Introduction to NGS Analysis (Day 2) Session IV

Lyda Hill department of Bioinformatics 2022 Nanocourse Series

Date & Time: May 19-20: 9AM-5PM (NG3.202)

Course Instructors: Bo Li, Daehwan Kim, Christopher Chaney, & **Micah Thornton**

UT Southwestern Medical Center

Day 2: Session IV (Using IGV, and Genotype Information to Assess Phenotype in Human Genomic Data)

UT Southwestern
Medical Center

What you will learn in this Session (2 Parts)

- *What is the Integrative Genomics Viewer (IGV)?*
  - *History & Background Information.*
  - *How to Access the IGV:*
    - *Local installation of IGV.*
    - *Accessing IGV using the BioHPC and training accounts (associated with this nanocourse).*

- *How to analyze human and other kinds of **DNA** sequencing data to determine where variants are.*
  - *How to load the proper data, and find information in online databases.*
  - *How to determine whether SNPs are hetero or homozygous*
    - *Introduction to checking for errors such as strand bias.*

- *How to analyze mouse and other kinds of **RNA** sequencing data to determine where variants are.*

UT Southwestern
Medical Center

Part 1: NGS Analysis with The Broad Institute IGV

Usage, Examples, and Exercises

$(\approx 3 - 3.5$ *hours*$)$

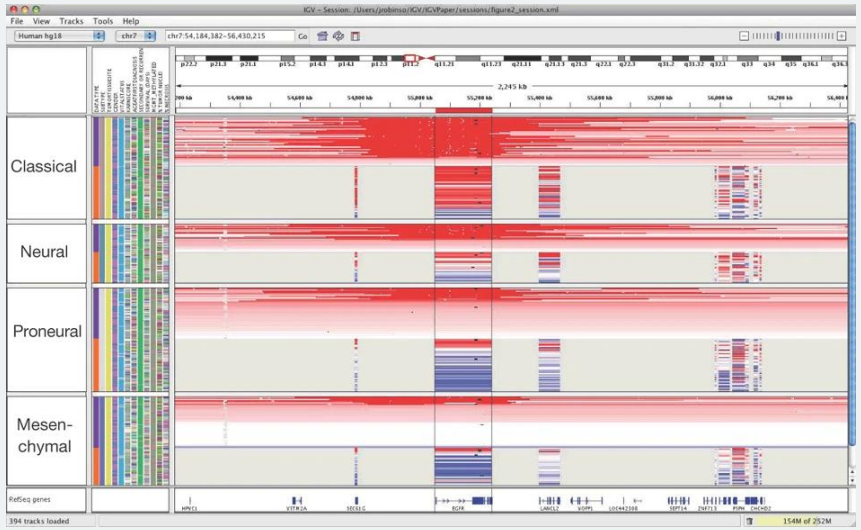
UT Southwestern
Medical Center

The Broad Institute Integrative Genomics Viewer

- *Developed and Maintained by the **Broad Institute**:*
  - *A **Joint Institute of Harvard and MIT**.*
  - *Founded in **2004** (from organizational remnants of the Human Genome Project) for the purposes of [from https://www.broadinstitute.org/about-us]:*
    - *"Illuminating Human Disease"*
    - *"Reading and Editing Genomes"*
    - *"Sharing Data and Tools"*
    - *"Building Communities"*
    - *"Developing Diagnostics and Treatments"*
    - *"Collaborating, Innovating, and Empowering"*

- ***Integrative Genomics Viewer (IGV)** is a software environment for analyzing NGS data*
  - *Provided and maintained by The Broad Institute, with IGV team based at UC San Diego, and MIT/Harvard Broad Institute.*
  - *Supported by funding from:*
    - *The National Cancer Institute (NCI)*
    - *The National Institutes of Health (NIH)*
    - *Informatics Technology for Cancer Research (ITCR)*
    - *Starr Cancer Consortium*
  - *Written in Java, hence allowing use of platform independent jvm.*

BROAD INSTITUTE



From Robinson, J., Thorvaldsdóttir, H., Winckler, W. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29,** 24–26 (2011). https://doi.org/10.1038/nbt.1754

UT Southwestern Medical Center

A C G T

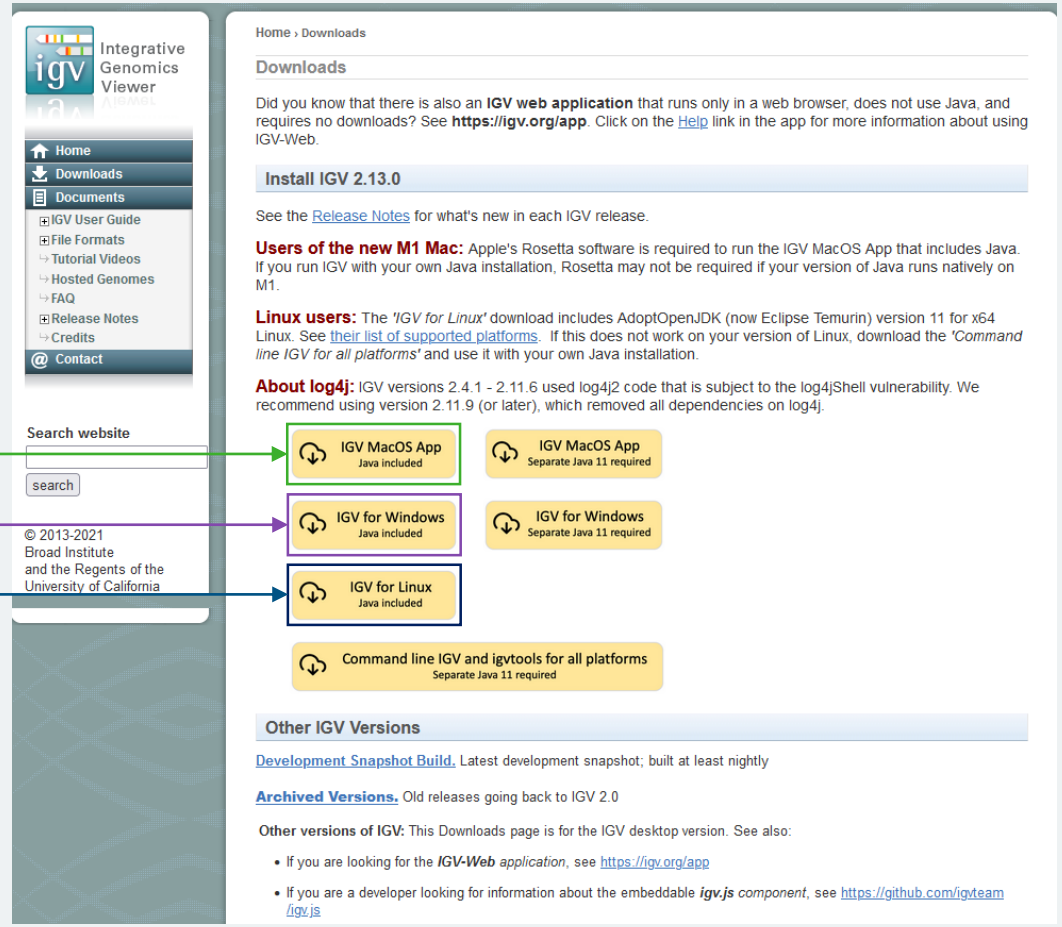
Locally Installing the IGV

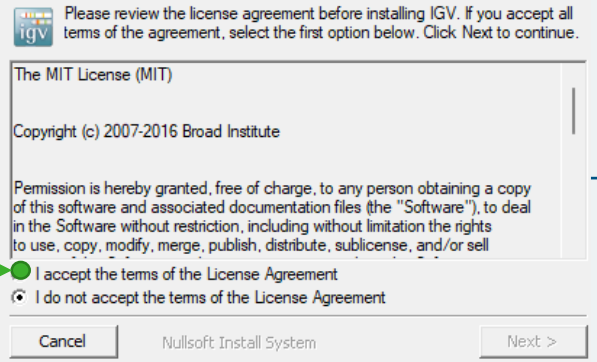
- *IGV is freely available for download and use, and since it was written using java, versions exist for all major operating systems (on which the JVM can run):*

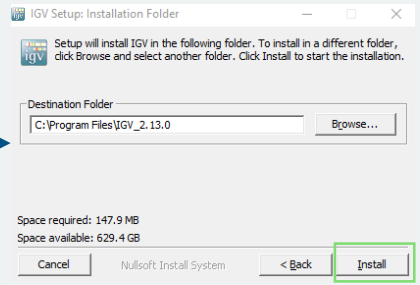
Mac OS

Windows

Linux







- Note: Must agree to The MIT License to install and use IGV.

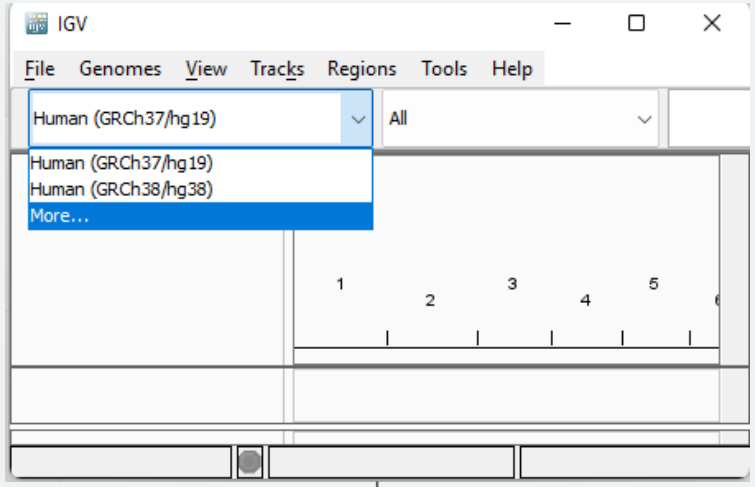
- Very non-restrictive license.
- May do many things, as long as copy of original license included.

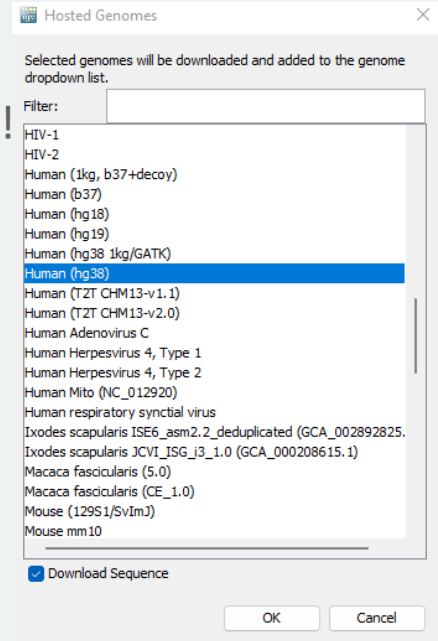
- More info: https://choosealicense.com/licenses/mit/

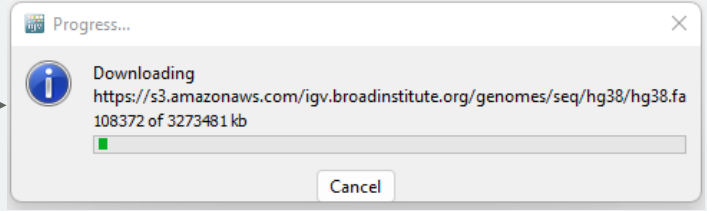
UT Southwestern Medical Center

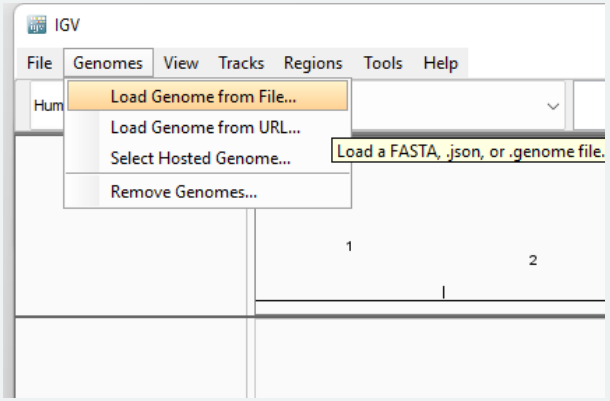
Installing new reference genomes in IGV

- Recall that reference genomes are very large files which contain FASTA format versions of the consensus of many samples of the same organism.
  - The most commonly used version of the human reference genome is called GRCh38, which was the co-ordinated result (by the Genome Reference Consortium) of many different studies including
    - 1000 Genomes Project
  - GRCh38 was released in December of 2013.
    - Most recent *version* February 2022
      - *Gapless except chromosome Y.*
  - Many different references available for download!
  - Can also install directly from downloaded or Manually assembled FASTA reference file.









