott[2], Darren W. Logan[2], Jane Loveland[1,2], Clayton E. Mathews[12], Richard Mott[13], Paul Muir[9], Stefanie Nachtweide[14], Fabio C. P. Navarro[9], Duncan T. Odom[15,16], Naomi Park[2], Sarah Pelan[2], Son K. Pham[17], Mike Quail[2], Laura Reinholdt[6], Lars Romoth[14], Lesley Shirley[2], Cristina Sisu[9,18], Marcela Sjoberg-Herrera[19], Mario Stanke[14], Charles Steward[2], Mark Thomas[2], Glen Threadgold[2], David Thybert[1,20], James Torrance[2], Kim Wong[2], Jonathan Wood[2], Binnaz Yalcin[4], Fengtang Yang[2], David J. Adams[2,23], Benedict Paten[3,23] and Thomas M. Keane[1,2,21,23*]

- The Collaborative Cross (CC) mouse genetic reference population that is widely used to study complex traits
- Each genome is a mosaic of 8 inbred founder strains

- Mouse reference genome GRCm38 is based on C57BL/6J strain
- Genomes of the other 7 founder strains were recently released

Collaborative Cross Graphical Genome (CCGG)



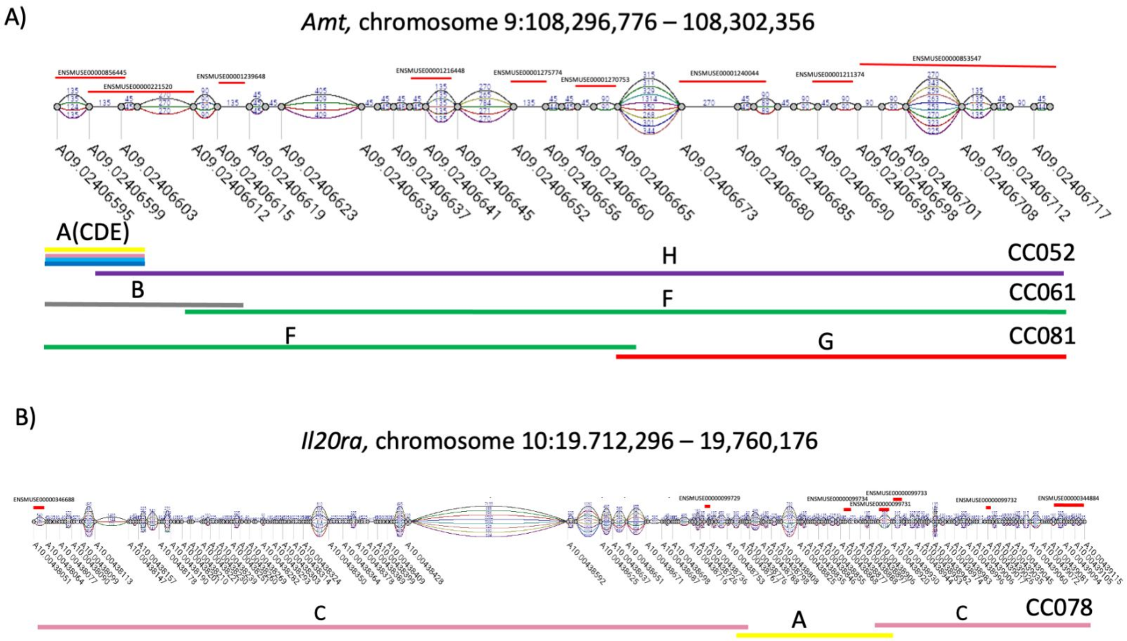
Input:

Genome assemblies of the 8 founder strains (Lilue, Jingtao, et al. 2018)
75 CC sequenced samples using 30x Illumina short-reads (Srivastava et al., 2017, Shorter et al. 2019)

Output:

A directed series-parallel graph with "*anchor*" **nodes** containing unique sequences present in every sample from the population, and **edges** representing haplotype diversity.

