**Visualizing Variants using the Integrative Genomics Viewer (IGV)**

Copyrighted © 2019Regie Santos-Cortez, Isabelle Schrauwen, Gao T. Wang, and Suzanne M. Leal

This exercise is meant to provide the student hands-on experience with IGV using exome data. While there are multiple ways to use IGV, our group primarily uses IGV to [a] visualize variants identified in .vcf, [b] check coverage of a specific gene or mapped region, and/or [c] search for variants that were potentially missed during variant calling including copy number variants.

A User Guide is available at <http://www.broadinstitute.org/igv/book/export/html/6>

For this exercise, the .bam files provided are limited to the genomic regions under study.

Start the tutorial by typing:

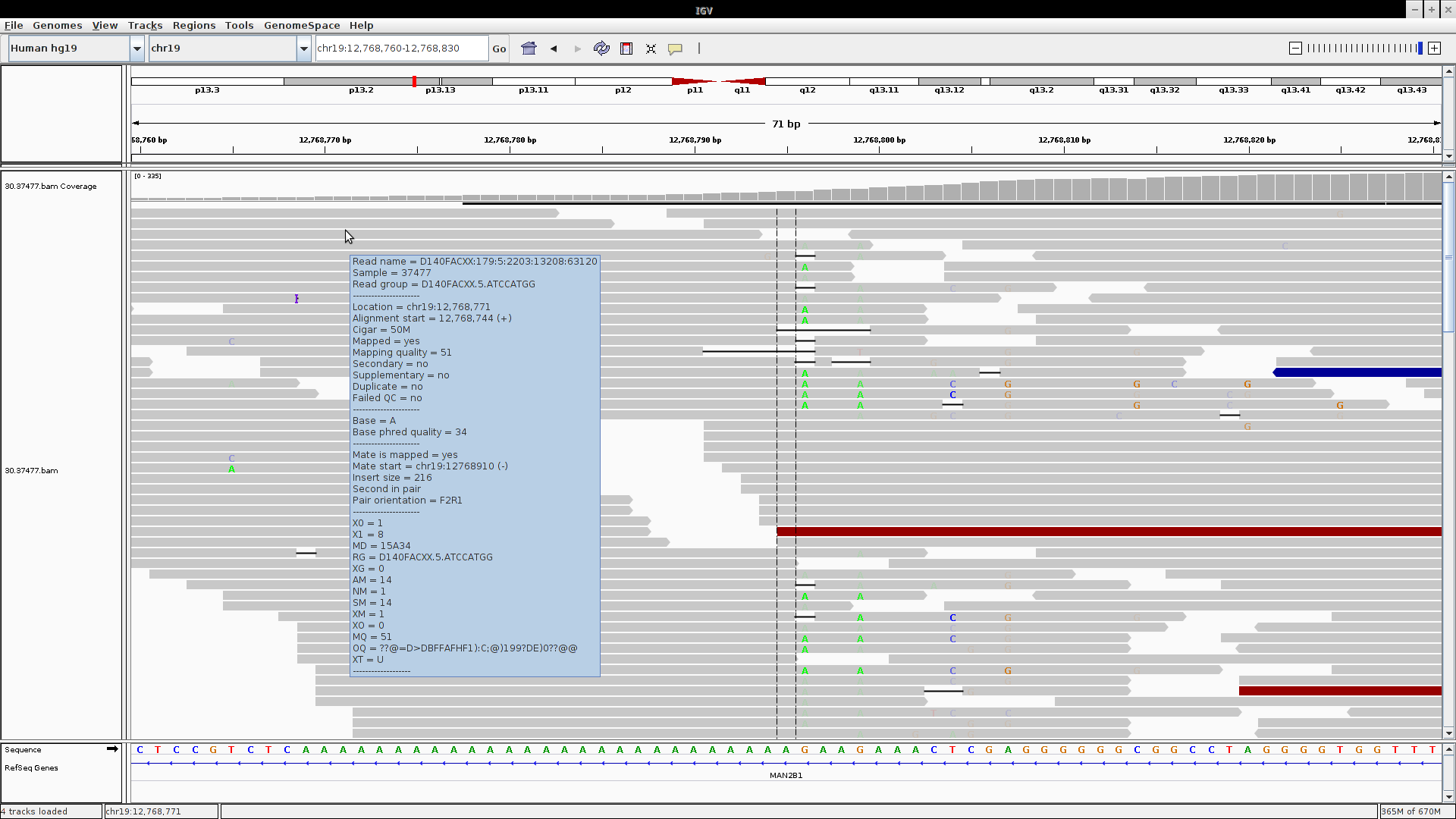
$ statgen-setup login --tutorial igv --my-name <my-name>

Now copy the http:// that you see in the output and paste it in the browser. This will lead you to an online and simplified version of IGV.

