



Application Note

Comparative analysis of three bovine genomes



Sample to Insight

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

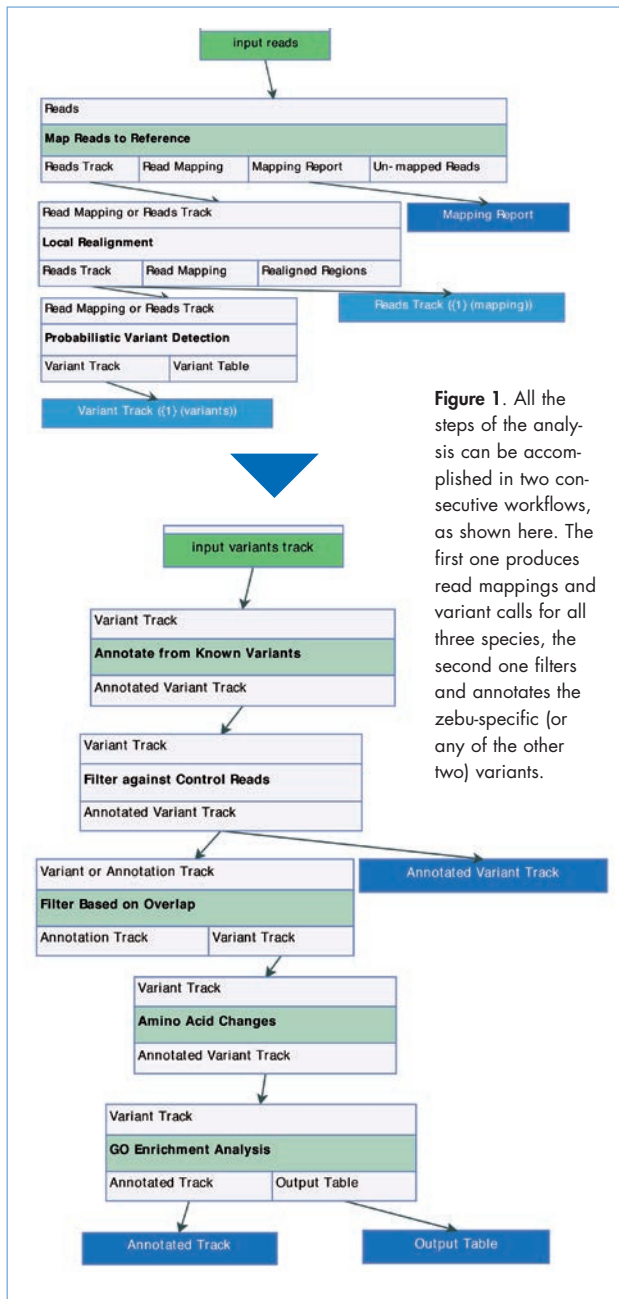


Figure 1. All the steps of the analysis can be accomplished in two consecutive workflows, as shown here. The first one produces read mappings and variant calls for all three species, the second one filters and annotates the zebu-specific (or any of the other two) variants.

Data

- Short reads for all three cattle species¹
- Bovine reference sequence, bovine gene annotations and bovine variation data from dbSNP/Ensembl²
- The GO annotation file for the bovine genome³

Software: CLC Genomics Workbench 7.0

Analysis Workflow

Preparing datasets

The bovine genome with annotations is downloaded using the automatic genome download in CLC Genomics Workbench. Reads are imported using Illumina import. Files can be imported simultaneously. We suggest appending species names to downloaded fastq files to simplify recognition of the species.

We import the bovine annotation file to the Workbench using the standard import tool for Gene Ontology Annotation File. To evaluate the need for trimming of sequencing reads prior to read mapping, a sequencing QC report is created. As examination of the report reveals no serious issues, the sequencing data can be used for read mapping without trimming. The sequencing QC report is run with default options.

Mapping the reads and detection of variants

The sequencing reads for each species are mapped separately to the reference genome. Read mappings are run with default parameter settings and with track and mapping reports as output. Mappings are then locally realigned with the Local Realignment tool. Mappings are realigned without guidance tracks. SNPs and small insertions and deletions are identified in the mapped sequencing reads by using the Probabilistic Variant Detection caller. Variant detection is run with default parameter settings.

Annotation of variants

By using the Annotate From Known Variants tools, additional information can be added to those called variants in the variant table that overlap with dbSNP variants track downloaded together with the reference genome.

Variant filtering

In this experiment, we are interested in finding zebu-specific variations. If only the variant tracks were available (i.e. imported from an external source), we could use the Filter Against Known Variants tool. Zebu should be set as target track, both bovine and bison chosen as known, and variants with no exact match in known tracks should be kept. Since we have the mappings as well, we can use the bovine and bison ones as control reads to filter the zebu-specific variants. The Filter against Control Reads tool is preferred as some variants may not have been called and can still be present in reads below the detection thresholds of the variant caller.

Further, we select only the variants that are present in the coding part of the genome. The Filter Based on Overlap Track tool and the CDS track from the reference are used. Filtering by amino acid changes using the Amino Acid Changes tool can be used to further reduce the number of variants by selecting only those variations that lead to non-synonymous changes in the coding part of the genome.

