Workshop Part I NGS basics

Christian Gilissen (Christian.gilissen@radboudumc.nl)
Galuh Astuti (Galuh.astuti@radboudumc.nl)
Department of Human Genetics, Radboud University Nijmegen Medical Centre

READ ME FIRST

- Actions you need to perform are indicated in **bold**
- Questions are designated by "**Q**" and are in *italics*.
- Screenshots are just examples, your specific output may look slightly different.
- At the end of this manual there are quick reference guides attached that can help you during this workshop.
- For this workshop you require:
  - This manual
  - Files in following folders:

```
1.NGS_basics
      |--- data
            |--- bam_files
```

  - Software:
    - IGV
- You can ask questions during the workshop!

Contents

Introduction

The purpose of this workshop is to explain to you the basics of Next Generation Sequencing with a focus on DNA sequencing. In this workshop you will familiarize yourself with the basic terminology, concepts and file formats in NGS.
During sequencing, the DNA nucleotide bases order is determined. A 'read' represents a sequence of nucleotides of each fragment in the library. NGS technologies can generate a massive number of sequence reads in a single experiment. However, no sequencing technology is perfect, and each instrument will generate different types and amounts of errors, such as incorrect nucleotides being called. These wrongly called bases are due to the technical limitations of each sequencing platform.

Therefore, it is necessary to understand, identify and exclude error-types that may impact the interpretation of downstream analysis. Sequence quality control is therefore an essential first step in your analysis.

All the raw sequencing reads are stored in fastq file format, which typically end with ".fq" or ".fastq". Fastq files have a specific format where every read contains 4 lines with the following information.

Line	Description
1	Always preceded by @ sign followed by the read/ sequence identifier
2	The actual DNA sequence
3	Always begins with a + sign followed by the sequence identifier **Often contains the same info as line 1*
4	Contains quality scores of each base. The number of quality values should match the number of bases of line 2. The quality values are denoted as ASCII characters where each character denotes a number indicating the quality of the base call in the sequence. Higher numbers indicate higher qualities. **Note:** You can look up the actual value of a character in an ASCII table (https://www.asciitable.com/)

Table 1. Information stored in a fastq file format

Note: Typically fastq files are compressed to keep them smaller and require less diskspace. The compressed version of the file is a filename that ends with ".fq.gz" or ".fastq.gz". These files can be decompressed (i.e. converted back to a normal fastq file) by the "gunzip" program which is installed on most Linux systems.

Part I: Exome sequence alignments

When the reads are of sufficient quality they can be aligned/mapped to a reference sequence to identify their origin. The most commonly used aligner is the Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/). Reads are aligned to the human reference genome, e.g. GRCh37 human genome assembly. Alignments of reads are typically stored in a sequence alignment map (SAM) format or its compressed version in bam or cram format. Alignment data can be viewed using the Integrative Genome Viewer (IGV).

Open your browser and go to: https://igv.org/app/. Load genome assembly by clicking 'Genome' and select Human (GRCh37/hg19).

Note: IGV loads Human (hg19) by default, but if you work with another version of the human genome, or another organism, you can change the genome by clicking the drop down menu in the upper-left.

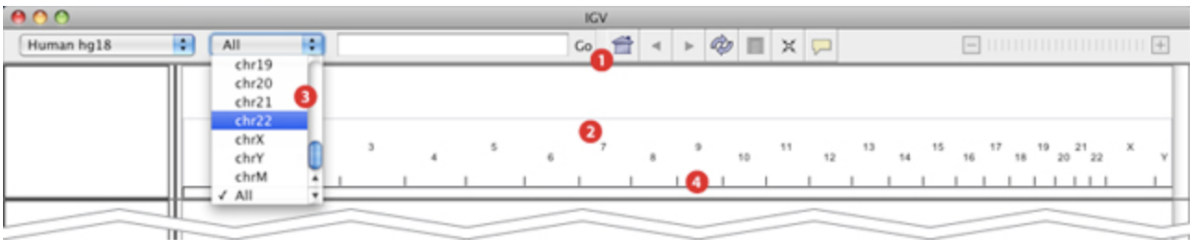


Figure 5. Navigating the view in IGV.

You will now be working with data from different sequencing platforms.

Open bam file by clicking 'Tracks' from IGV dropdown menu, select 'Local File' and load "1.NGS_basics/data/bam_files/hiSeq.bam" and "1.NGS_basics/data/bam_files/hiSeq.bam.bai" (see Figure 5).

This file only contains the sequence alignments to the (randomly chosen) *KMT2D* gene, to keep the file size small. A file containing the alignments to the entire exome would be roughly 10Gb in size, depending on the sequencing. The reads in this file were generated using an Illumina sequencer.

