Application Note Comparative analysis of three bovine genomes

(۲

~~{

Application note on comparative genomics data analysis, using Illumina exome data from *Bos taurus* (taurine), *Bos indicus* (zebu), and *Bison bison* (bison), analyzed and visualized through Genomics Gateway in CLC Genomics Workbench



CLC bio © Copyright

Comparative genomics data analysis using Genomics Gateway

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

Data

- Short reads for all three cattle species¹
- Bovine reference sequence², bovine gene annotations³ and bovine variation data⁴ from dbSNP/Ensembl downloaded from the Ensembl ftp server

¹Sequence reads for taurine, zebu and bison available at http://www.ebi. ac.uk/ena/data/view/SRP007095

² Reference sequence UMD3.1.64 at ftp://ftp.ensembl.org/pub/release-64/ fasta/bos_taurus/dna/

³ ftp://ftp.ensembl.org/pub/release-64/gtf/bos_taurus/

⁴ ftp://ftp.ensembl.org/pub/release-64/variation/gvf/bos_taurus/

Analysis Workflow



Define your reference genome and add annotations

In the first step we have to create the reference genome including all annotations. We import the downloaded taurine fasta files into CLC Genomics Workbench and create a track for the genome sequence using the "Define reference genome" function of Genomics Gateway.

The newly created track can be shown in a track set view by clicking "Create track list". You can zoom and add more tracks from the navigation area using drag-and-drop.

Afterwards, a track for gene annotations is created. This is carried out using the "Import tracks from file" function after downloading the gene annotations from the Ensembl website. If the track has been created previously, you can drag-and-drop the file from the navigation area to the open track list file in the workspace. These steps can be repeated in order to add additional tracks of interest. In this case we added a track including the variation data from Ensembl.

Map sequencing reads to the reference genome

The sequencing reads from each of the three cattle species are mapped separately to the reference genome using a Length fraction of 0.8 and a Similarity of 0.9.

Subsequently, the resulting read mappings are renamed in order to recognize the origin:

- Zebu mapping (Reads)
- Taurine mapping (Reads)
- Bison mapping (Reads)

The newly created tracks should of course be added to the open track list.

Variant calling

SNPs as well as small insertions and deletions are identified in the mapped sequencing reads by using the Probabilistic Variant Caller, and all SNV tracks – in this case for taurine, zebu and bison – can be added to the open track list for a better overview.

Filtering, annotating and comparing the results

In this experiment, we are interested in finding zebu-specific variations. Therefore, we merge all variations from the three SNV tracks (taurine, zebu and bison) and we can easily identify the zebu-specific variations.

		1,877,790	1.00	NOR O	1.077.022	II IIITAL		1077,000		1.0	7.00E	
	Sala suggest Annia 275 Milliona Inden Brightung (Physics) 199		TGTGACTCANGATT TGTGAC	TCOTAGOANCTEGA	TCATACTORCCATARATT ASTCATACTORCCAGE ASTCATACTORCCAGE		ADAGAGA	- 54	CATTOON	GCASTS	AAOGGACATCATT AAOTGACATCATT AADOSACATCATT	
	Institut magning Baselin 425430 magni Institut Telepage and State	AATTAGTTOTO	TOTOACTCAACATT	TCOCADDAAE1CCA TCOCADDAAE1 TCOCADCAAET TCOCASCAACTCO	AUTCATACT FACCACAAATT CTGACCACAAATT CACAAATT		A CAGA CAAAA	ARTTARD	CAT DCC T	CADCULTU CIC G	A CODUADA TOATT A CODUADA TOATT AA CAGACA TOATT	
	Binan maging Binain 1997 Miles Rodine Bin Jack and Jack		TOTOACTCAADA TOTOACTCAADATT TOT	ADDAASTCOA DAACTCOA TOGOROTAACTCOA	A DT CATA CTORC TA CRAATT	CARARC STA				AGCASTO AGCASTO	AAGGGACATCATT AAGGTACATCATT ACATCATT	
	Tantin marks front the factor sectors (11) bases for perception (12)											
	Peter sugging (Fronte 1874) Vielation and an over 2011, Income The Charles of Party Party											
	time regits field the factor and and 100 house Tricherces, 100											
Magel 1995 have be	Taxing and Bases and American Taxing a second second state from the Decomposition of the											1
142		4.5.				2						Ŧ
0												
No. (5)10 5	****		19w									8
Triaste Tage	Advance Postan Sanahar Taja	Information	Ciallo Concep Region	ni Dipitala			are .	Amadoup	-	409	Telature to 0	T
1 25410 25410 25710 25710 25710	Posts SP Vortin SP I INTER SP		114 101 94.2 12408 602 %.525 36 29 %.4	See mapping (Teach) 50% See mapping (Teach) 50%, Tautro mapping (Teach) 50% Start mapping (Teach) 50%	Textre reports Printly SPN, Discoursing P	insto 3Va	GREAK JONE GREAK JONE GREAK JONE GREAK JONE	a factoria a factoria a segmente	141 54 142	1121003	400 (hereit weiten 400 (hereit weiten 400 (hereit weiten	100
1000	entric ar		36 30 200 3	Built rapping (Read) (RM	Contraction of the local distance		ORDUH BOVE	a Brittle			Contraction of the local division of the loc	
100.00	righter str		10 10 Miles	The second second second second	In one hanned a line of times have a second di	and the second distance of	CONTRACT OF STREET	a second		- North St.	And in the second	