UT Southwestern Medical Center

Loading a Sequence Alignment Map File (aligned sequencing reads file)

Two Result Files necessary for display:

1. Binary Alignment Map (.bam)
2. Binary Alignment Index (.bai)

Example Queries and Reference (input reads to aligner)

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                  ATAGCT..............TCAGC
-r003                        ttagctTAGGC
-r001/2                                   CAGCGGCAT
```

Alignment tool:

- HISAT2
- Bowtie
- BWT
- STAR

Example SAM output (input to igv)

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

`samtools view file.sam > file.bam`

Fields of SAM

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
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	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Format Specification Images all defined in SAM format document, images from document available below, with more detail:

https://samtools.github.io/hts-specs/SAMv1.pdf
(SAM Specification)

`samtools sort file.bam file`
`samtools index file.bam`

BAI – Binary Alignment Index

5.2 The BAI index format for BAM files

Field	Description	Type	Value
magic	Magic string	char[4]	BAI\1
n_ref	# reference sequences	uint32_t	$< 2^{31}$
List of indices (n=n_ref)			
n_bin	# distinct bins (for the binning index)	uint32_t	≤ 37451
List of distinct bins (n=n_bin)			
bin	Distinct bin	uint32_t	≤ 37450
n_chunk	# chunks	uint32_t	*limited*[35]
List of chunks (n=n_chunk)			
chunk_beg	(Virtual) file offset of the start of the chunk	uint64_t	
chunk_end	(Virtual) file offset of the end of the chunk	uint64_t	
n_intv	# 16kbp intervals (for the linear index)	uint32_t	$\leq 2^{17}$
List of intervals (n=n_intv)			
ioffset	(Virtual) file offset of the first alignment in the interval	uint64_t	
n_no_coor (optional)	Number of unplaced unmapped reads (RNAME *)	uint64_t	

The index file may optionally contain additional metadata providing a summary of the number of mapped and placed unmapped read-segments per reference sequence, and of any unplaced unmapped read-segments.[36] This is stored in an optional extra metadata pseudo-bin for each reference sequence, and in the optional trailing n_no_coor field at the end of the file.

The pseudo-bins appear in the references' lists of distinct bins as bin number 37450 (which is beyond the normal range) and are laid out so as to be compatible with real bins and their chunks:

bin	Magic bin number	uint32_t	37450
n_chunk	# chunks	uint32_t	2
ref_beg	(Virtual) file offset of the start of reads placed on this reference	uint64_t	
ref_end	(Virtual) file offset of the end of reads placed on this reference	uint64_t	
n_mapped	Number of mapped read-segments for this reference	uint64_t	
n_unmapped	Number of unmapped read-segments for this reference	uint64_t	

BAM – Binary Encoding of SAM

BAM is compressed in the BGZF format. All multi-byte numbers in BAM are little-endian, regardless of the machine endianness. The format is formally described in the following table where values in brackets are the default when the corresponding information is not available; an underlined word in uppercase denotes a field in the SAM format.

Field	Description	Type	Value
magic	BAM magic string	char[4]	BAM\1
l_text	Length of the header text, including any NUL padding	uint32_t	$< 2^{31}$
text	Plain header text in SAM; not necessarily NUL-terminated	char[l_text]	
n_ref	# reference sequences	uint32_t	$< 2^{31}$
List of reference information (n=n_ref)			
l_name	Length of the reference name plus 1 (including NUL)	uint32_t	*limited*
name	Reference sequence name; NUL-terminated	char[l_name]	
l_ref	Length of the reference sequence	uint32_t	$< 2^{31}$
List of alignments (until the end of the file)			
block_size	Total length of the alignment record, excluding this field	uint32_t	*limited*
refID	Reference sequence ID, $-1 \leq$ refID $<$ n_ref; -1 for a read without a mapping position	int32_t	[-1]
pos	0-based leftmost coordinate (= POS − 1)	int32_t	[-1]
l_read_name	Length of read_name below (= length(QNAME) + 1)	uint8_t	
mapq	Mapping quality (=MAPQ)	uint8_t	
bin	BAI index bin, see Section 4.2.1	uint16_t	
n_cigar_op	Number of operations in CIGAR, see Section 4.2.2	uint16_t	
flag	Bitwise flags (= FLAG)[29]	uint16_t	
l_seq	Length of SEQ	uint32_t	*limited*
next_refID	Ref-ID of the next segment ($-1 \leq$ next_refID $<$ n_ref)	int32_t	[-1]
next_pos	0-based leftmost pos of the next segment (= PNEXT − 1)	int32_t	[-1]
tlen	Template length (= TLEN)	int32_t	[0]
read_name	Read name, NUL-terminated (QNAME with trailing '\0')[30]	char[l_read_name]	
cigar	CIGAR: op_len<<4|op. 'MIDNSHP=X'→'012345678'	uint32_t[n_cigar_op]	
seq	4-bit encoded read: '=ACMGRSVTWYHKDBN'→ [0,15]. See Section 4.2.3	uint8_t[(l_seq+1)/2]	
qual	Phred-scaled base qualities. See Section 4.2.3	char[l_seq]	
List of auxiliary data (until the end of the alignment block)			
tag	Two-character tag	char[2]	
val_type	Value type: AcCsSiIfZHB, see Section 4.2.4	char	
value	Tag value	(by val_type)	

UT Southwestern Medical Center

Finding and downloading Read Files for your analyses using SRA

One way: Get sra-toolkit for linux! (or windows subsystem for linux)

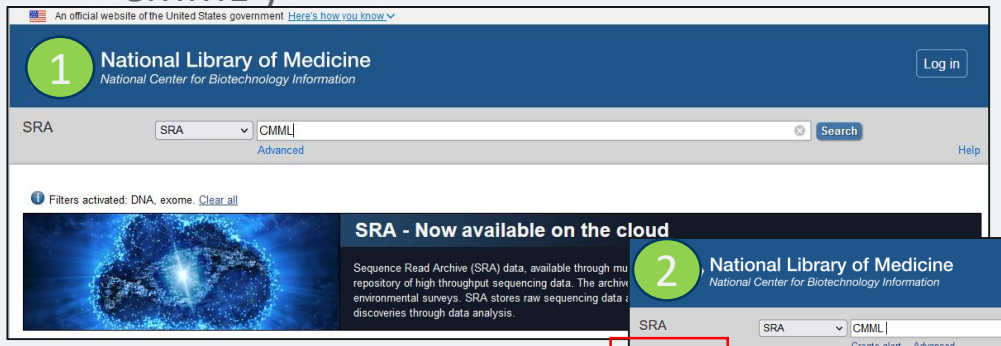
Note: Add to .bashrc for persistent use

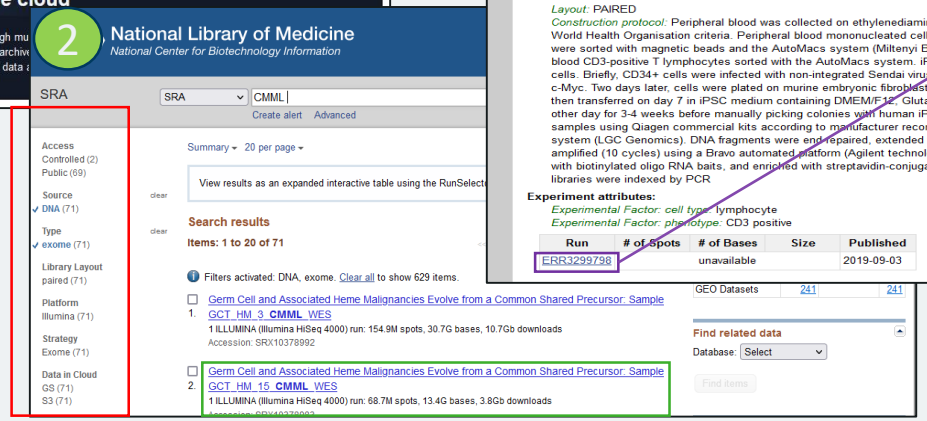
```
wget http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.4.1/sratoolkit.2.4.1-ubuntu64.tar.gz
tar xzvf sratoolkit.2.4.1-ubuntu64.tar.gz
export PATH=$PATH:/directory/sratoolkit.2.4.1-ubuntu64/bin
```

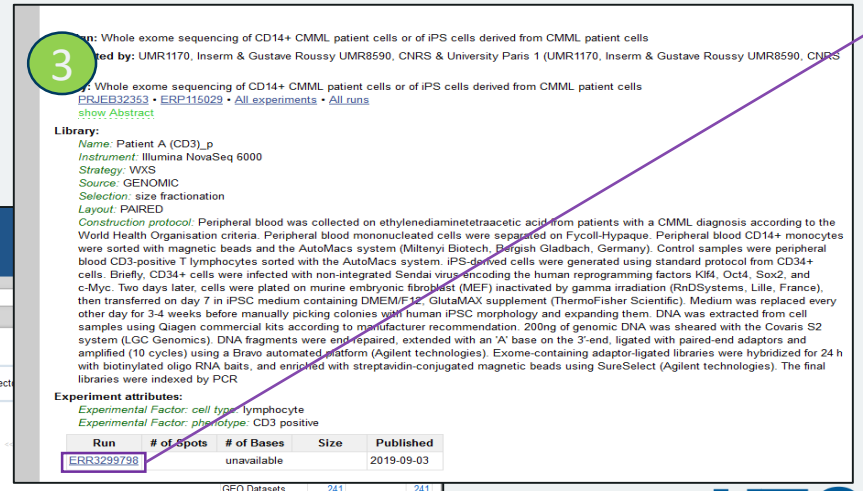
1. Go to the sequencing reads archive (NCBI SRA)
   1. https://www.ncbi.nlm.nih.gov/sra/
2. Create your search for the terms you are interested in investigating (Ie. “CMML”)