Note: IGV also requires an index file with the extension "**`.bam.bai`**". This file can be generated from the bam file using samtools and helps to speed up viewing in IGV.

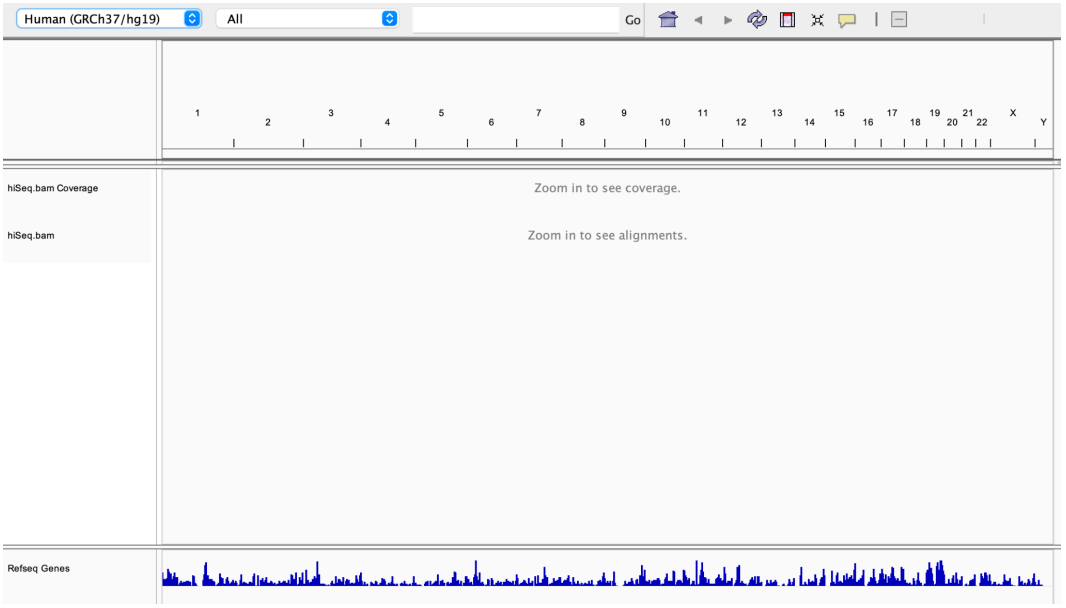


Figure 6. IGV with loaded bam file

Note: Notice the two track names on the left. For each bam file, IGV always displays two tracks: a coverage track (top) and the sequence alignment (bottom). Both only become visible after zooming in to a sufficient degree.

To go to a certain position or gene, type the position or gene name in the box on top.

Go to KMT2D gene and click '+' to zoom in.

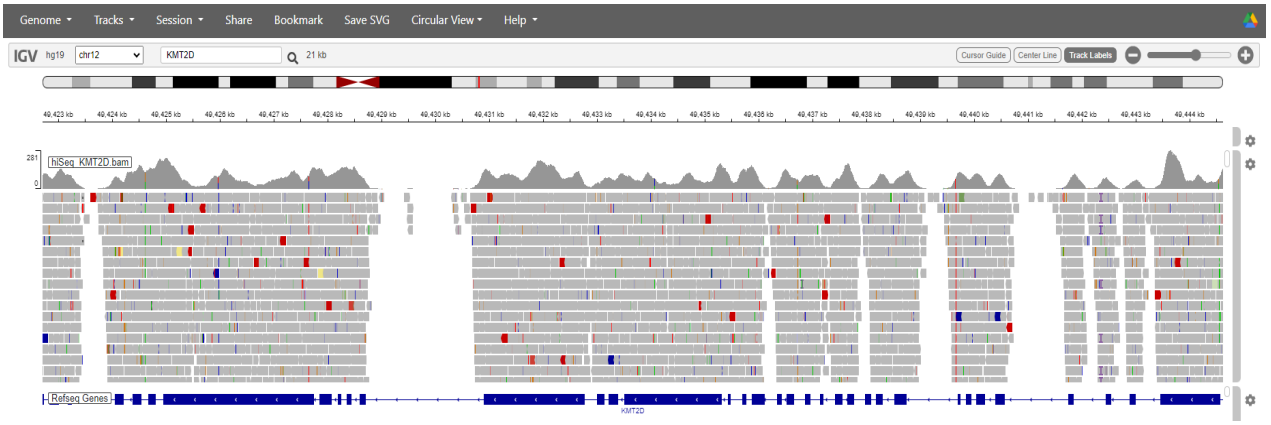


Figure 6. IGV view on KMT2D gene

Q7. *How do you recognize that this is exome sequencing data?*
(Hint: check the gene structure.)

You can zoom in/out by using the zoom bar on the top-right.

Zoom in to the single base pair level.

