




NCBI PowerScripting

Lecture 2:
Introduction to the eUtils



April 20-22, 2005






Outline

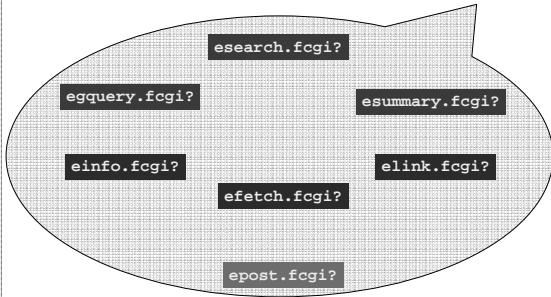
- The Basics: Syntax of eUtil URLs
- The Entrez Core
 - EGQuery, ESearch, ESummary
- Entrez Databases
 - EInfo, EFetch, ELink
- The Entrez History Server
 - EPost







The Base URL

`http://eutils.ncbi.nlm.nih.gov/entrez/eutils/eutil.fcgi?`





URL Parameters

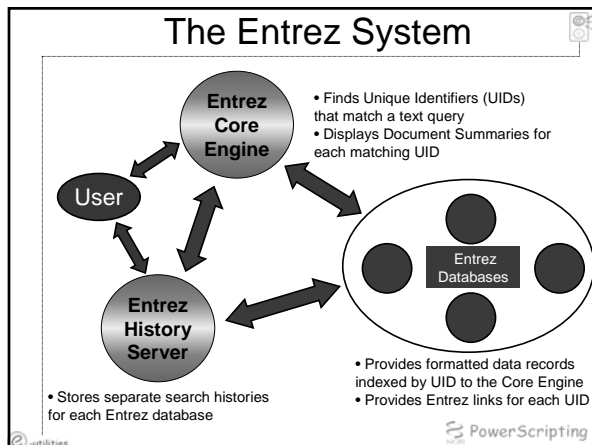
`BASE/ esearch.fcgi? db=nucleotide&term=mouse[orgn]`

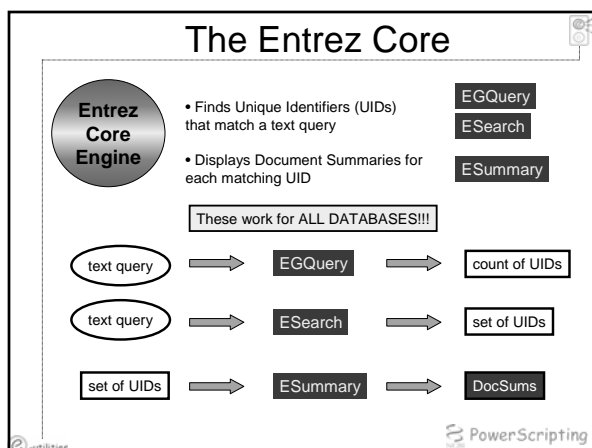
Parameters are separated by & symbols

`db = nucleotide`
`term = mouse[orgn]`

We need to know the following:

1. What parameters are available
2. What values they accept





EGQuery

Performs a global Entrez search across all databases

Why use it? To find the number of records matching a text query

INPUT term Entrez text query

BASE/ egquery.fcgi? term=mouse[orgn]

OUTPUT XML Number of records matching the query in each database

© utilities PowerScripting

EGQuery Output

```

<Result>
  <Term>mouse[orgn]</Term>
  <EGQueryResult>
    <ResultItem>
      <DbName>pubmed</DbName>
      <ResultName>PubMed</ResultName>
      <Count>0</Count>
      <Status>Term or Database is not found</Status>
    </ResultItem>
    <ResultItem>
      <DbName>pub</DbName>
      <ResultName>PBC</ResultName>
      <Count>0</Count>
      <Status>OK</Status>
    </ResultItem>
    <ResultItem>
      <DbName>journal</DbName>
      <ResultName>Journal</ResultName>
      <Count>0</Count>
      <Status>Term or Database is not found</Status>
    </ResultItem>
    <ResultItem>
      <DbName>med</DbName>
      <ResultName>Med</ResultName>
      <Count>0</Count>
      <Status>Term or Database is not found</Status>
    </ResultItem>
    <ResultItem>
      <DbName>nucleotide</DbName>
      <ResultName>Nucleotide</ResultName>
      <Count>103147</Count>
      <Status>OK</Status>
    </ResultItem>
    <ResultItem>
      <DbName>protein</DbName>
      <ResultName>Protein</ResultName>
      <Count>11974</Count>
      <Status>OK</Status>
    </ResultItem>
    <ResultItem>
      <DbName>genome</DbName>
      <ResultName>Genome</ResultName>
      <Count>1</Count>
      <Status>OK</Status>
    </ResultItem>
    <ResultItem>
      <DbName>structure</DbName>
      <ResultName>Structure</ResultName>
      <Count>1144</Count>
      <Status>OK</Status>
    </ResultItem>
  </EGQueryResult>
</Result>
  
