QIAGEN

Tutorial

Comparative Analysis of Three Bovine Genomes

September 15, 2016

Sample to Insight

CLC bio, a QIAGEN Company · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.clcbio.com · support-clcbio@qiagen.com

Comparative Analysis of Three Bovine Genomes

This tutorial takes you through some of the tools for identifying species-specific variants. As an example we use *Bos taurus* (cow), *Bos indicus* (zebu), and *Bison bison* (bison) data sets consisting of single GA and GAII exome sequencing short reads from an Illumina sequencer. For this tutorial, you will need to use CLC Genomics Workbench Version 9.5 or higher.

The goal is to identify zebu-specific (*Bos indicus*) exome variations that are not present in the bison (*Bison bison*) or bovine genome (*Bos taurus*) [Cosart et al., 2011]. Furthermore, we attempt to link the variations to altered pathways using the gene ontology tool that is built into the *CLC Genomics Workbench*.

Downloading the reference data

1. Download the bovine genome with annotations using the **Download** button found in the upper right corner of the *CLC Genomics Workbench* (figure 1). Select "Download Reference Genome Data".

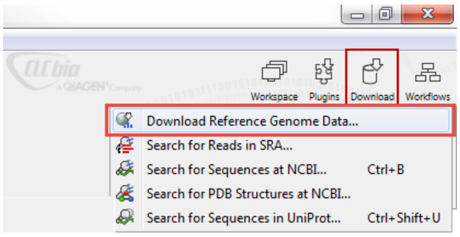
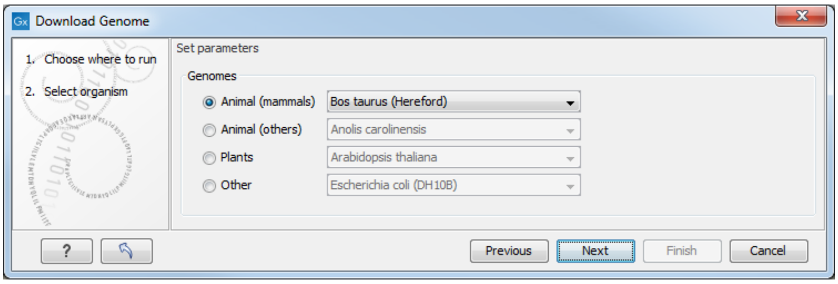

Workspace
Plugins
Download
Workflows
Download Reference Genome Data...
Search for Reads in SRA...
Search for Sequences at NCBI... Ctrl+B
Search for PDB Structures at NCBI...
Search for Sequences in UniProt... Ctrl+Shift+U

Figure 1: *The Download button is found in the upper right corner of the Workbench.*

2. In the wizard step shown in figure 2, select **Animal (mammals)** and **Bos taurus (Hereford)** from the drop-down list. Click on the button labeled Next.



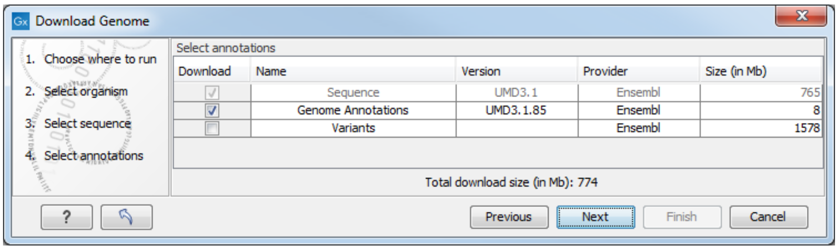
● Animal (mammals) Bos taurus (Hereford)
○ Animal (others) Anolis carolinensis
○ Plants Arabidopsis thaliana
○ Other Escherichia coli (DH10B)

Figure 2: *Select Animal (mammals) and Bos taurus (Hereford) from the drop-down list.*

3. Make sure that the option "Download genome sequence" is selected and click on the button labeled Next.

4. Select to download **Gene annotation** as shown in figure 3. We chose not to download Variants for this tutorial as it would increase the time needed to complete this tutorial. However, be aware that downloading variants would be needed if you wished to run the tool **Annotate from Known Variants** ( ) in subsequent analyses. Click on the button labeled Next.

5. Choose to **Save** the data and click on the button labeled Next.



Select annotations

Download	Name	Version	Provider	Size (in Mb)
☑	Sequence	UMD3.1	Ensembl	765
☑	Genome Annotations	UMD3.1.85	Ensembl	8
□	Variants		Ensembl	1578

Total download size (in Mb): 774

Figure 3: *Select Gene annotation.*

6. Specify where you would like to save your data. It can be a good idea to create a new folder for the tutorial, and a subfolder for the references. This can be done by clicking on the ( ) at the top of the wizard.

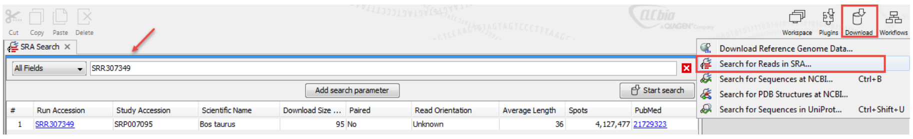
7. Click on the button labeled **Finish** to start downloading data. You can follow the process in the lower left corner.

Importing the reads

1. Download the read files using the tool:

   Download | **Search for Reads in SRA** ( )

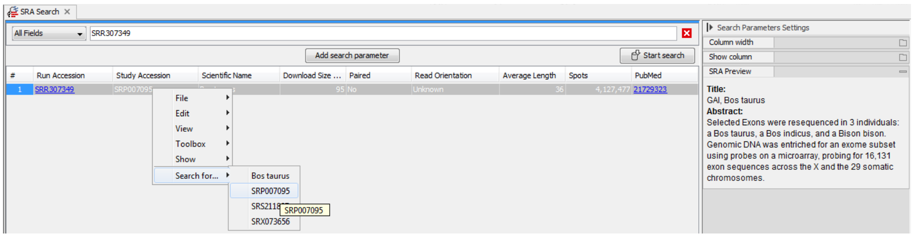
2. This will bring up the dialog shown in figure 4. Search for one of the run with the following accession number SRR307349.


Cut
Copy
Paste
Delete
CLCbio A QIAGEN Company
Workspace
Plugins
Download
Workflows
SRA Search
All Fields
SRR307349
Add search parameter
Start search
Download Reference Genome Data...
Search for Reads in SRA...
Search for Sequences at NCBI... Ctrl+B
Search for PDB Structures at NCBI...
Search for Sequences in UniProt... Ctrl+Shift+U

#	Run Accession	Study Accession	Scientific Name	Download Size ...	Paired	Read Orientation	Average Length	Spots	PubMed
1	SRR307349	SRP007095	Bos taurus	95	No	Unknown	36	4,127,477	21729323

Figure 4: *Search for a particular run using the Search for Reads in SRA tool.*

3. The search output one run. To find the other reads generated during this study, right click on the row of the search output table and choose the option "search for..." and select SRP007095, i.e., the accession number of the study (figure 5).