You can see the sequence of the reference genome down below. Bases matching the reference are not shown by default.

Right click on track (gray bars) and select the option "show all bases".

Now all bases including those that are the same as the reference are shown.

Turn this option off again.

In the search box type: chr12:49,444,545 to go directly to an SNV position. Right click on the reads, select "Color alignments by" and choose "read strand".

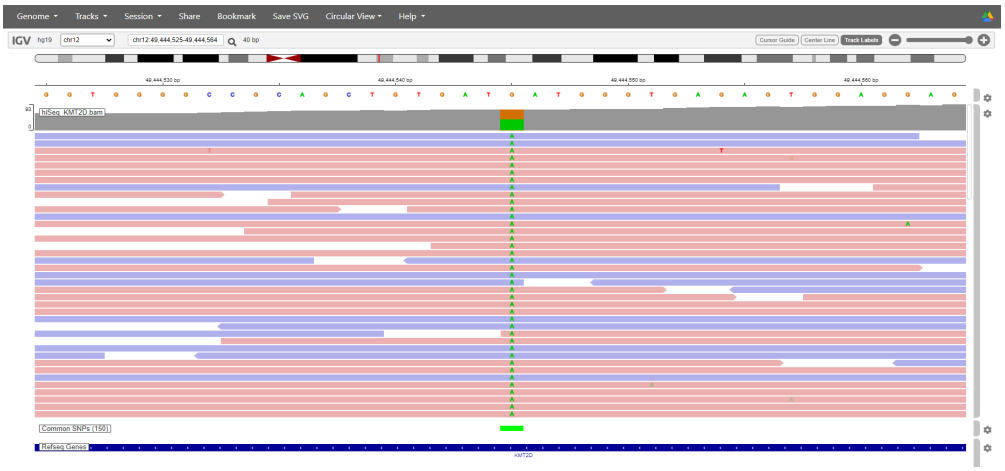


Figure 7. IGV view with color by read strand

Turn this option off again.

Q8. *What are the variant allele frequencies of this position and the variant is in homozygous or heterozygous state? homozygous*

(Hint: Count the total reads covering this position and how many reads showing a different base than the reference.)

Q9. *Is it a true variant? How does 'color by read strand' help the interpretation?*

Go to position chr12:49,424,243 to go directly to an SNV position. Right click and select sort alignments by base.

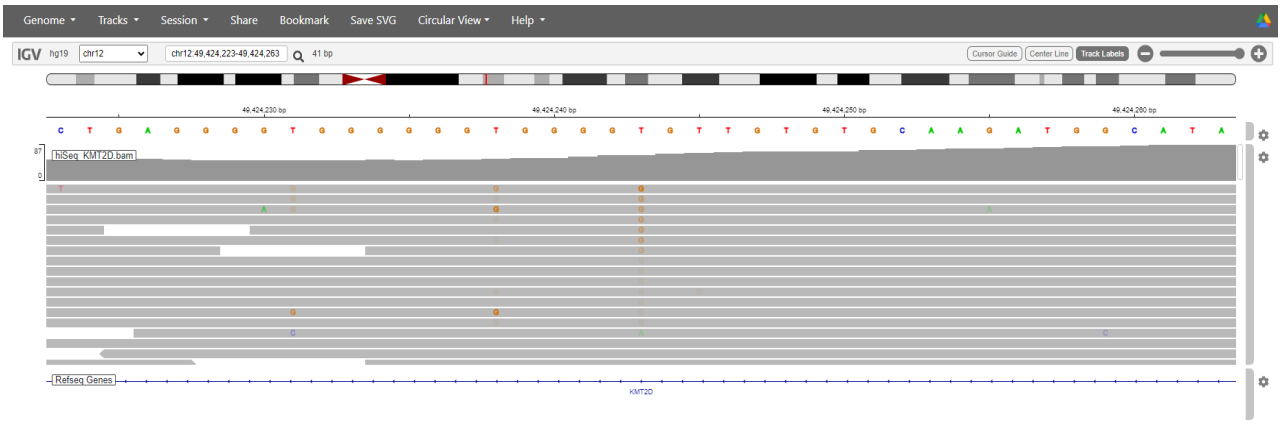


Figure 8. IGV view on chr12:49,424,243

Q10. *Is it a true variant? Why does the G show different shading?*

Navigate to region chr4:74,019,286-74,026,493. Click Tracks setting button and select 'View as pairs' and 'Color by pair orientation and insert size (TLEN)'.