Functional Regions and Comparative Analysis

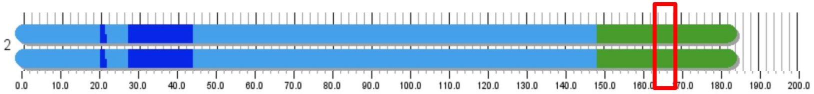


CC Recombination within Genes and Exons

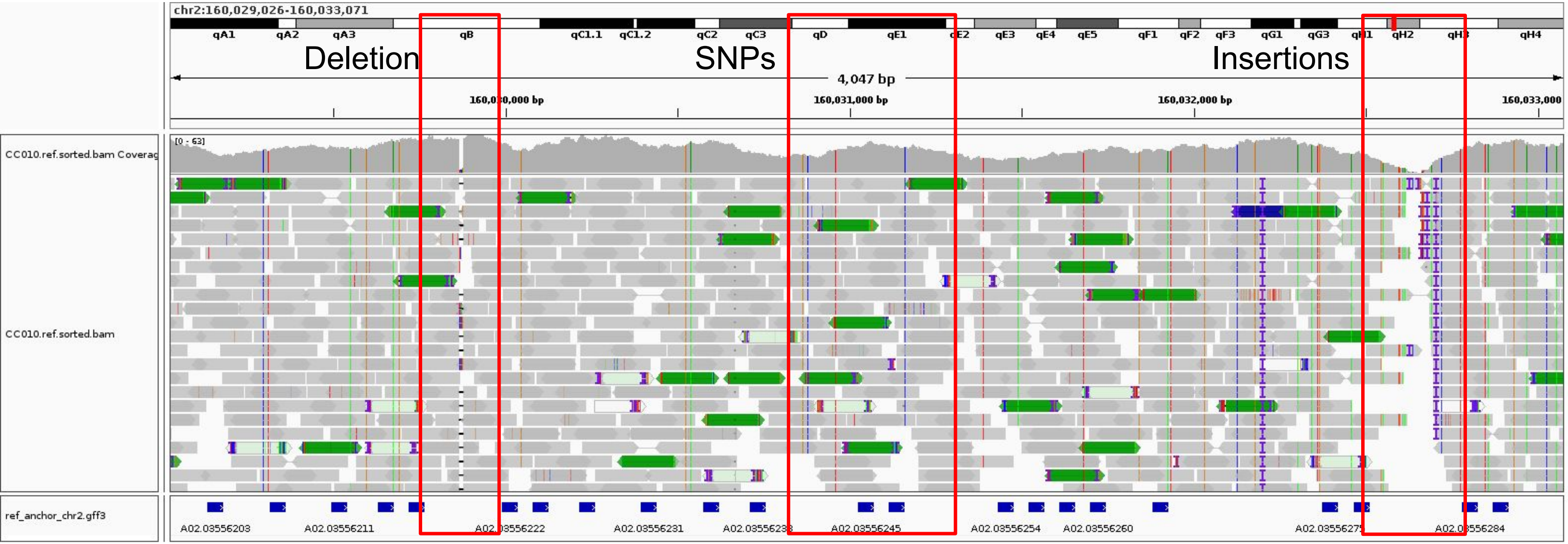
- Multiple CC strains have recombinations in a gene
- There are double recombinations in genes

