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 to do these exercises with a diploid yeast (because it would be small, but have heterozygous variants [I failed at this, it seems]) which had Illumina, Oxford Nanopore, and HiFi data [I failed at getting HiFi]. I found a paper from Sanger with Illumina, ONT, and PacBio (not HiFi) with a diploid S288C relative (it turns out to be a homozygous diploid). Since non-HiFi PacBio and ONT are very similar, I picked one. On the next slide is what happened.

The Ugly Mess Behind the Curtain

1. Decided to use ONT data. Downloaded from EBI.
2. ONT data was only submitted in ONT's fast5 format.
3. Found three programs to extract fasta from fast5, poretools, nanopolish, and poreseq.
4. Downloading poretools. It did not build. Went back to an earlier version of python. Build started but then failed (Guy did eventually succeed).
5. Switching to nanopolish. Nanopolish compiled. The current version of nanopolish no longer supports read extraction.
6. Tried to download an older version of nanopolish, but it was no longer supported in the github repo.
7. Tried to download poreseq. It no longer clones from github.
8. Downloaded the PacBio reads. They are in PacBio's hd5 format.
9. Found bax2bam to extract bam from hd5, downloaded it.
10. Current version of bax2bam is broken (globally, it appears).
11. Downloaded old version of bax2bam.
12. Extracted bam files.
13. Wrote perl script to extract raw fasta from the bam files.

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, one has Illumina data, and one has PacBio data.

The Yeast Data

The data we are working with come from the model yeast, *Saccharomyces cerevisiae*. We will look at the sequencing data of one strain, SK1, and compare it to the finished sequence of the reference strain, S288C. We have data from two sources: Illumina and Pacbio. We will align both and perform small variant calling with the Illumina and structural variant calling with the PacBio.

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/ Illumina/* PB/**

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 Illumina/ERR1938686_1.fastq.gz Illumina/ERR1938686_2.fastq > output.sam

This will take a bit 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

zcat Illumina/ERR1938686_1.fastq.gz | less

less output.sam

Convert Output to Binary

Our next step is to convert the sam file into binary format. There are two programs we can use to do this, GATK and samtools. The samtools syntax is a little easier, but the GATK tools are more comprehensive, and in some cases you might want to use other functions GATK provides that it can do simultaneously (as we will see here). Pick one of these (it doesn't matter which) and run it. If you have time (now or later), you can go back and do the other one with a different output file name.

Either of these should take about 2 minutes to run. Note the second step which is necessary with samtools!

Option 1: Run Samtools

Using samtools:

samtools sort –o output.bam output.sam

samtools index output.bam

Option 2: Run GATK

Using GATK:

gatk SortSam -I output.sam -O output.bam -SO coordinate --CREATE_INDEX true

Align the PacBio Data with Minimap2

Now we are going to align the Pacbio reads. The Pacbio reads are in PB/PB.fa

minimap2 –a –o output.sam S288C.fa PB/PB.fa

These should each take about a few minutes to align.

Binary Convert the Minimap2 Output

Now convert these into sorted bams also using either samtools or GATK, whichever you prefer.

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 either 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.

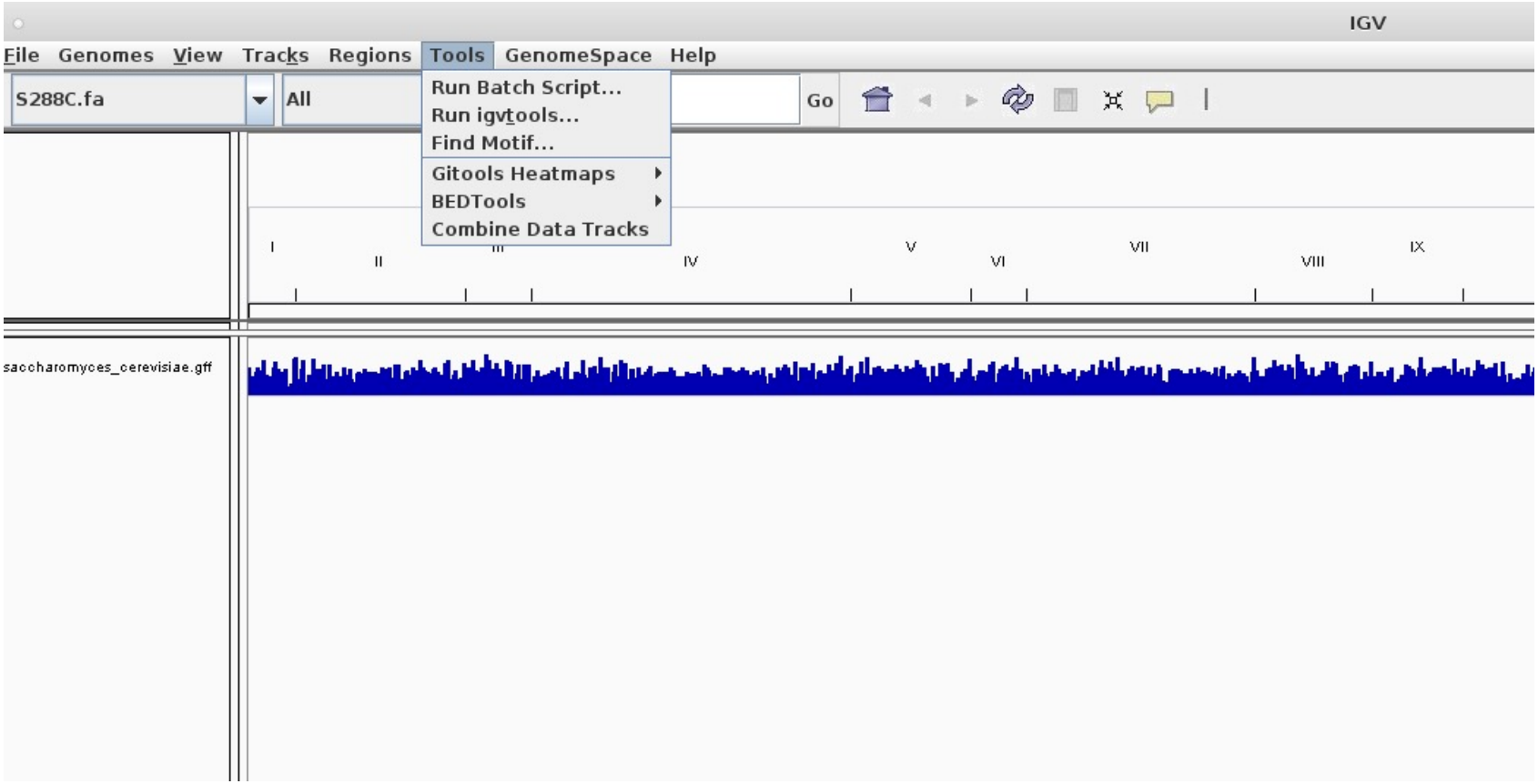
You will probably get some warning messages, but just persist and it will load properly in spite of them.