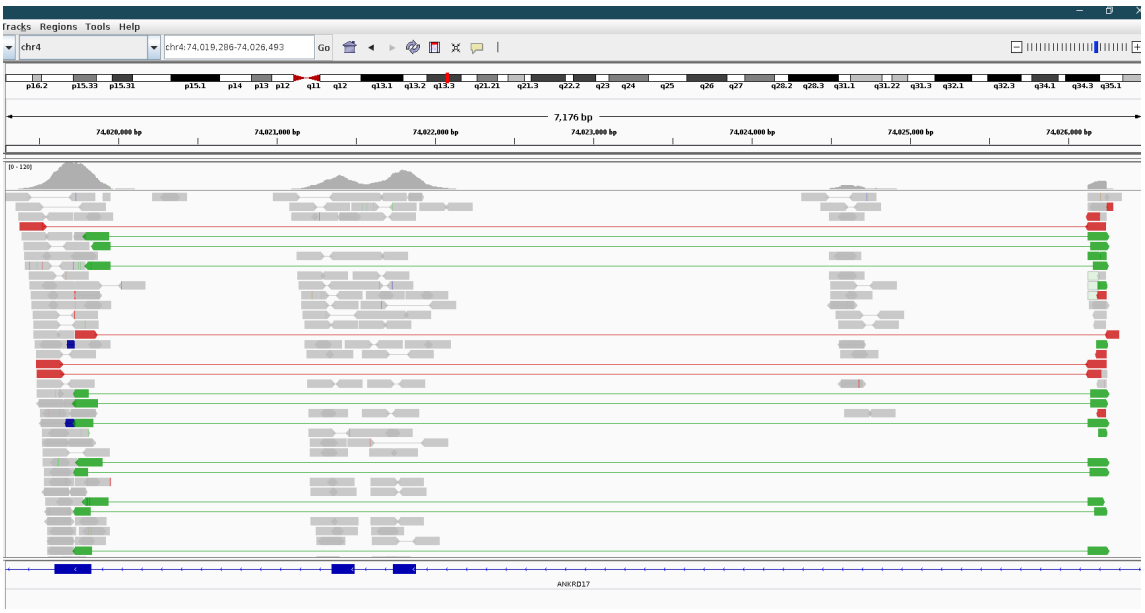


Figure 9. IGV view on ANKRD7 gene

Q11. *What do the red and green reads mean?*

Q12. *What kind of genetic aberration(s) do you observe in this sample?*

Go to the track setting and click 'View as pairs' and 'Color by pair orientation and insert size (TLEN)' to turn them off.

In the same bam file, navigate to chr6:29,857,060-29,857,419.

Figure 10. IGV view on HLA-H gene

Q13. *What do the transparent reads mean and how would you interpret it?*
(Hint: Check the mapping quality of transparent reads and compare with the gray reads.)

Navigate to chr4:73,984,549-73,984,660. Right click on the track and select sort by base.

