Alignment Workshop

Cesky Krumlov

May 25, 2022

The ~~Ugly~~ Mess Behind the Curtain

Before we start this, I want to tell a brief story about getting the data ready.

I wanted something small enough to align a whole genome that had Illumina, ONT, and PacBio HiFi data, that would have some complex genomic changes and also be heterozygous [failed at this]. The closest to this I could find was a project looking at three yeasts with MGI, ONT, and HiFi.* The data you will work with are downloaded from NCBI's Sequence Read Archive (SRA).

* Benchmarking of long-read sequencing, assemblers and polishers for yeast genome. Zhang *et al. Briefings in Bioinformatics*, Volume 23, Issue 3, May 2022, bbac146, https://doi.org/10.1093/bib/bbac146

The ~~Ugly~~ Mess Behind the Curtain

1. Went to SRA website and searched project IDs
2. Noted all the run level ids
3. Downloaded and installed the SRA toolkit
4. Use toolkit to download the SRA format files
5. Extract fastq/fasta files from SRA format
6. Wrote perl script to downsample all the files to ~35x genome coverage

Find the Data

Find the data we’re going to use for this workshop. It’s in your home directory

```
cd
```

From there you can go to where the data are:

```
cd workshop_materials/alignment
```

Look at what we have:

```
ls
```

There are 3 subdirectories, one has the reference genome and annotations, and the other two have data from two different yeasts.

The Yeast Data

The data we are working with come from the model yeast, *Saccharomyces cerevisiae*. We will mostly look at the sequencing data of one strain, CICC-1445, and compare it to the finished sequence of the reference strain, S288C. There is also sequencing data from the reference strain itself, which you can optionally align. We have data from three sources: MGI, ONT, and PacBio HiFi. We will align all three.

Optional: Make Data Read Only

Before we start working, you may want to protect yourself against mistakes. You will need to read, but not write to, all of the files in this directory during the exercise. To make sure you don’t accidentally delete these data, you can make them read only:

```
chmod 444 reference/* CICC-1445/*/* S288C/*/*
```

Decide Where to Work

It is generally good practice to make a directory separate from your raw data to run analyses. You don’t have to do this, but if you want to, you are on your own to do this. Throughout this exercise, the full paths to files will not be given. The raw data will always be in /home/genomics/workshop_materials/alignment (~/workshop_materials/alignment). Any files you create will be wherever you put them, and I will refer to them as, e.g., *output.sam*, but you should name this file something that is meaningful to you and distinct from other files you are creating.

Also, **do not copy/paste the commands from this document**. They will not work. You will have to form your own commands. Remember that you can use tab completion to find and confirm files that already exist.

Prepare to Run BWA

Now we can start working with the data. First, we will align the Illumina data using the program *bwa*. The bwa program should already be in your $PATH, so you should just be able to type the command and it should work.
In order to do alignments, bwa requires a special index (the Burrows-Wheeler transform of the data), so we start by making that:

```
bwa index –a bwtsw reference/S288C.fa
```

This should take about 15 seconds. After you are done, there should be 5 additional files in the alignment/reference directory of the form “S288C.fa.*”. The S288C.fa, the actual fasta sequence, is the name prefix, and the extensions are all the various pieces of the index. Note that you will never directly reference these files. You always specify the genome as S288C.fa, and bwa (or any other program using an index) will know how to find all of its index files based on that. However, if they are not there (or not matched), the program will fail.

Run BWA

Now we can run bwa. We’re going to use the “mem” algorithm to do the alignments:

```
bwa mem reference/S288C.fa CICC-1445/MGI/SRR17381667_35x_1.fastq CICC-1445/MGI/SRR17381667_35x_2.fastq > output.sam
```

This will take 8-10 minutes to run. Once you make sure it’s really running, you can take a short break.

Look at the Files

Now we can look at these files and see what’s in them.

We will all do this part together and I will walk you through what is in the files.

```
less reference/S288C.fa
```

```
less CICC-1445/MGI/SRR17381667_35x_1.fastq
```

```
less output.sam
```

Convert Output to Binary

Our next step is to convert the sam file into binary format. We will use samtools to do this.

