

# **THE LOST IDENTIFIERS, NAMES, AND SYMBOLS**

## **MANAGING NAMES, IDENTIFIERS AND THE LIKE IN BIOINFORMATICS**

Michael Dondrup

MCB Research Course: Integrative Bioinformatics

November 24, 2009

# Outline

1 Introduction

2 Mapping Identifiers

3 Exercises

# Definition Identifier

An identifier is a **unique expression** in a written format either by a code, by numbers or by the combination of both to distinguish variations from one to another among a class of substances, items, or objects...

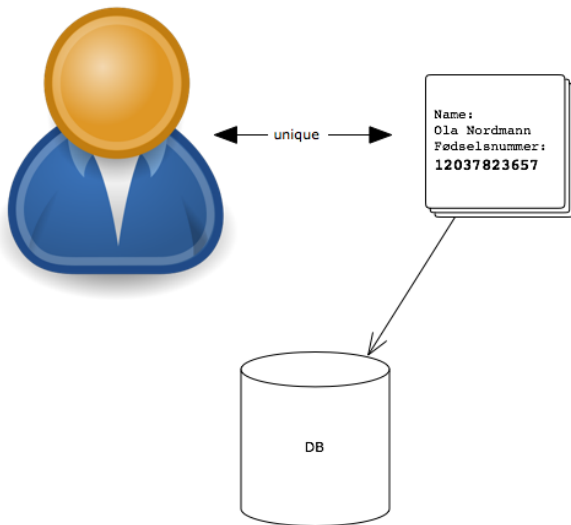
In computer science, Identifiers (IDs) are **lexical tokens** that name **entities**. The concept is analogous to that of a "name".

– Wikipedia

→ "a unique name for something".

# Examples

- Norwegian social security number



# Examples

- Norwegian social security number
- URI: `https://www.uib.no/rs/mcb`
- Digital Object Identifiers (DOI):  
`doi:10.1186/1752-0509-3-82`
- IUPAC symbols: ACGT
- GenBank accession number: AK090412
- Ensembl Gene/Transcript/Protein ID: ENSG00000241154  
ENST00000480017 ENSP00000420800
- Affymetrix probe id: 50780\_at

# Examples

- Norwegian social security number
- URI: `https://www.uib.no/rs/mcb`
- Digital Object Identifiers (DOI):  
`doi:10.1186/1752-0509-3-82`
- IUPAC symbols: `ACGT`
- GenBank accession number: `AK090412`
- Ensembl Gene/Transcript/Protein ID: `ENSG00000241154`  
`ENST00000480017` `ENSP00000420800`
- Affymetrix probe id: `50780_at`

# Examples

- Norwegian social security number
- URI: `https://www.uib.no/rs/mcb`
- Digital Object Identifiers (DOI):  
`doi:10.1186/1752-0509-3-82`
- IUPAC symbols: ACGT

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

# Examples

- Norwegian social security number
- URI: `https://www.uib.no/rs/mcb`
- Digital Object Identifiers (DOI):  
`doi:10.1186/1752-0509-3-82`
- IUPAC symbols: ACGT
- GenBank accession number: AK090412
- Ensembl Gene/Transcript/Protein ID: ENSG00000241154  
ENST00000480017 ENSP00000420800
- Affymetrix probe id: 50780\_at

# Examples

- Norwegian social security number
- URI: `https://www.uib.no/rs/mcb`
- Digital Object Identifiers (DOI):  
`doi:10.1186/1752-0509-3-82`
- IUPAC symbols: `ACGT`
- GenBank accession number: `AK090412`
- Ensembl Gene/Transcript/Protein ID: `ENSG00000241154`  
`ENST00000480017` `ENSP00000420800`
- Affymetrix probe id: `50780_at`