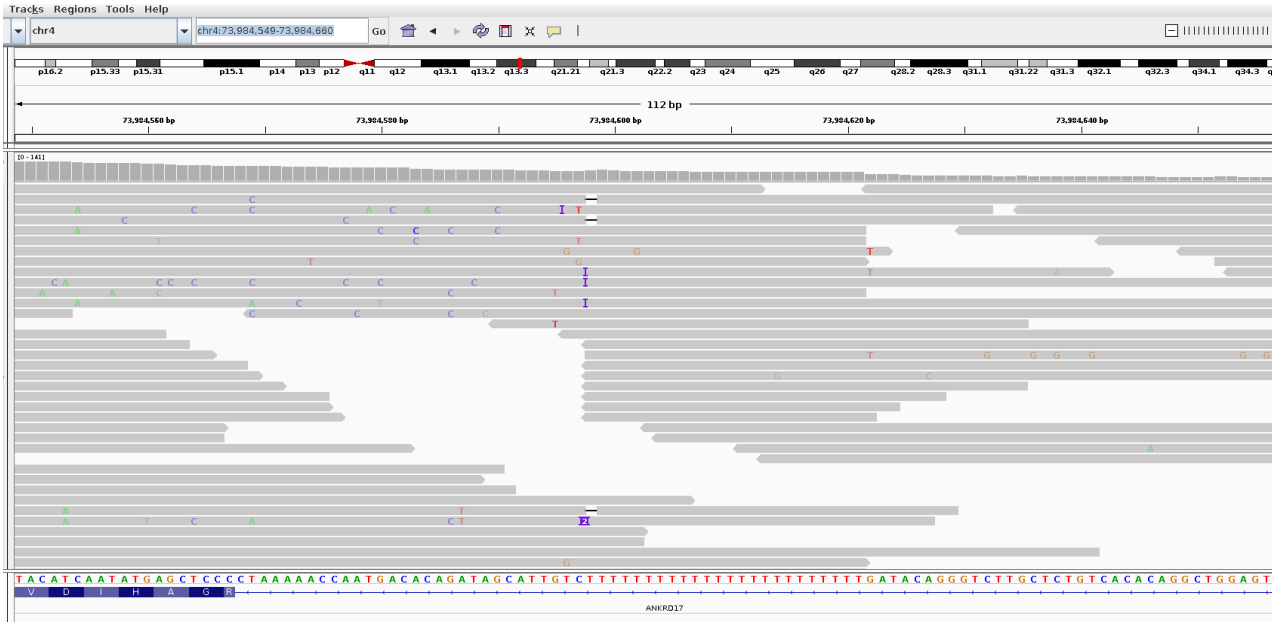


Figure 11. IGV view on ANKRD7 gene

Q14. *What do the purple* I *symbols represent?*

Q15. *What is the possible cause of this artefact?*

Navigate to KMT2D gene, zoom in and right click on the track with the hiSeq data and select the option "View as pairs". (Optionally also choose the option to sort the alignments by start location.)

This shows you that this is paired-end data, i.e. the two ends of a single DNA fragment is being read with a gap in between. The size of the gap is called the insert size.

Q16. *Can you estimate what the insert size is?*

Q17. *What do you think might be an advantage of having paired-end reads?*

Part II: Genome sequence alignments

Now also open the file "`1.NGS_basics/data/bam_files/novaSeq.bam`" in IGV.

This is sequencing data of the same region using the Illumina NovaSeq whole genome sequencing.

Note: These data are not from the same sample!

Now right click in both tracks and change the visualization from "expanded" to "squished".

Note: Changing the visualization can help by making nice screenshots of your variant or region of interest. The view you choose is just a personal choice.

Q18. *Compare the coverage tracks of both bam files, what differences do you see?*

(Hint: zoom out until you see gaps between the exons in the lower panel that says "RefSeq Genes".)

Scroll down to see the new alignment and zoom in to the single base pair level again.

Q19. *What can you say about the evenness of coverage? Which platform covers all of the coding sequence sufficiently for variant calling (e.g. 15x)?*

Part III: Long read sequence alignments

You have been looking at short read sequencing data up till now. Short-read sequencing is cost-effective and accurate, and is supported by a wide range of analysis tools and pipelines. However, as you may imagine, genetic variation that spans multiple consecutive bases (instead of a single base) is more difficult to detect with short read sequencing techniques. Long-read sequencing, or third-generation sequencing, allows for the detection of these larger structural variants (SVs), by improving the mapping accuracy.

Now also open the file "1.NGS_basics/data/bam_files/lrs.bam" and "1.NGS_basics/data/bam_files/lrs.bam.bai" in IGV.

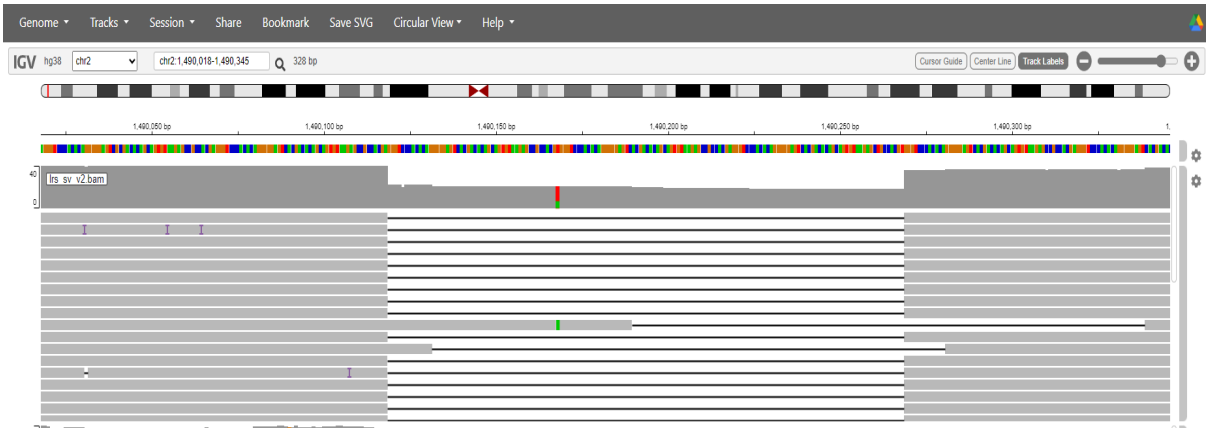
Q20. *First scroll down to see the new alignment, what can you say about the length of the reads?*
(Hint: When you press over a read, you can see detailed information about this read.)

Now right click in all three tracks again and change the visualization from "expanded" to "squished".