#	Run Accession	Study Accession	Scientific Name	Download Size ...	Paired	Read Orientation	Average Length	Spots	PubMed
1	SRR307349	SRP007095		95	No	Unknown	36	4,127,477	21729323

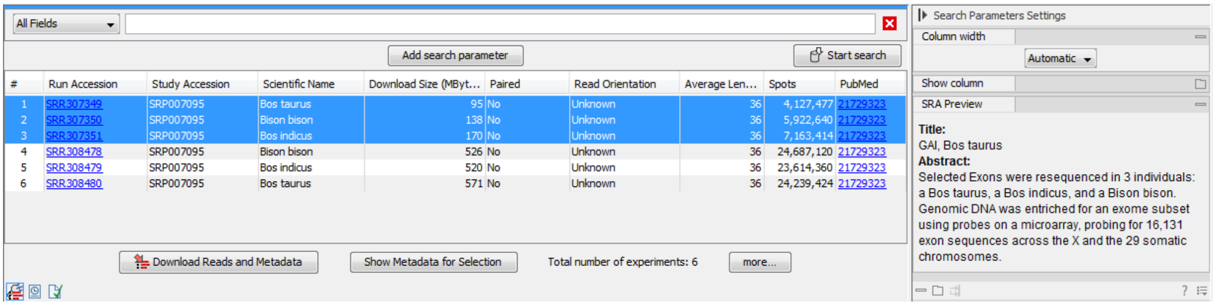
Title:
GAI, Bos taurus
Abstract:
Selected Exons were resequenced in 3 individuals: a Bos taurus, a Bos indicus, and a Bison bison. Genomic DNA was entriched for an exome subset using probes on a microarray, probing for 16,131 exon sequences across the X and the 29 somatic chromosomes.

Figure 5: *Search for more runs from the same experiment.*

4. You have now 6 different runs in the search output table. To minimize the tutorial download and analysis time, select the 3 runs that have the lowest Download Size and click on Download Reads and Metadata (figure 6).

5. In the download dialog that pops up, choose to discard read names and quality scores.

6. Save the data in the folder you created for this tutorial.



#	Run Accession	Study Accession	Scientific Name	Download Size (MByt...	Paired	Read Orientation	Average Len...	Spots	PubMed
1	SRR307349	SRP007095	Bos taurus	95	No	Unknown	36	4,127,477	21729323
2	SRR307350	SRP007095	Bison bison	138	No	Unknown	36	5,922,640	21729323
3	SRR307351	SRP007095	Bos indicus	170	No	Unknown	36	7,163,414	21729323
4	SRR308478	SRP007095	Bison bison	526	No	Unknown	36	24,687,120	21729323
5	SRR308479	SRP007095	Bos indicus	520	No	Unknown	36	23,614,360	21729323
6	SRR308480	SRP007095	Bos taurus	571	No	Unknown	36	24,239,424	21729323

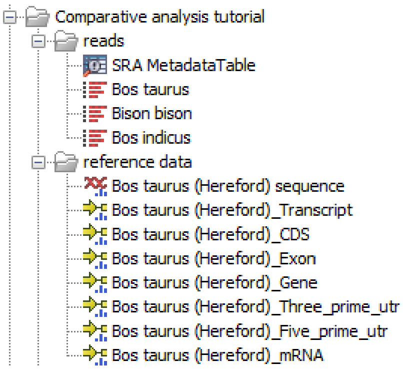
Total number of experiments: 6

Title:
GAI, Bos taurus
Abstract:
Selected Exons were resequenced in 3 individuals: a Bos taurus, a Bos indicus, and a Bison bison. Genomic DNA was entriched for an exome subset using probes on a microarray, probing for 16,131 exon sequences across the X and the 29 somatic chromosomes.

Figure 6: *Download in one click the runs and associated metadata.*

7. For clarity in subsequent analysis, rename the reads with the corresponding species name (right click on the read file in the Navigation Area).

When all data are downloaded you can see the following files in the **Navigation Area** as shown in figure 7.



- Comparative analysis tutorial
  - reads
    - SRA MetadataTable
    - Bos taurus
    - Bison bison
    - Bos indicus
  - reference data
    - Bos taurus (Hereford) sequence
    - Bos taurus (Hereford)_Transcript
    - Bos taurus (Hereford)_CDS
    - Bos taurus (Hereford)_Exon
    - Bos taurus (Hereford)_Gene
    - Bos taurus (Hereford)_Three_prime_utr
    - Bos taurus (Hereford)_Five_prime_utr
    - Bos taurus (Hereford)_mRNA

Figure 7: *The downloaded reference genome data.*

We are now ready to start analyzing the data.

Map the reads and detect variants

The data used in this tutorial do not contain adapters, so we do not need to run a trimming step. Thus, we proceed directly to mapping the reads to the reference.

We will map the reads to the reference genome using the "Batch" mode, which means that the three sets can be run in parallel (simultaneously).

1. To map reads to the reference, go to the toolbox:

   Toolbox | **NGS Core Tools** ( ) | **Map Reads to Reference** ( )

   In the first wizard step (figure 8), check the **Batch** box, then select the folder that holds the reads. Do not forget to check the batch box as it would generate a single read mapping with the reads of all three species. Click Next.

2. You should now see the batch overview window (figure 9). In this wizard step you can see the reads that were contained in the selected folder. In this case we will use all the reads in the analysis so you can just click Next.

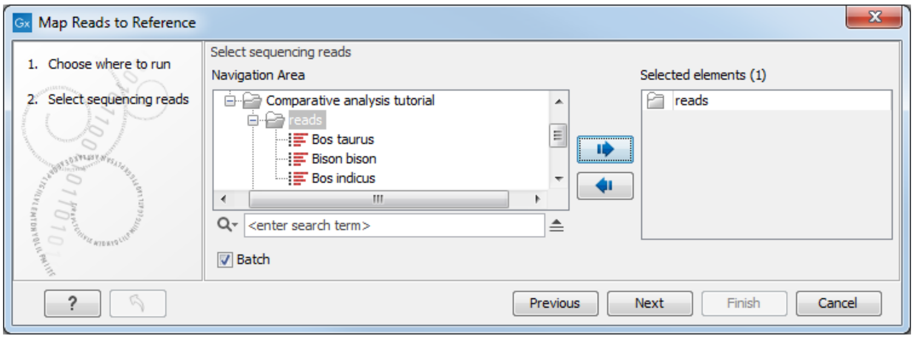


Figure 8: *Run the analysis in batch mode to generate one read mapping per species.*