```
Use 'prefetch' command to speed up fastq-dump
To download only a specific number of reads uses the -X flag
with fastq-dump (note prefetch fetches the entire file by
default, the -X flag for prefetch does something different.)
```

https://www.reneshbedre.com/blog/ncbi_sra_toolkit.html
(SRA Toolkit Download Full Tutorial)







fastq-dump -X <# reads/pairs> ERR3299798

This will download a fastq file,

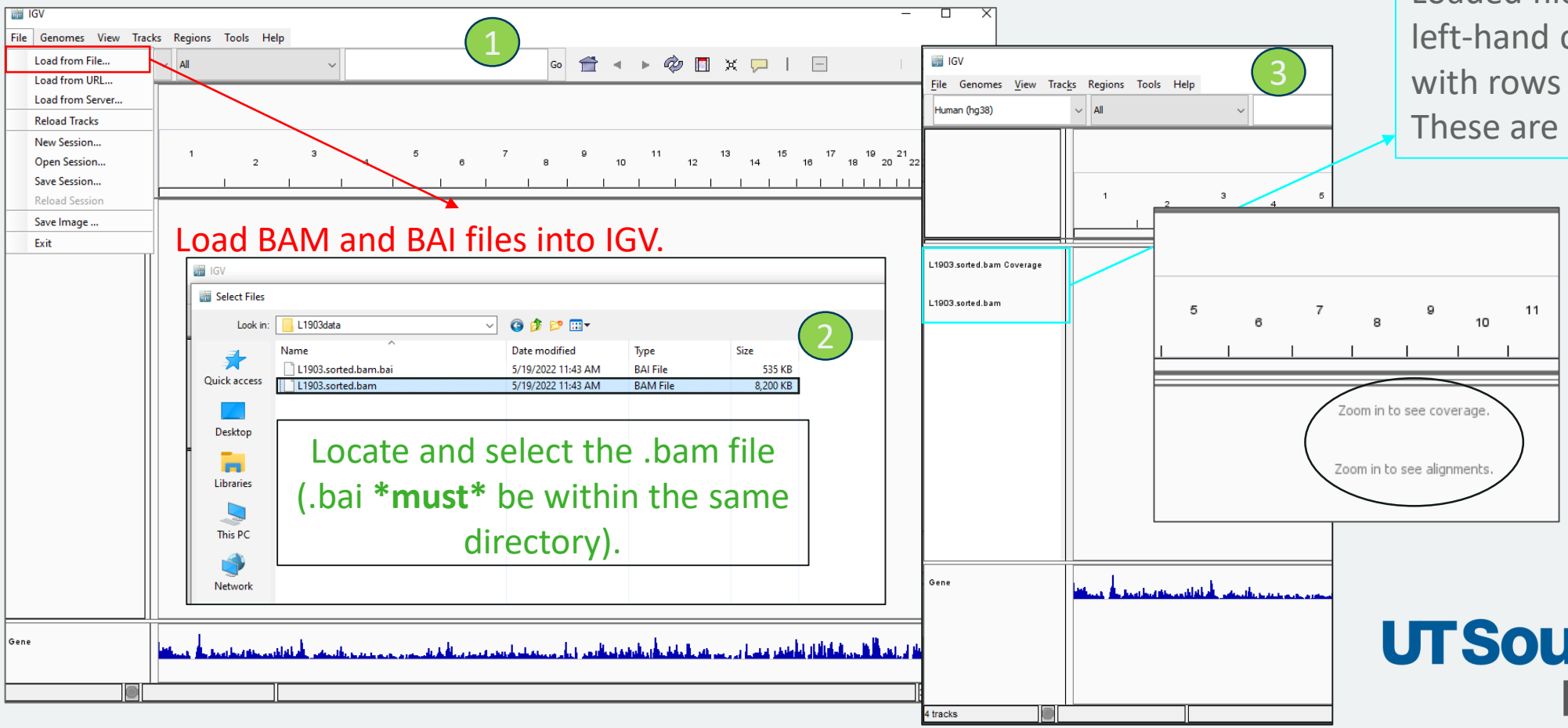
Caution, fastq files can be quite large!

3. Select a result from those displayed (Bonus, use filters to subset the results displayed)

UT Southwestern Medical Center

Loading a Sequence Alignment Map File (aligned sequencing reads file)

1. In this section we will look at some human sequencing reads that have been aligned to the human reference genome (GRCh38) using the HISAT2 software aligner.
2. The data has been converted into the binary format (BAM) using samtools, and an index has been properly created using the same (BAI).



Loaded files appear in the left-hand column (along with rows for coverage). These are called **TRACKS.**

To save memory,

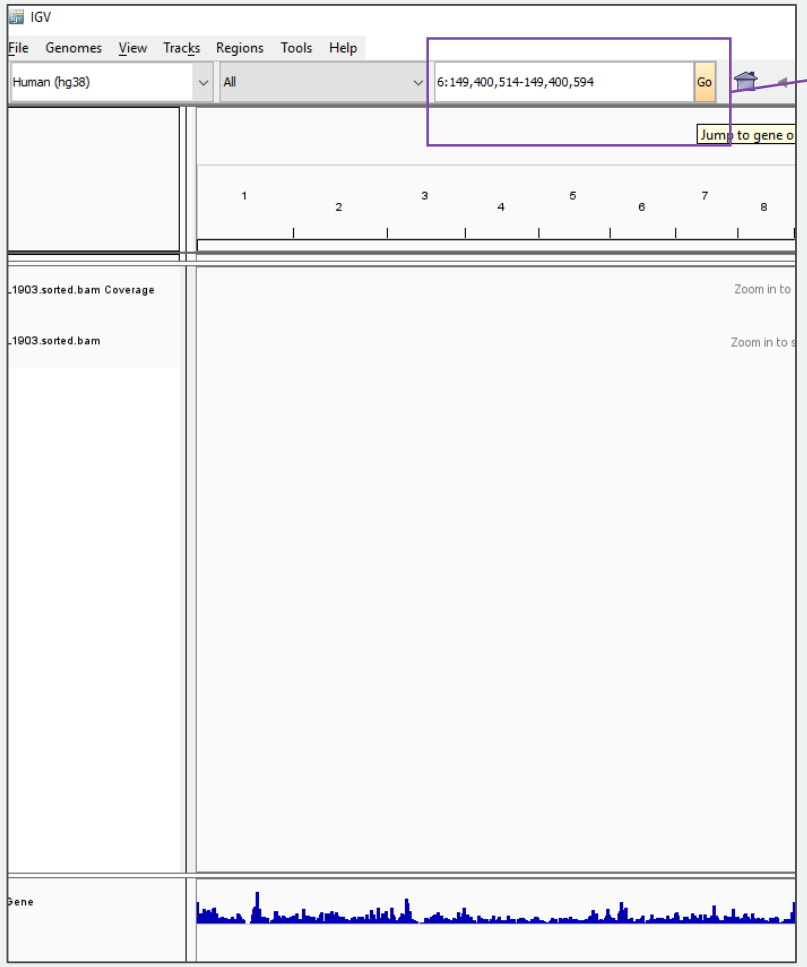
At this full genome level resolution (all chromosomes shown) reads not shown, must zoom in first.

UT Southwestern Medical Center

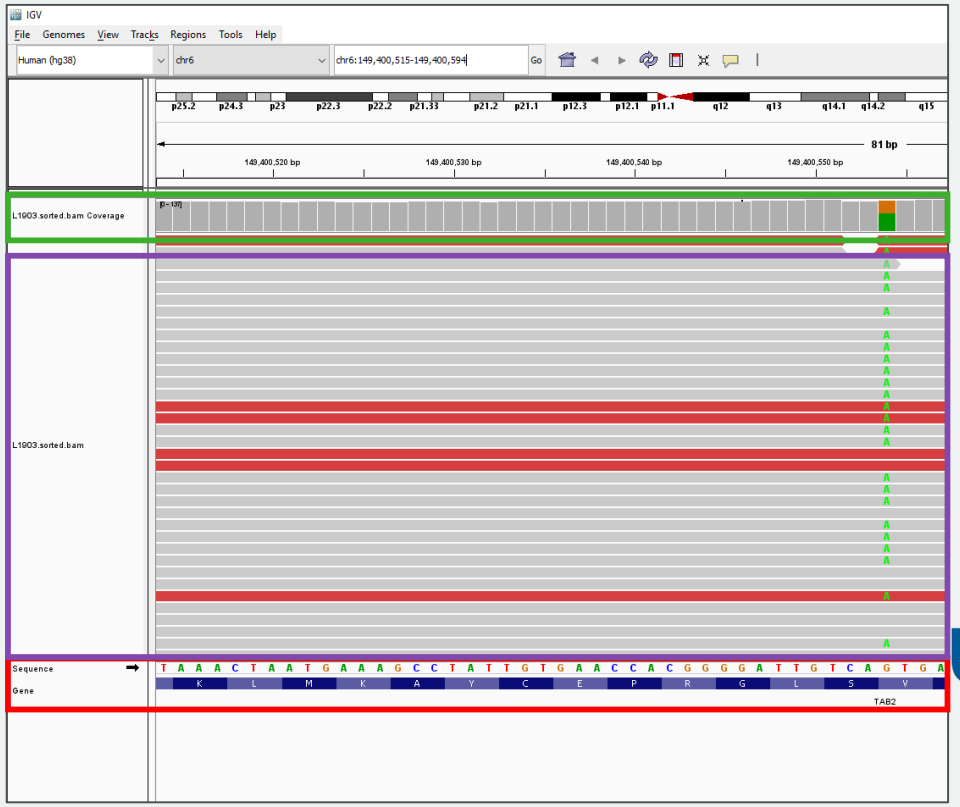
Selecting a new location in a reference genome with IGV

A
C
G
T

1. IGV Will not display reads at the full-reference level, (or in many cases even the chromosome level by default).
    1. We will change this in view>preferences soon.
2. For now, we can change both the window zoom level (range) and location by specifying them in the “Enter a gene or locus” search box in the center of the top tool bar.



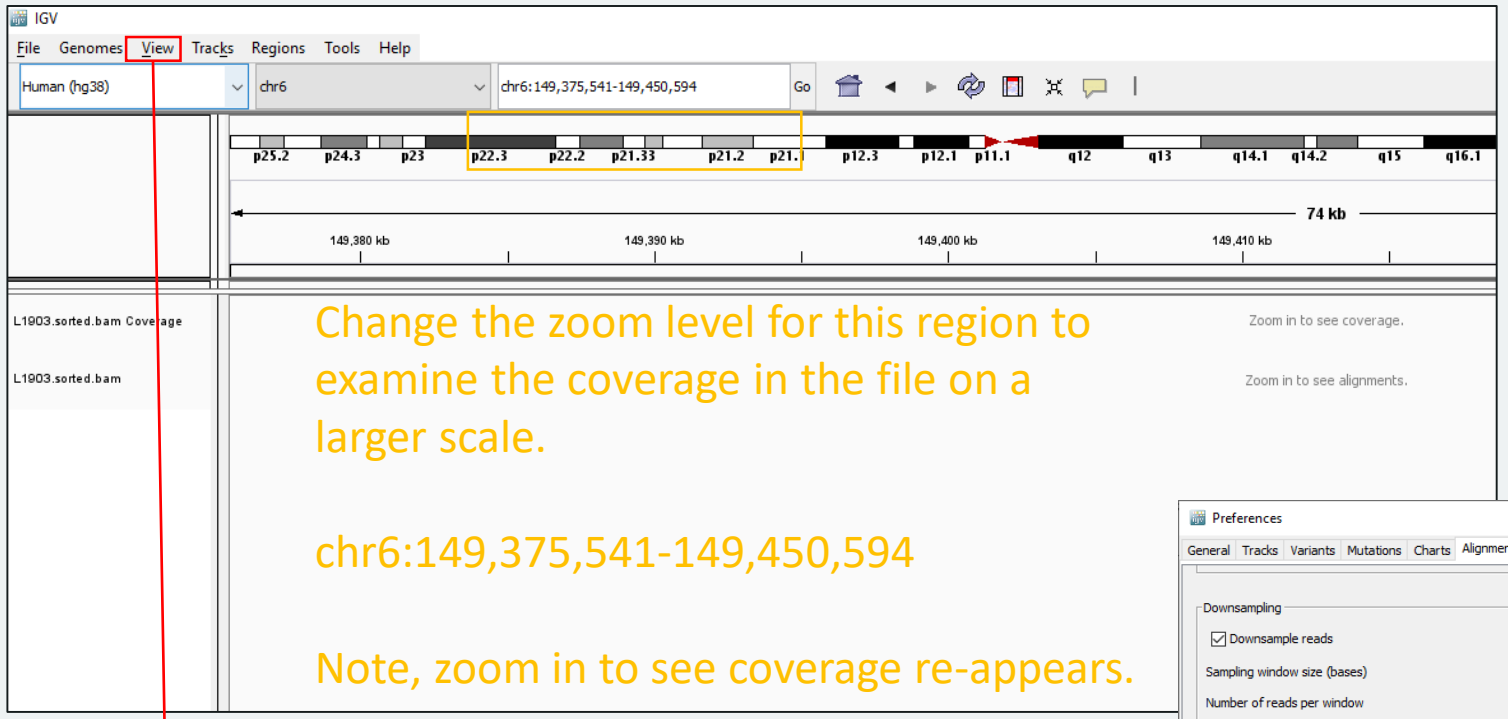
We would like to view between reference locus 149,400,514 and 149,400,594 (90 bases) on chromosome 6.
We can specify this as: “6: 149,400,514-149,400,594”



At this level of resolution (90 bp) we get coverage information from the bam/bai in the form of a histogram, read alignment information where reference matches are grey, and alternative bases are colored and labeled from the bam/bai, and reference sequence information from the reference selected.

UT Southwestern Medical Center

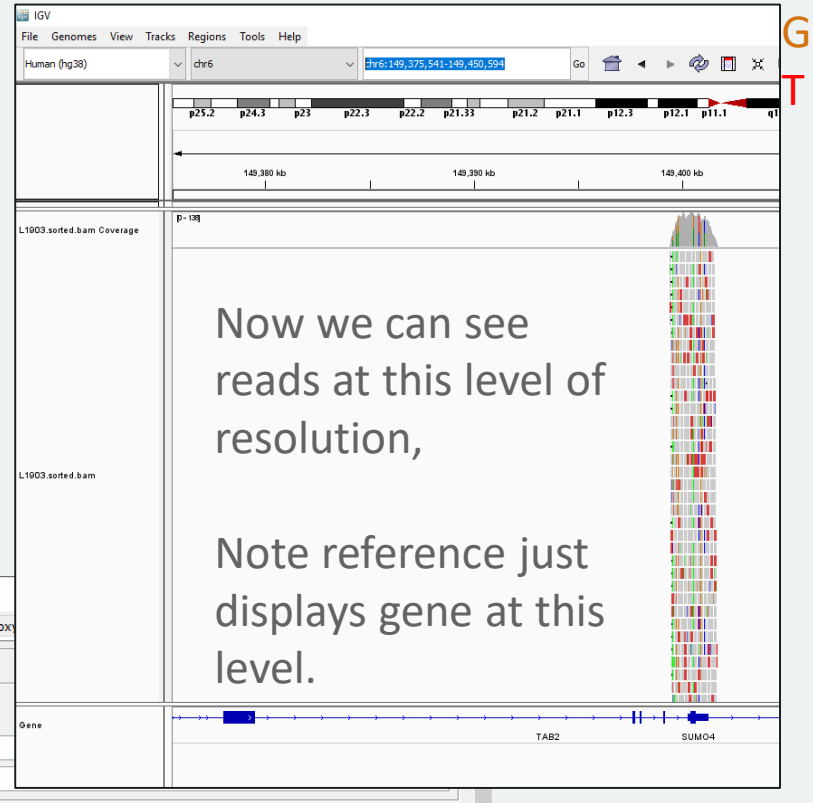
Selecting a new location in a reference genome with IGV (changing visibility range)



Change the zoom level for this region to examine the coverage in the file on a larger scale.

chr6:149,375,541-149,450,594

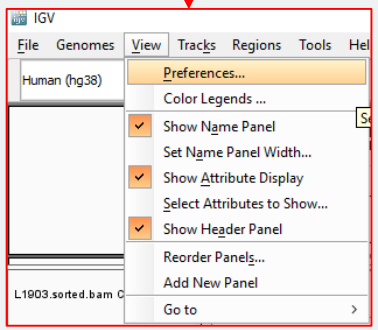
Note, zoom in to see coverage re-appears.

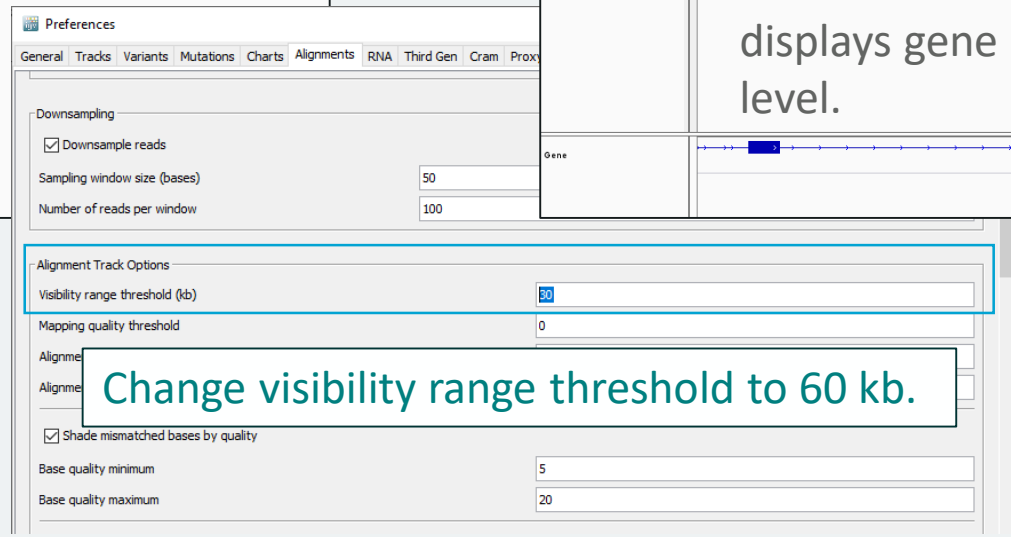


Now we can see reads at this level of resolution,

Note reference just displays gene at this level.

- To change the default for displaying coverage to allow viewing at this level go to
  - view>preferences>Alignments
  - In the Alignment Track Options box, change the Visibility range threshold (kb) from 30 to 60 and press “save”.

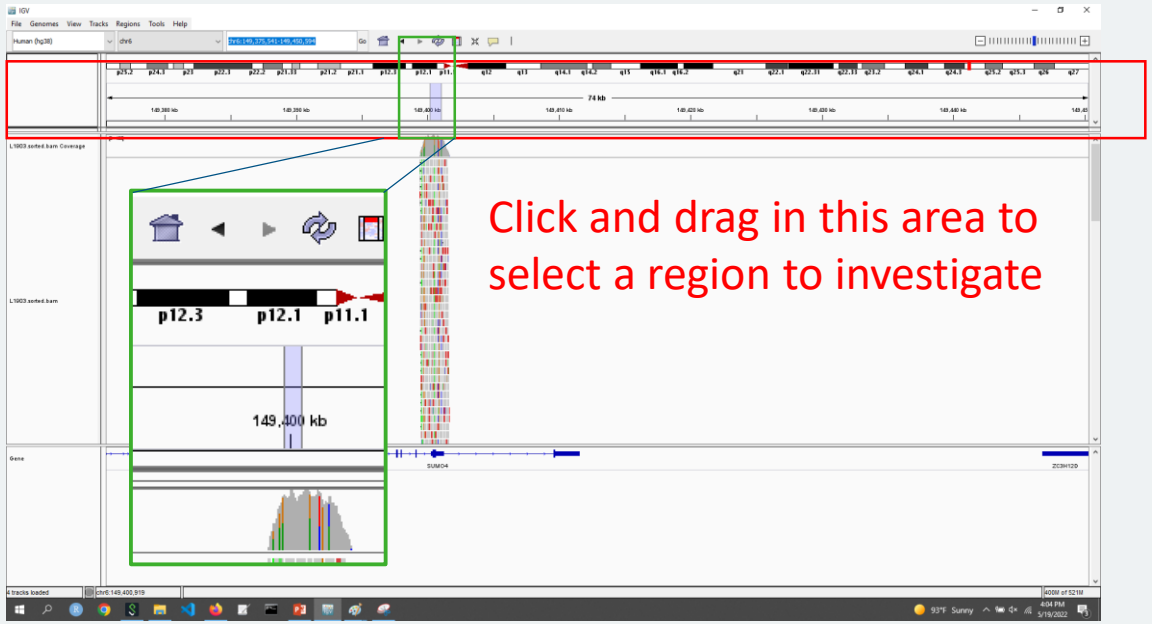


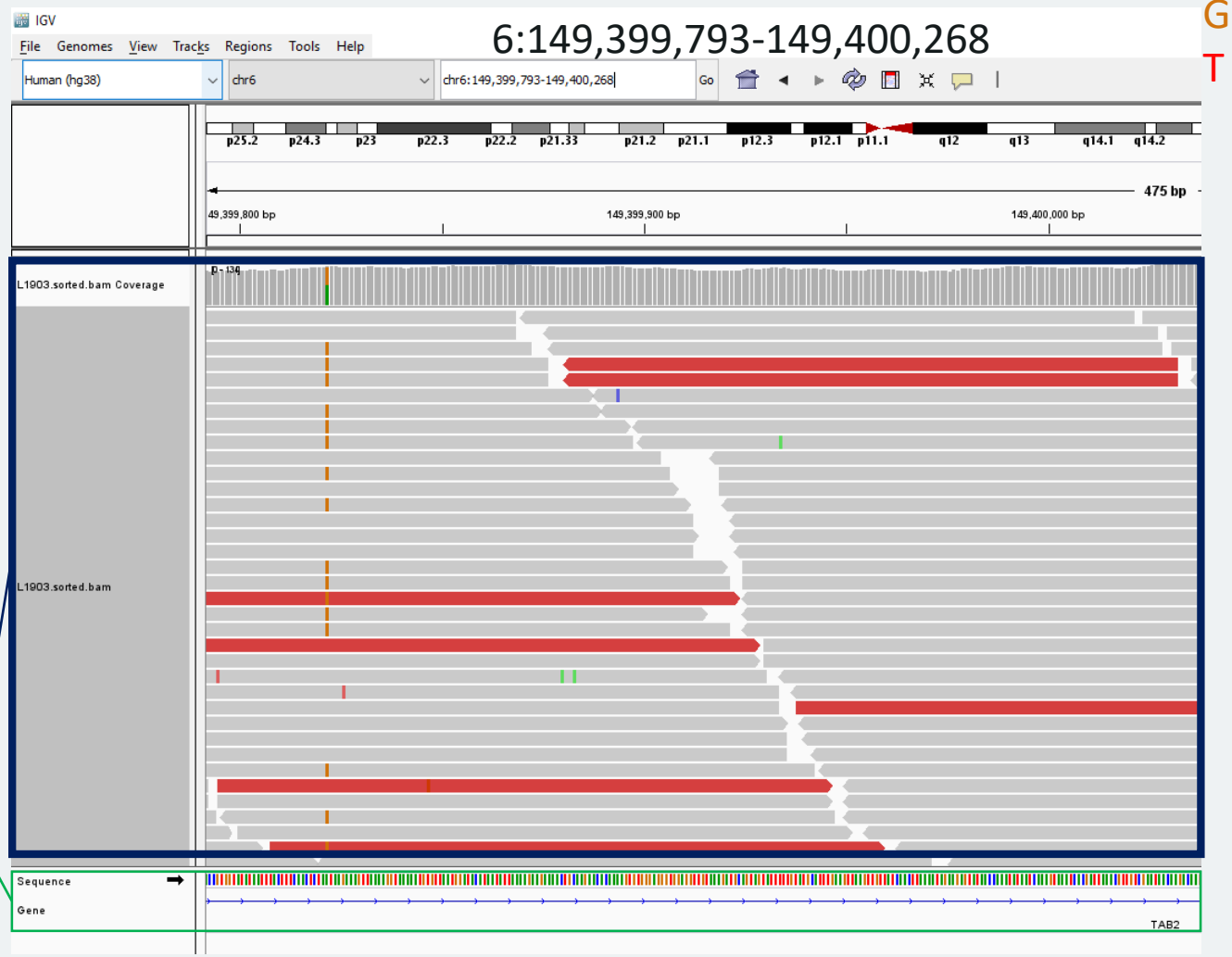


Change visibility range threshold to 60 kb.

UT Southwestern Medical Center

Navigating in the IGV Browser Window.



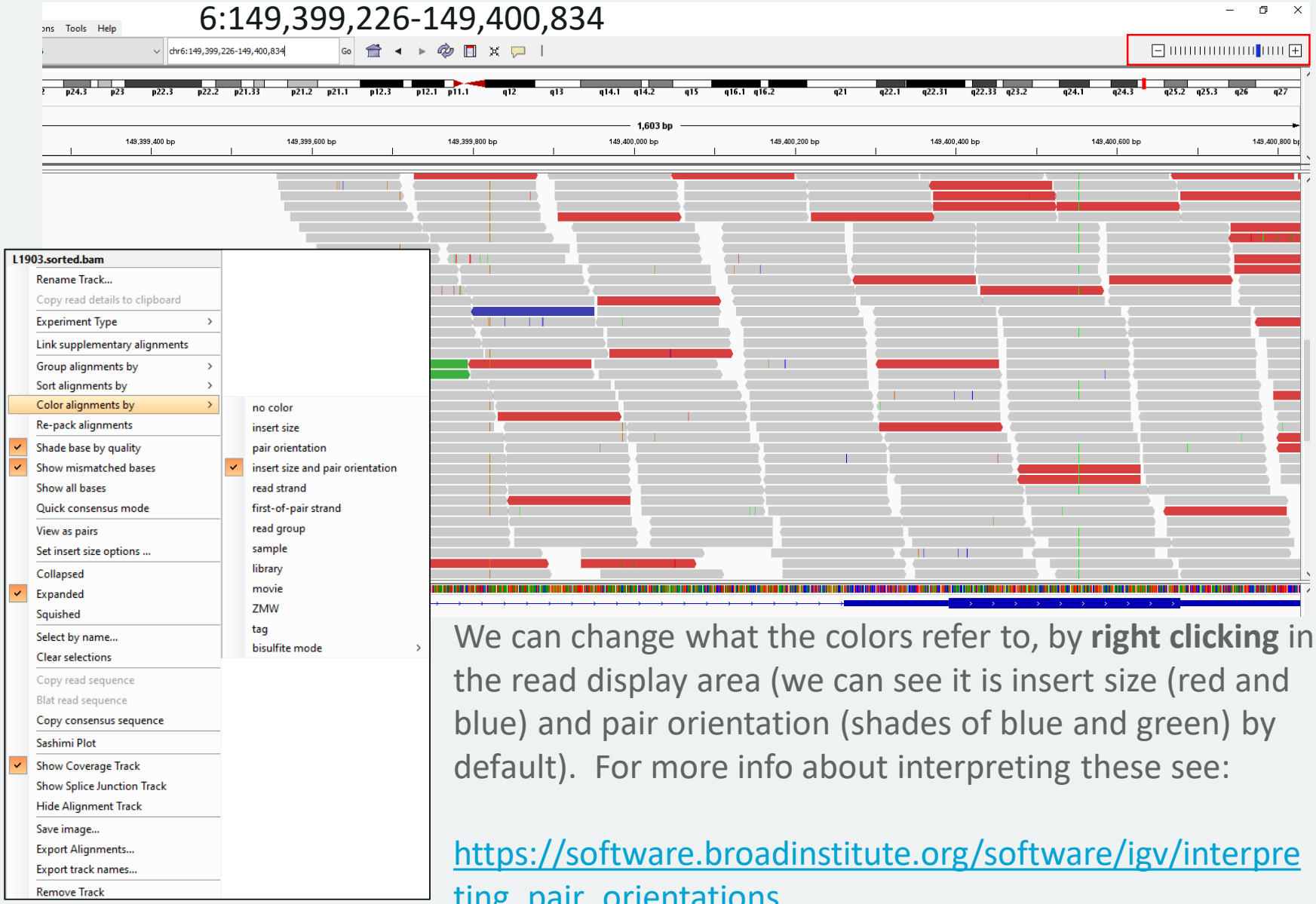


- At this level of resolution we can see bases are labeled by color in the reference
- Those reads with base calls differing from the reference are marked by color, those which are the same are grey.
- Click and drag in this region to pan manually in the sample.
- Scroll in this region to view more reads and files.

UT Southwestern Medical Center

Navigating in the IGV Browser Window.

Zoom in and out manually with '+' and '-' buttons here.



We can change what the colors refer to, by **right clicking** in the read display area (we can see it is insert size (red and blue) and pair orientation (shades of blue and green) by default). For more info about interpreting these see:

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

Notice at this level of resolution that reads are sometimes displayed with different colors.