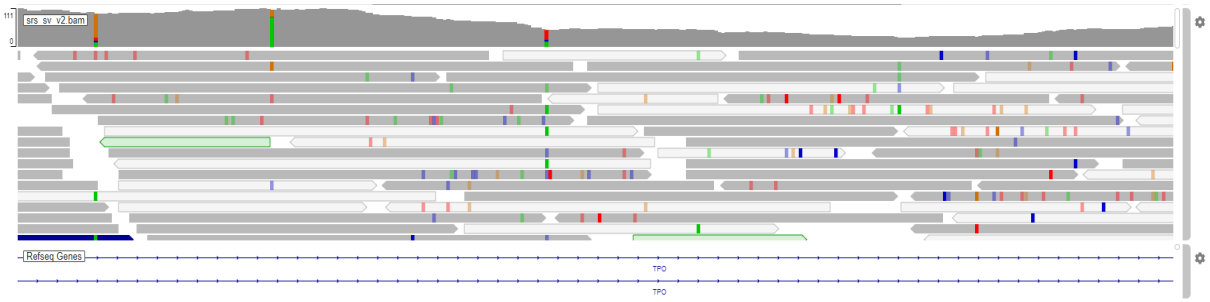
Q21. *What can you say about the amount of noise in the long-reads as compared to the short-reads? Which technique do you think is more suitable to identify SNVs?*
Q22. *Aside from SVs, what other use case can you think of where it would be more suitable to use long read sequencing?*
(Hint: Consider compound heterozygosity.)

Refresh the IGV web and load Human genome assembly (GRCh38/hg38). Open bam file by clicking 'Tracks' from IGV dropdown menu, select 'Local File' and load "1.NGS_basics/data/bam_files/lrs_sv.bam" and "1.NGS_basics/data/bam_files/lrs_sv.bam.bai". Navigate to chr2:1,490,020-1,490,350

Q23. *What kind of genetic aberrations can you observe in this region?*



Now load "1.NGS_basics/data/bam_files/srs_sv.bam" and "1.NGS_basics/data/bam_files/srs_sv.bam.bai". These two bam files are from the same individual.



Q24. *Why can't we see the same variant from this srs_sv.bam file?*

Quick reference guide to Next Generation DNA Sequencing

DNA sequencing is the process of determining the exact order of the bases A, T, C and G in a piece of DNA. Sanger sequencing is inherently a one-at-a-time technology - it generates a single sequence read of one region of DNA at a time. This works well in many applications. If you're sequencing one gene from one sample, Sanger sequencing is reliable, and generates long sequence reads that, under the right conditions, can average well over 500 nucleotide bases.

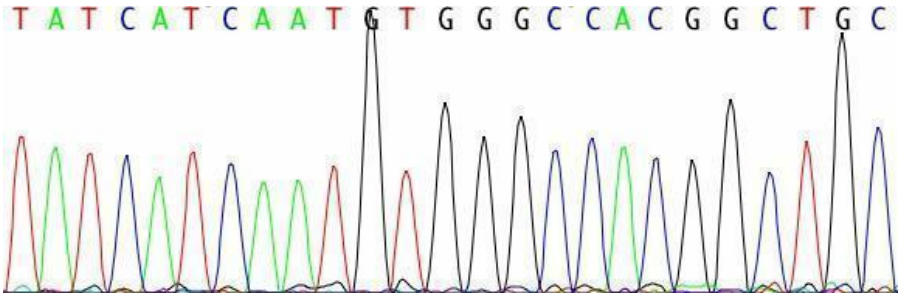


Figure Qref-NGS 1. Sanger sequencing chromatogram.

The high demand for low-cost sequencing has driven the development of high-throughput sequencing technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. High-throughput sequencing technologies (or Next generation sequencing (**NGS**)) are intended to lower the cost of DNA sequencing beyond what is possible with standard methods. Next-generation sequencing can't generate the nice, long sequence reads you get with Sanger sequencing, nor are the individual reads as accurate. Instead of 500 DNA bases or more, you just get about 35 -100 bases (depending on the exact instrument). But the difference is that you get lots and lots of sequence reads. Instead of just one long read from just one gene (or region of the genome), you get thousands of short, error-prone reads, from hundreds or thousands of different genes or genomic regions. Why exactly is this better? The individual reads may be short and error prone, but as they add up, you get accurate coverage of your DNA sample; thus you can get accurate sequence of many regions of the genome at once. The plot below shows a small region that was sequenced with 50bp reads. In the middle you can see a variant that was detected in a heterozygous state.