On the IGV window, bring the mouse pointer to the “Tracks” tab, then from the pop-up menu, choose “Load from URL”. In the data section write 4595\_19p.bam and in the index section write 4595\_19p.bam.bai. When the .bam file is loaded, the leftmost panel will show the ID/name of the .bam file. Under the tab “File”, there is another drop-down menu that lets you choose which genome reference to use: click the arrow and choose “Human hg19” (in this simplified online version it is already set to hg19 reference genome). IGV allows visualization of multiple files at once, but for this exercise we will only look at one .bam file at a time.

Family 4395 segregates autosomal recessive nonsyndromic hearing impairment that was mapped to chromosome 19pter-p13.12. The exome data in 4395.bam is from a hearing-impaired individual from family 4395. To visualize the mapped region, in the space provided on the second row to the left of the button that says “Go”, type in or copy-paste “chr19:1-15,640,972”, then press Enter or click on Go. You will see on the top panel the ideogram of chr19 for which the specified region is bounded by a red box. On the bottom panel you will also see the RefSeq genes and exons within the mapped region. For the middle panel, in order to show the exome data in greater resolution, you can slide the blue rectangle on the right uppermost panel towards the boxed ‘+’ sign. As you increase resolution, there will be less genes included in the lower panel and the reads per exon/base will start to show in the middle panel.

The screenshot on the next page is an example of what one would see in a typical .bam file using IGV at high resolution.



**1**

**4**

**3**

**2**

**6**

**5**

**7**

**[1]** The *upper panel* with vertical gray bars shows depth of coverage per variant site. Pointing the mouse arrow on each bar will open a pop-up box that shows the actual depth of coverage (number of reads) at the specific site, and the number and percentage of reads with a specific allele (see screenshot on page 3). The *black bar* below the vertical gray bars indicates sites where reads have been downsampled.

**[2]** An insertion was detected in a single read, as indicated by *purple* ‘**ƚ**’*.* When the mouse arrow is pointed on the insertion, a pop-up box shows the inserted bases. In this example, a single base A was inserted in the specific read.

**[3]** When the arrow is pointed at a specific read, a *pop-up box* appears with details on the specific read, e.g. read name, alignment start, mapping quality, base phred quality, insert size, etc.

**[4]** At high resolution the specific base or variant site that falls at the center of the IGV window is marked by *two vertical dashed lines.* When a variant site has an allele that is different from reference in more than 25% of reads, the alternate allele is indicated in a *different color.*

**[5]** A deletion is indicated by a *black horizontal bar within* *gaps* in a read.

**[6]** An insert size that is smaller than expected is represented as a *blue bar.*

**[7]** An insert size that is bigger than expected is represented as a *red bar.*

Note that in the previous screenshot, although there were several indicators flagging insertion, deletion or SNV, none of these were called, and variants at these sites are likely false-positives. If there is confidence in a variant, a good proportion of the reads would reflect the base change. Additionally the coverage panel would also indicate a homozygous or heterozygous variant depending on the proportion of reads with the alternate allele, which would be indicated by changes in color of the vertical bar at the specific variant site.

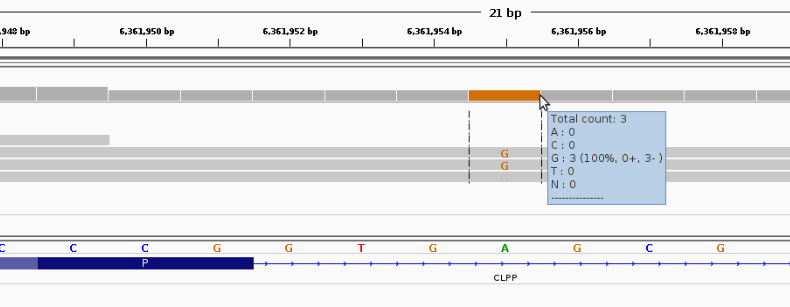
Within the mapped region on chr19, from the .vcf file there are several homozygous variants that are rare and predicted to be deleterious, listed below. Visualize each variant listed using IGV. To do this, go to the “Tracks” tab, select “Load from URL” and write 4395\_chr19.vcf under the data section, click Ok and you will see that the .vcf file will be loaded.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| hg19 chr19 position | Ref allele | Alt allele | Gene | Variant | GATK filter | ExAC All MAF<0.01 | Bioinformatics tools predicting deleteriousness |
| 12,763,267 | C | T | *MAN2B1* | p.Ile412Val | PASS | 0.0005 | CADD=15, MA, MT |
| 13,041,124 | C | T | *FARSA* | p.Arg139Gln | PASS | 0.0002 | CADD=22, MT, LRT |
| 14,030,642 | A | G | *CC2D1A* | p.Ile412Val | PASS | 0.002 | CADD=22, MA, MT, LRT, SIFT |
| GATK, Genome Analysis Toolkit; ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; CADD, scaled scores from Combined Annotation Dependent Depletion; MA, MutationAssessor; MT, MutationTaster; LRT, likelihood ratio test | | | | | | | |