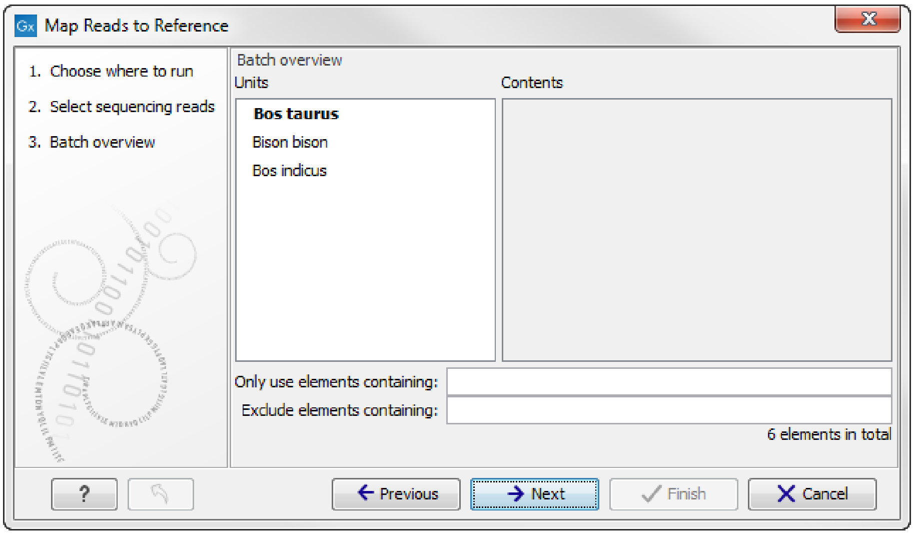


Figure 9: *Failing to see this dialog means you forgot to check the batch option previously!*

3. In the next wizard window (figure 10), click on the browse icon ( ) under **References** to select the *Bos taurus* (Hereford) reference sequence. Press OK. Leave the Reference masking parameters as set by default to "No masking" and click on Next.

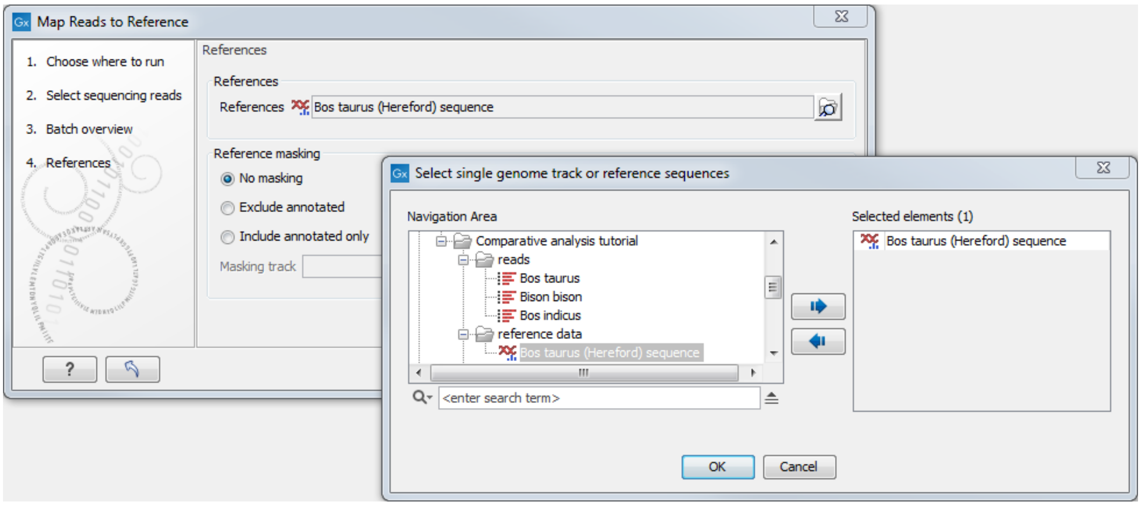


Figure 10: *Select the Bos taurus (Hereford) reference sequence.*

4. In the next wizard step, you can specify the **Mapping options**. We will use the default settings and click on the button labeled Next.

5. Select the options "Create reads track", "Create report" and "Save in specified location" using the option to "Create subfolders per batch unit". Specify where you want to save the data, for example a new folder called "Analysis" and click on the button labeled **Finish**.

Open the mapping summary reports from the **Navigation Area**. 100% of reads are mapped successfully in all three analyses. We can then proceed to a local realignment of the reads to improve the quality of the read mappings we have just generated.

1. To run the local realignment tool go to the toolbox:

   Toolbox | **NGS Core Tools** () | **Local Realignment**

2. Again we choose to run the analysis in **batch mode**, which allows us to analyze all samples in parallel. Select all the read mapping tracks by selecting the top folder as shown in figure 11 and click on the button labeled Next.

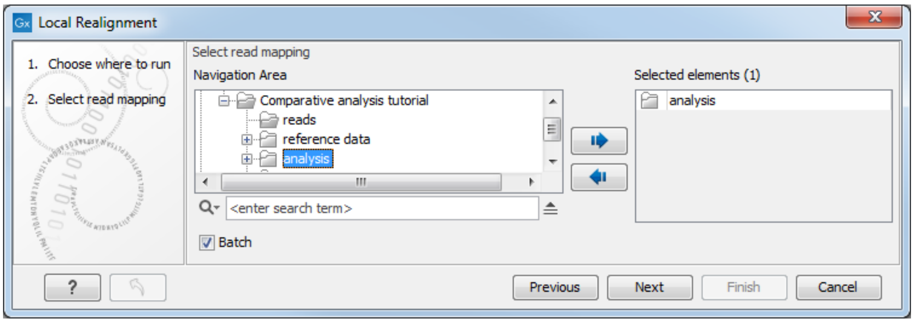


Figure 11: *Run local realignment for all species at the same time by using the batch function.*

3. The next dialog in which you can see the content of the selected folder confirms that you are working in batch mode. Click Next.

4. In the next wizard step, use the default settings (we will run the local realignment without a guidance track). Click on the button labeled Next.

5. Click on **Create reads track** (see figure 12) and choose to **Save in input folder**. Click on **Finish**.

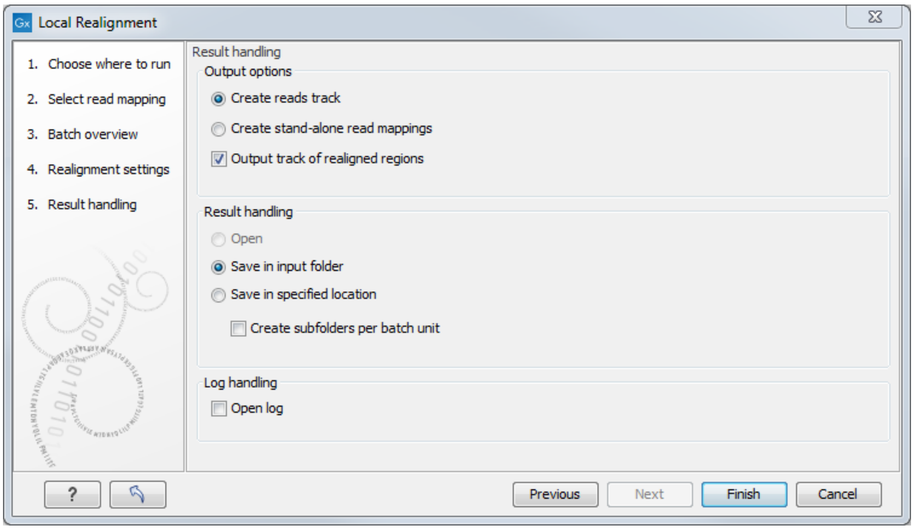
We are now ready to identify the variants. Note that we are generating here a variant track for all three species, when in fact we will use only the *Bos indicus* variant track in subsequent analyses.