Advantages of a Strain-Specific Genome

Align CC010 MRCA sequence data to GRCm38 and CC010 linear genome respectively

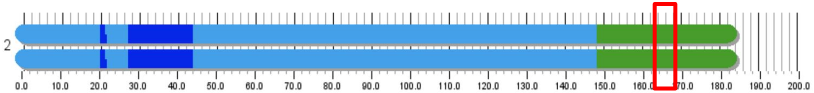


Aligned to reference GRCm38: Chr2:160.03Mb

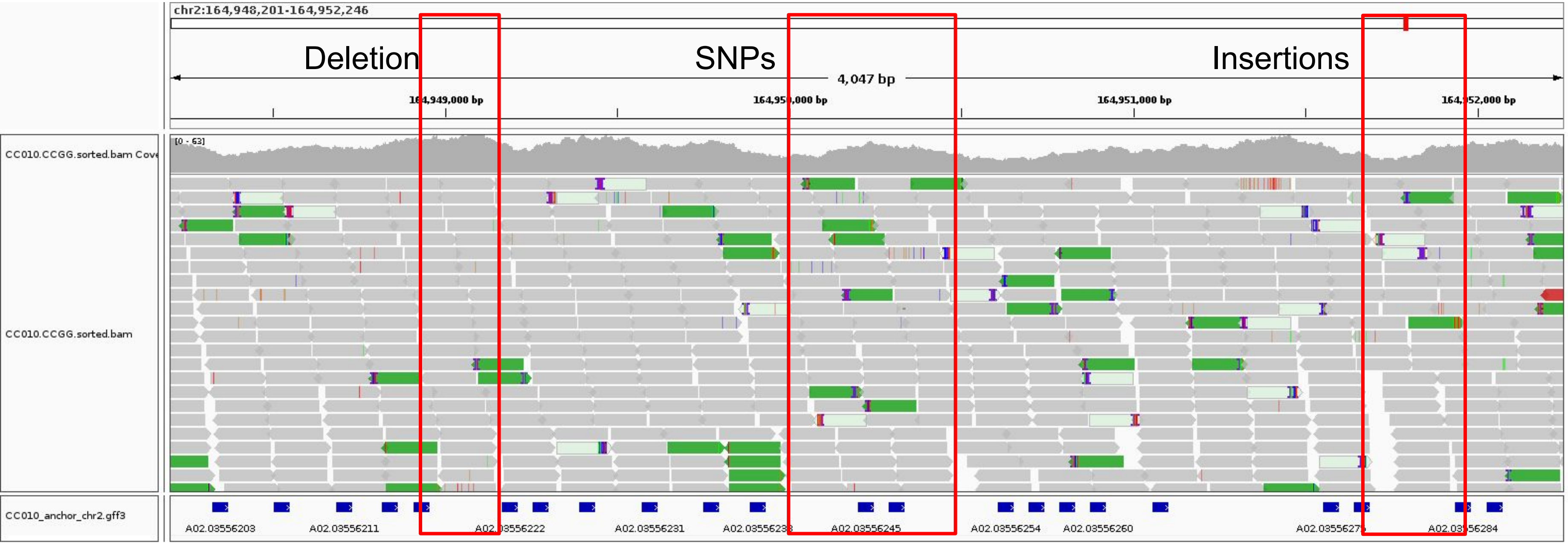


Advantages of a Strain-Specific Genome

Align CC010 MRCA sequence data to GRCm38 and CC010 linear genome respectively



Align to CC010 genome: Chr2:164.95Mb



Deletion

SNPs

Insertions

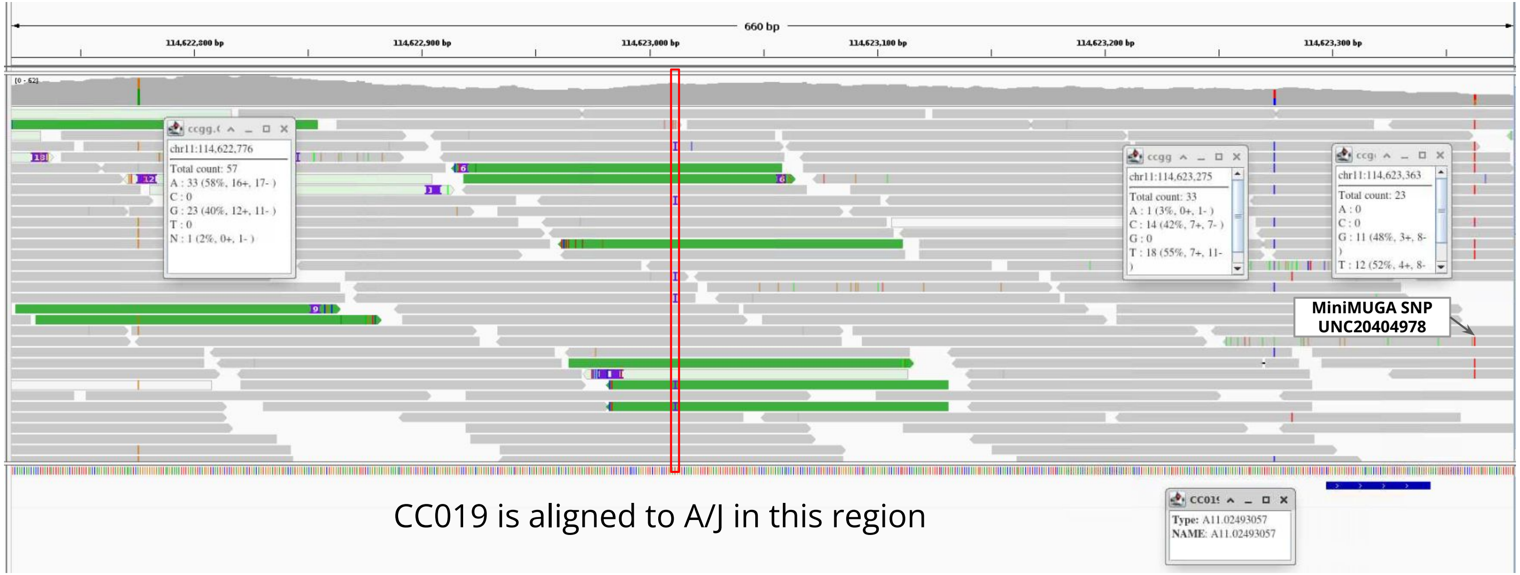
Deleti

CCGG to estimate Residual Heterozygosity

- MRCAs will be used to estimate allele frequencies in the segregating regions
- On 11:112M-114.5M (GRCm38) CC019 is segregating between A/J and WSB/EiJ

6 MRCAs
(12 chromo)
4 variants

A/J	WSB
33A	23G
34	11I
18T	14C
11G	12T
58%	42%



chr11:114,622,776
Total count: 57
A : 33 (58%, 16+, 17-)
C : 0
G : 23 (40%, 12+, 11-)
T : 0
N : 1 (2%, 0+, 1-)

chr11:114,623,275
Total count: 33
A : 1 (3%, 0+, 1-)
C : 14 (42%, 7+, 7-)
G : 0
T : 18 (55%, 7+, 11-
)

chr11:114,623,363
Total count: 23
A : 0
C : 0
G : 11 (48%, 3+, 8-
)
T : 12 (52%, 4+, 8-

