UCbase 2.0

Manual

Center for Genome Research
University of Modena and Reggio Emilia, Italy
http://www.cgr.unimore.it

Contents

1 Introduction

Improvements over our previous version of UCbase include searching tools for UCR Id, Gene,Pathology and sequence. In particular, it is possible to visualize if the genes in which UCRs are located have SNPs, splicing events or if they are involved in specific pathologies. To interrogate UCbase for genetic or non-genetic disorders, we have chosen to use the standardised Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). SNOMED CT is a systematic, computer-processable collection of medical terms, in human and veterinary medicine that provides codes, terms, synonyms and definitions that cover anatomy and diseases. It allows a consistent way to index, store, retrieve, and aggregate medical data across specialties and sites of care. Moreover, we includes linked every SNOMED term to NCI Thesaurus and OMIM synonym when possible, in order to allow the researchers to retrieve more information regarding a specific disorder. New features include also a query to search for specific SNPs using dbSNP Ids. The results show the UCRs located within the gene that has that specific SNP or mutation together with chromosomal coordinates, allelic frequency, validation and phenotype information. Moreover, the new BLAST query allows to match a sequence against the entire UCRs sequence database and includes now a tool that gives the opportunity to filter the results for UCRs located in genes or chromosome regions involved in specific pathologies.

2

How to search UCRs

2.1 Search for UCRs by Id

UCRs are 481 sequences annotated by numbers: uc1, uc2, uc3, etc. It is also possible to search for many UCRs simultaneously clicking on the UCR Id and the ctrl (for Windows users) or cmd (for Mac users) button. Thus they can be searched by their Ids and other information such as:

1. Chromosome coordinates (Chr, Start, End, Strand, type*)
2. ENSEMBL Gene Id
3. Gene Symbol
4. SNPs (located inside or 500bp up/downstream a specific UCR). SNPs located inside (within) the UCRs are always shown at the top of the table.
5. Alternative Splicing sites
6. Pathology

*In field type "e" is exonic, "n" is non-exonic and "p" is possibly exonic.

2.2 Search for UCRs by SNPs

New features include also a query to search for specific SNPs using dbSNP Ids (`http://www.ncbi.nlm.nih.gov/SNP/`). A search for a specific SNP returns all UCRs in which the polymorphism is located (inside or 500bp upstream/downstream a specific UCR) together with chromosomal coordinates, allelic frequency, validation, and phenotype information.

2.3 Search for UCRs by Genes

UCRs can be retrieved also by searching for the genes in which they are located. "Hugo Gene Symbol" nomenclature has been used to annotate these genes. In particular, it is possible to visualize if these genes have SNPs, splicing events or if they are involved in specific pathologies.

2.4 Search for UCRs by Pathologies

To interrogate UCbase for genetic or non-genetic disorders, we have chosen to use the standardised Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). SNOMED CT is a systematic, computer-processable collection of medical terms, in human and veterinary medicine that provides codes, terms, synonyms and definitions that cover anatomy and diseases. It allows a consistent way to index, store, retrieve, and aggregate medical data across specialties and sites of care.

2.5 Search for UCRs by Chromosome Coordinates

UCRs information can be retrieved giving the chromosome number, and the chromosome coordinates for a specific genomic region.

3 Match your sequence vs UCRs

The new BLAST query allows to match a sequence against the entire UCRs sequence database.

3.1 Filter for Pathology

UCbase includes now a tool that gives the opportunity to filter the results for UCRs located in genes or chromosome regions involved in specific pathologies.

4

Type your own query

This tool is dedicated to bioinformaticians who prefer to directly interrogate our database using MySQL language (`www.mysql.org`). The database scheme is shown in our data mining web page. An example of MySQL code and results is shown here below:

```
SELECT UC_NAME,NAME AS DISORDER FROM UC INNER JOIN (SELECT
   ENSEMBL_GENE_ID,NAME,LEVEL FROM RELATED_TO INNER JOIN (SELECT
   ID,NAME,LEVEL FROM PATHOLOGY WHERE NAME="retinoblastoma" OR
   NAME="neuroblastoma") AS B WHERE RELATED_TO.ID = B.ID)AS C WHERE
   UC.ENSEMBL_GENE_ID = C.ENSEMBL_GENE_ID group by uc_name, name
```

The query output (Table 1) shows all the UCRs that are included in the genes involved in retinoblastoma and neuroblastoma.

MySQL query output	
uc.102	neuroblastoma
uc.102	retinoblastoma
uc.103	neuroblastoma
uc.103	retinoblastoma
uc.104	neuroblastoma
uc.104	retinoblastoma

Table 4.1: MySQL output. Only the first six lines are shown.

```
SELECT UC_NAME FROM UC,GENE
WHERE UC.ENSEMBL_GENE_ID = GENE.ENSEMBL_GENE_ID AND
GENE_BIOTYPE="lincRNA"
```

The query above shows how to query the database in order to retrieve all the UCRs that are also long non-coding RNAs (lincRNAs).