1. Go to the toolbox:

   Toolbox | **Resequencing Analysis** () | **Variants Detection** | **Basic Variant Detection** ()

2. Using the batch function, select the top folder as in figure 13.

3. Click through the next wizard steps (by clicking on the button labeled Next) using the default settings. If you have changed some of the settings, you can always go back to the default settings by pressing the arrow button () in the lower left corner, which will return parameters to default settings.



Result handling

Output options

● Create reads track

○ Create stand-alone read mappings

☑ Output track of realigned regions

Result handling

○ Open

● Save in input folder

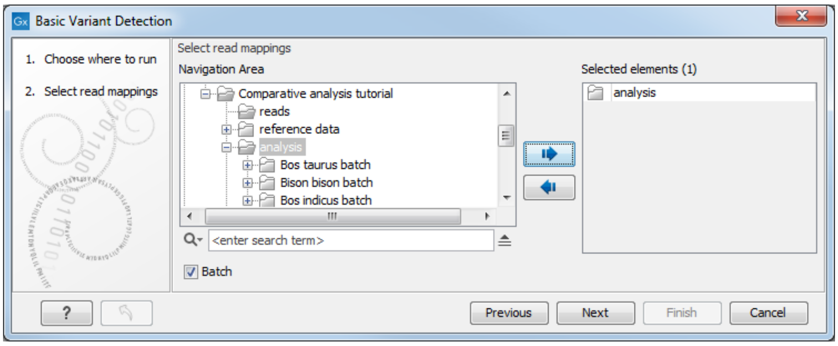
○ Save in specified location

□ Create subfolders per batch unit

Log handling

□ Open log

Figure 12: *Use the default settings and choose to save the results.*



Select read mappings

Navigation Area

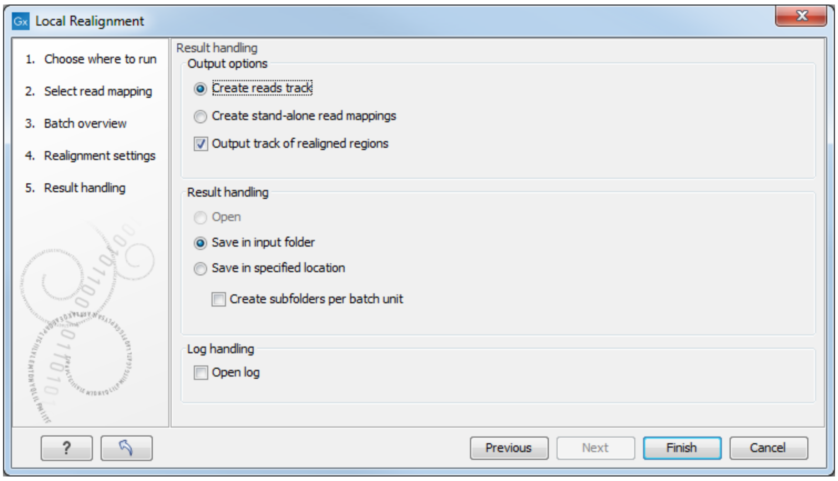
- Comparative analysis tutorial
  - reads
  - reference data
  - analysis
    - Bos taurus batch
    - Bison bison batch
    - Bos indicus batch

Selected elements (1)

analysis

☑ Batch

Figure 13: *Select the folder that holds the locally realigned reads.*



Result handling

Output options

● Create reads track

○ Create stand-alone read mappings

☑ Output track of realigned regions

Result handling

○ Open

● Save in input folder

○ Save in specified location

□ Create subfolders per batch unit

Log handling

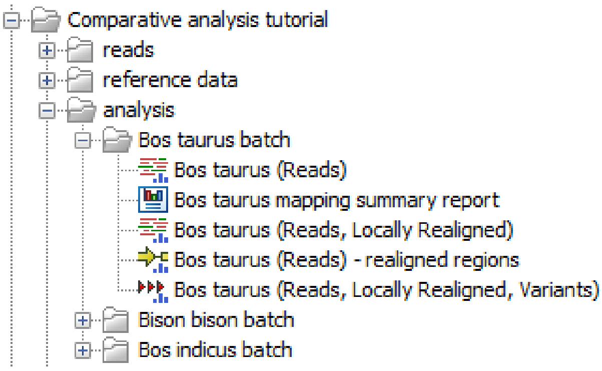
□ Open log

Figure 14: *Create a reads track and save the results in their respective input folders.*

4. In the last wizard step choose to **Create reads track** and **Save in input folder** before clicking on **Finish**(figure 14).

Take a look at the results, which should be organized as shown in figure 15.

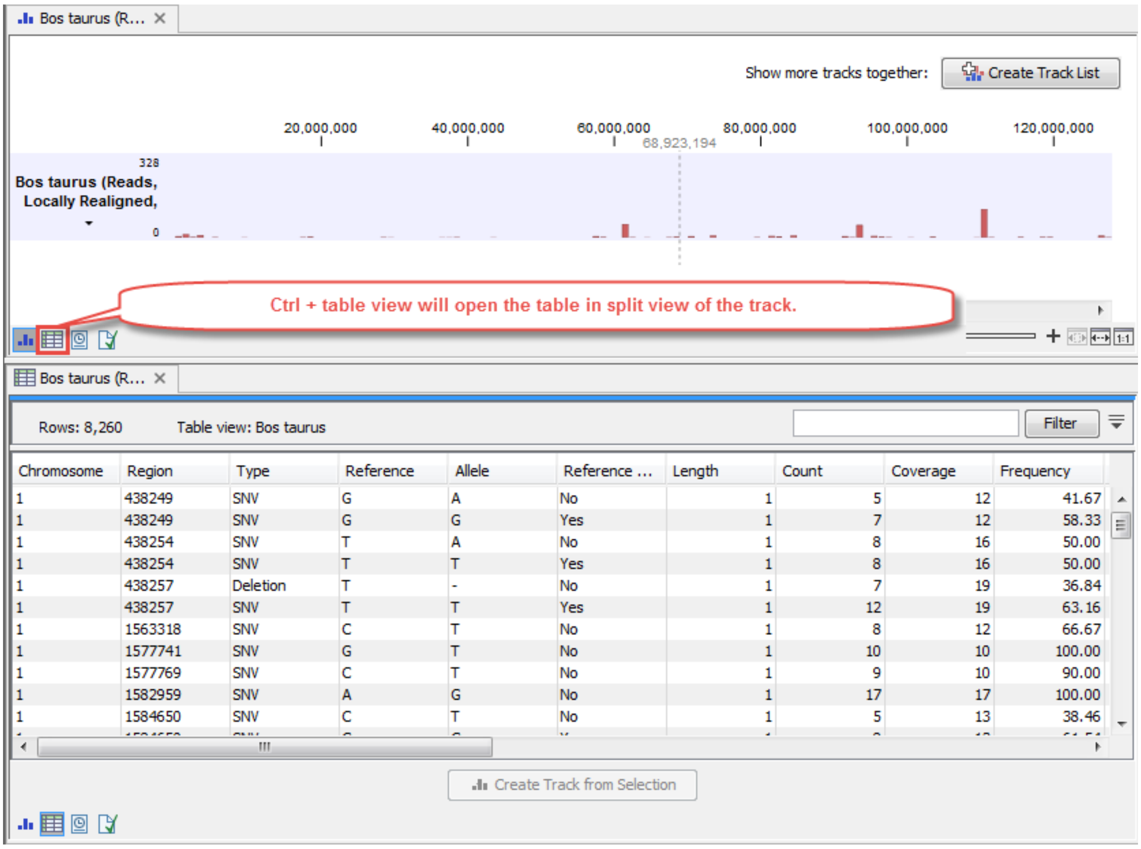
Let's take a look at the one of the variant tracks. Open the **Bos taurus (Reads) - locally realigned (Variants)** track by double-clicking on the name in the **Navigation Area**. Use the lower left corner