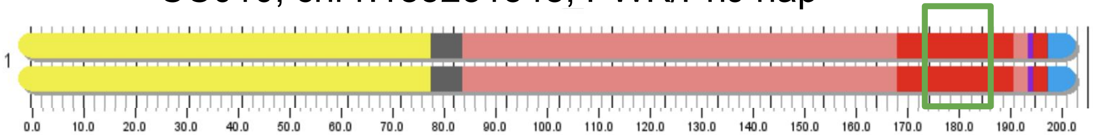
Type: A11.02493057
NAME: A11.02493057

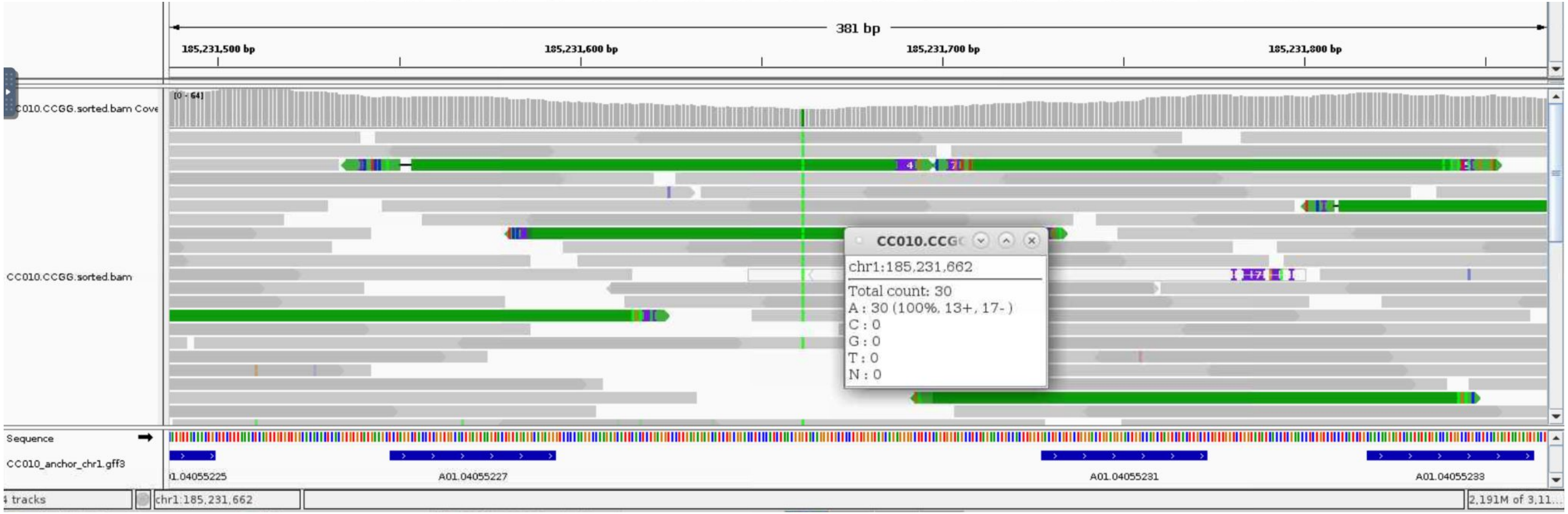
CC019 is aligned to A/J in this region

Private (de novo) CC Mutations

SNPs

CC010, chr1:185231548, PWK/PhJ hap





CC010.CCG

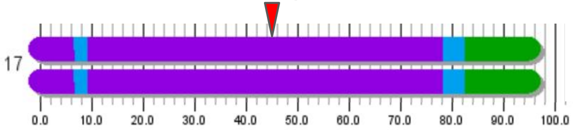
chr1:185,231,662

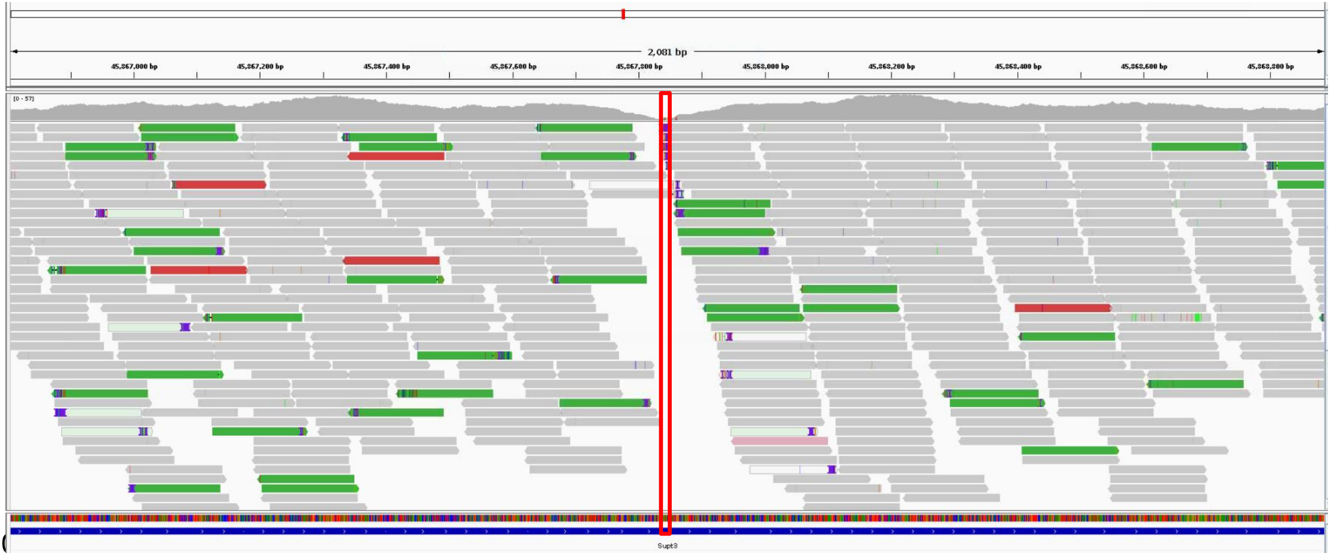
Total count: 30
A : 30 (100%, 13+, 17-)
C : 0
G : 0
T : 0
N : 0

Private (de novo) CC Mutations

CC019 has a ~6800 base TE insertion at chr17:45,010,174, on a WSB/EiJ background
We now have a catalog of over 10,000 non-reference SVs (Kashfeen's IMGC talk)

