Biological Annotation in R

Manchester R, 13th Nov, 2013

Nick Burgoyne
Bioinformatician, fiosgenomics
njburgoyne@hotmail.com
nick.burgoyne@fiosgenomics.com

Bioinformatics

- Information management in biology
- Application of statistics to experimental data
- Creation of databases to store it

- Post-genomic era
  - Sequence of a human genome exists
  - Sequence of many different species ex
  - New human genomes can generated in 2 weeks

Typical investigations

- Sequence data
  - What mutations are relevant for a sample?
  - Which differences correlate with a trait?
- Microarray
  - Which genes (when expressed) can be used as a classifier
- Generally
  - What other information is known, what is the structure of the gene in question?

Annotation Databases

- European bioinformatics institute
  - ebi.ac.uk
- National Centre for Biological information
  - ncbi.nlm.nih.gov
- Ensembl
  - ensembl.org
- *Catalogue of Somatic Mutations in Cancer*
  - cancer.sanger.ac.uk
- Mouse Genome Informatics
  - informatics.jax.org
- FlyBase

AnnotationDbi

- Microarrays
  - 10-100k features (genes, parts of genes)
  - Hope for ~10% genes of interest
    - Only care about a few
    - Sets of genes have known function together
  - Mapping them to information
    - Names, Symbols, Homologs
    - Functions, pathways they act in, function

Arrays and AnnotationDbi

- General database interface for annotation data

```
#A definition of species specific data
>require(org.Hs.eg.db)

#A set of data that is relevant to your microarray
>require(hgu95av2.db)

#A vector of probes of relevance
>probes
[1] "31882_at" "38780_at" "37033_s_at" "1702_at" "31610_at”
```

Arrays and AnnotationDbi

```
#Access information from these probes, symbols
>probes <- c("31882_at", "38780_at", "37033_s_at", "1702_at",
"31610_a")

>unlist(mget(probes, hgu95av2SYMBOL))

31882_at 38780_at 37033_s_at 1702_at 31610_at
"RRP9" "AKR1A1" "GPX1" "IL2RA" "PDZK1IP

#Access other aspects
>ls("package:hgu95av2.db")
…[5] "hgu95av2CHR" #The chromosome
…[7] "hgu95av2CHRLOC" #The location on the chromosome
…[15] "hgu95av2GO" #The functions of this probe
```

AnnotationDbi and BioBase etc

- Set of tools built around AnnotationDbi
- Allows for the annotation and analysis of function simply and easily
- Most array types are catered for
- Species specific data also exist (most model species)
- Even if the database doesn't exist your species, but is present in the ncbi repositories

```
>library(AnnotationForge)
```

What if I'm interested in other data?

- Other questions, such as the:
  - Is my mutant associated with Cancer?
  - Does it correspond to a known disorder?
  - What is the sequence of this region of the DNA?

- Hope that the data source you are using is on the biomart network!
  - http://www.biomart.org/
  - 46 databases, common interface, defined structure
  - Bioconductor: biomaRt

biomaRt

```
>library(biomaRt)

#What databases are available?
>head(listMarts())
biomart version
1 ensembl ENSEMBL 52 GENES (SANGER UK)
2 snp ENSEMBL 52 VARIATION (SANGER UK)
3 vega VEGA 33 (SANGER UK)
4 msd MSD PROTOTYPE (EBI UK)
5 uniprot UNIPROT PROTOTYPE (EBI UK)
6 htgt HIGH THROUGHPUT GENE TARGETING AND TRAPPING (SANGER UK)

#Choose a database
>mart = useMart("CosmicMart")
```

biomaRt

```
#What datasets are avaliable for each mart?
>head(listDatasets(mart)

 dataset description version
1 COSMIC65    COSMIC65
2 COSMIC66    COSMIC66
3 COSMIC64    COSMIC64

cos65 = useDataset("COSMIC65", mart = mart)
```

biomaRt

```
#How to filter the data
> head(listFilters(cos65))

                  name    description
1          id_sample       Sample ID
2        sample_name     Sample Name
3     sample_source  Sample Source
4 samp_gene_mutated Mutated Sample
5     tumour_source  Tumour Source
6           id_gene         Gene ID
```

biomaRt

```
#List the data that is available
head(listAttributes(cos65))

            name      description
1     id_sample COSMIC Sample ID
2   sample_name      Sample Name
3 sample_source    Sample Source
4 tumour_source    Tumour Source
5       id_gene          Gene ID
6     gene_name        Gene Name
```

biomaRt

```
#A set of gene names
genes <- c("KRAS", "BRAF")

#Get a list of cancers where these are mutated
mutations <- getBM(attributes = c("id_gene", "gene_name",
+ "site_primary"), filters = "gene_name", value = genes, mart =
+ cos65)

> head(mutations)
  id_gene gene_name    site_primary
1       4      KRAS     endometrium
2       4      KRAS        pancreas
3       4      KRAS   biliary_tract
4       2      BRAF large_intestine
5       2      BRAF         thyroid
6       4      KRAS large_intestine
```

biomaRt

```
genes <- "KRAS"
sites <- "prostate"

#Mutation data associated with KRAS and prostate cancer
mutations <- getBM(attributes = c("gene_name", "site_primary", "cds_mut_syntax",
+ "aa_mut_syntax", "tumour_source"), filters = c("gene_name", "site_primary"), value =
+ list(genes, sites), mart = cos66)
```

	gene_name	site_primary	cds_mut_syntax	aa_mut_syntax	tumour_source
1	KRAS	prostate	c.182A>G	p.Q61R	primary
2	KRAS	prostate	c.34G>T	p.G12C	NS
3	KRAS	prostate	c.38G>A	p.G13D	NS
4	KRAS	prostate	c.34G>A	p.G12S	NS
5	KRAS	prostate	c.181C>A	p.Q61K	NS
6	KRAS	prostate	c.37G>A	p.G13S	NS
7	KRAS	prostate	c.35G>A	p.G12D	NS
8	KRAS	prostate	c.35G>A	p.G12D	primary
9	KRAS	prostate	c.35G>T	p.G12V	NS
10	KRAS	prostate	c.35G>T	p.G12V	primary
11	KRAS	prostate	c.35G>T	p.G12V	metastasis

biomaRt

- Always request the filter value as an attribute
  - Not all values you search for exist
  - Results will be in random orders
  - The search parameters are totally excluded
- Don't expect the most up-to date answers
  - Often different (older) data than available directly
  - COSMIC is actually very good
- The connection can be very flaky
  - Only partial data sets may be returned