Alignment Workshop

Cesky Krumlov

January 11, 2018

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_methods

Look at what we have:
ls
The data we are working with come from the model yeast, *Saccharomyces cerevisiae*. We will look at the sequencing data of one strain, W303, and compare it to the finished sequence of the reference strain, S288C. We have data from two sources Illumina MiSeq and Pacbio RSII. The Pacbio reads further are raw and error corrected (these are not identical read sets; they are separate subsets of the same larger original dataset; any identical reads in the two sets are coincidental).

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 **

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_methods (~/workshop_materials/alignment_methods). Any files you create will be wherever you put them, and I will refer to them as, e.g., *<Illumina-S288C.sam>*. This <name> notation is a common placeholder for a specific named file (or directory) for which you are assumed to know the path/name but which the writer of the documentation could not know the name. **Do not actually put the <> in your commands;** remember that these are file redirects and you will do unfortunate things you did not expect.

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 S288C.fa

This should take about 15 seconds. After you are done, there should be 5 additional files in the alignment_methods 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 S288C.fa Illumina_R1.fastq Illumina_R2.fastq > <Illumina-S288C>.sam

This should take about 3 minutes.

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 S288C.fa

less Illumina_R1.fastq

less <Illumina-S288C>.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, Picard and samtools. The samtools syntax is a little easier, but the Picard tools are more comprehensive, and in some cases you might want to use other functions Picard provides that it can do simultaneous (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 <Illumina.samtools>.bam <Illumina-S288C>.sam

samtools index <Illumina.samtools>.bam

Option 2: Run Picard

Using picard:

picard-tools SortSam INPUT= <Illumina-S288C>.sam OUTPUT= <Illumina.picard>.bam SORT_ORDER=coordinate CREATE_INDEX=true

Align the PacBio Data with Blasr

Now we are going to align the Pacbio reads. The raw Pacbio reads are in PB.fasta, and the error-corrected Pacbio reads are in PBEC.fasta. Align each of these with blasr.

blasr PB.fasta S288C.fa –out <PB-S288C>.sam –sam

blasr PBEC.fasta S288C.fa –out <PBEC-S288C>.sam –sam

The uncorrected reads should take about 10 minutes to align, the corrected ones about 4 minutes. You have multiple cores on your virtual server, so you can actually run these at the same time and get them both done in the time it would take to run the slower one.

Binary Convert the Blasr Output

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

If You Ran Picard…

Note that if you tried Picard, you got an error. Picard by default strictly enforces compliance with the SAM/BAM specification on both input and output files, and blasr creates a non-compliant read group tag (it is missing both the sample name 'SM' and the library name 'LB'. If you want to run Picard, you need to include the extra argument '*VALIDATION_STRIGENCY=SILENT*'. This tells Picard to process the file even if it has format violating errors as long as it can make syntactically correct output.

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 or use the shortcut on your desktop. 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.