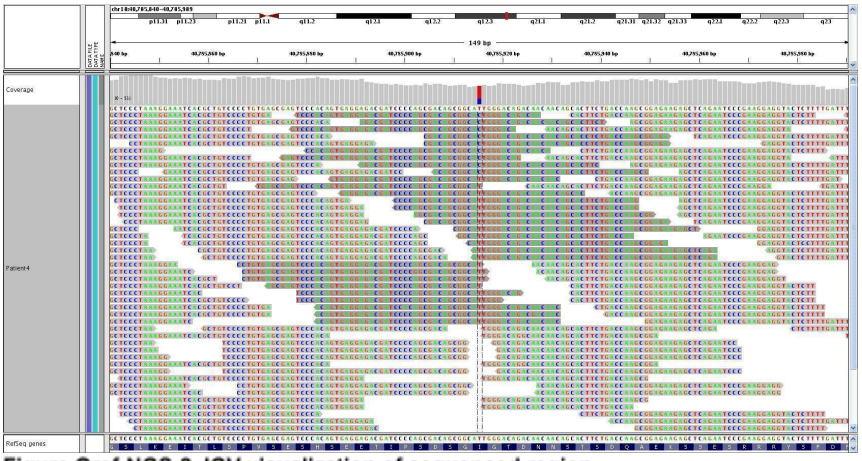


Figure Qref-NGS-2. IGV visualization of sequenced region.

Quick reference guide to IGV and .bam files

Bam files are binary files (faster access, smaller, but not human readable) that contain the alignment of reads to a reference. Bam files are the typical output of an aligner or mapper. You can extract data from a .bam file using software called **samtools** (http://samtools.sourceforge.net/). As an example you can convert a .bam file to a .sam file which is a (human readable) text file representation of the .bam file. Samtools also allows you to do other kinds of operations, like generate statistics, variant calling, generating pileups and others.)

The Integrated Genome Viewer (IGV) is a viewer which can view .bam files (see Figure Qref-NGS 2 on the previous page). To load a .bam file you also need a .bam.bai (index) file which can be generated for a .bam file using samtools. You can load a .bam file in IGV by dragging the file in the window. Select the appropriate reference genome and choose the coordinates (or gene name) that you are interested in. You can add additional annotation as well from the menu options.

Quick reference guide to the UCSC genome browser

The UCSC genome browser is an online website which allows you to look at the current assembly of the human genome. It combines many different sources of information to tell you something about the parts of the genome that you are looking at. E.g. whether there is a gene located at the position, or a known variant etc. These sources of information are called tracks.

Open a new window in your internet browser and go to URL: https://genome.ucsc.edu/ and click on the “Genome Browser”.

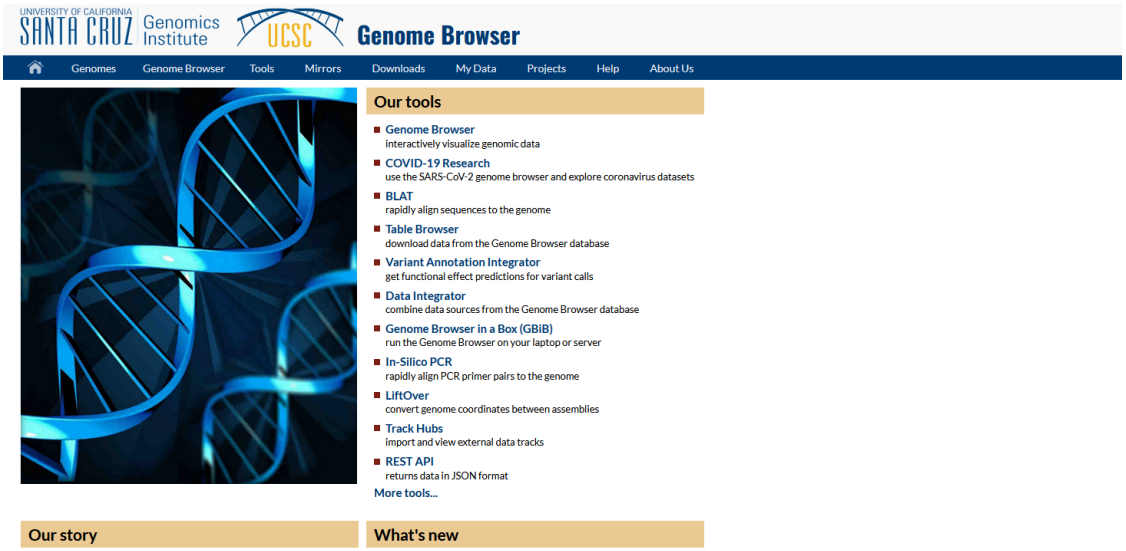


Figure Qref-UCSC-1. UCSC genome browser.

Note: You may be prompted to choose a mirror, if so choose the genome-euro.ucsc.edu mirror for faster access from within Europe.

Make sure you select the correct genome (human) and assembly of the genome (hg19), all the files that we use are based upon version hg19 of the human reference genome.