When looking at each of the three variants listed in the table above, the variant quality is good with coverage depth of at least 28× at each variant site and >95% of reads with the alternate allele, indicating a homozygous variant. However, two other families were identified to co-segregate syndromic hearing impairment with variants in another gene, *CLPP*, which lies within the mapped region for family 4395 with ARNSHI. It was therefore possible that none of the three variants called from the exome are causal of hearing impairment, rather a *CLPP* variant was missed by variant calling. By reviewing the exome data within the *CLPP* gene using IGV, a non-canonical splice variant was identified:

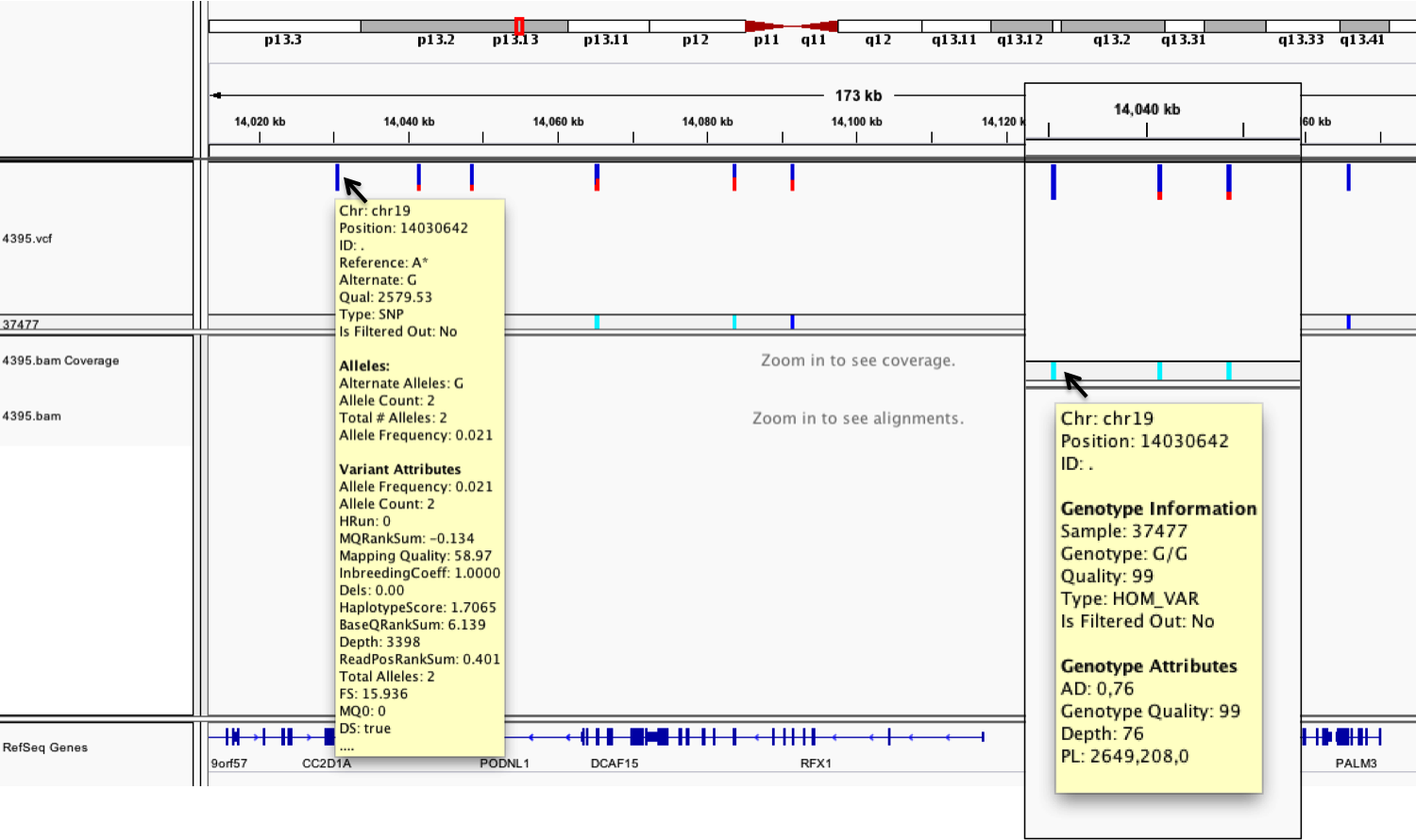
*CLPP* chr19:g.6,361,955A>G (c.270 +4A>G)