Red indicates there is possible evidence of deletion

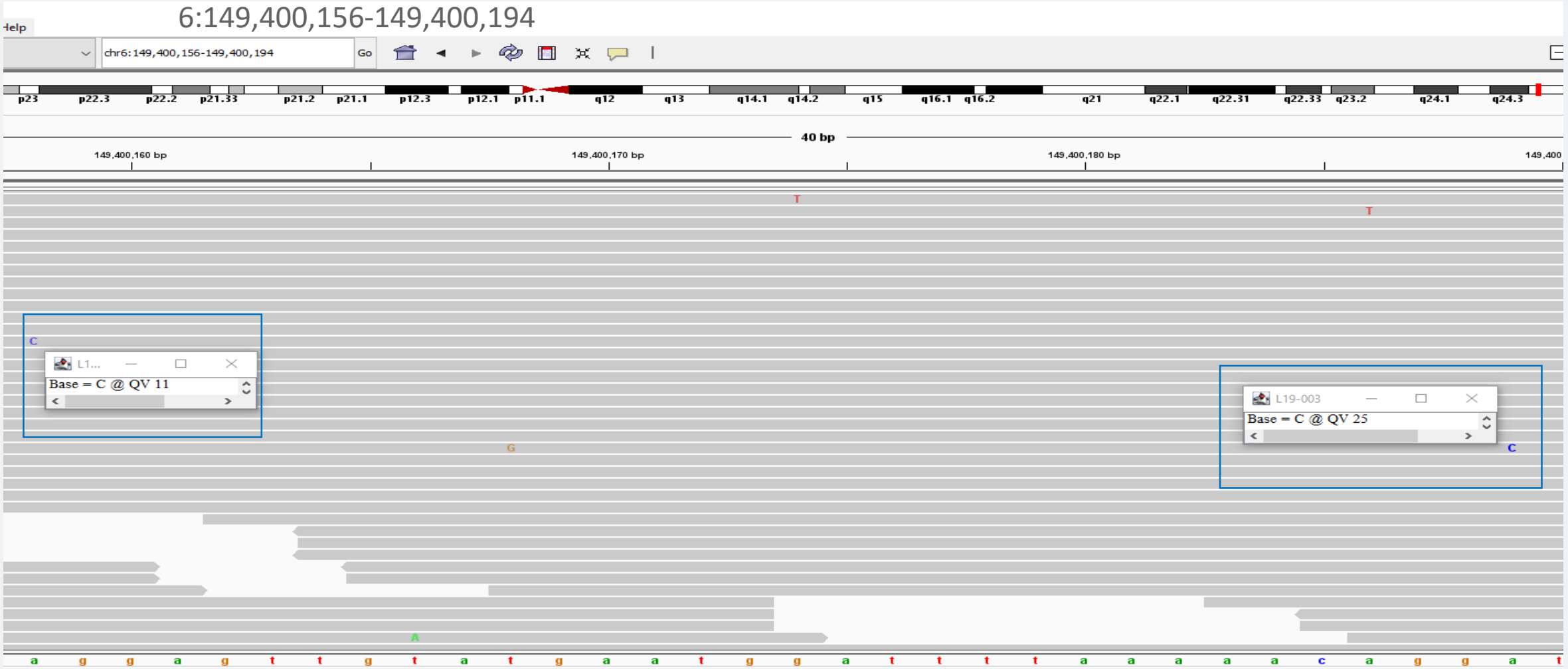
Blue indicates that there is possible evidence of an insertion

Green indicates orientation of the read orientation RL

UT Southwestern Medical Center

A
C
G
T

Navigating in the IGV Browser Window.



By Default, at this higher level of resolution the actual reference and base character is shown, we can see that they are also shaded lighter and darker by default depending on the quality value.

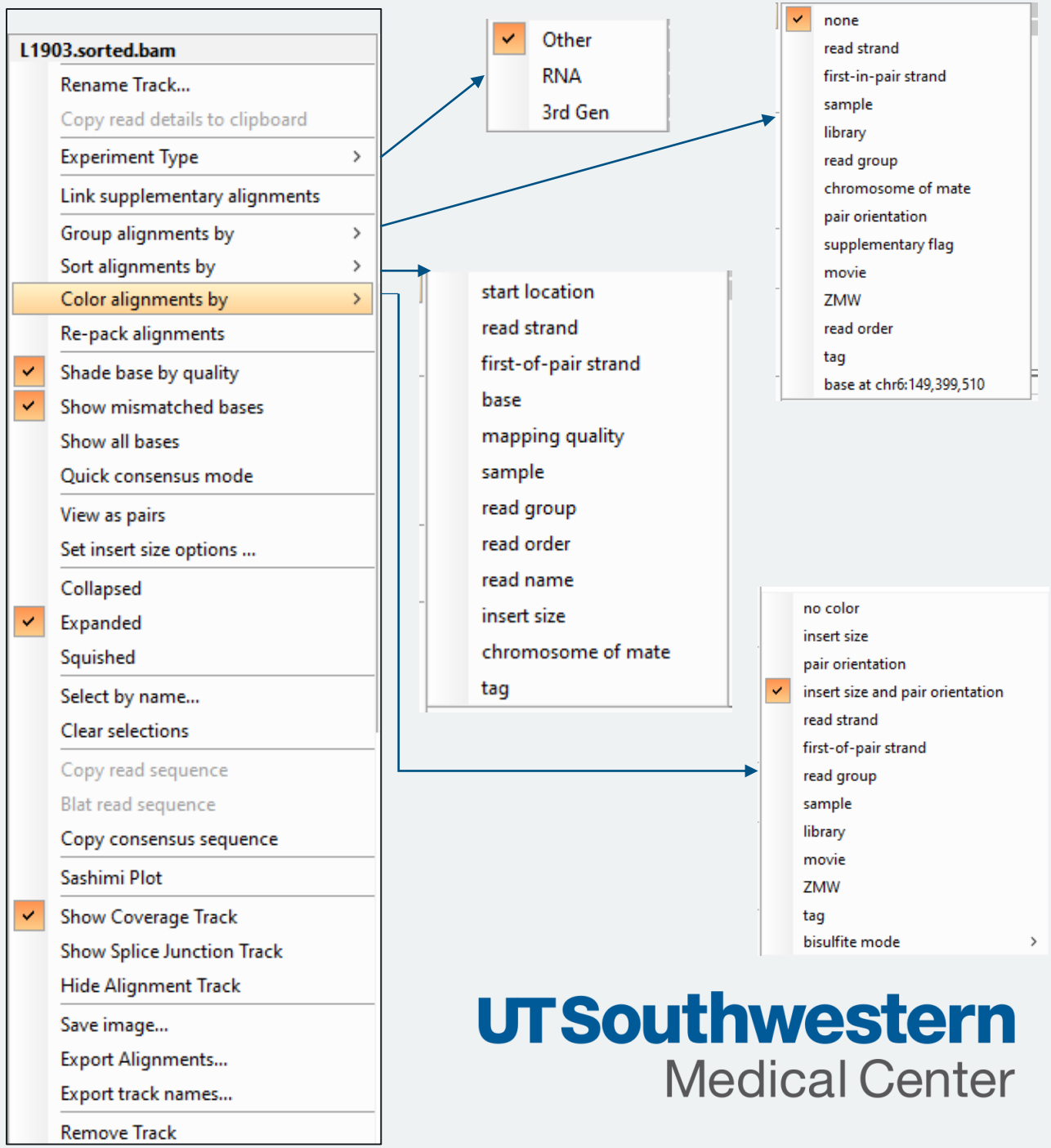
UT Southwestern Medical Center

A
C
G
T

Changing IGV Window Options

In the right click menu, we can toggle the options in the display window, for instance, we can:

1. Change the name on the left, with rename track…
2. Change the Experiment type (other, RNA, 3rd gen)
3. Add additional alignments
4. Group alignments,
5. Sort alignments,
6. Color alignments,
7. Repack – redisplay
8. Toggle shading characters (darker for higher quality)
9. Toggle Whether the mismatched bases are highlighted.
10. Toggle Whether all of the bases are shown
11. Toggle quick consensus mode (only highlight alternative bases If consensus of base calls at loci [mode] is different.
12. Toggle display to show pairs (and identify inserts)
13. Change insert options.

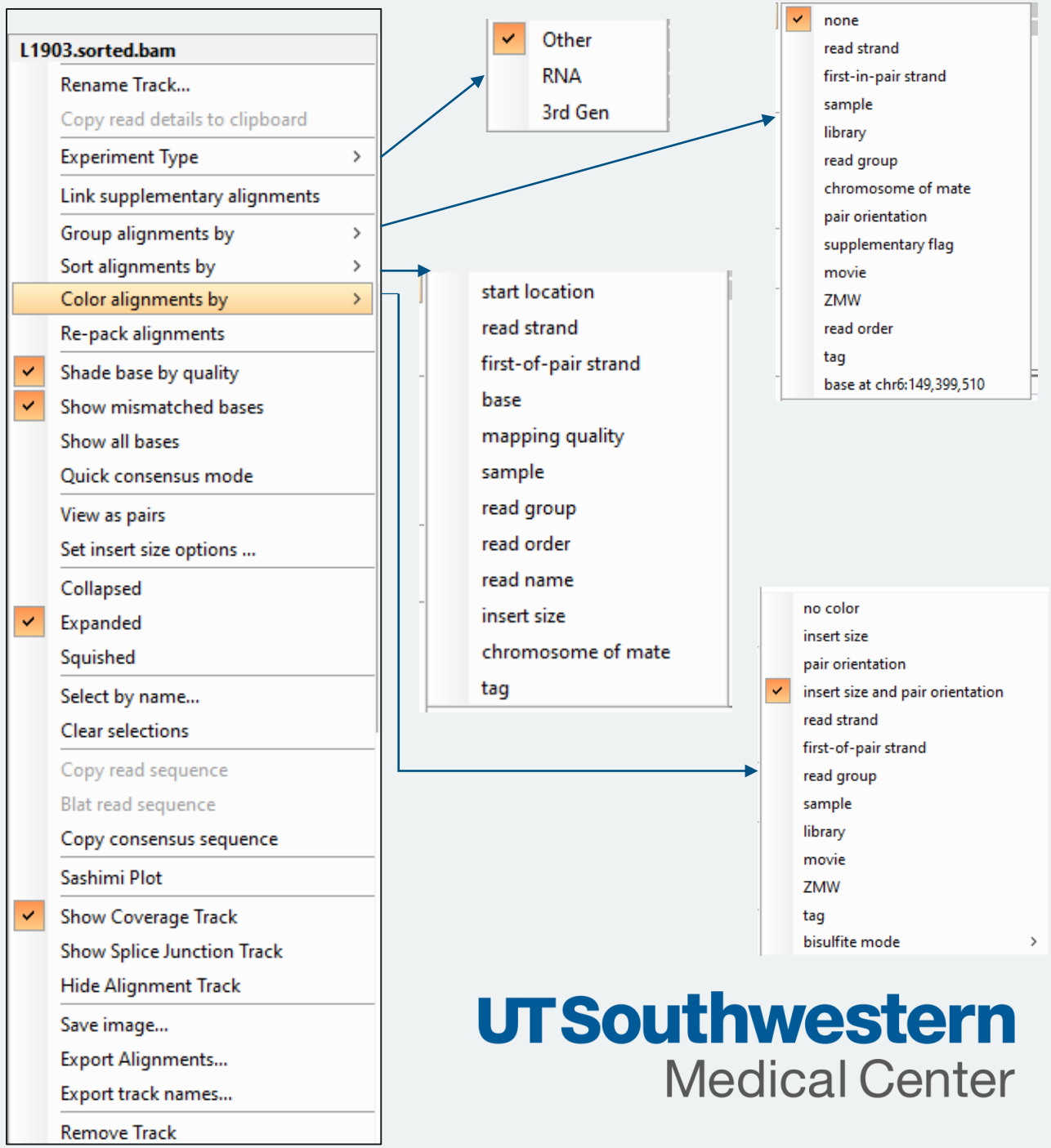


UT Southwestern Medical Center

Changing IGV Window Options (2)

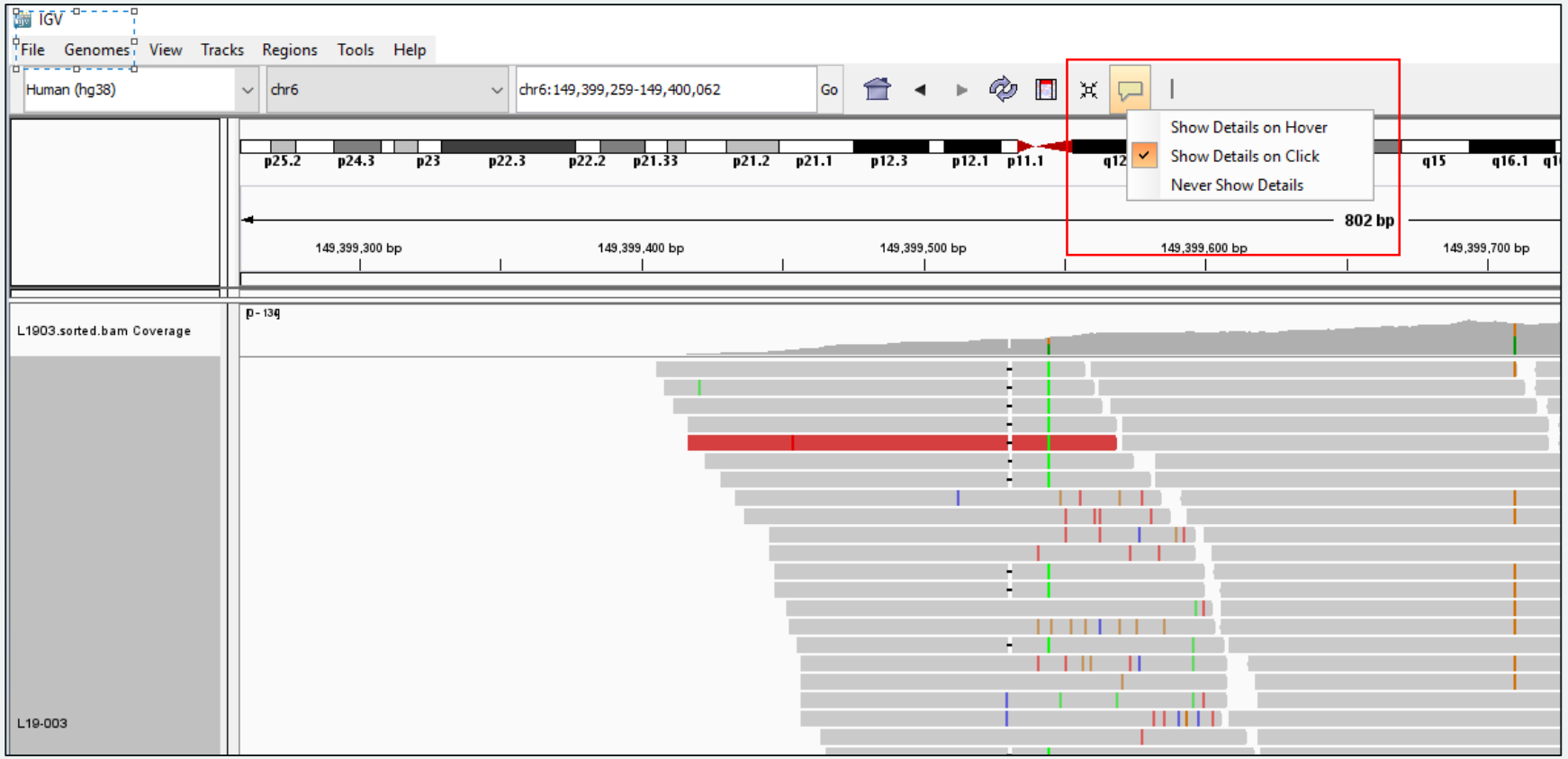
In the right click menu, we can toggle the options in the display window, for instance, we can:

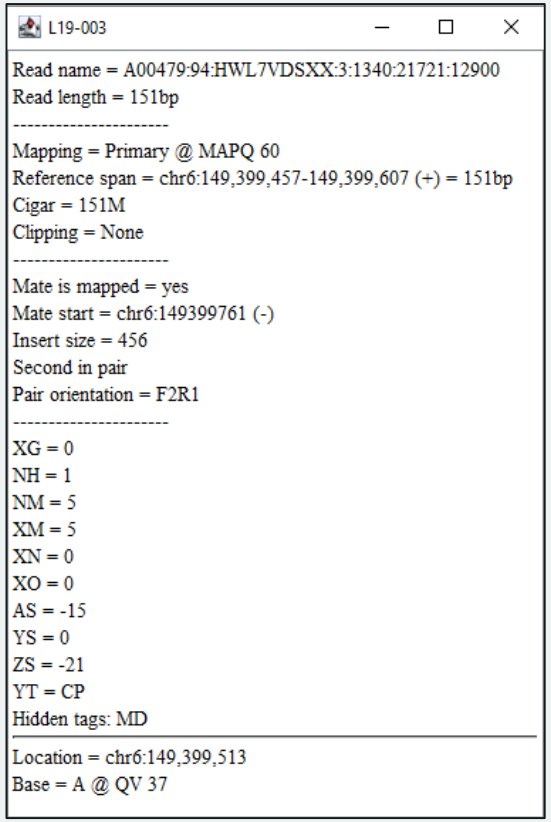
14. Change the vertical resolution (scrunch in the rows to show them all at once (squished), maximum detail (Expanded), and medium detail (Collapsed)).
15. Select a specific read by it's name [if you happened to know it].
16. Clear your selection.
17. Copy the consensus (only for the selected region).
18. Display a sashimi plot [more useful for RNA sequences].
19. Show and hide default tracks (junctions off by default, really only useful for RNA-Seq).
20. Can save images of the current display
21. You can select specific alignments, and export them.
22. If you have a lot of tracks you can export track names to read them back in later.
23. Remove the track (you will have to add it back later to interpret it again).



UT Southwestern Medical Center

Displaying info about a specific base call





A
C
G
T

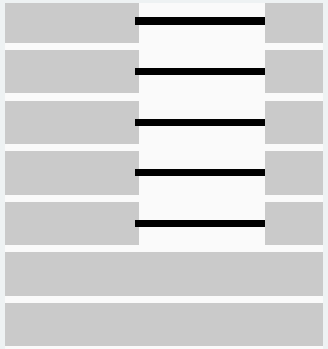
Information can be toggled to be displayed:

1. On mouse-button click. → Can be persisted to compare.
2. On mouse-cursor hover. → Not persistent!

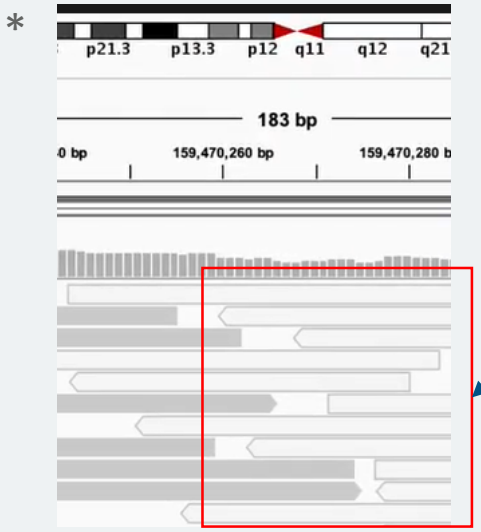
Recall that IGV is primarily a tool for displaying the results of sequence alignment, therefore the information we find here is what we would see in the SAM file.

UT Southwestern Medical Center

A few additional notes on basic analysis with IGV:

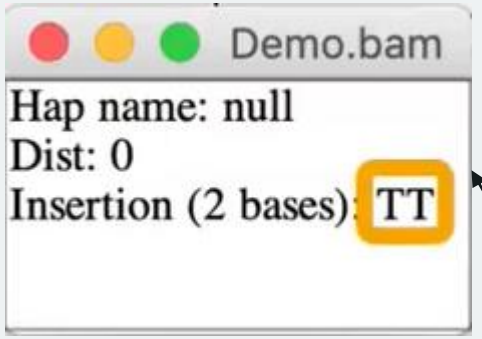


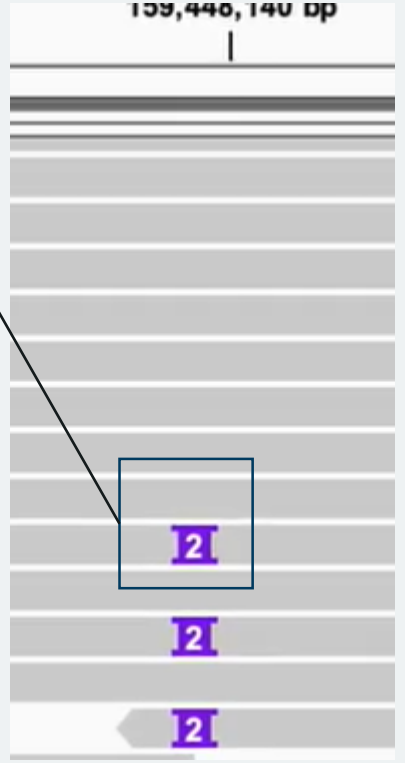
These symbols in the view window indicate gaps in the alignment, which is evidence of deletion.



Hollow reads represent low quality alignments

Insertions are shown by integers in these purple icons, when clicked the inserted bases will be displayed.

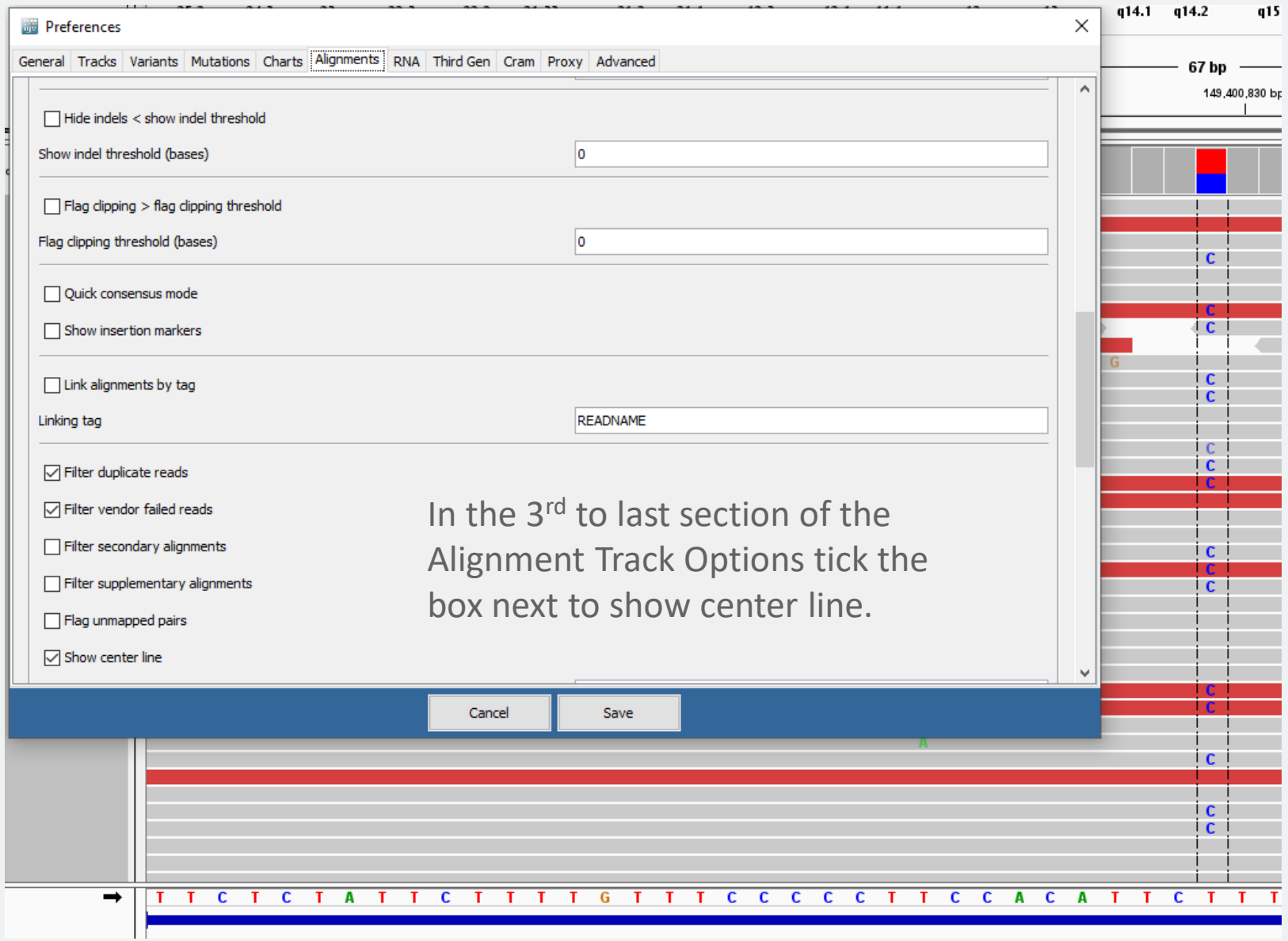




* From Official video tutorial: https://www.youtube.com/watch?v=E_G8z_2gTYM

UT Southwestern Medical Center

Display helpful centerline in view window



To access the settings for displaying the centerline on the plot (dotted vertical lines) go through:

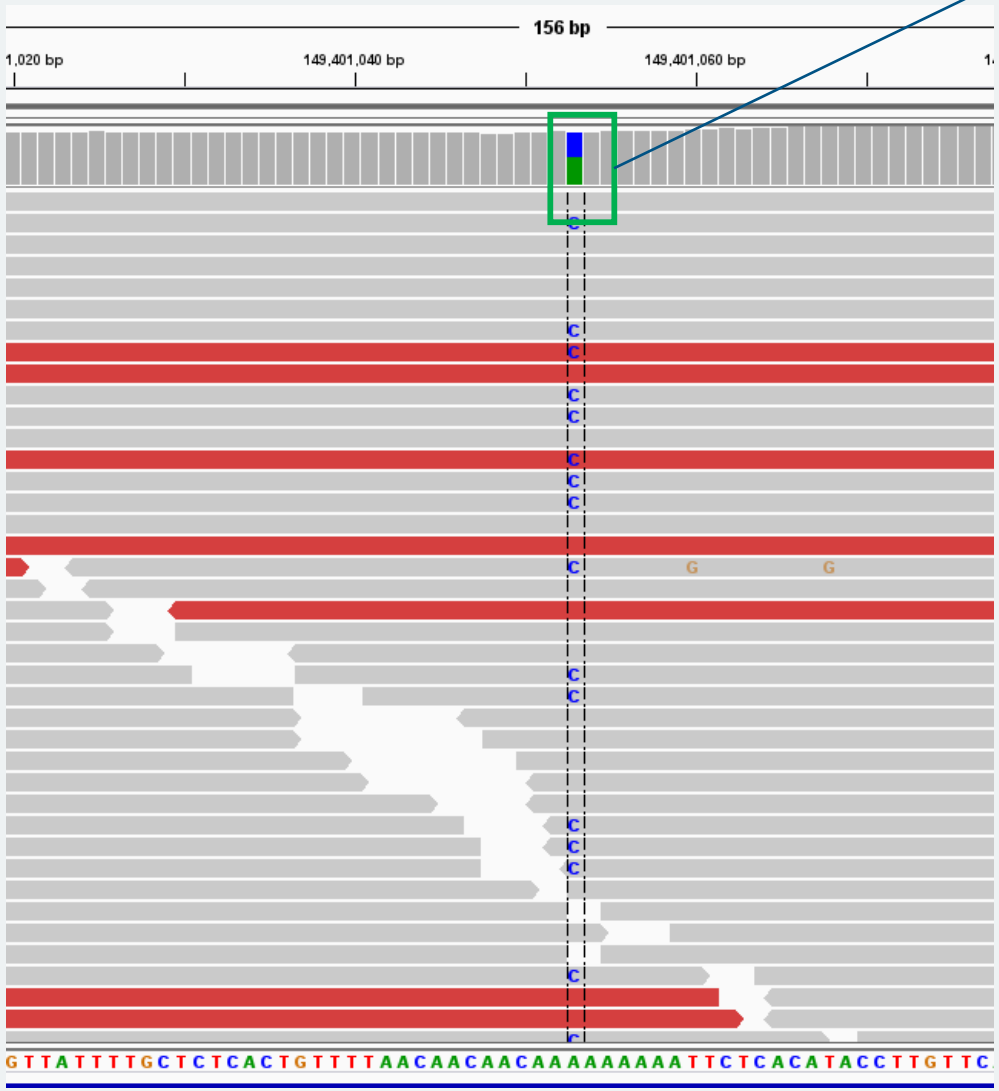
`view>preferences>alignments`

Note there are many useful options to explore here, depending on your experiment you may want to spend some time customizing your view window to display information in the most meaningful way.

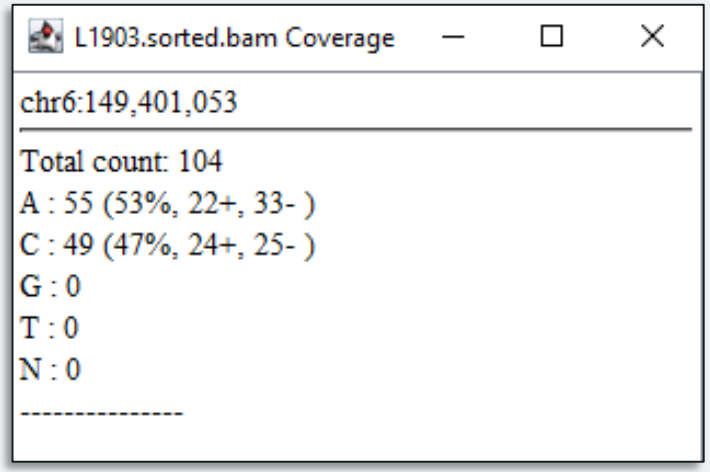
UT Southwestern Medical Center

Viewing SNV information.

chr6:149,400,976-149,401,130



Examine Histogram colors in the coverage TRACK for a quick peak at how the base call distribution is broken down for a specific position.