```

© utilities PowerScripting

ESearch

Performs an Entrez search in one specified database

Why use it? To find UIDs that match a text query

INPUT db Entrez database to search

term Entrez text query

BASE/ esearch.fcgi? db=nucleotide&term=mouse[orgn]

OUTPUT XML

- Total number of records matching the query
- Partial list of matching UIDs
- Term translations

© utilities PowerScripting

ESearch Output – UUIDs

```

<eSearchResult>
  <Count>6305267</Count>
  <RetMax>20</RetMax>
  <RetStart>0</RetStart>
  <IdList>
    <Id>49619226</Id>
    <Id>49615287</Id>
    <Id>49615286</Id>
    <Id>49615285</Id>
    <Id>49615204</Id>
    <Id>49615283</Id>
    <Id>49615282</Id>
    <Id>49615201</Id>
    <Id>49615280</Id>
    <Id>49615279</Id>
    <Id>49615278</Id>
    <Id>49615277</Id>
    <Id>49615276</Id>
    <Id>49615275</Id>
    <Id>49615274</Id>
    <Id>49615273</Id>
    <Id>49615272</Id>
    <Id>49615271</Id>
    <Id>49615270</Id>
    <Id>49615269</Id>
  </IdList>

```

Total number of records found

&retmax

&retstart

first record = &retstart

Matching UUIDs

quantity = &retmax

PowerScripting

Retrieval Parameters

These work for ESearch, ESummary, and EFetch

retstart First record to retrieve from UUID set (default = 0)

retmax Number of records to retrieve from UUID set

&retmax=4

(84, 23, 19, 55, 20, 96, 73) → (84, 23, 19, 55)

&retstart=2&retmax=4

(84, 23, 19, 55, 20, 96, 73) → (19, 55, 20, 96)

PowerScripting

ESearch Output – Translations

Term translations: equivalent to the Details link on the web

```

<TranslationSet>
  <Translation>
    <From>mouse[orgn]
    <From>mouse%5Borgn%5D</From>
    <To>%22Mus musculus%22%5BOrganism%5D</To>
  </Translation>
  <TranslationSet>
    <TranslationStack>
      <TermSet>
        <Term>"Mus musculus"[Organism]</Term>
        <Field>Organism</Field>
        <Count>6305267</Count>
        <Explode>Y</Explode>
      </TermSet>
    </TranslationStack>
  </TranslationSet>
</eSearchResult>

```

PowerScripting

Look at the Time!

Date format: YYYY/MM/DD

reldate=n Retrieve UIDs with dates less than n days prior to today

mindate Retrieves UIDs in this date range
maxdate These parameters must be used together!

datatype Type of date to examine:

Most Entrez databases only support two of these date types!

mdat Modification date: date the record was last touched
pdat Publication date: date the record was made public
edat Entrez date: date the record entered Entrez
cdat Creation date: date the record was created

PowerScripting

ESummary

Retrieves Document Summaries matching a set of UIDs

Why use it?