Computing Coverage

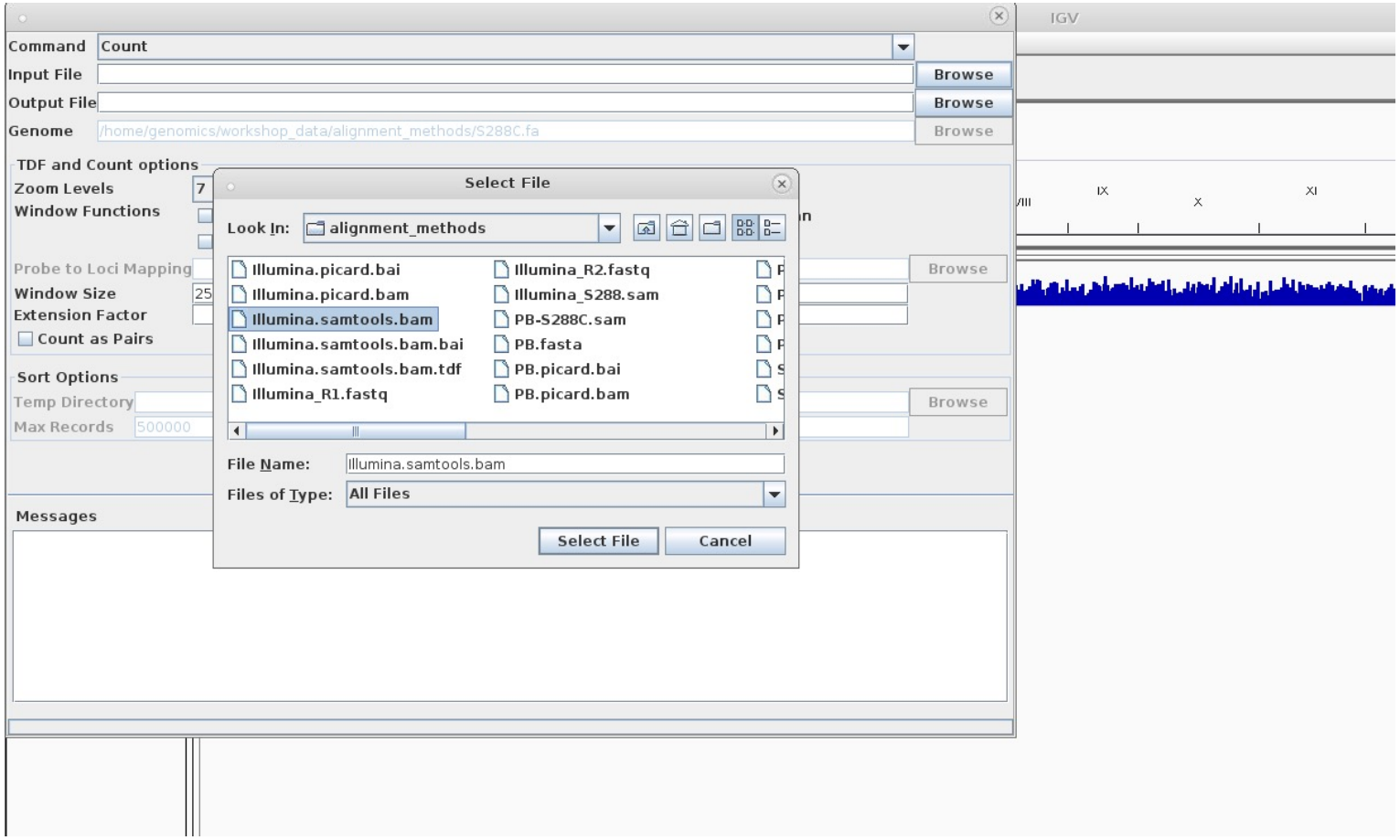
Now we want to use igv-tools to make coverage profiles for our alignments. Go to Menu->Tools->Run igvtools… Now navigate to where you put your bam files for Illumina and PB versus S288C. One at a time, select these as the input file. It will automatically set the output file. It should default to the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