- Comparative analysis tutorial
  - reads
  - reference data
  - analysis
    - Bos taurus batch
      - Bos taurus (Reads)
      - Bos taurus mapping summary report
      - Bos taurus (Reads, Locally Realigned)
      - Bos taurus (Reads) - realigned regions
      - Bos taurus (Reads, Locally Realigned, Variants)
    - Bison bison batch
    - Bos indicus batch

Figure 15: *Example of how all outputs have been saved to the batch specific folder created during the first batch analysis.*

buttons to open the variant track in table view. Open the table in split view by holding down the Ctrl key (Cmd on Mac) while clicking on the table icon (see the red arrow in figure 16). You will be able to see the variant track and the variant table in the same window.



Ctrl + table view will open the table in split view of the track.

Rows: 8,260 Table view: Bos taurus

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Count	Coverage	Frequency
1	438249	SNV	G	A	No	1	5	12	41.67
1	438249	SNV	G	G	Yes	1	7	12	58.33
1	438254	SNV	T	A	No	1	8	16	50.00
1	438254	SNV	T	T	Yes	1	8	16	50.00
1	438257	Deletion	T	-	No	1	7	19	36.84
1	438257	SNV	T	T	Yes	1	12	19	63.16
1	1563318	SNV	C	T	No	1	8	12	66.67
1	1577741	SNV	G	T	No	1	10	10	100.00
1	1577769	SNV	C	T	No	1	9	10	90.00
1	1582959	SNV	A	G	No	1	17	17	100.00
1	1584650	SNV	C	T	No	1	5	13	38.46

Figure 16: *Open the variant table in split view.*

The variant caller identifies around 8,000 variants in the bovine genome, about 13,000 in zebu, and about 23,000 in bison.

Identify Species specific Variants

We are now going to identify variants that are unique for the zebu species with the tool **Identify known mutations from sample mapping**.

1. Go to:

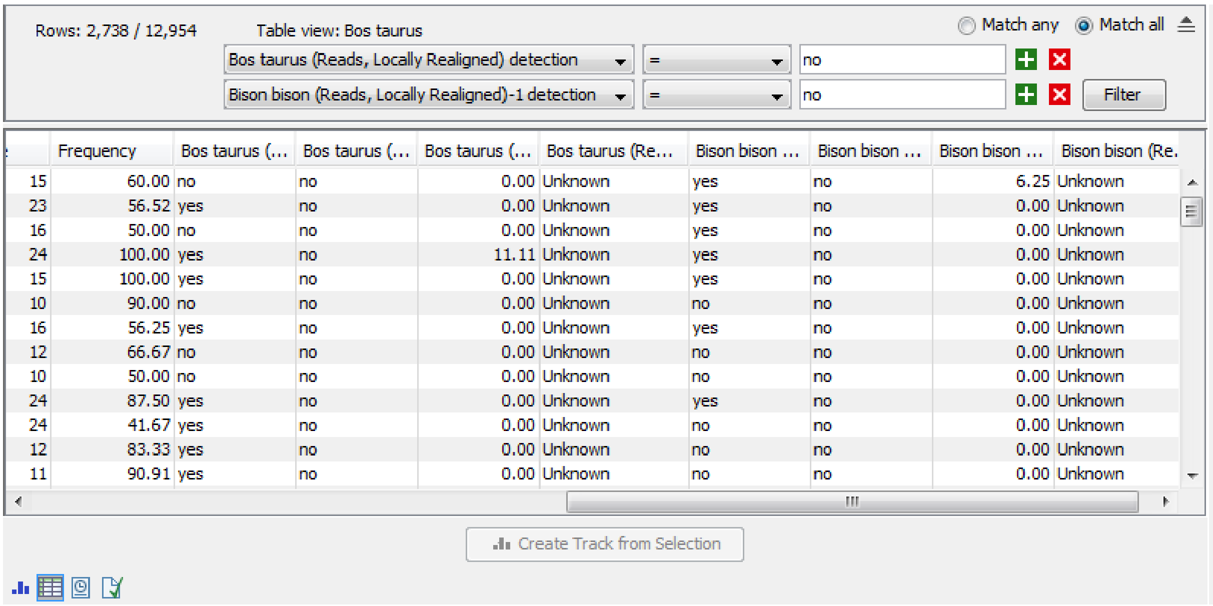
   Toolbox | **Resequencing Analysis** () | **Identify known mutations from sample mapping** ()

2. In the first wizard dialog, select the read mappings of *Bos taurus* and *Bison bison*. Click Next.

3. Select the zebu variant track **Bos indicus (Reads) - locally realigned (Variants)**. Leave the other parameters at their default value and click on the button labeled Next.

4. Create individual tracks and an overview track. Choose to Save them in a specified location (for example a new folder called zebu specific variants), and click on Finish.

5. Now open the file **Bos indicus (Reads, Locally Realigned, Variants) (MutationTest overview)**. We want to filter away from the table the variants that were also detected in the two other species, and keep only the ones that were not (the detection value is set to no). Open the filter and select "Bos taurus detection" in the first field, "=" in the second, and type "no" in the third field. Add an additional filter that you set up as you did the previous one but for "Bison bison detection" (see figure 17).