- To download data
- It's fast and can download large sets with one URL
- If EFetch does not support your database

INPUT

db Entrez database to search

id Set of UIDs

BASE/ **esummary.fcgi?** **db=nucleotide&id=49619226,49615287**

OUTPUT **XML** DocSums, often with more data than web Entrez provides

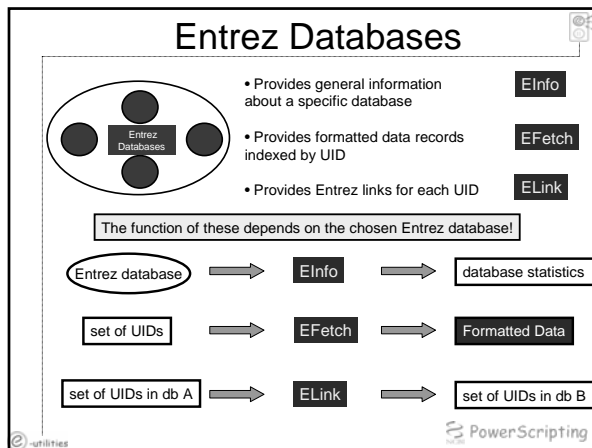
PowerScripting

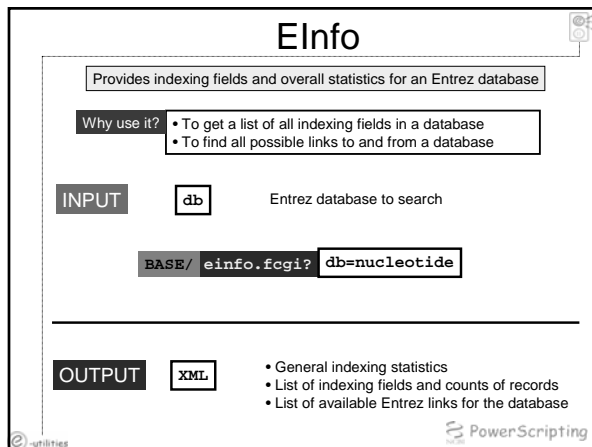
ESummary Output

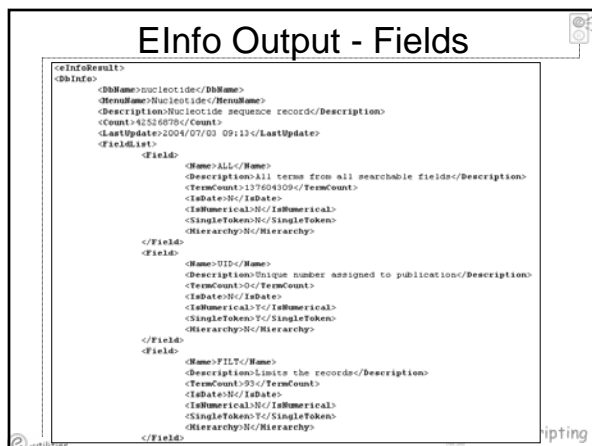
```
<eSummaryResult>
<DocSum>
  <Id>49619226</Id>
  <Item Name="Caption" Type="String">NM_008496</Item>
  <Item Name="Title" Type="String">Mus musculus lectin, galactose binding,
  <Item Name="Extra" Type="String">gi|49619226|ref|NM_008496.4|</Item>
  <Item Name="GI" Type="Integer">49619226</Item>
  <Item Name="CreateDate" Type="String">2000/01/04</Item>
  <Item Name="UpdateDate" Type="String">2004/07/02</Item>
  <Item Name="Flags" Type="Integer">0</Item>
  <Item Name="TaxId" Type="Integer">10090</Item>
</DocSum>
</eSummaryResult>
```

1: [NM_008496](#)
Mus musculus lectin, galactose binding, soluble 7 (Lgals7), mRNA
gi|49619226|ref|NM_008496.4|[49619226]

PowerScripting







EInfo Output - Links

```

<LinkList>
  <Link>
    <Name>nucleotide_comp_genome</Name>
    <Menu>Components to Genome</Menu>
    <Description>Genome(s) using this record as component</Description>
    <DbTo>genome</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_comp_nucleotide</Name>
    <Menu>Assembly</Menu>
    <Description>Link to master record</Description>
    <DbTo>nucleotide</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_gene</Name>
    <Menu>Gene Links</Menu>
    <Description>Link to related Genes</Description>
    <DbTo>gene</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_genome</Name>
    <Menu>Assembly to Genome</Menu>
    <Description>Genome record containing nucleotide sequence</Description>
    <DbTo>genome</DbTo>
  </Link>
  <Link>
    <Name>nucleotide_geo</Name>
    <Menu>GEO Profile Links</Menu>
    <Description>GEO records associated with nucleotide records</Description>
    <DbTo>geo</DbTo>
  </Link>
</LinkList>