# Examples

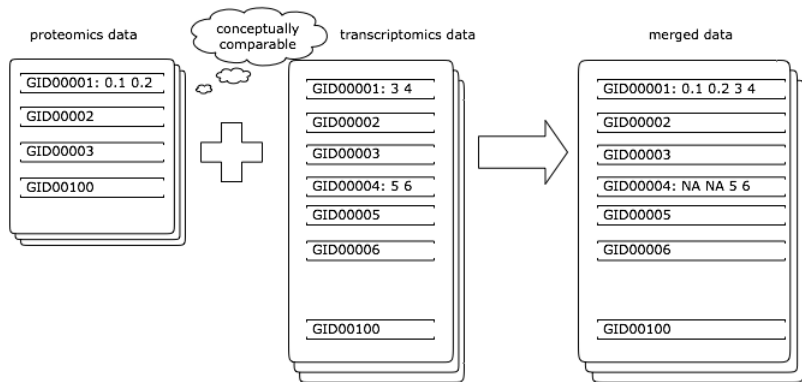
- Norwegian social security number
- URI: `https://www.uib.no/rs/mcb`
- Digital Object Identifiers (DOI):  
`doi:10.1186/1752-0509-3-82`
- IUPAC symbols: `ACGT`
- GenBank accession number: `AK090412`
- Ensembl Gene/Transcript/Protein ID: `ENSG00000241154`  
`ENST00000480017` `ENSP00000420800`
- Affymetrix probe id: `50780_at`

# Remarks

- Names, Symbols are not necessarily unique
- IDs are domain/database specific
- The same entity can have many different IDs
- The same token can be used for different things
- Closely related things (e.g.. Gene/Protein) have different IDs
- Cross-references exist in most databases
- IDs and Names can change over time (e.g. genome re-annotations, transcript re-mapping)

# Why bother about mapping identifiers?

## Merging data: the essence of integrative bioinformatics



# Remarks

- every mapping is a (possibly coarse) abstraction
- Gene  $\neq$  Transcript  $\neq$  Protein  $\neq$  Enzyme
- Gene  $\approx$  Transcript  $\approx$  Protein  $\approx$  Enzyme
- the mapping is most likely not "1 to 1"
- transcript/protein isoforms, protein/enzyme complexes
- let alone: transcription factor (binding sites), SNPs, RNAi, metabolites, motives

# Tools: BioMart

- <http://www.biomart.org>
- Data Warehouse of genome annotation data
- A query-oriented database for data-mining
- Most complete and up-to-date

# Tools: BioMart

[HOME](#)[MARTVIEW](#)[MARTSERVICE](#)[DOCS](#)[CONTACT](#)[NEWS](#)[CREDITS](#)[New](#) [Count](#) [Results](#)[URL](#) [XML](#) [Perl](#) [Help](#)

Dataset 1 / 49506 Genes  
Homo sapiens genes (GRCh37)

### Filters

Gene type : protein\_coding  
HGNC symbol: [ID-list specified]

### Attributes

Ensembl Gene ID  
Ensembl Transcript ID  
Associated Gene Name  
Gene Start (bp)  
Gene End (bp)  
Ensembl Protein ID

Export all results to

File

TSV

 Unique results only

Email notification to

View

10

rows as HTML

 Unique results only

Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name	Gene Start (bp)	Gene End (bp)	Ensembl Protein ID
ENSG00000124788	ENST00000379065	ATXN1	16299343	16761722	ENSP00000368355
ENSG00000124788	ENST00000436367	ATXN1	16299343	16761722	ENSP00000416360
ENSG00000124788	ENST00000450222	ATXN1	16299343	16761722	ENSP00000397260
ENSG00000124788	ENST00000405101	ATXN1	16299343	16761722	ENSP00000384776
ENSG00000124788	ENST00000483954	ATXN1	16299343	16761722	
ENSG00000124788	ENST00000473388	ATXN1	16299343	16761722	
ENSG00000124788	ENST00000479680	ATXN1	16299343	16761722	
ENSG00000124788	ENST00000495178	ATXN1	16299343	16761722	
ENSG00000124788	ENST00000496374	ATXN1	16299343	16761722	
ENSG00000124788	ENST00000467008	ATXN1	16299343	16761722	

### Dataset

[None Selected]

biomart version 0.7

# Tools: David

- <http://david.abcc.ncifcrf.gov>
- Database for Annotation, Visualisation and Integrated Data Mining
- Many useful tools including Gene ID converter, Pathway tools

## Analysis Wizard

[Tell us how you like the tool](#)  
[Contact us for questions](#)

← Step 1. Submit your gene list through left panel.