Structural Variants

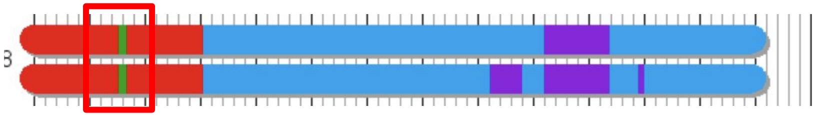


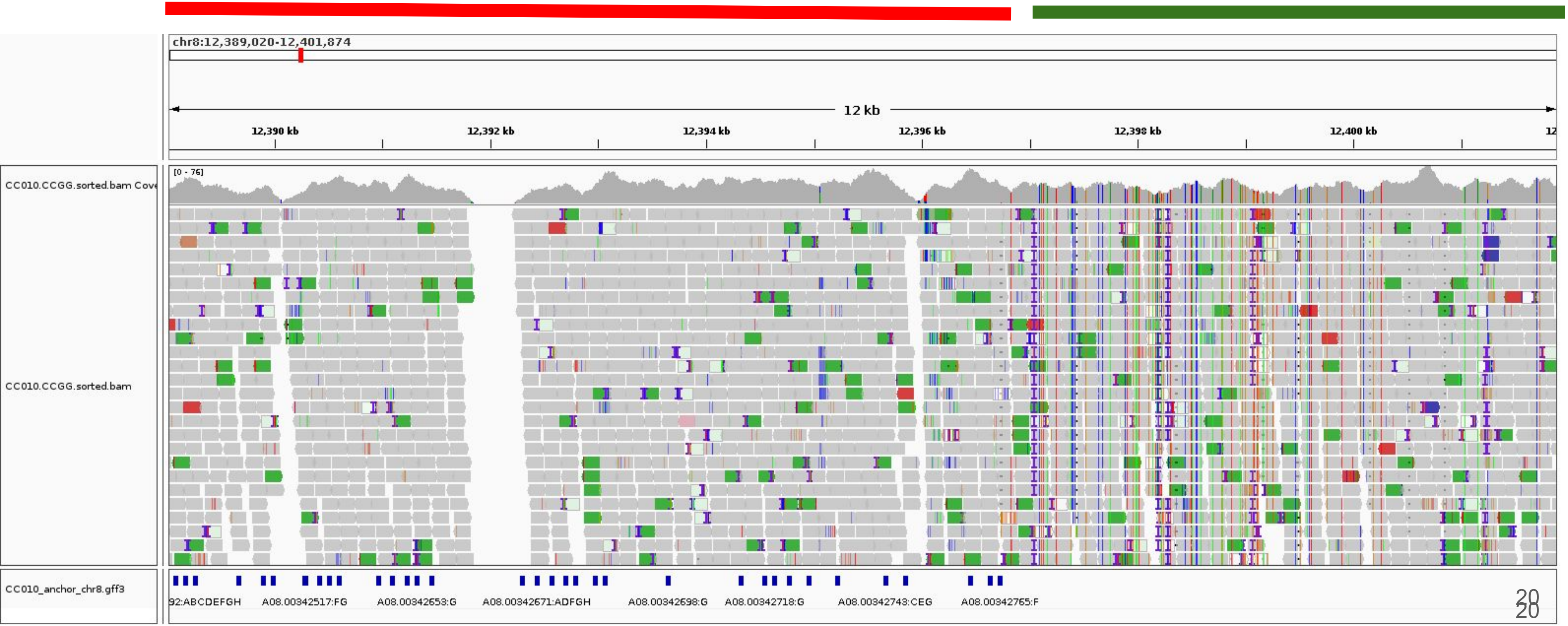


Recombination Boundary

chr8: 15.4Mb (GRCm38), Transition from G to F

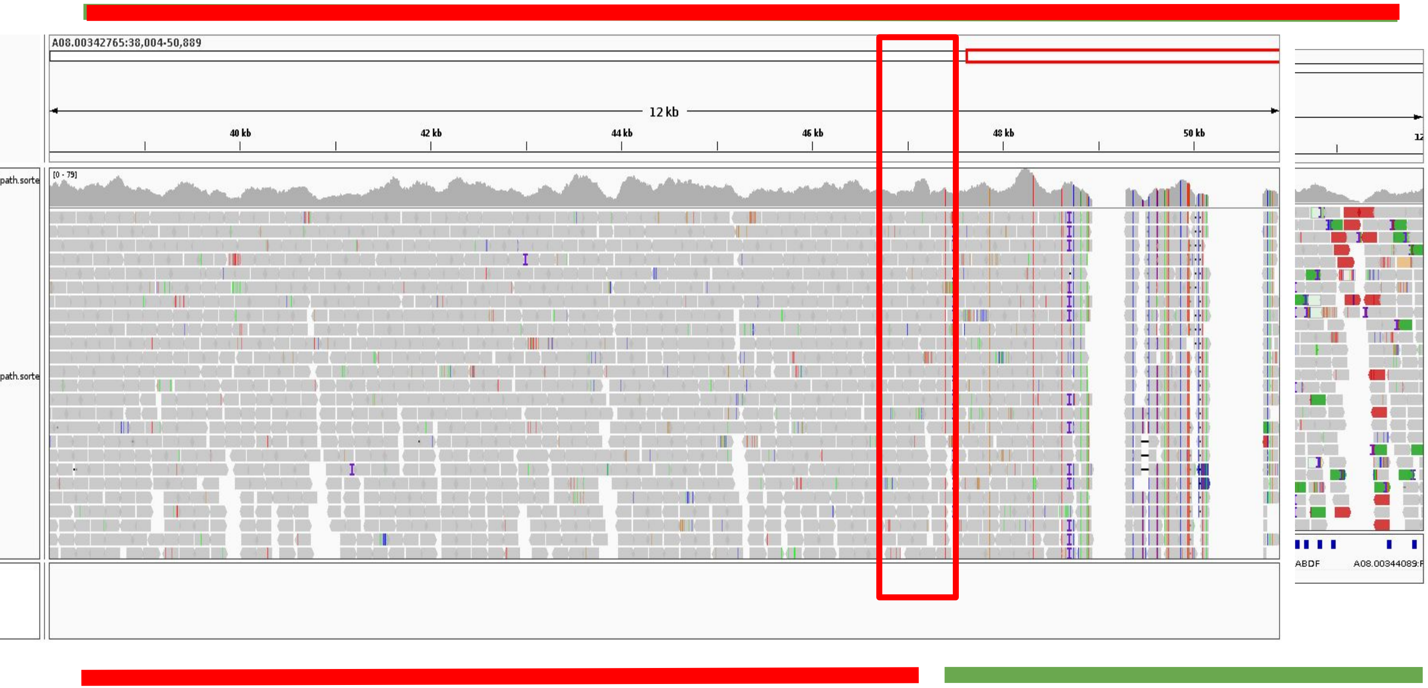
A08.00342765 - A08.00344047, 54.7kb long gap