This should take about 2 minutes to run. Note that it runs in 2 steps.

```
samtools sort –o output.bam output.sam
```

```
samtools index output.bam
```

Align the ONT Data with Minimap2

Now we are going to align the ONT reads. The ONT reads are in CICC-1445/ONT/SRR17382760_35x.fasta

```
minimap2 –a –o output.sam S288C.fa CICC-1445/ONT/SRR17382760_35x.fasta
```

These should take about 5 minutes to align.

Now convert these into sorted bams also using samtools.

Align the HiFi Data with Minimap2

Now we are going to align the HiFi reads. The HiFi reads are in CICC-1445/PB/SRR18210299_35x.fasta

```
minimap2 –a –o output.sam S288C.fa  CICC-1445/PB/SRR18210299_35x.fasta
```

These should take about 5 minutes to align.

Now convert these into sorted bams also using samtools.

Optional Extra Exercise

If you have gotten here and still have some extra time and would like more practice with aligning, there is a second set of data. The same paper also used sequence of S288C, the reference strain. In theory, this should be uninteresting, because all the reads should be identical the reference. If you just want to practice running the alignment steps, you can look in the S288C directory in the alignment directory and find parallel MGI, ONT, and HiFi datasets for S288C which you can also align.

Starting IGV

We will spend the rest of the time looking at alignments. For this we will use a tool call IGV (the Integrative Genomics Viewer). To launch IGV, you should be able to type igv.sh. It will pop up a new window, so if you launch it from the command line, you can place it in the background to free that terminal window.

Loading the Genome

We need to prep some data first.

Load the genome (Menu->Genomes->Load Genome from File…) S288C.fa from the reference directory. This should give you a graphical layout of the chromosome lengths at the top of the screen. We also want the annotations, so go to Menu->File->Load from File… and load reference/saccharomyces_cerevisiae.gff.

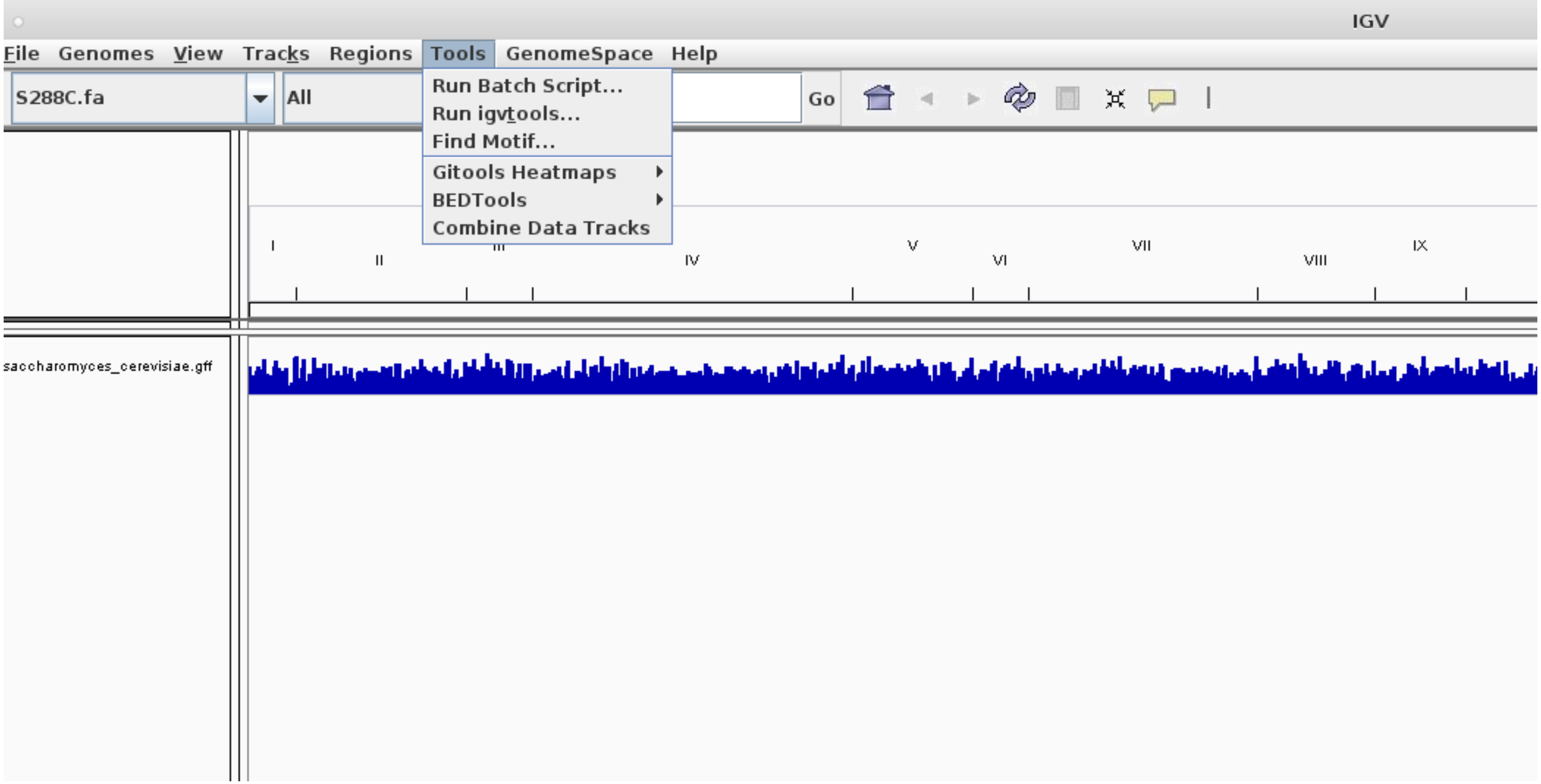
You will probably get some warning messages, but just persist and it will load properly in spite of them.