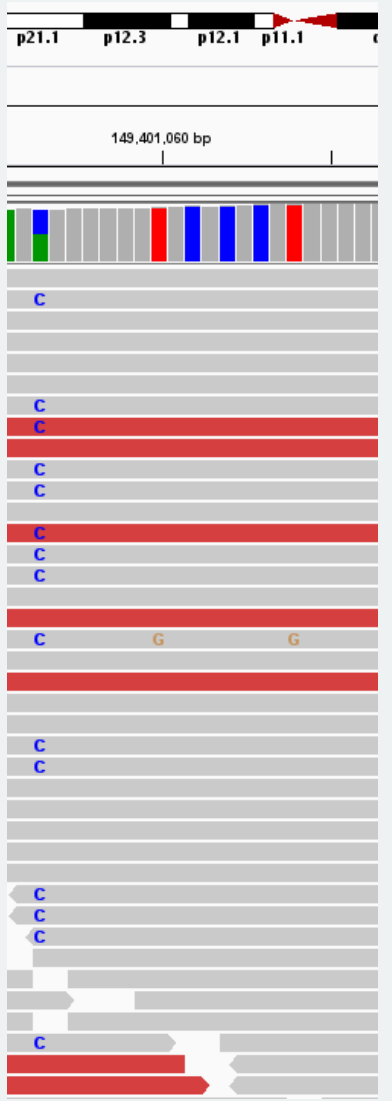


Click the bar in the coverage track for a more detailed breakdown of the distribution of base calls at a particular locus.

Note: Colored bars indicate imbalance in read base calls with frequency greater than a value defined by the user under

view>preferences >alignments

In the Coverage Track Options box, in the Coverage allele-fraction threshold change 0.2f (20 percent *f-floating point*) for example to 0.001f, ridiculous, but can now see even one mismatch.

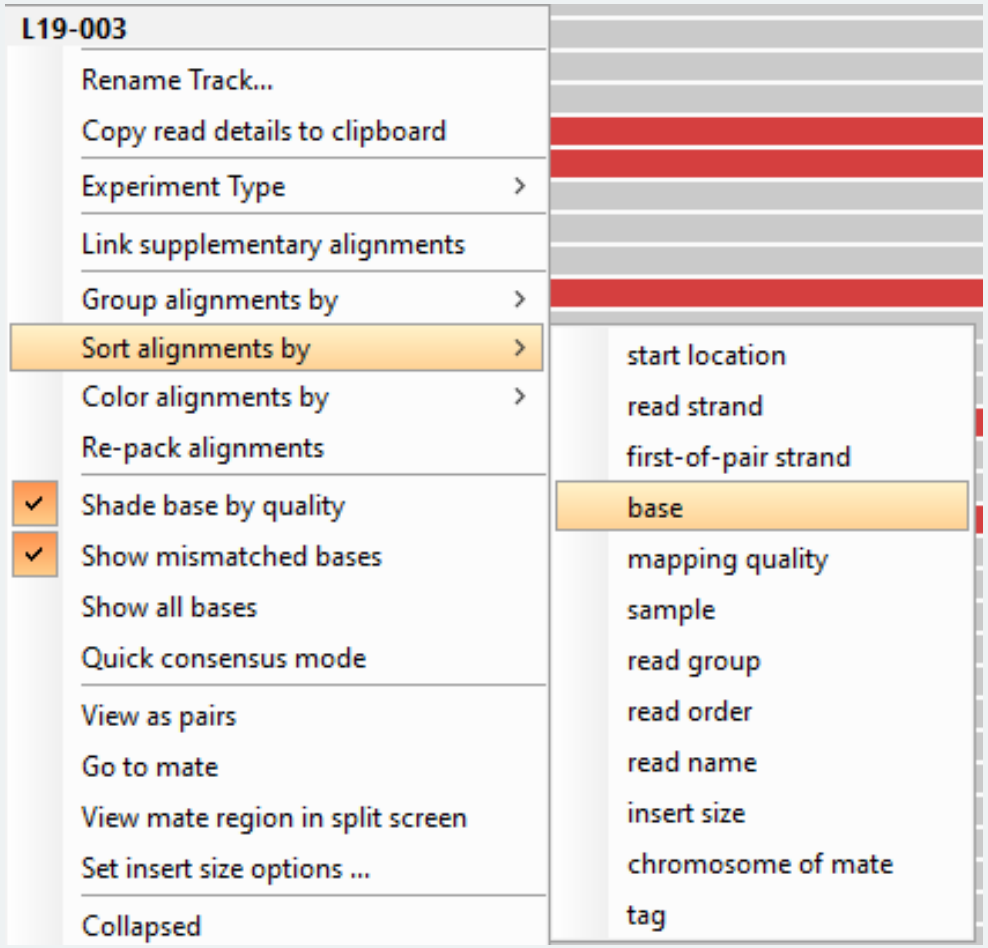


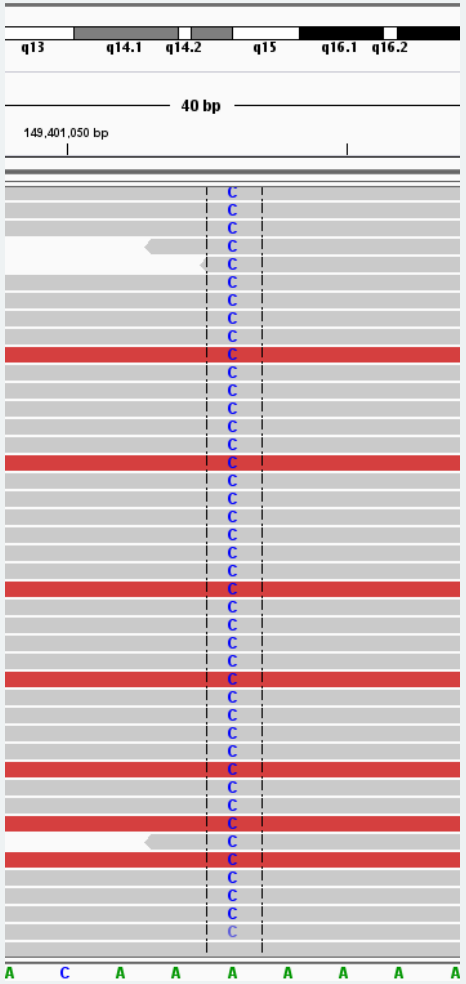
UT Southwestern Medical Center

A C G T

Viewing SNV information.

- Sorting reads by base can allow easy identification of snps.
- Note, these reads are also sorted in descending order of base call quality by default





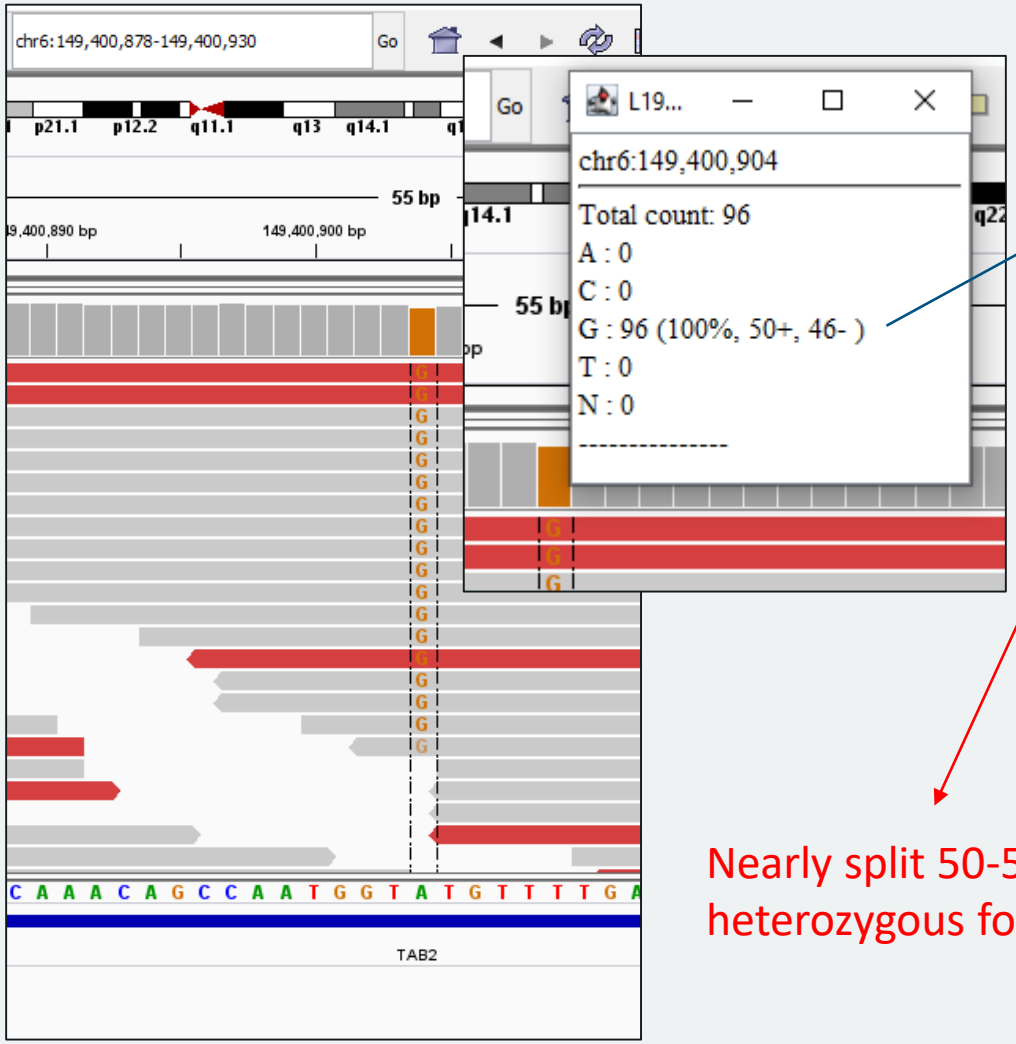
Be Cautious of setting values too low, This may make the visualization Misleading.

Recall, decreasing quality is evidenced by fading of the intensity of the color displayed.

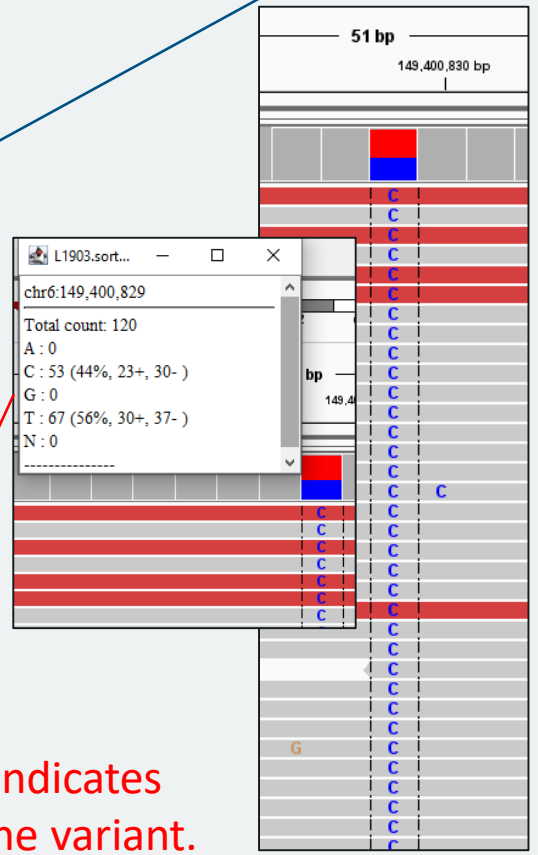
UT Southwestern Medical Center

Viewing SNV information.

- We can also use IGV to make determinations as to whether particular variants indicate homozygous or heterozygous.
- Consider Chr6:149,400,904

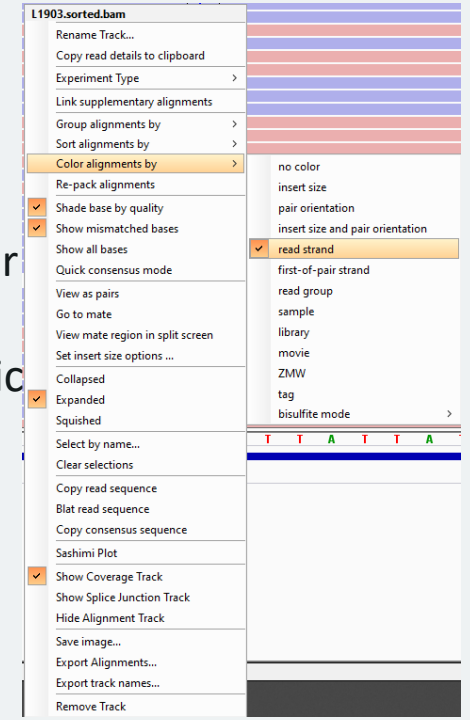


All reads mapping to location differ from reference, and therefore this variant (A->G) appears to be homozygous in the sample.



Nearly split 50-50 indicates heterozygous for the variant.

Can investigate strand bias (variant only appearing in near 50-50 ratio on strands of specific orientation) by coloring alignments by read strand.



UT Southwestern Medical Center

- Consider Chr6:149,400,829

Break

We will continue with an exercise after a short 5 minute break

UT Southwestern
Medical Center

Exercise 1 (1/2): Getting information about DNA variants from online databases

1. Look at the region: 6:149,400,514-149,400,594
    a. Can you see an SNV?
    b. At what position?
    c. How many reads support the alt allele?
    d. How many reads support the ref allele?
    e. Which gene is this region a part of?
    f. Use the following databases to find clinical information associated with this variant: MONDO (ID: 0010863), MedGen (ID: C1838260) and OMIM (ID: 600320)

2. If you have time, examine these other positions:
    a. 7:96,184,336-96,184,416
       MONDO (ID: 0011601), MedGen (ID: C1853942), OMIM (ID: 605814), and Orphanet (ID: ORPHA247598))
    b. 10:52,771,435-52,771,515
       MONDO (ID: 0013714), MedGen (ID: C3280586), and OMIM (ID: 614372)
    c. 16:23,619,946-23,620,026
       MONDO (ID: 0016419), MedGen (ID: C0006142), OMIM (ID: 114480), Orphanet (ID: ORPHA227535), and SNOMED_CT (ID: 254843006)

UT Southwestern Medical Center

(Exercises from workshop alignment by Dr. D. Kim)

Exercise 1 (2/2): Determining Likely Phenotypes from current literature.

1.Can you identify the most likely ethnicity?

- Please refer to Figure 3 in the following paper: Tao Huang et al. 2015, Genetic differences among ethnic groups
- When you search SNP location and allele frequency (e.g., rs6023406), use the dbSNP database available here

2. Can you identify the most likely eye color?

- Please refer to Table 1 in the following paper: Branicki et al. 2011, Model-based prediction of human hair color using DNA variants

3. Can you identify the most likely hair color?

- Please refer to Table 2 in the following paper: Spichenok et al. 2010, Prediction of eye and skin color in diverse populations using seven SNPs

(Exercises from workshop alignment by Dr. D. Kim)

UT Southwestern Medical Center

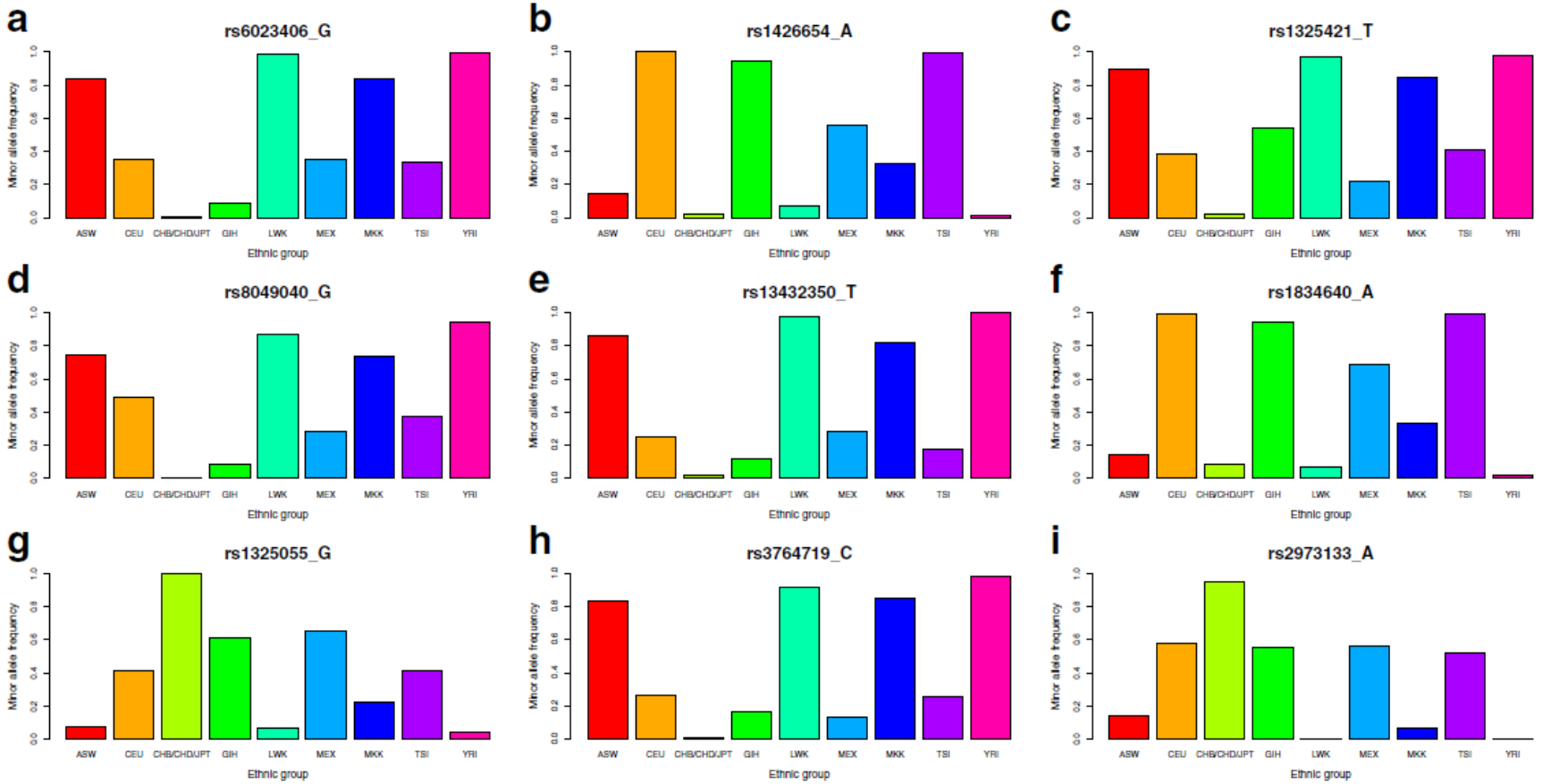


Fig. 3 The minor allele frequency of the top nine SNPs in each ethnic group. The minor allele frequencies of the top three SNPs, rs6023406 (**a**), rs1426654 (**b**), rs1325421 (**c**), rs8049040 (**d**), rs13432350 (**e**), rs1834640 (**f**), rs1325055 (**g**), rs3764719 (**h**), rs2973133 (**i**) in the nine ethnic groups were plotted. Each ethnic group has their own specific alleles. For example, the allele frequencies of rs6023406_G, rs1426654_A, rs1325421_T, rs8049040_G, rs13432350_T, rs1834640_A and rs3764719_C were very low, but those of rs1325055_G and rs2973133_A were very high in the Asian population (CHB/CHD/JPT)