Looking at the coverage at this variant site, depth of coverage was only 3 and there were only 2 reads that were homozygous for the alternate allele. This might explain why the variant was not present in the filtered .vcf. Sanger sequencing using all available DNA samples from family 4395 confirmed co-segregation of the *CLPP* c.270 +4A>G with hearing impairment.

****

The pop-up window clarifies that there are 3 reads aligned to that position, all have a “G” at this position and are oriented in the same direction (-). Lighter colored bases indicate low quality bases.

It is possible to upload your VCF file at the same time to see which variants were actually called. This makes it easy to find potentially uncalled, wrongly called, or filtered variants. Go to “file” – “load from file”, and select the 4395.vcf file. We have zoomed in on the *CC2D1A* variant mentioned above (chr19:14,030,642). By hovering over the variant, more information is shown.

****

**2**

**1**

**[1]** Each bar across the top of the plot shows the allele fraction for a single locus.

**[2]** The genotypes for each locus in each sample. Dark blue = heterozygous, Cyan = homozygous variant, Grey = reference. Filtered entries are transparent.

**A screenshot of a cell phone

Description automatically generated**When looking back at our causal mutation in *CLPP* (chr19:6,361,955) mentioned above, this variant was actually called but filtered (transparent).

**Exercise:**

Load the exome file 4433.bam and corresponding 4433.vcf file. The hearing impairment in family 4433 was mapped to chr11:76,862,816-76,903,568. Within the mapped interval lies a known gene for Usher syndrome which includes as features retinitis pigmentosa and hearing impairment, i.e. *MYO7A.* Because there was no rare deleterious *MYO7A* variant detected in .vcf, we wanted to check if a variant was missed by the calling algorithms.

Question 1: **Is coverage for all exons of *MYO7A* complete?** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Hint: Transcript 1 (NM\_000260.3), the biggest of 3 RefSeq transcripts for MYO7A, has 49 exons. We have verified that the coding regions and canonical splice sites for exons 3-35 and exons 40-49 were covered by exome sequencing. To easily get the hg19 positions for specific exons, go to the UCSC Genome Browser:* <http://genome.ucsc.edu/>

Question 2: **Was there a *MYO7A* variant missed by exome sequencing? \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

*Hint: The exons we checked for coverage were negative for additional variants.*

Mutalyzer: <https://mutalyzer.nl/position-converter>

<https://mutalyzer.nl/name-checker>

Question 3: **If there was a missed variant, is this variant rare and damaging?** **Specify variant MAF and bioinformatic prediction.** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Genome Aggregation Database: http://gnomad.broadinstitute.org/

MutationTaster: <http://www.mutationtaster.org/>

**References:**

IGV – [1] Robinson JT et al. Integrative Genomics Viewer. Nat. Biotechnol. 29, 24-26 (2011); [2] Thorvalsdsdóttir H., Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178-192 (2015).

Family 4395 – Jenkinson EM et al. Perrault syndrome is caused by recessive mutations in *CLPP*, encoding a mitochondrial ATP-dependent chambered protease. Am. J. Hum. Genet. 92, 605-613 (2013).

**Answers:**

Question 1: **Is coverage for all exons of *MYO7A* complete?**

Answer: The 5’- and 3’UTR on exons 1 and 49 were not completely covered. It is common for UTR to be excluded from exome sequence data.

Question 2: **Was there a *MYO7A* variant missed by exome sequencing?**

Answer: This splice variant on exon 37 was not called in the .vcf but was identified in the .bam using IGV –

*MYO7A* chr11:g.76914265A>G

NM\_000260.3: c.5326+3A>G

Number of reads at the variant site is 7, which might explain why the variant was missed. Aside from UTR variants non-canonical splice variants can also be easily missed either by exome sequencing or variant calling. This variant was Sanger-sequenced using all DNA samples available from family 4433 and was found to co-segregate with hearing impairment.

Question 3: **If there was a missed variant, is this variant rare and damaging?** **Specify variant MAF and bioinformatic prediction.**

Answer: The variant has zero frequency in gnomAD. A large majority of gnomAD individuals have <10x coverage at the variant site. MutationTaster predicts the splice variant to be disease-causing.