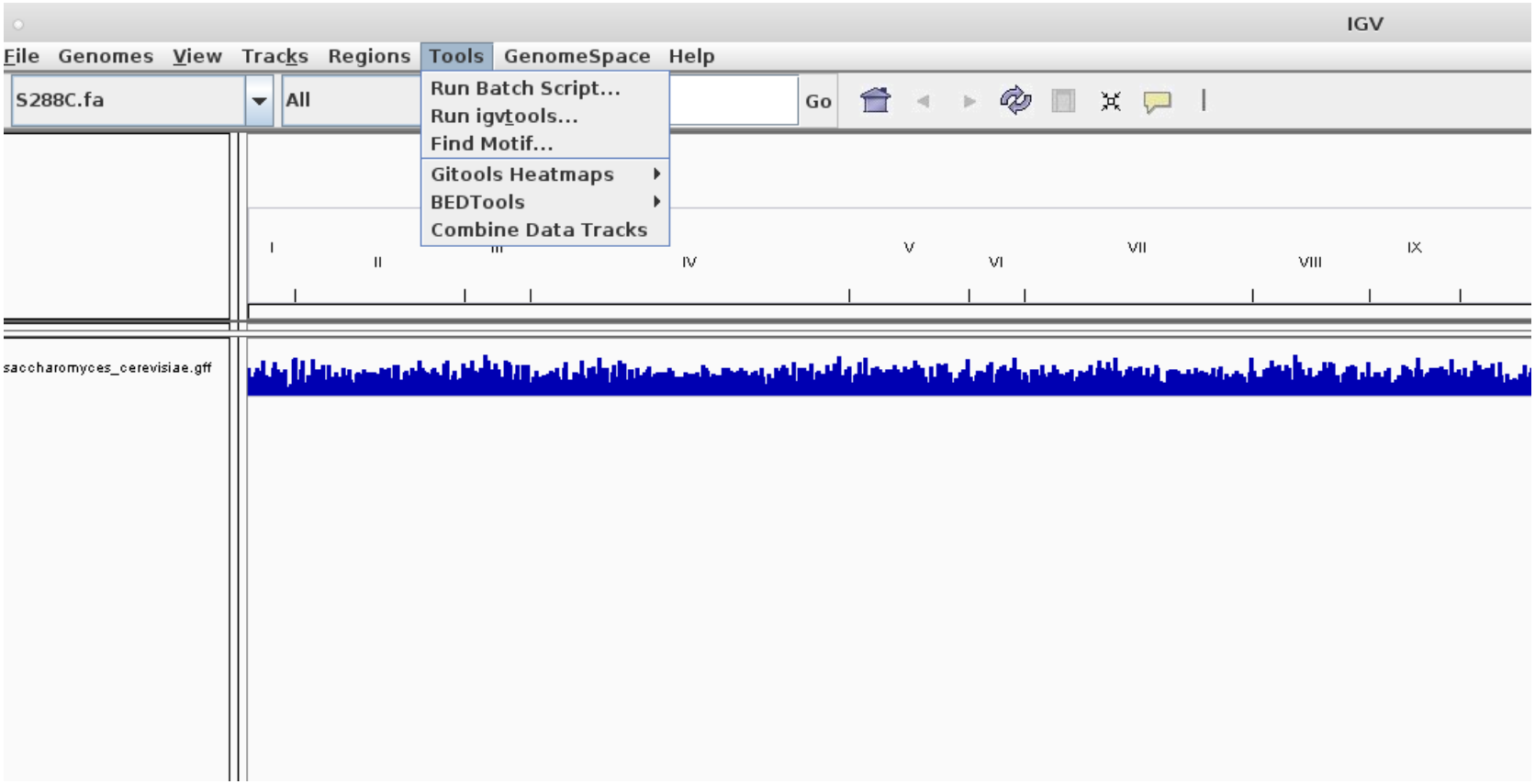
Load the genome (Menu->Genomes->Load Genome from File…) S288C.fa from the alignment_methods directory. This should give you a graphical layout of the chromosome lengths at the top of the screen. We also want the annotations, so go to Menu->File->Load from File… and load saccharomyces_cerevisiae.gff. *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 the 2 PB runs 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