```

© utilities PowerScripting

EFetch

Retrieves formatted data records matching a set of UIDs

Why use it? To download data records

INPUT

db Entrez database to search

id Set of UIDs

BASE/ `efetch.fcgi?` `db=nucleotide&id=49619226,49615287`

OUTPUT **Varied** Formatted data records

© utilities PowerScripting

Databases that Support EFetch

Literature	Sequences	Other
PubMed	Nucleotide	Gene
Journals	Protein	Taxonomy
PubMed Central	Genome	
	Popset	
	SNP	

© utilities PowerScripting

EFetch Formatting Parameters

rettype

Determines the type of data record returned (flat file, FASTA, EST, accession, etc.)

retmode

Determines the format (mode) of data record returned (text, HTML, XML)

Be warned!

- These settings are very dependent on the database
- These settings interact with one another
- Not all possible combinations are supported

utilities

PowerScripting

EFetch for PubMed

retmode

default = HTML text XML asn1

rettype

default = ASN.1 medline citation abstract uilist

To retrieve text abstracts:

&retmode=text&rettype=abstract

To retrieve HTML abstracts:

&retmode=html&rettype=abstract

To retrieve text medline:

&retmode=text&rettype=medline

To retrieve full XML:

&retmode=xml&rettype=

utilities

PowerScripting

EFetch for Sequences

retmode

default = text HTML XML

rettype

default = native gb gp gbwithparts fasta gss est acc

To retrieve text GenBank flat file:

&retmode=text&rettype=gb

To retrieve HTML GenPept flat file:

&retmode=html&rettype=gp

To retrieve text FASTA:

&retmode=text&rettype=fasta

To retrieve TinySeqXML:

&retmode=xml&rettype=fasta

To retrieve full XML:

&retmode=xml&rettype=native

utilities

PowerScripting

Retrieving Batches

A single EFetch URL can download a maximum of 500 records!

&retstart=0&retmax=500

&retstart=500&retmax=500

&retstart=1000&retmax=500

...

Retrieve successive batches of 500 records until the entire dataset is downloaded

Batch retrieval is easily implemented using a **for** loop:

```
for ($retstart=0; $retstart < $total; $retstart+=$retmax) {
  ... &retstart=$retstart&retmax=$retmax...
}
```

© utilities PowerScripting

ELink

Retrieves UUIDs in database B linked to a set of UUIDs in database A

Why use it?

- To find related data in another database
- To find neighbors within a database

INPUT

dbfrom	Entrez database to link <i>from</i>
db	Entrez database(s) to link <i>to</i> ; can be a list!
id	List of UUIDs
cmd	ELink command mode (default = neighbor)

BASE/ `elink.fcgi?` **dbfrom=nucleotide&db=protein&id=49619226**

OUTPUT **XML** Set(s) of linked UUIDs

© utilities PowerScripting

Computational Neighbors in ELink

Retrieves UUIDs linked to other UUIDs in the same database

dbfrom = **db**

BASE/ `elink.fcgi?` **dbfrom=protein&db=protein&id=15718680**

term Entrez query that ELink uses to limit the set of neighbors

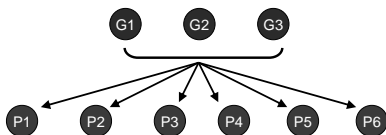
Supported databases:

pubmed	cdd
nucleotide	geo
protein	gds
domains	

© utilities PowerScripting

Passing One UID Set to ELink

dbfrom=gene&db=protein&id=G1,G2,G3

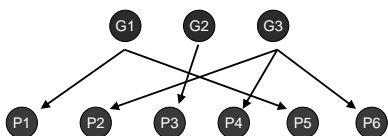


2-utilities

PowerScripting

Passing Multiple UID Sets to ELink

dbfrom=gene&db=protein&id=G1&id=G2&id=G3

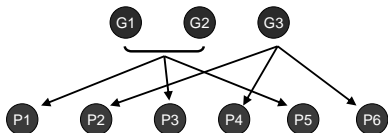


2-utilities

PowerScripting

Passing Multiple UID Sets to ELink

dbfrom=gene&db=protein&id=G1,G2&id=G3



2-utilities

PowerScripting

ELink Output

dbfrom=gene&db=protein&id=414758,414749,408191

```

<?xml version="1.0">
<eLinkResult>
  <LinkSet>
    <DbFrom>gene</DbFrom>
    <IdList>
      <Id>414758</Id>
      <Id>414749</Id>
      <Id>408191</Id>
    </IdList>
    <LinkSetDb>
      <DbTo>protein</DbTo>
      <LinkName>gene_protein</LinkName>
      <Link>
        <Id>40958541</Id>
      </Link>
      <Link>
        <Id>47169612</Id>
      </Link>
      <Link>
        <Id>28972145</Id>
      </Link>
      <Link>
        <Id>26337735</Id>
      </Link>
      <Link>
        <Id>26325192</Id>
      </Link>
    </LinkSetDb>
  </LinkSet>
</eLinkResult>
  
```

PowerScripting

ELink Output

dbfrom=gene&db=protein&id=414758&id=414749&id=408191

```

<?xml version="1.0">
<eLinkResult>
  <LinkSet>
    <DbFrom>gene</DbFrom>
    <IdList>
      <Id>414758</Id>
    </IdList>
    <LinkSetDb>
      <DbTo>protein</DbTo>
      <LinkName>gene_protein</LinkName>
      <Link>
        <Id>40958541</Id>
      </Link>
      <Link>
        <Id>26325192</Id>
      </Link>
      <Link>
        <Id>26337735</Id>
      </Link>
    </LinkSetDb>
  </LinkSet>
  <LinkSet>
    <DbFrom>gene</DbFrom>
    <IdList>
      <Id>414749</Id>
    </IdList>
    <LinkSetDb>
      <DbTo>protein</DbTo>
      <LinkName>gene_protein</LinkName>
      <Link>
        <Id>28972145</Id>
      </Link>
    </LinkSetDb>
  </LinkSet>
</eLinkResult>
  
```

```

<?xml version="1.0">
<eLinkResult>
  <LinkSet>
    <DbFrom>gene</DbFrom>
    <IdList>
      <Id>408191</Id>
    </IdList>
    <LinkSetDb>
      <DbTo>protein</DbTo>
      <LinkName>gene_protein</LinkName>
      <Link>
        <Id>47169612</Id>
      </Link>
    </LinkSetDb>
  </LinkSet>
</eLinkResult>
  
```

PowerScripting

The Entrez History Server

Stores separate search histories for each Entrez database

EPost

ESearch

ELink

The History Server represents the location of stored UID sets with two parameters:

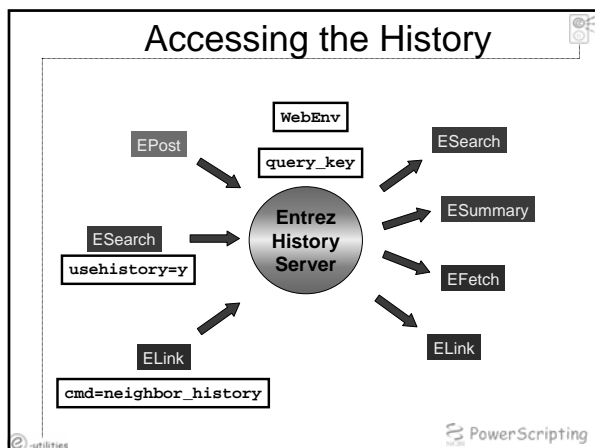
WebEnv

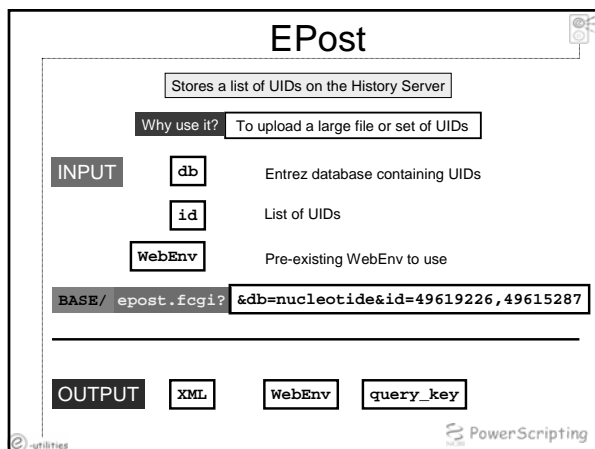
A string specifying a cookie assigned by the History Server