Rows: 2,738 / 12,954 Table view: Bos taurus

○ Match any ● Match all

Bos taurus (Reads, Locally Realigned) detection | = | no

Bison bison (Reads, Locally Realigned)-1 detection | = | no

	Frequency	Bos taurus (...	Bos taurus (...	Bos taurus (...	Bos taurus (Re...	Bison bison ...	Bison bison ...	Bison bison ...	Bison bison (Re.
15	60.00	no	no	0.00	Unknown	yes	no	6.25	Unknown
23	56.52	yes	no	0.00	Unknown	yes	no	0.00	Unknown
16	50.00	no	no	0.00	Unknown	yes	no	0.00	Unknown
24	100.00	yes	no	11.11	Unknown	yes	no	0.00	Unknown
15	100.00	yes	no	0.00	Unknown	yes	no	0.00	Unknown
10	90.00	no	no	0.00	Unknown	no	no	0.00	Unknown
16	56.25	yes	no	0.00	Unknown	yes	no	0.00	Unknown
12	66.67	no	no	0.00	Unknown	no	no	0.00	Unknown
10	50.00	no	no	0.00	Unknown	no	no	0.00	Unknown
24	87.50	yes	no	0.00	Unknown	yes	no	0.00	Unknown
24	41.67	yes	no	0.00	Unknown	no	no	0.00	Unknown
12	83.33	yes	no	0.00	Unknown	no	no	0.00	Unknown
11	90.91	yes	no	0.00	Unknown	no	no	0.00	Unknown

Figure 17: *You can filter for zebu specific variants in the overview table.*

6. Click on the button "Filter", select all remaining rows and click on "Create Track from Selection". Save the new track in the Navigation Area.

The output from this analysis is a new variant track that holds only the variants (less than 3000) that are specific for the zebu species Bos indicus called "(Reads, Locally Realigned, Variants) (MutationTest overview)_selection". You can also see that you have different types of variation, some of which are heterozygous and others homozygous. You can export tables from the Workbench in Excel formats or text formats. Click on the table so that the table you would like to export is the selected view, then use the **Export** () button in the toolbar to start the export. For small tables, you can also just **Copy** () the contents of the table and paste into a spreadsheet for further processing.

We will now look only at the variants that are present in the coding part of the genome. We can do this with the tool **Filter Based on Overlap**.

1. Go to:

   Toolbox | **Track Tools** () | **Annotate and Filter** () | **Filter Based on Overlap** ()

2. In the first wizard step, select the zebu specific variant track that you just saved as input. Click on the button labeled Next.

3. Select the **Bos taurus (Hereford)_CDS** track for overlap comparison and **Keep annotations that overlap**. Click on the button labeled Next.

4. Choose to save the results and click on the button labeled **Finish**.

The result is a variant track called "(Reads, Locally Realigned, Variants) (MutationTest overview)_selection (OF)" and containing a number in the range of 600 zebu-specific variants that occurs in the CDS regions.

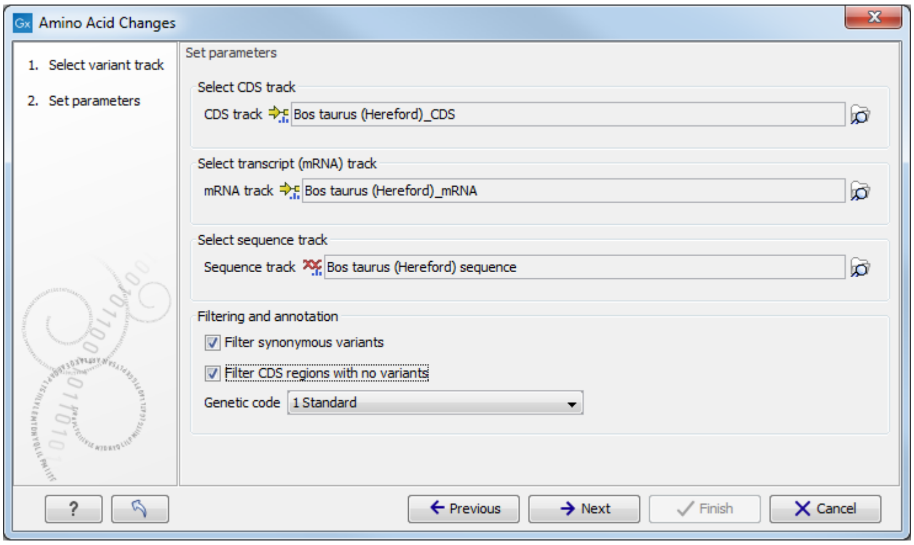
Now we would now like to look only at variants that cause amino acid changes. We can do this with the tool **Amino Acid Changes**.

1. To run the analysis, go to the toolbox:

   Toolbox | **Resequencing Analysis** () | **Functional Consequences** ()| **Amino Acid Changes** ()

2. In the first wizard step, select the most recent zebu specific variant track (the one filtered based on overlap with CDS) as input. Click on the button labeled Next.

3. Select the CDS, mRNA, and sequence tracks as shown in figure 18 and choose to detect only non-synonymous changes by ticking the box "Filter synonymous". Leave also the option "Filter CDS regions with no variants" checked. Click on the button labeled Next.



Set parameters

Select CDS track

CDS track Bos taurus (Hereford)_CDS

Select transcript (mRNA) track

mRNA track Bos taurus (Hereford)_mRNA

Select sequence track

Sequence track Bos taurus (Hereford) sequence

Filtering and annotation

☑ Filter synonymous variants

☑ Filter CDS regions with no variants

Genetic code 1 Standard

Figure 18: *Select CDS, mRNA, and sequence tracks and check Filter synonymous and Filter CDS regions with no variants.*

4. Choose to save the results and click on the button labeled **Finish**.

The result is a variant track called "(Reads, Locally Realigned, Variants) (MutationTest overview)_selection (OF, AAC)" with over 200 non-synonymous zebu-specific variants found in the CDS regions that cause amino acid changes.

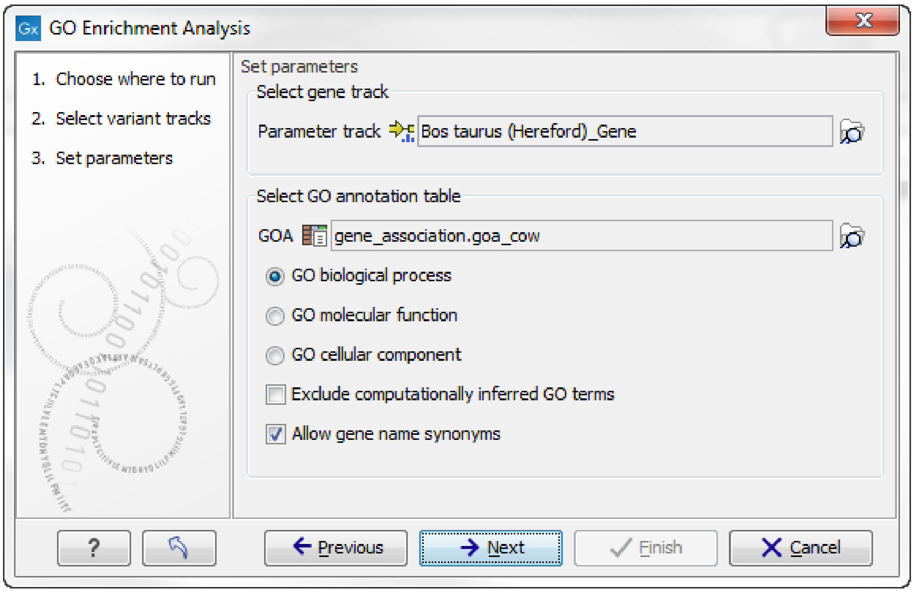
Overrepresentation analysis

To gain more insight into the underlying biology we are now going to perform a gene ontology overrepresentation analysis with the tool **GO Enrichment Analysis**. The Gene Ontology site `http://www.geneontology.org/GO.downloads.annotations.shtml` contains the bovine gene ontology annotation.