**new!**Note: Affy Exon IDs and Affy Gene Array IDs are now supported in DAVID, as "affy\_id" type.

An example:

Copy/paste IDs to "box A" -> Select Identifier as "Affy\_ID" -> List Type as "Gene List" -> Click "Submit" button

```
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
```

Upload **List** Background

### Upload Gene List

[Demolist 1](#) [Demolist 2](#)  
[Upload Help](#)

Step 1: Enter Gene List  
A: Paste a list

Or  
B: Choose From a File

Step 2: Select Identifier

Step 3: List Type  
Gene List   
Background

Step 4: Submit List

# Tools: KEGG


- <http://www.genome.jp/>
- KEGG has its own identifiers and search
- Useful in connection with enzymes
- Find the EC number of a gene more
- Not so good for batch conversion

# Tools: KEGG



Homo sapiens (human): 6310

Help

<b>Entry</b>	6310	CDS	H.sapiens
<b>Gene name</b>	ATXN1		
<b>Definition</b>	ataxin 1		
<b>Class</b>	<a href="#">BRITE hierarchy</a>		
<b>SSDB</b>	<a href="#">Ortholog</a> <a href="#">Paralog</a> <a href="#">Gene cluster</a> <a href="#">GFIT</a>		
<b>Motif</b>	Pfam: <a href="#">SSDP AXH</a> PROSITE: <a href="#">GLN_RICH AXH</a> <a href="#">Motif</a>		
<b>Other DBs</b>	NCBI-GI: <a href="#">51479158</a> NCBI-GeneID: <a href="#">6310</a> HGNC: <a href="#">10548</a> HPRD: <a href="#">03333</a> Ensembl: <a href="#">ENSG00000124788</a> OMIM: <a href="#">601556</a> UniProt: <a href="#">P54253</a> <a href="#">Q96FF1</a>		
<b>Structure</b>	PDB: <a href="#">1OAB</a> <a href="#">Thumbnails</a>  <a href="#">Jmol</a>		
<b>Position</b>	6p23		
<b>AA seq</b>	815 aa <a href="#">AA seq</a> <a href="#">DB search</a> MKSNQERSNECLPPKKREIPATSRSSSEKAPTLPSDNHRVEGTAWLPGNPGRGHGGRRH GPAGTSVELGLQQGIGLHKALSTGLDYSPPSAPRSVPVATTLPAAYATPQPGTPVSPVQY AHLPHTFQFIGSSQYSGTYASFIPSQLIPPTANPVTSAVASAAGATTPSQRSQLEAYSTL LANMGSLSQTPGHKAEQQQQQQQQQQQQHQHQHQQQQQQQQQQQQQHLSRAPGLITPGSP PAQQNQYVHISSSPQNTGRTASPPAIPVHLHPHQMTMIPHTLLTGPPSQVVMQYADSGSHF		

## All links

- Disease (3)
  - KEGG DISEASE (1)
  - OMIM (2)
- Genome (1)
  - KEGG GENOME (1)
- Gene (13)
  - NCBI-Gene (1)
  - NCBI-GI (8)
  - UNIGENE (1)
  - HGNC (1)
  - HPRD (1)
  - ENSEMBL-HSA (1)
- Protein sequence (9)
  - UniProt (2)
  - RefSeq(pep) (2)
  - IPI (4)
- DNA sequence (18)
  - RefSeq(nuc) (2)
  - GenBank (8)
  - EMBL (8)
- 3D Structure (1)
  - PDB (1)
- Protein domain (4)
  - Pfam (2)
  - PROSITE (2)
- All databases (48)

# Tools: MSigDB

- `http://www.broadinstitute.org/gsea/msigdb/index.jsp`
- A place for finding gene sets

- ▶ MSigDB Home
- ▶ About Collections
- ▶ Browse Gene Sets
- ▶ Search Gene Sets
- ▶ Annotate Gene Sets
- ▶ View Gene Families
- ▶ Help



## MSigDB

Molecular Signatures  
Database

### Overview

The Molecular Signatures Database (MSigDB) is a collection of gene sets for use with GSEA software. From this web site, you can

- ▶ **Search** for gene sets
- ▶ **Browse** gene sets
- ▶ **View annotations** by clicking a gene set name to display its gene set page; for example, [AKTPATHWAY](#)
- ▶ **Download** gene sets
- ▶ **Compute overlaps** between your gene set and other gene sets in MSigDB
- ▶ **Categorize** members of a gene set by gene families
- ▶ **Build an expression signature** of the gene set using a compendium of expression profiles

### Registration

Please [register](#) to download the GSEA software and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.

### Current Version

GSEA/MSigDB web site v2.0 released December 14 2007  
MSigDB database v2.5 updated April 7 2008, [Release notes](#).