Overrepresentation analysis

It could be interesting to investigate what the genes that only show variations in zebu have in common. This could indicate altered pathways in the zebu species. To analyse if any

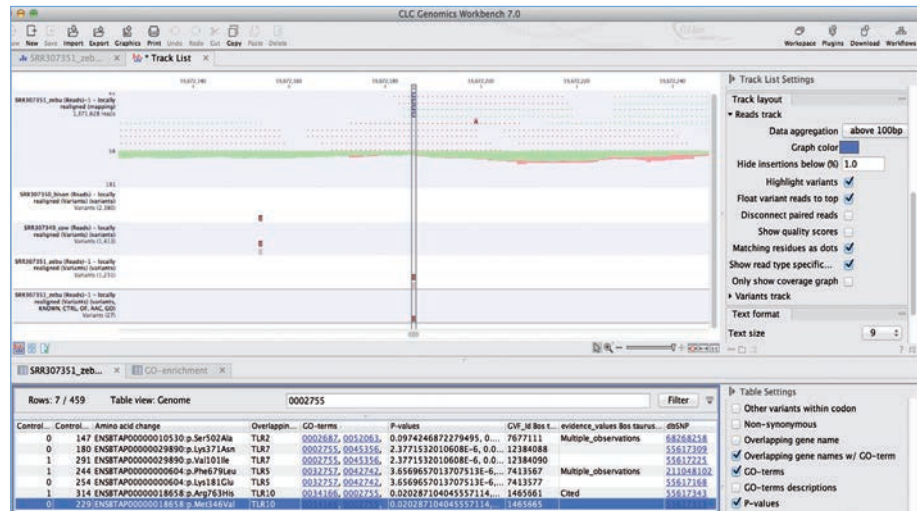


Figure 2. Track list view of zebu read mapping, variant tracks for three species, and filtered and annotated zebu-specific variant track. Partial table view of the latter shows additional annotation columns (amino acid change, GO terms, dbSNP matches); variants in GO entry 0002755 selected.

biological pathways are overrepresented in the zebu-unique variants, we use the GO Enrichment Analysis tool. The reference track needs to be specified with gene names and the GO track. In the resulting GO enrichment table, rows can be sorted by the calculated p-value. P-values are calculated from a hypergeometric distribution test comparing frequency of occurrence in one sample versus in all genes.

Results

In all three sets of read mappings, more than 99% of reads are mapped successfully. The variant caller identifies approximately 38,000 variants in each of the bovine and zebu genomes, and around 72,000 in the bison genome. And when the read mappings from bovine and bison species are used as control reads to filter variants only represented in zebu, it leaves 4733 unique zebu variants. Filtering of zebu-specific variants for those present in the coding regions and resulting in an amino acid change, leaves 459 variants.

Figure 2 shows a track list with zebu mapping, three original variant tracks and the final result of filtering; the bottom

part shows the tabular view of the final variant track. In that view, a subset is chosen with a 'filter' menu where GO-term for one of the pathways is chosen, and one variant row selected. The top view automatically zooms to a corresponding position in the track list, allowing easy examination of the variant. The tabular view also shows how various annotation and filtering steps add additional columns of information to the variant track table.

In the resulting table from the GO enrichment analysis sorted by p-values, the top rows are clearly dominated by pathways related to immune response (see Table 1). This matches previous findings showing that the zebu species responds differently to some infections compared to bison and taurine. Zebu also has a better innate immune response compared to the other two species (Freeman et al. 2008).

Automating the analysis as workflows

Most of the steps in this analysis can be combined into workflows. 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 bovinds.

More detailed steps of the analysis presented in this application note are also available in our online tutorial.