Computing Coverage

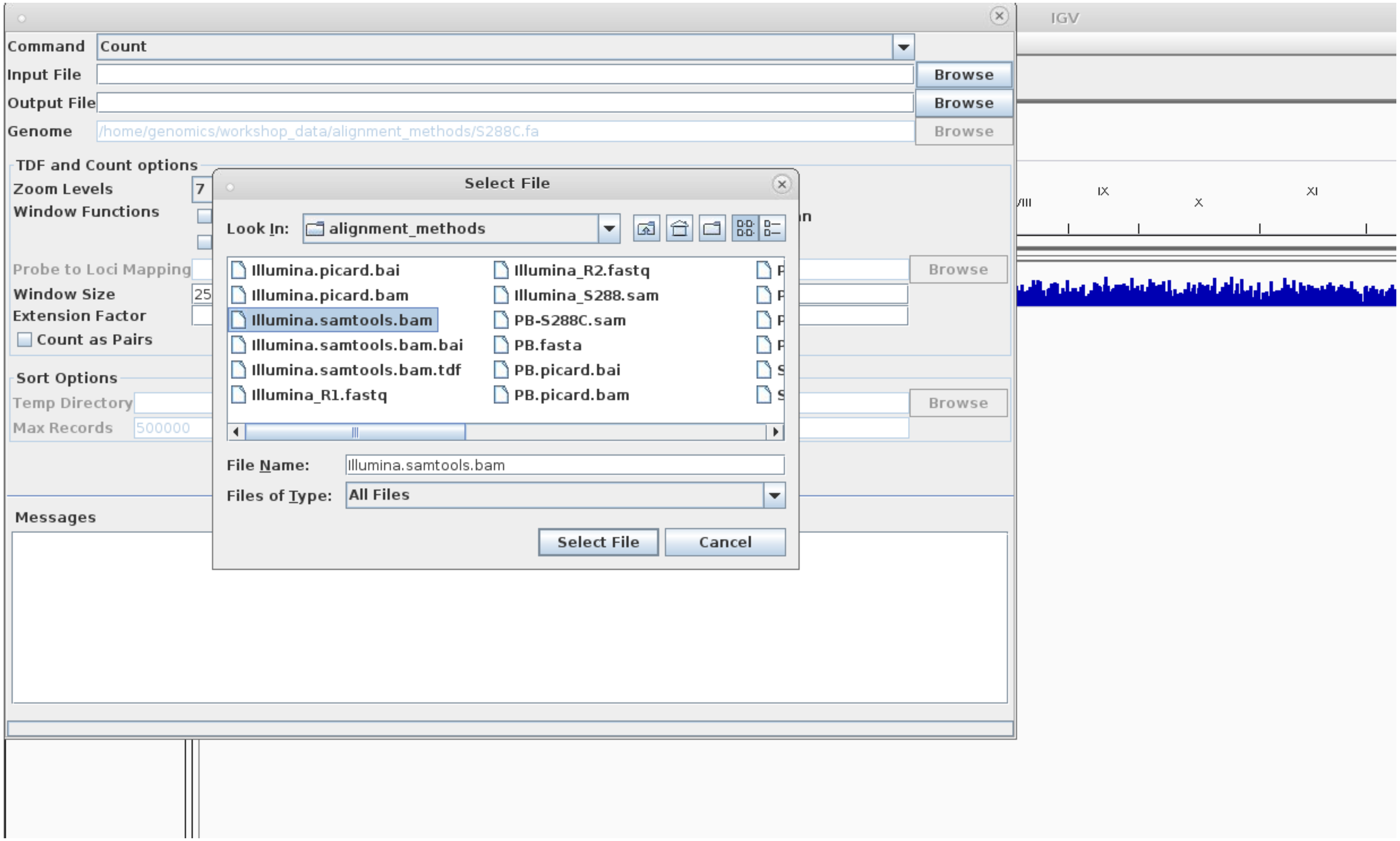
Now we want to use igv-tools to make coverage profiles for our alignments. Go to Menu->Tools->Run igvtools… Now navigate to where you put your bam files for MGI, ONT, and PB versus S288C. One at a time, select these as the input file. It will automatically set the output file. It should default to the “Count” function, and we should be fine with the default parameters. Hit run. When it has finished, do the same for the other files until you have done all three, then close that window.

