Application Number: 1 U54 HG004555-01

Project Title: Integrated human genome annotation: generation of a reference gene set

Data types and Data Release plans.

This document is a supplement to the above application, stating the data types that will be used and data release plan.

The principle output from this project will be coordinates of transcript structures on the reference human genome sequence. Each transcript will be labeled with metadata identifying its parent gene, its annotation status (what stage it has reached in the verification process involving computational, manual and experimental annotation) and its category, as defined by gene in C.1 (Known, Novel_CDS, Novel_transcript, Putative, Pseudogene, TEC) and by transcript under the categories given in D.2, Figure 7.

Both the consensus annotation and the raw annotation output from each contributing pipeline will be continuously available via DAS sources, so they can be displayed by any genome annotation application or website that is a DAS client. DAS sources can also be scripted against, allowing anyone who wishes to access the current annotation in real time. The consensus annotation set will combine functional transcripts and pseudogene annotation identified by metadata descriptions as described above (C.1).

All the DAS sources will be registered in the DAS registry (www.dasregistry.org) grouped together under the GENCODE project identifier. In addition lists of current DAS sources will be reported quarterly to the DCC.

In addition to this live data access, the project will also make date stamped dumps of the consensus annotation set in GTF format at the end of each quarterly reporting period and submit this to the DCC.

Part of the experimental verification of parts of the annotation will involve sequencing. The sequence of these products will be submitted to EMBL/GENBANK/DDBJ in a timely way and their accession numbers reported to the DCC. DAS sources showing these sequences aligned to the genome will also be provided.

March 2008

Dr Tim Hubbard
(Principle Investigator)
Head of Informatics
Wellcome Trust Sanger Institute

Mr David Davison
(Applicant Organisation Official)
Director of Corporate Services
Wellcome Trust Sanger Institute