| GO term | Description | Occurrence | Occurrence... | Genes (overlap) | P-values |
|---------|---|------------|---------------|--|----------|
| 0002755 | MyD88-dependent toll-like receptor signaling pathway | 6 | 4 | TLR10, TLR2, TLR5, TLR7 | 7.38E-6 |
| 0032753 | positive regulation of interleukin-6 production | 22 | 6 | IL18, IL32, NOD2, TLR2, TLR3, TLR7 | 3.64E-6 |
| 0034123 | positive regulation of interleukin-8 production | 13 | 5 | IL18, TLR2, TLR3, TLR5, TLR7 | 3.66E-6 |
| 0042346 | positive regulation of toll-like receptor signaling pathway | 4 | 3 | TLR2, TLR3, TLR5 | 3.23E-5 |
| 0042346 | positive regulation of NF-kappaB import into nucleus | 14 | 4 | IL18, TLR2, TLR3, TLR7 | 1.40E-4 |
| 0002252 | I-kappaB phosphorylation | 6 | 3 | TLR2, TLR3, TLR7 | 1.57E-4 |
| 0032722 | positive regulation of chemokine production | 7 | 3 | TLR2, TLR3, TLR7 | 2.70E-4 |
| 0050707 | regulation of cytokine secretion | 7 | 3 | TLR10, TLR2, TLR5 | 2.70E-4 |
| 0071224 | cellular response to peptidoglycan | 2 | 2 | NOD2, TLR2 | 4.09E-4 |
| 0050729 | positive regulation of inflammatory response | 19 | 4 | IL32, TLR10, TLR2, TLR3 | 4.99E-4 |
| 0044130 | negative regulation of growth of symbiont in host | 11 | 3 | MBL, NOD2, TLR2 | 1.20E-3 |
| 0002830 | positive regulation of type 2 immune response | 3 | 2 | IL33, NOD2 | 1.21E-3 |
| 0045356 | positive regulation of interferon-alpha biosynthetic process | 3 | 2 | TLR3, TLR7 | 1.21E-3 |
| 0042773 | ATP synthesis coupled electron transport | 4 | 2 | ND4L, ND5 | 2.39E-3 |
| 0048016 | inositol phosphate-mediated signaling | 4 | 2 | EDN2, NMR2 | 2.39E-3 |
| 0050766 | positive regulation of phagocytosis | 14 | 3 | MBL, NOD2, SLC11A1 | 2.53E-3 |
| 0071260 | cellular response to mechanical stimulus | 30 | 4 | IL18, TLR3, TLR5, TLR7 | 2.97E-3 |
| 0001867 | complement activation, lectin pathway | 5 | 2 | KRT1, MBL | 3.93E-3 |
| 0002024 | diet induced thermogenesis | 5 | 2 | ADRB2, CLIC5 | 3.93E-3 |
| 0042116 | macrophage activation | 5 | 2 | EDN2, SLC11A1 | 3.93E-3 |
| 0045080 | positive regulation of chemokine biosynthetic process | 5 | 2 | IL18, TLR3 | 3.93E-3 |
| 0045359 | positive regulation of interferon-beta biosynthetic process | 5 | 2 | TLR3, TLR7 | 3.93E-3 |
| 0006955 | immune response | 101 | 7 | BOLA-DQA5, BOLA, BOLA-DQB, CSF2, IL18, IL3, T... | 4.35E-3 |
| 0051092 | positive regulation of NF-kappaB transcription factor activity | 55 | 5 | IL18, NLRP3, NOD2, TLR2, TLR3 | 4.97E-3 |
| 0002925 | positive regulation of humoral immune response mediated by circulating immunogl... | 6 | 2 | NOD2, PTPRC | 5.82E-3 |
| 0006120 | mitochondrial electron transport, NADH to ubiquinone | 6 | 2 | MT-ND4, ND2 | 5.82E-3 |
| 0051607 | defense response to virus | 58 | 5 | IFNB2, NLRP3, PTPRC, TLR3, TLR7 | 6.24E-3 |
| 0019221 | cytokine-mediated signaling pathway | 19 | 3 | CCL2, EDN2, IL18 | 6.26E-3 |
| 0032760 | positive regulation of tumor necrosis factor production | 19 | 3 | NOD2, TLR2, TLR3 | 6.26E-3 |
| 0050830 | defense response to Gram-positive bacterium | 19 | 3 | MBL, NOD2, TLR2 | 6.26E-3 |
| 0006954 | inflammatory response | 86 | 6 | CCL2, NLRP3, SLC11A1, TLR10, TLR2, TLR3 | 7.88E-3 |
| 0002504 | antigen processing and presentation of peptide or polysaccharide antigen via MHC... | 7 | 2 | BOLA-DQA5, BOLA-DQB | 8.04E-3 |
| 0045078 | positive regulation of interferon-gamma biosynthetic process | 7 | 2 | TLR3, TLR7 | 8.04E-3 |
| 0048246 | macrophage chemotaxis | 7 | 2 | CCL2, EDN2 | 8.04E-3 |
| 0050718 | positive regulation of interleukin-1 beta secretion | 7 | 2 | NLRP3, NOD2 | 8.04E-3 |
| 0050730 | regulation of peptidyl-tyrosine phosphorylation | 7 | 2 | HRG, PDGFR | 8.04E-3 |
| 0006465 | signal peptide processing | 8 | 2 | CLN5, LONP2 | 0.01 |
| 0008306 | associative learning | 8 | 2 | BTG2, CLN8 | 0.01 |
| 0003214 | cardiac left ventricle morphogenesis | 9 | 2 | CPE, NPYSR | 0.01 |
| 0007338 | single fertilization | 9 | 2 | OVGP1, SLC22A16 | 0.01 |

Table 1. Results of GO enrichment analysis sorted by p-value in increasing order. The top-scoring term is highlighted, and corresponds to variants selected in table view of Figure 1.

References

- Cosart T., Beja-Pereira A., Chen S., Ng SB., Shendure J., Luikart G. (2011) Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* 12:347
 - Freeman A.R., Lynn D.J., Murray C., Bradley D.G. (2008) Detecting the effects of selection at the population level in six bovine immune genes. *BMC Genetics* 9:62
- 1 Sequence reads for bovine, zebu and bison available at <http://www.ebi.ac.uk/ena/data/view/SRP007095>; .fastq.gz read files are single GA and GAI exome sequencing short reads
 - 2 Available for download within CLC Genomics Workbench using the Download Genome tool
 - 3 http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/gene_association.goa_cow.gz?rev=HEAD