Index	Abbreviation	Full Name	Training Sample Size	Independent Test Sample Size
1	ASW	African ancestry in Southwest USA	74	13
2	CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	140	25
3	CHB/CHD/JPT	Han Chinese in Beijing, China/ Chinese in Metropolitan Denver, Colorado/Japanese in Tokyo, Japan	305	54
4	GIH	Gujarati Indians in Houston, Texas	86	15
5	LWK	Luhya in Webuye, Kenya	94	16
6	MEX	Mexican ancestry in Los Angeles, California	73	13
7	MKK	Maasai in Kinyawa, Kenya	156	28
8	TSI	Tuscan in Italy	87	15
9	YRI	Yoruban in Ibadan, Nigeria (West Africa)	173	30
Total			1188	209

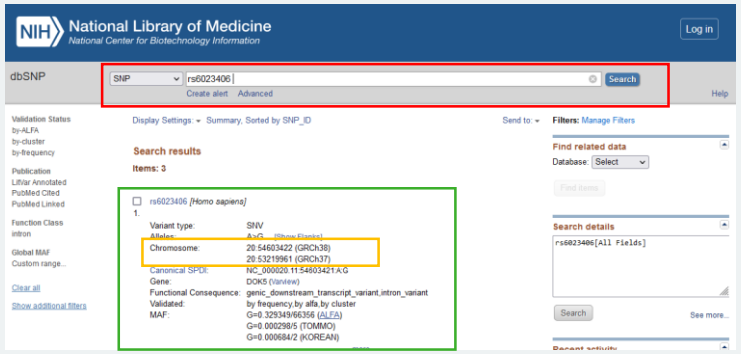
Huang, T., Shu, Y. & Cai, YD. Genetic differences among ethnic groups. *BMC Genomics* **16,** 1093 (2015). https://doi.org/10.1186/s12864-015-2328-0

Huang et al. Ethnicity SNP Exercise.

https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-2328-0

rs6023406

https://www.ncbi.nlm.nih.gov/snp



Search for SNP name, get info about location and variants! Including location. (chr20:54603422)

UT Southwestern Medical Center

Variant	Chr	Position	Gene	A	B	MAF	Color	OR	95% Lower CI	95% Upper CI	*P* val	Other
rs16891982	5	33987450	*SLC45A2*	G	C	0.02	Black	5.11	1.79	14.55	0.002	Yes
rs28777	5	33994716	*SLC45A2*	A	C	0.02	Black	7.05	2.23	22.25	0.001	Yes
rs26722	5	33999627	*SLC45A2*	G	A	0.02	Black	5.53	1.64	18.68	0.006	Yes
rs12203592	6	341321	*IRF4*	C	T	0.08	Black	2.35	1.22	4.54	0.011	Yes
rs9378805	6	362727	*IRF4*	A	C	0.45					0.103	
rs4959270	6	402748	*EXOC2*	C	A	0.46	Black	0.56	0.35	0.91	0.020	Yes
rs1408799	9	12662097	*TYRP1*	C	T	0.29					0.097	
rs2733832	9	12694725	*TYRP1*	T	C	0.40					0.177	
rs683	9	12699305	*TYRP1*	A	C	0.34					0.099	
rs35264875	11	68602975	*TPCN2*	A	T	0.23					0.158	
rs3829241	11	68611939	*TPCN2*	G	A	0.37					0.183	
rs2305498	11	68623490	*TPCN2*	G	A	0.27					0.230	
rs1011176	11	68690473	*TPCN2*	T	C	0.36					0.096	
rs1042602	11	88551344	*TYR*	C	A	0.28					0.255	
rs1393350	11	88650694	*TYR*	G	A	0.25	Brown	1.70	1.02	2.82	0.041	Yes
rs12821256	12	87852466	*KITLG*	T	C	0.09					0.052	
rs12896399	14	91843416	*SLC24A4*	G	T	0.44					0.064	
rs4904868	14	91850754	*SLC24A4*	C	T	0.46	Blond	0.64	0.43	0.97	0.037	
rs2402130	14	91870956	*SLC24A4*	A	G	0.16	D-blond	0.62	0.39	0.96	0.033	
rs1800407	15	25903913	*OCA2*	C	T	0.07	Red	3.23	1.07	9.76	0.038	
rs1800401	15	25933648	*OCA2*	C	T	0.06					0.105	
rs16950821	15	25957102	*OCA2*	C	T	0.12					0.170	
rs7174027	15	26002360	*OCA2*	C	T	0.12					0.146	
rs4778138	15	26009415	*OCA2*	T	C	0.16	Brown	1.80	1.02	3.19	0.043	Yes
rs4778241	15	26012308	*OCA2*	G	T	0.18					0.078	
rs7495174	15	26017833	*OCA2*	T	C	0.05					0.069	
rs12913832	15	26039213	*HERC2*	C	T	0.22	Black	3.33	1.99	5.57	4.3E−06	Yes
rs7183877	15	26039328	*HERC2*	C	A	0.07					0.124	
rs11635884	15	26042564	*HERC2*	T	C	0.01					0.135	
rs916977	15	26186959	*HERC2*	C	T	0.15	Red	0.34	0.18	0.65	0.001	Yes
rs8039195	15	26189679	*HERC2*	T	C	0.11	Red	0.30	0.14	0.64	0.002	Yes
MC1R_R	16		*MC1R*	wt	R	0.31	Red	12.64	7.03	22.74	2.5E−17	Yes
MC1R_r	16		*MC1R*	wt	r	0.20	Red	2.50	1.35	4.31	0.003	Yes
rs1805005	16	89985844	*MC1R*	G	T	0.08	Blond	2.99	1.52	5.86	0.001	Yes
Y152OCH	16	89986122	*MC1R*	A	C	0.00					0.982	
N29insA	16	89985753	*MC1R*	–	insA	0.01	Red	53.60	1.29	2221.72	0.036	
rs1805006	16	89985918	*MC1R*	C	A	0.00					0.476	
rs2228479	16	89985940	*MC1R*	G	A	0.09	Red	0.43	0.19	0.97	0.043	
rs11547464	16	89986091	*MC1R*	G	A	0.02	Red	3.35	1.04	10.76	0.042	
rs1805007	16	89986117	*MC1R*	C	T	0.11	Red	6.69	3.50	12.79	9.3E−09	Yes
rs1110400	16	89986130	*MC1R*	T	C	0.02					0.314	
rs1805008	16	89986144	*MC1R*	C	T	0.16	Red	5.69	3.31	9.78	3.2E−10	Yes
rs885479	16	89986154	*MC1R*	G	A	0.03	Blond	2.90	1.21	6.96	0.017	
rs1805009	16	89986546	*MC1R*	G	C	0.01	Red	31.85	2.61	388.28	0.007	Yes
rs1015362	20	32202273	*ASIP*	C	T	0.30	B-red	1.67	1.02	2.75	0.043	
