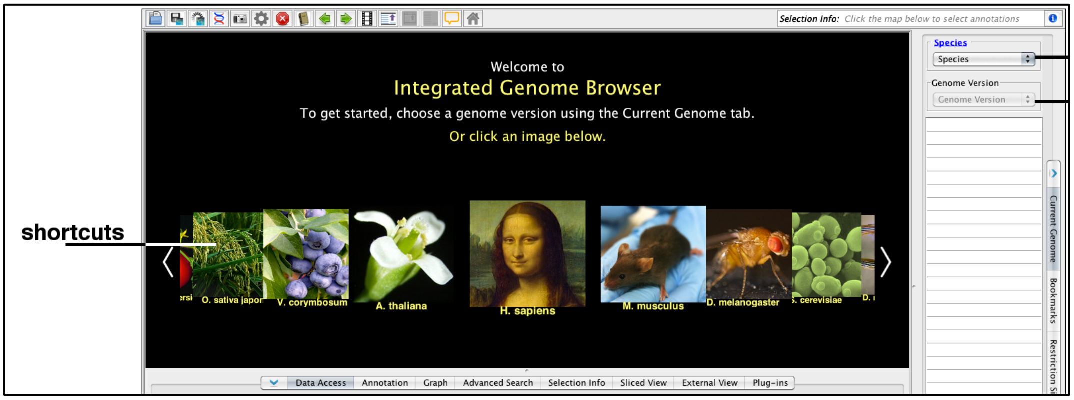
Downloading IGB and today's data

1. Go to http://bioviz.org, select Downloads and download the IGB installer.
2. Double-click the installer and follow the instructions to install IGB.
3. Download today's data by going to https://dl.dropboxusercontent.com/u/33652246/LabData.zip
4. Unzip the file.

Getting started with IGB

Select a genome

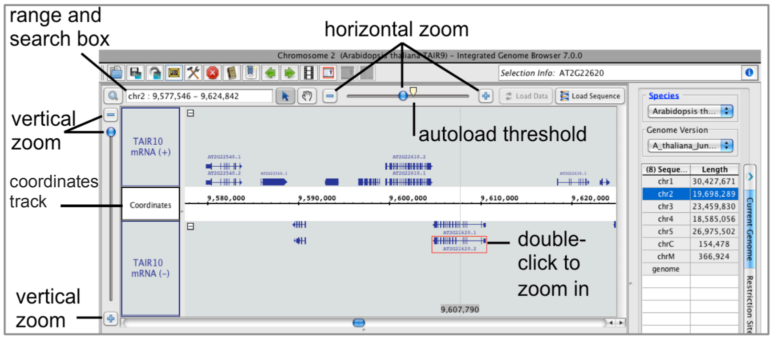
- Click on the image of a genome and it will automatically open.



Welcome to
Integrated Genome Browser
To get started, choose a genome version using the Current Genome tab.
Or click an image below.

Navigating a genome

- Use the **horizontal zoom** slider, or highlight an area of interest in the **Coordinates** track.
- To zoom in on a gene, enter its name in the **Search Box**, or double click on the gene.
- Use the **Vertical Zoom** to expand gene models.



(8) Seque...	Length
chr1	30,427,671
chr2	19,698,289
chr3	23,459,830
chr4	18,585,056
chr5	26,975,502
chrC	154,478
chrM	366,924
genome	

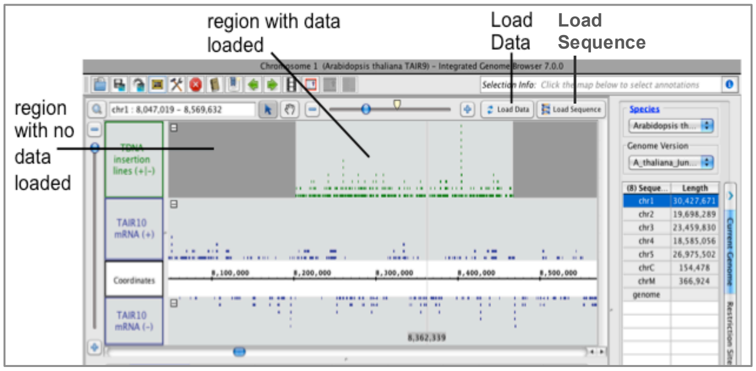
Loading data

- Open a file by dragging it directly into IGB. Or select it using **File > Open File…**

IGB loads data into tracks. Tracks will appear grayed out at first, as they do not load automatically.

To load data into tracks or to load sequence:

1. Zoom in to a region or gene of interest.
2. Click the **Load Data** button in the upper right corner to load data.
3. Click on **Load Sequence** to load genomic sequence.



(8) Seque...	Length
chr1	30,427,671
chr2	19,698,289
chr3	23,459,830
chr4	18,585,056
chr5	26,975,502
chrC	154,478
chrM	366,924
genome	

Identifying SNPs and insertions within genes and their promoters

1. Start IGB.
2. Select the chicken genome by clicking on G. gallus.
3. Practice navigating the chicken genome.

Q1) How many exons does the TBX6 gene have?