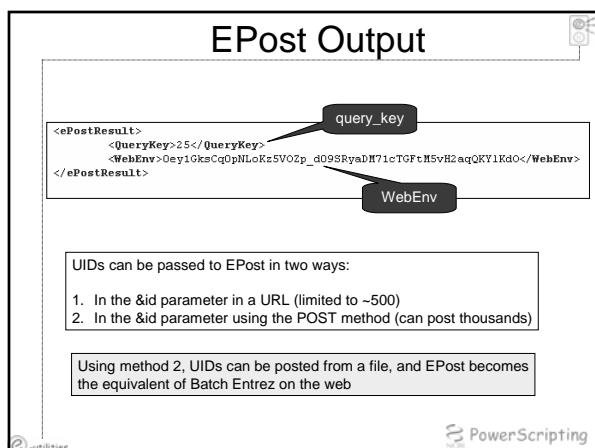
query_key

An integer equivalent to the History number on the web

PowerScripting







Performs an Entrez search in one specified database

INPUT

db

Entrez database to search

term

Entrez text query

usehistory

Flag (=y) for ESearch to access History

WebEnv

Pre-existing WebEnv to use

```
BASE/ esearch.fcgi?
```

db=nucleotide&term=mouse[orgn]&usehistory=y&WebEnv=A1C3

OUTPUT

XML

- Total number of records matching the query
- Partial list of matching UIDs
- Term translations

 PowerScripting

BASE/ [esearch.fcgi?](#)

db=nucleotide&term=mouse[orgn]&usehistory=y

```
<eSearchResult>
  <Count>6205267</Count>
  <RetMax>20</RetMax>
  <RetStart>0</RetStart>
  <QueryKey>24</QueryKey>
  <WebEnv>DyGfHNCt4n9f5Sm_WRY6R8U646NWQfay3bIdHt1cnZfH_pEzGo</WebEnv>
  <IdList>
    <Id>49615286</Id>
    <Id>49615287</Id>
    <Id>49615286</Id>
    <Id>49615285</Id>
    <Id>49615284</Id>
    <Id>49615203</Id>
    <Id>49615282</Id>
    <Id>49615201</Id>
    <Id>49615280</Id>
    <Id>49615279</Id>
    <Id>49615270</Id>
    <Id>49615277</Id>
    <Id>49615276</Id>
  </IdList>

```

query_key

WebEnv

 PowerScripting

```
cmd=neighbor_history&dbfrom=gene
&db=protein,pubmed&id=414758&id=414749
```

```
<linkHref!>
</linkSet>
<BdfFrom>gene/BdfFrom
</BdfFrom>
<Id>414750</Id>
</IdSet>
<linkSetBibliatory>
<BdfTo>protein/BdfTo
</BdfTo>
<linkName>gene.protein.LinkName
</linkName>
<linkSetBibliatory>
<BdfTo>pubmed/BdfTo
</BdfTo>
<linkName>gene.pubmed.LinkName
</linkName>
<WebEnv>GFh3z2qy50z2 coneHn EC9v10w9Tqy-3kTqyhe5tQ7q550 Q2w7110FRA4641-Y/Website
</WebEnv>
</linkSet>
<BdfFrom>gene/BdfFrom
</BdfFrom>
<Id>414749</Id>
</IdSet>
<linkSetBibliatory>
<BdfTo>protein/BdfTo
</BdfTo>
<linkName>gene.protein.LinkName
</linkName>
<linkSetBibliatory>
<BdfTo>pubmed/BdfTo
</BdfTo>
<linkName>gene.pubmed.LinkName
</linkName>
<WebEnv>GFh3z2qy50z2 coneHn EC9v10w9Tqy-3kTqyhe5tQ7q550 Q2w7110FRA4641-Y/Website
</WebEnv>
</linkSet>
</linkHref!>
```

Unique query_keys


Common WebEnv

© 2007 The Authors
Journal compilation © 2007 Blackwell Publishing Ltd

Scripting

Finally, Now for *Your* UID!

Please use both of these parameters in your URLs in case there are problems



tool

a unique name for your software package

email

your email address, so we can contact you...

&tool=mr.gene&email=funwithgenes@big.genomics.com

©-utilities PowerScripting
