Nature Precedings : doi:10.1038/npre.2009.3102.1 : Posted 20 Apr 2009

THE ROLE OF HAVANA AND COMMUNITIES IN THE MANUAL CURATION OF UNFINISHED VERTEBRATE GENOMES

D.R. Carvalho-Silva, Chen CK, Frankish A, Gilbert J, Gordon L, Hunt T, Larbaoui M, Loveland JE, Mudge J, Sehra H, Snow C, Steward C, Suner MM, Thomas M, Wilming L, Harrow J

Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

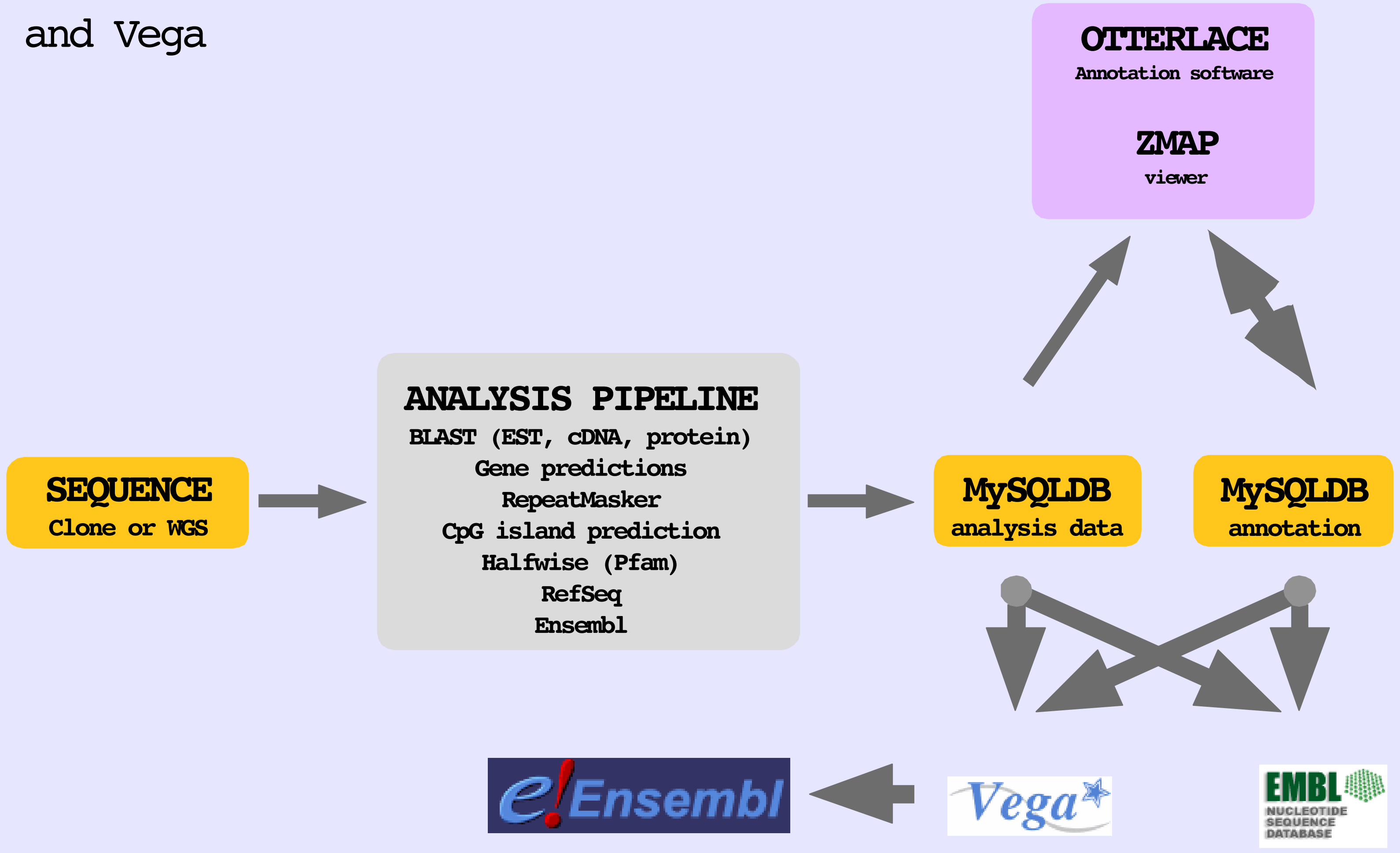
I) Introduction

The catalogue and classification of genes manually performed by specialised curators produce high quality annotation, but are costly, time consuming and used for just a handful of model genomes. Research communities working on other organisms rely on automated annotation and/or on annotation workshops (jamborees) where biologists come together with bioinformaticians to annotate their genes of interest. The HAVANA team at the Wellcome Trust Sanger Institute (WTSI) has hosted annotation jamborees for unfinished vertebrate genomes such as cow and pig in the last few years. With the next-generation sequencing technologies becoming more accessible, new genomes can be easily sequenced to a high depth coverage but will not be manually finished. For these genomes, manual annotation will cope better than the automated systems with the assembly problems inherent to unfinished sequences and therefore will be more accurate at identifying correct gene structures. Once this step is achieved, the curated genes will be incorporated and merged with the predicted genes in Ensembl to provide a consistent view of the landscape of vertebrate genomes.