1. Go to `http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associatio` `gene_association.goa_cow.gz?rev=HEAD` and download the bovine gene ontology annotation to your computer (1.6 MB).
2. Import the file into the Workbench with the import function found in the toolbar:
   Import | **Standard Import** .
3. Select **Gene Ontology Annotation File** as **Format** in the import wizard (or alternatively use **Automatic Import**, which will give the same result). The file does not have to be unzipped before import.
4. To analyze if any biological pathways are overrepresented for the detected zebu-unique variants, go to the toolbox and run the tool:
   Toolbox | **Resequencing Analysis** ( ) | **Functional Consequences** ( )| **GO Enrichment Analysis** ( )
5. In the first wizard step, select the variant track that was created with the **Amino Acid Changes** tool **Bos indicus (Reads) - locally realigned (Variants, CTRL, OF, AAC)** as input. Click on the button labeled Next.
6. Select the parameter track **Bos taurus (Hereford)_Gene** and the GO annotation table, uncheck **Exclude computationally inferred GO terms**, and check **Allow gene name synonyms** as shown in figure 19. Click on the button labeled Next.
7. Choose to save the results and click on the button labeled **Finish**.

You get two different types of output: a **GO-enrichment table** and a variant track that is essentially the same variant table as the one generated in the **Amino Acid Changes** analysis but with five extra columns (see the red box in the side panel in figure 20).

Open the GO-enrichment table by clicking on the file name in the **Navigation Area**. We can sort the rows by the calculated p-value (from a hypergeometric distribution test comparing frequency of occurrence in a sample vs in all genes) by clicking once on the column header "P-values". It is clear that in the top rows the pathways related to the immune response dominate (figure 21). This matches previous findings showing that zebu responds differently to some infections and also has a better innate immune response compared to the taurine and the bison species [Freeman et al., 2008].



GO Enrichment Analysis

1. Choose where to run
2. Select variant tracks
3. Set parameters

Set parameters

Select gene track

Parameter track: Bos taurus (Hereford)_Gene

Select GO annotation table

GOA: gene_association.goa_cow

● GO biological process

○ GO molecular function

○ GO cellular component

□ Exclude computationally inferred GO terms

☑ Allow gene name synonyms

Figure 19: *Select parameter track and GO annotation table as specified.*

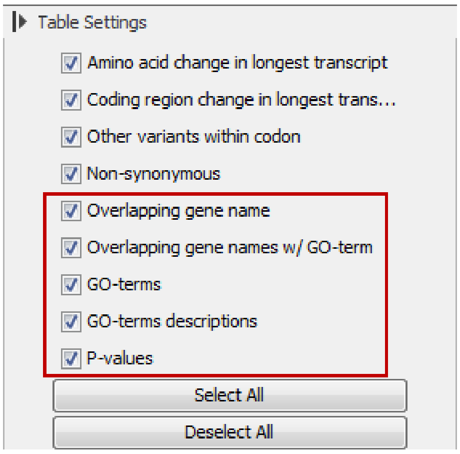


Table Settings

☑ Amino acid change in longest transcript

☑ Coding region change in longest trans...

☑ Other variants within codon

☑ Non-synonymous

☑ Overlapping gene name

☑ Overlapping gene names w/ GO-term

☑ GO-terms

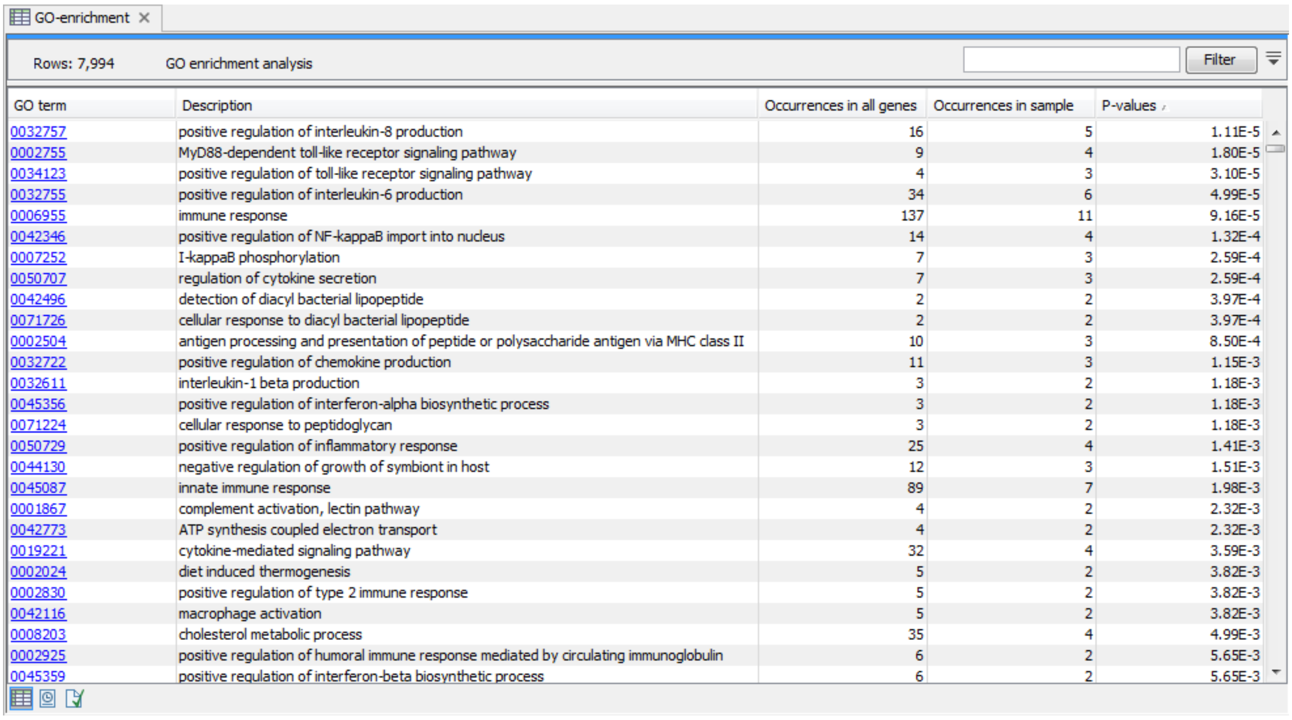
☑ GO-terms descriptions

☑ P-values

Figure 20: *Five extra columns have been added compared to the Amino Acid Changes table.*

Automating the analysis as workflows

Some of the steps in this analysis can be combined into workflows. This is shown in the two examples in (figure 22 and figure 23). These workflows, combined with batch mode of operation, greatly simplify the analysis, and can be used to repeat it, for instance, to look for similar variant enrichment in the other two species of bovines. For further information about how to create and run workflows, please see the tutorial "An Introduction to Workflows" that can be found here: http://www.clcbio.com/files/tutorials/Workflow-intro.pdf.



