RSSDB: A database of cryptic recombination signal sequences involved in V(D)J recombination.

Fabbri M.(1), Guffanti A.(2), Cocito A.(1,2), Furia L.(1), Fallini R.(2,3), McBlane F.(1,2)

(1) European Institute of Oncology, Milan Italy. (2) IFOM - The FIRC Institute of Molecular Oncology Foundation, Milan, Italy (3) Department of Molecular Sciences and Biotechnologies, University of Milan, Milano, Italy

Motivation

The antigen receptor repertoire is generated through rearrangements of Immunoglobulin (Ig) and T-cell receptor (TCR) gene segments into functional genes by a mechanism named V(D)J recombination. This is a two-step process: a cleavage step in which double-strand DNA cuts are made at specific sequences, followed by a joining step to repair the breaks. The coding segments of Ig and TCR genes are flanked by short recombination signal sequences (RSS). The RSS are recognized by a complex of the lymphocyte specific recombination proteins RAG1 and RAG2, which cleave the DNA between the coding sequence and the RSS. The broken coding strands are then rejoined to produce a rearranged gene. Mistakes in this process can generate chromosomal translocations that are involved in acute lymphoid leukemias and non-Hodgkin lymphomas. These errors include cleavage of DNA sequences, termed cryptic RSS, similar to functional RSS but located outside the Ig or TCR loci. We have screened the human and mouse genomes for the presence of putative cryptic RSS. To provide an initial list enriched in cryptic RSS, we used an original search algorithm. This primary set was then further filtered for biologically functional sequences using a published method. We have created a web-accessible database containing these putative recombination signals in the genome context. This is the first repository containing a genome-wide collection of RSS sequences. These sequence tags can be retrieved from a number of starting points including RSS type (with 12- or 23 bp spacer), chromosomal region, cytoband and gene identity. For visualization of our RSS search results, we have chosen to rely on an existing genome annotation knowledgebase and correlate our results with the gene structure, analysis, annotation and browsing features of the UCSC Genome Browser. Sequences of interest may also be searched for the occurrence of RSS and the corresponding tracks searched from within the genome database.

Methods

There are two types of canonical RSS sequences, either 39 or 28 bases long. However, these are highly degenerate and are therefore predicted to occur thousands of times throughout the genome. Both contain a conserved heptamer and a conserved nonamer separated by a less conserved spacer that is 23 or 12 bases long. In order to identify potentially functional RSS in the genome we performed a position-specific weighted motif match search, for which we designed a program (DNAGrab). Although this is not a novel algorithmic approach, our implementation is very fast (we can scan the entire human genome with the two profiles in one pass in less than 15 minutes on a mid-class PC) and provides a number of practical features that are not found in other programs. To build the RSS profile, we used a non-redundant set of functional RSSs. We searched human and mouse genomes with two profiles that are formed by the consensus for the heptamer and a consensus for the nonamer divided by a gap of 12 or 23 bases. To provide a list enriched in functional RSS, the DNAGrab results were further filtered by the RIC algorithm designed by Cowell and colleagues (2002).

Results

Considering both strands and both RSS types, DNAGrab searches produced 80,719,266 hits in human and 62,611,192 in mouse (approximately one every 40 bp). Of these only a small percentage passed the RIC statistical filter based on highest similarity to functional RSS (3,036,818 in human and 2,180,317 in mouse; approximately one every 1,000 to 1,200 bp). Since very few of the RSS known to participate in rare translocations passed the RIC threshold, we provide the filter as an option. Users of our database can obtain the selected RSS search data in UCSC-uploadable format or as a tab-separated table. The data will be automatically converted in BED format, suitable for uploading as a Custom Track to the Genome Browser. A text page with the track is written (transparently to the user) on the local bioinformatic server and the link is sent to the UCSC Genome Browser which will load and display the traces. Since the incoming Internet address seen by the UCSC Genome Browser corresponds to our Bioinformatic Server, external users that query our database will be able to visualize the tracks. Users will be able to navigate the Genome Browser (searching genes, locating chromosomal regions etc) and look at the density and location of our RSS matches with respect to the genomic features of their interest.

Contact email: marco.fabbri@ifom-ieo-campus.it

URL: http://bio.ieo-research.it/RSS/rss.html

Supplementary Information: Cowell LG, Davila M, Kepler TB, Kelsoe G. Genome Biol. 2002; 3(12) Identification and utilization of arbitrary correlations in models of recombination signal sequences.