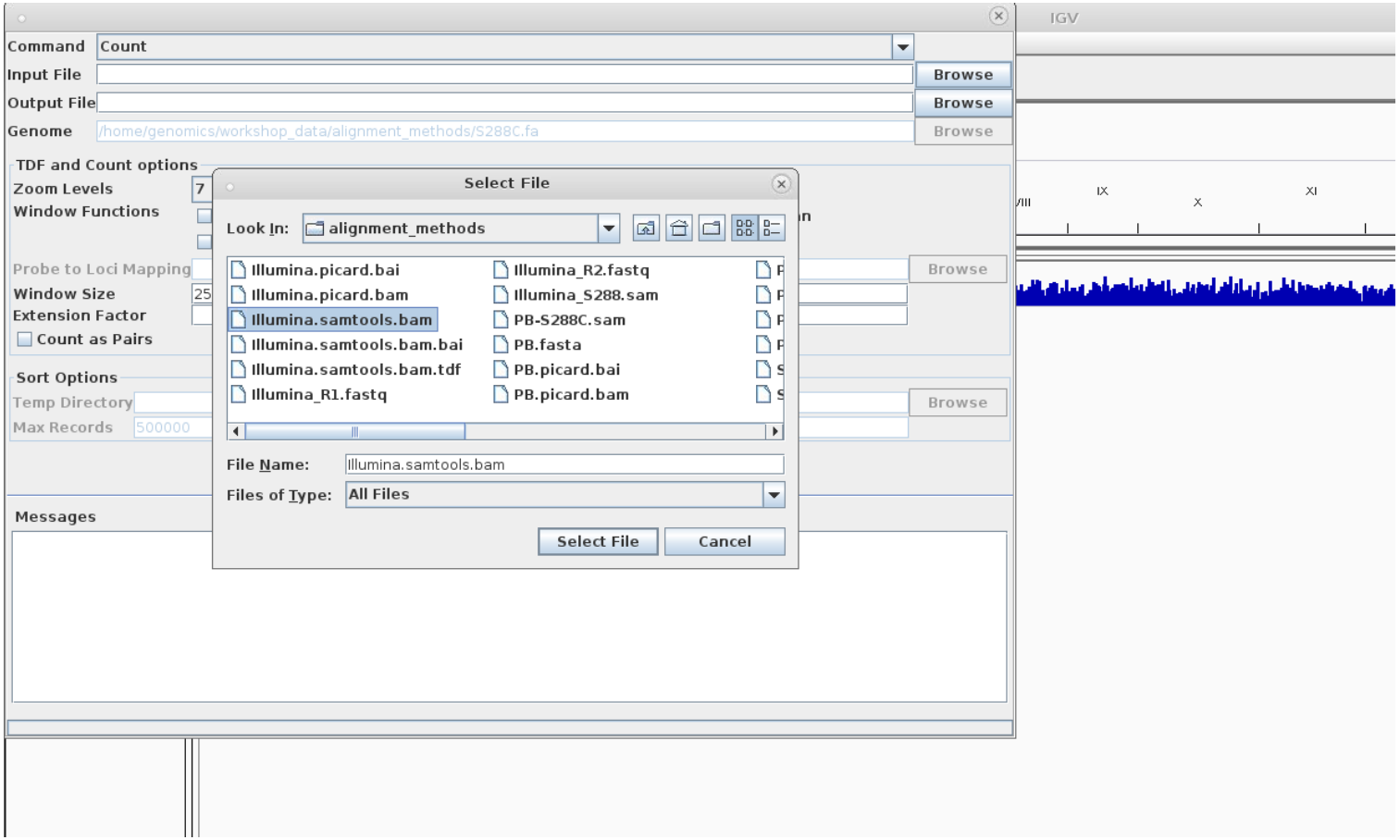


Computing Coverage

Now navigate to where you put your bam files for Illumina and the two PB versus S288C alignments. 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