Rows: 7,994 GO enrichment analysis

GO term	Description	Occurrences in all genes	Occurrences in sample	P-values
0032757	positive regulation of interleukin-8 production	16	5	1.11E-5
0002755	MyD88-dependent toll-like receptor signaling pathway	9	4	1.80E-5
0034123	positive regulation of toll-like receptor signaling pathway	4	3	3.10E-5
0032755	positive regulation of interleukin-6 production	34	6	4.99E-5
0006955	immune response	137	11	9.16E-5
0042346	positive regulation of NF-kappaB import into nucleus	14	4	1.32E-4
0007252	I-kappaB phosphorylation	7	3	2.59E-4
0050707	regulation of cytokine secretion	7	3	2.59E-4
0042496	detection of diacyl bacterial lipopeptide	2	2	3.97E-4
0071726	cellular response to diacyl bacterial lipopeptide	2	2	3.97E-4
0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	10	3	8.50E-4
0032722	positive regulation of chemokine production	11	3	1.15E-3
0032611	interleukin-1 beta production	3	2	1.18E-3
0045356	positive regulation of interferon-alpha biosynthetic process	3	2	1.18E-3
0071224	cellular response to peptidoglycan	3	2	1.18E-3
0050729	positive regulation of inflammatory response	25	4	1.41E-3
0044130	negative regulation of growth of symbiont in host	12	3	1.51E-3
0045087	innate immune response	89	7	1.98E-3
0001867	complement activation, lectin pathway	4	2	2.32E-3
0042773	ATP synthesis coupled electron transport	4	2	2.32E-3
0019221	cytokine-mediated signaling pathway	32	4	3.59E-3
0002024	diet induced thermogenesis	5	2	3.82E-3
0002830	positive regulation of type 2 immune response	5	2	3.82E-3
0042116	macrophage activation	5	2	3.82E-3
0008203	cholesterol metabolic process	35	4	4.99E-3
0002925	positive regulation of humoral immune response mediated by circulating immunoglobulin	6	2	5.65E-3
0045359	positive regulation of interferon-beta biosynthetic process	6	2	5.65E-3

Figure 21: *Sort on p-value by clicking once on the column header.*

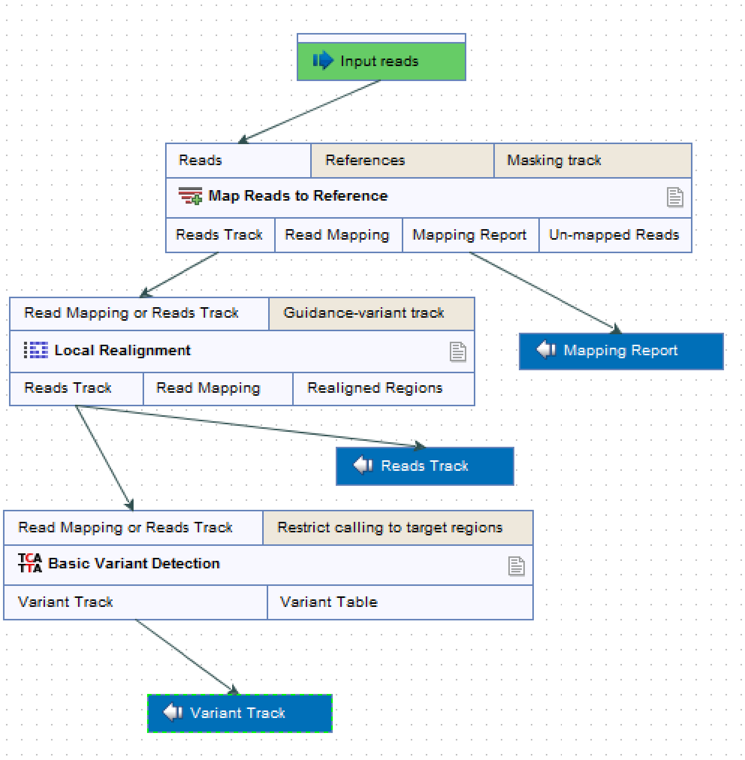


Figure 22: *Mapping of the reads and variant calling workflow.*

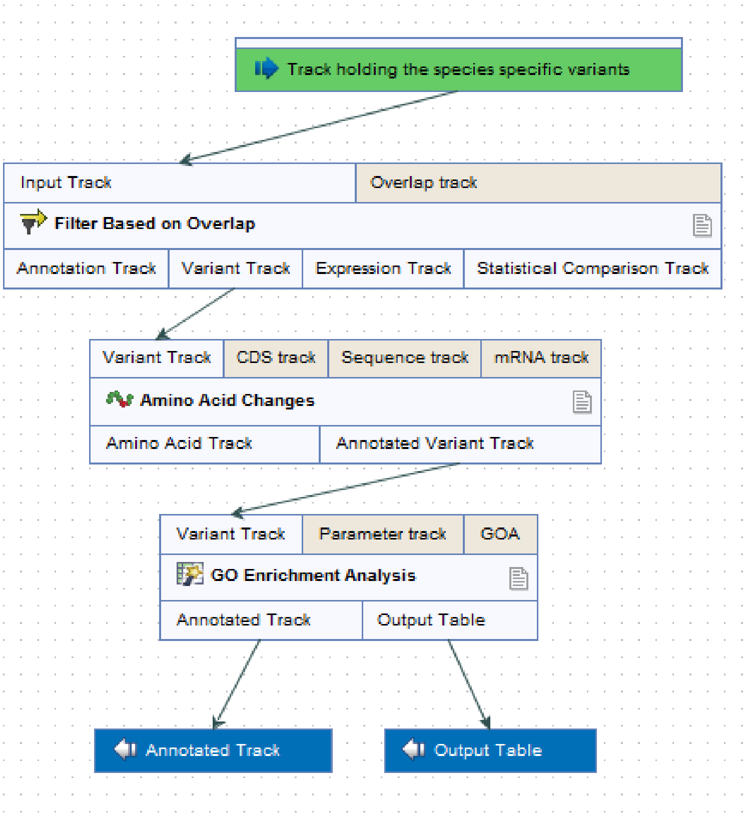


Figure 23: *Variant filtering and annotation workflow.*

Bibliography

[Cosart et al., 2011] Cosart, T., Beja-Pereira, A., Chen, S., Ng, S. B., Shendure, J., and Luikart, G. (2011). Exome-wide dna capture and next generation sequencing in domestic and wild species. *BMC Genomics*, 12:347.

[Freeman et al., 2008] Freeman, A. R., Lynn, D. J., Murray, C., and Bradley, D. G. (2008). Detecting the effects of selection at the population level in six bovine immune genes. *BMC Genet*, 9:62.