Figure 1: Table view on the merged variation track including different kinds of annotations

As this is an exome enrichment experiment, we only filter variations overlapping known transcripts in Ensembl. The enriched regions available from Agilent, Illumina etc. can also be used for filtering overlapping variations. Variations leading to amino acid changes are also annotated.

To deduce which annotations are already known in external databases, the "Filter against database variations" tool can be used to annotate these variations. The database track to be used for this is the variation data track from Ensembl created in the step defining our reference genome.

The merged annotation and variation track can be shown in a tabular view. This shows all variations annotated with overlapping gene names, amino acid changes, links to the dbSNP database and information about in which dataset (origin) the variation is present (Figure 1).

Selecting zebu-specific variations

To identify zebu-specific alleles, we will search for variations present in zebu, but not in taurine or bison.

From the merged SNV track that includes all annotations, we can identify zebu-specific variations by filtering the interactive table. In the advanced filtering options, you can deselect all annotations for taurine and bison to only show those related to zebu. Applying this filtering, only 2,685 variations are

Uniprot ID	Description					
Q5MACO_BOVIN	Interleukin 1, beta					
IL3_BOVIN	Interleukin 3					
TLR10_BOVIN	Toll-like receptor 10					
Q56GY5_BOVIN	Toll-like receptor 5					
Table 1: Some genes involved in immune response with variations only found in zebu						



Figure 2: Filtering of zebu-specific variations in the interactive table

displayed (Figure 2). We selected all the zebu-specific variations and created a new track including only this information.

In the new table of zebu-specific variations, some SNVs are false positives due to insufficient coverage of these variations in taurine and bison samples. The false positives can be removed using the "Filter variations against control reads" (the threshold of 3 reads for the variation to be removed is applicable to this dataset).

Investigating the functional role of mutated genes in zebu

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.

The GO annotation file for Bos taurus can be downloaded from the gene ontology download site⁵. Then we use the GO enrichment tool with the zebuspecific variation track as input file. The background file is the downloaded GO annotation file.

⁵ http://www.geneontology.org/GO.downloads.annotations.shtml

Results

From 19,564 single nucleotide variations identified in zebu, bison and taurine, 2,685 variations were identified to be present in zebu only. 663 of them were found to be non-synonymous variations.

An enrichment analysis of all genes overlapping these variations showed that mainly pathways for immune and defense response are affected. Considering only genes with non-synonymous variations, immune response is ranked at the top. 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*).

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

CLC bio · EMEA Finlandsgade 10-12 Katrinebjerg · DK-8200 Aarhus N Denmark Phone: +45 7022 5509

CLC bio · Americas 10 Rogers St # 101 Cambridge · MA 02142 USA Phone: +1 (617) 945 0178 CLC bio · AsiaPac 69 · Lane 77 · Xin Ai Road · 7th fl. Neihu District · Taipei · Taiwan 114 Taiwan Phone: +886 2 2790 0799