4. Zoom out so that you can see the entire TBX6 gene.
5. Click on the **Load Sequence** button in the upper right corner.
6. Zoom in on the first exon of TBX6.

Q2) What are the first three amino acids of TBX6?

Q3) What are the last three amino acids of TBX6?

Q4) What are the first ten base pairs in the SOX2 gene?

Q5) How many base pairs make up chromosome 2?

Q6) What is the id for the IRX1 gene?

Remember that genes can be on either the positive or negative strand. IGB displays these genes as two separate tracks, one for the positive and one for the negative.

7. Combine the two DNA strands together by clicking on the **+/-** under **Data Access -> Data Management Table**

Q7) What strand is the IRX2 gene on?

There should be two folders in the LabData file - chicken and human.

8. Open the genes folder within the chicken folder.

The rumpless_homozygote file is an Araucana that is homozygous for the rumpless mutation, while the tailed_homozygote file is an Araucana that is homozygous without the rumpless mutation.

9. Drag and drop the rumpless_homozygote.bam file into the main window in IGB.

This file contains reads from the whole genome sequencing of Araucana that mapped to the two genes that we are interested in - IRX1 and IRX2.
IRX1 can be found at coordinates chr2:87,127,233-87,133,900
IRX2 can be found at coordinates chr2:86,627,171-86,635,484

10. Go to the IRX1 coordinates.
11. Click on the **Load Data** button in the upper right corner.
12. Zoom in on the first exon of IRX1.

Q8) What is the length of the longest read?

13. Zoom out so you can see the first exon of IRX1 (chr2:87,131,457-87,132,575).
14. Click on the **Load Sequence** button in the upper right corner.

The reference sequence for the region you are currently looking at is now loaded. Note that the reads have gone from being multi-colored, to mostly being blue. Where the read's sequence matches the reference sequence, the read turns blue. When the read's sequence does not match the reference sequence, the base pair that does not match is highlighted.

15. Zoom out to the IRX1 coordinates (chr2:87,127,233-87,133,900).
16. Click on **Load Sequence**.
17. Zoom in on the middle of the second intron of IRX1 (chr2:87,133,422-87,133,466).

Note that at position chr2:87,133,431 all of the reads have an A highlighted. This indicates a SNP, as the reference sequence has C at this position.

18. Move a few base pairs to the left, so you can see position chr2:87,133,409.

Note that a single read has C, whereas the reference sequence has a G. Since there is only a single read that supports this C, it is most likely an error in the sequencing machine, and is not a SNP.

19. Zoom in in on the region upstream of IRX1 at position chr2:87,130,081-87,130,126. Here you can see several bases highlighted in red. The bases highlighted in red are small insertions, they are base pairs that are present in the read, but are not present in the reference sequence.

20. Zoom out to the IRX1 coordinates (chr2:87,127,233-87,133,900).

21. Go back to the genes folder, and drag and drop the tailed_genes.bam file into IGB.

22. Click on **Load Data**.

23. Zoom in on the middle of the second intron of IRX1 (chr2:87,133,422-87,133,466). Note that the SNP in the rumpless file is also present in the tailed file. This indicates that this SNP is shared by all Araucana, and has nothing to do with the rumpless phenotype. We're interested in SNPs and small insertions that are only present in either the rumpless file, or in the tailed file.

24. Identify 4 SNPs or insertions within the IRX1 region (chr2:87,127,233-87,133,900) and 4 SNPs or insertions within the IRX2 region (chr2:86,627,171-86,635,484) that are unique to either the rumpless or tailed Araucana.

Q9) IRX1

Position (in base pairs)	SNP or Insertion	In Exon (yes/no)	In Intron (yes/no)

Q10) IRX2

Position (in base pairs)	SNP or Insertion	In Exon (yes/no)	In Intron (yes/no)

Q11) Why would it be important to look for SNPs and insertions within the region upstream of IRX1 and IRX2?

Note that there is a third .bam file in the genes folder, labeled heterozygote.

Q12) Assuming you were able to find the rumpless mutation, and it was a SNP, what would you expect the mutation to look like in the heterozygote_genes file?

Q13) What is the phenotype of the heterozygous Araucana?

25. After you have finished answering questions 1-13, click on the eye icon for each file (rumpless and tailed) under the **Data Access** tab to hide the files.

Identifying large deletions and duplications in the genome

Now that we've identified SNPs and insertions within and around our genes of interest, we want to know if there are any large deletions or duplications in the area between the two genes.

26. Open the graphs folder within the chicken folder.

Within the graphs folder there are multiple files that end in .bedgraph.gz or .bedgraph.gz.tbi. These are bedgraph files that have been compressed (the .gz) and indexed (the .tbi).

These files are from the same Araucana chickens as the genes data.

27. Drag and drop the rumpless_homozygote.bedgraph.gz file into IGB.
28. Go to the following coordinates: chr2:86,478,905-87,202,642
29. Click on the **Load Data** button and then the **Load Sequence** button.

Using the original .bam files, I have created a .bedgraph file. This file adds up all of the reads that overlap each base pair. It allows us to quickly determine where we have many reads, and where we have very few reads.