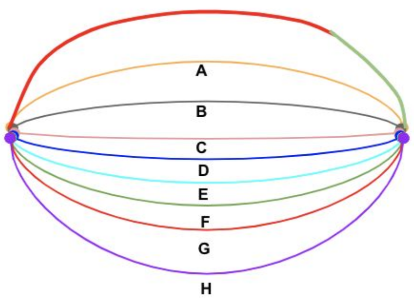


20

Recombination Boundary

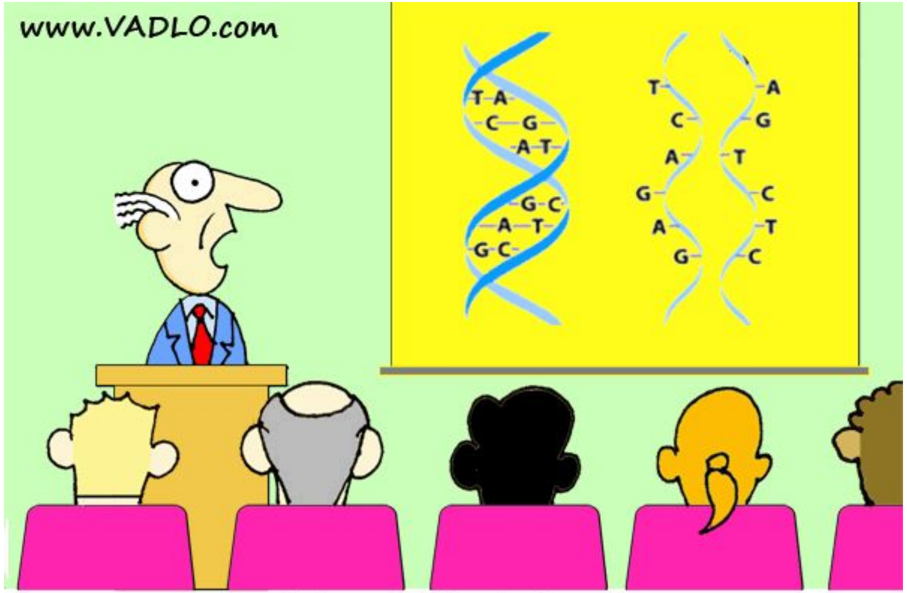


CC010 Recombinant Edge



Next Time

Visualizing, Interpreting, and Analyzing Alignment outputs



“Then Crick asked me,
‘Watson, should we name them A,C,T,G or A,B,C,D?”