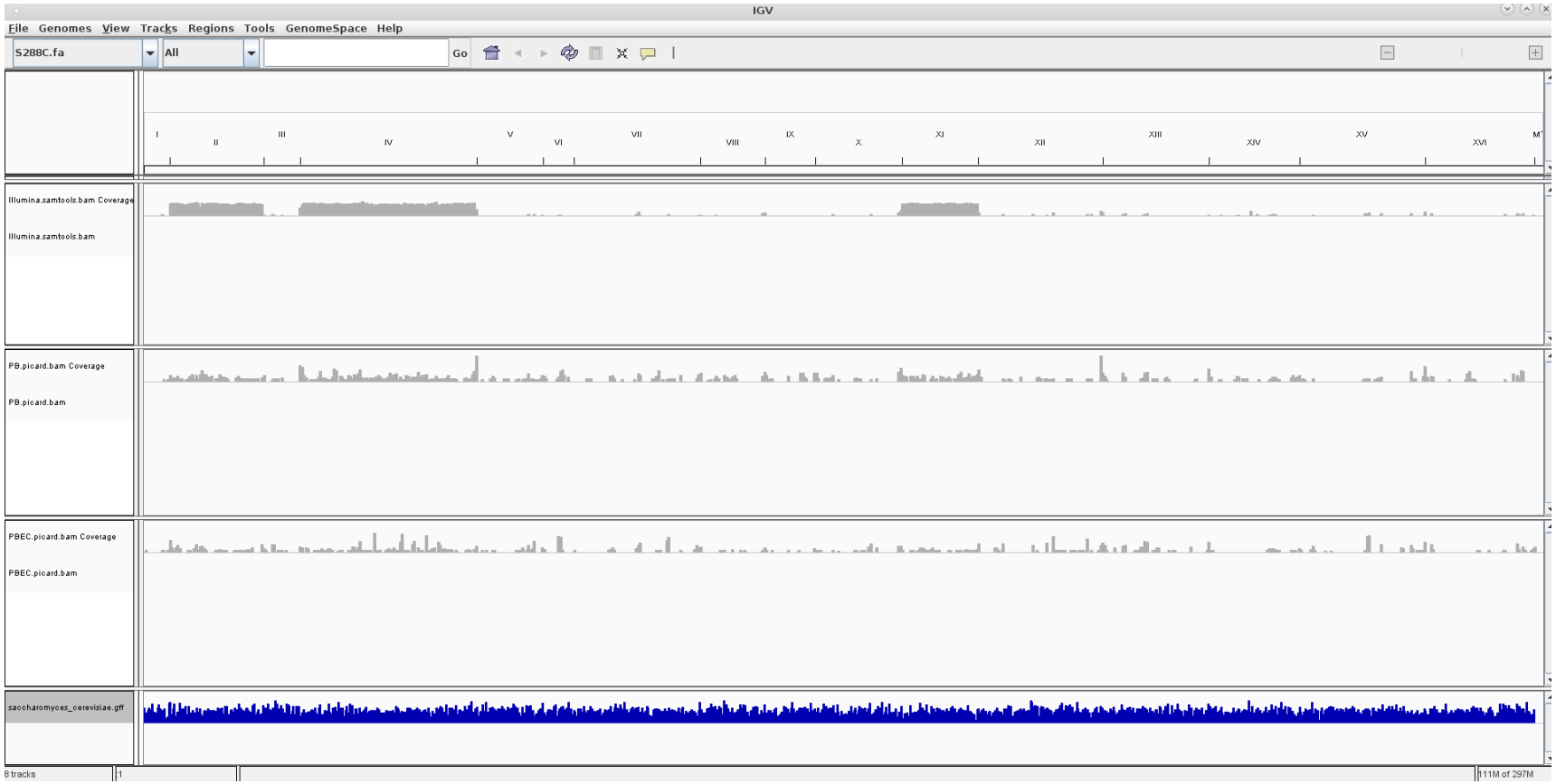


Load Sequence Data

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).

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.

The next few pages have some information about places that might be interesting to look at.

Things to See

You only started with reads that should align to chromosomes II, IV, and XI.

The Illumina reads have some off-target noise (e.g. I:160000-166000 or III: 82000-92000). What looks different about these reads?

The PacBio data are very sparse (it takes too long to align otherwise).

The next few pages have ranges on each of the chromosomes that seem interesting for various reasons.

Chromosome II

30000-35000

155000-185000

223000-227000

245000-275000

620000-710000

Chromosome IV

1-40000
160000-205000
215000-220000
300000-310000
355000-370000
375000-415000
437000-438000
510000-540000
635000-670000
822000-823000
870000-880000
980000-995000
1090000-1105000
1205000-1215000

Chromosome XI

100000-195000

202000-204000

372000-382000

408000-640000

517000-518000