(See screenshot next slide.)

Computing Coverage



Computing Coverage

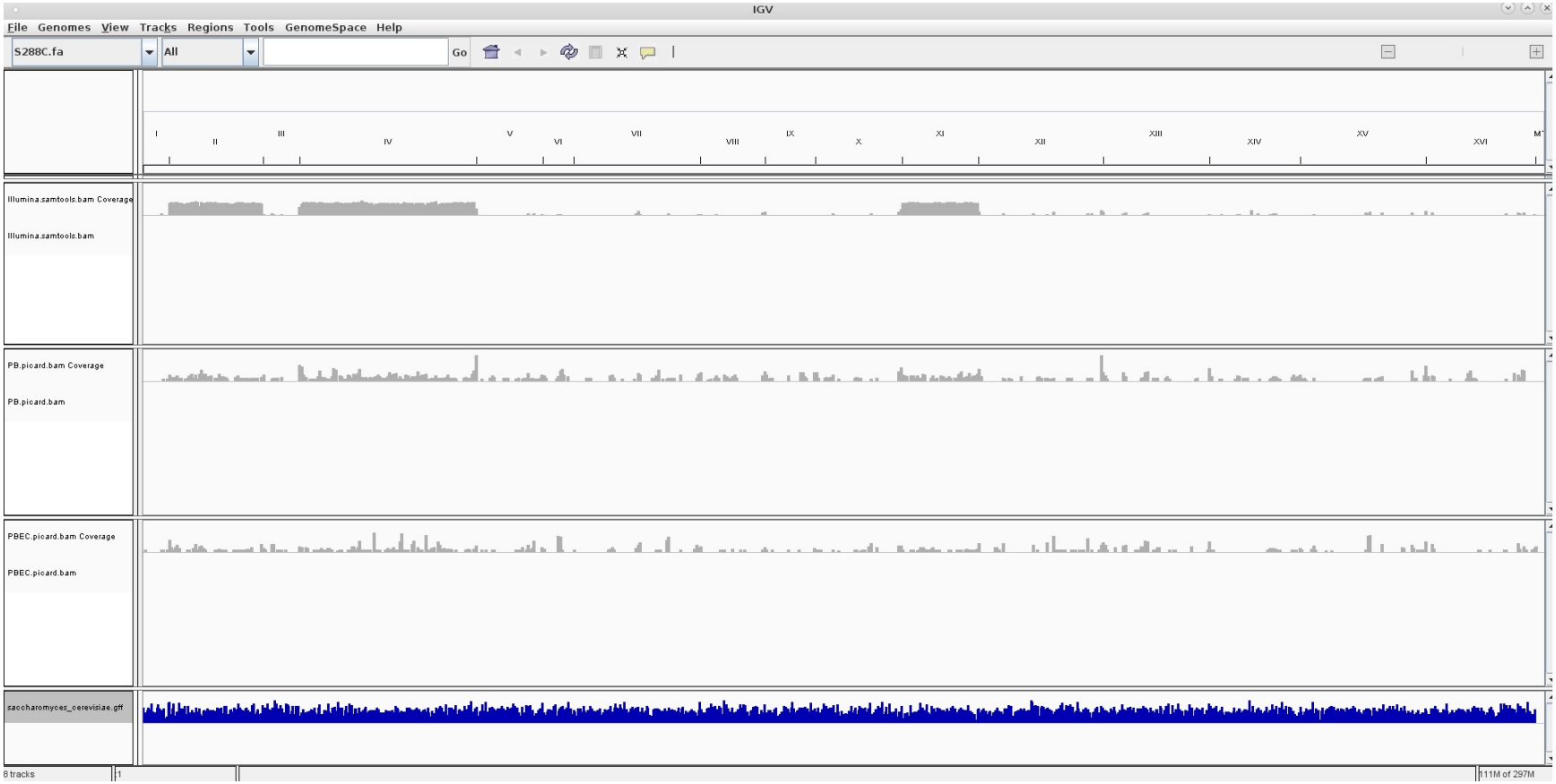


Load Sequence Data

Now load the 2 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).

Data Loaded

Now you should have something that looks like this:



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.