rs6058017	20	32320659	*ASIP*	A	G	0.13					0.211	
rs2378249	20	32681751	*ASIP*	A	G	0.18	Red	2.34	1.14	4.82	0.021	Yes

MAF minor allele frequency, *Color* the most significantly associated color, *OR* the allelic odds ratio for the minor B allele, shown only if $P < 0.05$, *P* val the *P* value adjusted for age and gender, *Other* if the SNP is also associated with other colors with $P < 0.05$

Branicki et al. Hair-Color SNP Exercise.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3057002/

Probably a good idea to check those statistically significant variants first.

Branicki, Wojciech et al. “Model-based prediction of human hair color using DNA variants.” *Human genetics* vol. 129,4 (2011): 443-54. doi:10.1007/s00439-010-0939-8

UT Southwestern Medical Center

Spichenok et al. Eye-Color SNP Exercise.

Table 2
Eye color predictor.

Gene	SNP ID	Genotype	Eye color (predicted)
HERC2	rs12913832	G/G	Not brown
HERC2	rs12913832	G/A	Not blue
HERC2	rs12913832	A/A	Not blue
HERC2	rs12913832	A/A	Not blue
IRF4	rs12203592	T/T	
HERC2	rs12913832	G/A	Green
IRF4	rs12203592	T/T	
HERC2	rs12913832	G/G	Green
SLC45A2	rs16891982	C/C	
HERC2	rs12913832	A/A or G/A	Brown
SLC45A2	rs16891982	C/C	
HERC2	rs12913832	A/A or G/A	Brown
OCA2	rs1545397	T/T	
HERC2	rs12913832	A/A or G/A	Brown
MC1R	rs885479	A/A	
HERC2	rs12913832	A/A or G/A	Brown
ASIP	rs6119471	G/G	

https://www.sciencedirect.com/science/article/pii/S1872497310001717?via%3Dihub

Spichenok O, Budimlija ZM, Mitchell AA, Jenny A, Kovacevic L, Marjanovic D, Caragine T, Prinz M, Wurmbach E. Prediction of eye and skin color in diverse populations using seven SNPs. Forensic Sci Int Genet. 2011 Nov;5(5):472-8. doi: 10.1016/j.fsigen.2010.10.005. Epub 2010 Nov 2. PMID: 21050833.

UT Southwestern Medical Center

Resources on this section:

- IGV video Tutorials.
  - https://www.youtube.com/watch?v=E_G8z_2gTYM
- IGV Manual, and Installation Tutorial
  - https://software.broadinstitute.org/software/igv/UserGuide
- Broad Institute Website
  - https://www.broadinstitute.org/
- Papers for Exercises
  - Ethnicity -https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-2328-0
  - Hair-Color - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3057002/
  - Eye-Color - https://www.sciencedirect.com/science/article/pii/S1872497310001717?via%3Dihub
- Databases
  - NCBI - https://www.ncbi.nlm.nih.gov/
  - SNP - https://www.ncbi.nlm.nih.gov/snp/
  - SRA - https://www.ncbi.nlm.nih.gov/sra/

UT Southwestern Medical Center

Part 2: NGS RNA-Seq Example and Exercises with IGV

$(\approx 0.5 - 1\ hour)$

UT Southwestern
Medical Center

We can also use HISAT2 to align RNASeq data and produce SAM/BAM with samtools

The file name and path to the real **RNA** single-end sequencing reads for this workshop is:
session3/reads/RNA.heart.e11.rep1.fastq.gz

To examine the file, you can optionally perform the following commands:

1.Extract a compressed read file and redirect to a text file
```
gzip -cd session3/reads/RNA.heart.e11.rep1.fastq.gz > RNA.heart.e11.rep1.fastq
```

2.Display the text file; to stop the command, type *ctrl+c*
```
cat RNA.heart.e11.rep1.fastq
```

3.Display the first 10 lines of the read file
```
head RNA.heart.e11.rep1.fastq
```

4.Display the first page of the read file; type *space* to go to the next page, and type *q* to quit
```
less RNA.heart.e11.rep1.fastq
```

5.Display the first page of the read file directly on the compressed file using pipe |
```
gzip -cd session3/reads/RNA.heart.e11.rep1.fastq.gz | less
```

6.Count the number of lines in a read file
```
wc -l RNA.heart.e11.rep1.fastq
```

7.Divide the number of lines we calculated above by 4 to get the number of reads
```
expr 113425724 / 4
```

(Exercises from workshop alignment by Dr. D. Kim)

Linux commands from Dr. Kim For information on manipulating sequences with the commandline.

UT Southwestern Medical Center

We can also use HISAT2 to align RNASeq data and produce SAM/BAM with samtools

1.We will use HISAT2 with a graph index to align reads as follows:

```
session3/programs/hisat2 -p 8 -x session3/indexes/mouse/genome_snp_tran -U
   session3/reads/RNA.heart.e11.rep1.fastq.gz > heart.e11.sam
```

Align to reference (Mouse transcriptome)

(Exercises from workshop alignment by Dr. D. Kim)

Linux commands from Dr. Kim For information on manipulating sequences with the commandline.

1.Look at the SAM file (use *space* to go to the next page and *q* to quit)

```
less heart.e11.sam
```

2.Convert the SAM file into a BAM file

```
session3/programs/samtools view -@ 8
   -bS heart.e11.sam > heart.e11.unsorted.bam
```

3.Create a sorted BAM file

```
session3/programs/samtools sort -@ 8 heart.e11.unsorted.bam
   -o heart.e11.sorted.bam
```

4.Make an index for the sorted BAM file

```
session3/programs/samtools index heart.e11.sorted.bam
```

5.Look at the sorted alignments

```
session3/programs/samtools view heart.e11.sorted.bam | less
```

6.Look at alignments between 61,989,341 and 61,990,361 on Chromosome 15

```
session3/programs/samtools view heart.e11.sorted.bam 15:61,989,341-61,990,361
```

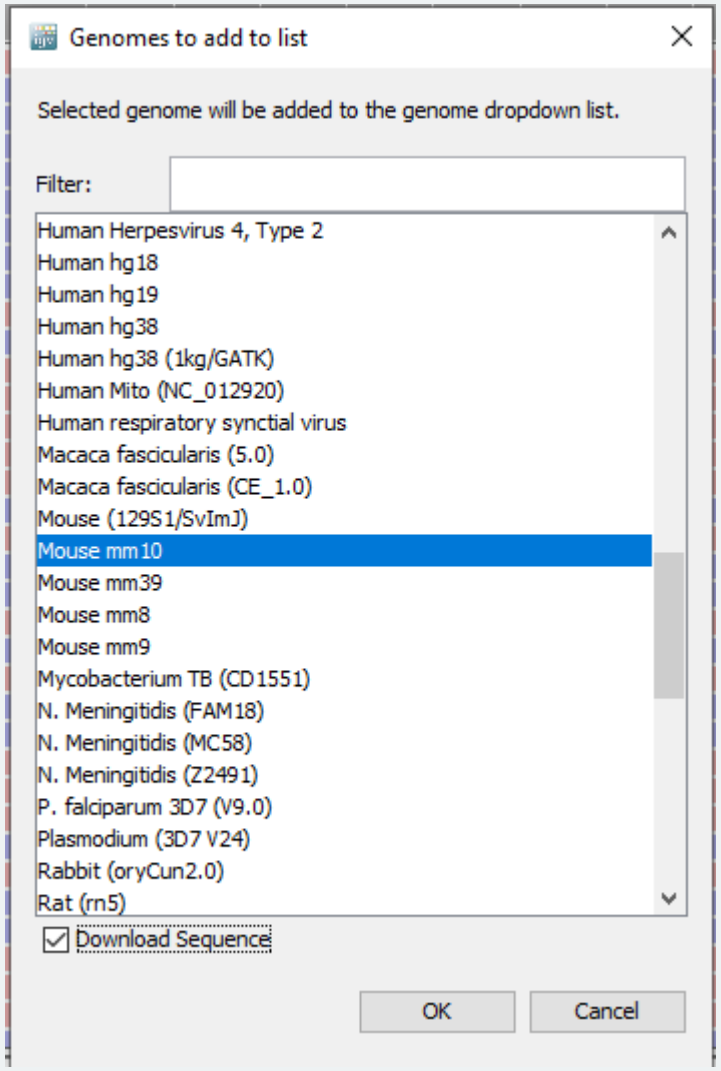
7.Look at bases of reads located at a particular locus to identify variants

```
session3/programs/samtools mpileup
   -f session3/indexes/mouse/genome.fa
   -r 15:61,989,341-61,990,361 heart.e11.sorted.bam
```

Use Samtools to produce binary sequence alignment map and index for bam file.

UT Southwestern Medical Center

Download the mouse Genome for displaying the RNA-Seq Alignment Results



Now we will look at the alignment using IGV (Integrative Genomics Viewer):

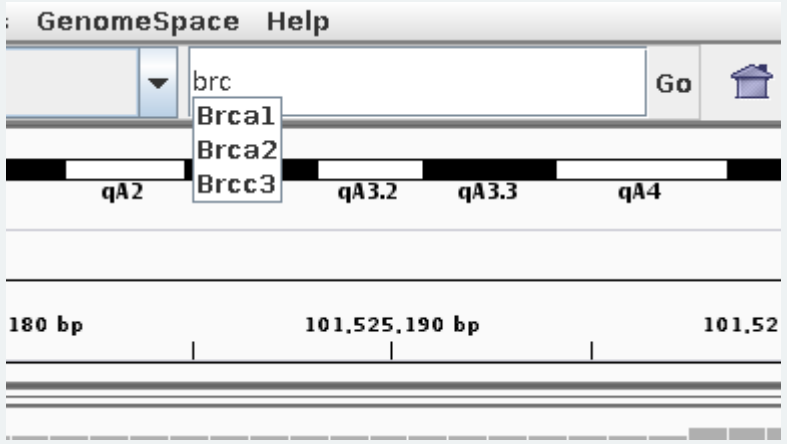
1.Run IGV

```
module add IGV/2.3.90
igv.sh
```

2.Load the genome using: Genomes -> Load from Server -> Select Mouse mm10

3.Open the BAM file using File -> Load From File in IGV and choose heart.e11.sorted.bam

Look through genomic regions for genes such as Brca1, Myc, and Cav1



UT Southwestern Medical Center

Lyda Hill Department of Bioinformatics 2022 Nanocourse Series
Introduction to NGS Analysis

Thank you for your attention & Interest!
Please do not forget to complete the course survey! ☺.

UT Southwestern Medical Center

Bonus: VCF File Format

1.1 An example

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

The Variant Call File File Format contains information about differences observed In a particular sample, these are often much smaller and contain much of the vital information to reconstruct an individual genome from a reference.

http://samtools.github.io/hts-specs/VCFv4.2.pdf

UT Southwestern Medical Center