II) Analysis and annotation pipeline

- Genomes analysed with an Ensembl-derived pipeline
- Analysis and annotation are stored in MySQL DBs
- Otterlace and Zmap communicate with the MySQL DBs
- Manual annotation feeding into and EMBL

and Vega



SEQUENCE
Clone or WGS

ANALYSIS PIPELINE
BLAST (EST, cDNA, protein)
Gene predictions
RepeatMasker
CpG island prediction
Halfwise (Pfam)
RefSeq
Ensembl

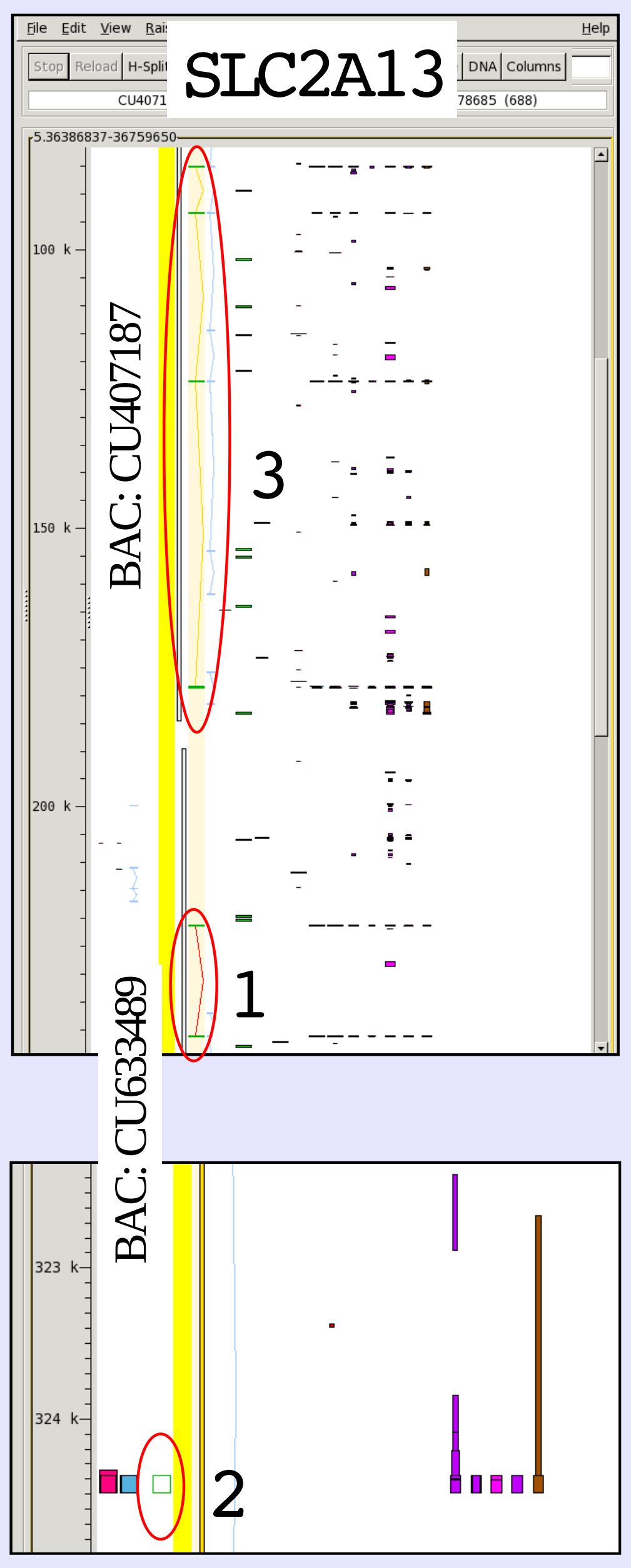
MySQLDB
analysis data

MySQLDB
annotation

OTTERLACE
Annotation software

ZMAP
viewer

III) Manual annotation scrutinise the pig assembly



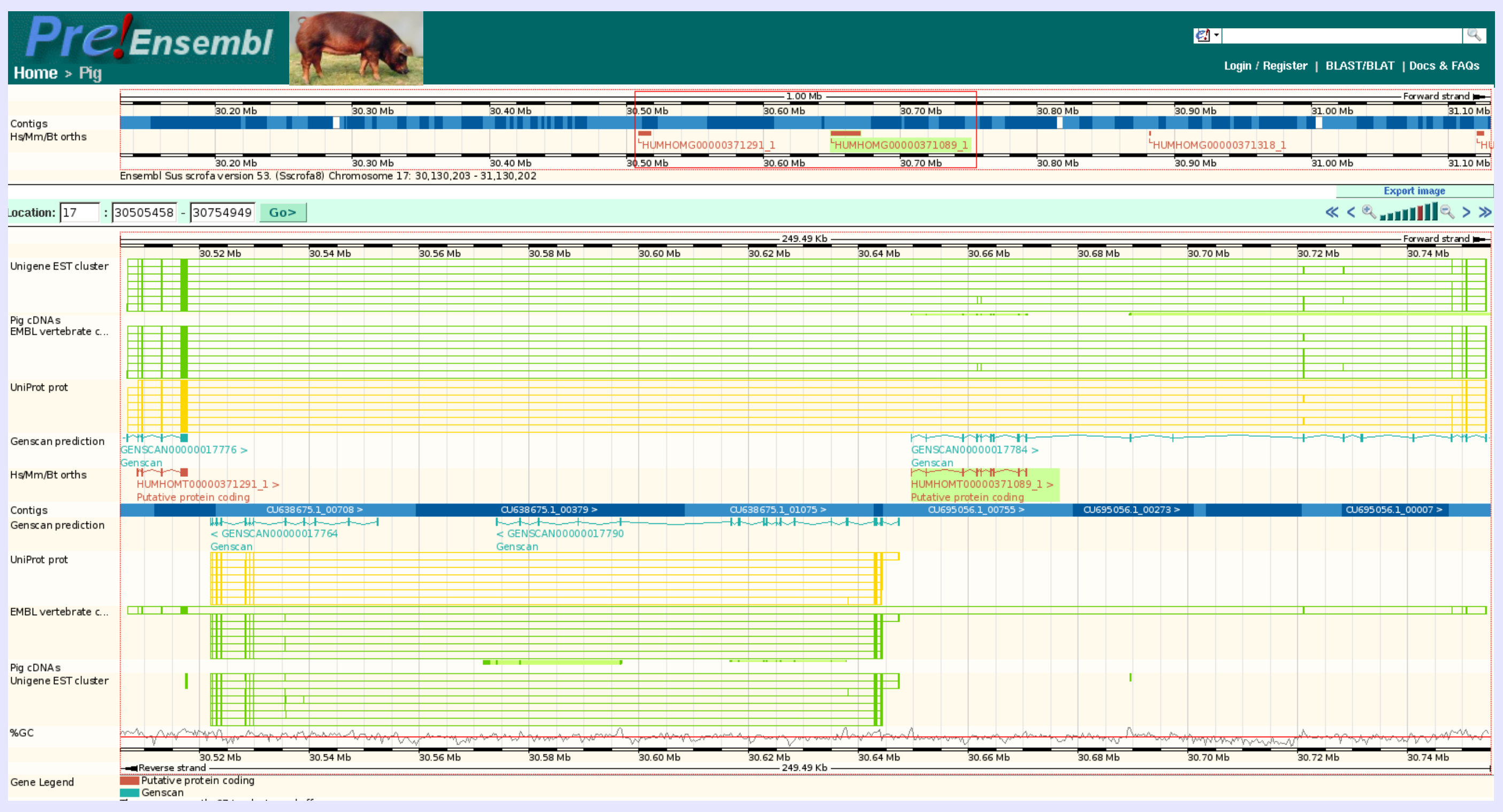
The fragments 1, 2 and 3 have been curated as part novel gene in pig which is homologous to the human *SLC2A13*. Two types of errors were detected:

A) Clones CU407187 and CU633489 are not in contiguous order. The sequence of fragment 1 corresponds to the 5' end of the human *SLC2A13*, whereas fragments 2 and 3 are homologous to the middle and 3' end of the human gene, respectively.

B) The annotated fragments of the pig *SLC2A13* are not located in the same strand. Both fragments 1 and 2 are on the same BAC but one is on the forward strand whereas the other is on the opposite strand.

In summary, BACs were placed in the wrong order and were also inverted

IV) Automated annotation in PreEnsembl



Initial BLAST analysis on the assemblies for chromosomes 1-18 and X of the pig genome has been performed and it is provided as an " early access" site on PreEnsembl. The preliminary gene build has been generated by alignments of the pig sequence to several public databases, *ab-initio* gene predictions, e-PCR based marker annotation and by aligning human peptides to the pig sequence. A new release of the pig assembly is scheduled for this month (April 2009).

V) The pig Jamboree at the WTSI

- 16th-18th July 2008
- 25 attendees
- 110 loci
- 148 variants

Aimed at teaching the pig community on how to use our annotation software and support them remotely, so they can complete the annotation of their regions of interest remotely.

VI) Concluding remarks

- Jamborees are a great opportunity for researchers to manually annotate genes in species with less funding;

- Low coverage genomes have higher numbers of poorly assembled regions, difficult to be annotated by automated means, but easily scrutinised by manual annotation;

- The current assembly of the pig genome is available in PreEnsembl. We envisage that soon the gene annotation in pig will be presented as a combination of Ensembl-Havana merged gene set as in human and mouse. The best of both manually and automated annotation worlds is upon us!