5

How to interpret results of standard and custom queries

Here below are listed the legend for each query result the researchers can obtain.

5.1 Search for UCR id results

Uc_name
This is the id of UCRs as published by Bejerano et al. in 2004.

Chr
This is the chromosome in which the UCR is located.

Start
This is the start position of the UCR (hg19).

End
This is the end position of the UCR (hg19).

Start
This is the start position of the UCR referred to Human genome hg18.

End
This is the end position of the UCR referred to Human genome hg18.

Strand
This is the strand in which the UCR is located. UCbase version 2.0 shows only the UCR sequences located in the positive strand (named 1) as default. However genes located in the negative strands are shown and the corresponding strand is named as -1.

Sequence
This is the sequence of the UCR located in the positive strand.

Type
In field type ”e” is exonic, ”n” is non-exonic and ”p” is possibly exonic.

Ensembl_gene_id
This is the Ensembl gene id (`www.ensembl.org`) in which the UCR is located.

Wiki_gene_name_up

Gene name of the gene upstream the UCR sequence.

Wiki_gene_name_down

Gene name of the gene downstream the UCR sequence.

Splicing_Event

This information shows a splicing event when located inside or overlapping a specific UCR. Splicing events can be:

1. inside (when a UCR is inside the exon)
2. includeFeature (when the exon is inside the UCR)
3. overlapStart (when the UCR overlaps the exon start)
4. ovarlapEnd (when the UCR overlaps the exon end)

Exon_name

These are the information about the exon involved in the alternative splicing of the longest isoform of the gene.

5.2 Search for UCRs by Gene results

Gene_biotype

This is the property of the region in which the UCR is located (coding, intergenic, etc).

Gene_start_pos

This is the start position of a specific UCR.

Gene_end_pos

This is the end position of a specific UCR.

Band

This is the chromosome band of a specific UCR.

Strand

This is the strand in which the gene is located named as 1 (positive strand) and -1 (negative strand).

Wikigene_name

This is the Wikigene name (`www.wikigenes.org/`).

G_c_perc

This is GC percentage of the gene.

5.3 Search for UCRs by SNP results

Refsnp_id
This is the SNP id (`www.ncbi.nlm.nih.gov/SNP/`).

Start
This is the start position of a specific SNP.

Allele
This is the allele change of the SNP.

Validation
This is type of validation of the SNP.

Minor_allele_freq
This is the allele frequency in a population as calculated by the 1000genomes project (`http://www.1000genomes.org/`).

Phenotype_desc
This is the phenotype description of the SNP (when available).

Clinical_significance
This is the clinical significance of the SNP (when available)

Strand
This is the strand of the SNP. It is set to 1 by default.

Chr
This is the chromosome of a specific SNP.

Start_uc
This is the start chromosome coordinate of a specific UCR.

Type
This is an information regarding the position of the SNPs from the UCR (within, upstream and downstream).

5.4 Search for UCR by Pathology

Id
This is the pathology SNOMED CT (`http://www.ihtsdo.org/snomed-ct/`) Id.

Name
This is the pathology SNOMED CT (`http://www.ihtsdo.org/snomed-ct/`) name.

Comment
This is the pathology SNOMED CT (`http://www.ihtsdo.org/snomed-ct/`) comment.

Definition
This is the pathology SNOMED CT (`http://www.ihtsdo.org/snomed-ct/`) definition.

Subset
This is the pathology SNOMED CT (`http://www.ihtsdo.org/snomed-ct/`) subset disorder name.

SNOMED
This is the link to SNOMED CT db.

NCI
This is the link to NCI Thesaurus db.

OMIM
This is the link to OMIM db.

5.5 Match your seq. vs UCRs query results (BLAST)

Using this query the researchers can align a specific sequence against all the UCRs sequences contained into UCbase 2.0. The aligning algorithm is BLAST version 2.2.26. BLAST is is an algorithm for comparing primary biological sequence information. Our BLAST search enables a researcher to compare a query sequence with the UCbase library of sequences, and identify library sequences that resemble the query sequence above a certain threshold. Moreover, it is possible to filter

Contacts

CGR - CENTER FOR GENOME RESEARCH - UNIVERSITY OF MODENA AND REGGIO EMILIA
CGR is one of the most important center for Genome Research in Italy. Our work is main focused on Next Generation Sequencing Analysis. Visit CGR website at http://www.cgr.unimore.it

INFORMATION SYSTEMS GROUP - UNIVERSITY OF MODENA AND REGGIO EMILIA
The work of the ISGroup, here at the Computer Engineering Department (DII) of the University of Modena and Reggio Emilia, mainly focuses on the design and development of new systems, algorithms and data structures for the access and management of Information.Visit ISGroup website at http://www.isgroup.unimo.it

Email: ucbase@unimore.it

Last update: April 15, 2014