(See screenshot next slide.)

Computing Coverage



Computing Coverage


Command: Count

Input File:

Output File:

Genome: /home/genomics/workshop_data/alignment_methods/S288C.fa

TDF and Count options

Zoom Levels: 7

Window Functions

Probe to Loci Mapping

Window Size: 25

Extension Factor

□ Count as Pairs

Sort Options

Temp Directory

Max Records: 500000

Messages

Select File

Look In: alignment_methods

Illumina.picard.bai
Illumina.picard.bam
Illumina.samtools.bam
Illumina.samtools.bam.bai
Illumina.samtools.bam.tdf
Illumina_R1.fastq
Illumina_R2.fastq
Illumina_S288.sam
PB-S288C.sam
PB.fasta
PB.picard.bai
PB.picard.bam

File Name: Illumina.samtools.bam

Files of Type: All Files

Load Sequence Data

Now load the 3 bams. Go back to Menu->File->Load from File… and select the bams (not the .bais or the .tdfs or the .bam.bais, but the .bams themselves; IGV will automatically find the bais and the tdfs).

After loading the ONT and HiFi reads, you may want to try going to View->Preferences and then go to the “Third Gen” tab and reset the top field, “Visibility Range Threshold” to 100 kb instead of 1000 kb so that the browser loads fewer reads at low resolution.

Browse!

Navigate around IGV and look at stuff. Some basic browsing:

Click the chromosome numbers to zoom to a chromosome.

Click the “Home” icon to go back to whole genome.

Click and drag in the coordinate window to select a zoom window.

Use the “railroad bars” in the upper right to rescale.

Individual reads will only appear when you are looking at 30kb of genome or less.

You can right-click on the track names to resize or reorder the tracks. You may want to change the compression level of the annotation track so that the different bands don’t stack on top of each other.

Some Interesting Regions

- MT: 1290-1520
- MT: 68,410-68,460
- II: 1-14,000
- IV: 1,510,000-1,531,933
- V: 430,000-520,000
- XII: 449,000-471,000