30. Go to the following coordinates: chr2:86,985,152-86,986,774

Here you can see a region with no reads. This indicates that there is a large deletion. Similarly, a large peak would indicate a duplication.

31. Drag and drop the tailed_homozygote.bedgraph.gz file into IGB.
32. Go to the following coordinates: chr2:86,478,905-87,202,642
33. Click on the **Load Data** button.

We now want to compare the rumpless versus tailed again, looking for differences in one file that are not present in the other. However, we can make this much easier by doing two things. First, we are only interested in big differences. In order to make the big differences stand out, we can take the $\log_{10}$ of each file.

34. Shift-click to select both tracks.
35. In the **Graph** tab, under **Single-Graph** there is a dropdown menu - use it to select **Log$_{10}$** and then click **Go**.

You should now have two new graphs, one for each file. Next we want to only look at the differences between our two graphs. We can do this by subtracting the tailed file from the rumpless file.

36. Shift-click to select both of the new tracks, making sure to select the rumpless track first, and the tailed track second.
37. In the **Graph** tab, under **Multi-Graph** there is a dropdown menu - use it to select **Diff** and then click **Go**.

You should now have one new graph. This graph is the difference in the number of reads between rumpless and tailed samples, at the $\log_{10}$ scale.

38. Hide the four older graphs, as we will not need to use them, by clicking on their respective eye icons under the **Data Access** tab.

Any point in the graph that is greater or less than zero indicates a difference in the number of reads between rumpless and tailed samples. Peaks over 1 indicate ten times as many reads in rumpless, whereas peaks under -1 indicate ten times fewer peaks in rumpless.

39. Go back to the following coordinates: chr2:86,985,152-86,986,774

The lack of reads in the rumpless file compared to the tailed file at this location is displayed as a strong negative peak. This indicates that there is a deletion in rumpless, which is not present in the tailed file. Interestingly, the chicken genome is poorly annotated - there are many undiscovered genes. It is possible that this deletion happens to fall within an unannotated gene, and this is what is causing the rumpless phenotype.

40. Zoom in on the deletion (chr2:86,985,671-86,986,255)

41. Highlight the base pairs that are within the coordinates of the deletion.
42. Right-click on the highlighted coordinates.
43. Click on **BLASTX nr protein database**

BLAST will use the sequence at the deletion to search a large database of genes, looking for similarities. It will return whatever gene is most similar to the sequence we blasted, thus telling us whether there were any unannotated genes at that location.

44. Identify 5 deletions or duplications that are greater than 1 or less than -1. BLAST each deletion/duplication to determine if it occurs at an unannotated gene.

Q14)

Position (basepair - basepair)	Deletion or Duplication	Gene Identified?

Identifying human disease markers

Now that we've thoroughly identified all of the possible SNPs, insertions, duplications, and deletions in Araucana, let's take a look at how we can use a similar approach in humans to identify disease and phenotypes.

45. Click on the **Home** icon at the top of IGB.
46. Click on the human genome.
47. In the upper right corner, click on the dropdown menu under **Genome Version**, and select H_sapiens_Feb_2009
48. Drag and drop the humanSNP.bed.gz file (located in the human folder within the LabData) into IGB.

The humanSNP.bed.gz file contains one million SNPs from my genome, as sequenced by 23andMe. These SNPs are a sampling of all of the SNPs currently known in humans, and are spread throughout the entire genome.

Q15) What does SNP stand for again?

SNPs can themselves cause disease if they occur within a gene. However, the majority of SNPs have no effect, as they fall within introns or outside of genes. Instead, we can use SNPs as markers, as there are millions of them, and they occur throughout the genome. By sequencing the SNPs of thousands of people, we can compare people with different diseases or phenotypes and see if they share any of the same SNPs. If everyone who has a particular phenotype also has the same set of SNPs, then those SNPs are associated with that phenotype.

49. Click on the +/- button for both the RefGene and humanSNP.bed file to put both the positive and negative tracks together.
50. Click on chr4 in the **Current Genome** tab.
51. Click on **Load Data**
52. Search for SNP: i4000397
53. Load the sequence for this region.

SNP i4000397 is associated with hemophilia. AA means a person is at risk of hemophilia, AG a carrier, and GG is not at risk of hemophilia.

54. Hover over the two SNPs to see what my genotype is.

Q16) Am I at risk for hemophilia, and are my parents at risk for hemophilia?

Q17) Is the reference sequence at risk for hemophilia, and why is this not surprising?

Many studies have now linked diseases and phenotypes to different SNPs.

55. Go to http://www.eupedia.com/genetics/medical_dna_test.shtml
56. Pick a phenotype or disease.
57. Copy the SNP id number (in the SNP column).
    *Make sure the disease of interest is one that is tested by 23andMe (look for A, A2, or A3 in the tested by column).
58. Go to whichever chromosome the SNP occurs on in IGB.
59. Click on **Load Data**.
60. Search for the SNP id in the search bar in the upper left.
61. Determine if I am at risk, a carrier, or not at risk for 5 diseases or phenotypes.

Q18)

SNP id	Name of Disease/Phenotype	At Risk, Carrier, Not at Risk