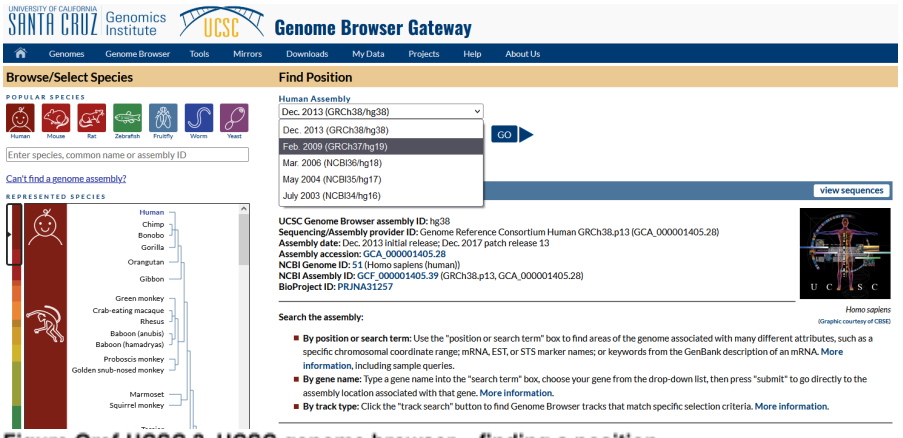


Figure Qref-UCSC 2. UCSC genome browser – finding a position.

In the box beneath the text “Position/Search Term”, you can enter a gene name, gene symbol or a genomic position. By pressing the “Go” button, your position or gene will be displayed in the browser on the specified assembly.

All the tracks are displayed below the browser window:

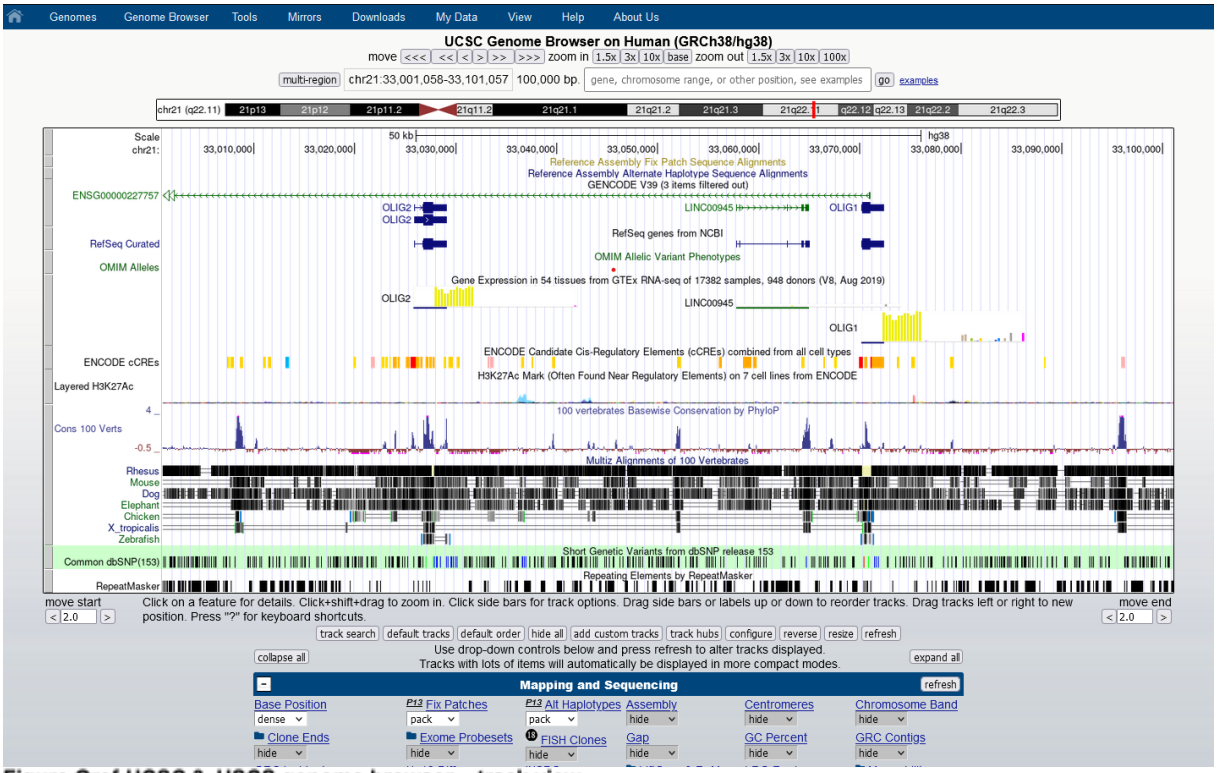


Figure Qref-UCSC 3. USCS genome browser – track view.