### Contributors

## Molecular Signatures Database

### Collections

The MSigDB gene sets are divided into five major collections:

**c1** **positional gene sets** for each human chromosome and each cytogenetic band.

**c2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

**c3** **motif gene sets** based on conserved *cis*-regulatory motifs from a comparative analysis of the human, mouse, rat and dog genomes.

**c4** **computational gene sets** defined by expression neighborhoods centered on 380 cancer-associated genes.

**c5** **GO gene sets** consist of genes annotated by the same GO terms.

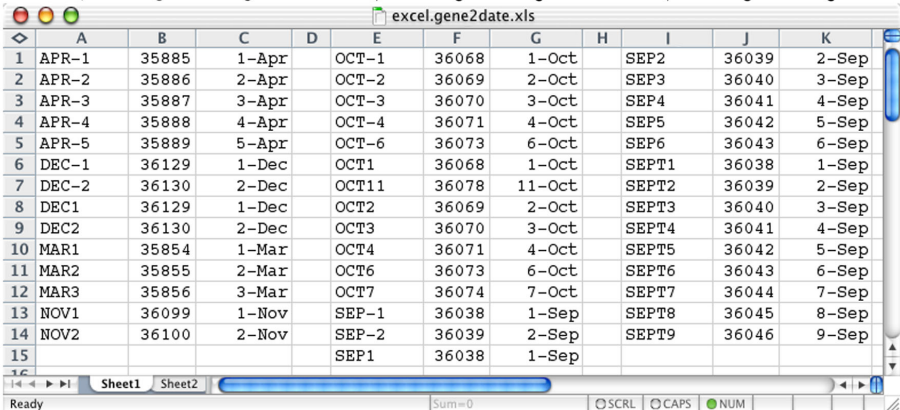
### Citing the MSigDB

# Why is Excel not a bioinformatics tool?

- Zeeberg *et al.*, Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics , BMC Bioinformatics, (2004)
- Excel can convert gene names to dates

# Why is Excel not a bioinformatics tool?

gene names      internal date format      default date format      gene names      internal date format      default date format      gene names      internal date format      default date format



	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

# Why is Excel not a bioinformatics tool?

The screenshot shows the NCBI LocusLink report for the NEDD5 gene. The page includes a search bar, navigation links, and detailed gene information. A red arrow points from the top of the page down to the '2-Sep' link in the Mouse Homology Maps table.

**NCBI LocusLink Report**

Search: LocusLink | Display: Brief | Organism: All

Query:

View: Hs NEDD5 | One of 1 Loci | Save All Loci

Click to Display: tRNA-Genomic Alignments (spanning 38716 bps)

**Homo sapiens Official Gene Symbol and Name (HGNC)**

**NEDD5: neural precursor cell expressed, developmentally down-regulated 5**

**LocusID: 4735**

**Overview** [Submit GeneRIF](#) ?

**Locus Type:** gene with protein product, function known or inferred

**Product:** neural precursor cell expressed, developmentally down-regulated 5

**Alternate Symbols:** DIFF6, SEPT2, hNedd5, KIAA0158

**Relationships** ?

**Mouse Homology Maps:**

Map	Map	Map	Map
NCBI vs. MGD	1 cM	<a href="#">2-Sep</a>	Hs Mm
UCSC vs. MGD	1 cM	<a href="#">Septu</a>	Hs Mm
UCSC vs. Hudson et al.	1 1319.34 cR	<a href="#">AW208991</a>	Hs Mm

**Map Information** ?

**Chromosome:** 2 **mv**

**Cytogenetic:** 2q37 **RefSeq**

**Markers:**

Chr.	Marker	Marker	Marker
Chr. 2:	<a href="#">D2S2576</a>	D2S2576	
Chr. 2	<a href="#">G19712</a>		mv
Chr. 2	<a href="#">A001W20</a>		mv
Chr. 2	<a href="#">D2S2850</a>	D2S2850	mv
Chr. 2	<a href="#">D2S2704</a>	D2S2704	mv
Chr. 2	<a href="#">G62117</a>		mv
Chr. 2	<a href="#">D15S049</a>	D15S049	

# Exercises

- 1 Find information about ATXN3 (gene symbol) using all bioinformatics tools presented.
- 2 Use Biomart to generate an identifier mapping table for the human genome and download it as a tab separated file.
- 3 Load the list into Excel without messing with the identifiers